Intra-tumor heterogeneity and evolution

by

Robert Austin Mathis

B.S. Biology
Haverford College, 2010

SUBMITTED TO THE DEPARTMENT OF BIOLOGY
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

SEPTEMBER 2017

Signature of Author ……

Robert Austin Mathis
Department of Biology
September 4, 2017

Certified by ……

Piyush B. Gupta
Professor of Biology
Thesis Supervisor

Accepted by ……

Stephen P. Bell
Professor of Biology
Co-Chair, Biology Graduate Committee

Intra-tumor heterogeneity and evolution

By

Robert Austin Mathis

Submitted to the Department of Biology on September 4, 2017
in partial fulfillment of the requirements for the Degree of Doctor of Philosophy

Abstract

Although the treatment of cancer is a major focus of biomedical research, many cancers are extremely hard to treat. Tumors likely resist treatment because each tumor is heterogeneous, and can evolve. Although tumor evolution has long been appreciated, it remains incompletely understood. In this thesis, I will explore two questions related to cancer heterogeneity and evolution: how evolution can affect plastic phenotypes, and the role of purifying selection in cancer evolution. Different cell states or phenotypes have been observed within tumors, and they are associated with treatment resistance and metastasis. The observation that these phenotypes are plastic leads to a conundrum: how can selection act on such an unstable phenotype? We determined that plasticity, in the form of cell state bias, varies widely across clones in a tumor. These different biases are heritable, with each cell faithfully passing its differentiation bias to its daughters. Simulations revealed that this makes plasticity an evolvable phenotype--- in a changing environment, an optimal state bias will be selected. The second question explored in this thesis is the role of purifying selection in cancer evolution. It is widely thought that tumor evolution is dominated by positive selection. We posited that, as in the evolution of species, purifying selection would prevent the fixation of deleterious mutations in tumors. Through computational analysis of tumor genomes, we determined that purifying selection acts to remove deleterious mutations. Genes under purifying selection must be important to tumors *in vivo*, as only mutations in these genes would be problematic. Consistent with this prediction, most genes under purifying selection in tumors were essential in cancer cell lines. To find genes essential to tumors but not generally cell-essential, we developed a method to find genes under increased purifying selection in one tumor type over others. This revealed a number of pathways under selection in melanomas, but not other tumor types, such as DNA damage pathways. By seeking genes important to tumors, but not generally essential, our analysis revealed potential therapeutic targets. Purifying selection offers an unprecedented view into which genes are essential to tumors *in vivo*, a finding predominantly inaccessible through experimentation.

Thesis Supervisor: Piyush B. Gupta
Title: Professor of Biology

Table of Contents

Acknowledgements

My successful PhD cannot be solely attributed to myself. I am greatly indebted to many individuals who provided me with assistance during this process. I must, of course, thank my many lab mates, who not only provided practical assistance, advice, and guidance, but also successfully created a positive and supportive working environment. Ethan Sokol, in particular, was an able and helpful participant in almost every project I conducted. Here, too, I should thank my advisor, Piyush Gupta, who successfully advised me during my PhD, and who recruited a fantastic group of people. My advisor and lab mates were also truly crucial during the seemingly inevitable moment of a graduate student's doubts. My committee, Peter Reddien and Mike Hemann, were also very helpful in steering both my projects and my career, through the entirety of my graduate tenure. I can honestly say that I owe much of my academic success to them. I should here thank those who, during their time at Merrimack pharmaceutical company, took the time to provide me with a productive internship experience, in particular Shawn Carrey and Jeff Kearns.

I could not fail to acknowledge those who assisted me in a variety of ways outside of the laboratory, providing practical and emotional support. My wife, Lara, was ever patient for my long stretches of time at lab, and helped provide me with confidence and direction. My family, similarly, encouraged me throughout my time here. My fellow graduate students and friends, colloquially known as the Biomansion, have been fantastic companions in the graduate student adventure. I truly believe that all of these individuals fantastically improved my graduate experience, and our adventures created many memories that I treasure greatly. I should here also thank those with whom I sang, and bicycled, sometimes simultaneously. These activities brought to me much peace and happiness during my studies, which I value quite greatly.

Chapter 1: Introduction

Chapter 1

1.1 Introduction

As cancer is the second leading cause of death in the United States [1], improving our understanding of cancer to improve the treatment of cancer patients is a major focus of biomedical research. However, significant challenges still lie before the field. For example, even with the much-heralded advancement of precision medicine, where treatments are matched to tumor genotypes, we find it difficult to predict if tumors will respond to targeted-therapy, and they rapidly become resistant to therapy. Many of these challenges relate to the complexity within tumors. Rather than being a discrete entity, it has become clear that each tumor consists of a number of distinct genetic clones. This genetic diversity is a suitable substrate for evolution, which can drive resistance and aggressiveness. A more thorough understanding of cancer heterogeneity and evolution will help illuminate paths towards the successful treatment of cancer.

Our modern understanding of cancer is based on a genetic model. First laid out by Nordling [2], and later refined by Knudson [3], this model suggests cancer is the result of mutations activating or breaking genes to increase proliferation. Multiple mutations are required for the development of cancer, often in multiple genes. These mutations do not occur simultaneously; the acquisition of multiple mutations under positive selection during tumor development is thought drive tumor evolution. In this model, successive clones are driven to take over the population of cells in the tumor through positive selection [4].

Recently, increased complexity has been added to this model of tumor evolution, stemming from the observation that individual tumors display marked genetic heterogeneity, in contrast to the homogeneity that would result from clonal dominance. Such observations, stemming from the analysis of single cells [5-7], distinct spatial regions [8, 9], and sorted subpopulations [10], all show that tumors

consist of many different clones of distinct genotypes. These clones can differ through the gains and losses of entire chromosomes (aneuploidies), copy-number changes of smaller regions, or through single point mutations.

In addition to the observation that tumors display genetic heterogeneity, orthogonal analyses have found heterogeneity in clonal behaviors. These analyses, mostly based off of tracking clones with induced genetic markers, have revealed extensive behavioral differences in clones' growth [11-13] and response to therapy [14-16]. It seems likely that these clonal differences in behavior are driven by the observed genetic heterogeneity. Genetic differences are passed on to cells' progeny, and the connection between genetic differences and different phenotypes is the backbone of our understanding of evolution.

Heterogeneity within tumors has been observed on yet another axis: cell state. In single cell RNA sequencing experiments, and immunohistochemistry, cells with markers of different states have been observed in breast and other cancers [10, 17-20]. In analogy to normal tissues, it seems that cancers are mixtures of cells in different states.

This thesis will attempt to improve our understanding of human tumors through detailed answers of two questions related to cancer heterogeneity and evolution.

First, which genes are important to tumors, *in vivo*? There is significant interest in understanding which genes are important to tumors (required for growth, for example), as their dependence on these genes could underlie novel treatments [21]. Unfortunately, many of the powerful tools used in cancer research, including cell lines and mouse models, cannot be used to find all of the genes that are truly important to human tumors, *in vivo*. Cell lines, cultured short- or long-term in two or three dimensions, are necessarily subject to very different environmental conditions from those *in vivo*. The murine immune system rejects any human cells implanted, so using a mouse xenograft

model of cancers requires knocking out the mouse's immune system; the human cells are also often unable to interact with mouse host factors, bringing them further away from their normal context [22]. The large costs and difficulties in testing patient-derived xenografts also makes them difficult to use to answer this question. While the strength of mouse models of cancer are not in doubt, where genetic lesions are introduced into particular cells to induce tumorigenesis, we cannot be certain they reflect human biology. Differences in tumor species, size, age, and possibly cell of origin all complicate these comparisons. Lastly, and obviously, screens for gene functionality in tumors cannot ethically be done in humans.

In order to get around these limitations, and possibly find genes important for tumor growth or maintenance, many resources have been expended to sequence the exons of human tumors [23]. Most of these efforts involve seeking recurrently-mutated genes, activated through mutation [21, 24]. In this case, as only a small number of missense mutations can lead to a novel, activated allele, a recurrent (repeated) mutation across different tumors leads to evidence of positive selection.

However, this strategy may not find all genes important to tumors. Differences in the rate of mutation across different regions of the genome can make it difficult to call whether any part of the genome has been recurrently mutated [24]. The heterogeneity across tumors, where tumors depend on different pathways or processes, has also made it more difficult to find evidence of important genes based on recurrent mutation. Indeed, many genes that are important in tumors may not be activated through mutation at all. In other words, there are likely many genes important to tumors that are not necessarily under positive selection, such as any cell-essential genes.

The second question asks the relationship between plastic phenotypes and cancer evolution. We know that cancer evolution underlies much of the difficulties of cancer treatment. We know that the expansion of some pre-existing clones, sometimes known to be driven by mutations, can cause

resistance to treatment [14, 16, 25, 26]. However, there is also evidence that resistance to therapy can be driven by the expansion of resistant cell states [15, 27-33]. These cell states are known to be plastic (here meaning changeable); tumor cells can differentiate between states, in a manner analogous to the differentiation of normal cells [34-37]. Additional evidence of plasticity *in vivo* comes from single cell RNA sequencing [19]. Plastic phenotypes, in contrast to stable genetic lesions, are not so clearly related to evolution. How can a clone be enriched based on the selection of a plastic phenotype? One analogy comes in the form of the evolution of bet hedging in other systems. In these systems, occurring in species as diverse as yeast and bacteria [38, 39], a fraction of each clone lies in a resistant, often slower-growing state, such as a spore. In this way, if the environment changes, killing the sensitive cells, each clone with a fraction of resistant cells will survive. If clones had different probabilities of creating resistant cells, the frequency of the selective environment would select for the optimal probability.

Understanding how differences in cell state representation can evolve in tumors is critical for understanding treatment resistance. The selection of a plastic phenotype would require clones with diverse plasticity, encoded in a stable, heritable matter.

1.2 Thesis Outline

In the second chapter, I describe our identification of purifying selection in human cancers, uncovered through the analysis of published exome sequencing data from human tumors [23]. Purifying selection is the evolutionary force that removes mutations that reduce the fitness of an organism, preventing the fixation of deleterious alleles [40]. Therefore, to find evidence of purifying selection, we had to look for mutations that disappeared, as they were removed from the tumor genome. This required calculating how many mutations we expected to find. We were able to control for differences in the rate of mutation across genes using whole genome sequencing of human tumors. We looked at the accumulation of mutations in non-coding regions of tumor genomes adjacent to genes, such as untranslated regions (UTRs) and introns. Such regions, compared to exons, have fewer functional

regions, letting us estimate the relative mutation rate of each gene, and determine how many mutations we would expect.

Using this methodology, we found strong evidence of purifying selection in human cancers. We saw an over-representation of conservative amino acid transitions [41], suggesting that less-conservative transitions were depleted by purifying selection. In an orthogonal analysis, we saw a depletion of mutations in expressed genes, as compared to silent genes, beyond that which is attributable to transcription-coupled repair.

Identifying genes under purifying selection in multiple cancer types, we determined that these genes tended to be essential in cancer cell lines, as identified through pooled CRISPR screens [42-44]. Purifying selection successfully predicted gene essentiality, consistent with its role in removing mutations from genes with important functions.

Although genes under purifying selection in cancer would be important for tumor growth or maintenance, we reasoned that these genes would not necessarily be good targets for treatment, as they were generally important for cell survival. We were able to develop a method to get around this limitation, by looking for genes under increased selection in one tumor type, as compared to other tumor types. This identified a variety of processes and pathways under increased selection in melanomas and in lung adenocarcinomas, as compared to other tumor types.

In the third chapter, I describe our identification of clonal diversity in phenotypic plasticity. After uncovering plastic cell state phenotypes in cancer, the question emerged whether such phenotypes could be under selection. As discussed above, although plasticity induces state heterogeneity in a population of cells, this very plasticity makes it unclear if there is a stable phenotype to be selected for or against.

With the additional knowledge that cells within individual tumors are extremely diverse on the genetic level, we hypothesized that plasticity, itself, could be a clonally-variable phenotype. By tracking the lineages of single cells using introduced DNA barcodes, we were able to determine that clones indeed displayed distinct cell state-proportions at equilibrium. Through single-cell cloning, we observed that each clone's cell state proportion was extremely stable over time, and immensely heritable (with a narrow-sense heritability of 0.89). After uncovering this phenomenon in a breast cancer cell line, we found further evidence for its existence *in vivo* using a single-cell RNA-seq data set from a human glioblastoma.

This clonally-inheritable difference in plasticity dramatically changed our predictions of how a population would respond to selection pressures. Using the data we collected from detailed clonal tracking of a breast cancer cell line, including each clone's growth rate and phenotypic equilibrium, we simulated the effect of many kinds of selection on this population. This revealed that clonal diversity in plasticity lets a population evolve a bet-hedging strategy in response to a varying (over time) selection. If such a selection targets cell states, the clone with an equilibrium cell state ratio best resisting the variable selection will become enriched.

1.3 Cancer genetics and evolution

1.3.1 The genetic model of cancer

Our current understanding of cancer is based on a genetic model, where genetic aberrations are responsible for changing a normal cell into a carcinogenic one. In this model, as laid out by Nordling, mutations breaking pathways or genes repressing growth, along with mutations activating genes to increase growth, force the expansion of a clone of cells that eventually becomes a tumor [2]. As uncovered in a thoughtful analysis by Knudson of patients presenting with mono- or bi-ocular

retinoblastomas, multiple mutations tend to be required for tumorigenesis [3]. This is due to the presence of tumor-repressive mechanisms in cells, which often must be inactivated biallelically.

Evidence for the genetic model of cancer comes from many vectors. First, while tumors tend to carry many mutations, making it difficult to identify the phenotype of any individual mutation, there is strong evidence of repeated mutational activation of the same genes in many distinct tumors [21]. This is evidence of strong positive selection for the same types of aberrations in many cancers. We would not expect mutations with no phenotype to be repeatedly dragged to enrichment (and therefore visible), despite their linkage to alleles under positive selection. Many of these recurrent mutant alleles are sufficient to induce cancer when introduced in mice [45] and *in vitro [46].*

Unsurprisingly, this method of increasing the growth of cells has been coopted by viruses. Some viruses are sufficient to cause oncogenesis by themselves; many such viruses have been observed to carry modified cellular genes (e. g. *v-myc*, *v-src*) [47]. In many cases, these same genes are found to be repeatedly mutated in cancers. This further suggests that the introduction or creation of certain alleles is sufficient to cause cancer.

Certain cancers are much more likely to happen in human populations with certain alleles. This heritable cancer predisposition suggests a role of specific alleles in carcinogenesis. For example, children born with one non-functional *Rb* allele are much more prone to bilateral retinoblastoma; while the same allele is found in the tumors of children with retinoblastoma in one eye, their probability of developing a tumor in their other eye is greatly reduced [3, 48]. This suggests the necessity of two *Rb* mutations for the genesis of retinoblastoma. Similarly, patients inheriting an inactivated *APC* allele have a dramatically increased risk of colorectal cancer [49]. Other alleles which increase the rate of acquiring mutations dramatically increase the probability of developing certain cancers. For example, alleles inhibiting nucleotide excision repair resulting in Xenoderma Pigmentosum; patients with these alleles

are extremely sensitive to ultraviolet light, and almost all develop skin cancers [50]. This suggest a role in mutation for the development of cancers, where increasing the rate of mutations increases the probability of an oncogenic mutation occurring.

These data, when combined, make a compelling argument for the genetic basis of cancer. Certain alleles appear to be necessary and sufficient for oncogenesis; these alleles are represented in populations with an increased probability of generating cancers, and are found in viruses capable of inducing cancers.

1.3.2 Introduction to cancer evolution

Cancer evolution refers here to the evolution of an individual tumor during its lifetime. The basic requirements for evolution are heterogeneity (differences between cells, here) and heritability. Mutations, which are readily transmitted to a cell's progeny, and occur in single cells (creating heterogeneity), are therefore readily able to be under selection and result in evolution. The expansion of a clone following the acquisition of a particular mutation is an example of positive selection. This is the idea that a mutation results in an increase in the rate of growth or survival for the cell which contains it. The increased growth or survival leads to the expansion of that lineage (Figure 1).

The positive selection of cell lineages is thought to underlie tumor development. As discussed above, cancers generally require greater than one mutation, usually affecting more than one gene. Even in the most mutagenic contexts, the rate of mutation is low enough that it is highly unlikely for these mutations to occur simultaneously. This led to the idea of cancers arising through a series of clonal successions, as advanced by Nowell [4]. In this model, a mutation results in the outgrowth of a clone due to faster growth, or increased survival. Another mutation in one cell within this lineage that further increases that cell's growth then creates another sub clone, which then, again, takes over the tumor. After multiple rounds of clonal successions, a tumor is created. One well studied model of this sort of

successive clonal evolution comes from colon cancer [51], where tumors appear to move step-wise through mutating a series of proteins, including the *APC* gene, mentioned above. An individual mutation will increase one stem cell clone's proliferation rate, increasing the probability that it will take over adjacent crypts, resulting in clonal expansion [52].

This sort of cancer evolution through positive selection seems to occur over long stretches of time. One piece of evidence comes from the observation of repeated mutations. Very specific recurrent mutations (such as *KRAS* G12D) can be repeatedly found in many different tumors [21]. Even though in many cases it can be observed that these regions are somewhat hyper mutable, the specificity of these mutations under positive selection suggests that a great many mutations occur in tumors (or pre-tumors). A cell would have to be extremely lucky to mutate that particular base (or the organism unlucky), suggesting that achieving these multiple rounds of succession may take quite a long time. Another line of evidence comes from pre-cancerous tissues. Many of these specific oncogenic lesions have been observed in normal tissues [53]. Expanding clones containing particular lesions can be found in organs such as the skin, although many will never progress to cancer. These snapshots of the early steps of cancer evolution suggest that tumor development could take an extremely long time.

Positive selection in cancer occurs predominantly in two contexts: oncogenes and tumor suppressors. Tumor suppressors are genes which function to repress tumors, and their inactivation is often required for tumorigenesis. Such inactivation can occur through mutation, silencing, or deletion. Tumor suppressors function through many mechanisms, including repressing growth signals (such as *PTEN*) or linking DNA damage to apoptosis (such as *TP53*) [54]. As these suppress tumor growth, the presence of one functional allele is often sufficient for suppression, requiring the inactivation of both alleles for tumorigenesis, classically modeled by *Rb* and *APC* [3, 51]. In contrast, oncogenes are genes whose activation increases tumor growth. This includes many genes where an aberrant increase in their expression results in an increase in proliferation (such as *MYC, ABL)*, where the protein has no novel

functionality. Other oncogenic alleles produce novel functionality, such as auto-activating growth factor receptors like EGFR or *BRAF*. In contrast to tumor suppressors, where the inactivation of both alleles is often necessary for tumorigenesis, in many cases the activation of a single oncogenic allele is sufficient to provide the tumorigenic function.

Evolution depends on heterogeneity, and tumors have been observed to be extremely heterogeneous on the genetic level. Analysis of tumor genomes have revealed many variants, including mutations, aneuploidies (gains and losses of entire chromosomes or chromosome arms), and copy number changes (smaller-scale gains and losses of chromosomal regions), that are variable across cells within a single tumor. Much of this has been revealed from single cell sequencing, or the sequencing of different regions of tumors [6-9, 19].

Besides its role in tumorigenesis, tumor evolution has a known role in resisting drug treatment. The selection for likely pre-existing clones with alleles conferring resistance to targeted therapy has been observed. Examples include the expansion of clones with EGFR mutations conferring resistance to targeted EGFR inhibition [26], or mutations conferring resistance to inhibition of the *BCR-ABL* oncogene [55]. The expansion of pre-existing clones resistant to EGFR inhibition *in vitro* suggested that, in many cases, the resistant alleles are already present before treatment begins [16].

1.3.3 Purifying selection in cancer

As shown in the previous discussion of cancer evolution, the study of cancer evolution is dominated by a study of positive selection. An equally strong force in the evolution of organisms is purifying selection, which is relatively understudied in tumor evolution.

Purifying selection acts to remove alleles that decrease the fitness of an organism or cell, often through reducing its proliferation or survival. The neutral theory of evolution, advanced by Kimura [40], suggests, among other propositions, that most of the observed differences between genes across

species (substitutions) have little to know phenotype. They are fixed in the population through random drift. This is because any problematic alleles, as they decrease an organism's fitness, would be kept from fixation in the population. This selection then prevents the accumulation of substitutions, and is therefore called purifying selection (Figure 2). Importantly, Kimura also proposed that this means more substitutions would occur in less-functional parts of molecules. This was in contrast to the contemporary (neo-Darwinist) view that most substitutions were fixed in a population through positive selection, so that more substitutions would occur in more-functional parts of molecules under positive selection. Evidence for purifying selection in humans comes, then, from determining if sequences are conserved in comparison to other organisms, suggesting a role of purifying selection in keeping them from changing. For example, genes that are essential in human cell lines, as determined through CRISPR knockout screens, show increased conservation across species [44].

Although most of the literature examining cancer evolution is focused on positive selection, there are a few examples of attempts to look for purifying selection. One group has pointed out, in contrast to the prediction of the neutral theory, many mutations appearing to be problematic are observed in tumors. They are thought to have been dragged to enrichment through linkage to positively-selected mutations [56]. This paper suggests that purifying selection is weak, relative to the strength of positive selection, although the entirely computational analysis depends on our ability to predict how problematic a mutation will be.

In contrast, the Hershberg group claimed to find some evidence of purifying selection in cancer, in the process of searching for evidence of positive selection [57]. This approach entailed analysis of dN/dS (the relative amount of missense vs synonymous mutations) to look for evidence of positive selection. Advanced by Kimura [40], this idea rests on the idea that synonymous mutations, which, due to the degeneracy of the codon code, encode the same amino acid, are likely to have a less-deleterious effect than missense mutations, which encode a different amino acid. While looking for genes with

more missense mutations than expected, to find genes under positive selection, the Hershberg group also found a few genes that had fewer missense mutations than expected. However, the analysis of dN/dS is fraught with difficulties due to the evidence of both positive and purifying selection on synonymous mutations [58], and an insufficient number of mutations to reach significance. Synonymous mutations have been known to have phenotypes, controlling gene expression related to mRNA stability, translation speed, and the probability of protein mis-folding [58-62]. This is reflected in the role of synonymous mutations driving the human disease cystic fibrosis, occurring in *CTFR*. In addition to the functionality of synonymous mutations, the dN/dS methods developed for comparing genes across species is thought to not be appropriate for analyzing samples within a population, such as tumors [63]. These difficulties are likely reflected in the errors in the Hershberg group's results. For example, they identified the gene *TTN* as showing very strong evidence of purifying selection; this gene is not even expressed in breast cancers (the cancers examined) nor any other non-muscular tissues, and was likely misidentified due to the fact it is the largest protein coding gene.

A last analysis actually sought to find genes under purifying selection across tumors, focusing on regions of hemizygocity (where only one copy of each gene is present) in tumors. They found strong evidence of purifying selection in a reduction of the number of estimated problematic mutations in the RNA polymerase II gene *POL2RA*, in those tumors where one copy had been deleted [64]. This is a gene where one copy is frequently lost in tumors, as it is next to the tumor suppressor *TP53* [65]. However, the group was unable to very successfully expand this analysis out to many other genes. As they required a large number of tumors to have lost one copy of a gene, only a small fraction of expressed genes (1,187) could be evaluated with their method. Of these, only 24 were identified as under purifying selection at a false discovery rate of 0.5, expecting half of those to be false positives.

The incentive for finding evidence of purifying selection in cancer stems from its ability to identify genes that are important in cancers, *in vivo*, a difficult place to find such information. Purifying

selection would only affect genes whose inhibition causes a reduction in fitness, thereby identifying genes important to tumor growth or maintenance (Figure 3). Such genes may turn out to be good targets for treatment, if they are not generally important. The ability to identify the tumor-maintenance functionality of genes *in vivo* is quite important. Although genome-wide CRISPR screens have recently been undertaken [42-44], increasing dramatically our ability to analyze which genes are actually essential in human cells, such experiments are not able to be carried out within human tumors. Unfortunately, our models for cancer biology (cell lines and mouse models) often fail recapitulate the tumor environment, including the presence of other, human cell types, which can change the essentiality of genes.

The challenge in finding purifying selection in cancer stems from the fact that the signal for this selection comes from missing mutations, requiring knowledge of how many mutations to expect (Figure 4). This is rendered more difficult by the observation that the rate of mutation varies across the genome, as discovered in the whole-genome sequencing of many cancers. Regional differences in mutation rate have been found to correlate with a number of factors, including DNA replication timing (when during the cell cycle a DNA region is copied) and chromatin accessibility [24, 66]. It should be noted that this was intuited from the observed mutations, so this analysis includes the probability of mutations occurring, and the probability that those mutations are immediately repaired, such that we never see them. The estimation is also not likely to include much evidence of purifying selection, since the estimate of mutation rates did not include any mutations from exons. Instead, the gene-by-gene estimations are mostly from introns, as well as untranslated regions. Although it has recently been confirmed that some non-coding mutations in the genome, related to controlling gene expression, can be under positive selection in cancer [67], this is not likely to impact the background mutation rate significantly. Additionally, it is well known that are important non-coding elements in untranslated regions can regulate gene expression [68] as well as protein localization [69], which may be under

purifying selection. Most of the signal comes from introns, however, and although introns include some functional elements that may be under purifying selection such as splice sites [70], this functional sequence represents a minority of the sequence. Therefore, although purifying selection may affect some of these sequences we used to estimate background mutation rate, we estimated that our comparator, exons, would display enough of an increase in purifying selection that we would be able to observe this signal.

An additional challenge for finding purifying selection is the number of mutations. Once we have calculated the number of mutations we would expect in a given gene, we must compare this to the number of mutations we observe. In order to find a significant depletion of observed mutations compared to the number expected (not conservatively, p value <0.01), with significance calculated by the cumulative distribution of the Poisson function, a relatively large number of mutations is required. This depends on the expected depletion: with a 50% depletion of mutations, one would have to observe 11 mutations (and expect 22); with a 25% depletion of mutations, one would have to observe 66 mutations (and expect 88). With the plan to test many genes for evidence of purifying selection, we would have to expect a fairly large depletion, and examine large numbers of mutations.

In order to observe enough mutations to observe a depletion, this analysis must be based on missense mutations--- mutations that change the coded amino acid. Such mutations are most common type due to the structure of codon amino acid coding. These mutations are widespread in cancers, and are believed to come from a large number of different sources [71]. Sources of mutation include increased proliferation (accompanied by an insufficient amount of nucleotides), activation of anti-viral mutation-inducing proteins (such as APOBEC enzymes [72]), and a variety of environmental exposures. Some known sources of environment-induced mutation includes ultraviolet light and tobacco exposure. However, analysis of mutations with dimensional reduction suggests that there are many sources of mutation of which we are currently unaware [71].

As discussed above, the feasibility of this study depends on a depletion of mutations from purifying selection. What depletion of mutations are we to expect? Assuming any missense mutation causing a reduced protein function would be under purifying selection, this would be related to the probability that a missense mutation would be problematic. An estimate for this comes from recent *in vitro* high-throughput mutagenesis studies, examining the proportion of missense mutations inhibiting enzymes. These studies revealed that proteins vary widely in their sensitivity to missense mutations. An *in vitro* mutagenesis of M. HaeIII (from *Haemophilus aegyptius*) showed that 66% of missense mutations were extremely deleterious to the enzyme's functionality, while 83% of missense mutations showed at least some effect) (Figure 5) [73]. This was an extremely sensitive assay of the enzyme's activity (methyltransferase). A different experiment examined the effects of missense mutations on the human DNA repair enzyme 3-methyladenine DNA glycosylase (AAG). Expressed in *Escherichia Coli,* approximately 34% of the missense mutations reduced the glycosylase function [74]. While varying across proteins, these data suggested that we could expect a reasonably high depletion of mutations.

It is worthwhile exploring our reasoning for why we may expect to observe purifying selection in cancer genome, starting with the presence of a second genome in each cell. First, for many genes the presence of one inactivated copy of a gene (in a diploid context) is sufficient to cause a phenotype; such genes are called haploinsufficient. Examination of nonsense mutations in human genomes has suggested that many genes are under purifying selection when one allele is rendered non-functional [75]. Additionally, despite the lack of sex-based recombination, there are many events in cancers that can cause the loss of a normal allele. The normal allele may be directly lost through aneuploidy (chromosome loss) [64]. Aneuploidy can be created through chromosome mis-segregation during mitosis, or DNA damage during replication [76, 77]. The outcome of mis-segregation can also result in copy-number neutral loss of heterozygosity: the loss of the normal allele, and duplication of the mutant allele [78, 79]. Lastly, approximately 10% of human genes are mono-allelically expressed, where only

one allele is expressed due to cis-regulation [80]. In this case, the normal allele can in no way compensate for the mutation. Through these ways and others, the normal allele is either rendered inactive, eliminated, or insufficient.

1.4 Cell state plasticity and evolution

1.4.1 Can cell state plasticity in cancer evolve?

This thesis will also explore cancer evolution from another angle: cell states. Can plasticity evolve in cancer? Evolution depends on the existence of heterogeneity. Selection can only act if individuals in a population (here, cells) are different. However, it is not clear if all types of heterogeneity in cancer can be under selection, due to the emergence in the literature of plasticity--- phenotypes not stably encoded by a cell.

1.4.2 Cell states and differentiation

Cell states are also known as differentiation states. A state can be modeled as an attractor creating a characteristic gene expression profile [81, 82]. This specific gene expression profile can be thought of as a point in high-dimensional gene expression space, where each dimension in this space is the expression of a particular gene. The topography of this space is governed by the underlying gene regulatory network, created by the effects of gene products on the expression of other gene products. Therefore, the network depends on which transcription factors are expressed, which cis-regulatory regions are accessible for factor binding, the chromatin and methylation landscape, and possibly many other factors. The cell's gene expression profile is a point in this gene expression space, lying somewhere on the topology created by the underlying gene regulatory network. The cell can move along this surface in gene expression space, which is the result of changing gene expression. A cell state is a stable point in that gene expression space, a basin of stability, also known as an attractor. Here, the underlying gene regulatory network makes a particular region of gene expression space stable.

The differentiation of cells--- where they move between states--- can also be modeled in this way. It is clear from embryonic biology and lineage tracking that cells can only do limited differentiation: the vast majority of possible cell state changes never occur (e. g. monocytes to neurons). Starting form an attractor, the underlying gene regulatory network, again, dictates the possible paths along which the cell's transcriptional profile can change. This limits the possible transitions a from a starting cell state. Changes in the gene regulatory profile, which push a cell's transcriptional profile, are thought to underlie differentiation. A differentiation signal changes the gene-regulatory network topology, resulting in destabilization of the attractor (state) where the cells began. The paths available then dictate the possible transitions, which can be biased based on this underlying gene regulatory network [81, 83]. Differentiation, then, is thought to be the result of cells moving again to a stable attractor state, after the destabilization of their starting position.

1.4.3 Cell states in breast cancer: the epithelial mesenchymal transition (and reverse)

The epithelial cells of breast cancer have been observed to be quite heterogeneous in regards to cell state [17, 27]. This mainly takes the form of the basic epithelial cells of breast cancer, and cells that appear to have undergone the epithelial-mesenchymal transformation (EMT). The EMT has been found and characterized in a variety of developmental processes, including mesoderm and neural crest formation in the embryo. The reverse process, mesenchymal-epithelial transition (MET), has been observed in the developing kidney [84]. EMT has also been associated with the stem cell potential of normal mammary epithelial cells, and the reverse process (MET) with differentiation [85]. In breast cancers, EMT'd (more-mesenchymal) cells are associated with treatment resistance and metastasis [33, 86, 87]. They have been found to be enriched in residual tumors post-therapy [33], and possibly for those breast cancer cells found in the blood [88].

More-mesenchymal and more-epithelial cells in breast cancer appear to be plastic; cells have been observed to transition between these phenotypes, in both directions. Most of these data come from analysis following sorted populations from cancer cell lines [35-37]. In contrast to the hierarchical stem-cell model thought to operate in many tissues and some cancers [89, 90], this plasticity seems to be bidirectional. This plasticity can be modeled as a Markov model, where the probability of transitioning depends on the cell's state [36]. In all systems set up in this way, the population will eventually come to an equilibrium, based on the relative probabilities of transition.

The plasticity of EMT has been found in some contexts to be related to regulation of the ZEB1 transcription factor, which in some normal and cancer cell lines is poised to be expressed [91]. This transcription factor, known to be involved in EMT, may serve to cause the observed plasticity through a bidirectional regulatory loop with the *MIR200* family of microRNAs.

