GeneGazer: A Toolkit Integrating Two Pipelines for Personalized Profiling and Biosignature Identification

JI-DUNG LUO[1*], YU-JIA CHANG[2,5,6,7*], CHUNG-MING CHANG[3], JENG-FU YOU[4], PO-LI WEI[5,6,7] and CHIUAN-CHIAN CHIOU[1,3]

[1]*Graduate Institute of Biomedical Sciences, College of Medicine, Chang Gung University, Taoyuan, Taiwan, R.O.C.;*
[2]*Graduate Institute of Clinical Medicine, College of Medicine, Taipei Medical University, Taipei, Taiwan, R.O.C.;*
[3]*Department of Biotechnology and Laboratory Science, College of Medicine, Chang Gung University, Taoyuan, Taiwan, R.O.C.;*
[4]*Division of Colonic and Rectal Surgery, Department of Surgery, Chang Gung Memorial Hospital, Linkou, Taiwan, R.O.C.;*
[5]*Division of General Surgery, Department of Surgery, Taipei Medical University Hospital, Taipei, Taiwan, R.O.C.;*
[6]*Department of Surgery, College of Medicine, Taipei Medical University, Taipei, Taiwan, R.O.C.;*
[7]*Cancer Research Center, Taipei Medical University Hospital, Taipei, Taiwan, R.O.C.*

Abstract. *Background: Next-generation sequencing provides useful information about gene mutations, gene expression, epigenetic modification, microRNA expression, and copy number variations. More and more computing tools have been developed to analyze this large quantity of information. However, to test and find suitable analytical tools and integrate their results is tedious and challenging for users with little bioinformatics training. In the present study, we assembled the computing tools into a convenient toolkit to simplify the analysis and integration of data between bioinformatics tools. Materials and Methods: The toolkit, GeneGazer, comprises of two parts: the first, named Gaze_Profiler, was designed for personalized molecular profiling from next-generation sequencing data of paired samples; the other, named Gaze_BioSigner, was designed for the discovery of disease-associated biosignatures from expressional and mutational profiles of a cohort study. Results: To demonstrate the capabilities of Gaze_Profiler, we analyzed a pair (colon cancer and adjacent normal tissues) of RNA-sequencing data from one patient downloaded from the Sequencing Read Archive database and used them to profile somatic mutations and digital gene expression. In this case, alterations in the RAS/RAF/MEK/ERK signaling pathway (activated by KRAS G13D mutation) and canonical WNT signaling pathway (activated by truncated APC) were identified; no EGFR mutation or overexpression was found. These data suggested a limited efficacy of cetuximab in the patient. To demonstrate the ability of Gazer_BioSigner, we analyzed gene-expression data from 192 cancer tissues downloaded from The Cancer Genome Atlas and found that the activation of cAMP/PKA signaling, OCT-3/4 and SRF were associated with colon cancer progression and could be potential therapeutic targets. Conclusion: GeneGazer is a reliable and robust toolkit for the analysis of data from high-throughput platforms and has potential for clinical application and biomedical research.*

*These Authors contributed equally to this work.

Abbreviations: cAMP, Cyclic adenosine monophosphate; APC, adenomatous polyposis coli; BIRC5, baculoviral IAP repeat containing 5; BRAF, B-RAF proto-oncogene, serine/threonine kinase; EGFR, epidermal growth factor receptor; ERK, extracellular signal-regulated kinase; FZD7, frizzled class receptor 7; IGF1R, insulin-like growth factor 1 receptor; KRAS, kirsten rat sarcoma viral oncogene homolog; MEK, mitogen-activated protein kinase kinase; MET, Met proto-oncogene, receptor tyrosine kinase; NF-Y, nucleic factor Y; OCT-3/4, octamer-binding transcription factor 3/4; PIK3CA, phosphotidylinositol-4,5-bisphosphate 3-kinase, catalytic subunit alpha; PKA, cAMP-dependent protein kinase; PTEN, phosphatase and tensin homolog; RAF, v-Raf murine sarcoma viral oncogene; SFRP5, secreted frizzled-related protein 5; SPRY2, sprouty homolog 2; SRF, serum response factor; RAS, rat sarcoma viral oncogene; TBP, TATA-Box binding protein; WNT, wingless-type MMTV integration site family.

Correspondence to: Dr. Chiuan-Chian Chiou, Department of Medical Biotechnology and Laboratory Science, College of Medicine, Chang Gung University, No.259, WenHua 1st road, KweiShan, TaoYuan, 333, Taiwan, R.O.C. Tel: +886 32118800 ext. 5204, Fax: +886 32118035, e-mail: ccchiou@mail.cgu.edu.tw

Key Words: Next-generation sequencing, personalized molecular profiling, biosignature, somatic mutations, digital gene expression.

High-throughput technologies, such as next-generation sequencing (NGS) and high-density microarrays, provide systematic information on molecular biological processes and lead to a comprehensive understanding of cancer genomes (1-3). Knowledge acquired with these sophisticated tools indicates that genotypes or expression profiles in cancer tissues are highly diverse among patients (4, 5). This molecular diversity leads to the activation of different pathways and cancer phenotypes, which subsequently affect therapeutic efficacy and prognosis. For instance, tumors with high *EGFR* expression are susceptible to cetuximab, the therapeutic antibody commonly used against metastatic colorectal cancer. Once a cancer cell acquires either mutation of *KRAS*, *BRAF* or *PIK3CA* mutation, *PTEN* loss, or overexpression of *MET* or *IGF1R*, it develops resistance to cetuximab by activating a signaling pathway other than EGFR (6). In the case of EGFR-independent tumors, cetuximab has limited therapeutic efficacy, and medications targeting other signaling pathways should be considered (7, 8). Therefore, the molecular profiling of genetic variations and gene expression for each patient would be useful for selecting for the most appropriate therapeutic approach (9-11).

NGS is a powerful tool for the molecular profiling of gene mutation, expression, methylation and microRNA profiles. More and more useful computing tools and pipelines have been developed for the analysis of NGS data (12). For instance, the BWA-GATK pipeline is often used for variant identification (13-15); Tophat-Cufflinks pipeline is widely applied to calculate the digital gene-expression profile (16); and VarScan (17, 18) and Somaticsnipper (19) were developed for identifying somatic variations. Most of these tools have been developed for text-based user interface and for the UNIX system only. Many commercial software kits, such as CLC Genome Workbench, Avadis, and Partek, provide a user-friendly graphic user interface and are available for various types of operating systems. These tools are powerful but often inflexible. In addition, using these programs and selecting for appropriate parameters is still a great challenge, especially for clinical technicians without bioinformatics training.

