SevenBridges

Scalable bioinformatics for discovery with RNA-seq

SCALABLE BIOINFORMATICS FOR DISCOVERY WITH RNA-SEQ

Advances in sequencing technologies are enabling researchers to identify RNA features that were undetectable just a few years ago. RNA-seq is a developing technology that requires a well-designed experimental strategy and a data analysis approach that scales. The Seven Bridges Platform enables rapid and computationally robust analysis of RNA-seq data, empowering researchers to pursue basic and clinical investigations related to the transcriptome.

IN THIS PAPER:

- RNA-seq's unique advantages compared to other RNA analysis technologies
- State-of-the-art approaches for quality control, sequence alignment, quantification, and visualization
- Challenges presented by RNA-seq analysis at scale, and how to overcome them
- How the Seven Bridges Platform supports RNA-seq analysis, by enabling scalable, reproducible and collaborative data analysis
- Specific RNA-seq use cases, including approaches for gene expression, SNP discovery, fusion identification, miRNA profiling, splice isoform identification, and single-cell transcriptomics
- Clinical applications of RNA-seq
- The future of RNA-seq, and the design of an optimum analysis system

TABLE OF CONTENTS

RNA-SEQ: THE NEW STANDARD FOR DISCOVERY

RNA-sequencing (RNA-seq) is quickly superseding older technologies for RNA analysis such as microarrays and quantitative PCR. By removing the requirement for previous knowledge of the sequences being analyzed, and increasing throughput by several orders of magnitude, RNA-seq is rapidly emerging as the most powerful tool for identifying novel SNPs, fusion genes, alternative splice isoforms and disease-associated gene expression profiles (**Box 1**).

In comparison to DNA, RNA is highly dynamic. The expressed sequences and their relative abundance can vary widely. The transcriptome is continually adjusted by numerous factors that respond to extracellular input and internal biochemistry. The study of the transcriptome can therefore provide unique insights. Moreover, RNA-based molecular diagnostics have the potential to apply broadly to diagnosis, prognosis, and therapeutic selection for numerous human diseases.

Because of RNA's unique dynamic role as a cellular messenger and regulator, RNA-seq has the potential to vastly expand the repertoire of disease-associated variants. In the hands of researchers, it is a uniquely powerful discovery tool.

BOX 1: ADVANTAGES OF RNA-SEQ

1. **REDUCED BIAS**

   RNA-seq is fundamentally different than its predecessors in that it is an open technology, meaning that it outputs sequence regardless of whether the researcher specifies the sequence in advance. A sequencer does not distinguish between well-characterized sequences and completely novel ones. In this way, RNA-seq dispenses with the need to design oligonucleotide probes of known sequence for hybridization with the experimental sample. RNA-seq can reveal unknown sequences present in the sample.

2. **HIGH THROUGHPUT**

   RNA-seq throughput is limited only by the available sequencing technology. Because sequencer throughput continues to increase as new methods and products become available, the amount of RNA-seq data that can be generated in a day is very large indeed. Sequencing throughput is not likely to be a bottleneck for groups who plan to use RNA-seq frequently. Rather, RNA extraction and library prep are more likely to be rate-limiting steps. Robotic automation of these steps is therefore used by the most active groups.

3. **DYNAMIC RANGE**

   RNA-seq's large dynamic range allows for better characterization of RNA abundance in samples where the expression difference between rare and commonly expressed transcripts may span several orders of magnitude.

4. **INFORMATION CONTENT**

   RNA-seq also allows for identification of functional variation at the posttranscriptional level, including splicing. RNA-seq can also be used to identify transcription start and stop sites. RNA-seq is therefore especially well suited to human genomics, since significant functional diversity is generated by splicing variation.

RNA-SEQ ANALYSIS BEST PRACTICES

No single best workflow exists for RNA-seq data analysis. Because RNA-seq experiments vary in their scope and content in accordance with experimental goals, optimal analysis methods can differ. In this paper we focus on delivering large-scale computational analysis of RNA-seq data; however other factors including library type, sequencer type, sequencing depth, and replicate number must all be carefully assessed in the context of research goals.

Generally, the stages of an RNA-seq experiment include experimental design, quality control, read alignment, quantification, and visualization. Additional analysis steps include differential expression, alternative splicing, functional analysis, gene fusion detection, or eQTL mapping (**Figure 1**).

Figure 1. Major analysis steps for RNA-seq. Pre-analysis for RNA-seq experiments includes experimental design, sequencing design, and quality control. Core analysis may include transcriptome profiling, differential expression, and functional profiling. Advanced analysis strategies vary depending on experimental goals. Adapted from Conesa et al. (2016) A survey of best practices for RNA-seq data analysis. Genome Biol. **17**: 13.

Quality control enters into analysis workflows at multiple points. Quality control for raw reads should include checks for sequence quality, GC content, the presence of adaptors, over-represented k-mers, and duplicated reads. Performing these checks early is critical for detecting contamination, PCR artifacts, and sequencing errors. We recommend **FastQC** for removing low-quality reads, adaptors, and ambiguous bases.

Quality control is also critical to ensure accurate alignment. The decision to map reads against a genome reference, transcriptome reference, or both encompasses both reference availability and experimental goals (**Figure 2**). Genome-guided methods for transcript assembly such as **Cufflinks** use a reference genome as a template to align and assemble reads, while de novo methods such as **Trinity** assemble transcripts directly from reads. We recommend checking the percentage of mapped reads as a measure of overall sequencing accuracy and as an indicator of foreign DNA contamination. It is also important to assess read coverage, both over exons and across the mapped strand. Excessive read accumulation at the 3′ end of transcripts is indicative of RNA degradation. GC content is also important for identifying PCR bias. **Picard**, **RSeQC**, and **Qualimap** are standard tools for alignment quality control.

Transcript quantification is the most common application for RNA-seq. Researchers might choose to focus on transcript quantification exclusively or take an approach with increased sensitivity towards the identification of novel transcripts. Most approaches count the number of reads that map to a given transcript, normalizing for length, total read number, and sequencing bias. Tools for transcript quantification include **Cufflinks, RSEM, eXpress, Sailfish**, and **kallisto**.

