Highly accurate RNA and DNA sequencing:
Application to longstanding questions in aging and cancer

Kate-Siobhan Reid-Bayliss

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington
2016

Reading Committee:
Lawrence Loeb, Chair
Peter Rabinovitch
Alan Weiner

Program Authorized to Offer Degree:
Molecular and Cellular Biology

©Copyright 2016
Kate-Siobhan Reid-Bayliss

University of Washington

Abstract

Highly accurate RNA and DNA sequencing:

application to longstanding questions in aging and cancer

Kate-Siobhan Reid-Bayliss

Chair of the Supervisory Committee:

Lawrence A. Loeb, MD, PhD

Department of Pathology

Accurate DNA replication and RNA transcription are critically important for proper cell functioning: the *fidelity* of these processes is crucial; infidelity can lead to cellular dysfunction and disease. The key problem in studying the fidelity of these processes is the accurate detection of rare DNA and RNA mutations, which result as a consequence of infidelity. Until recently, this has not been possible, as the high error rates of available methods has limited their ability accurately detect rare mutations among a preponderance of wildtype molecules. The solution to this problem, as the Loeb lab and others have found, is to perform single molecule sequencing of individually barcoded DNA and RNA molecules. In the present work, I present three projects which apply the use of barcoding individual DNA and RNA molecules in order to enable highly accurate and sensitive analyses of DNA replication and RNA transcription fidelity.

(i) The question of why CS patients don't get cancer despite being repair-deficient has puzzled scientists for decades. While many have speculated as to the cause, we have

applied Duplex Sequencing to definitively answer this question: CS patients fail to develop cancer because they do not accumulate mutations more quickly than repair-proficient individuals. In addition to finally solving this long-standing mystery, we provide novel insights into the mutagenic consequences of UV treatment in CS cells, at an unparalleled sensitivity.

(ii) The question of why GBM patients do so poorly and always recur has long plagued doctors and scientists. Here, we expand on the excellent clonal mutation work of our predecessors, revealing that the substantial inter- and intra-tumoral clonal heterogeneity is further compounded by considerable subclonal heterogeneity. We show that subclonal mutations are highly heterogenous within individual GBM tumors, between GBM tumors from different patients, as well as between primary and recurrent tumors from the same patient. Our findings of high subclonal heterogeneity in GBM tumors suggest that GBM patients do so poorly because their tumors already contain a reservoir of mutations that potentially enable them to adapt to any treatment currently available. This underlies the importance of expanding subclonal mutation studies of GBMs to better understand their mutational makeup.

(iii) The question of what, if any, contribution RNA mutations have to health and disease has been one that has remained unanswered for more than 50 years. RNA mutations have long been hypothesized to play roles in human health and disease, as well as in several other processes, including RNA virus evolution and bacterial resistance to antibiotics. Unfortunately, until now, it has been very difficult to study the hypothesis that transcriptional mutagenesis, resulting in RNA mutations, contributes to or drives these processes because there have not been the tools available to do so. I have, therefore, developed a method to accurately sequence RNAs. Here, I demonstrate that Accurate RNA Consensus Sequencing (ARC-seq) has inherent adaptability to enable increased

stringency, which eliminates a high level of damage-induced artifacts. I also show that RNA polymerase mutants induce increased transcriptional mutagenesis *in vivo*, with different mutants producing varying RNA mutation spectrums. Finally, I demonstrate the utility of ARC-seq to address questions on the biological importance of transcriptional mutagenesis *in vivo* by using ARC-seq to show that oxidative stress induces high levels of transcriptional mutagenesis in both mRNA and rRNA. Thus, ARC-seq will enable studies on how perturbing a cell's environment or machinery affects the fidelity of transcription and to what extent RNA mutations contribute to aging, cancer, and neurodegeneration, as well as the evolution and acquired resistance of viruses and bacteria.

Together the three projects encompassed in this thesis demonstrate the power of combining the use of barcoding individual DNA and RNA molecules in order to enable highly accurate and sensitive analyses of DNA replication and RNA transcription fidelity.

TABLE OF CONTENTS

ACKNOWLEDGMENTS

I would like to thank the scientific mentors with whom I have been fortunate to work throughout my young scientific career. Steve Stefanides, my mentor at Wenatchee Valley College, plucked me out of his intro chemistry class when I was still in running start at Wenatchee Valley College and taught me the wonders of asking and answering my own scientific questions rather than simply learning the discoveries of other scientists. He nurtured in me an intellectual curiosity and an unquenchable enthusiasm for science. It is largely a product of his passion for research and persistence that I ever began doing research and eventually decided to pursue an MD-PhD. Tina Negritto, my mentor at Pomona College, not only allowed me to pursue a project of my own devising but was incredibly supportive and patient with me as I struggled with my first solo foray into the world of failed experiments in the pursuit of a novel discovery. Larry Loeb, my mentor throughout my PhD, allowed me to pursue my "passion project" while keeping in place the "safety net" of more "grad-student-appropriate" projects. Larry gave me a lot of freedom and responsibility, and I feel as though my time in his lab has given me a true appreciation for the many challenges and rewards of being a scientist in these times. Thank you.

I would also like to thank my committee members, Gwen Garden, Dana Miller, Peter Rabinovitch, and Alan Weiner, for their guidance and support throughout my PhD. And thank you to James Cleaver, who brought the question of UV mutagenesis in CS to us and let me shape and lead the project as I saw fit. My undergraduate student, Julia Joo, was an immense help with the GBM project; she learned the Duplex Sequencing protocol exceptionally fast for such a young scientist and helped process samples and verify mutations. I'll be excited to learn where her future takes her. And I am immensely thankful to all the current and former members of the Loeb lab who have provided guidance, support, and friendship while I've worked my way thru my PhD. Thank you.

I would also like to thank my friends, those whom I've known since college and those who have more recently come into my life. They have been a constant source of adventure, laughter, and

support; they have helped me maintain my sanity and gain some much needed perspective on both science and life. Thank you to my significant other, Michael; he has been my champion as the end has approached, repeatedly reminding me of just how impressive a PhD is, and has helped me enjoy life even when there wasn't much time for it. Finally, I would like to thank my family, and especially my parents, for their ever-present love and support. They instilled in me a never-ending curiosity and stubborn perseverance that will hopefully continue to be of great service in my life. Thank you.

INTRODUCTION

The Central Dogma of Molecular Biology outlines the basic mechanisms underlying the molecular basis of life. DNA, the archive of our cells, is replicated to make more DNA, which is then passed on to daughter cells. DNA is also transcribed into RNA, the instruction book for the building blocks of our cells; RNA, in turn, is translated into proteins, the workhorses in our cells. Proper cell functioning is critically important upon these processes being accurate: the *fidelity* of these processes is crucial; infidelity can lead to cellular dysfunction and disease. The evidence of infidelity is the presence of DNA and RNA mutations, which arise due to mistakes in replication and transcription, respectively. In order to study the fidelity of DNA replication and RNA transcription on a large scale and address questions of their roles in health and disease, as well as to study their governing mechanisms, it is necessary to sequence individual molecules to detect rare mutations in a high throughput manner. The key problem in studying the fidelity of these processes is the accurate detection of rare mutations among a preponderance of wildtype molecules.

DNA FIDELITY

Historically, the only mutations we've been able to observe by DNA sequencing are those that are present in a substantial fraction of a tissue, referred to as clonal mutations. This is due to conventional DNA sequencing methods, including next-generation sequencing (NGS), being highly error prone [1]. Consequently, conventional NGS determines DNA sequences by forming a consensus of the DNA molecules sequenced, which prevents it from reliably detecting mutations below the level of about 1 in 20 wildtype sequences, limiting its detection ability to 5% clonality or greater. Rare mutations are lost during this consensus making. As such, conventional NGS prevents us from detecting a large number of rare mutations, which are the evidence of the infidelity of DNA replication.

While clonal and high frequency subclonal mutations are crucial to understanding many diseases, particularly heritable disease and cancer, the Loeb lab's interest lies in studying low frequency subclonal and rare mutations, as such mutations may serve as a reservoir for cellular adaptability and resistance [2]. Thus, investigating the role of subclonal and rare mutations in various disease contexts may give us insight into the mechanisms of disease initiation, resistance, and recurrence, particularly in cancer, and may potentially lead to new targets for treatment. Additionally, studying rare mutations will enable us to continue exploring the mechanisms that govern faithful DNA replication and repair.

The solution to the problem of accurate detection of rare mutations among a preponderance of wildtype molecules, that the Loeb lab and others have found, is to barcode individual molecules and then perform single molecule sequencing. Since DNA is inherently duplex in nature, with the strands of the helix being reverse-complement copies of each other, the Loeb lab reasoned that by maintaining the two strands' relationship to each other thru the sequencing process, one could leverage the power of the duplex to eliminate the artifacts that make conventional NGS so error prone [3, 4]. Duplex Sequencing accomplishes this by first tagging the duplex DNA with a double-stranded unique barcode (Fig. 1); this barcode contains 24 nucleotides of randomized sequence that uniquely identifies each double-stranded DNA molecule. The two strands of the DNA duplex are then amplified and sequenced. After sequencing, the copies of each strand are used to form single-stranded consensus sequences, which eliminate many but not all PCR and sequencing errors. The two strands of the original duplex are then compared and only mutations present in both strands are kept; this eliminates the remaining damage-induced and PCR artifacts. By eliminating damage-induced, PCR, and sequencing artifacts, Duplex Sequencing enables the detection of mutations as low as 1 in 10 million wild type sequences. This level of sensitivity enables Duplex Sequencing to detect subclonal and rare mutations with very high accuracy.

A trade-off of Duplex Sequencing's high sensitivity and accuracy, however, is that, due to the cost and capacity of current sequencing platforms, it is presently only feasible to sequence several kilobases of DNA at once. Thus, in order to apply Duplex Sequencing to detect rare mutations in human cells, it is necessary to capture a small subset of genes to study them at a depth sufficient to detect rare mutations. The consequence of this is that there may be mutations present in non-captured genes that are not detected in our studies.

RNA FIDELTY

The answer to the question of how to study RNA fidelity is more complicated. Unlike DNA, RNA is single stranded and, thus, contains only a single copy of its sequence. Additionally, in order to sequence RNA, it must first be reverse transcribed into complementary DNA (cDNA), which is a highly error-prone process that potentially contributes 10-fold more artifacts than there are true RNA mutations [5]. These reverse transcription errors, coupled with the damage-induced, PCR, and sequencing artifacts inherent to conventional NGS, results in conventional RNA sequencing (RNAseq) highly error prone. Consequently, RNAseq is only useful for studying gene expression changes (i.e. the abundance of various gene transcripts in cells) and as an indirect determination of DNA mutations in single-cell sequencing [6]; conventional RNAseq cannot reliably detect mutations in single RNA molecules, which are the evidence of the infidelity of transcription.

Given that multiple copies of RNA are transcribed from each gene, it may not be immediately obvious how transcriptional infidelity resulting in a mutation in one RNA molecule can affect a cell. But, there are situations in which it could. Some RNA mutations could act as initiators for pathological processes, such as carcinogenesis or antibiotic resistance, whereas others could promote protein misfolding and inhibit protein degradation, thus allowing protein accumulation and aggregation, such as in neurodegenerative disease. In the case of human disease, the neurodegeneration example is perhaps the most compelling. Take, for example, Alzheimer's disease. While familial forms of AD

have been linked to mutations in DNA, the vast majority of AD occurs sporadically, and no DNA mutation can be found. While we do not yet know why some people develop sporadic AD while others do not, one hypothesis is that RNA mutations could be the cause of the mutant proteins that misfold, promote misfolding of wildtype proteins, and, consequently, seed the aggregation of proteins into the plaques and tangles pathognomonic of Alzheimer's disease.

In addition to their potential role in human disease, RNA mutations have been hypothesized to play a role in several other situations, including RNA virus evolution and bacterial resistance to antibiotics [7-9]. Unfortunately, until now, it has been very difficult to study the hypothesis that RNA mutations contribute to or drive these processes because there have not been the tools available to do so. Thus, I developed a method to accurately sequence RNAs, which will enable the study of infidelity in transcription in any context. Accurate RNA Consensus Sequencing (ARC-seq) combines the use of unique barcodes for each RNA molecue and the generation of multiple cDNA copies per RNA molecule. This combination enables ARC-seq to eliminate artifacts due to cDNA synthesis, PCR errors, and sequencing errors, yielding the true RNA mutations. ARC-seq has inherent adaptability to enable increased stringency that eliminating a high level of damage-induced artifacts may require; and it is unbiased with respect to template. Also, any RNA from any source can be accurately sequenced. Thus, ARC-seq will enable studies on the role of transcriptional infidelity and its resultant RNA mutations in evolution, aging, carcinogenesis, and resistance.

In the following chapters, I present three projects that employ the use of single molecule sequencing of individually barcoded DNAs and RNAs to detect rare mutations. The first two projects employ Duplex Sequencing to address questions of (i) UV-induced mutagenesis in Cockayne Syndrome and (ii) subclonal mutational heterogeneity in Glioblastoma. The third project presents my newly developed ARC-seq method that enables highly accurate RNA sequencing for the study of RNA fidelity mechanisms.

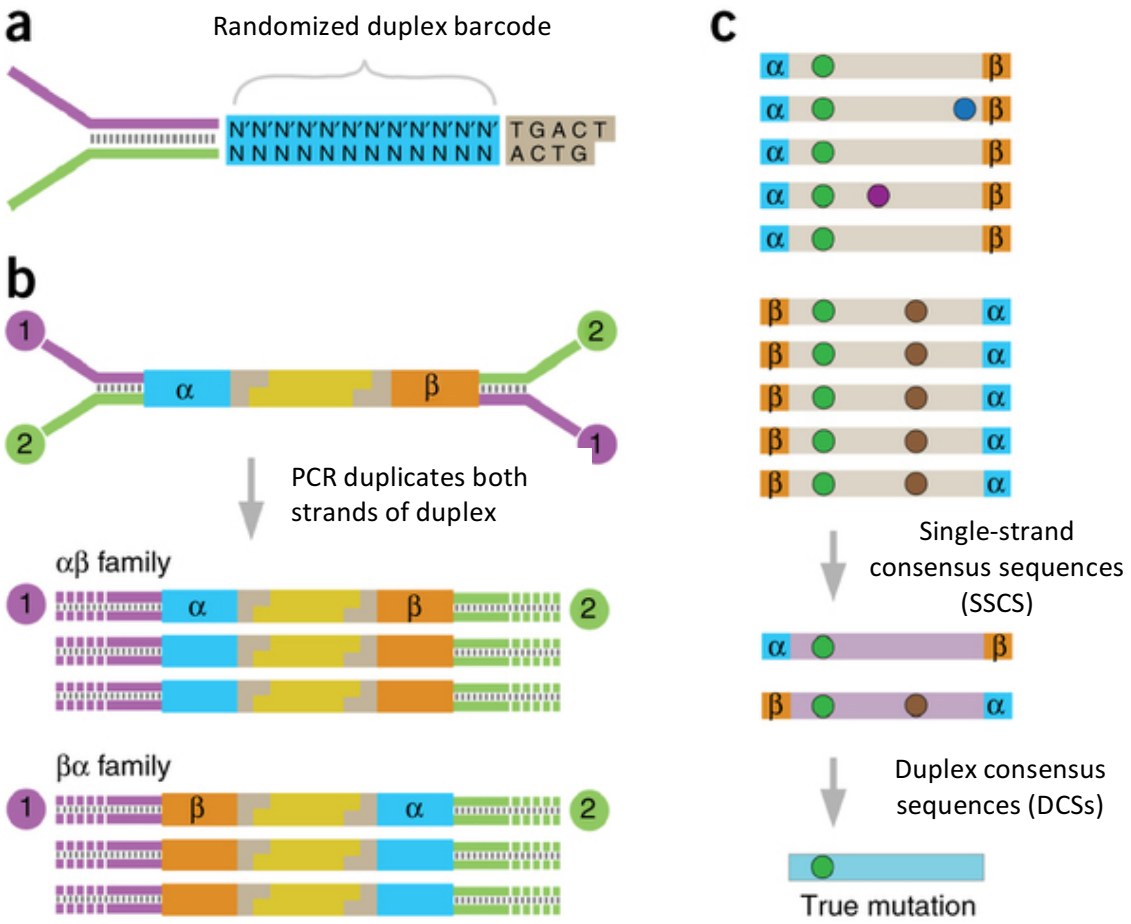


Figure 1: Overview of Duplex Sequencing. (a) Schematic of a Duplex Sequencing adapter, showing the random double-stranded barcode. (b) Ligation of the adapters to the sample DNA results in a unique 12-nt tag sequence on both ends of the molecule. PCR amplification of each strand of a DNA duplex results in two distinct, but related, PCR products. (c) Reads sharing unique α and β tag sequences are grouped together into tag families of form αβ or βα, and an SSCS is created for each tag family. Mutations are of three different types: sequencing mistakes (blue or purple dots); first-round PCR errors (brown dots); true mutations (green dots). Formation of the SSCS removes the first type of error, but not first-round PCR errors. Comparing SSCSs from the paired families with tags αβ and βα generates a DCS, which eliminates these first-round PCR errors. True mutations are scored if and only if they are present at the same position in both strands of the DNA. Figure is adapted from [10], © 2013 Kennedy *et al.*

CHAPTER 1: WHY COCKAYNE SYNDROME PATIENTS DON'T DEVELOP CANCER DESPITE THEIR REPAIR DEFICIENCY

ABSTRACT

Cockayne syndrome (CS) and Xeroderma pigmentosum (XP) are human photosensitive diseases with mutations in the nucleotide excision repair (NER) pathway, which repairs DNA damage from UV exposure. CS is mutated in the transcription-coupled repair (TCR) branch of NER, and exhibits developmental and neurological pathologies. The XP-C group of XP patients are mutated in the global genome repair (GGR) branch of NER and have a very high incidence of UV-induced skin cancer. Cultured cells from both diseases have similar sensitivity to UV-induced cytotoxicity, yet CS patients have never been reported to develop cancer, although they often exhibit photosensitivity. Since cancers are associated with increased mutations, especially when initiated by DNA damage, we examined UV-induced mutagenesis in both XP-C and CS cells, using Duplex Sequencing for high-sensitivity mutation detection. Duplex Sequencing detects rare mutagenic events, independent of selection and in multiple loci, enabling examination of all mutations rather than just those that confer major changes to a specific protein. We found telomerase-positive normal and CS-B cells had increased background mutation frequencies that decreased upon irradiation, purging the population of subclonal variants. Primary XP-C cells had increased UV-induced mutation frequencies compared to normal cells, consistent with their GGR deficiency. CS cells, in contrast, had normal levels of mutagenesis, despite their TCR deficiency. The lack of elevated UV-induced mutagenesis in CS cells reveals that their deficiency in TCR, though increasing cytotoxicity, is not mutagenic. The absence of cancer in CS patients, therefore, is due to the absence of UV-induced mutagenesis rather than enhanced lethality.

INTRODUCTION

The nucleotide excision repair (NER) syndromes, Xeroderma pigmentosum (XP) and Cockayne Syndrome (CS), lie at the extremes of increased cancer and neurodegeneration, respectively [11]. The XP-C class of XP patients have mutations in the DNA damage recognition proteins, DDB2 and XP-C, of global nucleotide excision repair (GGR) They are characterized by UV hypersensitivity, sun-induced cutaneous features, such as hypopigmentation and hyperpigmentation, and a greatly increased incidence of cancer; XP-C patients' susceptibility to skin cancer is increased by >1,000-fold [11-13]. In contrast, CS patients, with mutations in the RNA polymerase II cofactors CSA and CS-B, which recognize damage in transcribed regions (transcription-coupled repair, TCR), are characterized by neurological and developmental symptoms, such as early cessation of growth, microcephaly, mental retardation with dysmyelination, cachexia, and a greatly reduced life expectancy [11]. The reported average life expectancy of patients with CS is just 12 years [14]. Additionally, CS patients are also highly photosensitive, burning and blistering after only minutes of sun exposure (5). However, in stark contrast to the dramatic increase in skin cancer incidence in XP-C patients, no CS patient has ever been reported to develop cancer, skin or otherwise [11, 14-18]. Since cancer, especially in skin, is associated with mutagenesis [19, 20], we hypothesized that, unlike defects in GGR, which are associated with enhanced UV-induced mutagenesis in XP-C cells [21], the TCR defects in CS cells may not lead to increased mutagenicity.

Most sensitive mutagenesis studies in XP and CS cells have been confined to a few selectable genes, mutation of which confers drug resistance. Maher showed that UV-induced mutations in the *HPRT* gene that conferred resistance to 6-thioguanine (6-TG) were greatly enhanced in XP cells from both excision-defective and polymerase-defective groups [21, 22]. Similar studies in CS cells, however, failed to show an increase in UV-induced mutations in the *HPRT*, *T cell receptor* or *glycophorin A* gene loci [23]. In contrast, an episomal plasmid (pZ189), irradiated with UV and passed through CS cells showed increased levels of mutations [24, 25]. The limitations of these

methodologies include the small number of potential gene loci suitable for drug selection and the possibility that episomal vectors may not fully induce the DNA damage response of whole cells, thereby resulting in a high mutation frequency that is not representative of mutagenesis in chromosomal loci. These limitations complicate the disparate results of these previous studies, leaving unresolved the question of whether or not CS cells demonstrate elevated UV-induced mutagenesis.

With the development of next-generation sequencing came the potential to survey multiple genes simultaneously and independently of selection. However, standard next-generation sequencing methodologies are highly error-prone and limited to surveying mutations present at greater than 1 in 20 wildtype sequences [26]. To counteract the limitations of standard next-generation sequencing platforms, we (L.A.L, K.S.R-B) utilized Duplex Sequencing, a highly accurate sequencing method that is 100,000-fold more accurate than traditional next-generation sequencing methodologies. Due to its ability to remove sequencing artifacts resulting from DNA damage, as well as amplification and sequencing errors, Duplex Sequencing enables the detection of mutations as low as 1 in 10^8 nucleotides sequenced [27, 28].

To test our hypothesis that defects in TCR may not lead to increased UV-induced mutagenesis, unlike defects in GGR, we employed Duplex Sequencing for high-sensitivity mutation detection in primary cells derived from normal patients and patients with XP-C and CS. We find that, although primary XP-C and CS cells have similar sensitivities to UV-induced cell killing, surviving cells are radically different. XP-C survivors exhibit high levels of UV-induced mutations; CS cells do not.

RESULTS

UVC- and UVB-induced cytotoxicity in primary cells. Primary fibroblasts (Table S1), derived from normal adult skin (GM05659) and normal neonatal primary foreskin (NHF-D), and primary neonatal keratinocytes were exposed to UVB or UVC, cultured for 5-7 days, and harvested. Using the MTT (3-

(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide, Sigma-Aldrich, St Louis, MO) assay, the surviving fraction at the time of cell harvesting was calculated relative to untreated cells of the same genotype. Primary fibroblasts and keratinocytes were sensitive to killing by UVB and UVC, with keratinocytes showing increased sensitivity to both, relative to fibroblasts (Fig. 1A, 1B).

Increased UVC-induced cytotoxity in repair-deficient primary fibroblasts. Normal (GM05659 and NHF-D), XP-C (GM02997 and XP226BA), CS-A (GM17536 and GM01856), and CS-B (GM01428 and GM01629) primary fibroblasts were exposed to UVC, cultured for 5-7 days, and harvested. Using the MTT assay, the surviving fraction at the time of cell harvesting was calculated relative to untreated cells of the same genotype. XP-C, CS-A, and CS-B primary fibroblasts were markedly more sensitive to killing by UVC than normal primary fibroblasts (Fig. 1C), consistent with many previous studies reporting the enhanced UV sensitivity of XP and CS cells. One primary CS-A cell line, GM17536, appeared anomalous for CS-A and had higher survival than all other repair-deficient cells (Fig. 1C, S1A). We, therefore, excluded the GM17536 cell line from our subsequent mutational analyses.

UVC and UVB induce subclonal mutations in normal primary cells. To validate the use of Duplex Sequencing to detect mutagen-induced mutations, normal fibroblasts (GM05659, NHF-D) and keratinocytes, were exposed to UVB or UVC, cultured for 5-7 days, and harvested. Genomic DNA was then isolated and subjected to a single round of Duplex Sequencing [28]. Target genes were exonic regions of *NRAS*, *UMPS*, *PIK3CA*, *EGFR*, *BRAF*, *KRAS*, *F10*, *TP53*, and *TYMS* (Table 1), several of which were chosen for their importance in skin carcinogenesis.

The spectrum of subclonal (<20% clonal) mutations observed in normal fibroblasts and keratinocytes showed a dose-dependent increase in C:G→T:A transitions, as a function of UVB dose in primary keratinocytes and as a function of UVC dose in primary fibroblasts, especially in NHF-D

cells (Fig. 2A). In contrast, transversions were dose-independent. Of particular interest for UV-induced mutagenesis studies are C:G→T:A mutations at dipyrimidine sites (CpT, TpC, and CpC; hereafter referred to as Py-Py sites), as these are a known signature of UV-induced mutagenesis (22, 23). Importantly, when we examined the context of the C:G→T:A transitions, the majority of UV-induced mutations occurred at Py-Py sites and showed a dose response to UVB and UVC in all normal genotypes (Fig. 2B). These results are consistent with previous reports of UV-induced mutagenesis in reporter genes [21, 23-25] and validate our use of Duplex Sequencing for the detection of UV-induced mutagenesis. We chose to carry out subsequent experiments using UVC, as both UVB and UVC produce similar mutagenic photoproducts. In our hands the difference in cyclobutane dimer yield per $J.m^{-2}$ was approximately a factor of 7, measured by the cleavage of a plasmid using *M.luteus* UV endonuclease.

UVC induces an elevated subclonal mutation frequency in XP-C cells. Normal (GM05659 and NHF-D) and XP-C (GM2997 and XP226BA) primary fibroblasts were exposed to UVC, cultured for 5-7 days, harvested, and subjected to a single round of Duplex Sequencing [28]. XP-C cells showed elevated subclonal mutation frequencies, relative to normal fibroblasts (Fig. 3A). When we focused on C:G→T:A mutations at Py-Py sites, we found that XP-C cells accumulated more of these UV-specific mutations with increasing UVC dose, relative to normal cells (Fig 3B). These results are consistent with previous reports, using reporter genes, of elevated UV-induced mutagenesis in XP cells [21]. Additionally, normal cells showed a smaller shallow increase in UV-specific mutations with UVC dose, which is consistent with efficient repair that minimizes UV-induced mutations and skin cancer initiation in repair-proficient cells [20].

