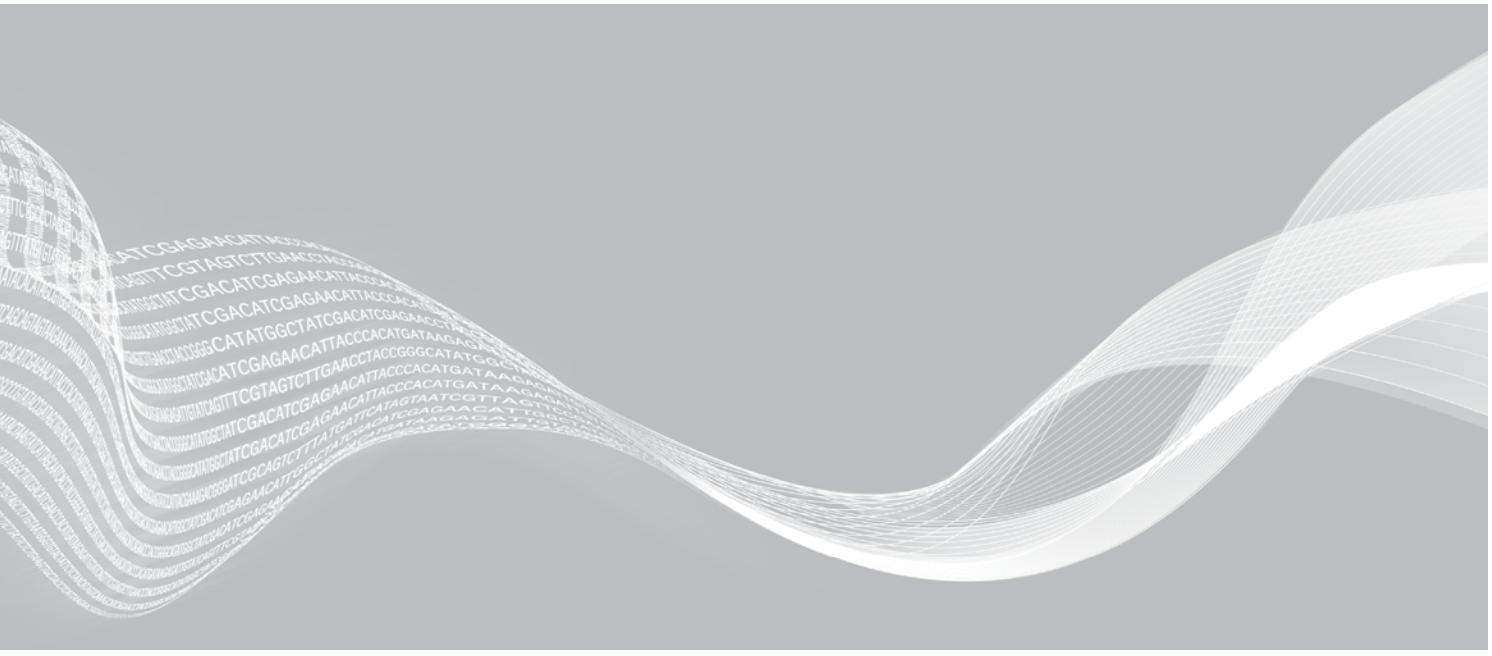
RNA-Based Drug Response Biomarker Discovery and Profiling

Section 3: Best Practices



Application solutions: RNA Drug Response Biomarker Discovery and Screening

RNA sequencing (RNA-Seq) is increasingly being utilized for the discovery of and profiling for RNA-based drug response biomarkers with the aim of improving the efficiency and success rate of the drug development process. While a number of technologies have been used for this application, the capabilities of RNA sequencing promise to be of particular benefit [1,3,4]. Consequently, there is a growing need to extend the accessibility of RNA sequencing-based workflow solutions for this application to a broader range of potential users, including those without prior experience with next-generation sequencing (NGS).

Towards that end, this document is designed to serve as a comprehensive resource for prospective users of any level of NGS experience who are considering adopting this application. It contains information that we have found to be particularly helpful to users across multiple stages of the process, from understanding the steps of an RNA sequencing workflow, to matching configuration options to specific program requirements, to preparing a plan for rapid navigation through the implementation process.

