MiPanda: A Resource for Analyzing and Visualizing Next-Generation Sequencing Transcriptomics Data

Yashar S. Niknafs*, Balaji Pandian*, Tilak Gajjar*, Zach Gaudette*, Kevin Wheelock*, Mitra P. Maz*, Rohan K. Achar*, Melinda Song*, Cory Massaro*, Xuhong Cao* and Arul M. Chinnaiyan*†,‡,§,¶,#

*Michigan Center for Translational Pathology, University of Michigan, Ann Arbor, MI, USA; †Department of Pathology, University of Michigan, Ann Arbor, MI, USA; ‡Department of Computational Medicine and Bioinformatics, Ann Arbor, MI, USA; §Howard Hughes Medical Institute, University of Michigan, Ann Arbor, MI, USA; †Rogel Cancer Center, University of Michigan, Ann Arbor, MI, USA; §Department of Urology, University of Michigan, Ann Arbor, MI, USA

Abstract

The Michigan Portal for the Analysis of NGS data portal (http://mipanda.org) is an open-access online resource that provides the scientific community with access to the results of a large-scale computational analysis of thousands of high-throughput RNA sequencing (RNA-seq) samples. The portal provides access to gene expression profiles, enabling users to interrogate expression of genes across myriad normal and cancer tissues and cell lines. From these data, tissue- and cancer-specific expression patterns can be identified. Gene-gene coexpression profiles can also be interrogated. The current portal contains data for over 20,000 RNA-seq samples and will be continually updated.

Neoplasia (2018) 20, 1144–1149

Introduction

Recently, there has been a rapid increase in the availability of high-quality tumor RNA sequencing (RNA-seq) datasets provided by multi-institutional consortia such as the TCGA, ICGC, GTEX, and CCLE [1–4]. These data have provided a robust resource for discovery and investigation of the driving events in various human pathologies from many different tissues. The GTEX consortium has provided a rich resource for expression in normal human tissues, and the TCGA consortium has provided a multitude of cancer RNA-seq samples from dozens of different cancer types. These datasets have enriched our understanding of cancer biology. Cancer genomics data have been interrogated extensively, and a multitude of mutation and copy-number aberrations have been identified as recurrent potential cancer drivers [5]. Additionally, however, gene expression patterns are also a meaningful data modality that can be leveraged to further investigate those genes most likely to be involved in oncogenesis.

Analysis of transcription enables identification of potential players in cancer that may not be discovered if only looking at genomic aberrations and also provides potential candidates for diagnostic and prognostic biomarkers. Historically, many microarray studies have been performed to identify genes differentially expressed in cancer, and tools including ONCOMINE [6] are available to enable the community to investigate expression profiles of genes in cancer. While tools such as ONCOMINE are powerful, microarray expression is limited because it can only investigate genes for which probes exist on the microarray platforms being used. Additionally, given the heterogeneity of the different microarray platforms being used, cross-study analysis is also limited (i.e., comparison of one microarray study to another is nontrivial). By compiling a vast RNA-seq compendium from human tissue and cell lines and performing systematic gene expression analyses, we open the door for...
transcriptomics analyses on an unprecedented scale of tens of thousands of samples. We are also able to integrate RNA-seq data from various sources as long as all RNA-seq libraries are of comparable quality and coverage to produce powerful cross-tissue lineage specificity analyses using the thousands of RNA-seq libraries available.

Various tools have been developed to provide the scientific community with easy access to the vast amount of cancer genomics analyses available including the UCSC Cancer Genome Browser [7], the cBio Portal [8], and others [9–11]. While immensely helpful to the scientific community, there is still need for a tool that provides powerful gene expression analyses and visualizations. Here we present an online portal, the Michigan Portal for the Analysis of NGS Data (MiPanda) (http://mipanda.org), built to enable the scientific community to begin investigating the robust transcriptional data in human normal and cancer samples (Figure 1). MiPanda is a product we are developing that will enable biomedical researchers to access the vast amount of next-generation sequencing data that have been generated in recent years. While much of these data are “publicly available,” they are stored in a raw, unprocessed format, preventing anyone without expertise in bioinformatics and powerful computing resources from interfacing with these data in a usable and meaningful way. Through a wide curation effort, we have obtained, annotated, and processed over 25,000 RNA-seq samples. Samples were processed using our RNA-seq analysis pipeline. These processed NGS data will be easily accessible using the MiPanda web tool, which facilitates data search and higher-level analyses. As the tool expands, it will integrate other genomics data modalities (e.g., exome capture mutation identification, copy number profiling, gene fusion identification, and immune-oncology).

