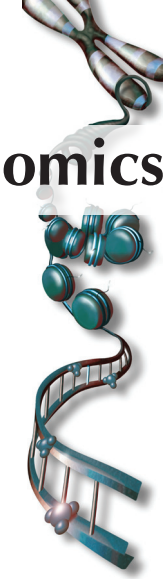
Methodology



Epigenomics



# scan_tcga tools for integrated epigenomic and transcriptomic analysis of tumor subgroups

**Aim:** The Cancer Genome Atlas contains multiple levels of genomic data (mutation, gene expression, DNA methylation, copy number variation) for 33 cancer types for almost 11,000 patients. However, a dearth of appropriate software tools makes it difficult for bench scientists to use these data effectively. **Materials & methods:** Here, we present a suite of flexible, fast and command line-based scripts that will allow retrieval and analysis of DNA methylation (tool: scan_tcga_methylation.awk), mRNA (tool: scan_tcga_mRNA.awk) and miRNA expression (tool: scan_tcga_miRNAs.awk) from cancer genome atlas network level 3 data. **Results:** We demonstrate the utility of these tools by analyzing DNA methylation and mRNA expression signatures of 60 frequently deregulated cancer genes and also of 30 miRNAs in primary (n = 102) and metastatic melanoma patients (n = 367). **Conclusion:** Our analysis illustrates the validity of the scan_tcga tools and reveals the epigenomic signatures and importance of identifying smaller patient subgroups with distinct molecular profiles.

First draft submitted: 21 May 2016; Accepted for publication: 25 July 2016; Published online: 14 September 2016

**Keywords:** cancer genome atlas • DNA methylation • gene expression • melanoma • miRNA

Aniruddha Chatterjee*[,‡,1,2], Peter A Stockwell[‡,3], Euan J Rodger[1,2], Matthew F Parry[4] & Michael R Eccles[1,2]

[1]Department of Pathology, Dunedin School of Medicine, University of Otago, 270 Great King Street, Dunedin 9054, New Zealand

[2]Maurice Wilkins Centre for Molecular Biodiscovery, level 2, 3A Symonds Street, Auckland, New Zealand

[3]Department of Biochemistry, University of Otago, 710 Cumberland Street, Dunedin 9054, New Zealand

[4]Department of Mathematics & Statistics, University of Otago, 710 Cumberland Street, Dunedin 9054, New Zealand

*Author for correspondence:
Tel.: +64 3 470 3455
aniruddha.chatterjee@otago.ac.nz

[‡]Authors contributed equally

Cancer is a group of complex diseases that arise from the accumulation of genomic, epigenomic and trancriptomic changes in cells [1]. Progressive advances in genomic and sequencing technologies have enabled the generation of multiple layers of -omics data at an unprecedented scale and rate [2] and have formed the basis of large-scale cancer genomic projects. As a result, there has been an explosion of genomic/epigenomic data for a number of human cancers. The efforts of international consortia have played a major role in the acquisition and generation of a plethora of cancer genome datasets. The Cancer Genome Atlas (TCGA) [3,4] is probably the most comprehensive initiative that has generated gene mutation, DNA methylation, mRNA, miRNA, protein and clinical information of more than 11,000 patients and comprising 33 human cancer types. Therefore, these datasets provide great opportunities for obtaining a multilayered view of cancer genomes and untangling the complexity of genomic landscapes in human cancers. It is exciting that these data are now available to explore cancer genomes at an unprecedented scale. However, the massive volume, multiple dimensions and varied formats of these datasets can be overwhelming [5] and they create substantial challenges for biologists to analyze, integrate and interpret.

The default gateway to download TCGA cancer datasets is the Data Portal [6], where different levels of data (level 1: raw, level 2: semi processed and level 3: processed data) can be downloaded. To analyze these datasets, appropriate tools and scripts are necessary. Initially, the main focus of cancer genome projects was to identify somatic mutations and therefore several tools were developed (e.g., COSMIC [7], Tumorscape [8], IntOGen [9,10] Oncoprint [11]) to

Future Medicine part of fsg

analyze and provide summary information of somatic alterations in cancers. However, tools for analyzing epigenomic and transcriptomic level information are very limited. Currently, a small number of web-based tools (such as cBio [11,12], Wanderer [13], canEvolve [14] Web-TCGA [15] and TCGA compass [16] are available to access TCGA data for further analysis. However, these tools have several limitations with regard to flexible and systematic analysis of epigenomic data. These tools provide an overview and present aggregated data for a whole cancer type. The major problem with these web-based tools is the lack of flexibility to analyze a subgroup of patients. For example, analysis of metastatic patients or epigenomic analysis of patients harboring a particular mutation type is not possible with these tools. Furthermore, the users are restricted to the analysis options provided in the tools and customized downstream analysis is not feasible with current tools. In addition, many of these web-based tools are based on curated databases and often it is not possible to access information recently released by TCGA. More recently, a new tool TCGAbiolinks that uses R statistical environment was made available [17], which will be an useful tool to the community. However, using all the different modules of the tools still requires expertise with R programming. Furthermore, the tool is designed for bulk analysis is less efficient if only a small number of genes or regions are intended to be investigated. It is possible to analyze subgroups in TCGAbiolinks; however, the groups are mainly predefined and the user has less flexibility in choosing or making their own subgroups for analysis. Recent research has elucidated substantial heterogeneity in tumor genomes and epigenomes [18] and the importance of investigating cancer subgroups is becoming more evident [19]. Therefore, it is crucial to have tools and analysis pipelines that allow interrogation of tumor subgroups at multiple levels.

Here we present a suite of flexible, fast, command line based tools that allows retrieval and analysis of DNA methylation (program: scan_tcga_methylation.awk), mRNA (program: scan_tcga_mRNA.awk) and miRNA expression (program: scan_tcga_miRNAs.awk) from TCGA data. Using these tools, it is possible to analyze TCGA data on simple desktop computers and obtain raw or processed values for further downstream analysis. We demonstrate the utility of these tools in retrieving meaningful biologically tenable data by analyzing DNA methylation and mRNA expression signatures of 60 frequently deregulated cancer genes and expression of 30 miRNAs in primary and metastatic melanoma patients. Our analysis reveals distinct epigenomic signatures in melanoma and provides evidence for the utility of these tools.

## Materials & methods

### Obtaining TCGA data

The TCGA data was downloaded using the Data Matrix interface of the TCGA Data Portal [6], which has now been moved to NCI's Genomic Data Commons [20]. TCGA level 3 datasets were obtained for both DNA methylation and RNA-Seq. The initial data matrix page provides easy selection options for choosing disease type, data type (methylation, expression, and mutation, among others), batch numbers (we downloaded all the batches associated with a data type) and sample preservation (FFPE, frozen or all). After choosing the described selections, the next page provides a matrix. All the desired samples were selected and an archive was built (this button will appear on top of the matrix). The third page will then require an email address and all these selected data can be obtained in compressed form (for faster download). The downloaded compressed folder can be unzipped using a command line script such as: gzip -dc tcgadownload.tar.gz | tar -xvf - or -dc route_to_the_downloaded_directory/tcgadownload.tar.gz | tar–xvf.

### Format of downloaded data

TCGA datasets broadly contain similar types of files or data structures. The uncompressed data folder will contain metadata (description of platform details, assay details, barcode, protocol reference and similar technical aspects that were used to generate the datasets), FILE_SAMPLE_MAP (.txt file containing file name and TCGA barcodes of the samples included in the datasets), file_manifest (.txt file containing barcode file name of the samples, center, platform details) and the level 3 data for a particular experiment (methylation, expression, among others) in a separate folder.

### TCGA barcode description

It is important to understand the barcoding system of TCGA in order to demultiplex samples for downstream analysis. A detailed description of a sample barcode is described in the TCGA website [21].

### Obtaining a list of sample barcodes for subsets of cancer patients

The complete barcode for each patient will differ due to the different analyte (DNA, RNA, protein), portions and vials used for analysis. The scripts described here accept any valid TCGA barcode for particular analysis. An easy and convenient way to extract the complete barcode of samples for different experiments is from the Broad Institute's GDAC resource [22]. We obtained the sample barcodes for solid normal tissue, primary and metastatic tumors for the SKCM dataset (TCGA data for SKCM) from the following source [23].

Similarly, just by clicking each hyperlink (browse samples option) it is possible to obtain sample barcodes for normal tissues, primary or metastatic tumors for each different type of experiment (methylation, miRSeq and RNA-Seq, among others). This information for each desired subgroup is separately copied into a separate tab delimited text file.

### Processing & analysis of DNA methylation data

For analyzing DNA methylation data using the scan_tcga_methylation.awk program, two input files were provided (Figure 1). The first one contained the TCGA barcodes of the group to be analyzed in a tab delimited text format (e.g., for SKCM primary tumor methylation SKCM_meth_barcode_primary.txt was used, these files are available with Supplementary data files 1, 2, 3 & 4). The second input file contained the list of regions to be analyzed (the file columns are: chromosome, start, end and gene name. Example input file, meth_inputregions_list.txt can be found in Supplementary data files 1, 2, 3 & 4). Using these two input files, scan_tcga_methylation.awk extracts DNA methylation information for the list of given regions for a given sample group (primary, metastatic or any other specified group). The output file is a large matrix with the first four columns being the same as the user provided and the other columns are the beta methylation values of the patients. In cases where multiple CpG sites are present in the region, mean beta methylation values of all the CpG sites for that region are provided as an output. If only one CpG site is to be investigated then the start and end coordinates should have a distance of 1 bp. scan_tcga_methylation_awk also provides an option of extending the input regions by any length (both upstream and downstream) by specifying margin and number of base pair to be extended in the command line. Furthermore, example test datasets along with scripts could be obtained from our GitHub repository [24].

