Selection and competition of somatic mutations in normal epithelia

Michael W. J. Hall

Clare College, University of Cambridge

This thesis is submitted for the degree of

Doctor of Philosophy

September 2021

Declaration

This thesis is the result of my own work and includes nothing which is the outcome of work done in collaboration except as declared in the preface and specified in the text. It is not substantially the same as any work that has already been submitted before for any degree or other qualification except as declared in the preface and specified in the text. This thesis does not exceed the prescribed word limit of 60,000 words for the Clinical Medicine and Clinical Veterinary Medicine Degree Committee. These limits exclude figures, photographs, tables, appendices and bibliography.

Michael Hall

September 2021

Preface

All experimental data sources referred to in this thesis were generated by others. Chapters 2[*], 3[†], 4[†], and 5[‡] of this thesis contain material that has been published or made publicly available as a preprint. These manuscripts are included in the Appendix.

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

[†] Hall MWJ, Shorthouse D, Jones PH, Hall BA. Investigating structure function relationships in the NOTCH family through large-scale somatic DNA sequencing studies. *bioRxiv*. 2020.

[†] Fowler JC, King C, Bryant C, Hall M, Sood R, Ong SH, et al. Selection of oncogenic mutant clones in normal human skin varies with body site. *Cancer Discovery*. 2020.

[‡] Abby E, Dentro SC, Hall MWJ, Fowler JC, Ong SH, Sood R, et al. *Notch1* mutation drives clonal expansion in normal esophageal epithelium but impairs tumor growth. *bioRxiv*. 2021:2021.06.18.448956.

[‡] Colom B, Herms A, Hall MWJ, Dentro SC, King C, Sood RK, et al. Precancer: Mutant clones in normal epithelium outcompete and eliminate esophageal micro-tumors. (accepted for publication in *Nature*, and preprint available at *bioRxiv*, 2021:2021.06.25.449880.)

In addition, I have contributed to the following publications that are closely related to the work included in this thesis:

Kostiou V, Zhang H, Hall MWJ, Jones PH, Hall BA. Methods for analysing lineage tracing datasets. *Royal Society Open Science*. 2021;8(5):202231.

Shorthouse D, Hall MW, Hall BA. Computational Saturation Screen Reveals the Landscape of Mutations in Human Fumarate Hydratase. *Journal of Chemical Information and Modeling*. 2021;61(4):1970-80.

Kostiou V, Hall MW, Jones PH, Hall BA. Different responses to cell crowding determine the clonal fitness of p53 and Notch inhibiting mutations in squamous epithelia. (accepted for publication in *Journal of The Royal Society Interface*, and preprint available at *bioRxiv*, 2020)

Colom B, Alcolea MP, Piedrafita G, Hall MWJ, Wabik A, Dentro SC, et al. Spatial competition shapes the dynamic mutational landscape of normal esophageal epithelium. *Nature Genetics*. 2020.

Murai K, Skrupskelyte G, Piedrafita G, Hall M, Kostiou V, Ong SH, et al. Epidermal Tissue Adapts to Restrain Progenitors Carrying Clonal p53 Mutations. *Cell Stem Cell*. 2018;23(5):687-99.e8.

Abstract

Tumourigenesis occurs when a series of genome alterations occur in the same group of cells and cause uncontrolled cell proliferation. Therefore, to understand the journey from healthy to cancerous tissue, it is important to study the accumulation and spread of mutations in pre-cancerous normal tissues. Recent studies have shown that apparently normal epithelium contains a high burden of mutations in cancer-associated genes. This thesis explores the behaviour of mutant clones in normal epithelium and how this affects cancer development.

The nature of mutant clonal growth and competition in normal epidermis has been a subject of debate. A study found that mutant clone sizes inferred from DNA sequencing of normal human eyelid skin were consistent with a mathematical model of neutral cell dynamics, appearing to contradict a genetic analysis of the same dataset that found several genes under positive selection. I investigate this debate using computational modelling that takes into account the tissue structure and experimental tissue-sampling methods. The results show that mutant clone sizes in skin and oesophagus are consistent with non-neutral clonal competition.

Further evidence for non-neutral selection in normal epithelium is found in patterns of mutations detected by DNA sequencing. By adapting a statistical method used for driver gene discovery, I look for enrichment or depletion of structural categories of missense mutations and find biologically meaningful patterns of selection in several proteins. The method can associate changes to protein structure or function with cell fitness, even in the absence of hotspot mutations and in the presence of passenger mutations. I demonstrate how the method is flexible and could be widely applicable, but can also produce misleading results if confounding sources of selection are not accounted for.

Clonal competition in normal oesophageal epithelium is dominated by *Notch1* loss-of-function mutations. I fit stochastic models of clonal dynamics to lineage tracing data to show that haploinsufficiency greatly accelerates *Notch1* mutant expansion and that the loss of the second *Notch1* allele provides a further strong selective advantage, consistent with the high frequency of *NOTCH1* loss-of-heterozygosity events observed in human oesophagus. Finally, I examine a consequence of the spread of these highly fit mutant clones in the normal tissue. I use a mathematical model to analyse the results of a series of experiments in mutagen-treated mouse oesophagus, finding that microscopic tumours can be eliminated by highly fit clones in the surrounding normal tissue.

Acknowledgements

First of all, I would like to thank my supervisors Phil Jones and Ben Hall for giving me the opportunity to do this PhD and for their encouragement and support throughout. I would also like to thank all members of the Jones Lab and the Hall lab for letting me work with them on their exciting projects, for patiently answering my questions about biology and many other topics, and for making these labs such friendly and welcoming places to work.

Particular thanks to Jo Fowler, Emilie Abby and Tomeu Colom for letting me use their experimental data as the basis for parts of this thesis and for their collaboration and helpful discussions. Thanks also to David Shorthouse for his help and discussions on the work in chapter 3. And special thanks to Vicky Kostiou for her careful proofreading of this thesis.

Finally, thank you to my family and to Hazel, for their constant support and encouragement.

Table of contents

List of figures

List of tables

List of abbreviations and acronyms

ABC	Approximate Bayesian computation
AUC	Area under the curve
BCC	Basal cell carcinoma
CA	Cellular automata
CDF	Cumulative distribution function
CLL	Chronic lymphocytic leukaemia
CP	Cellular Potts
cSCC	Cutaneous squamous cell carcinoma
CTL	Control
DBZ	Dibenzazepine
DEN	Diethylnitrosamine
DFE	Distribution of fitness effects
EGF	Epidermal growth factor
GFP	Green fluorescent protein
LFIM	Logarithm of the first incomplete moment
LGCA	Lattice gas cellular automata
NGS	Next generation sequencing
NICD	Notch intracellular domain
NMSC	Non-melanoma skin cancer
OAC	Oesophageal adenocarcinoma
OSCC	Oesophageal squamous cell carcinoma
PTE	Proportion of tumours eliminated
ROC	Receiver operating characteristic
SCC	Squamous cell carcinoma
SEM	Standard error of the mean
SP	Single progenitor
TPM	Transcripts per million
UV	Ultraviolet
VAF	Variant allele fraction
WF	Wright–Fisher

WT	Wild type
YFP	Yellow fluorescent protein

Chapter 1

Introduction

1.1 Introduction

Epithelial tissues in the adult body are maintained by a constant turnover of cells (1). The loss of cells from a tissue (for example, by the shedding of cells from the surface of the skin) must be exactly compensated for by the proliferation of stem cells. This balance is vitally important. An excess of differentiating cells can lead to tissue failure, whereas excessive production of proliferating cells is a hallmark of cancer (2). Carcinogenesis is thought to be a multi-stage process, where a series of genome alterations, such as mutations, increasingly disrupt the local homeostatic balance and ultimately lead to uncontrolled cell proliferation and cancer formation (3-6).

Approximately 90% of human cancers develop in epithelial tissues that line the body's surface and internal organs, protecting them from desiccation, radiation, and mechanical and chemical damage (7). Ongoing cell division and exposure to mutagens such as ultraviolet (UV) light or tobacco smoke leads to the accumulation of mutations in the DNA of epithelial stem cells (8). Two well-studied examples of epithelial tissues are the skin (epidermis) and the lining of the oesophagus. Both are squamous epithelia (made up of cells with a flattened shape) with a high rate of cell turnover (7). In recent years, DNA sequencing of aging normal healthy human skin and oesophageal epithelium has found that these tissues carry a high burden of somatic mutations (9-12). However, the tissues retain their normal appearance and function, meaning that the mutations present have not substantially altered the overall balance of cell production and loss.

The effect that mutations have on cell behaviour in the normal epidermis has been a subject of debate (9, 13-15). In particular, the debate centres on whether mutations are merely "passengers" that do not alter the usual rules of cell dynamics that govern the homeostatic maintenance of the tissue, or whether mutations can promote excess proliferation of the cells that carry them, driving the expansion of mutant clones. A greater spread of mutant cells would increase the chance of acquiring a combination of mutations that is oncogenic (3, 4, 16, 17). The presence of multiple expanding mutant clones within the same tissue may also lead to mutant clonal competition as clones collide and compete for the limited space in the proliferative compartment of the epithelium (18, 19). If cell dynamics are altered by mutations, it might be possible to develop interventions that manipulate cell dynamics or clonal competition to encourage the growth of benign clones and reduce the spread of more tumorigenic mutations (20). It is therefore important to study the dynamics and competition of mutant cells in normal epithelia. My thesis focuses on these processes in two epithelial tissues: the oesophageal epithelium and the skin.

1.2 Squamous epithelia and cancer

1.2.1 Stratified squamous epithelia

Both the epidermis and the oesophageal epithelium are stratified squamous epithelia. They consist of layers of densely packed sheets of cells, where the bottom layer sits on a basement membrane (Figure 1.1).

Mouse oesophageal epithelium is an architecturally simple and highly uniform tissue, containing no glands or papillae to interrupt the epithelial structure (21, 22). It is therefore an ideal experimental system for studying cell dynamics. Human oesophageal epithelium has more cell layers than mouse (22, 23) and contains tall papillary structures, which divide the basal layer into two zones: interpapillary and papillary (23). The epidermis has a slightly more complex architecture, with sweat glands and hair follicles punctuating the layers of keratinocytes (24, 25). Stem cells from hair follicles and sweat glands can contribute to the repair of the epidermis following injury, but do not play a role in the routine maintenance of the stratified epithelium (24-26).

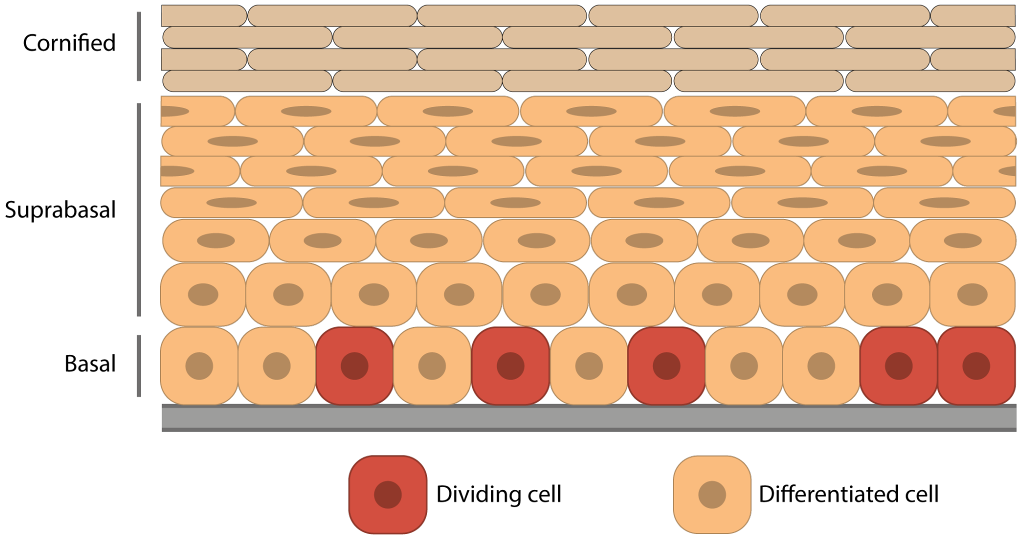

Figure 1.1 The structure of stratified squamous epithelium. Proliferation is constrained to the basal layer. Differentiated basal cells stratify into the suprabasal layers and migrate towards the tissue surface, where they are shed.

Although the oesophageal epithelium and epidermis differ in terms of architecture, they share a common mechanism of tissue maintenance. Cell division is restricted to the basal layer (21-23, 27, 28) (Figure 1.1). Once a basal cell has committed to differentiation, it stratifies out of the basal layer and migrates up through the suprabasal layers to the tissue surface. During the differentiation process, cells become wider and flatter and – except for those in the human oesophagus – lose their nuclei (Figure 1.1) (24, 29). Fully differentiated cells are shed from the tissue surface (21, 27), meaning a constant turnover of cells is required to maintain the tissue. In homeostasis, there is a balance between the quantity of cells shed from the tissue surface and the number of cells produced by proliferation in the basal layer (21, 27, 28).

1.2.2 Oesophageal and skin cancer

Oesophageal cancer is the eighth most common form of cancer and the sixth most common cause of cancer death (22). There are two main types of oesophageal cancer: oesophageal squamous cell carcinoma (OSCC) and oesophageal adenocarcinoma (OAC) (30). OSCC accounts for most cases of oesophageal cancer worldwide, although the incidence of OAC is rising and has overtaken the incidence of OSCC in Europe and North America (30). OAC is associated with gastroesophageal reflux disease (GERD), and is thought to develop from gastric cardia cells via a precursor lesion called Barrett's oesophagus (30, 31). As the name suggests, OSCC develops from the squamous epithelium that lines the oesophagus.

There are three major types of skin cancer: melanoma, basal cell carcinoma (BCC) and cutaneous squamous cell carcinoma (cSCC) (32). BCC and cSCC account for approximately 75% and 25% of non-melanoma skin cancer (NMSC) respectively (33, 34), and NMSC of any kind is approximately 20 times more common than melanoma (32). The primary environmental risk factor for skin cancer is exposure to UV light (32).

Cancer driver mutations produce a growth advantage in the cells that carry them and hence promote cancer development. Mutations in the genes *TP53* and *NOTCH1* are frequently detected in DNA-sequenced samples of squamous cell carcinoma, from both oesophagus and skin (33, 35-37). Therefore, based on the sequencing of tumour samples alone, both *TP53* and *NOTCH1* appear to be drivers of SCC. However, analysis of mutations in normal tissue suggests that each of these genes has a very different relationship with cancer, as discussed in the following section.

1.2.3 Mutations in normal tissues

DNA sequencing has revealed that normal human skin and oesophagus harbour large numbers of mutant clones, including many mutations in cancer genes (9-12). Similar studies have detected somatic mutations in many other normal tissues, including blood (38-40), endometrium (41, 42), colon (43, 44), small intestine (44), bladder (45, 46), liver (44, 47) and lung (48). The mutation burden of healthy tissues increases with age (10, 12, 44).

Of particular interest is the prevalence of cancer-associated mutations in normal tissues. For example, *TP53* was found to be mutated in 5–20% of the cells in normal oesophagus and in around 90% of OSCC tumours (10). This observation is consistent with the traditional multi-stage model of carcinogenesis: mutating *TP53* takes the normal tissue a step further towards cancer. In contrast, *NOTCH1* mutations occupy 30–80% of the healthy oesophageal tissue in middle-aged and older individuals, but only around 10% of OSCC tumours are *NOTCH1*-mutant (10). This raises the question whether *NOTCH1* is truly a driver of oesophageal cancer, or whether the ability of *NOTCH1* mutants to spread in the normal tissue merely means these mutants have an increased likelihood of appearing in a tumour as a passenger mutation (49).

The fact that *NOTCH1* mutations are less prevalent in tumours than in the normal tissue may even suggest that *NOTCH1* mutations have a protective effect against cancer (10).

Mirroring these human studies, studies in mouse skin and oesophagus have shown that mutant clones can spread extensively in the normal epithelia without causing major disruption to the tissue. It has been found that mutations in cancer driver genes such as *Trp53* and *Notch1* initially provide a strong competitive advantage over wild type cells (19, 50). However, the growth rate of the mutant clones slows over time due to cell crowding, competition with rival mutant clones or reaching the limits of the tissue capacity (19, 50). DNA sequencing following treatment with a chemical mutagen has also shown that the tissues can withstand a high mutation burden with no substantial changes to histology or tissue function (19, 51).

1.3 Methods for studying squamous epithelia

1.3.1 Lineage tracing experiments

The behaviour of cells in the epidermis and oesophageal epithelium can be studied using transgenic lineage tracing experiments. These experiments genetically label cells of interest in the tissue and track their progeny. A major benefit of this technique is that the cell behaviour is studied in the cells' native environment (52).

The mice used for transgenic lineage tracing contain two genetically modified genes. The first of these genes expresses the enzyme Cre recombinase, which is often fused to a mutant oestrogen receptor (ERT). The CreERT fusion protein is restricted to the cytoplasm until administration of the drug tamoxifen, which allows CreERT to enter the cell nucleus. There, Cre excises a stop cassette from the second modified gene, triggering the expression of a fluorescent label such as green fluorescent protein (GFP) (52) (Figure 1.2). The DNA of the labelled cell is altered, meaning that the progeny of the labelled cell also express the fluorescent protein. If low doses of tamoxifen are administered, only a small proportion of cells in the tissue are induced to express the label, and the descendants of individual labelled cells can be tracked (52) (Figure 1.2). In this thesis, these cell lineages derived from single cells are referred to as clones.

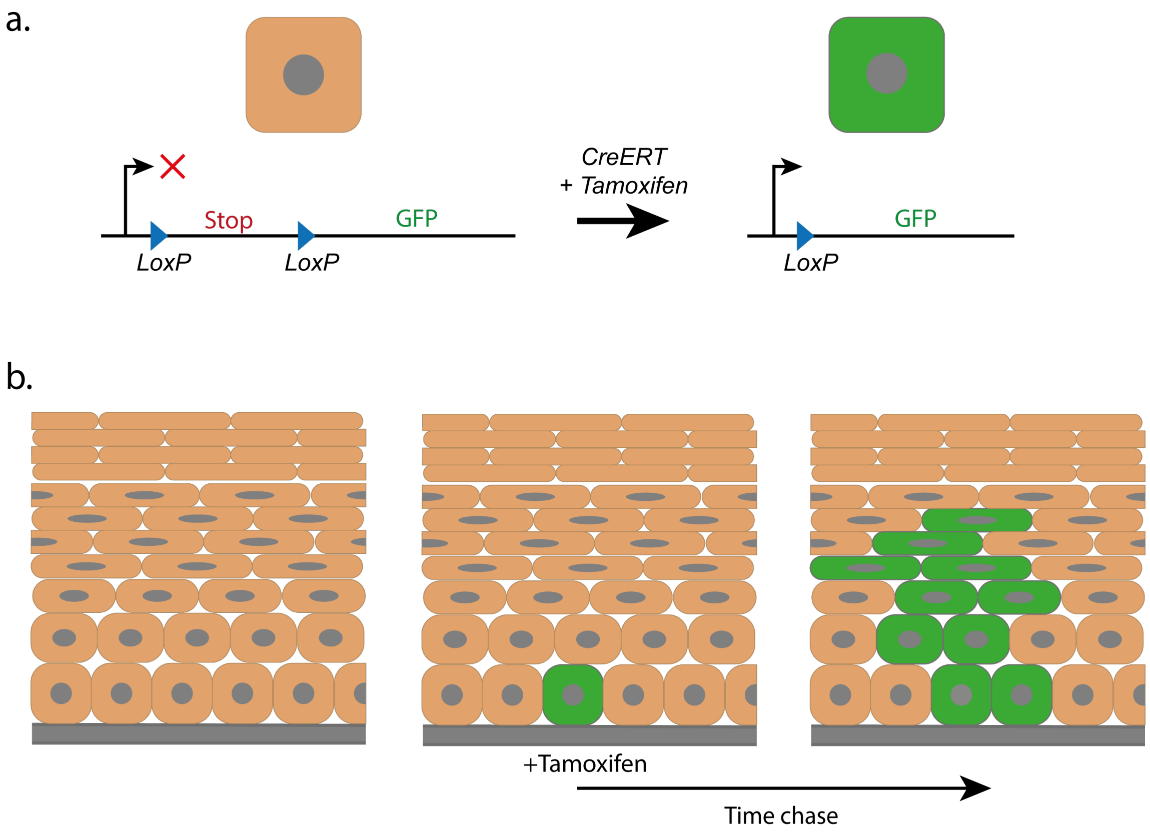

Figure 1.2 Transgenic lineage tracing. a) Administration of the drug tamoxifen allows Cre recombinase (fused to a mutant oestrogen receptor, ERT) to enter the cell nucleus. Cre excises a LoxP-flanked stop cassette from a second transgenic gene, triggering the expression of a fluorescent labelling protein such as GFP. This excision alters the DNA of the labelled cell, meaning that the cell's descendants also express GFP. **b)** If a low dose of tamoxifen is administered, only a small proportion of cells are induced to express the fluorescent label. The descendants of the original labelled cells can then be observed at a later time. Figure adapted from (52).

The fluorescent labels do not alter the wild type behaviour of cells (21, 27, 28). However, additional genetic mutations can be induced alongside the fluorescent label, enabling the growth of mutant clones to be studied (19, 50).

Modern live-imaging lineage tracing techniques can be used to directly observe each cell division outcome and track individual clones over time (53, 54). Unfortunately, in traditional lineage tracing, animals must be culled to observe the clone sizes, meaning that only a single snapshot of each clone can be obtained. To infer the rules governing cell behaviour, theoretical models of cell dynamics can be used to analyse the clone size distributions collected at various timepoints following induction (55) (section 1.4).

1.3.2 DNA sequencing

By its nature, transgenic lineage tracing can only be performed in genetically modified organisms. However, the sizes of mutant clones can also be inferred through DNA-sequencing experiments. Next generation sequencing (NGS) technology hugely increased the capacity of DNA sequencing compared with Sanger sequencing, as well as greatly reducing its cost (56). However, the improvement in scale came at the cost of sequencing accuracy (56).

Multiple NGS technology platforms are available (57). The sequencing methods used in these platforms differ in many details, but the general workflow is similar (57). DNA is extracted from the tissue sample and broken up into many short fragments (Figure 1.3a). Depending on the sequencing method, the DNA fragments may be amplified (duplicated many times) or particular regions of the genome may be targeted. The DNA fragments are attached to microbeads or a glass slide to be read by the sequencing machine (57). Only a fraction of the DNA fragments attach to a bead or the slide, meaning that not all the DNA from the original sample is sequenced. The sequencing outputs millions of short "reads", which are then processed by a bioinformatics pipeline to identify ("call") the mutations present in the tissue sample (Figure 1.3b).

The reads are mapped to the location on the reference genome that best matches the read sequence. Multiple reads are required at each targeted location in the genome so that mutations can be identified. For germline mutation calling, variants are identified as differences between the reference genome and the sequenced reads. On chromosomes with two copies, these germline mutations will either be homozygous (on both alleles, and therefore present in all sequenced reads at a location) or heterozygous (on one allele only, and therefore present in half of the sequenced reads). For somatic mutation calling, (at least) two tissue samples are taken (for example, a tumour sample and a normal blood or muscle sample), and the sequences of the two samples are compared to identify somatic variants (Figure 1.3b). Somatic mutations may only occur in a subset of the cells in a tissue sample. The fraction of sequenced reads containing the mutation will therefore depend on both the fraction of sampled cells carrying the mutation and whether the mutation is homozygous or heterozygous (and any copy number changes) (9).

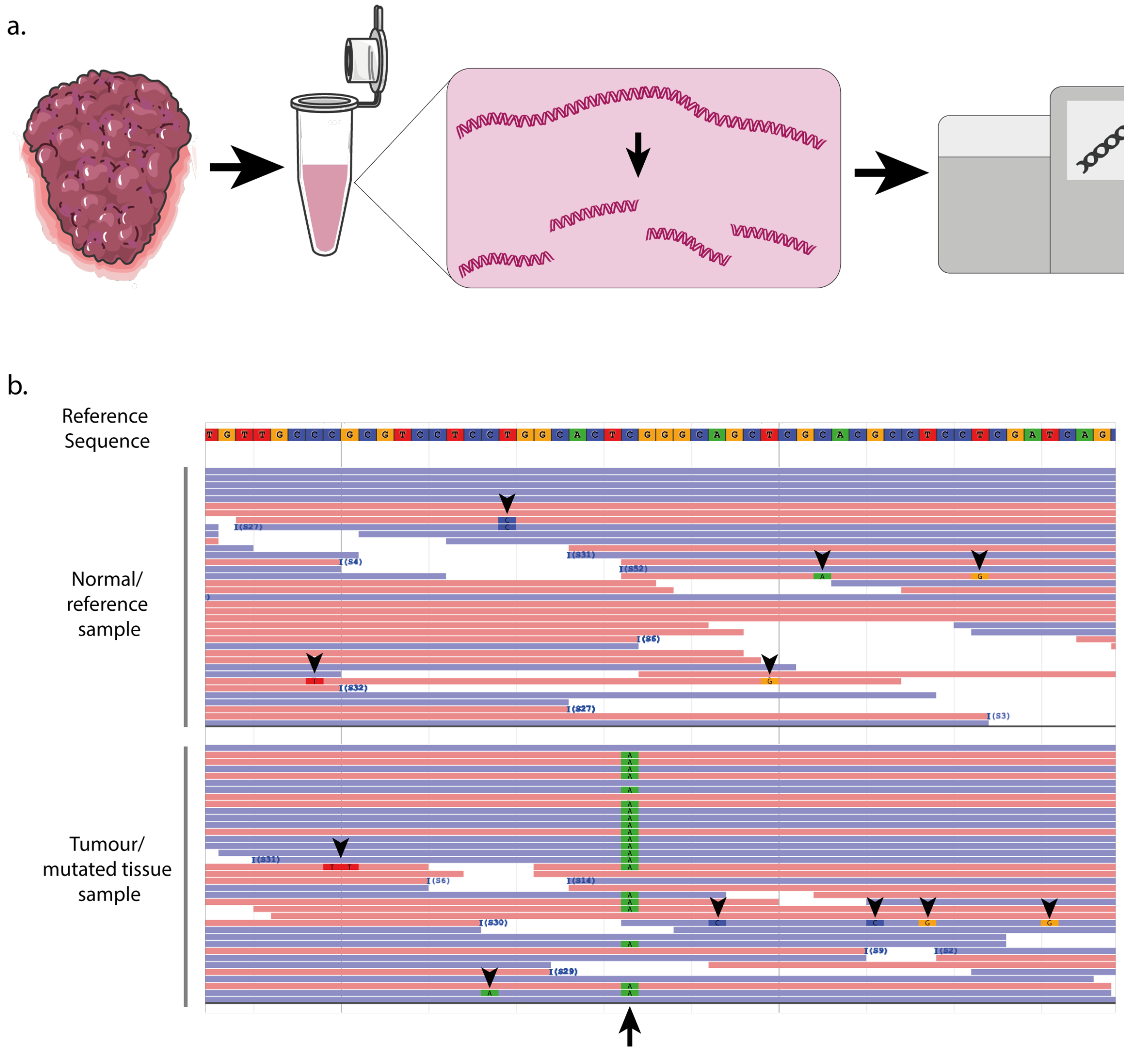

Figure 1.3 Next generation DNA sequencing. a) DNA is extracted from the tissue sample and broken into short fragments for sequencing. Images from smart.servier.com, licensed under CC BY 3.0, edited from the original. **b**) The sequences of the DNA fragments are mapped to a reference genome. Blue bars are DNA reads aligned to the forward strand, and red bars are reads aligned to the reverse strand. The highlighted bases (green, red, blue or orange) differ from the reference genome sequence. For analysis of a tumour or mutated normal tissue sample, a sample of normal, non-mutated tissue is taken from the same individual for comparison. Bases that frequently differ between the two samples are identified as somatic mutations (e.g. the mutation of C>A in the tumour sample, indicated with an arrow). Bases that differ only infrequently may be sequencing artefacts (arrowheads). Image generated using JBrowse (58)

Copy number changes are additional or missing copies of chromosomes or chromosome sections in a cell's DNA. Changes to copy number can alter the number of copies of DNA that a mutation appears on. Therefore, to estimate the fraction of sampled cells containing the mutation, the local copy number should also be considered (10). In normal human oesophagus, *NOTCH1* is frequently affected by copy-neutral loss-of-heterozygosity (LOH) events, but copy number changes affecting other genes are rare (10). Therefore, except for *NOTCH1* LOH in human oesophagus samples (Chapter 2), I have estimated mutant clone sizes in this thesis using the assumption of wild type copy number.

Errors can occur during various stages of DNA processing and sequencing (56). Error rates in NGS are far higher than those in the "first generation" technology of Sanger sequencing (57). To increase sensitivity to detect somatic mutations and filter out false calls, variant calling methods use the information from multiple reads covering the same genomic location, along with information regarding the background error rate at that location (59). However, because not all DNA in the tissue sample is sequenced, and because at low frequencies it is harder to distinguish a real mutation from a sequencing artefact, mutant clones occupying only a very small proportion of the tissue sample can be difficult to detect. The smallest size of mutant clone that can be reliably detected will depend on both the coverage of the DNA sequencing and the size of the tissue sample (9). New sequencing protocols are being developed to increase the sensitivity of mutation calling (60); however, this thesis does not use any data generated by these new protocols.

1.3.3 Mutational signatures

There are many processes that can introduce mutations into the DNA of cells in the body. These include errors made during the DNA replication of cell division, and mutations caused by exposure to mutagens – for example tobacco smoke or UV light (8). Different mutational processes produce different patterns of mutations, known as signatures. Single base substitution signatures have been identified using the trinucleotide context of the substitutions (including the bases immediately 5' and 3' of the mutated nucleotide) (Figure 1.4) (8). The pattern of mutations observed in a tissue sample is a combination of the signatures of all the mutational processes that have affected that tissue (Figure 1.4) (8).

The trinucleotide context is not the only factor that affects the likelihood of a nucleotide substitution (61). Nonetheless, the use of mutational spectra to model expected mutation patterns has been found to improve the statistical detection of driver genes (section 1.3.4 below) (62, 63).

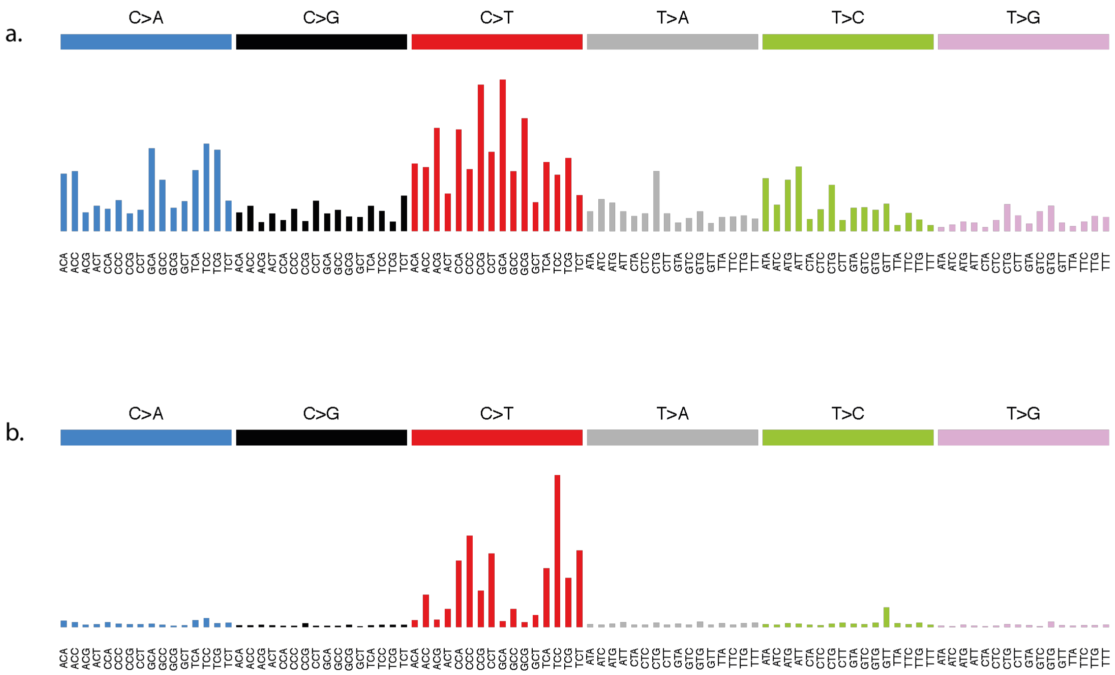

Figure 1.4 Mutational spectra. a) The trinucleotide mutational spectrum for normal human oesophagus (10). **b)** The trinucleotide mutational spectrum for normal human skin (11). The high numbers of C>T mutations are due to the mutational signature of ultraviolet light.

1.3.4 Identification of driver genes

In somatic evolution, ongoing genetic mutation (section 1.3.3) creates phenotypic diversity, which is then subject to natural selection (4, 17, 64). Positive selection increases the frequency of genotypes that convey a proliferative or survival advantage to a cell, and negative selection reduces the frequency of genotypes that are deleterious. Driver mutations provide the cells carrying them with a growth advantage and they occur, by definition, in “driver genes” (17). The term is usually used to describe genes that are frequently mutated in cancer and are thought to play a role in tumourigenesis or tumour expansion (17). However, driver mutations can also promote clonal expansion in normal tissues, and these are not necessarily the same mutations that promote the growth of cancers (10, 12, 65).

Cancer driver genes can be broadly categorised into two groups: tumour suppressors and oncogenes (66). Tumour suppressors protect against cancer by slowing cell division, repairing DNA or triggering cell death (66). Oncogenic mutations in tumour suppressors are those that inactivate the gene (66). In contrast, oncogenes promote cell growth; consequently, oncogenic mutations in oncogenes are those that increase the activity of the gene (66). Many tumour suppressor genes are recessive, i.e. both copies of the gene must be mutated before the mutant phenotype is observed (5, 67, 68). In contrast, many oncogenes are dominant, i.e. an activating mutation in a single allele of an oncogene is sufficient to produce the driver phenotype (67, 69). Many genes, however, do not fit neatly into these tumour suppressor or oncogene paradigms (67, 68).

Identifying driver genes is crucial for understanding the causes of cancer and for developing targeted cancer treatments; consequently, much research has been dedicated to developing methods for driver gene detection. Many of the methods start from a "neutral" null hypothesis model, which uses knowledge of patterns of mutation accumulation (section 1.3.3) to predict the mutations that would be expected to appear in a gene, assuming that no selection, positive or negative, was acting upon the gene. The observed data from DNA-sequencing studies is then statistically compared with the null model to identify genes in which the pattern of mutations significantly deviates from the expected neutral pattern. In studies of cancer, the focus is usually on detecting positively selected genes, as these may be the "drivers" of the cancer. Various types of mutational patterns can signal that a gene is not neutrally selected (Figure 1.5), which has led to the development of a variety of driver detection methods.

One commonly used method is dN/dS. Originally used to analyse the evolution of protein-coding genes (70), it has been adapted for use in the somatic context (62). It looks at the ratio of non-synonymous mutations (dN) to synonymous (dS) mutations. It is assumed that synonymous mutations have no effect on gene function and are therefore under neutral selection. If the dN/dS ratio is significantly higher than expected under the neutral hypothesis, i.e. if there is an excess of non-synonymous mutations, this indicates positive selection of the protein-altering mutations in the gene (Figure 1.5). A low dN/dS ratio signifies negative selection (Figure 1.5).

A similar approach compares observed and expected scores of mutational functional impact (63, 71). These scores are generated using machine-learning approaches to estimate which mutations will have a highly disruptive impact on gene function (e.g. nonsense mutations or mutations in highly evolutionarily conserved positions) or a low functional impact (e.g. synonymous mutations). A bias towards or away from high functional impact mutations in the observed data indicates positive and negative selection respectively (Figure 1.5). The principles behind the functional impact and dN/dS methods are discussed in more detail in Chapter 3.

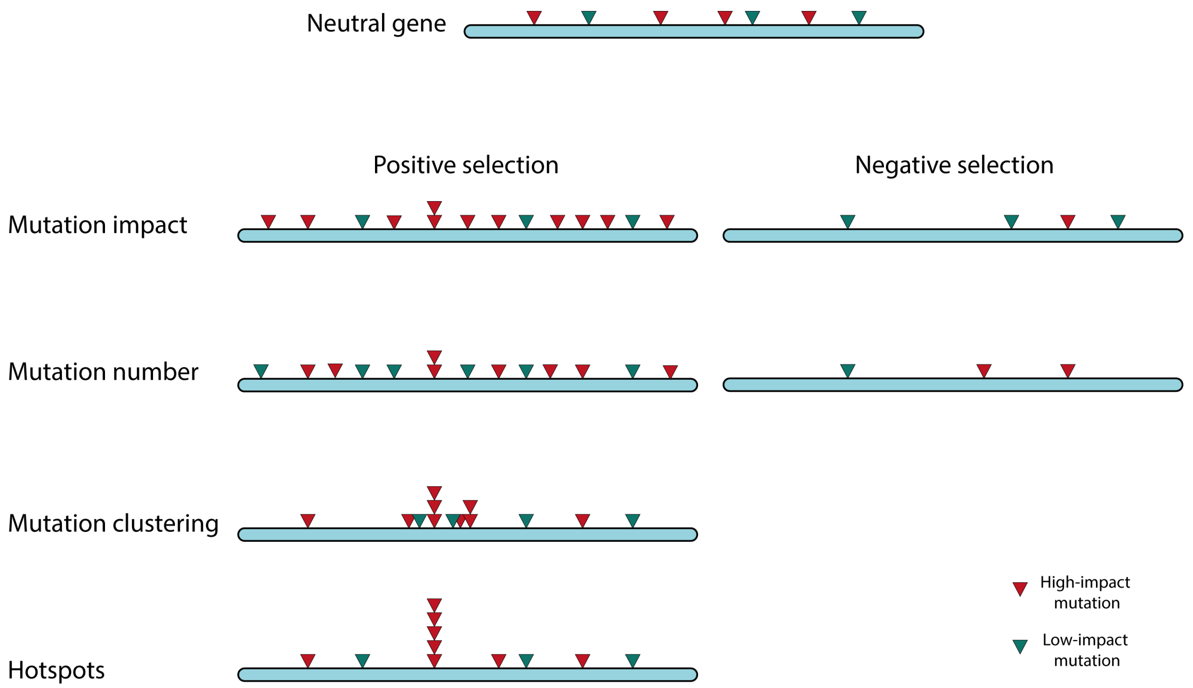

Figure 1.5 Approaches for driver gene detection. The neutral null hypothesis estimates the pattern of mutations that is expected to appear in the gene if it is not under positive or negative selection (top). Positive selection in a gene may be indicated by a relative excess of high-impact mutations (e.g. nonsense/missense mutations or high estimated functional impact) compared with low-impact mutations (synonymous mutations or low estimated functional impact); an overall excess of mutations; clustering of mutations either linearly in the gene sequence or in the 3D protein structure; or high-frequency hotspot mutations. Negative selection may be indicated by a lack of high-impact mutations relative to low-impact mutations, or an overall lack of mutations.

Other driver detection methods calculate an expected mutation rate for each gene based on a combination of factors, including the number of silent mutations, chromatin state and level of transcription (72, 73). In these methods, positively selected driver genes are identified by the presence of a larger number of mutations than expected by chance (72, 73) (Figure 1.5).

In some driver genes, the selected mutations cluster in key functional areas, meaning these genes can be identified by detecting linear clusters of mutations in the gene sequence (74), enrichment of mutations in particular protein domains (75), clusters of mutations in the 3D structure of the protein (76), or frequently mutated hotspots (77) (Figure 1.5). If the mutation clusters occur in well-characterised domains, these methods can associate selection with particular functional changes to a protein (78).

A combination of approaches can be used to detect a more comprehensive set of driver genes. This is achieved either by taking the union of the driver genes detected by each method, or by using the various data sources as inputs to a machine-learning algorithm (67, 79, 80).

1.4 Modelling cell dynamics in epithelia

Mathematical and computational models of epithelial cell dynamics have been developed to investigate how homeostasis is maintained and how mutant clones spread in the tissue. By comparing model predictions with experimental observations, it is possible to infer the rules that govern cell behaviour. In this section I describe and compare the three main stochastic models that I use in this thesis, and briefly discuss alternative models that have previously been applied to the epithelial tissue.

1.4.1 Single progenitor model

Lineage tracing experiments (section 1.3.1) in mouse skin and oesophagus have shown that the growth of wild type clones is consistent with a model of cell dynamics in which the tissue is maintained by a single population of functionally equivalent progenitor cells (21, 27, 28, 81, 82). Due to the assumption that the tissue only contains one type of dividing cell, as opposed to a hierarchy of multiple dividing cell types, this model is sometimes referred to as the single progenitor (SP) model. In this model, the basal layer of the epithelium consists of two types of cells: progenitor cells and differentiated cells (Figure 1.1). The differentiated cells do not divide, and will stratify into the suprabasal layers. Each progenitor cell divides to form either two progenitor cells, two differentiated cells, or one cell of each type (Figure 1.6a). The outcome of each cell division is random. This leads to a variance in clone sizes, even though all clones originate from functionally equivalent cells. Some clones lose all dividing cells and are lost from the tissue, while other clones expand (Figure 1.6b). This variance due to the stochastic nature of the system is known as drift (81). In homeostasis the proportions of the two symmetric division types are balanced so that the total cell population and proportion of each cell type are maintained.

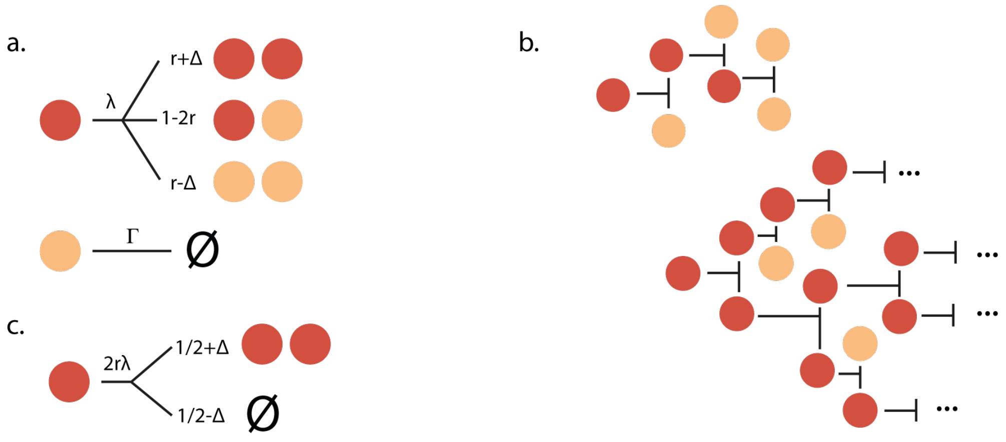

Figure 1.6 Single progenitor model. a) Progenitor cells (red) can divide symmetrically to produce two progenitor cells, two differentiated cells (beige) or asymmetrically to produce one cell of each type. The parameter λ is the division rate. The parameter r is the probability of each symmetric division type in neutral growth. The parameter Δ determines the imbalance between the production of dividing and non-dividing cells. For neutral growth, Δ=0. Differentiated cells stratify into the suprabasal layers at rate Γ. **b)** Through random chance, neutral clones can be lost when all cells in the clone differentiate (top) or they can grow to large sizes (bottom). **c)** Simplified single progenitor model that only considers changes to progenitor cell numbers.

The SP model reproduces the key hallmarks of lineage tracing data of neutral clones. The total number of labelled cells remains constant (Figure 1.7a), the number of surviving clones falls (Figure 1.7b) and the mean size of surviving clones increases linearly (Figure 1.7c). The clone size distributions widen over time and have a characteristic long-term scaling pattern (Figure 1.7d,e).

For the SP model, explicitly modelling clone sizes including both the progenitor cells and differentiated basal cells requires the use of either a complex solution to the master equations (83) or simulations that include both cell types (84). Alternatively, the full basal clone sizes can be approximated by calculating the clone size distribution including only the proliferating cells and scaling this distribution by a factor representing the proportion of differentiated basal cells (27). This simple scaling relationship between the full clone size distribution and the progenitor-cell-only clone size distribution means that key features of the full SP model can be well approximated by a simplified model that only considers the progenitor cells (13). In this simplified model, the asymmetric divisions do not affect the number of progenitor cells and therefore do not have to be explicitly included (Figure 1.6c).

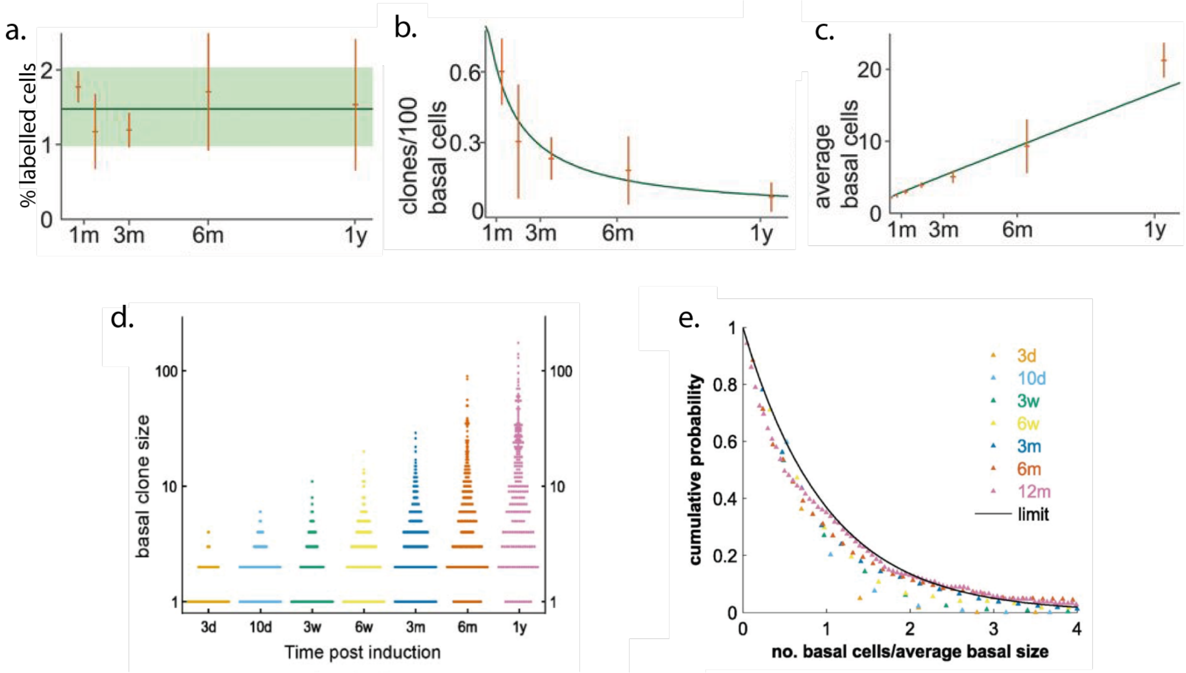

Figure 1.7 Lineage tracing hallmarks of the neutral single progenitor model. The average proportion of labelled cells is constant over time. Orange markers and error bars show observed data and SEM; the green line and shaded region show the mean and SEM across all time points. **b,c)** The number of surviving clones falls (**b**) and the average surviving clone size increases linearly (**c**) over time. Orange markers and error bars show observed data and SEM; green lines show the prediction of the single progenitor model fit to the data. **d)** Basal clone size distributions widen over time. **e)** The clone size distributions exhibit long-term scaling behaviour. The cumulative clone size distribution is plotted as a function of the clone size divided by the average clone size for each time point. Figures taken from (21).

1.4.1.1. More complex model variations

Various models of epithelial cell dynamics involving proliferative cell hierarchies or multiple stem cell populations have been proposed for mouse epithelium (85-87). However, the cell dynamics in these models are still assumed to be stochastic, and the clone size predictions are similar to those of the SP model (84). An analysis of multiple lineage-tracing experiments in murine epidermis and oesophagus epithelium found that, with the exception of the inter-scale compartment of tail skin, the SP model fits the data at least as well as the more complex models do (84).