CS cells fail to demonstrate elevated subclonal frequencies of UV-specific mutations upon UVC exposure. CS-A (GM01856) and CS-B (GM01428 and GM01629) primary fibroblasts were exposed

to UVC, cultured for 5-7 days, harvested, and subjected to a single round of Duplex Sequencing [28]. When we examined total subclonal mutations, primary CS cells appeared to show an initial increase in mutation frequency, with a reduction at higher UVC doses (Fig. 3C). However, in contrast to XP-C cells, CS-A and CS-B fibroblasts showed no elevation in C:G→T:A mutations at Py-Py sites, relative to normal fibroblasts (Fig. 3D). Indeed, when directly comparing XP-C and CS-A and -B cells (Fig. 3E), there is a marked increase in UV-specific mutations in XP-C fibroblasts but none in CS-A/B fibroblasts, despite their similar sensitivity to the cytotoxic effects of UVC (Fig. 1C). When we examined subclonal UV-specific mutation frequencies versus cell survival, the results suggested that CS cells might have an even lower mutation frequency than normal cells at equivalent survival (Fig S2).

Since UV-specific mutations did not account for the initial increase in total subclonal mutations seen in the CS cells (Fig. 3C), particularly the CS-B cells, we sought to determine if the UV-induced mutations in CS cells had an oxidative damage signature, as CS-B has been implicated in oxidative DNA damage repair [29-31]. Indeed, we found that the majority of UV-induced mutations in CS-B cells were G:C→T:A mutations, a signature of 8-oxo-dG-induced mutagenesis (Fig. 3F). This result is consistent with increased mutagenesis due to deficient oxidative DNA damage repair in CS cells.

Duplex Sequencing enables in-depth analyses of the mutagenic consequences of UVC. Since Duplex Sequencing allows us to study rare mutational events, we examined the distribution of UVC-induced mutations, combining all UV doses into a pool, designated UVC, and comparing these to the untreated cells, designated control (Table 2, Fig. S3 A-F). We determined the distribution of UVC-induced mutations between active and inactive genes, based on the gene expression status of each gene in skin (GeneCards, http://www.genecards.org/) and the distribution of mutations between the template (transcribed) and coding (nontranscribed) strands of active genes (Table 1, Table 2, Fig. S3

A-F). In XP-C cells, there was an increased ratio of C:G→T:A mutations in inactive genes, relative to active genes; there was also an elevated ratio of C→T mutations in the coding strand of active genes, relative to the template strand. These biases are consistent with the GGR deficiency of XP-C cells. In CS-B cells, there was little difference in the ratios of C:G→T:A mutations as a function of gene activity; there was, however, an elevated ratio of C→T mutations in the template strand, relative to the coding strand, consistent with the TCR deficiency of CS-B cells. Thus, although TCR deficiency influences the strand distribution of mutations, it does not increase the overall yield.

During our analysis, we observed 39 instances of multiple mutations within the same read (multiplets). Since the "classic" UV-induced mutation signature is the CC→TT mutation, we were intrigued by the presence of multiple other types of multiplet mutations (Table 3). While CC→TT mutations are the most frequent type of multiplet mutation observed, we encountered many other types, all of which occurred in UV-treated cells. Of 39 multiplet mutations, all but one occurred at or directly adjacent to a Py-Py (CpC, CpT, TpC, TpT) dinucleotide, consistent with the mutations resulting as a consequence of error-prone bypass of UV-induced damage. In addition to mutations that occurred within a doublet (i.e. CC → TT**)** or triplet (i.e. CTC → TTT), 6 of the multiplet mutations were two single mutations occurring 3-7 nucleotides apart.

DISCUSSION

XP and CS patients both have defects in NER. The XP-C class of XP patients display extreme UV sensitivity and are highly prone to develop skin, corneal, eyelid cancers due to their defects in GGR. CS patients, defective in the TCR branch of NER [11, 32], present a very different clinical picture, one of developmental defects and neurodegeneration; many but not all patients are also photosensitive, some developing blistering sunburns [15, 33]. In contrast to XP-C patients, no known CS patient has ever developed cancer [14, 15, 18]. Early studies of CS presented a discordant picture as to whether or not CS cells show an elevated UV-induced mutation frequency, relative to normal

cells, and differed depending on the method employed. Therefore, to definitively determine whether or not CS cells show an elevated UV-induced mutation frequency, we employed Duplex Sequencing, a highly accurate next-generation sequencing methodology that enables detection of rare mutagenic effects [28], to study UV-induced mutagenesis in primary cells derived from normal, XP-C, and CS-A and -B patients. In contrast to previous methods, our use of Duplex Sequencing [28] enabled us to study the mutagenic consequences of UV damage independent of selective pressures and at far greater detail than previously possible.

Our study of normal fibroblasts and keratinocytes validated our use of UVC to induce subclonal UV-specific mutations (C:G→T:A at Py-Py sites); we also validated our application of Duplex Sequencing to analyze the mutagenic consequences of UV in primary cultured cells absent selective pressures. Our analysis of the mutation spectrum in UVC-treated primary fibroblasts and UVB-treated primary keratinocytes revealed an elevated frequency of nearly all mutation subtypes in the keratinocytes, relative to the primary fibroblasts (Fig. 2). Interestingly, while the UV-induced C:G→T:A mutation showed the expected dose-response to UVB treatment, other mutations present in the untreated keratinocytes remained largely unchanged, indicating that these mutations were already present in the population. This increase in global subclonal mutations is not due to differences in culture duration between fibroblasts and keratinocytes, as the keratinocytes were used at a lower passage number than the fibroblasts. The most prevalent mutation type was the G:C→T:A transversion (Fig. 2A), which may be due to the mutagenicity of guanine oxidative products produced in culture under ambient oxygen concentrations [34-36].

When we analyzed the mutation frequency in unirradiated hTERT-immortalized normal (GM05659T) and CS-B (GM01428) cells, which had been maintained in culture for approximately 2 years, we found that their mutation frequency was over an order of magnitude above that in the corresponding primary fibroblasts (Fig. S4A). These cells had also developed aneuploidy and increased copy numbers (Fig. S5 A-C). Following UV irradiation, there was a greater than 8-fold

reduction in subclonal mutation frequency (Fig. S4A, B), in sharp contrast to our results with primary cells. We attribute this reduction to UV damage-induced "bottle-necking" of the population, resulting in a reduction in the population's subclonal mutation frequency. Such high mutation frequencies represent a caution in the use of immortalized cells for mutagenesis studies. While some reports claim that hTERT-immortalization is non-mutagenic and maintains diploidy during extended culture [37, 38], our observations, and those of others [39-41], suggest instead that continued *in vitro* proliferation under ambient oxygen can itself be mutagenic.

Confirming previous reports [21], our Duplex Sequencing analysis of XP-C primary cells revealed increased UV-specific mutations after UVC irradiation, relative to normal primary cells. This elevated UV-induced mutagenesis occurred primarily in inactive genes, as evident from the greater than 2-fold increase in C:G→T:A mutations in inactive genes versus active genes, consistent with defective GGR (Table 2, Fig. S3 C). The bias between the template (transcribed) and coding (non-transcribed) strands in XP-C cells was similar to that of the normal cells (Table 2, Fig. S3F) indicating that, despite the deficiency in GGR, TCR of the template strand is unaffected.

In contrast to XP-C primary cells, CS-B primary cells showed no increase in UV-specific mutations following UVC irradiation, relative to normal primary cells (Fig. 3D), despite having a survival profile akin to XP-C cells (Fig. 1C). Similar to normal cells, CS-B fibroblasts showed no bias in C:G→T:A mutation accumulation between active and inactive genes (Table 2, Fig. S3A and B), consistent with proficient GGR. However, in contrast to both normal and XP-C primary fibroblasts, both of which had reduced C→T mutations in the template strand of active genes, relative to the coding strand, CS-B cells had increased C→T mutations in the template strand (Table 2, Fig. S3D-F). This bias is consistent with defective TCR in CS-B primary cells.

An interesting observation in our in-depth spectrum studies is that, in contrast to normal and XP-C cells, CS-B primary fibroblasts accumulated more G:C→T:A mutations than C:G→T:A mutations upon UVC irradiation (Fig. 3F, Fig. S3B). Since G:C→T:A mutations are a signature of mutagenesis

induced by 8-oxo-dG, the most common oxidative lesion in cells [42], this observation alludes to CS-B's additional role in oxidative DNA damage repair [29-31], loss of which could result in increased oxidative damage-induced mutagenesis. Given the neurological involvement in CS, further studies on the mutagenic consequences of oxidative DNA damages may be worthwhile in understanding the pathologies seen in CS.

In addition to the gene- and strand-specific analyses afforded us by our Duplex Sequencing approach, we gained greater insight into the mutational consequences of UVC-induced damage, beyond that of C:G→T:A mutations at Py-Py sites. Specifically, we observed numerous types of multiplet mutations (Table 3). These included the "classic" signature of UV-induced mutagenesis, CC→TT, and also extended to triplets, such as CTC→TTT, and doublet mutations spaced 3 to 7 nucleotides apart. These multiplet mutations are likely the consequence of error-prone bypass polymerization during translesion synthesis and are consistent with the processivity of bypass polymerases persisting for several nucleotides after the bypass-requiring blocking lesion [43-45].

Our results reveal that UV-induced mutagenesis in CS cells is no higher than in normal cells (Figs 3D, S2). In normal individuals, the average age of skin cancer incidence is 55 years [12], 33 years beyond the average lifespan of a CS patient and, indeed, even 24 years longer than the longest lived CS patient on record (31 years, [46]). Although, increased exposure to sunlight or use of tanning beds in normal individuals can result in much earlier diagnosis of skin cancers (early in the 3rd decade of life) [47, 48], this is still a decade longer than the average lifespan of a CS patient. Thus, if CS cells accumulate mutations in response to UVC at the same rate as normal cells, CS patients simply do not live long enough to develop cancer. Furthermore, when analyzing UV-specific mutations, plotted relative to survival, it appears that UV-induced mutagenesis might even be lower in CS cells than in normal cells (Fig S2). This suggests that, even if CS patients could attain normal lifespans, they might never get cancer; TCR deficiency may even be protective against UVC-induced mutagenesis. Further

experiments examining normal versus CS cells would need to be conducted in order to determine if CS mutation frequencies are indeed lower than normal.

Our results are consistent with the presence of increased photosensitivity and the absence of skin cancer in CS. The question arises of why reduced repair of the template strand increases cell killing by UV damage but fails to increase mutagenesis. Since mutations need to be fixed by DNA replication across a photoproduct site, one possibility is that collisions between advancing replication forks and an arrested transcription complex trigger apoptosis in replicating CS cells [49]. A role for DNA replication arrest in CS cells has long been known [50, 51] and has been advanced as an apoptosis factor in CS [52, 53]. Noncanonical activation of ATM [54] may be the cause of such delay in DNA replication [50, 51], limiting mutagenic bypass of lesions. Furthermore, prolonged transcription blockage has itself been shown to trigger p53-independent apoptosis, mediated by the JNK pathway, in TCR-deficient cells [55]. Additionally, since TCR plays an important role in processing R-loops [56], the TCR deficiency in CS may result in elevated R-loop levels, posing persistent blocks to RNA polymerase progression and inducing apoptosis, further exacerbating cytotoxicity.

Interpreting our mutagenesis results in this light, we propose the following potential explanation for the divergent phenotypes of XP-C and CS despite their similar molecular defects. In XP-C, cells not killed by UV have increased mutations and undergo gene expression changes [57]. Consequently, genes become expressed which have not previously been subject to TCR, and likely harbor mutations due to error-prone bypass of UV-induced lesions during replication. The clonal expansion of these mutated cells (42) leads to the development of skin cancer. In contrast, in CS, cells not killed by UV have little if any increase in mutation load. The expansion of these non-mutated cells, to replace the cells lost to the cytotoxic effects of UV, does not lead to skin cancer, but, in contrast, may even decrease the spectrum of observed mutational heterogeneity.

Given that the phenotype of CS patients is weighted towards terminally differentiated cells and tissues with limited replicative potential, we can extend our explanation further. Because terminally differentiated cells attenuate GGR [58], there is no "backup" pathway for the defective TCR of CS cells. As such, bulky lesions and unresolved R loops induce apoptosis via irreversible transcription block. The few replicative cells in these tissues undergo replicative exhaustion [59], due to having to replace apoptotic cells more rapidly than in normal individuals, resulting in tissue senescence and contributing the neurodegeneration and progeria features seen in CS.

In conclusion, we have determined that, in human cells, defects in TCR fail to increase UV-induced mutagenesis as defects in GGR do. Thus, CS patients, defective in TCR, fail to develop cancer because they do not accumulate mutations more quickly than repair-proficient individuals.

METHODS

Normal, XP-C, and CS-A and –B human fibroblasts were obtained from the Coriell Institute in Camden, NJ (Table S1). One XP-C culture (XP226BA) was derived from discarded tissue after cancer surgery of patients in Guatemala [60]. The fibroblast culture NHF-D was a gift from D.Oh UCSF. A culture of pooled neonatal keratinocytes was developed in house. One normal (GM05659T) and one CS-B (GM01428T) culture were transfected with lentivirus expressing hTERT and grown continuously for at least 2 years.

To measure survival, cells were grown for 48hr in 96-well plates, drained of media, and then exposed to a range of doses of UVC (254nm) or UVB (280-320nm) using a battery of 5 fluorescent tubes for each wavelength. The UVB lamps were filtered to remove UVC. The plates were opaque to UVC, but additional shielding was used for UVB. Cells were then allowed to grow for 5-7 days and then harvested. Survival was measured colorimetrically with MTT (3-(4,5-Dimethylthiazol-2-yl)-2,5-diphenyltetrazolium bromide, Sigma-Aldrich, St Louis, MO) at 570nm. Relative survival was calculated from the ratios of exposed to unexposed wells, based on the average 570nm absorbance

in 4 to 6 wells per exposure condition. We chose to measure the survival at 5-7 days, which corresponded to the time of harvest for our mutagenesis analysis. The surviving cell numbers represent a combination of cell lysis, growth delays, and rates of regrowth.

To measure UV-induced mutagenesis, cultures of approximately 10^7 cells were washed in PBS, irradiated and grown for 5-7 days. Cells were then harvested by trypsin, washed in PBS and rapidly frozen in dry ice/methanol. DNA was isolated and mutations measured via one round of Duplex Sequencing [28], more fully described in the Supplementary Methods. Target genes were exonic regions of *NRAS*, *UMPS*, *PIK3CA*, *EGFR*, *BRAF*, *KRAS*, *F10*, *TP53*, and *TYMS*, several of which were chosen for their importance in skin carcinogenesis (Table 1). We required a minimum depth of 100 Duplex molecules to call a position, either mutant or not; all samples had a mid-exon peak depth of 1000-4000 Duplex molecules across all captured exons.

Acknowledgements.

Work supported by grants CA181771 and CA077852 (LAL and KSR-B) and the Academic Senate University of California San Francisco and the Lily Drake Cancer Research Fund of the University of San Francisco (JEC). We are grateful to D.Oh for the gift of NHF-D, and to D. Karentz University of San Francisco for advice and support. Work with human cells was approved by the UCSF Committee on Human Research IRB11-05993 (JEC).

Table 1. Genes captured for analyses by Duplex Sequencing

Gene name	Orientation	Transcription status[1]
NRAS	Reverse	Active
KRAS	Reverse	Active
BRAF	Reverse	Inactive
TP53	Reverse	Active
EGFR	Forward	Active
UMPS	Forward	Active
F10	Forward	Active
PIK3CA	Forward	Inactive
TYMS	Forward	Inactive

Table 2. Ratio of pooled C:G→T:A and C→T mutations after UV irradiation, relative to controls[1]

	Genotypes		
Gene expression status [2]	**Normal**	**CS-B**	**XPC**
Expressed	2.9	6.0	3.4
Not expressed	2.1	5.5	7.4
Strand of active genes, relative to transcription [3]			
Nontranscribed (coding)	5.8	4.2	5.3
Transcribed (template)	2.5	7.7	2.9

1. Since these ratios are calculated relative to controls of the same genotype the absolute numbers are cell-type dependent. The ratios should be compared according to gene activity or strand specificity for each cell type independently
2. C:G→T:A mutations at C:G basepairs
3. C→T mutations in active genes at cytosines

Table 3. UVC-induced multiplet mutations in normal, XP-C, CS-A, and CS-B fibroblasts.

	Active genes		Inactive genes	
	Exonic	**Intronic**	**Exonic**	**Intronic**
Normal, adult	CC→TT	CC→TT GAA→AAC AC→GT	AC→TT	
Normal, neonatal	CC→TT CC→TT		C→T and A→C in same read	T→C and C→T in same read
XPC(1)		G→T 7nt apart AC→TT TTT→ATC CCC→GCT CC→TT CC→TT AC→TT GTC→TTT CTC→TTT TC→AA CA→AG	GA→AT TG→CT	CTT→TTC
XPC(2)	CTC→TTT CTC→TTT CTC→TTA	AC→TG CC→AT A→T and A→G in same read AC→CT	ATC→TTT CAC→TTA AT→TA	
CSA(2)	C→G and C→G in same read	CA→AG	CC→TT	
CS-B(1)		CC→TT		
CS-B(2)	AA→GG C→T 6nt apart			

Figures

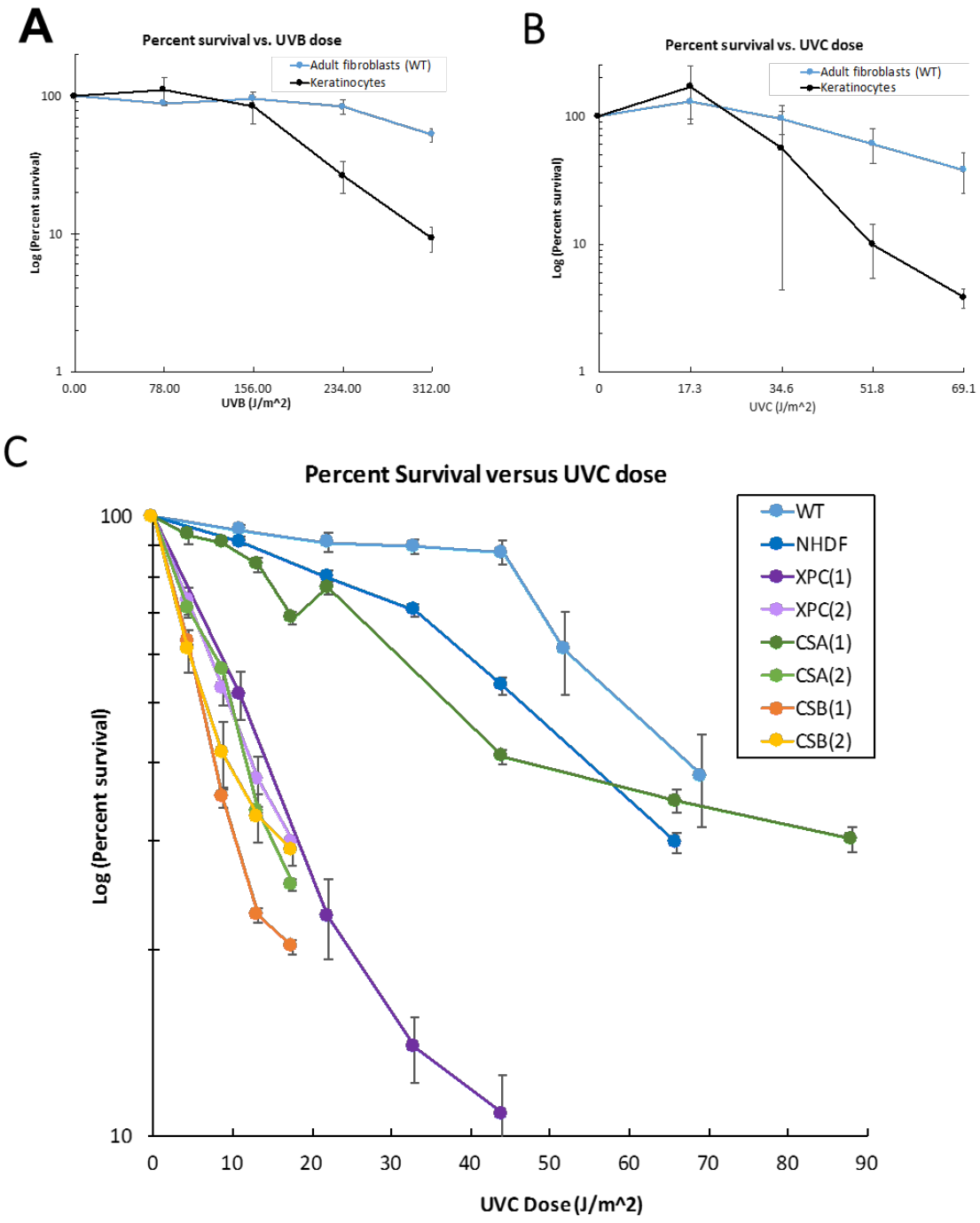


Figure 1. UV-induced cytotoxicity of primary fibroblasts and keratinocytes. A and B: Survival of normal fibroblasts (GM05659) and keratinocytes treated with UVB (A) and UVC (B). C: Survival of normal adult (WT), normal neonatal (NHF-D), XP-C (XPC(1) and (2)), CS-A (CSA(2)), CS-B (CS-B(1) and (2)), and undetermined (originally designated CS-A (CSA(1)) fibroblasts. Error bars represent standard deviation of two survival determinations.

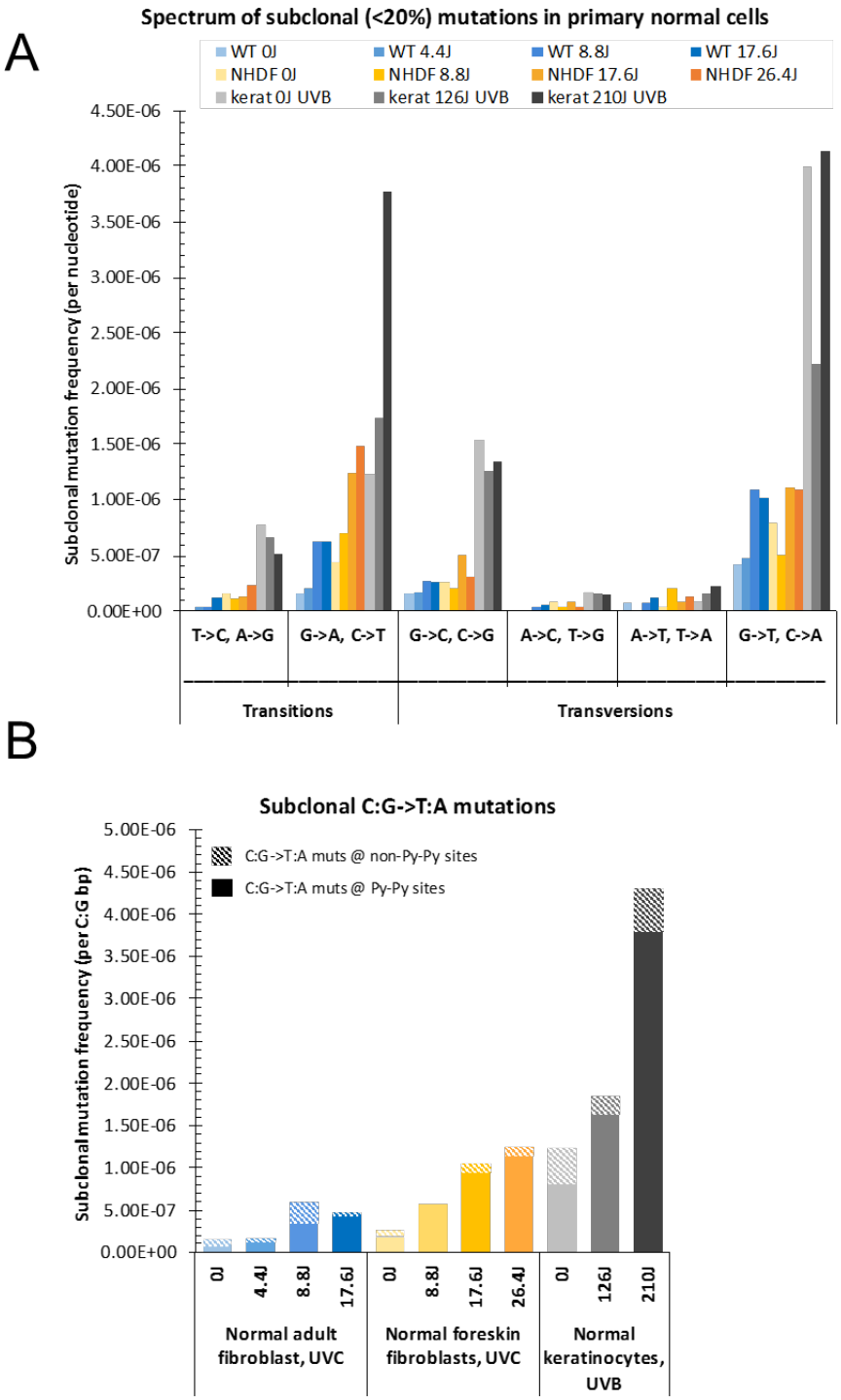


Figure 2. UV induces unselected subclonal (<20% clonal) mutations in normal primary fibroblasts and keratinocytes. A: Spectrum of subclonal mutations in adult (WT) and neonatal (NHF-D) fibroblasts treated with UVC and in keratinocytes (kerat) treated with UVB. B: Subclonal frequencies of UV-specific mutations in adult and neonatal fibroblasts treated with UVC and in keratinocytes treated with UVB; solid bars represent UV-specific mutations (C:G→T:A mutation at Py-Py sites); hashed bars represent C:G→T:A mutations at non-Py-Py sites. Frequencies were calculated by dividing the number of mutations of each type by the number of times the wildtype base of each mutation type was sequenced.

