Investigating the Evolutionary Dynamics of Drug Resistance in Colorectal Cancer

Freddie JH Whiting

A thesis submitted in partial fulfilment of
the requirements of the

Degree of
Doctor of Philosophy

School of Biological and Behavioural Sciences,
Queen Mary University of London
&
Centre for Genomics and Computational Biology,
Barts Cancer Institute

29 November, 2021

Statement of Originality

I, Freddie JH Whiting, confirm that the research included within this thesis is my own work or that where it has been carried out in collaboration with, or supported by others, that this is duly acknowledged below and my contribution indicated. Previously published material is also acknowledged below.

I attest that I have exercised reasonable care to ensure that the work is original, and does not to the best of my knowledge break any UK law, infringe any third party's copyright or other Intellectual Property Right, or contain any confidential material.

I accept that the College has the right to use plagiarism detection software to check the electronic version of the thesis.

I confirm that this thesis has not been previously submitted for the award of a degree by this or any other university.

The copyright of this thesis rests with the author and no quotation from it or information derived from it may be published without the prior written consent of the author.

Signature:

Freddie JH Whiting

Date: 29 November, 2021

This work was supported by a PhD fellowship from Life Sciences, Queen Mary University of London.

Abstract

Cancer resistance evolution was presumed to result from either a pre-existing or acquired mutation that survives treatment, re-populating the tumour following therapy. However, it appears cancer cells can adopt both genetic and non-genetic mechanisms to evade treatment, and a much broader range of evolutionary scenarios could drive resistance evolution.

Here, I first develop models that explicitly capture both genetic and non-genetic sources of phenotypic variation in cell populations evolving resistance to therapy. I show that, given different parameters controlling the change in a resistance phenotype per division and the relative fitness cost of resistance, I can distinguish between various evolutionary scenarios, including those that lead to the same proportion of resistance. I subsequently combine these theoretical models with a long-term drug-treatment experiment *in vitro*: I employ a high-resolution lineage tracing technique and metronomic chemotherapy exposure in two colorectal cancer cell models. In one cell-line - *HCT116* - the lineage distributions are consistent with a resistance phenotype being held at a low frequency by a high reversion phenotypic switching rate, or a high relative fitness cost. The other cell-line – *SW620* – exhibits a response that is consistent with a broad range of evolutionary scenarios, all of which have relatively lower switching rates and fitness costs, whilst maintaining the resistant phenotype at a higher frequency within the population.

My data show a role for either plasticity or a high fitness cost in the evolution of drug resistance in these colorectal cancer cell models. These results highlight the importance of including the diverse evolutionary scenarios that produce phenotypic differences within the population when modelling cancer cells' response to therapy. As stymieing resistance requires hampering a tumour's evolution, I argue that designing more effective treatment strategies will depend on accurately describing these diverse routes to resistance.

Acknowledgements

First, I would like to thank both of my supervisors, Trevor Graham and Richard Nichols, for their support throughout my PhD. You both struck a balance between help and freedom that made for an enjoyable and rewarding doctorate. Trevor, your vast knowledge of all things 'cancer evolution' has helped guide the PhD at every turn, and you provided a necessary calming influence whenever you sensed lab difficulties were causing me heartache. Richard, I have always looked forwards to our discussions over coffee, or jointly coding up the project's PopGen problems in R. Your guidance has ensured my work is clear and measured.

Thank you to all the members of the EvoCa lab group, past and present. Whilst too many to name you all individually, a special thanks to Max Mossner who patiently taught me the dark arts of PCR and NGS, Calum Gabbutt who aided my foray into the world of Bayesian modelling, Will Cross who ensured that Thursdays were for the pub, and Annie Baker and Chris Kimberley who both help maintain the lab's smooth operations whilst members come and go. I feel especially lucky to have been part of such a sociable team. Drinks on the Charterhouse lawn should never be underestimated as a source of motivating chats about cancer evolution or welcome encouragement when your project isn't going to plan.

I am grateful to my parents, Shirley and Peter, who have always supported me whatever the endeavour, have never told me which path to take, but instead encouraged me to do what I enjoy and do it well. Finally, I would like to thank Sophie. You have guaranteed that the last few years have been filled with fun and happy memories despite some trying and difficult times. I am always grateful for your love and support.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 The Evolution of Drug Resistance in Cancer

1.1.1 The Challenge of Resistance

As the development of sequencing technologies has revealed the high levels of heterogeneity within and between tumour's genomes, so too has the appreciation that these rapidly mutating systems are hotbeds for novel phenotypes. Frustratingly, this growing source of adaptive potential can hamper efforts to target cancer-specific phenotypes (Greaves 2018). Viewing tumours through an evolutionary lens as competing clonal lineages governed by Darwinian selection is now commonplace. As clinical intervention can be thought of as attempts to steer a tumour's evolution, the wealth of genetic and phenotypic diversity presents an obstacle to treatment by providing the raw adaptive material for a tumour to respond to these external challenges. One such response is drug resistance, where cancer cells eventually stop responding to treatment; resistance to therapy is still the primary obstacle to patient survival (Liedtke et al. 2008; Osborne and Schiff 2011; Panczyk 2014; Nikolaou et al. 2018). In combination with providing a source of adaptive variation, the vast number of differences that can distinguish a cancer cell from a healthy cell presents a problem of identification: an additional challenge of tackling drug resistance is also determining which of these changes is responsible for the phenotype that confers treatment resistance.

Cancer drug therapy consists of either traditional chemotherapeutics, newer targeted therapies, or some combination of the two. Whilst immunotherapy offers an exciting new avenue for targeting tumours by leveraging the body's immune system, treatment discussed in relation to this thesis is restricted to chemo- and targeted therapies. Whilst chemotherapy acts broadly, preferentially killing dividing cells by inhibiting cell division and DNA replication, targeted therapy involves killing cells by interfering with specific molecular pathways that are distorted exclusively in malignant cells. Despite substantial differences in specificity, evolution of drug resistance is common to both chemo- and targeted therapies (Holohan et al. 2013).

To effectively tackle the problem of resistance evolution, the first logical step is to identify which molecular changes in cancer cells has rendered them refractory to treatment. As such, research attention has often focused on the biochemical mechanisms of resistance. For chemotherapy, such as treatment with platinum compounds and antimetabolites, these mechanisms can include decreased uptake or increased expulsion of drug, altered proliferation and modifications to DNA damage repair (Siddik 2003; Usanova et al. 2011; Nikolaou et al. 2018). Diverse resistance mechanisms to targeted therapies also exist. The activation of parallel signalling pathways such as MET, an alternative trans-membrane tyrosine kinase that can activate the same pathways as Epidermal Growth Factor Receptor (EGFR) and, as such, confers resistance to EGFR-inhibitors (Tomasello et al. 2018); the amplification of drug targets such as BRAF amplification, which eludes BRAF inhibition by re-activating the target pathway – mitogen-activated protein kinase (MAPK) (Stagni et al. 2018); and mutations in target genes such as T790M, a point mutation in lung cancer that decreases the binding affinity of EGFR-TKI inhibitors to the ATP binding pocket of EGFR, allowing ATP to bind more efficiently and re-activating the oncogenic EGFR protein (Yun et al. 2008; Mok et al. 2017). Cells are not restricted to a single strategy; multiple resistance mechanisms can exist for a single drug (Shi et al. 2014), of which only few are needed for full resistance (Tegze et al. 2012), whilst cells can also adopt multiple mechanisms simultaneously (Cree and Charlton 2017). Treatment evasion is compounded further by pathways that confer resistance to more than one drug: multi-drug-resistance (MDR), which can effect both chemotherapies and targeted therapies simultaneously (Szakács et al. 2006). The most well studied of these is the ATP Binding Cassette (ABC) transporters: efflux pumps that decrease intracellular concentration of therapeutic agents (Robey et al. 2018).

Here, I have only touched on a few of the many mechanisms which have been shown to confer resistance to cancer cells. However, to effectively design strategies that aim to stymie resistance, knowledge of the resistance mechanisms alone is insufficient. The ongoing accrual of changes during neoplastic growth means cancer genomes can be thought of as 'moving targets' (Foo et al. 2013); a static snapshot of a tumour and the nature of its resistance is inadequate if these traits can continue to emerge throughout a tumour's lifespan. Research must therefore also focus on the behaviour of these resistance-conferring changes over time.

1.1.2 Pre-Existing vs De-Novo Resistance Mutations

The genetic instability intrinsic to cancer leads to the establishment of novel mutations during tumour cell divisions. By the time a tumour has become clinically detectable – approximately 10^9 cells – the combined population of tumour cells has had ample opportunity to accrue novel variants, of which some will have direct phenotypic consequences. Prior to recent advantages in sequencing above the level of the genome – including the mapping of RNA and epigenetic landscapes – next-generation sequencing sequencing meant that the diversity under scrutiny in tumours was genetic; attention was primarily focussed on mutations that controlled phenotypes relevant to cancer disease progression. When it comes to resistance evolution, early investigations tended to focus on describing the behaviour of variants that rendered treatment ineffective: resistance-conferring mutations. One pertinent question was whether resistance mutations arose during treatment, or whether they were pre-existing at the beginning of therapy, rendering resistance a *fait accompli* (To avoid confusion, I will speak of resistance mechanisms either solely as *'pre-existing'* prior to treatment, or arising *'de-novo'* following the onset of treatment).

Despite recent advances in the affordability and availability of whole-genome sequencing, studies have often circumvented these drawbacks by focussing on modelling the emergence of resistance, with the aim of answering the pre-existing vs *de-novo* dichotomy. Bozic and Nowak approached the question of whether a resistance-conferring mutation was present at the time of clinical detection, prior to the onset of treatment. They employed a simulation that assumes a genome-wide per-base-pair mutation rate of 10^{-9}/bp per cell division and 100 possible point mutations that can confer resistance. As long as the resistance mutations are not highly deleterious, their results show several resistant mutations should exist in a clinically detectable tumour of 10^9 cells (Bozic and Nowak 2014).

Diaz *et al.* combined similar mathematical modelling with longitudinal samples from colorectal cancer patients' tumours that eventually became resistant to an EGFR inhibitor (a targeted therapy) (Diaz et al. 2012). They also wished to understand whether resistance conferring mutations were present when treatment began, or arose *de-novo* during treatment. As KRAS mutations confer resistance to anti-EGFR therapy, a high-sensitivity assay identified mutants KRAS in longitudinal samples. The study then retrospectively estimated the probability that a resistance mutation was present at the onset of treatment by combining the longitudinal mutant KRAS status data from patients with a branching birth-death model of KRAS sequence evolution. Assuming 42 possible resistance-conferring mutations in the KRAS protein, and a per-base-pair mutation rate of 10^{-9} per cell-division, Diaz and colleagues' modelling also supported the hypothesis that resistance-conferring mutations were present prior to the onset of treatment.

Supporting the notion that genetic instability bestows cancer cells with a diverse adaptive arsenal, additional work using those patients' disease that did not accrue a KRAS mutation revealed that resistance to EGFR inhibition was instead conferred by an amplification of the MET gene. Using a PCR-based assay that targeted the MET amplification in patient blood samples, these genetic aberrations could also be found at very low frequencies before treatment began in the majority of patients studied (Bardelli et al. 2013).

Tomasetti *et al.* applied modelling techniques to gastrointestinal stromal tumours (GIST) to determine the likelihood that mutations conferring resistance to tyrosine kinase inhibitors (TKIs) were pre-existing (Tomasetti et al. 2013). They assume there are 10 possible resistance mutations – an estimate they describe as conservative – and, again, a per-base-pair mutation rate of 10^{-9} per-cell division and estimate the probability of pre-existing resistance as a function of tumour size in cm. Given these estimates in mutation rates and the number of target resistance mutations genome wide, Tomasetti and colleagues found that the probability of a resistance mutation existing at the onset of treatment in GIST was high.

The reliance on mathematical modelling to answer questions of pre-existence has partly been a function of technical constraints; sampling only portion of a tumour and the prohibitive cost of high-depth sequencing has often limited identifying variants present at an extremely low frequency. An *in vitro* solution to this problem was to adopt a novel lineage tracing technique in non-small cell lung cancer (NSCL) to enable the tracking of lineages found at extremely low frequency (Bhang et al. 2015). Semi-random nucleotide sequences were incorporated into cells using a lentiviral infection system. Next-generation sequencing of these markers allows lineages to be tracked at a much higher resolution than methods that rely on genome-wide variants for genealogical reconstruction. By complementing this experimental approach with mathematical modelling, they found that a few lineages were repeatedly shown to be refractory to treatment across replicates, consistent with some rare, pre-existing population of resistance. Lineage tracing technologies offer a remedy to some of the technical constraints encountered when cell relationships are inferred using whole-genome data, and are discussed in detail in section 1.4.

An alternative remedy to the problem of insufficient detection power, only made possible more recently by lower sequencing costs, is to employ ultra-deep sequencing. Blood cancers are especially amenable to evolutionary studies, where the nature of the cancer supports frequent, less invasive sampling, whilst ultra-deep sequencing helps mediate the technical difficulties introduced by the low ratio of malignant to healthy cells in patients' blood samples. In acute-lymphoblastic leukaemia, ultra-deep sequencing was used to track the status of mutations through diagnosis and treatment (Li et al. 2020). Whilst they show that in a subset of patients resistance associated mutations are present at diagnosis, ultra-deep sequencing (median depth 3669X) of serial samples revealed cases where resistance mutations were acquired sequentially, with various resistance mechanisms corresponding to the various compounds that made up the combined chemotherapy treatment regime.

Framing resistance evolution within the 'pre-existing vs de-novo mutation' paradigm has implications for how resistance should be tackled clinically. A lack of pre-existing mutations would suggest a strategy that minimised the opportunity for novel mutations might be optimal, whereas the presence of pre-existing resistance mutations could rule out the redundant application of cytotoxic treatments to patients, instead directing clinical attention to therapies that might still limit malignant growth (Schmitt et al. 2016).

Yet one common assumption has been that, if resistance is the product of one or a few mutations, the right combination of targeted therapies should provide an efficacious means of treatment (Diaz et al. 2012). The pervading model has often framed cancer as a population gradually accruing mutations, where a few sites render certain compounds ineffective, and effective treatment is built around circumventing these mutations. If resistance does proceed in this way, a tumour's adaptive potential is limited by the number of resistance mutations it can accrue, and this capability is in turn constrained by the number of possible genomic sites that confer resistance. Unfortunately, a growing body of work has revealed that this paradigm may be an over-simplification. Instead, it appears that cancer cells can exploit evolutionary options that are not solely the product of resistance-conferring genetic mutations.

1.1.3 Non-Genetic Phenotypic Variability and Drug-Tolerant Persisters

Whilst mutations undoubtedly play a pivotal role in driving the evolution of drug resistance, there is increasing evidence that non-genetic mechanisms can also contribute to treatment evasion. By non-genetic, I refer to any change to the phenotype that is *not* controlled by a change at the level of the genome. Such changes include stochastic changes in gene expression (Payne and Wagner 2019) and epigenetic changes, such as DNA methylation (Glasspool et al. 2006). Whilst these biological phenomena are 'invisible' to traditional genomic sequencing, they can still provide phenotypic variation upon which selection can act.

Whilst gene expression is tightly regulated in normal somatic cells by epigenetic modifications, in cancer these regulatory networks can break down, leading to elevated gene expression variation (Marusyk et al. 2012). Expression dysregulation is associated with drug resistance: targeted inhibition of KDM5 – a gene that regulated expression via chromatin modification - decreased transcription heterogeneity in ER+ breast cancer (Hinohara et al. 2018). This attenuation in gene expression variability led to a decrease in hormone-therapy resistance. It is possible that whilst tight regulation of transcription in healthy cells maintains tissue homeostasis, increasing gene expression variation in malignant cells grants them access to a wide array of phenotypes, some of which could be adaptive under the current conditions. Indeed, *in vitro* studies using glioblastoma showed that, whilst the resistant mutations known to the chosen therapy were absent, epigenetic changes associated with tolerance to tyrosine kinase inhibitors (TKIs) showed increased expression variation in genes under their regulatory control (Liau et al. 2017).

Despite the stochastic nature of the variability in gene expression, at the population level this behaviour can lead to stable proportions of different frequencies. This was shown in *in vitro* breast cancer cells, where one of three of the population's phenotypes were isolated. Following a subsequent growth step, the cells quickly approached the parental population's equilibrium phenotypic frequencies (Gupta et al. 2011).

One proposed mechanism by which increased variation in gene expression can lead to drug resistance is by prolonging survival, providing a larger window of opportunity during which selection can act: a small sub-population of cells that transiently exist in this state are refractory to treatment and have a higher probability of accruing stable, genetic resistant mutations (Brock 2009). This idea has gained traction as the number of studies identifying some form of 'drug tolerant persisters' (herein DTPs) in cancer have grown. By entering a state similar to diapause observed in multi-cellular organisms, DTPs incur a fitness cost by delaying proliferation, yet are simultaneously able to withstand otherwise lethal concentrations of cytotoxic drugs. As such, this strategy is sometimes referred to as ‘bet hedging’. Indeed, this route to resistance has been well documented in bacteria (Balaban et al. 2004; Cohen et al. 2013; Windels et al. 2019).

Some of the earlier, compelling evidence in favour of DTPs in cancer came from Sharma and colleagues who showed that EGFR tyrosine kinase inhibitors (TKIs) resistance in an *in vitro* model of NLSC arose from a population of DTPs (Sharma et al. 2010). Strikingly, the state was reversible, and KDM5 – the aforementioned chromatin modifier – was partly responsible for its maintenance. In this case, the transient nature and implication of an epigenetic regulator are consistent with the non-genetic maintenance of a DTP sub-population. Whilst Hata *et al.* show that a pre-existing mutation in NLSC confers resistance to TKIs, the same resistance mutation can also arise *de novo* in drug-tolerant cells (Hata et al. 2016). Furthermore, their results suggest that the resistance mutation was more refractory to further treatment if it arose *de novo* on a drug-tolerant genetic background. One potential explanation is that the DTP route to resistance allows more mutations to accrue and therefore produces a more robust phenotype. A study in acute-lymphoblastic leukaemia observed a worrying property of resistance evolution whereby chemotherapy treatment itself – namely, an antimetabolite, thiopurine – increased cells’ mutation rate (Li et al. 2020). Modelling supported a subset of these patients’ disease potentially arising from a sub-population of DTPs.

In an *in vitro* model of melanoma, single cells occasionally exist in a short-lived state where they transcribe higher levels of genes associated with resistance to the targeted drug vemuarfenib, a B-Raf inhibitor. After prolonged exposure, the cells gradually increased the number of resistant genes they expressed, and the resistant phenotype became stable: no longer reversible following a break from treatment (Shaffer, Dunagin, et al. 2017). Findings such as these undermine any paradigm of resistance evolution where treatment is only rendered ineffectual by the pre-existence of genetic resistance mutations. There is mounting evidence that cancer cells can call upon non-genetic means of resistance to survive initial rounds of treatment. Stochastic differences in gene expression mean genetically identical cells can exhibit distinct phenotypes (Payne and Wagner 2019); these mechanisms can provide rapid sources of adaptive variation not possible on the time-scale of stable, heritable mutations. Drug-tolerant persisters are an extreme example where a transient phenotype that is extremely costly in the absence of therapy provides a reservoir of cells refractory to treatment.

Finally, one promising line of evidence in the face of these diverse routes to resistance comes from work by Marin-Bejar and colleagues. They use patient-derived xenograft (PDX) models of melanoma to investigate resistance evolution to targeted therapy: BRAF/MEK inhibitors. PDXs could be categorised into groups that evolved resistance either through genetic or non-genetic means. Remarkably, replicate model tumours derived from the same patients followed the same trajectories (Marin-Bejar et al. 2021). This finding hints at the possibility that, although the various ways in which a tumour can become resistant are numerous, cancer cells may be predisposed to following one of several evolutionary routes. If the processes controlling these predispositions were found, it would introduce a measure of predictability and could offer a means to tailor treatment to tumours based on their molecular 'class' of resistance.

1.2 Insights from Evolutionary Biology

1.2.1 Cancer is an Evolutionary Disease

The established conceptual model of cancer is that of an evolutionary disease, subject to Darwinian selection; sub-populations (or sub-clones) of cells accrue genetic mutations that can produce variable, heritable phenotypes, leading to differential survival that manifests as changes in lineage frequency changes over time (Nowell 1976). As such, the theory developed to describe the principles of evolution in organisms can be adopted as a conceptual framework to understand cancer (Michor et al. 2004), and pertinent questions previously asked in evolutionary biology and population genetics can help build a predictive model of tumour evolution. Despite work highlighting the suitability of traditional population genetics for cancer (Ohtsuki and Innan 2017), it has often remained underutilised (Aktipis et al. 2011). To name one problem relevant to both fields, quantitative genetics has long sought to describe the selective consequences of new mutations: the distribution of fitness effects (DFE) (Eyre-Walker and Keightley 2007). High-depth, next-generation sequencing (NGS) now allows such questions to be tackled in cancer. For example, within a tumour, the imperfect replication of genomes during cell division means that the evolutionary history of cancer cell lineages are recorded in the number of mutations that any two cells share. By leveraging this information, modelling mutations within a population genetics framework has allowed the estimation of their selection coefficients (Williams et al. 2018). In respect to cancer therapy, as the production and effect size of mutations controlling resistance mechanisms can dictate the probability and strength of resistant phenotypes, robust estimates of such parameters could help inform how cancer cells might respond to treatment. Furthermore, observable phenotypic changes that are incompatible with the time-frames necessary for genetic mutations may well point to alternative biological phenomena. In such cases, theoretical expectations derived from traditional population genetics theory could provide a model with which other modes of phenotypic evolution are compared.

1.2.2 The Pace of Adaptive Evolution

A historically moot point in evolutionary biology was the relative contributions of evolutionary gradualism and saltation (Mayr 1989). Namely, does adaptive evolution occur via many small effect mutations or few, large effect macromutations? A well-established metaphor for considering adaptive evolution is that of an adaptive landscape (Wright 1932). Organisms occupy a position on a three-dimensional landscape: the x and y-axes represent a genotype, whilst the z-axis – the height of the point – represents the fitness of said genotype. A change in genotype – a mutation - can move the organism either towards or away from a peak, increasing or decreasing fitness. The adaptive landscape is not static: environmental changes and frequency-dependent effects can move the location of peaks, of which there can be numerous and, of course, in reality organisms are evolving through a high-dimensional space. Fisher's geometric analogy developed this metaphor of a landscape and showed that small changes in a genotype were more likely to approach the peak an organism was currently climbing, and therefore increase fitness. On the other hand, larger changes in genotype were more likely to overshoot the peak, decreasing fitness (Fisher 1958). Fisher argued that adaptive evolution was therefore likely the product of many mutations of small effect.

One criticism of Fisher's geometric argument is that organisms are approaching a single peak which only moves gradually over time (Orr and Coyne 1992). Yet in the case of cancer, something as detrimental as therapy likely places cells far from a nearby fitness peak on the adaptive landscape, and here large effect mutations may in fact allow them to more rapidly escape the low-fitness 'valley' they now occupy. One such 'mutation' which may allow cells to traverse the landscape more rapidly are chromosomal aberrations, typified by aneuploidy – an abnormal number of chromosomes. These large structural changes may provide cells with a quick and crude means to explore a wider assortment of phenotypes compared to single nucleotide substitutions, yet with the associated cost of altering multiple genes simultaneously. Indeed, there is evidence that aneuploidy allows yeast cells to rapidly explore a wider fitness landscape in adverse conditions: aneuploidy was associated with the rapid evolution of cytokinesis restoration in yeast cells where it had been artificially disrupted (Rancati et al. 2008); cells with an artificially introduced chromosomal amplification had a wider variance in fitness than single-gene amplifications when nutrients were limiting (Sunshine et al. 2015); and aneuploidy was an adaptive response to heat stress in experimental evolution of yeast cells (Yona et al. 2012). Furthermore, supporting the notion that aneuploidy is a rapid but, ultimately, a costly source of adaptive variation, the chromosomal gain was eventually replaced by a reversion to euploidy; higher fitness was subsequently achieved instead via the greater expression of certain heat tolerant genes (Yona et al. 2012).

Aneuploidy is found in the majority of common cancers (Sansregret et al. 2018). The rapid accumulation of genomic aberrations would increase the probability of a rapid traversal across the adaptive landscape. In fact, chromosomal copy-number-alterations (CNAs) can occur in a simultaneous, punctuated fashion in both prostate (Baca et al. 2013) and colorectal (Cross et al. 2018) cancer. The co-occurrence of these aberrations in tumours' evolutionary history suggest that these large-scale genomic aberrations can be adaptive in cancer. Whilst punctuated genomic changes need not lead to punctuated phenotypic change (Graham and Sottoriva 2017), there is evidence aneuploidy can produce phenotypic differences. Intermediate levels of aneuploidy confer worse patient outcome, suggesting that structural genomic changes lead to phenotypic changes that are adaptive to the cancer cells (Birkbak et al. 2011), whilst whole-genome doubling helps buffer cells against detrimental aneuploid phenotypes by decreasing the proportion of chromosomes gained or lost following structural changes (Dewhurst et al. 2014). Whilst there is some limited evidence that aneuploidy is associated with drug resistance in cancer cells (Swanton et al. 2009; Lee et al. 2011), more work is necessary to elucidate the relationship between large-scale chromosomal changes and the rate of a tumour's phenotypic evolution, including drug resistance.

Even when the rate of accrual of genetic changes in cancer has been described, this need not lead to a corresponding rate of change in a cell's phenotype. Indeed, such inferences are confounded by the non-linear relationship between genotype and phenotype, coined the genotype-phenotype map. For example, single mutations occurring within key regulatory genes can result in gross phenotypic changes that presage the progression of cancer (Drost et al. 2015). Describing the pace of adaptive *phenotypic* evolution may be therefore insufficient if relying solely on comparisons of genetic data. In fact, as discussed in section 1.1.3, a change in phenotype may occur independently of any genetic change via phenomena such as stochastic changes in gene expression. Like changes in chromosome copy number, these differences provide rapid sources of novel phenotypic variation not possible on the time-scale necessary for single-nucleotide point mutations. Charlebois and colleagues developed a model where cells had a probability of transiently shifting their gene expression profile (Charlebois et al. 2011). They modelled a continuous resistance phenotype where survival was contingent on cells expressing some threshold quantity of the hypothetical gene product. By increasing the variance in the cells' resistance gene expression, non-genetic sources of variability could push a higher proportion of a population into the phenotypic space that permitted survival during treatment.

It is worth noting that these different evolutionary scenarios could be modelled as points on a continuum, where the rate of phenotypic change is allowed to vary: if changes in phenotype were extremely low per cell-division, the model would resemble genetic mutations, whereas if the rates were much higher, it could instead capture stochastic phenotypic variation. This idea will receive more attention in the first of the results chapters.

1.2.3 The Evolution of Asexual Populations

Both tumour and microbial evolution see periods of rapid expansion of individuals lacking recombination. As such, similarities have frequently been drawn between the two. These shared features can influence how selection operates. For example, selection is less efficient in growing populations (Korolev et al. 2012) – differences in fitness are manifest as differences in expansion rates, whereas in fixed population sizes deleterious lineages are more likely to be driven to extinction. The lack of recombination also affects how evolution progresses. The study of large, asexually evolving populations has led to extensive theoretical and experimental work on the fate of mutations in populations lacking recombination (Gerrish and Lenski 1998; Fogle et al. 2008; Couce and Tenaillon 2015). Importantly, facets of asexual evolution can limit adaptive evolution.

Clonal interference occurs when beneficial mutations arising on different genetic backgrounds cannot recombine and therefore must compete (Gerrish and Lenski 1998). During the experimental evolution of *Saccharomyces cerevisiae*, clonal interference was shown to maintain multiple, competing lineages (Blundell, Schwartz, et al. 2019). Whilst diversity was initially predictable as single-mutant lineages arose, the dynamics became stochastic following the arrival of double-mutants. These double-mutants preceded a crash in diversity at highly variable intervals .

Muller's ratchet occurs when the accumulation of deleterious mutations cannot be ameliorated via recombination (Gabriel et al. 1993). In cancer, whole-genome doubling (WGD) appears to help limit the negative effects caused by the gradual accumulation of detrimental mutations (Lopez et al. 2019): loss of heterozygosity (LOH) in cancer can lead to deleterious mutations becoming present as both alleles. WGD appears to buffer this effect by limiting the potential for LOH.

Here I have outlined a few features of asexual evolution that might influence how tumours evolve. It is yet unclear to what extent these features of asexual evolution could be manipulated to limit adaption in the face of therapy. At the very least, a thorough understanding of their behaviour in cancer is necessary to describe the fate of beneficial mutations in full.

1.3 Evolutionarily Informed Strategies

1.3.1 Exploitable Evolutionary Constraints

As studies have increasingly investigated diseases through a Darwinian lens, it has become clear that natural axes of constraint imposed by evolutionary phenomena can provide windows of opportunity for therapeutic intervention: biophysical limits can restrict the direction in which traits can evolve, whilst phenomena such as pleiotropy and epistasis mean a new mutation can influence traits that are not under positive selection.

Supporting the idea that a patient's tumour must be viewed as a 'moving target', exploitable phenotypes are often temporary in nature, and include states that are less fit in non-treated environments or to other drug-treatments. These vulnerable phenotypes are often discussed with reference to a 'fitness trade-off' and the notion of a 'cost of resistance'. To avoid confusion, these concepts first deserve some clarification.

One way that a 'cost of resistance' can be invoked is by simply considering adaption to different environments. If 'non-treated' and 'drug-treatment' are two environmental conditions, we might consider a 'sensitive' phenotype, S, that has high relative fitness in the 'non-treated' environment and low relative fitness in the 'drug-treatment' environment. Similarly, we could imagine a 'resistant' phenotype, R, where the reverse is true. In this scenario, the term 'cost of resistance' captures the lower relative fitness of the resistant phenotype in the non-treated environment. As pointed out by Lenormand *et al.*, we could just as easily discuss the 'cost of sensitivity' by highlighting the negative relative fitness of the S phenotype in the drug-treated environment, although the concept is never framed in this fashion (Lenormand et al. 2018). Importantly, we can explain a 'cost of resistance' without any reference to epistasis, pleiotropy or life-history trade-offs.

Now, if we were to discuss *why* adaption to these different environments leads to different fitness effects of the phenotypes S and R, we could discuss possible trade-offs imposed by natural axes of constraint: fitness is composed of an individual's survival and reproduction advantage, and resource and biological limitations mean increasing one often comes at the cost of the other (Michod et al. 2006; Lenski 2017). For example, in a tumour, cells might pay the metabolic cost of increasing trans-membrane efflux pump number in exchange for more efficient expulsion of cytotoxic compounds (Kam et al. 2015). In the non-treated environment, cells will incur the metabolic cost whilst reaping none of the survival benefits. Such trade-offs are relevant to 'adaptive therapy', as discussed shortly.

Some of these axes of constraint will impose hard limits on the directions in which traits can evolve: cells have a limited metabolic budget - should energy be 'invested' in efflux pump expression or elsewhere? Nonetheless, some constraints may be surmountable given selection for a sufficient length of time. Pleiotropy can restrict the direction any single mutation can move an organism through 'trait-space'. Assuming cells have evolved to have high relative fitness in the untreated environment, phenotypic changes are more likely to be detrimental than positive (see Fisher's Landscape Analogy in section 1.2.2). As such, a mutation that confers high fitness in the drug-treated environment – a resistance mutation – is also likely to produce other changes that have negative fitness effects in both the treated and non-treated environments. For example, in cancer a large chromosomal aberration might change multiple genes simultaneously and could be interpreted as a 'quick and crude' means of creating adaptive variation. Given continued evolution in the drug-treated environment, we might expect these additional penalties to be mitigated over time by compensatory mutations. These mutations would reduce the pleiotropic cost, but a difference in relative fitness between the two treatment environments might still be interpreted as a 'cost of resistance'. Alternatively, in what would represent the worst-case (clinical) scenario, compensatory mutations might lead to a resistance phenotype that also retains full fitness in the untreated environment, precluding any clinical intervention which might aim to exploit these differences in fitness.

1.3.2 Competitive Release and Adaptive Therapy

In light of the well-studied evolutionary phenomenon of 'competitive release' (Connell 1961) where the growth of a species occurs rapidly when its competitor(s) is removed, the emergence of drug resistant subclones should come as no surprise. By removing the resistant cells' sensitive competitors with therapy, the resistant cells are free to utilise previously unattainable resources (West et al. 2018). Adaptive therapy (AT) is an evolutionary informed therapy that attempts to limit the competitive release of resistant sub-populations (Gatenby 2009). It capitalises on the idea that tumour cells may 'pay' for the resistant phenotype with some 'cost'. Again, to work, this need only assume that there is cell competition, and some fitness difference between the resistant and sensitive phenotypes in the two environments: treated and non-treated. Opposed to the current standard of 'maximum tolerated dose' (MTD), AT proposes 'drug vacations' to allow the faster growing sensitive cells to outgrow and hinder the expansion of drug-resistant sub-clones. Recent work has begun to investigate the efficacy of such a strategy.

A spatial agent-based model by Gallaher and colleagues showed that modulating the dose to encourage competition between sensitive and resistant cells can, in theory, prolong the time to treatment failure (Gallaher, Enriquez-Navas, et al. 2017). The optimal dosing strategies to ensure containment of resistant cells has been subject to theoretical investigation (Gluzman et al. 2020). These include the recent findings that containment strategy employed by AT can work even if resistant cells don't incur a fitness cost, as long as there is competition between resistant and sensitive cells (Viossat and Noble 2021). Furthermore, the competition experienced between resistant and sensitive cells is higher in scenarios where the turnover of cells is high (Strobl et al. 2020). That is, when the *sum* of the birth and death rates $(b + d)$ increases whilst fixing the difference between the two $(b - d)$ – the net growth rate. Whilst experimental evidence in favour of AT has been limited, there is evidence that resistant cells proliferate slower than their sensitive contemporaries in the absence of treatment, consistent with resistance incurring some relative fitness cost (Duan et al. 2018). This effect was also observed in resistant colorectal cancer cells *in vitro*, where different ratios of sensitive and resistant cells were either grown in monolayer cultures or organoids (3-

dimensional cultures of cells that preserve some rudimentary tissue structure of the organ of origin). When treated with various concentrations of a targeted therapy – a cyclin-dependent kinase inhibitor (CDKi) – it was only in the organoid cultures that the presence of sensitive cells led to lower numbers of resistant cells after therapy, an observation the authors argue is consistent with amplified competition between the two phenotypes (Bacevic et al. 2017). If true, this would support the hypothesis that spatial structure is another facet of tumour evolution necessary to successfully exploit the competition between resistant and sensitive cells.

1.3.3 Temporal Constraints and Collateral Sensitivity

As cells adapt to their new environment – drug-treatment – the aforementioned biological constraints can limit the direction of evolution and produce less fit phenotypes in a given environment. However, evolution towards a new fitness optimum often occurs as a stepwise process: the phenotypes a cell expresses on its adaptive route may only offer transient exploitable windows for intervention. Work has therefore begun to identify and characterise the dynamics at play cells traverse these phenotypes.

Opposed to broad acting cytotoxic chemotherapies, targeted therapies interact with specific gene pathways or molecules. We might therefore expect resistance to resemble a binary state, where a genetic change might change the shape of the treatment's protein target. However, in non-small cell lung cancer, resistance to ALK-inhibition – a targeted therapy – was not the product of a single or few nucleotide changes to the ALK protein sequence, but instead relied on the cumulative effect of numerous molecular changes. These changes included a point mutation, amplification of the drug-target gene and over expression of genes previously implicated in ALK-inhibition resistance (Vander Velde et al. 2020). Time-series single-cell analysis showed a gradual transition from sensitive to resistance and, importantly, only intermediate states were more sensitive to a second targeted therapy: lapatenib (a HER2 inhibitor). If full resistance was allowed to evolve, this window of opportunity disappeared. The order in which cells are exposed to therapies can also impact resistance evolution. Some breast cancer cells can persist despite exposure to the chemotherapy docetaxel – a taxane (analogous to the DTP phenotype discussed in section 1.1.3). However, administrating inhibitors of pathways activated in the persister cells – SFK and Hck – increases sensitivity to chemotherapy (Goldman et al. 2015). Notably, this effect was only effective if the SFK/Hck pathways were first activated with docetaxel, and not if the drugs were applied simultaneously.

Collateral sensitivity describes the phenomenon where evolving resistance to one line of treatment simultaneously renders cells more sensitive to another. For example, in a murine model of acute-lymphoblastic leukaemia, a single, pre-existing mutation in the V29LL locus provided resistance to a BCR-ALB1 inhibitor (B. Zhao et al. 2016). Resistance to the BCR-ALB1 inhibitor increased cells' response to other targeted therapies, indicative of collateral sensitivity. Importantly, however, if allowed to continue to evolve in the treatment conditions, cells accrued mutations that conferred resistance to the additional treatments. In other work that bridges collateral sensitivity and adaptive therapy, Dhawan and colleagues showed that the magnitude of sensitivity to other therapies in a resistant cell-line was contingent on the length of drug-holiday window between the first and second line of therapy (Dhawan et al. 2017). Surprisingly, the length of holiday and direction of sensitivity was drug-dependent. Whilst gaps in treatment provide respite for patients receiving cytotoxic compounds, it appears they can also mediate competition between resistant and sensitive cells and modulate the efficacy of a second choice of treatment.

Experimental evolution studies in bacteria provide some of the best evidence describing the dynamics of collaterally sensitive and 'costly' resistant phenotypes. In *E.coli*, one of two possible molecular routes to resistance to a range of antibiotics can dictate whether or not cells no longer exposed to treatment retain the resistant phenotype, whilst also regaining their ancestral fitness in the non-treated environment (Knopp and Andersson 2015). In *Pseudomonas aeruginosa*, investigators had previously identified collateral sensitivity to given combinations of clinically relevant antibiotics (Barbosa et al. 2019). Importantly, the order in which the therapies were added dictated whether or not cells retained collateral sensitivity, an outcome likely the product of epistasis between the resistant mutations.

In summary, evolutionary informed strategies can often be framed in terms of exploitable phenotypes that emerge during adaption to new environments; cells that are adapting to a given treatment may simultaneously accrue phenotypic changes that would be realised as fitness penalties if selection pressures were to be switched to non-treated or alternative therapy environments. For strategies such as adaptive therapy, their efficacy will depend on developing further the dynamics of resistance evolution. Features of interest include the rate of change between 'sensitive' and 'resistant' phenotypes, the resistant proportion of the population when treatment begins and the relative fitness of each phenotype in environments under clinical control. The best strategy for any given tumour may not just rely on its initial genetic and phenotypic makeup, but instead require identifying exploitable phenotypes as the tumour evolves. These vulnerable phenotypes can be transient, and therefore understanding if the trajectories on which they lie share common features or are repeatable will aid in designing effective evolutionary-informed therapies.

1.4 Investigating Evolutionary Dynamics with Lineage Tracing

1.4.1 Prospective Lineage Tracing

If we imagine a hypothetical population of cells with equal fitness, following a period of growth during which all cells divide and die with some given rates, we can consider the distribution of descendants each cell lineage has produced. Stochastic effects such as drift and random environmental change will cause variation around some expected lineage size (Greaves and Maley 2012; Basanta and Anderson 2013). If we now also permit cells have to have differences in fitness – either as higher birth rates, lower death rates, or some combination of these – there should be a higher, albeit more predictable component of variance in the lineage distributions (Williams et al. 2018; Graham and Sottoriva 2017). An important part of any experiment aiming to characterise the selection experienced by cells is teasing apart the 'predictable' effects of selection from the stochastic effects of drift. To do this in a population of cells, we require tools that reveal the relative success of individual lineages.

The first class of lineage tracing techniques – coined ‘lineage tracing’ - are prospective; those that introduce distinguishable markers prior to experimental manipulation. These techniques were first used to tackle questions in evolutionary development, where embryologists wished to reveal which cells gave rise to which tissues/organs (Kretzschmar and Watt 2012). These methods rely on visually distinguishable markers, such as radioactive tracers and inducible fluorescent proteins. The ongoing demand for spatially-explicit lineage tracing has given rise to the development of the number of possible distinguishable fluorescent markers. For example, the Brainbow construct supports roughly 100 unique fluorescent markers (Weissman and Pan 2014). Inducible recombination reporters such as the Cre-Lox system in mouse also allow such markers to be expressed in both tissue and time dependent manners (H. Kim et al. 2018).

Whilst visual markers offer a spatially explicit means to trace lineages, the number of possible markers is limited. An alternative approach sacrifices visual identification for resolution; artificial genetic sequences (which are referred to here as ‘barcodes’) can be created in libraries that consist of over one million unique barcodes (Bhang et al. 2015; Levy et al. 2015). Ideally barcodes should be identifiable, selectable and stably heritable. Common selection markers include resistance to a drug (e.g. puromycin) and fluorescent proteins (Lamprecht et al. 2017; Kebschull and Zador 2018). The most common technique for stable genomic integration is lentivirus infection. Whilst the selective neutrality of markers is potentially compromised by insertional mutagenesis: gene disruption via the random genomic site of lentivirus integration (Porter et al. 2014), the proportion of total insertions to total sites in the genome with a relevant, phenotypic consequence is small. Nonetheless, retrospective sequencing of the integration site with techniques such as LAM-PCR can help resolve any uncertainty by allowing researchers to amplify the sequences adjacent to the site of integration (Schmidt et al. 2007). As the genomic location of integration is (semi-)random, studies have even used the integration-site as a unique, heritable marker (Dieter et al. 2011; Giessler et al. 2017).

1.4.2 Retrospective Lineage Tracing

The second category of these methods exploit the ancestry recorded in the imperfect copying of a cell's genome (Graham and Sottoriva 2017). Such naturally occurring changes are both genetic and epigenetic; they include single-nucleotide variants (SNVs), chromosomal copy number variants (CNVs) and methylation patterns in non-expressed genes (Kester and Oudenaarden 2018). One drawback of these approaches is that, due to the small ratio of mutations to total genomic nucleotide positions, the resolution with which lineages can be delineated is limited by sequencing depth and breadth (Woodworth et al. 2017). These drawbacks can be ameliorated in part by using mitochondrial DNA variants, where the mutation rate is higher and genome size smaller (Ludwig et al. 2019). Nonetheless, the resolution limitations and lack of longitudinal sampling mean inferring evolutionary dynamics from naturally occurring genetic markers in primary tumour samples remains technically challenging.

1.4.3 High-Resolution Barcode Techniques

The first challenge faced when using high-complexity barcode pools is determining the number and distribution of unique sequences in the library prior to integration. Ideally, following transformation of the barcode plasmids into bacteria for amplification, colonies would be cultured in small batches, integration efficiency assessed and batches sequenced to create a confident list of the expanded library's barcodes prior to infection (Bystrykh and Belderbos 2016). However, in complex libraries this procedure is technically impractical, and a combination of counting transformed colonies and deep-sequencing of the library can provide an estimate of the number and distribution of unique barcodes.

Infecting cells with the barcode library involves a trade-off between the number of uniquely infected cells and the number of multiple integration events. A satisfactory trade-off is achieved by reducing the multiplicity of infection (m.o.i.): that is, by conducting the experiment so that only a small proportion of the cells are barcoded, it is possible to reduce the rate of multiple infections (the proportion of cells with more than one viral integration). It is convenient to assume that the number of integration events in each cell follows a Poisson distribution; we can then calculate the proportion of cells that should contain 0, 1, 2... etc. barcodes from the average integration rate (Fehse et al. 2004).

Cells are typically barcoded at the start of an experiment, and after the experimental treatment a survey is carried out to estimate the proportion of cells descended from each of the barcoded founder cells. The barcodes must therefore be isolated from genomic DNA. A PCR step simultaneously amplifies the barcodes to levels sufficient for next-generation sequencing (NGS). For this purpose, barcodes are flanked by universal sequences that allow the amplification primers to bind. As a superfluous number of PCR cycles has been shown to impair barcode identification, amplification primers can have (Illumina) sequencing primers integrated, precluding the PCR cycles barcodes would experience during a distinct library-preparation step.

The number of rounds of PCR necessary for barcode amplification and sequencing are minimized, since both processes introduce errors into barcode sequences. As the full set of unique random/semi-random barcodes are not known beforehand, these errors can artificially inflate the number of observed barcodes if erroneous sequences are mistaken as distinct, unique barcodes (Thielecke et al. 2017). As such, various computational techniques – termed 'barcode clustering' - have been developed to try and identify these errors, and re-group the reads with their putative 'parental' barcodes (Bhang et al. 2015; Zorita et al. 2015; L. Zhao et al. 2018; Tambe and Pachter 2019). Yet a systematic review comparing clustering algorithms is lacking. It is therefore prudent to compare the performance of these methods with simulated data, where the true 'parental' barcodes are known beforehand. Researchers can then choose the approach that best recovers true barcode frequencies under their chosen experimental parameters.

The clusters barcodes procured from a sequencing run have been subject to numerous selective bottlenecks, only some of which will be of biological interest. Therefore one must develop methods to model the noise in an experiment to distinguish stochastic changes due to sampling – the null distribution – from 'true' biological differences (Blundell and Levy 2014). Such an approach has already been adopted to simulate the binomial sampling of NGS (Williams et al. 2018) and, in the case of Levy and colleagues, the noise introduced into barcode distributions by PCR, NGS and culture bottlenecks in yeast experimental evolution (Levy et al. 2015).

1.5 Clinical and Molecular Features of Colorectal Cancer

1.5.1 Colorectal Cancer Treatment

Colorectal cancer (CRC) is the 3rd most common cancer in the world (Arnold et al. 2017), risk factors include family history, sedentary lifestyle, smoking and obesity, and it is responsible for roughly 10% of all cancer-related deaths (Kuipers et al. 2015). In healthy tissue, the epithelial layer is arranged into crypt-like glands which are replenished over time by stem cells at the base of the crypt (Humphries et al. 2013). CRC is the consequence of degeneration of the homeostatic regulation of this tissue. This transition from benign adenoma (pre-cancer) to malignant adenocarcinoma is believed to be the product of the sequential accumulation of specific genomic aberrations (Fearon and Vogelstein 1990). Genetically, the intra-tumour heterogeneity identified in CRC appears to be initiated early in the tumour's growth (Sottoriva et al. 2015). Whilst early disease (stage I) can be cured with surgical resection alone, later stages (II-III) are typically treated by surgical resection followed by adjuvant therapy (Nguyen and Duong 2018). The mainstay of treatment continues to be chemotherapy as a combination of leucovorin, 5-fluorouracil (5-Fu) and either oxaliplatin (FOLFOX in combination) or irinotecan (FOLFIRI in combination). There is moderate and strong evidence in favour of employing adjuvant chemotherapy for stage II and III CRC, respectively (Ragnhammar et al. 2001; Wilkinson et al. 2010). Chemotherapy has been combined with targeted therapies, most commonly EGFR-inhibitors (Shankaran et al. 2010). Response rates for combined therapies in a metastatic setting are variable, where overall/disease-free survival is measured in months (Wolpin and Mayer 2013). Surprisingly, a recent phase 3 trial found a significant *disadvantage* to adopting an EGFR-inhibitor (Cetuximab) in a pre-operative, intrahepatic metastatic setting (Bridgewater et al. 2020).

1.5.2 Microsatellite Instability (MSI)

There are two, usually distinct, molecular subtypes of colorectal cancer. The first – microsatellite instability (MSI) – is the product of a dysfunctional mismatch repairs system (dMMR) due to the inactivation of at least one mismatch repair genes, and occurs in roughly 15% of all colorectal cancers (Axel Walther et al. 2009). Microsatellites are tandem repeats of 1-5bp in length, of which there are 100,000s in the human genome, many of which play important roles in gene regulation (Gymrek et al. 2015). Microsatellite containing genes are prone to insertions and deletions (indels) in MSI tumours; the subsequent aberrations can disrupt gene expression (Jung et al. 2004). MMR genes include MLH1, MSH2, PRMS1 PMS2 and MSH6 (Wheeler and Bodmer 2000), and germline mutations in these genes are responsible for the most common form of familial colorectal cancer, Lynch Syndrome (Kuipers et al. 2015). Sporadic MSI colorectal cancers however are primarily due to hypermethylation of MLH1, which inactivates the gene's protein (Hudler 2012). The MSI cancers also have distinct clinical characteristics. They are usually found in the ascending colon, have higher lymphocyte infiltration – likely due to the higher diversity of neoantigens (Llosa et al. 2015) – and are less likely to progress to metastatic disease (Koopman et al. 2009).

1.5.3 Chromosomal Instability (CIN)

The other molecular sub-type of CRC is characterised by chromosomal instability (CIN), occurs in approximately 85% of all CRCs, they are nearly all exclusively microsatellite-stable (MSS), and it is defined by an abnormal numerical or structural chromosomal alterations (Grady and Pritchard 2014). The distinction is often not made between ongoing chromosomal instability and stable aneuploidy. This is despite evidence that sub-clonal differences indicative of CIN play an important role in tumour evolution (Vargas-Rondón et al. 2017). Here, however, I use CIN to refer broadly to aneuploid tumours where ongoing instability may not have been shown explicitly. CIN can be the product of many cellular defects including atypical mitotic checkpoint, assembly and microtubule dynamics (Gordon et al. 2012). These cellular defects lead to aneuploidy, copy number gains or losses and loss of heterozygosity (LOH) (Markowitz and Bertagnolli 2009). CIN CRC tumours are predominantly located in the descending colon, and whilst it is thought that CIN may increase the rate of adaption and therefore help confer resistance to therapy (Giam and Rancati 2015), the specifics of these processes have yet to be characterised in full.

1.5.4 The Consequences of Molecular Sub-Type

The biological consequences of either class of genomic aberration in colorectal cancer (CRC) is complex. There are variable, sometimes contradictory reports on the clinical impact of each molecular sub-type of colorectal cancer (in general MMRd tumours have a better prognosis though, probably as a consequence of the immune predation of neoantigens). For example, studies have reported mixed results regarding the effect MSI status has on the response to 5-Fu therapy (Reimers et al. 2013; Copija et al. 2017). Hveem and colleagues looked at 952 patients and, after stratifying tumours by MSI/non-MSI and stage, found that MSI cancers had better overall survival (OS) at stage II, but not stage III (Hveem et al. 2014). Patients with recurrent MSI colorectal tumours also appear to have worse overall survival (E. S. Kim 2016). One potential issue is that the low representation of patients with advanced dMMR CRC has led to limited statistical power in some studies (Koopman et al. 2009; Guastadisegni et al. 2010). CIN CRC tumours have a worse overall prognosis in stage II and II CRC, irrespective of adjuvant therapy (A. Walther et al. 2008). Overall, however, there is a consensus that MSS CRCs have a worse overall outcome than MSI CRCs (Popat et al. 2005; Axel Walther et al. 2009; Öhrling et al. 2010; Hutchins et al. 2011). Yet in several cancers, when CIN reaches some critical threshold the excess of genomic aberrations actually confers a better prognosis (Birkbak et al. 2011). Within CIN CRCs, the magnitude of aneuploidy doesn't appear to correlate with stage, yet there is a marked increase from adenoma to adenocarcinoma (Orsetti et al. 2014). Sequencing of multiple regions in adenomas and adenocarcinomas supports these findings (Cross et al. 2018).

Viewed through an evolutionary lens, the difference in scale of genomic aberrations found between MSI and CIN CRCs could have consequences for the rate of adaption. As discussed in section 1.2.2, experiments in yeast have shown that aneuploidy may act as a rapid, yet uneconomical source of adaptive variation in adverse environmental conditions. I hypothesise that CIN may similarly provide a means to quickly explore the adaptive landscape following exposure to chemotherapy; assuming large genetic changes likewise cause large phenotypic changes, aberrations on the chromosome level are more likely to move cells out of a fitness valley. However, cells with MSI may accrue mutations with smaller phenotypic effects, but at a faster rate.

1.6 Summary and Thesis Outline

Here, I have outlined some of the features and problems faced by those aiming to describe the evolutionary dynamics of drug resistance evolution in cancer. In particular, how progress in technologies that allow for analysis of molecular features above and beyond that of the genome have revealed that non-genetic factors appear to also play a role in drug resistance evolution. These various routes to resistance available to cancer cells are pertinent for evolutionary informed therapies. These strategies aim to exploit the evolutionary constraints and vulnerable phenotypes found in cancer as resistance emerges. Finally, I have discussed how lineage tracing can be employed to approach some of these questions, and described features of importance in my chosen model system - colorectal cancer.

In my thesis, I employ a prospective lineage tracing technology - the ClonTracer Barcode Library (Bhang et al. 2015) - in two colorectal cancer cell lines in a long-term drug-treatment experiment, *in vitro*. I develop theoretical expectations for how I expect these lineage distributions to look under various evolutionary scenarios. Finally, I compare how sequenced lineage distributions compare to simulated values to infer the rates with which cells transition to and from resistant phenotypes. Identifying likely evolutionary scenarios employed by these cells will point to broader strategies that tumour cells may adopt when insulted with cytoxic treatments. Ultimately, these modes of resistance evolution will dictate which evolutionary strategies will emerge as the most efficacious.

Chapter 2

Materials and Methods

2.1 Sequencing Experiment Designs

Prior to detailing the specific protocols undertaken for each step of my project, I will first briefly outline the different sequencing experiments I performed and the differences between them. The two letter experiment codes used to distinguish between the different experiments are in the respective headings. I use these codes in the subsequent detailed protocols to highlight where different methods were adopted.

2.1.1 Original Trial (OR)

This small trial experiment was performed to ensure that the ClonTracer library was behaving as expected *in vitro* and following sequencing. One of the two CRC cell-lines used in the project - *HCT116* - underwent the optimisation protocols (outlined subsequently) for infection with the expanded ClonTracer library. Cells were not treated with any drug-treatment *in vitro*, but instead were infected, expanded, assigned to three replicates and then grown for two passages. Cells were sampled at each splitting step for DNA extraction, barcode amplification and sequencing. PCR was optimised using the custom sequencing primers that are provided with the ClonTracer barcode library (outlined subsequently). The samples were sequenced on a single-end NextSeq 500 High Output Run (75 cycles). There were no issues identified with sequencing that required bespoke pipeline solutions downstream.

2.1.2 Pulse Run 1 (PR)

Following successful adoption of the ClonTracer library in the OR experiment, I re-optimised the infection and expansion step of the two CRC cell-lines - *HCT116* and *SW620* - for an experiment that now imposes metronomic drug-treatment *in vitro*. Cells were infected with the barcode library, expanded, before being split into respective replicates. Replicate samples are identified according to the following codes:

- `POTi` - a sample of cells sampled directly from the expanded, barcoded pool of cells, prior to assignment to control and drug-treatment replicates. *i* corresponds to replicate i of 8.

- `COi_PN` - a control replicate, passage N. These replicates are *not* subject to the metronomic drug-treatment during time *in vitro*. *i* corresponds to replicate i of 4.

- `DTi_PN` - a drug-treatment replicate, passage N. These replicates are given pulses of chemotherapy treatment - 5-Fu - followed by 'recovery' periods under normal culture conditions. *i* corresponds to replicate i of 4.

Again, cells were sampled at each stage of the experiment and labelled according to the sample codes above. For this first long-term treatment experiment, both cell-lines' control and drug-treatment replicates were grown until Passage 2 (`P2`). During time *in vitro*, cells were grown until any given replicate flask was approximately 80% confluent. Control treatment flasks were grown and passaged according to standard culture conditions (see: Tissue Culture Conditions). Drug-treatment flasks were instead subject to periods of therapy where the estimated IC_{50} values that had been previously estimated (see: Drug Assays) was applied for 3-5 days followed by a recovery period. The samples were sequenced on a paired-end NovaSeq S2 flowcell (50 cycles). Custom filtering steps were required after the identification of 'index-hopping' artefacts in the data (see: Sequencing Platform Specifications (PR) and see section 6.2 for more details).

2.1.3 Pulse Run 2 (QR)

The same experimental design as PR was repeated to assess the repeatability of the results and to procure additional time-points. The core design of this experiment was the same as the first (PR). The same codes are also used to distinguish between samples (`COi_PN`, etc...), however they now contain the `QR` prefix. Important differences between this second attempt and the first are as follows:

- Barcode infection optimisation was re-performed for each cell-line, and barcoding of the cells was repeated prior to the experiment to ensure repeatability of the technique.
- Cells were now successfully grown in all treatment replicates for 5 passages.
- The PCR protocol was adjusted and unique-dual index adapters were used across two flow-cells (see: Barcode Amplification and Adapter Integration (QR) for more details).
- Samples were now sequenced across two NovaSeq S2 flowcells (50 cycles) (see: Sequencing Platform Specifications (QR) for more details).
- Slight 'index-hopping' artefacts remained despite the adoption of unique dual-index adapters. Adjusted custom filtering steps were therefore necessary (see: Section 6.2 for more details).

