TED toolkit: a comprehensive approach for convenient transcriptomic profiling as a clinically-oriented application

Thahmina Ali[1], Baekdoo Kim[1], Carlos Lijeron[1], Olorunseun O. Ogunwobi[1,2,3], Raja Mazumder[5,6], and Konstantinos Krampis[1,2,4]

[1]**Weill Cornell Medicine - Belfer Research Building, Hunter College of The City University of New York, New York, NY**
[2]**Department of Biological Sciences, Hunter College of The City University of New York, NY**
[3]**Joan and Sanford I. Weill Department of Medicine, Weill Cornell Medical College, Cornell University, New York, NY**
[4]**Department of Physiology and Biophysics, Institute for Computational Biomedicine, Weill Cornell Medical College, Cornell University, New York, NY**
[5]**The Department of Biochemistry & Molecular Medicine The George Washington University Medical Center, Washington, DC**
[6]**The McCormick Genomic and Proteomic Center, The George Washington University, Washington, DC**

Corresponding author:
Konstantinos Krampis[1,2,4]

Email address: kk104@hunter.cuny.edu

ABSTRACT

In translational medicine, the technology of RNA sequencing (RNA-seq) continues to prove powerful, and transforming the RNA-seq data into biological insights has become increasingly imperative. We present the Transcriptomics profiler for Easy Discovery (TED) toolkit, a comprehensive approach to processing and analyzing RNA-seq data. TED is divided into three major modules: data quality control, transcriptome data analysis, and data discovery, with eleven pipelines in total. These pipelines perform the preliminary steps from assessing and correcting the quality of the RNA-seq data, to the simultaneous analysis of five transcriptomic features (differentially expressed coding, non-coding, novel isoform genes, gene fusions, alternative splicing events, genetic variants of somatic and germline mutations) and ultimately translating the RNA-seq analysis findings into actionable, clinically-relevant reports. TED was evaluated using previously published prostate cancer transcriptome data where we observed previously studied outcomes, and also created a knowledge database of highly-integrated, biologically relevant reports demonstrating that it is well-positioned for clinical applications. TED is implemented on an instance of the Galaxy platform (Galaxy page: http://galaxy.hunter.cuny.edu/u/bioitcore/p/transcriptomics-profiler-for-easy-discovery-ted-toolkit, Documentation Manual: http://ted.readthedocs.io/en/latest/index.html) as intuitive and reproducible pipelines providing a manageable strategy for conducting substantial transcriptome analysis in a routine and sustainable fashion for bioinformatics and clinical researchers alike.

INTRODUCTION

The modern sequencing technology, next generation sequencing (NGS) has expanded the analytical possibilities of the transcriptome in complete depth, the method known as RNA-sequencing (RNA-seq). RNA-seq can precisely determine the abundance of transcripts expressed in any RNA sample of study. Moreover, given the emergence of RNA-seq applications in many biomedical research areas, there are significant efforts in standardizing the method (1) within clinical settings. In the clinical laboratory, investigating the transcriptome has uncovered invaluable information of genetic mechanisms within a RNA sample of a conditioned or diseased individual (2, 3). The thorough view of the transcriptome

offered by RNA-seq offers ways for identifying disease causing bio-molecules of an individual that can serve as potential diagnostic indicators. This is especially applicable to complex diseases like cancer, where multiple bio-molecules contribute to its abnormal state, and findings through RNA-seq can be used as a reliable resource for therapeutic targets. In parallel with the considerable RNA-seq applications in the clinic, analyzing the RNA-seq data is essential, but delivering the biological insights unraveled from the analysis in the most informative means has become just as crucial. There are various data analysis programs most notably the Galaxy biomedical research platform (4) that addresses challenges such as the issues of accessibility and reproducibility. The platform provides an intuitive web based interface that serves as a workspace for data analysis in which researchers can import their data sets, and apply bioinformatics tools that are made available from the Galaxy toolshed (5) panel. Galaxy tools can run as standalone or chained together to create larger analyses transforming entire bioinformatics pipelines into automated "Galaxy workflows". By Galaxy offering the ability to create and perform automated analyses on a user interface fully operational on the web, bioinformatics analyses have become more approachable in doing all types of data analysis. Yet, there still does not exist a convenient framework mainstream enough to enable RNA sequencing analysis results in a way that readily lends itself to easy interpretation. The current approach of performing RNA sequencing analyses is difficult, especially for non-bioinformatics researchers for the following reasons: (i) analysis methods and protocols are organized in a non-uniformed manner; (ii) analysis methods dependencies, parameters or supporting data come across as undocumented (iii) analyses output is in raw file state that consist of incomprehensible results with no set process to interpret them. These aspects lead to prolonged complexity requiring a learning curve to understand and tackle them which in turn causes a distraction in performing the actual analysis, making standardizing RNA sequencing analysis as a diagnostic practice challenging.

The bioinformatics pipelines that have been developed on the Galaxy platform, have had a focus on automation and standardization, including several pipelines available for transcriptomic data analysis. For example, the Oqtans (6) workbench performs differential expression and enrichment analysis and the open pipelines for tumor genome profiling that consist of three separate analyses pipelines: exome, transcriptome and variant evaluation (7). In addition, the TRAPLINE pipeline (8) performs comparative transcriptomics analysis, identifying a set of differentially expressed genes and their corresponding protein-protein interactions. These Galaxy pipelines have accelerated the extensibility in the transcriptome data analysis, however, in order to visualize the outputs requires importing to external programs. For example, the TRAPLINE protein-protein interactions output requires the Cytoscape program for visualization, in which this method does not enable direct interpretation delivered straight from the analysis exclusively. There are other automated pipelines that are taking initiatives in striving to bring out the most informed data analysis, by way of a software application approach. RNAseq software methods such as RobiNA (9) which uses a biostatistical method and Grape (10), both of which provide an environment to analyze and visualize gene expression data but limited to solely performing differential gene expression analysis. The Chipster (11) platform houses a comprehensive collection of analysis tools that covers analysis other than gene expression, such as miRNA, methylation and others, yet has complicated installation procedures, as well as, technical navigation again requiring a learning curve for non-informatics individuals. There are methods that function on the web such as MeV (12) which is cloud based that is also limited to performing differential gene expression analysis and visualization and the functionalities offered stratify the data analysis with curations that consist of no annotative feature especially with biological content. Nevertheless, each of these applications still are contributors to the steps towards the potential for standardizing RNA sequencing within the reach of translational and diagnostic settings.