In addition to molecular profiling, high-throughput techniques also generate large amounts of data. Many archive databases, such as Sequence Read Archive (SRA), Gene Expression Omnibus (GEO), and The Cancer Genome Atlas (TCGA), have been established for storing omics data generated from these high-throughput techniques. These data provide important clues regarding disease-progression mechanisms, drug development, biosignature identification and biomarker discovery. For instance, general biosignatures in glioblastoma multiforme (20), ovarian carcinoma (21), colorectal adenocarcinoma (22), and squamous cell lung cancer (23) have been uncovered using the TCGA database. Many strategies, such as PARADIGM (24) and MuSiC (25), have been developed for identifying general disease-causing signatures. These observations indicate that mining information from such a massive data deposit would be useful in order to gain a comprehensive understanding of malignant diseases.

In the present study, we attempted to establish an easy and integrated toolkit that assembles bioinformatics resources for analyzing sequencing data and generating molecular profiles of the samples. The toolkit, named *GeneGazer*, comprises of two computing pipelines: the first one is for personalized molecular profiling through analysis of NGS data, and the second one is for biosignature identification (Figure 1). To demonstrate the capability of the toolkit, an RNA-sequencing dataset from SRA and a large dataset from TCGA were analyzed.

Materials and Methods

Resource requirements. The source codes for *GeneGazer* are available on request. The toolkit was developed and tested on Linux Ubuntu 12.04 operating system with the following software: *Burrows-Wheeler Aligner* (*BWA*, 0.6.2-r126), *SAMtools* (0.1.8), *TopHat* (2.0.3), *Bowtie* (0.12.7), *Cufflinks* (1.3.0), *MySQL* (5.5.35), and *JAVA* compiler. For *Gaze_Profiler*, at least two threads and four gigabytes of RAM are recommended. For *Gaze_BioSigner*, at least one thread and four gigabytes of RAM are required.

RNA-sequencing dataset. The RNA sequence dataset CC2 (SRP021221) was downloaded from the Sequence Reads Archive (SRA) database of the National Center of Biotechnology and Information (NCBI) database. This dataset consists of 59.2 million 100-bp reads from each tumor section and 35.4 million 75-bp reads from the adjacent normal section. All reads were paired-end sequences. The raw data were stored in the *raw_data* folder in *fastq* format for subsequent analysis.

Somatic mutation identification. This analysis was performed using *Gaze_Profiler*. The workflow is illustrated in Figure 2A. Firstly, five nucleotides from the 5'-end and 10 nucleotides from the 3'-end were trimmed. The trimmed data were stored in the *seq_data* folder. The data were aligned against human genome DNA version 19 (hg19, UCSC) using BWA (14, 15). The alignment result was stored in the *sam* format. The sam format data were sorted using *Picard-Utility* (http://broadinstitute.github.io/picard). The duplicated reads were removed with *Picard-Utility*. Indel realignment and quality recalculation were performed with *Genome analytic toolKit* (*GATK*) (13). Before analysis using *VarScan* (17, 18), the pileup files are generated with mpileup module in *SAMtools* (26). For scanning somatic variations, (i) the depth of sequencing of normal tissue should be greater than six; (ii) the depth of sequencing of tumor tissue should be greater than three; and (iii) the total depth of sequencing of normal and tumor tissue should be at least six. The significance of each somatic variant calling was evaluated by Fisher's exact test. The variants with p-value smaller than 0.05 were identified as somatic variants or loss of heterozygocity. The selected variants were subsequently annotated with *Annovar* (27, 28). The genes with missense, nonsense, or frameshift mutations were collected into an SQL database.

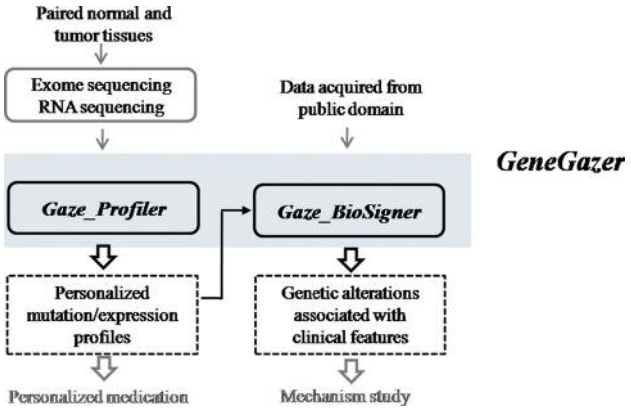

Figure 1. *The organization of GeneGazer.*

Geneexpression profile. "*Exp*" subroutine in *Gaze_Profiler*, based on the *TopHat-Cufflinks* pipeline (16), was applied to evaluate the gene-expression profile from the RNA-sequencing dataset. The trimmed sequencing data were aligned to GRCh37 (Ensemble) with *TopHat* (16). The alignment data were subsequently analyzed with *Cufflinks* (16). The read counts for each gene locus were calculated and normalized in fragments per kilobase of transcript per million mapped reads (FPKM) format. The genes whose expression was two-fold greater or lower in the tumor than in the paired normal tissue were selected as being up-/down-regulated and were collected into the *SQL* database.

Processing of data obtained from The Cancer Genome Atlas (TCGA) database. The expression profiles of patients with colon adenocarcinoma, including COAD.IlluminaHiSeq_RNASeqV2 (level 3.1.9.0) and COAD_IlluminaGA_RSEQV2 (level 3.1.1.0) were obtained from the TCGA database (https://tcga-data.nci.nih.gov/tcga/). The expression of each gene was centralized by Z-score transformation. The average and standard deviations of each gene in normal colon tissues were calculated. The Z-score of each target gene in tumor tissue was computed according to the equation: $Z\text{-score}_{tar}=(Expression_{tar}-Average_{tar})/SD_{tar}$, whereas $Expression_{tar}$ represented the level of expression of the target gene, $Average_{tar}$ represented the average target gene expression level in normal colon tissue, and SD_{tar} represented the standard deviation of the target gene expression level in normal tissue. All calculations were performed using the subroutine "*zscore*". The COAD_IlluminaGA_RSEQV2 (level 3.1.1.0) dataset was used for the following Gene-Set Enrichment Analysis (GSEA) (29).

Gene-set enrichment analysis. For GSEA, the required files were generated from the FPKM file. The phenotype data were recorded in the cls file; gene-annotation data were described in the chip file; and the expression data were in the gct file. All files were analyzed by *GSEA* program. The gene-set database was based on the curated Kyoto Encyclopedia of Genes and Genomes (KEGG) database (c2.cp.kegg.v5.0.symbols.gmt). The permutation type was set in "Gene Set". The default settings of the other parameters were used.