2.2 Cell Culture

2.2.1 Cell Lines and Tissue Culture Conditions

Two colorectal cancer cell-lines colorectal cell lines HCT116 (ATCC®CCL-247™) and SW620 (ATCC®CCL-227™) were both used for all data generating experiments in this project. Coloretcal cancer can be classified as having microsatellite instability (MSI) or being microstatellite stable (MSS), and these different classes of molecular instability have biological consequences (see Section 1.5 for more details). As such, one cell-line was chosen from each of these molecular sub-types: HCT116 is a near-diploid

MSI colon adenocarcinoma cell-line, whilst SW620 is an MSS cell-line derived form a colorectal lymph node metastasis. Although MSS, SW620 has experienced chromosomal instability (CIN) and has a highly aberrant karyotype (Berg et al. 2017). Cells were grown in 'standard conditions' which were as follows: growth medium, consisting of high glucose DMEM (Gibco - Life Technologies) supplemented with 10% Fetal Bovine Serum (FBS) and 2% Penicillin-Streptomycin (Gibco - Life Technologies) (from hereon in: 'full growth medium', unless stated otherwise). Cells were either grown in T-75 vented flasks (Corning®) with 12mL of full growth medium, or in T-175 vented flasks (Corning®) for the long-term drug-treatment experiment with 35mL of full growth medium. Flasks were grown inside tissue-culture incubators (Hercell VIOS 160i CO_2 incubator - Thermo Scientific™) at 37°C, 5% CO_2 and 95% relative humidity.

2.2.2 Splitting, Counting and Seeding Cells

When splitting cells for a subsequent passage or freezing, cells first had their current growth medium aspirated and were washed in either 10mL (T-75 flasks) or 20mL (T-175 flasks) of phosphate buffered saline (PBS - pH 7.4) (Gibco- Life Technologies). The PBS was then aspirated and 5mL (T-75 flasks) or 10mL (T-175 flasks) of Trypsin-EDTA (1.0X solution made with PBS) (Gibco- Life Technologies) was added. The flask was placed in the incubator at standard conditions for approximately 3 minutes. When the majority of cells were observed as having become detached from the bottom of the flask, equal volumes of full growth medium were added to the flask and aspirated up and down several times to maximise the disassociation of adherent cells. Cells were then centrifuged into a pellet at 300xg for 5 minutes. The leftover trypsin + growth medium was removed and the pellet was then re-suspended in 10mL of PBS. 10µL of the cell solution was then mixed with 10µL of 0.4% trypan blue (Gibco- Life Technologies) and added to a Countess™ Cell Counting Chamber Slides (Thermo Fisher Scientific). Cell numbers were then calculated using the Countess™ Automated Cell Counter. If passaging cells, the desired number of cells were diluted in full growth medium and added to the respective flask. If freezing, cells were processed as described below.

2.2.3 Freezing and Thawing Cells

When freezing cells, cells were first trypsinised and counted as described above. Cells were subsequently centrifuged into a pellet at X RPM for Y minutes. Cells were then resuspended in 1mL of freezing medium: 90% FBS and 10% Dimethyl Sulfoxide (DMSO) (Sigma-Aldrich) and placed inside 2mL cryovials (Corning®). Cryovials are then placed in a Mr.Frosty Freezing Container (Thermo Scientific) containing 100% isopropyl alcohol and placed into a -80°C freezer. To thaw cells, cryovials are removed from a freezing container and quickly placed in a water bath at 37°C. Cells are then quickly pipetted into a T-75 flask with 12mL of full growth medium which is replaced after 24hrs to remove any cell debris and cells that did not survive thawing.

2.2.4 Drug-Assays

The IC_{50} values of each CRC cell line (*HCT116* and *SW620*) were measured as follows: Cells were trypsinised into a single-cell solution and counted as outlined above. Cells were then seeded into a TC-treated 96-well adherent plate (Corning®) (*HCT116* - 8000cells/well and *SW620* - 10000 cells/well) in 200µL of full growth medium, leaving one column of wells free from cells to use as blanks to standardise the colorimetric reading. Cells were excluded from the peripheral wells to avoid edge-effects, and cell-lines were split across multiple plates to avoid plate-effects. After 1 day (+24hrs from start), the full growth medium was carefully aspirated from each well, and replaced with 200µL of full growth medium + drug stock for a range of concentrations. These drug-stock solutions were pre-made at 100x prior to the assays to allow an addition of 2µL to 198µL of full growth medium reaching the desired working concentrations on the day of the assay. Stocks were re-made every 2-4 months to avoid drug attenuation. After three subsequent days (+96hrs from start), CellTiter 96 ® AQueous One Soluton (MTS reagent - Promega Ltd) was thawed (pre-made aliquots of enough solution for 1 plate + 10% had been prepared) in a water bath at 37°C. 30µL of MTS reagent was added to all wells (excluding the edge wells which contain no cells). Once done, the plate(s) were added to the incubator (37°C, 5% CO_2 and 95% relative humidity) for 2.5 hours. After this, plates were removed from the incubator. A plate was gently tapped before being placed in the plate reader to distributed the MTS reagent evenly. The plate lid was removed and any bubbles were removed by gently piercing with a clean pipette tip. Finally, the colorimetric reading was taking at a wavelength of 490nm. The results were exported as a .csv file and the readings passed to a custom R script to calculate the dose-response curve. Specifically, a four parameter log-logistic model is fit to the observed response data using nonlinear least-squares estimation in the R package *'drc'* (Ritz et al. 2015).

2.3 ClonTracer Library Expansion, Production and Infection

2.3.1 Plasmid Expansion

Complex Library Pool Expansion

The ClonTracer library was a gift from Frank Stegmeier (Addgene #67267). The ClonTracer library (Bhang et al. 2015) only comes with enough plasmids for a single cell-line barcoding experiment. Therefore, the first necessary step was to expand the complex plasmid library via electroporation. This was performed as follows: Firstly, 23µL of electrocompenent *Escherichia coli* cells (MegaX DH10B™ T1R Electrocomp™ Cells (Life Technologies)) were added to 1µL of 100ng/µL ClonTracer library 5 times, for a total of 5x cuvettes - 500ng of library in total. According to estimates provided with the library, this would contain approximately 6.5x10^6 barcode molecules. The cuvettes were kept on ice. 2µL of cells + 100µL of SOC Recovery Medium were kept to one side on ice. A Gene Pulser electroporation system (XCell™, Bio-Rad) was set to the following settings: exponential decay wave, voltage at 2.0kV, resistance at 200Ω and capacitance to 25uF. The cuvettes were sequentially dried and carefully placed in the Gene Pulser before being pulsed - the time constant (TC) was recorded and confirmed to fall between the desired values of 4 and 5. 1mL of SOC Recovery Medium was added and to each cuvette before returning to ice. The contents of each cuvette were combined in a 50mL conical tube and incubated in the SOC Recovery Medium for 37°C whilst shaking at 225rpm for 45 minutes. This was repeated in parallel for the non-transformed cells in 100µL of SOC in a separate tube. After 1 hour, 100µL of the transformed cells were plated in 3x serial dilutions (1x10^{-3}, 1x10^{-6}, 1x10^{-7}) before being spread on LB agar + carbenicillin (100µg/mL) plates. On a 4th plate, non-transformed plates were spread as a control to ensure carbenicillin efficiency. These plates were incubated overnight. The remainder of the transformed cells were combined in SOC medium into a large flask with 500mL of LB medium supplemented with 100µg/mL carbenicillin (to select for transformed cells) and were incubated at 37°C shaking at 175rpm for 16.5 hours. Finally, the colonies on the agar plates were counted - the transformation efficiency was estimated such that, assuming there are approximately 1x10^6 unique barcodes, each unique barcode was represented approximately 20,250 times in the transformed, expanded library pool. The bacterial growth mix (500mL) was collected and a maxi-prep was performed according to the MAXIPREP protocol (QIAGEN), starting at the centrifugation at 6000xg step. Skipping the steps prior to centrifugation is crucial as, because we are dealing with a complex library pool, we avoid any dilution steps which might diminish the library complexity.

Envelope Protein and Packaging Plasmid Expansion

For virus production, three viral plasmid components are necessary: the complex library pool, an envelope protein - pCMV-VSV-G - and a lentiviral packaging plasmid - pCMV-dR8.2 dvpr (pCMV-VSV-G was a gift from Bob Weinberg (Addgene plasmid # 8454 ; http://n2t.net/addgene:8454 ; RRID:Addgene_8454) and pCMV-dR8.2 dvpr was a gift from Bob Weinberg (Addgene plasmid # 8455 ; http://n2t.net/addgene:8455 ; RRID:Addgene_8455)). The envelope protein and packaging plasmid arrived from Addgene as agar stabs. Bacteria were streaked for single colonies on on LB agar + carbenicillin (100µg/mL) plates. A single colony was picked and transfered to 150mL of LB broth with carbenicillin (100µg/mL) in 1L flasks for overnight expansions. The bacterial pellet was harvested the following day and a maxi-prep was performed according to the MAXIPREP protocol (QIAGEN).

2.3.2 Virus Production

Prior to infecting cells with the barcode library, it is necessary to produce the functioning virus by combining the viral components. This was undertaken as follows: a 10cm Tissue Culture Dish (Corning$^{®}$) was coated with 6mL of Poly-L-Lysine (Sigma-Aldrich) to enhance the adherant properties of the dish. This was left for 30 minutes before aspirating the Poly-L-Lysine and allowing the plate to dry for 2 hours. 2.5x10^6 HEK293T cells (ATCC® CRL-3216™) were seeded in 15.5mL of full growth medium. The next day (+24hrs since start), 16.2µL of Lipofectamine™ 2000 transfection reagent (Invitrogen) (pre-warmed to room temperature) was combined with 2.4µg of ClonTracer barcode library, 0.6µg of pCMV-VSV-G envelope protein and 2.4µg of pCMV-dR8.2 dvpr packaging plasmid, with high glucose DMEM (Gibco - Life Technologies) (importantly, *not* supplemented with FBS or PenStrep) up to a total of 600µL. This plasmid mix was left to incubate at room temperature for 20 minutes, following which, it was gently pipetted onto different areas of the 10cm plate containing the HEK293T cells. One day later (+48hrs since start), the medium was replaced with 6mL of fresh full growth medium (now including FBS and PenStrep). Two days later (+96hrs since start) the medium was sampled from the plate and passed through a 0.45µm syringe filter (CORNING) and aliquoted into cryovials before being labelled and stored in the -80°C freezer. This entire process was repeated for a total of 3x 10cm Tissue Culture Dishes (CORNING) to maximise virus production.

2.3.3 Puromycin Selection Optimisation

After infecting cells with the barcode virus, it is necessary to perform a selection step that kills any un-infected cells, leaving only barcoded cells. As the barcode plasmid contains a puromycin resistance gene, this is performed by adding to puromycin to the cells immediately following the infection step. For this step to be successful, two conditions must be met: firstly, the cells must not have reached confluence, otherwise they will stop actively dividing and the puromycin will be unable to kill any cells, independent of barcode status. Secondly, the puromycin concentration used must not be so high that it will kill any cells regardless of whether or not they have a puromycin resistance gene, but it must be high enough to effectively select for successfully barcoded cells. As such, an optimisation step which mediates this trade-off is performed as follows: each cell-line (*HCT116* and *SW620*) were seeded in multiple 6-well adhesive tissue-culture treated plates (CORNING), where each set of 3x plates contained the following number of cells: 0.01x10^6, 0.05x10^6, 0.1x10^6, 0.2x10^6 and 0.5x10^6, for a total of 15x 6-well plates. Each well contained cells diluted in 2mL of full growth medium. Two day later (+48hrs since start), *within* each plate, the following concentrations of puromycin were added to each well in 1mL of full growth medium: 0.0µg/mL, 0.4µg/mL, 0.6µg/mL, 0.8µg/mL, 1.2µg/mL, 2.0µg/mL (these were the *final* concentrations, in the now 3mL of medium/well). Three days later (+120hrs since start) the medium was aspirated from each well. Of the different cell-number replicates (x3 per each number), the number that led to approximately 80% confluency in the negative control well (0.0µg/mL) was chosen. Of these, the cell numbers were counted per well, and the puromycin concentration chosen that led to almost no living cells remaining (0% < cells remaining < 1%).

2.3.4 Virus Infection Optimisation

After the virus is ready for infection, and the number of cells and puromycin concentration have been optimised, it is necessary to determine the volume of viral supernatant to add. As viral infection can be assumed to follow a Poisson distribution (Fehse et al. 2004), adopting a multiplicity of infection of 10% helps mediate the trade-off between maximising the number of uniquely barcoded cells, whilst minimising the number of cells that have more than one barcode. Therefore, the supernatant volume that leads to 10% surviving cells following puromycin selection is estimated as follows: The cell numbers determined in the puromycin selection optimisation step are seeded into 6x 6-well adhesive tissue-culture treated plates (Corning®) in 2mL of full growth medium. One day later (+24hrs since start), the following amounts of virus supernatant are added + 8µg/mL Polybrene (Hexadimethrine Bromide - Sigma-Aldrich) in 1mL of full growth medium (the following volumes are added per well across 2x 6-well plates, bringing the total to 3x replicates per condition): 20µL virus + no puromycin, 5µL, 10µL, 20µL, 50µL, no virus [plate 1], 200uL virus + no puromycin, 100µL, 200µL, 500µL, 1000µL, no virus [plate 2]. (Note that the puromycin has not yet been added, and the range and number of replicates can be reduced when the volume of virus produced is a limiting factor). One day later (+48hrs since start), the growth medium containing the virus is removed, and fresh full growth medium containing the concentration of puromycin determined in the puromycin selection optimisation step (excluding the 'no puromycin' wells) is added. Three days later (+120hrs since start) the growth medium was removed and the cells counted. The volume of virus was chosen that achieved approximately 10% cell survival compared to the 'no puromycin' wells. For the full experiment, infection took place in a 150mm adherant cell culture dish (Corning®) - the optimised volumes and cell-numbers outlined above were therefore increased by a factor of 15 to account for this.

2.4 Long-Term Evolution Experiments

2.4.1 Experiment Set-Up

This stage is universal to all three experiments; OR, PR and QR. To begin the experiment, 3x 150mm adherant cell culture dishes (Corning®) were seeded with the number of cells optimised previously - 3 plates per cell-line (*HCT116* and *SW620*). After one day (+24hrs since start), the optimised virus volume was added to one of the three plates, per cell-line. After another day (+48hrs since start), the optimised concentration of puromycin that leads to 10% m.o.i (scaled up for the 150mm dishes) was added to the plate with virus, and one of the two remaining plates. The third plate (per cell-line) has the cells on this day counted - 10% of this cell count is the estimated number of cells infected with a barcode at the beginning of the experiment. This number was optimised, as discussed above, to ensure that $\sim 10^6$ cells are infected with a barcode. The two remaining 150mm plates were left for the puromycin selection step - the plate with no virus should eventually lead to 100% cell death, as optimised previously. The plate that had the optimised volume of virus supernatant added was left until the puromycin had killed all uninfected cells, and the remaining infected cells had expanded such that the plate was 80-90% confluent. At this point, the cells are harvested and counted. 1x10^6 cells were then seeded into each experiment's respective replicates in T-175 vented flasks (Corning®). Remaining unseeded cells were stored as `P0Ti` samples. In the two drug-pulse experiments, 'PR' and 'QR', the replicates consisted of four control replicates - `C0i` - and four drug-treatment replicates - `DTi`.

2.4.2 Experiment Maintenance

For the experiment 'OR', maintenance simply consisted of either freezing cells immediately prior to processing, or passaging for one or two passages under standard conditions, as outlined above. The two drug-treatment experiments - 'PR' and 'QR' were maintained as follows:

- Following two days, during which cells were allowed to settle and adhere to the bottom the flasks, the standard growth medium was removed and replaced with either:
    - For the drug-treatment (`DTi`) samples: 35mL of full growth medium with XµM of 5-fluorouracil (5-Fu), where XµM corresponds to the previously determined IC_{50} value for each cell-line.
    - For the control (`C0i`) samples: 35mL of standard DMEM with the same volume of DMSO (vehicle control) as the volume of 5-Fu added to the drug-treatment replicates
- Every 3-5 days, the 5-Fu/DMSO containing medium was removed and replaced with standard, full growth medium.
- When a flask was 80-90% confluent, cells were harvested, 1x10^6 cells were seeded into the following Passage and the remaining were frozen for downstream DNA extraction and barcode amplification (see below).
- This process was repeated for 2 ('PR') and 5 ('QR') passages for each experiment. This corresponded to 9-10 weeks for the `HCTbc` and `SW6bc` drug-treatment replicates, respectively, in 'PR', and 26-29 weeks for the `SW6bc` and `HCTbc` drug-treatment replicates, respectively, in 'QR'.

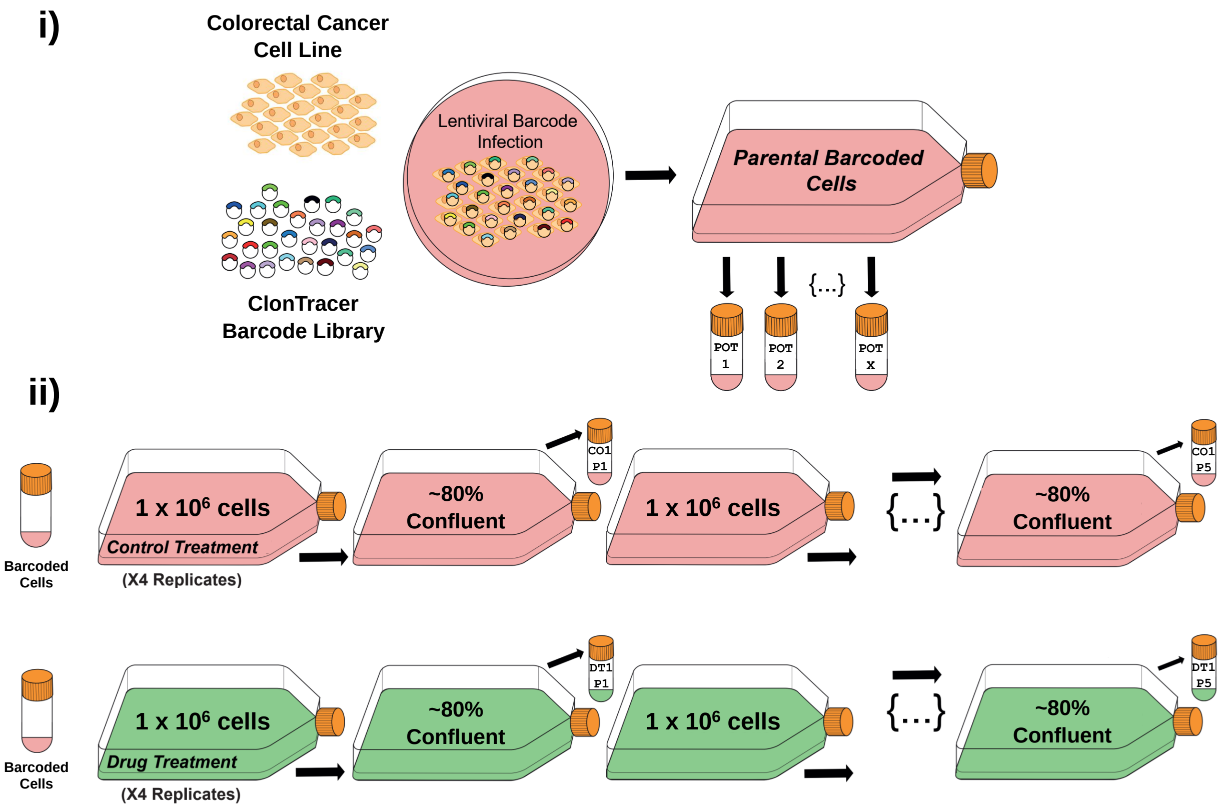


Figure 2.1: A schematic depicting the set-up and maintenance of the long-term evolution experiment during metronomic exposure to chemotherapy. i) Both colorectal cancer cell lines (*HCT116* and *SW620*, in parallel) were infected with the ClonTracer complex plasmid library ($m.o.i \sim 0.1$). ii) Barcoded cells are subject to repeated growth-bottleneck cycles for 4x replicates per control/drug treatment. Control and drug-treatmnets are subject to metronomic exposure to vehicle-control or IC_{50} values of 5-Fu, respectively (not shown). Sample names on cell sample cryovials correspond to those used in the main text.

2.5 Barcode Sequencing

2.5.1 DNA Extraction, Purification and Quantification

To extract DNA for barcode amplification, cells were defrosted and a volume of approximately $1x10^6$ cells was spun down into a pellet and re-suspended in 200µL of PBS. DNA was then extracted according to the DNeasy Blood & Tissue Kit ® (QIAGEN) protocol, following the cultured cell sub-section. I performed a second elution at 36°C to maximise the DNA extraction yield. DNA was eluted in DNase and RNase-Free PCR grade water. To quantify the concentration of DNA, the eluted volume was vortexed and then 1µL taken and measured using a Qubit™ 4 Fluorometer (Invitrogen™) according to the manufacturer's protocol. DNA concentrations were recorded and the DNA frozen at -20°C until future processing. Following PCR steps, I used a magnetic bead based clean-up system for the purification of amplified barcode DNA (CleanNGS - CleanNA). When mentioned in one of the protocols, the purification beads are used as follows: the beads are removed from the fridge and allowed to reach room temperature approximately 30mins prior to purification. A chosen ratio of beads:product is added (according to the desired size of purified fragments kept) and mixed well with the product. The volume is incubated at room temperature for 5 minutes. The sample is placed on a magnetic stand and the beads are separated from the supernatant for approximately 5 minutes. The supernatant is discarded, and the beads washed with 80% Ethanol for a total of two washes. All ethanol is removed and the DNA is eluted from the beads in DNase and RNase-Free PCR grade water and incubated at room temperature for approximately 5 minutes. Finally, the beads are separated from the DNA solution on the magnetic stand for a further 5 minutes, the DNA is serrated and stored for future use whilst the beads are discarded.

2.5.2 Barcode Amplification and Adapter Integration

Due to technical difficulties encountered with certain protocols and sequencing techniques, different protocols were adopted for each sequencing experiment.

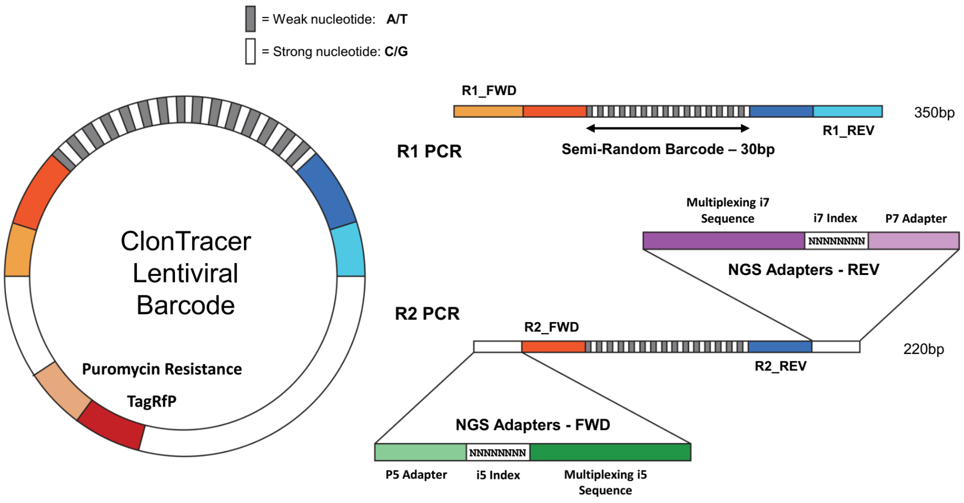
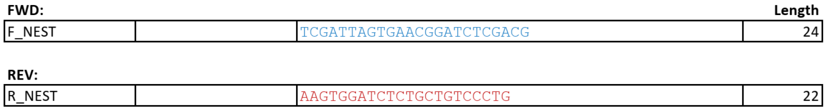


Figure 2.2: The structure of the ClonTracer lentivirus barcode construct. The regions targeted for round one and two of PCR are highlighted, as are the structure of the respective PCR products (RHS). Note the PCR targets shown correspond to the 'OR' and 'PR' sequencing experiments, whilst 'QR' targeted the 'R2' 'FWD' and 'REV' sequences directly.

OR

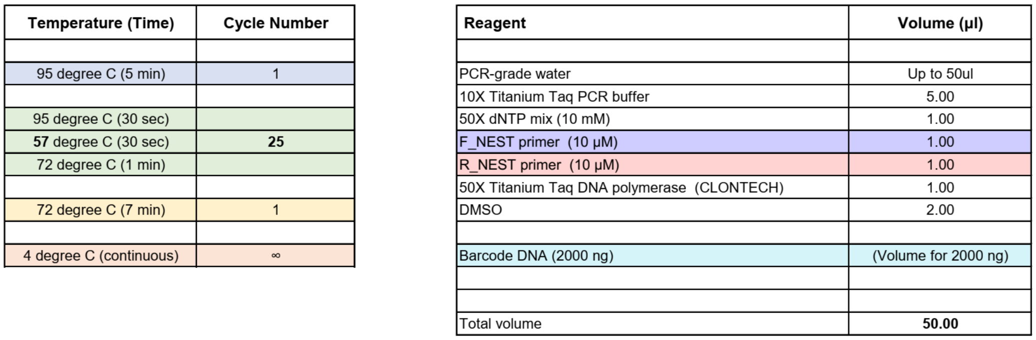
For the 'OR' sequencing experiment, due to secondary structures forming during the original ClonTracer protocol (Bhang et al. 2015), I adopted a nested PCR approach for the barcode amplification step. The first round PCR uses two universal forward and reverse primers to amplify the barcode sequence from cell gDNA (Table 2.1).

After amplifying the barcode sequences from cell gDNA using the PCR protocol outlined in Table 2.2, the round 1 product was diluted and used as input for a second round PCR. This PCR used primers that had Illumina™ index sequences already incorporated. This meant a subsequent library-preparation step could be skipped. Due to the limited number of samples being sequenced, indexes were only included in the reverse primer (Table 2.4). These forward and reverse index primers were used with the PCR protocol outlined in Table 2.3.



FWD:			Length
F_NEST		TCGATTAGTGAACGGATCTCGACG	24
REV:			
R_NEST		AAGTGGATCTCTGCTGTCCCTG	22

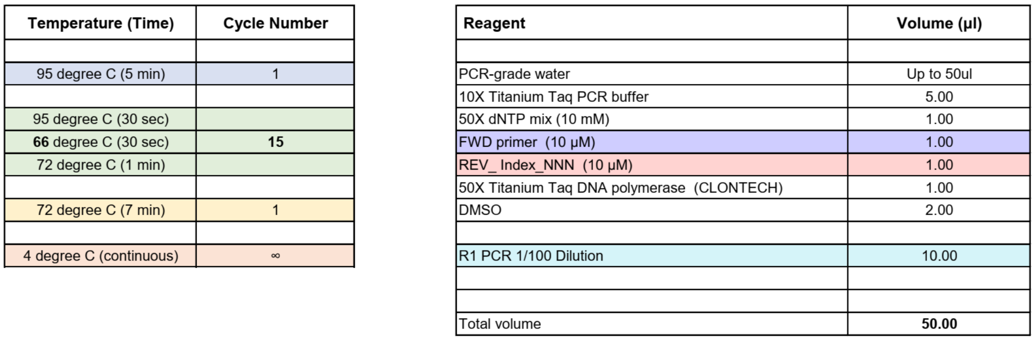
Table 2.1: Forward and Reverse round 1 amplification primers for the ‘OR’ sequencing experiment’s nested PCR protocol.



Temperature (Time)	Cycle Number
95 degree C (5 min)	1
95 degree C (30 sec)	
57 degree C (30 sec)	**25**
72 degree C (1 min)	
72 degree C (7 min)	1
4 degree C (continuous)	∞

Reagent	Volume (μl)
PCR-grade water	Up to 50ul
10X Titanium Taq PCR buffer	5.00
50X dNTP mix (10 mM)	1.00
F_NEST primer (10 μM)	1.00
R_NEST primer (10 μM)	1.00
50X Titanium Taq DNA polymerase (CLONTECH)	1.00
DMSO	2.00
Barcode DNA (2000 ng)	(Volume for 2000 ng)
Total volume	**50.00**

Table 2.2: PCR protocol used for the round 1 nested PCR with ‘OR’ sequencing experiment.



Temperature (Time)	Cycle Number
95 degree C (5 min)	1
95 degree C (30 sec)	
66 degree C (30 sec)	**15**
72 degree C (1 min)	
72 degree C (7 min)	1
4 degree C (continuous)	∞

Reagent	Volume (μl)
PCR-grade water	Up to 50ul
10X Titanium Taq PCR buffer	5.00
50X dNTP mix (10 mM)	1.00
FWD primer (10 μM)	1.00
REV_ Index_NNN (10 μM)	1.00
50X Titanium Taq DNA polymerase (CLONTECH)	1.00
DMSO	2.00
R1 PCR 1/100 Dilution	10.00
Total volume	**50.00**

Table 2.3: PCR protocol used for the round 2 nested PCR with the ‘OR’ sequencing experiment.

FWD:	Index:		Design:	Length
FWD	NA		AATGATACGGCGACCACCGAGATCTACACACTGACTGCAGTCTGAGTCTGACAG	54

P5 Illumina Adapter Fwd Amplification Sequence

Design:

CAAGCAGAAGACGGCATACGAGATNNNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT

P7 Illumina Adapter Index Multiplexing Read i7 Primer Rev Amplification Sequence

REV:	Index:	Rev_Comp_Index:		
Rev_Index_001	ACGATCGTGA	TCACGATCGT	CAAGCAGAAGACGGCATACGAGATACGATCGTGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_002	CTAGATCGTG	CACGATCTAG	CAAGCAGAAGACGGCATACGAGATCTAGATCGTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_003	GACTCGATCA	TGATCGAGTC	CAAGCAGAAGACGGCATACGAGATGACTCGATCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_004	TGACTAGCTC	GAGCTAGTCA	CAAGCAGAAGACGGCATACGAGATTGACTAGCTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_005	ATGCTCAGCA	TGCTGAGCAT	CAAGCAGAAGACGGCATACGAGATATGCTCAGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_006	CGATCTGCAT	ATGCAGATCG	CAAGCAGAAGACGGCATACGAGATCGATCTGCATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_007	GATAGCTGAC	GTCAGCTATC	CAAGCAGAAGACGGCATACGAGATGATAGCTGACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_008	TCAGCTACGT	ACGTAGCTGA	CAAGCAGAAGACGGCATACGAGATTCAGCTACGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_009	AGTACGCATG	CATGCGTACT	CAAGCAGAAGACGGCATACGAGATAGTACGCATGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_010	CACGTCGATA	TATCGACGTG	CAAGCAGAAGACGGCATACGAGATCACGTCGATAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_011	GTATCACGAC	GTCGTGATAC	CAAGCAGAAGACGGCATACGAGATGTATCACGACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_012	TCGCAGTACT	AGTACTGCGA	CAAGCAGAAGACGGCATACGAGATTCGCAGTACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_013	AGCGTCTGAT	ATCAGACGCT	CAAGCAGAAGACGGCATACGAGATAGCGTCTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_014	CAGCATGTCT	AGACATGCTG	CAAGCAGAAGACGGCATACGAGATCAGCATGTCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_015	GTACTCATCG	CGATGAGTAC	CAAGCAGAAGACGGCATACGAGATGTACTCATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_016	TCTGCAGCTA	TAGCTGCAGA	CAAGCAGAAGACGGCATACGAGATTCTGCAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_017	ACTGTACTCG	CGAGTACAGT	CAAGCAGAAGACGGCATACGAGATACTGTACTCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_018	CGACAGCTAT	ATAGCTGTCG	CAAGCAGAAGACGGCATACGAGATCGACAGCTATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_019	GTCATGCGTA	TACGCATGAC	CAAGCAGAAGACGGCATACGAGATGTCATGCGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_020	TAGTCGCATG	CATGCGACTA	CAAGCAGAAGACGGCATACGAGATTAGTCGCATGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_021	ATCGATGACG	CGTCATCGAT	CAAGCAGAAGACGGCATACGAGATATCGATGACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_022	CGATAGTCGT	ACGACTATCG	CAAGCAGAAGACGGCATACGAGATCGATAGTCGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_023	GAGCTGTATC	GATACAGCTC	CAAGCAGAAGACGGCATACGAGATGAGCTGTATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_024	TCTGATCGCA	TGCGATCAGA	CAAGCAGAAGACGGCATACGAGATTCTGATCGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_025	AGCATCGTCT	AGACGATGCT	CAAGCAGAAGACGGCATACGAGATAGCATCGTCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_026	CTACGTCTAG	CTAGACGTAG	CAAGCAGAAGACGGCATACGAGATCTACGTCTAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_027	GCTAGATGCT	AGCATCTAGC	CAAGCAGAAGACGGCATACGAGATGCTAGATGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_028	TCGAGTGCAT	ATGCACTCGA	CAAGCAGAAGACGGCATACGAGATTCGAGTGCATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_029	ACGCTGACAT	ATGTCAGCGT	CAAGCAGAAGACGGCATACGAGATACGCTGACATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_030	CATACAGTGC	GCACTGTATG	CAAGCAGAAGACGGCATACGAGATCATACAGTGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_031	GAGCACTAGT	ACTAGTGCTC	CAAGCAGAAGACGGCATACGAGATGAGCACTAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_032	TGCATGTAGC	GCTACATGCA	CAAGCAGAAGACGGCATACGAGATTGCATGTAGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_033	AGTGATCGAC	GTCGATCACT	CAAGCAGAAGACGGCATACGAGATAGTGATCGACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_034	CTGACATGCA	TGCATGTCAG	CAAGCAGAAGACGGCATACGAGATCTGACATGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_035	GTAGCAGATC	GATCTGCTAC	CAAGCAGAAGACGGCATACGAGATGTAGCAGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_036	TCACTATGCG	CGCATAGTGA	CAAGCAGAAGACGGCATACGAGATTCACTATGCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_037	ACTCGATACG	CGTATCGAGT	CAAGCAGAAGACGGCATACGAGATACTCGATACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_038	CGCATGATCA	TGATCATGCG	CAAGCAGAAGACGGCATACGAGATCGCATGATCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_039	GCAGATCACT	AGTGATCTGC	CAAGCAGAAGACGGCATACGAGATGCAGATCACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_040	TCGACTAGTG	CACTAGTCGA	CAAGCAGAAGACGGCATACGAGATTCGACTAGTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_041	ATCAGCGATG	CATCGCTGAT	CAAGCAGAAGACGGCATACGAGATATCAGCGATGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_042	CTGTATGAGC	GCTCATACAG	CAAGCAGAAGACGGCATACGAGATCTGTATGAGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_043	GTGACTGTCA	TGACAGTCAC	CAAGCAGAAGACGGCATACGAGATGTGACTGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_044	TACGCTGCAT	ATGCAGCGTA	CAAGCAGAAGACGGCATACGAGATTACGCTGCATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_045	AGCTGATGCA	TGCATCAGCT	CAAGCAGAAGACGGCATACGAGATAGCTGATGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_046	CTATGCACTG	CAGTGCATAG	CAAGCAGAAGACGGCATACGAGATCTATGCACTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_047	GCTCATGTCA	TGACATGAGC	CAAGCAGAAGACGGCATACGAGATGCTCATGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_048	TAGCGATCTG	CAGATCGCTA	CAAGCAGAAGACGGCATACGAGATTAGCGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_049	ACGTACTGCT	AGCAGTACGT	CAAGCAGAAGACGGCATACGAGATACGTACTGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_050	CATAGCATCG	CGATGCTATG	CAAGCAGAAGACGGCATACGAGATCATAGCATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94

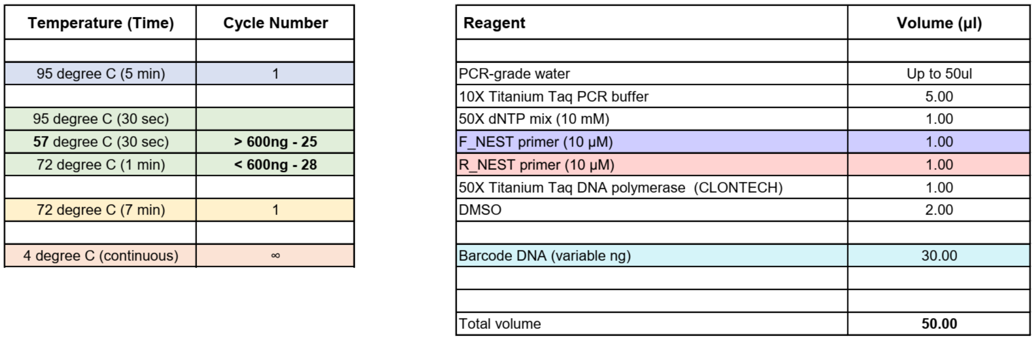
Table 2.4: Forward and Reverse round 2 amplification primers for 'OR' nested PCR protocol, with different components of each primer highlighted.

PR

For the ‘PR’ experiment, I again adopted a nested PCR protocol approach. However, due to the larger number of samples that needed to be sequenced simultaneously, the round 2 primers were adapted so that the forward primers now also contained a unique multiplexing sequencing (Table 2.7). This meant that it was now possible to sequence up to 400 different samples at once (8 fwd * 50 rev indexes) if necessary.

QR

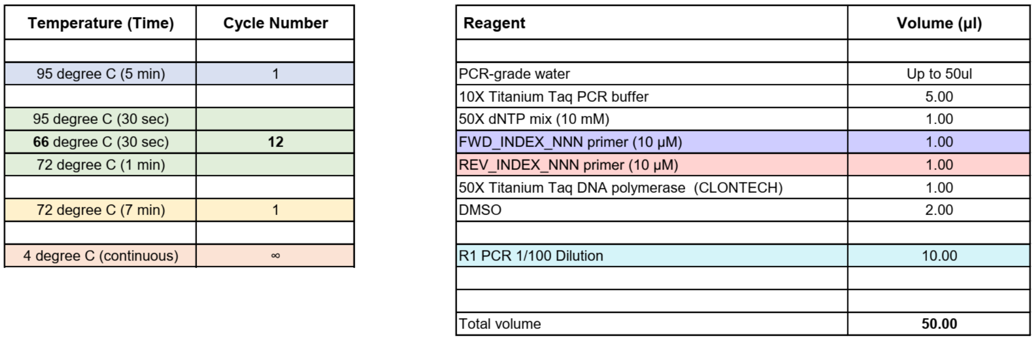
Finally, due to ‘index-hopping’ issues encountered when using the non-redundant combinations of forward and reverse index primers (see Results Chapter 3 for a more detailed description), a different approach was adopted for the third sequencing experiment, ‘QR’. Now, two universal amplification primers targeted the universal sequences flanking the semi-random barcode sequence in the ClonTracer construct (Bhang et al. 2015) directly (Table 2.9) using the PCR protcol outlined in Table 2.9. Following this, a bespoke adapter ligation protocol (Table 2.10) was optimised that used unique dual-indexes (IDT Illumina™ TruSeq UD Indexes 96) - this meant that each sample was now identified via its own unique forward *and* reverse indexes (Table 2.11).



Temperature (Time)	Cycle Number
95 degree C (5 min)	1
95 degree C (30 sec)	
57 degree C (30 sec)	**> 600ng - 25**
72 degree C (1 min)	**< 600ng - 28**
72 degree C (7 min)	1
4 degree C (continuous)	∞

Reagent	Volume (µl)
PCR-grade water	Up to 50ul
10X Titanium Taq PCR buffer	5.00
50X dNTP mix (10 mM)	1.00
F_NEST primer (10 µM)	1.00
R_NEST primer (10 µM)	1.00
50X Titanium Taq DNA polymerase (CLONTECH)	1.00
DMSO	2.00
Barcode DNA (variable ng)	30.00
Total volume	**50.00**

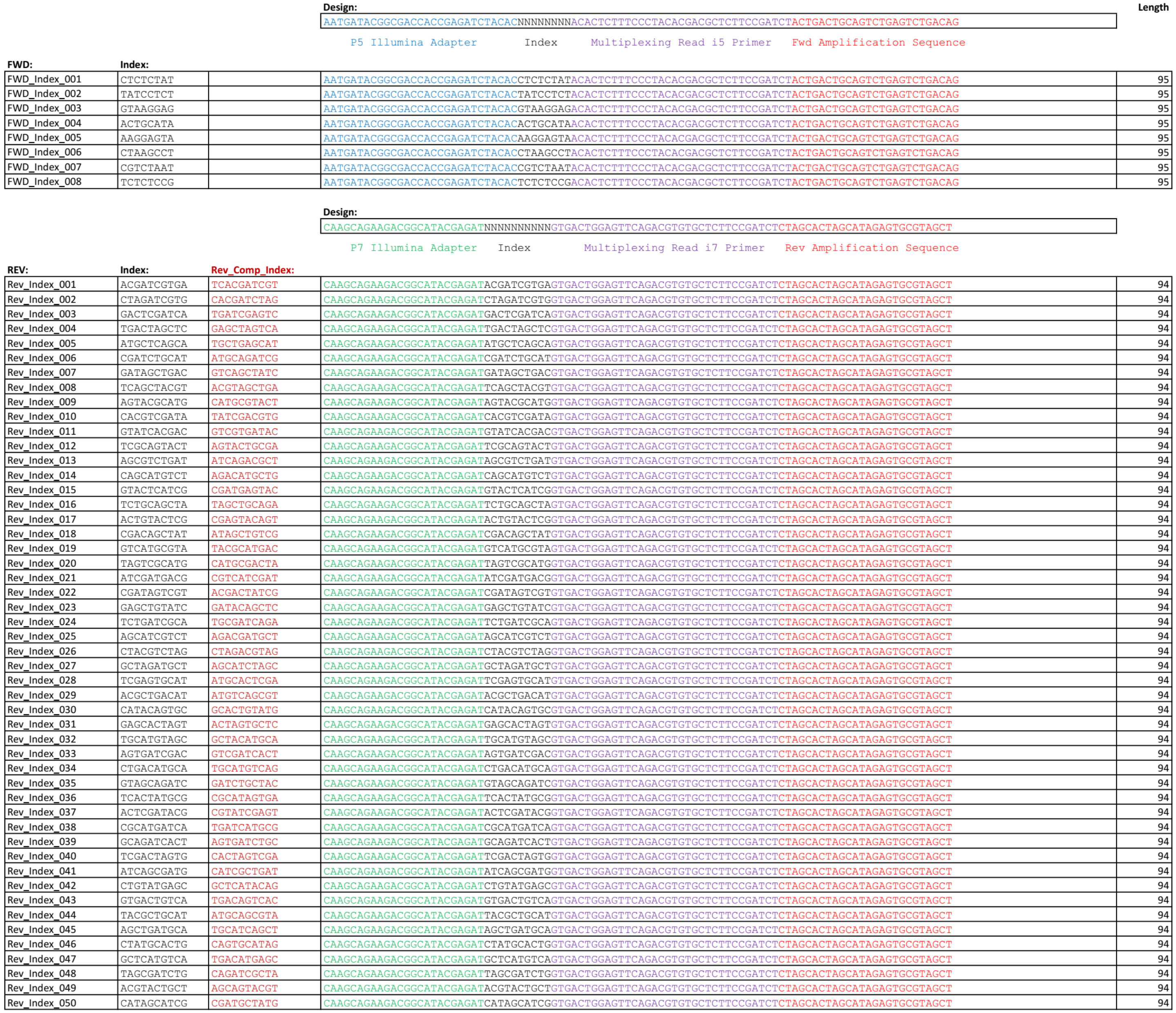
Table 2.5: PCR protocol used for the round 1 nested PCR with 'PR' sequencing experiment.



Temperature (Time)	Cycle Number
95 degree C (5 min)	1
95 degree C (30 sec)	
66 degree C (30 sec)	**12**
72 degree C (1 min)	
72 degree C (7 min)	1
4 degree C (continuous)	∞

Reagent	Volume (µl)
PCR-grade water	Up to 50ul
10X Titanium Taq PCR buffer	5.00
50X dNTP mix (10 mM)	1.00
FWD_INDEX_NNN primer (10 µM)	1.00
REV_INDEX_NNN primer (10 µM)	1.00
50X Titanium Taq DNA polymerase (CLONTECH)	1.00
DMSO	2.00
R1 PCR 1/100 Dilution	10.00
Total volume	**50.00**

Table 2.6: PCR protocol used for the round 2 nested PCR with the 'PR' sequencing experiment.



Design:

AATGATACGGCGACCACCGAGATCTACACNNNNNNNNACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGACTGCAGTCTGAGTCTGACAG

P5 Illumina Adapter   Index   Multiplexing Read i5 Primer   Fwd Amplification Sequence

FWD:	Index:			Length
FWD_Index_001	CTCTCTAT		AATGATACGGCGACCACCGAGATCTACACCTCTCTATACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGACTGCAGTCTGAGTCTGACAG	95
FWD_Index_002	TATCCTCT		AATGATACGGCGACCACCGAGATCTACACTATCCTCTACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGACTGCAGTCTGAGTCTGACAG	95
FWD_Index_003	GTAAGGAG		AATGATACGGCGACCACCGAGATCTACACGTAAGGAGACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGACTGCAGTCTGAGTCTGACAG	95
FWD_Index_004	ACTGCATA		AATGATACGGCGACCACCGAGATCTACACACTGCATAACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGACTGCAGTCTGAGTCTGACAG	95
FWD_Index_005	AAGGAGTA		AATGATACGGCGACCACCGAGATCTACACAAGGAGTAACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGACTGCAGTCTGAGTCTGACAG	95
FWD_Index_006	CTAAGCCT		AATGATACGGCGACCACCGAGATCTACACCTAAGCCTACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGACTGCAGTCTGAGTCTGACAG	95
FWD_Index_007	CGTCTAAT		AATGATACGGCGACCACCGAGATCTACACCGTCTAATACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGACTGCAGTCTGAGTCTGACAG	95
FWD_Index_008	TCTCTCCG		AATGATACGGCGACCACCGAGATCTACACTCTCTCCGACACTCTTTCCCTACACGACGCTCTTCCGATCTACTGACTGCAGTCTGAGTCTGACAG	95

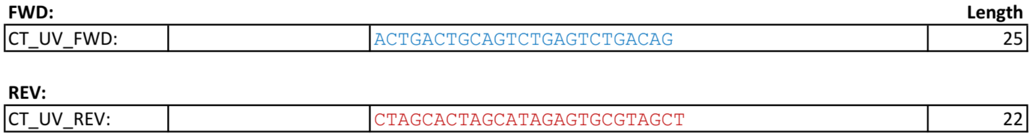
Design:

CAAGCAGAAGACGGCATACGAGATNNNNNNNNNNGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT

P7 Illumina Adapter   Index   Multiplexing Read i7 Primer   Rev Amplification Sequence

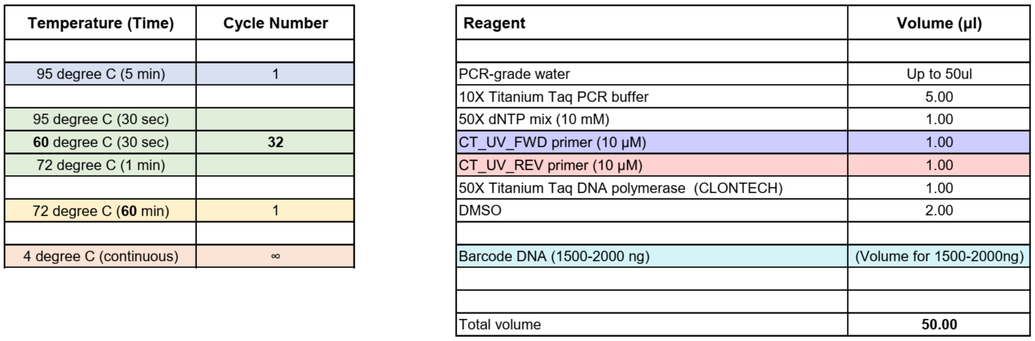
REV:	Index:	Rev_Comp_Index:		
Rev_Index_001	ACGATCGTGA	TCACGATCGT	CAAGCAGAAGACGGCATACGAGATACGATCGTGAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_002	CTAGATCGTG	CACGATCTAG	CAAGCAGAAGACGGCATACGAGATCTAGATCGTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_003	GACTCGATCA	TGATCGAGTC	CAAGCAGAAGACGGCATACGAGATGACTCGATCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_004	TGACTAGCTC	GAGCTAGTCA	CAAGCAGAAGACGGCATACGAGATTGACTAGCTCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_005	ATGCTCAGCA	TGCTGAGCAT	CAAGCAGAAGACGGCATACGAGATATGCTCAGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_006	CGATCTGCAT	ATGCAGATCG	CAAGCAGAAGACGGCATACGAGATCGATCTGCATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_007	GATAGCTGAC	GTCAGCTATC	CAAGCAGAAGACGGCATACGAGATGATAGCTGACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_008	TCAGCTACGT	ACGTAGCTGA	CAAGCAGAAGACGGCATACGAGATTCAGCTACGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_009	AGTACGCATG	CATGCGTACT	CAAGCAGAAGACGGCATACGAGATAGTACGCATGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_010	CACGTCGATA	TATCGACGTG	CAAGCAGAAGACGGCATACGAGATCACGTCGATAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_011	GTATCACGAC	GTCGTGATAC	CAAGCAGAAGACGGCATACGAGATGTATCACGACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_012	TCGCAGTACT	AGTACTGCGA	CAAGCAGAAGACGGCATACGAGATTCGCAGTACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_013	AGCGTCTGAT	ATCAGACGCT	CAAGCAGAAGACGGCATACGAGATAGCGTCTGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_014	CAGCATGTCT	AGACATGCTG	CAAGCAGAAGACGGCATACGAGATCAGCATGTCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_015	GTACTCATCG	CGATGAGTAC	CAAGCAGAAGACGGCATACGAGATGTACTCATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_016	TCTGCAGCTA	TAGCTGCAGA	CAAGCAGAAGACGGCATACGAGATTCTGCAGCTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_017	ACTGTACTCG	CGAGTACAGT	CAAGCAGAAGACGGCATACGAGATACTGTACTCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_018	CGACAGCTAT	ATAGCTGTCG	CAAGCAGAAGACGGCATACGAGATCGACAGCTATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_019	GTCATGCGTA	TACGCATGAC	CAAGCAGAAGACGGCATACGAGATGTCATGCGTAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_020	TAGTCGCATG	CATGCGACTA	CAAGCAGAAGACGGCATACGAGATTAGTCGCATGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_021	ATCGATGACG	CGTCATCGAT	CAAGCAGAAGACGGCATACGAGATATCGATGACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_022	CGATAGTCGT	ACGACTATCG	CAAGCAGAAGACGGCATACGAGATCGATAGTCGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_023	GAGCTGTATC	GATACAGCTC	CAAGCAGAAGACGGCATACGAGATGAGCTGTATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_024	TCTGATCGCA	TGCGATCAGA	CAAGCAGAAGACGGCATACGAGATTCTGATCGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_025	AGCATCGTCT	AGACGATGCT	CAAGCAGAAGACGGCATACGAGATAGCATCGTCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_026	CTACGTCTAG	CTAGACGTAG	CAAGCAGAAGACGGCATACGAGATCTACGTCTAGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_027	GCTAGATGCT	AGCATCTAGC	CAAGCAGAAGACGGCATACGAGATGCTAGATGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_028	TCGAGTGCAT	ATGCACTCGA	CAAGCAGAAGACGGCATACGAGATTCGAGTGCATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_029	ACGCTGACAT	ATGTCAGCGT	CAAGCAGAAGACGGCATACGAGATACGCTGACATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_030	CATACAGTGC	GCACTGTATG	CAAGCAGAAGACGGCATACGAGATCATACAGTGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_031	GAGCACTAGT	ACTAGTGCTC	CAAGCAGAAGACGGCATACGAGATGAGCACTAGTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_032	TGCATGTAGC	GCTACATGCA	CAAGCAGAAGACGGCATACGAGATTGCATGTAGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_033	AGTGATCGAC	GTCGATCACT	CAAGCAGAAGACGGCATACGAGATAGTGATCGACGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_034	CTGACATGCA	TGCATGTCAG	CAAGCAGAAGACGGCATACGAGATCTGACATGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_035	GTAGCAGATC	GATCTGCTAC	CAAGCAGAAGACGGCATACGAGATGTAGCAGATCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_036	TCACTATGCG	CGCATAGTGA	CAAGCAGAAGACGGCATACGAGATTCACTATGCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_037	ACTCGATACG	CGTATCGAGT	CAAGCAGAAGACGGCATACGAGATACTCGATACGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_038	CGCATGATCA	TGATCATGCG	CAAGCAGAAGACGGCATACGAGATCGCATGATCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_039	GCAGATCACT	AGTGATCTGC	CAAGCAGAAGACGGCATACGAGATGCAGATCACTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_040	TCGACTAGTG	CACTAGTCGA	CAAGCAGAAGACGGCATACGAGATTCGACTAGTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_041	ATCAGCGATG	CATCGCTGAT	CAAGCAGAAGACGGCATACGAGATATCAGCGATGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_042	CTGTATGAGC	GCTCATACAG	CAAGCAGAAGACGGCATACGAGATCTGTATGAGCGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_043	GTGACTGTCA	TGACAGTCAC	CAAGCAGAAGACGGCATACGAGATGTGACTGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_044	TACGCTGCAT	ATGCAGCGTA	CAAGCAGAAGACGGCATACGAGATTACGCTGCATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_045	AGCTGATGCA	TGCATCAGCT	CAAGCAGAAGACGGCATACGAGATAGCTGATGCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_046	CTATGCACTG	CAGTGCATAG	CAAGCAGAAGACGGCATACGAGATCTATGCACTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_047	GCTCATGTCA	TGACATGAGC	CAAGCAGAAGACGGCATACGAGATGCTCATGTCAGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_048	TAGCGATCTG	CAGATCGCTA	CAAGCAGAAGACGGCATACGAGATTAGCGATCTGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_049	ACGTACTGCT	AGCAGTACGT	CAAGCAGAAGACGGCATACGAGATACGTACTGCTGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94
Rev_Index_050	CATAGCATCG	CGATGCTATG	CAAGCAGAAGACGGCATACGAGATCATAGCATCGGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT	94

Table 2.7: Forward and Reverse round 2 amplification primers for ‘PR’ nested PCR protocol, with different components of each primer highlighted.



FWD:			Length
CT_UV_FWD:		ACTGACTGCAGTCTGAGTCTGACAG	25
REV:			
CT_UV_REV:		CTAGCACTAGCATAGAGTGCGTAGCT	22

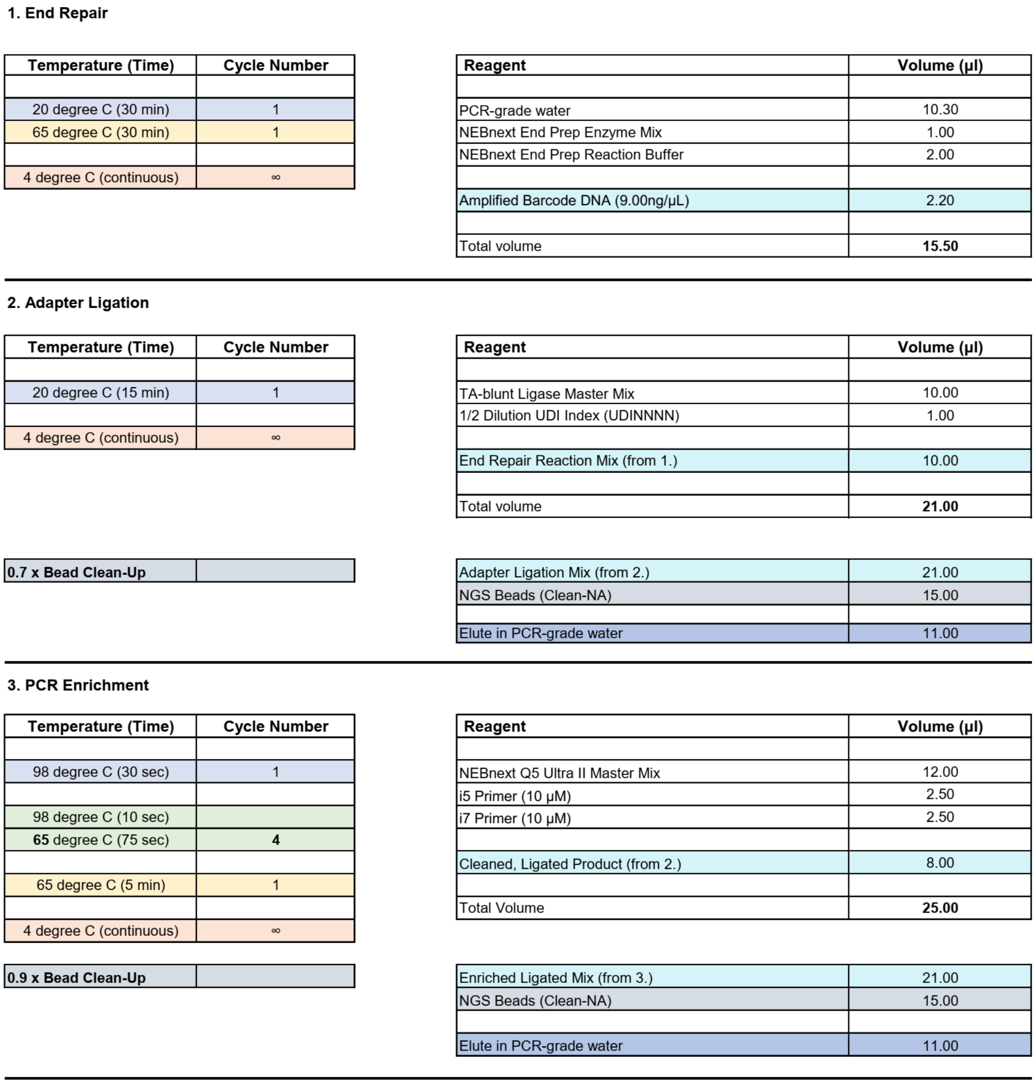
Table 2.8: PCR primers used for the barcode amplification PCR with the 'QR' sequencing experiment.