1.4.1.2. Non-neutral growth in the single progenitor model

Mutations may disrupt the balance between proliferation and differentiation (Figure 1.6a). Mutants that tip the scales towards division will cause clones to (on average) expand exponentially (88). However, this simple model contains no mechanism to limit the cell population, and any imbalance towards proliferation will therefore cause the total cell population to expand indefinitely (Figure 1.9e). Consequently, it may not be an appropriate model for normal tissues where mutant clones expand without altering the overall tissue architecture. This topic is discussed in more detail in Chapter 2.

1.4.1.3. 2D single progenitor model

The rules of the SP model have been applied to 2D sheets of cells representing the basal layer of epithelium. However, this has led to some unbiological behaviour in the model. In the SP model, the outcome of each cell division is determined independently, i.e. fates of neighbouring cells are not coordinated. This means there is no mechanism to ensure an even spread of dividing cells in the tissue, and patches of high and low cell density occur (89). The variation in cell density observed in the 2D SP models was greater than the density variation observed in microscopy images of murine oesophageal tissue (89). In high-density regions, dividing cells are surrounded by other dividing cells and have nowhere to place offspring. In low-density areas, differentiated cells that are not adjacent to a dividing cell leave the tissue without being replaced. To counteract this, feedbacks can be added to bias the cell fate towards maintaining local cell density (89). Although this is biologically plausible, it adds complexity to the model.

1.4.1.4. Single progenitor model summary

This model can recreate the key hallmarks of neutral clonal growth in mouse epithelia, and was crucial for showing that cell dynamics are consistent with a stochastic system rather than a deterministic stem cell–transit-amplifying cell system (27). However, when applied to the spatial tissue structure, or to mutant cells with growth biases, the SP model predicts outcomes that are not consistent with the biology of normal epithelial tissues.

Therefore I have also considered the Moran (90) and Wright–Fisher models (91). These models originated in the study of species evolution and are conceptually simple, share the stochastic nature of the SP model and are readily applicable in 2D space.

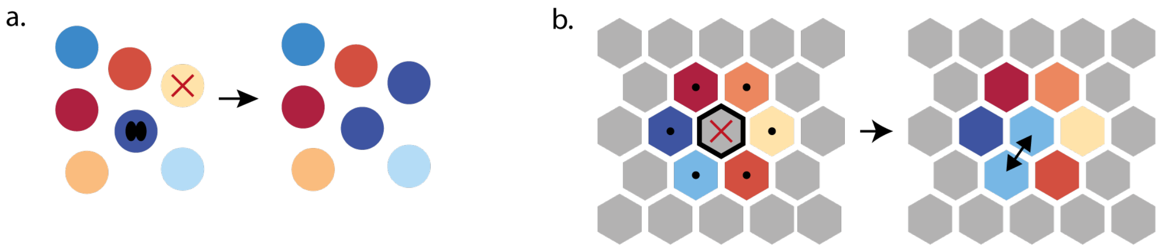

Figure 1.8 Moran model. a) One step of the Moran model. The pale-yellow cell dies and the dark-blue cell divides so the total population size remains constant. **b)** The Moran model in a 2D grid. A cell is selected to die (thick black border with red X) and a dividing cell is selected from the immediate neighbours (marked with black dots). The new cell takes the place of the dead cell.

1.4.2 Moran model

This model assumes there are a fixed number of cells in the population. At each step in the process, one cell is randomly selected to divide, and a different cell is randomly selected to "die" (Figure 1.8a). This means that the overall cell population is maintained, while the sizes of individual cell lineages can vary. The cells in this model can be thought of as equivalent to the progenitor cells in the SP model, and a similar scaling factor can be applied to account for the differentiated cells in the basal layer (51). Alternatively, the cell population in the model can be assumed to represent all cells in the basal layer, and cell loss in the model can be interpreted as the stratification of differentiated cells out of the basal layer. For large population sizes (where clone sizes are far smaller than the total population size), the results of the neutral Moran model are almost identical to the results of the neutral SP model (Figure 1.9a-d).

1.4.2.1. Selection in the Moran model

Each cell in the model has a "fitness". A higher cell fitness *relative* to the rest of the population will increase the probability that the cell will be selected to divide. Mutations can provide a growth advantage by increasing the fitness of a cell (which is then inherited by the cell's offspring).

At the initial stages of mutant clone growth, when the majority of the tissue is still wild type, the Moran model results are very similar to the results of the non-neutral SP model (if the equivalent parameter values are found in each model) (Figure 1.9e). However, for larger clone sizes or where multiple mutant clones coexist in the tissue (chapters 2 and 5), the results of the two models differ (Figure 1.9e). This is because of the fundamental difference between the SP model and the Moran model: that in the Moran model the key fate-determining property of a cell is its relative fitness, whereas in the SP model the cell fate bias is absolute, and each cell division is considered independently. The expansion of a mutant clone in the SP model has no effect on any other clone, and there are no external factors that can limit the expansion. In contrast, all cells in the Moran model are competing. The fixed population size means that the expansion of one clone must be compensated for by the loss of cells from other clones.

As the mutant clone approaches the size of the full population, its growth slows (Figure 1.9e). The chance that a wild type cell is randomly selected to die reduces as the wild type proportion of the population decreases. Therefore, an increasing number of division/death events involve a mutant cell replacing a mutant cell and thus the overall mutant cell population size does not change. The growth observed is sigmoidal (Figure 1.9e) and is similar to the long-term growth of mutations that expand without fundamentally altering the structure of the normal tissue (50).

The cell competition seen in the Moran model also has consequences for tissues in which multiple clones are present. The survival and growth prospects of a clone depend not only on the phenotype of the clone itself, but also on the phenotypes of the other clones in the surrounding tissue (Chapter 2) (51).

1.4.2.2. 2D Moran model

The Moran model can be easily adapted to run on a 2D grid (92). Each location in the grid is occupied by a cell. At each step of the process a cell is selected to die and is removed from the simulation, leaving an empty space in the grid. A cell from the immediate neighbourhood of the empty space is then selected to divide and fill the vacated grid location with a copy of itself (Figure 1.8b). The probability of the cell being picked to divide depends on its fitness relative to the other cells in the neighbourhood.

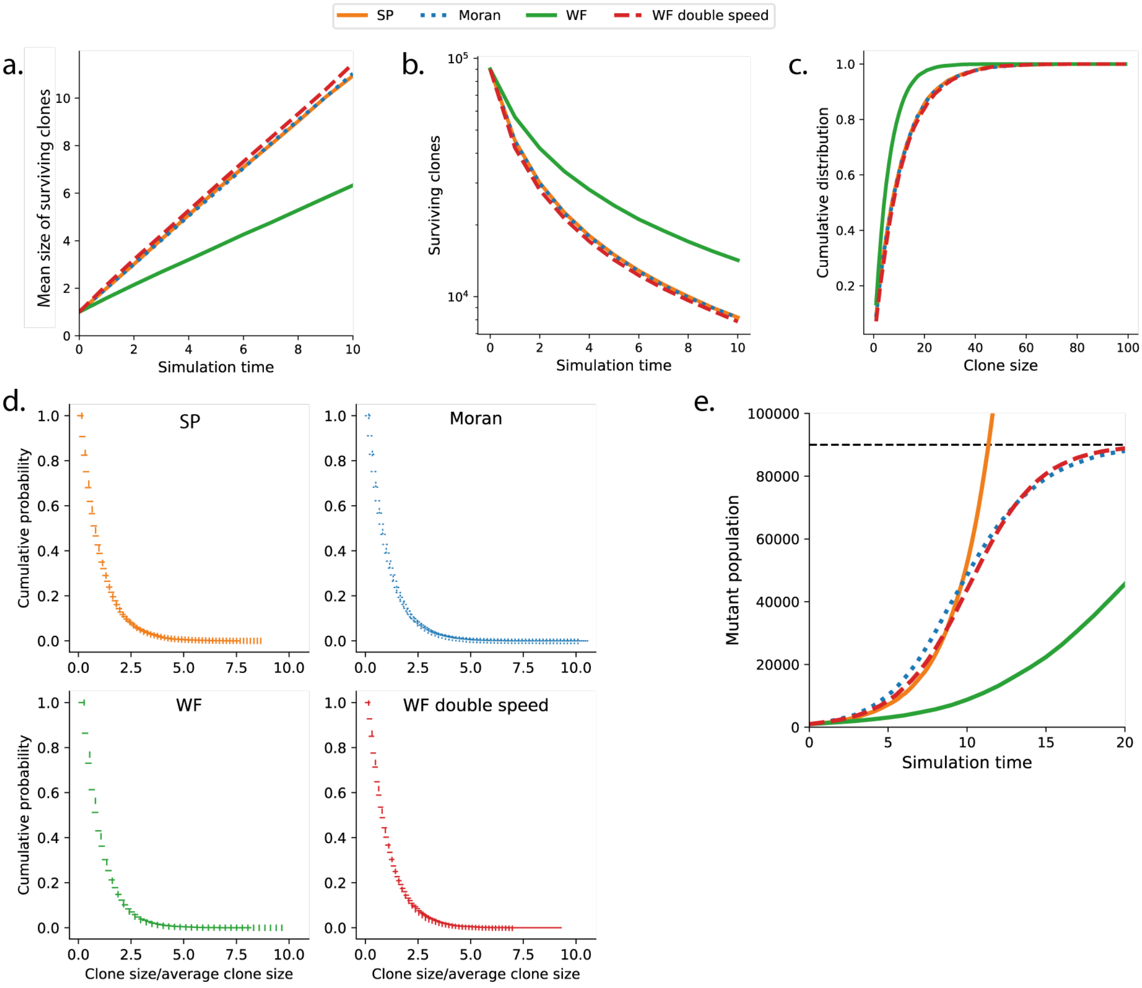

Figure 1.9 Comparison of outputs of non-spatial stochastic models. In all panels, SP refers to the simplified single progenitor model containing only progenitor cells (Figure 1.6c). Except for the WF double speed simulations, the average division rate is once per time unit. For the WF double speed simulations, the division rate was twice per time unit. **a-d)** Neutral simulations of 90000 clones starting from single cells. **a)** Mean size of surviving clones increases linearly. **b)** The number of surviving clones reduces over time. **c)** The SP model, the Moran model and the Wright–Fisher model (with division rate doubled to account for the difference in drift) produce highly similar clone size distributions. **d)** Scaling behaviour of the clone size distributions. The shape of the cumulative clone size distribution plotted as a function of the clone size divided by the average clone size remains approximately constant over time. Scaled cumulative clone size distributions are shown at two time points for each simulation: vertical lines represent the distribution after 5 time units; horizontal lines represent the distribution after 10 time units. **e)** Non-neutral growth starting from 1000 mutant cells. The horizontal dashed line shows the total cell population (90000) in the Moran and WF simulations; the non-mutant cells had fitness 1. In the SP model, the mutant had parameter Δ=10%; in the Moran simulation the mutant had fitness 1.5; in the WF simulations the mutant had fitness 1.25. SP, single progenitor. WF, Wright–Fisher.

At later time points, the growth of neutral clones in the 2D Moran model is slightly slower than that in the non-spatial Moran model (Figure 1.10a). This is because a clone can only grow at its boundary – for a clone to expand, one of the clone cells must divide and replace a cell that is not in the clone. As the clones grow larger, the proportion of cells on clone boundaries reduces, meaning a higher proportion of divisions involve a cell replacing another cell from the same clone. The rate of drift is therefore slightly reduced because a greater proportion of divisions do not alter clone sizes (Figure 1.10a-c).

This boundary growth in clones in the 2D Moran model also means that long-term non-neutral clonal expansion is slower than in a non-spatial Moran model (Figure 1.10e,f). The reduction in drift also reduces the loss of clones in the 2D Moran model. Together, this means that in simulations with multiple expanding non-neutral clones, the 2D Moran model results in a population made up of a larger number of smaller mutant clones than in the equivalent non-spatial Moran model (Figure 1.10e,f).

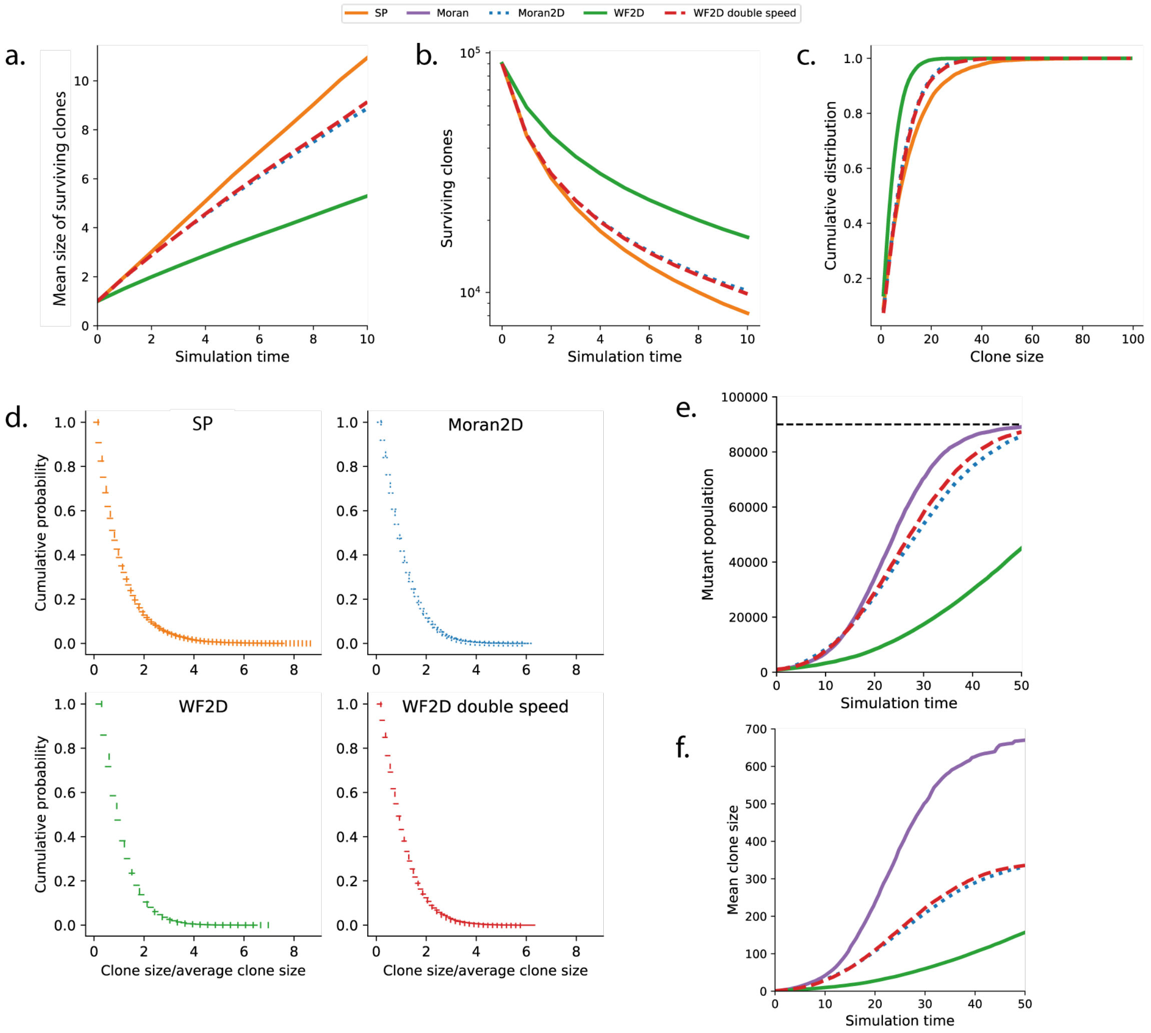

Figure 1.10 Comparison of output of spatial stochastic models. a-d) Neutral simulations in 300×300 cell hexagonal grids. A non-spatial SP model simulation is shown for comparison. Descriptions as in Figure 1.9a-d. **e-f)** Non-neutral simulations. Growth of 900 evenly-spaced single mutant cells in 300×300 cell hexagonal grids in 2D Moran and 2D WF models. Non-spatial Moran model shown for comparison. The horizontal dashed line shows the total cell population (90000) in the simulations; the non-mutant cells had fitness 1. The mutant had fitness 1.5 in the Moran and 2D Moran models and 1.2 in the WF2D models. SP, single progenitor. WF, Wright–Fisher.

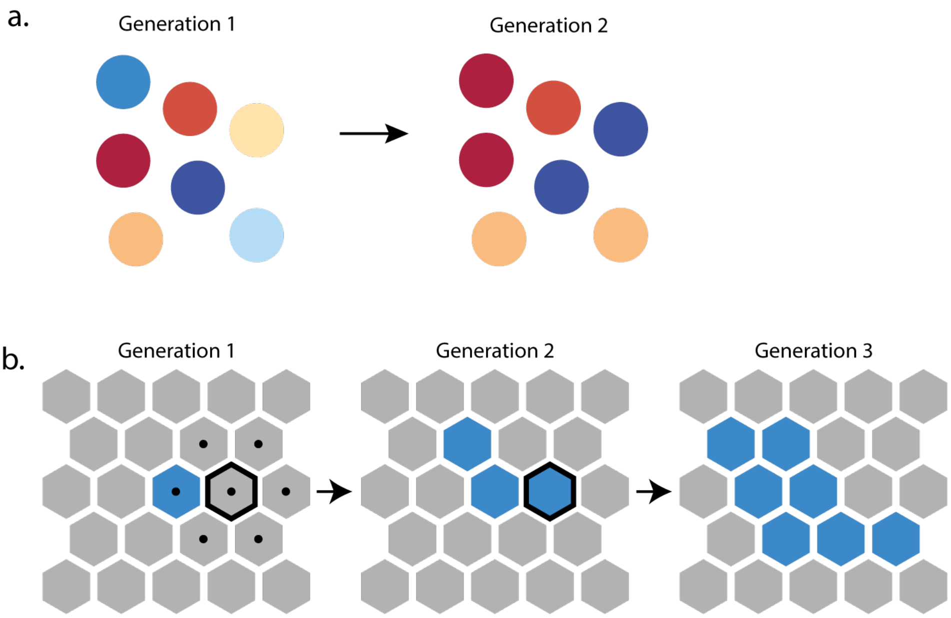

Figure 1.11 Wright–Fisher model. a) One step in a Wright–Fisher model. Each cell in Generation 2 picks a parent cell at random from the cells in Generation 1. **b)** The 2D Wright–Fisher model. Each cell in Generation 2 picks a parent cell from the neighbouring cells in the previous generation. The potential parent cells of the cell with the black border in Generation 2 are shown with black dots in Generation 1. A cell with a fitness advantage (blue cells) has a larger probability of being selected as a parent and therefore has a higher chance of expanding into a large clone.

1.4.3 Wright–Fisher model

The Wright–Fisher (WF) model is very similar to the Moran model, but instead of individual cells dividing one at a time, the entire population of cells is replaced in a single step to create the next generation of cells (91). In the new generation, each cell randomly "picks its parent" from the cells in the previous generation (Figure 1.11a).

Drift in the WF model is half that of the Moran model (90). This means that if the time between generations in the WF model is equal to the average division time in the Moran model, the growth of neutral clones (which depends on the drift rate) is halved. By doubling the division rate of the WF model, this produces the same trends as those observed for the Moran model (Figure 1.9a-d).

1.4.3.1. Selection in the Wright–Fisher model

The WF model has a similar mechanism of selection to the Moran model. The fitter a cell is relative to the rest of the population, the more likely it will be picked as a parent by cells in the next generation. When Moran and WF simulations are run using the same (non-neutral) value as the fitness parameter, the two models have different outputs. However, since this parameter does not have a direct biological counterpart, the actual value of the fitness parameter itself is not of great importance. If the appropriate corresponding fitness parameters are used, and the WF model division rate is double that of the Moran model, the outputs of the Moran and the WF models are very similar (Figure 1.9e).

1.4.3.2. 2D Wright–Fisher model

The WF model can be adapted to run on a 2D lattice following the same logic as the Moran model. A cell picks its parent from its immediate neighbourhood in the previous generation (Figure 1.11b). The neutral and non-neutral 2D WF models produce outputs very similar to the equivalent Moran processes (Figure 1.10). Again, to match the rate of drift between the two models, the division rate of the 2D WF model must be double that of the 2D Moran model.

1.4.4 Alternative spatial models of epithelial cell dynamics

In addition to the models described above, many other methods have been applied to study cell dynamics in epithelia. Each modelling approach will be best suited to studying a particular aspect of the tissue, from small-scale mechanical forces to large-scale cell density, and will strike a different balance between model complexity and computational efficiency.

Continuum models do not model individual cells; instead, cell boundaries are ignored and continuous variables are used to describe properties of regions of cells (93). These models can be used to study phenomena such as mechanical stress, migration, proliferation and cell adhesion (93), and are suited to large-scale analysis of tissue. However, because individual cells are not explicitly modelled, continuum models are less suitable for studying mutations or heterogeneity in the tissue (93, 94).

Cell-based models represent cells as discrete entities and generally model cell division, death and migration as stochastic processes (95). They are broadly split into two categories – lattice-based models and off-lattice models. Lattice-based models (such as the Moran and WF models described above) use a mesh or grid to represent the tissue. The simplest lattice-based models are cellular automata (CA), in which each lattice site can hold a single cell. The cells follow a set of rules that allow them to divide, die or migrate to other lattice sites. Even from simple rules, highly complex behaviour can emerge (96). Several variations of CA models have been used to model skin and oesophageal epithelium, with a variety of rules applied to model wild type and mutant cell behaviour (51, 86, 92, 97-99).

Other lattice-based models include lattice gas cellular automata (LGCA) and cellular Potts (CP) models. In LGCA models, each site in the lattice can be occupied by multiple cells. The number of cells within each lattice site and moving between each lattice site is tracked (95). LGCA models allow efficient simulation of large numbers of cells but lack the single-cell resolution of CA models (95). In CP models, individual cells occupy multiple contiguous lattice sites, which enables the modelling of individual cell morphology and mechanical properties of the system (100, 101). However, CP models are far more computationally expensive than CA models (102).

Off-lattice models do not restrict cells to fixed positions on a grid. The cells are free to move and interact based on the forces that the cells exert on each other (102). These models can simulate cells by tracking either the position of the cell centre (centre-based models) (51, 102, 103) or the boundaries of cells (vertex-based models) (102, 104), and are suited to studying the mechanics and cell morphology of the tissue (102).

1.4.5 Modelling summary

Biology is very complex, and theoretical models are necessarily simplified and abstracted by comparison (105-107). It is important to use models with an appropriate level of abstraction so that the model is complex enough to recreate the key behaviours of the true system, but simple enough that the model can be interpreted and used to make inferences and predictions (106, 108). In this thesis I err on the side of simplicity: I do not look to add complexity in order to match every detail of the experimental data, but instead use simple models that can illustrate the general principles underlying clonal dynamics.

The SP model provides a good fit to wild type clonal dynamics (84) and early exponential-like expansion of mutant clones (19, 50), but it is less appropriate for modelling the long-term expansion and competition of mutations (19) (Chapter 2). The Moran and WF models can simulate both non-neutral and spatial competition of mutations while maintaining the fixed population of the tissue. This assumption of a fixed cell population is appropriate for the modelling of normal epithelial tissues because homeostasis in wild type tissue means the cell population remains constant (21, 27), and even with a high burden of mutations, the normal tissues only display small increases in cell density and tissue thickness (10, 11, 19, 51).

The WF and Moran models are very similar. The major advantage of the WF model over the Moran is that it can be several orders of magnitude faster to run than the Moran model (Table 1.1). The high speed of the WF simulations enables the use of parameter-fitting methods that require large numbers of simulations to be run (Chapter 5).

Table 1.1 Timings of simulations. Neutral competition starting from 90,000 single-cell clones (in 300×300 hexagonal grids for the 2D simulations) over 25 average cell divisions (Moran) or 50 cell generations (WF). All simulations were written using the same underlying Python framework and were run on a single core of the same laptop.

	Moran	**WF**
Non-spatial	487s	0.25s
2D	79s	0.94s

1.5 Aims of the thesis

Mutations are widespread in normal epithelium of the skin and the oesophagus. The nature of somatic evolution and mutant clonal competition in these normal tissues has been a subject of debate, with the discussion centred around whether cell competition is neutral or non-neutral. This project had two main aims: first, to examine the evidence for non-neutral selection of mutations in DNA-sequencing datasets of normal human skin and oesophageal epithelium (chapters 2, 3 and 4); and second, to investigate the consequences of mutant clonal expansion and competition, in particular the impact on the earliest stages of cancer formation (Chapter 5).

My thesis is organised as follows:

In Chapter 2, I examine the debate regarding the neutrality or non-neutrality of mutations in human eyelid skin. Specifically, I use stochastic simulations of clonal competition that take into account the tissue structure and the experimental tissue-sampling method, enabling me to demonstrate how the experimental clone size distributions and the genetic analysis of driver genes are both consistent with non-neutral competition. I further validate the conclusions using a dataset of mutations from normal human oesophagus.

In Chapter 3 I introduce a statistical method for testing selection of functional categories of mutations and validate the method using *NOTCH1* mutations in normal skin and oesophagus.

In Chapter 4, I use the method described in Chapter 3 to examine patterns of mutations in five further genes sequenced in the normal human skin and oesophagus. I use these genes to demonstrate various benefits and limitations of the method, and present an extension to the method that enables comparison of selection between datasets. The results of chapters 3 and 4 provide further evidence of non-neutral clonal competition in the normal skin and oesophageal epithelium.

In Chapter 5 I investigate the expansion of super-competitive mutant clones in the mouse oesophagus. Firstly, I fit a stochastic spatial model of clonal growth to lineage tracing data, to show that cell fitness is strongly dependent on the level of *Notch1* signalling and that the haploinsufficiency of *Notch1* greatly enhances the spread of high-fitness *Notch1* loss-of-function mutant clones. I then use a mathematical model to analyse the disappearance of micro-tumours from mutagen-treated mouse oesophagus, showing that the tumours can be eliminated by highly competitive mutant clones in the surrounding tissue.

Chapter 2

Relating evolutionary selection and mutant clonal dynamics in normal epithelia

2.1 Introduction

Approximately 90% of human cancers arise from squamous epithelia (7), which consist of layers of keratinocytes (109) (Chapter 1, section 1.2.1). Cells are continually shed from the tissue surface and replaced by proliferation in the basal cell layer (Figure 2.1a). The proliferating cells accumulate mutations over time, and may generate mutant clones (110). If such clones persist within the otherwise normal tissue, they may acquire the multiple genomic alterations that lead to cancer (110). A key question is whether these large, persistent clones arise by neutral competition or are a consequence of cancer-associated mutations increasing the competitive fitness of mutant cells above that of wild type cells. If the former, little can be done to alter the risk of cancers emerging from mutant clones in normal tissue. However, if the founding clones of cancers emerge by competitive selection, it is possible that interventions that alter the fitness of mutant cells may decrease cancer risk. A dataset of mutant clones detected in normal human eyelid skin appears to contain conflicting evidence supporting both neutral and non-neutral mutant cell dynamics (9, 13). This "paradox" has yet to be resolved (14, 15), leading to uncertainty over the somatic mutant cell dynamics in normal epithelial tissues (111). My aim with this work was to unpick this apparent inconsistency.

The dataset in question is from a study of mutations in normal human eyelid skin epidermis (9). DNA was extracted from small samples of epidermis. About 500 DNA molecules of each targeted gene were sequenced and compared to the genome of the same tissue donor (Figure

2.1b). Somatic mutations were detected as altered sequences present in one or more samples (9). The proportion of altered DNA reads containing a mutation, the variant allele fraction (VAF), was assumed to be proportional to the size of the mutant clone (Figure 2.1b).

An analysis of inferred sizes of the mutant clones in the human eyelid data argues for neutral dynamics (13). Observed clone sizes can be compared to the predictions of candidate mathematical models of cell dynamics to determine the best fitting model (81). Lineage tracing experiments in homeostatic, unmutated mouse epidermis and oesophagus suggest that these tissues are maintained by a single population of equipotent progenitor cells (Figure 2.1c) (19, 21, 27). The outcome of individual cell divisions is unpredictable but on average 50% of the progeny of dividing cells differentiate, exit the proliferative compartment and are eventually shed from the tissue while 50% remain to divide again. Such balanced stochastic cell fate leads to wide variation in clone sizes while the total cell population remains constant. Mutant clone sizes observed in the human eyelid skin have been compared to predictions from this neutral stochastic model (13).

The comparison is made using the first incomplete moment of the clone size distribution (Chapter 2 methods), which has been used in several studies to shed light on mutant clone growth dynamics (13, 88, 98, 112, 113). In economics the first incomplete moment is used to study inequality – the value of the incomplete moment at £X shows how much of the wealth is held by those with a fortune of £X or higher (114). It has a similar role for the clone size analysis – it shows the proportion of the mutated cells that are in clones of size *x* or larger. Using the first incomplete moment has two advantages over using the clone size distribution directly. Firstly, the first incomplete moment reduces the fluctuations caused by low sample size (88) and secondly it simplifies the comparison of data to the neutral theory. The neutral model predicts that the first incomplete moment of mutant clone sizes will have a negative exponential form (13) – where there are many small clones and few large clones. A deviation from the exponential shape could indicate non-neutral competition – some clones have expanded to take over more than their expected share of the tissue. The logarithm of an exponential curve forms a straight line. Therefore it is proposed that there will be a clear distinction between the logarithm of the first incomplete moment (LFIM) from neutral and non-neutral competition: neutral competition will lead to a straight line whereas non-neutral competition will be indicated by a curved/kinked line (13). Using this criterion, the inferred mutant clone sizes from the human eyelid appear largely consistent with the neutral model (13).

However, the theory of neutral competition of cancer-associated mutations appears to be incompatible with results from several mouse and human studies that observed non-neutral mutant expansions in normal epithelial tissues (19, 50, 88). Furthermore, signs of non-neutral clonal competition in the eyelid mutational data can be detected using dN/dS analysis, a method from population genetics. This examines the ratio of protein-altering mutations (dN) to silent mutations (dS) for each gene (62). Once relevant corrections have been applied, a dN/dS ratio of 1 is indicative of neutral behaviour. A dN/dS value of less than 1 indicates the mutated gene has a negative effect on the competitive fitness of mutant cells compared with normal cells. However, if a disruption of the protein provides a growth advantage to the cell, then the number of protein-altering mutations that expand to a clone of detectable size will be increased, leading to a dN/dS ratio greater than 1. Analysis of the human eyelid mutations found 6 of the 74 sequenced genes had significantly raised dN/dS ratios ranging from 3 to over 30, consistent with mutations in those genes driving clonal expansion (9). Additionally, protein-altering missense mutations in some driver genes, e.g. *NOTCH1*, *NOTCH2* and *TP53*, were not randomly distributed but concentrated in functional domains. This suggests positive selection of function-altering mutations, which is also incompatible with neutrality (9).

In this work I show that in some conditions *non-neutral* competition can produce a straight-line LFIM, and therefore a straight LFIM alone is not a clear indicator of neutral dynamics. This work can thus reconcile the observed clone size distributions and positive genetic selection.

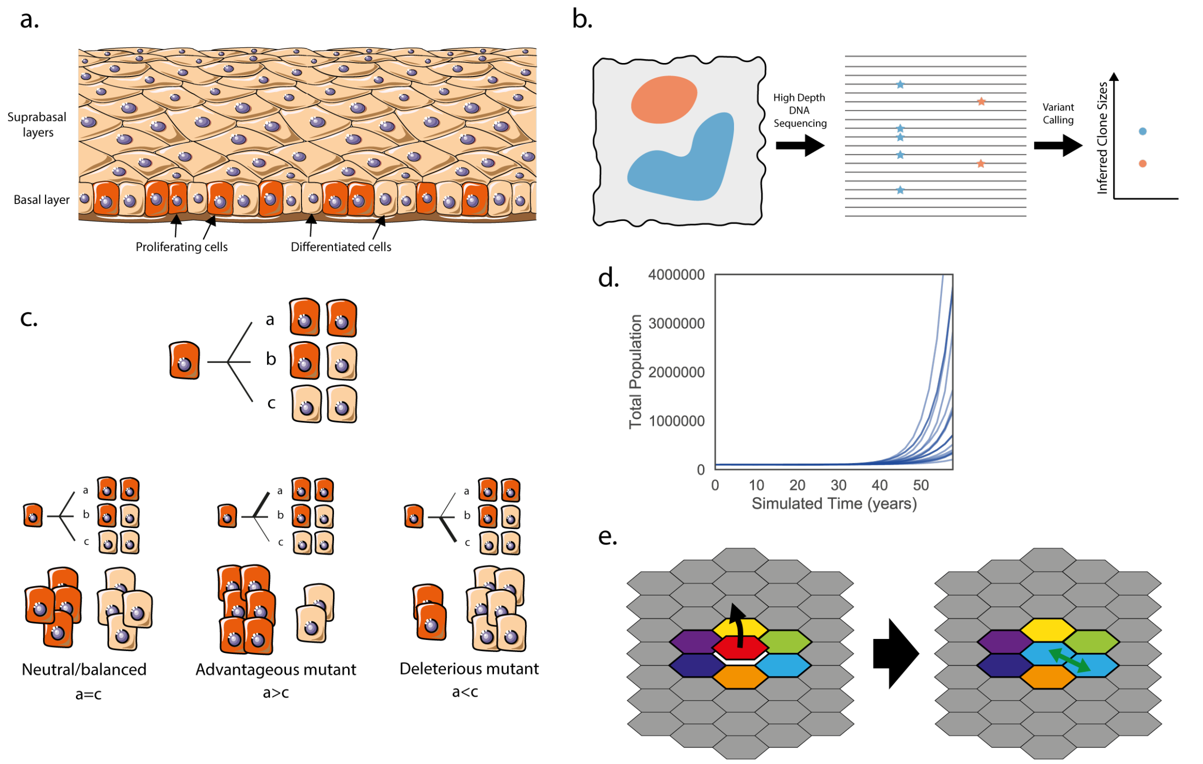

Figure 2.1 Data collection and cell dynamics. a) Proliferation occurs in the basal layer of the epithelium. After differentiation, cells migrate through the suprabasal layers before being shed. Image from smart.servier.com, licensed under CC BY 3.0, edited from original. **b)** DNA from a biopsy (left) containing mutant clones (red and blue) is sequenced and the variant allele fraction (proportion of reads containing the mutation, middle) used to infer clone sizes (right). **c)** The stochastic single progenitor model of cell dynamics. Each dividing cell (red) can produce two dividing cells (a), two non-dividing differentiated cells (brown) (c), or one of each type (b). In a homeostatic tissue or neutral clone, the probabilities of each symmetric division option are balanced (left). An advantageous mutation would increase the proportion of dividing cells produced (middle), and a deleterious mutation would increase the proportion of differentiated cells (right). Note that in the non-neutral case, the probabilities of each division type do not have to be fixed over time, but can depend on the cell context. **d)** Simulation of the model shown in c). If mutations introduce perpetual positive fate imbalances then the population will eventually explode. Total population of 20 simulations with mutations introducing only small fate imbalances drawn from N(mean = 0.25%, std = 1.25%). **e)** In the spatial Moran process, a differentiating cell (red) is replaced by the division of a neighbouring cell (light blue).

2.2 Results

I will start by considering several important constraints that apply to the mutant clones in normal epithelia. Firstly, the cellular structure and composition of the tissue remains at least approximately constant. Secondly, the proliferative compartments of the epidermis and oesophageal epithelium contain few barriers to mutant clone expansion. In a tissue with a high burden of mutations like the human eyelid, this means that expanding clones will soon collide and compete with each other as well as with unmutated cells. These two constraints were not included in the mathematical model used in the first incomplete moment analysis of the human eyelid data (13), meaning mutant clones in this model with a growth advantage (Figure 2.1c, middle) could expand without limit (Figure 2.1d).

2.2.1 Spatial constraints alter clone size distributions of non-neutral mutations

To address the effect of clonal competition I used a mathematical model drawn from the study of population genetics. I ran Moran-type simulations (90, 92) (Chapter 1, Chapter 2 methods) of cell competition on a 2D grid to represent the epidermis (Figure 2.1e). Cells lost through differentiation are replaced by the division of a neighbouring cell (Figure 2.1e), similar to behaviour observed in mouse epidermis (115). During each division, there is a small chance that one of the daughter cells will acquire a mutation (Chapter 2 methods).

Simulations of a 2D neutral model produced an approximately straight LFIM (Figure 2.2a). A non-neutral spatial model in which a small proportion of mutations change the fitness of a cell (Chapter 2 methods) may deviate from a neutral appearance by curving away from the straight line (Figure 2.2b). This is due to the contrast between the relatively large non-neutral clones and the smaller clones growing neutrally. Surprisingly, however, simulations with a higher proportion of non-neutral mutations may generate a straight line (Figure 2.2c). This is because almost all of the simulated tissue is taken over by non-neutral clones (Figure 2.2d). The only neutral mutations that persist are those that occur in clones with non-neutral mutations, carried as “passengers”. As all remaining clones exhibit similar behaviour, the LFIM is straightened. This shows that a straight-line LFIM does not necessarily imply neutral competition and is consistent with positive dN/dS ratios (Figure 2.2e). I concluded that since there is a high burden of mutant clones in the eyelid (9), the tissue is likely to be extensively colonised by non-neutral mutant cells, contributing to the apparently neutral appearance of the clone size distribution.

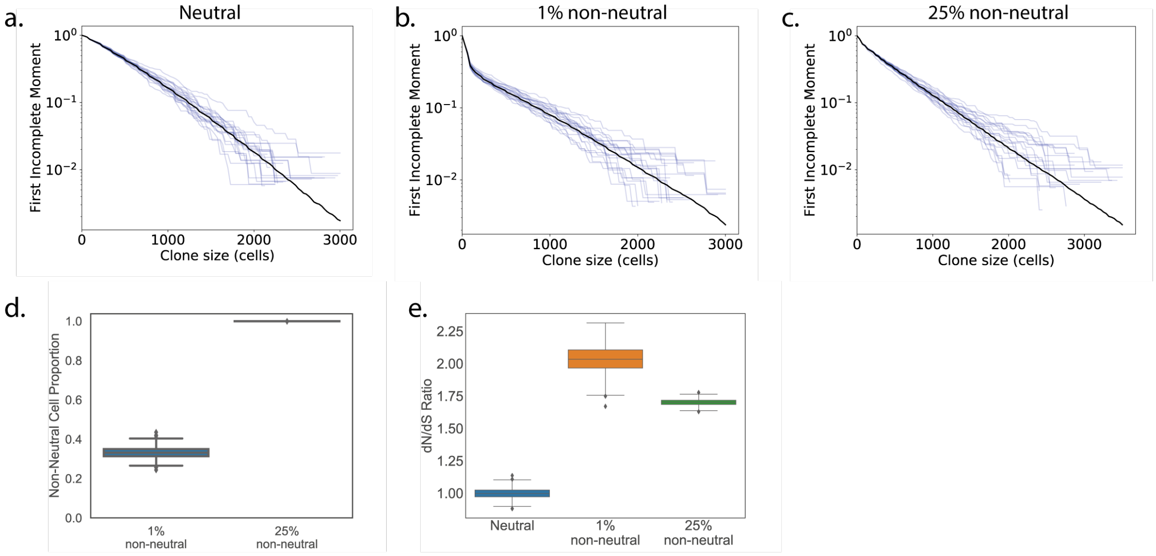

Figure 2.2 2D simulations of clonal competition. a-c) First incomplete moments of 2D simulations (Chapter 2 methods). The average of 1000 simulations is shown in black, a selection of 20 individual simulations is shown in blue. **a)** Neutral simulations. **b)** Simulations where 1% of mutations are non-neutral. A deviation from the straight line is seen at clone sizes of approximately 100 cells. **c)** Simulations where 25% of mutations are non-neutral. **d)** Proportion of cells at the end of the simulations with a fitness altered by non-neutral mutations. In the 25% non-neutral simulations, by the end of the simulation almost the entirety of the tissue has been colonised by non-neutral mutant clones. **e)** dN/dS values from the simulations shown in a-c). To enable this calculation for the neutral simulations, a proportion of neutral mutations were labelled as non-neutral but did not alter cell fitness.

2.2.2 Impact of sampling methods on measurement of clone size distributions

Another consideration that may impact the measurement of clone size distributions and hence inference of mutant clone dynamics is experimental design. In the human eyelid study, spatially separated tissue samples were collected (Figure 2.3a) (9). The area of the sample defines an upper limit on the reliable estimation of clone size. In the eyelid experiment, the area of each sample was less than that of the largest clones. The lower limit of clone size detection is also related to the sample area, since mutations present in only a small fraction of the cells in the sample may not be detected due to the technical noise in DNA sequencing (59). I simulated the combined effects of spaced samples in which only clones occupying 1% or more of the area of the sample can be detected (Chapter 2 methods). Figure 2.3b,c,d show these effects on the first incomplete moments of the simulations from Figure 2.2a,b,c respectively. The results lead me to conclude that these experimental factors may artefactually reduce a deviation of LFIM from a straight line caused by non-neutral competition (Figure 2.2b, Figure 2.3c).

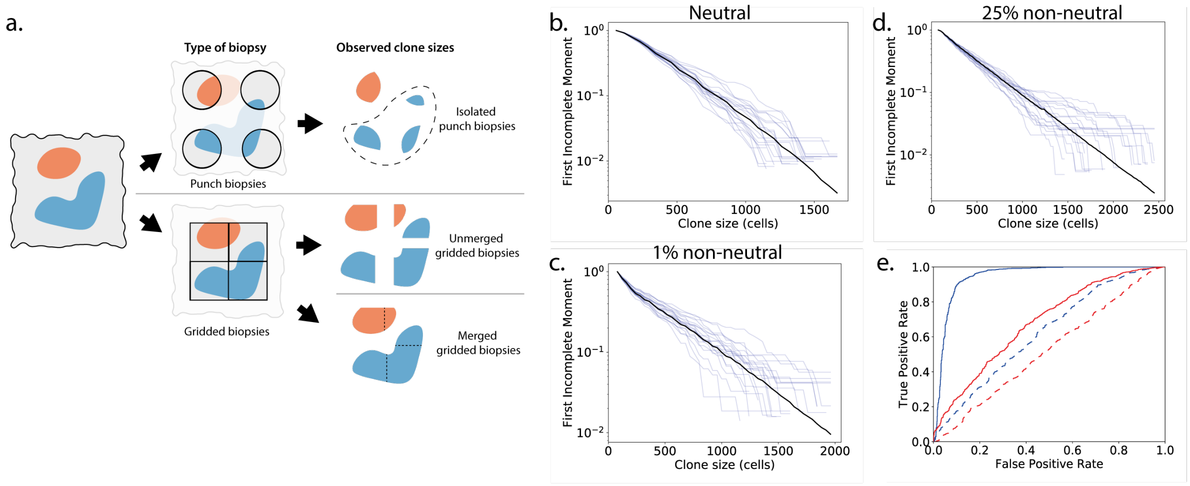

Figure 2.3 First incomplete moments of 2D simulations with biopsy sequencing. a) The method of taking biopsies can affect the observed mutant clone size distribution. Isolated punch biopsies (top) may not capture the entirety of a mutant clone; in the analysis in (13), clones which spanned multiple biopsies (shown in the dashed area) were excluded. Ungapped gridded biopsies (bottom) enable the reconstruction of larger clone sizes. **b-d)** The simulations from Figure 2.2a,b,c respectively, with the effects of isolated punch biopsies and sequencing. **e)** ROC curves using R^2 of the log first incomplete moment of the clone size distribution as the classifying statistic. Red, simulated biopsy and sequencing; blue, full data randomly subsampled to match biopsy plus sequencing sample sizes. Solid, 1% non-neutral; dash, 25% non-neutral. Area Under the Curve (AUC) is a measure of how successful the classifier is at distinguishing the two groups. A perfect classifier will have an AUC of 1. A random guess will have an AUC of 0.5. AUCs: Full data, 1% non-neutral, 0.94; biopsy plus sequencing, 1% non-neutral, 0.68; full data, 25% non-neutral, 0.62; biopsy plus sequencing, 25% non-neutral, 0.52.

2.2.3 Ability of LFIM to resolve neutral competition versus selection

I next tested how well the LFIM could discriminate between the neutral and non-neutral simulations using the coefficient of determination, R^2, to measure the straightness of a line (Chapter 2 methods). Although R^2 has documented weaknesses (116, 117) and has been found to perform poorly in a similar context (118), I used it here in order to be consistent with previous studies and to evaluate the LFIM test of neutrality as it was originally proposed (13, 98). For the LFIM to be a successful indicator of neutrality, the neutral simulations need to have a higher R^2 than the non-neutral simulations. Receiver operating characteristic (ROC) curves showed the accuracy of the LFIM as a test of neutrality depended on both the underlying shape of a non-neutral clone size distribution and on the experimental sampling method (Figure 2.3e). For example, using the LFIM was little better than a random guess (area under the curve, AUC = 0.52) when attempting to distinguish the neutral simulations from Figure 2.3b from the non-neutral simulations in Figure 2.3d, where the underlying shape of the non-neutral LFIM was largely straight (Figure 2.2c) and the clones sizes were inferred from simulated DNA sequencing of spatially separated biopsies.

2.2.4 Sensitivity of the model to the distribution of fitness effects

The distribution of fitness effects (DFE) of the mutations can affect the shape of the LFIM of mutant clone sizes, and therefore the ability of the LFIM to detect non-neutral competition. This is shown by the difference between including 1% or 25% non-neutral mutations in simulations (Figure 2.2b,c and Figure 2.3e). Here I briefly explore how sensitive the shape of the LFIM is to the DFE of the non-neutral mutations, and therefore whether the conclusions above were greatly influenced by the particular assumptions I have made. In particular I look at changing the shape of the non-neutral DFE, including additional deleterious mutations and changing the interaction of mutations (epistasis).

Several DFEs have been proposed based on theoretical predictions or experimental observations of mutations in evolving organisms (119). I ran simulations using three of the proposed distributions: normal (120, 121) (Figure 2.2b,c and Figure 2.3b,c), exponential (122) (Figure 2.4a,b) and uniform (123) (Figure 2.4c,d). In all of these cases the simulations produce similar shaped LFIMs.

In population genetics, it is generally assumed that a large majority of non-neutral mutations will be deleterious because an organism is already close to peak fitness (119). It is not clear to what extent this assumption applies for somatic mutations, since in healthy tissues individual cell fitness is secondary to the fitness of the organism as a whole (124). However, it may still be the case that many non-synonymous mutations are deleterious to cell fitness. I therefore ran simulations in which two thirds of non-neutral mutations reduce the fitness of a cell, but this has little impact on the shape of the LFIM (Figure 2.4e,f).

I have assumed in all cases so far that the effects of multiple mutations can be combined by simple addition of their fitness values. Here I look at diminishing returns (125) as an alternative form of epistasis. As a simple implementation of diminishing returns, I used the rule

$$new\ cell\ fitness = \text{maximum}\ (old\ cell\ fitness, new\ mutation\ fitness)$$

i.e. the fitness of a mutation will replace (rather than add to) the cell fitness if it is higher than the cell fitness, otherwise it has no effect. This rule means that a very fit clone will only rarely increase in fitness through a new mutation and the change is likely to be small. The results of these simulations are shown in Figure 2.4g,h and are similar to those using simple addition of mutation fitness.