1.4.4 Evolution acting on cell state plasticity

Could plasticity be another substrate for evolution? The very fact that plasticity involves cells transitioning between states--- an unstable phenotype--- makes it seem like a phenotype hard to be under selection. However, the fact that these stochastic transitions form an equilibrium allows for selection based on a phenotypic equilibrium. This assumes that the phenotypic equilibrium varies across different clones in a population, and that these differences are stably inherited.

There is robust evidence that cells can encode different probabilities of differentiation, resulting in a different equilibrium, within many tissue types. Much of this evidence comes from the hematopoietic system. An analysis of clones in an immortalized progenitor line showed that distinct hematopoietic progenitors had different differentiation probabilities, distinguishable based on *SCA* expression [92]. This analysis showed that these differences in differentiation probabilities were not

stable over long time periods, as sorted populations gradually reverted to reflect the parental distribution. However, this does suggest that cells can encode mechanisms to cause different differentiation probabilities. Similar differences in long term biases of hematopoietic lineage commitment have been found between distinct stem cells *in vivo*, based on tracking clones after reconstitution of the murine hematopoietic system [90, 93-96].

In addition to differences in differentiation bias across cells, in order for evolution to select for differentiation bias, it would have to be stably encoded. There is some evidence in the literature that clones can display stable differences in phenotypic equilibrium--- that is, that they are stably biased towards one state or another. In lung cancers, a very high-throughput analysis found clones that were repeatedly selected due to resistance to EGFR inhibition [16]. Interestingly, these clones displayed a more-mesenchymal phenotype compared to the general population from which they were derived. This suggests they may contain a clonally-heritable disposition towards mesenchymal states. A similar piece of evidence comes from cells from a breast tumor sorted by cell states [97]. This group sorted breast tumor cells into more-epithelial and more-mesenchymal based on the expressed of the surface markers CD24 and CD44. By looking for copy number differences between these sorted cells, they found genetic differences (copy number changes) between these sorted cells. Combined with the observation of plasticity, this suggested that some clones were predisposed towards one phenotype; it could also mean that there were certain clones that were not plastic in this context, and accumulated genetic differences. In contrast to this result, a similar analysis of cells sorted from a different breast tumor, but using high-throughput sequencing looking for mutations instead of copy number changes, did not find any differences [18]. Such a result is consistent with plasticity, and may reflect an inability to find small sub-clones.

1.4.5 Bet hedging mechanisms

An example of evolution acting on plasticity can be observed in a number of systems, based on the evolution of bet hedging. In melanomas, for example, there has long been evidence of so-called persisters, a cells state with increased therapy resistance [34]. Cells plastically turn into persisters at a low probability, repopulating a tumor after treatment is withdrawn. This is an example of a bet hedging mechanism.

Bet hedging mechanisms are found in a variety of organisms subject to variable environmental conditions. An organism or clone lacks the ability to predict when a state with increased resistance to a certain stress will be required, as the environment is constantly variable. For example, a colony of yeast has no way of knowing when its food source will run out. A solution to this, found in organisms as diverse as bacteria [98], viruses [99], and cancers [34], is to have some fraction of each clone produce a state that is resilient to a source of stress, even if that stress is not currently present. This means that each clone has some probability of creating cells (or viruses) in the resistant state, some cells in the non-resistant state. This allows each clone to survive a sudden onset of a stressful condition. Often resistant states are more sensitive to other environmental pressures, or cycle more slowly (e. g. a spore), so it is not advantageous to have the entire clone be in the resistant state. For example, EMT'd (more-mesenchymal) cells are more resistant to some stresses [33, 84], but more sensitive to other stresses such as endoplasmic reticulum stress [100]. Due to these different resistances among states, cancer plasticity can serve as a bet-hedging mechanism itself. Interestingly, the heterogeneity of plasticity would allow each clone to have a different amount of bet hedging, leading to the possibility of evolving a more effective bet hedging probability based on conditions.

1.4.6 Tracking clones *in vitro*

In order to test the clones within a tumor for heterogeneity and stability in plasticity, I needed to use a method to carefully distinguish the progeny of many clones from each other. Methods for keeping track of clones have long been based on adding labels to their DNA, which will be (probably) faithfully passed on to their daughter cells. Retroviruses, which insert their genomes into cells, can be used as these labels. To distinguish different labels, early trackers used differences in the viral integration site [95], which is pseudorandom, but generally biased to expressed genes for certain retrovirus types [101].

Initially, I attempted to adapt this method to modern technology with high efficiency, reading out the viral insertion sites with high-throughput sequencing. This attempt was based on adding a restriction enzyme recognition site to the viral genome, which cuts 20 base pairs away from its site, revealing the section of the genome adjacent to the virus. The method was adapted from a paper that used it to track transposon insertion sites [102], but I had to abandon the method when I determined that the mutation used to induce the restriction enzyme site inhibited viral production.

Instead, I used a system more recently developed, where a random DNA barcode inserted into the viral genome is used to label cells. This method, initially developed for tracking hematopoietic clones [93], is more straightforward due to our ability to amplify sequences flanked by known regions (here, the viral genome) using the polymerase chain reaction. Here, each virus contains a random DNA barcode, which can be inserted into the viral genome during viral assembly. This barcode is then introduced into cells by viral infection, and is faithfully passed on to its progeny. In this way, the lineage of each infected cell is labeled distinctly. By amplifying and sequencing the barcodes, one can determine the clonal composition of any pool of cells. This method has been used to track breast cancer

cells before [12, 13]. In order to track clones to test my hypothesis, I built my own pool of random DNA barcode.

1.4.7 Using clonal tracking to find differences in plasticity

Clonal tracing has been previously used to look for differences in differentiation probabilities. This was most famously applied in the hematopoietic system, where clonal tracing identified individual stem cells with distinct differentiation probabilities [93, 96]. A similar study sought to identify differences in mammary stem cell differentiation when challenged with transplantation [103]. These studies generally use the same strategy: after labeling individual cells, one waits to let them proliferate. During their proliferation, each cell faithfully transmits the label to its daughters. After each cell has grown into a clone (multiple cells), one can separate the resulting population of cells into the different differentiation states in question, usually with fluorescence-activated cell-sorting. Quantifying the labels in each sorted population allows for the evaluation of each clone's differentiation proportions. This method was what we used to test if clones had different biases in their differentiation. The stable identification of lineages with labels (barcodes) also let us track clones over time, to test the stability of these differences.

1.5 References

1. Xu, J., et al., *Mortality in the United States, 2015.* NCHS Data Brief, 2016(267): p. 1-8.
2. Nordling, C.O., *A new theory on cancer-inducing mechanism.* Br J Cancer, 1953. **7**(1): p. 68-72.
3. Knudson, A.G., Jr., *Mutation and cancer: statistical study of retinoblastoma.* Proc Natl Acad Sci U S A, 1971. **68**(4): p. 820-3.
4. Nowell, P.C., *The clonal evolution of tumor cell populations.* Science, 1976. **194**(4260): p. 23-8.
5. Navin, N., et al., *Tumour evolution inferred by single-cell sequencing.* Nature, 2011. **472**(7341): p. 90-4.
6. Wang, Y., et al., *Clonal evolution in breast cancer revealed by single nucleus genome sequencing.* Nature, 2014. **512**(7513): p. 155-60.
7. Paguirigan, A.L., et al., *Single-cell genotyping demonstrates complex clonal diversity in acute myeloid leukemia.* Sci Transl Med, 2015. **7**(281): p. 281re2.
8. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* N Engl J Med, 2012. **366**(10): p. 883-92.
9. Yates, L.R., et al., *Subclonal diversification of primary breast cancer revealed by multiregion sequencing.* Nat Med, 2015. **21**(7): p. 751-9.

10. Park, S.Y., et al., *Heterogeneity for stem cell-related markers according to tumor subtype and histologic stage in breast cancer.* Clin Cancer Res, 2010. **16**(3): p. 876-87.
11. Nolan-Stevaux, O., et al., *Measurement of Cancer Cell Growth Heterogeneity through Lentiviral Barcoding Identifies Clonal Dominance as a Characteristic of Tumor Engraftment.* PLoS One, 2013. **8**(6): p. e67316.
12. Nguyen, L.V., et al., *DNA barcoding reveals diverse growth kinetics of human breast tumour subclones in serially passaged xenografts.* Nat Commun, 2014. **5**: p. 5871.
13. Eirew, P., et al., *Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution.* Nature, 2015. **518**(7539): p. 422-6.
14. Ding, L., et al., *Clonal evolution in relapsed acute myeloid leukaemia revealed by whole-genome sequencing.* Nature, 2012. **481**(7382): p. 506-10.
15. Abubaker, K., et al., *Short-term single treatment of chemotherapy results in the enrichment of ovarian cancer stem cell-like cells leading to an increased tumor burden.* Mol Cancer, 2013. **12**: p. 24.
16. Bhang, H.E., et al., *Studying clonal dynamics in response to cancer therapy using high-complexity barcoding.* Nat Med, 2015.
17. Park, S.Y., et al., *Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype.* J Clin Invest, 2010. **120**(2): p. 636-44.
18. Klevebring, D., et al., *Sequencing of breast cancer stem cell populations indicates a dynamic conversion between differentiation states in vivo.* Breast Cancer Res, 2014. **16**(4): p. R72.
19. Patel, A.P., et al., *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science, 2014. **344**(6190): p. 1396-401.
20. Tirosh, I., et al., *Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma.* Nature, 2016.
21. Lawrence, M.S., et al., *Discovery and saturation analysis of cancer genes across 21 tumour types.* Nature, 2014. **505**(7484): p. 495-501.
22. Cassidy, J.W., C. Caldas, and A. Bruna, *Maintaining Tumor Heterogeneity in Patient-Derived Tumor Xenografts.* Cancer Res, 2015. **75**(15): p. 2963-8.
23. Weinstein, J.N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. **45**(10): p. 1113-20.
24. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-8.
25. Cara, S. and I.F. Tannock, *Retreatment of patients with the same chemotherapy: implications for clinical mechanisms of drug resistance.* Ann Oncol, 2001. **12**(1): p. 23-7.
26. Arena, S., et al., *Emergence of Multiple EGFR Extracellular Mutations during Cetuximab Treatment in Colorectal Cancer.* Clin Cancer Res, 2015. **21**(9): p. 2157-66.
27. Fillmore, C.M. and C. Kuperwasser, *Human breast cancer cell lines contain stem-like cells that self-renew, give rise to phenotypically diverse progeny and survive chemotherapy.* Breast Cancer Res, 2008. **10**(2): p. R25.
28. Li, X., et al., *Intrinsic resistance of tumorigenic breast cancer cells to chemotherapy.* J Natl Cancer Inst, 2008. **100**(9): p. 672-9.
29. Sharma, S.V., et al., *A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations.* Cell, 2010. **141**(1): p. 69-80.
30. Singh, A. and J. Settleman, *EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer.* Oncogene, 2010. **29**(34): p. 4741-51.
31. Saxena, M., et al., *Transcription factors that mediate epithelial-mesenchymal transition lead to multidrug resistance by upregulating ABC transporters.* Cell Death Dis, 2011. **2**: p. e179.

32. Del Vecchio, C.A., et al., *De-differentiation confers multidrug resistance via noncanonical PERK-Nrf2 signaling.* PLoS Biol, 2014. **12**(9): p. e1001945.
33. Creighton, C.J., et al., *Residual breast cancers after conventional therapy display mesenchymal as well as tumor-initiating features.* Proc Natl Acad Sci U S A, 2009. **106**(33): p. 13820-5.
34. Roesch, A., et al., *A temporarily distinct subpopulation of slow-cycling melanoma cells is required for continuous tumor growth.* Cell, 2010. **141**(4): p. 583-94.
35. Chaffer, C.L., et al., *Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state.* Proc Natl Acad Sci U S A, 2011. **108**(19): p. 7950-5.
36. Gupta, P.B., et al., *Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells.* Cell, 2011. **146**(4): p. 633-44.
37. Yang, G., et al., *Dynamic equilibrium between cancer stem cells and non-stem cancer cells in human SW620 and MCF-7 cancer cell populations.* Br J Cancer, 2012. **106**(9): p. 1512-9.
38. Newby, G.A. and S. Lindquist, *Blessings in disguise: biological benefits of prion-like mechanisms.* Trends Cell Biol, 2013. **23**(6): p. 251-9.
39. Beaumont, H.J., et al., *Experimental evolution of bet hedging.* Nature, 2009. **462**(7269): p. 90-3.
40. Kimura, M. and T. Ohta, *On some principles governing molecular evolution.* Proc Natl Acad Sci U S A, 1974. **71**(7): p. 2848-52.
41. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
42. Tzelepis, K., et al., *A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia.* Cell Rep, 2016. **17**(4): p. 1193-1205.
43. Hart, T., et al., *High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities.* Cell, 2015. **163**(6): p. 1515-26.
44. Wang, T., et al., *Identification and characterization of essential genes in the human genome.* Science, 2015. **350**(6264): p. 1096-101.
45. Jackson, E.L., et al., *The differential effects of mutant p53 alleles on advanced murine lung cancer.* Cancer Res, 2005. **65**(22): p. 10280-8.
46. Shih, C., et al., *Passage of phenotypes of chemically transformed cells via transfection of DNA and chromatin.* Proc Natl Acad Sci U S A, 1979. **76**(11): p. 5714-8.
47. Javier, R.T. and J.S. Butel, *The history of tumor virology.* Cancer Res, 2008. **68**(19): p. 7693-706.
48. Lohmann, D.R. and B.L. Gallie, *Retinoblastoma: revisiting the model prototype of inherited cancer.* Am J Med Genet C Semin Med Genet, 2004. **129c**(1): p. 23-8.
49. Kinzler, K.W. and B. Vogelstein, *Lessons from hereditary colorectal cancer.* Cell, 1996. **87**(2): p. 159-70.
50. Bradford, P.T., et al., *Cancer and neurologic degeneration in xeroderma pigmentosum: long term follow-up characterises the role of DNA repair.* J Med Genet, 2011. **48**(3): p. 168-76.
51. Fearon, E.R. and B. Vogelstein, *A genetic model for colorectal tumorigenesis.* Cell, 1990. **61**(5): p. 759-67.
52. Vermeulen, L., et al., *Defining stem cell dynamics in models of intestinal tumor initiation.* Science, 2013. **342**(6161): p. 995-8.
53. Martincorena, I., et al., *Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin.* Science, 2015. **348**(6237): p. 880-6.
54. Lee, E.Y. and W.J. Muller, *Oncogenes and tumor suppressor genes.* Cold Spring Harb Perspect Biol, 2010. **2**(10): p. a003236.
55. Pfeifer, H., et al., *Prevalence and dynamics of bcr-abl kinase domain mutations during imatinib treatment differ in patients with newly diagnosed and recurrent bcr-abl positive acute lymphoblastic leukemia.* Leukemia, 2012. **26**(7): p. 1475-81.

56. McFarland, C.D., et al., *Impact of deleterious passenger mutations on cancer progression.* Proc Natl Acad Sci U S A, 2013. **110**(8): p. 2910-5.
57. Ostrow, S.L., et al., *Cancer evolution is associated with pervasive positive selection on globally expressed genes.* PLoS Genet, 2014. **10**(3): p. e1004239.
58. Resch, A.M., et al., *Widespread positive selection in synonymous sites of mammalian genes.* Mol Biol Evol, 2007. **24**(8): p. 1821-31.
59. Quax, T.E., et al., *Codon Bias as a Means to Fine-Tune Gene Expression.* Mol Cell, 2015. **59**(2): p. 149-61.
60. Mauro, V.P. and S.A. Chappell, *A critical analysis of codon optimization in human therapeutics.* Trends Mol Med, 2014. **20**(11): p. 604-13.
61. Chamary, J.V. and L.D. Hurst, *Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals.* Genome Biol, 2005. **6**(9): p. R75.
62. Presnyak, V., et al., *Codon optimality is a major determinant of mRNA stability.* Cell, 2015. **160**(6): p. 1111-24.
63. Kryazhimskiy, S. and J.B. Plotkin, *The population genetics of dN/dS.* PLoS Genet, 2008. **4**(12): p. e1000304.
64. Van den Eynden, J., S. Basu, and E. Larsson, *Somatic Mutation Patterns in Hemizygous Genomic Regions Unveil Purifying Selection during Tumor Evolution.* PLoS Genet, 2016. **12**(12): p. e1006506.
65. Liu, Y., et al., *TP53 loss creates therapeutic vulnerability in colorectal cancer.* Nature, 2015. **520**(7549): p. 697-701.
66. Kazanov, M.D., et al., *APOBEC-Induced Cancer Mutations Are Uniquely Enriched in Early-Replicating, Gene-Dense, and Active Chromatin Regions.* Cell Rep, 2015. **13**(6): p. 1103-9.
67. Kim, K., et al., *Chromatin structure-based prediction of recurrent noncoding mutations in cancer.* Nat Genet, 2016. **48**(11): p. 1321-1326.
68. Mayr, C. and D.P. Bartel, *Widespread shortening of 3'UTRs by alternative cleavage and polyadenylation activates oncogenes in cancer cells.* Cell, 2009. **138**(4): p. 673-84.
69. Berkovits, B.D. and C. Mayr, *Alternative 3' UTRs act as scaffolds to regulate membrane protein localization.* Nature, 2015. **522**(7556): p. 363-7.
70. Lewandowska, M.A., *The missing puzzle piece: splicing mutations.* Int J Clin Exp Pathol, 2013. **6**(12): p. 2675-82.
71. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
72. Swanton, C., et al., *APOBEC Enzymes: Mutagenic Fuel for Cancer Evolution and Heterogeneity.* Cancer Discov, 2015. **5**(7): p. 704-12.
73. Rockah-Shmuel, L., A. Toth-Petroczy, and D.S. Tawfik, *Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations.* PLoS Comput Biol, 2015. **11**(8): p. e1004421.
74. Guo, H.H., J. Choe, and L.A. Loeb, *Protein tolerance to random amino acid change.* Proc Natl Acad Sci U S A, 2004. **101**(25): p. 9205-10.
75. Cassa, C.A., et al., *Estimating the selective effects of heterozygous protein-truncating variants from human exome data.* Nat Genet, 2017.
76. Bakhoum, S.F., et al., *DNA-damage response during mitosis induces whole-chromosome missegregation.* Cancer Discov, 2014. **4**(11): p. 1281-9.
77. Bester, A.C., et al., *Nucleotide deficiency promotes genomic instability in early stages of cancer development.* Cell, 2011. **145**(3): p. 435-46.

78. Marescalco, M.S., et al., *Genome-wide analysis of recurrent copy-number alterations and copy-neutral loss of heterozygosity in head and neck squamous cell carcinoma.* J Oral Pathol Med, 2014. **43**(1): p. 20-7.
79. Svobodova, K., et al., *Copy number neutral loss of heterozygosity at 17p and homozygous mutations of TP53 are associated with complex chromosomal aberrations in patients newly diagnosed with myelodysplastic syndromes.* Leuk Res, 2016. **42**: p. 7-12.
80. Gimelbrant, A., et al., *Widespread monoallelic expression on human autosomes.* Science, 2007. **318**(5853): p. 1136-40.
81. Mojtahedi, M., et al., *Cell Fate Decision as High-Dimensional Critical State Transition.* PLoS Biol, 2016. **14**(12): p. e2000640.
82. Moris, N., C. Pina, and A.M. Arias, *Transition states and cell fate decisions in epigenetic landscapes.* Nat Rev Genet, 2016. **17**(11): p. 693-703.
83. Huang, S., et al., *Bifurcation dynamics in lineage-commitment in bipotent progenitor cells.* Dev Biol, 2007. **305**(2): p. 695-713.
84. Yang, J. and R.A. Weinberg, *Epithelial-mesenchymal transition: at the crossroads of development and tumor metastasis.* Dev Cell, 2008. **14**(6): p. 818-29.
85. Guo, W., et al., *Slug and Sox9 cooperatively determine the mammary stem cell state.* Cell, 2012. **148**(5): p. 1015-28.
86. Polyak, K. and R.A. Weinberg, *Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits.* Nat Rev Cancer, 2009. **9**(4): p. 265-73.
87. Mani, S.A., et al., *The epithelial-mesenchymal transition generates cells with properties of stem cells.* Cell, 2008. **133**(4): p. 704-15.
88. Bulfoni, M., et al., *In patients with metastatic breast cancer the identification of circulating tumor cells in epithelial-to-mesenchymal transition is associated with a poor prognosis.* Breast Cancer Res, 2016. **18**(1): p. 30.
89. Bonnet, D. and J.E. Dick, *Human acute myeloid leukemia is organized as a hierarchy that originates from a primitive hematopoietic cell.* Nat Med, 1997. **3**(7): p. 730-7.
90. Ema, H., Y. Morita, and T. Suda, *Heterogeneity and hierarchy of hematopoietic stem cells.* Exp Hematol, 2014. **42**(2): p. 74-82.e2.
91. Chaffer, C.L., et al., *Poised chromatin at the ZEB1 promoter enables breast cancer cell plasticity and enhances tumorigenicity.* Cell, 2013. **154**(1): p. 61-74.
92. Chang, H.H., et al., *Transcriptome-wide noise controls lineage choice in mammalian progenitor cells.* Nature, 2008. **453**(7194): p. 544-7.
93. Lu, R., et al., *Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding.* Nat Biotechnol, 2011. **29**(10): p. 928-33.
94. Naik, S.H., et al., *Diverse and heritable lineage imprinting of early haematopoietic progenitors.* Nature, 2013. **496**(7444): p. 229-32.
95. Lemischka, I.R., D.H. Raulet, and R.C. Mulligan, *Developmental potential and dynamic behavior of hematopoietic stem cells.* Cell, 1986. **45**(6): p. 917-27.
96. Cheung, A.M., et al., *Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice.* Blood, 2013. **122**(18): p. 3129-37.
97. Shipitsin, M., et al., *Molecular definition of breast tumor heterogeneity.* Cancer Cell, 2007. **11**(3): p. 259-73.
98. Kussell, E., et al., *Bacterial persistence: a model of survival in changing environments.* Genetics, 2005. **169**(4): p. 1807-14.
99. Razooky, B.S., et al., *A hardwired HIV latency program.* Cell, 2015. **160**(5): p. 990-1001.
100. Feng, Y.X., et al., *Epithelial-to-mesenchymal transition activates PERK-eIF2alpha and sensitizes cells to endoplasmic reticulum stress.* Cancer Discov, 2014. **4**(6): p. 702-15.

101. Ambrosi, A., et al., *Estimated comparative integration hotspots identify different behaviors of retroviral gene transfer vectors.* PLoS Comput Biol, 2011. **7**(12): p. e1002292.
102. van Opijnen, T., K.L. Bodi, and A. Camilli, *Tn-seq: high-throughput parallel sequencing for fitness and genetic interaction studies in microorganisms.* Nat Methods, 2009. **6**(10): p. 767-72.
103. Nguyen, L.V., et al., *Clonal analysis via barcoding reveals diverse growth and differentiation of transplanted mouse and human mammary stem cells.* Cell Stem Cell, 2014. **14**(2): p. 253-63.

1.6 Figures

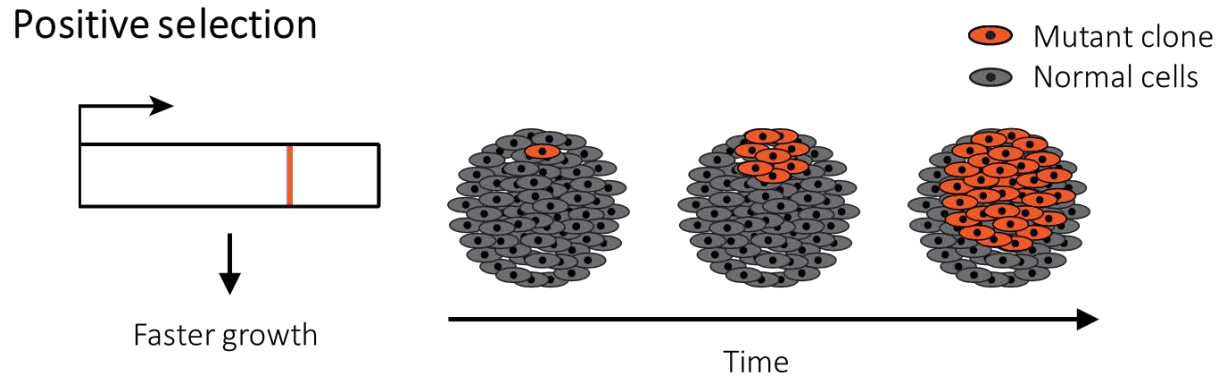


Figure 1. Positive selection acts on mutations that increase the growth rate of a clone.

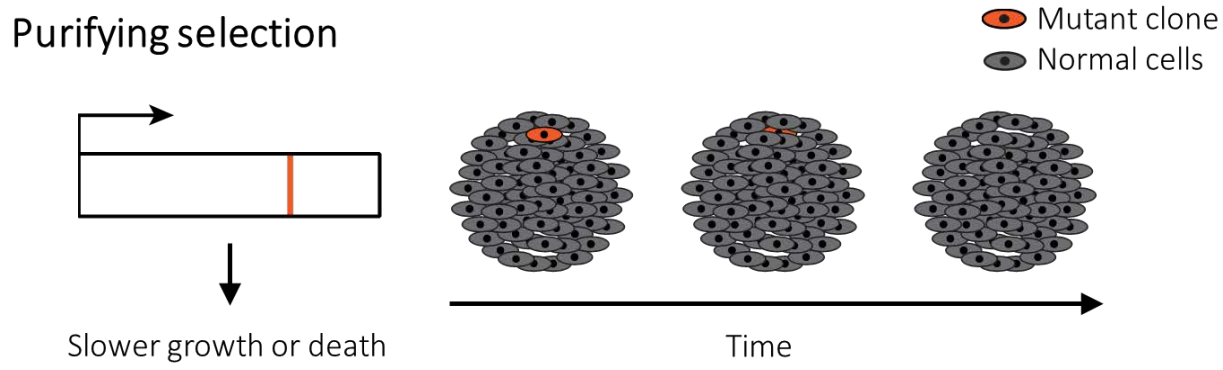


Figure 2. Purifying selection acts to remove mutations that decrease the growth rate of a clone.

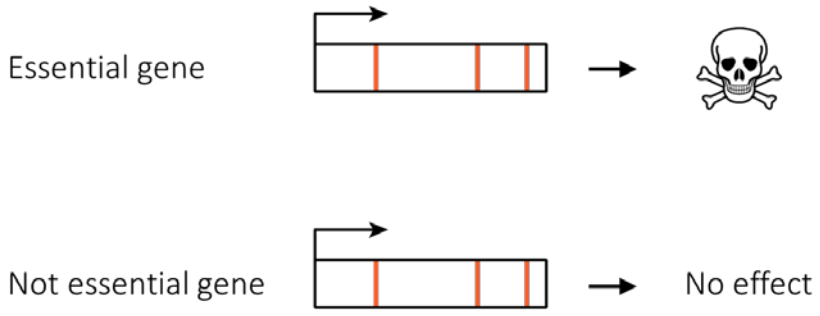


Figure 3. Purifying selection removes mutations from essential genes

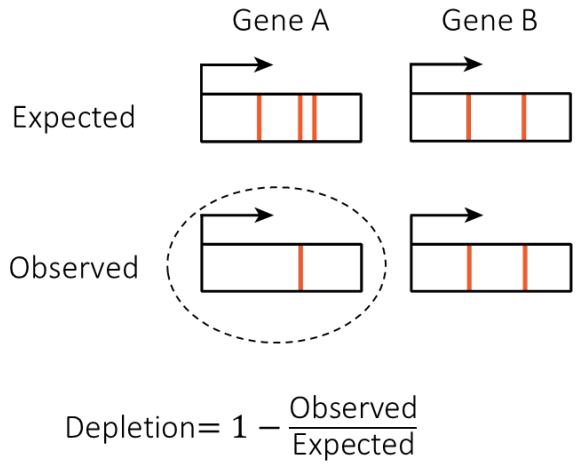


Figure 4. Evidence of purifying selection from observing fewer mutations than expected

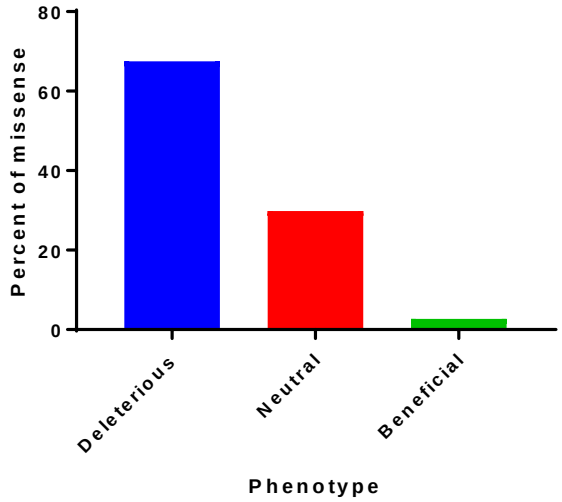


Figure 5. Most missense mutations cause a deleterious phenotype. Data evaluated from a published *in vitro* analysis of mutations in the enzyme M. HaeIII (from *Haemophilus aegyptius*). [73]

Chapter 2: Identification of Genes under Purifying Selection in Human Cancers

The material presented in this chapter was adapted from a manuscript currently in submission:

Mathis RA, Sokol ES, Gupta PB. Identification of Genes under Purifying Selection in Human Cancers. 2017.

Authors' contributions: All authors conceptualized the project and methodology, and participated in writing the manuscript. All authors read and approved the final manuscript. RM and ES wrote the software and perform the formal analysis. PG provided supervision.

Chapter 2

Introduction

Tumor formation is an evolutionary process driven by positive selection for somatic mutations that provide a competitive advantage to cancer cells [1-3]. While positive selection drives phenotypic change, it only enriches for a miniscule fraction of the mutations in tumor genomes [4, 5]. During species evolution, most newly arising mutations are deleterious, and are eliminated by negative (or purifying) selection before they can become substitutions fixed in the population of individuals [6-9]. In principle, negative selection could also impact cancer evolution [10, 11], and there is evidence of purifying selection in hemizygous regions of cancer genomes [12]. However, the extent to which this force shapes the pattern of somatic mutations in tumors is not known. In this study, we provide evidence that purifying selection is widespread in cancer genomes and acts to remove mutations from genes that contribute to the survival or growth of cancer cells. In this way, the pattern of mutations in patient tumors reveals the vulnerabilities of human cancers *in vivo*.

Results

Genes that are expressed or essential have fewer missense mutations (substitutions)

If purifying selection were significant during tumor evolution, it would reduce overall substitution rates by preventing the fixation of deleterious somatic mutations in genes contributing to tumor growth. To examine this possibility, we analyzed the mutational profiles of 5057 tumors of diverse cancer types sequenced by The Cancer Genome Atlas (TCGA) [13]. Since genes can only impact tumor growth if they are expressed, our first analysis was to compare substitution rates between expressed and non-expressed genes (Figure 1A). Each gene's exon mutation rate was normalized relative to its intron mutation rate; this controlled for gene-to-gene variations in mutation rates arising from differences in

chromatin accessibility and early-vs-late replication times, among other position factors [5] (Figure 1B). After controlling for all of these effects, expressed genes had significantly fewer substitutions than non-expressed genes across three tumor types— with a 57% reduction in melanomas ($p<10^{-20}$), a 51% reduction in lung adenocarcinomas ($p<10^{-20}$), and a 14% reduction in colorectal adenocarcinomas ($p<10^{-20}$) (Figure 1A). Absent this reduction, we estimate there would have been 167–416 additional mutations in the exons of expressed genes per tumor, depending on the cancer type. This depletion of missense mutations is similar to the 66-83% of missense mutations observed to impact a protein's functionality, based on experimental mutagenesis [14].

Transcription-coupled repair (TCR) [15] has been previously reported as a mechanism through which mutations are eliminated from expressed genes. To quantify TCR's effects, we compared substitution rates between transcribed (template) and non-transcribed (coding) strands in melanomas and lung adenocarcinomas. As expected, TCR lowered overall substitution rates in expressed genes. However, there was a 31-45% additional reduction that could not be accounted for by TCR (Figure 2A-B). These findings were consistent with a model in which mutations were being eliminated by purifying selection prior to their fixation.