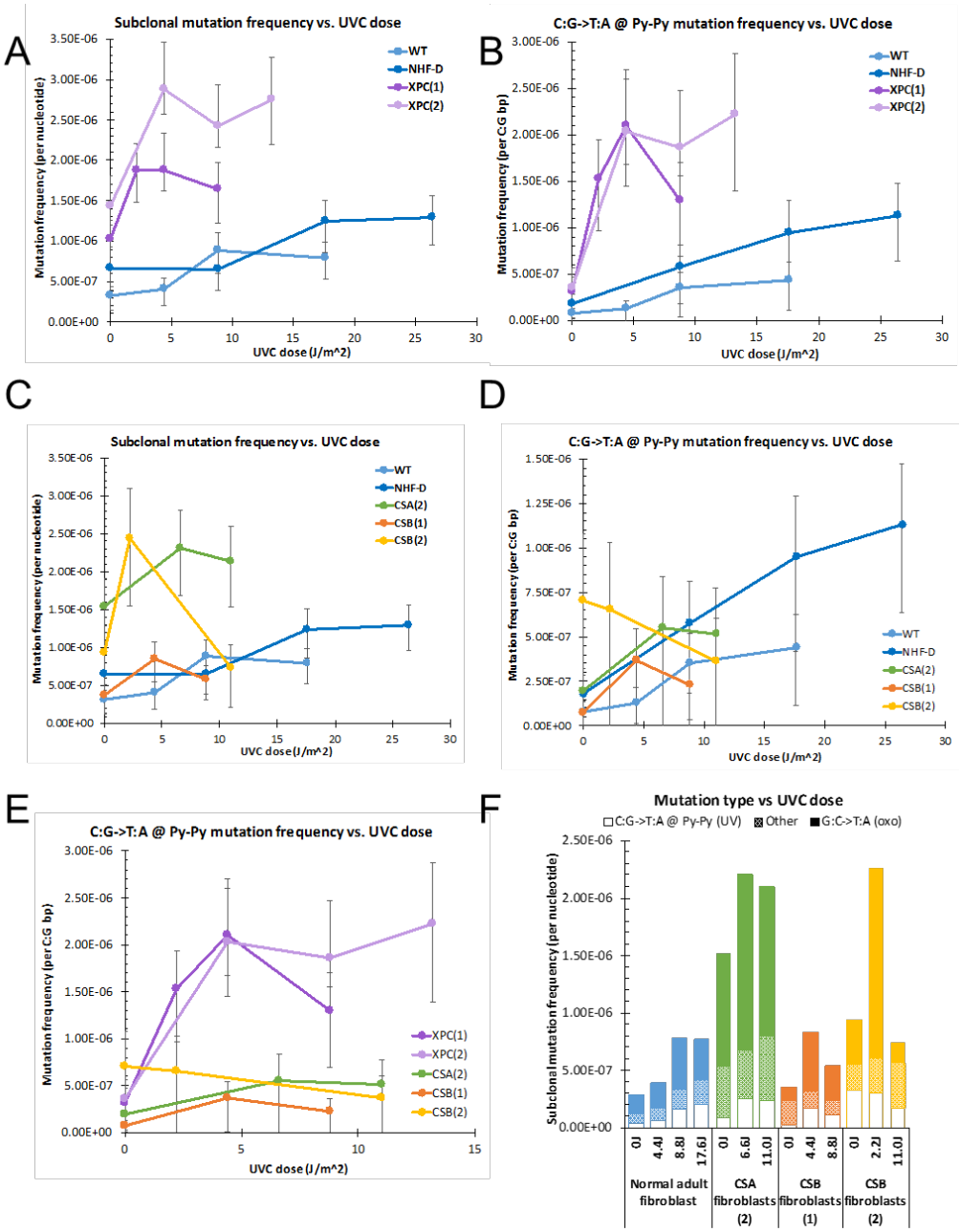


Figure 3. UVC induces increased UV-specific mutations in primary XP-C cells, relative to primary normal cells, but not in primary CS cells. A and B: Frequency of all subclonal (<20% clonal) mutations (A) and UV-specific mutations (B) in normal adult (WT) and neonatal (NHF-D) primary fibroblasts and in XP-C (XPC(1) and (2)) primary fibroblasts. C and D: Frequency of all subclonal mutations (C) and UV-specific mutations (D) in normal adult (WT) and neonatal (NHF-D) primary fibroblasts and in CS-A (CSA(2)) and CS-B (CS-B(1) and (2)) primary fibroblasts. E: UV-specific mutations in XP-C (XPC(1) and (2)) and CS-A (CSA(2)) and CS-B (CS-B(1) and (2)) primary fibroblasts. F: Subclonal frequencies of UV-specific mutations, oxidative-signature mutations, and all other mutations in primary neonatal (NHF-D), CS-A (CSA(2)), and CS-B (CS-B(1) and (2)) fibroblasts; open bars represent UV-specific mutations (C:G→T:A mutation at Py-Py sites); solid bars represent oxidative-signature mutations (G:C→T:A); hashed bars represent all other mutations. Frequencies were calculated by dividing the number of mutations of each type by the number of times the wildtype base of each mutation type was sequenced. Error bars represent 95% confidence intervals calculated from Wilson scores of the mutation frequency for each sample.

SUPPLEMENTARY INFORMATION

Cell cultures and identification of GM17536

The UV sensitivity and other characteristics of the cells were determined from survival data (Fig 1) and published data (https://catalog.coriell.org) (Table S1) [60, 61].

The source patient for the GM17536 cell line was originally described as "atypical Cockayne" and showed many of characteristics associated with the CS phenotype, including developmental delay, short lifespan, and ganglial calcifications, but little photosensitivity. The cell line was reported to show reduced reactivation of UV-damaged luciferase plasmids (https://catalog.coriell.org). Although it was originally classified as CS-A by the donor, this was by default, based on the exclusion of CS-B mutations and all groups of XP; it has since been reclassified as uncertain due to our observations.

To determine whether GM17536 cells have a significant sensitivity to UV, we determined its survival over a higher UVC dose range than that used for our mutagenesis analyses. GM17536 showed no increased UVC-induced cytotoxicity, relative to normal cells (Fig S1A). We also determined the sensitivity of GM17536 to illudin S, which is uniquely toxic for cells that lack TCR [62], and found it to show no more sensitivity than TCR-proficient normal (GM03440) and XP-C (XP226BA) cells (Fig S1B). In contrast, CS-A (GM01856) and CS-B (GM01428) cells were very sensitive to illudin, consistent with their TCR deficiencies. Additionally, the induction of subclonal UV-specific mutations (C:G→T:A at Py-Py sites) in GM17536 was within the normal cell range (Fig S1C).

The GM17536 cell line may, therefore, possess a mutation in another gene peripherally involved in CS that remains to be identified. Exome sequencing is planned to resolve this issue. It should be noted that photosensitivity is not always a hallmark of CS [15, 33], which is predominantly a developmental and neurological disease.

Duplex sequencing procedures and data analysis

To measure UV-induced mutations, we carried out Duplex Sequencing library preparations, as previously described [28] with the following minor modifications: (i) 3ug of total DNA was sheared using the Covaris AFA system with a duty cycle of 10%, intensity of 5, cycles/burst 100, time 20 seconds × 5, temperature of 4°C; (ii) prior to adapter ligation, the DNA was quantified and a 20 to 1 molar excess of adapters was used for the ligation step; (iii) after adapter ligation and clean up, the library was re-quantified and 1200 fmol of product was amplified by PCR for 16-18 cycles. The resulting libraries were then subjected to two sequential rounds of gene capture with 120 oligonucleotide probes against the exomes of *NRAS*, *UMPS*, *PIK3CA*, *EGFR*, *BRAF*, *KRAS*, *F10*, *TP53*, and *TYMS*, several of which were chosen due to their importance in skin carcinogenesis. The final libraries were then sequenced on an Illumina HiSeq 2000/2500 platform using 101bp paired-end reads.

The overall sequencing pipeline has been previously described [28]. Briefly, all raw sequencing reads are filtered based on the location of an expected fixed sequence within each read, as well as the requirement that each position of the 12 nucleotide random duplex tag sequence only contain one of the four canonical bases. Any reads not conforming to these criteria are discarded. The random tag sequences from each read pair are computationally concatenated into a 24 nucleotide random sequence and appended to the header of each read-pair. In order to remove potential artifactual mutations arising from the end-repair and ligation reactions, each read is trimmed by four additional bases. The reads are then aligned against build 19 of the human genome using the Burrows-Wheeler Aligner (BWA) [63]. Reads not aligning to the human genome are filtered out. A consensus sequence (SSCS) for each tag family (e.g. reads sharing identical tag sequences) is computationally determined using software written in-house. The consensus for any position in a read is considered undefined if the position is represented by fewer than three instances in the family or if less than 70% of the sequences at that position in the read are in agreement. The SSCS reads are then realigned to build 19 of the human genome using BWA. After filtering for unmapped reads,

duplex consensus (DCS) reads are constructed by pairing SSCS with their respective strand-mates by grouping the 24 nucleotide tag in read 1 with the appropriate 24 nucleotide tag in read 2 and the sequence identity at each position of the two reads are compared to one another. The sequence information is kept only if the base identity of both reads is identical. Next, to remove alignment artifacts common at the ends of reads, 10 nucleotides from each end of the duplex reads are soft-clipped using the Genome Analysis Toolkit. Finally, any nucleotides overlapping between the forward and reverse reads of the DCS families are clipped to prevent double-counting. Scripts to calculate mutation frequencies and locations, written in-house, are available upon request. Due to the high redundancy of the sequencing data in Duplex Sequencing, relatively few "genome equivalents" are sequenced (~1000-2000 genomes/sample); thus, our ability to detect deletions and rearrangements is limited and were, therefore, not evaluated in this study.

Subclonal mutation frequencies were initially plotted against UV dose to indicate the overall yield for each cell type (Fig 3). When we plotted subclonal mutation frequencies versus cell survival, the results suggested that CS cells might have an even lower mutation frequency than normal cells at equivalent survival (Fig S2). This possibility warrants further study.

We determined the distribution of mutations between active and inactive genes, based on the gene expression status of each gene in skin (Table 1, GeneCards, http://www.genecards.org/) and the distribution between transcribed and nontranscribed strands. This was done by pooling controls and all UV doses for each cell type (Table 2, Fig S3). In normal foreskin fibroblasts (NHF-D), the increase in C:G→T:A mutations in active genes was similar between active and inactive genes (Table 2, 2.9-fold increase over control versus 2.1-fold increase, respectively). This implies that, in normal cells, there is no bias in mutagenic processes between active and inactive genes (Fig. S3). CS cells showed a similar relationship between active and inactive genes (6.0-fold increase in C:G→T:A mutations in active versus 5.5-fold increase in inactive genes) (Fig. S3). In contrast, XP-C cells showed a 3.4-fold increase in C:G→T:A mutations in active genes and a 7.4-fold increase in C:G→T:A

mutations in inactive genes (Fig. S3). This increased mutagenesis in inactive genes of XP-C cells, relative to their active genes, is consistent with their defective GGR but still proficient TCR.

Next, we determined the distribution of mutations between the template (transcribed) strand and coding (non-transcribed) strand (Table 2, Fig S3 D-F). The strand distribution of mutations in normal cells showed that C:G→T:A mutations increased 5.8-fold on the coding strand versus only 2.5-fold on the template strand, consistent with proficient TCR of the template strand (Fig. S3 D-F). CS cells showed a similar increase in the coding strand (4.2-fold); however, the template strand of CS cells showed a 7.7-fold increase in C:G→T:A mutations. This is consistent with defective TCR of the template strand of active genes in CS cells. In contrast, XP-C cells were similar to normal cells, showing a 5.3-fold increase in C:G→T:A mutations in the coding strand versus a 2.9-fold increase in the template strand, consistent with functional TCR of the template strand (Fig. S3 D-F).

Mutation in hTERT cells – hormesis or repair?

Normal (GM05659) and CS-B (GM01428) cells were transfected with hTERT by lentivirus (pBabePuro-hTERT) and maintained in continuous culture for over 2 years. They maintained their relative UV sensitivities, the GM01428T line being much more UV sensitive compared to GM05659T (50% survival doses of 4.0 $J.m^{-2}$ and 25.5 $J.m^{-2}$ respectively). These cells were assayed for mutations following UV irradiation using Duplex Sequencing with a capture set against the exomic regions of *DNMT3a*, *IDH1*, *IDH2*, *TET2*, *NPM1*, *C/EBPa*, *PTPN11*, *c-KIT*, *BRAF*, *RUNX1*, *Hbb*, *NRAS*, *KRAS*, *TP53*, *WT1*, *U2AF1*, and *POLD1*. The immortalized normal and CS-B unirradiated cells both had a very high yield of mutations (Fig S4 A-C), several fold above the corresponding primary fibroblasts.

In contrast to our observation in primary fibroblasts, upon UV exposure, hTERT-immortalized normal and CS-B cells underwent an approximately 10-fold reduction in mutation frequency (Fig S4A), with no new mutations observed in the UV-treated cells. Rather, both normal and CS-B cells showed loss of subclones and normalization of clonal mutations to heterozygotes or

homozygotes (50% clonality and 100% clonality, respectively), consistent with UV-induced selection of the culture (Fig. S4B).

Immunohistochemistry analyses with a series of probes revealed that these cells were aneuploid and had increased copy numbers at several loci (Fig S5). Other groups have previously observed similar genomic instability in hTERT positive cells grown for extensive periods of time [39-41].

Similar decreases in mutations in response to low level radiation have previously been reported and termed "hormesis" [64]. Often the term has been used to imply that radiation elicits a repair process that mitigates the initial impact of radiation [64]. The concept has even been extrapolated to imply that radiation standards should be set at levels to take advantage of a supposed health benefit of low radiation levels [64]. Our studies provide the advantage of distinguishing between mutations and genomic instability initially present in the population of cells from the mutations induced by UV. Our observations clearly show that the observed reduction in mutations is a result of selection against cells with pre-existing mutations (Fig S4B) and not the consequence of *de novo* repair and mutagenesis. If these results can be extended to studies with ionizing radiation then the idea of a health benefit needs to be reconsidered, as our results suggest that cells already carrying a significant load of pre-existing mutations may be selectively vulnerable to further radiation-induced toxicity.

Table S1. Cells used for determination of UV survival and mutagenesis[1]

Genotype	Study Designation	Coriell Institute	Alternate	37% dose[2]	Mutations
Normal fibrobasts	WT(1)	GM05659		>44	+/+
	Wt	GM03440		>44	+/+
	NHF-D			>44	+/+
Normal keratinocytes	Kerat			32	+/+
XP-C	XPC(1)	GM02997	XP7CA	15.4	unavailable
	XPC(2)	XP226BA	XP226BA	13.4, 12.8	490delC homozygote
CS-A	CSA(2)	GM01856	CS3BE	12.3, 16.8	37G>T; 479C>T
CS-B	CSB(1)	GM01428	CS7SE	11.9	unavailable
	CSB(2)	GM01629	CS1BE	11, 13	Exon10 (2087C>T); exon 18 (3615delA)
Uncertain	CSA(1)	GM17536	CS210BE	Not UV sensitive	Assignment uncertain, not XP or CSB

1. Cell lines designated GMxxxxx were obtained from the NIGMS Human Genetic Cell Repository at the Coriell Institute for Medical Research. Normal neonatal fibroblasts NHF-D were from a single donor and donated by D.Oh UCSF; the normal neonatal keratinocytes were a pooled culture obtained in-house under CHR permit. XP226BA were developed in-house from discarded tissue from a cancer surgery [60].
2. Dose in $J.m^{-2}$ UVC

Supplementary figures

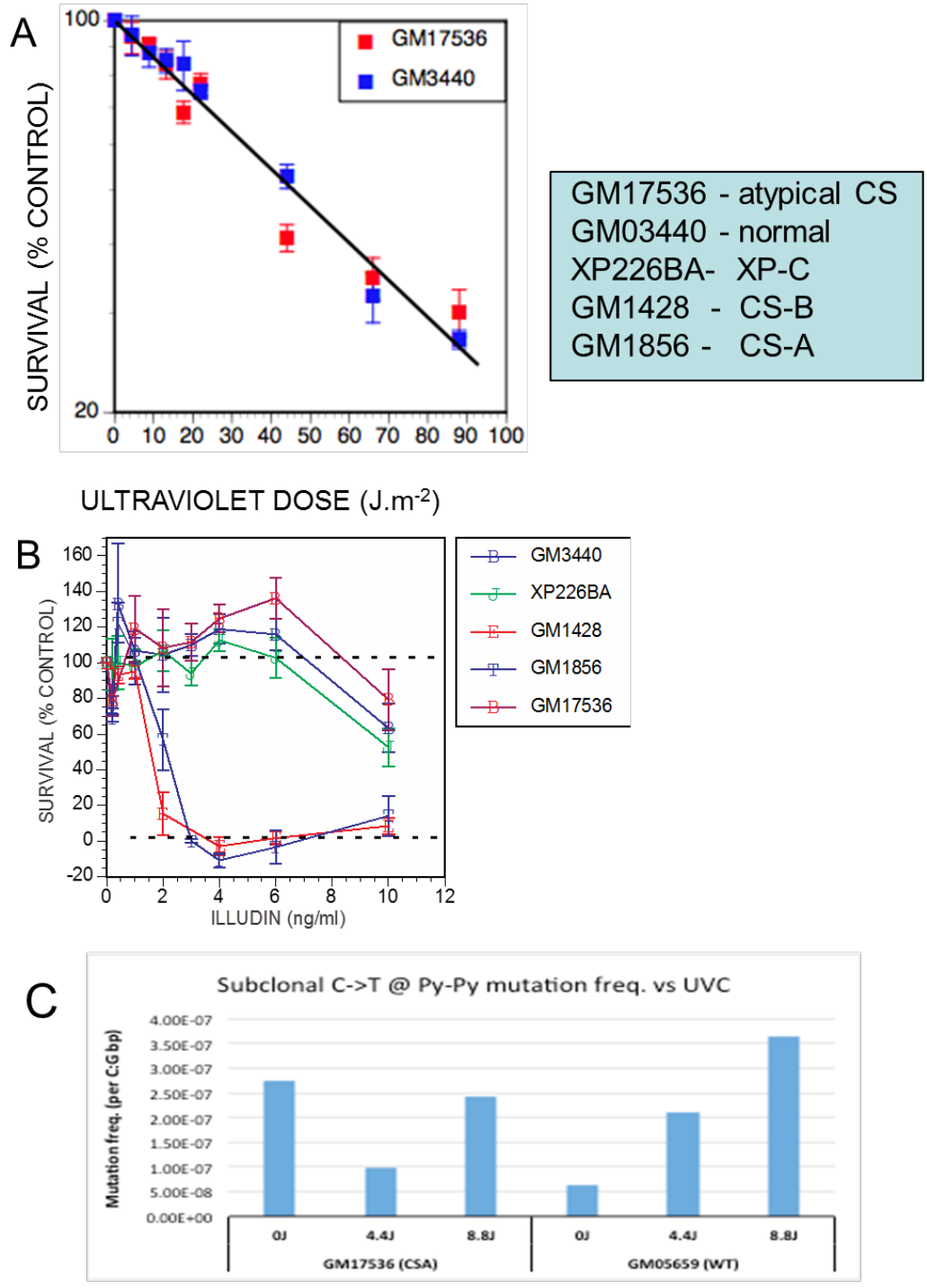


Figure S1. GM17536 shows no difference from normal when exposed to UV or illudin S. A: Survival of GM17536 (blue squares) and normal (red squares) primary fibroblasts exposed to increasing UV doses. B: Survival of GM17536, XP-C (XP226BA), CS-A (GM1856), CS-B (GM01428), and normal (GM03440) primary fibroblasts exposed to increasing concentrations of illudin S. C: Subclonal UV-specific mutations (C:G→T:A at Py-Py sites) in GM17536 and normal (GM05659) primary fibroblasts.

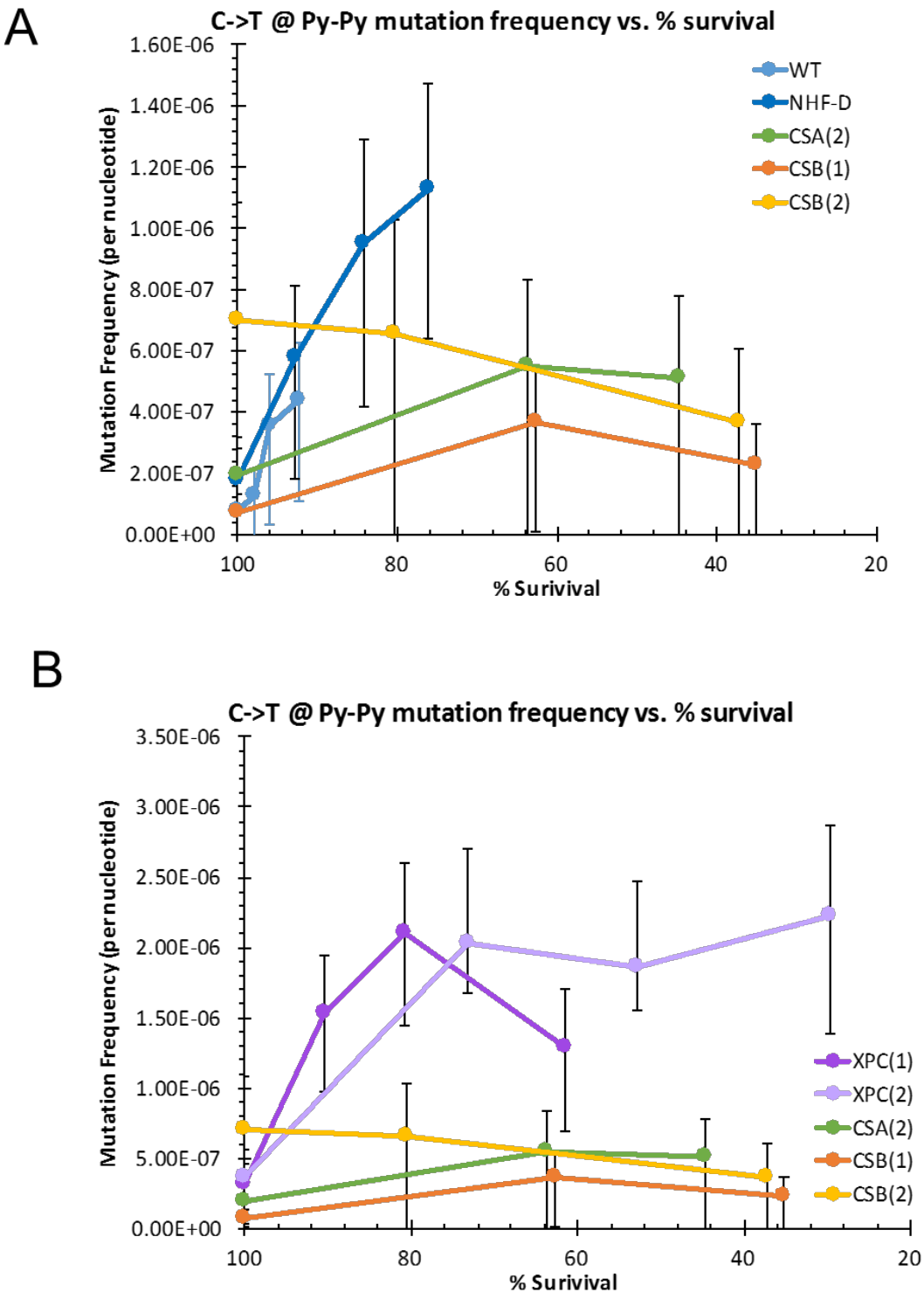


Figure S2. UVC-induced mutation frequencies versus survival in normal and CS-A and CS-B primary fibroblasts. A and B: C:G→T:A (A) and C:G→T:A mutations at CC:GG dinucleotides (B) in normal adult (WT(1)) and neonatal (NHF-D) fibroblasts and in CS-A (CSA(2)) and CS-B (CSB(1) and (2)) fibroblasts. Percent survivals were derived from cell survivals in Fig. 1C. Frequencies were calculated by dividing the number of mutations of each type by the number of times the wildtype base of each mutation type was sequenced.

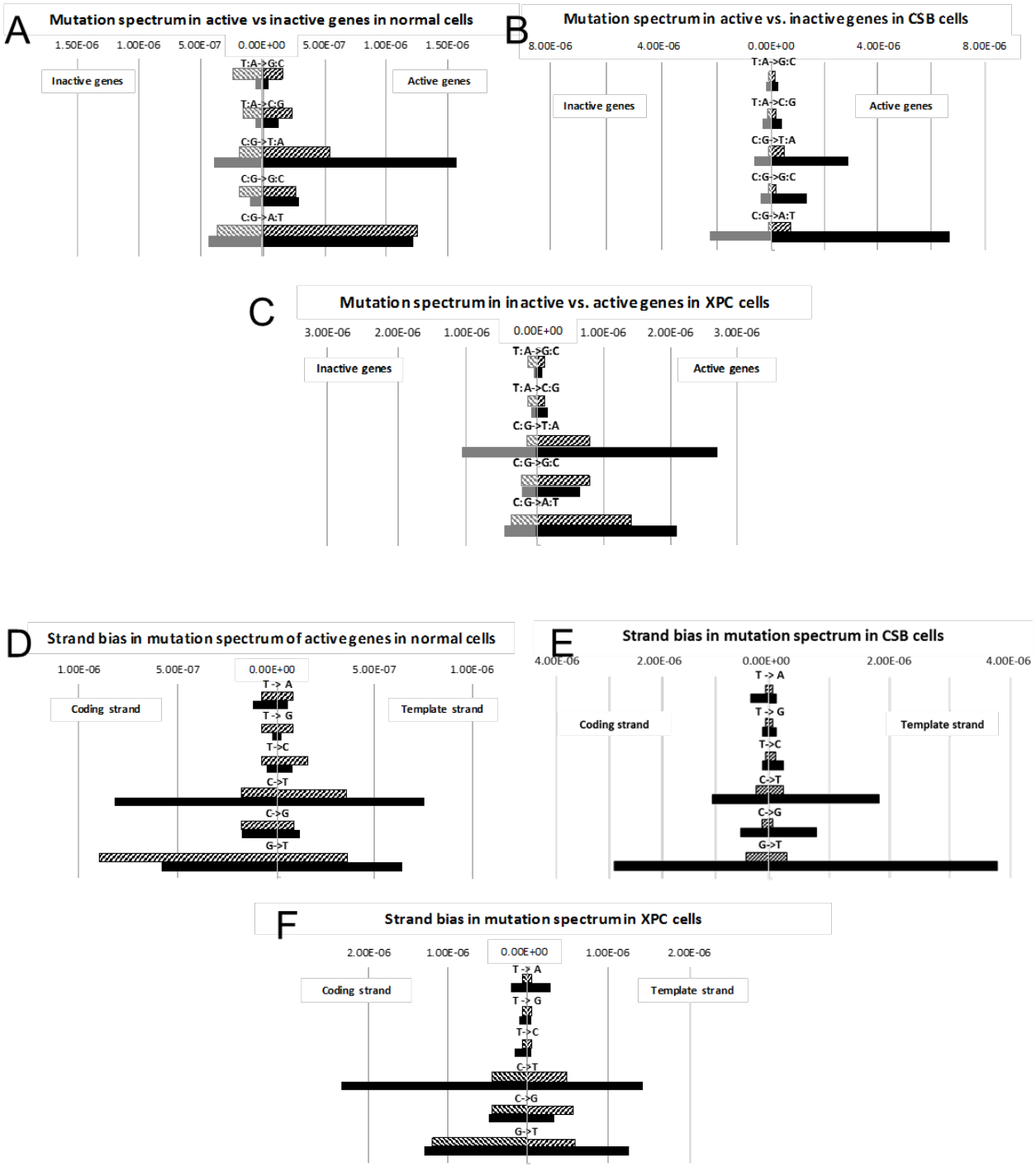
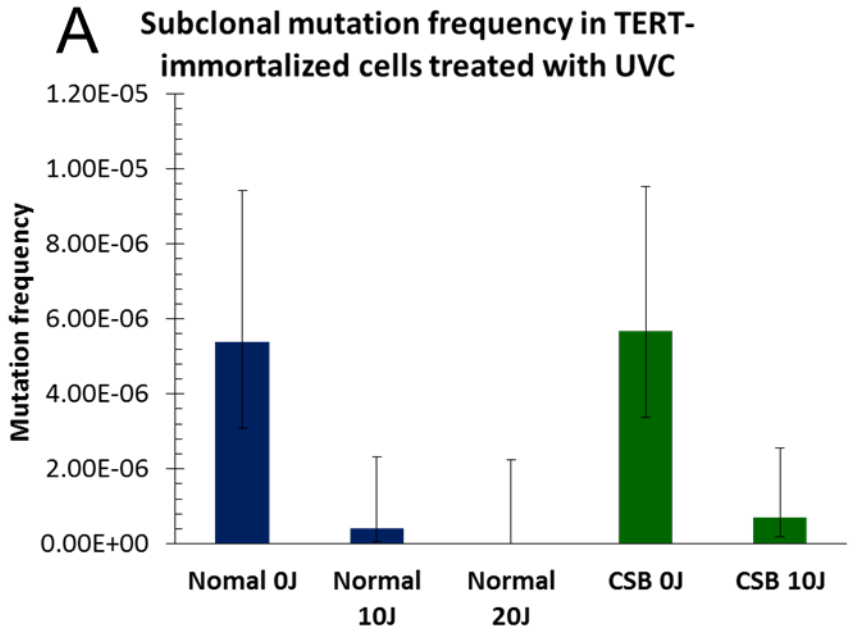


Figure S3. UVC-induced mutations accumulate preferentially in the inactive genes of XP cells and in the coding strand of the active genes of CS cells. A, B, and C: Mutation spectrum in active (black) versus inactive genes (grey) in primary normal neonatal (A), CS-B (B), and in XP-C fibroblasts. Hashed bars indicate mutations frequencies in control groups; solid bars indicate mutation frequencies in pooled UV-treated groups. D, E, F: Mutation spectrum in the coding (non-transcribed) strand versus the template (transcribed) strand primary normal neonatal (D), CS-B (E), and in XP-C fibroblasts (F). Frequencies were calculated by dividing the number of mutations of each type by the number of times the wildtype base of each mutation type was sequenced. Hashed bars indicate mutation frequencies in active genes of control groups; solid bars indicate mutation frequencies in active genes of pooled UV-treated groups.



