Assessment of Factors Impacting DuplexSeq Library Performance

Hobbs, CA[1], Hull, V[1], Smith-Roe, SL[2], Rivas, M[1], Sproul, C[1], Swartz, C[1], Marchetti, F[3], Yauk CL[4], and Auman, JT[1]

[1]Integrated Laboratory Systems, LLC, an Inotiv Company, Research Triangle Park, NC; [2]Division of Translational Toxicology, NIEHS, Research Triangle Park, NC; [3]Environmental Health Science and Research Bureau, Health Canada, Ottawa, Canada; [4]Department of Biology, University of Ottawa, Ottawa, Canada

Abstract

Duplex sequencing (DuplexSeq) is an ultra-accurate, error-corrected next generation sequencing (ecNGS) strategy that can be used to directly evaluate mutation frequency (MF) and spectrum in any organism. Efforts are currently underway throughout the genetic toxicology community to evaluate the utility of DuplexSeq for detection of gene mutations in vitro and in vivo, with a long-term goal of incorporating DuplexSeq (and other ecNGS platforms) into OECD test guidelines. Although the TwinStrand Bioscience DuplexSeq™ Mutagenesis rat, mouse, and human assay kits have proven to be very reliable, occasionally a sample fails library preparation or underperforms during sequencing. To minimize substantial cost associated with unsuccessful libraries, it will be useful to identify those parameters with greatest impact on performance outcome. We evaluated the laboratory's historical quality control results for libraries prepared from multiple tissue types of rats and mice and human cultured cells for correlations between various sequencing metrics (*e.g.*, raw reads, informative duplex bases, insert size, mean on-target duplex depth, peak tag family size on-target) and relationships of these metrics to input DNA mass and quality (integrity and purity). No clear association between poor DNA quality and sequencing performance was evident as the assay appears to tolerate a wide range of DNA integrity and purity. However, combination of lower DNA quality, low input quantity, and atypical library size profile increases the probability that the sequencing will be suboptimal (e.g., low duplex depth; high peak tag family size). Notably, informative mutation frequency results (*e.g.*, comparable to replicate animals) have been obtained from poor-performing libraries.

Introduction

Background:

DuplexSeq is an ultra-accurate, error-corrected, next-generation sequencing (ecNGS) strategy that detects 1 error in 10 to 100 million bases, which is within background levels of genetic mutation (Salk *et al.*, 2018, 2019). DuplexSeq can directly evaluate mutation frequency (MF) and the spectrum of mutations in any organism. The global genetic toxicology community is currently evaluating DuplexSeq for detecting gene mutations *in vitro* and *in vivo*, for interlaboratory reproducibility, and for sensitivity relative to established methods accepted by regulatory organizations. The long-term goal is to incorporate ecNGS technologies such as DuplexSeq into existing or new OECD test guidelines (Marchetti *et al.*, 2023).

TwinStrand Biosciences, Inc., has developed hybrid capture mutagenesis assay panel kits for human, rat, and mouse DNA that sample twenty, 2.4 kilobase (kb) target (bait) regions across ~50 kb of the genome. TwinStrand DuplexSeq™ Mutagenesis Assay kits have shown great promise for detection of chemical-induced mutation in tissues of transgenic mice (Dodge *et al.*, 2023; LeBlanc *et al.*, 2022) and standard laboratory rats (Smith-Roe *et al.*, 2023), as well as cultured cells (Cho *et al.*, 2023), with superb interlaboratory reproducibility.

Objective:

Although the DuplexSeq assay kits have proven reliable, occasionally a library does not perform as well as expected. Given the high cost of the reagent kits and sequencing, and time delays associated with repeating the library prep and sequencing, it would be useful to better understand the overall performance characteristics of the DuplexSeq mutagenesis assay and determine what parameters are most critical for success. Towards that goal, we evaluated the laboratory's historical quality control results for DuplexSeq libraries prepared from multiple tissue types of rats and mice and human cultured cells. Performance metrics evaluated include (with TwinStrand recommendations):