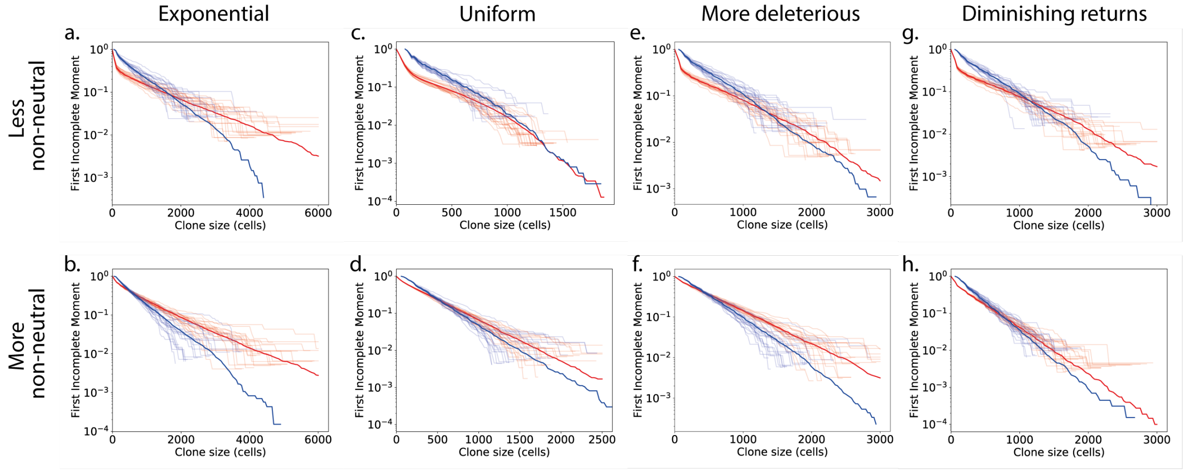

Figure 2.4 LFIM for alternate distributions of fitness effects. For all cases, the LFIMs using full clone sizes are shown in red and the LFIMs using clone sizes inferred from simulated biopsies and sequencing are shown in blue. 20 individual simulations are shown for each case and the mean of 100 simulations is shown in bold. Except where stated otherwise, the simulations in the top row are 1% non-neutral, and the simulations in the bottom row are 25% non-neutral. All simulations are run on 500 × 500 grids, have a division rate of 0.033 per week, a mutation rate of 0.015 per cell division and are run for 3000 weeks (~58 years). Unless stated otherwise, the cell fitness is the sum of the fitness of all mutations in the cell. **a,b)** The fitness of non-neutral mutations is drawn from an exponential distribution with parameter 0.1. **a)** 1% of mutations are non-neutral. **b)** 25% of mutations are non-neutral. **c,d)** The fitness of new mutations is drawn from a uniform distribution, *U*(0, 0.2). **c)** 1% of mutations are non-neutral. **d)** 25% of mutations are non-neutral. **e,f)** Addition deleterious mutations. **e)** The fitness of 1% of new mutations is drawn from *N*(mean = 0.1, std = 0.1) (mostly beneficial mutations) and the fitness of another 2% of new mutations is drawn from *N*(mean = -0.3, std = 0.1) (mostly deleterious mutations). The remaining 97% of mutations have no effect on cell fitness. **f)** The fitness of 25% of new mutations is drawn from *N*(mean = 0.1, std = 0.1) (mostly beneficial mutations) and the fitness of another 50% of new mutations is drawn from *N*(mean = -0.3, std = 0.1) (mostly deleterious mutations). The remaining 25% of mutations have no effect on cell fitness. **g,h)** Diminishing returns. The fitness of new mutations is drawn from *N*(mean=1.1, std=0.1). New cell fitness=max(old cell fitness, new mutation fitness), i.e. the fitness of a mutation will replace (rather than add to) the cell fitness if it is higher than the cell fitness, otherwise it has no effect. **g)** 1% of mutations are non-neutral. **h)** 25% of mutations are non-neutral.

The true effects of somatic mutations are likely to be far more complex than the examples I have explored here (126). However, in all cases I have simulated we see that the two key observations are not altered: Firstly, the use of biopsies reduces the curve in the LFIM and therefore reduces the ability of the LFIM to detect non-neutral competition. And secondly, a straight-line LFIM does not necessarily imply neutral competition, as shown in the highly non-neutral simulations.

2.2.5 Human oesophageal mutant clone sizes demonstrate non-neutral growth

To validate this analysis I drew on a second experiment that measured mutant clones in normal human oesophageal epithelium (10). dN/dS analysis revealed mutations in 14 of 74 genes sequenced were under significant positive selection (10). This study used an ungapped sampling strategy in which the epithelium was cut into gridded arrays of samples which were then deep DNA sequenced, allowing the areas of clones that extend over multiple samples to be determined (10) (Figure 2.3a). This key difference in design from the eyelid experiment allowed me to investigate the effect of sampling on the incomplete moment analysis of the eyelid data. I did this by comparing the gridded data with what would have happened if the oesophagus was sampled in the same manner as the eyelid skin (Figure 2.3a). The LFIMs for clone sizes estimated from both sampling approaches are shown in Figure 2.5. Taking Figure 2.5d as an example, the gapped sampling approach results in an LFIM that fits well with a straight line ($R^2 = 0.96$) and therefore appears consistent with neutral competition. However, if using the clone sizes based on a gridded approach, the LFIM deviates from the straight line ($R^2 = 0.78$), suggesting non-neutral competition may have occurred in the tissue. For each of the 9 individuals in the study, the LFIM exhibits a greater deviation from the straight line when using gridded samples than when using spaced samples.

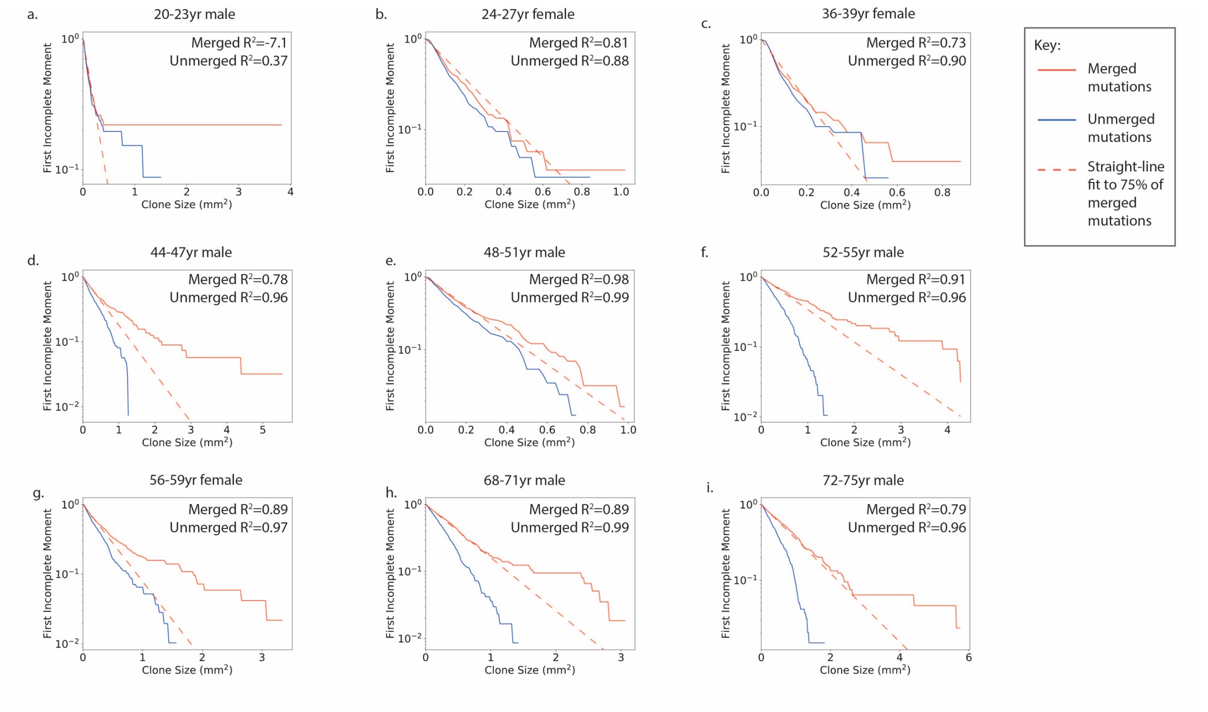

Figure 2.5 Normal human oesophagus. a-i) First incomplete moments of the human oesophagus mutation data for the nine individuals in the study (10). The clone sizes are either inferred from each 2mm^2 sample without merging (blue) or by using the gridded system to infer the size of mutations which span multiple samples using the methods of the original study (10) (red, solid). The extent of deviation from the straight line can be seen by comparing the data (solid) to the dashed red line, which shows a straight-line fit to the smallest 75% of clones in the merged case. Loss-of-heterozygosity copy number changes were frequently found to co-occur with protein-altering *NOTCH1* mutations (10) and to obtain conservative estimates of clone sizes I assumed this is the case for all protein-altering *NOTCH1* mutations. All other mutations on chromosomes 1-22 were assumed to be heterozygous. R^2 values can be negative because the line fitting is constrained to pass through the point (m, 1), where m was the smallest observed clone size. Ages given as a range for anonymisation purposes.

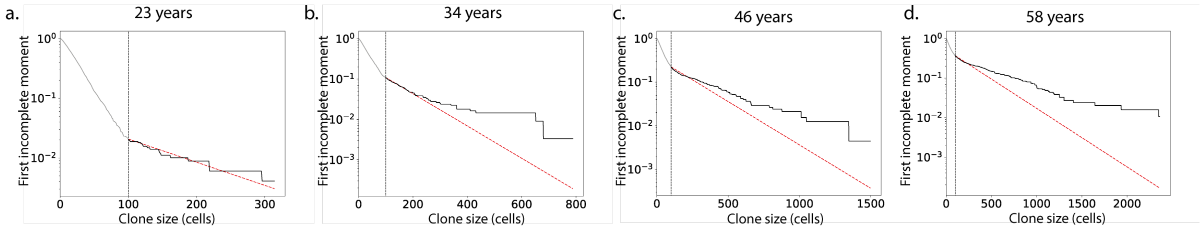

Figure 2.6 First incomplete moment over time. a-d) Curves in the LFIM may only be visible after sufficient time has passed, allowing both fast and slow growing clones to reach large enough sizes to be detectable through sequencing. Examples of the first incomplete moment for a simulation of non-neutral competition are shown for four timepoints. 1% of mutations are non-neutral with a fitness drawn from a normal distribution, *N*(mean = 0.1, std = 0.1). The vertical dashed line shows the detection limit, arbitrarily set at 100 cells. The section of the first incomplete moment that would not be visible due to the detection limit is shown in grey to the left of the line; the visible section is shown in black. The red dashed line is a straight line fit to the smallest 75% of visible mutant clones.

An intriguing pattern can be seen within the oesophageal data. With the exception of a few large clones, younger patients have less curved LFIMs, as shown by the deviation from a straight line fit to the smallest 75% of clones (Figure 2.5). Older individuals in contrast show a more distinct deviation from the line. This is consistent with simulations of non-neutral competition (Figure 2.6). At early timepoints, only the faster-growing non-neutral clones in the tail of the distribution are large enough to be observed in the simulated DNA sequencing (Figure 2.6a), leading to a straight-line LFIM. A curve is observed at later timepoints once the slower-growing clones reach a size large enough to be detected (Figure 2.6b,c,d).

2.2.6 Clone size as a marker of competitive selection of mutations

I have shown that clone size and LFIM alone cannot reliably classify clonal dynamics as neutral. This is due to a mixture of experimental limitations on the maximum and minimum sizes of clones and the fundamental effects of competition for space. In addition, where a curved LFIM is found, the position of the curve cannot simply discriminate the neutral and non-neutral mutant clones, although a trend of increasing proportions of non-neutral mutations at larger clone sizes is observed in both simulations (Figure 2.7a,b) and *in-vivo* experiments (10). Neutral mutations hitchhiking on non-neutral clones may grow to large sizes, meaning that analysis restricted to synonymous mutations and mutations in non-expressed genes (Chapter 2 methods), which are not identified as under selection by dN/dS analysis (10), may not reflect purely neutral dynamics (Figure 2.7c). This raises a question of how to meaningfully interpret clone sizes observed in a tissue. This is an important question as there remain a small number of metrics for assessing the neutrality of mutations. While the results here have demonstrated the risks of one specific interpretation of the LFIM, they also highlight the dangers of relying on a single measure of neutrality, especially if the underpinning mathematical assumptions are under-explored.

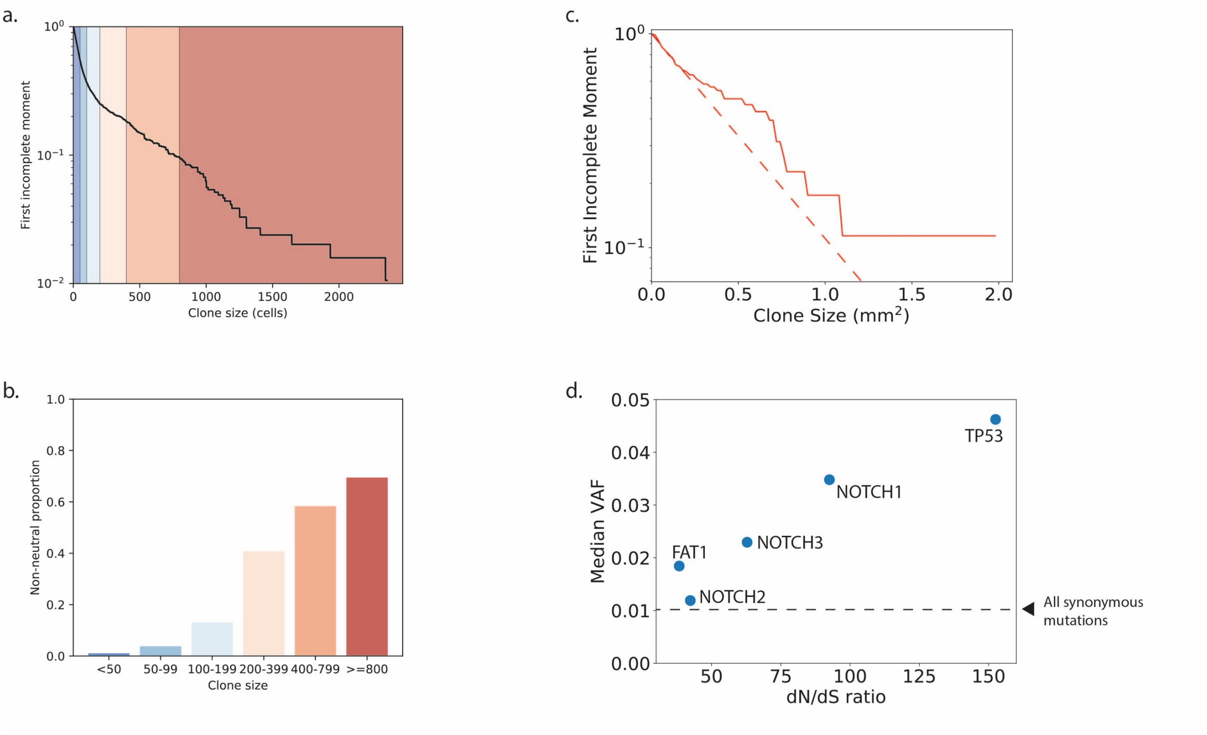

Figure 2.7 Clone size and selection. a-b) Proportion of non-neutral clones in different size ranges. **a)** The first incomplete moment of the clone size distribution from a simulation with 1% non-neutral mutations. Coloured regions correspond to ranges of clone sizes described in b. **b)** Proportion of non-neutral clones in each clone size interval. Colours correspond to the regions shaded in a. **c)** First incomplete moments of the human oesophagus mutation data for one individual, aged 72-75 (10), including only synonymous mutations and mutations in genes that are non-expressed (Chapter 2 methods). The synonymous mutation T125T in *TP53* was excluded as it has been found to affect splicing (62). Clone sizes which extend across multiple samples are merged using the methods of the original study (10). All mutations on chromosomes 1-22 were assumed to be heterozygous. The extent of deviation from the straight line can be seen by comparing the data (solid) to the dashed red line, which shows a straight-line fit to the smallest 75% of clones. **d)** Median variant allele fraction (VAF) for nonsense mutations in the five most significantly selected genes from the dN/dS analysis of the human oesophagus mutation data (10), plotted against the dN/dS ratio for nonsense mutations. Combined results for all individuals in the study. The dashed line shows the median VAF of all synonymous mutations. Many of these synonymous mutations are likely to be passengers on non-neutral clonal expansions, and therefore the line does not represent the median VAF of mutations that have grown solely under neutral drift. One-sided Mann–Whitney tests show that, aside from *NOTCH2* (p = 0.06), nonsense mutant clones in the genes shown are significantly larger than synonymous mutant clones (p-values < 0.0001).

In the specific case of the eyelid data, the original conclusion of non-neutral competition was supported by dN/dS analysis. While this is a widely used tool, this type of analysis is sensitive to the mutation model used for the neutral hypothesis (62), and detection of positive selection may be unreliable for some types of mutations in some genes. For example, almost all protein-truncating mutations inactivate the protein in which they occur. By contrast, missense mutations in some locations may reduce protein function, while in others they may generate a constitutively active mutant (127). In aggregate these effects may result in a dN/dS ratio close to 1.

Given the limitations of individual methods to assess neutrality, I speculated that combining discrete approaches may be more informative. To explore this I directly compared observed clone sizes and the associated dN/dS ratios of mutations in specific genes. I selected nonsense mutations from a panel of five mutated genes that were identified as under the strongest positive selection in normal oesophagus and which have well-characterised roles in cancer (*TP53, NOTCH1, NOTCH2, NOTCH3,* and *FAT1*) (Figure 2.7d). For 4 genes there is the expected relationship between clone size and selection; that is, mutations in genes under greater selection pressure grow into larger clones. However, *NOTCH2* clones are under selection according to dN/dS criteria but have a similar size to synonymous clones.

There are multiple possible explanations for this unexpected result for *NOTCH2.* The dN/dS ratio indicates mutations that promote clonal expansion to a sufficient size to be detected. However, the impact of a mutation on clonal behaviour may alter over time. This may occur if an initial expansion of a mutant clone increases the local cell density. If mutant cell proliferation is sensitive to this change of environment, the rate of clonal expansion may slow. Another potential mechanism is that the mutant clones grow initially due to an advantage over wild type cells, but are later constrained by the growth of neighbouring clones as the tissue is mutated over time. Both of these behaviours could lead to a high dN/dS ratio with only a modest increase in clone size, and are similar to observations of a *Trp53* missense mutation in mouse epidermis, where mutant clones have a strong competitive advantage over wild type cells but their expansion is constrained (19).

The reverse observation, large clone sizes accompanied by only a modest dN/dS ratio, may indicate that mutations in a small region or hotspot in the gene can lead to extensive clonal expansion but mutations in the rest of the gene are under weaker selective pressure. An example from the oesophagus data is *PIK3CA*, which has the largest median clone size of the 14 genes found to be under positive selection in human oesophagus, largely due to the multiple large clones of the hotspot mutation H1047R. This highlights the importance of not just considering the gene in which a mutation occurs, but also the location of the mutation in the structure of the protein. Where large-scale experiments have been performed, more than half of all non-synonymous point mutations in T4 lysozyme fail to substantially effect protein function (128), and mutations in the *TP53* DNA-binding domain were found to have a broad range of phenotypes (129). It follows that using the structure–function relationship to interpret and confirm the mechanism of frequently observed point mutations would support analysis and understanding of such datasets – this idea is explored further in chapters 3 and 4. Other factors such as epistatic interactions with other mutations (130) or age-related changes to the tissue microenvironment (131) could also lead to plastic and context-dependent mutant clone behaviour and a complex relationship between dN/dS ratios and clone sizes.

2.3 Discussion

I have presented two complementary explanations to resolve the apparent paradox regarding the dynamics of mutant clones in normal human eyelid skin. Both show how non-neutral competition can be consistent with a straight-line logarithm of the first incomplete moment of the inferred mutant clone size distribution – previously claimed to be an indication of neutral competition. Therefore, the mutant clone sizes observed in the normal human eyelid no longer appear to contradict the range of studies that suggest a number of mutations can drive non-neutral expansion of mutant clones in epithelia. I have also shown the benefits of using multiple orthogonal approaches to infer clone behaviour. Finding a consensus can provide a high degree of confidence in the analytical conclusions, and inconsistencies may reveal an issue with one of the methods or help to identify interesting outliers in the data.

The simulations used in this chapter only model a single basal layer of cells. This model setup clearly differs from the 3D nature of the real tissues, especially the human oesophageal epithelium in which there are multiple layers of proliferating cells (22). However, in order to spread in the normal tissue without altering the tissue structure, a mutant clone will still need to expand laterally within the proliferating cell layers and must compete for space with other clones. The 2D simulations will therefore still capture the general effects of spatial competition between clones in the 3D tissue, and the impact this has on the clone size distribution. Furthermore, the argument made in this chapter does not rely on the simulations accurately recreating the precise details of clonal competition in the real tissues. The simulations are only required to illustrate the overall impacts of spatial competition and experimental tissue sampling, and to demonstrate that the conclusions based on the previous non-spatial non-competitive cell dynamics model are not robust.

I have found that by considering spatial constraints of the tissue, non-neutral simulations can produce a straight-line LFIM, providing a counter example to the proposition that a straight-line LFIM implies neutral competition. I have also shown how the experimental method used to measure the sizes of mutant clones in the eyelid could hide signs of non-neutrality in the clone size distribution. Using isolated single samples which are too small in relation to the clone sizes will lead to underestimation of the size of a significant proportion of clones, as occurred in the eyelid experiment, and could lead to an apparently neutral clone size distribution. However, using over-large samples will reduce the ability to detect smaller clones, which can also lead to a straight-line LFIM because only the largest mutant clones are observed (98) (Figure 2.6a). By using a grid of adjacent samples, the larger clones can be more accurately measured without compromising the detection of smaller clones, and can reveal the signs of non-neutrality that would otherwise have been hidden.

I have discussed the effects of sampling in detail. However, there are other potential confounding factors that could appear during DNA sequencing experiments. For example, comparisons of clone sizes between genes may be confounded by variations in read coverage and the frequency of sequencing errors across genes, which could lead to different detection limits for small clones in different genes, and therefore different average clone sizes. Furthermore, multiple independent but identical mutations within the same sample would be observed as a single clone (although this is likely to be rare (13)) and some large clones may be caused by somatic mosaicism rather than positive selection (12).

At the same time as I was working on this project, other researchers were having a similar debate about selection of mutant subclones in tumours. It was proposed that neutral clonal competition in an exponentially expanding tumour would lead to a cumulative distribution of subclone sizes that follows a power law of the inverse of the allelic frequency (*1/f*) (132). Therefore plotting the cumulative number of mutations against the inverse of the allelic frequency would produce a straight line if the process was entirely neutral. The study then interpreted a straight line on this *1/f* plot as an indication of neutral growth (132). Subsequent work (including by the authors of the original study) highlighted various flaws in this argument. For example, it was argued that the test could only reject neutrality under certain conditions on the timing of the emergence of a positively selected subclone and the strength of the selection advantage (133, 134). It was also found that the shape of the *1/f* plot depended on the structure of the tumour (135, 136), including some circumstances in which non-neutral competition lead to a straighter line than neutral competition (136). In addition, dN/dS analysis detected significant positive selection in tumours that were classified as neutral using the 1/f test (133). More recently the authors of the original paper have used a computational model that includes the simulation of tumour structure and sampling methods to infer selection and growth parameters, while acknowledging the challenges that remain when trying to characterise a complex process using data that may be noisy and biased by the collection methods (135).

2.4 Chapter 2 methods

2.4.1 Moran simulations

The simulations were carried out on a 500 × 500 hexagonal lattice whose edges were wrapped to form a torus. Each cell was assigned a fitness value of 1 at the start of a simulation. Similar to a Moran process (90), one cell was randomly selected at each simulation step to differentiate (was removed from the simulation) and a neighbouring cell was selected to divide to fill the space, with fitter cells having a higher chance of dividing (Figure 2.1e). During each division, there was a chance that a mutation would occur in the new cell. If the cell did not mutate during division or if the mutation was neutral, the new cell produced would inherit the fitness of its parent cell. For the simulations in Figures 2.2,2.3,2.6 and 2.7, if the mutation was non-neutral, a random value drawn from a normal distribution, N(mean = 0.1, std = 0.1), was added to the fitness of a cell. In Figure 2.4, I show that the particular choice of distribution does not affect the conclusions of the analysis.

Estimates of cell cycle time in human tissues are hard to verify. However, I did not fit simulations to data, only demonstrated general properties of the models, hence the exact division rates used do not affect the conclusions of the analysis. For the neutral simulations, I used a division rate of 0.5 per week as estimated from the LFIM of clones in the human eyelid under the assumption of neutral competition (13). In the non-neutral simulations, the fittest clones can expand much faster than neutral, and therefore I reduced the division rate to 0.033 per week so that maximum mutant clone sizes were similar to the neutral simulations. The simulations ran for 3000 weeks (~58 years).

The somatic mutation rate for human tissue has been estimated at approximately 10^{-9} mutations per base pair per cell division (137) although it should be noted that exposure to UV light or mutagenic agents (such as stomach acid and alcohol) may substantially alter the mutation rate. With roughly 10^6 base pairs included in the targeted sequencing experiment (9), this leads to a mutation rate of 10^{-3} mutations per cell division which I use for the neutral simulations. I use a higher mutation rate of 1.5×10^{-2} mutations per cell division for the non-neutral simulations so that the total clone numbers were similar in the neutral and non-neutral simulations.

Clone sizes were defined by the number of cells containing each mutation at the end of the simulation.

2.4.2 Biopsy and sequencing simulations

Biopsies were simulated by taking 25 non-overlapping 70 × 70 cell squares from each grid. Assuming a density of basal cells of 10,000 per mm^2 (138) and that half of basal cells are progenitors, this would make the biopsies approximately 1 mm^2, similar in size to those used in the human eyelid (9).

Small clones may only appear in a very small proportion of sequenced DNA reads (if any) and are therefore hard to distinguish from sequencing errors (59), meaning they are not successfully detected as somatic mutations. To replicate this, I assumed a constant 1000x read depth and a requirement of 10 reads as a minimum to observe the mutant. Each mutant had a true frequency f, the proportion of cells which contained the mutant. I assumed all mutations were heterozygous, so the true VAF was given by $0.5f$. Each read then had a $0.5f$ chance of containing the mutation, so the total number of mutant reads observed, $reads_{obs}$, was given by a draw from a binomial distribution with $n = 1000$, $p = 0.5f$. If $reads_{obs}$ was greater than 10, I recorded the mutant as having a VAF of $reads_{obs}/1000$, otherwise the mutant was unobserved and not included in the results.

2.4.3 First incomplete moment test

I used the first incomplete moment as defined in (13).

$$\mu_1(n,t) = \frac{1}{\sum_{m=1}^{\infty} mP_m(t)} \sum_{m=n}^{\infty} mP_m(t)$$

where $P_m(t)$ is the proportion of surviving clones that have m cells at time t. The normalisation using the average clone size, $\sum_{m=1}^{\infty} mP_m(t)$, means that $\mu_1(n,t) = 1$ for all values of n smaller or equal to the smallest observed clone size. As in previous studies (13, 98), I used R^2, the coefficient of determination, to assess whether the log of the first incomplete moment was a straight line. The line fitting was constrained to pass through the point (m, 1), where m was the smallest observed clone size.

2.4.4 dN/dS

Neutral and non-neutral mutations were introduced into the simulations with a known ratio, a. dN/dS was calculated as follows:

$$dN/dS\ = \frac{N}{aS}$$

where *N* was the number of observed non-neutral clones and *S* was the number of observed neutral clones.

2.4.5 RNA expression

RNA levels for genes in the human oesophagus were obtained from RNA-seq data from the Human Protein Atlas (139) (available from www.proteinatlas.org, accessed 01/10/2018). I used this data to select a set of genes which are not expressed and are therefore highly unlikely to be selected for. It is not clear at which transcripts per million (TPM) value a gene would have sufficient expression to make selection possible. I therefore used a conservative threshold of 0.0 TPM. The genes with 0.0 TPM in the gene panel sequenced in the oesophagus (10) were *ADAM29, GRM3, KCNH5, MUC17, PTPRT, SCN11A, SCN1A* and *SPHKAP*. All genes under positive selection in the human oesophagus mutation data (10) have a non-zero TPM (Table 2.1).

Table 2.1 RNA levels for genes under positive selection in oesophagus. RNA-seq data from the Human Protein Atlas (139) for the genes under positive selection in human oesophagus (10).

Gene	**TPM**
NOTCH1	4.1
NOTCH2	22
NOTCH3	45.3
TP53	32.4
CUL3	60.1
FAT1	14.9
ARID1A	19.7
KMT2D	12.6
AJUBA	12
PIK3CA	7.3
ARID2	7
NFE2L2	267.1
TP63	60.9
CCND1	78.9

Chapter 3

Investigating protein structure–function relationships using somatic mutations: method and validation

3.1 Introduction

Mutational scanning experiments exhaustively mutate a protein or protein domain and measure the phenotypic consequences of each mutation (140). This approach can provide information both about the effects of individual mutations and about protein function and structure (140). However, these experiments are carried out *in vitro* (140), an environment which can substantially alter cell phenotype (141, 142). *In-silico* mutational scanning has become increasingly common as computational power increases and *in-silico* methods improve (143-146). This technique predicts the effects of mutations on protein function, but may not directly relate those results to effects on cell phenotype. Over the last decade, DNA sequencing has enabled the detection of vast numbers of mutations in normal tissue samples (9-12, 41, 43). These DNA-sequencing studies may be thought of as large-scale *in-vivo* mutational phenotype assays, examining the clonal growth advantage conveyed by mutations. By combining this sequencing information with *in-silico* mutational effect predictions, we can attempt to infer the changes to protein function that have *in-vivo* phenotypic consequences.

Mutations are acquired by cells as a result of exposure to mutagens (e.g. tobacco, alcohol or ultraviolet light) or cell-intrinsic processes (147). The pattern of mutations acquired, known as the mutational spectrum, will depend on the mutational processes involved (147). The chance of acquiring a particular single nucleotide substitution appears to depend (at least partially) on the adjacent nucleotides (147). Many of the mutations do not alter cell behaviour and are lost through neutral drift, the process of random cell births and loss in which, by chance, some cell lineages expand and others are lost (81). Positively selected (driver) mutations confer a proliferative advantage to the mutant cell, thus increasing the number of descendant cells that inherit the mutation (17). These driver mutations are therefore more likely to generate a clone large enough to be detected by DNA sequencing (62).

Methods to detect genes under selection make the assumption that mutations are generated according to the mutational spectrum (Figure 3.1). In non-selected (neutral) genes, where no mutations of any kind convey a growth advantage/disadvantage, the mutations detected by DNA sequencing will simply be an unbiased sample of the mutations produced by the spectrum. In genes under selection, it is assumed that certain types of mutations (missense/nonsense mutations (62), or mutations estimated by machine-learning approaches to have high functional impact (63, 71)) are more likely to alter cell phenotype than others (silent mutations, mutations with low estimated functional impact). Therefore, more high-impact mutations will be detected in a positively selected gene than would be expected under the neutral null hypothesis (62, 63, 71) (Figure 3.1). Conversely, negative selection in a gene will lead to the detection of fewer high-impact mutations than would be expected under the null hypothesis.

Based on this principle, in this chapter I develop a statistical methodology to look in detail at patterns of mutant selection within individual genes (Figure 3.1). By identifying common features of selected mutations we can gain information about how the wild type and mutant proteins are functioning. I demonstrate the method and explore its benefits and limitations using datasets of mutant clones found through DNA sequencing of normal human oesophageal epithelium (10) and skin (11). *NOTCH1* is one of the most frequently mutated genes in these studies and is a strong driver of clonal expansion in these tissues (9-12) (Chapter 5). Due to the large available sample size from these studies and the known functional impact of its frequently occurring mutations, *NOTCH1* is used in this chapter to introduce and validate the method. In Chapter 4, I apply the method to several other genes sequenced in the normal oesophagus and skin.

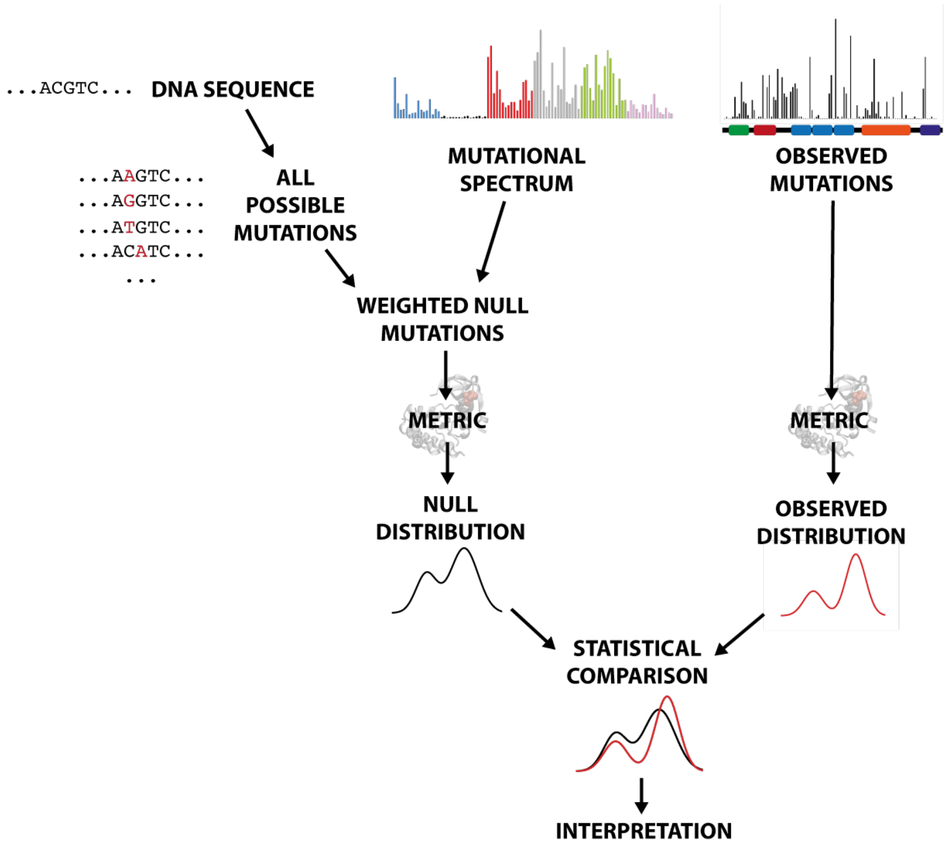

Figure 3.1 Schematic of the statistical test for a selected feature. The null hypothesis assumes that the mutations appear based on the mutational spectrum, and no selection or other bias is occurring. To generate the metric distribution under the null hypothesis, all possible single nucleotide mutations are generated for the gene or region to be tested, weighted by the mutational spectrum and scored using the metric. The alternative hypothesis is that some selection occurs and that it correlates with the tested metric, leading to a shift in the distribution of metric scores. The null and alternative distributions are compared using a statistical test (Chapter 3 Methods). The comparison determines if the mutations with a high/low metric score are more strongly selected than the rest of the mutations in the tested region, and therefore the results of the test must be interpreted while considering other causes of selection in the region.

NOTCH1 is a strong driver of clonal expansion in healthy skin and oesophageal epithelium (10, 11). NOTCH proteins are membrane-bound cell surface receptors (Figure 3.2) in a pathway that regulates cell fate (148). These genes are critical regulators of normal cellular function in development and adult tissues and are mutated to activate or block function in cancers (149, 150). The extracellular domains of NOTCH proteins contain up to 36 epidermal growth factor (EGF) repeats (Figure 3.2). Many of the EGF repeats bind to calcium ions, which add rigidity to the structure, help fix the relative orientation of adjacent EGF repeats (151), and are required for ligand binding (152). NOTCH ligands, from the Delta-like and Jagged families (153), are expressed by adjacent cells and bind to a subset of the EGF repeats (148) (Figure 3.2), with EGF11 and 12 of *NOTCH1* particularly crucial for ligand binding (148, 154). This binding triggers a cascade of cleavage events, resulting in the cleavage of the NOTCH transmembrane helix and the release of the NOTCH intracellular domain (NICD), which travels to the nucleus and forms part of a transcription factor complex that increases the expression of NOTCH target genes (148) (Figure 3.2).

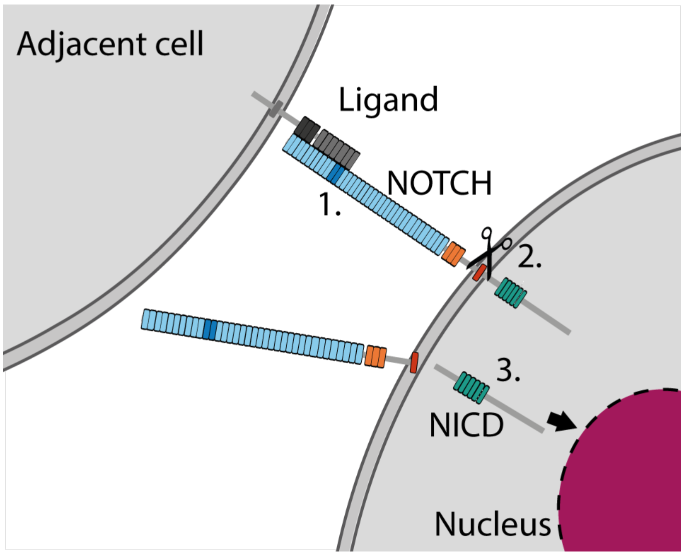

Figure 3.2 Mechanism of NOTCH activation. 1) NOTCH is activated when a ligand from an adjacent cell binds with a subset of the NOTCH EGF repeats. EGF repeats shown in blue, key ligand-binding region EGF11–12 shown in dark blue; LNR repeats, orange; transmembrane region, red; ankyrin repeats, green. 2) Ligand binding triggers a series of cleavage events, resulting in cleavage of the NOTCH transmembrane region (red) by γ-secretase. 3) This releases the intracellular domain of NOTCH (NICD) which travels to the nucleus to form part of a transcription factor.

The NICD is released by cleavage of the transmembrane helix of NOTCH at the S3 site by the multiprotein γ-secretase complex (155). The exact site of cleavage can vary, but the S3-V cleavage (between G1753 and V1754 in *NOTCH1*) produces the most stable form of NICD due to N-end rule degradation (156) and is therefore likely responsible for the majority of *NOTCH* signalling (148). Mutations in the vicinity of the S3 cleavage site may affect *NOTCH* signalling by reducing S3 cleavage efficiency (157) or by leading to the production of less stable NICD species (156).

The impact on ligand binding of recurrent mutations in *NOTCH1* EGF11–12 has previously been described (10), and there are large numbers of missense mutations in this region (Figure 3.3a, 905 and 308 missense mutations detected in normal skin and normal oesophagus respectively (10, 11)). I therefore use the ligand-binding region of *NOTCH1* to introduce the method and confirm that it can detect (and assign statistical significance to) the previously observed patterns of selection. I also demonstrate how excluding known forms of selection can help to detect weaker selection or less frequently selected features. In addition, I explore how the choice of mutational spectrum used in the null model affects the results for *NOTCH1* EGF11–12. Furthermore, I use the *NOTCH1* transmembrane region to illustrate how the combination of mutational spectrum and selection of functional impact leads to the pattern of mutations observed. Finally, I discuss a range of hypotheses that could account for the high concentration of missense mutations in the ligand-binding region of *NOTCH1*.

3.2 Results

3.2.1 *NOTCH1* EGF 11–12 missense mutations cause misfolding and disrupt ligand- and calcium-binding sites

The critical ligand-binding EGF repeats 11–12 of *NOTCH1* contain the highest concentration of missense mutations in the gene (Figure 3.3a) (10, 11). Recurrently mutated residues in this region include cysteines in disulphide bonds, buried glycines and hydrophobic packing residues (10). These would all be expected to affect the stability of the protein (158, 159) and could prevent the structure from folding into the correct shape to bind with the ligand (160, 161). I therefore started by testing whether there is selection for destabilising mutations.

The stability of a protein is determined by the protein folding free energy, ΔG, which is the difference in Gibbs free energy between the folded and unfolded form of a protein (162). A mutation may alter ΔG (this change is called ΔΔG) and therefore stabilise or destabilise the protein (162). I used FoldX (162) to calculate the ΔΔG for each possible single nucleotide missense mutation in NOTCH1 EGF11–12 (Figure 3.3b,c, Chapter 3 Methods). I constructed a null model of “neutral” selection, which assumes that the distribution of mutations in the region will depend solely on the mutational spectrum (Figure 3.1, Chapter 3 Methods). By comparing the distribution of ΔΔG values from the observed mutations with the distribution expected under the null hypothesis (Figure 3.1), I found a significant enrichment of destabilising mutations with high ΔΔG values (skin: $p<2e^{-5}$, n=905; oesophagus: $p<2e^{-5}$, n=308; two-tailed Monte Carlo test, Chapter 3 Methods, Figure 3.3d,e). However, we can see that many observed mutations, including recurrent hotspots, do not appear to be destabilising (Figure 3.3b,c). This suggests that some mutations are selected for reasons other than destabilising the protein structure.

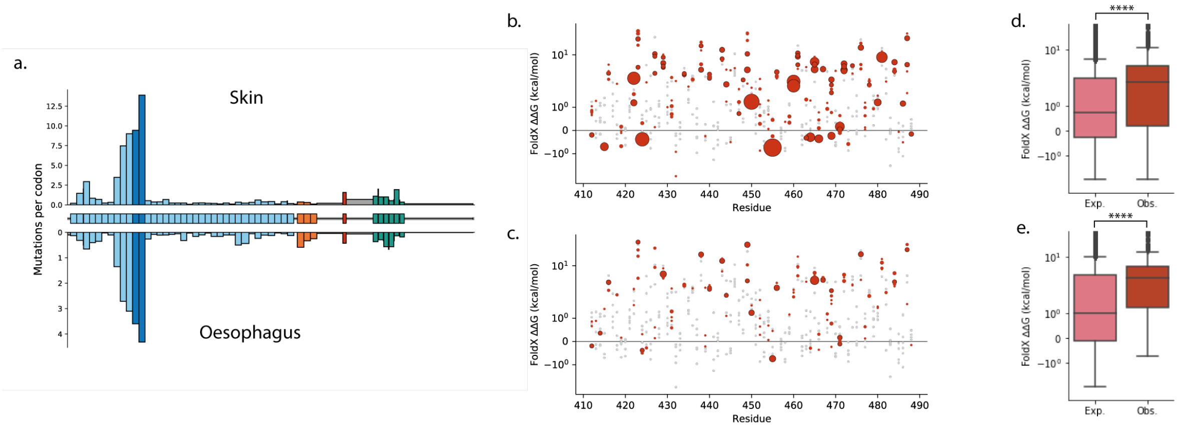

Figure 3.3 Destabilising missense mutations in NOTCH1 EGF11–12. a) Missense mutation frequency across the domains of *NOTCH1* in normal skin (top) and normal oesophagus (bottom). Domain definitions from UniProt (163). Where the gap between domains is only a single residue, mutations from this residue are included in the subsequent domain. EGF repeats, blue; EGF11–12, dark blue; LNR repeats, orange; transmembrane region, red; ankyrin repeats, green; other regions, grey. **b-c)** ΔΔG of mutations in NOTCH1 EGF11–12. Single nucleotide missense mutations that occur in the normal skin (**b**) or normal oesophagus (**c**), red, with marker size proportional to the number of times that mutation occurs. Single nucleotide missense mutations that do not occur in the dataset shown in grey. **d-e)** Distributions of calculated ΔΔG values of missense mutations. Distribution expected under the neutral null hypothesis, light red, and the distribution observed, dark red, in skin (**d**) and oesophagus (**e**). P-values calculated using a two-tailed Monte Carlo test for **d**,**e** (Chapter 3 Methods). ****P≤ 0.0001.

Another mechanism expected to inactivate NOTCH1 ligand binding is disruption of the ligand-binding interface. Using structures of rat NOTCH1 bound to the ligands JAG1 and DLL4 to define the residues on the ligand-binding interface (160, 161) (Chapter 3 Methods, Figure 3.4a, Figure 3.12, Table 3.3), I found that the observed proportion of missense mutations on the interface was similar to the null model and therefore the interface mutations were not significantly selected compared with the rest of EGF11–12 (Figure 3.4b,c; skin: expected=37%, observed=39%, p=0.12, n=905; oesophagus: expected=35%, observed=33%, p=0.72, n=308; two-tailed binomial test, Chapter 3 Methods). However, this does not indicate that mutations on the interface are under neutral selection, just that they are not under stronger selection than the bulk of missense mutations in EGF11–12. Orthogonal mechanisms may dominate the selection landscape and make it more difficult to identify enrichment of interface mutations. As I have already shown that highly destabilising mutations are positively selected, all mutations with high ΔΔG values (with ΔΔG>2kcal/mol, Chapter 3 Methods, results using a range of thresholds shown in Figure 3.6c-f) can be excluded from both the null model and the observed data and it can be tested whether, *within the non-destabilising mutations in EGF11–12*, there is an enrichment of interface mutations. Under this null model, there was a highly significant increase in the proportion of missense mutations on the ligand-binding interface

(Figure 3.4d,e, skin: expected=44%, observed=58%, $p=4e^{-9}$, n=462; oesophagus: expected=43%, observed=64%, $p=3e^{-5}$, n=107; two-tailed binomial test, Chapter 3 Methods).

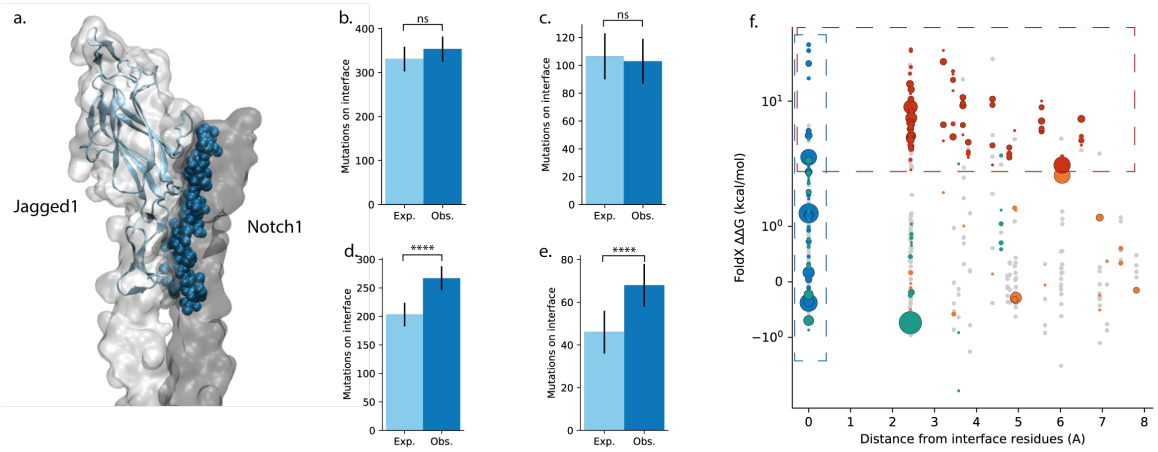

Figure 3.4 Ligand-binding interface mutations in NOTCH1 EGF11–12. **a)** Ligand interface residues (Chapter 3 Methods), blue spheres, shown on rat NOTCH1 bound to JAG1 (structure 5UK5) (161). **b-e)** Counts of NOTCH1 EGF11–12 mutations occurring on the ligand-binding interface under the neutral null hypothesis, light blue, and observed, dark blue. **b,c)** Null and observed counts including all missense mutations for skin (**b**) and oesophagus (**c**). **d,e)** Null and observed counts excluding destabilising mutations (ΔΔG>2kcal/mol) from both null model and observed data for skin (**d**) and oesophagus (**e**). **f)** Calculated ΔΔG plotted against distance from the NOTCH1 EGF11–12 ligand-binding interface residues. Single nucleotide missense mutations that occur in the skin data set with marker size proportional to the number of times that mutation occurs shown in green if the residue is calcium binding, blue if the residue is on the ligand-binding interface, red if the mutation has ΔΔG>2kcal/mol, orange otherwise. Single nucleotide missense mutations that do not occur in the skin data set shown in grey. Regions containing highly destabilising mutations (ΔΔG>2kcal/mol) and mutations on the ligand-binding interface shown with dashed red and blue boxes respectively. P-values for **b-e** calculated using a two-tailed binomial test (Chapter 3 Methods) and error bars in **b-e** show 95% confidence intervals (Chapter 3 Methods). ****P≤ 0.0001, ns P>0.05.

After excluding misfolding mutations and those on ligand-binding sites, there are still some missense mutations remaining (Figure 3.4f). Among these, the most frequent mutation in both skin and oesophagus is E455K, which forms part of a calcium-binding site that is known to be crucial for ligand binding (152, 164). Calcium binding is critical for the structural integrity of the EGF repeats (151, 152), however, FoldX predicts that many mutations affecting the calcium-binding sites would not be destabilising (Figure 3.4f). This may indicate that ΔΔG, as calculated by FoldX, is not fully capturing the disruption that mutations on calcium-binding residues are having on the protein structure. I therefore used MetalPDB (164) to define the calcium-binding residues in the structure of NOTCH1 EGF11–12 (Figure 3.5a, Chapter 3 Methods) and tested for enrichment of mutations on these sites. Similar to the results for the interface mutations, testing with all mutations in EGF11–12 did not detect a significant enrichment of mutations on the calcium-binding residues (skin: expected=19%, observed=18%, p=0.37, n=905; oesophagus: expected=16%, observed=14%, p=0.48, n=308; two-tailed binomial test, Chapter 3 Methods, Figure 3.5b,c). However, by excluding the FoldX-destabilising (ΔΔG>2kcal/mol) and ligand-binding interface mutations, I found that the calcium-binding mutations are highly selected compared to the remaining mutations in the region (skin: expected=18%, observed=49%, p=2e^{-22}, n=195; oesophagus: expected=20%, observed=59%, p=8e^{-8}, n=39; two-tailed binomial test, Chapter 3 Methods, Figure 3.5d,e).

