# Evolution and Vulnerabilities of Somatic Copy Number Alterations in Cancer

by

William J. Gibson

S.B Biological Engineering
S.B Philosophy
Massachusetts Institute of Technology, 2010

SUBMITTED TO THE DEPARTMENT OF HEALTH SCIENCES AND TECHNOLOGY IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY IN DIVISON OF HEALTH SCIENCES AND TECHNOLOGY

AT THE

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

AUGUST 2015 [September 2015]

Signature of Author: Signature redacted
Department of Health Sciences and Technology
August 1, 2015

Certified by: Signature redacted
Rameen Beroukhim
Assistant Professor, Department of Medicine, Harvard Medical School
Thesis Supervisor

Accepted by: Signature redacted
Emery N. Brown
Professor of Computational Neuroscience and Health Sciences and Technology
Director, Division of Health Sciences and Technology

Evolution and Vulnerabilities of Somatic Copy Number Alterations in Cancer

by

William J. Gibson

SUBMITTED TO THE DEPARTMENT OF HEALTH SCIENCES AND TECHNOLOGY ON AUGUST 1, 2015 IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF DOCTOR OF PHILOSOPHY IN HEALTH SCIENCES AND TECHNOLOGY

## ABSTRACT

Cancer is a Darwinian evolutionary process in which rounds of mutation and selection lead to increasingly fit clones. Understanding how cancers evolve and in particular how they form lethal metastases is critical to informing the design of new therapies. In the first part of this thesis, we performed whole-exome sequencing of paired endometrial cancer primaries and metastases to explore how tumors sample the genetic landscape. We find that mutations of PTEN, and TP53 occur early in the evolution of endometrial cancers, whereas BAF complex alterations occur late. We identified novel recurrent alterations in primary tumors, including mutations in the estrogen receptor cofactor *NRIP1* in 12% of patients.  Phylogenetic analyses in cases with multiple metastases indicated these metastases typically arose from one lineage of the primary tumor. We observed subclones within the sequenced part of the primary tumor that seeded metastases. We document extensive heterogeneity and genomic disruption across the various clinical stages in endometrial cancers. In the second part of this thesis we explore how the widespread genomic disruption observed in tumors can generate therapeutic opportunities. We use data from genome-scale shRNA screens to perform an unbiased analysis of all copy-number: gene dependency interactions. We identify a class of interactions called CYCLOPS interactions in which genomic loss of essential genes sensitizes cancer cells to their further suppression. We explore the properties of CYCLOPS genes and show that the splicing factor SF3B1 is one of them. Biochemical analyses showed that cancer cells harboring hemizygous loss of SF3B1 lack a buffer of SF3B1 present in cells whose SF3B1 locus is intact. These data provide evidence for the utility of developing non-oncogene targeted therapies as a means of advancing cancer therapeutics.

Thesis Supervisor: Rameen Beroukhim
Title: Assistant Professor, Department of Medicine, Harvard Medical School

## Table of Contents

# Acknowledgments

This thesis would not have been possible were it not for the mentorship, love and support of many people in my life.

I am deeply grateful to my mentor Rameen Beroukhim. Nearly all of the ideas and results in this thesis were the result of extensive back-and-forth discussions with him. His active concern for the well-being, intellectual growth and career success of all members of his lab is remarkable. Rameen has consistently challenged me and guided me in directions that have allowed me to grow the most. For that I am very thankful.

I want to thank the members of my thesis committee Leonid Mirny, William Hahn and Catherine Wu for their scientific and project advice. Their input was key to the progression of this work. I would also like to thank Franziska Michor, Robin Reed, Matthew Meyerson and Gad Getz for helpful discussions.

My collaborators in Norway have been truly spectacular to work with. Helga Salvesen has been an inspiration both as a physician-scientist and as a kind person. I also want to thank Erling Hoivik whose scientific insight was crucial to our project's success.

I was lucky to have joined such a collaborative and energetic lab. I want to thank all of my colleagues in the Beroukhim lab for providing such a scientifically rigorous environment without taking ourselves too seriously. In particular I want to thank Brent Paolella who has a wonderful collaborator and teacher. His advice on graduate school has helped me so much in finding my way.

I admire my colleagues at the Broad for their infinite patience with me. In particular, I want to thank Amaro Taylor-Weiner, Mara Rosenberg, Aaron Chevalier, Andrew Cherniak and others for making my experience there so fun and productive.

Several people at the MD/PhD program made my life significantly easier navigating this crazy place. I want to thank Amy Cohen who had to hold my hand through the F30 process. I also want to thank Steve Blacklow who recommended Rameen as a mentor and Loren Wolensky the tremendous energy he brings to this program. I

also want to thank all of the people at HST who have worked endlessly to make it such an amazing program. I feel crazy lucky to have had this opportunity.

My friends have made all the difference in keeping me sane and happy during these years. Thank you. And special thanks in particular to Lauren Klinker for her love, support and edits on this thesis.

Finally, I'd like to thank my parents Beth and Mike Gibson, my step-mom Sue Gibson, my brother Mike and my sister Grace. Your support over the years has meant everything to me.

# Chapter 1: Introduction

Cancer is fundamentally a loss of tissue organization. During development, individual cells must coordinate their growth and death to allow for the proper macroscopic organization required to create a functioning multicellular organism. In cancer, a single rogue cell and its descendants "forget" their role in tissue organization and grow uncontrolled. Eventually, these cells destroy tissue structures and compromise physiologic processes to cause the death of the entire organism.

## 1. Cancer as a disease of the genome

There are 14.2 million new cases and 8.2 million deaths due to cancer in the United States each year[1]. The proportion of deaths due to cancer has increased steadily as life expectancy has increased. While environmental exposures are linked to the development of cancers, many cancers may not be preventable. As humans defeat many of the diseases of the past, cancer looms as an inexorable probabilistic threat to all humans. These trends suggest that defining the ways in which cancer develops and investigating novel vulnerabilities will continue to benefit humans for the foreseeable future.

Cancers evolve from normal tissues through a multi-step process of mutation and selection. Hippocrates proposed that the body was composed of four fluids: black bile, yellow bile, blood and phlegm. For 1300 years since Hippocrates, the world believed that an excess of one of these humors, black bile, caused cancer. This humoral theory persisted until Stahl and Hoffman proposed that another liquid, "lymph" this time, was the cause of cancer. They proposed that cancer consisted of degenerating lymph that had fermented in parts of the body. Even the father of surgical oncology, John Hunter agreed that cancers arose from contaminated lymph. He reasoned that surgical resection of tumors did not cure patients because the cancerous fluid would always seep back into the wound like sap in a tree. Finally, in 1838, Johannes Müller showed that cancers are derived from normal cells, not fluid.

By the mid nineteenth century, scientists agreed that cancers are composed of cells from the host, but it was not clear whether they arose spontaneously or due to foreign influence. William Russell provided evidence in 1890 that unidentified intracellular organisms were found ubiquitously in cancerous tissues[2]. Today's pathologists believe that these "Russell bodies" represent accumulated immunoglobulins[3]. In 1926, Glover reported the isolation of a pleomorphic bacterium from a wide variety of cancers, in a work that was later refuted. Throughout the 1950s Livingston reported on the isolation of a similar universal cancer bacterium[4], although this work was also discredited. Finally, in 1984 Warren and Marshall reported on a spiral-shaped bacterium, *Helicobacter Pylori*, in the majority of patients with gastric ulcers[5]. The gastric ulcers in these patients often transformed into malignant cancers. Warren and Marshall received the Nobel Prize for establishing the first causal link between infection with a bacterium and cancer. By the time of Warren and Marshall's report, a separate line of investigation had clearly established that infection with viruses could aid in transforming normal human cells into cancer cells[6,7]. Recent studies have revisited the microbe hypothesis with new genetic tools[8] to establish novel associations between certain cancer types and microbial species[9].

While in rare cases, infectious agents can cause cancer, tumors principally arise due to somatic alterations to the genome. In 1975, scientists had established three tantalizing associations with the development of cancer: infection with certain viruses[10], family history of cancer and exposure to mutagens[11,12]. It was only in 1976 that Stehelin, Bishop and Varmus showed that somatic mutation of oncogenes was the dominant cause of cancers, thereby synthesizing these findings[13].

Next, scientists established that tumorigenesis is a multistep process. In 1985, Vogelstein and colleagues showed that cancers are born from a single lineage in the host (ie that cancer cells have a monoclonal origin)[14]. In 1988, Vogelstein and colleagues demonstrated that colonic polyps (early cancers) contained only some of the oncogenic lesions found in mature colon cancers[15]. These observations suggested that mature tumors require additional genetic "hits" beyond those that are required to form benign tumors. Several other lines of evidence indicated that mature tumors are derived from benign tumors. Pathologists had long observed tumors in various "stages" at the time of biopsy. This phenomenon was observed separately in histologic examination of colonic polyps and tumors, as well as benign nevi, malignant and metastatic melanoma biopsies[16].

Together, these observations indicated that the majority of cancers are born from a single lineage of cells whose genes have been altered to promote cellular growth. Since Vogelstein's demonstration of step-wise genetic evolution from pre-malignancy to frank carcinomas, scientists have studied the evolution of cancers in hopes that deciphering these patterns may help aid in the design and selection of therapeutics. A picture of primary cancer evolution has emerged that is both rule-based and bewilderingly complex.

If cancers arise through alterations to genes, several questions are salient: (i) What processes typically lead to the alteration of genes in cancer? (ii) Which pathways are the protein products of these genes typically involved in? and (iii) To what extent do all cells in the same tumor share the same gene alterations?

## 2. Alterations in cancer genomes

A primary focus of cancer research has aimed at identifying the ways in which cancer cells disrupt the genes they carry to promote cellular growth. In some cancers such as melanoma and lung cancers, the dominant genomic lesion is point mutation. The sources of these mutations, ultraviolet-induced pyrimidine dimers and smoking associated formation of DNA-adducts respectively have been well appreciated for decades[17]. Other environmental associations with mutagenesis, such as exposure to aflatoxin are well-established but collectively cannot explain the diversity of mutational signatures present in human tumors. Early evidence that cancer could be caused by mutations in genes responsible for the fidelity of DNA replications came from the study of families with predisposition to colorectal cancer. Mutations in DNA mismatch repair enzymes (*MSH2*[18], *MLH1*[19] and *MSH6*[20]) were found in the majority of families with Lynch syndrome. The tumors of these patients typically exhibited microsatellite instability (MSI), a consequence of the lost ability to repair DNA mismatches occurring during replication.

Systematic analyses of mutational signatures in cancer have recently highlighted several new sources of point mutation. One recent survey identified more than 20 different mutational signatures active in human tumors[21]. Mutations of the epsilon subunit of DNA polymerase cause a tremendous number of mutations in a subset of endometrial[22] and colorectal[23] cancers. Overactivity of the APOBEC family of cytidine deaminases is associated with a mutational signature present in 14.4% of cancers[21,24,25]. The APOBEC signature occurs across diverse cancer types including breast, bladder, and cervical cancers. One mutational signature describes the location rather than the nucleotide context of mutations. Hundreds to thousands of mutations are sometimes found on the same strand in very close proximity, a phenomenon called "kataegis". The majority of identified mutational signatures remain unexplained.

The predominant genomic lesion in other cancers is structural alterations to the genome, caused by genomic rearrangements. A subset of these rearrangements cause somatic copy number alterations (SCNAs). Some cancers are caused by a limited number of genomic rearrangements. The best known of these, perhaps is the *BCR-ABL* translocation that defines Chronic Myelogenous Leukemia genomes[26]. More recently, widespread structural rearrangements resulting from a single chromosomal shattering were demonstrated in a phenomenon called chromothripsis[27]. A structural analogue to kataegis called "chromoplexy" occurs in prostate cancers whereby multiple rearrangements across chromosomes occur in a single event[28].

Other cancer genomes exhibit gain or loss of genomic material encoding genes whose protein products affect the growth or survival of cancer cells. These somatic copy number alterations (SCNAs) are the predominant means of genomic disruption in some cancers, particularly those with low rates of mutation[29]. For instance, TP53 mutant ovarian tumors (96%) typically carry few other mutations in recurrently mutated genes, but have an average of 12.3 focal SCNAs involving recurrently altered regions[30]. Some of the genes most frequently involved in SCNAs have become the targets of effective therapeutics[31,32]. Because a single SCNA may affect many genes, deciphering the selective forces that generate SCNAs at a given locus beyond the background rate requires the study of many cancer genomes. This remains an active area of research. Further study of the structural alterations to genomes will likely reveal many more therapeutically tractable vulnerabilities.

Cancer cells may still evolve methods of altering gene function without altering their DNA content. Alterations to the epigenome are common across cancers and make up the majority of heritable alterations in some cancer types. One exciting early report suggested that hypomethylation of oncogene promoters was found specifically in cancer tissues[33]. Other laboratories did not replicate these findings. Instead, cancer epigenomes tend to display global hypermethylation. Hypermethylation of tumor suppressor promoters was first documented in 1989 in the retinoblastoma gene[34]. Recently, one of the mechanisms that can cause the CpG Island Hypermethylator Phenotype (CIMP) was demonstrated to be mutation of the enzyme *IDH1*[35]. The mutated *IDH1* produces an oncometabolite 2-HG, which itself is sufficient to establish CIMP. Instead of altering their methylation profiles, cancer cells may modify the function of their chromatin to drive oncogenic gene expression profiles. The most striking example of driver alterations in chromatin remodeling comes from pediatric rhabdoid tumors, in which 100% of cases exhibit mutation of *SMARCB1*, a member of the BAF chromatin-remodeling complex. Genes encoding components of the chromatin themselves are recurrently mutated in cancer; mutations in the gene encoding Histone H3.3 occur in one third of pediatric glioblastomas[36].

## 3. Evolution in primary cancers

Both the sets of genomic disruption and the pathways altered in the formation of primary tumors are diverse. Few cancers are acutely fatal, but in the absence of intervention, most will eventually evolve to kill their host. Whether facing therapy-driven selective pressures or competition within a nutrient-limited microenvironment, cancer cells frequently continue to evolve after the final selective sweep preceding initial transformation. The latency between presentation and lethality, and the changes therein present a possible window for understanding and intervention. Understanding how cancers progress from small, germinal neoplasms to lethal tumors is critical to decreasing the morbidity and mortality caused by cancer.

Cancer cells continue to evolve past the final selective sweep even in the absence of selective pressures. The tendency for cancers to evolve was observed clinically long before the tools to systematically assess evolution became available. In 1928, Maurice Richter described a case in which a 46 year old shipping clerk presented with a "chronic lymphatic leukemia [and] subsequently developed a rapidly growing, malignant tumor arising from reticular and reticuloendothelial cells of the lymphoid tissues". The now eponymous Richter's syndrome occurs in 5% of Chronic Lymphocytic Leukemia (CLL) patients[37], and describes the rapid transformation of CLL to Diffuse Large B-cell Lymphoma (DLBCL). Similarly, clinicians have long known that myelodysplastic syndrome can convert to acute myeloid leukemia (AML)[38].

However, the evolution of cancers in the face of therapy-induced selective pressures is even better documented. The first documented successes with chemotherapy were quickly reversed when relapses invariably occurred[39]. Response and resistance to therapy is a nearly universal theme in the treatment of cancer patients. Studies of the mechanisms of resistance that cancers evolve can yield important information about the design of therapies. For example, resistance develops in CML patients treated with imatinib monotherapy. Sequencing the *BCR-ABL* gene in resistant samples revealed frequent mutations in the *ABL* kinase domain[40]. The structural insights from these resistance mutations allowed for the design of second-line tyrosine-kinase inhibitors that successfully inhibited 14 of 15 resistance mutations[41]. Resistance to the EGFR inhibitor erlotinib is also frequently caused by point mutation. The T790M "gatekeeper" mutation in *EGFR* blocks erlotinib binding, but two recent clinical trials of novel inhibitors demonstrated remarkable success in treating T790M mutant cancers[42,43]. Acquired resistance to BRAF inhibition in melanoma has revealed alterations in the genes encoding other members of the MAPK pathways such as MEK and ERK. Therapies that anticipate these resistance mechanisms have improved outcomes for patients in clinical trials[44].

All of these resistance mechanisms have illustrated mechanisms by which cancer cells can allow for the continued flux of signals through the original oncogenic pathway in the presence of targeted therapy. However, resistance to blockade of oncogenic pathways by the activation of an entirely orthogonal pathway has been observed. One of the most striking examples of this phenomenon occurs in BRAF mutant melanomas in the face of MAPK blockade wherein amplification of the lineage specific transcription factor *MITF* occurs in a subset of resistant clones[45].

In a select few cases, the evolutionary trajectory of cancers has been characterized even past the death of the original host. A recent study of canine transmissible venereal tumor showed that complex cancer karyotypes remain stable over millennia of separate evolution[46]. A recent genome-wide survey of several strains of the extensively passaged HeLa cell line revealed a structurally stable genome as well[47].

## 4. Evolution of metastasis

While understanding primary cancer evolution is important, metastases account for approximately 90% of cancer mortality[48]. Metastasis to distant organs requires that tumor cells execute a complex set of actions termed "invasion-metastasis cascade". The first step in this cascade requires that cancer cells break free of their attachment to surrounding cancer and stromal cells. After the loss of adhesion, metastasis can occur by hematogenous or lymphatic spread. In hematogenous spread, cancer cells must migrate to and intravasate into a nearby blood vessel. The cells are then swept into the circulation where they must continue to survive in the presence of new biophysical forces and the absence of tissue attachment. The surviving cells may then extravasate out of the blood vessel and into a new target site. Metastasis is complete when the tumor cells have successfully invaded the foreign tissue and are able to continue to proliferate.

Metastasis of epithelial cancers starts with local invasion, wherein tumor cells are able to migrate past the basement membrane. The basement membrane serves an important barrier function in normal tissues, separating the luminal surface from the deeper regions containing blood vessels and lymph nodes. Cancers cells often secrete matrix metalloproteinases such as MMP-2[49] or MMP-9[50] to degrade this barrier, and travel through either singly or in concert. In other cancers, tumor cells coopt adjacent stromal cells into secreting matrix metalloproteinases on their behalf[51,52]. The dissolution of the basement membrane allows cancer cells to pass an important physical barrier and to release growth factors sequestered in the extracellular matrix[53]. The integrity of the basement membrane is so important to restricting the growth of carcinoma cells from their primary site that most pathologists will deem any hyperplastic growths that do not breach the basement membrane *benign*. Indeed, colorectal tumors with loss of basement membrane exhibited a far higher likelihood of metastasis in the five years following surgical removal compared to colorectal tumors with intact basement membranes[54].

After invading their surrounding stroma, cancer cells must intravasate into the bloodstream. Intravasation is facilitated by pro-angiogenic programs that cancer cells enact by secreting cytokines, or chemical signals, into the local environment. One of the most widely studied cytokines released by cancer cells is vascular endothelial growth factor (VEGF), which initiates the outgrowth of blood vessels to supply oxygen to the growing tumor[55]. This process is facilitated in part by the pro-antiogenic programs that cancer cells enact. The vessels recruited to tumors often have a permissive or "leaky" endothelium characterized by absence of pericyte coverage[56]. The proximity of permissive blood vessels to the primary tumor encourages the hematogenous spread of cells as they lose adhesion to the primary tumor.

Once cancer cells have entered the circulation, they encounter hydrodynamic forces to which they are poorly accustomed. Experimental models have demonstrated an increased probability of forming metastases when circulating tumor cells (CTCs) break off from the primary as clumps of cells[57] or attract platelets to accompany them. The lifetime of CTCs in the circulation is short; CTCs are often trapped in small diameter capillaries (3-8 μm) due to their large (~20 μm) diameters and relatively inflexible membranes compared to erythrocytes (~7 μm)[58]. These observations suggest that after entering the venous circulation, most CTCs become lodged in the capillary bed of the lungs within minutes, barely enough time for them to sense their loss of attachment to surrounding cells.

Having landed at a distant site, cancer cells must extravasate from the vessel into the foreign tissue. This process requires cancer cells to execute a complex set of interactions with the adjacent endothelial cells and pericytes. Alternatively, cancer cells may lodge in a capillary and continue to divide. The growth of a cluster of cancer cells within the vessel lumen can disrupt the structural integrity of the surrounding vessel, allowing cancer cells to physically break through the endothelial layer and push their way into the stroma of the foreign tissue.

In order to become a clinically detectable metastasis, cancer cells that have successfully invaded a foreign tissue must then found a colony of cells that continue to proliferate. Several lines of investigation suggest that colonization is the most stringent bottleneck for the formation of clinically detectable metastases. In one experiment, cells that had taken up fluorescent nanoparticles were injected into the portal circulations of mice[59]. When the cells divided, the fluorescence intensity of the two daughter cells would be halved. Examination of the mouse livers up to 11 weeks later revealed no metastatic colonies, but many single cancer cells whose fluorescence intensity was unchanged from the time of injection. When extracted from the livers and placed in a tissue culture dish, these cells were able to proliferate once more. These data indicate that cancer cells that have settled in a foreign tissue are rarely able to establish colonies, but often remain dormant.

Most human cancer cells remain dormant after metastatic dissemination. Because carcinomas originating in epithelial tissues stain strongly with antibodies against epithelial cytokeratins, they can be sensitively identified in otherwise uniformly mesenchymal tissues such as the bone marrow. Indeed, approximately 30% of breast cancer patients harbor hundreds of micrometastases in their bone marrow[60], yet less than half of these women will develop clinically apparent metastases within six years of diagnosis.

One strategy that allows cancer cells to accomplish many of the tasks in the invasion-metastasis cascade may be the epithelial-to-mesenchymal transition (EMT). During embryogenesis, cells within the developing embryo must transition into distinct cell types, sometimes requiring an EMT. During gastrulation for instance, ectodermal cells arrayed as sheets of epithelia must detach and migrate towards the center of the embryo to spawn the mesoderm[61]. In order to accomplish this feat, the epithelial cells of the ectoderm must undergo an EMT.

Fundamentally, EMT requires the loss of epithelial markers and the expression of mesenchymal markers. Principally, epithelial cells must lose E-Cadherin expression when undergoing an EMT[62]. E-Cadherin molecules from adjacent epithelial cells normally bind to one another to link the cells together[63] and sequester Beta-catenin molecules at the plasma membrane. When E-cadherin is repressed, it is replaced by N-cadherin[64,65], a cell-surface molecule expressed by mesenchymal cells that participates in similar homophilic interactions[66]. The presence of N-Cadherin on the plasma therefore increases the affinity of cancer cells for the surrounding stroma. Additionally, loss of E-

Cadherin allows beta-catenin to translocate to the nucleus and activate growth signaling. Indeed, tumor cells undergoing EMT exhibit high concentrations of nuclear beta-catenin.

EMT has been observed at the edges of tumors, where heterotypic interactions with "reactive" stroma can induce the transition with growth factors such as TGF-Beta[67,68]. Xenografted human breast cancer cells in immunocompromised mice stain positively with human-specific antibodies to the mesenchymal markers vimentin and αvβ6 integrin specifically at the invading edge of the tumor[69]. These experiments made clear that mesenchymal cells observed at the tumor margins often represented cancer cells themselves and not stromal cells.

Recent studies have demonstrated compelling evidence of EMT both in primary tumors and in CTCs isolated from breast cancer patients. Yu et al. used RNA in-situ hybridization assays to probe CTCs for expression of the 3 pooled mesenchymal markers and 7 pooled epithelial markers[70]. They were able to detect cancer cells that expressed both mesenchymal and epithelial markers within primary breast tumors. They also found that the proportion of cancer cells with mesenchymal expression patterns was greater in CTCs compared to cells in the primary tumor. Interestingly, the proportion of CTCs with predominantly mesenchymal transcripts varied in response to treatment. Further studies are required to determine the prognostic information of mesenchymal markers in CTCs and the biological underpinnings of mesenchymal changes in response to therapy.

Some of the most informative clues we have about the cellular physiology of metastasis come from clinical experience. Cancers originating in certain tissues display consistent biases in the tissues to which they eventually metastasize. This tissue tropism was first described in 1889 by Piaget in a survey of the sites of breast cancer metastasis[71].

Much of the observed tissue tropisms can be explained by anatomy. For example, two of the most common sites of metastasis across all cancers are the lungs and the brain. A simple explanation for this observation is that the pulmonary circulation offers the first capillary bed in which cancer cells entering the venous circulation may become trapped. Similarly, the prevalence of brain metastases can be explained the fact that the brain receives approximately 25% of the blood flow in the body. However, other tissue tropisms are more puzzling. Uveal melanoma for instance, almost always metastasizes to the liver first[72]. The "seed and soil" hypothesis states that favorable interactions between tumor cells, the "seeds", and certain foreign tissues, the "soil", are required for successful metastatic colonization. These forces are collectively responsible for the tissue tropisms we observe.

The requirements for a certain target tissue to be hospitable for a cancer cell are still poorly defined. The observation that breast cancers rarely metastasize to the contralateral breast[73] indicates that stromal similarity to the primary organ is not the only requirement of a fertile soil. One set of observations suggests that an inflamed stroma provides an environment rich with growth factors that enable cancer cells to colonize. During laparoscopic surgery, a port is inserted through the skin of the patient, creating a small wound with local inflammation. Port site metastasis is a recognized complication of these surgeries wherein cancer cells colonize the wound site[74]. Similarly, a review of oral metastases found that many occur after traumatic dental extractions at the site of the extracted tooth[75]. Together, these studies seem to indicate that inflammation produces an environment that is hospitable for the proliferation of metastatic cancer cells.

## 5. Genetic tools to study metastasis

One approach to studying metastasis is to compare the genetic features of tumors that metastasize to those that do not. Two main study designs can help address this question. The first such design entails surveying the genetic features of primary tumors in a population of patients, and associating these features with the development of metastasis[76]. Unfortunately, this type of study relies on the assumption that the genetic features detected in the primary tumor faithfully represent the features of the clone that gave rise to the metastasis. We now know that this assumption is often flawed. The second study design in this framework entails studying the genomes of biopsies obtained from metastases compared to biopsies of primary tumors that never metastasized, where neither the metastasis nor the primary samples are from the same individual[77]. This study design is the only way to associate certain genetic lesions occurring in primary tumors with the development of metastasis. It does not make the assumption that paired primaries faithfully represent metastases, but does suffer from the same potential confounders that can affect any epidemiologic study.

In contrast to the above study designs, the comparison of paired primary and metastatic tumors offers a useful internal control in each case. The data from paired tumors allows one to ask new questions such as: (i) What fraction of genetic alterations are shared between paired primaries and metastases? (ii) Are certain genetic alterations found uniquely and recurrently in metastases? (iii) Do all metastases typically descend from one lineage of the cancer cells in the primary tumor?

One of the first genome-scale studies of paired metastases and primaries focused on prostate cancer[78]. Array cGH and SNP6.0 analyses were performed on multiple metastases from the same patient, and in some cases, on a biopsy of the paired primary. The principle of clonal ordering was used to reconstruct phylogenetic trees[79]. This principle is based on fact that when a cancer cell acquires an alteration, all of its descendants inherit this same alteration. Comparing the sets of shared alterations between different tumor samples can therefore allow one to infer the ancestral structure of a set of related cancer samples. In this study, all metastases from the same patient shared copy number alterations, including common breakpoints. While primary prostate tumors are often multifocal, meaning that they consist of independently arising cancers, the common breakpoints identified in paired metastases indicated that all of the metastases derived from one such independent cancer. Unfortunately, these copy number data alone were unable to address the full spectrum of genetic alterations that exist uniquely in metastases and was unable to assess whether metastases arise from a single lineage within a cancer phylogeny. Since this study, other methods have been developed to reconstruct phylogenetic trees from both allelic[80] and total copy number data[81].

Two of the first genome-scale sequencing studies of paired metastases and primaries focused on pancreatic cancer[82,83]. Vogelstein and colleagues sequenced the exomes of seven pancreatic cancer metastases from seven patients (index lesions). At the time, genome-wide sequencing of additional tumor biopsies from these patients was prohibitively expensive. Therefore, the authors performed Sanger sequencing validation of the mutations identified in the index lesions in other metastases, and in three cases, various biopsies of the primary. An average of 64% of mutations identified in the index metastasis were detected in all tumor biopsies surveyed from the same patient. Mutations in well-known driver genes such as *TP53* and *KRAS* fell into this category of ubiquitous mutations.

Campbell et al conducted a parallel study that used structural variations rather than point mutations to draw phylogenetic trees. Whole genome sequencing of index lesions was performed to identify somatic rearrangements. Polymerase chain reaction assays were performed on other tumor biopsies to genotype them for all of the identified rearrangements. Both studies identified organ-specific clades within the phylogenetic tree, suggesting metastasis-to-metastasis spreading within a single foreign tissue.

By 2012, sequencing had become less expensive, enabling scientists to perform whole exome sequencing on a larger number of samples. Gerlinger et al capitalized on this advance, performing whole exome sequencing across multiple biopsies of two renal cell carcinoma primary tumors and, in one patient, across multiple metastases[84]. Separate biopsies of the primary tumor contained different alterations in driver genes of renal cell carcinomas, including instances in which the same driver gene harbored different mutations in different biopsies. Furthermore, all metastases descended from a single branch of the phylogenetic tree, in which the ancestral clone had undergone whole genome doubling. While not the first study of intratumoral heterogeneity in cancer, this was the first unbiased sequencing survey of somatic alterations in all biopsies. Similar studies of primary tumors followed, confirming the finding of intratumoral heterogeneity consisting of branched evolution[85-87].

In parallel to the spatial genomic heterogeneity discovered in multi-region sequencing studies, algorithmic advances were enabling the precise dissection of the subclonal populations within tumor biopsies. The accurate detection of subclones within tumor biopsies requires determination of the purity of the biopsy, and allele specific copy number at each genomic locus. One of the first algorithms to assemble this information was the Allele Specific Copy number Analysis of Tumors (ASCAT) algorithm[88]. The original ASCAT algorithm was able to derive allele specific copy number values and purity estimates from SNP array data, but did not include methods to quantify subclones within these populations. The subsequent development of ABSOLUTE[89] made several improvements that allowed for the quantification of subclones. ABSOLUTE integrated mutational allelic frequencies from sequencing data for more accurate estimation of tumor purity in samples with few copy number alterations. These variant allelic frequencies can be overlaid onto final estimates of purity and ploidy to estimate cancer cell fraction (CCF) within a biopsy carrying said mutation. These CCF values can be clustered to define subclonal populations of cancer cells within biopsies. Today, several such algorithms exist to perform these tasks based on both mutation[90] and copy number data[91]. Application of these algorithms to deep genome sequencing data allows for phylogenetic inference from a single tumor biopsy[92].

In 2015, the first evidence of tumor-self-seeding in humans was documented in prostate cancer[93]. A total of 51 biopsies from 10 patients with lethal metastatic prostate cancer were obtained on autopsy. Whole genome sequencing was performed on all biopsies. The authors used an n-dimensional Dirichlet process[92,94] to identify subclonal populations of cancer cells in and across biopsies. This analysis demonstrated that multiple subclones were shared across biopsies, a finding inconsistent with strictly branched evolution from single seeding events.

Two technological advances have opened a new window into metastasis: (i) the reliable capture of CTCs and (ii) the ability to sequence the DNA of single cells. CTCs were first observed in the blood of a deceased patient by Ashworth in 1869[95]. More than 130 years later, these cells were finally isolated in breast cancer patients, and their presence was associated with poorer survival[96]. These cells are often isolated based on their expression of the epithelial markers, such as cytokeratin[97] and EpCAM[98]. Various strategies for the isolation of CTCs have now been devised. In conjunction with advances that have allowed for the sequence of DNA from single cells[99], studies of clones that have completed some or all of the invasion-metastasis cascade have been performed. One such study in prostate cancer revealed that collectively, circulating tumor cells contained a set of mutations found in the metastasis that were not ubiquitously detected across exome sequenced cores of the primary tumor. Unfortunately, current technologies rely on whole genome amplification of DNA from single cells, and stochastic effects such as allelic drop-out can confound the interpretation of such analyses[100]. As such, the current data are insufficient to address the question of whether CTCs represent a genetically defined subpopulation of cancer cells from the primary tumor or are shed randomly from all tumor sites.

## 6. Approaches to cancer therapy

The efficacy of medical treatments for cancer has long relied on nonselective cytotoxicity. In 1896, a medical student named Emil Grubbe assembled the first X-ray machine in Chicago and noticed the skin peeling off of his hand after it was exposed to the machine's beam. He reasoned that if the beam could kill healthy cells, then perhaps it could also kill cancer cells. Later that year, he used it to experimentally treat Rose Lee's breast cancer. Her tumor showed a dramatic response, heralding the birth of radiotherapy. Toxic chemicals were the next modality to be employed in the treatment of cancer. Some of the earliest chemotherapies, the alkylating agents, spawned from mustard gas used in World War I. The grim history of these compounds reflects their toxicities. Radiotherapy and cytotoxic chemotherapies have saved many lives, but their nonselective nature has hastened the deaths of others and often left those who survive irreparably harmed. For this reason, the focus of cancer treatment strategies now emphasizes therapies that differentially impact normal cells and cancer cells.

Some of the most successful examples of cancer-selective agents are those whose therapeutic window consists of organ-type specificity. For example, therapies that influence the signaling of sex hormones have had an enormous impact on the treatment of breast and prostate cancers. Tamoxifen binds to and inhibits the transcriptional activity of the estrogen receptor and is now standard-of-care in estrogen-receptor positive breast cancers. Similarly, androgen deprivation therapy has long been the first-line treatment for prostate cancer. More recently, immunotherapeutic approaches have taken advantage of some of the cell surface markers expressed specifically by the cell of origin of certain cancers. One such therapy is Chimeric Antigen Receptor T-cell therapy (CAR-T). This strategy infects autologous T-cells with an immune-stimulating receptor that binds to CD19, a cell surface receptor expressed on all mature B-cells, including those found in chronic lymphocytic leukemia (CLL) and acute lymphocytic leukemia (ALL)[101-103]. Vaccines against antigens expressed specifically by certain organs, such as mammoglobin in breast tissue, are currently being developed[104].

A separate paradigm for cancer therapeutics consists of targeting oncogenic alterations present in cancer cells that are absent in normal cells. This strategy therefore offers a genetic therapeutic window. Therapeutics that interrupt these oncogene addictions now exist, mostly consisting of either including small molecules that inhibit aberrant tyrosine kinases or antibodies directed against overexpressed growth factor receptors.

## 7. Synthetic lethal therapeutics in cancer

Experimental studies show that approximately five oncogenic alterations are required to transform normal human cells into cancer cells[105]. Sequencing studies have revealed similar estimates, confirming that in many tumors the number of oncogenic driver alterations is small[106]. Unfortunately, many of the recurrent alterations in cancer genomes are not amenable to inhibition by small molecules. Entire classes of driver events, such as tumor suppressor loss and transcription factor amplification, are "undruggable" with current technology. These two sets of facts suggest that many cancer patients exist for whom the current paradigm of cancer drug discovery will offer no targeted therapies.

However, one can imagine that many of these driver genetic alterations, or clonal passenger alterations accompanying them, can cause unintended vulnerabilities in cancer cells that are not shared with normal cells. In model organisms, when mutations in two genes separately produce viable organisms, but when mutated together do not, the interaction between the two genes is said to be one of synthetic lethality. This principle was first proposed as a way of identifying therapeutic leads for cancer in 1997 by Hartwell and colleagues[107]. As proof of concept, Hartwell et al subjected 70 isogenic yeast strains with deletions in genes encoding proteins in the DNA repair pathway to FDA-approved chemotherapeutics. They sought to identify genetic modulators of sensitivity to each chemotherapeutic agent. Yeast strains deficient in proteins involved in post-replication DNA repair were especially sensitive to cisplatin. These data provided the first evidence that genetic mutations in cancer cells could predict heightened sensitivity to compounds without targeting a mutated oncogene.

The most therapeutically successful synthetic lethal interaction discovered thus far has been inhibition of the DNA single strand repair enzyme Poly-ADP-ribose Polymerase (PARP) in BRCA-deficient cancers. Cells incur double strand breaks as a result of exposure to ionizing radiation, but also at lower frequencies during each cell division[108]. Normally, the BRCA proteins help to resolve these DNA double-strand breaks as part of the homologous recombination pathway. As a result, germline mutations in either *BRCA1* or *BRCA2* predispose towards the biallelic loss of these "caretaker" genes and are associated with the development of breast and ovarian cancers[109]. Indeed, sporadic ovarian cancers also exhibit somatic mutations of either *BRCA1* (3%) or *BRCA2* (3%)[110].

PARP is responsible for repairing single stranded breaks in DNA. When these single stranded nicks go unrepaired, they cause stalled replication forks and cell death ensues[111]. In normal cells, BRCA can substitute for PARP to resolve these single stranded breaks, thereby sparing normal cells in the face of PARP inhibition. With these pathways in mind, small molecule inhibitors of PARP have been developed[112]. Preclinical models demonstrated remarkable sensitivity of BRCA deficient cancers to PARP inhibitors[113,114]. After early failures[115,116], these inhibitors have also enjoyed success in clinical trials[117]. Resistance to PARP inhibitors occurs by reversion of BRCA mutations[118,119], thereby confirming that they work through a synthetic lethal mechanism.

Other hypothesis-based synthetic lethal interactions have been explored. More than twenty years ago, Emil "Tom" Frei proposed that deletion events in cancer could affect the activity of essential enzymes, which may make chemotherapeutics targeting the same pathway more effective[120]. One compelling study demonstrating this principle showed that glioblastoma cell lines harboring deletions of one isoform of enolase were extremely sensitive to inhibition of the second isoform of enolase[121]. Our laboratory has recently shown that hemizygous deletion of essential genes sensitizes cancer cells to their further expression[122]. Similarly, frequent hemizygous deletions of *POLR2A* in colorectal cancer sensitize cancer cells to the small molecule inhibitor of POLR2A, α-amanitin[123].

## 8. Loss-of-function screens to uncover cancer vulnerabilities

Despite advances in our understanding of cellular biology, knowledge of the interactions between genes that would predict synthetic lethal opportunities is limited. Technological advances have allowed for the systematic perturbation of cellular processes and simultaneous measurement of effects. With these technologies, functional genotype-phenotype associations can be probed with relative ease. The first technology that ushered in this era of high-throughput screening was RNA-interference, which allowed for the silencing of any gene[124]. In particular, libraries of siRNAs or shRNAs allow for this screens where one assess the impact of decreasing a given gene's expression on a phenotype of interest. Use of siRNAs in a screening strategy require that siRNAs be arrayed into a multiwell tissue culture plate for reverse transfection into cells of interest. In contrast, shRNA libraries use plasmids that integrate into the genome following lentiviral infection to constitutively express shRNAs for gene silencing[125]. shRNA libraries does not require reverse transfection or multi-well tissue culture. The ability to encode barcodes that uniquely identify every shRNA allows one to measure the representation of cells expressing all shRNAs. Often, cells are infected with the shRNA library and then split into two pools where one is subjected to a treatment of interest and the other serves as a control. One can quantify the impact of shRNA expression on cellular proliferation by comparing the barcode representation in the control pool to the experimental pool. Next-generation sequencing enables discretized barcode representation data to be collected on millions of cells at once, allowing for shRNA screens to be conducted in pooled format[126,127]. Pooling makes screens easier, less expensive, and allows for the interrogation of *in-vivo* phenotypes. More involved experimental setups allow for screening of a variety of phenotypes aside from proliferation.

shRNA screens of cancer cells have identified candidate synthetic lethal vulnerabilities of previously "undruggable" cancer driver alterations. For example, KRAS has long been considered undruggable because of its lack of allosteric hydrophobic pockets and picomolar affinity for GTP/GDP[128]. Genome-scale shRNA screens have nominated PLK1[129], STK33[130], TBK1[131] and others[132-134] as vulnerabilities in KRAS mutant cancer cells. In addition, synthetic lethal vulnerabilities identified in RNAi screens have been proposed for TP53[135] and MYC[136].

While many shRNA screens focus on comparing two isogenic cell lines that differ in a variable of interest, more recently shRNA screens have been performed on large numbers of cell lines, each of which has undergone parallel genomic characterization[137-

[139]. These rich datasets enable scientists to identify associations of genetic dependencies with genetic features in an unbiased, post-hoc manner. With these growing datasets, an increasing appreciation for the technical shortcomings of shRNA screens has emerged[140].

