Characterization of Precancerous Mutations in Ulcerative Colitis-Associated Colorectal Cancer and High-Grade Serous Ovarian Cancer via Duplex Sequencing

Kathryn Terese Baker

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2018

Reading Committee:

Rosa Ana Risques, Chair

Raymond J. Monnat Jr.

Teresa A. Brentall

Program Authorized to Offer Degree:

Pathology

University of Washington

Abstract

Characterization of Precancerous Mutations in Ulcerative Colitis-Associated Colorectal Cancer and High-Grade Serous Ovarian Cancer via Duplex Sequencing

Kathryn Terese Baker

Chair of the Supervisory Committee:
Assistant Professor Rosa Ana Risques, Ph.D.
Department of Pathology

Humans have attempted to fight cancer since the beginning of recorded history, but advances in the early detection of cancer have begun to emerge only in the last several decades. Patient survival increases with earlier detection and cancer incidence decreases with efficacious screening tests, such as colonoscopy or the Pap test. Early detection has been difficult for many cancer types, however, because of non-unique symptoms, inaccessibility of certain tissues for biopsy, and false-positives from tests that lack acceptable levels of specificity. One method to overcome these issues is to analyze the genetic mutations of precancerous tissue. Cancer is considered a genetic disease, and molecular alterations may be specific to each cancer type. Despite the great utility of next-generation sequencing technology, many techniques are still plagued by significant error rates. Error rates range from 1/100 to 1/1000 for standard NGS, which precludes the identification of rare or very low frequency events that may signal early

tumorigenic processes. Ultra-accurate technology is necessary to confidently call these mutations. Duplex sequencing is a next-generation sequencing (NGS) method with unprecedented accuracy and sensitivity, with a detection rate of $1/10^7$ reads. Our use of this technology facilitated several applications for characterizing precancer and early cancer detection. The first is in the characterization of mitochondrial DNA mutations in colorectal cancer associated with ulcerative colitis, a preneoplastic inflammatory bowel disease that increases patient risk for tumorigenesis. Using DS to comprehensively catalogue these mutations, we found that mitochondrial DNA mutations increase in frequency and pathogenicity and appear to be positively selected in early dysplasia, but are removed in late dysplasia in cancer, implying that functional mitochondria are necessary for tumorigenic progression. Additionally, we posit that the presence of clonally expanded fields may serve as a predictive biomarker in this disease. The second application presented is for the early detection of *TP53* mutations in high-grade serous ovarian cancer, a disease for which there is currently not a reliable biomarker. Through the analysis of DNA collected from uterine lavage we were able to identify *TP53* tumor mutations in 80% of HGSOC patients. We also found a low level of background *TP53* mutations in all patients, including healthy controls that increased with patient age. These results demonstrate that careful calibration based on the accumulation of somatic mutations with age is necessary for an accurate biomarker. Finally, with the important modification of using CRISPR/Cas9 digestion as a target enrichment tool, CRISPR-DS is able to leverage the same accuracy of DS with the ability to use as little as 10ng of input DNA, opening the technology to further clinical applicability, as well as to use in any setting that requires low amounts of starting genetic material.

Dedication

For Danny Boncheff, without whose generosity this work would not have been possible

Acknowledgements

This work was supported by the following research grants:

JJS: NIH R01CA160674, T32HL007093, T32CA009515

JJS and LNW: NIH R44CA221426

KTB: NIH T32GM95421

LAL: NIH CA077852, CA193649

RAR: NIH R01CA181308, Mary Kay Foundation Grant 045-15, Rivkin Center for Ovarian Cancer Grant 567612

SRS: Cooperative Agreement Number W911NF-15-2-0127 from the Department of Defense Army Research Office/Defense Forensic Science Center (DFSC), W81XWH-16-1-0579 from the Department of Defense Congressionally Directed Medical Research Program

TAB: NIH R01CA160674

TS: Radiumhemmets Forskningsfonder 174261

This work was overseen by my graduate committee including: Su-in Lee; Marshall Horwitz; Teresa Brentall, whose clinical expertise was invaluable; and Raymond J. Monnat Jr., who has kindly and patiently offered his advice and knowledge since the first day of my graduate career. Additional thanks to William Mahoney, Scott R. Kennedy, Alan Herr, and Steve Berard whose assistance and mentorship were critical to my development as a scientist.

The work presented here would not have been possible without the help of my lab mates: Hye Son Yi, Jeffrey Krimmel, Daniela Nachmanson, Yan Liu, Shenyi Lian, Yuezheng Zhang, Jeanne Fredrickson, Jake Hoekstra, Elizabeth K. Schmidt, and Michael J. Hipp.

For their love and endless support, I thank my family: Ann and Tom Tamoria, Tom Baker and Debbie Baker, Wesley Baker, Clayton Baker, and Spencer Baker. Many thanks to my friends, without whom I wouldn't have had the stamina to finish this journey, especially my

fellow M3D/MBD students, my Zetas, Kelly and Danny Novet, Jennifer Alvarez, Kristi-Elize and Tre Charles, Kevin and Matt Clutario, Nick Perry, Maxine Reyna, and Brian Knox.

Finally, I am eternally grateful to Rosa Ana Risques, whose intelligence, patience, and enthusiasm I can only hope to emulate as a researcher and mentor in the future. You have pushed me to improve as a scientist and a person, and have believed in me even when I didn't. It has been an honor to be your first student.

Table of Contents

List of Figures

Page

Chapter 1

Chapter 2

Chapter 3

List of Tables

Abbreviations

BE – Barrett's esophagus

CE - chromoendoscopy

COX – cytochrome C oxidase subunit I

CRC – colorectal cancer

CRISPR-DS – CRISPR Duplex Sequencing

ddPCR – digital droplet PCR
DS – Duplex Sequencing
DCS – duplex consensus sequence
FFPE – formalin-fixed paraffin-embedded
GEE – general estimating equations
HGD – high-grade dysplasia
HGSOC – high-grade serous ovarian cancer
IBD – inflammatory bowel disease
IHC – immunohistochemistry
LGD – low-grade dysplasia
MAF – mutant allele frequency
mtDNA – mitochondrial DNA
NGS – Next-generation sequencing
NP – Non-Progressor
P – Progressor
PolyG – polyguanine mononucleotide repeat tract
RRSO – risk-reducing salpingo-oopherectomy
SSCS – single-strand consensus sequence
STIC – serous tubal intraepithelial carcinoma
TCGA – The Cancer Genome Atlas
UC – ulcerative colitis
UC-CRC – ulcerative colitis associated-colorectal cancer
VLF – very low frequency

Introduction

Early cancer detection

Cancer has been a recognized human ailment since the beginning of recorded history (1). From the early days of rudimentary surgical intervention (1,2), to the development of the first chemotherapy treatments in the 1940s (1,3), to the concerted efforts begun by The War on

Cancer in 1971 (4), many resources have been used to characterize and fight the disease. Advances in early detection and screening, or testing asymptomatic people, however, did not begin in earnest until the 1960s through the widespread use of the Papanicolaou test for cervical cancer and mammography for the detection of breast cancer (1). Indeed, until very recently, the NCI spent less than half the amount allocated to treatment on early detection efforts (5). Despite this, the utility of these strategies for early detection is quite evident (6), as the rate of survival of patients with cancer found at early stages is significantly higher than in those found with late stage or metastatic disease. Notable successful screening tests include the use of colonoscopy, which decreased the total incidence of colon cancer by ~20% since 1985 (7) , and the pap test, which has decreased the number of cervical cancer-related deaths by 70% (5). Projections for 5-year survival rates if tumors are detected while still confined to the organ of origin are up to 30% higher than current survival rates (6). Thus, early detection is increasingly recognized as crucial for improved patient survival.

Approaches to early cancer detection: from conventional tests to novel sequencing

Development of early detection tests has required the characterization of cancer at multiple biological levels. The detection of cancer-associated proteins in blood enabled some of the earlier tests, including prostate-specific antigen for prostate cancer and CA-125 for ovarian cancer (8). Radiographic imaging, including mammography and colonoscopy, relies on identification of gross morphological changes and has also become part of routine population screening strategies (8). These approaches face multiple major limitations, most importantly reduced specificity as not only cancers are detected but also benign, non-malignant lesions. For example, CA-125, the blood protein marker used for ovarian cancer diagnosis has an unacceptably high false-positive rate, which leads to unnecessary surgical procedures on healthy

patients (6). One potential way to increase specificity is to identify cancer-causing mutations. Cancer is a genetic disease. In other words, cancer is caused by DNA mutations that lead to a myriad of altered cellular processes that contribute to neoplasia and immortality. While some hereditary mutations, such as the *BRCA* susceptibility genes in breast and ovarian cancer (9) and mismatch repair genes in Lynch syndrome (10), predispose to neoplastic progression, the formation of a tumor relies on multiple somatic mutations, which are acquired through life and selected for their malignant properties following a process of Darwinian evolution (11,12).

From the perspective of early cancer detection, it is important to realize that cancerous mutations are often found in the normal-appearing tissue surrounding tumors, a phenomenon known as the field effect (13-17). Field effects precede the development of cancer and, therefore, the identification of these mutations in otherwise histologically normal tissue may be used as a biomarker for cancer progression (14). The study of preneoplastic, or precancerous diseases has proven extremely useful to characterize the field effect. These conditions feature lesions of abnormal cells with tumorigenic potential that predispose patients to cancer development. One of the most studied of these diseases is ulcerative colitis (UC), an inflammatory bowel disease (IBD) that predisposes patients to colorectal cancer (18). UC features intermediate phases of dysplasia and field effects that can be exploited to understand the molecular changes that occur during tumor progression (19). Our group has reviewed precancer in ulcerative colitis Chapter 1. A further example of a preneoplastic disease useful for studying the early stages of carcinogenesis is Barrett's esophagus, in which acid reflux from the stomach into the esophagus creates fields of abnormal cells featuring precancerous *TP53* mutations (20,21) and increases the risk esophageal adenocarcinoma (20-22).

While massive efforts have been made to understand the complexities of cancer genetics, such as The Cancer Genome Atlas (TCGA), much less is known about the genetics of precancer

(23). One major hurdle to developing a comparable understanding of precancer was overcoming technical limitations that prevented accurate detection of low abundance mutations that might signal early tumorigenic processes (24). While challenging, the characterization of these initial genetic alterations is essential in order to develop useful biomarkers for early detection.

In the last several decades, significant improvements have been made to high-throughput sequencing of nucleic acids with the development of next-generation sequencing (NGS) technology (24-28) One such next-generation sequencing technology developed at the University of Washington is a molecular barcoding method called Duplex Sequencing (DS)(29,30). DS relies on the use of double-stranded sequencing adapters, which enable identification not only of the parent DNA molecule, but which strand of that molecule a particular read is generated from. All reads from the same strand of a given molecule are computationally collapsed into a single-stranded consensus sequence (SSCS), which eliminates any sequencing errors. The two SSCS for each molecule are then condensed into a duplex consensus sequence (DCS), eliminating any PCR errors or mutations found on only one strand or the other. Only mutations found on both strands of the DCS are considered true mutations. While the error rate of standard NGS is between 1/100 and 1/1000, the theoretical error rate of DS is approximately $1/10^7$ (29-31). This extreme accuracy enables unprecedented sensitivity for the detection of low frequency mutations. Our group has demonstrated that DS is able to detect one cancer DNA mutation amongst 24,000 normal genomes (32).

The main objective of my research is to leverage this high sensitivity to advance the field of early cancer detection on two complementary fronts: 1) improve our understanding of the genetic alterations that take place in precancer by examining ulcerative colitis, a preneoplastic disease that serves as an excellent model of stepwise progression to cancer and that has already proven extremely useful to elucidate early genetic alterations that drive tumorigenesis; and 2)

develop highly sensitive biomarkers for the detection of early cancer via focusing on ovarian cancer, a deadly disease with an urgent unmet need for better diagnostic and predictive biomarkers.

Mitochondrial DNA mutations in Ulcerative Colitis

DS enables comprehensive characterization of low frequency mutations at a level never before possible. Thus, we took advantage of this technology to determine the role of mitochondrial DNA (mtDNA) mutations in ulcerative colitis tumorigenesis with the short-term goal of understanding their contribution to cancer and the long-term goal of using these mutations as a predictive biomarker of cancer progression in ulcerative colitis.

While the role of the nuclear genome in carcinogenesis has been extensively characterized at a genetic and epigenetic level, the contribution of mitochondrial DNA is less well understood and remains a controversy in the cancer field. It was posited that reactive oxygen species generated by oxidative phosphorylation causes damage to the histone-free mitochondrial genome, a double-stranded, circular stretch of genetic material separated from the nuclear genome (33). This damage, it was suggested, was the source of altered metabolism in cancer (34), leading to the Warburg effect, the name for the phenomenon in which cancers rely primarily on glycolysis for their energetic needs (33). Thus, mitochondrial dysfunction was thought to contribute to carcinogenesis (35-37). More recently, however, it has been shown that cancers do indeed rely on functional mitochondria (38). Validation of these theories proved challenging due to the error prone nature of most standard next generation sequencing technologies (31,39) and the complex biology of the mitochondrial genome, as there are multiple mitochondria per cell and multiple mitochondrial genomes per mitochondria. Because of the highly accurate detection rate of DS, however, our group has been able to effectively

characterize mitochondrial DNA mutations in several settings known to feature dysfunctional mitochondria (39,40).

With this advanced sequencing technology, our laboratory is able to interrogate the complex question of the role of mitochondrial DNA mutations in preneoplastic disease. We have long studied ulcerative colitis, which we suggested to be a disease of early colon aging based on its accelerated shortening of telomeres in colon epithelial cells (41). A distinctive feature of the aging colon is a loss of mitochondrial function and an increase in mitochondrial DNA mutations (39,42-44). Concordantly, mitochondrial dysfunction has also been previous identified in IBD (45). In order to characterize the role of mitochondria in UC carcinogenesis, we previously examined mitochondrial function in UC patients via immunohistochemistry (IHC) for proteins of the electron transport chain and quantified mtDNA copy number (46). We found that mitochondrial function and genomic copy number decline in early dysplasia and in the fields surrounding UC tumors, but both increase in later dysplasia, indicating that functional mitochondria might be required for malignant transformation. My goal was to use DS to determine whether this bimodal pattern of mitochondrial function is reflected in an inverse pattern of mutation accumulation; in other words, do mitochondrial DNA mutations accumulate in early dysplasia but decrease in number and frequency in high-grade dysplasia and cancer? This work is described here in Chapter 2.

Early detection of high-grade serous ovarian cancer

As more and more sensitive technology that accurately detects low frequency mutations is developed, the opportunity to create less invasive tests for malignancy increases. Previously, physicians would take biopsies of tissues suspected to be diseased based on blood tests or imaging (8). While biopsies could provide evidence of premalignant tissue or frank cancer, tissue

samples may be difficult to procure, either because the tissue is in an inaccessible portion of the body or because it is from a vital organ (47). A further complication arises for cancers that may arise and grow without detectable symptoms. These issues have popularized the idea of liquid biopsies, which is based on the fact that tumors shed whole cells and fragmented DNA into the bloodstream as well as surrounding tissues. Thus, easy-to-access samples such as blood and urine could be tested for the presence of cancer mutations in order to develop minimally invasive, early detection methods (48).

One disease in which better early detection methods are urgently needed is high-grade serous ovarian cancer (HGSOC). HGSOC is the most common type of ovarian cancer and has a dismal 5-year survival rate of 25%, which is due in large part to the difficulty of detecting the disease at early stages (49). By the time patients present with symptoms such as bloating or abdominal pain, the tumor has already metastasized (49). The putative initiating event for HGSOC is the development of foci of p53 mutations in the distal fallopian tube, which develop into serous tubal intraepithelial carcinoma (STIC), which then seeds into the ovarian epithelium (50-52). Over 96% of affected patients have *TP53* mutations (53,54) making these mutations an attractive target for the development early detection methods in HGSOC.

Previous work by our lab analyzed peritoneal fluid taken from the abdominal cavities of patients undergoing gynecological surgery for suspected masses or salpingo-oopherectomy as a risk-reducing prophylactic step in women with inherited high risk of ovarian cancer (32). Using DS, we were able to detect the cancer-driver *TP53* mutation in 94% of patients with HGSOC. Of note, very low frequency *TP53* mutations were found in both the peritoneal fluid and peripheral blood of cancer patients as well as control patients. These mutations increased with age, pointing toward an age-related level of biological background mutations (32). In spite of these background mutations, cancers and controls could be distinguished with 82% sensitivity and

90% specificity. With this success as a proof-of-principle, we further attempted to analyze Pap smear material, as pap smears are non-invasive and are already in use as a screening technique for cervical cancer. These attempts, however, did not yield a sensitivity higher than ~30% (unpublished results). Efforts made by other labs have also faced significant problems, with peak sensitivity at ~40% (55,56). Development of an early detection test in HGSOC requires use of a different sample type.

One such sample that has been previously analyzed and optimized for this purpose is uterine lavage (57). Our collaborators at the Medical University of Vienna developed a novel uterine lavage method that employs a three-way catheter that is inserted into the cervical canal and flushes the uterine cavity with saline solution with the intention of collecting a larger number of cells from the Mullerian ducts (57). They were able to detect HGSOC with a sensitivity of 80% using three separate detection strategies including two NGS methods and digital droplet PCR (ddPCR) (57). In Chapter 3, we have applied DS to uterine lavage fluid in the hopes of increasing early detection with a more streamlined methodology.

CRISPR-DS: modifying DS for clinical applicability

While DS improves accuracy and sensitivity versus standard NGS platforms (24) , several hurdles remain for translating this technology into the clinic. Standard DS relies on DNA sonication, which generates randomly sized fragments, and oligo-nucleotide hybridization capture, which has low efficiency for small target regions. Together, these two issues limit the recovery rate of DS, which proves problematic in samples for early detection, which may have a very low prevalence of mutated DNA as well as low DNA input. In order to improve upon this issue, our group developed CRISPR-DS, a modified DS pipeline that utilizes CRISPR-Cas9 digestion as a method of target enrichment. CRISPR-DS allows for analysis of small genomic

regions with even coverage, eliminates errors generated by sonication or biases due to random fragmentation, all while using up to 100-fold less input DNA than standard DS (Chapter 4). This technology may be adapted for many known genes or regions of interest, and therefore has many potential applications for early cancer detection.

References

1. Sudhakar A. History of Cancer, Ancient and Modern Treatment Methods. J Cancer Sci Ther **2009**;1:1-4
2. Lawrence Jr W. History of Surgical Oncology. Surgery: SpringerLink; 2005.
3. DeVita Jr VT, Chu E. A History of Cancer Chemotherapy. Cancer research **2008**
4. Hanahan D. Rethinking the war on cancer. The Lancet **2014**;383:558-63
5. Spinney L. Cancer: Caught in time. Nature **2006**;442:736
6. Etzioni R, Urban N, Ramsey S, McIntosh M, Schwartz S, Reid B*, et al.* Early detection: The case for early detection. Nature Reviews Cancer **2003**;3:243
7. Horner MJ, Ries LAG, Krapcho M, Neyman N, Aminou R, Howlader N*, et al.* SEER Cancer Statistics Review 1975-2006 - Previous Version - SEER Cancer Statistics. **2009**
8. Schiffman JD, Fisher PG, Gibbs P. Early detection of cancer: past, present, and future. American Society of Clinical Oncology Educational Book: ASCO; 2015.
9. Hall J, Lee M, Newman B, Morrow J, Anderson L, Huey B*, et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. Science **1990**;250:1684-9
10. Lynch HT, Snyder CL, Shaw TG, Heinen CD, Hitchins MP. Milestones of Lynch syndrome: 1895–2015. Nature Reviews Cancer **2015**;15:181
11. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science **2015**;349:1483-9
12. Nowell PC. The clonal evolution of tumor cell populations. Science **1976**;194:23-8
13. Braakhuis BJ, Tabor MP, Kummer JA, Leemans CR, Brakenhoff RH. A genetic explanation of Slaughter's concept of field cancerization: evidence and clinical implications. Cancer Res **2003**;63:1727-30
14. Salk JJ, Horwitz MS. Passenger mutations as a marker of clonal cell lineages in emerging neoplasia. Semin Cancer Biol **2009**;20:294-303
15. Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc Natl Acad Sci U S A **2013**;110:1999-2004

16. Slaughter DP, Research, Educational H, the Tumor Clinic of the University of Illinois College of M, the Presbyterian Hospital CI, Southwick HW*, et al.* "Field cancerization" in oral stratified squamous epithelium. Clinical implications of multicentric origin. Cancer **2017**;6:963-8
17. Cloos J, de Boer WP, Snel MH, van den Ijssel P, Ylstra B, Leemans CR*, et al.* Microarray analysis of bleomycin-exposed lymphoblastoid cells for identifying cancer susceptibility genes. Mol Cancer Res **2006**;4:71-7
18. Choi CR, Bakir IA, Hart AL, Graham TA. Clonal evolution of colorectal cancer in IBD. Nature reviews Gastroenterology & hepatology **2017**
19. Yashiro M. Ulcerative colitis-associated colorectal cancer. World J Gastroenterol **2014**;20:16389-97
20. Prevo LJ, Sanchez CA, Galipeau PC, Reid BJ. p53-mutant clones and field effects in Barrett's esophagus. Cancer Res **1999**;59:4784-7
21. Reid BJ, Prevo LJ, Galipeau PC, Sanchez CA, Longton G, Levine DS*, et al.* Predictors of progression in Barrett's esophagus II: baseline 17p (p53) loss of heterozygosity identifies a patient subset at increased risk for neoplastic progression. Am J Gastroenterol **2001**;96:2839-48
22. Brabender J, Marjoram P, Lord RV, Metzger R, Salonga D, Vallbohmer D*, et al.* The molecular signature of normal squamous esophageal epithelium identifies the presence of a field effect and can discriminate between patients with Barrett's esophagus and patients with Barrett's-associated adenocarcinoma. Cancer Epidemiol Biomarkers Prev **2005**;14:2113-7
23. Spira A, Yurgelun MB, Alexandrov L, Rao A, Bejar R, Polyak K*, et al.* Precancer Atlas to Drive Precision Prevention Trials. Cancer Res **2017**;77:1510-41
24. Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat Rev Genet **2018**;19:269-85
25. Krasnitz A, Kendall J, Alexander J, Levy D, Wigler M. Early Detection of Cancer in Blood Using Single-Cell Analysis: A Proposal. Trends in Molecular Medicine **2017**;23:594-603
26. Sokolenko AP, Imyanitov EN. Molecular Diagnostics in Clinical Oncology. Frontiers in Molecular Biosciences **2018**;5
27. Baslan T, Hicks J. Unravelling biology and shifting paradigms in cancer with single-cell sequencing. Nature Reviews Cancer **2017**;17:557
28. Zhang X, Marjani SL, Hu Z, Weissman SM, Pan X, Wu S. Single-Cell Sequencing for Precise Cancer Research: Progress and Prospects. Cancer research **2016**;76:1305-12
29. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH*, et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc **2014**;9:2586-606
30. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A **2012**;109:14508-13
31. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA. Accuracy of Next Generation Sequencing Platforms. Next Gener Seq Appl **2014**;1
32. Krimmel JD, Schmitt MW, Harrell MI, Agnew KJ, Kennedy SR, Emond MJ*, et al.* Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. Proc Natl Acad Sci U S A **2016**;113:6005-10
33. Chatterjee A, Dasgupta S, Sidransky D. Mitochondrial subversion in cancer. Cancer Prev Res (Phila) **2011**;4:638-54
34. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell **2011**;144:646-74

35. Sanchez-Arago M, Chamorro M, Cuezva JM. Selection of cancer cells with repressed mitochondria triggers colon cancer progression. Carcinogenesis **2010**;31:567-76
36. Larman TC, DePalma SR, Hadjipanayis AG, Cancer Genome Atlas Research N, Protopopov A, Zhang J*, et al.* Spectrum of somatic mitochondrial mutations in five cancers. Proc Natl Acad Sci U S A **2012**;109:14087-91
37. Yu M. Somatic mitochondrial DNA mutations in human cancers. Adv Clin Chem **2012**;57:99-138
38. Ward PS, Thompson CB. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. Cancer Cell **2012**;21:297-308
39. Kennedy SR, Salk JJ, Schmitt MW, Loeb LA. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet **2013**;9:e1003794
40. Hoekstra JG, Hipp MJ, Montine TJ, Kennedy SR. Mitochondrial DNA mutations increase in early stage Alzheimer disease and are inconsistent with oxidative damage. Annals of Neurology **2016**;80:301-6
41. Risques RA, Lai LA, Brentnall TA, Li L, Feng Z, Gallaher J*, et al.* Ulcerative colitis is a disease of accelerated colon aging: evidence from telomere attrition and DNA damage. Gastroenterology **2008**;135:410-8
42. Greaves LC, Elson JL, Nooteboom M, Grady JP, Taylor GA, Taylor RW*, et al.* Comparison of mitochondrial mutation spectra in ageing human colonic epithelium and disease: absence of evidence for purifying selection in somatic mitochondrial DNA point mutations. PLoS Genet **2012**;8:e1003082
43. Greaves LC, Barron MJ, Plusa S, Kirkwood TB, Mathers JC, Taylor RW*, et al.* Defects in multiple complexes of the respiratory chain are present in ageing human colonic crypts. Exp Gerontol **2010**;45:573-9
44. Greaves LC, Preston SL, Tadrous PJ, Taylor RW, Barron MJ, Oukrif D*, et al.* Mitochondrial DNA mutations are established in human colonic stem cells, and mutated clones expand by crypt fission. Proc Natl Acad Sci U S A **2006**;103:714-9
45. Novak EA, Mollen KP. Mitochondrial dysfunction in inflammatory bowel disease. Frontiers in cell and developmental biology **2015**;3:62
46. Ussakli CH, Ebaee A, Binkley J, Brentnall TA, Emond MJ, Rabinovitch PS*, et al.* Mitochondria and Tumor Progression in Ulcerative Colitis. Journal of the National Cancer Institute **2013**
47. Sholl LM, Aisner DL, Allen TC, Beasley MB, Cagle PT, Capelozzi VL*, et al.* Liquid Biopsy in Lung Cancer: A Perspective From Members of the Pulmonary Pathology Society. Archives of Pathology & Laboratory Medicine **2016**;140:825-9
48. Merker JD, Oxnard GR, Compton C, Diehn M, Hurley P, Lazar AJ*, et al.* Circulating Tumor DNA Analysis in Patients With Cancer: American Society of Clinical Oncology and College of American Pathologists Joint Review. Journal of Clinical Oncology **2018**;36:1631-41
49. Karst AM, Drapkin R. Ovarian Cancer Pathogenesis: A Model in Evolution. Journal of Oncology **2010**;2010:13
50. Finch A, Beiner M, Lubinski J, et al. Salpingo-oophorectomy and the risk of ovarian, fallopian tube, and peritoneal cancers in women with a brca1 or brca2 mutation. JAMA **2006**;296:185-92
51. Piek JMJ, van Diest PJ, Zweemer RP, Jansen JW, Poort-Keesom RJJ, Menko FH*, et al.* Dysplastic changes in prophylactically removed Fallopian tubes of women predisposed to developing ovarian cancer. The Journal of pathology **2001**;195:451-6

52. Lee Y, Miron A, Drapkin R, Nucci M, Medeiros F, Saleemuddin A*, et al.* A candidate precursor to serous carcinoma that originates in the distal fallopian tube. The Journal of pathology **2007**;211:26-35
53. The Cancer Genome Atlas Research N, Bell D, Berchuck A, Birrer M, Chien J, Cramer DW*, et al.* Integrated genomic analyses of ovarian carcinoma. Nature **2011**;474:609
54. Vang R, Levine DA, Soslow RA, Zaloudek C, Shih I-M, Kurman RJ. Molecular Alterations of TP53 are a Defining Feature of Ovarian High-Grade Serous Carcinoma: A Rereview of Cases Lacking TP53 Mutations in The Cancer Genome Atlas Ovarian Study. International Journal of Gynecological Pathology **2016**;35:48-55
55. Kinde I, Bettegowda C, Wang Y, Wu J, Agrawal N, Shih I-M*, et al.* Evaluation of DNA from the Papanicolaou Test to Detect Ovarian and Endometrial Cancers. Science Translational Medicine **2013**;5:167ra4-ra4
56. Wang Y, Li L, Douville C, Cohen JD, Yen T-T, Kinde I*, et al.* Evaluation of liquid from the Papanicolaou test and other liquid biopsies for the detection of endometrial and ovarian cancers. Science Translational Medicine **2018**;10
57. Maritschnegg E, Wang Y, Pecha N, Horvat R, Van Nieuwenhuysen E, Vergote I*, et al.* Lavage of the Uterine Cavity for Molecular Detection of Müllerian Duct Carcinomas: A Proof-of-Concept Study. Journal of Clinical Oncology **2015**;33:4293-300

Chapter 1: Precancer in Ulcerative Colitis: The Role of the Field Effect and its Clinical Implications

Kathryn T. Baker[1], Jesse J. Salk[2, 4], Teresa A. Brentnall[3] and Rosa Ana Risques[1]

University of Washington, Departments of [1]Pathology, [2]Medicine, Division of Hematology and Oncology and [3]Medicine, Division of Gastroenterology, Seattle, WA 98195. [4]TwinStrand Biosciences, Seattle, WA 98121.

Short title: Field Effect in Ulcerative Colitis

Keywords: Field effect, field cancerization, field defect, precancer evolution, preneoplastic, clonal expansion, ulcerative colitis, inflammatory bowel disease, early detection

Funding: JJS: R01CA160674, T32HL007093; TAB: R01CA160674. RAR: R01CA181308.

Financial disclosures: JJS is a founder and equity holder in TwinStrand Biosciences.

Published as:
Kathryn T Baker, Jesse J Salk, Teresa A Brentnall, Rosa Ana Risques; Precancer in ulcerative colitis: the role of the field effect and its clinical implications, *Carcinogenesis*, Volume 39, Issue 1, 12 January 2018, Pages 11–20, https://doi.org/10.1093/carcin/bgx117

Abstract

Cumulative evidence indicates that a significant proportion of cancer evolution may occur before the development of histological abnormalities. While recent improvements in DNA sequencing technology have begun to reveal the presence of these early preneoplastic clones, the concept of 'premalignant fields' was already introduced by Slaughter more than half a century ago. Also referred as the "field effect", "field defect", or "field cancerization", these terms describe the phenomenon by which molecular alterations develop in normal-appearing tissue and expand to form premalignant patches with the potential to progress to dysplasia and cancer. Field effects have been well characterized in ulcerative colitis, an inflammatory bowel disease that increases the risk of colorectal cancer. The study of the molecular alterations that define these fields is informative of mechanisms of tumor initiation and progression and has provided potential targets for early cancer detection. Herein, we summarize the current knowledge about the molecular alterations that comprise the field effect in UC and the clinical utility of these fields for cancer screening and prevention.

Summary

Cancer often arises in preneoplastic fields of histologically normal appearance. These fields have been extensively studied in ulcerative colitis, a cancer predisposing inflammatory bowel disease. Here we review the field effect in ulcerative colitis and its utility to improve early cancer detection.

Glossary

Field effect – Also known as field cancerization. The phenomenon by which molecular alterations develop in normal-appearing tissue and clonally expand to form patches of cells with potential to further progress to dysplasia or cancer. These patches are also called premalignant fields or field defects.

Dysplasia – Dysplasia refers to the presence of histological alterations within a tissue, which may signify a precancerous stage. In patients with UC, dysplasia is classified as Indefinite for Dysplasia (IND), Low Grade Dysplasia (LGD) or High-Grade Dysplasia (HGD) based on the extent of architectural and cytological abnormalities.

Clonal expansion – The propagation of a population of daughter cells from a single cell of origin. Clonal expansions that carry cancer-promoting molecular alterations that enhance cellular survival and proliferation form the basis of the field effect.

UC Non-Progressor – A patient with UC who has not developed dysplasia or cancer.

UC Progressor – A patient with UC that has developed dysplasia or cancer.

Introduction

Cancers evolve through an iterative process of mutation, selection, and clonal expansion (1). Most tumor types take multiple years to develop, during which time preneoplastic cells clonally expand and progressively acquire the molecular alterations necessary to allow them to fully escape growth control checkpoints and invade surrounding tissues. Precisely how much of this process occurs prior to the development of morphologically recognizable malignancy is unknown, but it has been estimated that cancer cells accumulate about half of their mutational load before tumor initiation (2). This implies that a substantial proportion of a cancer's development entails precancerous clonal evolution within histologically normal-appearing tissue.

During the last several years, new and extremely sensitive DNA sequencing technologies have begun to directly reveal the presence of these early clones (3,4). However, the notion that preneoplastic changes precede cancer was recognized more than a half century ago. Prior to even the modern understanding of DNA as the genetic material, Slaughter et al. proposed the concept of field cancerization to describe "preconditioned epithelium activated over an area in which multiple cell groups undergo a process of irreversible change towards cancer". This preneoplastic condition was originally described in oral carcinoma on the basis of subtle morphological abnormalities in surrounding tissue. It was postulated that this abnormal "field" underlied the relatively common finding of multiple synchronous primary tumors and the high frequency of local recurrence in head and neck squamous cell cancers, despite apparent complete tumor resection (5).

Although Slaughter and colleagues recognized field cancerization as an important clinical phenomenon, they lacked a mechanistic explanation. The following decades revealed examples of field cancerization associated with many other epithelial cancers and demonstrated that the fields could be characterized by defined molecular aberrations present in histologically normal tissue (6). Contemporary molecular tools have provided important insights into the basis of the cancer-prone phenotype of these fields. In a variety of examples (6,7), peritumoral fields were shown to encompass populations of clonally related cells that bear some, but not all, of the genetic changes of the tumor itself. The relative ease with which a second tumor or tumor relapse can occur within such a field simply reflects the relatively low evolutionary hurdle for one of these partially dysregulated cells to acquire the last molecular change(s) needed for full-blow malignancy.

While the specific nature of the fields, and the molecular mechanisms that initiate them, are likely to be different in different tissue types, the field concept illustrates the general notion of multistep carcinogenesis, wherein ancestral populations of preneoplastic cells can both precede and co-exist with a cancer (6). Such a field effect may be thought of as an early stage of the neoplastic process where selected mutant cells can incrementally enhance growth properties. Because these preneoplastic populations are often morphologically normal in appearance and may be very small, they frequently go undetected in sporadic tumors. Several preneoplastic diseases, however, often exhibit large preneoplastic fields and offer an excellent opportunity to study the initial stages of tumor development. Ulcerative colitis (UC) and Barrett's esophagus (BE) are among the best characterized. Fields can expand several centimeters in BE (8,9) and practically the entire length of the colon (~150cm) in UC (10). In both diseases, chronic inflammation generates extensive damage to epithelial cells, leading to increased cell replication and/or direct DNA damage. Subsequent mutations that alter growth control genes enable clonal expansions, which result in patches of cells that share identical mutations. In some cases, a single genetic

change can be found in all cells within an entire field many centimeters in size, indicating a single clonal founder cell. In other cases, multiple independent clones with distinct genetic signatures evolve simultaneously in response to the inflammatory environment.

A further advantage of the study of these preneoplastic diseases is that, in addition to non-dysplastic fields and overt cancer, intermediate degrees of dysplasia are routinely found during endoscopic surveillance. These intermediate stages help delineate with even finer precision the multistep sequence of tumor progression and provide a unique longitudinal window of opportunity to study early cancer and its patterns of evolution (11,12).

In this review, we focus on UC as a model of inflammation-mediated tumorigenesis in which the field effect has been extensively characterized. The search terms used to query the literature include field effect, field defect, and field cancerization in ulcerative colitis. We prioritized a broad discussion of the concept of field effect and its implications as opposed to a detailed recollection of articles describing precancerous fields. We first describe the disease's epidemiology, histologic sequence of neoplastic progression, and the types of molecular alterations that have, thus far, been identified in these fields. We then discuss the clinical implications of these fields with a special focus on their use as biomarkers of early or imminent cancer. Field effects have been identified in a growing variety of cancers including: breast (13) (14), head and neck (15), bladder (16) (17), colorectal (18), gastric (19) (20) (21), prostate (22) (23), lung (24) (25), skin (26) (27) (28), liver (29), ovarian (30), and cervical (31). Although UC is responsible for only a small fraction of the global burden of colon cancer, we posit that the lessons learned from studying tumor progression in this unique disease can be applied to the understanding, prevention, and clinical management of many other malignancies associated with field effects.

Colorectal Cancer Risk in Ulcerative Colitis

UC is one of the two major types of inflammatory bowel disease and is characterized by uninterrupted stretches of chronic inflammation of the colon mucosa. It affects roughly 1 million patients in the U.S. and its prevalence is increasing worldwide (32-34). The cause of UC remains to be fully determined, but a preponderance of evidence suggests that it is the result of a complex interaction between a dysregulated host immune system, the gut microbiome, and diet (35-39). A significant aspect of the management of UC is that it elevates the risk of colorectal cancer (CRC) (40) and cancer-related deaths, although improvements in surveillance methods appear to have decreased both the incidence (41) and mortality (42,43) of CRC in UC in recent years. The increased risk of CRC is attributable to multiple aspects of chronic inflammation and immune dysregulation (44,45). Patients whose inflammation is more severe (46) and more extensive (40,47) are more likely to develop CRC. Other risk factors include prolonged disease duration (47-50), concurrent diagnosis of primary sclerosing cholangitis, an autoimmune disorder of the biliary system (51-53), a family history of sporadic CRC (45,54,55), early age of UC onset (56-58), and extent of dysplasia (59).

Colorectal cancer development in the setting of UC differs from that of sporadic CRC in several respects (60). Histologically, adenomatous polyps typically precede sporadic CRC whereas UC-associated CRC (UC-CRC) often arises from flat dysplasia. Sporadic CRC is believed to be initiated by mutations in *APC*, followed by mutations in *KRAS* and *TP53* (61), although it is currently appreciated that this progression does not necessary follow a linear sequence (62). In UC, *TP53* mutations appear to be the initiating mutation in most lesions, although mutations in *KRAS* have also been identified as a founder event in a minority of cases (63). Recently, NGS-based studies confirmed a higher frequency of *TP53* mutations and lower frequency of *APC* and *KRAS* mutations in UC-CRC compared to sporadic CRC (64,65). Epidemiologically, the mean age

for CRC development in the general population is 64 years (66) versus 43 years for UC-CRC (33). In addition, the prognosis of UC-CRC is poorer than sporadic CRC, although it is unclear if this reflects the tumor biology itself, the average stage of disease at diagnosis, or other health challenges faced by UC-patients (67) including those related to immunosuppressive therapies.

A common trait between sporadic and UC-associated CRC appears to be the pivotal role of an abnormal intestinal microbiota as an initiating mechanism. In the last ten years, a large body of evidence has accumulated linking CRC with dysbiotic gut microbiota and dysregulated immunity, both in the context of sporadic CRC and inflammatory bowel diseases (IBD) (68-70). A dysbiotic microbiota contributes to tumor progression directly by generating reactive metabolites and carcinogens, and indirectly by disrupting the epithelial cell barrier in the host (70). This causes local intolerance to antigens of normal flora and leads to dysregulation of the adaptive and innate immune response and subsequent chronic inflammation (68). Diet appears to be an important contributor in this process, as it has a major influence on the gut flora and is transformed into metabolites that can have protective or promoting roles in tumor progression (71). In the case of patients with UC, genetic predisposition might contribute to a dysbiotic gut flora, causing the extensive chronic inflammation characteristic of this disease (68) and increasing the risk of tumor progression through the extensive cell proliferation required in repeated cycles of wound and repair (12,72).

The sequence of tumor progression in Ulcerative Colitis

Tumor progression in UC is clinically described as a multistep process defined by increasing degrees of histological abnormalities, progressing from no dysplasia, to low-grade dysplasia (LGD), high-grade dysplasia (HGD), and finally cancer (**Figure 1**). Although often represented linearly and sequentially, it is important to recognize that tumor evolution is usually

branched, not linear (73) and, in the case of UC tumorigenesis, not every dysplastic stage may be observed. This sequence might also occur independently in multiple locations in the colon. It is well established that in UC patients, multiple areas of the colon can simultaneously develop dysplastic changes and that independent synchronous cancers can evolve in parallel within these fields (74).

Recent mounting evidence has led to the appreciation that clinical patterns of progression to CRC in UC may differ based on the age of disease onset. UC has two peaks of incidence: early onset occurs between 25 and 35 years of age while late onset arises between 55 and 65 years of age (56,75). Clinical studies have recognized for some time that patients with late onset tend to have less extensive disease and a lower risk of CRC development (76-78). Interestingly, Brackmann *et al.* (58,59) reported that among UC patients with CRC, those with late age of onset tended to develop cancer without widespread dysplasia. Conversely, patients with early age of onset typically exhibited extensive dysplasia at CRC diagnosis. The presence of extensive dysplasia was also associated with longer disease duration prior to CRC development, higher probability of presenting with active inflammation (58), and worse CRC prognosis (59). These findings suggested that the development of CRC in patients with early onset disease is related to long exposure to inflammation and subsequent development of dysplasia. However, in patients with late onset of disease, CRC might arise independently of observable dysplasia or, alternatively, tumors might be fast growing and displace their original localized dysplastic fields in a clonal sweep. Our group has identified molecular evidence further supporting fundamental differences between early and late onset disease. We demonstrated that UC-cancer patients with early onset of disease have extensive fields of molecular abnormalities throughout their colons compared to UC-cancer patients who have late onset of disease. Specifically, large clonal populations with shortened telomeres could be found in multiple non-dysplastic areas of the colon of early onset

UC Progressors (patients with HGD or cancer), but this was almost never the case in the Progressors who had late onset of UC (79). While more research is needed to better distinguish these two modes of CRC progression in UC, their recognition as distinct entities has important clinical implications. Since patients with late onset of disease appear to develop cancer without a widespread field effect, these patients might benefit from partial colon resection instead of full colectomy, which is the current standard-of-care. The ability to safely use segmental resection of the colon in older UC patients would spare them significant morbidity.

Molecular alterations characterize preneoplastic fields in UC

The field effect was originally described based on regions of tissue sharing histological abnormalities (5), but the concept was later broadened to include clonal molecular abnormalities in otherwise histologically normal-appearing tissue, as it was recognized that multiple molecular changes produce Slaughter's original observation (6). More recently, the term has been used more broadly to describe an "etiologic" field effect, which takes into consideration the contribution of environmental and genetic factors that produce cancer susceptibility (80). The genetic component of UC is well established as well as the critical role of an altered microbiome, which dysregulates immunity and triggers inflammation. While the presence of these factors is not indicative of cancer progression, they contribute to the neoplastic process by producing a predisposing microenvironment. (80). Figure 1 illustrates the 3 conceptual types of field effects—etiological, molecular, and morphological—in the context of UC cancer progression. These fields represent increasing levels of cancer susceptibility that are operative in the colons of patients with UC, each level contributing to the development of the next. However, only molecular and dysplastic fields are indicative of an underlying preneoplastic process. Here we focus on molecular fields due to their importance to understand tumor progression and their potential for early cancer detection.

In this section, we first summarize the genetic alterations that have provided evidence of clonal field effects in UC. These alterations include somatic point mutations, chromosomal alterations, and passenger mutations in polyguanine tracts. We then describe a second set of molecular alterations that define preneoplastic fields in UC without an implicit assumption of clonality. These include telomere shortening, mitochondrial alterations, and epigenetic changes. Some of these alterations have provided meaningful clues in terms of the underlying molecular mechanisms that may drive tumor evolution in UC and for a more extensive review on that topic the readers are referred to a recent excellent article by Choi et al. (12). Of note, many of the molecular alterations discussed here, both clonal and non-clonal, harbor potential as UC cancer biomarkers. However, our goal is not the description of the biomarker value of each alteration, which has been previously done by us and others (81-83), but the review of the evidence for a field effect based on those alterations and the discussion of the clinical applicability of those fields for optimal cancer surveillance.

Clonal alterations

Point mutations

Mutations in *TP53* are the most common and best characterized single nucleotide variants in UC-associated preneoplastic fields. *TP53* mutations (84) and loss of heterozygosity (84,85) occur early in UC neoplastic development. UC patients with *TP53* mutations in non-dysplastic biopsies are four times more likely to progress to dysplasia and cancer (86). Additionally, there is a strong correlation between mutations in highly conserved regions of p53 and the histological progression from LGD to cancer in UC patients (82). In an early study by our group, we examined alterations in *TP53* by FISH in order to characterize the spatial pattern of these mutational events in individual cells within crypts (87). A detailed analysis of multiple crypts demonstrated that most *TP53* FISH abnormalities are shared by all the crypts within a colonic region, indicating

monoclonality. The observation of the same *TP53* alterations in the two branches of a crypt in fission strongly suggested that clonal expansion of mutated cells occurs by crypt fission and provided direct observation of how clonal fields propagate themselves in UC. In a later study, Leedham et. al. (63) identified several molecular alterations, including *TP53* and *KRAS* mutations, in individual, microdissected dysplastic crypts and adjacent non-dysplastic crypts. In one UC Progressor case, the same founding mutation was found in spatially separated tumors 14 cm apart from each other and in the non-dysplastic surrounding tissue, clearly demonstrating field cancerization (63). The authors also found clonally disparate tumors in another UC Progressor, supporting the idea that a common etiologic risk factor, i.e. inflammation, has sufficient carcinogenic potential to facilitate the emergence of multiple synchronous clonal fields throughout the colon.

Although it is rapidly becoming the new technical standard, we are aware of only two studies that have yet applied next-generation sequencing (NGS) to the study of somatic mutations in UC, and both were limited to the interrogation of tumors, rather than of preneoplastic fields (64,65). While these reports provide a useful baseline from which to compare the mutational landscape of sporadic versus UC-associated CRC, additional knowledge remains to be generated from applying this technology to a comprehensive study of the spatial and temporal pattern of mutation accumulation in preneoplastic fields.

Aneuploidy and Chromosomal Alterations

For more than 25 years, aneuploidy has been recognized as an early occurring alteration in UC carcinogenesis, often found even before the appearance of dysplasia (88,89). The presence of aneuploid fields is associated with a higher risk of progression to dysplasia (90), histological grade (90), disease duration (90), and the presence of PSC (91). More sensitive technologies, including

comparative genomic hybridization (92) and fluorescence *in situ* hybridization (FISH), revealed that chromosomal alterations occur early in UC tumorigenesis, often preceding histologically defined dysplasia (93) and affecting the entirety of the colon (10). The relative timing and frequency of numerical chromosomal alterations in UC differs significantly from those of sporadic CRC, supporting the conclusion that neoplastic progression follows distinct pathways in these diseases (93). More recently, CGH-array studies have demonstrated that in UC Progressors, chromosomal alterations can be found in distant normal-appearing biopsies (94) and the same alteration can be shared by multiple biopsies spanning most of the length of the colon (95). This indicates that the field effect can be very large and raises the fundamental question of how these clones propagate. The fields also appeared to be graded in nature, as copy number alterations increased in frequency and magnitude with proximity to dysplasia.

While the mechanisms for localized clonal expansions have been well characterized (12), it is unclear how a clone can generate the extensive pancolonic fields described above. Clones originate from stem cells that expand to occupy the whole crypt via niche succession and then crypts laterally expand by crypt-fission to generate monoclonal patches (12). Niche succession might occur by neutral drift (96), but the expansion of certain mutations beyond a crypt appears to involve selection (97,98). This process can generate clones of >10cm in size (63), but it appears unlikely that the same process would extent throughout the whole organ. Alternative hypotheses are convergent evolution and long-distance stem cell migration aimed at mucosal healing, as proposed in Choi *et al.* (12). Further investigation in this area is highly needed and would greatly benefit from more advanced methods of lineage detection, as explained in the next section.

Clonal Expansions detected by passenger mutations

The ability to recognize clonal expansions requires the presence and identification of one or more genetic changes that identically mark the clone's progeny as related to each other, yet distinct from adjacent cells. This lineage marker is typically a putative molecular driver of the clonal outgrowth. The challenge is, however, that just as with tumors, multiple genetic changes can drive the clonal expansion. These changes vary from one person to another, and even among different clones within a single individual. Thus, many of these fields might be undetectable if only screening for known driver mutations. An alternative approach is to focus on passenger mutations. As cells divide, replication errors produce mutations, the vast majority of which are functionally neutral and offer no selective advantage or disadvantage (7). These mutations are carried in the daughter cells as 'passenger' mutations and offer a much larger repertoire of somatic variants to enable the detection of clonal fields and elucidate lineage relationships.

Our group pioneered an approach for clone detection based on passenger mutations in polyguanine tracts (PolyG), which are highly mutable repetitive DNA sequences interspersed throughout the genome (99,100). We demonstrated that extensive clonal fields defined by PolyG mutations were present in histologically normal colonic biopsies of UC Progressors, but were almost entirely absent in UC Non-Progressors (patients without dysplasia). These fields were composed of thousands of cells that appeared microscopically normal, but had aberrantly proliferated from an original precursor cell, indicating the presence of an "occult" process of neoplastic evolution. In a later study these clonal fields were only found in patients with early onset of disease, suggesting that alternative pathways of progression without extensive clonal expansions might be operative in patients that develop UC later in life (100).

Other molecular alterations identified in preneoplastic fields

Telomere shortening

Telomere shortening is another well-documented, early event in UC tumorigenesis. While it is not informative about clonality, telomere shortening has provided insight about the extent of the field effect and mechanisms of tumor progression in UC (101-105). The colonic epithelium of patients with UC has shorter telomeres than age-matched non-UC patients (106,107). This shortening appears to occur within the first 8 years of disease duration (106), which, interestingly, coincides with the time at which clinical risk of CRC for UC patients increases. This suggests that the onset of cancer may depend upon telomeres becoming critically short. Telomere shortening occurs diffusely in the colonic epithelium of all UC patients (106), especially in those with more severe clinical phenotypes (108). Additionally, telomere shortening is more common in biopsies closer to dysplasia (102) and is more extreme in UC Progressors compared to Non-Progressors (102,104,105).

Mitochondrial dysfunction

Growing evidence indicates that epithelial cell mitochondrial dysfunction is present in UC (109), although it remains unclear whether this is the cause or consequence of the disease and whether it contributes to cancer progression. The role of mitochondria in cancer, in general, is controversial. One thought that prevailed for some time was that mutations in mitochondrial DNA (mtDNA) were somehow advantageous to tumors and clonally expanded into fields under positive selection (110). This was proposed to be the basis of the Warburg effect –the observation that most cancers use glycolysis for energy production (111-113). This idea has been challenged, however, by more recent studies arguing that many cancers do, in fact, rely on oxidative phosphorylation and mitochondrial function (114-116) and that mtDNA mutations either accumulate randomly and clonally expand as passengers without selective pressure, or are selected against (117-119).

In UC, genetic (120), proteomic (38,121-123), and metabolic (124-126) studies have identified mitochondrial alterations in colonic biopsies, both in non-dysplastic mucosa and in UC-associated cancers. Unfortunately, some results are contradictory and the relevance of these findings is still unclear. Our group previously demonstrated that in UC Progressors, the levels of cytochrome C oxidase subunit I (COX), a protein of complex I of the electron transport chain, decreased with proximity to dysplasia, indicating the presence of a gradient field effect. COX staining was completely absent in some dysplastic areas but, remarkably, was typically high in HGD and cancer (127). These results are concordant with mitochondrial dysfunction as a feature of UC colonic epithelium (109), but suggest that function might be restored later in progression to allow for the metabolic demands of cancer cells (116).

The mechanisms that trigger these processes are unknown. Field effects that include mitochondrial dysfunction can sometimes arise from clonal expansions of mtDNA mutations through crypt conversion and crypt fission, as previously characterized in the aging colon (128). Additionally, *PGC1α*, the master regulator of mitochondrial biogenesis, might mediate mitochondrial dysfunction. *PGC1α* expression is decreased in the intestinal epithelium of patients with UC. Notably, in mice, its deletion confers susceptibility to colitis, whereas restoration of the protein ameliorates the disease and restores mitochondrial integrity (129). *PGC1α* also offers a potential link between telomeres and mitochondria: in telomerase knockout mice, shortened telomeres trigger mitochondrial dysfunction via *TP53* and *PGC1α* signaling (130,131). This intriguing connection deserves further investigation in UC tumorigenesis, especially since both alterations exhibit a similar pattern of initial dysfunction followed by later recovery.

Epigenetic changes

Chronic inflammation is well known to play a role in the progression of cancer—UC-mediated CRC being only one of many examples (132). As noted above, the inflammatory state can mutate DNA and lead to disruption of growth control genes as well as accelerate mutation acquisition by increasing cell turnover. Additionally, chronic inflammation can epigenetically alter epithelial cells without a clonal relationship (132). Gloria et al. first identified DNA methylation changes in the setting of UC by measuring the incorporation of labeled methyl groups into DNA (133). They demonstrated that UC colonic DNA is globally hypomethylated compared to normal controls and suggested that epigenetic changes in UC colonic mucosa contribute to cancer progression. Since then, epigenetic changes have been more extensively studied as precursor lesions in UC. It has been observed that histone modification genes are overexpressed in UC and that the level of overexpression correlates with both disease extent and duration (134). In non-dysplastic, but inflamed, UC tissue, hypomethylation of bivalent H3K27me3-associated promoters facilitates the upregulation of cancer progression associated genes, including those associated with cell movement, death, survival, and proliferation. These inflammation-induced changes create a field of susceptibility that might predispose to cancer progression (135). Additionally, both dysplastic and normal-appearing UC Progressor epithelium features hypermethylation at CpG islands, similar to what is seen with aging, and might contribute to increased susceptibility (136-138). Other studies have observed a significantly higher methylation in genes associated with UC inflammation from UC Progressor non-dysplastic tissues when compared to UC Non-Progressor tissue (139). Collectively, these findings in preneoplastic fields are consistent with the epigenetic alterations found in CRC, in which both global hypomethylation and regional hypermethylation are observed (140). Thus, the current view is that epigenetic alterations play a role in the development and progression of UC (141). These alterations might create an epigenetic field effect

as a result of the clonal expansion of stem cells that carry epigenetic changes or non-clonally as a result of environmental exposures and inflammation (132,141,142).

Model of cancer progression in UC: accelerated colon aging?

The molecular alterations that define preneoplastic fields in UC are remarkably similar to the changes that occur in normal, aging tissue. Based on this, our group proposed the idea that UC can be thought of as a disease of accelerated colon aging. In normal individuals, telomere length declines progressively with age (81), but in UC the rate of decline is accelerated, especially within the first 8 years after disease diagnosis (106). On average, individuals with UC at age 40 carry colonocyte telomeres as short as non-UC individuals at age 60. This finding fits with the epidemiological observation that CRC develops at a mean age of 64 in the general population, but a mean age of 43 in UC patients--about 20 years earlier (33).

Similarly, chromosomal alterations (96,143), mitochondrial loss (144,145), and DNA methylation (146) changes have been reported in both normal aging colon and preneoplastic fields in UC. In the normal colon, these age-related alterations are attributed to the increased load of somatic mutations and molecular damage over time. While somatic mutations can lead to clonal fields through biased competition and expansion of mutated intestinal stem cells (147), extensive molecular damage can lead to non-clonal fields of cancer predisposing alterations, such as telomere shortening caused by oxidative damage (148).

We have made an effort to integrate the findings described above in our proposed model of carcinogenesis in UC (Fig. 1). We postulate that in UC, genetic susceptibility and an altered microbiome and immunity lead to chronic inflammation and concomitant increased cell turnover and oxidative damage. This accelerates the rate of mutation accumulation and molecular damage, thus effectively producing accelerated aging of the colon. When telomeres become critically short,

uncapped chromosome ends trigger a DNA damage response (149). In the presence of proficient *TP53*, this response results in the activation of cellular senescence; however, if *TP53* is mutated, cells bypass senescence and cell division continues, producing preneoplastic fields in which dysfunctional telomeres trigger end-to-end fusions and chromosomal instability (105). This instability and the progressive exhaustion of telomeres eventually leads cells to crisis and death (150). During this process, however, additional genetic and epigenetic alterations accumulate and might disrupt other tumor suppressors genes and activate proto-oncogenes. One of the many consequences of this can be mutational or epigenetic reactivation of telomerase. Reactivated telomerase rescues the cells from crisis, allows further proliferation under the drive of mutated oncogenes, and eventually results in the development of invasive malignancy (151). At that step cancer cells might also acquire proficient mitochondria to cover the metabolic needs to rapid growth (116).

Field effect implications: opportunities for early cancer detection

Preneoplastic fields are an indication of an emerging neoplastic process and, therefore, they could be used to improve cancer detection in UC. The current cancer surveillance system is based on colonoscopic screening for dysplasia and it is likely to be one of the factors contributing to the decrease in UC-associated CRC in recent years (44). However, this approach has limited sensitivity (152), fails to detect every patient at risk (153), and it is time consuming and expensive. A more efficient and sensitive system would be highly desirable, especially in view of the worldwide overall increase in UC incidence (154).

The American Gastroenterology Association recommends that colonoscopic surveillance should begin 8-10 years after disease diagnosis and should occur every 1-2 years depending on dysplasia findings (155,156). Until recently, the guidelines included collection of at

least 32 random quadrant biopsies to achieve 90% sensitivity for histological identification of dysplasia (90). However, gastroenterologists' non-adherence to colonoscopy guidelines, patients' non-compliance to the surveillance plan, and a lack of agreement between pathologists upon histological assessment reduce the efficacy of this surveillance approach (57,157,158). Over the last decade, the superiority of chromoendoscopy (CE) for the detection of dysplasia has been established (159-161). Whereas standard colonoscopic methods rely on the use of white light to visualize areas of dysplasia, CE uses dyes that stain the mucosa, increasing the contrast and the sensitivity to find dysplasia. CE is the method currently recommended for colonoscopic surveillance (162), but it is still limited by the requisite detection of morphological changes that are visible by endoscopy.

As described above, cumulative evidence demonstrates that molecular changes precede the morphological changes associated with dysplasia. Thus, focusing on the identification of preneoplastic fields may prove to be the most sensitive approach for identifying patients at risk of CRC progression. Figure 2 illustrates the potential use of the field effect to improve cancer colonoscopic surveillance in patients with UC. The premise is that CE with targeted biopsies is performed, in accordance to current recommendations (162), but in addition, two biopsies are collected at each colon segment for molecular analysis of field effects. The extension of the fields, as well as the degree of molecular alterations, might reflect the likelihood of cancer progression and could potentially be used to personalize colonoscopic surveillance, even in the absence of dysplastic findings. This hypothesis is based on the fact that cancer is a probabilistic process that depends not only on the rate of mutation, but also on the number of cells at risk (163). Thus, patients with pancolonic and multifocal fields might benefit from closer monitoring as compared with patients with localized, unifocal fields. Given the vast heterogeneity of genetic changes that can lead to UC-CRC (65), traditional genetic markers such as *TP53* mutations. are unlikely to be

universal detectors of preneoplastic fields. Passenger mutations, such as indels in PolyG tracts, overcome this problem and might offer a promising solution for the identification of preneoplastic progression in UC. This information could be integrated into predictive statistical models to identify patients at risk. In addition, studies of the field effect and its role in early cancer detection in UC could be supported by the use of mathematical modeling to quantify the extent and behavior of these fields and identify at-risk patients (164,165). Such computational approaches have been successfully applied to predict the size, shape, and distribution of fields (164), to predict cancer risk in Barrett's esophagus (166) and head and neck premalignant lesions (167), and to evaluate the prognostic value of putative biomarkers (165).

Moving forward, large studies are required to establish the value of molecular fields for cancer detection and prediction in a clinical setting. Fortunately, large repositories of archival UC biopsies already exist, which include multiple longitudinal colonoscopies for each patient, each containing multiple random colonic biopsies. Such repositories are an excellent resource for determining the biomarker potential of molecular fields for cancer prediction.

Conclusions

Here, we have discussed the concept of the field effect, the molecular alterations that define these fields, and their implications for early cancer detection in UC. Preneoplastic fields precede cancer progression in UC and have been characterized through the analysis of *TP53* mutations, chromosomal alterations, telomere shortening, mitochondrial dysfunction, and epigenetic alterations. Biologically, these fields reflect the early events that lead to tumor progression in UC and enable accurate spatial and temporal characterization of in vivo tumor evolution. Clinically, preneoplastic fields provide an opportunity to improve early cancer detection in UC and to personalize colonoscopic surveillance. Beyond its applications in UC, the study of field defects is

highly relevant to current efforts to understand ‘precancer’ in order to improve early detection and prevention of cancer (168).

