Published online 3 December 2015 *Nucleic Acids Research, 2016, Vol. 44, Database issue* ***D975–D979***
doi: 10.1093/nar/gkv1314

DriverDBv2: a database for human cancer driver gene research

I-Fang Chung[1,2,†], Chen-Yang Chen[1,†], Shih-Chieh Su[3], Chia-Yang Li[4,5], Kou-Juey Wu[3,6], Hsei-Wei Wang[7,8,9,10,*] and Wei-Chung Cheng[3,6,*]

[1]Institute of Biomedical Informatics, National Yang-Ming University, Taipei 11221, Taiwan, [2]Center for Systems and Synthetic Biology, National Yang-Ming University, Taipei, 11221, Taiwan, [3]Research Center for Tumour Medical Science, China Medical University, Taichung, 40402, Taiwan, [4]Department of Genome Medicine, College of Medicine, Kaohsiung Medical University, Kaohsiung 80708, Taiwan, [5]Center for Infectious Disease and Cancer Research, Kaohsiung Medical University, Kaohsiung 80708, Taiwan, [6]Graduate Institute of Cancer Biology, China Medical University, Taichung, 40402, Taiwan, [7]VGH-YM Genomic Research Center, National Yang-Ming University, Taipei 11221, Taiwan, [8]Institute of Clinical Medicine, Medical College, National Yang-Ming University, Taipei 11221, Taiwan, [9]Institute of Microbiology and Immunology, National Yang-Ming University, Taipei 11221, Taiwan and [10]Department of Education and Research, Taipei City Hospital, Taipei 10341, Taiwan

Received September 15, 2015; Revised October 31, 2015; Accepted November 10, 2015

ABSTRACT

We previously presented DriverDB, a database that incorporates ∼6000 cases of exome-seq data, in addition to annotation databases and published bioinformatics algorithms dedicated to driver gene/mutation identification. The database provides two points of view, 'Cancer' and 'Gene', to help researchers visualize the relationships between cancers and driver genes/mutations. In the updated DriverDBv2 database (http://ngs.ym.edu.tw/driverdb) presented herein, we incorporated >9500 cancer-related RNA-seq datasets and >7000 more exome-seq datasets from The Cancer Genome Atlas (TCGA), International Cancer Genome Consortium (ICGC), and published papers. Seven additional computational algorithms (meaning that the updated database contains 15 in total), which were developed for driver gene identification, are incorporated into our analysis pipeline, and the results are provided in the 'Cancer' section. Furthermore, there are two main new features, 'Expression' and 'Hotspot', in the 'Gene' section. 'Expression' displays two expression profiles of a gene in terms of sample types and mutation types, respectively. 'Hotspot' indicates the hotspot mutation regions of a gene according to the results provided by four bioinformatics tools. A new function, 'Gene Set', allows users to investigate the relationships among mutations, expression levels and clinical data for a set of genes, a specific dataset and clinical features.

INTRODUCTION

In the past few years, next generation sequencing (NGS) has revolutionized cancer genomic studies. Large-scale cancer genomic projects, such as The Cancer Genome Atlas (TCGA), have utilized different types of sequencing technology (such as RNA-seq and Exome-seq) in analysing cancer samples in order to provide distinct profiles of cancer biology. However, translating the different types of cancer genomic data into information that can be easily interpreted and accessed remains a challenge.

The integration of multi-dimensional genomic data has been crucial to our understanding of biologically and clinically relevant subtypes of cancer. One example of integrative analysis was the breast cancer study of TCGA, to show expression subtype-associated enrichment for cancer driver genes. For instance, the ERBB2-expression subtype is associated with the enrichment of TP53 and PIK3CA mutations (1). The recent unbiased genomic characterization of distinct cancers has also provided insights into the driving events in genetic subtypes of cancers that are not well understood. The integrative analysis of cancer genomics data can provide both mechanistic and biological insights into the role of driver genes in a specific cancer type (2). There are several tools, such as MAGI (3) and cBioportal (4,5), that allow for the exploration, annotation and integration of different kinds of cancer genomic data.

