D1048–D1054 *Nucleic Acids Research, 2014, Vol. 42, Database issue* *Published online 7 November 2013*
doi:10.1093/nar/gkt1025

DriverDB: an exome sequencing database for cancer driver gene identification

Wei-Chung Cheng[1,2], I-Fang Chung[3], Chen-Yang Chen[3], Hsing-Jen Sun[3], Jun-Jeng Fen[3,4], Wei-Chun Tang[3], Ting-Yu Chang[5], Tai-Tong Wong[1,2,*] and Hsei-Wei Wang[2,3,5,6,*]

[1]Pediatric Neurosurgery, Department of Surgery, Cheng Hsin General Hospital, Taipei 11220, Taiwan, [2]VGH-YM Genomic Research Center, National Yang-Ming University, Taipei 11221, Taiwan, [3]Institute of Biomedical Informatics, National Yang-Ming University, Taipei 11221, Taiwan, [4]Information Technology Office, Taipei Veterans General Hospital, Taipei 11217, Taiwan, [5]Institute of Microbiology and Immunology, National Yang-Ming University, Taipei 11221, Taiwan and [6]Department of Education and Research, Taipei City Hospital, Taipei 10341, Taiwan

Received August 15, 2013; Revised and Accepted October 7, 2013

ABSTRACT

Exome sequencing (exome-seq) has aided in the discovery of a huge amount of mutations in cancers, yet challenges remain in converting oncogenomics data into information that is interpretable and accessible for clinical care. We constructed DriverDB (http://ngs.ym.edu.tw/driverdb/), a database which incorporates 6079 cases of exome-seq data, annotation databases (such as dbSNP, 1000 Genome and Cosmic) and published bioinformatics algorithms dedicated to driver gene/mutation identification. We provide two points of view, 'Cancer' and 'Gene', to help researchers to visualize the relationships between cancers and driver genes/mutations. The 'Cancer' section summarizes the calculated results of driver genes by eight computational methods for a specific cancer type/dataset and provides three levels of biological interpretation for realization of the relationships between driver genes. The 'Gene' section is designed to visualize the mutation information of a driver gene in five different aspects. Moreover, a 'Meta-Analysis' function is provided so researchers may identify driver genes in customer-defined samples. The novel driver genes/mutations identified hold potential for both basic research and biotech applications.

INTRODUCTION

Next-generation sequencing (NGS) has greatly increased the identification of mutations in cancer genomes and allows researchers to profile the molecular characteristics of various cancer types. In the past few years, applying exome sequencing (exome-seq) in oncogenomics studies has become the norm (1). Also, enormous amounts of cancer genomics data have been generated from large-scale cancer projects (2) such as The Cancer Genome Atlas (TCGA), the International Cancer Genome Consortium (ICGC), the Therapeutically Applicable Research to Generate Effective Treatments (TARGET) and the Pediatric Cancer Genome Project (PCGP). Although NGS has already helped researchers discover huge amounts of aberrant events in cancer genomics, translating these data into information that can be easily interpreted and accessed is still challenging.

Cancers are primarily caused by the accumulation of genetic alterations and could be characterized by numerous somatic mutations. However, not all of these mutations are involved in tumorigenesis. Only a subset of mutations contributes to cancer development, whereas others make no or little important contribution. To crystallize this concept, the terms 'driver and 'passenger' mutation have been coined (3). The mutations that confer a selective growth advantage to the tumor cell are called 'driver' mutations (1). 'Passenger' mutations are defined as those which do not confer growth advantage but that do occur in a cell that coincidentally or subse-

*To whom correspondence should be addressed. Tel: +886 2 2826 7109; Fax: +886 2 28212880; Email: hwwang@ym.edu.tw
Correspondence may also be addressed to Tai-Tong Wong. Tel: +886 2 28264545; Fax: +886 2 28264533; Email: ch9321@chgh.org.tw

The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

quently acquires a driver mutation (4). In most solid tumors, an average of 33–66 genes with somatic mutations were found to alter their protein products, but the count of non-synonymous mutations varies across cancer types (1). More than 80% of mutations are missense (1), and these mutations vary highly in their functional impact depending on their position and function in the protein and the nature of the replacement amino acid. It remains a significant challenge to identify cancer driver mutations because many observed missense changes are neutral passenger mutations (5). Several computational algorithms have been developed to predict the functional impact of missense mutations based on concepts including evolutionary conservation, structural constraints and the physicochemical attributes of amino acids. In the last few years, machine learning methods have been developed to specifically predict cancer-driving deleterious mutations (6–8).