Transcription factor analysis. Genes with differential expression in distinct groups of patients were identified by the subroutine "*diff_zscore*". The gene list was imported into *MetaCore™* (THOMSON REUTERS, New York, NY, USA) for analysis of the transcription factors. The selected pathways were then integrated.

Results

Implementation. GeneGazer utilizes two automatic pipelines, both of which are encoded in two shell scripts and controlled by text-based user-interface in the Linux operating environment.

Gaze_Profiler is designed for identifying somatic genetic alterations from paired NGS data from a tumor and its normal tissue counterpart. The pipeline includes *BWA, Picard-tools, GATK, SAMtools, VarScan, Annovar, Tophat, Cufflinks*, and *GSEA*, as shown in Figure 2A. FASTQ-format RNA-sequencing or exome-sequencing data from paired normal and tumor samples from one patient are analyzed in this pipeline. Initially, low-quality base-calling at both the 5'-end and 3'-end of reads was trimmed by the subroutine "*prep*". To survey somatic mutations, the subroutine "*mut*" is applied to map the trimmed reads to the human genome (hg19, UCSC) using BWA. The alignment results are sorted and transformed into bam format with *Picard-tools*, and the duplicated reads are discarded. The realignment of insertion/deletion and the recalculation of mapping quality are processed with *GATK*. Thereafter, somatic variations are identified with *Varscan*, based on Fisher's exact test. These somatic variations are annotated with *Annovar*. Finally, the missense, nonsense, or frameshift mutations are extracted into a personalized mutational profile, which is subsequently imported into *MySQL* database. For such a paired sequencing dataset with 10 gigabytes of data output for each of the paired samples, the entire process can be completed within 24 hours using two threads and four gigabytes of RAM. The subroutine "*exp*" is designed to calculate differential gene expression, which is based on the *Tophat-Cufflinks* pipeline. The results are presented in FPKM format (16). The gene sets associated with tumor sections are selected using *GSEA*, and the differentially expressed genes, either overexpressed or down-regulated in the tumor section, are identified and stored in *MySQL* database for subsequent analysis. For such a sequencing dataset with 6 gigabyte data output for each of the paired samples, the generation of a complete differential gene expression profile still takes under 24 h with two threads and four gigabytes of RAM.

Gaze_BioSigner is a pipeline containing subroutines for data management and biosignature identification. The input data are genetic variations or gene expression in a cohort, obtained either from *Gaze_Profiler* or public domains. Its workflow is shown in Figure 2B. Molecular profiles from *Gaze_Profiler* are imported into *MySQL* database through the subroutine "*import*". The expression data from TCGA are imported through the subroutine "*zscore*". The clinical data

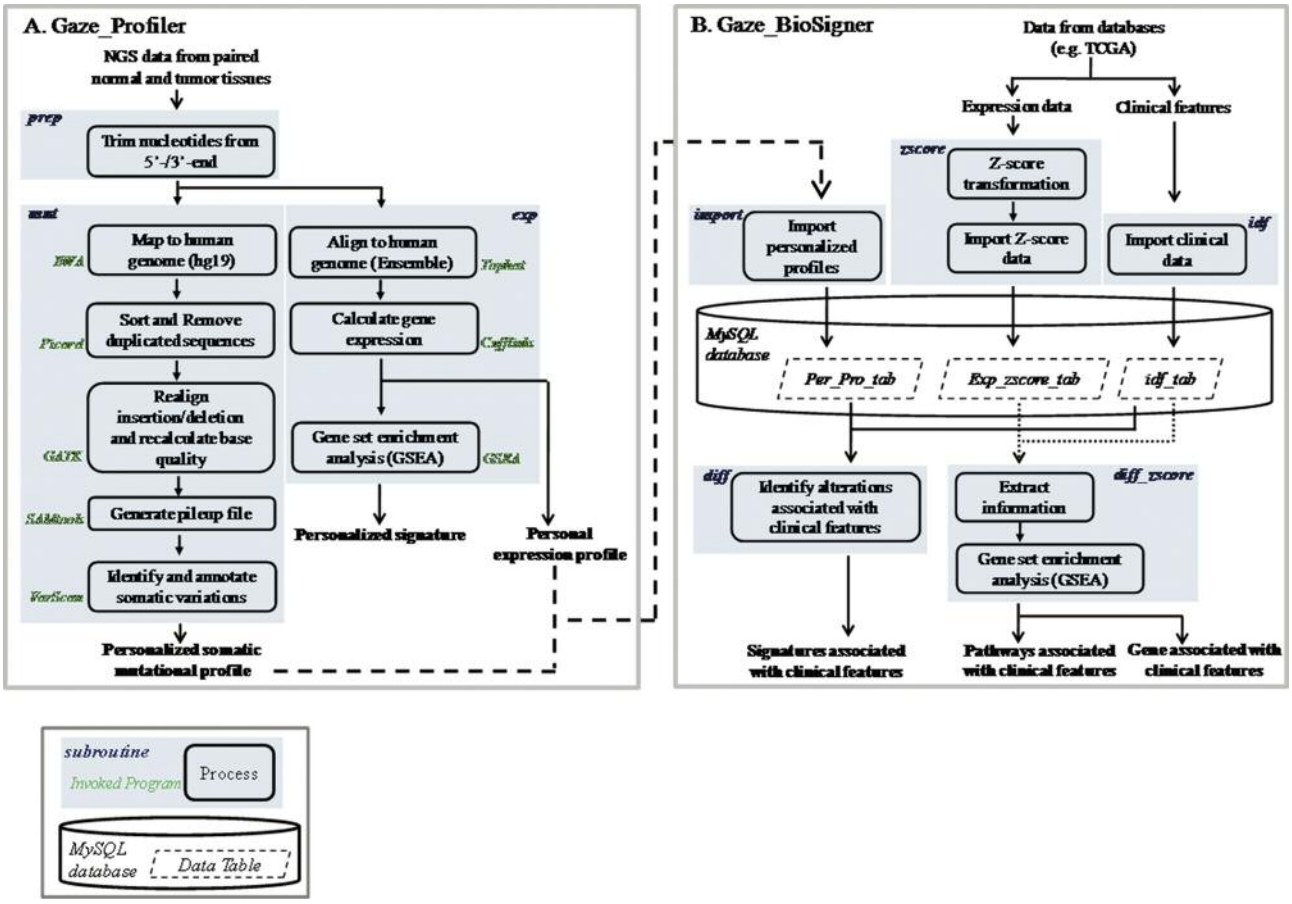

Figure 2. *The workflow of GeneGazer. The workflow of the two pipelines in GeneGazer, Gaze_Profiler (A) and Gaze_BioSigner (B), is shown. Blue squares represent subroutines each consisting of one or more invoked programs (identified in green).*

of each patient are stored in the database by the subroutine "*idf*". The associations between the genetic alterations and selected clinical features are calculated with the subroutine "*diff*", analyzed by Pearson's Chi-square test and odds ratio. Expression profiles can be further entered into subroutine "*diff_zscore*", which invokes *GSEA* and identifies the gene sets associated with selected clinical features.