Temperature (Time)	Cycle Number
95 degree C (5 min)	1
95 degree C (30 sec)	
60 degree C (30 sec)	**32**
72 degree C (1 min)	
72 degree C (**60** min)	1
4 degree C (continuous)	∞

Reagent	Volume (µl)
PCR-grade water	Up to 50ul
10X Titanium Taq PCR buffer	5.00
50X dNTP mix (10 mM)	1.00
CT_UV_FWD primer (10 µM)	1.00
CT_UV_REV primer (10 µM)	1.00
50X Titanium Taq DNA polymerase (CLONTECH)	1.00
DMSO	2.00
Barcode DNA (1500-2000 ng)	(Volume for 1500-2000ng)
Total volume	**50.00**

Table 2.9: PCR protocol used for the barcode amplification PCR with the 'QR' sequencing experiment.



1. End Repair

Temperature (Time)	Cycle Number
20 degree C (30 min)	1
65 degree C (30 min)	1
4 degree C (continuous)	∞

Reagent	Volume (µl)
PCR-grade water	10.30
NEBnext End Prep Enzyme Mix	1.00
NEBnext End Prep Reaction Buffer	2.00
Amplified Barcode DNA (9.00ng/µL)	2.20
Total volume	**15.50**

2. Adapter Ligation

Temperature (Time)	Cycle Number
20 degree C (15 min)	1
4 degree C (continuous)	∞

Reagent	Volume (µl)
TA-blunt Ligase Master Mix	10.00
1/2 Dilution UDI Index (UDINNNN)	1.00
End Repair Reaction Mix (from 1.)	10.00
Total volume	**21.00**

0.7 x Bead Clean-Up	

Adapter Ligation Mix (from 2.)	21.00
NGS Beads (Clean-NA)	15.00
Elute in PCR-grade water	11.00

3. PCR Enrichment

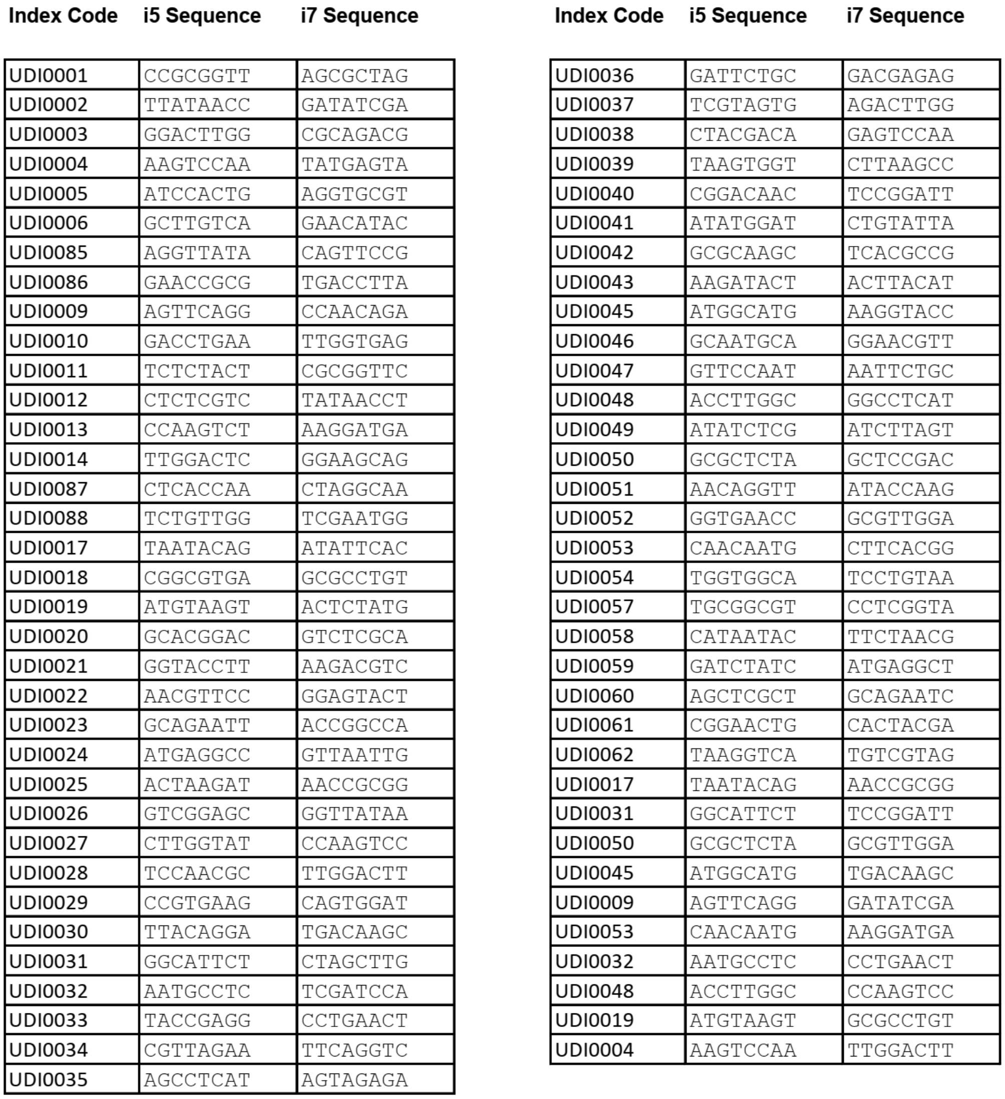
Temperature (Time)	Cycle Number
98 degree C (30 sec)	1
98 degree C (10 sec)	
65 degree C (75 sec)	**4**
65 degree C (5 min)	1
4 degree C (continuous)	∞

Reagent	Volume (µl)
NEBnext Q5 Ultra II Master Mix	12.00
i5 Primer (10 µM)	2.50
i7 Primer (10 µM)	2.50
Cleaned, Ligated Product (from 2.)	8.00
Total Volume	**25.00**

0.9 x Bead Clean-Up	

Enriched Ligated Mix (from 3.)	21.00
NGS Beads (Clean-NA)	15.00
Elute in PCR-grade water	11.00

Table 2.10: A bespoke adapter ligation protocol developed for unique dual-indexes (UDI) incorporation with the 'QR' sequencing experiment.



Index Code	i5 Sequence	i7 Sequence
UDI0001	CCGCGGTT	AGCGCTAG
UDI0002	TTATAACC	GATATCGA
UDI0003	GGACTTGG	CGCAGACG
UDI0004	AAGTCCAA	TATGAGTA
UDI0005	ATCCACTG	AGGTGCGT
UDI0006	GCTTGTCA	GAACATAC
UDI0085	AGGTTATA	CAGTTCCG
UDI0086	GAACCGCG	TGACCTTA
UDI0009	AGTTCAGG	CCAACAGA
UDI0010	GACCTGAA	TTGGTGAG
UDI0011	TCTCTACT	CGCGGTTC
UDI0012	CTCTCGTC	TATAACCT
UDI0013	CCAAGTCT	AAGGATGA
UDI0014	TTGGACTC	GGAAGCAG
UDI0087	CTCACCAA	CTAGGCAA
UDI0088	TCTGTTGG	TCGAATGG
UDI0017	TAATACAG	ATATTCAC
UDI0018	CGGCGTGA	GCGCCTGT
UDI0019	ATGTAAGT	ACTCTATG
UDI0020	GCACGGAC	GTCTCGCA
UDI0021	GGTACCTT	AAGACGTC
UDI0022	AACGTTCC	GGAGTACT
UDI0023	GCAGAATT	ACCGGCCA
UDI0024	ATGAGGCC	GTTAATTG
UDI0025	ACTAAGAT	AACCGCGG
UDI0026	GTCGGAGC	GGTTATAA
UDI0027	CTTGGTAT	CCAAGTCC
UDI0028	TCCAACGC	TTGGACTT
UDI0029	CCGTGAAG	CAGTGGAT
UDI0030	TTACAGGA	TGACAAGC
UDI0031	GGCATTCT	CTAGCTTG
UDI0032	AATGCCTC	TCGATCCA
UDI0033	TACCGAGG	CCTGAACT
UDI0034	CGTTAGAA	TTCAGGTC
UDI0035	AGCCTCAT	AGTAGAGA
UDI0036	GATTCTGC	GACGAGAG
UDI0037	TCGTAGTG	AGACTTGG
UDI0038	CTACGACA	GAGTCCAA
UDI0039	TAAGTGGT	CTTAAGCC
UDI0040	CGGACAAC	TCCGGATT
UDI0041	ATATGGAT	CTGTATTA
UDI0042	GCGCAAGC	TCACGCCG
UDI0043	AAGATACT	ACTTACAT
UDI0045	ATGGCATG	AAGGTACC
UDI0046	GCAATGCA	GGAACGTT
UDI0047	GTTCCAAT	AATTCTGC
UDI0048	ACCTTGGC	GGCCTCAT
UDI0049	ATATCTCG	ATCTTAGT
UDI0050	GCGCTCTA	GCTCCGAC
UDI0051	AACAGGTT	ATACCAAG
UDI0052	GGTGAACC	GCGTTGGA
UDI0053	CAACAATG	CTTCACGG
UDI0054	TGGTGGCA	TCCTGTAA
UDI0057	TGCGGCGT	CCTCGGTA
UDI0058	CATAATAC	TTCTAACG
UDI0059	GATCTATC	ATGAGGCT
UDI0060	AGCTCGCT	GCAGAATC
UDI0061	CGGAACTG	CACTACGA
UDI0062	TAAGGTCA	TGTCGTAG
UDI0017	TAATACAG	AACCGCGG
UDI0031	GGCATTCT	TCCGGATT
UDI0050	GCGCTCTA	GCGTTGGA
UDI0045	ATGGCATG	TGACAAGC
UDI0009	AGTTCAGG	GATATCGA
UDI0053	CAACAATG	AAGGATGA
UDI0032	AATGCCTC	CCTGAACT
UDI0048	ACCTTGGC	CCAAGTCC
UDI0019	ATGTAAGT	GCGCCTGT
UDI0004	AAGTCCAA	TTGGACTT

Table 2.11: Unique dual-indexes (IDT Illumina™ TruSeq UD Indexes 96) used with the 'QR' sequencing experiment.

2.5.3 Sequencing Library Concentration Quantification

For the ‘OR’ test experiment, the library concentration was estimated using a TapeStation (Agilent); samples were prepared and measured according to the manufacturer’s instructions. For experimental runs ‘PR’ and ‘QR’, the proportion of the cleaned-up library that was readily seqeunceable on an Illumina™ flow-cell, library concentrations were quantified using a KAPA Library Quantification Kit (Roche) according to the manufacturer’s instructions. Two different dilutions an order in magnitude apart were prepared for each library (1x library for ‘PR’, and 2x libraries for ‘QR’). Three replicates were processed in parallel for each library sample. Following performing qPCR with the library samples alongside the KAPA DNA standards, standard curves were used to convert Cq scores into average concentrations (pM), the average size-adjusted concentration of each library sample was calculated, averaging across all three replicates and each library dilution.

2.5.4 Sequencing Platform Specifications

OR

The ‘OR’ test experiment was sequenced on a NextSeq 500 High-Output flow cells (x4), with 75-cycles of single-end sequencing. Due to the low-diversity nature of the amplicon library, the amount of PhiX control library used was increased to 20%. Standard Illumina™ sequencing primers were used.

PR

The ‘PR’ long-term drug-treatment experiment was sequenced on a NovaSeq 6000 flow cell (x1), with 50-cycles of paired-end sequencing. Again, due to the low-diversity nature of the amplicon library, 20% of Phix control library was used. The NovaSeq reads the read 1 index directly of the flow-cell. However, due to the custom primers used for joint barcode amplification and sequencing primer incorporation (Table 2.7), a custom sequencing primer was used for read 2 (sequence:
`ACGTGTGCTCTTCCGATCTCTAGCACTAGCATAGAGTGCGTAGCT`).

QR

The ‘QR’ long-term drug-treatment experiment was sequenced on NovaSeq 6000 flow cells (x2), with 50-cycles of paired-end sequencing. Due to the low-diversity amplicon library, 15% of Phix control library was used. As - due to index-hopping issues experienced in ‘PR’ - we used UDIs and a standard NEBnext Ultra II library preparation, no custom sequencing primers were necessary.

2.6 Code and Software

Simulations were written in the language Julia (`https://julialang.org/`) whilst sequencing and simulation outputs and data visualisation was performed in the language R (`https://www.r-project.org/`). The code used to run the agent-based simulations can be found at the following GitHub repository:
`https://github.com/freddie090/Cancer_BarCode_Sim_CBC`

Chapter 3

Results Chapter 1 - Modelling the Evolutionary Dynamics of Drug Resistance

3.1 Summary

In this section, I develop models of resistance evolution. My experimental design permits time-series observations of uniquely labelled lineages during chemotherapy treatment *in vitro*. To build a quantitative understanding of the dynamics driving differences in lineage success, it is necessary to build theoretical expectations for lineage distributions under various evolutionary scenarios.

In my experimental design, lineages are subject to stochastic sampling that is the product of experimental and biological forces: cells are sampled into replicate subpopulations when passaging *in vitro*, whilst the birth-death process of cell growth can lead to the stochastic loss of lineages, especially when the population sizes are small (analogous to genetic drift with regard to allele frequency changes). These forces can lead to differences in lineage success without having to invoke selection. The first step is therefore to explicitly include these dynamics in my model. Subsequently, I can impose different hypotheses concerning the evolution of our phenotype of interest – drug resistance – on top of these sampling forces. In Results Chapter 4, by comparing true, sequenced lineage distributions with these expectations under i) different modes of resistance phenotype evolution, and ii) different parameter ranges within each of these modes, I aim to infer which dynamics are driving resistance evolution in our chosen colorectal cancer cell-lines.

3.2 Modelling Cell Turnover

Before I address possible models of resistance evolution, is is important to consider purely neutral dynamics. That is, given all the situation where all cells have equal birth and death rates, what differences do we expect to emerge in the lineage distributions purely due to the experimental and biological sampling steps?As cells are only uniquely labelled once, and the growing population is exposed to iterative sampling steps, we expect differences between lineages to increase over time. In fact, assuming a net positive growth rate and repeated population bottlenecks, as $t \to \infty$, we expect the number of lineages to monotonically decrease to a population consisting of only a single lineage.

To capture this underlying process of cell turnover, I employ a kinetic monte-carlo, agent-based model to simulate a stochastic birth-death process. Briefly, a cells are assigned birth and death rates. The maximum birth and death rates in the population are noted (b_{max} and d_{max}, respectively). A change in time, $\Delta(t)$ is drawn from the distribution $\Delta(t) = \frac{-1}{(b_{max}+d_{max} \cdot log(r)}$, where r is a random number drawn from $Uniform(0, (b_{max} + d_{max}))$. Figure 3.1 shows the transition rates for a cell given the value of this randomly drawn number. Unless stated otherwise, cells in my simulations are growth with uniform birth and death rates, and therefore $b_{max} = b$ and $d_{max} = d$.

The most powerful feature of experimental evolution is the capacity to observe evolution in parallel: exposing closely related descendants to the same selection pressures in replicates allows us to ask questions such as 'How repeatable is evolution in our chosen conditions?' and 'What is the rate of change of our phenotype of interest within our experimental time frame?'. Therefore, to ensure most lineages are represented numerous times prior to isolating sub-populations in replicates, cells experience a mutual expansion step immediately post-labelling with the lentivirus barcode. *In*

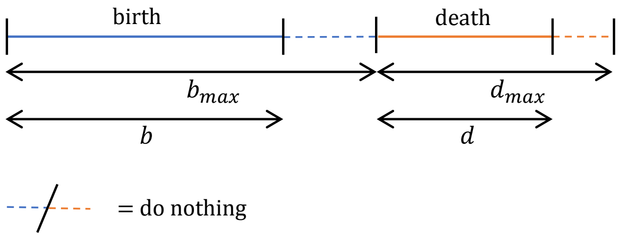


Figure 3.1: A schematic illustrating all the possible transition states for a cell. A number is drawn form the distribution $Uniform(0, (b_{max}+d_{max}))$. A birth occurs if this number falls in the solid blue interval, and a death if it falls in the solid orange interval (the dashed regions correspond to 'no event' occurring). Note that the probability of a birth or death event is directly proportional to a cell's b and d rates.

vitro, cells are then well-mixed and sampled into respective sub-populations (please see Materials and Methods and Figure 2.1 for more details).To ensure that I can infer the evolutionary dynamics from the model's lineage distributions, I design my simulation to reproduce important features the *in vitro* experiment. The most consequential of these is the shared expansion step. The simulation has analogous steps, where cells are uniquely labelled and grown together prior to being repeatedly sampled into isolated sub-populations (Figure 3.2). They are then grown apart, in parallel and subject to simulation-specific perturbations.

Figure 3.3 illustrates the stochastic nature of the birth-death process and successive sampling steps: lineages exhibit differences in frequency following the expansion step, and are randomly lost in some sub-populations. The population sizes simulated here are smaller than those chosen for the full experiment, in order to illustrate the differences that can emerge more clearly. There is a probability lineages are lost to drift early in the growth stage, whilst the stochasticity of success in the subsequent sub-populations is compounded by the probability of being lost during the sampling step.

Assigning uniform birth and death rates to all cells in a population produces a

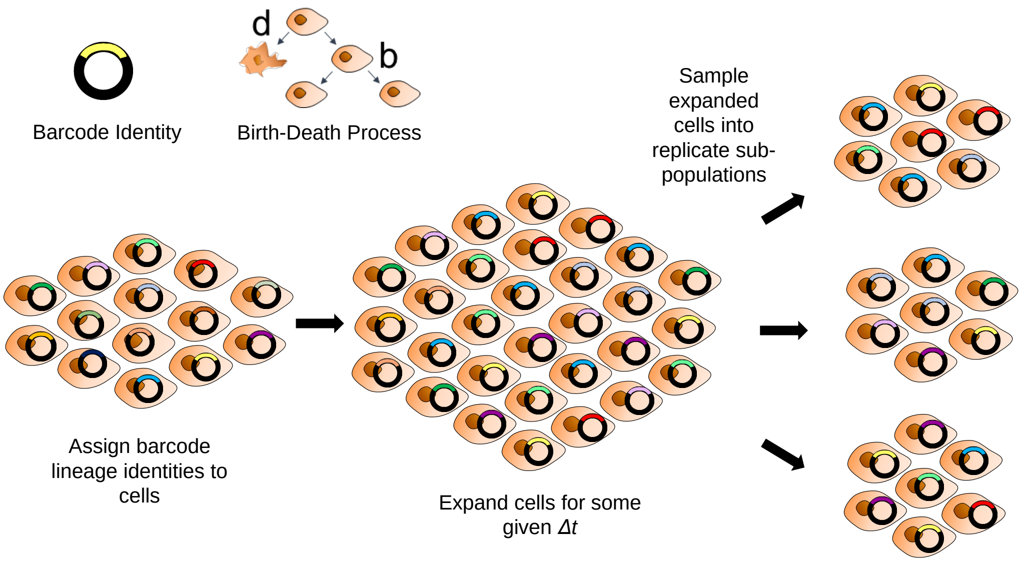

d
b
Barcode Identity
Birth-Death Process
Sample expanded cells into replicate sub-populations
Assign barcode lineage identities to cells
Expand cells for some given Δt

Figure 3.2: A schematic illustrating the components of the birth-death simulation that aim to closely resemble the *in vitro* experiment. Namely, cells are assigned a unique lineage identifier, share a mutual expansion step and are then subsequently sampled (without replacement) into replicate sub-populations.

model where variability in frequency is due solely to the neutral dynamics described previously; a lineage's relative success is purely stochastic. Importantly, whilst repeated iterations of the simulation will not lead to repeatability in any given lineage's success, we might still expect the same lineage to appear successful across replicates within a single simulation iteration without having to invoke selection. That is, when cells share a growth period at the beginning of a simulation (Figure 3.2) a lineage may by chance become more abundant. Subsequently, it is more likely to be at a higher frequency in multiple replicate sub-populations (Figure 3.4).

Without these theoretical considerations, one might arrive at specious conclusions such as 'a barcode found at high frequency across all replicates has a fitness advantage over its contemporaries'. As I have illustrated here, purely stochastic forces can lead to associations between sub-populations' lineage frequencies.

So far, I have outlined the foundation of the stochastic model. Cells are assigned

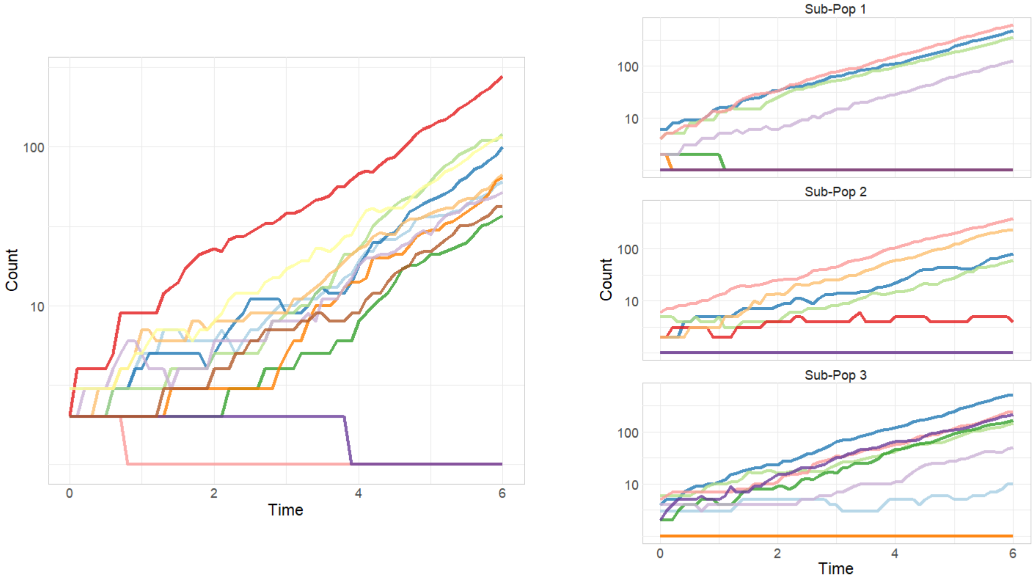


Figure 3.3: Growth trajectories of 12 randomly selected lineages (total lineages = 1000) during a shared expansion stage (LHS panel) and following subsequent sampling into and growth within three replicate sub-populations (RHS panels). All cells have uniform birth and death rates and counts of each lineage correspond to unique colours are shown in continuous time.

lineage identities, uniform birth and death rates and then grown according to the experimental design (Figure 3.2). Distributions and statistics of the populations' lineage counts and frequencies can then be obtained. As the real data consists of sequenced lineage distributions, these can then be compared directly.

3.3 Analytical Solutions to Lineage Growth and Sampling

Before I begin to model the evolution of resistance phenotypes we can make some initial predictions regarding how sub-lineages should be distributed immediately following the shared population's expansion step and upon sampling into multiple sub-populations.

First, we can ask how lineages are distributed given some birth (b) and death (d) rates and some period of time (t); the analytical solution to this birth-death process is

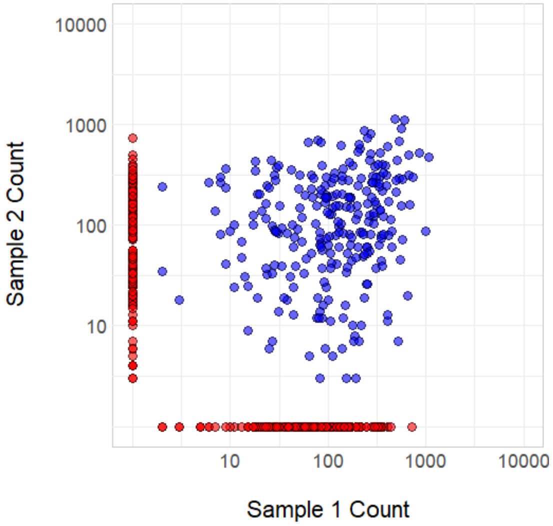


Figure 3.4: A pairwise comparison of lineage counts in two sub-populations (Sample 1 and Sample 2) following a shared expansion step then periods of isolated growth: cells are assigned unique lineage marker, grown together, and then split into two populations before being grown in isolation. All cells are grown with uniform birth and death rates. Shared lineages are shown in blue, and lineages unique to a sample in red.

known (Bailey 1990), where $p_{(n)}$ is the probability of seeing a lineage of size n at time t:

$$p_{(0)} = \alpha \tag{3.1}$$

$$p_{(n)} = (1-\alpha)(1-\beta)\beta^{(n-1)} \tag{3.2}$$

where

$$\alpha = \frac{d(e^{(b-d)t}-1)}{be^{(b-d)t}-d} \tag{3.3}$$

$$\beta = \frac{b(e^{(b-d)t}-1)}{be^{(b-d)t}-d} \tag{3.4}$$

We can also show that this $p.m.f$ describes the stochastic birth-death agent-based model described previously, exactly (Figure 3.5).

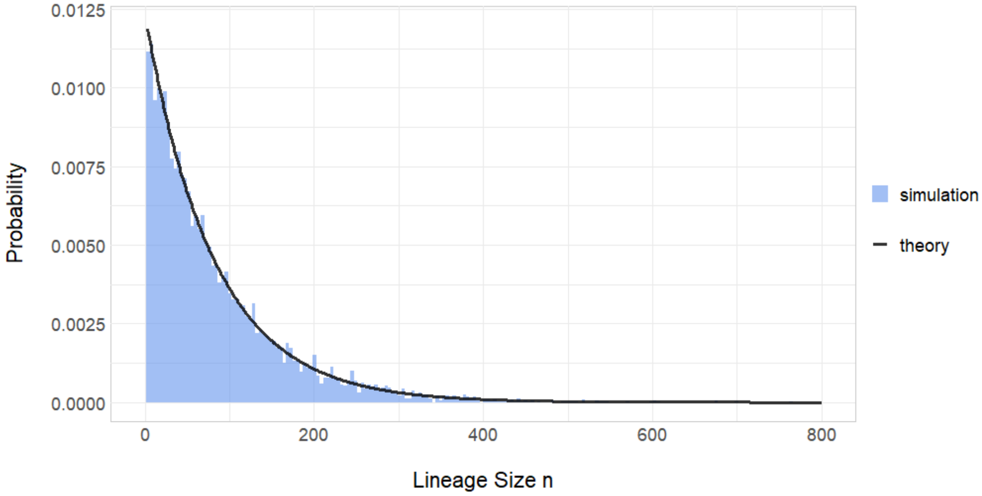


Figure 3.5: The analytical solution (black line) to the birth-death $p.m.f$ where $b = 0.893$, $d = 0.200$ and $\Delta t = 6.0$ and a corresponding lineage-size distribution (blue histogram) from a single iteration of the stochastic birth-death model.

The birth-death $p.m.f$ provides us with a theoretical expectation of lineage sizes following the expansion step, given some estimates of the population's birth and death rates. To derive a distribution for the probability of seeing a lineage k times following sampling, given it is present n times in the expanded population of size N, I make use of the following:

$$p(k) = \sum_{n=0}^{N} p(k|n)p(n) \tag{3.5}$$

where $p(n)$ is (5.2) if $n = 0$ and (5.1) if $n > 0$, and

$$p(k|n) = \frac{\binom{n}{k}\binom{N-n}{K-k}}{\binom{N}{K}} \tag{3.6}$$

if sampling without replacement (hypergeometric $p.m.f$), or

$$p(k|n) = \binom{K}{k}(\frac{n}{N})^k(1-\frac{n}{N})^{K-k} \tag{3.7}$$

if sampling with replacement (binomial $p.m.f$).

Whilst in practice, cells are sampled *without* replacement into their respective sub-populations, modelling this step as sampling *with* replacement allows use to use I binomial distributions, meaning each sub-population is $i.i.d.$ For the values of N, I and K used in the *in vitro* experiment, modelling sampling as a binomial distribution accurately estimates the number of lineages of size k (Figure 3.6).

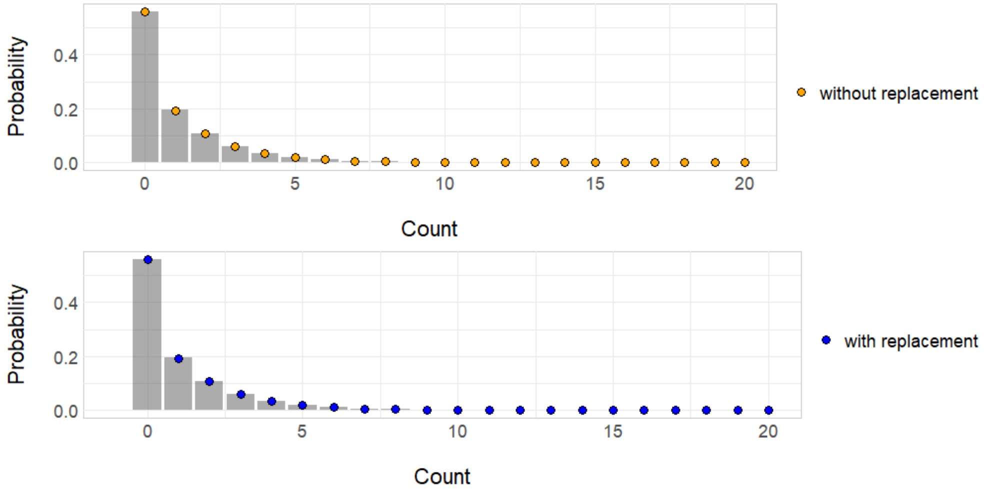


Figure 3.6: Simulated lineage size distributions (grey) within a sub-population following an expansion growth period and single round of sampling *without* replacement vs theoretical expectations given sampling *without* (orange points - hypergeometric) and *with* (blue points - binomial) replacement.

Not all lineages will be sampled into each replicate sub-population. This has implications for how dissimilar we expect each replicate to be (in the absence of selection). Modelling the lineage size distribution using (3.7) for each i of I sub-populations, we can ask: 'what is the probability of a lineage being sampled into i of I sub-populations?'. First, I make use of the complement:

$$p(k > 0) = 1 - \sum_{n=0}^{N} p(k = 0|n)p(n) \tag{3.8}$$

As each sub-population is modelled as one of i independent and identical distributions, we can model the probability that a chosen lineage makes it into i of I sub-populations as a second binomial distribution, where the probability of success $= p(k > 0)$ and the number of trials $= I$. Therefore the probability that a chosen lineage makes it into i of I sub-populations is:

$$p(i) = \sum_{n=0}^{N} \binom{I}{i} (p(k > 0|n)^{i}(1 - (p(k > 0|n)^{i}))^{I-i}p(n) \tag{3.9}$$

This distribution accurately estimates the number of lineages that make it into i of I sub-populations in my stochastic birth-death model. We can see that, in the absence of any selection, we expect approximately a third of all initial lineages to be lost to sampling effects, whilst most extant lineages will be sampled into all 4 sub-populations (Figure 3.7).

These analytical solutions show that the agent based model behaves as expected, whilst capturing important features of the *in vitro* experiment. The probability distributions can tell us i) the number of sub-populations any lineage is sampled into, and ii) the lineage size distributions within each sub-population immediately following sampling. I now proceed to develop models of resistance evolution within this framework.

3.4 Simple Models of Resistance

Cells in the control replicates were grown in the absence of the selection pressure of therapy, whereas the drug-treatment replicates were periodically exposed to therapy.

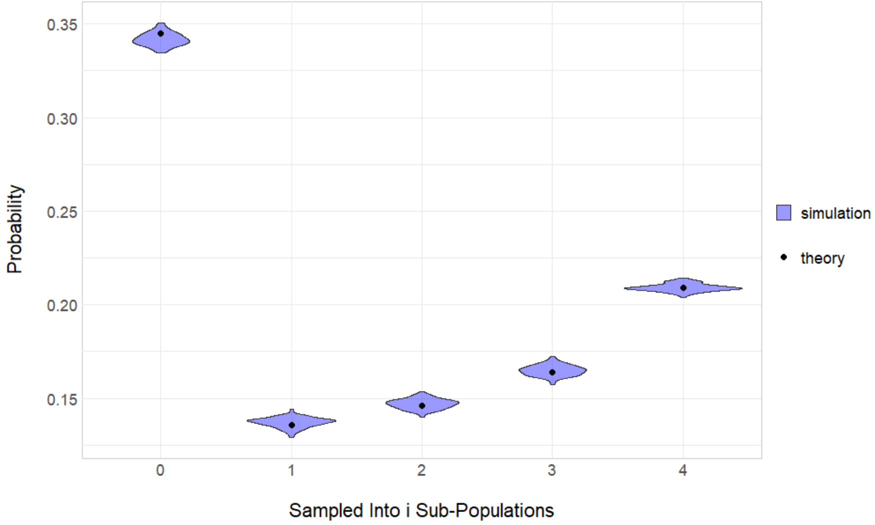


Figure 3.7: Simulated lineages found in i of I sub-populations following a shared expansion period and sampling *without* replacement (blue areas) vs the analytical expectations modelled using the sampling *with* replacement analytical distribution (black points) from Eq 3.9.

To model this process, I simulate the underlying population's birth-death process ('cell turnover') independently from the death incurred due to therapy ('drug-killing' step). In vitro, the cells are exposed to the drug treatment for several days followed by an equal period of recovery. In the simulations, this is replaced with an instantaneous drug-kill step. The resistant phenotype is modelled as a binary trait - cells are either sensitive ($\mathcal{R} = 0.0$) or resistant ($\mathcal{R} = 1.0$) – and death of the sensitive cells during the drug-killing step is deterministic – the probability of death $= (1.0 - \mathcal{R})$.

3.4.1 Pre-Existence and De-Novo Mutations

I begin by considering two very simple evolutionary scenarios: pre-existing resistance and de-novo resistance mutations. In the pre-existing resistance model, cells are simply assigned a resistance phenotype ($\mathcal{R} = 1.0$) at the beginning of the simulation with probability ρ. The resistance phenotype is completely heritable and immutable. As cells are also assigned a unique lineage identity at the beginning of the simulation, there is a one-to-one relationship between a population's lineage identities and resistant phenotypes. In the de-novo mutation resistance model, cells are all initially sensitive ($\mathcal{R} = 0.0$). They then have a probability of acquiring the resistant phenotype ($\mathcal{R} = 1.0$) with probability μ per division. Subsequent mutations – 'double hits' – have no further effect on the phenotype, and mutation to resistance is permanent; I do not permit back-mutations.

3.4.2 Limitations of the Simple Resistance Models

These two modes of resistance evolution – pre-existing resistance and de-novo mutations – have often been used to model drug resistance in cancer. However, these formulations lead to assumptions that oversimplify the biological realities, to the potential detriment of the investigator. Reflecting on the prospective change in phenotype frequencies over time, two questions arise:

1. Firstly, in the pre-existing resistance mode – if there are some proportion of cells that express the resistance phenotype at the onset of the experiment (either *in vitro* or *in silico*), by what process did this ratio of resistance to sensitive phenotypes arise? And why is it now absent for the duration of the experiment?

2. Similarly, in the de-novo mutation mode – why are mutations that produce a perceivable change in the resistance phenotype only now permitted to occur for the duration of the experiment?

One might suggest that the two modes could be combined. However, this new model would still make the implicit assumption that the population would continuously evolve towards a state of exclusively resistant cells. Yet the addition of treatment is often initially met with some degree of cell death, either *in vitro* or *in vivo*, an observation that would be rare if resistance was ubiquitous (In the subsequent Results Chapter I also address this question experimentally, and show the frequency of resistance appears stable over time). As a remedy to these limitations, I aim to develop models of phenotypic evolution where the frequency of a resistance phenotype in the population is the product of explicit evolutionary phenomena. I argue that these modifications improve

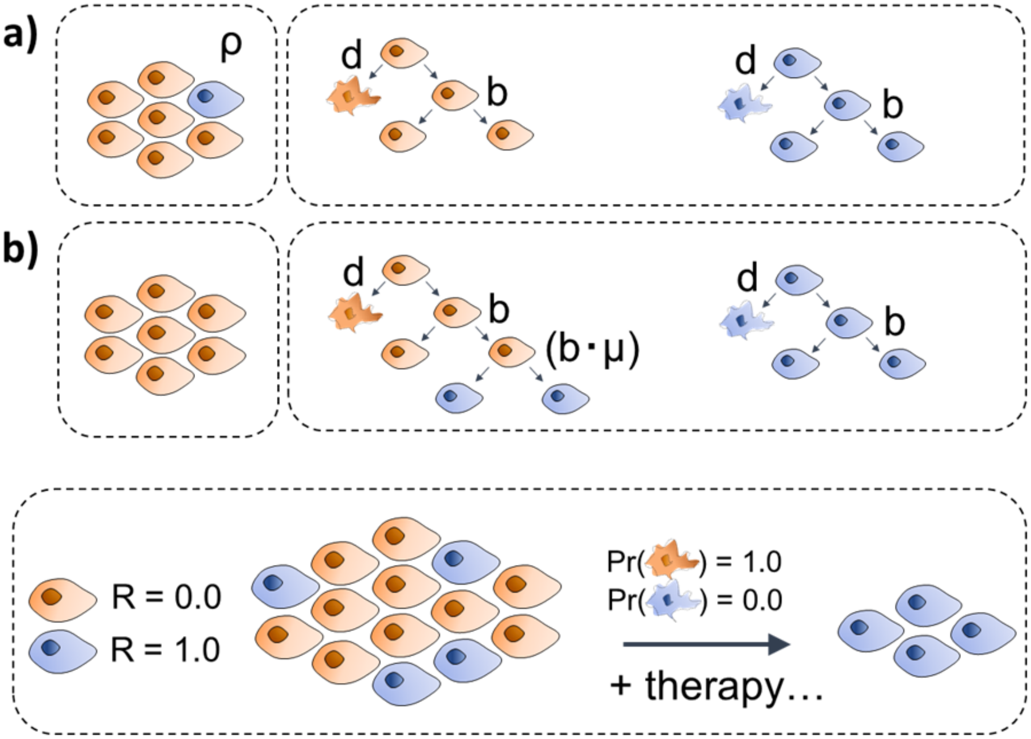


Figure 3.8: a) A schematic outlining the pre-existing resistance model. Cells are assigned the resistant phenotype (blue cells) with probability ρ, and the sensitive phenotype (orange cells) with probability $(1-\rho)$. All cells share birth and death rates: b and d, respectively. b) The same schematic for the de-novo mutation resistance model. All cells now start the simulation with the sensitive phenotype. However, there is now a probability per division - μ - that both daughter cells transition to the resistant phenotype. Drug-treatment induced death is modelled identically in a) and b) drug-death is deterministic: resistant cells ($\mathcal{R} = 1.0$) and sensitive cells ($\mathcal{R} = 0.0$) are killed with probability $1 - \mathcal{R}$.

the biological assumptions underpinning the model, and have important consequences for the theoretical expectations of lineage distributions following treatment.

3.5 Developing Models of Resistance

3.5.1 Mutation-Selection Balance

An evolutionary force that can act to maintain phenotypic variation is mutation-selection balance. Whilst mutations act to increase the frequency of a phenotype within a population, the change in phenotype incurs a fitness cost in an individual's current environment. As such, the forces of selection and mutation act to maintain some equilibrium frequency of the phenotype in question. Before developing the framework for this mode of resistance evolution, it is worth noting several features of mutation-selection balance. Firstly, if purifying selection is too weak, or the mutations too common (or vice-versa), I note that this process could still lead to an equilibrium phenotype frequency of either 0.0 or 1.0. Secondly, the equilibrium frequency of the phenotype is dependent on its fitness cost and, therefore, also dependent on the current selection pressures experienced by an individual. A change in environment may change the fitness of the chosen phenotype and, therefore, lead to a change in its frequency. As such, modelling a population as having reached some stable equilibrium implicitly assumes the current environment is also stable. Under controlled conditions *in vitro*, I argue that this is a safe assumption. Finally, whilst in classical population genetics mutation-selection balance is framed in terms of deleterious alleles, here I continue to focus on fitness cost and benefits in terms of a cell's phenotype. That is, I only assume that resistance incurs some fitness cost, without defining the molecular change responsible.

To model mutation-selection balance, I again define resistance as a binary phenotype, whereby cells are either sensitive ($\mathcal{R} = 0.0$) or resistant ($\mathcal{R} = 1.0$) and death of sensitive cells during the drug-killing step is deterministic: the probability of death = ($\mathcal{R} - 1.0$). Once more, cells have a probability of acquiring the resistant phenotype ($\mathcal{R} = 1.0$) with probability μ per division, subsequent mutations have no further effect on the phenotype, and back mutations are not permitted. However, a resistant cell now incurs some fitness cost δ, relative to the sensitive population. If sensitive cells have a net growth rate λ_r such that $\lambda_r = b_r - d_r$, δ is implemented as a relative fitness cost such that the resistant population grows with rate $\lambda_R = b_R - d_R = \lambda_r - (\delta\lambda_r)$ (Figure 3.9). The death incurred due to therapy is once again modelled instantaneously - with probability of death = $(\mathcal{R} - 1.0)$ - and separately from the population's underlying birth-death process.

3.5.2 Non-Genetic Phenotypic Variability

A second source of variability in phenotypes can occur even within isogenic populations. Non-genetic mechanisms mean populations can generate phenotypic diversity on time frames much shorter than genetic mutations. If this variability leads to cells switching to and from two different phenotypes – for example, sensitivity and resistance – with constant, albeit potentially asymmetrical rates, this too can produce equilibrium frequencies of a phenotype within a population. To model non-genetic sources of phenotypic variability, there remains some probability that cells acquire the resistant phenotype ($\mathcal{R} = 1.0$) per division, μ. However, there is now a second rate -σ - with which resistant cells revert to the sensitive phenotype ($\mathcal{R} = 0.0$) per division (Figure 3.9). Of note, this framework can still model forwards and back mutations, if μ and σ are very low per-division. Alternatively, if one or both of the transition rates are sufficiently high the model can simulate non-genetic variability, where cells transiently exist in a given phenotypic state for a few cell divisions.

3.5.3 Model Parameters Govern Phenotypic Change

Whilst studies often focus on modelling the genetic changes responsible for resistance, it is important to again highlight that I only model the resistance *phenotype* and make no explicit assumptions with regards to the molecular change responsible. This decision was made in light of the information available to me during the evolution experiment, where I only have access to the change in lineage frequencies over time. These dynamics will be the product of stochastic sampling and a selection on a cell's phenotype; phenotypes will be the product of heritable, genetic changes, other non-genetic mechanisms,

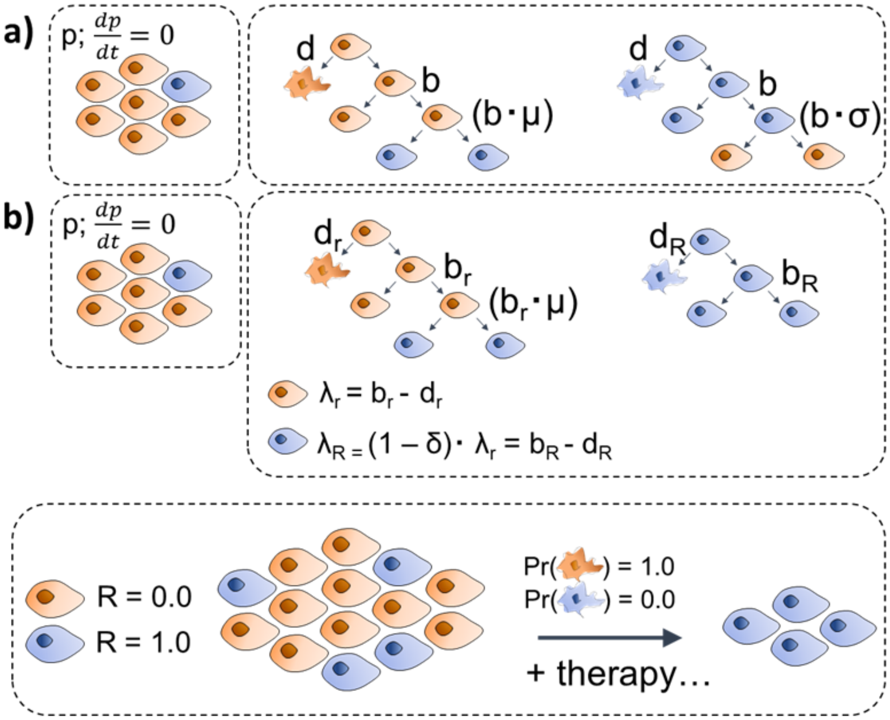


Figure 3.9: A schematic outlining the non-genetic phenotypic variability (a) and mutation-selection balance (b) models. Cells are assigned the resistant phenotype (blue cells: $\mathcal{R} = 1.0$) with probabilities derived from their equilibrium frequencies - p - and the sensitive phenotype (orange cells: $\mathcal{R} = 0.0$) are assigned with probability $((1-p))$. In the non-genetic phenotypic variability model, sensitive cells can become resistant with rate μ per-division, and resistant become sensitive with rate σ per-division. In the mutation-selection balance model, resistant cells incur a fitness cost such that their net growth rate, λ_R, is $(1-\delta)$ that of the sensitive cell net growth rate, λ_r. Drug-treatment induced death is modelled identically in a) and b) - drug-death is deterministic: resistant cells ($\mathcal{R} = 1.0$) and sensitive cells ($\mathcal{R} = 0.0$) are killed with probability $1 - \mathcal{R}$.

or a combination of each. I therefore make no assertion regarding either in this model of resistance evolution. Therefore, whilst I will refer to modelling 'mutations' that move a cell's resistant phenotype, these could in fact be interpreted as either heritable, genetic changes such as single-nucleotide polymorphisms (SNPs), insertions/deletions (indels) and chromosomal copy-number alterations (CNAs), or non-genetic cellular mechanisms of phenotypic diversity such as stochastic variation in gene expression. For brevity's sake, I employ the term 'mutation' to capture all of these potential changes in phenotype in the simulations, unless stated otherwise.

3.5.4 Equilibrium Phenotype Frequencies

In my models of resistance evolution that incorporate a cost of the resistant phenotype and non-genetic sources of phenotypic variability, we can describe the change in the proportion of resistant:sensitive cells in the population over time with the following pair of differential equations:

$$\frac{dn_R}{dt} = n_R(b_R - d_R) + n_r(2\mu b_r) - n_R(2\sigma b_R) \tag{3.10}$$

$$\frac{dn_r}{dt} = n_r(b_r - d_r) + n_R(2\sigma b_R) - n_r(2\mu b_r) \tag{3.11}$$

where n_R and n_r are the number of resistant and sensitive cells, respectively, b_R, d_R, b_r and d_r are the resistant and sensitive birth and death rates, μ is the probability of a sensitive cell producing two daughter resistant cells, per division, and σ is the probability of a resistant cell producing two daughter sensitive cells, per division.

The rationale for these developed models of resistance evolution was to capture biological forces that produce some phenotypic diversity within the population in a time-independent manner. Selection subsequently acts on this phenotypic variation following a change in selection pressure: in my case, the onset of drug-treatment. Now I have formalised the change in resistant and sensitive cells over time given some evolutionary parameters, I can ask what stable frequency of each phenotype these values would lead to in the population. Given $R_{eq} = \frac{n_R}{(n_R+n_r)}$, $\lambda_r = (b_r - d_r)$ and $\lambda_R = (b_R - d_R)$, I calculate the equilibrium frequency of the resistant phenotype in the population by rearranging (3.10) and (3.11) in terms of $\frac{dR_{eq}}{dt}$, setting to 0, and then solving for R_{eq}:

$$\frac{dR_{eq}}{dt} = (-\lambda_R + \lambda_r)R_{eq}^2 + (\lambda_R - 2\mu b_r - 2\sigma b_R - \lambda_r)R_{eq} + 2\mu b_r = 0 \tag{3.12}$$

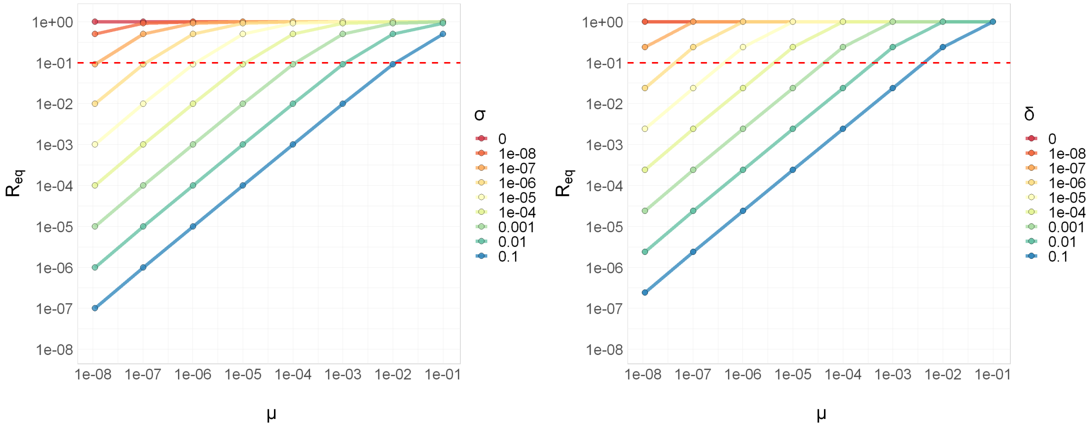


Figure 3.10: Solutions to the stable equilibrium frequency of resistance within a population given parameters that control the phenotype switching rate (μ and σ) and its relative fitness cost (δ): equilibrium proportions of the resistance phenotype - R_{eq} - are plotted as a function of different combinations of μ and σ (LHS panel) and of μ and δ (RHS panel) derived by solving the equation (3.12) for $\frac{dR_{eq}}{dt} = 0$. Parameter combinations that lead to $R_{eq} = 0.1$ in both models have been highlighted (red dashed lines).

Given that the difference between λ_R and λ_r is the relative cost of the resistance phenotype, δ, we can now explore the relationship between various combinations of μ, σ and δ and the equilibrium frequency of the resistance phenotype within the population, R_{eq} (Figure 3.10). It is clear that different combinations of μ - the rate of resistance conferring 'mutations' per division - and either the plasticity reversion rate to sensitivity,

σ, or the fitness cost incurred by the resistance phenotype, δ, can lead to different levels of standing phenotypic variation within the population.

One feature of a given cell population evolving with these evolutionary parameters is the observed proportion of resistance (a product of the equilibrium frequency). Higher proportions of resistance should lead to a high number of surviving lineages following drug-treatment, whilst low equilibrium frequencies will lead to fewer. I therefore expect combinations that lead to the same frequency of resistance to lead to similar losses of lineage diversity following therapy. However, we can also look at differences between scenarios that lead to the *same* frequency of resistance. Despite the same levels of population-wide resistance, differences in growth rates (given δ) or phenotypic switching (given σ) could lead to differences in how resistance is distributed amongst *lineages*. Differences in lineages that survive treatment between *replicate sub-populations*, evolving in parallel, could help distinguish between these scenarios. This idea is discussed in more detail shortly.

Given some equilibrium proportion of resistance we wish to investigate - for example, $R_{eq} = 0.1$ in Figure 3.10 (red dashed line) - parameter combinations can be chosen accordingly; controlling for the proportion of resistance prior to drug treatment enables the identification of differences in lineage distributions due solely to the mode of resistance. Of note, we can see in Figure 3.10 that when comparing identical parameter sets, combinations of μ and σ lead to lower equilibrium frequencies of resistance (R_{eq}) than those of μ and δ.

To ensure that these analytical solutions to the equilibrium proportions of resistance are predicting the stable proportions we'd expect populations to reach in the agent-based model, I simulated the same parameter combinations in the stochastic birth-death model where all cells started with the sensitive phenotype ($\mathcal{R} = 0.0$). They were then grown and subjected to population bottlenecks repeatedly. The proportion of resistance these populations approached was then compared to the expected equilibrium proportions derived by solving equation (3.12) for $\frac{dR_{eq}}{dt} = 0$. The agent-based model does approach the expected equilibrium proportions, with the caveat that the lower values of μ, σ and δ are more prone to stochastic dynamics and therefore lead to

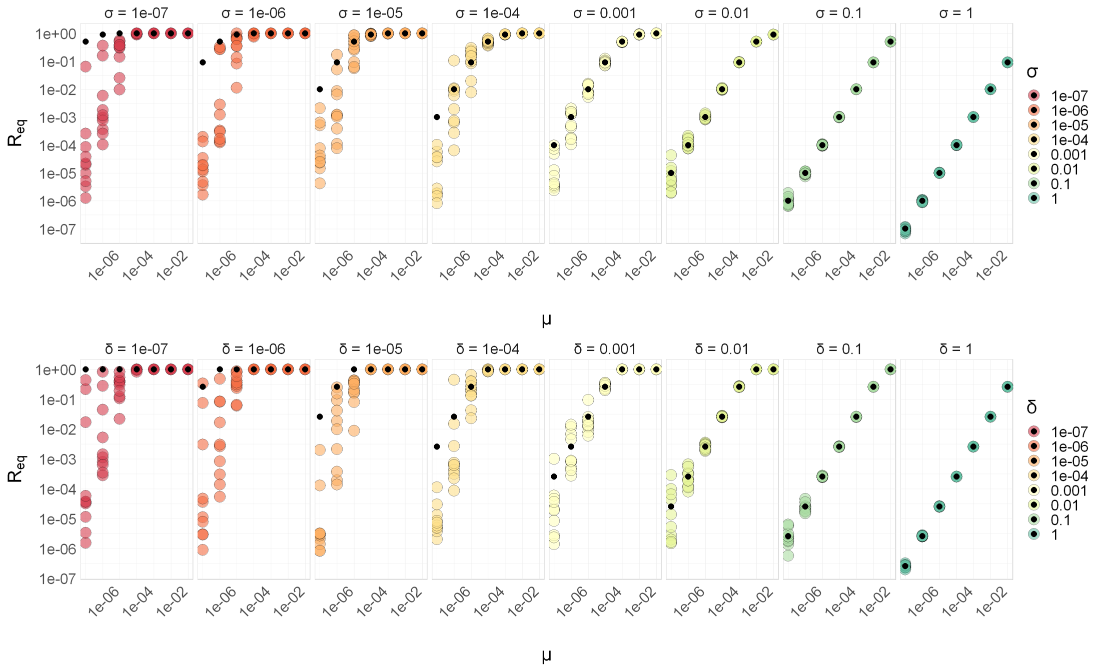


Figure 3.11: A comparison of the analytical solutions of the equilibrium frequency of resistance in the population with long-term simulated frequencies: equilibrium proportions of the resistance phenotype - R_{eq} - were derived by solving the equation (3.12) for $\frac{dR_{eq}}{dt} = 0$ (black points) and are plotted vs simulated values. Simulated values were calculated as the mean proportion of resistance in the final 1000 bottleneck-growth cycles given various values of μ and σ (top panel) and μ and δ (bottom panel).

more variable values of R_{eq} (Figure 3.11). This feature of the population dynamics is pertinent within the real experiment: within a given time-window, lower values of both μ and σ or δ will lead to fewer transitions from resistant to sensitive (and vice versa for the σ case) than higher values of the parameters, even when controlling for the expected proportion of resistant cells (e.g. dashed-red line in Figure 3.10). This difference has important consequences discussed in section 3.6.