Data Collection and Analysis

In its current form, MiPanda provides analysis and visualization of expression data from over 20,000 RNA-seq samples. This cohort is currently comprised of 1) normal human tissue from 21 tissue types from a combination of TCGA (n = 730) and GTEX (n = 9236) samples, 2) human primary cancer samples from the TCGA (n = 9496), 3) human metastatic cancer samples from the Michigan Center for Translational Pathology (n = 125) [12,13], and 4) human cancer cell lines from the CCLE (n = 935). Curation of the metadata associated with these samples was performed to standardize metadata fields across different cohorts so that they can be analyzed together (Figure 2). Raw sequencing reads were obtained and processed, yielding and expression estimates in TPM form for all samples. Combining the curated metadata fields with expression estimates provides a platform to generate powerful visualizations for expression data. Tissue and cancer status for all samples enables the identification of tissue- and cancer-specific expression patterns, as well as differential expression in a cancer versus normal context.

MiPanda Features

Currently, MiPanda provides the ability to visualize gene expression for 58,037 genes, comprised of 21,925 protein-coding genes, 14,651 pseudogenes, 14,181 lncRNAs, and 7280 other noncoding RNAs. Expression across the 20,967-sample cohort can be visualized in multiple ways, showcasing various groupings for the data. For example, expression can be visualized for all cancers across the different cancer types, for normal tissues separated by the tissue type, for cell lines separated by cell line tissue/cancer type, and combinations such as cancer versus normal (for tissues that have sufficient numbers of both cancer and normal).

At present, gene expression can be visualized for one gene at a time with boxplot depictions, or comparing the expression of two genes via a coexpression scatterplot. Selecting one gene enables identification of tissue and cancer expression patterns. One use case is a user with interest in knowing the distribution of expression for a given gene in normal human tissues. MiPanda can quickly display these data, for example, searching for the gluconeogenic enzyme PEPCK (PCK1) in normal tissues displays expression in the ~10,000-sample normal cohort, exhibiting expression of PEPCK predominantly in kidney, liver, and biliary tissue. Given that gluconeogenesis is a phenomenon occurring only in the kidney and liver, this expression pattern is appropriate (Figure 3A) and serves as a proof of concept for how MiPanda can be utilized to glean meaningful information regarding gene expression patterns.

MiPanda also provides information regarding expression patterns for cancer samples contrasted with normal samples. In this way, users can quickly identify genes whose expression is cancer specific. Leveraging such a large RNA-seq cohort confers the ability to identify global expression patterns across a multitude of tissues. For example, visualizing expression for the gene HER2 (ERBB2) using the “all
tissue cancer vs normal view displays a compelling largely cancer-specific expression pattern for HER2, wherein it is expressed highly in cancer only in a myriad of cancer types, while normal expression is significantly lower than cancer expression (Figure 3B). This phenomenon of identifying broad-scope cancer expression patterns by leveraging these data is emphasized through another example.

Figure 2. MiPanda sample distribution. Pie chart depictions of the distributions of MiPanda samples by (A) cohort, (B) cancer progression, and (C) tissue type.
Figure 3. Examples of expression visualization. (A) Boxplot depiction of the expression for *PCK1* (PEPCK) in a panel of normal tissues. Kidney, liver, and biliary are specifically indicated. (B) Expression profile of *ERBB2* (HER2) across all MiPanda tissue types that have at least five samples for both cancer and normal categories. (C) Expression profile for *FOLH1* (PSMA) across all MiPanda samples.
interrogating the expression of the gene FOLH1, which encodes for the protein prostate-specific membrane antigen (PSMA). PSMA’s compelling expression pattern has made it a valuable target for prostate cancer therapy and detection, as it is expressed specifically in prostate cancer tissue and has little to no expression in normal tissues or other cancer types [14]. This expression pattern can be seen clearly in MiPanda, emphasizing the impressive specificity of PSMA expression for prostate cancer, as it exhibits little to no expression in any of the other samples in the over 20,000-sample cohort (Figure 3C). Furthermore, expression analyses in single tissue types provide the opportunity to display data clearly to determine potential differential expression comparing cancer to normal. For example, the gene SPINK1 has been shown to be expressed in a subset of prostate cancers [15], while the majority of prostate cancer samples have low SPINK1 expression. MiPanda corroborates these data in a large 680-sample prostate cancer cohort, clearly depicting high expression in a small subset of “outlier” samples in prostate cancers only (Figure 4A).

