Investigation of critical RNA sequencing variables and their impact on fusion calling

Travis Sanders, Giulia Corbet, Josh Haimes, Martin Ranik, Thomas Harrison, Eduard Casas, Kailee Reed, Doug Wendel, Kristina Giorda, Jen Pavlica, Brian Kudlow
Watchmaker Genomics – Boulder, CO USA

Introduction

Gene fusions due to chromosomal rearrangement, duplication, or deletion can be important drivers of cancer and identification of gene fusions can play a valuable role in selecting targeted therapies. Targeted RNA sequencing via target enrichment (TE), which focuses sequencing reads on known fusion junctions is one of the most widely used methods for detecting fusion transcripts. However, it has limitations with respect to the discovery of novel fusions. Whole transcriptome analysis (WTA), where the entire transcriptome is interrogated, offers an opportunity to detect both known and novel fusions. However, there is no consensus in the field for best practices in library preparation, sequencing, or analysis for fusion detection in a WTA setting. In this study, we interrogated the effect of input mass and workflow alterations on gene fusion detection in RNA sequencing and compare targeted sequencing to WTA for known fusions.

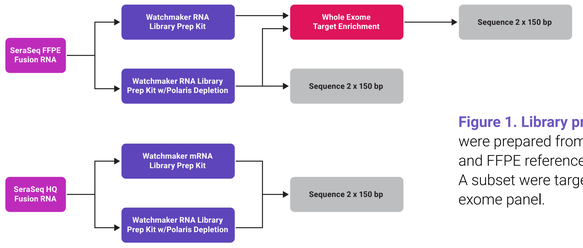
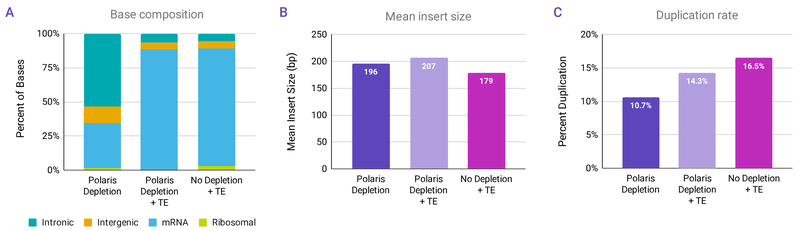
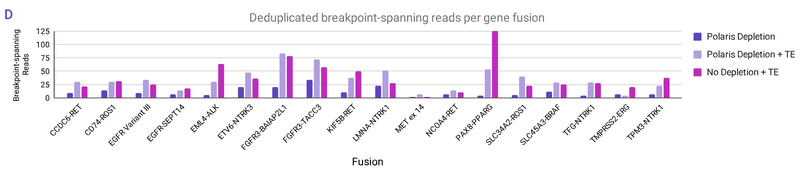


Figure 1. Library preparation workflow. Libraries were prepared from Seraseq® high quality (HQ) and FFPE reference RNA using various workflows. A subset were target-enriched with a whole exome panel.

TE Improves Known Fusion Calling Efficiency





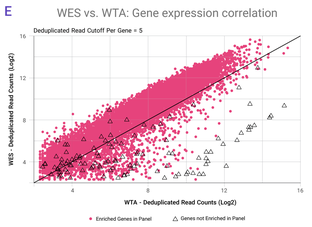
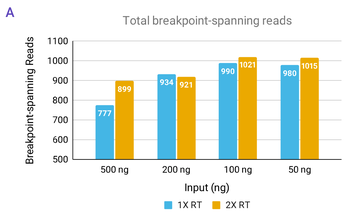


Figure 2. Target enrichment (TE) increases fusion calls. Libraries were prepared in duplicate from 100 ng Seraseq FFPE RNA with the Watchmaker RNA Library Prep Kit 1) Polaris Depletion upstream, 2) Polaris Depletion upstream and TE downstream, or 3) just TE downstream. Data are shown for **(A)** base composition, **(B)** mean insert size, **(C)** duplication rate, and **(D)** deduplicated breakpoint (BP)-spanning reads per gene fusion. Data shows that target enrichment increases breakpoint-spanning reads and thus sequencing economy for known fusion calling, and incorporation of Polaris Depletion upstream reduces rRNA contamination. **(E)** Expression correlation analysis between +/- TE libraries shows that target enrichment does alter relative gene expression to some degree. Subsampling was 28.5M read pairs.

More Input is Not Always Better for TE Fusion Calling



B

	500 ng FFPE RNA		50 ng FFPE RNA	
	1X RT	2X RT	1X RT	2X RT
CCD6-RET	13	24	28	37
CD74-ROS1	7	33	58	84
EGFR-SEPT14	16	44	47	61
EGFR Variant III	14	11	20	40
EML4-ALK	161	113	57	34
ETV6-NTRK3	32	44	45	53
FGFR3-BAIAP2L1	57	90	114	103
FGFR3-TACC3	41	62	80	103
KIF5B-RET	90	88	83	62
LMNA-NTRK1	20	25	60	52
MET ex 14 skipping	1	5	4	9
NCOA4-RET	5	10	26	26
PAX8-PPARG	197	202	162	120
SLC34A2-ROS1	2	17	37	50
SLC45A3-BRAF	33	33	46	61
TFG-NTRK1	39	46	38	43
TMPRSS2-ERG	11	13	23	22
TPM3-NTRK1	38	39	56	58
Total breakpoint spanning reads	777	899	980	1015

