Systematic associations between germ-line mutations and human cancers

Mohamad Al-Shammari

School of Electrical Engineering and Computer Science
Faculty of Engineering and Informatics
University of Bradford,
Bradford, West Yorkshire, BD7 1DP, UK
Email: malsham1@bradford.ac.uk

Desmond J. Tobin

Centre for Skin Sciences
Faculty of Life Sciences
University of Bradford,
Centre for Skin Sciences
Bradford, West Yorkshire, BD7 1DP, UK
Email: D.tobin@bradford.ac.uk

Yonghong Peng*

School of Electrical Engineering and Computer Science,
Faculty of Engineering and Informatics,
University of Bradford,
Bradford, West Yorkshire, BD7 1DP, UK
Email: Y.H.Peng@bradford.ac.uk
*Corresponding author

Abstract: The revolution in Big Data has opened the gate for new research challenges in biomedical science. The aim of this study was to investigate whether germ-line gene mutations are a significant factor in 29 major primary human cancers. Using data obtained from multiple biological databases, we identified 424 genes from 8879 cancer mutation records. By integrating these gene mutation records a human cancer map was constructed from which several key results were obtained. These include the observations that missense/nonsense and regulatory mutations might play central role in connecting cancers/genes, and tend to be distributed in all chromosomes. This suggests that, of all mutation classes missense/nonsense and regulatory mutation classes are over-expressed in human genome and so are likely to have a significant impact on human cancer etiology and pathomechanism. This offers new insights into how the distribution and interconnections of gene mutations influence the development of cancers.

Keywords: cancer; mutation; gene; chromosomes; pathways.

Reference to this paper should be made as follows: Al-Shammari, M., Tobin, D.J. and Peng, Y. (xxxx) 'Systematic associations between germ-line mutations and human cancers', *Int. J. Computational Biology and Drug Design*, Vol. x, No. x, pp.xxx–xxx.

Biographical notes: Mohamad Al-Shammari received his PhD in Bioinformatics from the Faculty of Engineering and Informatics, University of Bradford, UK in 2013. Currently, he is lecturing in the Faculty of Engineering at University of Bradford. His research interests include analysis of gene involvement in human primary cancers from the starting position of the mutation class that harbours the specific gene mutation. He uses data mining and systems biology approaches for the analytics of biological big data.

Desmond J. Tobin is Professor of Cell Biology and Director of the Centre for Skin Sciences (CSS) at University of Bradford. He holds a BSc from the National University of Ireland (Maynooth), a PhD from the University of London (St. John's Institute of Dermatology) and post-doctoral training from New York University Medical School's Department of Dermatology. Over the past 20 years he has researched in basic and applied skin sciences, he is a Fellow of Royal College of Pathologists, and Society of Biology amongst others.

Dr. Yonghong Peng is a Principal Investigator in Big Data Science and Technology (BDST) and Bio-Medical Informatics (BMI) at the University of Bradford, United Kingdom. Dr Peng is the Chair for Big Data Task Force (BDTF) of IEEE Computational Intelligence Society (IEEE CIS), and a member of Data Mining and Big Data Analytics Technical Committee of IEEE CIS. He is also a founding member of Technical SubCommittee on Big Data (TSCBD) of IEEE Communications Society, and a member of Big Data Task Force of China Information Industry Association (CIIA). Dr Peng is currently acting as an associate editor for IEEE Transaction on Big Data, a member of editorial board of International Journal of Big Data Intelligence, and an academic editor of PeerJ and PeerJ Computer Science.

This paper is a revised and expanded version of a paper entitled **[**Genomic Mutation Big Data Enable The Discovery of The Influence of Different Genes in 26 Human Cancers**]** presented at [BIBM, Belfast, November 2-5, 2014].

1 Introduction

Big data provides innovative approaches for capturing, processing, searching, analysing, storing, transferring and visualising large and complex datasets, where datasets become very hard to process using traditional database tools. Big data has become a frontier for many academic researchers and has found many applications, including in the private and public sectors. Recently, a rapid increase in online biological data has provided the possibility for scientists to gain insights from data analytics. Online biological repositories have become the significant source for biological records, examples including:

- The Online Mendelian Inheritance in Man (OMIM) (Amberger et al., 2009), which began in 1960s in 12 editions published books, moved into online database in 1987, and then moved onto the world wide web (WWW) in 1995. By July 2014 the database contained 22,435 genetics records, with each record containing a hundred to a thousand records attached to it.

