Novel DNA Standards for Assessing Technical Sensitivity and Reproducibility of Duplex Sequencing Mutagenesis Assays

Devon M. Fitzgerald[1], Jake Higgins[1], and Jesse Salk[1]
[1]TwinStrand Biosciences, Inc., Seattle, WA

TWINSTRAND® BIOSCIENCES

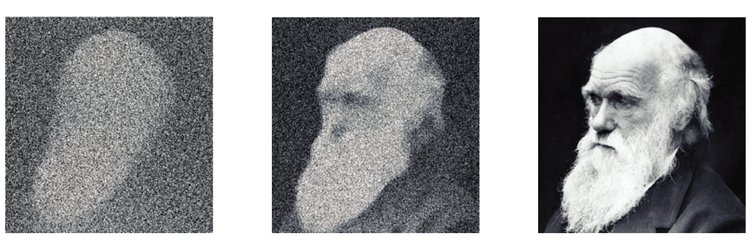
Mutagenesis Applications
(https://twinstrandbio.com)

Introduction

Duplex Sequencing (DS) uses double-stranded error correction to accurately detect and characterize ultra-rare somatic mutations, such as those caused by genotoxic exposures. DS has been used to assess the mutagenicity of a wide variety of compounds and shows high concordance with gold-standard transgenic rodent assays.

Here, we describe the creation of mixture-based DNA standards for technical validation of DS genomic safety assays. Experiments with these standards confirm their performance, show high intra-lab repeatability, and demonstrate the unprecedented technical sensitivity of DS.

Sequencing Errors Obscure Truth



Next-Generation Sequencing (NGS)

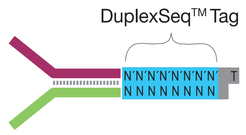
Single Strand Error-Corrected NGS

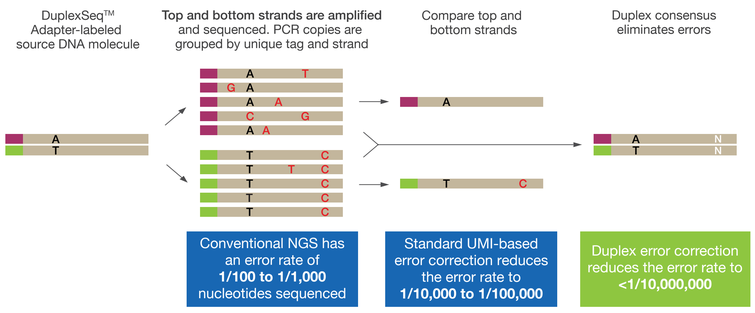
Duplex Sequencing

TwinStrand Duplex Sequencing® Technology

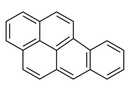
A DuplexSeq™ Adapter has:

1. Identical (or relatable) degenerate tags in each strand.
2. An asymmetry allowing independent strand identification.

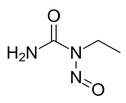




Methods: mutagens & DNA mixture preparation



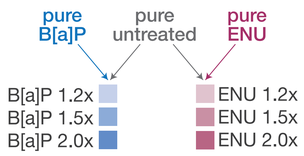
B[a]P
- By-product of burning
- Metabolized to epoxide which forms adducts at G



ENU
- "Model mutagen" and common positive control
- Direct alkylating agent

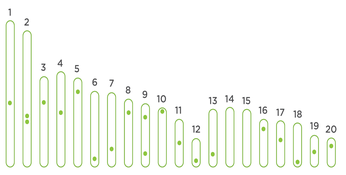
DS was performed on liver DNA from untreated rats and rats mutagenized with either B[a]P or ENU (1 animal each) to determine the mutation frequencies (MFs) of each pure sample.

DNA mixtures were made targeting mutation frequencies **1.2x** (20%), **1.5x** (50%), and **2x** (100%) higher than the untreated control, as illustrated (right).



Mixture calculations: $MF_{mixture} = (MF_{treated} - MF_{untreated}) * fraction_{treated} + MF_{untreated}$

Methods: DS library construction



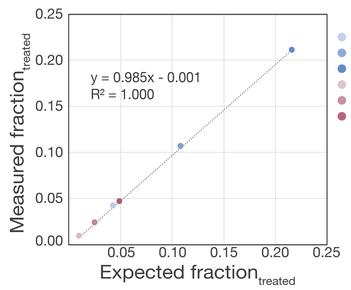
Rat Mutagenesis Panel:
- 20 x 2.4 kb regions throughout genome
- genic and intergenic regions included
- regions not predicted to be involved in positive or negative selection
- similar panels availible for **mouse** and **human**

DS libraries were prepared with the **TwinStrand DuplexSeq™ Mutagenesis Assay** (rat), using **1 ug** of **enzymatically fragmented** input DNA. Three to four technical replicate libraries were prepared for each DNA sample (pure components and mixtures).

Libraries were pooled and sequenced on one NovaSeq 6000 S4 flow cell and then downsampled to ~400 M PE reads per library prior to analysis. For some plots and analyses, consensus reads from technical replicate libraries were combined prior to variant calling.

