Colorectal cancer atlas: An integrative resource for genomic and proteomic annotations from colorectal cancer cell lines and tissues

David Chisanga[1], Shivakumar Keerthikumar[2], Mohashin Pathan[2], Dinuka Ariyaratne[2], Hina Kalra[2], Stephanie Boukouris[2], Nidhi Abraham Mathew[2], Haidar Al Saffar[2], Lahiru Gangoda[2], Ching-Seng Ang[3], Oliver M. Sieber[4,5], John M. Mariadason[6,7,8], Ramanuj Dasgupta[9], Naveen Chilamkurti[1] and Suresh Mathivanan[2,*]

[1]Department of Computer Science and Information Technology, La Trobe University, Bundoora, Victoria 3086, Australia, [2]Department of Biochemistry and Genetics, La Trobe Institute for Molecular Science, La Trobe University, Melbourne, Victoria 3086, Australia, [3]The Bio21 Molecular Science and Biotechnology Institute, University of Melbourne, Parkville, Victoria 3010, Australia, [4]Walter and Eliza Hall Institute of Medical Research, Parkville, Victoria 3052, Australia, [5]Faculty of Medicine, Dentistry and Health Sciences, Department of Medical Biology, University of Melbourne, Parkville, Victoria 3052, Australia, [6]Olivia Newton John Cancer Research Institute, Melbourne, Victoria 3084, Australia, [7]Ludwig Institute for Cancer Research, Melbourne-Austin Branch, Victoria 3084, Australia, [8]School of Cancer Medicine, La Trobe University, Melbourne, Victoria 3084, Australia and [9]Genome Institute of Singapore, A*STAR, 60 Biopolis Street, Singapore 138672, Singapore

ABSTRACT

In order to advance our understanding of colorectal cancer (CRC) development and progression, biomedical researchers have generated large amounts of OMICS data from CRC patient samples and representative cell lines. However, these data are deposited in various repositories or in supplementary tables. A database which integrates data from heterogeneous resources and enables analysis of the multidimensional data sets, specifically pertaining to CRC is currently lacking. Here, we have developed Colorectal Cancer Atlas (http://www.colonatlas.org), an integrated web-based resource that catalogues the genomic and proteomic annotations identified in CRC tissues and cell lines. The data catalogued to-date include sequence variations as well as quantitative and non-quantitative protein expression data. The database enables the analysis of these data in the context of signaling pathways, protein–protein interactions, Gene Ontology terms, protein domains and post-translational modifications. Currently, Colorectal Cancer Atlas contains data for >13 711 CRC tissues, >165 CRC cell lines, 62 251 protein identifications, >8.3 million MS/MS spectra, >18 410 genes with sequence variations (404 278 entries) and 351 pathways with sequence variants. Overall, Colorectal Cancer Atlas has been designed to serve as a central resource to facilitate research in CRC.

INTRODUCTION

Colorectal cancer (CRC) is the third most common form of cancer and has the fourth highest mortality rate in the world (1). In order to advance our understanding of the initiation and progression of this disease, biomedical researchers have performed global analyses of the genome, epigenome, transcriptome, proteome and metabolome of CRC patient samples and representative cell lines (2–5). According to The Cancer Genome Atlas Network (3), APC, TP53, KRAS, PIK3CA, FBXW7, SMAD4, TCF7L2 and NRAS are the most frequently mutated genes in CRC. Identification of these mutations and associated pathways has advanced our understanding of CRC, is enabling the sub-classification of this disease and is unveiling potential new avenues for treatment.

