Integration of whole exome sequencing and RNA-seq facilitates somatic mutation detection in tumor tissue samples

Margaret R. Heider, Chelsea Pinegar, Evan Janzen, Jian Sun, Gautam Naishadham, Adrian Reich, Chen Song | New England Biolabs Inc., Ipswich, MA 01938, USA

INTRODUCTION

Somatic mutations are critical targets for disease diagnosis, prognosis, and drug discovery. Whole genome sequencing and targeted sequencing approaches are crucial for identifying somatic mutations in cancer. Meanwhile, RNA-seq has emerged as a complementary tool for somatic mutation profiling, especially for variants of uncertain significance (VUS), to characterize cancer types for precision medicine. Therefore, a comparative study to determine the concordance between DNA-seq and different RNA-seq methods and data analysis tools will help define the optimal approaches for somatic variant detection.

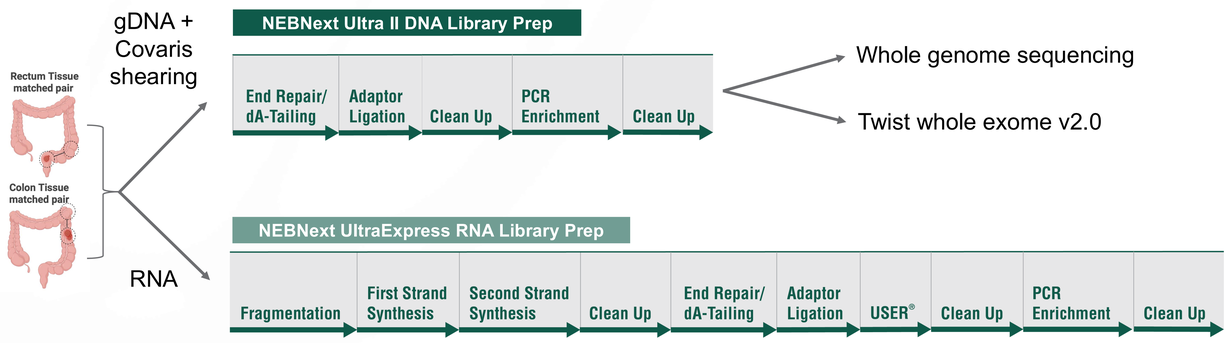
DNA and RNA were extracted from paired tumor and normal tissue samples from rectum and colon cancer patients. Libraries for Illumina sequencing were prepared using the NEBNext® Ultra™ II DNA library Prep Kit and the NEBNext UltraExpress® RNA Library Prep Kit. Direct RNA-seq (dRNA-seq) libraries were prepared for Oxford Nanopore Technologies (ONT). Target enrichment panels from Twist Bioscience were utilized to enrich Illumina libraries, using the whole exome panel for DNA libraries and the RNA exome panel for RNA libraries. Strelka and Arriba were used for somatic small variants and fusion gene calls from Illumina data, respectively. The fusion gene calls were compared to ONT dRNA-seq data. The single-nucleotide variant (SNV) somatic variant calls between RNA and DNA data were evaluated for concordance.

We have obtained copy number variation (CNV), SNV, and indels from DNA whole-exome sequencing (WES) data; SNVs, indels, and fusions from short-read RNA-seq datasets. The somatic mutation correlation between RNA and exome-enriched DNA data showed overlapping mutations and unique mutations in each dataset, which demonstrated RNA-seq is complementary to DNA-seq in somatic mutation detection. The exome-enriched RNA-seq data provided higher depth and sensitivity for the variants that occurred in target regions. In contrast, the non-enriched RNA-seq libraries provided additional variant information in intronic regions. The fusion variants in RNA-seq are valuable in confirming VUS detected in WES. dRNA-seq characterizes fusions and isoforms better than short-read RNA-seq alone. In contrast, dRNA-seq data is less sensitive than short-read RNA-seq for detecting variants with low allele frequency.

