Published online 14 January 2016 *Nucleic Acids Research, 2016, Vol. 44, No. 7* ***e62***
doi: 10.1093/nar/gkv1478

CrossHub: a tool for multi-way analysis of The Cancer Genome Atlas (TCGA) in the context of gene expression regulation mechanisms

George S. Krasnov[1,2,3], Alexey A. Dmitriev[1,*], Nataliya V. Melnikova[1], Andrew R. Zaretsky[4], Tatiana V. Nasedkina[1,2], Alexander S. Zasedatelev[1,2], Vera N. Senchenko[1] and Anna V. Kudryavtseva[1,2]

[1]Engelhardt Institute of Molecular Biology, Russian Academy of Sciences, Moscow 119991, Russia, [2]N.N. Blokhin Russian Cancer Research Center, Moscow 115478, Russia, [3]Orekhovich Institute of Biomedical Chemistry, Russian Academy of Medical Sciences, Moscow 119121, Russia and [4]M.M. Shemyakin-Yu.A. Ovchinnikov Institute of Bioorganic Chemistry, Russian Academy of Sciences, Moscow 117997, Russia

Received July 15, 2015; Revised November 20, 2015; Accepted December 5, 2015

ABSTRACT

The contribution of different mechanisms to the regulation of gene expression varies for different tissues and tumors. Complementation of predicted mRNA–miRNA and gene–transcription factor (TF) relationships with the results of expression correlation analyses derived for specific tumor types outlines the interactions with functional impact in the current biomaterial. We developed CrossHub software, which enables two-way identification of most possible TF–gene interactions: on the basis of ENCODE ChIP-Seq binding evidence or Jaspar prediction and co-expression according to the data of The Cancer Genome Atlas (TCGA) project, the largest cancer omics resource. Similarly, CrossHub identifies mRNA–miRNA pairs with predicted or validated binding sites (TargetScan, mirSVR, PicTar, DIANA microT, miRTarBase) and strong negative expression correlations. We observed partial consistency between ChIP-Seq or miRNA target predictions and gene–TF/miRNA co-expression, demonstrating a link between these indicators. Additionally, CrossHub expression-methylation correlation analysis can be used to identify hypermethylated CpG sites or regions with the greatest potential impact on gene expression. Thus, CrossHub is capable of outlining molecular portraits of a specific gene and determining the three most common sources of expression regulation: promoter/enhancer methylation, miRNA interference and TF-mediated activation or repression. CrossHub generates formatted Excel workbooks with the detailed results. CrossHub is freely available at https://sourceforge.net/projects/crosshub/.

INTRODUCTION

The Cancer Genome Atlas (TCGA) project is one of the largest available resources that accumulates genomic, transcriptomic and methylomic data for several types of cancer. During the first three years of the pilot phase (2006–2009), TCGA focused on large-scale studies of glioblastoma multiforme, lung and ovarian cancers (1). Today, TCGA includes omics data for more than 20 cancer types. For each of the most common cancers (lung, breast, prostate and others), TCGA collected genomic, methylomic and transcriptomic portraits of more than 300–500 samples. This makes TCGA a useful source of information for gene expression alteration (2), tumor molecular subtype classification (3,4), discovery of driver aberrations (5), identification of prognostic markers (6,7) and other applications.

Complementation of multidimensional omics projects with other resources can significantly increase the value of the results and highlight the most prominent associations. Integration of microRNA (miRNA) target prediction algorithms with the results of miRNA–mRNA expression correlation analysis can be used to identify the largest number of possible miRNA targets. This approach is implemented using the MiRGator resource (8). Additionally, complementation of ChIP-Seq data with the results of gene expression correlation studies increases the efficacy of identifying interactions between transcription factors (TFs) and target genes.

The ENCyclopedia of DNA Elements (ENCODE) is another large international project aiming to identify functional elements in the human genome and reveal relation-

[*]To whom correspondence should be addressed. Tel: +7 499 1356009; Fax: +7 499 1351405; Email: Alex_245@mail.ru

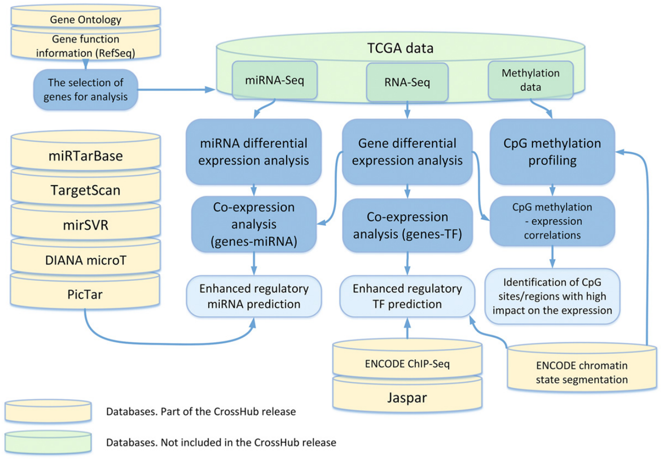
Gene Ontology
Gene function information (RefSeq)
The selection of genes for analysis
TCGA data
miRNA-Seq
RNA-Seq
Methylation data
miRTarBase
TargetScan
mirSVR
DIANA microT
PicTar
miRNA differential expression analysis
Gene differential expression analysis
CpG methylation profiling
Co-expression analysis (genes-miRNA)
Co-expression analysis (genes-TF)
CpG methylation - expression correlations
Enhanced regulatory miRNA prediction
Enhanced regulatory TF prediction
Identification of CpG sites/regions with high impact on the expression
ENCODE ChIP-Seq
Jaspar
ENCODE chromatin state segmentation
Databases. Part of the CrossHub release
Databases. Not included in the CrossHub release

Figure 1. CrossHub workflow. Complementation of ENCODE ChIP-Seq data and Jaspar predictions with TCGA expression correlation analysis allows the user to outline interactions with potential functional impacts to a specific cancer subtype. Similarly, combining miRNA target predictions with gene–miRNA expression correlation profiling (based on TCGA expression data) highlights gene–miRNA interactions, which likely take place for a particular tumor type. Expression-methylation correlation analysis allow identification of hypermethylated CpG sites or regions within promoters or enhancers (annotated with ENCODE) having the greatest potential impact on gene expression. In addition, CrossHub enables conventional differential expression (DE) and methylation analysis.

ships between these elements (9). ENCODE provides various types of data related to gene expression regulation: histone modification profiles (revealed by CHIP-Seq), open chromatin patterns (DNaseI assays and FAIRE-Seq), TF binding sites (TFBS) (ChIP-Seq), chromatin interactions (5C and ChIA-PET), DNA methylation (reduced representation of bisulfite sequencing and Illumina methyl-sensitive microarrays) and other features (10). ENCODE includes ChIP-Seq data for 160 TFs across 3–6 cell lines used as a core set. However, the binding of some TFs (CTCF, Pol II, RELA) has been profiled for more than 20 cell lines (11). Based on histone modification profiles and TFBSs, ENCODE offers annotation of genome segments (promoter, enhancer, insulator) for six cell lines using two different machine learning techniques (ChromHMM and Segway) (12).

Increasing the availability of large-scale data necessitates the creation of scalable platforms for multi-way analysis of these results. In the present work, we present CrossHub, a novel software tool that integrates multi-resource omics data. CrossHub was designed to analyze TCGA transcriptomic and epigenomic data in the context of ENCODE, Jaspar and various miRNA target prediction algorithms. This approach is intended to reveal gene expression regulation mechanisms such as methylation, TF-mediated transcription repression/activation and microRNA interference.

MATERIALS AND METHODS

CrossHub is a standalone Python-based application providing multiple methods for analyzing TCGA data (Figure 1). Users should download RNA-Seq, miRNA-Seq and methylation profiles (Illumina BeadChip) data from the TCGA Data Portal or other resources. CrossHub is released with dumps of ENCODE ChIP-Seq and Chromatin segmentation data, Jaspar matrix profiles and predictions, and five miRNA target databases (up to date as of November 2015). Otherwise, users can download source database files to parse them with CrossHub.

Differential expression (DE) analysis of genes and microRNA

CrossHub analyzes TCGA RNA-Seq data obtained using Illumina HiSeq, GA_{II} or other platforms. Two common approaches are used: expression analysis across pools of normal and tumor samples and across paired (tumor-normal) samples. The latter is considered to be the most reliable. To assess DE, CrossHub uses *t*-tests for dependent (paired) and independent (pooled) samples, taking into account the Poisson distribution of reads when evaluating dispersion and expression fold-change logarithms (*LogFC*). Next, Benjamini–Hochberg correction is introduced to estimate the false-discovery rate (FDR). Finally, all data are summarized in formatted Microsoft Excel workbooks. Selection of the genes of interest can be performed using Gene Ontology keywords.

