STOCHASTIC MODELING OF MITOCHONDRIAL POLYMERASE GAMMA REPLICATION AND NOVEL ALGORITHMS TO ENRICH RARE DISEASE ALLELES AND DETECT TUMOR SOMATIC MUTATIONS IN DEEP SEQUENCING DATA

By

ZHUO SONG

Dissertation

Submitted to Faculty of the

Graduate School of Vanderbilt University

in partial fulfillment of the requirements

for the degree of

DOCTOR OF PHILOSOPHY

in

Human Genetics

May, 2012

Nashville, Tennessee

Approved:

Professor David C. Samuels

Professor Chun Li

Professor Todd I. Edwards

Professor Ellen H. Fanning

Professor William S. Bush

Professor C. William Wester

To my beloved family infinitely supportive

ACKNOWLEDGMENTS

The works I present here were greatly improved by all the coauthors on the manuscripts from which this dissertation was adapted. These works were also guided and deeply enriched by input from my thesis committee (Dr. Chun Li, Dr. Todd I. Edwards, Dr. William S. Bush, C. William West, M.D., M.P.H., and Dr. Ellen Fanning).

I would like to especially thank my Ph.D. mentor, Dr. David C. Samuels for his unwavering support in my scientific training. Dr. Samuels has been an outstanding personal role model and has provided an exceptional training climate through his scientific advice and personal motivation. I cannot be thankful enough for the opportunities Dr. Samuels has given me to gain outside insight and share my work with other scientists both stateside and abroad.

I am grateful to all of those with whom I have had the pleasure to work during these projects. I would especially like to thank Dr. Chun Li, the chairman of my committee, who guided me into the genome sequencing area. He has shown me, by his example, what a good scientist should be.

These works would not have been possible without the financial support of National Institutes of Health through grants GM073744 (ZS, DCS) and R01HG004517 (ZS, CL).

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

INTRODUCTION

Part I Computational Biology Study of Mitochondrial DNA Replication

Mitochondria are essential for energy production through oxidative phosphorylation and have their own genome, which is maternally inherited and harbors 37 genes in a circular molecule of approximately 16.6 kb that is present in hundreds to thousands of copies per cell. The proteins involved in mtDNA replication, which are all encoded by the nuclear DNA, include the DNA polymerase gamma (pol γ) and its accessory protein POLG2, the single stranded DNA binding protein (mtSSB), the mtDNA helicase (Twinkle), and a number of accessory proteins and transcription factors (Graziewicz *et al.* 2004). The minimal protein set required to replicate mtDNA in vitro includes the two subunits of pol γ, the mitochondrial helicase and mtSSB (Korhonen *et al.* 2004). However, typically in experiments measuring the kinetics of pol γ, neither Twinkle nor mtSSB are added in the assay (Johnson and Johnson 2001; Johnson and Johnson 2001). Generally, these proteins are unnecessary since the kinetics assays only involve the replication of a short length of DNA.

The holoenzyme of human pol γ consists of a catalytic subunit (encoded by *POLG* at chromosomal locus 15q25) and a homodimer of its accessory subunit (encoded by *POLG2* at chromosomal locus 17q24.1) (Yakubovskaya *et al.* 2006). The catalytic subunit is a 140 kDa enzyme (p140) that has DNA polymerase, 3'-5' exonuclease and 5' dRP lyase activities. The accessory subunit is a 55 kDa protein (p55) required for tight DNA binding (Lim *et al.* 1999).

The function of pol γ is far more complicated than the function of a typical enzyme that converts a substrate into a product. For this reason, the analysis of the enzyme kinetics of pol γ must also be far more complicated. The function of pol γ is the replication of a complete

mitochondrial DNA molecule, while the enzyme kinetics data for pol γ are at the level of individual reactions at the base-pair level. One tool that can be used to bridge that large gap between the enzyme kinetics data and the final product of a replicated DNA molecule is simulation. A computational simulation can reproduce each individual reaction of the polymerase activity, at the lowest level for which we have experimental data on the relevant reaction kinetics of the polymerase. The simulation can tirelessly repeat this process for the several tens of thousands of reactions necessary for the replication of a single strand of human mtDNA. Since mutations introduced through polymerase errors are of great practical interest, we need to use a stochastic simulation approach that will allow such replication errors to occur with probabilities that are calculated from the experimentally measured enzyme kinetics.

Create mtDNA Replication Model of Pol γ Using The Gillespie Algorithm

The Gillespie algorithm (or Stochastic Simulation Algorithm) is a well-known Monte Carlo simulation method for chemical reactions (Gillespie 1976). This algorithm takes a defined list of possible reactions and uses the reaction rates, based on the measured enzyme kinetics data and the substrate concentrations, to calculate the probability of each reaction on the list. These probabilities are then used to randomly choose the sequence of reactions that occur. The original Gillespie algorithm is particularly useful for simulating reactions involving a small number of molecules, while more elaborate and approximate versions of the basic algorithm have been created in order to handle large numbers of molecules (Cao and Samuels 2009). For modeling pol γ activity we are generally concerned with only a single enzyme molecule, the one molecule of pol γ that is replicating a particular strand of mtDNA. In this special case, the original Gillespie algorithm is the preferred simulation method.

mtDNA deletion/depletion

The replication of DNA by a polymerase may not progress at a steady rate. How do replication pauses of pol γ relate to the commonly observed pathological phenotypes of mtDNA deletions and depletion? In theory, once a replication fork pauses, it can restart to continue the replication process without problems. However, if a pause is lengthy enough it might allow time for low-probability events such as double strand breaks. A double strand break can be repaired through blunt-ended rejoining, or homologous annealing of 5'- and 3'-repeat sequences (Krishnan *et al.* 2008), or nonhomologous end-joining (Graziewicz *et al.* 2006) that will form deleted mtDNA. If the double strand break cannot be repaired, it will cause a failed replication. Repeated failed replications eventually would lead to depletion of the amount of mtDNA within the cell. These concepts are illustrated in Figure 1. Under this hypothetical mechanism, mtDNA deletions and depletion may both be possible outcomes from the same initial event, the pausing of the polymerase. Of course, other mechanisms for deletion formation are also possible (Krishnan *et al.* 2008).

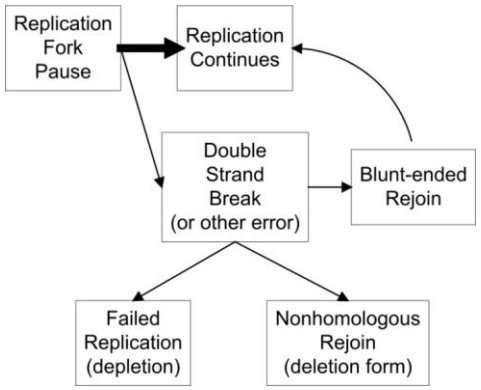

Figure 1. Possible consequences after mtDNA replication fork pauses.

Pathogenic Mutated Pol γ

Pathogenic mutations in *POLG* have been identified in patients with neurological or muscular diseases including progressive external ophthalmoplegia (PEO), Alpers syndrome, ataxia-neuropathy syndromes, idiopathic Parkinsonism, and nucleoside reverse-transcriptase inhibitor (NRTI) toxicity (Copeland 2008; Hudson and Chinnery 2006). All these diseases were characterized by mtDNA deletions and/or depletion in symptomatic tissues (Chan and Copeland 2009). PEO, as a mitochondrial disorder, is associated with the depletion of mtDNA and/or the accumulation of point mutations and deletions within mtDNA, caused by autosomal dominant mutations in *POLG* (Longley *et al.* 2005). In Alpers syndrome, the identified pol γ mutations are either homozygous A467T (MIM 174763.0002) or heterozygous A467T paired in *trans* with other mutations in *POLG* (Ferrari *et al.* 2005; Naviaux and Nguyen 2004).

The A467T mutation has been found in all of the major *POLG* related diseases, i.e. Alpers syndrome, ataxia-neuropathy syndromes and PEO (Chan and Copeland 2009). The frequency of the A467T mutation varies from less than 0.2% to approximately 1% in different European control populations (Horvath *et al.* 2006; Luoma *et al.* 2005; Van Goethem *et al.* 2001; Winterthun *et al.* 2005). In mitochondrial disease populations, however, the A467T allele was estimated to be the most common alleles associated with *POLG* diseases (Chan and Copeland 2009). Even the homozygous A467T substitution of *POLG* has been found in patients with highly varied phenotypes. For example, in Alpers syndrome (Boes *et al.* 2009; Utzig *et al.* 2007), two unrelated teenage boys with homozygous A467T *POLG* had ataxia and severe epilepsy. In contrast, in a case of concurrent progressive sensory ataxia, dysarthria, and ophthalmoparesis (McHugh *et al.* 2010), two siblings also homozygous for *POLG* A467T developed disease symptoms only late in life, with onset in their 40s. The kinetic parameters of the A467T mutant for both polymerase and exonuclease activity have been measured experimentally (Chan *et al.*

2005). Compared to the wild-type catalytic subunit, the A467T substitution increased the K_m of the enzyme 5-fold while also reducing the k_{pol} value approximately 5-fold for DNA synthesis. As determined by the ratio of k_{pol}/K_m, the A467T substitution reduces DNA synthesis efficiency to 4% of the wild-type activity. In contrast, the exonuclease activity of the A467T variant is only decreased 2-fold compared to the wild-type (Chan *et al.* 2005). Besides altering the kinetics of pol γ, the A467T variant in *POLG* also decreases the protein's physical association of the POLG2 subunit (Chan *et al.* 2005).

In addition to the identified pathogenic A467T substitution of pol γ in human, proof-reading deficient versions of the catalytic subunit of mtDNA polymerases have been created in two independent homozygous knock-in mouse models, one published in 2004 (Trifunovic *et al.* 2004) and the other one in 2005 (Kujoth *et al.* 2005). For simplicity, the first will be referred as Stockholm mouse model and the second will be referred as Madison mouse model. Both mouse models showed similar progeria-like phenotypes and shared the same D257A mutation on the second exonuclease domain of *PolgA* (the mouse homolog of *POLG*) that caused a profound reduction of the exonuclease activity but no decrease in DNA polymerase activity. Both mouse models (Kujoth *et al.* 2005; Trifunovic *et al.* 2004) were reported to have accumulated mtDNA point mutations. Using these two mouse models, other mitochondrial genome variations besides point mutation have been studied. For the Stockholm mouse model, after Trifunovic *et al.* (Trifunovic *et al.* 2004) reported increased mtDNA deletion, Bailey *et al.* (Bailey *et al.* 2009) using one- and two-dimensional agarose gel electrophoresis interpreted these non-replicating linear mtDNAs as the result of double-strand breaks of cyclic mtDNA caused by stalled replication intermediates. Ameur *et al.* (Ameur A. *et al.* 2011) further stated that mutator mice have abundant linear deleted mtDNA molecules but extremely low levels of circular mtDNA molecules with large deletions. For the Madison mouse model, Williams *et al.* (Williams *et al.*

2010) applying next-generation sequencing to native mtDNA did not observe the accumulation of mtDNA deletions but instead reported multiple copies of the mtDNA control region in brain or heart. However, Vermulst *et al.* (Vermulst *et al.* 2008; Vermulst *et al.* 2009) have identified mtDNA deletions as a critical driving factor behind the premature aging phenotype in Madison mouse models and have suggested there is a homology-directed DNA repair mechanism directly linked to mtDNA deletion formation. As this description shows, the data on the amount of deletions in these two mouse models is currently mixed and this is a research area that is rapidly developing.

Pol γ Dependent Toxicity of NRTI Drugs in HIV Treatment

In their activated tri-phosphorylated forms, each NRTI (nucleoside reverse transcriptase inhibitor) drug acts as a nucleotide analog interacting with the HIV viral reverse transcriptase (Anderson *et al.* 2004; Ray 2005). NRTIs are an alternative substrate to the natural nucleotides, which lacks the 3' OH group necessary for incorporation of the next nucleotide thereby terminating viral DNA strand elongation. Although NRTIs are effective drugs and have helped usher HIV into the category of a controllable chronic disease, they are also often toxic, inducing side effects such as lactic acidosis, peripheral neuropathy, lipoatrophy, and pancreatitis in patients. Intolerance of such side effects is a common reason for treatment discontinuation (d'Arminio Monforte *et al.* 2000). Any decrease in patient compliance to the treatment regimen is a serious concern that can lead to poor HIV virologic control, the increased potential for the development of HIV-1 genotypic drug resistance, and ultimately treatment failure. The first step in ameliorating these side effects and preventing them in future antiviral treatments is to understand the mechanisms behind the mitochondrial toxicity of the NRTIs that are in use today. As we discuss below, many mechanisms of the mitochondrial toxicity have been proposed. In

this paper we specifically consider the plausibility of the most widely accepted hypothesis for the toxicity mechanism for this class of drugs; interference of mitochondrial DNA replication by the activated drug.

In a study conducted by Martin et al. (Martin *et al.* 1994) the approved NRTIs were shown to inhibit various host DNA polymerases. After the HIV Reverse Transcriptase, the highest affinity of the NRTIs was for pol γ. This, along with the fact that many of the NRTI side-effects resemble symptoms of mitochondrial genetic disorders, implicated interaction with pol γ and subsequent depletion of mtDNA as a potential cause of NRTI toxicity giving rise to the pol γ hypothesis (Lewis *et al.* 2003). Indeed, experiments have demonstrated decreased mtDNA amounts in various tissue types of NRTI-treated HIV positive patients (Cherry *et al.* 2002; Cote *et al.* 2002; Haugaard *et al.* 2005; Walker *et al.* 2004). In addition, mtDNA depletion was observed in parallel with cell death, mitochondrial morphological changes, and increased lactate production in liver, heart, neuron, skeletal muscle, adipose, and blood cell cultures after incubation with different NRTIs (Azzam *et al.* 2006; Benbrik *et al.* 1997; Biesecker *et al.* 2003; Birkus *et al.* 2002; Cui *et al.* 1997; Galluzzi *et al.* 2005; Pan-Zhou *et al.* 2000; Setzer *et al.* 2005; Walker *et al.* 2002).

There are three possible pol γ dependent toxicity mechanisms that comprise the pol γ hypothesis: 1) direct inhibition of pol γ by NRTI-triphosphate without incorporation into the mtDNA, 2) chain termination of mtDNA replication following incorporation of the NRTI triphosphate, and 3) incorporation of the analog triphosphate into mtDNA without chain-termination allowing the NRTI to continue as a point mutation in mtDNA (Lewis *et al.* 2006).

However, there also exists a substantial body of data that are not consistent with toxicity mechanisms resulting in depletion of mtDNA and indicate a weak relationship between mtDNA copy number and nucleoside analog toxicity. Martin et al. (Martin *et al.* 1994) showed

no association between inhibition of pol γ by NRTIs and mtDNA depletion. Mitochondrial dysfunction has been observed *in vitro* in mouse muscle, white adipose, brain, liver, and heart tissue (Note *et al.* 2003), hepatoma cell lines (Walker *et al.* 2002) as well as CD4 cells (Setzer *et al.* 2005) after incubation with NRTIs although no significant decrease in mtDNA amount was observed. In clinical settings mtDNA depletion has been seen in parallel with normal cytochrome c oxidase activity, a sign of correct mitochondrial function (Piechota *et al.* 2006), and was not associated with lipoatrophy (McComsey *et al.* 2005) (although that study measured mtDNA depletion in blood samples, not fat cells). This warrants a deeper look at the data concerning the interaction of different NRTIs with pol γ.

To this end, we have simulated the DNA replication process of mitochondria. Using enzyme kinetics data gathered from Johnson et al. (Johnson *et al.* 2001), Feng et al. (Feng *et al.* 2004), and Hanes et al. (Hanes and Johnson 2007; Hanes *et al.* 2007) we have carried out a series of simulations of mtDNA replication in the presence of various nucleoside analogs that interact with pol γ (Table 3.2). These simulations bridge the gap between the basic enzyme kinetics data and the probability of failure of the mtDNA replication process.

Related Chapters

Chapter I describes the computational stochastic model of the replication of pol γ in detail, which is the 1st model on the single nucleotide incorporation level. This model could be a great tool for the analysis of enzyme kinetic data for mtDNA replication, which is capable of dealing with different substrates - natural nucleotide and nucleotide analogs and both wild-type and pathogenic mutated pol γ.

Chapter II studies replication pauses of pol γ. The stochastic model made it possible to quantify the pause lengths for the wild-type pol γ and for the pathogenic A467T and the

exonuclease deficient pol γ variants. Our model of the exo- polymerase had extremely long pauses, with a 30 to 300-fold increase in the time required for the longest single forward step compared to the wild-type, while the naturally occurring A467T variant showed at most a doubling in the length of the pauses compared to the wild-type. We identified the cause of these differences in the polymerase pausing time to be the number of disassociations occurring in each forward step of the polymerase.

Chapter III studies pol γ dependent side effect of NRTI drugs. The model has been used to analyze the inhibition of pol γ replication and toxicity caused by nucleoside analogs in antiretroviral treatment. The model predicts an approximate 1000 fold difference in the activated drug concentration required for a 50% probability of mtDNA strand termination between the activated di-deoxy analogs stavudine (d4T), zalcitabine (ddC), and didanosine (ddI) and the activated forms of the analogs lamuvidine (3TC), tenofovir (TDF), zidovudine (AZT/ZDV), emtricitabine (FTC), and abacavir (ABC). These predictions are supported by experimental and clinical data showing significantly greater mtDNA depletion in cell culture and patient samples caused by the di-deoxy analog drugs.

Part II Methodology for Sequencing Data Analysis

Enrich Rare Disease Alleles in GWAS Samples

In genetic epidemiology, a genome-wide association study (GWAS) is based on the premise that densely genotyped common alleles will have statistical power to detect causal associations with traits at nearby, ungenotyped common variants, through short-range linkage disequilibrium (LD). The foundation for this strategy is the common disease common variant (CDCV) hypothesis (Reich and Lander 2001). This approach has been proven to be very effective in different scenarios for mapping small genomic regions to traits (see the National Human Genome Research Institute Catalog of Published Genome-Wide Association Studies) (Manolio *et al.* 2008; McCarthy *et al.* 2008). However, the predominantly small effect sizes encountered thus far in investigations of most traits do not provide an explanation for a large proportion of the trait variance attributable to heritable factors (Maher 2008; Manolio *et al.* 2009). This phenomenon has been described by Purcell *et al.* who suggested that hundreds or thousands of SNPs may each have a very subtle influence on the risk of some psychiatric traits (Purcell *et al.* 2009). These observations seem to support an adjustment of the CDCV model to allow for the probability that rare alleles might also exert a major influence on common traits (Bodmer and Bonilla 2008; Pritchard 2001; Schork *et al.* 2009).

The common disease rare variant (CDRV), derived from CDCV, assumes that alleles with strong effects on traits are likely to be rare due to purifying selective pressure and recent time to coalescence due to the rapid expansion of human populations (Pritchard 2001). In addition, it has been shown in simulations that multiple rare alleles with strong effects can stochastically aggregate onto the haplotypic background of a common allele and produce genome-wide significant association signals, a scenario termed synthetic association (Dickson *et al.* 2010). The

CDRV hypothesis has further support coming from observational studies where the average allele frequency of SNPs with predicted effects on proteins was smaller than the average allele frequency of intronic or synonymous variants (Cargill *et al.* 1999; Gorlov *et al.* 2008; Wong *et al.* 2003). Estimates from human Mendelian traits, human–chimpanzee divergence data and human genetic variation suggested that ~53% of new missense mutations have mildly deleterious effects, and that up to 70% of low-frequency missense alleles are mildly deleterious (Kryukov *et al.* 2007).

A rare trait allele may not be annotated in the databases of common variants maintained by either dbSNP or the International HapMap Organization. As a consequence, the possibility of detecting that SNP will be excluded through imputation and subsequent association analysis. The constellation of causal alleles may also be unique for each population of human subjects, where sensitive functional gene or regulatory regions are perturbed by independent sets of rare mutations that occurred after geographic or cultural barriers led to increased genetic distance (Tishkoff *et al.* 2009). Thus, the same associated allele from GWAS across multiple ethnic groups does not necessarily imply the same underlying architecture of causal alleles in LD. Then resequencing becomes the best available means of discovering these rare SNPs in a GWAS sample and ultimately detecting the relationship between these alleles and traits.

To successfully discover the variants that determine trait susceptibility, it is necessary to directly capture all genetic mutations in a region (Cirulli and Goldstein 2010). This can be accomplished most efficiently by using next-generation sequencing technology to resequence subjects for the implicated loci (Service 2006), which has substantially decreased the financial cost of resequencing large genomic regions relative to Sanger sequencing technology. However, next-generation sequencing technologies are still not generally feasible for resequencing all the

subjects that were used to isolate a genomic region via GWAS. Thus, some strategies are necessary for employing sequencing technology that is cost-effective. One possibility is to resequence a small number of cases and controls or persons with extreme trait values and evaluate the observed genetic variation for association with traits to screen rare variants prior to larger genotyping experiments. This approach, however, will suffer from low statistical power at the screening step due to the infrequent exposure rate of rare alleles, and potentially suffer from inflated type I error rates (Li and Leal 2009) as a result of ascertainment bias in cases. Another alternative approach is not to attempt to associate alleles from resequencing data with the trait, but instead, to discover rare alleles by resequencing and then examine these SNPs with conventional genotyping methods in the entire available pool of study subjects. The limitation of this approach is the power to capture rare alleles in the targeted loci, which is directly related to the selection of subjects for the resequencing experiment (Li and Leal 2009). Using the information available at nearby trait-associated SNPs, targeted resequencing study designs can be tailored for efficient capture of rare disease alleles in small samples.

Detect Tumor Somatic Mutations in Sequencing Data

Cancer, as a class of complex genetic diseases, is responsible for one in eight deaths worldwide (Blecher *et al.* 2008). It is characterized by uncontrolled proliferation of malignant cells that can intrude into surrounding tissues and metastasize to distant organs because of somatically accumulated mutations and epigenetic changes (Esteller 2007; Stratton *et al.* 2009). Somatic genomic abnormalities of cancer encompass distinct classes of DNA and chromosomal changes, including base substitutions, insertions and deletions (indels), copy number variants (CNVs), and inter- and intra-chromosomal rearrangements. Exogenous DNA from viruses may also have been obtained in cells of some cancer types (Talbot and Crawford 2004).

Somatic mutations are distributed across all the genome (Stratton 2011). Only a small subset of them has been implicated in oncogenesis. The rate of somatic mutations may vary by cell type and can be influenced by mutagenesis exposures endogenously and exogenously. Generally, different cancer types have different mutation profiles. Some cancers display very disordered genomes while others have few genomic aberrations (Stratton 2011). It is crucial to understand the overall mutation profiles for different cancer types and in different patient individuals as mutation profiles may influence choice of treatments and a patient's reaction to a treatment. In early cytogenetic studies of cancer, most genomic abnormalities found were chromosomal translocations and CNVs. Thanks to the development of next-generation sequencing (NGS) technologies, we are on the edge of creating a complete catalogue of mutations for each cancer type (Singer 2011) by screening the cancer genome at the finest resolution.

Recently, the NGS technologies have been used to study somatic mutations in cancers, including melanoma (Pleasance *et al.* 2010), lung cancer (Lee *et al.* 2010; Pleasance *et al.* 2010), colorectal cancer (Timmermann *et al.* 2010), renal carcinoma (Varela *et al.* 2011), breast cancer (Ding *et al.* 2010; Shah *et al.* 2009), acute myeloid leukemia (Mardis *et al.* 2009), mesothelioma (Bueno *et al.* 2010), and prostate cancer (Berger *et al.* 2011).

In these studies, sequencing was typically carried out on high-throughput sequencers to generate billions of single-end or paired-end reads that are then aligned to the reference genome. Figure 2 shows short reads (paired-end) aligned against the reference genome. The number of reads that cover one site is defined as the depth or coverage of that site. The genotype of each site can be called using the information of the bases on the reads mapped that site.

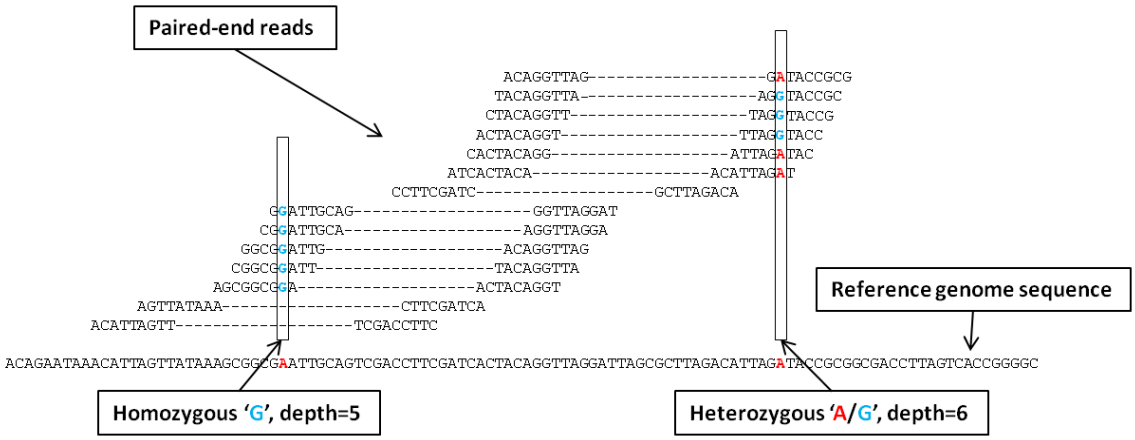

Figure 2. Paired-end reads aligned to the reference genome.

Although the methods for mutation detection differ in details, these studies all relied on comparing SNP calls across samples in their primary mutation screening. Several efficient analysis tools for NGS data have been developed. For example, the Genome Analysis Toolkit (GATK) (McKenna *et al.* 2010), is a suite of tools for sequence data processing and genotype calling, which provides a rich software library for developing additional analysis tools. The GATK Unified Genotyper and other SNP calling programs perform well on calling genotypes for homogeneous samples such as blood and normal tissues. However, in cancer genomes, sample cells taken from a tumor tissue may include non-cancer cells. Additionally, cancer cells in a tumor sample can be heterogeneous with respect to somatic mutations. Therefore, a somatic mutation may exist only in a fraction of cells in a tumor sample, decreasing the likelihood of detecting it through genotype calls. The resulting allelic imbalance may also lead to tumor genotype calls being flagged as problematic and excluded from analysis.

There are a few existing tools for detecting somatic mutations. However, they mostly rely on genotype calls for tumor samples. For instance, VarScan (Koboldt *et al.* 2009) screens data for mutation sites by comparing genotype calls between normal and tumor samples, and

SomaticSniper (Mardis *et al.* 2009) uses a genotype likelihood model (Li *et al.* 2009) to calculate the probability that the tumor and normal genotypes are different. VarScan (Koboldt *et al.* 2009) does use allele frequency information to evaluate significance, but only for sites that show genotype difference between normal and tumor samples.

Related Chapters

Chapter IV presents SampleSeq, an algorithm for enriching the yield of rare or uncommon disease alleles in a sample of unrelated study subjects by choosing subjects according to their observed associated alleles and trait information. When multiple regions are to be sequenced, SampleSeq selects subjects with a balanced representation of all the regions. SampleSeq can also estimate the sample size required to detect a hypothetical disease allele, and thus can optimize a resequencing study to preserve resources for subsequent genotyping or other investigations.

Chapter V introduces two novel algorithms for detecting two major types of somatic mutations, base substitution and loss of heterozygosity (LOH), by directly examining the sequence reads of tumor samples. We applied our methods, named Somatic Mutation Gleaner (SMUG), to whole exome sequencing data of eight breast cancer tumors and their matched blood samples, and detected somatic mutations that are missed by comparison of genotype calls between normal and tumor samples.

CHAPTER I

ANALYSIS OF ENZYME KINETIC DATA FOR mtDNA REPLICATION[1]

Introduction

The Function of Polymerase Gamma

Mitochondrial polymerase gamma is the sole DNA polymerase active in mitochondria (Kaguni 2004) and is responsible for the replication of mitochondrial DNA (mtDNA). Vertebrate polymerase gamma is composed of two subunits: a catalytic core, Pol γ-α, that contains the DNA polymerase and 3'-5' exonuclease activities, and an accessory subunit, Pol γ-β, which enhances catalytic activity and serves as a processivity factor during DNA synthesis (Fan *et al.* 2006).

The function of polymerase gamma is the replication of a complete mitochondrial DNA molecule, while the enzyme kinetics data for polymerase gamma are at the level of individual reactions at the base-pair level. To bridge the large gap between the enzyme kinetics data and the final product of a replicated DNA molecule, I build a computational simulation that can reproduce each individual reaction of the polymerase activity at the lowest level for which we have experimental data on the relevant reaction kinetics of the polymerase. The simulation can repeat this process for the several tens of thousands of reactions necessary for the replication of a single strand of human mtDNA. Since mutations introduced through polymerase errors are of great practical interest, we need to use a stochastic simulation approach that will allow such

[1] Song, Z., and Samuels, D. C. (2010). Analysis of enzyme kinetic data for mtDNA replication. *Methods* 51(4), 385-391. Author contributions: Conceived and designed the experiments: ZS DCS. Performed the experiments: ZS. Analyzed the data: ZS DCS. Wrote the paper: ZS DCS.

replication errors to occur with probabilities that are calculated from the experimentally measured enzyme kinetics.

The Gillespie Algorithm

The Gillespie algorithm is a well-known Monte Carlo simulation method for chemical reactions (Gillespie 1976). This algorithm takes a defined list of possible reactions and uses the reaction rates, based on the measured enzyme kinetics data and the substrate concentrations, to calculate the probability of each reaction on the list. These probabilities are then used to randomly choose the sequence of reactions which occur. The original Gillespie algorithm is particularly useful for simulating reactions involving a small number of molecules, while more elaborate and approximate versions of the basic algorithm have been created in order to handle large numbers of molecules (Cao and Samuels 2009).

Traditional continuous and deterministic biochemical rate equations are modeled as a set of coupled ordinary differential equations but these methods rely on bulk reactions that require the interactions of millions of molecules. In contrast, the Gillespie algorithm allows a discrete and stochastic simulation of a system with few reactants because every individual reaction is explicitly simulated. The physical basis of the Gillespie algorithm is the collision of molecules within a reaction vessel where the reaction environment is assumed to be well mixed. The general Gillespie algorithm can be summarized by the following series of steps:

Step 1, Initialization: Initialize the number of molecules in the system, the reaction kinetics constants, and the random number generators.

Step 2, Calculate Reaction Probabilities From Reaction Rates: Based on the substrate concentrations and the enzyme kinetics data, reaction rates R_i are calculated for each possible reaction, where *i* is an index denoting the particular reaction from the reaction list. The

probability P_i for each reaction is then calculated from the list of n reaction rates through the following equation.

$$P_i = R_i / R_{total} \tag{1}$$

The parameter R_{total} is the sum of all of the reaction rates, as follows.

$$R_{total} = \sum_{j=1}^{n} R_j \tag{2}$$

Note that the sum of the probabilities is 1, with this definition.

Step 3, Choose A Reaction: Use a pseudo-random number generator to generate a uniform random number $r_{reaction}$ in the range [0,1] and use this random number to choose the one reaction that occurs from the list of n possible reactions. The reaction number j is chosen if it satisfies the following condition.

$$\sum_{i=1}^{j-1} P_i < r_{reaction} \leq \sum_{i=1}^{j} P_i \tag{3}$$

Then a second uniform random number r_{time}, also in the range [0,1] is chosen. This random number is used to set the time τ required for this reaction by the following equation.

$$\tau = -\left(1/R_{total}\right)\ln\left(r_{time}\right) \tag{4}$$

It is important to note here that the sum of all of the reaction rates is used in defining the time τ, not just the rate of the one particular reaction that was chosen.

Step 4, Update the Variables: Increase the time by τ. Update the molecule count based on the reaction that occurred.

Step 5, Iterate the Process: Check to see if any stopping condition has been met. If not, return to step 2 to recalculate the reaction rates, which may have changed due to the reaction chosen and carried out in steps 3-4.

These five steps define the basic Gillespie method for stochastic simulation. For applications of the stochastic simulation that involve large numbers of interacting molecules this basic method is inefficient in many ways, though it is an exact solution of the relevant stochastic master equation. In those cases more advanced and efficient methods should be used (Cao and Samuels 2009). However, for modeling polymerase gamma activity we are generally concerned with only a single enzyme molecule, the one molecule of polymerase gamma that is replicating a particular strand of mtDNA. In this special case, the simple Gillespie algorithm outlined above is the preferred simulation method.

Applying the Gillespie Algorithm to Polymerase Gamma

The Gillespie algorithm is fundamentally a way of analyzing the outcome of a set of defined reactions which are competing with each other. The heart of this method is the definition of the reaction list. As we stated above, the function of polymerase gamma is complicated, and the reaction list could conceivably be defined in many different ways. However, our purpose is the analysis of the measured enzyme kinetics data, and for that purpose the available kinetics data are the primary determinant of the level of the reaction that can be modeled. There is no doubt that many of these kinetics values actually represent the cumulative effect of a number of subreactions. But unless there are enzyme kinetics data available for these subreactions, it is futile to attempt to model the subreactions. The available experimental kinetics data determine the lowest level of reactions which can be modeled reasonably.

Some DNA polymerases, such as the bacteriophage T7 DNA polymerase, have had their kinetics examined in great detail by transient-state analysis of single nucleotide incorporation

(Patel *et al.* 1991). However, the kinetics of the mitochondrial DNA polymerase gamma have not been measured in that high level of detail. A series of papers from the Johnson group (Johnson and Johnson 2001; Johnson and Johnson 2001; Johnson *et al.* 2001; Johnson *et al.* 2000) have reported kinetic parameters for polymerase gamma at the level of polymerization, exonuclease and the disassociation of the polymerase gamma holoenzyme from the mtDNA molecule. Based on these experiments, we have developed a stochastic model to simulate the replication of an individual strand of mtDNA based on the Gillespie algorithm (Wendelsdorf *et al.* 2009). This model operates at the level of individual nucleotide insertions in the new DNA strand, using a reference mtDNA sequence (generally the human reference sequence) for the template strand.

Of course, polymerase gamma does not operate in a vacuum. It uses deoxyribonucleoside triphosphates (dNTPs) as substrates and these chemicals are created through some combination of a salvage pathway within mitochondria and the transport of substrates or their precursors into mitochondria from the cytoplasm (Bradshaw and Samuels 2005; Saada 2004). One must choose an artificial boundary for every computational model and events outside of that boundary are ignored in the model. In the following description we will assume that only the actions of polymerase gamma are modeled and that the concentrations of the dNTPs within the mitochondria are held constant. Alternatively, one could reasonably extend the model to include the reactions of the salvage pathway. A practical limitation on this extension of the model is our current lack of knowledge of the correct transport kinetics of deoxyribonucleotides between the cytoplasm and the mitochondrial matrix (Kang and Samuels 2008). Another alternative would be to extend the model to include the function of other proteins critical to mtDNA replication, such as the helicase (Kaguni 2004). All such extensions of the model are limited by the availability of kinetics data for the relevant enzymes.