One important aspect of addressing these shortcomings is the design of algorithms to extract the most meaningful data possible from these screens. One such algorithm, ATARiS, was developed and applied to the Achilles Project data[141]. ATARiS works on the premise that when the predominant effects of a set of shRNAs targeting the same gene are due to on-target gene suppression, then these shRNAs should induce similar phenotypic effects across cell lines. The algorithm therefore searches for associations between independent hairpin sequences targeting the same gene. Hairpins that behave consistently across cell lines can then be grouped together to derive gene dependency scores.

More recently, CRISPR-CAS9 endonuclease technology has offered a new loss-of-function screening tool that addresses many of the shortcomings of shRNA screens. In 1987, Ishino et al, discovered perplexing 29-nucleotide repeats in Escherichia coli, and named these repeat structures CRIPSR[142]. These structures were later found to constitute one arm of a bacterial mechanism of immune memory[143]. Together with the CAS9 endonuclease, CRISPR guide RNAs are capable of cleaving dsDNA complimentary to the guide sequence. Whereas prior techniques for genome editing such as Zinc-finger-nucleases[144] and TALENs[145,146] required laborious cloning and optimization, generation of CRISPR guide RNAs is much simpler. This ease of generation allows for the prospective creation of large libraries of genome editing plasmids that are suitable for genome-scale loss-of-function screens. In 2014, the first CRISPR screens in human cells were performed[147,148].

CRISPR exhibit greater reproducibility between replicates, and suffer from fewer off-target effects than shRNA screens[147]. Depending on the phenotype of interest, one may prefer to study gene knockout using CRISPR rather than gene suppression with shRNAs. However, one drawback of CRISPR as a means to assess loss-of-function is that one third of deletions on haploid genes and 5/9 of deletions on diploid genes will result in at-least one in-frame deletion which may have potentially mild effects on gene function.

## 10. References

1 Ferlay, J. *et al.* Cancer incidence and mortality worldwide: sources, methods and major patterns in GLOBOCAN 2012. *International journal of cancer. Journal international du cancer* **136**, E359-386, doi:10.1002/ijc.29210 (2015).

2 Russell, W. An Address on a Characteristic Organism of Cancer. *British medical journal* **2**, 1356-1360 (1890).

3 Matthews, J. B. The immunoglobulin nature of Russell bodies. *British journal of experimental pathology* **64**, 331-335 (1983).

4 Livingston, V. W. & Alexander-Jackson, E. A specific type of organism cultivated from malignancy: bacteriology and proposed classification. *Annals of the New York Academy of Sciences* **174**, 636-654 (1970).

5 Marshall, B. J. & Warren, J. R. Unidentified curved bacilli in the stomach of patients with gastritis and peptic ulceration. *Lancet* **1**, 1311-1315 (1984).

6 Epstein, M. A., Achong, B. G. & Barr, Y. M. Virus Particles in Cultured Lymphoblasts from Burkitt's Lymphoma. *Lancet* **1**, 702-703 (1964).

7 Durst, M., Gissmann, L., Ikenberg, H. & zur Hausen, H. A papillomavirus DNA from a cervical carcinoma and its prevalence in cancer biopsy samples from different geographic regions. *Proceedings of the National Academy of Sciences of the United States of America* **80**, 3812-3815 (1983).

8 Kostic, A. D. *et al.* PathSeq: software to identify or discover microbes by deep sequencing of human tissue. *Nature biotechnology* **29**, 393-396, doi:10.1038/nbt.1868 (2011).

9 Kostic, A. D. *et al.* Genomic analysis identifies association of Fusobacterium with colorectal carcinoma. *Genome research* **22**, 292-298, doi:10.1101/gr.126573.111 (2012).

10 Rous, P. A Transmissible Avian Neoplasm. (Sarcoma of the Common Fowl.). *The Journal of experimental medicine* **12**, 696-705 (1910).

11 Ames, B. N. Dietary carcinogens and anticarcinogens. Oxygen radicals and degenerative diseases. *Science* **221**, 1256-1264 (1983).

12 Ames, B. N., Durston, W. E., Yamasaki, E. & Lee, F. D. Carcinogens are mutagens: a simple test system combining liver homogenates for activation and bacteria for detection. *Proceedings of the National Academy of Sciences of the United States of America* **70**, 2281-2285 (1973).

13 Stehelin, D., Varmus, H. E., Bishop, J. M. & Vogt, P. K. DNA related to the transforming gene(s) of avian sarcoma viruses is present in normal avian DNA. *Nature* **260**, 170-173 (1976).

14 Vogelstein, B., Fearon, E. R., Hamilton, S. R. & Feinberg, A. P. Use of restriction fragment length polymorphisms to determine the clonal origin of human tumors. *Science* **227**, 642-645 (1985).

15 Vogelstein, B. *et al.* Genetic alterations during colorectal-tumor development. *The New England journal of medicine* **319**, 525-532, doi:10.1056/NEJM198809013190901 (1988).

16 Clark, W. H., Jr. *et al.* A study of tumor progression: the precursor lesions of superficial spreading and nodular melanoma. *Human pathology* **15**, 1147-1165 (1984).

17 Pfeifer, G. P. Environmental exposures and mutational patterns of cancer genomes. *Genome medicine* **2**, 54, doi:10.1186/gm175 (2010).

18 Fishel, R. *et al.* The human mutator gene homolog MSH2 and its association with hereditary nonpolyposis colon cancer. *Cell* **75**, 1027-1038 (1993).

19 Papadopoulos, N. *et al.* Mutation of a mutL homolog in hereditary colon cancer. *Science* **263**, 1625-1629 (1994).

20 Miyaki, M. *et al.* Germline mutation of MSH6 as the cause of hereditary nonpolyposis colorectal cancer. *Nature genetics* **17**, 271-272, doi:10.1038/ng1197-271 (1997).

21 Alexandrov, L. B. *et al.* Signatures of mutational processes in human cancer. *Nature* **500**, 415-421, doi:10.1038/nature12477 (2013).

22 Cancer Genome Atlas Research, N. *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67-73, doi:10.1038/nature12113 (2013).

23 Cancer Genome Atlas, N. Comprehensive molecular characterization of human colon and rectal cancer. *Nature* **487**, 330-337, doi:10.1038/nature11252 (2012).

24 Taylor, B. J. *et al.* DNA deaminases induce break-associated mutation showers with implication of APOBEC3B and 3A in breast cancer kataegis. *eLife* **2**, e00534, doi:10.7554/eLife.00534 (2013).

25 Burns, M. B. *et al.* APOBEC3B is an enzymatic source of mutation in breast cancer. *Nature* **494**, 366-370, doi:10.1038/nature11881 (2013).

26 Rowley, J. D. Letter: A new consistent chromosomal abnormality in chronic myelogenous leukaemia identified by quinacrine fluorescence and Giemsa staining. *Nature* **243**, 290-293 (1973).

27 Stephens, P. J. *et al.* Massive genomic rearrangement acquired in a single catastrophic event during cancer development. *Cell* **144**, 27-40, doi:10.1016/j.cell.2010.11.055 (2011).

28 Baca, S. C. *et al.* Punctuated evolution of prostate cancer genomes. *Cell* **153**, 666-677, doi:10.1016/j.cell.2013.03.021 (2013).

29 Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nature genetics* **45**, 1127-1133, doi:10.1038/ng.2762 (2013).

30 Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature genetics* **45**, 1134-1140, doi:10.1038/ng.2760 (2013).

31 Piccart-Gebhart, M. J. *et al.* Trastuzumab after adjuvant chemotherapy in HER2-positive breast cancer. *The New England journal of medicine* **353**, 1659-1672, doi:10.1056/NEJMoa052306 (2005).

32 Romond, E. H. *et al.* Trastuzumab plus adjuvant chemotherapy for operable HER2-positive breast cancer. *The New England journal of medicine* **353**, 1673-1684, doi:10.1056/NEJMoa052122 (2005).

33 Feinberg, A. P. & Vogelstein, B. Hypomethylation distinguishes genes of some human cancers from their normal counterparts. *Nature* **301**, 89-92 (1983).

34 Greger, V., Passarge, E., Hopping, W., Messmer, E. & Horsthemke, B. Epigenetic changes may contribute to the formation and spontaneous regression of retinoblastoma. *Human genetics* **83**, 155-158 (1989).

35 Turcan, S. *et al.* IDH1 mutation is sufficient to establish the glioma hypermethylator phenotype. *Nature* **483**, 479-483, doi:10.1038/nature10866 (2012).

36 Schwartzentruber, J. *et al.* Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature* **482**, 226-231, doi:10.1038/nature10833 (2012).

37 Rossi, D. & Gaidano, G. Richter syndrome: molecular insights and clinical perspectives. *Hematological oncology* **27**, 1-10, doi:10.1002/hon.880 (2009).

38 Albitar, M. *et al.* Differences between refractory anemia with excess blasts in transformation and acute myeloid leukemia. *Blood* **96**, 372-373 (2000).

39 Farber, S. & Diamond, L. K. Temporary remissions in acute leukemia in children produced by folic acid antagonist, 4-aminopteroyl-glutamic acid. *The New England journal of medicine* **238**, 787-793, doi:10.1056/NEJM194806032382301 (1948).

40 Gorre, M. E. *et al.* Clinical resistance to STI-571 cancer therapy caused by BCR-ABL gene mutation or amplification. *Science* **293**, 876-880, doi:10.1126/science.1062538 (2001).

41 Shah, N. P. *et al.* Overriding imatinib resistance with a novel ABL kinase inhibitor. *Science* **305**, 399-401, doi:10.1126/science.1099480 (2004).

42 Sequist, L. V. *et al.* Rociletinib in EGFR-mutated non-small-cell lung cancer. *The New England journal of medicine* **372**, 1700-1709, doi:10.1056/NEJMoa1413654 (2015).

43 Janne, P. A. *et al.* AZD9291 in EGFR inhibitor-resistant non-small-cell lung cancer. *The New England journal of medicine* **372**, 1689-1699, doi:10.1056/NEJMoa1411817 (2015).

44 Long, G. V. *et al.* Combined BRAF and MEK inhibition versus BRAF inhibition alone in melanoma. *The New England journal of medicine* **371**, 1877-1888, doi:10.1056/NEJMoa1406037 (2014).

45 Van Allen, E. M. *et al.* The genetic landscape of clinical resistance to RAF inhibition in metastatic melanoma. *Cancer discovery* **4**, 94-109, doi:10.1158/2159-8290.CD-13-0617 (2014).

46 Murchison, E. P. *et al.* Transmissible [corrected] dog cancer genome reveals the origin and history of an ancient cell lineage. *Science* **343**, 437-440, doi:10.1126/science.1247167 (2014).

47 Adey, A. *et al.* The haplotype-resolved genome and epigenome of the aneuploid HeLa cancer cell line. *Nature* **500**, 207-211, doi:10.1038/nature12064 (2013).

48 Mehlen, P. & Puisieux, A. Metastasis: a question of life or death. *Nature reviews. Cancer* **6**, 449-458, doi:10.1038/nrc1886 (2006).

49 Vidal, M. *et al.* A role for the epithelial microenvironment at tumor boundaries: evidence from Drosophila and human squamous cell carcinomas. *The American journal of pathology* **176**, 3007-3014, doi:10.2353/ajpath.2010.090253 (2010).

50 Itoh, T. *et al.* Experimental metastasis is suppressed in MMP-9-deficient mice. *Clinical & experimental metastasis* **17**, 177-181 (1999).

51 Sternlicht, M. D. & Werb, Z. How matrix metalloproteinases regulate cell behavior. *Annual review of cell and developmental biology* **17**, 463-516, doi:10.1146/annurev.cellbio.17.1.463 (2001).

52 Egeblad, M. & Werb, Z. New functions for the matrix metalloproteinases in cancer progression. *Nature reviews. Cancer* **2**, 161-174, doi:10.1038/nrc745 (2002).

53 Kessenbrock, K., Plaks, V. & Werb, Z. Matrix metalloproteinases: regulators of the tumor microenvironment. *Cell* **141**, 52-67, doi:10.1016/j.cell.2010.03.015 (2010).

54 Spaderna, S. *et al.* A transient, EMT-linked loss of basement membranes indicates metastasis and poor survival in colorectal cancer. *Gastroenterology* **131**, 830-840, doi:10.1053/j.gastro.2006.06.016 (2006).

55 Neufeld, G. & Kessler, O. Pro-angiogenic cytokines and their role in tumor angiogenesis. *Cancer metastasis reviews* **25**, 373-385, doi:10.1007/s10555-006-9011-5 (2006).

56 Carmeliet, P. & Jain, R. K. Principles and mechanisms of vessel normalization for cancer and other angiogenic diseases. *Nature reviews. Drug discovery* **10**, 417-427, doi:10.1038/nrd3455 (2011).

57 Aceto, N. *et al.* Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell* **158**, 1110-1122, doi:10.1016/j.cell.2014.07.013 (2014).

58 MacDonald, I. C., Groom, A. C. & Chambers, A. F. Cancer spread and micrometastasis development: quantitative approaches for in vivo models. *BioEssays : news and reviews in molecular, cellular and developmental biology* **24**, 885-893, doi:10.1002/bies.10156 (2002).

59 Naumov, G. N. *et al.* Persistence of solitary mammary carcinoma cells in a secondary site: a possible contributor to dormancy. *Cancer research* **62**, 2162-2168 (2002).

60 Braun, S. *et al.* A pooled analysis of bone marrow micrometastasis in breast cancer. *The New England journal of medicine* **353**, 793-802, doi:10.1056/NEJMoa050434 (2005).

61 Nakaya, Y. & Sheng, G. Epithelial to mesenchymal transition during gastrulation: an embryological view. *Development, growth & differentiation* **50**, 755-766, doi:10.1111/j.1440-169X.2008.01070.x (2008).

62 Onder, T. T. *et al.* Loss of E-cadherin promotes metastasis via multiple downstream transcriptional pathways. *Cancer research* **68**, 3645-3654, doi:10.1158/0008-5472.CAN-07-2938 (2008).

63 Nagafuchi, A., Shirayoshi, Y., Okazaki, K., Yasuda, K. & Takeichi, M. Transformation of cell adhesion properties by exogenously introduced E-cadherin cDNA. *Nature* **329**, 341-343, doi:10.1038/329341a0 (1987).

64 Hsu, M. Y., Wheelock, M. J., Johnson, K. R. & Herlyn, M. Shifts in cadherin profiles between human normal melanocytes and melanomas. *The journal of investigative dermatology. Symposium proceedings / the Society for Investigative Dermatology, Inc. [and] European Society for Dermatological Research* **1**, 188-194 (1996).

65 Gravdal, K., Halvorsen, O. J., Haukaas, S. A. & Akslen, L. A. A switch from E-cadherin to N-cadherin expression indicates epithelial to mesenchymal transition and is of strong and independent importance for the progress of prostate cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **13**, 7003-7011, doi:10.1158/1078-0432.CCR-07-1263 (2007).

66 De Wever, O. *et al.* Critical role of N-cadherin in myofibroblast invasion and migration in vitro stimulated by colon-cancer-cell-derived TGF-beta or wounding. *Journal of cell science* **117**, 4691-4703, doi:10.1242/jcs.01322 (2004).

67 Xu, J., Lamouille, S. & Derynck, R. TGF-beta-induced epithelial to mesenchymal transition. *Cell research* **19**, 156-172, doi:10.1038/cr.2009.5 (2009).

68 Lamouille, S., Xu, J. & Derynck, R. Molecular mechanisms of epithelial-mesenchymal transition. *Nature reviews. Molecular cell biology* **15**, 178-196, doi:10.1038/nrm3758 (2014).

69 Weinberg, R. A. *The Biology of Cancer*. 2 edn, (Garland Science, 2014).

70 Yu, M. *et al.* Circulating breast tumor cells exhibit dynamic changes in epithelial and mesenchymal composition. *Science* **339**, 580-584, doi:10.1126/science.1228522 (2013).

71 Paget, S. The distribution of secondary growths in cancer of the breast. 1889. *Cancer metastasis reviews* **8**, 98-101 (1989).

72 Bedikian, A. Y. *et al.* Treatment of uveal melanoma metastatic to the liver: a review of the M. D. Anderson Cancer Center experience and prognostic factors. *Cancer* **76**, 1665-1670 (1995).

73 Chen, Y., Thompson, W., Semenciw, R. & Mao, Y. Epidemiology of contralateral breast cancer. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* **8**, 855-861 (1999).

74 Tseng, L. N. *et al.* Port-site metastases. Impact of local tissue trauma and gas leakage. *Surgical endoscopy* **12**, 1377-1380 (1998).

75 Hirshberg, A., Shnaiderman-Shapiro, A., Kaplan, I. & Berger, R. Metastatic tumours to the oral cavity - pathogenesis and analysis of 673 cases. *Oral oncology* **44**, 743-752, doi:10.1016/j.oraloncology.2007.09.012 (2008).

76 Tie, J. *et al.* KRAS mutation is associated with lung metastasis in patients with curatively resected colorectal cancer. *Clinical cancer research : an official journal of the American Association for Cancer Research* **17**, 1122-1130, doi:10.1158/1078-0432.CCR-10-1720 (2011).

77 Robinson, D. *et al.* Integrative clinical genomics of advanced prostate cancer. *Cell* **161**, 1215-1228, doi:10.1016/j.cell.2015.05.001 (2015).

78 Liu, W. *et al.* Copy number analysis indicates monoclonal origin of lethal metastatic prostate cancer. *Nature medicine* **15**, 559-565, doi:10.1038/nm.1944 (2009).

79 Merlo, L. M., Pepper, J. W., Reid, B. J. & Maley, C. C. Cancer as an evolutionary and ecological process. *Nature reviews. Cancer* **6**, 924-935, doi:10.1038/nrc2013 (2006).

80 Schwarz, R. F. *et al.* Phylogenetic quantification of intra-tumour heterogeneity. *PLoS computational biology* **10**, e1003535, doi:10.1371/journal.pcbi.1003535 (2014).

81 Letouze, E., Allory, Y., Bollet, M. A., Radvanyi, F. & Guyon, F. Analysis of the copy number profiles of several tumor samples from the same patient reveals the successive steps in tumorigenesis. *Genome biology* **11**, R76, doi:10.1186/gb-2010-11-7-r76 (2010).

82 Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114-1117, doi:10.1038/nature09515 (2010).

83 Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109-1113, doi:10.1038/nature09460 (2010).

84 Gerlinger, M. *et al.* Intratumor heterogeneity and branched evolution revealed by multiregion sequencing. *The New England journal of medicine* **366**, 883-892, doi:10.1056/NEJMoa1113205 (2012).

85 Zhang, J. *et al.* Intratumor heterogeneity in localized lung adenocarcinomas delineated by multiregion sequencing. *Science* **346**, 256-259, doi:10.1126/science.1256930 (2014).

86 de Bruin, E. C. *et al.* Spatial and temporal diversity in genomic instability processes defines lung cancer evolution. *Science* **346**, 251-256, doi:10.1126/science.1253462 (2014).

87 Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature genetics* **46**, 225-233, doi:10.1038/ng.2891 (2014).

88 Van Loo, P. *et al.* Allele-specific copy number analysis of tumors. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 16910-16915, doi:10.1073/pnas.1009843107 (2010).

89 Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* **30**, 413-421, doi:10.1038/nbt.2203 (2012).

90 Roth, A. *et al.* PyClone: statistical inference of clonal population structure in cancer. *Nature methods* **11**, 396-398, doi:10.1038/nmeth.2883 (2014).

91 Ha, G. *et al.* TITAN: inference of copy number architectures in clonal cell populations from tumor whole-genome sequence data. *Genome research* **24**, 1881-1893, doi:10.1101/gr.180281.114 (2014).

92 Nik-Zainal, S. *et al.* The life history of 21 breast cancers. *Cell* **149**, 994-1007, doi:10.1016/j.cell.2012.04.023 (2012).

93 Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature* **520**, 353-357, doi:10.1038/nature14347 (2015).

94 Cooper, C. S. *et al.* Analysis of the genetic phylogeny of multifocal prostate cancer identifies multiple independent clonal expansions in neoplastic and morphologically normal prostate tissue. *Nature genetics* **47**, 367-372, doi:10.1038/ng.3221 (2015).

95 TR., A. A case of cancer in which cells similar to those in the tumors were seen in the blood after death. **14**, 146-146 (1869).

96 Cristofanilli, M. *et al.* Circulating tumor cells, disease progression, and survival in metastatic breast cancer. *The New England journal of medicine* **351**, 781-791, doi:10.1056/NEJMoa040766 (2004).

97 Meng, S. *et al.* Circulating tumor cells in patients with breast cancer dormancy. *Clinical cancer research : an official journal of the American Association for Cancer Research* **10**, 8152-8162, doi:10.1158/1078-0432.CCR-04-1110 (2004).

98 Nagrath, S. *et al.* Isolation of rare circulating tumour cells in cancer patients by microchip technology. *Nature* **450**, 1235-1239, doi:10.1038/nature06385 (2007).

99 Lasken, R. S. Single-cell genomic sequencing using Multiple Displacement Amplification. *Current opinion in microbiology* **10**, 510-516, doi:10.1016/j.mib.2007.08.005 (2007).

100 Findlay, I., Ray, P., Quirke, P., Rutherford, A. & Lilford, R. Allelic drop-out and preferential amplification in single cells and human blastomeres: implications for preimplantation diagnosis of sex and cystic fibrosis. *Human reproduction* **10**, 1609-1618 (1995).

101 Maude, S. L. *et al.* Chimeric antigen receptor T cells for sustained remissions in leukemia. *The New England journal of medicine* **371**, 1507-1517, doi:10.1056/NEJMoa1407222 (2014).

102 Grupp, S. A. *et al.* Chimeric antigen receptor-modified T cells for acute lymphoid leukemia. *The New England journal of medicine* **368**, 1509-1518, doi:10.1056/NEJMoa1215134 (2013).

103 Porter, D. L., Levine, B. L., Kalos, M., Bagg, A. & June, C. H. Chimeric antigen receptor-modified T cells in chronic lymphoid leukemia. *The New England journal of medicine* **365**, 725-733, doi:10.1056/NEJMoa1103849 (2011).

104 Tiriveedhi, V. *et al.* Mammaglobin-A cDNA vaccination of breast cancer patients induces antigen-specific cytotoxic CD4+ICOShi T cells. *Breast cancer research and treatment* **138**, 109-118, doi:10.1007/s10549-012-2110-9 (2013).

105 Boehm, J. S., Hession, M. T., Bulmer, S. E. & Hahn, W. C. Transformation of human and murine fibroblasts without viral oncoproteins. *Molecular and cellular biology* **25**, 6464-6474, doi:10.1128/MCB.25.15.6464-6474.2005 (2005).

106 Vogelstein, B. *et al.* Cancer genome landscapes. *Science* **339**, 1546-1558, doi:10.1126/science.1235122 (2013).

107 Hartwell, L. H., Szankasi, P., Roberts, C. J., Murray, A. W. & Friend, S. H. Integrating genetic approaches into the discovery of anticancer drugs. *Science* **278**, 1064-1068 (1997).

108 Rothkamm, K., Kruger, I., Thompson, L. H. & Lobrich, M. Pathways of DNA double-strand break repair during the mammalian cell cycle. *Molecular and cellular biology* **23**, 5706-5715 (2003).

109 King, M. C., Marks, J. H., Mandell, J. B. & New York Breast Cancer Study, G. Breast and ovarian cancer risks due to inherited mutations in BRCA1 and BRCA2. *Science* **302**, 643-646, doi:10.1126/science.1088759 (2003).

110 Cancer Genome Atlas Research, N. Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615, doi:10.1038/nature10166 (2011).

111 McGlynn, P. & Lloyd, R. G. Recombinational repair and restart of damaged replication forks. *Nature reviews. Molecular cell biology* **3**, 859-870, doi:10.1038/nrm951 (2002).

112 Ratnam, K. & Low, J. A. Current development of clinical inhibitors of poly(ADP-ribose) polymerase in oncology. *Clinical cancer research : an official journal of the American Association for Cancer Research* **13**, 1383-1388, doi:10.1158/1078-0432.CCR-06-2260 (2007).

113 Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, 917-921, doi:10.1038/nature03445 (2005).

114 Bryant, H. E. *et al.* Specific killing of BRCA2-deficient tumours with inhibitors of poly(ADP-ribose) polymerase. *Nature* **434**, 913-917, doi:10.1038/nature03443 (2005).

115 Guha, M. PARP inhibitors stumble in breast cancer. *Nature biotechnology* **29**, 373-374, doi:10.1038/nbt0511-373 (2011).

116 O'Shaughnessy, J. *et al.* Iniparib plus chemotherapy in metastatic triple-negative breast cancer. *The New England journal of medicine* **364**, 205-214, doi:10.1056/NEJMoa1011418 (2011).

117 Fong, P. C. *et al.* Inhibition of poly(ADP-ribose) polymerase in tumors from BRCA mutation carriers. *The New England journal of medicine* **361**, 123-134, doi:10.1056/NEJMoa0900212 (2009).

118 Edwards, S. L. *et al.* Resistance to therapy caused by intragenic deletion in BRCA2. *Nature* **451**, 1111-1115, doi:10.1038/nature06548 (2008).

119 Lord, C. J. & Ashworth, A. Mechanisms of resistance to therapies targeting BRCA-mutant cancers. *Nature medicine* **19**, 1381-1388, doi:10.1038/nm.3369 (2013).

120 Frei, E., 3rd. Gene deletion: a new target for cancer chemotherapy. *Lancet* **342**, 662-664 (1993).

121 Muller, F. L. *et al.* Passenger deletions generate therapeutic vulnerabilities in cancer. *Nature* **488**, 337-342, doi:10.1038/nature11331 (2012).

122 Nijhawan, D. *et al.* Cancer vulnerabilities unveiled by genomic loss. *Cell* **150**, 842-854, doi:10.1016/j.cell.2012.07.023 (2012).

123 Liu, Y. *et al.* TP53 loss creates therapeutic vulnerability in colorectal cancer. *Nature* **520**, 697-701, doi:10.1038/nature14418 (2015).

124 Fire, A. *et al.* Potent and specific genetic interference by double-stranded RNA in Caenorhabditis elegans. *Nature* **391**, 806-811, doi:10.1038/35888 (1998).

125 Berns, K. *et al.* A large-scale RNAi screen in human cells identifies new components of the p53 pathway. *Nature* **428**, 431-437, doi:10.1038/nature02371 (2004).

126 Silva, J. M. *et al.* Profiling essential genes in human mammary cells by multiplex RNAi screening. *Science* **319**, 617-620, doi:10.1126/science.1149185 (2008).

127 Sims, D. *et al.* High-throughput RNA interference screening using pooled shRNA libraries and next generation sequencing. *Genome biology* **12**, R104, doi:10.1186/gb-2011-12-10-r104 (2011).

128 John, J. *et al.* Kinetics of interaction of nucleotides with nucleotide-free H-ras p21. *Biochemistry* **29**, 6058-6065 (1990).

129 Luo, J. *et al.* A genome-wide RNAi screen identifies multiple synthetic lethal interactions with the Ras oncogene. *Cell* **137**, 835-848, doi:10.1016/j.cell.2009.05.006 (2009).

130 Scholl, C. *et al.* Synthetic lethal interaction between oncogenic KRAS dependency and STK33 suppression in human cancer cells. *Cell* **137**, 821-834, doi:10.1016/j.cell.2009.03.017 (2009).

131 Barbie, D. A. *et al.* Systematic RNA interference reveals that oncogenic KRAS-driven cancers require TBK1. *Nature* **462**, 108-112, doi:10.1038/nature08460 (2009).

132 Steckel, M. *et al.* Determination of synthetic lethal interactions in KRAS oncogene-dependent cancer cells reveals novel therapeutic targeting strategies. *Cell research* **22**, 1227-1245, doi:10.1038/cr.2012.82 (2012).

133 Vicent, S. *et al.* Wilms tumor 1 (WT1) regulates KRAS-driven oncogenesis and senescence in mouse and human models. *The Journal of clinical investigation* **120**, 3940-3952, doi:10.1172/JCI44165 (2010).

134 Wang, Y. *et al.* Critical role for transcriptional repressor Snail2 in transformation by oncogenic RAS in colorectal carcinoma cells. *Oncogene* **29**, 4658-4670, doi:10.1038/onc.2010.218 (2010).

135 Baldwin, A. *et al.* Kinase requirements in human cells: V. Synthetic lethal interactions between p53 and the protein kinases SGK2 and PAK3. *Proceedings of the National Academy of Sciences of the United States of America* **107**, 12463-12468, doi:10.1073/pnas.1007462107 (2010).

136 Kessler, J. D. *et al.* A SUMOylation-dependent transcriptional subprogram is required for Myc-driven tumorigenesis. *Science* **335**, 348-353, doi:10.1126/science.1212728 (2012).

137 Cheung, H. W. *et al.* Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 12372-12377, doi:10.1073/pnas.1109363108 (2011).

138 Marcotte, R. *et al.* Essential gene profiles in breast, pancreatic, and ovarian cancer cells. *Cancer discovery* **2**, 172-189, doi:10.1158/2159-8290.CD-11-0224 (2012).

139 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).

140 Kaelin, W. G., Jr. Molecular biology. Use and abuse of RNAi to study mammalian gene function. *Science* **337**, 421-422, doi:10.1126/science.1225787 (2012).

141 Shao, D. D. *et al.* ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens. *Genome research* **23**, 665-678, doi:10.1101/gr.143586.112 (2013).

142 Ishino, Y., Shinagawa, H., Makino, K., Amemura, M. & Nakata, A. Nucleotide sequence of the iap gene, responsible for alkaline phosphatase isozyme conversion in Escherichia coli, and identification of the gene product. *Journal of bacteriology* **169**, 5429-5433 (1987).

143 Marraffini, L. A. & Sontheimer, E. J. CRISPR interference: RNA-directed adaptive immunity in bacteria and archaea. *Nature reviews. Genetics* **11**, 181-190, doi:10.1038/nrg2749 (2010).

144 Maeder, M. L. *et al.* Rapid "open-source" engineering of customized zinc-finger nucleases for highly efficient gene modification. *Molecular cell* **31**, 294-301, doi:10.1016/j.molcel.2008.06.016 (2008).

145 Cermak, T. *et al.* Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. *Nucleic acids research* **39**, e82, doi:10.1093/nar/gkr218 (2011).

146 Wood, A. J. *et al.* Targeted genome editing across species using ZFNs and TALENs. *Science* **333**, 307, doi:10.1126/science.1207773 (2011).

147 Shalem, O. *et al.* Genome-scale CRISPR-Cas9 knockout screening in human cells. *Science* **343**, 84-87, doi:10.1126/science.1247005 (2014).

148 Wang, T., Wei, J. J., Sabatini, D. M. & Lander, E. S. Genetic screens in human cells using the CRISPR-Cas9 system. *Science* **343**, 80-84, doi:10.1126/science.1246981 (2014).

# Chapter 2: The genomic evolution of endometrial carcinoma progression and abdominopelvic metastasis

## Abstract

Recent studies have detailed the genomic landscape of primary endometrial cancers, but their evolution into metastases has not been characterized. We performed whole-exome sequencing of 89 tumor biopsies including hyperplasias, primary tumors, and paired abdominopelvic metastases, to survey the evolutionary landscape of endometrial cancer. We identified novel recurrent alterations in primary tumors, including mutations in the estrogen receptor cofactor *NRIP1* in 12% of patients. We found that likely driver events tend to be homogenous across biopsy sites, with notable exceptions such as *ARID1A* mutations. Phylogenetic analyses in cases with multiple metastases indicated these metastases typically arose from one lineage of the primary tumor. These data indicate extensive genetic heterogeneity across different stages of endometrial cancer, reflecting changing evolutionary pressures during tumor progression.

Keywords: Cancer, Metastasis, Precursor, Endometrial cancer, Cancer genomics

## Introduction

Endometrial cancer is the most common pelvic gynecologic malignancy in industrialized countries; with increasing incidence attributed to the obesity epidemic[1]. Worldwide, there are annually more than 300,000 new cases and 75,000 deaths, with age standardized incidence rates ranging from 15/100.000 in developed regions to 6/100.000 in less developed regions[2]. Historically, endometrial cancers are divided into two major groups[3]: 75% of patients present with type I, endometrioid, tumors, often with adjacent regions with hyperplasia with atypia considered to represent precursor lesions. Type I tumors are often estrogen responsive and portend a good prognosis. Type II tumors are the non-endometrioid subtypes, including the carcinosarcomas, serous, clear cell and undifferentiated histologies, tending to occur in older, non-obese women. They are not estrogen responsive and carry a poor prognosis.

Recent large-scale sequencing studies of primary tumors of endometrioid and serous subtypes have provided evidence that the difference in phenotype is reflected in distinct molecular subgroups, also further refined in distinct molecular entities within each[4-6]. While these studies were able to detail the spectrum and patterns of somatic alterations across primary tumors, a comparative study of samples from endometrial hyperplasia, primary tumors, and paired metastatic lesions has not been performed. For example, it is not known whether metastases derive from the same or multiple lineages within the primary, and whether cancer cells require mutations that uniquely enable the metastatic phenotype. The extent to which genetic events observed in the primary biopsy reflect the heterogeneity that exists across the entire cancer is also unknown. Such information would be helpful in understanding the biological underpinnings of endometrial cancer progression and to determine treatment strategies that target features that are homogenous throughout individual cancers[7].

Here we address these questions in a collection of 89 extensively clinically annotated fresh frozen samples ranging from precursor lesions to primary tumors and paired abdominopelvic metastases from 45 cases. We analyzed somatic mutations and allelic copy number profiles between different biopsies from the same individual to reconstruct phylogenetic relationships and annotate putative cancer drivers across sites of disease. In addition, we reanalyzed data from The Cancer Genome Atlas (TCGA) using updated methods, which led us to identify novel recurrent mutations in *NRIP1* and patterns of microheterogeneity within biopsies that mimic heterogeneity across multiple tumor sites.

## Results

### Patient and sample cohort

Our cohort consisted of a population-based patient series from western Norway with extensive clinical annotation including complete follow-up information. We obtained fresh frozen tumor tissue from seven cases of complex atypical endometrial hyperplasia (CAH), 38 independent primary tumors, and 52 abdominopelvic metastases, totaling 89 biopsies from 45 individuals (Figure 1.1, Supplementary Table 1, and Supplementary Figures 1.1A-B). These included 23 endometrioid endometrial carcinomas (EEC) and 15 non-endometrioid endometrial carcinomas (NEEC). The median time between resection of the primary tumor and first metastasis was 17.5 months (range 1-99; median 17 months for EECs and 32 months for NEECs). Six and two patients received chemotherapy and external radiotherapy, respectively, between primary surgery and resection of metastasis.

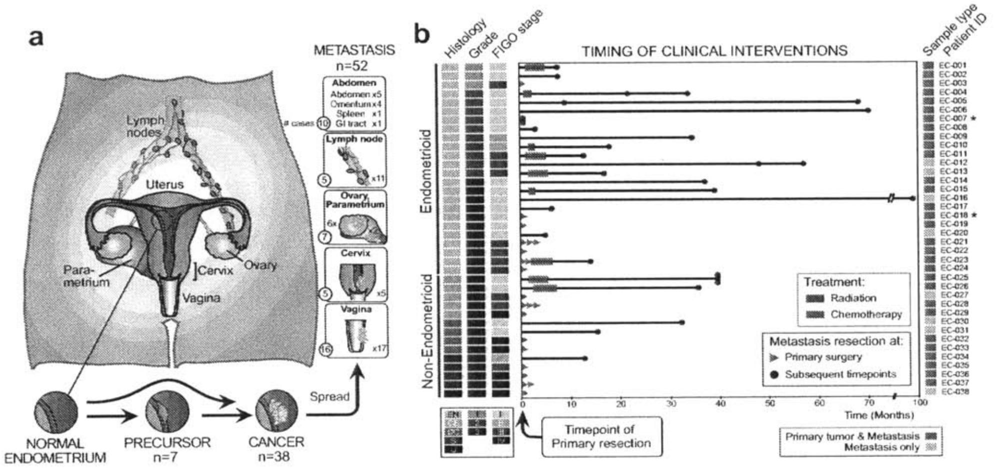
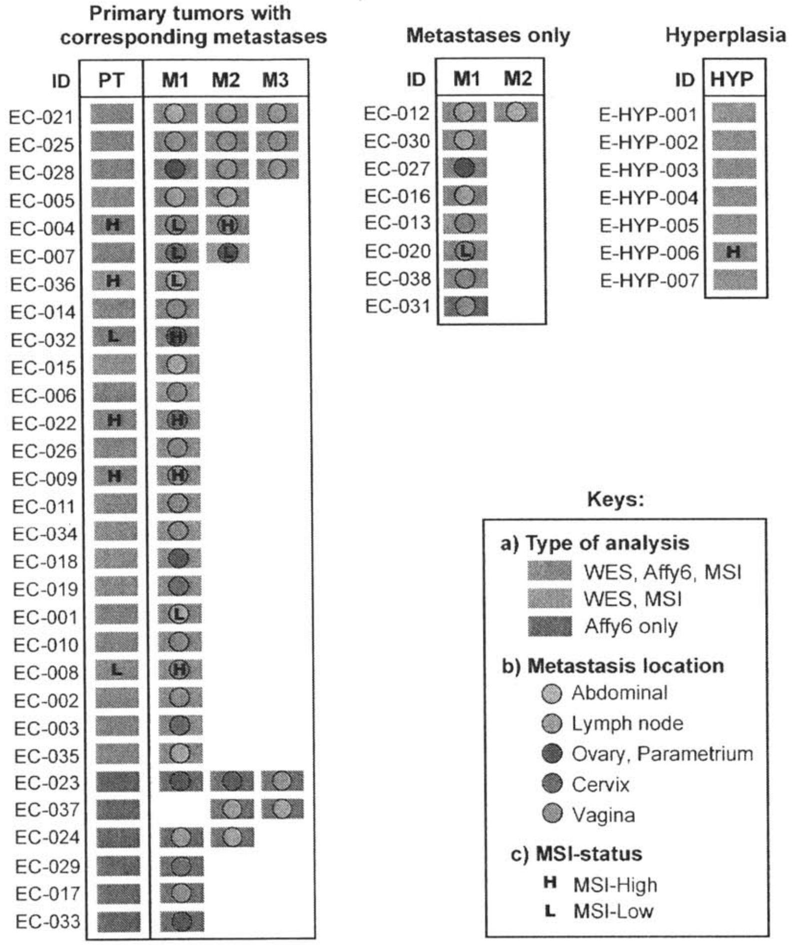


**Figure 1.1 Samples assessed.**

(a) Anatomic sites from which samples were obtained. (b) Histologic subtypes (E: endometrioid, CC: clear cell, SP: serous papillary and UN: undifferentiated carcinomas, and CS: carcinosarcomas), grade, FIGO 2009 stage at primary diagnosis, location of metastatic lesions, and timing of sampling and treatment after primary diagnosis. Stars on right indicate two cases that were clinically difficult to distinguish as metastatic or synchronous primary cancers at time of resection.



**Supplementary Figure 1.1. Overview of sample set and data generated.**

Samples were collected from hyperplasias (HYP; right), primary tumors (PT) with paired metastases from the same patient (left) and metastases without matching primary tumors (middle). Information about type of analysis (key a): Most biopsies underwent whole-exome sequencing (WES). Copy-number profiles were determined with Affymetrix SNP 6.0 arrays, in large overlapping with WES samples. Location of abdominal metastasis (M) for the samples are indicated by dots with colours corresponding to key b. For some patients multiple anatomically distinct metastases were obtained. Microsatellite instability status by a 7-marker panel was performed on samples analyzed by WES, as indicated by key c (No indication means MS-stable).

We performed whole exome sequencing (WES, mean 77x coverage) to characterize somatic mutations and copy-number alterations (SCNAs) for 72 biopsies from 38 of these individuals, including twenty-four with paired primary and metastatic lesions, five with more than one metastasis, and eight metastases without paired primaries, along with DNA from paired blood in all cases. All of these samples were also analyzed for microsatellite instability (MSI, Supplementary Figure 1.1D) by an established seven-marker panel, enabling classification according to the integrated molecular subgroups established by TCGA[4] (Figure 1.2A). We also analyzed SCNAs in 76 samples from 37 patients using Affymetrix SNP 6.0 arrays (Supplementary Figure 1.2). These included 59 samples from 30 patients that had also undergone WES, 10 additional metastases with paired primary tumors (from six cases, including three cases with more than one metastasis) and one unpaired metastasis (Supplementary Figure 1.1A).