Finally, the Algorithms 1, 2 and 3 integrate the ideas discussed so far and describe the simulation functions that: create a vector of cells with resistant phenotypes; grow cells and save the new cell vector; and grow cells whilst periodically killing cells accord-

Algorithm 1: Seed Cells

```
input  : Parameters controlling the total number of cells, the birth and
         death rates, and the resistant phenotype of each cell: $N$, $b$, $d$, $\rho$, $\mu$,
         $\sigma$, $\delta$, $limprobs$
output: A vector of $N$ uniquely barcoded cells, each with a resistant
         phenotype: $R \in [0.0, 1.0]$
for $i$ to $N$ do
    Assign cell a unique barcode, $i$;
    Assign cell birth and death rates, $b$ and $d$;
    if $limprobs = true$ then
        Assign $R$ to cell using equilibrium frequency, $R_{eq}$, given $\mu$, $\sigma$ and $\delta$
    else
        Assign $R$ to cell with probability $\rho$
    end
end
Return vector of cells;
```

Algorithm 2: Grow Cells

```
input  : A vector of cells, $cell_{vec}$,: where $cell_i$ has [$barcode$, $b$, $d$, $R$];
Parameters that control the resistant phenotype evolution: $\mu$, $\sigma$, $\delta$;
Parameters that control the growth period: $t_{max}$, $N_{max}$;
output: A new vector of grown cells: where $cell_i$ has [$barcode$, $b$, $d$, $R$];

set $t$ to 0;
while $t < t_{max}$ or $N < N_{max}$ do
    sample random uniform number, $ran_1 \in [0, (b + d)]$;
    sample single random cell, $ran_{cell}$ from $cell_{vec}$;
    set $b$ and $d$ according to $ran_{cell}.R$ and $\delta$;
    if $ran_1 < b$ then
        // birth event:  check if resistant phenotype changes
        sample random uniform number, $ran_2 \in [0, 1]$;
        if $ran_{cell}.R = 0.0$ then
            if $ran_2 < \mu$ then
                $ran_{cell}.R$ becomes 1.0
            end
        else if $ran_{cell}.R = 1.0$ then
            if $ran_2 < \sigma$ then
                $ran_{cell}.R$ becomes 0.0
            end
        Duplicate $ran_{cell}$ and add to $cell_{vec}$;
    else if $b <= ran_1 < (b + d)$ then
        // death event
        Remove $ran_{cell}$ from $cell_{vec}$;
    Update time;
    sample random uniform number, $ran_3 \in [0, 1]$;
    $\Delta t = \frac{-1}{(b+d)N_t} \cdot log(ran_3)$;
    $t = t + \Delta t$;
end
Return vector of cells;
```

Algorithm 3: Grow-Drug-Kill Cells

```
input  : Same input as *Grow Cells*, and also:;
$drug\_kill(true/false)$, $insta\_kill(true/false)$, $n\_pulse$, $\psi$
output: A new vector of grown cells: where $cell_i$ has [$barcode$, $b$, $d$, $R$];
A vector of total population sizes, $N_t$, per $n\_pulse$;

set $t$ to 0;
$\Delta t_{pulse} = t_{max}/n\_pulse$
while $t < t_{max}$ or $N < N_{max}$ do
    if $drug\_kill = true$ then
        if $insta\_kill = true$ then
            for $i$ to $N$ do
                Kill $cell_i$ with probability $(1 - cell_i.R) + \psi$;
            end
            Record the total population size;
        end
    end
    Grow cells using the *Grow Cells* function for $\Delta t_{pulse}$;
    if $drug\_kill = true$ then
        for $i$ to $N$ do
            Kill $cell_i$ with probability $(1 - cell_i.R) + \psi$;
        end
    end
    Record the total population size;
end
Return vector of cells;
Return vector of population sizes;
```

ing to their resistant phenotypes, respectively. Details on the code and languages used can be found in the Materials and Methods section.

Now I have shown that combinations of parameters in the models of resistance evolution lead to equilibrium frequencies of the resistance phenotype, I subsequently use the values of μ, σ, δ and λ_r to assign resistance ($\mathcal{R} = 1.0$) to cells at $t = 0$ (prior to the expansion step) with probability R_{eq} in the agent-based simulations. Cells are then grown according to their respective parameters.

3.6 Leveraging Within- and Between-Sub-Population Information

So far, I have outlined a model that assigns cells parameters that control the underlying growth of a population via birth and death rates. An additional set of parameters controls the distribution and rates of change of two mutually exclusive phenotypes of interest: drug-resistance and sensitivity, whilst the proportion of resistant:sensitive phenotypes are assigned according to the expected equilibrium frequencies at the start of the simulation. Within this framework, I can now ask whether we can distinguish between different evolutionary scenarios solely by comparing lineage distributions. To make these distinctions, there are three important sources of information available following the *in vitro* experiment:

1. The number of high-frequency lineages remaining within the replicate sub-populations following drug-treatment. This distribution will be the product of the standing variation of the resistant phenotype within the population (R_{eq}).

2. The difference in successful, high-frequency lineages between replicate sub-populations following drug-treatment. Mutual lineage success amongst sub-populations will be the product of the 'stability' of the resistant phenotype during i) the whole population's expansion step, and ii) drug-treatment within a replicate sub-population post-expansion.

3. The distribution of lineages within and amongst control-treatment sub-populations.

> Assuming that selection to non-drug conditions are weaker, these replicate populations will provide 'baseline' expectations with which the drug-treatment scenarios can be compared.

To leverage these sources of information, I require appropriate summary statistics. Ideally, these would condense my $(2 \cdot I \cdot Passage)$ different distributions $(I \cdot Passage$ per drug- and control-treatments) into two axes that summarise both the number and similarity of the lineage distributions. One approach would be to compare the total number of lineages within a sub-population vs a similarity index of lineages between sub-populations, e.g. the Jaccard index, which is the intersection divided by the union of two sets. However, the use of absolute lineage number has two drawbacks. The first is that I am interested primarily in the successful lineages: those that reach a high frequency within their sub-population. A large number of very low-frequency lineages would inflate the absolute count whilst contributing little to the population's current response to therapy. The second drawback is technical - errors in lineage-barcode amplification and sequencing can further inflate the number of low-frequency lineages, giving a false impression of the true number. A solution to both of these problems is to adopt diversity indices that accounts for both the number and relative frequency of lineages.

3.6.1 Within-Population Diversity

To capture the diversity of lineages within a sub-population, I adopt the Hill diversity indices of order q, ${}^{q}D$, defined as

$$ {}^{q}D = (\sum_{j=1}^{J} {p_j}^{q})^{1/(1-q)} \tag{3.13}$$

where p_j is the relative frequency of the j^{th} lineage, and q controls the leverage that high frequency lineages contribute to the index. Relative frequencies are scaled according to q, where values of $q < 1$ and $q > 1$ preferentially leverage low- and high-frequency lineages, respectively (Jost 2006; Roswell et al. 2020). When $q = 0$, ${}^{q=0}D$ is simply the total number of lineages (often referred to as 'species richness' in ecology). The exact solution for (3.13) when $q = 1$ does not exist, however as $q \rightarrow 1$,

$$^{q=1}D = exp(-\sum_{j=1}^{J} p_j log(p_j)) \tag{3.14}$$

which is also referred to as the Shannon diversity (the natural exponential of the Shannon entropy). When all lineages are present in equal proportions ($p_1 = p_2 = ... = p_J$) the values of ^qD are insensitive to q, and correspond to the total number of lineages.

^qD provides a way to circumvent the problems that arise When using raw lineage counts alone as a measure of diversity. Different values of ^qD between populations following drug-treatment should correspond to inter-population differences in the proportion of resistant lineages; if the resistance phenotype ($\mathcal{R} = 1.0$) is rare, we expect the majority of lineages to be lost during treatment, leading to low values of ^qD.

This apparently innocuous prediction that, following drug-treatment, a reduction in diversity will be a function of the population's resistant fraction rests on some biological assumptions that deserve scrutiny. Notably, there are two scenarios where this prediction might not hold. Firstly, if treatment acts by killing individuals with some set probability - independent of any phenotype - this could hypothetically lead to a diversity reduction of 0, where each lineage is depleted in equal measures. Alternatively, if for now I continue to assume resistance is a binary, heritable trait, a 0 reduction in a population's diversity could occur post-treatment if the resistance phenotype was equally distributed amongst all lineages. In practice, as long as the response to treatment is governed by some cell-specific phenotype which can be inherited by daughter cells and exhibits variability amongst individuals, these scenarios remain improbable.

3.6.2 Between-Population Diversity

Hill-diversity indices of the order q allow us to tune how strongly high-frequency lineages influence our statistics. By the same token, we can transform (x^q) and back-transform ($x^{1/(1-q)}$) lineage frequencies when quantifying the dissimilarity in diversity between populations. Specifically, this is captured in $^qD(\beta)$, the effective number of unique

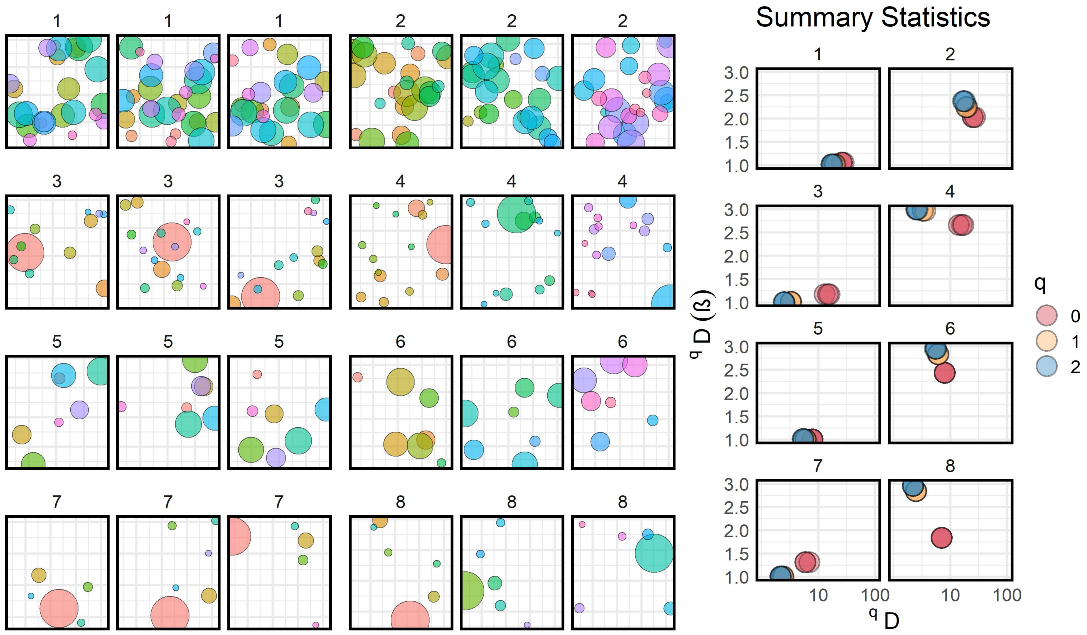


Figure 3.12: Simulated, illustrative lineage distributions highlighting how diversity summary statistics - ${}^{q}D$ (within-population diversity) and ${}^{q}D(\beta)$ (between-population diversity dissimilarity) - differ with total lineage numbers and abundances. The LHS panels represent various populations, where numbers denote evolutionary scenarios $(1-8)$, and panels hold distinct sub-populations (3 per scenario). Point size correspond to a given lineages frequency, the colour denotes lineage identity consistently between all panels and lineages are randomly distributed within each panel. ${}^{q}D$ vs ${}^{q}D(\beta)$ statistics (RHS panels) are shown for respective evolutionary scenarios $(1-8)$ for various orders of q $(q = 0, 1, 2)$.

populations or the true beta diversity of order q. ${}^qD(\beta)$ is the equivalent of the often adopted beta diversity, with the distinction that these transformations permit adjustment of the contribution of high-frequency lineages, whilst also having an intuitive range of $[1, I]$, where I is the total number of populations being compared. ${}^qD(\beta)$ can be calculated by partitioning multiple sub-population's diversity into two components:

$$ {}^qD(\beta) = \frac{{}^qD(\gamma)}{{}^qD(\alpha)} \tag{3.15} $$

where

$$ {}^qD(\gamma) = (\sum_{j_I=1}^{J_I} p_{j_I}{}^q)^{1/(1-q)} \tag{3.16} $$

where I is the total number of sub-populations being compared, J_I is the total number of lineages amongst all I populations, and p_{j_I} is the relative frequency of lineage j amongst all I populations, and

$$ {}^qD(\alpha) = (\frac{1}{I}\sum_{i}^{I}(\sum_{j}^{J}(p_{ij})))^{1/(1-q)} \tag{3.17} $$

where p_{ij} is the relative frequency of lineage j in population i.

In words, ${}^qD(\gamma)$ captures the total diversity of order q when pooling all I sub-populations, ${}^qD(\alpha)$ is the mean diversity of order q of all I sub-populations, and ${}^qD(\beta)$ is the ratio between the two. When $q = 0$, ${}^{q=0}D(\beta) = 1.0$ when all lineages are shared amongst all I sub-populations - there is one ‘effective’ unique population; and ${}^{q=0}D(\beta) = I$ when no lineages are shared amongst all I sub-population - there are I 'effective' unique populations.

Figure 3.12 illustrates how some differences within and between (hypothetical) lineage distributions are captured by differences in ${}^qD$ and ${}^qD(\beta)$. In particular, how increasing the value of q greatly reduces the diversity in samples where there are a few, dominant lineages (Figure 3.12, scenarios 3,4,7 and 8) and how this also produces greater differences in values of ${}^{q}D(\beta)$ in comparisons that have a few, dominant lineages that are shared (Figure 3.12, scenarios 3 and 7) and those where the high-frequency lineages are unique ((Figure 3.12, scenarios 4 and 8).

3.7 Simulation Results

I have now described two simple models of resistance evolution and their shortcomings which form the rationale for two developed modes of resistance evolution. These models calculate equilibrium phenotypic frequencies (resistant:sensitive) to assign the proportion of resistance to individual cells at the beginning of the simulation ($t = 0$) prior to evolving cells according to the simulation's respective evolutionary parameters. These models permit me to simulate analogous lineage distributions to those of the sequenced output of my *in vitro* experiment. Finally, I have outlined summary statistics that can capture two important axes of information between these lineages: within-population lineage diversity and between-population lineage diversity differences. The question remains as to whether I can now distinguish evolutionary scenarios of drug-resistance evolution by comparing lineage distributions.

The following simulation parameter values were chosen to resemble the *in vitro* experiment as closely as possible. In particular, these initial birth and death rates were chosen to be in a range of previously reported birth and death rates in cancer cell-lines *in vitro* (Acar, Nichol, Fernandez, et al. 2019; Russo et al. 2021), whilst erring on the side of higher cell turnover ($b+d$). As increasing the cell turnover increases the variance in lineage distributions by increasing the probability cell lineages are randomly lost to drift (see section 3.3), this ensured I didn't under-estimate the stochastic components of the experiment. For convenience, I also ensured that the net growth rate ($b - d$) $= log(2) \approx 0.693$. This means that a single time unit, $\Delta t = 1.0$ corresponds to a single population doubling.

- $N_0 = 10^6$ (the number of uniquely barcoded cells at $t = 0$.
- $b = 0.893$ (therefore $\lambda = (b - d) \approx log(2)$ and $\Delta t = 1 \approx 1$ population doubling.)

- $d = 0.200$ (therefore $\lambda = (b - d) \approx log(2)$ and $\Delta t = 1 \approx 1$ population doubling.)

- $t_{exp} = 6.0$ (Δt for shared expansion step prior to splitting into replicate sub-populations).

- $I = 4$ (the number of replicate sub-populations per control- and drug-treatment - therefore $2I$ total).

- $\Delta t_{DT} = 2.0$ (the time in between drug-kill events in the drug-treatment replicates)

- $N_{max} = 64 * 10^6$ (the carrying capacity of the replicate sub-populations - flasks in the *in vitro* experiment).

- $\rho, \mu, \sigma, \delta$ - these parameter values control the evolution of the resistance phenotype as described in the text, and are simulation dependent.

- $nsim = 10$ (number of simulation iterations per unique parameter combination).

3.7.1 Growth Kinetics

The growth kinetics of the simulations were contingent on the combination of evolutionary parameters μ, σ and δ. Figure 3.13 shows example population trajectories for a subset of simulations where $\mu = 10^{-6}$.

As death due to drug-treatment is deterministic in these versions of the model (death due to treatment is enforced with probability $(1 - \mathcal{R})$), nearly all of the selection for the resistant phenotype occurs immediately following the cells being sampled into their respective sub-populations: the proportion of surviving cells is determined by the fraction of resistant cells in the sub-population at that time - a function of the equilibrium frequency of resistance assigned at $t = 0$ (p) - and any resistance subsequently lost or gained in the shared expansion stage.

Whilst I model death due to treatment in this deterministic fashion, all of the drug-induced death nearly exclusively precedes the end of Passage 1. The only scenarios where there are additional deaths due to treatment beyond the first drug-kill step are when the switching rate from resistant to sensitive is high (e.g. $\sigma = 0.1$ in Figure 3.13). Here, enough cells have reverted to the sensitive phenotype by the next drug-kill step

(Δt_{DT}) that there is a detectable decrease in the population following the next drug-kill step. As the direction of sensitive to resistant evolution in the 'cost of resistance' scenarios ($\delta > 0$ - bottom row, Figure 3.13) is unidirectional and the death due to drug is deterministic, there is no analogous decrease (e.g. $\delta = 0.1$).

3.7.2 Distinguishing Evolutionary Scenarios with Lineage Distributions

The next step is to compare different evolutionary scenarios using the summary statistics I have chosen (discussed in section 3.6) that capture within- and between-population differences in lineage success; ${}^qD$ and ${}^qD(\beta)$. Due to reasons outlined earlier, for now I focus solely on the lineage distributions in Passage 2. The control replicates provide useful null expectations for lineage distributions *in vitro*; relative to the drug-treatment replicates, we expect selection to be weak as cells evolve in standard culture conditions. I therefore simulate the control treatment *in silico* by assuming uniform birth and death rates - neutral dynamics. I previously made some simple predictions given these dynamics by comparing simulated distributions to analytical solutions (Section 3.3). I now use the simulated Passage 2 control distributions to set a baseline expectation for diversity within a population's lineage distributions, and the level of divergence between multiple populations. Deviation from these values in the drug-treatment replicates represents the additional change in the within and between replicate diversity measures; changes above that expected given the biological and technical sampling steps alone.

Following the pooled control replicates' distributions, I can now compare values of ${}^qD$ and ${}^qD(\beta)$ for parameter combinations that capture an array of evolutionary scenarios. Figures 3.15 and 3.16 shows that a large number of parameter combinations lead to values that are indistinguishable from the control values (black points in Figures 3.15 and 3.16). These results make sense when instead highlighting each simulation output by the equilibrium frequencies of resistance (R_{eq}) assigned at the beginning of the simulation (Figures 3.17 and 3.18). Parameter combinations that lead to high equilibrium fractions of resistance - namely, high values of μ, and low values of σ and δ - are unaffected by drug-treatment, and therefore lineage relationships are simply

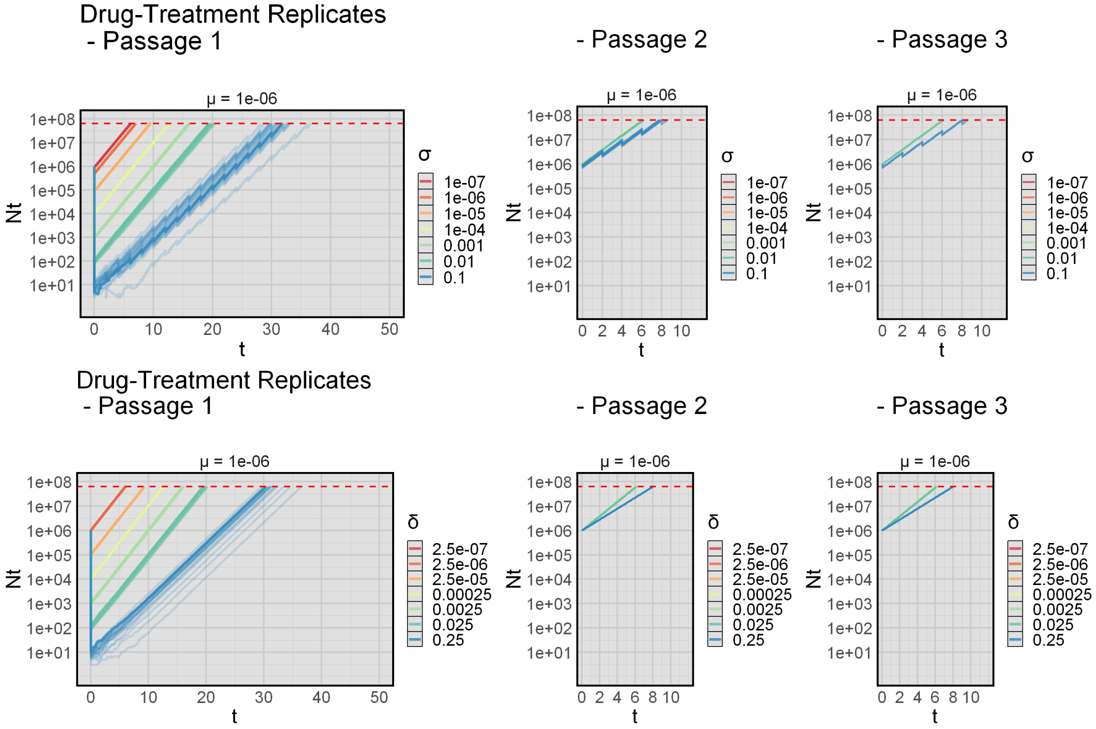


Figure 3.13: Example drug-treatment population trajectories for simulations where values of $\mu = 10^{-6}$ - following assignment to replicates, the total population size of each replicate sub-population was recorded at regular intervals throughout the simulation. Non-genetic sources of phenotypic variability simulations ($\sigma > 0.0$ - top row) and cost of resistance simulations ($\delta > 0.0$ - bottom row) where colour correspond to different values of σ and δ. Columns correspond to simulated passage number. The carrying capacity of the sub-populations has been marked in each panel (red-dashed line).

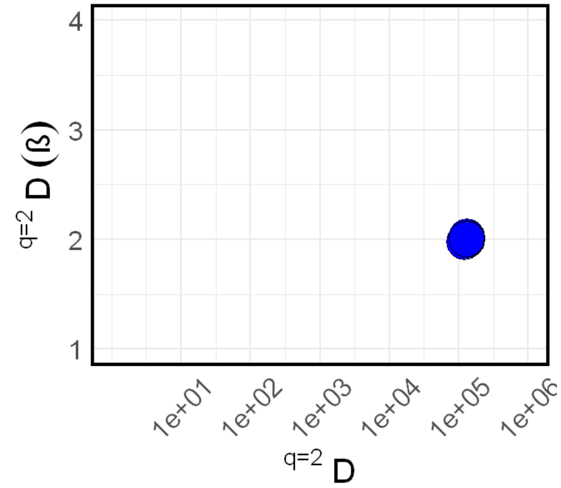


Figure 3.14: ${}^qD$ (within-replicate lineage diversity) vs ${}^qD(\beta)$ (between-replicate lineage diversity dissimilarity) of order $q = 2$ for the combined simulation's control (CO) replicates for all combined simulation outputs (σ and $\delta > 0.0$).

governed by the underlying birth-death process, as in the control replicates. Adopting these parameter values makes little biological sense as cells *do* respond to therapy *in vitro*. There are also combinations of parameters that consistently lead to too few resistant cells at the onset of drug-treatment for any cells to survive the experiment (blank panels in Figures 3.15 and 3.16). Again, these parameter values that consistently lead to extinction are of little interest; scenarios where all cells are killed by the drug-treatment aren't observed in the *in vitro* experiment.

First, I can exclude the scenarios that are indistinguishable from the control treatment dynamics. Secondly, although there are differences amongst the remaining parameter combinations, it is difficult to compare like-with-like: it isn't clear which combinations lead to different levels of resistance in the population at treatment onset. Figure 3.19 therefore instead organises parameter combinations so that rows now correspond to simulations that share an equilibrium frequency of resistance (R_{eq}). I expect these simulations (rows for a given R_{eq}) to have similar ratios of resistant:sensitive cells when drug-treatment begins. Indeed, the value of ${}^qD$ agrees with this prediction, where

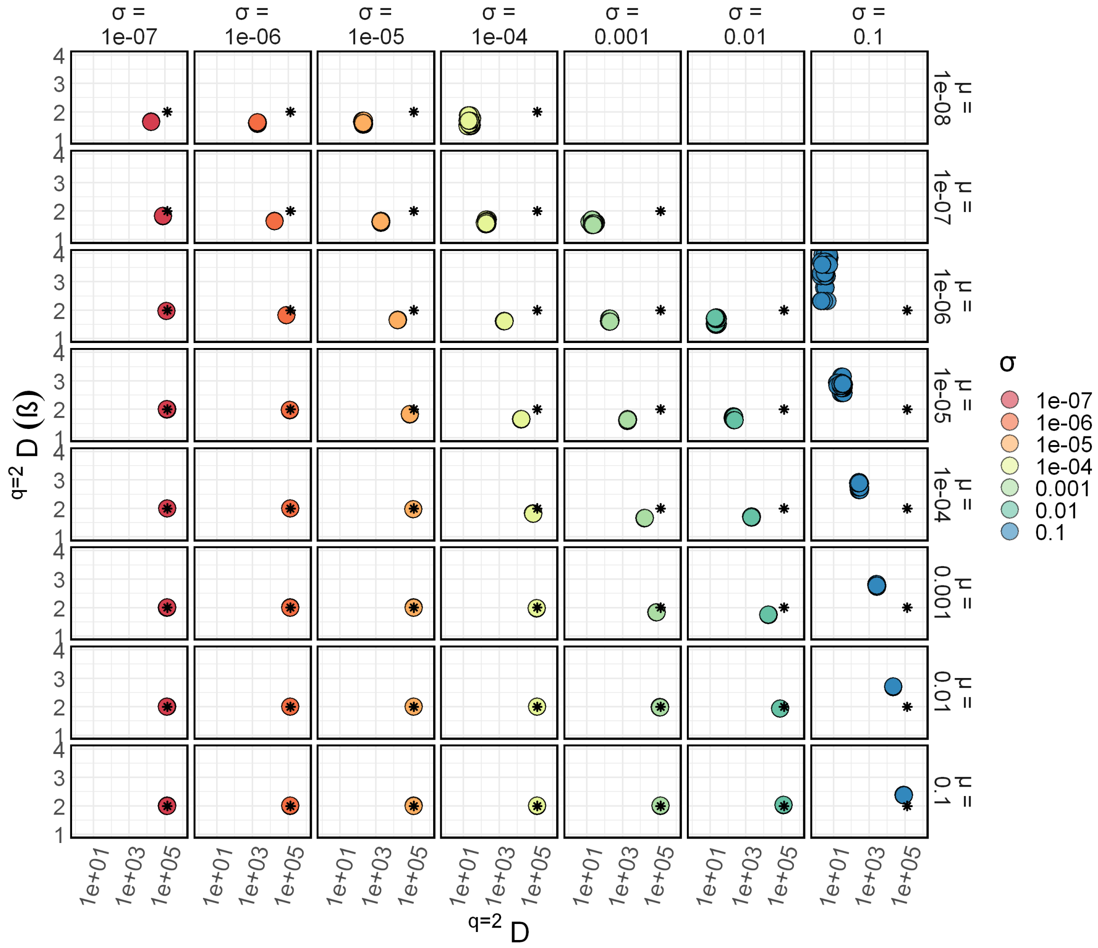


Figure 3.15: ${}^{q}D$ (within-replicate lineage diversity) vs ${}^{q}D(\beta)$ (between-replicate lineage diversity dissimilarity) of order $q = 2$ for the combined simulation's Passage 2 drug-treatment (DT_P2) replicates. Any given panel corresponds to a simulation set that was run using the combination of parameters that control the resistant phenotype's evolution: rows are different values of μ, whilst columns and the colour of points correspond to different values of σ. The black points in the bottom RHS of each individual panel correspond to the combined simulations' control treatments' mean values.

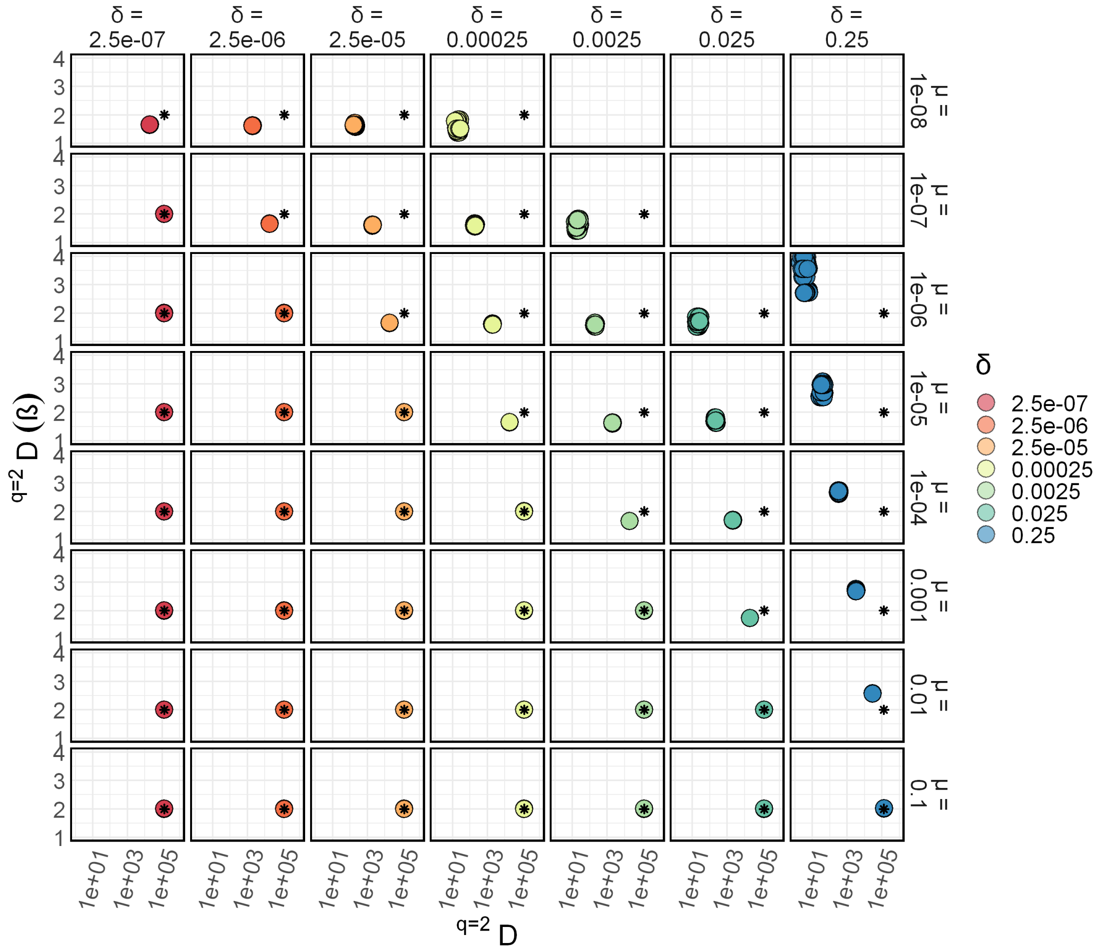


Figure 3.16: ${}^{q}D$ (within-replicate lineage diversity) vs ${}^{q}D(\beta)$ (between-replicate lineage diversity dissimilarity) of order $q = 2$ for the combined simulation's Passage 2 drug-treatment (DT_P2) replicates. Any given panel corresponds to a simulation set that was run using the combination of parameters that control the resistant phenotype's evolution: rows are different values of μ, whilst columns and the colour of points correspond to different values of δ. The black points in the bottom RHS of each individual panel correspond to the combined simulations' control treatment mean values.

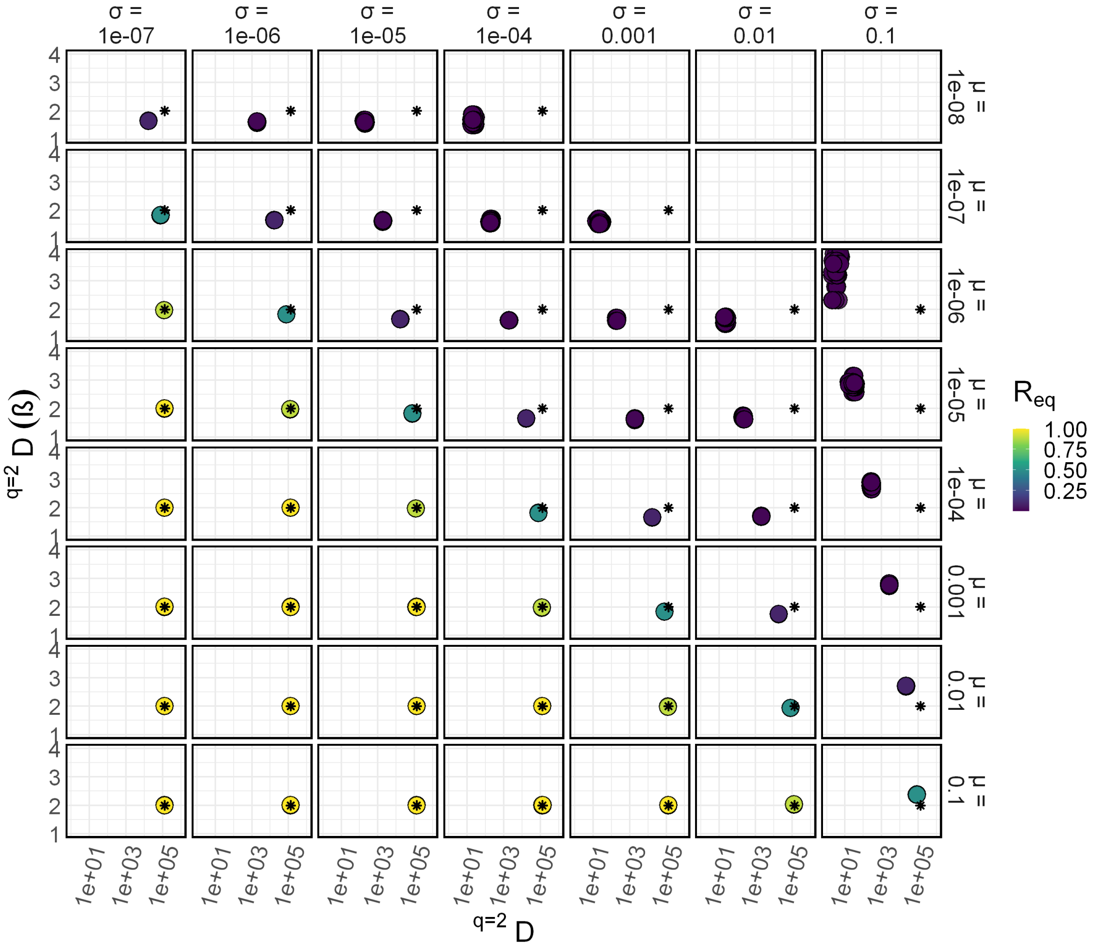


Figure 3.17: ${}^{q}D$ (within-replicate lineage diversity) vs ${}^{q}D(\beta)$ (between-replicate lineage diversity dissimilarity) of order $q = 2$ for the combined simulation's Passage 2 drug-treatment (DT_P2) replicates. Any given panel corresponds to a simulation set that was run using the combination of parameters that control the resistant phenotype's evolution: rows are different values of μ, whilst columns correspond to different values of σ. The colour of points are a given parameter combination's equilibrium frequency of resistance, R_{eq} $(0.0 - 1.0)$. The black points in the bottom RHS of each individual panel correspond to the combined simulations' control treatment mean values.

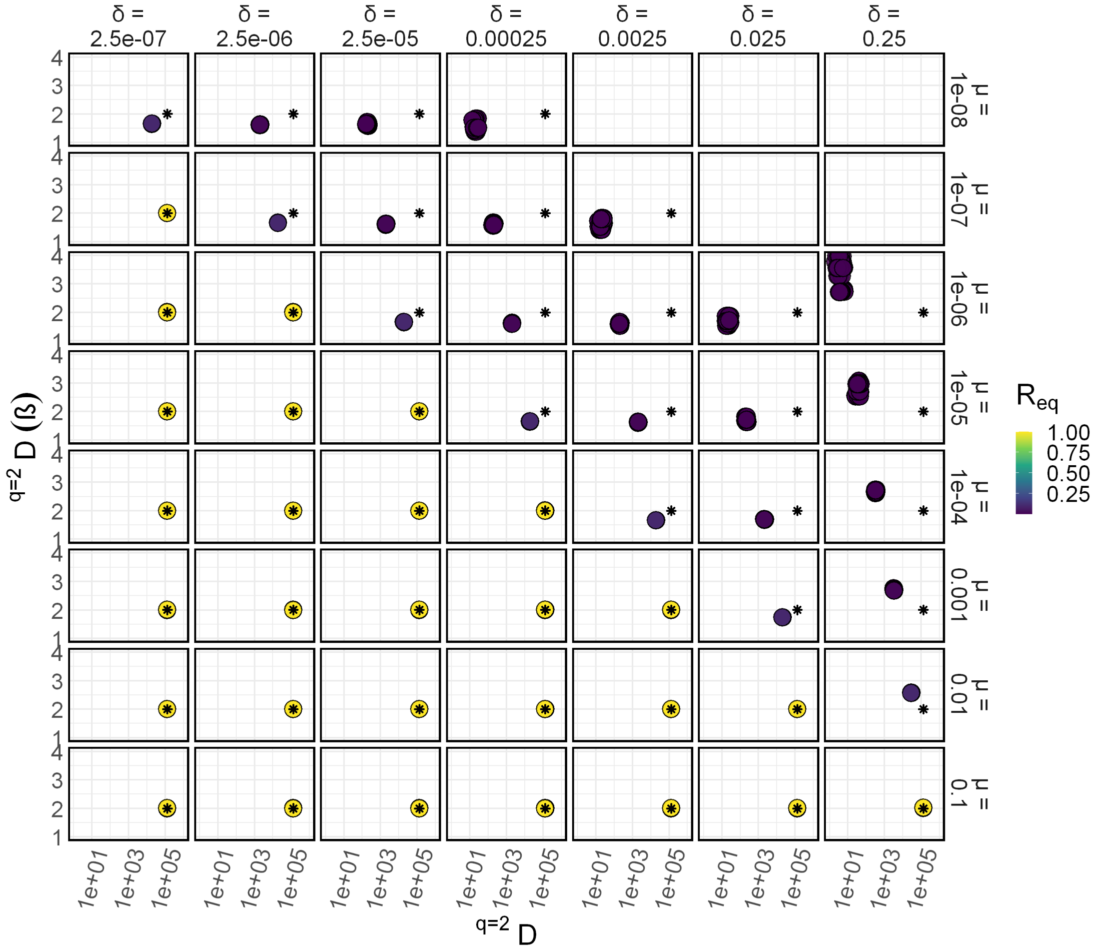


Figure 3.18: ^qD (within-replicate lineage diversity) vs $^qD(\beta)$ of order $q = 2$ for the combined simulation's Passage 2 drug-treatment (DT_P2) replicates. Any given panel corresponds to a simulation set that was run using the combination of parameters that control the resistant phenotype's evolution: rows are different values of μ, whilst columns correspond to different values of δ. The colour of points are a given parameter combination's equilibrium frequency of resistance, R_{eq} $(0.0 - 1.0)$. The black points in the bottom RHS of each individual panel correspond to the combined simulations' control treatment mean values.

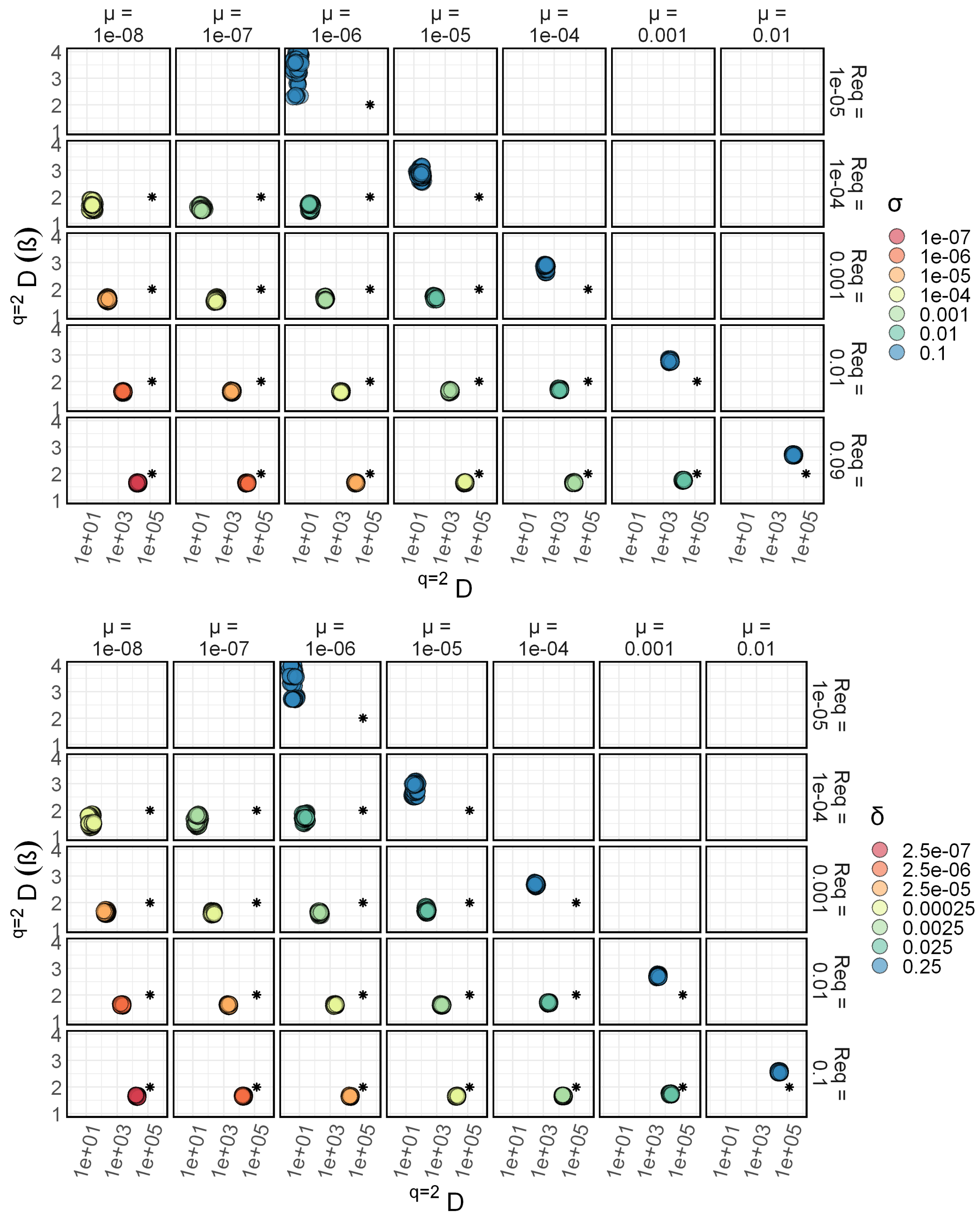


Figure 3.19: ${}^{q}D$ (within-replicate lineage diversity) vs ${}^{q}D(\beta)$ (between-replicate lineage diversity dissimilarity) of order $q = 2$ for the combined simulation's Passage 2 drug-treatment (DT_P2) replicates. Any given panel corresponds to a simulation set that was run using the combination of parameters that control the resistant phenotype's evolution: columns correspond to different values of μ, whilst rows correspond to values of σ (top panels) or δ (bottom panels) that lead to the same equilibrium frequencies of resistance (R_{eq}). The colour of points correspond to a simulation's values of σ and δ. The black points in the bottom RHS of each individual panel correspond to the combined simulations' control treatment mean values.

combinations that lead to smaller equilibrium frequencies (moving from the top to the bottom of Figure 3.19) lose more lineages to treatment, captured by smaller values of within-replicate ${}^{q}D$. At lower values of σ and δ, the differences between evolutionary scenarios (different panels in Figure 3.19) appear to be negligible; the dynamics in these cases are dominated by the lineages that were assigned pre-existing resistance according to the equilibrium frequency, R_{eq}. The low values of μ, σ and δ mean the probability of the phenotype switching per-division is low. The correlation between phenotype and lineage identity assigned at $t = 0$ remains high. Resistance is therefore shared amongst replicates following the mutual expansion step. The levels of dissimilarity between lineages following treatment therefore remain low (Figure 3.19).

Even when comparing evolutionary scenarios with identical equilibrium frequencies of resistance, some differences in values of ${}^{q}D$ vs ${}^{q}D(\beta)$ do emerge. Specifically, these differences are largest when R_{eq} is small, and values of σ or δ are high. These differences follow our expectations if we consider how these values influence the distribution of resistance amongst lineages during the shared expansion step. For high values of σ, the high number of cells that transition from resistant to sensitive per-division mean the proportion of the population's cells that are resistant become distributed widely amongst all extant lineages. For high values of δ, the fitness cost paid by cells bearing the resistant phenotype mean here, too, resistant is spread widely amongst lineages. To highlight these differences, Figures 3.20 and 3.21 show simulations where values of μ and δ (3.20) or μ and σ (3.21) lead to the same equilibrium values of resistance (R_{eq}). The top panel in each figure's two scenarios correspond to the distribution of resistant cells amongst all lineages that contain at least one resistant cell. Here, it is clear that, whilst each group has a similar proportion of resistance in the entire expanded population (LHS relative frequency columns in each plot), in the scenario where either δ or σ is high (top three panels in each figure), resistant cells are spread more evenly amongst numerous lineages (n.b. the log-scale for the counts). As such, it is much more likely that the resistant lineages sampled into each example replicate ('Replicate 1 & 2' in Figures 3.20 and 3.21) are unique to each replicate. Alternatively, if the phenotypic transition rates or fitness cost are low (the bottom three panels in each figure), a few lineages consist of all of the resistant cells in the expanded pool, and the probability of these same lineages being sampled into each replicate is high.

In the scenarios where resistance is spread evenly amongst different lineages, following drug-treatment, there is a high probability different resistant lineages now become successful in each replicate, engendering between-replicate dissimilarity in lineage diversity (high ${}^{q}D(\beta)$ - e.g. panels where $\sigma = 0.1$ and $\delta = 0.25$ in Figure 3.19). In the scenarios with high values of σ and δ *and* high values of μ ($\geq 10^{-5}$), the high equilibrium frequencies of resistance mean there are many resistant lineages consisting of numerous cells. This renders the evolutionary dynamics highly repeatable. However, as μ decreases ($\mu < 10^{-5}$) high values of σ and δ lead to lower equilibrium frequencies (R_{eq}). The dynamics become more stochastic, as illustrated by the increasing variance in values of ${}^{q}D(\beta)$: a resistant mutation occurring early in the shared expansion stage will have a higher probability of being sampled into multiple replicates, which will increase between-replicate similarity (and decrease ${}^{q}D(\beta)$) post-treatment.

3.8 Non-Deterministic Drug-Induced Death

So far I have discussed models where drug-treatment is simulated as an instantaneous, deterministic process: resistant cells ($\mathcal{R} = 1.0$) are killed with probability 0.0, and sensitive cells ($\mathcal{R} = 0.0$) are killed with probability 1.0. One drawback of this model is that it is impossible to recreate the growth kinetics of the corresponding *in vitro* experiment. That is, when death is deterministic, the selection for the resistant phenotype occurs exclusively during the first drug-kill step. Subsequent applications of the drug (*in silico*) only kill additional cells in the rare cases where σ is high enough that enough cells have reverted to sensitivity in the interim. In reality, metronomic chemotherapy using concentrations adopted in the *in vitro* experiment leads to observable cell death for numerous additions of treatment, not just the first.

A remedy to this feature of the model is to introduce variability into the process where cells are killed by drug-treatment. This could operate via two possible mechanisms: in the first, the resistance phenotype could be converted into a continuous trait, where the probability of death due to drug-treatment is now proportional to the

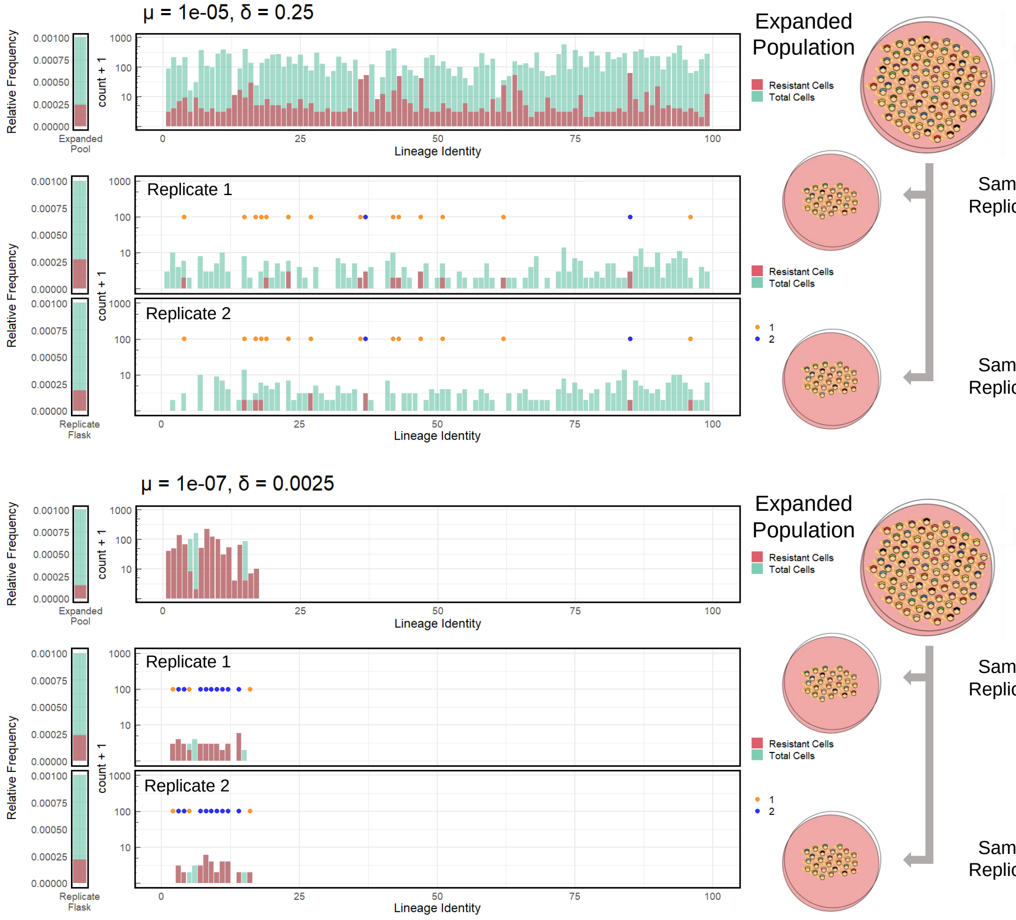


Figure 3.20: Simulated distributions of resistant and total cells in lineages in i) an expanded pool (top panel in each scenario), and ii) two sub-sampled replicates (bottom two panels in each scenario) - for two evolutionary scenarios: either $\mu = 10^{-5}$ and $\delta = 0.25$, or $\mu = 10^{-7}$ and $\delta = 0.0025$. Only lineages that have at least one resistant cell in the expanded population are shown. The colour proportion of each bar correspond to the total number of cells (green) and number of resistant cells (red). The frequency of the total population that is resistant is shown on the LHS of each panel, whilst the RHS shows the proportion of individual resistant lineages that express the resistant phenotype. Lineages where resistant cells are sampled into at least one replicate are highlighted by those that are either unique (orange points) or shared in each replicate (blue points).

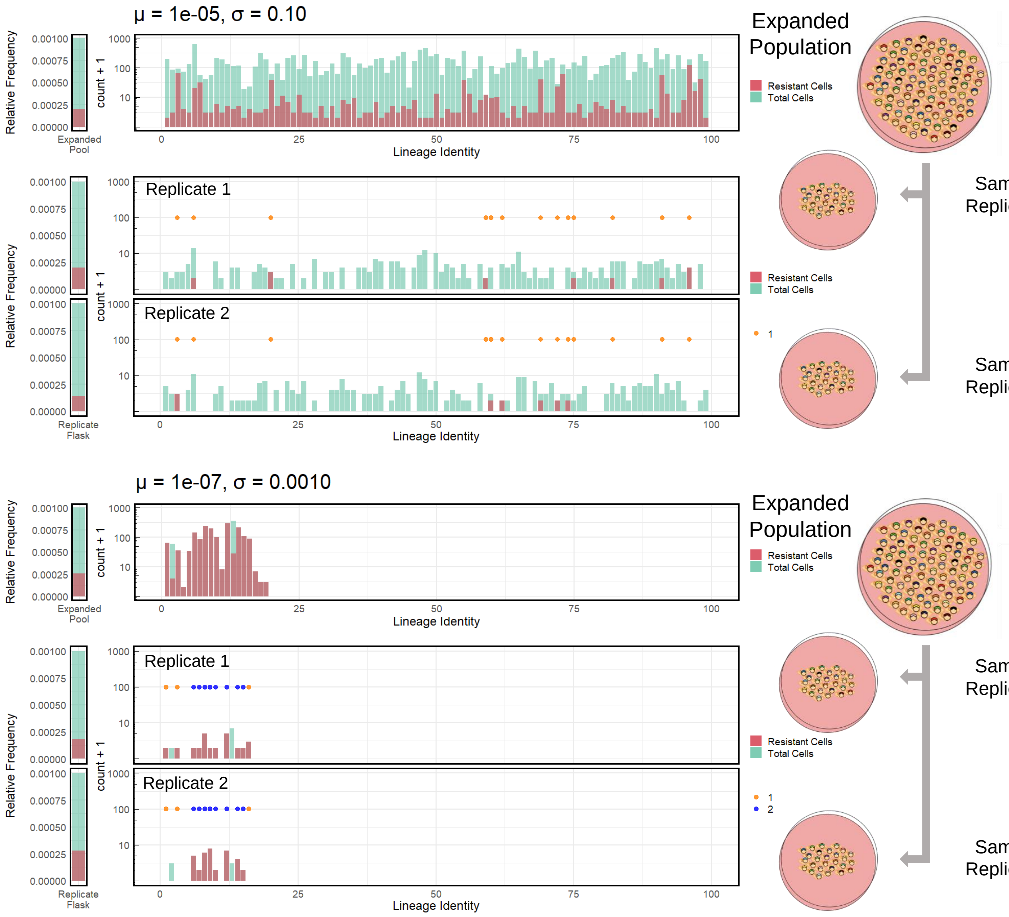


Figure 3.21: Simulated distributions of resistant and total cells in lineages in i) an expanded pool (top panel in each scenario), and ii) two sub-sampled replicates (bottom two panels in each scenario) - for two evolutionary scenarios: either $\mu = 10^{-5}$ and $\sigma = 0.10$, or $\mu = 10^{-7}$ and $\sigma = 0.0010$. Only lineages that have at least one resistant cell in the expanded population are shown. The colour proportion of each bar correspond to the total number of cells (green) and number of resistant cells (red). The frequency of the total population that is resistant is shown on the LHS of each panel, whilst the RHS shows the proportion of individual resistant lineages that express the resistant phenotype. Lineages where resistant cells are sampled into at least one replicate are highlighted by those that are either unique (orange points) or shared in each replicate (blue points).

‘strength’ of a cell’s resistance; the second possible approach could instead assume that there is some underlying stochasticity in whether drug-treatment kills *any* cell, whilst the resistance phenotype decreases this probability by some fraction. I choose to implement the second of these two mechanisms. It permits a range of hypotheses regarding how the phenotype governs resistance with the addition of just a single parameter: at one end, cells are equally likely to be killed irrespective of the chosen phenotype I choose to label ‘resistance’. At the other extreme, the stochasticity could be tuned to 0.0, and the model reverts to the deterministic approach. Furthermore, this approach limits the model complexity to one extra parameter.

Had I chosen the first of the two possible mechanisms, choosing how a population’s resistant phenotypes are distributed amongst individuals would be contingent on several further biological assumptions. I argue the model remains more tractable if I continue to model the resistant phenotype as a binary trait. Finally, if we assume that drug-treatment exerts a strong selective pressure, the demarcation of cells as resistant or sensitive becomes more appropriate as the proportion of mildly resistant cells that might have otherwise survived decreases.

3.8.1 Modelling Stochastic Drug-Induced Death

To model the stochastic component of drug-induced death, I introduce the parameter ψ. Drug-treatment is still imposed as an instantaneous event in the drug-treatment replicates. However, opposed to resistant cells being killed with probability 0.0, and sensitive cells being killed with probability 1.0, resistant cells are now killed with probability $(0.0 + \psi)$ and sensitive cells are killed with probability $(1.0 - \psi)$, where $\psi \in [0.0, 0.5]$. When $\psi = 0.0$, the simulation reverts to the deterministic model. When $\psi = 0.5$, cells are killed with probability 0.5, irrespective of their phenotype. As such, I can tune the contribution that a cell’s resistant phenotype contributes to the chance of survival during drug-treatment by varying ψ.

The following results were derived by setting $\psi = 0.3$. The remaining simulation parameters are unchanged form the previous results.