Personalized molecular profiling. To demonstrate the capability of *Gaze_Profiler*, we analyzed a paired (normal and tumor tissues) RNA-sequencing dataset from a patient with colon cancer, downloaded from the SRA database in NCBI. Mutational and expressional profiles of this dataset were generated with the *Gaze_Profiler* pipeline.

In this dataset, 70 somatic mutations were identified, including 18 missense substitutions, four nonsense substitutions and 48 frameshifts caused by insertions or deletions. Moreover, there were 108 genes up-regulated and 172 down-regulated in the tumor section. *GSEA* revealed that eight pathways were up-regulated and 22 pathways down-regulated in the tumor section (Figure 3A-C). Genes associated with these pathways were involved in the cell cycle, DNA repair, cytokine production, and cell adhesion. These results suggested that the tumor tissue in this patient had a phenotype of cell-cycle acceleration, DNA-repair activation, cytokine reduction and cell adhesion dysfunction.

Gaze_Profiler can extend its capability through utilization of other analytical tools not part of our pipeline. For example, to further investigate the pathways in which the genetic alterations occurred, the somatic genetic alterations for each patient were entered into *MetaCore™*, a comprehensive pathway analysis software. The results show that in the tumor tissue, a heterozygous KRAS G13D mutation and SPRY2 overexpression were identified, suggesting that the tumor tissue harbored the KRAS-driven activation of the RAS/RAF/MEK/ERK pathway. Moreover, the tumor also showed activation of the canonical WNT pathway, most likely due to the dysfunction of APC and SFRP5, activation of FZD7, and overexpression of BIRC5 (Figure 3D).

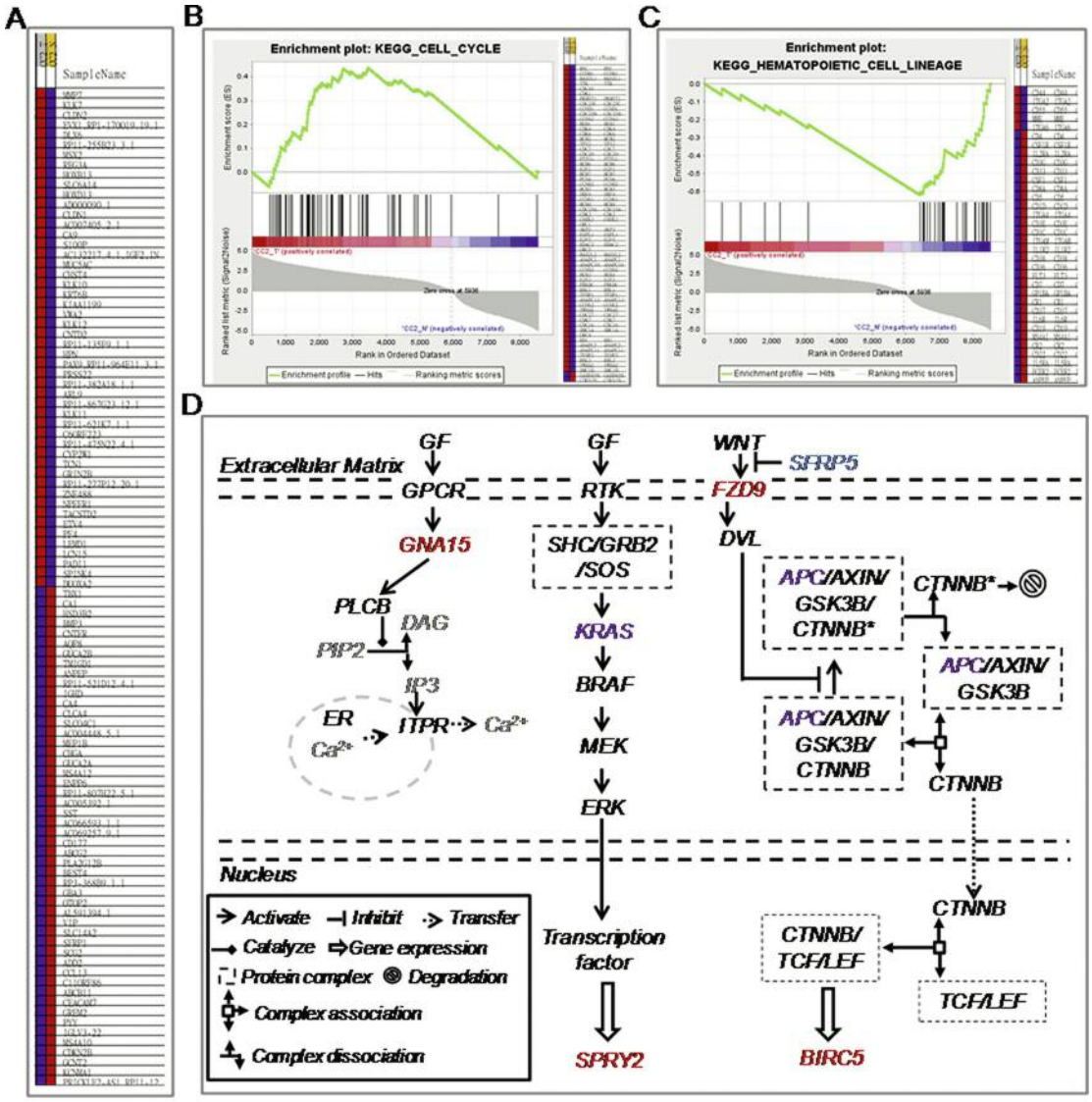

Figure 3. *The personal molecular profile of patient CC2 was analyzed by Gene-Set Enrichment Analysis (GSEA) and pathway analysis. This figure shows the top 20 differentially expressed genes (A), up-regulated gene sets (B), and down-regulated gene sets (C) identified by GSEA in the tumor section. After pathway analysis, activation of the canonical WNT signaling pathway and KRAS/RAF/MEK/ERK signaling pathway was identified in this patient (D). The genes up-regulated in the tumor are labeled in red; those down-regulated in the tumor in blue; non-synonymous somatic mutations in purple.*

Biosignature identification. To demonstrate the capability of *Gaze_BioSigner*, we downloaded the expression profiles of colon adenocarcinoma from the TCGA database (COAD_IlluminaGA_RSEQV2 level 3.1.1.0) and analyzed the molecular profiles associated with certain clinical features. This dataset includes 459 patients; among them, 192 patients with record of the tumor stage (111 at an early stage or stage I and II, and 81 at a late stage or stage III and IV). This dataset was used to identify possible genetic alterations that were associated with tumor stage. The expression profile of each patient was converted into a Z-score profile with the subroutine "*zscore*". The Z-score profile was transferred into another subroutine "*diff_zscore*", which invokes GSEA to identify stage-associated biosignatures.