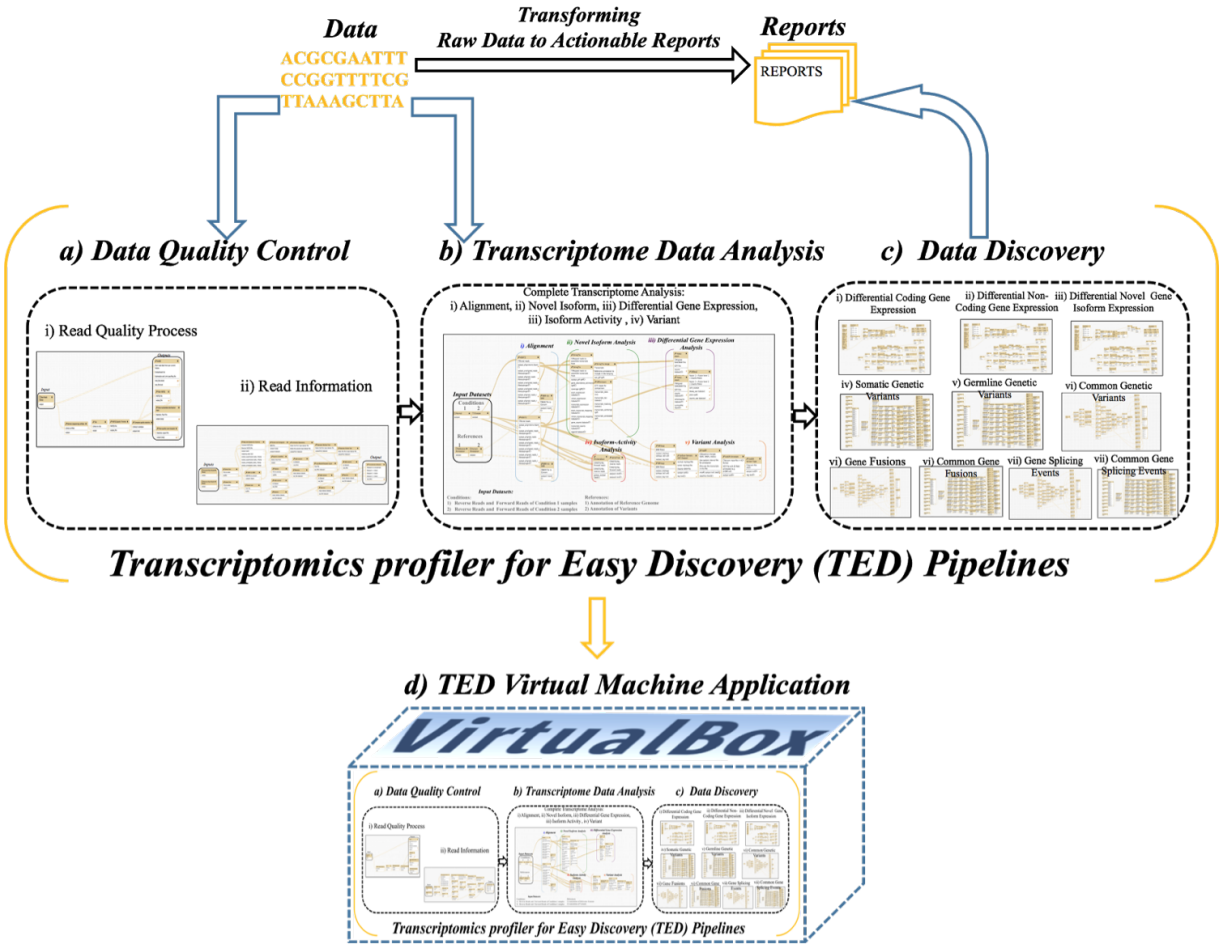
We propose a highly-integrated set of bioinformatics pipelines designed in the form of automated workflows, which are implemented into the Galaxy platform. The workflows are configured to perform quality control and analysis on RNAseq data, while also providing beyond the standard analysis in order to provide data discovery functionality. The entire set of workflows is packaged as a resource toolkit, termed Transcriptome profiler for Easy Discovery, or TED. TED has three fundamental modules, summarized in Fig. 1. The first module provides quality control of the RNAseq data which are preprocessing steps, as well as, acquiring information about the reads such as read length, insert size etc. The second module carries out analysis of differentially coding, non-coding and novel isoform gene expression, gene fusions, alternative splicing events, and genetic variants of somatic and germline mutations of the RNAseq data. And lastly, the third module transforms the analysis results produced from the second module into detailed, biologically interpreted annotated reports. TED joins these three modules together creating a

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3385v1 | CC BY 4.0 Open Access | rec: 1 Nov 2017, publ: 1 Nov 2017

knowledge database of prioritize biological outcomes, enabling users to obtain a comprehensive insight of the transcriptome analyzed from the RNA samples. TED becomes extensible to applications in clinical or diagnostic scenarios, allowing the user as a clinician or practitioner to leverage their experience to data mine the reports of analyzed results for discovery or indication of biological candidates to examine.