Definition of the Reaction List for Polymerase Gamma

In this model, polymerase gamma carries out four basic classes of reactions: DNA polymerase activity (inserting a correct or incorrect nucleotide into the new DNA strand), exonuclease activity (removing a nucleotide from the new DNA strand), disassociation of the polymerase from the DNA, and reassociation of the polymerase with the DNA molecule. In the **DNA polymerase reaction** a single nucleotide is added to the new DNA strand. This nucleotide may be the correct base indicated by the template strand, or it may be an incorrect one which does not complement the template. In the **exonuclease reaction** one nucleotide is removed from the new DNA strand. Again the removed nucleotide may be either a correct match to the template strand or an incorrect match. The exonuclease reaction is an error correction mechanism, as the rate for removal of incorrectly incorporated nucleotides is faster than that for correctly incorporated nucleotides. In the **disassociation reaction** the polymerase holoenzyme separates from the DNA molecule. The **reassociation reaction** re-attaches the polymerase to the DNA molecule after a disassociation event occurs.

In our basic model of polymerase gamma function at each position on the replicating mtDNA strand polymerase gamma will randomly undergo one of a set of six reactions; one correct polymerase reaction, three incorrect polymerase reactions, the exonuclease reaction, or disassociation (Figure I.1). Which reaction polymerase gamma undergoes at each time is determined by the probability of each reaction, calculated using the reactions rates as introduced in the Gillespie algorithm (Gillespie 1976) and described above. The probability for each possible reaction is proportional to the reaction rate, so that fast reactions occur more often than slow reactions. Currently, kinetics data for the reassociation reaction rate for polymerase gamma are not available to the best of our knowledge. For this model we assume that once disassociation occurs the only possible reaction is then reassociation of the

polymerase with the DNA template. Considering the lack of kinetics for the reassociation reaction, we make the choice to immediately reassociate the polymerase to the DNA after a disassociation event. This choice neglects the time that this reassociation event requires, but without kinetics data for this reaction little else can be done.

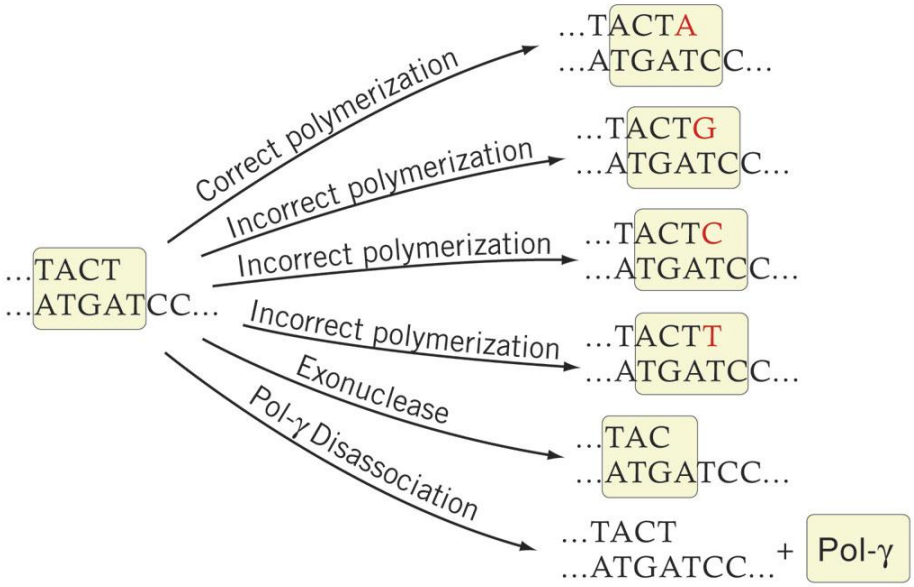

Figure I.1. **Diagram of the six basic competing reaction of the mitochondrial DNA polymerase**.

Michaelis-Menten kinetics were used for all of the DNA polymerization reactions. The exonuclease reaction and the disassociation reaction were set to have constant reaction rates based on the experimental data. For the two scenarios of a correctly inserted and incorrectly inserted previous nucleotide we have separate sets of kinetic parameters from Lee and Johnson (Lee and Johnson 2006) and Johnson and Johnson (Johnson and Johnson 2001; Johnson and Johnson 2001). These studies have reported an increase in exonuclease and disassociation rates, but a decrease in incorporation rates by polymerase gamma following an incorrect

incorporation. This is included in the simulation model by using two sets of enzyme kinetics parameters, one set for reactions following a correct incorporation and another set for reactions following an incorrect incorporation. The full set of enzyme kinetics data used in the model is listed in Tables I.1-I.5, which are reproduced from reference (Wendelsdorf *et al.* 2009). Details of how the parameter values were set or estimated based on experiments are given in the next section "Enzyme Kinetics Data" on pages 26-32.

Table I.1. **Kinetic parameters k_{cat} (s^{-1}) for base pairings when the previously inserted nucleotide pair is a correct Watson-Crick pair** ordered by template strand (rows) and the new strand (columns).

Base pairings	T	G	C	A
A	25	0.08	0.1	0.0036
C	0.012	37	0.003	0.1
G	0.16	0.066	43	0.042
T	0.013	1.16	0.038	45

Table I.2. **Kinetic parameters K_m (μM) for base pairings when the previously inserted nucleotide is a correct Watson-Crick pair**. The array order is the same as in Table I.1.

Base pairings	T	G	C	A
A	0.6	800	540	25
C	180	0.8	140	160
G	200	150	0.9	250
T	57	70	360	0.8

Table I.3. **Estimated k_{cat} (s^{-1}) kinetic parameters for base pairings when the previously inserted nucleotide forms a non-Watson-Crick pair**. The array order is the same as in Table I.1.

Base pairings	T	G	C	A
A	0.52	0.00052	0.00052	0.00052
C	0.00052	0.52	0.00052	0.00052
G	0.00052	0.00052	0.52	0.00052
T	0.00052	0.00052	0.00052	0.52

Table I.4. **Estimated K_m (μM) kinetic parameters for base pairings when the previously inserted nucleotide forms a non-Watson-Crick pair**. The array order is the same as in Table I.1.

Base pairings	T	G	C	A
A	404	40400	40400	40400
C	40400	404	40400	40400
G	40400	40400	404	40400
T	40400	40400	40400	404

Table I.5. **Reaction rates for the exonuclease and disassociation reactions of polymerase gamma when the previous base pair is a Watson-Crick pair or a non-Watson-Crick pair**.

Previous base pairs	**Exonuclease rate (s^{-1})**	**Disassociation rate (s^{-1})**
Watson-Crick pair	0.05	0.02
Non-Watson-Crick pair	0.4	0.2

Modified Nucleotide Reactions

The reaction list illustrated in Figure I.1 may be considered as a *minimum* list of reactions needed to model the activity of polymerase gamma. This list can be extended in a number of different ways. In addition to the four natural deoxyribonucleotide tri-phosphate substrates, several chemically altered forms of these substrates can also interact with polymerase gamma, albeit with altered, and generally poorer, enzyme kinetics. One important class of these chemically altered substrates is the nucleoside analogs commonly used in the treatment of HIV infection. Many of these medications have mitochondrial toxicity (Anderson *et al.* 2004; Lewis *et al.* 2003) which may be due to their interaction with polymerase gamma (Hanes *et al.* 2007; Johnson *et al.* 2001; Kaguni 2004; Martin *et al.* 1994; Wendelsdorf *et al.* 2009)

and their subsequent incorporation into mtDNA, generally causing the disruption of the mtDNA replication event. Due to this toxicity, which is very important in the long-term treatment of HIV infection, the kinetics data for the interaction of many nucleotide analog drugs have been measured (Feng *et al.* 2004; Hanes and Johnson 2007; Hanes *et al.* 2007; Johnson *et al.* 2001) (see Table I.6, reproduced in part from reference (Wendelsdorf *et al.* 2009). Details on how the parameter values were set or estimated based on experiments are given in the next section "Enzyme Kinetics Data" on pages 32-33).

Table I.6. **Enzyme kinetics parameter values used for polymerase gamma reactions with clinically relevant nucleoside analog drugs**.

NRTI	K_m (μM)	K_{cat} (s^{-1})	Exonuclease rate (s^{-1})	Disassociation rate (s^{-1})
ddC	0.041	0.660	<0.00002	0.02
ddA	0.022	0.310	0.0005	0.02
FIAU	2.9	24	0.06	0.02
d4T	0.045	0.24	0.0004	0.02
FTC(+)	0.79	0.84	0.0048	0.02
3TC(-)	9.2	0.125	0.015	0.02
3TC(+)	1.5	0.35	0.02	0.02
Acyclovir	6	1.03	0.0021	0.02
ddI	6.3	0.15	0.0007	0.02
TDF	40.3	0.21	0.0007	0.02
ABC	13	0.0018	0.0016	0.02
AZT_{2001}	187	0.2	0.001	0.02
AZT_{2007}	280	0.001	Not reported	0.02
FTC(-)	62.9	0.0086	0.0048	0.02

Enzyme kinetics data taken from references (Feng *et al.* 2004; Hanes and Johnson 2007; Hanes *et al.* 2007; Johnson *et al.* 2001).

Based on these data, we have included these alternate substrates in a model of polymerase gamma function (Wendelsdorf *et al.* 2009). Other naturally occurring alternative substrates could also be modeled. The most important of these are arguably ribonucleotide tri-

phosphates (Yasukawa *et al.* 2006) or oxidatively damaged substrates such as 8-oxo-dGTP (Pursell *et al.* 2008). The only practical limitation to the additional reactions that could be modeled is the availability of the reaction kinetics data for the new reaction.

Enzyme Kinetics Data

Michaelis-Menten kinetics for incorporation rates after a matched or an unmatched base pair

The reaction kinetics for polymerization reactions are given in Tables I.1-I.4. The reaction rates for the polymerization are calculated using a standard Michaelis-Menten equation, using assumed values for the four dNTP substrate concentrations. The dNTP concentrations are the major parameters for the simulation, and the results of the simulation depend greatly on the pattern of these four concentrations.

Given the complexity of polymerase gamma activity, it should not be surprising that even with the large set of kinetic parameters that have been measured our knowledge of the relevant enzyme kinetics parameters for this enzyme is still incomplete. For example, the data used in Tables I.1 and I.2 was measured only in the case where the previous base pair was an A:T pair. In principle the reaction rates could be different for other preceding bases, however in the absence of this data we have applied these reaction kinetics to all cases where the previous base pair is a standard Watson-Crick pairing. The data on kinetics following a non-Watson-Crick base pair are even more limited. The k_{cat} and K_m values reported by reference (Johnson and Johnson 2001) for this case were only determined for the correct pairing of a C opposite to a G in the template strand, where the previous base pairing was an incorrect T:T pairing. The kinetics values for other pairings, such as an A opposite to a T, are not available, and the

kinetics for any incorporations following incorrect pairings other than T:T are also not available. The last point is of perhaps great practical importance since the incorrect T:T pairing represents an A to T (or T to A) transversion, which is actually a rare event in mitochondrial DNA. It would be far more valuable to have kinetics after a G:T incorrect pairing, as this represents the very common A to G transition mutation. One important application of this analysis of the kinetics data for polymerase gamma is to discover such limitations in the experimental data.

Since there are holes in the available experimental data for polymerase gamma, we must make some practical assumptions to fill these holes. We defined an approximate model by setting the kinetic coefficients for all correct incorporations following an incorrect incorporation to be equal (Table I.3 and I.4) to the values reported for the C:G pairing. Reaction kinetics for the incorporation of an incorrect nucleotide following an incorrect nucleotide are not available at all, to the best of our knowledge. To fill this gap we chose to assume values for these missing parameters (Table I.3, off-diagonal elements) based on the observation from Table I.1 that k_{cat} values were approximately 1000 times less for non-Watson-Crick pairing compared to regular Watson-Crick pairings. Similarly, based on Table I.2 K_m values were estimated to be approximately 100 times greater for the non-Watson-Crick pairings and this was used to estimate K_m values in Table I.4 for the off-diagonal elements.

Reaction rates for the exonuclease and disassociation reactions

Reaction rates for the exonuclease and disassociation reactions of polymerase gamma are listed in Table I.5. As with the polymerase reaction, the exonuclease and disassociation reaction rates were measured for the two DNA template conditions where the previous base pair was a Watson-Crick pair or a non-Watson-Crick pair.

Reassociation of the polymerase with the DNA

Data regarding the reassociation reaction rate is not available to the best of our knowledge. For this model we assume that once disassociation occurs the only possible reaction is reassociation of the polymerase with the DNA template. That reassociation is modeled as occurring immediately, since we lack any experimental data on this process.

Details of the Experiments Measuring the Enzyme Kinetics of Polymerase Gamma

One must take care to consider the idealized conditions under which the experiments measuring the enzyme kinetics of polymerase gamma were conducted. Differences between these idealized conditions and the real in-vivo conditions must always be kept in mind. In this section we summarize the conditions of the experiments as described in references (Johnson and Johnson 2001; Johnson and Johnson 2001; Johnson *et al.* 2001; Johnson *et al.* 2000; Lee and Johnson 2006).

The polymerization rate kinetics

Lee and Johnson examined each incorporation of a nucleotide into the new mtDNA strand by polymerase gamma under all possible 16-base pair combinations conditions using synthetic DNA oligonucleotides and a recombinant holoenzyme of human mitochondrial DNA polymerase gamma. The recombinant holoenzyme consisted of a catalytic subunit containing a 29-amino acid truncation and an accessory subunit containing a His_6 tag and a 56-amino acid terminal truncation. The catalytic and accessory subunits were combined at a molar ratio of 1:5 to saturate the binding to reconstitute the holoenzyme. The kinetics of incorporation for all correct nucleotides were similar, with an average K_d of 0.8 μM and an average k_{pol} of 37 s^{-1}.

However, the kinetics of misincorporation varied widely. The ground state binding K_d of incorrect bases ranged from a low of 25 µM for an A:A mispair to a high of 360 µM for a C:T mispair. The rates of incorporation of incorrect bases varied from a low of 0.0031 s^{-1} for a C:C mispair to a high of 1.16 s^{-1} for a G:T mispair. (Lee and Johnson 2006)

Johnson and Johnson created an exonuclease-deficient (EXO^-) catalytic subunit to allow examination of the polymerization reaction in the absence of proofreading. Mutagenesis of Glu-200 to alanine (E200A) produced a modified enzyme with reduced exonuclease activity of single base excision from duplex DNA, but this was shown to bind DNA and to catalyze the correct base pair insertion with kinetics identical to the wild type enzyme. This mutant polymerase gamma enzyme was treated identically to the wild type for purification and experimental purposes. The wild-type polymerase was used to determine polymerization parameters for correct incorporation after a correct Watson-Crick match, for exonuclease experiments, and for the disassociation kinetics measurements following mismatched DNA. The exonuclease deficient mutant polymerase gamma was used for measuring the enzyme kinetics for the correct incorporations after a Watson-Crick mismatch and for the enzyme kinetics of incorrect incorporations (Johnson and Johnson 2001).

Johnson and Johnson preformed single nucleotide incorporation assays at various concentrations of the nucleotide substrates to examine the effects of nucleotide concentration on the incorporation rate. Burst conditions and a range of deoxyribonucleoside tri-phosphate (dNTP) concentrations were used to examine the enzyme kinetics for correct nucleotide incorporations. A reaction time course was determined for each concentration of nucleotide, and the product formation rate was plotted against time and fit to a burst equation. The burst rates (k) were then plotted against dNTP concentration and fit to a hyperbola to obtain the K_d

and the maximum rate of polymerization, k_{pol}, for each correct dNTP (Johnson and Johnson 2001).

Because the rate of incorporation after a non-Watson-Crick pair was comparable to the rate of exonuclease by polymerase gamma, Johnson and Johnson used the EXO^- polymerase to perform the experiment to determine the enzyme kinetics of a correct polymerization following an incorrect pairing. Incorporation of a dCTP was examined opposite a template G, following a T:T mismatch. The rate of the single turnover correct incorporation of dCTP was measured and the data were fit to a hyperbola, providing a K_d of 404 ± 51 μM and maximum incorporation rate of k_{pol} = 0.52 ± 0.03 s^{-1} (Johnson and Johnson 2001).

The wild-type human polymerase gamma cannot be used to measure the enzyme kinetics for incorrect polymerizations because the misincorporation is almost always corrected by the efficient exonuclease function. Therefore, Johnson and Johnson used EXO^- polymerase gamma to measure the kinetics of incorrect polymerizations. In order to limit the effects of disassociation these experiments were carried out with high polymerase gamma concentrations in excess of the template concentrations so that whenever disassociations did occur a rapid reassociation of another polymerase molecule was likely to occur, “canceling out” the disassociation event (Johnson and Johnson 2001). For each incorrect nucleotide, Lee and Johnson plotted the reaction product formation against time and fit to a single exponential. They plotted rates of polymerization determined from the single exponential against nucleotide concentration and fit to a hyperbola to determine the disassociation constant, K_d, and the maximum rate of polymerization, k_{pol} (Lee and Johnson 2006).

The exonuclease rate kinetics

To measure the exonuclease reaction rate, Johnson and Johnson reconstituted the wild-type polymerase gamma holoenzyme by mixing catalytic and accessory subunits in a 1:5 ratio. To quantify the excision reaction, the loss of full-length substrate primer due to exonuclease hydrolysis was plotted against time and fit to a single exponential (Johnson and Johnson 2001). To measure the exonuclease rate of polymerase gamma following a Watson-Crick matched base pair excision of the 3'-end base from the primer strand of correctly base-paired DNA was examined. The rate was obtained by fitting to a single exponential defining the kinetics of excision of the first base from the 3'-terminal of the primer and yielded a rate of 0.05 s^{-1} (Johnson and Johnson 2001).

To determine the effect of a mismatched non-Watson-Crick base pair on the exonuclease rate, Johnson and Johnson used three DNA substrates (containing a T:T, C:T or G:T mismatch) under conditions of excess polymerase concentration compared to template concentration. The decreases in concentrations of DNA were fit to single exponential curves. Excision of DNA containing the T:T, C:T and G:T mismatches occurred at rates of 0.40 ± 0.04 s^{-1}, 0.31 ± 0.01 s^{-1}, and 0.57 ± 0.03 s^{-1} respectively (Johnson and Johnson 2001). Since these values are similar and kinetics data were not available for all possible mismatched pairs, in our simulations we have chosen the value 0.4 s^{-1} to represent the exonuclease rate for all mismatched base pairs.

The disassociation rate kinetics

The rate-limiting step of single nucleotide incorporation is presumably the disassociation of the polymerase enzyme from the DNA. The steady-state rate is determined by the slowest step during polymerization, which must occur after the chemical reaction to account

for the observation of a pre-steady-state burst. Therefore the DNA disassociation rate of polymerase gamma holoenzyme is assumed to equal its steady-state rate of polymerization. The reported disassociation rate of polymerase gamma after a correct Watson-Crick base pair is 0.02 ± 0.001 s^{-1} (Johnson *et al.* 2000).

Johnson and Johnson used the EXO^- polymerase to examine the rate of release of polymerase gamma from a mismatched DNA substrate. The data was fit to a single exponential and yield a k_{off} of 0.18 ± 0.02 s^{-1} for DNA containing a 3' T:T mismatch (Johnson and Johnson 2001). This value is very close to the data listed in another paper (Johnson and Johnson 2001) that gave the disassociation rate of polymerase gamma after an incorrect non-Watson-Crick base pair as 0.2 s^{-1}.

Kinetics for nucleoside analogs substrates

To measure K_d and k_{pol} for each nucleoside analog triphosphate, Johnson *et al.* used an exonuclease-deficient (E200A) to reconstitute the human polymerase gamma holoenzyme. As in the previous sections, an excess of the polymerase was used in comparison to the template concentration. The kinetics of incorporation of nucleoside analog triphosphates were measured using this mutant, which was selected based upon studies showing that this single point mutation did not alter the kinetics of normal nucleotide incorporation (Johnson *et al.* 2001). To measure the exonuclease kinetics of polymerase gamma for nucleoside analog triphosphates, the wild-type polymerase holoenzyme was preincubated with DNA containing a 3'-terminal nucleoside analog to initiate the hydrolysis reaction (Johnson *et al.* 2001). No published study that we are aware of has been done to measure the disassociation rate of polymerase gamma following the incorporation of an analog triphosphate into the new mtDNA strand. With no

experimental data available, we assumed that the disassociation rates following an incorporated nucleoside analog were equal to the disassociation rates after a matched Watson-Crick pair.

The lack of experimental data on the disassociation rate of polymerase gamma following the incorporation of a nucleoside analog into the mtDNA may seem like a trivial limitation, but the analysis of the enzyme kinetics data through the stochastic simulation shows that this rate is actually critical to our understanding of the mitochondrial toxicity of these drugs (Wendelsdorf *et al.* 2009). Since the chemical alteration of these drugs prevents further polymerization after the nucleotide analog is incorporated into the DNA strand ("chain termination"), only two competing reactions are left, the exonuclease reaction which removes the nucleoside analog and the disassociation reaction which ends the mtDNA replication and defines the chain termination event. Therefore the disassociation rate following the analog is a fundamental determinant of the chain termination probability of the drug.

Problems

As the previous two sections illustrate, the primary problem with using a stochastic simulation to analyze the enzyme kinetics of polymerase gamma is the incomplete reaction kinetics data. The function of the polymerase, as represented in this simulation model, depends on the competition between different reactions which have different reaction rates. The outcome of that competition depends on the complete set of reaction rates. A change in the reaction rate of any single reaction alters the probability of all of the reactions (see equation 1). While this is an explicit feature of the simulation method, it is also a very realistic one and we should expect that the same principle should hold for the function of the real polymerase enzyme. Lack of knowledge about one reaction does not just affect that one aspect of the

polymerase function. It can in principle affect the balance between all of the other reactions which compete with that reaction. One very important practical exception to this problem is if the reaction with missing enzyme kinetics data is expected to have a very slow reaction rate in comparison to the reactions competing with it. In that case, uncertainty in the reaction kinetics of that relatively very slow reaction may cause negligible alterations in the probabilities of the competing reactions.

One important example of the limited kinetics data is that there is little data on how the template sequence may affect the reaction kinetics of polymerase gamma. As discussed earlier in section 3.1, the kinetics data that are currently available were taken from a very limited template of either a A:T pair or a T:T pair (to represent kinetics following a non-Watson-Crick pair). The lack of experimental data for other templates limits the ability of the model to analyze the potential effect of flanking sequences on mtDNA mutation rates, for example.

Equipment

No specialized hardware is required for carrying out these simulations. We have run the simulation code on both a relatively standard desktop PC under LINUX with a 4400 Intel CPU (dual core, 2.0 GHz with 2.0 GB of RAM) and on a central server (Intel Xeon CPU, 2.33 GHz). Typically, we simulate the replication of a single strand of mtDNA and repeat that simulate 10,000 times in order to gather statistics on rare events, such as specific point mutations. On the central server a set of 10,000 simulated mtDNA strand replications would take approximately 10 minutes of CPU time, and on the desktop PC would take approximately 20 minutes. The simulation that we wrote requires less than 7 MB of memory to run.

For software development, anyone wishing to carry out this model is faced with the basic choice of coding the algorithm in a standard programming language, or using one of several packages that exist for developing stochastic biochemical simulations. Two examples of these packages are StochKit (Li *et al.* 2008) and StochSim (Le Novere and Shimizu 2001). The advantage of using a prepared package such as these is that the basic algorithm is already written and debugged, and no programming knowledge is needed. However, much of the capabilities of the advanced packages, such as the relatively new StochKit, are aimed at developing approximate methods to handle simulations with large numbers of reacting molecules or to include spatial effects. These capabilities are not needed yet for a model of the polymerase gamma enzyme kinetics. Anyone using these packages must be careful to use the options which are relevant to this particular simulation, which will require at least some familiarity with the principles of stochastic simulations. We made the choice to write our own simulation code in C++. The decision to write our own code was based on the need for maximum flexibility in our definition of the model and in controlling the data output.

The Gillespie algorithm yields a series of reaction events, chosen from the reaction list of the model. For complete information about the simulation, one can choose to save this data in a number of ways. One simple choice of format is illustrated in Table I.7. Each line in the file represents one reaction event. Column 1 gives the event number. Column 2 gives the sequence position of the polymerase. Column 3 gives the template base at that position, Column 4 gives the reaction event chosen by the Gillespie algorithm, where and “A”, “C”, “G”, or “T” means the insertion of that nucleotide into the new DNA strand, a “^” indicates a disassociation event, and a “<” indicates an exonuclease reaction. Column 5 gives the time τ for that step, as defined by equation 4. Finally, column 6 gives the sequence context of the event in the format “template /

new strand". The total space needed for the file varies depending on the number of incorrect polymerization events which occur, but is generally in the range 0.5 to 1.0 MB.

Table I.7. Example format for output file of the simulation.

Event #	Sequence position	Template	Reaction	Reaction time, τ (s)	Sequence context
1012	990	T	A	0.017525	AGT/TC
1013	991	T	G	0.0809195	GTT/CA
1014	992	G	^	3.73018	TTG/AG
1015	992	G	^	0.491668	TTG/AG
1016	992	G	<	0.82397	TTG/AG
1017	991	T	A	0.0055691	GTT/CA
1018	992	G	C	0.0246826	TTG/AA

Carefully controlling the IO (the storage of data from the program onto the physical hard drive of the computer) is an important factor in the final speed of the code. This is a basic point, but it is also a common error. We chose to save in the RAM memory a time ordered list of every single reaction event that occurred in the simulation of the mtDNA strand, and then we analyzed that list for specific properties with a different analysis code, separate from the simulation code. It is a mistake to save each reaction event to a file on the hard drive as it is calculated in the simulation, as that process of initiating access to the file on the hard drive repeatedly a large number of times takes a very substantial amount of time. Instead, the list of reactions that occurred in a simulated mtDNA replication should be kept within the memory of the program and either downloaded to the hard drive in one piece at the end of the complete simulation run of a single strand replication or analyzed at the end of the simulation so that only summary statistics are saved to the hard drive.

In order to generate reliable statistics on rare events we run large numbers of repeated replications of a full mtDNA strand, typically 10,000 repetitions. Since a set of 10,000 mtDNA strand replications only takes about 10 minutes to run, and the disk space necessary for saving the 10,000 output files would be approximately 10 GB, we analyze the output of each simulated strand replication as it finishes, and then we save only the summary statistics for the specified quantity of interest for the full set of 10,000 simulated strand replications. These analysis programs were written in Perl, as the best choice for programs handling character data.

Troubleshooting

It is necessary to have quantitative and qualitative predictions against which the simulation output can be compared. There are statistical properties which the output data of the simulation must have, based on the assumptions of the algorithm. These properties should be used as checks to see if the algorithm has been properly implemented.

The distribution of reaction times

The simplest property to check is the distribution of the reaction times. From equation 4, the Gillespie algorithm assumes that the reaction times have an exponential distribution with a mean value of $1/R_{total}$, where R_{total} is the sum of all of the competing reaction rates. If you have chosen to keep the dNTP concentrations constant, then the R_{total} will also be constant and the distribution of the reaction times calculated by the simulation should have an exponential distribution with a mean value of $1/R_{total}$. We actually have two sets of reaction rates in this model, depending on whether the previous base was a correct Watson-Crick pairing or an incorrect Watson-Crick pairing. So for this test the reaction times should be separated into two

lists depending on whether the previous pairing was a Watson-Crick pairing or not. Each list of reaction times should have an exponential distribution with different means corresponding to the correct $1/R_{total}$ for that list, which can be calculated.

The mutation pattern under equal dNTP concentrations

When the four substrate dNTP concentrations are equal, it is easy to see from the experimental kinetics data in Table I.1 and I.2 that the incorrect polymerization event with the highest rate (and thus the greatest probability) is the polymerization of a G opposite to a T, forming an A to G transition mutation in the new strand. The second most common polymerization error should be the insertion of a T opposite to a G, forming the C to T transition on the new strand. All of the transversion mutations should occur at far lower rates than these two transition mutations, under equimolar dNTP concentration conditions. This can serve as a qualitative test of the accuracy of the simulation of the rare mutation events. Of course, if a different set of kinetics values are used than those given in Tables I.1 and I.2, then the prediction of the mutation pattern should change based on that data.

Using the Simulation to Analyze Polymerase Gamma Enzyme Kinetics Data

The process of the complete replication of a single strand of mitochondrial DNA by polymerase gamma is the result of tens of thousands of individual reactions. The exact sequence of reactions that occurs, the number and pattern of the point mutations introduced through polymerase errors, the number of exonuclease events and the number of disassociation events all have strong random components. The dependence of all of these important quantities on the basic measured kinetics of polymerase gamma is complex, and a direct stochastic simulation of

the polymerase activity, as described in this paper, is a thorough way of determining the end consequences of the measured kinetics values. This simulation should be thought of simply as an ordered way of determining the end result of a set of competing reaction rates. This process is of particular importance when one is considering the measured kinetics of a variant form of the polymerase gamma protein, which may be pathogenic. Many polymerase gamma variants are likely to have altered kinetics for more than one of the reactions illustrated in Figure I.1. The ultimate effect of multiple changes in the kinetic parameters of the polymerase may be quite difficult to predict correctly without the use of a simulation as a tool to test and examine hypotheses.

Future Works

Combined the pol γ replication model with a mitochondrial deoxynucleotide metabolism model

A computational model of mitochondrial deoxynucleotide metabolism and mtDNA synthesis has been built (Bradshaw and Samuels 2005). The model includes the transport of deoxynucleosides and deoxynucleotides into the mitochondrial matrix space, as well as their phosphorylation and polymerization into mtDNA. Different cytoplasmic deoxynucleotide concentrations are used to represent different simulated cell types (cancer, rapidly dividing, slowly dividing, and postmitotic cells). In our pol γ replication model, the mitochondrial dNTP concentrations are fixed values which are not change during the replication. Considering that the mitochondrial dNTP pools are fairly stable, it's a reasonable assumption. However, there are much less published data on concentrations of mitochondrial dNTP pools in patients. Most dNTP concentrations data were measured in cytoplasm. Therefore, in order to take the advantage of these available cytoplasmic dNTP pools data, we need to combine the two models together. By

building a combined model, we can analyze the effect of fluctuations of cytoplasmic dNTP pools on the replication fidelity of pol γ.

Generalizing this stochastic replication model to other polymerases

This computational stochastic model of replication of polymerase was designed for mitochondria pol γ. However, the model can be generalized to any replication process involving polymerase kinetics. One example is to simulate the template replication in Ion Torrent sequencer platform.

Ion Torrent uses the simple sequencing chemistry including natural nucleotides, with no enzymatic cascade, no fluorescence, no chemiluminescence, no optics, and no light involved in reactions. Briefly, when a nucleotide is incorporated into a strand of DNA by a polymerase, a hydrogen ion is released as a byproduct. The charge from that ion will change the pH of the solution, which can be detected by ion sensor. Ion Torrent sequencer uses the world's smallest solid-state pH meter to transfer chemical information to digital information, and then call the base.

Since there is no terminator, the replication of template strand by polymerase will continue without interruption. However, as more and more flow cycles occur, there will be more and more leading or lagging templates because replication is a dynamic process. Leading or lagging templates are those not in the same ‘phase’, which will cause signal noise called ‘phase error’. To distinguish the real signal out of background noise and accumulated ‘phase error’ noise as the read length gets longer is a real challenge. Ideally, in each time period, if there is way to know how many percentage of signal coming from ‘phase error’, that will be much helpful to analyze the signal. And this is exactly what a stochastic simulation of replication can help us.

CHAPTER II

REPLICATION PAUSES OF THE WILD-TYPE AND MUTANT MITOCHONDRIAL DNA POLYMERASE GAMMA: A SIMULATION STUDY[2]

Introduction

The holoenzyme of human pol γ consists of a catalytic subunit and a homodimer of its accessory subunit (Yakubovskaya *et al.* 2006). The catalytic subunit has DNA polymerase, 3'-5' exonuclease and 5' dRP lyase activities. The accessory subunit is required for tight DNA binding and processive DNA synthesis (Lim *et al.* 1999). Two modes of mtDNA replication have been proposed; an asynchronous strand displacement model (Brown *et al.* 2005; Shadel and Clayton 1997) and a strand-coupled bidirectional replication model (Holt *et al.* 2000).

The replication rate of DNA by a polymerase may not be stable. What is the relationship between replication pauses of pol γ and the commonly observed pathological phenotypes of mtDNA deletions and depletion? Theoretically, once a replication fork pauses, it can restart to continue the replication process without problems. But, a long pause might allow time for low-probability events such as double strand breaks. A double strand break can be repaired through blunt-ended rejoining, or homologous annealing of 5'- and 3'-repeat sequences (Krishnan *et al.* 2008), or nonhomologous end-joining (Graziewicz *et al.* 2006) that will form deleted mtDNA. If the double strand break cannot be fixed, it will cause a failed replication. Repeated failed

---

[2] Song, Z., Cao, Y., and Samuels, D. C. (2011). Replication Pauses of the Wild-Type and Mutant Mitochondrial DNA Polymerase Gamma: A Simulation Study. *PLoS Comput Biol* 7(11). Author contributions: Conceived and designed the experiments: ZS DCS. Performed the experiments: ZS. Analyzed the data: ZS DCS. Contributed reagents/materials/analysis tools: YC. Wrote the paper: ZS DCS. Algorithm design: YC ZS. Software: ZS.

replications would eventually lead to depletion of mtDNA within the cell. These concepts are illustrated in Figure 1 of Introduction Section (page 3). Under the mentioned hypothetical mechanism, mtDNA deletions and depletion may both be potential outcomes from the same initial event, the pausing of the polymerase. Certainly, other mechanisms for deletion formation are also possible (Krishnan *et al.* 2008).

Pathogenic mutations in *POLG* have been identified in patients with many neurological or muscular diseases (progressive external ophthalmoplegia (PEO), Alpers syndrome, ataxia-neuropathy syndromes, idiopathic parkinsonism) and nucleoside reverse-transcriptase inhibitor (NRTI) toxicity (Copeland 2008; Hudson and Chinnery 2006). All these diseases were characterized by mtDNA deletions and/or depletion in symptomatic tissues (Chan and Copeland 2009).

Longley reported that PEO is associated with the depletion of mtDNA and/or the accumulation of point mutations and deletions within mtDNA, caused by autosomal dominant mutations in *POLG* (Longley *et al.* 2005). Nearly all of the autosomal dominant PEO (adPEO) mutations in *POLG* were located in the polymerase domain of pol γ.