By introducing a probability that sensitive cells can survive a drug-treatment step -

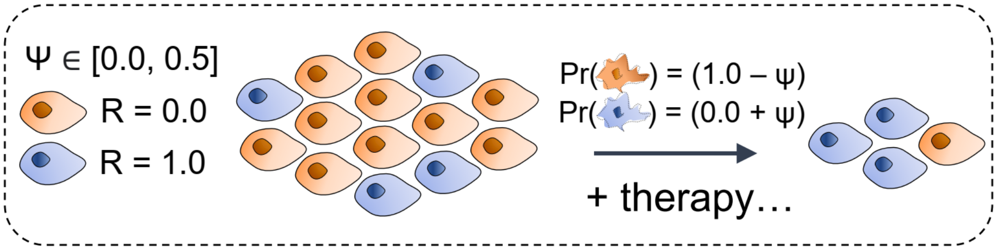


Figure 3.22: A schematic outlining the implementation of the stochastic component of drug-induced death. Sensitive cells are shown in orange ($\mathcal{R} = 0.0$) and resistant cells in blue ($\mathcal{R} = 1.0$). Drug-killing is determined by a cell's phenotype and the parameter ψ.

$\psi = 0.3$ - the number of cells that survive the initial pulse are now more similar across parameter ranges (Figure 3.23): the proportion of cells that survive the first treatment step is no longer just a product of R_{eq} (the equilibrium frequency of resistance), but is now $(R_{eq}(1 - \psi) + (1 - R_{eq})\psi)$ (as illustrated in Figure 3.22). As before, the time taken for all cells to fill each flask is protracted in scenarios where the probability of reverting to the sensitive phenotype is high (e.g. $\sigma = 0.1$) or the relative fitness cost of the resistant phenotype is high (e.g. $\delta = 0.25$). However, the replicates now take longer to fill each 'flask' as, even when all cells are resistant, each treatment step has a chance to kill resistant cells with probability ψ.

Despite the relatively generous introduction of a stochastic element into the drug-kill steps ($\psi = 0.3$), the majority of evolutionary scenarios lead to highly similar results (Figure 3.19 and Figure 3.24). It appears in most cases within- and between-replicate diversity is primarily governed by the distribution of resistance amongst lineages at the time of sampling the expanded cells into the respective replicates, and relaxing the stringency with which sensitive cells are killed by treatment has little impact on ${}^{q}D$ and ${}^{q}D(\beta)$. However, unlike the deterministic results, there are now scenarios where an equilibrium frequency of $R_{eq} = 0.0$ (no resistant cells are present at $t = 0.0$) that lead to simulation iterations that survive drug treatment (Figure 3.24). Here, as illustrated

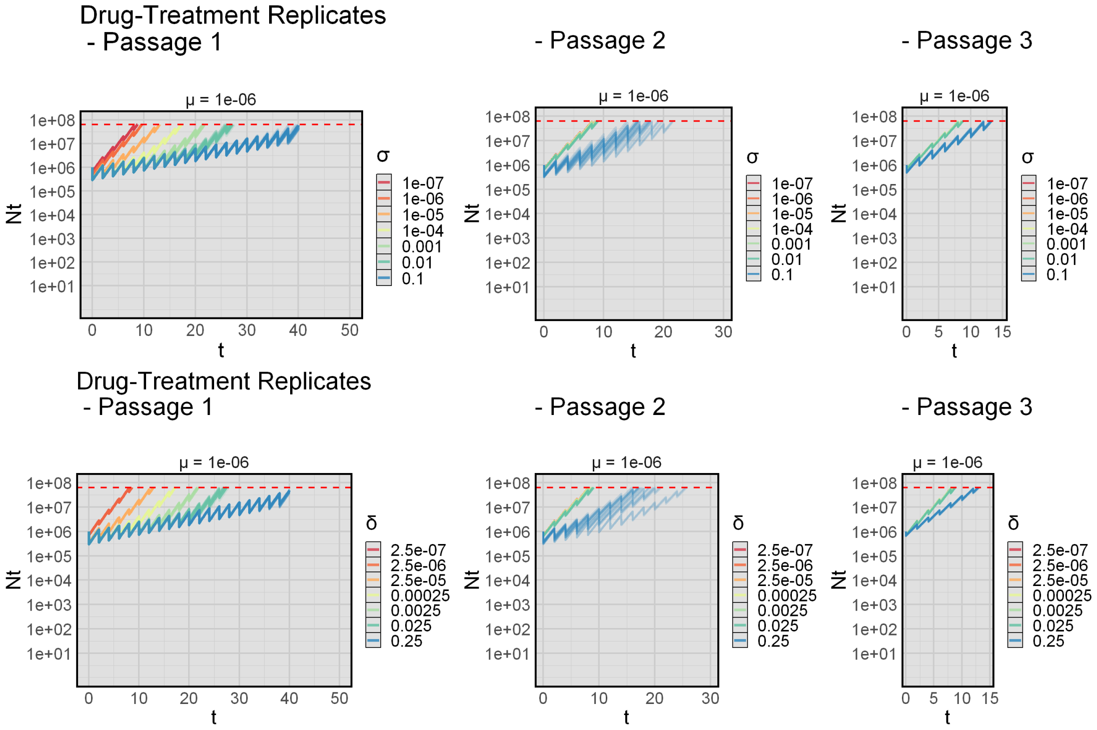


Figure 3.23: Example drug-treatment population trajectories for simulations where values of $\mu = 10^{-6}$ where $\psi = 0.3$ - following assignment to replicates, the total population size of each replicate sub-population was recorded at regular intervals throughout the simulation. Non-genetic sources of phenotypic variability simulations ($\sigma > 0.0$ - top row) and cost of resistance simulations ($\delta > 0.0$ - bottom row) where colour correspond to different values of σ and δ. Columns correspond to simulated passage number. The carrying capacity of the sub-populations has been marked in each panel (red-dashed line).

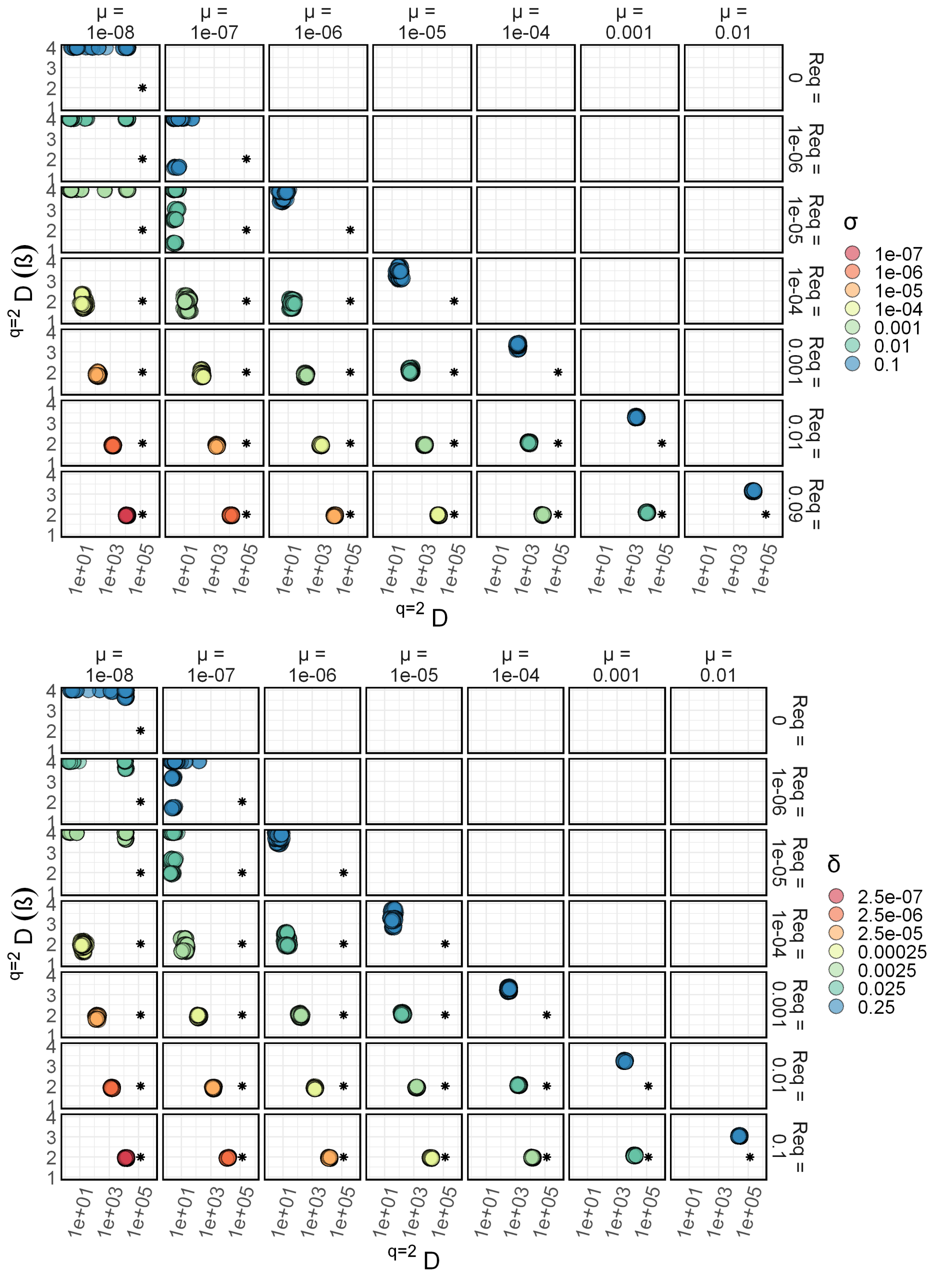


Figure 3.24: ${}^{q}D$ (within-replicate lineage diversity) vs ${}^{q}D(\beta)$ (between-replicate lineage diversity dissimilarity) of order $q = 2$ for the combined simulation's Passage 2 drug-treatment (DT_P2) replicates where $\psi = 0.3$. Any given panel corresponds to a simulation set that was run using the combination of parameters that control the resistant phenotype's evolution: columns correspond to different values of μ, whilst rows correspond to values of σ (top panels) or δ (bottom panels) that lead to the same equilibrium frequencies of resistance (R_{eq}). The colour of points correspond to a simulation's values of σ and δ. The black points in the bottom RHS of each individual panel correspond to the combined simulations' control treatment mean values.

by the high variance in values of ${}^qD$ and ${}^qD(\beta)$ the dynamics are highly stochastic; differences in lineage distributions are contingent on the emergence time of resistance mutations. The probability that sensitive cells can survive the drug-treatment steps now provides an additional window of opportunity following the first treatment stage for cells to transition to the resistant phenotype (controlled by μ). In some cases, the resistant mutation occurs early in the mutual expansion stage and is subsequently sampled into all replicate flasks, leading to high levels of between-replicate lineage similarity and correspondingly low values of ${}^qD(\beta)$. These early occurring resistance mutations also come to dominate the flask, leading to low within-flask diversity (${}^qD$). However, in the replicates with late-emerging mutations, they occur late in the expansion stage or within a flask and, therefore, the chance of being shared amongst replicates is highly unlikely. Furthermore, in these late-emerging replicates, numerous sensitive cells have been able to reach appreciable frequencies and contribute to the final within-replicate diversity. These dynamics lead to high levels of between-replicate dissimilarity (captured by high ${}^qD(\beta)$) and relatively high values of within-flask diversity (${}^qD$).

In these highly stochastic scenarios (top left panels in Figure 3.24), it is worth noting that the dynamics constrain the possible outputs in summary-statistic space (${}^qD$ vs ${}^qD(\beta)$). It is possible to have high within-replicate diversity (${}^qD$) but also high between-replicate differences (${}^qD(\beta)$) (top RHS of statistic space) if mutations occur late; replicates can also have low within population diversity and low within replicate diversity if the mutations occur early (bottom LHS of statistic space); however, it is *not* possible to have many, shared lineages after the treatment bottlenecks (bottom RHS of statistic space).

3.9 Discussion

The results in this chapter have built theoretical expectations for lineage distributions under a diverse array of evolutionary scenarios. These different scenarios are encoded in the different switching rates between phenotypes (μ and σ), or the fitness cost incurred by the resistant phenotype (δ). Previous studies that have simulated the evolution of drug resistance during an *in vitro* experiment have often assumed some pre-existing

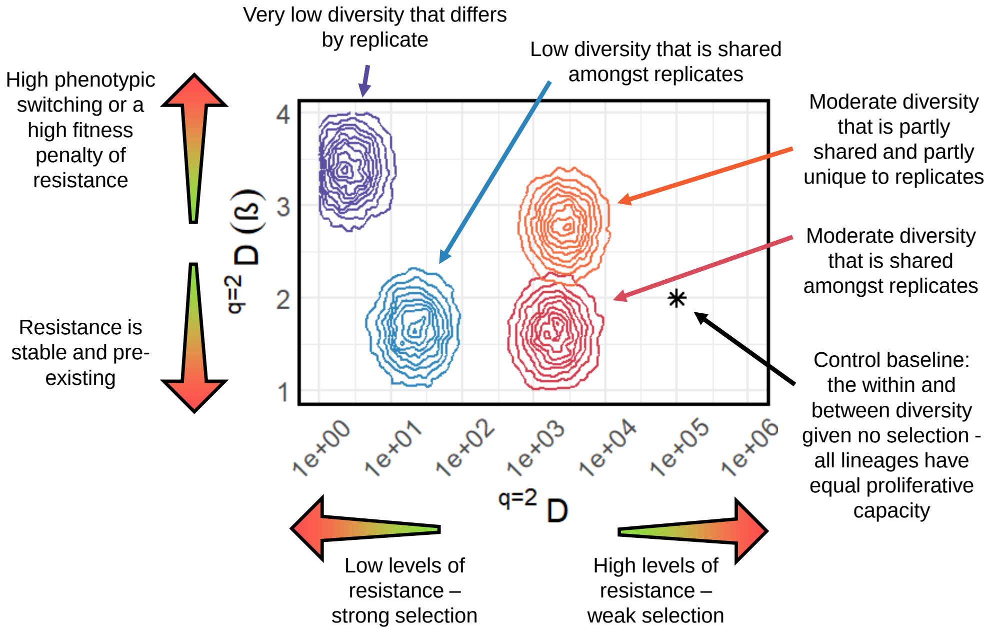


Figure 3.25: An illustration of how ${}^{q}D$ and ${}^{q}D(\beta)$ (within- and between-replicate diversity, respectively) change depending on the distribution of lineages amongst replicate flasks (smaller coloured arrows and density regions). In turn, the large arrows adjacent to the x and y axes illustrate how different mechanisms of phenotypic evolution lead to different values of ${}^{q}D$ and ${}^{q}D(\beta)$.

probability of resistance and some *de novo* mutation rate for mutations that confer resistance. However, these notions assume that the population is gradually approaching a state of pure resistance. They therefore have important consequences when modelling lineages' response to treatment *in vitro*: if cells are only ever transitioning from a sensitive to a resistant phenotype, the dynamics that emerge are contingent on the cell divisions experienced prior to the experiment. However, this assumption is rarely referenced, and the expansion time prior to the experiment therefore receives little attention (for example, Oren et al. 2021; Bhang et al. 2015).

Here, I have implemented a model that closely mirrors the design of my *in vitro* experiment. Furthermore, where equilibrium frequencies of a resistant phenotype are assigned to cells at the experiment's beginning, determined by the parameter combination in question. In many cases, where the rates are too low to lead to changes in phenotype prior to treatment, these dynamics are indistinguishable from the simpler, 'pre-existing resistance' model: lineages which succeed during treatment are those that were resistant at the time of lineage tagging. However, the proportion of resistant cells assigned at the start are now no longer arbitrary, but rather the product of the biological processes acting to produce and maintain differences in the resistance phenotype.

Scenarios where forces act to maintain the resistant phenotype at low frequencies - either a high phenotypic switching reversion rate to sensitivity, or a high fitness cost - leave distinct signatures in the combined within- and between-replicate lineage distributions: these dynamics mean each lineage has a low probability of harbouring numerous resistant cells. After sampling into separate replicate sub-populations and applying treatment, these dynamics are captured by high between-replicate differences. Such differences highlight the power of experimental evolution: namely, applying selection pressures on distinguishable, related individuals in parallel. Fortuitously, modes of resistance evolution highly similar to these readily discernible scenarios have been observed *in vitro*. High phenotypic switching rates from resistant to sensitive are analogous to transient phenotypic states highly refractory to treatment, as seen in *in vitro* models of melanoma (Shaffer, Dunagin, et al. 2017). Meanwhile high relative fitness costs of the resistant phenotype can mimic dormant, slow-cycling cells that have also been observed to provide a reservoir of treatment resistant cells (Russo et al. 2021; Oren et al. 2021).

By relaxing the assumption that death incurred due to drug-treatment is deterministic, the range of parameter values that are not lost to extinction increased. Surprisingly, the signatures left in the lineage distributions in the deterministic ($\psi = 0.0$) and partly stochastic case ($\psi = 0.3$) were strikingly similar. It appears that even given a generous probability of surviving each 'treatment step', repeated bouts of selection mean that the same dynamics emerge within and between replicates. In the partly stochastic case, sensitive cells that still survive therapy provide an additional window of opportunity for resistance to evolve. Specifically, the scenarios where the equilibrium frequency of resistance was previously too low to provide a sufficiently large number of resistant cells when the drug-treatment steps began can now potentially accrue resistant cells in the windows in between treatment steps. The stochastic nature of this subset of simulations is evident in the high variance of the two summary statistics, ${}^qD$ and ${}^qD(\beta)$.

Here I have chosen arbitrary birth and death rates (partly informed from previous studies) to observe the dynamics under various evolutionary scenarios. In Results Chapter 3, I develop a Bayesian model that leverages information in the non-treated replicates to infer the birth and death rates of my cell-lines. Subsequently, in Results Chapter 4, I will compare the theoretical expectations of the lineage distributions developed here to the observed, sequenced distributions derived from the *in vitro* experiment in my two chosen colorectal cancer cell-lines, *HCT116* and *SW620*.

Chapter 4

Results Chapter 2 - Optimising and Characterising Lineage Tracing and Drug-Treatment of Colorectal Cancer Cells In Vitro

4.1 Summary

Here I outline a variety of experiments undertaken to optimise the use of the barcoding lineage tracing technique *in vitro* as a means to investigate the evolutionary dynamics of drug resistance. As the complex barcode plasmid pool had to be expanded in-house, the behaviour of the lineage markers were assessed in my chosen colorectal cancer cell-lines (*HCT116* and *SW620*), whilst the plasmid pool itself was sequenced to a high depth to ensure complexity had been maintained. The models I have developed to resemble the dynamics of barcoded cells *in vitro* (see Results Chapter 1 for more details) rely on some assumptions concerning the rate at which cells have a unique lineage marker incorporated, and the selective bottleneck experienced by cells in standard culture conditions. These assumptions are tested here, empirically. Finally, before inferring specific evolutionary scenarios by comparing sequenced with simulated lineage distributions, I draw conclusions about the cell populations' response to treatment by measuring the

robustness of treatment response over time spent *in vitro*, and by observing cells directly during metronomic chemotherapy exposure.

4.2 Characterising the Expanded ClonTracer Library

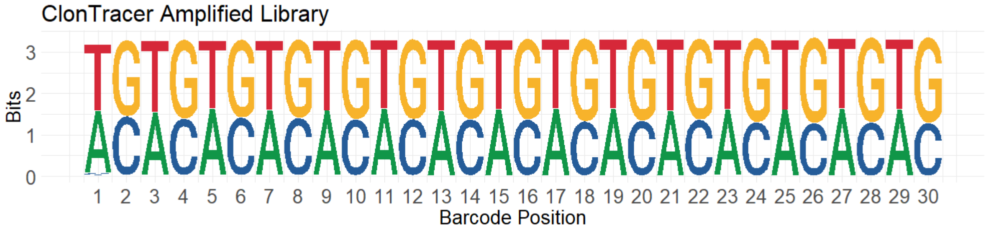


Figure 4.1: A nucleotide logo plot of the Expanded ClonTracer Library. The height of each nucleotide at each position corresponds to the frequency with which it was found in the pre-filtered sequenced barcode reads.

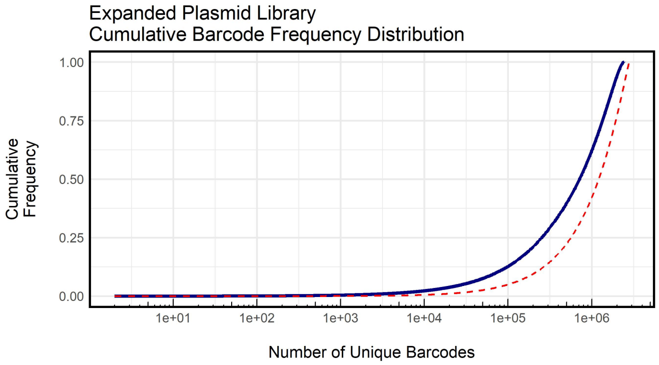


Figure 4.2: The cumulative frequency distribution of the unique lineages sequenced in the amplified ClonTracer Library. The red dashed line is the cumulative distribution given a hypothetical library with the same number of unique barcodes and identical ratios of each lineage.

After receiving the ClonTracer complex plasmid pool, it is necessary to expand the pool in electrocompetent bacteria to ensure there are enough barcode molecules for repeated rounds of viral infection (see Materials and Methods Chapter for more details). As this step may lead to an inadvertent bottleneck that reduces plasmid diversity, and this would alter the statistical properties of downstream lineage distributions, I sequenced the expanded plasmid pool to a high depth. Following sequencing, where approximately 10x10^6 barcode molecules were sequenced for a total of $\approx$ 140x10^6 reads, and subsequent barcode clustering to account for PCR amplification and sequencing errors, I identify $\approx$ 2.80x10^6 unique barcodes in the expanded plasmid pool, confirming that the library maintained a high barcode diversity. Figure 4.1 shows the distribution of nucleotides in the sequenced plasmid pool. The expanded library continues to adhere to the semi-random weak-strong pattern (`AT/CG`) that enables stringent downstream filtering of amplification and sequencing errors. The small proportion of barcodes that do not adhere to the pattern ($\approx 100,000$, or 5% of the total sequenced reads, not visible in the logo plot - the distribution shown is prior to filtering for the weak-strong pattern) is promising evidence that the majority of observed barcode reads are 'real'. Finally, there appears to have been very little bias introduced into the distribution of unique barcode lineages during the expansion process (Figure 4.2). Whilst due to the nature of the plasmid library production a skew in lineages is unavoidable, Figure 4.2 shows that the expanded library does not deviate strongly from a hypothetical library where all unique lineages are found at identical frequencies.

4.3 Investigating Barcode Infection Rate Assumptions

To ensure that the majority of barcodes contain a single, unique barcode, a low multiplicity of infection (0.1) was adopted when infecting cells with the lentivirus. Assuming barcode integration follows a Poisson distribution (where $\lambda = 0.1$), over 95% of infected cells will contain a single barcode. To ensure that the optimisation experiments that set the multiplicity of infection had been successful, single cell colonies from each cell-line were isolated from the barcoded pool of cells and then were allowed to expand for several weeks. Figure 4.3 shows the relative frequency of reads that contained a unique barcode sequence in each of these expanded colony samples. The majority of single colony samples do contain one, dominant barcode, and these reads adhere to the semi-random nucleotide sequence expected in the ClonTracer barcodes. The large number of low frequency read that do not adhere to this pattern highlight the utility of this filtering step. Most of the reads that do not belong to the dominant barcode are found at frequencies too low to pass the full filtering steps used later: any read where $(x/J) * K < 1$, where x is the observed number of reads for a given barcode, J is the total reads for that sample, and K is the number of input cells, is removed. Finally, whilst cells were seeded in dilutions that aimed to ensure 1cell/well, and wells were checked by eye, there is still a small probability 'doublet' cells made it into wells. That is, more than a single cell was the founder of the expanded colony. Therefore, the number of multiple integrations inferred from figure 4.3 is likely an over-estimate.

4.4 Confirming Expected Barcode Behaviour under Simple Culture Conditions

Assuming that selection in the absence of treatment is weak, there are some simple assumptions we can make about how the lineage distributions should be distributed. Assuming all cells have similar proliferative potential, the birth-death process leads to a distribution of lineages where most barcodes will be found at a low frequency, whilst the distribution also has a characteristic long tail: a small minority of barcodes will grow to high frequencies. The small, simple experiment under standard culture conditions (experiment code 'OR') was undertaken to first ensure that the barcode distributions adhered to these expectations. Most lineages are found at very low frequencies, even after two passages *in vitro*. The distributions shift as expected, where time spent growing leads to a longer tail as some lineages grow to higher frequencies. The characteristic 'bump' seen near the low end of the distributions is the product of the cells being subject to a round of sampling after the growth period: a sub-sample of cells are used for DNA extraction and barcode amplification. This behaviour can also be found in simulated distributions subject to the same sampling processes. Selection in standard conditions does indeed appear to be weak: by Passage 2, the most abundant barcode is still only found at a frequency of 5x10^{-4}. As 10^6 cells were used as input for these

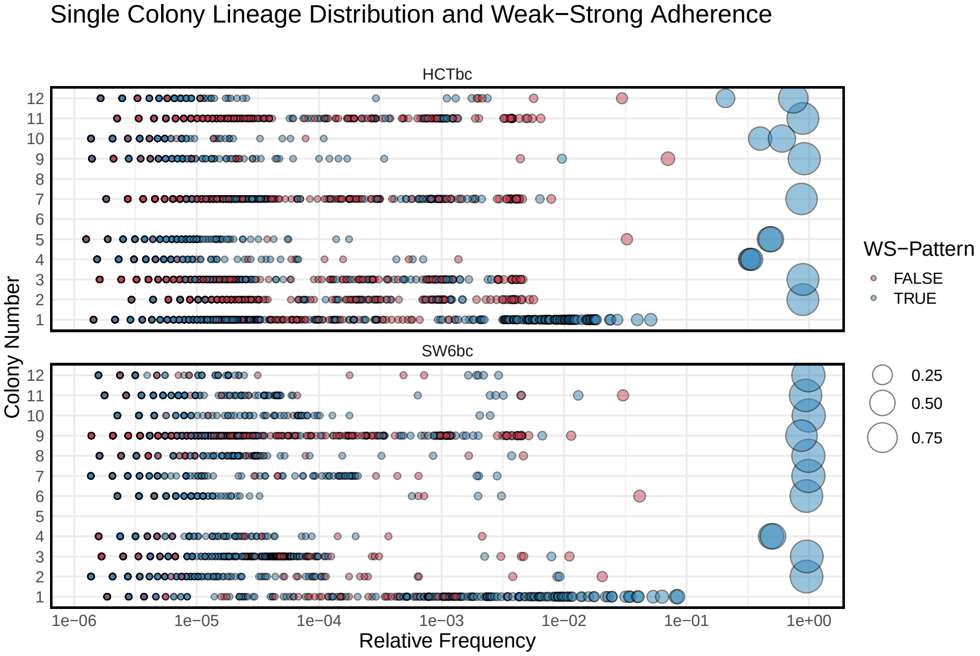


Figure 4.3: The read distributions of unique barcode sequences identified in expanded, single colony samples in two colorectal cancer cell-lines: `HCT116` (top panel) and `SW620` (bottom panel). Colony number 1 in each panel corresponds to a control well where approximately 100 cells were seeded. The x-axis position and size of the point corresponds to the relative frequency of the given barcode lineage, whilst the colour of each point corresponds to whether the given barcode sequence follows the ClonTracer semi-random weak-strong nucleotide pattern.

samples, this equates to only 500 cells. This high retention of diversity despite time in culture is critical to subsequently distinguish different responses to chemotherapy in the large, long-term drug treatment experiments.

4.5 Cell-Line Drug Response Stability In Vitro

One prediction of the model is that the cell populations used for the long-term experiment are close to equilibrium frequencies of the resistance phenotype. That is, given no selection for drug resistance, I expect the overall proportion of resistant cells to remain

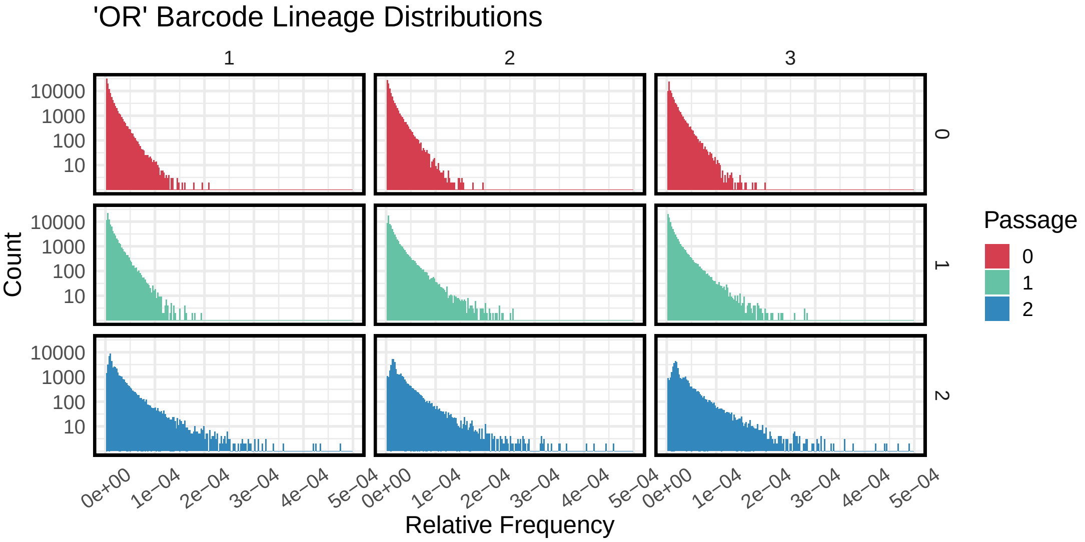


Figure 4.4: Barcode lineage distributions in `HCTbc` under standard culture conditions. Panel columns correspond to replicate number, whilst rows and colours correspond to passage number. Sequenced read counts have been filtered and normalised to relative frequencies.

roughly equal as time progresses, aside from stochastic fluctuations due to random drift. Figure 4.5 shows the IC$_{50}$ values for the chemotherapy drug 5-fluorouracil (5-Fu) of an early (`P4/P5`) and late (`P14/P15`) passage population of cells for each colorectal cell line (`HCT116` and `SW620`) used in the long-term experiment. Each cell-line's 50% inhibition concentration remains highly similar despite a long time in culture. As the selection pressure of my chosen drug-treatment was absent for the period of growth, we might expect the IC$_{50}$ values to gradually increase over time if mutations that were beneficial to standard culture conditions were pleiotropically linked to those that conferred resistance, or if the resistance phenotype was not costly and was subject to significant levels of drift in the time window observed. The relative stability of the IC$_{50}$ values supports the hypotheses that the proportion of resistant cells used for the initiation of the long-term evolution experiment has reached some phenotypic equilibrium, opposed to the cells being in the process of rapidly traversing some fitness landscape.

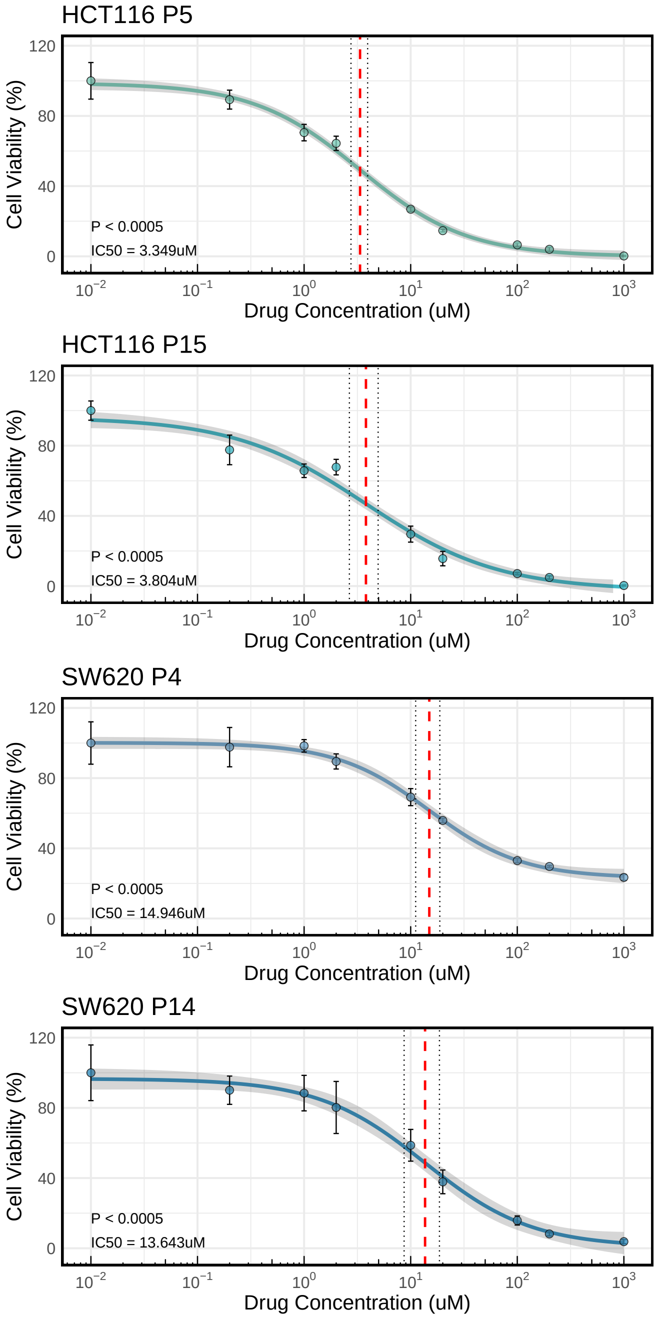


Figure 4.5: IC_{50} values for the drug 5-fluorouracil (5-Fu) in two colorectal cancer cell-lines: *HCT116* and *SW620* for early (P5 and P4) and late (P15 and P14) passage cell populations. The grey ribbon corresponds to the 95% confidence interval of the drug-response curve fit to the viability data. Error bars show +/-std.dev away from the observed viability mean per concentration. The estimated IC_{50} values are shown with the red dashed line and the 95% confidence interval by the black dotted lines.

4.6 Observable Culture Dynamics during Drug-Treatment In Vitro

Prior to sequencing barcodes from cultured cells that have been subjected to metronomic chemotherapy treatment *in vitro*, I observed their behaviour with bright-field microscropy. Even at this stage, there appeared to be differences in each cell-line's response to treatment. As cells are initially seeded at a relatively low density, after a period of time in culture, the position of cells on the bottom of a flask can be used as a coarse metric of relatedness: cells are adherent and grow out into adjacent, free space. Proximal individuals are therefore likely to be descendants of the same cell.

In `HCTbc`, the majority of cells died during the first few pulses of chemotherapy treatment. The remaining cells were larger and more irregular in shape (`HCTbc - DT early` in Figure 4.6) than their untreated counterparts (`HCTbc - CO` in Figure 4.6). Eventually, however, cells that resemble the untreated cells in the control replicates emerged in 'colony outgrowths'. These appear to emerge with surprisingly regularity (both between replicates within an experiment, and between experiments) and continue to populate the flask despite ongoing treatment.

In contrast, the `SW6bc` cells show a more uniform response to treatment: many cells appear following treatment with chemotherapy, however the regrowth in the recovery periods also appears more evenly distributed. The cells also appear more similar in appearance to their untreated contemporaries in the control replicates (`SW6bc - CO` vs `SW6bc - DT early and late` in Figure 4.6). This difference in each cell-line's response is also clear when observing the lineage trajectories for each experiment (Figure 4.6): the `HCTbc` drug-treatment replicates take much longer to populate most of the flask than the corresponding `SW6bc` drug-treatment replicates, as highlighted by the differences between Passage 0 and Passage 1 in Figure 4.6. The lineage trajectories also reveal slight differences between the two experiments. The drug-treatment replicates take longer to recover from treatment in the 'QR' Passage 1 than the corresponding samples in the 'PR' experiment. Although the aim was to maintain the conditions between the two experiments as consistent as possible, it is possible that on the time-scales necessary for resistance evolution stochastic differences led to a slightly more stringent bottleneck

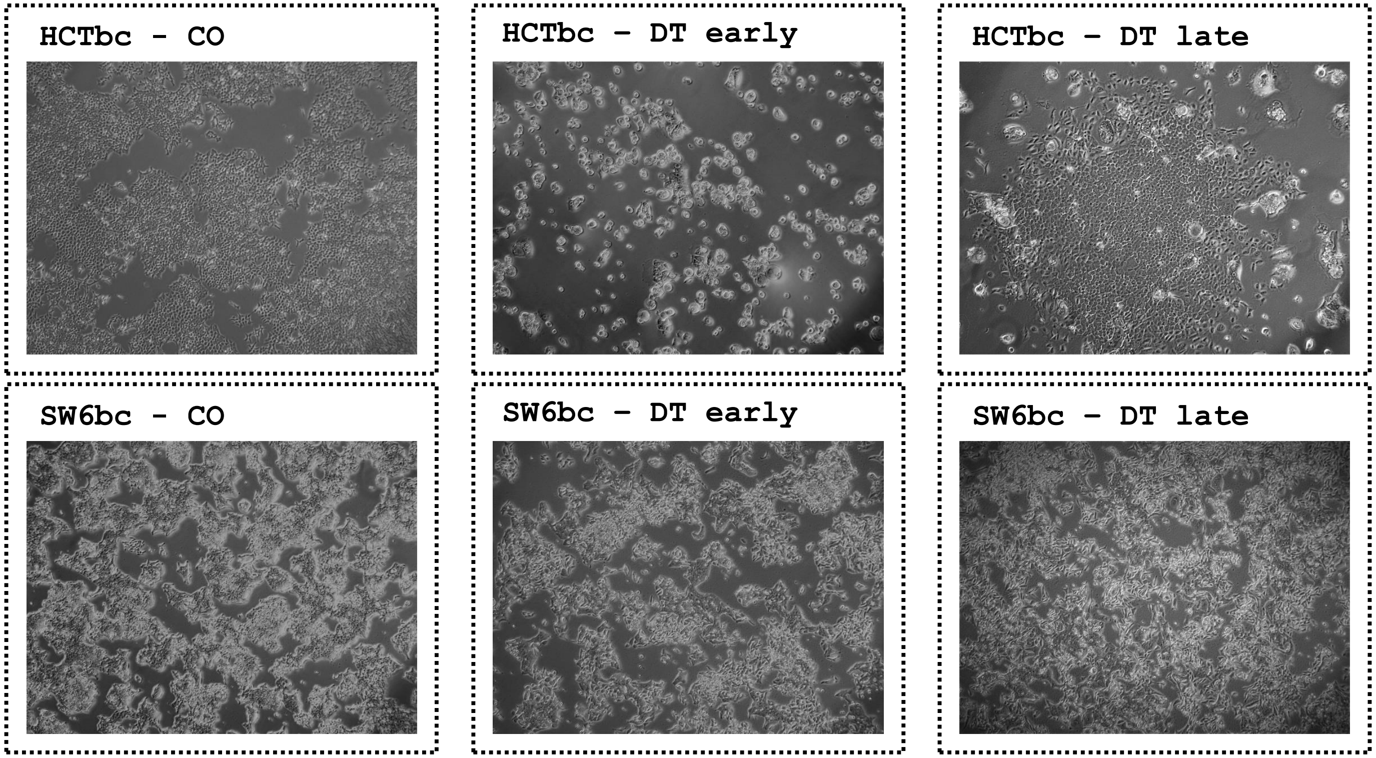


Figure 4.6: Bright-field tissue culture images taken from the ‘QR’ long-term drug-treatment experiment. Representative images are shown for late in the control treatment replicate, or early and late in the drug-treatment replicate, Passage 1, for each cell line, *HCT116* (top row) and *SW620* (bottom row).

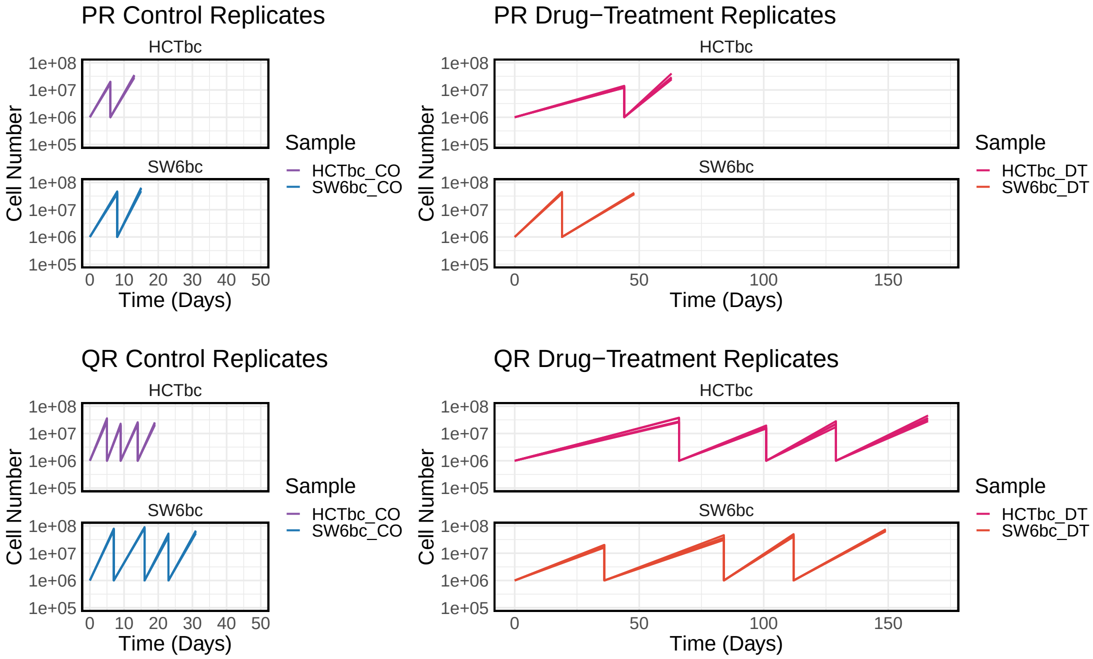


Figure 4.7: The total cell lineage trajectories for the two long-term drug-treatment experiments: ‘PR’ (for a total of 2 passages - top 4 panels) and ‘QR’ (for a total of 5 passages - bottom 4 panels).

in the second ‘QR’ experiment.

4.7 Barcode Clustering Method Comparisons

When amplifying and sequencing the semi-random nucleotide sequences (the barcodes), errors are introduced due to the incomplete fidelity of the polymerases used. As such, a crucial step in the bioinformatic pipeline when analysing the barcode sequences is a ‘clustering’ step, where statistical software attempts to ‘un-do’ the errors. As the adoption of lineage tracing experiments that employ barcoding technology have grown, several programs have been published that cluster sequences to produce a list of putative, true ‘parental’ barcodes without being computationally prohibitive (the high number of unique reads usually precludes calculating all pairwise distances between sequences).

Here, I compare three clustering methods: the software provided with the ClonTracer plasmid library, used in the original publication (Bhang et al. 2015), Bartender (L. Zhao et al. 2018) and Starcode (Zorita et al. 2015). The clustering methods were compared using simulated barcode reads: a distribution of barcodes was derived using the birth-death simulation, where $N_0 = 10^5$, $\Delta t = 6.0$, $b = 0.8$ and $d = 0.2$. The barcode identities were then converted into semi-random nucleotide sequences that adhered to the semi-random nucleotide pattern, with universal adjacent primer sequences. These nucleotide sequences then had errors introduced with rate $0.005/bp$ (a purposefully 'noisy' rate to assess the clustering methods with difficult sequences). Finally, these sequences were passed to a NGS simulator (ART: Huang et al. 2012) to produce FASTQ files with error profiles similar to those expected from an Illmina sequencer.

The ClonTracer method (herein 'ClonTracer') has no parameters that influence the clustering method. Therefore, only one clustered distribution was used for the comparison. Both Bartender and Starcode have a 'cutoff' parameter that removes any sequence found $< c$ times. This was set to 1 in both methods. These two methods also have a 'distance' parameter, d, that dictates the maximum number of mismatches between two putative clusters that can be merged. This was set to 2 in both methods. Finally, each method has a parameter that controls the stringency of clustering: in Bartender, this is controlled by z, where higher values result in more generous clustering (L. Zhao et al. 2018); in Starcode, this is controlled by r, where smaller values result in more generous clustering (Zorita et al. 2015).

The results of the clustering method comparisons are shown in Figures 4.8 and 4.9. Bartender and Starcode most faithfully reproduce the true frequencies of simulated lineage relative to ClonTracer. The ClonTracer method has the lowest dropout rate, but also calls the highest number of false positives. Overall, Bartender calls relatively low numbers of false positives and negatives. It also offers a high number of parameters to control the clustering process lacking in the ClonTracer software. As such, I choose to employ the Bartender method for all future clustering steps.

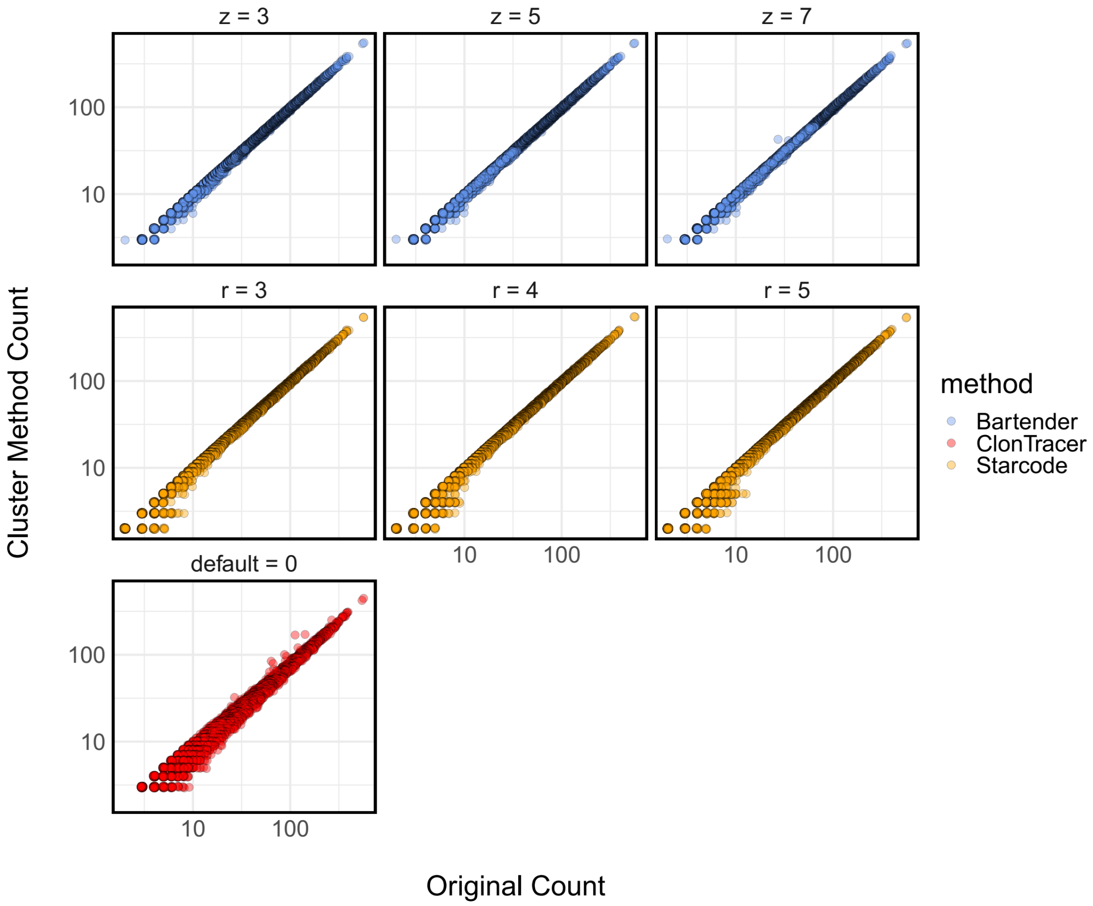


Figure 4.8: True, simulated counts (x-axis) vs estimated clustered counts (y-axis) for three clustering methods. Panels are labelled with the variable clustering parameter (n.b. the ClonTracer method has no user-defined variables).

4.8 Discussion

Here I have outlined some simple experiments that validate the adoption of the lentivirus lineage tracing technique as a suitable means to investigate drug resistance evolution *in vitro*. The modelling used to infer evolutionary scenarios in the long-term drug treatment experiments (discussed in the following section) are contingent on some assumptions concerning the statistical behaviour of the barcodes.

I have shown that the expanded plasmid pool retains a high level of diversity, ensuring I have enough unique barcode molecules to trace lineages *in vitro* at a high

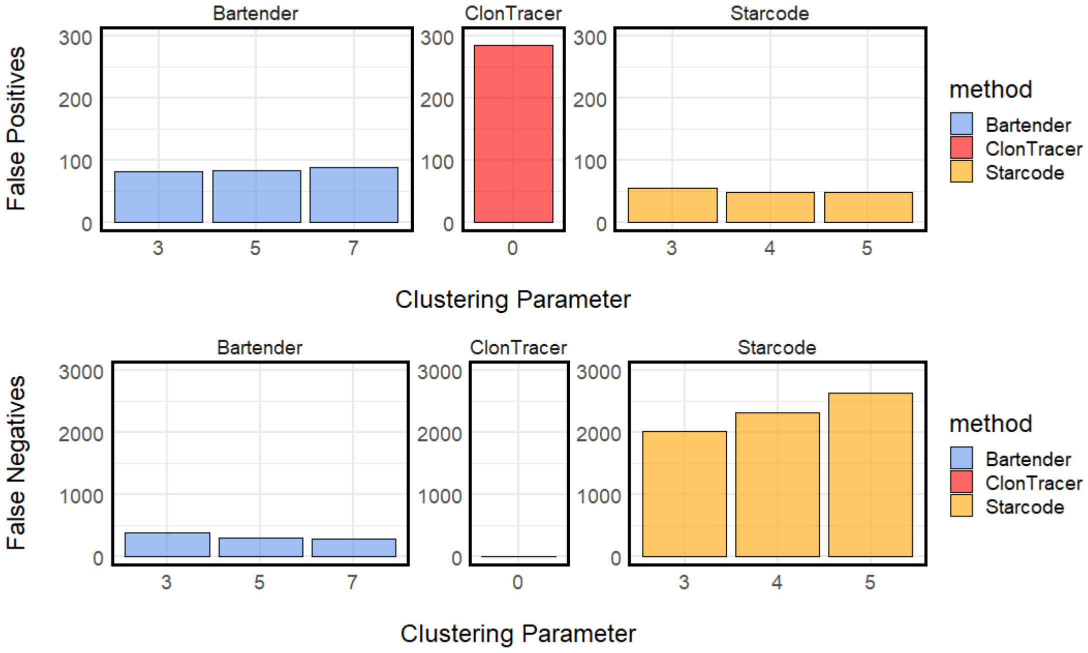


Figure 4.9: Comparisons of the number of false positives and false negatives for each of the clustering methods when used to cluster simulated barcode sequences. The x-axis correspond to user-defined clustering parameters: z in Bartender and r in Starcode (ClonTracer has no user-defined parameters).

resolution. Nearly all sequenced molecules also adhere to the semi-random weak-strong nucleotide pattern which assists in filtering out technical artefacts that accumulate during DNA extraction and barcode amplification, then sequencing. Most cells contain a single, dominant barcode post-infection. This minimises any additional statistical adjustments that might have been necessary had the infection led to multiple integrations per cell. I compare several barcode clustering techniques on simulated lineage distributions to ensure the chosen method maximises the concurrence between true and observable, sequenced lineage counts.

Following growth under simple culture conditions, the lineage distributions are consistent with theoretical predictions (as in Results Chapter 1): under weak selection, most lineages remain extremely rare, although as time progresses, some lineages reach high frequencies, whilst the overall number of lineages decreases as others are lost to drift. To observe the lineage distributions, the barcode construct must be extracted, amplified and sequenced on a flow-cell. The retention of high diversity even following these technical bottlenecks also validates the experimental pipeline as a means to track rare cell populations *in vitro*.

Chapter 5

Results Chapter 3 - Inferring Growth Rate Parameters with a Bayesian Noise Model

5.1 Introduction

Whilst the net growth rate of a growing population, $(b - d)$, is easy to derive by simply comparing the change in population size, $N_t - N_0$, in some given time window, Δt, the turnover of the population, $(b + d)$, is harder to infer. To faithfully recreate the growth dynamics of cells in simulations, estimates of these growth parameters are necessary. Estimating the death rate directly can require additional experiments that utilise, for example, live-imaging or FACS (Johnson et al. 2019; Russo et al. 2021). Here, I leverage information held in the initial expanded population of barcoded cells. Cells are uniquely barcoded, then subject to a mutual shared expansion step. Subsamples of this expanded pool are then processed and sequenced to assess the lineage distributions at $t = 0$ in the full long-term drug-treatment experiment (as illustrated in Figure 5.1. See Materials and Methods for full experimental schematic). The variance in the observed lineage distributions are a product of the birth-death process for a known Δt and subsequent technical bottlenecks (for barcode extraction, amplification and sequencing). Here, I develop a Bayesian model that accounts for these bottlenecks to independently recover the birth and death rates of each colorectal cancer cell-line *in vitro*; *HCT116* and *SW620*.

5.2 Birth-Death Process

First, I can make use of the *p.m.f.* for the birth-death process (Bailey 1990; Durrett 2015),

$$p_n(n) = (1-\alpha)(1-\beta)\cdot\beta^{(n-1)} \tag{5.1}$$

for $(n \geq 1)$, and

$$p_0(n) = \alpha \tag{5.2}$$

for $(n = 0)$, where

$$\alpha = \frac{d(e^{(b-d)t}-1)}{be^{(b-d)t}-d} \tag{5.3}$$

$$\beta = \frac{b(e^{(b-d)t}-1)}{be^{(b-d)t}-d} \tag{5.4}$$

Whilst the total number of cells following the mutual expansion step is known, N_t, this observed value is itself also a random variable that depends on $(b-d)$ (which dictates the mean of the total population, $\hat{N_t}$) and on $(b+d)$ (which dictates the variance of the total population, $\sigma^2_{N_t}$). I can therefore also make use of the following (Bailey 1990; Durrett 2015):

$$\hat{N_t} = N_0\cdot(e^{((b-d)\cdot t)}) \tag{5.5}$$

$$\sigma^2_{N_t} = N_0\cdot\frac{(b+d)}{(b-d)}\cdot e^{(b-d)\cdot t}\cdot(e^{(b-d)\cdot t}-1) \tag{5.6}$$

where N_0 is the number of cells at $t = 0$. If I could observe the barcode lineage frequencies directly, I could use a Bayesian model to infer b and d by jointly using the p_n

p.m.f. for each individual lineage, and assuming $N_t \sim Normal(\hat{N}_t, \sigma^2_{N_t})$. However, the lineage distribution is subject to two subsequent technical bottlenecks that influence the statistical properties of the observed, sequenced distributions.

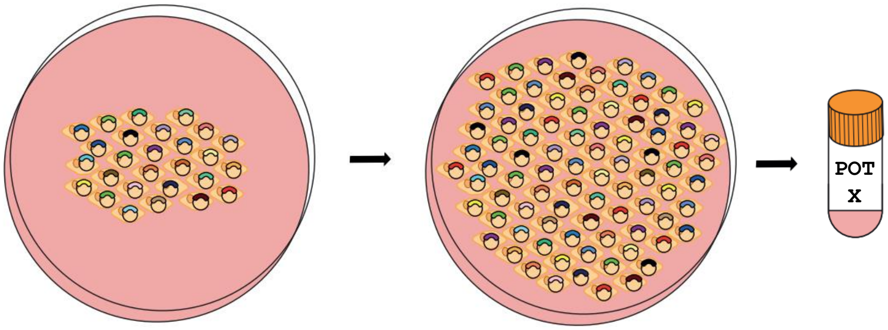


Figure 5.1: A schematic of the growth expansion and sampling of the `POT` samples used for the Bayesian growth parameter inference. Cells are barcoded, allowed to expand for a known Δt and then a fraction of these expanded cells, N_t, are sub-sampled (K) for barcode expansion and sequencing.