### Novel significantly mutated genes and hotspots in endometrial cancer

The burden of somatic genetic alterations in our primary tumors was consistent with endometrial cancers profiled by TCGA. We observed similar rates of somatic mutation (minimum 40 to maximum 13,717) and SCNAs (Figure 1.2A and Supplementary Figure 1.3A-B), and an inverse correlation between both ($P$=0.005; Figure 1.2B)[8]. However, mutation rates of some of the most frequently altered genes differed, with higher mutation rates for *PPP2R1A, FGFR2, PIK3CA*, and *ARID1A*, and lower rates for *PIK3R1* in our dataset compared to TCGA. This may reflect different strategies for sample inclusion, with TCGA enriching for serous and endometrioid grade

3 lesions, and our approach enriching for patients with systemic disease (Supplementary Figure 1.4).

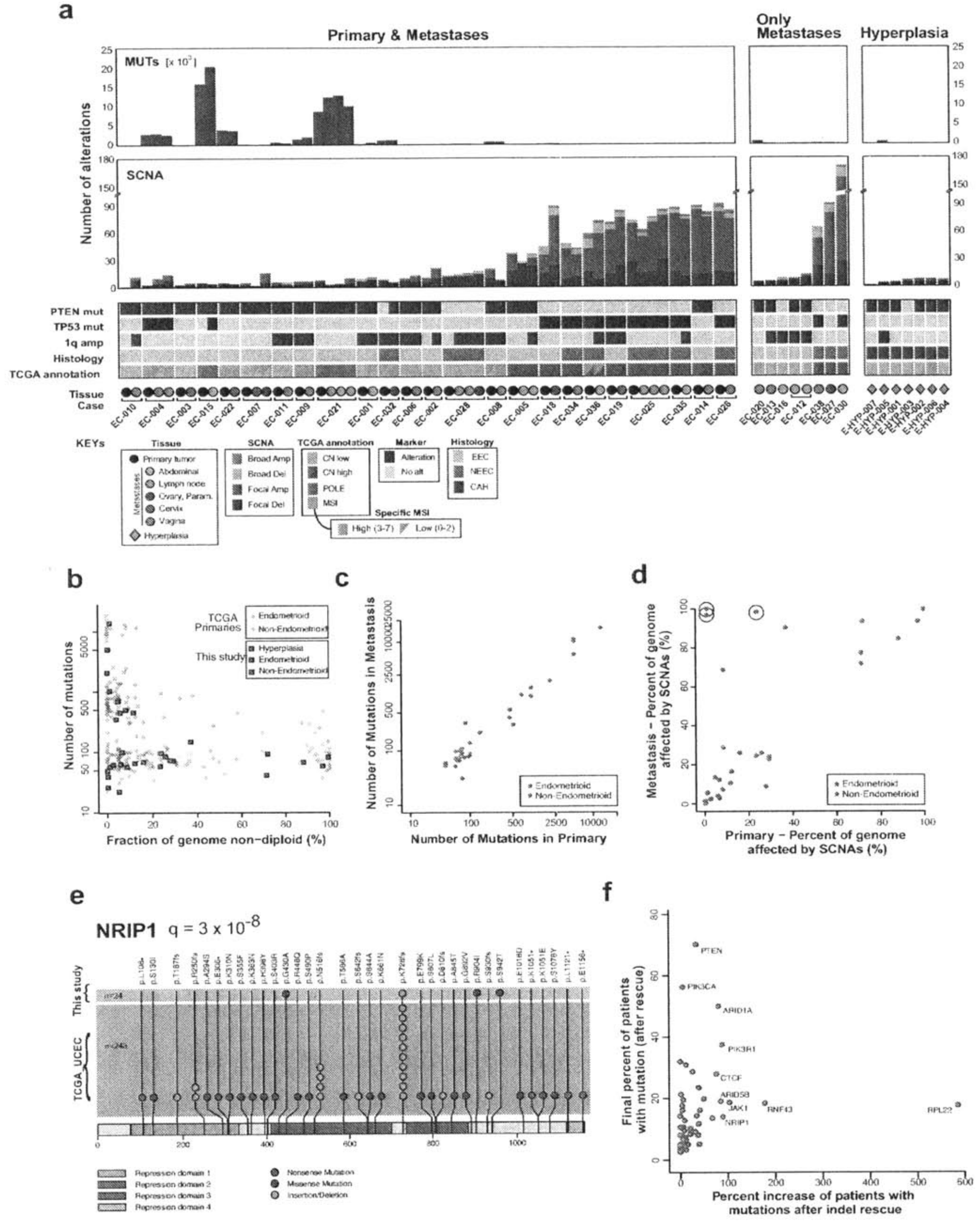
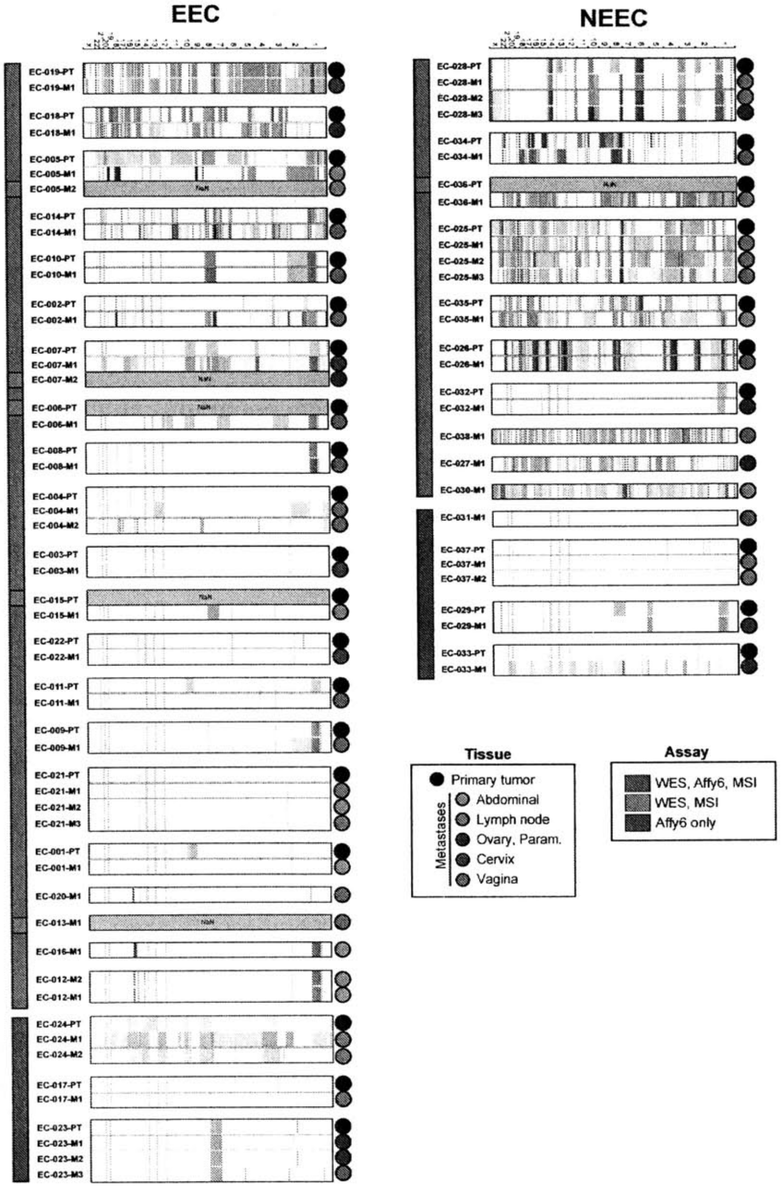


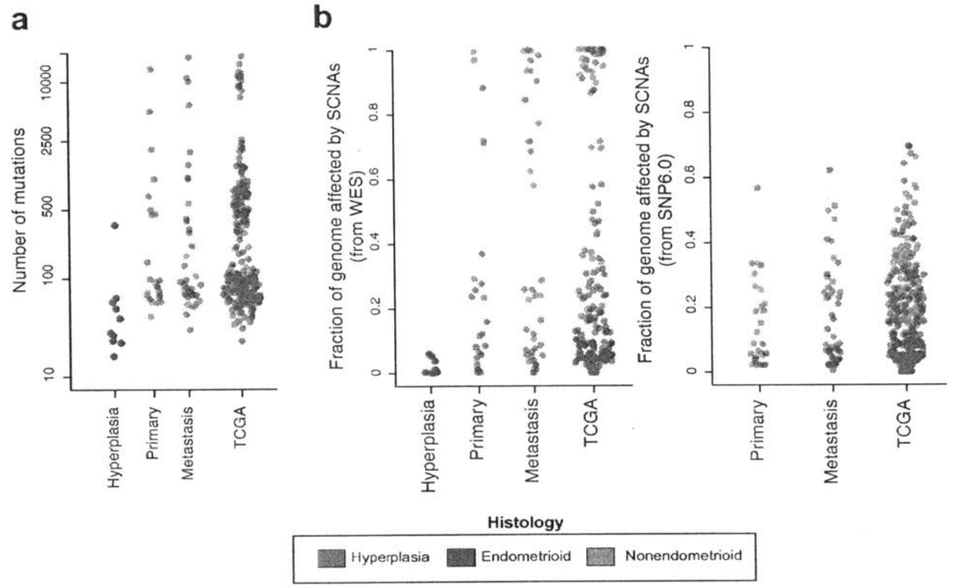
**Figure 1.2. Somatic genetic alterations in complex atypical hyperplasias and primary and metastatic endometrial carcinomas.**

(a) Number of exonic mutations (top) and SCNAs (middle) detected in each tumor biopsy. Tissue types and PTEN, TP53, and 1q amplification status are indicated on the bottom. (b) Number of mutations (y-axis) against fraction of the genome affected by SCNAs (x-axis) across atypical hyperplasias (green), primary lesions from endometrioid endometrial carcinoma (red) and non-endometrioid endometrial carcinoma (blue) primaries from our dataset (squares) and TCGA (dots). (c) Number of mutations detected in the primary tumor compared to their metastatic counterpart. (d) Fraction of the genome affected by SCNAs in metastases (y-axis) relative to paired primaries (x-axis). Circles indicate metastases that exhibit whole-genome doubling not observed in the primary biopsy. (e) Stick plot depicting mutations in NRIP1, a cofactor of the estrogen receptor. (f) Impact of indel rescue on the percentage of patients harboring mutations in known driver genes.



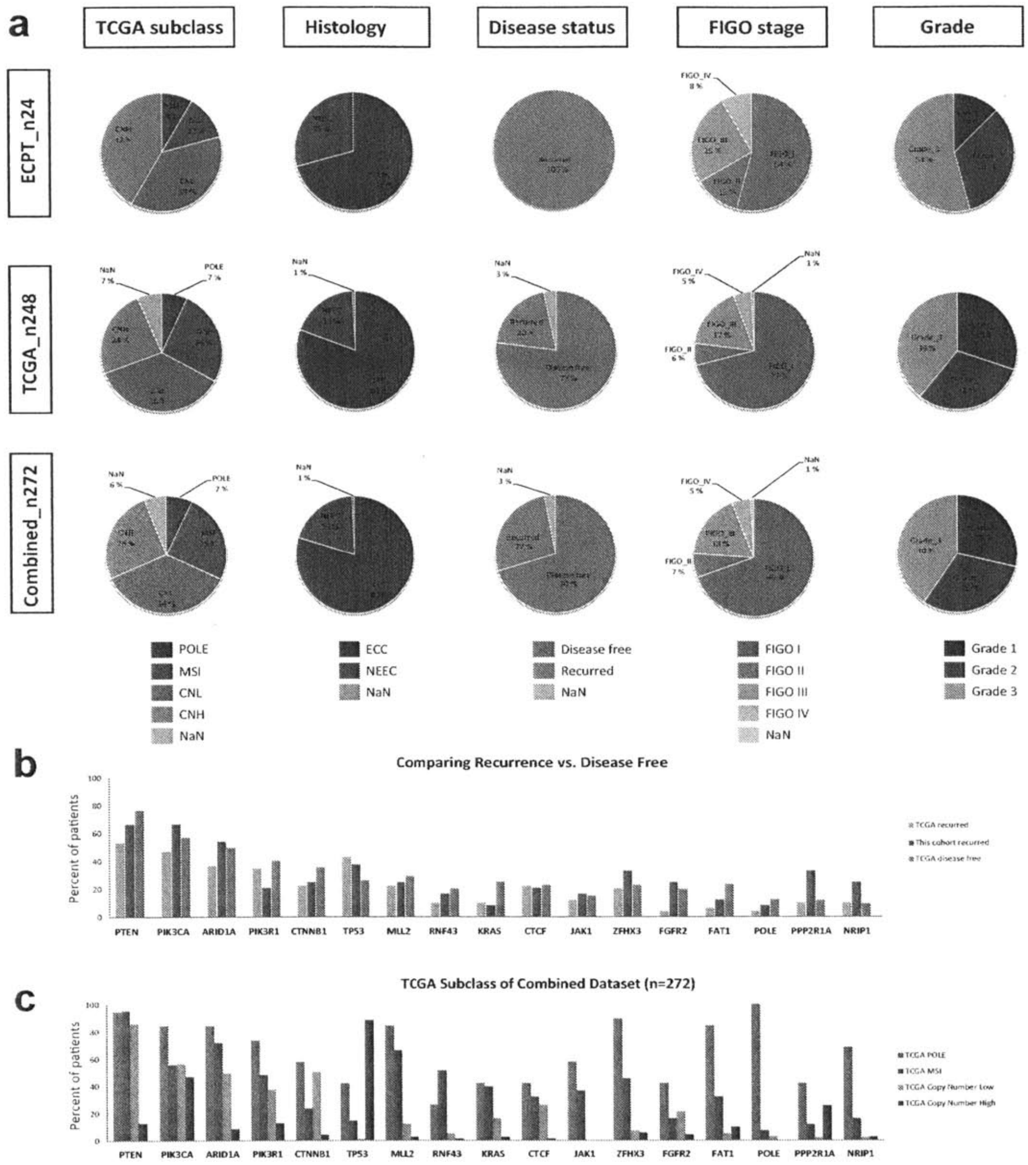
**Supplementary Figure 1.2. Somatic copy-number profiles by Affymetrix SNP 6.0 array.**

Heatmaps indicate amplifications in red and deletions in blue. Tumors are arranged by endometrioid endometrial carcinoma (EEC) vs. non-endometrioid endometrial carcinoma (NEEC) status. Assays performed and biopsy locations are indicated to the left and right of the heatmaps respectively.



**Supplementary Figure 1.3. Somatic alteration rates by tissue type.**

**a,** Number of exonic mutations detected in hyperplasias, primaries and metastases. Data from TCGA primary tumors are displayed on the right. **b,** Fraction of the genome that is affected by SCNAs in hyperplasias, primaries and metastases as determined by whole exome sequencing (left) or SNP6 array analysis (right). For WES-analyzed samples, allelic copy number profiles were computed and the fraction of the genome affected by SCNAs was defined as regions where there was not one copy of each allele. In this analysis, samples affected by whole genome doubling obtain higher values for the fraction of the genome affected by SCNAs than an analysis of total relative copy-number would yield. For samples analyzed with SNP 6.0 arrays, the fraction of the genome affected by SCNAs was defined as the fraction of the genome where the total relative copy level (normalized to a median of 2 per sample) was below 1.6 or above 2.5.



**Supplementary Figure 1.4. Cohort comparison to TCGA**

**a,** Clinical, histologic and molecular subgroup composition of our cohort and TCGA. Our cohort included more copy-number high subgroup tumors and more non-endometrioid tumors. (CNL= Copy-number low; CNH = Copy-number high. **b,** Percentage of patients harboring mutations in driver genes by disease status.

**c,** Percentage of patients with mutation in driver genes by TCGA molecular subgroup.

The burden of small insertions/deletions (indels) detected among primary endometrial cancers was higher than previously noted, particularly among microsatellite-unstable (MSI) carcinomas. MSI is prevalent among endometrial carcinomas and leads to high rates of these indels. However, due to their sequence context, these events are often observed at low-allelic fractions in the paired normal samples due to sequencing error, and are typically discarded by conservative analytic pipelines. We applied recently developed methods to detect highly recurrent indels that are enriched in tumor samples (Supplementary Figure 1.5A)[9]. As a result, we identified an average of 156 and 21 indels per MSI and non-MSI tumor, respectively, compared to 16 and 4.4 in prior analyses[10].

We used these new calls, as well as the combined dataset of our primary tumors and those of TCGA (272 primaries in total), to compile a catalogue of significantly mutated genes in primary endometrial cancer. We supposed these changes would increase our statistical power to detect novel driver alterations. We later use this catalogue to distinguish heterogeneity of likely driver vs passenger mutations between biopsies of primaries and metastases (see below).

We identified 49 genes that undergo significantly recurrent rates of mutation (Supplementary Table 2, Supplementary Figure 1.5B), including 16 that have not been previously described in endometrial cancer[4,10]. Of the 16 novel genes, four (*NFE2L2*, *ERBB2*, *U2AF1*, and *ALPK2*) have been found to be recurrently mutated in other primary cancer types using similar analytic methods to those used here[10].

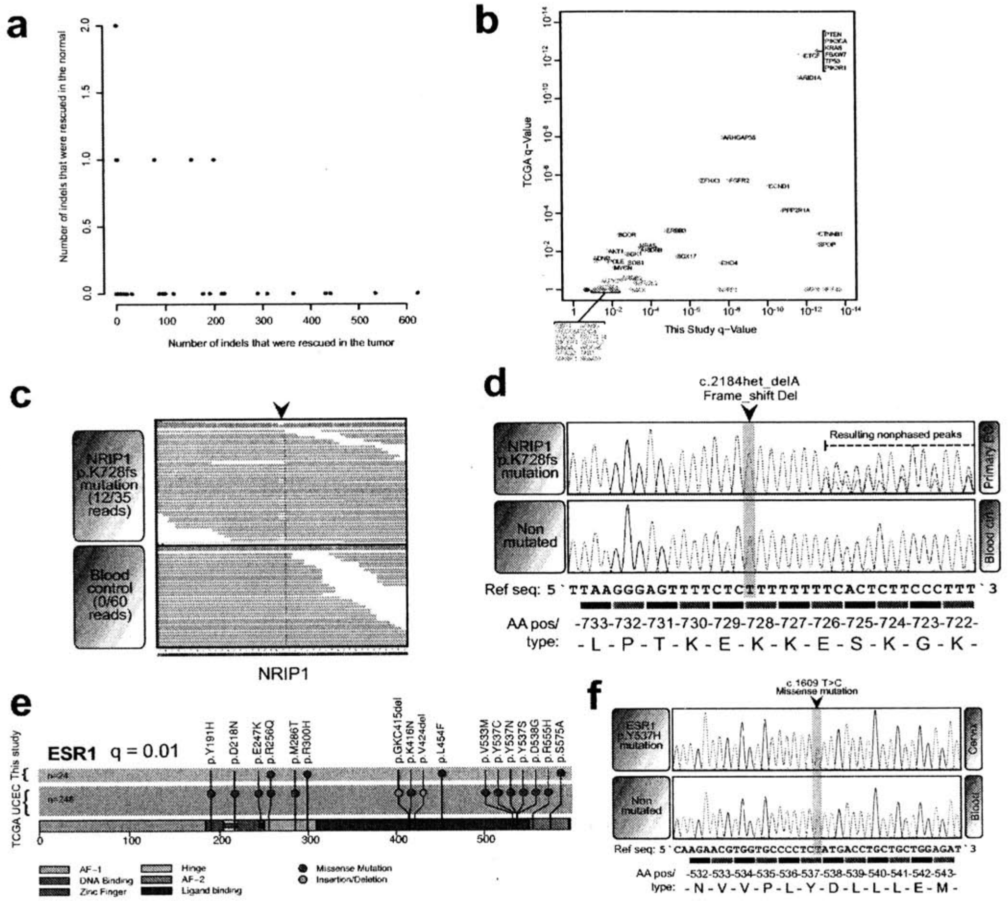
The other 12 novel significantly recurrently mutated genes included both *ESR1*, encoding the estrogen receptor alpha, and its binding partner *NRIP1*. Alterations in the estrogen signaling system, such as unopposed estrogen or alterations in estrogen receptor expression/function, are considered risk factors in endometrioid endometrial cancer[11], and recently recurrent rearrangements involving *ESR1* have been identified in breast cancer[12]. However, significantly recurrent point mutations in the estrogen pathway have not been previously described in cancers that had not received anti-estrogen therapy.

We found *NRIP1* mutations in 12.5% of cancers, concentrated in two highly recurrent sites, p.K728fs (n=11) and  p.N516fs (n=4) (Figure 1.2E, Supplementary Figure 1.5C-D). All of these recurrent site mutations were rescued indels, which may account for their absence in prior analyses, and all but two of these indels were in MSI samples (20% of MSI samples). Conversely, 14% of colorectal MSI samples analyzed by TCGA exhibited NRIP1 indels. We validated the existence of NRIP1 p.K728fs mutations by Sanger sequencing in an independent cohort of 37 primary endometrial cancers, of which two exhibited the mutation (Supplementary Figure 1.5D). NRIP1 is a coregulator that binds to the AF2 domain of the estrogen receptor and is essential for its transcriptional activity[13,14].

Mutations in *ESR1* were detected in 4% of cancers and clustered in the ligand-binding domain (Supplementary Figure 1.5E-F). These included Y537(C/N/S) mutations that have been shown to cause constitutive activation and resistance to tamoxifen therapy in breast cancer[15,16] and among breast cancers had only been detected after anti-estrogen therapy. However, the only patient in our cohort whose tumor harbored one of these mutations never received tamoxifen treatment, indicating that these mutations can occur outside this anti-estrogen therapy context.

Additional novel genes also included *MAX*, the binding partner of *MYC* family members. We identified two recurrently mutated sites in *MAX*: p.H28R (n=5) and p.R60Q (n=2, Supplementary Figure 1.6A). We also observed significantly recurrent mutations in *MYCN*, as previously noted[10], with a hotspot at p.P44L (n=5; Supplementary Figure 1.6B). Mutations of *MAX* and *MYCN* never co-occurred with each other or with amplifications of *MYC* or *MYCN* (Supplementary Figure 1.6D), though this mutual exclusivity did not reach statistical significance (p=0.36). The perturbation of *MYC* family members by translocations[17], amplifications[18] and viral mimicry[19] has been widely described, and indeed *MYCN* is recurrently amplified in endometrial cancers (Supplementary Figure 1.6C).

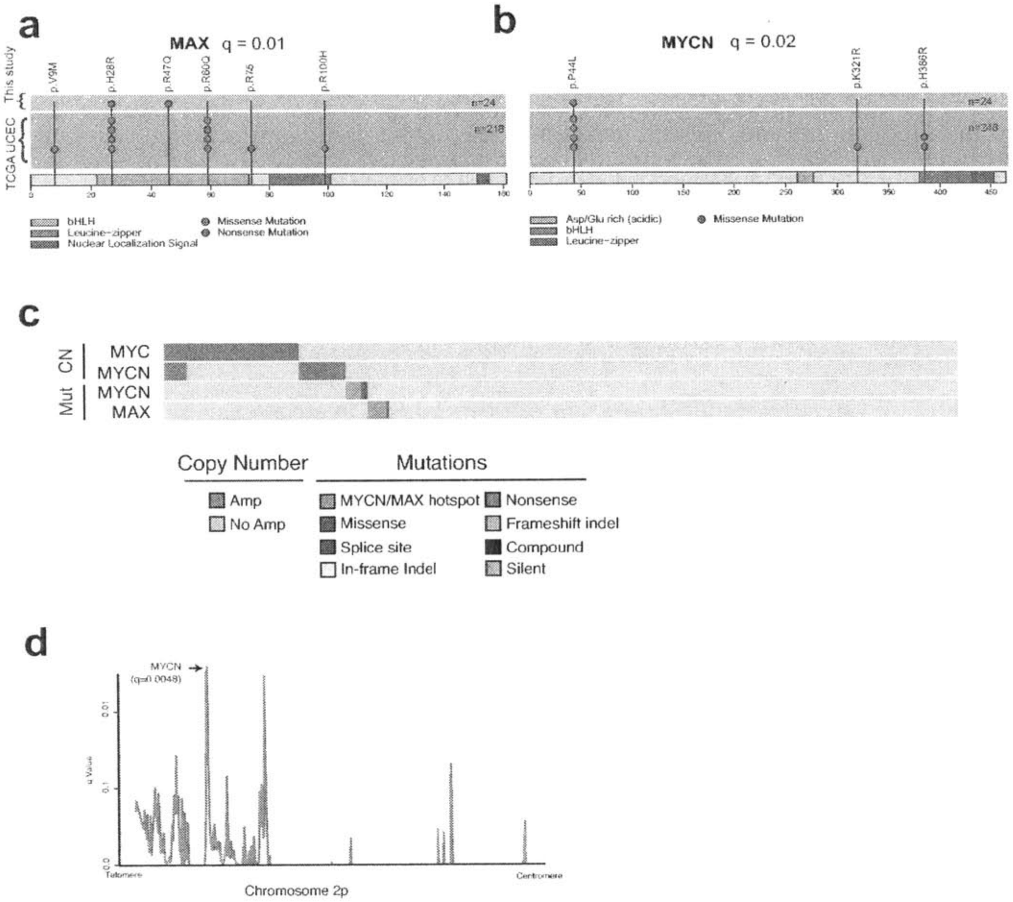
Even among genes previously noted to harbor significantly recurrent alterations, we often detected much higher rates of alteration than previously noted due to less conservative indel calling (Supplementary Table 2). For instance, we discovered *ARID1A* mutations in at least one biopsy from 49% of patients, a 40% increase over prior estimates. Genes in which polymerase slippage-associated indels have previously been identified, such as *RPL22*, *RNF43* and *JAK1*, showed even more dramatic gains (370%, 206%, and 163% increases respectively; Figure 1.2F). Conversely, the number of patients with biopsies exhibiting *PIK3CA* or *CTNNB1* mutations increased by only 5.6% and 0% respectively. Overall, we called 39% more mutations (all indels) across all genes and 54% more mutations in recurrently altered genes (p=0.28).



**Supplementary Figure 1.5. Indel rescue and estrogen pathway mutations.**

**a,** Accuracy of modification to indeLocator to "rescue" indels. Insertions/Deletions meeting modified indeLocator criteria were called in tumor-normal pairs (x-axis). To test the specificity of these calls, tumor/normal labels were swapped and the number of called indels was recorded for each normal-tumor pair (y-axis). A mean of 0.16 mutations were called in the swap, indicating a low false positive rate, whereas a mean of 62 mutations were rescued in the tumor-normal comparisons. **b,** Comparison of significance analysis results to TCGA. The 49 significantly mutated genes in our study (x-axis, orange dots) were plotted by their FDR q-value towards that obtained from the TCGA cohort (y-axis).

Genes along the x-axis are significantly mutated in this study but not in that of the TCGA. Red genes indicate novel significant genes in our study that are also recurrently mutated in other cancer types. Green genes indicate novel significant genes in our study that are not recurrently mutated in other cancer types. **c,** The NRIP1 p.K728fs mutation shown based on WES data for a vaginal metastasis in case EC-009 visualized by Integrated Genomics Viewer. **d,** Sanger sequencing chromatograms validating NRIP1 p.K728fs mutation in an external sample set of primary endometrial cancer. Two of 37 samples in the validation cohort carried the mutation. **e,** Stick plot of mutations observed in ESR1 by data source. Mutations cluster in the ligand-binding domain of the estrogen receptor. The Y537 mutations have previously been observed to confer tamoxifen resistance in breast cancers. However, we observed one patient in our cohort who developed a p.Y537H mutation without any prior hormonal therapy as shown in f. Chromatogram of Sanger sequencing validate this ESR1 p.Y537H hotspot mutation in this individual.



**Supplementary Figure 1.6. Frequency of co-alteration of MYC pathway copy number alterations and mutations.**

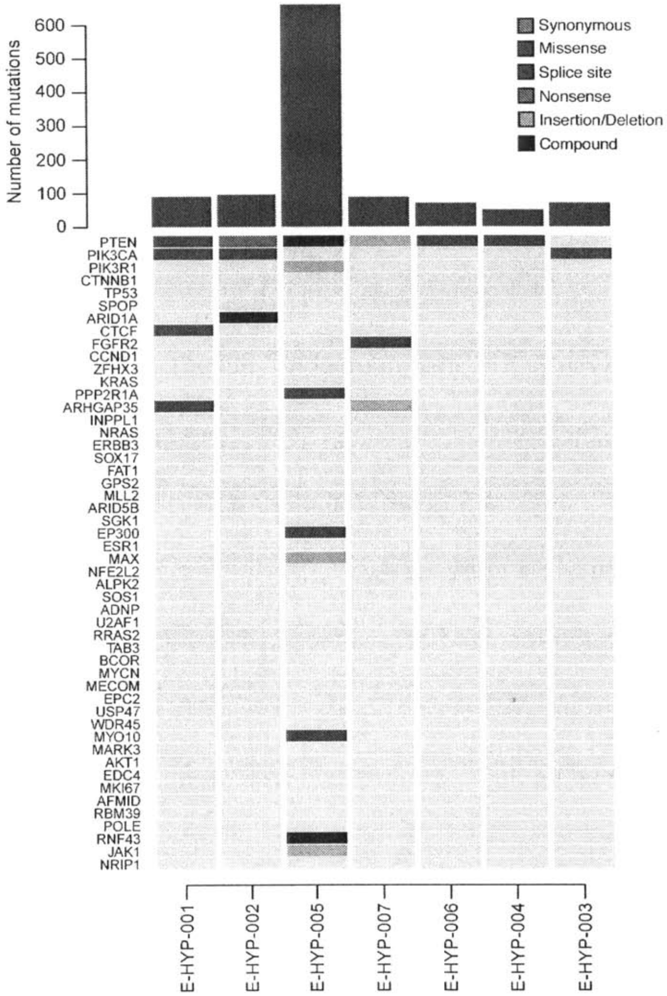
**a,** Stick plot of mutations in MAX by data source. We identified two hotspots in MAX, p.H28R (n=5) and p.R60Q (n=4). **b,** Stick plot of mutations in MYCN by data source. Two recurrent sites were identified, P44L (n=5) and H386R (n=2). **c,** All samples in our cohort and TCGA were analyzed to determine the patterns of coalteration between significantly altered MYC pathway members. Both MYCN and MAX mutations were mutually exclusive with other MYC pathway alterations. None of these anti-correlations reached statistical significance (p=0.29 and 0.36 for MYCN and MAX respectively). **d,**

MYCN is the only gene in a GISTIC peak in endometrial cancer. GISTIC q-values are plotted on the y-axis, with genomic position on the x-axis.

**PI3K pathway alterations predominate in complex atypical hyperplasias.**

Compared to primary tumors, complex atypical hyperplasias exhibited few somatic mutations with one highly-mutated exception (median 35 mutations per sample, range 15-345; Figure 1.2B; Supplementary Figure 1.6A). Less of the genome exhibited copy-number alterations in CAHs than primary tumors (median 0.4% of genome altered vs. median of 12.2% in primary tumors, $p=5x10^{-5}$; Supplementary Figure 1.3B).

The only significantly mutated genes in hyperplasias were *PTEN* and *PIK3CA*, and mutations in at least one of these (usually *PTEN*) were present in all seven samples (Supplementary Figure 1.7). Loss of heterozygosity of chromosome 10q, containing *PTEN*, was the only recurrent copy-number alteration (n=2). Other genes that were significantly mutated in primary endometrial cancers were found mutated in single hyperplasias, including *RNF43*, *FGFR2* and *ARID1A*; mutations of *ARHGAP35* occurred in two samples. Phosphatidylinositol 3-kinase (PI3K) pathway mutations have previously been shown to be prevalent in complex atypical endometrial hyperplasia[20]. These results indicate that no other genes are mutated at a similar rate.

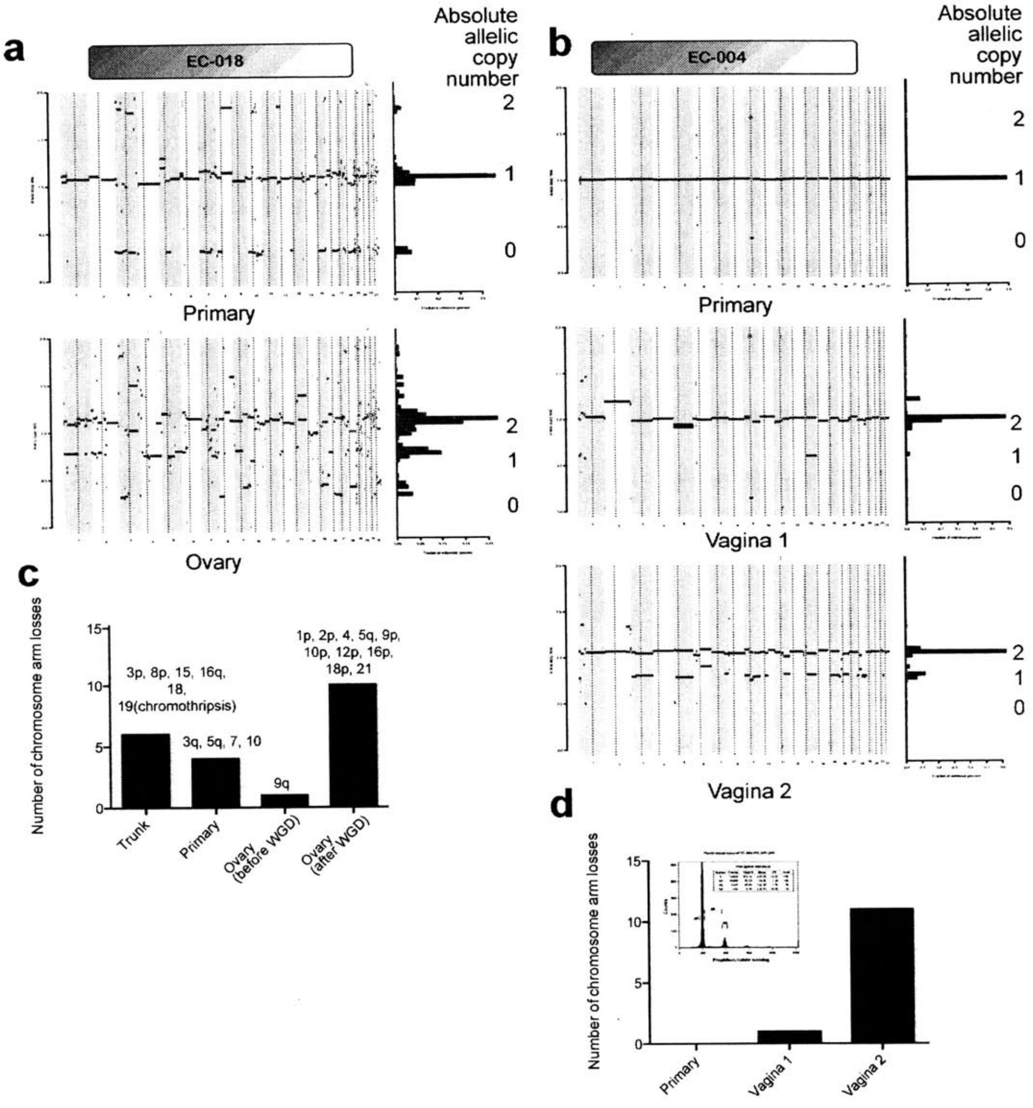


**Supplementary Figure 1.7. Mutations in endometrial hyperplasias.**
(top) Number of exomic mutations detected in endometrial hyperplasias. (bottom) Genes that were significantly mutated in primary tumors are depicted in the row of the comutation plot. Only PTEN and PIK3CA were mutated in every hyperplasia.

### Rates of genomic alteration are similar for primaries and metastases

Biopsies from primary tumors (PBs) and from their paired metastases (MBs) exhibited similar overall burdens of somatic genomic alteration (p=0.81). Metastases exhibited a median of 94.5 mutations per biopsy vs 93 in primaries (Supplementary Figure 1.3A); the number of mutations typically varied by only 6.2% between a primary and its paired metastasis (Figure 1.2C). Metastases exhibited a median of 14 SCNAs per tumor, vs 11 per primary (p=0.51); the number of SCNAs typically varied by 22% (Figure 1.2A). Similarly, the fraction of the genome that deviated from diploid was similar between most primaries and paired metastases, suggesting that aneuploidy develops early in tumorigenesis (Figure 1.2D).

Six phylogenies exhibited whole genome doubling (WGD), including two in which WGD was present in the MB but not in the paired PB. In both of these cases, the biopsies that underwent WGD exhibited more localized events. Indeed, in one case, we obtained two biopsies of the same metastasis at different times. Both exhibited WGD, but only the second biopsy exhibited increased rates of localized events (Supplementary Figure 1.8). Whole-genome doubling has previously been associated with increased rates of localized SCNAs in model systems[21] and in primary tumors[18,22]. These data support (though do not prove) a similar temporal relationship between WGD and subsequent localized SCNAs in human tumors.



**Supplementary Figure 1.8. Whole genome doubling (WGD) is followed by elevated rates of localized structural events**

a-b, Allelic copy number plots for phylogenies with non-truncal WGD for cases EC-018 and EC-004. While the primary biopsy for patient EC-004 harbored few SCNAs that would allow accurate determinate of biopsy purity, we were able to assign a purity of 67% from the variant allelic fractions of 1,027 mutations. At this purity, we were adequately powered to detect clonal SCNAs if they occurred. We detected neither WGD nor arm-level copy number events. c-d, Number of arm-level events occurring in each biopsy in a and b respectively. All single-copy losses were deemed to have occurred after WGD. Losses of two copies of the same allele could have occurred prior to or after WGD; we called all of these prior to WGD to be conservative. (inset, d) We performed flow cytometry to confirm that the primary in EC-004 was diploid. The second vaginal biopsy was obtained 12 months after the resection of the first vaginal metastasis and may therefore represent either recurrence of the first metastasis or an independent metastasis. Due to this ambiguity, EC-004 was excluded from multi-metastasis analyses.

**Half of mutations and SCNAs in metastases are not shared with paired primaries**

While the overall burden of somatic genomic alterations was similar between primaries and their matched metastases, only an average of 52% of the specific mutations (Figure 1.3A) and 50% of the SCNAs found in the MB were shared with the PB. Conversely, an average of 59% and 53% of mutations and SCNAs in the PB, respectively, were shared with each MB. The fraction of mutations that were unique to the MB tended to increase with the anatomical distance of the metastasis site from the endometrium (ρ=0.27; p=0.13; Supplementary Figure 1.9A), consistent with similar data in prostate tumors[23].

Overall mutation rates vary widely among endometrial cancers, but PBs with a high burden of somatic mutations typically shared a smaller fraction of mutations with their paired MB (p=0.006, Supplementary Figure 1.9B). Mutations that precede oncogenesis are likely to be ubiquitous throughout the tumor ("truncal"). These data indicate that the number of truncal mutations is more similar across endometrial cancers than indicated from single biopsy data, and that varying mutation rates detected across endometrial cancers from such data reflect in part varying rates of intratumoral heterogeneity.

Despite the dissimilarities between PBs and paired MBs, sufficient similarity exists to distinguish paired biopsies from a single cancer from synchronous primaries. We identified two cases (one is shown in Supplementary Figure 1.9C) in which two biopsies exhibited substantially different morphology, engendering clinical calls of synchronous primaries. In each case, however, sequencing of the biopsies revealed shared mutations, indicating the same cancer. Conversely, identical analyses of brain metastases have revealed clonally unrelated primary/metastasis pairs[24].

Among the 186 arm-level SCNAs (comprising most of a chromosome arm) we detected, 90 (48%) were heterogeneous across biopsies. Arm-level losses were more likely than gains to be truncal (58% vs 40%; p=0.02; Supplementary Figure 1.10A). Losses of 10q, harboring *PTEN* and 17p, harboring *TP53*, were truncal more often than expected given the overall rates (p=0.019 and 0.035). The most common arm-level gain, that of 1q, was only truncal in 6/12 cases (Supplementary Figure 1.10B). The heterogeneous events included both gains and losses of different alleles at the same genomic loci in the PB and MBs (Supplementary Figures 1.10C-F), indicating convergent evolution of SCNAs.