Amino acid substitutions with similar physicochemical traits are more acceptable during both tumor microevolution and species macroevolution

Mutations resulting in the substitution of amino acids with similar physicochemical properties (conservative substitutions) are less likely to be deleterious to protein function, relative to non-conservative substitutions [6, 16]. If this were the case in tumors, purifying selection should act less strongly on mutations resulting in conservative amino acid substitutions. To test this prediction, we segregated mutations into classes based on the amino acid substitutions that they generated. In total,

there were mutations in all of the 150 substitution classes that are possible by mutating a single base pair in codons. We quantified the strength of negative selection on each mutation-substitution class to identify pairs of amino acids (A_1 , A_2) that were most readily substituted in either direction (A_1 -> A_2 and A_2 -> A_1) in tumors (Figure 3A,B). This analysis identified several subsets of amino acids with similar physicochemical properties that were interchangeable in tumors: the hydrophobic amino acids isoleucine, leucine, valine, and methionine; the positively charged amino acids arginine, histidine, and lysine; and the positively charged and positive-polar amino acids arginine and glutamine. The analysis also identified several amino acids with similar structures but differing charges that were interchangeable (Gln⇔Glu and Asp⇔Asn), suggesting that such substitutions might minimize steric hindrances and be frequently tolerated. We conclude that mutations resulting in conservative substitutions were less often eliminated by purifying selection in tumors— presumably because they were less likely to disrupt protein folding or function.

The constraints imposed on protein folding and function during tumor microevolution might in principle be comparable to those imposed during the macroevolution of species. We therefore compared the amino acid substitutions that were tolerated in tumors with those that were most commonly tolerated across macro-evolutionary time scales. Surprisingly, we found that interchangeable amino acids identified using BLOcks of Amino Acid SUbstitution Matrix (BLOSUM; [17]) analysis— which quantifies substitutions within highly conserved protein domains across millions of years of species evolution— were nearly identical to those identified in the tumor analysis ($p < 7*10^{-6}$) (Figure 3B,C). However, this concordance was only observed if: (1) the macro-evolutionary analysis was performed for closely related proteins (BLOSUM90, but not BLOSUM45/62), and (2) the BLOSUM90 amino acid substitutions were limited to those that are possible by mutating a single DNA base in codons; both of these constraints reflect the fact that the substitution rates in tumors are much lower than those observed in comparisons across species. Moreover, this analysis revealed that several substitutions that

were well tolerated in tumors, which could not be understood on the basis of their physicochemical traits (e.g. Glu⇔Lys, Ser⇔Ala, Ala⇔Thr), were also more tolerated across the macro-evolutionary time scales associated with speciation, suggesting that they are in fact more permissible than others (Figure 3B). After considering these findings together with the functional observations above, we concluded that purifying selection has a significant role in shaping the global constellation of substitutions (fixed mutations) found in tumors.

Purifying selection targets genes that are important for tumor growth

Since these findings established that negative selection occurred at a genome-wide scale in tumors, we next asked whether we could identify individual genes that were substrates of purifying selection. We found genes associated with essential processes, such as transcription (*MED15*, *MED19*) and cell division (*ANAPC2*, *CEP72*) to be under purifying selection in tumors. However, we could not detect evidence of purifying selection in genes with too few mutations across the sequenced tumors. To work around this, we looked for purifying selection in sets of genes with known biological functions [18]. We found that genes which function in essential cellular processes— e.g. RNA metabolism and DNA replication— are under the strongest purifying selection across all tumor types (Figure 4A, B). In addition to these gene sets showing a depletion of mutations, we found that in each set under purifying selection, the majority of genes showed fewer mutations than expected (Figure 5C).

To support our observation that genes under purifying selection were enriched in essential cellular processes, we examined if these genes were known to be essential when perturbed. Assembling the results of three pooled CRISPR screens [19-21], we found that sets of genes essential in most tested cell lines were under greater purifying selection (Figure 4D). We also found that purifying selection has a

strong power to find essential genes (Figure 4E). This showed that genes under purifying selection in tumors are functionally essential, suggesting it reveals genes important for tumor growth or survival.

Although many genes under purifying selection across tumors are essential, such genes are not likely to be good targets for treating cancer, as they likely also have essential functions in normal cells. To get around this, we aimed to identify genes under increased selection in particular tumor types, relative to other tumors. To identify genes under increased purifying selection in specific tumor types, we developed a statistical approach that controlled for differences in the stage of tumor evolution, gene-specific variations in mutation rate within tumors, and genome-wide variations in mutation rates across tumors. Using this method, we could identify genes under purifying selection in specific tumor types—e.g. in lung tumors versus all other tumor types.

In lung adenocarcinomas, we identified 508 genes as strong substrates for purifying selection, enriched in 11 of the pathways in the network data exchange database (NDEx) (Figure 5, Figure 6A, Table 1) [22]. These included: 11 genes in pathways related to EGFR signaling (ERBB2/ERBB3, EGFR internalization, ERBB1 receptor proximal pathway, $p<3x10^{-3}$, Figure 5), and the AXL kinase. These pathways are both targeted by approved therapies for lung cancers: erlotinib/gefitinib (EGFR) and crizotinib (MET/AXL). Our analysis also identified *FGFR3*, a key driver of non-small cell lung cancer (NSCLC), which is activated by mutation in 6-8% of NSCLCs and is currently being explored as a therapy target [23-27].

In cutaneous melanomas, we identified 848 genes that were targets of purifying selection. Consistent with the established role of UV-induced damage in this cancer type, these included 27 genes in key pyrimidine dimer repair pathways: nucleotide-excision (*ERCC2*, *ERCC5*), base excision (*APEX2, POLE*), mismatch repair (*RFC1, RFC4*), and trans-lesion replication (*REV1, REV3L*) (Figure 6A,B, Table 2). While UV does not directly cause double-stranded breaks (DSBs), such breaks arise indirectly during NER

and are the primary cause of cell death [28]. Consistent with this, we identified a number of genes that repair DSBs in the ATM and Fanconi Anemia pathways (ATM & FANCONI pathways, $p < 4 \times 10^{-5}$; Table 2)—including two members of the core Fanconi Anemia complex (*FANCC, FANCL*), *ATM* and its phosphorylation target CHK2, and 2/3 proteins in the MRN complex (NBN and RAD50) (Figure 6B). We also identified all four components of the cohesin complex (SMC1A, SMC3, STAG2, RAD21), which, independently of its role in mediating sister chromatid cohesion, is phosphorylated by ATM and required for repairing DNA DSBs by homologous recombination [29-31]. Collectively, these findings indicate that purifying selection preserves the function of DNA repair pathways in melanomas. Because many of these pathways have established roles in promoting resistance to the DNA damage caused by radiation and chemotherapies [32-36]; this might explain why such therapies are almost completely ineffective when applied to melanomas.

We were again able to identify purifying selection on sets of genes with known biological function, this time looking for increased selection in a particular tumor type. Gene sets under increased purifying selection in melanomas are related to a number of pathways active in processes known to be important in melanomas (Figure 6C, Table 5). Describing 599/927 genes in these sets, we found many known pathways involved in melanoma growth and survival, such as the sonic hedgehog, WNT, NFκB, PI3K, EGFR, and INFγ pathways, and the proteasome [37-44]. We also found a pathway required for immune suppression in melanomas, TNFα [45].

Importantly, genes sets under increased purifying selection in melanomas contain fewer genes essential for cell viability when compared to gene sets under purifying selection in all tumors (Figure 6D). This showed that this method met its goal of finding genes under purifying selection that were not generally-essential.

When attempting to extend this analysis to other tumor types, we found that there were not enough passenger mutations identified to provide the statistical power needed for the analysis. How much more benefit would be obtained by sequencing additional tumors? Using numerical simulations, we estimated the number of new genes that would be discovered by sequencing 500 to 3000 additional tumors of each cancer type (Figure 7). For all tumor types, sequencing no more than 500-3000 additional tumors would be sufficient to discover nearly all of the genes under purifying selection that have yet to be identified. In addition, we established the optimal combination of tumor types to sequence that would maximize the number of new genes discovered as substrates of purifying selection (Figure 7).

Discussion

Our findings indicate that many genes are under purifying selection in specific tumor types, and are therefore not likely to be generally required for the survival of all cell types. Such genes could either be essential in particular cell types, or become essential as a consequence of synthetic interactions with the genetic or metabolic alterations associated with tumor formation, as recently reported [46, 47]. We propose calling genes under purifying selection in tumors 'enablers', to distinguish them from recurrently-mutated oncogenes.

Most efforts to uncover genes important to cancers, and therefore possible targets for cancer therapy, have focused on uncovering signals of positive selection [4, 5]. Such genes are recurrently mutated across tumors, allowing for their identification. Although this has been a fruitful approach, it would fail to uncover those genes important to cancers that are not activated through mutation. For this reason, we investigated the role of purifying selection in cancer, which we expected would act to remove mutations in genes important to cancers.

In contrast to the evolution of organisms, purifying selection in cancers has been posited to be relatively weak, as we observe many mutations in tumors with putative detrimental effects, presumed to be dragged to enrichment via linkage to positively-selected alleles [11, 48]. However, recent analysis of human and fly populations has determined that additional detrimental mutants display a more-than-additive impact on fitness, due to synergistic epistasis [49]. This increases the effect of each additional mutation on the clone's fitness. Synergistic epistasis, coupled with the experimentally-determined probability of a missense mutation being problematic for a protein's functionality (66-83%) [14], and the large number of mutations in cancers [4], suggests that many missense mutations in cancer could cause strong fitness defects. Such mutations would likely be under negative selection, particularly in the background of many other mutations. In addition, limited evidence of purifying selection has been found in hemizygous regions of cancer genomes [12]. These evidences support our observation of widespread purifying selection in tumors.

We developed an analysis to look for purifying selection without using the evolutionary biology technique dN/dS. dN/dS, the ratio of missense mutations to synonymous mutations [6], has been used to look for evidence of purifying selection in diverse organisms. However, it is inappropriate to apply methods of estimating dN/dS designed for comparisons of distinct species to samples within a single population, such as tumors [50]. Additionally, the strength of signal in this method depends on the assumption that synonymous mutations have a limited phenotypic effect. Recent scholarship has uncovered that many synonymous mutations are, in fact, not silent. The codons of genes are often under selection to match more abundant tRNAs; synonymous mutations therefore lead to inefficient coding, reducing the speed of translation, which can change the rate of protein production or protein conformation [51-53]. These changes can also result from the effects of synonymous mutations on secondary mRNA structure; there is evidence that many synonymous mutations affect mRNA stability and, through it, expression level [54-57]. The importance of codon choice is illustrated by the

observation that a synonymous mutations plays a major role in the phenotype of the most common cystic fibrosis variant, ΔF508 [58]. In the highly competitive environment of tumors, we would expect the phenotype of synonymous mutations to be even stronger. To avoid these potentially confounding factors, we instead estimated the expected number of missense mutations.

These findings show that purifying selection significantly influences the pattern of mutations in cancer genomes, reducing the rate at which substitutions accumulate in genes that are important for tumor growth. By comparing the spectrum and frequency of amino acid substitutions in cancers with those seen during the evolution of species, we found evidence of widespread purifying selection in cancers. This suggested that genes with fewer mutations than expected are under purifying selection in human cancers. In addition, these findings also show that genes under purifying selection in cancers are critical for the survival/proliferation of cancer cells in CRISPR screens, providing functional evidence that purifying selection prevents the accumulation of deleterious mutations in genes essential for tumor growth.

Genes that are important for the viability of cancer cells would not be attractive targets for therapy if they are generally required for the survival of all (including normal) cell types. We have therefore sought genes that are under increased purifying selection in only a subset of cancer types, thereby excluding genes that are essential for all cell types. This identified a variety of processes and pathways under increased selection in melanomas and in lung adenocarcinomas, as compared to other tumor types.

Using signatures of purifying selection to discover enablers provides an exciting opportunity to systematically identify hundreds of new vulnerabilities of cancer. As the vulnerabilities of human tumors will remain opaque to direct experimentation, and only approached by models, our observation

of purifying selection in cancers allows an unprecedented view into the dependencies of human cancers *in vivo*.

Methods

Tumor mutation data

Mutation Annotation Format files for 11 tumor types generated by The Cancer Genome Atlas (TCGA) were downloaded from the Broad Firehose[59]. Tumor types downloaded were lung adenocarcinoma (533 tumors), cutaneous melanoma (290 tumors), colorectal adenocarcinoma (489 tumors), bladder urothelial carcinoma (395 tumors), breast invasive carcinoma (977 tumors), glioma (796 tumors), uterine corpus endometrial carcinoma (248 tumors), head and neck squamous cell carcinoma (510 tumors), liver hepatocellular carcinoma (198 tumors), prostate adenocarcinoma (332 tumors), and stomach adenocarcinoma (289 tumors). Mutations were filtered to remove all but single base-pair missense mutations in exons.

Non-coding (intron) mutation data from were acquired from published analyses. [5]

RNA data

Level 3 normalized RNA sequencing data quantified with RNA-Seq by Expectation Maximization (RSEM)[60] were downloaded from the Broad Firehose[59]. These data are quartile-normalized RSEM count estimates.

Gene-length and sequence information

Gene length information was downloaded from UniProt (http://www.uniprot.org/), and coding sequences were downloaded from BioMart (http://www.biomart.org/).

Calculations:

Mutation rates in expressed and non-expressed genes

For tumor $t \in$ tumor type $T_i \in T$, where $T =$ {Lung adenocarcinoma, skin cutaneous melanoma, colorectal adenocarcinoma} (see Tumor mutation data, above); and for gene $g \in G$, where $G =$ all sequenced genes; and where $L_g =$ the length of gene g in amino acids (a.a.s);

$$m(g, T_i) = \sum_{t \in T_i} |\text{missense mutations in } g \text{ in } t|$$

Where

$$R(g, t) = \{\text{RNA sequencing counts (see RNA data) for gene } g \text{ in tumor } t \ \wedge g \in G \wedge t \in T_i\}$$

Define expressed genes

$$G_{e,T_i} = \{g : \ g \in G \wedge |\{t : \ R(g,t) \ > 8 \ \wedge t \in T_i \}| > 0.95 \ |T_i| \wedge m(g, T_i) \geq 1\}$$

and not-expressed genes as

$$G_{n,\,T_i} = \{g : \ g \in G \wedge \ |\{t : \ R(g,t) < 8 \ \wedge t \in T_i \}| > 0.95 \ |T_i| \wedge m(g, T_i) \geq 1\}$$

Determine an expected number of mutations for each gene by means of the gene's relative non-coding mutation rate, the average mutational rate in not expressed genes, and the length of the gene's coding sequence:

$$E(g, T_i) = nm(g) * \frac{|G|}{\sum_{\gamma \in G} nm(\gamma)} * \frac{\sum_{\gamma \in G_{n,T_i}} m(\gamma, T_i)}{\sum_{\gamma \in G_{n,T_i}} L_\gamma} * L_g$$

Where $nm(g)$ = the non-coding mutation rate for gene g calculated from published whole-genome sequencing of tumor samples [5].

To calculate the significance of the depletion in mutations in expressed genes,

$$\left\{\frac{m(g,T_i)}{E(g,T_i)} : g \in G_{e,T_i}\right\} \text{ and } \left\{\frac{m(g,T_i)}{E(g,T_i)} : g \in G_{e,T_i}\right\}$$

were compared with a two-tailed Wilcoxon Rank-Sum test.

The proportion of mutations depleted in expressed genes relative to not expressed genes was calculated as

$$Pd_{T_i} = 1 - \left(\frac{\sum_{g \in G_{e,T_i}} m(g,T_i)}{\sum_{g \in G_{e,T_i}} E(g,T_i)} * \frac{\sum_{g \in G_{n,T_i}} E(g,T_i)}{\sum_{g \in G_{n,T_i}} m(g,T_i)}\right)$$

The number of additional expressed mutations expected in sequenced tumors was calculated as

$$\frac{\sum_{g \in G_e} m(g,T_i)}{|T_i|} * \frac{1}{1 - Pd_{T_i}}$$

Determining the effect of intron mutation rate controls on mutation rate covariates

Using the intron mutation rate to estimate the background mutation rates of genes should ideally control for known gene mutation rate covariates, including replication time, chromatin accessibility, and GC nucleotide percentage. Where T_i = lung adenocarcinomas, observed and expected (intron-normalized) missense mutations were calculated for each expressed gene as in "Mutation rates in expressed and non-expressed genes," above. Replication time and chromatin accessibility of each gene were accessed from a published source [5]. The %GC nucleotides of each gene was determined from each gene's coding sequence.

To determine these relationships before controlling via the intron mutation rate, an expected number of mutations was calculated for each gene assuming a uniform mutation rate, or E^0(g):

$$E^0(g) = L_g * \frac{\sum_{\gamma \in G} m(\gamma)}{\sum_{\gamma \in G} L_\gamma}$$

For both the intron-normalized expected and the uniform mutation rate expected, Each covariate score for expressed genes was plotted against the $\log_2$ observed / expected mutations of those genes, and a linear regression determined.

Estimating the effects of transcription-coupled repair

To estimate the effect of transcription-coupled repair, mutation rates were quantified in the transcribed and not-transcribed strands. For each missense mutation μ, define the starting base $\left(B_\mu^0\right)$ and ending base $\left(B_\mu^1\right)$, and its indistinguishable complement with starting base $B_\mu'^0$ and $B_\mu'^1$. There are six kinds of recognizable base-pair transitions, as some are indistinguishable from a mutation in the opposite strand.

For G>T mutations in lung adenocarcinomas and C>T mutations in melanomas, mutation rates were calculated on a gene-by-gene basis in expressed and not-expressed genes.

Define f_{g,T_i}^{β} as the number of mutations in of the class $\beta^0 > \beta^1$ (e.g. C>T) in the transcribed (template) DNA strand of gene *g* in tumor type T_i, and $f_{g,T_i}^{\beta\prime}$ as the mutations in the class $\beta'^0 > \beta'^1$ in the not-transcribed (coding) DNA strand of gene *g* in tumor type T_i:

$$f_{g,T_i}^{\beta} = \sum_{t \in T_i} \left| \{ \mu : \mu \in \text{missense in } g \text{ in } t \ \wedge B_{\mu}^{0} = \beta^{0} \wedge B_{\mu}^{1} = \beta^{1} \} \right|$$

and

$$f_{g,T_i}^{\prime\beta} = \sum_{t \in T_i} \left| \{ \mu : \mu \in \text{missense in } g \text{ in } t \ \wedge B_{\mu}^{\prime 0} = \beta^{0} \wedge B_{\mu}^{\prime 1} = \beta^{1} \} \right|$$

Also define S_g^{β} and $S_g^{\prime\beta}$ as the number of sites that could mutate in the transcribed (template) DNA strand and not-transcribed (coding) DNA strand of gene *g* respectively, or

$$S_g^{\beta} = |\{\text{base } B : B \in g \ \wedge B = \beta^{0}\}|$$

$$S_g^{\prime\beta} = |\{\text{base } B : B \in g \ \wedge B = \beta^{\prime 0}\}|$$

Determine an expected number of mutations for each gene, and for each strand, by means of the gene's relative non-coding mutation rate, the average mutational rate in expressed genes, and the length of the gene's coding sequence of the base in question (for the template strand) or its complement (for the coding strand):

$$E(g, T_i, \beta) = nm(g) * \frac{|G|}{\sum_{\gamma \in G} nm(\gamma)} * \frac{\sum_{\gamma \in G_{n,T_i}} f_{\gamma,T_i}^{\beta} + f_{\gamma,T_i}^{\prime\beta}}{\sum_{\gamma \in G_{n,T_i}} S_{\gamma}^{\beta} + S_{\gamma}^{\prime\beta}} * S_g^{\beta}$$

And:

$$E'(g, T_i, \beta) = nm(g) * \frac{|G|}{\sum_{\gamma \in G} nm(\gamma)} * \frac{\sum_{\gamma \in G_{n,T_i}} f_{\gamma,T_i}^{\beta} + f_{\gamma,T_i}^{\prime\beta}}{\sum_{\gamma \in G_{n,T_i}} S_{\gamma}^{\beta} + S_{\gamma}^{\prime\beta}} * S_g^{\prime\beta}$$

To test the relative mutation rates of the non-transcribed (coding) strand,

$$\left\{\frac{f'^{\beta}_{g,T_i}}{E'(g,T_i,\beta)} : g \in G_{e,T_i}\right\}$$

and

$$\left\{\frac{f'^{\beta}_{g,T_i}}{E'(g,T_i,\beta)} : g \in G_{n,T_i}\right\}$$

were compared with a two-tailed Wilcoxon Rank-Sum test.

The percent depletion of mutations in expressed genes remaining after controlling for transcription coupled repair and noncoding mutation rates was computed by comparing the mutation rate of the not-transcribed strand of expressed genes and the mutation rate of the not-transcribed strand of not expressed genes, or:

$$100 * \left(1 - \frac{\sum_{g\in G_{e,T_i}} f'^{\beta}_{g,T_i}}{\sum_{g\in G_{e,T_i}} E'(g,T_i,\beta)} * \frac{\sum_{g\in G_{n,T_i}} E'(g,T_i,\beta)}{\sum_{g\in G_{n,T_i}} f'^{\beta}_{g,T_i}}\right)$$

For each observable transition β (e.g. G>A), where the starting base is defined as β^0 (or its complement β'^0), and the ending base as β^0 or its complement β'^0, the percent depletion of mutations in expressed genes was calculated in each strand. For the not-transcribed (coding) strand, this rate in tumor type T_i is:

$$100 * \left(1 - \frac{\sum_{g\in G_{e,T_i}} f'^{\beta}_{g,T_i}}{\sum_{g\in G_{e,T_i}} E'(g,T_i,\beta)} * \frac{\sum_{g\in G_{n,T_i}} E'(g,T_i,\beta)}{\sum_{g\in G_{n,T_i}} f'^{\beta}_{g,T_i}}\right)$$

While for the transcribed (template) strand, this rate in tumor type T_i is:

$$100 * \left(1 - \frac{\sum_{g\in G_{e,T_i}} f^{\beta}_{g,T_i}}{\sum_{g\in G_{e,T_i}} E(g,T_i,\beta)} * \frac{\sum_{g\in G_{n,T_i}} E(g,T_i,\beta)}{\sum_{g\in G_{n,T_i}} f^{\beta}_{g,T_i}}\right)$$

Finding conservative amino acid transitions from cancer mutation data

To determine the strength of selection on individual amino acid (a.a.) substitutions, a.a. substitution rates in lung adenocarcinomas from TCGA were examined. *T* is defined as the set of sequenced lung adenocarcinomas (see Tumor mutation data, above), and *G* is the set of sequenced genes.

Where

$$R(g,t) = \{\text{RNA sequencing counts (see RNA data)for gene } g \text{ in tumor } \ t \ \wedge g \in G \wedge t \in T\}$$

$$M(g,t) = \{\text{missense mutations in } g \text{ in } t\},$$

$$L_g = \text{length of g in amino acids}$$

define expressed genes as:

$$G_e = \{g : \ g \in G \wedge |\{t : \ R(g,t) \ > 8 \ \wedge t \in T \}| > 0.95 \ |T|\}$$

and not expressed genes as

$$G_n = \{g : \ g \in G \wedge \ |\{t : \ R(g,t) < 8 \ \wedge t \in T \}| > 0.95 \ |T|\}$$

Call

$$G_{e'} = G_e \setminus \left\{ g : g \in G \wedge \frac{\sum_{t \in T} |M(g,t)|}{L_g} \Big/ \frac{\sum_{t \in T} \sum_{g \in G} |M(g,t)|}{\sum_{g \in G} L_g} > 2 \right\}$$

Determine the matrix of transitions between each a.a. in expressed genes

$$S_{e',ij} \ = \sum_{g \in G_{e'}} \left| \bigcup_{t \in T} \{m : m \in M(g,t) \ \wedge \ \text{starting a.a. of } m = \text{a.a.}_i \ \wedge \text{ending a.a. of } m = \text{a.a.}_j\} \right|$$

and the matrix of transitions between each a.a. in not expressed genes

$$S_{n,ij} \ = \sum_{g \in G_n} \left| \bigcup_{t \in T} \{m : m \in M(g,t) \ \wedge \ \text{starting a.a. of } m = \text{a.a.}_i \ \wedge \text{ending a.a. of } m = \text{a.a.}_j\} \right|$$

$$(0 < i \leq j \leq 20)$$

Where

$$c(G, x) = \left| \bigcup_{g \in G} \{\text{codon } y : \text{codon } y \in g \ \wedge \ \text{codon } y = x\} \right|$$

And

$$S_p = \left\{ x : x \in \text{codons} \ \wedge x \in \text{codons that code for a.a.}_i \wedge x \in \left\{ \text{codons 1 missense from a.a.}_j \right\} \right\}$$

Compute the matrix of starting codon counts in expressed genes:

$$C_{e'}(i, j) = \sum_{x \in S_p} c(G_{e'}, x)$$

and compute the matrix of starting codon counts in not-expressed genes:

$$C_n\,(i, j) = \sum_{x \in S_p} c(G_n, x)$$

for $(0 < i \leq j \leq 20)$,

giving $S_{e'}$, S_n, $C_{e'}$, and C_n a size of 20 x 20.

Compute the average depletion of substitutions r such that

$$r = \frac{\sum S_n}{\sum C_n} * \frac{\sum C'_e}{\sum S'_e}$$

Use *r* to calculate an expected rate for each amino acid substitution in expressed genes, or

$$E_{e',ij} = \frac{S_{e',ij}\, r \ + \ S_{n,ij}}{r} \times \frac{C_{e',ij}}{C_{e',ij} + C_{n,ij}}$$

and in not-expressed genes

$$E_{n,ij} = \left(S_{e',ij}\, r \ + \ S_{n,ij}\right) \times \frac{C_{n,ij}}{C_{e',ij} + C_{n,ij}}$$

Using the expected and observed matrices, calculate a χ^2 statistic for each substitution, stored as matrix *X* such that

$$X_{ij} = \frac{\left(S_{e',ij} - E_{e',ij}\right)^2}{E_{e',ij}} + \frac{\left(S_{n,ij} - E_{n,ij}\right)^2}{E_{n,ij}}$$

Use *X* to compute p values with the χ^2 test with one degree of freedom, giving matrix *P*, where $P_{i,j} =$ the p value calculated from X_{ij}

Also calculate matrix *F* where

$$F_{ij} = \frac{S_{e',ij}}{C_{e',ij}} \times \frac{C_{n,ij}}{S_{n,ij}}$$

a.a.$_i$ and a.a.$_j$ are called substitutable if

$$F_{ij} \leq \frac{1}{r} \wedge P_{ij} \leq 0.251 \ \wedge F_{ji} \leq \frac{1}{r} \wedge P_{ji} \leq 0.251$$

Finding conservative amino acid transitions from BLOSUM

Substitutable amino acids from BLOSUM 90 were identified as pairs of amino acids with BLOSUM log-odds scores > 0. The significance of the overlap between substitutable amino acids identified from BLOSUM and those identified in tumors was calculated with the CDF of the hypergeometric distribution.

Identifying genes under purifying selection in multiple tumor types

To find genes under purifying selection in multiple tumor types, data from melanomas, lung adenocarcinomas, colorectal adenocarcinomas, liver hepatocellular carcinomas, gliomas, and breast

invasive carcinomas were used (forming set *T*). First, genes were only included in the analysis if they were called expressed in all tumor types, where

$$G_e = \{g: g \in G \ \wedge |\{T_i \in T : \ |\{t : t \in T_i \wedge R(g,t) > 8\}| > 0.95 * |T_i|\}| > |T|\}$$

An expected number of mutations was computed for each gene, or $E(g)$, based on each gene's non-coding / intron mutation rate in tumors subjected to whole-genome sequencing[5]:

$$E(g) = \sum_{T_i \in T} nm(g) * \frac{|G|}{\sum_{\gamma \in G} nm(\gamma)} * \frac{\sum_{\gamma \in G_{e,T_i}} m(\gamma, T_i)}{\sum_{\gamma \in G_{e,T_i}} L_\gamma} * L_g$$

Where $L_g =$ the length of gene g in amino acids, $nm(g)$ = the non-coding mutation rate for gene g calculated from published whole-genome sequencing of tumor samples [5], and

$$m(g, T_i) = \left| \left\{ \bigcup_{t \in T_i} \text{missense mutations in } g \text{ in } t \right\} \right|.$$

In this way, recurrent mutations (the same missense mutation observed more than once) within each tumor type were dropped from the analysis.

The expected number of mutations was compared to observed number of mutations, where

$$O(g) = \sum_{T_i \in T} m(g, T_i)$$

Genes were identified as under purifying selection (*N*) in these tumor types if they passed a p value and fold change cutoff:

$$N = \left\{ g: g \in G_e \ \wedge \frac{O(g)}{E(g)} < \frac{1}{2} \wedge \int_{x=0}^{x=O(g)} \frac{\left(E(g)\right)^x}{x!} e^{-\left(E(g)\right)} < 0.01 \right\}$$

Identifying gene sets under purifying selection in multiple tumor types

Gene sets were obtained from the Molecular Signature Database [61] version 5.1; sets examined were from the hallmark, canonical pathways, BioCarta, KEGG, Reactome, and GO subsets of the Molecular Signature Database, totaling 2834 sets, making $\bar{S}$, with set $S \in \bar{S}$. To find sets under purifying selection, mutations in these sets were examined in the melanoma, lung adenocarcinoma, and colorectal adenocarcinoma tumor types $\{T_i \in T\}$. For each tumor type, expressed genes were defined as genes

$$G_{e,T_i} = \left\{g : g \in G \wedge \left|\left\{t : R_{g,t} > 8 \wedge t \in T_i\right\}\right| > 0.95\,|T = T_i|\right\}$$

Sets were filtered so that they only contained genes with mutations in these tumors, so

$$\bigcup_{S \in \tilde{S}} (g \in S) \subseteq G$$

and so that

$$\tilde{S} = \left\{S : S \in \bar{S} \wedge 10 < |S| < 400 \wedge \forall\, T_i \in T\colon \frac{\left|\{g \in S \cap G_{e,T_i}\}\right|}{|S|} > 0.5\right\}$$

An expected number of mutations was computed for each set, where

$$E(S) = \sum_{g \in S} \sum_{T_i \in T} nm(g) * \frac{|G|}{\sum_{\gamma \in G} nm(\gamma)} * \frac{\sum_{\gamma \in G_{e,T_i}} m(\gamma, T_i)}{\sum_{\gamma \in G_{e,T_i}} L_\gamma} * L_g$$

Where $L_g =$ the length of gene g in amino acids, $nm(g)$ = the non-coding mutation rate for gene g calculated from published whole-genome sequencing of tumor samples [5], and

$$m(g, T_i) = \left|\left\{\bigcup_{t \in T_i} \text{missense mutations in } g \text{ in } t\right\}\right|$$

An observed number of mutations was also computed for each set, or $O(S)$, where

$$O(S) = \sum_{g \in S} \sum_{T_i \in T} m(g, T_i)$$

The difference between the observed and expected numbers of mutations for each set was determined through the CDF of the Poisson distribution, where

$$D(S) = \int_{x=0}^{x=O(S)} \frac{\left(E(S)\right)^x}{x!} e^{-\left(E(S)\right)}$$

To determine the significance of the depletion of mutations for each set, $1*10^4$ random sets (S_n^R) were generated for each set size, drawing from those genes in the union of all gene sets, so that

$$\forall S \in S_n^R \colon |S| = n \ \wedge |S_n^R| = 1 * 10^4 \wedge \bigcup_{S \in S_n^R} g \in S \subseteq \bigcup_{S \in \tilde{S}} g \in S$$

The significance of each set's depletion in mutation was evaluated by computing a p value, or $p(S)$, based on the depletion of random sets of the same size, so that.

$$p(S) = Pr\left(\left\{D(\sigma) : \ \sigma \in S_{|S|}^R\right\} \leq D(S)\right)$$

As many sets were depleted beyond even 10^4 random sets, an estimated p value was computed by regressing the randomly sampled sets. For each size set, the –log10 quantiles of those random sets with $p(S) < 0.01$ were fit with a linear regression vs the –log10(percentiles) that at which the quantiles were evaluated. This regression, generating for each set size slope *b* and constant *c*, was used to compute the revised p values, *P*(*S*), where

$$P(S) = b * p(S) + c$$

To correct for multiple hypothesis testing, a q-value was calculated using the method of Benjamini and Hochberg. [62]

Essentiality analysis of genes under purifying selection

The impact on cancer cell line growth of CRISPR-mediated knockout has been previously published [19-21]. In each of these three published pooled CRISPR screens, the investigators used differing methods to call whether a gene was essential. In each screen, a gene was called essential or not essential in each tested cell line. From this data, for each gene *g*, a score *C*(*g*) was recorded, or the proportion of tested cell lines in which this gene was deemed essential, based on the published results, from these three screens.