Application Overview	Workflow Introduction	Best Practices	Start-up Advice	Analysis Pipeline Review
An introduction to RNA-Seq drug response biomarker discovery and profiling	Key considerations, requirements and recommended components for multiple application use-cases	"How-to" guidance to facilitate workflow implementation	Tips from fellow application users and Illumina experts on how to get up and running quickly and smoothly	A screenshot-based walk-through from raw data through outputs needed to inform candidate assessment and prioritization

Section 3: Best Practices

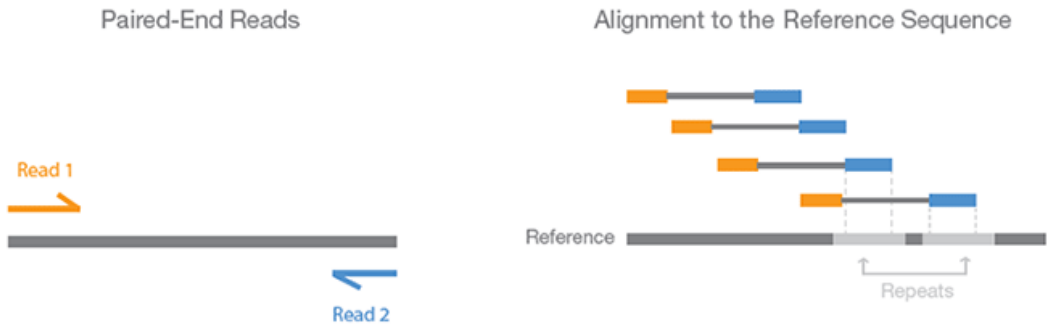
Sequencing parameters

For users who are new to NGS, one of the first considerations when beginning the adoption process is how to design a sequencing run. Our internal and field support teams respond regularly to a range of questions on this subject. The following section is geared to address these questions as they pertain to RNA-based drug response biomarker discovery and profiling.

Single-read vs paired-end sequencing

What it refers to:
These terms refer to whether the sequencing by synthesis (SBS) chemistry extends from only one end of a transcript insert captured in the library, or both.

Figure 4. Paired-End Sequencing and Alignment



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.

Source: http://www.illumina.com/content/dam/illumina-marketing/images/technology/paired-end-sequencing-figure.gif

Why it's important:
Paired-end sequencing will double the amount of data generated from each insert in the library, providing sequence data originating from each of the two ends. This will benefit the percent of reads output by the sequencer that can be uniquely aligned to the reference sequence. From the standpoint of this application, additional sequence context is provided that may benefit the ability to detect gene fusions, indels, and single nucleotide variants (SNVs), as well as call alternative transcript isoforms.

Recommendation:
The value of paired-end versus single-read sequencing depends on the goals of the experiment. For RNA-based drug response biomarker discovery and profiling, paired-end sequencing is recommended, as it maximizes your ability to capture candidate biomarkers in the categories previously discussed.

There are 2 exceptions to this recommendation:

- If you are running TruSeq RNA Access, are only interested in the detection of expression-based biomarkers, and only need to detect gene-level rather than expression-level data, single-read sequencing may meet your needs.
- If you are profiling using TruSeq Targeted RNA, given the targets are pre-defined and are pre-filtered for unique sequence context, only single-read sequencing is necessary.

Read length

What it refers to:

Read length refers to the number of bases for which the SBS chemistry extends along the library template and returns sequence data

Why it is important:

The read length represents continuous sequence data along the template, which—similar to paired-end versus single- read sequencing—offers benefits:

- It delivers a higher percent of reads that can be successfully, uniquely aligned to the reference sequence (ie, the more sequence context that exists, the less likely there will be multiple, identical sequences across the transcriptome).
- It is able to call features, such as indels, SNVs, and gene fusions.
- It can distinguish between and call alternative transcript isoforms.