### Processing & analyzing RNA-SeqV2 data

For this analysis, three input files are needed. The first file contains TCGA barcodes for the subset of samples to be analyzed (e.g., SKCM_mRNA_barcode_primary.txt, barcode files used in this analysis can be found in Supplementary data files 5, 6, 7 & 8). Barcodes from these files were used along with the FILE_SAMPLE_MAP.txt file (this file comes with TCGA level 3 expression data download) to locate appropriate matching samples in the RNA-Seq data folder. Unlike DNA methylation data, for RNA-Seq level 3 expression data, the name of the files (navigate from UNC__IlluminaHiSeq_RNASeqV2 to Level_3 to the files for each sample) does not contain TCGA sample barcodes. Furthermore, for each sample six different types of files are provided in TCGA RNASeqV2 datasets. Therefore this additional step of using FILE_SAMPLE_MAP is required to locate corresponding files for the user provided barcode and match against the correct expression data file to extract relevant information. Second, an input list of genes is required for which expression will be measured (file name: mRNA_inputgenes_list.txt in Supplementary data files 5, 6, 7 & 8). Using these input files, scan_tcga_mRNAs.awk extracts mRNA expression information for any given number of barcodes. scan_tcga_mRNAs.awk is able to extract raw read counts (i.e., raw_count, second column) and RSEM scaled estimate of a transcript (i.e., scaled_estimate, third column) for the input list of genes from .rsem.genes.results.txt files. Furthermore, the program also can extract normalized read counts for a gene (normalized_count, second column) from .rsem.genes.normalized_results.txt files. A point to be noted here is that the scaled transcript or normalized count for a gene as provided in TCGA level 3 data often contains an aggregate of counts from multiple different transcripts for a gene. Therefore, in many cases these values represent the sum of total expression for a gene but not expression profiles of individual transcripts. The users of scan_tcga_mRNAs.awk need to provide the option (wanted_field switch in the command line) of which information (or which field of data) is needed from level 3 data. As the output file is a large matrix with the first column being the provided gene names and the other columns being the expression value of the patients only one type of information can be retrieved at one time. To return $\log_2$ of the wanted field $\log_2 = 1$ can be specified in command line (default is $\log_2 = 0$. Detailed documentation and example commands can be found in in Supplementary data files 5, 6, 7 & 8). Furthermore, example test datasets along with scripts could be obtained from our GitHub repository [25].

### Processing & analyzing miRNA-Seq data

Analyzing sequencing based miRNA data using scan_tcga_miRNAs.awk requires two input files. The first containing the TCGA barcodes of the group to be analyzed in a tab delimited text format similar to those of methylation and expression analysis (e.g., SKCM_miRNA_barcode_primary.txt was used, these files are available with Supplementary data files 9, 10, 11 & 12. Second, an input list of miRNAs is required for which expression will be measured (file name: miRNA_input_list.txt in Supplementary data files 9, 10, 11 & 12). By default, scan_tcga_miRNAs.awk provides reads per million (RPM) values for a miRNA as RPM is the most accepted analysis unit for miRNA expression (i.e., mirna.quantification.txt files, third column).

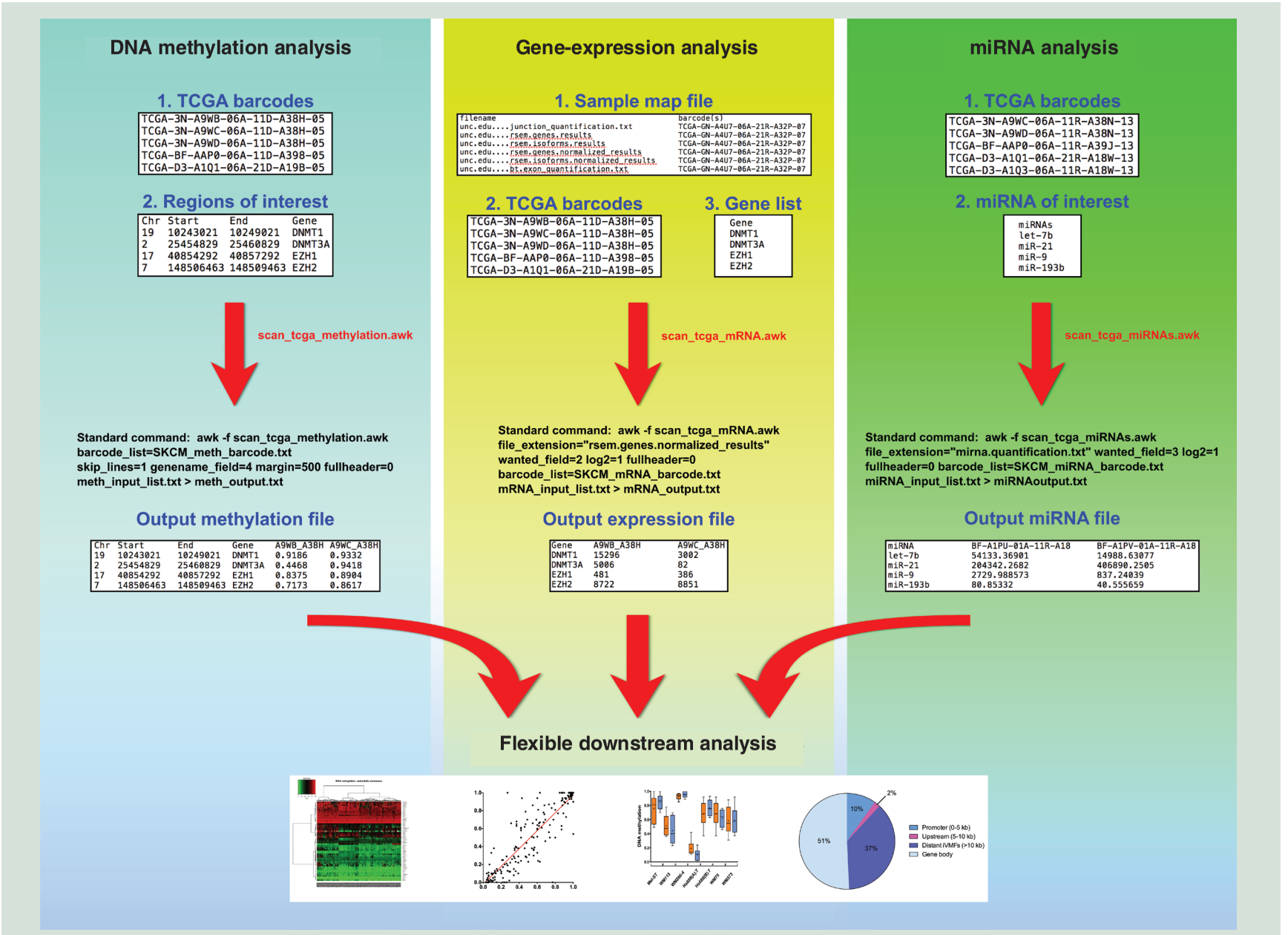



**Figure 1. Workflow for analyzing DNA methylation, gene expression and miRNA The Cancer Genome Atlas data.** For each type of analysis, the workflow requires tab-delimited text files of TCGA barcodes, a list of DNA regions/genes/miRNA of interest and a sample map file (for gene expression analysis). Running an awk script for each analysis set generates tab-delimited text output files that can be integrated and used for flexible downstream analysis.
TCGA: The Cancer Genome Atlas.

However, the program provides flexibility of returning other fields from datafiles and can be specified using the wanted_field option. To return $\log_2$ of the wanted field, $\log_2 = 1$ can be specified in command line (default is $\log_2 = 0$. In TCGA datasets, a constant header is present in miRNA identifiers and may not be present in the list to be provided by the user. For example, 'hsa-let-7b' will be found for an input of 'let-7b' with the default setting of mirprefix (='hsa-'). If identifiers are not to be prefixed in this way, mirprefix="" can be used in the command line. Furthermore, example test datasets along with scripts could be obtained from our GitHub repository [26].

### Downstream analysis

We obtained the matrix file output provided by the scan_tcga programs described here in text format and analyzed the data using standard operations and statistical tests in the R Studio environment (version 3.1.1) using standard R commands. The heatmaps shown are plotted using heatmap.2 function in R. All the analysis described here could be performed using publicly available R packages and scripts and can be obtained from the authors on request.

## Result & discussion

### scan_tcga programs for epigenomic analysis of sub groups within a cancer type

We have developed three independent programs to investigate DNA methylation, mRNA expression and miRNA expression in patient subgroups (e.g., primary, metastatic, patients of particular cancer stage, or molecular subgroups of patients identified independently).