Gene sets under purifying selection across all tumor types were identified as above ("Identifying gene sets under purifying selection in all tumor types"). Sets under purifying selection (S_p) were defined to be sets with q-values < 0.05, and an observed / expected number of mutations <0.8. Genes under purifying selection (G_p) were defined as the union of genes in sets under purifying selection (S_p), or $G_p = \bigcup_{S \in S_p} g \in S$; $S_p \subset \tilde{S}$, where S̃ represents those sets examined from the Molecular Signature Database (see above). Genes under purifying selection (G_p) was then compared to genes not under purifying selection (G_{np}), where $G_{np} = \bigcup_{S \in S_{np}} g \in S$; $S_{np} = \tilde{S} \setminus S_p$.

The essentiality of genes under purifying selection (G_p) was compared to the essentiality of genes not under purifying selection (G_{np}); the essentiality of each gene was defined based on its score *C*(*g*), as defined above, representing the proportion of tested cell lines in which this gene was deemed essential. To calculate the significance of the difference in essentiality between these groups of genes,

$$\{C(g) : g \in G_p\} \text{ and } \{C(g) : g \in G_{np}\}$$

were compared with a two-tailed Wilcoxon Rank-Sum test.

To determine the utility of purifying selection for finding essential genes, a receiver-operator characteristic curve was generated using genes ranked by their revised P values (see above). Each gene was given the lowest revised P value of the gene sets examined in which it was part. Genes that were

called true positives (essential) were defined as genes that were deemed essential in $\geq$ 5 / 7 examined cell lines in a pooled CRIPSR screen[21].

Identifying genes under tumor type-specific purifying selection

First, genes were only included in this analysis if they were not called unexpressed in all tumor types, where

$$G_e = G/\{g: g \in G \ \wedge |\{T_i \in T: \ |\{t: t \in T_i \wedge R(g,t) < 8\}| > 0.95 * |T_i|\}| > |T|\}$$

Genes with an increased mutation rate across tumors were also filtered out. An expected number of mutations was computed for each gene, or *En*(g), based on each gene's non-coding / intron mutation rate in tumors subjected to whole-genome sequencing[5]:

$$E_n(g) = \sum_{T_i \in T} nm(g) * \frac{|G|}{\sum_{\gamma \in G} nm(\gamma)} * \frac{\sum_{\gamma \in G_{e,T_i}} m(\gamma, T_i)}{\sum_{\gamma \in G_{e,T_i}} L_\gamma} * L_g$$

Where $L_g =$ the length of gene g in amino acids, $nm(g)$ = the non-coding mutation rate for gene g calculated from published whole-genome sequencing of tumor samples [5], and

$$m(g, T_i) = \left| \left\{ \bigcup_{t \in T_i} \text{missense mutations in } g \text{ in } t \right\} \right|$$

To identify genes under negative selection in a particular tumor type relative to other tumors, a different expected value was computed based on the mutation rate in all other tumors.

For tumor type $T_i \in T$ where $T = \{T_1, T_2, \ldots T_{11}\}$ or all tumor types listed above in Tumor Mutation Data, so $T_i \cap T_j = \emptyset$; and for gene $g \in G$ where G = all sequenced genes \ O, where

$$\widetilde{M_g}(t,k) = \{\text{top } k \text{ genes ranked by } m(g,t)/L_g \mid T_i\}$$

$$0 = \bigcup_{t \in T_i} \{g_i = \widetilde{M_g}(T_i, 10)\} \cup \left\{g : g \in G_e : \frac{\sum_{T_i \in T} m(g,T_i)}{E_n(g)} > 2.5\right\}$$

Compute the expected number of missense mutations in *g* in T_i, or

$$E(g,T_i) = \sum_{g \in G} m(g,T_i) \times \left(\sum_{\tau \in (T \setminus T_i)} m(g,T_i) \Big/ \sum_{\gamma \in G} \sum_{\tau \in (T \setminus T_i)} m(\gamma,T_i) \right)$$

The calculated expected number of mutations for each gene was used to identify those genes under negative selection in one tumor type relative to the others (*N*). Genes were called as under negative selection if they passed a fold-change and P value (calculated with the Poisson distribution) cutoff:

$$N = \left\{ g: g \in G \wedge \frac{m(g,t)}{E_{g,t}} > 2 \wedge \int_{x=0}^{x=m_{g,t}} \frac{{E_{g,t}}^x}{x!} e^{-E_{g,t}} < 0.01 \right\}$$

Identifying pathways enriched in genes under purifying selection in specific tumor types

Pathways under purifying selection were identified from the list of genes under purifying selection generated after filtering out recurrent mutations as detailed above. The overlap between genes under purifying selection and a database of pathway gene sets (NDEx) [22] was evaluated with a CDF of the hypergeometric distribution.

Identifying gene sets under purifying selection in specific tumor types

Gene sets under purifying selection in specific tumor types were identified the same way as those gene sets under purifying selection in multiple tumor types (as detailed above), with the following differences. First, the expected number of mutations in each gene was estimated based on comparing one tumor type to other tumor types, as in "identifying genes under tumor type-specific selection," above, where

$$E(g,T_i) = \sum_{g \in G} M(g,T_i) \times \left(\sum_{\tau \in (T \setminus T_i)} M(g,T_i) \Big/ \sum_{\gamma \in G} \sum_{\tau \in (T \setminus T_i)} M(\gamma,T_i) \right)$$

$T =$ {melanoma, lung adenocarcinoma, and colorectal adenocarcinoma}

All other analysis of the depletion of mutations, gene set filtering, statistical and multiple hypothesis control was identical to "Identifying gene sets under purifying selection in multiple tumor types," above.

For melanomas, gene sets were called to be under purifying selection if they had a q-value < 0.05 and an observed / expected mutation ratio < 0.5.

For lung adenocarcinomas, gene sets were called to be under purifying selection if they had a q-value < 0.1 and an observed / expected mutation ratio < 0.55.

Estimating the impact of sequencing more tumors

To evaluate the number of additional hits (individual genes identified as under purifying selection) we might find with more sequenced tumors, we down-sampled mutations by steps equivalent to the mutations of 40 average tumors in each tumor type, with 1000 replicates per down-sampling. The sampling was started from the dropping mutations equal to 80 random tumors and continued until the first step before the average number of hits returned was ≤ 1. Down-sampled data were fit to a four-parameter logistic curve ($R^2 \geq 0.99$):

$$f(x) = A + \frac{(B - A)}{1 + 10^{(C-x)\times D}}$$

These fits were used to predict the number of new hits that would be found by steps of 10 additional sequenced tumors, and used to find an optimal distribution of sequenced tumors across tumor types to maximize the number of new hits.

Determining the fraction of essential genes under purifying selection

Genes under purifying selection across tumor types (G_p) were defined as above, the union of genes in sets under purifying selection. Genes under increased purifying selection in melanomas (G_p^M) were defined similarly as the union of genes in sets under increased purifying selection in melanomas. Sets under increased purifying selection in melanomas were defined, above, in "Identifying gene sets under purifying selection in specific tumor types." Gene sets were called to be under increased purifying selection in melanomas if they had a q-value < 0.05 and an observed / expected mutation ratio < 0.5.

Essential genes were identified from three CRISPR pooled screens [19-21], as discussed in "Essentiality analysis of genes under purifying selection," above. In each screen, a gene was called essential or not essential in each tested cell line. From this data, for each gene g, a score $C_i(g)$ was recorded for screen i, or the number of tested cell lines in which this gene was deemed essential, based on the published results, in each screen. For each screen i, a gene was deemed essential if $C_i(g) \geq$ the number of cell lines tested in screen i - 2. The set of genes deemed essential in each screen i was then termed G_{es}^i. G_{es}^i was also filtered so that it only included genes that were members of the sets in $\tilde{S}$ (the filtered gene sets from the Molecular Signature Database, see above), as those were the only genes that could be called under purifying selection.

The proportion of genes in G_{es}^i that were under selection (members of G_p or G_p^M) was then assessed.

Code Availability

The code used in these analyses is publically posted to GitHub, and can be accessed at https://github.com/rmathisWI/Purifying-Selection.

Acknowledgements

We would like to acknowledge Peter Reddien, Peter Sabatini, and Gerry Fink for their help and suggestions on this manuscript.

Individuals doing this work were funded by the National Science Foundation Graduate Research Fellowship Program (1122374; ES), and the National Institutes of Health (2T32GM007287-36, RM). The funding sources played no role in the design, collection, analysis, or interpretation of the data, nor played any role in the writing of the manuscript.

References

1. Nowell, P.C., *The clonal evolution of tumor cell populations.* Science, 1976. **194**(4260): p. 23-8.
2. Greaves, M. and C.C. Maley, *Clonal evolution in cancer.* Nature, 2012. **481**(7381): p. 306-13.
3. Nordling, C.O., *A new theory on cancer-inducing mechanism.* Br J Cancer, 1953. **7**(1): p. 68-72.
4. Lawrence, M.S., et al., *Discovery and saturation analysis of cancer genes across 21 tumour types.* Nature, 2014. **505**(7484): p. 495-501.
5. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-8.
6. Kimura, M. and T. Ohta, *On some principles governing molecular evolution.* Proc Natl Acad Sci U S A, 1974. **71**(7): p. 2848-52.
7. Kimura, M., *The neutral theory of molecular evolution: a review of recent evidence.* Jpn J Genet, 1991. **66**(4): p. 367-86.
8. Kiezun, A., et al., *Deleterious alleles in the human genome are on average younger than neutral alleles of the same frequency.* PLoS Genet, 2013. **9**(2): p. e1003301.

9. Zollner, S., et al., *Evidence for extensive transmission distortion in the human genome.* Am J Hum Genet, 2004. **74**(1): p. 62-72.
10. McFarland, C.D., L.A. Mirny, and K.S. Korolev, *Tug-of-war between driver and passenger mutations in cancer and other adaptive processes.* Proc Natl Acad Sci U S A, 2014. **111**(42): p. 15138-43.
11. McFarland, C.D., et al., *Impact of deleterious passenger mutations on cancer progression.* Proc Natl Acad Sci U S A, 2013. **110**(8): p. 2910-5.
12. Van den Eynden, J., S. Basu, and E. Larsson, *Somatic Mutation Patterns in Hemizygous Genomic Regions Unveil Purifying Selection during Tumor Evolution.* PLoS Genet, 2016. **12**(12): p. e1006506.
13. Weinstein, J.N., et al., *The Cancer Genome Atlas Pan-Cancer analysis project.* Nat Genet, 2013. **45**(10): p. 1113-20.
14. Rockah-Shmuel, L., A. Toth-Petroczy, and D.S. Tawfik, *Systematic Mapping of Protein Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly Neutral Mutations.* PLoS Comput Biol, 2015. **11**(8): p. e1004421.
15. Hanawalt, P.C. and G. Spivak, *Transcription-coupled DNA repair: two decades of progress and surprises.* Nat Rev Mol Cell Biol, 2008. **9**(12): p. 958-70.
16. Grantham, R., *Amino acid difference formula to help explain protein evolution.* Science, 1974. **185**(4154): p. 862-4.
17. Henikoff, S. and J.G. Henikoff, *Amino acid substitution matrices from protein blocks.* Proc Natl Acad Sci U S A, 1992. **89**(22): p. 10915-9.
18. Liberzon, A., et al., *Molecular signatures database (MSigDB) 3.0.* Bioinformatics, 2011. **27**(12): p. 1739-40.
19. Wang, T., et al., *Identification and characterization of essential genes in the human genome.* Science, 2015. **350**(6264): p. 1096-101.
20. Hart, T., et al., *High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities.* Cell, 2015. **163**(6): p. 1515-26.
21. Tzelepis, K., et al., *A CRISPR Dropout Screen Identifies Genetic Vulnerabilities and Therapeutic Targets in Acute Myeloid Leukemia.* Cell Rep, 2016. **17**(4): p. 1193-1205.
22. Pratt, D., et al., *NDEx, the Network Data Exchange.* Cell Syst, 2015. **1**(4): p. 302-305.
23. Yin, Y., et al., *Rapid induction of lung adenocarcinoma by fibroblast growth factor 9 signaling through FGF receptor 3.* Cancer Res, 2013. **73**(18): p. 5730-41.
24. Wang, Y., et al., *Discovery and identification of new non-ATP competitive FGFR1 inhibitors with therapeutic potential on non-small-cell lung cancer.* Cancer Lett, 2014. **344**(1): p. 82-9.
25. Tiseo, M., et al., *FGFR as potential target in the treatment of squamous non small cell lung cancer.* Cancer Treat Rev, 2015. **41**(6): p. 527-39.
26. Semrad, T.J. and P.C. Mack, *Fibroblast growth factor signaling in non-small-cell lung cancer.* Clin Lung Cancer, 2012. **13**(2): p. 90-5.
27. Liao, R.G., et al., *Inhibitor-sensitive FGFR2 and FGFR3 mutations in lung squamous cell carcinoma.* Cancer Res, 2013. **73**(16): p. 5195-205.
28. Wakasugi, M., et al., *Nucleotide excision repair-dependent DNA double-strand break formation and ATM signaling activation in mammalian quiescent cells.* J Biol Chem, 2014. **289**(41): p. 28730-7.
29. Kim, S.T., B. Xu, and M.B. Kastan, *Involvement of the cohesin protein, Smc1, in Atm-dependent and independent responses to DNA damage.* Genes Dev, 2002. **16**(5): p. 560-70.
30. Yazdi, P.T., et al., *SMC1 is a downstream effector in the ATM/NBS1 branch of the human S-phase checkpoint.* Genes Dev, 2002. **16**(5): p. 571-82.

31. Kong, X., et al., *Distinct functions of human cohesin-SA1 and cohesin-SA2 in double-strand break repair.* Mol Cell Biol, 2014. **34**(4): p. 685-98.
32. Reed, E., *Platinum-DNA adduct, nucleotide excision repair and platinum based anti-cancer chemotherapy.* Cancer Treat Rev, 1998. **24**(5): p. 331-44.
33. Helleday, T., *Homologous recombination in cancer development, treatment and development of drug resistance.* Carcinogenesis, 2010. **31**(6): p. 955-60.
34. Begg, A.C., F.A. Stewart, and C. Vens, *Strategies to improve radiotherapy with targeted drugs.* Nat Rev Cancer, 2011. **11**(4): p. 239-53.
35. Dai, C.H., et al., *RNA interferences targeting the Fanconi anemia/BRCA pathway upstream genes reverse cisplatin resistance in drug-resistant lung cancer cells.* J Biomed Sci, 2015. **22**: p. 77.
36. Pennington, K.P., et al., *Germline and somatic mutations in homologous recombination genes predict platinum response and survival in ovarian, fallopian tube, and peritoneal carcinomas.* Clin Cancer Res, 2014. **20**(3): p. 764-75.
37. Boone, B., et al., *EGFR in melanoma: clinical significance and potential therapeutic target.* J Cutan Pathol, 2011. **38**(6): p. 492-502.
38. Gross, A., et al., *Expression and activity of EGFR in human cutaneous melanoma cell lines and influence of vemurafenib on the EGFR pathway.* Target Oncol, 2015. **10**(1): p. 77-84.
39. Jalili, A., et al., *NVP-LDE225, a potent and selective SMOOTHENED antagonist reduces melanoma growth in vitro and in vivo.* PLoS One, 2013. **8**(7): p. e69064.
40. Rubinfeld, B., et al., *Stabilization of beta-catenin by genetic defects in melanoma cell lines.* Science, 1997. **275**(5307): p. 1790-2.
41. Selimovic, D., et al., *Bortezomib/proteasome inhibitor triggers both apoptosis and autophagy-dependent pathways in melanoma cells.* Cell Signal, 2013. **25**(1): p. 308-18.
42. Ueda, Y. and A. Richmond, *NF-kappaB activation in melanoma.* Pigment Cell Res, 2006. **19**(2): p. 112-24.
43. Webster, M.R. and A.T. Weeraratna, *A Wnt-er migration: the confusing role of beta-catenin in melanoma metastasis.* Sci Signal, 2013. **6**(268): p. pe11.
44. Yaguchi, T., et al., *Immune suppression and resistance mediated by constitutive activation of Wnt/beta-catenin signaling in human melanoma cells.* J Immunol, 2012. **189**(5): p. 2110-7.
45. Wang, Y., et al., *Androgen receptor promotes melanoma metastasis via altering the miRNA-539-3p/USP13/MITF/AXL signals.* Oncogene, 2016.
46. Mavrakis, K.J., et al., *Disordered methionine metabolism in MTAP/CDKN2A-deleted cancers leads to dependence on PRMT5.* Science, 2016. **351**(6278): p. 1208-13.
47. Kryukov, G.V., et al., *MTAP deletion confers enhanced dependency on the PRMT5 arginine methyltransferase in cancer cells.* Science, 2016. **351**(6278): p. 1214-8.
48. Ostrow, S.L., et al., *Cancer evolution is associated with pervasive positive selection on globally expressed genes.* PLoS Genet, 2014. **10**(3): p. e1004239.
49. Sohail, M., et al., *Negative selection in humans and fruit flies involves synergistic epistasis.* Science, 2017. **356**(6337): p. 539-542.
50. Kryazhimskiy, S. and J.B. Plotkin, *The population genetics of dN/dS.* PLoS Genet, 2008. **4**(12): p. e1000304.
51. Shah, K., et al., *Synonymous codon usage affects the expression of wild type and F508del CFTR.* J Mol Biol, 2015. **427**(6 Pt B): p. 1464-79.
52. Quax, T.E., et al., *Codon Bias as a Means to Fine-Tune Gene Expression.* Mol Cell, 2015. **59**(2): p. 149-61.
53. Mauro, V.P. and S.A. Chappell, *A critical analysis of codon optimization in human therapeutics.* Trends Mol Med, 2014. **20**(11): p. 604-13.

54. Resch, A.M., et al., *Widespread positive selection in synonymous sites of mammalian genes.* Mol Biol Evol, 2007. **24**(8): p. 1821-31.
55. Chamary, J.V. and L.D. Hurst, *Evidence for selection on synonymous mutations affecting stability of mRNA secondary structure in mammals.* Genome Biol, 2005. **6**(9): p. R75.
56. Presnyak, V., et al., *Codon optimality is a major determinant of mRNA stability.* Cell, 2015. **160**(6): p. 1111-24.
57. Kudla, G., et al., *Coding-sequence determinants of gene expression in Escherichia coli.* Science, 2009. **324**(5924): p. 255-8.
58. Lazrak, A., et al., *The silent codon change I507-ATC->ATT contributes to the severity of the DeltaF508 CFTR channel dysfunction.* Faseb j, 2013. **27**(11): p. 4630-45.
59. Center, B.I.T.G.D.A., *Analysis-ready standardized TCGA data from Broad GDAC Firehose 2016_01_28 run.* Broad Institute of MIT and Harvard, 2016.
60. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC Bioinformatics, 2011. **12**: p. 323.
61. Subramanian, A., et al., *Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.* Proc Natl Acad Sci U S A, 2005. **102**(43): p. 15545-50.
62. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the royal statistical society. Series B (Methodological), 1995: p. 289-300.

Tables

Gene Set Size	Overlap Size	Hypergeometric P	Gene Set Names
57	8	2.94E-05	PID_FAK_PATHWAY
31	5	0.000202	PID_NCADHERINPATHWAY
40	5	0.00085	PID_ERBB2ERBB3PATHWAY
30	4	0.00142	PID_ERBB1_RECEPTOR_PROXIMAL_PATHWAY
44	5	0.001422	PID_RHOA_PATHWAY
64	6	0.002165	PID_CDC42_PATHWAY
48	5	0.002247	PID_ARF6_TRAFFICKINGPATHWAY
48	5	0.002247	PID_CERAMIDE_PATHWAY
49	5	0.0025	PID_ANGIOPOIETINRECEPTOR_PATHWAY
21	3	0.002598	PID_HEDGEHOG_2PATHWAY
35	4	0.002879	PID_ERBB1_INTERNALIZATION_PATHWAY

Table 1. Pathways enriched for genes under increased purifying selection in lung adenocarcinomas. Known pathways evaluated for enrichment of genes identified as under increased purifying selection in lung adenocarcinomas, relative to other tumors.

Gene Set Size	Overlap Size	Hypergeometric P	Gene Set Names
28	8	3.48E-06	PID_BARD1PATHWAY
45	10	5.75E-06	PID_FANCONI_PATHWAY
34	8	2.04E-05	PID_ATM_PATHWAY
29	7	4.28E-05	PID_CDC42_REG_PATHWAY
43	7	0.000805	PID_PS1PATHWAY
45	7	0.0011	PID_RHOA_REG_PATHWAY
38	6	0.001803	PID_ATR_PATHWAY
39	6	0.00211	PID_FOXM1PATHWAY
40	6	0.002455	PID_ERBB2ERBB3PATHWAY
51	7	0.002533	PID_CASPASE_PATHWAY
31	5	0.002933	PID_NCADHERINPATHWAY
22	4	0.003144	PID_NFKAPPABCANONICALPATHWAY
43	6	0.003753	PID_PLK1_PATHWAY
67	8	0.004139	PID_TELOMERASEPATHWAY
35	5	0.005508	PID_PI3KCIAKTPATHWAY

Table 2. Pathways enriched for genes under increased purifying selection in melanomas. Known pathways evaluated for enrichment of genes identified as under increased purifying selection in lung adenocarcinomas, relative to other tumors.

Figures

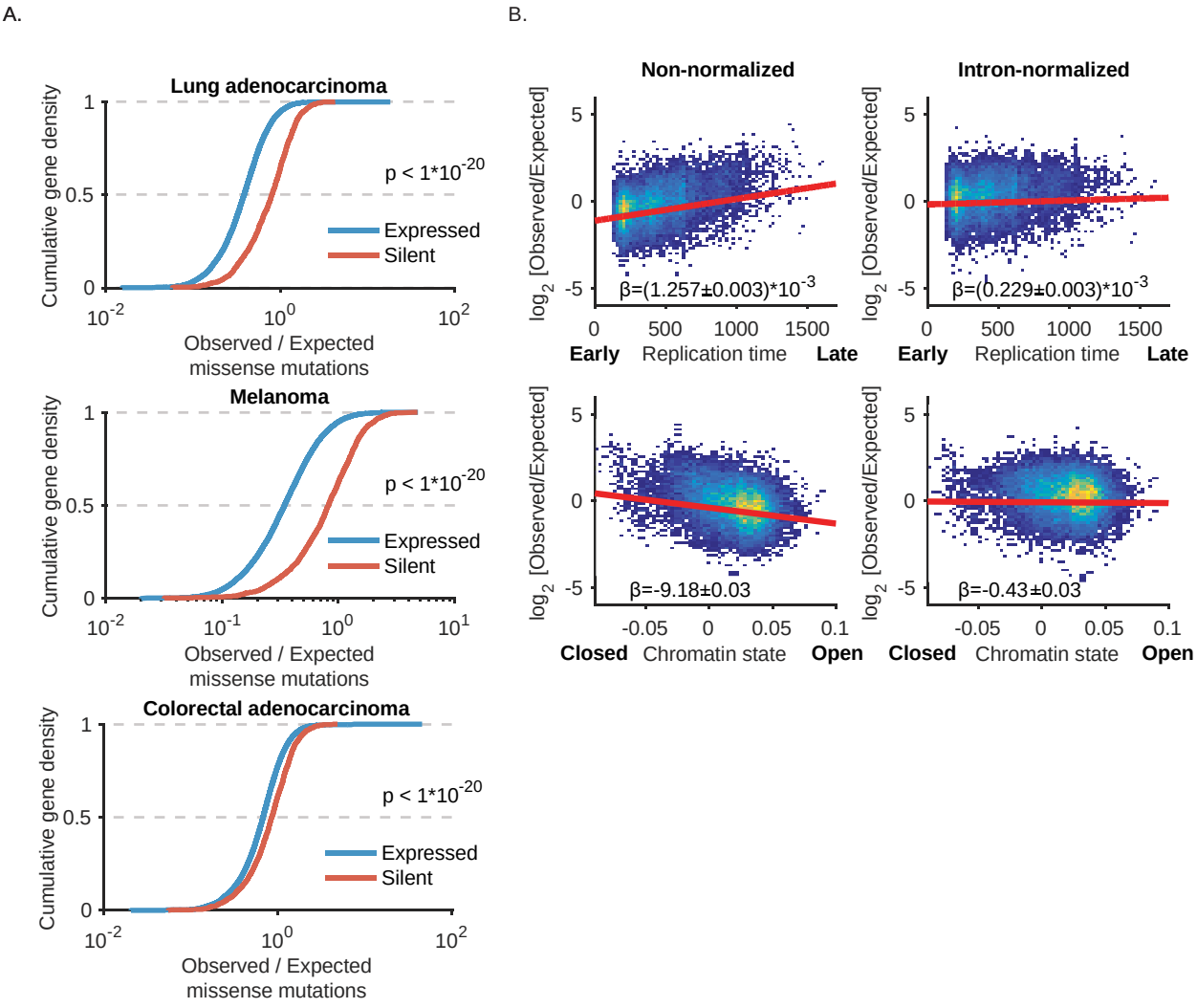


Figure 1. Genes that are expressed have fewer missense mutations (substitutions). (a) Cumulative distribution of observed missense mutations / expected (intron normalized) in expressed vs. silent genes in melanomas (n = 290), lung adenocarcinomas (n = 533) and colorectal adenocarcinomas (n = 489). For each tumor type, expressed genes were defined as having an estimated transcript count > 8 in ≥ 95% of tumors. Statistical significance was assessed using the Wilcoxon rank-sum test. **(b)** Expected mutation rates from intron mutations controls for various mutation covariates in lung adenocarcinomas, including %GC, replication timing, and chromatin accessibility. A linear regression is plotted of $\log_2$ observed over expected mutations for non-normalized and intron-mutation rate normalized based expected. The slope of the regression (β) is displayed with the 95% confidence interval. Color corresponds to density of genes in the scatter plots.

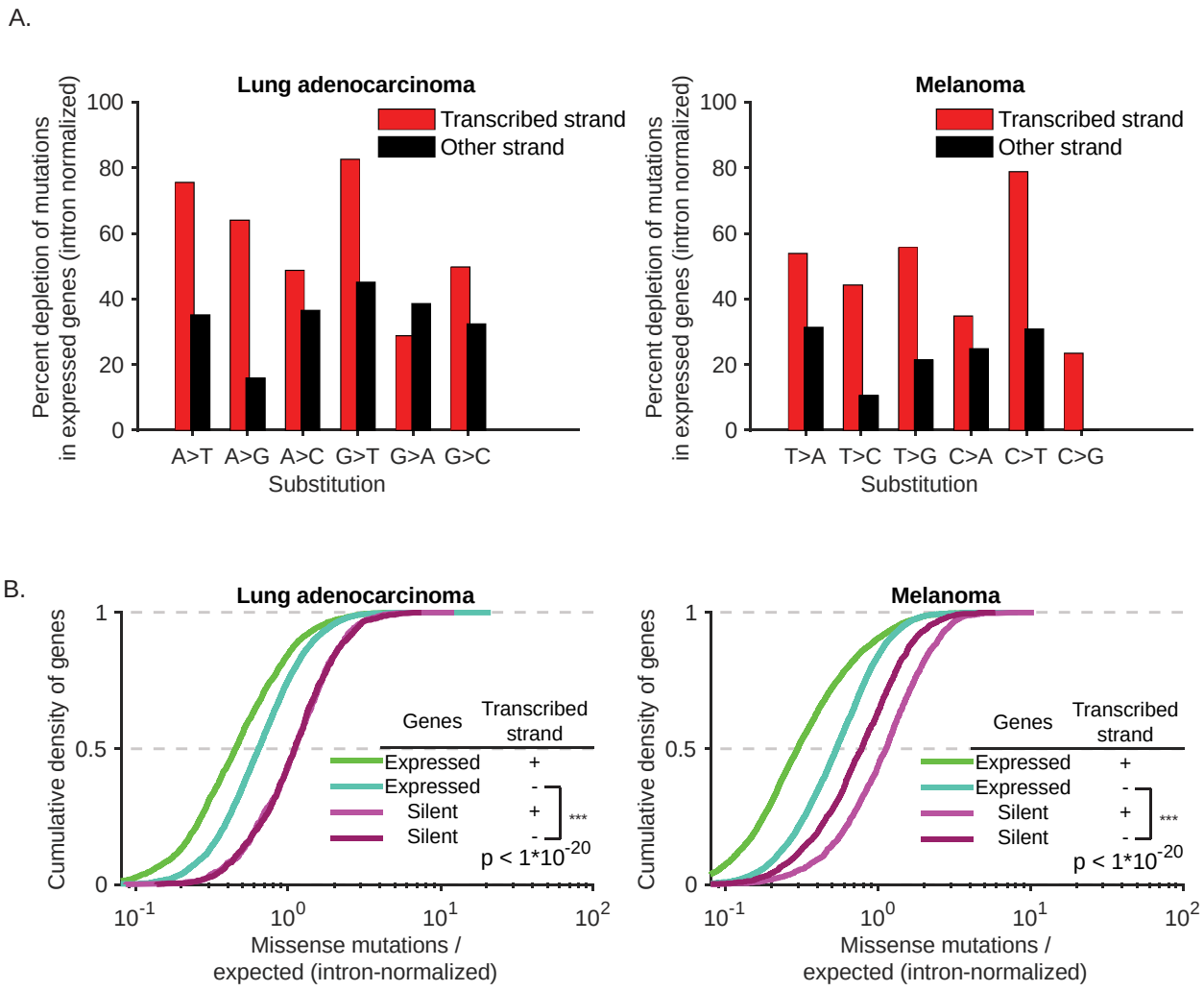


Figure 2. Strand bias in mutations. (a) Plots show percent depletion of missense mutations of each class of substitution in expressed vs silent genes, in both the transcribed (template) and not-transcribed (coding) strands. Gene mutation rates were normalized by each gene's relative intron mutation rate. **(b)** Cumulative distribution of observed / expected (intron-normalized) mutations in the transcribed and not-transcribed strands of expressed and silent genes. Plots represent G>T substitution rates in lung adenocarcinomas and C>T substitution rates in melanomas. Statistical significance was assessed using the Wilcoxon rank-sum test.