Visualization tools help promote rapid analysis of RNA-seq data by aggregating and indexing data while maintaining appropriate relationships to represent connections between data elements. Read pileup within a genome browser is perhaps the most commonly used visualization approach. Other available visualization tools include **CummeRbund, RSEM BAM2WIG, RSEM Plot Model, RSEM Plot Transcript Wiggles, AmpliconQC**, and **Report Renderer 2**.

Generating a dependable RNA-seq analysis strategy is non-trivial. The RNA-seq field has an overabundance of software tools and analysis options, with minimal agreement on best practice. Furthermore, laboratory-specific wrapper scripts and chained-together workflows are commonplace. While homemade solutions may be sufficient for small-scale analysis, these inefficiencies can have major impacts when experiments grow to include thousands of samples, and when it comes to submitting methods and results for regulatory approval.

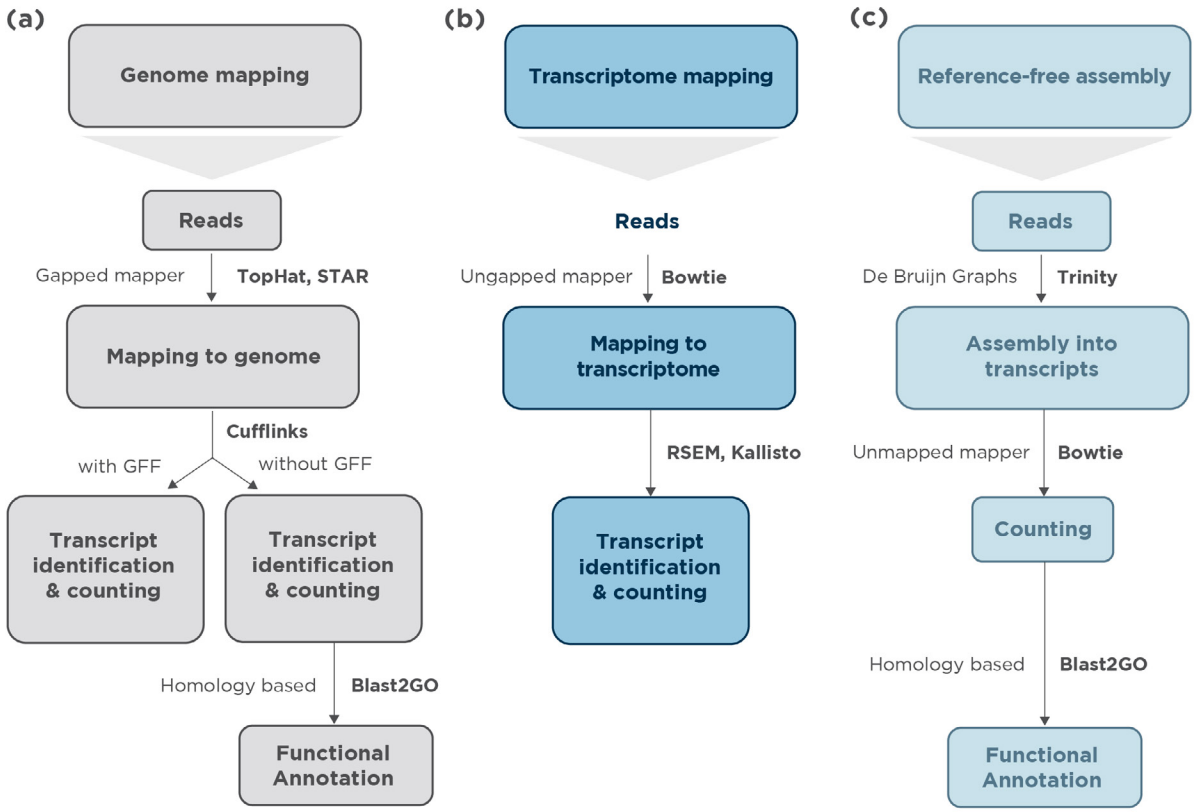

Figure 2. Strategies for read mapping and transcript identification. a.) For annotated genomes, reads are first mapped to the reference using a gapped aligner such as TopHat or STAR, followed by transcript discovery, quantitation, and annotation. b.) Experiments that do not require transcript discovery can map reads to a reference transcriptome using an ungapped aligner. c.) When a reference genome is not available, reads are assembled into contigs. The quantitation step then aligns reads back to these reference contigs. Adapted from Conesa et al. (2016) A survey of best practices for RNA-seq data analysis. Genome Biol. **17**:13.

THE SEVEN BRIDGES PLATFORM ENABLES RNA-SEQ ANALYSIS AT SCALE

The Seven Bridges Platform (sevenbridges.com/platform) is the cloud-based environment for conducting robust RNA-seq analysis at scale (**Figure 3**). It functions as a central hub for teams to store, analyze, and jointly interpret their experimental data. The Platform allocates storage and compute resources on demand, and optimizes processing by co-locating analysis workflows alongside genomic datasets.

