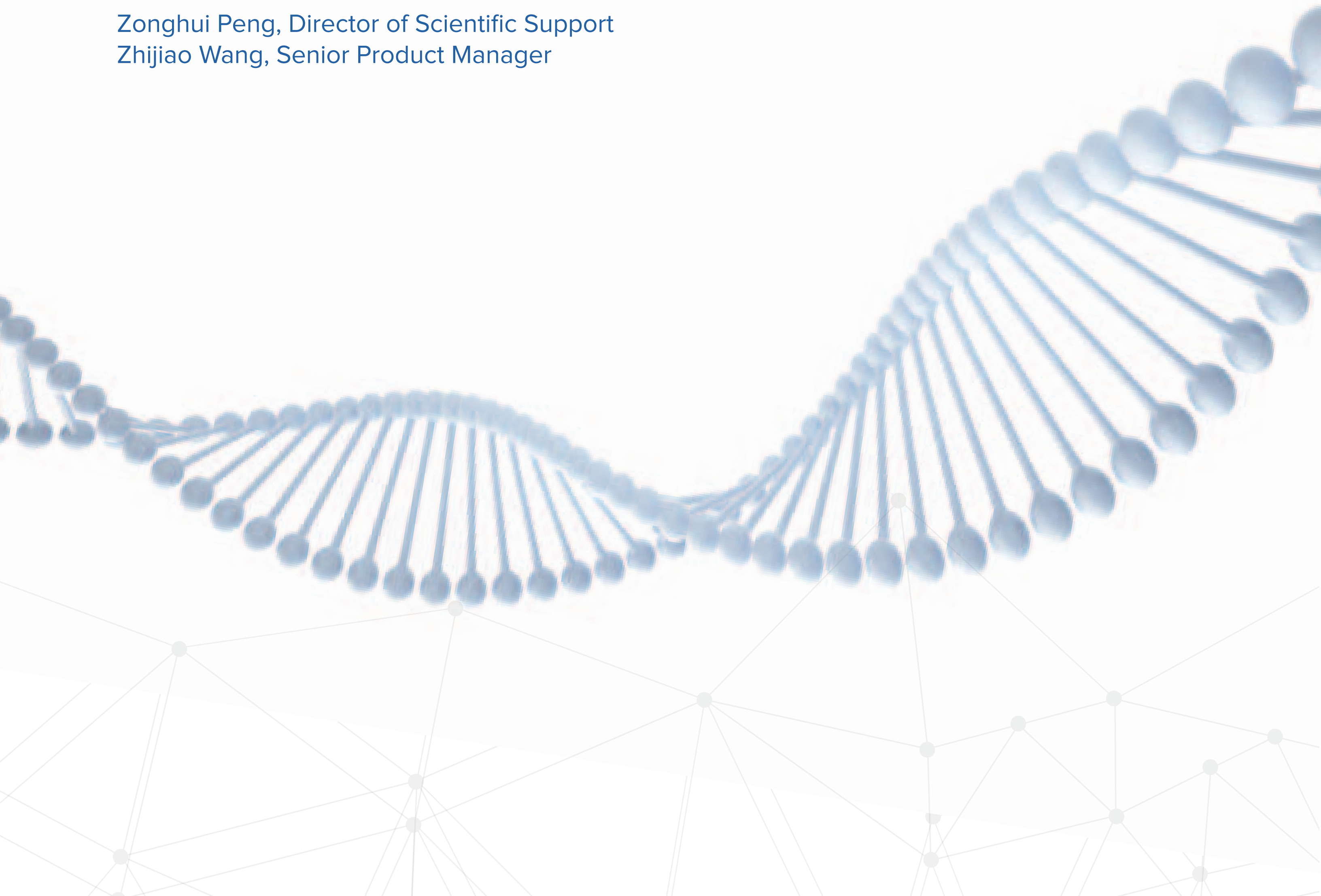
White Paper

GLOBAL GENOMICS SERVICES

State-of-the-art RNA Sequencing for Drug Discovery

Zonghui Peng, Director of Scientific Support
Zhijiao Wang, Senior Product Manager



Contents

1. Introduction

RNA molecules play important roles in various biological processes and have become an important point of focus in drug discovery and development. RNA sequencing has emerged as the most promising technology for both quantification and analysis; compared with RT-PCR and microarrays, sequencing offers a larger dynamic range and higher sensitivity while eliminating the bias inherent to microarrays.

Gene expression profiling is one major application of RNA sequencing that has been widely used by pharma R&D scientists in recent years, and the growing gene expression market has seen a significant shift from microarray technology to RNA sequencing because of the technical advantages and because the cost of sequencing has decreased to a level similar to that of microarray analysis. Meanwhile, variants at the RNA level, such as SNPs/InDels, gene fusions, and splicing events, have also been explored by researchers.

BGI Genomics is a leading provider of sequencing services for research and drug development. With a strong history of scientific contribution in the field of genomics, BGI Genomics offers world-renowned technological and global partnership experience to academic and pharmaceutical clients for projects of any size.

In this white paper, we describe the latest developments in RNA sequencing and demonstrate how BGI Genomics' experience sequencing more than 200,000 RNA samples has resulted in the best standardized processes using the newest technology in state-of-the-art laboratory infrastructure to consistently deliver high-quality results that are valuable for biomarker discovery and development.

2. Technical challenges and solutions

2.1 RNA sequencing quality control

RNA sequencing experiments may adapt to different input samples including whole blood, formalin-fixed paraffin-embedded (FFPE) tissue samples, fresh frozen tissues, and cultured cell lines. Each of these sample types requires an optimized sample preparation protocol to ensure that it can be analyzed successfully. Sample preparation also varies based on the different RNA subtype that will be analyzed, e.g., mRNA, lncRNA, or miRNA.