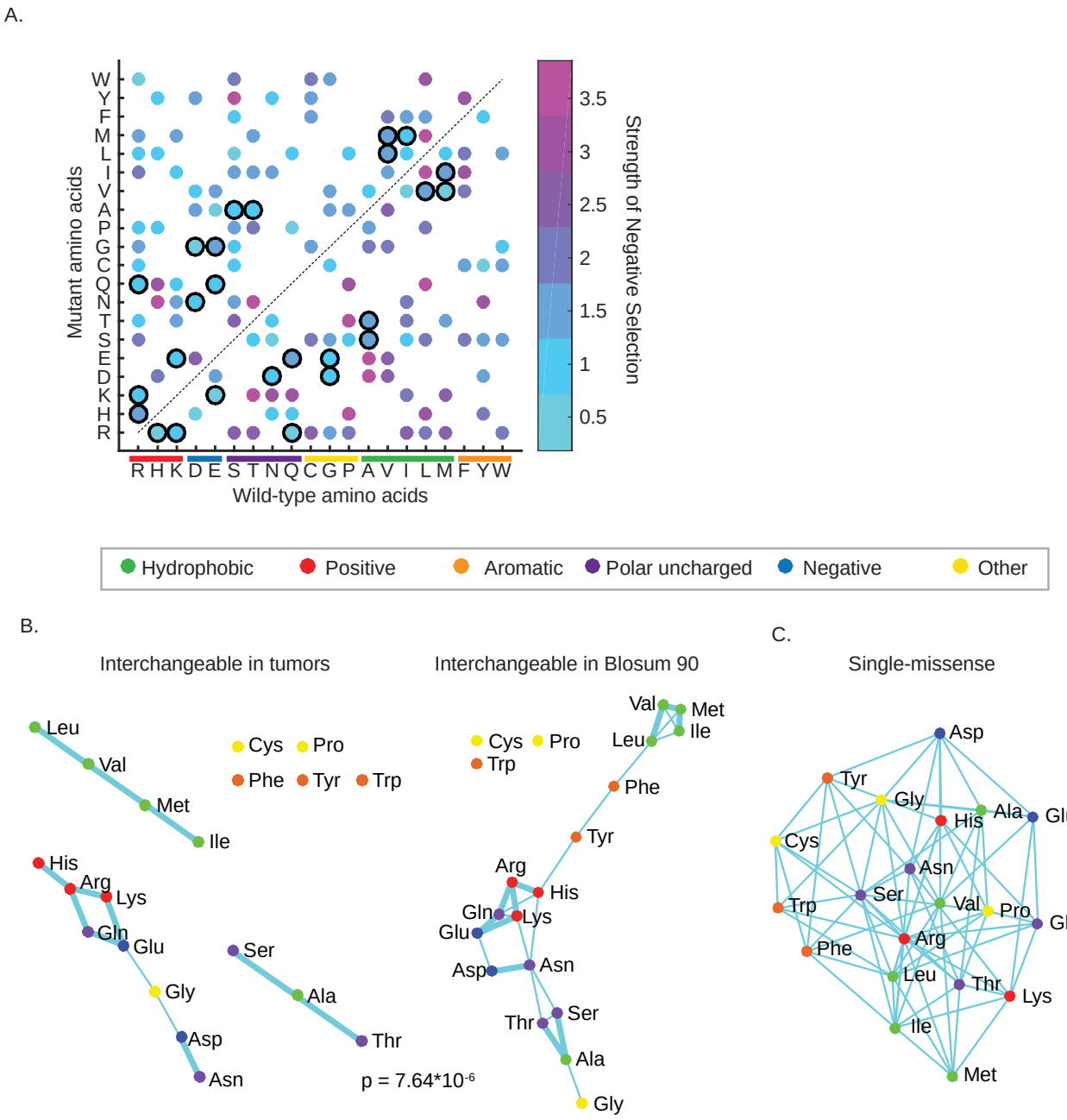
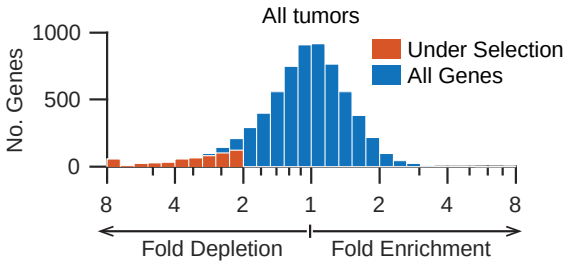


Figure 3. Amino acid substitutions with similar physicochemical traits are more acceptable during both tumor microevolution and species macroevolution. (a) Heat map showing the strength of negative selection on each observed pairwise amino acid substitution. Bold outlines highlight substitutions between interchangeable amino acids. Negative selection strength was quantified as described in methods. **(b)** Graph depicting amino acids interchangeable, color-coded based on their chemical properties. Edges shared between interchangeable amino acids identified from tumors (left) and organisms (right) are thickened; the overlap has a p-value of $7.64*10^{-6}$, as determined from the hypergeometric distribution. Amino acids interchangeable in organisms were defined as substitutions with a BLOSUM90 score ≥ 0. **(c)** Graph depicting all amino acid substitutions that are possible with a single missense mutation (center).

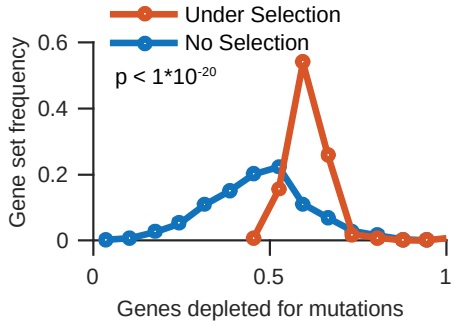
A.



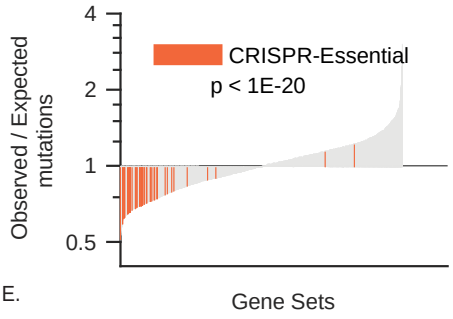
B.

Set Name	P value	Depletion of mutations	% Genes with reduced mutations
Reactome CDT1 association with the CDC6 ORC origin complex	$5.4*10^{-5}$	42.89%	65.96%
DNA polymerase activity	$8.1*10^{-5}$	38.65%	69.23%
Microtubule organizing center part	$2.4*10^{-4}$	37.85%	62.50%
Reactome assembly of the pre-replicative complex	$5.4*10^{-5}$	37.47%	61.82%
Nucleotidyltransferase activity	$5.7*10^{-6}$	37.21%	67.57%
Reactome M G1 transition	$6.8*10^{-6}$	36.14%	60.00%
Reactome transport of ribonucleoproteins into the host nucleus	$1.8*10^{-5}$	35.28%	65.38%
Reactome NEP NS2 interacts with the cellular export machinery	$2.1*10^{-5}$	34.95%	61.54%
Reactome metabolism of non-coding RNA	$5.7*10^{-6}$	34.78%	62.79%
Reactome influenza life cycle	$4.3*10^{-8}$	34.63%	65.87%
Reactome transport of mature mRNA derived from an intronless transcript	$1.1*10^{-5}$	34.56%	61.29%
Nuclear pore	$6.1*10^{-5}$	34.45%	63.33%
Reactome interactions of VPR with host cellular proteins	$2.6*10^{-5}$	34.42%	60.00%
Hallmark oxidative phosphorylation	$9.2*10^{-9}$	33.05%	62.37%
Reactome nonsense mediated decay enhanced by the exon junction complex	$1.4*10^{-4}$	33.01%	64.89%
Reactome metabolism of RNA	$1.7*10^{-12}$	32.96%	63.09%
Reactome transport of mature transcript to cytoplasm	$1.6*10^{-5}$	32.70%	61.90%
Reactome late phase of HIV life cycle	$9.1*10^{-8}$	32.32%	59.09%
Mitochondrial membrane	$1.9*10^{-4}$	32.25%	59.21%
Reactome metabolism of mRNA	$1.3*10^{-8}$	31.69%	63.73%

C.



D.



E.

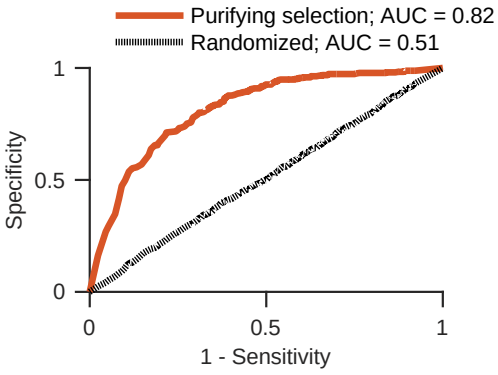


Figure 4. Purifying selection targets genes that are important for tumor growth. (a) Histogram showing the number of genes depleted for substitutions across all tumors (vs expectation from intron mutations). Genes with significant depletion (p<0.01 and >2-fold) are shown in red, and all genes in blue; all genes shown have >10 expected mutations and are expressed in all tumors. **(b)** Shown are the top 20 significant (Q<0.01) gene sets ranked by the depletion of mutations vs expected. P values were determined through sampling (see Methods). "% Genes with reduced mutations" represents the proportion of genes within each set with fewer mutations than expected. **(c)** Most genes in gene sets under purifying selection have fewer mutations than expected. Statistical significance was assessed using the Wilcoxon rank-sum test. **(d)** Most essential genes sets are under purifying selection. Gene sets were identified as essential if at least 60% of genes in the set were deemed essential in at least 60% of cell lines, based on 3 published pooled CRISPR screens. Statistical significance was assessed using the Wilcoxon rank-sum test. The background distribution of sets ordered by expected / observed mutations is in grey. **(e)** Receiver-operator characteristic (ROC) curves showing the predictive value of purifying selection, based on genes' estimated p values, for identifying essential genes. Also shown are random

genes. Essential genes were identified in a published pooled CRISPR screen in cancer cell lines. AUC, area under the curve.

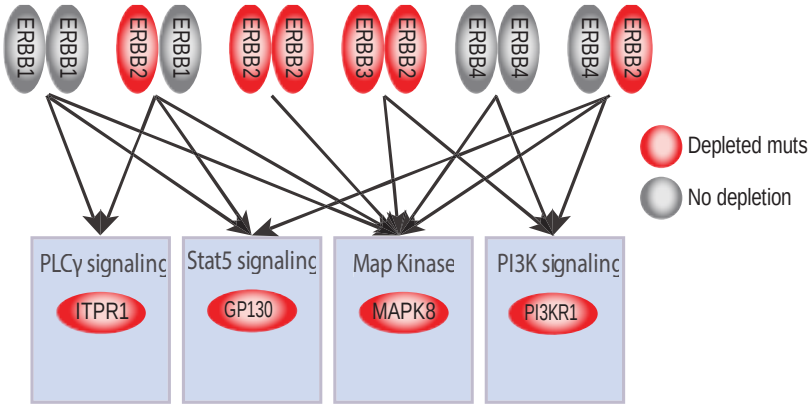


Figure 5. Genes under increased selection in Lung Adenocarcinomas. EGFR/ERBB signaling pathways highlighting genes (red) that are targets of purifying selection in lung adenocarcinomas ($p < 0.02$ with >2-fold depletion).

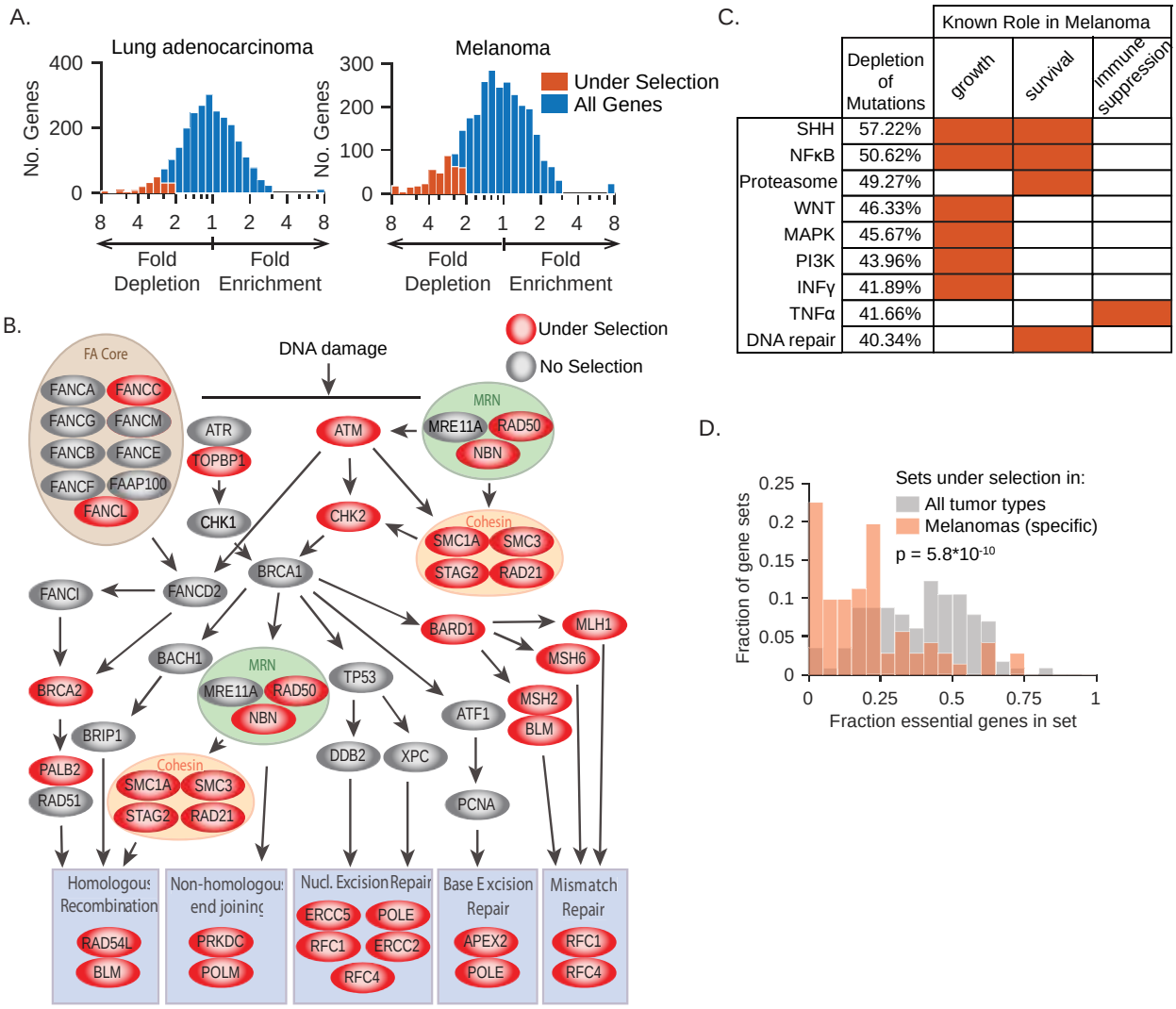


Figure 6. Purifying selection reveals tumor type-specific vulnerabilities (a) Histograms showing the number of genes depleted for substitutions across tumors of the indicated type relative to other tumor types. Genes with significant depletions ($p<0.02$ and >2-fold) are shown in red, and all genes in blue; all genes shown have at least 10 expected mutations and are expressed in all tumors. **(b)** Shown are DNA repair pathways, highlighting genes (red) that are targets of increased purifying selection in melanomas ($p < 0.02$ with >2-fold depletion). **(c)** Pathways under more purifying selection in melanomas are known to be important for growth, survival, immune suppression, or metastasis in melanomas. Shown are pathways identified from gene sets with increased purifying selection in melanomas relative to other tumor types ($Q < 0.05$ and > 50% depletion of mutations; pathways shown represent 519 / 882 genes under selection). "Depletion of mutations" represents the average percent depletion of mutations in genes in sets corresponding to each pathway. The depletion in melanomas is calculated relative to the expected number of mutations, based on the number of mutations observed in other tumor types. **(d)** Sets under increased purifying selection in melanomas have fewer essential genes compared to sets under purifying selection across all tumor types. Each histogram shows the proportion of sets under selection binned by the fraction of essential genes in each set. Genes essential in cancer cell lines were identified from published pooled CRISPR screens. The p value is calculated with a rank-sum test.

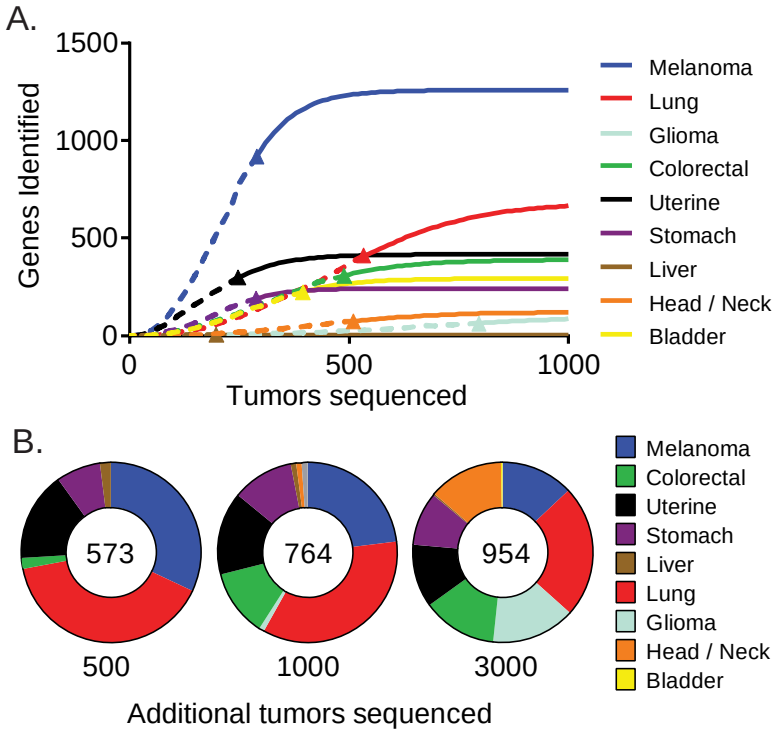


Figure 7. Estimating the effect of sequencing additional tumors. (a) The number of additional genes we expect to find under purifying selection if more tumors are sequenced was determined through fitting a curve to the number of genes found to be under purifying selection after down-sampling mutations. Shown in dotted lines are the hits found after down-sampling mutations (average of 1000 replicates). The predicted number of new hits from a fit to the down sampled data is shown in a solid line. A triangle indicates the observed number of hits and tumors sequenced. **(b)** Using the data from (a), we estimated the optimal set of tumors to sequence to find more genes under negative selection; the number of additional tumors sequenced is shown below each circle, and inside each circle is the number of new genes found.

Chapter 3: Cancer cells exhibit clonal diversity in phenotypic plasticity

The material presented in this chapter was adapted, with permission, from the following publication:

Mathis RA, Sokol ES, Gupta PB (2017). Cancer cells exhibit clonal diversity in phenotypic plasticity. Open Biology. doi: 10.1098/rsob.160283.

Authors' contributions: RAM performed the experiments and statistical analysis. RAM and PBG designed the experiments. ESS and RAM wrote the simulations. PBG, RAM and ESS wrote the manuscript.

Chapter 3

Introduction

The diversity of cancer cell phenotypes within individual tumors plays a major role in driving both drug resistance and tumor progression [1, 2]. For decades, the prevailing view has been that phenotypic diversity arises because tumors are mixtures of cancer cell clones with distinct yet heritable phenotypes. In this neo-Darwinian model, cancer cell phenotypes are genetically encoded and thus stably propagated to daughter cells [3-6]. In support of this model, there are significant genetic differences between different sections of a tumor, and even across different cells from the same tumors [5-9].

Phenotypic heterogeneity has been documented in breast tumors and breast cancer cell lines [10, 11]. Several recent reports have suggested that there are bi-directional transitions between cancer cells in distinct phenotypic states for various kinds of cancers [12-22]. For example, breast cancer cells in culture transition between mesenchymal (stem-like) and epithelial (differentiated) states [12-14, 16, 17, 23]. Analyses of cells within patient tumors also suggest that they transition between phenotypic states [18, 24]. In any population, random transitions of cells between phenotypic states will give rise to a stable equilibrium in which the different phenotypic states are represented at fixed proportions [12].

Since phenotypic plasticity has primarily been examined in populations of cancer cells, it is currently not known if this trait varies across the different cancer cell clones within a single population. Genetic analyses of phenotype states sorted from tumors and cell lines have led to conflicting conclusions regarding the contribution of genetic mutations to phenotypic plasticity [24-27]. While some studies have confirmed clonal relationships between states, a key question that remains open is if phenotypic plasticity can vary across the clones in a single cancer cell population [28].

Resolving this question— whether the clonal diversity of cancers influences their phenotypic plasticity— is fundamental to understanding cancer, and is also important from the perspective of developing combination cancer therapies. In particular, optimal combination chemotherapy designs will depend on whether the clones in a tumor have different capacities to transition between drug-sensitive and -resistant states.

Examining this question would require an experimental approach that can quantify phenotypic plasticity in hundreds of individual clones within a population of cancer cells. DNA barcodes combined with high-throughput sequencing have proven effective for tracking large numbers of clones in both normal and cancer cell populations [28-30]. Here, we apply DNA barcodes to quantify the extent to which phenotypic plasticity varies across hundreds of clones within in a single population of cancer cells.

Results

Labeling of Cancer Cell Clones with DNA Barcodes

To track the progeny of single cancer cells, we used retroviruses to stably introduce a random DNA sequence (or barcode) into their genome. These barcodes were introduced into $1*10^4$ MDA-MB-157 cells at a low multiplicity of infection (0.13), which we expected would label ~1300 individual clones (Figure 1A). After a brief drug selection for the infected cells, the barcoded clones were expanded in culture over a span of several months (Figure 2). Since the retrovirus pool contains ~$2.6*10^6$ random barcodes (Figure 1B), and only ~1300 cells were infected, there was a probability of 0.31 that more than one cell was independently infected with the same barcode, and a probability of $6.1*10^{-3}$ that 4 or more cells shared a barcode with other cells (Figure 1C). Accordingly, the number of copies of a given barcode sequence in the genomic DNA is directly proportional to the size of the corresponding clone in the population.

High-throughput sequencing of the barcodes from the pool of clones revealed that the barcodes were well-separated in DNA sequence space, with an average pair-wise Hamming distance of 10.5 base pairs. This is consistent with what one would expect if 1372 DNA sequences of length fourteen were randomly sampled from a space of 300 million possible sequences. Since the barcodes were well-separated in sequence space, it was straightforward to map reads to barcodes even in cases where point mutations arose through sequencing, consistent with the findings of others [28, 31, 32]; such reads were an average of 1.7 base pairs from their parent barcodes.

Clones Have Heterogeneous Phenotypic Ratios

We chose to barcode the MDA-MB-157 cell line because this line contains both epithelial and mesenchymal phenotypic states that can be robustly separated by fluorescence activated cell sorting (FACS). Using an antibody that recognizes Keratins 8 and 18, intracellular antigens which mark luminal epithelial cells in the mammary gland [33], we were able to separate the cells into Keratin 8/18 high or low fractions (Figure 3A, B). Importantly, this population of cells also contained roughly equal amounts of the two phenotypic states, with about 40% mesenchymal cells. Having a large minor population meant we were confident we could accurately detect clones with small amounts of progeny in the minor state.

To assess the proportion of cells with epithelial or mesenchymal phenotypes within each clone, we separated the barcoded population into epithelial and mesenchymal fractions with FACS (Figure 2). To assess clonal dynamics, the same population of cells was sampled once weekly for a total of three time points, each time separating these phenotypes (Figure 4A). After sorting, each population was further divided into equal halves before extracting and sequencing its DNA. Using high-throughput sequencing, we quantified the proportion of cells with epithelial and mesenchymal phenotypes for each of the 1372 barcoded clones in the population.

To estimate the magnitude of the technical error associated with sample preparation, sequencing, and analysis, we compared estimated clone sizes between each of the two sequenced partitions of these six populations (two sorted populations at three time points). The clone size estimated for each barcode was highly reproducible between these technical replicates, with an average Pearson correlation of 0.9119 across the 1372 clones detected (Figure 1D). These observations indicated that this experimental approach reproducibly quantified the numbers of cells corresponding to barcoded clones within the population.

Ordering clones by their fraction of epithelial cells revealed that the majority of clones produced progeny that were mixtures of cancer cells in the two phenotypic states (Figure 4B), with only 11% of clones consisting of only one lineage (statistically indistinguishable from mis-sorted cells). Although most clones exhibited such phenotypic plasticity, the ratio of epithelial to mesenchymal phenotypes varied significantly between clones (Shannon entropy = 3.5).

From the sequencing data we were able to distinguish three distinct classes of clones: clones with mostly epithelial cells, clones with mostly mesenchymal cells, and clones that were a mixture of cells in these two phenotypic states (Figure 4C). The majority of clones (89%) in the population gave rise to daughter cells in both the mesenchymal and epithelial states, with slightly more than half of the clones (64%) having a mesenchymal bias. We observed that the distribution of epithelial to mesenchymal ratios across clones was closely approximated by a log-normal distribution (Figure 4D, E), for all three time points. While most clones were comprised of both epithelial and mesenchymal cells, the proportion of progeny in these two states varied greatly between clones: 93% of clones had a bias significantly different from the bulk population proportions of the two states.

Although most clones had a mesenchymal bias, epithelial-biased clones tended to be larger (Figure 5A), resulting in approximately 60% of the population of cells being epithelial. Despite this

difference, there was only a weak correlation (0.06) between a clone's growth rate and Log2[E/M] ratio, although we found the differences in clones' growth rates to be stable across the time course (Figure 5B-D).

Phenotypic Ratios are Stably Inherited by Clonal Progeny

Although phenotypic ratios varied significantly across clones, they were highly stable for any given clone during the two weeks in culture, with an average Pearson correlation of 0.89 ($R^2 = 0.79$, all $p < 1*10^{-6}$) (Figure 4F). Additionally, 81% of clones' fraction epithelial differed by less than 0.15 over two weeks. This raised the possibility that the epithelial to mesenchymal ratio could be a quantitatively-inherited phenotype. To quantify the narrow-sense heritability of this trait, we generated 28 clonal subpopulations from individual cells expanded in culture over a span of six weeks. After expanding these clonal subpopulations, we used Sanger sequencing to determine their DNA barcodes. We also used flow cytometry to determine the phenotypic ratio in each of the cloned subpopulations, and compared this to the phenotypic ratio of the same clone in the parental pooled population (Figure 6, Figure 7A, Table 1). Regression analysis of these comparisons indicated that phenotypic plasticity was a highly heritable trait ($\rho = 0.89$) (Figure 7B). This heritability was considerably higher than that observed with 10^6 datasets with permutated barcode labels that randomized the relationship between the parental and cloned populations (Figure 7C). This finding indicated that phenotypic plasticities were stably inherited even through the rigors of single-cell cloning.

Phenotypic Plasticity Varies Across Clones in Primary Tumors

To assess if phenotypic plasticity varies across clones in patient tumors, we analyzed data from a recently published study that performed RNA-seq on single primary tumor cells [18]. By identifying clonal relationships between cells, and determining cell states, we could test whether these clones also had different cell state proportions. To identify clonal relationships, we looked for chromosomal gains

and losses using a sliding average of gene expression moving across chromosomes, modified from published methods [18, 34]. Although this limited resolution of genetic aberrations means we are likely missing genetic differences between clustered cells, we are confident the gains and losses of whole chromosomes reveal distinct clones. This analysis revealed a common gain in chromosome 7 and loss of chromosome 10 across tumor cells, aberrations commonly found in glioblastoma [35], as well as other changes, such as a gain of chromosome 5 or a loss of chromosome 14 or 13, that were only present in some cells (Figure 8). Hierarchical clustering grouped single cells into clones based on these inferred chromosomal gains and losses, resulting in four major clones (Figure 4).

We used the same single-cell RNA sequencing data to assign cells to cell states, using a published method based on the mean expression of gene sets defining different glioblastoma subtypes [18, 35], and increased "stem-ness" [18] (Figure 8). This analysis revealed that while each clone contained cells representing different glioblastoma subtypes, there were significant differences in subtype scores between clones, particularly of the mesenchymal subtype ($p < 6*10^{-5}$) and a stem-like state ($p < 8.0*10^{-4}$). Although this snapshot in time cannot tell us about the stability of these differences, this result suggests that clones within primary tumors have different cell-state proportions, consistent with our previous observations.

Combination Chemotherapies Enrich for Clones with Increased Phenotypic Plasticity

While there is significant interest in developing combination therapies that incorporate agents which selectively target the epithelial and mesenchymal states, the optimal design of such therapies is likely to depend on the mechanisms that give rise to phenotypic diversity in tumors. We therefore used computational simulations to model how such chemotherapies would affect tumors that are heterogeneous mixtures of clones with different phenotypic plasticities. These tumors were simulated to match the plasticities, sizes, and growth rates of the observed clones.

As expected, treatment with an epithelial-specific chemotherapy enriched for the more mesenchymal clones, whereas treatment with a mesenchymal-specific chemotherapy enriched for the more epithelial clones; both treatments selected for clones with reduced phenotypic plasticity. In contrast, a combination therapy that sequentially applied the epithelial- and mesenchymal-specific treatments enriched for clones with increased phenotypic plasticity, with a maximal enrichment for clones that were equal mixtures of cells in the epithelial and mesenchymal states (Figure 9A). In addition to selecting for clones with increased plasticity, this combination therapy was also significantly more effective at reducing tumor burden relative to either monotherapy (11- to 23-fold; Figure 5A).

Some of these effects could be magnified by increasing the number of cycles of combination chemotherapy. As the number of chemotherapy cycles was increased from 1 to 3, there was an increase in the enrichment of clones with higher plasticity (Figure 9B). Additionally, we observed increased tumor sizes with longer simulations, as treatments typically failed to prevent the outgrowth of few faster-growing clones. This was reflected in a dramatically increased variation in the number of surviving cancer cells across simulations.

We found that it was possible to enrich for any given phenotypic plasticity by altering the design of the combination chemotherapy. For example, if 7 treatments with an epithelial-specific agent were combined with 1 treatment with a mesenchymal-specific agent (instead of the 3:3 design considered above), there was a further enrichment of more-mesenchymal clones (Figure 9C). Conversely, if 7 treatments with a mesenchymal-specific agent were combined with 1 treatment with an epithelial-specific agent, the most strongly enriched clones were more-epithelial. Increasing the treatment imbalance only magnified this effect. However, the most effective combination therapies were balanced in treatment, and selected for clones with roughly equal mixtures of epithelial and mesenchymal cell types (Figure 9C). These observations indicated that plasticity was a clonal phenotype

that could be selected for (or against) by sequentially applying selection pressures for specific phenotypic states.

We next simulated how combination therapies that sequentially applied epithelial- and mesenchymal- specific treatments compared with therapies that alternated these treatments, while leaving unchanged the total dose of each therapy applied. Although the total dose of therapy applied stayed the same, a combination therapy that alternated between the mesenchymal- and epithelial-specific treatments was far more effective (48-fold) at reducing tumor size relative to the sequentially applied combination, while simultaneously greatly reducing the selection for clones with increased phenotypic plasticity (Figure 9D). Moreover, we found that doubling the rate at which the therapies were alternated — while halving their durations so as to maintain the same total dose of therapy applied — further reduced tumor size (Figure 9D). In contrast to the repeated sequential therapy design, repeating the alternating designs did not result in an increased tumor size, suggesting it prevented the enrichment of resistant clones. This observation demonstrates that the design of combination therapies has an enormous influence on their effectiveness, even in contexts where the total dose of therapy applied remains the same.

Discussion

In this study we used DNA barcodes to assess phenotypic plasticity across hundreds of clones in a single population of cancer cells. We found that the majority of cancer cell clones give rise to progeny in both the epithelial and mesenchymal states, and the ratio of epithelial and mesenchymal progeny differs between clones. Our results show that this ratio is stable within a clone, even over the course of weeks and through the rigors of single cell cloning.

We speculate that the marked stability of phenotypic ratios across many generations could be determined by genetic factors, as has been previously proposed [36]. Differences in many such factors

across clones would explain the log-normal distribution of phenotypic ratios we observed, if each factor had a small multiplicative effect on phenotypic ratio. As we found that each phenotypic state is a mixture of mostly the same clones, despite the bias of clones towards one state or another, it is not surprising that others rarely saw genetic differences between populations sorted by phenotype [24-27] .

Phenotypic switching can serve as a bet-hedging strategy allowing the survival of clones in diverse environments [37]. Phenotypic switching is prevalent across a variety of organisms, including prokaryotes [38, 39], yeasts [40, 41], and cancer cells [42]. In these examples, phenotype switching allows a clone to sample multiple phenotypes with different sensitivities and resistances, allowing the clone to survive in changing conditions. Since the epithelial and mesenchymal phenotypes we studied here are known to correlate strongly with sensitivity to most cancer therapies [43-45], phenotypic switching between these states would serve as an effective bet-hedging strategy for cancer cells. To be sure, cancer cells are not switching phenotypic states out of an awareness that this strategy will prove beneficial to them. Rather, as indicated by our simulations, cancer cell clones that undergo phenotypic switching have a competitive advantage and thus undergo a selective expansion when treated sequentially with therapies that selectively target the mesenchymal and epithelial states. The diversity of phenotypic plasticities observed across clones allows fluctuating environments to select for a subset of clones with bet-hedging strategies optimally suited to a particular environment. Thus, stably inherited differences in phenotypic plasticity enable tumors to evolve optimal bet-hedging strategies. Phenotypic switching is a powerful mechanism for overcoming selection pressures that vary over time— e.g. chemotherapy regimens— and is consistent with observations of changing phenotypic proportions in progressing tumors [11].

Supporting this interpretation, the enrichment of a particular set of clones based on cell state due to drug-induced selection has been observed *in vitro* [28]. Resistant clones of the HCC827 non-small

cell lung cancer cell line were observed to display a more mesenchymal phenotype than the parental cell line, suggesting that a heritable difference in cell state resulted in their expansion during selection.