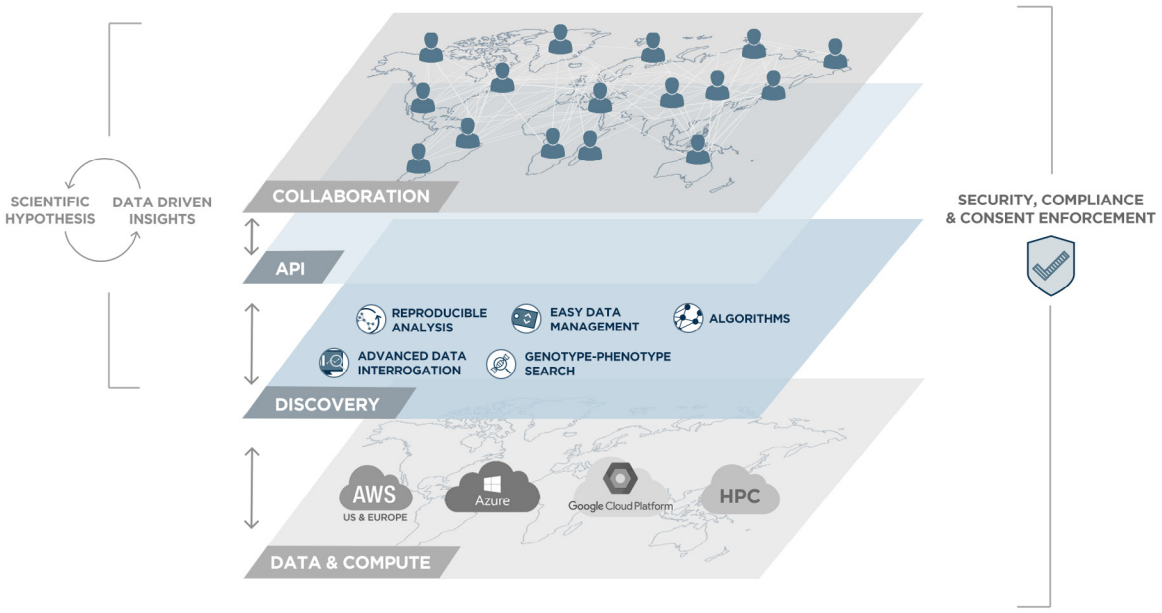

Figure 3. Overview of the Seven Bridges Platform. A Cloud-based data and compute infrastructure underlies the discovery layer, which is built around features to streamline data management, search, and analysis. An application programming interface and collaboration features ensure flexibility for users. Data security and regulatory compliance controls operate at all levels.

The Platform offers researchers an effective way of deploying RNA-seq to investigate the biology of human variation and disease, and to develop effective treatments.

- **Immediately analyze RNA-seq data at scale.** The Platform is designed to enable bioinformatic analysis at industrial scale. As RNA-seq experiments increase to thousands of samples, and hundreds of cell types, they need large-scale storage and compute. By working in the cloud, researchers can scale computation as needed to analyze their experimental data.

- **Find curated RNA-seq tools easily.** The Platform hosts many of the most popular RNA-seq tools, which researchers can immediately apply to their projects (**Box 2**). Our team of bioinformaticians maintains these tools and optimizes their performance for the cloud to ensure efficient analysis.

- **Use and modify preloaded workflows.** The Platform also hosts curated best practice RNA-seq workflows, enabling researchers—even those with minimal bioinformatic experience—to go from raw sequencing reads, all the way through visualization of differentially expressed genes and transcripts. Each workflow is fully customizable, via an intuitive drag-and-drop interface, or a powerful API.

- **Bring your own tools and workflows.** RNA-seq is an evolving technology, and no single best workflow exists for its data analysis. As state-of-the-art research often requires novel methods, the Platform enables researchers to deploy their own tools and custom workflows, described using the open-source Common Workflow Language (CWL).

- **Reproduce analyses, perfectly.** Whether done for the initial stages of discovery, or for clinical development, researchers need their RNA-seq analysis to be completely reproducible. The Platform enables reproducible research as default: all files, tool versions and parameters are automatically tracked and available for review at a click. By sharing CWL descriptions of methods, collaborators, regulatory bodies and journal editors have the ability to precisely reproduce analyses.

- **Collaborate easily across an organization.** A typical RNA-seq experiment involves multiple researchers spread across scientific disciplines. The Platform enables multiple researchers, analysts and clinicians, spread around the globe, to work side by side on a single RNA-seq project. They can comment on ongoing projects, and compare results in a single integrated environment.

- **Get the most from experimental data, alongside the world's data.** As RNA-seq techniques develop, experiments promise to generate petabytes of raw data. These data are most powerful when brought together in a single analysis environment. The Platform lets organizations store data centrally with upload straight from the sequencer, and enables researchers to design and execute experiments in based on all this information, including from hosted massive public datasets like The Cancer Genome Atlas and Cancer Cell Line Encyclopedia.

- **A Platform built by scientists, supported by scientists.** RNA-seq is an evolving technology, and there is no right way to analyze its data. The Seven Bridges Platform is designed to provide the flexibility that researchers need to pursue their discovery goals, instead of limiting users to preloaded pipelines. Our team of genome biologists, mathematicians, developers and bioinformaticians is available to support users, and to work alongside research teams to help them achieve their research goals.

BOX 2: RNA-SEQ ANALYSIS TOOLS AVAILABLE ON THE SEVEN BRIDGES PLATFORM

QUALITY CONTROL:
RSeQC; RNA-SeQC; FastQC; MultiQC

ALIGNERS:
STAR; TopHat; HISAT2; Bowtie/Bowtie 2

QUANTIFICATION:
Cuffquant (from Cufflinks package); HTSeq; RSEM; eXpress; Salmon

DIFFERENTIAL EXPRESSION ANALYSIS:
Cuffdiff (from Cufflinks package); DESeq2; DEXSeq

TRANSCRIPT ASSEMBLY:
Trinity; Cufflinks

FUSION GENES:
ChimeraScan; STAR-Fusion; deFuse; Chimera; Oncofuse

VISUALIZATION:
CummeRbund; Circos

UTILITY TOOLKITS:
Picard; SAMtools; BAMtools; SRA Toolkit; Eautils; Fastq-mcf; Flexbar; FQtrim

SINGLE-CELL:
Drop-seq tools; Seurat

RNA-SEQ RESEARCH APPLICATIONS

GENE EXPRESSION ANALYSIS

Gene expression is arguably the central factor influencing the generation of functional variety in biological systems. Differential expression analysis, which compares expression between multiple samples, allows researchers to investigate how transcript expression patterns affect phenotype. Whereas early approaches to compute differential expression used Poisson and negative binomial distributions, current tools perform integrated normalization and data transformation or take nonparametric approaches.

The ability to assess gene expression at many loci simultaneously has numerous clinical applications. Multigene expression assays such as OncotypeDx and MammaPrint are recommended for guiding treatment decisions in patients with breast cancer.[1,2] RNA-seq can provide novel insights on allelic imbalance and copy number variation when used to compare tumor and normal patient samples, and has the potential to outperform microarrays in analyzing cancer transcriptomes and predicting clinical endpoints.[3,4] Gene expression profiling is also applicable for predicting immune rejection in cardiac allograft recipients.[5]

Because the identification of differentially expressed genes and transcripts is one of the most common applications of RNA-seq, the Seven Bridges Platform offers optimized tools for the task. It also supports other common differential expression-related work, including detection of alterations in transcription start sites and alternate coding regions and identification of differential splicing between conditions. Differential expression workflows on the Seven Bridges Platform use **Cuffdiff** (**Figure 4**), **DESeq2**, and **DEXSeq** to identify significant differences between samples from read count distribution, while accounting for biological and technical variability.