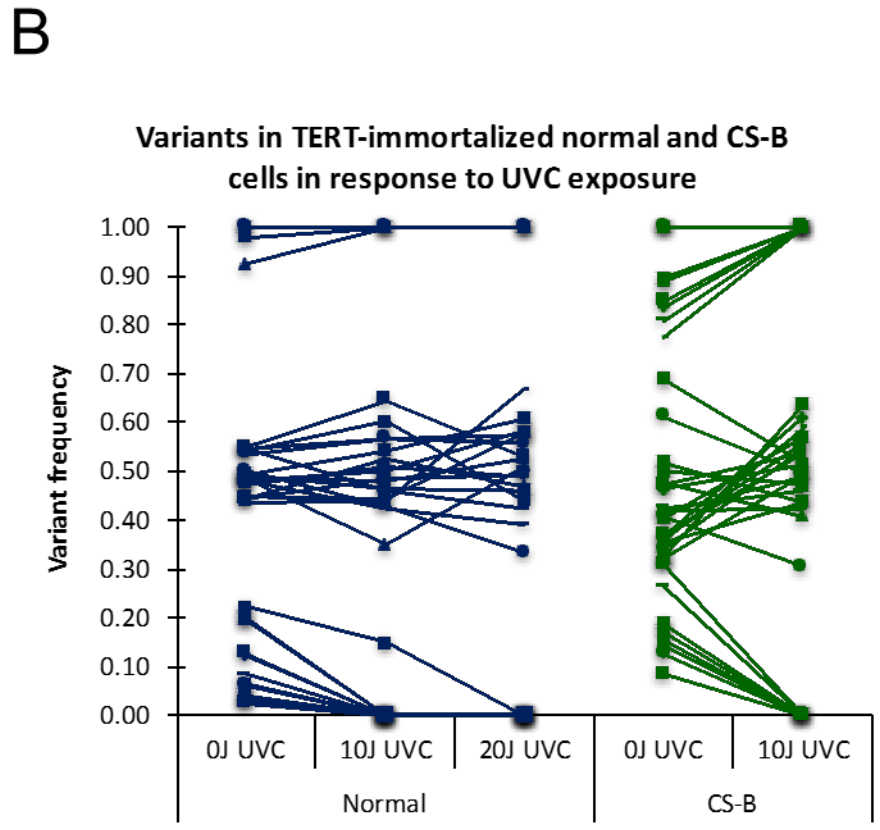
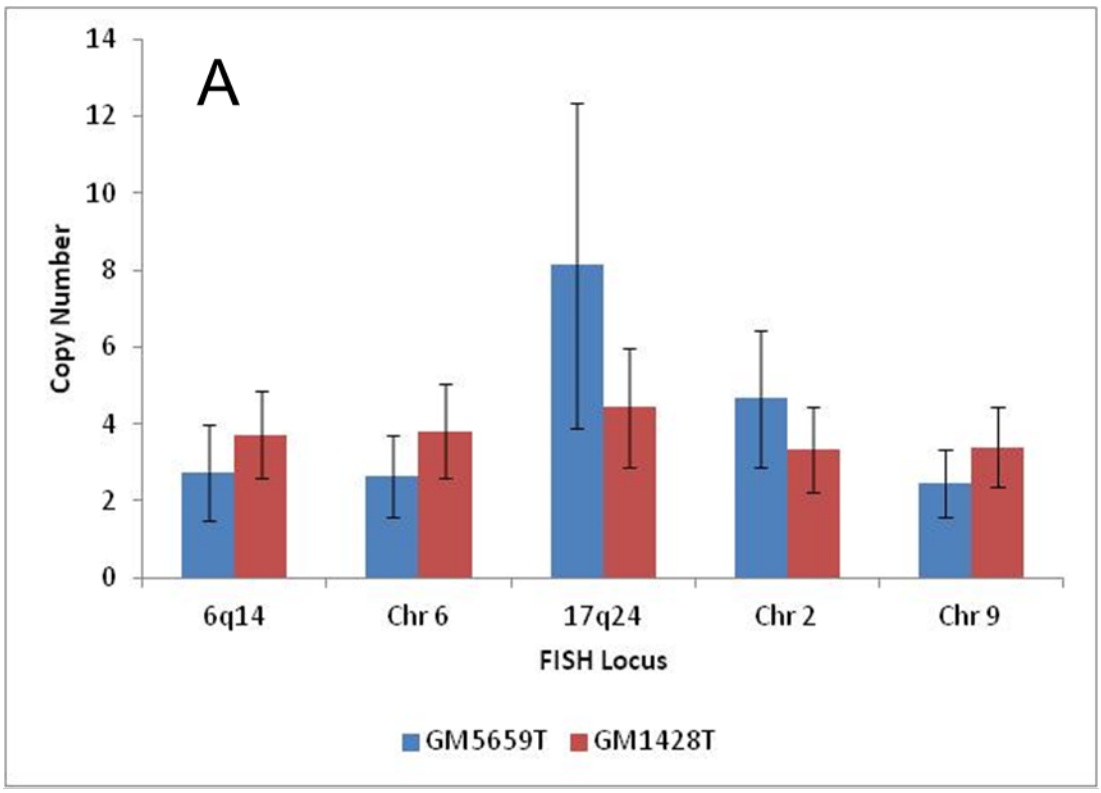
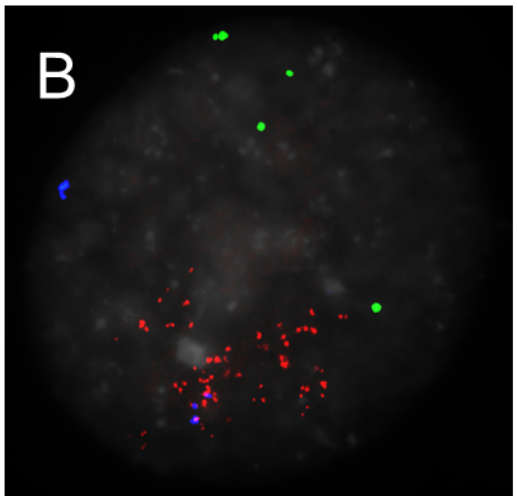


Figure S4. UV-induced mutation frequency changes in hTERT-immortalized cells. A: High mutation frequencies in untreated hTERT-immortalized normal (navy) and CS-B (green) and reduction by UV exposure. B: Reduction in subclonal variant frequencies with UV dose representing UV-induced population bottle-necking in hTERT-immortalized normal (navy) and CS-B (green) cells.





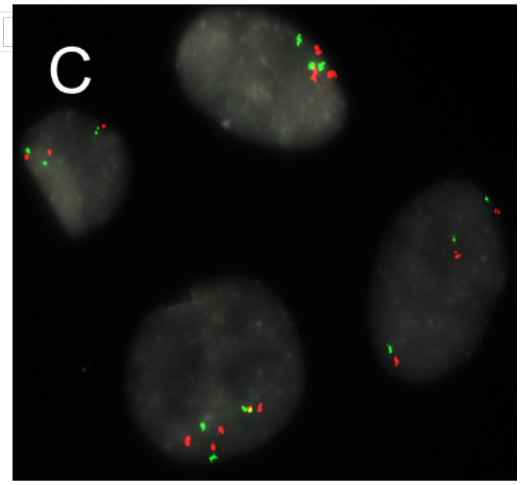


Figure S5. Gene amplification in hTERT positive cells. A: Copy number for selected genes in hTERT transfected fibroblasts that had been grown continuously for at least 2 years. Probes used are indicated in the X-axis. GM05659T (blue bars), GM01428T (red bars). Standard deviations shown. B: GM05659T cell labeled with probes for 17q24 (red), chromosome 6 (green). C: GM01429T labeled with probes for 6q14 (red) and chromosome 5 (green). Scale bar for 20 microns shown.

CHAPTER 2: SUBCLONAL MUTATIONS IN GLIOBLASTOMA

INTRODUCTION

Glioblastoma (GBM) accounts for 16% of all primary brain tumors and is the most common and aggressive brain malignancy in adults [65]. The incidence of GBM diagnosis increases with age, with an average incidence of 7.2 per 100,000, peaking at 14.64 per 100,000 between 75 and 84 years of age. Of all adult GBMs, 95% are primary, manifesting rapidly after a short clinical history and without signs of a preceding precursor lesion. The remaining 5% are secondary GBMs, which can develop from lower-grade tumors; these include diffuse astrocytoma (grade II) and anaplastic astrocytoma (grade III). Secondary GBMs are histologically indistinguishable from primary GBMs. GBM is universally incurable. Surgical resection followed by concurrent adjuvant radiotherapy and temozolamide is standard of care for newly diagnosed GBM [65-67]. While some subtypes initially respond better to treatment than others, all GBMs eventually recur, at which point few therapeutic options are available. The median survival is less than two years, with less than 5% of patients surviving longer than 5 years after initial diagnosis [66-68].

GBM is among the most highly clonally mutated cancers [69]. Analysis of clonal mutation data from The Cancer Genome Atlas (TCGA) [70], which encompasses thousands of exome sequences from 32 different cancer types to date, revealed that the average GBM contains as many as 1.5 to 32 single base clonal mutations per million basepairs. This exceptionally high clonal mutation frequency supports the long-standing hypothesis that cancers can exhibit a mutator phenotype [71, 72]. Clonal mutations in GBMs have been mapped to at least three core pathways: (i) the RTK signaling pathway involving the receptor tyrosine kinase/Ras/phosphatidylinositol-3-kinase (RTK/Ras/PI3K), (ii) the p53 pathway, and (iii) the retinoblastoma tumor suppressor pathways [70]. The recent TCGA analysis also revealed several signatures that may infer mutation origin [73]. Among the most frequent clonal mutations reported in GBM are C:G → T:A

substitutions, which can be mediated by (i) spontaneous or APOBEC/AID-induced deamination of cytidine to uracil [74-76], (ii) misincorporation by DNA polymerases [77], or (iii) depyrimidation of cytidine resulting in an abasic site that pairs with deoxyadenosine [78, 79]. Adding to this complexity, recent studies have revealed that GBMs exhibit geographical heterogeneity, in which different mutations are present in distinct regions of a tumor [80-82]. Importantly, because the TCGA analysis and studies of intratumoral heterogeneity were carried out on exome sequencing data, the mutations reported are clonal and, thus, are strongly biased by selection.

Subclonal mutations, which we define as those mutations present at less than 20% clonality, are thought to be passenger mutations, arising randomly and expanding independent of selective pressures. We hypothesize that subclonal mutations may empower malignant cells with the ability to evade host mechanisms and may serve as a reservoir for cancer progression, treatment resistance, and recurrence. The present study presents preliminary data addressing three questions concerning subclonal mutations in GBM: (i) Do subclonal mutations exhibit geographical heterogeneity in GBM? (ii) Does subclonal mutation load and/or spectrum correlate with GBM recurrence? (iii) Do pre-existing subclonal mutations undergo selection and clonal expansion upon recurrence of GBM?

RESULTS

GBM exhibits intratumoral subclonal heterogeneity. To determine if subclonal mutations differ between geographic regions of individual GBM tumors, as has been previously described for clonal mutations [80-82], we obtained two sections from distinct regions of three GBM tumors, extracted their genomic DNA, and prepared sequencing libraries via the Duplex Sequencing protocol [83]. Because of the high cost of sequencing and depth of sequencing necessary to analyze subclonal mutations by Duplex Sequencing, 13 genes (Table 1), chosen for their implication of GBM

pathogenesis and therapy resistance, were captured via two rounds of sequential capture, using biotinylated probes against the exonic regions of the genes, prior to sequencing.

The six samples (2 sections from each of 3 tumors) were analyzed for the presence of subclonal and clonal mutations. While conventional NGS has a limit of detection of 5% clonality due to artifacts resulting from errors during PCR and sequencing, Duplex Sequencing is able to detect subclonal events as low as 10^{-8}; thus, we detected numerous subclonal mutations in each of our samples (Fig. 1). In addition to the clonal mutations detected, determined by coding changes not found in dbSNP, we observed substantial variation between gene and depth of sequencing; upon further analysis, we discovered that EGFR had undergone substantial copy number increases in all of our samples; W22F and W33F had the greatest copy number increase, with 30X greater depth observed at the EGFR gene locus than all other genes analyzed, which had an average depth of ~1000-2000 nucleotides sequenced per position.

Upon analysis of subclonal (<20% clonal) variants, we observed substantial differences in mutation frequency between sections of two of the tumors, W22 and W53 (Fig. 2); in contrast, the two sections of tumor W33 had similar subclonal mutation frequencies. Additionally, the subclonal mutation spectrums (relative frequencies of each mutation type observed) differed between the two tumor sections of two of the tumors, W53 and W33, suggesting that the underlying mechanism(s) driving the expansion of these tumor may have differed between the two regionally distinct tumor sections. In contrast, tumor W22 exhibited similar subclonal mutation spectrums between its two regionally distinct sections.

Subclonal mutations in GBM show no consistent load or spectrum changes upon recurrence. Given that most GBM tumors are essentially treated the same, with surgery followed by concurrent radiation and temozolamide, we sought to determine if subclonal mutation load or spectrum correlates with GBM recurrence. To do so, we extracted and analyzed via Duplex Sequencing DNA

from 6 pairs of primary and matched recurrent GBM, capturing 13 genes implicated in GBM (Table 1). Upon analysis of subclonal mutations, differences between primary and recurrent sections were observed, however there did not appear to be a consistent pattern (Fig. 3.). Three of the six tumor pairs (pairs 1, 3, and 11) had reduced subclonal mutation frequency upon recurrence, whereas the other three pairs had increased subclonal mutation frequency upon recurrence. Additionally, examining the spectrums of the paired tumors, all six recurrent tumors exhibited a subclonal mutation spectrum distinct from its paired primary tumor. However, there was no consistent subclonal mutation spectrum change unique to the recurrent tumors.

When we more closely examined the tumor pairs, we again found no consistent changes in recurrent tumors relative to their paired primary tumors (Table 2). While all pairs shared at least some portion of the single nucleotide variants (SNVs) between the primary and recurrent tumor, the degree to which SNVs were lost or gained upon recurrence varied considerably. Some pairs had a large number of SNVs unique to the primary tumor (pairs 3 and 11), whereas others had relatively equivalent numbers of SNVs shared, unique to primary, and unique to recurrent (pairs 1, 7, and 12); pair 9 had a far greater number of shared SNVs and unique SNVs in the recurrent compared to the number of SNVs unique to the primary. Examining individual mutation changes, specifically those that are predicted to be pathogenic or may have the potential to be given the degree to which their clonality changed upon recurrence, revealed yet more inconsistencies. Six of the seven tumor pairs had at least one major coding mutation difference between the primary and recurrent tumor; pair 7 had none. Most of the changes appeared to result in loss of pathogenic clonal mutations, such as TP53 R174X in pair 1, which may have been a driving mutation in the primary tumor but was not observed in the recurrent. The only consistent change we observed was that no new clonal mutations arose in the recurrent GBMs, at least not in the 13 genes captured in this study.

DISCUSSION

Glioblastoma is the most common and aggressive malignant brain tumor in adults, with less than 5% of adults surviving 5 years after initial diagnosis [65-68]. Given the intratumoral clonal heterogeneity of GBMs [80-82], we hypothesized that subclonal heterogeneity may further complicate the mutational landscape of these tumors. We hoped that, by identifying the signature(s) of the subclonal mutations, we might gain insight into the underlying driver mechanisms of GBM. We, therefore, used Duplex Sequencing to sequence two sections of primary GBM tumors in an effort to identify subclonal intratumoral heterogeneity and to determine if distinct subclonal mutational signatures characterized different parts of the tumor. Between distinct geographic regions of the primary GBMs, we observed varying degrees of subclonal mutation burden, as well as differing spectrums between the two regions. These results suggest that the underlying mechanisms driving distinct regions of an individual GBM tumor may differ. While additional tumors need to be analyzed to confirm this to be the rule rather than the exception for GBMs, this preliminary finding suggests that future targeted therapy approaches will need to examine multiple tumor sections to determine how to best target the driving mechanisms of GBM pathogenesis to prevent recurrence.

Additionally, all GBMs eventually recur. Since relapse frequently occurs via tumor evolution, with the expansion of mutant clones that express proteins that empower cells with increased proliferative potential in the presence of specific chemotherapeutic agents, we hypothesized that these mutant clones may already be present in primary tumors at subclonal levels. To determine if resistance-promoting mutations are present in primary GBMs prior to treatment, we selected 13 genes implicated in the pathogenesis of GBM and analyzed them in paired primary and recurrent tumor pairs (Table 4). We determined if subclonal mutations underwent selection and clonal expansion from primary to recurrent tumor, since the presence of subclonal mutations in primary tumors that are implicated in treatment resistance could potentially help guide therapy decisions.

Despite a lack of consistency between types of changes observed in recurrent tumors versus their primary tumors, each tumor pair demonstrated differences. This suggests that the mechanisms driving selection of mutants underlying primary GBM pathogenesis and recurrence differ, which makes sense given that recurrent tumors are developing under the added stress of chemotherapy and radiation treatments. The large pool of subclonal mutations observed in the primary tumors appears to represent a reservoir of mutations that could promote recurrence and resistance to therapy, as evident from the increased clonality of some of the subclones. Further samples need to be analyzed so that tumors grouped by subtype and underlying pathology in the primary can be analyzed upon recurrence to determine if consistent mechanisms drive recurrence for a given primary tumor subtype.

Several low level subclonal mutations in the primary tumor appeared to undergo selection and expansion in the recurrent tumor (Table 2), as evident from the increased frequency of the PTENP ncRNA mutant in Pair 1, the TP53 V118E mutant in Pair 3, and the EGFR S703F mutant in Pair 9. However, there were also several mutations that decreased substantially in the recurrent tumor (such as the PTEN F154D and TP53X mutations in Pair 1, the EGFR T263P mutation in Pair 11, and the IDH1 R132G and TP53 R234C mutation in Pair 12), indicating selection against those clones. Whether or not this was due to loss of a significant fraction of the tumor due to surgery, since heterogeneity may be contributing to the differences observed, or whether this is due to treatment-induced selection, is unclear. Significant loss of SNVs is seen upon recurrence, consistent with removal of the tumor. However, the retention of a large number of SNVs in all six tumor pairs suggests that surgical resection incompletely removes the tumor. Given the diffuse nature of GBM tumors, this would not be surprising.

The lack of new clonal mutations in any of the recurrent GBMs could have three possible explanations. (i) Actionable molecular targets pre-exist treatment in both primary and recurrent GBM tumors. An implication of this possibility is that sequencing primary tumors and devising a

treatment based on the mutations present may be a viable option. (ii) Actionable targets specific to recurrence differ from the genes most commonly mutated in primary GBM (and captured in these experiments). This possibility implies that the pathology of recurrent GBM may be different than the primary. Therefore, if we want to prevent GBM recurrence, we need to understand what drives it and not assume the driving mechanisms are the same as those that drive initial GBM pathology. (iii) The current therapeutic approach, concurrent radiation and temozolamide, do not reduce the clonal pathogenic mutations in the primary. If this is the case, we would need to design new therapies to specifically target the mutated gene products present in GBM. These explanations are not mutually exclusive of each other. Given the complexity of GBMs, it would not be surprising if all three are the case.

In conclusion, this preliminary report on subclonal mutations in glioblastoma presents a bleak picture. In addition to the high clonal mutation burden of GBMs and intratumoral clonal heterogeneity already discovered, our results indicate that subclonal mutations in GBM appear to be geographically heterogenous, with differing spectrums between regions of a tumor, suggesting that multiple mechanisms may drive distinct part of the tumor. Additionally, upon recurrence, there appears to be little consistency between the primary and recurrent tumor or between different recurrent tumors, despite all GBMs receiving the same treatment. GBMs appear to exhibit substantial levels of heterogeneity between individual patients as well as in time and space within a single patient.

METHODS

Primary and recurrent GBM tumors were obtained from The Ivy Center for Advanced Brain Tumor Treatment at Swedish Cherry Hill Hospital in Seattle, WA. To measure mutation frequencies, DNA was isolated and sequenced via one round of Duplex Sequencing as previously described [28]. Target genes were exonic regions of *CDKN2A*, *CHEK2*, *EGFR*, *H3F3A*, *IDH1*, *PIK3CA*, *PIK3R1*, *PTEN*, *RB1*,

TERT, AND *TP53*, which were chosen based on the high incidence of clonal mutations in these genes in GBMs (Table 1); additionally, we included *MSH6* and *MGMT* in our gene capture set, given the potential for mutations in these genes to confer resistance to the first-line GBM chemotherapeutic agent, temozolamide [66, 67]. We required a minimum depth of 100 Duplex molecules to call a position, either mutant or not; all samples had a mid-exon peak depth of at least 1000 Duplex molecules across all captured exons.