A driver gene is defined as a gene whose dysfunction will cause tumorigenesis. Vogelstein *et al.* have demonstrated the fundamental difference between a driver gene and a driver mutation (1). Numerous computational methods to identify driver genes have been published; algorithms such as MutsigCV (9), MuSiC (10), Simon (11), OncodriverFM (12) and ActiveDriver (13) are based on the mutation frequency of an individual gene compared with the background mutation rate. However, background mutation rates among different genome regions and patients are highly variable (9). Recent studies have shown that the mutation rate varies in normal cells by more than 100-fold within the genome (14) and that such variation is higher in tumor cells (15). To correct for this bias, MutSigCV uses patient-specific mutation frequency and spectrum, as well as gene-specific background mutation rates. OncodriverFM incorporates the functional impacts of mutations as additional information. ActiveDriver identifies driver genes with statistically significant mutation rates in phosphorylation-specific regions. Other methods are based on the sub-network approach (16–24) that can identify groups of genes containing driver mutations directly from cancer mutation data either with or without prior knowledge of pathways or other information of protein/genetics interactions. This approach is successful particularly when the observed frequencies of passenger and driver mutations are indistinguishable, a situation wherein single gene tests fail. Moreover, sub-networks are believed to identify cancer driver genes with low recurrence (25). Most of sub-network based methods, such as MEMo (19), MDPFinder (16), Dendrix (17), Multi-Dendrix (18) and RME (24), identify driver genes with the characteristics of mutual exclusivity. Moreover, sub-network methods could additionally incorporate copy number variation (CNV) data for driver gene identification (16–19,22,24).

In this study, we present the DriverDB database, which incorporates a large amount (>6000 cases) of exome-seq data, annotation databases (such as dbSNP (26), 1000 Genome (27) and COSMIC (28)), and the various bioinformatics algorithms devoted to defining driver genes or mutations. DriverDB focuses on predicting driver genes by various algorithms and provides different aspects of the mutation profiles of an individual gene. We provide two view points, 'Cancer' and 'Gene', for benefiting researchers to visualize the relationships between cancers and driver genes/mutations. A 'Meta-Analysis' function is further included in the DriverDB for allowing researchers to identify driver genes of custom-defined samples according to clinical criteria.

MATERIALS AND METHODS

Dataset collection

As shown in Figure 1, DriverDB includes mutation profiles from 6079 tumor–normal pairs, including 4397 from TCGA, 861 from ICGC, 112 from PCGP, 238 from TARGET and 471 from published papers (denoted as 'others' in Figure 1). Detailed information for the datasets is provided in Supplementary Table S1. The mutation data and CNV data of these pairs were retrieved from the data portal of the projects or from the supplementary data of the published papers, and were then parsed using in-house Perl scripts. To ensure annotation consistency and to make the retrieval process more efficient, clinical information for each sample was manually curated, based on clinical data obtained as mentioned above. Each sample was re-annotated with 38 clinical characteristics. The summary of the clinical information is provided in Supplementary Table S2.

Mutation annotation

All mutations were mapped to known databases, and their functional impacts were predicted by numerous bioinformatics tools shown in the **Annotation** module in Figure 1. For annotating known variants, DriverDB incorporates the information collected from different databases including dbSNP, NHLBI GO ESP (29), 1000 genomes, COSMIC, ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/), NHGRI GWAS catalog (30), HGMD-PUBLIC (31) and OMIM (http://omim.org/). We used SnpEff (32) and VEP (33) to predict the effect of each mutation, such as non-synonymous coding, stop gained/lost and frame-shift. In addition, DriverDB scores the deleterious effects and functional impact by seven algorithms, including SIFT (34), PolyPhen2 (35), Condel (36), LRT (37), FATHMM (38), MutationAssessor (39) and MutationTaster (40). Furthermore, we scored each mutation by the number of algorithms that judge the mutation as deleterious (these numbers are denoted as 'Driver Score'). For example, the mutation g.178952085A>G of PIK3CA, which occurs in >100 patients from various cancer types, was identified as deleterious by seven algorithms; therefore, its Driver Score is 7.

Driver gene identification

DriverDB utilized eight computational methods to identify driver genes of cancer types (the **Cancer Driver Gene** module in Figure 1). Four methods, including MutsigCV, Simon, OncodriverFM and ActiveDriver, are