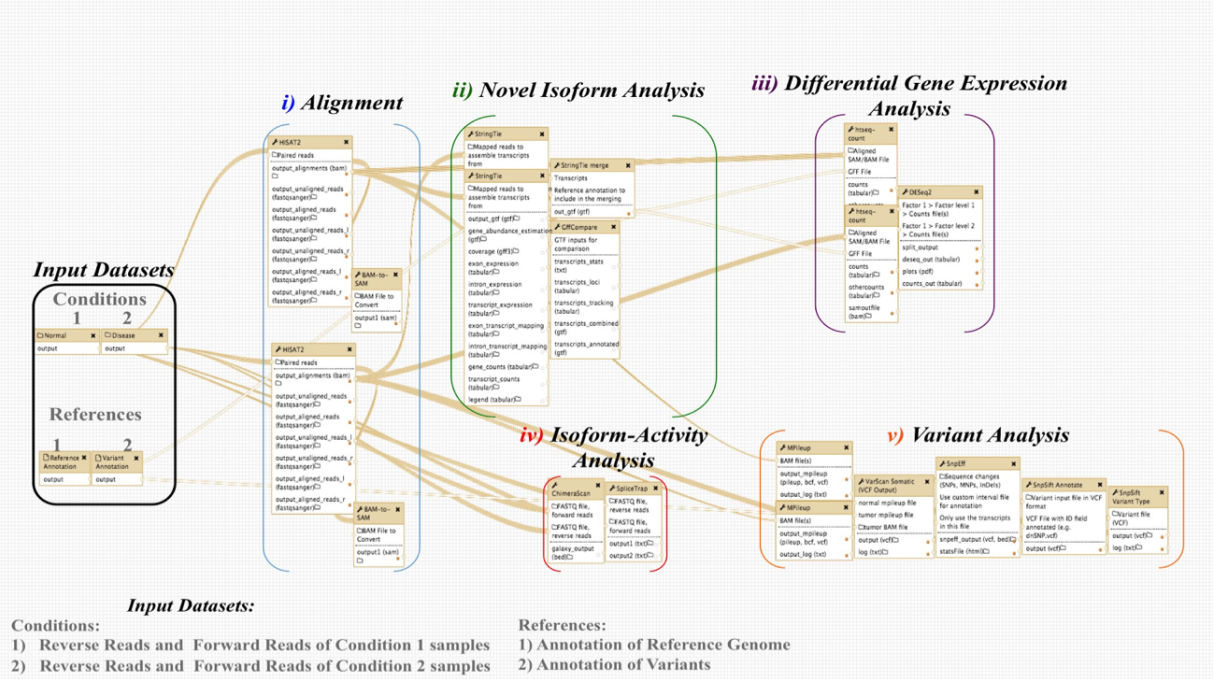
We document an example use case of TED with previously published prostate cancer transcriptome data (13) . We have developed a methodology that can provide the components of data analysis of complex RNA-seq datasets through a toolkit interface that is easy to access, handle in addition to a comprehensive data processing solution that is reusable and practical for users without extensive bioinformatics expertise.

Figure 1. Overview of the Transcriptomics Profiler for Easy Discovery (TED) toolkit



PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3385v1 | CC BY 4.0 Open Access | rec: 1 Nov 2017, publ: 1 Nov 2017

b) TED Transcriptome Data Analysis



Transcriptome Data Analysis (Fig.1b) is the second module of TED comprised of five data analysis pipelines i) Alignment, ii) Novel Isoform, iii) Differential Gene Expression, iv) Isoform-activity and v) Variant analysis. This module consist of 14 bioinformatics tools and 24 steps that will analyze any number of paired-end RNA sequencing data samples from two conditions.

METHODS

Availability

The TED toolkit is freely accessible on our local instance of the Galaxy platform via a url link: http://galaxy.hunter.cuny.edu/workflows/list_published or through our custom Galaxy page: http://galaxy.hunter.cuny.edu/ profiler-for-easy-discovery-ted-toolkit, that contains details of the RNAseq pipeline, datasets, and tutorials of the transcriptome analysis as well as described in our documentation manual: http://ted.readthedocs.io/en/latest/. A user can create an account (14) on our local Galaxy instance in order to have a private workflow workspace, then import and run the pipelines directly from the URL links above. Furthermore, for each new pipeline run, the results are saved in a separate Galaxy history (15) under the user's account, which additionally offers a sharing option of the output through a simple web link. A virtual machine (VM) (16) including Galaxy with the TED toolkit is also provided, with the tools and software dependencies preinstalled for download through the Data Libraries on our local Galaxy, under 'TED Virtual Machine (VM) Application': (http://galaxy.hunter.cuny.edu/library/list#folders/Fb56e686e7a485784) and instructions to set up and use the TED VM can be found in our documentation manual mentioned earlier.

Data Source

A total of 56 RNA-seq datasets were retrieved from the Array Express database of the European Bioinformatics Institute (https://www.ebi.ac.uk/arrayexpress/experiments/E-MTAB-567/samples/, EBI). The files correspond to 14 sequenced transcriptomes from tumor tissue samples of prostate cancer human patients and a technical replicate for each sample (total 28) in addition to 14 sequenced matched samples from the healthy tissue adjacent to the tumor tissue with replicates as well (additional 28). The samples were collected, prepared and sequenced as described in the study by Ren et al (13). For each tumor and healthy sample the dataset sequencing reads are paired-end, with replicates of each forward and reverse sequencing read data files also included in the analysis. The EBI RNA-seq datasets are also available for download through our local Galaxy Data Libraries, under 'TED toolkit Data Source': (http://galaxy.hunter.cuny.edu/library/list#folders/F862a7cb864998e85) as well as other supporting data

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3385v1 | CC BY 4.0 Open Access | rec: 1 Nov 2017, publ: 1 Nov 2017

such as the reference genome and reference annotation files.

Implementation

The TED toolkit was implemented on our local instance of the Galaxy platform: http://galaxy.hunter.cuny.edu/ and freely accessible via a url link as mentioned in the 'Availability' section above. The TED pipelines consist of distinct bioinformatics software components and utilities, in which they were either downloaded and installed to our Galaxy instance via the public Galaxy toolshed (https://toolshed.g2.bx.psu.edu/), or manually integrated (17) in our local Galaxy toolshed in which all of the necessary custom tool scripts and wrappers are published as a repository in the main Galaxy toolshed (https://toolshed.g2.bx.psu.edu/view/bioitcore/transcriptom) as well as in our public code repository on Github (https://github.com/BCIL/TED). All of the pipelines were assembled on Galaxy's workflow editor, by connecting the tools for the separate stages of the pipelines. In addition a virtual machine (VM) application was designed and build to include Galaxy with the TED pipelines, tools and software dependencies pre-installed for download and execution to the researcher's machine (http://www.virtualbox.org).