### scan_tcga_methylation.awk

For analyzing DNA methylation data for any region in the genome we developed the scan_tcga_methylation.awk program, which requires TCGA barcodes for samples and a list consisting of genomic coordinates of regions to be analyzed. This program provides methylation profiles for all the patients in a matrix in text format for easy investigation and downstream analysis. The region file requires a simple input of chromosome, start and end position and gene name (Figure 1). The program does not use the gene name information to retrieve methylation as the gene name supplied by the user could differ based on the annotation used. If multiple CpG sites are present within a given region, the scan_tcga_methylation.awk program provides an average methylation status of a given region. In addition, it is possible to investigate methylation patterns of the adjacent regions by specifying a margin in the command (i.e., how many additional base-pairs to be analyzed up or downstream).

### scan_tcga_mRNA.awk

For mRNA expression, we have developed scan_tcga_mRNA.awk that uses patient barcodes, a list of gene names and FILE_SAMPLE_MAP.txt file (from TCGA level 3 expression data) to provide either raw read counts or normalized counts for each patient in matrix format from RNASeqV2 data (Figure 1). For mRNA expression it is necessary to use the SAMPLE_MAP file since in TCGA RNA-Seq data, multiple files are provided for each patient. The map file helps to locate the barcode, the required file and allows the desired expression output to be retrieved. The scan_tcga_mRNA.awk also has the option of providing $\log_2$ of the expression values (e.g., normalized count), which is often required for downstream analysis.

### scan_tcga_miRNAs.awk

Similarly, for miRNA analysis we have developed scan_tcga_miRNAs.awk that returns RPM using a list of miRNA and respective barcodes (Figure 1). The program is also able to provide $\log_2$ of the RPM values.

### Implementation, speed & availability

The scripts are written in awk [27], a text processing language which is an integral part of all Unix-type operating systems. These include Linux dialects and MacOS X. The scripts have been developed and run under MacOS X 10.8–10.10 (Yosemite), but should function identically in other environments. All input files should have normal Unix line terminators ('\n'). Files originating from MicroSoft Excel or other sources may need pretreatment to correct for this (see 'Materials & methods' section). Each of the scan_tcga programs is able to retrieve the full TCGA barcode (fullheader = 1 option) or a four character unique patient identification (ID) as the column headers in the output. Also, for diagnostic purposes, it is possible to generate a list of any genes and files which are not found in the scan_tcga programs specifying unseen = 1. These programs can be used in standard desktop or laptop with an UNIX programming environment (e.g., the Terminal application in MacOS X).

The scan_tcga programs provide fast operation and data retrieval, for example, mRNA expression profiles for 102 primary and 358 metastatic patients (60 analyzed genes) were returned within 6.05 and 21.65 min, respectively (Table 1). Similarly, expression profiles of 30 miRNA in primary and metastatic patients' melanoma were obtained in 0.99 and 3.8 min, respectively. DNA methylation data for 102 patients (for 60 gene promoters) were obtained in 303.4 min (Table 1). The data retrieval process for DNA methylation requires additional steps (such as mapping each CpG sites within a region and then providing mean methylation for a region) and therefore, the mRNA and miRNA pro-

Table 1. Speed of operations of the scan_tcga programs.

| Operation | Program | Group analyzed | CPU time (s) |
|---|---|---|---|
| Retrieved promoter DNA methylation of 60 genes | scan_tcga_methylation.awk | Normal tissue (n = 2) | 728 |
| | | Primary (n = 102) | 18,204 |
| | | Metastatic (n = 367) | 65,336 |
| Retrieved $log_2$ of normalized expression count of 60 genes | scan_tcga_mRNAs.awk | Normal tissue (n = 1) | 3.5 |
| | | Primary (n = 102) | 363 |
| | | Metastatic n = 367) | 1299 |
| Retrieved $log_2$ of normalized expression count of 30 miRNAs | scan_tcga_miRNAs.awk | Normal tissue (n = 2) | 1 |
| | | Primary (n = 102) | 59 |
| | | Metastatic n = 367) | 228 |

The configuration of the computer used here is: operating system: MacOS 10.10.6. Dual Quad core Xeon processors, 32 GB RAM. The files were retrieved from our local storage server to process using scan_tcga program.
CPU: Central processing unit.

grams are relatively faster than scanning methylation. Furthermore, we have downloaded the TCGA level 3 data onto a server and accessed these from the server to generate the results. Storing in local hard disk will significantly improve the speed of scan_tcga programs. Large candidate gene analysis (hundreds or thousands of genes) using scan_tcga programs could be performed using local desktop computers. However, for obvious reasons, the run time will increase with a higher number of genes or patients (as the program is performing matching tasks for every gene and for every patient present in the dataset). An alternative option could be to perform these large operations in a highly configured computer or server to get the text format output from scan_tcga tools. These text outputs are relatively much smaller in size and downstream analysis could be done on these files without high computing power.

The scan_tcga programs are publicly available as GitHub repository. Following are the GitHub links for each of the tools, which consist of the program, detailed documentation and examples and a test dataset for repository. We recommend performing trial analysis with the test dataset and instructions to familiarize with the commands of the operations:

- scan_tcga_methylation [24];
- scan_tcga_mRNA [25];
- scan_tcga_miRNA [26].

## Demonstration of usage

### Methylation & mRNA expression analysis of 60 frequently deregulated genes in primary & metastatic melanoma

To demonstrate the utility of the tools for retrieving useful biological information, we set out to perform epigenomic analysis on 61 genes that have been curated based on the published literature and their involvement in defining epigenetic machinery and their implication in cancer. These consisted of genes involved in methylating machinery (DNA methyl transferase or DNMT family) demethylating (TET family proteins), the EMT (epithelial to mesenchymal transition) process and genes that were previously reported to be frequently deregulated (genetic mutation and/or aberrant methylation profiles) in cancer (list of genes used is shown in Supplementary Table 1). Out of these 61 genes, *MAGEA3* did not contain any analyzable CpGs in the promoter in TCGA-SKCM 450 K data. Therefore, it was excluded from the methylation analysis. We carried out analysis of promoter methylation and corresponding mRNA expression profiles for these 60 genes in primary (102 patients) and metastatic melanoma (367 patients) from TCGA data for skin cutaneous melanoma (TCGA-SKCM). Our goal was to analyze primary and metastatic melanoma as two groups and identify significant differences between these groups to demonstrate that biologically tenable results could be obtained using the suite of tools described in this article. We also present data for normal skin tissue from the TCGA-SKCM project.

### DNA methylation landscape of frequently deregulated genes in melanoma

Gene promoters were defined as regions within -5 kb to +1 kb from the transcription start site. For the genes that had multiple transcripts, we determined the promoter region based on the genomic coordinates of the main expressed transcript. The average promoter methylation status of normal skin tissue (n = 2), primary and metastatic melanoma is shown in Figure 2. Overall, the promoters of these genes were unmethylated in normal skin tissue as expected. However,