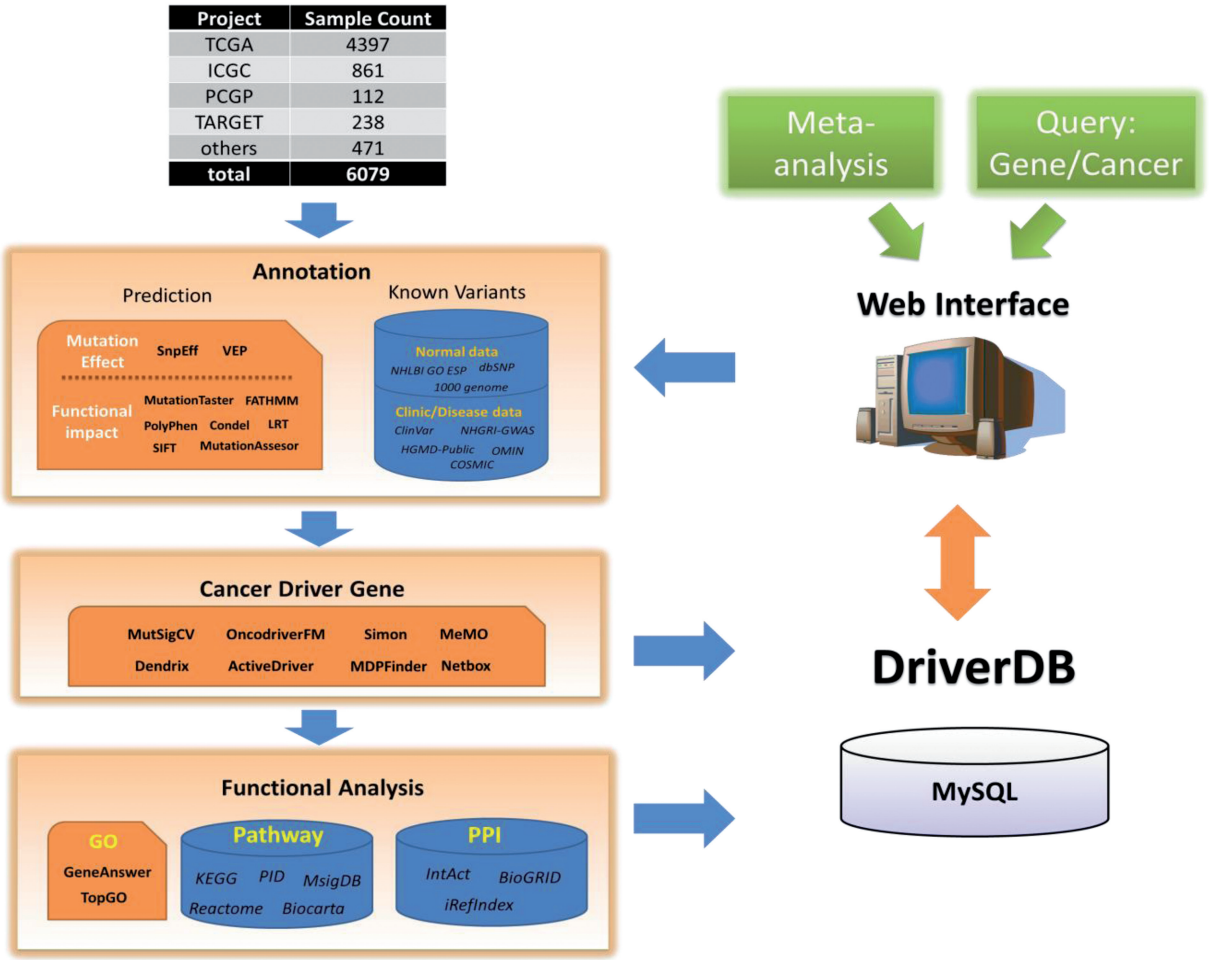

Figure 1. Schematic representation of data processing.

based on mutation frequencies and utilize all mutations to identify driver genes.

For the sub-network based methods, MEMo, Dendrix, MDPFinder and NetBox were used. We applied the following filters to remove mutations/genes from the analysis:

- Mutations whose effect impact was identified by SnpEff as 'Low' or 'Modifier.'
- Mutations denoted as common and not recorded in disease/clinical-related databases according to **Mutation annotation**.
- Potentially spurious genes reported by several studies (9,18).

Detailed criteria for each method are described in Supplementary Methods.

Functional analysis

For each set of driver genes identified by individual/multiple method(s) in a group of cancer samples, we provided three levels of biological interpretation (Gene Oncology, Pathways and Protein/Genetic Interaction) to help researchers to realize the relationships between driver genes. In the 'Gene Oncology' part, we used the topGO and GeneAnswers packages of Bioconductor to calculate the topology of the GO graph, as well as to visualize the many-to-many relationships between GO terms and genes. In the 'Pathway' analysis, we used collections from KEGG (41), PID (42), Biocarta (http://www.biocarta.com/), REACTOME (43) and MSigDB (44) to annotate driver genes. Detailed information for these eight collections is provided in Supplementary Table S3. The three databases, IntAct (45), BioGRID (46) and iRefIndex (47), were used to interpret the Protein/Genetic Interaction. We also performed classic Fisher's exact test and utilized –log(P value) to score each GO term and Pathway category in the Gene Oncology and Pathway analyses. For the 'Pathway' and 'Protein/Genetic Interaction' sections in the DriverDB web interface, the Cytoscape Web (48) tool was embedded for interactive network visualization.