RESULTS

TED is packaged as a toolkit that integrates eleven distinct pipelines, on the Galaxy workflow canvas (18) and currently supports analysis of paired-end RNA-seq datasets from the Illumina sequencing platform of the human organism and analyzes the transcriptomes of RNA-seq datasets from two conditions which outputs are available in the Galaxy history (15) for the user to view and use. Within TED, the set of pipelines are divided into three fundamental modules based on their functionality that includes Data Quality Control (Fig. 1a), Transcriptome Data Analysis (Fig. 1b), and Data Discovery (Fig. 1c). All of TED pipelines are available as published workflows on our Galaxy server mentioned in the 'Availability' section and can be imported into the workspace of a public or private Galaxy instance by using the generated workflow links we provide, so that users have the option to run the analysis on their own server (19). We also provide the TED toolkit, the Galaxy server and all of the required software dependencies preconfigured as a virtual machine image (Fig. 1d). This is to allow the entire TED toolkit components and units to operate on any type of physical machine and operating system, by loading and powering up its appliance image into a virtual machine application. In order to demonstrate the effectiveness and convenience of our comprehensive analysis toolkit for RNA-seq, TED was used to gain insight into the molecular pathogenesis of 14 human prostate cancer transcriptomes. Using the TED toolkit we identified a range of differentially expressed coding, non-coding, novel isoform genes, gene fusions, alternative splicing events, genetic variants of somatic and germline mutations in these datasets. The following results below will first describe the Transcriptome Data Analysis module to explain how TED analyzes RNA-seq datasets and present part of our results for the differentially expressed coding genes and produced from the Data Discovery pipeline.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3385v1 | CC BY 4.0 Open Access | rec: 1 Nov 2017, publ: 1 Nov 2017

Patient	Pvalue	Number of Identified Differentially Expressed Genes: Total Genes	Upregulated Genes	Downregulated Genes	Pearson Correlation of Identified Differentially Expressed Genes: Total Regulated Genes	Upregulated Genes	Downregulated Genes
1	$p \leq 0.01$	11008	5663	5345	0.93	0.92	0.96
	$p \leq 0.05$	11961	6133	5828	0.93	0.92	0.95
2	$p \leq 0.01$	10127	4865	5262	0.94	0.97	0.97
	$p \leq 0.05$	11287	5421	5866	0.87	0.98	0.97
3	$p \leq 0.01$	11513	5973	5540	0.93	0.92	0.96
	$p \leq 0.05$	12526	6436	6090	0.94	0.85	0.96
4	$p \leq 0.01$	11394	5676	5718	0.72	0.86	0.73
	$p \leq 0.05$	12199	6090	6109	0.72	0.86	0.73
5	$p \leq 0.01$	11352	5637	5715	0.85	0.95	0.89
	$p \leq 0.05$	12184	6074	6110	0.85	0.95	0.89
6	$p \leq 0.01$	11129	5533	5596	0.8	0.77	0.91
	$p \leq 0.05$	12088	6025	6063	0.8	0.77	0.91
7	$p \leq 0.01$	10954	5191	5763	0.97	0.93	0.99
	$p \leq 0.05$	11868	5670	6198	0.97	0.93	0.99
8	$p \leq 0.01$	11311	5839	5472	0.95	0.95	0.83
	$p \leq 0.05$	12126	6220	5906	0.96	0.95	0.83
9	$p \leq 0.01$	9910	4815	5095	0.34	0.96	0.35
	$p \leq 0.05$	10999	5366	5633	0.35	0.96	0.35
10	$p \leq 0.01$	5892	3289	2603	0.85	0.9	0.85
	$p \leq 0.05$	6870	3794	3076	0.85	0.9	0.85
11	$p \leq 0.01$	11557	5846	5711	0.85	0.93	0.86
	$p \leq 0.05$	12510	6371	6139	0.85	0.93	0.86
12	$p \leq 0.01$	9881	5121	4760	0.84	0.85	0.94
	$p \leq 0.05$	11077	5677	5400	0.83	0.85	0.9
13	$p \leq 0.01$	11372	5550	5822	0.85	0.87	0.9
	$p \leq 0.05$	12278	6040	6238	0.85	0.87	0.9
14	$p \leq 0.01$	9910	5095	4815	0.94	0.96	0.95
	$p \leq 0.05$	11883	6271	5612	0.85	0.98	0.94

Table 1. Quantitative Summary of Differentially Expressed Genes and Pearson Correlation Statistics

Transcriptome Data Analysis

The Transcriptome Data Analysis (Fig.1b) is the second module of TED comprised of five data analysis pipelines i) Alignment, ii) Novel Isoform, iii) Differential Gene Expression, iv) Isoform-activity and v) Variant analysis. This module consist of 14 bioinformatics tools and 24 steps that will analyze any number of paired-end RNA sequencing data samples from two conditions.

The Alignment (Fig. 1b.i.) pipeline uses the UCSC hg38 reference genome (20), with the HISAT2 (21) alignment program. HISAT2 uses an indexing technique to enable faster searches on the genome file, which consist of a global index that covers the whole genome and many other small indexes for regions that collectively cover the genome, to map whole reads entirely in the exons in which the Bowtie2 aligner handles many of the operations required to construct and search the genome indexes. HISAT2 identifies reads that span the exonic region as read alignments and the gaps between the spanning exonic regions as junction signals. The output from this step is a Binary Alignment file (BAM, (22)), containing the mapped exonic reads and their positions in the reference genome. To view the alignment file in text format, we a BAM-SAM conversion step is included in the pipeline. All parameters were left as default, except the minimum and maximum fragment length which are to be specified by the user that refers to the range of the fragment size of the sequencing reads. This information can be found in the Read Information pipeline that's part of the TED's first module Data Quality Control (not mentioned in this draft). Once the alignment step is complete, the pipeline proceeds in four different analyses paths, the first for the novel isoforms of the expressed genes, the second for differentially expressed genes (noncoding and coding), the third for alternative splicing events and gene fusions and lastly the fourth for genetic variants of the expressed genes.