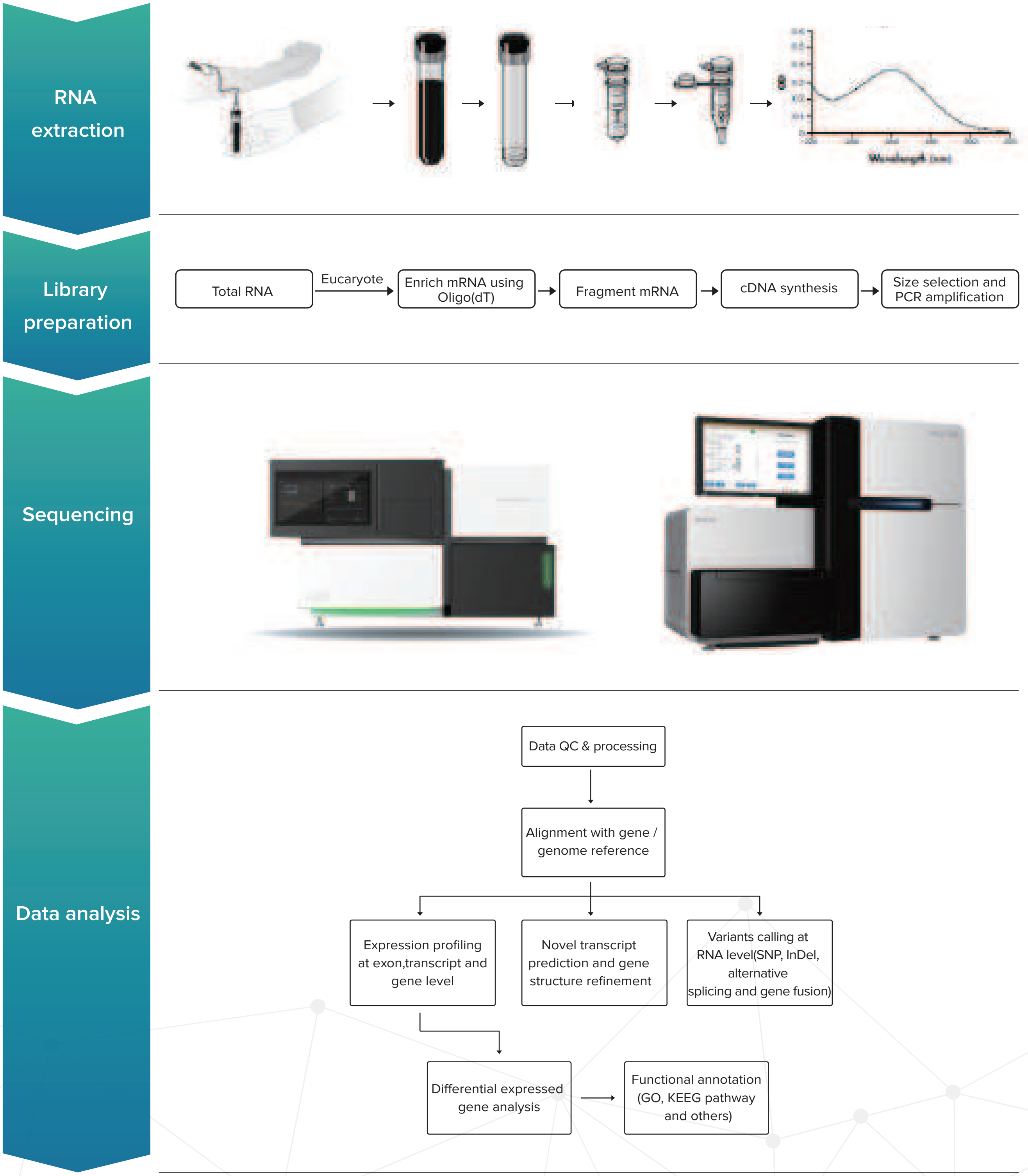
After samples are processed by sequencing machines, raw reads can be mapped to the reference genome, and various downstream analysis pipelines can be developed to quantify RNA molecules, discover novel transcripts, or better understand gene structures.

In 2014, the FDA coordinated a sequencing quality control consortium[1] to develop a comprehensive assessment of RNA-Seq experiments that includes accuracy, reproducibility, and information content measurements for splice junction discovery and differential expression profiling. BGI was a member of this consortium.

RNA sequencing process

RNA sequencing can be roughly divided into four steps: RNA extraction, library preparation, sequencing, and data analysis.

Different vendors have varying throughput capacity for each step, which can limit sample processing time. BGI has successfully extracted RNA from as many as 2,000 samples and completed library preparation and sequencing for up to 3,000 samples per month. In the following section, we use a whole blood sample as an example to illustrate the sequencing process and the associated quality control.