5.3 Sampling K Cells

The most stringent technical bottleneck is determined by the fact that I can not sequence all cells - I am limited by the amount of DNA it is feasible to extract and sequence. Instead, a sub-population of cells are sampled before their DNA is extracted, their barcode sequences amplified and finally sequenced. Whilst strictly sampling without replacement, this sampling step can be modelled as a Binomial sampling step, where, if I sample K cells from the N_t expanded cell pool, now the probability of seeing any lineage k times is

$$p(k) = \sum_{n=0}^{N_t} p(k|n) \cdot p(n) \tag{5.7}$$

where $p(n)$ is (5.2) if $n = 0$ and (5.1) if $n > 0$, and

$$p(k|n) \sim B(n = K, p = \frac{n}{N_t}) = \binom{K}{k}(\frac{n}{N_t})^k(1 - \frac{n}{N_t})^{K-k} \tag{5.8}$$

However, because $(\frac{n}{N_t}) << 1.0$, I can instead model this step as a Poisson distribution where the probability of seeing a any lineage k times (given it is at size n in the expanded cell pool of size N_t) becomes

$$p(k|n) \sim Pois(\lambda) = \frac{\lambda^k e^{-\lambda}}{k!} \tag{5.9}$$

where

$$\lambda = ((\frac{n}{Nt}) \cdot K) \tag{5.10}$$

If I want to determine the probability of seeing a any barcode lineage k times given the sampling K cells step, I can calculate the compound probability distribution by marginalising over n, as in 5.7.

So, now I have derived a p.m.f. for seeing any lineage represented k times in the total K sampled cells, given they were sampled from an expanded population of N_t cells, cells which all had birth and death rates b and d, respectively. As I also observe the total number of cells in the expanded pool - N_t - I can also simultaneously fit this as a free parameter, assuming $N_t \sim Normal(\hat{N}_t, \sigma^2_{N_t})$ (which relies on b and d - see 5.5 and 5.6). Yet an additional source of technical variance remains: the variance introduced when sampling the amplified barcode molecules on the flow-cell during sequencing.

5.4 Sampling J Reads

After amplifying and extraction, each cell's barcode is sequenced on a flow-cell. Whilst this process can be thought of sampling each barcode lineage $(\frac{k}{K} \cdot J)$ times, where k is the number of times the barcode lineage is represented in the pool of K sub-sampled cells (as per the previous section), and J is the total number of reads assigned to this sample on the flow-cell, again, because $(\frac{k}{K}) << 1.0$, we can model this as a second Poisson distribution, where

$$p(j|k) \sim Pois(\lambda) = \frac{\lambda^j e^{-\lambda}}{j!} \tag{5.11}$$

where

$$\lambda = ((\frac{k}{K}) \cdot J) \tag{5.12}$$

Whilst I could now construct a p.m.f. for observing j reads of a given barcode given some underlying birth and death rates (b and d), following which the ns and ks would be marginalised out, this nested marginalisation would not be computationally expedient. Therefore, as the variance introduced during a Poisson sampling step is known, I can introduce this into the p.m.f for $p(k)$ to account for the additional known noise introduced from the sequencing step.

5.5 Incorporating Sequencing Noise into Sampling K Cells

The expectation and variance of a Poisson where $\lambda = ((\frac{k}{K}) \cdot J)$ are

$$E[j|k] = var(j|k) = \lambda = (\frac{k}{K}) \cdot J \tag{5.13}$$

where I condition on k, that is, there being k cells of a given barcode lineage in the expanded K pool of cells. To calculate the additional noise, I must determine the variance of the compound distribution of sampling J barcode reads from K barcoded cells, assuming $p(j|k) \sim Pois((\frac{k}{K}) \cdot J)$. The variance of a compound distribution, accounting for all probabilities of $p(k)$, is

$$var(j) = E[var(j|k)] + var[E(j|k)] \tag{5.14}$$

(using the law of total variance), which becomes

$$var(j) = \sum_{k=0}^{K}(var(j|k) \cdot p(k)) + (\sum_{k=0}^{K}(E(j|k)^2 \cdot p(k)) - \sum_{k=0}^{K}(E(j|k) \cdot p(k))^2) \tag{5.15}$$

I can therefore calculate the variance introduced when sequencing J reads from a pool of K cells that themselves have been sampled from an expanded pool of N_t cells with birth and death rates b and d, respectively. This extra, *known* variance can then be introduced back into the p.m.f for sampling K cells by replacing the Poisson distribution with an over-dispersed Poisson; a Negative Binomial distribution (see bottom right-hand arrow in Figure 5.2).

The Negative Binomial can be parameterised in the following form:

$$p(k|n) \sim NegBinom(\mu, \phi) = \binom{k+\phi+1}{k}(\frac{\mu}{\mu+\phi})^k(\frac{\phi}{\mu+\phi})^\phi \tag{5.16}$$

where the first moment (the mean) is

$$E[k|n] = \mu = \frac{n}{N_t} \cdot K \tag{5.17}$$

and second moment (the variance) is

$$var(k|n) = \mu + \frac{(\mu)^2}{\phi} \tag{5.18}$$

As $\mu = (\frac{n}{N_t} \cdot K)$ is the mean of the standard Poisson, we can see that the *additional* variance in the over-dispersed Poisson (the Negative Binomial) is $\frac{(\mu)^2}{\phi}$. This additional variance can therefore be transformed into a value of ϕ for the negative binomial by rearranging (5.18) as follows:

$$\phi_n = \frac{\mu^2}{var(j|n) - \mu} \tag{5.19}$$

which can be thought of as a vector of ϕs for each possible n.

In summary, to introduce the additional noise added to K sampled cells when sequenced on a flow-cell to produce J sampled reads, the Poisson distribution $p(k|n) \sim Pois(\lambda = ((\frac{n}{Nt}) \cdot K))$ can be replaced with a Negative-Binomial, $p(j|n) \sim NegBinom(\mu = (\frac{n}{Nt}) \cdot K, \phi = \phi_n)$, where ϕ_n accounts for the additional, known variance introduced when sampling J reads from the K cells for a lineage at size n using the transform (5.19).

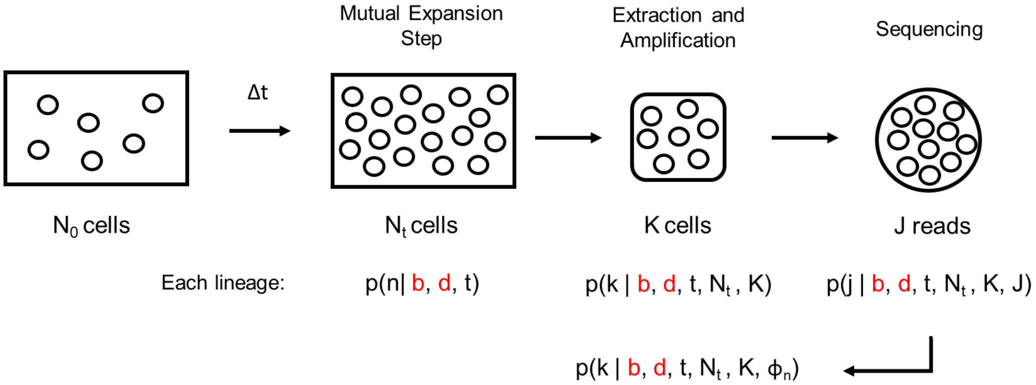


Figure 5.2: A schematic of the sampling steps and probability distributions used to infer the birth and death rates in the Bayesian model framework. Parameters in red (b and d) are unknown parameters that are fit by the model. The bottom arrow depicts the replacement of an explicit distribution for sampling J reads with an extra variance term in the $p(k)$ distribution.

Finally, as I observe the J sampled reads, it is necessary to transform them on to the scale K, where each transformed count, $k = j \cdot \frac{K}{J}$. I therefore have a likelihood function for the distribution of k normalised cell counts given the unknown parameters b and d. I subsequently fit this model within a Bayesian framework to simulated and experimental data.

5.6 Bayesian Model Development and Implementation

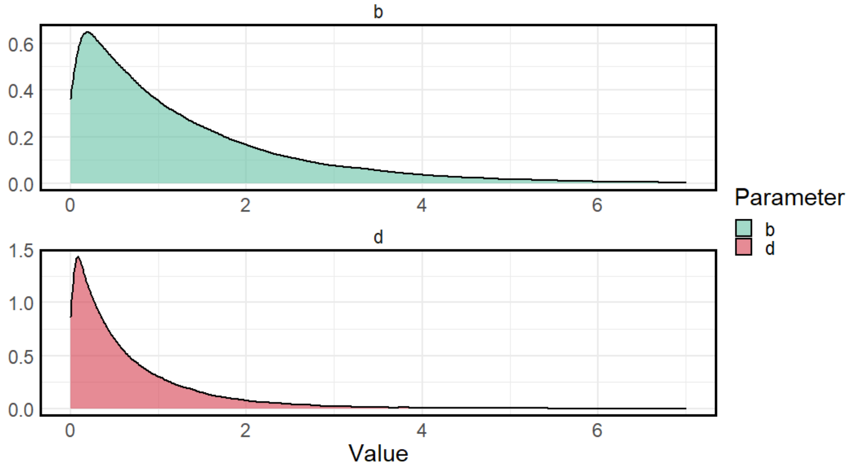
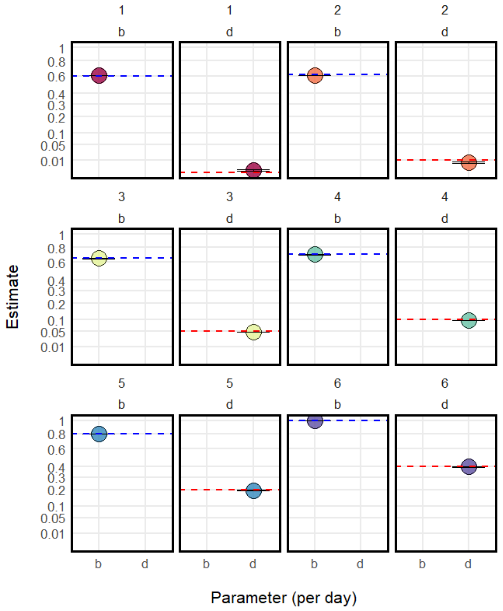


Figure 5.3: Prior distributions on the birth (b) and death (d) parameters (units = $days^{-1}$) used for the Bayesian inference on both the simulated and experimental data.

Employing the birth-death simulation model I developed in Results Chapter 1, I simulated several different lineage distributions using cell and read sampling numbers similar to those used in the *in vitro* experiment (see Materials and Methods for the specific sampling steps). For example, as depicted in Figure 5.2, the simulations assumed that there are 1x10^6 uniquely barcoded cells at $t = 0$. The cells are then grown for a given Δt, sampled K times without replacement (to simulate the sub-sampling and extraction step) and then sampled J times with replacement (to simulate the sampling of reads on the sequencer flow-cell). I then established a Bayesian model using the software Stan (in R - Rstan) and the likelihood function described above. Inference was run for 4000 iterations with 4 chains. The priors for the two parameters I aim to infer - the birth and death rates per day, b and d - are shown in Figure 5.3.

5.6.1 Simulated Data Inference

As shown by Figure 5.4, the model does well at recovering the true birth and death rates of the underlying cell population, despite the lineage distributions having been subject



Parameter	Posterior Mean	Posterior Std.Dev	R^	True Value	Simulation Number
b	0.6017	0.0003	1.001	0.60	1
d	0.0005	0.0003	1.002	0.00	1
b	0.6078	0.001	1.003	0.61	2
d	0.0066	0.0009	1.003	0.01	2
b	0.6483	0.001	1.001	0.65	3
d	0.0471	0.001	1.002	0.05	3
b	0.6986	0.0011	1.004	0.70	4
d	0.0975	0.001	1.004	0.10	4
b	0.7971	0.0012	1.004	0.80	5
d	0.196	0.0012	1.004	0.20	5
b	0.9977	0.0015	1.000	1.00	6
d	0.3968	0.0015	1.00000	0.40	6

Figure 5.4: Posterior estimates of the birth (b) and death (d) parameters (units = $days^{-1}$) vs the true values (blue and red dotted lines in plots). The LHS panels corresponds to posterior means with error bars +/- std.dev.

to subsequent sampling steps. At very low cell turnover values, the model struggles to dinstinguish very low levels of variance introduced via the birth-death process from those introduced due to the technical bottlenecks alone, and I therefore see slight over-estimates the death rate. For example, Simulation Number 1 in Figure 5.4 corresponds to a pure-birth process ($d = 0$). However, at rate ratios where $d/(b + d) \sim 0.04$, this slight over-estimate has largely disappeared. As I expect the true biological process in my cell-lines to deviate from a pure-birth process (even in standard culture conditions cell death is observed *in vitro*), even this slight bias is likely of little consequence in the true, sequenced data.

5.6.2 Experimental Data Inference

For each cell-line, I harvested 3x `POT` samples in the 'QR' experiment following the shared, mutual expansion stage (as illustrated in Figure 5.1). To infer the birth and death rates of each population, I simultaneously fit the Bayesian model outlined above to all 3x replicates simultaneously. Figure 5.5 shows the observed counts for one of the three replicates compared with the simulated distribution using 100 random draws from the posterior distribution. Figure 5.6 shows the full posterior distributions for all 4 chains of b and d per cell-line. The net growth rate $(b - d)$ of `QR_HCTbc` is slightly higher than `QR_SW6bc`, consistent with the longer passage times observed *in vitro*.

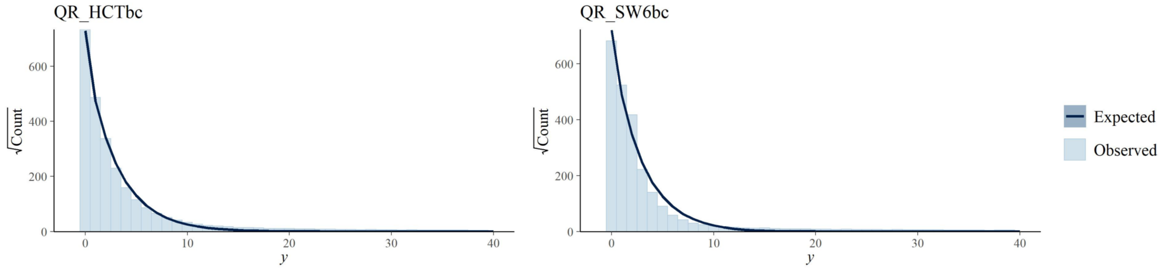
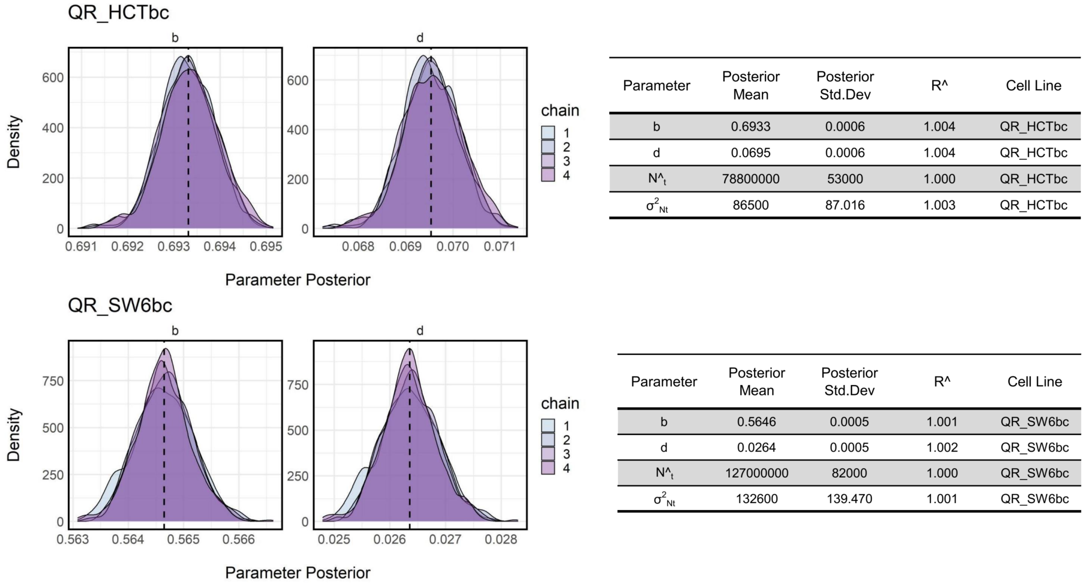


Figure 5.5: Posterior predictive distributions ('Expected') vs observed, normalised lineage counts for one of the three replicates used for model fitting for each barcoded cell line (`QR_HCTbc` and `QR_SW6bc`). Expected values were simulated by randomly sampling 100 draws from the posterior distributions of b and d. The y-axis depicts the square root of the given counts.

5.7 Discussion

Here I have developed and implemented a model within a Bayesian framework to infer the birth and death rates of my two cell-lines. Attempts to tease apart the birth and death rates separately have often required additional experiments, including live-imaging or determining the number of dead cells in a given time-window with FACS (Russo et al. 2021; Johnson et al. 2019). The variance within the observed lineage counts - made possible by the adoption of my lineage tracing technique - holds information on the cell turnover. As the total population size after a given period of time tells us about the net growth rate, I have designed a model that incorporates both sources of information to provide separate estimates of the birth and death rates.



Parameter	Posterior Mean	Posterior Std.Dev	R^	Cell Line
b	0.6933	0.0006	1.004	QR_HCTbc
d	0.0695	0.0006	1.004	QR_HCTbc
$N\hat{}_t$	78800000	53000	1.000	QR_HCTbc
σ^2_{Nt}	86500	87.016	1.003	QR_HCTbc

Parameter	Posterior Mean	Posterior Std.Dev	R^	Cell Line
b	0.5646	0.0005	1.001	QR_SW6bc
d	0.0264	0.0005	1.002	QR_SW6bc
$N\hat{}_t$	127000000	82000	1.000	QR_SW6bc
σ^2_{Nt}	132600	139.470	1.001	QR_SW6bc

Figure 5.6: Posterior distributions for the birth (b) and death (d) parameters for each barcoded cell-line (`QR_HCTbc` and `QR_SW6bc`). The posterior is plotted for each chain, and the mean value is denoted by the dotted line. $\hat{N_t}$ and $\sigma^2_{N_t}$ were also fit by the model and the mean and std.dev of their posterior distributions are also reported in the posterior summary tables (RHS).

Numerous studies have modelled the dynamics of resistance evolution *in vitro*. Due to the difficulty in teasing apart the two, birth and death rates have either been chosen based on previously published data (Acar, Nichol, Fernandez, et al. 2019) or have been simulated over a broad range of possible values (Bhang et al. 2015). By developing a model that explicitly mirrors features of my *in vitro* experiment - including a mutual expansion step and two technical sampling bottlenecks: extraction and sequencing - I recreate sampled lineage distributions where the true birth and death rates are known. In Results Chapter 1, I illustrated that the birth death probability distribution matches the lineage size distributions obtained from the stochastic birth-death model exactly. Here, I validate the model inference by accurately recovering the true birth and death rates from these simulated distributions.

One limitation is that the model assumes that all cells within the population have the same birth and death rates. Whilst cells instead undoubtedly have rates that follow some distribution, in the standard culture conditions we expect selection to be relatively weak; intra-population differences in proliferative rates should be small. Future work could extend the model to look at subsequent time-points (from the control replicates) and identify lineages that deviate strongly from the model's expectations. These deviations could then be used to infer the differences in relative fitness of the competing lineages, and therefore quantify the selection experienced by cells in the control conditions. Similar work in yeast has been employed to infer the distribution of fitness effects of newly arising mutations (Levy et al. 2015).

Chapter 6

Results Chapter 4 - Inferring the Evolutionary Dynamics of Drug Resistance during Long-Term Chemotherapy Experiments

6.1 Summary

In Results Chapter 1, I described and developed models that built expectations for lineage distributions under various evolutionary scenarios. I designed the models to capture important features of the *in vitro* experiment, including a shared mutual expansion step prior to splitting cells into respective treatment replicates, and subjecting half of the replicates - the drug-treatment arm - to periodic drug-treatment, where survival was contingent on a resistance phenotype. Here, I present the sequenced results from the analogous *in vitro* experiments, and use the evolutionary simulations to interpret the data. For a summary of the experimental set-up and which sample codes correspond to which replicate types, please refer to the Materials and Methods chapter. The Experimental codes PR and QR) also correspond to those also outlined in the Materials and Methods chapter. Briefly, the main distinguishing features of each sequencing experiment are as follows:

- Pulse Run 1 (PR) - Both colorectal cancer cell-lines (*HCT116* and *SW620*) were infected with the lineage tracing barcodes and grown in the drug-treatment long-term evolution experiment design (outlined in the Materials and Methods chapter) for a total of two passages - this included four control arms (subject to pulse vehicle control (DMSO) treatment) and four drug-treatment replicates (subject to pulse chemotherapy (5-Fu) treatment at IC_{50} values).
- Pulse Run 2 (QR) - The drug-treatment long-term evolution experiment was repeated in both cell-lines (*HCT116* and *SW620*) to ensure the technical application of the barcode library and the evolutionary dynamics observed were repeatable. All experimental parameters were the same as Pulse Run 1 (PR), however the experiment now lasted for a total of five passages.

The results sections covering both 'PR and QR' are prefaced with a section on solving technical difficulties encountered when sequencing the barcode amplicons on NGS platforms, before a detailed description of results pertinent to the evolutionary dynamics experienced by each cell-line during drug-treatment. I show how differences in the expected distributions of lineages within and between replicates for different modes of resistance evolution - captured by different parameters in the models of Results Chapter 1 - can be used to identify likely modes of evolution operating in the *in vitro* data.

6.2 Identifying and Rectifying Technical Artefacts in Sequenced Lineage Distributions

In each of the Pulse Run Experiments, when comparing the distribution of barcode lineages between samples, an unexpected relationship was observed. In both sequencing outputs, barcode lineages that rose to a high frequency in the drug-treatment replicates were also found in putatively unrelated replicates - namely, control replicates in the opposing cell-line - at a frequency that was strictly proportional to their frequency in the drug-treatment replicate. This pattern had no feasible biological explanation, and the strict concordance of the relationship hinted at a technical artefact that was occurring following the cell barcode DNA extraction and amplification, and was hence avoiding the noise introduced during these technical preparation steps. Additionally, in the PR experiment, this effect was most noticeable when comparing samples that shared either a forward or reverse multiplexing index, further supporting a technical effect that was occurring during sequencing.

6.2.1 PR Experiment

The clearest pattern that supported a technical artefact in the PR sequencing experiment was the lineage distribution relationship between two biologically unrelated samples. For example, as barcoding the cells was repeated independently for each cell line, we would not expect any correlation between successful barcodes in a passage 2 control sample in *SW620* (`PR_SW6bc_COi_P2`) and a drug-treatment passage 2 sample in *HCT116* (`PR_HCTbc_DTi_P2`). Yet certain comparisons of this nature yielded strong correlations. Confirming the *technical* nature of these inter-sample correlations, comparisons that didn't share a forward or reverse index did not yield such a relationship (indexes are unique nucleotide sequences incorporated into a sample's amplified barcode sequences to multiplex on a sequencing flow-cell).

The most parsimonious explanation for these patterns is a sequencing phenomenon known as 'index-hopping' (Costello et al. 2017). This occurs when sequence-able molecules incorporate a different forward or reverse index than the one they were assigned. As samples were distinguished via different combinations of forward and reverse indexes, this led to reads being wrongly assigned to a different sample if they shared the remaining forward or reverse index. Of note, these patterns are particularly conspicuous in my data due to certain drug-treatment samples having a few lineages that dominate the sample. These samples contain many millions of identical reads which, due to the index-hopping occurring proportional to any given barcode's read count, left an observable signature when 'hopping' into other samples. Fortunately, the 'hopping' process appeared to be repeatable when comparing biologically unrelated samples that either shared a forward or reverse index, and the conspicuous patterns in comparisons that contain extremely high frequency lineage counts provideded a means to estimate the rate at which index-hopping occurs. Therefore, to derive estimates of this rate , I made use of the following relationship:

$$f_j = n_j + p(f_k) \tag{6.1}$$

where sample j is a control-treatment replicate that shares either a forward or reverse index with sample k, a biologically *unrelated* drug-treatment replicate. f_j is the observed frequency of a barcode in sample j, p is the index-hopping probability, n_j is the 'true' frequency of a barcode in sample j prior to index-hopping, and f_k is the observed frequency in the drug-treatment replicate. To infer the hopping rate, I restricted myself to only using barcodes with an extremely high count in the drug-treatment replicates (f_k), as it is when comparing these lineages that the hopping signature was most noticeable.

To estimate the index-hopping rate from these specific comparisons, I made some assumptions. The first is that, because $n_j << n_k$, and because the comparison is between two biologically *unrelated* samples, I can derive an estimate by assuming $n_j \approx 0$. The second is that, because $f_k >> 0$, I can assume that $f_k \approx n_k$. That is, I can assume that the index-hopping had a negligible impact on the observed count in the drug-treatment sample barcodes that were found at an extremely high relative frequency (often $> 10^{-2}$).

After inspecting some different estimates of p for different index-combinations, it appeared that index-hopping occurred at slightly different rates depending on whether samples shared a forward or reverse index. Therefore, estimates of p were estimated separately as p_{fwd} and p_{rev} for forward and reverse indexes, respectively. By rearranging (6.1), I derived estimates of p_{fwd} and p_{rev} for various sample comparisons. I then 'un-hopped' these samples (described below), and compared the distributions of the new values of n_j. If these were too positive, the hopping signature remained, whilst too negative was interpreted as an over-estimate of the hopping rate.

Now that I had derived estiamtes of p_{fwd} and p_{rev} using the biologically unrelated sample comparisons, I could use these rules to 'undo' the hopping signature in biologically related samples. I assumed that the probability any given barcode molecule could hop into another sample can be represented by a probability matrix, M, where M_{jk} corresponds to the probability of any given barcode molecule hopping from sample j into sample k. If $j = k$, the probability equals the complement of the probability of hopping into all other samples one index away,

$$M_{jk} = 1 - (\sum_{k_{fwd}=1}^{K_{fwd}} p_{rev} + \sum_{k_{rev}=1}^{K_{rev}} p_{fwd}) \tag{6.2}$$

where K_{fwd} is the number of samples that are one forward index away, etc.

Therefore, finally, the relationship between observed counts and the putative 'true' counts prior to index-hopping in the i^{th} sample can be described by the two vectors f_i and n_i and the probability matrix M,

$$f_i = n_i * M \tag{6.3}$$

and I can rearrange to give the counts of each barcode, i, prior to hopping,

$$n_i = f_i * M^{-1} \tag{6.4}$$

To 'un-do' the hopping for all samples simultaneously, I repeated this process for each barcode and then used the corrected counts for downstream barcode clustering and statistical analysis.

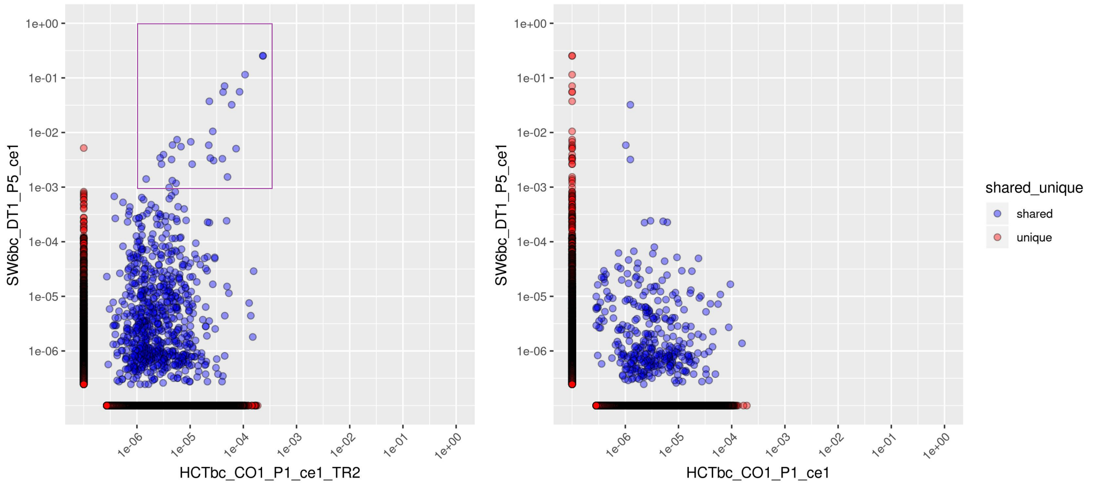


Figure 6.1: A comparison of two biologically unrelated samples: `HCTbc_CO1_P1` and `SW6bc_DT1_P5`. The plot on the LHS compares samples that were samples on the same flow-cell (`HCTbc_CO1_P1` is a technical replicate), whereas the plot on the RHS compares the two samples that were sequenced on *different* flow cells. Points are highlighted by whether they were shared between or unique to each sample in the respective comparisons.

6.2.2 QR Experiment

As a remedy to the index-hopping problems encountered in PR, I employed unique-dual indexes when sequencing samples in QR. These avoid the issue by assigning each sample a unique pair of both forward and reverse indexes. If a sample now wrongly incorporates a different forward or reverse index, it is now no longer correctly de-multiplexed and excluded from downstream analysis. Despite this technical modification, there were still slight signatures of a similar phenomenon occurring when comparing biologically unrelated samples, analogous to the comparisons outlined above. In fact, it was clear that this process now occurred on the sequencing flow-cell due to the sample layouts of my technical replicates: when one sample had a very high frequency of a few barcodes (drug-treatment replicates), biologically unrelated samples compared *within* a flow-cell exhibited the 'hopping' signature (purple highlighted box in Figure 6.1), whereas the same two samples that were sequenced on *different* flow-cells no longer displayed the relationship. Whilst the issue appeared much less severe than in the 'PR' experiment, the lack of any obvious relationship between index identity and hopping probability meant the solution had to be more general. The remedy involved making a list of 'trouble barcodes' that were found in any sample at a frequency $> 10^{-2}$ - the problem was dominated by barcodes found at these high frequencies. If any of these 'trouble barcodes' were found in another sample at a frequency $< 10^{-4}$, they were filtered out prior to downstream analysis. If the barcodes were found in another sample at a frequency $> 10^{-4}$, it was assumed that this observation was much more likely to be a real, shared observation. Whilst this process likely discards some true, low frequency barcode observations that are shared between samples, these low-frequency lineages will contribute little to the population's response to therapy. Furthermore, my adoption of diversity indices which can leverage high frequency barcodes diminishes the contribution of any residual technical mistakes missed by this filtering step.

6.3 The Dynamics of Drug Resistance Evolution

Following these filtering steps that rectify the technical artefacts that might influence the statistical behaviour of the barcode distributions, I now investigate the characteristics of the lineage distributions when comparing different replicates, treatments and experiments. I can now also make direct comparisons between my observed, experimental data with the theoretical assumptions developed in Results Chapter 1 to infer which evolutionary scenarios are driving the *in vitro* response to chemotherapy treatment in my two colorectal cell lines.

6.3.1 Cumulative Lineage Distributions Reveal Sample-Specific Bottlenecks

To compare the within-sample differences in lineage success, we can plot the cumulative frequency distributions: how many barcodes do we need to observe (x-axis) to have captured some cumulative proportion of the entire sample (y-axis). The shape of each cumulative distribution captures the inequality in barcode success within a sample, and the change between one time-point and the next is a product of the 'bottleneck' a sample has experienced. As control and drug-treatment replicates experience identical technical bottlenecks, the differences observed above and beyond those seen in the control replicates are the product of a cell population's response to treatment.

PR

In the PR experiment, the shapes of the cumulative distributions of both control treatments are similar: most lineages are found at very low frequencies, and the number of unique barcode lineages one needs to observe that make up 50% of each sample are in the order of 10^5 (Figure 6.2). In relation to the drug-treatment distributions, the changes between time-point one and two are minor. These observations are consistent with the two cell-lines experiencing very similar dynamics in the control-treatment replicates. The minor loss of lineages between time-points is a pattern expected from a small bottleneck - as the technical bottlenecks are the same for all replicates, these relative differences compared to the drug-treatment replicates are the result of small differences in competing lineages proliferative potential, or 'weak selection'.

Differences between each cell-line's drug-treatment cumulative distributions immediately point to differences in each population's response to drug-treatment (Figure 6.2). Namely, the `HCTbc_DT` replicates lose lineages much more rapidly; even by the first time-point, Passage 1, the top 50% of each sample is made up of lineages in the order of 10^2. By Passage 2, the variance between replicates has increased dramatically. for example, in `HCTbc_DT1`, one lineage now comprises more then 50% of the entire sample, whereas in `HCTbc_DT3`, there are still approximately 20 lineages making up this same cumulative fraction. In the `SW6bc_DT` replicates the bottleneck appears far less stringent. Thousands of barcode lineages make up the top 50% of both time-points. The repeatability of the dynamics also appear more consistent: the *between* replicate comparisons of `SW6bc_DT` cumulative distributions are almost identical (Figure 6.2).

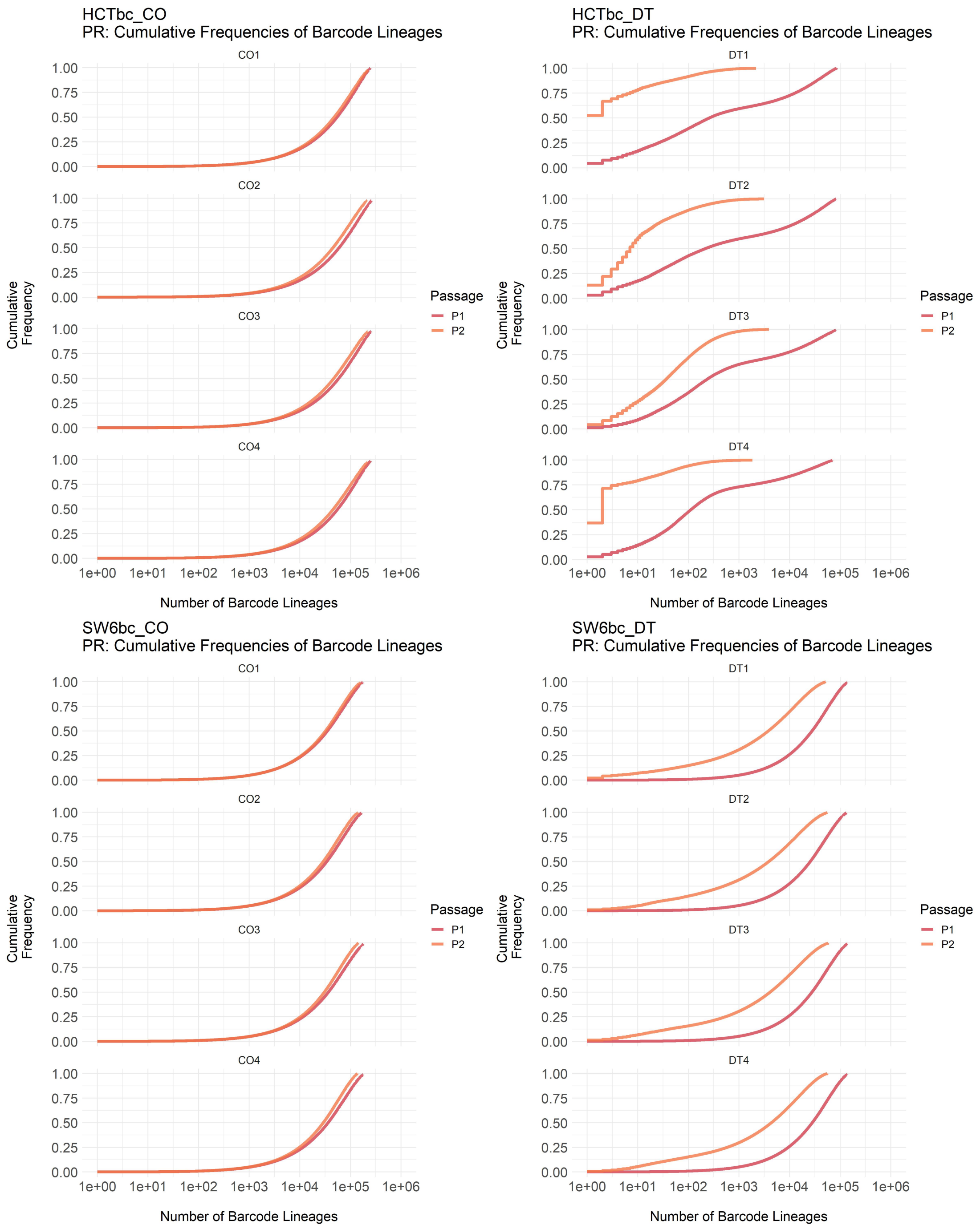


Figure 6.2: The cumulative frequency of sequenced barcode lineages as a function of the total number of unique barcode lineages in each replicate. Distributions shown for the 'PR' experiment. Panels correspond to different cell-lines - `HCTbc` and `SW6bc`, top and bottom row, respectively - and different treatment-types - control (`CO`) and drug-treatment (`DT`), left and right columns, respectively. Colours within each panel correspond to Passage number.

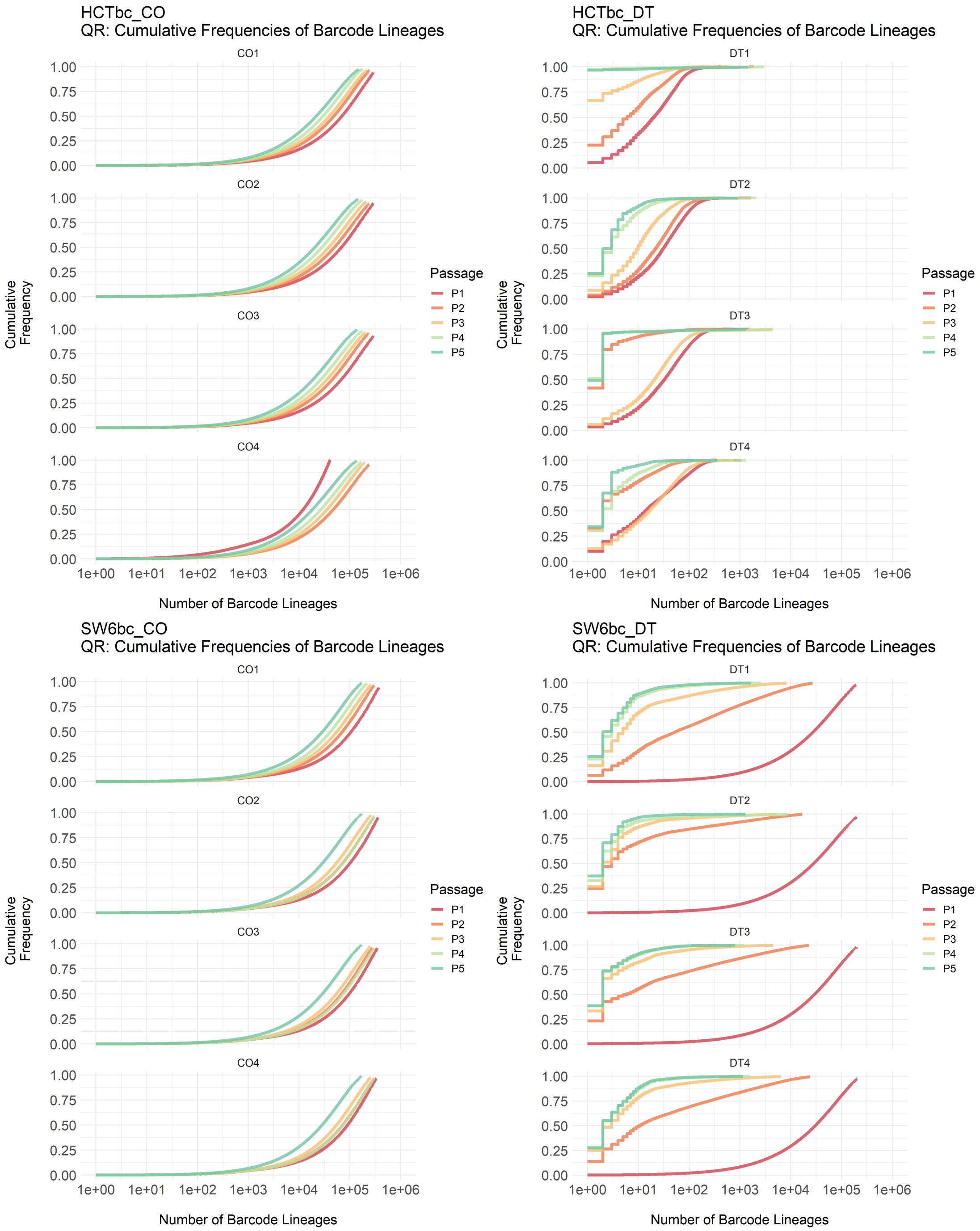


Figure 6.3: The cumulative frequency of sequenced barcode lineages as a function of the total number of unique barcode lineages in each replicate. Distributions shown for the 'QR' experiment. Panels correspond to different cell-lines - `HCTbc` and `SW6bc`, top and bottom row, respectively - and different treatment-types - control (`CO`) and drug-treatment (`DT`), left and right columns, respectively. Colours within each panel correspond to Passage number.

QR

Following the preliminary 'PR' drug-treatment experiment, the 'QR' experiment allowed me to assess the repeatability of the dynamics, and observe how they differ for an extended number of time-points, relative to 'PR'. The 'QR' control treatment cumulative frequency distributions are in agreement with the PR results: the loss in diversity between time-points is marginal, relative to the drug-treatment replicates (Figure 6.3, LHS plots). Most barcodes are found at a very low frequency, indicative of weak selection in the control environment: no lineages come to dominate the flasks within the evolutionary time observed in the experiment (n.b. the `QR_HCTbc_CO4_P1` sample that had a low read count, and hence shows an anomalous cumulative frequency shape).

The drug-treatment cumulative distributions are also broadly in agreement: relative to the control flasks, both cell-lines experience a high loss in diversity. Again, in keeping with the 'PR' experiment, in the first few passages the selective bottleneck appears far more stringent in `HCTbc` than in `SW6bc` (Figure 6.3, RHS plots). Also in agreement with 'PR', whereas the between-replicate differences in `SW6bc` appear strikingly repeatable, the `HCTbc` dynamics appear far more stochastic: there are clear between-replicate and time-point differences.

In `HCTbc_DT3` and `DT4`, more lineages are found at a higher frequency in Passage 2 than in Passage 3. Whilst a raise in diversity may seem to go against the expectations when selection is strong, these patterns are indicative of clonal interference, where a lack of recombination means beneficial mutations that arise on different genetic backgrounds are destined to compete. These patterns are only observed in two of the four `HCTbc` replicates, suggesting that the accrual of phenotypic changes that confer a growth advantage in the treated environment are highly stochastic. Nonetheless, this temporary gain in diversity has been lost by Passage 4, where it appears the resolution to observe ongoing dynamics has been lost: Passage 5 is almost indistinguishable in all 4 replicates in both cell-lines. Once a population has become dominated by one or a few lineages, the monotonic nature of experimental barcode loss means any ongoing dynamics occurring within a barcode lineage are no longer visible.

Overall, the bottleneck lineages experience appears less stringent in the 'PR' experiment than in ‘QR’: by Passage 2, both cell-lines drug-treatment replicates have lost a higher proportion of lineages in ‘QR’ than in ‘PR’. This pattern is shared between cell-lines despite their markedly different lineage distributions, suggesting that this disparity is a product of inter-experimental differences, opposed to intrinsic biological differences.

6.3.2 Within- and Between-Replicate Diversity Differences

In Results Chapter 1, I chose Hill Diversity indices - within-replicate (${}^{q}D$) and between-replicate differences (${}^{q}D(\beta)$) - that can reduce numerous replicate lineage distributions onto an informative statistic space. By focussing on the ${}^{q}D$ summary statistics of order $q = 2$, I leverage the high-frequency barcodes when calculating within- and between replicate diversity. This helps account for technical noise in sequencing barcodes which can inflate low-frequency counts, and preferentially focuses the statistics on lineages that have grown to dominate their respective replicate population. Reducing lineage distributions onto this statistical space allows for comparisons of within and between replicate differences simultaneously.

The within replicate lineage diversity (${}^{q}D$) decreases very slightly between time-points in the control replicates (moving down rows in the `CO` column, Figure 6.4). As highlighted in the cumulative frequency distributions, and supporting the notion that they provide a valuable ‘baseline’ with which to compare the drug-treatment samples, both cell-lines control replicates behave similarly: this is broadly true for the between-replicate differences also, shown by similar ${}^{q}D(\beta)$ values.

The striking difference between the two-cell lines’ responses to treatment are the extremely high dissimilarity between drug-treatment replicates in `HCTbc` (high values of ${}^{q}D(\beta)$), and the relative high similarity between drug-treatment replicates in `SW6bc` (low values of ${}^{q}D(\beta)$) (Figure 6.4). These differences point to either the same barcodes coming to dominate the replicate flasks, as in `SW6bc`, or different barcodes, as in `HCTbc`. In Passages 2 and 3 in the QR experiment, the variance in the `HCTbc` within-replicate diversity (the x-axis in Figure 6.4) suggests the route to success is stochastic. Yet despite these inter-time-point differences, dynamics appear highly repeatable between experimental runs (‘PR’ vs ‘QR’ panels). Following Passage 3, this variance in ${}^{q}D$ col-

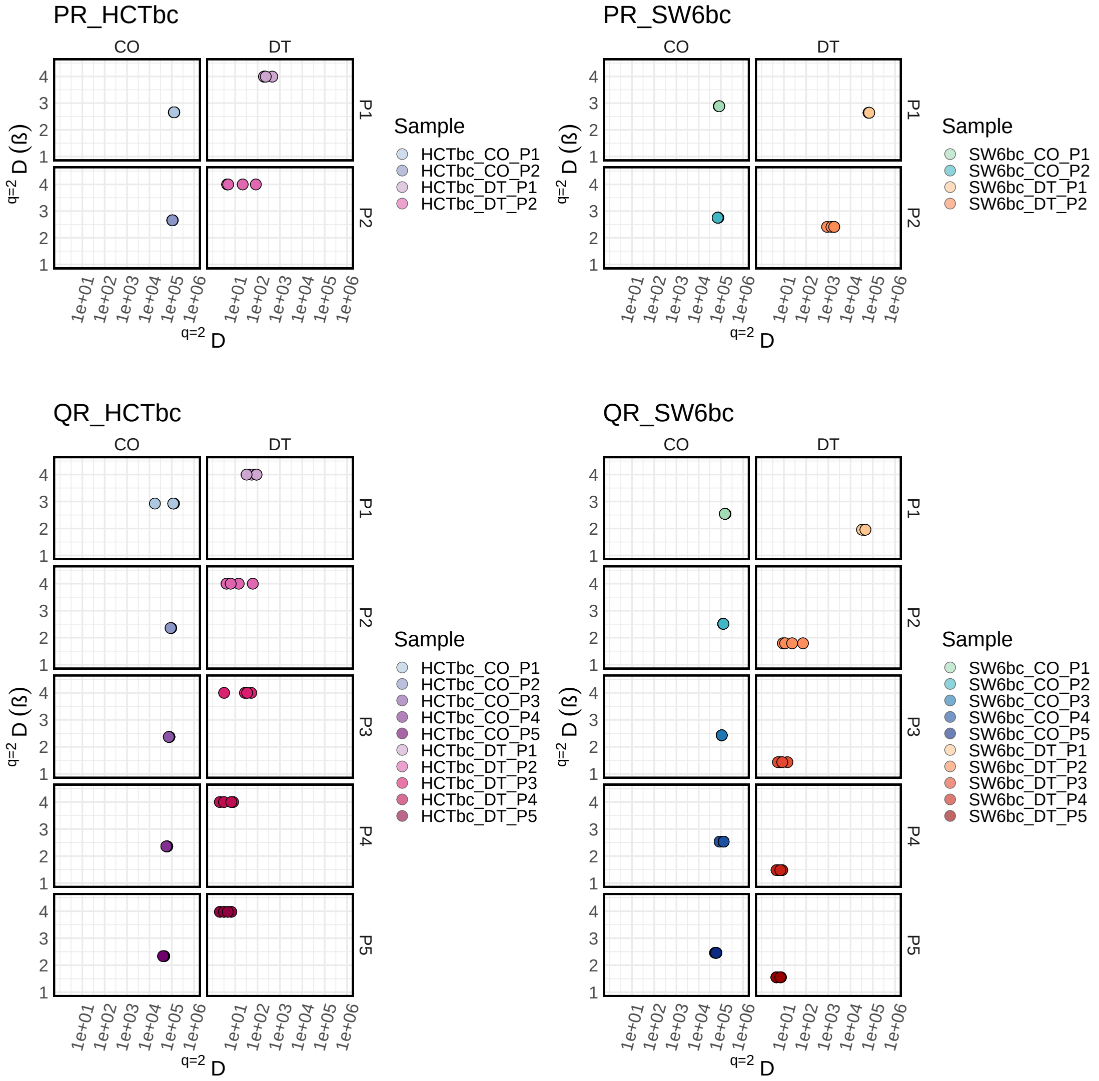


Figure 6.4: ${}^qD$ (within-replicate diversity) vs ${}^qD(\beta)$ (between-replicate differences in diversity) of order $q = 2$ for the 'PR' (top row) and 'QR' (bottom row) drug-treatment experiments, for barcoded coloretcal cancer cell-lines `HCTbc` (LHS) and `SW6bc` (RHS). Each point corresponds to a distinct replicate flask, and each separate panel and colour corresponds to treatment type - control (`CO`) and drug-treatment (`DT`) - and Passage (`P#`): moving from top to bottom corresponds to progressive time-points.

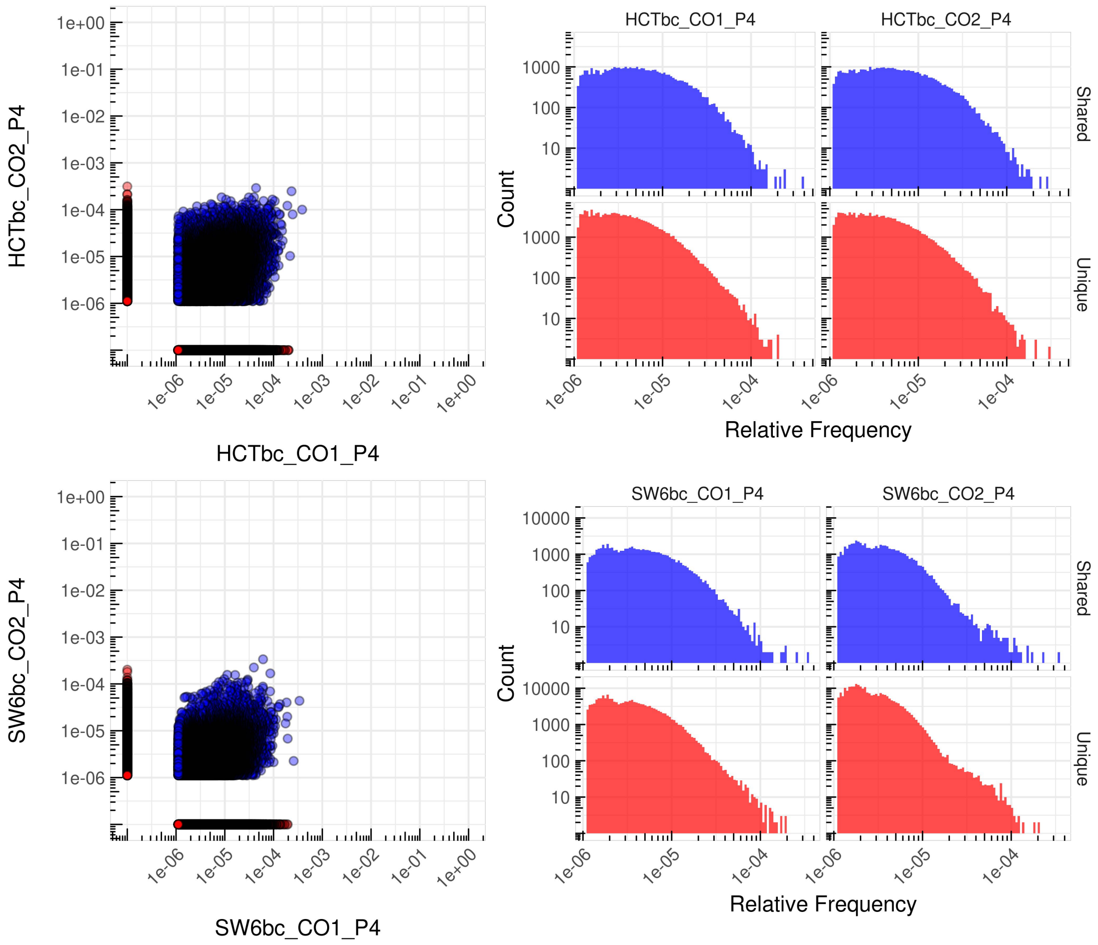


Figure 6.5: 'QR' experiment: Pairwise comparisons of sequenced barcode lineages between two control passage 4 replicates (`CON_P4`) from the *HCT116* and *SW620* cell-lines (top and bottom row, respectively). The shared lineages relative frequencies are shown in blue, whilst those unique to each replicate are shown in red. The histograms show the distribution of shared and unique lineages sorted by each sample.

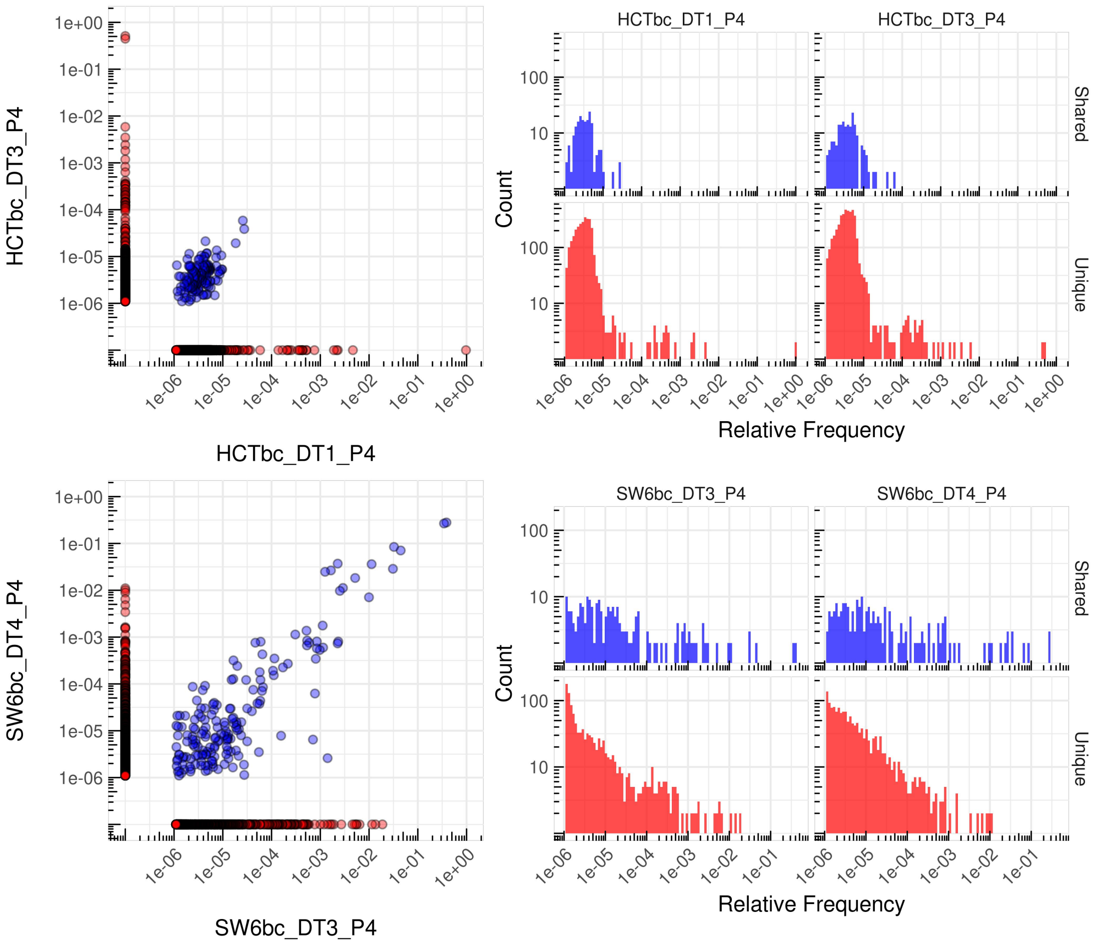


Figure 6.6: 'QR' experiment: Pairwise comparisons of sequenced barcode lineages between two drug-treatment passage 4 replicates (`DTN_P4`) from the *HCT116* and *SW620* cell-lines (top and bottom row, respectively). The shared lineages relative frequencies are shown in blue, whilst those unique to each replicate are shown in red. The histograms show the distribution of shared and unique lineages sorted by each sample.

lapses in later passages as the few successful lineages in each replicate come to dominate the flask.

Relative to HCTbc, the two experiments' SW6bc drug-treatment samples responses are in broad agreement: in each run, SW6bc replicates become more similar as the time-points progress (${}^{q}D(\beta)$ decreases), and a greater number of lineages are retained in the early passages. However, as observed in the cumulative frequency distributions, the bottleneck appears less severe in 'PR': more lineages survive treatment to drug-treatment Passage 2, and the selection for the *same* barcodes between replicates also appears less severe in 'PR'.