*To whom correspondence should be addressed. Tel: +886 4 22052121 (Ext. 7820); Fax: +886 4 22337425; Email: cwc0702@gmail.com
Correspondence may also be addressed to Hsei-Wei Wang. Tel: +886 2 28267109; Fax: +886 2 2821 2880; Email: hwwang@ym.edu.tw
†These authors contributed equally to the paper as first authors.

Mutations are random, but the occurrence of hotspot/clustered mutations is driven by positive selection, especially when the mutations are located in functional domains or in the residues that are important for 3D protein structures (2). The same mutations in hotspot mutation regions (HMRs) may be found as drivers in other cancers. Many driver mutations recurrently occur in the functional regions of proteins (for example, kinase domains or binding domains) (6) or interrupt active sites (for example, phosphorylation sites) (7). Hotspot regions can be grouped into two types (8), mutation clusters and hotspot domains. Hotspot domains are well-annotated domains with higher mutation rates than are found in the remaining regions of the protein. The documentation of a hotspot domain requires a prior annotation of previously known protein domain information for every transcript. Mutation clusters are small fractions of proteins that have accumulated a high number of mutations regardless of whether or not the clusters are located in functional domain regions of the protein. A mutation cluster may even have an extremely high mutation rate; for example, the V600E cluster of the BRAF gene has a very high mutation rate and is located in a tyrosine kinase domain. Some cancer driver genes (such as KRAS and BRAF) have only one HMR, but some (such as PIK3CA) may have two or more HMRs in distinct cancer types. HMRs are strong indicators for cancer in that mutations in these HMRs may promote cancer progression. Hence, it is important to identify HMRs in cancer biology. Several computational methods have been developed for identifying driver genes by defining HMRs (8–13).

Previously, we developed DriverDB (14), a database that incorporates ∼6000 cases of exome-seq data, in addition to annotation databases and published bioinformatics algorithms dedicated to driver gene/mutation identification. Here, we present DriverDBv2, an updated version of the database. In addition to including more exome-seq results (>7000 more datasets of exome-seq from TCGA, ICGC and published papers), we have incorporated seven more algorithms developed for driver gene identification in this updated version. Four of those seven methods identify driver genes according to the identification of HMRs. We also provide information on these HMRs in this updated database. Specifically, we have integrated >9500 RNA-seq into DriverDBv2 to provide expression profiles across cancer types. DriverDBv2 also contains a new function called 'GeneSet', which allows researchers to visualize the mutations, expression levels and clinical profiles of customer-defined genes, datasets and clinical data.

DATA COLLECTION AND PREPROCESSING

DriverDBv2 incorporates >7000 additional exome-seq datasets from TCGA, ICGC and published papers, as well as RNA-seq data from >9500 cancer-related samples (such as primary tumor, normal tissue and metastatic tissue) in TCGA. Detailed information on these datasets is described in Supplementary Table S1. All sequencing results, such as mutation and expression data, have been curated in uniform formats by an in-house script and then stored in our local MySQL server. All mutations are also functionally annotated as described in our previous study (14). For all clinical data downloaded from distinct studies using varied terminologies, we have standardized them using the Common Data Element (CDE) format, the standard elements of which are used in the validation of clinical data in TCGA, through manual curation according to the definition of terms (https://tcga-data.nci.nih.gov/docs/dictionary/).

DRIVER GENE AND HMR IDENTIFICATION

DriverDBv2 contains seven additional algorithms for driver gene identification. DriverNet (15) and DawnRank (16) utilize transcriptional networks to identify driver genes. The rationale of the two algorithms is that the impact of a potential driver gene can be determined by its effect on the genes that are regulated by it. COMDP (17) is based on mutual exclusivity to identify sets of driver genes mutated in known pathways. The other four algorithms, MSEA (8), e-Drivers (9), oncodriveCLUST (12) and iPAC (11), identify cancer driver genes by defining the HMRs. OncodriveCLUST and iPAC only identify mutation clusters and e-Driver only identifies hotspot domains, but MSEA can identify both types of HMRs. All HMRs identified by the four algorithms are integrated and illustrated in the 'Hotspot' panel of the 'Gene' section. The detailed criteria of the seven new algorithms are described in the Supplementary Methods.

