Plan Overview

A Data Management Plan created using DMPTool

Title: Mechanisms mitigating replicative DNA polymerase defects

Creator: Youri Pavlov

Affiliation: University of Nebraska Medical Center (unmc.edu)

Funder: National Institutes of Health (nih.gov)

Funding opportunity number: PAR-23-058

Grant: https://grants.nih.gov/grants/guide/pa-files/PAR-23-058.html

Template: NIH-Default DMSP

Project abstract:

DNA polymerases of the B-family play a central role in accurate genome duplication in eukaryotes. Alteration of replication machinery plays a significant role in cancer etiology. Mutations reducing the fidelity or changing replication parameters impose a high cancer risk. DNA polymerase ε (pol ε) is part of a multisubunit replicative helicase complex and contributes to the leading DNA strand synthesis. The inhibition of the catalytic subunit of pol ε, POLE, in basal-like breast carcinomas leads to rapid tumor cell decay. A specific characteristic of sensitive cancer cells is the hyperactivity of cyclin-dependent kinase CDK2, which regulates the replication initiation. The exact mechanism of the effect is unknown. We plan to investigate the mechanism of how CDKs control pol ε and other pols during DNA replication in the yeast model.

We recently found in yeast that the defect of polymerase activity of pol ε conferred by the absence of catalytically active N-terminal half of the Pol2 subunit is mitigated by mutations in the CDC28 gene, a homolog of human CDKs. It is possible that when part of Pol2 is missing, another DNA pol, pol δ, can substitute for pol ε activity. Cdc28 mutations may facilitate the recruitment of pol δ to synthesize the leading DNA strand. To decipher the suppression mechanism, we will investigate how cdc28 mutations affect replication, explore if suppression is specific for pol ε only, and find additional genetic factors that help cells overcome replication defects.

In Specific Aim 1, to explore the mechanism of suppression of pol ε defects by mutant Cdc28, we will perform a vertical scan for new suppressor mutations in the CDC28 gene to locate protein regions responsible for the

suppression. To find if the polymerase or exonuclease activity defects could be suppressed or if the effect is confined to the variant missing the N-terminal half of Pol2, we will examine which particular mutations in the POL2 gene are suppressed by cdc28 mutations. We will next characterize what function of Cdc28 is affected by suppressor changes; are they confer hyperactivity or hypoactivity? In Specific Aim 2, we will expand the analysis of the intertwining of replication and cell cycle control by examining if variants of Cdc28 suppress replication defects caused by mutants of pol α and pol δ and screen for additional suppressors of replication defects.

Start date: 07-01-2024

Last modified: 09-27-2023

Copyright information:

Mechanisms mitigating replicative DNA polymerase defects

We will investigate how CDK mutations affect replication. We will use baker's yeast model system. The project will generate novel yeast strains whose genomes will be sequenced by next generation sequencing (NGS). The following data files will be used or produced in the course of the project: a) the raw fastq instrument output files and “processed” fastq files representing demultiplexed, barcode-stripped, error-corrected sequencing reads. Raw data will be transformed to contig and contig scaffolds (draft genome assemblies). Depending on the assembler, contigs may have a “base call error” probability associated with each nucleotide in the contig’s DNA sequence. Contigs are typically “fasta” files (+/- associated “quality” files). Scaffold information or the scaffolds themselves are the collection of contigs with assigned chromosomal coordinates and orientations, with deduced gaps. Alternatively, scaffolds may order and orient contigs into chromosomes and fill-in predicted gaps with “N” symbols. a) Read alignments and variation detection data. Error-corrected reads from clones selected for sequencing will be aligned to the draft genome sequence of the relevant isogenic founder strain/line, and the resulting files used for the detection of sequence variants. Typically, these alignment and variation detection data are obtained in the form of binary “BAM” (binary form of the “sequence alignment/map” or SAM format) and “VCF” (variant call format) files, respectively, although there are several other formats for these types of files. As we are generating data from model organism (brewer’s yeast) concerns about privacy, confidentiality, and public security are not applicable.

We will generate an information how genotypic variation affects DNA replication. In addition to making all of the data produced by this study freely available to the scientific community, an equally important objective of our study will be to make all of our NGS data analysis workflows, from raw data to variant calls and their annotation and interpretation, reproducible and freely available via public NGS data analysis cloud computing/service providers. This will give our study an exceptionally important broader impact on mutagenesis research and teaching, as it will enable researchers and students to reproduce and extend our research findings.