Recommendation:

If you wish to detect candidate biomarkers based on SNVs, gene fusions, and transcript-level abundance measurement, 2x75 bp sequencing is recommended. Internal analysis has shown that this format is robust in the detection of these features, whereas extending the read length offers diminishing returns. This is largely due to the fact that both the TruSeq RNA Access and TruSeq Stranded Total RNA library prep methods generate a median insert size of 150 bp. The insert size refers to the length of the sequence template that is captured in the library. Given a 150 bp length, 2x75 sequencing will cover the complete sequence with 2 reads, starting on each end of the insert and meeting at the center.

Read depth

What it refers to:

Read depth refers to the number of reads, or sequencing output, allotted to each sample in a sequencing run.

Tissue-Specific Expression Detected with RNA-Seq

Source: http://www.illumina.com/content/dam/illumina-marketing/images/technology/mRNA_Seq_lg.gif

Why it's important:
The amount of sequencing data per sample will determine how much information one will be able to derive from the run. For example, sufficient sequence data is required to enhance the process as follows:

- Accurately measure expression levels.
- Delineate between transcript isoforms.
- Detect SNVs, gene fusions, and indels (particularly if not previously observed).

Recommendation:
A number of published manuscripts have included recommendations on per-sample read depth across a variety of NGS methods.[24–25] The depths that our internal groups routinely use for the methods highlighted for this application are captured in the table below. If you wish to analyze use-case datasets to inform your preferences for per-sample read depth, data are available on the BaseSpace® Suite (add link here to be provided by client).

Recommended read depth (FFPE)	Recommended usage	Recommended per-sample read depth (FFPE)
TruSeq RNA Access	Biomarker discovery – coding transcriptome	≤ 25M reads
TruSeq Stranded Total RNA	Biomarker discovery – coding + ncRNA transcriptome	≤ 100M reads
TruSight Pan-Cancer Panel	Biomarker profiling/focused discovery	≤ 3M reads
TruSeq RNA Access Custom	Biomarker profiling/focused discovery	Dependent on complexity
TruSeq Targeted RNA	Biomarker profiling	1000 reads per target/per sample

Other considerations impacting your project plan

Multiplexing

Multiplexing refers to the capability to run multiple samples in parallel through the use of uniquely identifiable indexes that allow reads generated from each sample to be analyzed separately. It allows for a more efficient use of sequencing output, such that it may be split to generate the desired read depth for each sample.

Flow cell configuration

Flow cells (see Sequencing in *Section 2: Workflow Introduction*) are designed in multiple configurations consisting of 1, 2, or 8 physically separated lanes depending on the platform. The configuration of the flow cell you use will impact both the output of the run and how that output is divided across the flow cell. Flow cell configuration can differ between sequencer modes, as described below.

Instrument mode

Certain instruments, including the NextSeq Series, HiSeq 2500, and MiniSeq Platform may be run on multiple modes. The mode will determine the speed of the run and/or the amount of output generated.

Cluster and SBS kit versions

Depending on the sequencer being used, there may be more than one version of the core reagent kits used to operate it. The version you select may impact the amount of sequencing output generated per run.

TruSeq RNA Access

Sequencer	Mode	Lanes per FC	FC capacity	Output/lane (clusters)	Samples/ lane	Samples/run	Reagent kit	Time
NextSeq Series	High-output (HO)	1	Single	400M	16	16	HO kit 150 cycles FC-404-1002	18 hrs
	Mid-output (MO)	1	Single	130M	5	5	MO kit 150 cycles FC-404-1001	15 hrs
HiSeq 2500	High-output	8	Dual	Up to 250M	10	80 (single FC) 160 (dual mode)	• HiSeq PE Cluster Kit v4 – cBot • HiSeq SBS Kit v4 FC-401-4002 (x3)*	~4 days
	Rapid run	2	Dual	Up to 150M	6	12 (single FC) 24 (dual mode)	• HiSeq PE Rapid Cluster Kit v2 • HiSeq Rapid SBS Kit v2 (50 cycle) (x3)*	~22 hrs
HiSeq 3000	N/A	8	Single	Up to 312M	12	96	• HiSeq 3000/4000 Cluster Kit PE-410-1001 • HiSeq 3000/4000 SR Cluster Kit GD-410-1001	~2 days
HiSeq 4000	N/A	8	Dual	Up to 312M	12	96 (single FC) 192 (dual FC)	• HiSeq 3000/4000 Cluster Kit PE-410-1001 • HiSeq 3000/4000 SR Cluster Kit GD-410-1001	~2 days

*Each kit provides reagents for 50 cycles. With a total of 150 cycles (at 2x75 bp) needed for the run, 3 kits (ie, x3) are needed per run.