- The HUGO gene nomenclature committee (HGNC) database, which holds a total of 33,000 symbols, each of which consists a further hundred or thousand records.
- The human gene mutation database (HGMD) (Stenson et al., 2009), which provides an online human mutation database containing a total of 148,413 records collected between 1997 and 2012.
- The genetic association database (GAD) (Duarte et al., 2007) which is an online library of published genetic association studies holding more than 130,000 entries of human genetic association studies.
- The cancer gene census database (COSMIC) (Stratton et al., 2009) that holds a total of 4,970,019 cancers related records.
- DAVID bioinformatics database (DAVID) (Sherman et al., 2007) that provide database and tools for online functional annotation.
- The biological general repository for interaction datasets (BioGRID) (Stark et al., 2011), which provides an online database containing a total of 684,996 genetic interaction reports.

The availability of aforementioned, biological data has opened up enormous potential for physicians, genetic counsellors and biomedical researchers further study of disease associations, to vastly improve our knowledge and understanding of how diseases are caused, and to eventually help the development of appropriate treatments.

In this study we curated various mutation datasets from multiple online biological databases, in order to investigate the influence of germ-line cancer mutation classes on the associated human cancers, considering them as major potential factors human cancer. We investigated the association of germ-line mutation classes of 424 genes with 26 primary human cancers. We systematically constructed and analysed a Human Cancer Map (HCM) and a Genome Wide Distribution Map based on 8870 germ-line cancer mutation records in the context of their distribution in the associated pathways and the relevance of biological factors. We applied Choen's Kappa (*k*) Coefficient to quantify the degree of agreement between two or more connected cancers on the basis of the shared gene associations. The analytic results suggest that, of all mutation classes missense/nonsense and regulatory mutation classes are over-expressed in the human genome and so are likely to have a significant impact on human cancer etiology and pathomechanism. It is also seen that several Chromosomes (Chr17, Chr1, Chr15, Chr2 and Chr3) tended to contribute to cancer genes disproportionally compared with other chromosome, whereas Chr21 and Chr-y did not show any contribution to any of the cancer-associated genes. Chromosome 17 carried the highest number of cancer-related genes with high mutation (14% of total), with chromosome X carrying the fewest (1%). This offers new insights into how the distribution and interconnections of gene mutations influence the development of cancers.

2 Materials and methods

2.1 Data integration

Data integration combine data obtained from different sources, with ‘cleaning-up’ to provide a unified view for users. This process is a significant element of this study, as it involves complex procedures to combine scientific data and findings from different depositories (Halevy, 2001). This process results in a new ‘data warehouse’, procedures of extracting data from various sources, and then transforming it to construct a database suitable for a particular operation, and loading it to be in a useable format (ETL) for data mining and analytics (Chaudhuri and Dayal, 1997). We applied the ETL process to perform data extraction from three complementary data sources, i.e., genetic association database (GAD), Sanger database (COSMIC) and (HGMD), and then transform it into a suitable dataset layout through cleaning, reformatting, standardisation, aggregation of multiple datasets.

2.2 The genetic association database (GAD)

The GAD (Becker et al., 2004), is an online library archive of published genetic association studies. The GAD provides a comprehensive, public, web-based repository of molecular, clinical and study parameters for more than 130,000 entries of human genetic association studies. These includes 17 different disease classes, such as cancer, aging, cardiovascular, chem dependency, developmental, haematological, immune, infection, metabolic, mitochondrial, neurological, normal variation, pharmacogenomics, psychiatric, renal, reproduction, and vision. Each entry of the GAD is saved as an independent record and is composed of 22 fields (or attributes), including gene symbol, entrez GeneID, chromosomal location, associated tag between genes and disorders (‘Y’, N), DNA position, *P*-value, reference and its corresponding PubMed is and OMIM id, etc.

The downloaded file (on 30 November, 2011) contained a total of 21,444 disorder records, each one of these records was tagged as positive or negative associations. We only selected the records that have positive association with cancers record, which resulted in total a set of 1908 records. These 1908 records contained a total of 486 unique genes and 480 duplicated primary or subtype cancer names.

2.3 The cancer gene census database (COSMIC)

The COSMIC (Futreal et al., 2004) is a collection of somatic and germ-line mutations with information on the mutation class. The mutation information based on those mutations implicated in the development of particular cancers. The COSMIC database stores a total of 4,970,019 mutations entries. Each entry of COSMIC is saved as a unique independent record and is composed of sixteen fields (attributes) including; gene symbol, name of gene, geneID, chromosome, chromosome band, somatic mutation, germ-line mutation, tumour type (somatic mutation), tumour type (germ-line mutation), cancer syndrome, tissue type, cancer molecular genetics, mutation subclasses, translocation, other syndromes.