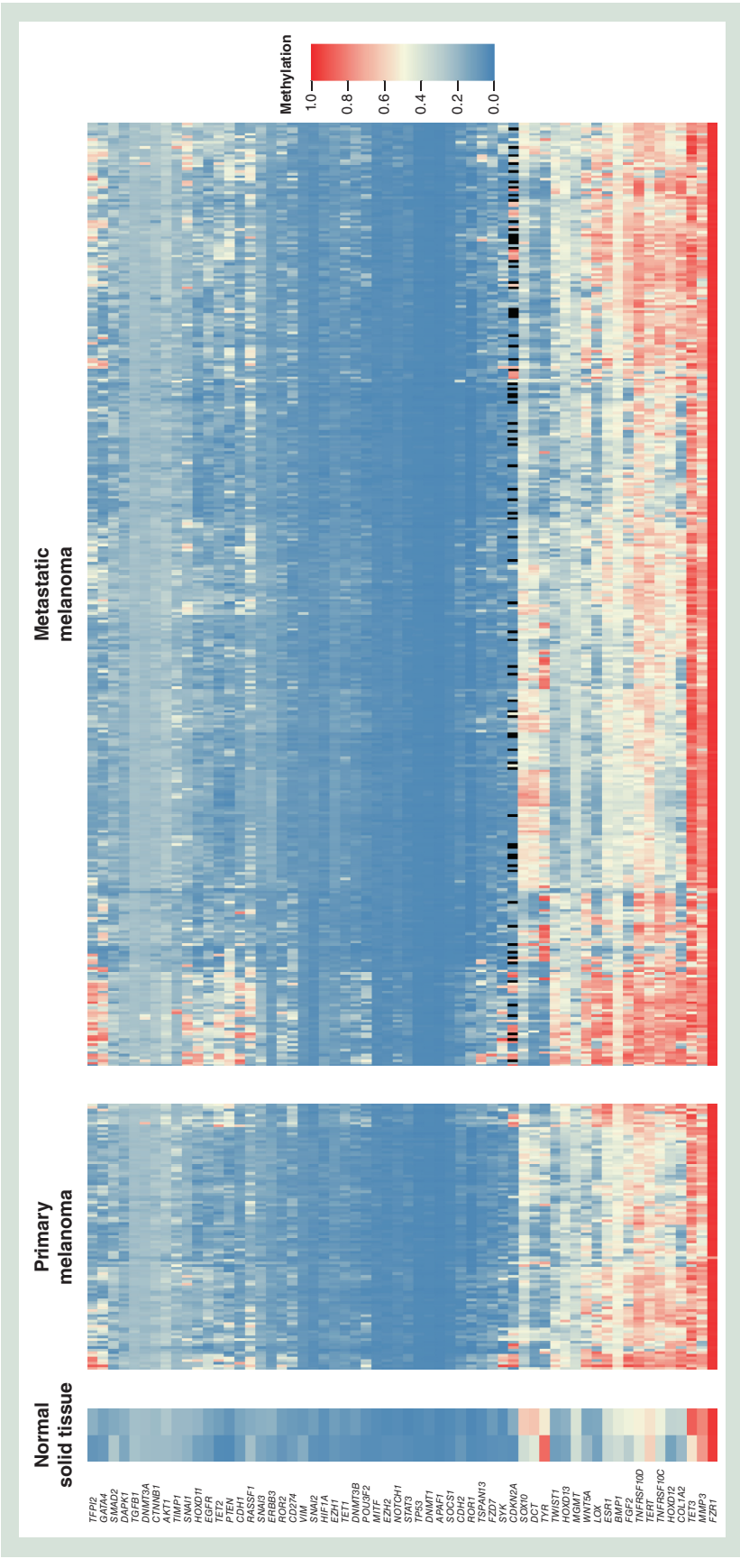



**Figure 2. DNA methylation profiles of normal skin tissue, primary and metastatic melanoma.** Heatmaps show promoter methylation (-5 kb to +1 kb from the transcription start site) of 60 frequently deregulated cancer genes in normal skin tissue (n = 2), primary (n = 102) and metastatic melanoma tumors (n = 367) retrieved from The Cancer Genome Atlas. Low (=0) to high (=1) methylation is shown as a continuous variable from a blue to red color (black = not detected). *MAGEA3* does not contain any analyzable CpGs in the promoter in The Cancer Genome Atlas-SKCM 450 K data. Therefore, it was excluded from the methylation analysis.

several genes showed high methylation level in normal skin tissue. These genes are *FZR1*, *MMP3* and *TET3*, *TERT*, *TYR*, *DCT* and *SOX10* (mean methylation >0.50). However, these observations should be interpreted with caution, as only two normal skin tissues were included.

The primary and metastatic melanoma patients show relatively more heterogeneous patterns of methylation (Figure 2) compared with normal skin tissue. Previously, *TNFSF10D*, *LOX* and *COL1A2* were reported to be highly methylated in melanoma [28,29]; our analysis confirmed promoter hypermethylation of these genes. *PTEN* has been reported to be methylated in ~60% of the melanomas [30,31]. We found promoter hypermethylation of *PTEN* in a relatively small proportion of primary and metastatic patients. Similarly, we found a low level of methylation in the *SYK* gene promoter, which was previously reported to be methylated in 30% of melanomas [29]. Our analysis also identified hypermethylation in the *HOXD12*, *TNFRSF10C*, *FGFR2* and *TERT* genes. These results confirm recent 450 K array methylation analysis (the same platform as used in TCGA methylation assays) that reported high levels of promoter methylation in these genes in melanoma [32,33].

Next, we performed differential methylation analysis between primary and metastatic samples and identified four genes that were significantly differentially methylated between the two groups (Wilcoxon Rank test, Bonferroni adjusted p-value < 0.00083). These genes are *CDH1* (median methylation = 11.4 and 13.8%, respectively, for primary and metastatic patients), *EZH2*, *NOTCH* and *TET3* (Figure 3 & Supplementary Table 2). Metastatic patients showed significant hypermethylation in all four genes compared with primary. Although statistically different, the majority of primary and metastatic patients showed similar levels of methylation in these genes. However,

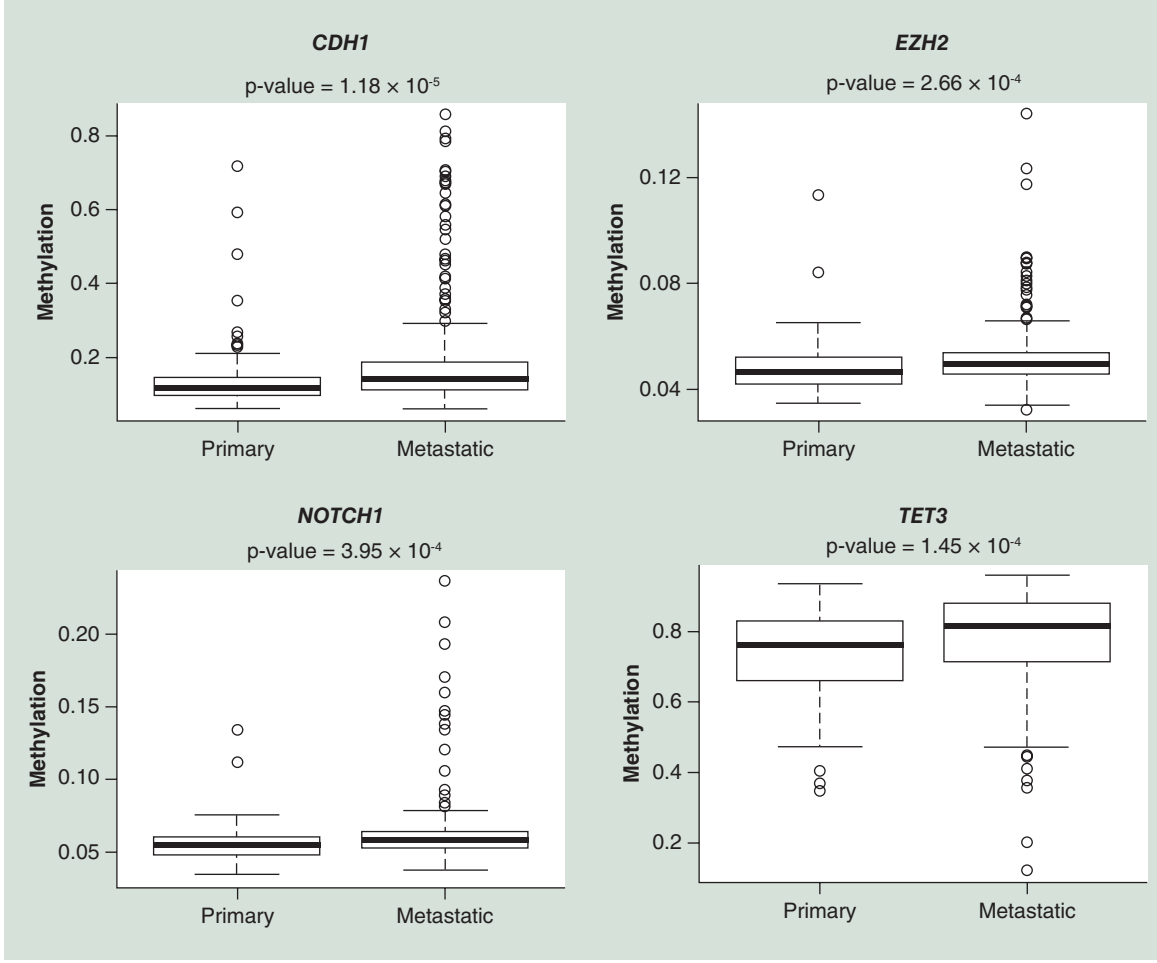



**Figure 3. Box plots representing DNA methylation profiles of four genes that showed significant differential methylation between primary and metastatic melanoma patients.** Primary patient (n = 99) and metastatic patients (n = 359). Y-axis represents DNA methylation in 0 (0%) to 1 (100%) scale. Statistical significance was derived using Wilcoxon Rank test followed by Bonferroni adjustment for multiple test correction at a significance level of 0.05 (i.e., p-value < 0.00083)