Alpers syndrome is a fatal early onset mitochondrial disorder characterized by tissue-specific mtDNA depletion (Chan *et al.* 2005). People identified pol γ mutations in Alpers syndrome are either homozygous A467T (MIM 174763.0002) or heterozygous A467T paired in *trans* with other mutations in *POLG* (Ferrari *et al.* 2005; Naviaux and Nguyen 2004).

The A467T mutation has been found in all of the major *POLG* related diseases (Chan and Copeland 2009). The frequency of the A467T mutation varies from less than 0.2% to approximately 1% in different European control populations (Horvath *et al.* 2006; Luoma *et al.* 2005; Van Goethem *et al.* 2001; Winterthun *et al.* 2005). However in mitochondrial disease populations, the A467T allele was estimated to be the most common alleles associated with

POLG diseases (Chan and Copeland 2009). People have found the homozygous A467T substitution of *POLG* in several mitochondrial disorders with highly varied phenotypes. For example, in Alpers syndrome (Boes *et al.* 2009; Utzig *et al.* 2007), two unrelated teenage boys with homozygous A467T *POLG* had ataxia and severe epilepsy. In contrast, in a case of concurrent progressive sensory ataxia, dysarthria, and ophthalmoparesis (McHugh *et al.* 2010), two siblings also homozygous for *POLG* A467T developed disease symptoms only late in life, with onset in their 40s.

The kinetic parameters of the A467T mutant for both polymerase and exonuclease activity have been measured experimentally (Chan *et al.* 2005). Compared to the wild-type catalytic subunit, the A467T substitution increased the K_m of the enzyme 5-fold while also reducing the k_{pol} value approximately 5-fold for DNA synthesis. As determined by the ratio of k_{pol}/K_m, the A467T substitution reduces DNA synthesis efficiency to 4% of the wild-type activity. In contrast, the exonuclease activity of the A467T variant is only decreased 2-fold compared to the wild-type (Chan *et al.* 2005). Besides altering the kinetics of polymerase γ, the A467T variant in *POLG* also decreases the protein's physical association of the POLG2 subunit (Chan *et al.* 2005).

People created two independent homozygous knock-in mouse models with proof-reading deficient versions of the catalytic subunit of mtDNA polymerases. One published in 2004 (Trifunovic *et al.* 2004) and the other one in 2005 (Kujoth *et al.* 2005). Both mouse models showed similar progeria-like phenotypes and shared the same D257A mutation on the second exonuclease domain of *PolgA* (the mouse homolog of *POLG*) that caused a profound reduction of the exonuclease activity but no decrease in DNA polymerase activity. Both mouse models (Kujoth *et al.* 2005; Trifunovic *et al.* 2004) were reported to have accumulated mtDNA point mutations.

Using these two mouse models, other mitochondrial genome variations besides point mutation have been studied. For the first mouse model, after Trifunovic *et al.* (Trifunovic *et al.* 2004) reported increased mtDNA deletion, Bailey *et al.* (Bailey *et al.* 2009) using one- and two-dimensional agarose gel electrophoresis interpreted these non-replicating linear mtDNAs as the result of double-strand breaks of cyclic mtDNA caused by stalled replication intermediates. Ameur *et al.* (Ameur A. *et al.* 2011) further stated that mutator mice have abundant linear deleted mtDNA molecules but extremely low levels of circular mtDNA molecules with large deletions. For the second mouse model, Williams *et al.* (Williams *et al.* 2010) applying next-generation sequencing to native mtDNA did not observe the accumulation of mtDNA deletions but instead reported multiple copies of the mtDNA control region in brain or heart. However, Vermulst *et al.* (Vermulst *et al.* 2008; Vermulst *et al.* 2009) have identified mtDNA deletions as a critical driving factor behind the premature aging phenotype in these mouse models and have suggested there is a homology-directed DNA repair mechanism directly linked to mtDNA deletion formation. As this description shows, the data on the amount of deletions in these two mouse models is currently mixed and this is a research area that is rapidly developing.

To study the replication process of pol γ, we have built a computational model of the activity of pol γ (Song and Samuels 2010; Wendelsdorf *et al.* 2009) (Figure II.1), based on the experimental kinetic data. Our computational model of the mitochondrial DNA replication process was based on the Stochastic Simulation Algorithm (Gillespie 1976; Gillespie 1977), a well-known Monte Carlo simulation method for chemical reactions. Using this stochastic simulation, we were able to quantify the pause lengths for the wild-type polymerase and for the pathogenic A467T and the exonuclease deficient pol γ variants.

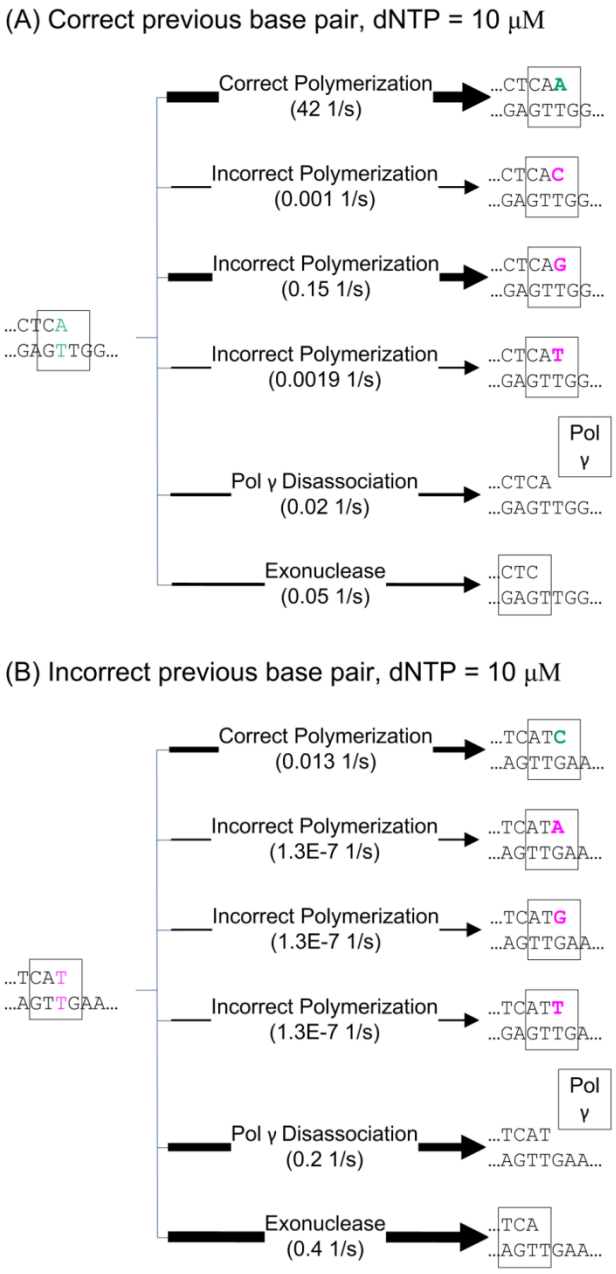

Figure II.1. Diagrams of the six competing reactions of pol γ. Example reaction rates are given below each reaction name. The line size approximately represents the reaction rate. The bigger the reaction rate, the wider the line. (A) For a correct previous base pair, the highest reaction rate by far is the correct polymerization. (B) For an incorrect previous base pair, the highest reaction rate is the exonuclease reaction, allowing proofreading. All reaction rates in this diagram are calculated for dNTP pool levels equimolar at 10 μM.

Methods

Pol γ replication model

In this model, pol γ carries out four basic types of reactions: DNA polymerase activity, exonuclease activity, disassociation of the polymerase from the DNA, and reassociation of the polymerase with the DNA molecule (Figure II.1). In the DNA polymerase reaction, pol γ adds one nucleotide to the new DNA strand. This nucleotide could be either a correct or incorrect (point mutation) base. In the exonuclease reaction, pol γ removes one nucleotide from the new DNA strand. The exonuclease reaction is an error correction mechanism, as the rate for removal of incorrectly incorporated nucleotides is much faster than that of correctly incorporated nucleotides, though the removal of both correct and incorrect nucleotides is allowed in the model. In the disassociation reaction, the polymerase separates from the DNA molecule. In the reassociation reaction (not shown in Figure II.1), the polymerase re-attaches to the DNA molecule after disassociation.

At each position on the replicating mtDNA strand, pol γ randomly undergoes one of a series of six possible competing reactions: one correct polymerase reaction, three incorrect polymerase reactions, the exonuclease reaction, or disassociation (Figure II.1). Which reaction pol γ undergoes at each time is determined by the propensity of each reaction, calculated using the reaction rates following the selection formulae as introduced in Gillespie (Gillespie 1976; Gillespie 1977). The simulation was implemented in C++ and run under LINUX.

Michaelis-Menten kinetics was used for all of the DNA polymerization reactions. The exonuclease reaction and the pol γ disassociation reaction were set to have constant reaction rates. For the two scenarios of a correctly inserted and incorrectly inserted previous nucleotide we had separate sets of kinetic parameters for each of the pol γ reactions taken from Lee and

Johnson (Lee and Johnson 2006) and Johnson and Johnson (Johnson and Johnson 2001; Johnson and Johnson 2001) (Tables 1.1-1.5). These studies have reported an increase in exonuclease and disassociation rates, but a decrease in incorporation rates by pol γ following an incorrect incorporation. This was included in the simulation model by using two sets of enzyme kinetics parameters, one set for reactions following a correct incorporation and another set for reactions following an incorrect incorporation. For an extended description of the simulation model, please see the Methods paper by Song and Samuels (Song and Samuels 2010).

Simulation method

In each simulation step, this algorithm calculates the propensity of each reaction and generates two random numbers to determine the time between reactions and the identity of the next reaction. The original Gillespie algorithm (Gillespie 1976; Gillespie 1977) calculates propensities based purely on the substrate species' population (i.e. integer copy number). In our method, since we are only concerned with the dynamics of pol γ, we do not have to convert the state variable of every species to its copy number. In this way, the propensity functions are calculated from the same formula as the reaction rates for the involved competing reactions. These propensities were then used to randomly choose the sequence of reactions that occur, at the level of individual molecular reactions following the Gillespie algorithm.

Pol γ kinetics parameters

Given the complexity of pol γ activity, it should not be surprising that even the large set of pol γ kinetic parameters that have been measured is still incomplete. The data used in Tables 1.1 and 1.2 was measured only in the case where the previous base pair was an A:T. In principle the reaction rates could be different for other preceding bases, however in the absence of this

data we have applied the measured reaction kinetics to all cases where the previous base pair is a standard Watson-Crick pairing. The data on kinetics following a non-Watson-Crick base pair are even more limited. The k_{pol} and K_m values reported (Johnson and Johnson 2001) for this case were only determined for the correct pairing of a C opposite to a G in the template strand, where the previous incorrect base pairing was a T:T. We define an approximate model by setting the kinetic coefficients for all correct incorporations following an incorrect incorporation to be equal to this reported value (Tables 1.3 and 1.4). Reaction kinetics for the incorporation of an incorrect nucleotide following an incorrect nucleotide are not available, to the best of our knowledge. Values for these parameters in the model (Table 1.3, off diagonal elements) were based on the observation from Table 1.1 that k_{pol} values were approximately 1000 times less for non-Watson-Crick pairing compared to regular Watson-Crick pairings. Similarly, based on Table 1.2 K_m values were estimated to be approximately 100 times greater for the non-Watson-Crick pairings and this was used to estimate K_m values in Table 1.4 for the off-diagonal elements.

There are currently two main groups (Chan *et al.* 2005; Johnson and Johnson 2001) that have independently measured the kinetics of wild-type polymerase γ. One (Johnson and Johnson 2001) employed pre-steady state measurement using single-turnover analysis to determine dNTP incorporation rate k_{pol} and the dissociation constant K_d of dNTP to bind to the polymerase-primer-template complex. The other one (Chan *et al.* 2005) employed steady state kinetic measurement that assumes that the complex of polymerase, primer and template behaves as a single enzyme and fit the data to Michaelis-Menten function to determine its k_{cat} and K_m. There are some disagreements of kinetic values between these two measurements and this inconsistency has recently been further addressed in the literature (Estep and Johnson 2011). We chose not to mix values of wild-type polymerase γ from different measurement

techniques, and instead we chose to continue to use the pre-steady state values (Johnson and Johnson 2001), as we have in several previous publications.

However, the A467T variant kinetics have only been published using the steady state measurement technique (Chan *et al.* 2005). To make the most reasonable estimation of the A467T kinetics based on published values consistent with the experimental methods (Johnson and Johnson 2001; Johnson and Johnson 2001) used to determine the wild-type POLG kinetics, we followed the observation of Bertram *et al* (Bertram *et al.* 2010) that even when polymerase enzyme kinetics measurements from different measurements were not in agreement, the dimensionless ratios of those kinetics constants often were in agreement. The A467T mutant enzyme is reported to have only 4% of wild-type DNA polymerase activity (the k_{cat} decreased 5-fold and K_m increased 5-fold compared to the wild-type) and 50% of the wild-type exonuclease activity (Chan *et al.* 2005). We used these proportions to estimate the A467T variant kinetics based on the wild-type kinetics measured by pre-steady state methods (Johnson and Johnson 2001; Johnson and Johnson 2001). Finally, we modeled an idealized exonuclease deficient polymerase by setting the exonuclease activity to zero and keeping the other kinetics values at the wild-type values. The exonuclease deficient mice are reported to have normal DNA synthesis capacity, but no detectable exonuclease activity (Trifunovic *et al.* 2004).

Data regarding the reassociation reaction rate is not available to the best of our knowledge. For this model we assume that once disassociation occurs the only possible reaction is then reassociation of the polymerase with the DNA template. Without reaction kinetics, the time required for this reaction cannot be calculated and is not included in this model. This approximation is not important to our results reported here on point mutation rates. The time required for reassociation would extend even further the pol γ pausing times that we estimate based on the computational model.

Substrate Concentrations

The polymerization reaction rates are functions of the deoxyribonucleotide triphosphate (dNTP) concentrations. However, concentrations of dNTP pools in mitochondria vary in different species, tissues, cell types and cell cycle phases. Measured dNTP pools in mitochondria of quiescent and cycling human fibroblasts are listed in Table II.1 (Ferraro *et al.* 2005). The units of picomoles per 10^6 cells were converted to μM by using an assumed 200 femtoliters of mitochondrial volume per cell (0.2 femtoliters each mitochondria, 1000 mitochondria per cell) (Pollak and Munn 1970). The four dNTP concentrations can be set individually in the model. In all cases reported here we have kept the four dNTP levels equal at concentrations of either 1 or 10 μM to approximately represent the conditions of postmitotic and dividing cells respectively (Ferraro *et al.* 2005). The dNTP levels were held constant throughout the simulated mtDNA replication.

Table II.1. Concentrations of dNTP pools in mitochondria of different human cells.

dNTP (μM)	Quiescent skin fibroblasts	Cycling skin fibroblasts
dATP	5.6	14.3
dCTP	2.5	20.4
dGTP	1.3	4.6
dTTP	1.5	19.9

Data taken from references (Ferraro *et al.* 2005)

mtDNA sequence

Vertebrate mitochondrial DNA has a highly asymmetric G content. The low-G strand is labeled the light strand, with the complement strand called the heavy strand. Sets of simulations were carried out separately for the light strand sequence (GenBank NC_001807) and the heavy

strand sequence, formed by inverting and complimenting the light strand sequence. We found no significant difference in the results reported in this paper between the light strand and the heavy strand. The data reported here are based on replication of the heavy strand, which takes the light strand as the template.

The time required for a single forward step

We define the time for pol γ to move from sequence position *i* to the next position *i+1*, which could include multiple intermediate reaction events, as the "single forward step time". For each simulated mtDNA strand replication, there are 16,570 single forward steps and 16,570 time intervals. We choose the longest single forward step time to represent the worst polymerase pausing event that happened in the strand replication simulation.

The number of simulations

We have three types of pol γ (the wild-type, A467T and exo$^-$) and two cell type conditions (dividing and postmitotic). Since the model is stochastic, the results vary between runs with identical parameter values. For each one of the six scenarios, we ran the simulation 10,000 times to determine the probability distribution of the model results.

Results

We modeled three types of pol γ, the wild type and two mutants (the human pathogenic variation A467T and the completely exonuclease deficient form exo$^-$). With frequencies of the A467T variant in different European control populations of 0.2-1% (Horvath *et al.* 2006; Luoma *et al.* 2005; Van Goethem *et al.* 2001; Winterthun *et al.* 2005), the naturally occurring heterozygous A467T alone, without paring in *trans* with other mutations in *POLG*, causes only mild, if any, pathogenic phenotypes (Ferrari *et al.* 2005; Naviaux and Nguyen 2004; Nguyen *et al.* 2005; Van Goethem *et al.* 2004). An exonuclease deficient POLG has never been reported naturally in humans or mice This severe POLG variant only occurs in the genetically engineered mouse models (Kujoth *et al.* 2005; Trifunovic *et al.* 2004). Our hypothesis is that the pathogenic severity of a mutated pol γ depends on its susceptibility to replication fork pauses, not just its rate of point mutation formation.

The single strand replication time

We began by measuring the full replication time of a single strand of mtDNA in this computational model. The replication time of a single strand is defined simply as the time needed for the simulation model to complete the replication of all 16,571 base pairs in the mtDNA reference sequence.

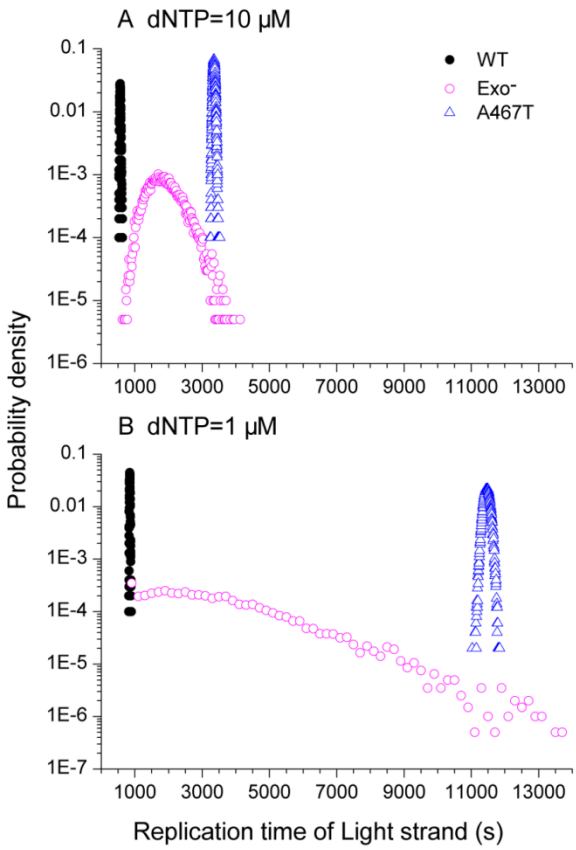

Figure II.2. Probability distribution of the replication time of a single strand of mtDNA. In these examples, the light strand was used as the template. (A) dNTP pool levels equimolar at 10 µM, representing dividing cells. The median values of the replication time distributions of the wild type (WT, black dots), exo$^-$ (magenta open circles), and the A467T variation (blue open triangles) are 563 s, 1837 s, and 3355 s respectively. (B) Equimolar 1 µM dNTP concentrations. The median values of the distributions of the WT, exo$^-$, and A467T variant are 849 s, 2946 s, and 11465 s respectively.

After running 10,000 simulations of replication of a single strand, the probability distributions of the replication time of wild-type and mutated pol γ are calculated (Figure II.2).As expected from the enzyme kinetics, the naturally occurring mutant A467T simulation had a much longer replication time than the wild-type pol γ simulation. However, surprisingly, the severe pathogenic mutant exo$^-$ had a highly variable replication time, ranging from times only slightly longer than that of the wild-type to replications taking longer than the A467T mutant

simulation. In contrast, both the wild-type and A467T simulations had very tight distributions, meaning that the replication time had little random variation. Under the condition of dividing cells (Figure II.2A, dNTP = 10 μM), the polymerase reaction rate of the A467T variant is approximately 7 times slower than the reaction rate of the wild-type pol γ. Considering the difference of kinetics values, the median strand replication time of 564 seconds for the wild-type and 3355 seconds for the A467T variant were reasonable.

Compared to dividing cells, postmitotic cells have lower dNTP levels and therefore pol γ in these cells has a lower polymerization rate and the replication process proceeds more slowly. Under this condition, we expected to see longer strand replication time (Figure II.2B, dNTP = 1 μM). The median of strand replication time of wild-type pol γ increased 1.5-fold to 849 seconds, compared to the rate at dNTP = 10 μM. For the exo$^-$ variant, the median replication time increased similarly from 1837 to 2946 seconds. For the A467T variant, the median replication time increased 3.4-fold to 11,465 seconds (3.2 hours).

The time required for the slowest single forward step of the polymerase

The total replication time (Figure II.2) is an average quantity, representing the summed effect of often tens of thousands of reactions. Pathogenic effects, such as deletion formation, may be more directly related to extreme single events that occur during the replication process than to this average quantity. To measure this, for each simulated mtDNA strand replication we recorded the time required for the longest single forward step of the polymerase (Figure II.3). See the Methods section for a precise definition of this measure. We interpret these slowest forward steps in the polymerase activity as polymerase pauses.

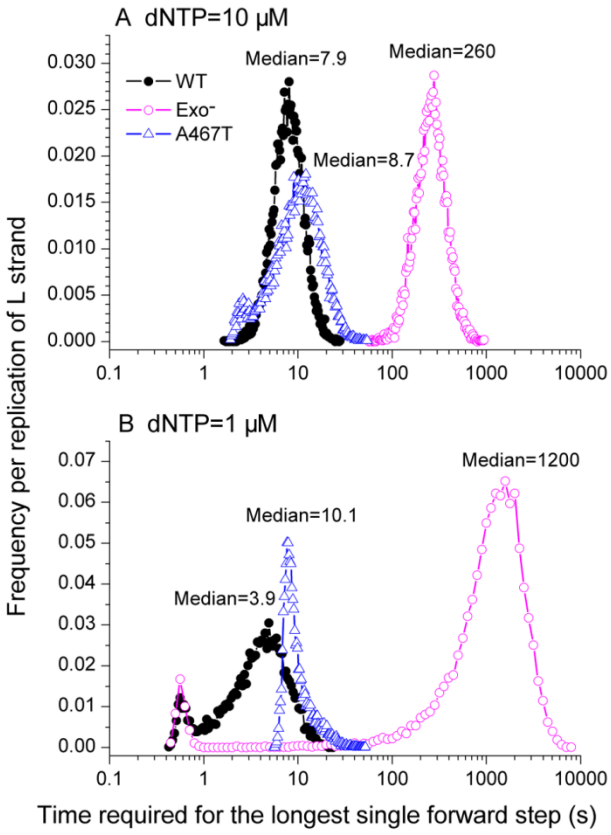

Figure II.3. Frequency distribution of the time required for the longest single forward step. The median values of each curve are shown. (A) High dNTP level representing dividing cells. (B) Low dNTP levels representing postmitotic cells.

The exo$^-$ simulation had extremely lengthy maximum single forward step times, corresponding to pol γ pauses (Figure II.3). Compared to the wild-type, the exo$^-$ polymerase had a 30 (Figure II.3A) to 300-fold (Figure II.3B) longer median of the longest single forward step time. However, the median of the A467T variant's longest single forward step time only increased slightly compared to the wild-type. This is reasonable because the slow polymerase reaction rate of A467T slowed down every forward step of the polymerase, without preferentially affecting the extremes.

For dividing cell conditions (Figure II.3A, dNTP = 10 μM), the median time of pausing of the wild-type polymerase was 7.9 seconds, while it was a similar value of 8.7 seconds for the A467T variant. However, the median time of pausing of the exo$^-$ polymerase was much longer at 260 seconds. For postmitotic cell conditions (Figure II.3B, dNTP = 1 μM), the median time of pausing of the wild-type polymerase decreased to 3.9 seconds. The median time of pausing of the A467T variant increased slightly to 10.1 seconds, however, the exo$^-$ pol γ had a dramatically increased median pausing time of 1200 seconds. Under both dNTP concentration conditions, the exo$^-$ polymerase had the slowest "longest single forward step time", implying the longest pauses and the greatest opportunity of deletions to occur.

The reaction events in the longest single forward step

We further identified what reaction events actually occurred in the longest single pausing step (Figure II.4 and II.5). The longest pausing steps usually occurred after incorrect polymerase events (Table II.2), which inserted a non-Watson-Crick paired nucleotide into the new DNA strand. The exceptions to this (i.e., pauses after correct incorporations) were always very brief pauses, which can be seen as the small peaks to the extreme left side in Figure II.3. In most of the cases of pausing after the incorrect nucleotide insertion, for both the wild-type and A467T variant, the longest single forward time step has one exonuclease reaction (to remove the incorrect nucleotide), two polymerase reaction (one to replace the removed nucleotide and a later one to move the polymerase forward to the next step), and a variable number of dissociation events (Figure II.4).

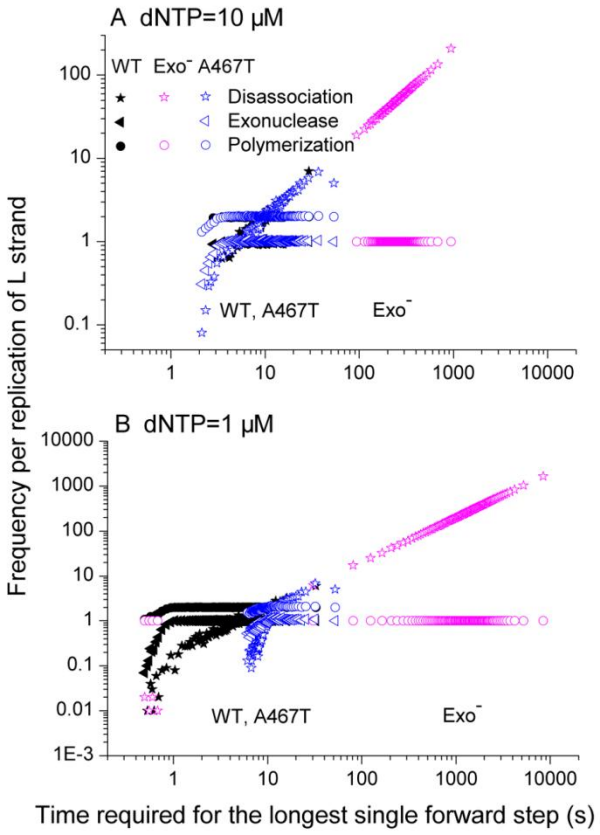

Figure II.4. Events in the longest single forward step. For each pol γ variant, 10,000 simulations were carried out and the probability distributions of the events were calculated. Symbol colors represent the pol γ variants (black = wild-type, magenta = exonuclease deficient, blue = A467T). Symbols represent the reaction events (star = disassociation, triangle = exonuclease, circle = polymerization).

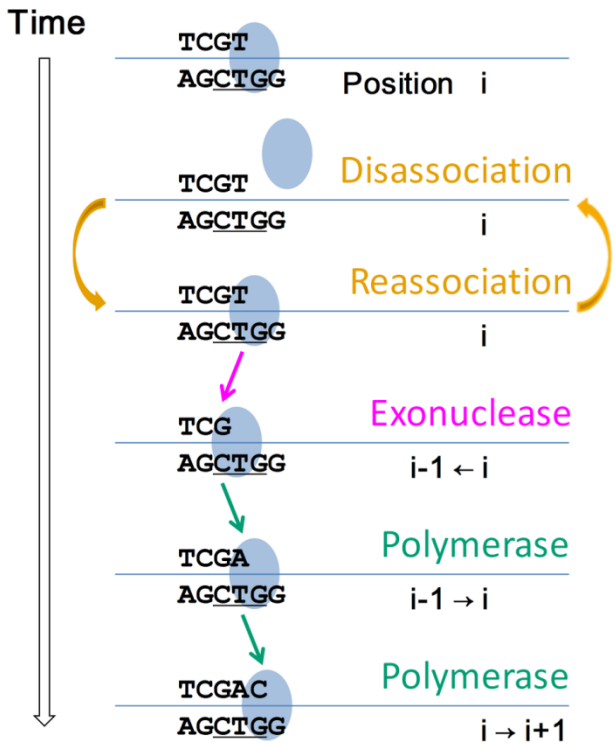

Figure II.5. An example of events in a typical longest forward step of the polymerase. For the wild-type (WT) and the pathogenic A467T pol γ variant, a typical single forward step includes several dissociations and reassociations, one exonuclease reaction and two polymerizations.

Table II.2. Strand replication time statistics.

Simulation template	dNTP (μM)	Mean replication time (s)	Probability of incorrect previous base pair	Median longest single forward step time (s)
Light strand	1	849.4	92.69%	3.9
Heavy strand	1	813.2	95.80%	4.4
Light strand	10	563.8	100.00%	7.9
Heavy strand	10	542.7	100.00%	8.4

For the exo$^-$ polymerase, there was one polymerase reaction and a variable but large number of dissociation events. In very rare cases, where pol γ replicated the whole strand without any error ever occurring, the longest single forward step time was very short and there was only one polymerase event and very few dissociation events.

We found that the pausing time of pol γ was roughly proportional to the number of disassociations events occurring within that pause (Figure II.4). The exo$^-$ pol γ had extremely long polymerase pauses, with a 30 to 300-fold increase in the time required for the longest single forward step of the polymerase compared to the wild-type polymerase, while the naturally occurring A467T pol γ variant had only a slight increase in the length of the pauses compared to the wild-type pol γ. It is the number of polymerase disassociations occurring in each forward step of the polymerase that causes the differences. Due to the lack of competition of the exonuclease activity in the exo$^-$ pol γ to remove the incorrect incorporation, disassociations of pol γ repeatedly recurred in each longest single forward step. The variation in the pausing time was almost completely due to variation in the number of disassociation events occurring during the pause (Figure II.4).

The effect of a partial loss of pol γ exonuclease activity

The exo$^-$ mutant is an extreme alteration of the polymerase function, which we have modeled as the complete loss of exonuclease activity. What if a mutated pol γ lost exonuclease activity only partially, not completely? To answer this question, we decreased the exonuclease reaction rate in the simulation gradually from 100% activity, equivalent to the wild-type, to 0%, equivalent to the exo$^-$ mutant simulation (Figure II.6). We calculated the median longest single forward step time (Figure II.6A) and the point mutation rate (Figure II.6B) for both dividing and postmitotic cell conditions. These two quantities were chosen to represent the two main mechanisms for generating mtDNA damage from a defective polymerase: deletions driven by polymerase pausing and point mutations. Surprisingly, pol γ has to lose approximately 50 to 90% of its wild-type exonuclease activity before there are significant consequences shifting either of these measures by more than a factor of two.

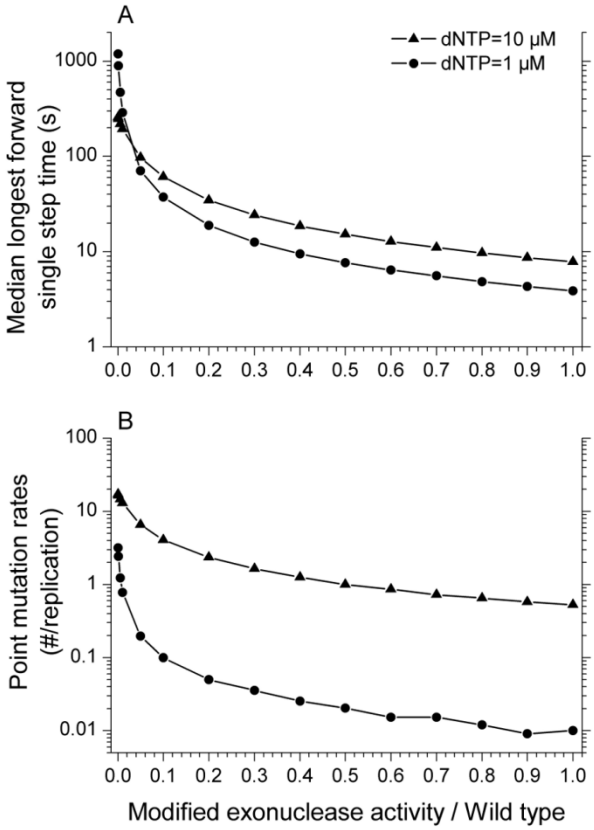

Figure II.6. The effect of varied exonuclease activity. (A) Median longest forward time step occurring within a replication from the mtDNA light strand template (taken over 10,000 repeated simulations). (B) Point mutation rate per strand replication. With the decreasing of exonuclease activity from 100% activity to zero, both the median value of longest forward single step time and point mutation rate increased nonlinearly.

Discussion

The mitochondrial DNA replication time has been measured in dividing cells in cell culture, and approximately one hour is required for replication of both strands (Clayton 1982). In our simulation, the light strand replication time under dividing cell condition (dNTP = 10 μM) was less than 10 minutes (Figure II.2A), and the heavy strand replication required a similar time (Table II.2). The standard asynchronous strand displacement replication model makes the double strand replication time (1 + 2/3) fold longer than that of single strand, where the 2/3 represents the fraction of the mtDNA sequence in the major arc between the heavy strand origin and light strand origin. Based on this estimate, the full mitochondrial genome replication time in our simulation was only 17 minutes. The simulated double strands replication time is 3-5 fold less than what has been reported in experiments (Clayton 1982). This fast replication time is a direct consequence of the measured enzyme kinetics values for pol γ. However, the kinetic parameters we used were measured *in vitro* (Johnson and Johnson 2001; Johnson and Johnson 2001), and the actual values *in vivo* may be different. Another reason why our simulated replication time is shorter than expected is that we could not include the reassociation time of pol γ with the DNA following a disassociation event, since there is no data on the reassociation rate. In addition, we cannot include the effect of the proteins mtSSB, Twinkle and other accessory proteins on the replication time, since these accessory proteins were not included in the basic experiments measuring the polymerase enzyme kinetics. All replication time in our simulation were underestimated for these reasons, which would make the pausing of pol γ in reality even worse than calculated from these simulations. Another relevant feature that cannot be included in this simulation is the role of DNA secondary structure in polymerase pausing.

These secondary structures would cause additional pausing of the polymerase, beyond that calculated in this simulation.