Well-designed bioinformatic pipelines, DNA WES and WGS, combined with short and long-read RNA-seq assays, can provide valuable somatic mutation information. Short-read RNA-seq can be sensitive for detecting low-frequency mutations and gene fusions. Long-read dRNA-seq is valuable for characterizing fusions. Utilizing both WES and RNA-seq approaches provides a deeper understanding of cancer-developing pathways for targeted therapy.

METHODS

Sequencing tumor-normal DNA and RNA from matched donors



- Genomic DNA and RNA were extracted from tumor and adjacent normal tissue (rectum and colon) from the same two donors (Biochain Institute)
- 100 ng gDNA was used for library prep with NEBNext Ultra II DNA Library Prep Kit
- 25 ng total RNA was used for library prep using NEBNext rRNA depletion kit v2 (human/mouse/rat) followed by NEBNext UltraExpress RNA Library Prep Kit
- All libraries were prepared using NEBNext Unique Dual Index UMI Adaptors and sequenced on the Illumina® NovaSeq® 6000 platform

RESULTS

High quality exome capture libraries from DNA and RNA

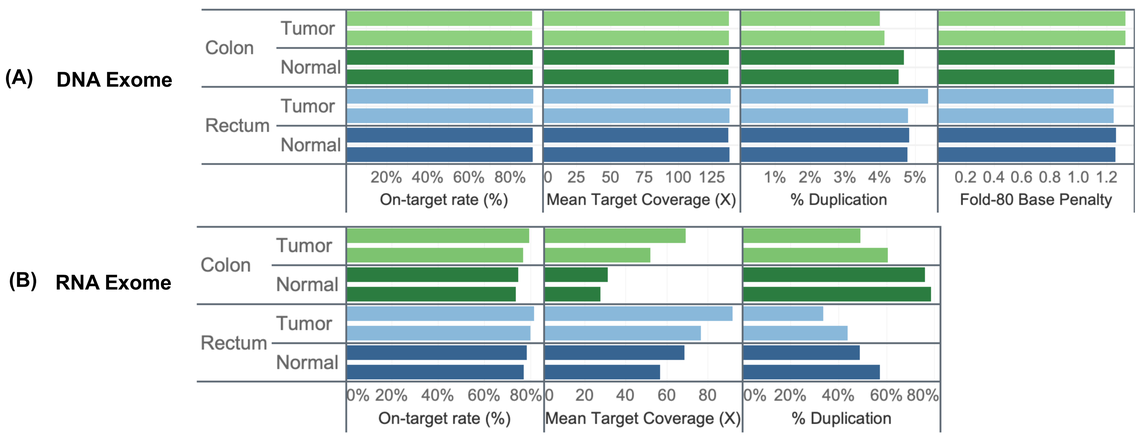


Figure 1. High quality exome capture libraries from both DNA and RNA inputs. (A) 100ng of DNA from matched tumor and normal, colon and rectum samples was sheared in duplicate using Covaris ML230 TPX tubes for library preparation with the NEBNext Ultra II DNA Library Prep Kit, NEBNext UDI UMI Adaptors, and 6 PCR cycles. Libraries were quantified using the Agilent HSD1000 Tapestation assay and then captured in an 8-plex reaction using the Twist exome 2.0 panel (Twist Bioscience). Captured libraries were sequenced on the NovaSeq6000 2x100 bp, down sampled to 700M read pairs, trimmed using fastp (v.0.20.0), and aligned using bwa-mem (v.0.7.17) to the GRCh38 reference. Duplicates were marked using Picard MarkDuplicates with UMI (v.2.18.29), and capture metrics assessed using Picard HS metrics (v.2.18.29). DNA libraries showed high on-target rate and high uniform coverage. (B) Total RNA (25ng) was extracted from paired normal/tumor, colon and rectum samples. The RNA samples were first depleted of ribosomal RNA using NEBNext® rRNA Depletion Kit v2 (Human/Mouse/Rat). Libraries were then prepared (in duplicate) using NEBNext UltraExpress® RNA Library Prep Kit and NEBNext Multiplex Oligos for Illumina (Unique Dual Index UMI Adaptors RNA Set 1) with 15 PCR cycles (in duplicate) for WTS and an 8-plex capture with the RNA exome panel (Twist Bioscience). UMI adaptors were used to distinguish and remove duplicates prior to mapping and alignment. The pre-capture and post-capture RNAseq libraries were sequenced on Illumina® NovaSeq® 6000 2x75 bp and down sampled to 40M PE reads for analysis. Reads were aligned to the GRCh38 reference genome using RNA STAR v2.7.8a, and reads with the same unique molecular identifier (UMI) and mapping coordinates were marked and removed using Picard MarkDuplicates v1.56.0. Capture metrics were assessed using Picard HS metrics (v.2.18.29) and showed high on-target rate and coverage.