MiPanda also offers the ability to visualize expression of two genes together in a scatterplot, allowing the user to identify coexpression patterns between genes. Given the broad tissue and cancer types within MiPanda, the coexpression visualization can rapidly identify gene-gene associations in various tissues and cancer types. For example, exploration of the estrogen receptor (ESR1) and one of its target genes, GATA3, exhibits correlation in breast normal and breast cancer samples, while these genes display low expression with little to no correlation in other tissues (Figure 4B). This analysis provides a proof of concept for the utility of MiPanda in displaying coexpression analyses in the large RNA-seq compendium.

Every plot generated in MiPanda can be downloaded as a publication-quality SVG file that can be edited for formatting, and can be easily integrated into manuscripts. Users can also modify plot format and labels to optimize the images for inclusion in manuscripts. Additionally, the raw expression data for each plot can be downloaded in Excel or CSV format if users desire to manipulate the data themselves. There is also a feature for users to download bulk expression data, selecting multiple genes at once and selecting the samples for which they want to download the expression data.

Future Directions
MiPanda utilized the wealth of cancer genomics and transcriptomics data available currently to provide analyses of the expression profiles of genes in normal and cancer tissues. Other widely used cancer genomics tools focus on mutational and copy number events driving cancers but do not provide an easy method for assessing expression levels of genes in cancer tissues. MiPanda enables the scientific community to evaluate the expression patterns of genes in cancer in an easy, user-friendly way. Future directions for our tool include

---

**Figure 4.** Outlier and co-expression. (A) Expression of SPINK1 in prostate samples, separated into primary cancer, metastasis, cell line, and normal categories. (B) Scatterplot depiction of GATA3 and ESR1 (ER) expression in the entire MiPanda cohort, colored by tissue type.
added features to survey gene sets and their enrichment in various samples, differential expression analyses for cancer versus normal analyses, and a tool to identify gene-gene correlations across all genes with subsequent gene set enrichment testing. In the longer term, we intend to add in genetic variant data for single nucleotide variants, indels, and copy number variants. Inclusion of these data will enable powerful associations of gene expression to these other genomics datasets.

Used alongside other robust tools such as the cBio and other genomics portals, MiPanda enables the scientific community to utilize vast amounts of data to interrogate scientific questions in cancer.

Methods

Data Collection

Sequencing samples and their associated metadata were obtained from dbGAP. Manual curation of the metadata was performed to create unified fields and unified values so that samples could be compared across studies.

Data Curation

Metadata for TCGA, GTEx, and CCLE were obtained. Fields accounted for in all datasets were selected for curation (e.g., tissue type). All datasets were curated such that TCGA, GTEx, and CCLE samples all contained standardized values for these fields.

RNA-seq Data Processing

RNA-sequencing reads were quantified to the human transcriptome (GENCODEv25) using Kallisto (v0.43.0) [16]. GENCODEv25 GTF was obtained from GENCODE [17], and transcriptome fasta file was produced using the rsem-prepare-reference function of RSEM (version 1.2.26) [18]. Kallisto index was generated using the kallisto index function. Transcript-level quantification was obtained using the kallisto quant function. Gene-level expression was obtained by summing the TPM values for all transcripts within each gene.

MiPanda

The web portal was developed with a backend consisting of a PostgreSQL database, a Node.js API server, and a front end client built using React.js. Plots were created using Plotly.js.

Acknowledgements

We thank Matthew Iyer for early transcriptomics work that was the precursor to the work presented here. We also thank Nicholas Molan for early web development work for the portal. A. M. C. is a Howard Hughes Medical Institute Investigator, A. Alfred Taubman Scholar, and American Cancer Society Professor. This work was supported by the Prostate Cancer Foundation, Early Detection Research Network (U01 CA214170), and NCI Prostate SPORE (P50 CA186786). Y. S. N. is supported by a Prostate Cancer Foundation Young Investigator Award.