In principle, the differences in cell-state proportions we observed in a patient glioblastoma could arise in the absence of cellular plasticity if the clones identified by our analyses consisted of sub-clones with stable and distinct phenotypes. However, we consider this unlikely since the existence of cellular plasticity in glioblastomas has been supported by several single-cell RNA sequencing studies of patient tumors [18]. While we used gene copy number differences to distinguish between the various clones, our analyses did not assess if these copy number distinctions played a functional role in determining clonal phenotypes.

The stable phenotypic plasticity of clones has implications for the design of combination treatments with phenotype-selective compounds. As conventional chemotherapeutics can cause the enrichment of a mesenchymal, resistant population [10, 43, 46], there have been significant efforts to develop therapies that target the resistant mesenchymal cells [47, 48]. Once developed, implementation of an appropriate treatment regimen will be important for the therapeutic success of these compounds. Even comparing combination therapies with the same total doses, our simulations showed the order and schedule of doses have profound effects on the effectiveness of the therapy. Strikingly, the most simple combination therapy schedule (one treatment, followed by the other) was also the worst performing, while more-rapid, repeated alternations between treatments were far more effective at reducing tumor burden. While changing selections enriched for more plastic clones, we found that even more rapid alternation would reduce clonal enrichment. These simulations suggest that, without due consideration of treatment schedule, the effectiveness of novel combination therapies could be undervalued. Additionally, our simulations underscore the importance of understanding heterogeneity and recommend alternations to be the most effective combination therapy.

Methods

Barcode library construction

Barcodes were synthesized as oligonucleotides from IDT (Coralville, IA), and are listed in Table 2 as ClonalBarcode5 and ClonalBarcode3. The oligonucleotides were annealed and ligated into pBabe Puro (Addgene #1764, Addgene, Cambridge, MA) that had been digested with BamHI-HF (New England Biolabs) and EcoRI-HF (New England Biolabs), treated with calf intestinal phosphatase (NEB), and purified with a PCR purification kit (Qiagen). 1µL of 150nM annealed clonal barcode was ligated to 190ng of digested pBabe Puro using T4 ligase (New England Biolabs) overnight at 16°. The ligation product was purified using 1X volume of AMPureXP beads (Beckman Coulter) as per the manufacturer's protocol, and eluted into 20µL. Four times, 2µL of purified ligation product was transformed into 40µL of DH5α Electromax *Echeria coli* (Fischer Scientific). Transformed bacteria were allowed to recover in 1mL SOC medium, pooled, and plated on LB Agar with 100µg/mL Ampicillin in two 245mm plates (Corning, Corning, NY). Some transformed mixture was diluted and plated for counting and colony estimation; this yielded an estimate of ~$1.7 * 10^6$ colonies. After overnight growth at 37°, colonies were scraped off and plasmid DNA extracted with a Gigaprep kit (Qiagen).

Cell culture, virus preparation, and infection

MDA-MB-157 cells (ATCC, Manassas, VA) and HEK293T cells were cultured in DMEM supplemented with 10% fetal bovine serum, Penicillin and Streptomycin, and GlutaMax (Thermo Fischer Scientific). Viral barcoding vectors were transfected into subconfluent HEK293T cells with pCL-10A1 retroviral packaging plasmid using Fugene 6 (Promega, Madison, WI) and viral supernatant was collected and concentrated with polyethylene glycol (PEG). For concentration, viral supernatant was spun at 931 gravities for 4 minutes and decanted into 1/5.5 volumes of sterile 50% PEG-3350 in PBS. After an overnight 4° incubation, the mixture was spun for 1455 gravities for 20 minutes, decanted, spun again at 524

gravities for 4 minutes, and the pellet resuspended in PBS with 1% bovine serum albumin and frozen at -80°. For infection, $1*10^4$ cells were incubated with concentrated virus and 30µg/mL protamine sulfate and spun at 1455 gravities for 1.5 hours. Viral concentration was optimized to infect approximately 10% of cells. After 48 hours cells were selected with 3µg/mL puromycin to kill uninfected cells. Barcoded cells were expanded without discarding cells until the population was at least $2*10^7$ cells, and subsequently split into supopulations no smaller than $2*10^6$ cells to maintain clonal representation.

Amplification and sequencing of barcode plasmid pool

Barcodes were amplified using PCR from 2ng of plasmid with 20 cycles of amplification, using ClBc_5_primer_AAG and ClBc_3_primer_CCT (see Table 2). PCR products were run on a 2% agarose gel, extracted using a gel extraction kit (Qiagen), and sequenced on a HiSeq 2000 (Illumina), TruSeq-DNA adaptors. The sequencing primer used was ClBc_seq_primer (see Table 2). Sequence data were analyzed with a custom Python script that first filtered by quality, where reads were only accepted if they contained fewer than 14 base pairs with a quality score < 25, and had no base pairs with a quality score <10. Additionally, reads were only accepted if they contained the index sequences marking each library and the sequences common to every barcode, and every base in those sequences had a quality score of > 25. This resulted in $2.4*10^6$ reads.

Estimation of plasmid pool complexity

Pool complexity was estimated based on published methods of estimating the number of classes based on sample coverage. [49] Where N equals the estimated number of barcodes, n = the sample size (2,447,204 reads), D = the number of unique barcodes observed (1,530,822), and f_1 = the number of barcodes observed only once (989,844), the sum of the probabilities of observed classes of barcodes was estimated as $\hat{C} = 1 - f_1/n \ = 0.5955$. This was used to estimate a lower bound on the number of unique barcodes in the plasmid pool, as $N = D/\hat{C} = 2{,}570{,}562$.

To estimate the accuracy of this estimation, random sequences of complexity *N* were randomly sampled (with replacement) *n* times, and the count of each unique sequence in the sample determined.

Poisson modeling of viral infection

Viral infection of cells was modeled based on the multiplicity of infection and assuming that the number of infections per cell followed a Poisson distribution, as has been observed by others. [50, 51]. Multiplicity of infection (MOI) was estimated from the estimated number of cells infected ($1.3 * 10^3$ out of $1*10^4$). Where *m* = MOI and $P(n)$ = the proportion of cells infected with *n* viruses, the MOI was estimated from $P(n > 0) = 1 - e^{-m}$ [50]. The MOI was therefore estimated as 0.139. A Poisson PDF was calculated from using this MOI as the μ parameter, and was used to estimate the number of cells infected with different numbers of barcodes.

Probability calculations of all cells uniquely barcoded, and simulations of barcodes in multiple cells

The probability at least two cells share a barcode after infection, or P(A), was calculated as 1 – P(A'), where P(A') is the probability that all cells have unique barcodes. This calculation is analogous to the so-called "Birthday Problem." [52] Where N = the estimated number of barcodes (from sequencing the barcode plasmid pool, 2,570,562) and c = the number of cells infected (1372),

$$P(A') = \prod_{i=1}^{c-1} \frac{N - i}{N}$$

To estimate the probability of different numbers of cells sharing barcodes with other cells, c barcodes were randomly sampled from N barcodes, $5*10^5$ times with replacement.

Intracellular flow cytometry

Cells were trypsinized, washed in DMEM supplemented with 10% fetal bovine serum (Sigma Aldrich, St. Louis, MO), washed 2x in phosphate buffered saline (PBS), and spun (as with all subsequent washes) at

524 gravities for 3 minutes. The pellet was disrupted by vortexing and the cells fixed by dripping in 2mL of ice-cold 70% ethanol while vortexing. Vortexing was continued for 30 seconds and the cells incubated overnight at 4 degrees. Cells were blocked by washing 3x in FACS buffer (FB), consisting of PBS supplemented with 6% fetal bovine serum (Sigma Aldrich). Cells were filtered through a 40µm filter, counted on a haemocytometer, resuspended to $1*10^6$ cells/mL in FB and stained with a 1:50 dilution of mouse anti K8/18, clone C51 (Cell Signaling, Danvers, MA), for 1-2 hours on ice. After washing 3x with FB, cells were stained in FB at a concentration of $1*10^6$ cells/mL and a 1:1000 dilution of goat anti mouse Fab Alexa Fluor 488 (Cell Signaling), incubating for 0.5 to 1 hour on ice in the dark. Cells were washed 3x in FB and resuspended at $1*10^6$ cells/mL in FB. Samples were run on a Fortessa (Becton Dickinson, Franklin Lakes, NJ), and flow cytometry data was analyzed with FlowJo (Tree Star, Ashland, OR).

Fluorescence activated cell sorting

For single cell cloning, clonally barcoded MDA-MB-157 cells were trypsinized, washed with PBS supplemented with 3% FBS, sent through a 40µm filter, and resuspended to $1*10^6$ cells/mL. A FACSaria (Becton Dickinson) was used to sort single cells into wells of 96 well plates, each well containing 100µL of DMEM with 10% FBS.

After expansion of the barcoded population of cells, a portion of the cells were stained for Keratin 8/18 expression and sorted via fluorescence activated cell sorting. Portions were separated out of the population and sorted at three time points each separated by a week (day 0, day 7, day 14).

For these sorts based on Keratin 8/18 expression, cells were stained as in intracellular flow cytometry, but stained at $1*10^7$ cells/mL in FB and with a 1:60 dilution of mouse anti K8/18, clone C51 (Cell Signaling), for 1-2 hours on ice. After secondary staining and washes, cells were resuspended at $1*10^7$

cells/mL in FB and sorted on a FACSAria (Becton Dickinson) set to maximize yield. Sorted samples were analyzed on the FACSAria to measure the proportion of cells missegregated, counted on a hemacytometer, and split in half. Barcodes were extracted from these cells as described below.

Extraction and amplification of barcodes from genomic DNA

Genomic DNA was collected with a DNeasy kit (Qiagen, Venlo, Netherlands). All genomic DNA was digested with BamHI and EcoRI (New England Biolabs, Ipswich, MA), using 3 units/µg and digesting for one hour at 37°. Digested DNA was directly purified from solution with a gel extraction kit (Qiagen), and barcodes size-selected with Agencourt AMPure XP beads (Beckman Coulter, Brea, CA). For size selection, one half volume of beads was added to the DNA mixture to bind to large DNA fragments, mixed by vortexing, and incubated at room temperature for 5 minutes. After precipitating the beads with a magnet, the supernatant containing small DNA was removed, and DNA was purified from the supernatant with a gel extraction kit (Qiagen) and quantified on a Nanodrop (Thermo Scientific). Barcodes were amplified from purified size-selected DNA using 25 cycles of PCR with ExTaq (Takara Bio, Kyoto, Japan), assembling the reaction mixture on ice. Template was added to a final concentration of 10ng/µL, and all size-selected DNA was used as template. This PCR step was used to also add library-specific index sequences (to allow for sequencing multiple samples in the same sequencing lane) and adaptor sequences for high-throughput sequencing. Index sequences were designed to have at least two differences from all other index sequences. Primer sequences are listed in Table 2. PCR products were purified with a PCR purification kit (Qiagen). Samples of 4, 2, and 1 µL of each PCR product were run on a 2% agarose gel and the intensity of the 131bp band quantified electronically. Samples' relative DNA concentration were computed with linear regression and the samples were combined in equimolar ratios. This combined library was run on a 2% agarose gel, and the 131bp band was purified with a gel extraction kit (Qiagen). The purified band was sequenced on a HiSeq2000 (Illumina, San Diego, CA); the resulting sequencing data is available at the NCBI SRA (SRX1175944).

Analysis of sequencing data (sorted cells)

Sequencing data were analyzed with custom Python scripts. Reads were first filtered by quality, where reads were only accepted if they contained fewer than 6 base pairs with a quality score < 25, and had no base pairs with a quality score <15. Additionally, reads were only accepted if they contained the index sequences marking each library and the sequences common to every barcode. These steps reduced $1.67*10^8$ reads to $1.01*10^8$ reads. This quality filtering procedure was more stringent than that used to analyze barcodes from the plasmid pool due to the increased cycles of amplification involved in library construction and lower starting pool complexity, which resulted in lower quality reads. Reads were then separated based on library-specific sequences that were introduced during PCR to distinguish samples. Taking reads for barcodes seen at least twice, we, as others, combined reads that could be connected with few mismatches, using the most abundant barcode to represent the group and giving it the abundance of the sum of the group's reads [29, 53]. To avoid erroneously combining barcodes that were by chance similar in sequence, we repeatedly iterated down the list of barcodes ordered by abundance, grouping together less abundant barcodes that were within one mismatch, and then repeating the process grouping together less abundant barcodes within two, three, and four mismatches. We then removed from analysis any barcodes that were not detected in any libraries from one or more time points.

These data were used to test for clones' bi-lineage potential and cell state bias, below, to allow for statistical analysis of clones based on the actual number of reads.

For further analysis, reads for each library were normalized by dividing by the sum of reads for that library multiplied by the fraction of the population consisting of that cell state at the time of sorting, being 60% for K8/18 high and 40% for K8/18 low. Any barcodes not found in a library were given a

fractional value of $1*10^{-6}$ for that library. To deal with sort contamination, for each barcode, and for each time point, we subtracted from each sorted library the average fraction of total reads of the other sorted populations multiplied by the fractional contamination observed in that sort from post-sort flow cytometry. Any barcode abundance thus brought to less than zero was given a value of $1*10^{-6}$. After determining in this way the size of each clone in each state, the results from the two sequenced replicates from each time point (see above) were combined by taking their mean.

Testing for clones' bi-lineage potential

To determine if clones did in fact have bi-lineage potential, we asked whether, for each clone, we could reject the hypothesis that there were as many reads as could be expected via sort contamination (mis-sorted cells), assuming each clone was composed entirely of cells in one state. The proportion of mis-sorted cells was determined via flow cytometry of the sorted cells (see above), here represented as $\sigma_{m,t}$ for the proportion of mesenchymal-sorted cells at time *t* that were actually mis-sorted epithelial cells, and similarly $\sigma_{e,t}$.

After sorting at three time points (days 0, 7, and 14; see above), each pool of sorted cells was split into two, and sequenced (see above). At each time point, the reads for each clone in each state were summed, creating r_e (*c, t*) for the summed epithelial (keratin 8/18 +) reads of clone *c* at time point *t,* and similarly r_m (*c, t*). As the count of observed reads for each clone were being compared to reads expected from sort contamination, the un-normalized reads from clones were used (see above).

In order to test for bi-lineage potential, we calculated for each state, at each time point, the expected probability of a read in the other state from sort contamination, assuming all cells were in the first state, or pC_e (*t*) for the probability of epithelial reads being mis-called as mesenchymal, assuming all cells were epithelial, and pC_m (*t*) for the probability of mesenchymal reads being mis-called as epithelial, assuming all cells were mesenchymal.

For ease of understanding the calculation of these probabilities, the reads from the epithelial-sorted population can be visualized as a combination of reads from real epithelial cells (totaling (1-σ_e) times the sum of epithelial-sorted reads) and reads from real mesenchymal cells (totaling σ_e times the sum of epithelial-sorted reads). Similarly, the reads from the mesenchymal-sorted population can be viewed as a combination of reads from real mesenchymal cells (totaling (1- σ_m) times the sum of mesenchymal-sorted reads) and reads from real epithelial cells (totaling σ_m times the sum of mesenchymal-sorted reads).

Therefore, the sum of reads from correctly-sorted epithelial cells (*E*) at time *t* is

$$E = \left(1 - \sigma_{e,t}\right) * \sum_{\kappa \in clones} r_e(\kappa, t)$$

The sum of reads from incorrectly-sorted epithelial cells (*E'*) at time *t* therefore is

$$E' = \sigma_{m,t} * \sum_{\kappa \in clones} r_m(\kappa, t)$$

The sum of reads from correctly- and incorrectly- sorted mesenchymal cells was calculated similarly.

pC_e is defined as the proportion of all epithelial reads that were mis-called as mesenchymal, which is equal to *E'* / (*E* +*E'*). Therefore, these probabilities were calculated as:

$$pC_e(t) = \frac{\sigma_{m,t} * \sum_{\kappa \in clones} r_m(\kappa, t)}{\sigma_{m,t} * \sum_{\kappa \in clones} r_m(\kappa, t) + \left(1 - \sigma_{e,t}\right) * \sum_{\kappa \in clones} r_e(\kappa, t)}$$

$$pC_m(t) = \frac{\sigma_{e,t} * \sum_{\kappa \in clones} r_e(\kappa, t)}{\sigma_{e,t} * \sum_{\kappa \in clones} r_e(\kappa, t) + \left(1 - \sigma_{m,t}\right) * \sum_{\kappa \in clones} r_m(\kappa, t)}$$

For each time point, and each clone, these probabilities were used to test the hypothesis that each clone was actually monolineage. To test the null hypothesis that all clones were epithelial, each clone was evaluated at 1- the CDF of a binomial distribution with $p = pC_e$ and $n = r_e(c, t) + r_m(c, t)$ for clone

c, evaluated at $x = r_m(c,t)$. Similarly, to test the null hypothesis that all clones were mesenchymal, each clone was evaluated at 1- the CDF of a binomial distribution with $p = pC_m$ and $n = r_e(c,t) + r_m(c,t)$ for clone *c,* evaluated at $x = r_e(c,t)$. These resulting p values from all states and all time points were corrected for multiple hypothesis testing using the Benjamini-Hochberg method [54]. The null hypothesis that a clone was monolineage at a particular time point was rejected to control the false discovery rate (FDR) at 0.05. A clone was declared monolineage if the null hypothesis was rejected at all time points for the same state, and in no time points for the other state. A clone was declared bi-lineage if the null hypothesis was rejected in both states (still at the FDR of 0.05) in at least 1 time point.

Testing for clones' cell state bias

To test if clones had bias in cell state, we attempted to reject the null hypothesis that each clone's cell state proportions matched the population cell state proportion. Again, pre-normalized reads (see above) were used. Reads from the two sequenced replicates of each sorted population were summed. For each clone, at each time point, the expected number of reads in the epithelial and mesenchymal states were calculated, or $E_e(c,t)$ and $E_m(c,t)$, respectively:

$$E_e(c,t) = \big(r_e(c,t) + r_m(c,t)\big) * \rho_t$$

$$E_m(c,t) = \big(r_e(c,t) + r_m(c,t)\big) * (1 - \rho_t)$$

$$\rho_t = \frac{\sum_{\kappa \in clones} r_e(\kappa,t)}{\sum_{\kappa \in clones} r_e(\kappa,t) + r_m(\kappa,t)}$$

The χ^2 test was used to examine the significance of the fit between the observed and expected reads. The upper CDF of the χ^2 distribution with one degree of freedom was evaluated at *x,* where:

$$x = \frac{\big(r_e(c,t) - E_e(c,t)\big)^2}{E_e(c,t)} + \frac{\big(r_m(c,t) - E_m(c,t)\big)^2}{E_m(c,t)}$$

These resulting p values from all time points were corrected for multiple hypothesis testing using the Benjamini-Hochberg method [54]. The null hypothesis that a clone at a time point had a cell state bias matching the population cell state proportions was rejected so as to control the false discovery rate at 0.05. Clones were declared significantly different from the population cell state proportion if this null hypothesis was rejected at all time points.

Fraction Epithelial and Log_2(Epithelial/Mesenchymal) ratio calculations

After normalizing the sequencing data of sorted cells (see above) to determine the size of each clone in the epithelial and mesenchymal states, and after subtracting those reads estimated to come from sort contamination, each clone's cell-state bias was determined, represented by the fraction of the clone that was epithelial (fraction epithelial) or the Log_2(Epithelial/Mesnechymal) ratio (Log_2(E/M)).

Here, $E_{c,t}$ equals the normalized fraction of cells of clone *c* in the epithelial state at time point *t*, and similarly $M_{c,t}$. These estimated cell counts are normalized such that the sum of epithelial and mesenchymal counts across clones at each time point equals one. For both of these calculations, dividing E by M or (E+M) cancels out this normalization factor, rendering the calculations equivalent to those using counts of cells. The fraction of clone *c* epithelial at time point *t* was calculated as:

$$\frac{E_{c,t}}{E_{c,t} + M_{c,t}}$$

And the Log_2(E/M) ratio for clone *c* at time point *t* was calculated as

$$\log_2 \frac{E_{c,t}}{M_{c,t}}$$

The Log_2(E/M) ratio for clone *c* averaging across the three time points was calculated as

$$\frac{1}{3} * \sum_{t=1}^{3} \log_2 \frac{E_{c,t}}{M_{c,t}}$$

Testing the significance of the correlation of clones' cell state bias across time points

The Pearson correlation of clones' Log2 E/M ratio across time points was determined to assess the stability of cell state bias. To determine the probability of randomly obtaining a correlation higher than the one observed in each of the three comparisons across time points, the barcode labels of one time point in each comparison were randomly shuffled. After each randomization the Pearson correlation was evaluated and the correlation coefficient ρ recorded. After $1*10^6$ such randomizations, the distribution of randomized ρ values was compared with the observed ρ, and the proportion of randomized ρ greater than observed ρ determined.

Clone Growth Rate calculations

From the frequency of splitting during cell culture, the population of cells was estimated to double approximately three times per week. The population of cells sorted at the first time point consisted of 2.9E7 cells, and this population growth rate was used to estimate the number of cells at the one and two weeks later, at the second and third time points. As the barcode sequencing information was used to calculate relative size of each clone as a fraction of the total population, these population cell numbers were used to compute the cell numbers of each clone at each time point through multiplication. Each clone's number of cells at the second and third time points ($N_{i,c}$) were compared to cell numbers from the first time point ($N_{1,c}$) to compute each clone's growth rate over these two intervals; the two rates were averaged to compute each clone's growth rate (k_c).

$$k_c = \frac{1}{2} * \sum_{i=1}^{2} \ln\left(\frac{N_{i,c}}{N_{0,c}}\right) * \frac{1}{7i}$$

Calculation of Shannon entropy of the distribution of clones' epithelial/mesenchymal ratio

After calculating the geometric-average log2(Epithelial/Mesenchymal) for each clone across the three examined time points, clones were binned from the minimum ratio to the maximum ratio in bins of width 1 (corresponding to a 2 fold change in ratio). The Shannon entropy of clones thus binned was calculated.

Flow cytometry of single-cell clones

Single-cell clones were trypsinized, washed in DMEM supplemented with 10% fetal bovine serum (Sigma Aldrich), and washed 2X in phosphate buffered saline (PBS). Cells were fixed as in intracellular flow cytometry. To serve as an internal staining control for the single-cell clones, pooled clones (the parental barcoded population) were fixed in the same way as the single-cell clones. This pooled clone population was resuspended at $1*10^6$ cells / mL in PBS and covalently stained with 1µL/mL of Blue Live/Dead Discrimination Dye (Thermo Fischer Scientific, Cambridge, MA) for 30 minutes on ice in the dark. Cells were blocked by washing 3x in FACS buffer, filtered through a 40µm filter, counted on a hemacytometer, and resuspended to $1*10^6$ cells/mL in FB. For analysis of single-cell clones, clones were mixed 1:1 with samples from the covalently stained pooled clones. Samples were then stained as in the intracellular flow cytometry protocol and run on a Fortessa (Becton Dickinson), and the flow cytometry data analyzed with FlowJo. After using FlowJo gating to remove debris, the flow cytometry data were exported and analyzed with a custom Python script. In brief, this script separated the cells of the pooled clones, and determined thresholds of K8/18 staining to gate E and M from this population such that the gates contained the same proportion of cells as the gates used for cell sorting. These gates were then used to determine the proportion of cells from the single-cell clone that would have been sorted as epithelial or mesenchymal to calculate the Log2 [epithelial/mesenchymal].

PCR and Sanger sequencing of barcodes from single-cell clones

Genomic DNA was collected with a DNeasy blood and tissue kit (Qiagen) as per the manufacturer's instructions. Barcodes were amplified with nested PCR, using two sets of primers (IDT, Coralville, IA) (first ClBc_A5 and ClBc_A3, then ClBc_B5 and ClBc_B3; see Table 2) to specifically amplify one band. The initial genomic DNA concentration was 1.2 ng/μL, and each round of PCR was 25 cycles. The PCR product was purified with a PCR purification kit (Qiagen) and sequenced with Sanger sequencing (Genewiz, South Plainfield, NJ).

Analysis of correlation of single-cell clone / pool phenotypic ratio

Narrow-sense heritability was calculated as the Pearson correlation coefficient [55]. 28 clones were deemed sufficient as power analysis showed a power of 0.96 to reject the null hypothesis at a significance of 0.01 for correlations of 0.7, calculated using the pwr package in R. To determine the probability of randomly obtaining a correlation higher than the one observed between single-cell clones' phenotypic proportion and those clones' phenotypic proportion in the pooled experiment, the barcode labels of single-cell clones were randomly shuffled between the single-cell clones using a custom script. After each randomization the Pearson correlation was evaluated and the correlation coefficient ρ recorded. After $1*10^6$ such randomizations, we compared the distribution of randomized ρ values with the observed ρ.

Glioblastoma single-cell RNA sequencing data

Normalized single-cell RNA sequencing data from primary glioblastoma tumors were obtained from the Gene Expression Omnibus (accession GSE57872) [18].

Copy number estimation from single-cell RNA sequencing data and clone separation

As has been previously described [18, 34], changes in copy number were estimated through analysis of single-cell RNA expression by chromosomal location. Mean normalized (by gene across cells) Log_2(TPM+1) RNA values from single cells were accessed from GSE57872 [18, 56]. RNA data from cells that were either identified as non-cancer cells or as cells from tumor MGH31. For the purposes of determining copy number variation, these data were thresholded, so that values >3 were set to 3, and values <-3 were set to -3. For each cell, we computed a sliding average of the expression of 101 genes moving down the list of genes with RNA data ordered by chromosomal location, to build a copy number variation profile (CNV profile). We then centered each cell's CNV profile at 0 (subtracting from each profile the mean value) to deal with any differences in expression remaining across cells. This meant that the CNV profile value for cell *j* at position *i* ($CNV_{i,j}$) is

$$CNV_{i,j} = \frac{CNV_{0\ i,j}}{\sum_{c \in cells} CNV_{\ i,c}} \text{ where } CNV_{0\ i,j} = \frac{1}{101} * \sum_{k=i-50}^{k=i+50} RNA_{k,j}\ ;\ \ RNA_{k,j} = \frac{\log_2(TPM_{k,j}+1)}{\sum_{c \in cells} \log_2(TPM_{k,c}+1)} * \frac{1}{|cells|}$$

To normalize to the average expression by chromosomal location, so as to deal with differences in expression across chromosomes, each cell's CNV profile was normalized using an averaged CNV profile from normal cells (CNV_{Base}), computed for each genomic location through an identical sliding-average strategy from normal neural cells identified in the same RNAseq data set [18]. For each cell, this normalized CNV (CNV_{norm}) was calculated as follows:

$$CNV_{norm\ i,j} = \begin{cases} CNV_{i,j} - CNV_{Base\ i}, \text{ if } CNV_{i,j} > CNV_{Base\ i} + 0.3 \\ CNV_{i,j} - CNV_{Base\ i}, \text{ if } CNV_{i,j} < CNV_{Base\ i} - 0.3 \\ 0, \text{ if } CNV_{Base\ i} - 0.3 < CNV_{i,j} < CNV_{Base\ i} + 0.3 \end{cases}$$

In this way CNV values were only recorded if they deviated significantly from the value obtained from normal cells, where a difference of 0.3 corresponds to a 23% change.

The cells' normalized CNV profiles were clustered via Ward's method, using the Euclidean distances between CNV profiles. This method clusters vectors by finding, at each step, the pair of clusters that leads to the minimum increase in within-cluster variability when the clusters are combined. In this way, the hierarchical clustering of CNV profiles clustered cells based on similar CNV profiles; cells were divided into clones based on this hierarchical clustering.

Subtype and stem-ness classification from single-cell RNA sequencing data

Cells were classified by subtype as previously described [18]. Cells were scored by subtype using published lists of genes enriched in each subtype [35], or marking cells with increased stem-ness[18]. Mean normalized (by gene, across cells) Log_2(TPM+1) RNA values for single cells were accessed from GSE57872 [18, 56]. For each cell, a score was calculated for each subtype. These scores for cell *i* for subtype *j* ($S_{i,j}$) were calculated by taking the average expression of classifier genes for subtype *j* (G_c) in cell *i,* and subtracting the mean expression of every gene in cell *i*:

$$S_{i,j} = \sum_{g \in G_c} RNA_{g,i} - \frac{1}{|G|} * \sum_{g \in G} RNA_{g,i}, \text{ where } RNA_{g,i} = \frac{\log_2(TPM_{g,i}+1)}{\sum_{c \in cells} \log_2(TPM_{g,c}+1)} * \frac{1}{|cells|}.$$

To evaluate each individual score for significance, we adapted a previously published method [18], evaluating the enrichment of each score relative to random sets. For each subtype, we made 100 random subtype-classifier gene sets from randomly sampling the set of sequenced genes. Each random set had the same number of genes as the real subtype gene set. We called each cell's subtype score as enriched or depleted by comparing it to these random scores. If the real score was greater than 95% of the random scores, we called it enriched; whereas, if the real score was less than 95% of the random scores, we called it depleted.

To determine if clones had different distributions of subtype scores, suggesting differences in plasticity, we evaluated the distributions of subtype scores across cells grouped in into clones by the clustering above with a Kruskal-Wallis test.

Simulations

Mechanistically, a "tumor" was seeded with 500 clones. Each clone was assigned a fraction epithelial, growth rate, and cell numbers matching a randomly chosen observed clone, so as to simulate the distribution of the observed growth rates and fraction epithelial. Both the growth rate and fraction epithelial parameters were inherent to the clone for the entirety of the simulation. Each clone's starting cell numbers were drawn from those observed at day 0 of sorting (see above). After instantiation of the tumor, growth was modeled under different treatment regimes, with 15 time points modeling a day.

The amount of cells for each clone after division were calculated as: $N_{s,i}^{0}(t) = N_{s,i}(t-1) * 2^{\frac{1}{D_i}}$, where $N_{s,i}(t)$ represents the number of cells of clone *i* in state *s* at time *t*, D_i represents the doubling time (in the time scale of the simulation, where 15 time points is equivalent to 1 day) of clone *i*.

At each time point, cells were also allowed to differentiate. Each clone's transition probabilities for going from epithelial to mesenchymal or mesenchymal to epithelial were defined so that 20% of cells changed state per division, and the ratio of transition probabilities matched the clone's defined equilibrium of cell states; in this way, the clones have stable cell state proportions at equilibrium and slowly return to equilibrium after the cell state ratios are perturbed through selection. This probability of differentiation was chosen to reflect those observed in other contexts [12]. The resulting number of cells of clone *i* in state *s* (where the other state is *s'*), or $N^{1}_{s,i}$, as a consequence of differentiation is as follows:

$$N_{s,i}^{1}(t) = N_{s,i}^{0}(t) + N_{s\prime,i}^{0}(t) * \frac{P_{s',i}}{D_i} - N_{s,i}^{0}(t) * \frac{P_{s,i}}{D_i}$$

Here, $P_{s,i}$ represents the probability that a cell of clone *i* differentiates from state *s* to state *s'*. This was calculated as follows, where R_i represents the equilibrium epithelial:mesenchymal ratio of clone *i*: $P_{M,i} = \frac{R_i * \psi}{1+R_i}$ and $P_{E,i} = \frac{\psi}{1+R_i}$. ψ here represents the probability of a cell differentiating during a division, or 0.2 as discussed above. If no treatment is simulated during this time point, $N^1_{s,i}$ is now the final count of cells for clone *i* in state *s* for time point t, or $N_{s,i}(t)$.

Treatments were applied as mesenchymal-specific or epithelial-specific, where a mesenchymal targeting therapy killed a 10-fold higher fraction of the mesenchymal cells compared to epithelial cells. This was chosen to match the relative effectiveness of certain *in vitro* compounds on cells in different differentiation states [57]. The number of cells remaining after death ($C^2_{s,i}$) for a treatment targeting state *s* is calculated as $N_{s,i}^{2}(t) = N_{s,i}^{1}(t) * (1 - \delta_s^s)$, and for a treatment targeting state *s'* (the other state), $N_{s,i}^{2}(t) = N_{s,i}^{1}(t) * (1 - \delta_s^{s\prime})$, where $\delta_s^{s\prime} = 0.01 = 0.1 * \delta_s^s$. In this case $N^2_{s,i}$ is now the final count of cells for clone *i* in state *s* for time point *t*, or $N_{s,i}(t)$.