TruSeq Stranded Total RNA

Sequencer	Mode	Lanes per FC	FC capacity	Output/lane (clusters)	Samples/ lane	Samples/run	Reagent kit	Time
NextSeq Series	High-output (HO)	1	Single	400M	4	4	HO kit 150 cycles FC-404-1002	18 hrs
	Mid-output (MO)	1	Single	130M	1	1	MO kit 150 cycles FC-404-1001	15 hrs
HiSeq 2500	High-output	8	Dual	Up to 250M	2	20 (single FC) 40 (dual mode)	• HiSeq PE Cluster Kit v4 – cBot • HiSeq SBS Kit v4 FC-401-4002 (x3)*	~ 4 days
	Rapid run	2	Dual	Up to 150M	1	3 (single FC) 6 (dual mode)	• HiSeq PE Rapid Cluster Kit v2 • HiSeq Rapid SBS Kit v2 (50 cycle) (x3)*	~ 22 hrs
HiSeq 3000	N/A	8	Single	Up to 312M	3	24	• HiSeq 3000/4000 Cluster Kit PE-410-1001 • HiSeq 3000/4000 SR Cluster Kit GD-410-1001	~ 2 days
HiSeq 4000	N/A	8	Dual	Up to 312M	3	25 (single FC) 50 (dual FC)	• HiSeq 3000/4000 Cluster Kit PE-410-1001 • HiSeq 3000/4000 SR Cluster Kit GD-410-1001	~ 2 days
NovaSeq Sereis	N/A	1	Dual	Up to 3.3B	33	33 (single FC) 66 (dual FC)	• NovaSeq 5000/6000 S2 Reagent Kit (200 cycles) 20012861 • NovaSeq 5000/6000 S2 Reagent Kit (100 cycles) 20012862	≤ 1 day

* Each kit provides reagents for 50 cycles. With a total of 150 cycles (at 2x75) needed for the run, 3 kits (ie, x3) are needed per run.

Profiling

TruSeq Targeted RNA

Sequencer	Mode	Lanes per FC	FC capacity	Output/lane (clusters)	Samples/ lane	Samples/run	Reagent kit	Time
MiniSeq	High-output	1	1	Up to 25M	Up to 384	Up to 384	HO kit 150 cycles FC-404-1002	~5 hrs
MiSeq	N/A	1	1	Up to 15M reads (v2 chemistry)			MiSeq reagent kit v2 (50 cycles) MS-102-2001	~5 hrs

TruSight Pan-Cancer Panel

Sequencer	Mode	Lanes per FC	FC capacity	Output/lane (clusters)	Samples/ lane	Samples/run	Reagent kit	Time
MiniSeq	High-output	1	1	Up to 25M	8	8	HO kit 150 cycles FC-404-1002	13 hrs
MiSeq	N/A	1	1	Up to 25M	5	5	MiSeq reagent kit v3 (150 cycles) MS-102-3001	21 hrs

TruSeq RNA Access Custom

Sequencer	Mode	Lanes per FC	FC capacity	Output/lane (clusters)	Samples/ lane	Samples/run	Reagent kit	Time
MiniSeq	High-output	1	1	Up to 25M	Depends on complexity		HO kit 150 cycles FC-404-1002	13 hrs
MiSeq	N/A	1	1	Up to 25M			MiSeq reagent kit v3 (150 cycles) MS-102-3001	21 hrs
NextSeq	Mid-output	1	1	Up to 130M reads			MO kit 150 cycles FC-404-1001	15 hrs

Considerations when transitioning from qPCR or GEX arrays

This section is geared to highlight some of the similarities and differences between qPCR, gene expression (GEX) arrays, and NGS, from the standpoint of how a transition will affect your day-to-day operations. If you have additional questions or would like to discuss any of the issues highlighted below, contact your regional Field Application Scientist.