WEB INTERFACE

Gene

As shown in Figure 1, we provide three new panels, 'Summary', 'Expression' and 'Hotspot', in the 'Gene' section of the updated database. In Figure 1, we used the gene TP53 as an example. For 'Summary', a heat map shows which bioinformatics tool identifies the gene as a driver gene in which cancer type (Figure 1A). The bar chart at the top of the heat map indicates the cumulative counts of tools. In the 'Hotspot' panel, a heat map shows the regions of the protein that are identified as HMRs across different cancer types (Figure 1B). The color used for a given region indicates the number of tools that identify that region as an HMR. The cumulative counts for the regions identified as HRMs are shown at the top of the heat map. Exon and domain information with protein coordinates are provided at the bottom of the heat map. For the 'Expression' panel, the expression profiles of the gene across cancer types by sample type and by mutation class are illustrated by boxplot in Figure 1C and D, respectively. The colors used in Figure 1C and D indicate the sample types (such as normal tissue and primary tumor) and mutation classes (such as truncating and in-frame mutations), respectively.

GeneSet

The new function, 'GeneSet', was designed to help researchers visualize the relationship among mutation, expression, and clinical information. Figure 2 is an example of KRAS, NRAS and RAF in colon adenocarcinoma samples from TCGA. As shown in Supplementary Figure S1, researchers could upload a set of genes, select a specific dataset and choose up to three clinical characteristics of the