The GSEA results are shown in the supplementary information (https://drive.google.com/open?id=0B3Nd1wE1_lrRaTl3NnFJTnRXWTA). Briefly, 98 pathways were up-regulated in patients with late-stage tumor. Only the gene set KEGG_Vasopression_regulated_water_absorption was significantly enriched. The genes in its core enrichment were

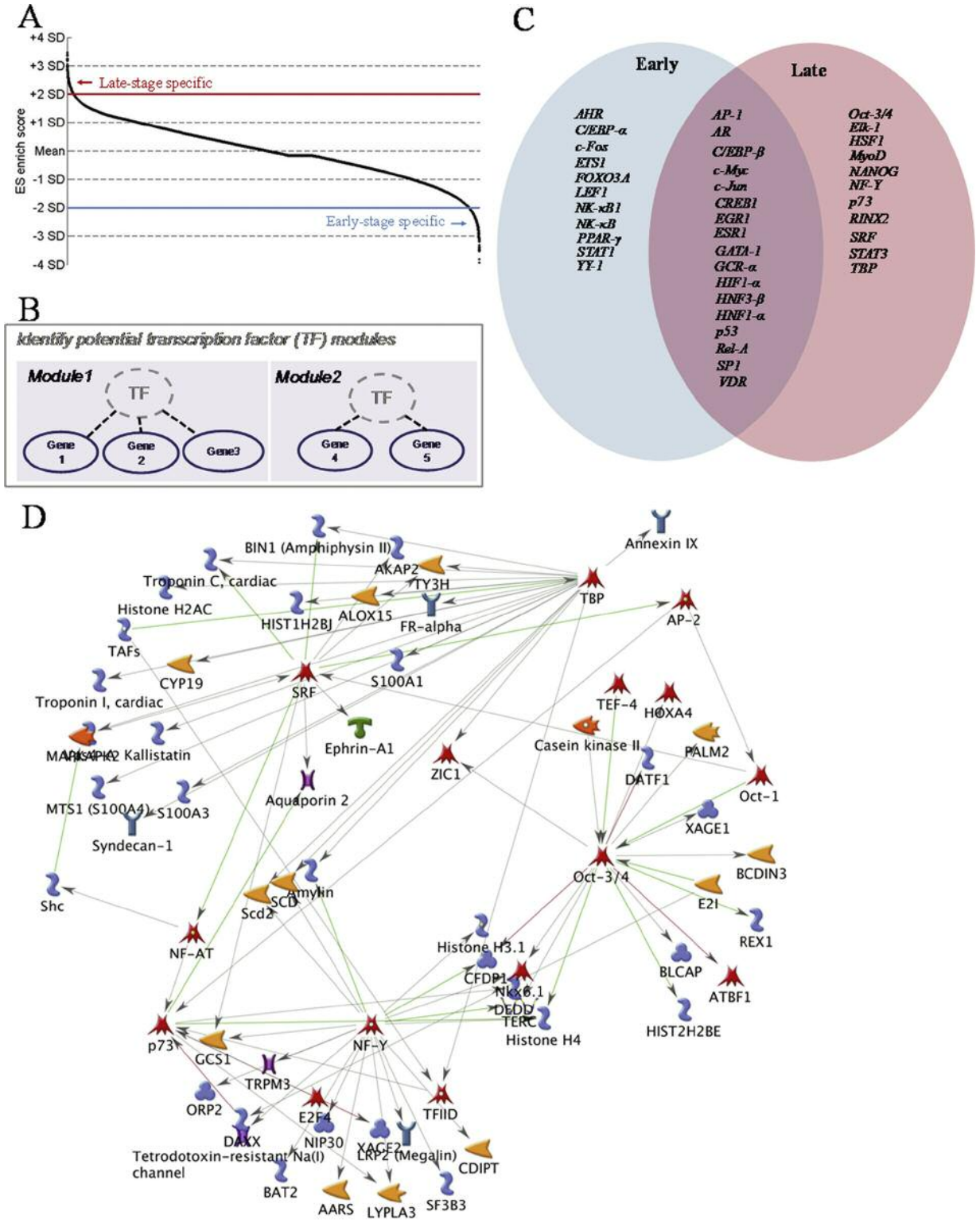

Figure 4. *Transcription factors activated in early- and late-stage colon adenocarcinoma. This figure describes the strategy for identifying the biosignatures associated with late-stage colon adenocarcinoma. The enriched score of each gene was calculated using Gene-Set Enrichment Analysis (GSEA) (A). The stage-specific gene expression was identified and could be traced back to transcription factors (B). The transcription factors activated in patients with late-stage or early-stage disease are shown (C). The network of late-stage-specific transcription factors and their downstream genes were plotted by Metacore™ (D). The results demonstrate that SRF, OCT3/4, p73, NF-Y and TBP are activated in late-stage colon adenocarcinomas.*

associated with cAMP/PKA signal transduction. This result suggested that cAMP/PKA activation could be important for tumor progression. However, 119 gene sets were up-regulated in patients with early-stage disease. Three of them were significantly enriched, including KEGG_p53_signaling_pathway, KEGG_Apoptosis, and KEGG_NOD_like_receptor_signaling_pathway. The genes in the core enrichment of these gene sets were associated with p53-dependent cell-cycle arrest, apoptosis, and immune response. These results indicated that maintaining the activation of these genes could be crucial to restricting tumor development.

In addition to pathway analysis with *GSEA*, we further surveyed stage-specific transcription factor activation according to the following strategies not involved in *GeneGazer*. Firstly, we extracted the enrich score (ES) of each gene, that was calculated in GSEA analysis, and calculated the average ($Mean_{ES}$) and standard deviation (SDES) of the ES. The genes with an ES larger than $Mean_{ES} + 2 \times SD_{ES}$ were defined as being related to late-stage disease. In contrast, the genes with an ES less than $Mean_{ES} - 2 \times SD_{ES}$ were defined as being related to early-stage disease (Figure 4A). We hypothesized that these differentially expressed genes were the result of altered transcription factors crucial for tumor progression (Figure 4B). These transcription factors can be identified using MetaCore™ . As shown in Figure 4C, 12 transcription factors were activated in late-stage, 12 were activated in early-stage, and 17 were activated throughout tumor progression. Combining the 12 late-stage-related transcription factors, the transcription factors involved in tumor progression included SRF, p73, NF-Y,TBP, and OCT3/4.