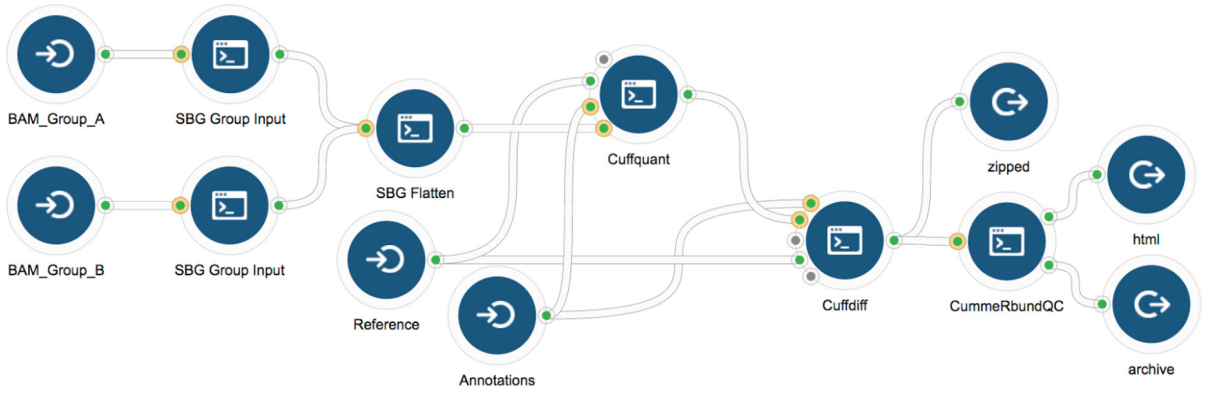

Figure 4. Cuffdiff differential expression workflow on the Seven Bridges Platform. This public workflow uses Cuffdiff to find changes in gene expression, splicing, and promoter usage. While the researcher can supply aligned reads as BAM or SAM files directly to Cuffdiff, using the Cuffquant utility allows them to quantify expression levels of all samples in parallel, and can significantly reduce the total running time of a differential expression analysis. Following quantitation, Cuffdiff performs differential expression tests between groups of samples. This workflow also uses CummeRbund for quality control.

SNP DISCOVERY

The most abundant form of genetic variation is single nucleotide polymorphisms (SNPs). Many SNPs, both heritable and somatic, are associated with disease, drug efficacy, and other clinically relevant factors. Although researchers often use whole exome sequencing for identifying variation at the single nucleotide level, RNA-seq, which is generally less expensive than whole exome sequencing and includes information on transcription levels, is also an excellent approach for SNP discovery. The Platform offers several ways of exploring variant calls from RNA-seq data. Users can modify workflows in a drag-and-drop graphical interface, allowing them to easily add SNP calling to other analysis workflows.

We recommend that whole exome sequencing and RNA-seq should be considered complementary methods. One study that compared SNP calls from whole exome sequencing and RNA-seq data found that RNA-seq identified SNPs outside of whole exome sequencing capture boundary regions and low coverage regions. Additionally, 55% of unique RNA-seq identified SNP calls bore mutational signatures of RNA editing.[6] Integration of both types of data is especially informative in heterogeneous samples such as tumor cells. In the context of results from ENCODE[6] and other studies that demonstrate pervasive transcription of many so-called 'noncoding' regions and the introduction of functionally relevant variation at the posttranscriptional level, RNA-seq should be considered a valuable tool for SNP discovery.negative binomial distributions, current tools perform integrated normalization and data transformation or take nonparametric approaches.

GENE FUSION DETECTION

Technical advances in next-generation sequencing, including better fusion detection algorithms, have allowed an expanded appreciation of the role of gene fusions as drivers of the altered regulatory signalling observed in many cancers.[7] Although putative fusions that arise from chromosomal rearrangement can be identified from DNA sequence data, the ability to compare expression levels makes RNA-seq the method of choice for novel fusion detection. Fusion expression levels have been shown to correlate with grade, stage, and prognosis in prostate tumors.[8] As of February 2017, 10,652 validated fusions are registered with the Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer.[9]

Computational algorithms that allow alignment across exon-exon junctions are central to bioinformatic approaches to gene fusion detection. The Seven Bridges Platform contains several curated tools for fusion genes, including **ChimeraScan** (**Figure 5a)**, **STAR-Fusion**, **deFuse**, **Chimera** and **Oncofuse**. **Oncofuse** estimates the tumor driver potential of a fusion set, while **ChimeraScan** is used for further study of detected fusions. Fusion gene discovery is comparable to novel isoform discovery in that reads contain sequence from regions that are nonadjacent on the reference genome. It is complicated by the greatly increased search space, as interchromosomal fusions are common (**Figure 5b**). Several studies have nevertheless validated RNA-seq for fusion discovery.[10–13]