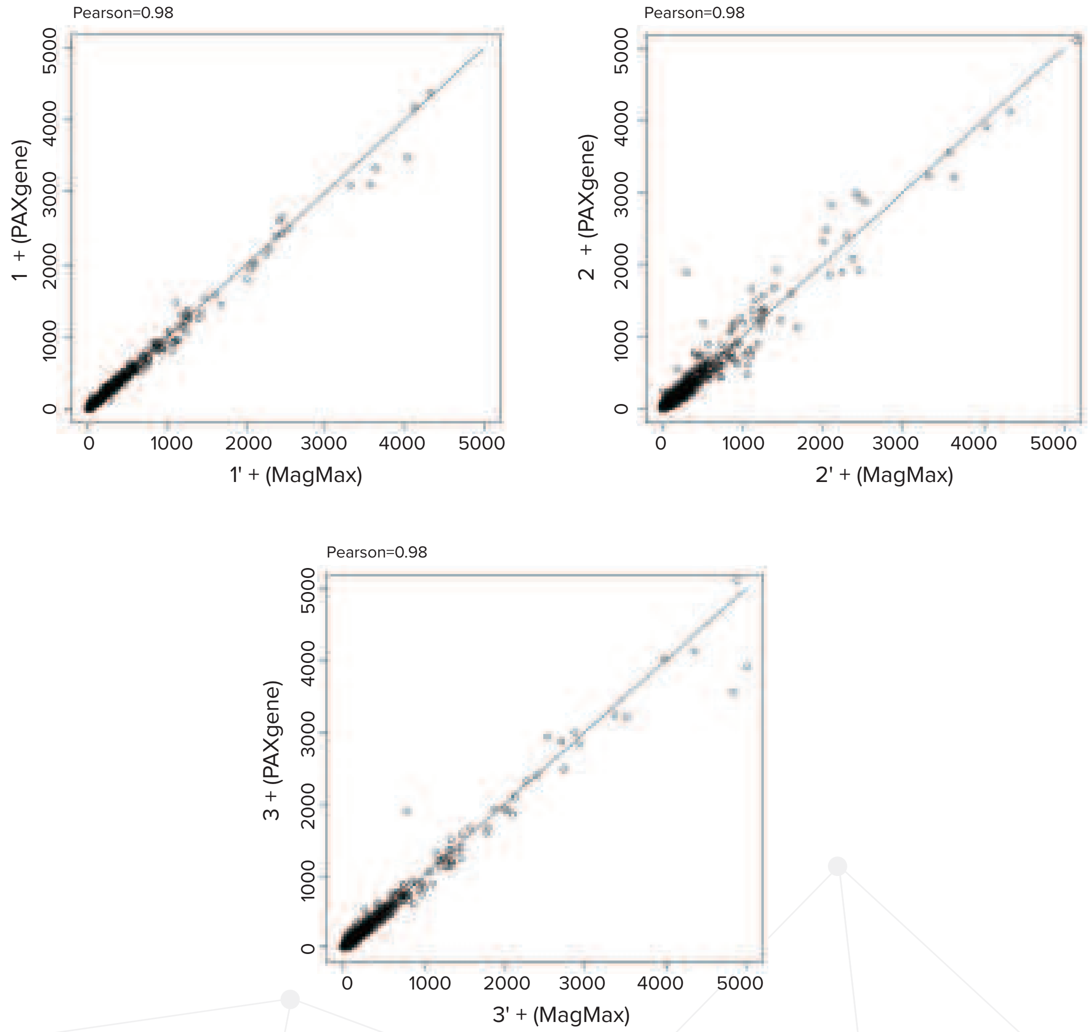

RNA extraction
Library preparation
Total RNA
Eucaryote
Enrich mRNA using Oligo(dT)
Fragment mRNA
cDNA synthesis
Size selection and PCR amplification
Sequencing
Data analysis
Data QC & processing
Alignment with gene / genome reference
Expression profiling at exon,transcript and gene level
Novel transcript prediction and gene structure refinement
Variants calling at RNA level(SNP, InDel, alternative splicing and gene fusion)
Differential expressed gene analysis
Functional annotation (GO, KEEG pathway and others)

RNA Sequencing workflow in BGI Genomics lab

Example: Whole blood sample RNA Sequencing

Step 1: RNA extraction

RNA can either be manually (Paxgene protocol) or automatically (MagMax protocol) extracted from whole blood samples collected in PAXgene tubes. The manual and automatic extraction protocols show strong correlation (Pearson correlation > 0.98).



Comparison between PAXgene RNA protocol and MagMax RNA protocol

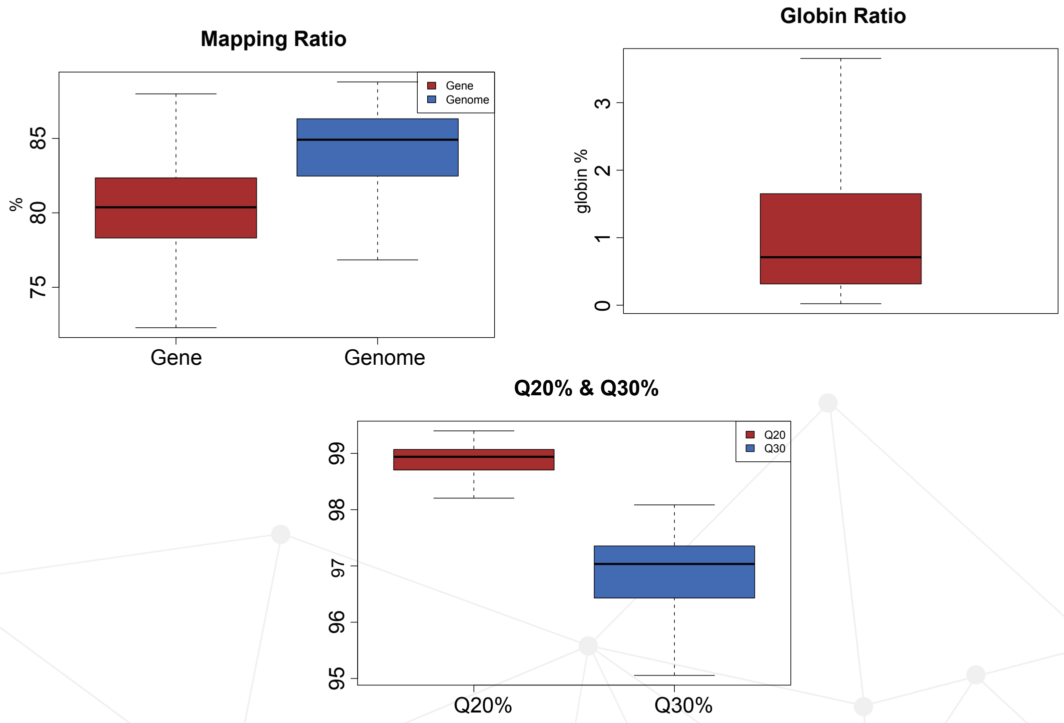
Step 2: Library preparation

For differential RNA expression profiling, two major sample types, those with an mRNA focus and those with a total RNA focus, were prepared. For mRNA library preparation, globin removal was performed to remove >95% of unwanted globin mRNA. For mRNA and lncRNA (total RNA focus) library preparation, an additional rRNA removal step was performed to remove cytoplasmic and mitochondrial rRNA.

Library preparation strategies

Protocol	Globin mRNA removal + poly-A enrichment	Globin mRNA and rRNA removal
Target	mRNA without globin mRNA	mRNA and lncRNA without globin mRNA or cytoplasmic/mitochondrial rRNA
Input	>200 ng Total RNA	>500 ng Total RNA
RIN	>7	>3

Samples with Q20 and Q30 scores



Data QC results for optimized BGI Genomics protocol for blood RNA

Step 3: Sequencing

To test our whole blood protocol, we randomly selected 500 blood samples for RNA extraction, library preparation, and sequencing. RNA reads were mapped to the gene (RefSeq) and genome (hg19). On average, over 80% of reads mapped to genes, and approximately 85% of reads mapped to the genome. Globin mRNA is present at less than 1%. Greater than 95% of the reads are Q30.

The raw outputs from the sequencer were processed through internal analysis pipelines to provide both RNA quantification and annotation information.