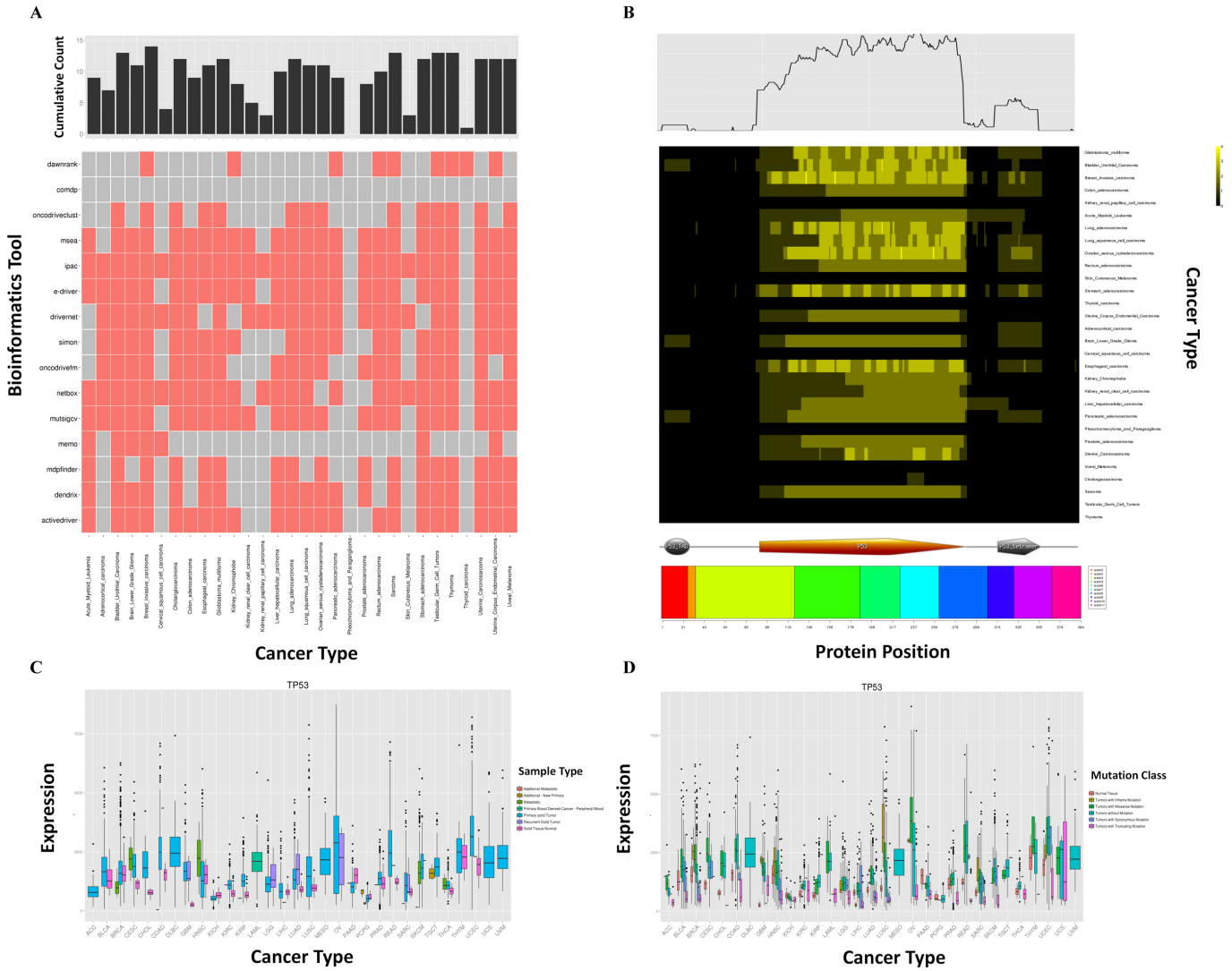

Figure 1. The three new features in the 'Gene' section. (**A**) The 'Summary' panel. A heat map shows which bioinformatics tool identifies the gene as a driver gene in which cancer type. The upper panel shows the cumulative counts of bioinformatics tools. (**B**) The 'Hotspot' panel. A heat map shows the regions of the protein identified as HMRs across different cancer types. The color used for a given region indicates the number of tools that identify that region as an HMR. The upper panel shows the cumulative counts of the regions identified as HRMs. Exon and domain information with protein coordinates are provided at the bottom of the heat map. (C and D) The 'Expression' panel. The expression boxplots of the gene across cancer types by sample type (**C**) and by mutation class (**D**). The colors in (C) and (D) indicate the sample types and mutation classes, respectively.

selected dataset. After the query is submitted, an integrative figure (Figure 2A) displays the relationship among the three kinds of information. For clinical plot, clinical data may be various and complex. To simplify this issue, we used the grayscale to indicate the level of data for each clinical characteristic and remove the figure legend. The red color indicates the value is not available. In addition, two expression boxplots show the expression of uploaded genes by sample type (Figure 2B) and by mutation class (Figure 2C). The raw data are available for download via a download link.

DISCUSSION

The integrated analysis of multi-dimensional genomic data is crucial to our understanding of cancer biology. DriverDBv2 seeks to integrate mutation and expression data to address several issues. For driver gene identification, Drivernet, MeMO and DawnRank, the tools used for identifying driver genes in DriverDBv2, utilize two types of data to predict cancer driver genes and may provide additional insights regarding those cancer driver genes. For a specific gene, the expression of the gene may differ in mutated cases as compared to normal cases. For example, a reduced expression of STAG2 in mutant cases has previously been reported (18,19). The 'Expression' panel in the 'Gene' section of DriverDBv2 shows the expression boxplots for a given gene in different cancer types by mutation class and by sample type. This function will be helpful when researchers would like to quickly evaluate an interesting gene in distinct cancer types or validate their wet lab results in silicon. Moreover, the new function 'GeneSet' further integrates mutations, expression levels and clinical