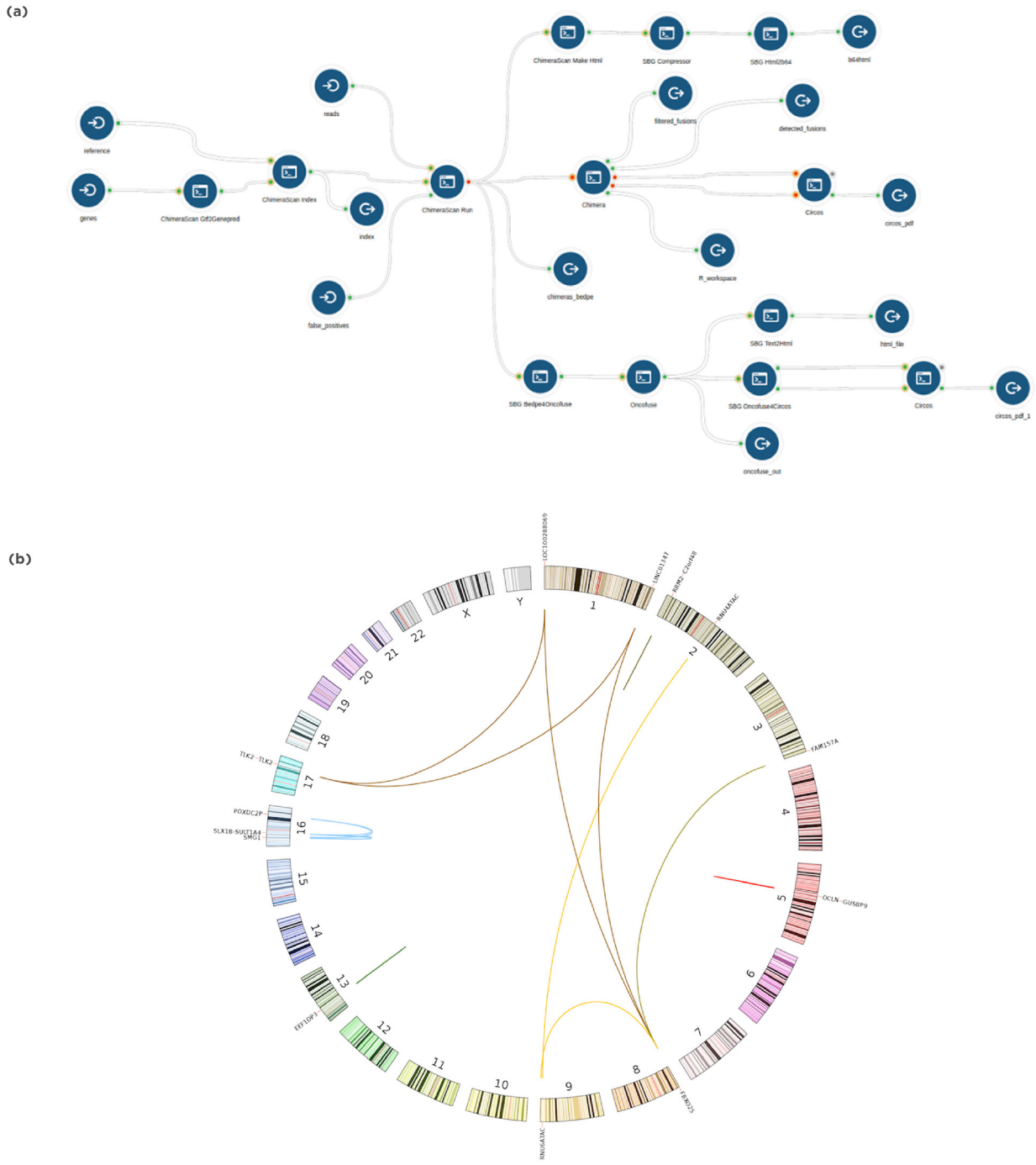

Figure 5. Fusion transcript detection with ChimeraScan workflow on the Seven Bridges Platform. (a) ChimeraScan is a software package that detects gene fusions in paired-end RNA seq data. It aligns paired-end reads to a combined genome-transcriptome reference. Read pairs that could not be aligned concordantly are trimmed into smaller segments and realigned. Trimming improves sensitivity by increasing the chance that neither read alignment spans a chimeric junction. Users can drag and drop, scatter, and adjust parameters easily from a graphical interface. All changes are automatically recorded. **Chimera** and **Oncofuse** are used for post-processing of candidate fusions. **(b)** Fusions in an ovarian cancer sample from the Cancer Genome Atlas visualized at whole-genome scale using **Circos**. Reference chromosome sequences are represented on the outer circle. Curved lines indicate intrachromosomal and interchromosomal rearrangement.

miRNAs AND OTHER sRNAs

MicroRNAs are short (18-25bp) noncoding RNAs that are evolutionarily conserved and function in gene regulation. MicroRNAs typically bind Argonaute proteins and target complementary mRNAs for downregulation. Because these short RNAs can modulate the expression of tens to hundreds of gene products, changes to miRNA-mediated gene regulation can significantly alter cellular functions. In cancer, miRNAs are frequently disregulated.[14] Furthermore, miRNA expression is closely correlated with tumor stage, and can be used to classify tumors that might otherwise be challenging to classify by histological or mRNA-based methods. A microarray-based method for detecting changes in cancer-associated miRNAs underlies Rosetta Genomics' Cancer Origin diagnostic assay.[15] With its unbiased approach and high throughput, RNA-seq is uniquely suited to identifying and quantifying microRNAs.

IDENTIFYING ALTERNATIVE SPLICE ISOFORMS

A great deal of functional diversity in humans emerges from alternative splicing, which allows a single gene to generate multiple protein products (**Figure 6**). For researchers wanting to explore the functional aspects of individual isoforms, computational approaches that integrate RNA-seq data are recommended.[16,17] Because developmental abnormalities and cancers frequently have alterations in splicing, RNA-seq data is valuable for identifying disease-associated splicing patterns. Importantly, alternative splicing has relevance throughout human disease. One study that used RNA-seq to study the brains of patients with Alzheimer's disease found that changes in splicing and promoter usage at the *APOE* gene correlate with the progression of neurodegeneration.[18] The Seven Bridges Platform contains multiple tools for assessing alternative splicing, including **Cuffdiff**, **DESeq2**, and **DEXSeq**.

There are two primary methods for identifying isoforms from RNA-seq data. The first is based on integrating differential expression information with isoform expression information. A second method compares read distribution over exons. These methods are most appropriate for studies concerned with the inclusion or exclusion of specific exons or protein domains. Currently, the short reads generated by many sequencers make detection of rare splice isoforms a major challenge.[19] To improve isoform detection, long-read sequencing may ultimately be the best choice.