References

[1] Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shimulevich I, Sander C, and Stuart JM (2013). The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet 45, 1113–1120.
[2] The GTEx Consortium (2013). The Genotype-Tissue Expression (GTEx) project. Nat Genet 45, 580–585.
[3] International Cancer Genome Consortium, Hudson TJ, Anderson W, Artez A, Barker AD, Bell C, Bernabé RR, Bhan MK, Calvo F, and Eerola I, et al (2010). International network of cancer genome projects. Nature 466, 993–998.
[4] Barretina J, Caponigro G, Stransky N, Venkatesan K, Margolin AA, Kim S, Wilson CJ, Lehár J, Kryukov GV, and Sonkin D, et al (2012). The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature 483, 603–607.
[5] Kandoh C, McLellan MD, Vandin F, Ye K, Niu B, Lu C, Xie M, Zhang Q, Micheal JP, and Wyczalkowski MA, et al (2013). Mutational landscape and significance across 12 major cancer types. Nature 502, 333–339.
[6] Rhodes DR, Yu J, Shanker K, Deshpande N, Varambally R, Ghosh D, Barrette T, Pandey A, and Chinnaiyan AM (2004). ONCOMINE: a cancer microarray database and integrated data-mining platform. Neoplasia 6, 1–6.
[7] Karolchik D, Barber GP, Casper J, Clawson H, Cline MS, Diekhans M, Dreszer TR, Fujita PA, Gurvadov L, and Haussler M, et al (2014). The UCSC Genome Browser database: 2014 update. Nucleic Acids Res 42, 764–770.
[8] Cerami E, Gao J, Dogrusoz U, Gross BE, Sumer SO, Aksoy BA, Jacobsen A, Byrne C, Heuer ML, and Larsson E, et al (2012). The cBio Cancer Genomics Portal: an open platform for exploring multidimensional cancer genomics data. Cancer Discov 2, 401–404.
[9] Deng M, Brägelmann J, Kryukov I, Saraiva-Agostinho N, and Perner S (2017). FirebrowseR: an R client to the Broad Institute’s Firehose Pipeline. Database 2017.
[10] Chandrashekar DS, Bashel B, BalaSubramanayam S, Creighton CJ, Ponce-Rodriguez I, Chavaththi BVSK, and Varambally S (2017). UALCAN: a portal for facilitating tumor subgroup gene expression and survival analyses. Neoplasia 19, 649–658.
[11] Samur MK, Yan Z, Wang X, Cao Q, Munshi NC, Li C, and Shah PK (2013). canEvolve: a web portal for integrating oncogenicities. PLoS One 8.
[12] Robinson DR, Wu YM, Lonigro RJ, Vats P, Cobain E, Everett J, Cao X, Rabbab E, Kumar-Sinha C, and Raymond V, et al (2017). Integrative clinical genomics of metastatic cancer. Nature 548, 297–303.
[13] Robinson D, Van Allen EM, Wu YM, Schulz N, Lonigro RJ, Mesquera JI, Montgomery B, Taplin ME, Pritchard CC, and Attard G, et al (2015). Integrative clinical genomics of advanced prostate cancer. Cell 161, 1215–1228.
[14] Hrachuk J, Von Hoff D, Mukkaram Ali M, Andrianova E, Auer J, Campbell T, De Witt D, Figa M, Figueiredo M, and Horhota A, et al (2012). Preclinical development and clinical translation of a PSMA-targeted docetaxel nanoparticle with a differentiated pharmacological profile. Sci Transl Med. 4.
[15] Tomlins SA, Rhodes DR, Yu J, Varambally S, Mehra R, Perner S, Demichelis F, E, Kumar-Sinha C, and Raymond V, et al (2017). Integrative clinical genomics of advanced prostate cancer. Cancer Cell 1228.
[16] Bray NL, Pimentel H, Melsted P, and Pachter L (2016). Near-optimal RNA-seq quantification. Nat Biotechnol 34, 525–527.
[17] Harrow J, Frankish A, Gonzalez JM, Tapanari E, Diekhans M, Kokocinski F, Aken BL, Barrell D, zadissa A, and Searle S, et al (2012). GENCODE: the reference human genome annotation for the ENCODE project. Genome Res 22, 1760–1774.
[18] Li B and Dewey CN (2011). RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinform 12, 323–338.