The Novel Isoform analysis pipeline consists of 3 bioinformatics tools performing 4 steps (Fig. 1b.ii.), with the Stringtie (23) software at its core, for reconstructing and quantifying the set of transcripts, and number of gene isoforms, from the aligned transcriptome read data with the annotations of the reference genome. The Stringtie assembler, uses as input the alignment file of mapped exonic reads produced from HISAT2. The approach this software takes is, it builds an alternative splice graph from overlapping reads in a given locus. This graph will contain nodes that corresponds to exons, and edges that corresponds to

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3385v1 | CC BY 4.0 Open Access | rec: 1 Nov 2017, publ: 1 Nov 2017

reads which connects the exons. Stringtie will identify a path in the generated splice graph that has the largest number of reads on the edges (highest weight). This selected path will resemble an assembled transcript and because the edge weight equals to the number of reads, StringTie estimates the coverage level for this transcript that can be used to estimate the transcript's abundance, thus performs assembly and quantification simultaneously for every identified transcript. After the procedure of associating the reads with the assembled transcripts completes, they are then removed and the graph will update to perform the iteration of the algorithm on the next transcript. Stringtie will generate a separate transcriptome assembly for each of the HISAT alignment input files in which will then merged together using the Stringtie merge software. This is to combine redundant transcript structures across the transcriptome assemblies and identify which transcript structure corresponds to which annotated transcript using a reference annotation file, from the UCSC hg38 reference annotations (20) in Gene Transfer Format (GTF) (24). The reference annotation file contains information about known genes and transcripts that will be used to annotate the origin and nature of each transcript in the transcriptome assemblies. Furthermore in our pipeline, the gffcompare utility (25) was used to determine the number of assembled transcripts in comparison to known transcripts in the merged transcriptome assembly. The gffcompare tool will use the same reference annotation file used in the Stringtie merge step and evaluate the assembled transcripts that matched with the annotated genes either fully, partially and which ones entirely novel for isoform discovery that are not annotated.

The Differential Gene Expression Analysis pipeline consist of 2 bioinformatics tools and 3 steps (Fig. 1b.iii.) to identify the transcripts that are differentially expressed between the two conditions of the RNA-seq experiment. This pipeline uses the htseq-count (26) tool from the HTSeq suite for counting the overlap of reads from the alignment files with annotation features, where each transcript is considered the union of all its exons. To count how many reads map to each transcript, the alignment files from HISAT2 are provided and the annotation file generated from the gffcompare tool which represents the transcripts present within the RNAseq datasets, as well as their location with non-redundant identifiers, and information regarding the origin. Then we will provide this information to DESeq2 (27) to generate normalized transcript counts (abundance estimates) and significance testing for differential expression.

The Variant Analysis (Fig. 1b.iii.) includes five tools and performs six steps, with the SAMtools Mpileup program (28), VarScan Somatic tool (29) and SnpEff suite (30). The pipeline receives as input the BAM files produced from TopHat2 from the Data Groom and Alignment stage in addition to a reference human genome, in order for the SAMtools Mpileup to collect summary information the likelihood of each possible genotype is computed from the data and stored in a file for future reference, in addition to pileup of read base differences in a binary Variant Call Format (VCF) (31) file of each dataset. The variant caller tool VarScan Somatic reads the Mpileup output files and produces germline, somatic, and Loss of heterozygosity (LOH) events at positions where both normal and prostate cancer datasets have sufficient read coverage. All parameters were left as default for Mpileup and VarScan Somatic, except for p-value significance threshold set to 0.01 for VarScan Somatic, in order to enable a more sensitive first-pass algorithm in determining positional variants, that occurred in the supplied normal and prostate cancer mpileup's. SnpEff is a variant annotation and effect prediction tool. It annotates and predicts the effects of genetic variants. The Isoform Level Analysis (Fig. 1b.iv.) pipeline consists of two bioinformatics tools and performs 4 steps that detect chimeric transcripts encoded by a fusion gene performed by the Chimerascan (32) tool and quantiying alternative splicing events performed by the SpliceTrap tool (33). The ChimeraScan tool in this pipeline aligns paired-end reads to a combined genome-transcriptome reference, to identify potential fusion breakpoints from fragments that align to distinct references, or distance genomic locations of the same reference which are referred to putative chimeric junction sequences. The junction sequences are then used as reference to realign candidate junction-spanning reads. Several output files will be produced and the key output file is a tabular text file named chimeras.bedpe. SpliceTrap detects alternative splicing in paired-end RNA-seq data by using a Bayesian inference approach, by quantifying for every exon the extent to which it is included, skipped or subjected to size variations due to alternative 3'/5' splice sites or intron retention.