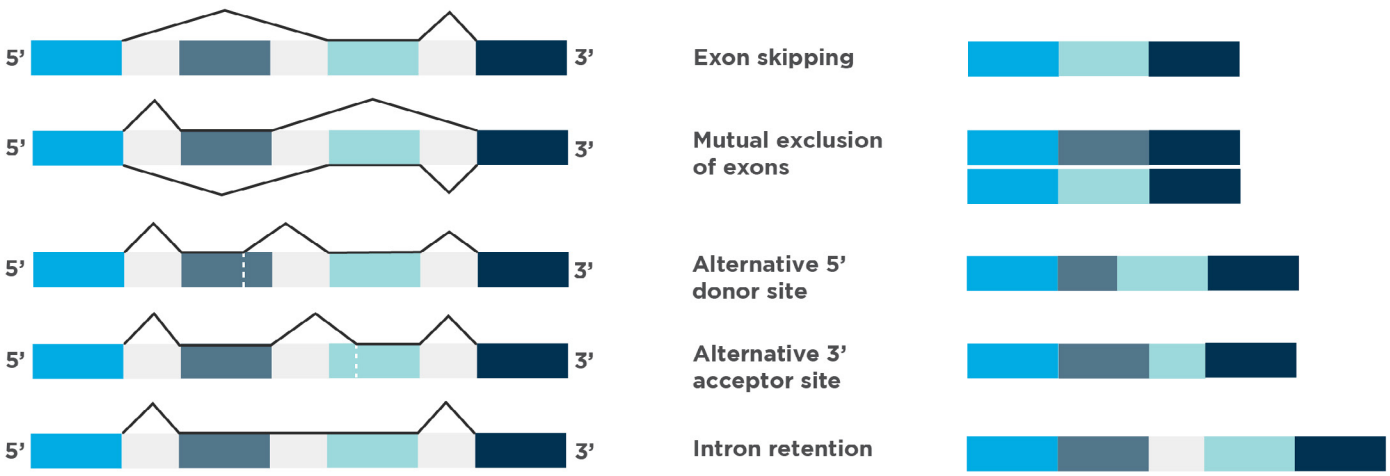

Figure 6. Alternative splicing generates a variety of processed RNAs through several mechanisms. Many human genes are notable for their highly intronic structure, which allows for variability in polypeptide sequence arising from tissue or cell type specific splicing paradigms.

SINGLE-CELL TRANSCRIPTOMICS

RNA-seq approaches are especially significant in their potential to better understand transcriptomics at the single cell level.[20] Protocols such as Smart-seq use amplification to generate an RNA library from a single cell, enabling the identification of new cell types within tissues and allowing researchers to disentangle the stochasticity of gene expression within a cell population. Other single-cell and subcellular RNA sequencing methods are in the development pipeline, with applications throughout oncology and infectious disease. Additional methods now exist to measure DNA methylation and open chromatin (ATAC-seq) at the single cell level.[21,22] For single-cell transcriptomics, the Seven Bridges Platform comes preloaded with tools for Drop-seq, a single-cell RNA-seq technology for high-throughput parallel expression profiling of individual cells, and **Seurat**, which assesses bias generated by droplets containing multiple cells, ambient RNA contamination, and conversion efficiency/transcript capture rate (**Figures 7 - 9**).

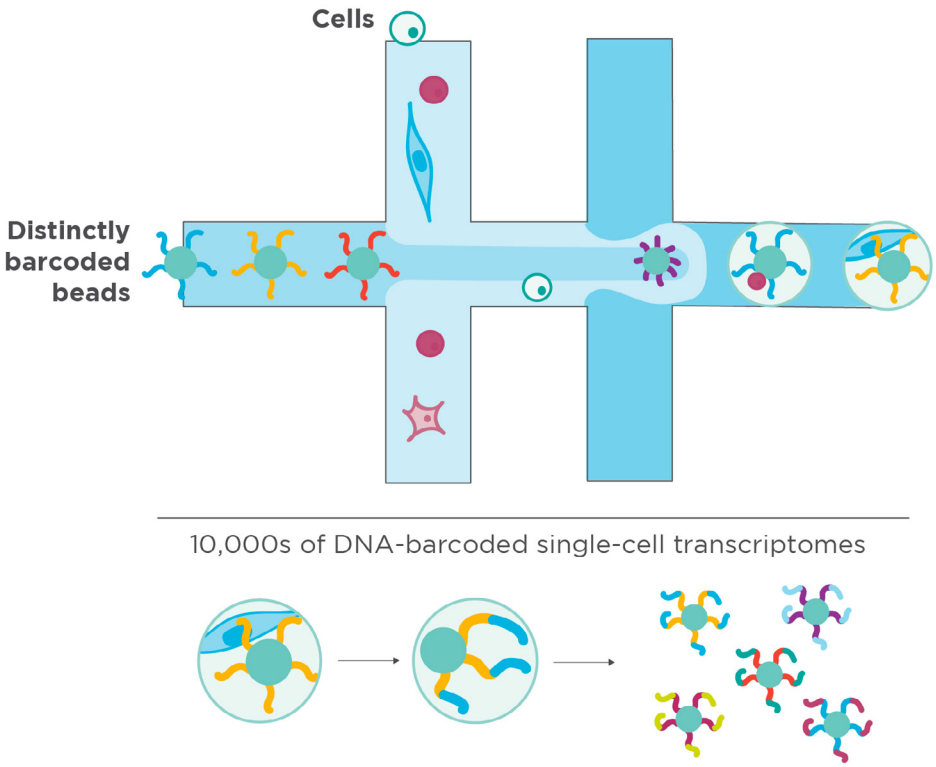

Figure 7. Drop-seq. Drop-seq generates single cell RNA-seq libraries rapidly in parallel by separating cells in a microfluidic platform along with uniquely barcoded beads.