Refining TF–gene functional associations with expression correlation analysis

The second feature provided by CrossHub is gene co-expression analysis complemented with ENCODE TF ChIP-seq data and Jaspar TFBS predictions. Jaspar is a database of nucleotide profiles describing the binding preferences of transcription factors. Jaspar includes 205 TFBS profiles in vertebrates (13). For each gene–TF association, CrossHub calculates an ENCODE score based on ChIP-Seq read count and the quantity of such observations. Next, CrossHub adjusts the score according to the distance from the transcription start site (TSS); the greatest score multipliers are assigned to the TSS-proximal regions ($-500\ldots+300$ bp). Additionally, CrossHub provides TFBS site annotation according to ENCODE genome segments annotation by ChromHMM and Segway (promoter/insulator/enhancer etc.). When possible, gene–TF associations are annotated with Jaspar predictions. Here, TFBS score is calculated based on the matrix-based Jaspar profile score, TFBS distance from TSS, and genome segments annotation. A correlation heatmap with a detailed description of the predicted TFBS is generated, and top gene–TF associations with the greatest ENCODE/Jaspar scores and highest/lowest correlation coefficients (r_s) are reported.

Refining microRNA target predictions with mRNA–miRNA expression correlation analysis

This analysis implements a similar two-way approach for identifying mRNA–miRNA regulation interactions. The results of expression correlation analysis are supplemented with miRNA target prediction databases (TargetScan, DIANA microT, mirSVR, PicTar) and experimental evidence of mRNA–miRNA interactions (miRTarBase) (14–18). For each gene–miRNA pair, a cumulative score is determined basing on the normalized scores of individual databases and relative reliability of the database. To determine the normalized DB score, CrossHub first ranks the pool of predicted miRNA binding sites with the internal DB score and determines which percentile P_{site} matches to a current miRNA binding site. Next, CrossHub calculates a score for this pair $S_{\mathrm{pair,DB}}$ according to the current database. If several miRNA binding sites are predicted for any gene or several mRNA isoforms are present, CrossHub merges this multi-hit as: $S_{\mathrm{pair,DB}} = 25 \cdot (\sum_{\mathrm{sites}} (1.25 - P_{\mathrm{site}}/100)^3)^{-1/3}$. Thus, if there is one binding site with the highest internal database score (100th percentile), $S_{\mathrm{pair,DB}} = 100$ for this gene–miRNA pair. The 75th percentile matches only to $S_{\mathrm{pair,DB}} = 50$. Next, CrossHub combines $\mathrm{S_{pair,db}}$ across several databases and calculates an overall score $\mathrm{S_{pair,main}}$: $S_{\mathrm{pair,main}} = \left(\sum_{\mathrm{all\,DB}} (S_{\mathrm{pair,DB}} W_{\mathrm{DB}})^{1.5}\right)^{1/1.5}$. Here, W_{DB} is the weight of a database reflecting its reliability. As described below, miRTarBase (strong experimental evidences only), PicTar and TargetScan (conservative miRNA binding sites only) and DIANA microT databases have the best consistency with the gene–miRNA co-expression analysis results. Basing on this criterion, we assigned the highest W_{DB} (3.0) to miRTarBase (strong experimental evidences), normal W_{DB} (0.8–1.2) to PicTar and TargetScan (conservative sites) and low W_{DB} (0.1–0.4) to mirSVR and databases of non-conservative sites. A coefficient of 1.5 is selected in order to optimize the balance between the significance of individual prediction scores and the predictions count.

The correlation table, database scores and top associations (with the highest scores and lowest correlation coefficients) are reported in CrossHub.

Methylation profiling

CrossHub performs differential CpG methylation analysis based on the data from Illumina Infinium HumanMethylation450 BeadChips (TCGA) corresponding to the 485 000 CpG sites in the human genome (∼17 CpG sites per gene). The distributions of β-values (ratio of methylated alleles) between pools of tumor and normal samples are compared, taking into account the heterogeneity of methylation patterns across the samples. A combined hyper/hypomethylation score for pooled samples is calculated based on the Mann–Whitney U test *P*-value and comparison of the 10, 25, 50, 75 and 90th β-value percentiles between normal and tumor pools. CrossHub evaluates the hyper/hypomethylation score for paired samples according to: (i) the mean Δβ-value between matched normal and tumor tissues; (ii) the frequency of cases with $|\Delta\beta| > 0.4$; and (iii) the *P*-value for Wilcoxon signed ranked test. The last method is more reliable and accurate; however, the quantity of microarray-analyzed paired samples is low for some cancers (e.g. rectum) and therefore is not suitable for rigorous statistical analysis. Finally, an overall gene hyper/hypomethylation score is calculated based on individual CpG scores and the number of these CpG sites. The selection of promoter and enhancer CpG dinucleotides is performed using ChromHMM and Segway data (ENCODE). For each CpG site, correlation coefficients are calculated between (i) the β-value and normalized gene expression in pools of normal and tumor samples and (ii) the expression *LogFC* and Δβ-value between matched normal and tumor tissues (for paired samples only). The second method is considered to be more reliable.

Differential expression (DE) and methylation analysis results as well as top gene–TF and gene–miRNA associations are summarized into one formatted Excel worksheet, representing a ‘portrait’ for each gene. This analysis is applicable to more than 15 cancer types represented in the TCGA.

RESULTS

Associations between methylation and downregulation

In the present work, we present CrossHub, an integrative tool aimed at analyzing multi-dimensional TCGA data in the context of ENCODE, Jaspar and other resources. First, CrossHub attempts to link gene expression and the methylation of CpG sites annotated as promoters or enhancers according to the ENCODE data. We tested the associations between promoter hypermethylation and gene expression for three cancer types: colon adenocarcinoma, lung squamous cell carcinoma and prostate adenocarcinoma. To shorten the gene list, we limited the analysis to genes encoding cytoplasmic proteins (selected with Gene Ontology keywords, total of 4489 genes). To discard low-coverage genes, we introduced a threshold: a gene should have at least 100 reads/sample for 30% of the total samples count. This threshold is adjustable (See Manual). Scatterplots illustrating promoter hypermethylation scores (*HMS*) and gene expression levels of *LogFC* are shown in Figure 2.

Prominent bias toward expression downregulation for genes with high hypermethylation scores is typically observed. However, some genes with high promoter hypermethylation scores are upregulated. A significant number of these have uncertain DE reliability score which is proportional to the absolute values of *LogFC* and logarithm of *FDR* (green, yellow circle colors, Figure 2); additionally, some have multiple promoters.

We split the analyzed genes into two groups (low *HMS* and high *HMS*) and then evaluated *LogFC* bias between them using Mann–Whitney U-test. Several *HMS* thresholds (T_{HMS}) were used for splitting (Figure 2). For each threshold, *LogFC* distribution bias was statistically significant (*P*-varied from $<10^{-16}$ to 10^{-6}). We also observed mean *LogFC* decrease with increasing T_{HMS}. Thus, as expected, CpG hypermethylation found by CrossHub using TCGA data is associated with gene expression downregulation.

Combining gene–TF expression correlation analysis and TF target prediction with ChIP-Seq

Next, we tested the agreement between two predictors of gene–TF functional relationships: whether gene-TF binding revealed by ChIP-Seq is associated with expression correlation bias. We chose two TF strongly upregulated in colon cancer, according to TCGA RNA-Seq data: Myc (c-Myc) and CBX3. Myc is well-known oncogenic protein responsible for the transduction of growth promoting signal (19). Myc activates transcription of many genes participating in cell cycle regulation and metabolism reprogramming, hypoxic adaptation, DNA replication and other processes