WEB INTERFACE

Cancer

The 'Cancer' section stored the calculated results of driver genes for a specific cancer type/dataset. First, users can define the data type(s) incorporated for driver gene

identification (the red rectangle in Supplementary Figure S1A) and then select a specific dataset, for example, 'Glioblastoma multiforme' (GBM). The result section will then indicate the detailed information of the specific dataset (red circle in Supplementary Figure S1B). Users can select a driver gene set identified by '*N*' methods (the 'Summary' in Supplementary Figure S1B; '*N*' is determined by a drop-down menu) or by individual methods according to the name of the method (Supplementary Figure S1B). For 'Summary', a heat map shows the relationship between genes and methods (Supplementary Figure S1C; the blue color indicates genes identified as driver genes by a method). For each driver gene set, there is a heat map showing a mutation profile of that driver gene set of samples (Supplementary Figure S1D). We also performed functional analysis in three levels of biological interpretation: 'Gene Ontology', 'Pathway' and 'Protein/Genetics Interaction'. In the 'Gene Ontology' analysis, **I** and **II** indicate the topology of GO graph by topGO and GeneAnswers, respectively, (Supplementary Figure S1E) whereas **III** and **IV** show the most significant GO terms and genes. The table in Supplementary Figure S1E lists the information of all the significant GO terms. In the 'Pathway' analysis, there are eight collections of gene sets from public databases including KEGG, REACTOME, MSigDB, PID and Biocarta. For each collection, there is a network visualization and a table displaying pathway categories of the driver genes that are involved (Supplementary Figure S1F). Finally, in the 'Protein/Genetics Interaction' part, the interactions between driver genes are illustrated according to three resources: BioGRID, IntAct and iRefIndex (Supplementary Figure S1G).

Gene

In this section, researchers can visualize the mutation data for a specific protein encoded by a gene in five different kinds of aspects: Mutation Profile, Mutation Percentage, Exon, Driver Score and Mutation Information (Supplementary Figure S2A). Here, we use the gene PIK3CA, which is identified as a driver gene in the 'Cancer' section, as an example. Bar chart colors in the sub-figures of Supplementary Figure S2 indicate the functional impact of a mutation, such as non-synonymous and frame-shift shown in Supplementary Figure S2B. For 'Mutation Profile' (Supplementary Figure S2C), a heat map shows the mutation rate calculated by the mutation count/sample count of a cancer, at different protein positions across several cancer types. We also provide exon and domain information with protein coordinates at the bottom of the heat map (Supplementary Figure S2C). Two bar charts located at the top and the left of the heat map indicate the sum of mutation rate according to protein position and cancer type, respectively. The 'Mutation Percentage' (Supplementary Figure S2D) is similar to Supplementary Figure S2C, but the number in the heat map is calculated by the following: (mutation count of a protein region/total mutation count of a cancer) $\times$ 100. The heights of the two bar charts at the left and the top of the heat map are normalized to the mutation count of a cancer type or a protein region, respectively. In the 'Exon' panel, the mutation counts and the mutation types of each exon are illustrated in Supplementary Figure S2E and S2F, respectively. For the 'Driver Score' part, Supplementary Figure S2G and S2H indicate the Driver Score (please see the 'Materials and Methods' section for details) distributions of exons and protein positions, respectively. All the mutation data of a specific protein are listed under 'Mutation Information' (Supplementary Figure S2A).

Meta-analysis

In addition to the stored calculated results, DriverDB allows researchers to identify driver genes of a user-defined, specific set of samples. As shown in Supplementary Figure S3, users can select one or multiple datasets in DriverDB. We provide a list of clinical criteria, such as ICD-O-3 histology, tumor stage, distant metastasis and lymph node status, to help researchers to select a sub-group of well-defined cancer samples according to one or multiple clinical parameters for driver gene identification. Users can overview the detailed clinical information of selected samples before submitting this job to the server for real-time calculation. The user will receive a notification email with a Result ID, and then visualized driver gene results in the 'Result and Download' section when the job is completed.

DISCUSSION

DriverDB makes the best of the massive amount of exome-seq data published in recent years by integrating driver gene analysis from numerous methods, as well as by providing visualizations of mutation information according to different aspects. As described in the 'Introduction' section, different bioinformatics algorithms have been developed to identify driver genes based on several assumptions and characteristics, each of which provides different points of view regarding driver genes. DriverDB integrates the analysis results of individual/multiple method(s) and provides three levels of biological interpretation: Gene Oncology, Pathway and Protein/Genetics Interaction. These visualization results will help users to quickly realize the relationships between driver genes. A representative example of driver genes identified in GBM is shown in Supplementary Figure S1. A total of 14 driver genes were identified (each gene by at least 4 methods), and nearly all samples had at least 1 deleterious mutation among these 14 genes. Ten genes (CDKN2A, EGFR, PTEN, TP53, CDK4, PIK3R1, NF1, PIK3CA, RB1 and IDH1) are known to be critical in GBM tumorigenesis (49,50). For the other four genes (ATRX, CHEK2, CPSF6 and COL6A3), our functional analysis shows that they are involved in cell cycle-related categories (Supplementary Figure S1F). Moreover, ATRX has been reported as the driver gene in pediatric glioblastomas (51) and neuroblastomas (52,53). CHEK2 is relevant to familial breast/ovarian cancer (54) and neuroblastomas (54). CPSF6 can either enhance the invasive capacities of or inhibit the proliferation of cancer cells (55). The spliced

variants and the aberrant methylation of COL6A3, are also related to cancers (56–58). Genes reported in other references but not included in our 14-gene list can be identified by less stringent criteria (such as those identified by at least three methods; for example PDGFRA, MDM2, MDM4 and CDKN2B).