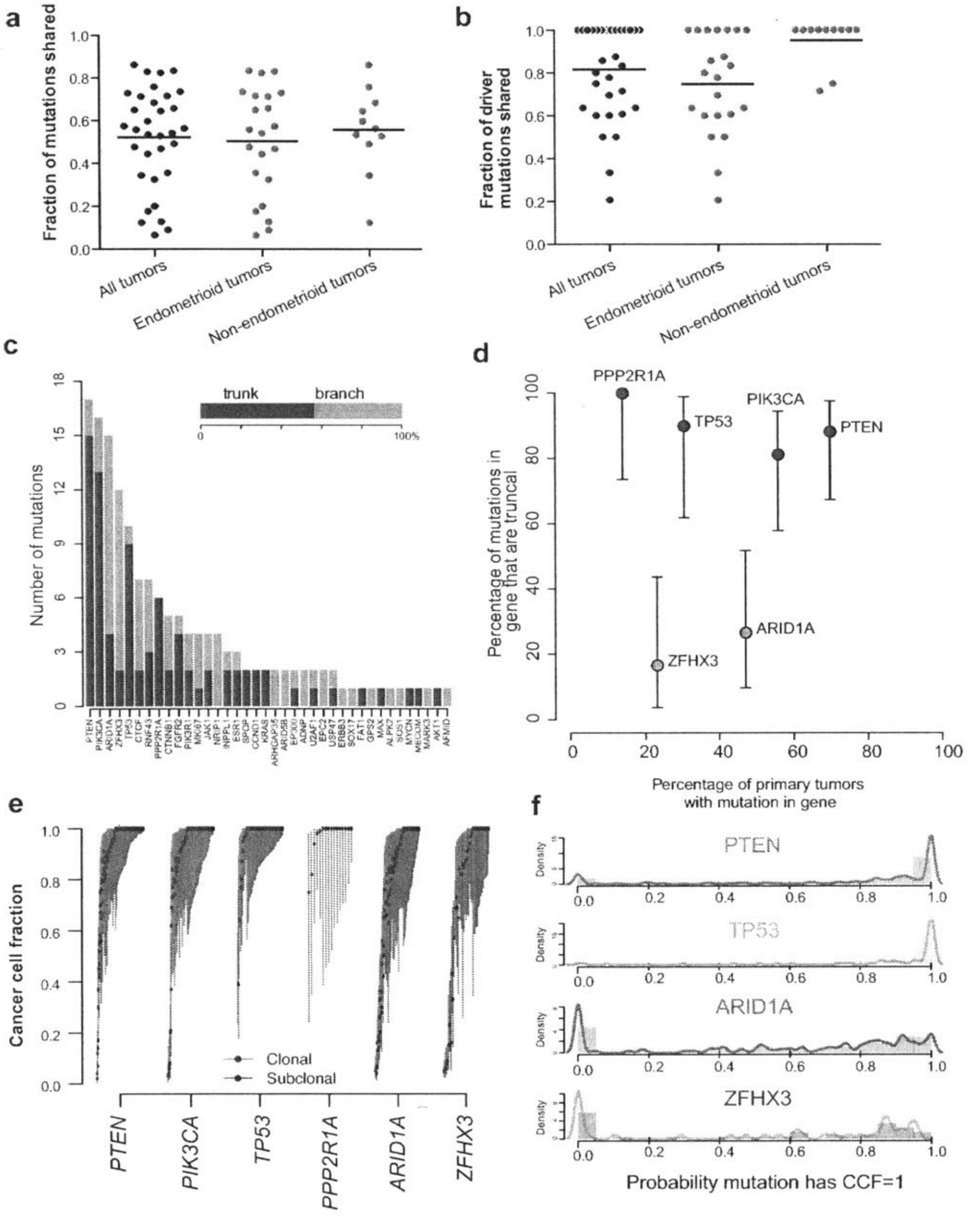
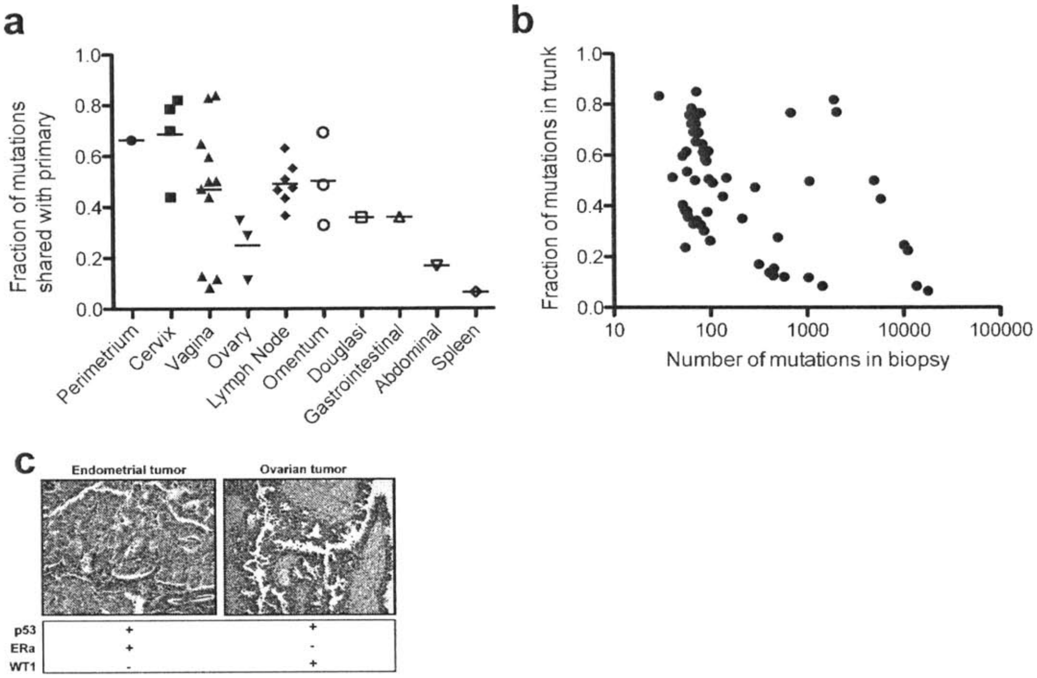


**Figure 1.3. Heterogeneity among somatic mutations.**

Fractions of (a) all mutations and (b) driver mutations detected in metastases that were truncal. Lines indicate the mean. (c) The number truncal and branch mutations involving the indicated driver genes. (d) Percentage of mutations that were detected in all paired biopsies in our dataset for frequently mutated driver genes. ARID1A and ZFHX3 are frequently mutated in the branches of phylogenies. (e) Distribution of the probability that each mutation detected in TCGA endometrial biopsies is clonal for the indicated genes.



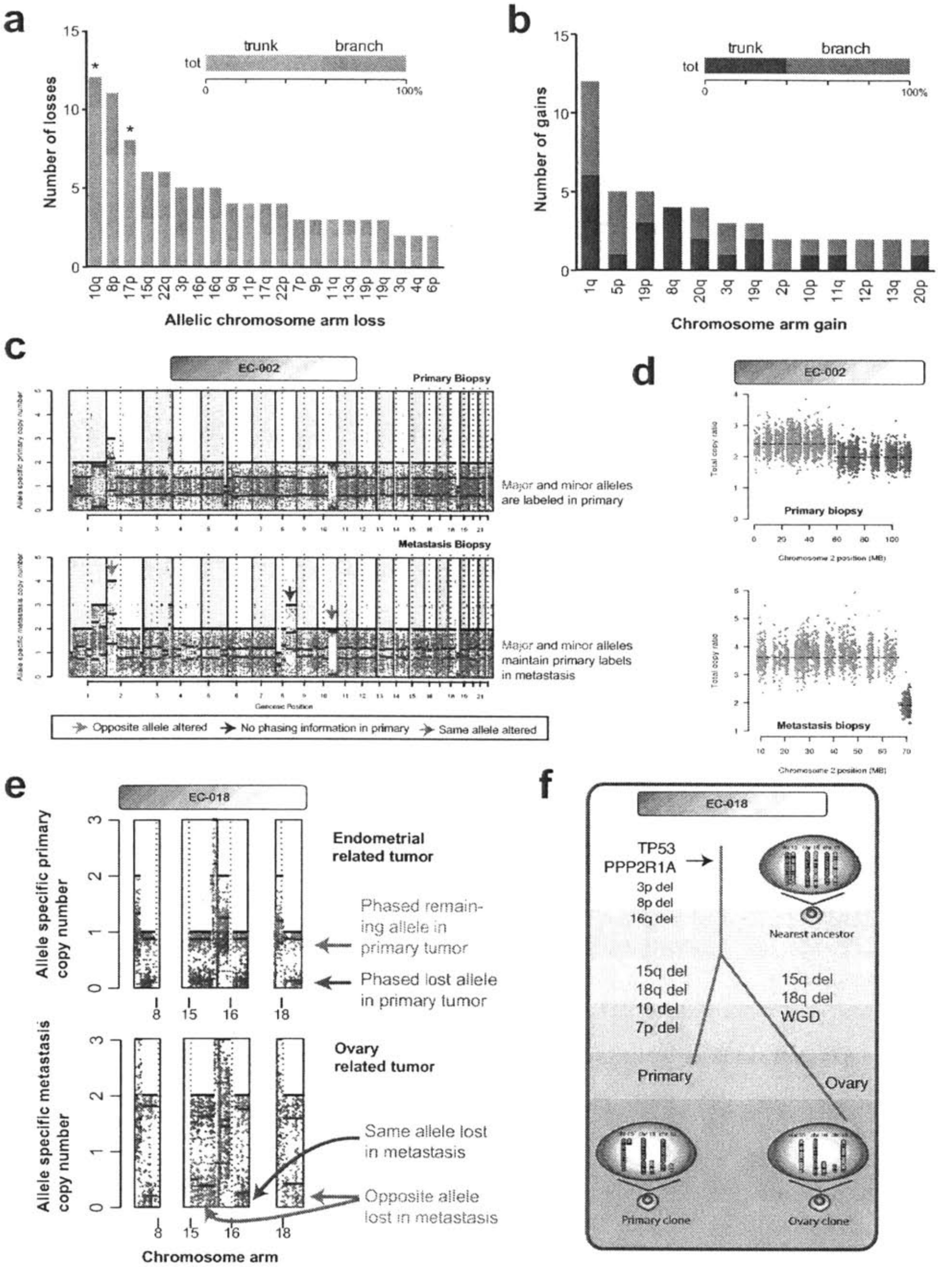
**Supplementary Figure 1.9. Trends in phylogenetic similarity.**

**a,** The fraction of mutations each metastasis shared with its paired primary decreased with the distance to the primary tumor. This trend was not statistically significant (p=0.27). For purposes of statistical analysis, sites were grouped into four categories:
close mets= Cervix, Perimetrium
gynecologic mets= Vagina,Ovary
semi-gynecologic mets= Douglasi, Lymph Node
abdominal mets= Omentum, Gastrointesinal, Spleen, Abdominal

**b,** The fraction of mutations each biopsy shared with its paired biopsies (y-axis) against overall mutation rate (x-axis). Highly mutated tumors shared proportionally fewer mutations across biopsies than less mutated tumors (p=0.0003)

**c,** EC-018 was reclassified during this study as two independent synchronous cancers by pathologists, but the two biopsies shared 76 mutations, indicating a single cancer. (left) Tissue (HE) section from the uterine cavity showed glandular morphology consistent with endometrioid adenocarcinoma. Immunohistochemistry (IHC) indicated the tumor to be positive for p53 and estrogen hormone receptor (ER alpha) and negative for WT1 (Wilms` Tumor 1).

(right) Tissue (HE) section from the right ovary showed a papillary morphology consistent with a high grade serous adenocarcinoma. Immunohistochemistry indicated that this tumor was positive for p53 and WT1 and negative for ER.



**Supplementary Figure 1.10. Phylogenetics of somatic copy number alterations.**

**a-b,** Phylogenetic placement of chromosome arm losses (a) and gains (b). * indicates bias towards truncal alteration (binomial test p<0.05).

**c,** Haplotype tracking across tumor biopsies for EC-002 reveals convergent SCNAs. Copy-number adjusted allelic fractions from germline heterozygous sites are plotted for the PB (top) and MB (bottom). All variants at germline heterozygous sites are labeled in the PB by whether more reads of either allele are present (major allele, red). These labels were maintained in the MB (the sites labeled as red in the PB are red in the MB) and the same analysis was performed in the MB. Where a genomic segment is deleted or amplified, a haplotype is revealed. Often, the same haplotype is gained/lost in both biopsies (eg 10q in this example, red arrowhead). Occasionally, genomic regions undergo convergent copy number gains/losses in which the opposite allele is gained/lost in the MB compared to the PB (eg 2p in this example, blue arrowhead).

**d,** The convergent SCNAs identified by allelic analysis in EC-002 also displayed different breakpoints in total copy ratio data.

**e,** Allele-specific copy-number profiles at germline heterozygous sites in primary and metastasis of EC-018.

**f,** Schematic outlining process of convergent loss of chromosome arms 15q and 18q indicated by data in e. WGD = Whole Genome Doubling.

### Rates of intratumoral heterogeneity among common drivers

An average PB shared 87% of its driver mutations with its paired MB (Figure 1.3B). We defined "driver mutations" as any non-silent mutation of a gene listed in Supplementary Table 2.2; we identified 1-22 (median 3) truncal driver mutations per patient. Among the 24 PBs, 11 (46%) contained driver mutations not detected in the paired MBs. The overlap among drivers exceeded the overlap in the overall number of mutations between primaries and metastases (mean 87% vs 59%, p=1.2 x $10^{-7}$). This suggests that the fraction of new mutations that are drivers decreases along the length of the evolutionary tree.

The rate at which driver mutations were shared across all biopsies varied by gene, ranging from 0%-100% (Figure 1.3C). For six genes, we had adequate power to determine whether mutations affecting them were truncal more or less often than the average rate among drivers ("trunk-biased" and "branch-biased" respectively, Figure 1.3D). Of these genes, *PTEN*, *PIK3CA*, *TP53*, and *PPP2R1A* were trunk-biased (Fisher's two-tailed p=0.03, p=0.1, p=0.05, p=0.04), suggesting they are early events and, in the cases of *PTEN* and *PIK3CA*, consistent with their prevalence among CAHs. The remaining two genes displayed significant branch-bias: *ARID1A* and *ZFHX3* (p=0.013 and p=0.005, respectively). Mutations in *ARID1A* were only truncal in 27% of phylogenies, vs. 60% among other drivers.

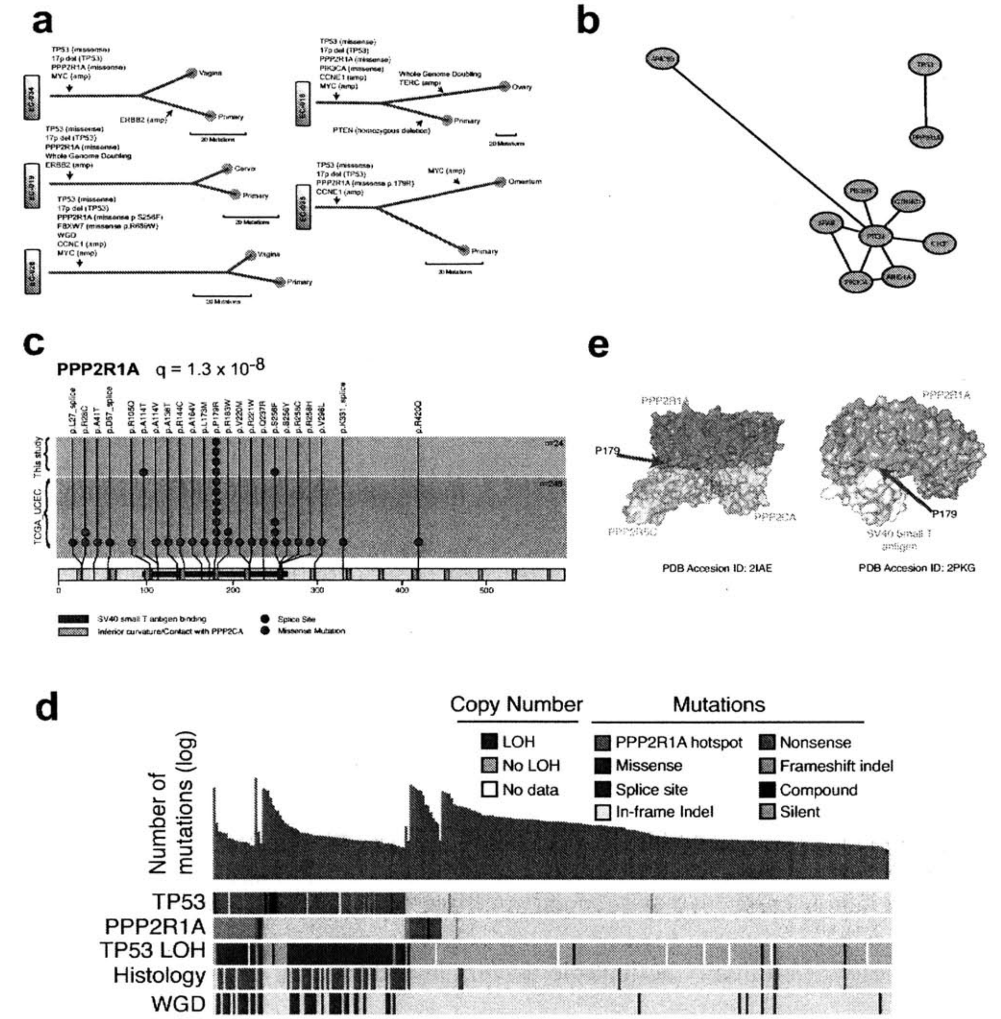
Analysis of heterogeneity within biopsies supports the finding of frequent heterogeneity of mutations in *ARID1A* and *ZHFX3*. We quantified the fraction of cancer cells that carry each mutation in the TCGA endometrial cancer dataset, using joint estimates of the allelic fractions and copy-numbers of those mutations[22,25,26]. While mutations of *PTEN* and *TP53* were almost exclusively clonal (81% and 92% of cases, respectively), only two-thirds of *ARID1A* and 61% of *ZFHX3* mutations were clonal (dissimilar rates from *PTEN*; $p=1.6\times10^{-11}$ and $6.9\times10^{-14}$ for *ARID1A* and *ZFHX3*, respectively; Figure 1.3E).

This relative heterogeneity of *ARID1A* mutations in endometrial cancers is mirrored by heterogeneity of mutations across BAF complex members in other cancers. The BAF complex, which contains ARID1A, regulates the organization of chromatin and contains several components that are significantly mutated across diverse cancer types[10,27]. Mutations of BAF complex members displayed the most phylogenetic heterogeneity of all known driver mutations in multi-region sequencing studies including mutations of *PBRM1* in renal cell carcinoma[28], *SMARCA4* in gliomas[29], and *ARID1A/SMARCB1* in meningiomas[30,31]. Collectively, these observations suggest that perturbation of BAF complex function enhances cancer cell fitness in the context of pre-existing alterations and tumor growth.

Among seven phylogenies with *PPP2R1A* mutations, five also exhibited *TP53* mutations ($p=0.02$, Supplementary Figure 1.11A), and in all cases both the *PPP2R1A* and *TP53* mutations were truncal. We validated the association between *PPP2R1A* and *TP53* mutations in the TCGA dataset using a network analysis approach that takes into account varying degrees of genomic instability across cancers[18,32] (Supplementary Figure 1.11B, Supplementary Table 3). The only gene whose mutation was positively correlated with *TP53* was *PPP2R1A* ($p=6\times10^{-7}$). These two gene nodes formed an isolated network corresponding to a subset of nonendometrioid tumors.

We determined from the analysis of all 272 endometrial cancers that *PPP2R1A* mutations tend to cluster in two hotspots: p.P179R (n=11) and p.S256F (n=4, Supplementary Figure 1.11C). The association between *PPP2R1A* and *TP53* mutations among the 272 primaries was also primarily due to *PPP2R1A* hotspot mutations: 17 of 19 tumors with *PPP2R1A* hotspot mutations exhibited *TP53* mutation ($p = 5.8\text{x}10^{-8}$; one of the remaining two tumors exhibited loss-of-heterozygosity at *TP53*, consistent with a cryptic inactivating event; Supplementary Figure 1.11D) whereas only three of 16 tumors with non-hotspot *PPP2R1A* mutations exhibited *TP53* mutations (p=0.56).

Among the 66 potentially clinically actionably genetic alterations[33] detected in our cohort, 50 were shared between PBs and MBs (Supplementary Figure 1.12). Ten of the alterations were detected uniquely in the paired MBs, whereas only six were detected in the paired PB. The enrichment of actionable alterations in MBs may represent sampling bias: in some cases, we sequenced multiple MBs and only a single PB. The relative homogeneity of clinically actionable drivers in endometrial cancer contrasts with recent analyses of brain metastases[24].



**Supplementary Figure 1.11. PPP2R1A and TP53 mutations cooperate to promote transformation.**

**a,** Phylogenies with recurrent PPP2R1A mutations. **b,** Network model of gene comutation. Distances between nodes reflect negative log q-values for correlated mutations. **c,** Stick plot of PPP2R1A mutations by data source. Due to the large number of likely passenger mutations in POLE/ultramutated tumors, these samples were removed. **d,** Comutation plot of TP53 mutation, PPP2R1A mutation, TP53 loss of heterozygosity (LOH), histology (blue: NEEC, grey: EEC) and whole genome doubling (WGD, purple). PPP2R1A hotspot mutations are positively correlated with TP53 loss (p= 6x10-8). **e,** The P179R mutation falls at the binding site of PPP2R1A to PPP2R5C (left).The P179R mutation overlaps the SV40 small T binding site to PPP2R1A (right)

**Multi-metastasis sequencing suggests most metastases arise from a single lineage in the typical metastatic endometrial cancer**

We determined phylogenetic relationships between tumor biopsies using both mutations and allelic copy-number alterations. In six of seven cases with multiple MBs, all MBs were more closely related to each other than to the PB (monophyly), suggesting the metastases all arose from a small fraction of the primary. In the seventh case, however, one of the MBs was more closely related to the PB than to the other MBs (polyphyly; Figure 1.4A, Supplementary Figures 1.13-1.14).

These results suggest that most metastases arise from one lineage in the primary tumor. In cases of two MBs and one PB, the phylogeny could have three configurations: either the PB is the most distantly related biopsy, or either of the MBs is the most distantly related biopsy. Assuming the MBs arise from independent lineages, all of these configurations would be equally likely, so that monophyly would be observed in one-third of cases. In cases of three MBs and one PB, the phylogenetic tree could include two clades with two members each (with a one-third probability[34]) or clade with one and three members, respectively (with a two-thirds probability[34]). Only the latter is consistent with monophyly, and only if the PB is in the clade by itself (with a probability of one-quarter), for a one-sixth probability of monophyly overall. Three of our seven cases included two MBs and the other four included three MBs. We would typically expect one or two (expectation value 1.7) cases of monophyly in this setting, a significantly different result from the six of seven cases of monophyly observed (p=0.001).

Even the existence of two independent lineages, each contributing half of observable metastases in the typical metastatic endometrial cancer, is inconsistent with our observed data. In such a case, and considering phylogenies with two MBs, these MBs would derive from different lineages in half of cases, each associated with a one-third rate of monophyly, and from the same lineage in half of cases. We assume that if the MBs derive from the same lineage, they necessarily exhibit monophyly (a conservative assumption: it is possible that the PB would by chance represent that same lineage within the primary tumor, in which case polyphyly would still be possible). Using similar considerations for phylogenies with three MBs, we calculate that four of the seven phylogenies would exhibit monophyly, a significantly different result from that observed (p=0.018). The single observed case of polyphyly might represent a cancer with more than one independent metastatic lineage, or a case in which the PB happened to sample descendants of the metastatic lineage within the primary tumor.

We performed similar calculations on phylogenies from prostate and pancreatic cancers for which genome-level sequencing had been performed on two to ten MBs and one to nine paired PBs[23,35,36]. Prostate and pancreatic cancers exhibited polyphyly in one of five and three of five cases, respectively. Like endometrial cancers, the results for the prostate cancers are inconsistent with equal contributions to metastasis of two or more lineages (p=0.05). Pancreatic cancer metastases exhibit monophyly approximately as often as expected if they arose from two independent lineages (p=0.98).

The finding that observed metastases in endometrial cancers tend to be associated with a single lineage indicates that cells that metastasize share a feature that is associated with genetic ancestry. This may be a genetic event that enables metastasis (see below), but it is consistent with other explanations. For example, members of a single lineage may happen to be located in an environment that is conducive to metastasis[37], or by chance have initiated a first metastasis, an act that may somehow enable them to form additional metastases.

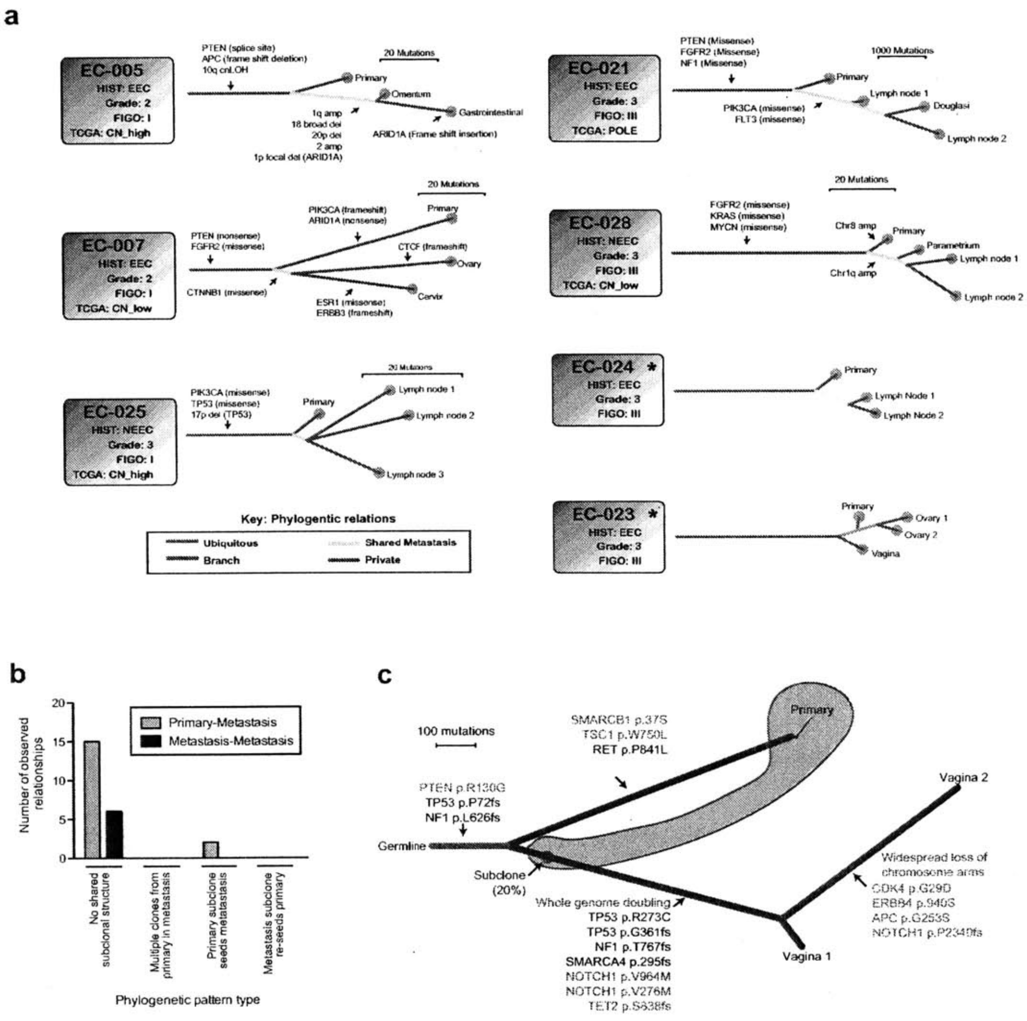
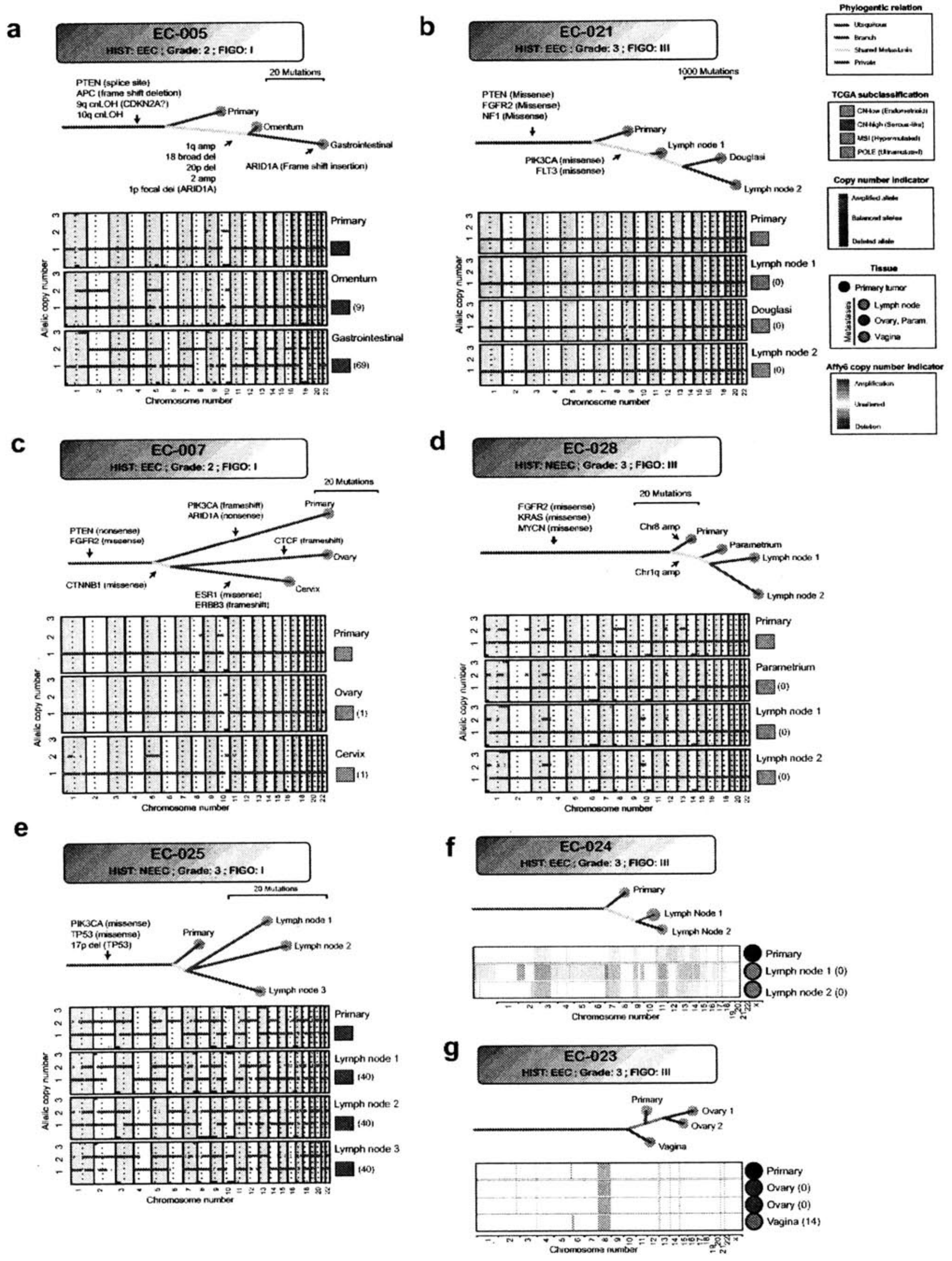
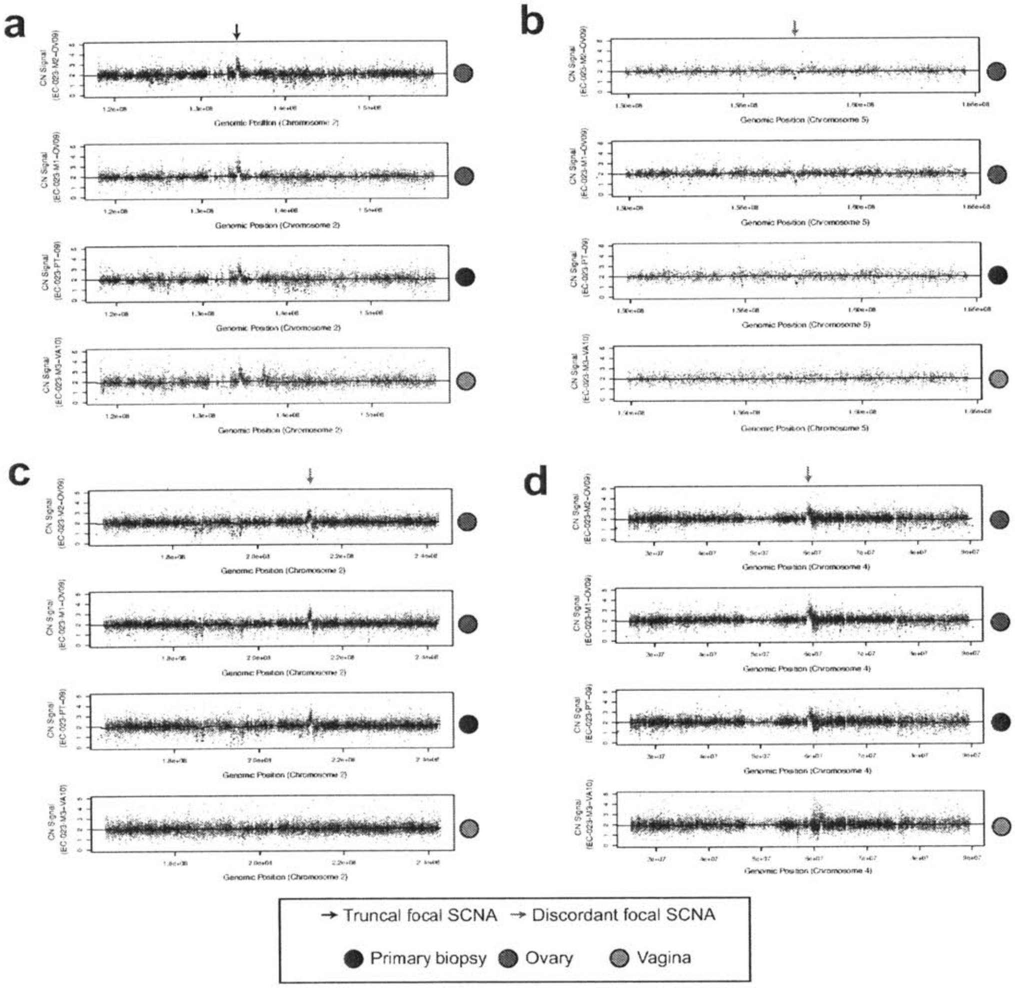


**Figure 1.4. Phylogenetic trees for tumors with more than one metastasis.**
(a) The labeled alterations constitute a subset of the alterations that distinguish between the indicated branches. Asterisks indicate trees that were derived from SNP6.0 array data. (b) Summary of the 2D phylogenetic pattern observed across biopsy comparisons. (c) (top) 2D phylogenetic plots depicting the seeding of metastases from a subclone detected in the primary biopsy for case EC-004. (bottom) Proposed evolution of EC-004. CCF densities for the mutations supporting the metastasizing subclone (purple) are presented in Supplementary Figure 1.17.



**Supplementary Figure 1.13. Phylogenetic trees and copy number alterations for trees with multiple metastases.**

**a-e,** (Top) WES-derived tumor phylogenies from the combined mutation and copy number alteration profiles from individual biopsies. The trees display the evolutionary relationship between primary-metastasis samples for a single patient: alterations ubiquitous to all biopsies are blue, yellow indicates mutations shared by all metastases, and red indicates mutations specific to only one biopsy (private). **a-e,** (Bottom) Allelic copy number profiles. Molecular classifications for all primary and metastatic lesions are denoted in colored squares to the right of each copy number panel in line with the TCGA uterine cancer classification scheme, followed by time (months) between resection of primary and metastatic lesions. Histological type, grade and FIGO stage is given for each individual. Further clinical details are in Supplementary Table 1. Scale bars for mutations are given for each case. Non-silent driver mutations are labeled in trees **a,d**, and **e**. Due to the extremely high burden of mutations in b and c, only specific mutations are labeled (those with amino acid changes labeled as recurrent in COSMIC). **f-g,** Phylogenies of multi-met tumors profiled with Affymetrix SNP 6.0 Arrays.



**Supplementary Figure 1.14. Focal copy number alterations demonstrate separate origins of metastasis in patient EC-1495.**

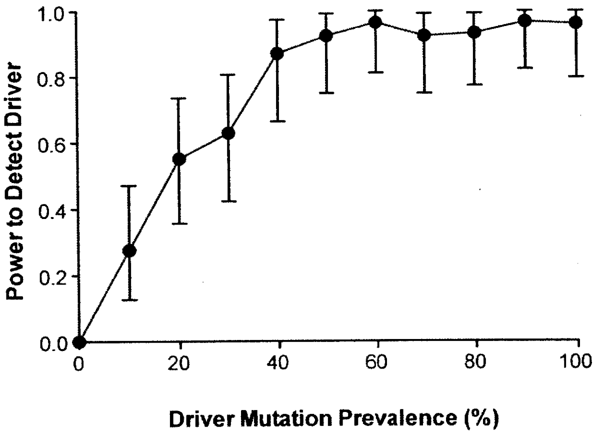
**a,** A focal SNCA on chromosome 2 that is shared across all biopsies demonstrates that all biopsies were of comparable purity and similarly able to detect SCNAs.

**b-d,** Focal SCNAs that are shared by two ovarian metastases and the primary biopsy but absent in the biopsy from a vaginal metastasis. These SCNAs indicate that the ovarian biopsies and the primary biopsy have a shared common ancestor that did not give rise to the vaginal biopsy.

**No evidence of ubiquitous metastasis-specific mutations**

We did not discover any significantly recurrent metastasis-specific mutations. It is possible that infrequent metastasis-specific drivers remain undetected. To assess the power to detect a hypothetical metastasis specific driver that occurs in a given percentage of metastases, we "spiked" mutations corresponding to a hypothetical driver into our dataset and then assessed whether we recovered the gene in question (Supplementary Figure 1.15). Our power exceeded 90% for genes mutated in at least 50% of metastases and remained greater than 50% for genes mutated in at least 20% of metastases. These results indicate there are no metastasis-specific exomic mutations that recur in greater than 50% of abdominopelvic metastases.

We also observed no significant excess of known driver mutations among metastases. Among our 24 phylogenies, 19 exhibited the same number of driver mutations in the primary and metastatic biopsies, four exhibited more drivers in the metastasis, and one exhibited more drivers in the primary (p=0.38).