The COSMIC data was downloaded on 1st February, 2012. We extracted 76 genes from 27,829 cancer-related mutation records, based on their germ-line mutations.

2.4 The human gene mutation database (HGMD)

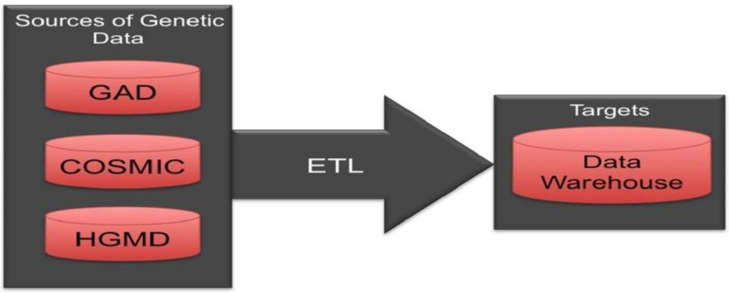
The HGMD (Stenson et al., 2009) is a large depository of data on human germ-line mutations with details of mutation classes. The mutation data includes point mutations of a single base pair with insertions and deletions, regulatory and splicing-relevant regions of, micro-deletions (indels), repeat variants, gross lesions (deletions, insertions and duplications) and complex rearrangements. The mutation data was stored as independent records and presented on a gene-wise basis. It also provides access to the mutation classes data via a hypertext link, including additional data sources (i.e., Genome Database (GDB), Online Mendelian inheritance in Man (OMIM), HUGO Gene Nomenclature Committee (HGNC), Entrez Gene, GeneCards, GeneAtlas, GeneClinics, UniGene, SwissProt and the Human Protein Reference Database from each gene page.

In order to extract data from the web-based application of HGMD, a gene symbol is needed. The gene symbols used here were based on merging the GAD and COSMIC gene data, then performing a filtering process to remove duplicated genes. In total 520 unique genes were collected which represent cancer-associated genes. The HGMD was then extracted based on the gene symbols from the HGMD web-based applications. The data included gene symbol, gene descriptions, chromosome location, cDNA sequence ID, mutation subclasses, total number of mutations for each gene, total number of specific class of mutation for each of the gene, related phenotypes, unique mutation ID, PubMed references, and some additional information about the gene. The search was further expanded by using cancer names obtained from the GAD to increase the chance of finding new cancer-related gene and mutation records. As a result a total of
15,264 records were collected, with a combination of cancers and some non-cancers disorders for each of the 10 different mutation classes. The data were then cross-checked to ensure there was no duplication of the data and none of the records were eliminated by the ETL process (see Figure 1). Table 1 summarises the total number records for each mutation class.

Table 1 Classes of mutations and the total number of collected records for cancers and non-cancer disorders

Classes of mutations	*Total number of mutation records*	*Non-cancer mutations*	*Cancers mutations*
Missense nonsense	7456	4371	3085
Small deletion	3280	1181	2099
Splicing	1386	295	1091
Gross deletions	1227	325	902
Small insertions	1190	374	816
Regulatory	280	91	189
Small indels	177	34	143
Gross insertions	150	19	131
Comples rearrangements	77	11	66
Repeat variations	42	16	25
Total	15,264	6717	8547

Figure 1 Extract, transform, and load (ETL). The left hand side three main biological databases used in this project. The ETL is the process to extract data, and prepare the data suitable for further analytics (see online version for colours)



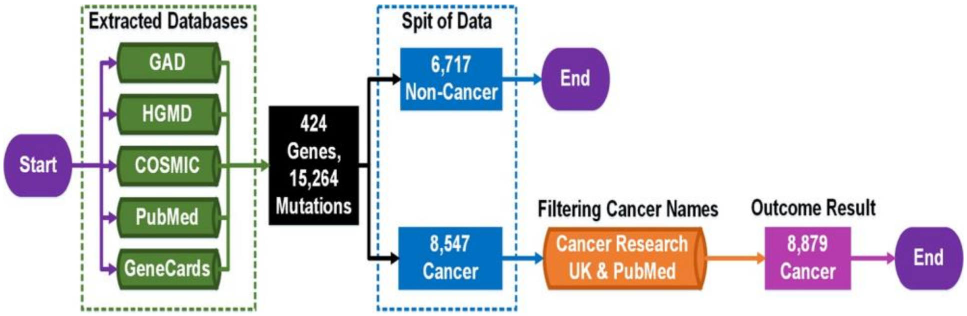
2.5 Data preparation

A total of 15,264-curated mutation records were extracted from the HGMD database and were found to be highly complex in its first iteration. A set of procedures was employed for the data pre-processing, including the elimination of non-cancer records, and merging the sub-types of cancers. Non-cancer records were eliminated from the list using published literature studies in the following databases (e.g., PubMed, OMIM and other online libraries) (Amberger et al., 2009). This process eliminated a total of 6717 non-cancer records from the original list (see Table 1).