Each treatment cycle killed a fraction of the cells for 30 time points, simulating a course of therapy for two days, which was followed with 20 time points of no treatment. The simulation was ended after the conclusion of treatment-rest periods; the number and pattern of treatment-rest periods varied among the simulations. A variety of treatment combinations were simulated as detailed in the results. To compare the results of different simulations, clones were binned by their fraction epithelial. Each clone's fold change in cell numbers during the simulation (fc_i for clone *i*) was computed as

$$fc_i = \frac{N_{i,s,}(t_{end}) + N_{i,s\prime,}(t_{end})}{N_{i,s,}(t_0) + N_{i,s\prime,}(t_0)}$$

Where $N_{s,i}(t)$ represents the number of cells of clone i in state s at time t, t_0 is the time point of the start of simulated treatments, and t_{end} the time point of the end of simulated treatments. For each bin of clones by fraction epithelial, the median fold change in cell numbers for the clones in each bin was computed. The sum of cells across clones at the last point was also computed for each simulated treatment. Simulations were repeated 500 times, and the 0.1, 0.5, and 0.9 quantiles of the median clone fold change for each bin across simulations were recorded. Similarly, the 0.1, 0.5, and 0.9 quantiles of the sum of cell numbers across simulations were recorded.

Acknowledgements

We acknowledge the Whitehead Institute flow cytometry facility for assistance with flow cytometry and cell sorting, and the MIT BioMicro Center and Whitehead Institute Genome Technology Core for assistance with high-throughput sequencing.

This work was funded by the National Science Foundation Graduate Research Fellowship Program (1122374; ESS), the Richard and Susan Smith Family Foundation, the Breast Cancer Alliance, and the National Institutes of Health (2T32GM007287-36, RM).

References

1. Heppner, G.H., *Tumor heterogeneity.* Cancer Res, 1984. **44**(6): p. 2259-65.
2. Meacham, C.E. and S.J. Morrison, *Tumour heterogeneity and cancer cell plasticity.* Nature, 2013. **501**(7467): p. 328-37.
3. Nowell, P.C., *The clonal evolution of tumor cell populations.* Science, 1976. **194**(4260): p. 23-8.
4. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* N Engl J Med, 2012. **366**(10): p. 883-92.
5. Shah, S.P., et al., *The clonal and mutational evolution spectrum of primary triple-negative breast cancers.* Nature, 2012. **486**(7403): p. 395-9.
6. Yates, L.R., et al., *Subclonal diversification of primary breast cancer revealed by multiregion sequencing.* Nat Med, 2015. **21**(7): p. 751-9.
7. Wang, Y., et al., *Clonal evolution in breast cancer revealed by single nucleus genome sequencing.* Nature, 2014. **512**(7513): p. 155-60.
8. Navin, N., et al., *Tumour evolution inferred by single-cell sequencing.* Nature, 2011. **472**(7341): p. 90-4.
9. Eirew, P., et al., *Dynamics of genomic clones in breast cancer patient xenografts at single-cell resolution.* Nature, 2015. **518**(7539): p. 422-6.
10. Fillmore, C.M. and C. Kuperwasser, *Human breast cancer cell lines contain stem-like cells that self-renew, give rise to phenotypically diverse progeny and survive chemotherapy.* Breast Cancer Res, 2008. **10**(2): p. R25.

11. Park, S.Y., et al., *Heterogeneity for stem cell-related markers according to tumor subtype and histologic stage in breast cancer.* Clin Cancer Res, 2010. **16**(3): p. 876-87.
12. Gupta, P.B., et al., *Stochastic state transitions give rise to phenotypic equilibrium in populations of cancer cells.* Cell, 2011. **146**(4): p. 633-44.
13. Chaffer, C.L., et al., *Normal and neoplastic nonstem cells can spontaneously convert to a stem-like state.* Proc Natl Acad Sci U S A, 2011. **108**(19): p. 7950-5.
14. Chaffer, C.L., et al., *Poised chromatin at the ZEB1 promoter enables breast cancer cell plasticity and enhances tumorigenicity.* Cell, 2013. **154**(1): p. 61-74.
15. Chang, H.H., et al., *Transcriptome-wide noise controls lineage choice in mammalian progenitor cells.* Nature, 2008. **453**(7194): p. 544-7.
16. Roesch, A., et al., *A temporarily distinct subpopulation of slow-cycling melanoma cells is required for continuous tumor growth.* Cell, 2010. **141**(4): p. 583-94.
17. Yang, G., et al., *Dynamic equilibrium between cancer stem cells and non-stem cancer cells in human SW620 and MCF-7 cancer cell populations.* Br J Cancer, 2012. **106**(9): p. 1512-9.
18. Patel, A.P., et al., *Single-cell RNA-seq highlights intratumoral heterogeneity in primary glioblastoma.* Science, 2014. **344**(6190): p. 1396-401.
19. Phillips, S., et al., *Cell-state transitions regulated by SLUG are critical for tissue regeneration and tumor initiation.* Stem Cell Reports, 2014. **2**(5): p. 633-47.
20. Schwitalla, S., et al., *Intestinal tumorigenesis initiated by dedifferentiation and acquisition of stem-cell-like properties.* Cell, 2013. **152**(1-2): p. 25-38.
21. Polyak, K. and R.A. Weinberg, *Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits.* Nat Rev Cancer, 2009. **9**(4): p. 265-73.
22. Marjanovic, N.D., R.A. Weinberg, and C.L. Chaffer, *Cell plasticity and heterogeneity in cancer.* Clin Chem, 2013. **59**(1): p. 168-79.
23. Mani, S.A., et al., *The epithelial-mesenchymal transition generates cells with properties of stem cells.* Cell, 2008. **133**(4): p. 704-15.
24. Klevebring, D., et al., *Sequencing of breast cancer stem cell populations indicates a dynamic conversion between differentiation states in vivo.* Breast Cancer Res, 2014. **16**(4): p. R72.
25. Shipitsin, M., et al., *Molecular definition of breast tumor heterogeneity.* Cancer Cell, 2007. **11**(3): p. 259-73.
26. Balic, M., et al., *Genetic and epigenetic analysis of putative breast cancer stem cell models.* BMC Cancer, 2013. **13**: p. 358.
27. Park, S.Y., et al., *Cellular and genetic diversity in the progression of in situ human breast carcinomas to an invasive phenotype.* J Clin Invest, 2010. **120**(2): p. 636-44.
28. Bhang, H.E., et al., *Studying clonal dynamics in response to cancer therapy using high-complexity barcoding.* Nat Med, 2015.
29. Nguyen, L.V., et al., *DNA barcoding reveals diverse growth kinetics of human breast tumour subclones in serially passaged xenografts.* Nat Commun, 2014. **5**: p. 5871.
30. Wagenblast, E., et al., *A model of breast cancer heterogeneity reveals vascular mimicry as a driver of metastasis.* Nature, 2015. **520**(7547): p. 358-62.
31. Cheung, A.M., et al., *Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice.* Blood, 2013. **122**(18): p. 3129-37.
32. Lu, R., et al., *Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding.* Nat Biotechnol, 2011. **29**(10): p. 928-33.
33. Taylor-Papadimitriou, J., et al., *Keratin expression in human mammary epithelial cells cultured from normal and malignant tissue: relation to in vivo phenotypes and influence of medium.* J Cell Sci, 1989. **94 (Pt 3)**: p. 403-13.

34. Tirosh, I., et al., *Single-cell RNA-seq supports a developmental hierarchy in human oligodendroglioma.* Nature, 2016.
35. Verhaak, R.G., et al., *Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1.* Cancer Cell, 2010. **17**(1): p. 98-110.
36. Huang, S., *Genetic and non-genetic instability in tumor progression: link between the fitness landscape and the epigenetic landscape of cancer cells.* Cancer Metastasis Rev, 2013. **32**(3-4): p. 423-48.
37. Kussell, E. and S. Leibler, *Phenotypic diversity, population growth, and information in fluctuating environments.* Science, 2005. **309**(5743): p. 2075-8.
38. Beaumont, H.J., et al., *Experimental evolution of bet hedging.* Nature, 2009. **462**(7269): p. 90-3.
39. Kussell, E., et al., *Bacterial persistence: a model of survival in changing environments.* Genetics, 2005. **169**(4): p. 1807-14.
40. Acar, M., J.T. Mettetal, and A. van Oudenaarden, *Stochastic switching as a survival strategy in fluctuating environments.* Nat Genet, 2008. **40**(4): p. 471-5.
41. Newby, G.A. and S. Lindquist, *Blessings in disguise: biological benefits of prion-like mechanisms.* Trends Cell Biol, 2013. **23**(6): p. 251-9.
42. Sharma, S.V., et al., *A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations.* Cell, 2010. **141**(1): p. 69-80.
43. Singh, A. and J. Settleman, *EMT, cancer stem cells and drug resistance: an emerging axis of evil in the war on cancer.* Oncogene, 2010. **29**(34): p. 4741-51.
44. Feng, Y.X., et al., *Epithelial-to-mesenchymal transition activates PERK-eIF2alpha and sensitizes cells to endoplasmic reticulum stress.* Cancer Discov, 2014. **4**(6): p. 702-15.
45. Tiwari, N., et al., *EMT as the ultimate survival mechanism of cancer cells.* Semin Cancer Biol, 2012. **22**(3): p. 194-207.
46. Li, X., et al., *Intrinsic resistance of tumorigenic breast cancer cells to chemotherapy.* J Natl Cancer Inst, 2008. **100**(9): p. 672-9.
47. Visvader, J.E. and G.J. Lindeman, *Cancer stem cells: current status and evolving complexities.* Cell Stem Cell, 2012. **10**(6): p. 717-28.
48. Gupta, P.B., et al., *Identification of selective inhibitors of cancer stem cells by high-throughput screening.* Cell, 2009. **138**(4): p. 645-59.
49. Anne Chao, S.-M.L., *Estimating the Number of Classes via Sample Coverage.* Journal of the American Statistical Association, 1992. **87**(417): p. 210-217.
50. Arai, T., et al., *Dose-dependent transduction of vesicular stomatitis virus G protein-pseudotyped retrovirus vector into human solid tumor cell lines and murine fibroblasts.* Virology, 1999. **260**(1): p. 109-15.
51. Ellis, E.L. and M. Delbrück, *The growth of bacteriophage.* The Journal of general physiology, 1939. **22**(3): p. 365-384.
52. Mathis, F.H., *A generalized birthday problem.* SIAM Review, 1991. **33**(2): p. 265-270.
53. Nguyen, L.V., et al., *Clonal analysis via barcoding reveals diverse growth and differentiation of transplanted mouse and human mammary stem cells.* Cell Stem Cell, 2014. **14**(2): p. 253-63.
54. Benjamini, Y. and Y. Hochberg, *Controlling the false discovery rate: a practical and powerful approach to multiple testing.* Journal of the royal statistical society. Series B (Methodological), 1995: p. 289-300.
55. Griffiths, A.J.F., Wessler, Susan R. , Carroll, Sean B. , Doebley, John., *Introduction to Genetic Analysis*. 2012, New York: W. H. Freeman and Company.
56. Li, B. and C.N. Dewey, *RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome.* BMC bioinformatics, 2011. **12**(1): p. 1.

57. Germain, A.R., et al., *Identification of a selective small molecule inhibitor of breast cancer stem cells.* Bioorganic & medicinal chemistry letters, 2012. **22**(10): p. 3571-3574.

Tables

Barcode	Single Cell Clone log_2[E/M]	Pool log_2[E/M]
AACATTTTACTGAG	3.345	0.891
AAGTTCGAACTTAG	1.687	1.934
AAGTTCGAACTTAG	1.895	1.934
CAGTTCATCTAGTC	-9.522	-3.157
CATATAATTCAGGC	6.583	5.292
CGCTGAGTAAAAGG	2.737	0.817
CGTATTGAGGTTCC	-5.642	-4.257
CTGGCGGAACGTGG	1.932	1.285
CTTCACACTAATTT	2.626	-0.987
GAGACGAATGGGTA	3.329	3.188
GATACTAGCTACTA	-0.825	-1.357
GCAACGGACGCGAC	3.457	1.310
GCAACGGACGCGAC	2.833	1.310
GCATCCAGCCTCCT	-2.721	0.031
GGATTTAGGACTAC	-8.617	-3.929
GGGGCTAGGTTGGG	-1.141	-1.084
GGGTTCTAGGCAGG	-6.932	-4.074
GGTGGGGGTGGAGT	0.504	-0.905
GTATCGGGGGAACG	-9.928	-5.013
GTATCGGGGGAACG	-9.179	-5.013
GTTTTAACATGGGG	-1.584	-2.766
TAAAGAATACCGCT	2.837	1.391
TATAACTATTCCGA	3.105	1.432
TGACTCATCGTTTA	-0.953	1.166
TGACTCATCGTTTA	0.379	1.166
TGGCAATTACTATT	-1.599	1.280
TGGCAATTACTATT	-0.955	1.280
TTTTTAACATTGAT	1.559	1.090

Table 1: Phenotypic ratio of single cell clones

ClBc_5_primer_AAG	AATGATACGGCGACCACCGAGTAGACGGAGCGGACAACACTAAGA CAGG
ClBc_5_primer_CAT	AATGATACGGCGACCACCGAGTAGACGGAGCGGACAACACTCATA CAGG
ClBc_5_primer_TAG	AATGATACGGCGACCACCGAGTAGACGGAGCGGACAACACTTAGA CAGG
ClBc_5_primer_CAA	AATGATACGGCGACCACCGAGTAGACGGAGCGGACAACACTCAAA CAGG
ClBc_5_primer_TTC	AATGATACGGCGACCACCGAGTAGACGGAGCGGACAACACTTTCA CAGG
ClBc_5_primer_ATC	AATGATACGGCGACCACCGAGTAGACGGAGCGGACAACACTATCA CAGG
ClBc_5_primer_GTT	AATGATACGGCGACCACCGAGTAGACGGAGCGGACAACACTGTTA CAGG
ClBc_5_primer_GTA	AATGATACGGCGACCACCGAGTAGACGGAGCGGACAACACTGTAA CAGG
ClBc_3_primer_TCC	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGGCTCTGGATT GTCAGCGTTCCCGTGC
ClBc_3_primer_TCG	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGGCTCTGGATT GTCAGCGTTCGCGTGC
ClBc_3_primer_GCT	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGGCTCTGGATT GTCAGCGTGCTCGTGC
ClBc_3_primer_CCT	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGGCTCTGGATT GTCAGCGTCCTCGTGC
ClBc_3_primer_GGA	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGGCTCTGGATT GTCAGCGTGGACGTGC
ClBc_3_primer_CGA	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGGCTCTGGATT GTCAGCGTCGACGTGC
ClBc_3_primer_AGC	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGGCTCTGGATT GTCAGCGTAGCCGTGC
ClBc_3_primer_AGG	CAAGCAGAAGACGGCATACGAGCTCTTCCGATCTGGCTCTGGATT GTCAGCGTAGGCGTGC
ClBc_seq_primer	CCGAGTAGACGGAGCGGACAACACT
ClonalBarcode5	5’phos/GATCCTAGACGGAGCGGACAACACTGACACAGGNNNNNNN GAGAGNNNNNNNGCACGTGTACGCTGACAATCCAGAGCCG
ClonalBarcode3	5’phos/AATTCGGCTCTGGATTGTCAGCGTACACGTGCNNNNNNNC TCTCNNNNNNNCCTGTGTCAGTGTTGTCCGCTCCGTCTAG
CLBc_A5	CTCTGCAGAATGGCCAACC
CLBc_A3	CTTCTGGAATAGCTCAGAGGCCGAG
CLBc_B5	GCACCTTTAACCGAGACCTCATCAC
CLBc_B3	GACTTTCCACACCTGGTTGC

Table 2: Table of primers and oligonucleotide sequences

Figures

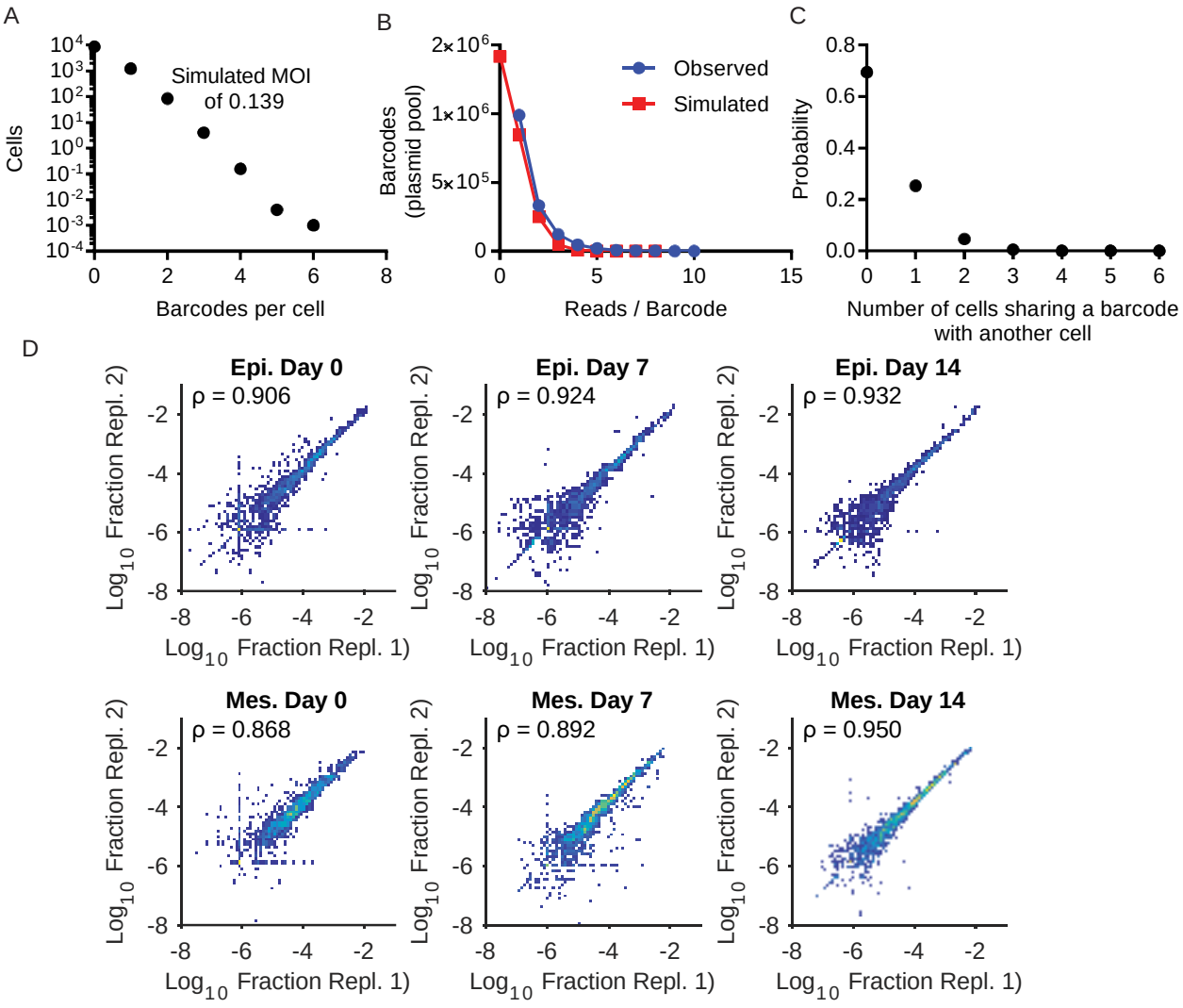


Figure 1. Tracking clones with DNA barcodes.
(a) Simulation of the infection conditions from the calculated MOI using a Poisson distribution suggests very few cells (6.8%) received more than one barcode. **(b)** High-throughput sequencing results of the pool of barcodes, plotted as the number of barcodes observed with each number of reads vs the reads per barcode. Also plotted is the expected number of unique barcodes sequenced assuming equal abundance of barcodes and $2.6*10^6$ total unique barcodes, the complexity calculated from the observed distribution (see Materials and Methods). **(c)** Plot of the probability (y) of a certain number of infected cells (x) sharing a barcode with another cell (via simulation). **(d)** Technical replicates of barcodes sampled from the same population of labeled cells show good reproducibility. Each barcode's estimated fraction of the total population is plotted in two halves of one population, split before extracting DNA. Barcodes not found in one replicate were given values of $1*10^{-6}$ (see Methods). Color indicates the density of points; abbreviations are epithelial cells (Epi.), mesenchymal cells (Mes.), technical replicate (Repl.).

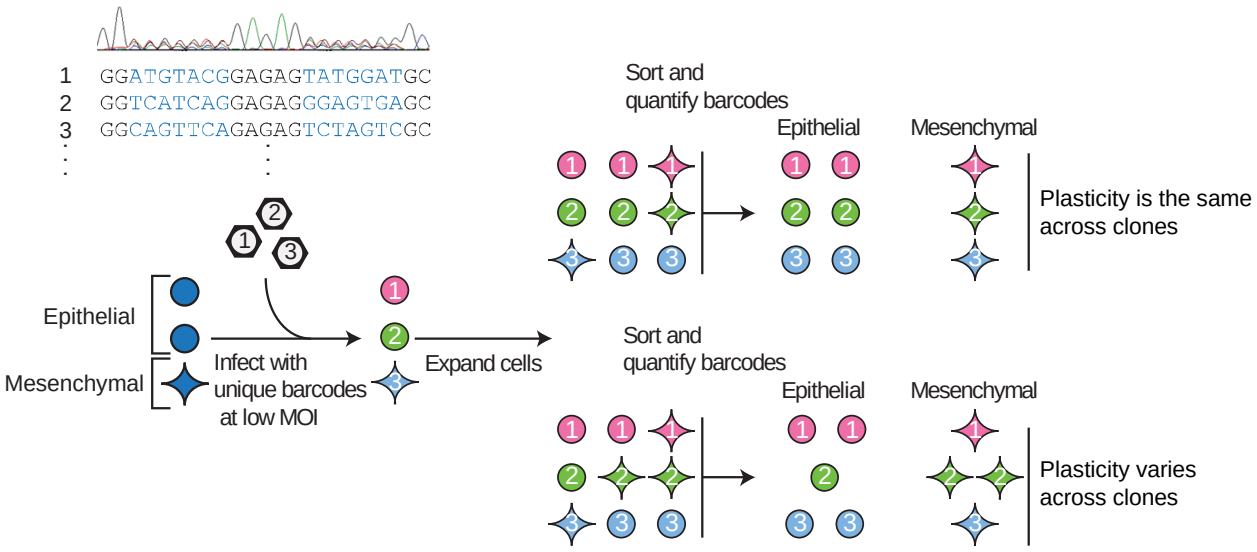


Figure 2. DNA barcode strategy to examine plasticities across clones.
DNA barcodes, stably introduced into cells, can be used to track the progeny of individual cells. Sorting these cells, and quantifying the abundance of each barcode in both cell states, will distinguish if clones vary in phenotypic plasticity. Included in the schematic is a Sanger sequence trace of the pool of barcode plasmids.

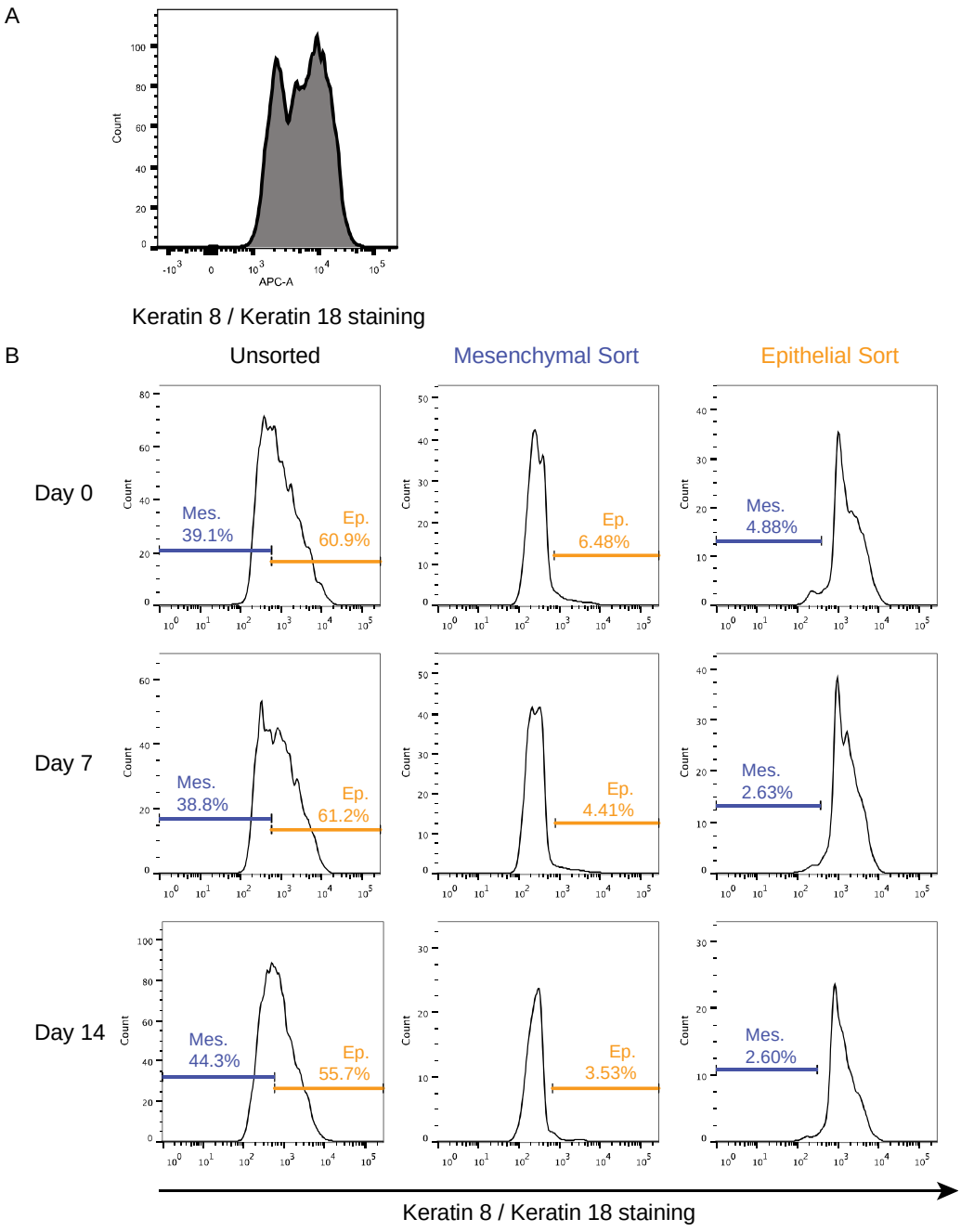


Figure 3. Fluorescence activated cell sorting based on keratin expression
(a) Histogram of barcoded MDA-MB-157 cells based on Keratin 8 / Keratin 18 expression as assayed by flow cytometry. **(b)** Post-sort flow cytometry of sorted populations; cells sorted in error were computationally subtracted from the barcode data. Shown are unsorted cells, and each sorted fraction for each time point: Keratin 8 / Keratin 18 high cells (Epithelial, Ep.) and Keratin 8 / Keratin 18 low cells (Mesenchymal, Mes.).

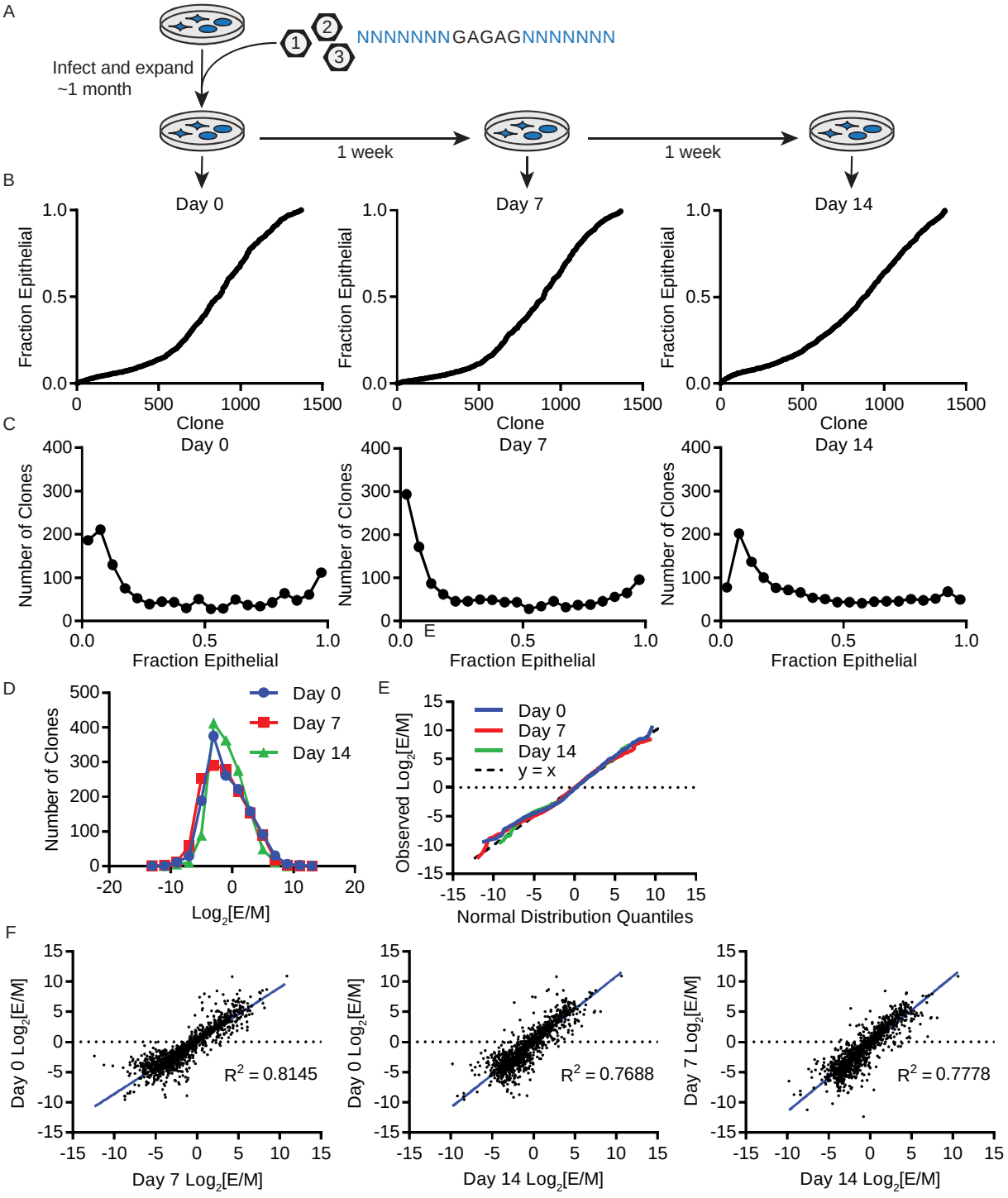


Figure 4. Phenotypic plasticity varies across clones.
(a) Cells stably transduced with DNA barcodes were expanded over ~1 month into a large population. Three times over two weeks, samples of the population were sorted by cell state, and each clone's representation was quantified in these sorted states. **(b)** The fraction of each clone in the epithelial state (fraction epithelial) is plotted for each time point, showing the diversity of phenotypic plasticities among clones. Clones are sorted in ascending order by their fraction epithelial, calculated from the

mean clone size in duplicate sorts. **(c)** Histograms of clones, binned by the fraction of each clone that is epithelial, shows the distribution of clonal plasticity. Each plot is from a different time point. **(d)** Histogram of clones binned by their $\log_2$ ratio of epithelial (E) to mesenchymal (M) cells (Log_2[E/M]), with one plot for each time point, showing phenotypic plasticity is approximately log normally distributed across clones. **(e)** Quantile/Quantile plot with the distribution of Log_2[E/M] across clones plotted against the quantiles of a normal distribution with the same mean and standard deviation, with one line for each time point. Also plotted is the line y = x (dashed line), representing a perfect normal distribution. **(f)** Each clone's Log_2[E/M] in one time point plotted against its $\log_2$[E/M] in another time point, showing that clones have stable phenotypic plasticity. The R^2 (squared Pearson correlation coefficient) is shown, and a linear regression of the data is plotted in blue. Each clone's $\log_2$[E/M] was calculated from the mean clone size in duplicate sorted E and M states.