Due to the significant advancements in high-throughput technologies, vast amounts of multidimensional data relevant to the biology of CRC have been generated. To extract meaningful biological insights from these data, researchers previously needed to collate data from a large number of studies. To facilitate this process, a series of databases have been created. For example, cancer gene mutations are currently catalogued in databases including TCGA (3), COS-

Table 1. Colorectal cancer atlas statistics

Protein entries	62 251
MS/MS spectra	8 378 422
Primary tissues	13 711
Cell lines	165
Genes with sequence variants	18 410
Gene sequence variants	404 278
Pathways with genes having sequence variants	351
Pathways with genes having no sequence variants	1657
Cell lines with drug sensitivity	27
PTMs	88 819
PTMs affected by sequence variants	1631
Protein–protein interactions	253 700

MIC (6), TumorPortal (7), IntOGen (8), Network of Cancer Genes (9) and TSGene (10). These databases provide valuable information of gene variations for a number of tumor types including CRC, however, they are not specifically designed to integrate sequence variations with proteomic data. NetGestal (11) is a web-based framework that allows for integration of OMIC data from multiple species in the context of biological networks (12) and contains data pertaining to human CRC from TCGA. However, there is currently no user-friendly online resource specifically pertaining to CRC which catalogues genomic and proteomic data from literature, databases and TCGA, integrates the sequence variations with protein domain, post-translational modifications and protein–protein interactions.

Here, we describe Colorectal Cancer Atlas (http://www.colonatlas.org), an integrated web-based resource which catalogues genomic and proteomic data from CRC tissues and cell lines. Data catalogued include; quantitative and non-quantitative protein expression, sequence variations, cellular signaling pathways, protein–protein interactions, Gene Ontology terms, protein domains and post-translational modifications (PTMs). Data pertaining to genomic sequence variations and protein expression have been manually curated from the scientific literature and collated from other publicly available databases. Colorectal Cancer Atlas is designed to enable a user to search for a specific mutation in any particular cell line, and search for cell lines with and without specific mutations. Currently, Colorectal Cancer Atlas contains data for >13 711 primary CRC tissues, >165 CRC cell lines, 62 251 protein identifications, >8.3 million MS/MS spectra, >18 410 genes with sequence variations, 404 278 sequence variation entries, 351 pathways with sequence variants, 88 819 PTMs and 253 700 protein–protein interactions (Table 1).

DATABASE ARCHITECTURE AND WEB INTERFACE

Colorectal Cancer Atlas is a web-based application developed using Zope2 (version 2.8.7–1), a python-based web framework. The back end database is MySQL (version 5.0.95), a well-established open source database. The web pages were developed using Hyper Text Markup Language (HTML) in combination with JavaScript for front end functionality, while Python (version 2.4.3), a scripting language was used for database connectivity. JavaScript modules include DataTables (version 1.10.4) for the development of interactive data tables, Data-Driven Documents (D3JS) for the development of interactive protein–protein interaction networks, and Highcharts (version 4.1.6) for the development of interactive heat maps and column charts.