The simulation results gave us a clear picture of how pol γ pauses may happen. For wild-type and the A467T pol γ variant, the longest single forward step always starts with an incorrect nucleotide incorporation. For example if this happens at position *i* along the DNA strand then an exonuclease event will happen with high probability at position *i*, removing the incorrect nucleotide, and the pol γ moves backward one step. After this, two following polymerase events are needed to move the polymerase forward to the position *i+1*. In this case, the longest single forward time step from position *i* to *i+1* requires at least three reaction events, one exonuclease reaction and two correct polymerizations, as well as any disassociation and reassociation events that may have happened between these events (Figure II.5). In principle, there could be *N* exonuclease events and *N+1* polymerase events in each single forward step. However, 10,000 repeated simulations showed that there were almost always only one exonuclease event and two polymerase events (Figure II.4). For the exo^- polymerase, once an incorrect incorporation happens at position *i*, the polymerase keeps repeating a disassociation-and-reassociation cycle until a polymerization happens and moves the polymerase forward to position *i+1* (leaving behind a point mutation). The disassociation-and-reassociation cycle repeats because the rate of polymerization following the incorrect nucleotide incorporation is very low. In the wild-type polymerase, the slow polymerization following an incorrect nucleotide incorporation allows time for the exonuclease reaction to occur with high probability, repairing the DNA replication error.

Several experiments (Chan *et al.* 2005; Fan *et al.* 2006; Johnson *et al.* 2000) have measured the processivity of polymerase gamma. However, there is a fundamental difference between processivity as measured in these experiments and the concept of “polymerase

pausing" as defined in our simulation. The polymerase processivity experiments (Chan *et al.* 2005; Fan *et al.* 2006; Johnson *et al.* 2000) always use an unlabeled DNA trap to deliberately limit the polymerase to the very first disassociation event. At the first disassociation event the polymerase is captured by the DNA trap with very high probability and the polymerase reassociation is prevented, ending the DNA replication. In contrast, we define pausing as the longest time required to move the polymerase forward one nucleotide along the DNA strand, a process that we show in the simulation often includes several polymerase disassociation-reassociation cycles. Those extreme cases where multiple disassociation-reassociation cycles occur (by random chance) are the slowest step in the full mtDNA replication, which we take as a representation of polymerase pausing. As we show in Figure II.4, the difference between the wild-type POLG and the exonuclease deficient POLG comes in the number of polymerase disassociation-reassociation cycles that occur in the slowest forward step of the polymerase on the DNA strand. This difference would not be seen in the processivity experiments where reassociation is deliberately prevented by the experimental design. Because of this fundamental difference, the processivity experiments are not in disagreement with this simulation model. A further difference between the processivity experiments and the polymerase pausing is that the processivity experiments only measure the behavior of the polymerase over hundreds to a few thousand bases (Fan *et al.* 2006; Johnson *et al.* 2000), while the polymerase pausing (as we define it) is a measure of the most extreme replication event (the slowest forward step) over 16,600 bases. Those rare events simply could not be observed in the much shorter processivity experiments.

A recent study (Williams *et al.* 2010) reported at least two orders of magnitude increase of point mutations in exo$^-$ mice compared to controls in postmitotic tissues – brain and heart. In Figure II.6B, under the postmitotic cell condition (dNTP = 1 μM), the point mutation rate with

100% wild-type exonuclease activity was 0.01 (number of uncorrected errors per strand replication) and the rate increased to 3.2 for the completely exonuclease deficient polymerase simulation. This is consistent with the experimentally measured difference between the exonuclease deficient mouse and the wild-type mouse (Williams *et al.* 2010). In an exo$^-$ mouse model (Trifunovic *et al.* 2004), the measured mutation load of the exo$^-$ polymerase γ was 3-13 mutations per 10 Kb (approximates to 5-21 mutations per replication) for brain and heart cells (which can be taken as postmitotic) and liver cells (which can be taken as mainly mitotic). This experimental data is consistent with our simulation results (Figure II.6B), which shows 3.2 mutations per replication for postmitotic cells (dNTP = 1 μM) and 17 mutations per replication for mitotic cells (dNTP = 10 μM).

Based on this simulation model, a reduced DNA binding effect caused by variants in the accessory subunit might also be another reason to cause pausing of the polymerase. There is growing interest in variants in *POLG2,* the gene that encodes accessory subunit p55 of polymerase gamma (Young *et al.* 2009). Reduced DNA binding is not obviously related to mtDNA depletion, but our simulations clearly showed that the number of disassociation events that occur during the pause directly determines the length of the polymerase pause. Thus, based on this simulation model, reduced DNA binding of the polymerase to the mtDNA molecule could greatly increase the length of pauses of the polymerase both by increasing the probability of disassociation events and by slowing the rate of the reassociation reaction.

In addition to the altered mtDNA polymerization kinetics, the A467T variant in *POLG* also has altered kinetics in the formation of the protein complex with POLG2 (Chan *et al.* 2005). We deliberately chose to model just the process of replication of the mtDNA, and not the preceding process of the formation of the polymerase protein complex from POLG and POLG2. Extending the model to include the stage of formation of the polymerase complex would greatly

extend the complexity of the model, and would distract from the purpose of the model, which was to represent the process of mtDNA replication after replication is initiated. The kinetics values used in this model are from experiments (Chan *et al.* 2005; Johnson and Johnson 2001; Johnson and Johnson 2001; Johnson *et al.* 2000) using POLG and POLG2 together.

Several presumably pathogenic amino acid substitutions in pol γ have been identified in PEO patients, including G923D, R943H (neither are currently in OMIM) and A957S (MIM 174763.0014). Although all of these have been characterized biochemically (Graziewicz *et al.* 2004; Ponamarev *et al.* 2002) with some limited kinetics, none of them has had its exonuclease activity measured yet. Without that critical data, we cannot simulate these variants. Recently, the pathogenic pol γ mutation Y955C (MIM 174763.0001), identified in Autosomal Dominant Progressive External Ophthalmoplegia, was reported to stall at dATP insertion sites (Atanassova *et al.* 2011). Unfortunately, many important kinetic values of the Y955C pol γ variant are not available, such as the exonuclease and dissociation reaction rates, although its polymerization kinetics have been reported (Estep and Johnson 2011; Ponamarev *et al.* 2002).

Future Works

Other pathogenic pol γ mutators

Y955C mutation was the first adPEO mutations to be discovered(Ponamarev *et al.* 2002). It has been analyzed the effects of the mutation on the kinetics and fidelity of DNA synthesis by the purified human mutant polymerase in complex with its accessory subunit(Ponamarev *et al.* 2002). As compared with the wild-type enzyme in the presence of the accessory subunit, the Y955C pol γ exhibited a 4-fold reduction in overall DNA polymerase activity in a standard DNA polymerase assay. Steady state kinetic analyses indicated that the Y955C enzyme retains a wild-

type catalytic rate (k_{cat}) but suffers a 45-fold decrease in apparent binding affinity for the incoming nucleoside triphosphate (K_m). The full mutator effect of the Y955C substitution was revealed by genetic inactivation of the exonuclease, and error rates for certain mismatches were elevated by 10–100-fold. This enhanced mutagenesis is mitigated by a functional intrinsic exonuclease activity resulting in only a 2-fold mutator effect for base pair substitutions by the exonuclease proficient Y955C enzyme.

After A467T, the Y955C is the next valuable pol γ mutator to be analyzed. Because its K_m value decreased dramatically, Y955C is expected to more sensitive to the substrate concentrations. It will be interested to simulate and compare the point mutation rate and the longest single forward step time of Y955C to wild-type pol γ. Though the exonuclease reaction rate of the Y955C mutation has not been measured, we still can make reasonable assumptions about this rate to explore its pathogenic effect in adPEO.

CHAPTER III

AN ANALYSIS OF ENZYME KINETICS DATA FOR MITOCHONDRIAL DNA STRAND TERMINATION BY NUCLEOSIDE REVERSE TRANSCRIPTION INHIBITORS[3]

Introduction

Recent guidelines for highly active anti-retroviral treatment (HAART) regimens of HIV-positive patients recommend two drugs of the nucleoside reverse transcriptase inhibitor (NRTI) class (Table III.1) (Antiretroviral_Guidelines_for_Adults_and_Adolescents_Panel 2006). This class currently consists of: stavudine (d4T), lamivudine (3TC), zidovudine (AZT), zalcitabine (ddC), didanosine (ddI), abacavir (ABC), emtricitabine (FTC) and tenofovir (TDF, a nucleotide analog). Though zalcitabine (ddC) is still technically approved for treatment its distribution in the United States was discontinued by Roche in 2006. In their activated tri-phosphorylated forms, each NRTI acts as a nucleotide analog interacting with the HIV viral reverse transcriptase as an alternative substrate to the natural nucleotides (Anderson *et al.* 2004; Ray 2005). Each of these analogs lacks the 3' OH group necessary for incorporation of the next nucleotide thereby terminating viral DNA strand elongation.

---

[3] Wendelsdorf, K. V., Song, Z., Cao, Y., and Samuels, D. C. (2009). An Analysis of Enzyme Kinetics Data for Mitochondrial DNA Strand Termination by Nucleoside Reverse Transcription Inhibitors. *PLoS Comput Biol* 5(1). Author contributions: Conceived and designed the experiments: YC DCS. Performed the experiments: KVW ZS. Analyzed the data: KVW ZS. Contributed reagents/materials/analysis tools: YC. Wrote the paper: KVW ZS DCS.

Table III.1. Nucleoside and nucleotide analogs used in this study.

Drug	Abbreviation	Natural nucleoside	Comment
Abacavir	ABC	dG	The activated form is CBV-TP
Acyclovir	ACV	dG	Used in treatment against Herpes viruses including HSV1 and 2, chickenpox and herpes zoster.
Didanosine	ddI	dA	Must be aminated to become ddA, which is activated form. The nonactivated form, however, has also been shown to incorporate into mtDNA.
Dideoxyadenosine	ddA	dA	The active form of ddI.
Emtricitabine	FTC(-)	dC	The unnatural enantiomer that is approved for treatment of HIV.
Emtricitabine	FTC(+)	dC	The natural enantiomer that is more toxic and not approved for treatment.
Lamivudine	3TC(+)	dC	This natural enantiomer of 3TC is not used in treatment, but is used in studies that look at effects of configuration on toxicity and efficacy.
Lamivudine	3TC(-)	dC	The unnatural enantiomer that is approved for treatment of HIV.
Stavudine	d4T	T	
Tenofovir	PMPA, TDF	dAMP	A nucleotide analog
Zalcitabine	ddC	dC	Not currently recommended for clinical use.
Zidovudine	AZT	T	

NRTIs are effective drugs that have helped usher HIV into the category of a controllable chronic disease, however, they are also often toxic, inducing side effects such as lactic acidosis, peripheral neuropathy, peripheral lipoatrophy, and pancreatitis in patients. Intolerance of such side effects is a common reason for treatment discontinuation (d'Arminio Monforte *et al.* 2000). It is a serious concern that any decrease in patient compliance to the treatment regimen can lead to an increase in viral resistance and ultimately to treatment failure. The 1st step in ameliorating these side effects and preventing them in future antiviral treatments is to understand the mechanisms behind the mitochondrial toxicity of the NRTIs that are in use today. People have proposed many mechanisms of mitochondrial toxicity which we discuss below. In

this paper we specifically consider the plausibility of the most widely accepted hypothesis for the NRTIs' toxicity mechanism - interference of mitochondrial DNA replication.

The polymerase γ hypothesis

Polymerase γ (pol γ) is the only polymerase responsible for mitochondrial DNA replication. While pol γ is not believed to directly regulate mtDNA levels, pathogenic mutations in the gene *POLG* do affect the stability of mtDNA and cause mtDNA depletion (Copeland 2008). Polymorphisms found in the *POLG* gene in the human population may cause a natural variability in the activity of this complex enzyme and may conceivably play a role in patient variability in NRTI drug toxicities.

Martin et al. (Martin *et al.* 1994) reported that the approved NRTIs inhibited various host DNA polymerases. Behind the HIV Reverse Transcriptase, the highest affinity of the NRTIs was for polymerase γ. Moreover, many of the NRTI side-effects resemble symptoms of mitochondrial genetic disorders. All these facts implicated interaction with polymerase γ and subsequent depletion of mtDNA as a potential cause of NRTI toxicity giving rise to the polymerase γ hypothesis (Lewis *et al.* 2003).

Several groups have conducted experiments that demonstrate decreased mtDNA amounts in various tissue types of NRTI-treated HIV positive patients (Cherry *et al.* 2002; Cote *et al.* 2002; Haugaard *et al.* 2005; Walker *et al.* 2004). In addition, people observed mtDNA depletion in parallel with cell death, mitochondrial morphological changes, and increased lactate production in liver, heart, neuron, skeletal muscle, adipose, and blood cell cultures after incubation with different NRTIs (Azzam *et al.* 2006; Benbrik *et al.* 1997; Biesecker *et al.* 2003; Birkus *et al.* 2002; Cui *et al.* 1997; Galluzzi *et al.* 2005; Pan-Zhou *et al.* 2000; Setzer *et al.* 2005; Walker *et al.* 2002).

However, other groups published a substantial body of data that are not consistent with toxicity mechanisms resulting in depletion of mtDNA. Martin et al. (Martin *et al.* 1994) showed no association between inhibition of polymerase γ by NRTIs and mtDNA depletion. Mitochondrial dysfunction has been observed *in vitro* in mouse muscle, white adipose, brain, liver, and heart tissue (Note *et al.* 2003), hepatoma cell lines (Walker *et al.* 2002) as well as CD4 cells (Setzer *et al.* 2005) after incubation with NRTIs although no significant decrease in mtDNA amount was observed. Particularly, incubation of liver and skeletal muscle cells with ddC, ddI, d4T, and AZT show a higher rate of lactate production in the presence of AZT, but the least amount of mtDNA depletion (Birkus *et al.* 2002; Pan-Zhou *et al.* 2000). In clinical practices mtDNA depletion has been seen in parallel with normal cytochrome c oxidase activity, a sign of correct mitochondrial function (Piechota *et al.* 2006), and was not associated with lipoatrophy (McComsey *et al.* 2005) (although that study measured mtDNA depletion in blood samples, not fat cells). All these findings indicate a weak relationship between mtDNA copy number and nucleoside analog toxicity, which warrants a deeper look at the data concerning the interaction of different NRTIs with polymerase γ.

Lewis *et al.* suggested three possible polymerase γ dependent toxicity mechanisms that comprise the polymerase γ hypothesis, which are (1) direct inhibition of polymerase γ by NRTI-triphosphate without incorporation into the mtDNA, (2) chain termination of mtDNA replication following incorporation of the NRTI triphosphate, and (3) incorporation of the analog triphosphate into mtDNA without chain-termination allowing the NRTI to continue as a point mutation in mtDNA (Lewis *et al.* 2006).

To study the polymerase γ toxicity, we have simulated the DNA replication process of mitochondria to examine the impact of NRTIs. Using enzyme kinetics data gathered from Johnson et al. (Johnson *et al.* 2001), Feng et al. (Feng *et al.* 2004), and Hanes et al. (Hanes and

Johnson 2007; Hanes *et al.* 2007) we have carried out a series of simulations of mtDNA replication in the presence of various nucleoside analogs that interact with polymerase γ (Table I.6). These simulations bridge the gap between the basic enzyme kinetics data and the probability of failure of the mtDNA replication process.

Methods

The drugs included in this study

Thirteen analogs were used in the simulations (Table III.1). These included eight drugs of the NRTI class currently approved for human treatment- stavudine (d4T), lamivudine (3TC(-)), zidovudine (AZT), zalcitabine (ddC), didanosine (ddI) (whose active form is dideoxyadenosine (ddA) triphosphate) , abacavir (ABC), emtricitabine (FTC(-)) and tenofovir (TDF) (Antiretroviral_Guidelines_for_Adults_and_Adolescents_Panel 2006), and one anti-herpes drug, acyclovir (ACV). In addition we modeled the effects of the natural enantiomers of FTC(+) and 3TC(+) that have been used to explore a possible role of stereochemistry in the efficacy of strand termination (Feng and Anderson 1999), and ddI in its non-activated form. Since this study focuses on strand termination, we have not included FIAU, an anti-hepatitis B drug that tragically resulted in the deaths of five patients in phase 2 trials and whose toxicity is believed to be due to errors in mtDNA replication (Colacino 1996; Lewis *et al.* 1996), though not necessarily through strand termination (Johnson *et al.* 2001; Lewis *et al.* 1996).

Computational Model

Our computational model of the mitochondrial DNA replication process is based on the Stochastic Simulation Algorithm (Gillespie 1976; Gillespie 1977). The model is based on four reactions; DNA polymerase activity, exonuclease activity, disassociation of the polymerase from the DNA, and reassociation of the polymerase with the DNA molecule (Figure III.1). In the DNA polymerase reaction pol γ adds one nucleotide to the new DNA strand. This nucleotide may be the correct or incorrect (point mutation) base indicated by the template strand. In this model this includes the incorporation of nucleoside analog triphosphates. In the exonuclease reaction pol γ removes one nucleotide from the new DNA strand. This includes the removal of nucleoside analogs from the DNA strand. The exonuclease reaction is an error correction mechanism, as the rate for removal of incorrectly incorporated nucleotides is typically faster than that of correctly incorporated nucleotides. In the disassociation reaction the polymerase separates from the DNA molecule. In the reassociation reaction the polymerase re-attaches to the DNA molecule after disassociation. At each position on the replicating mtDNA strand, pol γ will randomly undergo one of the first three reactions (polymerase, exonuclease, or dissociation). Which reaction pol γ undergoes is determined by the probability of each reaction, calculated using the reactions rates and Michaelis-Menten kinetics.

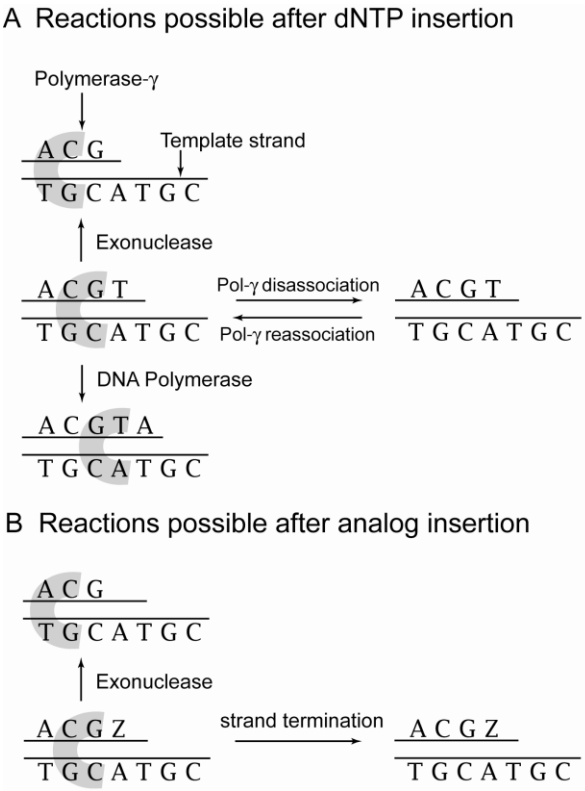

Figure III.1. Schematic diagrams of the polymerase γ reactions in this model. A. The four reactions possible following a correct incorporation. DNA Polymerase: polymerase γ adds one nucleotide to the new DNA strand. This nucleotide may be the correct or incorrect (point mutation) base pair for template strand, or a nucleotide analog. Exonuclease: polymerase γ removes one nucleotide (correct or incorrect match) from the new DNA strand. This is an error correction mechanism. Disassociation: The DNA polymerase can separate from the DNA molecule. Reassociation: The DNA polymerase re-attaches to the DNA molecule after disassociation. B. The possible reactions after insertion of analog (Z to represent AZT in this example): exonuclease activity or dissociation. Our model does not allow further polymerization or reassociation once an analog is inserted and not removed by an exonuclease reaction.

For the two scenarios of a correctly inserted and incorrectly inserted previous nucleotide we have separate sets of kinetic parameters for each of the pol γ reactions (Johnson and Johnson 2001; Johnson and Johnson 2001; Lee and Johnson 2006). These studies have reported an increase in exonuclease and disassociation rates, but a decrease in incorporation rates by pol γ following an incorrect incorporation. This is included in the simulation model by using two sets of enzyme kinetics parameters, one set for reactions following a correct

incorporation and another set for reactions following an incorrect incorporation. Kinetic parameters for the natural nucleotide (dNTP) interaction with pol γ are available in Tables I.1-I.5. As data regarding the reassociation reaction rate are not available our model assumes that after a disassociation event occurs the reassociation reaction follows, except in the special case discussed immediately below. Since the rate for the reassociation reaction is not available, the time required for that reaction is not calculated in this model. This approximation is not important to our results reported here which focus on strand termination probabilities.

Upon incorporation of an analog into the new DNA strand the next polymerase reaction is blocked. The exonuclease reaction can still occur, removing the analog molecule. However, if a disassociation reaction occurs before the analog can be removed, we assume that reassociation of the DNA polymerase is also blocked and the mtDNA replication event is disrupted resulting in strand termination (Figure III.1). There has been some speculation that the drugs, in particular AZT, may be inserted into a replicating mtDNA strand without causing strand termination. In this model we take the conservative assumption that all NRTIs that are inserted in the mtDNA strand and not subsequently removed cause strand termination.

Parameters included in the model for incorporation of each analog by pol γ were the concentration necessary for binding of 50% of available pol γ (K_m), the rate of polymerization (k_{pol}),, and the rate of excision (V_{exo}) of each analog by pol γ. The parameters k_{pol} and K_m were estimated from the maximum rate of incorporation by pol γ (k_{cat}) and the dissociation constant from pol γ (K_d), respectively, obtained under pre-steady state conditions (Feng *et al.* 2004; Hanes and Johnson 2007; Johnson *et al.* 2001). A recent publication shows that pyrophosphate release from AZT is uniquely slow during polymerization and that kinetics measured during steady-state conditions give a more accurate k_{pol} estimation (Hanes and Johnson 2007). These measurements were carried out on AZT due to the fact that under pre-steady state conditions a

decrease in incorporation rate was observed with increased AZT concentration indicating reversible binding. This pattern was not seen with any of the other analogs studied (d4T, 3TC(-), AZT, ddC, ddI, ddA, ABC, TDF, and 3TC(+)) and for this reason reanalysis of the enzyme kinetics for those drugs was not performed in that experiment. Given this continuing evolution in our understanding of the AZT kinetics we carried out two simulations for AZT insertion using the two available published sets of parameters determined under steady-state conditions in the 2007 paper by Hanes and Johnson (Hanes and Johnson 2007) and pre-steady state conditions published in the 2001 paper by Johnson et al. (Johnson *et al.* 2001). We distinguished the results using these two parameter sets as AZT_{2001} and AZT_{2007}. These parameter values, as well as those for the other analogs, are given in Table I.6.

Triphosphorylated mitochondrial natural nucleotide (dNTP) levels

The polymerization reaction rates are functions of the dNTP concentrations. For this calculation we consider three sets of dNTP concentrations, representing high, medium and low concentration conditions. Mitochondrial dNTP levels were estimated following the observations of Rampazzo et al and Ferraro et al (Ferraro *et al.* 2005; Rampazzo *et al.* 2004) (Table III.2). The units of picomole of mitochondrial dNTP per mg of mitochondria or picomoles per 10^6 cells were converted to µM by using an assumed mitochondrial volume of 0.2 femtoliters and density measurements from Pollak and Munn (Pollak and Munn 1970). It should be noted that these density measurements considered mitochondria as discrete entities not taking in to account any change in mitochondrial size due to organelle fission and fusion processes. We use these values only as estimates, in order to define the three categories of dNTP concentrations given below.

Table III.2. Mitochondrial dNTP concentrations and K_m values

dNTP level	dATP (μM)	dCTP (μM)	dGTP (μM)	dTTP (μM)
high	22.5	28	19.5	26
medium	1.675	1.644	0.47	0.76
low	0.1675	0.1644	0.047	0.076
K_m with polymerase γ*	0.8	0.9	0.8	0.6

*from reference (Lee and Johnson 2006).

High dNTP levels: As an estimate for the natural nucleotide concentrations within mitochondria of actively dividing cells, concentrations of natural nucleotides in mitochondrial pools from a cycling cell culture of 3T3-L1TK1+ (a mouse fibroblast line) were used. (Rampazzo *et al.* 2004). The units of dNTP measurement were converted to μM units by estimating the volume of a 3T3-L1 cell from images in Friis et al. (Friis *et al.* 2005).

Medium dNTP Levels: As an estimate of the natural nucleotide concentrations within mitochondria of resting or slowly dividing cells values for rat liver cells were used (Ferraro *et al.* 2006).

Low dNTP levels: A third set of simulations were carried out using natural nucleotide levels at 1/10th those estimated for the liver cells. This is meant to represent the low dNTP concentrations in postmitotic cells.

Simulation sets

Vertebrate mitochondrial DNA has a highly asymmetric G content. The low-G strand is labeled the light strand, with the complement strand called the heavy strand. Sets of simulations were carried out separately for the light strand sequence (NCBI, gi 17981852) and the heavy strand sequence. On each template three separate simulation sets were carried out using the

high, medium, or low natural nucleotide concentrations described above and varying the concentration of activated triphosphorylated analog. Each simulation was repeated 1000 times. The number of simulated mtDNA strand replications ending in a strand termination event (caused by a nucleoside analog incorporation) was recorded. The concentrations of the four dNTP pools and the activated analog were held constant throughout each simulated replication.

Results

The purpose of this study was to explore the plausibility of the pol γ hypothesis by calculating the probability of insertion of the nucleotide analog into the replicating mtDNA strand, leading to strand termination. We constructed a model based on the mitochondrial genome length and sequence, mitochondrial dNTP concentrations, and the measured enzyme kinetics of pol γ. With this model we simulated mitochondrial DNA replication in the presence of different analogs determining the dose response curves and the IC_{50} values of DNA termination for each drug. The simulation results were compared to reports of mitochondrial toxicity, and specifically to reports of mtDNA depletion.

Dose response curves and IC_{50} values

By measuring the probability of strand termination in the simulation as a function of the activated drug concentration, dose response curves for each drug were calculated. Figure III.2 shows the dose response curves obtained for the strand termination probability of each clinically approved analog as a function of the analog mitochondrial concentrations. The concentration at which these dose response curves passed 50% defined the IC_{50} values for each activated drug (Table III.3). In our model, replication was terminated once an analog was

inserted and failed to be removed by exonuclease activity. Based on these simulated IC_{50} values the list of analogs in the order of decreasing probability of mtDNA strand termination on the light strand was:

ddC = ddA = d4T > FTC(+) > 3TC(+) > ACV > ddI > 3TC(-) = TDF >= AZT_{2001} >> FTC(-) > ABC = AZT_{2007}, in which ">>" indicates a 10 fold difference or more, ">" indicates a 2 to 10 fold difference and "=" indicates a less than 2 fold difference. Note that ddA is the activated form of ddI. Of this list only d4T, ddI, ddC, 3TC(-), TDF, AZT, ABC, ACV and FTC(-) are approved for therapeutic use. The IC_{50} list showed that the "di-deoxy drugs", meaning ddC, d4T, and ddA, had the highest probability of causing mtDNA strand termination during replication while FTC(-), ABC and AZT_{2007} showed the least. Of those drugs approved for HIV treatment there was an observed difference of more than 800 fold between the di-deoxy drugs (ddC, ddA, and d4T) and other approved drugs (3TC, TDF, AZT, ABC, and FTC(-)) in the activated drug concentration necessary for 50% probability of mtDNA strand termination.

The only difference seen in the simulation of heavy strand replication (Table III.4) was that acyclovir had a slightly higher probability of termination than 3TC(+) and ABC had approximately equal probability of termination as FTC(-). Since there was little difference in the results for the two strands of the mtDNA molecule, we concentrated on results from the light strand. For the readers' convenience in interpreting these IC_{50} values, reported ranges of intracellular concentrations (Barry *et al.* 1996; Becher *et al.* 2004; Becher *et al.* 2002; Kewn *et al.* 2002; Moore *et al.* 1999; Piliero 2004; Pruvost *et al.* 2005) for activated nucleoside analog drugs measured in peripheral blood mononuclear cells in patients are given in Table III.5. Where necessary, values were converted to units of μM using the conversion of Kewn et al (Kewn *et al.* 2002). However it should be kept in mind that these concentrations are intracellular values, not the concentration values in the mitochondria which may be different.

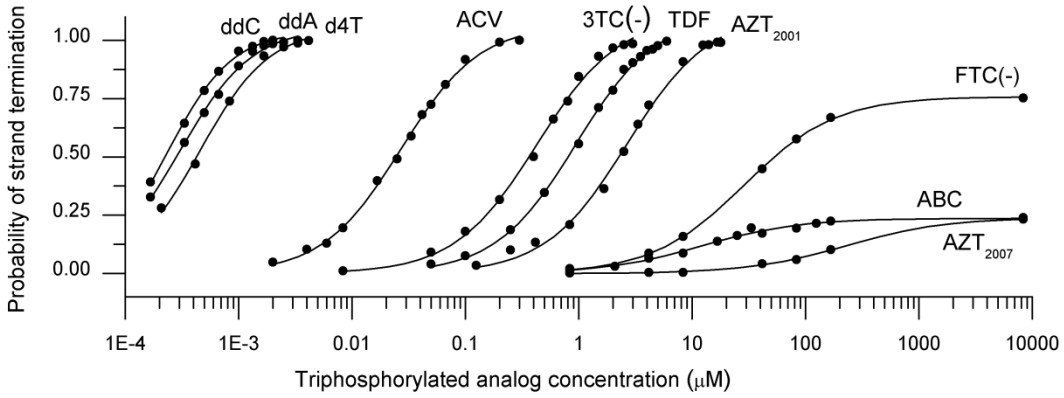

Figure III.2. The dose-response curves for incorporation probability of analogs approved for treatment. Circles are the probability of strand termination calculated from a set of 1000 simulations for each point and the curves are dose-response curves fit to the data points. There is a large difference of incorporation probability between the di-deoxy drugs (ddC, ddI, and d4T) and other approved anti-retrovirals (3TC, TDF, AZT, ABC, and FTC). The anti-herpes drug, acyclovir, falls in the middle of these two extremes, but shows little mitochondrial toxicity in clinical use possibly due to the fact it is dependent upon viral proteins for activation. AZT_{2001} probabilities were determined using kinetic parameters from reference (Johnson *et al.* 2001) and AZT_{2007} probabilities were determined using newly reported kinetic parameters from reference (Hanes and Johnson 2007).

Table III.3. IC_{50} values for light strand termination calculated from the simulation. Analogs are listed in order of increasing IC_{50}. IC_{50} values are calculated for three different sets of mitochondrial dNTP levels; high, medium and low, as defined in Table III.2. NA – not applicable.

Analog triphosphate	IC_{50} values (μM)			Reported mtDNA depletion
	High dNTP	Medium dNTP	Low dNTP	
ddC-TP	$3.42x10^{-4}$	$2.38x10^{-4}$	$5.80x10^{-5}$	Yes
ddA-TP	$4.42x10^{-4}$	$2.89x10^{-4}$	$7.47x10^{-5}$	Yes
d4T-TP	$7.92x10^{-4}$	$4.65x10^{-4}$	$9.76x10^{-5}$	Yes
FTC(+)-TP	$5.90x10^{-3}$	$4.20x10^{-3}$	$9.80x10^{-4}$	No data
3TC(+)-TP	$3.54x10^{-2}$	$2.26x10^{-2}$	$6.30x10^{-3}$	No data
Acyclovir-TP	$6.78x10^{-2}$	$2.58x10^{-2}$	$3.81x10^{-3}$	No data
ddI-TP	0.25	0.17	$4.60x10^{-2}$	No data
3TC(-)-TP	0.56	0.40	$8.40x10^{-2}$	No
TDF-TP	1.18	0.87	0.22	No
AZT_{2001}-TP	3.95	2.38	0.50	No
FTC(-)-TP	778.9	58.59	7.57	No
ABC-TP	NA	NA	8.10	No
AZT_{2007}-TP	NA	NA	NA	No

Table III.4. IC_{50} values (μM) for the strand termination on the mtDNA heavy strand. NA – not applicable, meaning that the probability of strand termination never exceeded 50%.

Analog triphosphate	IC_{50} values (μM)		
	High dNTP	Medium dNTP	Low dNTP
d4T-TP	$6.75x10^{-4}$	$3.60x10^{-4}$	$7.50x10^{-5}$
ddA-TP	$5.20x10^{-4}$	$4.00x10^{-4}$	$9.80x10^{-5}$
ddC-TP	$8.90x10^{-4}$	$5.80x10^{-4}$	$1.30x10^{-4}$
FIAU-TP	$1.44x10^{-3}$	$8.30x10^{-4}$	$1.60x10^{-4}$
FTC(+)-TP	$1.40x10^{-2}$	$1.00x10^{-2}$	$2.20x10^{-3}$
acyclovir-TP	$2.85x10^{-2}$	$1.14x10^{-2}$	$1.79x10^{-3}$
3TC(+)-TP	$8.8x10^{-2}$	$5.60x10^{-2}$	$1.36x10^{-2}$
ddI-TP	0.33	0.23	$5.75x10^{-2}$
3TC(-)-TP	1.50	1.00	0.19
TDF-TP	1.45	1.06	0.26
AZT_{2001}-TP	3.00	1.94	0.37
AZT_{2007}-TP	NA	NA	143
CBV-TP	NA	NA	2.22
FTC(-)-TP	NA	NA	30.9

Table III.5: Reported intracellular concentrations of the activated (triphosphate) form of the nucleoside and nucleotide analogs, measured in peripheral blood mononuclear cells in patients.

Analog triphosphate	Intracellular concentration range (mM)
3TC-TP	35.4 - 51.2
FTC(-)-TP	0.40 - 4.50
AZT-TP	0.84 - 1.2
ABC-TP	0.04 - 0.75
TDF-TP	0.27 - 0.39
ddA-TP	0.013 – 0.078
d4T-TP	0.016 – 0.082
Acyclovir-TP	Not available
ddC-TP	Not available

Abacavir, Emtricitabine, and Zidovudine$_{2007}$

For most analogs, the simulated dose response curve increases to 100% probability of strand termination if the analog concentration is raised high enough. AZT$_{2007}$, ABC and FTC(-) behaved differently from the other analogs in that they reached the point of saturation below 100% probability of strand termination (Figure III.2), and in some cases the strand termination probability saturated below 50%, meaning that no IC$_{50}$ values could be defined in those cases (the blank entries in Table III.3). These three analogs interact so poorly with pol γ that over the finite length of the mtDNA sequence (approximately 16,600 base pairs) these analogs have too few chances to incorporate into the growing mtDNA strand for the probability of strand termination to approach 100%, even with very large concentrations of the activated drug in the mitochondrion.

When the recently revised steady-state derived parameters for k_{cat} and K_m for AZT (Hanes and Johnson 2007) were used in the simulation, AZT$_{2007}$ did not reach a 50% probability of strand termination in the presence of normal to high dTTP levels, instead saturating at a 23% probability. This grouped AZT$_{2007}$ with ABC as having the least probability of causing termination of the replicating mtDNA.

