SINGULAR GENOMICS

APPLICATION NOTE

RNA Sequencing on the G4™

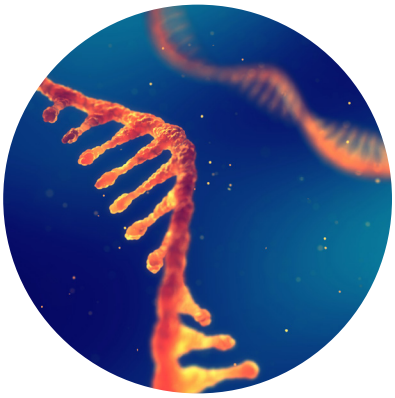
- The unique flow cell design and scalable capacity of the G4 empowers labs to process 6-48 samples per run across 1 to 4 flow cells and 4 to 16 lanes.
- The G4 Sequencing Platform delivers accurate RNA sequencing data highly correlated with the industry standard benchtop sequencer in just 6 to 15 hours.

Introduction

RNA sequencing (RNA-Seq) uses next generation sequencing (NGS) technology to assess differential gene expression, detect novel transcripts, and characterize new splice variants or cell types.[1] These studies provide greater insight into the genetic mechanisms that differentiate normal and diseased cells. RNA-Seq is used in clinical oncology research and diagnostics as a tool to characterize tumor phenotypes and to develop more effective therapies. Additionally, RNA-Seq is useful in other clinical applications, such as transplant medicine and infectious diseases.[3]

RNA-Seq workflows begin with the isolation of RNA from biological material, for example blood samples, fresh, frozen, or formalin fixed and paraffin embedded (FFPE) tissue, or cell lines. After isolation, NGS library preparation begins with reverse transcription of RNA into cDNA and the addition of adapter sequences. Strand-specificity in cDNA libraries enables the preservation of antisense transcripts.[4] After sequencing, reads are processed for transcriptome profiling, which involves comparing reads to existing annotations for transcript discovery or mapping and quantifying reads to a reference genome or transcriptome.[5] The analysis of differential expression and alternative splicing enables characterization of cell-types, cellular activity mechanisms, and other phenotypic information.

The G4™ Sequencing Platform is a highly versatile benchtop sequencer suitable for demanding research and clinical research applications. The G4 leverages a novel, 4 color rapid SBS chemistry and advanced optical and fluidics engineering to deliver unmatched power and versatility in key applications like RNA sequencing. The G4 is compatible with existing upstream RNA-Seq library preparation kits and outputs demultiplexed FASTQ files compatible with existing bioinformatic pipelines.



RNA-Seq Parameters

The G4 provides users with the flexibility to multiplex samples within a lane or a flow cell depending on experimental design. Examples of expected sequencing output, run time, accuracy, quality, and throughput by lane, flow cell, and run are shown in **Table 1**. Run time, accuracy, and quality metrics are indicative of G4 specifications.

Flow Cell Type	F2		F3*		
Read Length	2 x 100	2 x 50	2 x 100	2 x 50	1 x 50
Run Time (Hours)	12-15	8-10	12-15	8-10	6-8
Reads	150-165M per FC 600-660M per run		300-330M per FC 1,200-1,320M per run		
Quality	75-90% Bases ≥ Q30				
Accuracy	99.6-99.9%				
Samples / Lane	1.5		3		
Samples / FC	6		24		
Samples / Run[a]	24		48		

Table 1: RNA-Seq sequencing parameters
[a]RNA-Seq assumptions are based on 25M read sequencing per sample.
*Planned for next release.

Methods

All human RNA samples were from Thermo Fisher Scientific: Universal Human Reference RNA (Cat #QS0639, lot #301095-000). Poly(A)-selection (NEBNext PolyA mRNA Magnetic Isolation Module; Cat. #E7490) steps were carried out per manufacturer guidelines, using 1 µg of input RNA from each sample containing ERCC spike-in Mix 1 (Cat. #4456740; 2 µL of 1:100 dilution).

Poly-A selected libraries were made using NEBNext® Ultra™ Directional RNA Library Prep Kit for Illumina® (Cat. #E7760), with minor modifications. PCR amplification (8 cycles) was carried out using Singular Genomics indexed PCR primers at 2 µM final concentration each. Libraries were purified using SparQ PureMag beads (Cat. #95196) with a 0.9X ratio, quantified with a Qubit 4 Fluorometer and an Agilent TapeStation. Libraries (331 bp ± 17 bp) were multiplexed and sequenced on a Singular Genomics G4 with a 2x100 cycle format with F2 flow cells.