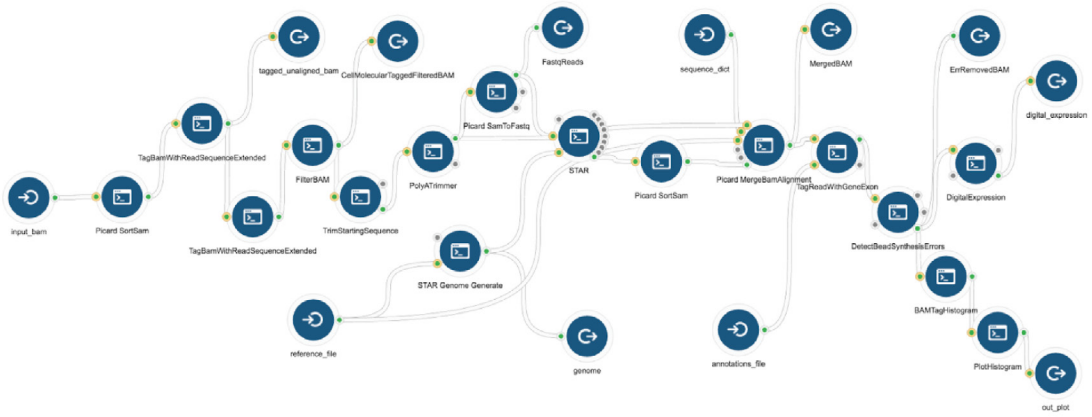

Figure 8. Single cell transcriptomics workflow on the Seven Bridges Platform. The workflow consists of a set of tools for tagging, alignment, filtering and QC of unaligned BAM files, which preserve information of origin of a read in custom tags in a BAM header. Users can modify the workflow and adjust parameters using an intuitive graphical interface. Workflow versions are automatically recorded, allowing for reproducibility in future analyses.

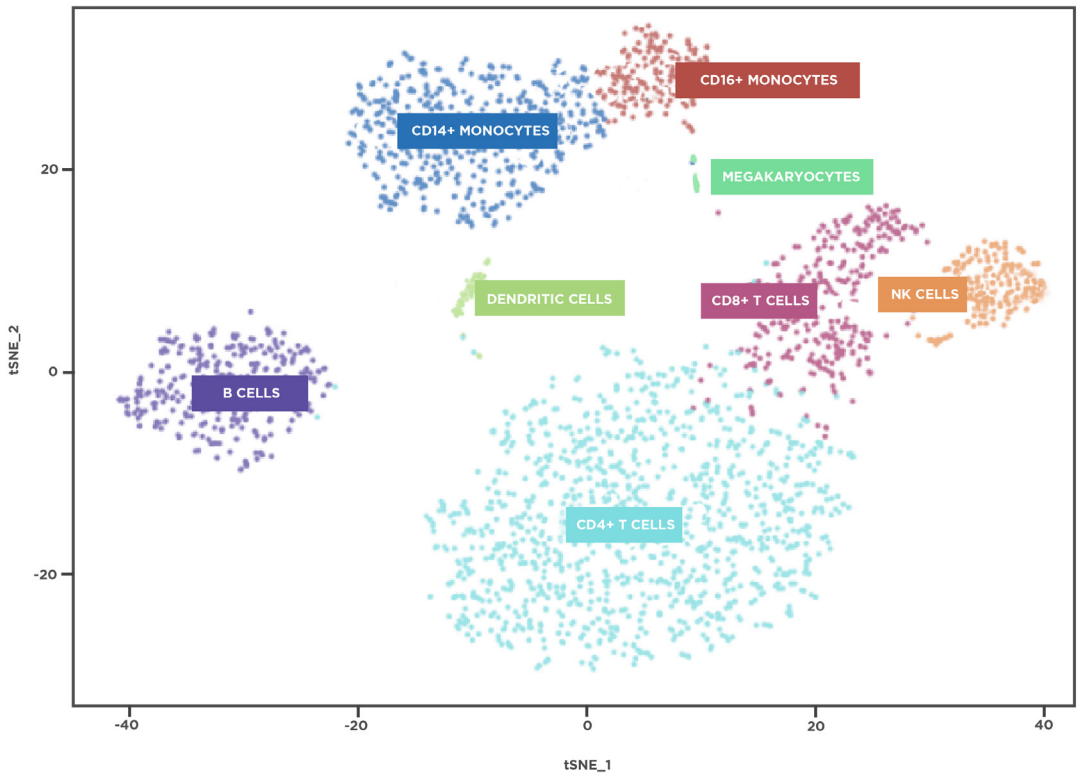

Figure 9. Seurat clustering of 68,000 peripheral blood mononuclear cells by t-distributed stochastic neighbor embedding. Seurat performs downstream analysis, extracting digital expression matrices and clustering cells based on expression. Seurat works well for many single-cell isolation techniques including inDrop, 10c Chromium, Fluidigm C1, and Drop-seq.

CLINICAL APPLICATIONS OF RNA-SEQ

The clinical utility of RNA-based measures is established for multiple human diseases. These include qRT-PCR detection of viral RNA for influenza, dengue virus, HIV, and ebola virus. Diagnostic mRNA-based tests include AlloMap,[23] Cancer Type ID,[24] and Afirma's Thyroid Nodule Assessment.[25] Several prognostic tests are currently used in the clinic for breast, prostate, and colon cancer. Tests that detect fusion transcripts in cancer include tests for RUNX1-RUNx1T1 fusion in acute myeloid leukemias, BCR-ABL fusion, and ExoDx's test for EML4-ALK fusion.[26] Although these tests use qPCR or array-based methods, applying RNA-seq to these clinical tests has the potential to improve accuracy and breadth of application. RNA-seq is already in use for fusion detection in FoundationOne's Heme assay.[27]

Systematic efforts are underway to assess the accuracy, reproducibility, and information content of RNA-seq technologies across platforms and in comparison to microarray and qPCR platforms. The FDA-directed Sequencing Quality Control project used Illumina HiSeq, Life Technologies SOLiD, and Roche 454 platforms at multiple laboratory sites to assess RNA-seq performance for junction discovery and differential expression profiling and compare it to microarray and quantitative PCR (qPCR) data using complementary metrics.[28] They found that measurements of relative expression were accurate and reproducible across multiple sites if appropriate data filters were used. This study discovered unannotated exon-exon junctions at all sequencing depths and identified gene-specific biases in all three platforms. In general, performance is dependent on the workflow used. Studies of this type point to the unique value of RNA-seq for transcript analysis and are a necessary step to design technologies that will reach the accuracy and reproducibility requirements necessary to transition RNA-seq from the lab to the clinic.