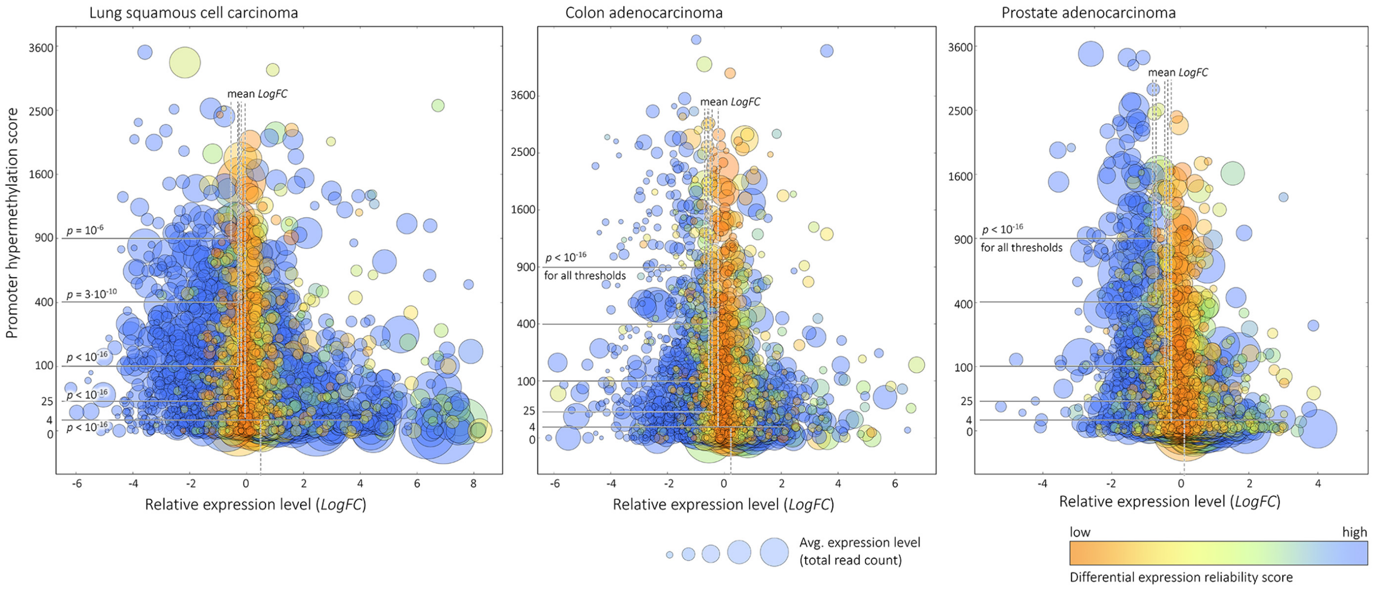

Figure 2. Associations between the promoter hypermethylation score (*HMS*) and the logarithm of gene expression level changes in tumors (*LogFC*). Circle colors indicate gene DE reliability score, which is proportional to the absolute values of *LogFC* and logarithm of false-discovery rate (FDR). Circle size is proportional to square root of total read count for a gene. For all three cancers, a significant increase in the ratio of downregulated genes was observed for genes with positive promoter hypermethylation scores. We selected several *HMS* thresholds (T_{HMS}) to prove the statistical significance of differences between distribution of *LogFC* for genes with $HMS \geq T_{HMS}$ and $HMS < T_{HMS}$. Vertical dashed lines indicate mean *LogFC* for these groups. Average *LogFC* decreases with increasing *HMS*.

(19–21). Myc is one of the most extensively studied TF within the ENCODE ChIP-Seq project.

The second TF, CBX3 (Chromobox homolog 3) binds lamin B receptor, an integral membrane protein found in the inner nuclear membrane, and may be responsible either for transcription activation or suppression *via* maintenance of heterochromatin (22). Studies with CBX3-knockdown assays revealed that depletion of CBX3 resulted in downregulation of a subset of genes co-localized with CBX3; loss of CBX3 leads to a dramatic accumulation of unspliced nascent transcripts (22). CBX3 is crucial for reprogramming of somatic cells into induced pluripotent stem cells (23) and can act as a marker of tumor stem cells (24). CBX3 gene fusions and overexpression were found in cancer (25–27). We tested associations between ChIP-Seq score for these TFs and co-expression with the potential target genes using two samplings. One of Myc metabolic targets is the activation of glycolysis (20,21,28), and the first sampling includes glucose metabolism-related genes (430 genes). The second sampling includes genes encoding extracellular proteins (3600 genes).

Figure 3 illustrates the distribution of genes based on two parameters: adjusted ENCODE ChIP-Seq score (S) and the Spearman correlation coefficient (r_s) between expression levels of a gene and *Myc* or *CBX3* for colon adenocarcinoma TCGA RNA-Seq dataset. As for the previous analysis, we used several score thresholds (T_S) to split genes into low-score and high-score groups (Figure 3). We found that the bias of distribution of r_s between the groups was statistically significant for each threshold (P varied from $<10^{-16}$ to 0.01), and mean r_s for high-score group increased with increasing T_S. This clearly illustrates a linkage between the ChIP-Seq score S and Spearman r_{s} and the informativeness of both these predictors. Their use in combination will improve the accuracy of identification of potential TF–gene functional relationships.

In contrast to ENCODE, Jaspar revealed only slightly noticeable differences in the Spearman r_{s} distribution between low-score and high-score genes (data not shown). Thus, ENCODE ChIP-Seq data have greater predictive value of TF–gene interactions than Jaspar.

Integration of miRNA target prediction with gene–miRNA expression correlation analysis

Finally, we tested the association between the prediction of microRNA target genes and Spearman correlation coefficients for the gene–miRNA expression levels. For this analysis, we selected genes encoding extracellular proteins (3600 genes) and top overexpressed miRNAs, according to the results of TCGA colon cancer miRNA-Seq dataset analysis. We have excluded few miRNAs with extremely high abundance in both tumor and normal tissues (>2 billion reads). Firstly, we tested the dependency of mean r_s on miRNA target prediction score threshold (T_{mS}) for individual databases (Figure 4A–E). The 2D-histogram illustrates the distribution density of paired gene–miRNA relationships depending on their expression correlation coefficients (horizontal axis, colon cancer TCGA dataset) and target prediction scores ($S_{\mathrm{pair,DB}}$, vertical axis). Mean r_s is expected to decrease with increasing T_{mS}, and this criterion permits to assess the confidence of miRNA target prediction databases. This trend was the most prominent for TargetScan conservative sites, DIANA microT (Figure 4B and C) and PicTar (data not shown). Mean r_s difference (d) between zero-score gene–miRNA pairs (e.g. not predicted)

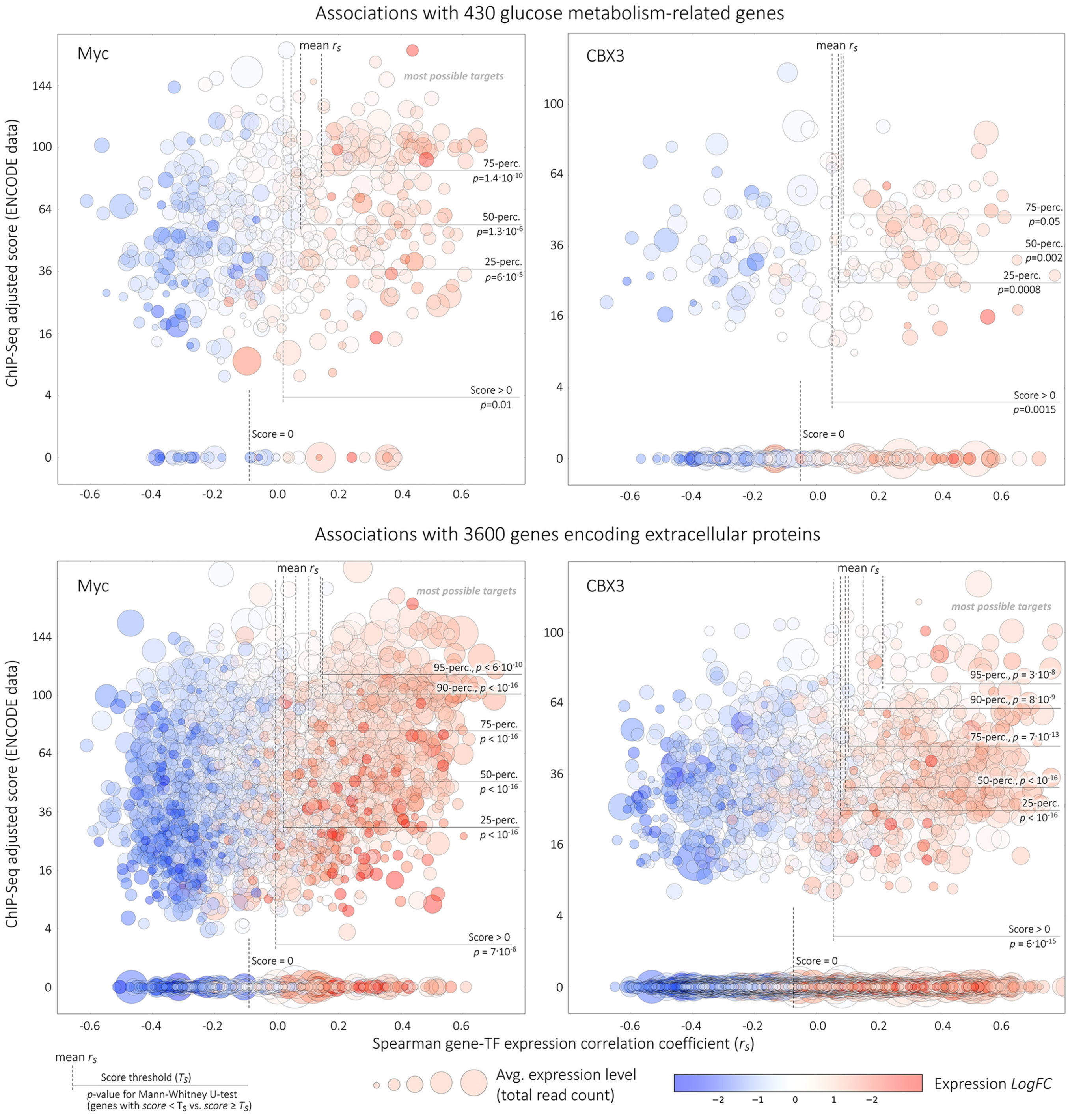

Figure 3. Distribution of genes for two parameters: ENCODE ChIP-Seq transcription factor (TF) binding score and Spearman gene–TF expression correlation coefficients (r_S; colon cancer TCGA dataset). Two samplings were analyzed: genes participating in the glucose transport and metabolism (top) and genes encoding extracellular proteins (bottom). Genes with no ChIP-Seq evidence of TF binding are marked with zero score. Circle size is proportional to square root of total read count for a gene. Circle color indicates gene expression level change in tumor. The analysis was performed for two TF strongly upregulated in colon cancer: well-known oncogenic protein Myc and CBX3 which is less extensively studied in the context of cancer. We compared distributions of r_S between genes that passed and did not pass score thresholds (T_S). Several T_S were selected: >0 (any positive score), 25th, 50th, 75th and 90th score percentiles. Vertical dashed lines indicate mean values of r_S for these groups. For each T_S we observed statistically significant difference between the distributions of r_S indicating linkage of these characteristics: ChIP-Seq score and TF–gene co-expression.