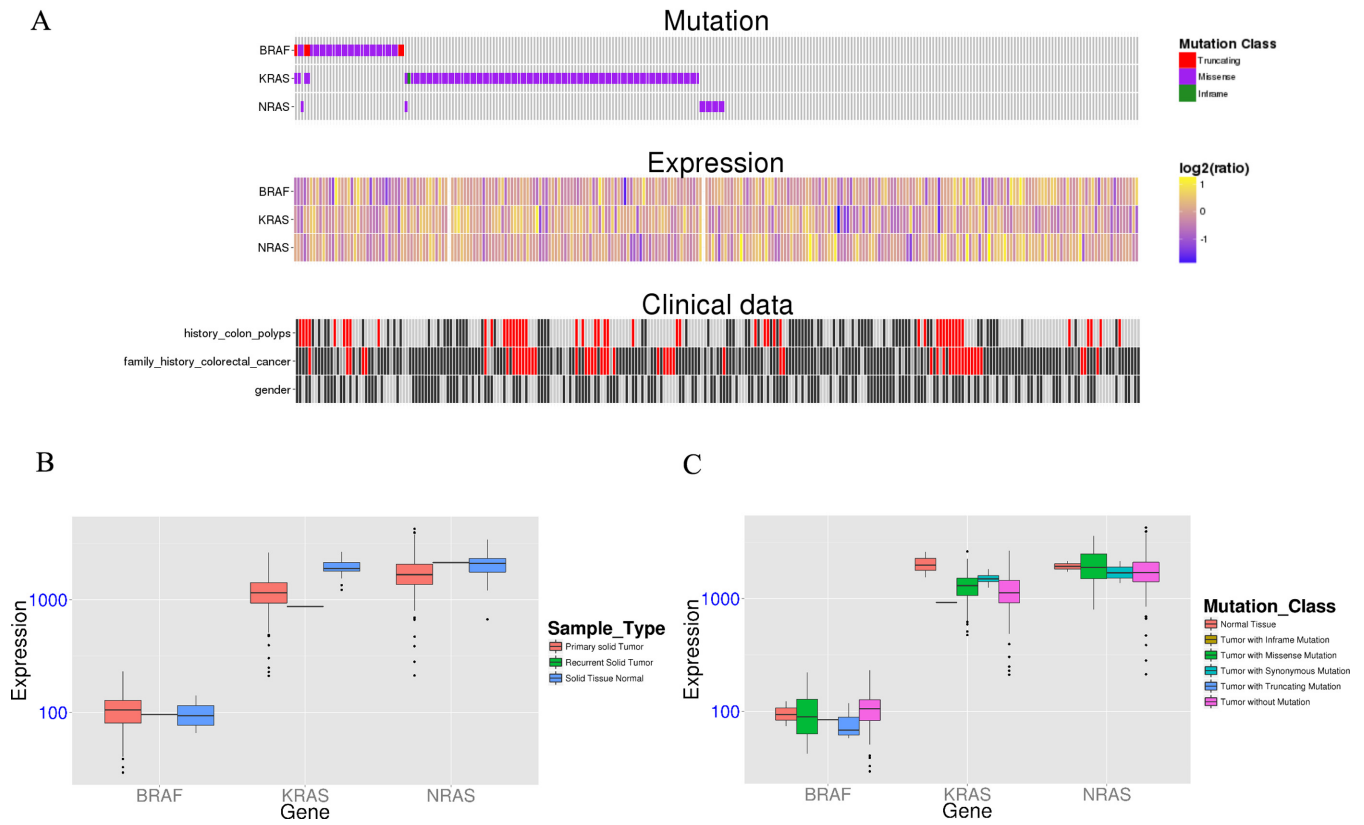

Figure 2. The new function, 'GeneSet'. (**A**) An integrative figure displays the relationship between mutation, expression levels, and clinical information. For the clinical plot, the grayscale indicates the level of data for each clinical characteristic and the red color indicates the value is not available. (B and C) Two expression boxplots show the expression of uploaded genes in terms of sample type (**B**) and mutation class (**C**).

information for visualization. It has previously been noted that the co-occurrence of a mutated gene with the abnormal expression of another gene may be related to a specific phenotype. The example of abnormal MITF expression with mutated BRAF has been used to illustrate this concept (20). When MITF overexpression occurs in isolation, it does not affect the proliferation of immortalized melanocytes; however, it does affect their proliferation when it also occurs with the expression of the BRAF V600E mutant, which co-occurs with abnormal MITF expression. The 'GeneSet' panel could help explore this relationship. Furthermore, we have also provided the raw data for the integrative figure in 'GeneSet' for further analysis. To answer whether a gene is a driver in cancer, DriverDBv2 provides the new panel, 'Summary', in 'Gene' section. This panel shows which bioinformatics tool identifies the gene as a driver gene in which type of cancer (Figure 1A)