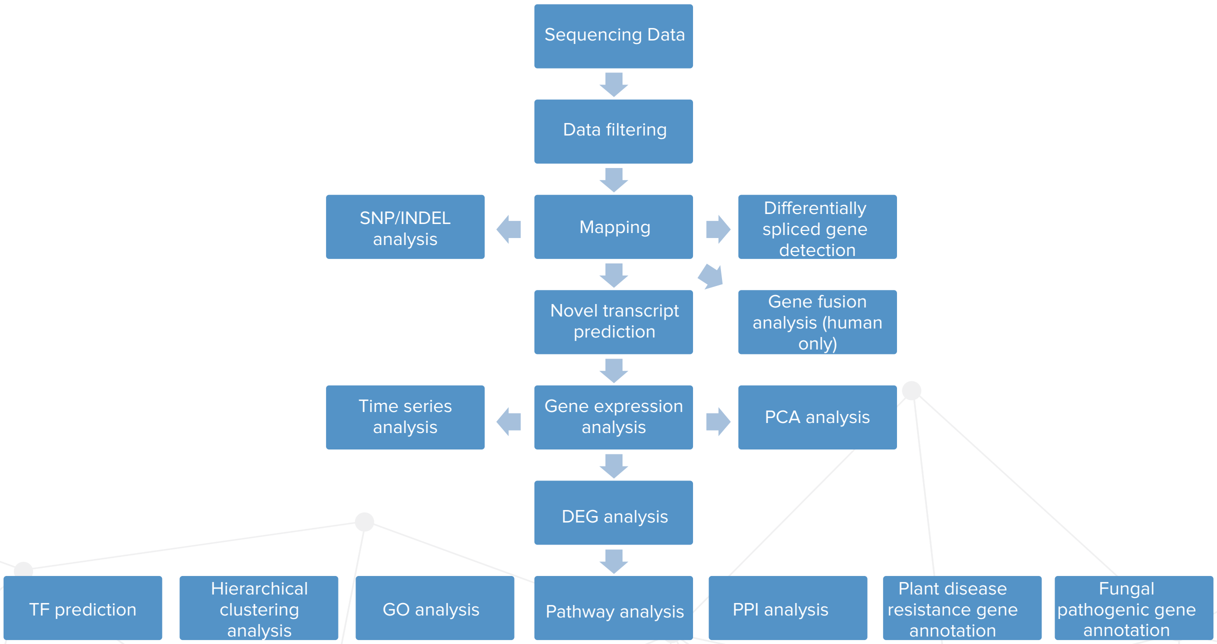
Step 4: Data analysis

In addition to the standard pipeline, we have engaged in customized solutions. Our current R&D efforts are focusing on the following:

1) Variant calling, such as SNPs, InDels, and gene fusions;

2) Integrated analysis, including small RNA-RNA-Seq integrated analysis, lncRNA-small RNA integrated analysis, and RNA Seq-proteomic quantification integrated analysis;

3) The validation and prioritization of mutations detected at the RNA level, e.g., neoantigen prediction.


Sequencing Data
Data filtering
SNP/INDEL analysis
Mapping
Differentially spliced gene detection
Novel transcript prediction
Gene fusion analysis (human only)
Time series analysis
Gene expression analysis
PCA analysis
DEG analysis
TF prediction
Hierarchical clustering analysis
GO analysis
Pathway analysis
PPI analysis
Plant disease resistance gene annotation
Fungal pathogenic gene annotation

BGI Genomics standard bioinformatics pipeline

Internal experiment to test reproducibility

We conducted an internal experiment including 16 different batches of samples, each batch consisting of ~100 samples. Each sample batch included 1 blood-derived RNA standard control, and the batches were processed between July 2016 and December 2016. We compared the RNA control samples between different batches and consistently found a Pearson correlation of 0.99 or higher.

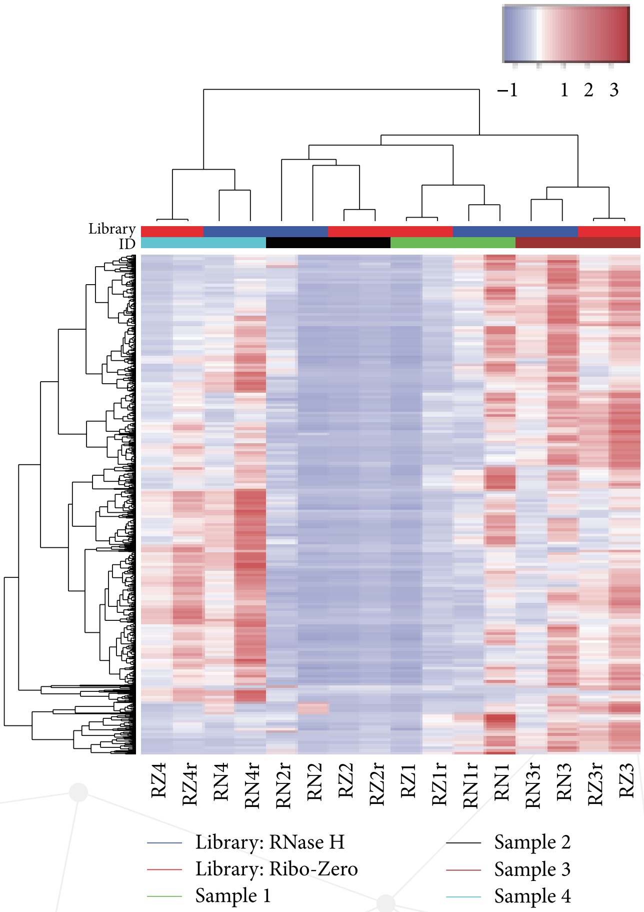
Reproducibility of blood RNA sequencing among batches