- informative duplex bases – the number of duplex consensus bases examined (exclusive of any ambiguous bases); target ≥500 million – 1 billion duplex bases
- mean on-target duplex depth – the mean coverage of the target region following duplex consensus alignment
- peak tag family size on-target – modal number of reads representing each strand in duplex consensus families; a median peak family tag size of 10 represents a balance between optimizing data yield and sequencing cost
- total mutant duplex bases – the total sum of mutation counts from duplex consensus bases; 10-15 mutations are considered ideal for simple frequency measurements and ≥100 mutations are needed to assess trinucleotide spectra

We also looked for relationships between these metrics and the following experimental parameters (with TwinStrand recommendations):

- input DNA mass – 50 ng to 2 µg; note that the combination of DNA input and number of flow cell clusters impact the performance metrics listed above
- DNA quality (integrity and purity) – DNA Integrity Number (DIN) >7 or intact high molecular weight band on a 0.5% agarose gel
- median insert size – insert sizes associated with high quality data range from 200 to 350 bp

Methods

Experimental Models:

- Sprague Dawley rats
- CD-1 mice; MutaMouse
- Rodent tissues – liver; bone marrow; lung; mammary gland; tumors
- Human – EpiIntestinal™ 3D tissue (MatTek); cultured cells of unknown origin

DNA Isolation and QC:

- Isolation performed in-house: Monarch® Genomic DNA Purification Kit (New England BioLabs, Ipswich, MA, USA)
- DNA samples shipped to Inotiv: Qiagen DNeasy; DNA Extraction Kit (Agilent Technologies, Santa Clara, CA), or unspecified non-phenol method
- Integrity assessment: electrophoresis on a 0.8% agarose gel stained with ethidium bromide (0.15 µg/mL) or 1x SYBR® Safe DNA gel stain in Tris-borate-EDTA buffer.
- Purity assessment: Nanodrop microvolume UV spectrophotometer (Thermo Fisher Scientific, Inc., Waltham, MA, USA); if the 260/230 absorbance ratio was below 1.7, the DNA sample underwent a clean-up step using CleanNGS SPRI Beads (Bulldog Bio, Portsmouth, NH, USA), except when DNA yield was low.
- Concentration determination: Qubit Broad Range DNA assay (Life Technologies, Carlsbad, CA, USA) and Qubit 3.0 fluorometer (Life Technologies)

DNA Library Preparation:

- Input genomic DNA was 500, 650, or 1000 ng, with some exceptions due to available yield. When library preparation was split over multiple batches, each batch contained samples from each treatment group to minimize any potential batch effects. Sample cross-contamination was avoided by appropriate spacing of samples in 96-well plates and using multi-channel pipettors with single-use pipettor tips.
- DNA libraries were prepared using TwinStrand DuplexSeq™ Mutagenesis v1.0 kits and Enzymatic Fragmentation Module (Mouse-50, Rat-50, Human-50; TwinStrand Bioscience, Seattle, WA, USA) in accordance with the instruction manual.
- Enzymatic digestion was used to generate fragments with overhanging adenine to which thymidine-containing adapter sequences were ligated. These adapters have double-stranded molecular barcodes to permit downstream identification of each starting DNA molecule. Indexing primer sequences were then ligated to the ends of the double-stranded adapters so that DNA fragments would hybridize and bind to the sequencing flow cell. A limited polymerase chain reaction (PCR) amplification was followed by a series of capture and washing steps to remove free primers and indexes.

Quantitation and Sizing of DNA Libraries:

- Quantitation during library preparation utilized the dsDNA Broad Range or High Sensitivity Qubit assay (Thermo Fisher Scientific/Life Technologies, Eugene, OR, USA) and Qubit fluorometer. In some cases, the libraries underwent additional quantification using the KAPA Library Quantification kit (Roche, Pleasanton, CA, USA).
- A Bioanalyzer using a High Sensitivity DNA chip (Agilent Technologies, Santa Clara, CA, USA) was used to determine the approximate size range and mean library peak size.
- Libraries were normalized to 15-20 nM with low TE (10 mM Tris-HCl pH 8.0, 0.1 mM EDTA) or resuspension (10 mM Tris-HCl, pH 8.5 with 0.1% Tween-20) buffer.