Data Discovery

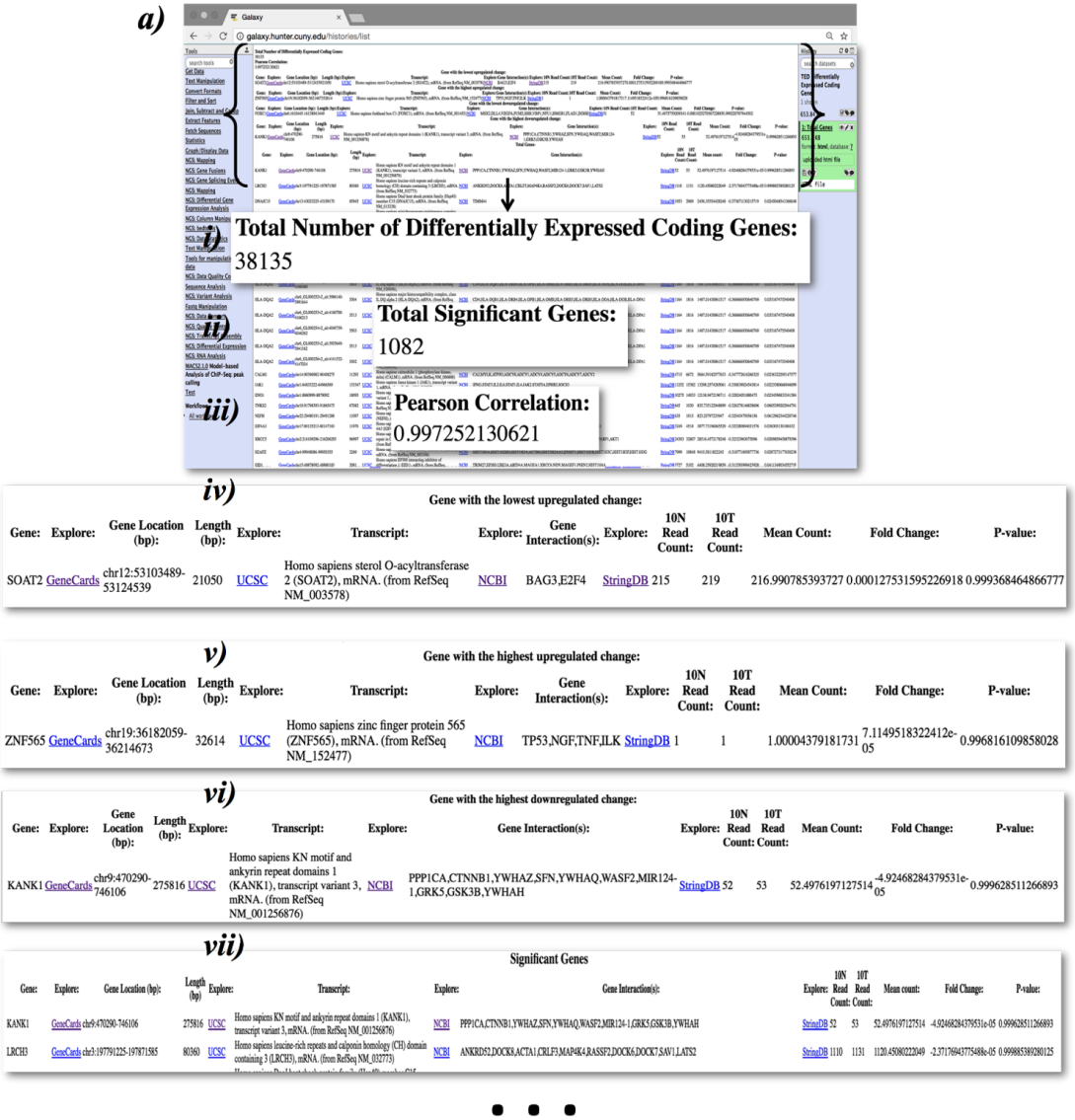
The Data Discovery module consists of eleven pipelines and utilizes a highly structured approach for aggregating and summarizing the results produced from the transcriptome data analysis module for easy assessment, interpretation and downstream discovery. The eleven pipelines in this module generate

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3385v1 | CC BY 4.0 Open Access | rec: 1 Nov 2017, publ: 1 Nov 2017

HyperText Markup Language (HTML) reports which can also be referred to as 'actionable reports,' that transforms the data results into thorough, concise and intuitive information reports, consist of differential coding gene expression (Fig. 1c.i.), differential non-coding gene expression (Fig. 1c.ii.), differential novel gene isoform expression (Fig. 1c.iii.), somatic genetic variants (Fig. 1c.iv.), germ line genetic variants (Fig. 1c.v.), comparison of genetic variants between samples (Fig. 1vi.), gene fusions (Fig. 1vii.), comparison of gene fusions between samples (Fig. 1c.viii.), gene splicing events (Fig. 1c.ix.), and (Fig. 1c.x.) comparison of gene splicing events between samples. The following below will describe the differential expression reports showing an example report of differential coding gene expression in Figure 2.

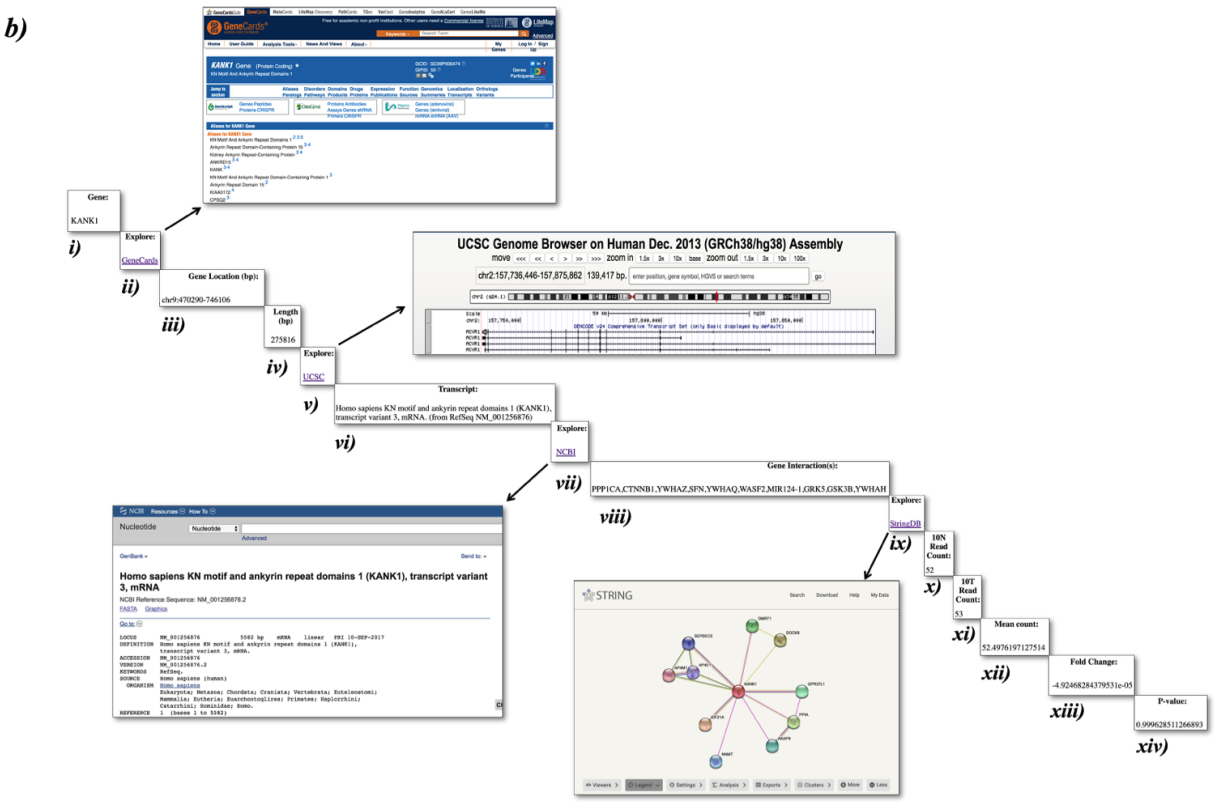
Differential Expression Report

Figure 2. TED Data Discovery Analysis Report: Differentially Expressed Coding Genes