The 'Gene' section is designed to help researchers to visualize the mutation data of a driver gene. The representative example is PIK3CA, a well-known driver gene in GBM as well as in other cancers (Supplementary Figure S2). It is easy to find that there are two hotspot mutation regions (at the middle and the end of the protein), especially in the 'Mutation Percentage' figure (Supplementary Figure S2D). The two well-known driver genes, BRAF and KRAS, also have the same characteristics (Supplementary Figure S4). However, a driver gene may have distinct hotspot mutation regions in different cancers. For example, unlike lung cancers that carry EGFR mutations at the kinase domain (KD), activation of EGFR in GBM occurs through mutation at the extracellular domain (59). This has been noted as the reason that GBM with mutations in the extracellular domain respond poorly to EGFR inhibitors (e.g., erlotinib) that target the active kinase conformation (59). This phenomena was recaptured by our calculation and was present in the 'Mutation Profile' of EGFR in DriverDB (Supplementary Figure S5).

In the 'Gene' section, bar chart colors indicate the functional impact of a mutation, which can help to convey important information. For example, FLT3 has been reported to be mutated in approximately one-third of patients in acute myeloid leukemia and has two hotspot regions: one consists of internal tandem duplication (ITD) mutations of 3–400 bp (always in-frame), and the other consists of point mutations at aspartic acid 835 of the KD (60). Such mutation information for FLT3 can be easily obtained in DriverDB (Supplementary Figure S6).

Several studies have assessed the performance of existing tools for predicting deleterious mutations, and the results have demonstrated that identifying cancer-driving mutations remains a significant challenge (5,61). Hence, we used the 'Driver Score', which integrates the information from seven computational tools, to describe the deleterious level of a mutation and to highlight the hotspot mutation region. For example, the Driver Score distribution of the cancer-related gene 'MLL2' implies that the third region of the MLL2 protein plays a more important role than other positions (Supplementary Figure S7). In summary, in the 'Gene' section of DriverDB, researchers can easily be informed when mutations are concentrated in one/some specific protein position(s)/domain(s)/exon(s)/cancer(s).

The 'Meta-analysis' section allows a user to re-define a group of samples from one/multiple datasets and then identify driver genes for selected samples. It has been noted that mutations are accumulated during tumor progression. Different driver mutations may be used to convert a normal cell to a tumor cell, or to turn a benign tumor into a malignant one. The timing of mutations is relevant to metastasis, and there are mutations that occur during this process (1). Thus, if we could define samples by a clarified biological or clinical goal, we would have the opportunity to identify a specific set of driver genes for a distinct question. To achieve this, DriverDB offers a list of clinical characteristics to define samples and provides a high degree of freedom for researchers to utilize the huge amount of sequencing data. For example, in Supplementary Figure S3 we selected only 180 samples from TCGA breast cancer project. Their lymphonode pathologic spread and ICD oncology of histology are 'N0' and 'infiltrating duct carcinoma, NOS', respectively.

A number of databases and frameworks have been developed to integrate large-scale genomic data (2), including cBioportal (62,63) and IntOGen (64). cBioportal contains datasets from TCGA and provides gene-based search capabilities to interactively explore multidimensional cancer genomics data. IntOGen is a framework that integrates multidimensional data for the identification of genes and biological modules involved in cancer development. DriverDB incorporates a large-scale data mining work using these algorithms in one go, presents summarized driver genes, and provides different kinds of aspects for mutation visualization. Another unique part of DriverDB is that it also helps researchers to identify driver genes in a customer-defined manner.

NGS has become the norm for large-scale cancer research, and cancer exome-seq results will accumulate rapidly in the next few years. For example, TCGA will examine over 11,000 samples for 20 cancer types by the end of 2014. Due to the Publication Guidelines of TCGA (http://cancergenome.nih.gov/abouttcga/policies/publicationguidelines), parts of data from TCGA are excluded in DriverDB. As time goes by, data from TCGA, as well as from other cancer projects/literatures, will have no publication limitations and will be incorporated into updated DriverDB. We envision that these novel driver genes or mutations identified and stored in DriverDB will hold great potential for both basic research and biotech product development.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We acknowledge the contribution of Chien Shu and Kun Zhang from University of California at San Diego for the analysis of exome-seq; and thank the National Center for High-performance Computing for computer time and facilities.

FUNDING

National Science Council [NSC; NSC101-2320-B-010-059-MY3, NSC101-2627-B-010-003 and NSC101-2321-B-010-011]; Veterans General Hospitals University System of Taiwan (VGHUST) Joint Research Program; Tsou's Foundation [GHUST102-G7-3-2]; National Health Research Institutes (NHRI) [NHRI-EX102-10254SI];

Taipei Veteran General Hospital [Cancer Excellence Center Plan DOH102-TD-C-111-007]; Taipei City Hospital [10201-62-070]; National Yang-Ming University [Ministry of Education, Aim for the Top University Plan]; UST-UCSD International Center for Excellence in Advanced Bioengineering sponsored by the Taiwan NSC I-RiCE Program (in part) [NSC101-2911-I-009-101]. Funding for open access charge: NSC.