Reads were demultiplexed using fgbio's DemuxFastqs (v1.4.0-c00bb7f-SNAPSHOT) with min-mismatch-delta and max-mismatches set to 1. Nf-Core/RNASeq v3.6 (10.5281/zenodo.1400710) was used to perform all downstream bioinformatic processing using the `remove_ribo_rna` option for rRNA removal. All other parameters were left as their defaults. All default containers were used except STAR's was changed to quay.io/biocontainers/star:2.7.10a--h9ee0642_0 and Salmon to quay.io/biocontainers/salmon:1.7.0--h10bb6b4_1. Summary statistics and plotting were conducted in R version 4.1.2, using tidyverse_1.3.1, ggpubr_0.4.0, and ggplot2_3.3.5.

Results

REPLICATE LIBRARY SEQUENCING RESULTS ON G4

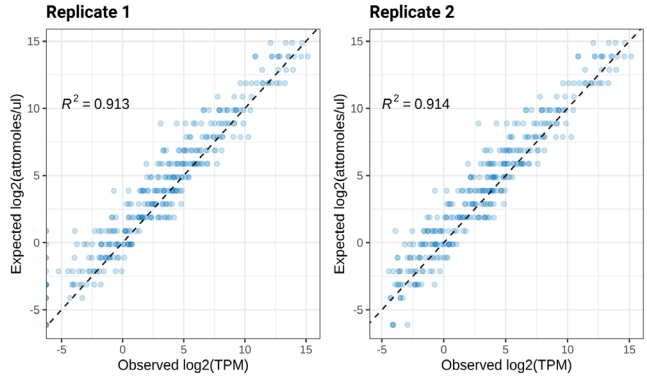
RNA sequencing run metrics from a study comparing replicate runs of Universal Human Reference (UHR) RNA samples are summarized in **Table 2**. Sample replicates were run using a 2 x 100 bp configuration. Replicates yielded 37-55M paired-reads with Q30 base quality scores for all replicates exceeding 86%.

Run Metrics	UHR Rep 1	UHR Rep 2
Read Configuration	2 x 100bp	2 x 100bp
Paired-Reads (M)	37M	55M
% Bases ≥ Q30 R1	87%	88%
% Bases ≥ Q30 R2	86%	89%

Table 2: RNA-Seq on the G4: Replicate run metrics.

Spike-recovery studies were conducted using ERCC spike-in Mix 1 into UHR replicates. Results are shown in Figures 1A & 1B. **Figure 1A** shows that a high correlation between expected and observed ERCC counts was achieved in both replicates (R1, R^2=0.913; R2, R^2=0.914). **Figure 1B** shows very high correlation (R^2=0.99) across transcript counts in replicate runs of Poly-A selected libraries and sequenced on the G4.

(1A) High Expected vs. Observed ERCC Count Correlation



(1B) High Correlation Across Transcript Counts

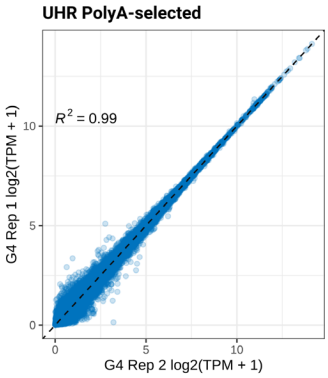
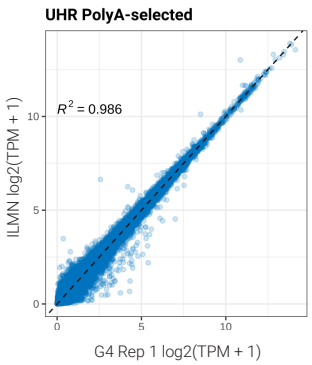
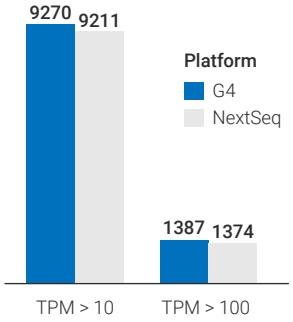


Figure 1A-1B: High correlation across replicates. (1A) Expected vs. observed ERCC count correlation. (1B) Correlation across transcript counts.