GENOMIC DATA SETS

Colorectal Cancer Atlas catalogues gene sequence variations present in primary CRC tissues and cell lines which were collated by manual curation of the scientific literature. In addition, the database contains genomic variations identified in CRC cell lines sequenced in-house. For cell lines, where available, the gender and age of the patient is provided, along with the specific cell type, doubling time, culture properties and stage of cancer. This information was obtained from the Cancer Cell Line Encyclopedia (13), ATCC (http://www.atcc.org), COSMIC database and literature. Sequence variation details including the type of sequence variants, putative mutational effects, nucleotide change and amino acid changes are displayed.
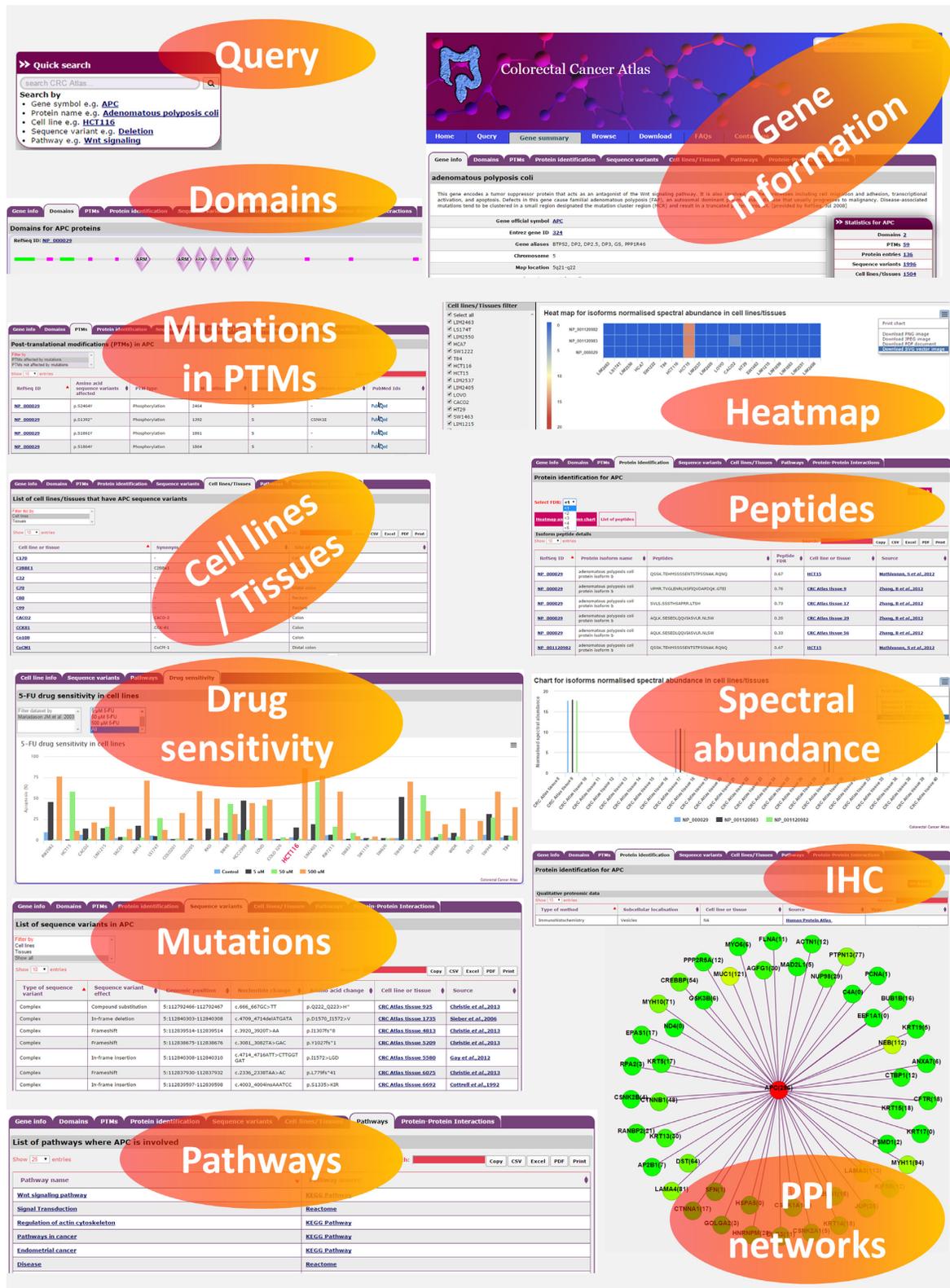
PROTEOMIC DATA SETS

Colorectal Cancer Atlas also catalogues proteomic data collated from multiple resources including the scientific literature (e.g. Zhang *et al.* (5)), Human Protein Atlas (14), Human Proteinpedia (15) and Human Protein Reference Database (16). Experimental techniques used in generating these data included mass spectrometry, Western blotting, immunohistochemistry, confocal microscopy, immunoelectron microscopy and fluorescence-activated cell sorting (FACS). In addition, publicly available label-free quantitative mass spectrometry data for CRC cell lines and tissues were re-analyzed using an in-house proteomics pipeline in order to provide standardized data. The proteomics pipeline involved conversion of raw mass spectrometry data files into the Mascot Generic File Format (MGF) using MsConvert with peak picking (17). The MGF files were then searched using X! Tandem (Sledgehammer edition version 2013.09.01.1) (18) against a target and decoy Human RefSeq protein database. Peptides were further filtered using <5% false discovery rate (FDR) as a cut-off, and quantified using the Normalized Spectral Abundance Factor (NSAF) method (19).