For a high-resolution comparison of how lineages differ between replicates, we can plot the pairwise differences between two given samples. Figure 6.5 shows such a comparison for two Passage 4 control replicates from the 'QR' experiment for each cell-line. Even by the 4th time-point, no lineages have risen above $10^{-}3$ in relative frequency, and there are many shared barcodes (blue histograms in Figure 6.5) throughout the lineage distribution. Figure 6.6 instead shows the analogous comparison for two drug-treatment replicates for each cell-line. Unlike comparisons between individual control replicates (Figure 6.5), where the frequencies and proportions of each lineages are broadly the same, drug-treatment differences between cell-lines is stark: there is a strong correlation in the highly successful lineages in the SW6bc_DT replicates, where the barcodes that come to dominate each flask are shared amongst replicates (blue points in Figure 6.6 - *n.b.* the log-scale for the relative frequencies). In HCTbc_DT, however, even fewer lineages dominate any individual replicate flask, and those that do are unique to each replicate (red points in Figure 6.6). The two summary statistics I employ can be thought of as reducing these sources of information into two axes: the ${}^{q}D$ axis (within-replicate diversity) captures the degree to which any individual replicate is dominated by a few, successful lineages, whilst the ${}^{q}D(\beta)$ axis (between-replicate difference in diversity) captures the 'sharedness' of these successful lineages between all pairwise comparisons, simultaneously.

Parameter	Value	Cell Line
b (days^{-1})	0.693	HCTbc
d (days^{-1})	0.070	HCTbc
N_0	1 x 10^6	HCTbc
N_{max}	30 x 10^6	HCTbc
t_{exp}	7.0	HCTbc
l	4	HCTbc
Δt_{DT} (days)	4.0	HCTbc
nsim	10	HCTbc
ρ, μ, σ, δ	(variable)	HCTbc

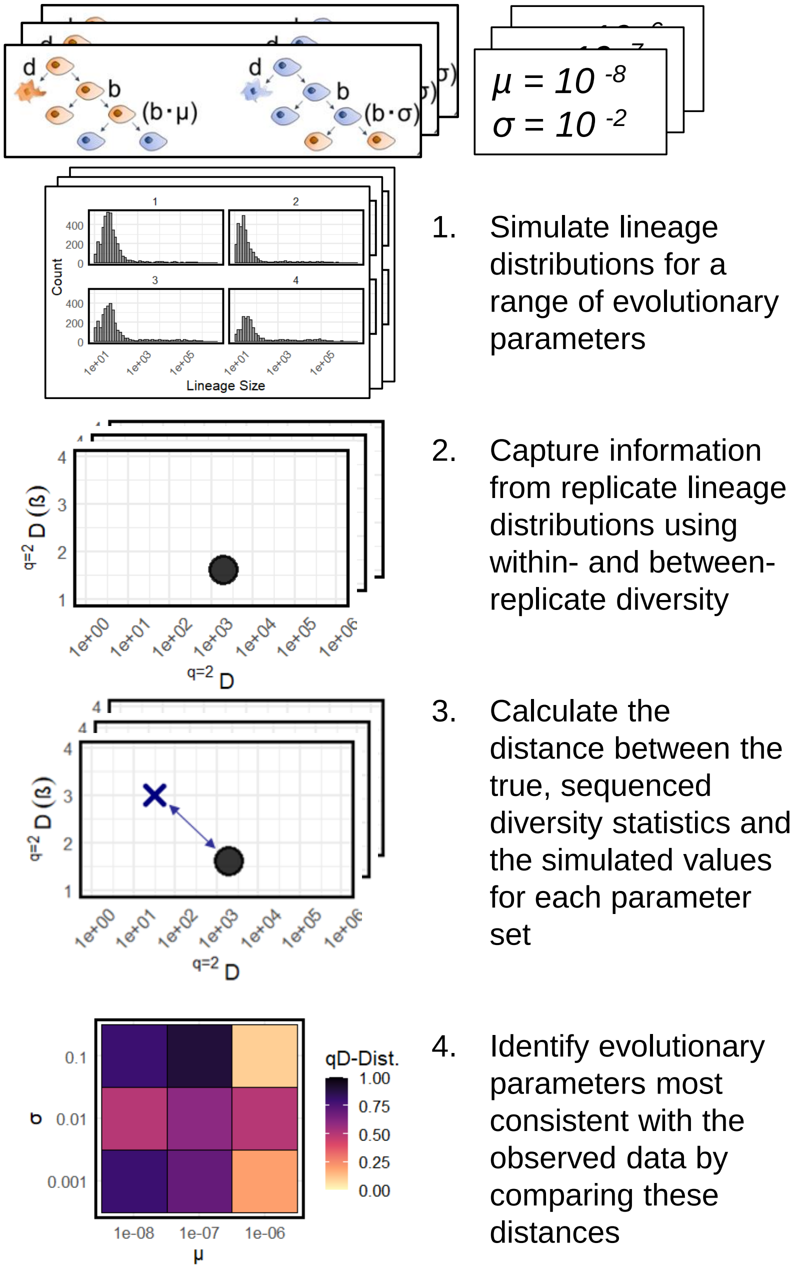
Parameter	Value	Cell Line
b (days^{-1})	0.565	SW6bc
d (days^{-1})	0.026	SW6bc
N_0	1 x 10^6	SW6bc
N_{max}	40 x 10^6	SW6bc
t_{exp}	9.0	SW6bc
l	4	SW6bc
Δt_{DT} (days)	4.0	SW6bc
nsim	10	SW6bc
ρ, μ, σ, δ	(variable)	SW6bc

Table 6.1: Parameters used for each cell-line specific set of simulations (parameters correspond to those outlined in Results Chapter 1, and birth and death rates are those inferred using the Bayesian model in Results Chapter 3).

6.3.3 Inferring Evolutionary Scenarios with Simulated Lineage Distributions

Whilst comparisons between cell-lines' and time-points' lineage distributions can provide some qualitative evidence as to the population-level dynamics of resistance evolution, it can be hard to distinguish differences due to stochastic experimental sampling from true, biological differences. As such, I now compare the summary statistics of the sequenced lineages with the simulated values under various evolutionary scenarios. Figure 6.7 shows the workflow for identifying the most likely combination of evolutionary parameters in the sequenced data. Briefly, the simulated lineage distributions (4 per treatment, as in the *in vitro* experiment) are condensed into within- and between-replicate lineage diversities (${}^qD$ and ${}^qD(\beta)$, respectively). This process is repeated for a range of parameter values. The euclidean distance between the simulated and sequenced points in diversity statistic space is calculated for each parameter set, and the distances compared to identify those most consistent with the cell-line's response (heatmap in figure 6.7).

To illustrate the distances in summary statistics between the simulated lineage distributions and sequenced values, I have plotted them simultaneously for drug-treatment Passage 2 for a range of evolutionary scenarios, excluding those indistinguishable from



1. Simulate lineage distributions for a range of evolutionary parameters

2. Capture information from replicate lineage distributions using within- and between-replicate diversity

3. Calculate the distance between the true, sequenced diversity statistics and the simulated values for each parameter set

4. Identify evolutionary parameters most consistent with the observed data by comparing these distances

Figure 6.7: A schematic illustrating how evolutionary scenarios are inferred with simulated lineage distributions. A small set of hypothetical parameters are chosen (μ and σ) for illustrative purposes.

the control replicates, and sorted the panels so rows correspond to equilibrium frequencies of resistance (Figures 6.8-6.11 and 6.13-6.16 - filled coloured points are simulated values, crosses are sequenced values). However, to leverage the multiple Passages *in vitro* and *in silico*, the similarity in summary statistics over all time-points can be combined by taking the average euclidean distance between each simulated and sequenced replicate for a given set of parameter values. The values of ${}^{q}D$ and ${}^{q}D(\beta)$ are normalised so that each fall between 0 and 1. The distance in the ${}^{q}D$ axis is taken in log-transformed space, as is shown plotted in Figures 6.8-6.11 and 6.13-6.16, and then averaged across all time-points. The final distances are shown as the heatmap values qD-$Dist.$ in Figures 6.12 and 6.17.

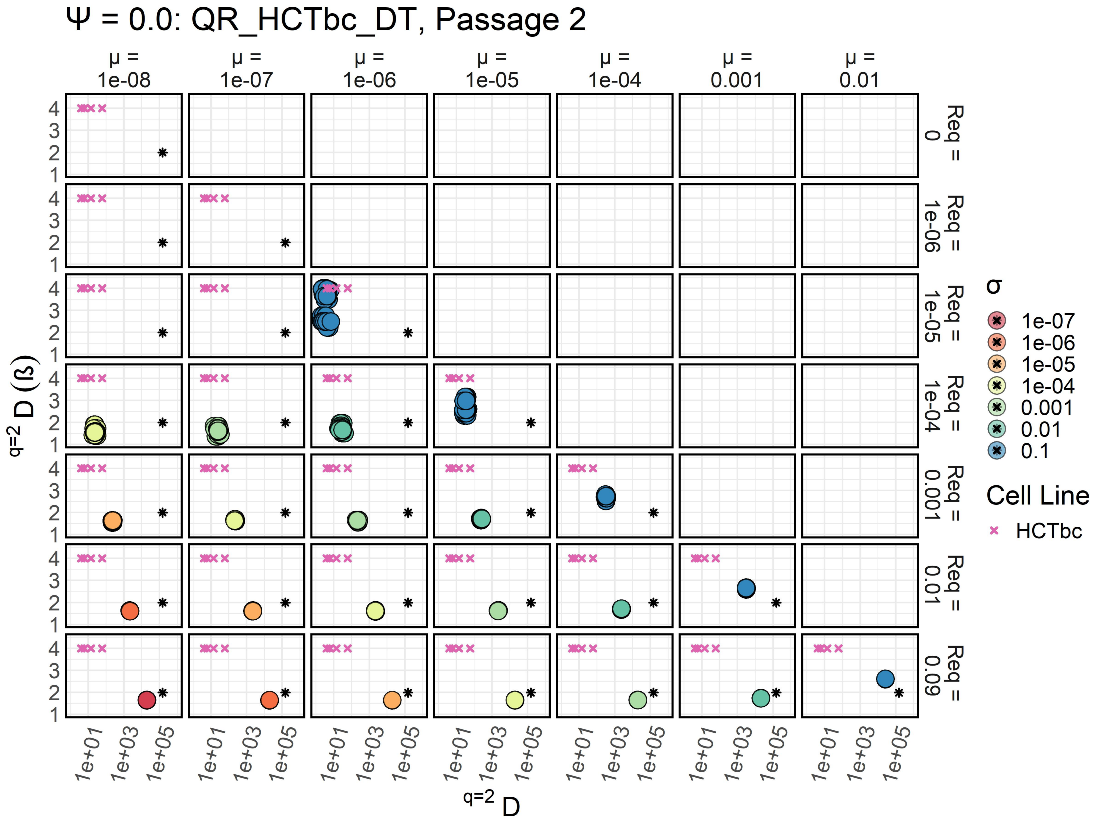


Figure 6.8: ${}^{q}D$ (within-replicate diversity) vs ${}^{q}D(\beta)$ (between-replicate diversity dissimilarity) of order $q = 2$ for the combined HCTbc simulation's Passage 2 drug-treatment (`DT_P2`) replicates. Any given panel corresponds to a simulation set that was run using the combination of parameters that control the resistant phenotype's evolution: columns correspond to values of μ, whilst rows correspond to values of σ that lead to the same equilibrium frequencies of resistance (R_{eq}). The black points in the bottom RHS of each individual panel corresponds to the control treatments' mean value. The coloured crosses correspond to the equivalent statistic values in the 'QR' HCTbc sequenced drug-treatment (`HCT_DT_P2`) replicates. The drug-treatment was modelled as deterministic ($\psi = 0.0$).

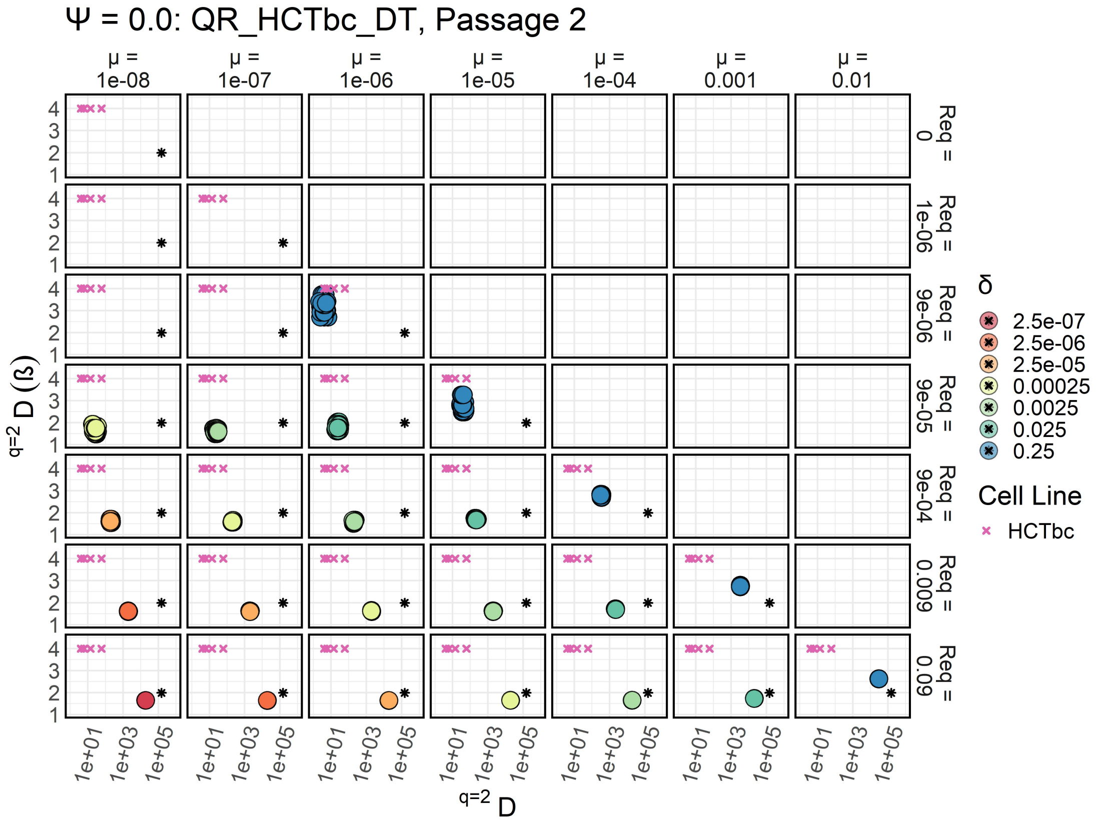


Figure 6.9: ${}^{q}D$ (within-replicate diversity) vs ${}^{q}D(\beta)$ (between-replicate diversity dissimilarity) of order $q = 2$ for the combined HCTbc simulation's Passage 2 drug-treatment (DT_P2) replicates. Any given panel corresponds to a simulation set that was run using the combination of parameters that control the resistant phenotype's evolution: columns correspond to values of μ, whilst rows correspond to values of δ that lead to the same equilibrium frequencies of resistance (R_{eq}). The black points in the bottom RHS of each individual panel corresponds to the control treatments' mean value. The coloured crosses correspond to the equivalent statistic values in the 'QR' HCTbc sequenced drug-treatment (HCT_DT_P2) replicates. The drug-treatment was modelled as deterministic ($\psi = 0.0$).

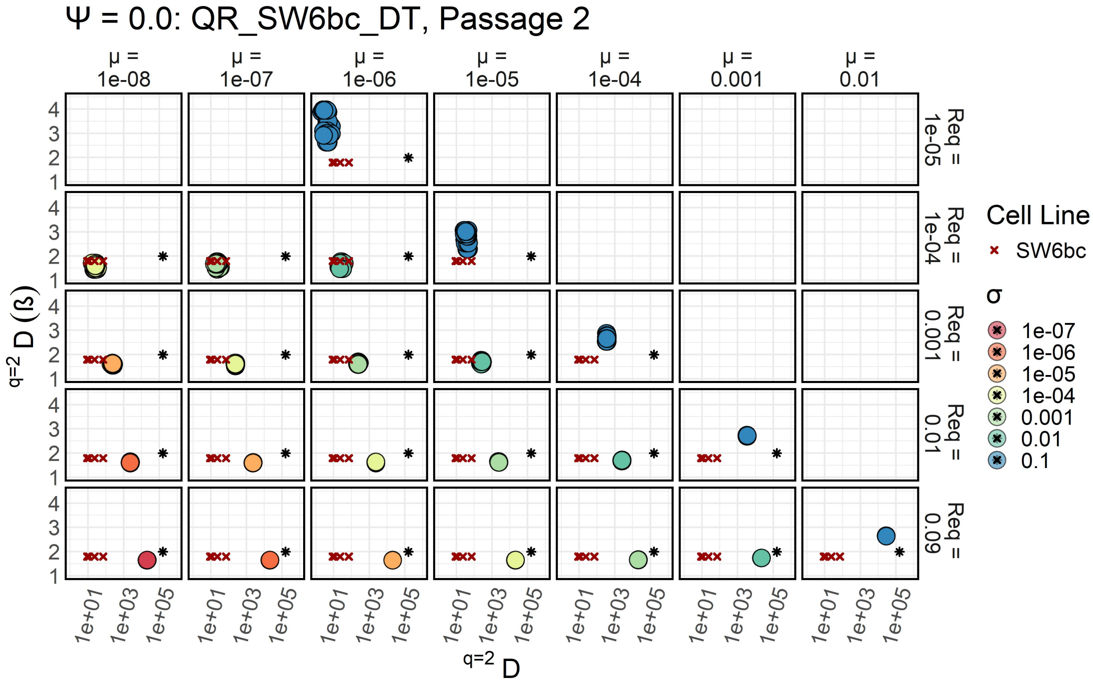


Figure 6.10: ${}^{q}D$ (within-replicate diversity) vs ${}^{q}D(\beta)$ (between-replicate diversity dissimilarity) of order $q = 2$ for the combined SW6bc simulation's Passage 2 drug-treatment (DT_P2) replicates. Any given panel corresponds to a simulation set that was run using the combination of parameters that control the resistant phenotype's evolution: columns correspond to values of μ, whilst rows correspond to values of σ that lead to the same equilibrium frequencies of resistance (R_{eq}). The black points in the bottom RHS of each individual panel corresponds to the control treatments' mean value. The coloured crosses correspond to the equivalent statistic values in the 'QR' SW6bc sequenced drug-treatment (SW6_DT_P2) replicates. The drug-treatment was modelled as deterministic ($\psi = 0.0$).

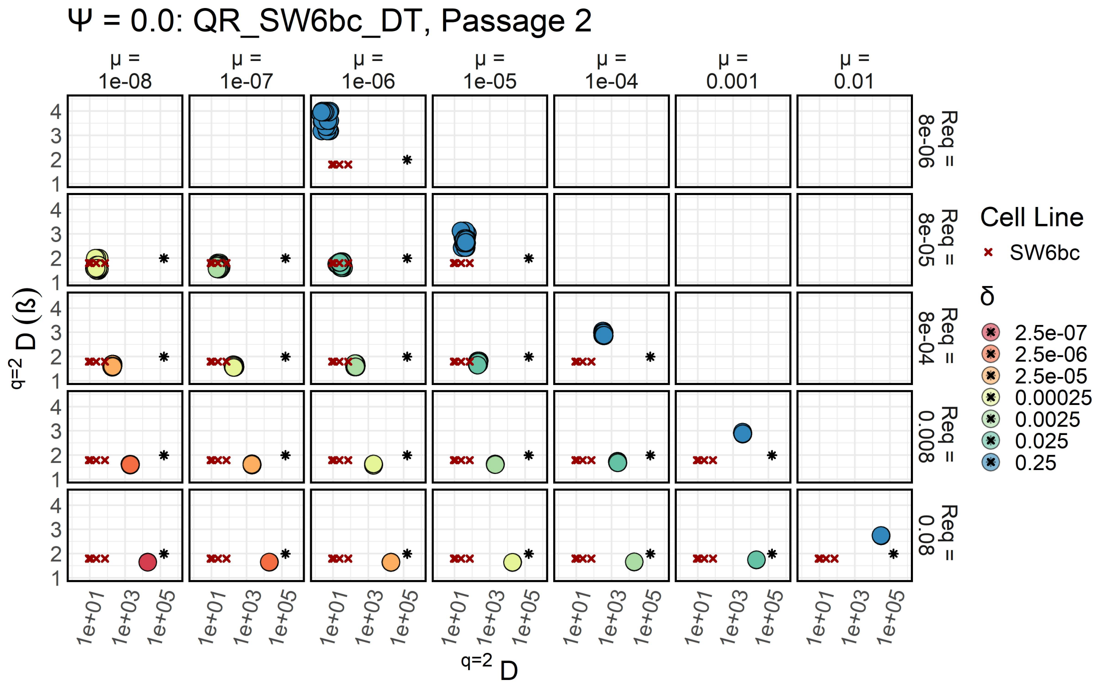


Figure 6.11: ${}^{q}D$ (within-replicate diversity) vs ${}^{q}D(\beta)$ (between-replicate diversity dissimilarity) of order $q = 2$ for the combined SW6bc simulation's Passage 2 drug-treatment (DT_P2) replicates. Any given panel corresponds to a simulation set that was run using the combination of parameters that control the resistant phenotype's evolution: columns correspond to values of μ, whilst rows correspond to values of δ that lead to the same equilibrium frequencies of resistance (R_{eq}). The black points in the bottom RHS of each individual panel corresponds to the control treatments' mean value. The coloured crosses correspond to the equivalent statistic values in the 'QR' SW6bc sequenced drug-treatment (SW6_DT_P2) replicates. The drug-treatment was modelled as deterministic ($\psi = 0.0$).

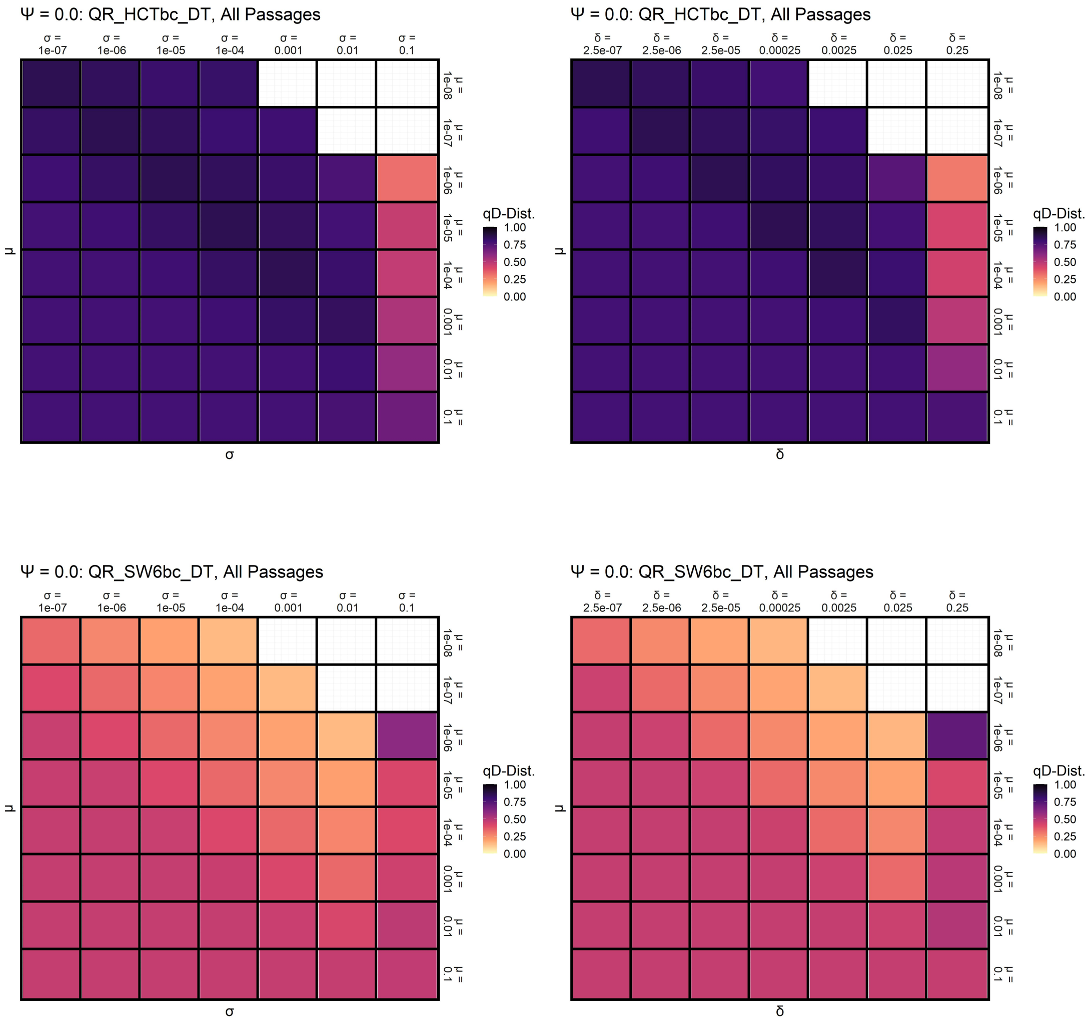


Figure 6.12: The normalised, average distance in log-statistic space ($log_{10}(^qD)$ vs $^qD(\beta)$) between simulated and sequenced drug-treatment Passages' lineage distributions from the 'QR' experiment over a range of parameter values that control the evolution of the resistant parameter: rows and columns within each panel correspond to values of μ and σ (LHS panels) or δ (RHS panels), respectively. Simulated values are derived from the deterministic drug-kill outputs ($\psi = 0.0$) using either `HCTbc` (top panels) or `SW6bc` (bottom panels) specific parameters. White panels correspond to parameter values where all simulations' end states were extinction. The drug-treatment was modelled as deterministic ($\psi = 0.0$).

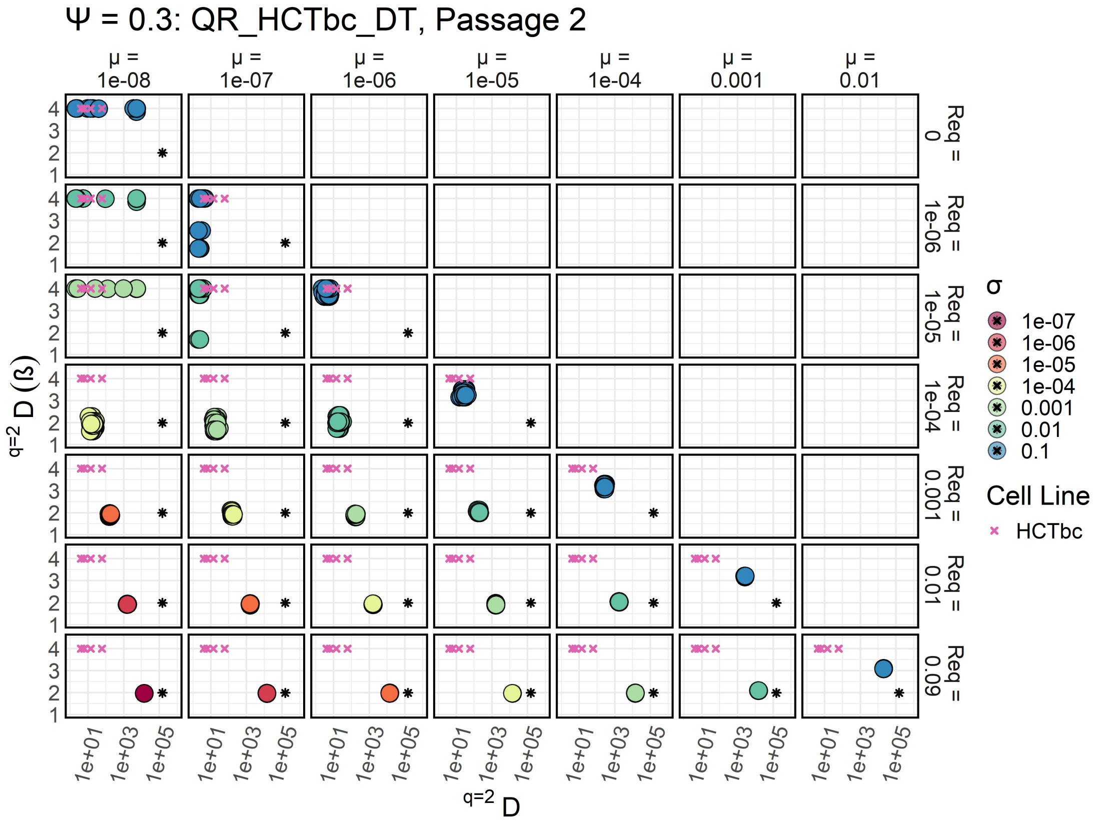


Figure 6.13: ${}^{q}D$ (within-replicate diversity) vs ${}^{q}D(\beta)$ (between-replicate diversity dissimilarity) of order $q = 2$ for the combined HCTbc simulation's Passage 2 drug-treatment (DT_P2) replicates. Any given panel corresponds to a simulation set that was run using the combination of parameters that control the resistant phenotype's evolution: columns correspond to values of μ, whilst rows correspond to values of σ that lead to the same equilibrium frequencies of resistance (R_{eq}). The black points in the bottom RHS of each individual panel corresponds to the control treatments' mean value. The coloured crosses correspond to the equivalent statistic values in the 'QR' HCTbc sequenced drug-treatment (HCT_DT_P2) replicates. The drug-treatment was modelled with a stochastic component ($\psi = 0.3$).

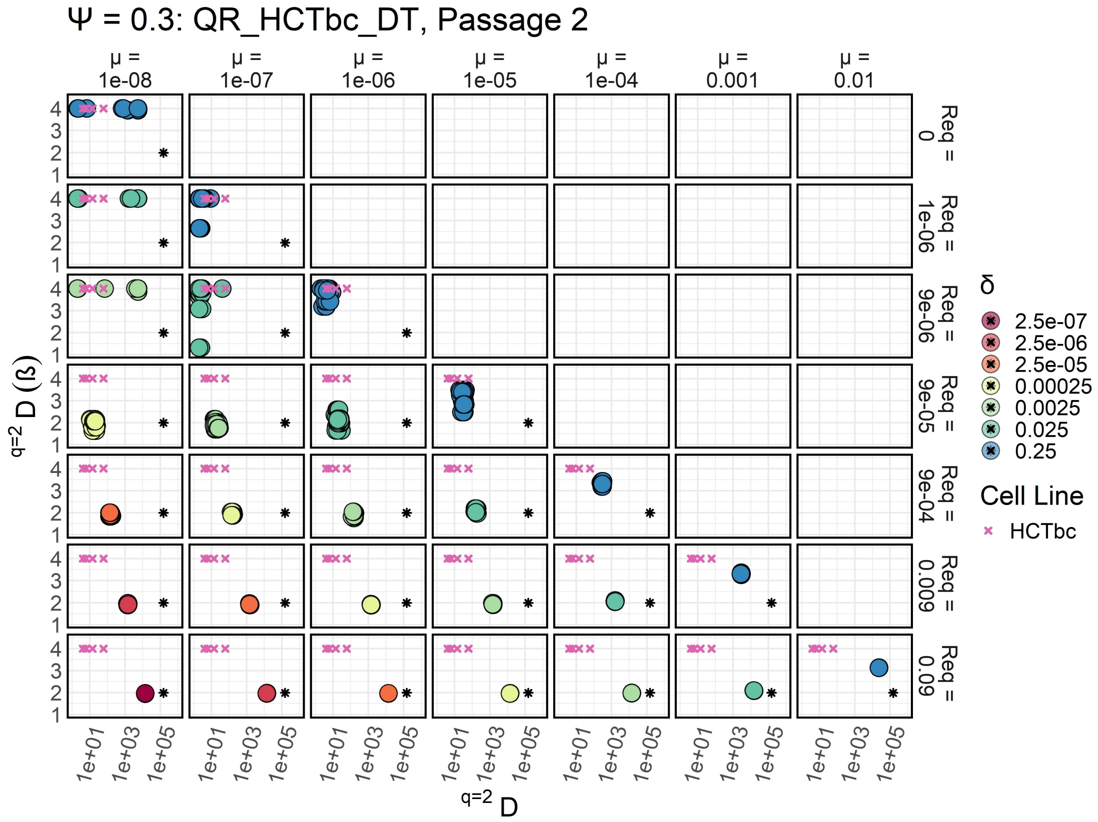


Figure 6.14: ^{q}D (within-replicate diversity) vs $^{q}D(\beta)$ (between-replicate diversity dissimilarity) of order $q = 2$ for the combined HCTbc simulation's Passage 2 drug-treatment (DT_P2) replicates. Any given panel corresponds to a simulation set that was run using the combination of parameters that control the resistant phenotype's evolution: columns correspond to values of μ, whilst rows correspond to values of δ that lead to the same equilibrium frequencies of resistance (R_{eq}). The black points in the bottom RHS of each individual panel corresponds to the control treatments' mean value. The coloured crosses correspond to the equivalent statistic values in the 'QR' HCTbc sequenced drug-treatment (HCT_DT_P2) replicates. The drug-treatment was modelled with a stochastic component ($\psi = 0.3$).

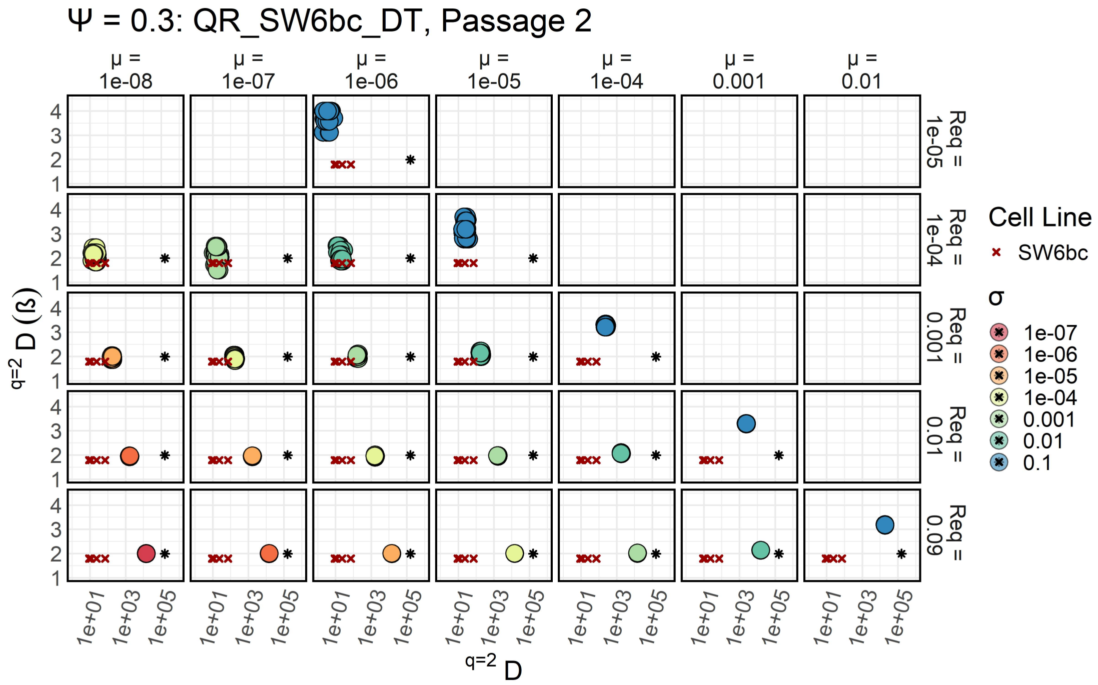


Figure 6.15: ^qD (within-replicate diversity) vs $^qD(\beta)$ (between-replicate diversity dissimilarity) of order $q = 2$ for the combined SW6bc simulation's Passage 2 drug-treatment (DT_P2) replicates. Any given panel corresponds to a simulation set that was run using the combination of parameters that control the resistant phenotype's evolution: columns correspond to values of μ, whilst rows correspond to values of σ that lead to the same equilibrium frequencies of resistance (R_{eq}). The black points in the bottom RHS of each individual panel corresponds to the control treatments' mean value. The coloured crosses correspond to the equivalent statistic values in the 'QR' SW6bc sequenced drug-treatment (SW6_DT_P2) replicates. The drug-treatment was modelled with a stochastic component ($\psi = 0.3$).

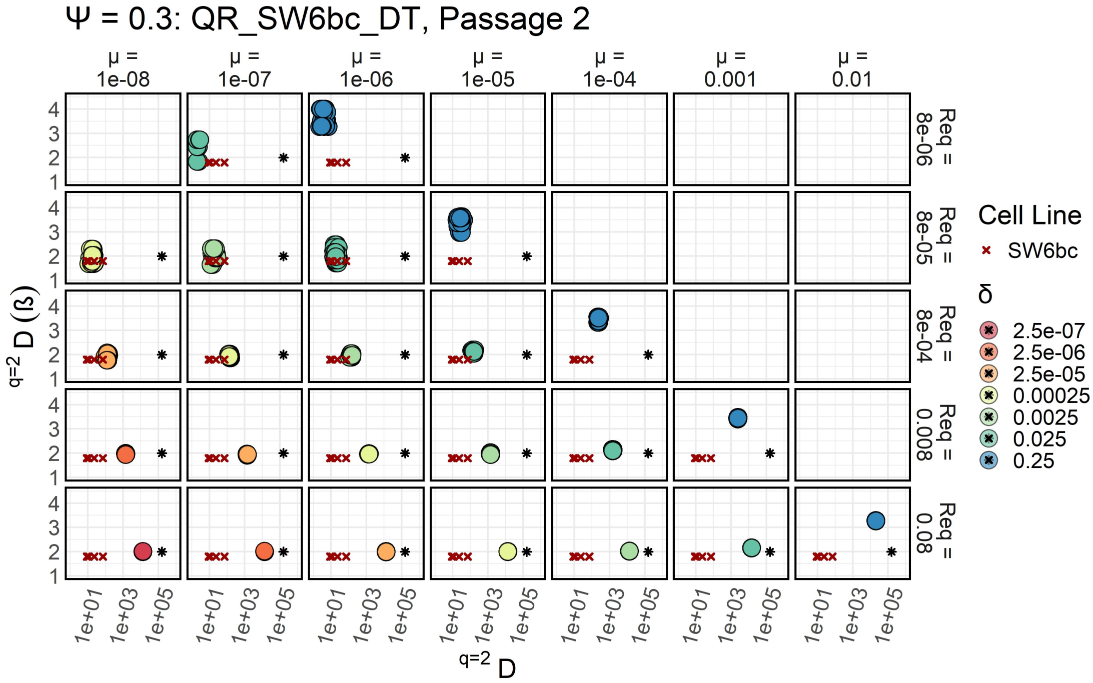
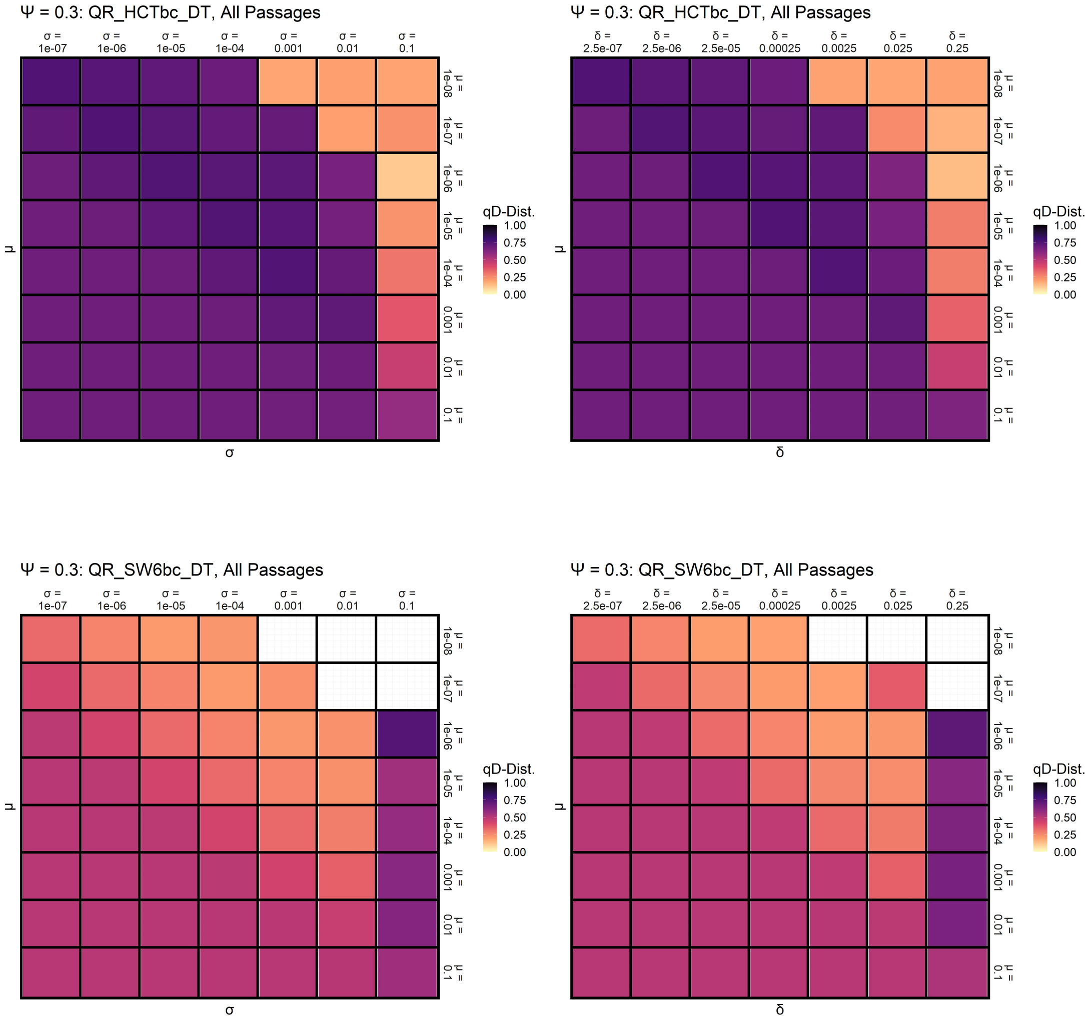


Figure 6.16: ${}^{q}D$ (within-replicate diversity) vs ${}^{q}D(\beta)$ (between-replicate diversity dissimilarity) of order $q = 2$ for the combined SW6bc simulation's Passage 2 drug-treatment (DT_P2) replicates. Any given panel corresponds to a simulation set that was run using the combination of parameters that control the resistant phenotype's evolution: columns correspond to values of μ, whilst rows correspond to values of δ that lead to the same equilibrium frequencies of resistance (R_{eq}). The black points in the bottom RHS of each individual panel corresponds to the control treatments' mean value. The coloured crosses correspond to the equivalent statistic values in the 'QR' SW6bc sequenced drug-treatment (SW6_DT_P2) replicates. The drug-treatment was modelled with a stochastic component ($\psi = 0.3$).


Ψ = 0.3: QR_HCTbc_DT, All Passages

Ψ = 0.3: QR_HCTbc_DT, All Passages

Ψ = 0.3: QR_SW6bc_DT, All Passages

Ψ = 0.3: QR_SW6bc_DT, All Passages

Figure 6.17: The normalised, average distance in log-statistic space ($log_{10}(^qD)$ vs $^qD(\beta)$) between simulated and sequenced drug-treatment Passages' lineage distributions from the 'QR' experiment over a range of parameter values that control the evolution of the resistant parameter: rows and columns within each panel correspond to values of μ and σ (LHS panels) or δ (RHS panels), respectively. Simulated values are derived from the deterministic drug-kill outputs ($\psi = 0.3$) using either HCTbc (top panels) or SW6bc (bottom panels) specific parameters. White panels correspond to parameter values where all simulations' end states were extinction.

By comparing to the simulated ${}^qD$ vs ${}^qD(\beta)$ summary statistics, each cell line's drug treatment replicates (HCTbc_DT and SW6bc_DT) can be categorised into two evolutionary scenarios: SW6bc drug-treatment replicates are consistent with a broad range of parameter values that all lead to similar equilibrium frequencies of resistance: $R_{eq} \approx 10^{-4}$ (Figures 6.10 and 6.11). These results remain true for the non-deterministic version of the simulation ($\psi = 0.3$, Figures 6.15 and 6.16). HCTbc is instead most consistent with an equilibrium frequency of resistance an order of magnitude lower: $R_{eq} \approx 10^{-5}$. The range of likely scenarios is also much narrower in HCTbc: in both the deterministic ($\psi = 0.0$) and stochastic versions ($\psi = 0.3$) of drug-induced death, the sequenced data are consistent with a 'resistance-conferring-mutation rate' (μ) of 10^{-6}, and either a reversion 'mutation' rate (σ) of 0.1, or a relative fitness cost of resistance (δ) of 0.25 (see the lowest *qD-Dist* values in Figures 6.12 and 6.17). These values imply that the resistance phenotpye is either costly (δ) or transient (σ). In fact, if the transience or the cost of the resistant phenotype is this high in the SW6bc simulations, these are the only parameter combinations that lead to an equilibrium frequency of $R_{eq} = 10^{-4}$ that are *not* consistent with the sequenced SW6bc results, for both the deterministic and non-deterministic models (Figures 6.12 and 6.17).

Relaxing the model so that death due to treatment is no longer deterministic ($\psi = 0.3$) broadens the range of parameter values that lead to an equilibrium frequency of $R_{eq} = 10^{-5}$ in the HCTbc simulations (Figure 6.17). Yet notably, if I consider other scenarios where $R_{eq} = 10^{-5}$, but with lower combinations of μ, σ or δ (rows $R_{eq} = 1e{-}05$ and $9e-06$ in Figures 6.13 and 6.14), the dynamics become too stochastic to reproduce the sequenced HCTbc_DT results as consistently as when $\mu = 10^{-6}$. The parameter combinations where *no* cells are assigned resistance at the beginning of the simulations - $\mu = 10^{-8}$ and $\sigma = 0.1$ or $\delta = 0.25$ also fail to consistently capture the HCTbc_DT results for the same reason: the model rules out any scenario where resistance arises solely *de novo* during the experiment. In both cell-lines, the similarity of the signatures left in the lineage distributions by either mode of phenotypic variability maintenance - either the reversion to sensitive phenotypic switching rate, $\sigma > 0.0$, or the relative fitness cost of resistance, $\delta > 0.0$ - mean the simulation cannot distinguish between the two scenarios.

6.4 Discussion

In this chapter I have combined theoretical expectations for lineage distributions under various evolutionary scenarios with sequenced data from a long-term evolutionary experiment in two colorectal cancer cell-lines. I have shown that the two cell-lines exhibit distinct population-dynamics in response to chemotherapy treatment *in vitro*.

In *HCT116*, a single case of each evolutionary scenario emerges as most consistent with the simulated distributions: either $\mu = 10^{-6}$ and $\sigma = 0.1$ or $\mu = 10^{-6}$ and $\delta = 0.25$. Notably, the model is able to distinguish between scenarios that lead to the same equilibrium frequencies of resistance (R_{eq}) and, therefore, the same proportion of resistance (on average) when treatment begins. In terms of within- and between-replicate diversity (${}^qD$ and ${}^qD(\beta)$), no other combination of parameters led to a consistently low enough number of successful lineages that differed between all replicates.

The model results that best capture the sequenced lineage distributions show that resistance is maintained at a much lower frequency in *HCT116*, which is consistent with an equilibrium frequency of $R_{eq} \approx 10^{-5}$, than in *SW620*, where it is closer to $R_{eq} \approx 10^{-4}$. As such, the overall selection experienced by the *HCT116* cell populations appears much stronger, and the dynamics were therefore more variable; when the population dynamics are dominated by a few resistant cells and waiting times for infrequent phenotypic transitions, the stochastic forces of lineage drift predominate early growth. In *HCT116*, this led initially to higher variance in between-replicate diversity measures: ${}^qD$. Whilst within-experiment replicate comparisons are indicative of stochastic dynamics in *HCT116*, the change in diversity measures when comparing the two experiments - 'PR' and 'QR' – are remarkably similar. It is possible that whilst any one replicate's successful lineages are the product of stochastic waiting times, the trajectory the combined sub-populations take towards resistance is broadly repeatable.

In *SW620* - a CIN CRC cell-line - the sequenced results are consistent with a wide range of parameters that lead to the same equilibrium frequency of resistance ($R_{eq} \approx 10^{-4}$) and, therefore, similar proportions of resistant cells when treatment begins. In the case where the phenotype frequency is controlled by the transition probabilities to and from resistance, the rates with which a cell's phenotype can move (per-division) - μ and σ - are too slow to break the correlation between lineage identity and phenotype (resistance). When, instead, it is a cost that controls the proportion of resistance, the relative fitness cost - δ - is too moderate to maintain resistant lineages at extremely low frequencies. At the start of the experiment, each cell is assigned a unique lineage marker. In both cases (σ and δ), the phenotype that lineages are assigned at the beginning of the experiment - either resistant or sensitive - does not change during the mutual expansion step, the assignment to replicate flasks nor treatment beginning. As such, both scenarios lead to identical signatures in the lineage distributions, and these results can be interpreted as a stable, resistant phenotype 'pre-existing' in the initial population. Importantly, however, opposed to being some arbitrary fraction, in my model the fraction of resistance is an emergent property of the simulations parameters: namely, the switch from sensitive to resistant, μ, the reverse rate, σ, or the relative fitness cost of resistance, δ. Although multiple parameter values lead to the same equilibrium frequency in *SW620*, the values of σ the model supports - between 10^{-4} and 10^{-2} - are too low to invoke some form of transcriptional memory, where epigenetic changes have been shown to invoke phenotype transitions that are maintained for several divisions (Shaffer, Emert, et al. 2018). Yet these rates are also likely too high to invoke genetic mutations (resistance conferring mutations are often discussed within the $10^{-9} - 10^{-7}$ range (Bhang et al. 2015; Acar, Nichol, Fernandez-Mateos, et al. 2020). Therefore, it is more likely that the scenarios where resistance is maintained in the population by some fitness cost ($\delta > 0.0$) responsible for the observed *SW620* results.

Comparisons of features such as the cumulative frequency distributions of lineages can provide clues as to how the resistant phenotype is distributed amongst cells within the population, whilst correlations between replicates can indicate how resistance is maintained within lineages. However, the conclusions drawn from these features alone are qualitative in nature. Furthermore, when the frequency of resistance in the population is small, it is difficult to distinguish stochastic sampling effects that are experimentally imposed from those that are due to the random nature of phenotype transitions, whether they are under genetic or non-genetic control. Here I have shown that by comparing diversity statistics of sequenced distributions with theoretical models where parameters control a resistance phenotype, I gain quantitative insights into the dynamics of resistance evolution. Interestingly, the two colorectal cancer cell-lines I investigate appear to exhibit quantitatively distinct responses to chemotherapy *in vitro*: both the frequency of resistance within each population and the stability or cost of the resistant phenotype differ.

Chapter 7

Summary and Outlook

7.1 Summary

In this thesis I set out to investigate the evolutionary dynamics of drug resistance in colorectal cancer. Important determinants of these dynamics are the rates at which cells transition between resistant and sensitive phenotypes and the relative fitness cost incurred by resistance. By estimating these values, I aimed to understand the underlying molecular processes controlling them. Namely, pre-existing and *de-novo* genetic mutations, and epigenetic innovations. I developed models that incorporated these features of resistance evolution, optimised an *in vitro* long-term experiment to test the predictions of the model, and then combined these results to understand how colorectal cancer cells evolve resistance when challenged with chemotherapy.

I designed the evolutionary model to simultaneously capture a range of rates, from those that resemble genetic mutations to those that are consistent with transient phenotypic states. In results chapter 1, I develop theoretical expectations for how lineages should be distributed under these various evolutionary scenarios. By choosing two summary statistics that capture the within and between replicate diversity (${}^{q}D$ and ${}^{q}D(\beta)$, respectively), I was able to distinguish between evolutionary scenarios, even when they led to the same equilibrium frequency of resistance. These differences are only perceivable due to the model recapitulating the most powerful facet of the *in vitro* experiment: namely, the parallel evolution of closely related, distinguishable individuals under the same selection pressures. The modelling revealed that when the rate of change from resistant to sensitive was frequent enough, or the relative fitness cost was high, resistance was kept at a low enough frequency in *individual lineages* that, following sampling into replicate sub-populations, the probability of lineage success being replicate-specific was high.

In results chapters 2 and 3, I characterised the behaviour of my chosen lineage tracing technology (ClonTracer - Bhang et al. 2015) *in vitro*. My model was contingent on the statistical behaviour of the lineage 'barcodes' used to track individual cell lineages. For example, importantly: the expanded complex plasmid library retained a high level of diversity and adhered to the semi-random nucleotide pattern that aids downstream filtering of sequenced barcodes; most infected cells contained a single, dominant barcode; and under standard culture conditions selection appeared weak, as demonstrated by the retention of many lineages at low frequencies. I leveraged information contained in control treatment cells sampled immediately after infection and expansion to infer the cell lines' birth and death rates. These rates informed the subsequent simulations that were used to distinguish evolutionary scenarios in the empirical, *sequenced* distributions. Finally, even prior to analysing sequenced barcode data, the behaviour of each cell line in response to treatment *in vitro* indicated there were intriguing differences between the two cell-lines: the difference in growth rates between drug-treatment passages and the existence of apparent drug-resistant 'colonies' in *HCT116* suggested a more stringent selective bottleneck and the accrual of rare, 'jackpot' events, when compared to the other cell line investigated, *SW620*.

Finally, in results chapter 4, I combined all of these results to infer the evolutionary dynamics of each cell-line to chemotherapy treatment during a long-term *in vitro* experiment. In *SW620*, the results were consistent with a range of parameter combinations that led to same proportion of pre-existing resistance. The similarity in these scenarios can be understood by considering the model of resistance evolution: if the probability of a cell switching between resistant and sensitive phenotypes per cell-division is not high enough, the correlation between lineage identity and phenotype remains constant during the shared expansion step, and it is not possible to distinguish between scenarios using differences in between-replicate diversity. Similarly, if the relative fitness cost of resistance is too low, the cost has negligible effect during the expansion step, and the dynamics are instead simply the product of the phenotypes cells were expressing when labelled with the lineage markers.

In *HCT116* - an MSI CRC cell-line – the difference in treatment response observed *in vitro* when compared to *SW620* was verified during analysis of the barcode data. Here, one evolutionary scenario was consistently supported by the data: the resistant phenotype was kept at low frequencies by either a high phenotypic switching rate from resistant to sensitive (per cell-division: $\sigma = 0.1$) or a high relative fitness cost ($\delta = 0.25$). The results excluded other parameter combinations that led to the same equilibrium frequency of resistance; the few numbers of lineages that became dominant in different replicates could not be explained by a resistant phenotype that was less costly or more stable during cell-division. In this final section, I will discuss the relevance of these results in the broader context of resistance evolution in cancer, and how these findings might influence the design of more effective treatment strategies.

7.2 Evolutionary Dynamics and Mechanisms

The paradigm of drug resistance in cancer has been one of resistance mutations that are either pre-exsiting or acquired (Wang et al. 2018; Iwasa et al. 2006; Misale et al. 2012; Diaz et al. 2012). Following treatment, the lineage harbouring the resistance mutation survives and eventually expands leading to disease progression. Here, I have shown that these dynamics are unlikely to be responsible for the evolution of drug resistance *in vitro* in my two colorectal cancer cell models.

The high rates at which the model predicts *HCT116* either reverts from the resistant to sensitive phenotype ($\sigma = 0.1$) or the relative fitness cost incurred by resistance ($\delta = 0.25$) are both consistent with recently identified responses cancer cells employ to evade treatment. Given the high transition from resistance to sensitivity, Shaffer and colleagues have shown that melanoma cells can exhibit a ‘transcriptional memory’ whereby a sub-population of individuals can express certain genes for several divisions. At any given time, a minority of cells can transcribe a set of genes associated with resistance at a higher level than the rest of the population. Following the addition of treatment, these differences are revealed via strong selection against any cell not expressing these resistance phenotypes (Shaffer, Dunagin, et al. 2017). Subsequent work has also shown that the coordinated expression of these genes can persist for several divisions (approximately 2-3) (Emert et al. 2020). My model can give rise to these dynamics by constraining the amount of time cells exist in the resistant state via the high reversion probability to sensitivity per cell division. One possible explanation for the high levels of between-replicate diversity differences observed in *HCT116* is that the resistant phenotype is the product of a rare subset of cells briefly expressing a set of genes that confer a higher fitness in the presence of 5-fluorouracil (my chemotherapeutic agent).

Whilst resistance is a binary trait in my models, others have modelled this stochastic variation in gene expression as a continuous trait, where resistance is conferred by expression above some threshold (Charlebois et al. 2011). In reality, these transient shifts in expression are continuous phenomenon. However, as selection in my experiment is strong in the drug-treatment environment – as illustrated by the low numbers of successful lineages – I argue that modelling resistance as a binary trait can be an acceptable sacrifice in favour of model tractability: the more stringent selection, the higher the expression threshold necessary for survival, and the lower the proportion of surviving cells that exhibit moderate expression levels.