THE FUTURE OF RNA-SEQ ANALYSIS

RNA-seq is a rapidly developing technology. The upcoming years will likely see an extensive expansion of single cell techniques, as well as methods for assessing RNA post-transcriptional modification (epi-transcriptomics) and RNA tertiary structure. Additionally, numerous RNA-based diagnostic technologies have the potential to become more powerful. Challenges for the future of RNA-seq include developing analytical methods that account for technical challenges and developing clinical technologies that produce accurate therapeutic information.

For RNA-seq to fully transition to the clinic, both experimental and bioinformatic solutions will be required. From a bioinformatic standpoint, reproducibility, standardization of analysis workflows, and established standards for sequencing depth that ensure accurate variant calling under clinical conditions are table stakes. Ultimately, analysis systems for RNA-seq should enable researchers to identify functionally relevant variation at the transcriptional level, and provide a flexible platform for reproducible discovery-focused research. The Seven Bridges Platform offers organizations and researchers a complete system for gaining insight from their RNA-seq experiments to drive progress in health care.

Seven Bridges is accelerating research and development innovation at the world's leading research organizations by delivering systems that connect genomic data assets, computational infrastructure, algorithms and teams. Contact us to discuss how our Platform can be used to develop and deploy advanced RNA-seq workflows that support your discovery program, from initial research all the way to the clinic.

TEAM@SEVENBRIDGES.COM
SEVENBRIDGES.COM

SevenBridges

REFERENCES

1. ESMO. Primary Breast Cancer: ESMO Clinical Practice Guidelines | ESMO. Available at: http://www.esmo.org/Guidelines/Breast-Cancer/Primary-Breast-Cancer. (Accessed: 9th March 2017)
2. Ignatiadis M, E. al. St Gallen International Expert Consensus on the primary therapy of early breast cancer: an invaluable tool for physicians and scientists. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/26063634. (Accessed: 9th March 2017)
3. Zhang, W. et al. Comparison of RNA-seq and microarray-based models for clinical endpoint prediction. Genome Biol. 16, 133 (2015).
4. Tuch, B. B. et al. Tumor Transcriptome Sequencing Reveals Allelic Expression Imbalances Associated with Copy Number Alterations. PLoS One 5, e9317 (2010).
5. Deng, M. C. et al. Noninvasive Discrimination of Rejection in Cardiac Allograft Recipients Using Gene Expression Profiling. Am. J. Transplant 6, 150–160 (2006).
6. ENCODE: Encyclopedia of DNA Elements – ENCODE. Available at: https://www.encodeproject.org/. (Accessed: 9th March 2017)
7. Teixeira, M. R. Recurrent Fusion Oncogenes in Carcinomas. CRO 12, (2006).
8. Font-Tello A, E. al. Association of ERG and TMPRSS2-ERG with grade, stage, and prognosis of prostate cancer is dependent on their expression levels. - PubMed - NCBI. Available at: https://www.ncbi.nlm.nih.gov/pubmed/25939480. (Accessed: 9th March 2017)
9. Mitelman Database of Chromosome Aberrations and Gene Fusions in Cancer. Available at: https://cgap.nci.nih.gov/Chromosomes/Mitelman. (Accessed: 9th March 2017)
10. Maher, C. A. et al. Transcriptome sequencing to detect gene fusions in cancer. Nature 458, 97–101 (2009).
11. Ku, G. M. et al. Research Resource: RNA-Seq Reveals Unique Features of the Pancreatic □-Cell Transcriptome. Mol. Endocrinol. 26, 1783 (2012).
12. Cavelier, L. et al. Clonal distribution of BCR-ABL1 mutations and splice isoforms by single-molecule long-read RNA sequencing. BMC Cancer 15, 45 (2015).
13. Berger, M. F. et al. Integrative analysis of the melanoma transcriptome. Genome Res. 20, 413–427 (2010).
14. Lujambio, A. & Lowe, S. W. The microcosmos of cancer. Nature 482, 347–355 (2012).

15. Cancer Origin Test | RosettaGXTM by Rosetta Genomics®. Available at: http://rosettagx.com/testing-services/cancer-origin. (Accessed: 9th March 2017)

16. Li, H.-D., Menon, R., Omenn, G. S. & Guan, Y. The emerging era of genomic data integration for analyzing splice isoform function. Trends Genet. 30, 340–347 (2014).

17. Eksi, R. et al. Systematically Differentiating Functions for Alternatively Spliced Isoforms through Integrating RNA-seq Data. PLoS Comput. Biol. 9, (2013).

18. Twine, N. A., Janitz, K., Wilkins, M. R. & Janitz, M. Whole Transcriptome Sequencing Reveals Gene Expression and Splicing Differences in Brain Regions Affected by Alzheimer's Disease. PLoS One 6, e16266 (2011).

19. Mehta, S. et al. A Study of TP53 RNA Splicing Illustrates Pitfalls of RNA-seq Methodology. Cancer Res. 76, 7151–7159 (2016).

20. Conesa, A. et al. A survey of best practices for RNA-seq data analysis. Genome Biol. 17, 13 (2016).

21. Buenrostro, J. D., Wu, B., Chang, H. Y. & Greenleaf, W. J. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. Curr. Protoc. Mol. Biol. 109, 21.29.1–9 (2015).

22. Smallwood, S. A. et al. Single-cell genome-wide bisulfite sequencing for assessing epigenetic heterogeneity. Nat. Methods 11, 817–820 (2014).

23. brandonvd. Homepage - Allomap. Allomap Available at: http://www.allomap.com/. (Accessed: 9th March 2017)

24. CancerTYPE ID | The Most Widely Adopted Molecular Test for Metastatic Patients with Diagnostic Ambiguity. Available at: https://www.cancertypeid.com/. (Accessed: 9th March 2017)

25. Clinical Studies. Available at: https://www.afirma.com/dossier/clinical-studies/. (Accessed: 9th March 2017)

26. Lung Cancer | Exosome Diagnostics. Available at: http://www.exosomedx.com/lung-cancer-0. (Accessed: 9th March 2017)

27. FoundationOne. Available at: http://foundationone.com/learn.php. (Accessed: 9th March 2017)

28. SEQC/MAQC-III Consortium. A comprehensive assessment of RNA-seq accuracy, reproducibility and information content by the Sequencing Quality Control Consortium. Nat. Biotechnol. 32, 903–914 (2014).