Conflict of interest statement. None declared.

REFERENCES

1. 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.
2. Shyr,D. and Liu,Q. (2013) Next generation sequencing in cancer research and clinical application. *Biol. Proced. Online*, **15**, 4.
3. Stratton,M.R., Campbell,P.J. and Futreal,P.A. (2009) The cancer genome. *Nature*, **458**, 719–724.
4. Bozic,I., Antal,T., Ohtsuki,H., Carter,H., Kim,D., Chen,S., Karchin,R., Kinzler,K.W., Vogelstein,B. and Nowak,M.A. (2010) Accumulation of driver and passenger mutations during tumor progression. *Proc. Natl Acad. Sci. USA*, **107**, 18545–18550.
5. Gnad,F., Baucom,A., Mukhyala,K., Manning,G. and Zhang,Z. (2013) Assessment of computational methods for predicting the effects of missense mutations in human cancers. *BMC Genomics*, **14(Suppl. 3)**, S7.
6. Carter,H., Chen,S., Isik,L., Tyekucheva,S., Velculescu,V.E., Kinzler,K.W., Vogelstein,B. and Karchin,R. (2009) Cancer-specific high-throughput annotation of somatic mutations: computational prediction of driver missense mutations. *Cancer Res.*, **69**, 6660–6667.
7. Zhao,H., Yang,Y., Lin,H., Zhang,X., Mort,M., Copper,D.N., Liu,Y. and Zhou,Y. (2013) DDIG-in: discriminating between disease-associated and neutral non-frameshifting micro-indels. *Genome Biol.*, **14**, R23.
8. Li,M.X., Kwan,J.S., Bao,S.Y., Yang,W., Ho,S.L., Song,Y.Q. and Sham,P.C. (2013) Predicting mendelian disease-causing non-synonymous single nucleotide variants in exome sequencing studies. *PLoS Genet.*, **9**, e1003143.
9. Lawrence,M.S., Stojanov,P., Polak,P., Kryukov,G.V., Cibulskis,K., Sivachenko,A., Carter,S.L., Stewart,C., Mermel,C.H., Roberts,S.A. *et al.* (2013) Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature*, **499**, 214–218.
10. Dees,N.D., Zhang,Q., Kandoth,C., Wendl,M.C., Schierding,W., Koboldt,D.C., Mooney,T.B., Callaway,M.B., Dooling,D., Mardis,E.R. *et al.* (2012) MuSiC: identifying mutational significance in cancer genomes. *Genome Res.*, **22**, 1589–1598.
11. Youn,A. and Simon,R. (2011) Identifying cancer driver genes in tumor genome sequencing studies. *Bioinformatics*, **27**, 175–181.
12. Gonzalez-Perez,A. and Lopez-Bigas,N. (2012) Functional impact bias reveals cancer drivers. *Nucleic Acids Res.*, **40**, e169.
13. Reimand,J. and Bader,G.D. (2013) Systematic analysis of somatic mutations in phosphorylation signaling predicts novel cancer drivers. *Mol. Syst. Biol.*, **9**, 637.
14. Michaelson,J.J., Shi,Y., Gujral,M., Zheng,H., Malhotra,D., Jin,X., Jian,M., Liu,G., Greer,D., Bhandari,A. *et al.* (2012) Whole-genome sequencing in autism identifies hot spots for de novo germline mutation. *Cell*, **151**, 1431–1442.
15. Nik-Zainal,S., Alexandrov,L.B., Wedge,D.C., Van Loo,P., Greenman,C.D., Raine,K., Jones,D., Hinton,J., Marshall,J., Stebbings,L.A. *et al.* (2012) Mutational processes molding the genomes of 21 breast cancers. *Cell*, **149**, 979–993.
16. Zhao,J., Zhang,S., Wu,L.Y. and Zhang,X.S. (2012) Efficient methods for identifying mutated driver pathways in cancer. *Bioinformatics*, **28**, 2940–2947.
17. Vandin,F., Upfal,E. and Raphael,B.J. (2012) De novo discovery of mutated driver pathways in cancer. *Genome Res.*, **22**, 375–385.