Specificity constant

A common measurement for the relative likelihood of strand termination by each analog is the specificity constant (Johnson *et al.* 2001) determined by the ratio k_{cat}/K_m for the incorporation of an analog by pol γ. This is a common measurement used for predicting the discrimination of analogs by pol γ over the natural nucleotide substrate (Feng *et al.* 2004; Hanes and Johnson 2007; Hanes *et al.* 2007; Johnson *et al.* 2001). The drawback in taking this measurement of direct interaction with pol γ as a predictor for the actual incorporation into the

replicating mtDNA strand is that the specificity constant does not consider exonuclease activity, mitochondrial dNTP levels, nor strand length, all of which can affect the probability that an analog will be incorporated. All of these factors of the system are integrated into our computational model and the resulting IC_{50} values. Previous studies (Feng *et al.* 2004; Hanes and Johnson 2007; Hanes *et al.* 2007; Johnson *et al.* 2001) provide a list of increasing specificity constants of: ddC > ddA > d4T >> ACV > 3TC(-) > TDF > AZT >> ABC = FTC(-). The order of this list agreed quite well with our list given above based on simulated mtDNA strand termination. This agreement validates the use of k_{cat}/K_m values as an appropriate proxy for the relative probability of incorporation of these NRTIs by pol γ. A quantitative comparison between the specificity constant and our calculated IC_{50} for strand termination is given in Figure III.3.

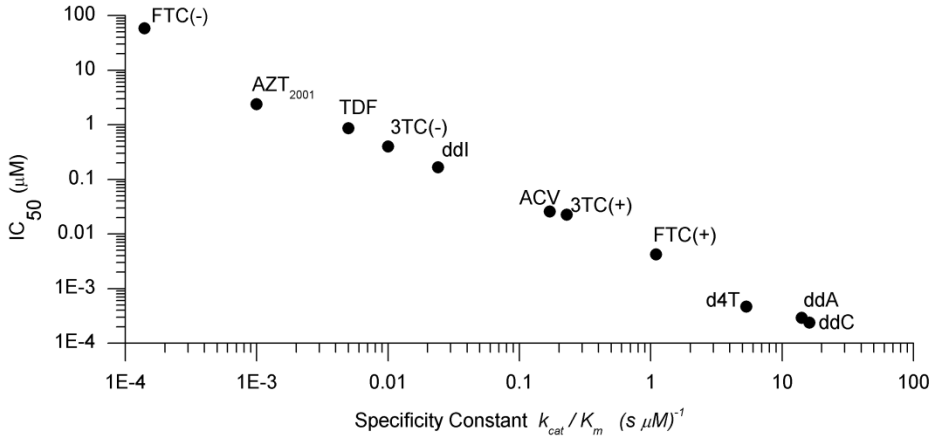

Figure III.3. The relationship between the IC_{50} values and the specificity constant, k_{cat}/K_m. The specificity constant is a measurement of direct interaction with polymerase γ often used as prediction for analog incorporation. The IC_{50} values are a more direct measure of incorporation probability taking exonuclease activity and other features into account. This relationship shows that the specificity constant is a useful proxy for incorporation probability.

Removal of nucleoside analogs from the mtDNA

The very low exonuclease reaction rate for each analog is the primary reason why the specificity constant serves as a reasonable prediction of mtDNA strand termination. The exonuclease reaction rates used in this study were taken from Johnson et al. (Johnson *et al.* 2001). Low excision rates for NRTIs have also been documented in the case of ddC (Longley and Mosbaugh 1991) and using yeast mtDNA polymerase with ddC and AZT (Eriksson *et al.* 1995). However, 3TC has a non-negligible measured exonuclease rates (Table I.6) (Johnson *et al.* 2001). Whenever an analog is inserted into the DNA strand, our pol γ model assumed that only the exonuclease and pol γ dissociation reactions can occur. Based on this model, in Table III.6 we give the predicted probability P_{exo} of the analog removal

$$P_{exo} = V_{exo} / (V_{exo} + V_{dis})$$

where V_{exo} is the rate of exonuclease reaction for the analog and V_{dis} is the rate of disassociation of the polymerase. To test these predictions, we carried out a set of simulations with the analog exonuclease reactions removed. The ratio of the IC_{50} value in the full model to the IC_{50} value in the exonuclease deficient model was in very good agreement with the $1\text{-}P_{exo}$ values (Table III.6). As predicted, only the two 3TC forms showed significant effects from the removal of the analog exonuclease reaction. Even in these cases the effect of the exonuclease reaction only shifted the IC_{50} value by a factor of 2 or less.

Table III.6. The effect of the exonuclease reaction for each nucleoside or nucleotide analog, in order of increasing V_{exo}.

Analog	V_{exo} (s^{-1})	P_{exo}	1-P_{exo}	IC_{50} / IC_{50}(exo-)
ddC	0.00002	0.00040	0.9996	1
d4T	0.0004	0.0079	0.9921	0.98
AZT_{2001}	0.0004	0.0079	0.9921	1.038
AZT_{2007}	Not reported	-	-	-
ddA	0.0005	0.01	0.99	1.086
ddI	0.0007	0.014	0.986	1.059
TDF	0.0007	0.014	0.986	0.942
ABC	0.0016	0.031	0.969	-
Acyclovir	0.0021	0.040	0.96	0.919
FTC(-)	0.0048	0.088	0.912	0.964
FTC(+)	0.0048	0.088	0.912	0.867
3TC(-)	0.015	0.23	0.77	0.725
3TC(+)	0.02	0.29	0.71	0.69

Effects of multiple nucleoside analogs

The current therapy for HIV infections involves a combination of nucleoside analog drugs, along with another class of drug such as a protease inhibitor. It has been reported that combining nucleoside analogs increases toxicity (Chan *et al.* 2007; Venhoff *et al.* 2007). The pol γ model is a series of reactions occurring as the DNA polymerase moves along the template strand. At each position on the DNA strand different nucleoside analogs would be able to be incorporated into the DNA strand. For example, AZT triphosphate molecules would only have a reasonable rate of incorporation opposite an A on the template strand, while 3TC triphosphate molecules would only have a reasonable rate of incorporation opposite a G on the template. Considering this, it is unlikely that there could be a combined effect of two analogs of different nucleosides on strand termination through the pol γ interaction alone. To test this, we modeled the effects of two analogs, AZT and 3TC, separately and in combination (Figure III.3). The combination of AZT and 3TC has been shown to have enhanced toxicity (Chan *et al.* 2007; Venhoff *et al.* 2007), though neither of these two studies found any significant mtDNA depletion

associated with this toxicity. If we define P_{AZT} as the probability of strand termination from a given concentration of AZT triphosphate and P_{3TC} as the probability from a given 3TC triphosphate concentration, then the combination of the two drugs should result in a strand termination probability of

$$P_{AZT+3TC} = 1-(1-P_{AZT})(1-P_{3TC})$$

This equation assumes there is no interaction between the two nucleoside analog drugs (this is known as the Webb fractional effect (Martinez-Irujo *et al.* 1998)). Note that $P_{AZT+3TC}$ is here defined as one minus the probability that neither AZT nor 3TC independently cause strand termination. A set of 1000 simulations was repeated 10 times, using the medium dNTP concentrations defined in Table III.2, and mean and standard deviations for the probabilities P_{AZT}, P_{3TC} and $P_{AZT+3TC}$ were measured. The results for $P_{AZT+3TC}$ were consistent with the probability expected assuming no interaction between the two drugs (Figure III.4). This indicates that any synergistic effects of multiple NRTIs on mitochondrial toxicity are not consequences of direct strand termination. Alternative explanations for synergistic effects may include competitive inhibition of deoxynucleotide phosphorylation, which is outside the limits of this computational model.

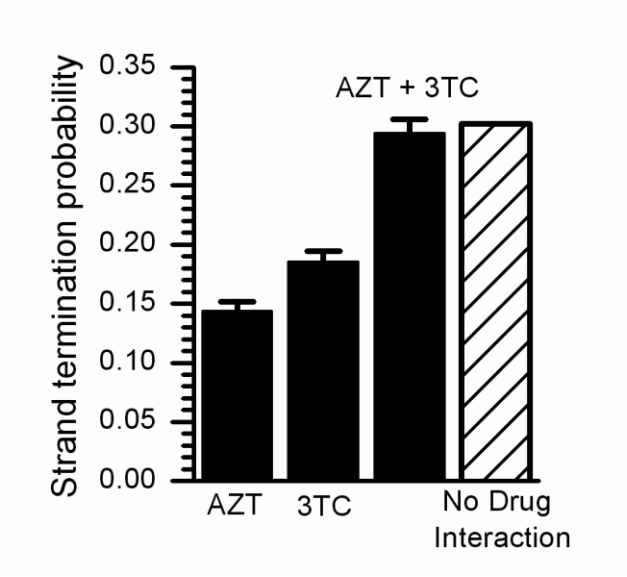

Figure III.4. The computed probability of strand termination for an AZT and 3TC combination. Computed probabilities are shown for AZT-TP alone (at 0.5 µM), 3TC-TP alone (at 0.1 µM) and the combination of both drugs at those concentrations (solid bars). The error bars represent standard deviations from 10 repeated sets of 1000 simulations. Also shown is the predicted probability of strand termination for the combined drugs (hatched bar) with the assumption of no interaction between the two drugs, calculated from the equation given in the text.

Discussion

Our simulated IC_{50} values of mtDNA strand termination for AZT-TP, CBV-TP (ABC), FTC(-)-TP and TDF-DP were approximately 1,000-fold or more higher than those of the di-deoxy drugs. This agrees with the lack of observed mtDNA depletion in liver, fat, and PBMC samples from patients on regimens comprised of NRTIs that are not di-deoxy analogs (Cherry *et al.* 2002; Cote *et al.* 2002; Walker *et al.* 2004). The herpes drug Acyclovir (ACV) fell between the di-deoxy drugs and the others in terms of probability of causing mtDNA strand termination. ACV is not associated with mitochondrial toxicity clinically. This is believed to be due to the fact that the drug must be activated by a viral-encoded kinase (Coen and Schaffer 1980). As herpes

replication tends to occur in waves with long latent periods, this may not lead to the long term effects on mitochondria seen with NRTI use for HIV treatment.

The dideoxy drugs

The dideoxy NRTIs (ddC, ddA, and d4T) showed the greatest risk of strand termination in our simulations, indicated by their predicted low IC_{50} values. This agrees with previous studies showing they are the NRTIs most associated with mtDNA depletion *in vitro* (Biesecker *et al.* 2003; Birkus *et al.* 2002; Cote *et al.* 2003; Cui *et al.* 1997; Pan-Zhou *et al.* 2000; Setzer *et al.* 2005) and mitochondria-related toxicities clinically placing them as alternative drugs for adults and adolescents in the federal guidelines (Antiretroviral_Guidelines_for_Adults_and_Adolescents_Panel 2006). In both muscle and subcutaneous fat biopsies of HIV+ patients, mtDNA levels were significantly lower in those on dideoxy drug regimens as opposed to those on non-dideoxy NRTI regimens (Cherry *et al.* 2002; Walker *et al.* 2004). Even though the toxic side effects of dideoxy drugs are well known and the *in vitro* effects on tissue mtDNA levels of these drugs are in agreement with our simulation results, the very low IC_{50} values for these drugs of approximately 3×10^{-4} μM warrant discussion. The low IC_{50} value for ddC is in agreement with findings that this drug is not readily metabolized in the cell to its active form (Cui *et al.* 1997; Piliero 2004), implying that the concentrations of the activated drug in the cell may be quite small. There is evidence, however, that d4T and ddI are activated to a significant degree as the concentration of their triphosphorylated forms in patient peripheral blood mononuclear cells are above the predicted IC_{50} values by approximately 100-fold (Becher *et al.* 2004; Becher *et al.* 2002) (Table III.5, note that the experimentally measured value is the activated drug concentration in the cytoplasm, not in the mitochondria). Given that d4T and ddI are still recommended drugs for HAART

(Antiretroviral_Guidelines_for_Adults_and_Adolescents_Panel 2006) they are obviously tolerable to a large number of patients who do not experience the serious side-effects of lactic acidosis and neuropathy. One plausible explanation for this tolerance in the face of the striking affinity of these drugs for pol γ is that there exists a significant barrier to dideoxy drug entry into the mitochondrion or drug activation within the mitochondrion allowing activated dideoxy drug mitochondrial concentrations to remain low in the majority of patients treated with these drugs. We know of no reports of measured levels of the triphosphate form of these dideoxy drugs within mitochondria.

AZT

The experimental data indicates that AZT interacts poorly with pol γ as shown by the high K_m and low k_{cat} values (Table I.6) for this drug. An explanation for the slow rate of incorporation of AZT was recently published (Hanes and Johnson 2007). AZT demonstrates unusually slow pyrophosphate release upon incorporation by pol γ rendering polymerization readily reversible even upon binding to the template:primer molecule. In the cases of natural nucleotides this subreaction is fast enough to be considered negligible indicating that pre-steady state k_{pol} values are a good approximation for k_{cat}. Yet, in the case of AZT, this slow pyrophosphate release rate has a significant effect on k_{cat} so that a measurement of k_{pol} during steady-state conditions is more appropriate for estimating k_{cat}. The k_{pol} and K_d determined in the recent Hanes and Johnson study (Hanes and Johnson 2007) with steady-state conditions, that theoretically take the slow pyrophosphate release rate in to account, indicate a k_{cat} 100-fold lower than that determined from the k_{pol} calculated under pre-steady state conditions (Johnson *et al.* 2001). We carried out the simulation for both sets of parameters separately and in both cases AZT shows a poor probability of mtDNA strand termination (Figure III.2). The IC_{50} values

generated by this study show that AZT should not be toxic through mtDNA strand termination as it has a higher IC_{50} value than 3TC(-) and TDF, neither of which are associated with mitochondrial toxicity.

The low probability of strand termination by AZT is supported by the fact that although AZT has consistently been associated with positive markers for mitochondrial toxicity, substantial evidence exists that the extent of AZT-induced mitochondrial toxicity is disproportional to the amount of mtDNA depletion it causes. The analogs ddC, d4T and ddI (activated to ddA) cause significantly more mtDNA depletion and decreased protein subunit expression of various electron transfer chain proteins with essential subunits encoded in the mtDNA, as would be expected from their increased interaction with pol γ compared to other analogs (Figure III.2). Yet AZT still manages to demonstrate a cytotoxicity that is equal to or greater than ddA, ddC, and d4T at comparable concentrations in various studies. In human liver and cardiac muscle cells incubation with AZT lead to cytotoxicity and increased lactate levels with no sign of mtDNA depletion (Cihlar *et al.* 2002; Lund *et al.* 2007; Pan-Zhou *et al.* 2000). Similar results are seen in blood cells and adipose cells (Mallon *et al.* 2005; Setzer *et al.* 2005; Stankov *et al.* 2007). Szabados et al. (Szabados *et al.* 1999) showed significant toxic effects on cardiac muscle cells including increased ROS, abnormal mitochondrial structure, and decreased ATP/ADP ratio after two weeks of exposure of cells in medium with no effects on mtDNA levels. In fact, AZT is actually associated with slight increase in mtDNA levels in cell culture (de Baar *et al.* 2007; Hobbs *et al.* 1995), PBMCs (Cote *et al.* 2002), and liver tissue samples (Walker *et al.* 2004). Our model, however, does not address subreactions that influence pol γ binding of the analog, meaning our results cannot disprove the possibility that AZT toxicity is due to deactivating pol γ either through irreversible binding or induction of a conformational change in the enzyme. However, the high K_d determined by both Hanes and Johnson and Johnson et al.

(Hanes and Johnson 2007; Johnson *et al.* 2001), along with the cited studies showing toxicity independent of mtDNA depletion, make this an improbable mode of toxicity. It is our conclusion that based on the measured kinetic coefficients of AZT with pol γ that AZT toxicity is not dependent upon mtDNA strand termination. Indeed, various pol γ independent hypotheses have been proposed for AZT mitochondrial toxicity. These include inhibition of the enzymes of the mitochondrial salvage pathway causing nucleotide pool imbalances (Lynx *et al.* 2006), binding to ADP-ATP translocator (Valenti *et al.* 2000), and direct inhibition of components of the electron transport chain (Pereira *et al.* 1998).

TDF

Tenofovir is associated with renal dysfunction without significant mtDNA depletion (Karras *et al.* 2003; Zimmermann *et al.* 2006). In a retrospective study of HIV positive patients taking TDF and those not taking TDF, no significant differences in mtDNA levels of kidney biopsies were observed (Cote *et al.* 2006). Similarly, in human renal proximal tubule cells (Vidal *et al.* 2006), TDF was not associated with cytotoxicity, mtDNA depletion, or COII mRNA depletion. In our simulations mitochondrial TDF triphosphate IC_{50} values were in the range 0.2 to 1.2 μM, depending on the natural dNTP levels. Since these concentrations are not unusually high, our conclusion is that Tenofovir might be able to cause some moderate mtDNA depletion, depending on how well the activated drug is concentrated within the mitochondrion.

Conclusions

A number of hypothesis with supporting evidence have been proposed for NRTI toxicity experienced during HAART. Possible pol γ mediated pathways include the direct inhibition of pol

γ by NRTI-triphosphate without incorporation of the analog; chain termination by incorporation of NRTI triphosphate into mtDNA; and incorporation without chain termination of the analog-triphosphate allowing it to remain as a point mutation in mtDNA. Our model only addresses the case of chain termination. There is not enough data on the subreactions that comprise analog binding to pol γ for this model to explore the possibility that some analogs cause toxicity through inhibition of the pol γ enzyme directly, either by irreversible binding or induction of conformational change, as opposed to strand termination.

The specificity constant, k_{cat}/K_m, (Johnson *et al.* 2001) is commonly used as an approximate indicator of mitochondrial toxicity through strand termination of mtDNA. Before this model, this has been a bit of a leap as the specificity constant does not take genome length, exonuclease activity, nor dNTP concentration into account, and no direct predictions or measurements of strand termination probabilities have previously been given. We fill this gap in our understanding by providing a model that includes all of these factors and that predicts strand termination probabilities. The consistence between our simulation model results and the qualitative ordered list of NRTI drug toxicity based on the specificity constant is a validation of the model results. However, the simulation model goes far beyond the specificity constant by predicting IC50 values and quantitative dose-response curves (Fig. 2) for these drugs. Furthermore, the specific definition of strand termination used in this model raises the hypothesis that dissociation of Polymerase γ after an NRTI is incorporated into the mtDNA strand is a critical step in strand termination. In this particular model we chose to define strand termination as the dissociation of Polymerase-γ after the incorporation of an NRTI, under the assumption that re-association of the polymerase after the NRTI could not occur. Of course, it is possible that in-vivo there may be other currently unknown factors which may alter the polymerase γ dissociation kinetics (or any other kinetics for that matter) from the measured

values. If our assumption that pol γ re-association is blocked after NRTI incorporation was changed, and re-association of the polymerase was to be allowed, then more exonuclease events of the NRTIs would occur. However, it is not clear to us then what the definition of "strand termination" would be since the exonuclease activity would eventually remove all incorporated NRTIs given enough time. Johnson et al (Johnson *et al.* 2001) used that assumption, where all NRTIs incorporated into the mtDNA were eventually removed by exonuclease activity, to define a toxicity index based on a calculation of the amount of additional time required for these NRTI exonuclease events. Based on this definition of a toxicity index, Johnson et al (Johnson *et al.* 2001) also defined an ordered list of NRTI drug toxicity which was similar to our list and similar to the lists based on the specificity constants. An important use of any computational model is to raise questions for further experimental study. This simulation raises the following questions. Is strand termination defined by the dissociation of the polymerase after insertion of an activated NRTI? If not, what is the proper definition of strand termination?

NRTI toxicity appears perplexingly specific to cell type (Antiretroviral_Guidelines_for_Adults_and_Adolescents_Panel 2006) and the mechanism for this tissue specificity is currently unclear. As natural nucleotide concentrations within the mitochondrion can differ greatly across cell types, we sought to observe how incorporation of NRTIs may differ in the presence of varying mitochondrial dNTP levels, which we broke down into three sample categories; high dNTP levels, medium dNTP levels and low dNTP levels (Table III.2). Although the IC_{50} values for strand termination measured in this simulation did depend on the concentrations of the natural nucleotide triphosphates, the relative ordering of the nucleoside analogs by IC_{50} was the same for all three dNTP conditions (Table III.3). The simulations made with low dNTP concentrations, representing post-mitotic tissues, did have lower IC_{50} values, consistent with a greater sensitivity of these tissues to damage by nucleoside

analog drugs. However, these results would not explain why some tissues are susceptible to toxicity from a particular analog. We have limited this simulation model to the activity of the mitochondrial DNA polymerase acting on the tri-phosphate form of the four natural deoxyribonucleosides and the tri-phosphate form of the drugs. Since the tissue dependence of the toxicity of the drugs was not reproduced in this model, this implies that the source of this tissue dependence lies outside the bounds of this particular model. This includes the possible interference of the various phosphate states of the drugs with the metabolism within the mitochondria that produces the natural deoxyribonucleoside tri-phosphates, potentially altering the relative levels of the four natural dNTPs.

Although mitochondrial toxicity from NRTIs is common, the more severe forms of this toxicity are certainly not universal. Current research is revealing that the gene for polymerase γ is the site of a large number of mutations and polymorphisms that alter its enzyme kinetics and function (Graziewicz *et al.* 2007; Graziewicz *et al.* 2006; Horvath *et al.* 2006; Hudson and Chinnery 2006). The natural variability in this crucial gene may be an important source of the individual variation in the susceptibility of patients to this toxicity, and perhaps to the phenotypic variation which occurs. Although the interaction of nucleoside analogs with polymerase γ has been recognized for almost 15 years now (Martin *et al.* 1994), we still know surprisingly little about the levels of activated drugs within mitochondria (Lynx *et al.* 2006) or about the transport mechanism by which these drugs enter the mitochondrion (Kang and Samuels 2008; Lindhurst *et al.* 2006).

CHAPTER IV

ENRICHING TARGETED SEQUENCING EXPERIMENTS FOR RARE DISEASE ALLELES[4]

Introduction

Genome-wide association studies (GWAS) are based on the premise that densely genotyped common alleles will have statistical power to detect causal associations with traits at nearby, ungenotyped common mutations through short-range linkage disequilibrium (LD). The basis for this strategy is the common disease common variant (CDCV) hypothesis (Reich and Lander 2001). This approach has been proven to be effective in many scenarios for mapping small genomic regions to traits (see the National Human Genome Research Institute Catalog of Published Genome-Wide Association Studies) (Manolio *et al.* 2008; McCarthy *et al.* 2008). However, the predominantly small effect sizes encountered thus far in investigations of most traits have provided no explanation for a large proportion of the trait variance attributable to heritable factors (Maher 2008; Manolio *et al.* 2009). Some effort to describe this phenomenon has suggested that hundreds or thousands of SNPs may each have a very subtle influence on the risk of some psychiatric traits (Purcell *et al.* 2009). These observations seem to support an adjustment of the CDCV model to allow for the possibility that rare alleles might also exert a major influence on common traits (Bodmer and Bonilla 2008; Pritchard 2001; Schork *et al.* 2009). This modification of CDCV, known as common disease rare variant (CDRV), postulates that

---

[4] Edwards, T. L., Song, Z., and Li, C. (2011). Enriching targeted sequencing experiments for rare disease alleles. *Bioinformatics* 27(15), 2112-2118. Author contributions: Conceived and designed the experiments: TLE CL. Performed the experiments: ZS. Analyzed the data: ZS TLE. Contributed reagents/materials/analysis tools: TLE CL. Wrote the paper: TLE CL. Algorithm design: TLE CL Software: ZS.

alleles with strong effects on traits are likely to be rare due to purifying selective pressure and recent time to coalescence due to the rapid expansion of human populations (Pritchard 2001). Additionally, it has been shown in simulations that multiple rare alleles with strong effects can stochastically aggregate onto the haplotypic background of a common allele and produce genome-wide significant association signals, a scenario termed synthetic association (Dickson *et al.* 2010). Further support for the CDRV hypothesis comes from observational studies where the average allele frequency of SNPs with predicted effects on proteins was smaller than the average allele frequency of intronic or synonymous variants (Cargill *et al.* 1999; Gorlov *et al.* 2008; Wong *et al.* 2003). Estimates from human Mendelian traits, human–chimpanzee divergence data and human genetic variation suggested that ~53% of new missense mutations have mildly deleterious effects, and that up to 70% of low-frequency missense alleles are mildly deleterious (Kryukov *et al.* 2007).

A rare trait allele may not be annotated in the databases of common variants maintained by the International HapMap Organization or dbSNP, thereby excluding the possibility of detecting that SNP through imputation and subsequent association analysis. The constellation of causal alleles may also be unique for each population of human subjects, where sensitive functional gene or regulatory regions are perturbed by independent sets of rare mutations that occurred after geographic or cultural barriers led to increased genetic distance (Tishkoff *et al.* 2009). Thus, the same associated allele from GWAS across multiple ethnic groups does not necessarily imply the same underlying architecture of causal alleles in LD. Furthermore, our simulations suggest that rare disease-causing variants may not be captured at all by the modest samples of each population isolate from the 1000 Genomes Project, regardless of the high error rates for rare genotype calls from that study due to low coverage. Resequencing is

then the best available means of discovering these rare SNPs in a GWAS sample and ultimately detecting the relationship between these alleles and traits.

To successfully discover the mutations that determine trait susceptibility, detailed assays that directly capture all genetic variation in a region are required (Cirulli and Goldstein 2010). This can be accomplished most efficiently using next-generation sequencing technology to resequence subjects for the implicated loci (Service 2006). While next-generation sequencing technologies have substantially decreased the financial cost of resequencing large genomic regions relative to Sanger sequencing technology, it is still not generally feasible to resequence all the subjects that were used to isolate a genomic region via GWAS. Thereby, some strategy is necessary for employing sequencing technology that is cost-effective. One possibility is to resequence a small number of cases and controls or persons with extreme trait values and evaluate the observed genetic variation for association with traits to screen rare variants prior to larger genotyping experiments; however, this approach will suffer from low statistical power at the screening step due to the infrequent exposure rate of rare alleles, and potentially suffer from inflated type I error rates (Li and Leal 2009) as a result of ascertainment bias in cases. An alternative approach is not to attempt to associate alleles from resequencing data with the trait, but to discover rare alleles by resequencing, and then assay these SNPs with conventional genotyping methods in the entire available pool of study subjects. The effectiveness of this approach will be limited by the power to capture rare alleles in the targeted loci, which is directly related to the selection of subjects for the resequencing experiment (Li and Leal 2009). For SNP discovery, targeted resequencing study designs can be tailored for efficient capture of rare disease alleles in small samples, by using the information available at nearby trait-associated SNPs.

In this chapter, we present SampleSeq, an algorithm for enriching the yield of rare or uncommon disease alleles in a sample of unrelated study subjects by choosing subjects according to their observed associated alleles and trait information. When multiple regions are to be sequenced, SampleSeq selects subjects with a balanced representation of all the regions. SampleSeq can also estimate the sample size required to detect a hypothetical disease allele, and thus can optimize a resequencing study to preserve resources for subsequent genotyping or other investigations.

Methods

We first describe our method for selecting subjects for sequencing a single region. We then extend the method to sequencing multiple regions. Finally, we describe simulation strategies for evaluating our method.

Sequencing a single region

Let *A* be a disease-associated common SNP, with alleles *A* and *a* and allele frequencies p_A and p_a, respectively. Let *D* be the true disease SNP close to SNP A, with alleles *D* and *d* and allele frequencies p_D and p_d, respectively. SNP *D* may not have been genotyped in previous stages of the investigation, and the common SNP *A* serves as a proxy for SNP *D*. The assumption of a single disease SNP simplifies the derivation, but it does not appear to be necessary as will be shown in our simulation results. As our method seeks to calculate the expected count of disease variants for each subject by conditioning on his SNP *A* genotype and affection status, we assume genotypes at SNP *A* are available for all subjects, and further that resequencing will be performed at sufficient depth to accurately call rare genotypes in small sample sizes. Suppose

allele a is the 'risk' allele, in positive LD with allele d and is either the major or minor allele at SNP A, and allele d is the real disease variant. When d is a rare variant, it is reasonable to assume that it originated on the background of allele a and almost no recombination has since occurred between them; we describe the rationale for this assumption in section Discussion. Then $p_d<p_a$, and the four haplotype frequencies are p_{dA}, $p_{da}=p_d-p_{dA}$, $p_{Da}=p_a-p_d+p_{dA}$ and $p_{DA}=p_A-p_{dA}$. Since almost no recombination has occurred between the two loci, it is reasonable to assume p_{dA} is much smaller than p_d, otherwise the LD between the two loci would be too weak to make SNP A a good proxy and be identified in a GWAS. When $p_{dA}\approx 0$, we have $p_{da}\approx p_d$, $p_{Da}\approx p_a-p_d$ and $p_{DA}\approx p_A$. This assumption is not required for the calculations below, although it is implemented in the current version of our software for ease of computation. Our simulations did not have this requirement either (see section Simulation strategy). Let G_a and G_d be the genotypes at the loci: G_a=0,1,2 for *AA*, *Aa*, *aa* and G_d=0,1,2 for *DD*, *Dd*, *dd*, respectively. We assume Hardy–Weinberg equilibrium (HWE) at the SNPs in the population. Let Y be the disease status, 1 for cases and 0 for controls. Let $f_i=P(Y=1|G_d=i)$ (i=0,1,2) be the penetrances for genotypes *DD*,*Dd* and *dd*, respectively, and K be the disease prevalence in the population.

Our goal is to calculate the expected count of allele d, $E(G_d|G_a,Y)$, given each subject's genotype at SNP A and affection status, and select subjects accordingly. To achieve this, we first calculate $P(G_d=g|G_a,Y)$ for g=0,1,2. Note that

$$P(G_d \mid G_a, Y) = \frac{P(G_d, G_a, Y)}{P(G_a, Y)}$$

where the denominator is $P(G_a,Y)=\sum G_d P(G_d,G_a,Y)$ and the numerator is $P(G_d,G_a,Y)=P(G_d,G_a)P(Y|G_d,G_a)=P(G_d,G_a)P(Y|G_d)$. The genotype probability$P(G_d,G_a)$ is a function of haplotype frequencies under HWE. The probability$P(Y|G_d=i)=f_i$ when Y=1, and $1-f_i$ when Y=0. We now show how to obtain f_i. Note that

$$P(Y=1,G_a=i)=P(G_a=i\,|\,Y=1)P(Y=1)=P(G_a=i\,|\,Y=1)K$$

where $P(G_a=i|Y=1)$ is the case frequency for genotype $G_a=i$. Since

$$\begin{aligned}P(Y=1,G_a=i)&=\sum_g P(Y=1,G_a=i,G_d=g)\\&=\sum_g P(Y=1\,|\,G_d=g)P(G_a=i,G_d=g)\\&=\sum_g f_g P(G_a=i,G_d=g)\end{aligned}$$

we have

$$\begin{cases}P(G_a=0\,|\,Y=1)K= & p_{DA}^2 f_0+ & 2p_{DA}p_{dA}f_1+ & p_{dA}^2 f_2\\ P(G_a=1\,|\,Y=1)K= & 2p_{DA}p_{Da}f_0+ & (2p_{DA}p_{Da}+2p_{Da}p_{dA})f_1+ & 2p_{dA}p_{da}f_2\\ P(G_a=2\,|\,Y=1)K= & p_{Da}^2 f_0+ & 2p_{Da}p_{da}f_1+ & p_{da}^2 f_2\end{cases}$$

and can solve for f_0, f_1, f_2 using these linear equations.

In the above calculation, the case genotype frequencies $P(G_a=i|Y=1)$ can be estimated from the data at hand. The haplotype frequencies depend on the allele frequencies at SNPs *A* and *D*. The SNP A allele frequencies *pA* and *pa* can be estimated as weighted averages of case and control allele frequencies; for example, $\hat{p}_A=K\hat{p}_{A,case}+(1-K)\hat{p}_{A,control}$, where $\hat{p}_{A,case}$ and $\hat{p}_{A,control}$ are the frequencies of allele *A* in the cases and controls, respectively. When the disease prevalence is very low, $\hat{p}_A\approx\hat{p}_{A,control}$. The investigator needs to specify p_d, for which we will show that often a range is sufficient. We also need the information on disease prevalence *K*, which often is available from external sources and also can be specified as a range.

Once we have calculated $P(G_d|G_a,Y)$, the expected count of allele *d* can be easily calculated as $E(G_d\,|\,G_a,Y)=\sum_g E(G_d=g\,|\,G_a,Y)g=P(G_d=1|G_a,Y)+2P(G_d=2\,|\,G_a,Y)$.

If we focus on a single region, then the subjects can be ranked according to their expected count of allele *d*. The top ranked subjects can be selected for sequencing to ensure the

highest chance of detecting rare disease variants. We will discuss stopping criteria and sample size determination at the end of the next section.

Sequencing multiple regions

In practice, investigators may want to fine map multiple regions simultaneously. Our method can be extended for this scenario. We assume there are *M* regions to sequence and they are unlinked to each other. For subject *i* and region *j* (*i*=1,…,*n* and *j*=1,…,*M*), let G_{ijd} and G_{ija} be the genotypes at the real disease SNP and the reported associated common SNP, respectively. Because the regions are unlinked, the above calculations can be carried out separately for each region, with $E_{ij}=E(G_{ijd}|G_{ija},Y_i)$. One might want to rank the subjects according to the expected number of disease variants over all regions, $E_i=\sum_{k=1}^{M}E_{ik}$ and select top ranked subjects. However, as the regions can differ in key characteristics such as risk allele frequency and strength of disease association, the top ranked subjects may contribute unevenly to the regions. As a result, this selection strategy may lead to overrepresentation of one region and lack of representation for another. A more efficient procedure is to select the top ranked subjects one at a time, each time tallying the cumulative expected count of disease variants for each region, denoted by C_j for region *j*. Once a region *j* has reached $C_j \geq c$, a prespecified target number of disease variants, we re-rank the remaining subjects based on $\sum_{k:C_k<c}E_{ik}$, calculated by excluding the region, and continue to select top ranked subjects. This process is repeated every time a region reaches $C \geq c$.

We may stop the process when all regions have reached $C \geq c$. The number of selected subjects is the sample size needed to have $C \geq c$ for all regions. If the number of selected subjects is fewer than planned, the resources could be preserved for subsequent follow-up. If the

investigator wants to select more subjects, he may either raise the target value *c* and redo the selection or continue selecting from the remaining subjects according to their E_i. Although this algorithm allows investigators to determine the sample size needed to reach $C \geq c$ for all regions, as the disease variant frequency *pd* that is used in calculation of E_{ik} may be different than the real disease variant frequency, the target value *c* may be far from the true number of disease variants in the selected subjects, as will be seen in our simulation results. However, our simulations also showed that even when p_d was misspecified, SampleSeq performed well compared to the alternative approaches we simulated.