For each of the pipelines generating differential coding gene, non-coding gene, and novel gene isoform expression reports takes inputs the output data produced from the HTseq tool and Deseq2 tools of the transcriptome data analysis module and generates three html reports Upregulated Genes, Downregulated and Total regulated genes. Figure 2a illustrates an example of the total regulated html report for differential coding gene expression of two sample RNAseq datasets (one sample for each condition), in which the

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3385v1 | CC BY 4.0 Open Access | rec: 1 Nov 2017, publ: 1 Nov 2017



pipeline will populate the report with the following information in 8 parts: i) total number of differentially expressed coding genes (Fig. 2a.i.), ii) total number of significant genes (Fig. 2a.ii.), iii) Pearson correlation statistic between the genes (Fig. 2a.iii.), iv) the gene with the highest upregulated differential fold change in FPKM (Fig. 2a.iv.), v) the gene with the lowest upregulated differential fold change in FPKM (Fig. 2a.v.), vi) the gene with the lowest downregulated differential fold change in FPKM (Fig. 2a.vi.), vii) the gene with the highest differential fold change in FPKM (Fig. 2a.vii.), and the viii) list of all the identified differentially expressed genes (Fig. 2a.viii.). The information of the gene with the highest differential fold changes, the lowest differential fold changes, and the list of all the differentially expressed genes are arranged in 14 columns (Fig. 2b). The 14 columns specify, i) the gene name (Fig. 2b.i), ii) link to GeneCards (34) a database of predicted human genes that provides concise genomic related information, iii) the chromosomal location in which the gene resides, iv) the gene length, v) link to UCSC Genome Browser (35) displaying the genomic location and other genomic data, vi) gene transcript description name, vii) link to NCBI nucleotide database (36) providing gene and transcript data from several sources, viii) list of genes involved in interaction with the differentially expressed gene ix) link to StringDB (37) database of known and predicted gene pathway networks displaying direct and indirect interactions x) read count of gene for sample 1 in normal condition xi) read count of gene for sample 1 for experimental condition xii) mean read count of the gene from both samples of both conditions xiii) fold change of gene between the conditions and xiv) significant statistic in pvalue.

In our study of analyzing the 14 prostate cancer transcriptomes, we identified from the generated reports, differentially expressed genes between paired prostate tumor and normal samples based on two separate criteria's: pvalue <=0.05 and pvalue <=0.01 (Table 1). We analyzed at both pvalue cutoffs of <=0.01 and <=0.05 patients 1, 3, 8, 10, 11, 12, and 14, exactly half of the patients of the study group, portrayed more up regulated genes expressed than down regulated genes at an exhibited correlation coefficient of 0.85< for each of the upregulated genes and downregulated genes. This observation emphasizes past findings in differential gene expression prostate cancer studies of the distribution of upregulated genes being larger than downregulated genes (38, 39).

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3385v1 | CC BY 4.0 Open Access | rec: 1 Nov 2017, publ: 1 Nov 2017

DISCUSSION

Compared to other RNA-seq and transcriptome analysis resources (6–9, 11, 12) that has its capabilities of reaching to a vast number of different scientific settings, the TED toolkit offers potential to reaching largely to a translational and diagnostic setting searching for a starting point of a preliminary overview of their RNAseq data that will lead to a discovery process with at most ease and minimal effort. In many clinical perspectives, RNA-seq delivers specific and sensitive genomic signatures but due to the lack of easy-to-use pipelines that can process in a transparent and streamlined fashion is limiting the expansion of RNA-seq from becoming a clinical diagnostic tool. Thus, the TED toolkit was intentionally designed as a Galaxy webserver since it allows inexperienced users to easily access advanced analysis tools processing the complex transcriptome analysis that will prepare unified outputs on a versatile workbench. Aside from TED being hosted on an accessible and intuitive system, it is also framed as a discovery platform that will structure the analysis results in html reports with analytical statistics and prioritization set with annotations and resource links to extremely comprehensive databases of disease and non-disease related information. This methodology offers a basic assessment of a RNA-seq study with initial details that are coherent as shown earlier with the differential gene expression results and can aid in the direction of a targeted discovery process to help come further to a conclusive clinical interpretation. Therefore, the TED toolkit holds strength to be a reliable, convenient and central protocol covering the majority aspects of transcriptome analytical results that is suitable to cater well within the reach of translational and diagnostic settings.

ACKNOWLEDGMENTS

The authors would like to thank all members of the Bioinformatics Core Infrastructures and Krampis Lab for their feedback during manuscript preparation.

REFERENCES

1. Kamalakaran,S., Varadan,V., Janevski,A., Banerjee,N., Tuck,D., McCombie,W.R., Dimitrova,N. and Harris,L.N. (2013) Translating next generation sequencing to practice: Opportunities and necessary steps. Mol. Oncol., 7, 743–755.

2. Cummings,B.B., Marshall,J.L., Tukiainen,T., Lek,M., Donkervoort,S., Foley,A.R., Bolduc,V., Waddell,L.B., Sandaradura,S.A., O'Grady,G.L., et al. (2017) Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. Sci. Transl. Med., 10.1126/scitranslmed.aal5209.

3. Christensen,S.M., Dillon,L.A.L., Carvalho,L.P., Passos,S., Novais,F.O., Hughitt,V.K., Beiting,D.P., Carvalho,E.M., Scott,P., El-Sayed,N.M., et al. (2016) Meta-transcriptome Profiling of the Human-Leishmania braziliensis Cutaneous Lesion. PLoS Negl. Trop. Dis., 10, e0004992.

4. Goecks,J., Nekrutenko,A. and Taylor,J. (2010) Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol., 11, R86.

5. Galaxy Tool Shed [Galaxeast Wiki].

6. Schultheiss,S.J., Jean,G., Behr,J., Drewe,P., Görnitz,N., Kahles,A., Mudrakarta,P., Sreedharan,V.T., Zeller,G. and Rätsch,G. (2011) Oqtans: a Galaxy-integrated workflow for quantitative transcriptome analysis from NGS Data. BMC Bioinformatics, 12, A7.