COLORECTAL CANCER ATLAS PROVIDES AN INTEGRATED VIEW OF MULTIPLE DATA TYPES

Colorectal Cancer Atlas provides an integrated view of the sequence variations and the proteomic data. Mass spectrometry-based quantitative proteomic data are depicted as heat maps and column charts in the respective molecular pages (Figure 1), and users are able to filter the data sets based on the FDR. The database also contains protein expression data generated using immunohistochemistry, Western blotting, FACS, confocal and immunoelectron microscopy. The database also includes protein data derived from various cellular fractions including the nucleus, cytoplasm, membrane, the secretome (20) and exosomes (21) (from ExoCarta (22)).

The integration of sequence variants with proteomic data is designed to facilitate the prediction of functional effects of the protein. For each gene, Colorectal Cancer Atlas

Figure 1. Snapshot of Colorectal Cancer Atlas features. An overview of proteomic and genomic data features for APC gene is displayed. A user can query the database using a gene symbol or a protein name. A gene information page will provide the users with details pertaining to protein domains, post-translational modifications (PTM), reported mutations in cell lines/tissues, quantitative protein expression, pathway, protein–protein interaction (PPI) and cell line drug sensitivity.

Figure 2. PTMs and domains in β-catenin are affected due to mutation. Snapshot of β-catenin molecular page is displayed. The PTMs affected by mutations can be viewed in the tab PTMs. Mutations in β-catenin at positions important for phosphorylation (S33, S37, T41 and S45) allows for the stabilization of β-catenin and constitutive activation of the Wnt signaling pathway. The upstream kinases responsible for the phosphorylation is also provided along with the literature reference. Likewise, mutations in the armadillo domain can be viewed by correlating the sequence variants and the domain span regions. For example, mutations in the armadillo domain (p.R582) in β-catenin have been described which have been reported to alter the binding of β-catenin to TCF4 (24).

enables parallel visualization of CRC-associated sequence variants with quantitative protein expression across CRC cell lines and tissues. In addition, PTMs, and protein domains affected by the sequence variation can be visualized (Figure 1), enabling the potential effect of sequence variants on protein function to be easily ascertained. For example, β-catenin mutations in positions S33, S37, T41 and S45 occur in CRC, all of which are critical for phosphorylation (23). Mutations in these serine/threonine residues allow for the stabilization of β-catenin and constitutive activation of the Wnt signaling pathway. Similarly, Colorectal Cancer Atlas displays sequence variations in known protein domains which can provide valuable insight into the putative effect on protein function. For example, mutations in the armadillo domain (R582) in β-catenin have been described which have been reported to alter the binding of β-catenin to TCF4 (24) (Figure 2).

Colorectal Cancer Atlas also provides a graphical representation of known protein interactions (obtained from BioGrid (25) and Human Protein Resource Database (16)), where each protein is depicted as a node with a specific colour and intensity corresponding to the number of sequence variants in the encoding gene (Figure 1). Furthermore, Colorectal Cancer Atlas integrates biological pathways with gene sequence variants. Biological Pathways were obtained from Reactome (26), KEGG (27), Cell map and HumanCyc. For example, as shown in Figure 1, sequence variants in APC are implicated in dysregulation of the Wnt signaling pathway and actin cytoskeletal remodeling. Finally, Colorectal Cancer Atlas contains data on 5-flurouracil (5-FU) drug sensitivity for CRC cell lines curated from the literature (studies using at least three CRC cell lines (28)). Users can view the sensitivity profile of a cell line of interest relative to other CRC cells.

ACCESSING COLORECTAL CANCER ATLAS

Users can search Colorectal Cancer Atlas through the home, query or browse pages (Supplementary Figure S1). In addition, the website features a navigation menu and a search box at the top of the page. The database can be queried by gene symbol, Entrez Gene ID, protein name, cell line name or pathway. The browse page provides users with the option to access the database by categorized lists of genes, sequence variations, cell lines and techniques. The browse page allows the users to search for sequence variations in genes of interest and displays them in interactive color-coded table format. The gene information page includes gene details, associated GO terms, sequence variations (displayed in an interactive table), domain details, PTMs, a protein data page leading to experimental techniques and quantitative data with an interactive heat map, a column chart for spectral abundance and a list of detected peptides. Other information includes a list of cell lines and tissues that contain sequence variants in a given gene, a list of pathways in which the gene is involved, and an interactive protein–protein interaction network for the protein encoded by the gene. The cell line page provides details of the cell line, an interactive table of gene sequence variants identified in the cell line, an interactive table of dysregulated pathways and 5-FU drug sensitivity profile. Data curated

in Colorectal Cancer Atlas are available as tab-delimited files and is free for download to all users. Using the custom database option, the tab delimited data can also be uploaded into FunRich (29), a functional enrichment analysis tool to identify classes of genes/proteins that are overrepresented in a specific category.