Missing and imputed genotypes

In practice, missing genotypes exist due to various reasons. In SampleSeq, when genotype G_{ija} is unavailable, E_{ij} is calculated as a weighted average

$$Ei_j = \sum_g E(G_{ijd} \mid G_{ija} = g, Y_i) P(G_{ija} = g \mid Y_i),$$

where $P(G_{ija}=g|Y_i)$ is the estimated genotype frequency of *g* in cases or controls, depending on the value of Y_i. Similarly, a missing genotype may be imputed from the haplotype distribution of the population and observed haplotypes in the study subjects. When genotype G_{ija} is imputed, E_{ij} can be calculated as a weighted average using the posterior probabilities of the imputed SNP as weights:

$$Ei_j = \sum_g E(G_{ijd} \mid G_{ija} = g, Y_i) P(G_{ija} = g \mid G_F),$$

where *GF* denotes flanking marker genotypes.

Simulation strategy

We simulated case–control data with one or multiple disease regions, each harboring one or multiple disease variants with additive effects on trait risk. For rare variants, additive effect is practically equivalent to dominant effect as there are mostly only two genotypes, *DD* and *Dd*. To simulate realistic sequence-level genetic data from human populations, we employed the coalescent simulation software cosi, with parameters developed to calibrate the LD profile of simulated data to the observed LD profile from human populations (Schaffner *et al.* 2005). Additionally, we used the recombination map from the International HapMap Project to model the probability of recombination in specified genomic regions. We randomly chose five disease regions between 125 kb and 250 kb in length, and for each region, we used cosi to generate a pool of 25 000 haplotypes. It is possible that associations between rare variants and common proxies might extend over longer physical distances than 250 kb; however, these simulations were computationally intensive to perform on a large scale. We note that there is no size limitation for our method and software.

We simulated three scenarios: (i) CDRV with one rare disease variant per region; (ii) synthetic association (Dickson *et al.* 2010) with 10 rare disease variants per region; and (iii) CDCV with one common disease variant per region. For the CDRV scenario, we simulated various settings of prevalence (K=0.01, 0.05, 0.1, 0.2) and disease variant minor allele frequencies (MAFs, range 0.0025–0.01, denoted as MAF_{min} and MAF_{max}), with odds ratios (ORs) in the range 2–6 (denoted by OR_{min} and OR_{max}). The OR of a disease variant was determined according to its MAF through the following formula:

$$OR_i = OR_{min} + \left(1 - \frac{MAF_i - MAF_{min}}{MAF_{max} - MAF_{min}}\right)(OR_{max} - OR_{min}).$$

For the synthetic association scenario, we simulated one level of prevalence (*K*=0.01), and placed 10 random rare disease alleles with MAF in the range 0.0025–0.01 in each of five independent genomic regions, with OR in the range 2–6. For the CDCV scenario, we simulated one prevalence (*K*=0.01) with OR in the range 1.1–1.5, and various disease allele frequencies (0.01, 0.05, 0.15, 0.25).

For each combination of prevalence and ORs, a disease model was established with

$$P(Y=1|G_1,G_2,...,G_M)=\frac{e^{\beta_0+\beta_1G_1+\beta_2G_2+...+\beta_MG_M}}{1+e^{\beta_0+\beta_1G_1+\beta_2G_2+...+\beta_MG_M}},$$

where (G_1,G_2,…,G_M) is an individual's joint genotype at the *M* disease SNPs, and the coefficients β_j were determined based on the prevalence and ORs.

To simulate a control, a pair of haplotypes from each region was randomly drawn, and a random number in (0, 1) was drawn and compared to the penetrance from the disease model to determine if the subject is a control. To simulate cases, the probability of a having a multilocus genotype conditional on being a case was calculated for all possible genotypes from the *M* disease SNPs using the equation:

$$P(G_1,G_2,...,G_M|Y=1)=\frac{P(Y=1|G_1,G_2,...,G_M)P(G_1,G_2,...,G_M)}{P(Y=1)}.$$

A multilocus genotype across all disease SNPs was then randomly selected according to this conditional distribution, and haplotypes consistent with that genotype were randomly chosen from the haplotype pools simulated by cosi.

For each scenario and each prevalence level, we generated 100 replicates of 2000 cases and 2000 controls. We then identified a proxy marker (i.e. SNP *A*) for each region with 1.1<OR<1.5 and 0.2<MAF<0.4. These criteria were chosen to emulate typical associations from GWAS, and to allow the association between disease and SNP *A* to arise naturally as a result of LD between SNPs *A* and *D*, which were calibrated to resemble the LD profile of European-

ancestry populations. On average the D′ between SNPs *A* and *D* was 0.88, with a range of 0.8–1 across all CDRV simulations; in other words, p_{dA} could be non-zero in our simulated data. For the synthetic association scenario, we did not impose any restrictions on the relationship between SNP *A* and the real trait SNPs, so that some of the risk alleles might fall on the low-risk background of SNP *A*.

In addition to SampleSeq, we also considered other approaches to selecting the same number of subjects, including (i) random selection of controls; (ii) random selection of cases; (iii) selection of subjects ranked by dosage of proxy marker risk alleles; and (iv) selection of cases ranked by dosage of proxy marker risk alleles. For all these approaches, we counted the total number of disease variants per region that were captured in the selected subjects for sample sizes from 50 to 500 in increments of 50 subjects. For the simulated data, we also counted the maximum number of disease variants that can be carried for each given sample size.

Results

For all simulated scenarios, we calculated the number of rare disease alleles captured. For the CDRV scenarios, where the allele frequency of the trait locus is well-estimated, the SampleSeq algorithm consistently provided higher yields of captured disease alleles than the other methods for all sample size thresholds (Figures IV.1-IV.4, Tables IV.1-IV.4). These results demonstrate the benefit in efficiency that SampleSeq can provide over the other alternatives. The yield of rare disease alleles provided by SampleSeq is a little higher over all sample sizes than by ranking case subjects by their burden of risk alleles. The other three alternative approaches were less efficient than SampleSeq by large percentages. Among the four alternative approaches, those relying on the burden of the proxy marker risk alleles were better

than those not using this information, and those focusing on cases were better than those not limited to cases. These results are as expected as both the burden of marker risk alleles and disease status are informative for the likelihood of carrying real disease variants. As SampleSeq is able to appropriately combine these two pieces of information, it often results in a more efficient selection of subjects than the alternatives. When only one piece of information was used, using the burden of proxy marker alleles performed similarly to the random selection of cases, with the former being slightly better when the prevalence was K=0.01 and 0.05 and the latter slightly better when K=0.1 and 0.2. We also observed that as trait prevalence increased, the total number of captured rare disease alleles decreased (Tables IV.1-IV.4). This was due to the fact that the same number of disease variants with similar effects would account for a high fraction of heritability for a low prevalence disease than for a high prevalence disease, which resulted in a higher likelihood for a patient of a low prevalence disease to carry a disease variant in the targeted regions in our simulations. We also simulated a single region of size 1 Mb with a single disease variant; the results followed the same pattern (data not shown).

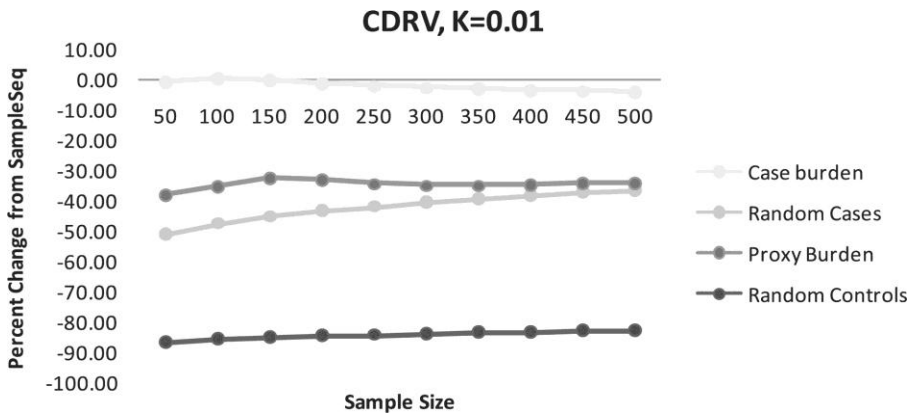

Figure IV.1. (CDRV) Percent change of disease alleles captured compared to SampleSeq using alternative methods, *K*=0.01, assuming a disease allele frequency of 0.01 in calculation of E_{ik}.

Table IV.1 (CDRV) Ratio of average counts of disease alleles captured to the average maximum possible disease alleles (percent change relative to SampleSeq), *K*=0.01, assuming p_d=0.01 in calculation of E_{ik}

Size	SampleSeq	Case Burden	Proxy Burden	Random Cases	Random Controls
50	0.31	0.31(-0.61)	0.19(-38.20)	0.15(-51.43)	0.04(-86.93)
100	0.33	0.33(0.52)	0.21(-35.34)	0.17(-47.78)	0.05(-85.95)
150	0.37	0.37(-0.11)	0.25(-32.41)	0.20(-45.43)	0.05(-85.32)
200	0.40	0.40(-1.33)	0.27(-32.96)	0.22(-43.67)	0.06(-84.84)
250	0.41	0.41(-1.98)	0.27(-34.40)	0.24(-42.16)	0.06(-84.44)
300	0.42	0.41(-2.56)	0.27(-34.99)	0.25(-40.86)	0.07(-84.09)
350	0.43	0.41(-2.93)	0.28(-34.99)	0.26(-39.69)	0.07(-83.77)
400	0.43	0.41(-3.46)	0.28(-34.77)	0.26(-38.53)	0.07(-83.46)
450	0.43	0.41(-3.65)	0.28(-34.25)	0.27(-37.44)	0.07(-83.16)
500	0.43	0.41(-4.06)	0.28(-34.03)	0.27(-36.68)	0.07(-82.96)

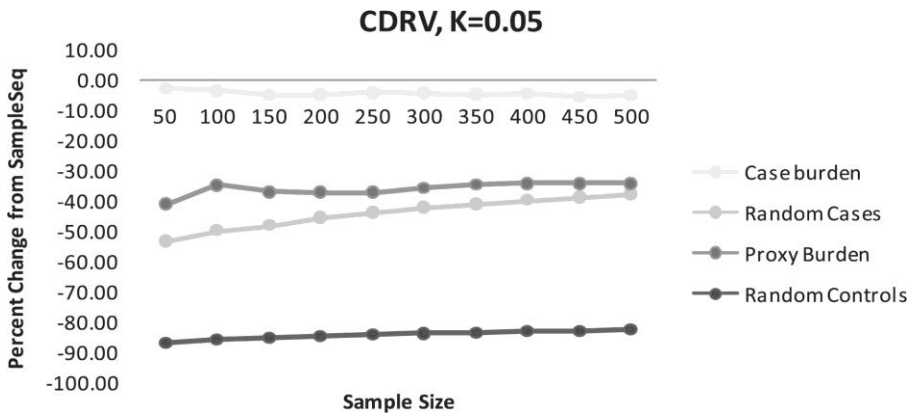

Figure IV.2. (CDRV) Percent change of disease alleles captured compared to SampleSeq using alternative methods, *K*=0.05, assuming a disease allele frequency of 0.01 in calculation of E_{ik}.

Table IV.2 (CDRV) Ratio of average counts of disease alleles captured to the average maximum possible disease alleles (percent change relative to SampleSeq), *K*=0.05, assuming p_d=0.01 in calculation of E_{ik}

Size	SampleSeq	Case Burden	Proxy Burden	Random Cases	Random Controls
50	0.28	0.27(-2.70)	0.17(-41.18)	0.13(-53.60)	0.04(-87.09)
100	0.34	0.33(-3.35)	0.22(-34.89)	0.17(-49.94)	0.05(-86.07)
150	0.37	0.35(-4.96)	0.23(-36.96)	0.19(-48.25)	0.05(-85.60)
200	0.38	0.36(-4.60)	0.24(-37.13)	0.20(-45.65)	0.06(-84.87)
250	0.38	0.36(-3.90)	0.24(-37.21)	0.21(-43.84)	0.06(-84.37)
300	0.38	0.36(-4.06)	0.24(-35.75)	0.22(-42.16)	0.06(-83.90)
350	0.38	0.36(-4.71)	0.25(-34.44)	0.22(-41.07)	0.06(-83.60)
400	0.38	0.36(-4.44)	0.25(-34.11)	0.23(-39.82)	0.06(-83.25)
450	0.38	0.36(-5.33)	0.25(-34.21)	0.23(-38.95)	0.06(-83.01)
500	0.37	0.35(-5.03)	0.24(-34.27)	0.23(-37.72)	0.06(-82.67)

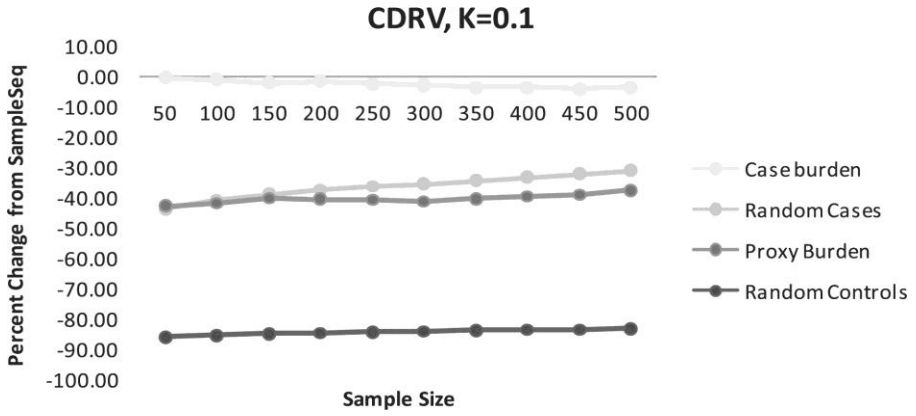

Figure IV.3. (CDRV) Percent change of disease alleles captured compared to SampleSeq using alternative methods, *K*=0.1, assuming a disease allele frequency of 0.01 in calculation of E_{ik}.

Table IV.3 (CDRV) Ratio of average counts of disease alleles captured to the average maximum possible disease alleles (percent change relative to SampleSeq), *K*=0.1, assuming p_d=0.01 in calculation of E_{ik}

Size	SampleSeq	Case Burden	Proxy Burden	Random Cases	Random Controls
50	0.24	0.24(-0.41)	0.14(-42.71)	0.13(-43.79)	0.03(-86.22)
100	0.27	0.26(-0.83)	0.16(-41.74)	0.16(-41.02)	0.04(-85.54)
150	0.27	0.27(-2.04)	0.16(-39.88)	0.17(-38.98)	0.04(-85.04)
200	0.28	0.27(-1.68)	0.16(-40.30)	0.17(-37.46)	0.04(-84.67)
250	0.28	0.27(-2.33)	0.16(-40.49)	0.18(-36.29)	0.04(-84.38)
300	0.28	0.27(-2.76)	0.16(-41.03)	0.18(-35.44)	0.04(-84.17)
350	0.28	0.27(-3.59)	0.16(-40.24)	0.18(-34.51)	0.04(-83.94)
400	0.27	0.26(-3.53)	0.16(-39.47)	0.18(-33.18)	0.04(-83.62)
450	0.27	0.26(-4.08)	0.17(-38.85)	0.18(-32.41)	0.05(-83.43)
500	0.29	0.28(-3.64)	0.18(-37.29)	0.20(-31.05)	0.05(-83.10)

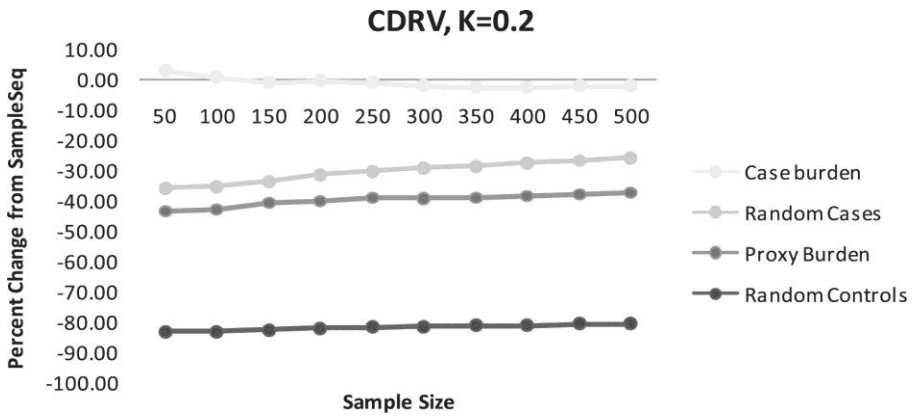

Figure IV.4. (CDRV) Percent change of disease alleles captured compared to SampleSeq using alternative methods, *K*=0.2, assuming a disease allele frequency of 0.01 in calculation of E_{ik}.

Table IV.4 (CDRV) Ratio of average counts of disease alleles captured to the average maximum possible disease alleles (percent change relative to SampleSeq), *K*=0.2, assuming p_d=0.01 in calculation of E_{ik}

Size	SampleSeq	Case Burden	Proxy Burden	Random Cases	Random Controls
50	0.20	0.20(3.01)	0.11(-43.57)	0.13(-35.88)	0.03(-83.38)
100	0.22	0.22(0.76)	0.13(-43.18)	0.15(-35.03)	0.04(-83.16)
150	0.23	0.23(-0.81)	0.14(-40.73)	0.15(-33.51)	0.04(-82.76)
200	0.23	0.23(-0.16)	0.14(-40.10)	0.16(-31.34)	0.04(-82.20)
250	0.23	0.23(-0.92)	0.14(-38.93)	0.16(-30.15)	0.04(-81.89)
300	0.23	0.22(-2.06)	0.14(-39.29)	0.16(-29.22)	0.04(-81.65)
350	0.23	0.22(-2.56)	0.14(-39.07)	0.16(-28.40)	0.04(-81.44)
400	0.23	0.22(-2.53)	0.14(-38.59)	0.16(-27.55)	0.04(-81.22)
450	0.24	0.24(-2.09)	0.15(-38.03)	0.18(-26.61)	0.05(-80.98)
500	0.27	0.26(-2.01)	0.17(-37.42)	0.20(-25.69)	0.05(-80.74)

We note that SampleSeq is sensitive to very low values of p_d as the algorithm is involved with solving linear equations, for which the solutions will be highly variable due to nearly singular matrices at very small values of p_d Our experience is that p_d should be at least $\frac{20}{2(n_{case}+n_{control})}$. For example, to select subjects from a pool of $n_{case}+n_{control}=2000$ subjects, setting p_d =0.005 is good but the performance will become less optimal for p_d <0.005. This limitation is computational. Our simulations showed that assuming p_d =0.01 was relatively robust to misspecification of the true frequency of *d* within the range we simulated, and performed well over all scenarios (data not shown).

When we simulated the synthetic association (SA) scenario (Figure IV.5, Table IV.5), we observed similar patterns as for the CDRV scenario, although all methods captured a higher proportion of the maximum number of possible disease alleles than the CDRV scenario. SampleSeq captured an average of 75% of the maximum possible disease alleles over all sample sizes. In our simulated scenario, the random selection of cases performed much better than using the burden of proxy marker alleles. This is most likely due to the large number of disease alleles to be found among the cases on both allelic backgrounds of SNP A compared to the number of causal alleles in controls.

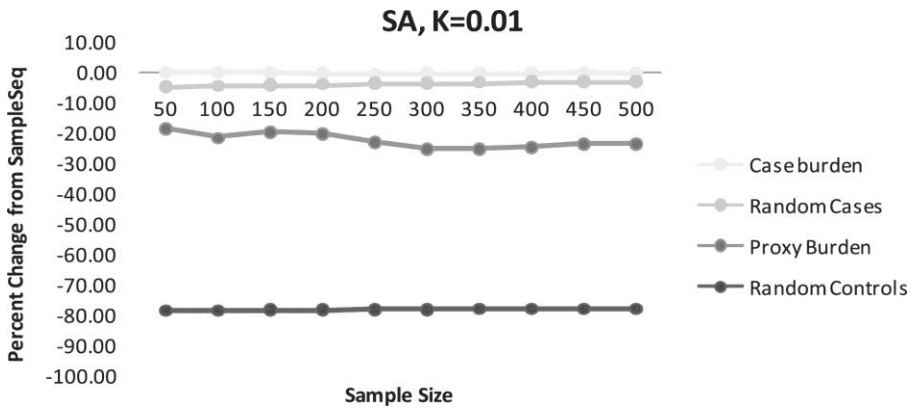

Figure IV.5. (SA) Percent change of disease alleles captured compared to SampleSeq using alternative methods, *K*=0.01, assuming a disease allele frequency of 0.01 in calculation of E_{ik}.

Table IV.5 (SA) Ratio of average counts of disease alleles captured to the average maximum possible disease alleles (percent change relative to SampleSeq), *K*=0.01, assuming p_d=0.01 in calculation of E_{ik}

Size	SampleSeq	Case Burden	Proxy Burden	Random Cases	Random Controls
50	0.66	0.66(-0.11)	0.54(-18.58)	0.63(-4.95)	0.14(-78.65)
100	0.69	0.69(-0.08)	0.54(-21.58)	0.65(-4.58)	0.15(-78.57)
150	0.72	0.72(-0.18)	0.58(-19.79)	0.69(-4.30)	0.15(-78.51)
200	0.74	0.74(-0.33)	0.59(-20.40)	0.71(-4.08)	0.16(-78.46)
250	0.76	0.75(-0.40)	0.58(-22.80)	0.73(-3.88)	0.16(-78.41)
300	0.77	0.76(-0.28)	0.57(-25.03)	0.74(-3.69)	0.17(-78.37)
350	0.77	0.77(-0.31)	0.58(-25.22)	0.75(-3.52)	0.17(-78.33)
400	0.78	0.78(-0.26)	0.59(-24.47)	0.76(-3.32)	0.17(-78.29)
450	0.80	0.80(-0.21)	0.61(-23.69)	0.78(-3.15)	0.17(-78.25)
500	0.82	0.81(-0.32)	0.62(-23.58)	0.79(-3.08)	0.17(-78.23)

For the CDCV scenario, we compared SampleSeq to the alternative methods when the disease alleles were not rare, but we assumed that p_d =0.01 in our calculations. In these experiments, SampleSeq was slightly more efficient than the burden of proxy alleles in cases, and was slightly less efficient than the burden of proxy risk alleles regardless of case status (Figure IV.6, Table IV.6). This was also true when p_d was close to the true frequency of *d*. Also notable was the generally smaller magnitude of differences among the other three methods. These results are due to the presence of many disease alleles in both cases and controls, as a result of the subtle effect sizes simulated in this scenario.

We further summarized the results of these simulations by comparing the average expected counts of disease alleles as determined by SampleSeq, *E*(*d*), to the average count of observed disease alleles, *d*, for each scenario and sample size (Table IV.7). As the calculation of E_{ik} is based on a hypothetical disease variant frequency (p_d=0.01 in our calculations), *E*(*d*) may not match the true number of disease alleles. For the CDRV scenarios, where the frequency of *d* was between 0.0025 and 0.01, the ratio of the average expected to observed alleles across sample sizes were 1.7 for *K*=0.01, 2.0 for *K*=0.05, 2.9 for *K*=0.1, and 2.7 for *K*=0.2. When the allele frequency of *d* was constrained to fall within the range 0.009–0.011, and p_d was set to 0.01, the ratio of *E(d)* to *d* was 1.3 for *K*=0.01. However, for the SA scenario, the ratio of average *E*(*d*) to *d* was 0.41, although this value is counting the observation of all disease alleles in a region, where there were 10 disease alleles in the simulation per region. Also for the CDCV scenario, the ratio of average *E(d)* to *d* was 0.27, demonstrating that our method is based on finding rare disease alleles, and that if the disease alleles are common, they will occur much more often than expected by SampleSeq assuming a rare p_d.

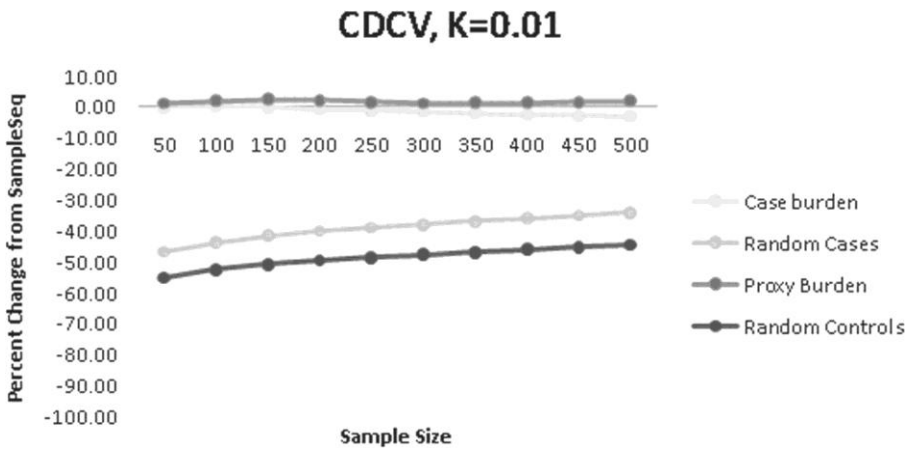

Figure IV.6. (CDCV) Percent change of disease alleles captured compared to SampleSeq using alternative methods, *K*=0.01, assuming a disease allele frequency of 0.01 in calculation of E_{ik}.

Table IV.6 (CDCV) Ratio of average counts of disease alleles captured to the average maximum possible disease alleles (percent change relative to SampleSeq), *K*=0.01, assuming p_d=0.01 in calculation of E_{ik}

Size	SampleSeq	Case Burden	Proxy Burden	Random Cases	Random Controls
50	0.9	0.90(-0.13)	0.91(1.22)	0.48(-46.88)	0.40(-55.27)
100	0.91	0.92(0.26)	0.93(1.82)	0.51(-43.86)	0.43(-52.72)
150	0.91	0.91(-0.18)	0.94(2.45)	0.53(-41.74)	0.45(-50.94)
200	0.91	0.90(-0.92)	0.93(2.13)	0.54(-40.22)	0.46(-49.66)
250	0.9	0.89(-1.35)	0.92(1.57)	0.55(-39.06)	0.46(-48.68)
300	0.91	0.90(-1.66)	0.92(1.21)	0.57(-38.06)	0.48(-47.84)
350	0.92	0.90(-1.95)	0.93(1.28)	0.58(-37.04)	0.49(-46.98)
400	0.93	0.91(-2.48)	0.94(1.41)	0.59(-36.13)	0.50(-46.21)
450	0.93	0.90(-2.90)	0.94(1.62)	0.60(-35.17)	0.51(-45.40)
500	0.93	0.90(-3.15)	0.95(1.92)	0.61(-34.18)	0.51(-44.57)

Table IV.7. The average cumulative expected count of disease alleles, denoted *E*(*d*), and the actual observed average count of disease alleles, denoted *d*, for each scenario in Figures IV.1–IV.6, denoted IV.1–IV.6, for each of 10 sample sizes.

Size	CDRV							
	IV.1 *E(d)*	**IV.1 *d***	**IV.2 *E(d)***	**IV.2 *d***	**IV.3 *E(d)***	**IV.3 *d***	**IV.4 *E(d)***	**IV.4 *d***
50	60.61	33.01	57.77	27.37	54.19	17.20	41.11	13.29
100	112.43	61.40	106.16	50.73	100.19	32.44	76.66	26.24
150	159.77	88.14	150.05	73.62	142.48	47.04	109.77	38.54
200	203.92	113.85	190.93	93.46	182.28	61.19	140.98	49.65
250	245.86	138.59	229.21	113.05	219.95	75.08	170.25	61.00
300	286.01	162.67	266.24	131.73	256.05	88.91	197.91	72.24
350	323.93	186.10	302.01	150.83	290.30	102.26	225.60	83.32
400	360.35	208.67	335.85	168.82	322.98	114.54	252.02	94.10
450	394.54	230.64	376.42	187.19	353.79	127.40	276.51	104.51
500	429.20	253.22	398.36	203.89	385.89	138.76	300.98	114.68

Size	SA		CDCV	
	IV.5 *E(d)*	**IV.5 *d***	**IV.6 *E(d)***	**IV.6 *d***
50	126.25	250.83	91.49	326.19
100	236.61	499.71	171.66	617.27
150	339.02	747.42	249.07	892.26
200	433.72	994.22	318.45	1159.36
250	523.99	1240.19	385.29	1421.55
300	611.43	1485.35	449.50	1678.27
350	696.42	1729.79	509.24	1926.32
400	779.41	1972.88	569.92	2170.27
450	859.65	2215.47	634.62	2405.16
500	936.72	2459.95	676.23	2632.34

Although E_{ik} *changes* as the hypothetical disease variant frequency changes, using these estimates to rank subjects when an incorrect allele frequency is used in the calculation is still an effective means of selecting subjects. To demonstrate this, we calculated *E(d)* under p_d =0.1, 0.01 and 0.001, using the simulation parameters from the experiment in Figure IV.1, and present their correlation coefficients in Table IV.8. While the magnitude of the value of *E(d)* is proportional to the assumed value of p_d, the ranking of subjects is similar even when p_d is misspecified.

Table IV.8. Correlation coefficient between $E(d)$ from three settings of p_d.

	p_d=0.1	p_d=0.01	p_d=0.001
p_d=0.1	-	0.908	0.841
p_d=0.01	0.938	-	0.988
p_d=0.001	0.928	0.999	-

Five regions, one rare disease variant per region, K=0.01. Correlations for cases are in the upper triangle and those for controls are in the lower triangle.

Discussion

Targeted resequencing using next-generation sequencing technology allows investigators to fine map regions identified in GWAS to localize true variants. Since it is generally not feasible for an investigator to sequence everybody in a large GWAS, questions arise as to the optimal design of follow-up studies aimed at identifying novel, particularly rare, variants that may explain the GWAS signals: the optimal balance between numbers of subjects, depth of sequencing and sizes of regions; follow-up by further sequencing of selected variants or imputation; whether to use DNA pooling or family-based designs; choice of specific subjects for sequencing, etc. (D.Thomas and F.Yang, personal communication). We developed SampleSeq to address the last issue.

We have conducted a simulation study of several scenarios that have been postulated to represent the genetic architecture of common complex traits in human populations. We explored individual rare variants with strong effects, the synthetic association scenario with multiple rare variants per region and the CDCV model to evaluate our approach for capturing causal alleles. We demonstrated that SampleSeq can estimate the count of rare causal alleles in a sample of subjects from a case–control study, estimate the sample size required to capture a

specified number of alleles in each region of interest and select subjects to optimize and balance the capture of alleles across an arbitrary number of regions.

When designing a next-generation resequencing study, a compromise must be struck between read depth and sample size. Regardless of the balance between these parameters, the allele frequency in the sample will be the primary determinant of whether genotypes are called accurately. By increasing the frequency of a disease allele in a sample of subjects, the accuracy of genotype calls and the chance that any resequencing study design will detect the presence of that allele will be improved. To increase the chance of detecting disease variants in a targeted resequencing study, an intuitive strategy is to select cases according to the dosage of risk alleles at the reported associated SNP (Thomas *et al.* 2009). Our results showed that this is indeed a good strategy compared to random selection of cases or controls. However, because of incomplete penetrance, a control subject homozygous for risk alleles at several loci may have a higher chance of carrying a real disease variant than a case subject who is heterozygous for some of those risk alleles. SampleSeq allows us to quantify their probabilities of carrying real disease variants and then select subjects accordingly.

We observe that compared to random controls, samples of random cases have much better performance for discovering rare disease alleles, which is consistent with previous studies (Li and Leal 2009). Some investigators may choose to evaluate a set of controls in order to perform screening with association tests before proceeding to large-scale variant-based genotyping. This is likely the most effective strategy when resources are abundant for resequencing studies and sample sizes are large. However, when sample sizes are small, we would expect most of the ability to detect the presence of rare disease alleles in the population to come from the cases. As the majority of samples selected by SampleSeq will be cases, in some situations, it may be reasonable to resequence some controls to augment the SampleSeq

selection. The control subjects could then be used to screen variants for frequency differences and prioritize for genotyping. This comparison would be biased due to the frequency enrichment achieved by SampleSeq, but could help discern the SNPs that should be tested for association with the trait with unbiased approaches, such as genotyping in the full cohort. However, reallocating resources to sequence additional controls would also lower the chance of seeing real disease variants in the cases. If the number of subjects that can be resequenced is small, we advocate also using the sequence context and putative biological impact of variants to prioritize SNPs for genotyping, as there will not be a large amount of statistical information for comparing rare variant frequencies in small samples.

Some recent research has shown that association testing from sequence data may provide slightly more statistical power than variant-based genotyping on a per-subject basis (Liu and Leal 2010) using two recently developed tests of association (Li and Leal 2009; Madsen and Browning 2009). However, we note that due to the large difference in the cost of resequencing to the cost of variant-based genotyping, on a per-unit of resources basis, many more subjects could be genotyped with variant-based methods than could be resequenced. Thereby, the statistical power to detect an association might be considerably better in a large sample of variant-based genotypes than in a small sample of sequence-based genotypes, utilizing the same resources. The goal of this work is to optimize the resources expended for resequencing studies, preserving DNA samples and financial assets for subsequent steps in investigations.

The key element of the model that provides SampleSeq with a performance advantage over counting common risk alleles is the assumption of rare ancestral recombination between SNPs A and D. We assumed that disease variant *d* originated on the 'risk' allele *a* background, which resulted in three haplotypes, *AD*, *aD* and *ad*. To break up the LD between the SNPs through recombination, the recombination event needs to occur in the double heterogyzotes,

for which the frequency is quite low as *d* is rare. Moreover, if the two SNPs are close enough to have very low recombination fraction between them, then the chance of breaking up the LD between the SNPs will be small. This is supported by our simulation data; the recombinant haplotype frequency averaged 3.7×10^{-4} across all our CDRV haplotype pools, suggesting $pdA\approx0$ is a reasonable assumption. It is implemented in our software for the ease of computation. In simulations where this assumption was badly violated, such as the CDCV scenario, the performance of SampleSeq was still competitive with the burden of risk alleles in cases.

We also noted that as the prevalence in our simulations increased, but the ORs of rare disease variants was held constant, the proportion of cases not carrying any risk alleles at any of the target disease loci increased. This observation is a result of our simulation strategy, but it is perhaps worthy of note that high-prevalence traits may require many more rare risk alleles than low-prevalence traits for the CDRV model to account for most of the trait heritability for a highly heritable common trait. Thereby, if there are not a large number of associated regions identified for a high-prevalence trait, it is possible that the yield of rare disease alleles from a resequencing study of that trait may be small, as additional trait variation may be due to untargeted regions or environmental influences.