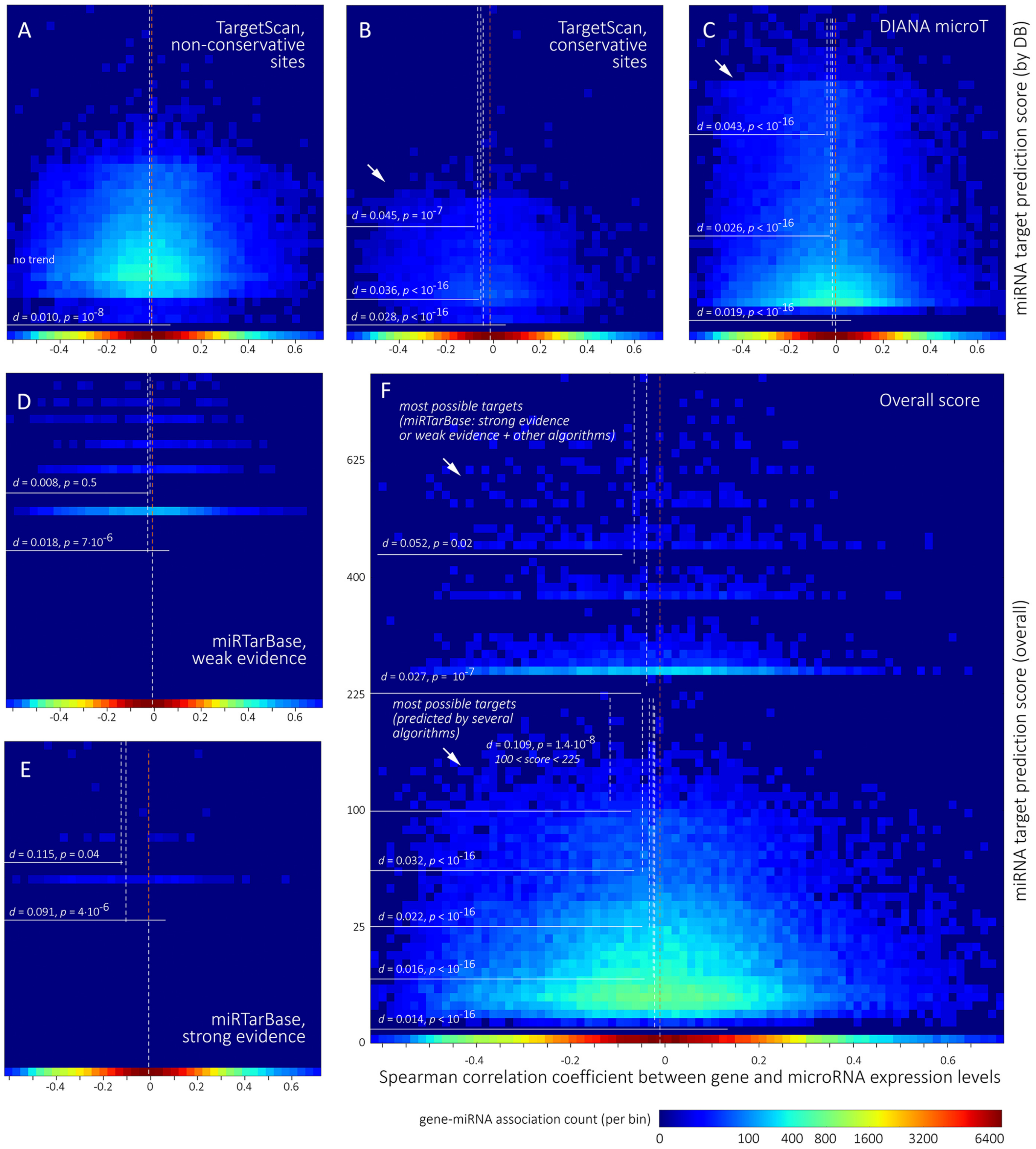

Figure 4. Distribution density of gene–microRNA pairs in expression level correlation coefficients (r_s) and miRNA binding site scores according to TargetScan (**A** and **B**), DIANA microT (**C**), miRTarBase (**D** and **E**) and overall score according to several algorithms (**F**). TargetScan (conservative sites), DIANA microT showed the greatest mean r_s bias among the analyzed prediction algorithms. Distribution density is slightly asymmetrical for these databases, especially for high scores ($d = 0.02$–0.045). These areas are marked with an arrow. However, miRNA–gene relationships predicted by several algorithms showed a prominent r_s bias ($d = 0.109$; overall score range 100—225; F). This region represents the maximum number of true miRNA–gene relationships with functional impact. Another region with a significant r_s bias ($d = 0.052$, overall score >400) mainly includes miRNA–gene relationships with strong experimental evidence or weak evidence coupled to miRNA binding site prediction by one or more algorithms.

and pairs that have passed T_{mS} filter was relatively high for TargetScan (any T_{mS}; $d = 0.028–0.045$, Mann–Whitney U-test $P < 10^{-16}$–10^{-7}) and DIANA microT (medium and high T_{mS}; $d = 0.026–0.043$, $P < 10^{-16}$). TargetScan database of non-conservative sites and mirSVR database demonstrated very modest r_s bias ($d = 0.01$) with no trend of its increasing with elevating T_{mS}. A significant r_s bias ($d = 0.05$, $P = 0.003$ for pairs with score >0) was observed for several datasets of PicTar: conservative in vertebrates, in mammals and birds, or in mammals only (data not shown). Finally, miRTarBase demonstrated a significant r_s bias ($d = 0.091$, $P = 7{\cdot}10^{-6}$) for associations with strong experimental evidence (Figure 4D and E). Basing on these findings, we selected database weights W_{DB} for calculating overall target prediction score (see 'Materials and Methods' section).

Figure 4F illustrates distribution density of gene–miRNA pairs on Spearman correlation coefficient r_s and overall target prediction score $S_{pair,main}$. It can be seen that r_s bias increases with increasing T_{mS}. Score range $100 < S_{pair,main} < 225$ is characterized with the greatest r_s bias ($d = 0.109$, $P = 10^{-7}$). This range should contain the maximum number of true relationships with functional impact; miRNA–gene relationships that are predicted with several algorithms are included in this score range. mirTarBase targets (weak evidence) are located in a separate group with higher scores but have lesser r_s bias ($d > 0.027$). However, miRNA–gene relationships with strong experimental evidence (or with weak evidence but also predicted by one the algorithms) are characterized with strong r_s bias ($d = 0.052$, top of the 2D-histogram). Notably, CrossHub reports all miRNA–gene scores for each database in the output files.

Thus, both ChIP-Seq data on transcription factors binding sites and miRNA target prediction data demonstrate particular consistency with the correlation analysis results. Although each of these predictors is weak, their combination into one prediction algorithm is expected to significantly increase the accuracy of mRNA–miRNA or gene–TF functional relationship discovery (8). In the present study, we demonstrated that the relationships with conservative miRNA binding sites identified simultaneously by several algorithms showed the greatest concordance with co-expression.

DISCUSSION

Cancer is a complex disease manifesting in transcriptomic, proteomic and epigenomic aberrations. The linkage of multidimensional genomic data helps to elucidate molecular mechanisms inherent in various tumor types (29–32). TCGA has been one of the starting points for dozens of studies, including validation of cancer susceptibility allele variants (33) and prognostic signatures (6), cancer molecular characteristics (34), studies of cancer angiogenesis (35), and cancer proteogenomic studies, one of the most recent fields of research (36–38). In this article, we present a novel tool for linking TCGA data and other resources in the context of gene expression regulation mechanisms: CpG methylation, transcription factors and microRNA.

Currently, nearly 20 different microRNA target prediction algorithms have been developed (39,40). However, none provide reliable results. No prediction method is consistently superior to the others (41). Different algorithms use different criteria sets, with most using seed match, site conservation, free energy and site accessibility (39). Three tools, DIANA-microT, mirSVR (miRanda) and TargetScan are considered to be the most suitable and are the most popular because of their wide range of capabilities, reliance on relatively updated versions of miRBase and ease of use (39).