TABLES

Table 1: Captured genes for subclonal mutation studies of GBM, with references for gene choice

Gene	Reference
CDKN2A	[70, 84, 85]
CHEK2	[70, 84, 85]
EGFR	[70, 84, 85]
H3F3A	[84, 85] [70]
IDH1	[84-86] [70]
MGMT	[66, 84, 85]
MSH6	[66, 84, 85]
PIK3CA	[84, 85] [70]
PIK3R1	[84, 85] [70]
PTEN	[84, 85] [70]
RB1	[84, 85] [70]
TERT	[84, 85, 87]
TP53	[84, 85] [70]

Table 2: Key clonal and subclonal mutation differences in clonal and subclonal mutations in recurrent GBM, relative to primary

	Major mutation differences between primary and recurrent GBM[1]				Summary of SNV distributions[2]
	Mutation	**Primary freq.**	**Recurrent freq.**	**Freq. difference**	
Pair 1	PTEN F154D TP53 R174X[3] PTENP ncRNA[4]	51% 53% 14%	15% 0% 29%	↓ 36% ↓ 53% ↑ 25%	48 shared 53 unique primary 40 unique recurrent
Pair 3	EGFR R832H CHEK2P2, ncRNA TP53 V118F[3]	37% 10% 1%	59% 0% 11%	↑ 22% ↓ 10% ↑ 10%	70 shared 146 unique primary 34 unique recurrent
Pair 7	No major differences				29 shared 23 unique primary 39 unique recurrent
Pair 9	TP53 D242N	<1%	9%	9%	67 shared 34 unique primary 104 unique recurrent
Pair 11	EGFR S703F[4] EGFR T263P	1% 88%	32% 34%	↑ 31% ↓ 54%	58 shared 127 unique primary 58 unique recurrent
Pair 12	TERT H412Y IDH1 R132G[3] TP53 R234C[3]	50% 35% 75%	44% 20% 54%	↓ 6% ↓ 15% ↓ 21%	53 shared 61 unique primary 65 unique recurrent

1. Pathogenic, undetermined, and non-dbSNP mutations with a ≥5% clonality change in recurrent tumor, relative to primary
2. All single nucleotide variants (SNV) detected by Duplex Sequencing
3. Known pathogenic SNV in COSMIC [84, 85]
4. Undetermined clinical significance in dbSNP [88]

FIGURES

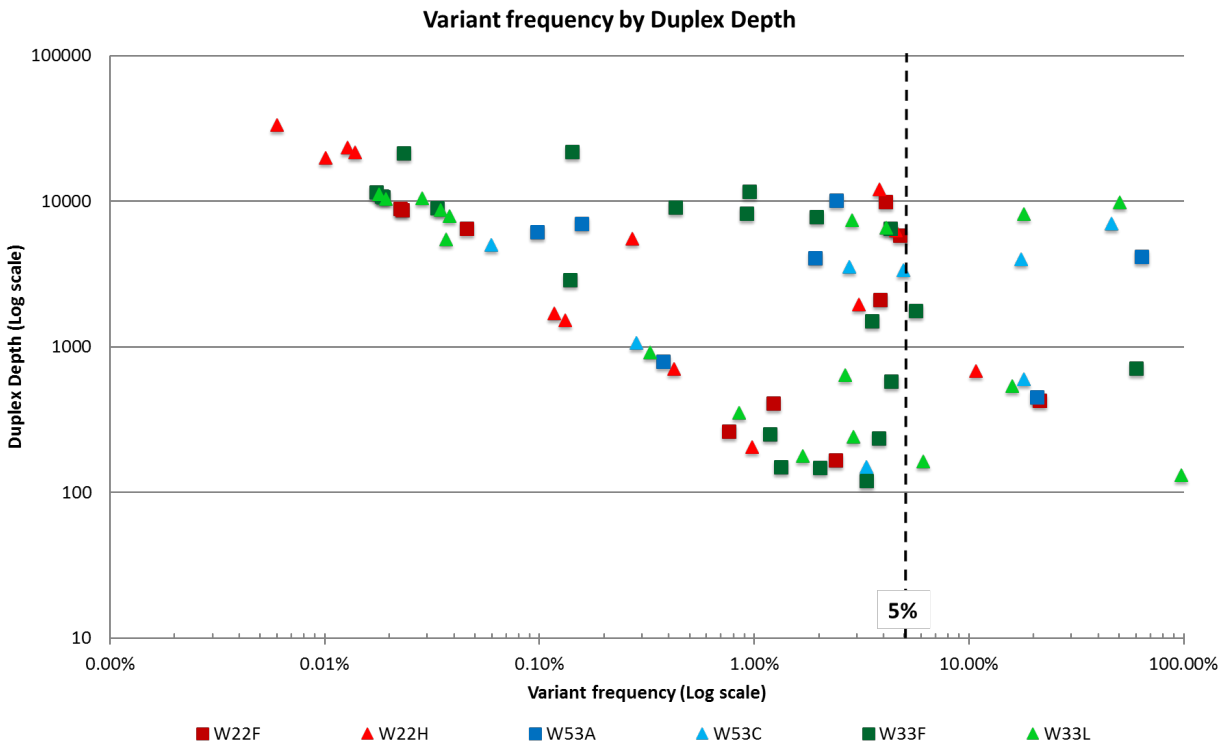


Figure 1: Numerous subclonal mutations are detected in primary GBM via Duplex Sequencing. Two sections from each of three GBM tumors (W22, W53, and W33) were analyzed via Duplex Sequencing. All subclonal (<20% clonal) mutations and non-dbSNP mutations are plotted, relative to the Duplex Sequencing depth obtained for each mutation. All data points above 10,000 are EGFR mutations.

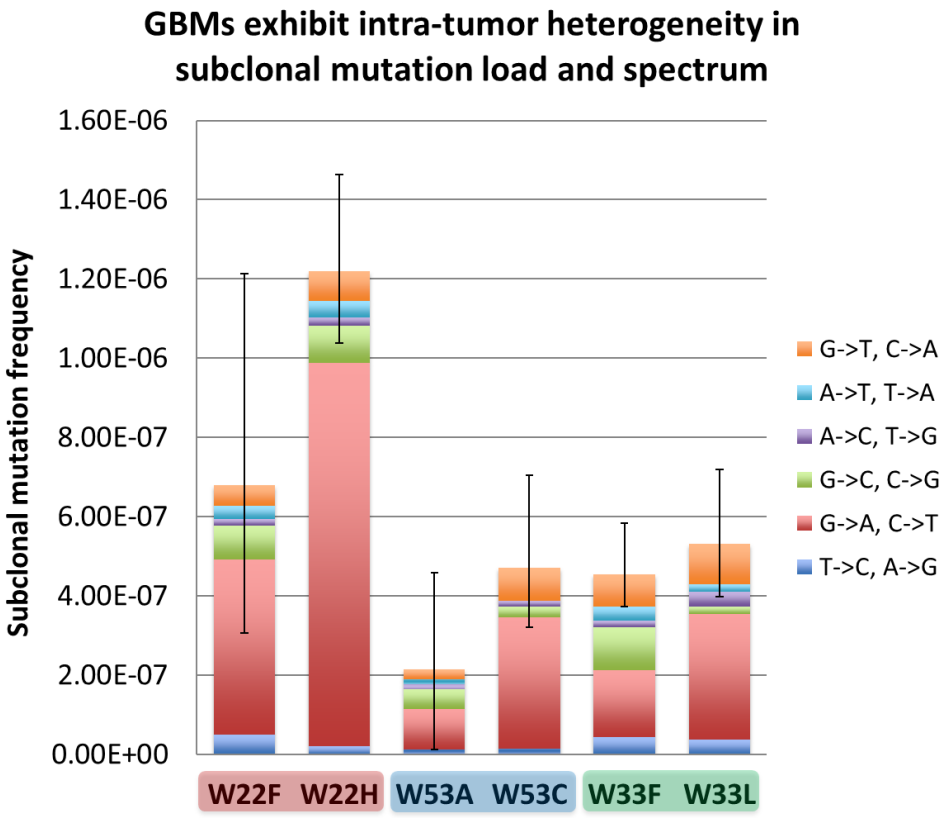


Figure 2: Intratumoral subclonal heterogeneity in primary GBM. Two sections from each of three GBM tumors (W22, W53, and W33) were analyzed via Duplex Sequencing. Mutations present at less than 20% clonality (subclonal) were analyzed by mutation type.

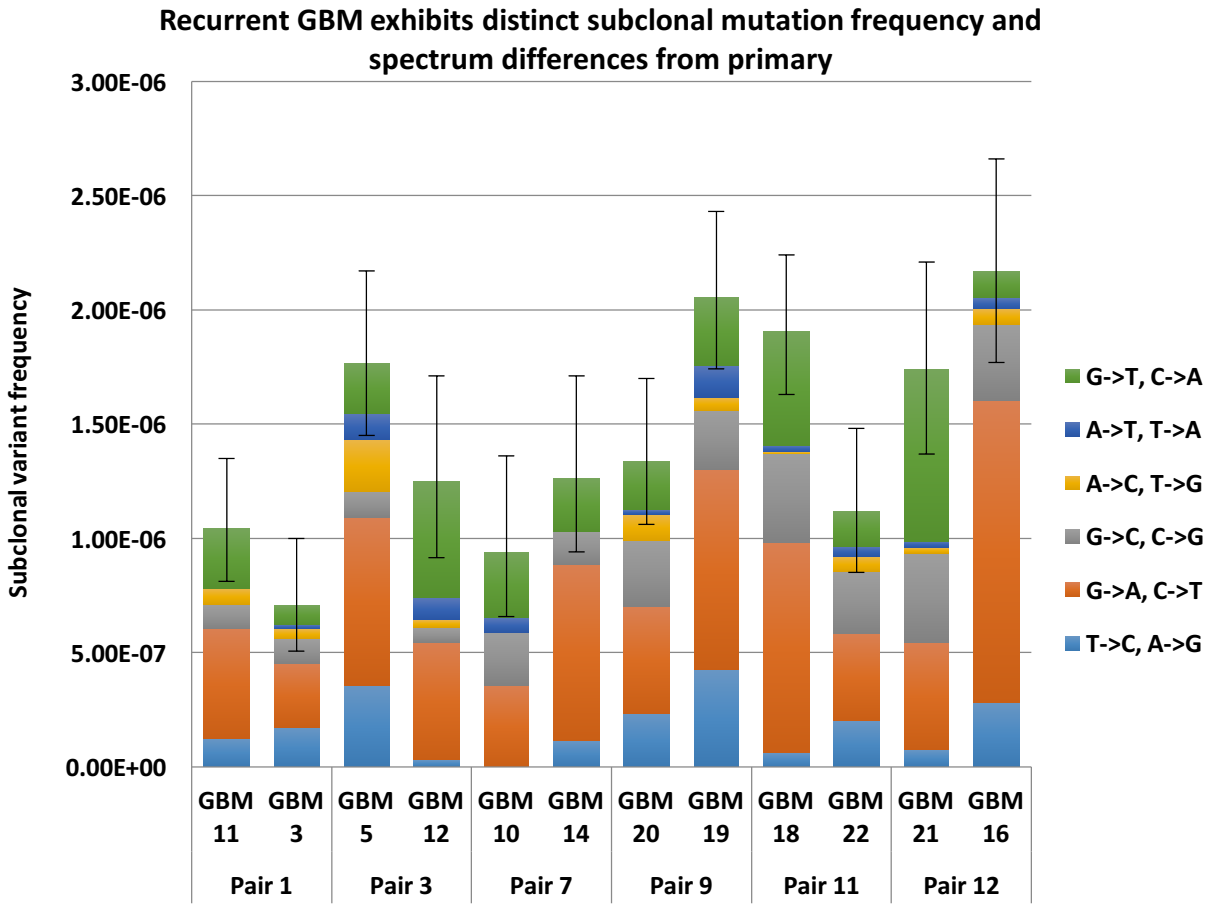


Figure 3: Subclonal heterogeneity between primary and recurrent GBM. DNA from each of 6 primary and recurrent GBM tumor pairs was analyzed via Duplex Sequencing. Mutations present at less than 20% clonality (subclonal) were analyzed by mutation type.

CHAPTER 3: DEVELOPMENT OF ACCURATE RNA CONSENSUS SEQUENCING (ARC-SEQ) FOR HIGH-FIDELITY RNA MUTATION DETECTION

ABSTRACT

Mistakes make during gene transcription, termed transcriptional mutagenesis (TM), can result in mutant proteins with altered properties. TM has long been hypothesized to play a role in aging and age-related diseases, including cancer and neurodegeneration, as well as in the evolution of viruses and bacteria. Despite this long-standing hypothesis, scientists have made little progress over the last 50 years in elucidating the importance of TM in human health and disease, as methods for detecting TM have, until recently, been lacking. In order to study TM, I have developed a highly sensitive, high-throughput method, termed Accurate RNA Consensus sequencing (ARC-seq), to measure RNA mutations. ARC-seq utilizes an adaptor to uniquely identify each RNA molecule and generates multiple cDNA copies per RNA molecule, allowing artifacts introduced during cDNA synthesis, PCR, and sequencing to eliminated, yielding the true RNA mutations. Herein, I describe the development of ARC-seq and demonstrate its ability to fully correct artifacts resulting from reverse transcription, amplification, and sequencing errors. Additionally, I demonstrate its utility for studying *in vivo* TM due to RNA polymerase mutations and oxidative stress. ARC-seq enables the highly accurate study of RNA mutations at the single-molecule level from any RNA source, regardless of organism or type of RNA, and has the flexibility to increase stringency, as may be required to sequence highly damage RNAs. Thus, ARC-seq will enable sensitive studies of how perturbing a cell's environment or machinery affects the fidelity of transcription. Additionally, ARC-seq can be used to elucidate the contribution of RNA mutations to aging, cancer, and neurodegeneration, as well as the evolution and acquired resistance of viruses and bacteria.

INTRODUCTION

DNA mutations play an important role in mammalian aging and disease. In particular, coding mutations can disrupt the function of proteins and alter the physiology of cells. Consequently, mutations can have profound effects on mammalian health. Germline mutations, for example, can cause congenital disorders, whereas somatic mutations can drive carcinogenesis and aging [89, 90]. A wealth of evidence implicates inherited and somatic mutations in human disease, yet there are numerous sporadic diseases for which no genetic cause has yet been found.

In addition to genetic mutations, mistakes during transcription and translation can introduce errors in a protein's sequence [91, 92]. We do not know the extent to which infidelity of these processes contributes to disease, yet recent evidence suggests that they could play a crucial role in multiple disease processes. Infidelity of transcription, termed transcriptional mutagenesis (TM), in which RNA polymerase errs during transcription, is of particular interest, as multiple proteins are translated from each mRNA copy [93]. Indeed, using a phenotypic switching model in *Escherichia coli*, Gordon, *et al.* recently demonstrated that the effect of a transcription error is amplified nonlinearly over the original stochastic event, whereas a translation error effects only a single protein [94]. Thus, a single mutant transcript has the potential to have a profound effect on cellular function, depending on the consequence of the nucleotide change in the protein's sequence and the function of the protein.

Many studies have investigated the behavior of RNA polymerase II (RNAPII), the primary polymerase responsible for transcribing protein-coding genes, in various sequence contexts; *in vitro* studies estimate the fidelity of RNAPII transcription to be $\sim 1 \times 10^{-5}$ [95-99]. However, this rate can increase dramatically during transcription of damaged templates or repetitive sequence. While bulky lesions largely block transcription elongation, essentially forcing the cells to repair the damage before transcription can resume (see [100] for a review), RNAPII efficiently bypasses many smaller lesions, with the consequences of bypass varying widely with lesion type and, in some cases,

sequence context (see [101] for an overview). For example, RNAPII correctly incorporates adenine across from thymine glycol [102]. 5-hydroxyuracil and dihydrouracil, on the other hand, miscode frequently, inducing the misincorporation of adenine (G→A transition)[103, 104], which can lead to the production of mutant proteins in non-dividing cells [105]. 7,8-dihydro-8-oxoguanine (8-oxoG) also causes misincorporation of an adenine into the mRNA transcript (C→A transversion). In *E. coli*, this occurs approximately 60% of the time [106], whereas in mammalian cells, transcriptional bypass of 8-oxoG is highly dependent on sequence context and the lesion's position relative to the promoter [107]. Additionally, in human cells, error-prone transcriptional bypass of an 8-oxoG lesion at a key residue in *HRAS* can lead to production of mutant Ras protein and constitutively activate the downstream MAP kinase cascade [108]. Furthermore, O6-methylguanine (O6-MeG) causes misincorporation of uracil (C→U transitions) *in vitro*, with a frequency as high as 67%, and can result in altered protein function in human cells [109].

In addition to DNA damage, repetitive sequences can also induce transcriptional errors in the form of slippage, leading to frameshift mutations [110, 111]. Several studies have shown that transcriptional slippage within polyA tracts of genes can correct genetic mutation-induced phenotypes, lessening the severity of diseases such as canine cyclic neutropenia, hypobetalipoproteinemia, and hemophilia A [112-115]. Transcriptional slippage can occur in other repeat sequence motifs as well. Homozygous Brattleboro rats have a single base deletion in the vasopressin gene, resulting in a lack of functional vasopressin mRNA and protein; Evans, *et al.* found functional mRNA and protein produced in a small but increasing proportion of hypothalamic cells as the rats aged and determined the mechanism to be transcriptional slippage at a GAGAG motif within the vasopressin gene, resulting in a ΔGA [116]. Similarly, another study determined that mutant amyloid-beta and ubiquitin aggregates from Alzheimer's disease (AD) brains arose due to mutant mRNAs with a ΔGA in one or more of the proteins' GAGAG motifs [117]. Following the discovery of prion-like seeding of protein aggregation in several neurodegenerative diseases, including AD, PD,

and ALS, several groups put forth the prion hypothesis of neurodegenerative diseases [118-120]. Coupled with the observation that TM can produce mutant proteins, it is conceivable that TM past DNA damage can produce misfolded proteins that can then seed the transformation of normal proteins, leading to pathogenic aggregation, and triggering the pathologies characteristic of AD and PD.

Despite the wealth of evidence for TM from *in vitro* fidelity assays and assays using highly expressed reporter genes, it is difficult to draw conclusions about role the of TM *in vivo*. The results of these studies cannot be easily extrapolated to what may be occurring in cells, where transcription factors, repair enzymes, chromatin, and gene expression levels modulate transcriptional fidelity. Consequently, in order to elucidate the roles of TM in aging, disease, drug-resistance, and evolution, it is likely necessary to study RNA molecules transcribed *in vivo*.

RNA sequencing (RNAseq) is an immensely useful method for studying gene expression and genetic mutations. However, it is inadequate for determining RNA mutations due, in large part, to the method's requirement to convert the RNA to cDNA prior to sequencing. Reverse transcriptase (RT), an inherently error-prone polymerase, performs this conversion. While the error rate of RNA polymerase is $\sim 1x10^{-5}$ [95-99], the error rate of RT is $\sim 10^{-4}$ [5]. Thus, RT introduces 10-fold more artificial errors during cDNA synthesis than the RNA polymerase does during transcription, effectively swamping out the mRNA mutations. Consequently, because of the high rate of RT errors, it is difficult to determine if sequenced mutations are due to TM or due to the RT or other sequencing errors. Thus, in order to study TM *in vivo*, I have developed a highly accurate sequencing method, termed Accurate RNA Consensus sequencing (ARC-seq) to measure RNA mutations with single-molecule sensitivity. ARC-seq uniquely combines the use of an adaptor to barcode each RNA and the generation of multiple cDNA copies of each RNA molecule prior to sequencing. This combination enables the elimination of sequence artifacts due to cDNA synthesis, PCR errors, and sequencing errors, yielding the true RNA mutations resulting from TM *in vivo*.

RESULTS

Development of a sensitive, highly accurate method to detect RNA mutations. In seeking to accurately measure RNA mutations with high sensitivity, I had to address three obstacles to accurate RNA sequencing: (i) RNA must be reverse transcribed to DNA before it can be sequenced; (ii) PCR amplification of DNA can introduce artifacts; (iii) high-throughput sequencing is highly error-prone. In order to overcome these obstacles, ARC-seq combines the use of unique barcodes for each RNA molecule and the generation of multiple cDNA copies per individual RNA molecule. This combination enables the elimination of artifacts due to cDNA synthesis, PCR errors, and sequencing errors, yielding the true RNA mutations. To do so, ARC-seq first ligates a single stranded barcoded adaptor onto each fragmented RNA molecule; this adaptor contains 16 nucleotides of randomized sequence that uniquely identify individual RNA molecules (Fig. 1A). Next, ARC-seq generates multiple cDNA copies of each barcoded RNA by circularizing each barcoded RNA molecule and performing rolling-circle reverse transcription (RC-RT); after restricting the resultant multimeric cDNA molecule into monomers, each cDNA copy of the original RNA molecule is uniquely tagged and then amplified via PCR. After sequencing, using bioinformatics, the cDNA tags are then used to generate a PCR consensus sequence, eliminainge artifacts due to sequencing and PCR errors (Fig. 1B). Finally, the RNA barcode is used to generate a cDNA consensus sequence, eliminating reverse transcription and damage-induced artifacts, regenerating the original RNA sequence and revealing any true RNA mutations. ARC-seq has a theoretical background of 0.01^n, where *n* is the number of cDNA copies produced from each RNA molecule, which can be increased as needed to account for the quality of the RNA.

ARC-seq effectively corrects RT, PCR, and sequencing artifacts. In order to validate ARC-seq's effectiveness at eliminating artifacts due to reverse transcription, PCR errors, and sequencing errors, three types of RNAs were generated by *in vitro* transcription, using T7 RNA polymerase (Fig. 2A): (i)

high-fidelity RNA, which was generated using pristine DNA template and has an expected true mutation frequency of $3x10^{-5}$ [98]; (ii) damaged RNA, which was generated by treating the high-fidelity RNA with an oxidizing agent and should have the same true RNA mutation frequency as the high-fidelity RNA ($3x10^{-5}$); (iii) mutated RNA, which was generated from a DNA template oxidatively damaged in order to induce mistakes during transcription, which is expected to have an elevated mutation frequency. These RNAs were then sequenced via ARC-seq. At a cDNA family size of 1, which corresponds to conventional RNAseq, the error frequency of the high-fidelity RNA was approximately $2x10^{-4}$, approximately 10-fold greater than the expected true RNA mutation frequency (Fig. 2B). Additionally, the error frequency of the damaged RNA template was increased about 3-fold over the high-fidelity RNA, consistent with the high error rate of conventional RNAseq, especially on damaged RNA templates.

In contrast, by requiring 5 cDNA copies per RNA molecule, ARC-seq was able to reveal the true RNA mutation frequency of the high-fidelity RNA to be $\sim 2x10^{-5}$. Furthermore, by requiring 6 cDNA copies per RNA molecule to form a consensus sequence and, therefore, increasing the stringency of its error-correction, ARC-seq was able to fully correct for damage-induced artifacts and reveal the true RNA mutation frequency of the damaged high-fidelity RNA to also be $\sim 2x10^{-5}$. Consistent with ARC-seq eliminating errors without mistakenly removing true RNA mutations, even with the high stringency of 6 cDNA copies per RNA molecule to form a consensus sequence, the true RNA mutation frequency of the mutated RNA remains more than 10-fold greater than the high-fidelity RNA. Thus, by repeatedly sequencing the same RNA molecule, ARC-seq is able to eliminate damage-induced and sequencing artifacts, revealing true RNA mutations.

ARC-seq reveals the frequency and spectrum of RNA mutations *in vivo*: Recent studies have described several mutants of *Saccharomyces cerevisiae* (yeast) that have reduced *in vitro* RNA synthesis fidelity [22, 23]. *Rpb1 E1103G* is a point mutant of the catalytic domain of RNA polymerase

II and confers dependence on TFIIS [23]. *ΔRpb9* is a deletion mutant of a transcription factor that is known to enhance the fidelity of transcription in yeast [41]. To establish ARC-seq's utility for measuring *in vivo* RNA mutations, I used it to study TM in these yeast mutants. Stationary phase wildtype yeast was found to have a RNA mutation frequency of ~$4.6x10^{-5}$ (Fig. 3A). In contrast, both RNA polymerase mutants had elevated RNA mutation frequencies, ~$6.6x10^{-5}$ and $9.5x10^{-5}$ for E1103G and ΔRpb9, respectively. Interestingly, these mutation frequencies were revealed by requiring only 3 cDNA copies per RNA molecule and were unchanged with increasing stringency, even when I required as many as 7 cDNAs per RNA. Furthermore, examining the spectrum of mutations revealed subtle differences between the three yeast strains (Fig. 3B). While C→U mutations are the most frequently observed RNA mutation in all three yeasts, E1103G had an increase in A→G mutations, and ΔRpb9 has an increase in both A→G mutations and A→U mutations, consistent with deficiencies in different aspects of RNA transcription in the two mutants.

ARC-seq reveals that oxidative stress induces TM *in vivo.* DNA damage due to oxidative stress is well known to induce DNA mutations, and *in vitro* studies of RNA polymerase behavior at DNA lesions indicate that it behaves similarly to DNA polymerases. Thus, in order to determine if oxidative stress induces elevated TM, log-phase wildtype and ΔRpb9 yeast were treated with 50μM H_2O_2 for 30 minutes and then their RNAs were extracted and sequenced via ARC-seq. In mRNA, the wildtype mutation frequency increased from $5.6x10^{-5}$ to $1.3x10^{-4}$ and from $8.2x10^{-5}$ to $1.5x10^{-4}$ in ΔRpb9 yeast, following oxidative stress. C→U mutations increased from 4.6 $x10^{-5}$ to $9.17x10^{-5}$ in wildtype yeast and from $6.98x10^{-5}$ to $1.04x10^{-4}$ in ΔRpb9 yeast, consistent with oxidative stress-induced cytosine deamination. Additionally, C→A mutations increased from $5.39x10^{-6}$ to $2.12x10^{-5}$ in wildtype yeast and from $5.28x10^{-6}$ to $1.54x10^{-5}$ in ΔRpb9 yeast, consistent with error-prone transcriptional bypass of an 8-oxoG lesion, a common DNA lesion resulting from H_2O_2 treatment (Fig. 4A).

Since the ΔRpb9 mutant is a RNA polymerase II mutant, which should largely only effect mRNA transcription fidelity, I examined the rRNA of the wildtype and ΔRpb9 yeast to ensure that the differences between the wildtype and mutant yeast were specific to RNA polymerase II transcription and to determine how oxidative stress effects the transcriptional fidelity RNA polymerases I and III in yeast. The control ncRNA mutation frequencies of wildtype and ΔRpb9 yeast were similar, $2.1x10^{-5}$ and $2.8x10^{-5}$, respectively (Fig. 4B), consistent with the key difference between the wildtype and mutant being in mRNA transcription fidelity. Following oxidative stress, ncRNA mutation frequencies increased to $1.2x10^{-4}$ and $1.1x10^{-4}$ in wildtype and ΔRpb9 yeast, respectively, consistent oxidative stress-induced TM of rRNA genes. Similar to mRNA, C→U mutations increased from 1.1 $x10^{-5}$ to 5.8 $x10^{-5}$ in wildtype yeast and from 1.8 $x10^{-5}$ to 5.2 $x10^{-5}$ in ΔRpb9 yeast; C→A mutations increased from 4.8 $x10^{-6}$ to 3.3 $x10^{-5}$ in wildtype yeast and from 6.1 $x10^{-5}$ to 1.5 $x10^{-5}$ in ΔRpb9 yeast, consistent with oxidative stress-induced cytosine deamination and TM across from oxidative DNA lesions.

DISCUSSION

Errors during gene transcription can result in mutant proteins with altered properties. Consequently, TM has long been hypothesized to play a role in aging and age-related diseases, including cancer and neurodegeneration, as well as in the evolution of viruses and bacteria [101, 121-123] . Despite this long-standing hypothesis, scientists have made little progress over the last 50 years in elucidating the importance of TM in human health and disease, as methods for detecting TM have, until recently, been lacking. The sources of artifacts in conventional RNA sequencing are cDNA synthesis errors due to reverse transcription, PCR errors, and sequencing errors. Recently, recognizing the importance of RNA mutation detection, several groups have sought to address one or multiple of these issues in an effort to accurately sequence RNA [124-127]. However, these methods have either failed to address all three sources of artifacts [124-126], are too inefficient for

in vivo studies, or they've been highly laborious [127, 128] and, consequently, limited to studying highly mutated RNAs, such as those from RNA viruses.

In developing ARC-seq, I reasoned that, by generating multiple cDNA copies per RNA molecule, I could eliminate sequencing and RT errors. This required uniquely tagging each RNA molecule prior to cDNA synthesis and then sequencing multiple cDNA copies per original RNA molecule. Furthermore, by uniquely tagging each cDNA copy (in addition to the RNA barcoding), I reasoned that I would be able to distinguish between cDNA duplicates of a single RNA molecule and PCR duplicates of a single cDNA copy. This is important to eliminate PCR errors.