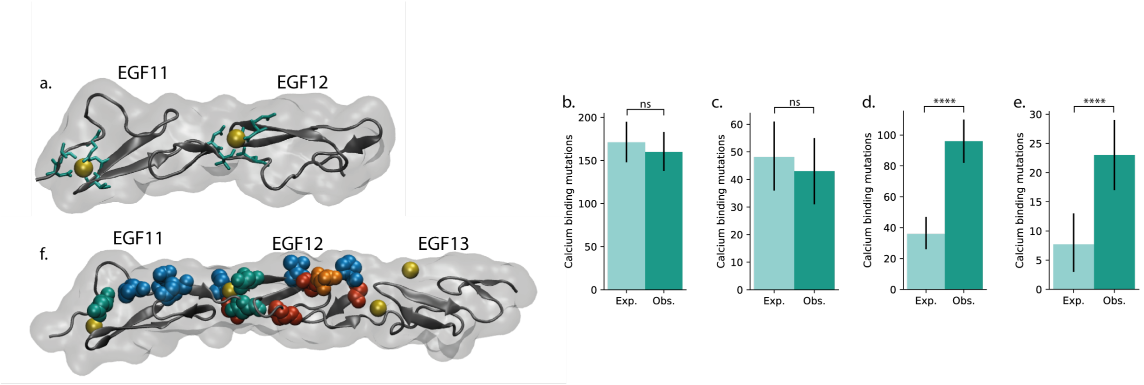

Figure 3.5 Calcium-binding residue mutations in NOTCH1 EGF11–12. **a)** Calcium-binding residues in NOTCH1 EGF11–12 (Chapter 3 Methods) shown on structure 2VJ3 (165). Calcium-binding residues shown in green, Ca^{2+} ions shown in yellow. **b-e)** Counts of NOTCH1 EGF11–12 mutations that are on calcium-binding residues under the neutral null hypothesis, light green, and observed, dark green. **b,c)** Null and observed counts including all missense mutations for skin (**b**) and oesophagus (**c**). **d,e)** Null and observed counts excluding destabilising mutations (ΔΔG>2kcal/mol) and ligand-binding interface mutations from both null model and observed data for skin (**d**) and oesophagus (**e**). **f)** Residues containing more than 10 missense mutations in skin shown on structure 2VJ3 (165). Note the none of the recurrent mutations occur on EGF13 at the right-hand end of the structure. Residues coloured based on the category of the mutations on that residue: ligand-binding interface residues, blue; calcium-binding residues, green; destabilizing residues (mutations with ΔΔG>2kcal/mol), red; D464N, orange, does not fit into the previous categories. Calcium ions are shown in yellow. P-values for **b-e** calculated using a two-tailed Monte Carlo test (Chapter 3 Methods) and error bars show 95% confidence intervals (Chapter 3 Methods). ****P≤ 0.0001, ns P>0.05.

This example of missense mutations in *NOTCH1* EGF11–12 has confirmed that the statistical method can detect selection of known functional consequences of mutations. Approximately 70% of potential single nucleotide mutations in EGF11–12 belong to the three categories of mutational impact examined here, and are therefore likely to inactivate *NOTCH1*. 89% and 95% of observed missense mutations are within the three categories in the skin (Figure 3.4f, **Figure 3.5**f) and oesophagus data respectively. The small proportion of mutations remaining may be weakly selected or neutral passenger mutations, may be marginally outside of the category definitions selected here, or may be selected due to a functional impact not tested here. For example, the most frequent mutation in *NOTCH1* EGF11–12 in skin that is not in the above categories is P460L (34 mutant clones), which has a ΔΔG value marginally below the chosen threshold of 2 kcal/mol (Figure 3.4f). The next most frequent uncategorised mutation, D464N (17 mutant clones) does not clearly belong to any of the three categories (Figure 3.4f,

Figure 3.5f). The repeating method of testing for selection, excluding selected categories and testing again for new selected features may be useful for quickly classifying the majority of mutations and identifying outliers for further investigation.

Testing for selection of each mutation category while excluding the other categories leads to more significant results in the majority of cases, even though the exclusion can substantially reduce the sample size (Figure 3.6, Table 3.1). This demonstrates the importance of considering the exact set of mutations included when using a test which detects *relative* selection.

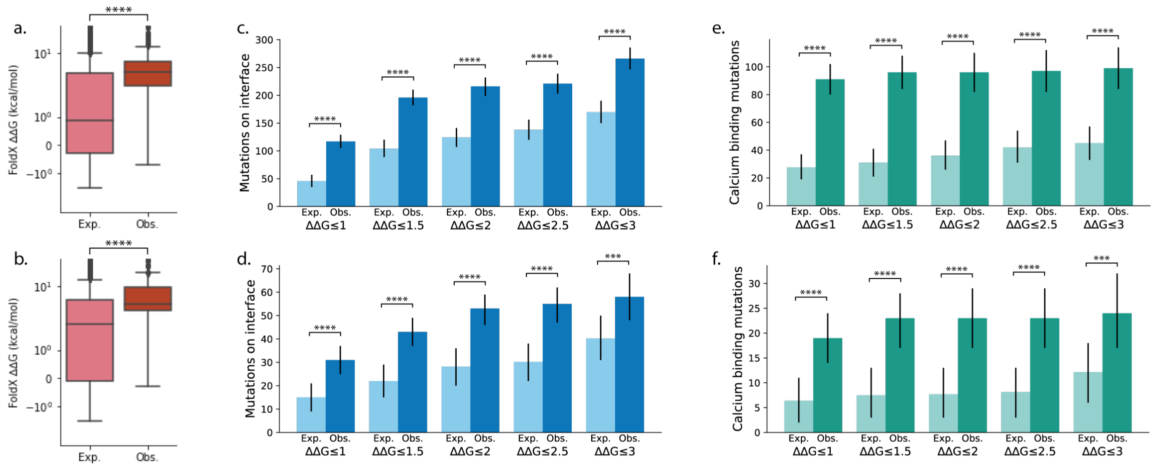

Figure 3.6 Selection of mutation categories in NOTCH1 EGF11–12 while excluding mutations in the other categories. a-b) Expected and observed distributions of ΔΔG in skin (**a**) and oesophagus (b) excluding all mutations on the ligand-binding interface and calcium-binding residues. Distribution expected under the neutral null hypothesis, light red, and the distribution observed, dark red. c-**d)** Expected (light blue) and observed (dark blue) number of mutations on the ligand-binding interface excluding calcium-binding residues and excluding mutations with ΔΔG above the thresholds shown. e-**f)** Expected (light green) and observed (dark green) number of mutations on calcium-binding residues excluding the ligand-binding interface and excluding mutations with ΔΔG above the thresholds shown. P-values calculated using a two-tailed Monte Carlo test for **a,b** and using a two-tailed binomial test for **c-f** (Chapter 3 Methods). Error bars in **c-f** show 95% confidence intervals (Chapter 3 Methods). ****P≤ 0.0001, ***P≤0.001.

Table 3.1 Statistical tests of selection in *NOTCH1* EGF11–12. Tests either include all missense mutations in the region or exclude mutations in other selected categories from both the null model and observed data. FoldX ΔΔG p-values are calculated using a two-tailed Monte Carlo test; ligand-binding and calcium-binding p-values are calculated using a two-tailed binomial test (Chapter 3 Methods).

Tested feature	Tissue	Excluded mutations	n	Expected proportion	Observed proportion	P-value
FoldX ΔΔG	Skin	None	905	-	-	$2e^{-5}$
		Ligand-binding and calcium-binding	452	-	-	$2e^{-5}$
	Oesophagus	None	308	-	-	$2e^{-5}$
		Ligand-binding and calcium-binding	181	-	-	$2e^{-5}$
Ligand-binding interface	Skin	None	905	37%	39%	0.12
		ΔΔG>2kcal/mol and calcium-binding	315	39%	69%	$2.4e^{-25}$
	Oesophagus	None	308	35%	33%	0.72
		ΔΔG>2kcal/mol and calcium-binding	69	41%	77%	$1.7e^{-9}$
Calcium-binding	Skin	None	905	19%	18%	0.37
		ΔΔG>2kcal/mol and ligand-binding	195	18%	49%	$2.2e^{-22}$
	Oesophagus	None	308	16%	14%	0.48
		ΔΔG>2kcal/mol and ligand-binding	39	20%	59%	$8.4e^{-8}$

3.2.2 Choice of mutational spectrum assumptions

The analysis above has demonstrated how the choice of mutations included in the test can dramatically alter the results. Another choice the researcher needs to make is how to define the mutational spectrum used to construct the null model (Figure 3.1). In this section I explore a range of spectrum assumptions and their impact on the statistical tests of selection in *NOTCH1* EGF11–12. I then use the transmembrane region of *NOTCH1* as an illustration of how the combination of the mutational spectrum and selection leads to the pattern of mutant clones observed.

Cells acquire somatic mutations through exposure to mutagens or during (sometimes defective) cellular processes (147). Different mutational processes have their own characteristic 'signatures' of nucleotide substitution frequencies that depend on the nucleotide change itself and the surrounding nucleotide context (147). The pattern of mutations observed in a sample, known as the mutational spectrum, is a mix of the signatures of the contributing mutational processes (147).

The probability of a particular somatic mutation appearing in a sequenced sample depends on both the rate at which the mutation occurs in cells and the strength of selection on the mutation once it has occurred. The method in this chapter attempts to separate these two factors, and therefore requires a model of how often each mutation would appear in the absence of selection (the null hypothesis model). A common method to do this is to assume that, within a gene, each mutation in the same spectrum category has the same mutation rate (62, 63, 71). For example, using a trinucleotide spectrum, two AAA trinucleotide sequences at different locations within the same gene are assumed to have the same probability of mutating into ACA.

A variety of spectrum assumptions have been used in driver detection models. Using a trinucleotide mutational spectrum to generate a null model has been found to improve driver detection compared to more simplistic substitution models (62). However, it has also been suggested that a pentanucleotide context may be more appropriate for the signature of ultraviolet light (62). Transcription coupled repair may produce a strand bias in the mutational spectrum (62) which can be taken into account by separating the transcribed and non-transcribed strand of a gene. Previous studies have assumed either that all coding regions in the genome share the same mutational spectrum (62, 63), or that each gene has its own spectrum (71). Null models may improve in the future as work continues to determine the factors that influence mutation rate (61).

Each of these assumptions alters the number of mutation rate parameters that need to be estimated. The more mutation rate parameters there are, the more thinly spread the observed mutations, and the noisier the estimation of each rate will become. The choice of spectrum assumptions is therefore a compromise between a potentially more accurate representation of the context that determines mutation rate (larger spectrum) and more stable estimates of each rate parameter (smaller spectrum).

Selected mutations within the data may distort the mutational spectrum. This could be a larger problem when using gene specific spectra or for datasets containing only a small number of genes because the selected mutations may make up a larger proportion of the total data. To some extent this selection bias can be reduced by removing duplicate mutations before calculating the spectrum (166). For high mutation loads, where duplicate mutations are likely even under neutral selection, this deduplication may have its own distorting effect.

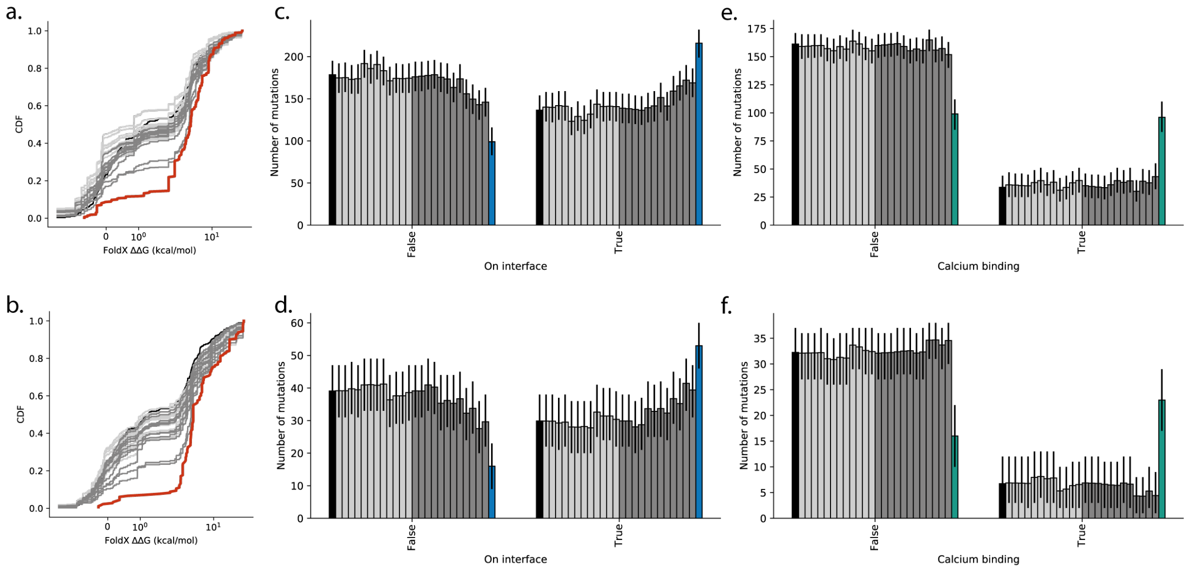

Figure 3.7 Effect of mutational spectrum assumptions on statistical tests of selection in NOTCH1 EGF11–12. a,b) Cumulative distributions of the observed ΔΔG values, red, and the null distributions, grey and black, for missense mutations in NOTCH1 EGF11–12. Mutations on the ligand-binding interface and calcium-binding residues excluded from both the null and observed distributions. The observed distributions are significantly different from all shown null distributions: **a)** skin, p<0.0001, n=452; **b)** oesophagus, p<0.004, n=181; two-tailed Monte Carlo test, Chapter 3 Methods. Null distribution using a spectrum with all mutations with equal probability shown in black, null distributions with spectra calculated from mutations in all genes in the dataset shown in light grey, and null distributions with spectra calculated only from mutations in NOTCH1 shown in dark grey. **c,d)** Counts of NOTCH1 EGF11–12 missense mutations on the ligand-binding interface under the null hypotheses, grey or black, and observed, blue. Mutations with ΔΔG>2kcal/mol and mutations on calcium-binding residues are excluded. Colours of null hypotheses as in **a**. Error bars show 95% confidence intervals (Chapter 3 Methods). The observed counts are significantly different from all shown null models: **c)** skin, $p<5e^{-7}$, n=315; **d)** oesophagus, p<0.005, n=69; two-tailed binomial test, Chapter 3 Methods. **e,f)**. Counts of NOTCH1 EGF11–12 missense mutations on calcium-binding residues under the null hypotheses, grey or black, and observed, green. Mutations with ΔΔG>2kcal/mol and mutations on the ligand-binding interface are excluded. Colours of null hypotheses as in **a**. The observed counts are significantly different from all shown null models: **e)** skin, $p<2e^{-16}$, n=195; **f)** oesophagus, $p<3e^{-7}$, n=39; two-tailed binomial test, Chapter 3 Methods. Error bars show 95% confidence intervals (Chapter 3 Methods). Full list of p-values shown in Table 3.2.

I ran the tests for selection of destabilising, ligand interface, and calcium-binding mutations in *NOTCH1* EGF11–12 (section 3.2.1 above) using a range of different spectrum assumptions to see how robust the signs of selection are to a variety of null hypotheses. In all cases, the results remained significant (Figure 3.7, Table 3.2). The least significant results tended to be from spectra with large numbers of mutational categories calculated from only the mutations in

NOTCH1 (Figure 3.7, Table 3.2). These are likely to be examples of overly detailed spectra for the data (71), with up to 3072 separate rate parameters, calculated in the case of the oesophagus from less than 1300 single nucleotide substitutions in *NOTCH1*. As the number of mutations per rate parameter reduces, recurrent hotspots will have a larger influence over their own expected mutation rate in the null model. In the extreme case, where each possible mutation has a unique rate parameter, the expected frequencies in the null model would simply match the observed data.

Table 3.2 P-values of statistical tests of selection in NOTCH1 EGF11–12 under null models that use different assumptions for the mutational spectrum. For each test, mutations in the other two categories (ΔΔG>2kcal/mol, ligand-binding, or calcium-binding) were excluded from both the null model and the observed data. The "Even" spectrum assigns an equal probability to every mutation. The "global" spectra calculate the mutation rates using all exonic single nucleotide mutations in the dataset. The "transcript" spectra calculate the mutation rates using only the exonic single nucleotide mutations in *NOTCH1*. The number in the spectrum name is the number of mutation rates in the spectrum. "6" and "12" do not use a wider nucleotide context, "96" and "192" use a trinucleotide context, and "1536" and "3072" use pentanucleotide contexts. "12", "192" and "3072" distinguish between the transcribed and non-transcribed strands, while "6", "96" and "1536" do not. "dedup" indicates that duplicate mutations (defined as having the same chromosomal position, reference base and mutant base) are removed before calculating the spectrum. The two-tailed Monte Carlo test was used for the ΔΔG tests, the two-tailed binomial test was used for the ligand-binding and calcium-binding tests (Chapter 3 Methods). The lowest possible p-value from the Monte Carlo tests used was $2e^{-5}$. The global_192 spectrum (shaded) has been used throughout chapters 3 and 4 unless otherwise specified.

	Skin			**Oesophagus**		
Spectrum	**ΔΔG**	**Ligand-binding**	**Calcium-binding**	**ΔΔG**	**Ligand-binding**	**Calcium-binding**
Even	$2e^{-5}$	$2e^{-19}$	$2e^{-24}$	$2e^{-5}$	$3e^{-8}$	$6e^{-9}$
global_6	$2e^{-5}$	$1e^{-17}$	$2e^{-22}$	$2e^{-5}$	$3e^{-8}$	$9e^{-9}$
global_6_dedup	$2e^{-5}$	$7e^{-18}$	$1e^{-22}$	$2e^{-5}$	$3e^{-8}$	$9e^{-9}$
global_12	$2e^{-5}$	$6e^{-17}$	$4e^{-23}$	$2e^{-5}$	$1e^{-8}$	$8e^{-9}$
global_12_dedup	$2e^{-5}$	$2e^{-17}$	$3e^{-23}$	$2e^{-5}$	$1e^{-8}$	$7e^{-9}$
global_96	$2e^{-5}$	$6e^{-26}$	$9e^{-21}$	$2e^{-5}$	$2e^{-9}$	$2e^{-7}$
global_96_dedup	$2e^{-5}$	$5e^{-23}$	$3e^{-19}$	$2e^{-5}$	$1e^{-9}$	$2e^{-7}$
global_192	$2e^{-5}$	$2e^{-25}$	$2e^{-22}$	$2e^{-5}$	$2e^{-9}$	$8e^{-8}$
global_192_dedup	$2e^{-5}$	$1e^{-21}$	$2e^{-20}$	$2e^{-5}$	$1e^{-9}$	$1e^{-7}$
global_1536	$2e^{-5}$	$3e^{-16}$	$4e^{-27}$	$2e^{-5}$	$7e^{-7}$	$5e^{-11}$
global_1536_dedup	$2e^{-5}$	$2e^{-17}$	$1e^{-24}$	$2e^{-5}$	$2e^{-7}$	$2e^{-10}$
global_3072	$2e^{-5}$	$2e^{-17}$	$7e^{-21}$	$2e^{-5}$	$2e^{-7}$	$2e^{-9}$
global_3072_dedup	$2e^{-5}$	$1e^{-17}$	$3e^{-19}$	$2e^{-5}$	$4e^{-8}$	$4e^{-9}$
transcript_6	$2e^{-5}$	$2e^{-18}$	$3e^{-23}$	$2e^{-5}$	$3e^{-8}$	$9e^{-9}$

transcript_6_dedup	$2e^{-5}$	$1e^{-18}$	$9e^{-24}$	$2e^{-5}$	$3e^{-8}$	$8e^{-9}$
transcript_12	$2e^{-5}$	$7e^{-19}$	$3e^{-24}$	$2e^{-5}$	$2e^{-9}$	$7e^{-9}$
transcript_12_dedup	$2e^{-5}$	$2e^{-19}$	$9e^{-25}$	$2e^{-5}$	$4e^{-9}$	$4e^{-9}$
transcript_96	$2e^{-5}$	$4e^{-18}$	$2e^{-22}$	$2e^{-5}$	$3e^{-6}$	$3e^{-9}$
transcript_96_dedup	$2e^{-5}$	$4e^{-17}$	$3e^{-19}$	$2e^{-5}$	$1e^{-6}$	$2e^{-9}$
transcript_192	$2e^{-5}$	$3e^{-13}$	$1e^{-20}$	$2e^{-5}$	$3e^{-6}$	$9e^{-9}$
transcript_192_dedup	$2e^{-5}$	$3e^{-17}$	$2e^{-19}$	$2e^{-5}$	$7e^{-7}$	$5e^{-9}$
transcript_1536	$2e^{-5}$	$1e^{-10}$	$7e^{-28}$	$2e^{-5}$	$8e^{-5}$	$8e^{-13}$
transcript_1536_dedup	$2e^{-5}$	$9e^{-9}$	$1e^{-19}$	$2e^{-5}$	$2e^{-5}$	$6e^{-13}$
transcript_3072	$8e^{-5}$	$5e^{-7}$	$7e^{-21}$	0.004	0.004	$5e^{-11}$
transcript_3072_dedup	$4e^{-5}$	$9e^{-8}$	$1e^{-16}$	$4e^{-5}$	0.0009	$1e^{-12}$

The overall statistical results of *NOTCH1* EGF11–12 were not substantially altered by the mutational spectrum assumptions – even assuming all mutations had equal mutation rates produced similar results to far more complex spectra (Figure 3.7, Table 3.2). However, this does not mean the mutation spectrum is irrelevant, and the transmembrane region of *NOTCH1* provides an illustration of the importance of considering the mutational spectrum when analysing the observed patterns of mutation.

There are far fewer mutations in the transmembrane region than the EGF repeats of *NOTCH1* (Figure 3.3a), and the majority of missense mutations occur on a single residue (Figure 3.8a,b). In this case, I could not look for common selected features since it lacked the required diversity of different mutations (see section 3.2.3 below). The hotspot residue, G1753, is on a highly conserved residue adjacent to the S3 cleavage site at which γ-secretase cleaves the NICD from the extracellular domain of NOTCH1 (167) (Figure 3.8a,b). Preventing cleavage or shifting the cleavage site to produce an NICD with a shorter half-life would inactivate, or at least reduce, the *NOTCH1* signal (148, 156, 157).

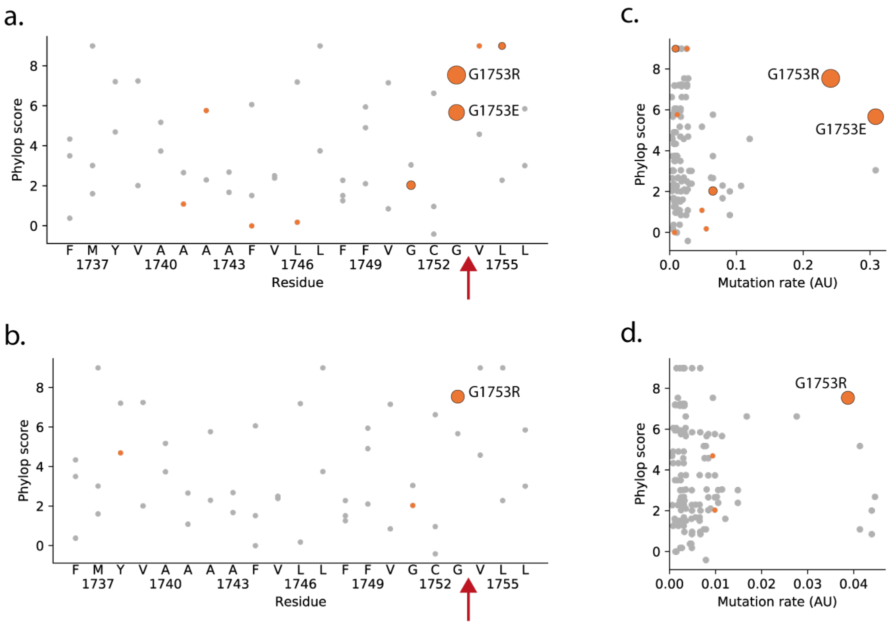

Figure 3.8 Missense mutations in the transmembrane helix of NOTCH1. a,b) Conservation scores from Phylop2 shown on y-axis (Chapter 3 Methods). Single nucleotide missense mutations that occur in the normal human skin (**a**) and oesophagus (**b**), orange, with marker size proportional to the number of times that mutation occurs. Single nucleotide missense mutations that do not occur in the dataset shown in grey. Location of the key S3-V cleavage site indicated by a red arrow. **c,d)** Phylop2 conservation score plotted against relative expected mutation rate (AU=arbitrary units) for missense mutations in the transmembrane helix of NOTCH1 for skin (**c**) and oesophagus (**d**). Single nucleotide missense mutations that occur in the dataset, orange, with marker size proportional to the number of times that mutation occurs. Single nucleotide missense mutations that do not occur shown in grey. The hotspot mutations (G1753R/E) are both highly conserved and have a high expected mutation rate.

Conservation of protein sequence across species is likely to be correlated with function, with residues that are more important for normal protein function more likely to be conserved (160, 168-170). G1753 is highly conserved, but so are other residues in the transmembrane region that were not mutated in the skin or oesophagus datasets (Figure 3.8a,b). This possibly suggests that mutations on these non-mutated residues were not selected as strongly as G1753R/E, despite the high conservation scores. However, looking at both the conservation score and the mutational spectrum, we see that G1753R and G1753E are the only mutations that are both on highly conserved nucleotides and that have high expected mutation rates (Figure 3.8c,d), plausibly explaining why it was these mutations in particular that appeared recurrently in the data sets.

This shows the importance of considering the mutational spectrum when investigating which mutations are selected for. Within *NOTCH1* EGF11–12, unsurprisingly, mutations with a higher expected mutation rate were generally more mutated (Figure 3.9). However, low sample size, imprecise mutation rate estimation, and potential within-category differences in functional impact mean that it is hard to precisely predict which individual mutations will be highly mutated (Figure 3.9). This illustrates one advantage of testing for mutation function selection using the bulk set of mutations instead of focusing on individual hotspot mutations. The following section discusses another advantage of using the bulk set of mutations: that hotspots can lead to false inference of functional selection.

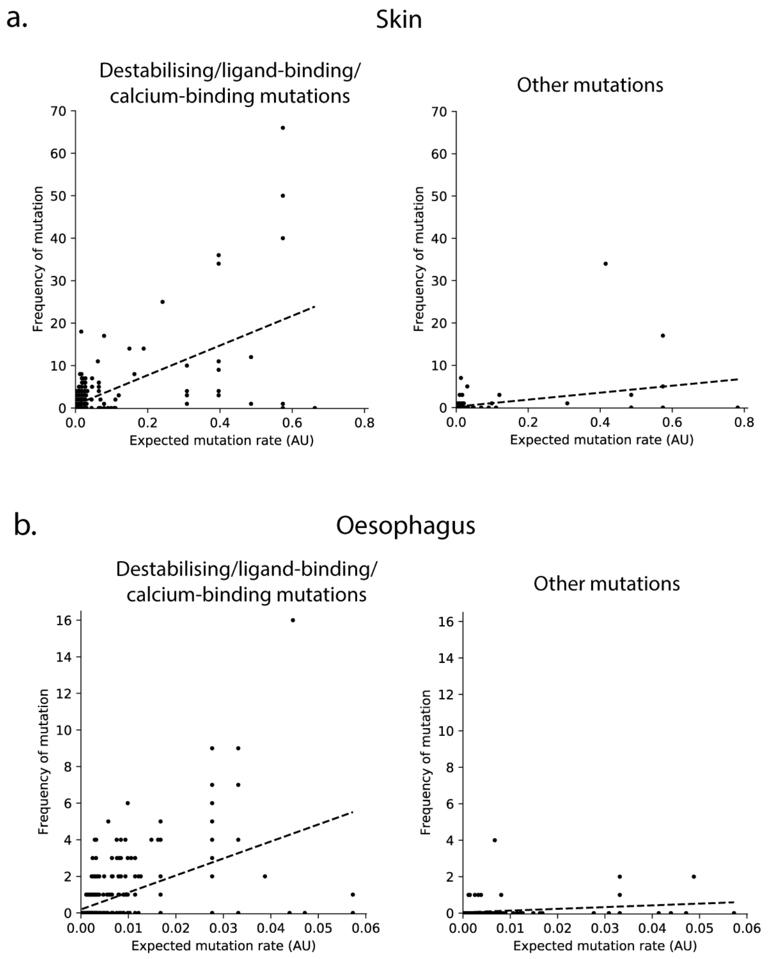

Figure 3.9 Expected and observed mutation rates. Expected mutation rate (Chapter 3 Methods, AU=arbitrary units) vs number of observations for each potential mutation in NOTCH1 EGF11–12 in normal human skin (**a**) and oesophagus (**b**). Destabilising mutations (ΔΔG>2kcal/mol), ligand-binding mutations and calcium-binding mutations shown together on the left-hand scatter plots, all other mutations shown on the right. There is a trend of increasing number of occurrences with higher expected mutation rate (skin, destabilising, ligand-binding and calcium-binding: R^2=0.36, p-value=$3e^{-36}$; skin, other: R^2=0.13, p-value=$5e^{-6}$; oesophagus, destabilising, ligand-binding and calcium-binding: R^2=0.24, p-value=$4e^{-23}$; oesophagus, other: R^2=0.05, p-value=0.007; two-tailed Wald Test for zero slope using SciPy stats.linregress (171)). However, the observed frequency of individual mutations is not very predictable based on only the expected mutation rate and the functional impact category.

3.2.3 Correlation, causation and hotspots

The statistical method tests for a shift in the distribution of a metric between the null model and the observed mutations (Figure 3.1). Ideally, any shift detected will be due to selection of the particular feature that is being tested. However, hotspots can distort the distribution, regardless of the metric used, and can even provide evidence of selection when using random numbers to score the mutations (71). This is not necessarily a problem when just looking for evidence of selection acting on a gene. However, the aim of this work is to infer whether the particular feature is selected, and distortions due to hotspots may give a false impression of the importance of a feature (Figure 3.10a,b)

A simple approach is to check that the trend of selection is still apparent after deduplicating the recurrent mutations (Figure 3.10c,d). P-values from this approach may not be reliable because hotspots may be expected to occur if the feature is strongly selected, and deduplicating mutations both reduces the sample size and could distort the mutational spectrum used for the null model (see section 3.2.2 above). However, it can still be a useful method to explore the influence of hotspots.

I used two null models: one with a spectrum calculated from the full data, as used elsewhere in this chapter; the other with a spectrum calculated using deduplicated data (section 3.2.2 above). The trends of selection in *NOTCH1* EGF11–12 for destabilising mutations, ligand-binding interface mutations and calcium-binding mutations are still apparent and statistically significant after the removal of duplicate mutations (Figure 3.10d-l). This provides more confidence that the significant results obtained for *NOTCH1* EGF11–12 (Figure 3.3, Figure 3.4, Figure 3.5, Figure 3.6) are due to the selection of the features identified.

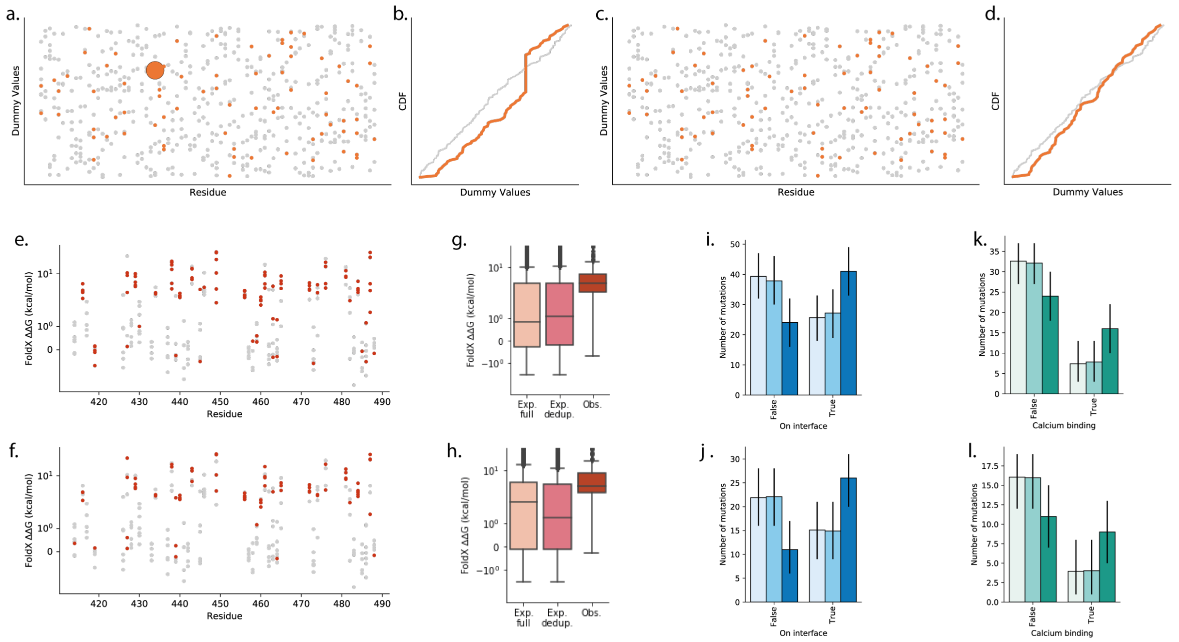

Figure 3.10 The effect of hotspot mutations. a-d) Demonstration of how a hotspot mutation can lead to rejection of the null hypothesis even when the metric has no biological meaning. A synthetic set of mutations was generated, where 74 mutations appear once, and one hotspot mutation appears 25 times. In place of a metric based on a biological feature, random scores between zero and one are generated for each mutation. **a)** Scatter plot of the example mutations. "Observed" mutations shown in orange, "unobserved" mutations in the region shown in grey. The hotspot mutation is the large orange marker. **b)** Cumulative distributions of the metric for the null model, grey, and the "observed" mutations. The large vertical jump in the "observed" CDF is caused by the hotspot mutation. The "observed" distribution is significantly different from the null (p=0.005, two-tailed Monte Carlo test, Chapter 3 Methods). **c)** Scatter plot of the example mutations after removing duplicate mutations. "Observed" mutations shown in orange, "unobserved" mutations in the region shown in grey. The hotspot mutation is now only counted once. **d)** Cumulative distributions of the metric for the null model, grey, and the "observed" mutations after removing duplicate mutations (p=0.3, two-tailed Monte Carlo test, Chapter 3 Methods). **e-l)** Reanalysis of the selection of features in NOTCH1 EGF11–12 after removing duplicate mutations. **g-l**, two null models are tested, either using the full set of mutations to calculate the mutation spectrum (lightest shade, leftmost bar/boxplot) or deduplicating the mutations prior to calculating the spectrum (middle shade, middle bar/boxplot). Observed mutations shown in the darkest shade, rightmost bar/boxplot. **e-h)** ΔΔG of mutations in NOTCH1 EGF11–12. All mutations on the ligand-binding interface or calcium-binding residues excluded from both the null model and observed data. Single nucleotide missense mutations that occur in the skin (**e**) and oesophagus (**f**) datasets, red. Single nucleotide missense mutations that do not occur in the data shown in grey. **g,h)** Distribution of calculated ΔΔG values of missense mutations in skin (**g**) and oesophagus (**h**). Distribution expected under the neutral null hypothesis using all mutations for the spectrum, left, distribution expected under the neutral null hypothesis using deduplicated mutations for the spectrum, middle, and the distribution observed after removing duplicate mutations, right. Skin: p<2e-5 for both null models, n=113; oesophagus: p<2e-5 for both null models, n=76; two-tailed Monte Carlo test (Chapter 3 Methods). **i,j)** Expected and observed counts of mutations on or off the ligand-binding interface of NOTCH1 EGF11–12, where mutations with a calculated ΔΔG>2 kcal/mol and mutations on calcium-binding residues have been excluded. Skin: p<0.0007 for both null models, n=65; oesophagus: p<0.0004 for both null models, n=37; two-tailed binomial test (Chapter 3 Methods). Error bars show 95% confidence intervals (Chapter 3 Methods). **k,l)** Expected and observed counts of mutations on calcium-binding residues in NOTCH1 EGF11–12 where mutations with a calculated ΔΔG>2 kcal/mol or on the ligand-binding interface have been excluded. Skin: p<0.005 for both null models, n=40; oesophagus: p<0.01 for both null models, n=20; two-tailed binomial test (Chapter 3 Methods). Error bars show 95% confidence intervals (Chapter 3 Methods).

3.2.4 Multiple hypotheses could explain the distribution of mutations across *NOTCH1*

The examples covered so far in this chapter have only looked at small regions of the gene. More information is potentially available from the distribution of mutations across the whole gene. For example, the ligand-binding EGF repeats were much more mutated than the rest of *NOTCH1* (Figure 3.3a). There are a number of possible explanations for this, which are not mutually exclusive.

Firstly, it could be a result of a variable background mutation rate across the gene; that is to say, some regions of the gene are inherently more prone to mutation. Synonymous mutations have been used to estimate background mutation rates (172) as they are assumed to be neutrally selected. There is a significantly higher number of synonymous mutations in the EGF11–12 region than the rest of the gene (Figure 3.11a,b, skin: expected=3%, observed=9%, p=5.6e^{-9}, n=404; oesophagus: expected=3%, observed=19%, p=0.0005, n=31; two-tailed binomial test, Chapter 3 Methods), suggesting that a variable mutation rate could be contributing to the uneven distribution of missense mutations. However, some synonymous mutations may have been created in the same mutation event as a nearby missense mutation and not successfully merged into a single multi-nucleotide variant (11), meaning the synonymous mutation is carried along as a passenger. It is also possible that some of the synonymous mutations are not functionally neutral (173) and therefore may be increased or decreased in number due to positive or negative selection, meaning they may not give a true indication of mutation rate. Secondly, it may be that the DNA sequence of the highly mutated regions has a higher probability of acquiring mutations due to the mutational spectrum. However, this does not appear to be the case for missense mutations (Figure 3.11c,d). Thirdly, it could be that a higher proportion of potential mutations are functionally impactful in the ligand-binding region than other sections of the gene. For example, we have seen that approximately two-thirds of potential single nucleotide substitutions in the EGF11–12 repeats are likely to be selected. If this proportion is lower in the rest of the gene, this would result in a smaller number of observed mutations there. Fourthly, the effects of a disruptive mutation may be milder in the non-ligand-binding regions. It might be that disruption of ligand binding completely stops the *NOTCH1* signal, but, for example, a mutation in the transmembrane helix reduces rather than entirely stops the signal (156). Lastly, it has been suggested that NOTCH receptors may form homodimers or clusters when they bind to a ligand (174, 175). Potentially, this could mean that a mutant in the ligand-binding region could disrupt the activity of the wild type allele as well as the mutant allele, reducing the *NOTCH1* signal more than having completely lost one *NOTCH1* allele (Figure 3.11e). This dominant negative effect has been observed for some missense mutations in the *TP53* DNA-binding domain (176). The last two hypotheses suggest that the fitness of *NOTCH1* mutant clones depends on the dose of *NOTCH1* signalling. This is consistent with the frequent observation of loss-of-heterozygosity copy number variants associated with *NOTCH1* mutations (10), which strongly suggests that losing both wild type alleles of *NOTCH1* provides a stronger growth advantage than losing a single allele. This *Notch1* dose-dependent fitness was also confirmed to be the case in lineage tracing studies of mice (Chapter 5).

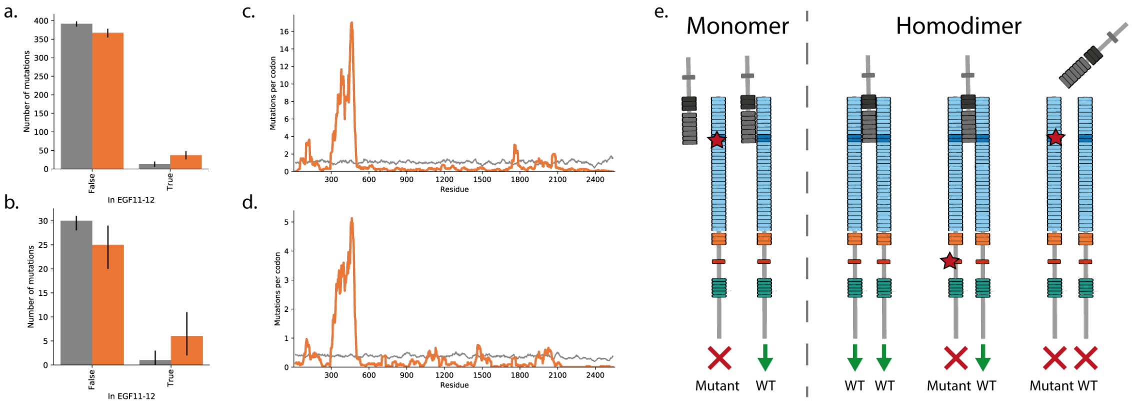

Figure 3.11 Potential causes of uneven missense mutation distribution across NOTCH1. a,b) The counts of synonymous mutations expected under the neutral null model, grey, and observed, orange, in EGF11–12 in skin (**a**) and oesophagus (**b**) (skin: p=5.6e^{-9}, n=404; oesophagus: p=0.0005, n=31; two-tailed binomial test, Chapter 3 Methods). Error bars show 95% confidence intervals (Chapter 3 Methods). **c,d)** Sliding windows showing the distribution of missense mutations across *NOTCH1* expected under the neutral null model, grey, and observed, orange, in skin (**c**) and oesophagus (**d**). **e)** Potential impact of mutations if NOTCH1 (shown with domains as in Figure 3.3a) binds to the ligand (grey) as a monomer or a homodimer. If NOTCH1 binds to the ligand as a monomer (left), then a mutation in the ligand-binding EGF repeats may entirely stop signalling from the mutant allele (red cross) but the wild type (WT) allele still functions normally (green arrow). If NOTCH1 binds the ligand as a homodimer (right), then a mutant away from the ligand-binding region may stop signalling from the mutant allele but not affect the WT allele. However, a mutant in the ligand-binding region may prevent the ligand from binding, and therefore stop signalling from the WT allele as well. If NOTCH1 binds to the ligand in pairs, a heterozygous *NOTCH1* mutant would mean only one quarter of NOTCH1 pairs would be WT-WT, so a mutant that prevents ligand binding could reduce *NOTCH1* signalling by three-quarters.

3.3 Discussion

In this chapter I adapted a statistical method for cancer driver gene discovery to look for selected features of mutations in a gene. By using this method, structural and functional information can be drawn from the increasingly large amount of DNA sequencing data available. The results can be used to support driver gene detection methods such as dN/dS by showing that the functional impact of the enriched mutations is biologically meaningful. This can be especially useful if the reliability of the driver gene detection method is in doubt (see Chapter 2). Manual investigation of hotspot mutations can provide similar information (10) but has disadvantages: it can be time-consuming, does not leverage the information provided by rarer mutations and does not statistically test the selection of mutation features.

Some caution must be applied when interpreting the results of the method. Although protein misfolding and ligand binding are well-known processes that control protein activity (162, 177), correlation does not mean causation, and significant selection may be found for a feature that correlates (coincidentally or otherwise) with the true selected feature. This is also a test for selection *relative to the rest of the tested region*, and does not necessarily directly translate to positive or negative selection. I have shown that if there are multiple selected features in a region, testing for one individual feature at a time may lead to misleading results, but that this can be corrected by conditioning against the other, confounding selected features. This potential for misleading results also highlights an advantage of using general functional impact scores when looking for driver genes, rather than focusing on single impact types (63, 71). A future improvement would be to incorporate an estimate for background mutation rate in a region (62, 72) and hence test for absolute rather than relative selection.

To separate initial mutation incidence from mutational selection, the method assumes that nucleotide substitutions occur based on the mutational spectrum of the dataset. This follows the assumptions used in many driver discovery tools to improve the null model of mutations expected under neutral selection. Altering the form of the mutational spectrum does not alter the conclusions of the analysis of *NOTCH1* EGF11–12, suggesting that the inference of selection is robust and not an artefact of inappropriate spectrum assumptions. However, the mutational spectrum should not be ignored, as a correlation is seen between the expected mutation rate and the observed mutation frequency, and the mutations observed in the transmembrane region illustrate the principle that mutations with a high expected mutation rate and a strongly selected functional impact are more likely to be observed in the sequencing data. Together, the results highlight the importance of examining mutagenic processes alongside the biophysical mechanism of mutant action.

The statistical method introduced here is used in the next chapter to analyse patterns of selection in five more genes sequenced in the normal skin and oesophagus. This serves to further investigate the evidence for non-neutral clonal selection in DNA sequencing datasets of normal epithelia, and to explore the strengths and weaknesses of the statistical method in more depth.

3.4 Chapter 3 Methods

3.4.1 Data

I used mutations detected in normal human oesophagus (10) and normal human skin (11). Both studies used a grid of adjacent tissue samples. Large clones could spread over multiple samples. To avoid double counting of such clones, I used the mutations list where mutations that were seen repeatedly in nearby samples were assumed to be from a single clone and were merged (10, 11) (Chapter 2, Figure 2.3). The data are available from the original publications (10, 11).

3.4.2 Protein structures

The 2VJ3 structure of NOTCH1 EGF11-12 (165) was used for the ΔΔG analysis, to define calcium-binding residues and for structural distance calculations.

Residue numbers in the PDB files were aligned to residue numbers in the UniProt protein sequences using SIFTS (178). Distances between residues in structures were calculated using the Python package MDAnalysis (179).

3.4.3 FoldX

For each PDB file, the FoldX command *RepairPDB* was run to minimise steric clashes and optimize residue orientation. Then the FoldX command *PositionScan* was run for every residue of the protein chains of interest in the structure. This command mutates each residue to all other amino acids and calculates the ΔΔG value for each mutation. Default FoldX settings were used for both the *RepairPDB* and *PositionScan* commands.

Some analyses required a threshold to discriminate destabilising mutations from non-destabilising mutations. Unless otherwise noted, I used a threshold ΔΔG value of 2 kcal/mol because this has been used in previous studies to define mutations which are highly destabilising (180-183).

3.4.4 Ligand-binding interface residues

The ligand-binding interface residues in EGF11 and EGF12 have been identified for the rat NOTCH1 bound to the ligands JAG1 and DLL4 (160, 161). *NOTCH* genes and ligands are highly conserved between species meaning that the results from the rat protein can be applied to human NOTCH1 (169) (Figure 3.12). The ligand-binding surface is very similar for both ligands (161) and I therefore chose to use the union of both sets of ligand-binding residues (Figure 3.4a, Table 3.3).

```
Human   DVDECSLGANPCEHAGKCINTLGSFECQCLQGYTGPRCEIDVNECVSNPCQNDATCLDQIGEFQCICMPGYEGVHCE
Rat     DVDECALGANPCEHAGKCLNTLGSFECQCLQGYTGPRCEIDVNECISNPCQNDATCLDQIGEFQCICMPGYEGVYCE
```

Figure 3.12 Alignment of EGF11–12 of human and rat NOTCH1 protein sequences. Residues 412 to 488 of each sequence shown, residues which differ highlighted in yellow.

Table 3.3 Ligand-binding interface residues of NOTCH1 EGF11–12. Based on Figure S3 of (161).

	Interface residues
NOTCH1 EGF11–12	413, 415, 418, 420, 421, 422, 423, 424, 425, 435, 436, 444, 447, 448, 450, 451, 452, 454, 466, 467, 468, 469, 470, 471, 475, 477, 478, 479, 480

3.4.5 Calcium-binding mutations

The calcium-binding residues in EGF11–12 of NOTCH1 were defined using MetalPDB (164) and the 2VJ3 (165) structure (Table 3.4).