CrossHub integrates data from four bioinformatics-based target prediction algorithms, including TargetScan, DIANA-microT, mirSVR and PicTar. This is implemented in the miRNA body-map and miRWalk databases (42,43). miRWalk database includes both mRNA–miRNA interactions predicted using miRWalk's algorithm and information from other resources, such as TargetScan, RNA22 and PicTar (43,44). The next milestone is to combine bioinformatics miRNA target prediction algorithms with gene–miRNA expression correlation results obtained from ever-growing deep sequencing databases (45). Here, TCGA was the optimal resource, as it integrates both RNA-Seq and miRNA-Seq data from several dozen cancer types. This approach is implemented in miRGator (8,46), SigTerms (47) and TopKCEMC (48). Although the accuracy of each of the prediction methods is rather weak, their combination significantly increased the overall prediction efficacy (8). Generally, miRNA target prediction based on sequence analysis is not tissue- or cancer-specific. Each tumor type can harbor very different patterns of mRNA–miRNA relationships; moreover, different microRNAs may show opposite properties, such as being either pro- or anti-oncogenic in various tumors (49,50). Here, gene–miRNA expression correlation analysis can make a prediction more sample-specific to extract interactions in the analyzed tissue or tumor type. However, only conservative sites predicted with three or more algorithms were consistent with the expression correlation results. No tendency of median bias or over-representation of mRNA–miRNA associations with significant negative expression correlation was observed for non-conservative sites (Figure 4). This enabled us to highlight site conservation and its identification with several algorithms as the major criteria for mRNA–miRNA interaction prediction.

Recent studies have increased the amount of data describing mRNA–microRNA interactions. One of the first approaches of experimental miRNA target identification assumed immunoprecipitation of miRNA–ribonucleoprotein complexes, isolation and microarray analysis of associated mRNAs (40,51,52). Next-generation of miRNA target discovery approaches are mainly represented with CLIP-seq (or HITS-CLIP) and PAR-CLIP, which assume ultraviolet crosslinking of RNA–protein complexes, immunoprecipitation and sequencing (53,54). These approaches allow scientists to identify various miRNA targets and exact miRNA binding sites in the mRNAs (40). The most recent development is CLASH (crosslinking, ligation and sequencing of hybrids), enabling direct mapping of miRNA–mRNA binding sites without precursory target prediction (55). Using CLASH, frequent (up to 60%) non-canonical seed interactions containing bulged or mismatched nucleotides were found, creating additional challenges for their prediction *in silico* (56). In its current state, CrossHub includes data from miRTarBase covering 21 high-throughput CLIP-seq and one CLASH human datasets (14), representing a valuable addition to bioinformatics-based miRNA target prediction algorithms.

In general, knowledge of microRNA targets allows prediction of their potential functional impacts. In contrast to conventional protein-coding genes, which typically exhibit only a limited set of functions, microRNA effects show large variety and context dependency. Having up to several hundred potential targets, microRNA plays dramatically opposite roles in different cells, tissues and tumors, as described above (49,50,57,58). However, several methods to elucidate the functional impact of microRNAs and highlight the most likely affected signal pathways have been developed. This approach is implemented in the DIANA-miRPath and miRNA body-map resources (42,59). DIANA-miRPath is based on data from DIANA-TarBase, another database of mRNA–miRNA interactions supported with experimental evidence (60). DIANA-miRPath enables Gene Ontology and KEGG pathways-centric evaluation of the functional impact of microRNAs. These tools can facilitate analysis to outline a set of potentially affected pathways.

CrossHub integrates ENCODE ChIP-Seq data, and Jaspar TFBS predictions with TCGA gene expression data (RNA-Seq). This approach, which is very similar to the miRNA target screening analysis implemented, may represent a good method for refining potential TF targets. Similarly to microRNA, TF-mediated expression regulation demonstrates pronounced context dependency, even when a TF targets the same genes. Additionally, most TF are involved in the regulation of several pathways, including many transcription factors, which have up to several thousand potential targets in the human genome (61–64). However, additional analysis is required to highlight gene–TF relationships, which take place in the tissue and have functional impacts resulting in gene expression alteration.

A standard procedure for evaluating TF effects on gene expression and cellular states assumes TF knockdown or TF transfection assays with subsequent gene DE profiling (65–67). Supporting the results of gene DE profiling with ChIP-Seq data significantly improves this approach, allowing us to refine TF target lists predicted only from gene DE profiling after TF knockdown/activation (68). Expression correlation analysis also represents an option, with either gene-TF (69,70) or correlation of expression of multiple genes regulated by a common TF (71,72).

Approximately 2000 potential transcription factors, coactivators, co-repressors and chromatin remodeling complexes are known (73). In the current state, the ENCODE project offers nearly comprehensive ChIP-Seq data for 160 major transcription factors for 3–6 cell lines obtained from normal tissues (HUVEC, GM12878, H1-hESC), cancer (HeLa-S3, HepG2) and leukemia (K562). Gene-TF binding and expression regulation mechanisms are highly context-specific, and ChIP-Seq results cannot be simply extrapolated for any tumor type. While ChIP-Seq data with wide tissue/cell coverage provided information regarding the global presence of TFBS in the genome, the gene expression profiling or correlation analysis performed for the current biological material revealed gene-TF associations with functional impacts prevailing for a particular type of tissue, tumor or cell type (68). However, correlation analysis can provide reliable results only for a large sample; here, TCGA represents an ideal source of expression data, as it contains hundreds of tumor samples for more than 15 cancer types.

ENCODE and Jaspar describe different sets of transcription factors with only partial overlap. Jaspar, as opposed to ENCODE, showed no significant differences in Spearman r_s distribution between low-score and high-score genes. Thus, ENCODE ChIP-Seq data may have more predictive power for TF–gene interactions compared to Jaspar.

Multidimensional genomic approaches outperform one-dimensional approaches in multiple aspects (74,75). High quality, coverage and accessibility of TCGA data has inspired the creation of several integrative tools aimed at analyzing pan-cancer data in the pathway-centric, functional and clinical contexts. Zodiac represents an outstanding tool that enables the prediction of interactions and relationships in cancer. Zodiac uses a global map of known possible genes or protein interactions and refines this data according to likelihood models derived from TCGA data. As a result, Zodiac outlines interactions that likely take place in a particular cancer subtype (76). Like Zodiac, two other well-known pathway-centric approaches SPIA (77) and PARADIGM (78) allow inferring pathway alteration specific for a patient or a tissue basing on the gene expression data. Analysis of perturbed interaction graphs and comparing to the experimental evidences allows uncovering pathways affected by DNA mutations. Such approach is implemented in PINE algorithm (79). Pathway-centric approaches are perfect for the identification of activated interaction subnetworks or interaction graph alterations. In contrast to these tools, CrossHub implements gene-centric approach that is useful to identify unknown regulatory mechanisms of a specific gene set with no need for these genes to be a part of known interaction network.

Multifaceted TCGA data served as a basis for identifying genomic and transcriptomic prognostic and cancer risk signatures (75,80,81). TCGA allowed us to identify potential prognostic markers of breast invasive carcinoma, glioblastoma multiforme, acute myeloid leukemia and lung squamous cell carcinoma. Interestingly, RNA-Seq gene expression profiling was found to have the highest prediction power rather than epigenomic and miRNA expression data: there were no substantial improvements in prediction when adding an additional genomic measurement after gene expression and clinical covariates were included in the model (75). In contrast, when evaluating the expression of individual genes or microRNA, miRNAs showed high prediction values (81–83).

Methylation profiling in the context of gene expression analysis enables identification of CpG sites with a maximal impact on gene expression regulation. It is known that the effect of methylation of various CpG sites across the island is not equal; specific CpG island regions or even single CpG pairs are known to significantly impact gene activity (84–87). Methylation of a single CpG dinucleotide, and to a lesser extent its nearest neighbors, was found to play a crucial role in the expression regulation of protein kinase gene *ZAP-70* involved in T-cell signaling and determining the prognosis of chronic lymphocytic leukemia (88). Similarly, methylation of a single intronic CpG was shown to dramatically affect the expression of peroxisomal membrane protein PMP24. Methylation of this CpG disrupted DNA–

protein interactions and suppressed gene expression (87). In contrast, methylation of CpG island shore regions may contribute to the upregulation of promoter activity and gene expression (89,90). Based on TCGA RNA-Seq data, methylation profiles and ENCODE genome segments annotation, CrossHub enables researchers to outline promoter regions and CpG sites demonstrating maximal contributions to gene expression downregulation. CrossHub reports a direct link to the UCSC genome browser to identify histone methylation patterns and other regulation signatures, and provides methylation-expression correlation and normal-tumor differential methylation tracks that are uploadable to the UCSC browser.

TCGA provides two types of methylation profiling data. First, data derived with methyl-specific microarrays, Illumina BeadChip 27K and much more representative 450K correspond to several distinct regions of CpG islands and some intronic CpG sites. Second, the results of whole genome bisulfite sequencing studies showed single-nucleotide resolution. However, these data are available only for a very limited number of samples covering several cancer types. Although a number of techniques have been developed to enrich DNA libraries with CpG-rich fragments, both targeted CpG-island and whole genome bisulfite sequencing procedures remain laborious and expensive.