18. Leiserson,M.D., Blokh,D., Sharan,R. and Raphael,B.J. (2013) Simultaneous identification of multiple driver pathways in cancer. *PLoS Comput. Biol.*, **9**, e1003054.
19. Ciriello,G., Cerami,E., Sander,C. and Schultz,N. (2012) Mutual exclusivity analysis identifies oncogenic network modules. *Genome Res.*, **22**, 398–406.
20. Cerami,E., Demir,E., Schultz,N., Taylor,B.S. and Sander,C. (2010) Automated network analysis identifies core pathways in glioblastoma. *PLoS ONE*, **5**, e8918.
21. 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.
22. Trifonov,V., Pasqualucci,L., Dalla Favera,R. and Rabadan,R. (2013) MutComFocal: an integrative approach to identifying recurrent and focal genomic alterations in tumor samples. *BMC Syst. Biol.*, **7**, 25.
23. Dand,N., Sprengel,F., Ahlers,V. and Schlitt,T. (2013) BioGranat-IG: a network analysis tool to suggest mechanisms of genetic heterogeneity from exome-sequencing data. *Bioinformatics*, **29**, 733–741.
24. Miller,C.A., Settle,S.H., Sulman,E.P., Aldape,K.D. and Milosavljevic,A. (2011) Discovering functional modules by identifying recurrent and mutually exclusive mutational patterns in tumors. *BMC Med. Genomics*, **4**, 34.
25. Vandin,F., Upfal,E. and Raphael,B.J. (2012) Finding driver pathways in cancer: models and algorithms. *Algorithm. Mol. Biol.*, **7**, 23.
26. Sherry,S.T., Ward,M.H., Kholodov,M., Baker,J., Phan,L., Smigielski,E.M. and Sirotkin,K. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
27. Abecasis,G.R., Auton,A., Brooks,L.D., DePristo,M.A., Durbin,R.M., Handsaker,R.E., Kang,H.M., Marth,G.T. and McVean,G.A. (2012) An integrated map of genetic variation from 1,092 human genomes. *Nature*, **491**, 56–65.
28. Forbes,S.A., Bindal,N., Bamford,S., Cole,C., Kok,C.Y., Beare,D., Jia,M., Shepherd,R., Leung,K., Menzies,A. *et al.* (2011) COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.*, **39**, D945–D950.
29. Tennessen,J.A., Bigham,A.W., O'Connor,T.D., Fu,W., Kenny,E.E., Gravel,S., McGee,S., Do,R., Liu,X., Jun,G. *et al.* (2012) Evolution and functional impact of rare coding variation from deep sequencing of human exomes. *Science*, **337**, 64–69.
30. Hindorff,L.A., Sethupathy,P., Junkins,H.A., Ramos,E.M., Mehta,J.P., Collins,F.S. and Manolio,T.A. (2009) Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc. Natl Acad. Sci. USA*, **106**, 9362–9367.
31. Cooper,D.N., Stenson,P.D. and Chuzhanova,N.A. (2006) The Human Gene Mutation Database (HGMD) and its exploitation in the study of mutational mechanisms. *Curr. Protoc. Bioinformatics*, **Chapter 1**, Unit 1 13.
32. Cingolani,P., Platts,A., Wang le,L., Coon,M., Nguyen,T., Wang,L., Land,S.J., Lu,X. and Ruden,D.M. (2012) A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. *Fly*, **6**, 80–92.
33. McLaren,W., Pritchard,B., Rios,D., Chen,Y., Flicek,P. and Cunningham,F. (2010) Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics*, **26**, 2069–2070.
34. Ng,P.C. and Henikoff,S. (2001) Predicting deleterious amino acid substitutions. *Genome Res.*, **11**, 863–874.
35. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
36. Gonzalez-Perez,A. and Lopez-Bigas,N. (2011) Improving the assessment of the outcome of nonsynonymous SNVs with a consensus deleteriousness score, Condel. *Am. J. Hum. Genet.*, **88**, 440–449.