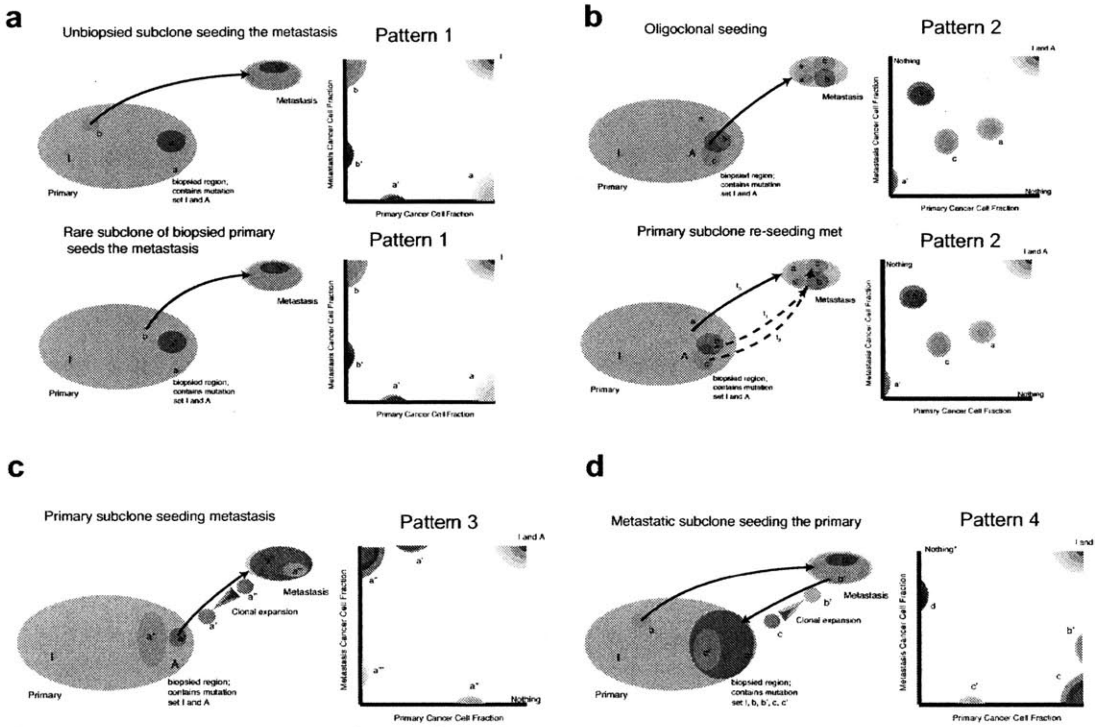


**Supplementary Figure 1.15. Power to detect metastasis specific drivers in this study.** Hypothetical drivers of metastasis were spiked into the set of mutations detected only in metastasis biopsies that were used to search for metastasis-specific drivers. The fraction of hypothetical drivers was varied and the rate at which MUTSIGCV2 recovered the driver as significant (q<0.25) is plotted above. Error bars reflect binomial 95% confidence intervals on the fraction of hypothetical drivers recovered.

### Heterogeneity within primary and metastasis biopsies reveals a seeding clone

We integrated allelic fraction data for each mutation with its local copy-number to determine the fraction of cancer cells carrying each mutation[22]. Different mutations tended to cluster around similar cancer cell fractions, indicating the presence of subclonal populations. Subclonal mutations were detected in all of the primary biopsies and 97% (32/33) of the MBs. A median of 20.7% and 21.3% of mutations in biopsies from metastases and primaries, respectively, were subclonal.

We focused on mutation clusters that were detected in more than one biopsy, but subclonal in at least one of them, as these may indicate seeding patterns from one biopsy to another (Figure 1.4B, Supplementary Figure 1.16). In one patient, this analysis identified a clone within the PB that seeded the MBs (Figure 1.4C, Supplementary Figure 1.17). We did not, however, find evidence of either oligoclonal seeding of metastases or re-seeding of either metastases or primaries. These results are consistent with a near-ubiquitous 'branched-sibling' relationship between primary tumors and paired metastases observed previously[24].


a
Unbiopsied subclone seeding the metastasis
Pattern 1
Metastasis
Primary
biopsied region; contains mutation set I and A
Rare subclone of biopsied primary seeds the metastasis
Pattern 1
Metastasis Cancer Cell Fraction
Primary Cancer Cell Fraction
b
Oligoclonal seeding
Pattern 2
Nothing
I and A
Primary subclone re-seeding met
Pattern 2
c
Primary subclone seeding metastasis
Pattern 3
Clonal expansion
I and A
Nothing
d
Metastatic subclone seeding the primary
Pattern 4
Nothing
I and b
biopsied region; contains mutation set I, b, b', c, c'

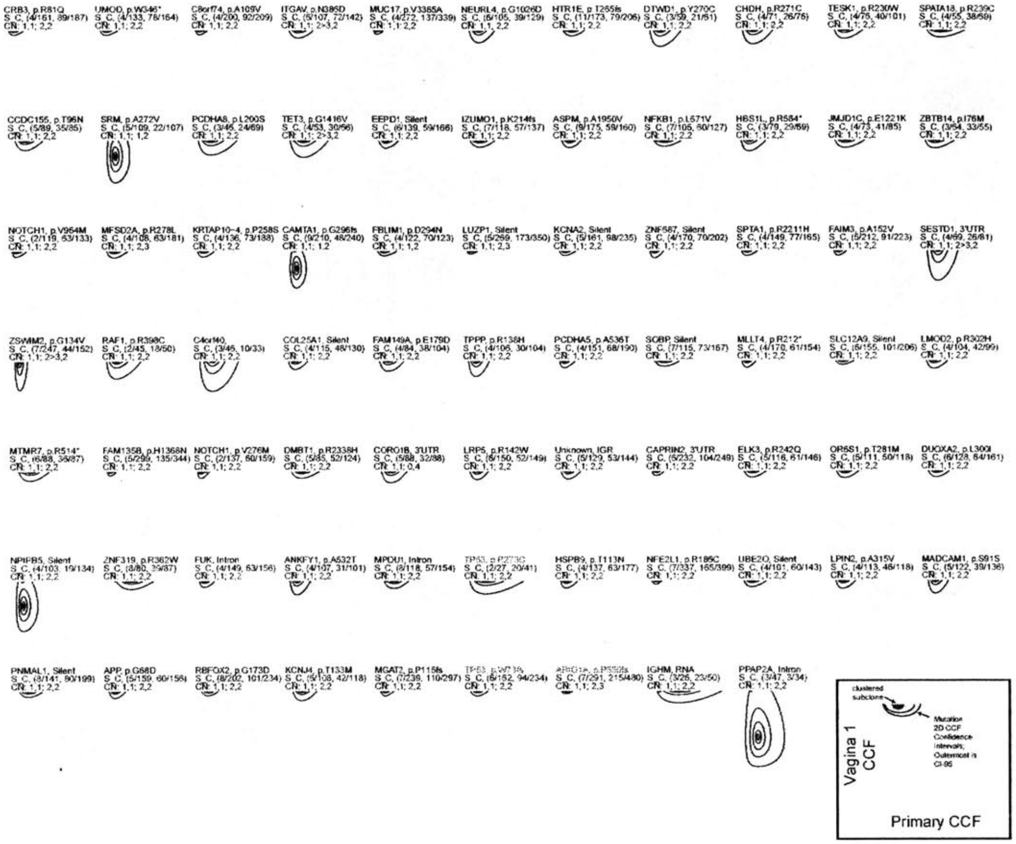
**Supplementary Figure 1.16. Possible phylogenetic patterns in analysis of 2D mutation clonality**

**a,** Where an unbiopsied subclone of the primary seeds a metastasis, no subclonal mutations in the primary can be observed in the metastasis. A similar pattern would be expected if the metastasis was seeded by a subclone of the biopsied part of the primary tumor that was below the limit of detection.

**b,** Oligoclonal seeding of the metastasis by genetically distinquishable cells within the primary would demonstrate a pattern in which subclonal mutations were shared between the primary and the metastasis. Similarly, if a distinct subclone present within the primary biopsy re-seeded the metastasis at a later time, subclonal mutations would be shared between both the biopsies from the primary and the metastasis.

**c,** If a subclone of the biopsied portion of the primary tumor seeded the metastasis, then the subclone of the primary would appear clonal in the metastasis (a', purple cluster). Importantly, no private clonal mutations may be contained in the primary biopsy, as these mutations would also be clonal in the metastasis.

**d,** If a subclone of the biopsied portion of the metastasis reseeds the biopsied portion of the primary and performs a clonal sweep, the pattern in d would occur.



**Supplementary Figure 1.17. Mutations assigned to primary subclone that seeded metastases in EC-004**

2D confidence intervals on the cancer cell fraction (CCF) of mutations assigned to the cluster of mutations that became clonal in the metastasis but was subclonal in the primary for patient EC-004. The CCF of the mutation in the primary biopsy is plotted on the x-axis and the CCF of the mutation in the metastasis biopsy (Vagina 1) is plotted on the y-axis. The outermost ellipse represents the 95% confidence interval. Candidate driver mutations are highlighted in red.

## Discussion

We present the first genome-wide analysis of genetic changes through endometrial cancer progression including hyperplasias, primary tumors, and paired metastases. We observed striking heterogeneity between biopsies of paired primary and metastatic tumors, with only half of mutations shared on average between any two biopsies. These biopsies did not fully sample these tumors, implying higher overall levels of heterogeneity than what we measured.

Across primary endometrial cancers, we identified 16 novel significantly mutated genes, owing in part to less stringent exclusion of indels in MSI tumors. Among these was NRIP1, which was mutated in 12.5% of tumors. NRIP1 is an obligate cofactor of the estrogen receptor[14], and germline SNPs near *NRIP1* have also been associated with ER-positive breast cancer in GWAS studies[38]. These data suggest that *NRIP1* alterations are common drivers of primary endometrial cancer oncogenesis. However, variations in indel rates across the genome are not well-described, and NRIP1 alterations were also seen in MSI colorectal cancers. Further characterization of the functional effects of these alterations is necessary.

The varying rates in heterogeneity across mutations in different genes indicate the order in which these mutations are acquired during tumor evolution. In particular, likely drivers of primary oncogenesis tend to be more homogenous than likely passengers. Among the drivers, mutations in *PIK3CA*, *PTEN*, *TP53*, and *PPP2R1A* occurred earlier on average in tumor evolution. In the case of *PIK3CA* and *PTEN*, these findings from advanced cancers mirror the findings in hyperplasias, which almost exclusively exhibited

PI3K pathway mutations, supporting hyperplasias as a genetically appropriate model of early stages of advanced endometrial cancer[20].

Conversely, mutations of the BAF chromatin remodeling complex subunit *ARID1A* were frequently heterogeneous. Recent studies have provided new therapeutic leads, such as EZH2 inhibition[39], to target *ARID1A*-deficient cancers. The heterogeneity of *ARID1A* mutations raises questions regarding the likely efficacy of such therapeutic interventions. However, it should be noted that convergent evolution involving *ARID1A* mutations was also observed, suggesting more homogenous pathway activation than indicated by looking at specific mutations in *ARID1A*. The finding of heterogeneity in mutations affecting the BAF complex are in accordance with several other cancer sequencing studies[28-31], but contrast with the role of BAF complex mutations in malignant pediatric rhabdoid tumors in which *SMARCB1* mutation is typically the sole driver of oncogenesis[40].

We also observed different levels of heterogeneity across cancers. In particular, we observed higher rates of heterogeneity between biopsies in tumors with more mutations in each biopsy, suggesting a single biopsy can indicate overall levels of heterogeneity across an entire tumor.

The finding that most of the biopsied metastases appeared to arise from the same lineage could indicate that acquiring metastatic potential is a rare event in endometrial cancers. It is possible, however, that it also reflects the timing of metastasis, whereby additional lineages metastasize later but do not have time to grow to macroscopic sizes by the time the first metastases are detected. Moreover, our analyses involved solely abdominopelvic metastases, many of which were nodal. It is possible that metastases to other sites arise from other lineages. Multi-metastasis sequencing across more sites of metastasis is necessary to comprehensively characterize phylogenetic patterns of metastasis in endometrial and other cancers. Multi-metastasis sequencing across larger numbers of patients may also reveal subpopulations of endometrial cancer in which metastases arise from a larger number of lineages.

Our data did not support the hypothesis that seeding originated from multiple clones, nor that re-seeding occurred, as has previously been detected in models of breast cancer[41] and in prostate tumors[23]. However, our data are insufficient to reject the hypothesis that these events occurred. Our biopsies of the primary tumor are unlikely to reflect the full diversity of the primary, and some of the diversity within biopsies might only be revealed with higher sequencing depth. Similarly, our data can neither refute nor confirm recent mouse studies demonstrating seeding of metastases by clumps of adjacent tumor cells[42]. These clumps would probably not appear as distinct entities in phylogenetic analysis, because of their spatial proximity and hence likely identical genetic makeup. In contrast to our finding of consistent branched evolution, a recent study of recurrent gliomas following surgery and temozolomide therapy[29], found a variety of evolutionary relationships, including branched, linear or "intermediate-mixed" phylogenies. However, the glioma data reflect recurrence after local resection rather than metastasis. The bottlenecks involved in recurrence may be less stringent, allowing persistence of multiple subclones that may repopulate the tumor.

The finding that metastases arise from a limited number of lineages within the primary tumor suggests that the ability to metastasize is not a common feature among primary endometrial cancer cells—further suggesting that metastasis requires additional events to those that are required for primary oncogenesis. However, we did not identify metastasis-specific drivers. It may be that the drivers of metastasis are intergenic, epigenetic, or environmental events that are not well-assessed by whole-exome sequencing[37]. It is also possible that there is a great diversity of genetic events that contribute to metastasis, each occurring in a small subset of metastatic cancers, and that we had insufficient power to detect such rare events. Genomic analysis of much larger numbers of paired primary and metastasis samples is therefore warranted.

## Methods

### Sample collection and description

The investigations within this study were approved by the Norwegian Social Science Data Services (15501), the local Institutional Review Board at Haukeland University Hospital, Bergen, Norway (REK-number) and the Broad Institute, Cambridge MA, USA. All patients consented to inclusion in this study. Samples were collected from Sept 2002-Sept 2012. Biopsies were snap-frozen in liquid nitrogen and stored at -80°C. Tumor purity was assessed by sections obtained by microtome prior to DNA extraction. Blood samples were collected for reference as normal controls.

### Assessment of microsatellite instability (MSI) status

MSI testing was performed on all samples submitted to whole exome sequencing using the marker set employed by TCGA[4]. DNA was whole-genome amplified using the GenomePlex Complete Whole Genome Amplification kit (Sigma Aldrich). The probe set consisted of BAT25, BAT26, BAT40, TGFBRII, D2S123, D5S346 and D17S250. No markers were positive in the normal controls (blood). None of the patients in this study were diagnosed with hereditary nonpolyposis colorectal cancer (HNPCC).

### Exome Sequencing and SNP Array profiling

Genomic DNA was isolated from frozen tissues using the Qiagen DNAamp kit or a standard proteinase K protocol. Samples were sequenced on an Illumina HiSeq-2000 to an average of 77x depth. Affymetrix SNP 6.0 arrays were used for a subset of samples. Three hyperplasias, out of 10 profiled, were found to have both an exceptionally low purity (less than 25% per ABSOLUTE analysis[22]) and low burden of mutations. Upon manual review, the mutations whose allelic fractions were higher than 10% were enriched in regions with low mapping quality. These three samples were therefore excluded from further analyses.

*ESR1* p.Y537H and *NRIP1* p.K728fs mutations were validated by Sanger sequencing using an Applied Biosystem 3730XL Analyzer, as previously described[43].

### Somatic mutation calling

Somatic mutations were called with MuTect[44]. OxoG artifacts were removed using the Broad Institute OxoG3 filter[45]. Insertions and deletions (indels) were called with Indelocator. Additional indels were rescued according to the following previously established criteria: at least 50 reads in both the tumor and normal, > 0.2 allelic fraction for the variant read in the tumor, and <0.05 allelic fraction for the variant read in the normal[9]. To ensure the fidelity of this approach, we swapped tumor and normal labels to determine the false positive rate of indel calls. A median of 0 and maximum of 2 indels were falsely called exome-wide in this approach (Supplementary Figure 1.5A). In contrast, we rescued more than 100 mutations in 13 tumor samples using this approach.

### Copy number analysis

Relative copy number profiles from Affymetrix SNP 6.0 arrays were determined as previously described[18]. Relative copy number profiles from exome sequencing data were determined by normalizing exome coverage data to values from blood controls and generates segmented copy-number profiles. These were paired with germline heterozygous sites to obtain allele-specific relative copy-number profiles, as previously described[24,46]. The relative allele-specific copy number profiles were paired with exome mutation data for each tumor sample as input to ABSOLUTE for final determination of discrete allele-specific copy number profiles. The sequence of events that led to each allelic copy number profile was inferred using a maximum parsimony approach[18].

### Mutation correlations analysis

The mutations detected in primary cancers from this cohort were combined with mutations detected in endometrial cancers profiled by TCGA[4] to detect correlations and anticorrelations between mutated genes, using a previously described approach that maintains the marginal counts of both the number of mutations within each sample and the number of events within each gene[32]. Ultramutated samples and rescued indels in MSI tumors were excluded from this analysis. P-values were calculated using 10,000 permutations of the observed data. The network of correlated interactions was plotted using Cytoscape where the negative-log of the q value for positive correlation is proportional to the spring constant of an edge between two nodes.

### Detection of subclones within biopsies

For each mutation, cancer cell fractions were calculated by ABSOLUTE. Information from local allelic copy number, biopsy purity and variant allele counts were integrated to calculate a posterior distribution over cancer cell fractions. PHYLOGIC[24] jointly clusters the distributions of cancer cell fractions across each biopsy. The number of predicted subclones is likely to be smaller than the real number due to imprecision of calculated posterior distributions on cancer cell fractions from 78x median sequencing depth. As such, our estimates of subclonal heterogeneity are conservative.

For the analysis of cancer cell fraction of mutations in TCGA data, we performed ABSOLUTE across all tumor samples in the TCGA endometrial dataset. ABSOLUTE computes a probability distribution of the cancer cell fraction of each mutation and includes a probability that each mutation is subclonal. To exclude the possibility that passenger mutations in hypermutated samples could confound our analysis, we excluded hypermutated samples (>1000 detected mutations) from this analysis.

We determined from this analysis that subclones were not shared across biopsies (with the sole exception of the seeding clone in EC-004), enabling us to draw biopsy-level phylogenetic trees.

### Phylogenetic tree reconstruction

To improve the mutation calls in each biopsy, we implemented a "force-calling" procedure. The union of all mutations observed in a given phylogeny was obtained. For each biopsy, raw sequencing reads were re-examined at all sites for evidence of the alterative allele originally called by the mutation caller. This procedure effectively rescues true mutations that failed to reach the threshold of Mutect in a given biopsy.

Phylogenetic trees were constructed using an implementation of clonal ordering. Force-called mutations were converted into a binary incidence matrix depending on their absence/presence in a set of paired biopsies. We calculated the power to detect each mutation in each biopsy based on local allelic copy number and purity. Where a mutation was not detected in one biopsy, but power to detect it was less than 0.95, the mutation was excluded from the incidence matrix and separately annotated. A distance matrix was computed from the final incidence matrix using the following distance metric:

$$d_{a,b} = \frac{1}{1 + \overrightarrow{m_a} \cdot \overrightarrow{m_b}}$$

where $\overrightarrow{m_a}$ corresponds to the binary vector of mutations in biopsy *a* and $\overrightarrow{m_b}$ is the vector describing biopsy *b*. Hierarchical clustering of this distance matrix was performed using the complete linkage method in R.

**Haplotype tracking across tumor biopsies**

Germline heterozygous sites were determined from exome sequencing of the normal blood control sample. The allelic fraction of these sites was determined at each site in the exome in all paired tumor samples (primaries and metastases). Purity estimates (*p*) from ABSOLUTE were used to generate purity-adjusted minor allelic fractions (*mAF*) at each site.

$$mAF_{corrected} = \frac{mAF - \frac{1}{2}(1 - p)}{p}$$

These allelic fractions were multiplied by the local total copy number ($CN_T$) by ABSOLUTE to graph a point estimate for each SNP of the major and minor tumor alleles. A point estimate for the minor allelic copy number (*mACN*) at each site was calculated as follows:

$$\widehat{mACN} = CN_T \times mAF_{corrected}$$

The major allele in the reference tumor for each site was defined as whichever allele count was greater (variant vs. reference). The expected major allele at each SNP was colored red in the resulting plot. In the test tumor, the same major and minor alleles estimated from the reference (primary) tumor are used and colored accordingly. Haplotype tracking was performed across every pair-wise comparison in the cohort. Resulting plots were manually reviewed for discordant tumor haplotype alterations.

Instances in which the haplotype undergoing copy loss/gain in the test tumor is opposite the haplotype undergoing loss/gain in the reference tumor indicate separate events in the genetic history of the tumor. Raw plots for selected chromosomes for case EC-018 are shown in Supplementary Figure 1.10.

### Significance analysis of primary endometrial cancers

We combined the force-called mutation lists (without indel rescue) from our primary tumors with the mutations from the TCGA[10]. We applied MUTSIG2CV on this list of mutations. Genes with q-values less than 0.1 were considered significantly mutated.

Separately, we combined the force-called mutation lists (with indel rescue) from our primary tumors with the mutations from TCGA that included indel rescue[9]. We considered any genes that were mutated in greater than 10% of samples and whose q-value was less than $10^{-5}$ as significant.

**Significance analysis of metastasis-associated drivers**

For each phylogeny in our dataset, we selected the set of mutations that were detected in every paired metastasis biopsy that were not detected in the biopsy of the primary tumor. We applied MUTSIG2CV to this set of mutations and considered any mutations whose q-value was less than 0.25 as significant.

**Power to detect metastasis-associated drivers**

We used an empirical approach to determine our power to detect mutations that conferred the ability to metastasize. We used the list of metastasis-specific mutations that we previously constructed as a pool into which we "spiked" hypothetical driver gene mutations at decreasing frequencies. We then assessed the rate at which these hypothetical driver genes were recovered as significant ($q<0.25$).

**Percentage of mutations in driver genes found in all biopsies**

To calculate the percentage of mutations in each driver gene that were truncal, we used force-calling mutation lists annotated with detection power from ABSOLUTE. We then determined whether the mutation was detected in all biopsies from the same patient. If the mutation was present in all biopsies, then the number of trunk mutations was incremented by one. If a mutation was not detected in a given biopsy, and power to detect a single alternate read of the mutation was greater than 0.8, the number of branch mutations was incremented by one. If there was not sufficient power to detect the mutation in one or more biopsies lacking the mutation, then the mutation was not counted towards the trunk or branch counts. To exclude that possibility that passenger mutations in driver genes could confound our analysis, two phylogenies with POLE exonuclease mutations and ultramutated genomes (15,095 and 30,601 mutations detected) were excluded from this analysis.

## Supplementary Note

Here we describe the methods we used to determine the most likely number of lineages in the primary tumor that contribute to metastases that are large enough to biopsy ("metastatic-potential lineages", or MPLs), and the likely fraction $F$ of the primary tumor occupied by each of these of these MPLs.

### Estimation of the fraction of the primary tumor capable of metastasis assuming a single lineage with metastatic potential

To simplify the explanation, we first consider $F$, and specifically in the case that the primary contains only one MPL. In that case, if $F$ is small, biopsies of the primary tumor (PBs) are unlikely to have sampled the MPL. Therefore, all metastasis biopsies (MBs) will be more closely related to each other than any is to the PB. Conversely, if $F$ is large, PBs may have sampled the MPL, and in some cases an MB may be more closely related to the PB than to the other MBs.

We use Bayes' theorem to calculate the posterior distribution on $F$ given the trees $T$ that we observe:

$$P(F|T) = \frac{P(T|F)P(F)}{P(T)} \qquad (1)$$

where $P(T|F)$ is the probability of observing $T$ given $F$.

We consider specifically the tree structure $T^m$, in which all MBs exhibit monophyly (i.e. they are more closely related to each other than to the PB). In that case, either the PB was within the MPL but nevertheless more distant from the MBs than the MBs are from each other, or the PB was not within the MPL. We write this as follows:

$$P(T^m|F) = P(M \cap D|F) + P(\neg M|F) \tag{2}$$

where *P(M)* is the probability that the PB falls within the MPL and *P(D)* is the probability that the PB is the most distant biopsy in the phylogeny. By definition,

$$P(M|F) = F \tag{3}$$

and therefore:

$$P(\neg M|F) = 1 - F \tag{4}$$

Furthermore,

$$P(M \cap D|F) = P(M|F) * P(D|M,F) = F * P(D|M,F) \tag{5}$$

To determine P(D|M,F), we consider all possible tree structures as equally likely if the PB falls within the MPL. Among these, D is only possible if all the MBs form an unbroken clade in the phylogenetic tree. Such a state requires that the tree structure contains an unbroken clade with a number of leaves equal to the number of MBs, and further requires that the PB happens to occupy the most distant leaf.

We call the state in which the tree structure contains an unbroken clade with a number of leaves equal to the number of MBs state U. We determined its probability $P(U)$ using the Yule-Harding-Kingman (YHK) process (for a more advanced discussion and proof, see Zhu et al[34]). In the case of three biopsies (two MBs and one PB), $P(U) = 1$. In the case of four biopsies (three MBs and one PB), $P(U) = 2/3$.

Assuming U applies, to achieve D the primary biopsy would also have to occupy the most distant leaf. Given a tree structure, the probability of a given labeled tree is calculated by considering all unique permutations of its leaves. With n biopsies overall, including the one PB and n-1 MBs, the probability $P(D|U)$ that the PB is the most distant leaf is thus $1/n$.

Therefore,

$$P(D) = \frac{P(U)}{n_{biopsies}} \tag{6}$$

For the case of 2 metastases and 1 primary biopsy, *P(D)*=1/3. For the case of 3 metastases and 1 primary biopsy, *P(D)* = 1/6 because the P(U) under YHK is 2/3 and P(D|U) is 1/4.

Combining the equations above, we obtain:

$$P(T_m|F) = \frac{F * P(U)}{n_{biopsies}} + (1 - F). \tag{7}$$

Necessarily,

$$P(\neg T_m|F) = 1 - P(T_m|F). \tag{8}$$

To calculate P(F|T), we use equations 1 and 6 and iteratively observe trees. After each observation we update the prior, P(F), to be the posterior *P*(*F*|*T*) given the trees before, as follows:

$$P\left(F\middle|\vec{T}_{0\ldots i}\right) = \frac{P(T_i|F)P(F|\vec{T}_{0\ldots i-1})}{P(T)}; P(F|T_o) = 1, \tag{9}$$

where $T_i$ represents the ith tree and $P\left(F\middle|\vec{T}_{0\ldots i}\right)$ represents all trees through $T_i$. Note that this process assumes the same $F$ and $n_{mpl}$ apply to all trees. We discuss below how variations between trees might affect our results.

**Joint estimation of F and the number of MPLs**

In the setting of more than one MPL, we can obtain monophyly either because all MBs derive from the same MPL (we call this state M1), or if the MBs derive from different MBLs (state M2) but the PB nevertheless occupies the most distant leaf on the tree. This implies:

$$P\left(T_m|n_{mpl}, F\right) = P(M_1)\times P(T^m\ |M_1, F) + P(M_2)\times P(T^m|M_2, F), \tag{10}$$

Where $n_{mpl}$ refers to the number of MPLs. Note that $P(M_1)$ and $P(M_2)$ do not depend on F.

$P(T^m \mid M_1, F)$ is given by equation 7 above, except that in this case F must be adjusted for the fact that it is divided between all MPLs. For simplicity, we assume this division is equal, so that:

$$P(T_m|M_1, F) = \frac{F * P(U)}{n_{mpl} \times n_{biopsies}} + (1 - \frac{F}{n_{mpl}}). \qquad (11)$$

Calculating $P(T^m \mid M_2, F)$, $P(M_1)$, and $P(M_2)$ depend critically on the number of MBs and PBs. We will describe the calculation for two MBs and one PB (indicated by subscripts 2m and 1p below). In this case,

$$P(M_1)_{2m,1p} = \frac{1}{n_{mpl}}, \qquad (12)$$

$$P(M_2)_{2m,1p} = 1 - P(M_1)_{2m,1p} = \frac{n_{mpl} - 1}{n_{mpl}}, \qquad (13)$$

and

$$P(T_m|M_2, F)_{2m,1p} = 1/3 \qquad (14)$$

because by definition, the distinct MPLs evolved independently and therefore share no increased relation to one another than a random biopsy of the primary tumor. Moreover, equation 11 reduces to:

$$P(T_m|M_1, F)_{2m,1p} = \frac{1}{3}\frac{F}{n_{mpl}} + \left(1 - \frac{F}{n_{mpl}}\right) = 1 - \frac{2}{3}\frac{F}{n_{mpl}}. \qquad (15)$$

Putting these together,

$$P(T_m|n_{mpl}, F)_{2m,1p} = \frac{1}{n_{mpl}}\left(1 - \frac{2}{3}\frac{F}{n_{mpl}}\right) + \frac{n_{mpl}-1}{n_{mpl}}\left(\frac{1}{3}\right). \qquad (16)$$

Considering multiple trees, we can calculate:

$$P(\vec{T}|F, n_{mpl}) = \prod_{i=1}^{n_{trees}} P(T_i|F, n_{mpl}). \qquad (17)$$

We then calculate the probability distribution over $F$ and $n_{mpl}$ given $\vec{T}$ using Bayes' rule:

$$P(F, n_{mpl}|\vec{T}) = \frac{P(\vec{T}|n_{mpl}, F)P(F, n_{mpl})}{P(\vec{T})}. \qquad (18)$$

We assume that the prior probability distributions over $F$ and $n_{mpl}$ are independent, so that:

$$P(F, n_{mpl}) = P(F) \times P(n_{mpl}). \qquad (19)$$

We assume a uniform prior over $F$ [i.e., $P(F) = 1$ for all values between 0 and 1]. In theory, $n_{mpl}$ can range from 1 to infinity. In this case, a uniform prior would be improper (non-normalizable). We therefore used Rissanen's universal prior on integers[47] as $P(n_{mpl})$. Rissanen's prior also accounts for the increased relative precision of posterior estimates on $n_{mpl}$ when $n_{mpl}$ is small.

**Results and effects of variations in $F$ and $n_{mpl}$ between trees**

Applied to the endometrial cancer dataset, these calculations indicate the typical primary cancer has only one MPL with a 92% probability, and that this MPL comprises a substantial fraction of the primary tumor with an expectation value of 30%.

The result that $n_{mpl}$ is most likely one is due to the finding that six of the seven endometrial cancer trees we evaluated exhibited monophyly, a result that is unlikely if two or more MPLs were present and generated similar numbers of metastases that could be biopsied. Even in the case of a single tree with two MBs and one PB, and assuming $F$ is near-zero (the assumptions that maximize estimates of $n_{mpl}$), the presence of two MPLs would tend to generate polyphyly in two-thirds of cases. The finding that six of seven trees exhibited monophyly, in some cases with more than two metastases, would be highly unlikely under such circumstances (p=0.007). This likelihood decreases if more MPLs are assumed (p=0.001 for infinite MPLs). For these reasons, it seems likely that the $n_{mpl}$ is one or near-one for most of these trees and, by generalization, the typical metastatic endometrial cancer.

The finding that $F$ has an expectation value of 30% rather than near-zero is due to the presence of one tree that did not exhibit monophyly for the MBs. In the models above and assuming $n_{mpl}$ is one in all cases, this could only occur if the PB sampled the MPL, which is only likely if $F$ is non-zero.

However, it is possible that the one tree with polyphyly reflected a tumor with more than one MPL, even if many of the other tumors had only one MPL. In such a case, F could be near-zero and polyphyly nevertheless be observed in that one tree. For these reasons, the conclusion that $F$ in not near-zero is less reliable than the conclusion than $n_{mpl}$ is one or near-one for most trees.

## References

1 Bhaskaran, K. *et al.* Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5.24 million UK adults. *Lancet* **384**, 755-765, doi:10.1016/S0140-6736(14)60892-8 (2014).

2 Ferlay J, S. I., Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F. *GLOBOCAN 2012 v1.0, Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet]*, <http://globocan.iarc.fr> (

3 Bokhman, J. V. Two pathogenetic types of endometrial carcinoma. *Gynecologic oncology* **15**, 10-17 (1983).

4 Cancer Genome Atlas Research, N. *et al.* Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67-73, doi:10.1038/nature12113 (2013).

5 Salvesen, H. B. *et al.* Integrated genomic profiling of endometrial carcinoma associates aggressive tumors with indicators of PI3 kinase activation. *Proceedings of the National Academy of Sciences of the United States of America* **106**, 4834-4839, doi:10.1073/pnas.0806514106 (2009).

6 Dutt, A. *et al.* Drug-sensitive FGFR2 mutations in endometrial carcinoma. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 8713-8717, doi:10.1073/pnas.0803379105 (2008).

7 Salvesen, H. B., Haldorsen, I. S. & Trovik, J. Markers for individualised therapy in endometrial carcinoma. *The Lancet. Oncology* **13**, e353-361, doi:10.1016/S1470-2045(12)70213-9 (2012).

8 Ciriello, G. *et al.* Emerging landscape of oncogenic signatures across human cancers. *Nature genetics* **45**, 1127-1133, doi:10.1038/ng.2762 (2013).

9 Giannakis, M. *et al.* RNF43 is frequently mutated in colorectal and endometrial cancers. *Nature genetics* **46**, 1264-1266, doi:10.1038/ng.3127 (2014).

10 Lawrence, M. S. *et al.* Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature* **505**, 495-501, doi:10.1038/nature12912 (2014).

11 Woodruff, J. D. & Pickar, J. H. Incidence of endometrial hyperplasia in postmenopausal women taking conjugated estrogens (Premarin) with medroxyprogesterone acetate or conjugated estrogens alone. The Menopause Study Group. *American journal of obstetrics and gynecology* **170**, 1213-1223 (1994).

12 Veeraraghavan, J. *et al.* Recurrent ESR1-CCDC170 rearrangements in an aggressive subset of oestrogen receptor-positive breast cancers. *Nature communications* **5**, 4577, doi:10.1038/ncomms5577 (2014).

13 Cavailles, V. *et al.* Nuclear factor RIP140 modulates transcriptional activation by the estrogen receptor. *The EMBO journal* **14**, 3741-3751 (1995).

14 Rosell, M. *et al.* Complex formation and function of estrogen receptor alpha in transcription requires RIP140. *Cancer research* **74**, 5469-5479, doi:10.1158/0008-5472.CAN-13-3429 (2014).

15 Toy, W. *et al.* ESR1 ligand-binding domain mutations in hormone-resistant breast cancer. *Nature genetics* **45**, 1439-1445, doi:10.1038/ng.2822 (2013).

16 Robinson, D. R. *et al.* Activating ESR1 mutations in hormone-resistant metastatic breast cancer. *Nature genetics* **45**, 1446-1451, doi:10.1038/ng.2823 (2013).

17 Croce, C. M. *et al.* Transcriptional activation of an unrearranged and untranslocated c-myc oncogene by translocation of a C lambda locus in Burkitt. *Proceedings of the National Academy of Sciences of the United States of America* **80**, 6922-6926 (1983).

18 Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature genetics* **45**, 1134-1140, doi:10.1038/ng.2760 (2013).

19 Evan, G. I., Lewis, G. K., Ramsay, G. & Bishop, J. M. Isolation of monoclonal antibodies specific for human c-myc proto-oncogene product. *Molecular and cellular biology* **5**, 3610-3616 (1985).

20 Hayes, M. P. *et al.* PIK3CA and PTEN mutations in uterine endometrioid carcinoma and complex atypical hyperplasia. *Clinical cancer research : an official journal of the American Association for Cancer Research* **12**, 5932-5935, doi:10.1158/1078-0432.CCR-06-1375 (2006).

21 Ganem, N. J., Godinho, S. A. & Pellman, D. A mechanism linking extra centrosomes to chromosomal instability. *Nature* **460**, 278-282, doi:10.1038/nature08136 (2009).

22 Carter, S. L. *et al.* Absolute quantification of somatic DNA alterations in human cancer. *Nature biotechnology* **30**, 413-421, doi:10.1038/nbt.2203 (2012).

23 Gundem, G. *et al.* The evolutionary history of lethal metastatic prostate cancer. *Nature*, doi:10.1038/nature14347 (2015).

24 Brastianos, P. *et al.* Genomic characterization of brain metastases reveals branched evolution and potential
therapeutic targets. *Cancer Discovery* (2015).

25 Landau, Dan A. *et al.* Evolution and Impact of Subclonal Mutations in Chronic Lymphocytic Leukemia. *Cell* **152**, 714-726, doi:10.1016/j.cell.2013.01.019 (2013).

26 Lohr, J. G. *et al.* Widespread genetic heterogeneity in multiple myeloma: implications for targeted therapy. *Cancer cell* **25**, 91-101, doi:10.1016/j.ccr.2013.12.015 (2014).

27 Kadoch, C. *et al.* Proteomic and bioinformatic analysis of mammalian SWI/SNF complexes identifies extensive roles in human malignancy. *Nature genetics* **45**, 592-601, doi:10.1038/ng.2628 (2013).

28 Gerlinger, M. *et al.* Genomic architecture and evolution of clear cell renal cell carcinomas defined by multiregion sequencing. *Nature genetics* **46**, 225-233, doi:10.1038/ng.2891 (2014).

29 Johnson, B. E. *et al.* Mutational analysis reveals the origin and therapy-driven evolution of recurrent glioma. *Science* **343**, 189-193, doi:10.1126/science.1239947 (2014).

30 Abedalthagafi, M. S. *et al.* ARID1A and TERT promoter mutations in dedifferentiated meningioma. *Cancer Genet*, doi:10.1016/j.cancergen.2015.03.005 (2015).

31 Torres-Martin, M. *et al.* Whole exome sequencing in a case of sporadic multiple meningioma reveals shared NF2, FAM109B, and TPRXL mutations, together with unique SMARCB1 alterations in a subset of tumor nodules. *Cancer Genet*, doi:10.1016/j.cancergen.2015.03.012 (2015).

32 Ding, L. *et al.* Somatic mutations affect key pathways in lung adenocarcinoma. *Nature* **455**, 1069-1075, doi:10.1038/nature07423 (2008).

33 Van Allen, E. M. *et al.* Whole-exome sequencing and clinical interpretation of formalin-fixed, paraffin-embedded tumor samples to guide precision cancer medicine. *Nature medicine* **20**, 682-688, doi:10.1038/nm.3559 (2014).

34 Zhu, S., Degnan, J. H. & Steel, M. Clades, clans, and reciprocal monophyly under neutral evolutionary models. *Theoretical population biology* **79**, 220-227, doi:10.1016/j.tpb.2011.03.002 (2011).

35 Campbell, P. J. *et al.* The patterns and dynamics of genomic instability in metastatic pancreatic cancer. *Nature* **467**, 1109-1113, doi:10.1038/nature09460 (2010).

36 Yachida, S. *et al.* Distant metastasis occurs late during the genetic evolution of pancreatic cancer. *Nature* **467**, 1114-1117, doi:10.1038/nature09515 (2010).

37 Valastyan, S. & Weinberg, R. A. Tumor metastasis: molecular insights and evolving paradigms. *Cell* **147**, 275-292, doi:10.1016/j.cell.2011.09.024 (2011).

38 Ghoussaini, M. *et al.* Genome-wide association analysis identifies three new breast cancer susceptibility loci. *Nature genetics* **44**, 312-318, doi:10.1038/ng.1049 (2012).

39 Bitler, B. G. *et al.* Synthetic lethality by targeting EZH2 methyltransferase activity in ARID1A-mutated cancers. *Nature medicine* **21**, 231-238, doi:10.1038/nm.3799 (2015).

40 Lee, R. S. *et al.* A remarkably simple genome underlies highly malignant pediatric rhabdoid cancers. *The Journal of clinical investigation* **122**, 2983-2988, doi:10.1172/JCI64400 (2012).

41 Kim, M. Y. *et al.* Tumor self-seeding by circulating cancer cells. *Cell* **139**, 1315-1326, doi:10.1016/j.cell.2009.11.025 (2009).

42 Aceto, N. *et al.* Circulating tumor cell clusters are oligoclonal precursors of breast cancer metastasis. *Cell* **158**, 1110-1122, doi:10.1016/j.cell.2014.07.013 (2014).

43 Pardanani, A. *et al.* IDH1 and IDH2 mutation analysis in chronic- and blast-phase myeloproliferative neoplasms. *Leukemia* **24**, 1146-1151, doi:10.1038/leu.2010.77 (2010).

44 Cibulskis, K. *et al.* Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nature biotechnology* **31**, 213-219, doi:10.1038/nbt.2514 (2013).

45 Costello, M. *et al.* Discovery and characterization of artifactual mutations in deep coverage targeted capture sequencing data due to oxidative DNA damage during sample preparation. *Nucleic acids research* **41**, e67, doi:10.1093/nar/gks1443 (2013).

46 Stachler, M. *et al.* Paired Exome Analysis of Barrett's Esophagus and Adenocarcinoma. *Nature genetics* (2015).

47 Rissanen, J. A universal prior for integers and estimation by minimum description length. *The Annals of statistics*, 416-431 (1983).