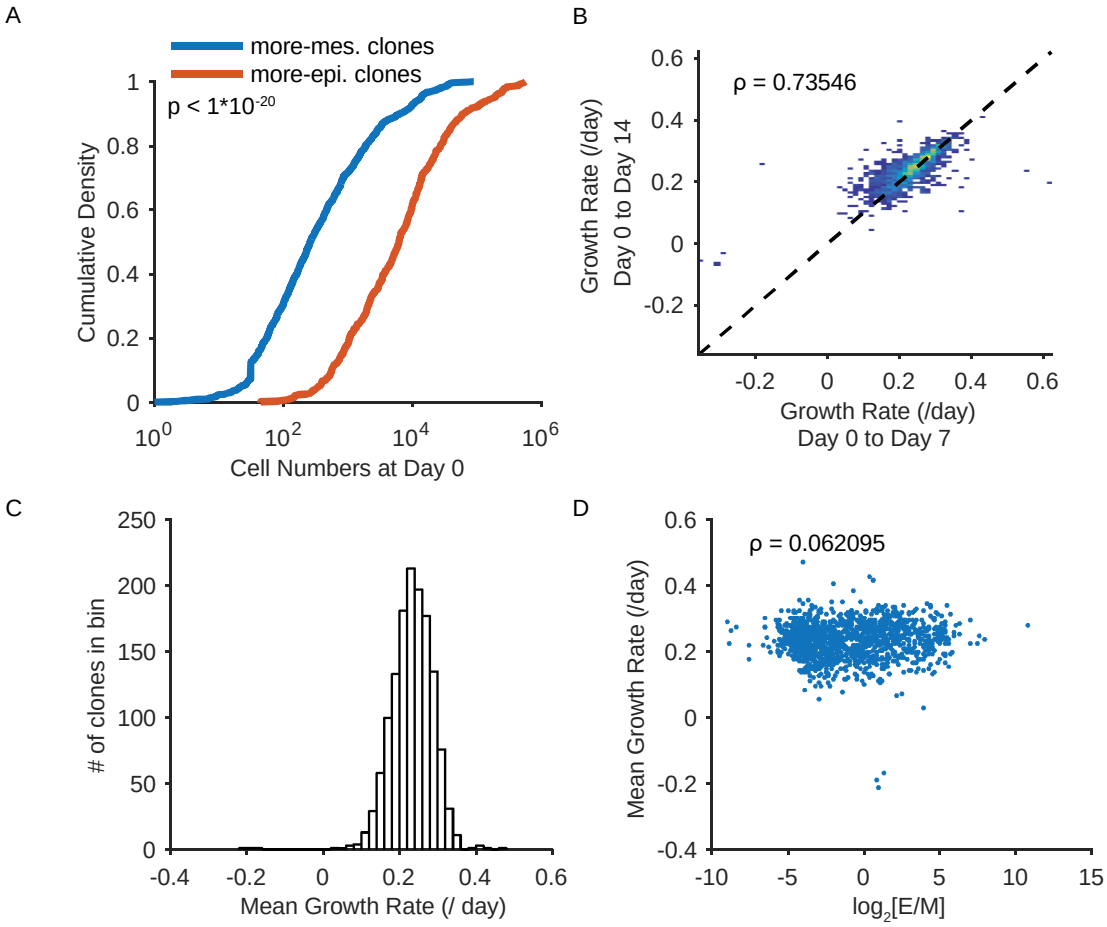


Figure 5. Growth rates of clones

(a) Estimated cell numbers at day 0 of more-mesenchymal (more-mes.) and more-epithelial (more-epi.) clones. The p value displayed is from a rank-sum test of these two groups. **(b)** Growth rate (k, per day) of clones, determined from the change in each clone's cell numbers from day 0 to day 7, or to day 14. Each dot corresponds to a clone; overlapping dots produce a different color. The Pearson correlation (ρ) is displayed, and the line y=x is plotted (dotted line). **(c)** Histogram of the average growth rate of each clone. **(d)** Plot of each clone's average growth rate vs its $\log_2$(Epithelial/Mesenchymal) ratio ($\log_2$[E/M]). The Pearson correlation (ρ) is displayed.

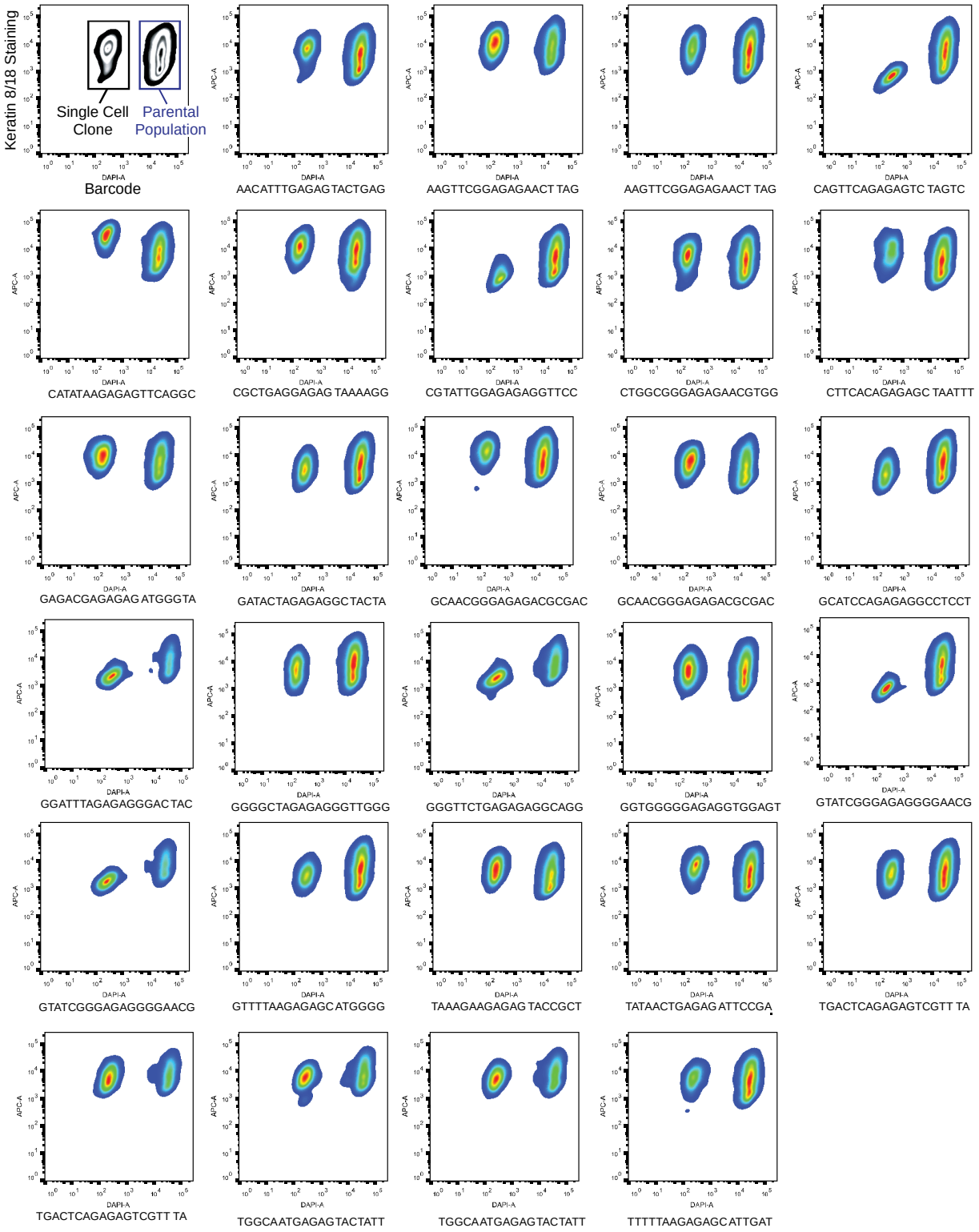


Figure 6: Flow cytometry of single cell clones

Flow cytometry of single-cell clones (see Figure 7), with Keratin 8 / Keratin 18 staining on the y axis, and the covalent stain marking the spiked-in polyclonal/parental population on the x axis. The upper-left plot is a key showing the different populations. Each clone's barcode is displayed under each clone.

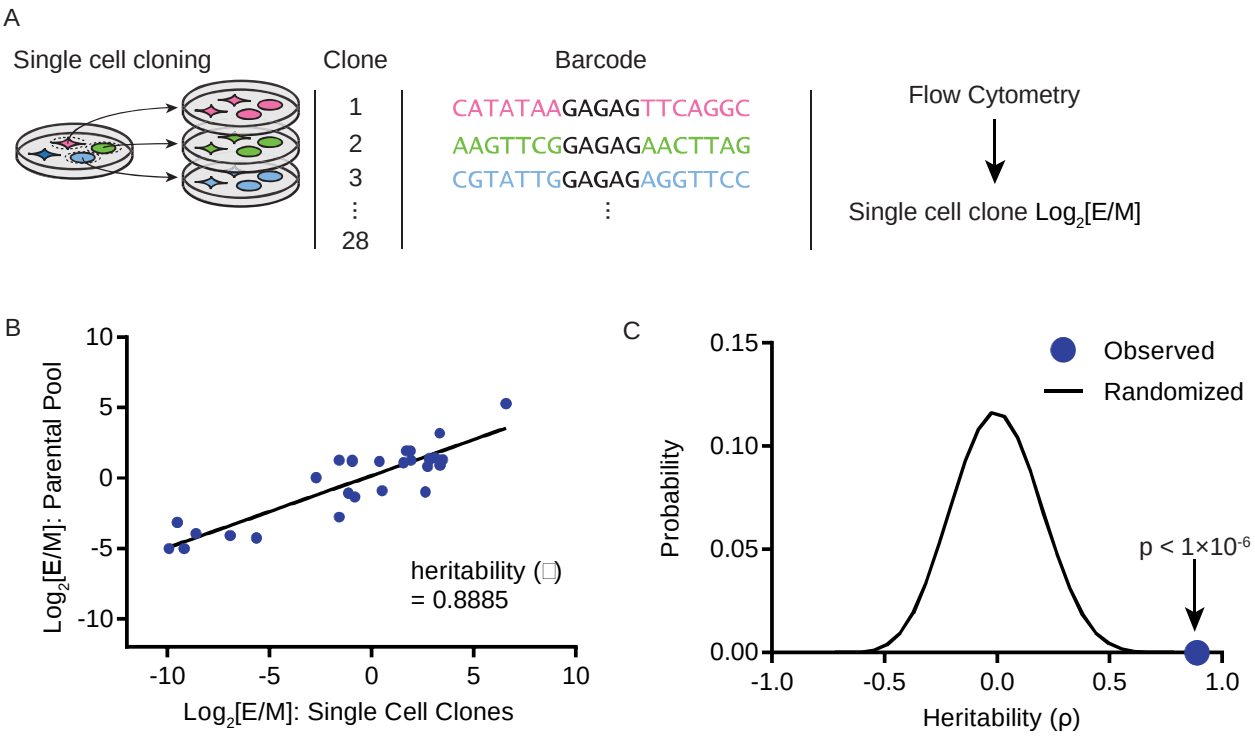


Figure 7. Phenotypic plasticity is stably inherited.
(a) Clonal subpopulations were generated from single cells, and each clone's phenotypic ratio was evaluated with flow cytometry. **(b)** Each single cell clone's $\log_2$ ratio of epithelial to mesenchymal cells (Log_2[E/M]) is plotted against the Log_2[E/M] of the same clone in the parental pooled population. The Pearson correlation coefficient is shown, estimating the narrow-sense heritability. In black is a linear regression of the data. Phenotypic ratio is a heritable phenotype. **(c)** Barcode labels were randomized 10^6 times, and the Pearson correlation was calculated for each iteration; the observed correlation (blue circle) was higher than all of the randomizations, suggesting the heritability of phenotypic ratio is not likely due to random chance.

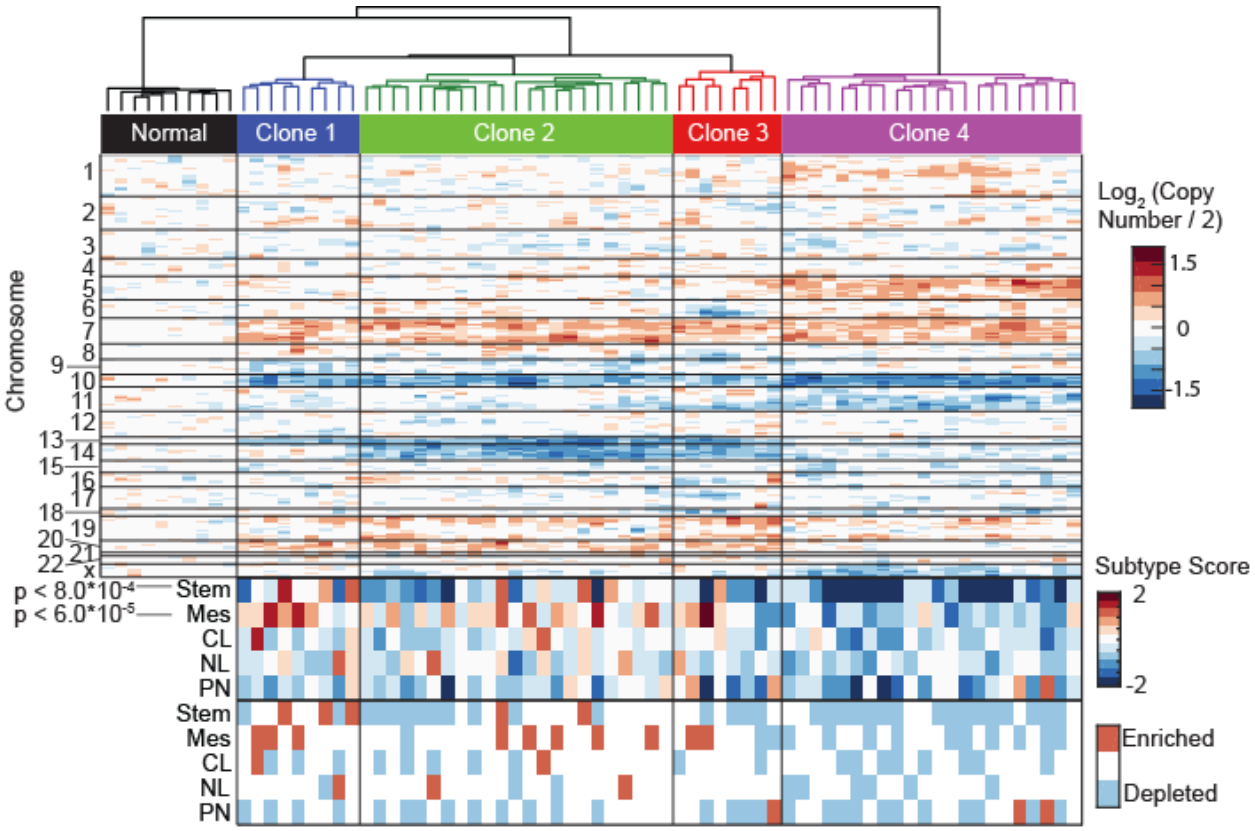


Figure 8. Phenotypic plasticity varies across clones in vivo.
At top is a heat map of copy number variation of single cells estimated from single-cell RNA sequencing data of a primary glioblastoma [18]. Shown is a heatmap of the averaged, normalized expression of a sliding window of 100 genes moving across chromosomes, revealing chromosomal gains and losses. Each value shows the estimated $\log_2$(copy number / 2) for genes in the window. Based on these data, clones were grouped by hierarchical clustering based on Ward clustering of the Euclidean distances between clones, shown above. Below, each cell was given a score based on the average expression of a set of classifier genes for different tumor subtypes[35] and a score for glioblastoma stemness [18], shown as a heatmap. CL, classical; NL, neural; PN, proneural; Stem, stem-like; Mes, mesenchymal. Kruskal-Wallis tests showed differential representation of the mesenchymal (Mes) subtype ($p<6\times10^{-5}$) and the stemness score (Stem) ($p<8\times10^{-4}$) among clones. At the bottom, each cell's subtype scores are evaluated for significance compared to the background of gene expression in that cell. Scores higher or lower than 95% of gene sets were marked as enriched or depleted.

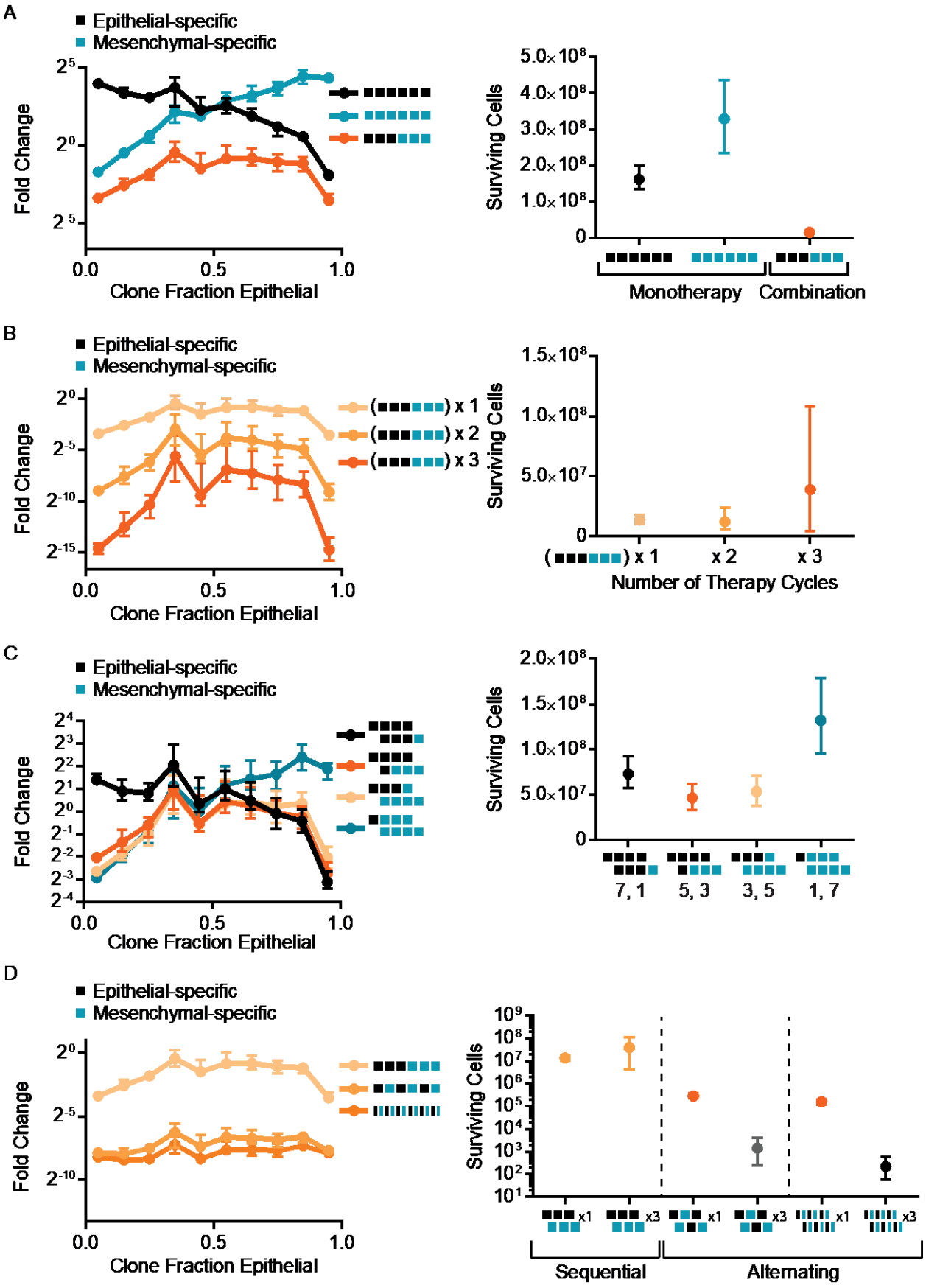

A
Epithelial-specific
Mesenchymal-specific
Fold Change
Clone Fraction Epithelial
Surviving Cells
Monotherapy
Combination
B
Epithelial-specific
Mesenchymal-specific
(■■■■■■) x 1
(■■■■■■) x 2
(■■■■■■) x 3
Fold Change
Clone Fraction Epithelial
Surviving Cells
(■■■■■■) x 1
x 2
x 3
Number of Therapy Cycles
C
Epithelial-specific
Mesenchymal-specific
Fold Change
Clone Fraction Epithelial
Surviving Cells
7, 1
5, 3
3, 5
1, 7
D
Epithelial-specific
Mesenchymal-specific
Fold Change
Clone Fraction Epithelial
Surviving Cells
x1
x3
x1
x3
x1
x3
Sequential
Alternating

Figure 9. Combination chemotherapy enriches for clones with increased phenotypic plasticity
Simulations of clones treated with different patterns of combination therapies that include mesenchymal-specific (blue squares) and epithelial-specific treatments (black squares) (see Materials and Methods). **(a-d)** Left, the median fold change in clone size during the course of treatment for clones binned by the fraction of their progeny in the epithelial state. Displayed is the median and 90%–10% range of observed medians across 500 simulations. Right, the number of cancer cells surviving at the end of the simulation for each treatment; displayed is the median and 90%–10% range of observed cell numbers across 500 simulations. **(a)** Combination therapy (orange curve) enriches for clones with increased plasticity, while monotherapies enrich for clones in predominately one or the other state. Fewer cells survive combination therapy. **(b)** Increasing the cycles of combination therapies (as shown in (a)) further enriches for clones with increased plasticity. However, resistant populations eventually emerge. The number of cycles of each combination therapy is indicated. **(c)** Different patterns of combination therapy, with varying proportions of epithelial- and mesenchymal-specific treatments, enrich for different, particular plasticities. **(d)** More rapid alternation between therapies reduces the enrichment for more plastic clones and more effectively reduces cancer cell numbers. Repeated alternating therapy also prevents the outgrowth of resistant clones, in contrast to repeated sequential therapy. The number of cycles of each therapy is indicated.

Chapter 4: Conclusions

Summary

In this work, I summarize our discovery that clones with a single breast cancer cell line display distinct cell state equilibria, or biases. Not only do clones have diverse biases, they faithfully pass on this bias to their daughter cells. This stable and heterogeneous phenotype allows for selection to act on the clones, and cause evolution of the population. In particular, as these states can serve as bet hedging mechanisms, the diversity among clones allows them to sample different bet hedging probabilities, allowing the evolution of an optimal bet-hedging strategy. Simulating a variety of conditions based on our observed data, we were able to identify treatment strategies that could slow the evolution of resistance.

I have also described here our finding of widespread purifying selection in cancers. This was revealed through the enrichment of conservative amino acid transitions, and purifying selection successfully identifying essential genes. In addition to uncovering a novel tumor evolution phenomenon, this finding also illuminates genes important for tumor growth or maintenance. As human tumors, the true target of cancer biology understanding, remain experimentally intractable, this finding allows an unprecedented view into the importance of genes in *in vivo* tumors. We extended this analysis by seeking genes and pathways under increased purifying selection in one tumor type over others, thereby excluding generally-essential genes. This identified diverse pathways under increased purifying selection in specific tumor types, such as DNA repair mechanisms in melanomas.

Future directions

The utility of purifying selection for uncovering pathways and genes important to cancer has already led to many novel hypotheses. Some of these are sure to be only findable *in vivo*, and may be very difficult to validate in pre-clinical models. For example, the apparent sensitivity of melanomas to

the inhibition of DNA repair mechanisms could be due to their unique exposure *in vivo* to ultraviolet radiation. Purifying selection could also provide false negatives; some pathways that are truly important may not be found to be under purifying selection. This could be because some pathways are resistant to purifying selection, due, for example, to the presence of multiple copies of genes, or functionally redundant paralogs, resisting a loss of functionality. Still, I hope that purifying selection may serve as a very helpful tool for hypothesis generation and preclinical validation, particularly for the development of novel therapeutics.

While we were able to show evidence of purifying selection in human tumors, our ability to detect purifying selection in single genes is limited. It may be that a better way to detect purifying selection could be constructed. Such a method could be developed by including more kinds of mutations, including nonsense, frame-shift, and possibly synonymous mutations with a predicted phenotype. In addition, a far better constructed model for the expected number of mutations can likely be constructed using more information, such as copy number, allele frequency, chromatin state, expression level, DNA methylation, and nucleotide context, all of which could affect the probability or phenotype of a mutation [1, 2]. These combined approaches might increase our ability to discern if a gene is under purifying selection. This power could also increase with the sequencing of more tumors, or, DNA sequencing of individual tumor cells.

In this work, we implemented a method comparing tumors of different types, to identify genes and pathways under increased purifying selection in certain tumor types. This allowed us to identify tumor-type specific essential genes, which were not generally essential. With increased power to identify purifying selection, or more sequenced tumor genomes, we may be able to do other, more complex analyses comparing smaller populations. This could include comparing the dependencies of tumors with distinct genotypes (eg RAS mutant, vs RAS normal cancers), allowing the development of

novel treatments for personalized medicine. These alterations, under positive selection, could induce other genes to be essential; examples of this have been found with CRISPR and shRNA screening of cancer cell lines [3, 4]. Additionally, with the right kind of information, one could divide tumors not just by tumor genotypes, but by the genotype or phenotype of the patients. It may be that some patient alleles confer unique tumor dependencies. Similarly, the tumors of patients with other conditions (such as insulin resistance) could show unique dependencies. These and other questions could be tested with purifying selection, especially with the increased power from more sequenced tumors.

Purifying selection could further be used to reveal synthetic vulnerabilities in cancers, where the loss of one gene causes another to become essential. Examples of this include the essentiality of ARID1B in an ARID1A mutant context, and the essentiality of BRM in a BRG-1 mutant context [5, 6]. In particular, if one gene is mutated, the other would then become subject to purifying selection. This would cause mutually-exclusive purifying selection; whichever gene is lost first, would cause the other to be under purifying selection. This would be most easily used to test already generated synthetic lethal hypotheses, such as those generated through *in-vitro* CRISPR screens or through understanding functional information, due to the otherwise large number of pair-wise tests required across all expressed genes. Such synthetic-lethal pairs might not just be combined to lesions affecting individual genes, but rather genes suddenly essential after the loss or gain of chromosomes or other large amplifications or deletions. As many aneuploidies are recurrent in certain tumor types (e.g. glioblastomas [7]), there could be enough information to test for such genes. This information could inform the generation of targeted therapies based on tumor genome information, particularly with the recurrence of certain chromosomal aneuploidies in certain tumor types.

An important caveat that remains unexplored is how much, if any, of the purifying selection we observed in tumors occurred late in tumorigenesis. As we have limited ability to observe mutations with

a low allelic-frequency (such as those in small clones or single cells), and are sampling only a small fraction of a tumor with a biopsy, it may be that most of the mutations we observe are in the dominant, or truncal clone [8-10]. These mutations, then, are those dragged to enrichment alongside those under positive selection. While such an event must be necessary for the formation of a tumor, it remains unclear if this is only evidence from the last clonal expansion, or if there is continual restructuring of the clones in a tumor, such that we may see evidence of late-occurring purifying selection. In fact, it may even be that some of the signal of purifying selection we observe could be coming from pre-malignant lesions, in those contexts where normal cells are likely to be frequently mutated, such as melanocytes (ultraviolet radiation) and lung epithelia (smoking). These and other questions may be robustly answered in the future with the development of single-cell DNA sequencing, if this technique is eventually capable of confidently identifying mutations in single cells. Such sequencing, applied to single cells late in tumors, will allow us to see many more mutations in each tumor. This is likely an enormous number of mutations, as each cell acquires mutations during each division, invisible in the sequencing of the population [11]. In addition to giving us much more statistical power, this could be used to reveal if there are differences between those genes under purifying selection in early tumor development (truncal mutations) or later tumor development (low allelic-frequency mutations). Similarly, the sequencing of normal cells will let us determine the mutational burden of normal cells in tissues, potentially revealing the presence of purifying selection in these cells. This would generate a large amount of important information, able to be used to define the essentialities in normal tissues. Comparing this to those genes under purifying selection in tumors would allow the finding of truly tumor-specific vulnerabilities.

The striking heritability of each clone's differentiation bias--- somehow quantitatively encoded in each of its daughter cells--- defies an obvious mechanism. Although genetic or epigenetic

mechanisms are both known to encode heritable phenotypes, I find it hard to understand how they would encode such a quantitative, single cell-level, heritable phenotype.

It is possible that this phenomenon, the stable encoding of cell state bias, could actually not be a cancer-specific phenomenon, and occur in normal tissues. A similar experiment performed on normal mammary epithelial cells revealed a similar heterogeneity in differentiation bias, although they did not determine whether this bias was heritable [12]. Similar lineage biases were observed in the hematopoietic system [13-16]. Still, our observation here is distinct in that, as far as we could tell from our single cell cloning, each cell faithfully passes on its bias to both of its daughters.

It remains to be seen how widespread this phenomenon (the heritability of lineage bias) is across tumors. The limited single cell mRNA sequencing analysis presented here does suggest that, in contrast to the rather laborious single cell tracking experiment, single-cell sequencing could answer this question. As this technology becomes cheaper and more reliable, this question may be answered, particularly if the sequencing of both the DNA and mRNA from a single cell becomes feasible, allowing the assignment of each cell to a clone and cell state.

The behaviors we described suggest that circumventing tumor heterogeneity to better treat patients will be more difficult than expected. Already determined to be more difficult by the observation of bidirectional plasticity, this work shows that tumors have a much greater ability to optimize plasticity through evolution than we expected. Our modeling based on the parameters we observed in this particular cell line, namely each clone's growth rate and lineage bias, did suggest that it may be possible to optimize combination therapies to suppress the outgrowth of resistant clones. However, tumors differ widely in their characteristics, and it seems unlikely that one strategy will be effective for treating all tumors. Circumventing tumor heterogeneity remains a challenge.

References

1. Alexandrov, L.B., et al., *Signatures of mutational processes in human cancer.* Nature, 2013. **500**(7463): p. 415-21.
2. Lawrence, M.S., et al., *Mutational heterogeneity in cancer and the search for new cancer-associated genes.* Nature, 2013. **499**(7457): p. 214-8.
3. Hart, T., et al., *High-Resolution CRISPR Screens Reveal Fitness Genes and Genotype-Specific Cancer Liabilities.* Cell, 2015. **163**(6): p. 1515-26.
4. McDonald, E.R., 3rd, et al., *Project DRIVE: A Compendium of Cancer Dependencies and Synthetic Lethal Relationships Uncovered by Large-Scale, Deep RNAi Screening.* Cell, 2017. **170**(3): p. 577-592.e10.
5. Hoffman, G.R., et al., *Functional epigenetics approach identifies BRM/SMARCA2 as a critical synthetic lethal target in BRG1-deficient cancers.* Proc Natl Acad Sci U S A, 2014. **111**(8): p. 3128-33.
6. Helming, K.C., et al., *ARID1B is a specific vulnerability in ARID1A-mutant cancers.* Nat Med, 2014. **20**(3): p. 251-4.
7. Li, B., et al., *Genomic estimates of aneuploid content in glioblastoma multiforme and improved classification.* Clin Cancer Res, 2012. **18**(20): p. 5595-605.
8. Nowell, P.C., *The clonal evolution of tumor cell populations.* Science, 1976. **194**(4260): p. 23-8.
9. Yates, L.R., et al., *Subclonal diversification of primary breast cancer revealed by multiregion sequencing.* Nat Med, 2015. **21**(7): p. 751-9.
10. Gerlinger, M., et al., *Intratumor heterogeneity and branched evolution revealed by multiregion sequencing.* N Engl J Med, 2012. **366**(10): p. 883-92.
11. Tomlinson, I., P. Sasieni, and W. Bodmer, *How many mutations in a cancer?* Am J Pathol, 2002. **160**(3): p. 755-8.
12. Nguyen, L.V., et al., *Clonal analysis via barcoding reveals diverse growth and differentiation of transplanted mouse and human mammary stem cells.* Cell Stem Cell, 2014. **14**(2): p. 253-63.
13. Chang, H.H., et al., *Transcriptome-wide noise controls lineage choice in mammalian progenitor cells.* Nature, 2008. **453**(7194): p. 544-7.
14. Lu, R., et al., *Tracking single hematopoietic stem cells in vivo using high-throughput sequencing in conjunction with viral genetic barcoding.* Nat Biotechnol, 2011. **29**(10): p. 928-33.
15. Naik, S.H., et al., *Diverse and heritable lineage imprinting of early haematopoietic progenitors.* Nature, 2013. **496**(7444): p. 229-32.
16. Cheung, A.M., et al., *Analysis of the clonal growth and differentiation dynamics of primitive barcoded human cord blood cells in NSG mice.* Blood, 2013. **122**(18): p. 3129-37.