Overview Comparison – Platforms for RNA-based Drug Response Biomarker Discovery and Profiling

	qPCR	GEX arrays	RNA sequencing (profiling)	RNA sequencing (discovery)
Total assay time	3–5 hrs	3–4 days	1 day	1.0–3.5 days
Target complexity	Focused	Transcriptome-scale or focused	Focused	Transcriptome scale
Data returned	Ct score (cycles to reach fluorescence threshold)	Fluorescence intensity	Digitally quantified sequence data	Digitally quantified sequence data
Dynamic range	+++++	+++	+++++	+++++
Cost/sample	$ – $$	$$ – $$$	$ – $$	$$ – $$$
Capturable biomarker categories	• Gene-level abundance • Transcript-level abundance (known transcripts included in array design) • Gene fusions (known)	• Gene-level abundance • Transcript-level abundance (known transcripts selected in assay design)	• Gene-level abundance • Transcript-level abundance (known + novel*) • SNVs (known and novel*) • Gene fusions • (known and novel*)	• Gene-level abundance • Transcript-level abundance (known and novel)* • SNVs (known and novel)* • Gene fusions (known and novel)*

* Focused regions

Feature discovery

How will novel candidates be filtered and prioritized?

An important aspect of transitioning from qPCR and RNA-Seq is the fact that the scope of data that will be produced and require analysis and interpretation is not defined at the beginning of the experiment. While the ability to detect novel features of the transcriptome is a considerable advantage in the context of biomarker discovery, harnessing and managing this capability requires workstream adjustments. For example, it will be necessary to define the level of evidence required to support a novel fusion or alternative transcript isoform that appears to have predictive value. Even if a novel transcript isoform is confirmed as “real,” will it make the cut as a biomarker candidate if the function is poorly understood? These and similar questions must be addressed.

How can I integrate RNA-Seq data with existing qPCR or GEX array datasets?

In many cases, transitioning to an RNA-Seq-based method for drug response biomarker discovery will require having the ability to integrate RNA-Seq data with existing datasets generated using qPCR, GEX arrays or other platforms. BaseSpace Correlation Engine includes a workstream that incorporates integration into the core workflow.

Data storage

What files will my NGS workflow produce?

From the point at which raw data is generated by the sequencer through biomarker candidate prioritization, data is progressively processed into multiple file types. The table below describes each of these files types and gives examples of how they are used.

File type	Description	When it is produced	How it is used
BCL (Base Call File)	File generated by Illumina DNA sequencing instruments; saves base call and quality score information for each cycle processed	During Illumina sequencing, real time analysis output from a sequencing run is a set of quality-scored base call files (*.bcl) generated from the raw image files	Converted to FASTQ files
FASTQ	Text-based file that stores nucleotide sequence and corresponding quality scores (Q scores)	After Illumina sequencing, either at command line (blc2fastq), on BaseSpace, or via on-instrument software	Focused
SAM (Sequence Alignment/ Map)	SAM (file format) is a text-based format for storing biological sequences aligned to a reference sequence. The acronym SAM stands for Sequence Alignment/ Map. It is widely used for storing data, such as nucleotide sequences, generated by NGS technologies	During alignment	Store alignment data
BAM (Binary Alignment/Map)	Compressed binary files containing sequence alignment data (see page 27 of MiSeq Reporter software guide, link above to alignment data)	During alignment	Store alignment data. BAM files are the primary input for the variant-calling step.
VCF (Variant Call Format)	Text files containing SNVs, indels, and other structural variants	During variant-calling, after alignment	To identify variants (in comparison to a reference)

What data do you want to keep?

Data storage requirements are influenced by a number of variables, including what data you intend to store. When considering what files will need to be kept versus discarded, consider these factors:

- What file types are “redundant,” in that one file can be re-created from another?
- How can each file type be used?
- What is the iterative value of keeping each file type vs the impact on storage requirements?
- Are there company-specific requirements that dictate what files must be retained?