Figure 3. Impact of input mass and reverse transcriptase (RT) on fusion calling. (A) Total deduplicated BP-spanning reads for TE libraries generated with 50 – 500 ng of total Seraseq FFPE RNA, summed across all fusions. Inputs >100 ng are detrimental to the number of BP-spanning reads, though increasing RT concentration can salvage performance at higher inputs. **(B)** Deduplicated BP-spanning reads per fusion. Red indicates fewer than 3 BP-spanning reads per fusion. Individual replicates shown for 500 and 200 ng, average of duplicates for 100 and 50 ng. Subsampling was 42M read pairs.

rRNA Depletion Prior to TE Benefits High Inputs

Figure 4. High input masses benefit from rRNA depletion prior to target enrichment. (A) Total breakpoint-spanning reads for individual libraries at 500 ng, average of duplicates at 100 ng. **(B)** Duplication rate is lower for rRNA-depleted libraries at 500 ng, whereas rates are comparable at 100 ng. Subsampling was 42M read pairs.

Modified SPRI Maintains Library Complexity and May Enhance *de novo* Fusion Calling

Figure 5. Longer inserts are achievable and may benefit *de novo* fusion calling. (A) Median insert size of BP-spanning reads increases with decreasing post-PCR cleanup ratio. Magnitude of increase agrees with insert size across whole library. *De novo* fusion calling requires longer overlaps on each arm of the fusion; thus, longer inserts are expected to be beneficial for *de novo* applications. **(B)** Library complexity, as measured by unique genes detected, is not impacted by reduced SPRI ratio. **(C)** Percentage of fusions detected with 3 or more BP-spanning reads. Duplicate libraries were generated with 100 ng of Seraseq FFPE or High Quality Fusion RNA. Subsampling was 28.5M read pairs.

mRNA-seq Produces More Fusion-Supporting Reads

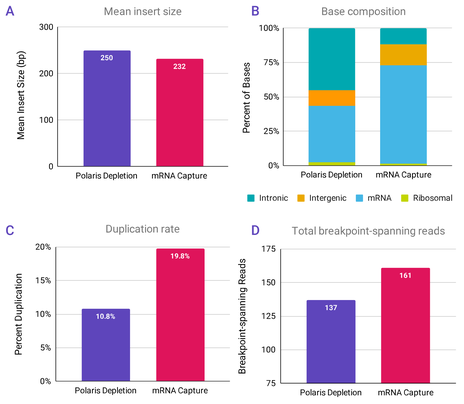


Figure 6. mRNA capture improves fusion calling with high-quality samples. Duplicate libraries were prepared with 100 ng high-quality Seraseq RNA using either the Polaris Depletion or mRNA capture workflow. Data are shown for **(A)** mean insert size, **(B)** base composition, **(C)** duplication rates, and **(D)** total fusion BP-spanning reads. Results indicate the Polaris libraries are more complex but cover noncoding regions, whereas mRNA libraries are focused on exonic regions — leading to increased coverage of gene fusions with equivalent sequencing depth.

Downsampling was 28.5M read pairs.

Materials and Methods

RNA samples: RNA was extracted from Seraseq FFPE Fusion RNA v4 Reference Material, which contains 18 clinically relevant gene fusions. Seraseq Fusion RNA v4 Reference Materials was used for high-quality RNA. For high-quality RNA, the RIN was 9.8 and DV200 was 94.6%. For FFPE RNA, the RIN was 1.8 and DV200 was 91%.

Library Prep: Libraries were prepared using the Watchmaker RNA Library Prep Kit, Watchmaker RNA Library Prep Kit with Polaris Depletion, or Watchmaker mRNA Library Prep Kit. Input masses ranged from 50 – 500 ng. The IDT xGEN adapter system was used.

Target Enrichment: Libraries were target-enriched using the Twist Exome 2.0 panel and Twist Hybridization Capture Reagents. 200 ng per library was used for capture. Equivalent RNA mass libraries were pooled together to control for capture variation.

Sequencing: Libraries were sequenced 2 x 150 bp on a NovaSeq 6000.

Data Analysis: Subsampling was performed with Seqtk, followed by quality control using FastQC and adapter trimming with Cutadapt. Reads are aligned with the STAR aligner, and the duplicates are marked using GATK's Picard tool. Gene expression is measured through gene/transcript abundance estimation with featureCounts and a cutoff of 5 unique reads. Various metrics concerning RNA quality, coverage biases, GC content, read distribution, and potential contamination are subsequently assessed using tools like Picard, RSeQC, and Kraken2, among others.

Fusion Calling: In order to identify all reads spanning a fusion breakpoint, fusion reference sequences were generated by taking at most 500 bp on each side of the breakpoint and added into a background of the transcriptome coming from gencode v44 to generate a mapping reference. Downsampled reads with adapter sequences removed were mapped to the reference using bwa mem. BP-spanning reads were annotated if they met the following conditions: The record must have mapped in a proper pair, the record must have 10 base pairs flanking each side of the breakpoint, the record must have a mapping score no less than 30, the record must be a primary alignment, the record can have no more than 4 mismatched bases from the reference, finally, the record could not have mapped any deletions, insertions, or softclips.

Conclusions and Next Steps

Watchmaker Genomics offers several library preparation solutions which effectively enable fusion calling. Workflow modifications such as increased RT or post-PCR SPRI ratio may improve fusion calling at high inputs or for *de novo* applications.

Next step: Assess whether or not library prep workflow modifications that are best for TE fusion detection also perform best with *de novo* fusion calling using clinical samples.

M252