To facilitate the interpretation and reuse of the data, a README file and data dictionary will be generated and deposited into a repository along with all shared datasets. The README file will include method description, instrument settings, RRIDs of resources such as antibodies, model organisms, cell lines, plasmids, and other tools (e.g., software, databases, services), and Protocol DOIs issued from protocols.io. The data dictionary will define and describe all variables in the dataset.

For the analysis, we will use the following Bioinformatics tools.

Data processing. Raw reads of the five yeast strains will be filtered by AfterQC (v0.9.7), adapters, primers sequences, and low-quality nucleotides ($Q < 20$) removed. Reads shorter than 35 nucleotides after trimming will be discarded. Raw and filtered sequences will be explored with FastQC to calculate and visualize sequence quality metrics. The resulting filtered reads will be then aligned to the S288c reference (R64-3-1) genome using bwa-mem (v0.7.17) with default parameters. Samtools (v.1.17) and picard (v2.27.5) will be used to index the reference and create a dictionary. Mapped reads will be sorted with samtools (v1.17). Duplicated reads will be marked with Picard MarkDuplicates. Alignment statistics were assessed using picard and samtools. The alignment results will be used to identify SNPs and indels within each genome by GATK HaplotypeCaller (v4.3.0.0) with sample_ploidy set to 1 or 2 depending on yeast strain ploidy. GATK VariantFiltration will be used to filter the vcf files. The variant call set will be filtered by excluding multiallelic variants, thresholds DP <10 for read depth, and allele depth AD <5 for the alternative allele by bcftools (v1.17). Variants will be annotated using snpEFF (v5.1) and Ensembl Variant Effect Predictor (VEP) (v109). Genomic information will be visualized with CoMut [69].

These tools are free-distributed software. Our policy will be to make all reasonable attempts to deposit the data from this study in public databases and make it freely available to the scientific community as soon as the results are published.

In accordance with FAIR Principles for data, we will use open file formats (e.g. JPEG, DOC, VCF, CSV, TXT, PDF, HTML, BAM) and persistent unique identifiers (PIDs) for resources (e.g., organisms, plasmids, antibodies, cell lines, software tools, and databases) and DOIs for protocols using protocols.io. Our policy will be to make all reasonable attempts to deposit the data from this study in public databases and make it freely available to the scientific community as soon as the results are published.

Aggregate NGS data of the study will be deposited into Sequence Read Archive. DNA constructs, proteins and yeast strains generated by this project will be made freely available to the broaderresearch community, either before or immediately after publication. We will maintain reagents under conditions that prevent contamination and that preserve their value as unique research resources. We will assume responsibility for distributing the newly generated reagents, and will fill requests in a timely fashion and at no cost, except forstandard maintenance and transportation expenses. In addition, we will provide relevant protocols and published genetic and phenotypic data upon request. Should any intellectual property arise which requires a patent, we will ensure that the technology (materials and data) remains widely available to the research community in accordance with the NIH Principles and Guidelines document.

We will use Persistent Unique Identifiers (PIDs) to improve data findability across all dissemination outputs. PIDs used will include ORCID iDs for people, DOIs for outputs (e.g., datasets, protocols), Research Resource IDentifiers (RRIDs) for resources, and Research Organization Registry (ROR) IDs and funder IDs for places, as much as possible to make data identifiable and findable. We will also use indexed metadata, such as MeSH terms with a unique URL to make scientific data easily findable. We will keep our ORCID Records up to date with DOIs for our datasets and publications, ROR, and funder IDs to increase findability.

All scientific data generated from this project will be made available as soon as possible, and no later than the time of publication or the end of the funding period, whichever comes first. The duration of preservation and sharing of the data will be a minimum of 10 years after the funding period.

We will not use human subjects or animal models in our study.

Data will be available with our restrictions , because we use yeast model.

n/a, no human subjects.

Lead PI, Youri Pavlov, ORCID:0000-0003-1179-5796, will be responsible for the day-to-day oversight of lab/team data management activities and data sharing. Broader issues of DMS Plan compliance oversight and reporting will be handled by the PI as part of general [campus(es)] stewardship, reporting, and compliance processes.