References

1. Nowell PC. The clonal evolution of tumor cell populations. Science **1976**;194:23-8
2. Tomasetti C, Vogelstein B, Parmigiani G. Half or more of the somatic mutations in cancers of self-renewing tissues originate prior to tumor initiation. Proc Natl Acad Sci U S A **2013**;110:1999-2004
3. Martincorena I, Campbell PJ. Somatic mutation in cancer and normal cells. Science **2015**;349:1483-9
4. Krimmel JD, Schmitt MW, Harrell MI, Agnew KJ, Kennedy SR, Emond MJ*, et al.* Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. Proc Natl Acad Sci U S A **2016**;113:6005-10
5. Slaughter DP, Southwick HW, Smejkal W. Field cancerization in oral stratified squamous epithelium; clinical implications of multicentric origin. Cancer **1953**;6:963-8
6. Braakhuis BJ, Tabor MP, Kummer JA, Leemans CR, Brakenhoff RH. A genetic explanation of Slaughter's concept of field cancerization: evidence and clinical implications. Cancer Res **2003**;63:1727-30
7. Salk JJ, Horwitz MS. Passenger mutations as a marker of clonal cell lineages in emerging neoplasia. Semin Cancer Biol **2009**;20:294-303
8. Brabender J, Marjoram P, Lord RV, Metzger R, Salonga D, Vallbohmer D*, et al.* The molecular signature of normal squamous esophageal epithelium identifies the presence of a field effect and can discriminate between patients with Barrett's esophagus and patients

with Barrett's-associated adenocarcinoma. Cancer Epidemiol Biomarkers Prev **2005**;14:2113-7
9. Prevo LJ, Sanchez CA, Galipeau PC, Reid BJ. p53-mutant clones and field effects in Barrett's esophagus. Cancer Res **1999**;59:4784-7
10. Rabinovitch PS, Dziadon S, Brentnall TA, Emond MJ, Crispin DA, Haggitt RC*, et al.* Pancolonic chromosomal instability precedes dysplasia and cancer in ulcerative colitis. Cancer Res **1999**;59:5148-53
11. Reid BJ. Early events during neoplastic progression in Barrett's esophagus. Cancer biomarkers : section A of Disease markers **2011**;9:307-24
12. Choi CR, Bakir IA, Hart AL, Graham TA. Clonal evolution of colorectal cancer in IBD. Nature reviews Gastroenterology & hepatology **2017**
13. Rivenbark AG, Coleman WB. Field cancerization in mammary carcinogenesis - Implications for prevention and treatment of breast cancer. Exp Mol Pathol **2012**;93:391-8
14. Heaphy CM, Griffith JK, Bisoffi M. Mammary field cancerization: molecular evidence and clinical importance. Breast Cancer Res Treat **2009**;118:229-39
15. Jaiswal G, Jaiswal S, Kumar R, Sharma A. Field cancerization: concept and clinical implications in head and neck squamous cell carcinoma. Journal of experimental therapeutics & oncology **2014**;10:209-14
16. Cheng L, Davidson DD, Maclennan GT, Williamson SR, Zhang S, Koch MO*, et al.* The origins of urothelial carcinoma. Expert review of anticancer therapy **2010**;10:865-80
17. Hoglund M. Bladder cancer, a two phased disease? Semin Cancer Biol **2006**;17:225-32
18. Amaro A, Chiara S, Pfeffer U. Molecular evolution of colorectal cancer: from multistep carcinogenesis to the big bang. Cancer Metastasis Rev **2016**;35:63-74
19. Maeda M, Moro H, Ushijima T. Mechanisms for the induction of gastric cancer by Helicobacter pylori infection: aberrant DNA methylation pathway. Gastric cancer : official journal of the International Gastric Cancer Association and the Japanese Gastric Cancer Association **2016**;20:8-15
20. Rugge M, Capelle LG, Cappellesso R, Nitti D, Kuipers EJ. Precancerous lesions in the stomach: from biology to clinical patient management. Best Pract Res Clin Gastroenterol **2013**;27:205-23
21. Graham TA, McDonald SA, Wright NA. Field cancerization in the GI tract. Future Oncol **2011**;7:981-93
22. Nonn L, Ananthanarayanan V, Gann PH. Evidence for field cancerization of the prostate. Prostate **2009**;69:1470-9
23. Squire JA, Park PC, Yoshimoto M, Alami J, Williams JL, Evans A*, et al.* Prostate cancer as a model system for genetic diversity in tumors. Advances in cancer research **2011**;112:183-216
24. Beane J, Mazzilli SA, Tassinari AM, Liu G, Zhang X, Liu H*, et al.* Detecting the Presence and Progression of Premalignant Lung Lesions via Airway Gene Expression. Clin Cancer Res **2017**;23:5091-100
25. Gomperts BN, Spira A, Massion PP, Walser TC, Wistuba, II, Minna JD*, et al.* Evolving concepts in lung carcinogenesis. Seminars in respiratory and critical care medicine **2011**;32:32-43
26. Stern RS, Bolshakov S, Nataraj AJ, Ananthaswamy HN. p53 mutation in nonmelanoma skin cancers occurring in psoralen ultraviolet a-treated patients: evidence for heterogeneity and field cancerization. J Invest Dermatol **2002**;119:522-6

27. Verkouteren JAC, Ramdas KHR, Wakkee M, Nijsten T. Epidemiology of basal cell carcinoma: scholarly review. The British journal of dermatology **2017**;177:359-72
28. Merkel EA, Gerami P. Malignant melanoma of sun-protected sites: a review of clinical, histological, and molecular features. Lab Invest **2017**;97:630-5
29. Utsunomiya T, Shimada M, Morine Y, Tajima A, Imoto I. Specific molecular signatures of non-tumor liver tissue may predict a risk of hepatocarcinogenesis. Cancer Sci **2014**;105:749-54
30. Kobayashi H, Iwai K, Niiro E, Morioka S, Yamada Y, Ogawa K*, et al.* The conceptual advances of carcinogenic sequence model in high-grade serous ovarian cancer. Biomedical reports **2017**;7:209-13
31. Chu TY, Shen CY, Lee HS, Liu HS. Monoclonality and surface lesion-specific microsatellite alterations in premalignant and malignant neoplasia of uterine cervix: a local field effect of genomic instability and clonal evolution. Genes Chromosomes Cancer **1999**;24:127-34
32. Kappelman MD, Moore KR, Allen JK, Cook SF. Recent trends in the prevalence of Crohn's disease and ulcerative colitis in a commercially insured US population. Digestive diseases and sciences **2012**;58:519-25
33. Eaden JA, Abrams KR, Mayberry JF. The risk of colorectal cancer in ulcerative colitis: a meta-analysis. Gut **2001**;48:526-35
34. Hanauer SB. Inflammatory bowel disease: epidemiology, pathogenesis, and therapeutic opportunities. Inflamm Bowel Dis **2006**;12 Suppl 1:S3-9
35. Palmieri O, Creanza TM, Bossa F, Palumbo O, Maglietta R, Ancona N*, et al.* Genome-wide Pathway Analysis Using Gene Expression Data of Colonic Mucosa in Patients with Inflammatory Bowel Disease. Inflamm Bowel Dis **2015**;21:1260-8
36. Ye BD, McGovern DP. Genetic variation in IBD: progress, clues to pathogenesis and possible clinical utility. Expert Rev Clin Immunol **2016**;12:1091-107
37. Grivennikov SI, Greten FR, Karin M. Immunity, inflammation, and cancer. Cell **2010**;140:883-99
38. Hsieh SY, Shih TC, Yeh CY, Lin CJ, Chou YY, Lee YS. Comparative proteomic studies on the pathogenesis of human ulcerative colitis. Proteomics **2006**;6:5322-31
39. Payne CM, Bernstein C, Dvorak K, Bernstein H. Hydrophobic bile acids, genomic instability, Darwinian selection, and colon carcinogenesis. Clinical and experimental gastroenterology **2008**;1:19-47
40. Dyson JK, Rutter MD. Colorectal cancer in inflammatory bowel disease: what is the real magnitude of the risk? World J Gastroenterol **2012**;18:3839-48
41. Castano-Milla C, Chaparro M, Gisbert JP. Systematic review with meta-analysis: the declining risk of colorectal cancer in ulcerative colitis. Aliment Pharmacol Ther **2014**;39:645-59
42. Soderlund S, Brandt L, Lapidus A, Karlen P, Brostrom O, Lofberg R*, et al.* Decreasing time-trends of colorectal cancer in a large cohort of patients with inflammatory bowel disease. Gastroenterology **2009**;136:1561-7; quiz 818-9
43. Herrinton LJ, Liu L, Levin TR, Allison JE, Lewis JD, Velayos F. Incidence and mortality of colorectal adenocarcinoma in persons with inflammatory bowel disease from 1998 to 2010. Gastroenterology **2012**;143:382-9
44. Garg SK, Loftus EV, Jr. Risk of cancer in inflammatory bowel disease: going up, going down, or still the same? Curr Opin Gastroenterol **2016**;32:274-81
45. Loftus EV, Jr. Clinical epidemiology of inflammatory bowel disease: Incidence, prevalence, and environmental influences. Gastroenterology **2004**;126:1504-17

46. Itzkowitz SH, Yio X. Inflammation and cancer IV. Colorectal cancer in inflammatory bowel disease: the role of inflammation. Am J Physiol Gastrointest Liver Physiol **2004**;287:G7-17
47. Levin B. Risk of cancer in ulcerative colitis. Gastrointestinal endoscopy **1999**;49:S60-2
48. Lakatos L, Mester G, Erdelyi Z, David G, Pandur T, Balogh M*, et al.* Risk factors for ulcerative colitis-associated colorectal cancer in a Hungarian cohort of patients with ulcerative colitis: results of a population-based study. Inflammatory bowel diseases **2006**;12:205-11
49. Beaugerie L. Clinical, serological and genetic predictors of inflammatory bowel disease course. World journal of gastroenterology **2012**;18:3806
50. Madanchi M, Zeitz J, Barthel C, Samaras P, Scharl S, Sulz MC*, et al.* Malignancies in Patients with Inflammatory Bowel Disease: A Single-Centre Experience. Digestion **2016**:1-8
51. Zheng HH, Jiang XL. Increased risk of colorectal neoplasia in patients with primary sclerosing cholangitis and inflammatory bowel disease: a meta-analysis of 16 observational studies. Eur J Gastroenterol Hepatol **2016**;28:383-90
52. Brentnall TA, Haggitt RC, Rabinovitch PS, Kimmey MB, Bronner MP, Levine DS*, et al.* Risk and natural history of colonic neoplasia in patients with primary sclerosing cholangitis and ulcerative colitis. Gastroenterology **1996**;110:331-8
53. Shetty K, Rybicki L, Brzezinski A, Carey WD, Lashner BA. The risk for cancer or dysplasia in ulcerative colitis patients with primary sclerosing cholangitis. Am J Gastroenterol **1999**;94:1643-9
54. Askling J, Dickman PW, Karlen P, Brostrom O, Lapidus A, Lofberg R*, et al.* Colorectal cancer rates among first-degree relatives of patients with inflammatory bowel disease: a population-based cohort study. Lancet **2001**;357:262-6
55. Velayos FS, Loftus EV, Jr., Jess T, Harmsen WS, Bida J, Zinsmeister AR*, et al.* Predictive and protective factors associated with colorectal cancer in ulcerative colitis: A case-control study. Gastroenterology **2006**;130:1941-9
56. Jess T, Simonsen J, Jorgensen KT, Pedersen BV, Nielsen NM, Frisch M. Decreasing risk of colorectal cancer in patients with inflammatory bowel disease over 30 years. Gastroenterology **2012**;143:375-81 e1; quiz e13-4
57. Lutgens MW, van Oijen MG, van der Heijden GJ, Vleggaar FP, Siersema PD, Oldenburg B. Declining risk of colorectal cancer in inflammatory bowel disease: an updated meta-analysis of population-based cohort studies. Inflamm Bowel Dis **2013**;19:789-99
58. Brackmann S, Andersen SN, Aamodt G, Roald B, Langmark F, Clausen OP*, et al.* Two distinct groups of colorectal cancer in inflammatory bowel disease. Inflamm Bowel Dis **2008**;15:9-16
59. Brackmann S, Aamodt G, Andersen SN, Roald B, Langmark F, Clausen OP*, et al.* Widespread but not localized neoplasia in inflammatory bowel disease worsens the prognosis of colorectal cancer. Inflamm Bowel Dis **2009**;16:474-81
60. Neumann H, Vieth M, Langner C, Neurath MF, Mudter J. Cancer risk in IBD: how to diagnose and how to manage DALM and ALM. World J Gastroenterol **2011**;17:3184-91
61. Cho KR, Vogelstein B. Genetic alterations in the adenoma--carcinoma sequence. Cancer **1992**;70:1727-31
62. Sprouffske K, Pepper JW, Maley CC. Accurate reconstruction of the temporal order of mutations in neoplastic progression. Cancer Prev Res (Phila) **2011**;4:1135-44

63. Leedham SJ, Graham TA, Oukrif D, McDonald SA, Rodriguez-Justo M, Harrison RF*, et al.* Clonality, founder mutations, and field cancerization in human ulcerative colitis-associated neoplasia. Gastroenterology **2008**;136:542-50.e6
64. Yaeger R, Shah MA, Miller VA, Kelsen JR, Wang K, Heins ZJ*, et al.* Genomic Alterations Observed in Colitis-Associated Cancers Are Distinct From Those Found in Sporadic Colorectal Cancers and Vary by Type of Inflammatory Bowel Disease. Gastroenterology **2016**;151:278-87.e6
65. Robles AI, Traverso G, Zhang M, Roberts NJ, Khan MA, Joseph C*, et al.* Whole-Exome Sequencing Analyses of Inflammatory Bowel Disease-Associated Colorectal Cancers. Gastroenterology **2016**;150:931-43
66. Amersi F, Agustin M, Ko CY. Colorectal cancer: epidemiology, risk factors, and health services. Clinics in colon and rectal surgery **2005**;18:133-40
67. Ou B, Zhao J, Guan S, Lu A. Survival of Colorectal Cancer in Patients With or Without Inflammatory Bowel Disease: A Meta-Analysis. Dig Dis Sci **2015**;61:881-9
68. DuPont AWD, Herbert L. The intestinal microbiota and chronic disorders of the gut. Nature Reviews Gastroenterology and Hepatology **2011**;8:523-31
69. Zhu Y, Michelle Luo T, Jobin C, Young HA. Gut microbiota and probiotics in colon tumorigenesis. Cancer Lett **2011**;309:119-27
70. Arthur JC, Jobin C. The struggle within: microbial influences on colorectal cancer. Inflamm Bowel Dis **2010**;17:396-409
71. O'Keefe SJ. Diet, microorganisms and their metabolites, and colon cancer. Nature reviews Gastroenterology & hepatology **2016**;13:691-706
72. Romano M, F DEF, Zarantonello L, Ruffolo C, Ferraro GA, Zanus G*, et al.* From Inflammation to Cancer in Inflammatory Bowel Disease: Molecular Perspectives. Anticancer Res **2016**;36:1447-60
73. Gerlinger M, McGranahan N, Dewhurst SM, Burrell RA, Tomlinson I, Swanton C. Cancer: evolution within a lifetime. Annu Rev Genet **2014**;48:215-36
74. Harpaz N, Ward SC, Mescoli C, Itzkowitz SH, Polydorides AD. Precancerous lesions in inflammatory bowel disease. Best Pract Res Clin Gastroenterol **2013**;27:257-67
75. Baars JE, Kuipers EJ, van Haastert M, Nicolai JJ, Poen AC, van der Woude CJ. Age at diagnosis of inflammatory bowel disease influences early development of colorectal cancer in inflammatory bowel disease patients: a nationwide, long-term survey. J Gastroenterol **2012**
76. Charpentier C, Salleron J, Savoye G, Fumery M, Merle V, Laberenne JE*, et al.* Natural history of elderly-onset inflammatory bowel disease: a population-based cohort study. Gut **2013**
77. Quezada SM, Cross RK. Association of age at diagnosis and ulcerative colitis phenotype. Dig Dis Sci **2012**;57:2402-7
78. Hou JK, Feagins LA, Waljee AK. Characteristics and Behavior of Elderly-onset Inflammatory Bowel Disease: A Multi-center US Study. Inflammatory Bowel Diseases **2016**
79. Salk JJ, Bansal A, Lai LA, Crispin DA, Ussakli CH, Horwitz MS*, et al.* Clonal Expansions and Short Telomeres Are Associated with Neoplasia in Early-onset, but not Late-onset, Ulcerative Colitis. Inflamm Bowel Dis **2013**
80. Lochhead P, Chan AT, Nishihara R, Fuchs CS, Beck AH, Giovannucci E*, et al.* Etiologic field effect: reappraisal of the field effect concept in cancer predisposition and progression. Mod Pathol **2014**;28:14-29

81. Risques RA, Rabinovitch PS, Brentnall TA. Cancer surveillance in inflammatory bowel disease: new molecular approaches. Curr Opin Gastroenterol **2006**;22:382-90
82. Thorsteinsdottir S, Gudjonsson T, Nielsen OH, Vainer B, Seidelin JB. Pathogenesis and biomarkers of carcinogenesis in ulcerative colitis. Nature reviews Gastroenterology & hepatology **2011**;8:395-404
83. Chen R, Lai LA, Brentnall TA, Pan S. Biomarkers for colitis-associated colorectal cancer. World J Gastroenterol **2016**;22:7882-91
84. Brentnall TA, Crispin DA, Rabinovitch PS, Haggitt RC, Rubin CE, Stevens AC*, et al.* Mutations in the p53 gene: an early marker of neoplastic progression in ulcerative colitis. Gastroenterology **1994**;107:369-78
85. Burmer GC, Rabinovitch PS, Haggitt RC, Crispin DA, Brentnall TA, Kolli VR*, et al.* Neoplastic progression in ulcerative colitis: histology, DNA content, and loss of a p53 allele. Gastroenterology **1992**;103:1602-10
86. Lashner BA, Shapiro BD, Husain A, Goldblum JR. Evaluation of the usefulness of testing for p53 mutations in colorectal cancer surveillance for ulcerative colitis. Am J Gastroenterol **1999**;94:456-62
87. Chen R, Rabinovitch PS, Crispin DA, Emond MJ, Bronner MP, Brentnall TA. The initiation of colon cancer in a chronic inflammatory setting. Carcinogenesis **2005**;26:1513-9
88. Porschen R, Robin U, Schumacher A, Schauseil S, Borchard F, Hengels KJ*, et al.* DNA aneuploidy in Crohn's disease and ulcerative colitis: results of a comparative flow cytometric study. Gut **1992**;33:663-7
89. Lindberg JO, Stenling RB, Rutegard JN. DNA aneuploidy as a marker of premalignancy in surveillance of patients with ulcerative colitis. Br J Surg **1999**;86:947-50
90. Rubin CE, Haggitt RC, Burmer GC, Brentnall TA, Stevens AC, Levine DS*, et al.* DNA aneuploidy in colonic biopsies predicts future development of dysplasia in ulcerative colitis. Gastroenterology **1992**;103:1611-20
91. Holzmann K, Klump B, Borchard F, Gregor M, Porschen R. Flow cytometric and histologic evaluation in a large cohort of patients with ulcerative colitis: correlation with clinical characteristics and impact on surveillance. Dis Colon Rectum **2001**;44:1446-55
92. International HIVCS, Pereyra F, Jia X, McLaren PJ, Telenti A, de Bakker PI*, et al.* The major genetic determinants of HIV-1 control affect HLA class I peptide presentation. Science **2010**;330:1551-7
93. Aust DE, Willenbucher RF, Terdiman JP, Ferrell LD, Chang CG, Moore DH, 2nd*, et al.* Chromosomal alterations in ulcerative colitis-related and sporadic colorectal cancers by comparative genomic hybridization. Hum Pathol **2000**;31:109-14
94. Bronner MP, Skacel M, Crispin DA, Hoff PD, Emond MJ, Lai LA*, et al.* Array Comparative Genomic Hybridization in Ulcerative Colitis Neoplasia: Single Non-Dysplastic Biopsies Distinguish Progressors from Non-Progressors. Mod Pathol **2010**;23:1624-33
95. Lai LA, Risques RA, Bronner MP, Rabinovitch PS, Crispin D, Chen R*, et al.* Pan-colonic field defects are detected by CGH in the colons of UC patients with dysplasia/cancer. Cancer Lett **2012**;320:180-8
96. Kang H, Shibata D. Direct measurements of human colon crypt stem cell niche genetic fidelity: the role of chance in non-darwinian mutation selection. Front Oncol **2013**;3:264
97. Martincorena I, Luscombe NM. Non-random mutation: the evolution of targeted hypermutation and hypomutation. Bioessays **2013**;35:123-30

98. Preston SL, Wong WM, Chan AO, Poulsom R, Jeffery R, Goodlad RA*, et al.* Bottom-up histogenesis of colorectal adenomas: origin in the monocryptal adenoma and initial expansion by crypt fission. Cancer Res **2003**;63:3819-25
99. Boyer JC, Yamada NA, Roques CN, Hatch SB, Riess K, Farber RA. Sequence dependent instability of mononucleotide microsatellites in cultured mismatch repair proficient and deficient mammalian cells. Hum Mol Genet **2002**;11:707-13
100. Salk JJ, Salipante SJ, Risques RA, Crispin DA, Li L, Bronner MP*, et al.* Clonal expansions in ulcerative colitis identify patients with neoplasia. Proc Natl Acad Sci U S A **2009**;106:20871-6
101. O'Sullivan J, Risques RA, Mandelson MT, Chen L, Brentnall TA, Bronner MP*, et al.* Telomere length in the colon declines with age: a relation to colorectal cancer? Cancer Epidemiol Biomarkers Prev **2006**;15:573-7
102. Risques RA, Lai LA, Himmetoglu C, Ebaee A, Li L, Feng Z*, et al.* Ulcerative colitis-associated colorectal cancer arises in a field of short telomeres, senescence, and inflammation. Cancer Res **2011**;71:1669-79
103. Brentnall TA. Molecular underpinnings of cancer in ulcerative colitis. Curr Opin Gastroenterol **2003**;19:64-8
104. Friis-Ottessen M, Bendix L, Kolvraa S, Norheim-Andersen S, De Angelis PM, Clausen OP. Telomere shortening correlates to dysplasia but not to DNA aneuploidy in longstanding ulcerative colitis. BMC Gastroenterol **2014**;14:8
105. O'Sullivan JN, Bronner MP, Brentnall TA, Finley JC, Shen WT, Emerson S*, et al.* Chromosomal instability in ulcerative colitis is related to telomere shortening. Nat Genet **2002**;32:280-4
106. Risques RA, Lai LA, Brentnall TA, Li L, Feng Z, Gallaher J*, et al.* Ulcerative colitis is a disease of accelerated colon aging: evidence from telomere attrition and DNA damage. Gastroenterology **2008**;135:410-8
107. Kinouchi Y, Hiwatashi N, Chida M, Nagashima F, Takagi S, Maekawa H*, et al.* Telomere shortening in the colonic mucosa of patients with ulcerative colitis. J Gastroenterol **1998**;33:343-8
108. Tahara T, Shibata T, Okubo M, Kawamura T, Sumi K, Ishizuka T*, et al.* Telomere length in non-neoplastic colonic mucosa in ulcerative colitis (UC) and its relationship to the severe clinical phenotypes. Clin Exp Med **2014**;15:327-32
109. Novak EA, Mollen KP. Mitochondrial dysfunction in inflammatory bowel disease. Frontiers in cell and developmental biology **2015**;3:62
110. Chatterjee A, Dasgupta S, Sidransky D. Mitochondrial subversion in cancer. Cancer Prev Res (Phila) **2011**;4:638-54
111. Sanchez-Arago M, Chamorro M, Cuezva JM. Selection of cancer cells with repressed mitochondria triggers colon cancer progression. Carcinogenesis **2010**;31:567-76
112. Larman TC, DePalma SR, Hadjipanayis AG, Cancer Genome Atlas Research N, Protopopov A, Zhang J*, et al.* Spectrum of somatic mitochondrial mutations in five cancers. Proc Natl Acad Sci U S A **2012**;109:14087-91
113. Yu M. Somatic mitochondrial DNA mutations in human cancers. Adv Clin Chem **2012**;57:99-138
114. Ward PS, Thompson CB. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. Cancer Cell **2012**;21:297-308
115. Smolkova K, Plecita-Hlavata L, Bellance N, Benard G, Rossignol R, Jezek P. Waves of gene regulation suppress and then restore oxidative phosphorylation in cancer cells. The international journal of biochemistry & cell biology **2010**;43:950-68

116. Zong WX, Rabinowitz JD, White E. Mitochondria and Cancer. Mol Cell **2016**;61:667-76
117. Ju YS, Alexandrov LB, Gerstung M, Martincorena I, Nik-Zainal S, Ramakrishna M*, et al.* Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. Elife **2014**;3
118. Pereira L, Soares P, Maximo V, Samuels DC. Somatic mitochondrial DNA mutations in cancer escape purifying selection and high pathogenicity mutations lead to the oncocytic phenotype: pathogenicity analysis of reported somatic mtDNA mutations in tumors. BMC Cancer **2012**;12:53
119. Stewart JB, Alaei-Mahabadi B, Sabarinathan R, Samuelsson T, Gorodkin J, Gustafsson CM*, et al.* Simultaneous DNA and RNA Mapping of Somatic Mitochondrial Mutations across Diverse Human Cancers. PLoS Genet **2015**;11:e1005333
120. Nishikawa M, Oshitani N, Matsumoto T, Nishigami T, Arakawa T, Inoue M. Accumulation of mitochondrial DNA mutation with colorectal carcinogenesis in ulcerative colitis. Br J Cancer **2005**;93:331-7
121. Brentnall TA, Pan S, Bronner MP, Crispin DA, Mirzaei H, Cooke K*, et al.* Proteins That Underlie Neoplastic Progression of Ulcerative Colitis. Proteomics Clin Appl **2009**;3:1326
122. Chen R, Pan S, Lai K, Lai LA, Crispin DA, Bronner MP*, et al.* Up-regulation of mitochondrial chaperone TRAP1 in ulcerative colitis associated colorectal cancer. World J Gastroenterol **2014**;20:17037-48
123. May D, Pan S, Crispin DA, Lai K, Bronner MP, Hogan J*, et al.* Investigating neoplastic progression of ulcerative colitis with label-free comparative proteomics. J Proteome Res **2010**;10:200-9
124. Santhanam S, Rajamanickam S, Motamarry A, Ramakrishna BS, Amirtharaj JG, Ramachandran A*, et al.* Mitochondrial electron transport chain complex dysfunction in the colonic mucosa in ulcerative colitis. Inflamm Bowel Dis **2012**;18:2158-68
125. Santhanam S, Venkatraman A, Ramakrishna BS. Impairment of mitochondrial acetoacetyl CoA thiolase activity in the colonic mucosa of patients with ulcerative colitis. Gut **2007**;56:1543-9
126. Sifroni KG, Damiani CR, Stoffel C, Cardoso MR, Ferreira GK, Jeremias IC*, et al.* Mitochondrial respiratory chain in the colonic mucosal of patients with ulcerative colitis. Mol Cell Biochem **2010**;342:111-5
127. Ussakli CH, Ebaee A, Binkley J, Brentnall TA, Emond MJ, Rabinovitch PS*, et al.* Mitochondria and Tumor Progression in Ulcerative Colitis. Journal of the National Cancer Institute **2013**
128. Greaves LC, Preston SL, Tadrous PJ, Taylor RW, Barron MJ, Oukrif D*, et al.* Mitochondrial DNA mutations are established in human colonic stem cells, and mutated clones expand by crypt fission. Proc Natl Acad Sci U S A **2006**;103:714-9
129. Cunningham KE, Vincent G, Sodhi CP, Novak EA, Ranganathan S, Egan CE*, et al.* Peroxisome Proliferator-activated Receptor-gamma Coactivator 1-alpha (PGC1alpha) Protects against Experimental Murine Colitis. J Biol Chem **2016**;291:10184-200
130. Sahin E, Colla S, Liesa M, Moslehi J, Muller FL, Guo M*, et al.* Telomere dysfunction induces metabolic and mitochondrial compromise. Nature **2011**;470:359-65
131. Sahin E, DePinho RA. Axis of ageing: telomeres, p53 and mitochondria. Nat Rev Mol Cell Biol **2012**;13:397-404
132. Chiba T, Marusawa H, Ushijima T. Inflammation-associated cancer development in digestive organs: mechanisms and roles for genetic and epigenetic modulation. Gastroenterology **2012**;143:550-63

133. Gloria L, Cravo M, Pinto A, de Sousa LS, Chaves P, Leitao CN*, et al.* DNA hypomethylation and proliferative activity are increased in the rectal mucosa of patients with long-standing ulcerative colitis. Cancer **1996**;78:2300-6
134. Gerceker E, Boyacioglu SO, Kasap E, Baykan A, Yuceyar H, Yildirim H*, et al.* Never in mitosis gene A-related kinase 6 and aurora kinase A: New gene biomarkers in the conversion from ulcerative colitis to colorectal cancer. Oncol Rep **2015**;34:1905-14
135. Hahn MA, Li AX, Wu X, Yang R, Drew DA, Rosenberg DW*, et al.* Loss of the polycomb mark from bivalent promoters leads to activation of cancer-promoting genes in colorectal tumors. Cancer Res **2014**;74:3617-29
136. Issa JP, Ahuja N, Toyota M, Bronner MP, Brentnall TA. Accelerated age-related CpG island methylation in ulcerative colitis. Cancer Res **2001**;61:3573-7
137. Kim BN, Yamamoto H, Ikeda K, Damdinsuren B, Sugita Y, Ngan CY*, et al.* Methylation and expression of p16INK4 tumor suppressor gene in primary colorectal cancer tissues. Int J Oncol **2005**;26:1217-26
138. Kang K, Bae JH, Han K, Kim ES, Kim TO, Yi JM. A Genome-Wide Methylation Approach Identifies a New Hypermethylated Gene Panel in Ulcerative Colitis. International journal of molecular sciences **2016**;17
139. Garrity-Park MM, Loftus EV, Jr., Bryant SC, Smyrk TC. A Biomarker Panel to Detect Synchronous Neoplasm in Non-neoplastic Surveillance Biopsies from Patients with Ulcerative Colitis. Inflammatory bowel diseases **2016**;22:1568-74
140. Jones PA, Baylin SB. The epigenomics of cancer. Cell **2007**;128:683-92
141. Ventham NT, Kennedy NA, Nimmo ER, Satsangi J. Beyond gene discovery in inflammatory bowel disease: the emerging role of epigenetics. Gastroenterology **2013**;145:293-308
142. Luo Y, Yu M, Grady WM. Field cancerization in the colon: a role for aberrant DNA methylation? Gastroenterology report **2014**;2:16-20
143. Hsieh JC, Van Den Berg D, Kang H, Hsieh CL, Lieber MR. Large chromosome deletions, duplications, and gene conversion events accumulate with age in normal human colon crypts. Aging Cell **2013**;12:269-79
144. Greaves LC, Barron MJ, Plusa S, Kirkwood TB, Mathers JC, Taylor RW*, et al.* Defects in multiple complexes of the respiratory chain are present in ageing human colonic crypts. Exp Gerontol **2010**;45:573-9
145. Greaves LC, Elson JL, Nooteboom M, Grady JP, Taylor GA, Taylor RW*, et al.* Comparison of mitochondrial mutation spectra in ageing human colonic epithelium and disease: absence of evidence for purifying selection in somatic mitochondrial DNA point mutations. PLoS Genet **2012**;8:e1003082
146. Kaz AM, Wong CJ, Dzieciatkowski S, Luo Y, Schoen RE, Grady WM. Patterns of DNA methylation in the normal colon vary by anatomical location, gender, and age. Epigenetics **2014**;9:492-502
147. Snippert HJ, Schepers AG, van Es JH, Simons BD, Clevers H. Biased competition between Lgr5 intestinal stem cells driven by oncogenic mutation induces clonal expansion. EMBO Rep **2013**;15:62-9
148. von Zglinicki T. Oxidative stress shortens telomeres. Trends Biochem Sci **2002**;27:339-44
149. d'Adda di Fagagna F, Reaper PM, Clay-Farrace L, Fiegler H, Carr P, Von Zglinicki T*, et al.* A DNA damage checkpoint response in telomere-initiated senescence. Nature **2003**;426:194-8

150. Shay JW. Role of Telomeres and Telomerase in Aging and Cancer. Cancer Discov **2016**;6:584-93
151. DePinho RA. The age of cancer. Nature **2000**;408:248-54
152. Fornaro R, Caratto M, Caratto E, Caristo G, Fornaro F, Giovinazzo D*, et al.* Colorectal Cancer in Patients With Inflammatory Bowel Disease: The Need for a Real Surveillance Program. Clin Colorectal Cancer **2016**;15:204-12
153. Mooiweer E, van der Meulen-de Jong AE, Ponsioen CY, van der Woude CJ, van Bodegraven AA, Jansen JM*, et al.* Incidence of Interval Colorectal Cancer Among Inflammatory Bowel Disease Patients Undergoing Regular Colonoscopic Surveillance. Clin Gastroenterol Hepatol **2015**;13:1656-61
154. Molodecky NA, Soon IS, Rabi DM, Ghali WA, Ferris M, Chernoff G*, et al.* Increasing incidence and prevalence of the inflammatory bowel diseases with time, based on systematic review. Gastroenterology **2012**;142:46-54 e42; quiz e30
155. Farraye FA, Odze RD, Eaden J, Itzkowitz SH. AGA technical review on the diagnosis and management of colorectal neoplasia in inflammatory bowel disease. Gastroenterology **2010**;138:746-74, 74.e1-4; quiz e12-3
156. Yashiro M. Ulcerative colitis-associated colorectal cancer. World J Gastroenterol **2014**;20:16389-97
157. Zisman TL, Bronner MP, Rulyak S, Kowdley KV, Saunders M, Lee SD*, et al.* Prospective study of the progression of low-grade dysplasia in ulcerative colitis using current cancer surveillance guidelines. Inflamm Bowel Dis **2012**
158. Ullman TA. Colonoscopic surveillance in inflammatory bowel disease. Curr Opin Gastroenterol **2005**;21:585-8
159. Wu L, Li P, Wu J, Cao Y, Gao F. The diagnostic accuracy of chromoendoscopy for dysplasia in ulcerative colitis: meta-analysis of six randomized controlled trials. Colorectal Dis **2010**;14:416-20
160. Subramanian V, Mannath J, Ragunath K, Hawkey CJ. Meta-analysis: the diagnostic yield of chromoendoscopy for detecting dysplasia in patients with colonic inflammatory bowel disease. Aliment Pharmacol Ther **2010**;33:304-12
161. Marion JF, Waye JD, Israel Y, Present DH, Suprun M, Bodian C*, et al.* Chromoendoscopy Is More Effective Than Standard Colonoscopy in Detecting Dysplasia During Long-term Surveillance of Patients With Colitis. Clin Gastroenterol Hepatol **2015**;14:713-9
162. Soetikno R, Kaltenbach T, McQuaid KR, Subramanian V, Kumar R, Barkun AN*, et al.* Paradigm Shift in the Surveillance and Management of Dysplasia in Inflammatory Bowel Disease (West). Digestive endoscopy : official journal of the Japan Gastroenterological Endoscopy Society **2016**;28:266-73
163. Muir LN, B. Peto's paradox and the hallmarks of cancer: constructing an evolutionary framework for understanding the incidence of cancer. **2015**
164. Foo J, Leder K, Ryser MD. Multifocality and recurrence risk: a quantitative model of field cancerization. J Theor Biol **2014**;355:170-84
165. Dhawan A, Graham TA, Fletcher AG. A Computational Modeling Approach for Deriving Biomarkers to Predict Cancer Risk in Premalignant Disease. Cancer Prev Res (Phila) **2016**;9:283-95
166. Curtius K, Wong CJ, Hazelton WD, Kaz AM, Chak A, Willis JE*, et al.* A Molecular Clock Infers Heterogeneous Tissue Age Among Patients with Barrett's Esophagus. PLoS Comput Biol **2016**;12:e1004919

167. Ryser MD, Lee WT, Ready NE, Leder KZ, Foo J. Quantifying the Dynamics of Field Cancerization in Tobacco-Related Head and Neck Cancer: A Multiscale Modeling Approach. Cancer Res **2016**;76:7078-88
168. Spira A, Yurgelun MB, Alexandrov L, Rao A, Bejar R, Polyak K*, et al.* Precancer Atlas to Drive Precision Prevention Trials. Cancer Res **2017**;77:1510-41

Figure Legends

Figure 1. Proposed model of carcinogenesis in UC This model integrates the etiological, molecular, and dysplastic field effects with known cellular and molecular events that contribute to the different stages of carcinogenesis in UC. The arrow indicates the temporal direction of dysplastic progression and the color gradient reflects the increasing risk of cancer at the different stages of the process. Cancer arises within dysplastic and/or molecular fields. Molecular fields precede dysplastic fields and are more extensive, thus offering an excellent opportunity for precancer detection.

Figure 2. Implications of the field effect on UC colonoscopic surveillance By analyzing several screening biopsies procured along the colon, field effects can be identified and could possibly be used to predict cancer progression risk. The colon diagrams represent four clinical scenarios in which dysplasia may not be detected by chromoendoscopy but in which screening biopsies could identify molecular fields: A) pancolonic field, B) multifocal fields, C) extensive field, D) localized field. Cancer is a probabilistic process that depends on the rate of mutation and the number of cells at risk. Thus, we postulate that large, multifocal fields are likely to be associated with higher risk of cancer progression compared to small, unifocal fields. Studies with large numbers of patients are needed to validate this model, quantify the progression risk accordingly, and integrate it with other known cancer risk factors in UC. If such approach was developed, it would facilitate tailoring the time interval between colonoscopies to the level of risk for each patient.

Figures

Figure 1. Proposed model of carcinogenesis in UC

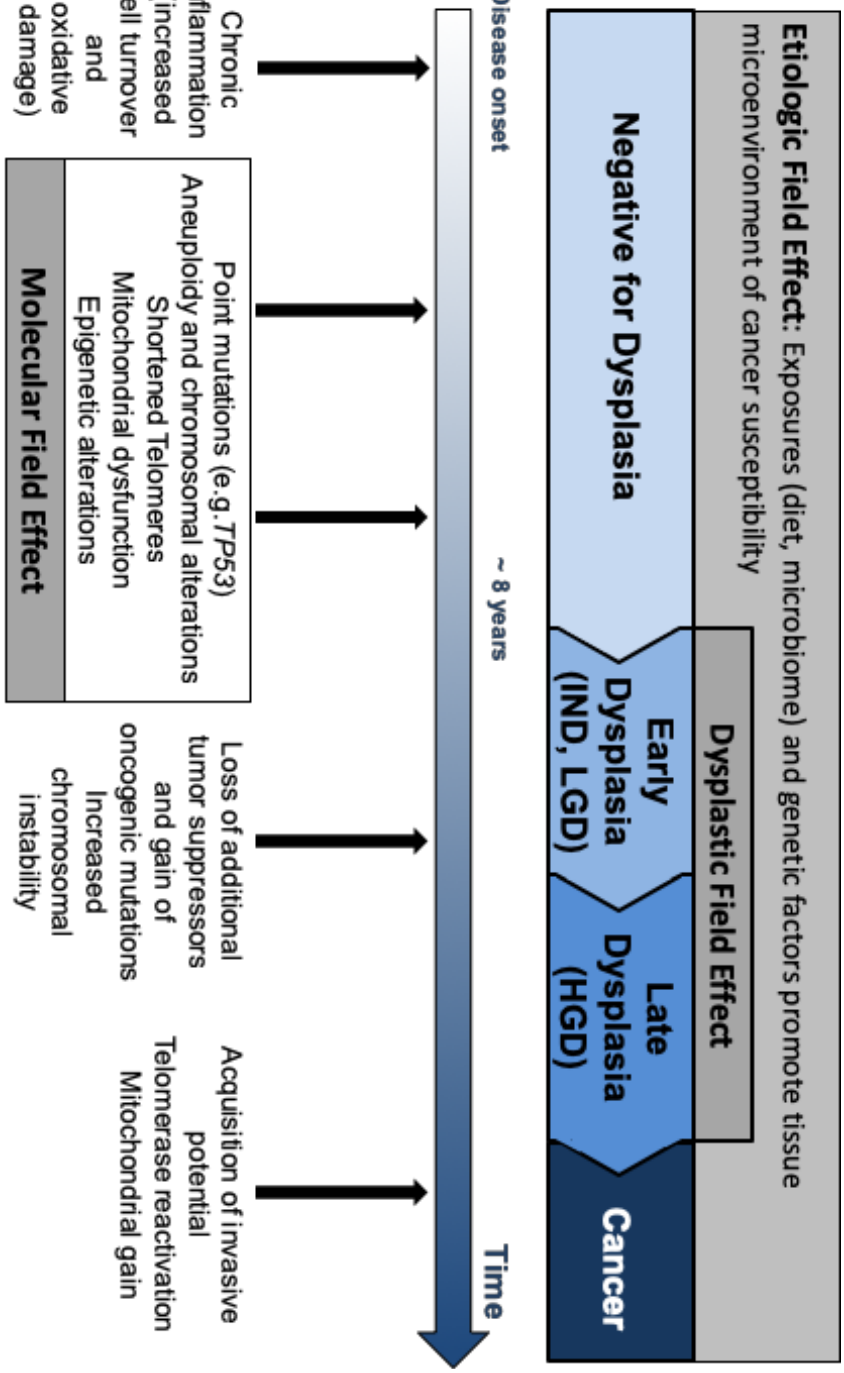
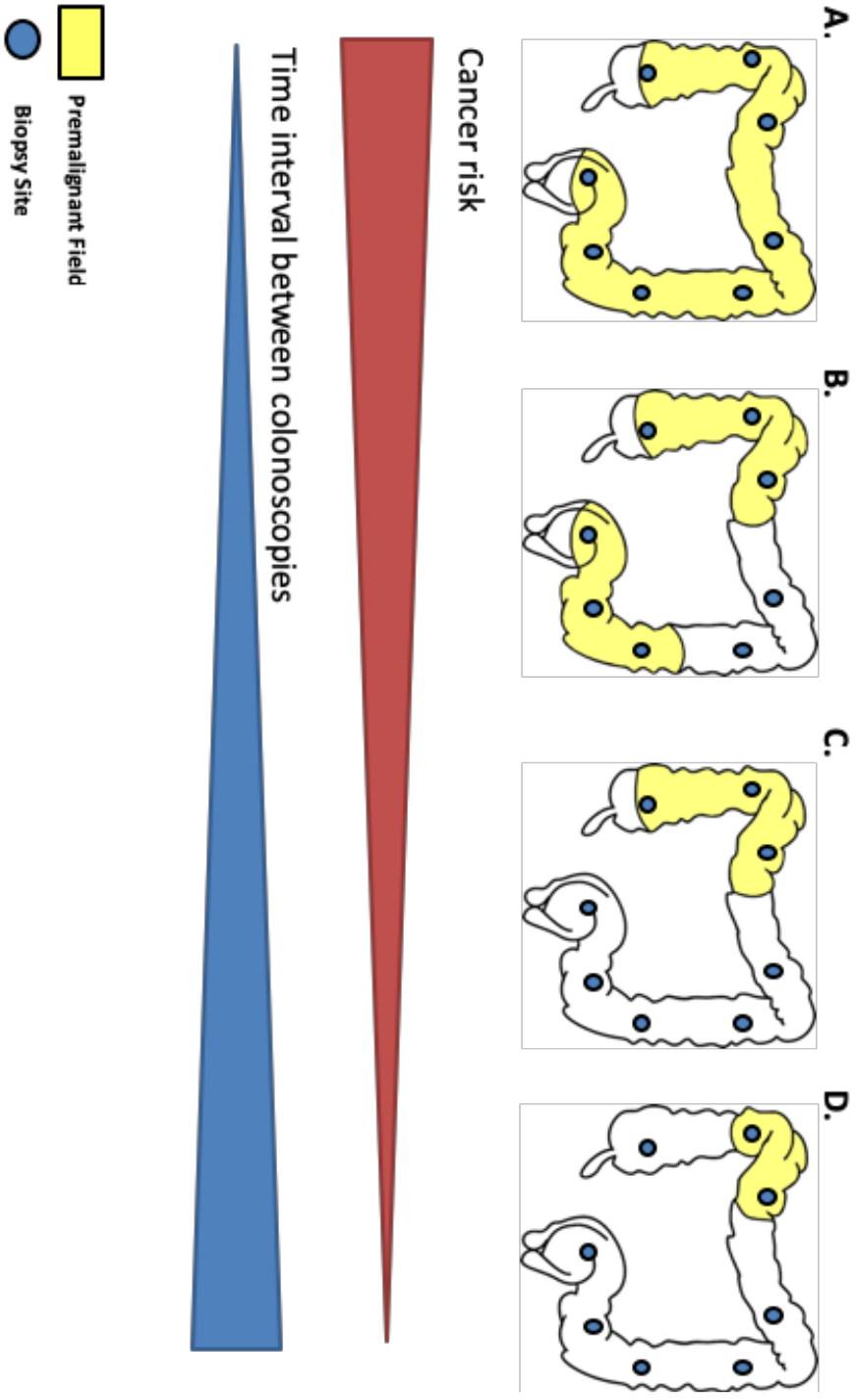


Figure 2. Implications of the field effect on UC colonoscopic surveillance



Chapter 2: Mitochondrial DNA mutations are associated with ulcerative colitis preneoplasia but tend to be negatively selected in cancer

Kathryn T. Baker[1], Daniela Nachmanson[1], Shilpa Kumar[1], Mary J. Emond[2], Cigdem Ussakli[1, 3*], Teresa A. Brentnall[4], Scott R. Kennedy[1], Rosa Ana Risques[1]

[1]Department of Pathology, University of Washington, Seattle, WA 98195, USA. [2]Department of Biostatistics, University of Washington, Seattle, WA 98195, USA. [3]Department of Laboratory Medicine, University of Washington, Seattle, WA 98195, USA. [4]Division of Gastroenterology, Department of Medicine, University of Washington, Seattle, WA, 98195, USA
*Current address: PhenoPath, Seattle, WA 98103

Running title: Mitochondrial DNA mutations in ulcerative colitis carcinogenesis

Keywords: Mitochondrial DNA, Ulcerative Colitis, Inflammatory Bowel Disease, Duplex Sequencing, Next-Generation Sequencing

Financial support: This work supported by NIH Grants R01CA181308 (RAR) and R01CA160674 (TAB). KTB was a recipient of the predoctoral fellowship from the Molecular Medicine Predoctoral Training Program, NIH T32GM95421.

Conflict of interest: SRK is a consultant and equity holder for TwinStrand Biosciences Inc. RAR shares equity in NanoString Technologies Inc. and is the principal investigator on a NIH SBIR R44CA221426 subcontract research agreement with TwinStrand Biosciences Inc.

Abbreviation list
CRC – colorectal cancer
DCS – double strand consensus sequence
DS – Duplex Sequencing
FFPE – formalin-fixed paraffin-embedded
GEE – general estimating equations
HGD – high-grade
LGD – low-grade
MAF – mutant allele frequency
mtDNA – mitochondrial DNA
NP – Non-Progressor
P – Progressor
UC – ulcerative colitis
VLF – very low frequency

Published as:
Mitochondrial DNA mutations are associated with ulcerative colitis preneoplasia but tend to be negatively selected in cancer. Kathryn T Baker, Daniela Nachmanson, Shilpa Kumar, Mary J Emond, Cigdem Ussakli, Teresa A Brentnall, Scott R Kennedy and Rosa Ana Risques
Mol Cancer Res November 16 2018 **DOI:** 10.1158/1541-7786.MCR-18-0520

Abstract

The role of mitochondrial DNA (mtDNA) mutations in cancer remains controversial. Ulcerative Colitis (UC) is an inflammatory bowel disease that increases the risk of colorectal cancer and involves mitochondrial dysfunction, making it an ideal model to study the role of mtDNA in tumorigenesis. Our goal was to comprehensively characterize mtDNA mutations in UC tumorigenesis using Duplex Sequencing, an ultra-accurate next generation sequencing method. We analyzed 46 colon biopsies from non-UC and UC patients with and without cancer, including biopsies at all stages of dysplastic progression. mtDNA was sequenced at a median depth of 1,364x. Mutations were classified by mutant allele frequency: clonal >0.95, subclonal 0.01-0.95, and very low frequency (VLF)<0.01. We identified 208 clonal and subclonal mutations and 56,764 VLF mutations. Mutations were randomly distributed across the mitochondrial genome. Clonal and subclonal mutations increased in number and pathogenicity in early dysplasia but decreased in number and pathogenicity in cancer. Most clonal, subclonal, and VLF mutations were C>T transitions in the heavy strand of mtDNA, which likely arise from DNA replication errors. A subset of VLF mutations were C>A transversions, which are probably due to oxidative damage. VLF transitions and indels were less abundant in the non D-loop region and decreased with progression. Our results indicate that mtDNA mutations are frequent in UC preneoplasia but negatively selected in cancers.

Implications: While mitochondrial DNA mutations might contribute to early UC tumorigenesis, they appear to be selected against in cancer, suggesting that functional mitochondria might be required for malignant transformation in UC.

Introduction

While the role of nuclear DNA mutations in cancer has been extensively characterized, the contribution of mitochondrial DNA (mtDNA) mutations to carcinogenesis remains unclear. For some time, the prevailing hypothesis has been that mtDNA mutations contribute to tumor progression by impairing oxidative phosphorylation and promoting aerobic glycolysis, a feature of cancer cells known as the Warburg effect (1-3). Mounting evidence, however, has challenged this idea by revealing that cancer cells rely on oxidative phosphorylation and functional mitochondria for ATP production and rapid cell growth (4,5). Recent studies also demonstrate that mtDNA mutations accumulate randomly and clonally expand without selective pressure or, if deleterious, they are selected against (6,7). These results call into question a driving role of mtDNA mutations in tumor progression and their contribution to the Warburg effect.

Ulcerative colitis (UC) is an inflammatory bowel disease that serves as an excellent model for studying mtDNA mutations in preneoplastic progression. UC causes chronic inflammation of the colonic epithelium and affected patients have an elevated risk for colorectal cancer (CRC) (8-10). Tumorigenesis in this disease follows a distinct pattern of progression from negative for dysplasia (Neg) to low-grade dysplasia (LGD), high-grade dysplasia (HGD), and finally cancer. In patients that develop colorectal cancer, molecular alterations are found not only in dysplastic tissue but in histologically normal tissue surrounding dysplasia (11-13) indicating the presence of a field effect, or field cancerization (9,14). These premalignant fields offer a unique opportunity to study the early molecular events that contribute to tumor progression as well as their evolution across all dysplastic stages into malignancy.

Mitochondrial dysfunction has been demonstrated in UC (15), but there is conflicting literature regarding its contribution to cancer progression (14). The conflict might arise from the fact that mitochondrial alterations could play different roles in early and late disease. Using

cytochrome C oxidase subunit I (COXI) immunohistochemistry, our group previously reported mitochondrial loss in premalignant lesions but a recovery of normal levels of mitochondria in cancer (16). Based on these observations we hypothesized that while dysfunctional mitochondria might contribute to early dysplasia, functional mitochondria are essential at the cancer stage. Further, this dual pattern might be mirrored in mtDNA mutations, with an increase of mutations in early dysplasia followed by negative selection of mtDNA mutations in cancer.

Previous studies of the role of mtDNA in cancer have used next-generation sequencing (NGS) technologies to analyze mutations (6,7,17,18). However, conventional NGS has an error rate of 1 in 100-1000 (19), which precludes the accurate detection of mutations with mutant allele frequency (MAF) <0.01 (20). The detection of very low frequency mtDNA mutations is essential to characterize the underlying mutagenic processes as well as to detect small clones that might arise during carcinogenesis. Thus, in this study we have utilized a double-strand molecular tagging method called Duplex Sequencing (DS) (Fig. 1A, B), which performs error-correction by scoring only mutations found on both strands of DNA independently (21). The estimated error rate is less than 1 in 10 million, which enables the identification of mutations at frequencies as low as 0.0001 (21,22).

Here we have applied this highly accurate technology to identify the presence of mutations in mtDNA with high confidence. Our goal was to uncover the underlying mechanism of mtDNA mutagenesis in UC and to clarify the role of these mutations in cancer progression. We analyzed the mtDNA of 46 colon biopsies at all histological stages of progression and detected thousands of mutations. We characterized these mutations by frequency, location, type, pathogenicity, and mutational context, thus producing a comprehensive, high-resolution analysis of mtDNA mutations in preneoplastic progression.

Materials and Methods

Patients and Biopsies

The study included 10 patients: 7 with UC and 3 non-UC controls. Four of the patients with UC had progressed to HGD or cancer (Progressors) and the remaining 3 were cancer and dysplasia free (Non-Progressors, NP) (Table 1, Supplementary Table S1 and Supplementary Methods). Fresh frozen samples were collected at colectomy (UC patients) or colonoscopy (controls) in accordance with Human Subjects Guidelines and the appropriate Institutional Review Board at the University of Washington. A total of 46 colon biopsies were analyzed from these patients, including 36 biopsies that represented all histological grades in UC Progressors (Fig. 1C, Supplementary Table S1 and Supplementary Fig. S1). Thus, we considered six biopsy types in total: normal, NP, Neg, LGD, HGD and cancer (the last four corresponding to biopsies from Progressors). The biopsies from UC Progressors were selected based on the colon maps generated upon colectomy (Fig. 1C) with the criteria of covering different histological grades and different areas of progression. Formalin- fixed paraffin embedded (FFPE) biopsies adjacent to the frozen biopsies used for analysis were stained with hematoxylin and eosin and examined under a light microscope for acute inflammation (cryptitis and the presence of neutrophils in the epithelium) and chronic inflammation (lymphocytes in the lamina propria). For acute inflammation, scores were assigned the following numeric equivalents: none – 1, mild – 2, moderate – 3. For chronic inflammation, scores were assigned the following numeric equivalents: none – 1, low – 2, high – 3. Epithelial isolation and DNA extraction were performed as part of prior studies via EDTA shake-off, which yields ~90% enrichment for epithelial cells (13,23)(Supplementary Methods).

Duplex Sequencing

For each sample, between 50-150ng of colonic epithelium DNA were processed for DS of mtDNA as previously described (21,24) (Fig. 1A, B). DNA was end repaired, A-tailed, and ligated

to DS adapters (IDT, Coralville, IA, USA) (Supplementary Methods). To determine the optimal input of ligated DNA for amplification, samples were qPCR amplified with a DS adapter specific primer (MWS13, 5'- AATGATACGGCGACCACCGAG-3') and a primer from an internal mitochondrial sequence (MitoRev, 5'-GCGCTTACTTTGTAGCCTTCA-3') (both by IDT) and titrated against a standard DNA sample. DNA was then captured using the NimbleGen SeqCap Target Enrichment kit (Roche, Basel, Switzerland) or the xGen Lockdown Target Enrichment kit (IDT) with probes specific for the mitochondrial genome. Samples were indexed, pooled, and sequenced using 2x100 bp paired-end reads on the Illumina HiSeq 2500 or 2x150 bp paired-end reads on the Illumina NextSeq 550.

Data Processing

Raw data files were processed as in previous studies (24,25) (https://github.com/risqueslab/DuplexSequencingScripts) with some modifications. First, consensus making was performed prior to the alignment of reads. Second, paired read information was retained. Finally, duplex consensus sequence (DCS) reads were aligned using BWA-MEM with default parameters (bio-bwa.sourceforge.net) to a version of human reference genome v37 (GRCh37) (ncbi.nlm.nih.gov/grc/human) according to the revised Cambridge reference, which corrects for an error at base 3107 in previous versions. The Genome Analysis Toolkit (GATK) version 3.6 (software.broadinstitute.org/gatk) Indel-Realigner was used to perform local realignment of each mapped read. GATK Clip-Reads was used to clip 10bp from both the 5' and 3' end of each read to remove low quality reads and artifacts created during end repair and A-tailing. DCS reads with more than 5% indeterminate bases (Ns) were removed. Indeterminate bases occur when there is no consensus. Positions with less than 100 DCS reads were not considered for analysis. The fgbio (https://github.com/fulcrumgenomics/fgbio) tool ClipOverlappingReads was then used to clip any overlapping bases from paired reads.

All samples were sequenced to an average depth of at least 600X. The frequency of Ns was calculated for each position along the genome. For each sample, positions with N≥0.1 were excluded from analysis, but this never represented more than 0.5% of the mtDNA positions. The haplotype of each patient was identified with the Haplogrep tool (http://haplogrep.uibk.ac.at). To stratify the frequency of mutational events, clonality cutoffs were established based on MAF; very low frequency (VLF) mutations MAF<0.01; subclonal mutations, MAF≥0.01 and <0.95; and clonal mutations MAF≥0.95. Clonal/subclonal mutations represent different degrees of clonal expansion within the colonic tissue. In contrast, VLF mutations could represent small clones or unique *de novo* events, since they were often supported by a single mutated DCS read. Of note, mutations identified in a single DCS read have very low probability of being artefactual ($<10^{-7}$) (20) because they are independently identified in the two complementary strands of DNA and are produced by the consensus of at least 6 raw reads (3 for each DNA strand). Thus, VLF mtDNA mutations capture the ongoing mutagenic processes at the molecular level as well as small clonal expansions while clonal/subclonal mutations quantify large clonal expansions.

Clonal and Subclonal Mutation Analysis

Clonal and subclonal mutations were analyzed jointly and compared across the spectrum of biopsy types in the study, e.g.: normal, NP, Neg, LGD, HGD, and cancer. Mutations found in all samples from a given colon and with >75% of samples having a mutation frequency ≥0.80 were considered constitutional to the patient and removed from consideration. For colons where only one sample was analyzed, mutations with a frequency ≥0.99 that were commonly identified polymorphisms in the human population were also considered constitutional and thus removed. Clonal and subclonal mutation location was visualized using the Circlize package in R (https://CRAN.R-project.org/package=circlize). Clonal and subclonal mutations were compared across samples based on D-loop mutation frequency, clonality, number of mutations per biopsy,

pathogenicity based on MitImpact (26) and mutational signature (Supplementary Methods). The mutational signature analysis was based on the substitution rate, which calculated the number of observed mutations of each type (e.g. C>A, C>G, C>T, T>A, T>C, T>G) in each mtDNA strand and divided it by the number of expected mutations assuming equal probability for all substitutions.

Very Low Frequency Mutation Analysis

Mutations with a MAF<0.01 were considered VLF. Different mutations identified at the same nucleotide position were independently counted. Similar to clonal/subclonal mutations, VLF mutations were compared across the 6 biopsy types in the study. However, there were 3 major differences in the analysis. First, to calculate the frequency of MAF<0.01 mutations in each biopsy, the number of mutations was divided by the total amount of mtDNA nucleotides sequenced in each biopsy. This was critical to correct for sequencing depth because higher depth results in finding more VLF mutations. Second, to calculate the frequency of each mutation type, the number of mutations for each possible nucleotide substitution was divided by the number of times that nucleotide was sequenced in each given sample. This takes into consideration the depth of sequencing of each sample, as well as the nucleotide composition of the mtDNA. This calculation was done separately for mutations in the D-loop and non D-loop. Third, due to the much larger number of VLF mutations than clonal/subclonal mutations, the mutational signature analysis could be performed taking into consideration not only the 6 possible nucleotide substitutions in the heavy and light strand of DNA, but also the trinucleotide context of each substitution, for a total of 96 substitution types in each strand.

Statistical Analysis

To account for the possibility of correlation between observations from the same individual (or biopsy), we applied the method of generalized estimating equations (GEE). However, GEE relies on large-sample theory for the validity of the estimates, particularly the standard error estimates. Because the sample size here is modest, we also applied resampling with GEE (see Supplementary Methods).

Results

Duplex Sequencing identifies abundant mtDNA mutations in UC biopsies

Mutations in mtDNA were identified by performing DS on DNA extracted from colonic epithelium from 46 biopsies covering different stages of preneoplastic and neoplastic progression (Table 1). Samples were sequenced at a median depth of 1,364x with a minimum depth of ~600x (Supplementary Table S2). Because DS enables ultra-accurate deep sequencing (21), we were able to detect and classify mtDNA mutations in 3 groups according to their MAF: clonal ≥0.95; subclonal ≥0.01 and <0.95, and very low frequency (VLF) <0.01. We used Haplogrep2 (haplogrep.uibk.ac.at) to identify each patient's haplotype (Supplementary Table S1), which allowed us to discount haplotype specific polymorphisms and constitutional polymorphisms. In total we identified 208 clonal/subclonal mutations, and 56,764 VLF mutations (Table 1).

Clonality increases with progression

The overall distribution of clonal and subclonal mutations across the mitochondrial genome as well as their MAF is shown in Fig. 2A. While most mutations were low frequency ($0.01<MAF<0.1$), a subset of mutations appeared at larger frequencies ($MAF>0.1$). The proportion of these large frequency mutations as well as their MAF increased with progression (Fig. 2B),

consistent with larger clones progressively expanding during tumorigenesis. A detailed analysis of these mutations revealed that in the colons from UC Progressors, some mutations were shared at different frequencies in adjacent and relatively distant biopsies (~25cm), often spanning colonic epithelium of different histological grades (Supplementary Fig. S2). These findings confirm the clonal nature of the expansions and the presence of large fields of cancerization in UC (14). The analysis of individual biopsies (Supplementary Fig. S3A) indicated that the majority of UC Progressor biopsies (29/36 = 80.5%) harbored a clonal expansion in which a mtDNA mutation was present at MAF>0.1, whereas these expansions were less frequent in colon from UC Non-Progressors or non-UC colon (1/10 = 10%) ($p=9x10^{-5}$ by Fisher's exact test). Importantly, the number of mutations within both the MAF>0.1 and MAF>0.01 categories did not correlate with the total amount of DCS nucleotides sequenced (Supplementary Fig. S4A, indicating that differences in sequencing depth did not explain the variation in number of subclonal mutations observed across biopsies. Clonal/subclonal mutations were slightly higher in older UC patients (Supplementary Fig. S5A) however, at all ages, they were more frequent in UC Progressors than in Non Progressors. There were no associations between clonal and subclonal mutations and sex, disease duration, active inflammation, and chronic inflammation (Fig. S5B-E). Within histological grades, the number of mutations was not associated to inflammation scores (Supplementary Fig. S5F).

Clonal and subclonal mutations are randomly distributed in the coding region but tend to cluster in the D-loop with advanced disease

Clonal and subclonal mutations appeared randomly distributed across the mtDNA coding region (Fig. 2A), an observation that was confirmed by plotting the number of mutations in each mtDNA encoded gene sorted by ascending size (Fig. 2C). Larger genes had more mutations and no significant clustering by gene was observed (p=0.36 by χ^2 test of homogeneity). The proportion

of D-loop mutations, however, increased with progression (Fig. 2D). The D-loop is a non-coding region that represents 6.7% of the mitochondrial genome, but as much as 19%, 14%, and 26% of clonal/subclonal mutations in LGD, HGD, and cancer, respectively, were found in the D-loop. Mutations in tRNA and rRNA did not significantly change with progression, but the percentage of mutations in the coding region decreased in cancers. Individual analysis of all the biopsies in the study confirmed that these results were not driven by a single biopsy or by biopsies from a single colon (Supplementary Fig. S3B). These results suggest that mtDNA mutations in the coding region are selected against in UC cancer progression.

Clonal and subclonal mutations display a mutational signature indicative of mtDNA replication errors

Previous studies have demonstrated that most mtDNA mutations that accumulate with aging and cancer correspond to C>T transitions that occur almost exclusively in the heavy strand of the mtDNA (6,7,27). These mutations are attributed to mtDNA replication errors. To determine whether the same mutational mechanisms are operative in the inflammatory setting of UC, we quantified the mutation substitution rate for each of the six possible mutation types in each of the two strands of mtDNA. The mutation substitution rate was calculated as the ratio of the number of observed mutations divided by the number of expected mutations. Indeed, C>T transitions in the heavy strand were the most predominant type of mutation across all 6 biopsy types, observed between 8 to 16-fold times more than what would be expected by chance (Fig. 2E). Of note, clonal and subclonal mutations did not show a significant contribution from C>A transversions, the signature of oxidative damage.

The number of clonal/subclonal mutations spikes in early stages of progression but decreases in later stages

To better characterize the role of mtDNA mutations in UC clonal expansions, we performed a detailed analysis of the number, MAF, and mutational consequence of mtDNA mutations by biopsy type (Supplementary Fig. S6). We observed that normal colon biopsies had low frequency subclonal mutations that were either non-coding or synonymous. UC Non-Progressors also featured low frequency subclonal mutations, but they were often non-synonymous. The number and the frequency of mutations dramatically increased in negative for dysplasia biopsies from UC Progressors compared to normal and Non-Progressors. However, with advanced progression, the proportion of mutations with high MAF increased (Fig. 2B) but the overall number of mutations appeared to decrease. To better quantify this finding, we compared the mean number of mtDNA mutations for each biopsy type (Fig. 3A). While biopsies from normal and UC Non-Progressor colon only harbored, on average, about 2 mtDNA mutations (MAF>0.01), this number increased to 5 and 6 in biopsies from UC Progressors negative for dysplasia and LGD, respectively. However, the number of mutations decreased in HGD and even more in cancers. This decrease was statistically significant ($p=0.014$ for linear effect over LGD, HGD, CA; $p = 3.6 \times 10^{-5}$ for a quadratic effect (inverse V-shape) over all biopsy types, by GEE permutation tests). Overall these results indicate that (1) clonal expansions that carry mtDNA mutations are a feature of UC preneoplastic progression, (2) the maximum number of mutations is achieved in LGD and decreases in HGD and cancer, showing an inverse V-shape that is in agreement with prior findings of mitochondrial alterations in UC (16).

Clonal and subclonal mutations are enriched for non-synonymous and pathogenic mutations in LGD but not in cancer

We next quantified the frequency of non-synonymous mutations for each biopsy type (Fig. 3B). Interestingly, for all UC biopsy types except LGD the frequency of non-synonymous mutations was less than 71%, which is the expected frequency given the composition of the mitochondrial genome. For LGD, however, the frequency was 81%. While the test for a decreasing linear trend from LGD to HGD to cancer was not significant (p=0.11), there was a nominally significant difference when comparing the frequency for LGD, 81%, to non-synonymous frequency over all other grades, 62% (p=0.026, n=109 total mutations). These results suggest that damaging mtDNA mutations might be positively selected in LGD, but they appear to be selected against at other stages.

While non-synonymous mutations are a first indication of potential for pathogenicity, they often lead to amino acid changes that are inconsequential. Thus, a better estimate of pathogenicity can be achieved by utilizing computational algorithms to predict the functional impact of a specific missense variant. To comprehensively address this issue, we used MitImpact 2.9 (mitimpact.css-mendel.it) (26), which is a collection of pre-computed pathogenicity predictions for all possible nucleotide changes that cause non-synonymous substitutions in human mitochondrial protein coding genes. We interrogated six different algorithms (Polyphen2, Fathmmw, CADD, Mutation Assessor, SIFT, and Provean) that categorized missense clonal and subclonal mutations into different pathogenicity groups. Two of the algorithms, Polyphen2 (28) and FatHmmW (29), identified significant differences with progression (Fig. 3C and 3D, p=0.025 and p=0.006 for decrease in pathogenicity from LGD to CA by GEE permutation tests, respectively). The six algorithms measure different aspects of pathogenicity using different mathematical approaches and, thus, they vary in their predictions. Polyphen2 predicts structural and functional impact of missense mutations using a probabilistic classifier whereas FatHmmW predicts functional impact

by combining sequence conservation with hidden Markov models. Interestingly, both algorithms showed increased pathogenicity in early stages of progression and decreased pathogenicity in cancer (Fig. 3C and 3D). These findings complement our previous observations based on number of mutations (Fig. 3A) and frequency of non-synonymous mutations (Fig. 3B). Overall, these data indicate that the clones in early progression tend to carry more mtDNA mutations and these are more pathogenic. However, the clones that eventually evolve to cancer tend to carry mutations that are not coding or non-pathogenic, suggesting selection against deleterious mtDNA mutations.

VLF mutations display mutational signatures corresponding to mtDNA replication errors and oxidative damage

In contrast to clonal and subclonal mutations, VLF mutations were very abundant in all biopsies (Table 1) and their number was highly associated with the total amount of sequenced nucleotides (Supplementary Fig. S4B). Thus, to compare between biopsies we calculated the VLF mutation frequency as the number of VLF mutations divided by the total DCS nucleotide sequenced. A subset of biopsies showed a disproportionately large number of VLF mutations, which corresponded mostly to C>A transversions, the signature caused by oxidative damage (Supplementary Fig. S7). All the biopsies from UC Non-Progressors, the normal biopsy with Hirschprung disease, and 7/10 biopsies from one of the UC Progressors harbored a high frequency of C>A mutations in both the heavy and the light strand of mtDNA (Supplementary Fig. S7). VLF mutations were not associated with age, sex, disease duration, acute inflammation or chronic inflammation (Supplementary Fig. S8A-E). Importantly, within Non-Progressors, the high level of C>A mutations was not associated with higher levels of inflammation in those biopsies (Supplementary Fig. S8F).

To further investigate the mutational signatures operative in VLF mutations, we analyzed the trinucleotide context in which each of the 6 possible substitutions occurred in the heavy or light strand of the mtDNA. This analysis generates 96 possible mutational events (6 substitutions x 16 flanking nucleotide combinations) and has been extensively used to elucidate mutagenic processes in both nuclear and mitochondrial tumor DNA (6,7,30). The combined analysis of all samples revealed two overlapping mutational signatures (Fig. 4): 1) C>A transversions in both strands of DNA and independent of nucleotide context; and 2) C>T transitions in the heavy strand of DNA and T>C transitions in the light strand of DNA, both with markedly increased frequency in certain trinucleotide contexts. Specifically, C>T in the heavy strand were enriched in NpCpG contexts and T>C in the light strand were enriched in NpTpC contexts. These mutational events correspond to the ones previously identified in mitochondrial DNA from cancer samples (6,7) and have been attributed to DNA replication errors. Mutational signature analysis by biopsy type (Supplementary Fig. S9) demonstrated that the mtDNA "replication error" signature is not exclusive to cancers but is also found in preneoplastic biopsies and normal colon. In biopsies from UC Non-Progressors, the signature was also present but overshadowed by an excess of C>A transversions (Supplementary Fig. S9).

VLF transitions and indels are more common in the D-loop than non D-loop and decrease with progression

Because of the prominent role of C>A mutations in some biopsies, the comparison of mutation frequency between biopsies and within D-loop and non D-loop regions was best performed by separating transitions and transversions (Fig. 5). We observed that in Non-Progressors, not only transversions were disproportionally high, so were transitions, pointing to an excessive mutational load beyond oxidative damage. For all biopsy types, the frequency of

transitions was lower in the non-D-loop region than in the D-loop (Fig. 5A, mean difference=$7.8x10^{-6}$, $p<1x10^{-9}$). Remarkably, in the non-D-loop, C>T transitions in the heavy strand and T>C transitions in the light strand, which correspond to the predominant mtDNA mutational signature, significantly decreased with progression (Supplementary Fig. S10C, $p=6.6x10^{-4}$). Compared to transitions, transversions (Fig. 5B) displayed a smaller difference in frequencies in the D-loop and non-D-loop over all biopsy types (mean difference=$2.3x10^{-6}$, p=0.04) and did not show any changes with progression. Indels (Fig. 5C) showed a similar pattern to transitions, presenting at much higher frequency in D-loop than in non-D-loop ($p<1x10^{-6}$). They were highest in NP and decreased with progression both in the non-D-loop (p=0.0017) as well as in the D-loop (p=0.014).

To further investigate the mutational pattern by biopsy type within the D-loop and non D-loop region, we separated transitions and transversions into the corresponding nucleotide substitutions in each strand of DNA (Supplementary Fig. S10). This analysis allowed us to determine that the C>T and T>C strand biases were exclusive of the non D-loop region, in agreement with prior findings in aging and cancer (6,27) (Supplementary Fig. S10A-C). Transitions in the D-loop had no strand bias and did not change in frequency with progression. However, transitions in the non-D-loop were strongly biased according to the "replication error" signature previously described (Fig. 4) and sharply declined with progression. In contrast, transversions, which were almost exclusively C>A, were found at similar frequencies in the heavy and light strand and in the D-loop and non-D-loop region (Supplementary Fig. S10D-F), consistent with the widespread effect of oxidative damage.

VLF mutations are randomly distributed in the coding region and tend to be enriched for synonymous mutations during progression

The mean frequency of non-synonymous and synonymous mutations was constant across all genes indicating that mutations accumulated randomly, without any detectable clustering by gene (Fig. 6A). The same was true when tested for each biopsy type (Supplementary Fig. S11) indicating no preferential incidence of VLF mutations in any given gene during progression. Regarding the percentage of non-synonymous mutations, we observed a decreasing trend during progression ($p=0.017$, Fig. 6B), although there was substantial variability within biopsy type.

Discussion

The contribution of mtDNA mutations to tumorigenesis has been an area of controversy for some time. The work presented here helps to explain this contribution in the context of UC associated colorectal tumorigenesis: mtDNA mutations increase in early UC carcinogenesis, but appear to be selected against in cancer. Previous studies of mtDNA mutations have been limited by the sensitivity issues inherent to standard NGS (20) and few have been able to detect mutations with MAF<0.1 (6). The accuracy of DS allowed us to obtain reliable estimates of MAF ranging from 1 down to 0.0005. This provided a comprehensive analysis of mtDNA mutations with progression because we could accurately quantify not only the number of mutations but their clonality. In addition, because VLF mutations are extremely frequent in the mitochondrial genome, in spite of the relatively low number of biopsies in the study, we identified thousands of mutations that enabled us to perform detailed mutational signature analyses.

The main finding of our study is the selection against mtDNA mutations in cancer compared to early stages of progression, which was supported by multiple lines of evidence. In cancers, we observed: 1) fewer mtDNA mutations, both subclonal and VLF, 2) decreased

proportion of distinct subclonal and VLF mutations (transitions and indels) in the coding region, 3) fewer non-synonymous mutations, 4) fewer subclonal pathogenic mutations. Although our findings are derived from UC-associated cancer, they are in agreement with a prior report of decreased mtDNA mutagenesis in sporadic colorectal cancer (18) and with the detailed mutational analysis of mtDNA from TCGA data, which demonstrated negative selection of deleterious mitochondrial mutations in cancers (6,7). Collectively, these results support the notion that cancer cells require functional mitochondria. This notion is consistent with a novel view of mitochondria as essential organelles in cancer (4,5), which not only supply energy and intermediate metabolites, but are critical to enable the metabolic reprogramming characteristic of cancer cells (31).

A limitation of our study is the small number of UC patients. However, multiple biopsies were included from each patient and a large number of mutations were analyzed, enabling a detailed characterization of the mutational profile of mtDNA in UC tumorigenesis. Importantly, our results are in agreement with our previous work in UC. We previously demonstrated the widespread loss of mitochondrial function in UC via immunohistochemical staining for mitochondrial proteins (16). In UC Progressors, we identified a v-shaped pattern with maximum mitochondrial loss in LGD and a recovery of normal levels in cancer. This pattern was confirmed by mtDNA copy number quantification (16). Based on these findings, we hypothesized an initial increase and a later decrease in the burden of mtDNA mutations over the course of dysplastic progression. Our results have now confirmed this hypothesis, strongly suggesting that, while mitochondrial dysfunction might be associated with the earlier stages of the disease, cancer cells tend to feature functional mitochondria. Several non-exclusive mechanisms are possible: 1) mitochondria with damaged DNA might be removed by autophagy and mitochondrial biogenesis activated via PGC1α (16); 2) a genetic bottleneck might be bypassed only by premalignant cells

with intact mitochondria; or 3) cells with damaged mtDNA might acquire whole functional mitochondria by horizontal transfer from neighboring tissue (32).