Discussion

In this study, we developed a computing toolkit, GeneGazer, which integrates two pipelines, *Gaze_Profiler* and *Gaze_BioSigner*, designed to discover personalized molecular profiles and the biosignatures associated with clinical features, respectively. All analyses were automatically processed after assigning the required parameters via a simple text-based user interface, and the results were output as format-ready for further analysis with other software. We used an RNA-sequencing dataset from a tumor tissue and its paired normal tissue to demonstrate the capability of *Gaze_Profiler*. The pipeline identified activation of the RAS/RAF/MEK/ERK pathway and canonical WNT pathway in the tumor tissue. In addition, we used mutation and expression datasets of colon cancer in TCGA to demonstrate the capability of *Gaze_BioSigner*. The pipeline identified that the activation of SRF, OCT-3/4, NF-Y, p73, and TBP was highly associated with the progression of colon adenocarcinoma. These results suggested that *GeneGazer* is a useful and reliable toolkit for biomedical studies.

The features of GeneGazer. As a user-friendly and reliable toolkit, *GeneGazer* offers the following features: (i) the integration of computing tools in a simple user interface; (ii) easy interface with other out-sourcing toolkits; and (iii) reliable strategies for identification of somatic variation. *GeneGazer* integrates many well-established computing tools into a simple text-based user interface. Once the necessary parameters are assigned, the personalized molecular profiling or biosignature discovery will be automatically processed. The results from *GeneGazer* are ready to be analyzed with other computing tools that are not incorporated in the pipeline. For example, to annotate each variant and estimate its effect at the protein level, the somatic mutations identified by *Gazer_Profiler* were stored in a vcf4-format text file, which is directly accessible to VarioWatch (http://genepipe.ncgm.sinica.edu.tw/variowatch/main.do) (30). In addition, the personalized molecular profile is directly available to DAVID Bioinformatics Resources (http://david.abcc.ncifcrf.gov) (31, 32) and *MetaCore™* . With these two bioinformatics sources, the pathways associated with genetic alterations can be identified. This information provides an insight into the mechanisms of these genetic alterations.

To identify somatic variations, many computing tools are available, such as *MutTect* in *GATK* (13), mpileup in *SAMtools* (26), *Somaticsniper* (19), and *VarScan* (17, 18). The first three are based on genotype likelihood and the latter, *VarScan*, is based on Fisher's exact test. We chose to integrate *VarScan* in our pipeline as it is more conservative and reliable for paired samples with different read counts or insufficient sequencing depth in some alleles, which is a common occurrence in RNA-sequencing data.

Application of GeneGazer in personalized medicine. Personalized molecular profiling generated by GeneGazer can provide useful information for personalized medicine, therapeutic approach, and prognosis. In the example of the CC2 patient, the canonical WNT and RAS/RAF/MEK/ERK signaling pathway was activated; no EGFR overexpression, *BRAF* or *PIK3CA* mutation, nor *PTEN* loss were identified. This profile suggested that the CC2 patient is very likely to be unresponsive to therapy with cetuximab, a commonly used therapy targeting EGFR in colon cancer (33). Instead, tankyRASe inhibitor (for blocking the canonical WNT signaling pathway) (34) or MEK inhibitor (for blocking the RAS/RAF/MEK/ERK pathway) (35) may be an alternative therapeutic choice. More and more information on drug–protein interactions is being gathered (36), and computational modeling has been applied to evaluate the effects of genetic alterations on phenotypic changes (37). By combining this information, personalized molecular profiling can suggest reasonable therapeutic strategies for individual patients.

Application of GeneGazer in cancer research. By using *Gaze_BioSigner*, we identified late-stage-specific biosignatures from colon adenocarcinoma data in the TCGA database. Activated cAMP/PKA pathway, p53-mediated apoptosis and immune response were associated with late-stage disease. The role of p53 and immune response in colon cancer has been well-documented, but the activation of the cAMP/PKA pathway is less well-understood. The activation of cAMP/PKA pathway has been found in many types of cancer (38-40), including of the stomach (38), prostate (41, 42), thyroid (41), and breast (43-45). Our finding suggests that the cAMP/PKA pathway is also important in late-stage colon cancer.

In addition, five transcription factors, namely SRF, p73, NF-Y, TBP and OCT-3/4, have been linked to late-stage colon cancer based upon the results of the *Gazer_BioSigner* pipeline and *MetaCore™* . SRF is an important regulator of tumor progression. Its activation has been identified in prostate (46-48), lung (49), and gastric (50) cancer, hepatocellular carcinoma (51, 52), and papillary thyroid cancer (53). OCT-3/4 has been shown to accelerate prostate cancer progression and aggressiveness (54). NF-Y has been found to support tumor proliferation and progression in many types of cancers, including colon cancer (55). The p53 family protein p73 has been shown to act as a tumor suppressor in many types of cancers (56); however, its activation is highly associated with colon cancer progression (57). TBP is associated with ovarian cancer progression (58). These results suggest that these transcription factors are involved in colon cancer progression and may be potential therapeutic targets.

Conclusion

In the present study, we demonstrated that *GeneGazer* is a reliable and robust toolkit not only for personalized profiling but also for biosignature discovery. This information could be useful for personalized medication and mechanism study. Thus, GeneGazer could have potential for clinical application and biomedical research.

Acknowledgements

The Authors would like to acknowledge Professor Petrus Tang and Professor Kwan-Wu Chen from Chang Gung University and Professor Yuen-Shin Lin from National Taiwan Chiao-Tung University for kindly giving advice and making suggestions. This work was supported by the Chang Gung Molecular Medicine Research Center, Ministry of Education (grant EMRPD 170291), and Taipei Medical University.

References

1 Young RA: Biomedical discovery with DNA arrays. Cell *102*: 9-15, 2000.

2 Pfeifer GP and Hainaut P: Next-generation sequencing: emerging lessons on the origins of human cancer. Curr Opin Oncol *23*: 62-68, 2011.

3 Katsios C, Papaloukas C, Tzaphlidou M and Roukos DH: Next-generation sequencing-based testing for cancer mutational landscape diversity: clinical implications? Expert Rev Mol Diagn *12*: 667-670, 2012.