Key Performance Metrics:
- **PE reads per library:** 396 M +/- 0.6 M
- **Mean duplex depth:** 31,000 +/- 1,200
- **Informative duplex bases:** 1.9 B +/- 0.08 B
- **Peak Tag Family Size:** 11+/- 1

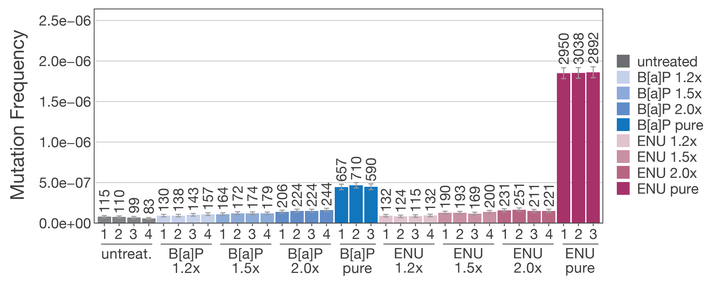
DNA mixture confirmation: germline variants



Germline variant profiles were used to assess whether mixtures were prepared correctly. A tool designed to detect accidental cross-contamination between samples was used to measure the fraction of DNA from B[a]P- or ENU-treated rats in each mixture, using combined data from four technical replicates per mixture.

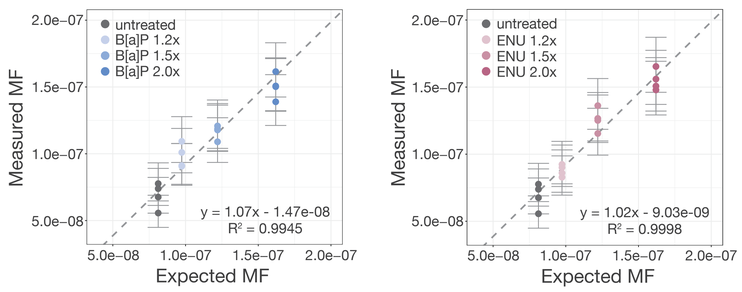
Each point represents one mixture, with the **expected** vs. **measured fraction**$_{treated}$ plotted on the x- and y-axes, respectively. The best-fit line and R^2 value confirm that the mixtures were prepared accurately.

Mutation frequency overview



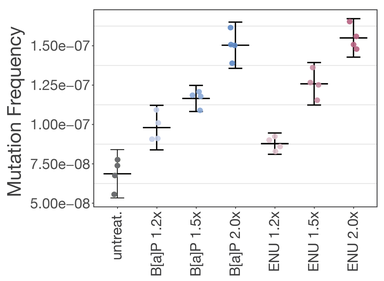
Bars represent **per-library mutation frequencies**, calculated using the minimum assumption (# unique variants divided by # error-corrected bases sequenced). Numbers above each bar are the number of unique variants detected in the library and error bars show Wilson binomial confidence intervals (95%).

Mutation frequency: linearity across mixtures



Each point represents a technical replicate library, with the **expected** vs. **measured mutation frequencies (MF)** plotted on the x- and y-axes, respectively. Error bars are Wilson binomial confidence intervals (95%). The best-fit lines and R^2 values indicate high accuracy and linearity of mutation frequency measurements, even at relatively low mutation frequencies.

Mutation frequency: repeatability



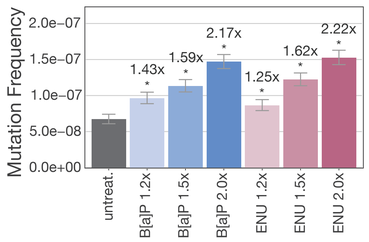
Sample	Mean MF	SD MF	Rel. SD
untreated	6.9e-08	9.9e-09	14%
B[a]P 1.2x	9.8e-08	9.3e-09	9%
B[a]P 1.5x	1.2e-07	5.5e-09	5%
B[a]P 2.0x	1.5e-07	9.2e-09	6%
B[a]P pure	4.5e-07	1.1e-08	2%
ENU 1.2x	8.8e-08	4.2e-09	5%
ENU 1.5x	1.3e-07	8.5e-09	7%
ENU 2.0x	1.5e-07	7.7e-09	5%
ENU pure	1.9e-06	7.0e-09	0.4%

To visualize technical repeatability of the assay, **mutation frequencies** of technical replicates are plotted as points, with 95% confidence intervals calculated using a t-distribution.

The table shows summary statistics for mutation frequency measurements of technical replicates, demonstrating high intra-lab repeatability, even at relatively low mutation frequencies.