We detected signs of positive selection for mtDNA mutations in LGD, including enrichment for nonsynonymous and pathogenic mutations. Others have reported mtDNA mutations in precancerous lesions, suggesting a potential contribution to early transformation (1). It is well known that the carcinogenic process in UC is histologically and genetically different from sporadic colorectal cancer (14) and is possible for mitochondria to play differential roles in these processes. However, it is remarkable that in sporadic colorectal carcinogenesis, mtDNA mutations have also been observed to increase in adenomas and decrease in colon cancer (18), in agreement with our data. Thus, it appears that the opposite role of mtDNA mutations in early and late cancer might occur in sporadic carcinogenesis as well and might explain some of the contradictions in the field.

Regarding the causes of mtDNA mutations, our results support mtDNA replication as the major mechanism of mutation, in agreement with previous results from cancer (6,7,33) and aging (27). The resemblance of the mutational signature reported here to those reported by Ju *et al.* is striking (6) and indicates that the same mutational processes operative in the mitochondria of tumors take place in the mitochondria of normal, inflamed, and preneoplastic colon. Although the exact mechanism of mutagenesis is unknown, based on the exclusive non D-loop location of the strand bias (also observed here), Ju *et al.* proposed three explanations: a) the parent heavy strand might be more prone to cytosine and adenine deamination while being single-stranded during replication, b) endogenous POLG errors might occur on the leading strand preferentially, and c) different repair mechanisms might be at play in the leading vs. lagging strand (6). A major difference with Ju *et al.* is that in a subset of samples in our study we did observe an important contribution from oxidative damage, although only at the level of VLF mutations. In the presence

of reactive oxygen species (ROS), guanine oxidizes to 8-oxo-guanine, which results in C>A transversions (34). Oxidative damage is an important pathogenic factor in UC (35,36), and, thus, we expected to observe this mutational signature. Surprisingly, C>A mutations were not widespread among all biopsies from UC patients, but were restricted to biopsies from UC Non-Progressors, most biopsies (negative, LGD and cancer) from a single Progressor patient, and the non-UC colon that had Hirschprung disease. These results were not explained by variability in inflammation levels since C>A mutations, as well as mtDNA mutations in general, were not associated with acute or chronic inflammation. However, these results might be explained by interindividual or interregional variations in the generation of ROS or in the production of antioxidant defenses, an interesting hypothesis that deserves further investigation with a larger number of patients.

In the context of UC Progressors, a critical finding is the presence of large clonal expansions in non-dysplastic epithelium. mtDNA mutations with MAF>0.1 were abundant in non-dysplastic biopsies from UC Progressors but were rare in Non-Progressors. While these clonal expansions might be driven by a pathogenic mtDNA mutation that confers a selective advantage, in many cases mtDNA mutations might be carried as passengers and arrive to homoplasmy by genetic drift (6,37). In any case, these mutations could be used as markers of clonal expansions, which are an essential component of preneoplastic fields in UC (9,14). This study was not designed to assess differences between Progressors and Non-Progressors and the number of cases in each group is insufficient to make these group comparisons. However, we have previously demonstrated the potential value of clonal expansions to detect UC cancer progression (23,38) and the utility of mtDNA for this purpose warrants further investigation.

Our findings support a model in which mtDNA mutations accumulate and clonally expand in early tumorigenesis but are subject to purifying selection in cancer (Supplementary Fig. S12).

During normal aging, mtDNA mutations accumulate and clonally expand in the colon epithelium (39), but this process might be accelerated in UC due to the increased cellular proliferation necessary to regenerate the ulcerated epithelium. This increased cellular proliferation would lead to extensive replication of the mtDNA, which appears to be the main cause of mutation not only in UC tumorigenesis, but also in most cancers (6). Cells with pathogenic mtDNA mutations might clonally expand in early progression, leading to multiple small clones. However, progression to malignancy appears to be characterized by a decrease in the number and pathogenicity of mtDNA mutations, possibly due to the outgrowth of one or few clones carrying non-pathogenic mtDNA mutations that drift to homoplasmy. Further research is necessary to elucidate the role of mitochondrial epigenetic regulation and metabolic reprogramming during this process (16) and to determine to what extent this model is applicable to other cancer types.

Author's Contributions

Conception and design: KT Baker, RA Risques

Development of methodology: KT Baker, D Nachmanson, S Kumar, SR Kennedy, RA Risques

Acquisition of data (provided animals, acquired and managed patients, provided facilities, etc.): C Ussakli, TA Brentnall

Analysis and interpretation of data (e.g. statistical analysis, biostatistics, computational analysis): KT Baker, D Nachmanson, S Kumar, MJ Emond, RA Risques

Writing, review, and/or revision of the manuscript: KT Baker, RA Risques

Administrative, technical, or material support (i.e., reporting or organizing data, constructing databases): KT Baker, RA Risques

Study supervision: RA Risques

Data Access

Sequencing data that supports the findings of this study have been deposited in the Sequence Read Archive (SRA: SRP139857, BioProject ID: PRJNA449763).

Acknowledgements

We thank Jesse J. Salk and Jeffrey D. Krimmel for their preliminary contributions to this work, Jake G. Hoekstra and Monica Sanchez-Contreras for their advice and expertise for analyzing mitochondrial DNA, Kelly Jin for her assistance with data visualization, and Rebecca Ortega for her helpful comments and suggestions.

References

1. Chatterjee A, Dasgupta S, Sidransky D. Mitochondrial subversion in cancer. Cancer Prev Res (Phila) **2011**;4:638-54
2. Larman TC, DePalma SR, Hadjipanayis AG, Cancer Genome Atlas Research N, Protopopov A, Zhang J*, et al.* Spectrum of somatic mitochondrial mutations in five cancers. Proc Natl Acad Sci U S A **2012**;109:14087-91
3. Yu M. Somatic mitochondrial DNA mutations in human cancers. Adv Clin Chem **2012**;57:99-138
4. Zong WX, Rabinowitz JD, White E. Mitochondria and Cancer. Mol Cell **2016**;61:667-76
5. Wallace DC. Mitochondria and cancer. Nat Rev Cancer **2012**;12:685-98
6. Ju YS, Alexandrov LB, Gerstung M, Martincorena I, Nik-Zainal S, Ramakrishna M*, et al.* Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. Elife **2014**;3
7. Stewart JB, Alaei-Mahabadi B, Sabarinathan R, Samuelsson T, Gorodkin J, Gustafsson CM*, et al.* Simultaneous DNA and RNA Mapping of Somatic Mitochondrial Mutations across Diverse Human Cancers. PLoS Genet **2015**;11:e1005333
8. Hanauer SB. Inflammatory bowel disease: epidemiology, pathogenesis, and therapeutic opportunities. Inflamm Bowel Dis **2006**;12 Suppl 1:S3-9
9. Choi CR, Bakir IA, Hart AL, Graham TA. Clonal evolution of colorectal cancer in IBD. Nature reviews Gastroenterology & hepatology **2017**
10. Dyson JK, Rutter MD. Colorectal cancer in inflammatory bowel disease: what is the real magnitude of the risk? World J Gastroenterol **2012**;18:3839-48
11. Brentnall TA, Crispin DA, Rabinovitch PS, Haggitt RC, Rubin CE, Stevens AC*, et al.* Mutations in the p53 gene: an early marker of neoplastic progression in ulcerative colitis. Gastroenterology **1994**;107:369-78
12. Rabinovitch PS, Dziadon S, Brentnall TA, Emond MJ, Crispin DA, Haggitt RC*, et al.* Pancolonic chromosomal instability precedes dysplasia and cancer in ulcerative colitis. Cancer Res **1999**;59:5148-53
13. Risques RA, Lai LA, Himmetoglu C, Ebaee A, Li L, Feng Z*, et al.* Ulcerative colitis-associated colorectal cancer arises in a field of short telomeres, senescence, and inflammation. Cancer Res **2011**;71:1669-79
14. Baker KT, Salk JJ, Brentnall TA, Risques RA. Precancer in ulcerative colitis: the role of the field effect and its clinical implications. Carcinogenesis **2018**;39:11-20
15. Novak EA, Mollen KP. Mitochondrial dysfunction in inflammatory bowel disease. Frontiers in cell and developmental biology **2015**;3:62
16. Ussakli CH, Ebaee A, Binkley J, Brentnall TA, Emond MJ, Rabinovitch PS*, et al.* Mitochondria and Tumor Progression in Ulcerative Colitis. Journal of the National Cancer Institute **2013**
17. He Y, Wu J, Dressman DC, Iacobuzio-Donahue C, Markowitz SD, Velculescu VE*, et al.* Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. Nature **2010**;464:610-4
18. Ericson NG, Kulawiec M, Vermulst M, Sheahan K, O'Sullivan J, Salk JJ*, et al.* Decreased mitochondrial DNA mutagenesis in human colorectal cancer. PLoS Genet **2012**;8:e1002689
19. Goodwin S, McPherson JD, McCombie WR. Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet **2016**;17:333-51

20. Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat Rev Genet **2018**;19:269-85
21. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A **2012**;109:14508-13
22. Krimmel JD, Schmitt MW, Harrell MI, Agnew KJ, Kennedy SR, Emond MJ*, et al.* Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. Proc Natl Acad Sci U S A **2016**;113:6005-10
23. Salk JJ, Salipante SJ, Risques RA, Crispin DA, Li L, Bronner MP*, et al.* Clonal expansions in ulcerative colitis identify patients with neoplasia. Proc Natl Acad Sci U S A **2009**;106:20871-6
24. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH*, et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc **2014**;9:2586-606
25. Nachmanson D, Shenyi L, Schmidt EK, Hipp MJ, Baker KT, Zhang Y*, et al.* Targeted genome fragmentation with CRISPR/Cas9 enables fast and efficient enrichment of small genomic regions and ultra-accurate sequencing with low DNA input (CRISPR-DS). Genome Res **2018**;10:1589-1599
26. Castellana S, Ronai J, Mazza T. MitImpact: an exhaustive collection of pre-computed pathogenicity predictions of human mitochondrial non-synonymous variants. Hum Mutat **2015**;36:E2413-22
27. Kennedy SR, Salk JJ, Schmitt MW, Loeb LA. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet **2013**;9:e1003794
28. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P*, et al.* A method and server for predicting damaging missense mutations. Nat Methods. United States **2010**. p 248-9.
29. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ*, et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat **2013**;34:57-65
30. Alexandrov LB, Stratton MR. Mutational signatures: the patterns of somatic mutations hidden in cancer genomes. Curr Opin Genet Dev **2014**;24:52-60
31. Hanahan D, Weinberg RA. Hallmarks of cancer: the next generation. Cell **2011**;144:646-74
32. Dong LF, Kovarova J, Bajzikova M, Bezawork-Geleta A, Svec D, Endaya B*, et al.* Horizontal transfer of whole mitochondria restores tumorigenic potential in mitochondrial DNA-deficient cancer cells. Elife **2017**;6
33. Ahn EH, Lee SH, Kim JY, Chang CC, Loeb LA. Decreased Mitochondrial Mutagenesis during Transformation of Human Breast Stem Cells into Tumorigenic Cells. Cancer Res **2016**;76:4569-78
34. Delaney S, Jarem DA, Volle CB, Yennie CJ. Chemical and biological consequences of oxidatively damaged guanine in DNA. Free radical research **2012**;46:420-41
35. Roessner A, Kuester D, Malfertheiner P, Schneider-Stock R. Oxidative stress in ulcerative colitis-associated carcinogenesis. Pathol Res Pract **2008**;204:511-24
36. Jena G, Trivedi PP, Sandala B. Oxidative stress in ulcerative colitis: an old concept but a new concern. Free radical research **2012**;46:1339-45
37. Coller HA, Bodyak ND, Khrapko K. Frequent intracellular clonal expansions of somatic mtDNA mutations: significance and mechanisms. Ann N Y Acad Sci **2002**;959:434-47

38. Salk JJ, Bansal A, Lai LA, Crispin DA, Ussakli CH, Horwitz MS*, et al.* Clonal Expansions and Short Telomeres Are Associated with Neoplasia in Early-onset, but not Late-onset, Ulcerative Colitis. Inflamm Bowel Dis **2013**
39. Greaves LC, Preston SL, Tadrous PJ, Taylor RW, Barron MJ, Oukrif D*, et al.* Mitochondrial DNA mutations are established in human colonic stem cells, and mutated clones expand by crypt fission. Proc Natl Acad Sci U S A **2006**;103:714-9

Tables

Table 1. Study Design and Mutation Counts

Patient Type	Number of Patients	Dysplastic Grade	Number of Biopsies	Number of mtDNA Mutations	
				MAF>0.01	MAF<0.01
Normal	**3**	Negative	3	5	6790
UC Non-Progressor	**3**	Negative	7	14	25495
UC Progressor	**4**	Negative	15	77	7384
		Low Grade	9	59	5026
		High Grade	4	22	5417
		Cancer	8	31	6652
TOTAL	**10**		46	208	56764

Figure Legends

Figure 1. Experimental design A, Schematic of the Duplex Sequencing method. Duplex Sequencing adapters contain a molecular tag consisting of a randomized nucleotide sequence (represented as α (cyan) and β(orange)) and two universal priming sites (purple and green). DNA is fragmented (yellow) and ligated to Duplex Sequencing adapters. **B,** Reads from the same strand of a DNA molecule are used to produce a single-strand consensus sequence (SSCS). Then the two complementary SSCS generated from the same original DNA molecule are condensed into a double-strand consensus sequence (DCS). Only mutations found on the two complementary SSCS are considered true mutations. **C,** Colon maps for each of the Progressor patients as diagrammed by the pathologist after colectomy. Each box corresponds to an individual biopsy and is color-coded according to histological findings: Neg: negative for dysplasia, IND: indefinite for dysplasia, LGD: low-grade dysplasia, HGD: high-grade dysplasia, and cancer. Biopsies are named based on the coordinates defined by columns (letters) and rows (numbers). Columns correspond to the diameter of the colon divided into 3-4 sections and are ~2cm apart. Rows indicate colon levels and are evenly spaced ~2-5cm along the length of the organ. For each colon, the biopsies analyzed are indicated with a box with the biopsy name.

Figure 2. Clonal and subclonal mtDNA mutations A, Genomic positions of clonal and subclonal mtDNA mutations. Biopsy types are color-coded. MAF from 0.01-1.0 is indicated by position along the vertical axis. **B,** Distribution of clonal and subclonal mutations by MAF in each biopsy type. **C,** Association between number of mutations and gene size (bp). Biopsy types are color-coded. **D,** Distribution of clonal and subclonal mutations by region compared to the expected distribution based on composition of the mitochondrial genome. **E,** Substitution rate for

clonal and subclonal mutations shown by biopsy type. MAF: mutant allele frequency, NP: Non-Progressor, Neg: negative for dysplasia, LGD: low-grade dysplasia, HGD: high-grade dysplasia.

Figure 3. Comparison of clonal and subclonal mutations by biopsy type

A, Mean number of clonal and subclonal mtDNA mutations for each biopsy type. Error bars indicate standard error of the mean. p-value corresponds to the linear effect by GEE permutation tests. **B,** The proportion of synonymous and non-synonymous clonal and subclonal mutations is shown by biopsy type. Dashed line indicates expected percentage of non-synonymous mutations in the mtDNA in the absence of selection. **C, D** Proportion of clonal and subclonal mutations of each biopsy type predicted to have pathogenic impact by Polyphen 2 (C) and FatHmmW (D) algorithms. p-values were calculated using GEE permutation tests.

Figure 4. Very low frequency (MAF<0.01) mtDNA mutational signature The substitution rate of each of the 96 possible substitution classes is shown for mtDNA mutations with MAF<0.01. Substitution rate is calculated as the ratio of the number of observed mutations to the number of expected mutations. Black arrows indicate CpG and TpC trinucleotides, which are amongst the most frequently mutated in the heavy strand and light strand of mtDNA, respectively. MAF: mutant allele frequency.

Figure 5. Quantification of very low frequency mutations (MAF<0.01) in D-loop *vs.* non-D-loop and by progression The total frequency of mutations for variants with MAF<0.01 is shown for transitions (**A**), transversions (**B**), and indels (**C**) for each biopsy type and by D-Loop or Non D-Loop regions. p-values were calculated using GEE permutation tests.

Figure 6. Very low frequency (MAF<0.01) mutation selection A, Non-synonymous and synonymous mutation frequency for each mtDNA encoded gene plotted by gene size. Mutation frequency was calculated as the number of mutations with MAF<0.01 (non-synonymous or synonymous) within each given gene divided by the total number of DCS nucleotides sequenced. **B,** Proportion of non-synonymous mutations (MAF<0.01) for each biopsy by histological grade. p-values were calculated using a GEE permutation test. MAF: mutant allele frequency.

Figures

Figure 1. Experimental Design

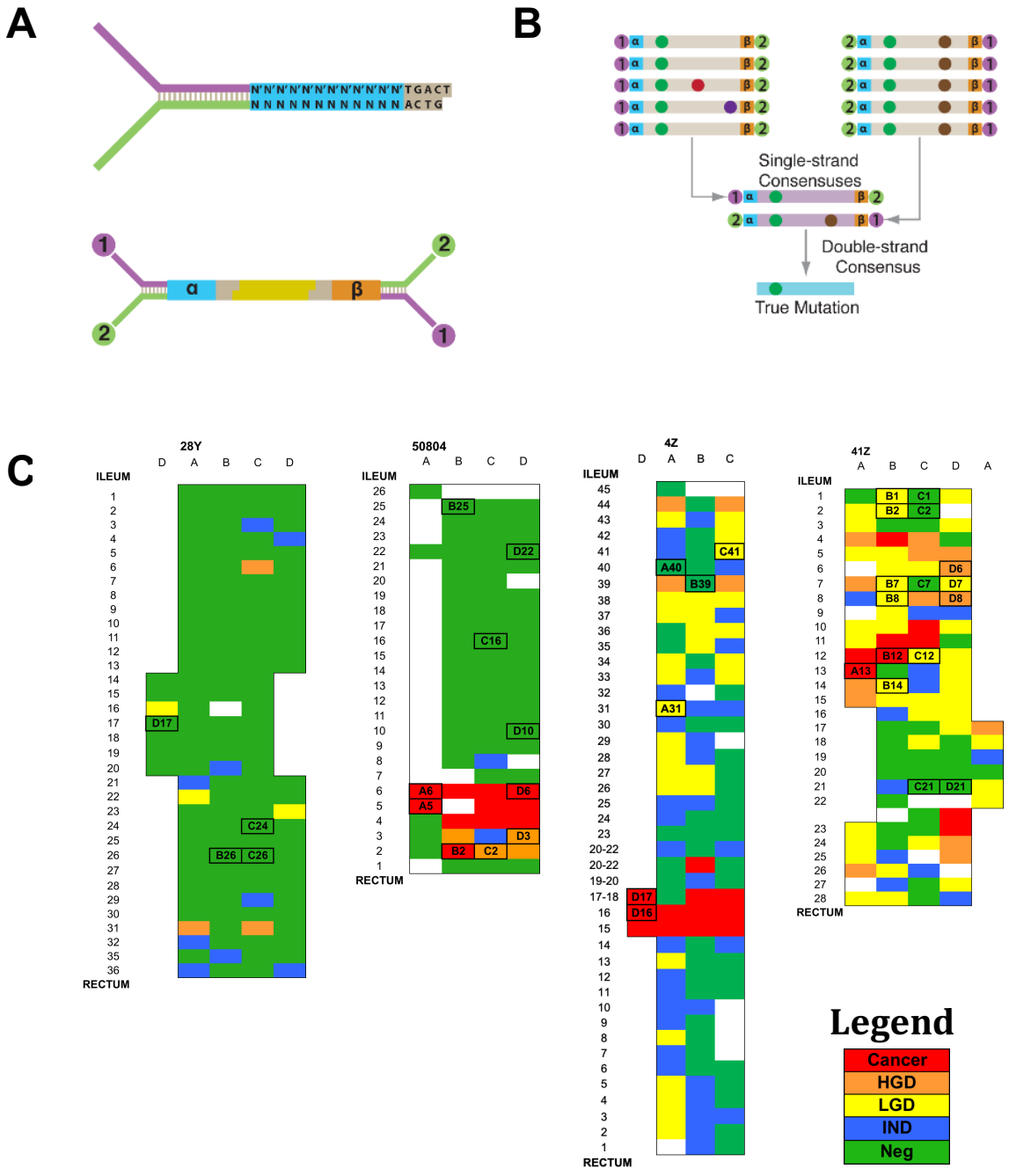


Figure 2. Clonal and subclonal mtDNA mutations

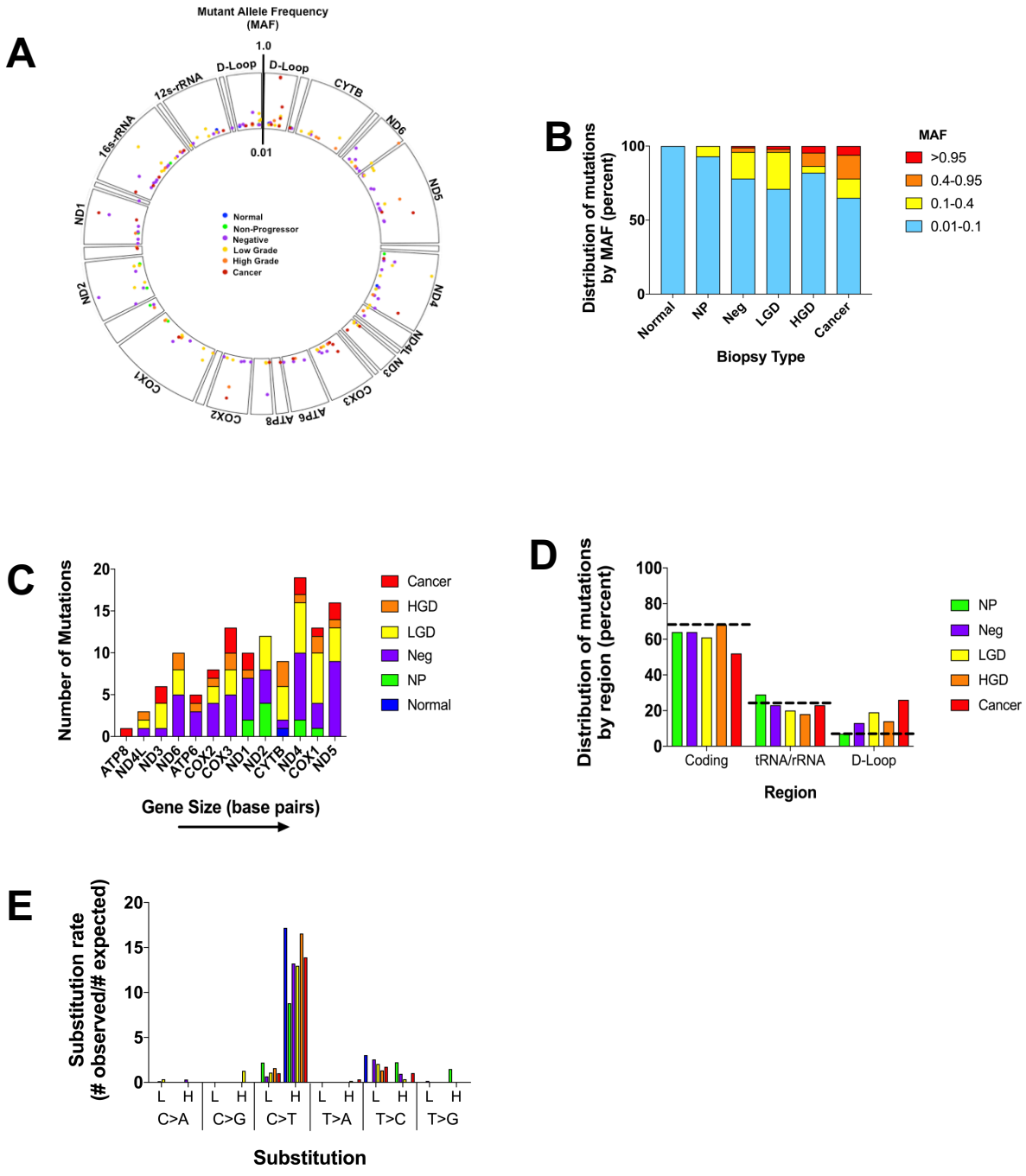


Figure 3. Comparison of clonal and subclonal mutations by biopsy type

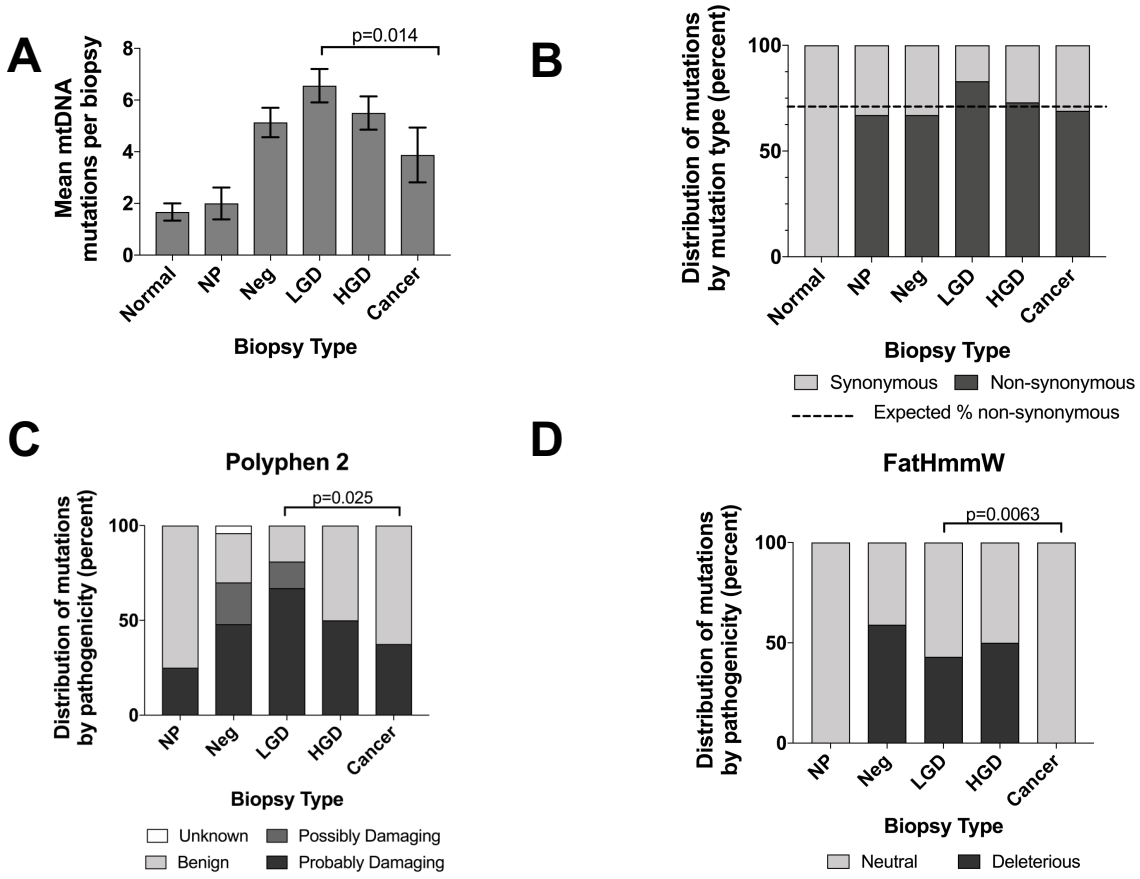


Figure 4. Very low frequency (MAF<0.01) mtDNA mutational signature

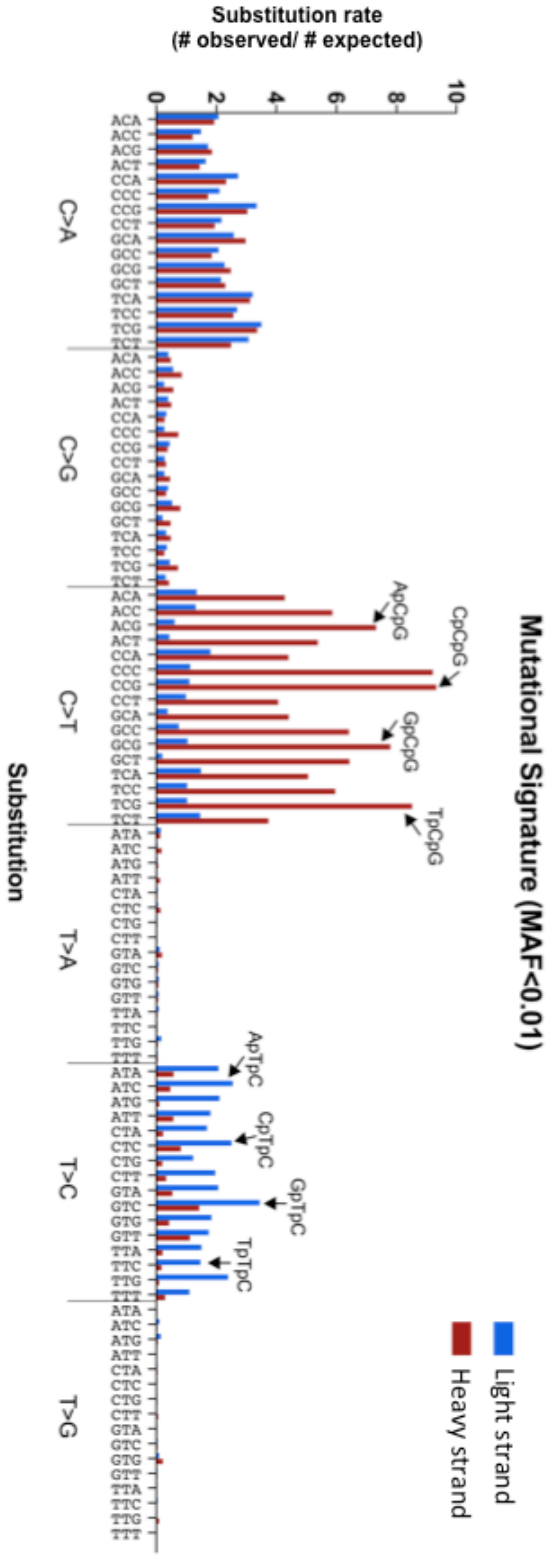


Figure 5. Quantification of very low frequency mutations (MAF<0.01) in D-loop *vs.* non-D-loop and by progression

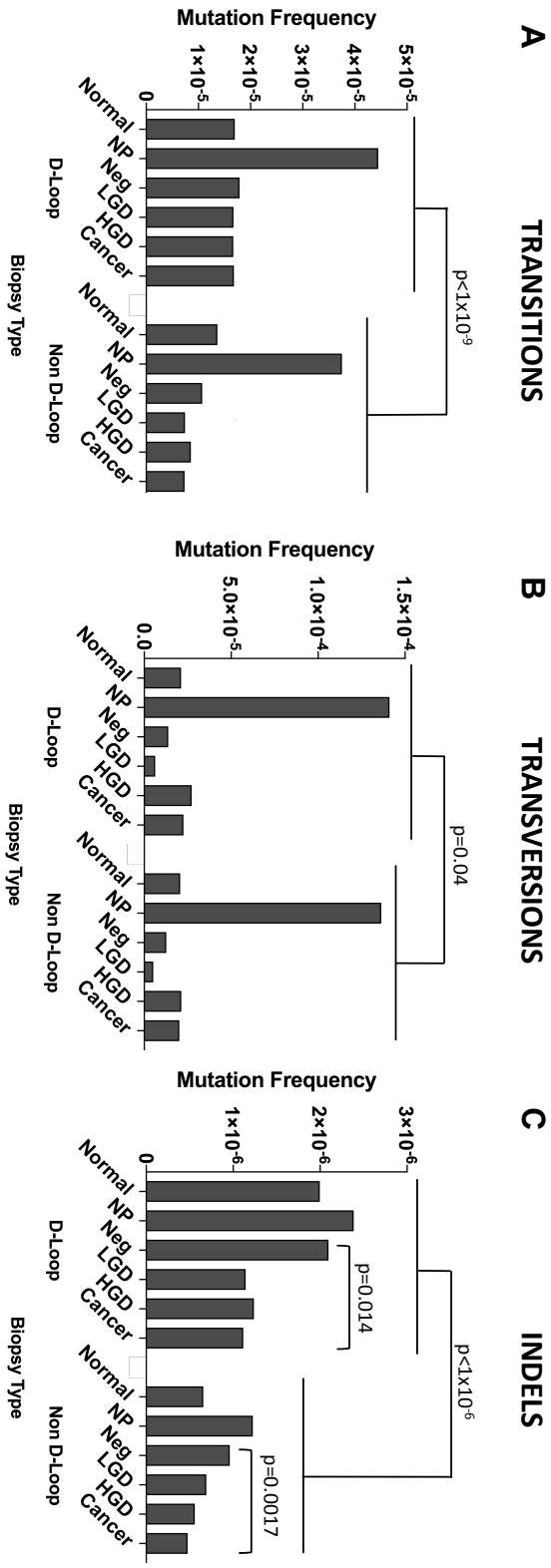
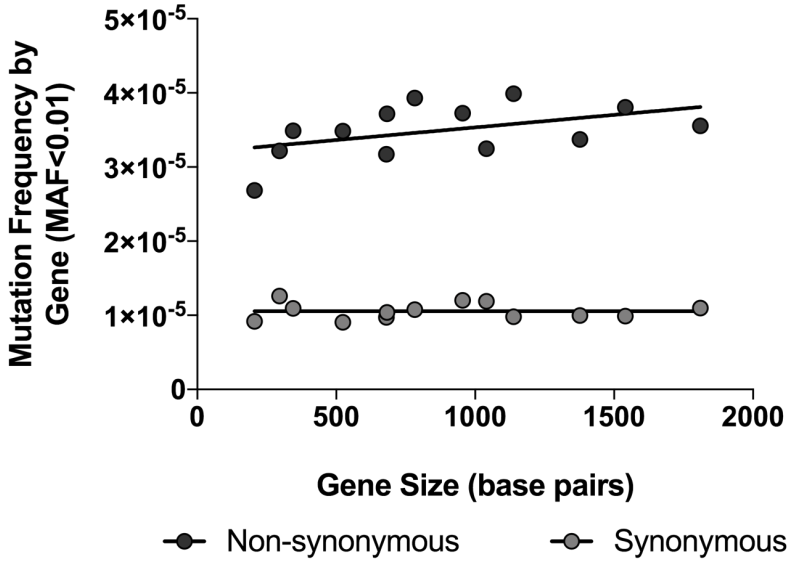
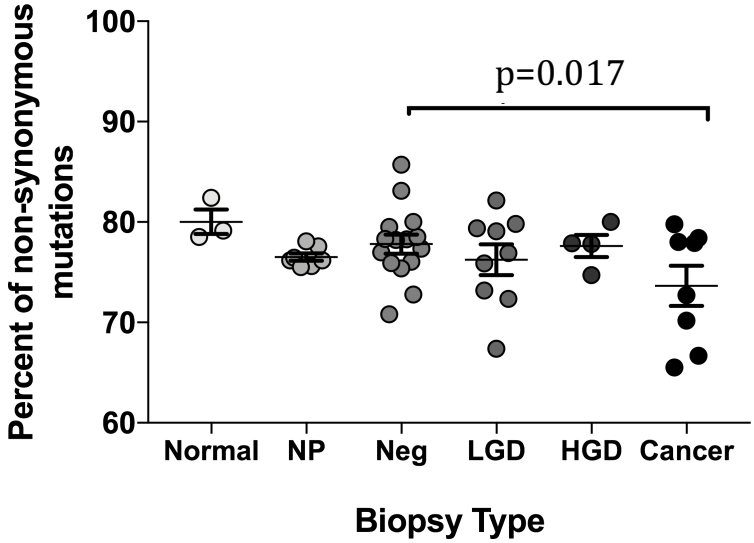


Figure 6. Very low frequency (MAF<0.01) mutation selection

A


Mutation Frequency by Gene (MAF<0.01)
5×10-5
4×10-5
3×10-5
2×10-5
1×10-5
0
0
500
1000
1500
2000
Gene Size (base pairs)
Non-synonymous
Synonymous

B


Percent of non-synonymous mutations
100
90
80
70
60
p=0.017
Normal
NP
Neg
LGD
HGD
Cancer
Biopsy Type

Supplementary Materials and Methods

Patients and samples

Biopsies from 3 non-UC patients, 3 UC Non-Progressors, and 4 UC Progressors were analyzed (Supplementary Table 1). The 3 non-UC controls included one patient with Hirschsprung's disease and two who had constipation. They were purposely selected in the younger age spectrum to avoid the effect of age-related mtDNA mutations. The reason for this non-UC group was to serve as a negative control, not to perform direct comparisons between mtDNA mutations in non-UC and UC patients. UC Progressor colons were chosen because detailed colon maps were available (Fig. 1C), they had been extensively analyzed in prior studies (1, 2), and epithelial DNA was available. Importantly, these colons included varying degrees and extension of dysplasia.

Biopsies from non-UC colons were collected during colonoscopy while UC samples were collected at colectomy. Samples were frozen at -80°C and stored in Minimum Essential Medium with 10% DMSO until used. For UC patients, a portion of colonic tissue adjacent to each biopsy was fixed in formalin, embedded in paraffin, sectioned, stained with hematoxylin-eosin, and histologically assessed by a GI pathologist to determine dysplastic grade and inflammation.

Epithelial cell and DNA isolation

In prior studies (1, 2) epithelial cells were isolated from colonic tissue via EDTA shake-off as previously described (3). Then DNA was extracted using silica filtration columns (Qiagen, Hilden, Germany), quantified via a Qubit fluorimeter (ThermoFisher Scientific, Waltham, MA, USA) or Nanodrop UV spectroscopy (ThermoFisher Scientific) and stored at -80°C.

Duplex Sequencing

Duplex Sequencing was performed as previously described (4, 5) with the following modifications. Upon sonication, end repair and A-tailing were performed using End Repair/dA-Tailing modules (New England Bioscience, Ipswich, MA, USA) and a GeneAmp PCR System 9700 (Applied Biosystems). Prepared DNA was then ligated to DS adapters (IDT, San Jose, CA, USA) using NEBNext Ultra II Ligation module. DS adapters are asymmetrical, double-stranded DNA molecules that contain the sequences required for the Illumina system as well as a random, double-stranded, 10bp sequence that uniquely tags each DNA molecule. Non-ligated DNA was then removed via size-selection magnetic bead wash (selecting for 300-700bp range) with Agencourt Ampure XP beads (Beckman Coulter, Brea, CA, USA) with a 0.8:1 DNA to bead ratio (0.8x) and eluted in molecular biology grade water. To determine optimal input of ligated DNA for amplification, samples were amplified with a DS adapter specific primer (MWS13, 5'-AATGATACGGCGACCACCGAG-3') and a primer from an internal mitochondrial sequence (MitoRev, 5'-GCGCTTACTTTGTAGCCTTCA-3') (both by IDT) using the Kapa Real Time Library Amplification Kit (Kapa Biosystems, Wilmington, MA, USA). qPCR reactions were performed using a Rotorgene 3000 (Corbett Research, Cambridge, UK) using the following cycling conditions: 95°C for 10 minutes, 40 cycles of 95°C for 10 seconds (sec), 60°C for 15 sec, and 72°C for 20 sec, followed by a 60°C-98°C melt step. The cycle threshold of each sample was compared to that of a DNA standard created from a previously sequenced mtDNA sample that yielded optimal reads per tag and sequencing depth. Ligated DNA was then diluted to the optimal input amount and amplified using the Kapa KK2702 kit, the two DS adapters specific primers MWS13 and MWS20 (IDT), and the following cycling conditions: 95°C for 4 minutes, a variable number of cycles of 95°C for 20 sec, 60°C for 20 sec, and 72°C for 15 sec. Samples were taken out of the Rotor-gene 3000 before the qPCR curve plateaued to prevent the formation

of DNA concatemers. A second 0.8x bead wash was performed in the same manner as the first to purify the amplified DNA. DNA was then captured using the NimbleGen SeqCap Target Enrichment kit (Roche, Basel, Switzerland) or the xGen Lockdown Target Enrichment kit (IDT) with probes specific for the mitochondrial genome (IDT). A third 1.8x bead wash was performed to clean the DNA. Samples were then indexed for sequencing via qPCR using MWS13 and custom 6 bp indexing primers (IDT) and the Kapa Real Time Library Amplification Kit on the Rotor-gene 3000. A final 0.8x bead was performed as before. Individual DNA sample concentrations were quantified using a Qubit fluorimeter (Life Technologies, Carlsbad, CA, USA). Samples were then analyzed with a 4200 TapeStation (Agilent, Santa Clara, CA, USA) to determine DNA fragment length, and the presence of any unwanted DNA fragments or contaminants. An additional bead wash was performed as needed to purify samples. Sequencing libraries were pooled by adding sample DNA in ratios consistent with the desired sequencing depth for each sample. The pool was then quantified via qPCR using a library quantification kit (Kapa Biosystems) and the Rotor-gene 3000 with the following cycling conditions: 95°C for 5 minutes, 40 of cycles of 95°C for 30 seconds and 60°C for 45 seconds, followed by a 60°C-98°C melt step. The DNA library pool concentration was then adjusted for sequencing. Sequencing was performed using 2x100 bp pair-end reads on the Illumina HiSeq 2500 or 2x150 bp pair-end reads on the Illumina NextSeq 550.

Data processing

Raw data files were analyzed as previously described (5, 6) with several modifications. Paired-end information was retained. Consensus making was performed prior to alignment. Sequences were first compiled using the Unified Consensus Maker, a custom python script (https://github.com/risqueslab/DuplexSequencingScripts) that compares the sequences of all the reads that share the same molecular tag in order to produce a single stranded consensus sequence

(SSCS). Because the molecular tags are double-stranded, each SSCS sequence can be then compared to its complement, introducing an additional step of error correction that filters all mutations that are not present in the two complementary DNA molecules. These highly accurate double-stranded consensus sequences, or DCS reads, were aligned using BWA-MEM with default parameters (bio-bwa.sourceforge.net) to human reference genome version 37 (GRCh37), (ncbi.nlm.nih.gov/grc/human) a revised version of the Cambridge reference that corrects for an error at base 3107 in previous versions. The Genome Analysis Toolkit (GATK) version 3.6 (software.broadinstitute.org/gatk) Indel-Realigner was used to perform local realignment of each read. Regions with PolyC sequences, which caused significant alignment and mutation calling issues, were excluded from the bed file. The excluded regions include the following positions: 301-317, 13052-13064, 15536-15548, and 16183-16196. GATK Clip-Reads was used to clip 10 bp at both the 5' and 3' end of each read to remove artifacts created during end repair and A-tailing. DCS reads with more than 5% indeterminate bases (Ns) were removed. Positions with less than 100 DCS reads were not considered for analysis. The fgbio (https://github.com/fulcrumgenomics/fgbio) tool ClipOverlappingReads was used to clip any overlapping bases from paired reads. The generated output file includes the mutation type (including insertions and deletions) and frequency at each position, the resulting base and amino acid change, and clonal and subclonal mutations that can be used for haplotype analysis.

Clonal and Subclonal Mutation Analysis

Total numbers of clonal and subclonal mutations numbers were counted (Table 1, Table S2). Mutations found in all samples from a given colon and with the majority of samples having a mutation frequency of 0.99 or above were considered constitutional to the patient and removed from consideration. For colons where only one sample was analyzed, synonymous and non-coding mutations with a frequency of 0.99 or above were also considered constitutional and thus

removed. The remaining clonal and subclonal mutations were compared across samples based on clonality (MAF 1-10%, 10-40%, 40-95%, >95%). Mutation location was visualized using the Circlize package in R (https://CRAN.R-project.org/package=circlize).

Very Low Frequency Mutation Analysis

To calculate the frequency of mutations in the D-loop and non-D-loop, the number of mutations was divided by the total number of nucleotides sequenced in the D-loop and in the non-D-loop, to take into consideration the fact that the D-loop is much smaller and differences in sequencing depth between samples. To calculate the frequency of each mutation type, the number of mutations of each type (e.g. C>A, C>G, C>T, T>A, T>C, T>G) was divided by the total number of sequenced C nucleotides or G nucleotides in the heavy and light strand of the mtDNA. This takes into consideration the nucleotide composition of the mtDNA as well as the depth of sequencing of each sample. Different mutations identified at the same nucleotide position were independently counted.

Determination of mutational signatures

We calculated the expected frequency of each substitution within each of the possible 16 trinucleotides by taking into account the trinucleotide composition of the mtDNA and assuming equal probability for all three substitutions. We then estimated the substitution rate by dividing the number of observed substitutions by the number of expected substitutions.

Associations between frequency of mutations and gene size

For each biopsy, we calculated the number of synonymous and non-synonymous mutations identified in each gene. We then divided the number of mutations by the total number of nucleotides sequenced for each respective gene in order to correct by gene size and sequencing depth between samples. The average frequency for all the biopsies in the study as

well as for all the biopsies within each group (e.g. normal, NP, Negative, LGD, HGD, and cancer) was then calculated separately for synonymous and non-synonymous mutations and plotted against gene size.

Mutation Pathogenicity

Analysis of non-synonymous mitochondrial mutation consequences and the resulting pathogenicity were assessed by comparing against the MitImpact database version 2.9 (7), which compiles pathogenicity calls from a variety of online tools. Only clonal and subclonal mutations were included for pathogenic analysis. MitImpact contains information on all possible missense mutations that could occur in coding regions of the mitochondrial genome. Thus, in this analysis indels or mutations found in tRNA coding or control regions were removed from consideration. Annotations from Polyphen2 (8) and Fathmmw (9) were used to predict mutational pathogenicity.

Statistical analyses

For clonal and subclonal mutations, a standard two-way table χ^2 test was used to test the null hypothesis of homogeneity of mutation rates (proportions) across genes using the number of mutant sites and number of non-mutant sites, which takes into account gene size. To account for possible correlation between outcomes within the same patient (including within the same biopsy), we applied the method of generalized estimating equations (GEE). While the number of patients in the study is modest (N=10), the "effective sample" size is affected by the number of biopsies, number of mutations analyzed and depth of sequencing (size of sample interrogated). The effective sample size is close to the number of patients if all results within a patient are very similar but is close to the number of mutations in the analysis if results have little correlation within individual. GEE uses an empirical estimator of the standard error in regression models, an

estimator that includes a component for correlation among results within individual and arrives at a result that reflects the effective sample size. However, GEE also relies on a large number of individuals (large sample theory) for validity of the estimators. Because of the modest number of individuals in the study, we did not want to rely on large sample assumptions. Instead, resampling was used with GEE in order to obtain standard error (SE) estimators. Specifically, observations were permuted within individual (1000 times for each test) to retain any correlation within individual and the standard deviation of the 1000 estimators was used as the SE (10). This method provided a much more conservative p-value than the GEE SE estimator, justifying the approach. For tests of outcomes for an inverted V-shape or a linear decrease from LGD to HGD to cancer, we used models with either a quadratic polynomial or only a linear term restricted to biopsies in the last three groups, respectively. For analysis of VLF mutations, the same GEE permutation method was used except the number of bases interrogated was taken into account via a Poisson offset or binomial denominator, as appropriate.

References

1. Salk JJ, Salipante SJ, Risques RA, Crispin DA, Li L, Bronner MP, et al. Clonal expansions in ulcerative colitis identify patients with neoplasia. Proceedings of the National Academy of Sciences of the United States of America. 2009;106(49):20871-6.
2. Risques RA, Lai LA, Himmetoglu C, Ebaee A, Li L, Feng Z, et al. Ulcerative colitis-associated colorectal cancer arises in a field of short telomeres, senescence, and inflammation. Cancer Res. 2011;71(5):1669-79.
3. Rabinovitch PS, Dziadon S, Brentnall TA, Emond MJ, Crispin DA, Haggitt RC, et al. Pancolonic chromosomal instability precedes dysplasia and cancer in ulcerative colitis. Cancer Res. 1999;59(20):5148-53.

4. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proceedings of the National Academy of Sciences of the United States of America. 2012;109(36):14508-13.
5. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, et al. Detecting ultralow-frequency mutations by Duplex Sequencing. Nature protocols. 2014;9(11):2586-606.
6. Nachmanson D, Shenyi L, Schmidt EK, Hipp MJ, Baker KT, Zhang Y, et al. Targeted genome fragmentation with CRISPR/Cas9 enables fast and efficient enrichment of small genomic regions and ultra-accurate sequencing with low DNA input (CRISPR-DS). Genome Res 2018;10:1589-1599.
7. Castellana S, Ronai J, Mazza T. MitImpact: an exhaustive collection of pre-computed pathogenicity predictions of human mitochondrial non-synonymous variants. Hum Mutat. 2015;36(2):E2413-22.
8. Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nat Methods. 7. United States2010. p. 248-9.
9. Shihab HA, Gough J, Cooper DN, Stenson PD, Barker GL, Edwards KJ, et al. Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. Hum Mutat. 2013;34(1):57-65.
10. Efron BTRJ. An Introduction to the Bootstrap: Chapman & Hall/CRC; 1993.

Supplementary Tables

Table S1. Patient Information

Case	Patient Type	Highest Grade	Haplotype	Sex	Age (years)	Disease Duration (years)	Disease Activity	Number of Biopsies by Grade					
								Negative for dysplasia			Low Grade Dysplasia	High Grade Dysplasia	Cancer
12N	Normal	*NA*	T2a1a8	Female	30	*NA*	*NA*	1					
13N	Normal*	*NA*	K1a1b1a	Male	12	*NA*	*NA*	1					
18N	Normal	*NA*	H11a2a	Female	22	*NA*	*NA*	1					
3062514	Non-progressor	*NA*	H1c1	Male	59	0.25	Active		1				
3062587	Non-progressor	*NA*	H3	Male	23	8	Mild		3				
182J	Non-progressor	*NA*	H3	Female	46	17	Moderate		3				
28Y	Progressor	HGD	H56b	Male	48	10	Moderate			4			
50804	Progressor	Cancer	K2a5	*DU*	*DU*	*DU*	*DU*			4		2	4
4Z	Progressor	Cancer	T2b	Male	31	4	Unknown			2	2		2
41Z	Progressor	Cancer	U5a1a1	Female	51	13	Severe			5	7	2	2
							Total Biopsies	3	7	15	9	4	8
							MAF>0.01	5	14	77	59	22	31
						Total Mutations	**MAF<0.01**	6790	25495	7384	5026	5417	6652

NA - Not applicable, *DU* - Data unavailable

* Hirschprung Disease

Table S2. Biopsy Information

										Number of mutations	
Case	**Patient Type**	**Biopsy**	**Dysplastic Grade**	**Acute inflammation**	**Chronic inflammation**	**PSC**	**% DCS mapped**	**DCS nucleotides sequenced**	**DCS depth**	**MAF >0.01**	**MAF <0.01**
12N	Normal	12N	Negative	*NA*	*NA*	*NA*	99.8	30,184,042	1,822	2	619
13N	Normal	13N	Negative	*NA*	*NA*	*NA*	99.8	79,226,031	4,782	1	5,371
18N	Normal	18N	Negative	*NA*	*NA*	*NA*	99.8	45,499,639	2,746	2	800
3062514	NP	C6	Negative	mild	high	*DU*	99.5	22,474,307	1,356	5	4,686
3062587	NP	C6	Negative	none	low	*DU*	99.6	18,504,290	1,117	3	4,476
3062587	NP	C7	Negative	none	low	*DU*	99.6	28,857,768	1,741	1	1,584
3062587	NP	C8	Negative	none	none	*DU*	99.5	11,027,421	666	2	4,505
182J	NP	A15	Negative	none	low	*DU*	99.6	25,329,670	1,529	2	2,592
182J	NP	B4	Negative	none	low	*DU*	99.5	17,643,045	1,065	0	5,059
182J	NP	C10	Negative	none	low	*DU*	99.6	22,023,637	1,329	1	2,593
28Y	P	B26	Negative	none	low	No	99.1	11,398,272	688	9	132
28Y	P	C24	Negative	none	low	No	99.6	20,358,789	1,229	3	168
28Y	P	C26	Negative	moderate	high	No	99.7	42,583,801	2,570	5	342
28Y	P	D17	Negative	none	none	No	99.2	11,465,838	692	8	166
50804	P	A5	Cancer	*DU*	*DU*	*DU*	99.6	27,668,019	1,670	2	149
50804	P	A6	Cancer	*DU*	*DU*	*DU*	99.7	24,237,061	1,463	0	2,025
50804	P	B2	Cancer	*DU*	*DU*	*DU*	99.0	12,210,267	737	3	785
50804	P	B25	Negative	*DU*	*DU*	*DU*	98.2	10,233,330	618	1	1,205
50804	P	C2	HGD	*DU*	*DU*	*DU*	99.7	19,849,601	1,198	4	671
50804	P	C16	Negative	*DU*	*DU*	*DU*	99.8	21,124,369	1,275	5	2,435
50804	P	D3	HGD	*DU*	*DU*	*DU*	99.8	31,390,621	1,895	6	2,253
50804	P	D6	Cancer	*DU*	*DU*	*DU*	99.7	50,815,555	3,067	10	1,924
50804	P	D10	Negative	*DU*	*DU*	*DU*	96.7	10,397,971	628	6	100
50804	P	D22	Negative	*DU*	*DU*	*DU*	83.1	22,717,025	1,371	7	180
4Z	P	A31	LGD	mild	high	No	99.8	11,999,286	724	5	80
4Z	P	A40	Negative	mild	high	No	99.8	11,155,416	673	3	96
4Z	P	B39	Negative	none	high	No	99.8	11,514,668	695	4	89
4Z	P	C41	LGD	none	high	No	99.8	11,977,064	723	7	75
4Z	P	D16	Cancer	none	none	No	99.8	11,923,752	720	5	105
4Z	P	D17	Cancer	none	none	No	99.5	9,877,477	596	4	59
41Z	P	A13	Cancer	none	high	Yes	99.7	75,137,215	4,538	5	1,266
41Z	P	B1	LGD	none	high	Yes	99.8	53,215,145	3,212	6	2,630
41Z	P	B2	LGD	none	high	Yes	83.8	9,816,110	592	7	140
41Z	P	B7	LGD	mild	low	Yes	99.7	22,299,726	1,346	6	234
41Z	P	B8	LGD	mild	high	Yes	99.6	19,242,426	1,161	7	229
41Z	P	B12	Cancer	none	low	Yes	99.8	45,994,840	2,776	2	339
41Z	P	B14	LGD	none	high	Yes	99.8	37,332,002	2,253	11	227
41Z	P	C1	Negative	none	high	Yes	99.8	27,752,296	1,675	7	379
41Z	P	C2	Negative	none	low	Yes	99.8	35,202,461	2,125	5	347
41Z	P	C7	Negative	none	high	Yes	99.7	47,715,509	2,880	4	993
41Z	P	C12	LGD	mild	high	Yes	99.7	29,191,093	1,762	4	332
41Z	P	C21	Negative	none	high	Yes	99.7	80,571,808	4,863	7	493
41Z	P	D6	HGD	none	high	Yes	99.8	110,013,341	6,640	7	958
41Z	P	D7	LGD	none	high	Yes	99.8	115,704,629	6,983	6	1,079
41Z	P	D8	HGD	none	high	Yes	99.8	145,141,921	8,760	5	1,535
41Z	P	D21	Negative	none	high	Yes	99.7	17,829,255	1,076	3	259

Abbreviations: DCS - duplex consensus sequence; PSC - primary sclerosing cholangitis; MAF - mutant allele frequency; NA - not applicable; DU - data unavailable

Table S3. Mutation Information

Case	Patient Type	Sample	Biopsy	Dysplastic Grade	Position	Ref Base	Gene	Coding	MAF	Frequency Classification	Substitution	Gene Consequence	Mutation Type	Amino Acid Change
12N	Normal	12N	12N-12N	Negative	65	T	Control_region (HVS2)	noncoding	0.0168	subclonal	D	deletion	indel	
12N	Normal	12N	12N-12N	Negative	11626	T	ND4	coding	0.0145	subclonal	TC	synonymous	synonymous	
13N	Normal	13N	13N-13N	Negative	366	G	Control_region (HVS2)	noncoding	0.0111	subclonal	GA	not coding	not coding	
18N	Normal	18N	18N-18N	Negative	1303	G	12S_rRNA	tRNA/rRNA	0.0726	subclonal	GA	not coding	not coding	
18N	Normal	18N	18N-18N	Negative	16145	G	Control_region HVS1	noncoding	0.0475	subclonal	GA	not coding	not coding	
3062514	NP	C6	3062514-C6	Negative	513	G	Control_region HVS3	noncoding	0.0221	subclonal	D	deletion	indel	
3062514	NP	C6	3062514-C6	Negative	2702	G	16S_rRNA	tRNA/rRNA	0.011	subclonal	GA	not coding	not coding	
3062587	NP	C6	3062587-C6	Negative	3577	A	ND1	coding	0.0151	subclonal	AC	Met > Leu	nonsynonymous	M91L
3062587	NP	C6	3062587-C6	Negative	5330	C	ND2	coding	0.0118	subclonal	CT	synonymous	synonymous	
3062514	NP	C6	3062514-C6	Negative	5800	A	tRNA-Cys	tRNA/rRNA	0.109	subclonal	AG	not coding	not coding	
3062514	NP	C6	3062514-C6	Negative	7236	G	COX1	coding	0.0103	subclonal	GA	Asp > Asn	nonsynonymous	D445N
3062514	NP	C6	3062514-C6	Negative	10830	G	ND4	coding	0.0109	subclonal	GA	Trp > Ter	nonsynonymous	W24Ter
3062587	NP	C6	3062587-C6	Negative	11711	G	ND4	coding	0.0207	subclonal	GA	Ala > Thr	nonsynonymous	A318T
3062587	NP	C7	3062587-C7	Negative	5330	C	ND2	coding	0.0144	subclonal	CT	synonymous	synonymous	
3062587	NP	C8	3062587-C8	Negative	3577	A	ND1	coding	0.0158	subclonal	AC	Met > Leu	nonsynonymous	M91L
3062587	NP	C8	3062587-C8	Negative	5330	C	ND2	coding	0.0102	subclonal	CT	synonymous	synonymous	
182J	NP	A15	182J-A15	Negative	4720	G	ND2	coding	0.0105	subclonal	GA	Trp > Ter	nonsynonymous	W84Ter
182J	NP	A15	182J-A15	Negative	15922	A	tRNA-Thr	tRNA/rRNA	0.0281	subclonal	AG	not coding	not coding	
182J	NP	C10	182J-C10	Negative	15922	A	tRNA-Thr	tRNA/rRNA	0.0118	subclonal	AG	not coding	not coding	
28Y	P	B26	28Y-B26	Negative	72	T	Control_region (HVS2)	noncoding	0.0406	subclonal	TC	not coding	not coding	
28Y	P	B26	28Y-B26	Negative	152	T	Control_region HVS2	noncoding	0.0467	subclonal	TC	not coding	not coding	
28Y	P	B26	28Y-B26	Negative	583	G	tRNA-Phe	tRNA/rRNA	0.0145	subclonal	GA	not coding	not coding	
28Y	P	B26	28Y-B26	Negative	3737	T	ND1	coding	0.0278	subclonal	TC	Val > Ala	nonsynonymous	V144A
28Y	P	B26	28Y-B26	Negative	4754	A	ND2	coding	0.0753	subclonal	AG	synonymous	synonymous	
28Y	P	B26	28Y-B26	Negative	10685	G	ND4L	coding	0.0871	subclonal	GA	synonymous	synonymous	
28Y	P	B26	28Y-B26	Negative	10993	G	ND4	coding	0.23	subclonal	GA	synonymous	synonymous	
28Y	P	B26	28Y-B26	Negative	11651	G	ND4	coding	0.0611	subclonal	GA	Val > Met	nonsynonymous	V298M
28Y	P	B26	28Y-B26	Negative	11866	A	ND4	coding	0.0346	subclonal	I	insertion	indel	
28Y	P	C24	28Y-C24	Negative	513	G	Control_region HVS3	noncoding	0.0122	subclonal	D	not coding	not coding	
28Y	P	C24	28Y-C24	Negative	2573	G	16S_rRNA	tRNA/rRNA	0.0123	subclonal	GA	not coding	not coding	
28Y	P	C24	28Y-C24	Negative	5177	G	ND2	coding	0.119	subclonal	GA	synonymous	synonymous	
28Y	P	C26	28Y-C26	Negative	390	A	Control_region	noncoding	0.269	subclonal	AG	not coding	not coding	
28Y	P	C26	28Y-C26	Negative	2054	T	16S_rRNA	tRNA/rRNA	0.347	subclonal	TC	not coding	not coding	
28Y	P	C26	28Y-C26	Negative	9531	A	COX3	coding	0.0138	subclonal	I	insertion	indel	
28Y	P	C26	28Y-C26	Negative	12814	G	ND5	coding	0.0585	subclonal	GA	Ala > Thr	nonsynonymous	A160T
28Y	P	C26	28Y-C26	Negative	14207	G	ND6	coding	0.278	subclonal	GA	Thr > Ile	nonsynonymous	T156I
28Y	P	D17	28Y-D17	Negative	1313	A	12S_rRNA	tRNA/rRNA	0.0102	subclonal	AG	not coding	not coding	
28Y	P	D17	28Y-D17	Negative	2128	G	16S_rRNA	tRNA/rRNA	0.012	subclonal	GA	not coding	not coding	
28Y	P	D17	28Y-D17	Negative	2631	G	16S_rRNA	tRNA/rRNA	0.0126	subclonal	GA	not coding	not coding	
28Y	P	D17	28Y-D17	Negative	3323	T	ND1	coding	0.0136	subclonal	TC	Leu > Pro	nonsynonymous	L6P
28Y	P	D17	28Y-D17	Negative	6825	G	COX1	coding	0.233	subclonal	GA	Ala > Thr	nonsynonymous	A308T
28Y	P	D17	28Y-D17	Negative	10993	G	ND4	coding	0.0171	subclonal	GA	synonymous	synonymous	
28Y	P	D17	28Y-D17	Negative	13608	T	ND5	coding	0.0359	subclonal	TC	synonymous	synonymous	
28Y	P	D17	28Y-D17	Negative	14649	T	ND6	coding	0.0163	subclonal	TC	Ser > Gly	nonsynonymous	S9G
50804	P	A5	50804-A5	Cancer	2022	G	16S_rRNA	tRNA/rRNA	0.0158	subclonal	GA	not coding	not coding	
50804	P	A5	50804-A5	Cancer	8270	C	non-coding	coding	0.0154	subclonal	D	deletion	indel	
50804	P	B2	50804-B2	Cancer	7978	C	COX2	coding	0.807	subclonal	CT	synonymous	synonymous	
50804	P	B2	50804-B2	Cancer	10946	A	ND4	coding	0.0111	subclonal	I	insertion	indel	
50804	P	B2	50804-B2	Cancer	16207	A	Control_region HVS1	noncoding	0.0229	subclonal	AG	not coding	not coding	
50804	P	B25	50804-B25	Negative	8249	G	COX2	coding	0.0114	subclonal	GA	Gly > Ter	nonsynonymous	G222Ter
50804	P	C16	50804-C16	Negative	1958	G	16S_rRNA	tRNA/rRNA	0.0317	subclonal	GT	not coding	not coding	
50804	P	C16	50804-C16	Negative	5910	G	COX1	coding	0.0104	subclonal	GA	Ala > Thr	nonsynonymous	A3T
50804	P	C16	50804-C16	Negative	14279	G	ND6	coding	0.0269	subclonal	GA	Ser > Leu	nonsynonymous	S132L
50804	P	C16	50804-C16	Negative	15106	G	CYTB	coding	0.0156	subclonal	GA	synonymous	synonymous	
50804	P	C16	50804-C16	Negative	16181	A	Control_region HVS1	noncoding	0.0604	subclonal	AG	not coding	not coding	
50804	P	C2	50804-C2	High Grade	4449	G	tRNA-Met	tRNA/rRNA	0.0425	subclonal	GA	not coding	not coding	
50804	P	C2	50804-C2	High Grade	7978	C	COX2	coding	0.576	subclonal	CT	synonymous	synonymous	
50804	P	C2	50804-C2	High Grade	15242	G	CYTB	coding	0.0141	subclonal	GA	Gly > Ter	nonsynonymous	G165Ter
50804	P	C2	50804-C2	High Grade	16264	C	Control_region HVS1	noncoding	0.109	subclonal	CT	not coding	not coding	
50804	P	D3	50804-D3	High Grade	2536	G	16S_rRNA	tRNA/rRNA	0.0258	subclonal	GA	not coding	not coding	
50804	P	D3	50804-D3	High Grade	5703	G	tRNA-Asn	tRNA/rRNA	0.0614	subclonal	GA	not coding	not coding	
50804	P	D3	50804-D3	High Grade	8902	G	ATP6	coding	0.0223	subclonal	GA	Ala > Thr	nonsynonymous	A126T
50804	P	D3	50804-D3	High Grade	9713	G	COX3	coding	0.0114	subclonal	GA	synonymous	synonymous	
50804	P	D3	50804-D3	High Grade	11866	A	ND4	coding	0.0164	subclonal	I	insertion	indel	
50804	P	D3	50804-D3	High Grade	15306	T	CYTB	coding	0.0114	subclonal	TC	Phe > Ser	nonsynonymous	F186S
50804	P	D6	50804-D6	Cancer	214	A	Control_region HVS2	noncoding	0.0123	subclonal	AG	not coding	not coding	
50804	P	D6	50804-D6	Cancer	1440	G	12S_rRNA	tRNA/rRNA	0.0127	subclonal	GA	not coding	not coding	
50804	P	D6	50804-D6	Cancer	2269	G	16S_rRNA	tRNA/rRNA	0.0102	subclonal	GA	not coding	not coding	
50804	P	D6	50804-D6	Cancer	3143	T	16S_rRNA	tRNA/rRNA	0.0201	subclonal	TC	not coding	not coding	
50804	P	D6	50804-D6	Cancer	4412	G	tRNA-Met	tRNA/rRNA	0.026	subclonal	GA	not coding	not coding	
50804	P	D6	50804-D6	Cancer	6791	A	COX1	coding	0.0915	subclonal	AG	synonymous	synonymous	
50804	P	D6	50804-D6	Cancer	8951	T	ATP6	coding	0.0161	subclonal	TC	Val > Ala	nonsynonymous	V142A
50804	P	D6	50804-D6	Cancer	10068	G	ND3	coding	0.151	subclonal	GA	Ala > Thr	nonsynonymous	A4T