As next-generation sequencing technology matures, the need for targeted resequencing of association study-implicated regions for fine-mapping of mutations may eventually expire. However, for researchers who do not have access to tremendous financial resources or the most current sequencing platforms, targeted resequencing followed by variant-based genotyping of candidate SNPs is likely the most direct and cost-efficient means of fine-mapping of causal rare mutations. Additionally, this approach capitalizes on previous discoveries, rather than pursuing agnostic resequencing of whole genomes or exomes. While agnostic approaches to discovery will and should be taken, we believe there is also a role for hypothesis-based resequencing

studies in human genetic epidemiology in the foreseeable future. The SampleSeq software is available at http://biostat.mc.vanderbilt.edu/SampleSeq

CHAPTER V

EFFICIENT DETECTION OF TUMOR SOMATIC MUTATIONS USING NEXT-GENERATION SEQUENCING DATA[5]

Introduction

Cancer is a class of complex genetic diseases that are responsible for one in eight deaths worldwide (Blecher *et al.* 2008). It is characterized by uncontrolled proliferation of malignant cells that can intrude into surrounding tissues and metastasize to distant organs because of somatically accumulated mutations and epigenetic changes (Esteller 2007; Stratton *et al.* 2009). Cancer somatic mutations encompass distinct classes of DNA and chromosomal changes, including base substitutions, insertions and deletions (indels), copy number variants (CNVs), and inter- and intra-chromosomal rearrangements. Cells in some cancer types may also have obtained exogenous DNA from viruses (Talbot and Crawford 2004).

Somatic mutations are distributed across the genome (Stratton 2011). Only a small subset of mutations has been implicated in oncogenesis. The type and rate of somatic mutations may vary by cell type and can be influenced by mutagenesis exposures endogenously and exogenously. As a result, different cancer types have different mutation profiles. Some cancer types display very disordered genomes while others have few genomic aberrations (Stratton 2011). It is important to understand the overall mutation profiles for different cancer types and

---

[5] Song, Z., Long, J. He, J., Shi, J., Shu, X., Cai, Q., Zheng. W., and Li, C. (2011). Efficient Detection of Tumor Somatic Mutations using Next-Generation Sequencing Data. (submitted). Author contributions: Conceived and designed the experiments: ZS CL. Performed the experiments: ZS. Analyzed the data: ZS CL. Contributed reagents/materials/analysis tools: JL JH JS XS QC WZ. Wrote the paper: ZS CL. Algorithm design: ZS CL Software: ZS.

in different patients as they may influence choice of treatment as well as a patient's reaction to a treatment. In early cytogenetic studies of cancer, most mutations found were chromosomal translocations and CNVs. Thanks to the development of next-generation sequencing (NGS) technologies, we are on the edge of creating a complete catalogue of mutations for each cancer type (Singer 2011) by screening the cancer genome at the finest resolution.

The NGS technologies have recently been used to study somatic mutations in melanoma (Pleasance *et al.* 2010), lung cancer (Lee *et al.* 2010; Pleasance *et al.* 2010), colorectal cancer (Timmermann *et al.* 2010), renal carcinoma (Varela *et al.* 2011), breast cancer (Ding *et al.* 2010; Shah *et al.* 2009), acute myeloid leukemia (Mardis *et al.* 2009), mesothelioma (Bueno *et al.* 2010), and prostate cancer (Berger *et al.* 2011). Although the methods for mutation detection differ in details, these studies all relied on comparing SNP calls across samples in their primary mutation screening.

Efficient analysis tools for NGS data have been developed. The Genome Analysis Toolkit (GATK) (McKenna *et al.* 2010), for example, is a suite of tools for sequence data processing and genotype calling, and provides a rich software library for developing additional analysis tools. The GATK Unified Genotyper and some other SNP calling programs do an excellent job of calling genotypes for homogeneous samples such as blood and normal tissues. In cancer genomes, however, cells taken from a tumor tissue may include non-cancer cells. In addition, cancer cells in a tumor sample can be heterogeneous with respect to somatic mutations. As a result, a somatic mutation may exist only in a fraction of cells in a tumor sample, lowering the likelihood of detecting it through genotype calls. The resulting allelic imbalance may also lead to tumor genotype calls being flagged as problematic and excluded from analysis. However, existing tools for detecting somatic mutations mostly rely on genotype calls for tumor samples. For example, VarScan (Koboldt *et al.* 2009) screens data for mutation sites by comparing genotype calls

between normal and tumor samples, and SomaticSniper (Mardis *et al.* 2009) uses a genotype likelihood model (Li *et al.* 2009) to calculate the probability that the tumor and normal genotypes are different. VarScan (Koboldt *et al.* 2009) does use allele frequency information to evaluate significance, but only for sites that show genotype difference between normal and tumor samples.

We develop and implement novel algorithms for detecting two major types of somatic mutations, base substitution and loss of heterozygosity (LOH), by directly examining the sequence reads of tumor samples. We applied our methods, named Somatic Mutation Gleaner (SMUG), to whole exome sequencing data of eight breast cancer tumors and their matched blood samples, and detected somatic mutations that are missed by comparison of genotype calls between normal and tumor samples.

Method

We consider patients of the same type of cancer, each with NGS data available for a tumor sample and a normal sample (e.g., from blood or normal tissue). In SMUG, we first process each normal-tumor sample pair and then aggregate information across patients. Before applying our methods, general filtering criteria can be employed such as setting thresholds for base, mapping, and genotype quality scores. The depth for a site is defined as the number of remaining bases at the site after filtering.

Detection of Base Substitutions

We seek to identify sites that are homozygous in the normal sample but have mutant alleles in the tumor sample. For each patient, we first walk through the tumor sequence data to look for sites with more than one type of nucleotide bases, and then check normal sample genotypes and keep the sites that are homozygous in the normal sample. As base substitutions can appear as an artifact during sample preparation (e.g., PCR error), sequencing (e.g., reading error), and data processing (e.g., alignment error), one may want to ignore sites with very few copies or very low fraction of "mutant" alleles; this will reduce noise as well as the amount of computation down the road. To obtain genotypes in the normal sample, forced calling may be necessary as SNP callers often do not output results for sites deemed monomorphic.

For each selected site j, the fraction of mutant alleles could be estimated as x_j/n_j, where n_j is the depth at the site and x_j is the number of bases at the site that are different from the allele in the normal sample homozygous genotype. Both n_j and x_j are counts of bases after filtering. However, this ratio can have very different accuracy depending on the denominator; for example, a site with $x = 30$ and $n = 200$ may be more likely to carry a real somatic mutation than a site with $x = 3$ and $n = 20$ although they have the same fraction of mutant alleles. This issue can be addressed by employing an empirical Bayes (EB) approach, in which the mutant allele count is assumed to follow a binomial distribution, $x_j \sim \text{Bin}(n_j, \vartheta_j)$, where ϑ_j is the true fraction of mutant alleles and follows a prior distribution $\text{Beta}(\alpha, \beta)$. The posterior distribution for ϑ_j is $\text{Beta}(x_j + \alpha, n_j - x_j + \beta)$ and the EB-estimate of the fraction of mutant alleles is the posterior mean $r_j = (x_j + \alpha)/(n_j + \alpha + \beta)$. The adjustment will be small for sites with high depth and relatively high for sites with low depth. As a result, the EB-estimate for a site with $x = 30$ and $n = 200$ is often higher than that for a site with $x = 3$ and $n = 20$. The parameters α and β are obtained as the maximum likelihood estimates of the marginal likelihood function,

$\prod f(x_j|\alpha,\beta,n_j)$,where $f(x|\alpha,\beta,n) = \int \binom{n}{x} \theta^x (1-\theta)^{n-x} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \theta^{\alpha-1}(1-\theta)^{\beta-1} d\theta$ and the product is over all the selected sites. The parameters α and β may vary across samples.

A site identified by this approach may also carry "mutant" alleles in the normal sample even if it is called homozygous with a genotype quality score high enough to pass filtering. For example, for one of our patients (patient 8), position 28241838 on chromosome 20 was called homozygous C/C in blood but there were 55 copies of C and 7 copies of T after filtering by base, mapping, and genotype quality scores. The presence of multiple copies of T could be due to contamination, somatic mutation in the normal sample, sequencing error, or alignment error. We implement two filters for removing these sites. One is to apply the procedure described above to the normal sample and calculate the adjusted EB-estimate, $r_{1j} - r_{0j}$, where r_{1j} and r_{0j} are the EB-estimates for the tumor and normal samples, respectively, at site *j*. Another is to calculate the p-value for a t-statistic for each site, $D/\sqrt{V} = \left(\frac{x_1}{n_1} - \frac{x_0}{n_0}\right) / \sqrt{\left(\frac{1}{n_1} + \frac{1}{n_0}\right) p(1-p)}$, where x_1 is the number of mutant alleles and n_1 is the depth in the tumor sample, x_0 and n_0 are their counterparts for the normal sample, and $p = (x_1 + x_0)/(n_1 + n_0)$. Our experience shows that the former works reasonably well while the latter can result in a misleadingly significant p-value for a site with high depth.

A somatic mutation that occurs in multiple patients is likely to be relevant to tumor development or progression. We thus aggregate information across patients for each base substitution site. One measure is the total adjusted EB-estimates across the patients for whom the site was selected. Another measure is the p-value for the composite t-statistic $\sum D / \sqrt{\sum V}$, where the sums are over the patients for whom the site was selected. These measures are analogous to the filters developed above. Our program can output sorted site- and gene-level summaries that are ready for further annotation using programs such as ANNOVAR (Wang *et al.* 2010) or PolyPhen-2 (Adzhubei *et al.* 2010).

The results will be useful for selection of sites for further replication; for example, one may select sites that have adjusted EB-estimate ≥0.1 in ≥20% of the patients. Its false discovery rate (FDR), the fraction of false positives among the selected sites, can be estimated using a permutation procedure. Let S_i be the set of qualifying sites (e.g., sites with adjusted EB-estimate ≥0.1) for patient *i*. Let *S* be the union of S_i over all patients. For each patient, we permute the sites in *S*; different patients are permuted independently. Then we apply the same selection criterion (e.g., qualifying sites in ≥20% of the patients) to the permuted data. This procedure is repeated multiple times, and the average number of selected sites across permutation replicates is an estimate of the number of false positive sites. The FDR can then be estimated as the number of false positive sites divided by the number of sites selected from the real data.

Detection of Regions with Loss of Heterozygosity

Ideally, LOH could be detected by looking for sites that are heterozygous in the normal sample and homozygous in the tumor sample. However, as for base substitutions, a tumor sample may be a mixture of tumor and non-tumor cells, and loss of an allele may not occur in all tumor cells. As a result, an LOH region may not manifest a total loss of an allele in the tumor sample, and thus may not be detectable through comparison of genotype calls between normal and tumor samples. A partial loss of an allele will result in a departure of allele counts from 50:50. We develop a method to capture this information.

For each patient, we focus on sites that are heterozygous in the normal sample. If there is no LOH, the tumor sample base counts should follow a binomial distribution with probability 0.5 for each of the two alleles in the normal sample. A significant departure from this distribution is evidence of LOH, which can be quantified by defining an LOH score. We first compute a two-tailed p-value for the observed base counts and convert it to the Phred scale, *Ph*

= –10log(p), rounded to the nearest integer and capped at 99. The LOH score is defined as Ph – $E(Ph \mid d)$, where $E(Ph \mid d)$ is the expected Phred score under no LOH given depth d at the site. The expected Phred score depends on depth because the distribution of Phred scores varies as the depth changes. At depth d, there are $d + 1$ possible outcomes, $x = 0, \ldots, d$, each having a probability p_x and a Phred score Ph_x under no LOH. The expected Phred score is the weighted average of these $d + 1$ scores, $E(Ph \mid d) = p_0 Ph_0 + \ldots + p_d Ph_d$.

The LOH score defined above allows us to identify potential LOH sites and regions. For example, one may apply a threshold s_{LOH} to LOH scores and select sites with LOH score ≥ s_{LOH} as potential LOH sites. An LOH region can then be defined as a contiguous region that contains ≥ n_{LOH} LOH sites without interruption, that is, all normal sample heterozygous sites in the region have an LOH score ≥ s_{LOH}. Higher values of the thresholds s_{LOH} and n_{LOH} result in more stringent criteria. The boundaries of an LOH region can be conservatively defined as the positions of its first and last LOH sites. We note that correct identification of an LOH region relies both on successful detection of normal sample heterozygous sites inside the region and on strong evidence of LOH in the tumor sample. When there is not enough information (e.g., due to low or no coverage), a normal sample heterozygous site may not be reliably called or the evidence of LOH may not be strong enough in the tumor sample. In this situation, an LOH region may not be detected due to a lack of supporting LOH sites, and multiple LOH regions may be identified as a single one due to a lack of sites breaking them apart.

The overall evidence of LOH for a site can be obtained by adding individual LOH scores over patients heterozygous in the normal sample. This is justified by the fact that the p-values are independent across subjects and under no LOH, $-2\ln(p) \sim \chi^2_2$, a chi-squared distribution with two degrees of freedom. Since –10log(p) = –2ln(p)×(5/ln10) and the sum of k independent χ^2_2 statistics follows χ^2_{2k}, the sum of Phred scores over k patients is expected to follow

$(5/\ln 10)\chi^2_{2k}$ under no LOH. Thus the total LOH score can reflect the overall evidence of LOH. Our program can output sorted site- and gene-level LOH results that are ready for further annotation.

Results from the Shanghai Breast Cancer Study

We applied SMUG to whole exome sequencing data for blood and tumor samples from eight breast cancer patients. We evaluated the performance of SMUG and the approach of comparing genotype calls between normal and tumor samples; the latter will be referred to as CALL for brevity. The same filtering criteria were used in both SMUG and CALL: mapping quality score ≥ 20, base quality score ≥ 20, and genotype quality score ≥ 20. Similar patterns of results were observed when alternative threshold values were used.

Samples and Exome Sequencing

We selected eight breast cancer patients from the Shanghai Breast Cancer Study (SBCS) for whole exome sequencing of their blood and tumor DNA samples. The SBCS is a large-scale, population-based case-control study conducted in urban Shanghai (Gao *et al.* 2000). Seven patients have early-onset (28-32 years old) breast cancer and one has early-onset (39 years old) plus first-degree family history of breast cancer. Their breast cancer TNM (Tumor, Node, Metastasis) stages ranged from I to III (Table V.1). Tumor DNA was extracted from buffy coat and snap frozen breast cancer tissue using QIAmp DNA Kit (Qiagen, Valencia, CA) following the manufacturer's protocols.

Table V.1. Data summary. Patient characteristics: age (years old), family history, tumor stage. Exome data: number of aligned reads, median depth in target regions, α and β. Blood genotypes: overall and heterozygous consistency rate with chip-based SNP calls.

Patient	Age	Family history	TNM stage	Blood						Tumor			
				#aligned ($\times 10^6$)	Depth	Consistency rate (%)	Het consistency rate (%)	α	β	#aligned ($\times 10^6$)	Depth	α	β
1	31	0	IIa	29.7	43	99.31	97.99	1.22	28.14	10.6	16	4.41	29.80
2	28	0	III	34.6	46	99.71	99.35	1.56	33.19	16.9	25	7.66	67.98
3	31	0	IIa	31.5	44	99.51	98.58	1.22	31.95	16.3	22	4.13	29.10
4	32	0	IIb	36.9	48	99.51	98.72	1.28	27.50	10.1	15	6.04	44.20
5	32	0	IIb	34.4	48	99.65	99.06	0.88	21.91	15.8	23	6.46	52.77
6	30	0	I	35.7	49	99.58	98.85	0.75	20.48	18.7	27	5.88	49.39
7	28	0	IIa	32.6	43	99.63	99.08	0.61	17.14	12.5	19	9.93	84.08
8	39	1	III	31.3	43	99.22	98.75	1.06	28.94	16.4	23	3.59	25.19

DNA enrichment was done using Agilent SureSelect Human All Exon kit v1 (Gnirke *et al.* 2009), which was designed to target 165,637 genomic regions (37.8 million bases, or 1.22% of the human genome). The consensus coding sequence database has 27.8 million bases, 97.4% of which are covered by the assay's target regions. Exome sequencing was conducted at the HudsonAlpha Institute for Biotechnology. The data were 72-base paired-end reads generated from Illumina GA IIx machines (Bentley *et al.* 2008). Each sample was run on a single lane of a flow cell. The median depth in target regions was on average 45x for the blood samples and 21x for the tumor samples (Table V.1).

We shifted the Illumina base quality score to the Sanger scale and performed initial alignment to the NCBI human reference genome version 36 using BWA (Li and Durbin 2009). We then marked duplicates with Picard, carried out regional realignment and quality score recalibration using GATK (McKenna *et al.* 2010). We used GATK Unified Genotyper to call SNPs, which were then filtered to remove low quality calls. The blood and tumor samples were processed separately. The blood DNA samples were also genotyped using the Affymetrix 6.0 chip in our previous genome-wide association study (Zheng *et al.* 2009). The chip-based and sequence-based SNP calls were compared for SNPs overlapping the two platforms, and the consistency rate was over 99% for all samples, suggesting very good quality of our sequencing data (Table V.1).

Detection of Base Substitutions

For the CALL approach, we obtained genotype calls for the blood and tumor samples. Sites that were heterozygous in a tumor sample but had no call in the corresponding blood sample were force-called in the blood sample to check if the blood genotype was homozygous

with high quality. For each patient, the CALL list of base substitutions consisted of sites that were homozygous in blood and heterozygous in tumor after filtering.

For the SMUG approach, we screened each tumor sample for sites that had depth ≥5, at least 3 "mutant" alleles with fraction ≥ 0.05. These putative mutation sites were force-called in blood samples. For each patient, her SMUG list consisted of the putative mutation sites that were homozygous in blood after filtering. The parameters α and β were estimated (Table V.1) and the EB-estimates of the fraction of mutant alleles were calculated for all SMUG sites.

In both approaches, we focused on mutation sites that were observed in at least two patients. For SMUG sites, we required ≥ 0.15 total adjusted EB-estimates over mutant carriers. SMUG identified 1,111 sites and CALL found 614. Among the 465 sites that were identified by both methods, 221 had the same number of supporting patients, 228 had more supporting patients in SMUG, and only 16 had more supporting patients in CALL. SMUG identified 646 additional sites that were missed by CALL. There were also 149 sites that were identified by CALL but not by SMUG; all these sites had either very few mutant alleles or very low adjusted fraction of mutant alleles.

Further examination of the base counts at the identified sites provided some insight on why SMUG performed differently from CALL. Table V.2 contains a few examples. (1) Position 180434779 on chromosome 3 is a well-known non-synonymous breast cancer somatic mutation site in the *PIK3CA* gene. Both SMUG and CALL identified evidence of base substitution at this site in three patients. (2) In many situations, a low fraction of mutant alleles makes it difficult to detect mutations through genotype calls. For position 42589657 on chromosome 17 in the *CDC27* gene, SMUG found evidence of non-synonymous somatic mutation in six patients while CALL could detect only three of them. This is because the tumor samples of the other three patients were called homozygous, although they each had multiple copies of the putative

mutant alleles. (3) A low fraction of mutant alleles can also result in allelic imbalance, which can lead to low quality genotype calls for tumor samples. One example is position 132744647 on chromosome 2, for which CALL had low quality genotype calls for two tumor samples. (4) We also examined the sites that were identified by CALL but missed by SMUG. These sites failed the criteria we used in SMUG due to either too few "mutant" alleles or very small adjusted EB-estimate; one example is position 30958603 on chromosome 10 (Table V.2).

We also compared the results from SMUG and CALL with the Catalogue of Somatic Mutations in Cancer (COSMIC) database (Bamford *et al.* 2004; Forbes *et al.* 2011) release v52, which contains 19,224 sites and 8,951 genes related to human cancers, among which 1,638 sites and 1,030 genes were related to breast cancer. SMUG identified two COSMIC sites and one in breast cancer, the *PIK3CA* site described above; CALL only identified the *PIK3CA* site. At the gene level, the SMUG sites were inside 67 COSMIC genes (10 breast cancer genes), while the CALL sites hit 41 COSMIC genes (7 breast cancer genes) (Table V.3). We further stratified the identified sites by their annotated functions. SMUG consistently hit more COSMIC genes than CALL did for non-synonymous, synonymous, and stop gain mutations. We did not find any stop loss substitution in COSMIC genes (Table V.3).

Table V.2. Examples of sites for comparison between SMUG and CALL for detection of base substitutions. Base counts and genotype calls for both blood and tumor samples for all identified base substitutions are shown.

Chr	Position	Total EB-estimate	Pt. 1	Pt. 2	Pt. 3	Pt. 4	Pt. 5	Pt. 6	Pt. 7	Pt. 8
3	180434779	0.5932	86,0,0,0	112,0,0,0	80,0,0,0					
	(PIK3CA)		15,0,7,0	40,0,12,0	19,0,34,0					
	(nonsyn)		A/A;A/G	A/A;A/G	A/A;A/G					
17	42589657	0.6168			0,0,0,68	0,0,0,81	0,0,0,74	0,0,0,87	0,0,0,70	0,0,0,70
	(CDC27)				0,0,7,40	0,0,4,16	0,0,7,45	0,0,3,53	0,0,3,39	0,0,5,55
	(nonsyn)				T/T;T/G	T/T;T/G	T/T;T/G	T/T;T/T	T/T;T/T	T/T;T/T
2	132744647	0.5615		0,0,15,0		0,0,26,0	0,0,25,0	0,0,12,0	1,0,30,0	0,0,14,0
				3,0,36,0		6,0,26,0	7,0,40,0	3,0,21,0	9,0,43,0	3,0,30,0
				G/G;G/G		G/G;A/G	G/G;A/G	G/G;./.	G/G;A/G	G/G;./.
10	30958603	-	*0,1,0,67*		*1,5,0,56*					
			0,2,0,11		*0,6,0,22*					
			T/T;C/T		T/T;C/T					
SNP1		0.4108			0,1,0,27	0,0,0,48	0,3,0,41			
					0,6,0,13	0,5,0,17	0,9,0,21			
					T/T;C/T	T/T;C/T	T/T;C/T			
SNP2		0.3252				0,0,30,1			0,0,28,2	0,0,34,1
						0,0,8,3			0,0,21,3	0,0,15,6
						G/G;G/T			G/G;./.	G/G;G/T
SNP3		0.2885				2,0,26,0	1,0,29,0			
						9,0,11,0	7,0,14,0			
						G/G;A/G	G/G;A/G			

Under patients: Row 1: counts of bases A, C, G, T, respectively, in blood; Row 2: base counts in tumor; Row 3: blood genotype followed by tumor genotype. No call is denoted as "./.". Base counts and genotypes are after filtering. Base substitutions identified by CALL are marked by a line above genotypes. Base substitutions missed by SMUG are in italics.

Table V.3. Comparison of SMUG and CALL for detection of base substitutions at COSMIC sites and genes.

	CALL		SMUG	
	Total	**Breast cancer**	**Total**	**Breast cancer**
COSMIC sites	1	1	2	1
COSMIC genes:	41	7	67	10
Non-synonymous	25	5	45	6
Synonymous	41	7	67	10
Stop gain	0	0	2	0
Stop loss	0	0	0	0

Base Substitution Result Validation and Replication

We collected DNA samples for normal breast tissue and tumor breast tissue for 3 patients (patients 1-3) in this study and 104 additional patients in the Shanghai breast cancer study. These samples were genotyped for a set of SNPs in a different study. Three SNPs were in the top list of our SMUG results (called SNP1, SNP2, and SNP3, respectively in Table V.2). We evaluated these three SNPs to see if our results could be validated for patients 1-3 and replicated in the independent set of 104 patients.

Genotyping was performed using the iPLEX Sequenom MassArray platform (Cai *et al.* 2011; Long *et al.* 2010; Zheng *et al.* 2009). Polymerase chain reaction (PCR) and extension primers were designed using the MassARRAY Assay Design software (Sequenom, Inc). PCR and extension reactions were performed according to the manufacturer's instructions, and extension product sizes were determined by mass spectrometry using the Sequenom iPLEX system. On each 96-well plate, two negative controls (water), two blinded duplicates, and two HapMap samples were included. The median consistency rate was 100% for both the duplicate samples and the HapMap samples when compared with their genotypes in the HapMap database.

After quality control filtering, 93 of the 104 patients had genotype at SNP1 for both normal and tumor breast tissues. Among them, 11 patients were homozygous in normal tissue and heterozygous in tumor tissue, resulting in an estimated mutation incidence rate of 11.8%. For SNP2, 67 patients had genotype available for both tissues and 5 (7.5%) had different genotypes between tissues. For SNP3, 81 patients had genotype available for both tissues and 9 (11.1%) had different genotypes between tissues. We note that a mutation incidence rate is calculated across patients, which is different from the fraction of mutant alleles calculated within a patient. Since the normal and tumor tissues from a patient may be genetically more similar than between her blood and tumor tissue, the number of differences in genotype between blood and tumor tissue could be higher than those observed above. Therefore these replication results are encouraging and show the potential of our method.

The Sequenom genotype data also allowed us to see if our results at SNP1-SNP3 for patients 1-3 could be validated. As shown in Table V.2, only patient 3 had a base substitution detected at SNP1. This was validated by our Sequenom data: patient 3 was homozygous in the normal tissue and heterozygous in the tumor tissue. For all the other eight combinations of patient and SNP, both normal and tumor tissues were homozygous.

Allelic Imbalance for Base Substitution Sites

As we explained earlier, a tumor sample may contain a mixture of tumor and non-tumor cells, and a somatic mutation may not be present in all tumor cells. Thus, a somatic mutation may be present in only a fraction of cells in a tumor sample. The resulting allelic imbalance can make it inefficient to identify somatic mutations through genotype calls, as was previously shown. We evaluated the extent of allelic imbalance by comparing two types of sites that were called heterozygous in tumor: those with the same heterozygous genotype in blood and those

homozygous in blood. The former should mostly be genuinely heterozygous in tumor with no allelic imbalance. The latter were base substitutions identified by the CALL approach, and should mostly display allelic imbalance if the phenomenon we described above was present.

We calculated the p-value for allelic departure from 50:50 at these sites in tumor. The p-value should follow a uniform distribution over the interval [0,1] if there is no allelic imbalance and should be near zero if allelic imbalance exists. Figure V.1 contains the results for tumor heterozygous sites with depth ≥ 20 for patient 2; the results for the other patients were similar. The distribution for the base substitution sites (right panel) shows strong allelic imbalance, with $p \leq 0.05$ for 87% of the sites. In contrast, among the genuinely heterozygous sites (left panel), only 26% had $p \leq 0.05$; the distribution was consistent with a mixture of sites with no allelic imbalance (78%) and sites with allelic imbalance (22%). We expect these allelic imbalance sites are potential LOH sites, although their LOH scores might not be high enough to pass our LOH detection threshold.

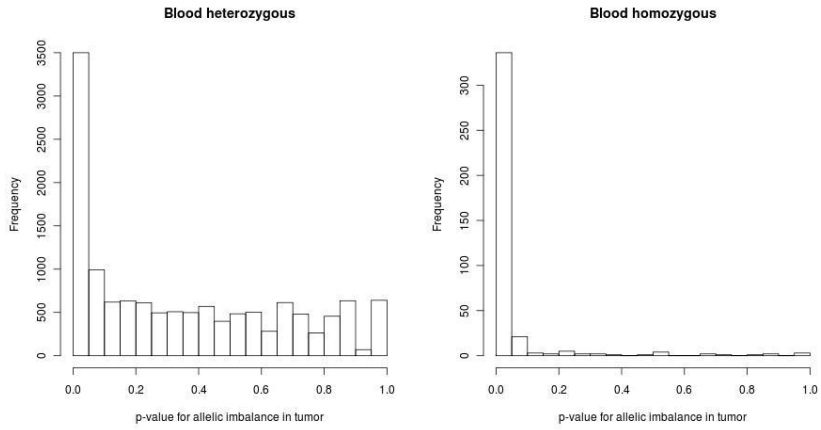

Figure V.1. Allelic imbalance at base substitution sites in tumor samples.

These results demonstrate that most base substitution sites can manifest significant allelic imbalance in tumor samples, at least in our tumor DNA samples. This may generate a

dilemma for the CALL approach since allelic imbalance has been a commonly used tool for genotype quality control. Filtering genotypes based on allelic imbalance may lower the chance of detecting base substitutions by CALL. We have shown that CALL identified 614 base substitution sites in our data. When an allelic balance filter (AB ≤ 0.75) was used, only 310 sites remained. When the allelic balance filter was applied only to blood samples but not tumor samples, 395 sites were identified by CALL.

Detection of Regions with Loss of Heterozygosity

For exome sequencing data, heterozygous sites outside exons may not be reliably called due to gaps of coverage for many genomic regions. In this situation, as was explained in the Methods section, an LOH region may not be detectable due to a lack of supporting sites, and multiple LOH regions may be identified as a single one due to a lack of sites breaking it apart. This is true for both SMUG and CALL. Nonetheless we applied both approaches to our exome data and compared their performance.

For the CALL approach, we first obtained genotype calls for blood and tumor samples. For each patient, her CALL list of LOH sites consisted of sites that were heterozygous in blood and homozygous in tumor. A CALL LOH region was defined as a contiguous region that contained at least two LOH sites and no interrupting site (i.e., sites heterozygous in both blood and tumor) in the region. The boundaries of an LOH region were the positions of its first and last LOH sites.

For the SMUG approach, we first obtained genotype calls for blood samples using the same filtering criteria as in the CALL approach. For each patient, we calculated the tumor LOH score for every site heterozygous in blood, and generated her SMUG list of LOH sites as those sites with LOH score ≥30. A SMUG LOH region was defined as a contiguous region that contained

at least two LOH sites and no interrupting site (i.e., sites heterozygous in blood but with LOH score <30). Its boundaries were defined similarly as above.

SMUG reported an average of 354 LOH regions per patient, while CALL reported 51 (Table V.4). SMUG also reported more LOH regions covering the COSMIC genes (Table V.4). LOH has been widely reported in breast cancer.

Table V.4. Comparison of SMUG and CALL for detection of LOH regions.

Patient	Call			SMUG		
	# of LOH regions	Inside COSMIC genes	Inside support intervals	# of LOH regions	Inside COSMIC genes	Inside support intervals
1	34	3	4	234	28	10
2	37	4	2	314	34	15
3	59	8	5	519	130	30
4	33	3	2	230	26	8
5	43	7	3	321	28	21
6	58	6	4	399	45	21
7	55	4	3	265	23	13
8	87	16	4	554	118	27
Average	50.8	6.4	3.4	354.5	54	18.1

Compiling data from 151 LOH studies of breast cancer, Miller et al. (Miller *et al.* 2003) developed a likelihood-based approach to identify loci with preferential loss. They reported 19 support intervals (Table V.5) that had strong evidence for LOH in breast cancer, and stated that the list was not exhaustive due to the limited resolution of LOH mapping. SMUG identified more LOH regions falling in these intervals than CALL did (Table V.4). As detailed in Table V.5, SMUG identified a total of 145 LOH regions that fell in 12 of the 19 support intervals, while CALL found 27 LOH regions that hit 4 support intervals. For all the support intervals with at least one

identified LOH region, SMUG identified more patients with LOH regions falling in the interval. These results suggested that SMUG is more powerful than CALL at detecting LOH regions.

Table V.5. Comparison of SMUG and CALL for detection of LOH regions in the 19 support intervals reported in Miller et al.

	Support interval (position in Mb)	CALL		SMUG		Candidate TSGs (position in Mb)
		#Patients	#LOH regions	#Patients	#LOH regions	
1p36.3	0-10.5					*SKI* (1.5), *TP73* (3.5)
2q22.1	134.5-137.7					*FLJ11857* (135.5), *LRP1B* (138.5)
3p24.1	22.1-40.4	1	1	4	8	*TGFBR2* (32.6)
3p14.2	59.3-63.6			1	1	*FHIT* (63.5)
4q35.1	182.9-189.2			3	3	*FAT* (187.6)
6q25.1	151.8-170.3	2	2	5	5	*ESR1* (159.3), *IGF2R* (167.3)
7q31.2	115.7-119.0					*TES* (116.2), *CAV1* (116.5), *ST7*/*MET* (117.0)
8p21.3	18.5-23.0					*N33* (15.6), *PDGFRL* (17.8), *LZTS1* (18.6), *FGF17* (20.4), *DBC2* (23.2)
9p21.3	26.2-27.5			1	1	*CDKN2A/p16* (23.8), *IFNA* (24.5)
13q14.11	38.8-42.1			2	2	*BRCA2* (38.2), *AS3* (38.5)
16q22.1	69.4-72.2			4	6	*CDH3* (71.3), *CDH1* (71.5)
16q24.1	72.9-89.8			4	8	*WWOX* (83.9)
16q24.3	90.3-94.3					*CDH13* (88.5), *CBFA2T3* (92.5), *CDH15* (92.7), *FANCA* (93.2)
17p13.3	0-1.5					*ABR* (.8), *OVCA1* (1.7), *HIC1*(1.7)
17p13.2	0-22.8	6	8	8	49	*TP53* (8.0), *MAP2K4* (13.6)
18p11.32	0-3.0	7	16	8	60	*DAL1/EPBB41L3* (5.8)
18q21.2	51.7-54.8					*DPC4* (51.1), *DCC* (53.3)
19p13.3	4.6-6.9			1	1	*SAFB* (7.6)
21q11.1	15.5-18.8			1	1	*BTG3* (15.6)

Discussion

We have developed SMUG, a set of methods for detecting base substitutions and LOH regions using NGS data for normal-tumor pairs. The methods are applicable to both whole genome and whole exome sequencing designs. We demonstrated application of SMUG using whole exome data for blood and tumor samples from eight breast cancer patients, and showed that SMUG detected more somatic mutations than the conventional approach that relies on comparing genotype calls between normal and tumor samples. For three base substitution sites identified by SMUG, we had genotypes for the normal and tumor breast tissues from an independent set of breast cancer patients, which replicated our finding that these sites may be breast cancer somatic mutation sites. The result is encouraging, although further investigation is warranted. We are following them up in a separate study.

Somatic mutations may not be detectable through genotype comparison because a tumor sample may contain a significant fraction of non-tumor cells and a somatic mutation may not be present in all tumor cells of a tumor sample. Although laser capture microdissection can be used to purify the selection of tumor cells in sample preparation, its usage may be limited for certain tissues and the selected tumor cells may still be heterogeneous with respect to somatic mutations.

With multiple sequence reads at each position, the NGS technologies allow a quantitative examination of mutations. While we take advantage of the quantitative nature of NGS data for tumor samples, our methods rely on genotype calls for normal samples. This probably is a good choice when the normal samples are homogeneous. For a homogeneous diploid sample with only three genotypes underlying sequencing data, genotype calling may

serve as a de-noising tool. In fact, current genotype callers such as that in GATK can achieve a very good accuracy.