The choice is typically whether to keep .bcl files, FASTQ files, or BAM files, or some combination thereof. These file types overlap considerably in the data they contain, so often it is more efficient to keep only 1–2 types.

Sequencing data is initially stored by the sequencer as .bcl files. This is a proprietary file type that is translated by bcl2fastq to FASTQ formatted files. The conversion is one-way, and base-calling is already complete by this step. BCL files contain essentially the same information as their equivalent FASTQ files.

After alignment, sequencing data is typically stored in BAM files. A BAM file contains both the sequencing information resident in the FASTQ file and alignment information. It is possible to convert BAM files to their FASTQ equivalents using commonly available tools, such as Picard and Bamtools, so the BAM file can also substitute for the FASTQ file.

Data security

This is a critical consideration for many users, particularly within the pharmaceutical industry. Historically, onsite data storage and analysis have been an industry requirement, and for many users this may still be the case. While this is still the case for some companies, multiple Top 5 pharmaceutical companies have engaged with us on their security concerns, and based on their findings, have implemented cloud-based data storage and analysis. As a result of these analyses, a number of users have revisited their policies and have integrated cloud-based solutions.

Following, we have highlighted common questions we have received regarding cloud-based data security and have addressed in pharmaceutical industry audits.

How is my data secured?

BaseSpace Informatics Sequence Hub utilizes Amazon Web Services (AWS) for its cloud-based services. AWS data security protocols, including data encryption during transfer and at rest, require Application Program Interface (API) key signatures, and meet many other security standards (see Data Storage table above). BaseSpace Informatics Sequence Hub Enterprise adds an additional security layer that includes a private domain, single sign-on, and access permissions.

Cloud security highlights – BaseSpace Informatics Suite

Transfer of data to BaseSpace is encrypted using the AES256 standard
All traffic is over Secure Sockets Layer (SSL)
All service methods require API key signatures
Requests are monitored for abuse and IPs and/or API keys can be blacklisted
AWS security standards for data at rest include: • SOC 1/SSAE 16/ISAE 3402 • FISMA moderate • PCI DSS Level 1 • ISO 27001 • FIPS 140-2
AWS data integrity standards include: • Synchronous storage across multiple facilities • Regular data integrity checks • Automatic self-healing

Who controls my data?

You will always have control over your own data. In BaseSpace Informatics Sequence Hub, you control who has access to your data at all times, and you can grant access to specific users or workgroups as needed.

What if a user leaves my organization?

BaseSpace Informatics Sequence Hub Enterprise allows an administrator in your organization to add and remove users. You control which users have access to the platform at any time.