The occurrence of hotspot mutations is driven by positive selection and is a strong indicator for cancer in that mutations in hotspot regions may promote cancer progression. Hence, it is important to identify HMRs in cancer biology. It has been noted that some driver genes have one or more HMRs. For example, mutations in PIK3CA form two clusters in the helical and catalytic domains (2,21). In extreme cases, driver genes have highly recurrent substitutions that change the same amino acid, such as in the case of the arginine at codon 132 in IDH1 (22) and the V600 mutation in BRAF (23). Jia *et al.* investigated known cancer genes from the Cancer Gene Census (CGC) (24) collection and investigated mutations from COSMIC database (25). They found that the known driver genes from CGC genes were detected through mutation analysis in previous studies; approximately 51% of the CGC genes can be detected through mutation hotspot analysis (8). This high proportion of genes with HMRs supports the feasibility of predicting additional cancer genes based on mutation clustering patterns. DriverDBv2 integrates the information of HMRs in distinct cancer types through the utilization of four bioinformatics tools and illustrates the results in the 'Hotspot' panel of the 'Gene' section. The information thus provided tells researchers whether the driver gene that they are interested in has the same or distinct HMRs in different cancer types.

In this updated version, we have integrated exome-seq and RNA-seq data to identify cancer driver genes and HMRs from larger-scale cancer sequencing data. DriverDBv2 provides researchers with easy access to different aspects of information regarding cancer driver genes. In the future, we will incorporate more different kinds of genomics data in further updates to DriverDB, so that the database will continue to be an informative and valuable source of data on cancer driver genes.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENT

We are grateful to the National Center for High performance Computing for computer time and facilities and thank the TCGA research network for the availability of data.

FUNDING

Ministry of Science and Technology (MOST) [104-2221-E-010-012, 104-2320-B-039-053]; China Medical University, Taiwan [CMU104-N-14]; National Yang-Ming University, Taiwan (a grant from the Ministry of Education, Aim for the Top University Plan). Funding for open access charge: Ministry of Science and Technology in Taiwan.
Conflict of interest statement. None declared.