Sample	blood control _1	blood control _2	blood control _3	blood control _4	blood control _5	blood control _6	blood control _7	blood control _8	blood control _9	blood control _10	blood control _11	blood control _12	blood control _13	blood control _14	blood control _15	blood control _16
blood control _1	1.0000															
blood control _2	0.9996	1.0000														
blood control _3	0.9995	0.9999	1.0000													
blood control _4	0.9994	0.9999	0.9999	1.0000												
blood control _5	0.9992	0.9997	0.9996	0.9998	1.0000											
blood control _6	0.9996	0.9998	0.9996	0.9998	0.9997	1.0000										
blood control _7	0.9997	0.9997	0.9996	0.9997	0.9997	0.9997	1.0000									
blood control _8	0.9917	0.9920	0.9923	0.9922	0.9922	0.9922	0.9918	1.0000								
blood control _9	0.9994	0.9998	0.9997	0.9999	0.9998	0.9998	0.9996	0.9921	1.0000							
blood control _10	0.9997	0.9998	0.9996	0.9997	0.9996	0.9998	0.9999	0.9919	0.9997	1.0000						
blood control _11	0.9986	0.9999	0.9992	0.9991	0.9990	0.9990	0.9989	0.9925	0.9990	0.9989	1.0000					
blood control _12	0.9991	0.9997	0.9996	0.9999	0.9998	0.9997	0.9996	0.9919	0.9999	0.9997	0.9989	1.0000				
blood control _13	0.9995	0.9996	0.9997	0.9996	0.9992	0.9995	0.9993	0.9920	0.9997	0.9994	0.9987	0.9993	1.0000			
blood control _14	0.9995	0.9997	0.9996	0.9997	0.9994	0.9998	0.9995	0.9920	0.9999	0.9997	0.9988	0.9997	0.9998	1.0000		
blood control _15	0.9990	0.9997	0.9995	0.9998	0.9997	0.9996	0.9994	0.9918	0.9999	0.9995	0.9987	0.9999	0.9995	0.9997	1.0000	
blood control _16	0.9995	0.9991	0.9991	0.9988	0.9986	0.9988	0.9994	0.9915	0.9988	0.9993	0.9983	0.9986	0.9989	0.9988	0.9983	1.0000

FFPE samples

All FFPE samples were required to have an RNA concentration ≥ 1 ng/μl and a total quantity ≥ 20 ng. Their DV200 values vary, but most samples have greater than a 30% pass rate.

Despite the challenges associated with FFPE processing, over 10,000 FFPE samples were successfully processed between 2014 and 2017. Four major methods have been developed for FFPE. Three involve rRNA depletion protocols, such as the use of RNase H and Ribo-Zero. An exon capture protocol has been widely used recently because it captures the most mRNA lacking rRNA. In 2016, Guo Y. et al. published a comparison between the RNase H and Ribo-Zero protocols using 4 FFPE samples[2]. Their data showed that both protocols show good technical reproducibility for FFPE samples.



Unsupervised clustering for all detected RNAs.
Samples clustered first by replicates and then by the rRNA depletion method.

Low input solutions

Most of our standard protocols require at least 100 ng of input material. However, we can optimize SMARTer or NuGEN protocols for quantities less than 5 ng. The final output will depend on the specific samples and may vary from study to study. BGI Genomics has sequenced thousands of low-input samples, and high-quality data can be guaranteed if the optimized protocol is followed.

Low-input RNA sequencing data from BGI Genomics

Sample name	Total RNA input (ng)	Total bases of data	Mapping to genome	Mapping to gene	Gene number	Reproducibility
A1	9.26	2,400,000,120	88.96%	45.11%	18,652	S=0.9491 P=0.9074 Gene num=17,388
A2	10	2,400,000,120	88.77%	47.13%	18,134	
B1	2.91	2,400,000,120	85.89%	54.72%	17,219	S=0.9330 P=0.9606 Gene num=15,817
B2	1.11	2,400,000,120	87.94%	49.40%	17,115	
C1	10	2,400,000,120	89.74%	38.62%	18,907	S=0.9063 P=0.9479 Gene num=18,371
C2	10	2,400,000,120	95.18%	16.49%	19,466	

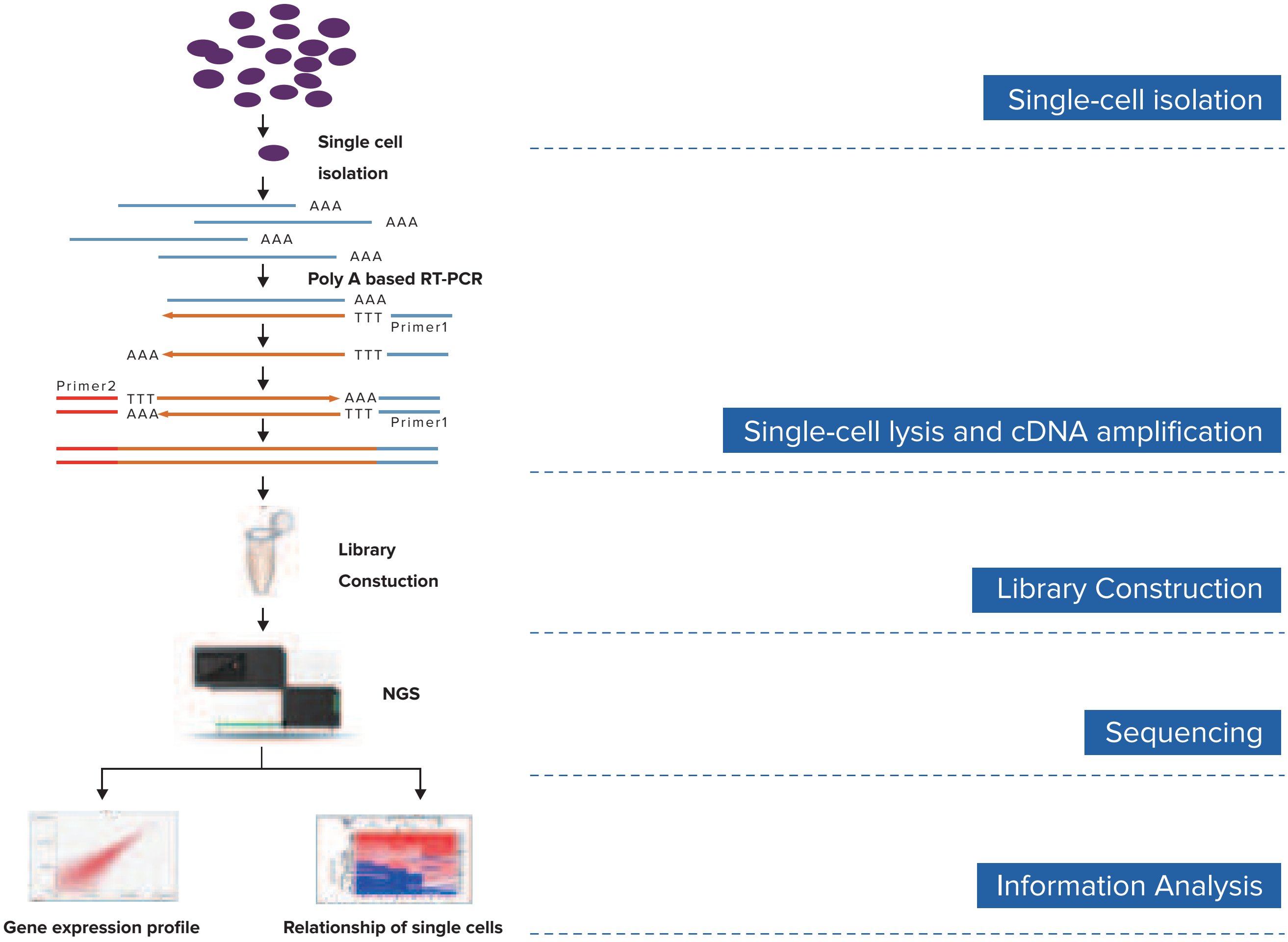
Note: “S” denotes Spearman parameter; “P” denotes Pearson parameter; “Gene num” denotes detected gene number.