References

1. Zhao S, Fung-Leung W-P, Bittner A, Ngo K, Liu X. Comparison of RNA-seq and microarray in transcriptome profiling of activated T cells. *PLoS ONE*. 2014;9(1):e78644. doi:10.1371/journal.pone.0078644.
2. Wang ZL, Zhang CB, Cai JQ, Li QB, Wang Z, Jiang T. Integrated analysis of genome-wide DNA methylation, gene expression and protein expression profiles in molecular subtypes of WHO II-IV gliomas. *J Exp Clin Cancer Res*. 2015;34:127. doi: 10.1186/s13046-015-0249-z.
3. Atak ZK, Gianfelici V, Hulselmans G, et al. Comprehensive analysis of transcriptome variation uncovers known and novel driver events in T-cell acute lymphoblastic leukemia. *PLoS Genet*. 2013;9(12):e1003997.
4. Kumar-Sinha C, Kalyana-Sundaram S, Chinnaiyan AM. Landscape of gene fusions in epithelial cancers: seq and ye shall find. *Genome Med*. 2015;7:129.
5. Ishikawa R, Amano Y, Kawakami M, et al. The chimeric transcript RUNX1–GLRX5: a biomarker for good postoperative prognosis in Stage IA non-small-cell lung cancer. *Jpn J Clin Oncol*. 2016;46(2):185-189.
6. Lu L, Zhang H, Pang J, Hou G, Lu M, Gao X. ERG rearrangement as a novel marker for predicting the extra-prostatic extension of clinically localised prostate cancer. *Oncol Lett*. 2016;11(4):2532-2538.
7. Perez-Gracia JL, Sanmamed MF, Bosch A, et al. Strategies to design clinical studies to identify predictive biomarkers in cancer research. *Cancer Treat Rev*. 2017;53:79-97.
8. Kantae V, Krekels EHJ, Esdonk MJV, et al. Integration of pharmacometabolomics with pharmacokinetics and pharmacodynamics: towards personalized drug therapy. *Metabolomics*. 2017;13(1):9.
9. Fang B, Mehran RJ, Heymach JV, Swisher SG. Predictive biomarkers in precision medicine and drug development against lung cancer. *Chin J Cancer*. 2015;34(7):295-309.
10. Zhao X, Modur V, Carayannopoulos LN, Laterza OF. Biomarkers in Pharmaceutical Research. *Clin Chem*. 2015;61(11):1343-1353.
11. Mishra PJ. Non-coding RNAs as clinical biomarkers for cancer diagnosis and prognosis. *Expert Rev Mol Diagn*. 2014;14(8):917-919.
12. Costa C, Giménez-Capitán A, Karachaliou N, Rosell R. Comprehensive molecular screening: from the RT-PCR to the RNA-seq. *Transl Lung Cancer Res*. 2013;2(2):87-91.
13. Perkins JR, Antunes-Martins A, Calvo M, et al. A comparison of RNA-seq and exon arrays for whole genome transcription profiling of the L5 spinal nerve transection model of neuropathic pain in the rat. *Molecular Pain*. 2014;10:7.
14. Brewer CT, Chen T. PXR variants: the impact on drug metabolism and therapeutic responses. *Acta Pharmaceutica Sinica B*. 2016.
15. Bracco L, Kearsey J. The relevance of alternative RNA splicing to pharmacogenomics. *Trends Biotechnol*. 2003;21(8):346-353.
16. Barrie ES, Smith RM, Sanford JC, Sadee W. mRNA Transcript diversity creates new opportunities for pharmacological intervention. *Mol Pharmacol*. 2012;81(5):620-630.
17. Ling H, Fabbri M, Calin GA. MicroRNAs and other non-coding RNAs as targets for anticancer drug development. *Nat Rev Drug Discov*. 2013;12(11):847-865.
18. Rönnau CG, Verhaegh GW, Luna-Velez MV, Schalken JA. Noncoding RNAs as novel biomarkers in prostate cancer. *Biomed Res Int.* 2014;2014:591703.
19. Moorman AV. New and emerging prognostic and predictive genetic biomarkers in B-cell precursor acute lymphoblastic leukemia. *Haematologica*. 2016;101:407-416.
20. Nalejska E, Mączyńska E, Lewandowska MA. Prognostic and predictive biomarkers: tools in personalized oncology. *Mol Diagn Ther*. 2014;18(3):273-284.
21. Shang C, Guo Y, Zhang H, Xue YX. Long noncoding RNA HOTAIR is a prognostic biomarker and inhibits chemosensitivity to doxorubicin in bladder transitional cell carcinoma. *Cancer Chemother Pharmacol*. 2016;77(3):507-513.
22. McCleland ML, Mesh K, Lorenzana E, et al. CCAT1 is an enhancer-templated RNA that predicts BET sensitivity in colorectal cancer. *J Clin Invest*. 2016;126(2):639-652. doi:10.1172/JCI83265.
23. Zhou M, Ye Z, Gu Y, Tian B, Wu B, Li J. Genomic analysis of drug resistant pancreatic cancer cell line by combining long non-coding RNA and mRNA expression profiling. *Int J Clin Exp Pathol*. 2015;8(1):38-52.
24. Zhao W, He X, Hoadley KA, Parker JS, Hayes DN, Perou CM. Comparison of RNA-seq by poly (A) capture, ribosomal RNA depletion, and DNA microarray for expression profiling. *BMC Genomics*. 2014;15:419.
25. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. *Nat Biotechnol*. 2014:32(9):903-914.

Illumina • 1.800.809.4566 toll-free (US) • +1.858.202.4566 tel • techsupport@illumina.com • www.illumina.com

For Research Use Only. Not for use in diagnostic procedures.

illumina®