Case	Patient Type	Sample	Biopsy	Dysplastic Grade	Position	Ref Base	Gene	Coding	MAF	Frequency Classification	Substitution	Gene Consequence	Mutation Type	Amino Acid Change
50804	P	D6	50804-D6	Cancer	10326	T	ND3	coding	0.0296	subclonal	TC	Ser > Pro	nonsynonymous	S90P
50804	P	D6	50804-D6	Cancer	15914	A	tRNA-Thr	tRNA/rRNA	0.0153	subclonal	AT	not coding	not coding	
50804	P	D10	50804-D10	Negative	2083	T	16S_rRNA	tRNA/rRNA	0.035	subclonal	TC	not coding	not coding	
50804	P	D10	50804-D10	Negative	2182	G	16S_rRNA	tRNA/rRNA	0.0161	subclonal	GA	not coding	not coding	
50804	P	D10	50804-D10	Negative	2927	C	16S_rRNA	tRNA/rRNA	0.0126	subclonal	CT	not coding	not coding	
50804	P	D10	50804-D10	Negative	3243	A	tRNA-Leu	tRNA/rRNA	0.141	subclonal	AG	not coding	not coding	
50804	P	D10	50804-D10	Negative	12127	G	ND4	coding	0.176	subclonal	GA	synonymous	synonymous	
50804	P	D10	50804-D10	Negative	13345	G	ND5	coding	0.0217	subclonal	GA	Ala > Thr	nonsynonymous	A337T
50804	P	D22	50804-D22	Negative	5774	T	tRNA-Cys	tRNA/rRNA	0.0404	subclonal	TC	not coding	not coding	
50804	P	D22	50804-D22	Negative	7762	G	COX2	coding	0.0147	subclonal	GA	synonymous	synonymous	
50804	P	D22	50804-D22	Negative	9182	G	ATP6	coding	0.0483	subclonal	GA	Ser > Asn	nonsynonymous	S219N
50804	P	D22	50804-D22	Negative	9798	T	COX3	coding	0.0603	subclonal	TC	Phe > Leu	nonsynonymous	F198L
50804	P	D22	50804-D22	Negative	11222	G	ND4	coding	0.0104	subclonal	GA	Val > Met	nonsynonymous	V155M
50804	P	D22	50804-D22	Negative	12392	T	ND5	coding	0.0367	subclonal	TC	Ile > Thr	nonsynonymous	I19T
50804	P	D22	50804-D22	Negative	15915	G	tRNA-Thr	tRNA/rRNA	0.0109	subclonal	GA	not coding	not coding	
4Z	P	A31	4Z-A31	Low Grade	2470	G	16S_rRNA	tRNA/rRNA	0.0153	subclonal	GA	not coding	not coding	
4Z	P	A31	4Z-A31	Low Grade	11866	A	ND4	coding	0.0113	subclonal	I	insertion	indel	
4Z	P	A31	4Z-A31	Low Grade	12384	T	ND5	coding	0.256	subclonal	I	insertion	indel	
4Z	P	A31	4Z-A31	Low Grade	16148	C	Control_region HVS1	noncoding	0.0165	subclonal	CT	not coding	not coding	
4Z	P	A31	4Z-A31	Low Grade	16390	G	Control_region	noncoding	0.0296	subclonal	GA	not coding	not coding	
4Z	P	A40	4Z-A40	Negative	534	C	Control_region HVS3	noncoding	0.0123	subclonal	CT	not coding	not coding	
4Z	P	A40	4Z-A40	Negative	14384	G	ND6	coding	0.106	subclonal	GA	Ala > Val	nonsynonymous	A97V
4Z	P	A40	4Z-A40	Negative	15959	G	tRNA-Pro	tRNA/rRNA	0.0144	subclonal	GA	not coding	not coding	
4Z	P	B39	4Z-B39	Negative	3082	G	16S_rRNA	tRNA/rRNA	0.0122	subclonal	GA	not coding	not coding	
4Z	P	B39	4Z-B39	Negative	4490	C	ND2	coding	0.342	subclonal	CT	synonymous	synonymous	
4Z	P	B39	4Z-B39	Negative	11866	A	ND4	coding	0.0494	subclonal	I	insertion	indel	
4Z	P	B39	4Z-B39	Negative	13289	G	ND5	coding	0.0134	subclonal	GA	Gly > Asp	nonsynonymous	G318D
4Z	P	C41	4Z-C41	Low Grade	5294	C	ND2	coding	0.0242	subclonal	CA	Ser > Ter	nonsynonymous	S275Ter
4Z	P	C41	4Z-C41	Low Grade	6766	G	COX1	coding	0.0212	subclonal	GA	Trp > Ter	nonsynonymous	W288Ter
4Z	P	C41	4Z-C41	Low Grade	8353	T	tRNA-Lys	tRNA/rRNA	0.0135	subclonal	TC	not coding	not coding	
4Z	P	C41	4Z-C41	Low Grade	9810	G	COX3	coding	0.0241	subclonal	GA	Gly > Ser	nonsynonymous	G202S
4Z	P	C41	4Z-C41	Low Grade	10068	G	ND3	coding	0.012	subclonal	GA	Ala > Thr	nonsynonymous	A4T
4Z	P	C41	4Z-C41	Low Grade	11079	T	ND4	coding	0.198	subclonal	TC	Ile > Thr	nonsynonymous	I107T
4Z	P	C41	4Z-C41	Low Grade	11150	G	ND4	coding	0.989	subclonal	GC	Ala > Pro	nonsynonymous	A131P
4Z	P	D16	4Z-D16	Cancer	9438	G	COX3	coding	0.25	subclonal	GA	Gly > Ser	nonsynonymous	G78S
4Z	P	D16	4Z-D16	Cancer	9477	G	COX3	coding	0.012	subclonal	I	insertion	indel	
4Z	P	D16	4Z-D16	Cancer	16291	C	Control_region HVS1	noncoding	0.996	clonal	CT	not coding	not coding	
4Z	P	D16	4Z-D16	Cancer	16389	G	Control_region	noncoding	0.0107	subclonal	GA	not coding	not coding	
4Z	P	D16	4Z-D16	Cancer	16391	G	Control_region	noncoding	0.402	subclonal	GA	not coding	not coding	
4Z	P	D17	4Z-D17	Cancer	1719	G	16S_rRNA	tRNA/rRNA	0.269	subclonal	GA	not coding	not coding	
4Z	P	D17	4Z-D17	Cancer	9438	G	COX3	coding	0.0774	subclonal	GA	Gly > Ser	nonsynonymous	G78S
4Z	P	D17	4Z-D17	Cancer	16291	C	Control_region HVS1	noncoding	0.983	subclonal	CT	not coding	not coding	
4Z	P	D17	4Z-D17	Cancer	16391	G	Control_region	noncoding	0.168	subclonal	GA	not coding	not coding	
41Z	P	A13	41Z-A13	Cancer	3882	G	ND1	coding	0.0862	subclonal	GA	synonymous	synonymous	
41Z	P	A13	41Z-A13	Cancer	8368	G	ATP8	coding	0.0379	subclonal	GA	synonymous	synonymous	
41Z	P	A13	41Z-A13	Cancer	11988	T	ND4	coding	0.619	subclonal	TC	Met > Thr	nonsynonymous	M410T
41Z	P	A13	41Z-A13	Cancer	13393	G	ND5	coding	0.039	subclonal	GA	Glu > Lys	nonsynonymous	E353K
41Z	P	A13	41Z-A13	Cancer	13558	G	ND5	coding	0.628	subclonal	GA	Ala > Thr	nonsynonymous	A408T
41Z	P	B1	41Z-B1	Low Grade	822	G	12S_rRNA	tRNA/rRNA	0.0935	subclonal	GA	not coding	not coding	
41Z	P	B1	41Z-B1	Low Grade	2421	G	16S_rRNA	tRNA/rRNA	0.0389	subclonal	GA	not coding	not coding	
41Z	P	B1	41Z-B1	Low Grade	6063	T	COX1	coding	0.231	subclonal	TC	Tyr > His	nonsynonymous	Y54H
41Z	P	B1	41Z-B1	Low Grade	10488	A	ND4L	coding	0.0538	subclonal	AG	Asn > Asp	nonsynonymous	N7D
41Z	P	B1	41Z-B1	Low Grade	14364	G	ND6	coding	0.0156	subclonal	GA	synonymous	synonymous	
41Z	P	B1	41Z-B1	Low Grade	15059	G	CYTB	coding	0.0153	subclonal	GA	Gly > Ter	nonsynonymous	G104Ter
41Z	P	B2	41Z-B2	Low Grade	528	T	Control_region HVS3	noncoding	0.242	subclonal	TC	not coding	not coding	
41Z	P	B2	41Z-B2	Low Grade	567	A	Control_region HVS3	noncoding	0.121	subclonal	I	insertion	indel	
41Z	P	B2	41Z-B2	Low Grade	1669	G	tRNA-Val	tRNA/rRNA	0.0225	subclonal	GA	not coding	not coding	
41Z	P	B2	41Z-B2	Low Grade	8027	G	COX2	coding	0.025	subclonal	I	insertion	indel	
41Z	P	B2	41Z-B2	Low Grade	8167	T	COX2	coding	0.0145	subclonal	TC	synonymous	synonymous	
41Z	P	B2	41Z-B2	Low Grade	8348	A	tRNA-Lys	tRNA/rRNA	0.0277	subclonal	AG	not coding	not coding	
41Z	P	B2	41Z-B2	Low Grade	10098	G	ND3	coding	0.0236	subclonal	GA	Ala > Thr	nonsynonymous	A14T
41Z	P	B7	41Z-B7	Low Grade	366	G	Control_region (HVS2)	noncoding	0.0122	subclonal	GA	not coding	not coding	
41Z	P	B7	41Z-B7	Low Grade	10076	T	ND3	coding	0.113	subclonal	TC	synonymous	synonymous	
41Z	P	B7	41Z-B7	Low Grade	12417	C	ND5	coding	0.0399	subclonal	I	insertion	indel	
41Z	P	B7	41Z-B7	Low Grade	12442	G	ND5	coding	0.248	subclonal	GA	Val > Met	nonsynonymous	V36M
41Z	P	B7	41Z-B7	Low Grade	15708	G	CYTB	coding	0.233	subclonal	GA	Ser > Asn	nonsynonymous	S320N
41Z	P	B7	41Z-B7	Low Grade	16290	C	Control_region HVS1	noncoding	0.0151	subclonal	CT	not coding	not coding	
41Z	P	B8	41Z-B8	Low Grade	1074	G	12S_rRNA	tRNA/rRNA	0.0112	subclonal	GA	not coding	not coding	
41Z	P	B8	41Z-B8	Low Grade	4830	G	ND2	coding	0.233	subclonal	GA	Gly > Ser	nonsynonymous	G121S
41Z	P	B8	41Z-B8	Low Grade	11372	G	ND4	coding	0.0122	subclonal	GC	Val > Leu	nonsynonymous	V205L
41Z	P	B8	41Z-B8	Low Grade	11711	G	ND4	coding	0.0209	subclonal	GA	Ala > Thr	nonsynonymous	A318T
41Z	P	B8	41Z-B8	Low Grade	14503	T	ND6	coding	0.0117	subclonal	I	insertion	indel	
41Z	P	B8	41Z-B8	Low Grade	14865	G	CYTB	coding	0.0119	subclonal	GA	Cys > Tyr	nonsynonymous	C39Y
41Z	P	B8	41Z-B8	Low Grade	15092	G	CYTB	coding	0.0721	subclonal	GA	Gly > Ser	nonsynonymous	G115S

Case	Patient Type	Sample	Biopsy	Dysplastic Grade	Position	Ref Base	Gene	Coding	MAF	Frequency Classification	Substitution	Gene Consequence	Mutation Type	Amino Acid Change
41Z	P	B12	41Z-B12	Cancer	366	G	Control_region (HVS2)	noncoding	0.0394	subclonal	GA	not coding	not coding	
41Z	P	B12	41Z-B12	Cancer	3882	G	ND1	coding	0.835	subclonal	GA	synonymous	synonymous	
41Z	P	B14	41Z-B14	Low Grade	366	G	Control_region (HVS2)	noncoding	0.0237	subclonal	GA	not coding	not coding	
41Z	P	B14	41Z-B14	Low Grade	533	A	Control_region HVS3	noncoding	0.096	subclonal	AT	not coding	not coding	
41Z	P	B14	41Z-B14	Low Grade	1079	G	12S_rRNA	tRNA/rRNA	0.172	subclonal	GC	not coding	not coding	
41Z	P	B14	41Z-B14	Low Grade	1282	G	12S_rRNA	tRNA/rRNA	0.0276	subclonal	GA	not coding	not coding	
41Z	P	B14	41Z-B14	Low Grade	2534	G	16S_rRNA	tRNA/rRNA	0.704	subclonal	GA	not coding	not coding	
41Z	P	B14	41Z-B14	Low Grade	2566	C	16S_rRNA	tRNA/rRNA	0.151	subclonal	CT	not coding	not coding	
41Z	P	B14	41Z-B14	Low Grade	2871	T	16S_rRNA	tRNA/rRNA	0.0137	subclonal	TC	not coding	not coding	
41Z	P	B14	41Z-B14	Low Grade	5302	T	ND2	coding	0.129	subclonal	TC	Ile > Thr	nonsynonymous	I278T
41Z	P	B14	41Z-B14	Low Grade	6164	C	COX1	coding	0.0105	subclonal	CT	synonymous	synonymous	
41Z	P	B14	41Z-B14	Low Grade	9646	C	COX3	coding	0.0128	subclonal	CT	Ala > Val	nonsynonymous	A147V
41Z	P	B14	41Z-B14	Low Grade	9820	G	COX3	coding	0.0106	subclonal	GA	Gly > Glu	nonsynonymous	G205E
41Z	P	C1	41Z-C1	Negative	195	T	Control_region HVS2	noncoding	0.0123	subclonal	TC	not coding	not coding	
41Z	P	C1	41Z-C1	Negative	366	G	Control_region (HVS2)	noncoding	0.0356	subclonal	GA	not coding	not coding	
41Z	P	C1	41Z-C1	Negative	669	T	12S_rRNA	tRNA/rRNA	0.0323	subclonal	TC	not coding	not coding	
41Z	P	C1	41Z-C1	Negative	1200	G	12S_rRNA	tRNA/rRNA	0.0157	subclonal	GA	not coding	not coding	
41Z	P	C1	41Z-C1	Negative	3424	G	ND1	coding	0.0109	subclonal	GA	Val > Met	nonsynonymous	V40M
41Z	P	C1	41Z-C1	Negative	7829	C	COX2	coding	0.0285	subclonal	CT	Arg > Cys	nonsynonymous	R82C
41Z	P	C1	41Z-C1	Negative	8348	A	tRNA-Lys	tRNA/rRNA	0.666	subclonal	AG	not coding	not coding	
41Z	P	C2	41Z-C2	Negative	366	G	Control_region (HVS2)	noncoding	0.0264	subclonal	GA	not coding	not coding	
41Z	P	C2	41Z-C2	Negative	3607	G	ND1	coding	0.022	subclonal	GA	Gly > Ser	nonsynonymous	G101S
41Z	P	C2	41Z-C2	Negative	8715	T	ATP6	coding	0.106	subclonal	TC	synonymous	synonymous	
41Z	P	C2	41Z-C2	Negative	12417	C	ND5	coding	0.019	subclonal	I	insertion	indel	
41Z	P	C2	41Z-C2	Negative	14560	G	ND6	coding	0.0101	subclonal	GA	synonymous	synonymous	
41Z	P	C7	41Z-C7	Negative	9636	A	COX3	coding	0.134	subclonal	AG	Ile > Val	nonsynonymous	I144V
41Z	P	C7	41Z-C7	Negative	11900	G	ND4	coding	0.142	subclonal	GA	Val > Met	nonsynonymous	V381M
41Z	P	C7	41Z-C7	Negative	13177	G	ND5	coding	0.0135	subclonal	GA	Gly > Ser	nonsynonymous	G281S
41Z	P	C7	41Z-C7	Negative	13628	T	ND5	coding	0.0428	subclonal	TG	Leu > Arg	nonsynonymous	L431R
41Z	P	C12	41Z-C12	Low Grade	366	G	Control_region (HVS2)	noncoding	0.0256	subclonal	GA	not coding	not coding	
41Z	P	C12	41Z-C12	Low Grade	4869	C	ND2	coding	0.0106	subclonal	CA	Gln > Lys	nonsynonymous	Q134K
41Z	P	C12	41Z-C12	Low Grade	12125	G	ND4	coding	0.06	subclonal	GA	Gly > Ter	nonsynonymous	G456Ter
41Z	P	C12	41Z-C12	Low Grade	16390	G	Control_region	noncoding	0.213	subclonal	GA	not coding	not coding	
41Z	P	C21	41Z-C21	Negative	4869	C	ND2	coding	0.958	subclonal	CA	Gln > Lys	nonsynonymous	Q134K
41Z	P	C21	41Z-C21	Negative	7609	T	COX2	coding	0.0131	subclonal	TC	synonymous	synonymous	
41Z	P	C21	41Z-C21	Negative	8959	G	ATP6	coding	0.0763	subclonal	GA	Glu > Lys	nonsynonymous	E145K
41Z	P	C21	41Z-C21	Negative	9531	A	COX3	coding	0.0127	subclonal	I	insertion	indel	
41Z	P	C21	41Z-C21	Negative	9755	G	COX3	coding	0.115	subclonal	GA	synonymous	synonymous	
41Z	P	C21	41Z-C21	Negative	10068	G	ND3	coding	0.0336	subclonal	GA	Ala > Thr	nonsynonymous	A4T
41Z	P	C21	41Z-C21	Negative	12773	G	ND5	coding	0.0129	subclonal	GA	Gly > Asp	nonsynonymous	G146D
41Z	P	D6	41Z-D6	High Grade	366	G	Control_region (HVS2)	noncoding	0.0249	subclonal	GA	not coding	not coding	
41Z	P	D6	41Z-D6	High Grade	3054	G	16S_rRNA	tRNA/rRNA	0.0472	subclonal	GA	not coding	not coding	
41Z	P	D6	41Z-D6	High Grade	6734	G	COX1	coding	0.0707	subclonal	GA	synonymous	synonymous	
41Z	P	D6	41Z-D6	High Grade	7293	G	COX1	coding	0.0464	subclonal	GA	Ala > Thr	nonsynonymous	A464T
41Z	P	D6	41Z-D6	High Grade	13227	C	ND5	coding	0.449	subclonal	CT	synonymous	synonymous	
41Z	P	D6	41Z-D6	High Grade	14666	T	ND6	coding	0.966	subclonal	TC	Tyr > Cys	nonsynonymous	Y3C
41Z	P	D6	41Z-D6	High Grade	15777	G	CYTB	coding	0.0152	subclonal	GA	Ser > Asn	nonsynonymous	S343N
41Z	P	D7	41Z-D7	Low Grade	366	G	Control_region (HVS2)	noncoding	0.0244	subclonal	GA	not coding	not coding	
41Z	P	D7	41Z-D7	Low Grade	5917	G	COX1	coding	0.0192	subclonal	GA	Arg > His	nonsynonymous	R5H
41Z	P	D7	41Z-D7	Low Grade	6734	G	COX1	coding	0.0402	subclonal	GA	synonymous	synonymous	
41Z	P	D7	41Z-D7	Low Grade	7293	G	COX1	coding	0.019	subclonal	GA	Ala > Thr	nonsynonymous	A464T
41Z	P	D7	41Z-D7	Low Grade	13227	C	ND5	coding	0.14	subclonal	CT	synonymous	synonymous	
41Z	P	D7	41Z-D7	Low Grade	14666	T	ND6	coding	0.298	subclonal	TC	Tyr > Cys	nonsynonymous	Y3C
41Z	P	D8	41Z-D8	High Grade	366	G	Control_region (HVS2)	noncoding	0.0269	subclonal	GA	not coding	not coding	
41Z	P	D8	41Z-D8	High Grade	3565	A	ND1	coding	0.0682	subclonal	I	insertion	indel	
41Z	P	D8	41Z-D8	High Grade	9531	A	COX3	coding	0.0194	subclonal	I	insertion	indel	
41Z	P	D8	41Z-D8	High Grade	10599	G	ND4L	coding	0.0125	subclonal	GA	Ala > Thr	nonsynonymous	A44T
41Z	P	D8	41Z-D8	High Grade	14503	T	ND6	coding	0.0143	subclonal	I	insertion	indel	
41Z	P	D21	41Z-D21	Negative	366	G	Control_region (HVS2)	noncoding	0.0236	subclonal	GA	not coding	not coding	
41Z	P	D21	41Z-D21	Negative	3882	G	ND1	coding	0.629	subclonal	GA	synonymous	synonymous	
41Z	P	D21	41Z-D21	Negative	6528	C	COX1	coding	0.0177	subclonal	CT	synonymous	synonymous	

Supplementary Figure Legends

Figure S1. Histology of UC colon tissue Representative biopsies from each dysplastic group are shown. White light microscope pictures were taken from hematoxylin and eosin stained slides of FFPE tissue. FFPE blocks were made from tissue adjacent to the frozen biopsies used for mutational analysis. UC: Ulcerative Colitis. NP: Non-Progressor. P: Progressor. Neg: Negative for Dysplasia. LGD: Low Grade Dysplasia. HGD: High Grade Dysplasia.

Figure S2. Clonal expansions Colon maps for each of the Progressor colons as diagrammed by the pathologist after colectomy. Each box corresponds to an individual biopsy and is color-coded according to histological findings: Neg: negative for dysplasia, IND: indefinite for dysplasia, LGD: low-grade dysplasia, HGD: high-grade dysplasia, and cancer. Biopsies are named based on the coordinates defined by columns (letters) and rows (numbers). Columns correspond to the diameter of the colon divided into 3-4 sections and are ~2cm apart. Rows indicate colon levels and are evenly spaced ~2-5cm along the length of the organ. For each colon, the biopsies analyzed are indicated with a box with the biopsy name. Several mtDNA mutations were found in multiple biopsies from the same colon at varying frequencies, indicating clonal expansions. Biopsies sharing mutations are connected with brackets and the respective mutant allele frequencies for each commonly mutated gene are indicated in tables.

Figure S3. Clonal and subclonal mutations by biopsy Each bar corresponds to a biopsy in the study. **A,** The total number of clonal and subclonal mutations per biopsy is shown by level of Mutant Allele Frequency (MAF) **B,** The total number of clonal and subclonal mutations is shown by location in the mtDNA, including tRNA or rRNA, non-coding region and coding region. NP:

Non Progressor, Neg: negative for dysplasia, LGD: low-grade dysplasia, HGD: high-grade dysplasia.

Figure S4. Association between mutation number and total DCS nucleotides sequenced A, Number of clonal and subclonal mutations compared to total number of DCS nucleotides sequenced for each biopsy. Mutations are color coded by MAF. **B,** Number of VLF mutations compared to total number of DCS nucleotides sequenced for each biopsy. Because of the high number of C>A mutations found in some samples, but not others, correlations by C>A high and low are shown separately. C>A high was defined as biopsies in which the frequency of C>A mutations in the heavy or light strand of mtDNA was > $5x10^{-5}$. Both C>A low and C>A high biopsies showed a significant association between number of mutations and total DCS nucleotides sequenced (C>A low: r=0.89, Spearman p-value< 0.001. C>A high: r=0.43, Spearman p-value=0.035). DCS: Duplex Consensus Sequence. MAF: mutant allele frequency; VLF: very low frequency

Figure S5. Association of clonal and subclonal mutations with clinical variables
Associations between clonal and subclonal mutations and clinical variables including age (A), sex (B), disease duration (C), acute inflammation (D) and chronic inflammation (E) are shown. (F) The number of clonal and subclonal mutations for each biopsy is shown, with biopsies grouped by patient. Along the x-axis, acute and chronic inflammation scores are indicated. For acute inflammation, scores were assigned the following numeric equivalents: none – 1, mild – 2, moderate – 3. For chronic inflammation, scores were assigned the following numeric equivalents: none – 1, low – 2, high – 3.

Figure S6. MAF of individual clonal and subclonal mutations The MAF of each individual mutation (MAF>0.01) is shown for each biopsy. Biopsies are grouped in the x-axis by biopsy type. Each column corresponds to one biopsy and the circles correspond to different mtDNA mutations identified in that biopsy. Mutations are color coded to indicate whether they are non-coding, synonymous, or non-synonymous. MAF: mutant allele frequency.

Figure S7. C>A Mutation frequencies by biopsy The VLF C>A mutation frequency on either strand, as well as all other combined mutations, is shown for each biopsy analyzed, organized by dysplastic grade. VLF: very low frequency

Figure S8. Association of very low frequency mutations with clinical variables Associations between VLF mutations and clinical variables including age (A), sex (B), disease duration (C), acute inflammation (D) and chronic inflammation (E) are shown. (F) The number of clonal and subclonal mutations for each biopsy is shown, with biopsies grouped by patient. Along the x-axis, acute and chronic inflammation scores are also shown. For acute inflammation, scores were assigned the following numeric equivalents: none – 1, mild – 2, moderate – 3. For chronic inflammation, scores were assigned the following numeric equivalents: none – 1, low – 2, high – 3.

Figure S9. Mutation signature by mutation type The substitution rate for each of the 96 possible substitution classes is shown for each biopsy type. L: light strand, H: heavy strand.

Figure S10. VLF transitions and transversions by mtDNA region and biopsy type The mutation frequency for each biopsy type is shown by either transition or transversion mutations

and by D-loop and non-D-loop region. In the non-D-loop, the mtDNA mutational signature detected with transitions (C>T in heavy strand and T>C in light strand) significantly decreased with progression from Neg to Cancer (* $p<6.6 \times 10^{-4}$ for the "within-substitutions trend", calculated with GEE permutation tests with log transformed values). VLF: very low frequency.

Figure S11. VLF mutation frequency and gene size For each biopsy type, non-synonymous and synonymous mutation frequency for each mtDNA-encoded gene was plotted by gene size. Mutation frequency was calculated as the number of mutations with MAF<0.01 (non-synonymous or synonymous) within each given gene divided by the total number of DCS nucleotides sequenced. VLF: very low frequency

Figure S12. Model of mitochondrial mutation progression during UC carcinogenesis Normal colons may accumulate a small number of unique mutations or very small clones. Non-Progressors accumulate larger numbers of very low frequency (VLF) mutations, indicated by circles of different colors, which represent mitochondrial genomes harboring different mutations. In UC Progressors, more mutations with greater abundance begin to amass due to clonal expansion in negative for dysplasia tissue, reaching a peak in low grade in terms of number and diversity of clones. This stage might act as a bottleneck, as further progression to high-grade dysplasia and cancer appears to be characterized by a decrease of deleterious mutations and the outgrowth of clones carrying non-deleterious mitochondrial DNA mutations.

Supplementary Figures

Figure S1. Histology of UC colon tissue

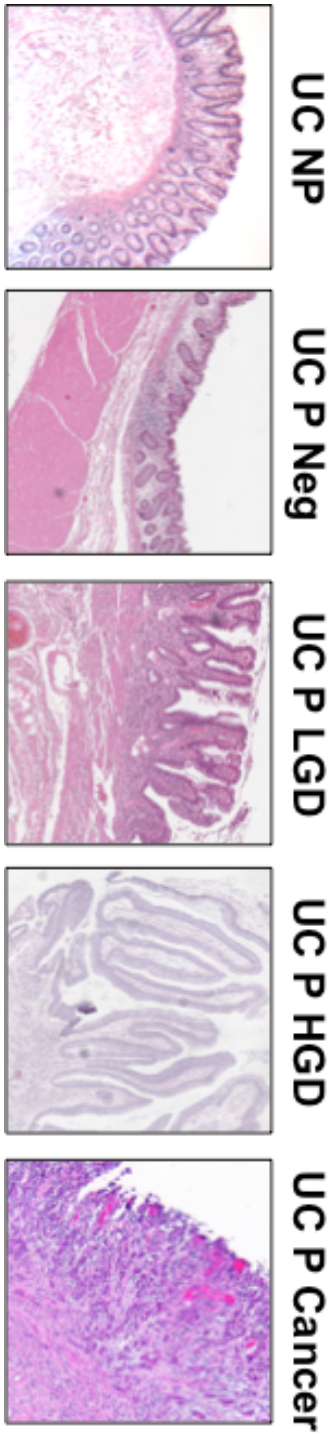
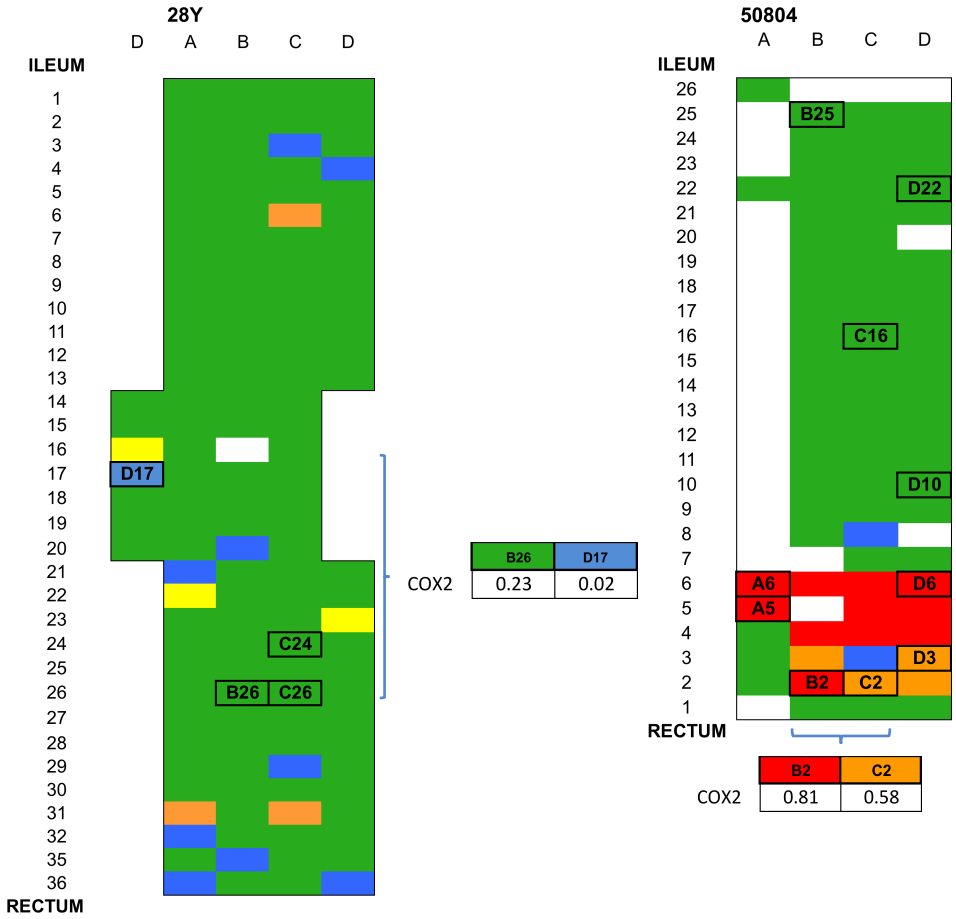
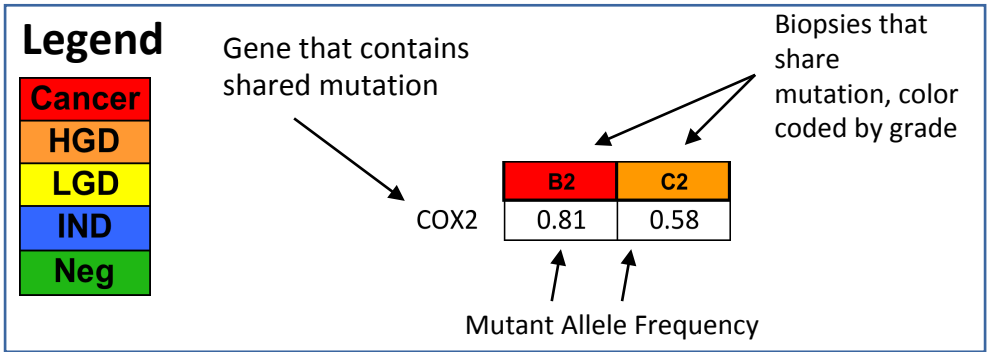


Figure S2. Clonal expansions


28Y
D A B C D
ILEUM
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30 31 32 35 36
RECTUM
D17
C24
B26 C26
COX2
B26 D17
0.23 0.02
50804
A B C D
ILEUM
26 25 24 23 22 21 20 19 18 17 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
RECTUM
B25
D22
C16
D10
A6
A5
D6
D3
B2 C2
COX2
B2 C2
0.81 0.58


Legend
Cancer
HGD
LGD
IND
Neg
Gene that contains shared mutation
Biopsies that share mutation, color coded by grade
COX2
B2 C2
0.81 0.58
Mutant Allele Frequency

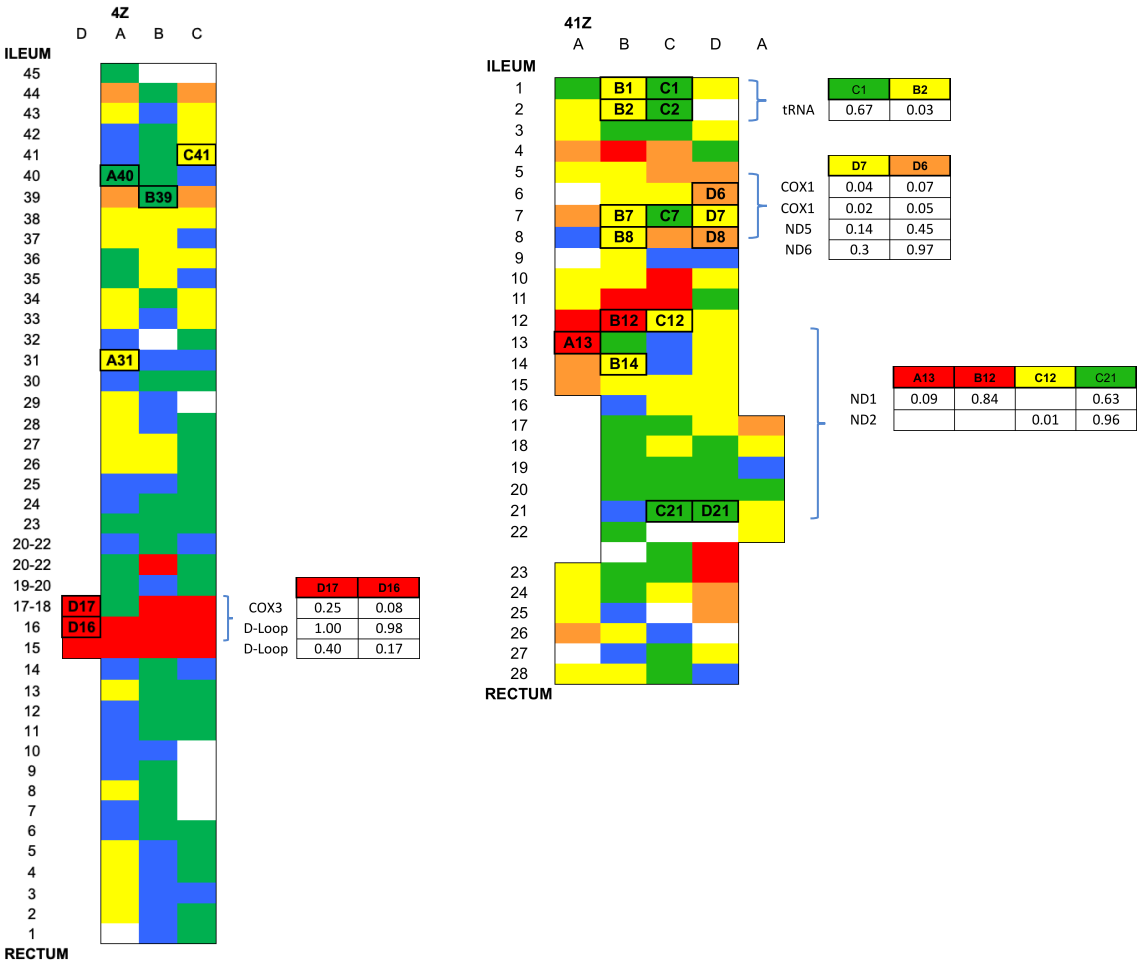

4Z
D A B C
ILEUM
45 44 43 42 41 40 39 38 37 36 35 34 33 32 31 30 29 28 27 26 25 24 23 20-22 20-22 19-20 17-18 16 15 14 13 12 11 10 9 8 7 6 5 4 3 2 1
RECTUM
C41
A40
B39
A31
D17
D16
COX3 D17 0.25 D16 0.08
D-Loop 1.00 0.98
D-Loop 0.40 0.17
41Z
A B C D A
ILEUM
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28
RECTUM
B1 C1
B2 C2
B7 C7 D7
B8 D8
D6
B12 C12
A13
B14
C21 D21
tRNA C1 0.67 B2 0.03
COX1 D7 0.04 D6 0.07
COX1 0.02 0.05
ND5 0.14 0.45
ND6 0.3 0.97
ND1 A13 0.09 B12 0.84 C12 C21 0.63
ND2 0.01 0.96

Figure S3. Clonal and subclonal mutations by biopsy

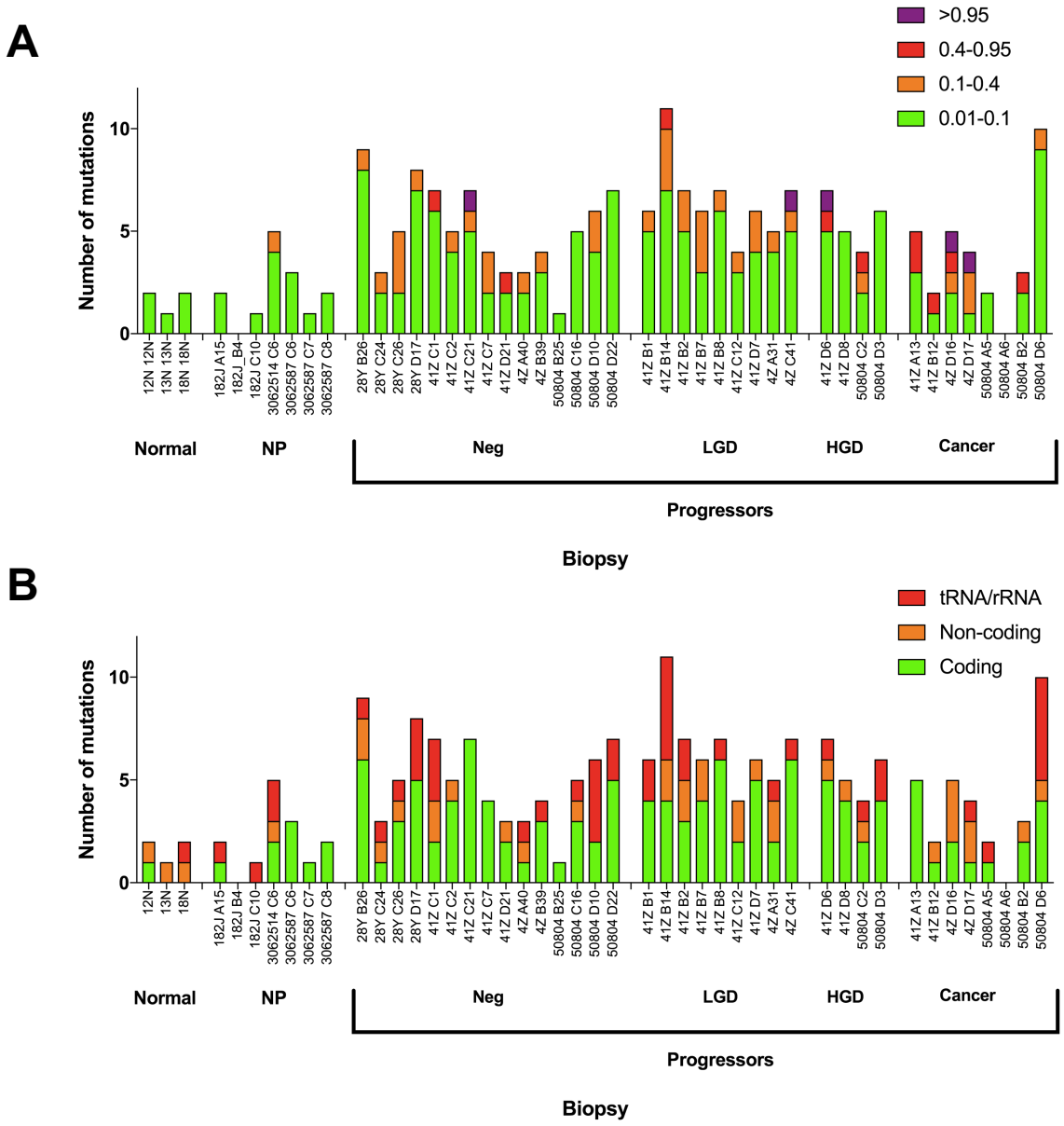


Figure S4. Association between mutation number and total DCS nucleotides sequenced

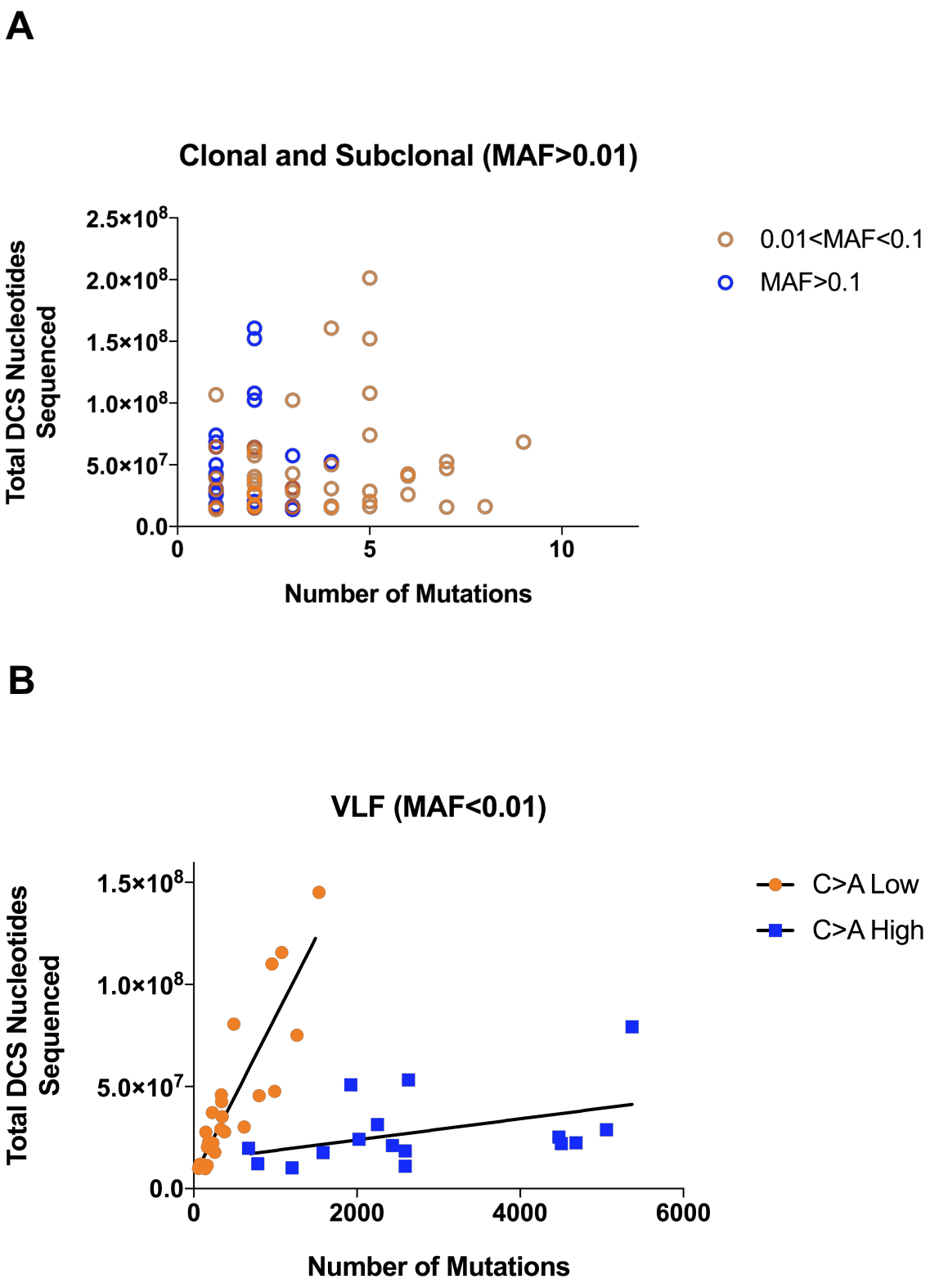


Figure S5. Association of clonal and subclonal mutations with clinical variables

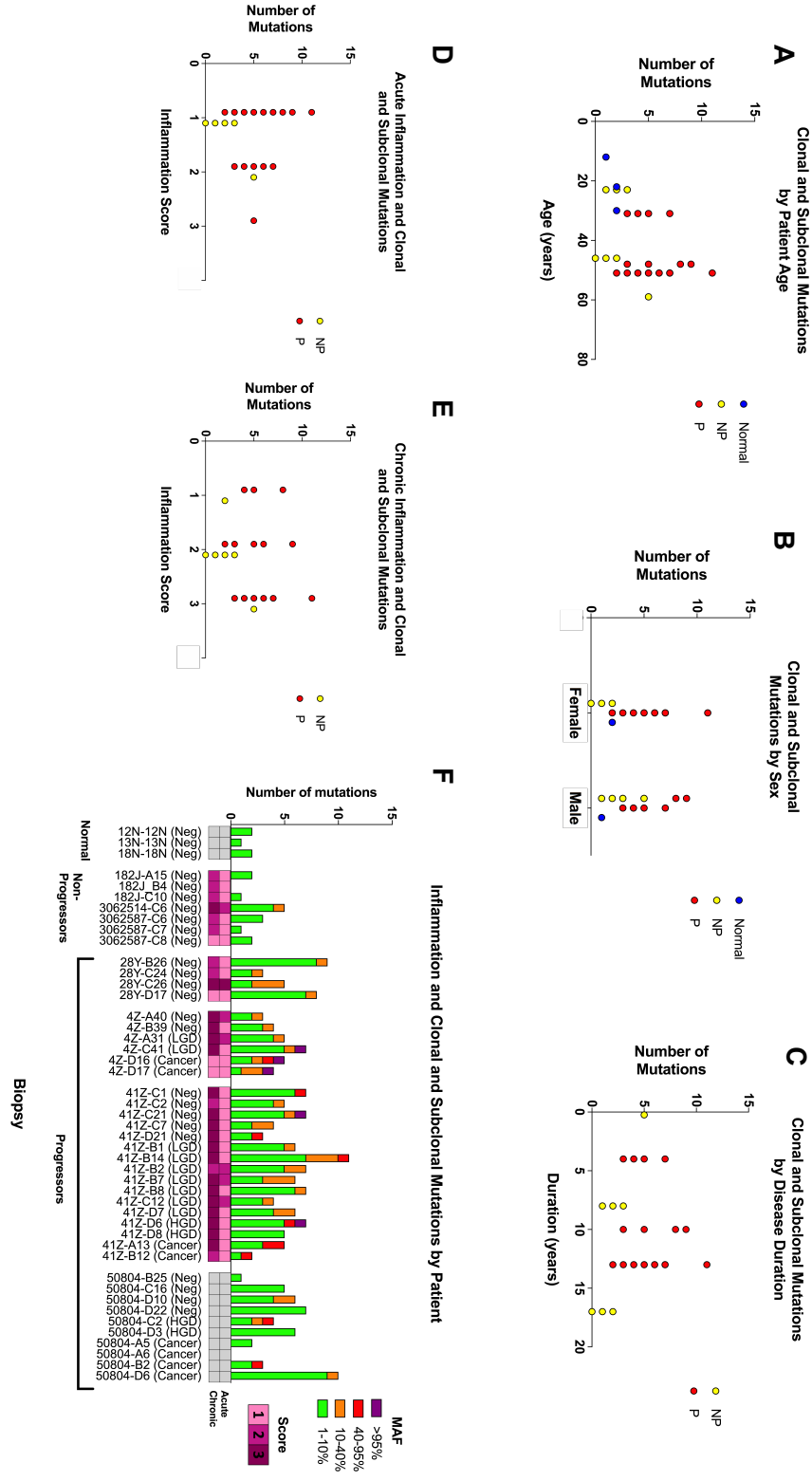


Figure S6. MAF of individual clonal and subclonal mutations

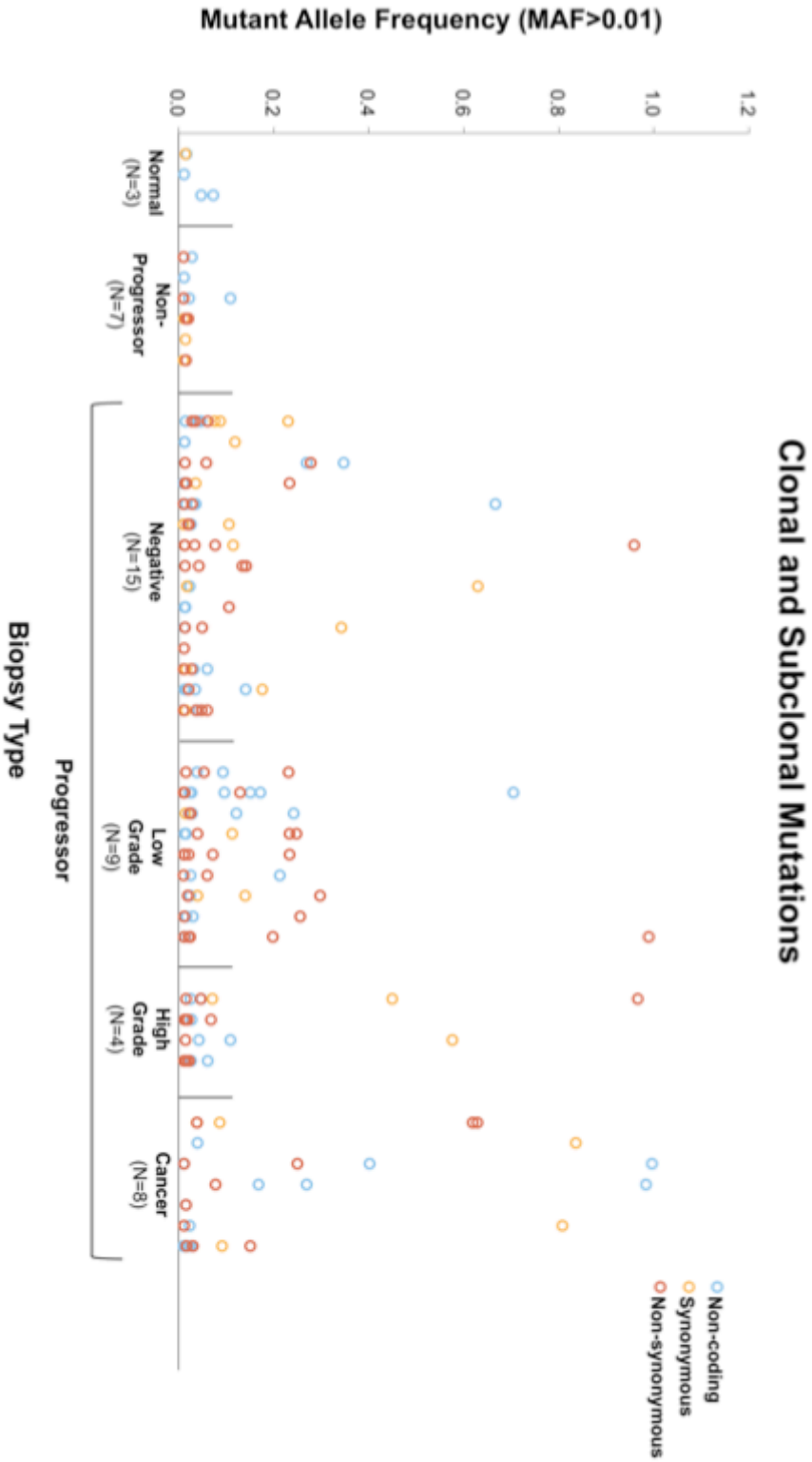


Figure S7. C>A Mutation frequencies by biopsy

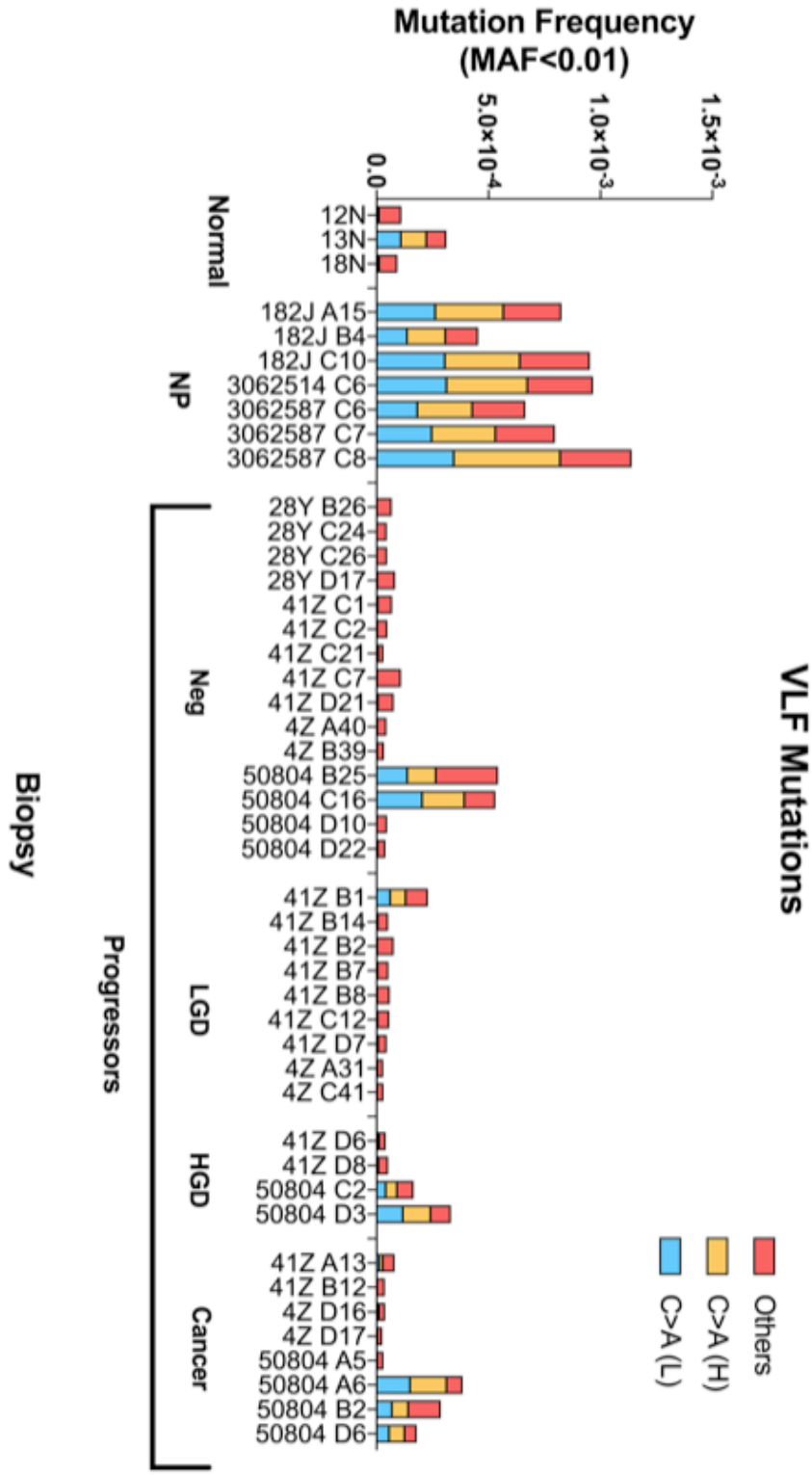

Mutation Frequency
(MAF<0.01)
0.0
5.0×10-4
1.0×10-3
1.5×10-3
VLF Mutations
Others
C>A (H)
C>A (L)
Biopsy
Normal
12N
13N
18N
NP
182J A15
182J B4
182J C10
3062514 C6
3062587 C6
3062587 C7
3062587 C8
Progressors
Neg
28Y B26
28Y C24
28Y C26
28Y D17
41Z C1
41Z C2
41Z C21
41Z C7
41Z D21
4Z A40
4Z B39
50804 B25
50804 C16
50804 D10
50804 D22
LGD
41Z B1
41Z B14
41Z B2
41Z B7
41Z B8
41Z C12
41Z D7
4Z A31
4Z C41
HGD
41Z D6
41Z D8
50804 C2
50804 D3
Cancer
41Z A13
41Z B12
4Z D16
4Z D17
50804 A5
50804 A6
50804 B2
50804 D6

Figure S8. Association of very low frequency mutations with clinical variables

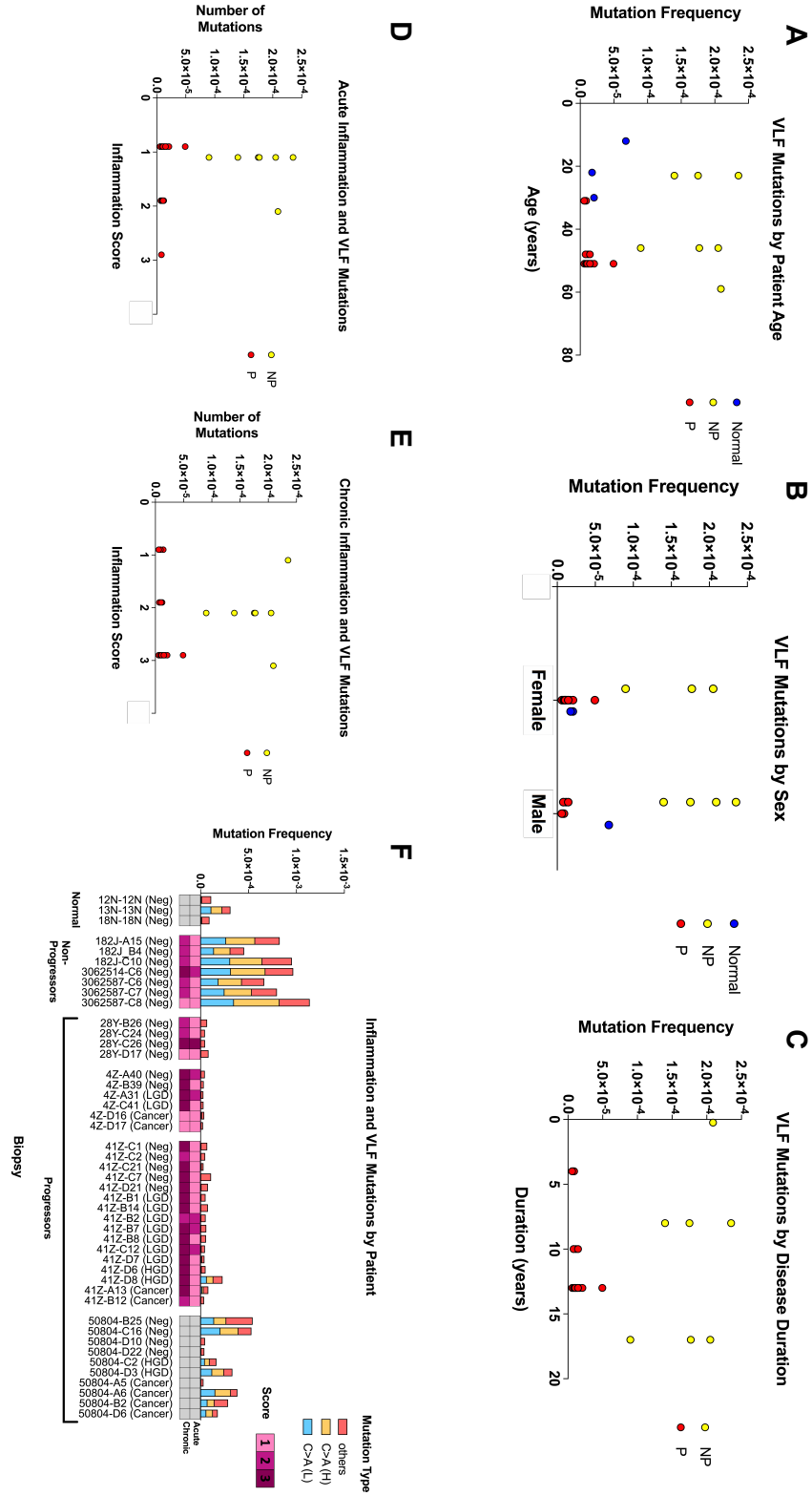


Figure S9. Mutation signature by mutation type

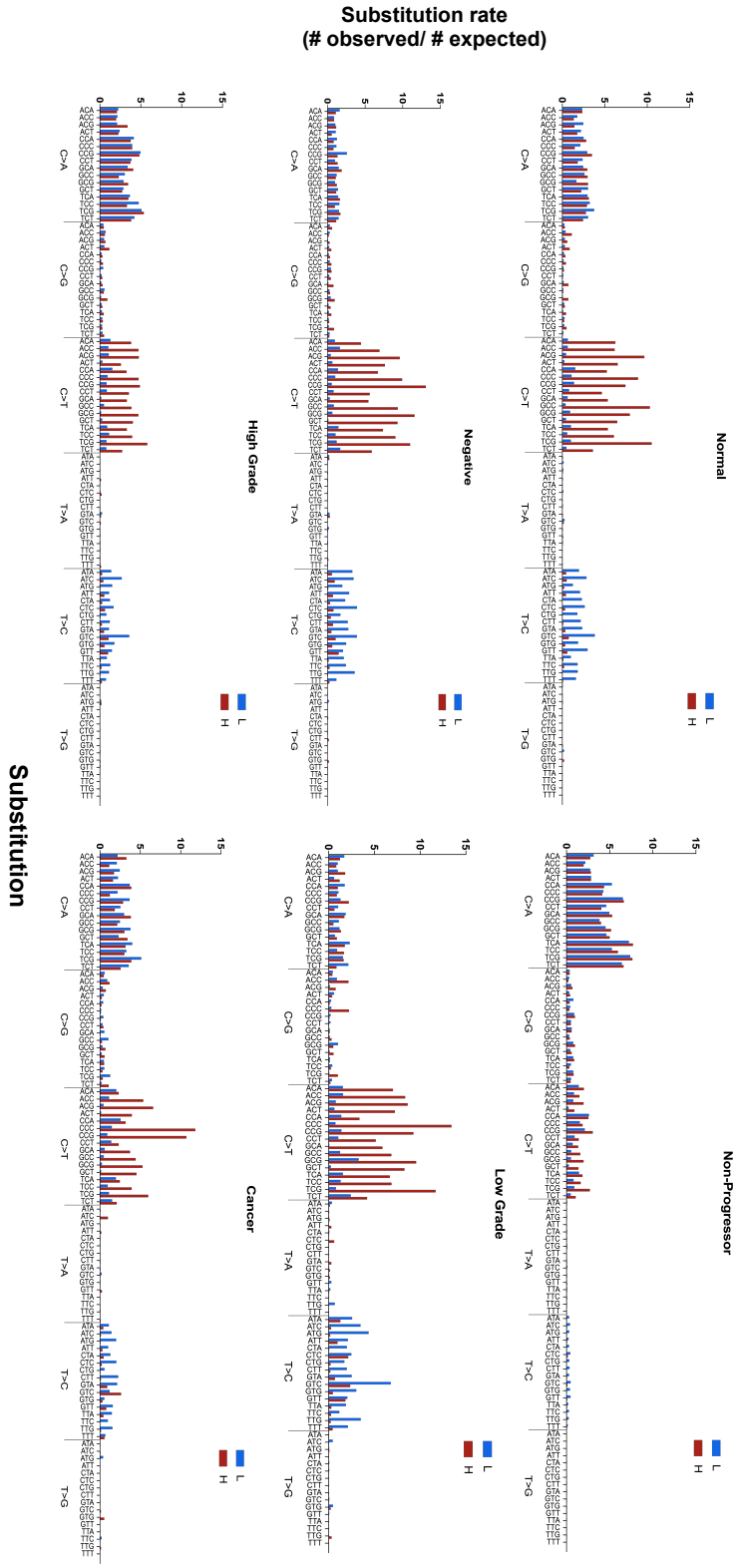


Figure S10. VLF transitions and transversions by mtDNA region and biopsy type

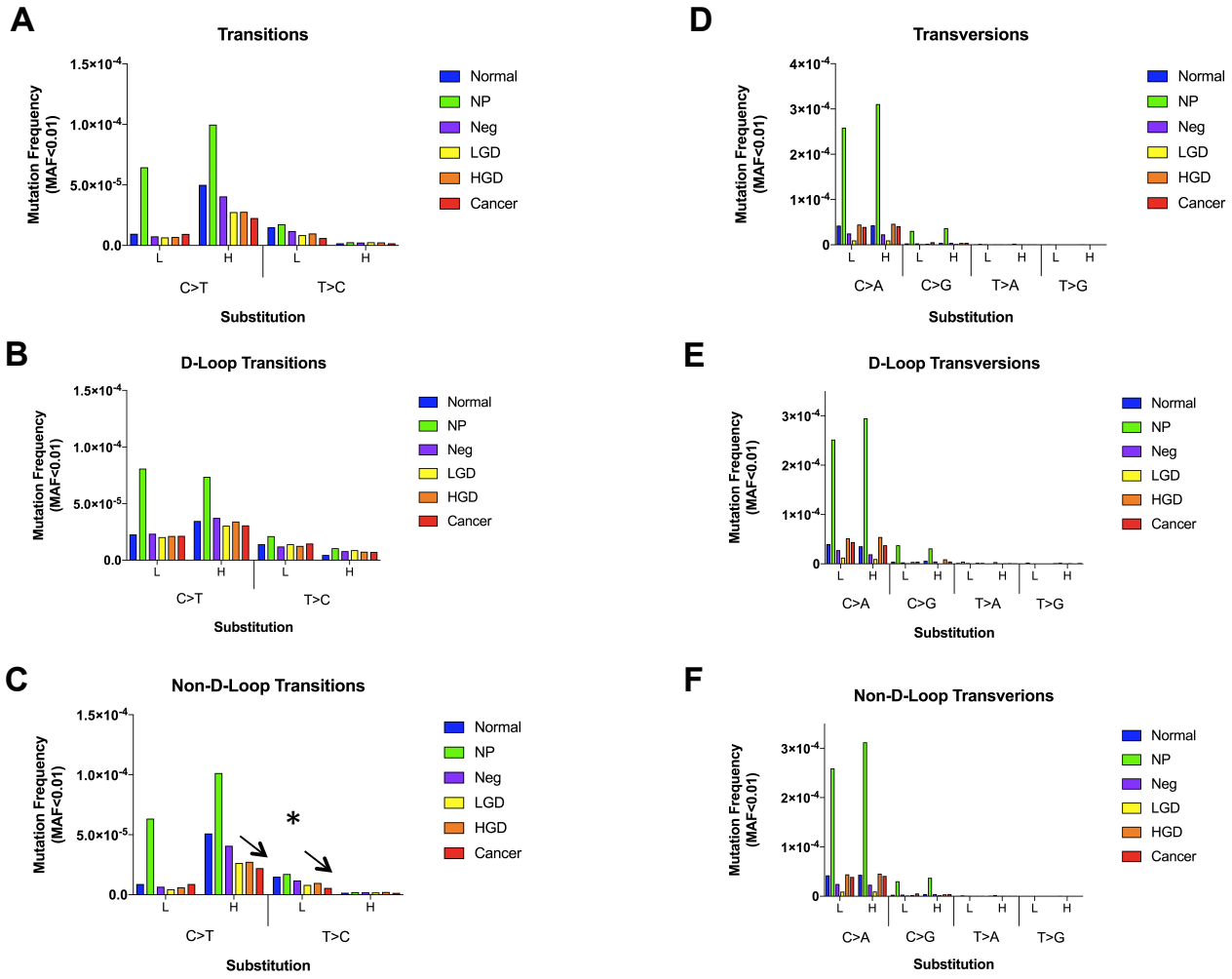


Figure S11. VLF mutation frequency and gene size

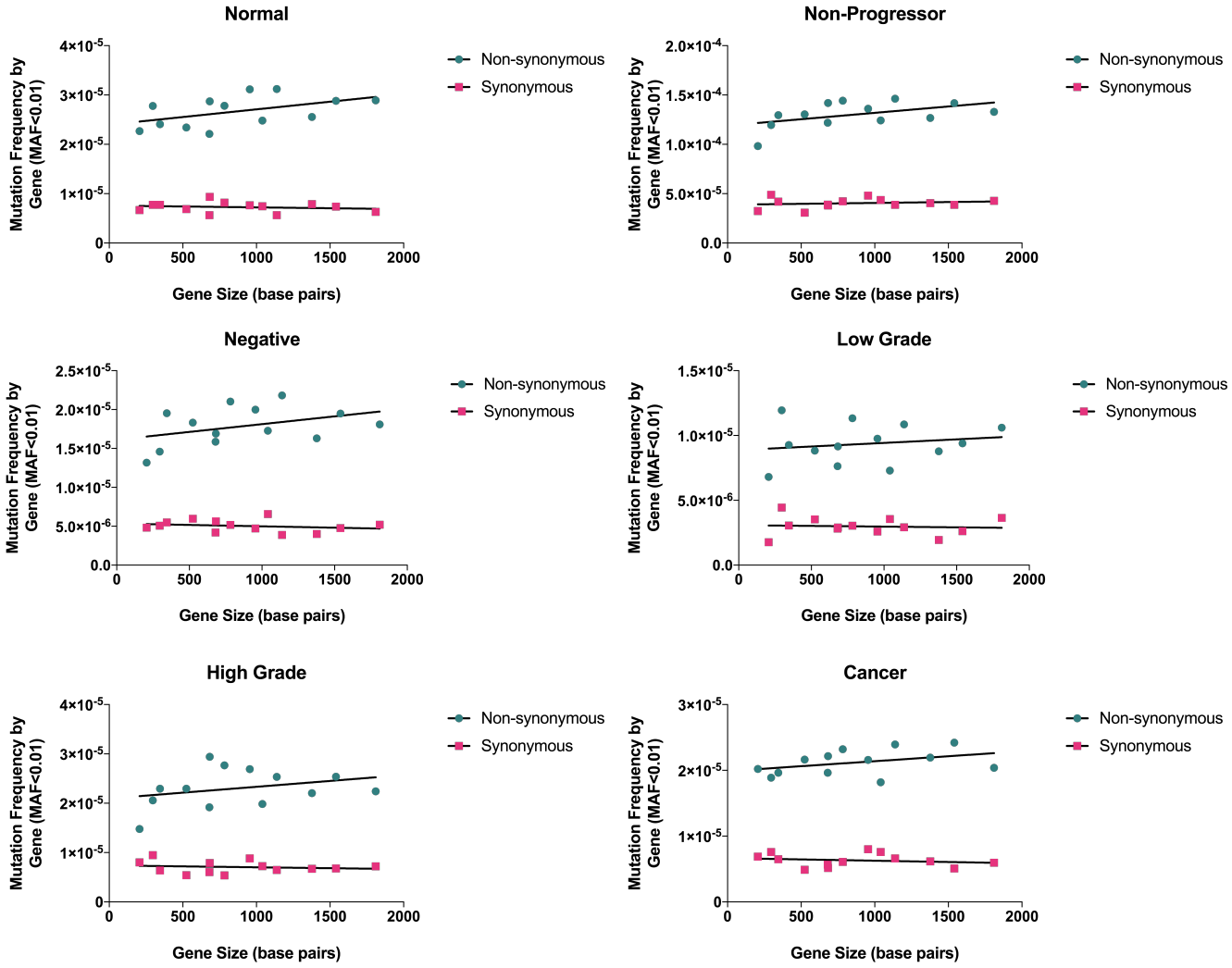
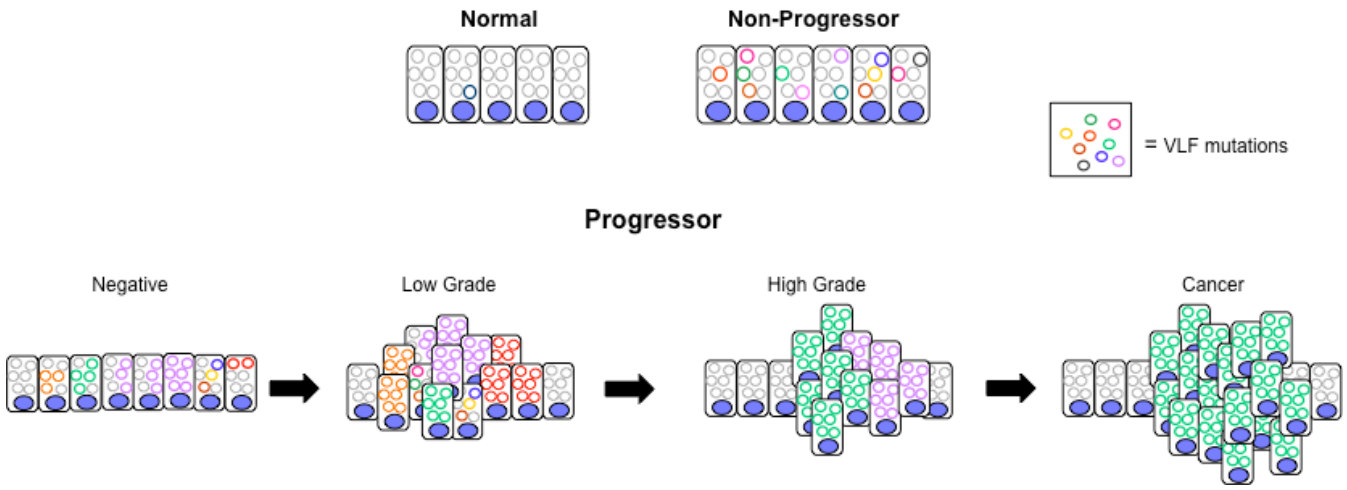


Figure S12. Model of mitochondrial mutation progression during UC carcinogenesis



Chapter 3: Ultra-sensitive sequencing for cancer detection reveals progressive clonal selection in normal tissue over a century of human lifespan

Jesse J. Salk[1,2], Kaitlyn Loubet-Senear[3†], Elisabeth Maritschnegg[4††], Charles C. Valentine[2], Lindsey N. Williams[2], Reinhard Horvat[5], Adriaan Vanderstichele[6], Daniela Nachmanson[3†††], Kathryn T. Baker[3], Mary J. Emond[7], Emily Loter[8], Thierry Soussi[9,10,11], Lawrence A. Loeb[3], Robert Zeillinger[4], Paul Speiser[4] and Rosa Ana Risques[3*]

[1] Division of Medical Oncology, University of Washington, Seattle, WA 98195, USA.
[2] TwinStrand Biosciences, Seattle, WA 98121, USA.
[3] Department of Pathology, University of Washington, Seattle, WA 98195, USA.
[4] Molecular Oncology Group, Department of Obstetrics and Gynecology, Comprehensive Cancer Center – Gynecologic Cancer Unit, Medical University of Vienna, Vienna, Austria.
[5] Department Pathology, Medical University of Vienna, Vienna, Austria.
[6] Department of Gynecologic Oncology, Leuven Cancer Institute, University Hospitals Leuven, KU, Leuven, Belgium.
[7] Department of Statistics, University of Washington, Seattle, WA 98195, USA.
[8] Department of Pathology, Seattle Children's Hospital, Seattle, WA 98105, USA.
[9] Sorbonne Université, UPMC Univ Paris 06, F- 75005 Paris, France.
[10] Department of Oncology-Pathology, Karolinska Institutet, Stockholm, Sweden.
[11] INSERM, U1138, Centre de Recherche des Cordeliers, Paris, France.
† Current address: Department of Molecular and Cellular Biology, Harvard University, Cambridge, MA, USA
†† Current address: VIB-KU Leuven Center for Brain & Disease Research, Belgium
††† Current address: Bioinformatics and Systems Biology Graduate Program, University of California San Diego, La Jolla, CA, USA

*To whom correspondence should be addressed: rrisques@uw.edu

SHORT TITLE: Cancer-like mutations increase in normal aging

KEYWORDS

Duplex Sequencing, error-corrected sequencing, deep sequencing, next generation sequencing, NGS, cancer driver mutation, *TP53,* clonal evolution, preneoplastic, early detection, cancer screening, cancer biomarker, high-grade serous ovarian cancer, HGSOC, uterine lavage, aging, gynecologic oncology.