Table 3.4 Calcium-binding residues of NOTCH1 EGF11–12. Based on MetalPDB (164) and the structure 2VJ3 (165).

	Calcium-binding residues
NOTCH1 EGF11–12	412, 413, 415, 431, 432, 435, 452, 453, 455, 469, 470

3.4.6 Conservation scores

Phylop conservation scores (184) for each position were downloaded via the UCSC genome browser, http://hgdownload.soe.ucsc.edu/goldenPath/hg19/phyloP100way/ (185).

3.4.7 Images of protein structures

VMD (186) was used to visualise protein structures and images were rendered using Tachyon (187).

3.4.8 Mutational spectrum calculation

Firstly, all genes containing an exonic mutation in the data sets were found using exon locations in GRCh37.p13 downloaded from Ensembl Biomart (188). For each of these genes, the longest transcript was selected and alternative transcripts discarded. A trinucleotide context was calculated in the direction of the protein transcription for every nucleotide in the coding sequence of each transcript and applied to each observed mutation. For each data set, a mutation rate was calculated for each single nucleotide substitution type in each trinucleotide context by dividing the total number of observed mutations of that trinucleotide change by the number of times the trinucleotide context occurs in the included transcripts.

The mutational spectrum can be assumed to be gene specific (71), to depend on a larger nucleotide context (62), to be distorted by selected hotspot mutations (166), or to be symmetrical in terms of the transcribed strand direction (e.g. AAA>ACA is assumed to have the same mutation rate as to TTT>TGT, so there are 96 possible trinucleotide changes instead of 192) (63). The above method can therefore be adapted by changing the number of context bases, whether to test using a global spectrum for the dataset or one calculated from just the transcript to test, deduplicating recurrent (and therefore possibly selected) mutations before calculating the spectrum, or by assuming there is no difference between mutation rate for +/– strands. I show the results of changing these assumptions in section 3.2.2.

Except where otherwise stated, all calculations were made using a trinucleotide, transcribed strand non-symmetric (192) spectrum calculated from all genes in the dataset.

3.4.9 Monte Carlo test

Each possible single nucleotide change was enumerated for the region to be tested. For each of these potential mutations, a relative mutation rate (see section 3.4.8 above) and a metric score, e.g. ΔΔG, was calculated. The null hypothesis of neutral selection assumed that the probability of observing each mutation, and therefore the distribution of metric scores, is determined solely by the mutational spectrum. A cumulative distribution of scores under the null hypothesis was calculated, the score for each possible mutation converted to a cumulative distribution function (CDF) score, and the sum of these CDF scores taken as the test statistic. Using the summed CDF score instead of the mean of the raw scores for the test statistic means that the test is less sensitive to outlier values. Using the median would also be less sensitive to outliers than the mean, but would be inappropriate for testing discrete metrics. The best choice of statistic to use in the test may depend on the particular data and the question being asked (for example, if the outlier mutations are crucial and reliably scored, then mean may be more appropriate). However, the CDF score sum is a robust option that works for the metric scores tested in this project.

Let *n* be the number of observed mutations in the region to test. A large number *N* (here N=100000) of random draws of *n* values from the null distribution of CDF scores was made, and for each draw, *i,* the sum of the CDF scores, s_i, was calculated. The sum of the CDF scores for the observed mutations, s_{obs}, was also calculated. Then, *b*, the number of the s_i values that were more extreme than s_{obs}, was counted. This was converted into a two-tailed p-value by multiplying the exact Monte Carlo p-value (189) by two, i.e.

$$p = 2 \times \frac{(b+1)}{N+1}$$

where

$$b = \min\left(\sum_{i=0}^{N} \begin{cases} 1, & s_i \leq s_{obs} \\ 0, & s_i > s_{obs} \end{cases}, \quad \sum_{i=0}^{N} \begin{cases} 0, & s_i < s_{obs} \\ 1, & s_i \geq s_{obs} \end{cases} \right)$$

The minimum p-value possible with N=100000 is just under 2e^{-5} (2/100001).

The central limit theorem means that, under certain conditions, the null distribution from this Monte Carlo test can be approximated by a normal distribution. The normal approximation is faster to calculate and allows for more extreme p-values to be calculated (which would otherwise require a much larger number of random draws to calculate using the Monte Carlo test). However, as this normal approximation is not appropriate for small sample sizes and skewed score distributions (190), it has not been used in this thesis.

3.4.10 Binomial test

In the special case that there are only two possible values for the metric score (e.g. on/not on an interface), the binomial test can be used. This means very small p-values can be accurately calculated and it is faster to run than the Monte Carlo test. Similar to the Monte Carlo test, the null model is calculated from the mutational spectrum, the nucleotide sequence of the gene or gene region, and the number of observed mutations. I used the Python package SciPy (171) to calculate the p-values. The method to calculate the two-tailed p-value differs slightly from the method described above for the Monte Carlo test, but otherwise the results for Monte Carlo test with infinitely large N would be equivalent to the results of the binomial test.

3.4.11 Confidence intervals

95% confidence intervals were calculated by taking 10000 random samples from the null distribution or by bootstrapping 10000 random samples with replacement from the observed data.

Chapter 4

Inferring structure–function relationships in multiple proteins using somatic mutations

4.1 Introduction

In the previous chapter I introduced a method for testing selection of functional categories of mutations within a gene. The method was validated and its assumptions explored using the example of *NOTCH1* mutations in normal skin and oesophagus. Having established the overall validity of the method and identified its potential pitfalls, in this chapter I apply it to several further genes sequenced in normal human oesophagus and skin. To show the flexibility of the approach, the genes chosen exhibit a mix of positive and negative selection, and selection of loss-of-function and gain-of-function mutations. I also demonstrate how the method can be adapted to statistically compare mutational patterns in the same gene across different datasets.

I start by looking at the patterns of missense mutations in three genes – *NOTCH2*, *NOTCH3,* and *TP53* – that have positive selection of loss-of-function truncating (nonsense and splice) mutations. *NOTCH2* and *NOTCH3* are both members of the Notch family of genes, and are highly similar to *NOTCH1*, analysed in the previous chapter. *TP53* is another strong driver of clonal expansion in both the normal skin and oesophagus. I finish by analysing the missense mutations in two genes in which truncating mutations are under negative selection, *PIK3CA* and *FBXW7.*

4.2 Results

4.2.1 Similar patterns of selection in EGF11–12 of *NOTCH1* and *NOTCH2*

NOTCH1 and *NOTCH2* are highly similar genes, with the same mechanism of activation (Chapter 3, Figure 3.2). Both are drivers of clonal expansion in skin and oesophagus and both have a peak of missense mutations at the ligand-binding EGF repeats 11 and 12 (Figure 4.1a) (10, 11, 169). Therefore, based on the results for *NOTCH1*, I considered destabilising, ligand-binding interface and calcium-binding mutations as potentially selected groups in *NOTCH2* EGF11–12 (Figure 4.1b,c; total missense mutations in region: skin 178, oesophagus 38). To remove confounding sources of selection, I tested each category while excluding mutations in the other two categories from both the null and observed distributions. Destabilising mutations (excluding ligand-binding and calcium-binding mutations; skin: $p<2e^{-5}$, n=98; oesophagus: $p<2e^{-5}$, n=25; two-tailed Monte Carlo test, Chapter 3 Methods, Figure 4.1d,e), ligand-interface mutations (excluding $\Delta\Delta G>2$kcal/mol and calcium-binding mutations; skin: expected=27%, observed=56%, $p=2e^{-5}$, n=48; oesophagus: expected=23%, observed=67%, p=0.03, n=6; two-tailed binomial test, Chapter 3 Methods, Figure 4.1f,g), and calcium-binding mutations (excluding $\Delta\Delta G>2$kcal/mol and ligand-binding mutations; skin: expected=14%, observed=34%, p=0.003, n=32; oesophagus: expected=15%, observed=60%, p=0.03, n=5; two-tailed binomial test, Chapter 3 Methods, Figure 4.1h,i) were significantly enriched compared to the null models. The results show that selection of missense mutations in the *NOTCH2* ligand-binding EGF repeats follows similar patterns to selection in *NOTCH1* in both skin and oesophagus.

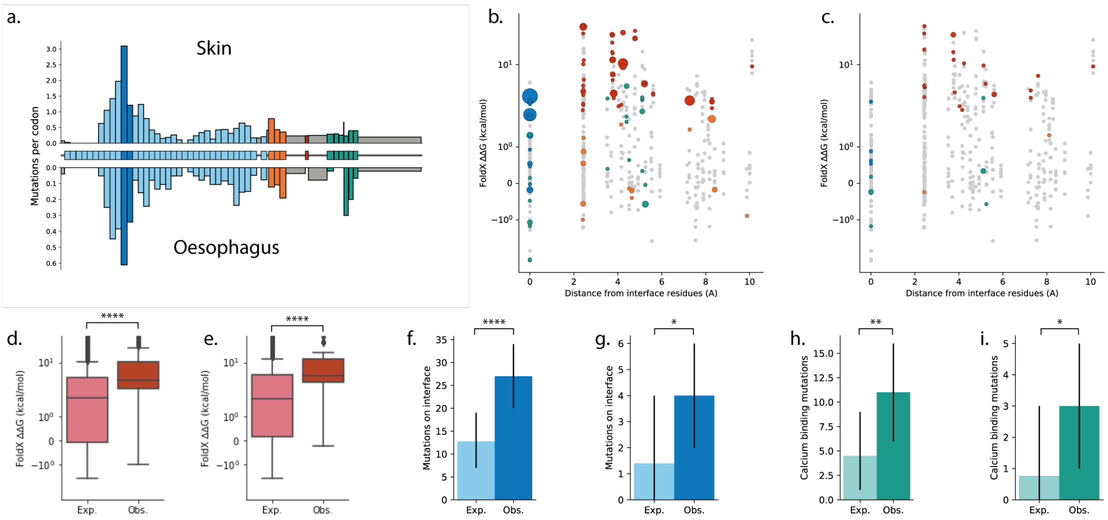

Figure 4.1 Selected features of missense mutations in NOTCH2 EGF11–12. **a)** Missense mutation frequency across the domains of *NOTCH2* in normal skin (top) and normal oesophagus (bottom). Domain definitions from UniProt (163). Where the gap between domains is only a single residue, mutations from this residue are included in the subsequent domain. EGF repeats, blue; EGF11–12, dark blue; LNR repeats, orange; transmembrane region, red; ankyrin repeats, green; other regions, grey. **b,c)** Calculated ΔΔG plotted against distance from the NOTCH2 EGF11–12 ligand-binding interface for skin (**b**) and oesophagus (**c**). Single nucleotide missense mutations that occur in the dataset, with marker size proportional to the number of times that mutation occurs, shown in green if the residue is calcium binding, blue if the residue is on the ligand-binding interface, red if the mutation has ΔΔG>2kcal/mol, orange otherwise. Single nucleotide missense mutations that do not occur in the dataset shown in grey. **d,e)** Distribution of calculated ΔΔG values of missense mutations after excluding mutations on the ligand-binding interface and calcium-binding residues. Distribution expected under the neutral null hypothesis, light red, and the distribution observed, dark red. **f,g)** Counts of NOTCH2 EGF11–12 mutations occurring on the ligand-binding interface, having excluded destabilising mutations (with calculated ΔΔG > 2 kcal/mol) and calcium-binding mutations, in skin (**f**) and oesophagus (**g**). Expected counts under the neutral null hypothesis, light blue; counts observed, dark blue. **h,i)** Counts of NOTCH2 EGF11–12 mutations occurring on the calcium-binding residues, having excluded destabilising mutations (with calculated ΔΔG > 2 kcal/mol) and ligand-binding mutations, in skin (**h**) and oesophagus (**i**). Expected counts under the neutral null hypothesis, light green; counts observed, dark green. P-values calculated using a two-tailed Monte Carlo test for **d,e** and using a two-tailed binomial test for **f-i** (Chapter 3 Methods). Error bars in **f-i** show 95% confidence intervals (Chapter 3 Methods). ****P≤0.0001, **P≤0.01, *P≤0.05.

4.2.2 Selection in *NOTCH3* differs in skin and oesophagus

For *NOTCH1* and *NOTCH2*, the patterns of selection in skin and oesophagus were highly consistent. *NOTCH3* is highly similar to *NOTCH1* and *NOTCH2*, and has the same mechanism of activation (Chapter 3, Figure 3.2). For *NOTCH3*, however, differences can be seen between the two tissues. The gene is positively selected in both data sets, but while in the oesophagus the selection acts on both missense and truncating mutations (10), in skin the selection seems to act mostly on truncating mutations, with missense mutations appearing to be under neutral selection (11). In the oesophagus data (145 missense mutations), there is a peak of missense mutations in EGF 10–11 of *NOTCH3*, which is the ligand-binding region equivalent to EGF11–12 of *NOTCH1* and *NOTCH2* (191) (Figure 4.2a). However, there is not a clear peak in the skin data (527 missense mutations) where the missense mutations are spread more evenly across the entire gene (Figure 4.2a).

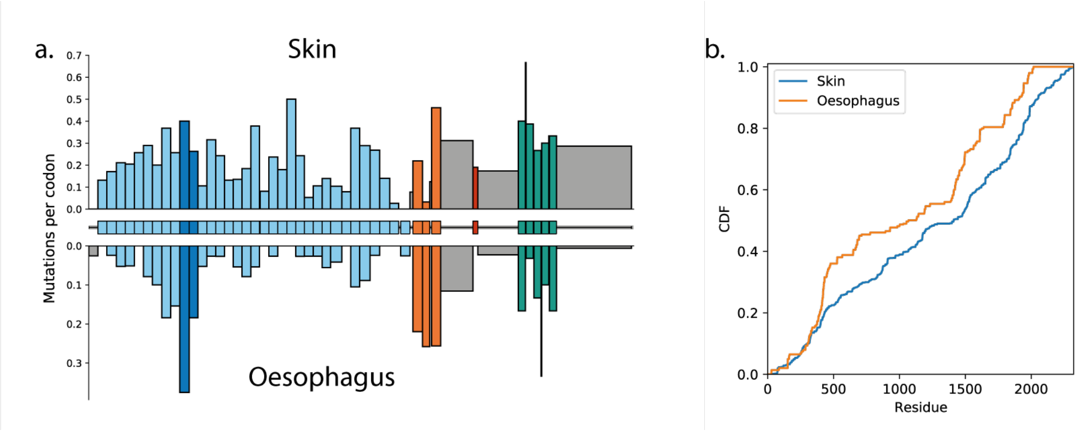

Figure 4.2 Distributions of missense mutations across *NOTCH3* in skin and oesophagus. a) Missense mutation frequency across the domains of *NOTCH3* in normal skin (top) and normal oesophagus (bottom). Domain definitions from UniProt (163). Where the gap between domains is only a single residue, mutations from this residue are included in the subsequent domain. EGF repeats, blue; EGF11–12, dark blue; LNR repeats, orange; transmembrane region, red; ankyrin repeats, green; other regions, grey. **b)** Cumulative distribution of missense mutations across *NOTCH3* in skin and oesophagus, adjusted for the mutational spectrum. The mutation distributions significantly differ in the two tissues, p=0.003, weighted Anderson–Darling Test, Chapter 4 Methods.

The distribution of mutations can be compared in the two datasets by compensating for the difference in mutational spectrum (Chapter 4 Methods, Chapter 1 Figure 1.4). There is a significant difference between the distributions (Figure 4.2b, p=0.003, weighted Anderson–Darling test, Chapter 4 Methods). Of the six genes analysed in this chapter and Chapter 3, only *NOTCH3* and *PIK3CA* (see section 4.2.5 below) have significant differences between the distribution of missense mutations in skin and oesophagus after multiple-test correction (*NOTCH3* q=0.019, *PIK3CA* q=0.035, weighted Anderson–Darling test with Benjamini–Hochberg correction (192), Chapter 4 Methods). This result for *NOTCH3* is consistent with the dN/dS analysis that finds strong selection of missense mutations in the oesophagus but not in skin. As seen in the oesophagus data, strong selection of missense mutations is likely to appear as a large enrichment of mutations in key functional regions such as the ligand-binding EGF repeats. In the absence of such selection, missense mutations are likely to be more scattered across the gene, as seen in the skin data.

Unfortunately, the number of missense mutations in *NOTCH3* EGF 10–11 is low in both data sets (skin n=26; oesophagus n=22), and no significant trends were found when testing for selection of the mutation categories selected in *NOTCH1* and *NOTCH2* (p>0.13 for all tests in both skin and oesophagus, equivalent tests to those run on NOTCH2 EGF11–12 above, analysed using two homology models of NOTCH3 EGF 10–11, Chapter 4 Methods).

NOTCH1 and NOTCH2 are more structurally similar to each other than they are to NOTCH3 (193), so it is interesting to note that the mutation selection patterns in *NOTCH1* and *NOTCH2* are more similar to each other than they are to the selection patterns in *NOTCH3*. The functions and tissue distribution of the Notch family of genes are complex, and some aspects remain poorly understood (194). There is some overlap in the function of the Notch family genes, but each gene also has distinct functions (193). Furthermore, Notch activity depends both on which particular pairs of ligands and receptors are engaged and on the metalloproteinase availability in the tissue (194). Altogether, it is therefore unsurprising that we see both similarities and differences in mutational selection between the Notch genes and between tissues. The function of the Notch genes in the epithelial tissues, and how this relates to selection operating on the genes, is an interesting area for further investigation.

4.2.3 Missense mutations in *TP53* destabilise the protein and disrupt DNA binding

Along with *NOTCH1*, *TP53* is one of the two most significantly selected genes in both the oesophagus and skin (10, 11). Truncating mutations in *TP53* are strongly selected (10, 11) and most missense mutations in *TP53* occur in the DNA-binding domain (total missense mutations in DNA-binding domain: skin 898, oesophagus 380; Figure 4.3a). I therefore hypothesised that, similar to *NOTCH1* and *NOTCH2*, the missense mutations in the DNA-binding domain are inactivating *TP53* by either causing protein misfolding or by affecting the interaction surface of p53 (in this case the binding interface with DNA). I found strong enrichment of destabilising mutations (excluding mutations within 5Å of the DNA molecule; skin: $p<2e^{-5}$, n=760; oesophagus: $p<2e^{-5}$, n=338; two-tailed Monte Carlo test, Chapter 3 Methods, Figure 4.3b,c,f) and of mutations close to the DNA molecule (excluding mutations with ΔΔG>2 kcal/mol; skin: $p<2e^{-5}$, n=395; oesophagus: $p<2e^{-5}$, n=154; two-tailed Monte Carlo test, Chapter 3 Methods, Figure 4.3d,e,f). Approximately 42% of potential mutations in the *TP53* DNA-binding domain are either destabilising (ΔΔG>2 kcal/mol) or within 5Å of the bound DNA molecule, and together these two categories contain 71% and 69% of observed mutations in skin and oesophagus respectively.

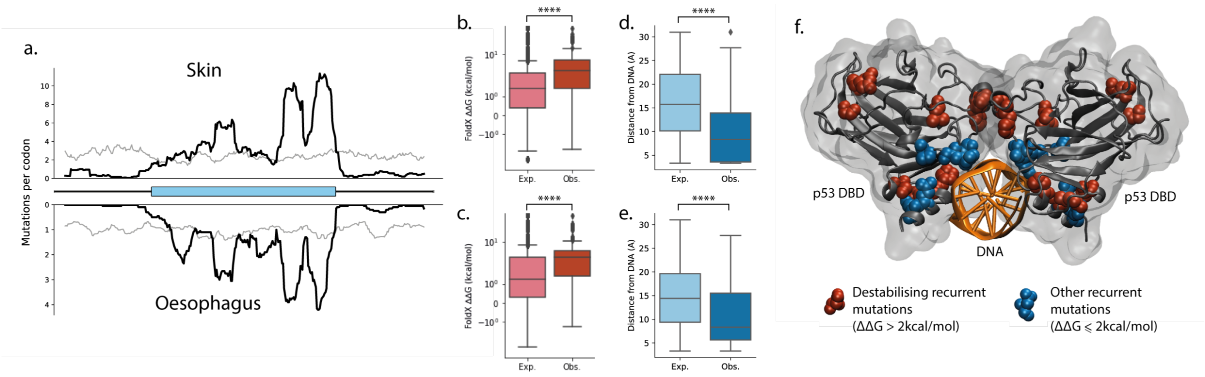

Figure 4.3 Selection of missense mutations in *TP53*. a) Sliding window of missense mutation frequency across *TP53* in skin (top) and oesophagus (bottom). Observed distribution shown with a bold black line, the expected distribution based on the mutational spectrum shown with a thin grey line. DNA-binding domain shown as a blue bar. Domain definition from UniProt (163). **b,c)** Distribution of calculated ΔΔG values of missense mutations in skin (**b**) and oesophagus (**c**) after excluding mutations within 5Å of the DNA molecule. Distribution expected under the neutral null hypothesis, light red, and the distribution observed, dark red. **d,e)** Distribution of distances of missense mutations from DNA in skin (**d**) and oesophagus (**e**) after excluding mutations with ΔΔG > 2 kcal/mol. Distribution expected under the neutral null hypothesis, light blue, and the distribution observed, dark blue. **f)** Structure of two copies of the p53 DNA-binding domain (DBD) bound to DNA (orange). Residues containing missense mutations that occur at least 10 times in the skin data set are highlighted. Highly destabilizing mutations (ΔΔG > 2 kcal/mol) shown in red. Non-destabilizing mutations are shown in blue. ****P≤0.0001, two-tailed Monte Carlo test, Chapter 3 Methods.

4.2.4 Selection of destabilising mutations correlates with selection of nonsense mutations

In *NOTCH1*, *NOTCH2* and *TP53*, I found positive selection of missense mutations that are likely to reduce normal protein function. In each case this includes mutations that strongly destabilise the protein. Among the genes sequenced in the normal skin, there is a strong correlation between selection of destabilising mutations and selection of truncating mutations (Figure 4.4**,** Pearson's correlation coefficient=0.89, two-tailed p=8e^{-6}), as might be expected for two types of mutation which produce a similar outcome (loss of function). In the next two sections I will investigate selection of missense mutations in genes (*PIK3CA* and *FBXW7*) with negative selection of truncating mutations.

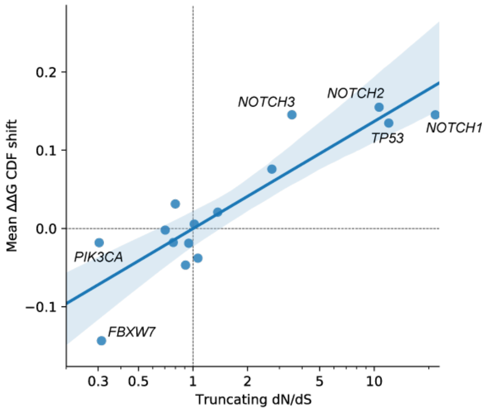

Figure 4.4 Correlation between the dN/dS ratio of truncating mutations and selection of destabilising missense mutations in normal skin. The selection of destabilising missense mutations is shown as the shift in the distribution of ΔΔG values between the neutral null model and the observed data. The CDF values were used instead of the raw ΔΔG values to reduce the influence of extreme outliers (Chapter 3 Methods). The ΔΔG CDF shift is the difference between the mean of the null and observed distributions of ΔΔG CDF values. It is a value between -0.5 (all observed mutations have the minimum ΔΔG possible in the structure) and 0.5 (all observed mutations have the maximum ΔΔG possible in the structure). A ΔΔG CDF shift of 0 means that the mean values of the null and the observed ΔΔG CDF distributions are equal. Structures of wild type proteins that contain at least 50 missense mutations in the skin dataset were used for the ΔΔG calculations (Chapter 3 Methods), and the average ΔΔG CDF shift of all structures for each gene was calculated. Pearson's correlation coefficient=0.89, two-tailed p=8e^{-6}, mean ΔΔG CDF shift vs logarithm of truncating dN/dS ratio. Blue line and shaded region show the linear regression and 95% confidence interval of the regression estimate calculated and plotted using the Python package Seaborn (195).

4.2.5 Positive and negative selection in *PIK3CA*

The lipid-kinase PI3K is made up of the subunits p110α, encoded by *PIK3CA*, and p85α, which binds to p110α to inhibit PI3K activity (196) (Figure 4.5a). In the normal skin, dN/dS analysis found missense mutations were under approximately neutral selection, while nonsense and splice mutations are negatively selected (11) (Figure 4.4). In contrast, missense mutations in the normal oesophagus are positively selected, with recurrent mutations of the H1047R hotspot (10) (Figure 4.5b).

I compared the observed missense mutations to those catalogued in Clinvar (197), a curated database of disease associated genetic mutations. In both skin and oesophagus, I found a significant enrichment of mutations matching those labelled in Clinvar as pathogenic/likely pathogenic (skin: expected=2%, observed=12%, p=4e^{-9}, n=167; oesophagus: expected=2%, observed=45%, p=3e^{-25}, n=49; two-tailed binomial test, Chapter 3 Methods, Figure 4.5c,d). In particular, the matching mutations were associated with various cancers and overgrowth diseases known to be driven by *PIK3CA* activating mutations.

To further investigate the missense mutations, I looked for enrichment of mutations in key regions of the protein that undergo conformational changes when PI3K is activated by a phosphopeptide or binds to a membrane (198) (Table 4.1). These regions include the locations of many cancer-linked *PIK3CA* mutations such as G106V, G118D, N345K, E542K, E545K and H1047R. In skin there is enrichment in residues 100–119, 444–473, and 962–980 (Figure 4.5e-g,j, Table 4.1). Although there are many recurrent mutations in the gene (Figure 4.5b), these trends are still apparent when removing duplicate mutations (Table 4.2**,** section 3.2.3 in Chapter 3), with only the 444–473 region not quite reaching statistical significance. Residues 100–119 are part of the link between the adapter-binding domain (ABD) and the Ras-binding domain (RBD). Mutations in this region have been found to cause conformational changes normally associated with membrane binding and to increase kinase activity (198). Residues 444–473 are on the interface between p110α and p85α (Figure 4.5j), the disruption of which can lead to activation of PI3K (198). Residues 962–980 are in the kinase domain and are involved with membrane binding (198, 199). Together with the enrichment of mutations in Clinvar, these analyses provide strong evidence that, despite the apparently neutral results of dN/dS analysis, selection is acting on *PIK3CA* missense mutations in normal skin. The analyses suggest that there is negative selection of inactivating mutations (shown by the dN/dS ratio of truncating mutations of less than one, Figure 4.4) and positive selection of activating mutations. The combination of these opposing directions of selection may contribute to the inference of neutral selection when missense mutations in *PIK3CA* are analysed as a whole (62).

In oesophagus there is enrichment of missense mutations in residues 100–119 (also enriched in skin) and 1039–1055 (Figure 4.5h-j, Table 4.1). Residues 1039–1055 are at the C-terminal end of the kinase domain and contain the strongly activating H1047R mutation (198). Despite the high frequency of hotspots – the H1047R hotspot in particular – both regions remain significantly selected after removing duplicate mutations (Table 4.2). There is a significant difference in the distribution of missense mutations between oesophagus and skin after correcting for differences in the mutation spectra (p=0.01, weighted Anderson–Darling test, Chapter 4 Methods). As the type of activating mutation can alter the mutant phenotype (198), the differences in missense selection may indicate that the selected phenotypes of *PIK3CA* mutations differ in the two tissues.

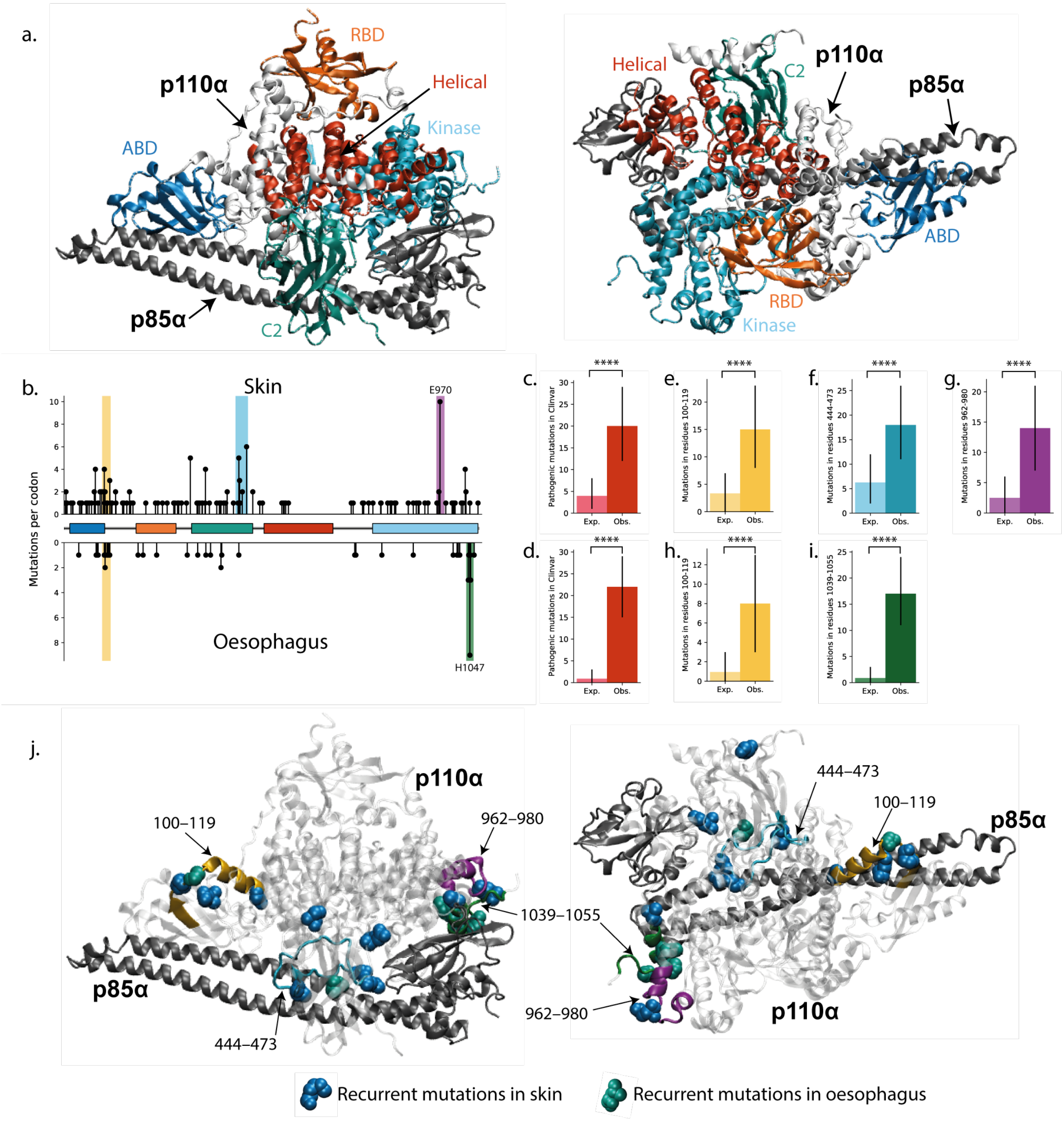

Figure 4.5 Selection of missense mutations in *PIK3CA*. a) The structure of PI3K showing the domains of p110α. Domain definitions from UniProt (163). **b)** Missense mutations in normal skin (top) and oesophagus (bottom). The most commonly mutated residues are labelled. Domain definitions from UniProt (163). ABD-domain, dark blue; RBD-domain, orange; C2 domain, green; helical domain, red; kinase domain, light blue. Shaded areas show the regions with significant enrichment of missense mutations as shown in **e-j**. Residues 100–119, yellow; residues 444–473, blue; residues 962–980, purple; residues 1039–1055, green. **c,d)** Missense mutations annotated as Likely Pathogenic/Pathogenic in Clinvar (197) in skin (**c**) and oesophagus (**d**). **e-i)** Missense mutations in regions of *PIK3CA* in skin (**e-g**) and oesophagus (**h-i**). Expected counts under the neutral null hypothesis, left bar; counts observed, right bar. **j)** Recurrent mutations shown on the structure of PI3K. p110α, white; p85α, grey. p110α residues 100–119 shown in yellow, 444–473 shown in cyan, 962–980 shown in purple and 1039–1055 shown in green. Residues mutated at least three times in the skin dataset shown as blue spheres. Residues mutated at least twice in the oesophagus dataset shown as green spheres. Error bars in **c-i** show 95% confidence intervals (Chapter 3 Methods). ****P≤0.0001, two-tailed binomial test, Chapter 3 Methods.

Table 4.1 Tests for enrichment of missense mutations in *PIK3CA* regions in normal skin and oesophagus. Regions based on (198). N=total missense mutations in test, Expected=percentage of missense mutations expected on the given residues under the neutral null hypothesis, Observed=percentage of missense mutations observed on the given residues, P=two-tailed binomial p-value, Chapter 3 Methods, Q=Benjamini–Hochberg multiple-test corrected p-value (192). Shaded rows show statistically significant results.

Skin					
Residues	**N**	**Expected**	**Observed**	**P**	**Q**
100–119	167	2%	9%	$2e^{-6}$	$9e^{-6}$
120–127	167	1%	1%	0.3	0.5
335–342	167	1%	0%	0.6	0.7
343–350	167	1%	2%	0.1	0.3
444–473	167	4%	11%	$6e^{-5}$	0.0002
532–551	167	2%	0%	0.06	0.2
720–744	167	2%	1%	0.2	0.4
848–859	167	1%	1%	0.7	0.7
930–956	167	2%	1%	0.5	0.6
962–980	167	1%	8%	$3e^{-7}$	$3e^{-6}$
1039–1055	167	1%	1%	0.7	0.7
Oesophagus					
Residues	**N**	**Expected**	**Observed**	**P**	**Q**
100–119	49	2%	16%	$4e^{-6}$	$2e^{-5}$
120–127	49	1%	0%	1	1
335–342	49	1%	0%	1	1
343–350	49	1%	2%	0.4	1
444–473	49	3%	2%	1	1
532–551	49	2%	0%	1	1
720–744	49	2%	0%	1	1
848–859	49	2%	0%	1	1
930–956	49	3%	0%	0.6	1
962–980	49	1%	0%	1	1
1039–1055	49	2%	35%	$7e^{-18}$	$8e^{-17}$

Table 4.2 Tests for enrichment of missense mutations in *PIK3CA* regions in normal skin and oesophagus, after deduplicating recurrent mutations and while excluding other selected regions in the tissue. Only significant regions from Table 4.1 tested. N=total missense mutations in test, Expected=percentage of missense mutations expected on the given residues under the neutral null hypothesis, Observed=percentage of missense mutations observed on the given residues, P=two-tailed binomial p-value, Chapter 3 Methods. Shaded rows show statistically significant results.

Skin				
Residues	**N**	**Expected**	**Observed**	**P**
100–119	114	2%	9%	0.0002
444–473	112	4%	7%	0.08
962–980	109	2%	5%	0.03
Oesophagus				
Residues	**N**	**Expected**	**Observed**	**P**
100–119	30	2%	23%	$1e^{-6}$
1039–1055	31	2%	26%	$6e^{-8}$

4.2.6 Signs of selection in *FBXW7*, but functional impact remains unclear

FBXW7 is one component of the Skp, Cullin, F-box containing (SCF) E3 ubiquitin ligase complex (200). FBXW7 recognises the substrate to be targeted for ubiquitination and subsequent degradation (200). It has around 90 target substrates, including TP53 and NOTCH1 (201).

There appears to be selection against loss-of-function mutations in normal skin, where nonsense and splice mutations are negatively selected (dN/dS=0.31, p=0.025, (11)) and there is a significant deficit of destabilising missense mutations (p=0.0001, n=62, two-tailed Monte Carlo test, Chapter 3 Methods, Figure 4.6a). However, the observed missense mutations are located significantly closer to the substrate binding site (202, 203) than expected under the neutral null hypothesis ($p<2e^{-5}$, n=62, two-tailed Monte Carlo test, Chapter 3 Methods, Figure 4.6b,c), suggesting that there is positive selection of a change to wild type *FBXW7* function. Re-testing the skin data while excluding mutations near the substrate binding site finds that there is no significant selection for destabilising mutations (excluding mutations within 8Å of the FBXW7 substrate; p=0.55, n=30, two-tailed Monte Carlo test, Chapter 3 Methods, Figure 4.6d). This suggests that the apparently strong negative selection of destabilising mutations seen when testing with the full set of missense mutations was, in fact, a result of positive selection of non-destabilising mutations around the binding site.

The peak of missense mutations in the oesophagus appears to mirror that of the skin data (Figure 4.6e), but unfortunately there are only a small number of *FBXW7* mutations in the oesophagus data set and the statistical tests are non-significant (FoldX ΔΔG: p=0.59, n=19, Figure 4.6f; distance to substrate binding site: p=0.096, n=19, Figure 4.6g; two-tailed Monte Carlo test, Chapter 3 Methods).

The most common *FBXW7* mutation hotspots in cancer – R465, R479 and R505 (200) – are also located at the substrate binding site (202) (Figure 4.6c). Mutations to these residues abrogate the ability of FBXW7 to bind to its substrates (200, 203). Inactivating *FBXW7* mutations are common in cancers driven by *NOTCH1* activating mutations, such as T-cell acute lymphoblastic leukaemia (T-ALL) and chronic lymphocytic leukaemia (CLL) (203). The mutant FBXW7 cannot bind to NOTCH1 NICD, leading to an accumulation of NICD similar to that caused by *NOTCH1* activating mutations (203). However, in the normal epithelia, there is strong selection for *NOTCH1* loss-of-function mutations (Chapter 3), so selection of mutations which *increase NOTCH1* activity would be surprising. In fact, despite their similar location in the protein structure, there is no overlap between the missense mutations that appear in the normal tissues and those that appear in T-ALL and CLL (COSMIC v91 (204), T-ALL n=89, CLL n=8, normal skin n=102, normal oesophagus n=20). This does not appear to be due to the mutational spectrum, as the most common cancer hotspot mutations (R465C, R465H, R479Q, R505C, combined total of zero occurrences in the normal skin) together would be expected to be mutated at least as often as the L559F mutation (11 occurrences) (Figure 4.6h). Therefore, in the normal skin, there does not appear to be selection for loss of FBXW7 binding to NOTCH1 NICD.

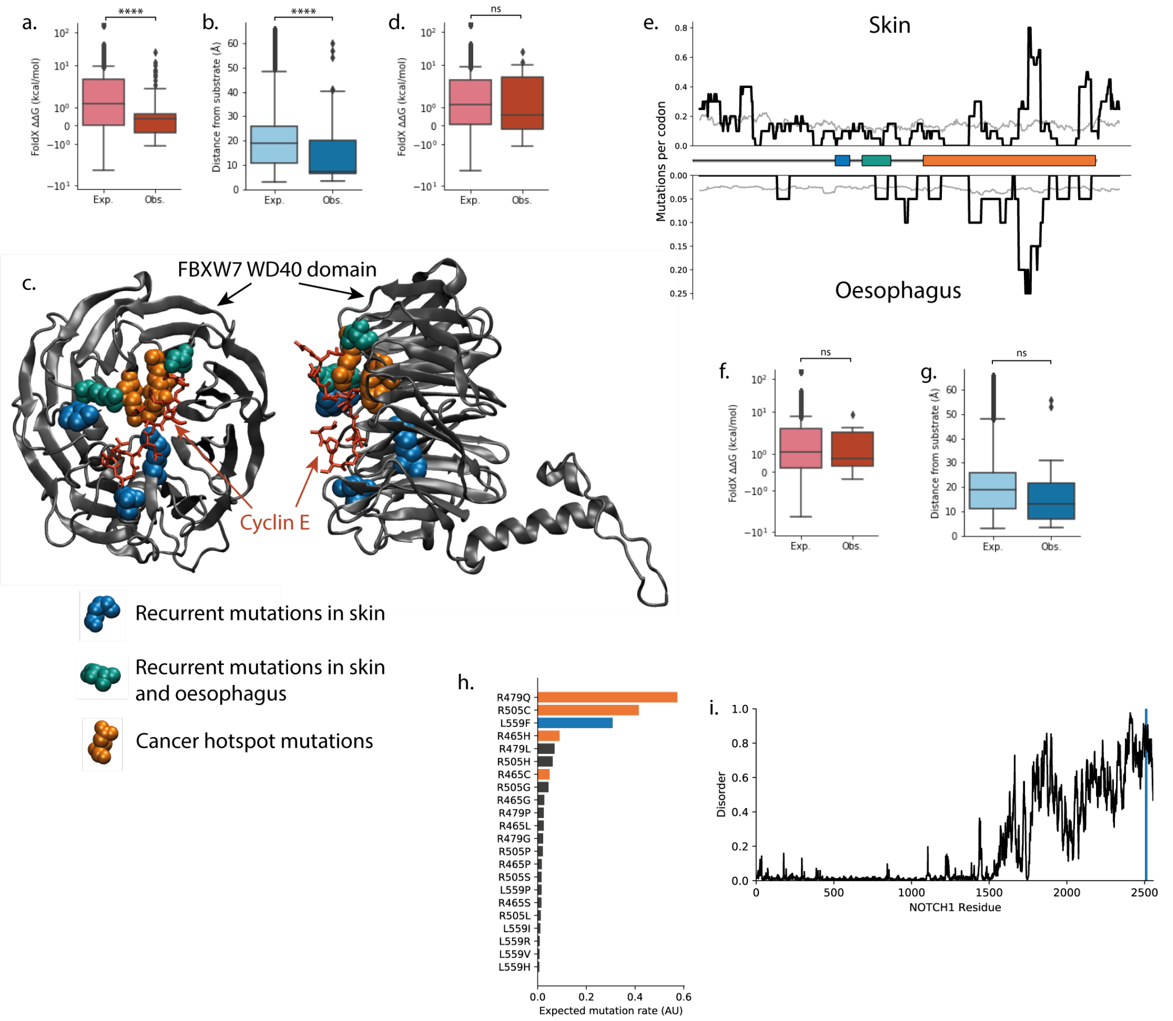

Figure 4.6 Missense mutations in *FBXW7*. a) Distribution of calculated ΔΔG values of missense mutations in FBXW7 in normal skin. Distribution expected under the neutral null hypothesis, light red, and the distribution observed, dark red. **b)** Distribution of distances of missense mutations in skin from the FBXW7 substrate (in this instance a 12-residue section of Cyclin E in PDB 2OVQ (202)). Distribution expected under the neutral null hypothesis, light blue, and the distribution observed, dark blue. **c)** Structure of the FBXW7 WD40 domain (PDB 2OVQ (202)) bound to Cyclin E (red). Commonly mutated residues are highlighted. Blue and green residues contain at least four missense mutations in the skin. Green residues also contain at least two missense mutations in oesophagus. The three main hotspot residues in cancer (R465, R479 and R505) shown in orange. **d)** Distribution of calculated ΔΔG values of missense mutations in skin, where all mutations within 8Å of the FBXW7 substrate have been excluded from both the null model and the observed data. Distribution expected under the neutral null hypothesis, light red, and the distribution observed, dark red. **e)** Sliding window of missense mutation frequency across *FBXW7* in skin (top) and oesophagus (bottom). Observed distribution shown with a bold black line, the expected distribution based on the mutational spectrum shown with a thin grey line. Dimerization domain, blue bar; F-box domain, green bar; WD40 domain, orange bar. **f)** Distribution of calculated ΔΔG values of missense mutations in oesophagus. Distribution expected under the neutral null hypothesis, light red, and the distribution observed, dark red. **g)** Distribution of distances of missense mutations in oesophagus from the FBXW7 substrate. Distribution expected under the neutral null hypothesis, light blue, and the distribution observed, dark blue. **h)** Expected relative mutation rates (arbitrary units) in skin of single nucleotide missense mutations on the residues R465, R479, R505 and L559. The four most common mutations in T-ALL and CLL are shown in orange, and the most common mutation in normal skin shown in blue. **i)** Disorder (calculated by IUPred2A (205)) of NOTCH1. FBXW7 targets a highly disordered section of NOTCH1, highlighted in blue. P-values in **a,b,d,f,g** calculated using the Monte Carlo test (Chapter 3 Methods). ****P≤0.0001, ns P>0.05.

It is tempting to speculate that, since *FBXW7* loss of function is a driver alongside *NOTCH1* gain-of-function mutations in T-ALL and CLL, selection of *NOTCH1* loss-of-function mutations in the normal epithelia would be accompanied by selection of gain-of-function *FBXW7* mutations. If *FBXW7* mutations lead to an increase in ubiquitination and degradation of *NOTCH1* NICD, then those mutant clones might have a growth advantage due to a reduction in *NOTCH1* signalling. However, due to the large number of *FBXW7* target substrates, several of which are driver genes in the normal epithelia, it is hard to narrow down to a single hypothesis. For example, an accumulation of c-Myc due to *Fbxw7* loss has been found to increase proliferation in keratinocytes (206). However, those cells also differentiate earlier due to accumulation of NOTCH1 NICD (206). If the mutations in normal skin and oesophagus selectively abrogate FBXW7–c-Myc binding without disrupting FBXW7–NOTCH1 binding it may lead to clonal expansion.

One approach to investigate these hypotheses further would be to model *in silico* (or measure experimentally) the interactions of *FBXW7* mutants with its various substrate proteins. It would then be possible to test for enrichment of mutations in the skin and oesophagus that increase or decrease interactions with particular substrates. However, the substrates can be disordered (Figure 4.6i), and predicting interactions of flexible or disordered proteins is a particularly hard problem (207). Methods may be inaccurate (207) and/or computationally expensive (208). This highlights a couple of limitations of the statistical method described in this chapter. It is limited by the quality of the *in-silico* methods, and the requirement to model every possible mutation in a region of a protein can require prohibitively large amounts of computation.

4.3 Discussion

In this chapter I have found biologically plausible and statistically significant patterns of selection in several proteins, which strengthens the evidence for non-neutral competition in normal epithelia. The statistical method can associate changes to protein structure or function with cell fitness, even in the absence of hotspot mutations and in the presence of passenger mutations. For example, in *NOTCH2* EGF11−12, no mutations occurred more than twice in the normal oesophagus dataset. However, by considering those mutations in bulk, I have identified three statistically significant features of selected missense mutations (the same features that are selected for in *NOTCH1* EGF11−12).

A strength and a limitation of the method is its use of external sources to provide scores for each mutation. Although the wide choice of potential metrics allows the analysis to be tailored to the gene and question at hand, it also makes the method less straightforward to use than driver discovery tools that have a single standard analysis. Since scores are required for all mutations in the region of interest, whether observed in the data or not, the metric calculations must be computationally cheap (or pre-computed scores must be available). The robustness of results will be limited by the accuracy of the metric calculations, but analysing mutations in bulk can help minimise the impact of incorrect scores for individual mutations. Furthermore, a quick-to-run but imperfect metric can be useful to detect broad trends of selection and to identify subsets of mutations that merit further investigation using more rigorous methods.

For proteins that are sufficiently ordered that their structures can be determined, selection of loss-of-function nonsense and splice mutations is frequently accompanied by selection of destabilising missense mutations. A standard workflow for analysing missense mutations in these genes might start by identifying if destabilising mutations are selected for before searching for other selected mutation features. Many different mutations in a gene may break protein function, but generally only a small proportion of potential mutations in genes are likely to be gain-of-function (67). This means that selection for gain-of-function mutations in a gene can lead to hotspots or small clusters of mutations (67), such as those seen in *PIK3CA* and *FBXW7*. However, as demonstrated by the pattern of mutations in *NOTCH1* (Chapter 3), clusters of mutations and hotspot mutations are also seen in genes where loss-of-function mutations are selected. Selection for gain-of-function mutations may be accompanied by selection against loss-of-function mutations, such as nonsense mutations or missense mutations that cause misfolding. This may dilute the signs of selection when analysing missense mutations across the gene as a whole, meaning the selection may be harder to detect using some driver detection methods.

The method presented in this and the previous chapter may not be appropriate for use as a high-throughput analysis tool. Significant departures from the null model may indicate the presence of non-neutral selection in a gene, but running the analysis without properly considering confounding sources of selection can lead to misleading inference of the mutation features under selection. Instead, the method is better suited to exploring individual genes in detail. If the genes of interest are not known prior to analysis of a sequencing dataset, one of the many driver detection tools available (or a combination of tools that use orthogonal methods (67, 209)) could be used to select a subset of genes for further investigation. Existing driver detection methods can provide some information regarding the selected changes to protein function (67), but it is an ongoing challenge to understand the precise biological function of driver mutations in genes (67). The method presented here could represent a small step in addressing that challenge.