# Chapter 3: Genome-wide copy number synthetic lethal analysis identifies partial copy loss of SF3B1 as a novel cancer vulnerability

## Abstract

One of the hallmarks of cancer is genomic instability, resulting in widespread somatic copy number alterations. We use genome-scale shRNA screening data to perform an unbiased analysis of all copy-number:gene-dependency interactions. We find an enriched class of interactions in which hemizygous loss of essential genes sensitizes cells to their further suppression. We validate one of these interactions for the splicing factor SF3B1. Cancer cells harboring loss of SF3B1 lack a reservoir of SF3b complex that protects cells whose SF3B1 locus is intact in the face of SF3B1 suppression. These data provide evidence that copy-number synthetic lethal approaches may serve as a useful means for developing novel cancer therapeutics.

## Introduction

During the course of tumorigenesis, the majority of cancers undergo genomic structural alterations affecting the number of copies of each gene [1]. Such somatic copy number alterations (SCNAs) result in either gains or losses of genetic material affecting many regions across a cancer's genome. Genes located within SCNAs may promote the oncogenic phenotype and are thought to be cancer "driver events", which undergo positive selection due to their effects on oncogenes or tumor suppressor genes that drive tumor development.

Frequently, SCNAs encompass broad chromosomal regions and affect hundreds to thousands of neighboring genes, including essential genes, that do not affect the oncogenic phenotype are considered to be “passenger events”. Much effort has been placed on the development of therapeutics that target cancer driver events including those arising from SCNAs, such as amplifications of *MYC* [2], and *CCND1* [3]. However, a systematic and genome-wide evaluation of the landscape of copy-number associated cancer vulnerabilities has not been conducted to date. Besides driver gene-associated SCNA vulnerabilities, targeting non-driver genes in cancer represents a potential new approach for cancer treatment that could be extended to a large number of novel therapeutic targets.

Non-driver dependencies can arise in cancer as a result of structural alterations during tumorigenesis. Compared to normal cells, cancer cells can rely on SCNAs to drive altered signaling pathways that maintain their oncogenic state. For example, inactivation of a tumor suppressor gene by deletion can initiate tumorigenesis but often coincides with loss of an entire chromosome arm, or chromosome. These broad events inadvertently delete hundreds to thousands of presumably normal genes [4], some of which are essential to cell survival. It is possible that structural alterations create an opportunity to selectively kill tumor cells that harbor hemizygously deleted essential genes while normal diploid cells can tolerate partial gene suppression (Figure 2.1A).

We surveyed the landscape of copy-number associated vulnerabilities by integrating copy-number data from cancer cell lines with a genome-wide shRNA screen for cell viability. We determined which genes, when altered by SCNAs, result in cancer dependencies when the same or different gene is suppressed. We find that many of the SCNA-associated dependencies result from copy-loss of a gene sensitizing cells to further suppression of that same gene. We previously named these CYCLOPS (*c*opy number alterations *y*ielding *c*ancer *l*iabilities *o*wing to *p*artial los*s*) genes. Amongst the CYCLOPS gene list, one of the most significantly enriched pathways identified was the spliceosome.

The spliceosome, which removes introns and ligates exons from nascent pre-mRNAs [5], represents a novel target for anti-cancer therapies [6-8]. While spliceosome function is essential for cell survival, recent evidence suggests that mutation of spliceosome components may be driver events in many cancer types [9-12]. Small molecule inhibitors of mRNA splicing have shown the ability to potently kill cancer cells [13]. The existence of spliceosome inhibitors raises the possibility of a "therapeutic window" for pharmacologic targeting of cancer cells either harboring loss of spliceosome-associated CYCLOPS genes or dependent on a splicing factor driver mutation.

Here we identify common genomic features of CYCLOPS genes and characterize the U2 snRNP component, *SF3B1*, as a CYCLOPS gene. We find that hemizygous loss of CYCLOPS genes are common in cancer genomes and represent a subset of cell-essential genes with genomic features unique from other essential genes. We demonstrate that cells with partial copy loss of *SF3B1* are highly vulnerable to further *SF3B1* suppression. The identification of *SF3B1* as a CYCLOPS gene highlights a previously unrecognized cancer vulnerability and suggests that inhibition of WT SF3B1 can be a potential therapeutic strategy for treatment of *SF3B1*$^{loss}$ tumors.

## Results

### Analysis of Genome-Wide Copy-Number Induced Cancer Dependencies

We interrogated copy-number associated vulnerabilities genome-wide across 179 cell lines by integrating dependency data from Project Achilles [14] with copy-number profiles generated across 1.8 million loci [15] (Figure 2.1A). The dependency data represented the effects on proliferation of 55,416 shRNAs against 11,589 genes, measured in each line as z-scored fold-changes in cells carrying shRNAs against each targeted gene [16]. For every pair of genes, we calculated Pearson correlations between the copy-number of the first gene and the dependency score of the second. We excluded correlations involving dependency scores of genes for which we did not observe consistent results between at least two shRNAs, resulting in 6,192 gene dependency scores and 156,076,513 pairings in total. We calculated p-values for each correlation and false discovery rate q-values to correct for multiple hypotheses.

We identified 2,375 significant copy-number:gene-dependency interactions with q<0.1 (Table 1). In 2,309 cases, these represented redundant data resulting from associations between a gene dependency profile and identical copy-number profiles from neighboring genes. After excluding these, we identified 66 independent significant copy-number:gene-dependency interactions.

Approximately two-thirds (69.7%) of these significant interactions involved genes on separate chromosomes (*trans* interactions). *Trans* interactions had both positive (23/46) and negative Pearson correlations (23/46), indicating instances where cells were sensitized to suppression of some gene by either copy-loss or gain respectively from a different genomic region. In contrast, all but one interaction (19/20) between genes on the same chromosome (*cis* interactions) had positive Pearson correlations and indeed involved copy-loss of a gene being associated with sensitivity to cells to further suppression of that same gene. As a result, 63.6% of all 66 significant copy-number:gene-dependency interactions were associated with copy-loss rather than gain (p=0.009).

## Enhanced Identification of CYCLOPS Genes

The genes involved in cis interactions whereby copy-loss implies further dependency on that same gene have previously been termed CYCLOPS (*c*opy number alterations *y*ielding *c*ancer *l*iabilities *o*wing to *p*artial los*s)* genes. Although they represent less than 0.003% of all potential interactions, they constitute nearly one-third (28.8%) of all significant interactions.

The prevalence of CYCLOPS vulnerabilities is partly the result of frequent genomic loss in cancer genomes. We evaluated the fraction of the genome that undergoes relative copy-loss across 10,570 cancers spanning 31 cancer types profiled by The Cancer Genome Atlas, and found that 18.9% of the genome undergoes loss in the average cancer (Figure S2.1A), mainly due to losses encompassing chromosome arms or entire chromosome (Figure S2.1B) [17]. Indeed, loss of a tumor suppressor often involves such arm-level losses (Figure S2.1C). The fraction of the genome lost ranged from an average of 1.3% in thyroid cancer to 34.4% in ovarian cancer (Figure S2.1D).

To enhance the identification of CYCLOPS genes, we specifically compared gene dependencies for each gene between cell lines with and without loss of that gene. We permuted class labels to calculate p-values and False Discovery Rate q-values, and considered $q<0.1$ to be significant. This analysis identified 124 candidate CYCLOPS genes (Table S2).

Gene set enrichment analysis revealed candidate CYCLOPS genes were most highly enriched for members of the spliceosome and the proteasome (Figure S2.1E), consistent with previous analyses [4]. Of the 124 CYCLOPS genes, 11 are members of the proteasome and 20 are members of the spliceosome. These genes were distributed across all autosomes (Figure S2.1F) and were biased towards areas of frequent copy-loss ($p = 2.2x10^{-16}$), perhaps resulting from greater power to detect CYCLOPS genes in these regions.

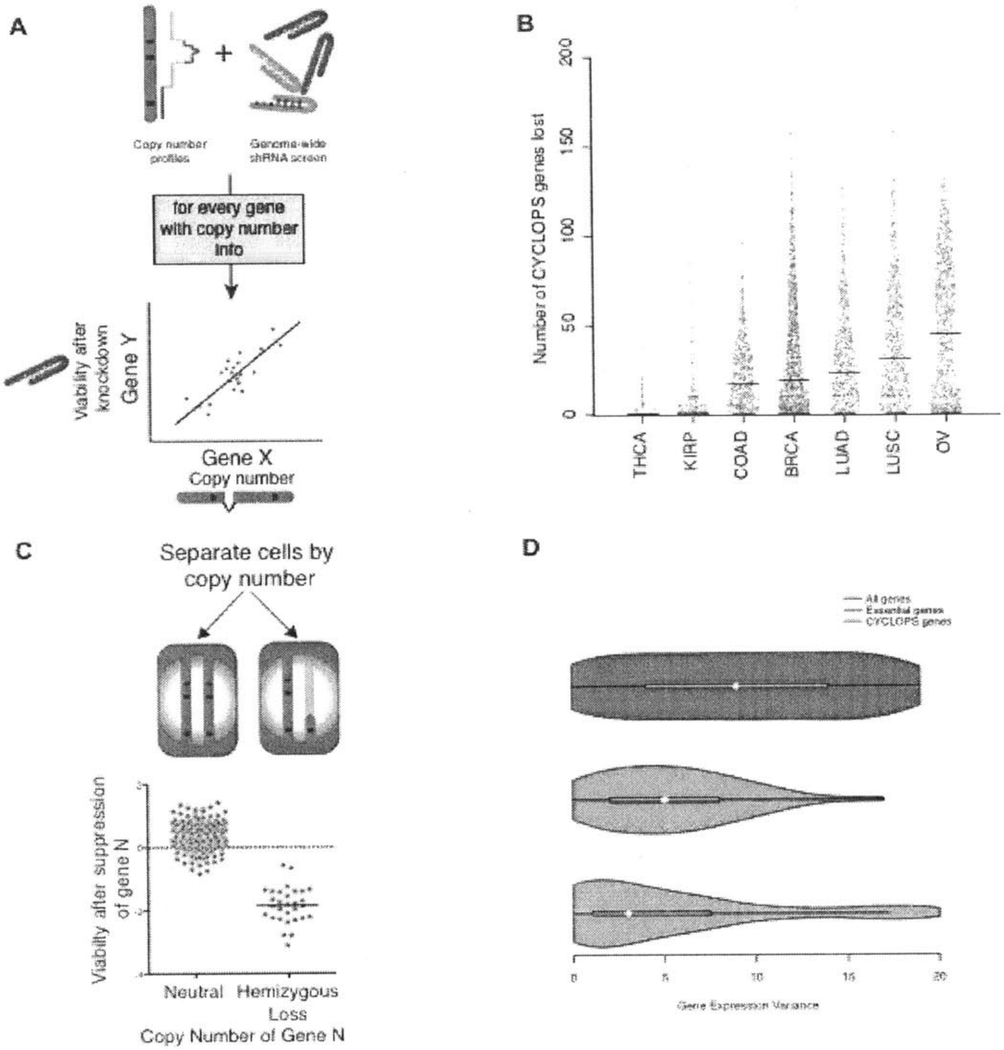


**Figure 2.1: Analysis of synthetic lethal interactions with copy-number. (A)** Schematic describing the approach to identify copy-number induced cancer dependencies. **(B)** The number of CYCLOPS genes lost per tumor in the TCGA for indicated tumor types. Horizontal black lines represent mean number of genes per tumor type. **(C)** Schematic describing the approach to identify CYCLOPS genes. **(D)** Variance of gene expression between all genes, essential genes and CYCLOPS genes.

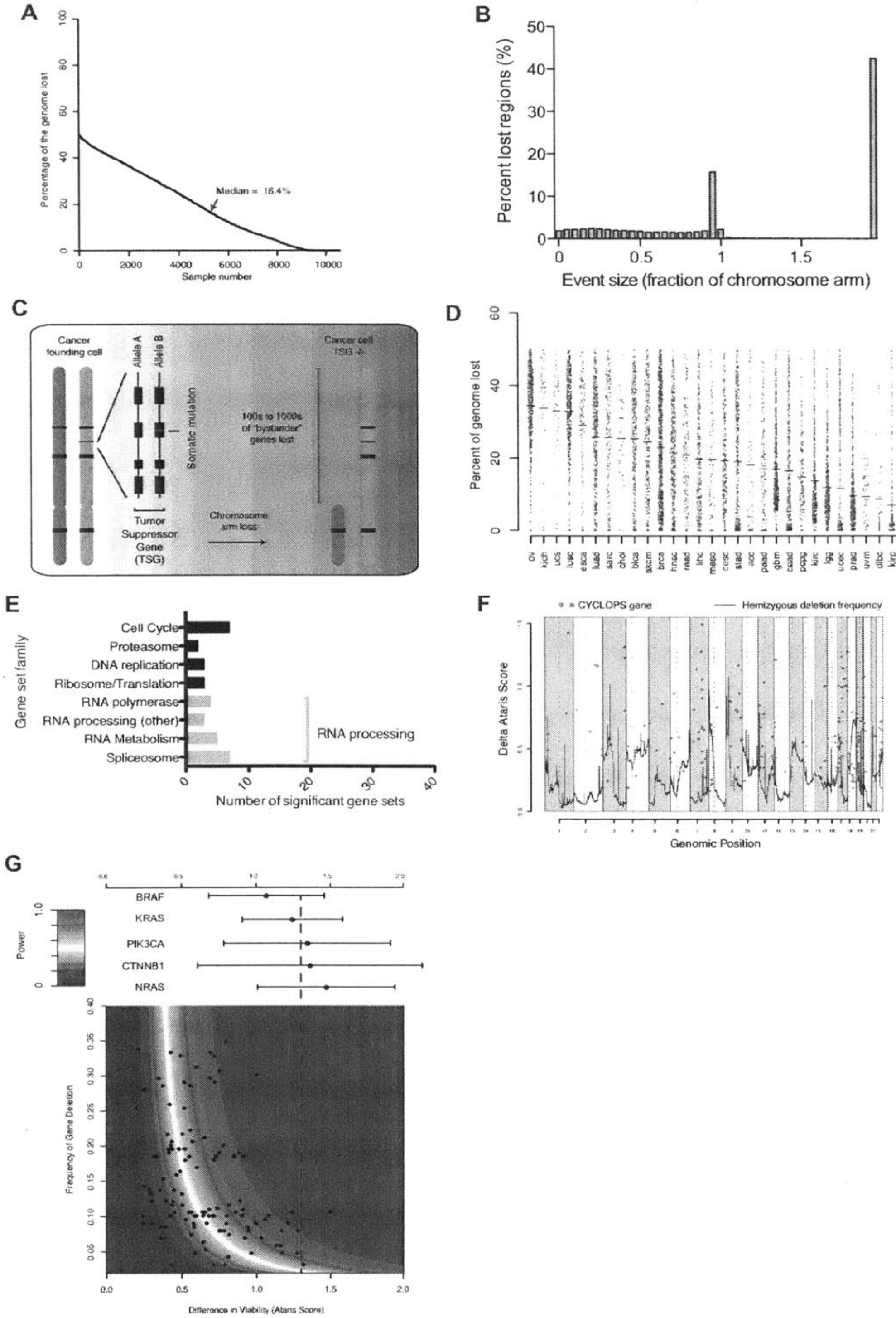

A
Percentage of the genome lost
Median = 16.4%
Sample number
0
20
40
60
80
100
0
2000
4000
6000
8000
10000
B
Percent lost regions (%)
Event size (fraction of chromosome arm)
0
10
20
30
40
50
0
0.5
1
1.5
C
Cancer founding cell
Allele A
Allele B
Tumor Suppressor Gene (TSG)
Somatic mutation
100s to 1000s of "bystander" genes lost
Chromosome arm loss
Cancer cell TSG -/-
D
Percent of genome lost
0
20
40
60
ov
kich
ucs
lusc
esca
luad
sarc
chol
blca
skcm
brca
hnsc
read
lihc
meso
cesc
stad
acc
paad
gbm
coad
pcpg
kirc
lgg
ucec
prad
uvm
dlbc
kirp
laml
thca
E
Gene set family
Cell Cycle
Proteasome
DNA replication
Ribosome/Translation
RNA polymerase
RNA processing (other)
RNA Metabolism
Spliceosome
RNA processing
Number of significant gene sets
0
10
20
30
40
F
CYCLOPS gene
Hemizygous deletion frequency
Delta Ataris Score
Genomic Position
G
Power
0
1.0
BRAF
KRAS
PIK3CA
CTNNB1
NRAS
0.0
0.5
1.0
1.5
2.0
Frequency of Gene Deletion
0.05
0.10
0.15
0.20
0.25
0.30
0.35
0.40
Difference in Viability (Ataris Score)

**Figure S2.1, related to Figure 2.1:** **(A)** Fraction of the genome undergoing copy loss in 10,570 tumors analyzed in the TCGA. **(B)** The fraction of deleted regions resulting from deletion events of the indicated size. Most deleted regions undergo loss as a result of losses involving whole chromosomes. **(C)** Schematic depicting the hemizygous loss of passenger essential genes during biallelic inactivation of a tumor suppressor gene. **(D)** The percent of the genome that is hemizygously lost in the indicated tumor types **(E)** Summary of significantly enriched gene sets among CYCLOPS genes **(F)** Distribution of CYCLOPS genes throughout the genome **(G)** (top) The average difference in dependency scores of cell lines harboring mutations in the indicated genes when said gene is suppressed. For CTNNB1, APC mutations were considered in the two-class comparison. (bottom)

### Strength and Prevalence of CYCLOPS vulnerabilities

To determine the relative strength of CYCLOPS dependencies in the Achilles dataset, we compared CYCLOPS effects to that of the average strength of known oncogene-induced dependencies. We calculated the strength of oncogene dependencies in the Achilles dataset after suppression of the known driver gene between cell lines with the driver gene alteration to those without the driver gene alteration (genes used were BRAF, KRAS, NRAS, CTNNB1 or PIK3CA mutant cell lines). At the typical rate of genomic loss (18.9%), the CYCLOPS analysis obtained 99.5% power to detect cancer vulnerabilities with 50% of the strength of the average oncogene dependency (Figure S2.1G). In total we find 3 CYCLOPS genes with dependency scores greater than or equal to the average mutated oncogene induced dependency.

Partial copy-loss of CYCLOPS genes are frequent events in cancer genomes. Among the 7,232 TCGA cancers with ABSOLUTE data, 71.6% of tumors harbored loss of at least one CYCLOPS gene (Figure 2.1B). The average number of CYCLOPS genes lost per tumor ranged from 1 in thyroid cancer to 47 in ovarian cancer. These data suggest that a large fraction of patients may benefit from therapies that target CYCLOPS vulnerabilities. Taken together these data suggest that CYCLOPS vulnerabilities are common and often lead to oncogene-equivalent vulnerabilities.

### Genomic Features Associated with CYCLOPS Genes

CYCLOPS genes displayed significantly less variability in gene expression values across tissues. We surveyed RNA-sequencing data from 2,342 samples across 42 tissues in the GTEX database to quantify the relative variance in expression of CYCLOPS genes.

For every gene in the genome, we calculated their expression variance across all samples and ranked their variance among the 20 genes with the most similar expression level. The expression of both essential genes and CYCLOPS genes varied significantly less than the average gene analyzed ($p=3.4x10^{-12}$ and $p=1.8x10^{-15}$ respectively), but CYCLOPS gene expression varied even less than essential genes (Figure 2.1D, p=0.07). These data suggest that CYCLOPS genes are consistently expressed across tissues, consistent with their role as a subset of cell-essential genes.

### *SF3B1* is a CYCLOPS gene

*SF3B1* was among the most significant candidate genes in our CYCLOPS analysis (Table 1). *SF3B1* is an mRNA splicing factor that directs the U2 snRNP to intronic branch-point sequences to determine 3' splice-site selection[18]. Cells with *SF3B1* copy-loss exhibited significantly reduced viability to *SF3B1* suppression while cells without *SF3B1* copy-loss did not from our CYCLOPS analysis (mean dependency scores of -1.14 and 0.01 respectively, $p<1x10^{-5}$).

*SF3B1* is partially lost in 10.8% of the 10,570 cancers from the TCGA PanCan dataset (see Methods for definitions of copy number states). Losses were most frequent in invasive breast adenocarcinoma (20.2%), urothelial bladder carcinoma (31.8%) and chomophobe kidney carcinoma (71.2%). Genomic deletions of *SF3B1* are rarely focal events (2.1% of cancers) and are never homozygous (0/10,570 cancers), consistent with characterization of *SF3B1* as an essential gene[19,20]. Similarly, analysis of copy number alterations from 1042 cancer cell lines in CCLE indicated 24.1% of cell lines harbor hemizygous *SF3B1* deletion, including 16/61 (26.2%) of breast cancer cell lines, but never homozygous loss (0/1042 cell lines).

We confirmed the vulnerability of *SF3B1*$^{loss}$ cells to *SF3B1* suppression in both breast and hematopoietic lineages. We used at least two hairpins against *SF3B1* targeting separate regions of the gene and found they generated similar levels of *SF3B1* knockdown across ten breast cell lines (Figure S2.2A). We then tested the growth of six of these lines using CellTiterGlo, including three lines with *SF3B1* copy-loss (*SF3B1*$^{loss}$) and three without either loss or gain of the gene (*SF3B1*$^{neutral}$). Upon *SF3B1* suppression, the *SF3B1*$^{loss}$ cells exhibited significant growth defects but the *SF3B1*$^{neutral}$ cells did not (Figure 2.2A).

We generated isogenic *SF3B1*$^{loss}$ cells from the *SF3B1*$^{neutral}$ cell line Cal51 using two independent methods by CRISPR (see Methods). The first cell line contained a frame shift mutation inactivating one SF3B1 allele (CRISPR$^{\text{frameshift-loss}}$). In a second cell line, we generated a deletion of one copy of the *SF3B1* locus by co-expressing two sgRNAs, one upstream targeting a heterozygous SNP, and one downstream of *SF3B1* (CRISPR$^{\text{copy-loss}}$). In both cases (collectively called CRISPR$^{\text{loss}}$), CRISPR-mediated SF3B1 loss resulted in decreased growth upon *SF3B1* suppression relative to cells that were generated in parallel but did not produce inactivating alleles (CRISPR$^{\text{neutral}}$ cells; Figure 2.2A and S2.2B).

We confirmed the vulnerability of the *SF3B1*$^{loss}$ cells to *SF3B1* suppression using a GFP-competition assay in which we compared the proliferation rate of uninfected cells co-cultured with cells infected with a vector that co-expressed GFP and an shRNA targeting either *LacZ* or *SF3B1*. The expression of *LacZ* or *SF3B1* shRNAs did not result in significant changes in proliferation of *SF3B1*$^{neutral}$ cells in seven cell lines, including the non-transformed mammary cell line, MCF10A (Figure 22.B and S2.2C). However, *SF3B1*$^{loss}$ cells expressing *SF3B1* shRNAs we not compatible with long term culture (Figure 2.2B and S2.2C) and are consistent with our previous growth assays (Figure 2.2A).

The *SF3B1* CYCLOPS vulnerability is not recapitulated by suppression of other SF3B complex members. We calculated Pearson correlation coefficients between Achilles RNAi sensitivity data of each SF3B complex member in *SF3B1*$^{loss}$ cells. Besides SF3B1 suppression, we did not find strong correlations between copy-loss of *SF3B1* and sensitivity to suppression of a different SF3b member, suggesting that CYCLOPS effects are specific to each gene hemizygously lost and suppression of that same gene (Figure S2.2D)

Likewise, although multiple SF3b complex members are candidate CYCLOPS genes (Table 1), copy-loss of these genes does not confer susceptibility to SF3B1 suppression (Figure S2.2D). Pearson correlation coefficients were calculated using Achilles RNAi sensitivity data for each SF3b complex component and $\log_2$ copy number ratios for another SF3b component. No association reached statistical significance.

Suppression of *SF3B1* leads to both cell cycle arrest and apoptosis in *SF3B1*$^{loss}$ but not *SF3B1*$^{neutral}$ lines. We generated cultures containing a tetracycline inducible system expressing hairpins targeting *Luciferase* or *SF3B1* (TR-shSF3B1#3 and an additional hairpin, TR-shSF3B1#5, Figure S2.2E), enabling us to discriminate *SF3B1* suppression from infection with shRNA vectors. Consistent with stable *SF3B1* suppression, inducible *SF3B1* suppression retards *SF3B1*$^{loss}$ cell growth and does not affect *SF3B1*$^{neutral}$ growth (Figure S2.2F) and reduces cell viability in *SF3B1*$^{loss}$ cells but not in *SF3B1*$^{neutral}$ cells (Figures 2.2C and 2.2D). *SF3B1*$^{loss}$ cells had significantly increased proportions of cells in G2/M phase after *SF3B1* suppression, which did not occur in *SF3B1*$^{neutral}$ cells (Figure 2.2E). Subsequent to G2/M arrest, *SF3B1*$^{loss}$ cells further exhibited a significant induction in apoptosis as determined by increased number of AnnexinV/PI-positive cells that were not observed in *SF3B1*$^{neutral}$ cells (Figure 2.2F).

Expression of exogenous *SF3B1* rescued the loss of cell viability in *SF3B1*$^{loss}$ cells, confirming the specificity of our shRNAs. We used a lentiviral construct expressing a codon-optimized *SF3B1* ORF, which is resistant to shRNA suppression, fused to an IRES GFP sequence (*SF3B1*$^{WT}$-IRES-GFP). When placed in competition, cells infected and not infected with *SF3B1*$^{WT}$-IRES-GFP maintained constant ratios over 10 days (Figure 2G), suggesting that short-term expression of *SF3B1* does not alter cellular fitness in either *SF3B1*$^{neutral}$ or *SF3B1*$^{loss}$ cells. Next, we concomitantly suppressed endogenous *SF3B1* in all cells and expressed *SF3B1*$^{WT}$-IRES-GFP in ~50% of cells. While *SF3B1*$^{neutral}$ cells were not affected by *SF3B1* suppression, *SF3B1*$^{loss}$ cells expressing *shSF3B1* were not compatible with long-term culture. Importantly, *SF3B1*$^{loss}$ cells expressing both *shSF3B1* and *SF3B1*$^{WT}$-IRES-GFP persisted in culture (Figure 2.2H), indicating that re-expression of *SF3B1* is sufficient to prevent cell death mediated by shRNA suppression of *SF3B1*. Furthermore, *SF3B1*$^{loss}$ cells expressing both *shSF3B1* and *SF3B1*$^{WT}$-IRES-GFP had a 20 fold increased in GFP fluorescence, suggesting that the exogenous *SF3B1* construct was more highly expressed in *SF3B1*$^{loss}$ cells after suppression of endogenous *SF3B1* (Figure S2.2G). Furthermore, stable exogenous *SF3B1* expression is sufficient to restore the proliferation of *SF3B1*$^{loss}$ cells expressing shRNAs targeting *SF3B1* (Figure 2.2I and Figure S2.2H). Taken together, data summarized in Figure 2.2 support the characterization of *SF3B1* as a CYCLOPS gene.

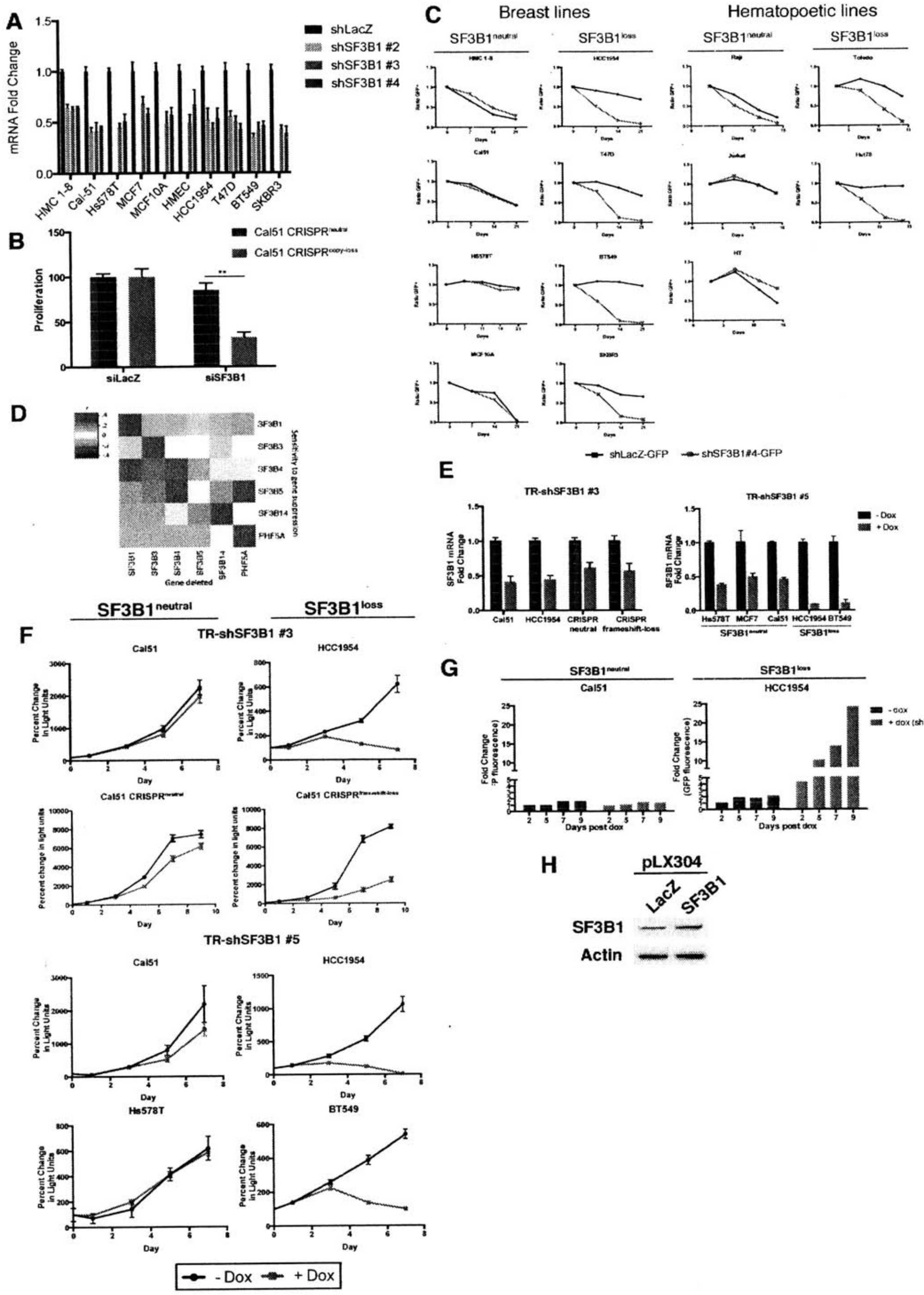


**Figure S2.2, related to Figure 2.2: (A)** Quantitative RT-PCR of *SF3B1* expression from the indicated cell lines expressing shLacZ or shSF3B1 shRNAs. **(B)** Proliferation of

SF3B1 copy-loss cells induced by CRISPR after treatment with siRNAs targeting LacZ or SF3B1. **(C)** The effect of *SF3B1* suppression on 6 breast and 5 hematopoietic cell lines expressing shLacZ-GFP (black) or shSF3B1#4-GFP (grey). **(D)** Heatmap of Pearson correlation coefficients indicating the relationship between copy loss of SF3B complex members (columns) and sensitivity of those cells to suppression of SF3B complex members by shRNA (rows). **(E)** Quantitative RT-PCR of *SF3B1* expression in cells expressing doxycycline-inducible *SF3B1* shRNAs. **(F)** Cell Titer Glo growth assays in cells expressing doxycycline-inducible *SF3B1* shRNAs. **(G)** GFP fluorescence quantification from cells expressing SF3B1-IRES-GFP constructs without and with suppression of endogenous *SF3B1* by doxycycline. **(H)** SF3B1 immunoblot from HCC1954 cells expressing LacZ or SF3B1.

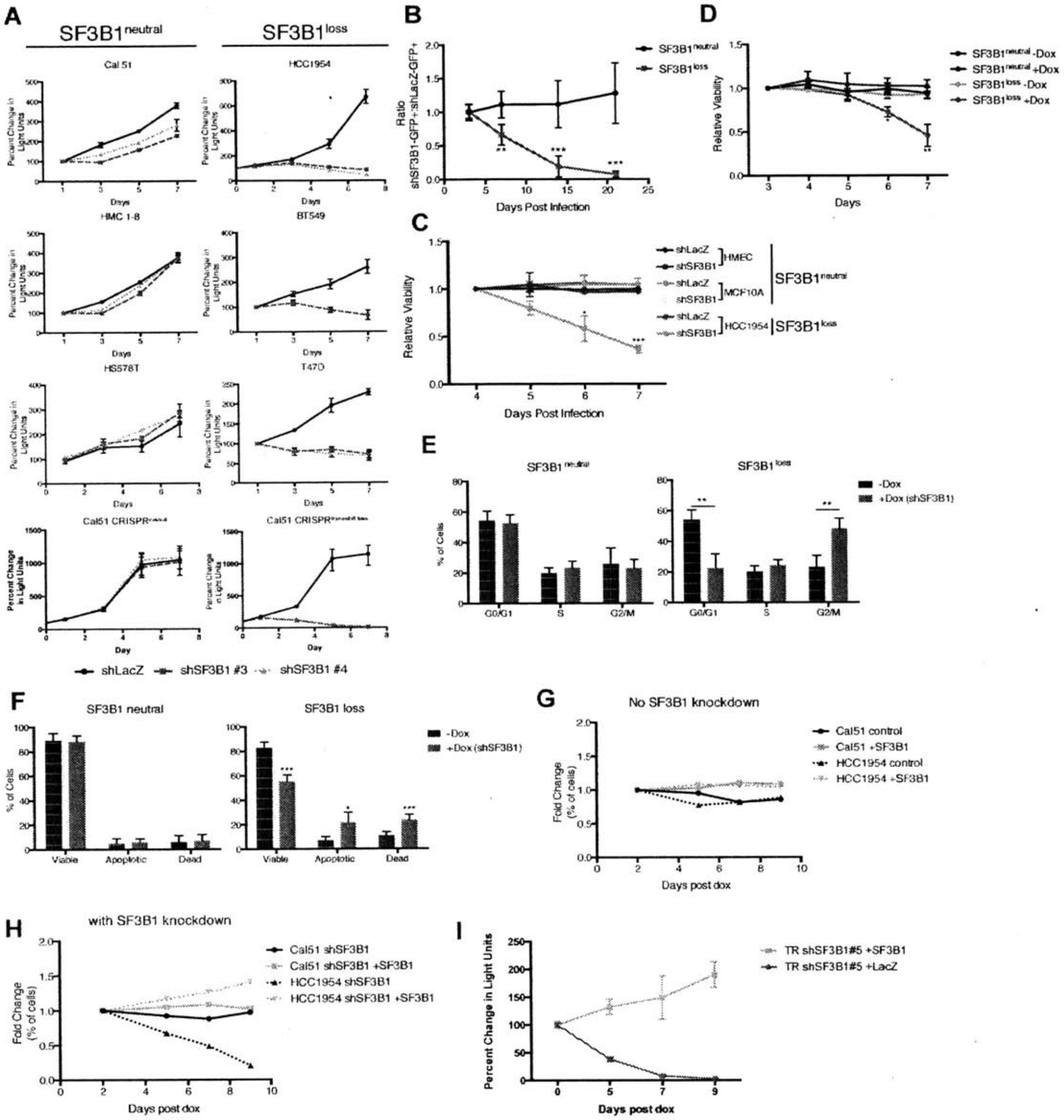


**Figure 2.2: Characterization of SF3B1 as a CYCLOPS gene. (A)** Cell-titer glo growth assays in breast cancer cell lines expressing shLacZ (black) or shSF3B1 (red and orange). Data represents mean +/- SD, SF3B1$^{neutral}$ (n=6), SF3B1$^{loss}$ (n=5). **(B)** The effect of *SF3B1* suppression on the ratio of GFP+ cells expressing either shLacZ-GFP or shSF3B1-GFP. Error bars represent +/- SD. **(C)** Propidium iodide viability from cells

expressing shLacZ or two SF3B1 hairpins (shSF3B1#3 and shSF3B1#4). Error bars represent +/- SD. **(D)** Propidium iodide viability from cells expressing doxycycline activated SF3B1 hairpins (TR-shSF3B1#3 and TRshSF3B1#5), n=3 for each group. **(E)** Cell cycle distribution upon SF3B1 suppression. Data are mean +/- SD, n=3 for each group. **(F)** Quantification of apoptosis upon *SF3B1* suppression by AnnexinV/PI flow cytometry. Data are mean +/- SD. **(G)** Change in ratio of cells expressing SF3B1-GFP or uninfected cells. Representative experiment performed in duplicate. **(H)** Change in ratio of cells expressing shSF3B1 with or without expression of SF3B1-GFP. Representative experiment from two biological replicates. **(I)** Cell Titer Glo growth assay in LacZ or SF3B1 expressing *SF3B1*$^{loss}$ cells upon *SF3B1* suppression. For all panels, *p<0.05 **p<0.01 ***p<0.001. See also Figure S2.

### *SF3B1*$^{neutral}$ cells contain excess SF3B1 beyond the requirement for survival

Analyses of *SF3B1* mRNA indicates that *SF3B1*$^{neutral}$ cells tolerate partial *SF3B1* suppression because they express more SF3B1 than they require. In both TCGA breast adenocarcinoma data (777 samples) [21] and the Cancer Cell Line Encyclopedia (947 cell lines) [15], *SF3B1*$^{neutral}$ samples had significantly higher expression of *SF3B1* mRNA relative to *SF3B1*$^{loss}$ samples (Figure 2.3A and S3A; TCGA Mann-Whitney $p<1x10^{-4}$, CCLE Mann-Whitney $p<1x10^{-4}$), suggesting excess mRNA over requirements for survival. We validated that *SF3B1*$^{neutral}$ breast cancer cell lines (n=7) express approximately twice as much *SF3B1* mRNA as *SF3B1*$^{loss}$ cells (n=5) by quantitative PCR (Figure 2.3B; $p<1x10^{-4}$) and found similar *SF3B1* expression changes between the CRISPR$^{neutral}$ and CRISPR$^{loss}$ lines; Figure S2.3B).

These differences in *SF3B1* mRNA expression were recapitulated at the protein level. Among breast cancer lines, Western blots indicated increased SF3B1 protein expression in *SF3B1*$^{neutral}$ compared to *SF3B1*$^{loss}$ cells (Figure 2.3C) and these differences were recapitulated in CRISPR$^{neutral}$ vs. CRISPR$^{frameshit-loss}$ cells (Figure 2.3D).

These observations suggest that *SF3B1*$^{neutral}$ cells tolerate partial *SF3B1* suppression because moderate SF3B1 suppression leaves them with sufficient residual protein for survival. Indeed, immunoblots of *SF3B1*$^{neutral}$ cells after SF3B1 suppression indicated detectable SF3B1 levels, whereas no protein could be detected in *SF3B1*$^{loss}$ cells after SF3B1 suppression (Figure 2.3E).

A systematic analysis of shRNA-induced mRNA suppression across *SF3B1*$^{neutral}$ and *SF3B1*$^{loss}$ lines indicated that *SF3B1* mRNA levels can be reduced by ~60% from *SF3B1*$^{neutral}$ cell basal levels before proliferation defects are apparent (Figure 2.3F). We suppressed *SF3B1* using shRNAs with different potency to generate a range of *SF3B1* suppression in neutral and copy-loss cells. Although similar reductions in SF3B1 expression were obtained in *SF3B1*$^{neutral}$ and *SF3B1*$^{loss}$ lines, the elevated basal levels of *SF3B1* expression in *SF3B1*$^{neutral}$ lines enabled them to retain sufficient *SF3B1* for proliferation despite shRNA expression.