CONCLUSIONS

Here we present CrossHub, a tool aimed to outline molecular portrait of a specific gene and elucidate the three most important gene expression regulation axes: promoter or enhancer methylation, microRNA interference and impact of transcription factors. CrossHub uses the combination of gene–TF co-expression analysis with ENCODE ChIP-Seq data to reveal most possible gene–TF interactions with functional impact and the combination of gene–miRNA co-expression analysis with several miRNA target prediction algorithms to uncover most possible gene–miRNA relationships. ENCODE ChIP-Seq data shows greater consistency with expression correlation results compared to Jaspar transcription factor binding site predictions. Similarly, data for conservative miRNA binding sites predicted simultaneously with several algorithms are consistent with the co-expression analysis. This indicates informativeness of these fundamentally different predictors. Use of them in combination improves the accuracy of identification of functional gene–TF or gene–miRNA relationships. CrossHub has a scalable design intended to analyze more various cancer types available in TCGA. This tool may be a starting point for integrating the data of several major projects such as TCGA and ENCODE. The software with databases dumps is freely available at Sourceforge, http://sourceforge.net/p/crosshub/.

ACKNOWLEDGEMENTS

Authors thank the EIMB RAS 'Genome' center (http://www.eimb.ru/RUSSIAN_NEW/INSTITUTE/ccu_genome_c.php), Orekhovich Institute of Biomedical Chemistry and M.M. Shemyakin–Yu.A. Ovchinnikov Institute of Bioorganic Chemistry for the opportunity to use computational resources.

FUNDING

Development and testing of the modules related to DE, TF target predictions and correlation analysis was supported by the Russian Science Foundation grant no. 14-35-00107; Development and testing of miRNA target prediction and methylation modules was supported by the Russian Science Foundation grant no. 14-15-01083. Funding for open access charge: Funds of authors.
Conflict of interest statement. None declared.

REFERENCES

1. Tomczak,K., Czerwinska,P. and Wiznerowicz,M. (2015) The Cancer Genome Atlas (TCGA): an immeasurable source of knowledge. *Contemp. Oncol.*, **19**, A68–A77.
2. Guo,Y., Sheng,Q., Li,J., Ye,F., Samuels,D.C. and Shyr,Y. (2013) Large scale comparison of gene expression levels by microarrays and RNAseq using TCGA data. *PLoS One*, **8**, e71462.
3. Cancer Genome Atlas, N. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.
4. Cancer Genome Atlas, N. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.
5. Chen,Y., McGee,J., Chen,X., Doman,T.N., Gong,X., Zhang,Y., Hamm,N., Ma,X., Higgs,R.E., Bhagwat,S.V. *et al.* (2014) Identification of druggable cancer driver genes amplified across TCGA datasets. *PLoS One*, **9**, e98293.
6. Kim,Y.W., Koul,D., Kim,S.H., Lucio-Eterovic,A.K., Freire,P.R., Yao,J., Wang,J., Almeida,J.S., Aldape,K. and Yung,W.K. (2013) Identification of prognostic gene signatures of glioblastoma: a study based on TCGA data analysis. *Neuro Oncol.*, **15**, 829–839.
7. Ricketts,C.J., Hill,V.K. and Linehan,W.M. (2014) Tumor-specific hypermethylation of epigenetic biomarkers, including SFRP1, predicts for poorer survival in patients from the TCGA Kidney Renal Clear Cell Carcinoma (KIRC) project. *PLoS One*, **9**, e85621.
8. Cho,S., Jang,I., Jun,Y., Yoon,S., Ko,M., Kwon,Y., Choi,I., Chang,H., Ryu,D., Lee,B. *et al.* (2013) MiRGator v3.0: a microRNA portal for deep sequencing, expression profiling and mRNA targeting. *Nucleic Acids Res.*, **41**, D252–D257.
9. Qu,H. and Fang,X. (2013) A brief review on the Human Encyclopedia of DNA Elements (ENCODE) project. *Genomics Proteomics Bioinformatics*, **11**, 135–141.
10. Consortium,E.P. (2011) A user's guide to the encyclopedia of DNA elements (ENCODE). *PLoS Biol.*, **9**, e1001046.
11. Consortium,E.P. (2012) An integrated encyclopedia of DNA elements in the human genome. *Nature*, **489**, 57–74.
12. Ernst,J. and Kellis,M. (2012) ChromHMM: automating chromatin-state discovery and characterization. *Nat. Methods*, **9**, 215–216.
13. Mathelier,A., Zhao,X., Zhang,A.W., Parcy,F., Worsley-Hunt,R., Arenillas,D.J., Buchman,S., Chen,C.Y., Chou,A., Ienasescu,H. *et al.* (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, **42**, D142–D147.
14. Hsu,S.D., Tseng,Y.T., Shrestha,S., Lin,Y.L., Khaleel,A., Chou,C.H., Chu,C.F., Huang,H.Y., Lin,C.M., Ho,S.Y. *et al.* (2014) miRTarBase update 2014: an information resource for experimentally validated miRNA-target interactions. *Nucleic Acids Res.*, **42**, D78–D85.
15. Friedman,R.C., Farh,K.K., Burge,C.B. and Bartel,D.P. (2009) Most mammalian mRNAs are conserved targets of microRNAs. *Genome Res.*, **19**, 92–105.
16. Maragkakis,M., Reczko,M., Simossis,V.A., Alexiou,P., Papadopoulos,G.L., Dalamagas,T., Giannopoulos,G., Goumas,G., Koukis,E., Kourtis,K. *et al.* (2009) DIANA-microT web server: elucidating microRNA functions through target prediction. *Nucleic Acids Res.*, **37**, W273–W276.
17. Betel,D., Koppal,A., Agius,P., Sander,C. and Leslie,C. (2010) Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol.*, **11**, R90.
18. Krek,A., Grun,D., Poy,M.N., Wolf,R., Rosenberg,L., Epstein,E.J., MacMenamin,P., da Piedade,I., Gunsalus,K.C., Stoffel,M. *et al.*