RESULTS (CONTINUED)

Sensitive SNV detection in DNA and RNA libraries

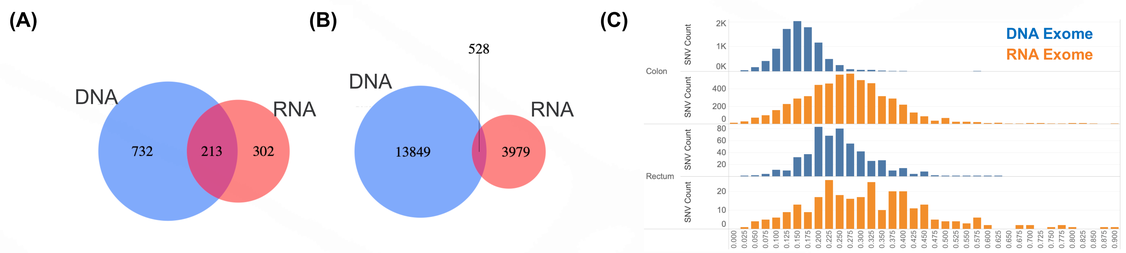


Figure 2. Correlating SNV calls between DNA and RNA exome libraries and comparing the observed variant allele fraction for overlapping variants. DNA and RNA exome library mapping and duplicate marking by UMI were done according to Figure 1, with consensus UMI sequence used for DNA libraries. Tumor-normal variant calling was done using strelka2 (v 2.9.10) for both DNA and RNA libraries. Observed SNV calls are shown for DNA (purple) and RNA (red) exome libraries for rectum (A) and colon (B) tumor samples with shared variants in the overlap of the Venn diagrams. (C) The variant allele fraction (VAF) distribution for the shared set of DNA and RNA SNVs are shown.

SNV and CNV detection in whole genome sequencing libraries (WGS)

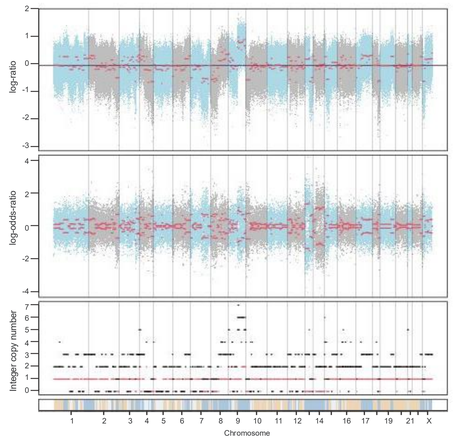


Figure 3. Copy number variation (CNV) detection in whole genome sequencing data from tumor-normal pairs. 100ng of fresh frozen tissue DNA from matched tumor and normal rectum (left) and colon (not shown) were prepared in duplicate using Covaris ML230 shearing (TPX tubes) and NEBNext Ultra II DNA Library Prep using NEBNext UMI UDI Adaptors. Whole genome sequencing was done on the Illumina® NovaSeq® 6000 using 2x150 bp reads to 23X coverage (rectum) and 18X coverage (colon, data not shown) using 400M read pairs. FACETs was used for allele-specific copy number analysis (ASCN) (version 0.5.6). At each position, logR is defined by the log-ratio of total read depth in the tumor versus that in the normal and logOR is defined by the log-odds ratio of the variant allele count in the tumor versus in the normal. 157 segments were called as somatic CNVs in the rectum tumor sample. About 1.32Gbp regions in the rectum tumor genome contain CNV gain, loss or copy-neutral LOH. Follow up analysis using RNA-seq libraries and differential expression could be used to correlate and identify the impact of these copy number changes on gene expression.