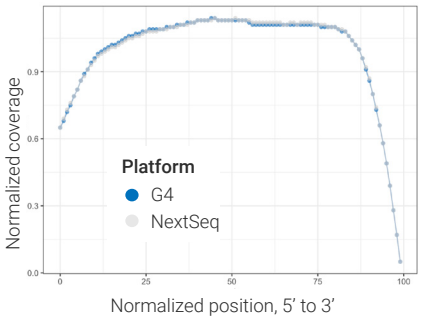
(2A) Correlation of transcript counts



(2B) Number of detected genes



(2C) Gene body coverage uniformity



(2D) Read distribution across genic and intergenic regions

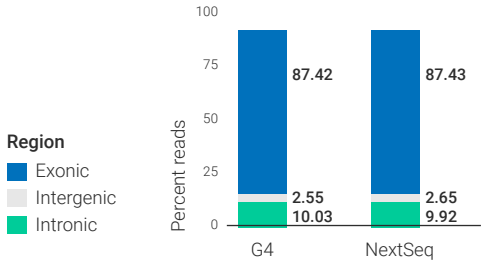


Figure 2A-2D: High correlation between G4 and NextSeq™ 2000. (2A) Correlation of transcript counts. (2B) Number of detected genes. (2C) Gene body coverage uniformity. (2D) Read distribution across genic and intergenic regions.

COMPARISON OF G4 AND NEXTSEQ™ 2000 RNA SEQUENCING

UHR Poly-A selected replicate libraries were run on the G4 and Illumina® NextSeq 2000 platforms. **Figure 2A** shows a high correlation of transcript counts (R^2=0.986) between the systems. The number of detected genes between the platforms was measured and compared (**Figure 2B**) in transcripts per million (TPM), both for expression levels of >10 TPM and >100 TPM. A comparable number of genes was detected by each platform, at each expression level.

Read coverage uniformity across genes was compared between platforms (**Figure 2C**). The normalized coverage plots by gene position show nearly identical coverage uniformity between platforms. Finally, read distribution across exonic, intergenic, and intronic regions were compared (**Figure 2D**). The percentage of reads obtained from the various genomic regions was highly comparable between the platforms.

Conclusion

RNA sequencing data generated by the G4 demonstrates performance comparable to the Illumina® NextSeq 2000 platform. Notably, the gene expression data generated by the G4 and NextSeq 2000 are highly correlated.

The G4 is a plug-and-play solution for RNA-Seq workflows that is compatible with existing laboratory ecosystems. The unique flow cell flexibility and unmatched run times of the G4 offer labs the ability to scale operations to match demand and reduce turnaround times on results.

RNA sequencing with the G4 provides users with added flexibility to tailor run sizes and flow cell configurations to the sample set, rather than holding samples to massively pool onto large flow cells. Less waste, reduced turnaround times and controlled costs can be realized by incorporating the G4 into your RNA sequencing operations.

*FASTQ files from this study are available by request for additional analysis.

REFERENCES

1. Challenges. *Frontiers in Genetics vol.* 11 220 (2020).
2. Buzdin, A. *et al.* RNA sequencing for research and diagnostics in clinical oncology. *Seminars in Cancer Biology vol.* 60 311–323 (2020).
3. Byron, S. A., Van Keuren-Jensen, K. R., Engelthaler, D. M., Carpten, J. D. & Craig, D. W. Translating RNA sequencing into clinical diagnostics: Opportunities and challenges. *Nature Reviews Genetics vol.* 17 257–271 (2016).
4. Levin, J. Z. *et al.* Comprehensive comparative analysis of strand-specific RNA sequencing methods. *Nat. Methods* **7**, 709–715 (2010).
5. Conesa, A. *et al.* A survey of best practices for RNA-seq data analysis. *Genome Biology vol.* 17 1–19 (2016).

SINGULAR GENOMICS

Get in Touch with the Customer Care Team

The purchase of a G4 comes with the assistance of an expert team to help you every step of the way. Our customer care team will assist you with order placement and can rapidly address any questions. Our field service engineers (FSE) ensure a successful installation and provide instrument support and our field application scientists (FAS) conduct training and validation of the desired application. Singular Genomics is committed to top-tier frontline support of users when and where it's needed.

RNA-Seq data is available by request. Please contact customer support at **care@singulargenomics.com**.

Begin Your Journey with G4

Contact our sales team to learn more about how the G4 can transform your sequencing workflows.

Website: www.singulargenomics.com
Email: care@singulargenomics.com
Call: +1 442-SG-CARES (442-742-2737)
Address: 3010 Science Park Rd, San Diego, CA 92121