Sequencing:

- Libraries were sequenced at Azenta, Psomagen, or the NIEHS Core Sequencing facility using the Illumina NovaSeq 6000 or NovaSeq X Plus platform.
- Libraries were pooled together at equimolar concentrations and loaded on S4 or 10B (2 × 150 bp) flow cells to approximate the equivalent of 7.5 - 16 (S4) or 6 (10B) samples per lane, to target ~313 - 667 million paired-end reads per sample (assuming capacity of the flow cell is 20 billion paired-end reads). For context, libraries prepared with 650 ng of input DNA, pooled and sequenced to target ~417 million paired-end reads per sample are expected to yield ~1–1.25 billion duplex bp per sample.

Analysis of Duplex Sequencing Results:

- FastQ files containing sequence data for the pooled libraries were demultiplexed and uploaded to the DNAnexus cloud-based platform (dnanexus.com). Raw sequencing data were processed using the TwinStrand DuplexSeq FASTQ to VCF Parallel App that performed consensus making, consensus read postprocessing, and variant calling.
- The TwinStrand app on the DNAnexus platform generated a DuplexSeq™ Assay Performance Report and a Mutagenesis Assay Report. Only somatic variants with variant allele fraction (VAF) < 0.01 were included in the mutation frequency (MF) calculations. MF was calculated as the ratio of mutant duplex bases to total duplex bases for each sample. MFs were calculated using the minimum method (MF_{min}) on the assumption that multiple observations of the same mutation result from cell division rather than multiple independent mutation events (Dodge *et al.* 2023).

Results

1) No clear association between poor DNA quality and sequencing performance was evident as the assay appears to tolerate a wide range of DNA integrity and purity.
2) The combination of lower DNA quality, low input quantity, and atypical library insert size may increase the probability that the sequencing will be suboptimal.
3) Informative MF and spectra results (*e.g.*, comparable to biological replicates) have been obtained from under-performing libraries.

Conclusion

No single experimental factor critical to the success of DuplexSeq library preparation and sequencing could be identified, confirming the overall robustness of the assay.

References

Cho, E., Swartz, C. D., Williams, A., Rivas, M. V., Recio, L., Witt, K. L., Schmidt, E. K., Yaplee, J., Smith, T., Van, P., Lo, F. Y., Valentine III, C. C., Salk, J. J., Marchetti, F., Smith-Roe, S. L., & Yauk, C. L. (2023). Error-corrected duplex sequencing enables direct detection and quantification of mutations in human TK6 cells with strong inter-laboratory consistency. Mutat Res Genet Toxicol Environ Mutagen 889, 503649. https://doi.org/10.1016/j.mrgentox.2023.503649.

Dodge, A. E., LeBlanc, D. P., Zhou, G., Williams, A., Meier, M. J., Van, P., Lo, F. Y., Valentine III, C. C., Salk, J. J., Yauk, C. L., & Marchetti, F. (2023). Duplex sequencing provides detailed characterization of mutation frequencies and spectra in the bone marrow of MutaMouse males exposed to procarbazine hydrochloride. Arch Toxicol 97, 2245-2259. doi: 10.1007/s00204-023-03527-y.

LeBlanc, D. P. M., Meier, M., Lo, F. Y., Schmidt, E., Valentine III, C., Williams, A., Salk, J. J., Yauk, C. L., & Marchetti, F. (2022). Duplex sequencing identifies genomic features that determine susceptibility to benzo(a)pyrene-induced in vivo mutations. BMC Genomics 23(1), 542. https://doi.org/10.1186/s12864-022-08752-w

Marchetti, F., Cardoso, R., Chen, C. L., Douglas, G. R., Elloway, J., Escobar, P. A., Harper, Jr. T., Heflich, R. H., Kidd, D., Lynch, A. M., Myers, M. B., Parsons, B. L., Salk, J. J., Settivari, R. S., Smith-Roe, S. L., Witt, K. L., Yauk, C., Young, R. R., Zhang, S., & Minocherhomji, S. (2023). Error-corrected next-generation sequencing to advance nonclinical genotoxicity and carcinogenicity testing. Nat Rev Drug Disc 22, 165–166. https://doi.org/10.1038/d41573-023-00014-y

Salk, J. J., Schmitt, M. W., and Loeb, L. A. (2018). Enhancing the accuracy of next-generation sequencing for detecting rare and sub-clonal mutations. Nat Rev Genet 19: 269–285. https://doi.org/10.1038/nrg.2017.117