Available as preprint:
Ultra-sensitive sequencing for cancer detection reveals progressive clonal selection in normal tissue over a century of human lifespan. Jesse J Salk, Kaitlyn Loubet-Senear, Elisabeth Maritschnegg, Charles C Valentine, Lindsey N Williams, Reinhard Horvat, Adriaan Vanderstichele, Daniela Nachmanson, Kathryn T Baker, Mary J Emond, Emily Loter, Thierry Soussi, Lawrence A Loeb, Robert Zeillinger, Paul Speiser, Rosa Ana Risques. bioRxiv 457291; doi: https://doi.org/10.1101/457291

Abstract

High accuracy next-generation DNA sequencing promises a paradigm shift in early cancer detection by enabling the identification of mutant cancer molecules in minimally invasive body fluid samples. We demonstrate 80% sensitivity for ovarian cancer detection using ultra-accurate Duplex Sequencing to identify *TP53* mutations in uterine lavage. However, in addition to tumor DNA, we also detect low frequency *TP53* mutations in nearly all lavages from women with and without cancer. These mutations increase with age and share the selection traits of clonal *TP53* mutations commonly found in human tumors. We show that low frequency *TP53* mutations exist in multiple healthy tissues, from newborn to centenarian, and progressively increase in abundance and pathogenicity with older age across tissue types. Our results illustrate that subclonal cancer evolutionary processes are a ubiquitous part of normal human aging and great care must be taken to distinguish tumor-derived, from age-associated mutations in high sensitivity clinical cancer diagnostics.

Introduction

Worldwide more than a quarter of a million new cases of ovarian cancer are diagnosed each year and two thirds of these women die from the disease (*1*). This high mortality is largely due to the high frequency of metastasis before symptoms lead to diagnosis and a lack of effective screening methods. More than 60% of cases are diagnosed at an advanced stage, when the 5-year survival rate is only 29% (*2*). In contrast, survival for women with localized disease is 92%, indicating that early ovarian cancer detection could vastly decrease mortality, yet diagnosis at this stage is rare.

Despite significant efforts, unlike colonoscopies, pap smears and mammograms that are evidence-based, mortality-reducing methods for early detection of colon, cervical and breast cancer, respectively, no ovarian cancer screening technique has demonstrated sufficient sensitivity and specificity for use in the general population (*3*). The approach most explored involves a combination of testing for the serum protein CA-125 and transvaginal ultrasound, but the US Preventive Services Task Force recommends against its use (*4*) because it has not been shown to reduce mortality and may result in harms due to false positives, such as unnecessary surgeries in cancer-free women (*5*). Better tools for early ovarian cancer detection remains an urgent and unmet clinical need.

In the last several years an increasing number of cancers have been found to shed cells or DNA into blood or other body fluids where they can be non-invasively detected, a concept often termed “liquid biopsy” (*6*). Proof-of-principle for this approach in ovarian cancer screening was initially accomplished via identification of tumor-derived mutations in DNA extracted from routine Pap smear fluid (*7*). Although the sensitivity for ovarian cancer detection was only 41%, these findings supported the exciting possibility that ovarian cancer could be detected based on the genetic identification of cancer cells disseminated into the gynecological tract. A follow-up study

recently reported that improved sensitivity, up to 63%, could be obtained by combining mutation detection in Pap tests and plasma. In addition, sampling with an intrauterine brush also improved sensitivity, probably due to increased tumor cell recovery by more proximal collection to the anatomical site of tumors (*8*).

An alternate means for tumor cell collection, developed by members of our team several years ago, consists on trans-cervical lavage of the uterine cavity (**Fig. 1A**). This method improves the efficiency of collection by rinsing all surfaces, including those near, and even into, the fallopian tubes where most early serous ovarian cancers are believed to initiate (*9*). This lavage technique demonstrated 80% sensitivity for ovarian cancer detection. The challenge, however, was that cancer-derived mutations, particularly those from early stage tumors, were often present in a very small fraction of the total lavage DNA. To detect these mutations, digital droplet PCR (ddPCR) was required, which is an extremely sensitive method but not practical for prospective screening because the tumor mutation needs to be known *a priori*.

Next-generation DNA sequencing (NGS) is a widely used, variant-agnostic form of mutation detection, but has a background error rate of up to ~1%, which precludes confident detection of lower frequency mutations (*10*). Of the mutations comprising the 80% sensitivity achieved in our previous study, conventional NGS missed approximately one quarter (*9*). Currently, the most sensitive NGS method is Duplex Sequencing (DS) (*10*), which employs double-stranded molecular barcodes for error correction and decreases the error rate of sequencing from 10^{-3} to $<10^{-7}$ (*11, 12*). We previously demonstrated that DS can detect ovarian cancer-derived mutations in DNA extracted from peritoneal fluid at frequencies as low as one tumor mutation per 25,000 normal genomes (*13*). The extreme sensitivity of the technique also led to the discovery of prevalent, yet very low frequency (<0.01%) *TP53* mutations in both the peritoneal fluid and peripheral blood from healthy women. These 'biological background' mutations resembled *TP53*

mutations found in cancers, but appeared to result from the normal aging process. This observation was among the first of an emerging body of literature that has identified age-related, cancer-associated mutations within non-cancerous tissue (*14*).

In the present study, we combine the most sensitive reported sampling technique for ovarian cancer detection, uterine lavage, with the highest accuracy sequencing technology available, DS. High- grade serous ovarian carcinoma (HGSOC) is both the most common and most deadly histological type, accounting for 70-80% of ovarian cancer deaths (*15*). With this in mind, we focus our sequencing efforts solely on *TP53* because >98% of HGSOCs carry an inactivating mutation in this relatively small and well-characterized gene (*16, 17*) and because *TP53* mutations have been routinely found in microscopic carcinoma-*in-situ* precursor lesions in the fallopian tubes, indicating that they are among the earliest genetic events in HGSOC formation (*18, 19*).

The primary goal of this study is to demonstrate the clinical and technical feasibility of using DS to deeply sequence *TP53* from uterine lavage as a potential screening test for ovarian cancer detection. We capitalize on the extreme sensitivity of ultra-accurate DS to identify cancer-derived mutations as well as to uniquely detect low frequency, age-related mutations that might impact diagnostic specificity. To better understand the extent and nature of these 'biological background' mutations, we perform a detailed characterization of somatic *TP53* mutations in multiple gynecologic tissues from women without ovarian cancer of ages spanning a century of human lifetime.

Results

Study design and technology rational

Prior work by members of our group demonstrated 80% sensitivity to detect ovarian cancer, including small volume, early stage disease, by using a combination of uterine lavage and mutation analysis via massively parallel sequencing and ddPCR(*9*). While highly sensitive, ddPCR requires baseline knowledge of the specific mutations sought and thus is not practical for cancer screening. In the current study, we tested whether ultra-high accuracy DS could identify ovarian cancer mutations in uterine lavage with sensitivity comparable to ddPCR, but without prior knowledge of the tumor mutation.

We used DS to examine the coding region of *TP53* in DNA extracted from the lavage cell pellet of 10 women with ovarian cancer and 11 controls under blinded conditions. Most of these samples were included in the original study (*9*) (**Table S1**). DS employs special adapters with double-stranded molecular barcodes, which allow the identification of sequencing reads that were derived from both strands of each starting DNA molecule. Mutations are only scored if they are present in the majority of reads from both DNA strands, effectively eliminating sequencing and PCR artifacts (**Fig. 1B**). The estimated error rate of DS is below one-in-ten-million (*11*), which allows for extreme sensitivity and specificity of mutation detection when carrying out high depth sequencing (**Fig. S1**).

Duplex Sequencing detects ovarian cancer mutations in uterine lavages with high sensitivity

To illustrate the superior accuracy of DS compared with standard Illumina sequencing, an example of a uterine lavage sample (case 6) processed by both methods is shown side-by-side in **Fig. 1C-D**. Whereas every nucleotide position in the gene artefactually appears mutated in 0.1-1%

of molecules with standard sequencing, DS eliminates these tens of thousands of erroneous mutations to reveal the known tumor mutation at a mutant allele frequency (MAF) of 0.15%, a value very close to the frequency previously determined by ddPCR (0.12%, case 6 in **Table 1**) (*9*).

Among the 10 lavages from women with ovarian cancer, we identified the expected tumor mutation (fuchsia bars) in 8, matching the 80% sensitivity of the previous study (**Fig. 1E**). In the subset of these lavages that had been analyzed by conventional NGS and/or ddPCR, we confirmed tumor mutations at similar allele frequencies in most cases (**Table 1**, top). In addition to the tumor mutations, in nearly all lavages from women with and without tumors we identified very low frequency *TP53* mutations (blue bars) (**Fig. 1F**). To confirm that these mutations were not due to technical errors, two of the mutations identified in controls (lavages con2 and con7) were assessed by ddPCR (**Table 1**, bottom). This orthogonal assay demonstrated that these mutations, present at a comparable frequency <0.1% by both assays, were authentic.

Although *TP53* background mutations were common, their MAF was always below 1% (**Fig. 1E-F**), which could be used as a threshold to optimally identify patients with ovarian cancers from patients cancer-free. In this, albeit small, pilot study, a 1% threshold yielded a sensitivity of 70% and specificity of 100%, which outperformed other published tests for ovarian cancer detection (*8*). Furthermore, in the lavages where the tumor mutation was identified, its frequency was at least 10-fold above the highest background mutation in that individual.

TP53 mutations in uterine lavage increase with age

To better understand the basis of *TP53* background mutations we examined the association of their abundance with age. When patients were ordered by ascending age (**Fig. 1E-F**), it appeared that older patients carried more mutations. However, the number of mutations found depends on total number of nucleotides sequenced (**Fig. S2**), which was variable across samples and tended to be higher in controls due to increased sequencing depths (**Table S2**). To compensate for this

variation, for each sample we calculated the total *TP53* mutation frequency by dividing the number of mutations identified in uterine lavage (including exons and flanking intronic sequences) by the total number of Duplex bases sequenced. For patients with cancer, we excluded the tumor mutation from this calculation in order to fairly reflect only *TP53* background mutations. For patients with ovarian cancer, as well as cancer-free control patients, the *TP53* mutation frequency significantly increased with age (**Fig. 2**, p=0.0006 for ovarian cancer, p=0.001 for controls, Spearman's correlation test). This trend was identical to prior observations of *TP53* background mutation frequency in peritoneal fluid and peripheral blood (*13*).

TP53 mutations in uterine lavage are not random, but rather are positively selected

The *TP53* gene is a tumor suppressor, the genetic disruption of which facilitates cell proliferation in tumors, even when only one allele is mutated (*20*). An age-associated increase in ultra-low frequency *TP53* background mutations could result from random, age-related mutagenic processes or, alternatively, from mutagenesis coupled with clonal selection of pathogenic variants. To distinguish between these possibilities, we performed a detailed analysis of traits of selection among the 112 age-associated *TP53* background mutations collectively identified among all 21 patients (**Table S3**, **Fig. 3**).

A metric of selection widely used in evolutionary biology, and recently incorporated into cancer genomics, is the dN/dS ratio (*21*). For a given coding region, this ratio compares the relative proportion of non-synonymous and synonymous mutations observed versus what would be expected based on random mutagenesis across all bases. A value of 1 indicates that the relative ratio observed is consistent with a random process and, in aggregate, coding changes in a region have neither a strong positive or negative impact on cell growth or survival. Values above and below 1, on the other hand, indicate positive and negative selection, respectively. The dN/dS ratio among background uterine lavage *TP53* mutations in cases and controls was 4.4 and 2.9,

respectively, indicating enrichment for non-synonymous mutations that confer a growth or survival advantage (**Fig. 3A**). The excess of non-synonymous mutations was not driven by a subset of outlier samples, but rather was uniformly observed across nearly all lavage samples (**Fig. S3A**).

We next examined metrics of selection related to the genic location of mutations. Background *TP53* mutations were not randomly distributed along the gene but clustered in certain regions of biological significance. First, nearly a quarter of *TP53* lavage background mutations occurred in the context of methylation-sensitive CpG dinucleotides, which is remarkable given the fact that these dinucleotides comprise less than 5% of the coding region of *TP53* (**Fig. 3B**, $p=5.8x10^{-10}$ for controls and $p=1.7x10^{-5}$ for ovarian cancer mutations, by Fisher's exact test). Mutations were also enriched in exons 5 to 8, which encode the DNA binding domain of the protein (**Fig. 3C**, $p=3.3x10^{-6}$ for controls and p=0.002 for ovarian cancer mutations, by Fisher's exact test). The most significant enrichment, however, was observed in *TP53* cancer-associated hotspot codons, which are the codons most recurrently observed mutated in cancer sequence databases. We considered the 9 most abundantly mutated codons in the UMD database (April 2017 version) (*20*). These codons encompass only 2.3% of the coding region of *TP53*, yet more than 25% of lavage background mutations clustered within these 27 base pairs (**Fig. 3D**, $p=5.1x10^{-17}$ for controls and $p=3.9x10^{-9}$ for ovarian cancer mutations, by Fisher's exact test), and among these, all were non-synonymous. The biases seen for each characteristic were not driven by a subset of outlier samples, but were distributed homogeneously across samples in both groups (**Fig. S3B-D**).

To assess the impact of these mutations on *TP53* protein function, we took advantage of Seshat, a recently developed online tool for *TP53* analysis that provides comprehensive mutational information including prediction of impact on protein activity as well as pathogenicity (*22*). We queried all background *TP53* mutations identified in the 21 lavages and color-coded them according to 5 binned categories of protein activity and predicted pathogenicity. Nearly all samples

carried at least one *TP53* mutation that inactivated the protein totally or partially (**Fig. 3E**) and/or was predicted to be pathogenic (**Fig. 3F**). The unambiguous signature across six distinct metrics of positive selection within the ultra-low frequency *TP53* mutations observed in all lavages, regardless of cancer status, indicate that these mutations expanded under strong selective pressure and are not the result of technical errors.

TP53 mutations in uterine lavage resemble mutations in cancer

We next compared the features of selection of *TP53* mutations identified in lavages to *TP53* mutations from cancers. For this analysis, we used all the cancer mutations present in the UMD cancer database (April 2017, n=71,051). We determined the percentage of these mutations that reside at CpG sites, cancer hotspots, and exons 5-8, as well as the percentage of mutations that impact protein activity or are predicted to be pathogenic. For each trait, we compared the distribution of mutations in the theoretical absence of selection, in our 21 uterine lavages, and in the cancer database (**Fig. 4A**). Remarkably, for all traits, *TP53* background mutations from uterine lavages far more strongly resembled *TP53* mutations in the cancer database than the pattern expected by random chance. We also used a feature of Seshat that categorizes *TP53* mutations according to their frequency in the UMD database. Nearly all uterine lavage samples harbored *TP53* mutations listed as frequent or very frequent in the database (**Fig. 4B**).

To further characterize background *TP53* mutations in uterine lavage in comparison to those in cancers, we compared mutation type and spectrum distribution as well as location along the gene. Non-cancer derived mutations in uterine lavages from women with and without cancer were predominantly missense, similar to mutations in the database (**Fig. 4C**), and displayed a mutational spectrum enriched in G>A and C>T transitions, comparable to cancer mutations (**Fig. 4D**). Most strikingly, the distribution of low-frequency *TP53* background mutations from just 21 women along the length of the gene is a mirror image of the distribution of *TP53* mutations from

more than 71,000 different tumors included in the database (**Fig. 4E**). Thus, somatic *TP53* mutations recovered from cells sloughed into the uterine cavity from normal healthy women are not random, but appear to emerge from an evolutionary process of mutation, selection and clonal expansion akin to what takes place in tumors, but within normal tissue.

TP53 mutations are common in healthy tissues from middle age women

These striking results prompted us to consider what the tissue origin of the mutation-bearing cells in the uterine lavages might be. To address this question, we sequenced *TP53* from DNA obtained from pre-operative uterine lavage and peripheral blood as well as multiple gynecological tissues collected postoperatively following total hysterectomy/bilateral salpingo-oophorectomy for symptomatic fibroids (benign leiomyomas) from two middle age women (**Fig. 5A, Table S4**). DNA was extracted and processed for DS as previously, except that samples were sequenced to a higher average depth (**Table S5**).

We identified *TP53* mutations in all samples from a 56-year-old woman and in all but two samples from a 46-year-old woman (**Fig. 5B, Table S6**). When we compared the mutation frequency across all samples, several interesting observations emerged. First, the uterine lavage from the 56 year old had a mutation frequency that appeared disproportionally high, both when compared to that of most other tissues and when compared with the lavage of the 46 year old. However, when compared to the mean values of uterine lavages from control women of similar ages (50-56 and 40-46 year old) from first part of the study, the frequencies by age were quite similar (**Fig. 5B**).

Moreover, the distribution of mutations according to each trait of positive selection (type, frequency in the cancer database, predicted activity and pathogenicity, exon clustering, CpG clustering, and enrichment for cancer hotspots) was comparable to the lavages previously analyzed (**Fig. S4**). Both lines of reasoning support the conclusion that the elevated frequency of mutations

in the uterine lavage of the 56 year old is not artefactual and confirm the previously observed age effect.

For other tissues, however, we did not observe an obvious increase of *TP53* mutation frequency between 46 and 56 years of age. There was substantial variability in the mutation content of different tissues and of different biopsies of the same tissue, which reflects either a stochastic effect or the imprecision of macrodissection for obtaining exactly comparable tissue samples (for example, the depth of endometrium harvested or how distal the tubal fimbriae were cut). No single tissue stood out as obviously more mutation prone than another, nor could any tissue be identified as a dominant source of the mutations found in lavages.

TP53 mutations increase in number and cancer-like features during normal human aging

With the hope of observing a stronger aging mutational signal, we looked to tissue samples from greater extremes of age. While the procurement of such material was challenging, we managed to obtain several gynecologic tissues at autopsy from a neonate who died from a congenital vascular malformation and from a 101 year old female who died of natural causes (**Table S4**). Together with the middle age samples, this unique specimen collection represents the full breadth of a century of the human lifespan.

Although the tissue types available were not fully identical across all four subjects, the pattern of *TP53* mutations, nevertheless, yielded unique insights. To help more intuitively visualize this multiparametric data, in **Fig. 6** we annotated all mutations found among the different tissues of the four subjects as color coded boxes for each feature of selection: red for “cancer-like”, blue for non-cancer-like. The number of columns of colored boxes per sample reflects the total number of mutations identified. When viewed in this format it is apparent that *TP53* mutations are not only more abundant with age but are also more “cancer-like”. Mutations found in older tissues are disproportionately observed in cancers and are predicted to inactivate the protein or be otherwise

pathogenic. In contrast, mutations found in the newborn are rarely found in cancer, tend to preserve the protein activity, and are not predicted to be pathogenic.

Different tissues and different biopsies within the same tissue showed substantial variability in both the number of mutations and their cancer-like features. In aggregate, fallopian tube epithelium appeared to be a "hot" tissue with high number of mutations and a high percentage of cancer-like mutations. However, in the 56 year old, one fallopian tube sample harbored only a single synonymous mutation, consistent with the notion of "hot" and "cold" zones within a tissue. This was similarly seen in the two distinct endometrial biopsies of the centenarian.

In addition to a larger number of clones, with aging we would also predict an increase in the size of clones, as would be reflected by a higher MAF of each variant found. However, in this study, some samples were sequenced at a lower depth, which may lead to outlier (low event count) biases in the calculation of MAFs, thus precluding a fair comparison between samples (**Fig. S5**). Despite this caveat, two large clones were clearly seen within the peripheral blood leukocytes of the 101-year-old (**Fig. S5** and **Table S6**, c.659A>G MAF: $1.2x10^{-2}$, and c.455C>T MAF: $4.5x10^{-3}$). Interestingly, the exact *TP53* mutation that defined each of these clones was also detected at lower frequencies in peritoneal and endometrial samples from the same subject, revealing an apparent contribution of leukocyte DNA to those tissue samples. (**Fig. S6**).

In fact, this cross-tissue mutation sharing was common in the 101-year-old woman, suggesting that aged leukocytes might indeed harbor relatively large clones that recurrently contribute to the mutations found when sequencing other biopsies. Mutation sharing was less prevalent in middle age women. While very low frequency mutations are often hard to replicate due to the low precision of the measurement resulting purely from sampling statistics (not technical accuracy), it is important to keep in mind that certain mutations might also be recurrently identified simply because they are hotspots, and thus common origin cannot automatically be assumed. For

example, the hotspot mutation c.659A>G was identified in the large blood clone of the 101 year old woman as well as in a myometrium sample and a fallopian tube biopsy of the 46 year woman (**Fig. S6**). The processing of these particular samples was done on different days, making a cross-contamination explanation improbable.

As already considered, an important limitation of this study was the different depth of sequencing achieved across samples, due to the inherent variability in library preparations as well as differences in DNA availability. Because numerically more mutations will be identified in samples with more sequencing (**Fig. S7A**), it is essential to compare samples based on their mutation frequency, which is a sequencing-normalized value calculated as the number of mutations divided by the number of total Duplex error-corrected nucleotides sequenced (**Fig. S7B**). *TP53* mutation frequency tended to be higher at older ages in the three tissue types shared by the neonate and the centenarian (leukocytes, peritoneum, and endometrium), although there was substantial variability across samples.

TP53 mutations in newborn tissue are random, yet become positively selected over a lifetime

As further illustration of the increase of cancer-like mutations with aging, we divided all *TP53* mutations into two binary categories: "common in cancer" and "not common in cancer", with the former being defined by those classified as "frequent" or "very frequent" in the UMD cancer database (**Table S7**). When plotted by age, the progressive enrichment for cancer-like mutations was easily apparent, especially in certain tissues such as fallopian tube and leukocytes (**Fig. 7A**).

We then examined the five traits of selection previously calculated for the uterine lavage study but for *TP53* mutations found in the newborn, middle age, and centenarian tissues (**Fig. 7B, Fig S8**). Remarkably, for mutations found in newborn, all five traits yielded values consistent with

random processes (e.g. absence of selection), yet in middle age, and even more so in centenarian tissue, values reflected selection to an extent that neared that seen in mutations from tumors in the UMD database. Analysis of mutation type (**Fig. 7C**) revealed a decrease of synonymous mutations with age (in fact, no synonymous mutations were identified in centenarian tissue, **Fig. S8, Table S6**).

Lastly, regarding mutation spectrum, we noted an interesting preponderance of C and G mutations in newborn tissue, which progressively shifted towards an increased representation of A and T mutations in centenarian tissue, more similar to the pattern in cancers. The significance of this shift is unknown as it could represent both biases in the nucleotide composition of the gene at selectable hotspots, as well as differential age-associated mutagenic processes (*23*), which disproportionately contribute to the clonal mutation burden of tumors because tumors mostly arise in the elderly.

TP53 mutations in cfDNA and peritoneal fluid follow the same patterns as solid tissue

To explore the abundance of *TP53* mutations to liquid biopsies of clinical interest, we analyzed plasma-derived cell-free DNA (cfDNA) and peritoneal fluid from the 46-year-old woman. *TP53* mutations were identified in both, with cancer-like features similar to what was observed for solid tissue biopsies and leukocytes (**Fig. S9**). None of the mutations identified in cfDNA or peritoneal fluid overlapped with mutations identified in leukocytes, uterine lavage or any of the solid tissues analyzed (**Fig. S6).** Peritoneal fluid is routinely collected for disease staging during gynecological surgery and we previously demonstrated that it carries *TP53* background mutations with cancer-like features (*13*). cfDNA had not been analyzed previously by DS. The fact that one of the identified mutations is pathogenic and commonly found in cancers (**Fig. S9, Table S6**) raises important concern over specificity in cancer-screening studies based purely on mutation detection in plasma.

Discussion

We have demonstrated that uterine lavage coupled with DS offers a promising solution for ovarian cancer detection based on a minimally invasive sampling approach that can practically be integrated into routine gynecologic primary care (*24*). The technique improves upon past mutation-based screening efforts through use of (1) a collection method that recovers cancer cells very close to the anatomical site of the tumor and (2) an ultra-accurate DNA sequencing technology that can resolve exceptionally low frequency mutations. In this small study, we were able to achieve remarkable sensitivity and specificity using a mutation allele frequency threshold for differentiating cancer cases from non-cancer controls. This was possible without the allele-specific PCR technique required in our past lavage study (*9*) which, from the perspective of a screening test, impractically necessitated prior knowledge of the specific tumor mutation being sought. While validation through substantially larger prospective trials will be critical to support widespread clinical use, we have established the technical foundation for such studies, which are now enrolling in both Europe and the US (LUDOC II - ClinicalTrials.gov Identifier: NCT02518256, LUSTIC - ClinicalTrials.gov Identifier: NCT02039388).

However, the most profound finding of this work is not the biomarker performance of the technique itself, but the incidentally found mutational patterns that reflect a somatic evolutionary process that appears operative throughout much of human life in normal tissues. Specifically, we identified widespread low frequency *TP53* mutations that were heavily enriched for pathogenic variants. This enrichment reflects a process of natural selection that favors the survival and proliferation of cells with mutations that are identical to those observed in cancer, but as part of routine aging. The unambiguous selection signature is supported by multiple orthogonal metrics

and cannot be explained by technical errors; both the biological and diagnostic implications are substantial.

One of the main reasons that cancer biomarkers fail to reach the clinic is their inability to achieve the extremely high specificity required for screening (*25*). This is critical for cancers with low incidence and that require an invasive procedure to follow up positive screening tests, such is the case of ovarian cancer (*3*). Harms due to false positives and a lack of proven reduction in mortality are the reasons for the recent recommendation against the use of CA-125 and transvaginal sonography for screening asymptomatic women (*4*). In recent years, DNA mutation-based cancer screening from plasma or other body fluids has emerged as a promising method to detect cancer based on the supposition that cancer-associated mutations found in liquid biopsies are a specific indication of cancer somewhere in the body (*26*). Here we demonstrate that with a sufficiently low technical background, biologically real cancer-associated mutations can be found in every tissue tested: cancer-associated mutations are, in fact, far from cancer-specific.

The detection of cancer-associated mutations in normal tissues is not entirely new (*14*). In 2014, a series of major publications reported that mutations commonly found in acute myeloid leukemia occur as minority subclones in the blood of ~10% of healthy elderly individuals–a phenomena dubbed Clonal Hematopoesis of Indeterminate Potential (CHIP) (*27-29*). A year later Martincorena and colleagues observed hundreds of tiny clones carrying cancer-associated driver mutations on sun-exposed eyelids (*30*), a finding recently replicated in normal aged esophagus (*31*). Cancer-associated mutations have been similarly reported in abnormal, but non-cancerous tissues including endometriosis (*32, 33*) and benign dermal nevi (*34*). The use of laser capture microdissection in recent studies has revealed that as many as 1% of normal colorectal crypts of middle-age individuals (*35*) and >50% of normal endometrial glands of middle-age women (*33*) carry mutations in cancer driver genes.

The relationship between oncogenic mutations and cancer has been known for decades, yet the delay in appreciating their presence outside of cancer can be largely attributed to technical limitations. The advent of NGS enabled surveying wide swaths of the genome and detection of mutations present clonally or as modest size subclones. In the above studies, standard whole exome or multi-gene NGS were able to identify driver mutations because of unique scenarios where clones were either relatively large (CHIP in a subset of very elderly) (*27, 28*) or spatially coherent and comprising a sizeable percentage of cells when very small biopsies were taken (*30, 33, 35*). With higher accuracy NGS techniques able to resolve lower frequency subclones, later studies have found CHIP mutations at lower levels in most middle-aged adults (*36, 37*). Using ultra-high accuracy DS, we found extremely low frequency cancer-associated *TP53* mutations in both blood and peritoneal fluid of healthy women undergoing abdominal surgery and, in both sample types, the abundance of mutations increased with age (*13*). A subsequent study that employed uterine lavage for endometrial cancer detection found pathogenic mutations in cancer driver genes in lavages of cases as well as controls (*38*).

The present study adds a new layer of knowledge by identifying extremely low frequency *TP53* mutations in uterine lavages and multiple normal tissues. DS has an error rate below one-in-ten-million, so even mutations that are seen in a single nucleotide among hundreds of millions of other sequenced nucleotides could be confidently identified. In aggregate, we sequenced 319,576,913 unique *TP53* coding nucleotides and identified 292 non-cancer derived "biological background" mutations. The average background mutation frequency of $9x10^{-7}$ in this study is at least ten thousand times below the background error threshold of standard NGS and dozens of times below that of other error correction methods (*10*) (**Fig. S1**).

We examined 10 different tissue/sample types from a unique cohort of individuals spanning more than a century of the human lifespan and assessed the pattern of mutations found

using multiple different metrics of selection. A significant and novel finding was that, not only do *TP53* mutations increase in abundance with age, but the relative representation of random mutations versus cancer-associated mutations transitioned from almost entirely the former to almost entirely the latter from birth to end-of-life. Moreover, the extent of mutation frequency increase varied considerably by sample type and tissue. Uterine lavage samples carried nearly an order of magnitude more mutations than any other tissues examined by a woman's mid 50's. The basis of the rapid increase from women in their mid 40's to those in their mid 50's could plausibly relate to the onset of menopause. The majority of cells collected in uterine lavage are of endometrial origin and the cessation of cyclic sloughing at menopause might eliminate a physiologic means of purging mutated cells. Of note, none of the 10 *TP53* coding mutations identified in the endometrium of the 56-year old woman was detected in the uterine lavage, which contained as many as 25 mutations. This could be explained by the fact that the endometrial biopsies analyzed contained not only superficial surface epithelium, which are the cells collected by the lavage, but deep tissue as well, which might have a lower mutational load.

The biology seen here is likely only the tip of the iceberg and much further work remains to be done. We only examined tissue sets from four individuals at a very coarse spacing along the aging continuum. We focused on only gynecologic tissues and did not have perfectly matching samples for each subject, given how challenging these unique specimens were to obtain at the extremes of age. Furthermore, we only sequenced the coding region of a single gene, albeit the one most commonly mutated in cancer.

The implications of our findings are important as related to the physiology of aging, but also as a cautionary message for mutation-based cancer biomarkers of all varieties. At the same time as we have shown that high sensitive NGS methods are essential for maximally sensitive mutation-based cancer diagnosis, we have also illustrated a substantial specificity challenge related

to biology, not technology, the extent of which has been under-appreciated. This is not limited to one or a few tissues, rather, it seems to be ubiquitous among the epithelial, mesenchymal, and hematopoietic cells lineages we investigated. Moreover, the same selection patters of mutations were found in minimally invasively collected body fluids, including both uterine lavage and cfDNA. This suggests that ongoing large scale efforts to develop universal "liquid biopsy" cancer screening tests via deep sequencing of cfDNA from plasma need to be approached with great caution (*26, 39*).

Despite the extent of biological background mutations, as a non-invasive cancer test our approach worked remarkably well. We identified 80% of tumor mutations and 70% of those were above the 1% MAF threshold we used to distinguish cases from controls. The only tumor mutation missed below this threshold corresponded to a 42-year-old woman, one of the youngest in the study. Younger women tended to have fewer background mutations and mutations with lower MAF, which suggest that, moving forward, sensitivity could be increased by using age-adjusted thresholds. In addition, specificity could be improved by: uniform lavage collection at the luteal phase in premenopausal women, sequencing of peripheral blood to identify and exclude CHIP clones that might be present in lavage and longitudinal assessment of mutations to identify MAF increases over time. An important consideration is that we have demonstrated detection of intermediate and late stage cancers, but the most critical target for screening is early stage cancers because they are most curable. In that regard, monitoring of high-risk populations, such as *BRCA1* and *BRCA2* carriers, may be the highest impact near-term clinical implementation.

The sensitivity improvements lent by new sequencing technologies are forcing a far more nuanced genetic definition of what distinguishes a cancer cell from simply an old cell. Our results show that CHIP clones are merely one relatively easy-to-detect manifestation of a far broader phenomena that appears to extend to most, if not all, tissues in the body. From a biomarker

perspective, the fact that those who are at greatest risk of cancer and for whom cancer screening holds the most benefit (older adults) are also the population with the most cancer-like age-associated background mutations, is particularly inconvenient. Ongoing improvements will be needed to find ways to maximize specificity through careful MAF threshold calibration and combination with orthogonal biomarkers. Further investigation into the significance of biological background mutations from the perspective of human aging and biology is similarly warranted. For example, does the frequency of mutations and extent of pathologic shift provide a chronologically independent empiric measure of age? Could it be used to integrate a lifetime load of intrinsic and extrinsic mutagenic exposures and predict future cancer risk?

While the notion that our somatic genomes are steadily evolving towards neoplasia with each passing decade might viewed as disheartening, an alternative perspective is that, in spite of this, most people do *not* develop overt cancer in their lifetime. This serves as a reminder of just how much remains unknown about the body's many complex mechanisms of tumor suppression– a toolkit that we can perhaps augment with future technologies for cancer prevention.

Material and Methods

Experimental Design

We performed two complementary studies. The objective of the uterine lavage study was to determine the ability of DS to detect ovarian cancer through deep sequencing of *TP53* mutations in uterine lavage. This study included 10 patients with high-grade serous ovarian cancer (cases) and 11 with benign gynecological masses (controls). The objective of the normal tissue study was to characterize somatic *TP53* mutations that accumulate during aging. This study included tissue from two newborn subjects (one newborn male only provided blood), two middle age women (ages

46 and 56 years old) and one centenarian woman (101 years old). Clinico-pathological information for all subjects is listed in **Table S1**.

Samples

In the first study, we analyzed uterine lavages collected by a trans-cervical catheter (**Table S1**). Lavages were collected immediately pre-operatively as previously described (*9*). Lavage samples were centrifuged at 300x g for 10 minutes at room temperature and DNA was isolated from the cell pellet (QIAamp MinElute Kit, Qiagen, Hilden, Germany). Patients were recruited in three institutions: Medical University of Vienna (Austria), Charles University Pilsen (Czech Republic) and University Hospitals Leuven (Belgium). Sample procurement was performed in accordance with the institutional review boards of the Medical University of Vienna (EK#1148/2011 and EK#1766/2013), the Catholic University Leuven (B322201214864/S54406) and the Medical Faculty Hospital Pilsen (No 502/2013).

In the second study, multiple gynecological tissues were collected per **Table S4**. Not all sample types were available for all subjects. Newborn and centenarian tissue was collected at autopsy while tissue from middle age women was collected following hysterectomy. Peripheral blood was unavailable from the female newborn and to compensate we sequenced spleen from this subject as well as peripheral blood sample from a neonatal male (7 weeks old). The two newborn autopsies were performed at Seattle Children's' Hospital and tissues were collected under research IRB #52304. For the two middle age women, uterine lavage was collected with the same procedure as in the first study. In addition, for the 46-year-old woman, cfDNA was collected preoperatively and peritoneal lavage collected intraoperatively. Both operations were performed at the Medical University of Vienna and samples were collected with informed consent and according to approved IRB EK# 1152/2014. For the centenarian case, tissue was obtained via rapid autopsy from Tissue for Research Inc. and processed at the University of Washington under IRB waiver 2016-52304.

All samples were collected using sterile new instruments between biopsies and frozen over liquid nitrogen immediately after collection and stored at -80˚C until DNA extraction.

Digital Droplet Polymerase Chain Reaction

Lavage DNA from 5 ovarian cancer cases and 2 controls was analyzed by ddPCR (**Table 1**). In ovarian cancer lavage, ddPCR amplified the tumor mutation whereas in benign lavage, the assay targeted two mutations previously identified by DS at frequencies below 0.1%. ddPCR was performed with the QX100 Droplet Digital PCR system (Bio-Rad Laboratories, Hercules, CA) using custom TaqMan SNP Genotyping Assays (Life Technologies, Carlsbad, CA) designed using Primer Express 3.0 software (ThermoFisher). 10-20ng of DNA were used in each reaction and samples were analyzed at least in duplicates. A positive control and a wild-type control were included in every run.

Duplex Sequencing

Duplex Sequencing was performed as previously described with minor modifications (*12*). Briefly, DNA was sonicated, end-repaired, A-tailed, and ligated with DS adapters using the KAPA HyperPrep library kit (Roche Sequencing, Pleasanton, CA). After initial amplification, 120 bp biotinylated oligonucleotide probes (Integrated DNA Technologies, Coralville, Iowa) were used to capture the coding region of *TP53*. Two successive rounds of captures were performed to ensure sufficient target enrichment, as previously described (*40*). Indexed libraries were pooled and sequenced on an Illumina HiSeq2500 or NextSeq 550. Sequencing reads were aligned to hg19 then reads sharing a common molecular tag in both distinct strand orientations were grouped and assembled into an error-corrected Duplex Consensus Sequence as previously described (*12*).

The total number of Duplex nucleotides sequenced for each uterine lavage and tissue sample is listed in **Tables S2** and **S5**. In aggregate, we sequenced 587,169,708 unique nucleotides,

319,576,913 of which corresponded to coding nucleotides. We targeted a median Duplex depth of ~1000x. Three tissue biopsies were excluded because of insufficient depth. Because Duplex reads correspond to original DNA molecules, Duplex depth indicates the total number of haploid genomes sequenced. For each sample, *TP53* mutation frequency was calculated as the number of identified mutations divided by the total number of Duplex nucleotides sequenced. For each individual mutation, mutant allele frequency (MAF) was calculated as number of mutated Duplex bases divided by the total Duplex depth at a given nucleotide position. Mutations identified as SNPs in the 1000 genome database with were excluded from mutation analysis. All mutations were manually reviewed with the Integrative Genome Viewer (IGV).

Characterization of TP53 mutations using Seshat and the UMD TP53 database

The final list of mutations from all samples in the study (uterine lavage study n=166, normal tissue study n=264) was converted into a Variant Call Format (VCF) file and submitted to Seshat (https://p53.fr/*TP53*-database/seshat), a web service that performs *TP53* mutation annotation using data derived from the UMD *TP53* database (*22*). This database is the most updated and comprehensive repository of *TP53* variants. From the Seshat output (included as **Database 1** for uterine lavage mutations and **Database 2** for normal tissue mutations), the following variables were extracted: cDNA variant, HG19 Variant, Variant Classification, Frequency, Activity, Pathogenicity, Exon, Codon, CpG, Mutational event and Variant Comment. These variables were used to annotate the DS pipeline-generated mutational calls in **Table S3** (uterine lavage) and **Table S6** (normal tissue). The human genomic reference hg19 (GRCH37) was used for data reporting. Mutations occurring in the coding region and adjacent splice sites were selected for mutational analysis (uterine lavage n=112, normal tissue n=180).

Mutations were annotated based on type (missense, nonsense, splice, synonymous), mutation spectrum (each of the 12 possible nucleotide substitutions), localization to CpG

dinucleotides, localization in exons 5 to 8 (encoding the protein's DNA binding domain), localization to mutational hotspot (9 most common mutated codons in the UMD *TP53* database: 175, 179, 213, 220, 245, 248, 249, 273, 282), frequency of the mutation in the cancer database, functional activity, and predicted pathogenicity. Functional activity was assessed by a transcriptional activity chart assay for 3,000 variants performed by Kato et al. (*41*). Pathogenicity was based on multiple predictive algorithms included from dbNSFP (*42*) as well as functional activity (*22*). For the last 3 variables (frequency in cancer database, activity and pathogenicity), mutations were aggregated into binned groups (**Table S7**).

TP53 cancer database mutational analysis

From the UMD *TP53* database (April 2017 version), we selected the set of 71,051 mutations reported within human tumors of all types (mutations from cell lines, normal and premalignant tissue were excluded). Then we determined the distribution of mutations in the following categories: CpG, hotspots, exons 5-8, activity, pathogenicity, mutation type, and mutation spectrum. These values were used as a comparator for *TP53* mutations identified in uterine lavage and normal tissues.

TP53 mutations without selection

To assess the distribution of *TP53* mutations in the absence of selection, we generated a list of all possible mutations in the gene coding region (n=3,546) *in silico*. Then we submitted this list to Seshat to determine the distribution of mutations in the same categories as above. The values obtained represent the distribution of all possible *TP53* mutations in the absence of selection and were used as a comparator for *TP53* mutations identified in uterine lavage and normal tissue.

dN/dS calculation

The dN/dS ratio measures the ratio of non-synonymous vs. synonymous mutations taking into consideration observed and expected values for a given genomic region. It was calculated as (n/s)/(N/S), where n is the total number of observed nonsynonymous mutations, s is the total number of observed synonymous mutations, N is the total number of possible nonsynonymous mutations, and S is the total number of possible synonymous mutations. For *TP53*, N=2,715 and S=831 based on the 3,546 possible mutations in its coding region. A dN/dS ratio equal to 1 indicates neutral selection, <1 indicates negative selection, and >1 indicates positive selection.

Statistical analyses

Correlations were tested with Spearman's rank test due to high variability in the outcomes (nonnormality). The comparison of *TP53* mutational traits in uterine lavage of controls and cancers *vs.* non-selected mutations was performed with Fisher's exact tests. All tests were two-sided at an alpha level (type 1 error rate) of 0.05. Statistical analyses were performed with SPSS and R.

Acknowledgments: We thank members of the Loeb, Risques, Kennedy, and Swisher labs at the University of Washington for helpful discussions as well as participating members of the LUSTIC and LUDOC clinical trials. Most importantly we deeply thank the patients and families who volunteered to provide clinical samples, without which this research would not have been possible.

Funding: NIH grant R01CA181308 to RAR; Mary Kay Foundation grant 045-15 to RAR; Rivkin Center for Ovarian Cancer grant 567612 to RAR; T32CA009515 for JJS; R44CA221426 to JJS and LNW; CA077852 and CA193649 to LAL; Radiumhemmets Forskningsfonder 174261 to TS.

Author contributions: JJS, RZ, PS, EM and RAR designed the study; EM, EL, RH and AV procured the samples; JJS, KLS, EM, and LNW processed the samples; JJS, CCV, DN, KB, TS and RAR contributed to data analysis and visualization; ME and RAR performed statistical analyses; LAL, RZ, and PS contributed expertise and invaluable critical discussion; JJS and RAR wrote the article.

Competing interests: JJS and LAL are founders and equity holders at TwinStrand Biosciences Inc. JJS, CCV, LNW are employees and equity holders at TwinStrand Biosciences Inc. PS is a founder and equity holder in Ovartec Inc. RZ is a founder and equity holder in Oncolab GmbH. RAR shares equity in NanoString Technologies Inc. and is the principal investigator on an NIH SBIR subcontract research agreement with TwinStrand Biosciences Inc.

Data and materials availability: Sequencing data that supports the findings of this study have been deposited in the Sequence Read Archive (BioProject ID: PRJNA503496). Software for DS data analysis is available at https://github.com/risqueslab.

References

1. F. Bray *et al.*, Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin*, (2018).
2. R. L. Siegel, K. D. Miller, A. Jemal, Cancer Statistics, 2017. *CA Cancer J Clin* **67**, 7-30 (2017).
3. C. W. Drescher, G. L. Anderson, The Yet Unrealized Promise of Ovarian Cancer Screening. *JAMA oncology* **4**, 456-457 (2018).
4. USPSTF *et al.*, Screening for Ovarian Cancer: US Preventive Services Task Force Recommendation Statement. *JAMA* **319**, 588-594 (2018).
5. J. T. Henderson, E. M. Webber, G. F. Sawaya, Screening for Ovarian Cancer: Updated Evidence Report and Systematic Review for the US Preventive Services Task Force. *Jama* **319**, 595-606 (2018).
6. L. A. Diaz, Jr., A. Bardelli, Liquid biopsies: genotyping circulating tumor DNA. *J Clin Oncol* **32**, 579-586 (2014).
7. I. Kinde *et al.*, Evaluation of DNA from the Papanicolaou test to detect ovarian and endometrial cancers. *Sci Transl Med* **5**, 167ra164 (2013).
8. Y. Wang *et al.*, Evaluation of liquid from the Papanicolaou test and other liquid biopsies for the detection of endometrial and ovarian cancers. *Sci Transl Med* **10**, (2018).
9. E. Maritschnegg *et al.*, Lavage of the Uterine Cavity for Molecular Detection of Mullerian Duct Carcinomas: A Proof-of-Concept Study. *J Clin Oncol*, (2015).
10. J. J. Salk, M. W. Schmitt, L. A. Loeb, Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat Rev Genet* **19**, 269-285 (2018).
11. M. W. Schmitt *et al.*, Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci U S A* **109**, 14508-14513 (2012).
12. S. R. Kennedy *et al.*, Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9**, 2586-2606 (2014).
13. J. D. Krimmel *et al.*, Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. *Proc Natl Acad Sci U S A* **113**, 6005-6010 (2016).
14. R. A. Risques, S. R. Kennedy, Aging and the rise of somatic cancer-associated mutations in normal tissues. *PLoS Genet* **14**, e1007108 (2018).
15. D. D. Bowtell *et al.*, Rethinking ovarian cancer II: reducing mortality from high-grade serous ovarian cancer. *Nat Rev Cancer* **15**, 668-679 (2015).
16. TCGA, Integrated genomic analyses of ovarian carcinoma. *Nature* **474**, 609-615 (2011).
17. R. Vang *et al.*, Molecular Alterations of TP53 are a Defining Feature of Ovarian High-Grade Serous Carcinoma: A Rereview of Cases Lacking TP53 Mutations in The Cancer Genome Atlas Ovarian Study. *Int J Gynecol Pathol* **35**, 48-55 (2016).
18. J. Chien *et al.*, TP53 mutations, tetraploidy and homologous recombination repair defects in early stage high-grade serous ovarian cancer. *Nucleic Acids Res* **43**, 6945-6958 (2015).
19. E. Kuhn *et al.*, TP53 mutations in serous tubal intraepithelial carcinoma and concurrent pelvic high-grade serous carcinoma--evidence supporting the clonal relationship of the two lesions. *J Pathol* **226**, 421-426 (2012).
20. B. Leroy, M. Anderson, T. Soussi, TP53 mutations in human cancer: database reassessment and prospects for the next decade. *Human mutation* **35**, 672-688 (2014).
21. M. Nei, T. Gojobori, Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol Biol Evol* **3**, 418-426 (1986).

22. T. Tikkanen *et al.*, Seshat: A Web service for accurate annotation, validation, and analysis of TP53 variants generated by conventional and next-generation sequencing. *Human mutation*, (2018).
23. L. B. Alexandrov *et al.*, Clock-like mutational processes in human somatic cells. *Nat Genet* **47**, 1402-1407 (2015).
24. E. H. Maritschnegg, F; Pecha, N; Bouda, J; Trillsch, F;, C. V. Grimm, A; Agreiter, C; Harter, P;, E. V. Obermayr, I; Zeillinger, R; Speiser, P, Uterine and Tubal Lavage for Earlier Cancer Detection Using an Innovative Catheter. *Int J Gynecol Cancer* **In press**, (2018).
25. E. P. Diamandis, The failure of protein cancer biomarkers to reach the clinic: why, and what can be done to address the problem? *BMC Med* **10**, 87 (2012).
26. A. M. Aravanis, M. Lee, R. D. Klausner, Next-Generation Sequencing of Circulating Tumor DNA for Early Cancer Detection. *Cell* **168**, 571-574 (2017).
27. G. Genovese *et al.*, Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N Engl J Med* **371**, 2477-2487 (2014).
28. S. Jaiswal *et al.*, Age-related clonal hematopoiesis associated with adverse outcomes. *N Engl J Med* **371**, 2488-2498 (2014).
29. M. Xie *et al.*, Age-related mutations associated with clonal hematopoietic expansion and malignancies. *Nat Med* **20**, 1472-1478 (2014).
30. I. Martincorena *et al.*, Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880-886 (2015).
31. I. Martincorena *et al.*, Somatic mutant clones colonize the human esophagus with age. *Science*, (2018).
32. M. S. Anglesio *et al.*, Cancer-Associated Mutations in Endometriosis without Cancer. *N Engl J Med* **376**, 1835-1848 (2017).
33. K. Suda *et al.*, Clonal Expansion and Diversification of Cancer-Associated Mutations in Endometriosis and Normal Endometrium. *Cell Rep* **24**, 1777-1789 (2018).
34. A. H. Shain *et al.*, The Genetic Evolution of Melanoma from Precursor Lesions. *N Engl J Med* **373**, 1926-1936 (2015).
35. H. Lee-Six *et al.*, The landscape of somatic mutation in normal colorectal epithelial cells. *bioRxiv*, (2018).
36. R. Acuna-Hidalgo *et al.*, Ultra-sensitive Sequencing Identifies High Prevalence of Clonal Hematopoiesis-Associated Mutations throughout Adult Life. *Am J Hum Genet* **101**, 50-64 (2017).
37. A. L. Young, G. A. Challen, B. M. Birmann, T. E. Druley, Clonal haematopoiesis harbouring AML-associated mutations is ubiquitous in healthy adults. *Nat Commun* **7**, 12484 (2016).
38. N. Nair *et al.*, Genomic Analysis of Uterine Lavage Fluid Detects Early Endometrial Cancers and Reveals a Prevalent Landscape of Driver Mutations in Women without Histopathologic Evidence of Cancer: A Prospective Cross-Sectional Study. *PLoS Med* **13**, e1002206 (2016).
39. C. Alix-Panabieres, K. Pantel, Clinical Applications of Circulating Tumor Cells and Circulating Tumor DNA as Liquid Biopsy. *Cancer discovery* **6**, 479-491 (2016).
40. M. W. Schmitt *et al.*, Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat Methods* **12**, 423-425 (2015).
41. S. Kato *et al.*, Understanding the function-structure and function-mutation relationships of p53 tumor suppressor protein by high-resolution missense mutation analysis. *Proc Natl Acad Sci U S A* **100**, 8424-8429 (2003).

42. X. Liu, C. Wu, C. Li, E. Boerwinkle, dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human mutation* **37**, 235-241 (2016).

Tables

Table 1. Comparison of *TP53* mutant allele frequencies by standard NGS, Duplex Sequencing, and digital droplet PCR

						Mutant Allele Frequency of tumor *TP53* mutations		
Patient ID	**Age**	**Histology**	**FIGO**	**Tumor mutation-DNA**	**Tumor mutation-protein**	**Standard NGS MAF**	**Duplex Sequencing MAF**	**ddPCR MAF**
case 1	67	HGSOC	IIIC	c.578A>C	p.H193P	15.25%	15.63%	14.45%
case 2	69	HGSOC	IIIB	c.646G>A	p.V216M	2.22%	2.12%	n.a.
case 3	51	Signet ring†	na	c.844C>T	p.R282W	22.76%	22.22%	25.55%
case 4	54	HGSOC	IV	c.785G>T	p.G262V	5.05%	4.88%	n.a.
case 5	32	HGSOC	IIIC	c.743G>A	p.R248Q	n.a.	0.98%	0.07%
case 6	42	HGSOC	IIIA	c.503A>G	p.H168R	Negative	0.15%	0.12%
case 7	55	HGSOC	IIIA	c.815A>T	p.V272E	Negative	Negative	n.a.
case 8	48	HGSOC	IIIC	c.734G>T	p.G245V	Negative	2.26%	0.20%
case 9	67	HGSOC	IV	c.1024C>T	p.R342*	Negative	Negative	n.a.
case 10	68	HGSOC	IIIC	c.535C>T	p.H179Y	n.a.	9.95%	n.a.
						Mutant Allele Frequency of background *TP53* mutations		
Patient ID	**Age**	**Histology**	**FIGO**	**Non cancer mutation-DNA**	**Non cancer mutation-protein**	**Standard NGS MAF**	**Duplex Sequencing MAF**	**ddPCR MAF**
con 2	70	benign	-	c.733G>A	p.G245S	n.a.	0.08%	0.08%
con 7	79	benign	-	c.817C>T	p.R273C	n.a.	0.05%	0.07%

NGS: Next Generation Sequencing; MAF: Mutant Allelle Frequency; ddPCR, digital droplet PCR; n.a: not assessed; † Spread to ovary
TP53 cDNA reference: NM_000546.5, TP53 protein reference: NP_000537.3

Figures and Legends

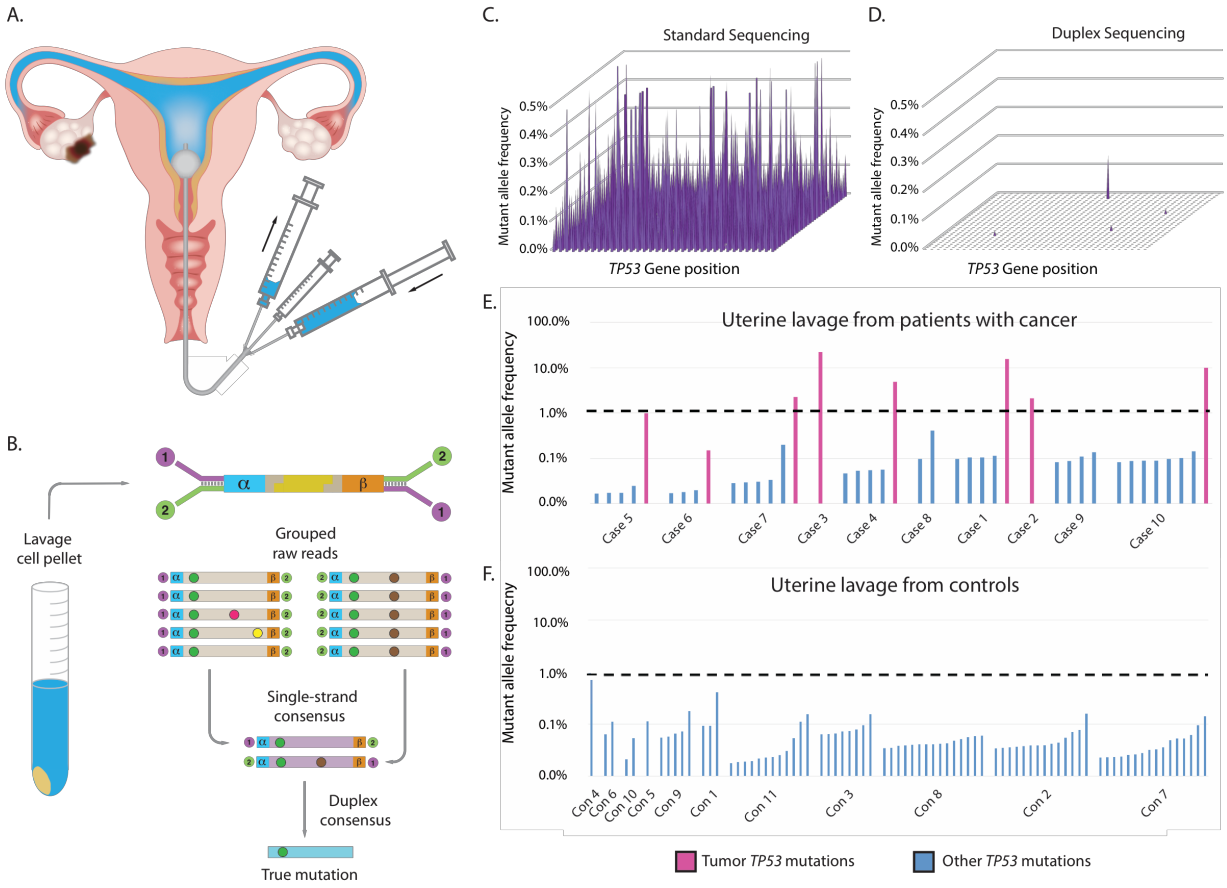


Figure 1. Detection of ovarian cancer using uterine lavage plus Duplex Sequencing (A) Uterine lavage is carried out by passing a small catheter through the cervix followed by concurrent flushing and aspiration with 10 mL of saline as described (*24*). (B) After cell isolation by lavage centrifugation, DNA is extracted, fragmented, and ligated with specialized Duplex Sequencing adapters that include degenerate molecular tags (α and β). Following amplification, hybrid capture and sequencing, reads sharing the same barcodes are grouped into families and mutations are scored only if present in both strands of each original DNA molecule. (C) Each spot on the 2 dimensional surface represents one of the 1179 coding positions in *TP53*. The height of each peak indicates the mutant allele frequency (MAF) at each position as determined by conventional NGS, which shows false mutations at every position. (D) DS of the same sample (case 6 below) eliminates errors and reveals only true mutations. (E) *TP53* mutations identified by DS in uterine lavage from women with ovarian cancer and (F) cancer-free (controls). Fuchsia bars represent the matching tumor mutation and blue bars represent 'biological background' mutations. Mutations are sorted by ascending MAF within each patient and patients are sorted by age. Dashed lines indicate the optimal threshold to distinguish patients with and without ovarian cancer (sensitivity: 70%, specificity: 100%).

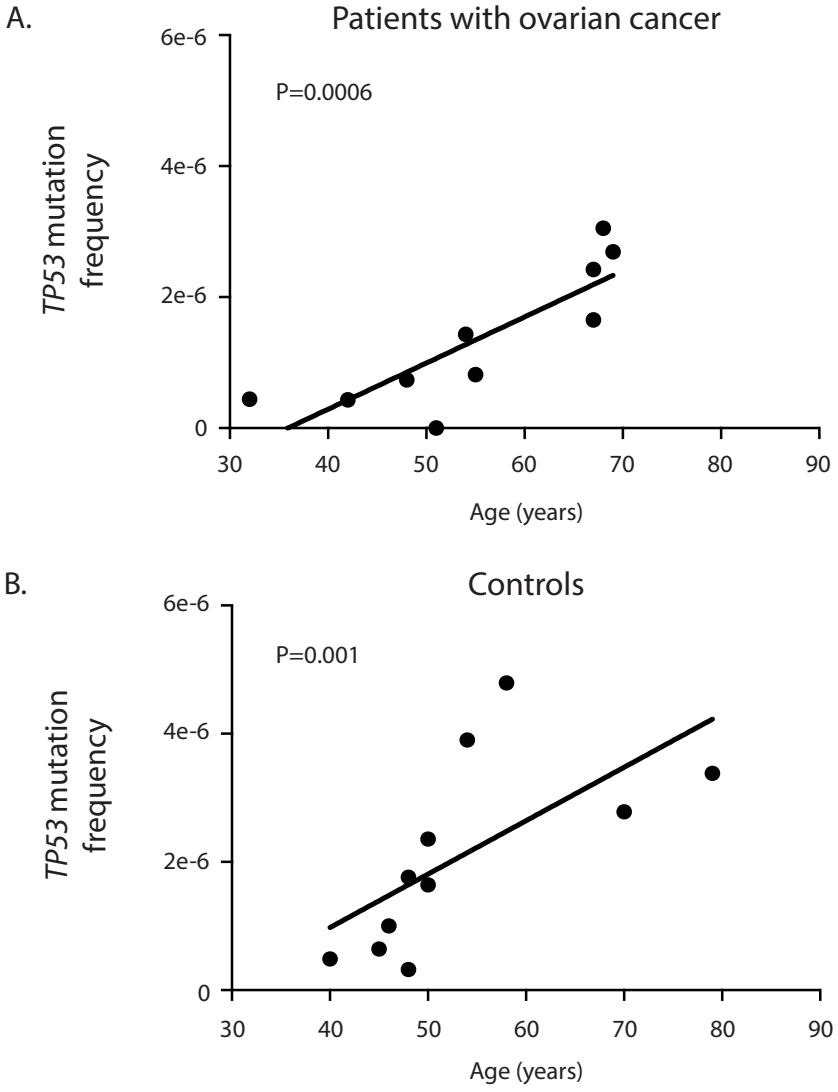


Figure 2. The frequency of *TP53* mutations in uterine lavage increases with age Frequency is calculated as the total number of unique *TP53* mutations identified in each sample (including exons and flanking intronic regions) divided by the total number of Duplex nucleotides sequenced. (A) Uterine lavage samples from patients with ovarian cancer, n=10, r=0.89, p=0.0006 by Spearman's correlation test. In these patients, the total count of *TP53* mutations excludes the tumor mutation identified in the lavage in order to only represent *TP53* background mutations. (B) Uterine lavage samples from control patients without cancer, n=11, r=0.83, p=0.001 by Spearman's correlation test.