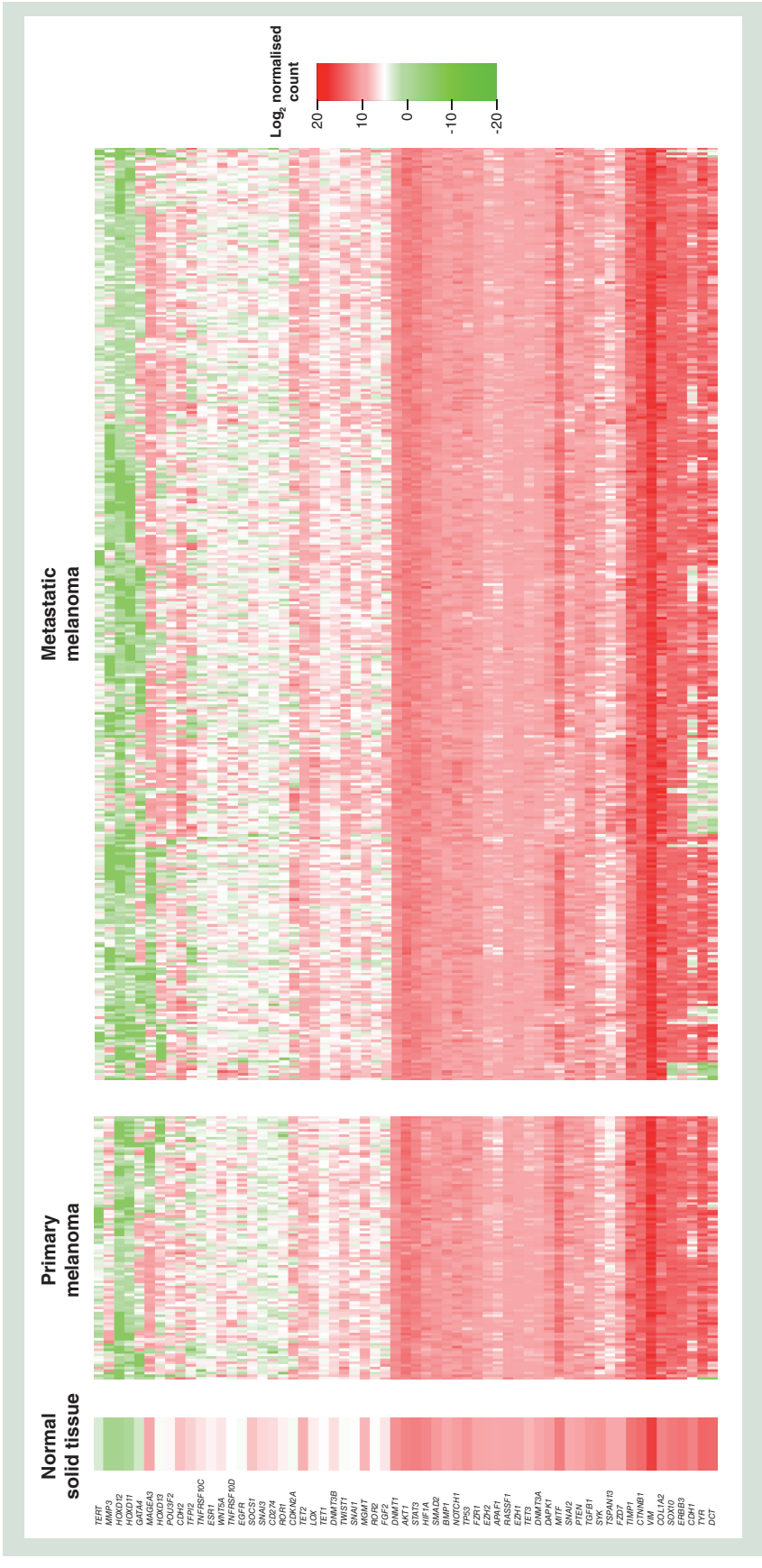



**Figure 4. Gene expression (mRNA) profiles of normal skin tissue, primary and metastatic melanoma.** Heatmaps show expression of 61 frequently deregulated cancer genes in normal skin tissue (n = 1), primary (n = 102) and metastatic melanoma tumors (n = 367) retrieved from The Cancer Genome Atlas. $Log_2$ normalized counts of low (= -20) to high (= 20) gene expression are shown as a continuous variable from a green to red color. Although *MAGEA3* was excluded from methylation and methylation–mRNA relationship analysis, we have shown the mRNA expression profile of the gene in this figure.

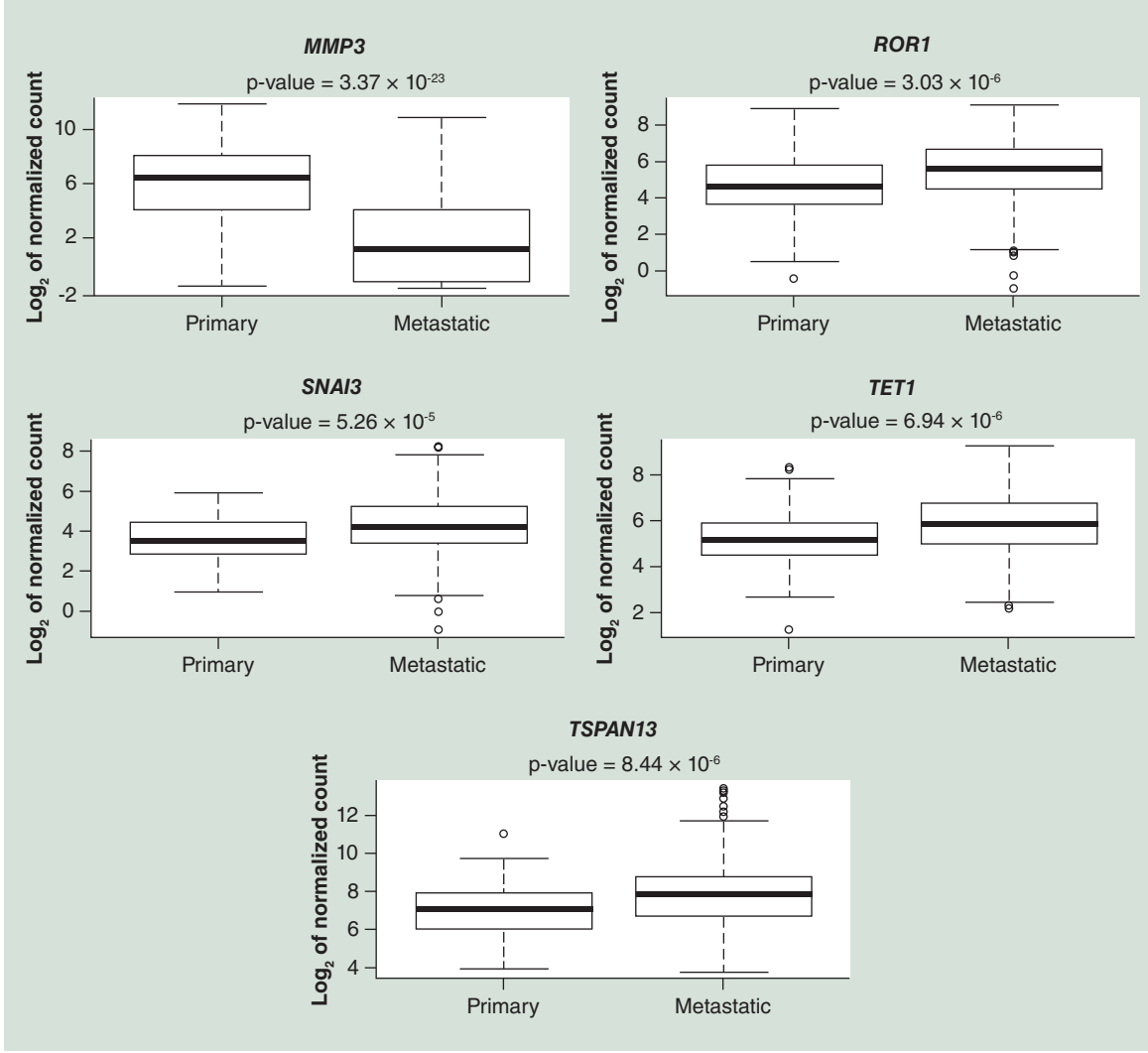



**Figure 5. Box plots representing mRNA expression profiles of five genes that showed significant differential expression between primary and metastatic melanoma patients.** Primary patient (n = 99) and metastatic patients (n = 359). Y-axis represents $\log_2$ of the normalized counts for mRNA expression. Statistical significance was derived using Wilcoxon Rank test followed by Bonferroni adjustment for multiple test correction at a significance level of 0.05. These genes showed a fold difference of 1.5 or higher in their normalized read counts between primary and metastatic patients.

a small subgroup of patients showed a strikingly different methylation pattern, giving rise to the overall difference. Principal component analysis (PCA) further suggested that overall the methylation patterns between primary and metastatic patients are similar in these 60 genes (Supplementary Figure 1).

## Gene expression profiles of primary & metastatic melanoma patients

In TCGA, for RNA-Seq experiments, six files are provided for each sample. The scan_tcga_mRNA.awk program is able to retrieve normalized counts or raw counts for a gene for any given sample (see methods for details). For quantification of the mRNA expression and differential expression analysis, the most relevant information is the normalized count of expression for a gene (this is the default output option for scan_tcga_mRNA.awk). The normalized count of expression for a gene does not provide expression profiles of individual transcripts. These values are representative of the expression of the gene as a whole (i.e., a combination of different transcripts). The $\log_2$ normalized expression level of the analyzed genes for normal skin tissue, primary and metastatic melanoma are shown in Figure 4 (although *MAGEA3* was excluded from methylation and methylation–mRNA relationship analysis, we have shown the mRNA expression profile of the gene in this figure). These data demonstrate mRNA expression patterns between primary and metastatic melanoma patients are more variable than DNA methyla-