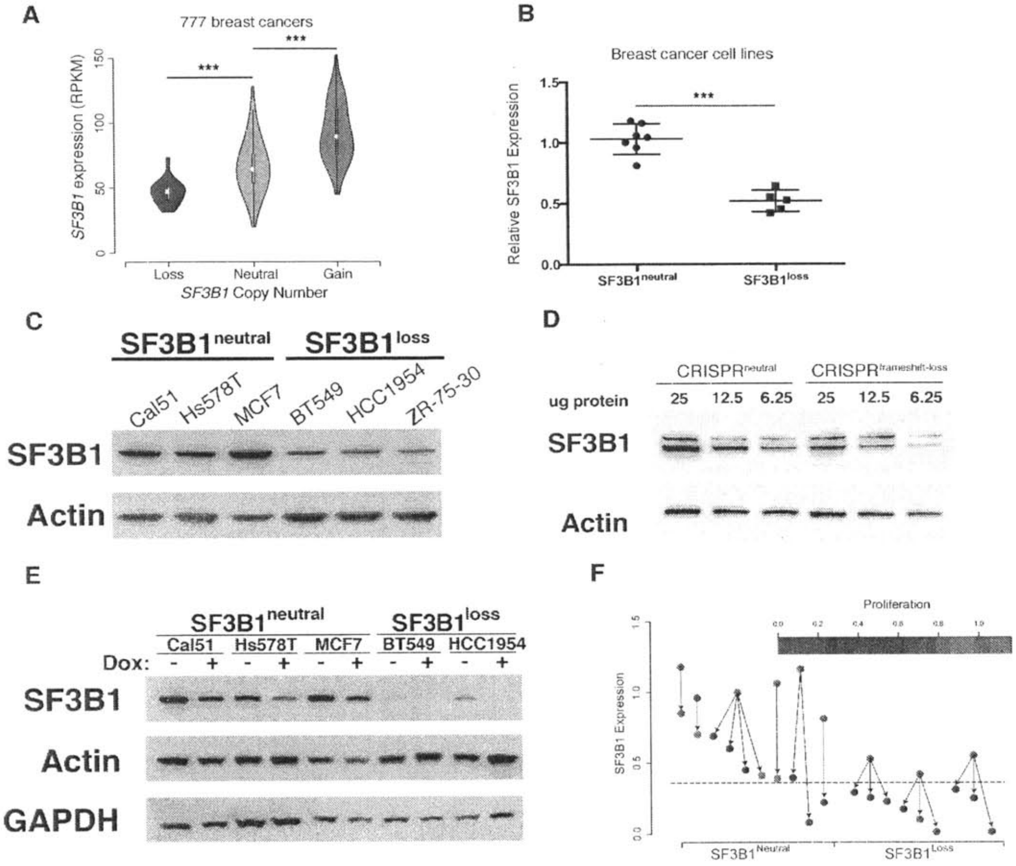


**Figure 2.3: SF3B1$^{loss}$ cells have reduced SF3B sub-complex to buffer U2 snRNP assembly.** **(A)** *SF3B1* expression from 777 breast adenocarcinomas segregated by *SF3B1* copy number. **(B)** Quantitative RT-PCR of *SF3B1* expression in breast cancer cell lines. Data points represent individual cell lines, horizontal line indicates mean and error bars are +/- SD. **(C)** SF3B1 levels in breast cancer cell lines by western blot. **(D)** Serial dilution of lysates from CRISPR$^{neutral}$ and CRISPR$^{frameshift\text{-}loss}$ cells probed by immunoblot for SF3B1. **(E)** SF3B1 immunoblot from SF3B1$^{neutral}$ and SF3B1$^{loss}$ cells without and with TR-shSF3B1#5 induction. **(F)** Schematic combining data indicating reduction in proliferation by cell titer glo assay (red=high proliferation, blue=low proliferation), and relative level of SF3B1 expression before and after SF3B1 suppression detected by

qPCR. Data points with multiple arrows represent individual cell lines with more than one SF3B1 shRNA assayed. Dashed line represents the minimum threshold of SF3B1 expression required for survival. For all panels, *p<0.05 **p<0.01 ***p<0.001. See also Figure S3.

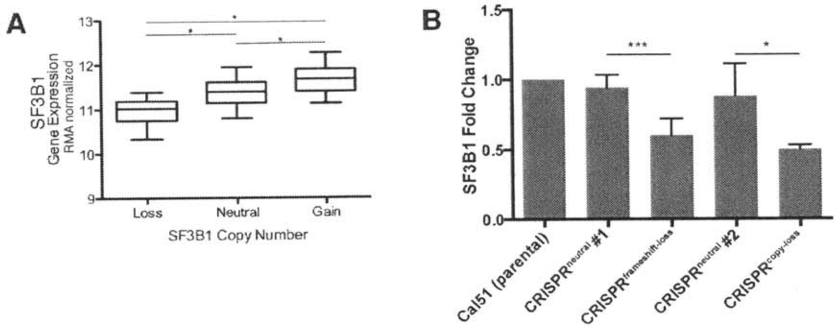


**Figure S2.3, related to Figure 2.3:** **(A)** *SF3B1* expression from 974 cell lines classified by *SF3B1* copy-number status. *p<0.0001. **(B)** SF3B1 RNA expression measured by qPCR in isogenic cells engineered to be SF3B1. CRISPR cell lines refer to cell lines in which one allele was inactivated by CRISPR.

### *SF3B1* copy-loss selectively reduces the abundance of the SF3b complex

We next asked whether the reduction of SF3B1 protein expression in SF3B1$^{loss}$ cells preferentially altered specific SF3B1-containing complexes. SF3B1 is a component of the seven-member SF3b sub-complex of the U2 snRNP. Incorporation of SF3b into the U2 snRNP 12S "core" forms the 15S U2 snRNP, which combines with SF3a to form the full 17S U2 snRNP (Figure 2.4A) [22,23].

We therefore interrogated expression levels of native SF3B1-containing complexes from whole-cell extracts by glycerol gradient sedimentation and gel filtration chromatography. We were able to resolve protein complexes from 29-650 kDa and 650-2,000 kDa using 10-30% glycerol gradients and Sephacryl S-500 gel filtration chromatography, respectively (Figure S4A-B). This enabled resolution of SF3B1-containing complexes ranging from monomers (155 kDa) to the SF3b sub-complex (450 kDa) to the 15S and 17S U2 snRNPs (790 and 987 kDa, respectively) [23]. We compared these elution profiles between patient-derived and isogenic *SF3B1*$^{loss}$ and *SF3B1*$^{neutral}$ cells.

We observed substantially lower levels of SF3B1-containing complexes in the *SF3B1*$^{loss}$ cells in glycerol gradient fractions corresponding to ~450 kDa (fractions 4-6; Figure 2.4B-C) and in gel filtration chromatography fractions corresponding to the lowest masses (Figure S2.4C). We then asked if *SF3B1* copy-loss reduced all SF3B1-containing complexes equally by comparing dilution series from pooled gradient fractions 4-6, 12-14 and 25. Western blots from diluted fractions revealed *SF3B1*$^{loss}$ cells had dramatically reduced SF3B1 complexes in fractions 4-6 and 12-14, but only a modest, if any, reduction in fraction 25 (Figure 2.4D). Importantly, suppression of *SF3B1* in *SF3B1*$^{neutral}$

cells phenocopies the reduction of *SF3B1* in precursor complexes observed in *SF3B1*$^{loss}$ cells (Figure 2.4E). Immunoprecipitation of SF3B1 from glycerol gradient fractions 24-25 resulted in the co-precipitation of the U2 snRNP components SNRPB2 and SF3A3, suggesting these fractions contain the fully assembled U2 snRNP (Figure S2.4E).

Molecular characterization of the protein complex components suggests SF3b is in excess in *SF3B1*$^{neutral}$ cells. Independent evaluation by native western blotting from pooled glycerol gradient fractions corroborated the loss of a single SF3B1 complex approximately 450 kDa in mass in *SF3B1*$^{loss}$ cells (Figure 2.4G), corresponding to the theoretical molecular weight of the SF3b complex. To examine the components of the excess 450 kDa complex present in *SF3B1*$^{neutral}$ cells, we immunoprecipitated SF3B1 from glycerol gradient fractions 4-6 in copy-neutral cells. SF3B1 immunoprecipitation resulted in the co-precipitation of SF3B3 and SF3B4, but not U2 snRNP members SNRPB2 and SF3A3 (Figure S2.4E). Based on the estimated mass of the complex and presence of SF3B3 and SF3B4 by immunoprecipitation, we conclude that copy loss of SF3B1 reduces the abundance of the SF3b complex as a precursor to U2 snRNP formation.

Conversely, it appears that U2 snRNP levels are only modestly decreased in *SF3B1*$^{loss}$ lines. At >790 kD, the U2snRNP would be expected to be in fraction 25 of the glycerol gradients, in which SF3B1 levels were similar between *SF3B1*$^{neutral}$ and *SF3B1*$^{loss}$ lines (Figure 2.4D). U2 snRNA levels are known to track with U2 snRNP levels, and we also did not observe a significant difference in U2 snRNA abundance between *SF3B1*$^{neutral}$ and *SF3B1*$^{loss}$ lines, although there was a trend towards lower expression in the *SF3B1*$^{loss}$ lines (Figure 2.4F p=0.39, two-tailed t-test). These data suggest that copy-loss of *SF3B1* only modestly affects U2 snRNP abundance but substantially decreases levels of U2 snRNP precursor complexes.

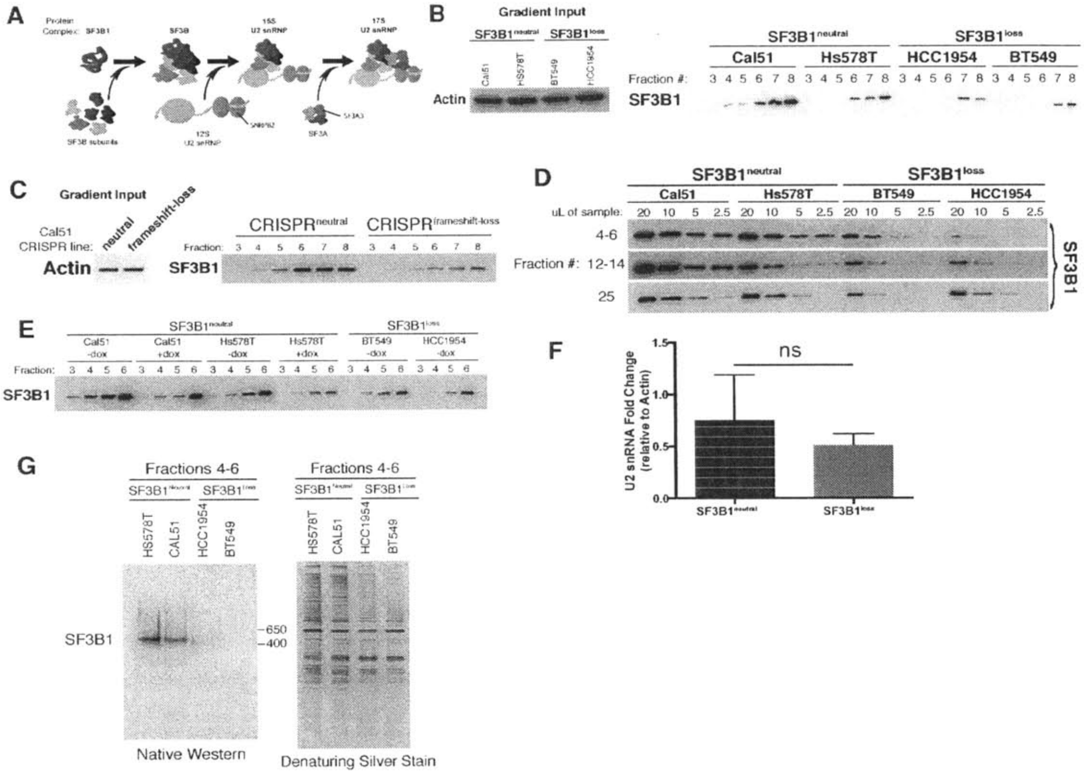


**Figure 2.4: SF3B1$^{loss}$ cells have reduced SF3B sub-complex to buffer U2 snRNP assembly. (A)** Diagram of U2 snRNP assembly. Glycerol gradient fractionation of native whole-cell lysates probed by western blot in breast cancer cell lines **(B)** and isogenic copy cells generated by CRISPR **(C)**. **(D)** Serial dilution of pooled glycerol gradient fractions probed for SF3B1 by immunoblot. **(E)** Glycerol gradient fractions from *SF3B1$^{neutral}$* cells without and with *SF3B1* suppression compared to *SF3B1$^{loss}$* without suppression. **(F)** Quantitative RT-PCR for U2 snRNA expression in three *SF3B1$^{neutral}$* and three *SF3B1$^{loss}$* breast cancer cell lines. Ns = not significant, p=0.39, data represent mean +/- sd. **(G)** (left) SF3B1 Native PAGE immunoblot of pooled glycerol gradient fractions. (right) denaturing silver stain of total protein from pooled fractions shown on right.

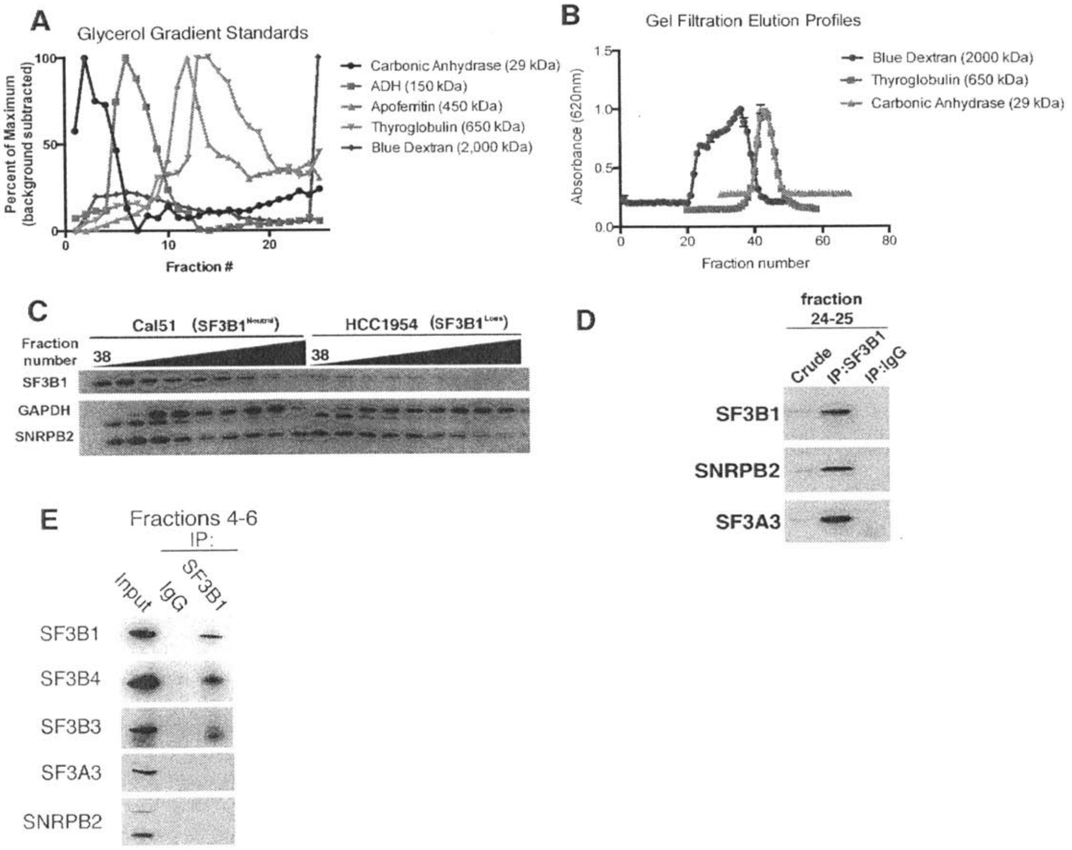


**Figure S2.4, related to Figure 2.4:** **(A)** Sedimentation of mass standards in 10-30% glycerol gradients. **(B)** Elution profiles of mass standards in gel filtration chromatography columns. **(C)** Immunoblot of indicated gel filtration fractions. GAPDH (upper band) and SNRPB2 (lower two bands) represent markers for complexes <700 kDa and spliceosome precursors respectively. **(D)** Immunoblot after SF3B1 immunoprecipitation from pooled glycerol gradient fractions 24-25. **(E)** Immunoblot after SF3B1 immunoprecipitation from pooled glycerol gradient fractions 4-6.

### SF3B1 suppression selectively reduces U2 snRNP abundance in *SF3B1$^{loss}$* cells

We reasoned that for the viability of cells to be affected by SF3B1 suppression, splicing function and therefore U2 snRNP complex assembly must be impaired by SF3B1 deficiency. We therefore subjected whole cell extracts to glycerol density gradients and immunoblotted for proteins to identify the stages of U2 snRNP assembly. Upon *SF3B1* suppression, *SF3B1$^{loss}$* extracts contained decreased abundance of SNRPB2 and SF3A3 from the gradient inputs, suggesting a reduction in the fully assembled U2 snRNP (Figure 2.5A). We then examined if partial *SF3B1* suppression preferentially removed precursor sub-complexes instead of the assembled U2 snRNP. Upon *SF3B1* suppression, there was a more substantial reduction of SF3B1 in fractions 4-6 than fraction 25 suggesting that *SF3B1* knockdown reduced smaller SF3B1-containing protein complexes, likely corresponding to the SF3B sub-complex, instead of the assembled U2 snRNP (Figure 2.5B-C). Further, *SF3B1$^{loss}$* cells exhibited dramatic reductions in SF3A3 and SNRPB2 in fraction 25 that do not occur in *SF3B1$^{neutral}$* cells (Figure 2.5C). These data indicate that upon *SF3B1* suppression, copy-loss cells have decreased amounts of fully assembled U2 snRNP that does not occur in copy-neutral cells.

We verified the observation of decreased U2 snRNP abundance in copy-loss cells after *SF3B1* suppression by gel filtration chromatography and quantification of U2 snRNA expression. Native protein extracts from *SF3B1$^{neutral}$* and *SF3B1$^{loss}$* cells after *SF3B1* suppression were fractionated on Sephacryl-S500 columns assayed for U2 snRNP components SF3B1 and SNRPB2 in fractions containing complexes from 700-2,000 kDa that contain assembled U2 snRNP [24]. Immunoblots of fractions from copy-neutral cells after *SF3B1* suppression contained both SF3B1 and SNRPB2, while copy-loss cells had no detectable U2 snRNP components (Figure 2.5D-E). Quantitative PCR from *SF3B1*$^{loss}$ cells resulted in significantly reduced U2 snRNA expression after *SF3B1* suppression that was not observed in *SF3B1*$^{neutral}$ cells (Figure 2.5F), indicating reduced U2 snRNP abundance in copy-loss cells.

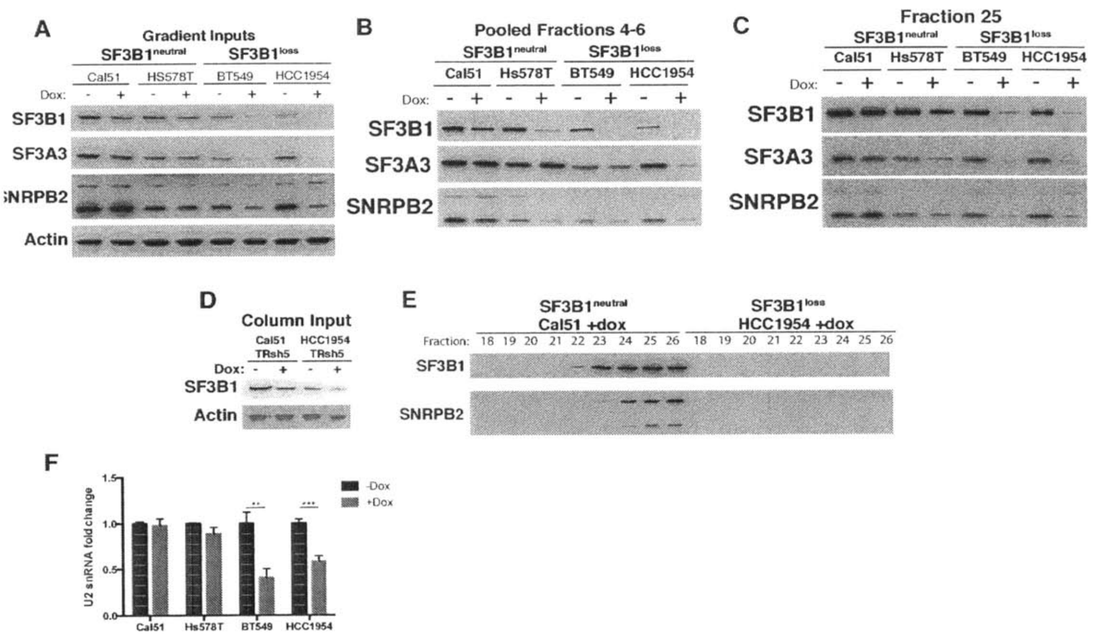


**Figure 2.5: Reduced spliceosome precursors and U2 snRNP abundance upon *SF3B1* suppression in *SF3B1*$^{loss}$ cells. (A)** Western immunoblots without and with *SF3B1* suppression prior to glycerol gradient fractionation. **(B)** Western immunoblots from pooled glycerol gradient fractions 4-6 (protein complexes ~150-450 kDa). **(C)** Western immunoblots from glycerol gradient fractions 25 (protein complexes >650 kDa). **(D)** Immunoblot from lysates prior to gel filtration chromatography. **(E)** Immunoblot of gel filtration fractions 18-26 (protein complexes >650kDa) from lysates with *SF3B1* suppression. **(F)** Quantitative RT-PCR for U2 snRNA expression without and with *SF3B1* suppression. For all panels, *p<0.05 **p<0.01 ***p<0.001. See also Figure S5.

**RNA-sequencing reveals loss of splicing function after *SF3B1* suppression**

*SF3B1* is a well-established mRNA splicing factor [25,26], therefore we evaluated transcriptomes of both *SF3B1*$^{loss}$ and *SF3B1*$^{neutral}$ cells before and after suppression of SF3B1. Suppression of *SF3B1* in *SF3B1*$^{neutral}$ cells reduced *SF3B1* expression comparably to steady-state levels in *SF3B1*$^{loss}$ cells. Upon *SF3B1* suppression, 513 genes were differentially expressed at an FDR <10% and only 306 genes were differentially expressed in *SF3B1*$^{neutral}$ cells. These data are consistent with the hypothesis that *SF3B1* suppression more severely impacts the transcriptome of *SF3B1*$^{loss}$ cells.

defects in mRNA splicing occurred at time points prior to the reduced viability observed in *SF3B1*$^{loss}$ cells (day 4 post *SF3B1* suppression, Figure 2C-D). We performed RNA-sequencing to characterize the transcriptome-wide effects of *SF3B1* suppression in *SF3B1*$^{neutral}$ and *SF3B1*$^{loss}$ cells. The baseline mRNA expression of *SF3B1* in *SF3B1*$^{neutral}$ cells was greater than in *SF3B1*$^{loss}$ cells (Figure 2.6A), consistent with previous experiments (Figure 2.3A-B; S2.3A-B).

Due to the role of *SF3B1* in pre-mRNA splicing, we sought to quantify the extent of splicing disruption in *SF3B1*$^{neutral}$ and *SF3B1*$^{loss}$ cells. Intron retention has been reported upon treatment of cells with spliceosome inhibitors [6] and in patients harboring *SF3B1* mutations [10]. We used juncBase [27] and a novel statistical framework to analyze 50,600 splice junctions for intron retention in *SF3B1*$^{neutral}$ and *SF3B1*$^{loss}$ cells upon SF3B1 suppression (see Methods). All cells showed evidence of increased intron retention following *SF3B1* suppression ($p<10^{-5}$). However, splicing was significantly more affected in *SF3B1*$^{loss}$ cells compared to *SF3B1*$^{neutral}$ cells. Upon *SF3B1* suppression, 632 transcripts in *SF3B1*$^{loss}$ cells showed evidence of significantly ($q<0.01$) increased intron retention relative to *SF3B1*$^{neutral}$, whereas only 11 transcripts showed evidence of increased intron retention in the reverse direction (Figure 2.6C, p=4.9x10$^{-171}$).

We confirmed the alterations in mRNA splicing by RT-PCR. Primers were designed for two ubiquitously expressed genes that flanked short introns amenable to PCR detection if they are improperly retained (Figure 2.6D). Upon *SF3B1* knockdown, *SF3B1*$^{loss}$ cells contained *RPS18* and *CALR* transcripts with retained introns that were not observed in *SF3B1*$^{neutral}$ cells (Figure 2.6E), consistent with our RNA-seq analysis. Furthermore, loss of *SF3B1* expression can potently alter the alternative splicing of *MCL1*, converting it to a shorter isoform (MCL1-s) defective in anti-apoptotic function [28]. RT-PCR of *SF3B1*$^{neutral}$ and *SF3B1*$^{loss}$ cells after *SF3B1* suppression resulted in significantly increased ratio of the MCL1-s isoform only in *SF3B1*$^{loss}$ cells (Figure 2.6F-G).

We next examined if alterations to the organization of nuclear speckles occurred after *SF3B1* suppression by SC-35 immunofluorescence. Spliceosome components, including SF3B1, are thought to assemble and function in sub-nuclear compartments known as nuclear speckles [29]. We performed an unbiased quantification of the number and size of SC-35$^{+}$ speckles per nucleus using a custom image analysis pipeline with CellProfiler software [30]. *SF3B1*$^{neutral}$ cells did not display changes in SC-35+ speckles after *SF3B1* suppression, however *SF3B1*$^{loss}$ nuclei contained significantly fewer speckles and increased speckle area (Figure 2.6H-J). The formation of enlarged 'mega-speckles' were previously observed in cells treated with either mRNA splicing or transcriptional inhibitors [6,31] and suggest defects in spliceosome assembly and function. The presence of defective alternative splicing, intron retention and formation of mega-speckles uniquely in *SF3B1*$^{loss}$ cells after *SF3B1* suppression suggests gross defects in mRNA splicing.

Taken together, these data demonstrate that defects in pre-mRNA processing (Figure 2.6B-G) and nuclear speckle localization (Figure 2.5H-J) occur only in copy-loss cells upon *SF3B1* suppression. These splicing defects are a result of decreased U2 snRNP abundance (Figure 2.5) after *SF3B1* suppression and suggest that SF3B1 is a limiting factor to U2 snRNP assembly and function in *SF3B1*$^{loss}$ cells.

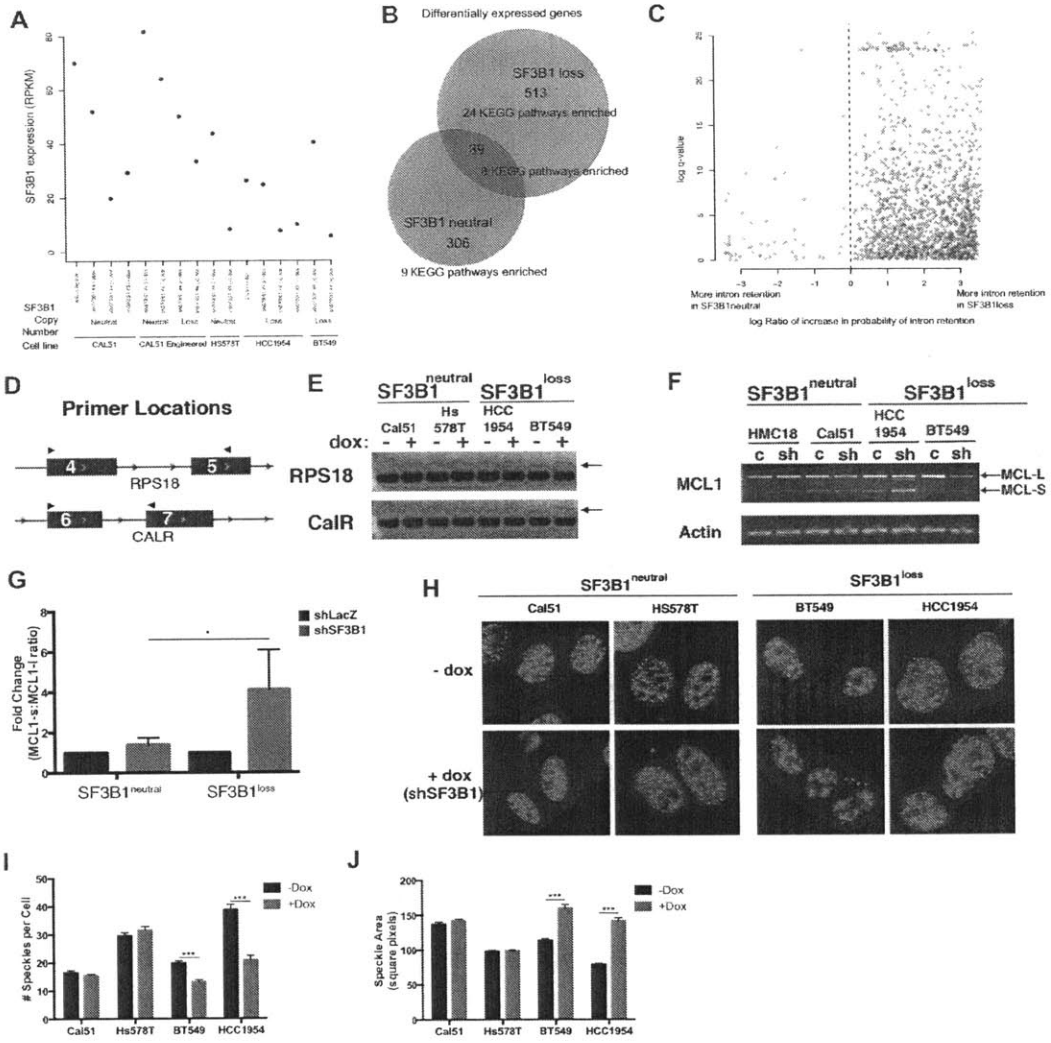


**Figure 2.6: Transcriptome-wide analysis reveals reduced pre-mRNA processing in *SF3B1$^{loss}$* cells after *SF3B1* suppression.** **(A)** SF3B1 expression quantified by RNA-Sequencing upon SF3B1 suppression in *SF3B1$^{neutral}$* and *SF3B1$^{loss}$* cell lines. **(B)** Number of differentially expressed genes upon *SF3B1* suppression ($q < 0.1$) and the number of enriched KEGG pathways amongst indicated gene set ($q < 0.05$). **(C)** (x-axis) The relative risk of intron retention in *SF3B1$^{neutral}$* and *SF3B1$^{loss}$* cells after SF3B1 suppression. (y-axis) Significance of the difference in effect of SF3B1 suppression on

intron retention between *SF3B1*$^{neutral}$ and *SF3B1*$^{loss}$. **(D)** Schematic of primer locations. Arrowheads indicate the locations of primers used for intron retention. **(E)** RT-PCR for *RPS18* and *CALR* in cells without and with *shSF3B1* induction by doxycycline. Arrows indicate PCR products corresponding to retained introns. **(F)** Representative RT-PCR from SF3B1$^{neutral}$ and SF3B1$^{loss}$ cells after *SF3B1* knockdown. "c" are LacZ control hairpins, "sh" are shSF3B1#4 hairpins. **(G)** Densitometric quantification of the ratio of MCL1-S:MCL1-L from 3 biological replicates. **(H)** Immunofluorescent images of nuclear spackles by anti-SC35 (SRSF2) staining. Scale bars = 10 uM. **(I)** Quantification of number of nuclear speckles per cell in panel (F). **(J)** Quantification of nuclear speckle area in panel (F). For all panels, *p<0.05 **p<0.01 ***p<0.001. See also Figure S6.

### Suppression of SF3B1 reduces tumor growth in *SF3B1*$^{loss}$ xenografts

To explore the potential of SF3B1 as a therapeutic target, we evaluated existing SF3b inhibitors that target SF3B1. We tested the sensitivity of CRISPR-induced copy loss cells, and *SF3B1*$^{neutral}$ cells with partial SF3B1 suppression to treatment with spliceosome inhibitors. Neither approach exhibited increased sensitivity to SF3b inhibitors or NSC95397, a compound reported to inhibit splicing activity by an SF3b-independent mechanism (Figure S2.5A-D)[32]. Therefore, we evaluated the effect of SF3B1 suppression on xenograft tumor growth using the doxycycline-regulated shRNA system.

Suppression of *SF3B1* in *SF3B1*$^{loss}$ cells reduces xenograft growth *in vivo*. We generated luciferase-labeled cell lines from the CRISPR$^{frameshift\text{-}loss}$ and CRISPR$^{neutral}$ cells containing TR-shSF3B1#3. Animals were placed on doxycycline upon detection of established tumors. Suppression of *SF3B1* only reduced the growth of xenografts from CRISPR$^{\text{frameshift-loss}}$ cells and did not affect growth of CRISPR$^{\text{neutral}}$ cells or CRISPR$^{\text{frameshift-loss}}$ cells without doxycycline (Figure 2.7A-B).

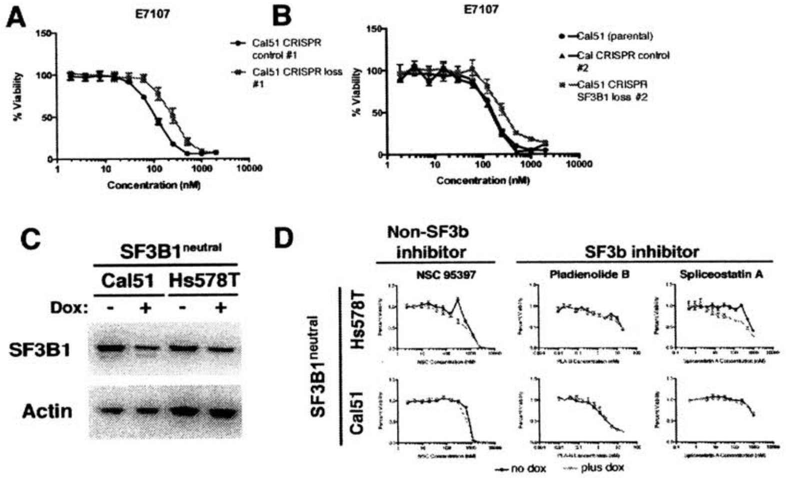


**Figure S2.7, related to Figure 2.7: (A-B)** Drug sensitivity curves for isogenic *SF3B1*$^{neutral}$ and *SF3B1*$^{loss}$ cells after treatment with E7107. **(C)** SF3B1 immunoblot demonstrating knockdown for experiment in D. **(D)** Drug sensitivity curves for indicated splicing inhibitors in cells without and with *SF3B1* suppression.

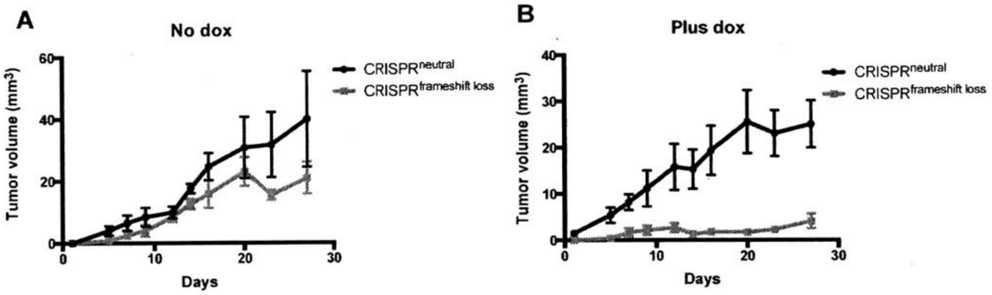


**Figure 2.7: SF3B1 suppression inhibits the growth of *SF3B1*$^{loss}$ cells in-vivo. (A)** Growth of nude mouse xenografts of isogenic CRISPR$^{\text{frameshift-loss}}$ and CRISPR$^{\text{neutral}}$ cell

lines. **(B)** Growth of nude mouse xenografts of isogenic CRISPR$^{\text{frameshift-loss}}$ and CRISPR$^{\text{neutral}}$ cell lines with suppression of SF3B1 using a tetracycline inducible shRNA.

## Discussion

### Genomic features affecting CYCLOPS genes

We identified 170 CYCLOPS genes from the Achilles RNAi viability screen across 216 cell lines. However we expect these findings to under estimate the number of true CYCLOPS genes in cancer genomes. Our ability to precisely suppress gene expression by approximately 50% is limited by the characteristics of shRNAs used in the Achilles data set. Genes could not be evaluated if the hairpins targeting their expression either did not effectively suppress expression, or completely ablated their expression. Indeed, the number of genes for which we could accurately evaluate cell viability across all cell lines was 8,321. Furthermore, we are unlikely to have identified tumor-type specific CYCLOPS genes due to the relatively small numbers of cell lines screened in each cell lineage.

Shared molecular features of among CYCLOPS genes may have aided in their identification. We identified that CYCLOPS genes tend to be more uniformly expressed throughout normal tissues (Figure 1D), suggesting that their consistent expression across cell lines aided in the accurate determination of cell line sensitivity after gene suppression. CYCLOPS genes also frequently comprise components of multi-protein complexes. It is possible that the stoichiometric equilibrium of precursors to multi-protein complexes are disrupted in cancer cells but do not affect the assembled complex's function. Therefore, aneuploid cancer cells may not need to compensate for expression changes as a result of copy number alterations. However after further gene suppression, essential multi-protein complexes can have reduced function and unveil CYCLOPS vulnerabilities.

### SF3B1 as a CYCLOPS gene and therapeutic target

The observation that *SF3B1*$^{loss}$ cells depend on the remaining *SF3B1* expression for survival suggests *SF3B1* is a novel therapeutic target in cancers harboring hemizygous *SF3B1* deletion. A recent unexpected finding also suggests that *SF3B1* mutated cancer cells do not depend on the mutated *SF3B1* allele, but rather the remaining wild-type copy [33]. Taken together, these data support the notion that SF3B1 inhibitors may have therapeutic benefit in cancers with SF3B1 copy-loss or mutation. Surprisingly, current compounds that target SF3B1 and the SF3B sub-complex do not appear to differentially kill *SF3B1*$^{loss}$ cells (Figure 2.4H and S2.4G) and whether these compounds are more effective in SF3B1 mutated cancers remains unclear.

Compounds with different mechanisms of action than current "SF3B1 inhibitors" likely will be required to treat SF3B1-dependent cancers. The observation that hemizygous loss of *SF3B1* reduces SF3B precursor sub-complexes, and not assembled U2 snRNP, suggests that current inhibitors function to inhibit splicing by disrupting U2 snRNP activity. These observations are consistent with previous work demonstrating that E7107, a pladienolide D derivative, prevents an ATP-dependent conformational change in the U2 snRNP[34]. Instead, we predict that chemical approaches to disrupt the incorporation of SF3B1 into spliceosomeal complexes, or down-regulate SF3B1 expression would be needed to target the SF3B1 CYCLOPS vulnerability. However substantial opportunities still exist to use existing splicing inhibitors in cancer therapy and perhaps use the existing pharmacophores to develop compounds that can target the SF3B1 CYCLOPS vulnerability.

**Potential approaches to target CYCLOPS genes**

Partial genomic loss of CYCLOPS genes occur frequently in cancers. It is possible that many non-specific chemotherapies, such as flavopiridol which inhibits transcription, exert some of their cytotoxic effects by further reducing CYCLOPS gene expression in cells with hemizygous loss of CYCLOPS genes. Other existing compounds may also be selectively cytotoxic in aneuploid cancers when expression of the target of inhibition is reduced. This is particularly appealing when considering many CYCLOPS genes function in the same essential cellular complexes such as the spliceosome, proteasome and ribosome. It is possible that current spliceosome, proteasome or ribosome inhibitors could selectively treat cancers undergoing partial copy-loss of the small-molecule target. Further investigation into the development and understanding of inhibitors targeting these pathways could provide novel opportunities for identifying patient populations that can benefit from these drugs.

Recent work also suggests that partial reduction of essential gene expression results in a "therapeutic window" that can be targeted pharmacologically; an important observation for the development of CYCLOPS-based therapies. For example, treatment of myeolodysplastic syndrome harboring hemizygous deletion of *CSNK1A1* confers sensitivity to CSNK1A1 inhibitors [35]. Similarly, partial gene deletion of *POLR2A* can sensitize cancer cells to treatment with alpha-amanitin, an inhibitor of POLR2A [36]. Our work and the aforementioned studies support the idea that partial inhibition of essential genes represents novel cancer targets and that their inhibition can be well tolerated to provide a therapeutic window.