37. Chun,S. and Fay,J.C. (2009) Identification of deleterious mutations within three human genomes. *Genome Res.*, **19**, 1553–1561.
38. Shihab,H.A., Gough,J., Cooper,D.N., Day,I.N. and Gaunt,T.R. (2013) Predicting the functional consequences of cancer-associated amino acid substitutions. *Bioinformatics*, **29**, 1504–1510.
39. Reva,B., Antipin,Y. and Sander,C. (2011) Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic Acids Res.*, **39**, e118.
40. Schwarz,J.M., Rodelsperger,C., Schuelke,M. and Seelow,D. (2010) MutationTaster evaluates disease-causing potential of sequence alterations. *Nat. Methods*, **7**, 575–576.
41. Kanehisa,M., Goto,S., Sato,Y., Furumichi,M. and Tanabe,M. (2012) KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res.*, **40**, D109–D114.
42. Schaefer,C.F., Anthony,K., Krupa,S., Buchoff,J., Day,M., Hannay,T. and Buetow,K.H. (2009) PID: the Pathway Interaction Database. *Nucleic Acids Res.*, **37**, D674–D679.
43. Croft,D., O'Kelly,G., Wu,G., Haw,R., Gillespie,M., Matthews,L., Caudy,M., Garapati,P., Gopinath,G., Jassal,B. *et al.* (2011) Reactome: a database of reactions, pathways and biological processes. *Nucleic Acids Res.*, **39**, D691–D697.
44. Subramanian,A., Tamayo,P., Mootha,V.K., Mukherjee,S., Ebert,B.L., Gillette,M.A., Paulovich,A., Pomeroy,S.L., Golub,T.R., Lander,E.S. *et al.* (2005) Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc. Natl Acad. Sci. USA*, **102**, 15545–15550.
45. Kerrien,S., Aranda,B., Breuza,L., Bridge,A., Broackes-Carter,F., Chen,C., Duesbury,M., Dumousseau,M., Feuermann,M., Hinz,U. *et al.* (2012) The IntAct molecular interaction database in 2012. *Nucleic Acids Res.*, **40**, D841–D846.
46. Chatr-Aryamontri,A., Breitkreutz,B.J., Heinicke,S., Boucher,L., Winter,A., Stark,C., Nixon,J., Ramage,L., Kolas,N., O'Donnell,L. *et al.* (2013) The BioGRID interaction database: 2013 update. *Nucleic Acids Res.*, **41**, D816–D823.
47. Razick,S., Magklaras,G. and Donaldson,I.M. (2008) iRefIndex: a consolidated protein interaction database with provenance. *BMC Bioinformatics*, **9**, 405.
48. Lopes,C.T., Franz,M., Kazi,F., Donaldson,S.L., Morris,Q. and Bader,G.D. (2010) Cytoscape Web: an interactive web-based network browser. *Bioinformatics*, **26**, 2347–2348.
49. Verhaak,R.G., Hoadley,K.A., Purdom,E., Wang,V., Qi,Y., Wilkerson,M.D., Miller,C.R., Ding,L., Golub,T., Mesirov,J.P. *et al.* (2010) Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell*, **17**, 98–110.
50. Cancer Genome Atlas Research Network. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature*, **455**, 1061–1068.
51. Schwartzentruber,J., Korshunov,A., Liu,X.Y., Jones,D.T., Pfaff,E., Jacob,K., Sturm,D., Fontebasso,A.M., Quang,D.A., Tonjes,M. *et al.* (2012) Driver mutations in histone H3.3 and chromatin remodelling genes in paediatric glioblastoma. *Nature*, **482**, 226–231.
52. Pugh,T.J., Morozova,O., Attiyeh,E.F., Asgharzadeh,S., Wei,J.S., Auclair,D., Carter,S.L., Cibulskis,K., Hanna,M., Kiezun,A. *et al.* (2013) The genetic landscape of high-risk neuroblastoma. *Nat. Genet.*, **45**, 279–284.
53. Cheung,N.K. and Dyer,M.A. (2013) Neuroblastoma: developmental biology, cancer genomics and immunotherapy. *Nat. Rev. Cancer*, **13**, 397–411.
54. Rashid,M.U., Muhammad,N., Faisal,S., Amin,A. and Hamann,U. (2013) Constitutional CHEK2 mutations are infrequent in early-onset and familial breast/ovarian cancer patients from Pakistan. *BMC Cancer*, **13**, 312.
55. Yu,K., Ganesan,K., Tan,L.K., Laban,M., Wu,J., Zhao,X.D., Li,H., Leung,C.H., Zhu,Y., Wei,C.L. *et al.* (2008) A precisely regulated gene expression cassette potently modulates metastasis and survival in multiple solid cancers. *PLoS Genet.*, **4**, e1000129.
56. Maekawa,R., Sato,S., Yamagata,Y., Asada,H., Tamura,I., Lee,L., Okada,M., Tamura,H., Takaki,E., Nakai,A. *et al.* (2013) Genome-wide DNA methylation analysis reveals a potential mechanism for the pathogenesis and development of uterine leiomyomas. *PLoS ONE*, **8**, e66632.
57. Gardina,P.J., Clark,T.A., Shimada,B., Staples,M.K., Yang,Q., Veitch,J., Schweitzer,A., Awad,T., Sugnet,C., Dee,S. *et al.* (2006) Alternative splicing and differential gene expression in colon cancer detected by a whole genome exon array. *BMC Genomics*, **7**, 325.
58. Arafat,H., Lazar,M., Salem,K., Chipitsyna,G., Gong,Q., Pan,T.C., Zhang,R.Z., Yeo,C.J. and Chu,M.L. (2011) Tumor-specific expression and alternative splicing of the COL6A3 gene in pancreatic cancer. *Surgery*, **150**, 306–315.
59. Vivanco,I., Robins,H.I., Rohle,D., Campos,C., Grommes,C., Nghiemphu,P.L., Kubek,S., Oldrini,B., Chheda,M.G., Yannuzzi,N. *et al.* (2012) Differential sensitivity of glioma- versus lung cancer-specific EGFR mutations to EGFR kinase inhibitors. *Cancer Disc.*, **2**, 458–471.
60. Small,D. (2006) FLT3 mutations: biology and treatment. *Hematology / the Education Program of the American Society of Hematology. American Society of Hematology. Education Program*, 178–184.
61. Gonzalez-Perez,A., Deu-Pons,J. and Lopez-Bigas,N. (2012) Improving the prediction of the functional impact of cancer mutations by baseline tolerance transformation. *Genome Med.*, **4**, 89.
62. 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 Disc.*, **2**, 401–404.
63. 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.
64. Gundem,G., Perez-Llamas,C., Jene-Sanz,A., Kedzierska,A., Islam,A., Deu-Pons,J., Furney,S.J. and Lopez-Bigas,N. (2010) IntOGen: integration and data mining of multidimensional oncogenomic data. *Nat. Methods*, **7**, 92–93.