Other technologies (partners)

We provide other technology solutions, including full-length single-cell RNA sequencing, high-throughput single-cell RNA sequencing on a 10X genomics platform, isoform sequencing (ISO-Seq), and gene quantification with NanoString.

Full-length single-cell RNA sequencing

Single-cell RNA sequencing was first developed by Tang et al.[3] BGI Genomics established a mouth-controlled pipetting method to isolate a single cell for subsequent single-cell whole exome sequencing[4, 5]. By combining single-cell isolation and transcriptome amplification, BGI Genomics offers full-length single-cell RNA sequencing that can reveal the whole transcriptome of a single cell.



Workflow of BGI Genomics single-cell RNA sequencing

10X Genomics single-cell RNA sequencing

In 2016, 10X Genomics launched the Chromium™ Single Cell 3′ Solution for 3' end gene quantification. This is a fully automatic and high-throughput platform based on partitioning and molecular barcoding technology. It can profile 100 – 80,000+ cells in a single run and requires only one day for library preparation from a cell suspension; this library can then be sequenced on the Illumina or BGISEQ platforms. Special parameters must be chosen for sequencing libraries prepared with this technology. Since its release, approximately 30 studies using this technology have been published.

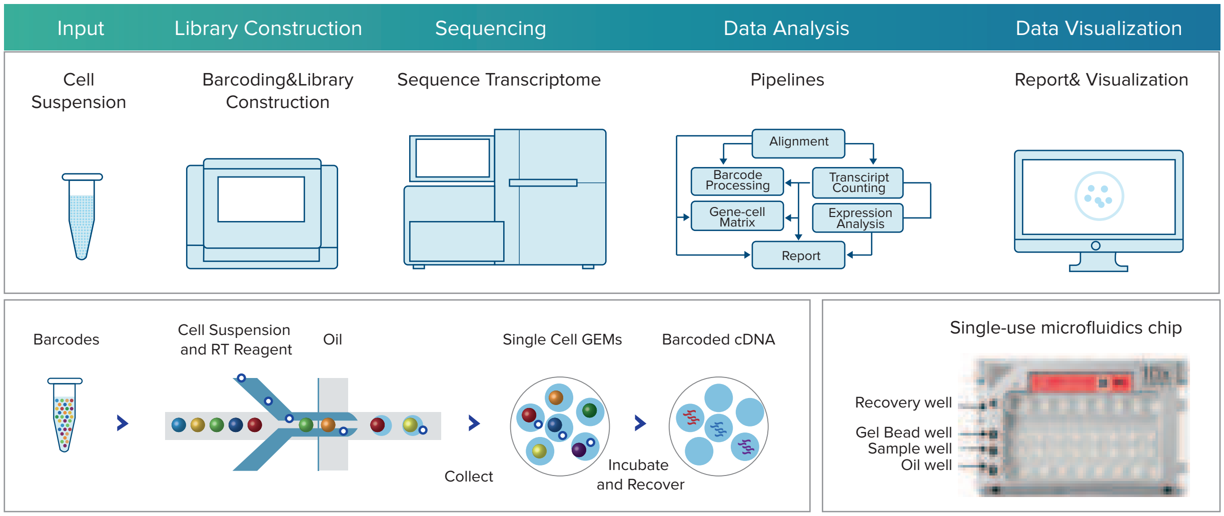


Figure 7. Workflow of 10X Genomics single-cell RNA sequencing

ISO-Seq on PacBio Sequel System

PacBio Sequel can generate ≥ 10 kb reads (Sequel 2.0 chemistry) with the longest reads > 60 kb, which allows us to sequence transcripts without fragmentation. Thus, ISO-Seq can identify known and novel alternative splicing more accurately than short reads-based platforms. No additional reconstruction relying on paired-end short reads is needed. Unlike traditional RNA sequencing, ISO-Seq directly detects the expression level of transcripts and provides better precision.

Gene quantification on NanoString nCounter FLEX System

We also partner with NanoString to provide alternative measurements of RNA profiling. The Nanostring nCounter FLEX system can detect as many as 800 molecules in one reaction based on its novel digital barcode technology and hybridization. Two ~50-base probes—a capture probe and a reporter probe—are designed for each gene. After hybridization, the probe and its target gene form a complex. The gene is then quantified by counting molecular "barcodes" without needing PCR. Because of this advantage, NanoString is widely used in biomarker development. Normally, RNA sequencing is used to reveal the whole transcriptome and identify potential biomarkers. These biomarkers can then be validated using the NanoString platform. Furthermore, a companion diagnostic could be developed using the NanoString system.