### Cellular and molecular basis mediating cell death in *SF3B1$^{loss}$* cells

*SF3B1$^{loss}$* cells arrest in G2/M and subsequently initiate apoptosis upon *SF3B1* suppression (Figure 2.2E-F), however the molecular mechanisms mediating these processes remain largely unknown. Recent work suggests that the sister chromatid cohesion factor, *CDCA5* (Sororin), is highly dependent on SF3B1 for its mRNA splicing [37], and may provide a mechanism for M-phase arrest upon splicing factor suppression. Our data corroborate those findings, as *CDCA5* is one of the most differentially expressed genes in *SF3B1$^{loss}$* cells after *SF3B1* suppression.

Apoptosis induction is crucial for mediating the effects of the *SF3B1* CYCLOPS vulnerability. We have yet to characterize the molecular components required for apoptosis induction upon *SF3B1* suppression in *SF3B1$^{loss}$* cells. Our observations and the work of others on the role of alternative splicing in apoptosis induction suggests that BH3-domain containing proteins, including MCL1 and BCL2L1, are potently regulated by alternative splicing (Figure 2.5A; [28,38,39]. The role of these genes as essential downstream effectors of cell death after *SF3B1* suppression remains unexplored.

## Experimental Procedures

### Analysis of Genome-Wide Copy-Number Induced Cancer Dependencies

For each gene dependency score in Achilles, pearson correlation p-values were calculated for association with copy number at 22,202 loci using cor.test in R. All analyses were conducted in the R programming language. Correction for multiple hypotheses was performed with the Benjamini and Hochberg method.

### CYCLOPS Analysis

Gene level, $\log_2$ relative copy number from 1043 cell lines were downloaded from the CCLE portal (http://www.broadinstitute.org/ccle, data version 5/27/2014). Samples were assigned one of two classifications: 1) copy loss cell lines had $\log_2$ relative copy number ratio <=-0.35, and 2) copy neutral cell lines had $\log_2$ relative copy number ratio >=-0.35. 214 cell lines with copy number data were also profiled for genome-wide RNAi viability in Project Achilles (version 2.4.3). Data copy number data and RNAi viability data were integrated using R statistical software.

Evaluation of differential sensitivity to gene suppression based on copy loss was done using lineage controlled permutation tests. Briefly, cell lines from each tumor type were classified as copy neutral or copy loss for each gene. The relative viability between the two classes were calculated by using the mean ATARiS score for each group [16]. Significance was determined by permuting class labels between the two groups. Correction for multiple hypotheses was performed using the Bejamini and Hochberg method.

### Classification of length and amplitude for copy number alterations

For relative $\log_2$ normalized copy number data analyzed from 10,570 tumors from TCGA (including hyperdiploid samples), the following thresholds were used for copy number classification: homozygous loss $\log_2$ values <-1.07, hemizygous loss $\log_2$ <=-0.1. Broad SCNA events are defined as a copy number region of greater than half a chromosome arm, all other smaller events are considered focal. For cancer cell lines used in functional studies: copy-loss cells had $\log_2$ copy number <=-0.35, and copy neutral cells had $\log_2$ copy number >-0.2 and <0.2.

**Tissue Culture**

Human cancer cell lines were maintained in RPMI-1640 supplemented with 10% fetal bovine serum and 1% penicillin and streptomycin and were assayed to be free of mycoplasma. Non-transformed MCF10A and HMEC cells were cultured in Mammary Epithelial Growth Medium (CC-3150, Lonza). For cells expressing tetracycline-regulated shRNAs, tetracycline-approved fetal bovine serum (Clonetech) was used.

**Correlation analysis of copy-loss of SF3B genes with cell dependencies upon suppression of other SF3B complex genes**

We determined relative copy number and ATARiS gene dependency scores after knockdown of each SF3B complex member for 189 cancer cell lines. We performed linear regression analysis and calculated Pearson correlation coefficients for copy number of each SF3B complex gene with knockdown of every SF3B component.

**Quantitative and reverse transcription PCR**

RNA was extracted using the RNeasy extraction kit (Qiagen) and subjected to on-column DNase treatment. cDNA was synthesized with the Superscript II Reverse Transcriptase kit (Life Technologies) with no reverse transcriptase samples serving as negative controls. Gene expression was quantified by Power Sybr Green Master Mix (Applied Biosystems). Primers for all genes were determined to be equally efficient over

5 serial two-fold dilutions. Gene expression values were normalized to *ACTB* and the fold change calculated by the $\Delta\Delta C_t$ method. For quantification of the U2 snRNA, the above method was used except total cellular RNA was extracted with Trizol (Life Technologies). A table containing primer information and sequences can be found as Table S4.

**shRNAs targeting SF3B1**

Lentiviral expression constructs for shRNA-mediated suppression of *SF3B1* were obtained through the RNAi-consortium (http://www.broadinstitute.org/rnai/public/). The clone ID’s and names used in our studies are as follows: shSF3B1 #2 (TRCN0000320576), shSF3B1 #3 (TRCN0000320566), shSF3B1 #4 (TRCN0000350273), shSF3B1 #5 (TRCN0000320636).

**Generation of Inducible *SF3B1* shRNA expression system.**

Sense and anti-sense oligonucleotides were annealed and cloned into the *AgeI* and *EcoRI* restriction sites of the pLKO-Tet-puro vector (Addgene, plasmid #21915). The oligonucleotide sequences were:

LacZ (sense) 5’-

CCGGTGTTCGCATTATCCGAACCATCTCGAGATGGTTCGGATAATGCGAACATTTTTG,

LacZ (anti-sense) 5’-

AATTCAAAAATGTTCGCATTATCCGAACCATCTCGAGATGGTTCGGATAATGCGAACA,

TR-shSF3B1#3 (sense) 5’-

CCGGCAACTCCTTATGGTATCGAATCTCGAGATTCGATACCATAAGGAGTTGTTTTTG, TR-shSF3B1#3 (anti-sense) 5’-

AATTCAAAAACAACTCCTTATGGTATCGAATCTCGAGATTCGATACCATAAGGAGTTG,

TR-shSF3B1#5 (sense) 5’-

CCGGCCTCGATTCTACAGGTTATTACTCGAGTAATAACCTGTAGAATCGAGGTTTTTG,

TR-shSF3B1#5 (anti-sense) 5’-

AATTCAAAAACCTCGATTCTACAGGTTATTACTCGAGTAATAACCTGTAGAATCGAGG

**Cellular Growth Assays**

Cells were plated in 96 well plates at 1000 cells per well. Cell number was inferred by ATP-dependent luminescence by Cell Titer Glo (Promega) and normalized to the relative luminescence on the day of plating. For short-term lentiviral infections, cells were infected 24 hours prior to plating.

**GFP Competition Assays**

Oligonucleotides encoding *LacZ* or *SF3B1* shRNA#4 hairpin sequences were annealed and cloned into the pLKO.1 derivative vector TRC047 (pLKO.3pgw) and verified by Sanger sequencing. Cells were infected with serial dilutions of virus to achieve ~50% GFP-positive cells. Cells with approximately equivalent ratios of GFP-positive –and negative cells were assayed by flow cytometry 3 days post infection and at subsequent time-points. The fold change in GFP+ cells was normalized to the percentage present 3 days after infection. For competition assays re-introducing exogenous *SF3B1*, we expressed a human codon-optimized *SF3B1* by lentivirus. Cells were infected as described above and treated with doxycycline two days after infection.

**Propidium Iodide Cell Viability Assays**

Cells were treated with either short-term lentiviral infection or tetracycline-inducible *SF3B1* shRNAs. After treatment, cells were trypsinized and pelleted including any cells in suspension. Cells were resuspended in propidium iodide viability staining solution (1x PBS, 1% BSA, 2.5 ug/mL propidium iodide) and quantified by live-cell flow cytometry. The change in viability was normalized to the percent of viable cells quantified on the first day of the assay.

**Determination of Cell Cycle Distribution by Propidium Iodide**

Cells were trypsinized, washed and fixed with ice-cold 70% ethanol for a minimum of 15 minutes at 4C. Cells were incubated in propidium iodide cell cycle staining solution (1x PBS, 1% BSA, 50 ug/mL propidium iodide, 100ug/mL RNAse A) for 15 minutes and analyzed by flow cytometry. Debris and aggregates were gated out and cell cycle stage was quantified using Modfit (Varity Software House).

**Annexin-V Apoptosis Assays**

Cellular apoptosis was quantified by live-cell flow cytometry using Alexa-Fluor 488 conjugated Annexin-V (Life Technologies) and propidium iodide. Cells were incubated in Annexin binding buffer containing propidium iodide (10 mM Hepes, 140 mM NaCl, 2.5 mM $CaCl_2$, 2.5 ug/mL propdium iodide) for 15 min, washed and resuspended in FACS buffer (1x PBS, 1% BSA and 50 mM EDTA). Determination of the stage of apoptosis by gating was as follows: viable cells (Annexin-$V^-$/$PI^-$), early apoptosis (Annexin-$V^+$/$PI^-$), late apoptosis (Annexin-$V^+$/$PI^+$), and dead cells (Annexin-$V^-$/$PI^+$).

## Generation of heterozygous *SF3B1*$^{loss}$ cells by CRISPR-Cas9

Short guide RNAs targeting the first constitutively expressed coding exon of *SF3B1* (exon 2) were designed with the aid of the Zhang laboratory's web-based application (http://crispr.mit.edu/). Sense and anti-sense oligonucelotides were annealed and cloned into *BbsI* site of pX458 (Addgene) and verified by Sanger sequencing. Oligonucleotide sequences were as follows: 5' CACCGCATAATAACCTGTAGAATCG (forward), 5' 5'AAACCGATTCTACAGGTTATTATGC (reverse). pX458 was transfected with LipoD293 (SignaGen) into the diploid breast cancer cell line, Cal51. 3-4 days post transfection, single GFP+ cells were sorted by FACS and plated at low density for single cell cloning.

19 monoclonal cell lines were genotyped for Cas-9 induced mutations by Sanger sequencing cloned PCR products. All monocolonal lines had either no mutations or harbored biallelic mutations in *SF3B1*. The genotypes of the Cal51 CRISPR cell lines used from this method of generation were: *SF3B1*$^{delT36/delT36}$ (CRISPR$^{\text{neutral}}$ #1) and SF3B1$^{\text{delT36/A23fsX20}}$ (CRISPR$^{\text{frameshift-loss}}$).

A Cas9 construct co-expressing two sgRNAs was used to delete a 57 kb region encoding SF3B1. The guide RNA targeting the 5' upstream of SF3B1 used a mismatch from a heterozygous SNP (rs3849362) in Cal51 to bias towards mono-allelic deletion of *SF3B1*. Oligonucleotides were cloned in a similar fashion as above (with BbsI overhangs). The sequences are as follows: For the 5' guide targeting SNP, 5' CACCGCGCATTATAGATTATGGCCC (forward) and 5' AAACGGGCCATAATCTATAATGCGC (reverse). For the 3' targeting guide:

5'CACCGCGGAGTTTCATCCGTTACAC (forward),

AAACGTGTAACGGATGAAACTCCGC (reverse)

The control cell line (CRISPR$^{\text{neutral}}$ #2), was a monoclonal culture derived from a cell that expressed the Cas9 construct, but did not have the SF3B1 deletion. The induced *SF3B1*$^{loss}$ line (CRISPR$^{copy\text{-}loss}$ #2) was validated by PCR to harbor a 55 kb deletion encoding SF3B1.

**Western Blotting**

For denaturing protein immunoblots, cells were washed in ice cold PBS and lysed in 1x RIPA buffer (10mM Tris-Cl Ph 8.0, 1 mM EDTA, 1% Triton X-100, 0.1% SDS and 140 mM NaCl) supplemented with 1x protease and phosphatase inhibitor cocktail (PI-290, Boston Bioproducts). Lysates were sonicated in a bioruptor (Diagenode) for 5 minutes (medium intensity) and cleared by centrifugation at 15000 x g for 15 min at 4C. Proteins were electrophoresed on polyacrylamide gradient gels (Life Technologies) and detected by chemiluminescence. For native western blotting, cells were washed in ice cold PBS and lysed in 1x sonication buffer (10% Glycerol, 25 mM HEPES pH 7.4, 10 mM $MgCl_2$) supplemented with protease and phosphatase inhibitors. Coomassie blue native PAGE western blots were run according the manufacturer's instructions (Life Technologies).

**SF3B1 Gene expression analysis from TCGA and CCLE datasets**

Relative copy number and Affymetrix expression data for *SF3B1* were downloaded from the CCLE portal (http://www.broadinstitute.org/ccle/home). TCGA breast adenocarcinoma data were downloaded from the cBioPortal (http://www.cbioportal.org/public-portal/index.do) [40,41]. For both datasets, samples lacking either gene expression or copy-number were removed. As described above copy-loss was defined as samples with $\log_2$ normalized relative copy number of <-0.35, copy gain was defined as >=0.3.

**Glycerol Gradient Sedimentation**

Glycerol gradient sedimentation was performed as previously described with slight modifications for use with whole-cell lysates [42]. Briefly, linear 10-30% glycerol gradients were formed by layering a 10% glycerol gradient buffer (20 mM Hepes-KOH (pH 7.9), 150 mM NaCl, 1.5 mM $MgCl_2$ 10% glycerol) atop a 30% glycerol solution with identical salt concentrations. Gradients were formed using a Gradient Station (Biocomp Instruments) according to manufacturers instructions. Cells were lysed in "IP lysis buffer" (50mM Tris, 150 mM NaCl and 1% Triton X-100). 400 uL containing 1-3 mg of crude lysate was loaded per gradient in SW55 centrifuge tubes and spun at 55,000 RPM for 3.5 hours at 4C. 200 uL fractions were collected by manually pipetting from the top of the gradient. Recombinant proteins of known mass were run in parallel gradients as controls.

**Gel Filtration Chromatography**

Sephacryl S-500 (17-0613-05, GE Healthcare) columns were packed into a 50 x 1.5 cm column and equilibrated with column buffer (10 mM Tris, 60 mM KCl, 25 mM EDTA, 1% Triton X-100 and 0.1% sodium azide). Whole-cell lysates were collected in IP lysis buffer (as described above) and incubated with 0.5mM ATP, 3.2 mM $MgCl_2$ and 20 mM creatine phosphate (di-Tris salt) and incubated for 20 min at 30C to dissociate multi-snRNP spliceosomal complexes. 2 mL of lysate containing 5 mg of protein was loaded on columns and 90 1.5 mL fractions were collected overnight at 4C.

### Immunoprecipitation

Immunoprecipitations were performed with pooled glycerol gradient fractions. The Fc region of mouse anti-SF3B1 (Medical and Biological Laboratories, D221-3) was directionally cross-linked to protein G Dynabeads (Life Technologies) using 20 mM dimethyl pimelimidate (DMP). IgG isotype controls were cross-linked and processed identically. Proteins were eluted with elution buffer (15% glycerol, 1% SDS, 50mM tris-HCl, 150mM NaCl pH 8.8) at 80C and subjected to western blot analysis.

### Nuclear speckle quantification by SC-35 Immunofluorescence with CellProfiler image analysis

Cells were plated on 35 mm glass bottom dishes with #1.5 cover glass (D35-14-1.5-N, In Vitro Scientific). Cells were fixed and stained with anti-SC-35 antibody (S4045, Sigma-Aldrich) at 1:1000 dilution and detected with Alexafluor488 secondary antibody at 1:1000 (Life Technologies). Nuclei were counterstained with Hoescht dye. Monochromatic images were captured under identical conditions and pseudo-colored using Photoshop. A custom image analysis pipeline was empirically adapted from a pre-existing pipeline designed for detecting H2AX foci using CellProfiler [30]. Measurements of nuclear speckles were generated from at least 15 random microscopic fields. A minimum of 100 nuclei identified by CellProfiler were used for quantitation per treatment.

### Library preparation and RNA-sequencing

Total RNA was extracted with the RNeasy mini extraction kit (Qiagen) and treated by on-column DNAse digestion. RNA quality was determined with a bioanalyzer

(Agilent) and samples with RIN values >7 were used for sequencing. mRNA were enriched with the NEBNext Poly(A) mRNA Magnetic Isolation Module (New England BioLabs, #E7490S). Library preparations for paired end sequencing were performed using the NEBNext Ultra Directional RNA Library Prep Kit for Illumina (New England BioLabs, #E7420S) according to manufacturer's specifications. Samples were pooled and 75bp paired reads were generated using a NextSeq 500 sequencer (Illumina). Approximately 50 million reads per sample were generated.

**RNA sequencing analysis**

FASTQ files were aligned using TOPHAT v1.4 with parameters "--mate-inner-dist 300 --mate-std-dev 500 --no-sort-bam --no-convert-bam -p 4". juncBase was used to identify read counts at splice junctions.

**Author Contributions**

Conceptualization, R.B., W.J.G, and B.R.P.; Methodology, W.J.G., B.R.P., R.R.; Investigation, W.J.G., B.R.P., L.M.U. R.L.; Resources, P.C., E.O., D.H.; Writing-Original Draft, W.J.G., B.R.P, R.B.; Supervision, R.B., R.R.

**Acknowledgements**

We would like to thank Yong Yu and members of the Reed lab for helpful discussion. This work was supported in part by NIH grants K08 CA122833 (R.B.), F32 CA180653 (B.R.P.), F30 CA192725 (W.J.G), and R01 GM043375 (R.R.), the Sontag Foundation (R.B.), Grey Matters (R.B.), the Friends for Life Fellowship (B.R.P.). R.B. is a consultant for Novartis Pharmaceuticals.

## References

1 Beroukhim, R. *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* **463**, 899-905, doi:10.1038/nature08822 (2010).

2 Bandopadhayay, P. *et al.* BET bromodomain inhibition of MYC-amplified medulloblastoma. *Clin Cancer Res* **20**, 912-925, doi:10.1158/1078-0432.CCR-13-2281 (2014).

3 Musgrove, E. A., Caldon, C. E., Barraclough, J., Stone, A. & Sutherland, R. L. Cyclin D as a therapeutic target in cancer. *Nat Rev Cancer* **11**, 558-572, doi:10.1038/nrc3090 (2011).

4 Nijhawan, D. *et al.* Cancer vulnerabilities unveiled by genomic loss. *Cell* **150**, 842-854, doi:10.1016/j.cell.2012.07.023 (2012).

5 Wahl, M. C., Will, C. L. & Luhrmann, R. The spliceosome: design principles of a dynamic RNP machine. *Cell* **136**, 701-718, doi:10.1016/j.cell.2009.02.009 (2009).

6 Kotake, Y. *et al.* Splicing factor SF3b as a target of the antitumor natural product pladienolide. *Nature chemical biology* **3**, 570-575, doi:10.1038/nchembio.2007.16 (2007).

7 Kaida, D. *et al.* Spliceostatin A targets SF3b and inhibits both splicing and nuclear retention of pre-mRNA. *Nature chemical biology* **3**, 576-583, doi:10.1038/nchembio.2007.18 (2007).

8 Webb, T. R., Joyner, A. S. & Potter, P. M. The development and application of small molecule modulators of SF3b as therapeutic agents for cancer. *Drug discovery today*, doi:10.1016/j.drudis.2012.07.013 (2012).

9 Harbour, J. W. *et al.* Recurrent mutations at codon 625 of the splicing factor SF3B1 in uveal melanoma. *Nat Genet* **45**, 133-135, doi:10.1038/ng.2523 (2013).

10 Wang, L. *et al.* SF3B1 and other novel cancer genes in chronic lymphocytic leukemia. *The New England journal of medicine* **365**, 2497-2506, doi:10.1056/NEJMoa1109016 (2011).

11 Imielinski, M. *et al.* Mapping the hallmarks of lung adenocarcinoma with massively parallel sequencing. *Cell* **150**, 1107-1120, doi:10.1016/j.cell.2012.08.029 (2012).

12 Yoshida, K. *et al.* Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64-69, doi:10.1038/nature10496 (2011).

13 Mizui, Y. *et al.* Pladienolides, new substances from culture of Streptomyces platensis Mer-11107. III. In vitro and in vivo antitumor activities. *The Journal of antibiotics* **57**, 188-196 (2004).

14 Cheung, H. W. *et al.* Systematic investigation of genetic vulnerabilities across cancer cell lines reveals lineage-specific dependencies in ovarian cancer. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 12372-12377, doi:10.1073/pnas.1109363108 (2011).

15 Barretina, J. *et al.* The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature* **483**, 603-607, doi:10.1038/nature11003 (2012).

16 Shao, D. D. *et al.* ATARiS: computational quantification of gene suppression phenotypes from multisample RNAi screens. *Genome research* **23**, 665-678, doi:10.1101/gr.143586.112 (2013).

17 Zack, T. I. *et al.* Pan-cancer patterns of somatic copy number alteration. *Nature genetics* **45**, 1134-1140, doi:10.1038/ng.2760 (2013).

18 Wang, C. *et al.* Phosphorylation of spliceosomal protein SAP 155 coupled with splicing catalysis. *Genes Dev* **12**, 1409-1414 (1998).

19 An, M. & Henion, P. D. The zebrafish sf3b1b460 mutant reveals differential requirements for the sf3b1 pre-mRNA processing gene during neural crest development. *The International journal of developmental biology* **56**, 223-237, doi:10.1387/ijdb.113383ma (2012).

20 Isono, K., Mizutani-Koseki, Y., Komori, T., Schmidt-Zachmann, M. S. & Koseki, H. Mammalian polycomb-mediated repression of Hox genes requires the essential spliceosomal protein Sf3b1. *Genes Dev* **19**, 536-541, doi:10.1101/gad.1284605 (2005).

21 Network, T. C. G. A. Comprehensive molecular portraits of human breast tumours. *Nature* **490**, 61-70, doi:10.1038/nature11412 (2012).

22 Kramer, A., Gruter, P., Groning, K. & Kastner, B. Combined biochemical and electron microscopic analyses reveal the architecture of the mammalian U2 snRNP. *J Cell Biol* **145**, 1355-1368 (1999).

23 van der Feltz, C., Anthony, K., Brilot, A. & Pomeranz Krummel, D. A. Architecture of the spliceosome. *Biochemistry* **51**, 3321-3333, doi:10.1021/bi201215r (2012).

24 Luo, M. J. & Reed, R. Identification of RNA binding proteins by UV cross-linking. *Current protocols in molecular biology / edited by Frederick M. Ausubel ... [et al.]* **Chapter 27**, Unit 27 22, doi:10.1002/0471142727.mb2702s63 (2003).

25 Gozani, O., Potashkin, J. & Reed, R. A potential role for U2AF-SAP 155 interactions in recruiting U2 snRNP to the branch site. *Mol Cell Biol* **18**, 4752-4760 (1998).

26 Zhou, Z., Licklider, L. J., Gygi, S. P. & Reed, R. Comprehensive proteomic analysis of the human spliceosome. *Nature* **419**, 182-185, doi:10.1038/nature01031 (2002).

27 Brooks, A. N. *et al.* Conservation of an RNA regulatory map between Drosophila and mammals. *Genome Res* **21**, 193-202, doi:10.1101/gr.108662.110 (2011).

28 Moore, M. J., Wang, Q., Kennedy, C. J. & Silver, P. A. An alternative splicing network links cell-cycle control to apoptosis. *Cell* **142**, 625-636, doi:10.1016/j.cell.2010.07.019 (2010).

29 Spector, D. L. & Lamond, A. I. Nuclear speckles. *Cold Spring Harb Perspect Biol* **3**, doi:10.1101/cshperspect.a000646 (2011).

30 Kamentsky, L. *et al.* Improved structure, function and compatibility for CellProfiler: modular high-throughput image analysis software. *Bioinformatics* **27**, 1179-1180, doi:10.1093/bioinformatics/btr095 (2011).

31 Loyer, P., Trembley, J. H., Lahti, J. M. & Kidd, V. J. The RNP protein, RNPS1, associates with specific isoforms of the p34cdc2-related PITSLRE protein kinase in vivo. *Journal of cell science* **111 ( Pt 11)**, 1495-1506 (1998).

32 Berg, M. G. *et al.* A quantitative high-throughput in vitro splicing assay identifies inhibitors of spliceosome catalysis. *Mol Cell Biol* **32**, 1271-1283, doi:10.1128/MCB.05788-11 (2012).

33 Zhou, Q. *et al.* A chemical genetics approach for the functional assessment of novel cancer genes. *Cancer Res*, doi:10.1158/0008-5472.CAN-14-2930 (2015).

34 Folco, E. G., Coil, K. E. & Reed, R. The anti-tumor drug E7107 reveals an essential role for SF3b in remodeling U2 snRNP to expose the branch point-binding region. *Genes Dev* **25**, 440-444, doi:10.1101/gad.2009411 (2011).

35 Schneider, R. K. *et al.* Role of casein kinase 1A1 in the biology and targeted therapy of del(5q) MDS. *Cancer Cell* **26**, 509-520, doi:10.1016/j.ccr.2014.08.001 (2014).

36 Liu, Y. *et al.* TP53 loss creates therapeutic vulnerability in colorectal cancer. *Nature* **520**, 697-701, doi:10.1038/nature14418 (2015).

37 Sundaramoorthy, S., Vazquez-Novelle, M. D., Lekomtsev, S., Howell, M. & Petronczki, M. Functional genomics identifies a requirement of pre-mRNA splicing factors for sister chromatid cohesion. *EMBO J* **33**, 2623-2642, doi:10.15252/embj.201488244 (2014).

38 Massiello, A., Roesser, J. R. & Chalfant, C. E. SAP155 Binds to ceramide-responsive RNA cis-element 1 and regulates the alternative 5' splice site selection of Bcl-x pre-mRNA. *FASEB J* **20**, 1680-1682, doi:10.1096/fj.05-5021fje (2006).

39 Schwerk, C. & Schulze-Osthoff, K. Regulation of apoptosis by alternative pre-mRNA splicing. *Mol Cell* **19**, 1-13, doi:10.1016/j.molcel.2005.05.026 (2005).

40 Cerami, E. *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer discovery* **2**, 401-404, doi:10.1158/2159-8290.CD-12-0095 (2012).

41 Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* **6**, pl1, doi:10.1126/scisignal.2004088 (2013).

42 Klaus Hartmuth, M. A. v. S. a. R. L. *Ultracentrifugation in the Analysis and Purification of Spliceosomes Assembled In Vitro*. (Wiley, 2012).

# Chapter 4: Perspectives and Future Directions

## Therapeutic implications of intratumoral heterogeneity

An important goal of cancer research is to understand mechanisms of cancer evolution and in particular, metastasis. We performed exome sequencing of endometrial cancers to assess the extent of intratumoral heterogeneity and to catalogue the alterations in metastases. We found that mutations in certain cancer driver genes such as ARID1A were frequently heterogeneous between biopsies, suggesting targeted therapies aimed at these alterations may be less effective.

We have argued that therapies that attempt to reverse the effects of mutations that are not shared by all cells in the tumor are likely to have limited efficacy. Two ideas contribute to the intuition behind this argument: (i) the fitness of cancer cells that do not harbor mutations in these genes is unlikely to be affected by a drug that reverses the effect of the mutation in question; and (ii) cancer cells that do harbor a driver gene mutation that is heterogeneous may not display functional dependence on that alteration, as their ancestors were able to undergo malignant proliferation before acquisition of the alteration. However, one can imagine conditions under which these assumptions are not true, and the presence of a tumor subclone may be essential to tumor growth. For example, a subpopulation of cancer cells may provide non-cell-autonomous fitness advantage to neighboring tumor cells. A recent study assessed the growth of breast cancer xenografts with different proportions of 18 genetically engineered subclonal populations[1]. The tumors grew more rapidly and were more metastatic when a small proportion of the tumor was composed of a subclone that secreted defined growth factors or cytokines, such as IL-11 or CCL5. When subclones expressing other soluble factors were mixed with the parental cell line, the tumors had different histologies as well[1]. These findings suggest that therapeutics targeting a subclonal population may be effective in slowing disease progression.

Therapeutic targeting of heterogeneous mutations may be particularly effective in hematopoietic cancers due to their unique population dynamics. Cancer cells must compete with one another for resources in a nutrient-limited environment, and in hematopoietic cancers, the population is well-mixed. Therefore, a single clone with greater fitness can perform relatively rapid clonal sweeps[2]. One can imagine that the proportion of a leukemia cells representing a subclone harboring a druggable driver gene mutation could increase over time. These cancer cells may then assume some degree of oncogene-addiction to the new alteration, rendering susceptible to pharmacologic inhibition of said alteration. In such a case, treatment with a targeted therapy could kill off a substantial portion of the tumor cells and allow the patient sufficient time to undergo bone marrow transplantation. Indeed, case reports depicting this scenario have been presented.

### Implications for metastasis-specific driver genes

One could imagine that mutation of certain genes involved in cell migration or differentiation could promote metastasis. We searched for evidence of such mutations in endometrial tumors, but were unable to find any genes with sufficient statistical evidence of positive selection by mutation. It is possible that these mutations occur, but that we lacked statistical power to detect mutations associated with metastasis. There may be many such genes that are recurrently mutated in metastases, but the individual frequency of mutation of each gene is so low that very large number of primary-metastasis pairs would need to be sequenced to detect any statistically significant metastasis-specific driver genes. It is also possible that phenotypic programs, such as the epithelial-mesenchymal transition, are the predominant means to metastatic potential. A frequently occurring metastasis-specific driver mutation would be useful as in shedding light on the functional bottlenecks to metastasis. With precise information on the biological functions required for metastatic cells, one could in principle design drugs that inhibit some or all of these activities. If there are no frequently occurring metastasis-specific mutations, the possibility of developing drugs to inhibit the metastatic phenotype seems further away.

**The future of CYCLOPS genes**

In the absence of metastasis-specific alterations, the best way to treat metastases is to find new ways to treat cancers as a whole. One way to identify new therapeutic vulnerabilities is to search for vulnerabilities associated with patterns of somatic copy number alteration in cancer genomes. We identified a class of interaction called CYCLOPS vulnerabilities in which copy loss of a gene sensitizes cells to its further suppression.

Since we began work on this project, several other groups have identified other CYCLOPS vulnerabilities. A recent study identified *POLR2A* as a CYCLOPS gene[3]. *POLR2A* lies on chromosome 17p in close proximity to *TP53*. Thus, *POLR2A* is frequently co-deleted with *TP53* loss. The authors showed that colorectal cell lines with loss of *POLR2A* were more sensitive to *POLR2A* inhibition than cell lines that had two copies of the gene. Importantly, they showed that alpha-amanitin, a small molecule inhibitor of *POLR2A* was able to reproduce the vulnerability demonstrated by RNA interference. *POLR2A* was a statistically significant CYCLOPS gene in our survey, but it was not included as a focus of our research effortss.

A second study has validated a CYCLOPs gene in the commonly deleted region of 5q implicated in 5q-del myelodysplastic syndrome[4]. Copy loss of casein kinase, encoded by *CSNK1A1*, sensitizes cells to further suppression of *CSNK1A1*. In this case as well, the authors were able to identify a small compound, lenalidomide, that was able to degrade casein kinase and specifically inhibit the growth cells with hemizygous deletion of the gene. We did not identify *CSNK1A1* as a CYCLOPS gene in our analysis, most likely because none of the cell lines we analyzed harbored hemizygous deletions of *CSNK1A1*. However, we found that amplifications of *CSNK1A1* were associated with greater resistance to its suppression, consistent with features of a CYCLOPS gene.

Our failure to find *CSNK1A1* highlights the limited sensitivity of our analysis to find CYCLOPS genes. Our analysis of CYCLOPS genes was based on the copy number profiles and gene dependency scores of 179 cell lines. The current release of Achilles data includes 384 cell lines, which should give us much more power to detect CYCLOPS vulnerabilities. In addition, our prior analysis used data from a 55K array of shRNAs, which has since been expanded to 100K shRNAs. The current release of gene solutions includes 21,901 gene solutions compared to the previous 9,047. Therefore, we are now able to analyze many more genes for evidence of CYCLOPS vulnerability than before. At this sample size, it is possible that we will now be able to identify lineage-specific CYCLOPS vulnerabilities.

Future experiments should include determining whether the inhibition of multiple CYCLOPS genes on the same hemizygously deleted chromosome leads to additive or synergistic effects. A major goal of future research will be to translate the observations of these vulnerabilities based on gene suppression into small molecules. The finding that certain thalidomide derivatives can modulate ubiquitin ligase activity to degrade proteins is an exciting lead. Perhaps similar approaches can be adapted to induce the therapeutic degradation of CYCLOPS genes. Other exciting new pharmacologic strategies, such as protacs, provide a general framework for the creation of compounds that can degrade any protein to which a small molecule can bind[5]. Alternatively, advances in the delivery of nucleic acids could allow for RNA interference to be used directly. These are just some of the many therapeutic possibilities that could be used to target CYCLOPS vulnerabilities.

**Potential issues with CYCLOPS genes**

There are two main concerns with using CYCLOPS vulnerabilities to treat cancers. First, the therapeutic window created by these vulnerabilities may be small. Second, many CYCLOPS genes may not represent readily druggable targets.

One concern about CYCLOPS vulnerabilities is that the therapeutic window they create may be too small for drugs to be effective. Without feedback mechanisms, hemizgous loss of a CYCLOPS gene would result in a two-fold reduction in the amount of protein present. Such a reduction in the amount of protein present may not provide an adequate therapeutic window between normal cells and cancer cells. One reason that a half-reduction in protein may not be adequate, is that different tissues may display different requirements for the CYCLOPS gene being targeted. Indeed, certain tissue types may normally express the CYCLOPS gene at similar or even lower concentrations than cancer cells harboring hemizygous deletion of the gene. Certain tissues may express normal concentrations of CYCLOPS genes, but may be more dependent on the function of these genes for cell survival than cancer cells originating from other tissues are. Indeed, some Mendelian diseases associated with mutations in one allele of CYCLOPS genes display phenotypes specific to certain tissues[6]. Therefore, tissue-specific sensitivities to perturbation of CYCLOPS gene function may dominate the therapeutic window afforded by these vulnerabilities and limit the doses of therapeutic compounds that can be safely administered to patients. Conversely, lineage-based expression differences could be used to our advantage if the cancer comes from a lineage with low expression of a CYCLOPS gene. The lineage-controlled permutations we performed were able to minimize the probability of identifying tissue-specific vulnerabilities that are frequently deleted in the tissue in question. However, lineage-controlled permutations may nominate CYCLOPS candidates where different tissues display different dependencies on the gene in question. The current data that we have do not support the concern about variation in the dependence of different tissues on CYCLOPS gene function, as most CYCLOPS genes were statistically significant in both lineage-controlled and non-lineage-controlled permutation tests.

Another concern regarding CYCLOPS vulnerabilities is that few of these genes represent readily druggable targets. A large proportion of CYCLOPS genes tend to be members of macromolecular complexes. As such, many CYCLOPS genes play predominantly structural roles rather than enzymatic roles. For example, in the case of

SF3B1, the relative expression of the fully assembled spliceosome is unaffected in SF3B1$^{loss}$ cells. Therefore, small molecules that inhibit the enzymatic action of the fully assembled spliceosome will not strike at the relative vulnerability between diploid cells and cancer cells with loss of SF3B1. Optimally, we would devise a strategy that targets the relative depletion of SF3B1 in subcomplexes. Two possible strategies include degrading the protein encoded by the CYCLOPS gene product, or preventing the protein from associating assembling with other complex members. Unfortunately, current technology does not allow for the reliable production of small molecules to induce degradation or inhibit protein-protein interactions[7]. Therefore, in many cases, it is unlikely that small molecules will be able to reproduce the vulnerability unveiled by gene suppression.

## Promise of allele-specific vulnerabilities in cancer genomes

Many cancers inactivate tumor suppressor genes through either copy-neutral loss of heterozygosity (LOH) or uncompensated copy loss, both of which produce allelic loss. One strategy to target this distinction would be to create allele-specific inhibitors against essential genes that harbor frequent polymorphisms in humans. The main advantage of allele-specific inhibitors is that the therapeutic window is categorical, not quantitative as in CYCLOPS vulnerabilities.

One could imagine that coding differences between two germline alleles of essential genes could cause differential binding to a small molecule inhibitor. We have been performing a search for these variations across human populations. The search involves several steps, which are described below.

First, we must define essential genes. We have used a variety of data sources as evidence of cell-essentiality for human genes. For instance, we are using data from CRISPR screens to find genes that severely impact cellular fitness when inactivated. We are also examining disruptions observed in cancer genomes. If a gene is biallelically inactivated in any cancer genome, either through inactivating mutation or gene deletion, the gene cannot be essential in human cells. Similarly, we can leverage sequencing data from populations with a degree of inbreeding, such as Iceland[8], to attempt to identify genes with homozygous mutation in healthy individuals. Orthogonal evidence comes from shRNA screens and the literature. We have used machine learning tools trained on these features to identify approximately 2,000 high confidence essential genes in humans.

A second task was to map human variation onto these essential genes. We have used the Exome Aggregation Consortium (ExAC) to find common variation in these essential genes. Next, we had to prioritize candidates based on the probability that we can identify a variant that might represent a druggable difference between two alleles. An optimal candidate would occur in a region of an essential protein that is close to the active site, and would have pre-identified small molecules that bind to the region in question. We used databases of catalytic active sites in proteins, structural database and pharmacology databases to prioritize our candidates. We also manually reviewed the structures of these candidates to identify hydrophobic pockets into which a small molecule could bind.

In the future, we will attempt to find small molecules that distinguish between two alleles of an essential gene. A potential method would be to screen small molecule libraries against recombinant protein from the essential gene in question. A counterscreen against the same essential gene with the variant allele could also be performed. DNA encoded libraries could be particularly useful for this[9]. One could also generate isogenic cell lines with CRISPR technology whose only difference is the variant in question. In-vitro screens for cell viability could then be performed with candidate small molecules as well.

## Conclusion

We have performed a sequencing analysis of endometrial cancer metastasis. We showed that certain driver genes such as *ARID1A* tend to be heterogenous in these tumors and that metastases tend to be more related to one another than would be expected by chance. Despite the latter finding, we were unable to identify and genes that were significantly recurrently mutated specifically in metastases. These data shed light on the biology of metastasis and have important implications for the design of therapeutics.

We have also performed an analysis of cancer vulnerabilities associated with SCNAs. CYCLOPS vulnerabilities are the most enriched class of vulnerabilities associated with SCNAs. We validated this vulnerability for a core splicing factor, *SF3B1*. A reservoir of *SF3B1* protein protected *SF3B1*$^{neutral}$ cells from the effects of SF3B1 suppression. Future studies will focus on translating these vulnerabilities into effective therapeutics. We hope that these studies will inspire future progress on synthetic lethal opportunities in cancer therapy.

## References

1 Marusyk, A. *et al.* Non-cell-autonomous driving of tumour growth supports sub-clonal heterogeneity. *Nature* **514**, 54-58, doi:10.1038/nature13556 (2014).

2 Landau, D. A. *et al.* Evolution and impact of subclonal mutations in chronic lymphocytic leukemia. *Cell* **152**, 714-726, doi:10.1016/j.cell.2013.01.019 (2013).

3 Liu, Y. *et al.* TP53 loss creates therapeutic vulnerability in colorectal cancer. *Nature* **520**, 697-701, doi:10.1038/nature14418 (2015).

4 Kronke, J. *et al.* Lenalidomide induces ubiquitination and degradation of CK1alpha in del(5q) MDS. *Nature* **523**, 183-188, doi:10.1038/nature14610 (2015).

5 Sakamoto, K. M. Protacs for treatment of cancer. *Pediatric research* **67**, 505-508, doi:10.1203/PDR.0b013e3181d35017 (2010).

6 Wieland, I. *et al.* Refinement of the deletion in 7q21.3 associated with split hand/foot malformation type 1 and Mondini dysplasia. *Journal of medical genetics* **41**, e54 (2004).

7 Arkin, M. R. & Wells, J. A. Small-molecule inhibitors of protein-protein interactions: progressing towards the dream. *Nature reviews. Drug discovery* **3**, 301-317, doi:10.1038/nrd1343 (2004).

8 Gudbjartsson, D. F. *et al.* Large-scale whole-genome sequencing of the Icelandic population. *Nature genetics* **47**, 435-444, doi:10.1038/ng.3247 (2015).

9 Mannocci, L. *et al.* High-throughput sequencing allows the identification of binding molecules isolated from DNA-encoded chemical libraries. *Proceedings of the National Academy of Sciences of the United States of America* **105**, 17670-17675, doi:10.1073/pnas.0805130105 (2008).