FUTURE DIRECTIONS

Colorectal Cancer Atlas will be continuously updated with more studies as they become available and additional features. Studies currently being curated include Wnt signaling activity determined by the TOPFLASH assay, and genomic and proteomic data generated from patient derived xenografts.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

FUNDING

REFERENCES

1. Jemal,A., Bray,F., Center,M.M., Ferlay,J., Ward,E. and Forman,D. (2011) Global cancer statistics. *CA Cancer J. Clin.*, **61**, 69–90.
2. Atlas,T.C.G. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.
3. Cancer Genome Atlas Network. (2012) Comprehensive molecular characterization of human colon and rectal cancer. *Nature*, **487**, 330–337.
4. Sadanandam,A., Lyssiotis,C.A., Homicsko,K., Collisson,E.A., Gibb,W.J., Wullschleger,S., Ostos,L.C., Lannon,W.A., Grotzinger,C., Del Rio,M. *et al.* (2013) A colorectal cancer classification system that associates cellular phenotype and responses to therapy. *Nat. Med.*, **19**, 619–625.
5. Zhang,B., Wang,J., Wang,X., Zhu,J., Liu,Q., Shi,Z., Chambers,M.C., Zimmerman,L.J., Shaddox,K.F., Kim,S. *et al.* (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature*, **513**, 382–387.
6. 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.
7. Lawrence,M.S., Stojanov,P., Mermel,C.H., Robinson,J.T., Garraway,L.A., Golub,T.R., Meyerson,M., Gabriel,S.B., Lander,E.S. and Getz,G. (2014) Discovery and saturation analysis of cancer genes across 21 tumour types. *Nature*, **505**, 495–501.
8. 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.
9. An,O., Pendino,V., D'Antonio,M., Ratti,E., Gentilini,M. and Ciccarelli,F.D. (2014) NCG 4.0: the network of cancer genes in the era of massive mutational screenings of cancer genomes. *Database*, **2014**, doi:10.1093/database/bau015.
10. Zhao,M., Sun,J. and Zhao,Z. (2013) TSGene: a web resource for tumor suppressor genes. *Nucleic Acids Res.*, **41**, D970–D976.

11. Shi,Z., Wang,J. and Zhang,B. (2013) NetGestalt: integrating multidimensional omics data over biological networks. *Nat. Methods*, **10**, 597–598.

12. Zhu,J., Shi,Z., Wang,J. and Zhang,B. (2015) Empowering biologists with multi-omics data: colorectal cancer as a paradigm. *Bioinformatics*, **31**, 1436–1443.

13. Barretina,J., Caponigro,G., Stransky,N., Venkatesan,K., Margolin,A.A., Kim,S., Wilson,C.J., Lehar,J., Kryukov,G.V., Sonkin,D. *et al.* (2012) The Cancer Cell Line Encyclopedia enables predictive modelling of anticancer drug sensitivity. *Nature*, **483**, 603–307.

14. Uhlen,M., Fagerberg,L., Hallstrom,B.M., Lindskog,C., Oksvold,P., Mardinoglu,A., Sivertsson,A., Kampf,C., Sjostedt,E., Asplund,A. *et al.* (2015) Proteomics. Tissue-based map of the human proteome. *Science*, **347**, 1260419.

15. Mathivanan,S., Ahmed,M., Ahn,N.G., Alexandre,H., Amanchy,R., Andrews,P.C., Bader,J.S., Balgley,B.M., Bantscheff,M., Bennett,K.L. *et al.* (2008) Human Proteinpedia enables sharing of human protein data. *Nat. Biotechnol.*, **26**, 164–167.