The alternative mode of resistance that might best describe the *HCT116* data is the existence of a sub-population of quiescent cells, often coined ‘drug-tolerant persisters’ (DTPs). Numerous studies have shown that a rare population of cells can survive high concentrations of cytotoxic treatments whilst entering a state where cell division stops, or is reduced greatly (Marin-Bejar et al. 2021; Sharma et al. 2010; Liau et al. 2017). In patient-derived xenograft (PDX) models of colorectal cancer, this state was shown to be equipotent; all cells had equal capacity to become DTPs (Rehman et al. 2021). The high fitness cost incurred by resistant cells in a subset of my simulations can recreate dynamics similar to those experienced by cells in a quiescent state. Like the results of Rehman and colleagues (2021), my model permits cells to transition from the sensitive to resistant phenotype with equal probabilities. Whilst the highest relative fitness I consider ($\delta = 0.25$) means cells are dividing too fast to be considered 'quiescent', recent work by Oren *et al.* has shown that it is only the small proportion of DTPs that can continue to actively divide that go on to drive resistance (Oren et al. 2021). As such, if a sub-population of DTPs in *HCT116* are responsible for the observed dynamics, it is only this actively dividing fraction that would re-populate the replicate and be present in the sequenced lineage distributions.

One hypothesis for how DTPs might aid resistance evolution is by providing a reservoir of cells that can survive treatment long enough to accrue additional mutations that confer full resistance (Brock et al. 2009). My model simply distinguishes resistant cells from sensitive. Under this formulation, there is no distinction between cells that can merely survive the drug-treated environment – DTPs - and those that can actively proliferate in the presence of treatment - the traditional definition of resistance. If *HCT116* cells follow this pathway to resistance, where persistence is followed by a second molecular 'event' that confers full resistance, my model would underestimate the standing variation of a persister phenotype. Instead, the equilibrium frequency of resistance most consistent with my results - $R_{eq} \approx 10^{-5}$ - would represent the fraction of the population that make it from persistence to resistance. In fact, two of the *HCT116* drug-treatment replicates (in the 'QR' experiment) show evidence of clonal interference: namely, a rise in the number of successful lineages despite ongoing selection in the therapy condition. One explanation could be that tolerance acts as a short-term solution to survive treatment. Over time, evolution selects for resistance mutations that confer a higher fitness in the presence of chemotherapy. As the cells are an asexually evolving population, different beneficial mutations must compete on alternative genetic backgrounds. This could drive a transient increase in diversity, as more lineages accrue resistant mutations. Subsequently, the arrival of 'double mutant' lineages would lead to a concomitant drop in diversity again. Such dynamics were shown to explain a crash in diversity in *Saccharomyces cerevisiae* (Blundell, Schwartz, et al. 2019) (and can be observed in the time-series lineage distributions in Jasinska et al. 2020).

Following the long-term drug-treatment experiment, I now have a list of 'successful' lineages that either went on to dominate all/most replicates (in *SW620*) or each individual replicate (in *HCT116*). Future *in vitro* experiments could isolate these lineages from the original, expanded population (the `POT` samples). Isolated cells could be seeded in individual wells, expanded briefly, and then exposed to treatment. Such experiments resemble the famous 'Luria–Delbrück experiment' in *E.coli* and have now been employed in cancer several times to characterise the behaviour of cell's resistant phenotypes over time (Shaffer, Dunagin, et al. 2017; Russo et al. 2021). The variance between single-cell colonies that become resistant could help tease apart whether resistance is either costly, or the product of a transient phenotypic state. Additionally, a recent study was designed to tease apart the persister and resistance phenotypes: Russo and colleagues fit experimental cell-number trajectories during treatment to a Bayesian model and were able to derive estimates for the rate at which cells transitioned to persisters, as well as the rate at which these persister cells became fully resistant (Russo et al. 2021). If *HCT116* do adopt a persister phenotype, I could assess whether there is any overlap between cells that become persisters in a subsequent single-cell isolation experiment, and the resistant lineages in the original experiment. If the cells are primed to become persisters, there should be lineages that are shared in each of these groups. It might be that the accrual of additional changes that grant full resistance following the persister phenotype are stochastic, and it is these dynamics that drive differences in lineage success between replicates.

Future experiments can also be designed to better distinguish between either scenario. One approach is to leverage information held in the control replicates: if cells that are resistant confer some fitness cost in the non-treated environment, they should be less successful, on average, in the control replicates. Due to the various sources of biological and technical noise that arise in the lineage distributions - discussed throughout the thesis - this will require the development of further statistical models. Another approach would involve testing the proliferative capacity of cells derived from the original expanded pool *in vitro*, prior to any treatment exposure. If lower growth rates in the absence of drug correlated with success in the drug-treatment replicates in the original experiment, this would provide compelling evidence that resistance came with a relative fitness cost. Finally, single-cell RNA sequencing experiments could assess the gene expression of cells in the original expanded pool. Techniques have been developed that allow for the simultaneous genotyping and RNA-sequencing of single cells (Nam et al. 2019). I could therefore extract the barcode identity of cells as well as their expression data. The variance profile of certain genes between individual cells could provide evidence that the mechanism of resistance was driven by transient expression, as has been shown elsewhere (Shaffer, Emert, et al. 2018).

In contrast to persister dynamics or transcriptional memory, *SW620* could be described by simpler dynamics, where the lineages that came to dominate all replicates were resistant at the beginning of the experiment. *In vitro* lineage tracing has previously shown that resistance was the product of a rare pre-existing sub-population in lung cancer (Bhang et al. 2015; Acar, Nichol, Fernandez-Mateos, et al. 2020). However, opposed to a single lineage harbouring a single resistance mutation, in *SW620*, where it appears resistance could be genetically controlled, numerous lineages survive to appreciable frequencies despite 5 Passages of drug exposure *in vitro*. Whilst the stability of the phenotype during the experiment does strongly suggest that the mechanism is under genetic control, a next, logical step will be to identify what molecular features endow these cells with the ability to grow readily despite ongoing chemotherapy.

A difference between the two cell-lines employed in this study are the underlying class of genomic instability: *HCT116* is classed as microsatellite unstable (MSI), where a non-functional mismatch repair system leads to high numbers of insertions and deletions in the short tandem repeat regions of the genome (Wheeler and Bodmer 2000). In *SW620*, cells are chromosomally unstable (CIN), leading to high levels of aneuploidy. An unanswered question is to what extent these differences lead to different tempos of adaptive evolution. My results show that there is higher standing variation in *SW620* for resistance: the modelling predicts there is an equilibrium frequency of resistance of approximately 10^{-4}, compared to *HCT116*, where it is approximately $10^{-}5$. One possibility is that the chromosomal aberrations in *SW620* are more likely to hit genes that confer a selective benefit in the presence of chemotherapy. Work in yeast has shown that large chromosomal changes can act as ‘quick and crude’ sources of adaptive variation given strong selection that follows a sudden change in environmental pressures (Yona et al. 2012; Sunshine et al. 2015). In contrast, the insertions and deletions in *HCT116* have a lower probability of hitting any gene sequence involved in a resistance phenotype. This could explain why they might employ a survival mechanism that extends the time window for possible resistance mutations to accrue, such as the DTP phenotype.

One observation which is often taken as evidence in favour of non-genetic mechanisms of resistance is the lack of any clear genetic bottleneck in either lineage tracing markers (Rehman et al. 2021), or in phylogenies recreated from genetic sequencing data (Turati et al. 2021; Echeverria et al. 2019). My results imply that changes in diversity alone may be insufficient to infer whether the resistance mechanism is genetically controlled. A drop in any given population's diversity (as captured by the ${}^{q}D$ statistic I adopted) following treatment is a product of how widespread resistance is in the cell population. It is instead the difference between replicate populations' diversity that have evolved in parallel - ${}^{q}D(\beta)$ - that can capture how quickly cells in any given lineage transition from a sensitive to resistant phenotype, sources of information generally unique to experimental evolution. To give an example, cells that transiently express resistance genes appear to be rare in a population (Emert et al. 2021). As such, if a tumour survives an initial round of treatment due to the presence of these cells, this could lead to only several genetic lineages being observed in subsequent samples, despite no genetic mechanism controlling resistance. This has important consequences for attempting to track the mode of resistance via genetic sequencing of serially derived patient samples: If genetic bottlenecks cannot be taken as *prima facie* evidence of resistant phenotypes under genetic control, other metrics will have to be employed to distinguish which mechanism is at play.

7.3 Clinical Outlook

The mechanisms consistent with the dynamics observed in *HCT116* represent a challenge for effective treatment. If cells with no pre-existing resistance mutation can enter transient phenotypic states that are refractory to treatment – either by briefly expressing certain genes or becoming quiescent - this broadens the scope of potential targets that therapy must aim to limit or circumvent. One method would be to target the pathways that lead to resistance emerging and combine these drugs with chemo- and targeted therapies. Limiting cell transcription variability already forms the basis of one class of therapy. *KDM5* is a gene that encodes a histone demethylase, a family of enzymes involved in transcription regulation via epigenetic modifications. Hinohara and colleagues have shown that KDM5 inhibition in breast cancer can reduce resistance to endocrine therapies by reducing heterogeneity in gene transcription (Hinohara et al. 2018). It is possible that the 'transcriptional memory' observed elsewhere is also the product of dysfunctional epigenetic regulation, and that by restoring it the probability of cells expressing the resistant set of genes is diminished.

If instead cells adopt a DTP phenotype, treatments could aim to limit the probability that cells enter the quiescent state. Recent work in an *in vitro* model of lung cancer identified metabolic pathways that permitted a sub-population of persister cells to divide in the presence of targeted therapies (Oren et al. 2021). These included pathways involved in tolerating reactive oxygen species (ROS) and increased fatty acid oxidation (FAO), whilst the presence of DTPs and their reliance on these pathways was confirmed in several other cell-lines. Encouragingly, inhibition of glutathione synthesis – an intracellular compound which aids ROS tolerance – decreased the proportion of actively dividing persister cells. As persister cells appear to provide a means by which cancer cells evade chemo- targeted therapies, a promising avenue of treatment could be to first limit the emergence of DTPs by targeting persister-specific pathways before additional cytotoxic treatment.

The heterogeneity of tumours' genetic and phenotypic makeup has now been well documented (PCAWG Transcriptome Core Group et al. 2020). This has led to an appreciation that a 'one-size-fits-all' approach is insufficient. Instead, attention has shifted to personalised cancer treatment, where therapy is tailored to the molecular profile of a patient's tumour. This strategy relies, in part, on the predictability of cancer evolution: if there is no way to project the response of intervention by observing a tumour's current state, then there is little utility in a bespoke strategy for each patient‘s tumour. The stable, pre-existing resistance phenotype supported by the data in *SW620* is likely more amenable to a personalised strategy: the molecular features responsible for resistance could be identified prior to treatment, and the choice of therapies tailored accordingly. Choosing treatments based on a cancer's genetic makeup is already commonplace. For example, patients receiving the targeted therapy cetuximab in colorectal cancer are screened for a KRAS mutant, which renders the treatment ineffective (Grady and Pritchard 2014). If *SW620* resistance is the product of a genetic alteration, a similar strategy could be employed to pre-emptively circumvent resistant sub-populations.

The results from *HCT116* provide a greater challenge for implementing personalised treatments. Firstly, if every cell has an equal capacity to enter the resistant phenotype – as is consistent with my results, and those elsewhere (Rehman et al. 2021) – there is no rare sub-population that screening might aim to identify and target with alternative treatments. An important question is therefore whether molecular features predispose tumours to adopt non-genetic or DTP modes of resistance evasion. Whilst the lineages that become successful in each replicate were different, the *HCT116* dynamics were highly repeatable between each of my experiments. It is possible that, although low levels of resistance in the pool of cells lead to stochastic between-replicate dynamics, the tumour the cell-line was derived from was 'primed' to follow a given evolutionary route. Promising work using PDX models of melanoma show that replicate experiment tumours derived from the same patient do repeatedly adopt either genetic or non-genetic mechanisms of resistance, including a putative DTP phenotype (Marin-Bejar et al. 2021).

Adaptive therapy is an evolutionary informed approach where treatment is modulated to encourage competition between resistant and sensitive sub-populations, thereby extending the time before the tumour consists of solely resistant cells. Whilst there have been numerous theoretical discussions concerning the conditions under which adaptive therapy would be effective (Viossat and Noble 2021; West et al. 2018; Strobl et al. 2020; Gallaher, Brown, et al. 2018), empirical tests of the assumptions have been fewer (although see Bacevic et al. 2017). In the context of the two mechanisms consistent with

HCT116 in my results, adaptive therapy could benefit from either a high fitness cost of resistance or a high reversion rate to sensitivity. It has been shown that the efficacy of adaptive therapy is heightened if the resistant phenotype incurs a growth penalty (Viossat and Noble 2021). Alternatively, if cells have a relatively high probability of reverting to a sensitive phenotype, the number of sensitive cells with which the resistant sub-population will have to compete will increase during drug-holidays. As such, my results from *HCT116* provide additional empirical evidence that the conditions necessary for adaptive therapy may well operate in certain cancer cell populations.

7.4 Conclusion

In summary, I have shown that by combining evolutionary models, lineage tracing technology and a long-term drug-treatment experiment *in vitro*, I can draw quantitative conclusions regarding the evolution of a resistant phenotype. By formulating the simulations such that the evolutionary parameters can take a wide range of values, the model can simultaneously capture dynamics that resemble previously identified resistance mechanisms, including stable, pre-existing resistance, *de novo* mutations, transcriptional memory, and drug-tolerant persisters. This was crucial to distinguish two diverse routes to resistance employed by two colorectal cancer cell models. The framing of a tumour's response to therapy has often been perceived rigidly, where a tumour either responds to treatment, or resistance mutations render therapy a *fait accompli.* I hope my work has contributed to the burgeoning evidence that a broader conceptual model is necessary to capture the evolution of drug resistance in cancer.

Bibliography

Acar, Ahmet, Daniel Nichol, Javier Fernandez, et al. (2019). “Exploiting evolutionary herding to control drug resistance in cancer”. In: *bioRxiv*. DOI: 10.1101/566950.

Acar, Ahmet, Daniel Nichol, Javier Fernandez-Mateos, et al. (2020). “Exploiting evolutionary steering to induce collateral drug sensitivity in cancer”. In: *Nature Communications* 11.1, pp. 1–14. ISSN: 20411723. DOI: 10.1038/s41467-020-15596-z. URL: http://dx.doi.org/10.1038/s41467-020-15596-z.

Aktipis, C. Athena et al. (2011). “Overlooking Evolution: A Systematic Analysis of Cancer Relapse and Therapeutic Resistance Research”. In: *PLoS ONE* 6.11. ISSN: 19326203. DOI: 10.1371/journal.pone.0026100.

Arnold, Melina et al. (2017). “Global patterns and trends in colorectal cancer incidence and mortality”. In: *Gut*. ISSN: 14683288. DOI: 10.1136/gutjnl-2015-310912.

Baca, Sylvan C. et al. (2013). “Punctuated evolution of prostate cancer genomes”. In: *Cell* 153.3, pp. 666–677. ISSN: 00928674. DOI: 10.1016/j.cell.2013.03.021. URL: http://dx.doi.org/10.1016/j.cell.2013.03.021.

Bacevic, Katarina et al. (2017). “Spatial competition constrains resistance to targeted cancer therapy”. In: *Nature Communications* 8.1. ISSN: 20411723. DOI: 10.1038/s41467-017-01516-1. URL: http://dx.doi.org/10.1038/s41467-017-01516-1.

Bailey, N T J (1990). *The elements of stochastic processes with applications to the natural sciences*. ISBN: 0471523682. DOI: 10.2307/2004121.

Balaban, Nathalie Q. et al. (2004). “Bacterial persistence as a phenotypic switch”. In: *Science*. ISSN: 00368075. DOI: 10.1126/science.1099390.

Barbosa, Camilo et al. (2019). "Evolutionary stability of collateral sensitivity to antibiotics in the model pathogen pseudomonas aeruginosa". In: *eLife* 8, pp. 1–22. ISSN: 2050084X. DOI: 10.7554/eLife.51481.

Bardelli, Alberto et al. (2013). "Amplification of the MET receptor drives resistance to anti-EGFR therapies in colorectal cancer". In: *Cancer Discovery* 3.6, pp. 658–673. ISSN: 21598274. DOI: 10.1158/2159-8290.CD-12-0558.

Basanta, David and Alexander R.A. Anderson (2013). "Exploiting ecological principles to better understand cancer progression and treatment". In: *Interface Focus* 3.4. ISSN: 20428901. DOI: 10.1098/rsfs.2013.0020.

Berg, Kaja CG et al. (2017). "Multi-omics of 34 colorectal cancer cell lines - a resource for biomedical studies". In: *Molecular Cancer* 16.1, pp. 1–16. ISSN: 14764598. DOI: 10.1186/s12943-017-0691-y.

Bhang, Hyo Eun C. et al. (2015). "Studying clonal dynamics in response to cancer therapy using high-complexity barcoding". In: *Nature Medicine* 21.5, pp. 440–448. ISSN: 1546170X. DOI: 10.1038/nm.3841.

Birkbak, Nicolai J. et al. (2011). "Paradoxical relationship between chromosomal instability and survival outcome in cancer". In: *Cancer Research* 71.10, pp. 3447–3452. ISSN: 00085472. DOI: 10.1158/0008-5472.CAN-10-3667.

Blundell, Jamie R. and Sasha F. Levy (2014). "Beyond genome sequencing: Lineage tracking with barcodes to study the dynamics of evolution, infection, and cancer". In: *Genomics* 104, pp. 1–14. ISSN: 10898646. DOI: 10.1016/j.ygeno.2014.09.005.

Blundell, Jamie R., Katja Schwartz, et al. (2019). "The dynamics of adaptive genetic diversity during the early stages of clonal evolution". In: *Nature Ecology and Evolution* 3.2, pp. 293–301. ISSN: 2397334X. DOI: 10.1038/s41559-018-0758-1. URL: http://dx.doi.org/10.1038/s41559-018-0758-1.

Bozic, Ivana and Martin A. Nowak (2014). "Timing and heterogeneity of mutations associated with drug resistance in metastatic cancers". In: *Proceedings of the National Academy of Sciences of the United States of America* 111.45, pp. 15964–15968. ISSN: 10916490. DOI: 10.1073/pnas.1412075111.

Bridgewater, John A. et al. (2020). “Systemic chemotherapy with or without cetuximab in patients with resectable colorectal liver metastasis (New EPOC): long-term results of a multicentre, randomised, controlled, phase 3 trial”. In: *The Lancet Oncology* 21.3, pp. 398–411. ISSN: 14745488. DOI: 10.1016/S1470-2045(19)30798-3.

Brock, Amy et al. (2009). “Non-genetic heterogeneity a mutation-independent driving force for the somatic evolution of tumours”. In: *Nature Reviews Genetics* 10.5, pp. 336–342. ISSN: 14710056. DOI: 10.1038/nrg2556.

Bystrykh, Leonid V and Mirjam E Belderbos (2016). “Clonal Analysis of Cells with Cellular Barcoding : When Numbers and Sizes Matter”. In: *Stem Cell Heterogeneity.* Ed. by Kursad Turksen. 1516th ed. New York: Humana Press, New York, NY. Chap. 4, pp. 57–89. ISBN: 978-1-4939-6550-2. DOI: https://doi.org/10.1007/978-1-4939-6550-2.

Charlebois, Daniel A. et al. (2011). “Gene expression noise facilitates adaptation and drug resistance independently of mutation”. In: *Physical Review Letters* 107.21, pp. 1–5. ISSN: 00319007. DOI: 10.1103/PhysRevLett.107.218101.

Cohen, Nadia R. et al. (2013). *Microbial persistence and the road to drug resistance.* DOI: 10.1016/j.chom.2013.05.009.

Connell, Joseph H. (1961). “The Influence of Interspecific Competition and Other Factors on the Distribution of the Barnacle Chthamalus Stellatus”. In: *Ecology.* ISSN: 00129658. DOI: 10.2307/1933500.

Copija, Angelika et al. (2017). “Clinical significance and prognostic relevance of microsatellite instability in sporadic colorectal cancer patients”. In: *International Journal of Molecular Sciences* 18.1, pp. 1–12. ISSN: 14220067. DOI: 10.3390/ijms18010107.

Costello, Maura et al. (2017). “Characterization and remediation of sample index swaps by non-redundant dual indexing on massively parallel sequencing platforms”. In: *bioRxiv*, pp. 1–10. DOI: 10.1101/200790.

Couce, Alejandro and Olivier A. Tenaillon (2015). “The rule of declining adaptability in microbial evolution experiments”. In: *Frontiers in Genetics* 6.MAR, pp. 1–6. ISSN: 16648021. DOI: 10.3389/fgene.2015.00099.

Cree, Ian A. and Peter Charlton (2017). "Molecular chess? Hallmarks of anti-cancer drug resistance". In: *BMC Cancer* 17.1, pp. 1–8. ISSN: 14712407. DOI: 10.1186/s12885-016-2999-1. URL: http://dx.doi.org/10.1186/s12885-016-2999-1.

Cross, William et al. (2018). "The evolutionary landscape of colorectal tumorigenesis". In: *Nature Ecology and Evolution* 2.10, pp. 1661–1672. ISSN: 2397334X. DOI: 10.1038/s41559-018-0642-z.

Dewhurst, Sally M. et al. (2014). "Tolerance of whole- genome doubling propagates chromosomal instability and accelerates cancer genome evolution". In: *Cancer Discovery* 4.2, pp. 175–185. ISSN: 21598274. DOI: 10.1158/2159-8290.CD-13-0285.

Dhawan, Andrew et al. (2017). "Collateral sensitivity networks reveal evolutionary instability and novel treatment strategies in ALK mutated non-small cell lung cancer". In: *Scientific Reports* 7.1, pp. 1–9. ISSN: 20452322. DOI: 10.1038/s41598-017-00791-8. URL: http://dx.doi.org/10.1038/s41598-017-00791-8.

Diaz, Luis A. et al. (2012). "The molecular evolution of acquired resistance to targeted EGFR blockade in colorectal cancers". In: *Nature* 486.7404, pp. 537–540. ISSN: 00280836. DOI: 10.1038/nature11219.

Dieter, Sebastian M. et al. (2011). "Distinct types of tumor-initiating cells form human colon cancer tumors and metastases". In: *Cell Stem Cell* 9.4, pp. 357–365. ISSN: 19345909. DOI: 10.1016/j.stem.2011.08.010.

Drost, Jarno et al. (2015). "Sequential cancer mutations in cultured human intestinal stem cells". In: *Nature* 521.7550, pp. 43–47. ISSN: 14764687. DOI: 10.1038/nature14415.

Duan, Guihua et al. (2018). "Increased Glutamine Consumption in Cisplatin-Resistant Cells Has a Negative Impact on Cell Growth". In: *Scientific Reports* 8.1, pp. 1–11. ISSN: 20452322. DOI: 10.1038/s41598-018-21831-x. URL: http://dx.doi.org/10.1038/s41598-018-21831-x.

Durrett, Richard (2015). "Branching Process Models of Cancer". In: *Branching Process Models of Cancer*. ISBN: 9783319160641. DOI: 10.1007/978-3-319-16065-8{_}1.

Echeverria, Gloria V. et al. (2019). "Resistance to neoadjuvant chemotherapy in triple-negative breast cancer mediated by a reversible drug-tolerant state". In: *Science*

Translational Medicine 11.488. ISSN: 19466242. DOI: 10.1126/scitranslmed.aav0936.

Emert, Benjamin L et al. (2020). "Retrospective identification of rare cell populations underlying drug resistance connects molecular variability with cell fate". In: *bioRxiv*.

— (2021). "Variability within rare cell states enables multiple paths toward drug resistance". In: *Nature Biotechnology* 39.7, pp. 865–876. ISSN: 15461696. DOI: 10.1038/s41587-021-00837-3. URL: http://dx.doi.org/10.1038/s41587-021-00837-3.

Eyre-Walker, Adam and Peter D. Keightley (2007). "The distribution of fitness effects of new mutations". In: *Nature Reviews Genetics* 8.8, pp. 610–618. ISSN: 14710056. DOI: 10.1038/nrg2146.

Fearon, Eric R. and Bert Vogelstein (1990). *A genetic model for colorectal tumorigenesis.* DOI: 10.1016/0092-8674(90)90186-I.

Fehse, B. et al. (2004). "Pois(s)on - It's a question of dose.." In: *Gene Therapy* 11.11, pp. 879–881. ISSN: 09697128. DOI: 10.1038/sj.gt.3302270.

Fisher, R. A. (1958). "The genetical theory of natural selection, 2nd edn." In: *Dover Publication, New York.* ISSN: 0016-6731. DOI: 10.1111/jeb.12566.

Fogle, Craig A. et al. (2008). "Clonal interference, multiple mutations and adaptation in large asexual populations". In: *Genetics* 180.4, pp. 2163–2173. ISSN: 00166731. DOI: 10.1534/genetics.108.090019.

Foo, Jasmine et al. (2013). "Cancer as a moving target: Understanding the composition and rebound growth kinetics of recurrent tumors". In: *Evolutionary Applications* 6.1, pp. 54–69. ISSN: 17524563. DOI: 10.1111/eva.12019.

Gabriel, W. et al. (1993). "Muller's Ratchet and Mutational Meltdowns". In: *Evolution.* ISSN: 00143820. DOI: 10.2307/2410218.

Gallaher, Jill A, Joel Brown, et al. (2018). "The dynamic tumor ecosystem : how cell turnover and trade-offs affect cancer evolution". In:

Gallaher, Jill A, Pedro M Enriquez-Navas, et al. (2017). "Adaptive vs continuous cancer therapy: Exploiting space and trade-offs in drug scheduling". In: *bioRxiv* 3, p. 128959. DOI: 10.1101/128959. URL: https://www.biorxiv.org/content/early/2017/05/30/128959.full.pdf+html.

Gatenby, Robert A (2009). “A change of strategy in the war on cancer”. In: *Nature* 459.7246, pp. 508–509. ISSN: 00280836. DOI: 10.1038/459508a.

Gerrish, Phillip J and Richard E Lenski (1998). “The fate of competing beneficial mutations in an asexual population.” In: *Genetica* 102-103.1-6, pp. 127–144. ISSN: 0016-6707. DOI: 10.1023/A:1017067816551. arXiv: 0005074v1 [astro-ph].

Giam, Maybelline and Giulia Rancati (2015). “Aneuploidy and chromosomal instability in cancer: A jackpot to chaos”. In: *Cell Division* 10.1, pp. 1–12. ISSN: 17471028. DOI: 10.1186/s13008-015-0009-7. URL: ???.

Giessler, Klara M et al. (2017). “Genetic subclone architecture of tumor clone-initiating cells in colorectal cancer”. In:

Glasspool, R. M. et al. (2006). “Epigenetics as a mechanism driving polygenic clinical drug resistance”. In: *British Journal of Cancer* 94.8, pp. 1087–1092. ISSN: 00070920. DOI: 10.1038/sj.bjc.6603024.

Gluzman, Mark et al. (2020). “Optimizing adaptive cancer therapy: Dynamic programming and evolutionary game theory”. In: *Proceedings of the Royal Society B: Biological Sciences* 287.1925. ISSN: 14712954. DOI: 10.1098/rspb.2019.2454.

Goldman, Aaron et al. (2015). “Temporally sequenced anticancer drugs overcome adaptive resistance by targeting a vulnerable chemotherapy-induced phenotypic transition”. In: *Nature Communications* 6, pp. 1–13. ISSN: 20411723. DOI: 10.1038/ncomms7139. URL: http://dx.doi.org/10.1038/ncomms7139.

Gordon, David J. et al. (2012). *Causes and consequences of aneuploidy in cancer.* DOI: 10.1038/nrg3123.

Grady, William M. and Colin C. Pritchard (2014). “Molecular alterations and biomarkers in colorectal cancer”. In: *Toxicologic Pathology* 42.1, pp. 124–139. ISSN: 01926233. DOI: 10.1177/0192623313505155.

Graham, Trevor A. and Andrea Sottoriva (2017). “Measuring cancer evolution from the genome”. In: *Journal of Pathology* 241.2, pp. 183–191. ISSN: 10969896. DOI: 10.1002/path.4821.

Greaves, Mel (2018). “Nothing in cancer makes sense except.” In: *BMC Biology* 16.1, pp. 1–8. ISSN: 17417007. DOI: 10.1186/s12915-018-0493-8.

Greaves, Mel and Carlo C. Maley (2012). “Clonal evolution in cancer”. In: *Nature* 481.7381, pp. 306–313. ISSN: 00280836. DOI: 10.1038/nature10762.

Guastadisegni, Cecilia et al. (2010). “Microsatellite instability as a marker of prognosis and response to therapy: A meta-analysis of colorectal cancer survival data”. In: *European Journal of Cancer*. ISSN: 09598049. DOI: 10.1016/j.ejca.2010.05.009.

Gupta, Piyush B. et al. (2011). “Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells”. In: *Cell* 146.4, pp. 633–644. ISSN: 10974172. DOI: 10.1016/j.cell.2011.07.026. URL: http://dx.doi.org/10.1016/j.cell.2011.07.026.

Gymrek, Melissa et al. (2015). “Abundant contribution of short tandem repeats to gene expression variation in humans”. In: *Nature Genetics*. ISSN: 15461718. DOI: 10.1038/ng.3461.

Hata, Aaron N. et al. (2016). “Tumor cells can follow distinct evolutionary paths to become resistant to epidermal growth factor receptor inhibition”. In: *Nature Medicine* 22.3, pp. 262–269. ISSN: 1546170X. DOI: 10.1038/nm.4040.

Hinohara, Kunihiko et al. (2018). “KDM5 Histone Demethylase Activity Links Cellular Transcriptomic Heterogeneity to Therapeutic Resistance”. In: *Cancer Cell* 34.6, pp. 939–953. ISSN: 18783686. DOI: 10.1016/j.ccell.2018.10.014.

Holohan, Caitriona et al. (2013). “Cancer drug resistance: An evolving paradigm”. In: *Nature Reviews Cancer* 13.10, pp. 714–726. ISSN: 1474175X. DOI: 10.1038/nrc3599. URL: http://dx.doi.org/10.1038/nrc3599.

Huang, Weichun et al. (2012). “ART: A next-generation sequencing read simulator”. In: *Bioinformatics* 28.4, pp. 593–594. ISSN: 13674803. DOI: 10.1093/bioinformatics/btr708.

Hudler, Petra (2012). *Genetic aspects of gastric cancer instability.* DOI: 10.1100/2012/761909.

Humphries, Adam et al. (2013). “Lineage tracing reveals multipotent stem cells maintain human adenomas and the pattern of clonal expansion in tumor evolution”. In: *Proceedings of the National Academy of Sciences of the United States of America* 110.27. ISSN: 00278424. DOI: 10.1073/pnas.1220353110.

Hutchins, Gordon et al. (2011). "Value of mismatch repair, KRAS, and BRAF mutations in predicting recurrence and benefits from chemotherapy in colorectal cancer". In: *Journal of Clinical Oncology.* ISSN: 0732183X. DOI: 10.1200/JCO.2010.30.1366.

Hveem, T. S. et al. (2014). "Prognostic impact of genomic instability in colorectal cancer." In: *British journal of cancer* 110.8, pp. 2159–2164. ISSN: 15321827. DOI: 10.1038/bjc.2014.133.

Iwasa, Yoh et al. (2006). "Evolution of resistance during clonal expansion". In: *Genetics* 172.4, pp. 2557–2566. ISSN: 00166731. DOI: 10.1534/genetics.105.049791.

Jasinska, Weronika et al. (2020). "Chromosomal barcoding of E. coli populations reveals lineage diversity dynamics at high resolution". In: *Nature Ecology and Evolution* 4.3, pp. 437–452. ISSN: 2397334X. DOI: 10.1038/s41559-020-1103-z.

Johnson, Kaitlyn E. et al. (2019). "Cancer cell population growth kinetics at low densities deviate from the exponential growth model and suggest an Allee effect". In: *PLoS Biology* 17.8, pp. 1–29. ISSN: 15457885. DOI: 10.1371/journal.pbio.3000399.

Jost, Lou (2006). "Entropy and diversity". In: *Oikos* 113.2, pp. 363–375.

Jung, Barbara et al. (2004). "Loss of Activin Receptor Type 2 Protein Expression in Microsatellite Unstable Colon Cancers". In: *Gastroenterology.* ISSN: 00165085. DOI: 10.1053/j.gastro.2004.01.008.

Kam, Yoonseok et al. (2015). "Sweat but no gain: Inhibiting proliferation of multidrug resistant cancer cells with ersatzdroges". In: *International Journal of Cancer* 136.4, E188–E196. ISSN: 10970215. DOI: 10.1002/ijc.29158.

Kebschull, Justus M. and Anthony M. Zador (2018). "Cellular barcoding: lineage tracing, screening and beyond". In: *Nature Methods* 15.11, pp. 871–879. ISSN: 15487105. DOI: 10.1038/s41592-018-0185-x. URL: http://dx.doi.org/10.1038/s41592-018-0185-x.

Kester, Lennart and Alexander van Oudenaarden (2018). "Single-Cell Transcriptomics Meets Lineage Tracing". In: *Cell Stem Cell* 23.2, pp. 166–179. ISSN: 18759777. DOI: 10.1016/j.stem.2018.04.014. URL: https://doi.org/10.1016/j.stem.2018.04.014.

Kim, Eric S. (2016). “Chemotherapy Resistance in Lung Cancer”. In: *Advances in Experimental Medicine and Biology* 890, pp. 37–56. ISSN: 22148019. DOI: 10.1007/978-3-319-24932-2{_}3. URL: http://dx.doi.org/10.1007/978-3-319-24932-2_3%5Cnhttp://link.springer.com/chapter/10.1007%2F978-3-319-24932-2_3%5Cnhttps://www.ncbi.nlm.nih.gov/pubmed/26703798%0Ahttp://link.springer.com/10.1007/978-3-319-24932-2_3.

Kim, Hyeonhui et al. (2018). “Mouse Cre-LoxP system: general principles to determine tissue-specific roles of target genes”. In: *Laboratory Animal Research.* ISSN: 1738-6055. DOI: 10.5625/lar.2018.34.4.147.

Knopp, Michael and Dan I. Andersson (2015). “Amelioration of the fitness costs of antibiotic resistance due to reduced outer membrane permeability by upregulation of alternative porins”. In: *Molecular Biology and Evolution* 32.12, pp. 3252–3263. ISSN: 15371719. DOI: 10.1093/molbev/msv195.

Koopman, M. et al. (2009). “Deficient mismatch repair system in patients with sporadic advanced colorectal cancer”. In: *British Journal of Cancer* 100.2, pp. 266–273. ISSN: 00070920. DOI: 10.1038/sj.bjc.6604867.

Korolev, Kirill S et al. (2012). “Selective sweeps in growing microbial populations”. In: *Phys. Biol* 9.2. ISSN: 15378276. DOI: 10.1038/jid.2014.371.

Kretzschmar, Kai and Fiona M. Watt (2012). “Lineage tracing”. In: *Cell* 148.1-2, pp. 33–45. ISSN: 00928674. DOI: 10.1016/j.cell.2012.01.002. URL: http://dx.doi.org/10.1016/j.cell.2012.01.002.

Kuipers, Ernst J. et al. (2015). “Colorectal cancer”. In: *Nature Reviews* 1, pp. 127–135. DOI: 10.1038/nrdp.2015.65. URL: http://dx.doi.org/10.1038/nrdp.2015.65.

Lamprecht, Sebastian et al. (2017). “Multicolor lineage tracing reveals clonal architecture and dynamics in colon cancer”. In: *Nature Communications* 8.1, pp. 1–8. ISSN: 20411723. DOI: 10.1038/s41467-017-00976-9. URL: http://dx.doi.org/10.1038/s41467-017-00976-9.

Lee, Alvin J.X. et al. (2011). “Chromosomal instability confers intrinsic multidrug resistance”. In: *Cancer Research* 71.5, pp. 1858–1870. ISSN: 00085472. DOI: 10.1158/0008-5472.CAN-10-3604.

Lenormand, Thomas et al. (2018). "Cost of resistance: an unreasonably expensive concept". In: *Rethinking Ecology* 3, pp. 51–70. ISSN: 2534-9260. DOI: 10.3897/rethinkingecology.3.31992.

Lenski, Richard E (2017). "Experimental evolution and the dynamics of adaptation and genome evolution in microbial populations". In: *The ISME Journal* 11.10, pp. 2181–2194. ISSN: 1751-7362. DOI: 10.1038/ismej.2017.69. URL: http://www.nature.com/doifinder/10.1038/ismej.2017.69.

Levy, Sasha F. et al. (2015). "Quantitative evolutionary dynamics using high-resolution lineage tracking". In: *Nature* 519.7542, pp. 181–186. ISSN: 14764687. DOI: 10.1038/nature14279.

Li, Benshang et al. (2020). "Therapy-induced mutations drive the genomic landscape of relapsed acute lymphoblastic leukemia". In: *Blood* 135.1, pp. 41–55. ISSN: 15280020. DOI: 10.1182/blood.2019002220.

Liau, Brian B. et al. (2017). "Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance". In: *Cell Stem Cell* 20.2, pp. 233–246. ISSN: 18759777. DOI: 10.1016/j.stem.2016.11.003.

Liedtke, Cornelia et al. (2008). "Response to neoadjuvant therapy and long-term survival in patients with triple-negative breast cancer". In: *Journal of Clinical Oncology*. ISSN: 0732183X. DOI: 10.1200/JCO.2007.14.4147.

Llosa, Nicolas J. et al. (2015). "The vigorous immune microenvironment of microsatellite instable colon cancer is balanced by multiple counter-inhibitory checkpoints". In: *Cancer Discovery* 5.1, pp. 43–51. ISSN: 21598290. DOI: 10.1158/2159-8290.CD-14-0863.

Lopez, Saioa et al. (2019). "Whole Genome Doubling mitigates Muller's Ratchet in Cancer Evolution". In: *bioRxiv* 3, p. 513457. DOI: 10.1101/513457. URL: https://www.biorxiv.org/content/early/2019/01/07/513457.

Ludwig, Leif S. et al. (2019). "Lineage Tracing in Humans Enabled by Mitochondrial Mutations and Single-Cell Genomics". In: *Cell*. ISSN: 10974172. DOI: 10.1016/j.cell.2019.01.022.

Marin-Bejar, Oskar et al. (2021). "Evolutionary predictability of genetic versus non-genetic resistance to anticancer drugs in melanoma". In: *Cancer Cell* 39.8, pp. 1135–1149. ISSN: 18783686. DOI: 10.1016/j.ccell.2021.05.015. URL: https://doi.org/10.1016/j.ccell.2021.05.015.

Markowitz, Sanford D. and Monica M. Bertagnolli (2009). "Molecular basis of colorectal cancer". In: *New England Journal of Medicine* 361.25, p. 2449. ISSN: 15334406. DOI: 10.1056/NEJMra0804588.

Marusyk, Andriy et al. (2012). "Intra-tumour heterogeneity: A looking glass for cancer?" In: *Nature Reviews Cancer* 12.5, pp. 323–334. ISSN: 1474175X. DOI: 10.1038/nrc3261.

Mayr, Ernst (1989). "Speciational evolution or punctuated equilibria". In: *Journal of Social and Biological Systems* 12.2-3, pp. 137–158. ISSN: 01401750. DOI: 10.1016/0140-1750(89)90041-9.

Michod, Richard E. et al. (2006). "Life-history evolution and the origin of multicellularity". In: *Journal of Theoretical Biology.* ISSN: 00225193. DOI: 10.1016/j.jtbi.2005.08.043.

Michor, Franziska et al. (2004). "Dynamics of cancer progression". In: *Nature Reviews Cancer* 4.3, pp. 197–205. ISSN: 1474175X. DOI: 10.1038/nrc1295.

Misale, Sandra et al. (2012). "Emergence of KRAS mutations and acquired resistance to anti-EGFR therapy in colorectal cancer". In: *Nature* 486.7404, pp. 532–536. ISSN: 00280836. DOI: 10.1038/nature11156. URL: http://dx.doi.org/10.1038/nature11156.

Mok, TS et al. (2017). "Osimertinib or platinum-pemetrexed in EGFR T790M-Positive lung cancer". In: *New England Journal of Medicine.* ISSN: 15334406. DOI: 10.1056/NEJMoa1612674.

Nam, Anna S. et al. (2019). "Somatic mutations and cell identity linked by Genotyping of Transcriptomes". In: *Nature* 571.7765, pp. 355–360. ISSN: 14764687. DOI: 10.1038/s41586-019-1367-0. URL: http://dx.doi.org/10.1038/s41586-019-1367-0.

Nguyen, Ha and Hong-Quan Duong (2018). “The molecular characteristics of colorectal cancer: Implications for diagnosis and therapy (Review)”. In: *Oncology Letters*, pp. 9–18. ISSN: 1792-1074. DOI: 10.3892/ol.2018.8679. URL: http://www.spandidos-publications.com/10.3892/ol.2018.8679.

Nikolaou, Michail et al. (2018). “The challenge of drug resistance in cancer treatment: a current overview”. In: *Clinical and Experimental Metastasis* 35.4, pp. 309–318. ISSN: 15737276. DOI: 10.1007/s10585-018-9903-0. URL: http://dx.doi.org/10.1007/s10585-018-9903-0.

Nowell, Peter C. (1976). “The clonal evolution of tumor cell populations”. In: *Science*. ISSN: 00368075. DOI: 10.1126/science.959840.

Öhrling, Katarina et al. (2010). “Mismatch repair protein expression is an independent prognostic factor in sporadic colorectal cancer”. In: *Acta Oncologica*. ISSN: 0284186X. DOI: 10.3109/02841861003705786.

Ohtsuki, H and H Innan (2017). “Allele Frequency Spectrum in a Cancer Cell Population”. In: XXX.January. DOI: 10.1534/genetics.XXX.XXXXXX.

Oren, Yaara et al. (2021). “Cycling cancer persister cells arise from lineages with distinct programs”. In: *Nature* June 2020. ISSN: 0028-0836. DOI: 10.1038/s41586-021-03796-6.

Orr, H. Allen and Jerry A. Coyne (1992). “The Genetics of Adaptation: A Reassessment”. In: *The American Naturalist* 140.5, pp. 725–742. ISSN: 0003-0147. DOI: 10.1086/285437.

Orsetti, Béatrice et al. (2014). “Impact of chromosomal instability on colorectal cancer progression and outcome”. In: *BMC Cancer* 14.1, pp. 1–13. ISSN: 14712407. DOI: 10.1186/1471-2407-14-121.

Osborne, C. Kent and Rachel Schiff (2011). “Mechanisms of Endocrine Resistance in Breast Cancer”. In: *Annual Review of Medicine*. ISSN: 0066-4219. DOI: 10.1146/annurev-med-070909-182917.

Panczyk, Mariusz (2014). *Pharmacogenetics research on chemotherapy resistance in colorectal cancer over the last 20 years.* DOI: 10.3748/wjg.v20.i29.9775.

Payne, Joshua L. and Andreas Wagner (2019). “The causes of evolvability and their evolution”. In: *Nature Reviews Genetics* 20.1, pp. 24–38. ISSN: 14710064. DOI: 10.1038/s41576-018-0069-z. URL: https://www.nature.com/articles/s41576-018-0069-z?utm_source=researcher_app&utm_medium=referral&utm_campaign=MKEF_USG_Researcher_inbound.

PCAWG Transcriptome Core Group et al. (2020). “Genomic basis for RNA alterations in cancer.” In: *Nature* 578.7793, pp. 129–136. ISSN: 1476-4687. DOI: 10.1038/s41586-020-1970-0. URL: http://www.ncbi.nlm.nih.gov/pubmed/32025019.

Popat, Sanjay et al. (2005). *Systematic review of microsatellite instability and colorectal cancer prognosis.* DOI: 10.1200/JCO.2005.01.086.

Porter, Shaina N. et al. (2014). “Lentiviral and targeted cellular barcoding reveals ongoing clonal dynamics of cell lines in vitro and in vivo”. In: *Genome biology* 15.5, R75. ISSN: 1474760X. DOI: 10.1186/gb-2014-15-5-r75.

Ragnhammar, P. et al. (2001). *A systematic overview of chemotherapy effects in colorectal cancer.* DOI: 10.1080/02841860151116367.

Rancati, Giulia et al. (2008). *Aneuploidy Underlies Rapid Adaptive Evolution of Yeast Cells Deprived of a Conserved Cytokinesis Motor.* DOI: 10.1016/j.cell.2008.09.039.

Rehman, Sumaiyah K. et al. (2021). “Colorectal Cancer Cells Enter a Diapause-like DTP State to Survive Chemotherapy”. In: *Cell* 184.1, pp. 226–242. ISSN: 10974172. DOI: 10.1016/j.cell.2020.11.018. URL: https://doi.org/10.1016/j.cell.2020.11.018.

Reimers, M. S. et al. (2013). “Biomarkers in precision therapy in colorectal cancer”. In: *Gastroenterology Report* 1.3, pp. 166–183. DOI: 10.1093/gastro/got022.

Ritz, Christian et al. (2015). “Dose-response analysis using R”. In: *PLoS ONE.* ISSN: 19326203. DOI: 10.1371/journal.pone.0146021.

Robey, Robert W. et al. (2018). *Revisiting the role of ABC transporters in multidrug-resistant cancer.* DOI: 10.1038/s41568-018-0005-8.

Roswell, Michael et al. (2020). “A conceptual guide to measuring species diversity”. In: *Oikos* 128.5, pp. 659–667. ISSN: 16000706. DOI: 10.1111/oik.07202.

Russo, Mariangela et al. (2021). “Drug-induced colorectal cancer persister cells show increased mutation rate”. In: *bioRxiv*, pp. 1–59. DOI: https://doi.org/10.1101/2021.05.17.444478.

Sansregret, Laurent et al. (2018). “Determinants and clinical implications of chromosomal instability in cancer”. In: *Nature Reviews Clinical Oncology* 15.3, pp. 139–150. ISSN: 17594782. DOI: 10.1038/nrclinonc.2017.198.

Schmidt, Manfred et al. (2007). “High-resolution insertion-site analysis by linear amplification-mediated PCR (LAM-PCR)”. In: *Nature Methods* 4.12, pp. 1051–1057. ISSN: 15487091. DOI: 10.1038/nmeth1103.

Schmitt, Michael W. et al. (June 2016). “The influence of subclonal resistance mutations on targeted cancer therapy”. In: *Nature Reviews Clinical Oncology* 13.6, pp. 335–347. ISSN: 1759-4774. DOI: 10.1038/nrclinonc.2015.175. URL: http://www.nature.com/articles/nrclinonc.2015.175.

Shaffer, Sydney M, Margaret C Dunagin, et al. (2017). “Rare cell variability and drug-induced reprogramming as a mode of cancer drug resistance”. In: *Nature* 546.7658, pp. 431–435. ISSN: 14764687. DOI: 10.1038/nature22794. URL: http://dx.doi.org/10.1038/nature22794.

Shaffer, Sydney M, Benjamin L Emert, et al. (2018). “Memory sequencing reveals heritable single cell gene expression programs associated with distinct cellular behaviors”. In: *bioRxiv*.

Shankaran, V. et al. (2010). “Predicting Response to EGFR Inhibitors in Metastatic Colorectal Cancer: Current Practice and Future Directions”. In: *The Oncologist*. ISSN: 1083-7159. DOI: 10.1634/theoncologist.2009-0221.

Sharma, Sreenath V. et al. (2010). “A Chromatin-Mediated Reversible Drug-Tolerant State in Cancer Cell Subpopulations”. In: *Cell* 141.1, pp. 69–80. ISSN: 00928674. DOI: 10.1016/j.cell.2010.02.027. URL: http://dx.doi.org/10.1016/j.cell.2010.02.027.

Shi, Hubing et al. (2014). “Acquired resistance and clonal evolution in melanoma during BRAF inhibitor therapy”. In: *Cancer Discovery* 4.1. ISSN: 21598274. DOI: 10.1158/2159-8290.CD-13-0642.

Siddik, Zahid H. (2003). *Cisplatin: Mode of cytotoxic action and molecular basis of resistance*. DOI: 10.1038/sj.onc.1206933.

Sottoriva, Andrea et al. (2015). “A big bang model of human colorectal tumor growth”. In: *Nature Genetics* 47.3, pp. 209–216. ISSN: 15461718. DOI: 10.1038/ng.3214. URL: http://dx.doi.org/10.1038/ng.3214.

Stagni, Camilla et al. (2018). “BRAF gene copy number and mutant allele frequency correlate with time to progression in metastatic melanoma patients treated with MAPK inhibitors”. In: *Molecular Cancer Therapeutics* 17.6, pp. 1332–1340. ISSN: 15388514. DOI: 10.1158/1535-7163.MCT-17-1124.

Strobl, Maximilian et al. (2020). “Turnover modulates the need for a cost of resistance in adaptive therapy”. In: *bioRxiv*. DOI: 10.1101/2020.01.22.914366.

Sunshine, Anna B. et al. (2015). “The Fitness Consequences of Aneuploidy Are Driven by Condition-Dependent Gene Effects”. In: *PLoS Biology* 13.5, pp. 1–34. ISSN: 15457885. DOI: 10.1371/journal.pbio.1002155.

Swanton, C. et al. (2009). “Chromosomal instability determines taxane response”. In: *Proceedings of the National Academy of Sciences* 106.21, pp. 8671–8676. ISSN: 0027-8424. DOI: 10.1073/pnas.0811835106. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.0811835106.

Szakács, Gergely et al. (2006). “Targeting multidrug resistance in cancer”. In: *Nature Reviews Drug Discovery* 5.3, pp. 219–234. ISSN: 14741776. DOI: 10.1038/nrd1984.

Tambe, Akshay and Lior Pachter (2019). “Barcode identification for single cell genomics”. In: pp. 1–9.

Tegze, Bálint et al. (2012). “Parallel evolution under chemotherapy pressure in 29 breast cancer cell lines results in dissimilar mechanisms of resistance”. In: *PLoS ONE*. ISSN: 19326203. DOI: 10.1371/journal.pone.0030804.

Thielecke, Lars et al. (2017). “Limitations and challenges of genetic barcode quantification”. In: *Scientific Reports* 7.August 2016, pp. 1–14. ISSN: 20452322. DOI: 10.1038/srep43249. URL: http://dx.doi.org/10.1038/srep43249.

Tomasello, Chiara et al. (2018). “Resistance to EGFR inhibitors in non-small cell lung cancer: Clinical management and future perspectives”. In: *Critical Reviews in On-*

cology/Hematology 123.January, pp. 149–161. ISSN: 18790461. DOI: 10.1016/j.critrevonc.2018.01.013. URL: https://doi.org/10.1016/j.critrevonc.2018.01.013.

Tomasetti, Cristian et al. (2013). "Why tyrosine kinase inhibitor resistance is common in advanced gastrointestinal stromal tumors". In: *F1000Research* 2. ISSN: 1759796X. DOI: 10.12688/f1000research.2-152.v1.

Turati, Virginia A. et al. (2021). "Chemotherapy induces canalization of cell state in childhood B-cell precursor acute lymphoblastic leukemia". In: *Nature Cancer* 2.8, pp. 835–852. ISSN: 26621347. DOI: 10.1038/s43018-021-00219-3.

Usanova, S. et al. (2011). *Cisplatin sensitivity of testis tumour cells is due to deficiency in interstrand-crosslink repair and low ercc1-xpf expression.* DOI: 10.1016/j.juro.2011.04.047.

Vander Velde, Robert et al. (2020). "Resistance to targeted therapies as a multifactorial, gradual adaptation to inhibitor specific selective pressures". In: *Nature Communications* 11.1. ISSN: 20411723. DOI: 10.1038/s41467-020-16212-w. URL: http://dx.doi.org/10.1038/s41467-020-16212-w.

Vargas-Rondón, Natalia et al. (2017). "The Role of Chromosomal Instability in Cancer and Therapeutic Responses". In: *Cancers* 10.1, p. 4. ISSN: 2072-6694. DOI: 10.3390/cancers10010004. URL: http://www.mdpi.com/2072-6694/10/1/4.

Viossat, Yannick and Robert Noble (2021). "A theoretical analysis of tumour containment". In: *Nature Ecology & Evolution.* ISSN: 2397334X. DOI: 10.1038/s41559-021-01428-w. URL: http://dx.doi.org/10.1038/s41559-021-01428-w.

Walther, A. et al. (2008). "Association between chromosomal instability and prognosis in colorectal cancer: A meta-analysis". In: *Gut* 57.7, pp. 941–950. ISSN: 00175749. DOI: 10.1136/gut.2007.135004.

Walther, Axel et al. (2009). "Genetic prognostic and predictive markers in colorectal cancer". In: *Nature Reviews Cancer* 9.7, pp. 489–499. ISSN: 14741768. DOI: 10.1038/nrc2645.

Wang, Qiang et al. (2018). "PIK3CA mutations confer resistance to first-line chemotherapy in colorectal cancer". In: *Cell Death and Disease* 9.7. ISSN: 20414889. DOI: 10.1038/s41419-018-0776-6.

Weissman, Tamily A. and Y. Albert Pan (2014). *Brainbow: New resources and emerging biological applications for multicolor genetic labeling and analysis.* DOI: 10.1534/genetics.114.172510.

West, J. et al. (2018). "Capitalizing on competition: An evolutionary model of competitive release in metastatic castration resistant prostate cancer treatment". In: *Journal of Theoretical Biology* 455, pp. 249–260. ISSN: 10958541. DOI: 10.1016/j.jtbi.2018.07.028.

Wheeler, JMD and WF Bodmer (2000). "DNA mismatch repair genes and colorectal cancer". In: *Gut* 47, pp. 148–153.

Wilkinson, Neal W. et al. (2010). "Long-term survival results of surgery alone versus surgery plus 5-fluorouracil and leucovorin for stage ii and stage iii colon cancer: Pooled analysis of NSABP C-01 through C-05. A baseline from which to compare modern adjuvant trials". In: *Annals of Surgical Oncology.* ISSN: 10689265. DOI: 10.1245/s10434-009-0881-y.

Williams, Marc J. et al. (2018). "Quantification of subclonal selection in cancer from bulk sequencing data". In: *Nature Genetics* 50.June, pp. 1–9. ISSN: 15461718. DOI: 10.1038/s41588-018-0128-6. URL: http://dx.doi.org/10.1038/s41588-018-0128-6.

Windels, Etthel Martha et al. (2019). "Bacterial persistence promotes the evolution of antibiotic resistance by increasing survival and mutation rates". In: *ISME Journal* 13.5, pp. 1239–1251. ISSN: 17517370. DOI: 10.1038/s41396-019-0344-9. URL: http://dx.doi.org/10.1038/s41396-019-0344-9.

Wolpin, Brian M and Robert J Mayer (2013). "Systemic treatment of colorectal cancer". In: *Gastroenterology* 134, pp. 1296–1310. ISSN: 03008142. DOI: 10.1053/j.gastro.2008.02.098.

Woodworth, Mollie B. et al. (2017). "Building a lineage from single cells: Genetic techniques for cell lineage tracking". In: *Nature Reviews Genetics* 18.4, pp. 230–244.

ISSN: 14710064. DOI: 10.1038/nrg.2016.159. URL: http://dx.doi.org/10.1038/nrg.2016.159.

Wright, Sewall (1932). "The roles of mutation, inbreeding, crossbreeding and selection in evolution." In: *Proc. 6th International Congress of Genetics* 1, pp. 356–366.

Yona, A. H. et al. (2012). "Chromosomal duplication is a transient evolutionary solution to stress". In: *Proceedings of the National Academy of Sciences* 109.51, pp. 21010–21015. ISSN: 0027-8424. DOI: 10.1073/pnas.1211150109. URL: http://www.pnas.org/cgi/doi/10.1073/pnas.1211150109.

Yun, Cai Hong et al. (2008). "The T790M mutation in EGFR kinase causes drug resistance by increasing the affinity for ATP". In: *Proceedings of the National Academy of Sciences of the United States of America* 105.6, pp. 2070–2075. ISSN: 00278424. DOI: 10.1073/pnas.0709662105.

Zhao, Boyang et al. (2016). "Exploiting Temporal Collateral Sensitivity in Tumor Clonal Evolution". In: *Cell* 165.1, pp. 234–246. ISSN: 10974172. DOI: 10.1016/j.cell.2016.01.045. URL: http://dx.doi.org/10.1016/j.cell.2016.01.045.

Zhao, Lu et al. (2018). "Bartender: A fast and accurate clustering algorithm to count barcode reads". In: *Bioinformatics* 34.5, pp. 739–747. ISSN: 14602059. DOI: 10.1093/bioinformatics/btx655.

Zorita, Eduard et al. (2015). "Starcode: Sequence clustering based on all-pairs search". In: *Bioinformatics* 31.12, pp. 1913–1919. ISSN: 14602059. DOI: 10.1093/bioinformatics/btv053.