Applying ARC-seq to the determination of TM induced during *in vitro* transcription, I found that ARC-seq was able to fully eliminate artifacts due to RT, PCR, and sequencing errors, revealing the true RNA frequency of the *IVT* RNA to be $\sim 2\text{x}10^{-5}$ (Fig. 2B), consistent with published findings of T7 RNA polymerase's fidelity [98]. Importantly, by increasing the stringency of ARC-seq by requiring an additional cDNA copy to form a cDNA consensus sequence, ARC-seq was even able to fully eliminate damage-induced artifacts in the high-fidelity RNA sequencing data. ARC-seq's adaptability to increase stringency is important for determining TM in highly damaged RNAs, such as those extracted from tissues, as *in vivo* RNAs are subject to a cellular environment that may induce oxidative RNA damage as well as other RNA lesions.

Upon sequencing RNAs from wildtype and RNA polymerase II-mutant yeast, I found that ARC-seq revealed a stable RNA mutation frequency with as few as 3 cDNA copies per RNA molecule (Fig. 3A); this was surprising, given the higher stringency required when sequencing *in vitro* transcribed RNAs. At this relatively low stringency, ARC-seq revealed mutation frequency and spectrum differences between the wildtype and two mutant yeast strains (Fig. 3B). The most frequent mutation in all three yeast strains was the C→U mutation, with the majority of the mutation frequency differences between the mutants and wildtype being due to increases in this mutation. Additionally, subtle differences between the spectrums of mutations in the three yeast

strains were apparent, consistent with the fidelity of RNA polymerase II differing between the strains.

Applying ARC-seq to the biological question of the consequences of oxidative stress on TM, ARC-seq revealed that both wildtype and ΔRpb9 yeast showed elevated TM following treatment with H_2O_2, relative to controls. Specifically, C→A mutations increased ~4-fold in wildtype and ~ 3-fold in ΔRpb9, consistent with error-prone transcriptional bypass of an 8-oxoG lesion, the most common DNA damage resulting from oxidative stress [129, 130] (Fig. 4A). In addition to the increased C→A mutations, G→A mutations increased ~2-fold in wildtype yeast and by ~50% in ΔRpb9 yeast. Deamination of cytosine in DNA readily results upon oxidation of cytosine; thus, these mutations could be a consequence of TM across a deaminated cytosine. Alternatively, oxidative stress within the cell could result in deamination of cytosine in RNA, yielding the same mutation upon sequencing. I am unable to distinguish between these two possibilities with my current bioinformatic pipeline. However, in terms of the resulting consequence for translation of the mutated or deaminated RNA, the protein change will be the same, regardless of if the source of the G->A mutation is TM across from a deaminated cytosine in DNA or deamination of cytosine in the RNA. Further bioinformatic development will allow enable me to distinguish between these two events.

Verifying that the ΔRpb9 mutant is specifically defective in RNA polymerase II transcriptional fidelity, the rRNA frequency of wildtype and ΔRpb9 yeast were approximately equal (Fig. 4B). Similar to the mRNA, following oxidative stress, a dramatic increase in mutation frequency occurred in rRNA of both wildtype and ΔRpb9 yeast, largely due to increased G→U and C→U mutations, consistent with oxidative stress-induced TM in the rRNA of these yeast. Surprisingly, the untreated rRNA mutation frequency of wildtype yeast was ~2-fold lower than its mRNA mutation frequency, largely due to decreased C→U mutations. This potentially suggests that either rDNA is more readily repaired than protein-coding gene regions in the genome or that the

fidelity of rRNA synthesis in higher than that of mRNA synthesis. Given that rRNA is longer lived and involved in protein translation, both of these potential implications make sense. While a mutated mRNA may be translated multiple times, yielding a pool of mutant proteins, codon redundancy limits the impact of an individual mutation and, even if a codon change results, there is still only that one protein species affected by the TM event. In contrast, a mutated rRNA could disrupt the function or fidelity of the ribosome, potentially creating many more mutant proteins, which would be a worse consequence for the cell; therefore, rDNA genome regions may be more closely guarded against the persistence of DNA damage or RNA polymerases I and III may have higher fidelity than RNA polymerase II. Further studies combining examination of DNA damage distribution coupled with TM studies of rRNA and mRNA may help distinguish between these possibilities.

Importantly, the H_2O_2 dose applied in my oxidative stress experiment resulted in greater than 95% survival. The high levels of TM observed suggests that DNA damage, whether due to exogenous agents or endogenous perturbations, could have profound yet unappreciated consequences for cells. Indeed, the oxidative damage theory of aging proposes that the accumulation of oxidative DNA damage over the course of an organism's life contributes to decline of cellular functions and eventual death[131]. In the brain, increased oxidative DNA damage and decreased repair capacities correlate with cognitive decline, age-related neuronal dysfunction, and neuronal death [132, 133]. Since neurons are terminally differentiated, transcriptional mutagenesis may be the consequence of increased damage and decreased repair, leading to cellular dysfunction and death. Additionally, multiple studies implicate increased oxidative DNA damage as a driving pathological mechanism in several neurodegenerative disorders, including AD [134-136]. TM may be the mechanism responsible for the cellular dysfunction and protein aggregation in AD. Error-prone transcription of damaged DNA by RNAPII could produce dysfunctional proteins, seeding protein aggregation and, ultimately, result in cell death.

In conclusion, Accurate RNA Consensus sequencing (ARC-seq), enables the highly accurate study of RNA mutations at the single-molecule level from any RNA source, regardless of organism or type of RNA, and has the flexibility to increase stringency, as may be required to sequence highly damage RNAs. ARC-seq will enable sensitive studies of how perturbing a cell's environment or machinery affects the fidelity of transcription. Additionally, ARC-seq can be used to elucidate the contribution of RNA mutations to aging, cancer, and neurodegeneration, as well as the evolution and acquired resistance of viruses and bacteria.

METHODS

In vitro transcribed (IVT) RNAs were generated from a single-stranded m13mp18 DNA template via an established protocol [137], using T7 RNA polymerase. To generate damaged IVT RNA, following transcription, the high-fidelity RNA was treated with 100μM H_2O_2 and $FeCl_3$ to induce oxidative DNA damage, according to an established protocol [138]. To generate mutated IVT RNA, the m13mp18 DNA template was treated with 1mM H_2O_2 prior to transcription.

Wild type and E1103G yeast were a gift from Mikhail Kashlev at the NIH/NCI, and ΔRpb9 yeast were a gift from Jeffrey Strathern at the NIH/NCI. To measure TM in yeast, log-phase or stationary-phase yeast were pelleted, washed with cold 1X PBS, and re-pelleted. The cell walls were then digested by incubating cells in a buffer containing sorbitol and 100U of Zymolyase, according to an established protocol [139]. RNAs were then extracted, enriching for mRNA, using the Dynabead mRNA Direct Kit from Ambion. Extracted RNAs were stored in TElow buffer, made with DEPC-treated nuclease-free water, with 100U of murine RNAse inhibitor from NEB added, at -80°C.

To measure RNA mutations, RNA libraries were prepped via the ARC-seq protocol, detailed in full in the Supplementary Information. Briefly, fragmented RNAs were end-repaired, pre-adenylated, and ligated to a custom RNA adaptor containing 16 nucleotides of randomized sequence as well as primer sequences necessary for sequencing on the Illumina platform. Adapted

RNAs were then circularized and subjected to rolling-circle reverse transcription to generated multimeric cDNAs. cDNA multimers were then restricted into cDNA monomoers, each of which was subsequently indexed via 5'-overhang extension, using a primer containing an additional 8 nucleotides of randomized sequence, as well as the remaining primer sequence necessary for Illumina platform sequencing. Finally, the indexed cDNA monomers were amplified and sequenced on an Illumina HiSeq 2200, using the dual indexing protocol.

Bioinformatic processing of the resulting raw sequence data combines the use of commonly available tools as well as software written in-house. Briefly, all raw sequencing reads are filtered based on the requirement that each position of the 16 nucleotide RNA barcode in read 4 and the 8 nucleotide cDNA tag in read 3 be only one of the four canonical bases; any reads not conforming to this criterion are discarded. The 16 nucleotide RNA barcode and 8 nucleotide cDNA tag are computationally appended to the header of each read-pair. The reads are then aligned against S228C yeast reference genome, using the Burrows-Wheeler Aligner (BWA) [63]. Reads not aligning to the yeast genome are filtered out. A PCR consensus sequence for cDNA tag family (e.g. reads sharing identical tag sequences) is computationally determined using software written in-house. The PCR consensus for any position is considered undefined if the position if less than 99% of the sequences at that position are in agreement. The PCR consensus sequences are then realigned to the yeast genome using BWA. After filtering for unmapped reads, cDNA consensus reads are then computationally determined using software written in-house. The cDNA consensus for any position is considered undefined if the position is represented by fewer than *n* instances in the family or if less than 70% of the sequences at that position in the read are in agreement; *n* represents the number of cDNA copies generated from each RNA molecule and can be adjusted to increase assay stringency if the RNA template is damaged. Next, to remove alignment artifacts common at the ends of reads, 8 nucleotides from the end of each end of the cDNA consensus reads are soft-clipped using

the Genome Analysis Toolkit. Finally, mutation frequencies and locations are determined, using scripts written in-house. All scripts are available upon request.

FIGURES

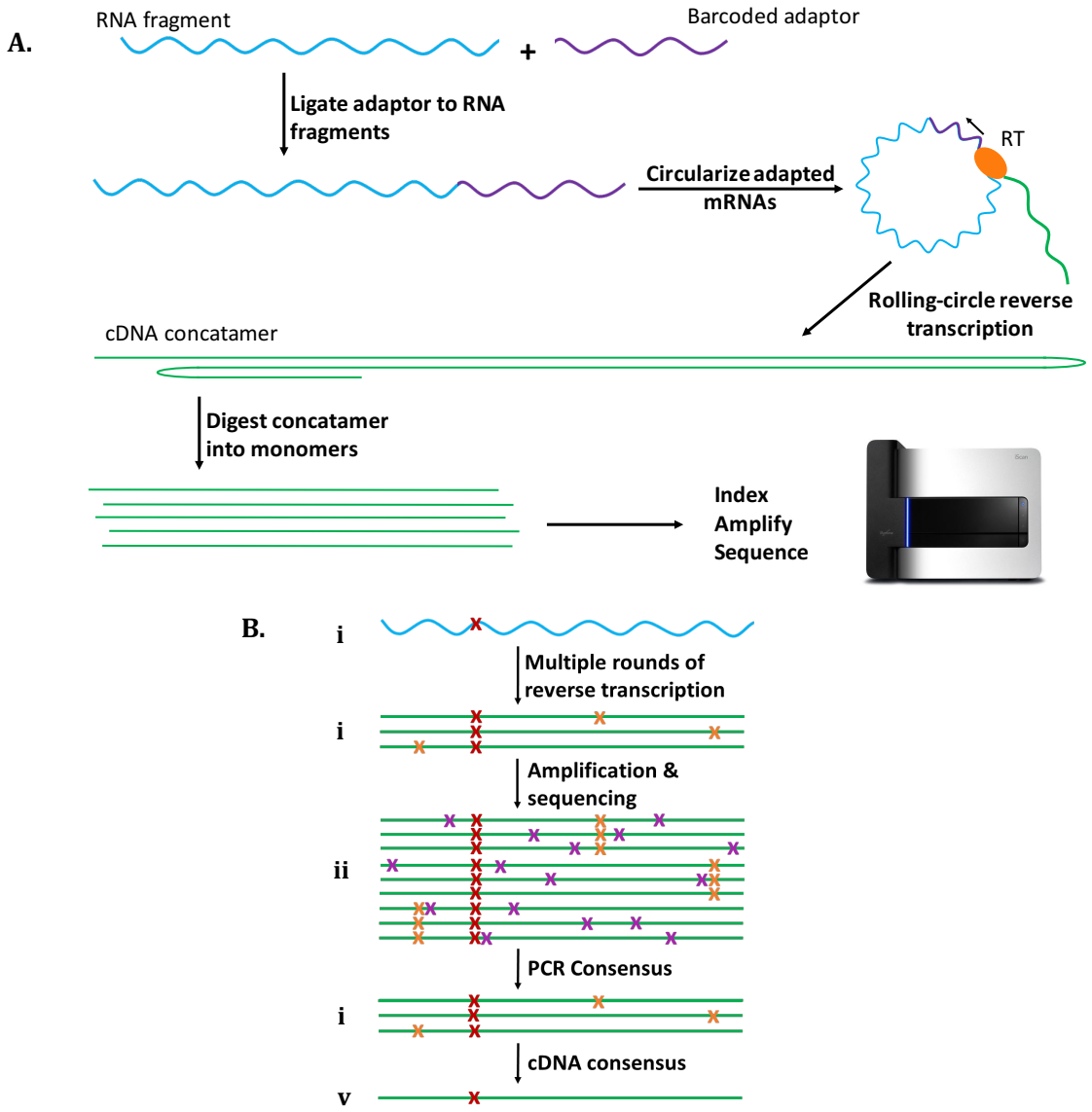


Figure 1: Overview of ARC-seq method and advantages. A: Each RNA is ligated to an adaptor containing a unique barcode. Ligated RNAs are then circularized and subjected to rolling-circle reverse transcription, generating a multimeric cDNA from each RNA molecule. cDNA multimers are then restricted into monomers, which represent cDNA copies of the original RNA molecule. Each cDNA is then tagged with a unique tag, amplified, and sequenced. B: Illustration of error-correction by ARC-seq. (i) A single RNA molecule containing a true RNA mutation (red); this molecule is barcoded. (ii) Rolling-circle reverse transcription generates multiple cDNA copies from each ligated RNA molecule, introducing random errors (orange); (iii) Amplification and sequencing amplifies the existing errors and introduces new errors (purple), further obscuring the true RNA mutation. The level of artifacts present in standard RNAseq data is illustrated at this level. (iv) After sequencing, cDNA tags are bioinformatically matched and a consensus sequence is generated for each cDNA copy, eliminating many amplification and sequencing artifacts. (v) Finally, the RNA barcodes are matched and a consensus sequence is generated from the cDNA copies, which regenerates the original RNA molecule's sequence, revealing the true RNA mutation.

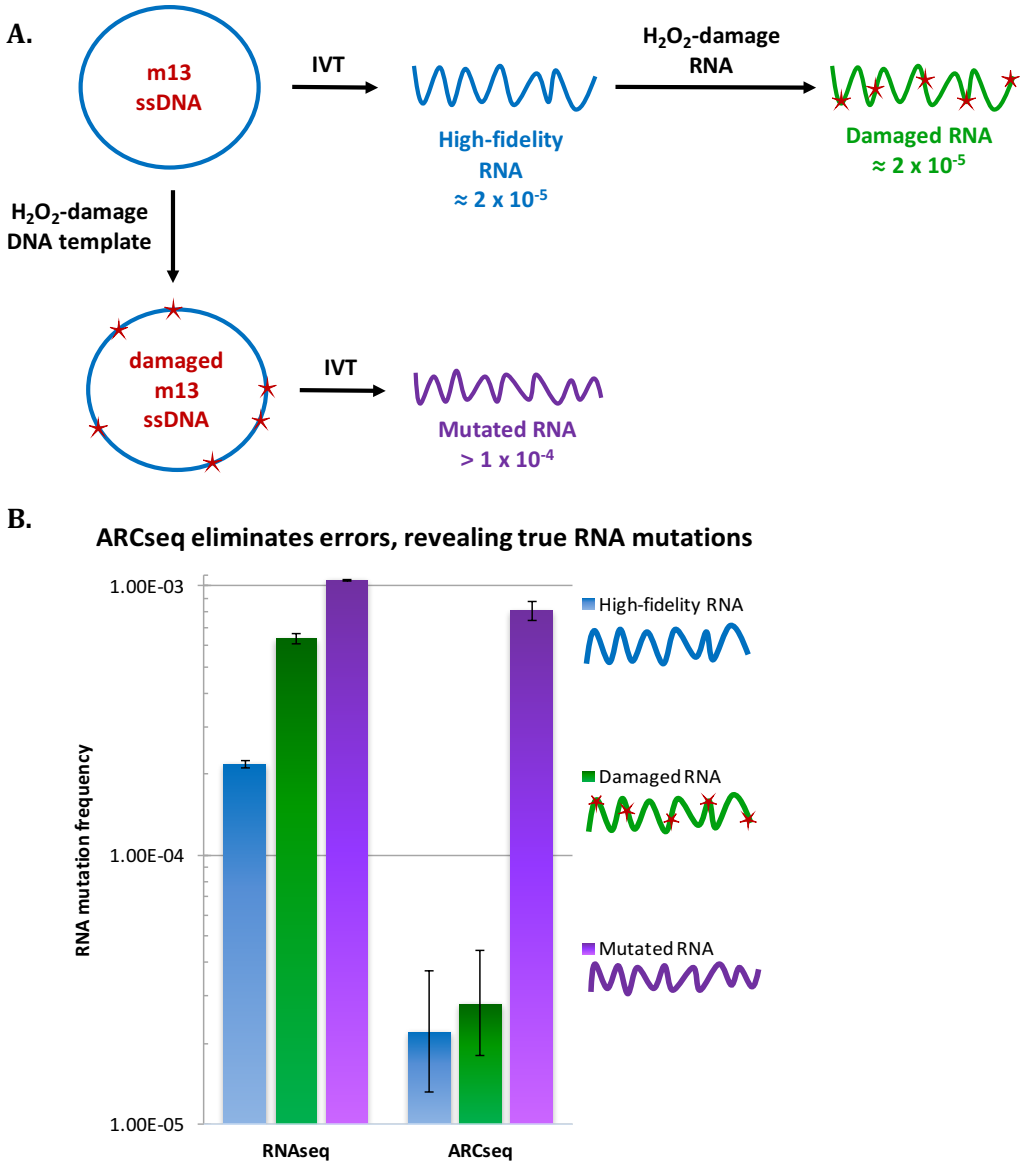


Figure 2: ARC-seq eliminates damage-induced, reverse transcription, PCR, and sequencing artifacts, revealing true RNA mutations. High-fidelity (blue), damaged (green), and mutated (purple) RNAs, were generated by *in vitro* transcription by T7 RNA polymerase (A) and sequenced via ARC-seq (B). While conventional RNAseq has a high level of artifacts, with increased artifacts observed in the damaged RNA template, ARC-seq is able to fully correct damage-induced artifacts, revealing the true RNA mutation frequency to be ~$2x10^{-5}$, without artificially removing true RNA mutations. Error bars represent Wilson scores of 95% confidence.

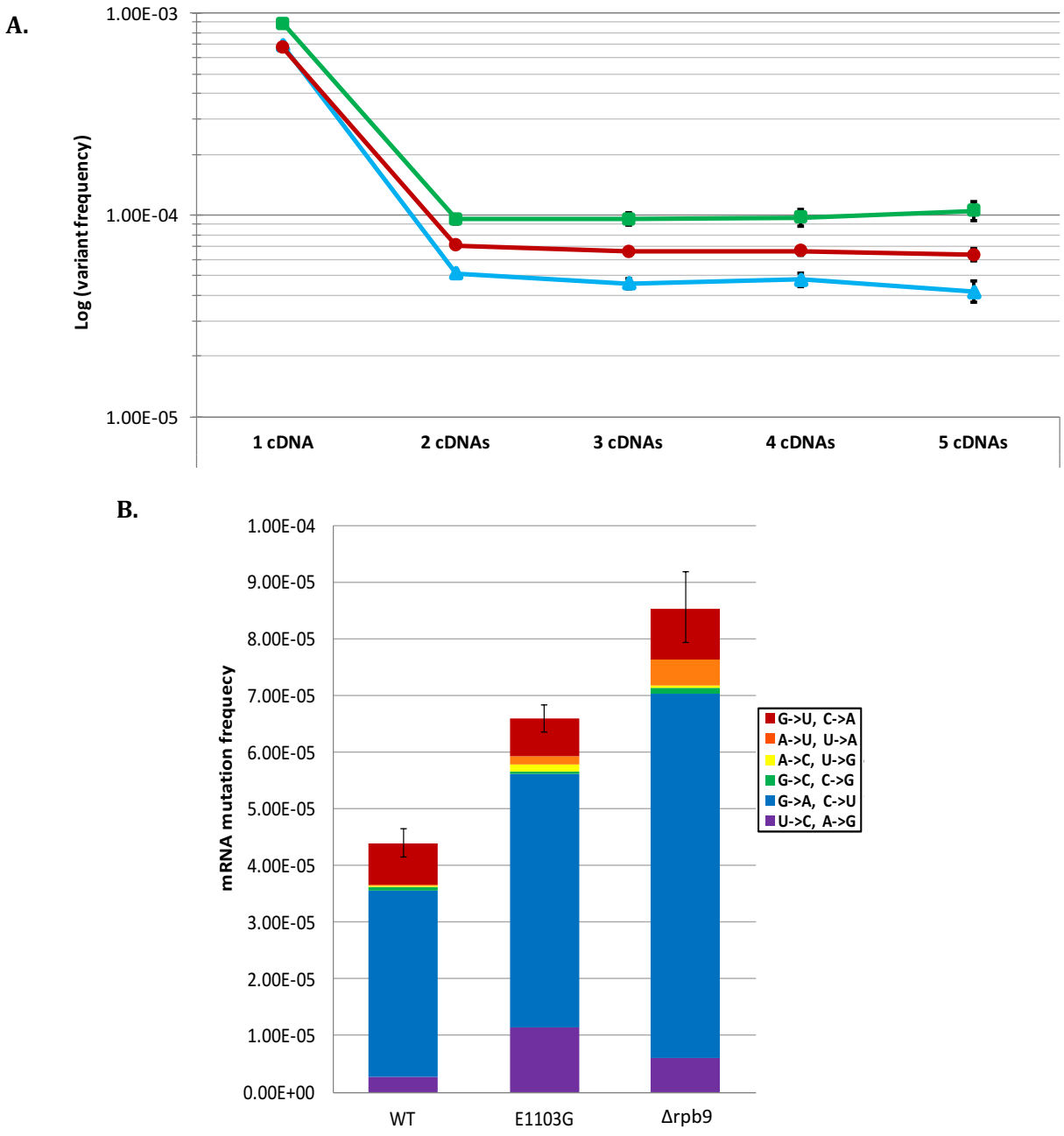


Figure 3: ARC-seq reveals differences in RNA mutation frequencies and spectrum between yeast RNA polymerase mutants. RNAs from wildtype (WT), E1103G, and ΔRpb9 yeast were sequenced via ARC-seq, with the number of cDNA copies per RNA molecule required to generate a consensus sequence varied from 1 thru 5. A: RNA mutation frequency stabilizes at 3 cDNAs per RNA molecule, revealing RNA mutation frequency differences between wildtype and the two mutants. B: ARCseq reveals subtle differences between the mRNA mutation spectrum of wildtype and mutant yeasts. Error bars represent Wilson scores of 95% confidence.

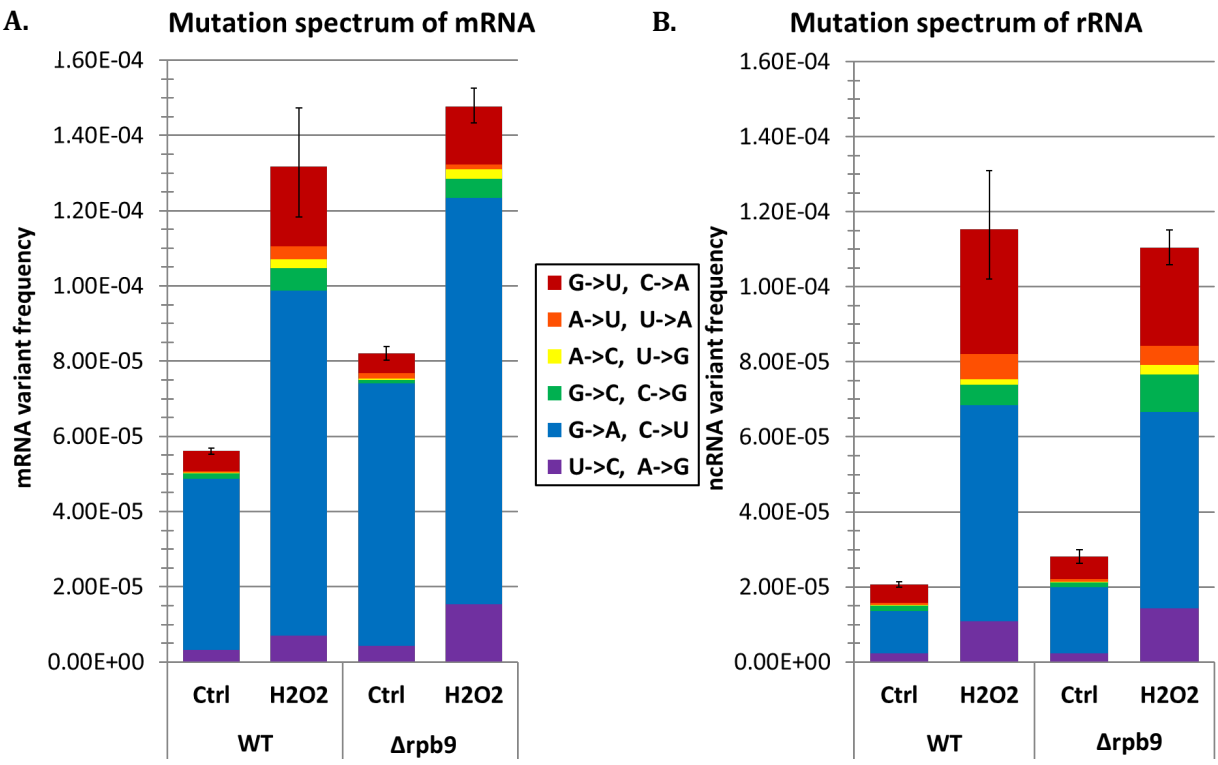


Figure 4: ARC-seq reveals differences in TM after oxidative stress between yeast RNA fidelity mutants and between RNA type. Wildtype (WT) and ΔRpb9 yeast were exposed to H_2O_2 and then their RNAs were sequenced via ARC-seq. A: mRNA frequency and spectrum in untreated (ctrl) and 50uM-treated (H2O2) wildtype and ΔRpb9 yeast. B: rRNA frequency and spectrum in untreated (ctrl) and 50uM-treated (H_2O_2) wildtype and ΔRpb9 yeast. Error bars represent Wilson scores 95% confidence.