Salk, J. J., & Kennedy, S. R. (2020). Next-generation genotoxicology: using modern sequencing technologies to assess somatic mutagenesis and cancer risk. Environ Mol Mutagen 61(1), 135–151. https://doi.org/10.1002/em.22342

Smith-Roe, S. L., Hobbs, C. A., Hull, V., Auman, J. T., Recio, L., Streicker, M. A., Rivas, M. V., Pratt, G. A., Lo, F. Y., Higgins, J. E., Schmidt, E. K., Williams, L. N., Nachmanson, D., Valentine, C. C., Salk, J. J., Witt, K. L. (2023). Adopting duplex sequencing technology for genetic toxicity testing: A proof of-concept mutagenesis experiment with N-ethyl-N-nitrosourea (ENU)-exposed rats. *Mutat Res Genet Toxicol Environ Mutagen* 891, 503669. doi.org/10.1016/j.mrgentox.2023.503669

Funding

This work was funded in part by the Division of Translational Toxicology, National Institutes of Environmental Health Sciences, National Institutes of Health, US Department of Health and Human Services under contract HHSN273201300009C (genetic toxicity testing). Some tissues were provided from a project funded by the Burroughs Wellcome Fund and Health Canada's Chemicals Management Plan.

Figure 1. Distribution of Historical Performance Metric Data

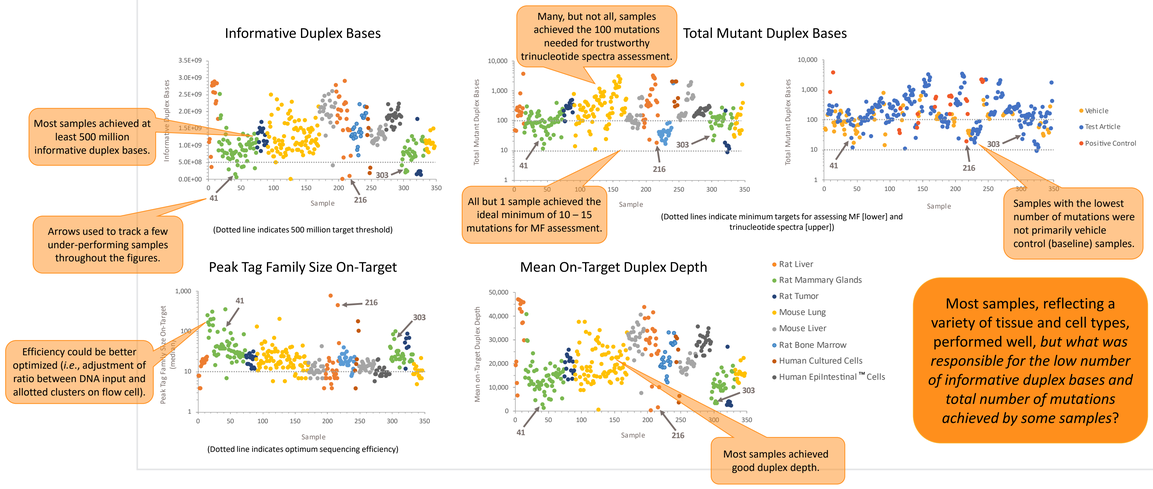

Figure 2. Impact of DNA Integrity and Purity on Performance Metrics

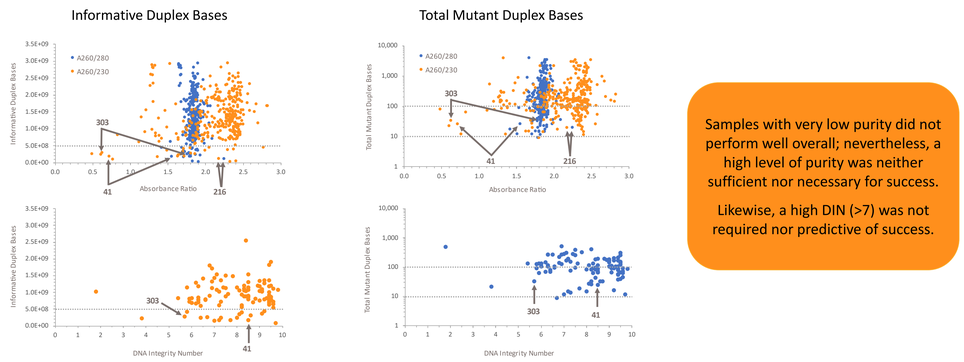

Figure 3. Impact of DNA Input and Library Insert Size on Performance Metrics

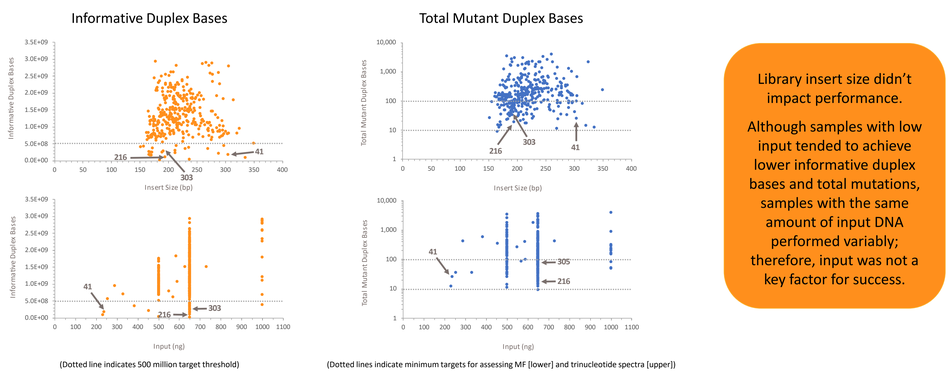

Table 1. Characteristics of Some Samples that Under-performed