3. Applications of RNA-Seq in R&D

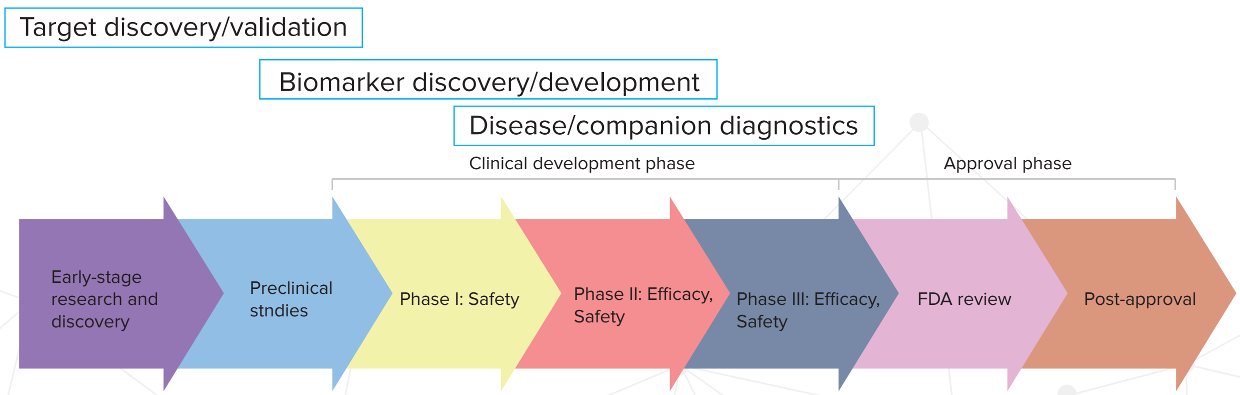
RNA sequencing can be applied in research & development efforts, which have traditionally relied on microarray analysis, and it can be useful in unique fields that can only be studied using this novel technology. For example, RNA sequencing can be used to detect fusion transcripts (RNA molecules that are joined from different genes), alternative splicing variants (particular exons of a gene are either included in or excluded from the RNA transcript), or novel transcripts (unknown mRNA or small RNA molecules from past studies). These findings may have broad applications in drug discovery.

RNA Sequencing, a more robust method for gene expression profiling

Microarray technology is an established method based on hybridization in which probes are synthesized based on known transcripts/genes. Comparison studies between RNA-Seq and microarrays reveal an overall high correlation for relative gene expression[6, 7]. In some cases, RNA-Seq provides a broader dynamic range that can enable the detection of low-level transcripts, and many gene expression events can be identified by RNA-Seq that are not represented by probes on microarrays[8]. RNA-Seq also allows the discovery of novel transcripts, e.g., splice/structural variants. Because RNA sequencing also reveals the nucleotide sequence of transcripts, useful information related to SNPs and InDels can also be obtained.

RNA sequencing in drug discovery and development

RNA sequencing can be useful for target discovery, target validation, biomarker discovery, biomarker development, disease monitoring, and companion diagnostics.

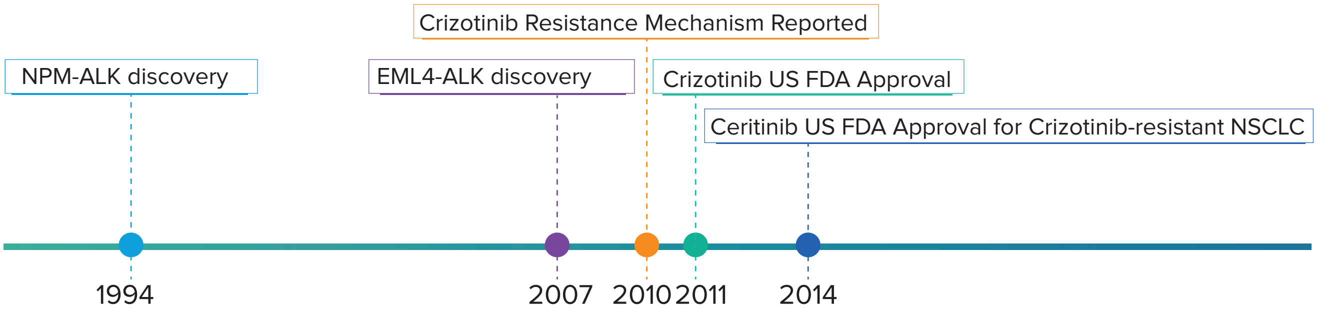

Target discovery/validation
Biomarker discovery/development
Disease/companion diagnostics
Clinical development phase
Approval phase
Early-stage research and discovery
Preclinical stndies
Phase I: Safety
Phase II: Efficacy, Safety
Phase III: Efficacy, Safety
FDA review
Post-approval

Adjuested from O' Driscoll the drug discovery and development pipeline

Applications of RNA sequencing in drug discovery

Fusion Transcript Target Discovery

The discovery of the EML4-ALK fusion gene was originally identified by a classic transformation assay from a single patient and validated in a small group of NSCLC patients[9]. Crizotinib is a tyrosine kinase inhibitor (TKI) targeting c-MET, ALK, and ROS1, and a diagnostic assay for ALK fusions enables the study of Crizotinib in ALK-positive patients. Crizotinib was approved in 2011, four years after the fusion gene was discovered.


Crizotinib Resistance Mechanism Reported
NPM-ALK discovery
EML4-ALK discovery
Crizotinib US FDA Approval
Ceritinib US FDA Approval for Crizotinib-resistant NSCLC
1994
2007
2010
2011
2014

History of Ceritinib development

EML4-ALK is more easily identified by DNA sequencing (rearrangements) or RNA sequencing (fusion transcript). Other fusion transcripts have been identified by RNA-sequencing, e.g., TMPRSS2-ERG and TMPRSS2-ETV1 in prostate cancer[10]. A number of sequencing panels in the clinical space include known fusion genes. For example, the FoundationACT Panel contains assays for ALK, EGFR, FGFR3, PDGFRA, RET, and ROS1.

RNA Sequencing in Biomarker Discovery

Biomarker discovery has generally been accomplished using microarray experiments. However, the microarray platforms only cover a limited number of splicing variants. It is therefore a common practice to focus on the average expression or the consensus transcript in biomarker discovery. A recent study using RNA-Seq showed that specific isoform expression levels were associated with drug response, and this information would not have been observed by focusing on overall gene expression[11]. This result illustrates the potential of RNA-Seq in biomarker discovery and suggests new opportunities.

BGI Genomics RNA sequencing assays for biomarker discovery

Assay type	Sample type
Transcriptome Sequencing	Tissue, FFPE, blood, cells, and others
ISO-Seq on PacBio	Tissue, blood, cells, and others
RNA-Seq (Quantification)	Tissue, FFPE, blood, cells, and others
Long Non-coding RNA (lncRNA) Sequencing	Tissue, FFPE, blood, cells, and others
Single-cell RNA Sequencing	Tissue, blood, cells, and others
Prokaryotic Transcriptome Sequencing	Culture medium, strains, and others
Targeted RNA-Seq (NanoString platform)	Tissue, FFPE, blood, cells, and others
Metatranscriptome	Soil, stool, blood, and others

Note: For each sample type, BGI Genomics provides optimized/proprietary protocols for RNA extraction, library preparation and data analysis.