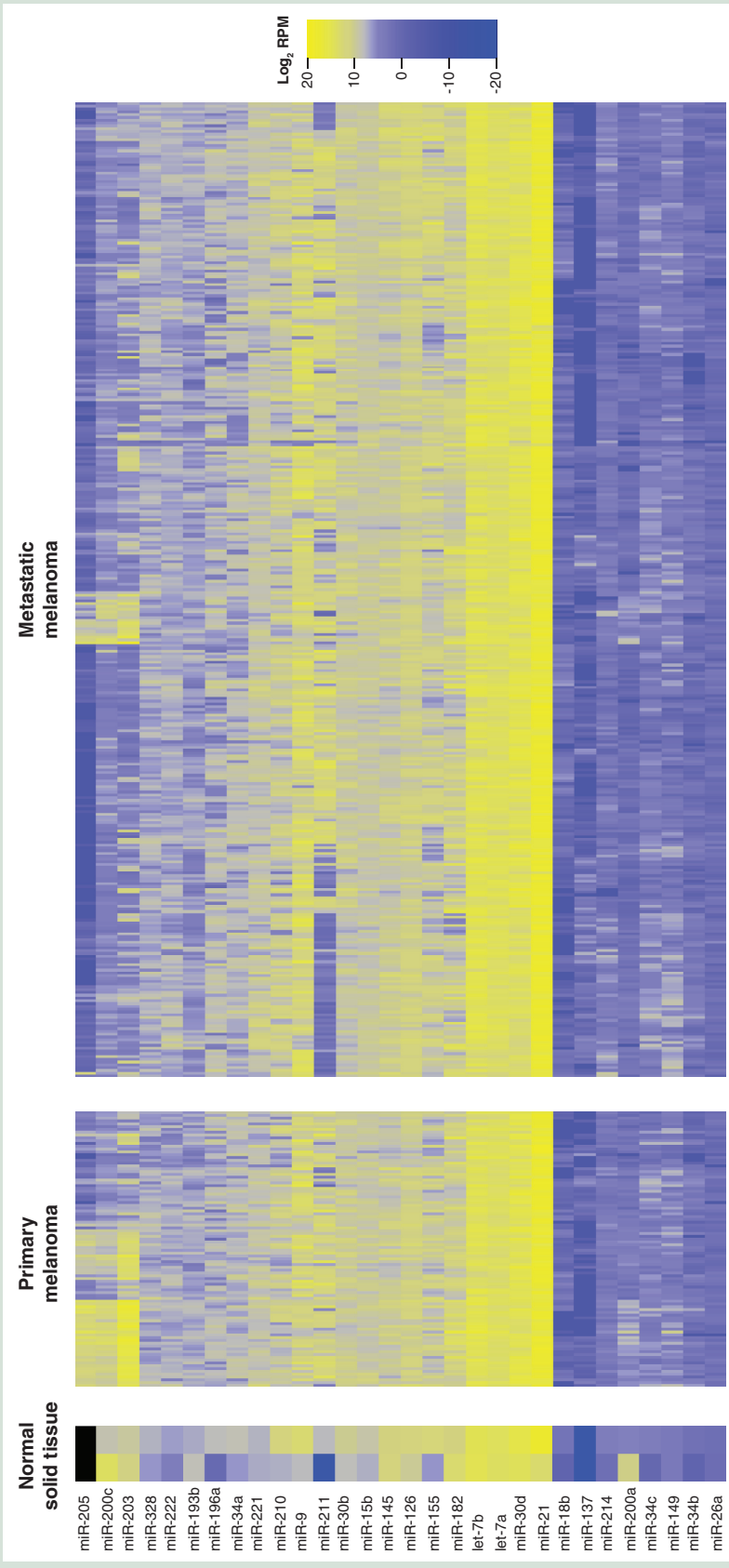



**Figure 6. miRNA expression profiles of normal skin tissue, primary and metastatic melanoma.** Heatmaps show expression of 30 frequently deregulated miRNA in normal skin tissue (n = 2), primary (n = 102) and metastatic melanoma tumors (n = 367) retrieved from The Cancer Genome Atlas. $Log_2$ RPM of low (= -20) to high (= 20) miRNA expression are shown as a continuous variable from a blue to yellow color (black = not detected).
RPM: Reads per million.

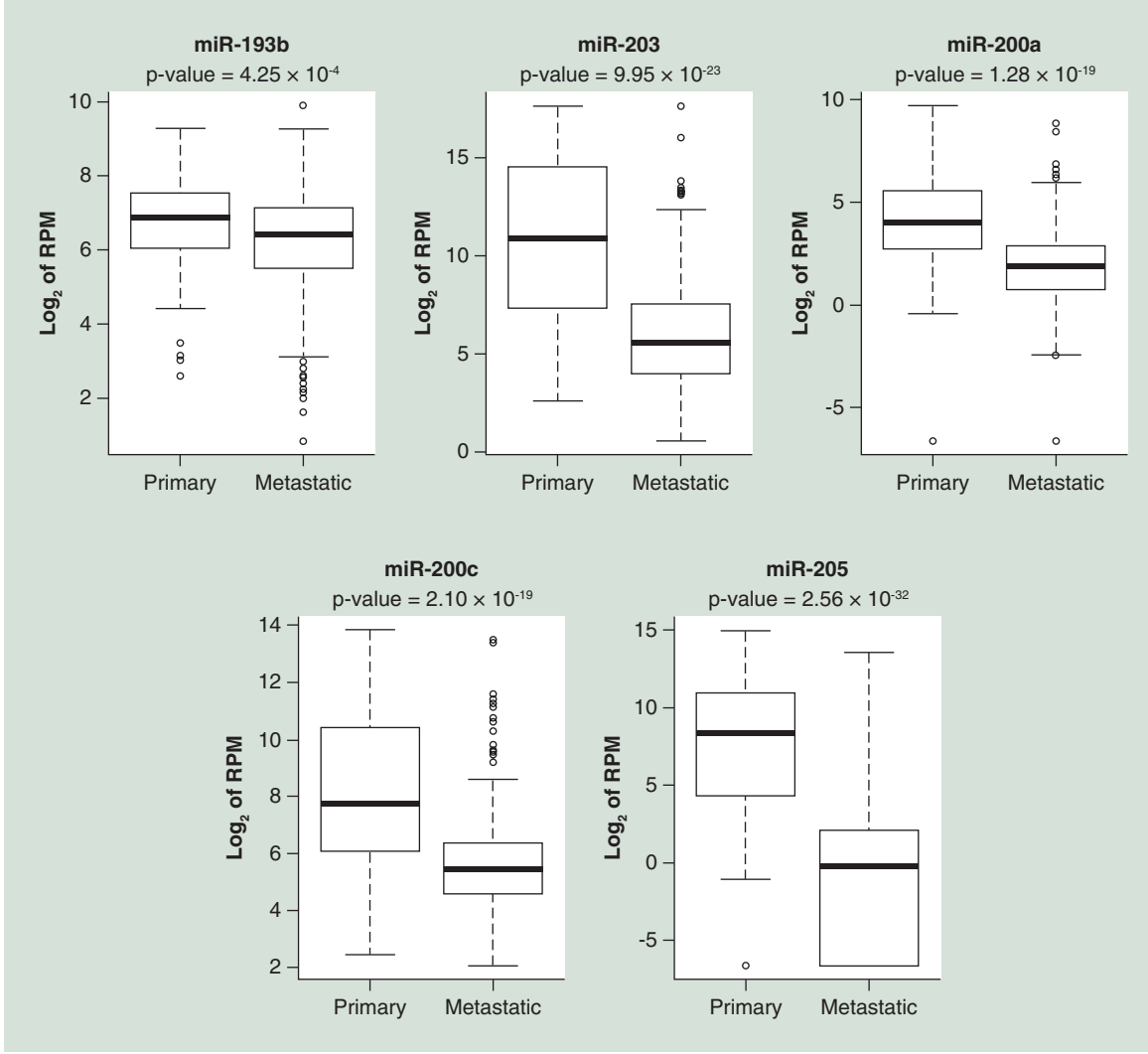



**Figure 7. Box plots representing miRNA expression profiles for five miRNAs that showed significant differential expression between primary and metastatic melanoma patients.** Primary patient (n = 99) and metastatic patients (n = 359). Y-axis represents $\log_2$ of the RPM. Statistical significance was derived using Wilcoxon rank test followed by Bonferroni adjustment for multiple test correction. The adjusted p-value was (0.05/30) = 0.00167 (30 is the number of investigated miRNA).
RPM: Reads per million.

tion patterns between the two groups. PCA confirmed this observation (Supplementary Figures 1 & 2).

When we performed differential expression analysis, we found 18 genes that showed statistically significant difference (adjusted p-value cut off < 0.00082) between primary and metastatic melanomas (Supplementary Figure 3, 4 & Supplementary Table 3). However, when we applied an expression fold difference cut of ≥1.5 on these genes, five genes met the criteria. Out of these five genes, *ROR1*, *SNAI3*, *TET1*, *TSPAN13* were significantly upregulated in metastatic compared with primary (Figure 5). Upregulation of *ROR1* in melanoma was reported recently [34], while overexpression of *TSPAN13* was reported in other cancers [35] and *SNAI3* involved in the EMT process, however, its role in melanoma is not well documented yet [36]. Intriguingly, loss of 5-hydroxymethylcytosine and downregulation of *TET1* in melanoma compared with nevi was reported as a key feature of melanoma [37]. However, our analysis reveals that in TCGA patients, *TET1* is upregulated in tumor progression (i.e., metastasis) contradictory to the previous finding (Figure 5). *MMP3* showed a striking downregulation in metastatic tumors ($\log^2$ of the median read count = 1.15) compared with primary ($\log^2$ of the median read count = 6.46). Matrix metalloproteinases (MMPs) family mediates gene expression changes and plays a role in modulating tumor microenvironment during tumor progression [38] and study of specific genes in this family is an active area of research. Downregulation of *MMP3* expression

and subsequent suppression of tumor metastasis was reported in esophageal squamous cell carcinoma [39]. Role of *MMP3* in suppressing tumor invasion and metastasis was also reported in *in vivo* model [40]. However, its specific role in the metastatic cascade in melanoma is not known and our results calls for further investigation of its function in melanoma. Our results demonstrate substantial heterogeneity in gene expression profiles in melanoma. Although overall statistical significance is reached (due to the high number of patients analyzed here, improving the power for detecting significant differences) for several genes, a small sub group of patients demonstrates large differences. This indicates that within primary or metastatic melanoma groups there are small subgroups that demonstrate distinct profiles for certain genes. These results call for further work to identify small subgroups of patients (within melanoma primary or metastatic groups) that demonstrate distinct profiles in certain genes.