REFERENCES

1. Cancer Genome Atlas Network. (2012) Comprehensive molecular portraits of human breast tumours. *Nature*, **490**, 61–70.
2. Watson,I.R., Takahashi,K., Futreal,P.A. and Chin,L. (2013) Emerging patterns of somatic mutations in cancer. *Nat. Rev. Genet.*, **14**, 703–718.
3. Leiserson,M.D., Gramazio,C.C., Hu,J., Wu,H.T., Laidlaw,D.H. and Raphael,B.J. (2015) MAGI: visualization and collaborative annotation of genomic aberrations. *Nat. Methods*, **12**, 483–484.
4. Cerami,E., Gao,J., Dogrusoz,U., Gross,B.E., Sumer,S.O., Aksoy,B.A., Jacobsen,A., Byrne,C.J., Heuer,M.L., Larsson,E. *et al.* (2012) The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov.*, **2**, 401–404.
5. Gao,J., Aksoy,B.A., Dogrusoz,U., Dresdner,G., Gross,B., Sumer,S.O., Sun,Y., Jacobsen,A., Sinha,R., Larsson,E. *et al.* (2013) Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Sci. Signal.*, **6**, pl1.
6. Prior,I.A., Lewis,P.D. and Mattos,C. (2012) A comprehensive survey of Ras mutations in cancer. *Cancer Res.*, **72**, 2457–2467.
7. Reimand,J., Wagih,O. and Bader,G.D. (2013) The mutational landscape of phosphorylation signaling in cancer. *Scientific Rep.*, **3**, 2651.
8. Jia,P., Wang,Q., Chen,Q., Hutchinson,K.E., Pao,W. and Zhao,Z. (2014) MSEA: detection and quantification of mutation hotspots through mutation set enrichment analysis. *Genome Biol.*, **15**, 489.
9. Porta-Pardo,E. and Godzik,A. (2014) e-Driver: a novel method to identify protein regions driving cancer. *Bioinformatics*, **30**, 3109–3114.
10. Ryslik,G.A., Cheng,Y., Cheung,K.H., Bjornson,R.D., Zelterman,D., Modis,Y. and Zhao,H. (2014) A spatial simulation approach to account for protein structure when identifying non-random somatic mutations. *BMC Bioinformatics*, **15**, 231.
11. Ryslik,G.A., Cheng,Y., Cheung,K.H., Modis,Y. and Zhao,H. (2013) Utilizing protein structure to identify non-random somatic mutations. *BMC Bioinformatics*, **14**, 190.
12. Tamborero,D., Gonzalez-Perez,A. and Lopez-Bigas,N. (2013) OncodriveCLUST: exploiting the positional clustering of somatic mutations to identify cancer genes. *Bioinformatics*, **29**, 2238–2244.
13. Van den Eynden,J., Fierro,A.C., Verbeke,L.P. and Marchal,K. (2015) SomInaClust: detection of cancer genes based on somatic mutation patterns of inactivation and clustering. *BMC Bioinformatics*, **16**, 125.
14. Cheng,W.C., Chung,I.F., Chen,C.Y., Sun,H.J., Fen,J.J., Tang,W.C., Chang,T.Y., Wong,T.T. and Wang,H.W. (2014) DriverDB: an exome sequencing database for cancer driver gene identification. *Nucleic Acids Res.*, **42**, D1048–D1054.
15. Bashashati,A., Haffari,G., Ding,J., Ha,G., Lui,K., Rosner,J., Huntsman,D.G., Caldas,C., Aparicio,S.A. and Shah,S.P. (2012) DriverNet: uncovering the impact of somatic driver mutations on transcriptional networks in cancer. *Genome Biol.*, **13**, R124.
16. Hou,J.P. and Ma,J. (2014) DawnRank: discovering personalized driver genes in cancer. *Genome Med.*, **6**, 56.
17. Zhang,J., Wu,L.Y., Zhang,X.S. and Zhang,S. (2014) Discovery of co-occurring driver pathways in cancer. *BMC Bioinformatics*, **15**, 271.
18. Gerstung,M., Pellagatti,A., Malcovati,L., Giagounidis,A., Porta,M.G., Jadersten,M., Dolatshad,H., Verma,A., Cross,N.C., Vyas,P. *et al.* (2015) Combining gene mutation with gene expression data improves outcome prediction in myelodysplastic syndromes. *Nat. Commun.*, **6**, 5901.
19. Kon,A., Shih,L.Y., Minamino,M., Sanada,M., Shiraishi,Y., Nagata,Y., Yoshida,K., Okuno,Y., Bando,M., Nakato,R. *et al.* (2013) Recurrent mutations in multiple components of the cohesin complex in myeloid neoplasms. *Nat. Genet.*, **45**, 1232–1237.
20. Garraway,L.A., Widlund,H.R., Rubin,M.A., Getz,G., Berger,A.J., Ramaswamy,S., Beroukhim,R., Milner,D.A., Granter,S.R., Du,J. *et al.* (2005) Integrative genomic analyses identify MITF as a lineage survival oncogene amplified in malignant melanoma. *Nature*, **436**, 117–122.
21. Vogelstein,B., Papadopoulos,N., Velculescu,V.E., Zhou,S., Diaz,L.A. Jr and Kinzler,K.W. (2013) Cancer genome landscapes. *Science*, **339**, 1546–1558.
22. Parsons,D.W., Jones,S., Zhang,X., Lin,J.C., Leary,R.J., Angenendt,P., Mankoo,P., Carter,H., Siu,I.M., Gallia,G.L. *et al.* (2008) An integrated genomic analysis of human glioblastoma multiforme. *Science*, **321**, 1807–1812.
23. Davies,H., Bignell,G.R., Cox,C., Stephens,P., Edkins,S., Clegg,S., Teague,J., Woffendin,H., Garnett,M.J., Bottomley,W. *et al.* (2002) Mutations of the BRAF gene in human cancer. *Nature*, **417**, 949–954.
24. Futreal,P.A., Coin,L., Marshall,M., Down,T., Hubbard,T., Wooster,R., Rahman,N. and Stratton,M.R. (2004) A census of human cancer genes. *Nat. Rev. Cancer*, **4**, 177–183.
25. Forbes,S.A., Beare,D., Gunasekaran,P., Leung,K., Bindal,N., Boutselakis,H., Ding,M., Bamford,S., Cole,C., Ward,S. *et al.* (2015) COSMIC: exploring the world's knowledge of somatic mutations in human cancer. *Nucleic Acids Res.*, **43**, D805–D811.