For detection of base substitutions, we proposed an empirical Bayes approach to remove the effect of depth on estimation of the fraction of mutant alleles. Since systemic sequencing and alignment errors may also appear as base substitutions in tumor, we developed filters that use the normal sample data as a baseline. We also showed that strong allelic imbalance can exist for base substitution sites. This raises an issue for the conventional CALL approach of comparing genotype calls because allelic imbalance has been a common tool for filtering genotypes. If one chooses to use the CALL approach, it might be more effective not to use allelic imbalance as a filtering criterion, at least not for tumor samples. Other genotype filtering tools developed for homogeneous samples also may not work well for tumor samples. One example is transition/transversion ratio, which may have different target values for tumor mutations than those established for germ line polymorphisms.

For detection of LOH regions, we proposed to calculate LOH scores to facilitate identification of LOH regions. Our method focuses on sites that are heterozygous in the normal sample. This probably works well for large LOH regions that harbor multiple heterozygous sites in the normal sample, but it may not be as powerful at detecting short LOH regions. Short LOH regions could be detected by studying depth using algorithms similar to those for detecting CNVs. However, as depth varies along the genome and across samples, appropriate normalization will be needed. We plan to extend our method to take depth into account for detecting LOH regions. The incorporation of depth information may also allow us to distinguish deletion LOH and copy-neutral LOH.

Our methods are especially useful for non-purified tumor samples that contain non-tumor cells. The fraction of non-tumor cells in a tumor sample is often unknown but could be

estimated. Methods have been proposed to use SNP chip data to estimate the fraction of non-tumor cells in a tumor sample. We plan to develop methods for estimating tumor cell fraction using NGS data and to incorporate this information into our mutation detection algorithms.

We have focused on two major types of somatic mutations: base substitution and LOH. Other types of mutations such as indels, inter- and intra-chromosomal rearrangements, and CNVs, can also be similarly detected quantitatively. We are developing methods for detecting these types of mutations. The methods described in the paper have been implemented in software Somatic Mutation Gleaner (SMUG), which is publicly available at http://biostat.mc.vanderbilt.edu/SMUG.

Future Works

There are some limitations of the current implementation of SMUG. For detection of base substitutions, at this point, SMUG cannot detect sites that are homozygous in normal but also homozygous in tumor with different allele. We plan to improve and extend SMUG to deal with this situation.

The detected LOH regions could be incorrect as well. Due to lack of supporting sites that are heterozygous in blood sample and has high LOH score in tumor sample, an LOH region may not be detected. On the other hand, multiple LOH regions may be identified as a single one due to lack of supporting sites breaking them apart. We plan to extend SMUG to take depth into account for detecting LOH regions. The incorporation of local depth information may also allow us to distinguish deletion LOH and copy-neutral LOH regions.

BIBLIOGRAPHY

CCDS. *http://www.ncbi.nlm.nih.gov/projects/CCDS/CcdsBrowse.cgi*.

Picard. *http://picard.sourceforge.net/*.

Adzhubei, I. A., et al. (2010). A method and server for predicting damaging missense mutations. *Nat. Methods* **7**(4): 248-249.

Ameur A., et al. (2011). Ultra-deep sequencing of mouse mitochondrial DNA: mutational patterns and their origins. *PLoS Genet.* **7**(3): 1-9.

Anderson, P. L., et al. (2004). The cellular pharmacology of nucleoside- and nucleotide-analogue reverse-transcriptase inhibitors and its relationship to clinical toxicities. *Clin. Infect. Dis.* **38**(5): 743-753.

Antiretroviral_Guidelines_for_Adults_and_Adolescents_Panel (2006) "Guidelines for the Use of Antiretroviral Agents in HIV-1-infected Adults and Adolescents. www.aidsinfo.nih.gov/ContentFiles/AdultandAdolescentGL.pdf." **October 10**, 11-20.

Atanassova, N., et al. (2011). Sequence-specific stalling of DNA polymerase gamma and the effects of mutations causing progressive ophthalmoplegia. *Hum. Mol. Genet.* **20**(6): 1212-1223.

Azzam, R., et al. (2006). Adverse effects of antiretroviral drugs on HIV-1-infected and -uninfected human monocyte-derived macrophages. *J. Acquir. Immune Defic. Syndr.* **42**(1): 19-28.

Bailey, L. J., et al. (2009). Mice expressing an error-prone DNA polymerase in mitochondria display elevated replication pausing and chromosomal breakage at fragile sites of mitochondrial DNA. *Nucleic Acids Res.* **37**(7): 2327-2335.

Bamford, S., et al. (2004). The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Brit. J. Cancer* **91**(2): 355-358.

Barry, M. G., et al. (1996). The effect of zidovudine dose on the formation of intracellular phosphorylated metabolites. *AIDS* **10**(12): 1361-1367.

Becher, F., et al. (2004). Monitoring of didanosine and stavudine intracellular trisphosphorylated anabolite concentrations in HIV-infected patients. *AIDS* **18**(2): 181-187.

Becher, F., et al. (2002). Improved method for the simultaneous determination of d4T, 3TC and ddI intracellular phosphorylated anabolites in human peripheral-blood mononuclear cells using high-performance liquid chromatography/tandem mass spectrometry. *Rapid. Commun. Mass Spectrom.* **16**(6): 555-565.

Benbrik, E., et al. (1997). Cellular and mitochondrial toxicity of zidovudine (AZT), didanosine (ddI) and zalcitabine (ddC) on cultured human muscle cells. *J. Neurol. Sci.* **149**(1): 19-25.

Bentley, D. R., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature* **456**(7218): 53-59.

Berger, M. F., et al. (2011). The genomic complexity of primary human prostate cancer. *Nature* **470**(7333): 214-220.

Bertram, J. G., et al. (2010). DNA Polymerase Fidelity: Comparing Direct Competition of Right and Wrong dNTP Substrates with Steady State and Pre-Steady State Kinetics. *Biochemistry* **49**(1): 20-28.

Biesecker, G., et al. (2003). Evaluation of mitochondrial DNA content and enzyme levels in tenofovir DF-treated rats, rhesus monkeys and woodchucks. *Antiviral. Res.* **58**(3): 217-225.

Birkus, G., et al. (2002). Assessment of mitochondrial toxicity in human cells treated with tenofovir: Comparison with other nucleoside reverse transcriptase inhibitors. *Antimicrob. Agents. Chemother.* **46**(3): 716-723.

Blecher, E., et al. (2008) "Global Cancer Facts & Figures 2nd Edition."

Bodmer, W. and C. Bonilla (2008). Common and rare variants in multifactorial susceptibility to common diseases. *Nat. Genet.* **40**(6): 695-701.

Boes, M., et al. (2009). Proof of progression over time: Finally fulminant brain, muscle, and liver affection in Alpers syndrome associated with the A467T POLG1 mutation. *Seizure-Eur. J. Epilep.* **18**(3): 232-234.

Bradshaw, P. C. and D. C. Samuels (2005). A computational model of mitochondrial deoxynucleotide metabolism and DNA replication. *Am. J. Physiol. Cell Physiol.* **288**(5): C989-1002.

Brown, T. A., et al. (2005). Replication of mitochondrial DNA occurs by strand displacement with alternative light-strand origins, not via a strand-coupled mechanism. *Genes Dev.* **19**(20): 2466-2476.

Bueno, R., et al. (2010). Second Generation Sequencing of the Mesothelioma Tumor Genome. *Plos One* **5**(5).

Cai, Q., et al. (2011). Genome-wide association study identifies breast cancer risk variant at 10q21.2: results from the Asia Breast Cancer Consortium. *Hum. Mol. Genet.*: September 20. [Epub ahead of print].

Cao, Y. and D. C. Samuels (2009). Discrete Stochastic Simulation Methods for Chemically Reacting Systems. *Methods Enzymol.* **454**: 115-140.

Cargill, M., et al. (1999). Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* **22**(3): 231-238.

Chan, S. S. L. and W. C. Copeland (2009). DNA polymerase gamma and mitochondrial disease: Understanding the consequence of POLG mutations. *BBA-Bioenergetics* **1787**(5): 312-319.

Chan, S. S. L., et al. (2005). The common A467T mutation in the human mitochondrial DNA polymerase (POLG) compromises catalytic efficiency and interaction with the accessory subunit. *J. Biol. Chem.* **280**(36): 31341-31346.

Chan, S. S. L., et al. (2007). Mitochondrial toxicity in hearts of CD-1 mice following perinatal exposure to AZT, 3TC, or AZT/3TC in combination. *Environ.Mol. Mutagen.* **48**(3-4): 190-200.

Cherry, C. L., et al. (2002). Exposure to dideoxynucleosides is reflected in lowered mitochondrial DNA in subcutaneous fat. *J. Acquir. Immune Defic. Syndr.* **30**(3): 271-277.

Cihlar, T., et al. (2002). Tenofovir exhibits low cytotoxicity in various human cell types: comparison with other nucleoside reverse transcriptase inhibitors. *Antiviral. Res.* **54**(1): 37-45.

Cirulli, E. T. and D. B. Goldstein (2010). Uncovering the roles of rare variants in common disease through whole-genome sequencing. *Nat. Rev. Genet.* **11**(6): 415-425.

Clayton, D. A. (1982). Replication of animal mitochondrial-DNA. *Cell* **28**(4): 693-705.

Coen, D. M. and P. A. Schaffer (1980). Two distinct loci confer resistance to acycloguanosine in herpes simplex virus type 1. *Proc. Natl. Acad. Sci. U S A* **77**(4): 2265-2269.

Colacino, J. M. (1996). Mechanisms for the anti-hepatitis B virus activity and mitochondrial toxicity of fialuridine (FIAU). *Antiviral. Res.* **29**(2-3): 125-139.

Copeland, W. C. (2008). Inherited mitochondrial diseases of DNA replication. *Annu. Rev. Med.* **59**: 131-146.

Cote, H. C., et al. (2002). Changes in mitochondrial DNA as a marker of nucleoside toxicity in HIV-infected patients. *N. Engl. J. Med.* **346**(11): 811-820.

Cote, H. C., et al. (2006). Exploring mitochondrial nephrotoxicity as a potential mechanism of kidney dysfunction among HIV-infected patients on highly active antiretroviral therapy. *Antivir. Ther.* **11**(1): 79-86.

Cote, H. C. F., et al. (2003). Mitochondrial : nuclear DNA ratios in peripheral blood cells from human immunodeficiency virus (HIV)-infected patients who received selected HIV antiretroviral drug regimens. *J. Infect. Dis.* **187**(12): 1972-1976.

Cui, L. X., et al. (1997). Effect of nucleoside analogs on neurite regeneration and mitochondrial DNA synthesis in PC-12 cells. *J. Pharmacol. Exp. Ther.* **280**(3): 1228-1234.

d'Arminio Monforte, A., et al. (2000). Insights into the reasons for discontinuation of the first highly active antiretroviral therapy (HAART) regimen in a cohort of antiretroviral naive patients. I.CO.N.A. Study Group. Italian Cohort of Antiretroviral-Naive Patients. *AIDS* **14**(5): 499-507.

de Baar, M. P., et al. (2007). Effects of apricitabine and other nucleoside reverse transcriptase inhibitors on replication of mitochondrial DNA in HepG2 cells. *Antiviral. Res.* **76**(1): 68-74.

Dickson, S. P., et al. (2010). Rare variants create synthetic genome-wide associations. *PLoS Biol.* **8**(1): e1000294.

Ding, L., et al. (2010). Genome remodelling in a basal-like breast cancer metastasis and xenograft. *Nature* **464**(7291): 999-1005.

Eriksson, S., et al. (1995). Efficient incorporation of anti-HIV deoxynucleotides by recombinant yeast mitochondrial DNA polymerase. *J. Biol. Chem.* **270**(32): 18929-18934.

Esteller, M. (2007). Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat. Rev. Genet.* **8**(4): 286-298.

Estep, P. A. and K. A. Johnson (2011). Effect of the Y955C Mutation on Mitochondrial DNA Polymerase Nucleotide Incorporation Efficiency and Fidelity. *Biochemistry* **50**(29): 6376-6386.

Fan, L., et al. (2006). A novel processive mechanism for DNA synthesis revealed by structure, modeling and mutagenesis of the accessory subunit of human mitochondrial DNA polymerase. *J. Mol. Biol.* **358**(5): 1229-1243.

Feng, J. Y. and K. S. Anderson (1999). Mechanistic studies comparing the incorporation of (+) and (-) isomers of 3TCTP by HIV-1 reverse transcriptase. *Biochemistry* **38**(1): 55-63.

Feng, J. Y., et al. (2004). Relationship between antiviral activity and host toxicity: Comparison of the incorporation efficiencies of 2 ',3 '-dideoxy-5-fluoro-3 '-thiacytidine-triphosphate analogs by human immunodeficiency virus type 1 reverse transcriptase and human mitochondrial DNA polymerase. *Antimicrob. Agents. Chemother.* **48**(4): 1300-1306.

Ferrari, G., et al. (2005). Infantile hepatocerebral syndromes associated with mutations in the mitochondrial DNA polymerase-gamma A. *Brain* **128**: 723-731.

Ferraro, P., et al. (2006). Mitochondrial deoxynucleotide pool sizes in mouse liver and evidence for a transport mechanism for thymidine monophosphate. *Proc. Natl. Acad. Sci. U S A* **103**(49): 18586-18591.

Ferraro, P., et al. (2005). Mitochondrial deoxynucleotide pools in quiescent fibroblasts - A possible model for mitochondrial neurogastrointestinal encephalomyopathy (MNGIE). *J. Biol. Chem.* **280**(26): 24472-24480.

Forbes, S. A., et al. (2011). COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**: D945-D950.

Friis, M. B., et al. (2005). Cell shrinkage as a signal to apoptosis in NIH 3T3 fibroblasts. *J. Physiol.* **567**(Pt 2): 427-443.

Galluzzi, L., et al. (2005). Changes in mitochondrial RNA production in cells treated with nucleoside analogues. *Antivir. Ther.* **10**(1): 191-195.

Gao, Y. T., et al. (2000). Association of menstrual and reproductive factors with breast cancer risk: Results from the Shanghai Breast Cancer Study. *Int. J. Cancer* **87**(2): 295-300.

Gillespie, D. T. (1976). General method for numerically simulating stochastic time evolution of coupled chemical-reactions. *J. Comput. Phys.* **22**(4): 403-434.

Gillespie, D. T. (1977). Exact stochastic simulation of coupled chemical-reactions. *J. Phys. Chem.* **81**(25): 2340-2361.

Gnirke, A., et al. (2009). Solution hybrid selection with ultra-long oligonucleotides for massively parallel targeted sequencing. *Nat. Biotechnol.* **27**(2): 182-189.

Gorlov, I. P., et al. (2008). Shifting paradigm of association studies: value of rare single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **82**(1): 100-112.

Graziewicz, M. A., et al. (2007). The DNA polymerase gamma Y955C disease variant associated with PEO and parkinsonism mediates the incorporation and translesion synthesis opposite 7,8-dihydro-8-oxo-2'-deoxyguanosine. *Hum. Mol. Genet.* **16**(22): 2729-2739.

Graziewicz, M. A., et al. (2004). Structure-function defects of human mitochondrial DNA polymerase in autosomal dominant progressive external ophthalmoplegia. *Nat. Struct. Mol. Biol.* **11**(8): 770-776.

Graziewicz, M. A., et al. (2006). DNA polymerase gamma in mitochondrial DNA replication and repair. *Chem. Rev.* **106**(2): 383-405.

Hanes, J. W. and K. A. Johnson (2007). A novel mechanism of selectivity against AZT by the human mitochondrial DNA polymerase. *Nucleic Acids Res.* **35**(20): 6973-6983.

Hanes, J. W., et al. (2007). Enzymatic therapeutic index of acyclovir: Viral versus human polymerase gamma specificity. *J. Biol. Chem.* **282**(34): 25159-25167.

Haugaard, S. B., et al. (2005). Depleted skeletal muscle mitochondrial DNA, hyperlactatemia, and decreased oxidative capacity in HIV-infected patients on highly active antiretroviral therapy. *J. Med. Virol.* **77**(1): 29-38.

Hobbs, G. A., et al. (1995). Cellular targets of 3'-azido-3'-deoxythymidine: an early (non-delayed) effect on oxidative phosphorylation. *Biochem. Pharmacol.* **50**(3): 381-390.

Holt, I. J., et al. (2000). Coupled leading- and lagging-strand synthesis of mammalian mitochondrial DNA. *Cell* **100**(5): 515-524.

Horvath, R., et al. (2006). Phenotypic spectrum associated with mutations of the mitochondrial polymerase gamma gene. *Brain* **129**: 1674-1684.

Hudson, G. and P. F. Chinnery (2006). Mitochondrial DNA polymerase-gamma and human disease. *Hum. Mol. Genet.* **15**: R244-R252.

Johnson, A. A. and K. A. Johnson (2001). Exonuclease proofreading by human mitochondrial DNA polymerase. *J. Biol. Chem.* **276**(41): 38097-38107.

Johnson, A. A. and K. A. Johnson (2001). Fidelity of nucleotide incorporation by human mitochondrial DNA polymerase. *J. Biol. Chem.* **276**(41): 38090-38096.

Johnson, A. A., et al. (2001). Toxicity of antiviral nucleoside analogs and the human mitochondrial DNA polymerase. *J. Biol. Chem.* **276**(44): 40847-40857.

Johnson, A. A., et al. (2000). Human mitochondrial DNA polymerase holoenzyme: Reconstitution and characterization. *Biochemistry* **39**(7): 1702-1708.

Kaguni, L. S. (2004). DNA polymerase gamma, the mitochondrial replicase. *Annu. Rev. Biochem.* **73**: 293-320.

Kang, J. and D. C. Samuels (2008). The evidence that the DNC (SLC25A19) is not the mitochondrial deoxyribonucleotide carrier. *Mitochondrion* **In Press**.

Karras, A., et al. (2003). Tenofovir-related nephrotoxicity in human immunodeficiency virus-infected patients: three cases of renal failure, Fanconi syndrome, and nephrogenic diabetes insipidus. *Clin. Infect. Dis.* **36**(8): 1070-1073.

Kewn, S., et al. (2002). Development of enzymatic assays for quantification of intracellular lamivudine and carbovir triphosphate levels in peripheral blood mononuclear cells from human immunodeficiency virus-infected patients. *Antimicrob. Agents. Chemother.* **46**(1): 135-143.

Koboldt, D. C., et al. (2009). VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* **25**(17): 2283-2285.

Korhonen, J. A., et al. (2004). Reconstitution of a minimal mtDNA replisome in vitro. *EMBO J.* **23**(12): 2423-2429.

Krishnan, K. J., et al. (2008). What causes mitochondrial DNA deletions in human cells? *Nat. Genet.* **40**(3): 275-279.

Kryukov, G. V., et al. (2007). Most rare missense alleles are deleterious in humans: implications for complex disease and association studies. *Am. J. Hum. Genet.* **80**(4): 727-739.

Kujoth, G. C., et al. (2005). Mitochondrial DNA mutations, oxidative stress, and apoptosis in mammalian aging. *Science* **309**(5733): 481-484.

Le Novere, N. and T. S. Shimizu (2001). STOCHSIM: modelling of stochastic biomolecular processes. *Bioinformatics* **17**(6): 575-576.

Lee, H. R. and K. A. Johnson (2006). Fidelity of the human mitochondrial DNA polymerase. *J. Biol. Chem.* **281**(47): 36236-36240.

Lee, W., et al. (2010). The mutation spectrum revealed by paired genome sequences from a lung cancer patient. *Nature* **465**(7297): 473-477.

Lewis, W., et al. (2003). Mitochondrial toxicity of NRTI antiviral drugs: An integrated cellular perspective. *Nat. Rev. Drug Discov.* **2**(10): 812-822.

Lewis, W., et al. (2006). Antiretroviral nucleosides, deoxynucleotide carrier and mitochondrial DNA: evidence supporting the DNA pol gamma hypothesis. *AIDS* **20**(5): 675-684.

Lewis, W., et al. (1996). Fialuridine and its metabolites inhibit DNA polymerase gamma at sites of multiple adjacent analog incorporation, decrease mtDNA abundance, and cause

mitochondrial structural defects in cultured hepatoblasts. *Proc. Natl. Acad. Sci. U S A* **93**(8): 3592-3597.

Li, B. and S. M. Leal (2009). Discovery of rare variants via sequencing: implications for the design of complex trait association studies. *PLoS Genet.* **5**(5): e1000481.

Li, H., et al. (2008). Algorithms and software for stochastic simulation of biochemical reacting systems. *Biotechnol. Prog.* **24**(1): 56-61.

Li, H. and R. Durbin (2009). Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**(14): 1754-1760.

Li, H., et al. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**(16): 2078-2079.

Lim, S. E., et al. (1999). The mitochondrial p55 accessory subunit of human DNA polymerase gamma enhances DNA binding, promotes processive DNA synthesis, and confers N-ethylmaleimide resistance. *J. Biol. Chem.* **274**(53): 38197-38203.

Lindhurst, M. J., et al. (2006). Knockout of Slc25a19 causes mitochondrial thiamine pyrophosphate depletion, embryonic lethality, CNS malformations, and anemia. *Proc. Natl. Acad. Sci. U S A* **103**(43): 15927-15932.

Liu, D. J. and S. M. Leal (2010). Replication Strategies for Rare Variant Complex Trait Association Studies via Next-Generation Sequencing. *Am. J. Hum. Genet.* **87**(6): 790-801.

Long, J. R., et al. (2010). Identification of a Functional Genetic Variant at 16q12.1 for Breast Cancer Risk: Results from the Asia Breast Cancer Consortium. *Plos Genetics* **6**(6).

Longley, M. J., et al. (2005). Consequences of mutations in human DNA polymerase gamma. *Gene* **354**: 125-131.

Longley, M. J. and D. W. Mosbaugh (1991). Properties of the 3' to 5' exonuclease associated with porcine liver DNA polymerase gamma. Substrate specificity, product analysis, inhibition, and kinetics of terminal excision. *J. Biol. Chem.* **266**(36): 24702-24711.

Lund, K. C., et al. (2007). Absence of a universal mechanism of mitochondrial toxicity by nucleoside analogs. *Antimicrob. Agents. Chemother.* **51**(7): 2531-2539.

Luoma, P. T., et al. (2005). Functional defects due to spacer-region mutations of human mitochondrial DNA polymerase in a family with an ataxia-myopathy syndrome. *Hum. Mol. Genet.* **14**(14): 1907-1920.

Lynx, M. D., et al. (2006). 3'-Azido-3'-deoxythymidine (AZT) inhibits thymidine phosphorylation in isolated rat liver mitochondria: a possible mechanism of AZT hepatotoxicity. *Biochem. Pharmacol.* **71**(9): 1342-1348.

Madsen, B. E. and S. R. Browning (2009). A Groupwise Association Test for Rare Mutations Using a Weighted Sum Statistic. *Plos Genetics* **5**(2).

Maher, B. (2008). Personal genomes: The case of the missing heritability. *Nature* **456**(7218): 18-21.

Mallon, P. W., et al. (2005). In vivo, nucleoside reverse-transcriptase inhibitors alter expression of both mitochondrial and lipid metabolism genes in the absence of depletion of mitochondrial DNA. *J. Infect. Dis.* **191**(10): 1686-1696.

Manolio, T. A., et al. (2008). A HapMap harvest of insights into the genetics of common disease. *J. Clin. Invest.* **118**(5): 1590-1605.

Manolio, T. A., et al. (2009). Finding the missing heritability of complex diseases. *Nature* **461**(7265): 747-753.

Mardis, E. R., et al. (2009). Recurring Mutations Found by Sequencing an Acute Myeloid Leukemia Genome. *N. Engl. J. Med.* **361**(11): 1058-1066.

Martin, J. L., et al. (1994). Effects of antiviral nucleoside analogs on human DNA polymerases and mitochondrial DNA synthesis. *Antimicrob. Agents. Chemother.* **38**(12): 2743-2749.

Martinez-Irujo, J. J., et al. (1998). Analysis of the combined effect of two linear inhibitors on a single enzyme. *Biochem. J.* **329**: 689-698.

McCarthy, M. I., et al. (2008). Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat. Rev. Genet.* **9**(5): 356-369.

McComsey, G., et al. (2005). Extensive investigations of mitochondrial DNA genome in treated HIV-infected subjects: beyond mitochondrial DNA depletion. *J. Acquir. Immune Defic. Syndr.* **39**(2): 181-188.

McHugh, J. C., et al. (2010). Sensory ataxic neuropathy dysarthria and ophthalmoparesis (sando) in a sibling pair with a homozygous p.A467T POLG mutation. *Muscle Nerve* **41**(2): 265-269.

McKenna, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**(9): 1297-1303.

Miller, B. J., et al. (2003). Pooled analysis of loss of heterozygosity in breast cancer: a genome scan provides comparative evidence for multiple tumor suppressors and identifies novel candidate regions. *Am. J. Hum. Genet.* **73**(4): 748-767.

Moore, K. H., et al. (1999). The pharmacokinetics of lamivudine phosphorylation in peripheral blood mononuclear cells from patients infected with HIV-1. *AIDS* **13**(16): 2239-2250.

Naviaux, R. K. and K. V. Nguyen (2004). POLG mutations associated with Alpers' syndrome and mitochondrial DNA depletion. *Ann. Neurol.* **55**(5): 706-712.

Nguyen, K. V., et al. (2005). POLG mutations in Alpers syndrome. *Neurology* **65**(9): 1493-1495.

Note, R., et al. (2003). Mitochondrial and metabolic effects of nucleoside reverse transcriptase inhibitors (NRTIs) in mice receiving one of five single- and three dual-NRTI treatments. *Antimicrob. Agents. Chemother.* **47**(11): 3384-3392.

Pan-Zhou, X. R., et al. (2000). Differential effects of antiretroviral nucleoside analogs on mitochondrial function in HepG2 cells. *Antimicrob. Agents. Chemother.* **44**(3): 496-503.

Patel, S. S., et al. (1991). PRE-STEADY-STATE KINETIC-ANALYSIS OF PROCESSIVE DNA-REPLICATION INCLUDING COMPLETE CHARACTERIZATION OF AN EXONUCLEASE-DEFICIENT MUTANT. *Biochemistry* **30**(2): 511-525.

Pereira, L. F., et al. (1998). Mitochondrial sensitivity to AZT. *Cell Biochem. Funct.* **16**(3): 173-181.

Piechota, J., et al. (2006). Nuclear and mitochondrial genome responses in HeLa cells treated with inhibitors of mitochondrial DNA expression. *Acta Biochim. Pol.* **53**(3): 485-495.

Piliero, P. J. (2004). Pharmacokinetic properties of nucleoside/nucleotide reverse transcriptase inhibitors. *J. Acquir. Immune Defic. Syndr.* **37 Suppl 1**: S2-S12.

Pleasance, E. D., et al. (2010). A comprehensive catalogue of somatic mutations from a human cancer genome. *Nature* **463**(7278): 191-U173.

Pleasance, E. D., et al. (2010). A small-cell lung cancer genome with complex signatures of tobacco exposure. *Nature* **463**(7278): 184-U166.

Pollak, J. K. and E. A. Munn (1970). The isolation by isopycnic density-gradient centrifugation of two mitochondrial populations from livers of embryonic and fed and starved adult rats. *Biochem. J.* **117**(5): 913-919.

Ponamarev, M. V., et al. (2002). Active site mutation in DNA polymerase gamma associated with progressive external ophthalmoplegia causes error-prone DNA synthesis. *J. Biol. Chem.* **277**(18): 15225-15228.

Ponamarev, M. V., et al. (2002). Active site mutation in DNA polymerase gamma associated with progressive external ophthalmoplegia causes error-prone DNA synthesis. *Journal of Biological Chemistry* **277**(18): 15225-15228.

Pritchard, J. K. (2001). Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* **69**(1): 124-137.

Pruvost, A., et al. (2005). Measurement of intracellular didanosine and tenofovir phosphorylated metabolites and possible interaction of the two drugs in human immunodeficiency virus-infected patients. *Antimicrob. Agents. Chemother.* **49**(5): 1907-1914.

Purcell, S. M., et al. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**(7256): 748-752.

Pursell, Z. F., et al. (2008). Trace amounts of 8-oxo-dGTP in mitochondrial dNTP pools reduce DNA polymerase gamma replication fidelity. *Nucleic Acids Res.* **36**(7): 2174-2181.

Rampazzo, C., et al. (2004). Mitochondrial deoxyribonucleotides, pool sizes, synthesis, and regulation. *J. Biol. Chem.* **279**(17): 17019-17026.

Ray, A. S. (2005). Intracellular interactions between nucleos(t)ide inhibitors of HIV reverse transcriptase. *AIDS Rev.* **7**(2): 113-125.

Reich, D. E. and E. S. Lander (2001). On the allelic spectrum of human disease. *Trends Genet.* **17**(9): 502-510.

Saada, A. (2004). Deoxyribonucleotides and disorders of mitochondrial DNA integrity. *DNA Cell Biol.* **23**(12): 797-806.

Schaffner, S. F., et al. (2005). Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* **15**(11): 1576-1583.

Schork, N. J., et al. (2009). Common vs. rare allele hypotheses for complex diseases. *Curr. Opin. Genet. Dev.* **19**(3): 212-219.

Service, R. F. (2006). Gene sequencing. The race for the $1000 genome. *Science* **311**(5767): 1544-1546.

Setzer, B., et al. (2005). Mitochondrial toxicity of nucleoside analogues in primary human lymphocytes. *Antivir. Ther.* **10**(2): 327-334.

Shadel, G. S. and D. A. Clayton (1997). Mitochondrial DNA maintenance in vertebrates. *Annu. Rev. Biochem.* **66**: 409-435.

Shah, S. P., et al. (2009). Mutational evolution in a lobular breast tumour profiled at single nucleotide resolution. *Nature* **461**(7265): 809-U867.

Singer, E. (2011). Cancer's Genome. *Technol. Rev.* **114**(1): 46-50.

Song, Z. and D. C. Samuels (2010). Analysis of enzyme kinetic data for mtDNA replication. *Methods* **51**(4): 385-391.

Stankov, M. V., et al. (2007). Relationship of mitochondrial DNA depletion and respiratory chain activity in preadipocytes treated with nucleoside reverse transcriptase inhibitors. *Antivir. Ther.* **12**(2): 205-216.

Stratton, M. R. (2011). Exploring the Genomes of Cancer Cells: Progress and Promise. *Science* **331**(6024): 1553-1558.

Stratton, M. R., et al. (2009). The cancer genome. *Nature* **458**(7239): 719-724.

Szabados, E., et al. (1999). Role of reactive oxygen species and poly-ADP-ribose polymerase in the development of AZT-induced cardiomyopathy in rat. *Free Radic. Biol. Med.* **26**(3-4): 309-317.

Talbot, S. J. and D. H. Crawford (2004). Viruses and tumours - an update. *Eur. J. Cancer* **40**(13): 1998-2005.

Thomas, D. C., et al. (2009). Methodological Issues in Multistage Genome-Wide Association Studies. *Stat. Sci.* **24**(4): 414-429.

Timmermann, B., et al. (2010). Somatic Mutation Profiles of MSI and MSS Colorectal Cancer Identified by Whole Exome Next Generation Sequencing and Bioinformatics Analysis. *Plos One* **5**(12).

Tishkoff, S. A., et al. (2009). The Genetic Structure and History of Africans and African Americans. *Science* **324**(5930): 1035-1044.

Trifunovic, A., et al. (2004). Premature ageing in mice expressing defective mitochondrial DNA polymerase. *Nature* **429**(6990): 417-423.

Utzig, N., et al. (2007). Clinical course of a boy with Alpers syndrome due to homozygosity for A467T mutation in POLG gene. *Acta Neuropathol.* **114**(3): 71.

Valenti, D., et al. (2000). AZT inhibition of the ADP/ATP antiport in isolated rat heart mitochondria. *Int J Mol Med* **6**(1): 93-96.

Van Goethem, G., et al. (2001). Mutation of POLG is associated with progressive external ophthalmoplegia characterized by mtDNA deletions. *Nat. Genet.* **28**(3): 211-212.

Van Goethem, G., et al. (2004). POLG mutations in neurodegenerative disorders with ataxia but no muscle involvement. *Neurology* **63**(7): 1251-1257.

Varela, I., et al. (2011). Exome sequencing identifies frequent mutation of the SWI/SNF complex gene PBRM1 in renal carcinoma. *Nature* **469**(7331): 539-542.

Venhoff, N., et al. (2007). Mitochondrial toxicity of tenofovir, emtricitabine and abacavir alone and in combination with additional nucleoside reverse transcriptase inhibitors. *Antivir. Ther.* **12**(7): 1075-1085.

Vermulst, M., et al. (2008). DNA deletions and clonal mutations drive premature aging in mitochondrial mutator mice. *Nat. Genet.* **40**(4): 392-394.

Vermulst, M., et al. (2009). On Mitochondria, Mutations, and Methodology. *Cell Metab.* **10**(6): 437-437.

Vidal, F., et al. (2006). In vitro cytotoxicity and mitochondrial toxicity of tenofovir alone and in combination with other antiretrovirals in human renal proximal tubule cells. *Antimicrob. Agents. Chemother.* **50**(11): 3824-3832.

Walker, U. A., et al. (2004). Depletion of mitochondrial DNA in liver under antiretroviral therapy with didanosine, stavudine, or zalcitabine. *Hepatology* **39**(2): 311-317.

Walker, U. A., et al. (2002). Increased long-term mitochondrial toxicity in combinations of nucleoside analogue reverse-transcriptase inhibitors. *AIDS* **16**(16): 2165-2173.

Wang, K., et al. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res.* **38**(16): 7.

Wendelsdorf, K. V., et al. (2009). An analysis of enzyme kinetics data for mitochondrial DNA strand termination by nucleoside reverse transcription inhibitors. *PLoS Comput. Biol.* **5**(1): 11.

Williams, S. L., et al. (2010). The mtDNA mutation spectrum of the progeroid Polg mutator mouse includes abundant control region multimers. *Cell Metab.* **12**(6): 675-682.

Winterthun, S., et al. (2005). Autosomal recessive mitochondrial ataxic syndrome due to mitochondrial polymerase gamma mutations. *Neurology* **64**(7): 1204-1208.

Wong, G. K., et al. (2003). A population threshold for functional polymorphisms. *Genome Res.* **13**(8): 1873-1879.

Yakubovskaya, E., et al. (2006). Functional human mitochondrial DNA polymerase gamma forms a heterotrimer. *J. Biol. Chem.* **281**(1): 374-382.

Yasukawa, T., et al. (2006). Replication of vertebrate mitochondrial DNA entails transient ribonucleotide incorporation throughout the lagging strand. *EMBO J.* **25**(22): 5358-5371.

Young, M. J., et al. (2009). Biochemical analysis of POLG2 variants associated with mitochondrial disease. *Mitochondrion* **10**(2): 96.

Zheng, W., et al. (2009). Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.* **41**(3): 324-328.

Zimmermann, A. E., et al. (2006). Tenofovir-associated acute and chronic kidney disease: a case of multiple drug interactions. *Clin. Infect. Dis.* **42**(2): 283-290.