### Relationship of promoter DNA methylation & mRNA expression in frequently deregulated genes in primary & metastatic melanoma

Next, we assessed the link between promoter DNA methylation and mRNA expression in the analyzed genes. When we performed an aggregated analysis of all the genes, we found the overall correlation between DNA methylation and expression of the associated gene was negative. This association was stronger for metastatic tumors (Spearman correlation of rho = -0.17; p-value < $2.2 \times 10^{-16}$, Supplementary Figure 5) compared with the primary tumors (Spearman correlation of rho = -0.122; p-value < $2.2 \times 10^{-16}$, Supplementary Figure 6). This is in concordance with the general perception that highly methylated genes are expressed at low levels and therefore negatively correlated, and vice versa [41]. Next, we analyzed relationship of mRNA expression and promoter methylation for individual genes. This analysis revealed 23 genes that showed significant negative correlation (after multiple test correction using bonferroni correction at a significance level of 0.05 for the combined data of primary and metastatic melanoma patients) with methylation and corresponding mRNA expression. These genes are (ranked as from low to high p-values): *FGF2*, *TNFRSF10D*, *CD274* (or *PD-L1*), *TFPI2*, *CDH1*, *COL1A2*, *SNAI1*, *ERBB3*, *TYR*, *SOX10*, *LOX*, *VIM*, *DCT*, *RASSF1*, *TSPAN13*, *SNAI2*, *CDH2*, *FZD7*, *CDKN2A*, *TET2*, *DNMT3B*, *SYK*, *TET1*. Similar to the aggregated analysis, we found that even at an individual gene level, these negative correlations are stronger for metastatic melanoma patients compared with the primary. The correlation analyzes of the individual genes are presented in Supplementary Table 4 and the relationship plots are depicted in Supplementary Figures 7, 8 & 9.

However, although this was an overall pattern, several genes behaved differently. Surprisingly, we identified significant positive correlation (i.e., high methylation in promoters are associated with high mRNA expression) for five genes (after multiple test correction using bonferroni correction at a significance level of 0.05 for the combined data of primary and metastatic melanoma patients). These genes are *GATA4* (rho = 0.45; p-value = $5.55 \times 10^{-24}$), *HOXD12* (rho = 0.42; p-value = $6.59 \times 10^{-21}$), *ESR1* (rho = 0.36; p-value = $3.84 \times 10^{-15}$), *TWIST1* (rho = 0.32, p-value = $3.35 \times 10^{-12}$), *MGMT* (rho = 0.28, p-value = $7.99 \times 10^{-10}$) (Supplementary Figures 7, 8, 9 & Supplementary Table 4). As the normalized expression level of a gene in TCGA level 3 data could be a combination of several transcripts, we further investigated whether the expression of these analyzed transcripts were derived from the methylated promoters. We were able to confirm that the analyzed transcript for *MGMT* (transcript ID: uc001lkh.2), *HOXD12* (transcript ID: uc010zev.1) and two transcripts for *GATA4* (transcript IDs: uc003wuc.2 and uc011kxc.1) were derived from the methylated promoters in melanoma patients. The report of high promoter methylation and high expression is relatively very rare in literature and these observations calls for future research further validate these findings and to reveal the role of methylated DNA in facilitating the corresponding transcriptional program.

### miRNA profiling of primary & metastatic melanoma tumors

We developed scan_tcga_miRNAs.awk for comprehensive profiling of miRNAs from TCGA data. For miRNAseq data in TCGA, two types of file are provided, they are mirna.quantification.txt and isoform.quantification.txt. scan_tcga_miRNAs.awk extracts RPM from mirna.quantification.txt files as the default. However, it is possible to retrieve other information if desired by specifying wanted_field in the command line. To demonstrate the utility of the software we curated a list of 30 miRNAs that were previously reported to be deregulated in melanoma (see the list of analyzed miRNAs in Supplementary Table 5) [42]. The expression levels ($\log^2$ of RPM) of the analyzed miRNAs for normal skin tissue, primary and metastatic melanomas are shown in Figure 6. The separation of primary and metastatic patients was more pronounced in PCA for miRNA than mRNA expression or DNA methylation profiles (Supplementary Figure 10). We observed that a group of miRNAs were consistently expressed either at a high level (e.g., let7a/b, miR-21 and miR-30d, Figure 6) or a low level (e.g., miR-34b, miR-26a, miR-137, miR-18b). These results are consistent with previous reports of deregulation of

these miRNAs in melanoma [43–48]. Furthermore, a group of miRNA showed distinct expression profiles between primary and melanoma patients. We identified miR-193b, miR-203, miR-200a, miR-200c, miR-205 expression to be significantly downregulated in metastatic patients compared with primary (Figure 7 & Supplementary Table 6). Our analysis confirmed the previous report of miR-193b downregulation in metastatic melanoma [49]. In addition, downregulation of miR-203, miR-200a, miR-200c, miR-205 was previously reported in several melanoma cell lines [50–52]. Our data reveal that the loss of expression in these miRNAs occurs predominantly in metastatic melanomas compared with primary.

## Conclusion

We provide a suite of intuitive command line based tools to analyze different layers of epigenomic data from TCGA. These programs provide users with flexibility to choose a subgroup of patients, retrieve large TCGA data and perform comprehensive analysis of the epigenome. The scan_tcga tools are currently designed to exclusively interrogate TCGA level 3 data. These datasets are processed and presented in standard format. However, for independent additional datasets if the raw data are processed in similar format to TCGA, scan_tcga tools could be adopted for analysis of those files. Expanding the scan_tcga tools for analyzing other types of datasets with different formats will be a subject for future developments.

The output files of scan_tcga tools are in an easily accessible text format and readily usable for bench scientists without bioinformatics knowledge. If advanced analysis is required, the output files could be directly read into the R environment for further analysis and plotting with several publicly available packages. The matrix format text files generated by scan_tcga tools are compatible with other tools for analysis. We believe the scan_tcga tools are complimentary to the existing R bioconductor based tools such as TCGAbiolinks [17], TCGA2STAT [53] or RTCGAToolbox [54]. The future development of scan_tcga will include options for retrieving multiple fields in one operation and additional function for statistical tests. We illustrate the usage of scan_tcga tools by analyzing primary and metastatic melanoma patients. Our analysis confirms previous reports of aberrant methylation and expression patterns of several genes and miRNA. In addition, the current analysis reveals novel patterns of differential methylation and differential expression of mRNA and miRNA between primary and metastatic melanoma patients. scan_tcga tools and documentation are distributed as part of this article and also are freely accessible at GitHub with relevant test datasets.

### Supplementary data

To view the supplementary data that accompany this paper please visit the journal website at: www.futuremedicine.com/doi/full/10.2217/epi-2016-0063

### Financial & competing interests disclosure

A Chatterjee and MR Eccles would like to gratefully acknowledge the funding support from the New Zealand Institute for Cancer Research Trust. This research was also supported by funding from the HS & JC Anderson Trust, University of Otago Dunedin School of Medicine, the Maurice & Phyllis Paykel Trust and the Maurice Wilkins Centre for Molecular Biodiscovery. The authors have no other relevant affiliations or financial involvement with any organization or entity with a financial interest in or financial conflict with the subject

### Executive summary

- The Cancer Genome Atlas contains multiple levels of genomic data. However, lack of computational methods to investigate subgroups of patients makes it difficult for bench scientists to use these data effectively.
- We present three command line based scripts to perform multi-omic analysis of The Cancer Genome Atlas data.
- We have developed scan_tcga_methylation.awk (DNA methylation), scan_tcga_mRNA.awk (for mRNA analysis from RNA-Seq data) and scan_tcga_miRNAs.awk (for miRNA expression).
- Using these tools we have analyzed 60 frequently deregulated cancer genes in primary and metastatic melanomas. We identified hypermethylation of *CDH1*, *EZH2*, *NOTCH* and *TET3* promoters in metastatic melanomas compared to primary. For mRNA expression levels, we found *ROR1*, *SNAI3*, *TET1* and *TSPAN13* were significantly upregulated in metastatic compared to primary, while we identified significant downregulation of the *MMP3* gene in metastatic patients. Although higher promoter methylation was associated with lower expression generally, we identified significant positive correlation (i.e., high methylation in promoters associated with high mRNA expression) for five genes.
- Our miRNA analysis revealed that miR-193b, miR-203, miR-200a, miR-200c and miR-205 expression to be significantly downregulated in metastatic patients compared to primary.
- scan_tcga tools and documentation are distributed as part of this article and are also freely accessible at GitHub with relevant test datasets.

matter or materials discussed in the manuscript apart from those disclosed.

No writing assistance was utilized in the production of this manuscript.

## References


1 Stratton MR, Campbell PJ, Futreal PA. The cancer genome. *Nature* 458(7239), 719–724 (2009).

2 Pareek CS, Smoczynski R, Tretyn A. Sequencing technologies and genome sequencing. *J. Appl. Genet.* 52(4), 413–435 (2011).