SUPPLEMENTARY INFORMATION

Accurate RNA Consensus Sequencing (ARCseq) Protocol

Sequences of primers and adaptor for protocol

- RCRT-01
  - ARCseq adaptor
  - **5'- NNNNNNNNNNNNNNNN AGAUCGGAAGAGCACACGUCUGAACUCCAGUCAC GGCGCGCC CCUACACGACGCUCUUCCGAUCU GA -3'**
- RCRT-02
  - Primer for rolling-circle reverse transcription
  - **5' – TGCATTGATGGTGCCTACAG AGATCGGAAGAGCGTCGTGTAGG - 3'**
- AscIX oligo
  - Single-stranded oligo annealed in for multimer digestion
  - **5' - TGAACTCCAGTCACGGCGCGCCCCTACACGACGCT - 3'**
- RCRT-05 index2
  - Primer to index cDNA duplicates to distinguish btw cDNA and PCR duplicates
  - **5' – AATGATACGGCGACCACCGAGA TCT TTCA NNNNNNNNNNN ACACTCTTTCCCTACACGACGCTCTTCCGATCT GA – 3'**
- P5
  - Illumina 5'-primer
  - **5' - AATGATACGGCGACCACCGAGA – 3'**
- MWS-20
  - Illumina 3'-primer
  - **5' – GTGACTGGAGTTCAGACGTGTGC – 3'**
- MWS-21
  - Illumina indexing primer
  - **5' – CAAGCAGAAGACGGCATACGAGAT-XXXXXX-GTGACTGGAGTTCAGACGTGTGC – 3'**

Required reagents

- RNAClean XP beads
  - Agencourt, A66514
- AMPure XP beads
  - Agencourt, A63882
- RNA Fragmentation Reagent
  - ThermoFisher Scientific, AM8740
- RNAse inhibitor, murine
  - NEB, M0314L
- Shrimp Alkaline Phosphatase (rSAP)
  - NEB, M0371L
- OptiKinase
  - Affymerix, 78334Z
- DNAse, RNAse free
- 5'-DNA adenylation kit
  - NEB, E2610L
- T4 RNA Ligase 2, KQ, truncated
  - NEB, M0373L
- T4 RNA Ligase 1, high concentration
  - M0437M
- Protoscript II RT
  - NEB, M0368X

- AscI
  - NEB, R0558L
- BssHII
  - NEB, R0199L
- KAPA HiFi HotStart DNA Polymerase
  - Kapa Biosystems, KK2102
- Agilent TapeStation Reagents
  - High-sensitivity RNA and DNA screen tapes and reagents: 5067- -5584, -5585, -5579, and -5580

mRNA extraction

For mRNA extraction, use a method that minimized RNA damage and maximizes mRNA yield and purity.

Recommended kit (minimal mRNA damage and easy to use): Ambion's Dynabeads mRNA DIRECT Purification Kit (ThermoScientific, cat. no. 61011)

Store extracted mRNA in >50uL DEPC-treated TlowE + 200U RNAse inhibitor

Fragmentation using Ambion RNA Fragmentation Reagents

1. Preheat thermocycler to 70° C and hold
2. Aliquot 18uL each mRNA sample into a PCR tube
3. Add **2uL Fragmentase Solution**, mix & spin briefly
4. Incubate at **70° C** for 1-5 minutes, determined empirically for RNA source
5. Place on ice, add **2uL Stop Solution**, and mix

End-repair

1. Dephosphorylation of 5'- and 3'- ends of fragmented mRNA
   a. Mix the following components in a sterile nuclease-free PCR tube:

	1X	**___X MM**
Fragmented mRNA	22	---
10X Phosphatase buffer	3	
NFW	2	
RNAse inhibitor	1	
rSAP (shrimp alk. phosp)	2	
Total	***30***	*(8uL/smpl)*

   b. Mix & briefly spin to collect all liquid from the sides of the tube
   c. Place in a thermocycler, with the heated lid on, and run the following program:
   30 min @ 37° C
   Δ-inactivate @ 65° C x 5 min
   Hold @ 4° C

2. 5'-phosphorylation w/OptiKinase (PNK mutant) + DNAse treatment

DNA contamination is not uncommon even with good mRNA/rRNA ratios, post extraction; to remove any DNA that may still be present, add DNAse during repair of the 5'-ends

a. Add the following components to the PCR tubes from 1c above:

	1X	**___X MM**
P-tx'd mRNA from 1c	30	---
10X PNK buffer	10	
NFW	52	
***100**mM ATP*	1	
RNAse inhibitor	1	
T4 PNK	4	
DNAse	2	
Total	***183***	*(70uL/smpl)*

b. Mix & briefly spin to collect all liquid from the sides of the tube; split smpls btw 2 tubes each

c. Place in a thermocycler, with the heated lid on, and run the following program:

60 min @ 37° C

Add **5uL** **50mM** EDTA and mix, to inactivate DNAse

3. Sample clean-up & size-selection
   a. Add **150uL** (1.5X vol) Ampure's RNA CleanXP beads (pre-warmed 30' at room temperature) & mix well
   b. Incubate at room temperature for 2min then briefly spin down to collect smpl
   c. Place on magnet >3min (until all beads gather)
   d. Discard supernatant
   e. Wash beads with 200uL **80%** EtOH, freshly made (ON magnet)
   f. Discard supernatant; use P10 pipette to remove residual EtOH
   g. Incubate @ RT for **3min** to allow residual EtOH to evaporate (do NOT over-dry beads!)
   h. Remove from magnet & add **20uL** **DEPC-tx'd TlowE** & mix to elute RNA from beads
   i. Incubate 2min at room temperature then place on magnet
   j. Transfer supernatant to new tube

4. Quantification (***ESSENTIAL***)
   a. Mix 1uL high-sensitivity RNA sample dye with 1uL of RNA and 1uL NFW in PCR tubes
   b. Incubate at 72° C x 3 min then place on ice for 2min
   c. Run mix on TapeStation, using high-sensitivity RNA tapes
   d. Quantify the region 150-800fmol

Adaptor Ligation

1. 5'-adenylation of RNAs
   a. Make-up RNA + NFW mixes shown in table above in nuclease-free PCR tubes
   b. Add 10uL of the following master mix (MM) to each tube:

	1X	**____X MM**
End-repaired RNA + NFW	20	---
10X 5'-adenylation buffer	4	

DEPC-tx'd NFW	9	
***1mM** ATP*	4	
RNAse inhibitor	2	
Mth ligase (50pmol/uL)	1	
Total	***40***	*(20uL/smpl)*

*** NOTE: Mth ligase must be equimolar or better to RNA; the above recipe has Mth in 10-fold excess*

c. Mix & briefly spin to collect all liquid from the sides of the tube
d. Place in a thermocycler, with the heated lid on, and run the following program:
60 min @ 37° C
Δ-inactivate @ 80° C x 5 min
Hold @ 25° C

2. Sample clean-up
   a. Add **60uL** (1.5X vol) Ampure's RNA CleanXP beads (pre-warmed 30min at room temperature) & mix well
   b. Incubate at room temperature for 3min then briefly spin down to collect smpl
   c. Place on magnet >2min (until all beads gather)
   d. Discard supernatant
   e. 2 X 200uL **80%** EtOH washes, freshly made (ON magnet)
   f. Discard supernatant; use P10 pipette to remove residual EtOH
   g. Incubate @ RT for **3min** to allow residual EtOH to evaporate (do NOT over-dry beads!)
   h. Remove from magnet & add **20uL** **DEPC-tx'd NFW** & mix to elute RNA from beads
   i. Incubate 2min at room temperature then place on magnet
   j. Transfer supernatant to new tube

3. 5'-ligation of RCRT-01 adaptor, 1 rxn/smpl (PEG enables crowding, so no need to split smpls)

*** NOTE: **new** "100nM RCRT-01" = 33fmol/uL, according to the TapeStation*
∴ 1uM RCRT-01 = 330 fmol/uL

a. Add 4.85uL "1uM RCRT-01" adaptor to each RNA from 2k above
b. Incubate RNA + adaptor at **72° C** x 3 min then on **ice** x 2 min

	1X	**15X MM**
pApp RNA	20	---
***"1uM RCRT-01"** adaptor* **	4.85	---
DEPC-tx'd NFW	5.65	
10X RNA Ligase buffer	6	
50% PEG-8000	18	
RNAse inhibitor	1.5	
Rnl2 KQ	4	
Total	***60***	*(35.2uL/smpl)*

*** Note: Want >**4x molar ratio** of RCRT-01 adaptor to end-repaired RNA
c. Add 35.2 uL MM shown above and aliquot 15uL each RNA rxn mix into **4** tubes
d. Incubate @ 16° C overnight (19h)

- > 16h is preferable

4. Sample clean-up & size-selection, using *Zymo Research's RNA Clean & Concentrator kit*
   a. Combine 60uL (1X vol.) RNA Bind Buffer w/60uL (1X vol.) 100% EtOH & add mix to RNA ligation reaction from 3d
   b. Follow kit instructions for bind, prep, and wash steps
      - This should clean all nucleic acids <200nt away from larger, desired (i.e. ligated) products
   c. Elute RNA/DNA in 35uL NFW

Circularization of adapted RNAs

1. 5'-phosphorylation of adapted RNAs by PNK

* Note: RCRT-01 adaptor does not come phosphorylated on the 5'-end in order to minimize self-ligation during RNA:adaptor ligation; thus, it needs to be phosphorylated prior to circularization

   a. Add 15uL of the following MM to each adapted RNA from 4k above:

	1X	**____X MM**
RCRT-01- adapted RNA	35	---
DEPC-tx'd NFW	7.5	
10X PNK buffer	5	
100mM ATP	0.5	
RNAse inhibitor	1	
T4 PNK	2	
Total	***50***	*(15uL/smpl)*

   b. Incubate at **37° C** X **60 min**
   c. 25° C hold

2. Sample clean-up
   a. IMPORTANT: transfer reactions to a *NEW* PCR tube
   b. Add 50uL (1X vol) Ampure XP beads (pre-warmed 30min at room temperature) & mix well
   c. Incubate at room temperature for 3min then briefly spin down to collect smpl
   d. Place on magnet >2min (until all beads gather)
   e. Discard supernatant
   f. Wash beads **2X** w/200uL **75%** EtOH, freshly made, w/vortexing/magnet shift
   g. Discard supernatant; use P10 pipette to remove residual EtOH
   h. Incubate @ RT for **3min** to allow residual EtOH to evaporate (do NOT over-dry beads!)
   i. Remove from magnet & add **20uL** **DEPC-tx'd NFW** & mix to elute RNA from beads
   j. Incubate 2min at room temperature then place on magnet
   k. Transfer supernatant to new tube

3. Circularization of 5'-phosphorylated, RCRT-01-adapted RNAs, 2 rxns/smpl
   a. Incubate RNA from 2k @ **72° C** x 3 min then on **ice** x 2 min
   b. Add 20 uL MM shown below to each RNA

	1X	_____X MM
RCRT-01- adapted RNA	20	---
DEPC-tx'd NFW	3.5	
10X RNA Ligase buffer	4	
1mM ATP	2	
50% PEG-8000	8	
RNAse inhibitor	1	
Rnl1, high concentration	1.5	
Total	***40***	*(20uL/smpl)*

c. Incubate @ **37°** C X **2h**
d. Cool to 16° C
e. Add **1uL** add'l Rnl1 HC enzyme
e. Incubate @ **16° C overnight** (21.5h)
   - > 16h is preferable

4. Sample clean-up
   a. Add 60uL NFW to increase volume to 100uL
   b. Add 100uL (1X vol) Ampure's RNA CleanXP beads (pre-warmed 30min at room temperature) & mix well
   c. Incubate at room temperature for 3min then briefly spin down to collect smpl
   d. Place on magnet >2min (until all beads gather)
   e. Discard supernatant
   f. Wash beads with 200uL **80%** EtOH, freshly made (ON magnet)
   g. Discard supernatant; use P10 pipette to remove residual EtOH
   h. Incubate @ RT for **3min** to allow residual EtOH to evaporate (do NOT over-dry beads!)
   i. Remove from magnet & add **20uL** **NFW** & mix to elute RNA from beads
   j. Incubate 2min at room temperature then place on magnet
   k. Transfer supernatant to new tube

Rolling-circle reverse transcription

1. Setup overnight RC-RT reactions, 4 rxns/smpl
   a. Add 4uL 10uM RCRT-02 RT-primer to each circularized RNA from 4l above
   b. Incubate RNA + primer @ **72° C** x 3 min then on **ice** x 2 min
   c. Add 56uL of the following MM to each RNA+primer mix and aliquot 20uL each RNA rxn mix into **4** tubes

	1X	___X MM
Circularized RNA	20	---
10uM RCRT-02 primer	4	---
DEPC-tx'd NFW	28	
10X ProtoScript buffer	8	
10mM dNTPs	6	
10X DTT	8	
RNAse inhibitor	2	
ProtoScript RT	4	
Total	***80***	*(56uL/smpl)*

   d. Incubate @ **42° C overnight** (21.5h)

- > 20h is preferable

2. Size-select RC-RT rxn products
    a. Transfer smpls to 1.5mL epi tube
    b. Add 120uL NFW to increase volume to 200uL
    c. Add **120uL** (**0.6X** vol) Ampure XP beads (pre-warmed 30min at room temperature) & mix well
    d. Incubate at room temperature for 3min then briefly spin down to collect smpl
    e. Place on magnet >2min (until all beads gather)
    f. Discard supernatant
    g. Wash beads with **500uL 80%** EtOH, freshly made, *vortexing* to resuspend beads
    h. Discard supernatant; repeat wash w/**200uL 80%** EtOH, vortexing again
    i. Discard supernatant; use P10 pipette to remove residual EtOH
    j. Incubate @ RT for **3min** to allow residual EtOH to evaporate (do NOT over-dry beads!)
    k. Remove from magnet & add **35uL** **NFW** & mix to elute cDNA from beads
    l. Incubate 2min at room temperature then place on magnet
    m. Transfer supernatant to PCR tube

Digest multimeric cDNAs into monomers w/AscI and BssHII

* Anneal RC ssDNA → cut → repeat 2x

1. Setup AscI digestions of multimeric cDNAs from 2m above, 2 rxns/smpl
    a. Add the **23uL** of following MM to cDNAs from 2m above:

	1X	**____X MM**
Multimeric cDNA	35	---
NFW	6.5	
10X CutSmart buffer	5	
100X BSA	0.5	
10uM AscI X oligo	1	
AscI	2	---
Total	***50***	*(23uL/smpl)*

    b. Heat cDNA + MM to **90° C X 3min**, anneal @ **65° C X 30sec**, cool to **37° C**
    c. Add **2uL AscI** to each tube
    d. Incubate rxns @ **37° C X 1h**
    e. REPEAT step b.)
    f. Add **1uL add'l AscI** to each tube
    g. Incubate rxns @ **37° C X 30min**
    h. Add **1uL** add'l **10uM AscI X oligo** to each tube
    i. Heat cDNA + MM to **90° C X 3min**, anneal @ **65° C X 30sec**, cool to **50° C**
    j. Add **1uL BssHII** to each tube
    k. Incubate rxns @ **50° C X 30min**

5. Sample clean-up
    a. Add **50uL** (**1X vol**) Ampure XP beads (pre-warmed 30min at room temperature) & mix well
    b. Incubate at room temperature for 3min then briefly spin down to collect smpl

c. Place on magnet >2min (until all beads gather)
d. Discard supernatant
e. Wash beads **2X** w/200uL **80%** EtOH, freshly made, w/pipette-resuspension of beads
f. Discard supernatant; use P10 pipette to remove residual EtOH
g. Incubate @ RT for **3min** to allow residual EtOH to evaporate (do NOT over-dry beads!)
h. Remove from magnet & add **40uL** **TlowE** & mix to elute cDNA from beads
i. Incubate 2min at room temperature then place on magnet
j. Transfer supernatant to new tube

Index2 Extensions (tail-in PCR-ID)

1. Setup extension reactions, with the RCRT Index2 primer, to tail-in the PCR-ID (2rxns/smpl)
   a. Add 60uL of the following MM to each monomerized cDNA smpl

	1X	**___X MM**
NFW	34	
5X Kapa Buffer	20	
10mM dNTPs	2	
10uM RCRT-05 index2	2	
Kapa HiFi poly.	2	
monomeric cDNA	40	---
Total	***100***	*(60uL/smpl)*

   b. Mix & briefly spin to collect all liquid from the sides of the tube; *aliquot 50uL into each of **2** tubes*
   c. Place in a thermocycler, with the heated lid on, and run the following program:

95° C X 4min

Then **3** cycles:

- 98° C X 20sec
- 57° C X 40sec
- 72° C X 4min

Hold @ 25° C

2. Sample clean-up
   a. IMPORTANT: pool 2rxns/smpl aliquots into a single PCR tube
   b. Add **80uL** (**0.8X vol**) Ampure XP beads (pre-warmed 30min at room temperature) & mix well
   c. Incubate at room temperature for 3min then briefly spin down to collect smpl
   d. Place on magnet >2min (until all beads gather)
   e. Discard supernatant
   f. Wash beads **2X** w/200uL **80%** EtOH *(ON MAGNET)*
   g. Discard supernatant; use P10 pipette to remove residual EtOH
   h. Incubate @ RT for **3min** to allow residual EtOH to evaporate (do NOT over-dry beads!)
   i. Remove from magnet & add **40uL** **TlowE** & mix to elute cDNA from beads
   j. Incubate 2min at room temperature then place on magnet
   k. Transfer supernatant to new tube

PCR1

1. Setup PCR 1 to amplify the extended products from 2k
    a. Add 60uL of the following MM to each extended DNA smpl and aliquot 50uL into each of **2** tubes

	1X	___X MM
NFW	24	
5X Kapa Buffer	20	
10mM dNTPs	4	
10uM P5 primer	4	
20uM MWS-20 primer	2	
12.5X Sybr	4	
Kapa HiFi poly.	2	
extended DNAs	40	---

    b. Mix & briefly spin to collect all liquid from the sides of the tube
    c. PCR conditions

95° C X 4min
X cycles:
98° C X 20sec
60° C X 20sec
72° C X 15sec

Pull reaction(s) as they approach saturation
Note cycle no. for each sample

** For 100fmol input into adaptor ligation, samples should come up around 11-13 cycles; significantly later could indicate issues with previous steps

Smpl. no.	No. of cycles
1	
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	
13	
14	

2. Sample clean-up
    a. IMPORTANT: pool 2rxns/smpl aliquots into a single *NEW* PCR tube
    b. Add **80uL (0.8X vol)** Ampure XP beads (pre-warmed 30min at room temperature) & mix well
    c. Incubate at room temperature for 3min then briefly spin down to collect smpl
    d. Place on magnet >2min (until all beads gather)
    e. Discard supernatant
    f. Wash beads **2X** w/200uL **80%** EtOH *(ON MAGNET)*
    g. Discard supernatant; use P10 pipette to remove residual EtOH
    h. Incubate @ RT for **3min** to allow residual EtOH to evaporate (do NOT over-dry beads!)
    i. Remove from magnet & add **50uL** **TlowE** & mix to elute cDNA from beads
    j. Incubate 2min at room temperature then place on magnet
    k. Transfer supernatant to new tube

Index PCR

IMPORTANT: In this PCR, the second primer, MWS-20, is replaced with a sample-specific index primer, MWS-21-*index*. Each sample gets a **different** primer, which has a different index. This allows multiplexing samples on sequencing lanes and demuxing them later.

1. Amplify captured DNA, running ONE tube per reaction
   a. Mix the following components:

	1X	**___X MM**
NFW	13	
5X Kapa Buffer	10	
10mM dNTPs	2	
10uM P5	2	
10uM MWS-21-index	**2**	---
Kapa HiFi poly.	1	
DNA	20	---

Add 2uL sample-specific MWS-21-index to each DNA
Add 28uL MM to each DNA+index smpl & mix well; spin briefly to collect

   a. PCR conditions

95° C X 4min
6 cycles:
98° C X 20sec
60° C X 20sec
72° C X 15sec

Pull reaction(s) as they approach saturation

2. Sample clean-up & size-selection

*NOTE: the bead wash steps for post-index PCR clean-up are more robust because the index primer tends to dimerize, which will waste sequencing space if residual primer dimers are present in the sequencing mix.

   a. Add **40uL** (0.8X vol) Ampure XP beads (pre-warmed 30min at room temperature) & mix well
   b. Incubate at room temperature for 3min then briefly spin down to collect smpl
   c. Place on magnet >2min (until all beads gather)
   d. Discard supernatant
   e. Wash beads 2 X with 200uL 80% EtOH & pipete up-and-down to resuspend
   f. Place on magnet >2min & discard super
   g. Use P10 pipette to remove residual EtOH
   h. Remove from magnet & add **30uL** NFW & mix to elute DNA from beads
   i. Incubate 2min at room temperature then place on magnet
   j. Transfer supernatant to new tube

Sequencing mix prep

1. Post-ligation quantitation
    2. Dilute 1:100
    3. Mix 2uL high-sensitivity sample dye with 2uL of DNA (1:100 dilution) in PCR tubes
    4. Run mix on TapeStation, using high-sensitivity tapes
    5. Quantify the region 200-950fmol

Smpl. ID	*Index*	*1/100X fmol/uL*	*fmol/uL*	*frac. wanted (of 2 lanes)*	*uL 1/100X needed*	*5X vol for mix*	*5x fmol*

uL smpl ==

uL NFW ==

== total fmol in mix

== fmol/uL in mix

CONCLUSIONS AND PERSPECTIVES

Proper cell functioning is critically important the processes of DNA replication and RNA transcription being accurate: the *fidelity* of these processes is crucial; infidelity can lead to cellular dysfunction and disease. The evidence for replication and transcription infidelity are present in DNA and RNA mutations, respectively. In order to study the fidelity of DNA replication and RNA transcription *in vivo*, it is necessary to sequence individual DNA and RNA molecules to detect rare mutations in a high throughput manner. Until recently, this has not been possible, as accurate detection of rare mutations among a preponderance of wildtype molecules has been limited by the high error rate of available methods. Conventional sequencing methods are only able to reliably detect mutations present in 5% or more of molecules sequenced. Rare mutations are lost during this consensus making; consequently, we have been unable to accurately measure the frequency of rare events, which are the evidence of the fidelity of replication and transcription. The solution to this problem is to perform single molecule sequencing of individually barcoded DNA and RNA molecules.

Here I have presented three projects which apply the use of barcoding individual DNA and RNA molecules in order to enable highly accurate and sensitive analyses of DNA replication and RNA transcription fidelity. Duplex Sequencing is an immensely powerful method, the potential of which is only recently starting to be realized. The two studies employing its use herein reveal the power of Duplex Sequencing to address questions that have until now remained a mystery due to the lack of tools available. The question of why CS patients don't get cancer has puzzled doctors and scientists for decades. While many have speculated as to the cause, we have finally definitively answered the question: CS patients fail to develop cancer because they do not accumulate mutations more quickly than repair-proficient individuals. The broader implication of this finding is that it provides evidence supporting the existence of a mutator phenotype in cancer (XP-C patients accumulate excessive UV-induced mutations and, consequently, develop cancer at incredibly high rates), as well as potentially indicating that allowing some mutations to occur is crucial for normal organism survival (CS patients

fail to repair or bypass UV-induced damage and, consequently, undergo excessive cell death, leading to tissue senescence and accelerated aging). This small study was a nice example of the balance between mechanisms governing aging and cancer and how imbalance in either direction can have profound consequences.

The question of why GBM patients do so poorly and always recur has long plagued doctors and scientists. Here, we expand on the excellent clonal mutation work of our predecessors, revealing that the substantial inter- and intra-tumoral clonal heterogeneity is further compounded by considerable subclonal heterogeneity. It is disconcerting that GBMs have so many levels of heterogeneity: between patients, between regions of a tumor, and between primary and recurrent tumors within each patient. Our results emphasize the need for in-depth analyses of individual patient's tumors to determine which mutations are already present and devise treatment strategies unique to each patient's tumor. While costs and treatment options may preclude this from being a viable option at this time, our hope is that with further analyses and correlating clinical data, patient history, and mutation (clonal and subclonal) data, we can devise treatment options that may one day prove effective in GBM treatment; the pipe dream being that we'll discover how GBM occur in the first place and be able to put in place monitoring systems to detect early stage GBM and prevent it becomes problematic.

In my quest to address the question of the role of RNA fidelity *in vivo*, the challenge was first to figure out how to do so, as no method yet existed. When I began developing an accurate RNA sequencing method, my goal was simply to be able to generate and link at least two copies of cDNA derived from each RNA molecule. This proved more difficult than I anticipated. Initially, I attempted to take advantage of flu's RNA-directed RNA polymerase to first generate a double-stranded RNA from which I could make two cDNAs. However, flu RNA polymerase proved more error-prone than I expected based on the *in vitro* studies of its fidelity. Next, I sought to take advantage of a biotinylated adaptor to recover RNA between cDNA synthesis steps, thereby enabling the same RNA to be reverse

transcribed multiple times. Despite my best efforts to maximize the yield of each round of reverse transcription, however, this proved highly inefficient; I was only somewhat consoled by the publication of a similar approach, examination of which revealed a similar inefficiency problem.

Finally, I discovered that reverse transcriptase strand-displaces and could, therefore, be used to generate a multimeric cDNA from a circularized RNA. Eureka! I had already optimized the ligation conditions for multiple RNA ligases; two of which proved to be immensely useful in my method. After a few more hurdles, remembering that biotin can sterically hinder enzymes and, thus, interfere with efficient reverse transcription and realizing that I would need a way to distinguish between PCR duplicates of a single cDNA and cDNA duplicates of a single RNA, I finally succeeded in developing an efficient method of generating multiple cDNA copies from single RNA molecules. To ensure that my method was able to fully eliminate artifacts and, thus, was indeed a superior approach than the methods that had been published during my time developing my own, I painstakingly quality controlled every step of the protocol to ensure its efficiency and accuracy was optimized. And, thus, after two years of failure and frustration, I finally succeeded in developing ARC-seq.

The question of what, if any, contribution RNA mutations have in health and disease has been one that has remained unanswered for more than 50 years. ARC-seq will enable studies on how perturbing a cell's environment or machinery affects the fidelity of transcription and to what extent RNA mutations contribute to aging, cancer, and neurodegeneration, as well as the evolution and acquired resistance of viruses and bacteria. While I was not able to apply ARC-seq to all the studies I had in mind when I first began developing an accurate RNA sequencing method, my hope is that others will see potential power and utility and pursue studies of the roles of transcriptional infidelity in health and diseases.