I have also presented a statistical method for detecting differences in mutational patterns between datasets, accounting for the difference in mutational spectrum. Future refinements of the method might involve more clearly identifying the particular regions that differ (210) and generalising to metrics other than mutation location – for example comparing ΔΔG distributions between data sets. Comparison between data sets is not a trivial task, however, since many factors can influence the apparent selection. For example, the tissue sample size and sequencing depth used in an experiment will effect which clone sizes can be detected (Chapter 2), and the inferred strength of selection can depend on the size of the clones analysed (10).

Altogether, the analysis presented here and in the previous chapter demonstrates the rich vein of information available in large DNA sequencing data sets. By combining these data with functional interpretation of mutations we can infer the selection of functional changes in proteins, and hence learn about both the protein structure–function relationship and the role of the protein in the tissue sequenced. The method can be used as an *in-vivo* validation of results of *in-vitro* studies, and could be a useful method to explore selection of mutation features in existing data sets prior to conducting further experiments. This is an approach that will be widely applicable for genes or domains that are positively or negatively selected in somatic contexts, whether in cancer or normal tissue.

4.4 Chapter 4 Methods

The methods are largely as described in Chapter 3. Only methods specific to this chapter are described below.

4.4.1 Protein structures

The following structures were used for the individual protein analysis of FoldX ΔΔG and structural distance calculations:

NOTCH2 EGF11–12	5MWB (169)
TP53 DNA-binding domain	2AC0 chain A (211)
PIK3CA	4L1B (212)
FBXW7	2OVQ (202)

For the comparison of nonsense dN/dS values and FoldX ΔΔG, UniProt (163) was used to find all protein structures for genes sequenced in skin that had resolution ≤ 2.6Å, no mutations in the protein sequence, and containing at least 50 missense mutations in the skin data set. To reduce computational expense, only a single chain (the first alphabetically for the protein of interest) was run per protein structure. The structures run were:

TP53	1TSR (213), 1TUP (213), 1YCS (214), 2AC0 (211), 2ADY (211), 2AHI (211), 2ATA (211), 2OCJ (215), 2XWR (216), 2YBG (217), 3IGL (218), 3KMD (219), 3KZ8 (218), 4HJE (220), 4QO1 (221), 4XR8 (222), 5BUA (223), 5MCT (224), 5MCU (224), 5MCV (224), 5MCW (224), 5MF7 (224), 5MG7 (224), 6FJ5 (224)
NOTCH1	1YYH (225), 2F8Y (226), 2VJ3 (165), 3ETO (227), 3L95 (228), 4CUD (229), 4D0E (229), 5FMA (151), 5L0R (230)
GRIN2A	5H8F (231), 5H8H (231), 5H8N (231), 5H8Q (231), 5I2N (232), 5KCJ (231), 5KDT (232), 5TP9 (233), 5TPA (233)
PIK3CA	4L1B (212), 4L23 (212), 5DXT (234), 5UBR (235), 5XGI (236), 6GVF (237), 6PYS (238), 7K6M (239)
ERBB4	2R4B (240), 3BCE (241), 3U2P (242)
FBXW7	2OVQ (202), 2OVR (202), 5V4B (243)
ERBB2	1N8Z (244), 2A91 (245), 5MY6 (246)

EGFR	1MOX (247), 1YY9 (248), 4UV7 (249)
GRM3	3SM9 (250), 4XAR (251)
TP63	3QYN (252), 3US0 (253)
NOTCH3	4ZLP (254), 5CZX (255)
EPHA2	3FL7 (256), 7KJA (257)
BAI3	4DLO (258)
PTPRT	2OOQ (259)
NOTCH2	5MWB (169)

4.4.2 Homology models

Swiss Model (260) was used to create two models of NOTCH3 EGF10−11 based on the template structures 2VJ3 of NOTCH1 EGF11−12 (165) and 5MWB of NOTCH2 EGF11−12 (169).

4.4.3 Ligand-binding interface residues

The ligand-binding residues for NOTCH2 are based on conservation with the rat NOTCH1 ligand interface (169) (Table 4.3). Ligand-binding residues in NOTCH3 EGF10−11 were defined as the equivalent residues of those in NOTCH1 (shifted by 21 residues, Table 4.3).

Table 4.3 Ligand-binding interface residues of NOTCH2 EGF11−12 and NOTCH3 EGF10–11. Ligand-binding interface residues of NOTCH2 EGF11−12 from (169). NOTCH3 EGF10−11 residues are based on conservation with NOTCH1 EGF11−12.

	Interface residues
NOTCH2 EGF11−12	418, 421, 424, 425, 426, 428, 429, 439, 440, 452, 454, 456, 470, 472, 473, 481
NOTCH3 EGF10−11	392, 394, 397, 399, 400, 401, 402, 403, 404, 414, 415, 423, 426, 427, 429, 430, 431, 433, 445, 446, 447, 448, 449, 450, 454, 456, 457, 458, 459

4.4.4 Calcium-binding mutations

The calcium-binding residues in EGF11−12 of NOTCH2 were defined using MetalPDB (164) and the 5MWB (169) structure (Table 4.4). Based on conservation, the calcium-binding residues in NOTCH3 EGF10–11 were defined as the equivalent residues of those in NOTCH1 (shifted by 21 residues) (Table 4.4).

Table 4.4 Calcium-binding residues of NOTCH2 EGF11−12 and NOTCH3 EGF10–11. Based on MetalPDB (164) and the structure 5MWB (169) for NOTCH2. NOTCH3 EGF10–11 residues are based on conservation with NOTCH1 EGF11−12.

	Calcium-binding residues
NOTCH2 EGF11−12	415, 416, 418, 435, 436, 439, 456, 457, 459, 473, 474
NOTCH3 EGF10–11	391, 392, 394, 410, 411, 414, 431, 432, 434, 448, 449

4.4.5 Protein disorder

The IUPred2A server (https://iupred2a.elte.hu) (205) was used to calculate the “long disorder” of each residue in *NOTCH1*.

4.4.6 Weighted Anderson–Darling test

The distribution of mutations across a gene is dependent on the gene sequence and the mutational spectrum (Figure 4.7a,b). Therefore, there could be differences in the mutation distribution between data sets, even if selection is identical (Figure 4.7a,b). This difference can be corrected for by weighting the number of mutations on each residue by the inverse of the expected mutation rate (Figure 4.7c,d). There are several statistical tests available for comparing weighted distributions, with the Anderson–Darling test reported to be the most powerful (261). I have used the implementation of the weighted Anderson–Darling test in ROOT (261, 262) through the pyROOT interface (263).

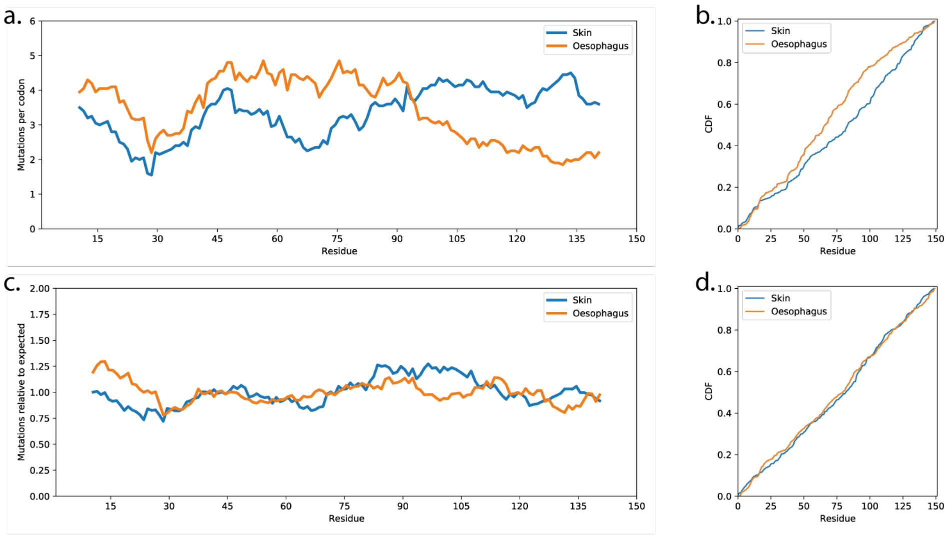

Figure 4.7 Example of comparing the mutation distributions in a neutral gene in two data sets with different mutational spectra. The artificial “gene” has a sequence of randomly drawn nucleotides with proportions that differ in different regions (first 100 bases: A 1%, C 1%, G 49%, T 49%; next 200 bases: equal proportions of all nucleotides; last 150 bases: A 49%, C 49%, G 1%, T 1%). 500 single base mutations were randomly drawn using the normal skin and oesophagus trinucleotide mutational spectra applied to the gene sequence. **a)** Sliding window of mutation density across the gene. **b)** Cumulative distribution of mutations across the gene. The distributions are significantly different, p=6.5e^{-6}, Anderson–Darling test. **c)** Sliding window of mutation density relative to the expected density calculated using the mutational spectra. **d)** Cumulative distribution of mutations across the gene after weighting the mutations by the inverse of the expected mutation rate for each residue. The weighted distributions are not significantly different, p=0.77, weighted Anderson–Darling test.

Chapter 5

Spread of super-competitive mutant clones can eliminate micro-tumours

5.1 Introduction

As discussed in previous chapters, normal oesophageal epithelium contains a large number of somatic mutations, with a mutational burden that increases with age (10, 12). The mutational landscape of the tissue is shaped by competition between expanding mutant clones, which can result in the removal of “loser” clones by highly fit “winner” clones (51). In rare cases, mutant cells in the oesophageal epithelium can develop into oesophageal squamous cell carcinoma (OSCC) (264). During the initial stages of growth, microscopic OSCC tumours will be bordered by mutant clones in the surrounding normal tissue. It is therefore of interest to investigate how mutations grow in the normal tissue, whether they interact with emerging tumours, and how that may impact the development of cancer.

The most widespread mutations in aged human oesophagus are *NOTCH1* mutants (10, 12), which cover 30 to 80% of the tissue by middle age (10). Mutations in *NOTCH1* are strongly enriched for loss-of-function mutations, including nonsense mutations, splice-site mutations, and missense mutations in key ligand-binding domains (10). The majority of large *NOTCH1* mutant clones have lost both wild type copies of the gene (10), which suggests strong selection for biallelic loss of *NOTCH1*. Mutations in the Notch pathway are also highly competitive in mouse oesophagus (50, 51). To investigate the competitive advantage conveyed by monoallelic and biallelic loss of *Notch1* in the oesophageal epithelium, lineage tracing experiments of heterozygous and homozygous *Notch1* null mutants were carried out in mice. I fit simple spatial simulations of clonal dynamics to the lineage tracing data to infer the “fitness” of these two *Notch1* genotypes. I then use the results to demonstrate how *Notch1* dose-dependent fitness effects the spread of *Notch1* mutants in the tissue.

Notch1 mutant clones’ domination of clonal competition in normal tissue raises the question of whether these highly fit clones have an effect on the early stages of cancer. Microscopic pre-malignant tumours are formed in the oesophagus when mice are given high doses of the mutagen diethylnitrosamine (DEN) in their drinking water (265). Some of these tumours grow to large sizes and acquire features of transformation such as dysplasia, inflammation and angiogenesis, and, 18 months after DEN treatment, can develop into squamous cell carcinoma (265). However, the majority of the micro-tumours are lost from the tissue (265). It was speculated that aggressive clones in the surrounding normal tissue may be responsible for the removal of many of the tumours, and a series of experiments were carried out to investigate this hypothesis: the density of tumours in the tissue was measured from 10 days to 18 months post-DEN treatment; additional highly fit clones were induced in the tissue; a *Notch1* inhibitor was administered; and DNA sequencing of the tumours was carried out at early and late timepoints. I constructed an abstracted mathematical model of tumour–clone interaction to analyse the results of these experiments and investigate whether micro-tumours are eliminated by competition with highly fit clones in the surrounding tissue.

5.2 Results

5.2.1 Cell fitness is strongly dependent on *Notch1* dosage

To investigate the growth of *Notch1* mutant clones, a series of lineage tracing experiments was carried out in mouse oesophagus by researchers in Prof Phil Jones’ laboratory team (Chapter 1 section 1.3.1) (266). The transgenic mice used contained an allele of *Notch1* which could be deleted by administering inducing drugs. The mice also contained a separate conditional allele of yellow fluorescent protein (YFP), which is used to track cell lineages and does not alter cell dynamics. Upon administration of a low dose of induction drugs, a small fraction of cells was genetically altered to express YFP, or the mutant *Notch1* allele, or, through random chance, to express both YFP and mutant *Notch1* (Figure 5.1). The genetic changes are inherited by any offspring cells (Figure 5.1). Three different strains of mice were used: a wild type control which did not contain the transgenic *Notch1* allele; a heterozygous mouse with one wild type and one conditional mutant allele of *Notch1*; and a homozygous mouse with two conditional mutant alleles of *Notch1* (both *Notch1* alleles are deleted together in induced cells).

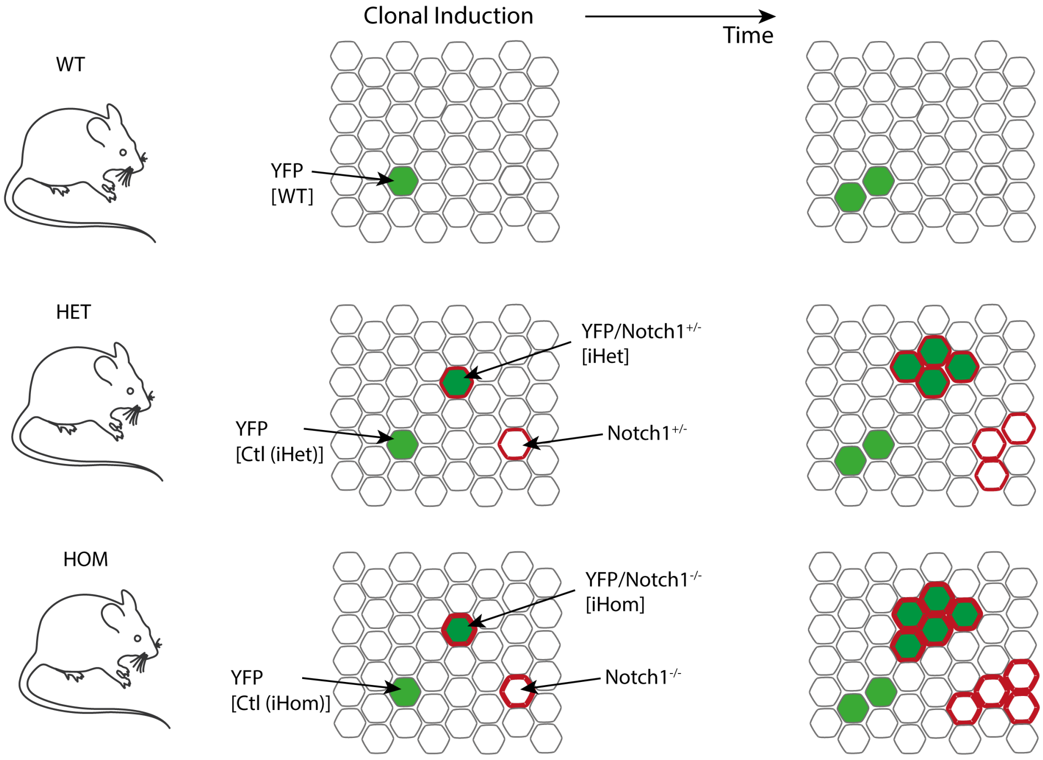

Figure 5.1 Lineage tracing system for *Notch1* mutant clones. Three strains of mice were used: wild type (top); heterozygous mice containing a single conditional *Notch1* mutant allele (middle); and homozygous mice where both copies of *Notch1* are the conditional mutant allele (bottom). All mice contained a conditional allele of yellow fluorescent protein (YFP, green). On administration of a low dose of induction drugs, most cells remained wild type (white), a small proportion of cells were genetically altered to express YFP, the mutant *Notch1* gene (red outlines) or both. The genetic changes are inherited by offspring cells. The YFP staining was used to identify clones (right). Staining for the NOTCH1 protein was able to distinguish *Notch1* mutant clones (iHet or iHom clones) from wild type clones (Ctl iHet or Ctl iHom) in the same mice. Mice images adapted from smart.servier.com, licensed under CC BY 3.0.

The number of basal cells in YFP-labelled wild type (WT) clones, heterozygous *Notch1* mutant (iHet) clones and homozygous *Notch1* mutant (iHom) clones were counted at several time points following induction (Figure 5.1). The observed iHet clones were significantly larger than WT clones, and iHom clones were larger again than the iHet clones (Figure 5.2). The iHom clones grew so quickly that for time points later than 30 days post-induction many of the clones had merged, meaning that the sizes of individual clones could not be reliably counted.

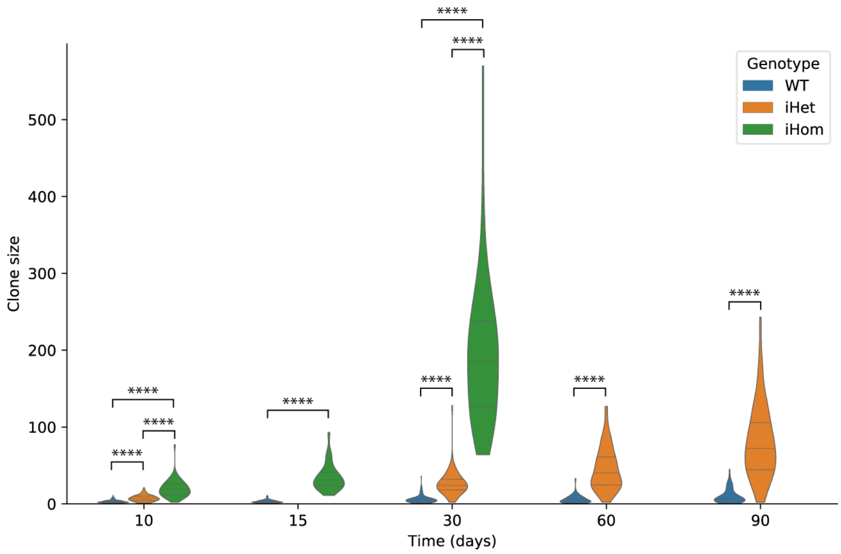

Figure 5.2 Lineage tracing data of *Notch1* mutant clones. Basal clone sizes of wild type (blue), heterozygous *Notch1* null (orange) and homozygous *Notch1* null (green) clones. **** $p < 0.0001$, two-tailed Mann–Whitney U test. Data from (266).

Previous studies have shown that clonal competition in normal oesophageal epithelium is consistent with a two-dimensional spatial model in which cell fate is determined by relative cell fitness in local cell neighbourhoods (51). If a cell is fitter than its neighbours, it is more likely to divide, and if it is less fit it is more likely to differentiate. This leads to clonal expansion driven by competition at the clone edges (51). In the middle of large clones, or at the boundaries of clones with similar fitness, the cell competition is approximately neutral (51). Here I have used two-dimensional Wright–Fisher style simulations to model clonal dynamics (Chapter 5 Methods). This model produces highly similar outputs to the Moran-style models in previous studies (Chapter 1 section 1.4.3.2), but is more computationally efficient. This efficiency enables the use of parameter-fitting methods that require the running of thousands of individual simulations.

I used approximate Bayesian computation (ABC) to fit the simulations to the clonal data (Chapter 5 Methods). There were three parameters to define in the model: the fitness of mutant cells, the induction level of the mutant cells, and the time between cell generations. Since *Notch1* mutant cells and wild type cells divide at a similar rate (266), I fixed the generation time to equal the average division time of wild type cells in the mouse oesophagus. Fitting of the wild type clone size data found that this division rate was appropriate for the growth rate of the neutral clones (Figure 5.3a,b, Table 5.1).

This leaves two parameters that are not fixed: the fitness of mutations and the induced proportion of cells at the start of the simulation. The induction proportion must be considered because as the clones expand they can collide and restrict each other's growth (51, 267). In a highly induced tissue, fit mutant clones will have less room to expand and will be smaller at later time points than in a sparsely induced tissue (Chapter 2). The other parameter, mutant fitness, is the key parameter I wish to compare between the genotypes. It represents how much of a growth advantage the mutant has compared to the surrounding non-mutant cells.

Firstly, the wild type clones (including the wild type YFP-labelled clones in the heterozygous and homozygous conditional *Notch1* mutant mice, Figure 5.1) have inferred fitness values closely centred on neutrality (fitness=1) (Figure 5.3a,b, Table 5.1). Both iHet and iHom clones are fitter than wild type clones, with iHom clones substantially fitter than iHet clones (Figure 5.3a,b, Table 5.1). Should this genotype–fitness relationship also apply in humans, this large increase in fitness conveyed by the second *Notch1* mutation may explain the strong selection and high prevalence of "double hit" *NOTCH1* mutant clones in aged human oesophagus (10, 266).

In the case of neutral clones, the induction proportion was not constrained by the fitting (Figure 5.3a, Table 5.2) because these clones have the same fitness as the surrounding wild type cells and therefore do not affect the growth of any other clones in the tissue. The inferred induction proportion for the iHet clones was larger than for the iHom clones (Figure 5.3a, Table 5.2). In simulations of the best fitting iHet parameters, the mutant clones were colliding at the later time points, reducing the clone sizes (Figure 5.3d). This high density of clones and extensive clonal collision was not occurring to such an extent in the experimental data (266), suggesting that the simple model is not fully capturing all details of the iHet clonal dynamics and may be overestimating the fitness of large iHet clones. However, the conclusions of the inferred fitness comparison are still valid: that the iHet clones have a clear growth advantage over wild type cells, and that the iHom clones have a much larger growth advantage than the iHet clones.

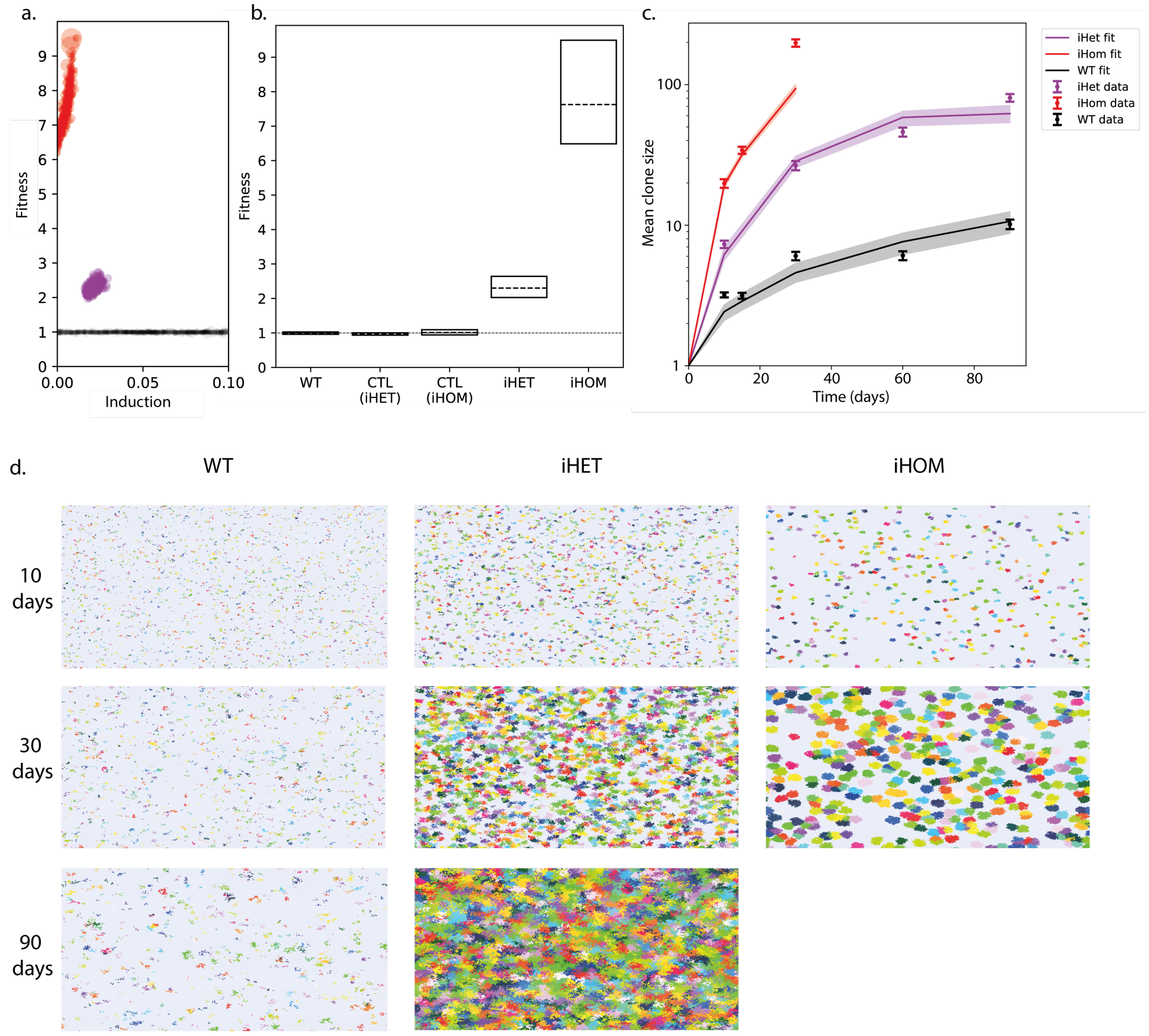

Figure 5.3 ***Notch1* mutant clone model-fitting results. a)** The inferred induction proportion and inferred fitness values from ABC fitting to lineage tracing data (Chapter 5 Methods) for iHom (red), iHet (purple) and WT (black) animals. Each dot shows an individual "accepted" parameter set. **b)** Distributions of acceptable values of the fitness parameter. Dashed lines and box show median inferred fitness and 95% credible intervals. A fitness of 1 is neutral. **c)** Mean clone sizes from simulations of the parameters at the peak of the acceptable distributions (Chapter 5 Methods). Median and 95% confidence intervals of 100 simulations shown for the simulation curves. Mean ± SEM are shown for the experimental data. **d)** Snapshots of simulations run with best fitting parameters for WT (left), iHet (middle) and iHom (right) clones. The later timepoints for the iHet show a denser crowding of mutant clones than was present in the experimental data.

Table 5.1 Inferred fitness parameter values from ABC fitting to lineage tracing data.

	Median inferred fitness	95% credible interval	
		Lower	Upper
WT	0.99	0.96	1.03
WT in Het mice	0.96	0.93	0.99
WT in Hom mice	1.03	0.96	1.12
iHet	2.3	2.0	2.6
iHom	7.0	6.2	8.6

Table 5.2 Inferred induction proportion parameter values from ABC fitting to lineage tracing data.

	Median inferred induction	95% credible interval	
		Lower	Upper
WT	0.054	0.002	0.099
WT in Het mice	0.051	0.004	0.098
WT in Hom mice	0.052	0.003	0.097
iHet	0.023	0.017	0.029
iHom	0.0045	0.0001	0.0085

5.2.2 Haploinsufficiency of *Notch1* greatly accelerates the spread of *Notch1*-null mutant cells

It is clear from the lineage tracing experiments and the model-fitting results that *Notch1* is haploinsufficient, i.e. the heterozygous mutation gives a growth advantage to the clone. This growth advantage of the heterozygous mutant increases the chance of a double *Notch1* mutant occurring because a higher proportion of the tissue becomes primed with the first mutation.

To illustrate the large difference haploinsufficiency makes to *Notch1* mutant tissue takeover, I ran simulations of haplo*insufficient* and haplo*sufficient* versions of *Notch1*. In the haplosufficient simulations, a cell with a single mutant *Notch1* allele has the same fitness as wild type cells, and a cell with both alleles of *Notch1* mutated increases the fitness to the best fitting iHom clone fitness (section 5.2.1). In the haploinsufficient simulations, the first *Notch1* mutation provides a fitness advantage equal to the best fitting iHet fitness, then a mutation to the other *Notch1* allele of the same cell increases the fitness to the best fitting iHom fitness.

When wild type mice were aged, a small number of spontaneous *Notch1* mutant clones appeared and spread in the tissue (266). Mutant clones in which both copies of *Notch1* were lost (*Notch1*$^{-/-}$ clones) could be identified by staining for NOTCH1 protein. To parallel this experiment, I ran simulations with a very low mutation rate. In the simulations, each cell has a 0.0005% chance of gaining a *Notch1* mutation in each generation, with the mutation rate picked so that the haplosufficient simulations approximate the experimental data (Figure 5.4a). Each mutation in the simulations can either appear on allele 1 or allele 2 of *Notch1* with 50% chance of each. A cell with a single mutant allele (Notch1$^{+/-}$) has the heterozygous fitness of that simulation type (see above), a cell with both alleles mutated has the homozygous fitness. A second mutation to the same allele does not change the cell fitness. I ran the simulations for 5000 days or until all cells in the tissue had both *Notch1* alleles mutated.

The simulations showed that it would take substantially longer for the tissue to be swept by the homozygous mutant clones if *Notch1* was haplosufficient (Figure 5.4). This shows how the haploinsufficiency of *Notch1* is a key property that allows mutations in this gene to dominate clonal competition in normal oesophageal epithelium.

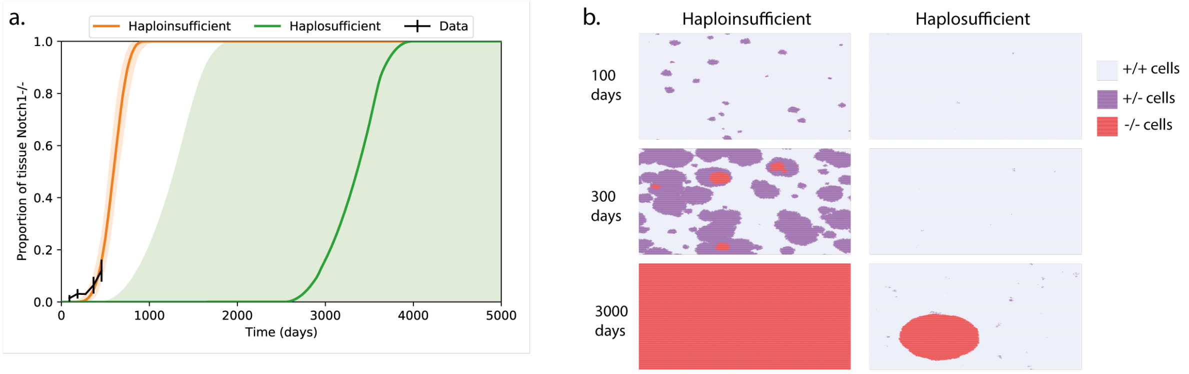

Figure 5.4 Simulations of haplosufficient and haploinsufficient mutant clone growth. a) Proportion of tissue covered by homozygous *Notch1* mutant (*Notch1*$^{-/-}$) clones over time in simulations using the best fit fitness parameter for iHom fitness for the *Notch1*$^{-/-}$ clones . Heterozygous *Notch1* mutants (*Notch1*$^{+/-}$) are either assumed to be neutral (haplosufficient, green) or to have the best fitting fitness parameter to the experimental analysis of iHet clones (haploinsufficient, orange). Curves show median of 100 simulations and shaded areas show 95% confidence intervals. **b)** Snapshot images at 100 days, 300 days and 3000 days of a representative simulation from a. On the left, *Notch1* is assumed to be haploinsufficient (*Notch1*$^{+/-}$ cell fitness inferred from fitting to lineage tracing data), on the right *Notch1* is assumed to be haplosufficient (*Notch1*$^{+/-}$ cells have the same fitness as wild type cells).

5.2.3 Super-competitive clones can remove pre-neoplastic micro-tumours

As shown above, *Notch1* mutant clones spread aggressively through the normal oesophageal epithelium. This raises the question of whether these clones in the normal tissue affect the early development of cancer. When mice are administered high doses of the mutagen DEN in their drinking water, small micro-tumours form in the oesophageal epithelium (265) (Figure 5.5a,b). Hundreds of these tiny tumours, which can be as small as ~20 cells, are present in each oesophagus 10 days after the end of DEN treatment (265). However, once mutagen treatment ends, the number of tumours falls dramatically (265) (Figure 5.5c). It was found that tumours were not lost due to apoptosis, abnormal proliferation of tumour cells or elimination by the immune system (265).

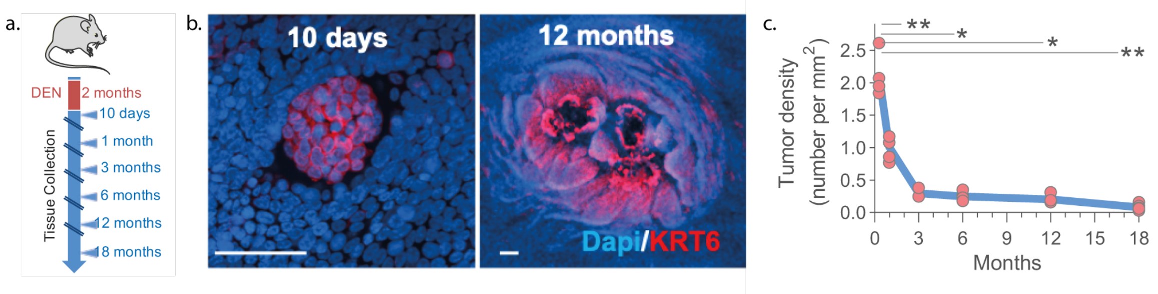

Figure 5.5 Diethylnitrosamine (DEN) treatment causes tumours to form in the oesophageal epithelium. a) Mice are treated with DEN and tissues collected at the indicated time points. **b)** Confocal images of tumours collected from oesophagus at the indicated time points after DEN treatment (scale-bars: 50µm). Tissues stained for Dapi (blue) and Keratin 6 (KRT6, red). **c)** Tumour density post-DEN treatment. Each dot is one mouse. P-values are from a two-sided Mann–Whitney test vs 10 days. * P<0.05, ** P<0.01. All figures taken from (265).

The tumours grow amongst a dense patchwork of highly fit clones competing for their place in the tissue (51). In the normal tissue, competition between mutant clones results in the loss of "loser" clones while the "winner" clones expand (19, 51). It was hypothesised that the tumours themselves could be involved in this competitive process, and that the substantial drop in tumour numbers could be due to "loser" tumours being outcompeted by the mutant clones surrounding them. Researchers in Prof Phil Jones' laboratory team ran a series of experiments to determine whether competition with clones in the surrounding normal tissue could be the cause of the high rate of tumour loss. I use a mathematical model as a framework to explore the results of these experiments and determine whether tumour survival is affected by mutant clones in the surrounding normal epithelium.

Here I describe a general model of tumour loss that allows the generation of testable predictions without requiring details of cell dynamics in tumours or mutant clones in the epithelium. I am interested in particular in whether the tumours are lost due to tumour-intrinsic mechanisms, or if the tumours are removed by mutant clones in the adjacent normal epithelium. As I discuss the factors affecting tumour loss, the principles of the model are described in mathematical terms so that I can later construct a numerical version of the model. This numerical version of the model is intended as a demonstration of the model principles, rather than a parameter- or model-fitting exercise.

In section 5.2.3.1, I describe a previously proposed stochastic model of tumour dynamics and apply it to the tumour density following DEN treatment. This model shows how tumour loss might occur due to the stochastic nature of the cell dynamics in oesophageal epithelium. However, it does not consider any interaction between tumour cells and the surrounding normal tissue. In section 5.2.3.2, I show that tumour loss is increased by the induction of highly fit mutant clones in the normal epithelium. This justifies incorporating into the model the effects that mutant clones surrounding the tumours could have in tumour loss. To further challenge this hypothesis, in section 5.2.3.3, I show how reducing the competitive imbalance between tumours and highly fit mutant clones in the normal tissue increases the number of surviving tumours. If competition with surrounding clones can remove early neoplasms, we might expect this competition to act as a selective pressure on tumours. This is the topic of section 5.2.3.4, where I discuss how if certain tumour genotypes reduce the competitive imbalance between the tumour and mutant clones in the surrounding tissue, those tumour genotypes will have a higher chance of survival and will become enriched in the tumour population. In section 5.2.3.5, I substitute simple mathematical equations into the model to numerically illustrate the principles described in the previous sections. I show how the various experimental results in this study are all consistent with this model in which mutant clones in the normal tissue contribute to the loss of tumours. Section 5.2.3.6 is a brief summary of conclusions from the mathematical analysis.

5.2.3.1. Stochastic model of tumour dynamics

I started by considering an existing model of cell dynamics in mouse oesophageal epithelial tumours to examine whether it is capable of explaining the pattern of tumour loss over time. This model was based on the shallow proliferative cell-differentiated cell hierarchy that operates in the normal oesophagus (21, 84).

In a previous study, oesophageal tumour growth in DEN- and Sorafenib-treated mice was found to be consistent with a stochastic model of cell dynamics (268). In this model, proliferating cells divide to form a pair of proliferating daughter cells, a pair of non-dividing daughter cells, or one cell of each type (Chapter 1 section 1.4.1). In the normal tissue, the probabilities of each symmetric division type are balanced, so that the total number of proliferating cells remains approximately constant. In the tumours, there was a bias towards producing more dividing than differentiated cells, causing the average size of tumours to increase over time (265, 268). Due to the stochastic nature of the process, some tumours expand in size, while in others all basal cells differentiate and the tumour is shed and lost from the tissue (27, 268). This variation in outcome due to random chance is known as drift.

The stochastic system here parallels the continuous time Markov process used to model the dynamics of mutant clones with imbalanced cell fates in normal epithelium (19, 88) (Chapter 1 section 1.4.1.2). I can therefore modify the equations used in these studies to calculate the probability that a tumour will lose all of its proliferating cells due to stochastic drift. The clones in normal epithelium are assumed to originate from single cells (19, 88), so the equations must be altered for this study to allow for the likely possibility that the tumours contain multiple proliferating cells at the end of the DEN treatment.

Asymmetric divisions do not alter the number of proliferating cells, which means we only need to consider the symmetric divisions. Letting P and D represent a proliferating and a differentiated cell respectively, the simplified model considering only symmetric divisions becomes (88)

$$P \xrightarrow{2r\lambda} \begin{cases} P + P & Prob.\ \dfrac{1}{2} + \Delta \\ D + D & Prob.\ \dfrac{1}{2} - \Delta \end{cases}$$

where λ is the division rate, $2r$ is the fraction of symmetric divisions, and Δ is the probability imbalance between symmetric division and symmetric differentiation. If Δ=0, equal proportions of proliferating and differentiating cells are produced, as occurs in neutral growth of wild type clones (21). The value of Δ can range from -0.5 (all symmetric divisions produce two differentiated cells) to 0.5 (all symmetric divisions produce two proliferating cells).

The probability that a cell lineage starting from a single cell will go extinct by time *t*, $\alpha(t)$, is given by (269)

$$\alpha(t) = \frac{\left(\frac{1}{2} - \Delta\right) e^{2dt\Delta} - \left(\frac{1}{2} - \Delta\right)}{\left(\frac{1}{2} + \Delta\right) e^{2dt\Delta} - \left(\frac{1}{2} - \Delta\right)}$$

where d is the symmetric division rate of tumour cells ($r\lambda$).

In this model, each cell is assumed to behave independently of all others, and therefore the extinction probability of a population starting with *n* proliferating cells equals the probability that *n* independent cell lineages starting from single cells all become extinct, i.e. $\alpha(t)^n$. If we assume that fully differentiated tumours are lost from the tissue, then the survival probability of a tumour is given by 1 minus the extinction probability:

$$p_{surv}(t) = 1 - \alpha(t)^n \qquad (1)$$

where *n* is the initial number of proliferating cells in each tumour (assumed here to be the same for all tumours). With *n*=1, Eq.1 matches the equation used in previous studies for the survival probability of clones that originate from single proliferating cells (19, 88). For a full derivation of the equations used here and in the cited studies of imbalanced clone dynamics, see ref (269).

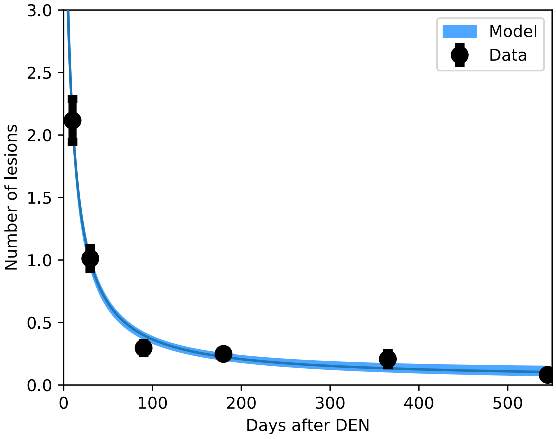

Figure 5.6 Tumour drift model (with no interaction with mutant clones in surrounding normal tissue). Model fit to tumour density (number of tumours per mm^2 of oesophageal epithelium) following DEN treatment (Chapter 5 Methods). The mean value and range between the minimum and maximum values of the model run with the accepted parameters from ABC fitting are shown.

Fitting Equation 1 to the tumour density following DEN treatment results in a steep initial drop in tumour numbers followed by a slower downward trend, consistent with the experimental data (Figure 5.6). The median accepted parameters found from fitting the model to the data (Chapter 5 Methods) were (95% credible interval lower bound, upper bound) d=0.33/day (0.27, 0.36), Δ=0.003 (0.001, 0.005), n=1.2 (1.0, 1.5), and initial tumour density (immediately following DEN treatment)=4.8/mm^2 (4.1, 5.6), though these parameters should not be interpreted as estimates of the true biological values (see sections below).

However, I will show in the sections below that this model is not capable of explaining the results seen in the full range of experiments, and therefore must be rejected (or at least adjusted) to account for the clones in the surrounding tissue.

5.2.3.2. Elimination of tumours by highly competitive mutant clones in the surrounding normal epithelium

Although the stochastic model of tumour drift defined above is sufficient to describe the pattern of tumour loss following DEN treatment, it does not rule out alternative causes of tumour loss. The normal epithelium surrounding the tumours contains a patchwork of competing mutant clones (51, 265). Therefore, we speculated that, like clones in the surrounding normal tissue, tumours may be displaced by highly fit mutant clones.

I considered a general model in which a highly competitive mutation, *M*, is induced in the tissue following DEN treatment. I assume that clones of the mutant *M* are able to remove tumours that they encounter. Let $p_M(t)$ be the survival probability of a tumour assuming that the mutant *M* is the only cause of tumour loss. Let p_{other} be the survival probability of a tumour based on all other sources of tumour loss (such as drift [section 5.2.3.1], and it may also include tumours removed by DEN-created clones [section 5.2.3.3 below]). For simplicity, I assume that tumour loss due to the mutant *M* is independent of all other causes of tumour loss. The combined survival probability of a tumour is then given by

$$p_{surv}(t) = p_{other}(t)p_M(t) \tag{2}$$

Let *M(t)* be the proportion of tissue covered by the mutant at time *t.* The following additional assumptions are made for the sake of simplicity:

1. The proportion of tissue covered by *M* increases monotonically.
2. Tumours are spread randomly across the tissue.
3. Tumours are removed instantly with probability 1 when the mutant *M* colonizes the location of the tumour (see the end of this section for a discussion of this assumption).

This means

$$p_M(t) = 1 - M(t) \tag{3}$$

and the combined probability of tumour survival is then given by

$$p_{surv}(t) = p_{other}(t)\big(1 - M(t)\big) \tag{4}$$

To make it easier to compare mice in which the initial density of tumours may vary, we looked at the proportion of tumours eliminated (*PTE*) by *M*, using the tumour density in the non-M-mutant regions of the tissue to estimate the tumour density in the full tissue in the absence of *M*.

$$PTE = 1 - \frac{total}{Mneg} \tag{5}$$

where *total* is the density of tumours over the full tissue, and *Mneg* is the density of tumours in the M-negative region.

The expected *PTE* for two models - where the *M* mutant removes tumours it encounters (PTE_M) and where tumour loss is independent of the mutant in the surrounding tissue ($PTE_{\neg M}$), are

$$PTE_M = 1 - \frac{p_{surv}(t)}{p_{other}(t)} = M(t) \tag{6}$$

$$PTE_{\neg M} = 1 - \frac{p_{other}(t)}{p_{other}(t)} = 0 \tag{7}$$

In other words, if *M* clones are able to remove tumours in a similar manner to which highly fit clones are able to displace weaker clones in the normal tissue, then we will see a reduction in tumour numbers proportional to the spread of the *M* mutant. By definition, models in which tumour survival is independent of the surrounding tissue ($PTE_{\neg M}$), such as the stochastic tumour drift model described in section 5.2.3.1, predict that tumour numbers will be unaffected by the spread of the *M* clones (Figure 5.7a).

To test these predictions, a colleague ran an experiment using an inducible DN-Maml1 mutation that prevents Notch signalling and forms rapidly expanding clones when induced in murine oesophageal epithelium (50, 51). When DN-Maml1 mutant clones are induced in the normal epithelium, the density of tumours is significantly reduced compared to uninduced control tissues (Figure 5.7b,c) (265). The density of tumours is not altered in regions of the induced oesophagus which are not occupied by DN-Maml1 mutant clones (265) (Figure 5.7c), and therefore the tumour density reduction is occurring only in the DN-Maml1 mutant areas. Furthermore, as predicted under the assumption that DN-Maml1 clones can remove tumours they encounter from the tissue, the data shows a strong correlation between DN-Maml1 clone spread and tumour loss (Figure 5.7a). The results of the experiment therefore indicate that DN-Maml1 clones are contributing to the loss of tumours from the tissue (Figure 5.7d).

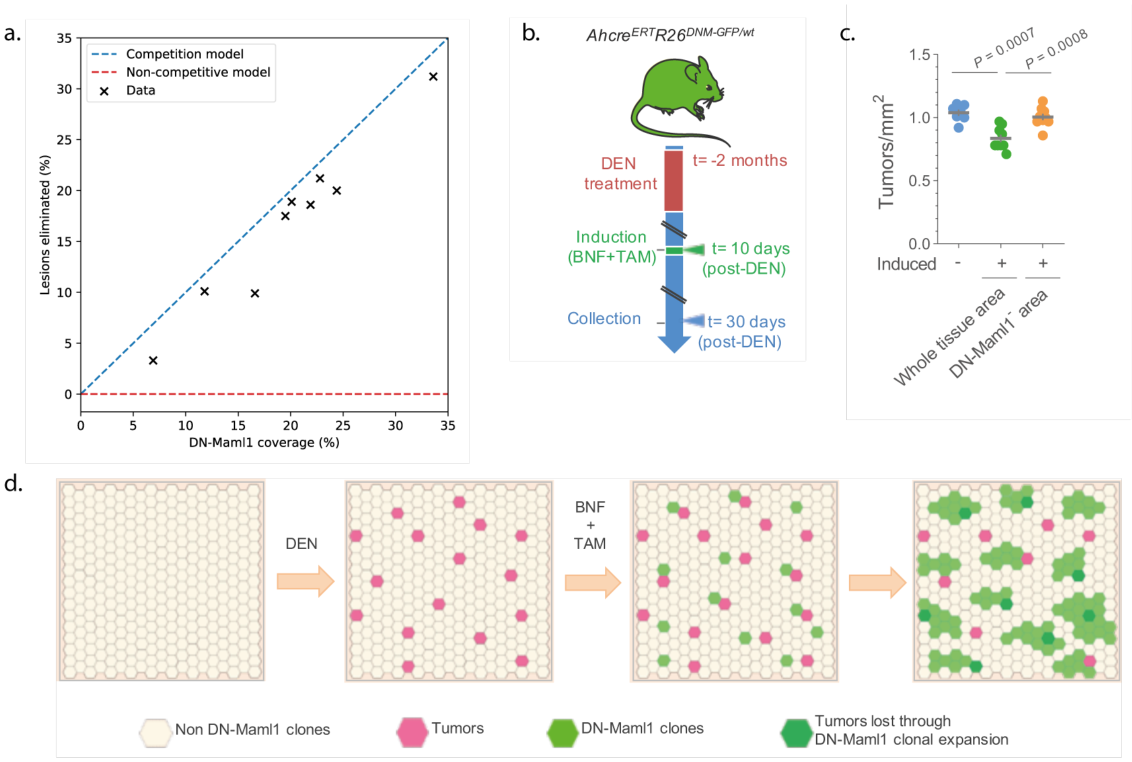

Figure 5.7 Tumour loss when highly fit mutant clones (such as DN-Maml1 clones) are induced in oesophageal epithelium following DEN treatment. a) Predictions of a model in which the induced mutant clones immediately remove tumours they encounter (blue) and a model in which tumour loss is independent of clones in the surrounding tissue (red). Experimental data from protocol in b shown as crosses. **b)** 10 days after the end of DEN treatment, highly fit DN-Maml1 clones were induced and tissues collected 20 days later. **c)** Number of tumours per mm^2 of oesophagus in control non-induced mice and induced DN-Maml1 mice. In the induced mice, the tumour density is not reduced in the area not taken over by DN-Maml1 clones (DN-Maml1$^-$ area). Error bars are mean ± s.e.m, p-values from two-tailed Mann–Whitney test. **d)** Cartoon illustrating the elimination of tumours by DN-Maml1 clones. The two months of DEN treatment generates tumours, then DN-Maml1 clones are induced by administering the drugs ß-napthoflavone (BNF) and tamoxifen (TAM). Expanding DN-Maml1 clones remove less fit tumours. **b,c,d** taken from (265).