7. Goecks,J., El-Rayes,B.F., Maithel,S.K., Khoury,H.J., Taylor,J. and Rossi,M.R. (2015) Open pipelines for integrated tumor genome profiles reveal differences between pancreatic cancer tumors and cell lines. Cancer Med., 4, 392–403.

8. Wolfien,M., Rimmbach,C., Schmitz,U., Jung,J.J., Krebs,S., Steinhoff,G., David,R. and Wolkenhauer,O. (2016) TRAPLINE: a standardized and automated pipeline for RNA sequencing data analysis, evaluation and annotation. BMC Bioinformatics, 17, 21.

9. Lohse,M., Bolger,A.M., Nagel,A., Fernie,A.R., Lunn,J.E., Stitt,M. and Usadel,B. (2012) RobiNA: a user-friendly, integrated software solution for RNA-Seq-based transcriptomics. Nucleic Acids Res., 40, W622-7.

10. Knowles,D.G., Roder,M., Merkel,A. and Guigo,R. (2013) Grape RNA-Seq analysis pipeline environment. Bioinformatics, 29, 614–621.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3385v1 | CC BY 4.0 Open Access | rec: 1 Nov 2017, publ: 1 Nov 2017

11. Kallio,M.A., Tuimala,J.T., Hupponen,T., Klemelä,P., Gentile,M., Scheinin,I., Koski,M., Käki,J. and Korpelainen,E.I. (2011) Chipster: user-friendly analysis software for microarray and other high-throughput data. BMC Genomics, 12, 507.

12. Howe,E.A., Sinha,R., Schlauch,D. and Quackenbush,J. (2011) RNA-Seq analysis in MeV. Bioinformatics, 27, 3209–3210.

13. Ren,S., Peng,Z., Mao,J.-H., Yu,Y., Yin,C., Gao,X., Cui,Z., Zhang,J., Yi,K., Xu,W., et al. (2012) RNA-seq analysis of prostate cancer in the Chinese population identifies recurrent gene fusions, cancer-associated long noncoding RNAs and aberrant alternative splicings. Cell Res., 22, 806–821.

14. Galaxy.

15. Histories.

16. Oracle VM VirtualBox.

17. Adding custom tools to Galaxy.

18. Data,B.S., Pipelines,A. and Biolinux,C. Galaxy Workflow Composition Canvas. 3.

19. ToolShed Workflow Sharing.

20. NCBI - WWW Error 404 Diagnostic.

21. Pertea,M., Kim,D., Pertea,G.M., Leek,J.T. and Salzberg,S.L. (2016) Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat. Protoc., 11, 1650–1667.

22. Sam,T., Format,B.A.M. and Working,S. (2015) Sequence Alignment / Map Format Specification. SAM/BAM Format Specif. Work. Gr., 10.1016/j.ymeth.2012.07.021.

23. Pertea,M., Pertea,G.M., Antonescu,C.M., Chang,T.-C., Mendell,J.T. and Salzberg,S.L. (2015) StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol., 33, 290–295.

24. GTF2.2: A Gene Annotation Format.

25. GFF utilities.

26. Anders,S., Pyl,P.T. and Huber,W. (2015) HTSeq–a Python framework to work with high-throughput sequencing data. Bioinformatics, 31, 166–169.

27. Love,M.I., Huber,W. and Anders,S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol., 15, 550.

28. Li,H., Handsaker,B., Wysoker,A., Fennell,T., Ruan,J., Homer,N., Marth,G., Abecasis,G. and Durbin,R. (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics, 25, 2078–2079.

29. Koboldt,D.C., Zhang,Q., Larson,D.E., Shen,D., McLellan,M.D., Lin,L., Miller,C.A., Mardis,E.R., Ding,L. and Wilson,R.K. (2012) VarScan 2: Somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res., 22, 568–576.

30. Cingolani,P., Platts,A., Wang,L.L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly (Austin)., 6, 80–92.

31. Danecek,P., Auton,A., Abecasis,G., Albers,C.A., Banks,E., DePristo,M.A., Handsaker,R.E., Lunter,G., Marth,G.T., Sherry,S.T., et al. (2011) The variant call format and VCFtools. Bioinformatics, 27, 2156–2158.

32. Iyer,M.K., Chinnaiyan,A.M. and Maher,C.A. (2011) ChimeraScan: A tool for identifying chimeric transcription in sequencing data. Bioinformatics, 27, 2903–2904.

33. Wu,J., Akerman,M., Sun,S., McCombie,W.R., Krainer,A.R. and Zhang,M.Q. (2011) SpliceTrap: a method to quantify alternative splicing under single cellular conditions. Bioinformatics, 27, 3010–3016.

34. Rebhan,M., Chalifa-Caspi,V., Prilusky,J. and Lancet,D. (1997) GeneCards: integrating information about genes, proteins and diseases. Trends Genet., 13, 163.

35. Kent,W.J., Sugnet,C.W., Furey,T.S., Roskin,K.M., Pringle,T.H., Zahler,A.M. and Haussler, a. D. (2002) The Human Genome Browser at UCSC. Genome Res., 12, 996–1006.

36. NCBI Resource Coordinators (2017) Database Resources of the National Center for Biotechnology Information. Nucleic Acids Res., 45, D12–D17.

37. Szklarczyk,D., Franceschini,A., Wyder,S., Forslund,K., Heller,D., Huerta-Cepas,J., Simonovic,M., Roth,A., Santos,A., Tsafou,K.P., et al. (2015) STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res., 43, D447-52.

38. Wei,Q., Li,M., Fu,X., Tang,R., Na,Y., Jiang,M. and Li,Y. (2007) Global analysis of differentially expressed genes in androgen-independent prostate cancer. Prostate Cancer Prostatic Dis., 10, 167–174.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3385v1 | CC BY 4.0 Open Access | rec: 1 Nov 2017, publ: 1 Nov 2017

395 39. Savli,H., Szendröi,A., Romics,I. and Nagy,B. (2008) Gene network and canonical pathway
396 analysis in prostate cancer: a microarray study. Exp. Mol. Med., 40, 176.

PeerJ Preprints | https://doi.org/10.7287/peerj.preprints.3385v1 | CC BY 4.0 Open Access | rec: 1 Nov 2017, publ: 1 Nov 2017