3 Zhang K, Wang H. [Cancer Genome Atlas Pan-cancer Analysis Project]. *Zhongguo Fei Ai Za Zhi* 18(4), 219–223 (2015).

4 Cancer Genome Atlas Research N, Weinstein JN, Collisson EA *et al.* The Cancer Genome Atlas Pan-Cancer analysis project. *Nat. Genet.* 45(10), 1113–1120 (2013).

5 Pennisi E. Human genome 10th anniversary. Will computers crash genomics? *Science* 331(6018), 666–668 (2011).

6 TCGA Data Portal. https://tcga-data.nci.nih.gov/tcga

7 Forbes SA, Beare D, Gunasekaran P *et al.* COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.* 43, D805–D811 (2015).

8 Beroukhim R, Mermel CH, Porter D *et al.* The landscape of somatic copy-number alteration across human cancers. *Nature* 463(7283), 899–905 (2010).

9 Rubio-Perez C, Tamborero D, Schroeder MP *et al. In silico* prescription of anticancer drugs to cohorts of 28 tumor types reveals targeting opportunities. *Cancer Cell* 27(3), 382–396 (2015).

10 Gonzalez-Perez A, Perez-Llamas C, Deu-Pons J *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nat. Methods* 10(11), 1081–1082 (2013).

11 Gao J, Aksoy BA, Dogrusoz U *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.* 6(269), pl1 (2013).

12 Cerami E, Gao J, Dogrusoz U *et al.* The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.* 2(5), 401–404 (2012).

13 Diez-Villanueva A, Mallona I, Peinado MA. Wanderer, an interactive viewer to explore DNA methylation and gene expression data in human cancer. *Epigenetics Chromatin* 8, 22 (2015).

14 Samur MK, Yan Z, Wang X *et al.* canEvolve: a web portal for integrative oncogenomics. *PLoS ONE* 8(2), e56228 (2013).

15 Deng M, Bragelmann J, Schultze JL, Perner S. Web-TCGA: an online platform for integrated analysis of molecular cancer data sets. *BMC Bioinformatics* 17, 72 (2016).

16 ICTS Compass. https://research.icts.uiowa.edu/compass/index.html

17 Colaprico A, Silva TC, Olsen C *et al.* TCGAbiolinks: an R/Bioconductor package for integrative analysis of TCGA data. *Nucleic Acids Res.* 44(8), e71 (2016).

18 Alizadeh AA, Aranda V, Bardelli A *et al.* Toward understanding and exploiting tumor heterogeneity. *Nat. Med.* 21(8), 846–853 (2015).

19 Easwaran H, Tsai HC, Baylin SB. Cancer epigenetics: tumor heterogeneity, plasticity of stem-like states, and drug resistance. *Mol. Cell* 54(5), 716–727 (2014).

20 NCI's Genomic Data Commons. https://gdc.nci.nih.gov

21 National Cancer Institute: NCI Wiki. https://wiki.nci.nih.gov/display/TCGA/TCGA+barcode

22 Broad Institute. http://gdac.broadinstitute.org

23 Broad Institute: Skin Cutaneous Melanoma (SKCM) Samples Report. http://gdac.broadinstitute.org

24 Scan_tcga_methylation. https://github.com/peterstockwell

25 Scan_tcga_mRNAs. https://github.com/peterstockwell

26 Scan_tcga_miRNAs. https://github.com/peterstockwell

27 Aho BWK A. V., Weinberger P. J. *The AWK Programming Language.* Addison-Wesley, MA, USA (1988).

28 Liu S, Ren S, Howell P, Fodstad O, Riker AI. Identification of novel epigenetically modified genes in human melanoma via promoter methylation gene profiling. *Pigment Cell Melanoma Res.* 21(5), 545–558 (2008).

29 Muthusamy V, Duraisamy S, Bradbury CM *et al.* Epigenetic silencing of novel tumor suppressors in malignant melanoma. *Cancer Res.* 66(23), 11187–11193 (2006).

30 Mirmohammadsadegh A, Marini A, Nambiar S *et al.* Epigenetic silencing of the *PTEN* gene in melanoma. *Cancer Res.* 66(13), 6546–6552 (2006).

31 Lahtz C, Stranzenbach R, Fiedler E, Helmbold P, Dammann RH. Methylation of *PTEN* as a prognostic factor in malignant melanoma of the skin. *J. Invest. Dermatol.* 130(2), 620–622 (2010).

32 Marzese DM, Scolyer RA, Huynh JL *et al.* Epigenome-wide DNA methylation landscape of melanoma progression to brain metastasis reveals aberrations on homeobox D cluster associated with prognosis. *Hum. Mol. Genet.* 23(1), 226–238 (2014).

33 De Araujo ES, Pramio DT, Kashiwabara AY *et al.* DNA methylation levels of melanoma risk genes are associated with clinical characteristics of melanoma patients. *Biomed. Res. Int.* 2015, 376423 (2015).

34 Fernandez NB, Lorenzo D, Picco ME *et al.* ROR1 contributes to melanoma cell growth and migration by regulating N-cadherin expression via the PI3K/Akt pathway. *Mol. Carcinog.* doi:10.1002/mc.22426 (2015) (Epub ahead of print).

35 Arencibia JM, Martin S, Perez-Rodriguez FJ, Bonnin A. Gene expression profiling reveals overexpression of *TSPAN13* in prostate cancer. *Int. J. Oncol.* 34(2), 457–463 (2009).

36 Gras B, Jacqueroud L, Wierinckx A *et al.* Snail family members unequally trigger EMT and thereby differ in their ability to promote the neoplastic transformation of mammary epithelial cells. *PLoS ONE* 9(3), e92254 (2014).

37 Lian CG, Xu Y, Ceol C *et al.* Loss of 5-hydroxymethylcytosine is an epigenetic hallmark of melanoma. *Cell* 150(6), 1135–1146 (2012).

38 Kessenbrock K, Plaks V, Werb Z. Matrix metalloproteinases: regulators of the tumor microenvironment. *Cell* 141(1), 52–67 (2010).

39 Zhu YH, Liu H, Zhang LY *et al.* Downregulation of LGI1 promotes tumor metastasis in esophageal squamous cell carcinoma. *Carcinogenesis* 35(5), 1154–1161 (2014).

40 Kwon M, Lee SJ, Wang Y *et al.* Filamin A interacting protein 1-like inhibits WNT signaling and MMP expression to suppress cancer cell invasion and metastasis. *Int. J. Cancer* 135(1), 48–60 (2014).

41 Jones PA. Functions of DNA methylation: islands, start sites, gene bodies and beyond. *Nat. Rev. Genet.* 13(7), 484–492 (2012).

42 De Unamuno B, Palanca S, Botella R. Update on melanoma epigenetics. *Curr. Opin. Oncol.* 27(5), 420–426 (2015).

43 Gaziel-Sovran A, Segura MF, Di Micco R *et al.* miR-30b/30d regulation of GalNAc transferases enhances invasion and immunosuppression during metastasis. *Cancer Cell* 20(1), 104–118 (2011).

44 Seftor RE, Seftor EA, Gehlsen KR *et al.* Role of the alpha v beta 3 integrin in human melanoma cell invasion. *Proc. Natl Acad. Sci. USA* 89(5), 1557–1561 (1992).

45 Luo C, Tetteh PW, Merz PR *et al.* miR-137 inhibits the invasion of melanoma cells through downregulation of multiple oncogenic target genes. *J. Invest. Dermatol.* 133(3), 768–775 (2013).

46 Stark MS, Bonazzi VF, Boyle GM *et al.* miR-514a regulates the tumour suppressor NF1 and modulates BRAFi sensitivity in melanoma. *Oncotarget* 6(19), 17753–17763 (2015).

47 Levati L, Pagani E, Romani S *et al.* MicroRNA-155 targets the SKI gene in human melanoma cell lines. *Pigment Cell Melanoma Res.* 24(3), 538–550 (2011).

48 Dar AA, Majid S, Rittsteuer C *et al.* The role of miR-18b in MDM2–p53 pathway signaling and melanoma progression. *J. Natl Cancer Inst.* 105(6), 433–442 (2013).

49 Chen J, Feilotter HE, Pare GC *et al.* MicroRNA-193b represses cell proliferation and regulates cyclin D1 in melanoma. *Am. J. Pathol.* 176(5), 2520–2529 (2010).

50 Noguchi S, Mori T, Otsuka Y *et al.* Anti-oncogenic microRNA-203 induces senescence by targeting E2F3 protein in human melanoma cells. *J. Biol. Chem.* 287(15), 11769–11777 (2012).

51 Elson-Schwab I, Lorentzen A, Marshall CJ. MicroRNA-200 family members differentially regulate morphological plasticity and mode of melanoma cell invasion. *PLoS ONE* 5(10), pii: e13176 (2010).

52 Dar AA, Majid S, De Semir D, Nosrati M, Bezrookove V, Kashani-Sabet M. miRNA-205 suppresses melanoma cell proliferation and induces senescence via regulation of E2F1 protein. *J. Biol. Chem.* 286(19), 16606–16614 (2011).

53 Wan YW, Allen GI, Liu Z. TCGA2STAT: simple TCGA data access for integrated statistical analysis in R. *Bioinformatics* 32(6), 952–954 (2016).

54 Samur MK. RTCGAToolbox: a new tool for exporting TCGA Firehose data. *PLoS ONE* 9(9), e106397 (2014).