In the experiment, there remained a small number of tumours in close contact with DN-Maml1 mutant regions (265). This may indicate that there is a lag time between contact with DN-Maml1 clones and tumour removal and that the surviving tumours seen in the DN-Maml1 areas are in the process of being removed. The clear significance of the experimental results and the small number of tumours surviving in the DN-Maml1 mutant regions suggest that lag time is small compared to the timescale of the experiment. It may also be the case that a small proportion of tumours are able to survive despite the competition with DN-Maml1 clones (see section 5.2.3.4 below).

5.2.3.3. Reducing competitive imbalance

Having shown that the induction of highly fit mutant clones following DEN treatment can eliminate tumours, the next question to address is whether mutant clones already present in the DEN-treated tissue are able to remove tumours too. Fit clones present in the normal epithelium might be able to outcompete the tumours, eliminating them from the tissue. By removing the competitive advantage of those clones, we can examine the impact they are having on tumour survival.

We assume there is a type of highly fit mutant clone, *N*, in the DEN-treated tissue that is able to remove tumours. As in the section above (Equation 2), we assume that the removal of clones by mutant clones *N* is independent of other causes of tumour loss. Tumour survival probability is given by

$$p_{surv}^{ctl}(t) = p_{other}(t)p_N^{ctl}(t) \tag{8}$$

where, similar to above, $p_{other}(t)$ is the survival probability of a tumour based on all sources of tumour loss other than elimination by *N* clones, and $p_N^{ctl}(t)$ is the survival probability of a tumour assuming that the mutant *N* is the only cause of tumour loss. I assume that, without intervention, *N* clones spread progressively throughout the tissue and outcompete tumours they encounter, so $p_N^{ctl}(t_2) < p_N^{ctl}(t_1)$ for $t_2 > t_1$.

If the fitness of the surrounding tissue and tumours can be raised to a similar level as *N* clones, this would both prevent the spread of the *N* clones across the tissue (reducing the number of tumours directly competing with *N* clones) and reduce the elimination of tumours that are already adjacent to *N* clones, as they will now be competing neutrally (Figure 5.8a). Assuming that the intervention completely levels the fitness of *N* clones with the rest of the tissue and tumours, the loss of tumours due to *N* clones will cease during that period, i.e. if the intervention starts at time t_1 and lasts until t_2, $p_N^{int}(t_2) = p_N^{int}(t_1) = p_N^{ctl}(t_1) > p_N^{ctl}(t_2)$ and

$$p_{surv}^{int}(t_2) = p_{other}(t_2)p_N^{int}(t_2) \; > \; p_{other}(t_2)p_N^{ctl}(t_2) = p_{surv}^{ctl}(t_2)$$

where, for the experiment in which the intervention is applied, $p_{surv}^{int}(t)$ is the overall tumour survival probability and $p_N^{int}(t)$ is the tumour survival probability related to the mutant *N*.

Therefore, if we can remove or reduce the competitive imbalance between a mutant that can remove tumours and the rest of the tissue (including the tumours), then we should see an increase in surviving tumours compared to control experiments (Figure 5.8a).

Notch1 mutant clones dominate competition in the normal oesophageal epithelium (51), even at early time points (265), and therefore is a good candidate mutation to test this prediction. The Notch inhibitor dibenzazepine (DBZ) prevents Notch signalling and would affect all cells in the tissue, effectively raising the fitness of all Notch wild type clones and tumours to the level of the Notch mutant clones (50) (Figure 5.8a). The competitive advantage of Notch mutants is thus removed during the DBZ intervention. Furthermore, by raising all cells to a high background level of fitness, this may also reduce the relative fitness advantage conveyed by mutations which work independently of the Notch pathway (270).

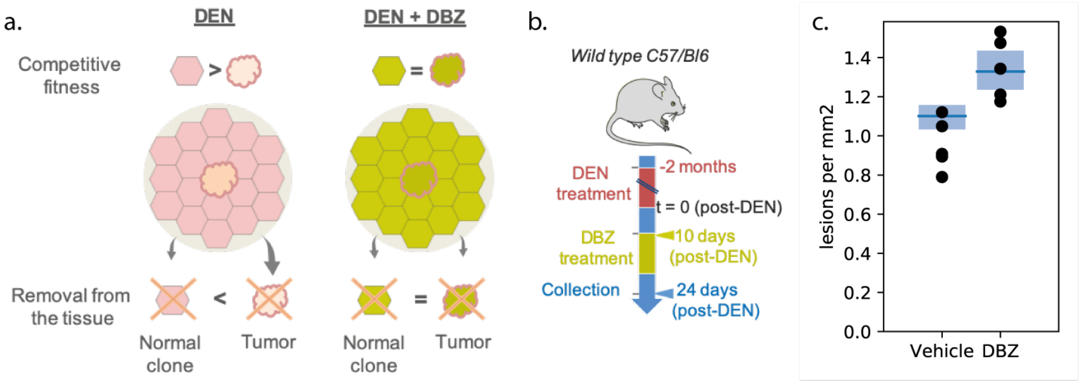

Figure 5.8 Reducing competitive imbalance reduces tumour loss. a) If the adjacent mutant clones have a higher fitness than the tumour, the tumour has a high probability of being outcompeted and removed from the tissue (left). Administration of the Notch inhibitor dibenzazepine (DBZ) increases the fitness of all cells to that of a *Notch1* null cell, levelling the competition between tumour and surrounding cells, and reducing the probability that a tumour will be removed (right). **b)** Protocol: wild type mice were treated with DEN for two months. Ten days after DEN withdrawal, DBZ or a vehicle control was administered for 14 days, after which the tissues were collected. **c)** Numerical example of the model (section 5.2.3.5) showing increased tumour survival when DBZ is administered compared to a vehicle control. Model fit to tumour density data using approximate Bayesian computation (section 5.2.3.5 and Chapter 5 Methods). The mean value and range between the minimum and maximum values of the model run with the accepted parameters are shown. Experimental data shown as black dots. **a,b** taken from (265).

The predicted increase in surviving tumours was seen when DBZ was administered between 10 days (t_1) and 24 days (t_2) after DEN treatment (Figure 5.8b,c). This suggests that competition from clones in the surrounding tissue is removing a substantial proportion of the tumours lost in the first few weeks following DEN treatment.

5.2.3.4. Selection pressure on tumours from competition with surrounding clones in the normal epithelium

There are genetic differences between the tumours sequenced 10 days and 1 year after DEN treatment. In particular, mutations in *Notch1* and *Atp2a2* are more prevalent in the 1 year tumours than the 10 day tumours (Figure 5.9a,b) (265). This genetic change over time could be consistent with ongoing selection of mutant subclones within tumours (64). However, the elimination of early neoplasms by mutant clones in the surrounding tissue could also act as a selective pressure on tumours.

As described in section 5.2.3.3 above, the competitive fitness of a tumour compared to the surrounding clones in the normal tissue will affect the survival prospects of the tumour. If certain tumour genotypes have a competitive fitness comparable to or higher than the fittest mutant clones in the surrounding tissue, they would be more likely to survive. For example, *Notch1* is the dominant mutant gene in normal tissue, occupying almost the entire tissue 12 months after DEN treatment (51). We might expect that, similar to the *Notch1* mutant clones in the normal tissue, tumours which are also *Notch1* mutant would be able to resist displacement by the mutant clones in the surrounding tissue. Consistent with this, there was an increase in the proportion of *Notch1* mutant tumours from the 10-day to the 1-year time point (265) (Figure 5.9b). There was also a large increase in the proportion of *Atp2a2* mutant tumours (265) (Figure 5.9b), suggesting that these too may be able to resist displacement by clones in the surrounding tissue.

To explore this hypothesis, I expanded the model to include two tumour phenotypes: sensitive and resistant. Sensitive tumours can be removed by clones in the surrounding tissue, while resistant tumours cannot. The mutation rate is expected to be low following the cessation of DEN treatment (few mutations spontaneously occur in untreated aged mice (51)), and therefore I assume that, post-DEN treatment, sensitive tumours do not evolve resistance through mutation. The probability of tumour survival is then given by

$$p_{surv}(t) = S(0)p_{other}(t)p_N(t) + R(0)p_{other}(t) \tag{9}$$

where $S(0)$ and $R(0)$ are the proportions of tumours at day 0 following DEN treatment which are sensitive and resistant respectively.

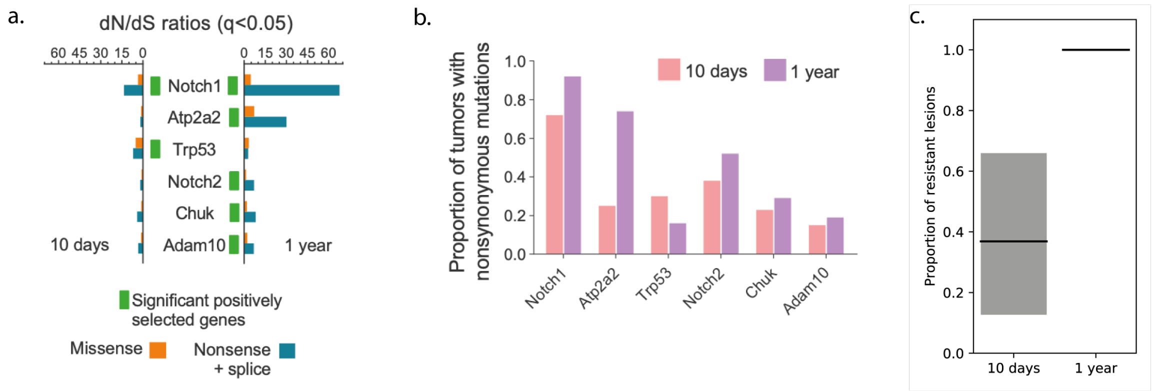

Figure 5.9 Selection of tumour genotype. a) dN/dS analysis of mutations found in tumours collected 10 days after DEN treatment and 1 year after DEN treatment. **b)** Proportion of tumours containing non-synonymous mutations in each indicated gene. **c)** Numerical example of the model from section 5.2.3.5 fit to experimental tumour density data using approximate Bayesian computation (Chapter 5 Methods). The inferred proportion of tumours resistant to displacement by mutant clones in the model increases from 10 days to 1 year after DEN treatment. The mean value and range between the minimum and maximum outputs of the model run with the accepted parameters are shown. **a,b** taken from (265).

The proportion of surviving tumours which are resistant to displacement by mutant clones, $R(t)$, is given by

$$R(t) = \frac{R(0)}{R(0) + S(0)p_N(t)} \tag{10}$$

Assuming that the initial numbers of sensitive and resistant tumours are non-zero, and that, as I assumed earlier, $p_N(t)$ is a decreasing function of time, then the proportion of surviving tumours which are resistant will increase over time (Figure 5.9c).

The increasing proportion of the surviving tumours which are *Notch1* and/or *Atp2a2* mutant is consistent with the hypothesis that these mutations may increase the tumour's ability to resist displacement by clones in the surrounding tissue. The presence of a subset of tumours which are able to resist displacement by clones also could explain why a small fraction of tumours are able to survive over 1 year after DEN treatment (Figure 5.10) when the surrounding normal tissue is almost entirely populated by highly fit mutant clones (51).

5.2.3.5. Numerical model example

So far, I have mostly defined general properties of the tumour survival probabilities rather than given specific equations for their values. This has allowed me to make testable predictions without requiring the details of the clonal or tumour dynamics to be defined.

Here I substitute feasible functions into Equation 9 to construct a numerical expression of the model. The functions are intended to be simple and introduce only a small number of model parameters, and the purpose here is to illustrate the concepts described in the previous sections rather than accurately and verifiably model the data.

Firstly, I need to define p_{other}, the non-clone related probability of tumour survival. Because apoptosis, abnormal proliferation of tumour cells and the immune system do not contribute to tumour loss (265), I assume that p_{other} is simply the survival probability based on drift (Equation 1). Secondly, I need to define $p_N(t)$, the probability of tumour survival based on highly fit mutant clones in the surrounding tissue. Following the assumptions in sections 5.2.3.3 and 5.2.3.4, I assume that there is a single mutant population *N* capable of removing tumours and that *N* mutant clones remove tumours as soon as they occupy the location of the tumour in the tissue. The model start time occurs after the end of DEN treatment, and so much of the tissue may already be occupied by the *N* clones. Therefore, I modify Equation 3 to account for the sensitive tumours existing in the remaining non-*N* proportion of the tissue:

$$p_N(t) = \frac{1 - N(t)}{1 - N(0)} \tag{11}$$

This still leaves the definition of $N(t)$, the growth pattern of the mutant clones. Growth of mutant clones in oesophageal epithelium have previously been modelled using branching processes (19), but these don't consider competition between clones and limitations of the tissue size (51, 267). Cellular automata simulations have also been used to model clones in this tissue (51) (sections 5.2.1, 5.2.2), but this does not allow for easy integration with the mathematical formulation. Instead, I used the logistic equation, which captures the key features of clonal spread: fast growth at early time points when mutant cells are mostly competing with surrounding wild type cells, slower growth at late time points when the tissue is already largely mutant, and an upper bound on total mutant spread (clones cannot grow beyond the capacity of the tissue) (51). Therefore,

$$N(t) = \frac{1}{1 + \left(\frac{1 - x_0}{x_0}\right) e^{-kt}} \tag{12}$$

where x_0 is the initial proportion of tissue covered by N mutant clones and k is the clone growth rate. To represent the experiment in which the DBZ Notch inhibitor is applied between 10 and 24 days after DEN, I can define

$$N_{DBZ}(t) = \begin{cases} N(t) & t < 10 \\ N(10) & 10 \leq t \leq 24 \end{cases} \tag{13}$$

The full model can then be constructed by substituting these expressions into equation 9.

There are 7 independent parameters in this model: d, Δ, n, x_0, k, $S(0)$, and the initial tumour density (Table 5.3). I used ABC to find the parameter combinations for which the model most closely matched the mean tumour density from 10 days to 18 months after DEN treatment and the mean tumour density in CTL and DBZ experiments at 24 days after DEN treatment (Chapter 5 Methods). The parameters were constrained as listed in Table 5.3. The results are shown in Figure 5.8c, Figure 5.9c and Figure 5.10. The median acceptable parameters found were (95% credible interval lower bound, upper bound) *d*=0.29/day (0.17, 0.35)*,* Δ=0.020 (0.007, 0.047), *n*=1.4 (1.0, 2.5), x_0=0.80 (0.42, 0.99), *k*=0.032 (0.019, 0.040), *S(0)*=0.64 (0.44, 0.82) and initial tumour density=4.2/mm^2 (3.4, 5.4), although, given the simplifications and approximations used in this numerical example of the model, they should not be interpreted as estimates of their biological counterparts.

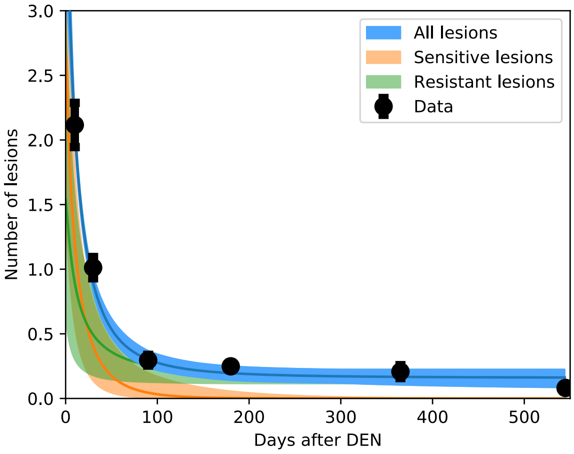

Figure 5.10 Numerical example of the mathematical model fit to experimental data. Tumour density (number of tumours per mm^2 of oesophageal epithelium) decreases following DEN treatment. Model fit to tumour density data using approximate Bayesian computation (Chapter 5 Methods). The mean value and range between the minimum and maximum outputs of the model run with the accepted parameters are shown. Experimental data shown in black, shown as mean and s.e.m.

The numerical example of the modelling principles demonstrates how the elimination of (a subset of) tumours by mutant clones in the surrounding normal tissue leads to the experimental observations of decreasing tumour numbers following DEN treatment, higher tumour survival when competitive imbalance is reduced (DBZ experiment), and a selection pressure on tumour genotype

5.2.3.6. Summary of micro-tumour removal by mutant clones

When DN-Maml1 clones are induced following tumour initiation, they can remove micro-tumours from the normal tissue. Following this result, I have shown that the results of a series of further experiments are all consistent with the hypothesis that mutant clones in the normal tissue remove tumours. Because the data are insufficient to fully characterize the mutant clone dynamics, tumour growth or the tumour–mutant interaction, the interpretation of the experiments is made using a highly abstract model. The numerical version of the model is provided to demonstrate how all the experimental data are consistent with each other and with the model, and to illustrate how the abstract model can be fleshed out with further details, even if those details are based on currently unverified assumptions.

Together, the experimental data and modelling indicates that tumours in the DEN-treated oesophageal epithelium are eliminated by mutant clones in the surrounding normal tissue Figure 5.11. The data also suggests that the tumour genotype influences the chance of a tumour surviving, possibly by allowing the tumours to better compete with mutant clones in the surrounding tissue Figure 5.11.

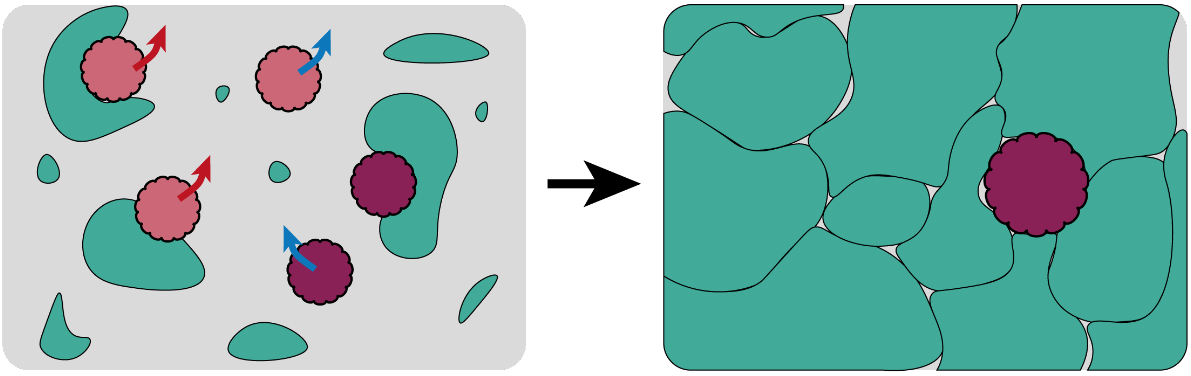

Figure 5.11 Summary of the tumour-loss model. At early time points following DEN treatment (left), many small tumours (pink and purple) exist in the tissue, alongside mutant clones in the surrounding normal tissue (green). Tumours can be removed from the tissue through competition with the mutant clones (red arrows), or through tumour-intrinsic mechanisms (blue

arrows). Certain tumour genotypes may be better able to resist competition with mutant clones (resistant tumours, purple; sensitive tumours, pink). At late time points (right), the tissue is almost entirely occupied by highly fit mutant clones. A small number of larger, resistant-genotype tumours remain in the tissue.

Table 5.3 Parameters for the numerical example of the model and initial bounds for the approximate Bayesian computation parameter fitting.

Parameter	Definition	Lower bound	Upper bound	Notes
d	Symmetric division rate of proliferating tumour cells.	0	2.5/week	Histone dilution revealed the tumours had a similar division rate to the normal tissue (approximately 2.5 times a week (51)). This was used as an upper bound on the rate of symmetric divisions.
Δ	Cell fate imbalance of tumour cells.	-0.5	0.5	
n	Number of proliferating cells per tumour at the end of the DEN treatment.	1	100	
x_0	Proportion of tissue covered by tumour-removing clones at the end of the DEN treatment (i.e. the start of the modelling period).	0	1	
k	Growth rate of tumour-removing clones.	0	0.04	Upper bound is the fit to DN-Maml1 mutant clone expansion in wild type background (data from (50)).
$S(0)$	Initial proportion of tumours that can be removed by clones in the surrounding tissue.	0	1	
Initial tumour density	Tumours per mm^2 at the end of DEN treatment.	0	20	At 10 days, there are approximately 2 tumours per mm^2 (Figure 5.5c). Allowed upper bound to be order of magnitude higher immediately following DEN treatment.

5.3 Discussion

Notch1 loss-of-function mutant clones in the oesophageal epithelium are highly competitive and can colonise vast areas without markedly altering the histology of the tissue (10, 12, 51, 266). In this chapter, the fitting of cell competition models to lineage tracing has shown that the loss of the second *Notch1* allele provides a large increase in fitness. If similar behaviour were to occur in humans, it could explain the high frequency of *NOTCH1* loss-of-heterozygosity events observed in human oesophagus (10). Furthermore, I have demonstrated that the haploinsufficiency of *Notch1* greatly accelerates the spread of *Notch1* mutant clones. Heterozygous missense mutations in *Trp53* and *Pik3ca* – mouse orthologues of positively selected genes in normal human oesophagus – also convey a growth advantage in mouse epithelium (19, 271). An avenue for further research would be to investigate whether dominant mutations or haploinsufficiency is required for a gene to be detected as a driver of clonal expansion in normal epithelium. The results in this chapter also imply that *Notch1* mutations may have a protective role against cancer. The results suggest that highly competitive clones, such as *Notch1* mutants, are capable of removing micro-tumours and hence may be contributing to the maintenance of histologically normal and healthy tissue. More generally, the results suggest that interventions affecting competition in the normal tissue may be able to influence the early stages of cancer.

The mathematical and computational models in this chapter are deliberately highly abstract and the conclusions are based on very broad patterns of behaviour. Additional complexity can be added to models to improve fitting of the finer details of the data. However, it can be hard to distinguish between different models of cell dynamics based only on clone size distributions from lineage tracing experiments (84). Therefore, more complex models may require further experimental data to validate their mechanisms. Future work might involve more precise determination of the rules of cell competition that provide *Notch1* mutations with a growth advantage (99). It would be particularly useful to characterise the competition between *Notch1* mutant clones in normal tissue and cells in micro-tumours that may no longer obey the rules governing the normal tissue. Illuminating the mechanisms by which "normal" cells in the tissue remove multicellular neoplasms might suggest interventions that encourage the maintenance of normal histology in highly mutated aging tissue.

5.4 Chapter 5 Methods

5.4.1 Two-dimensional Wright–Fisher simulations

I used simulations based on a Wright–Fisher model (91) constrained to a hexagonal 2D grid (11) (Chapter 1, Figure 1.11b). Each cell, $C_{a,b}$, in the grid represents a cell in the basal layer of the oesophageal epithelium. The subscripts denote the cell in location a in the grid and generation b of the simulation. I used a fixed size (500×500 cells) hexagonal grid with periodic boundary conditions.

Each cell has a fitness, $F_{a,b}$. In each step of the simulation, each cell $C_{a,b}$ in the new generation picks a parent cell from its immediate neighbourhood $\mathcal{N}_a$ (the six adjacent cells plus the cell in the same grid position) in the previous generation (Chapter 1, Figure 1.11b). The chance of a cell being picked as the parent depends on its fitness relative to the other potential parent cells in the neighbourhood,

$$P\big(C_{a,b}\, parent\ is\ C_{a',b-1}\big) = \begin{cases} \dfrac{F_{a',b-1}}{\sum_{j\in\mathcal{N}_a} F_{j,b-1}} & a' \in \mathcal{N}_a \\ 0 & a' \notin \mathcal{N}_a \end{cases}$$

A cell can produce multiple offspring in the next generation and the offspring cells inherit the fitness of their parent cells. In this way, a clone with high fitness can expand over multiple generations (Chapter 1, Figure 1.11b). At the start of a simulation, a small proportion of cells in the grid were randomly selected to be the induced mutant cells. In each simulation, a single fitness value was assigned to all the mutant cells. The rest of the cells in the simulation were given a fitness of 1. I tracked the sizes of the clones that grew from the mutant cells.

5.4.2 Approximate Bayesian computation

To fit the model parameters to the data I used approximate Bayesian computation (ABC) (272, 273). In this fitting method, prior distributions for the parameters are defined. These define the range of values that each parameter may take, and can include any "prior information" that may be known about the parameter values before fitting to the dataset. As I did not have any prior information to include, I used uniform distributions as the prior for all parameters.

ABC generates estimates of the parameter values as follows. A parameter set is randomly drawn from the prior distributions, and a simulation run with those parameters. The simulation results are measured against the observed data using summary statistics. If the "distance" between the simulation and the experimental data is less than a set threshold, those simulation parameters are accepted, otherwise they are rejected. Repeating this process many times with parameter sets randomly drawn from the prior distributions builds up a set of accepted parameters, which approximates the posterior parameter distribution (273).

I used a version of ABC based on sequential Monte Carlo sampling (ABC-SMC) (274), implemented in the Python package PyABC (275). In this method, multiple generations are run with increasingly stringent thresholds for acceptance, with each new generation of parameters based on the accepted parameter combinations from the previous generation. In this manner, the method efficiently identifies the regions of parameter space that produce model outputs most similar to the observed data.

5.4.2.1. Fitting of *Notch1* mutant clone dynamics

For the prior distributions for the two model parameters, I used uniform distributions across wide intervals: mutant fitness between 0 and 50, induction between 0% and 10% of cells. I used the Kolmogorov-Smirnov statistic (276) to compare the simulated and experimentally observed basal clone size distributions. I used 100 simulated clones at each time point, and summed the Kolmogorov-Smirnov statistics calculated for each time point to get the total distance between the simulation and the data. I ran 15 generations using a population of 1000 particles (the required number of accepted parameter sets per generation).

5.4.2.2. Fitting of micro-tumour loss

I ran the ABC-SMC fitting for 25 generations with a population of 10000 (a larger population size is required for this model than for the fitting of *Notch1* clonal dynamics, in order to adequately sample the higher dimensional parameter space here). For the summary statistic, I used the total squared distance of the mean model prediction from the mean observed tumour density. For the tumour drift model with no interaction with mutant clones in surrounding normal tissue, the fit was to the tumour density following DEN treatment. For the model including clone interaction, the fit was to the tumour density following DEN treatment and the tumour density in the DBZ experiment. Uniform distributions using the bounds in Table 5.3 were used as the initial prior distributions of the parameters.

Chapter 6

Discussion

6.1 Strong and consistent evidence for non-neutral selection in normal skin and oesophageal epithelium

One of the aims of this thesis was to resolve a debate on the nature of mutant clonal dynamics in normal human epidermis (9, 13-15, 111). A dN/dS analysis of the mutations detected through DNA sequencing of eyelid skin found six genes under significant positive selection (9). However, an analysis of clone sizes inferred from the same dataset found that the clone sizes were consistent with a model of neutral cell dynamics, suggesting that mutations do not alter the growth or survival of the cells that carry them (13).

By using a computational model of cell dynamics that takes account of the sheet-like structure of the epithelial tissue and the experimental tissue-sampling method, I have shown that the clone sizes inferred from the eyelid skin were consistent with non-neutral clonal competition. This work therefore resolves the apparent paradox presented by the discrepancy between the dN/dS analysis and the mutant clone sizes (15). The predictions of the computational model were validated using a DNA-sequencing dataset from normal human oesophageal epithelium (10). Like the eyelid skin, the oesophageal dataset exhibited strong signs of positive selection when analysed using dN/dS. In addition, the oesophageal dataset illustrated how the choice of experimental tissue-sampling method can reduce signs of non-neutrality in the mutant clone size distribution.

I found further evidence of non-neutral selection in normal skin and oesophageal epithelium by examining the patterns of missense mutations in protein structures of driver genes. The mutations in oesophageal epithelium and a larger DNA-sequencing dataset of normal human skin (11) exhibited highly significant enrichment of mutations with biologically meaningful impacts on protein function. The results were consistent with the dN/dS analysis of these datasets. Together with previous studies that found evidence of non-neutral expansion of mutant clones in human and mouse epidermis and oesophagus (9, 10, 12, 19, 50, 88, 271), the work in this thesis shows that the many strands of available evidence paint a consistent picture of widespread non-neutral mutant clonal competition and selection in the normal epidermis and oesophageal epithelium.

6.2 Spatial clonal competition has implications for cancer development

In chapters 2 and 5 of this thesis, I used computational models of epithelial cell dynamics based on spatial competition between mutant clones, i.e. for a clone to expand it must displace other cells from the tissue. The use of these spatial competition models is justified by experimental results in mouse skin and oesophagus. In normal mouse skin, the expansion of *Trp53* mutant clones is reversed when they encounter more aggressive mutant clones (19). And, more recently, a study in normal mouse oesophageal epithelium provided clear evidence of spatial competition between mutant clones (51).

As shown in Chapter 5, one remarkable and exciting consequence of clonal competition is the elimination of microscopic tumours by mutant clones in the surrounding normal tissue. Previous studies had shown that wild type cells can eliminate neighbouring mutant cells to preserve the health of the tissue and defend against cancer formation (277-279). However, in this study we have shown that *mutant* cells can also play a role in protecting tissue homeostasis. This study also demonstrated that the elimination of tumours can be altered by manipulating clonal competition (in this case, administering the Notch inhibitor dibenzazepine increased tumour survival). This shows the feasibility of intervening in precancerous clonal competition to reduce cancer incidence (10). These interventions would be designed to alter the environment in which clonal competition occurs, so that the growth of benign, or even beneficial tumour-eliminating clones, would be encouraged at the expense of more tumorigenic mutations.

Notch pathway mutations are highly competitive in the normal tissue because the cells that carry them proliferate at the expense of their neighbours (50, 266). Although these mutations tip the balance of cell fate in their favour, they do not altogether break the rules that maintain tissue homeostasis. In tumours, however, the rules that govern the normal tissue environment, and that *Notch1* mutants thrive under, are broken. Indeed, the mutations that drive tumour growth, or are highly competitive in the tumour environment, may not be the same mutations that are highly competitive in the normal tissue environment (280). For example, several genes are mutated more frequently in normal oesophageal epithelium than in OSCC tumours (10, 12). Some mutations may not contribute to tumourigenesis, but appear in tumours simply because they were present in the normal tissue prior to tumour formation. Signs of positive selection in the normal tissue, such as a high dN/dS ratio, may also be inherited by tumours and could erroneously lead to the classification of a non-tumourigenic gene as a cancer driver. This may be the case for *NOTCH1*, which has been described as a driver of OSCC (10, 281), but is mutated at a higher frequency in the normal tissue than in cancer (10, 12) and complete loss of *Notch1* even appears to reduce the growth of oesophageal tumours (266). We must therefore be careful how we interpret "driver mutations" in cancer. To learn which mutations are true cancer drivers, the genetic landscape of the normal tissue should be examined too.

6.3 Using computational methods to investigate clonal competition and selection in normal epithelia

Theoretical models can be used to turn a qualitative description of a biological system into quantitative predictions that can be experimentally tested. Theoretical models construct a simplified and abstracted version of the biological system they represent; selecting the appropriate degree of simplification is a crucial step in the modelling process (105, 106). More complex models can be harder to solve mathematically or can be more computationally expensive, and they can also be harder to understand and interpret. Parameter sweeps can be used to explore the full range of model outputs, but these sweeps also become less practical to run and visualise as the complexity of the model increases. The usefulness of adding complexity to models is also limited by the data available to verify the model assumptions and test the model predictions.

The modelling of clonal dynamics is further complicated by the effects of mutations. Different advantageous mutations may increase cell fitness through different mechanisms, leading to different patterns of clonal expansion (99). Even mutations within the same gene can have a wide range of phenotypes (129, 198). Complex interactions may occur between mutations in the same cell (98, 282, 283) or in neighbouring cells (51, 284), and the mutant clone phenotype may even depend on the order in which the mutations appear (285, 286). The behaviour of the mutant clone can also change over time, reacting to a changing local environment that may be altered by the mutant itself (19, 287). Exploring the consequences of adaptive mutant behaviour, while still using models which are simple enough to fit to data and interpret, will be an ongoing challenge in this field of research.

New experimental innovations, such as live imaging of cells *in vivo* (54, 288) or improved *in-vitro* culture systems, may make the use of more complex models of epithelial cell dynamics more feasible. Improved experimental methods would allow the measurement of more aspects of the biological system and could enable more of the cell dynamics to be directly observed rather than inferred through model-fitting. The use of new technology could therefore ensure that more model assumptions and parameters are founded on experimental observations, and would allow more model predictions to be tested.

To reduce model complexity, the models of clonal competition used in this thesis are highly abstracted. For example, the model of tumour–mutant clone competition in Chapter 5 makes very few assumptions about the mechanics of tumour growth, mutant clone dynamics, or tumour–mutant clone interaction. The abstract nature of the model means it could be used to interpret experimental data from a highly complex biological system even though most of the detailed complexity of the system was unknown and unmeasured. One drawback of highly simplified models is that they may only produce a rough approximation of the experimental data, as illustrated in this thesis by the imperfect fit of the 2D Wright–Fisher model to the expansion of heterozygous *Notch1* mutant clones. Because the model has only two adjustable parameters, there is a limit to how much it can be tweaked to precisely match the data. As long as model results are not over-interpreted, slight model inaccuracy is not necessarily a problem.

However, if model abstraction and simplification discards crucial features of the biological system, the key outputs of the model may no longer match the biology it is intended to represent. This is exemplified by the analysis of mutant clone sizes in normal human eyelid skin presented in Chapter 2: previous analysis using a very simple model of cell dynamics concluded that mutations did not alter cell behaviour, but by increasing the complexity of the model to include details of the tissue structure and experimental method, the interpretation of the data is fundamentally changed.

Some model outputs can also be easily misinterpreted. For example, in the statistical method presented in chapters 3 and 4, enrichment of the same mutational feature can be non-significant or highly significant depending on how the data are sliced. Interpreting the result requires an awareness of other selected features that may exist in the tested region. Nonetheless, the method could prove useful if paired with other methods to validate the results.

Each experimental approach, and each method of analysis, will have certain flaws and limitations. However, each new method that provides an alternative view of the biological system can be a useful addition to the scientist's arsenal. Considering a broad range of theoretical and experimental approaches can be highly beneficial. Agreement between multiple methods can provide more confidence than a single method, while differing conclusions between methods can raise important questions for further investigation.

6.4 Future Directions

Several directions for future work could follow from this thesis. As mentioned in section 6.3 above, developments in experimental methods should enable testing and verification of more complex theoretical models of cell dynamics. The combination of experiments and modelling will enable detailed exploration of the spread of and competition between the many different mutations that frequently appear in normal tissues, and build up a clearer picture of the dynamic clonal landscape in pre-cancerous tissue. Illuminating the mechanisms that govern clonal competition in normal tissues could lead to interventions that can steer somatic evolution in a benign direction. There also remains much to discover about the earliest stages of tumour growth: how tumours emerge from normal tissue, how they grow and how they do or do not survive in a highly mutated tissue environment. Here, as shown in Chapter 6, theoretical models can be used to help design and interpret experiments.

There is plenty of scope for the statistical method presented in Chapters 3 and 4 to be improved. Some metric scores may be accompanied by an uncertainty or a reliability score. These could be incorporated in the model distributions by weighting the scores by their reliability. Mutation rates can vary based on epigenetic factors and location in the gene (61, 289); therefore, as better predictions of per base mutation rate become available, further improvements could be made to the null models. As shown in Chapter 4, the statistical method can be adapted to compare the selection on a gene between tissues, or between normal tissue and cancer. However, this is a particularly difficult challenge, because technical differences between the data are hard to distinguish from selection differences. Consequently, further work is required to test the robustness of this method. There are vast and increasing amounts of data that the technique in Chapters 3 and 4 could be applied to, both in normal tissues and cancer, and it could easily be adapted to interpret mutational scanning data.

References

1. Pellettieri J, Alvarado AS. Cell Turnover and Adult Tissue Homeostasis: From Humans to Planarians. Annual Review of Genetics. 2007;41(1):83-105.

2. Doupé DP, Jones PH. Cycling progenitors maintain epithelia while diverse cell types contribute to repair. BioEssays. 2013;35(5):443-51.

3. Armitage P, Doll R. The age distribution of cancer and a multi-stage theory of carcinogenesis. British journal of cancer. 1954;8(1):1.

4. Nowell P. The clonal evolution of tumor cell populations. Science. 1976;194(4260):23-8.

5. Knudson AG. Mutation and cancer: statistical study of retinoblastoma. Proceedings of the National Academy of Sciences. 1971;68(4):820-3.

6. Renan MJ. How many mutations are required for tumorigenesis? Implications from human cancer data. Mol Carcinog. 1993;7(3):139-46.

7. Frank SA. Dynamics of cancer: incidence, inheritance, and evolution: Princeton University Press; 2007.

8. Alexandrov LB, Nik-Zainal S, Wedge DC, Aparicio SAJR, Behjati S, Biankin AV, et al. Signatures of mutational processes in human cancer. Nature. 2013;500(7463):415-21.

9. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S, et al. High burden and pervasive positive selection of somatic mutations in normal human skin. Science. 2015;348(6237):880-6.

10. Martincorena I, Fowler JC, Wabik A, Lawson ARJ, Abascal F, Hall MWJ, et al. Somatic mutant clones colonize the human esophagus with age. Science. 2018.

11. Fowler JC, King C, Bryant C, Hall M, Sood R, Ong SH, et al. Selection of oncogenic mutant clones in normal human skin varies with body site. Cancer Discovery. 2020.

12. Yokoyama A, Kakiuchi N, Yoshizato T, Nannya Y, Suzuki H, Takeuchi Y, et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. Nature. 2019;565(7739):312-7.

13. Simons BD. Deep sequencing as a probe of normal stem cell fate and preneoplasia in human epidermis. Proceedings of the National Academy of Sciences. 2016;113(1):128-33.

14. Martincorena I, Jones PH, Campbell PJ. Constrained positive selection on cancer mutations in normal skin. Proceedings of the National Academy of Sciences. 2016;113(9):E1128-E9.

15. Simons BD. 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. 2016;113(9):E1130-E1.

16. Braakhuis BJM, Tabor MP, Kummer JA, Leemans CR, Brakenhoff RH. A Genetic Explanation of Slaughter's Concept of Field Cancerization. Cancer Research. 2003;63(8):1727.

17. Stratton MR, Campbell PJ, Futreal PA. The cancer genome. Nature. 2009;458(7239):719-24.

18. Martens EA, Kostadinov R, Maley CC, Hallatschek O. Spatial structure increases the waiting time for cancer. New J Phys. 2011;13:115014.

19. Murai K, Skrupskelyte G, Piedrafita G, Hall M, Kostiou V, Ong SH, et al. Epidermal Tissue Adapts to Restrain Progenitors Carrying Clonal p53 Mutations. Cell Stem Cell. 2018;23(5):687-99.e8.

20. Martincorena I. Somatic mutation and clonal expansions in human tissues. Genome Medicine. 2019;11(1):35.

21. Doupé DP, Alcolea MP, Roshan A, Zhang G, Klein AM, Simons BD, et al. A single progenitor population switches behavior to maintain and repair esophageal epithelium. Science (New York, NY). 2012;337(6098):1091-3.

22. Alcolea MP. Oesophageal Stem Cells and Cancer. In: Birbrair A, editor. Stem Cell Microenvironments and Beyond. Cham: Springer International Publishing; 2017. p. 187-206.

23. Seery JP. Stem cells of the oesophageal epithelium. Journal of cell science. 2002;115(9):1783-9.

24. Moreci RS, Lechler T. Epidermal structure and differentiation. Current Biology. 2020;30(4):R144-R9.

25. Page ME, Lombard P, Ng F, Göttgens B, Jensen KB. The epidermis comprises autonomous compartments maintained by distinct stem cell populations. Cell stem cell. 2013;13(4):471-82.

26. Ito M, Liu Y, Yang Z, Nguyen J, Liang F, Morris RJ, et al. Stem cells in the hair follicle bulge contribute to wound repair but not to homeostasis of the epidermis. Nature medicine. 2005;11(12):1351-4.

27. Clayton E, Doupé DP, Klein AM, Winton DJ, Simons BD, Jones PH. A single type of progenitor cell maintains normal epidermis. Nature. 2007;446(7132):185-9.

28. Doupé DP, Klein AM, Simons BD, Jones PH. The Ordered Architecture of Murine Ear Epidermis Is Maintained by Progenitor Cells with Random Fate. Developmental Cell. 2010;18(2):317-23.

29. Rosekrans SL, Baan B, Muncan V, van den Brink GR. Esophageal development and epithelial homeostasis. American Journal of Physiology-Gastrointestinal and Liver Physiology. 2015;309(4):G216-G28.

30. Rustgi AK, El-Serag HB. Esophageal Carcinoma. New England Journal of Medicine. 2014;371(26):2499-509.

31. Nowicki-Osuch K, Zhuang L, Jammula S, Bleaney CW, Mahbubani KT, Devonshire G, et al. Molecular phenotyping reveals the identity of Barrett's esophagus and its malignant transition. Science. 2021;373(6556):760.

32. Apalla Z, Lallas A, Sotiriou E, Lazaridou E, Ioannides D. Epidemiological trends in skin cancer. Dermatol Pract Concept. 2017;7(2):1-6.

33. Pickering CR, Zhou JH, Lee JJ, Drummond JA, Peng SA, Saade RE, et al. Mutational Landscape of Aggressive Cutaneous Squamous Cell Carcinoma. Clinical Cancer Research. 2014;20(24):6582.

34. Madan V, Lear JT, Szeimies R-M. Non-melanoma skin cancer. The Lancet. 2010;375(9715):673-85.

35. Inman GJ, Wang J, Nagano A, Alexandrov LB, Purdie KJ, Taylor RG, et al. The genomic landscape of cutaneous SCC reveals drivers and a novel azathioprine associated mutational signature. Nature Communications. 2018;9(1):3667.

36. Sasaki Y, Tamura M, Koyama R, Nakagaki T, Adachi Y, Tokino T. Genomic characterization of esophageal squamous cell carcinoma: Insights from next-generation sequencing. World J Gastroenterol. 2016;22(7):2284-93.

37. Li XC, Wang MY, Yang M, Dai HJ, Zhang BF, Wang W, et al. A mutational signature associated with alcohol consumption and prognostically significantly mutated driver genes in esophageal squamous cell carcinoma. Annals of Oncology. 2018;29(4):938-44.

38. Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG, et al. Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. New England Journal of Medicine. 2014;371(26):2488-98.

39. Lee-Six H, Øbro NF, Shepherd MS, Grossmann S, Dawson K, Belmonte M, et al. Population dynamics of normal human blood inferred from somatic mutations. Nature. 2018;561(7724):473-8.

40. Xie M, Lu C, Wang J, McLellan MD, Johnson KJ, Wendl MC, et al. Age-related mutations associated with clonal hematopoietic expansion and malignancies. Nature Medicine. 2014;20(12):1472-8.

41. Moore L, Leongamornlert D, Coorens THH, Sanders MA, Ellis P, Dentro SC, et al. The mutational landscape of normal human endometrial epithelium. Nature. 2020;580(7805):640-6.

42. Suda K, Nakaoka H, Yoshihara K, Ishiguro T, Tamura R, Mori Y, et al. Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. Cell Reports. 2018;24(7):1777-89.

43. Lee-Six H, Olafsson S, Ellis P, Osborne RJ, Sanders MA, Moore L, et al. The landscape of somatic mutation in normal colorectal epithelial cells. Nature. 2019;574(7779):532-7.

44. Blokzijl F, de Ligt J, Jager M, Sasselli V, Roerink S, Sasaki N, et al. Tissue-specific mutation accumulation in human adult stem cells during life. Nature. 2016;538(7624):260-4.

45. Lawson ARJ, Abascal F, Coorens THH, Hooks Y, O'Neill L, Latimer C, et al. Extensive heterogeneity in somatic mutation and selection in the human bladder. Science. 2020;370(6512):75.

46. Li R, Du Y, Chen Z, Xu D, Lin T, Jin S, et al. Macroscopic somatic clonal expansion in morphologically normal human urothelium. Science. 2020;370(6512):82.

47. Brunner SF, Roberts ND, Wylie LA, Moore L, Aitken SJ, Davies SE, et al. Somatic mutations and clonal dynamics in healthy and cirrhotic human liver. Nature. 2019;574(7779):538-42.

48. Yoshida K, Gowers KHC, Lee-Six H, Chandrasekharan DP, Coorens T, Maughan EF, et al. Tobacco smoking and somatic mutations in human bronchial epithelium. Nature. 2020;578(7794):266-72.

49. Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proceedings of the National Academy of Sciences. 2013;110(6):1999.

50. Alcolea MP, Greulich P, Wabik A, Frede J, Simons BD, Jones PH. Differentiation imbalance in single oesophageal progenitor cells causes clonal immortalization and field change. Nature cell biology. 2014;16(6):615.

51. Colom B, Alcolea MP, Piedrafita G, Hall MWJ, Wabik A, Dentro SC, et al. Spatial competition shapes the dynamic mutational landscape of normal esophageal epithelium. Nature Genetics. 2020.

52. Alcolea MP, Jones PH. Lineage Analysis of Epidermal Stem Cells. Cold Spring Harbor Perspectives in Medicine. 2014;4(1).

53. Park S, Greco V, Cockburn K. Live imaging of stem cells: answering old questions and raising new ones. Current Opinion in Cell Biology. 2016;43:30-7.

54. Rompolas P, Mesa KR, Kawaguchi K, Park S, Gonzalez D, Brown S, et al. Spatiotemporal coordination of stem cell commitment during epidermal homeostasis. Science. 2016;352(6292):1471-4.

55. Blanpain C, Simons BD. Unravelling stem cell dynamics by lineage tracing. Nature Reviews Molecular Cell Biology. 2013;14(8):489-502.

56. Metzker ML. Sequencing technologies — the next generation. Nature Reviews Genetics. 2010;11(1):31-46.

57. Hari R, Parthasarathy S. Next Generation Sequencing Data Analysis. In: Ranganathan S, Gribskov M, Nakai K, Schönbach C, editors. Encyclopedia of Bioinformatics and Computational Biology. Oxford: Academic Press; 2019. p. 157-63.

58. Buels R, Yao E, Diesh CM, Hayes RD, Munoz-Torres M, Helt G, et al. JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biology. 2016;17(1):66.

59. Gerstung M, Papaemmanuil E, Campbell PJ. Subclonal variant calling with multiple samples and prior knowledge. Bioinformatics. 2014;30(9):1198-204.

60. Abascal F, Harvey LMR, Mitchell E, Lawson ARJ, Lensing SV, Ellis P, et al. Somatic mutation landscapes at single-molecule resolution. Nature. 2021;593(7859):405-10.

61. Gonzalez-Perez A, Sabarinathan R, Lopez-Bigas N. Local Determinants of the Mutational Landscape of the Human Genome. Cell. 2019;177(1):101-14.

62. Martincorena I, Raine KM, Gerstung M, Dawson KJ, Haase K, Van Loo P, et al. Universal Patterns of Selection in Cancer and Somatic Tissues. Cell. 2017;171(5):1029-41.e21.

63. Mularoni L, Sabarinathan R, Deu-Pons J, Gonzalez-Perez A, López-Bigas N. OncodriveFML: a general framework to identify coding and non-coding regions with cancer driver mutations. Genome Biology. 2016;17(1):128.