(2005) Combinatorial microRNA target predictions. *Nat. Genet.*, **37**, 495–500.
19. Dang,C.V. (2012) MYC on the path to cancer. *Cell*, **149**, 22–35.
20. Dang,C.V., Le,A. and Gao,P. (2009) MYC-induced cancer cell energy metabolism and therapeutic opportunities. *Clin. Cancer Res.*, **15**, 6479–6483.
21. Krasnov,G.S., Dmitriev,A.A., Snezhkina,A.V. and Kudryavtseva,A.V. (2013) Deregulation of glycolysis in cancer: glyceraldehyde-3-phosphate dehydrogenase as a therapeutic target. *Expert Opin. Ther. Targets*, **17**, 681–693.
22. Smallwood,A., Hon,G.C., Jin,F., Henry,R.E., Espinosa,J.M. and Ren,B. (2012) CBX3 regulates efficient RNA processing genome-wide. *Genome Res.*, **22**, 1426–1436.
23. Sridharan,R., Gonzales-Cope,M., Chronis,C., Bonora,G., McKee,R., Huang,C., Patel,S., Lopez,D., Mishra,N., Pellegrini,M. *et al.* (2013) Proteomic and genomic approaches reveal critical functions of H3K9 methylation and heterochromatin protein-1gamma in reprogramming to pluripotency. *Nat. Cell Biol.*, **15**, 872–882.
24. Saini,V., Hose,C.D., Monks,A., Nagashima,K., Han,B., Newton,D.L., Millione,A., Shah,J., Hollingshead,M.G., Hite,K.M. *et al.* (2012) Identification of CBX3 and ABCA5 as putative biomarkers for tumor stem cells in osteosarcoma. *PLoS One*, **7**, e41401.
25. Xu,X., Xu,L., Gao,F., Wang,J., Ye,J., Zhou,M., Zhu,Y. and Tao,L. (2014) Identification of a novel gene fusion (BMX-ARHGAP) in gastric cardia adenocarcinoma. *Diagn. Pathol.*, **9**, 218.
26. Han,S.S., Kim,W.J., Hong,Y., Hong,S.H., Lee,S.J., Ryu,D.R., Lee,W., Cho,Y.H., Lee,S., Ryu,Y.J. *et al.* (2014) RNA sequencing identifies novel markers of non-small cell lung cancer. *Lung Cancer*, **84**, 229–235.
27. Slezak,J., Truong,M., Huang,W. and Jarrard,D. (2013) HP1gamma expression is elevated in prostate cancer and is superior to Gleason score as a predictor of biochemical recurrence after radical prostatectomy. *BMC Cancer*, **13**, 148.
28. Krasnov,G.S., Dmitriev,A.A., Lakunina,V.A., Kirpiy,A.A. and Kudryavtseva,A.V. (2013) Targeting VDAC-bound hexokinase II: a promising approach for concomitant anti-cancer therapy. *Expert Opin. Ther. Targets*, **17**, 1221–1233.
29. Pal,B., Chen,Y., Bert,A., Hu,Y., Sheridan,J.M., Beck,T., Shi,W., Satterley,K., Jamieson,P., Goodall,G.J. *et al.* (2015) Integration of microRNA signatures of distinct mammary epithelial cell types with their gene expression and epigenetic portraits. *Breast Cancer Res.*, **17**, 85.
30. Senchenko,V.N., Kisseljova,N.P., Ivanova,T.A., Dmitriev,A.A., Krasnov,G.S., Kudryavtseva,A.V., Panasenko,G.V., Tsitrin,E.B., Lerman,M.I., Kisseljov,F.L. *et al.* (2013) Novel tumor suppressor candidates on chromosome 3 revealed by NotI-microarrays in cervical cancer. *Epigenetics*, **8**, 409–420.
31. Dmitriev,A.A., Kashuba,V.I., Haraldson,K., Senchenko,V.N., Pavlova,T.V., Kudryavtseva,A.V., Anedchenko,E.A., Krasnov,G.S., Pronina,I.V., Loginov,V.I. *et al.* (2012) Genetic and epigenetic analysis of non-small cell lung cancer with NotI-microarrays. *Epigenetics*, **7**, 502–513.
32. Gnad,F., Doll,S., Manning,G., Arnott,D. and Zhang,Z. (2015) Bioinformatics analysis of thousands of TCGA tumors to determine the involvement of epigenetic regulators in human cancer. *BMC Genomics*, **16** (Suppl. 8), S5.
33. Wang,Z., Rajaraman,P., Melin,B.S., Chung,C.C., Zhang,W., McKean-Cowdin,R., Michaud,D., Yeager,M., Ahlbom,A., Albanes,D. *et al.* (2015) Further Confirmation of Germline Glioma Risk Variant rs78378222 in TP53 and Its Implication in Tumor Tissues via Integrative Analysis of TCGA Data. *Hum. Mutat.*, **36**, 684–688.
34. Brodie,S.A., Li,G. and Brandes,J.C. (2015) Molecular characteristics of non-small cell lung cancer with reduced CHFR expression in The Cancer Genome Atlas (TCGA) project. *Respir. Med.*, **109**, 131–136.
35. Gore,J., Craven,K.E., Wilson,J.L., Cote,G.A., Cheng,M., Nguyen,H.V., Cramer,H.M., Sherman,S. and Korc,M. (2015) TCGA data and patient-derived orthotopic xenografts highlight pancreatic cancer-associated angiogenesis. *Oncotarget*, **6**, 7504–7521.
36. Zhang,B., Wang,J., Wang,X., Zhu,J., Liu,Q., Shi,Z., Chambers,M.C., Zimmerman,L.J., Shaddox,K.F., Kim,S. *et al.* (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature*, **513**, 382–387.
37. Woo,S., Cha,S.W., Na,S., Guest,C., Liu,T., Smith,R.D., Rodland,K.D., Payne,S. and Bafna,V. (2014) Proteogenomic strategies for identification of aberrant cancer peptides using large-scale next-generation sequencing data. *Proteomics*, **14**, 2719–2730.
38. Cole,C., Krampis,K., Karagiannis,K., Almeida,J.S., Faison,W.J., Motwani,M., Wan,Q., Golikov,A., Pan,Y., Simonyan,V. *et al.* (2014) Non-synonymous variations in cancer and their effects on the human proteome: workflow for NGS data biocuration and proteome-wide analysis of TCGA data. *BMC Bioinformatics*, **15**, 28.
39. Peterson,S.M., Thompson,J.A., Ufkin,M.L., Sathyanarayana,P., Liaw,L. and Congdon,C.B. (2014) Common features of microRNA target prediction tools. *Front. Genet.*, **5**, 23.
40. Ekimler,S. and Sahin,K. (2014) Computational methods for MicroRNA target prediction. *Genes*, **5**, 671–683.
41. Dweep,H., Sticht,C. and Gretz,N. (2013) In-silico algorithms for the screening of possible microRNA binding sites and their interactions. *Curr. Genomics*, **14**, 127–136.
42. Tsang,J.S., Ebert,M.S. and van Oudenaarden,A. (2010) Genome-wide dissection of microRNA functions and cotargeting networks using gene set signatures. *Mol. Cell*, **38**, 140–153.
43. Dweep,H., Sticht,C., Pandey,P. and Gretz,N. (2011) miRWalk–database: prediction of possible miRNA binding sites by "walking" the genes of three genomes. *J. Biomed. Inform.*, **44**, 839–847.
44. Dweep,H., Gretz,N. and Sticht,C. (2014) miRWalk database for miRNA-target interactions. *Methods Mol. Biol.*, **1182**, 289–305.
45. Li,Z., Qin,T., Wang,K., Hackenberg,M., Yan,J., Gao,Y., Yu,L.R., Shi,L., Su,Z. and Chen,T. (2015) Integrated microRNA, mRNA, and protein expression profiling reveals microRNA regulatory networks in rat kidney treated with a carcinogenic dose of aristolochic acid. *BMC Genomics*, **16**, 365.
46. Nam,S., Kim,B., Shin,S. and Lee,S. (2008) miRGator: an integrated system for functional annotation of microRNAs. *Nucleic Acids Res.*, **36**, D159–D164.
47. Gunaratne,P.H., Creighton,C.J., Watson,M. and Tennakoon,J.B. (2010) Large-scale integration of MicroRNA and gene expression data for identification of enriched microRNA-mRNA associations in biological systems. *Methods Mol. Biol.*, **667**, 297–315.
48. Lin,S. and Ding,J. (2009) Integration of ranked lists via cross entropy Monte Carlo with applications to mRNA and microRNA Studies. *Biometrics*, **65**, 9–18.
49. Li,P., Sheng,C., Huang,L., Zhang,H., Huang,L., Cheng,Z. and Zhu,Q. (2014) MiR-183/-96/-182 cluster is up-regulated in most breast cancers and increases cell proliferation and migration. *Breast Cancer Res.*, **16**, 473.
50. Cao,L.L., Xie,J.W., Lin,Y., Zheng,C.H., Li,P., Wang,J.B., Lin,J.X., Lu,J., Chen,Q.Y. and Huang,C.M. (2014) miR-183 inhibits invasion of gastric cancer by targeting Ezrin. *Int. J. Clin. Exp. Pathol.*, **7**, 5582–5594.
51. Easow,G., Teleman,A.A. and Cohen,S.M. (2007) Isolation of microRNA targets by miRNP immunopurification. *RNA*, **13**, 1198–1204.
52. Beitzinger,M., Peters,L., Zhu,J.Y., Kremmer,E. and Meister,G. (2007) Identification of human microRNA targets from isolated argonaute protein complexes. *RNA Biol.*, **4**, 76–84.
53. Hafner,M., Lianoglou,S., Tuschl,T. and Betel,D. (2012) Genome-wide identification of miRNA targets by PAR-CLIP. *Methods*, **58**, 94–105.
54. Hafner,M., Landthaler,M., Burger,L., Khorshid,M., Hausser,J., Berninger,P., Rothballer,A., Ascano,M. Jr, Jungkamp,A.C., Munschauer,M. *et al.* (2010) Transcriptome-wide identification of RNA-binding protein and microRNA target sites by PAR-CLIP. *Cell*, **141**, 129–141.
55. Kudla,G., Granneman,S., Hahn,D., Beggs,J.D. and Tollervey,D. (2011) Cross-linking, ligation, and sequencing of hybrids reveals RNA-RNA interactions in yeast. *Proc. Natl. Acad. Sci. U.S.A.*, **108**, 10010–10015.
56. Helwak,A., Kudla,G., Dudnakova,T. and Tollervey,D. (2013) Mapping the human miRNA interactome by CLASH reveals frequent noncanonical binding. *Cell*, **153**, 654–665.
57. Huhn,D., Kousholt,A.N., Sorensen,C.S. and Sartori,A.A. (2014) miR-19, a component of the oncogenic miR-17