16. Keshava Prasad,T.S., Goel,R., Kandasamy,K., Keerthikumar,S., Kumar,S., Mathivanan,S., Telikicherla,D., Raju,R., Shafreen,B., Venugopal,A. *et al.* (2009) Human protein reference database–2009 update. *Nucleic Acids Res.*, **37**, D767–D772.

17. Chambers,M.C., Maclean,B., Burke,R., Amodei,D., Ruderman,D.L., Neumann,S., Gatto,L., Fischer,B., Pratt,B., Egertson,J. *et al.* (2012) A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.*, **30**, 918–920.

18. Craig,R. and Beavis,R.C. (2004) TANDEM: matching proteins with tandem mass spectra. *Bioinformatics*, **20**, 1466–1467.

19. Paoletti,A.C., Parmely,T.J., Tomomori-Sato,C., Sato,S., Zhu,D., Conaway,R.C., Conaway,J.W., Florens,L. and Washburn,M.P. (2006) Quantitative proteomic analysis of distinct mammalian Mediator complexes using normalized spectral abundance factors. *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 18928–18933.

20. Mathivanan,S., Ji,H., Tauro,B.J., Chen,Y.S. and Simpson,R.J. (2012) Identifying mutated proteins secreted by colon cancer cell lines using mass spectrometry. *J. Proteomics*, **76**, 141–149.

21. Keerthikumar,S., Gangoda,L., Liem,M., Fonseka,P., Atukorala,I., Ozcitti,C., Mechler,A., Adda,C.G., Ang,C.S. and Mathivanan,S. (2015) Proteogenomic analysis reveals exosomes are more oncogenic than ectosomes. *Oncotarget*, **6**, 15375–15396.

22. Keerthikumar,S., Chisanga,D., Ariyaratne,D., Al Saffar,H., Anand,S., Zhao,K., Samuel,M., Pathan,M., Jois,M., Chilamkurti,N. *et al.* (2015) ExoCarta: a web-based compendium of exosomal cargo. *J. Mol. Biol.*, doi:10.1016/j.jmb.2015.09.019.

23. Wang,Z., Vogelstein,B. and Kinzler,K.W. (2003) Phosphorylation of beta-catenin at S33, S37, or T41 can occur in the absence of phosphorylation at T45 in colon cancer cells. *Cancer Res.*, **63**, 5234–5235.

24. Fasolini,M., Wu,X., Flocco,M., Trosset,J.Y., Oppermann,U. and Knapp,S. (2003) Hot spots in Tcf4 for the interaction with beta-catenin. *J. Biol. Chem.*, **278**, 21092–21098.

25. Chatr-aryamontri,A., Breitkreutz,B.-J., Oughtred,R., Boucher,L., Heinicke,S., Chen,D., Stark,C., Breitkreutz,A., Kolas,N., O'Donnell,L. *et al.* (2015) The BioGRID interaction database: 2015 update. *Nucleic Acids Res.*, **43**, D470–D478.

26. Milacic,M., Haw,R., Rothfels,K., Wu,G., Croft,D., Hermjakob,H., D'Eustachio,P. and Stein,L. (2012) Annotating Cancer Variants and Anti-Cancer Therapeutics in Reactome. *Cancers*, **4**, 1180.

27. Kanehisa,M., Goto,S., Sato,Y., Kawashima,M., Furumichi,M. and Tanabe,M. (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–D205.

28. Mariadason,J.M., Arango,D., Shi,Q., Wilson,A.J., Corner,G.A., Nicholas,C., Aranes,M.J., Lesser,M., Schwartz,E.L. and Augenlicht,L.H. (2003) Gene expression profiling-based prediction of response of colon carcinoma cells to 5-fluorouracil and camptothecin. *Cancer Res.*, **63**, 8791–8812.

29. Pathan,M., Keerthikumar,S., Ang,C.S., Gangoda,L., Quek,C.Y., Williamson,N.A., Mouradov,D., Sieber,O.M., Simpson,R.J., Salim,A. *et al.* (2015) FunRich: An open access standalone functional enrichment and interaction network analysis tool. *Proteomics*, **15**, 2597–2601.