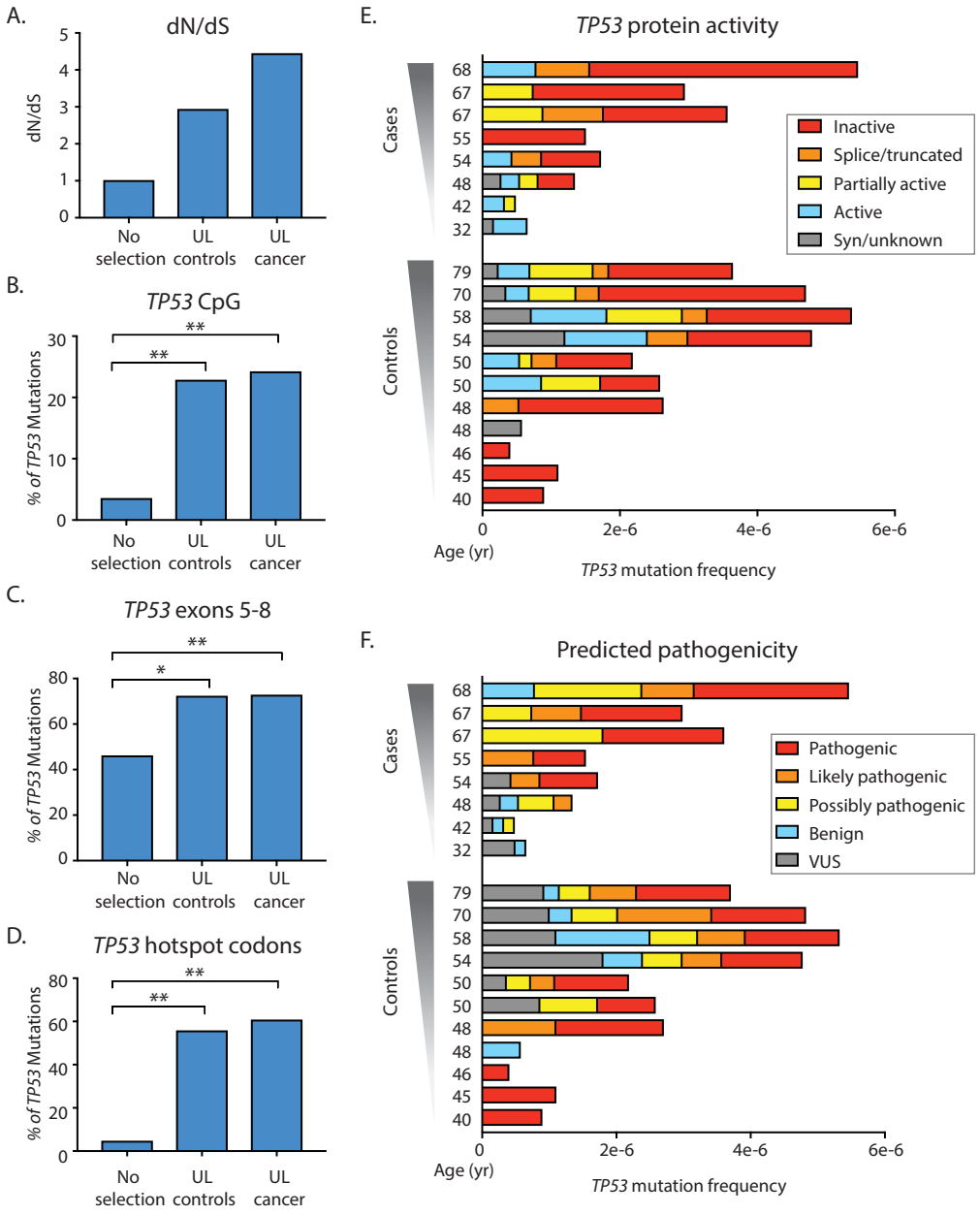


Figure 3. Evidence of positive selection in *TP53* background mutations from uterine lavages (A) dN/dS. Values above 1 in uterine lavage from controls and cases correspond to an excess of non-synonymous mutations suggestive of positive selection. (B) Percent of *TP53* mutations localized in CpG dinucleotides. (C) Percent of *TP53* mutations localized in exons 5-8. (D) Percent of *TP53* mutations localized in hotspot codons. For C-D: *TP53* mutations identified in uterine lavage from controls and cancer significantly exceed expected values without selection. * p-value<0.01 ** p-value<0.0001 by Fisher's exact test, n=79 for uterine lavage controls and n=33 for uterine lavage cancer. (E) Protein activity and (F) predicted pathogenicity color-coded as 5 groups from Seshat data. Patients are sorted by ascending age. For each patient, *TP53* Mutation Frequency is calculated as the number of mutations in the coding region divided by the total number of Duplex nucleotides sequenced in that region. Two cancer patients with reduced sequencing depth and no identified *TP53* mutations are excluded from the analysis. Nearly all cases and controls carry mutations that have impact in protein activity and predicted pathogenicity. UL: Uterine lavage.

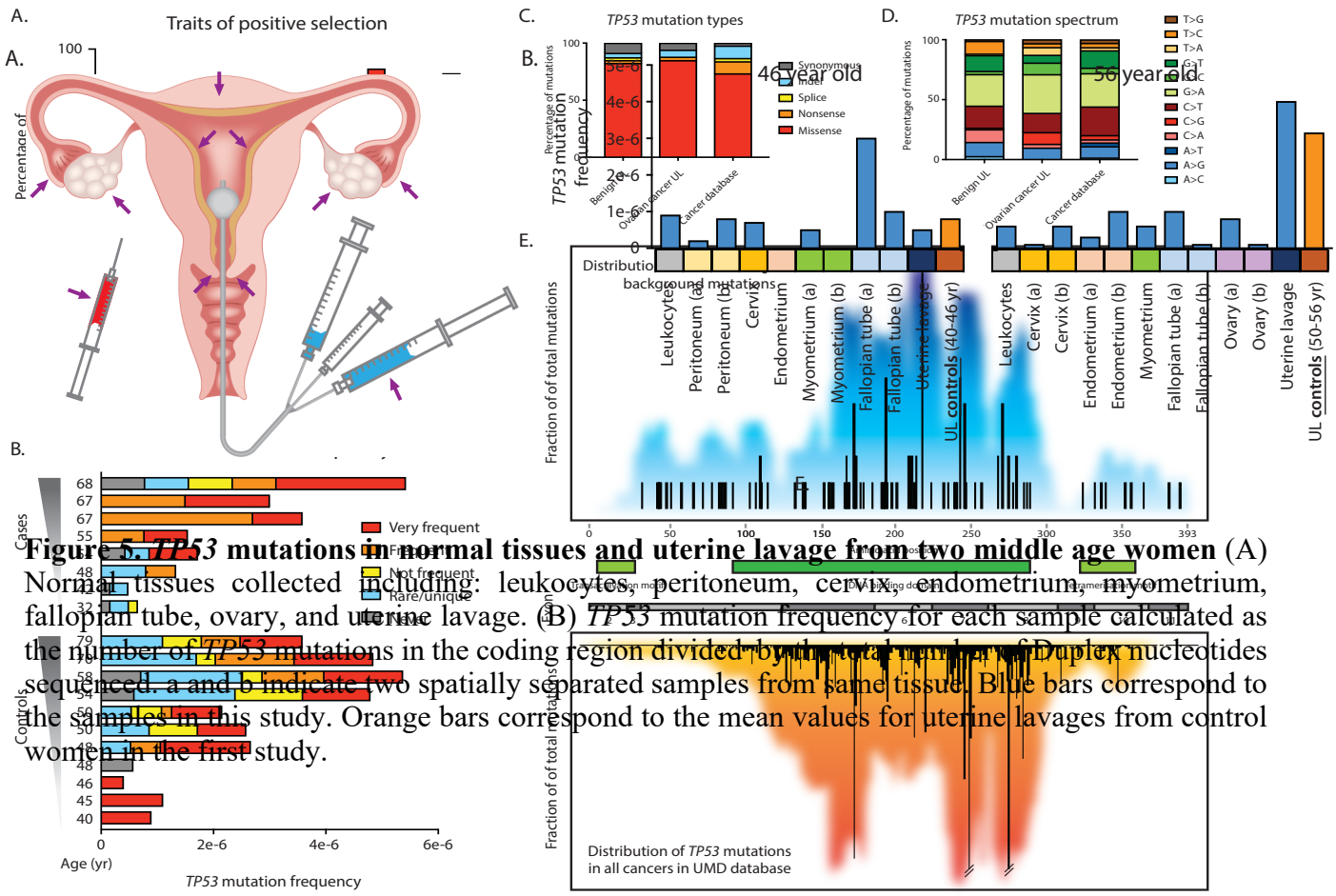


Figure 5. *TP53* mutations in normal tissues and uterine lavage from two middle age women (A) Normal tissues collected including: leukocytes, peritoneum, cervix, endometrium, myometrium, fallopian tube, ovary, and uterine lavage. (B) *TP53* mutation frequency for each sample calculated as the number of *TP53* mutations in the coding region divided by the total number of Duplex nucleotides sequenced. a and b indicate two spatially separated samples from same tissue. Blue bars correspond to the samples in this study. Orange bars correspond to the mean values for uterine lavages from control women in the first study.

Figure 4. *TP53* mutations in uterine lavage are very similar to *TP53* mutations found in human cancers (A) Traits of positive selection are compared between all possible mutations in the *TP53* coding region (No Selection, n=3,546), *TP53* background mutations found in uterine lavage (n=112), and *TP53* mutations reported in the UMD cancer database (n=71,051). (B) Uterine lavage mutations in cases and controls color-coded by their abundance in the UMD *TP53* database. For each sample, *TP53* Mutation Frequency is calculated as the number of mutations in the coding region divided by the total number of Duplex nucleotides sequenced in that region. Most samples harbor *TP53* mutations that are common in the database. (C) *TP53* mutation type and (D) *TP53* mutation spectrum are compared between mutations identified in uterine lavage from controls (n= 79), uterine lavage from cases (n=33) and the UMD database (n=71,051). (E) *TP53* mutation distribution map for uterine lavage (top panel) and UMD cancer database (bottom panel). Bars quantify the frequency of mutations at each codon. Colored background reflects a 20 BP sliding window average of mutation density around each position. Gene exons and protein domains are indicated in the middle section. UL: Uterine lavage.

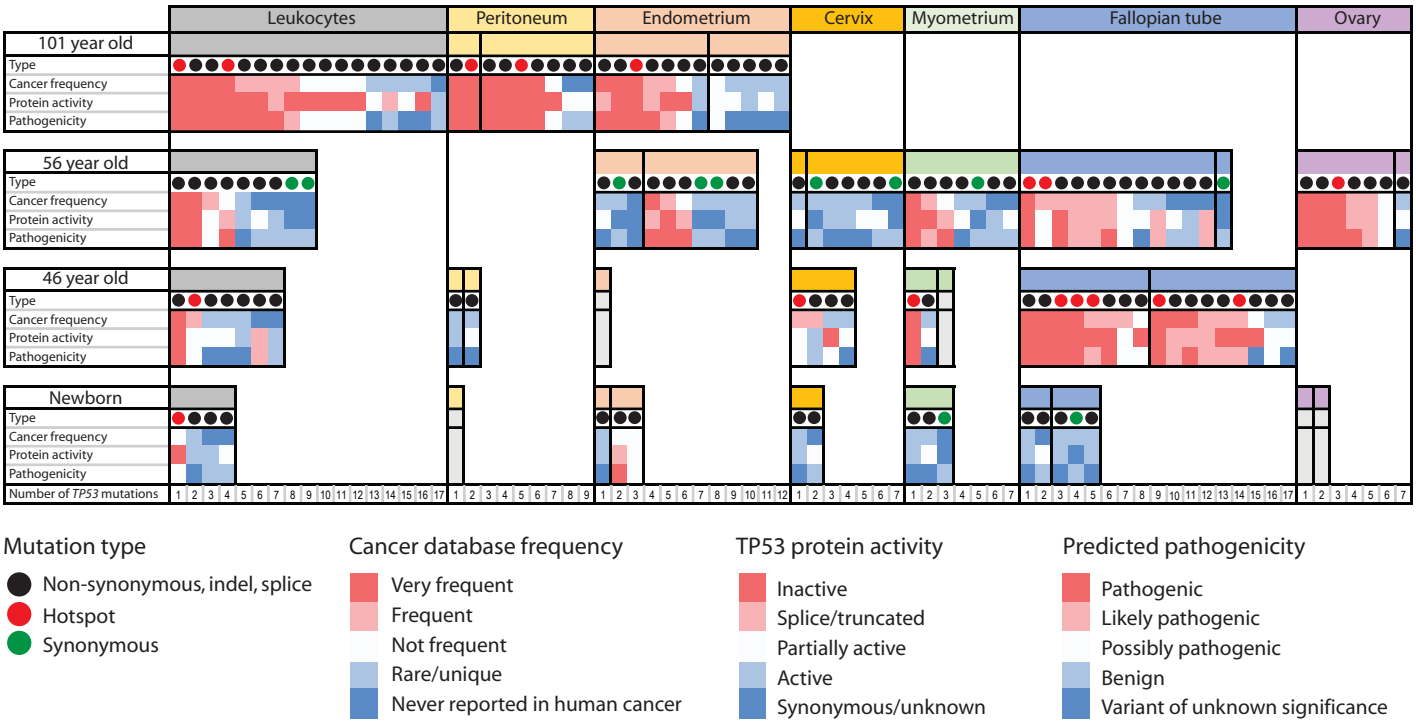


Figure 6. Characterization of *TP53* mutations in normal tissues over a century of the human lifespan *TP53* mutations identified by DS in leukocytes and gynecological tissue are indicated as columns within each tissue. Each mutation is characterized by 4 parameters: type (synonymous, non-synonymous and hotspots); frequency in cancer; protein activity; and predicted pathogenicity. The last 3 parameters are color-coded with red indicating 'cancer-like' mutations and blue indicating benign mutations. Vertical black lines separate different biopsies from the same tissue. Within each biopsy, mutations are ordered left to right by decreasing cancer frequency. Biopsies that were sequenced but no mutations were identified are shown in grey.

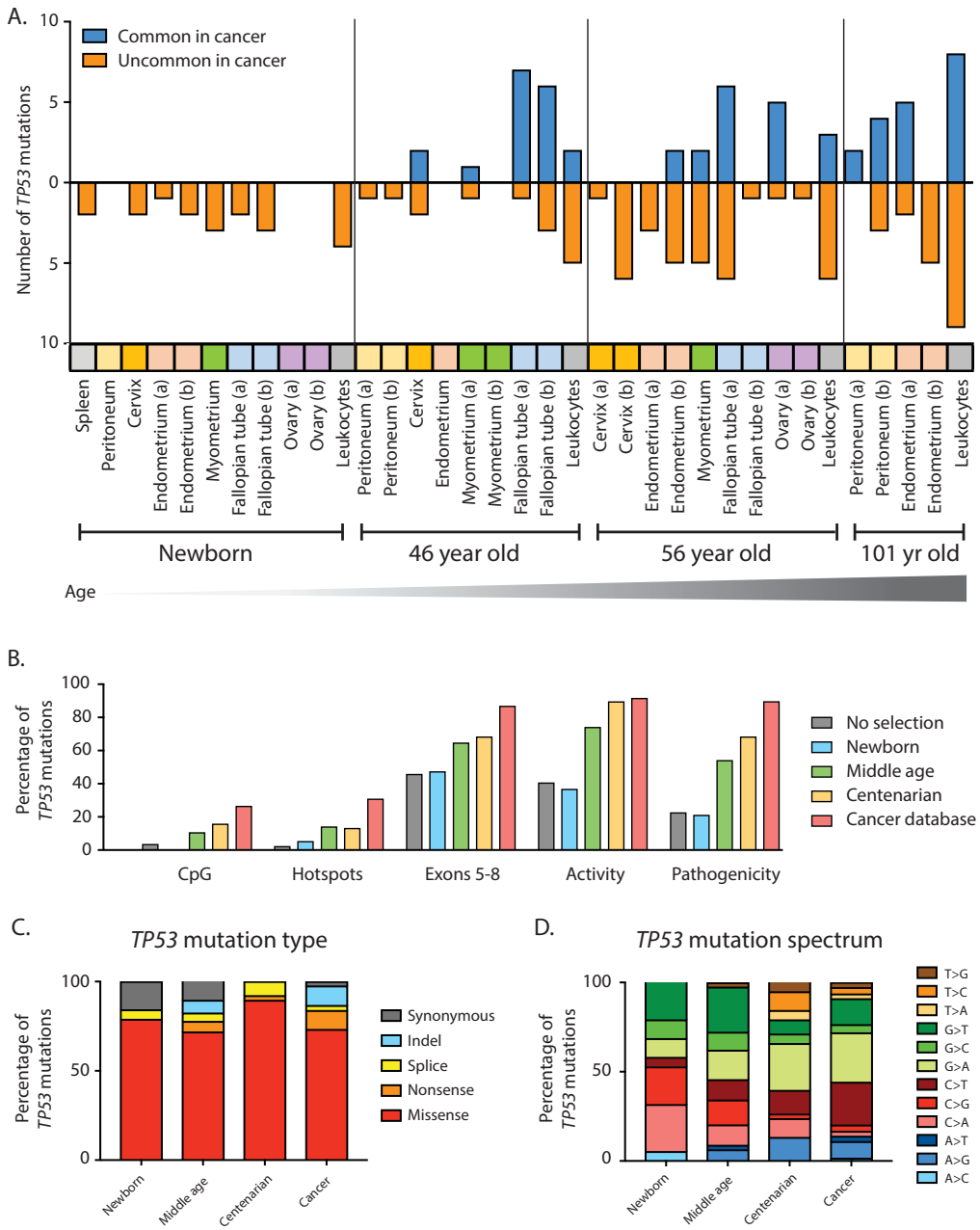


Figure 7. Cancer-associated *TP53* mutations are positively selected during normal aging (A) Across a variety of human tissues, *TP53* mutations accumulate with age and are progressively enriched for mutations commonly found in cancers. Tissues are color-coded. 'a' and 'b' indicate two biopsies from the same tissue. (B) Traits of positive selection are compared between all possible mutations in the *TP53* coding region (n=3,546), *TP53* mutations found in newborn (n= 19), middle age (n=85), and centenarian (n= 38), and *TP53* mutations reported in the UMD cancer database (n=71,051). (C) Distribution of *TP53* mutation type and (D) mutation spectrum for newborn, middle age, and centenarian mutations (n=19, 85, and 38, respectively) compared to UMD cancer database (n=71,051).

Supplementary Figure Legends

Figure S1. Comparison of mutation detection limit by sequencing accuracy for different NGS methods The positive predictive value (PPV) is the fraction of all mutations detected that are true positives. PPV depends on the target mutation frequency and the error rate of each method (given in parentheses). Mutations present at frequencies $>10^{-2}$ are confidently detected by all methods. However, PPV drops for mutations at lower frequencies, as these mutations represent a progressively smaller fraction of all detected mutations and thus become increasingly obscured by sequencing errors. For standard NG and single-strand molecular tagging methods, critical losses in PVV (<0.9) occur at mutation frequencies of ~10^{-2} and 10^{-3} (red and blue dotted lines), respectively. Thus, these values represent the limit for mutation detection of these methods. For Duplex Sequencing, however, the extremely low error rate allows accurate detection of mutations at frequencies below $\leq 10^{-6}$.

Figure S2. Association between number of independent TP53 mutations detected and total number of Duplex nucleotides sequenced The total number of TP53 mutations found (including exons and flanking intronic regions) in the 21 uterine lavages of the study was plotted against the total number of Duplex nucleotides sequenced in each sample. More TP53 mutations were identified in samples with more nucleotides sequenced (p=0.008 by Spearman's correlation test).

Figure S3. TP53 mutation frequency and characteristics by age for individual patient lavages in case-control study Data is parsed by mutation type, CpG dinucleotide site, exon type, and cancer-associated hotspots. Patients are divided in cases (women with ovarian cancer) and controls (women without ovarian cancer) and ordered by age within each group. For each

patient, TP53 mutations frequency was calculated as the number of TP53 mutations identified in the coding region divided by the total number of Duplex nucleotides sequenced in that region. For each trait, the fraction of mutations corresponding to each of the categories of analysis is indicated by color.

Figure S4. TP53 mutation frequency and characteristics by age including uterine lavages from the two middle age women in the normal tissue study The two new lavages correspond to a 46-year-old woman and a 56-year-old woman and are indicated by arrows. TP53 mutations type, frequency in cancer database, activity, pathogenicity, exon 5-8 location, CpG location, and hotspot location are indicated by color, with warm colors indicating 'cancer-like' features. The TP53 mutation frequency and the distribution of mutational cancer-like traits in the two new lavages are very similar to the data obtained for women of comparable age in the first part of the study.

Figure S5. Mutant allele frequency as a function of Duplex sequencing depth Each dot corresponds to a TP53 mutation identified in normal tissue. Mutations are color coded by subject. MAF is calculated as the number of times a mutation was observed divided by the depth of sequencing at the given position. Because MAF is inversely associated with depth, mutations identified in biopsies sequenced at a lower depth (mostly from the 46-year-old woman and newborn) present with higher MAF.

Figure S6. Analysis of mutations shared across multiple tissue samples within the same individual For each individual, all analyzed samples are listed and color-coded by tissue type. Mutations identified in more than one biopsy are indicated in columns with ratios provided for

the biopsies in which the mutation was identified. The ratios indicate the number of Duplex reads with the given mutation divided by the depth of sequencing in that position. Mutations found at MAF>1% and MAF>0.1% are indicated. It should be noted that the first mutation listed for the 101-year-old woman and the 46-year-old woman is the same and correspond to codon 220, one of the most common hotspots in TP53.

Figure S7. TP53 mutation frequency by tissue type (A) Association between number of TP53 mutations and total number of Duplex nucleotides sequenced in the TP53 coding region. Dots correspond to samples and are color-coded by individual of origin. (B) For each sample, TP53 mutation frequency was calculated as the number TP53 mutations identified in the coding region divided by the total number of Duplex nucleotides sequenced in that region. Subject age is indicated in the X-axis.

Figure S8. TP53 mutation frequency and characteristics by age for individual tissue samples Characterization of TP53 mutations identified in normal tissue from newborn, middle-aged, and centenarian females. TP53 mutation type, frequency in cancer database, activity, pathogenicity, CpG location, exon 5-8 location, and hotspot location are color-coded as leveled in the corresponding legends, with warm colors indicating “cancer-like” features.

Figure S9. TP53 mutation characteristics within non-invasively collected body fluids (A) heatmap indicating mutation type, frequency in caner database, impact on protein activity, and predicted pathogenicity for each of the TP53 mutations identified in leukocytes, cfDNA, peritoneal fluid, and uterine lavage. Categories are color-coding are the same as for Figure 6. Each column represents a mutation. (B) Comparison of TP53 mutation frequency in all tissues

collected from the 46-year-old woman, including liquid biopsies. (C) Comparison of Tp53 mutational features in all tissues collected from the 46-year-old woman including liquid biopsies (indicated by arrows).

Supplementary Figures

* **Note**: The following supplementary information has not been provided here, but is available in the published work:

- https://www.biorxiv.org/content/early/2018/11/04/457291
- **Database S1.** Seshat's long form analysis of *TP53* mutations identified by DS in uterine lavage
- **Database S2.** Seshat's long form analysis of *TP53* mutations identified by DS in normal tissue

- **Table S3.** *TP53* mutations detected by Duplex Sequencing in uterine lavage
- **Table S6.** *Tp53* mutations detected by Duplex Sequencing in normal tissue

Figure S1. Comparison of mutation detection limit by sequencing accuracy for different NGS methods

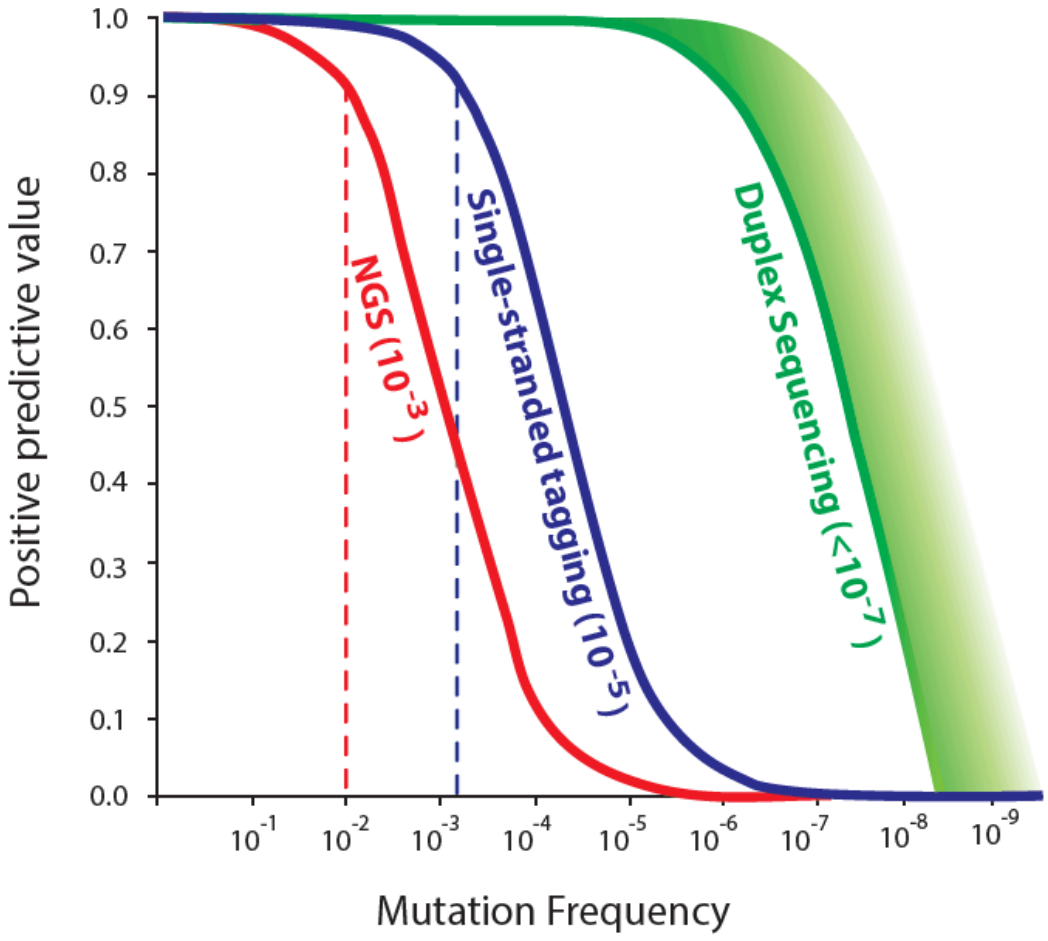


Figure S2. Association between number of independent *TP53* mutations detected and total number of Duplex nucleotides sequenced

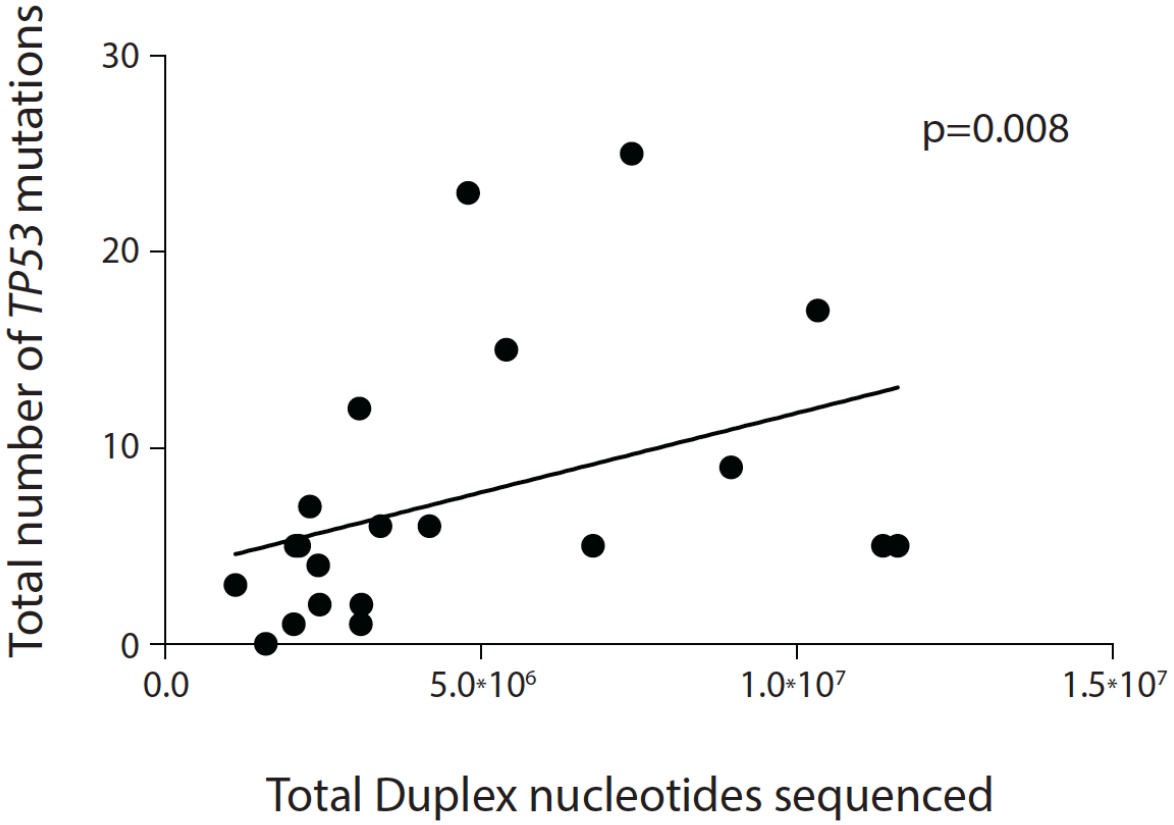


Figure S3. TP53 mutation frequency and characteristics by age for individual patient lavages in case-control study

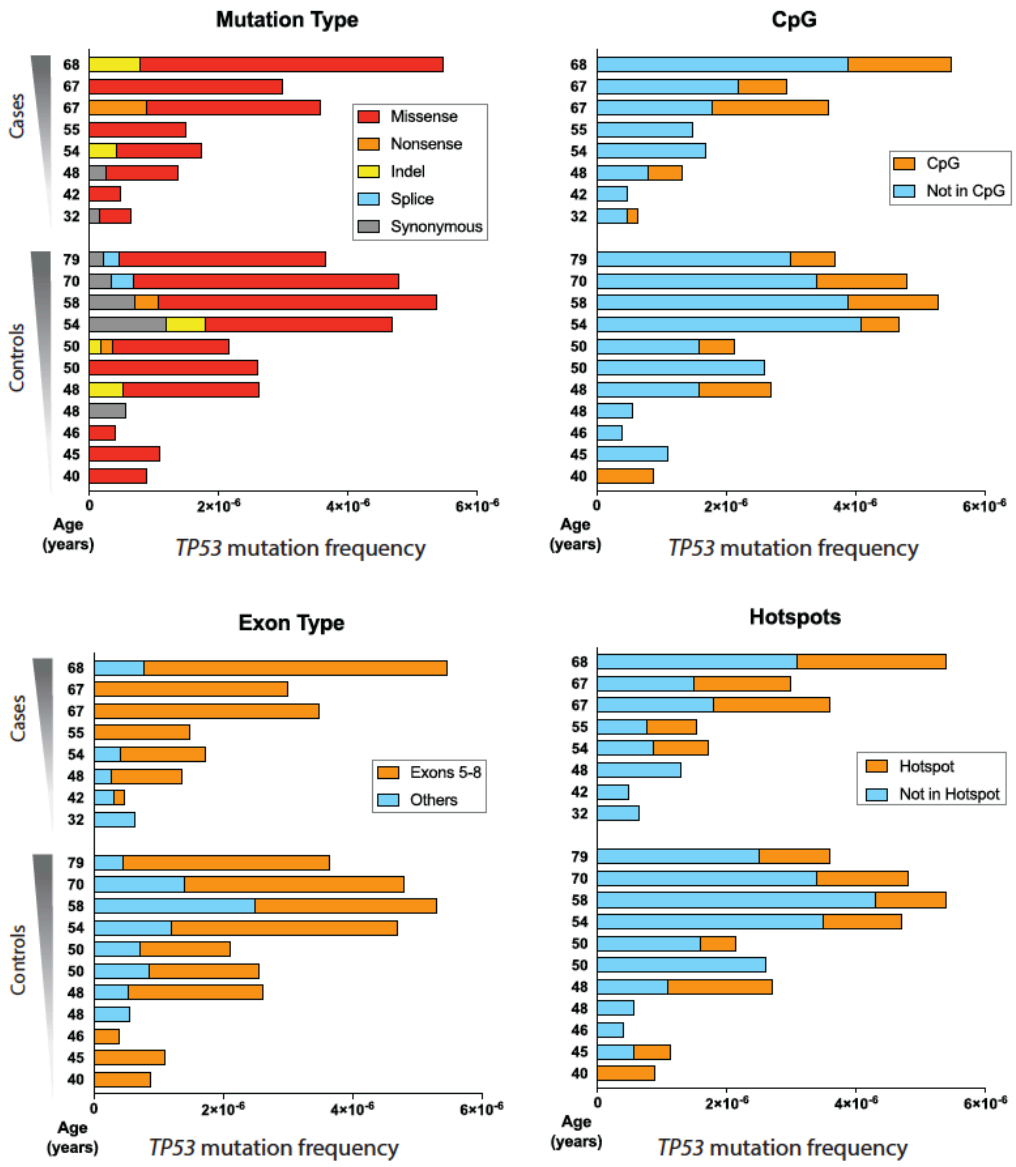


Figure S4. *TP53* mutation frequency and characteristics by age including uterine lavages from the two middle age women in the normal tissue study

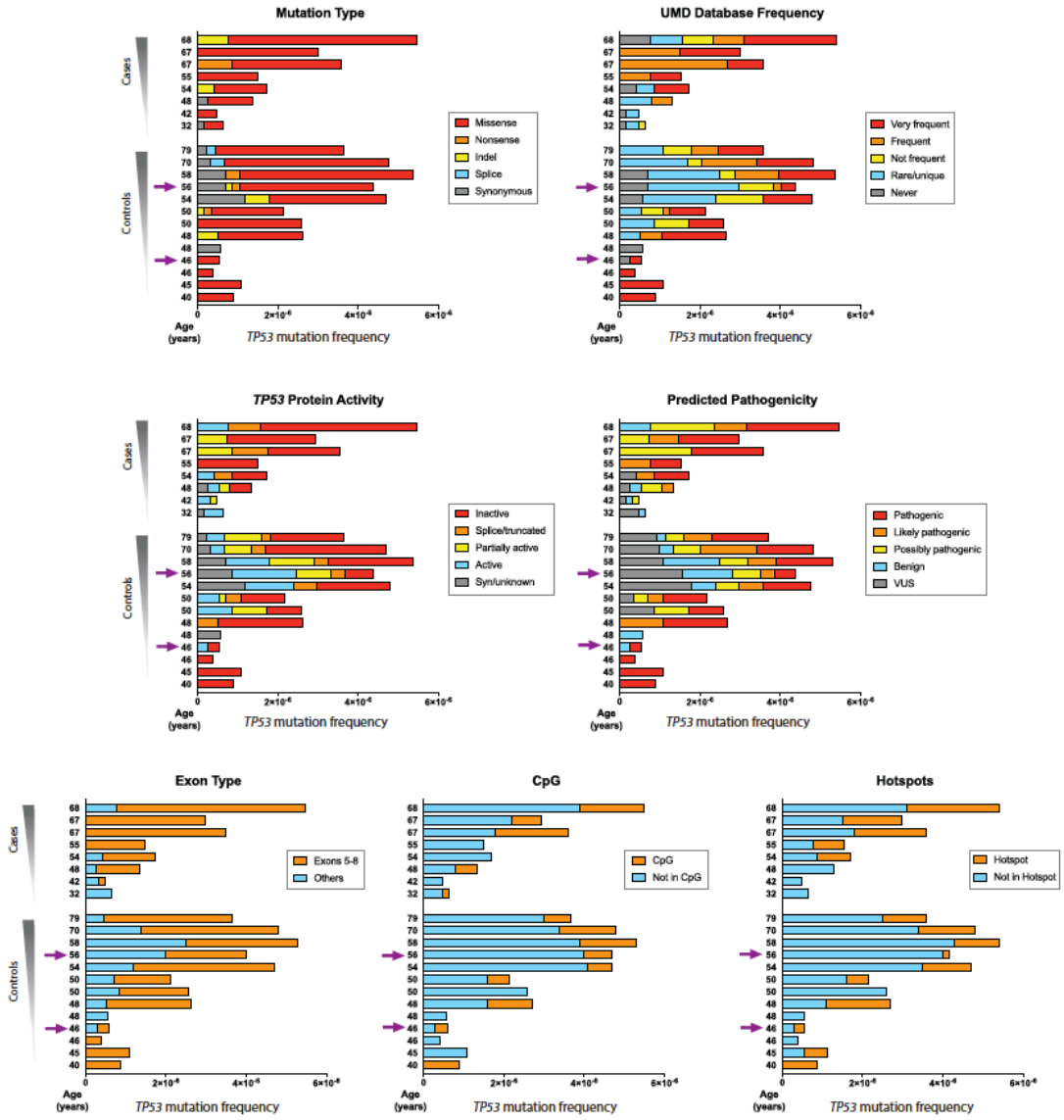


Figure S5. Mutant allele frequency as a function of Duplex sequencing depth

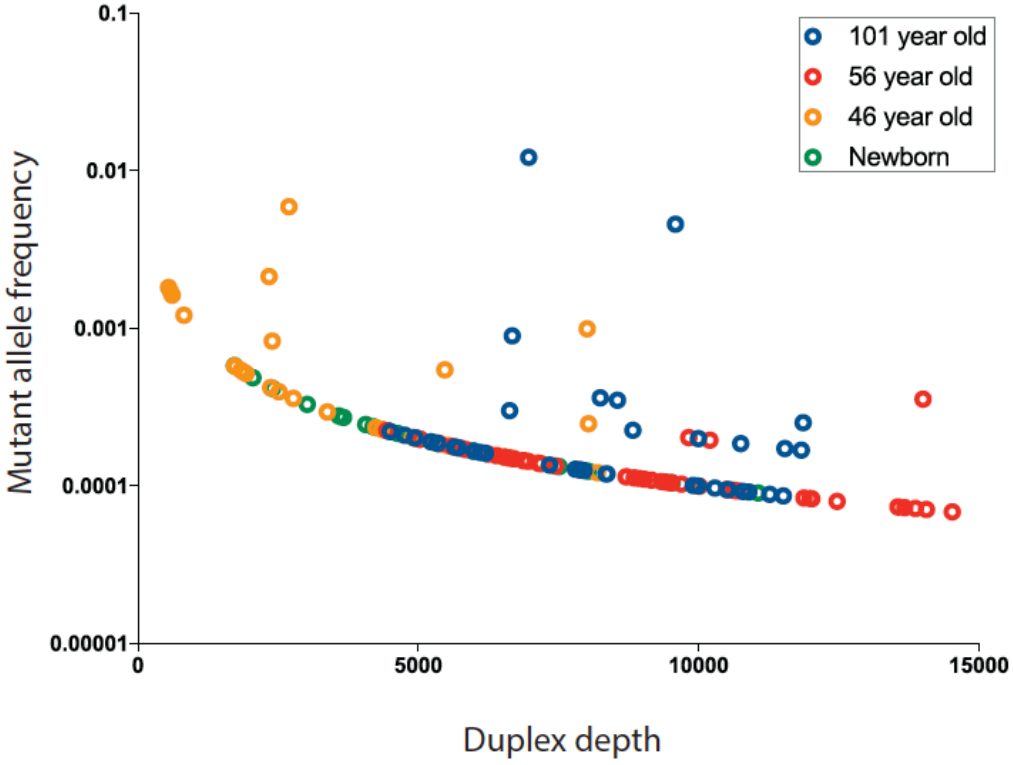


Figure S6. Analysis of mutations shared across multiple tissue samples within the same individual

TISSUE	MUTATIONS FOUND IN DIFFERENT TISSUES OF SAME INDIVIDUAL					
101 yo	**c.659A>G**	**c.596G>A**	**c.517G>A**	**c.455C>T**	**c.389T>G**	**c.149T>C**
A001-Leukocytes	85/6971	2/11543		44/9587	1/10797	1/11512
A001-Peritoneum (a)	1/4501			1/4939		
A001-Peritoneum (b)	1/5344	3/8248	1/7342	2/6633		
A001-Endometrium (a)	1/5712		1/7803		1/7961	
A001-Endometrium (b)						1/9988

56 yo	**c.151G>T**
A004-Leukocytes	1/13560
A004-Cervix (a)	
A004-Cervix (b)	
A004-Endometrium (a)	
A004-Endometrium (b)	
A004-Myometrium	
A004-FT (a)	
A004-FT (b)	
A004-Ovary (a)	
A004-Ovary (b)	
A004-Uterine lavage	1/4843

46 yo	**c.659A>G**	**c.524G>T**
A006-Leukocytes		1/6665
A006-Peritoneum (a)		
A006-Peritoneum (b)		
A006-Cervix		1/4349
A006-Endometrium		
A006-Myometrium (a)	16/2694	
A006-Myometrium (b)		
A006-FT (a)	1/1918	
A006-FT (b)		
A006-Uterine lavage		
A006-Peritoneal fluid		
A006-ctDNA		

	MAF>1%
	MAF>0.1%

Figure S7. *TP53* mutation frequency by tissue type

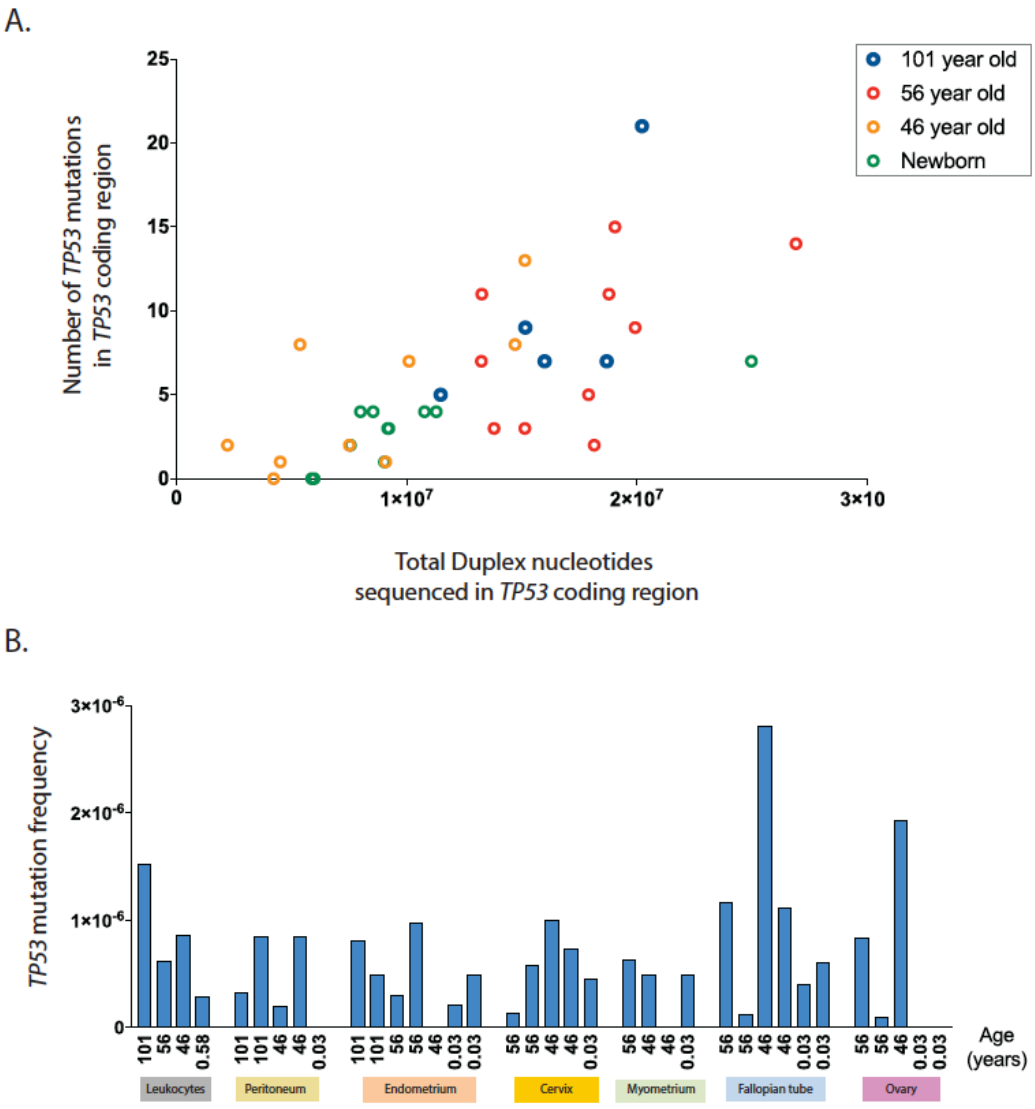
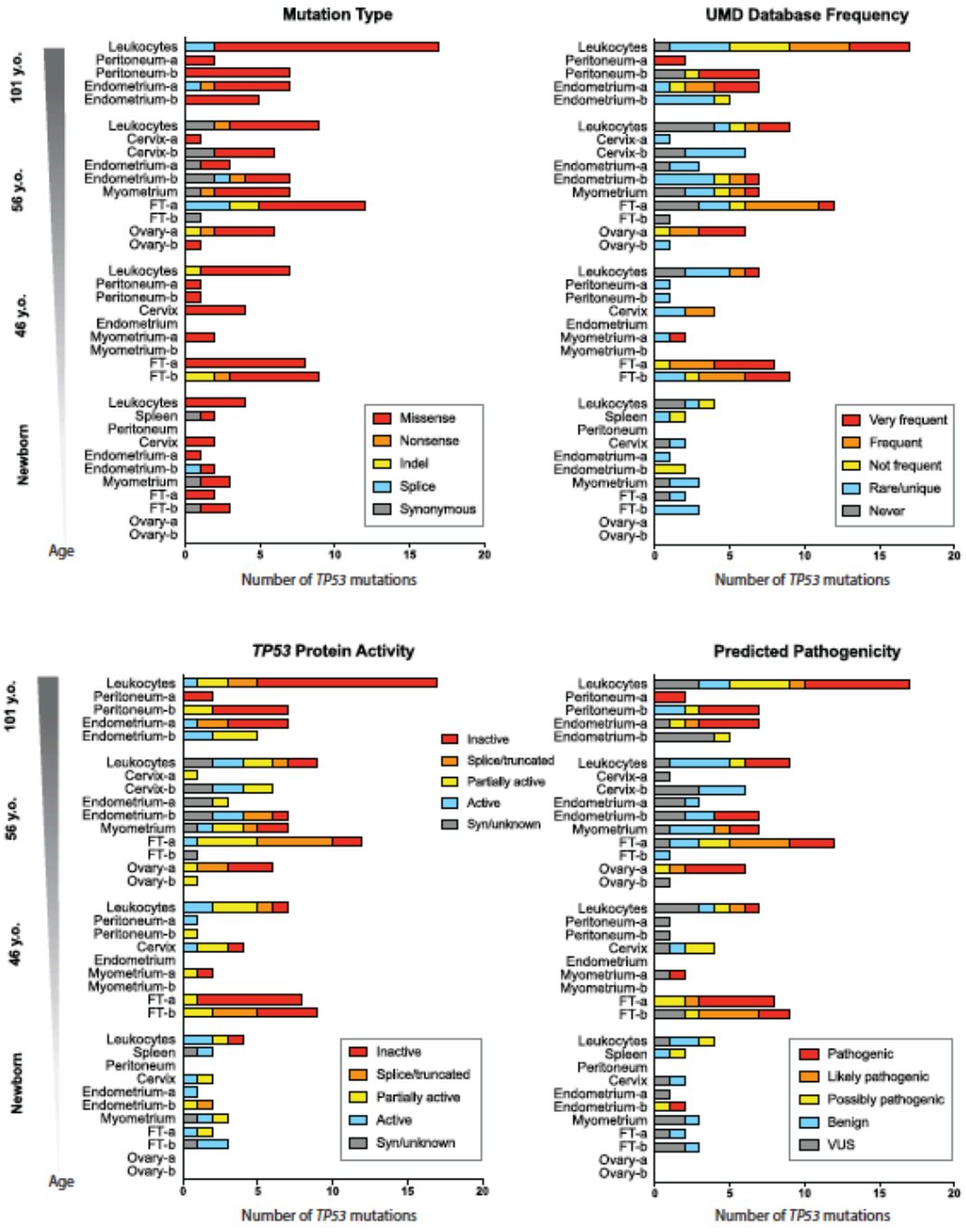


Figure S8. *TP53* mutation frequency and characteristics by age for individual tissue samples


Mutation Type
UMD Database Frequency
TP53 Protein Activity
Predicted Pathogenicity
101 y.o.
56 y.o.
46 y.o.
Newborn
Age
Leukocytes
Peritoneum-a
Peritoneum-b
Endometrium-a
Endometrium-b
Cervix-a
Cervix-b
Myometrium
FT-a
FT-b
Ovary-a
Ovary-b
Cervix
Endometrium
Myometrium-a
Myometrium-b
Spleen
Peritoneum
Missense
Nonsense
Indel
Splice
Synonymous
Very frequent
Frequent
Not frequent
Rare/unique
Never
Inactive
Splice/truncated
Partially active
Active
Syn/unknown
Pathogenic
Likely pathogenic
Possibly pathogenic
Benign
VUS
0
5
10
15
20
Number of TP53 mutations

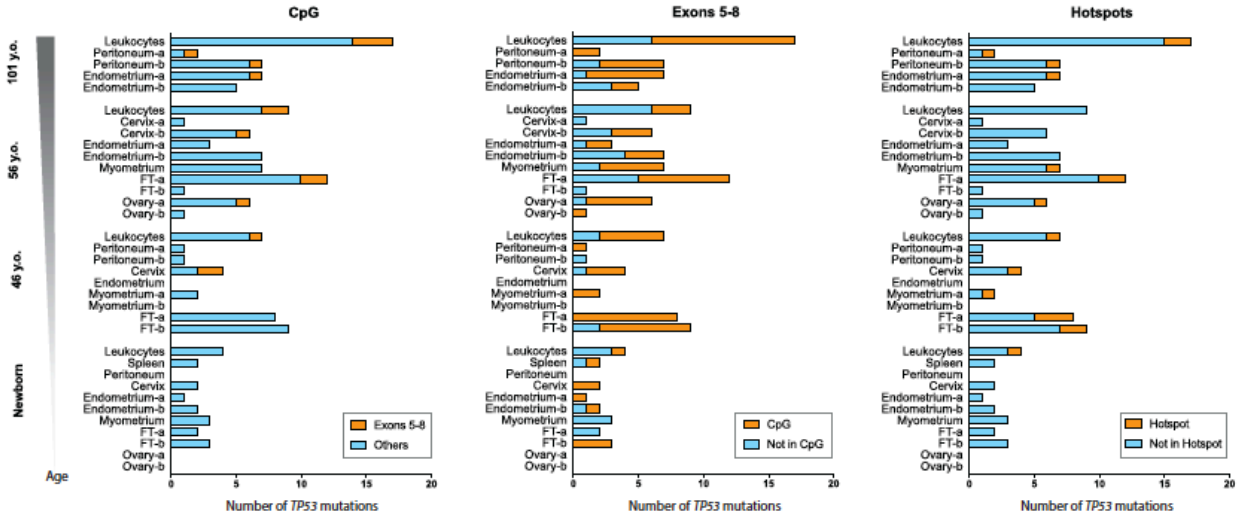
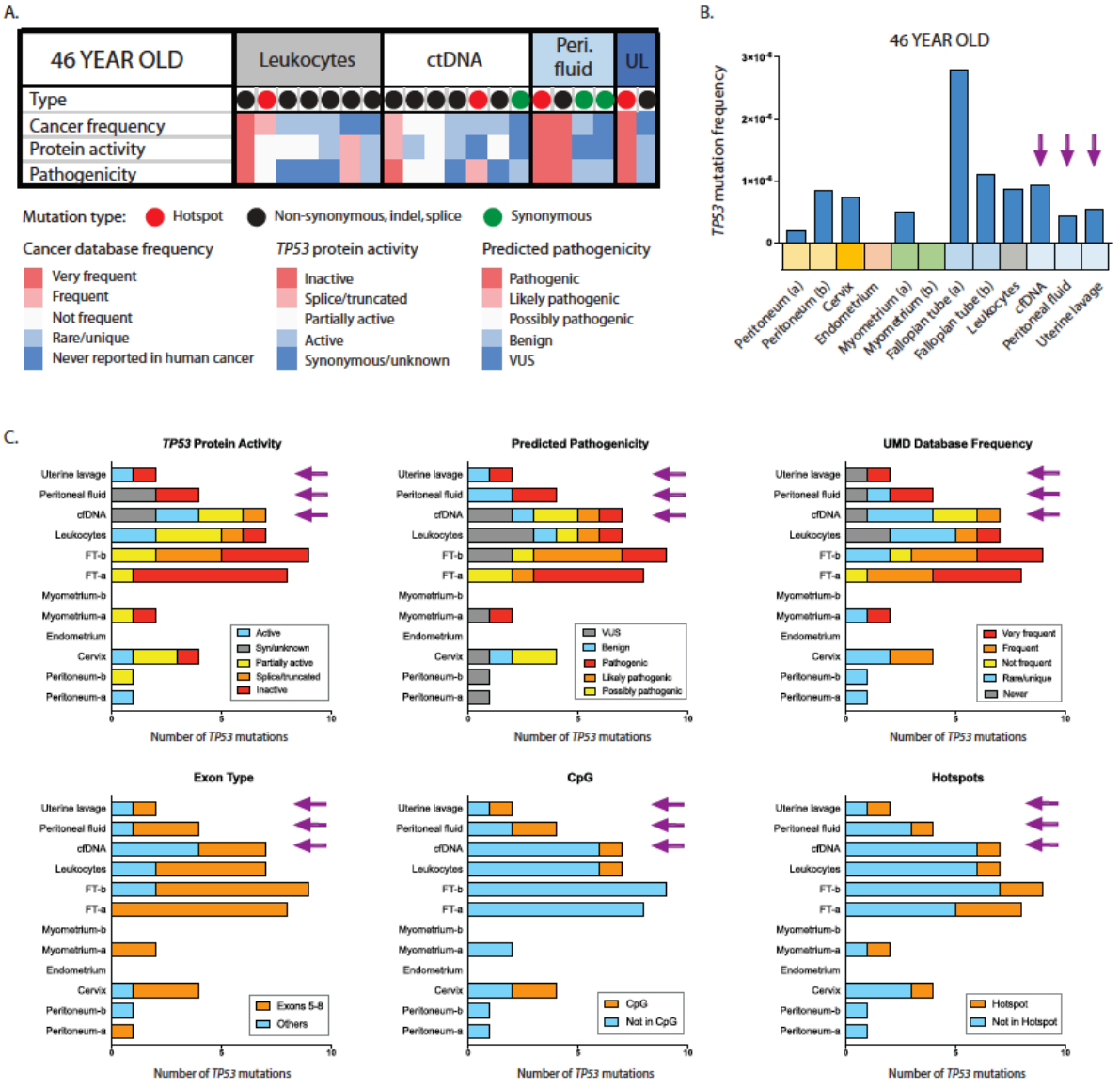


Figure S9. *TP53* mutation characteristics within non-invasively collected body fluids


A.
46 YEAR OLD
Leukocytes
ctDNA
Peri. fluid
UL
Type
Cancer frequency
Protein activity
Pathogenicity
Mutation type: Hotspot Non-synonymous, indel, splice Synonymous
Cancer database frequency
Very frequent
Frequent
Not frequent
Rare/unique
Never reported in human cancer
TP53 protein activity
Inactive
Splice/truncated
Partially active
Active
Synonymous/unknown
Predicted pathogenicity
Pathogenic
Likely pathogenic
Possibly pathogenic
Benign
VUS
B.
46 YEAR OLD
TP53 mutation frequency
Peritoneum (a)
Peritoneum (b)
Cervix
Endometrium
Myometrium (a)
Myometrium (b)
Fallopian tube (a)
Fallopian tube (b)
Leukocytes
cfDNA
Peritoneal fluid
Uterine lavage
C.
TP53 Protein Activity
Predicted Pathogenicity
UMD Database Frequency
Exon Type
CpG
Hotspots
Uterine lavage
Peritoneal fluid
cfDNA
Leukocytes
FT-b
FT-a
Myometrium-b
Myometrium-a
Endometrium
Cervix
Peritoneum-b
Peritoneum-a
Active
Syn/unknown
Partially active
Splice/truncated
Inactive
VUS
Benign
Pathogenic
Likely pathogenic
Possibly pathogenic
Very frequent
Frequent
Not frequent
Rare/unique
Never
Exons 5-8
Others
CpG
Not in CpG
Hotspot
Not in Hotspot
Number of TP53 mutations

Supplementary Tables

Table S1. Clinico-pathological characteristics of patients

Study	Patient group	Patient ID	Age	Menopause	Smoking	Diagnosis	In prior study	CA125	FIGO	Grade
Uterine lavage	Ovarian Cancer	case 1	67	yes	never	HGSOC	yes	417.4	IIIC	G3
Uterine lavage	Ovarian Cancer	case 2	69	yes	never	HGSOC	yes	27.2	IIIB	G3
Uterine lavage	Ovarian Cancer	case 3	51	perimenopausal	never	Signet ring, spread ovaries	yes	156.4	na	na
Uterine lavage	Ovarian Cancer	case 4	54	yes	prior	HGSOC	yes	1985	IV	G3
Uterine lavage	Ovarian Cancer	case 5	32	no	never	HGSOC	yes	114.3	IIIC	G3
Uterine lavage	Ovarian Cancer	case 6	42	no	current	HGSOC	yes	25.8	IIIA	G3
Uterine lavage	Ovarian Cancer	case 7	55	yes	current	HGSOC	yes	46.5	IIIA	G3
Uterine lavage	Ovarian Cancer	case 8	48	no	never	HGSOC	yes	266.1	IIIC	G3
Uterine lavage	Ovarian Cancer	case 9	67	yes	never	HGSOC	yes	1288	IV	G3
Uterine lavage	Ovarian Cancer	case 10	68	yes	never	HGSOC	no	442.6	IIIC	G3
Uterine lavage	Control	con 1	50	no	never	Benign ovarian cyst	no	na	na	na
Uterine lavage	Control	con 2	70	yes	prior	VAIN III	no	na	na	na
Uterine lavage	Control	con 3	54	yes	never	Benign ovarian cyst	no	na	na	na
Uterine lavage	Control	con 4	40	no	current	Uterine fibroma	no	na	na	na
Uterine lavage	Control	con 5	48	no	current	Uterine fibroma	no	na	na	na
Uterine lavage	Control	con 6	45	no	current	Benign ovarian cyst	no	na	na	na
Uterine lavage	Control	con 7	79	yes	never	Uterine fibroma	no	na	na	na
Uterine lavage	Control	con 8	58	yes	never	Benign ovarian cyst	no	na	na	na
Uterine lavage	Control	con 9	48	yes	never	Benign ovarian cyst	no	na	na	na
Uterine lavage	Control	con 10	46	no	current	Benign ovarian cyst	no	na	na	na
Uterine lavage	Control	con 11	50	no	current	Benign ovarian cyst	no	na	na	na

Table S2. Uterine lavage Duplex Sequencing coverage

Patient group	Patient ID	Total Duplex Nucleotides	Duplex Nucleotides in Coding Region	Median Depth Coding Region
Ovarian Cancer	case 1	2,067,165	1,130,319	959
Ovarian Cancer	case 2	1,114,686	613,433	520
Ovarian Cancer	case 3	1,591,699	861,824	731
Ovarian Cancer	case 4	4,183,394	2,313,739	1,962
Ovarian Cancer	case 5	11,366,916	6,142,239	5,210
Ovarian Cancer	case 6	11,599,106	6,173,923	5,237
Ovarian Cancer	case 7	2,450,431	1,294,458	1,098
Ovarian Cancer	case 8	6,772,145	3,769,668	3,197
Ovarian Cancer	case 9	2,421,453	1,349,645	1,145
Ovarian Cancer	case 10	2,294,407	1,284,941	1,090
Control	con 1	2,119,288	1,166,096	989
Control	con 2	5,398,366	2,956,801	2,508
Control	con 3	3,077,667	1,695,730	1,438
Control	con 4	2,039,443	1,125,027	954
Control	con 5	3,099,332	1,756,846	1,490
Control	con 6	3,105,613	1,750,731	1,485
Control	con 7	7,388,066	4,348,432	3,688
Control	con 8	4,800,008	2,811,580	2,385
Control	con 9	3,405,550	1,900,200	1,612
Control	con 10	8,963,986	4,951,373	4,200
Control	con 11	10,334,636	5,589,891	4,741

Table S4. Clinico-pathological characteristics of individuals that provided normal tissue

					NUMBER OF SAMPLES ANALYZED PER TISSUE TYPE											
Study	**Patient ID**	**Age**	**Menopause**	**Diagnosis**	**Leukocytes**	**Peritoneum**	**Endometrium**	**Cervix**	**Myometrium**	**Fallopian tube**	**Ovary**	**Spleen**	**Uterine lavage**	**cfDNA**	**Peritoneal fluid**	**Grand Total**
Normal tissue	A001	101	yes	Hip fracture after a fall, elderly	1	2	2									5
Normal tissue	A004	56	yes	Uterine leyomyoma	1		2	2	1	2	2		1			11
Normal tissue	A006	46	no	Uterine leyomyoma	1	2	1	1	2	2			1	1	1	12
Normal tissue	A007	7 months	na	ɔnchopulmonary dysplasia, prematur	1											1
Normal tissue	A008	2 weeks	na	Vein of Galen malformation		1	2	1	1	2	2	1				10
				TOTAL	4	5	7	4	4	6	4	1	2	1	1	39

Table S5. Normal tissue Duplex Sequencing coverage

Patient ID	Age	Tissue	Sample	Total Duplex Nucleotides	Duplex Nucleotides in Coding Region	Depth Coding Region
A001	101 years	Leukocytes	N1-1	20231850	11142821	9427
A001	101 years	Peritoneum	N1-2	11488272	6135358	5191
A001	101 years	Peritoneum	N1-3	15159370	8213620	6949
A001	101 years	Endometrium	N1-4	16002401	8650922	7319
A001	101 years	Endometrium	N1-5	18700925	10189381	8620
A004	56 years	Leukocytes	N1-15	26919646	14541451	12302
A004	56 years	Cervix	N1-8	13805787	7494026	6340
A004	56 years	Cervix	N1-9	18798010	10315701	8727
A004	56 years	Endometrium	N1-6	17911377	9755747	8254
A004	56 years	Endometrium	N1-7	13268169	7187740	6081
A004	56 years	Myometrium	N1-14	19930038	10974764	9285
A004	56 years	Fallopian tube	N1-10	19055091	10251270	8673
A004	56 years	Fallopian tube	N1-11	15144724	8155593	6900
A004	56 years	Ovary	N1-12	13248888	7136033	6037
A004	56 years	Ovary	N1-13	18160136	9997998	8459
A004	56 years	Uterine lavage	N1-18	9962705	5704924	4827
A006	46 years	Leukocytes	N2-23	14714971	8079419	6835
A006	46 years	Peritoneum	N2-21	9090845	4895509	4142
A006	46 years	Peritoneum	N2-20	2216635	1180319	999
A006	46 years	Cervix	N2-17	10113295	5407948	4575
A006	46 years	Endometrium	N2-14	4226008	2243645	1898
A006	46 years	Myometrium	N2-13	7520919	4008688	3391
A006	46 years	Myometrium	N2-22	4516257	2516123	2129
A006	46 years	Fallopian tube	N2-18	5371583	2850068	2411
A006	46 years	Fallopian tube	N2-19	15142991	8084403	6840
A006	46 years	Uterine lavage	N2-25-26	6740383	3664977	1550
A006	46 years	Peritoneal lavage	N2-24	16568058	9044752	7652
A006	46 years	ctDNA	N2-27	13061449	7510724	6354
A007	7 months	Leukocytes	N2-11	24989245	13635782	11536
A008	2 weeks	Spleen	N2-10	10785100	5776247	4887
A008	2 weeks	Peritoneum	N2-9	9043367	4808762	4068
A008	2 weeks	Cervix	N2-1	8000614	4313391	3649
A008	2 weeks	Endometrium	N2-4	8549897	4552872	3852
A008	2 weeks	Endometrium	N2-5	7559543	4041398	3419
A008	2 weeks	Myometrium	N2-6	11292962	6042979	5113
A008	2 weeks	Fallopian tube	N2-7	9179478	4873742	4123
A008	2 weeks	Fallopian tube	N2-8	9233815	4914615	4158
A008	2 weeks	Ovary	N2-2	5882208	3117987	2638
A008	2 weeks	Ovary	N2-3	5989339	3178318	2689

Table S7. Postprocessing of Seshat analytical variables into categorical variables

Seshat variable	Field description	Groups	5 group labels	2 group labels
Comment_1_Frequency	This single nucleotide variant is very frequent	1	very frequent	common in cancer
Comment_1_Frequency	This frameshift variant is very frequent	1	very frequent	common in cancer
Comment_1_Frequency	This single nucleotide variant is frequent	2	frequent	common in cancer
Comment_1_Frequency	This frameshift variant is frequent	2	frequent	common in cancer
Comment_1_Frequency	This single nucleotide variant is not frequent	3	not frequent	not common in cancer
Comment_1_Frequency	This single nucleotide variant is rare	4	rare/unique	not common in cancer
Comment_1_Frequency	This single nucleotide variant is unique	4	rare/unique	not common in cancer
Comment_1_Frequency	This frameshift variant is rare	4	rare/unique	not common in cancer
Comment_1_Frequency	This frameshift variant is unique	4	rare/unique	not common in cancer
Comment_1_Frequency	This variant has never been identified in human cancer	5	never	not common in cancer
Comment_2_Activity	Inactive	1	inactive	impaired activity
Comment_2_Activity	Splice site mutation; high probability of splicing defect	2	splice/truncated	impaired activity
Comment_2_Activity	The activity of truncated p53 is assumed to be nil	2	splice/truncated	impaired activity
Comment_2_Activity	This synonymous mutation is known to impair TP53 splicing	2	splice/truncated	impaired activity
Comment_2_Activity	Partial activity	3	partially active	impaired activity
Comment_2_Activity	Fully active	4	active	active/ likely active
Comment_2_Activity	Hyper active	4	active	active/ likely active
Comment_2_Activity	Synonymous mutation; unknown consequences	5	syn/unknown	active/ likely active
Comment_2_Activity	The consequences of this intronic mutation are unknown	5	syn/unknown	active/ likely active
Comment_2_Activity	The consequences of this mutation in the 3'UTR are unknown	5	syn/unknown	active/ likely active
Comment_2_Activity	The consequences of this mutation in the 5'UTR are unknown	5	syn/unknown	active/ likely active
Comment_2_Activity	The consequences of this mutation targeting isoforms beta are unknown	5	syn/unknown	active/ likely active
Comment_2_Activity	The consequences of this mutation targeting isoforms gamma are unknown	5	syn/unknown	active/ likely active
Pathogenicity	Pathogenic	1	pathogenic	expected pathogenic
Pathogenicity	Likely Pathogenic	2	likely pathogenic	expected pathogenic
Pathogenicity	Possibly pathogenic	3	possibly pathogenic	expected pathogenic
Pathogenicity	Benign	4	benign	unlikely pathogenic
Pathogenicity	VUS	5	VUS	unlikely pathogenic

Chapter 4: Targeted genome fragmentation with CRISPR/Cas9 enables fast and efficient enrichment of small genomic regions for sequencing

Running title: CRISPR-based target enrichment for efficient NGS

Daniela Nachmanson[1], Shenyi Lian[1†], Elizabeth K. Schmidt[1], Michael J. Hipp[1], Kathryn T. Baker[1], Yuezheng Zhang[1], Maria Tretiakova[1], Kaitlyn Loubet-Senear[1], Brendan F. Kohrn[1], Jesse J. Salk[2‡], Scott R. Kennedy[1]*, Rosa Ana Risques[1]*

[1]Department of Pathology, University of Washington, Seattle, WA 98195, USA.
[2]Department of Medicine, Division of Hematology and Oncology, University of Washington, Seattle, WA 98195, USA.

†Current address: Key laboratory of Carcinogenesis and Translational Research (Ministry of Education/Beijing), Department of Pathology, Peking University Cancer Hospital & Institute, Beijing, PR, China.
‡ Current address: TwinStrand Biosciences, Seattle, WA 98121, USA.