64. Greaves M, Maley CC. Clonal evolution in cancer. Nature. 2012;481:306.

65. Kakiuchi N, Ogawa S. Clonal expansion in non-cancer tissues. Nature Reviews Cancer. 2021;21(4):239-56.

66. Chow AY. Cell cycle control by oncogenes and tumor suppressors: driving the transformation of normal cells into cancerous cells. Nature Education. 2010;3(9):7.

67. Martínez-Jiménez F, Muiños F, Sentís I, Deu-Pons J, Reyes-Salazar I, Arnedo-Pac C, et al. A compendium of mutational cancer driver genes. Nature Reviews Cancer. 2020;20(10):555-72.

68. Payne SR, Kemp CJ. Tumor suppressor genetics. Carcinogenesis. 2005;26(12):2031-45.

69. Klein G, Klein E. Evolution of tumours and the impact of molecular oncology. Nature. 1985;315(6016):190-5.

70. Nei M, Gojobori T. Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Molecular biology and evolution. 1986;3(5):418-26.

71. Brandes N, Linial N, Linial M. Quantifying gene selection in cancer through protein functional alteration bias. Nucleic Acids Research. 2019;47(13):6642-55.

72. Lawrence MS, Stojanov P, Polak P, Kryukov GV, Cibulskis K, Sivachenko A, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature. 2013;499(7457):214-8.

73. Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, et al. MuSiC: Identifying mutational significance in cancer genomes. Genome Research. 2012;22(8):1589-98.

74. Tamborero D, Gonzalez-Perez A, Lopez-Bigas N. OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. Bioinformatics. 2013;29(18):2238-44.

75. Porta-Pardo E, Godzik A. e-Driver: a novel method to identify protein regions driving cancer. Bioinformatics. 2014;30(21):3109-14.

76. Kamburov A, Lawrence MS, Polak P, Leshchiner I, Lage K, Golub TR, et al. Comprehensive assessment of cancer missense mutation clustering in protein structures. Proceedings of the National Academy of Sciences. 2015;112(40):E5486.

77. Chang MT, Asthana S, Gao SP, Lee BH, Chapman JS, Kandoth C, et al. Identifying recurrent mutations in cancer reveals widespread lineage diversity and mutational specificity. Nature Biotechnology. 2016;34(2):155-63.

78. Porta-Pardo E, Kamburov A, Tamborero D, Pons T, Grases D, Valencia A, et al. Comparison of algorithms for the detection of cancer drivers at subgene resolution. Nature Methods. 2017;14(8):782-8.

79. Bailey MH, Tokheim C, Porta-Pardo E, Sengupta S, Bertrand D, Weerasinghe A, et al. Comprehensive Characterization of Cancer Driver Genes and Mutations. Cell. 2018;173(2):371-85.e18.

80. Collier O, Stoven V, Vert J-P. LOTUS: A single- and multitask machine learning algorithm for the prediction of cancer driver genes. PLOS Computational Biology. 2019;15(9):e1007381.

81. Klein AM, Simons BD. Universal patterns of stem cell fate in cycling adult tissues. Development. 2011;138(15):3103-11.

82. Lim X, Tan SH, Koh WLC, Chau RMW, Yan KS, Kuo CJ, et al. Interfollicular epidermal stem cells self-renew via autocrine Wnt signaling. Science. 2013;342(6163):1226-30.

83. Antal T, Krapivsky PL. Exact solution of a two-type branching process: models of tumor progression. Journal of Statistical Mechanics: Theory and Experiment. 2011;2011(08):P08018.

84. Piedrafita G, Kostiou V, Wabik A, Colom B, Fernandez-Antoran D, Herms A, et al. A single-progenitor model as the unifying paradigm of epidermal and esophageal epithelial maintenance in mice. Nature Communications. 2020;11(1):1429.

85. Mascré G, Dekoninck S, Drogat B, Youssef KK, Brohée S, Sotiropoulou PA, et al. Distinct contribution of stem and progenitor cells to epidermal maintenance. Nature. 2012;489(7415):257-62.

86. Aragona M, Sifrim A, Malfait M, Song Y, Van Herck J, Dekoninck S, et al. Mechanisms of stretch-mediated skin expansion at single-cell resolution. Nature. 2020;584(7820):268-73.

87. Sada A, Jacob F, Leung E, Wang S, White BS, Shalloway D, et al. Defining the cellular lineage hierarchy in the interfollicular epidermis of adult skin. Nature Cell Biology. 2016;18(6):619-31.

88. Klein AM, Brash DE, Jones PH, Simons BD. Stochastic fate of p53-mutant epidermal progenitor cells is tilted toward proliferation by UV B during preneoplasia. Proceedings of the National Academy of Sciences. 2010;107(1):270-5.

89. Kostiou V. The role of space in homeostasis and preneoplasia in stratified squamous epithelia: University of Cambridge; 2020.

90. Moran PAP, editor Random processes in genetics. Mathematical Proceedings of the Cambridge Philosophical Society; 1958: Cambridge University Press.

91. Wright S. Evolution in Mendelian Populations. Genetics. 1931;16(2):97-159.

92. Williams T, Bjerknes R. Stochastic Model for Abnormal Clone Spread through Epithelial Basal Layer. Nature. 1972;236(5340):19-21.

93. Milde F, Tauriello G, Haberkern H, Koumoutsakos P. SEM++: A particle model of cellular growth, signaling and migration. Computational Particle Mechanics. 2014;1(2):211-27.

94. Osborne JM, Fletcher AG, Pitt-Francis JM, Maini PK, Gavaghan DJ. Comparing individual-based approaches to modelling the self-organization of multicellular tissues. PLOS Computational Biology. 2017;13(2):e1005387.

95. Van Liedekerke P, Palm MM, Jagiella N, Drasdo D. Simulating tissue mechanics with agent-based models: concepts, perspectives and some novel results. Computational Particle Mechanics. 2015;2(4):401-44.

96. Hoekstra AG, Kroc J, Sloot PM. Simulating complex systems by cellular automata: Springer; 2010.

97. Maley CC, Forrest S. Exploring the Relationship between Neutral and Selective Mutations in Cancer. Artificial Life. 2000;6(4):325-45.

98. Lynch M, Lynch C, Craythorne E, Liakath-Ali K, Mallipeddi R, Barker J, et al. Spatial constraints govern competition of mutant clones in human epidermis. Nature Communications. 2017;8(1):1119.

99. Kostiou V, Hall MW, Jones PH, Hall BA. Different responses to cell crowding determine the clonal fitness of p53 and Notch inhibiting mutations in squamous epithelia. bioRxiv. 2020.

100. Graner F, Glazier JA. Simulation of biological cell sorting using a two-dimensional extended Potts model. Physical review letters. 1992;69(13):2013.

101. R. Noppe A, Roberts AP, Yap AS, Gomez GA, Neufeld Z. Modelling wound closure in an epithelial cell sheet using the cellular Potts model. Integrative Biology. 2015;7(10):1253-64.

102. Metzcar J, Wang Y, Heiland R, Macklin P. A Review of Cell-Based Computational Modeling in Cancer Biology. JCO Clinical Cancer Informatics. 2019(3):1-13.

103. Sütterlin T, Tsingos E, Bensaci J, Stamatas GN, Grabe N. A 3D self-organizing multicellular epidermis model of barrier formation and hydration with realistic cell morphology based on EPISIM. Scientific Reports. 2017;7(1):43472.

104. Fletcher Alexander G, Osterfield M, Baker Ruth E, Shvartsman Stanislav Y. Vertex Models of Epithelial Morphogenesis. Biophysical Journal. 2014;106(11):2291-304.

105. Brodland GW. How computational models can help unlock biological systems. Seminars in Cell & Developmental Biology. 2015;47-48:62-73.

106. Torres NV, Santos G. The (Mathematical) Modeling Process in Biosciences. Frontiers in Genetics. 2015;6(354).

107. Servedio MR, Brandvain Y, Dhole S, Fitzpatrick CL, Goldberg EE, Stern CA, et al. Not Just a Theory—The Utility of Mathematical Models in Evolutionary Biology. PLOS Biology. 2014;12(12):e1002017.

108. Noble R. These few lines [Internet]2016. [18/08/2021]. Available from: https://thesefewlines.wordpress.com/2016/11/20/the-box-einstein-surface-of-mathematical-models/.

109. Dotto GP, Rustgi Anil K. Squamous Cell Cancers: A Unified Perspective on Biology and Genetics. Cancer Cell. 2016;29(5):622-37.

110. Martincorena I, Campbell Peter J. Somatic mutation in cancer and normal cells. Science. 2015;349(6255):1483-9.

111. Abbosh C, Venkatesan S, Janes SM, Fitzgerald RC, Swanton C. Evolutionary dynamics in pre-invasive neoplasia. Current Opinion in Systems Biology. 2017;2:1-8.

112. Teixeira VH, Nadarajan P, Graham TA, Pipinikas CP, Brown JM, Falzon M, et al. Stochastic homeostasis in human airway epithelium is achieved by neutral competition of basal cell progenitors. Elife. 2013;2:e00966.

113. Lan X, Jörg DJ, Cavalli FM, Richards LM, Nguyen LV, Vanner RJ, et al. Fate mapping of human glioblastoma reveals an invariant stem cell hierarchy. Nature. 2017;549(7671):227-32.

114. Butler RJ, McDonald JB. Using incomplete moments to measure inequality. Journal of Econometrics. 1989;42(1):109-19.

115. Mesa KR, Kawaguchi K, Cockburn K, Gonzalez D, Boucher J, Xin T, et al. Homeostatic Epidermal Stem Cell Self-Renewal Is Driven by Local Differentiation. Cell Stem Cell. 2018.

116. Kiser MM, Dolan JW. Selecting the best curve fit. LC GC NORTH AMERICA. 2004;22(2):112-7.

117. Pardoe I SL, Young D. Stat 462, R-squared Cautions: The Pennsylvania State University; [Available from: https://online.stat.psu.edu/stat462/node/98/.

118. Williams MJ, Werner B, Heide T, Curtis C, Barnes CP, Sottoriva A, et al. Quantification of subclonal selection in cancer from bulk sequencing data. Nature Genetics. 2018;50(6):895-903.

119. Eyre-Walker A, Keightley PD. The distribution of fitness effects of new mutations. Nature Reviews Genetics. 2007;8:610.

120. Kassen R, Bataillon T. Distribution of fitness effects among beneficial mutations before selection in experimental populations of bacteria. Nature Genetics. 2006;38:484.

121. McDonald Michael J, Cooper Tim F, Beaumont Hubertus JE, Rainey Paul B. The distribution of fitness effects of new beneficial mutations in Pseudomonas fluorescens. Biology Letters. 2011;7(1):98-100.

122. Orr HA. The Distribution of Fitness Effects Among Beneficial Mutations. Genetics. 2003;163:1519-26.

123. Rokyta DR, Beisel CJ, Joyce P, Ferris MT, Burch CL, Wichman HA. Beneficial Fitness Effects Are Not Exponential for Two Viruses. Journal of Molecular Evolution. 2008;67(4):368.

124. Cannataro VL, McKinley SA, St. Mary CM. The implications of small stem cell niche sizes and the distribution of fitness effects of new mutations in aging and tumorigenesis. Evolutionary Applications. 2016;9(4):565-82.

125. MacLean RC, Perron GG, Gardner A. Diminishing returns from beneficial mutations and pervasive epistasis shape the fitness landscape for rifampicin resistance in Pseudomonas aeruginosa. Genetics. 2010;186(4):1345-54.

126. Lipinski KA, Barber LJ, Davies MN, Ashenden M, Sottoriva A, Gerlinger M. Cancer Evolution and the Limits of Predictability in Precision Cancer Medicine. Trends in Cancer. 2016;2(1):49-63.

127. Davoli T, Xu Andrew W, Mengwasser Kristen E, Sack Laura M, Yoon John C, Park Peter J, et al. Cumulative Haploinsufficiency and Triplosensitivity Drive Aneuploidy Patterns and Shape the Cancer Genome. Cell. 2013;155(4):948-62.

128. Rennell D, Bouvier SE, Hardy LW, Poteete AR. Systematic mutation of bacteriophage T4 lysozyme. Journal of Molecular Biology. 1991;222(1):67-88.

129. Kotler E, Shani O, Goldfeld G, Lotan-Pompan M, Tarcic O, Gershoni A, et al. A Systematic p53 Mutation Library Links Differential Functional Impact to Cancer Mutation Pattern and Evolutionary Conservation. Molecular Cell. 2018;71(1):178-90.e8.

130. Ashworth A, Lord Christopher J, Reis-Filho Jorge S. Genetic Interactions in Cancer Progression and Treatment. Cell. 2011;145(1):30-8.

131. Rozhok AI, DeGregori J. Toward an evolutionary model of cancer: Considering the mechanisms that govern the fate of somatic mutations. Proceedings of the National Academy of Sciences. 2015;112(29):8914.

132. Williams MJ, Werner B, Barnes CP, Graham TA, Sottoriva A. Identification of neutral tumor evolution across cancer types. Nature Genetics. 2016;48(3):238-44.

133. Tarabichi M, Martincorena I, Gerstung M, Leroi AM, Markowetz F, Dentro SC, et al. Neutral tumor evolution? Nature Genetics. 2018;50(12):1630-3.

134. Bozic I, Paterson C, Waclaw B. On measuring selection in cancer from subclonal mutation frequencies. PLOS Computational Biology. 2019;15(9):e1007368.

135. Chkhaidze K, Heide T, Werner B, Williams MJ, Huang W, Caravagna G, et al. Spatially constrained tumour growth affects the patterns of clonal selection and neutral drift in cancer genomic data. PLOS Computational Biology. 2019;15(7):e1007243.

136. Noble R, Burri D, Kather JN, Beerenwinkel N. Spatial structure governs the mode of tumour evolution. bioRxiv. 2019:586735.

137. Milholland B, Dong X, Zhang L, Hao X, Suh Y, Vijg J. Differences between germline and somatic mutation rates in humans and mice. Nature Communications. 2017;8:15183.

138. Jones PH, Harper S, Watt FM. Stem cell patterning and fate in human epidermis. Cell. 1995;80(1):83-93.

139. Uhlén M, Fagerberg L, Hallström BM, Lindskog C, Oksvold P, Mardinoglu A, et al. Tissue-based map of the human proteome. Science. 2015;347(6220).

140. Fowler DM, Fields S. Deep mutational scanning: a new style of protein science. Nature Methods. 2014;11(8):801-7.

141. Bara JJ, Richards RG, Alini M, Stoddart MJ. Concise Review: Bone Marrow-Derived Mesenchymal Stem Cells Change Phenotype Following In Vitro Culture: Implications for Basic Research and the Clinic. STEM CELLS. 2014;32(7):1713-23.

142. Sun T-T. Altered phenotype of cultured urothelial and other stratified epithelial cells: implications for wound healing. American Journal of Physiology-Renal Physiology. 2006;291(1):F9-F21.

143. Bromberg Y, Rost B. Comprehensive in silico mutagenesis highlights functionally important residues in proteins. Bioinformatics. 2008;24(16):i207-i12.

144. Abildgaard AB, Stein A, Nielsen SV, Schultz-Knudsen K, Papaleo E, Shrikhande A, et al. Computational and cellular studies reveal structural destabilization and degradation of MLH1 variants in Lynch syndrome. Elife. 2019;8:e49138.

145. Shorthouse D, Hall MW, Hall BA. Computational Saturation Screen Reveals the Landscape of Mutations in Human Fumarate Hydratase. Journal of Chemical Information and Modeling. 2021;61(4):1970-80.

146. Raimondi D, Orlando G, Tabaro F, Lenaerts T, Rooman M, Moreau Y, et al. Large-scale in-silico statistical mutagenesis analysis sheds light on the deleteriousness landscape of the human proteome. Scientific Reports. 2018;8(1):16980.

147. Alexandrov LB, Kim J, Haradhvala NJ, Huang MN, Tian Ng AW, Wu Y, et al. The repertoire of mutational signatures in human cancer. Nature. 2020;578(7793):94-101.

148. Kopan R, Ilagan MXG. The Canonical Notch Signaling Pathway: Unfolding the Activation Mechanism. Cell. 2009;137(2):216-33.

149. Aster JC, Pear WS, Blacklow SC. The Varied Roles of Notch in Cancer. Annu Rev Pathol. 2017;12:245-75.

150. Lloyd-Lewis B, Mourikis P, Fre S. Notch signalling: sensor and instructor of the microenvironment to coordinate cell fate and organ morphogenesis. Current Opinion in Cell Biology. 2019;61:16-23.

151. Weisshuhn Philip C, Sheppard D, Taylor P, Whiteman P, Lea Susan M, Handford Penny A, et al. Non-Linear and Flexible Regions of the Human Notch1 Extracellular Domain Revealed by High-Resolution Structural Studies. Structure. 2016;24(4):555-66.

152. Cordle J, RedfieldZ C, Stacey M, van der Merwe PA, Willis AC, Champion BR, et al. Localization of the Delta-like-1-binding Site in Human Notch-1 and Its Modulation by Calcium Affinity. Journal of Biological Chemistry. 2008;283(17):11785-93.

153. Steinbuck MP, Winandy S. A Review of Notch Processing With New Insights Into Ligand-Independent Notch Signaling in T-Cells. Frontiers in Immunology. 2018;9:1230.

154. Hambleton S, Valeyev NV, Muranyi A, Knott V, Werner JM, McMichael AJ, et al. Structural and Functional Properties of the Human Notch-1 Ligand Binding Region. Structure. 2004;12(12):2173-83.

155. De Strooper B, Annaert W, Cupers P, Saftig P, Craessaerts K, Mumm JS, et al. A presenilin-1-dependent γ-secretase-like protease mediates release of Notch intracellular domain. Nature. 1999;398(6727):518-22.

156. Tagami S, Okochi M, Yanagida K, Ikuta A, Fukumori A, Matsumoto N, et al. Regulation of Notch Signaling by Dynamic Changes in the Precision of S3 Cleavage of Notch-1. Molecular and Cellular Biology. 2008;28(1):165.

157. Xu T-H, Yan Y, Kang Y, Jiang Y, Melcher K, Xu HE. Alzheimer's disease-associated mutations increase amyloid precursor protein resistance to γ-secretase cleavage and the Aβ42/Aβ40 ratio. Cell Discovery. 2016;2(1):16026.

158. Yue P, Li Z, Moult J. Loss of Protein Structure Stability as a Major Causative Factor in Monogenic Disease. Journal of Molecular Biology. 2005;353(2):459-73.

159. Ittisoponpisan S, Islam SA, Khanna T, Alhuzimi E, David A, Sternberg MJE. Can Predicted Protein 3D Structures Provide Reliable Insights into whether Missense Variants Are Disease Associated? Journal of Molecular Biology. 2019;431(11):2197-212.

160. Luca VC, Jude KM, Pierce NW, Nachury MV, Fischer S, Garcia KC. Structural basis for Notch1 engagement of Delta-like 4. Science. 2015;347(6224):847.

161. Luca VC, Kim BC, Ge C, Kakuda S, Wu D, Roein-Peikar M, et al. Notch-Jagged complex structure implicates a catch bond in tuning ligand sensitivity. Science. 2017;355(6331):1320.

162. Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. Nucleic Acids Research. 2005;33(suppl_2):W382-W8.

163. The UniProt C. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Research. 2018;47(D1):D506-D15.

164. Putignano V, Rosato A, Banci L, Andreini C. MetalPDB in 2018: a database of metal sites in biological macromolecular structures. Nucleic Acids Research. 2017;46(D1):D459-D64.

165. Cordle J, Johnson S, Zi Yan Tay J, Roversi P, Wilkin MB, de Madrid BH, et al. A conserved face of the Jagged/Serrate DSL domain is involved in Notch trans-activation and cis-inhibition. Nature Structural & Molecular Biology. 2008;15(8):849-57.

166. Goncearenco A, Rager SL, Li M, Sang Q-X, Rogozin IB, Panchenko AR. Exploring background mutational processes to decipher cancer genetic heterogeneity. Nucleic Acids Research. 2017;45(W1):W514-W22.

167. Yang G, Zhou R, Zhou Q, Guo X, Yan C, Ke M, et al. Structural basis of Notch recognition by human γ-secretase. Nature. 2019;565(7738):192-7.

168. Fay JC, Wu C-I. Sequence divergence, functional constraint, and selection in protein evolution. Annual review of genomics and human genetics. 2003;4(1):213-35.

169. Suckling RJ, Korona B, Whiteman P, Chillakuri C, Holt L, Handford PA, et al. Structural and functional dissection of the interplay between lipid and Notch binding by human Notch ligands. The EMBO Journal. 2017;36(15):2204-15.

170. Cooper GM, Goode DL, Ng SB, Sidow A, Bamshad MJ, Shendure J, et al. Single-nucleotide evolutionary constraint scores highlight disease-causing mutations. Nature methods. 2010;7(4):250-1.

171. Virtanen P, Gommers R, Oliphant TE, Haberland M, Reddy T, Cournapeau D, et al. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nature Methods. 2020.

172. Wong CC, Martincorena I, Rust AG, Rashid M, Alifrangis C, Alexandrov LB, et al. Inactivating CUX1 mutations promote tumorigenesis. Nature Genetics. 2014;46(1):33-8.

173. Sauna ZE, Kimchi-Sarfaty C. Understanding the contribution of synonymous mutations to human disease. Nature Reviews Genetics. 2011;12(10):683-91.

174. Kelly DF, Lake RJ, Middelkoop TC, Fan H-Y, Artavanis-Tsakonas S, Walz T. Molecular Structure and Dimeric Organization of the Notch Extracellular Domain as Revealed by Electron Microscopy. PLOS ONE. 2010;5(5):e10532.

175. Nandagopal N, Santat LA, LeBon L, Sprinzak D, Bronner ME, Elowitz MB. Dynamic Ligand Discrimination in the Notch Signaling Pathway. Cell. 2018;172(4):869-80.e19.

176. Willis A, Jung EJ, Wakefield T, Chen X. Mutant p53 exerts a dominant negative effect by preventing wild-type p53 from binding to the promoter of its target genes. Oncogene. 2004;23(13):2330-8.

177. Brender JR, Zhang Y. Predicting the Effect of Mutations on Protein-Protein Binding Interactions through Structure-Based Interface Profiles. PLOS Computational Biology. 2015;11(10):e1004494.

178. Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, Martin M, et al. SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource allows 40-fold increase in coverage of structure-based annotations for proteins. Nucleic Acids Research. 2018;47(D1):D482-D9.

179. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. MDAnalysis: A toolkit for the analysis of molecular dynamics simulations. Journal of Computational Chemistry. 2011;32(10):2319-27.

180. Worth CL, Preissner R, Blundell TL. SDM--a server for predicting effects of mutations on protein stability and malfunction. Nucleic acids research. 2011;39(Web Server issue):W215-W22.

181. Juritz E, Fornasari MS, Martelli PL, Fariselli P, Casadio R, Parisi G. On the effect of protein conformation diversity in discriminating among neutral and disease related single amino acid substitutions. BMC Genomics. 2012;13(4):S5.

182. David A, Sternberg MJE. The Contribution of Missense Mutations in Core and Rim Residues of Protein–Protein Interfaces to Human Disease. Journal of Molecular Biology. 2015;427(17):2886-98.

183. Tokuriki N, Stricher F, Serrano L, Tawfik DS. How Protein Stability and New Functions Trade Off. PLOS Computational Biology. 2008;4(2):e1000002.

184. Siepel A, Bejerano G, Pedersen JS, Hinrichs AS, Hou M, Rosenbloom K, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome research. 2005;15(8):1034-50 %@ 88-9051.

185. Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, et al. The human genome browser at UCSC. Genome research. 2002;12(6):996-1006 %@ 88-9051.

186. Humphrey W, Dalke A, Schulten K. VMD: Visual molecular dynamics. Journal of Molecular Graphics. 1996;14(1):33-8.

187. Stone JE. An efficient library for parallel ray tracing and animation. 1998.

188. Cunningham F, Achuthan P, Akanni W, Allen J, Amode M R, Armean IM, et al. Ensembl 2019. Nucleic Acids Research. 2018;47(D1):D745-D51.

189. Phipson B, Smyth Gordon K. Permutation P-values Should Never Be Zero: Calculating Exact P-values When Permutations Are Randomly Drawn. Statistical Applications in Genetics and Molecular Biology2010.

190. Smith ZR, Wells CS, editors. Central limit theorem and sample size. annual meeting of the Northeastern Educational Research Association, Kerhonkson, New York; 2006.

191. Peters N, Opherk C, Zacherle S, Capell A, Gempel P, Dichgans M. CADASIL-associated Notch3 mutations have differential effects both on ligand binding and ligand-induced Notch3 receptor signaling through RBP-Jk. Experimental Cell Research. 2004;299(2):454-64.

192. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. Journal of the royal statistical society Series B (Methodological). 1995:289-300.

193. Bellavia D, Checquolo S, Campese AF, Felli MP, Gulino A, Screpanti I. Notch3: from subtle structural differences to functional diversity. Oncogene. 2008;27(38):5092-8.

194. Bray SJ. Notch signalling in context. Nature Reviews Molecular Cell Biology. 2016;17(11):722-35.

195. Waskom ML. Seaborn: statistical data visualization. Journal of Open Source Software. 2021;6(60):3021.

196. Huang C-H, Mandelker D, Schmidt-Kittler O, Samuels Y, Velculescu VE, Kinzler KW, et al. The Structure of a Human p110α/p85α Complex Elucidates the Effects of Oncogenic PI3Kα Mutations. Science. 2007;318(5857):1744.

197. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic acids research. 2018;46(D1):D1062-D7.

198. Burke JE, Perisic O, Masson GR, Vadas O, Williams RL. Oncogenic mutations mimic and enhance dynamic events in the natural activation of phosphoinositide 3-kinase p110α (&lt;em&gt;PIK3CA&lt;/em&gt;). Proceedings of the National Academy of Sciences. 2012;109(38):15259.

199. Gkeka P, Evangelidis T, Pavlaki M, Lazani V, Christoforidis S, Agianian B, et al. Investigating the Structure and Dynamics of the PIK3CA Wild-Type and H1047R Oncogenic Mutant. PLOS Computational Biology. 2014;10(10):e1003895.

200. Yeh C-H, Bellon M, Nicot C. FBXW7: a critical tumor suppressor of human cancers. Molecular Cancer. 2018;17(1):115.

201. Yumimoto K, Nakayama KI. Recent insight into the role of FBXW7 as a tumor suppressor. Seminars in Cancer Biology. 2020;67:1-15.

202. Hao B, Oehlmann S, Sowa ME, Harper JW, Pavletich NP. Structure of a Fbw7-Skp1-Cyclin E Complex: Multisite-Phosphorylated Substrate Recognition by SCF Ubiquitin Ligases. Molecular Cell. 2007;26(1):131-43.

203. Close V, Close W, Kugler SJ, Reichenzeller M, Yosifov DY, Bloehdorn J, et al. FBXW7 mutations reduce binding of NOTCH1, leading to cleaved NOTCH1 accumulation and target gene activation in CLL. Blood. 2019;133(8):830-9.

204. Forbes SA, Beare D, Boutselakis H, Bamford S, Bindal N, Tate J, et al. COSMIC: somatic cancer genetics at high-resolution. Nucleic Acids Research. 2017;45(D1):D777-D83.

205. Mészáros B, Erdős G, Dosztányi Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. Nucleic Acids Research. 2018;46(W1):W329-W37.

206. Ishikawa Y, Hosogane M, Okuyama R, Aoyama S, Onoyama I, Nakayama KI, et al. Opposing functions of Fbxw7 in keratinocyte growth, differentiation and skin tumorigenesis mediated through negative regulation of c-Myc and Notch. Oncogene. 2013;32(15):1921-32.

207. Lensink MF, Velankar S, Wodak SJ. Modeling protein–protein and protein–peptide complexes: CAPRI 6th edition. Proteins: Structure, Function, and Bioinformatics. 2017;85(3):359-77.

208. Peterson LX, Roy A, Christoffer C, Terashi G, Kihara D. Modeling disordered protein interactions from biophysical principles. PLOS Computational Biology. 2017;13(4):e1005485.

209. Frankell AM, Jammula S, Li X, Contino G, Killcoyne S, Abbas S, et al. The landscape of selection in 551 esophageal adenocarcinomas defines genomic biomarkers for the clinic. Nature Genetics. 2019;51(3):506-16.

210. Goldman M, Kaplan DM. Comparing distributions by multiple testing across quantiles or CDF values. Journal of Econometrics. 2018;206(1):143-66.

211. Kitayner M, Rozenberg H, Kessler N, Rabinovich D, Shaulov L, Haran TE, et al. Structural Basis of DNA Recognition by p53 Tetramers. Molecular Cell. 2006;22(6):741-53.

212. Zhao Y, Zhang X, Chen Y, Lu S, Peng Y, Wang X, et al. Crystal Structures of PI3Kα Complexed with PI103 and Its Derivatives: New Directions for Inhibitors Design. ACS Medicinal Chemistry Letters. 2014;5(2):138-42.

213. Cho Y, Gorina S, Jeffrey PD, Pavletich NP. Crystal structure of a p53 tumor suppressor-DNA complex: understanding tumorigenic mutations. Science. 1994;265(5170):346.

214. Gorina S, Pavletich NP. Structure of the p53 Tumor Suppressor Bound to the Ankyrin and SH3 Domains of 53BP2. Science. 1996;274(5289):1001.

215. Wang Y, Rosengarth A, Luecke H. Structure of the human p53 core domain in the absence of DNA. Acta Crystallographica Section D. 2007;63(3):276-81.

216. Natan E, Baloglu C, Pagel K, Freund SMV, Morgner N, Robinson CV, et al. Interaction of the p53 DNA-Binding Domain with Its N-Terminal Extension Modulates the Stability of the p53 Tetramer. Journal of Molecular Biology. 2011;409(3):358-68.

217. Arbely E, Natan E, Brandt T, Allen MD, Veprintsev DB, Robinson CV, et al. Acetylation of lysine 120 of p53 endows DNA-binding specificity at effective physiological salt concentration. Proceedings of the National Academy of Sciences. 2011;108(20):8251.

218. Kitayner M, Rozenberg H, Rohs R, Suad O, Rabinovich D, Honig B, et al. Diversity in DNA recognition by p53 revealed by crystal structures with Hoogsteen base pairs. Nature Structural & Molecular Biology. 2010;17(4):423-9.

219. Chen Y, Dey R, Chen L. Crystal Structure of the p53 Core Domain Bound to a Full Consensus Site as a Self-Assembled Tetramer. Structure. 2010;18(2):246-56.

220. Chen Y, Zhang X, Dantas Machado AC, Ding Y, Chen Z, Qin PZ, et al. Structure of p53 binding to the BAX response element reveals DNA unwinding and compression to accommodate base-pair insertion. Nucleic Acids Research. 2013;41(17):8368-76.

221. Bethuyne J, De Gieter S, Zwaenepoel O, Garcia-Pino A, Durinck K, Verhelle A, et al. A nanobody modulates the p53 transcriptional program without perturbing its functional architecture. Nucleic Acids Research. 2014;42(20):12928-38.

222. Martinez-Zapien D, Ruiz FX, Poirson J, Mitschler A, Ramirez J, Forster A, et al. Structure of the E6/E6AP/p53 complex required for HPV-mediated degradation of p53. Nature. 2016;529(7587):541-5.

223. Vainer R, Cohen S, Shahar A, Zarivach R, Arbely E. Structural Basis for p53 Lys120-Acetylation-Dependent DNA-Binding Mode. Journal of Molecular Biology. 2016;428(15):3013-25.

224. Golovenko D, Bräuning B, Vyas P, Haran TE, Rozenberg H, Shakked Z. New Insights into the Role of DNA Shape on Its Recognition by p53 Proteins. Structure. 2018;26(9):1237-50.e6.

225. Ehebauer Matthias T, Chirgadze Dimitri Y, Hayward P, Martinez Arias A, Blundell Tom L. High-resolution crystal structure of the human Notch 1 ankyrin domain. Biochemical Journal. 2005;392(1):13-20.

226. Nam Y, Sliz P, Song L, Aster JC, Blacklow SC. Structural Basis for Cooperativity in Recruitment of MAML Coactivators to Notch Transcription Complexes. Cell. 2006;124(5):973-83.

227. Gordon WR, Roy M, Vardar-Ulu D, Garfinkel M, Mansour MR, Aster JC, et al. Structure of the Notch1-negative regulatory region: implications for normal activation and pathogenic signaling in T-ALL. Blood. 2009;113(18):4381-90.

228. Wu Y, Cain-Hom C, Choy L, Hagenbeek TJ, de Leon GP, Chen Y, et al. Therapeutic antibody targeting of individual Notch receptors. Nature. 2010;464(7291):1052-7.

229. Taylor P, Takeuchi H, Sheppard D, Chillakuri C, Lea SM, Haltiwanger RS, et al. Fringe-mediated extension of &lt;em&gt;O&lt;/em&gt;-linked fucose in the ligand-binding region of Notch1 increases binding to mammalian Notch ligands. Proceedings of the National Academy of Sciences. 2014;111(20):7290.

230. Li Z, Fischer M, Satkunarajah M, Zhou D, Withers SG, Rini JM. Structural basis of Notch O-glucosylation and O–xylosylation by mammalian protein–O-glucosyltransferase 1 (POGLUT1). Nature Communications. 2017;8(1):185.

231. Hackos David H, Lupardus Patrick J, Grand T, Chen Y, Wang T-M, Reynen P, et al. Positive Allosteric Modulators of GluN2A-Containing NMDARs with Distinct Modes of Action and Impacts on Circuit Function. Neuron. 2016;89(5):983-99.

232. Volgraf M, Sellers BD, Jiang Y, Wu G, Ly CQ, Villemure E, et al. Discovery of GluN2A-Selective NMDA Receptor Positive Allosteric Modulators (PAMs): Tuning Deactivation Kinetics via Structure-Based Design. Journal of Medicinal Chemistry. 2016;59(6):2760-79.

233. Villemure E, Volgraf M, Jiang Y, Wu G, Ly CQ, Yuen P-w, et al. GluN2A-Selective Pyridopyrimidinone Series of NMDAR Positive Allosteric Modulators with an Improved in Vivo Profile. ACS Medicinal Chemistry Letters. 2017;8(1):84-9.

234. Heffron TP, Heald RA, Ndubaku C, Wei B, Augistin M, Do S, et al. The Rational Design of Selective Benzoxazepin Inhibitors of the α-Isoform of Phosphoinositide 3-Kinase Culminating in the Identification of (S)-2-((2-(1-Isopropyl-1H-1,2,4-triazol-5-yl)-5,6-dihydrobenzo[f]imidazo[1,2-d][1,4]oxazepin-9-yl)oxy)propanamide (GDC-0326). Journal of Medicinal Chemistry. 2016;59(3):985-1002.

235. Qin L-Y, Ruan Z, Cherney RJ, Dhar TGM, Neels J, Weigelt CA, et al. Discovery of 7-(3-(piperazin-1-yl)phenyl)pyrrolo[2,1-f][1,2,4]triazin-4-amine derivatives as highly potent and selective PI3Kδ inhibitors. Bioorganic & Medicinal Chemistry Letters. 2017;27(4):855-61.

236. Song K, Yang, X., Zhao, Y., Jian, Z. Crystal structure of PI3K complex with an inhibitor, doi:10.2210/pdb5XGI/pdb. 2018.

237. Ouvry G, Clary L, Tomas L, Aurelly M, Bonnary L, Borde E, et al. Impact of Minor Structural Modifications on Properties of a Series of mTOR Inhibitors. ACS Medicinal Chemistry Letters. 2019;10(11):1561-7.

238. Fradera X, Methot JL, Achab A, Christopher M, Altman MD, Zhou H, et al. Design of selective PI3Kδ inhibitors using an iterative scaffold-hopping workflow. Bioorganic & Medicinal Chemistry Letters. 2019;29(18):2575-80.

239. Cheng H, Orr STM, Bailey S, Brooun A, Chen P, Deal JG, et al. Structure-Based Drug Design and Synthesis of PI3Kα-Selective Inhibitor (PF-06843195). Journal of Medicinal Chemistry. 2021;64(1):644-61.

240. Wood ER, Shewchuk LM, Ellis B, Brignola P, Brashear RL, Caferro TR, et al. 6-Ethynylthieno[3,2-d]- and 6-ethynylthieno[2,3-d]pyrimidin-4-anilines as tunable covalent modifiers of ErbB kinases. Proceedings of the National Academy of Sciences. 2008;105(8):2773.

241. Qiu C, Tarrant MK, Choi SH, Sathyamurthy A, Bose R, Banjade S, et al. Mechanism of Activation and Inhibition of the HER4/ErbB4 Kinase. Structure. 2008;16(3):460-7.

242. Liu P, Bouyain S, Eigenbrot C, Leahy DJ. The ErbB4 extracellular region retains a tethered-like conformation in the absence of the tether. Protein Science. 2012;21(1):152-5.

243. Yalla K, Elliott C, Day JP, Findlay J, Barratt S, Hughes ZA, et al. FBXW7 regulates DISC1 stability via the ubiquitin-proteosome system. Molecular Psychiatry. 2018;23(5):1278-86.

244. Cho H-S, Mason K, Ramyar KX, Stanley AM, Gabelli SB, Denney DW, et al. Structure of the extracellular region of HER2 alone and in complex with the Herceptin Fab. Nature. 2003;421(6924):756-60.

245. Garrett TPJ, McKern NM, Lou M, Elleman TC, Adams TE, Lovrecz GO, et al. The Crystal Structure of a Truncated ErbB2 Ectodomain Reveals an Active Conformation, Poised to Interact with Other ErbB Receptors. Molecular Cell. 2003;11(2):495-505.

246. Huyvetter M, De Vos J, Xavier C, Pruszynski M, Sterckx YGJ, Massa S, et al. &lt;sup&gt;131&lt;/sup&gt;I-labeled Anti-HER2 Camelid sdAb as a Theranostic Tool in Cancer Treatment. Clinical Cancer Research. 2017;23(21):6616.

247. Garrett TPJ, McKern NM, Lou M, Elleman TC, Adams TE, Lovrecz GO, et al. Crystal Structure of a Truncated Epidermal Growth Factor Receptor Extracellular Domain Bound to Transforming Growth Factor α. Cell. 2002;110(6):763-73.

248. Li S, Schmitz KR, Jeffrey PD, Wiltzius JJW, Kussie P, Ferguson KM. Structural basis for inhibition of the epidermal growth factor receptor by cetuximab. Cancer Cell. 2005;7(4):301-11.

249. Lim Y, Yoo J, Kim M-S, Hur M, Lee EH, Hur H-S, et al. GC1118, an Anti-EGFR Antibody with a Distinct Binding Epitope and Superior Inhibitory Activity against High-Affinity EGFR Ligands. Molecular Cancer Therapeutics. 2016;15(2):251.

250. Wernimont AK, Dong, A., Seitova, A., Crombet, L., Khutoreskaya, G., Edwards, A.M., Arrowsmith, C.H., Bountra, C., Weigelt, J., Cossar, D., Dobrovetsky, E. Crystal Structure of Metabotropic glutamate receptor 3 precursor in presence of LY341495 antagonist, doi: 10.2210/pdb3SM9/pdb. 2011.

251. Monn JA, Prieto L, Taboada L, Pedregal C, Hao J, Reinhard MR, et al. Synthesis and Pharmacological Characterization of C4-Disubstituted Analogs of 1S,2S,5R,6S-2-Aminobicyclo[3.1.0]hexane-2,6-dicarboxylate: Identification of a Potent, Selective Metabotropic Glutamate Receptor Agonist and Determination of Agonist-Bound Human mGlu2 and mGlu3 Amino Terminal Domain Structures. Journal of Medicinal Chemistry. 2015;58(4):1776-94.

252. Chen C, Gorlatova N, Kelman Z, Herzberg O. Structures of p63 DNA binding domain in complexes with half-site and with spacer-containing full response elements. Proceedings of the National Academy of Sciences. 2011;108(16):6456.

253. Chen C, Gorlatova N, Herzberg O. Pliable DNA Conformation of Response Elements Bound to Transcription Factor p63*. Journal of Biological Chemistry. 2012;287(10):7477-86.

254. Xu X, Choi Sung H, Hu T, Tiyanont K, Habets R, Groot Arjan J, et al. Insights into Autoregulation of Notch3 from Structural and Functional Studies of Its Negative Regulatory Region. Structure. 2015;23(7):1227-35.

255. Bernasconi-Elias P, Hu T, Jenkins D, Firestone B, Gans S, Kurth E, et al. Characterization of activating mutations of NOTCH3 in T-cell acute lymphoblastic leukemia and anti-leukemic activity of NOTCH3 inhibitory antibodies. Oncogene. 2016;35(47):6077-86.

256. Himanen JP, Yermekbayeva L, Janes PW, Walker JR, Xu K, Atapattu L, et al. Architecture of Eph receptor clusters. Proceedings of the National Academy of Sciences. 2010;107(24):10860.

257. Lechtenberg BC, Gehring, M.P., Pasquale, E.B. Crystal structure of the EphA2 intracellular KD-SAM domains, doi: 10.2210/pdb7KJA/pdb. 2021.

258. Araç D, Boucard AA, Bolliger MF, Nguyen J, Soltis SM, Südhof TC, et al. A novel evolutionarily conserved domain of cell-adhesion GPCRs mediates autoproteolysis. The EMBO Journal. 2012;31(6):1364-78.

259. Barr AJ, Ugochukwu E, Lee WH, King ONF, Filippakopoulos P, Alfano I, et al. Large-Scale Structural Analysis of the Classical Human Protein Tyrosine Phosphatome. Cell. 2009;136(2):352-63.

260. Waterhouse A, Bertoni M, Bienert S, Studer G, Tauriello G, Gumienny R, et al. SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic Acids Research. 2018;46(W1):W296-W303.

261. Trusina J, Franc J, Novotný A, editors. Generalization of Homogeneity Tests for Weighted Samples and Their Implementation in ROOT. Journal of Physics: Conference Series; 2020: IOP Publishing.

262. Brun R, Rademakers F. ROOT — An object oriented data analysis framework. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment. 1997;389(1):81-6.

263. Galli M, Tejedor E, Wunsch S, editors. A New PyROOT: Modern, Interoperable and More Pythonic. EPJ Web of Conferences; 2020: EDP Sciences.

264. Zhang Y. Epidemiology of esophageal cancer. World journal of gastroenterology: WJG. 2013;19(34):5598.

265. Colom B, Herms A, Hall MWJ, Dentro SC, King C, Sood RK, et al. Precancer: Mutant clones in normal epithelium outcompete and eliminate esophageal micro-tumors. bioRxiv. 2021:2021.06.25.449880.

266. Abby E, Dentro SC, Hall MWJ, Fowler JC, Ong SH, Sood R, et al. &lt;em&gt;Notch1&lt;/em&gt; mutation drives clonal expansion in normal esophageal epithelium but impairs tumor growth. bioRxiv. 2021:2021.06.18.448956.

267. Hall MWJ, Jones PH, Hall BA. Relating evolutionary selection and mutant clonal dynamics in normal epithelia. Journal of The Royal Society Interface. 2019;16(156):20190230.
268. Frede J, Greulich P, Nagy T, Simons BD, Jones PH. A single dividing cell population with imbalanced fate drives oesophageal tumour growth. Nature Cell Biology. 2016;18:967.
269. David GK. On the Generalized "Birth-and-Death" Process. The Annals of Mathematical Statistics. 1948;19(1):1-15.
270. Orr HA. THE POPULATION GENETICS OF ADAPTATION: THE DISTRIBUTION OF FACTORS FIXED DURING ADAPTIVE EVOLUTION. Evolution. 1998;52(4):935-49.
271. Herms A, Colom B, Piedrafita G, Murai K, Ong SH, Fernandez-Antoran D, et al. Levelling out differences in aerobic glycolysis neutralizes the competitive advantage of oncogenic *PIK3CA* mutant progenitors in the esophagus. bioRxiv. 2021:2021.05.28.446104.
272. Tavaré S, Balding DJ, Griffiths RC, Donnelly P. Inferring coalescence times from DNA sequence data. Genetics. 1997;145(2):505-18.
273. Beaumont MA, Zhang W, Balding DJ. Approximate Bayesian computation in population genetics. Genetics. 2002;162(4):2025-35.
274. Toni T, Welch D, Strelkowa N, Ipsen A, Stumpf MP. Approximate Bayesian computation scheme for parameter inference and model selection in dynamical systems. Journal of the Royal Society Interface. 2009;6(31):187-202.
275. Klinger E, Rickert D, Hasenauer J. pyABC: distributed, likelihood-free inference. Bioinformatics. 2018;34(20):3591-3.
276. Smirnov NV. Estimate of deviation between empirical distribution functions in two independent samples. Bulletin Moscow University. 1939;2(2):3-16.
277. Brown S, Pineda CM, Xin T, Boucher J, Suozzi KC, Park S, et al. Correction of aberrant growth preserves tissue homeostasis. Nature. 2017;548(7667):334-7.
278. Ellis SJ, Gomez NC, Levorse J, Mertz AF, Ge Y, Fuchs E. Distinct modes of cell competition shape mammalian tissue morphogenesis. Nature. 2019;569(7757):497-502.
279. Kon S, Ishibashi K, Katoh H, Kitamoto S, Shirai T, Tanaka S, et al. Cell competition with normal epithelial cells promotes apical extrusion of transformed cells through metabolic changes. Nature Cell Biology. 2017;19(5):530-41.
280. Wijewardhane N, Dressler L, Ciccarelli FD. Normal Somatic Mutations in Cancer Transformation. Cancer Cell. 2021;39(2):125-9.
281. Zhang L, Zhou Y, Cheng C, Cui H, Cheng L, Kong P, et al. Genomic Analyses Reveal Mutational Signatures and Frequently Altered Genes in Esophageal Squamous Cell Carcinoma. The American Journal of Human Genetics. 2015;96(4):597-611.
282. Saito Y, Koya J, Araki M, Kogure Y, Shingaki S, Tabata M, et al. Landscape and function of multiple mutations within individual oncogenes. Nature. 2020;582(7810):95-9.
283. Lehner B. Molecular mechanisms of epistasis within and between genes. Trends in Genetics. 2011;27(8):323-31.
284. Esposito A. Cooperation of partially transformed clones: an invisible force behind the early stages of carcinogenesis. Royal Society Open Science.8(2):201532.
285. Kent DG, Green AR. Order Matters: The Order of Somatic Mutations Influences Cancer Evolution. Cold Spring Harbor perspectives in medicine. 2017;7(4):a027060.
286. Clarke M, Woodhouse S, Piterman N, Hall BA, Fisher J, editors. Using State Space Exploration to Determine How Gene Regulatory Networks Constrain Mutation Order in Cancer Evolution2018.
287. Sandoval M, Ying Z, Beronja S. Interplay of opposing fate choices stalls oncogenic growth in murine skin epithelium. eLife. 2021;10:e54618.

288. Farrelly O, Kuri P, Rompolas P. In vivo genetic alteration and lineage tracing of single stem cells by live imaging. Skin Stem Cells: Springer; 2018. p. 1-14.
289. Monroe JG, Srikant T, Carbonell-Bejerano P, Becker C, Lensink M, Exposito-Alonso M, et al. Mutation bias reflects natural selection in Arabidopsis thaliana. Nature. 2022.

Appendix

This thesis includes material from the following published articles and preprints

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

Hall MWJ, Shorthouse D, Jones PH, Hall BA. Investigating structure function relationships in the NOTCH family through large-scale somatic DNA sequencing studies. *bioRxiv*. 2020.

Fowler JC, King C, Bryant C, Hall M, Sood R, Ong SH, et al. Selection of oncogenic mutant clones in normal human skin varies with body site. *Cancer Discovery*. 2020.

Abby E, Dentro SC, Hall MWJ, Fowler JC, Ong SH, Sood R, et al. *Notch1* mutation drives clonal expansion in normal esophageal epithelium but impairs tumor growth. *bioRxiv*. 2021:2021.06.18.448956.

Colom B, Herms A, Hall MWJ, Dentro SC, King C, Sood RK, et al. Precancer: Mutant clones in normal epithelium outcompete and eliminate esophageal micro-tumors. (accepted for publication in *Nature*, and preprint available at *bioRxiv*, 2021:2021.06.25.449880.)