Detecting and validating fusions using RNA exome and direct RNA ONT library prep

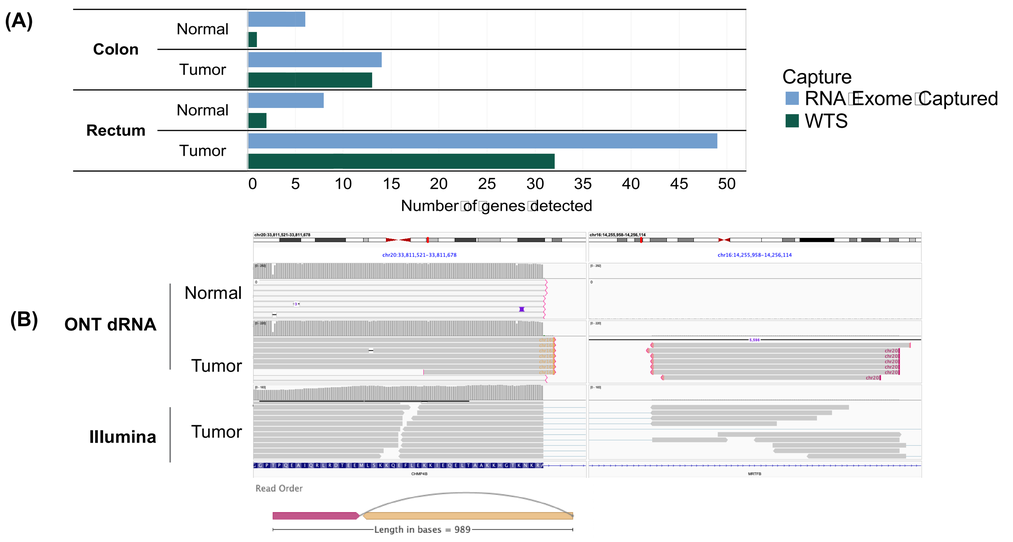


Figure 4. Fusion genes were identified using short read sequencing and validated using ONT dRNA sequencing. (A) Total number of fusion genes observed (all replicates) with high or medium confidence by tissue and sample type using Arriba (v2.5.0) after UMI deduplication. Fusions in-common between normal and tumor samples were eliminated from analysis. (B) Direct RNA libraries were prepared using the Direct RNA Sequencing Kit SQK-RNA004 (ONT) and rectum tumor and normal samples following manufacturer-recommended protocol. Reverse transcription steps were performed using Induro RT (55 °C for 20 min). Libraries were sequenced on PromethION R10 flow cells and down sampled to equivalent total read depths. Supplementary alignments of Induro-generated direct RNA sequencing data were visualized in IGV to identify potential fusion transcripts. One potential fusion transcript was identified in the rectum tumor sample spanning regions of chr16 and chr20 that was not observed in the normal sample. Reads from these regions were observed in Illumina sequencing data, but the fusion transcripts were not observed, possibly due to shorter read lengths precluding alignment.

CONCLUSIONS

- DNA library prep using NEBNext Ultra II DNA enables sensitive detection of SNVs and small indels (not shown) when combined with exome capture. Combining SNV detection and coverage differences in whole genome sequencing enables robust copy number detection. Next steps include correlating copy number changes with differential expression analysis in the RNA-Seq data.
- RNA-seq with RNA exome and WTS (not shown) enabled less sensitive SNV detection with altered VAF distribution, which could be due to lower coverage, but did show some concordance with DNA-seq. Next steps include identifying technical and biological reasons underlying variants uniquely identified in RNA libraries.
- Fusion detection is uniquely enabled by RNA-Seq. Direct RNA sequencing on ONT provides strong corroborating evidence for fusions identified in short read data. Follow up analysis includes identifying any genomic changes in DNA WGS that might contribute to some classes of fusions observed.

ACKNOWLEDGEMENTS

Thank you to the NEB Sequencing core facility for their technical assistance.