*These authors contributed equally

Key words: Target enrichment, Duplex Sequencing, Next-Generation Sequencing, NGS, CRISPR/Cas9

Published as:
Nachmanson D, Shenyi L, Schmidt EK, Hipp MJ, Baker KT, Zhang Y, et al. Targeted genome fragmentation with CRISPR/Cas9 enables fast and efficient enrichment of small genomic regions and ultra-accurate sequencing with low DNA input (CRISPR-DS). Genome Res 2018;10:1589-1599

Abstract

Next-generation sequencing methods suffer from low recovery, uneven coverage, and false mutations. DNA fragmentation by sonication is a major contributor to these problems because it produces randomly sized fragments, PCR amplification bias, and end artifacts. In addition, oligonucleotide-based hybridization capture, a common target enrichment method, has limited efficiency for small genomic regions, contributing to low recovery. This becomes a critical problem in clinical applications, which value cost-effective approaches focused on the sequencing of small gene panels. To address these issues, we developed a targeted genome fragmentation approach based on CRISPR/Cas9 digestion that produces DNA fragments of similar length. These fragments can be enriched by a simple size selection, resulting in targeted enrichment of up to ~49,000-fold. Additionally, homogenous length fragments significantly reduce PCR amplification bias and maximize read usability. We combined this novel target enrichment approach with Duplex Sequencing, which employs double-strand molecular tagging to correct for sequencing errors. The approach, termed CRISPR-DS, enables efficient target enrichment of small genomic regions, even coverage, ultra-accurate sequencing, and reduced DNA input. As proof-of-principle, we applied CRISPR-DS to the sequencing of the exonic regions of *TP53* and performed side-by-side comparisons with standard Duplex Sequencing. CRISPR-DS detected previously reported pathogenic *TP53* mutations present as low as 0.1% in peritoneal fluid of women with ovarian cancer, while using 10 to 100-fold less DNA than standard Duplex Sequencing. Whether used as stand-alone enrichment or coupled with high-accuracy sequencing methods, CRISPR-based fragmentation offers a simple solution for fast and efficient small target enrichment.

Introduction

In the past decade, NGS has revolutionized the fields of biology and medicine. However, standard NGS suffers from two major problems that negatively impact multiple applications: the limited efficiency of current target selection methods and the high error rate of the sequencing process. Targeted genome enrichment is essential to many applications that do not require whole genome sequencing and it is performed either by PCR or by hybridization capture. PCR is simple and efficient but does not scale well and suffers from biases that result in uneven coverage and false mutation calls (Kebschull and Zador 2014; Samorodnitsky et al. 2015). Hybridization capture improves coverage uniformity and mutation call accuracy but has low recovery, especially when the target region is small, which leads to the requirement of larger amounts of DNA (Samorodnitsky et al. 2015). An additional complication is that DNA is typically fragmented by sonication which introduces DNA damage resulting in sequencing errors (Park et al. 2017). Moreover, the heterogeneous fragment sizes generated by sonication are subject to PCR bias and contribute to uneven coverage. An alternative option to sonication is enzymatic fragmentation. This method resolves some issues but introduces different artifacts that also result in sequencing errors (Knierim et al. 2011). Thus, at the library preparation step, both methods of target selection suffer important limitations that lead to non-optimal sequencing outcomes, including uneven coverage, introduction of false mutations, and low recovery.

The second major problem of NGS is the high error rate inherent to the sequencing process. Illumina currently offers the most accurate sequencing platform with an estimated error rate of 10^{-3} (Goodwin et al. 2016). This error rate, however, translates into millions of false calls in each sequencing run and precludes the detection of low frequency mutations, which is critical for applications such as forensics, metagenomics, and oncology (Salk et al. 2018). Sequencing errors can be significantly reduced by the use of molecular barcodes, which are random DNA sequences

attached to the original DNA molecules before or during PCR. Single-stranded molecular barcodes, also known as 'unique molecular identifiers' or UMIs, produce a consensus with the reads derived from one DNA strand (Kinde et al. 2011) whereas double-stranded molecular barcodes introduce an additional level of correction by allowing the comparison of independent consensus sequences derived from the two complementary strands of the original DNA molecule (Schmitt et al. 2012). This additional level of correction is essential for removing polymerase errors occurring in the first round of PCR and subsequently propagated to all reads derived from a given DNA strand (Arbeithuber et al. 2016) . Duplex Sequencing (DS), the method that pioneered double-strand molecular barcodes (Kennedy et al. 2014; Schmitt et al. 2012), has an estimated error rate $<10^{-7}$, two orders of magnitude less than single-strand molecular barcode methods. This translates into the confident detection of mutations present at frequencies $<10^{-5}$, whereas single-strand molecular barcode methods show substantial decrease in accuracy at $\lesssim 10^{-3}$(Salk 2018). The extreme sensitivity of DS has been employed in a variety of applications including the detection of very low frequency somatic mutations in cancer and aging (Ahn et al. 2016; Kennedy et al. 2013; Krimmel et al. 2016; Reid-Bayliss et al. 2016; Hoekstra et al. 2016).

DS successfully addresses the problem of sequencing errors, but it suffers from the limitations of hybridization capture, which is required to perform target selection while preserving the strand recognition of molecular barcodes. As described above, hybridization capture is highly inefficient when selecting small targets (Winters et al. 2017), estimated at only 5-10% of reads are on-target for targets <50kb (Schmitt et al. 2015) . In DS, as well as in other panel-based sequencing approaches, the region of interest is usually designed to be small as a cost-effective trade-off for higher sequencing depth. In this situation, a successful approach for target enrichment is to perform two consecutive rounds of capture (Schmitt et al. 2015). However, this approach results in a time consuming, costly, and inefficient protocol that requires large amounts of DNA (Kennedy et al.

2014). For example, in DS at least 1µg of DNA has historically been needed to produce depths >3,000x (Krimmel et al. 2016), which is prohibitive in many applications that rely on small samples.

Here we present CRISPR-DS, a new method that addresses the two main problems of NGS: limited efficiency of target selection and high error rate. Target selection is facilitated by an enrichment of the regions of interest using the CRISPR/Cas9 system. *In vitro* digestion with CRISPR/Cas9 has been proven to be a useful tool for multiplexed excision of large megabase fragments and repetitive sequence regions for PCR-free NGS (Bennett-Baker and Mueller 2017; Shin et al. 2017). We reasoned that targeted *in vitro* CRISPR/Cas9 digestion could be used to excise similar length fragments covering the area of interest, which could then be enriched by size selection prior to library preparation. We designed this method to enable target enrichment while simultaneously eliminating sonication-related errors and biases arising from random genome fragmentation. In addition, by pairing this approach with double-strand molecular barcoding, we aimed to produce a method that preserves the sequencing accuracy of DS while increasing the recovery rate, thus enabling low DNA input and a simplified protocol for translational applications.

Results

Design of CRISPR-DS based on CRISPR/Cas9 target fragmentation and double strand molecular barcodes

CRISPR-DS is based on *in vitro* CRISPR/Cas9 excision of target sequences to generate DNA molecules of uniform length, which are then enriched by size selection. The versatility, specificity, and multiplexing capabilities of the CRISPR/Cas9 system enable its application for the excision of any target region of interest by simply designing guide RNAs (gRNA) to the desired cutting points. As a proof of principle, we developed the method for sequencing the exons of *TP53*. Further, in order to achieve high recovery and sequencing accuracy, we combined it with DS. The

main steps of the protocol are illustrated in Figure 1. First, target regions are excised from genomic DNA by multiplexed *in vitro* CRISPR/Cas9 digestion (Fig. 1a), followed by enrichment of the excised fragments by size-selection using SPRI beads (Fig. 1b). The selected fragments are then coupled with the double-strand molecular barcodes used in DS (Fig. 1c) (Kennedy et al. 2014) . These fragments are then amplified and captured with biotinylated hybridization probes as previously described for DS (Kennedy et al. 2014), with the exception that only one round of hybridization capture is required due to the prior enrichment of target fragments (see below). Finally, the library is sequenced and the resulting reads are analyzed to perform error correction based on the consensus sequences of both strands of each DNA molecule (Fig. 1d) (Kennedy et al. 2014). Due to the requirement of only one round of hybridization capture, the workflow of CRISPR-DS is almost one day shorter than standard-DS (Fig. 2, Supplementary Fig. 1), enabling a more cost-efficient and applicable method.

CRISPR/Cas9 cut fragments can be designed to be of homogenous length, reducing PCR bias and producing uniform coverage

Typically, genome fragmentation is performed with sonication, which generates randomly sized fragments that have different amplification efficiencies (Dabney and Meyer 2012) . Short fragments are preferentially amplified, resulting in uneven coverage of the regions of interest and decreased recovery. In DS, amplification bias introduces an additional problem because short fragments produce an excess of PCR copies that do not further aid error reduction. To produce a consensus, only three PCR copies of the same molecule are required. Additional copies waste resources because they produce sequencing reads but do not generate additional data. By using CRISPR/Cas9, gRNA can be designed such that restriction with Cas9 produces fragments of predefined, homogeneous size. We reasoned that these fragments would eliminate PCR bias,

leading to homogeneous sequencing coverage and minimizing wasted reads that are PCR copies of the same original molecule.

To test this approach, we designed gRNAs to specifically excise the coding regions and their flanking intronic sequence of *TP53* (Fig. 1a). Fragment length was designed to be ~500bp in order to maximize read space of an Illumina MiSeq v3 600 cycle kit while allowing for sequencing of the molecular barcode (10 bp) and 3'-end clipping of 30bp to remove low-quality bases produced in the later sequencing cycles. gRNAs were selected based on the highest specificity score that produced appropriate fragment length (Supplementary Table 1, Supplementary Data 1) (Hsu et al. 2013) . The fragment comprising exon 7 was designed to be shorter than the rest (336 bp) to avoid a homopolymeric run of T's in the flanking intronic region which induced poor base quality in reads that span this region (Supplementary Fig. 2).

We performed a side-by-side comparison of library performance (Fig. 3a-c) and sequencing coverage (Fig. 3d) of a sample DNA processed with CRISPR-DS vs. standard-DS (see Material and Methods). Standard-DS for *TP53* had been previously performed using sonication and published protocols (Kennedy et al. 2014; Krimmel et al. 2016). Visualization of the resulting sequencing library by gel electrophoresis showed that CRISPR restriction produced distinct bands/peaks (Fig. 3a-b) corresponding to the predesigned size of target fragments as opposed to the characteristic "smear" of libraries prepared by sonication. The discrete peaks allow confirmation of correct library preparation and target enrichment, preventing the sequencing of suboptimal libraries. Sequencing and mapping of the libraries demonstrated that targeted Cas9 restriction results in well-defined DNA fragments corresponding to the expected size (Fig. 3d). Importantly, these fragments exhibited extremely uniform sequencing depth. In contrast, sonicated DNA fragments resulted in significant variability in depth across target regions. Because DS reads correspond to individual DNA molecules, the uniform depth achieved by CRISPR-DS indicates a

homogenous representation of the original genomic DNA in the final sequencing output, confirming the proper excision of all fragments.

The ability to uniformly control the DNA insert size should not only provide homogenous depth, but also a more uniform number of copies of each molecule, minimizing the waste of unnecessary reads to produce a consensus sequence. We examined this possibility by counting the number of PCR copies for each molecular barcode and plotting it as a function of the DNA fragment size (Fig. 3c). Sonicated DNA exhibited a strongly negative association between DNA fragment size and the number of PCR copies as expected due to the fact that small DNA fragments are preferentially amplified (Fig. 3c, *blue*). In contrast, targeted fragmentation produced a consistent number of PCR copies for all fragments, including the smaller exon 7 fragment (Fig. 3c, *red*).

CRISPR/Cas9 cut fragments can be designed to be of optimal length to maximize read usage

An additional disadvantage of the variable fragment size produced by sonication is inefficient read usage: fragments that are too short generate overlapping reads that waste sequencing space, whereas fragments that are too long get sequenced on the ends, leaving captured but un-sequenced DNA in the middle (Fig. 4a). The programmable nature of Cas9 can be leveraged to reduce the amount of data "lost" by generating optimal length fragments tailored to the preferred number of sequencing cycles. To illustrate the improvement in read usage, we quantified the amount of deviation from the optimal fragment size (defined as the total number of sequencing cycles minus the total length of the molecular barcodes and 3'-end clipping) of seven samples independently processed with sonication and targeted fragmentation. Sonication produced significant variability in the amount of deviation from the optimal fragment size with a large fraction of fragments being twice the optimal size for one of the samples (Fig. 4b,c; Supplementary

Fig. 3). Indeed, only 9.1±4.2% of reads had inserts that were within 10% deviation from the optimal fragment length. Even samples with more stringent size selection had only ~61% of reads within the 10%-deviation window (Fig. 4c; Supplementary Fig. 3). In contrast, the same samples fragmented with Cas9 had 71.0±3.2% of reads within the same window range, with the vast majority of the reads outside the window being due the purposefully shorter Exon 7 fragment (Fig. 4b,c; Supplementary Fig. 2, 3). Exclusion of exon 7 from this analysis improved the percent of reads within the 10%-deviation window to 94.3±2.1%. These data indicate that targeted fragmentation can tightly control the fragment size to optimize read usage, thereby increasing the efficiency of sequencing.

CRISPR/Cas9 fragmentation enables target enrichment by size selection, eliminates one round of hybridization capture, and increases sequencing yield

While performing two rounds of capture substantially increases the number of on-target reads for standard-DS and other small target applications, the process is time consuming, expensive, and requires additional PCR steps that introduce further bias (Schmitt et al. 2015). We hypothesized that target enrichment via size selection of CRISPR/Cas9 digested fragments would sufficiently enrich for on-target DNA fragments and eliminate the need for a second capture. To test this hypothesis, we performed CRISPR/Cas9 digestion of targeted *TP53* exons (Fig. 1a) on a range of DNA input amounts (10-250ng) followed by SPRI size selection to remove undigested high molecular weight DNA fragments (>1kb in size). The selected DNA fragments were ligated to DS adapters, PCR amplified, and sequenced (see Material and Methods). No hybridization capture or any other type of target enrichment was performed. Mapping of raw reads revealed between 0.2% to 5% reads on-target (*i.e.* covering *TP53)* (Table 1). Given the fact that the *TP53* target region only amounts to 0.0001% of the human genome, this corresponds to ~2,000x to

50,000x enrichment, which matches or exceeds what is typically achieved with solution based hybridization for small target size (Schmitt et al. 2015; Winters et al. 2017). Notably, lower DNA inputs showed the highest enrichment, potentially reflecting more efficient digestion or improved removal of off-target, high molecular weight DNA fragments when they are in lower abundance.

These results suggested that a simple size selection step could be used in lieu of a targeted hybridization enrichment step. To test this possibility, we performed a side-by-side comparison of standard-DS (Kennedy et al. 2014) (both with one and two rounds of hybridization capture) and CRISPR-DS with only one round of hybridization capture. Three input amounts of the same control DNA extracted from normal human bladder tissue were sequenced in parallel for each of the methods. CRISPR-DS with one round of capture achieved >90% raw reads on-target (Fig. 5a), a significant improvement over standard-DS which only achieved ~5% raw reads on-target with a single capture, consistent with prior work (Schmitt et al. 2015) . In an independent experiment, we tested the reproducibility of this result with three different DNA samples that were sequenced with CRISPR-DS using one and two rounds of capture (Supplementary Fig. 4). Confirming the prior result, the three samples produced >90% raw reads on target using only one round of capture. The second round of capture only minimally increased raw reads on-target and is, therefore, unnecessary.

The side-by-side comparison of CRISPR-DS vs. standard-DS also demonstrated a substantial increase in recovery using CRISPR-DS. Sequencing recovery, also referred to as yield, is typically measured as the fraction or percentage of sequenced genomes equivalents compared to input genomes. Consistent with prior studies (Krimmel et al. 2016; Schmitt et al. 2012), standard-DS produced a recovery rate of ~1% across the different inputs, while CRISPR-DS recovery rate ranged between 6 and 12% (Fig. 5b). Notably, 25ng of DNA prepared with CRISPR-DS produced a post-processing depth comparable to 250ng with standard-DS, indicating that size

selection for excised fragments not only removes a step from the library preparation, but increases the recovery of input DNA, thereby enabling deep sequencing with greatly reduced DNA requirements.

Validation of CRISPR-DS recovery in an independent set of samples, including low quality DNA

We further confirmed the performance of CRISPR-DS in an independent set of 13 DNA samples extracted from bladder tissue (Supplementary Table 3). We used 250ng and obtained a median DCS depth of 6,143x, corresponding to a median recovery rate of 7.4%, in agreement with the prior experiment. Reproducible performance was demonstrated with technical replicates for two samples (B2 and B4, Supplementary Table 3). All samples had >98% reads on-target after consensus making, but the percentage of on-target raw reads ranged from 43% to 98%. We noticed that the low target enrichment corresponded to samples with DNA Integrity Number (DIN) <7. DIN is a measure of genomic DNA quality ranging from 1 (very degraded) to 10 (not degraded) (Jung et al. 2014). We reasoned that degraded DNA compromises enrichment by size selection, and the poor yield could be mitigated by removing low molecular weight DNA prior to CRISPR/Cas9 digestion. To test this hypothesis, we used the pulse-field feature of the BluePippin system to select high molecular weight DNA (> 8kb) from two samples with degraded DNA (DINs 6 and 4). This pre-enrichment resulted in successful removal of low molecular weight products and increased on-target raw reads by 2-fold and DCS depth by 5-fold (Supplementary Fig. 5). These results indicate that enrichment of high molecular weight DNA could be used as a solution for successful CRISPR-DS performance in partially degraded DNA.

Validation of CRISPR-DS for the detection of low-frequency mutations

Since CRISPR-DS uses a double-strand barcoding technique identical to standard-DS, it should theoretically have the same ability to identify low-frequency mutations. To prove this point, we sequenced a defined mixture of mutations with both CRISPR-DS and standard-DS. Two samples with known *TP53* variants were mixed at dilutions of 1:1, 1:10, 1:100 and 1:1000. Because the spiked-in sample was heterogenous, this experiment yielded a wide range of expected MAF to be compared with the MAF obtained by CRISPR-DS and standard-DS (Supplementary Fig. 6). The two methods showed very high correlations between expected and observed MAF (CRISPR-DS R^2=0.980, standard-DS R^2=0.984), as well as very high correlation between observed MAFs with each method (R^2=0.996). Most importantly, both methods could detect mutations at frequencies of ~0.001 and CRISPR-DS, but not standard-DS, detected an expected mutation at frequency of 0.0005. No additional, unexpected mutations were detected with any of the methods. Thus, these data demonstrate that the extremely high sensitivity and accuracy of double-strand molecular barcoding employed by standard-DS is preserved with CRISPR-DS.

To validate the sensitivity of CRISPR-DS with clinical samples, we analyzed four peritoneal fluid samples collected during gynecological surgery from women with ovarian cancer and previously analyzed for *TP53* mutations using the standard-DS protocol (Krimmel et al. 2016). The tumor mutation was previously identified in the four samples: in one sample at a high frequency (68.5%) and at a very low frequency (around or below 1%) in the remaining 3 samples. CRISPR-DS detected the tumor mutation in all samples at frequencies comparable to what was reported in the original study (Table 2) (Krimmel et al. 2016). In addition to the tumor mutation, standard-DS also revealed the presence of extremely low frequency (<0.1%) *TP53* mutations in these samples. These “biological background” mutations are not tumor-derived, but age-related (Krimmel et al. 2016). Standard-DS detected between one and five biological background mutations in each of the

samples, representing an overall mutation frequency of about ~$1x10^{-6}$. Similarly, CRISPR-DS identified biological background mutations in the 4 samples at a comparable overall mutation frequency (Supplementary Fig. 7). These results indicate that CRISPR-DS matches the performance of standard-DS in clinical samples (Krimmel et al. 2016).

Table 2 also illustrates a critical advantage of CRISPR-DS compared to standard-DS in terms of translational applicability: the reduced requirement of input DNA. Standard-DS of these peritoneal fluid samples required between 3-10 μg of DNA to compensate for the ~1% recovery rate of standard-DS and to achieve the high depth necessary to detect low frequency tumor mutations. With CRISPR-DS, we only used 100ng of DNA (30-100 fold less than what was used for standard-DS), and obtained comparable DCS depth to standard-DS (Table 2). Recovery rates ranged between 6 and 12%, as in prior experiments (Fig. 5 and Supplementary Table 3). These results represent an efficiency increase of 15x-200x compared to standard-DS with the same DNA. Notably, CRISPR-DS not only preserved sensitivity for mutation detection, increased sequencing recovery, and reduced DNA input, but also shortened the protocol by nearly one day (Supplementary Fig. 1), making it a more cost effective option for accurate deep sequencing of samples with limited DNA amounts.

Discussion

Here we have developed a new approach for target enrichment based on CRISPR/Cas9 fragmentation followed by size selection and we have combined this approach with DS, producing a new method called CRISPR-DS. CRISPR-DS merges the increased efficiency provided by CRISPR-based targeted genome fragmentation with the high accuracy of sequencing provided by double strand molecular barcodes, thus enabling ultra-accurate sequencing of small target regions using minimal DNA inputs. In addition to CRISPR-DS, the CRISPR-based target

enrichment approach can be used in combination with other methods for targeted sequencing to improve recovery of small targets and to reduce PCR bias and uneven coverage arising from random fragment sizes.

Targeted sequencing remains a cost effective alternative to whole genome-sequencing, especially when high depth is desired (Samorodnitsky et al. 2015). In multiple applications, such as oncology, the goal is to sequence a small panel of relevant genes with high accuracy in order to find low frequency mutations. While the selected target panel can be amplified by PCR, this method creates uneven coverage and false mutations, thus hybridization capture is typically preferred (Samorodnitsky et al. 2015). Hybridization capture improves coverage uniformity and removes certain artefactual mutations but does not resolve these issues completely. A major disadvantage in hybridization-based sequencing methods is the reliance on sonication for genome fragmentation, which generates DNA fragments of random size. We have demonstrated that this size heterogeneity produces two problems that can be solved by replacing sonication with CRISPR-based genome fragmentation. The first problem is PCR bias, which results in the preferential amplification of short DNA fragments. PCR bias leads to wasted reads that contain an excess of PCR copies of the same molecule. While these reads can be removed bioinformatically (Li 2011), the amplification advantage of certain molecules can lead to uneven coverage and reduced recovery (Kozarewa et al. 2015). In methods that employ molecular barcodes, such as DS, three PCR copies are typically sufficient to generate a consensus sequence (Kennedy et al. 2014). Thus, additional sequencing of PCR copies does not produce additional data and only wastes resources. We have demonstrated that with CRISPR-based fragmentation all fragments amplify similarly. This homogeneous amplification translates into uniform coverage across all targeted regions, a critical feature when the goal is to detect low frequency mutations in selected panel of genes.

The second problem associated with the heterogeneous fragment sizes relates to reduced data yield at the read level. Because sonication allows minimal control over fragment size, a large proportion of fragments are typically too short or too long compared to the optimal length size determined by the number of sequencing cycles. When reads are too short, paired-end reads overlap and the middle region is double-sequenced. Conversely, when reads are too long, the middle part of the DNA fragment, which may contain a variant or region of interest, remains un-sequenced. This inefficient read usage is solved with CRISPR-based target selection because the fragments are tailored to the desired read length.

CRISPR-based target fragmentation also offers two additional advantages. First, homogeneously sized DNA fragments can be visualized to confirm library target enrichment prior to sequencing. In sonication-based hybridization capture, the gel electrophoresis for a target-enriched library looks identical to a library with no target enrichment. This issue can result in the costly waste of a sequencing run where the majority of reads are in off-target regions. We show that the defined fragment lengths created by CRISPR-based digestion produce distinct peaks, which are easily visualized and confirm that the sequencing library is target-enriched. A second advantage of Cas9 digestion over sonication is the elimination of sonication-induced sequencing errors (Park et al. 2017) and the preservation of double stranded DNA at the ends of fragments. Sonication produces ssDNA at the end of molecules, which is susceptible to damage and converted into "pseudo-dsDNA" by end repair. This process has the potential to introduce false variant calls, but it is prevented by CRISPR-DS because Cas9 produces blunt ends, which do not require end repair.

In the context of small target sequencing by hybridization-capture, the major advantage introduced by CRISPR-based target enrichment is increased recovery, that is, percentage of input genomes that produce sequencing data. Hybridization capture is notably inefficient, especially for

small target regions (Schmitt et al. 2015; Winters et al. 2017). As demonstrated with our experiments and in agreement with prior studies, the average recovery rate of DS is ~1% which translates to at least 1 μg of DNA being needed to produce an average depth of ~3,000x. This recovery is improved 10-fold by the addition of CRISPR-based target enrichment and the elimination of one round of capture. We have demonstrated that by simply excising the genomic regions of interest and performing size selection, we can achieve a level of enrichment comparable to a single round of capture. By performing this step prior to library preparation, only one round of hybridization capture is needed, greatly minimizing DNA loss and increasing recovery. Therefore, using CRISPR-based target enrichment prior to DS achieves the same depth with 10 times less DNA.

Though CRISPR-DS addresses several needs in targeted NGS, it could still benefit from optimizations. First, improvements could be made to increase the recovery of degraded samples. Currently, in order to perform efficient target enrichment with CRISPR/Cas9 digestion and size selection, degraded samples must be pre-processed to remove low molecular weight fragments. We performed this pre-processing using electrophoretic size selection with the BluePippin system. However, to minimize loss of DNA, high molecular weight DNA could be selected with alternative methods such as micro-column filters. Second, although CRISPR-DS is highly efficient with a wide range of DNA inputs (10-500ng), we noticed that the best recovery was achieved with smaller starting DNA amounts (10-25ng). Since our goal was to achieve higher depth with low DNA input, this was not problematic. However, further efforts can be directed to improve recovery from larger DNA inputs, since this would be required if extremely high (>10,000x) target depths are desired. Lastly, although CRISPR-DS provides an effective solution for small-target region deep sequencing, the method becomes costly for deep sequencing of large genomic regions, an inherent problem of deep sequencing. Nevertheless, fragmentation by CRISPR/Cas9 followed by size

selection as a generic target enrichment technique can easily be scaled to many genomic regions as each region only requires the addition of the appropriate gRNAs for target excision. Sequencing of larger fragments can be achieved by tiling the gRNA along the desired fragment. Regarding the additional cost arising from the synthesis of gRNAs, it is important to note that a typical synthesis results in enough gRNA for thousands of cutting reactions. Over time this upfront cost becomes minimal compared with the substantial savings in time and reagents generated by the elimination of the second round of hybridization capture and by a more efficient use of sequencing space. Thus, CRISPR-DS becomes more economical than standard-DS in the long term, especially for small to moderate size panels (1-100kb) that are deployed on large numbers of samples.

In conclusion, we have demonstrated that CRISPR/Cas9 fragmentation followed by size selection enables efficient target enrichment by increasing the recovery of hybridization capture and eliminating the need for a second round of capture for small target regions. In addition, it eliminates PCR bias, maximizes the use of sequencing resources, and produces homogeneous coverage. This fragmentation method can be applied to multiple sequencing modalities that suffer from these problems. Here we have applied it to DS in order to produce CRISPR-DS, an efficient, highly accurate sequencing method with significantly reduced input DNA requirements. CRISPR-DS has broad application for the sensitive identification of mutations in situations in which samples are DNA-limited, such as forensics and early cancer detection.

Methods

Samples

The samples analyzed included de-identified human genomic DNA from peripheral blood, bladder with and without cancer, and peritoneal fluid DNA from a prior study (Krimmel et al. 2016). Only peritoneal fluid samples had patient information available, which was necessary to confirm the tumor mutation. The peritoneal fluid samples were obtained from the University of

Washington Gynecologic Oncology Tissue Bank, which collected specimens and clinical information after informed consent under protocol number 27077 approved by the University of Washington Human Subjects Division institutional review board. Frozen bladder samples were obtained from the University of Washington Genitourinary Cancer Specimen Biorepository and from unfixed or frozen autopsy tissue with waiver of consent under protocol number 52389 approved by the Fred Hutchinson Cancer Research Center Human Subjects Division institutional review board. The remainder of the study samples were used solely to illustrate technical aspects of the technology, no patient information was available, and interpretation of the mutational status of *TP53* is not reported. DNA was extracted with the QIAamp DNA Mini kit (Qiagen, Inc., Valencia, CA, USA) with care being taken to not heat the sample above the recommended 56°C, which is essential to preserve the double-stranded nature of each DNA molecule prior to ligation of DS adapters. DNA was quantified with a Qubit HS dsDNA kit (ThermoFisher Scientific). DNA quality was assessed with Genomic TapeStation tapes (Agilent, Santa Clara, CA) and DNA integrity numbers (DIN) were recorded. Peripheral blood DNA and peritoneal fluid DNA had DIN>7 reflecting good quality DNA with no degradation. Bladder samples, however, were purposely selected to include different levels of DNA degradation. Samples B1 to B13 had DINs between 6.8 and 8.9 and were successfully analyzed by CRISPR-DS (Supplementary Table 3). Samples B14 and B16 had DINs of 6 and 4, respectively, and were used to demonstrate pre-enrichment of high molecular weight DNA with the BluePippin system (see below and Supplementary Fig. 5).

CRISPR guide design

CRISPR/Cas9 uses a gRNA to identify the site of cleavage. gRNAs are composed of a complex of CRISPR RNA (crRNA), which contains the ~20bp unique sequence responsible for target recognition, and a trans-activating crRNA (tracrRNA), which has a universal sequence (Ran

et al. 2013). To select the best gRNAs to excise *TP53* exons we used the CRISPR MIT design website (http://CRISPR.mit.edu). The selection criteria were: (1) production of fragments of ~500bp covering exons 2-11 of *TP53* and (2) highest MIT website score (Supplementary Table 1 and Supplementary Data 1). Additionally, we recommend avoiding gRNAs that cover sites with known polymorphisms or mutational hotspots as this could potentially decrease the affinity of the gRNA and lead to reduced fragment depth. For exon 7, a smaller size fragment was required in order to avoid a proximal poly-T repeat (Supplementary Fig. 2). We designed a total of 12 gRNA, which excised *TP53* into 7 different fragments (Figure 1a). All gRNA had scores >60. 10 gRNAs were successful with the first chosen sequence and 2 had to be redesigned due to poor cutting. Initially, the quality of the cut was assessed by reviewing the alignment of the final DCS reads with Integrative Genomics Viewer (Robinson et al. 2011). Successful guides produced a typical coverage pattern with sharp edges in region boundaries and proper DCS depth (Figure 3d). Unsuccessful guides led to a drop in DCS depth and the presence of long reads that spanned beyond the expected cutting point. In order to simplify and speed up the assessment of guides, especially with scores <80, as well as to assess the activity of the Cas9/gRNA complex over time, we designed a synthetic GeneBlock DNA fragment (IDT, Coralville, IA) that included all gRNA sequences interspaced with random DNA sequences (Supplementary Data 2). 3ng of GeneBlock DNA were digested with each of the gRNAs using the CRISPR/Cas9 in vitro digestion protocol described below. After digestion, the reactions were analyzed by TapeStation 4200 (Agilent Technologies, Santa Clara, CA, USA) (Supplementary Fig. 8). The presence of predefined fragment lengths confirms: (1) proper gRNA assembly; (2) the ability of the gRNA to cleave the designed site; (3) proper nuclease activity of Cas9.

CRISPR/Cas9 *in vitro* digestion of genomic DNA

The *in vitro* digestion of genomic DNA with *S. pyogenes* Cas9 Nuclease requires the formation of a ribonucleoprotein complex, which both recognizes and cleaves a pre-determined site. This complex is formed with gRNAs (crRNA + tracrRNA) and Cas9. For multiplex cutting, the gRNAs can be complexed by pooling all the crRNAs, then complexing with tracrRNA, or by complexing each crRNA and tracrRNA separately, then pooling. The second option is recommended by manufacturers because it eliminates competition between crRNAs, however, in the limited set of gRNAs tested here both methods of complexing were comparable. Decreased efficiency over time has been observed due to degradation of Cas9 and gRNA. Thus, exposure to room temperature and repeated cycles of freeze-thawing should be avoided. The crRNAs and tracrRNAs (IDT, Coralville, IA) were complexed into gRNAs by incubating 5 min at 95°C and then 30nM of gRNAs were incubated with Cas9 nuclease (NEB, Ipswich, MA) at ~30nM, 1x NEB Cas9 reaction buffer, and water in a volume of 23-27 µL at 25°C for 10 min. Then, 10-250ng of DNA was added for a final volume of 30 µL. The reaction was incubated overnight at 37°C and then heat shocked at 70°C for 10 min to inactivate the enzyme.

Size Selection

Size selection for the predetermined fragment length is critical for target enrichment prior to library preparation. AMPure XP Beads (Beckman Coulter, Brea, CA, USA) were used to remove off-target, un-digested high molecular weight DNA. After heat inactivation, the reaction was combined with a 0.5x ratio of beads, briefly mixed, and then incubated for 3 min to allow the high molecular weight DNA to bind. The beads were then separated from the solution with a magnet and the solution containing the targeted DNA fragment length was transferred into a new

tube. This was followed by a standard AMPure 1.8x ratio bead purification eluted into 50 µL of TE Low to exchange the buffer and remove small DNA contaminants.

A-tailing and ligation

The fragmented DNA was A-tailed and ligated using the NEBNext Ultra II DNA Library Prep Kit (NEB, Ipswich, MA) according to manufacturer's protocol. The NEB end-repair and A-tailing (ERAT) reaction was incubated at 20°C for 30 min and 65°C for 30 min. Note that end-repair is not needed for CRISPR-DS because Cas9 produces blunt ends, but the ERAT reaction was used for convenient A-tailing. The NEB ligation master mix and 2.5µl of DS adapters at 15 µM were added and incubated at 20°C for 15 min according to the manufacturer's instructions. Instead of relying on in-house manufactured adapters using previously published protocols (Kennedy et al. 2014; Schmitt et al. 2012), which tend to exhibit substantial batch-to-batch variability, we used a commercial adapter prototype of the structure shown in Figure 1c that were synthesized externally through arrangement with TwinStrand Biosciences. The two differences from the previous adapters are: (1) 10bp random double stranded molecular tag instead of 12bp and (2) substitution of the previous 3' 5bp conserved sequence by a simple 3'-dT overhang to ligate onto the 5'-dA-tailed DNA molecules. Upon ligation, the DNA was cleaned by a 0.8X ratio AMPure Bead purification and eluted into 23 µL of nuclease free water.

PCR

The ligated DNA was amplified using KAPA Real-Time Amplification kit with fluorescent standards (KAPA Biosystems, Woburn, MA, USA). 50µl reactions were prepared including KAPA HiFi HotStart Real-time PCR Master Mix, 23µl of previously ligated and purified DNA and DS primers MWS13, 5'- AATGATACGGCGACCACCGAG-3', and MWS20, 5'-GTGACTGGAGTTCAGACGTGTGC-3' (Kennedy et al. 2014; Schmitt et al. 2012) at a final

concentration of 2 μM. The reactions were denatured at 98°C for 45 sec and amplified with 6-8 cycles of 98°C for 15 sec, 65°C for 30 sec, and 72°C for 30 sec, followed by final extension at 72°C for 1 min. Samples were amplified until they reached Fluorescent Standard 3, which typically takes 6-8 cycles depending on the amount of DNA input. Reaching Fluorescent Standard 3 produces a sufficient and standardized number of DNA copies into capture across samples and prevents over-amplification. A 0.8X ratio AMPure Bead wash was performed to purify the amplified fragment and eluted into 40μL of nuclease free water.

Capture and post-capture PCR

TP53 xGen Lockdown Probes (IDT, Coralville, IA) were used to perform hybridization capture for *TP53* exons as previously reported with minor modifications (Krimmel et al. 2016). From the pre-designed IDT *TP53* Lockdown probes, we selected 21 probes that cover the entire *TP53* coding region (exon 1 and part of exon 11 are not coding) (Supplementary Table 2). Each CRISPR/Cas9 excised fragment was covered by at least 2 probes and a maximum of 5 probes (Supplementary Data 1). To produce the capture probe pool, each of the probes for a given fragment was pooled in equimolar amounts, producing 7 different pools, one for each fragment. The pools were mixed again in equimolar amounts, except for the pools for exon 7 and exons 8-9, which were represented at 40% and 90%, respectively. The decrease of capture probes for those exons was implemented after observing consistent overrepresentation of these exons at sequencing. The final capture pool was diluted to 0.75 pmol/μl. Of note, it is essential to dilute the capture pool in low TE (0.1 mM EDTA) and to aliquot it in small volumes suitable for 2-3 uses. Excessive rounds of freeze-thaw severely impact the efficiency of the protocol. Hybridization capture was performed according to the IDT protocol, except for 3 modifications. First, we used blockers MWS60, 5'-AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC

TIIIIIIIIIIIITGACT-3’ and MSW61, 5’-GTCAIIIIIIIIIIIIAGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-3’, which are specific to DS adapters. Second, we used 75µl of Dynabeads M-270 Streptavidin beads instead of 100µl. Third, the post-capture PCR was performed with the KAPA Hi-Fi HotStart PCR kit (KAPA Biosystems, Woburn, MA, USA) using MWS13 and indexed primer MWS21 at a final concentration of 0.8 µM. The reaction was denatured at 98°C for 45 sec and then amplified for 20 cycles at 98°C for 30 sec, 60°C for 45 sec, and 72°C for 45 sec, followed by extension at 72°C for 60 sec. The PCR product was purified with a 0.8X AMPure Bead wash.

Sequencing

Samples were quantified using the Qubit dsDNA HS Assay Kit, diluted, and pooled for sequencing. The sample pool was visualized on the Agilent 4200 TapeStation to confirm library quality. The TapeStation electropherogram should show sharp, distinct peaks corresponding to the fragment length of the designed CRISPR/Cas9 cut fragments (Figures 3b-c). This step can also be performed for each sample individually, prior to pooling, to verify the performance of each individual sample. The final pool was quantified using the KAPA Library Quantification kit (KAPA Biosystems, Woburn, MA, USA). The library was sequenced on the MiSeq Illumina platform using a v3 600-cycle kit (Illumina, San Diego, CA, USA), as specified by the manufacturer. For each sample, we allocated ~7-10% of a lane corresponding to ~2 million reads. Each sequencing run was spiked with approximately 1% PhiX control DNA.

Standard-DS experiments

Three amounts of DNA (25ng, 100ng, and 250ng) from normal human bladder sample B9 were sequenced with standard-DS with one round and two rounds of capture to provide direct comparison with CRISPR-DS. Standard-DS was performed as previously described (Kennedy et

al. 2014), with the exception that the KAPA Hyperprep kit (KAPA Biosystems, Woburn, MA, USA) was used for end-repair and ligation and the KAPA Hi-Fi HotStart PCR kit (KAPA Biosystems, Woburn, MA, USA) was used for PCR amplification. Hybridization capture was performed with xGen Lockdown probes that covered *TP53* exons 2-11, the same that were used for CRISPR-DS. Samples were sequenced on ~10% of a HiSeq 2500 Illumina platform to accommodate shorter fragment lengths. Data analysis was perform with the standard-DS analysis pipeline (https://github.com/risqueslab/DuplexSequencingScripts).

CRISPR-DS target enrichment experiments

Two different experiments were performed to characterize CRISPR-DS target enrichment. The first experiment consisted of comparing one vs. two rounds of capture. Three DNA samples were processed for CRISPR-DS and split in half after one hybridization capture. The first half was indexed and sequenced and the second half was subject to an additional round of capture, as required in the original DS protocol. Then the percentage of raw reads on-target (covering *TP53* exons) was compared for one vs. two captures. The second experiment assessed the percentage of raw reads on-target without performing hybridization capture to determine the enrichment produced exclusively by size selecting CRISPR excised fragments. Fold-enrichment was calculated as the fraction of on-target raw reads divided over the expected fraction of on-target reads given the size of the target region (bases in the target region/total genome bases). Different DNA amounts (from 10ng to 250ng) of three different samples were processed with the protocol described above until first PCR, that is, prior to hybridization capture. Then the PCR product was indexed and sequenced. The percentage of raw reads on-target was calculated and the fold enrichment was estimated considering the size of the targeted region, which is 3,280bp.

Pre-enrichment for high molecular weight DNA

Selection of high molecular weight DNA improves the performance of degraded DNA in CRISPR-DS. We performed this selection using a BluePippin system (Sage Science, Beverly, MA). Two bladder DNAs with DINs of 6 and 4 were run using a 0.75% gel cassette and high-pass setting to obtain >8kb fragments. Size selection was confirmed by TapeStation (Supplementary Fig. 5a). Then 250ng of DNA before BluePippin and 250ng of DNA after BluePippin were processed in parallel with CRISPR-DS. The percentage of raw reads on-target as well as average DCS depth was quantified and compared (Supplementary Fig. 5b.). Alternative methods for size selection such as AMPure beads might be suitable to perform this enrichment.

Data processing

A custom bioinformatics pipeline was created to automate analysis from raw FASTQ files to text files (Supplementary Fig. 7). The primary modification of this pipeline is performing consensus making prior to alignment rather than after, as previously described for DS analysis (Kennedy et al. 2014; Schmitt et al. 2012). In this pipeline, consensus is executed by custom python and bash scripts. After consensus calling, the resulting processed FASTQ files are aligned to the reference genome of interest, in this case human reference genome v38, using bwa-mem v.0.7.4 (Li and Durbin 2009) with default parameters. Mapped reads are re-aligned with GATK Indel-Realigner and low quality bases are clipped from the ends with GATK Clip-Reads (https://software.broadinstitute.org/gatk/). Because of the expected decrease in read quality in the latest cycles of sequencing, we performed a conservative clipping of 30 bases from the 3' end and another 7 bases from 5' end were clipped to avoid the occasional extra overhang left by incorrectly synthesized adapters. In addition, overlapping areas of read-pairs, which in our *TP53* design spanned ~80bp, are trimmed back using fgbio ClipOverlappingReads

(https://github.com/fulcrumgenomics/fgbio). Software for CRISPR-DS is available at https://github.com/risqueslab/CRISPR-DS.

Data analysis

Recovery rate (also called fractional genome-equivalent recovery) was calculated as average DCS depth (sequenced genomes) divided by number of input genomes (1ng of human genomic DNA corresponds to ~330 haploid genomes). The number of on-target raw reads was calculated by counting the number of reads within 100bp window on either side of the CRISPR/Cas9 cut sites. Optimal fragment size (Figures 4b-c and Supplementary Fig. 3) was calculated as the sequencing read length minus the barcode sequence and minus clipped off bases for poor quality at the ends of reads. For peritoneal fluid samples sequenced with both CRISPR-DS and standard-DS, *TP53* biological background mutation frequency was calculated as the number of *TP53* mutations in *TP53* exons 4 to 10 (excluding the tumor mutation) divided by the total number of nucleotides sequenced in those exons. The 95% confidence intervals were calculated in R using the Clopper-Pearson 'exact' method for binomial distributions.

Software Availability
Software for CRISPR-DS data analysis is available at https://github.com/risqueslab/CRISPR-DS.

Data Access
Sequencing data that supports the findings of this study have been deposited in the Sequence Read Archive (BioProject ID: PRJNA412416).

Competing Interest Statement
SRK is a consultant and equity holder for TwinStrand Biosciences Inc. JJS is a founder and equity holder in TwinStrand Biosciences Inc. RAR is the principal investigator on a NIH SBIR R44CA221426 subcontract research agreement with TwinStrand Biosciences Inc.

Acknowledgements
Research reported in this publication was supported by grants from the NIH under award numbers R01CA160674 and R01CA181308 to RAR; Mary Kay Foundation grant 045-15 to RAR; and Rivkin Center for Ovarian Cancer grant 567612 to RAR. Cooperative Agreement Number W911NF-15-2-0127 from the Department of Defense Army Research Office/Defense Forensic Science Center(DFSC), as well as grant W81XWH-16-1-0579 from the Department of Defense Congressionally Directed Medical Research Program to SRK. The views and conclusions

contained in this document are those of the authors and should not be interpreted as representing the official policies, either expressed or implied, of the Army Research Office, DFSC, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation hereon.

We thank Shilpa Kumar for assistance with computational analysis, Emily Kohlbrenner for technical support and helpful discussions, Penny Faires for critical reading and copy editing of the manuscript, and the Genitourinary Cancer Specimen Biorepository for providing access to bladder cases (Director Dr. Colm Morrissey, PhD). We thank the University of Washington Gynecologic Oncology Tissue Bank for providing peritoneal fluid DNA and the Brigham and Women's Hospital/Harvard Cohorts Biorepository for sending archived samples from the Nurses' Health Study for pilot testing.

Authors' Contributions

S.R.K. conceived the idea; D.N., S.R.K., and R.A.R. developed the method; D.N. and R.A.R. designed the experiments; D.N., S.L., E.K.S., M.J.H., K.B., K.L.S., B.F.K., R.A.R, and S.R.K. carried out experiments and/or performed data analysis; M.T. provided samples and scientific input; Y.Z., and J.S. contributed to assay development and provided invaluable critical discussion; D.N., S.R.K., and R.A.R. wrote the paper.

References

Ahn EH, Lee SH, Kim JY, Chang CC, Loeb LA. 2016. Decreased mitochondrial mutagenesis during transformation of human breast stem cells into tumorigenic cells. *Cancer Res* **76**: 4569–4578.

Arbeithuber B, Makova KD, Tiemann-Boege I. 2016. Artifactual mutations resulting from DNA lesions limit detection levels in ultrasensitive sequencing applications. *DNA Res* **23**: 547–559.

Bennett-Baker PE, Mueller JL. 2017. CRISPR-mediated isolation of specific megabase segments of genomic DNA. *Nucleic Acids Res* **45**.

Dabney J, Meyer M. 2012. Length and GC-biases during sequencing library amplification: A comparison of various polymerase-buffer systems with ancient and modern DNA

sequencing libraries. *Biotechniques* **52**.

Goodwin S, McPherson JD, McCombie WR. 2016. Coming of age: ten years of next-generation sequencing technologies. *Nat Rev Genet* **17**: 333–351. https://www.ncbi.nlm.nih.gov/pubmed/27184599.

Hoekstra JG, Hipp MJ, Montine TJ, Kennedy SR. 2016. Mitochondrial DNA mutations increase in early stage Alzheimer disease and are inconsistent with oxidative damage. *Ann Neurol* **80**: 301–306.

Hsu PD, Scott DA, Weinstein JA, Ran FA, Konermann S, Agarwala V, Li Y, Fine EJ, Wu X, Shalem O, et al. 2013. DNA targeting specificity of RNA-guided Cas9 nucleases. *Nat Biotechnol* **31**: 827–832.

Jung H, Ji S, Song S, Park Y, Yang J SE. 2014. The DNA Integrity Number (DIN) provided by the genomic DNA ScreenTape assay allows for streamlining of NGS on FFPE tissue samples. *Appl Note Nucleic Acid Anal*.

Kebschull JM, Zador AM. 2014. Sources of PCR-induced distortions in high-throughput sequencing datasets Sources of PCR-induced distortions in high-throughput sequencing datasets. *bioRxiv* 0–19.

Kennedy SR, Salk JJ, Schmitt MW, Loeb LA. 2013. Ultra-Sensitive Sequencing Reveals an Age-Related Increase in Somatic Mitochondrial Mutations That Are Inconsistent with Oxidative Damage. *PLoS Genet* **9**.

Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH, Prindle MJ, Kuong KJ, Shen JC, Risques RA, et al. 2014. Detecting ultralow-frequency mutations by Duplex Sequencing. *Nat Protoc* **9**: 2586–2606.

Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B. 2011. Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci* **108**: 9530–9535.

http://www.pnas.org/cgi/doi/10.1073/pnas.1105422108.

Knierim E, Lucke B, Schwarz JM, Schuelke M, Seelow D. 2011. Systematic comparison of three methods for fragmentation of long-range PCR products for next generation sequencing. *PLoS One* **6**.

Kozarewa I, Armisen J, Gardner AF, Slatko BE, Hendrickson CL. 2015. Overview of target enrichment strategies. *Curr Protoc Mol Biol* **2015**.

Krimmel JD, Schmitt MW, Harrell MI, Agnew KJ, Kennedy SR, Emond MJ, Loeb LA, Swisher EM, Risques RA. 2016. Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic *TP53* mutations in noncancerous tissues. *Proc Natl Acad Sci* **113**: 6005–6010. http://www.pnas.org/lookup/doi/10.1073/pnas.1601311113.

Li H. 2011. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**: 2987–2993.

Li H, Durbin R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**: 1754–1760.

Park G, Park JK, Shin SH, Jeon HJ, Kim NKD, Kim YJ, Shin HT, Lee E, Lee KH, Son DS, et al. 2017. Characterization of background noise in capture-based targeted sequencing data. *Genome Biol* **18**.

Ran FA, Hsu PD, Wright J, Agarwala V, Scott DA, Zhang F. 2013. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc* **8**: 2281–2308.

Reid-Bayliss KS, Arron ST, Loeb LA, Bezrookove V, Cleaver JE. 2016. Why Cockayne syndrome patients do not get cancer despite their DNA repair deficiency. *Proc Natl Acad Sci* **113**: 10151–10156. http://www.pnas.org/lookup/doi/10.1073/pnas.1610020113.

Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G, Mesirov JP.

2011. Integrative genomics viewer. *Nat Biotechnol* **29**: 24–26.

Salk JJ, Schmitt MW, Loeb LA. 2018. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. *Nat Rev Genet* **19**: 269–285.

Samorodnitsky E, Jewell BM, Hagopian R, Miya J, Wing MR, Lyon E, Damodaran S, Bhatt D, Reeser JW, Datta J, et al. 2015. Evaluation of Hybridization Capture Versus Amplicon-Based Methods for Whole-Exome Sequencing. *Hum Mutat* **36**: 903–914.

Schmitt MW, Fox EJ, Prindle MJ, Reid-Bayliss KS, True LD, Radich JP, Loeb LA. 2015. Sequencing small genomic targets with high efficiency and extreme accuracy. *Nat Methods* **12**: 423–425.

Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. 2012. Detection of ultra-rare mutations by next-generation sequencing. *Proc Natl Acad Sci* **109**: 14508–14513. http://www.pnas.org/cgi/doi/10.1073/pnas.1208715109.

Shin G, Grimes SM, Lee H, Lau BT, Xia LC, Ji HP. 2017. CRISPR-Cas9-targeted fragmentation and selective sequencing enable massively parallel microsatellite analysis. *Nat Commun* **8**.

Winters M, Monroe C, Barta JL, Kemp BM. 2017. Are we fishing or catching? Evaluating the efficiency of bait capture of CODIS fragments. *Forensic Sci Int Genet* **29**: 61–70.

Tables

Table 1: Target enrichment due to size selection

Sample	DNA Input (ng)	Reads On Target Pre-Capture (%)	Fold Enrichment
B9	25	0.76%	7,527
	200	0.25%	2,452
	250	0.21%	2,037
PF1	10	2.85%	28,139
	25	1.99%	19,583
	100	0.68%	6,667
	250	0.70%	6,878
PF5	10	5.05%	49,794
	25	0.96%	9,456
	100	0.34%	3,321
	250	0.22%	2,217

Table 2. Comparison of Standard-DS vs. CRISPR-DS for four different samples with *TP53* mutations

Method	Sample	Input DNA (ng)	Raw Reads On Target	Median Final Depth*	Recovery (%)	Tumor Mutation	Mutant Allele Fraction
Standard-DS	PF1	9,196	92.4%	2742	0.09%	chr17:g.7578275G>A	68.5%
	PF2	3,000	92.8%	5381	0.54%	chr17:g.7577548C>T	1.2%
	PF3	10,186	95.9%	1866	0.06%	chr17:g.7578403C>T	1.6%
	PF4	7,436	95.4%	2029	0.08%	chr17:g.7578526C>T	0.6%
CRISPR-DS	PF1	100	76.6%	2039	6.18%	chr17:g.7578275G>A	68.4%
	PF2	100	94.3%	2831	8.58%	chr17:g.7577548C>T	1.0%
	PF3	100	87.6%	3801	11.52%	chr17:g.7578403C>T	0.4%
	PF4	100	96.5%	2194	6.65%	chr17:g.7578526C>T	0.1%

*After final Duplex Sequencing data processing is performed

Figures

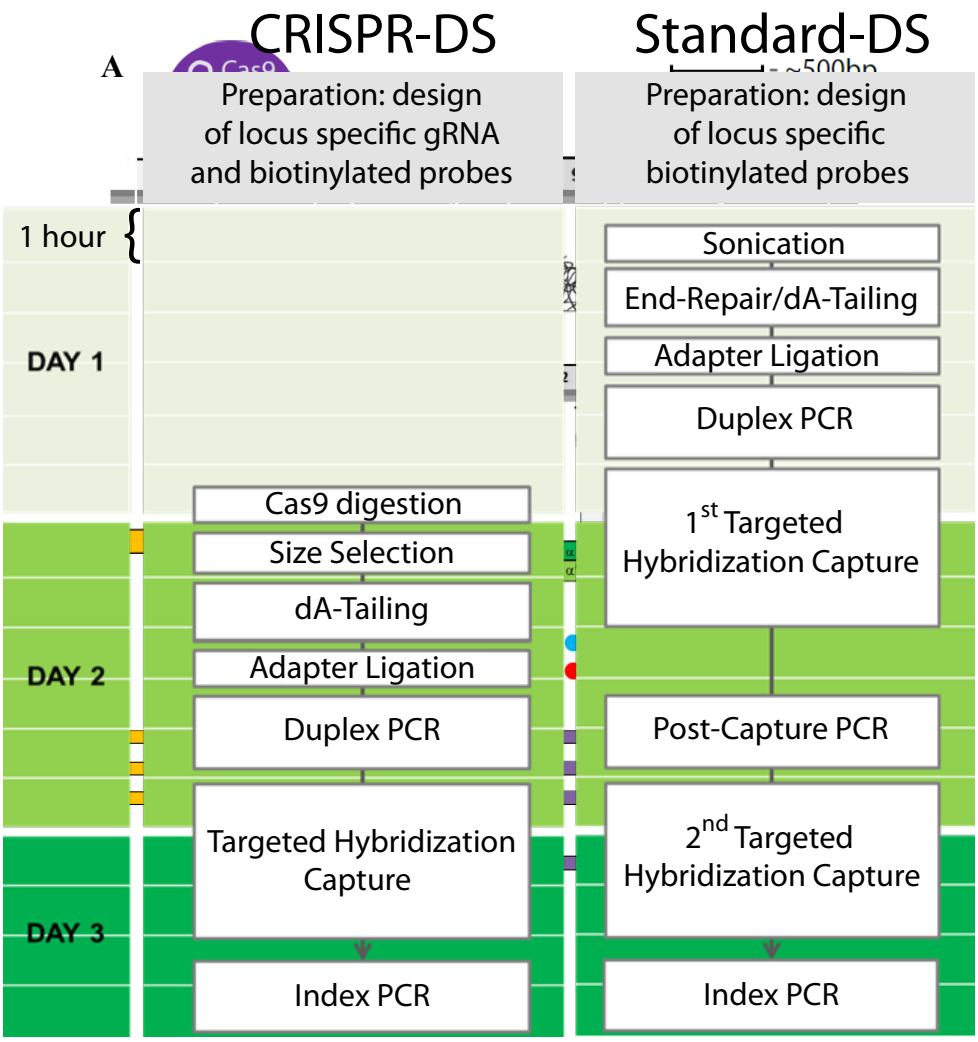


Figure 1. Schematic representation of key aspects of CRISPR-DS (A) CRISPR/Cas9 digestion of *TP53*. Seven fragments containing all *TP53* coding exons were excised via targeted cutting using gRNAs. Dark grey represents reference strand and light grey represents the anti-reference strand. (B) Size selection using 0.5x SPRI beads. Uncut, genomic DNA binds to the beads and allows the recovery of the homogenously sized excised fragments in solution. (C) Double-stranded DNA molecule fragmented and ligated with double-stranded DS-adapters. Adapters contain 10-bp of random, complementary nucleotides and a 3'-dT overhang. (D) Error correction by DS. After creating Single-Strand Consensus Sequence (SSCS) reads, SSCS reads derived from same original DNA molecule are compared with one another to create a Double-Strand Consensus Sequence (DCS). Only mutations found in both SSCS reads are counted as true mutations in DCS reads.

Figure 2. Comparison of library preparation protocols for standard-DS vs. CRISPR-DS The primary differences between the CRISPR-DS and standard-DS library preparation are the fragmentation methods and the number of hybridization capture steps. Instead of fragmentation by sonication as performed in standard-DS, CRISPR-DS relies on an *in vitro* excision of target regions by CRISPR/Cas9 followed by size selection for the excised fragments. While this method requires additional preparation to design locus specific gRNAs, this is a one step process that then reduces the protocol by nearly a day. The reduction is achieved by the elimination of the second round of hybridization capture, which is required for sufficient target enrichment in the standard-DS protocol but not in CRISPR-DS. Colored boxes represent 1h of time.

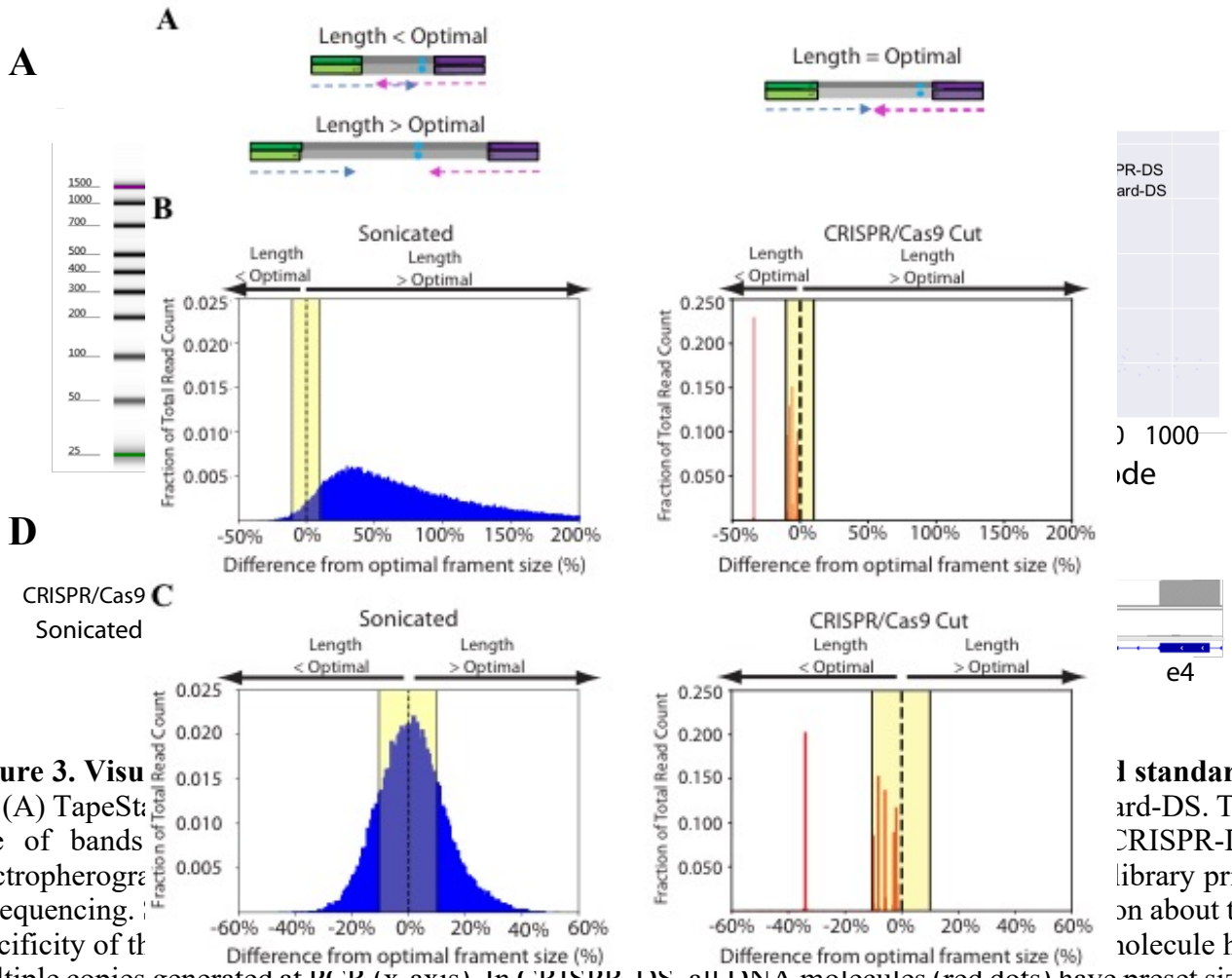


Figure 3. Visu **l standard-DS** (A) TapeSt ard-DS. The size of bands CRISPR-DS electropherogra library prior to sequencing. on about the specificity of th olecule has multiple copies generated at PCR (x-axis). In CRISPR-DS, all DNA molecules (red dots) have preset sizes (y-axis) and generate similar number of PCR copies. In standard-DS, sonication shears DNA into variable fragment lengths (blue dots). Smaller fragments amplify better and generate an excess of copies that waste sequencing resources. (D) Integrative Genomics Viewer of *TP53* coverage with DCS reads generated by CRISPR-DS and standard-DS. CRISPR-DS shows distinct boundaries that correspond to the CRISPR/Cas9 cutting points and an even distribution of depth across positions, both within a fragment and between fragments. Standard-DS shows the typical 'peak' pattern generated by random shearing of fragments and hybridization capture, which leads to variable coverage.

Figure 4. CRISPR/Cas9 fragmentation produces optimal fragment lengths (A) Sonication produces fragments that are either too short or too long, corresponding to redundant or lost information, respectively. CRISPR-DS produces optimally sized fragments, which are perfectly covered by the sequencing reads. (B-C) Comparison of histograms of the insert sizes of two samples prepared with standard-DS (*blue*, left panels), which uses sonication for fragmentation, and CRISPR-DS (*red*, right panels), which uses CRISPR/Cas9 digestion for fragmentation. The x-axis represents the percent difference from the optimally sized fragment, e.g. fragment size that matches the sequencing read length after adjustments for molecular barcodes and clipping. Yellow shading highlights range of fragment sizes, which are within 10% difference from optimal size.

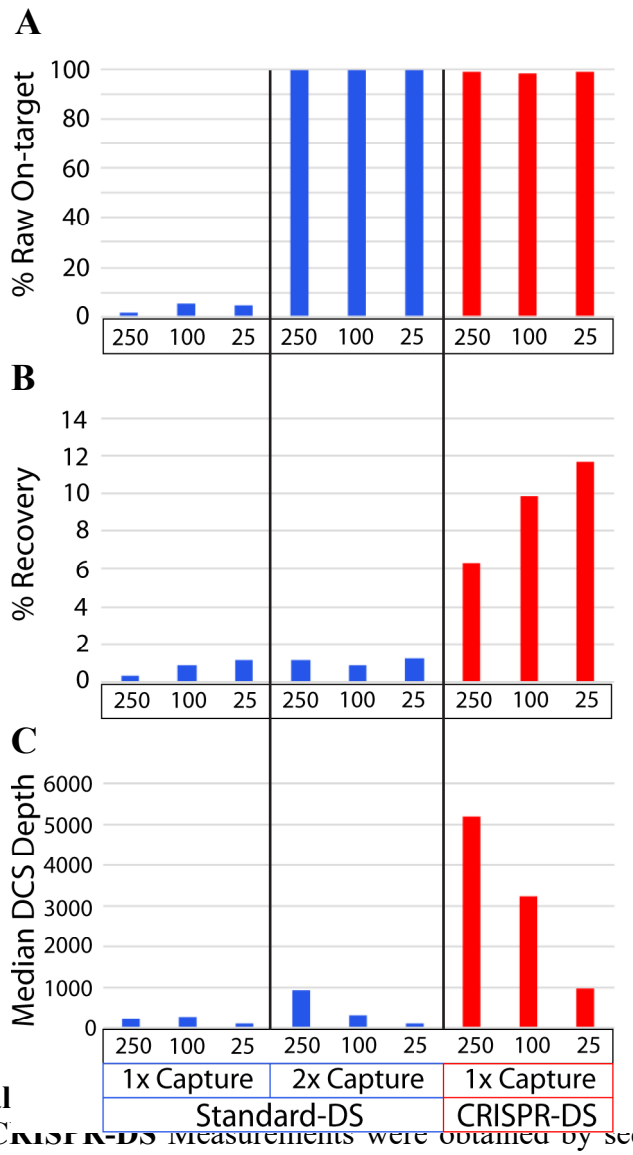


Figure 5. Technical **NA sequenced with both standard-DS and CRISPR-DS** Measurements were obtained by sequencing samples prepared with standard-DS (*blue*) using one and two rounds of hybridization capture and CRISPR-DS (*red*) with only one round of hybridization capture. (A) The percentage of raw sequencing reads on-target (covering *TP53*) post-capture(s) was comparable between Standard-DS with two rounds of capture and CRISPR-DS with one round of capture, demonstrating the target enrichment efficiency of the novel method. (B) Percentage recovery was calculated as the percentage of genomes in input DNA that produced DCS reads. CRISPR-DS increases recovery thanks to the initial CRISPR-based target enrichment, which eliminates one round of hybridization capture. (C) After creating DCS reads, the median DCS depth across all targeted regions was calculated for each input amount. The increased recovery enabled by CRISPR-DS translates into 5-10 times more sequencing depth for the same input DNA.

Supplementary Figure Legends

Figure S1. Timeline of library preparation for CRISPR-DS and standard-DS CRISPR-DS reduces the time required for library preparation by nearly an entire day, with just a short set-up for the CRISPR/Cas9 digestion on Day 1. ERAT: End-Repair and A-Tailing; BW; bead wash.

Figure S2. Homopolymer region produces suboptimal sequencing near TP53 exon 7 The IGV plot shows suboptimal sequencing quality in standard-DS reads that contain the Poly-T repeat. To avoid this problem, the CRISPR-DS gRNA had to be placed in the small region between the Poly-T repeat and the beginning of exon 7, constraining the size of that fragment.

Figure S3. Fraction of reads within 10% of optimal insert size: CRISPR-DS vs. standard-DS The fraction of reads within 10% of optimal insert size corresponds to the yellow highlighted regions in Fig. 4b, c. Markers with matching colors between the two fragmentation methods represent the same sample. This comparison includes seven samples in the study that were analyzed with both methods: 4 peritoneal fluid DNAs (Table 2) and one normal bladder control DNA processed with 250ng, 100ng, and 25ng input DNA (Figure 5).

Figure S4. Target enrichment for CRISPR-DS with one vs. two captures Three de-identified blood DNA samples were processed for CRISPR-DS and split in half after one hybridization capture. The first half was indexed and sequenced and the second half was subject to an additional round of capture, as required in the original DS protocol. After one capture, all three samples had nearly ~90% of raw reads on-target, indicating successful enrichment of the target region. Performing an additional capture slightly increased enrichment to 99%. However, this minimal gain is unnecessary as 90% of raw reads on-target is sufficient enrichment for

successful analysis. Note that raw reads are converted to DCS reads, which are typically >98% on-target.

Figure S5. Pre-enrichment for high molecular weight (MW) DNA with BluePippin Two samples with degraded DNA (B14, DIN=6; B16, DIN=4) were run on a BluePippin 0.75% gel cassette using high-pass setting to obtain >8kb fragments. (a) Genomic TapeStation demonstrating successful removal of lower MW DNA products by BluePippin. (b) Comparison of percentage of on-target raw reads and DCS depth for the same DNA sequenced before and after BluePippin pre-enrichment for high MW DNA. Pre-enrichment resulted in ~2-fold increase in percentage of on-target raw reads and about a 5-fold increase on average DCS depth.

Figure S6. Comparison of mutant allele fraction (MAF) detected by CRISPR-DS and standard-DS Two samples with known MAF for 6 different variants were mixed at 1:1, 1:10, 1:100 and 1:1000. The resulting mixtures were sequenced with both CRISPR-DS and standard-DS. (a) The expected MAF as well as the resulting MAF by each technique, for each variant, are reported in this table. ND: not detected. Note that the mutation at frequency 0.0001 was not expected to be detected with either method because the mean depth of sequencing was ~2,000x. (b-c) Expected MAF of mixture samples vs. MAF from sequencing samples with CRISPR-DS (red) and standard-DS (blue), with least-squares regression line shown. R2 = 0.980 and 0.984 respectively. (d) MAF of mixture samples sequenced with CRISPR-DS on x-axis and MAF by standard-DS on y-axis, least-squares regression line shown, R2 = 0.996.