4 Wood LD, Parsons DW, Jones S, Lin J, Sjoblom T, Leary RJ, Shen D, Boca SM, Barber T, Ptak J, Silliman N, Szabo S, Dezso Z, Ustyanksky V, Nikolskaya T, Nikolsky Y, Karchin R, Wilson PA, Kaminker JS, Zhang Z, Croshaw R, Willis J, Dawson D, Shipitsin M, Willson JK, Sukumar S, Polyak K, Park BH, Pethiyagoda CL, Pant PV, Ballinger DG, Sparks AB, Hartigan J, Smith DR, Suh E, Papadopoulos N, Buckhaults P, Markowitz SD, Parmigiani G, Kinzler KW, Velculescu VE and Vogelstein B: The genomic landscapes of human breast and colorectal cancers. Science *318*: 1108-1113, 2007.

5 Wood LD: Pancreatic cancer genomes: toward molecular subtyping and novel approaches to diagnosis and therapy. Mol Diagn Ther *17*: 287-297, 2013.

6 Therkildsen C, Bergmann TK, Henrichsen-Schnack T, Ladelund S and Nilbert M: The predictive value of KRAS, NRAS, BRAF, PIK3CA and PTEN for anti-EGFR treatment in metastatic colorectal cancer: A systematic review and meta-analysis. Acta Oncol *53*: 852-864, 2014.

7 Schmoll HJ and Stein A: Colorectal cancer in 2013: Towards improved drugs, combinations and patient selection. Nat Rev Clin Oncol *11*: 79-80, 2014.

8 Stein A and Bokemeyer C: How to select the optimal treatment for first line metastatic colorectal cancer. World J Gastroenterol *20*: 899-907, 2014.

9 Gonzalez de Castro D, Clarke PA, Al-Lazikani B and Workman P: Personalized cancer medicine: molecular diagnostics, predictive biomarkers, and drug resistance. Clin Pharmacol Ther *93*: 252-259, 2013.

10 Li-Pook-Than J and Snyder M: iPOP goes the world: integrated personalized Omics profiling and the road toward improved health care. Chem Biol *20*: 660-666, 2013.

11 Stricker T, Catenacci DV and Seiwert TY: Molecular profiling of cancer – the future of personalized cancer medicine: a primer on cancer biology and the tools necessary to bring molecular testing to the clinic. Semin Oncol *38*: 173-185, 2011.

12 Pabinger S, Dander A, Fischer M, Snajder R, Sperk M, Efremova M, Krabichler B, Speicher MR, Zschocke J and Trajanoski Z: A survey of tools for variant analysis of next-generation genome sequencing data. Brief Bioinform *15*: 256-278, 2014.

13 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, Garimella K, Altshuler D, Gabriel S, Daly M and DePristo MA: The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res *20*: 1297-1303, 2010.

14 Li H and Durbin R: Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics *25*: 1754-1760, 2009.

15 Li H and Durbin R: Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics *26*: 589-595, 2010.

16 Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, Pimentel H, Salzberg SL, Rinn JL and Pachter L: Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. Nat Protoc *7*: 562-578, 2012.

17 Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK and Ding L: VarScan: variant detection in massively parallel sequencing of individual and pooled samples. Bioinformatics *25*: 2283-2285, 2009.
18 Koboldt DC, Zhang Q, Larson DE, Shen D, McLellan MD, Lin L, Miller CA, Mardis ER, Ding L and Wilson RK: VarScan 2: somatic mutation and copy number alteration discovery in cancer by exome sequencing. Genome Res *22*: 568-576, 2012.
19 Larson DE, Harris CC, Chen K, Koboldt DC, Abbott TE, Dooling DJ, Ley TJ, Mardis ER, Wilson RK and Ding L: SomaticSniper: identification of somatic point mutations in whole genome sequencing data. Bioinformatics *28*: 311-317, 2012.
20 Cancer Genome Atlas Research N: Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature *455*: 1061-1068, 2008.
21 Cancer Genome Atlas Research N: Integrated genomic analyses of ovarian carcinoma. Nature *474*: 609-615, 2011.
22 Cancer Genome Atlas N: Comprehensive molecular characterization of human colon and rectal cancer. Nature *487*: 330-337, 2012.
23 Cancer Genome Atlas Research N: Comprehensive genomic characterization of squamous cell lung cancers. Nature *489*: 519-525, 2012.
24 Vaske CJ, Benz SC, Sanborn JZ, Earl D, Szeto C, Zhu J, Haussler D and Stuart JM: Inference of patient-specific pathway activities from multi-dimensional cancer genomics data using PARADIGM. Bioinformatics *26*: i237-245, 2010.
25 Dees ND, Zhang Q, Kandoth C, Wendl MC, Schierding W, Koboldt DC, Mooney TB, Callaway MB, Dooling D, Mardis ER, Wilson RK and Ding L: MuSiC: identifying mutational significance in cancer genomes. Genome Res *22*: 1589-1598, 2012.
26 Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G and Durbin R: The Sequence Alignment/Map format and SAMtools. Bioinformatics *25*: 2078-2079, 2009.
27 Wang K, Li M and Hakonarson H: ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res *38*: e164, 2010.
28 Yang H and Wang K: Genomic variant annotation and prioritization with ANNOVAR and wANNOVAR. Nat Protoc *10*: 1556-1566, 2015.
29 Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, Paulovich A, Pomeroy SL, Golub TR, Lander ES and Mesirov JP: Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc Natl Acad Sci USA *102*: 15545-15550, 2005.
30 Cheng YC, Hsiao FC, Yeh EC, Lin WJ, Tang CY, Tseng HC, Wu HT, Liu CK, Chen CC, Chen YT and Yao A: VarioWatch: providing large-scale and comprehensive annotations on human genomic variants in the next generation sequencing era. Nucleic Acids Res *40*: W76-81, 2012.
31 Dennis G Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC and Lempicki RA: DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol *4*: P3, 2003.
32 Huang da W, Sherman BT and Lempicki RA: Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*: 44-57, 2009.
33 Bardelli A and Siena S: Molecular mechanisms of resistance to cetuximab and panitumumab in colorectal cancer. J Clin Oncol *28*: 1254-1261, 2010.
34 Lau T, Chan E, Callow M, Waaler J, Boggs J, Blake RA, Magnuson S, Sambrone A, Schutten M, Firestein R, Machon O, Korinek V, Choo E, Diaz D, Merchant M, Polakis P, Holsworth DD, Krauss S and Costa M: A novel tankyrase small-molecule inhibitor suppresses APC mutation-driven colorectal tumor growth. Cancer Res *73*: 3132-3144, 2013.
35 Misale S, Arena S, Lamba S, Siravegna G, Lallo A, Hobor S, Russo M, Buscarino M, Lazzari L, Sartore-Bianchi A, Bencardino K, Amatu A, Lauricella C, Valtorta E, Siena S, Di Nicolantonio F and Bardelli A: Blockade of EGFR and MEK intercepts heterogeneous mechanisms of acquired resistance to anti-EGFR therapies in colorectal cancer. Sci Transl Med *6*: 224ra226, 2014.
36 Law V, Knox C, Djoumbou Y, Jewison T, Guo AC, Liu Y, Maciejewski A, Arndt D, Wilson M, Neveu V, Tang A, Gabriel G, Ly C, Adamjee S, Dame ZT, Han B, Zhou Y and Wishart DS: DrugBank 4.0: shedding new light on drug metabolism. Nucleic Acids Res *42*: D1091-1097, 2014.
37 Tang J and Aittokallio T: Network pharmacology strategies toward multi-target anticancer therapies: from computational models to experimental design principles. Curr Pharm Des *20*: 23-36, 2014.
38 Cho-Chung YS: cAMP signaling in cancer genesis and treatment. Cancer Treat Res *115*: 123-143, 2003.
39 Cho-Chung YS, Nesterova M, Becker KG, Srivastava R, Park YG, Lee YN, Cho YS, Kim MK, Neary C and Cheadle C: Dissecting the circuitry of protein kinase A and cAMP signaling in cancer genesis: antisense, microarray, gene overexpression, and transcription factor decoy. Ann NY Acad Sci *968*: 22-36, 2002.
40 McCarty MF: A role for cAMP-driven transactivation of EGFR in cancer aggressiveness - therapeutic implications. Med Hypotheses *83*: 142-147, 2014.
41 Wu D, Zhau HE, Huang WC, Iqbal S, Habib FK, Sartor O, Cvitanovic L, Marshall FF, Xu Z and Chung LW: cAMP-responsive element-binding protein regulates vascular endothelial growth factor expression: implication in human prostate cancer bone metastasis. Oncogene *26*: 5070-5077, 2007.
42 Comuzzi B, Lambrinidis L, Rogatsch H, Godoy-Tundidor S, Knezevic N, Krhen I, Marekovic Z, Bartsch G, Klocker H, Hobisch A and Culig Z: The transcriptional co-activator cAMP response element-binding protein-binding protein is expressed in prostate cancer and enhances androgen- and anti-androgen-induced androgen receptor function. Am J Pathol *162*: 233-241, 2003.
43 Son J, Lee JH, Kim HN, Ha H and Lee ZH: cAMP-response-element-binding protein positively regulates breast cancer metastasis and subsequent bone destruction. Biochem Biophys Res Commun *398*: 309-314, 2010.
44 Spina A, Di Maiolo F, Esposito A, D'Auria R, Di Gesto D, Chiosi E, Sorvillo L and Naviglio S: Integrating leptin and cAMP signalling pathways in triple-negative breast cancer cells. Front Biosci (Landmark Ed) *18*: 133-144, 2013.
45 Zhang M, Xu JJ, Zhou RL and Zhang QY: cAMP responsive element binding protein-1 is a transcription factor of lysosomal-associated protein transmembrane-4 Beta in human breast cancer cells. PLoS One *8*: e57520, 2013.