Sample	Sample Type	DNA Isolated In-House?	DNA Isolation Kit	$A_{260/280}$	$A_{260/230}$	Cleaned Up or Bead Conc.?	DIN	Agarose Gel QC	Input Mass (ng)	Median Insert Size (bp)	Mean On-Target Duplex Depth	Peak Tag Family Size On-Target	Informative Duplex Bases	Total Mutant Duplex Bases
31	Rat Mammary Glands	No	Unknown	1.73	1.70	No	8.6	-	650	212	4,524	99	303,644,148	33
41	Rat Mammary Glands	No	Unknown	1.55	0.72	Yes	8.5	-	238	305	2,885	113	166,578,435	25
43	Rat Mammary Glands	No	Unknown	1.51	0.77	Yes	9.7	-	231	336	1,382	363	79,059,076	12
45	Rat Mammary Glands	No	Unknown	1.42	2.05	No	9.2	-	650	215	6,522	57	432,533,957	17
51	Rat Mammary Glands	No	Unknown	1.88	2.30	Yes	9.1	-	650	373	5,004	94	369,173,739	68
54	Rat Mammary Glands	No	Unknown	1.82	1.71	No	8.2	-	650	152	4,859	66	393,043,671	26
71	Rat Mammary Glands	No	Unknown	1.71	2.00	No	8.1	-	650	230	5,188	86	345,284,976	28
126	Mouse Lung	Yes	Monarch gDNA Isolation Kit	1.83	2.41	No	-	Intact	500	238	464	209	30,387,114	240
191	Mouse Liver	Yes	Monarch gDNA Isolation Kit	1.69	2.77	No	-	Intact	650	298	7,546	44	433,572,938	156
216	Rat Liver	Yes	Monarch gDNA Isolation Kit	2.25	2.18	No	-	Intact	650	195	1,512	455	105,374,427	19
297	Rat Mammary Glands	No	Unknown	1.83	1.80	No	5.8	Partially Degraded	650	157	5,379	43	422,345,554	93
302	Rat Mammary Glands	No	Unknown	1.87	0.61	No	3.8	Degraded	650	212	3,317	88	219,393,273	22
303	Rat Mammary Glands	No	Unknown	1.78	0.63	No	5.7	Partially Degraded	650	192	4,079	66	280,333,461	33
305	Rat Mammary Glands	No	Unknown	1.81	0.49	No	6.4	Partially Degraded	650	189	3,214	102	246,175,893	80

Figure 4. Results of Some Under-Performing Samples (↓) Compared to Group Counterparts