Track record for RNA Seq biomarker discovery at BGI Genomics

Type of biomarker	Biomarker candidate	Disease	Publication	Date
Expression	TIMP1 overexpression	Breast cancer	Tumor Biology	2013
	ALDH2, CCNE1, and SMAD3	Upper tract urothelial carcinoma	BMC Cancer	2014
	JUN, TNFAIP3, TOB1, GIMAP4, GIMAP6, TRMT112, NR4A2, CD69, and TNFSF8	Paroxysmal nocturnal hemoglobinuria	Journal of Immunology	2017
Fusion	TM-PRSS2-ERG	Prostate cancer	Cell Research	2012
lncRNA	NBAT-1	Neuroblastoma	Cancer Cell	2014
	FENDRR and LINC00511	Lung adenocarcinoma	Lung Cancer	2017

Between 2011 and 2017, we processed a total of 216,620 samples by RNA sequencing with consistent growth (20-30% compound annual growth rate).

Track record of BGI Genomics RNA sequencing projects

Sample type	Sample number	Success rate	Percentage
FFPE	10,831	90%	5%
Blood	32,493	98%	15%
Cell line	21,662	98%	10%
Total RNA	151,634	98%	70%
Total	216,620		100%

Note: "Total RNA" includes the extracted RNA samples from FFPE, blood, and other normal samples.

RNA Sequencing in biomarker development and companion diagnostics

In the development of personalized medicine, the ultimate goal for biomarkers discovered by RNA sequencing is to provide high predictive value for diagnosing, monitoring, and stratifying cancer patients. After performing high-throughput biomarker screening via RNA sequencing, researchers normally proceed with research use only (RUO) assay development, companion diagnostic (CDx) assay development & validation, and finally application for FDA approval. For example, in 2014, Scott et al. published a report showing that a 20-gene expression set is a useful biomarker for stratifying patients with diffuse large B-cell lymphoma (DLBCL). After evaluation and according to the results of this scientific paper, Celgene has been collaborating with NanoString to develop a CDx test for REVLIMID that will be used to screen patients who are being enrolled in a pivotal study for the treatment of DLBCL[12]. A similar example exists for Merck; Merck and NanoString are initiating a collaboration using a gene-expression biomarker for CDx assay development for KEYTRUDA.

4. Conclusion

As a highly sensitive and accurate NGS tool, RNA sequencing enables researchers to study the transcriptome in a more precise and revolutionary way; RNA sequencing reveals otherwise undetected expression changes, enables novel isoform quantification, and exposes gene splicing/fusion events and RNA editing that occurs in response to different types of disease, especially in oncology.

Here, we summarized the technical challenges of RNA extraction, library preparation, sequencing, and data analysis. To overcome these challenges, BGI Genomics has established the most strict and comprehensive quality control systems and developed proprietary/optimized protocols, especially for clinical patient samples (whole blood). The rigor of our pipeline is demonstrated by our sample data. For challenging samples, such as highly degraded FFPE, or extremely low quantity samples, optimized and validated protocols are available from BGI Genomics. Additionally, to meet researchers' particular study needs, such as novel alternative splicing detection, gene expression profiling at the single cell level, novel transcript detection, or non-PCR amplification-based gene expression screening, BGI Genomics has developed and effectively used an optimized single-cell assay, the 10x Genomics platform, PacBio ISO-Seq technology, and the NanoString platform. Furthermore, examples of how RNA sequencing can be applied in drug development, including for target discovery, biomarker discovery, biomarker development, disease monitoring, and companion diagnostics, are briefly described. As one of the largest global NGS service providers, BGI Genomics is always updating our service portfolio and contributing our expertise to technology development.

By continuing to improve and innovate protocols/methods for RNA sequencing technologies, more unbiased and more consistent results can be generated. The RNA sequencing technologies applied in pharmaceutical R&D may have many additional uses and further our understanding of how the transcriptome changes in response to disease and treatment. Finally, with contributions from other related experts in the NGS field, we believe RNA sequencing can deeply penetrate the different processes and stages of the drug development pipeline to allow pharmaceutical companies and researchers to comprehensively use this technology in the right time with the right parameters.

References

1. SEQC Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control consortium. Nature biotechnology. 2014.
2. Guo, et al. RNA Sequencing of Formalin-Fixed, Paraffin-Embedded Specimens for Gene Expression Quantification and Data Mining. International Journal of Genomics. 2016.
3. Tang, et al. mRNA-Seq whole-transcriptome analysis of a single cell. Nat. Methods. 2009.
4. Hou, et al. Single-Cell Exome Sequencing and Monoclonal Evolution of a JAK2-Negative Myeloproliferative Neoplasm. Cell. 2012.
5. Xu, et al. Single-Cell Exome Sequencing Reveals Single-Nucleotide Mutation Characteristics of a Kidney Tumor. Cell. 2012.
6. Su, et al. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nature Biotechnology, 2014.
7. Zhao et al. Comparison of RNA-Seq and Microarray in Transcriptome Profiling of Activated T Cell. Plos One. 2014.
8. Zhang, et al. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome Biology. 2015.
9. Soda, et al. Identification of the transformingEML4–ALK fusion gene in non-small-cell lung cancer. Nature. 2007.
10. Tomlins, et al. Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer. Science. 2005.
11. Birzele, et al. CD44 Isoform Status Predicts Response to Treatment with Anti-CD44 Antibody in Cancer Patients. Clinical Cancer Research. 2015.
12. http://investors.nanostring.com/releasedetail.cfm?ReleaseID=852251

Request for more information

For more information, or to discuss how we can meet your specific project requirements, please visit www.bgi.com, write to info@bgi-international.com or contact your local BGI representative.

BGI Americas
One Broadway,
Cambridge, MA 02124,
USA
Tel: +1 617 500 2741

BGI Europe
Ole Maaløes Vej 3,
DK-2200 Copenhagen N
Denmark
Tel: +45 7026 0806

BGI Asia-Pacific
16 Dai Fu Street,
Tai Po Industrial Estate,
New Territories, Hong Kong
Tel: +852 36103510

Published January 2018.

All Services and Solutions are for research use only.

BGISEQ-500 Services are executed at our service laboratory in China.

We Sequence, You Discover

BGI