46 O'Hurley G, Prencipe M, Lundon D, O'Neill A, Boyce S, O'Grady A, Gallagher WM, Morrissey C, Kay EW and Watson RW: The analysis of serum response factor expression in bone and soft tissue prostate cancer metastases. Prostate *74*: 306-313, 2014.

47 Verone AR, Duncan K, Godoy A, Yadav N, Bakin A, Koochekpour S, Jin JP and Heemers HV: Androgen-responsive serum response factor target genes regulate prostate cancer cell migration. Carcinogenesis *34*: 1737-1746, 2013.

48 Heemers HV: Identification of a RhoA- and SRF-dependent mechanism of androgen action that is associated with prostate cancer progression. Curr Drug Targets *14*: 481-489, 2013.

49 Walker T, Nolte A, Steger V, Makowiecki C, Mustafi M, Friedel G, Schlensak C and Wendel HP: Small interfering RNA-mediated suppression of serum response factor, E2-promotor binding factor and survivin in non-small cell lung cancer cell lines by non-viral transfection. Eur J Cardiothorac Surg *43*: 628-633; discussion 633-624, 2013.

50 Zhao M, Xu H, He X, Hua H, Luo Y and Zuo L: Expression of serum response factor in gastric carcinoma and its molecular mechanisms involved in the regulation of the invasion and migration of SGC-7901 cells. Cancer Biother Radiopharm *28*: 146-152, 2013.

51 Kim KR, Bae JS, Choi HN, Park HS, Jang KY, Chung MJ and Moon WS: The role of serum response factor in hepatocellular carcinoma: an association with matrix metalloproteinase. Oncol Rep *26*: 1567-1572, 2011.

52 Kwon CY, Kim KR, Choi HN, Chung MJ, Noh SJ, Kim DG, Kang MJ, Lee DG and Moon WS: The role of serum response factor in hepatocellular carcinoma: implications for disease progression. Int J Oncol *37*: 837-844, 2010.

53 Kim HJ, Kim KR, Park HS, Jang KY, Chung MJ, Shong M and Moon WS: The expression and role of serum response factor in papillary carcinoma of the thyroid. Int J Oncol *35*: 49-55, 2009.

54 Chang CC, Shieh GS, Wu P, Lin CC, Shiau AL and Wu CL: Oct-3/4 expression reflects tumor progression and regulates motility of bladder cancer cells. Cancer Res *68*: 6281-6291, 2008.

55 Yamanaka K, Mizuarai S, Eguchi T, Itadani H, Hirai H and Kotani H: Expression levels of NF-Y target genes changed by CDKN1B correlate with clinical prognosis in multiple cancers. Genomics *94*: 219-227, 2009.

56 Rufini A, Agostini M, Grespi F, Tomasini R, Sayan BS, Niklison-Chirou MV, Conforti F, Velletri T, Mastino A, Mak TW, Melino G and Knight RA: p73 in Cancer. Genes Cancer *2*: 491-502, 2011.

57 Su XL, Ouyang XH, Yan MR and Liu GR: p73 expression and its clinical significance in colorectal cancer. Colorectal Dis *11*: 960-963, 2009.

58 Ribeiro JR, Lovasco LA, Vanderhyden BC and Freiman RN: Targeting TBP-Associated Factors in Ovarian Cancer. Front Oncol *4*: 45, 2014.

Received October 13, 2015
Revised December 15, 2015
Accepted December 16, 2015