Mutation frequency: assay sensitivity

< 0.05 < 0.001 < 0.01 < 0.0001

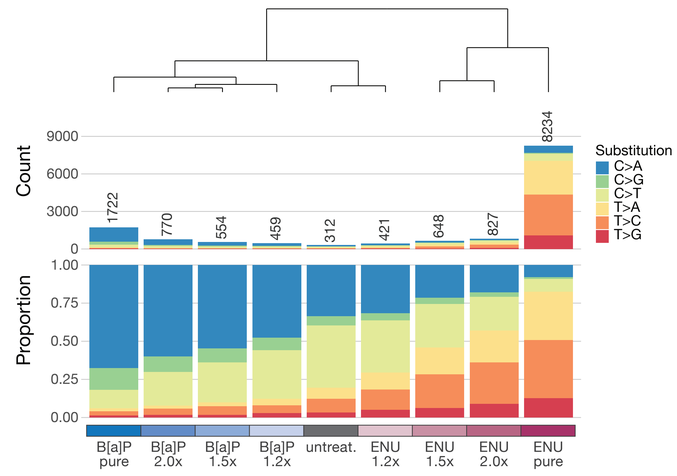


Bars show **per-sample mutation frequencies** from combined technical replicates, with Wilson binomial confidence intervals (95%). Numbers above bars represent measured fold-changes, compared to untreated. All mixture sample mutation frequencies are significantly higher than the untreated sample, with asterisks indicating FDR-adjusted p-values < 0.05 (Pearson's chi-squared test).

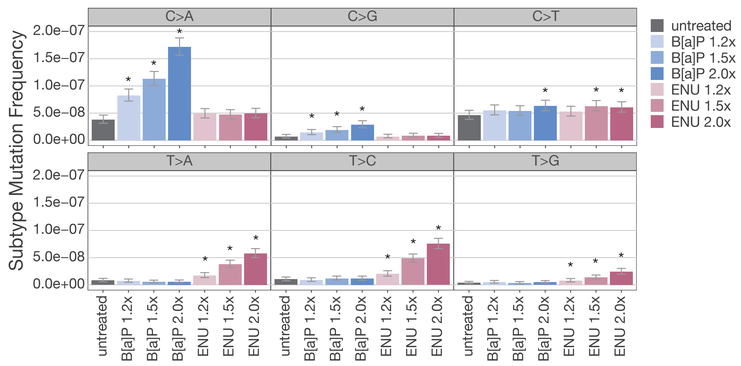
The table shows p-values for pair-wise comparisons between individual technical replicate libraries. Single libraries are sufficient to detect ~1.5x increase in MF.

			untreat. 1 2 3 4	% of pairwise p-value < 0.05	
~1.2x	B[a]P (1.43x)	1 2 3 4		75%	62.5%
	ENU (1.25x)	1 2 3 4		50%	
~1.5x	B[a]P (1.59x)	1 2 3 4		100%	100%
	ENU (1.62x)	1 2 3 4		100%	
~2.0x	B[a]P (2.17x)	1 2 3 4		100%	100%
	ENU (2.22x)	1 2 3 4		100%	

Single base substitution spectra

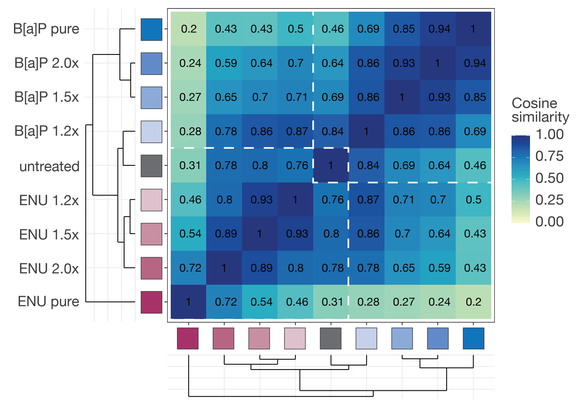


Unsupervised clustering of **single base subtitution spectra** shows step-wise similarity between the mixed and pure DNA samples (combined technical replicates). Numbers above the bars in the top panel are the number of unique single nucleotide variants detected in each sample.



Bars show **subtype mutation frequencies** for each sample (combined technical replicates) with Wilson binomial confidence intervals (95%). Asterisks indicate FDR-adjusted p-values < 0.05 (Pearson's chi-squared test). Each mixture has multiple subtype frequencies that are significantly higher than untreated, with different subtypes increasing in B[a]P and ENU mixture series.

Trinucleotide spectra: cosine similarity



Trinucleotide spectra were determined for each sample (combined technical replicates) and **cosine similarity** was calculated for all pair-wise combinations. Clustering along both axes of the heatmap recapitulates known relationships between samples and shows that trinucleotide spectra are sensitive to subtle changes in mutation frequency, such as those simulated by the DNA mixtures in this study.

Conclusions

- DNA mixtures were prepared and proper mixing was confirmed by analysis of germline variants.
- DNA mixtures had predictable and highly repeatable MFs, demonstrating the **accuracy and reproducibility** of DS-based measurements.
- A 1.25x increase in MF was detectable using combined replicate data and a ~1.5x increase was detectable with single libraries, demonstrating the **technical sensitivity** of DS.
- Single-base and trinucleotide spectra recapitulated known relationships between samples and revealed significant differences between compounds.
- The mixture-based DNA standards described here will facilitate further validation of DS-based genomic safety assays, including as part of an ongoing multi-site reproducibility study.

Contact: https://twinstrandbio.com/contact/
dfitzgerald@twinstrandbio.com