Figure S7. Comparison of TP53 biological background mutation frequency measured by Standard-DS and CRISPR-DS Four peritoneal fluid samples previously analyzed with

Standard-DS were processed with CRISPR-DS. TP53 biological background mutation was calculated as the number of TP53 mutations not including the tumor mutation divided by the total number of nucleotides sequenced. Error bars correspond to the 95% confidence intervals calculated using the Clopper-Pearson 'exact' method for binomial distributions.

Figure S8. Overview of CRISPR-DS data processing Raw FastQ files from the Illumina platform MiSeq sequencer are used to create Single Stranded Consensus Sequence (SSCS) FastQ files and double-stranded consensus sequence (DCS) FastQ files. DCS FastQ files are then aligned to the reference genome using BWA-mem, and then post-processed using GATK and fgbio software. After pileup and filtering for the region of interest, a custom python script creates a text file ('mutpos' file) with mutant calls.

Figure S9. Control CRISPR/Cas9 digestion of TP53 gRNAs (a) A 500-bp synthetic dsDNA was designed to contain the 12 TP53 gRNA sequences used in the study. Each guide is 23-bp long, and they are separated by 17 bp of spacer sequence. The colored boxes represent the gRNA sequence and the grey boxes represent spacer DNA sequence. (b) List of the 12 TP53 gRNAs used in the study (as in Supplementary Table 1) and their predicted fragment lengths after cutting the synthetic DNA. (c) TapeStation gel image shows distinct bands corresponding to the expected fragment lengths for each of the gRNAs, demonstrating successful cutting.

Supplementary Figures

Figure S1. Timeline of library preparation for CRISPR-DS and standard-DS

	CRISPR-DS	Time Required (min)	Active Time Required (min)	~TOTAL Time required (hrs)	Standard-DS	Time Required (min)	Active Time Required (min)	~TOTAL Time required (hrs)
	Preparation time: design of locus specific gRNA and biotinylated probes				Preparation time: design of locus specific biotinylated probes			
Day 1	Prepare Cas9 Digest	30	30	0.5	Prepare to Sonicate	45	15	5.2
					Sonicate	3	3	
					Prepare ERAT	10	10	
					End repair	60	-	
					Prepare ligation	10	10	
					Ligation	15	-	
					0.8X Ampure BW	20	20	
					Prepare PCR	10	10	
					PCR	20	-	
					0.8X Ampure BW	20	20	
					Prepare Wash Buffers	10	10	
					Prepare Samples	10	10	
					Lyophilize	60	-	
					Prepare to hybridize	20	20	
	Overnight CRISPR-Cas9 Digestion				Overnight Hybridization			
Day 2	0.5X Ampure BW	15	15	5	Heated Washes	120	120	5.2
	1.8X Ampure BW	20	20		Prepare	10	10	
	Prepare ERAT	10	10		PCR	60	-	
	End repair	60	-		0.8X Ampure BW	20	20	
	Prepare ligation	10	10		Prepare Wash Buffers	10	10	
	Ligation	15	-		Prepare Samples	10	10	
	0.8X Ampure BW	20	20		Lyophilize	60	-	
	Prepare PCR	10	10		Prepare to hybridize	20	20	
	PCR	20	-					
	0.8X Ampure BW	20	20					
	Prepare Wash Buffers	10	10					
	Prepare Samples	10	10					
	Lyophilize	60	-					
	Prepare to hybridize	20	20					
	Overnight Hybridization				Overnight Hybridization			
Day 3	Heated Washes	120	120	3.5	Heated Washes	120	120	3.5
	Prepare PCR	10	10		Prepare PCR	10	10	
	PCR	60	-		PCR	60	-	
	0.8X Ampure BW	20	20		0.8X Ampure BW	20	20	
	* Ready to Sequence*				* Ready to Sequence*			

Figure S2. Homopolymer region produces suboptimal sequencing near TP53 exon 7

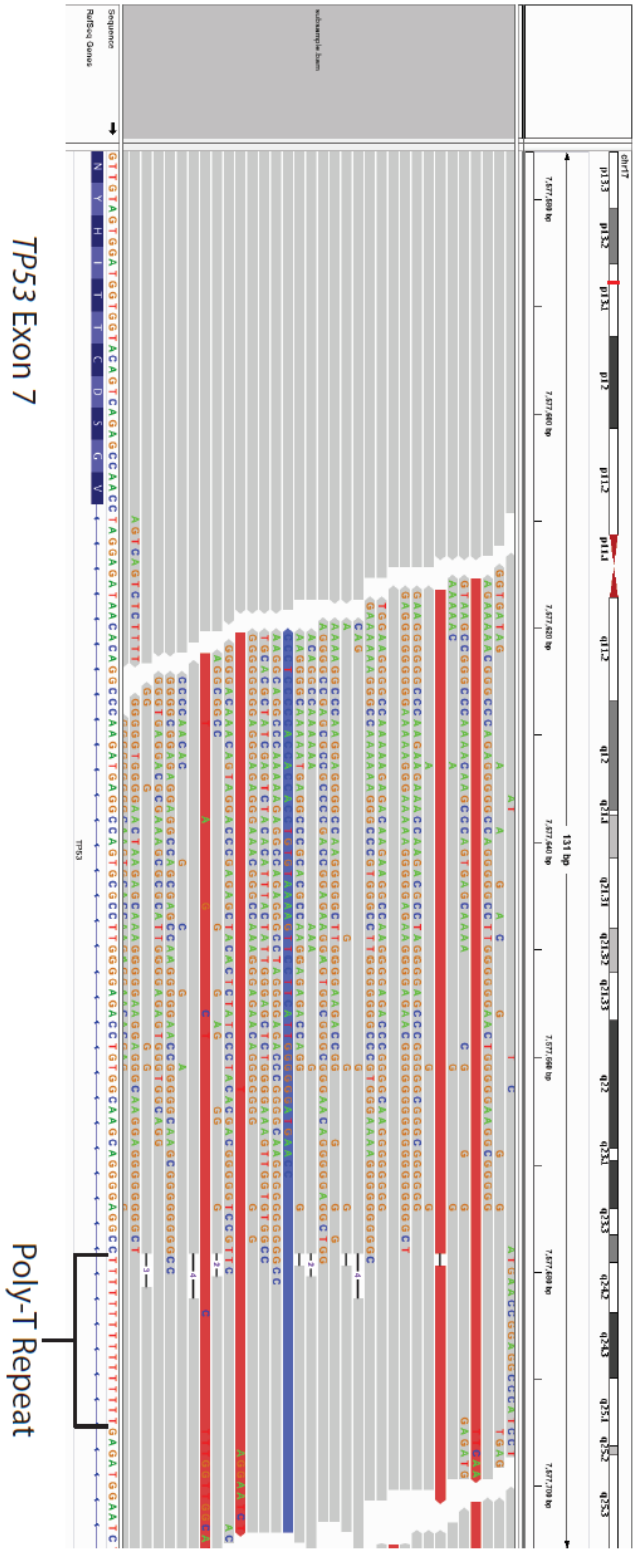


Figure S3. Fraction of reads within 10% of optimal insert size: CRISPR-DS vs. standard-DS

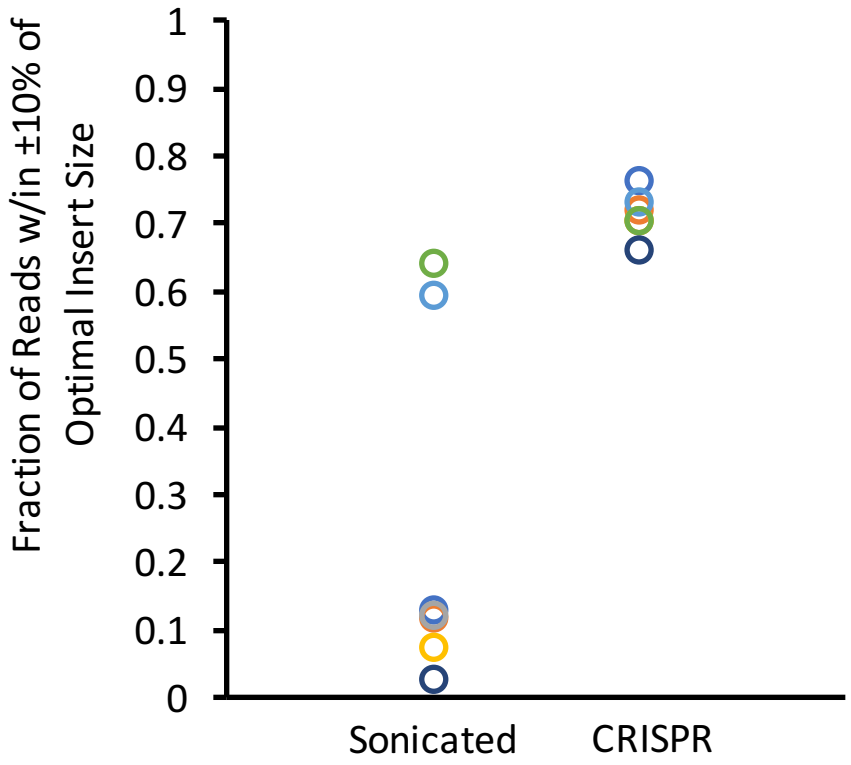


Figure S4. Target enrichment for CRISPR-DS with one vs. two captures

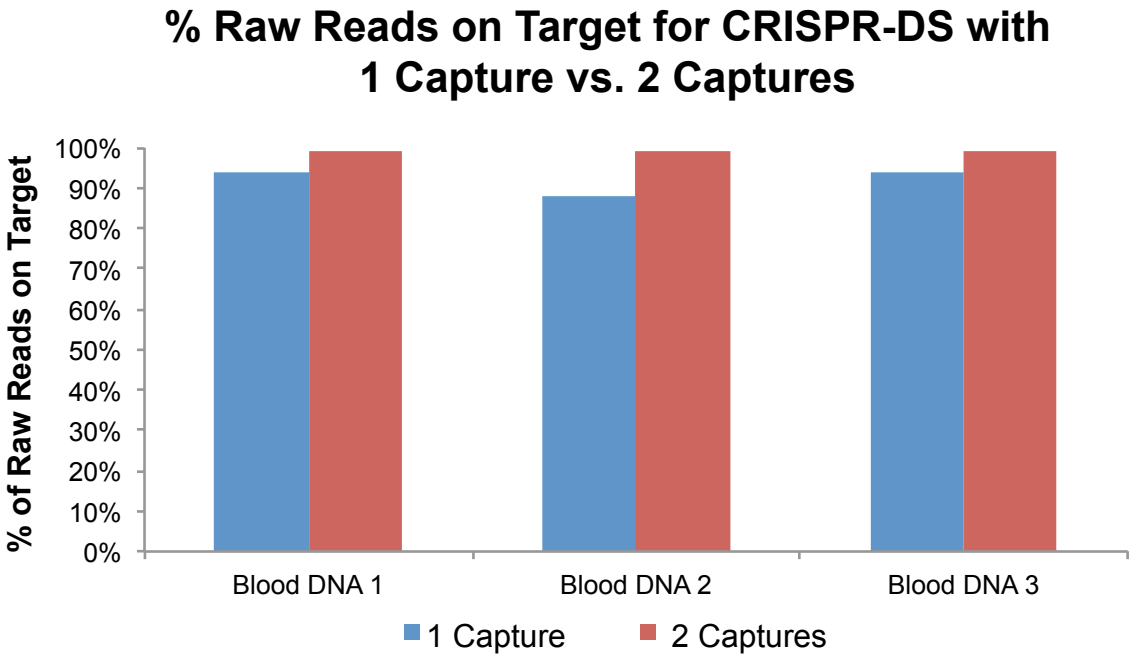


Figure S5. Pre-enrichment for high molecular weight DNA with BluePippin

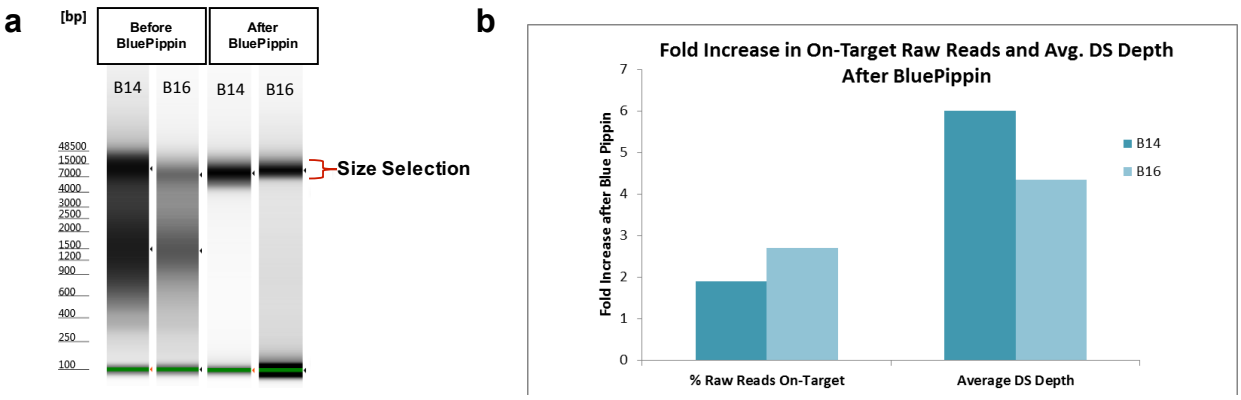


Figure S6. Comparison of mutant allele fraction detected by CRISPR-DS and standard-DS

a

MIXTURE	POSITION	VARIANT	EXPECTED MAF	CRISPR-DS MAF	STANDARD-DS MAF
1:1	7577407	A>C	0.610	0.596	0.607
1:1	7577427	G>A	0.610	0.595	0.626
1:1	7578275	G>A	0.219	0.218	0.199
1:1	7579472	G>C	0.338	0.330	0.349
1:1	7579619	G>T	0.052	0.046	0.047
1:1	7579801	G>C	0.346	0.365	0.342
1:10	7577407	A>C	0.522	0.509	0.528
1:10	7577427	G>A	0.522	0.509	0.566
1:10	7578275	G>A	0.044	0.032	0.027
1:10	7579472	G>C	0.466	0.473	0.472
1:10	7579619	G>T	0.010	0.009	0.008
1:10	7579801	G>C	0.477	0.445	0.438
1:100	7577407	A>C	0.502	0.513	0.492
1:100	7577427	G>A	0.502	0.514	0.476
1:100	7578275	G>A	0.004	0.001	0.002
1:100	7579472	G>C	0.494	0.505	0.488
1:100	7579619	G>T	0.001	0.002	0.001
1:100	7579801	G>C	0.506	0.524	0.463
1:1000	7577407	A>C	0.500	0.486	0.458
1:1000	7577427	G>A	0.501	0.487	0.445
1:1000	7578275	G>A	0.0004	0.0005	ND
1:1000	7579472	G>C	0.497	0.488	0.510
1:1000	7579619	G>T	0.0001	ND	ND
1:1000	7579801	G>C	0.509	0.495	0.489

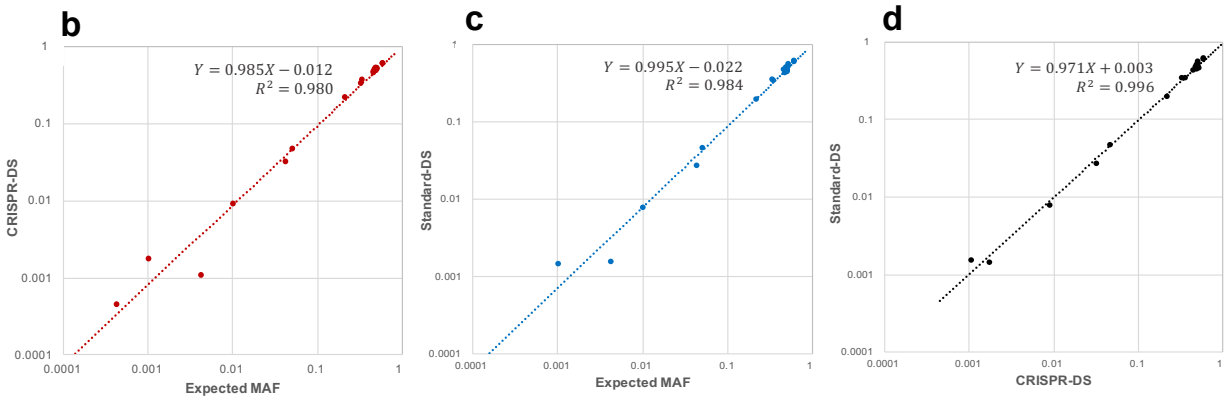


Figure S7. Comparison of TP53 biological background mutation frequency measured by Standard-DS and CRISPR-DS

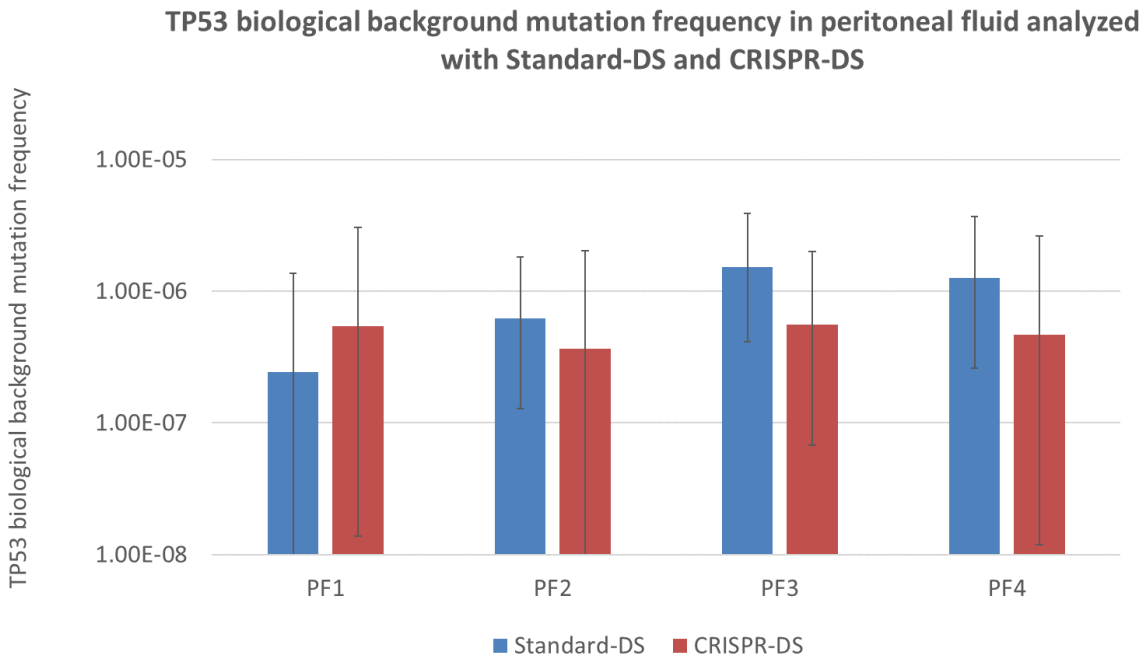


Figure S8. Overview of CRISPR-DS data processing

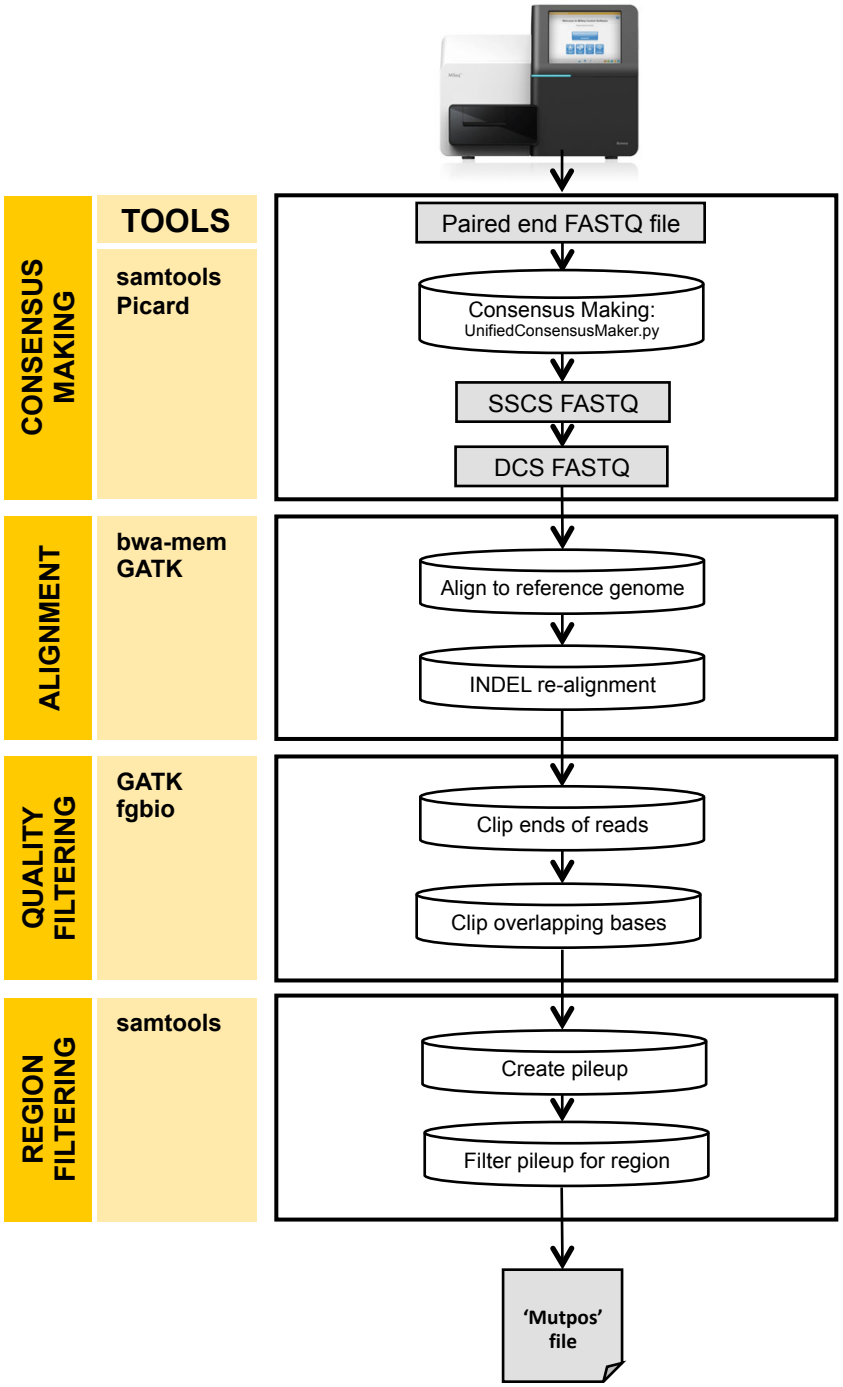
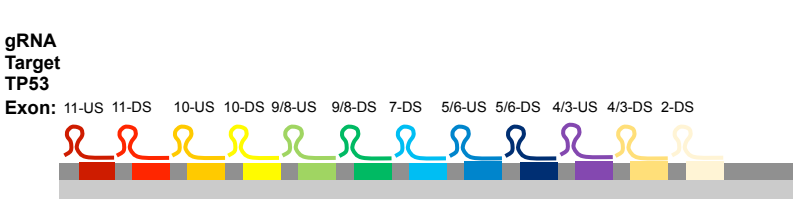


Figure S9. Control CRISPR/Cas9 digestion of TP53 gRNAs

a

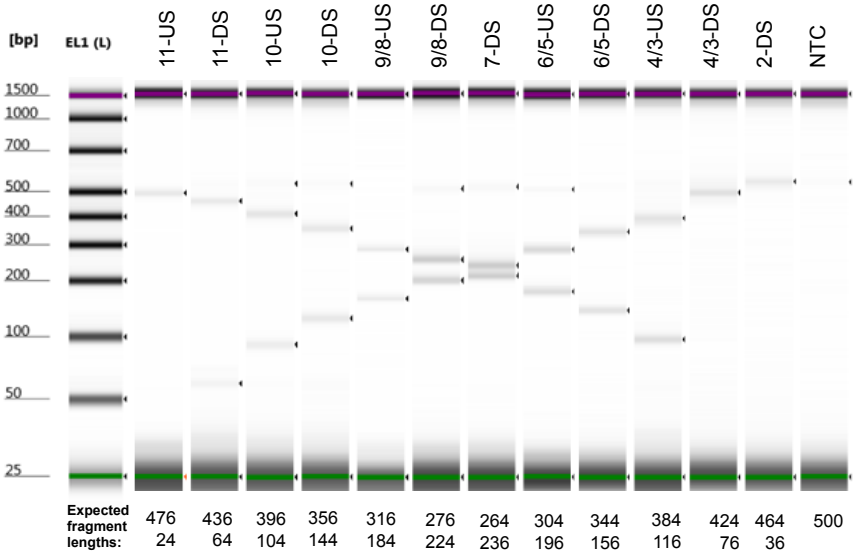


SYNTHETIC DNA WITH TP53 gRNA sequences

b

TP53 guide RNA	TP53 Exon(s) Targeted	Left Fragment (bp)	Right Fragment (bp)
TP5e11_US	11	24	476
TP5e11_DS	11	64	436
TP5e10_US	10	104	396
TP5e10_DS	10	144	356
TP5e9-8_US	9,8	184	316
TP5e9-8_DS	9,8	224	276
TP5e7_DS.v2	7	264	236
TP5e6-5_US	5,6	304	196
TP5e6-5_DS	5,6	344	156
TP5e4-3_US.v2	4,3	384	116
TP5e4-3_DS	4,3	424	76
TP5e2_DS	2	464	36

c



Supplementary Tables

Supplementary Table 1. crRNA sequences for TP53 CRISPR/Cas9 digestion

Target description:	Name:	Sequence plus pam site:	Position start:	Position end:	Zhang score
TP53 - upstream of exon 11	TP53e11_US	GTGGGCCCCTACCTAGAATG**TGG**	7572606	7572628	79
TP53 - downstream of exon 11	TP53e11_DS	ATTCCCGTTGTCCCAGCCTT**AGG**	7573118	7573096	70
TP53 - upstream of exon 10	TP53e10_US	TGGTTATAGGATTCAACCGG**AGG**	7573754	7573776	91
TP53 - downstream of exon 10	TP53e10_DS	CTGATTGCAATCTCCGCCTC**TGG**	7574261	7574283	86
TP53 - upstream of exons 9-8	TP53e9-8_DS	CGGCATTTTGAGTGTTAGAC**TGG**	7576792	7576814	80
TP53 - downstream of exons 9-8	TP53e9-8_US	CTTTGGGACCTCTTAACCTG**TGG**	7577324	7577302	80
TP53 - downstream of exon 7	TP53e7_DS.v2	CAGGTCTCCCCAAGGCGCAC**TGG**	7577660	7577638	81
TP53 - upstream of exons 6-5	TP53e6-5_US	GCACATCTCATGGGGTTATA**GGG**	7578050	7578072	84
TP53 - downstream of exons 6-5	TP53e6-5_DS	CAGGGGAGTACTGTAGGAAG**AGG**	7578545	7578567	61
TP53 - upstream of exons 4-3	TP53e4-3_US.v2	TGCACGGTCAGTTGCCCTGA**GGG**	7579317	7579295	81
TP53 - downstream of exons 4-3	TP53e4-3_DS	ATGGAATTTTCGCTTCCCAC**AGG**	7579751	7579773	79
TP53 - downstream of exon 2	TP53e2_DS	TGGGAATAGGGTGCACATTT**AGG**	7580242	7580220	66

Supplementary Table 2. *TP53* hybridization capture probes

Targeted exon:	IDT probe name:	IDT probe sequence:
Exon11	TP53_e11.A.2	CCCCGGGACAAAGCAAATGGAAGTCCTGGGTGCTTCTGACGCACACCTATTGCAAGCAAGGGTTCAAAGACCCAAAACCCAAAATGGCAGGGGAGGGAGAGATGGGGGTGGGAGGCTGTC
Exon11	TP53_e11.A.1	AGTGGGGAACAAGAAGTGGAGAATGTCAGTCTGAGTCAGGCCCTTCTGTCTTGAACATGAGTTTTTTATGGCGGGAGGTAGACTGACCCTTTTTGGACTTCAGGTGGCTGTAGGAGACAG
Exon10	TP53_e10.1	ACTCACCTGGAGTGAGCCCTGCTCCCCCCTGGCTCCTTCCCAGCCTGGGCATCCTTGAGTTCCAAGGCCTCATTCAGCTCTCGGAACATCTCGAAGCGCTCACGCCCACGGATCTGCAGC
Exon10	TP53_e10.1_1	AATCCTATGGCTTTCCAACCTAGGAAGGCAGGGGAGTAGGGCCAGGAAGGGGCTGAGGTCACTCACCTGGAGTGAGCCCTGCTCCCCCCTGGCTCCTTCCCAGCCTGGGCATCCTTGAGT
Exon10	TP53_e10.1_2	TCCAAGGCCTCATTCAGCTCTCGGAACATCTCGAAGCGCTCACGCCCACGGATCTGCAGCAACAGAGGAGGGGGAGAAGTAAGTATATACACAGTACCTGAGTTAAAAGATGGTTCAAGT
Exon9-8	TP53_e8.1	AAGAGGTCCCAAGACTTAGTACCTGAAGGGTGAAATATTCTCCATCCAGTGGTTTCTTCTTTGGCTGGGGAGAGGAGCTGGTGTTGTTGGGCAGTGCTAGGAAAGAGGCAAGGAAAGGTG
Exon9-8	TP53_e7.2	GCATAACTGCACCCTTGGTCTCCTCCACCGCTTCTTGTCCTGCTTGCTTACCTCGCTTAGTGCTCCCTGGGGGCAGCTCGTGGTGAGGCTCCCCTTTCTTGCGGAGATTCTCTTCCTCTG
Exon9-8	TP53_e7.1	TGCGCCGGTCTCTCCCAGGACAGGCACAAACACGCACCTCAAAGCTGTTCCGTCCCAGTAGATTACCACTACTCAGGATAGGAAAAGAGAAGCAAGAGGCAGTAAGGAAATCAGGTCCTA
Exon7	TP53_e6.1	TGACCTGGAGTCTTCCAGTGTGATGATGGTGAGGATGGGCCTCCGGTTCATGCCGCCCATGCAGGAACTGTTACACATGTAGTTGTAGTGGATGGTGGTACAGTCAGAGCCAACCTAGGA
Exon7	TP53_e6.2	ATGTGATGAGAGGTGGATGGGTAGTAGTATGGAAGAAATCGGTAAGAGGTGGGCCCAGGGGTCAGAGGCAAGCAGAGGCTGGGGCACAGCAGGCCAGTGTGCAGGGTGGCAAGTGGCTCC
Exon6-5	TP53_e5.1	GACCTCAGGCGGCTCATAGGGCACCACCACACTATGTCGAAAAGTGTTTCTGTCATCCAAATACTCCACACGCAAATTTCCTTCCACTCGGATAAGATGCTGAGGAGGGGCCAGACCTAA
Exon6-5	TP53_e5.1_1	CTGGAGGGCCACTGACAACCACCCTTAACCCCTCCTCCCAGAGACCCCAGTTGCAAACCAGACCTCAGGCGGCTCATAGGGCACCACCACACTATGTCGAAAAGTGTTTCTGTCATCCAA
Exon6-5	TP53_e4.3	CACCATCGCTATCTGAGCAGCGCTCATGGTGGGGGCAGCGCCTCACAACCTCCGTCATGTGCTGTGACTGCTTGTAGATGGCCATGGCGCGGACGCGGGTGCCGGGCGGGGGTGTGGAAT
Exon6-5	TP53_e4.3_1	GAATCAGAGGCCTGGGGACCCTGGGCAACCAGCCCTGTCGTCTCTCCAGCCCCAGCTGCTCACCATCGCTATCTGAGCAGCGCTCATGGTGGGGGCAGCGCCTCACAACCTCCGTCATGT
Exon6-5	TP53_e4.2	CAACCCACAGCTGCACAGGGCAGGTCTTGGCCAGTTGGCAAAACATCTTGTTGAGGGCAGGGGAGTACTGTAGGAAGAGGAAGGAGACAGAGTTGAAAGTCAGGGCACAAGTGAACAGAT
Exon4-3	TP53_e3.3	ATTGAAGTCTCATGGAAGCCAGCCCCTCAGGGCAACTGACCGTGCAAGTCACAGACTTGGCTGTCCCAGAATGCAAGAAGCCCAGACGGAAACCGTAGCTGCCCTGGTAGGTTTTCTGGG
Exon4-3	TP53_e3.2	AAGGGACAGAAGATGACAGGGGCCAGGAGGGGGCTGGTGCAGGGGCCGCCGGTGTAGGAGCTGCTGGTGCAGGGGCCACGGGGGGAGCAGCCTCTGGCATTCTGGGAGCTTCATCTGGAC
Exon4-3	TP53_e3.1	CTGGGTCTTCAGTGAACCATTGTTCAATATCGTCCGGGGACAGCATCAAATCATCCATTGCTTGGGACGGCAAGGGGGACTGTAGATGGGTGAAAAGAGCAGTCAGAGGACCAGGTCCTC
Exon4-3	TP53_e2.3	GCCCCCCAGCCCTCCAGGTCCCCAGCCCTCCAGGTCCCCAGCCCAACCCTTGTCCTTACCAGAACGTTGTTTTCAGGAAGTCTGAAAGACAAGAGCAGAAAGTCAGTCCCATGGAATTTT
Exon2	TP53_e2.2	CGCTTCCCACAGGTCTCTGCTAGGGGGCTGGGGTTGGGGTGGGGGTGGTGGGCCTGCCCTTCCAATGGATCCACTCACAGTTTCCATAGGTCTGAAAATGTTTCCTGACTCAGAGGGGGC
Exon2	TP53_e2.1	TCGACGCTAGGATCTGACTGCGGCTCCTCCATGGCAGTGACCCGGAAGGCAGTCTGGCTGCTGCAAGAGGAAAAGTGGGGATCCAGCATGAGACACTTCCAACCCTGGGTCACCTGGGCC

Supplementary Table 3. CRISPR-DS sequencing results for 13 samples processed with 250ng input DNA

Sample ID	DIN	DNA input (ng)	# Raw reads	% of Raw Reads on target	# DCS reads	% of DCS Reads on target	DCS depth	Recovery rate
B1	6.8	250	7751046	44.0%	68906	100.0%	6143.2	7.4%
B2a	6.9	250	4575484	43.0%	37984	99.1%	3386.4	4.1%
B2b	6.9	250	4855458	47.5%	42815	99.1%	3817.1	4.6%
B3	8.2	250	4214290	85.8%	30847	98.8%	2750.1	3.3%
B4a	8.8	250	4200814	84.4%	85822	99.0%	7651.3	9.3%
B4b	8.8	250	4581646	86.6%	84051	99.1%	7493.4	9.1%
B5	8.5	250	3938328	98.4%	101201	98.7%	9022.4	10.9%
B6	8.7	250	4640288	78.0%	69002	98.8%	6151.7	7.5%
B7	7.6	250	4230402	91.2%	60950	98.8%	5433.9	6.6%
B8	7.0	250	3869654	93.6%	38586	98.9%	3440.1	4.2%
B9	8.9	250	4594068	96.6%	75089	99.2%	6694.4	8.1%
B10	8.6	250	5764098	79.0%	61303	99.1%	5465.3	6.6%
B11	8.5	250	5764650	80.9%	71381	99.3%	6363.8	7.7%
B12	7.9	250	5234650	85.9%	40092	99.4%	3574.3	4.3%
B13	7.0	250	3737110	74.0%	71138	99.1%	6284.8	7.6%

Supplementary Data

Data S1. TP53 sequence with crRNA and capture probes

TP53 tumor protein p53 [Homo sapiens (human)]
NC_000017.11 Chromosome 17 Reference GRCh38.p2 Primary Assembly
Positive Strand, exons only (not entire sequence)

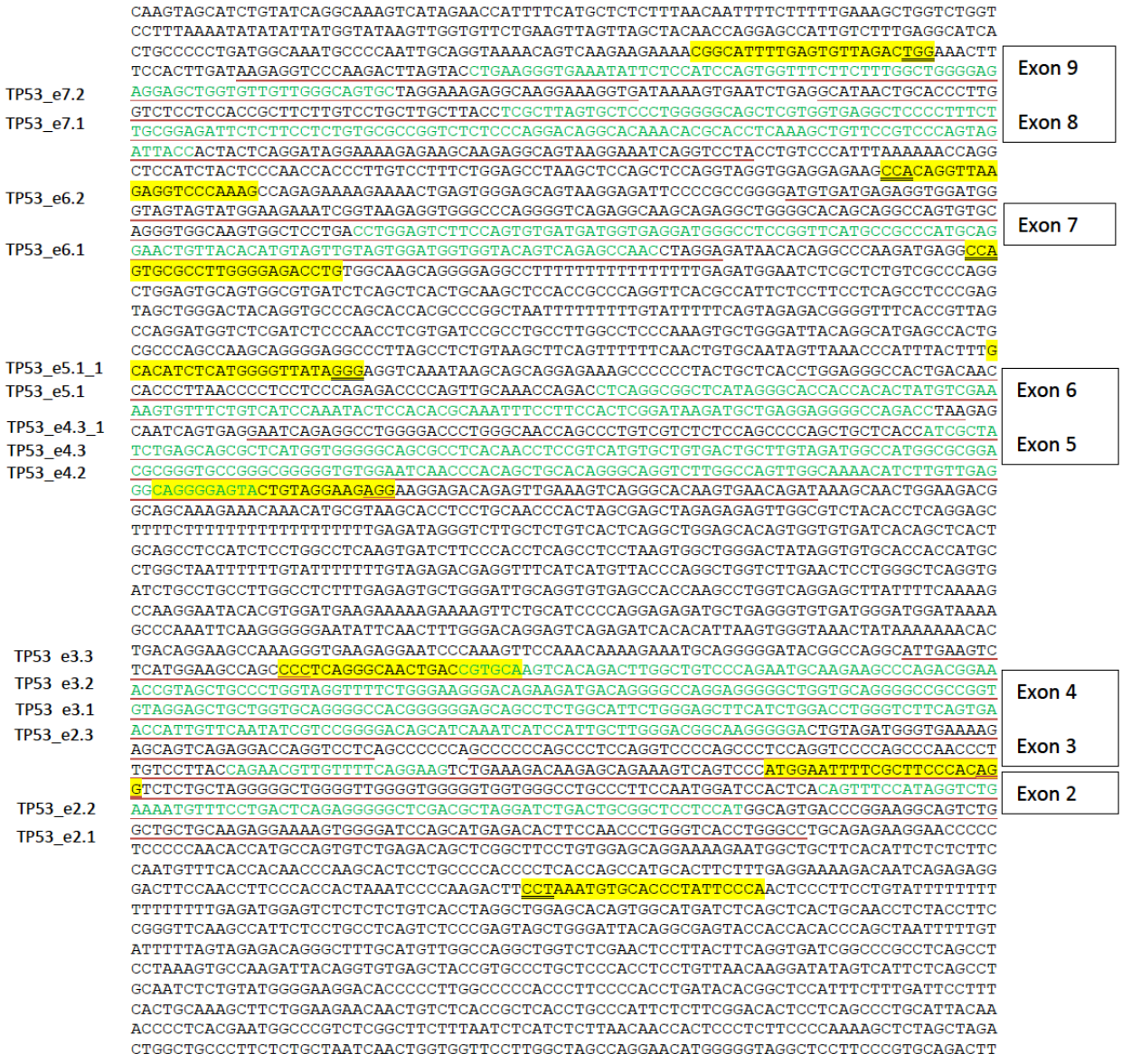
Key:
Green letters: Coding region. Exon names are indicated in right margin and boxed together when they are cut in the same fragment
Yellow highlight: Cas9 cut site, PAM sequence are underlined with double line
Red underline: Biotinylated probes. Probes name are indicated in left margin

…
ATACAAGAGATGAAATCCTCCAGGGTGTGGGATGGGGTGAGATTTCCTTTTAGGTACTAAGGTTCACCAAGAGGTTGTCAGA
CAGGGTTTGGCTGGGCCAGCAGAGACTTGACAACTCCCTCTACCTAACCAGCTGCCCAACTGTAGAAACTACCAACCCACCG
ACCAACAGGGAGAGGGAACAAGCACCCTCAAGGGGGTCAAGTTCTAGACCCCATGTAATAAAAGGTGGTTTCAAGGCCAGAT
GTACATTATTTCATTAACCCTCACAATGCACTCTGTGAGGTAGGTGCAAATGCCAGCATTTCACAGATATGGGCCTTGAAGT
TAGAGAAAATTCAACAGTGAGGGACAGCTTCCCTGGTTAGTACGGTGAAGTGGGCCCCTACCTAGAATGTGGCTGATTGTAA
ACTAACCCTTAACTGCAAGAACATTTCTTACATCTCCCAAACATCCCTCACAGTAAAAACCTTAAAATCTAAGCTGGTATGT
CCTACTCCCCATCCTCCTCCCCACAACAAAACACCAGTGCAGGCCAACTTGTTCAGTGGAGCCCCGGGACAAAGCAAATGGA
TP53_e11.A.2 AGTCCTGGGTGCTTCTGACGCACACCTATTGCAAGCAAGGGTTCAAAGACCCAAAACCCAAAATGGCAGGGGAGGGAGAGAT
TP53_e11.A.1 GGGGGTGGGAGGCTGTCAGTGGGGAACAAGAAGTGGAGAATGTCAGTCTGAGTCAGGCCCTTCTGTCTTGAACATGAGTTTT
TTATGGCGGGAGGTAGACTGACCCTTTTTGGACTTCAGGTGGCTGTAGGAGACAGAAGCAGGGAGGAGAGATGACATCACAT
GAGTGAGAGGGTCTGTGCCCCTTTTCCCTGACCAATGCTTTGAAGGGCCTAAGGCTGGGACAACGGGAATTCAAATCAAGAT
GGTGGCCACACCCCATGCAAATATGTTTACTGAGCACCTCAGAGTATTAGTGTGTATTAGTCTCGTAATCTTCCCTTACCCC
ATTTTACTTTATTTATCTTTTTTGAGACGGAGTTTCACTCTTGTTGCCCAGGCTGGAGTGTAATGGTGAGATCTCAGCTCAC
CGCAACCTCTGCCTCCCGGGTTCAAGCGATTCTCCTGCCTCAGCCTCCCGAGTAGGTAGCTGGGATTACAGGCATGCATCAC
CACGCCCGGCTACTTTTGTATTTTTAGTAGAGATGGGGTTTCTCCATGTTGGTCAGGCTGGGCTCAAACTCCCGACCTCAGG
TGATCCACTCGCCTTGGCCTCCCAGAGTGTGGGATTCGTGAGCCACTGCGCCCGGCCCCCTTACCCCATTTTATATATAAGG
AAACTGAGTTTGACGGGGGTCACCTAGGACCTGCCGGTGCATGGCAGGGCTGAGTATATGACCTGAAACTCTGGCTGTATTC
AGTATTACACAATTATTAGGCCCCTCCTTGAGACCCTCCAGCTCTGGGCTGGGAGTTGCGGAGAATGGCAAAGAAGTATCCA
CACTCGTCCCTGGGTTTGGATGTTCTGTGGATACACTGAGGCAAGAATGTGGTTATAGGATTCAACCGGAGGAAGACTAAAA
AAATGTCTGTGCAGGGCTGGGACCCAATGAGATGGGGTCAGCTGCCTTTGACCATGAAGGCAGGATGAGAATGGAATCCTAT
TP53_e10.1_1 GGCTTTCCAACCTAGGAAGGCAGGGGAGTAGGGCCAGGAAGGGGCTGAGGTCACTCACCTGGAGTGAGCCCTGCTCCCCCCT
TP53_e10.1 GGCTCCTTCCCAGCCTGGGCATCCTTGAGTTCCAAGGCCTCATTCAGCTCTCGGAACATCTCGAAGCGCTCACGCCCACGGA
TP53_e10.1_2 TCTGCAGCAACAGAGGAGGGGGAGAAGTAAGTATATACACAGTACCTGAGTTAAAAGATGGTTCAAGTTACAATTGTTTGAC
TTTATGACGGTACAAAAGCAACATGCATTTAGTAGAAACTGCACTTCAAGTACCTATACAGCTGACTTTTAAAAATATTTAT
TTATTTATTTTGAGATGGGGTCTCACTCTGTTGCCCAGGCGGGAGTGCAATGGTGCAATCTTGGCTGATTGCAATCTCCGCC
TCTGGGGTTCAAGTGATTCTTGTGCCTCAGCCTCCCGAGTAGCTGGGACTACAGGCGTGTGCTACCACACCTGGCTAATTTT
TGTGTTTTTAGTAGAGATGGGGCTTCACCATGTTAGCCAGGCTGGTTTCCAACTCCTGACGTCAGGTGATCTACCCACCTCC
ACCTCCCAAAGTGCTGGGATTACAGGTGTGAGCCACTGTGCCCGGCCCTTTTTTAAATTTTAGAGATGATGTCTTGCTATGT
TGTTCAGGCTGGACTCAAACTCTTGGGCTCAAGAGATCCTCCTGCCTTAGCCTCTCAAGTAACTGGGACTACATGTGCATGC
GACTGTGCCTCGTTTCTTTTCTTTTTTTTCTGAGACGGAGTCTCACTCTATCGCCCAGGCTGGAGTGCAGTGGCGCCATCTT
GGCTCCCTGCAACCTCCGCCTCCTGGTTCAAGCGATTCTCCTGCCTCAGCCTCCCAAGTAGCTGGGATTACAGGCACCTGCC
ATCACGCCCGGTTAATTTTTGTATTTTAGTAGAGACGGGGTTTCACCATGTTGGCTAGGCTGGTCTTGAACTCCTGACCTCA
GGTGATCCACCCGCCTCAGCCTCCCGAAATGCTGGGATTACAGGCGTGAGCCAGTGCGCCTGGCCTTTTCTTTTTTTGAGTC
TCGCTCTGCGCCCAGGCTGTGCCTGGCTCGACTGTGCCTCCTTTCATGCAACCATGCTGTTTCTCACTTTCAGTAACAATAT
TCAATAAATCACATGAGATATACAACATTTTATTACTATAAAAAGGGCTTTGTGTTAGATGACTTTGCCCAACTGTAGGGTA
ACTTAAATGCTCTGAACACGTTTCAAGTAGGCTAGGGCTGAGTGTGGTAGCTCATGCCTGTAACCCCAATACTTGGGGAGGC
TGAGGTGGAAGGATTGATTGAGCCCAGGGGTTTGATACCAGCATGGGCAACGTAGCAAGACCTTGACTTCACAGAAAATAAA
AAATTAGCTGGGTGTCGTGGCATGTGCCTGTAGTCCTAGCTACTTGGGAGGGTGAAATCACCGGAGCCCAGGGAGGTCAAGG
CTGCAGTGAGCTGAGATGGTGCCACTGCACTCTAGCCTGAGTGACAGAGTGAGACTCTGTCTTTAAATAAATAAATAAAAAT
TAGCCGGGCGTGGTGGCTCACACCTGTAATCCCAGCACTTTGGGAGGCCGAGGCGGGCGGATCACATGGTCAGAAGTTCGAG
ACCAGCCTGGCCAACATGGTGAAACCCTGTCTCTACTAAAAATACAAAAATTAGCTGGGCGTGGTAGCAGGCGCTTGTAGTC
CTAGCTATTCGGGAGGCTGAGGCAGGAGAATCACTTGAACCCAGGAGGCAGAGGTTGCAGTGAGCCGAGATCATGCCACTGC
ACTCCAGCCTGGGCGACAGAGTGAGACTGAGTCTCAAAAAAATAAAATAAAATAAAATAAAAATAAATAAATAAAAATTAGC
CAGGCATGGTGGTGCAGGCCTGTAGTTGAAGCAACTTGGGAGGCTGAGCTGGGAGGATGGATGGAGCCTGGGAGGTGGAGGC
TGCAGTGAGCTGTGACTGCACTACTGCACTCTATCCAGCCTGGGTGACAGAGCAAGACCTTGTCTCAAAAAAGTAGGCTAGA
GACCAGCCTGGGCAACATAGTGAGACTCTATCTATCTACAAAAAATTTTAAAAATTAGCTGGGTATGGTGGTGTATGCCTGT
GGTCCTAGCTACTGGGGAGGCAGAGTTAGGGGGATTGCTTGAGCCCAGGAGGGTATAATGAGCTATGATCACATCACTGTAA
TCCAGCCTGGGCAACAGAGCAAGATGCTGTCTCCATTAAAAATAAAATAAAAGTAGGCTAGGCAGGCCGGGTGCGGTGGCTC
ACGCCTGTAATCCCAGCACTTTGGGAGGCCAAGGCAGGCAGATCACAAGGTCAGGAGTTCGAGACTAGCCTGGCCAACATGG
TGAAACCTCATCTCTACTAAAAAAAAAAATAAATAAATAACAAAAAATTAGCTGGGCGTCGGGGCAGGTGCCTGTAATCCCA
GCTACTCAGTGGGCTGAGGCAGGAGAATCGCTTGAACCCAGAAGGCGGAGGTTGCAGTGAGCCGAGATCCCGCCACTGCACT
CCAGCCTGGGTGACAGAGTGAGACTCTGTCTCCAAAAAAAAAAAAAAAAAAAGCAGGCTAGGCTAAGCTATGATGTTCCTTA
GATTAGGTGTATTAAATCCATTTTCAACTTACAATATTTTCAACTTACGACGAGTTTATCAGGAAGTAACACCATCGTAAGT

Exon 11

Exon 10

TP53 tumor protein p53 [Homo sapiens (human)]
NC_000017.11 Chromosome 17 Reference GRCh38.p2 Primary Assembly
Positive Strand, exons only (not entire sequence)



Data S2. GeneBlock sequence

Geneblock fragment- 500bp with all of the gRNA target sequences.

GCTGAGTGTGGGCCCCTACCTAGAATGTGGGACGGAGTCTCACTCTAATTCCCGTTGTCCCAGCCTTAG
GCCCAGGCTGGAGTGCAGTGGTTATAGGATTCAACCGGAGGCGCCATCTTGGCTCCCTCTGATTGCAAT
CTCCGCCTCTGGACCTCCGCCTCCTGGTTCGGCATTTTGAGTGTTAGACTGGGATTCTCCTGCCTCAGCCT
TTGGGACCTCTTAACCTGTGGCCAAGTAGCTGGGATTACAGGTCTCCCCAAGGCGCACTGGGCACCTGC
CATCACGCCGCACATCTCATGGGGTTATAGGGGTAGAGACGGGGTTTCACAGGGGAGTACTGTAGGAA
GAGGTGTTGGCTAGGCTGGTCTGCACGGTCAGTTGCCCTGAGGGAACTCCTGACCTCAGGTATGGAATT
TTCGCTTCCCACAGGTCAGCCTCCCGAAATGCTGGGAATAGGGTGCACATTTAGGGTGGTAGCTCATGC
CTGTAACCCCAATGTC

Spacer Sequences 17bp (from intronic area DS of TP53 exon 10)
GACGGAGTCTCACTCTA
CCCAGGCTGGAGTGCAG
CGCCATCTTGGCTCCCT
ACCTCCGCCTCCTGGTT
GATTCTCCTGCCTCAGC
CCAAGTAGCTGGGATTA
GCACCTGCCATCACGCC
GTAGAGACGGGGTTTCA
TGTTGGCTAGGCTGGTC
AACTCCTGACCTCAGGT
TCAGCCTCCCGAAATGC

Beginning spacer sequence (7bp):
GCTGAGT
Ending spacer sequencer (30bp):
GTGGTAGCTCATGCCTGTAACCCCAATGTC

Conclusions

The results presented here contribute to the understanding of precancer and early cancer detection in several contexts. Chapters 1 and 2 were dedicated to expanding our knowledge of the early genetic events that lead to cancer through the study of mtDNA mutations in ulcerative colitis-associated preneoplasia. Chapter 3 used *TP53* mutations in HGSOC as a means for the detection of this deadly cancer type using minimally invasive biopsies and illustrated the need for understanding age-related somatic mutations in normal tissue in order to calibrate potential cancer biomarkers. Chapter 4 described a modification to the DS methodology that increases the efficiency of the technology while maintaining its high degree of accuracy. Altogether, this work demonstrates that ultra-accurate sequencing methods have the power to improve our understanding of the genetic events that occur during tumorigenesis and to facilitate the development of non-invasive liquid biopsies.

The data presented in Chapters 1 and 2 focused on the role of clonal expansions and mitochondrial DNA mutations in UC tumorigenesis. In Chapter 1, we discussed the contribution of the field effect in ulcerative colitis as an indicator of early cancer progression. We posit that identification of clonally expanded fields by the detection of their molecular alterations might be useful for risk stratification and could enable individually tailored treatment plans in which patients with very few or small clonal fields might be able to undergo surveillance colonoscopy less frequently than patients with multiple, large fields. Such a paradigm shift in treatment may help reduce costs per patient and increase provider and patient compliance. This is a departure from the current prevailing surveillance methods, which rely entirely on visual identification of dysplastic lesions. While these methods have decreased the number of patients that do progress to cancer by recommending patients that develop a certain degree of dysplasia undergo a colectomy, interval cancers still occur. Our work in Chapter 1 suggests that molecular tests that

identify field effects may help decrease morbidity for these patients through the proposed risk stratification theme.

In Chapter 2, we used DS to extensively characterize mitochondrial DNA mutations at the different dysplastic stages of ulcerative colitis tumorigenesis. We found that mitochondrial DNA mutations increase in number and frequency from negative to low grade dysplasia and then a decrease in number in high grade and cancer, in agreement with our hypothesis based on prior mitochondrial protein measurements by IHC. High grade and cancer biopsies, however, retain larger, mostly synonymous clones. Mutations in early dysplasia were also more likely to be deleterious, as evidenced by the higher proportion of nonsynonymous and likely pathogenic mutations. These appear to be removed by a bottleneck event in later dysplasia and cancer, restoring mitochondrial function and allowing the expansion of passenger clones.

Our results may reconcile controversy in the field of mitochondrial mutations in cancer. Some believe that oxidative damage must cause mitochondrial DNA mutations that render the organelles non-functional and that this process must contribute to carcinogenesis (1-4) while others believe that functional mitochondria are necessary for tumorigenic progression (5). Our results partially support both theories, as they demonstrate abundant mitochondrial DNA mutations in early tumorigenesis, as well as a decrease in the overall number of mutations and their pathogenic traits in cancer. Our results, however, disagree with a major role of oxidative damage in the production of mtDNA mutations. While some biopsies showed contribution by oxidative damage, the majority of the mutations we identified were caused by DNA replication errors. We find, therefore, that mtDNA mutations are accumulated and tolerated in early preneoplasia up unto a threshold, at which point functional mitochondria must be restored. While a direct extrapolation cannot definitely be made from an inflammatory, preneoplastic disease to

cancer generally, our results do correspond to findings in many, non-inflammatory cancer types (6,7) as well as in aging tissues (8).

mtDNA mutations have historically been neglected for study in favor of the nuclear genome due to the high number of mtDNA copies within each cell which poses a formidable technical challenge to analyze accurately. Improvements like DS have led to closer examination of the role of mitochondria in tumorigenesis, but this area of research is still growing. Our work is novel, not only because it comprehensively characterizes these mutations in precancer, but because it exploits the stepwise progression of UC preneoplasia to elucidate the changes in mitochondrial mutations over the course of cancer progression

There are several directions in which this work could be continued. First, sequencing of many samples from many more UC colons at different stages of progression would elucidate the size and number of fields of mitochondrial DNA mutations and allow correlations to be made between progression and the extent of the fields with these mutations. Such an approach could potentially lead to the development of biomarkers for early cancer *detection* based on the presence, number, and size of clonally expanded fields. Second, sequencing of longitudinal biopsies from patients at each of their colonoscopies or colectomies over the duration of their disease would facilitate building a natural history of the changes to mitochondrial mutations over the course of UC carcinogenesis in real time. An understanding of the development of mutations over the course of progression could potentially lead to improved methods early cancer *prediction* by building a catalogue of the expected mutational landscape at each stage of cancer development. A further avenue of interest is single-cell sequencing of mitochondrial DNA mutations. Such an endeavor would allow accurate resolution of heteroplasmy at the single cell level and would be informative of the bottleneck process by which functional mitochondria are restored and large, synonymous clones are fixed. Such studies would test our proposed model of

progression as described in Chapter 2, and may point to which, if any, of the three suggested mechanisms of mitochondrial restoration is at play in UC carcinogenesis.

While mtDNA mutations harbor biomarker potential, the sequencing of the whole mitochondrial genome with DS is expensive, which currently limits its clinical utility. For the purpose of developing an affordable, clinically usable biomarker based on clonal expansions, our group is focusing on sequencing mononucleotide guanine repeat tracts (PolyG), which are repeats of 15-20 guanine bases found in non-coding regions of the genome (9,10). Previous work by our group has shown by Sanger sequencing that clonal insertions and deletions in these regions are indicative of UC progression (9). Because PolyG mutations are non-coding, they serve as passenger indicators of the extent of proliferation. With the accuracy and sensitivity of DS, we will be able to resolve even subclonal mutation populations (11,12). In addition to being highly mutagenic, these tracts are convenient because they are very short, thus requiring less total coverage than the entire mitochondrial genome and fewer sequencing resources overall. This allows for multiplexing of many PolyG regions. Of note, some of our purported fields of mitochondrial DNA mutations correspond to fields found via PolyG analysis. We are hopeful that PolyG analysis may be useful in combination with current surveillance methods to improve the prediction of UC progression and prevention of interval cancer.

Our work in Chapter 3 describes the application of DS to the early detection of HGSOC using uterine lavage. We were able to detect known tumor-causing mutations with 80% sensitivity; 70% of those identified were found about the 1% MAF cut-off established to distinguish cases from controls. This work is among the most successful attempts to detect ovarian cancer in a research setting (13), a disease for which there is still no effective screening test. The potential consequence of this work is the development of a clinically viable early detection method that could save patient lives.

Of note, with increasing age, we found higher mutation frequency and stronger positive selection for mutations in gynecological tissue as well as peripheral blood. This trend was identified not only in HGSOC patients, but also in normal tissue from control patients, from newborn to centenarian. This work, along with previous studies that found similar background mutations in otherwise healthy individuals (14-16) demonstrates the vital need to identify and classify age-related background mutations to remove these as potential confounders for early detection. While DS increases the detection rate of cancer-causing mutations, our results serve as a cautionary tale for the development of biomarkers. Calibration of such early detection tests must take into account the somatic evolution operative throughout human lifespan. Greater success in early detection is likely to be achieved through combination with complementary biomarkers, such as proteomics.

Continued work on this potential cancer prediction tool will require a much larger sample size than that analyzed here. Efforts are ongoing to validate our findings in many more patients. We have shown that we can detect intermediate and late stage cancers, but a viable biomarker must be able to detect tumors at early stages when they are most curable. Because most HGSOCs are not found until late stage, it will be more realistic to focus on high-risk populations such as *BRCA* mutation carriers.

In addition to these ongoing efforts for early detection of HGSOC, our group is also working on the development of tests for cancer prediction in high-risk women. We are currently analyzing *TP53* mutations in the blood and fallopian tissue of *BRCA* wild type patients (Group 1), *BRCA* mutant patients without HGSOC (Group 2), and *BRCA* mutant patients with HGSOC (Group 3) using CRISPR-DS as described in Chapter 4. Our goal is to determine whether a constitutionally higher level of background *TP53* mutations for a given age is associated with higher risk of ovarian cancer. If so, the measurement of *TP53* mutational burden in blood could

be used as a potential tool for risk stratification and cancer risk management in *BRCA* carriers. While successfully sequencing fallopian tube tissue has proved challenging due to the low amount of DNA available in such small OCT biopsies as we have used, we have begun to resolve mutations for these samples. We hope to determine whether women with higher mutational loads in blood also have higher mutational load in the fallopian tubes. More importantly, we hypothesize that we will see an increase in the overall burden of *TP53* mutations from Group 1 to Group 2 and finally Group 3. If our final data does support our hypothesis, our work, in conjunction with the concept of age-related calibration, may lead to the development of a non-invasive cancer prediction test for HGSOC. Efforts in this direction will be bolstered to improvements to the CRISPR-DS method, to be described below.

Finally, the results presented in Chapter 4 describe the development of a modified DS method, which harnesses the power of CRISPR/Cas9 digestion not for genome editing, but for target enrichment. Isolation of the *TP53* exons of interest by size selection prior to library preparation increased our recovery rate by 10 to 100-fold. The main benefit of CRISPR-DS is that it leverages the extreme accuracy of DS while lowering the required amount of input DNA. Standard DS requires ~500-1000ng of DNA for an average read depth of ~3000x. CRISPR-DS, however, can achieve this depth with only 10ng, opening DS to a range of clinical applications that rely on minimal amounts of DNA. For instance, CRISPR-DS could be used to detect very small, mutated clones within non-dysplastic biopsies or to detect subclones with mutations that confer therapy resistance within cancer biopsies.

A further benefit of this methodology is that it can be applied in any number of settings in which the gene of interest in known. For example, in UC, one could feasibly analyze *TP53* mutations, which are known to be early driver mutations in UC carcinogenesis, and then correlate these mutations with the extent of PolyG clonal fields. Additionally, one could

sequence genes associated with DNA repair mechanisms to see whether mutations in these genes correlate with the extent of accumulated DNA repair errors in the mitochondrial genome (17-19). The multiple applications of ultra-accurate sequencing have already been reviewed (20) and range from forensics to antimicrobial resistance and immunological mosaicism. The development of CRISPR-DS enhances the success and translational potential of most of those applications by improving yield, reducing DNA input requirements, and increasing cost-effectiveness.

While CRISPR-DS greatly increases the efficiency of DS, there is still DNA lost during the hybridization capture step of library preparation. In order to enable an even wider rage of applications, including analysis of degraded DNA from FFPE tissue alterations must be made to further reduce DNA losses, ideally through removal of hybridization capture from DS library preparation. Next steps in this work require the implementation a PCR-based CRISPR-DS approach. In such an approach, the exons of interest would still be excised using Cas9-mediated digestion, but instead of relying on hybridization capture, nested PCR reactions would generate sufficient copies of the molecules of interest. These molecules would then be isolated through a simple size-selection step, thus preserving much more DNA. We anticipate that our efforts on this front will be applicable to a broad range of biological and medical questions and will open up archives of FFPE tissue for analysis.

The work presented in this thesis not only demonstrates two applications of DS for early cancer detection, but makes findings critical to several fields. In the field of UC, our suggestions could shift the surveillance methods used for detection while improving patient quality of life and decreasing provider workload. Further, we have provided one of the first descriptions of the contribution of mtDNA mutations to UC carcinogenesis, shedding light on the early mutagenic processes in this disease. We have also possibly bridged the divide between those who believe

mitochondria are necessarily dysfunctional in cancer progression and those who believe function mitochondria are vital to tumor development. Finally, we have used two iterations of the DS method to better understand background variation in normal and precancerous tissues and have worked toward the development of a viable biomarker for HGSOC detection and prediction. This work illustrates that cancer is an evolutionary process fostered by genetic mutations that accumulate through life. Ultra-accurate sequencing has enabled us to detect these mutations with unprecedented resolution, increasing our knowledge of the dynamics of somatic mutations in aging, precancer, and cancer. Based on this knowledge, we now appreciated the existence of background aging mutations and have improved our ability to develop cancer tests that take into consideration these mutations and are optimized for early cancer detection and prediction. We hope these efforts will improve patient survival and lead to the development of tests that take us closer to our ultimate goal of eliminating cancer at an earlier stage.

References

1. Chatterjee A, Dasgupta S, Sidransky D. Mitochondrial subversion in cancer. Cancer Prev Res (Phila) **2011**;4:638-54
2. Sanchez-Arago M, Chamorro M, Cuezva JM. Selection of cancer cells with repressed mitochondria triggers colon cancer progression. Carcinogenesis **2010**;31:567-76
3. Larman TC, DePalma SR, Hadjipanayis AG, Cancer Genome Atlas Research N, Protopopov A, Zhang J*, et al.* Spectrum of somatic mitochondrial mutations in five cancers. Proc Natl Acad Sci U S A **2012**;109:14087-91
4. Yu M. Somatic mitochondrial DNA mutations in human cancers. Adv Clin Chem **2012**;57:99-138
5. Ward PS, Thompson CB. Metabolic reprogramming: a cancer hallmark even warburg did not anticipate. Cancer Cell **2012**;21:297-308
6. Ju YS, Alexandrov LB, Gerstung M, Martincorena I, Nik-Zainal S, Ramakrishna M*, et al.* Origins and functional consequences of somatic mitochondrial DNA mutations in human cancer. Elife **2014**;3
7. Ericson NG, Kulawiec M, Vermulst M, Sheahan K, O'Sullivan J, Salk JJ*, et al.* Decreased mitochondrial DNA mutagenesis in human colorectal cancer. PLoS Genet **2012**;8:e1002689
8. Kennedy SR, Salk JJ, Schmitt MW, Loeb LA. Ultra-sensitive sequencing reveals an age-related increase in somatic mitochondrial mutations that are inconsistent with oxidative damage. PLoS Genet **2013**;9:e1003794
9. Salk JJ, Salipante SJ, Risques RA, Crispin DA, Li L, Bronner MP*, et al.* Clonal expansions in ulcerative colitis identify patients with neoplasia. Proc Natl Acad Sci U S A **2009**;106:20871-6
10. Salipante SJ, Horwitz MS. Phylogenetic fate mapping. Proc Natl Acad Sci U S A **2006**;103:5448-53
11. Kennedy SR, Schmitt MW, Fox EJ, Kohrn BF, Salk JJ, Ahn EH*, et al.* Detecting ultralow-frequency mutations by Duplex Sequencing. Nat Protoc **2014**;9:2586-606
12. Schmitt MW, Kennedy SR, Salk JJ, Fox EJ, Hiatt JB, Loeb LA. Detection of ultra-rare mutations by next-generation sequencing. Proc Natl Acad Sci U S A **2012**;109:14508-13
13. Wang Y, Li L, Douville C, Cohen JD, Yen T-T, Kinde I*, et al.* Evaluation of liquid from the Papanicolaou test and other liquid biopsies for the detection of endometrial and ovarian cancers. Science Translational Medicine **2018**;10
14. Jaiswal S, Fontanillas P, Flannick J, Manning A, Grauman PV, Mar BG*, et al.* Age-Related Clonal Hematopoiesis Associated with Adverse Outcomes. New England Journal of Medicine **2014**;371:2488-98
15. Krimmel JD, Schmitt MW, Harrell MI, Agnew KJ, Kennedy SR, Emond MJ*, et al.* Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. Proc Natl Acad Sci U S A **2016**;113:6005-10
16. Martincorena I, Roshan A, Gerstung M, Ellis P, Van Loo P, McLaren S*, et al.* Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. Science **2015**;348:880-6
17. Wisnovsky S, Jean SR, Liyanage S, Schimmer A, Kelley SO. Mitochondrial DNA repair and replication proteins revealed by targeted chemical probes. Nature Chemical Biology **2016**;12:567
18. Alexeyev M, Shokolenko I, Wilson G, LeDoux S. The Maintenance of Mitochondrial DNA Integrity—Critical Analysis and Update. **2013**
19. Druzhyna NM, Wilson GL, LeDoux S, P. Mitochondrial DNA repair in aging and disease. **2008**

20. Salk JJ, Schmitt MW, Loeb LA. Enhancing the accuracy of next-generation sequencing for detecting rare and subclonal mutations. Nat Rev Genet **2018**;19:269-85