BIBLIOGRAPHY

1. Fox, E.J., et al., *Accuracy of Next Generation Sequencing Platforms.* Next Gener Seq Appl, 2014. **1**(1): p. 106.
2. Burrell, R.A. and C. Swanton, *Tumour heterogeneity and the evolution of polyclonal drug resistance.* Mol Oncol, 2014. **8**(6): p. 1095-111.
3. Kennedy, S.R. and D.A. Erie, *Templated nucleoside triphosphate binding to a noncatalytic site on RNA polymerase regulates transcription*, in *Proc Natl Acad Sci USA*. 2011. p. 10.1073/pnas.1011274108.
4. Schmitt, M.W., et al., *Detection of ultra-rare mutations by next-generation sequencing.* Proc Natl Acad Sci U S A, 2012. **109**(36): p. 14508-13.
5. Ji, J.P. and L.A. Loeb, *Fidelity of HIV-1 reverse transcriptase copying RNA in vitro.* Biochemistry, 1992. **31**(4): p. 954-8.
6. Wang, Z., M. Gerstein, and M. Snyder, *RNA-Seq: a revolutionary tool for transcriptomics.* Nat Rev Genet, 2009. **10**(1): p. 57-63.
7. Lauring, A.S., J. Frydman, and R. Andino, *The role of mutational robustness in RNA virus evolution.* Nat Rev Microbiol, 2013. **11**(5): p. 327-36.
8. Goldsmith, M. and D.S. Tawfik, *Potential role of phenotypic mutations in the evolution of protein expression and stability.* Proc Natl Acad Sci U S A, 2009. **106**(15): p. 6197-202.
9. Morreall, J.F., L. Petrova, and P.W. Doetsch, *Transcriptional mutagenesis and its potential roles in the etiology of cancer and bacterial antibiotic resistance.* J Cell Physiol, 2013. **228**(12): p. 2257-61.
10. Kennedy, S.R., et al., *Ultra-Sensitive Sequencing Reveals an Age-Related Increase in Somatic Mitochondrial Mutations That Are Inconsistent with Oxidative Damage.* Plos Genetics, 2013. **9**(9).
11. Cleaver, J.E., E.T. Lam, and I. Revet, *Disorders of nucleotide excision repair: the genetic and molecular basis of heterogeneity. .* Nat Rev Gen, 2009. **10**: p. 756-768.
12. Kraemer, K.H., M.M. Lee, and J. Scotto, *DNA repair protects against cutaneous and internal neoplasia:evidence from xeroderma pigmentosum.* Carcinogenesis, 1984. **5**: p. 511-514.
13. Bradford, P.T., et al., *Cancer and neurologic degeneration in xeroderma pigmentosum: long term follow-up characterises the role of DNA repair.* J Med Genet, 2011. **48**: p. 168-176.
14. Nance, M.A. and S.A. Berry, *Cockayne syndrome:review of 140 cases.* Amer J Med Genet, 1992. **42**: p. 68-84.
15. Zhang, W.R., et al., *Survey of Cockayne patients reports no skin cancers despite DNA repair deficiency.* J Amer Acad Dermatol, 2016 (in press).
16. Lehmann, A.R., *Cockayne syndrome and trichothiodystrophy: defective repair without cancer.* Cancer reviews, 1987. **7**: p. 82-103.
17. Wilson III, D.M. and V.A. Bohr, *Special issue on the segmental progeria Cockayne syndrome.* Mech Aging Dev, 2013. **134**: p. 159-160.
18. Kubota, M., et al., *Nationwide survey of Cockayne syndrome in Japan: Incidence, clinical course and prognosis.* PediatrInt, 2015. **57**: p. 339-347.
19. Loeb, L., *Mutator phenotype may be required for multistage carcinogenesis.* Cancer Research, 1991. **51**: p. 3075-3079.
20. Martincorena, I., et al., *High burden and pervasive positive selection of somatic mutations in normal human skin.* Science, 2015. **348**: p. 880-886.
21. Maher, V.M., et al., *DNA excision repair processes in human cells can eliminate the cytotoxic and mutagenic consequences of ultraviolet irradiation.* Mutation Research, 1979. **62**: p. 311-323.

22. Maher, V.M., et al., *Frequency of ultraviolet light-induced mutations is higher in xeroderma pigmentosum variant cells than in normal human cells.* Nature, 1976. **261**: p. 593-595.
23. Lin, Y.W., et al., *Somatic cell mutation frequency at the HPRT, T-cell antigen receptor and glycophorin A loci in Cockayne syndrome.* Mutat Res, 1995. **337**: p. 49-55.
24. Parris, C.H. and K.H. Kraemer, *Ultraviolet-light induced mutations in Cockayne syndrome cells are primarily caused by cyclobutane dimer photoproducts while repair of other photoproducts is normal.* Proceedings of the National Academy of Sciences USA, 1993. **90**: p. 7260-7264.
25. Muriel, W.J., J.R. Lamb, and A.R. Lehmann, *UV mutation spectra in cell lines from patients with Cockayne's syndrome and ataxia telangiectasia using the shuttle vector pZ189.* Mutat Res, 1991. **254**: p. 119-123.
26. Fox, E.J., et al., *Accuracy of Next Generation Sequencing Platforms.* Next Gener Seq Appl, 2014. **1**: p. 1000106.
27. Schmitt, M.W., et al., *Detection of ultra-rare mutations by next-generation sequencing.* Proc Natl Acad Sci USA, 2012. **109**: p. 14508-14513.
28. Kennedy, S.R., et al., *Detecting ultralow-frequency mutations by Duplex Sequencing.* Nat Protoc, 2014. **9**: p. 2586-2606.
29. Aamann, M.D., et al., *Multiple interaction partners for Cockayne syndrome proteins: implications for genome and transcriptome maintenance.* Mech Ageing Dev., 2013. **134**: p. 212-224.
30. Spivak, G. and P.C. Hanawalt, *Host cell reactivation of plasmids containing oxidative DNA lesions is defective in Cockayne syndrome but normal in UV-sensitive syndrome fibroblasts.* DNA Repair (Amst), 2006. **5**: p. 13-22.
31. D'Errico, M., et al., *The role of CSA in the response to oxidative DNA damage in human cells.* Oncogene, 2007. **26**: p. 4336-4343.
32. Hoeijmakers, J.H., *Genome maintenance mechanisms for preventing cancer.* Nature, 2001. **411**: p. 366-374.
33. Pasquier, L., et al., *Wide clinical variability among 13 new Cockayne syndrome cases confirmed by biochemical assays.* Arch Dis Child, 2006. **91**: p. 178-182.
34. Cheng, K.C., et al., *8-Hydroxyguanine, an abundant form of oxidative DNA damage, causes G---- T and A----C substitutions.* J Biol Chem, 1992. **267**: p. 166-172.
35. Shigenaga, M.K., C.J. Gimeno, and B.N. Ames, *Urinary 8-hydroxy-2'-deoxyguanosine as a biological marker of in vivo oxidative DNA damage.* Proc Natl Acad Sci USA., 1989. **86**: p. 9697-9701.
36. Gajewski, E., et al., *Modification of DNA bases in mammalian chromatin by radiation-generated free radicals.* Biochem, 1990. **29**: p. 7876-7882.
37. Jiang, X.R., et al., *Telomerase expression in human somatic cells does not induce changes associated with a transformed phenotype.* Nat Genet, 1999. **21**: p. 111-114.
38. Morales, C.P., et al., *Absence of cancer-associated changes in human fibroblasts immortalized with telomerase.* Nat Genet, 1999. **21**: p. 115-118.
39. Milyavsky, M., et al., *Prolonged culture of telomerase-immortalized human fibroblasts leads to a premalignant phenotype.* Cancer Res, 2003. **63**: p. 7147-7157.
40. van Waarde-Verhagen, M.A., H.H. Kampinga, and M.H. Linskens, *Continuous growth of telomerase-immortalised fibroblasts: how long do cells remain normal?* Mech Ageing Dev, 2006. **127**: p. 85-87.
41. MacKenzie, K.L., et al., *Mass cultured human fibroblasts overexpressing hTERT encounter a growth crisis following an extended period of proliferation.* Exp Cell Res., 2000. **259**: p. 336-350.
42. Lee, D.H. and G.P. Pfeifer, *Translesion synthesis of 7,8-dihydro-8-oxo-2'-deoxyguanosine by DNA polymerase eta in vivo.* Mutat Res, 2008. **641**: p. 19-26.

43. Washington, M.T., et al., *Fidelity and processivity of Saccharomyces cerevisiae DNA polymerase eta.* J Biol Chem, 1993. **274**: p. 36835-36838.
44. Washington, M.T., et al., *The mechanism of nucleotide incorporation by human polymerase eta differs from that of the yeast enzyme.* Mol Cell Biol, 2003. **23**: p. 8316-8322.
45. Tang, M., et al., *Roles of E. coli DNA polymerases IV and V in lesion-targeted and untargeted SOS mutagenesis.* Nature, 2000. **404**: p. 1014-1018.
46. Rapin, I., et al., *Cockayne syndrome in adults: review with clinical and pathologic study of a new case.* J Child Neurol, 2006. **21**: p. 991-1006.
47. Armstrong, B.K. and A. Kricker, *The epidemiology of UV induced skin cancer.* Journal of Photochemistry and Photobiology B: Biology, 2001. **63**: p. 8-18.
48. Ferrucci, L.M., et al., *Indoor tanning and risk of early-onset basal cell carcinoma.* J Am Acad Dermatol, 2012. **67**: p. 552-562.
49. Brash, D.E., et al., *The DNA damage signal for Mdm2 regulation, Trp53 induction, and sunburn cell formation in vivo originates from actively transcribed genes.* J Invest Dermatol., 2001. **117**: p. 1234-1240.
50. Lehmann, A.R., S.Kirk-Bell, and L.Mayne, *Abnormal kinetics of DNA synthesis in ultraviolet light-irradiated cells from patients with Cockayne syndrome.* Cancer Research, 1979. **39**: p. 4237-4241.
51. Cleaver, J.E., *Normal reconstruction of DNA supercoiling and chromatin structure in cockayne syndrome cells during repair of damage from ultraviolet light.* Am J Hum Genet., 1982. **34**: p. 566-575.
52. McKay, B.C., et al., *Ultraviolet light-induced apoptosis is associated with S-phase in primary human fibroblasts.* DNA Repair (Amst). 2002. **1**: p. 811-820.
53. Carvalho, H., et al., *Effect of cell confluence on ultraviolet light apoptotic responses in DNA repair deficient cells.* Mutat Res, 2003. **544**: p. 159-166.
54. Tresini, M., et al., *The core spliceosome as target and effector of non-canonical ATM signalling.* Nature, 2015. **523**: p. 53-58.
55. Hamdi, M., et al., *DNA damage in transcribed genes induces apoptosis via the JNK pathway and the JNK-phosphatase MKP-1.* Oncogene, 2005. **24**: p. 7135-7144.
56. Sollier, J., et al., *Transcription-coupled nucleotide excision repair factors promote R-loop-induced genome instability.* Mol Cell, 2014. **56**: p. 777-785.
57. Bowden, N.A., et al., *Understanding Xeroderma Pigmentosum Complementation Groups Using Gene Expression Profiling after UV-Light Exposure.* Int J Mol Sci, 2015. **16**: p. 15985-15996.
58. Nouspikel, T. and P. Hanawalt, *Terminally differentiated human neurons repair transcribed genes but display attenuated global DNA repair and modulation of repair gene expression.* Molecular Cell Biology, 2000. **20**: p. 1562-1570.
59. Hayflick, l., *The limited in vitro lifetime of human diploid cell strains.* Exp Cell Res, 1965. **37**: p. 614-636.
60. Cleaver, J.E., et al., *Xeroderma pigmentosum group C in an isolated region of Guatemala.* J Invest Dermatol, 2007. **127**: p. 493-496.
61. Ridley, A.J., et al., *Characterisation of novel mutations in Cockayne syndrome type A and xeroderma pigmentosum group C subjects.* J Hum Genet., 2005. **50**: p. 151-154.
62. Jaspers, N.G., et al., *Anti-tumour compounds illudin S and irofulven induce DNA lesions ignored by global repair and exclusively processed by transcription- and replication-coupled repair pathways.* DNA Repair (Amst) 2002. **1**: p. 1027-1038.
63. Li, H. and R. Durbin, *Fast and accurate short read alignment with Burrows–Wheeler transform.* Bioinformatics, 2009. **25**: p. 1754–1760.
64. Calabrese, E.J., D.Y. Shamoun, and J.C. Hanekamp, *Cancer risk assessment: Optimizing human health through linear dose-response models.* Food Chem Toxicol., 2015. **81**: p. 137-40.

65. Dolecek, T.A., et al., *CBTRUS statistical report: primary brain and central nervous system tumors diagnosed in the United States in 2005-2009.* Neuro Oncol, 2012. **14 Suppl 5**: p. v1-49.
66. Stupp, R., et al., *Effects of radiotherapy with concomitant and adjuvant temozolomide versus radiotherapy alone on survival in glioblastoma in a randomised phase III study: 5-year analysis of the EORTC-NCIC trial.* Lancet Oncol, 2009. **10**(5): p. 459-66.
67. Stupp, R., et al., *Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma.* N Engl J Med, 2005. **352**(10): p. 987-96.
68. Parker, N.R., et al., *Molecular heterogeneity in glioblastoma: potential clinical implications.* Front Oncol, 2015. **5**: p. 55.
69. Alexandrov, L.B., et al., *Deciphering signatures of mutational processes operative in human cancer.* Cell reports, 2013. **3**(1): p. 246-59.
70. TCGA, *Comprehensive genomic characterization defines human glioblastoma genes and core pathways.* Nature, 2008. **455**(7216): p. 1061-8.
71. Loeb, L.A., *Human cancers express mutator phenotypes: origin, consequences and targeting.* Nature reviews. Cancer, 2011. **11**(6): p. 450-7.
72. Loeb, L.A., C.F. Springgate, and N. Battula, *Errors in DNA replication as a basis of malignant changes.* Cancer research, 1974. **34**(9): p. 2311-21.
73. Burns, M.B., N.A. Temiz, and R.S. Harris, *Evidence for APOBEC3B mutagenesis in multiple human cancers.* Nat Genet, 2013. **45**(9): p. 977-83.
74. Roberts, S.A., et al., *An APOBEC cytidine deaminase mutagenesis pattern is widespread in human cancers.* Nature genetics, 2013. **45**(9): p. 970-6.
75. Burns, M.B., N.A. Temiz, and R.S. Harris, *Evidence for APOBEC3B mutagenesis in multiple human cancers.* Nature genetics, 2013. **45**(9): p. 977-83.
76. Kuong, K.J. and L.A. Loeb, *APOBEC3B mutagenesis in cancer.* Nature genetics, 2013. **45**(9): p. 964-5.
77. Kunkel, T.A., *DNA replication fidelity.* J Biol Chem, 2004. **279**(17): p. 16895-8.
78. Loeb, L.A., *Apurinic sites as mutagenic intermediates.* Cell, 1985. **40**(3): p. 483-4.
79. Sagher, D. and B. Strauss, *Insertion of nucleotides opposite apurinic/apyrimidinic sites in deoxyribonucleic acid during in vitro synthesis: uniqueness of adenine nucleotides.* Biochemistry, 1983. **22**(19): p. 4518-26.
80. Kumar, A., et al., *Deep sequencing of multiple regions of glial tumors reveals spatial heterogeneity for mutations in clinically relevant genes.* Genome Biol, 2014. **15**(12): p. 530.
81. Sottoriva, A., et al., *Intratumor heterogeneity in human glioblastoma reflects cancer evolutionary dynamics.* Proc Natl Acad Sci U S A, 2013. **110**(10): p. 4009-14.
82. Misra, A., et al., *Extensive intra-tumor heterogeneity in primary human glial tumors as a result of locus non-specific genomic alterations.* J Neurooncol, 2000. **48**(1): p. 1-12.
83. Kennedy, S.R., et al., *Detecting ultralow-frequency mutations by Duplex Sequencing.* Nat Protoc, 2014. **9**(11): p. 2586-606.
84. *COSMIC database online*. 2016; Available from: cancer.sanger.ac.uk.
85. Forbes, S.A., et al., *COSMIC: exploring the world's knowledge of somatic mutations in human cancer.* Nucleic Acids Res, 2015. **43**(Database issue): p. D805-11.
86. Yan, H., et al., *IDH1 and IDH2 mutations in gliomas.* N Engl J Med, 2009. **360**(8): p. 765-73.
87. Killela, P.J., et al., *TERT promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal.* Proc Natl Acad Sci U S A, 2013. **110**(15): p. 6021-6.
88. *Database of Single Nucleotide Polymorphisms (dbSNP). Bethesda (MD): National Center for Biotechnology Information, National Library of Medicine. (dbSNP Build ID: {build ID}).* *Available from: http://www.ncbi.nlm.nih.gov/SNP/*.
89. Vijg, J. and Y. Suh, *Genome instability and aging.* Annu Rev Physiol, 2013. **75**: p. 645-68.

90. Hanahan, D. and R.A. Weinberg, *The hallmarks of cancer.* Cell, 2000. **100**(1): p. 57-70.
91. Drummond, D.A. and C.O. Wilke, *The evolutionary consequences of erroneous protein synthesis.* Nat Rev Genet, 2009. **10**(10): p. 715-24.
92. Bregeon, D. and P.W. Doetsch, *Transcriptional mutagenesis: causes and involvement in tumour development.* Nature Reviews Cancer, 2011. **11**(3): p. 218-U88.
93. Selbach, *Global quantification of mammalian gene expression control.* 2011.
94. Gordon, A.J., et al., *Heritable change caused by transient transcription errors.* PLoS Genet, 2013. **9**(6): p. e1003595.
95. Blank, A., et al., *An RNA polymerase mutant with reduced accuracy of chain elongation.* Biochemistry, 1986. **25**(20): p. 5920-8.
96. Rosenberger, R.F. and J. Hilton, *The Frequency of Transcriptional and Translational Errors at Nonsense Codons in the Lacz Gene of Escherichia-Coli.* Molecular & General Genetics, 1983. **191**(2): p. 207-212.
97. Ninio, J., *Connections between Translation, Transcription and Replication Error-Rates.* Biochimie, 1991. **73**(12): p. 1517-1523.
98. Remington, K.M., et al., *Highly Mutagenic Bypass Synthesis by T7 RNA Polymerase of Site-specific Benzo[a]pyrene Diol Epoxide-adducted Template DNA.* Journal of Biological Chemistry, 1998. **273**(21): p. 13170-13176.
99. Kireeva, M.L., et al., *Transient reversal of RNA polymerase II active site closing controls fidelity of transcription elongation.* Mol Cell, 2008. **30**(5): p. 557-66.
100. Guo, J., P.C. Hanawalt, and G. Spivak, *Comet-FISH with strand-specific probes reveals transcription-coupled repair of 8-oxoGuanine in human cells.* Nucleic Acids Res, 2013. **41**(16): p. 7700-12.
101. Doetsch, P.W., *Translesion synthesis by RNA polymerases: occurrence and biological implications for transcriptional mutagenesis.* Mutat Res, 2002. **510**(1-2): p. 131-40.
102. Tornaletti, S., et al., *Effect of thymine glycol on transcription elongation by T7 RNA polymerase and mammalian RNA polymerase II.* J Biol Chem, 2001. **276**(48): p. 45367-71.
103. Liu, J., W. Zhang, and P.W. Doetsch, *Rna-Polymerase Bypass at Sites of Dihydrouracil - Implications for Transcriptional Mutagenesis.* Molecular and Cellular Biology, 1995. **15**(12): p. 6729-6735.
104. Wang, D., D.A. Kreutzer, and J.M. Essigmann, *Mutagenicity and repair of oxidative DNA damage: insights from studies using defined lesions.* Mutation Research-Fundamental and Molecular Mechanisms of Mutagenesis, 1998. **400**(1-2): p. 99-115.
105. Viswanathan, A., H.J. You, and P.W. Doetsch, *Phenotypic change caused by transcriptional bypass of uracil in nondividing cells.* Science, 1999. **284**(5411): p. 159-62.
106. Bregeon, D., et al., *Transcriptional mutagenesis induced by uracil and 8-oxoguanine in Escherichia coli.* Molecular Cell, 2003. **12**(4): p. 959-970.
107. Pastoriza-Gallego, M., J. Armier, and A. Sarasin, *Transcription through 8-oxoguanine in DNA repair-proficient and Csb(-)/Ogg1(-) DNA repair-deficient mouse embryonic fibroblasts is dependent upon promoter strength and sequence context.* Mutagenesis, 2007. **22**(5): p. 343-351.
108. Saxowsky, T.T., et al., *8-Oxoguanine-mediated transcriptional mutagenesis causes Ras activation in mammalian cells.* Proc Natl Acad Sci U S A, 2008. **105**(48): p. 18877-82.
109. Burns, J.A., et al., *O6-methylguanine induces altered proteins at the level of transcription in human cells.* Nucleic Acids Res, 2010. **38**(22): p. 8178-87.
110. Kashkina, E., et al., *Template misalignment in multisubunit RNA polymerases and transcription fidelity.* Mol Cell, 2006. **24**(2): p. 257-66.
111. Wagner, L.A., et al., *Transcriptional Slippage Occurs during Elongation at Runs of Adenine or Thymine in Escherichia-Coli.* Nucleic Acids Research, 1990. **18**(12): p. 3529-3535.

112. Benson, K.F., et al., *Paradoxical homozygous expression from heterozygotes and heterozygous expression from homozygotes as a consequence of transcriptional infidelity through a polyadenine tract in the AP3B1 gene responsible for canine cyclic neutropenia.* Nucleic Acids Res, 2004. **32**(21): p. 6327-33.
113. Linton, M.F., V. Pierotti, and S.G. Young, *Reading-frame restoration with an apolipoprotein B gene frameshift mutation.* Proc Natl Acad Sci U S A, 1992. **89**(23): p. 11431-5.
114. Linton, M.F., et al., *Reading-frame restoration by transcriptional slippage at long stretches of adenine residues in mammalian cells.* Journal of Biological Chemistry, 1997. **272**(22): p. 14127-14132.
115. Young, M., et al., *Partial correction of a severe molecular defect in hemophilia A, because of errors during expression of the factor VIII gene.* American Journal of Human Genetics, 1997. **60**(3): p. 565-573.
116. D A Evans, A.A.v.d.K., M A Sonnemans, J P Burbach, F W van Leeuwen, *Frameshift mutations at two hotspots in vasopressin transcripts in post-mitotic neurons.* Proc Natl Acad Sci U S A, 1994. **91**(13): p. 6059-63.
117. van Leeuwen, F.W., *Frameshift Mutants of Amyloid Precursor Protein and Ubiquitin-B in Alzheimer's and Down Patients.* Science, 1998. **279**(5348): p. 242-247.
118. Guest, W.C., et al., *Generalization of the prion hypothesis to other neurodegenerative diseases: an imperfect fit.* J Toxicol Environ Health A, 2011. **74**(22-24): p. 1433-59.
119. Polymenidou, M. and D.W. Cleveland, *The seeds of neurodegeneration: prion-like spreading in ALS.* Cell, 2011. **147**(3): p. 498-508.
120. Trojanowski, J.Q. and V.M. Lee, *"Fatal attractions" of proteins. A comprehensive hypothetical mechanism underlying Alzheimer's disease and other neurodegenerative disorders.* Ann N Y Acad Sci, 2000. **924**: p. 62-7.
121. Orgel, L.E., *The maintenance of the accuracy of protein synthesis and its relevance to ageing.* Proc Natl Acad Sci U S A, 1963. **49**: p. 517-21.
122. Orgel, L.E., *The maintenance of the accuracy of protein synthesis and its relevance to ageing: a correction.* Proc Natl Acad Sci U S A, 1970. **67**(3): p. 1476.
123. Martin, G.M. and S.L. Bressler, *Transcriptional infidelity in aging cells and its relevance for the Orgel hypothesis.* Neurobiol Aging, 2000. **21**(6): p. 897-900; discussion 903-4.
124. Jabara, C.B., et al., *Accurate sampling and deep sequencing of the HIV-1 protease gene using a Primer ID.* Proceedings of the National Academy of Sciences of the United States of America, 2011. **108**(50): p. 20166-20171.
125. Gout, J.F., et al., *Large-scale detection of in vivo transcription errors.* Proc Natl Acad Sci U S A, 2013. **110**(46): p. 18584-9.
126. Imashimizu, M., et al., *Direct assessment of transcription fidelity by high-resolution RNA sequencing.* Nucleic Acids Res, 2013. **41**(19): p. 9090-104.
127. Acevedo, A., L. Brodsky, and R. Andino, *Mutational and fitness landscapes of an RNA virus revealed through population sequencing.* Nature, 2014. **505**(7485): p. 686-90.
128. Acevedo, A. and R. Andino, *Library preparation for highly accurate population sequencing of RNA viruses.* Nat Protoc, 2014. **9**(7): p. 1760-9.
129. Valavanidis, A., T. Vlachogianni, and C. Fiotakis, *8-hydroxy-2' -deoxyguanosine (8-OHdG): A critical biomarker of oxidative stress and carcinogenesis.* J Environ Sci Health C Environ Carcinog Ecotoxicol Rev, 2009. **27**(2): p. 120-39.
130. Fortini, P., et al., *8-Oxoguanine DNA damage: at the crossroad of alternative repair pathways.* Mutat Res, 2003. **531**(1-2): p. 127-39.
131. Bohr, V.A., *Repair of oxidative DNA damage in nuclear and mitochondrial DNA, and some changes with aging in mammalian cells.* Free Radical Biology and Medicine, 2002. **32**(9): p. 804-812.

132. Bohr, V.A., O.P. Ottersen, and T. Tonjum, *Genome instability and DNA repair in brain, ageing and neurological disease.* Neuroscience, 2007. **145**(4): p. 1183-6.
133. Gredilla, R., et al., *Differential age-related changes in mitochondrial DNA repair activities in mouse brain regions.* Neurobiol Aging, 2010. **31**(6): p. 993-1002.
134. Gabbita, S.P., M.A. Lovell, and W.R. Markesbery, *Increased nuclear DNA oxidation in the brain in Alzheimer's disease.* J Neurochem, 1998. **71**(5): p. 2034-40.
135. Markesbery, W.R. and M.A. Lovell, *DNA oxidation in Alzheimer's disease.* Antioxid Redox Signal, 2006. **8**(11-12): p. 2039-45.
136. Wang, J., et al., *Increased oxidative damage in nuclear and mitochondrial DNA in Alzheimer's disease.* J Neurochem, 2005. **93**(4): p. 953-62.
137. Korencic, D., D. Soll, and A. Ambrogelly, *A one-step method for in vitro production of tRNA transcripts.* Nucleic Acids Res, 2002. **30**(20): p. e105.
138. McBride, T.J., B.D. Preston, and L.A. Loeb, *Mutagenic spectrum resulting from DNA damage by oxygen radicals.* Biochemistry, 1991. **30**(1): p. 207-13.
139. Klassen, R., et al., *A modified DNA isolation protocol for obtaining pure RT-PCR grade RNA.* Biotechnol Lett, 2008. **30**(6): p. 1041-4.