approximately 92 cluster, targets the DNA-end resection factor CtIP. *Oncogene*, **34** , 3977–3984.
58. Li,J., Yang,S., Yan,W., Yang,J., Qin,Y.J., Lin,X.L., Xie,R.Y., Wang,S.C., Jin,W., Gao,F. *et al.* (2015) MicroRNA-19 triggers epithelial-mesenchymal transition of lung cancer cells accompanied by growth inhibition. *Lab. Invest.*, **95**, 1056–1070.
59. Vlachos,I.S., Zagganas,K., Paraskevopoulou,M.D., Georgakilas,G., Karagkouni,D., Vergoulis,T., Dalamagas,T. and Hatzigeorgiou,A.G. (2015) DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res.*, **43**, W460–W466.
60. Vlachos,I.S., Paraskevopoulou,M.D., Karagkouni,D., Georgakilas,G., Vergoulis,T., Kanellos,I., Anastasopoulos,I.L., Maniou,S., Karathanou,K., Kalfakakou,D. *et al.* (2015) DIANA-TarBase v7.0: indexing more than half a million experimentally supported miRNA:mRNA interactions. *Nucleic Acids Res.*, **43**, D153–D159.
61. Guo,Y., Chang,H., Li,J., Xu,X.Y., Shen,L., Yu,Z.B. and Liu,W.C. (2015) Thymosin alpha 1 suppresses proliferation and induces apoptosis in breast cancer cells through PTEN-mediated inhibition of PI3K/Akt/mTOR signaling pathway. *Apoptosis*, **20**, 1109–1121.
62. Maiese,K. (2015) mTOR: Driving apoptosis and autophagy for neurocardiac complications of diabetes mellitus. *World J. Diabetes*, **6**, 217–224.
63. Migliardi,G., Sassi,F., Torti,D., Galimi,F., Zanella,E.R., Buscarino,M., Ribero,D., Muratore,A., Massucco,P., Pisacane,A. *et al.* (2012) Inhibition of MEK and PI3K/mTOR suppresses tumor growth but does not cause tumor regression in patient-derived xenografts of RAS-mutant colorectal carcinomas. *Clin. Cancer Res.*, **18**, 2515–2525.
64. Everett,L.J., Le Lay,J., Lukovac,S., Bernstein,D., Steger,D.J., Lazar,M.A. and Kaestner,K.H. (2013) Integrative genomic analysis of CREB defines a critical role for transcription factor networks in mediating the fed/fasted switch in liver. *BMC Genomics*, **14**, 337.
65. Salgado,M.C., Meton,I., Anemaet,I.G. and Baanante,I.V. (2014) Activating transcription factor 4 mediates up-regulation of alanine aminotransferase 2 gene expression under metabolic stress. *Biochim. Biophys. Acta*, **1839**, 288–296.
66. Hoeth,M., Niederleithner,H., Hofer-Warbinek,R., Bilban,M., Mayer,H., Resch,U., Lemberger,C., Wagner,O., Hofer,E., Petzelbauer,P. *et al.* (2012) The transcription factor SOX18 regulates the expression of matrix metalloproteinase 7 and guidance molecules in human endothelial cells. *PLoS One*, **7**, e30982.
67. Haoues,M., Refai,A., Mallavialle,A., Barbouche,M.R., Laabidi,N., Deckert,M. and Essafi,M. (2014) Forkhead box O3 (FOXO3) transcription factor mediates apoptosis in BCG-infected macrophages. *Cell. Microbiol.*, **16**, 1378–1390.
68. Qin,J., Li,M.J., Wang,P., Zhang,M.Q. and Wang,J. (2011) ChIP-Array: combinatory analysis of ChIP-seq/chip and microarray gene expression data to discover direct/indirect targets of a transcription factor. *Nucleic Acids Res.*, **39**, W430–W436.
69. Campa,V.M., Baltziskueta,E., Bengoa-Vergniory,N., Gorrono-Etxebarria,I., Wesolowski,R., Waxman,J. and Kypta,R.M. (2014) A screen for transcription factor targets of glycogen synthase kinase-3 highlights an inverse correlation of NFkappaB and androgen receptor signaling in prostate cancer. *Oncotarget*, **5**, 8173–8187.
70. Hu,Z., Zhu,L., Tan,M., Cai,M., Deng,L., Yu,G., Liu,D., Liu,J. and Lin,B. (2015) The expression and correlation between the transcription factor FOXP1 and estrogen receptors in epithelial ovarian cancer. *Biochimie*, **109**, 42–48.
71. Marco,A., Konikoff,C., Karr,T.L. and Kumar,S. (2009) Relationship between gene co-expression and sharing of transcription factor binding sites in Drosophila melanogaster. *Bioinformatics*, **25**, 2473–2477.
72. Mahdevar,G., Nowzari-Dalini,A. and Sadeghi,M. (2013) Inferring gene correlation networks from transcription factor binding sites. *Genes Gene. Syst.*, **88**, 301–309.
73. Yamamizu,K., Piao,Y., Sharov,A.A., Zsiros,V., Yu,H., Nakazawa,K., Schlessinger,D. and Ko,M.S. (2013) Identification of transcription factors for lineage-specific ESC differentiation. *Stem Cell Rep.*, **1**, 545–559.
74. Dellinger,A.E., Nixon,A.B. and Pang,H. (2014) Integrative pathway analysis using graph-based learning with applications to TCGA colon and ovarian data. *Cancer Inform.*, **13**, 1–9.
75. Zhao,Q., Shi,X., Xie,Y., Huang,J., Shia,B. and Ma,S. (2015) Combining multidimensional genomic measurements for predicting cancer prognosis: observations from TCGA. *Brief. Bioinform.*, **16**, 291–303.
76. Zhu,Y., Xu,Y., Helseth,D.L. Jr, Gulukota,K., Yang,S., Pesce,L.L., Mitra,R., Muller,P., Sengupta,S., Guo,W. *et al.* (2015) Zodiac: a comprehensive depiction of genetic interactions in cancer by integrating TCGA data. *J Natl. Cancer Inst.*, **107**, djv129.
77. Tarca,A.L., Draghici,S., Khatri,P., Hassan,S.S., Mittal,P., Kim,J.S., Kim,C.J., Kusanovic,J.P. and Romero,R. (2009) A novel signaling pathway impact analysis. *Bioinformatics*, **25**, 75–82.
78. Vaske,C.J., Benz,S.C., Sanborn,J.Z., Earl,D., Szeto,C., Zhu,J., Haussler,D. and Stuart,J.M. (2010) Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. *Bioinformatics*, **26**, i237–245.
79. Wilentzik,R. and Gat-Viks,I. (2015) A statistical framework for revealing signaling pathways perturbed by DNA variants. *Nucleic Acids Res.*, **43**, e74.
80. Braun,R., Finney,R., Yan,C., Chen,Q.R., Hu,Y., Edmonson,M., Meerzaman,D. and Buetow,K. (2013) Discovery analysis of TCGA data reveals association between germline genotype and survival in ovarian cancer patients. *PLoS One*, **8**, e55037.
81. Wong,H.K., Fatimy,R.E., Onodera,C., Wei,Z., Yi,M., Mohan,A., Gowrisankaran,S., Karmali,P., Marcusson,E., Wakimoto,H. *et al.* (2015) The Cancer Genome Atlas Analysis Predicts MicroRNA for Targeting Cancer Growth and Vascularization in Glioblastoma. *Mol. Ther.*, **23**, 1234–1247.
82. Wang,Z., Cai,Q., Jiang,Z., Liu,B., Zhu,Z. and Li,C. (2014) Prognostic role of microRNA-21 in gastric cancer: a meta-analysis. *Med. Sci. Monit.*, **20**, 1668–1674.
83. Dong,Y., Yu,J. and Ng,S.S. (2014) MicroRNA dysregulation as a prognostic biomarker in colorectal cancer. *Cancer Manag. Res.*, **6**, 405–422.
84. Fedorova,M.S., Kudryavtseva,A.V., Lakunina,V.A., Snezhkina,A.V., Volchenko,N.N., Slavnova,E.N., Danilova,T.V., Sadritdinova,A.F., Melnikova,N.V., Belova,A.A. *et al.* (2015) Downregulation of OGDHL expression is associated with promoter hypermethylation in colorectal cancer. *Molecular Biology.*, **49**, 608–617.
85. Loginov,V.I., Dmitriev,A.A., Senchenko,V.N., Pronina,I.V., Khodyrev,D.S., Kudryavtseva,A.V., Krasnov,G.S., Gerashchenko,G.V., Chashchina,L.I., Kazubskaya,T.P. *et al.* (2015) Tumor suppressor function of the SEMA3B gene in human lung and renal cancers. *PLoS One*, **10**, e0123369.
86. Nile,C.J., Read,R.C., Akil,M., Duff,G.W. and Wilson,A.G. (2008) Methylation status of a single CpG site in the IL6 promoter is related to IL6 messenger RNA levels and rheumatoid arthritis. *Arthritis Rheum.*, **58**, 2686–2693.
87. Zhang,X., Wu,M., Xiao,H., Lee,M.T., Levin,L., Leung,Y.K. and Ho,S.M. (2010) Methylation of a single intronic CpG mediates expression silencing of the PMP24 gene in prostate cancer. *Prostate*, **70**, 765–776.
88. Claus,R., Lucas,D.M., Stilgenbauer,S., Ruppert,A.S., Yu,L., Zucknick,M., Mertens,D., Buhler,A., Oakes,C.C., Larson,R.A. *et al.* (2012) Quantitative DNA methylation analysis identifies a single CpG dinucleotide important for ZAP-70 expression and predictive of prognosis in chronic lymphocytic leukemia. *J. Clin Oncol.*, **30**, 2483–2491.
89. Rao,X., Evans,J., Chae,H., Pilrose,J., Kim,S., Yan,P., Huang,R.L., Lai,H.C., Lin,H., Liu,Y. *et al.* (2013) CpG island shore methylation regulates caveolin-1 expression in breast cancer. *Oncogene*, **32**, 4519–4528.
90. Bockmuhl,Y., Patchev,A.V., Madejska,A., Hoffmann,A., Sousa,J.C., Sousa,N., Holsboer,F., Almeida,O.F. and Spengler,D. (2015) Methylation at the CpG island shore region upregulates Nr3c1 promoter activity after early-life stress. *Epigenetics*, **10**, 247–257.