Secondly, the naming of the cancers were re-processed to enable consistency. For example, 'Leukaemia' has two subclasses – 'Acute Leukaemia' and 'Chronic Leukaemia', and each one of these have a variety of subtypes. On the other hand, many of the cancers were duplicated because of the presence of synonyms in the list e.g., 'Bowel' cancer, 'Colorectal' cancer and 'Colon' cancer. These entries were merged into the primary cancer term e.g., 'Bowel cancer'. To complete this merging accurately a hierarchy list of cancer designations for each of the cancer types was created (i.e., primary, secondary, tertiary and quaternary subclass cancer names) to construct a family tree for the primary cancers.

The family tree of cancers was constructed based on terms used by Cancer Research UK (Grzybowska et al., 2002). The family tree was used to inform a re-arrangement of the data, which builds two maps, namely: a HCM and a genome-wide distribution map (GWD). After this data pre-processing a total of 8879 mutation records remain, which contained a total of 26 primary unique cancers and a set of 424 unique genes. Each of the 8879 mutation records was stored as an independent record (see Figure 2). The dataset was manipulated to construct a table containing the following entries: gene symbol, unique mutation ID, sub class of mutation, cancers disorder name, total number of mutations for each gene, total number of mutations for each of the 10 different classes of mutations, and the associated PubMed references.

Figure 2 Algorithmic framework for the extractions and preparation of gene mutations data. The start point at the top of the image shows the databases used for finding the cancer-associated genes, and the mutation records. The data is then split into cancer records and non-cancer related records. This is followed by filtering and merging the data based on primary cancer names (see online version for colours)



2.6 *Cohen's Kappa (k) coefficient*

To evaluate the association between two cancers, we performed a statistical analysis on the agreement of each cancer pair based on the shared genes and mutation class. The Cohen's kappa coefficient (Carletta, 1996; Cohen, 1960) is used to quantify the agreement between two individuals (*a* and *b*):

$$k(a,b) = \frac{A(a,b) - P(a,b)}{1 - P(a,b)}$$

where $A(a, b)$ is the relative observed agreement between *a* and *b*, and $P(a, b)$ is the hypothetical probability of chance agreement, using the observed data to calculate the probabilities of each observer randomly saying each category. If two individuals are in complete agreement then the Cohen's kappa coefficient *k* is equal to 1, while if there is no agreement between the two individuals then $k = 0$.

For example Table 2 is the observation of five genes in two cancers *a* and *b*. Table 3 show the agreements and disagreement between the observation on cancer *a* and *b*. The Cohen's kappa coefficient is calculated involving the following three steps:

- Calculate the observed percentage agreement between *a* and *b*,

$$A(a,b) = \frac{C(1,1) + C(0,0)}{C(*,*)} = \frac{3+1}{5} = 0.8. \quad (1)$$

- Calculate the probability of random agreement between *a* and *b*

$$P(a,b) = \frac{C(*,1)C(1.*) + C(*,0)C(0,*)}{C(*,*)C(*,*)} = \frac{4\times3 + 1\times2}{5\times5} = 0.56. \quad (2)$$

- Calculate the degree of *k*,

$$k(a,b) = \frac{A(a,b) - P(a,b)}{1 - P(a,b)} = \frac{0.8 - 0.56}{1 - 0.56} = 0.55 \quad (3)$$

Table 2 The observation of five genes in two cancers (*a* and *b*). It is assigned to be 1 if a gene is related to a cancer, otherwise 0 is assigned

	Gene 1	*Gene 2*	*Gene 3*	*Gene 4*	*Gene 5*
Cancer *a*	1	1	1	0	0
Cancer *b*	1	1	1	1	0

where $A(a, b)$ represent the observed percentage agreement for each connected cancers in the network and $P(a, b)$ the probability of random agreement for each two interconnected cancers in the network. $K(a, b)$ is the Cohen's Kappa coefficient.

In this study, we classified the degree of association between two cancers into four categories: 'very highly connected', 'highly connected', and 'moderately connected' and 'poorly connected'. Two cancers are defined as being very highly connected if the Cohen's kappa coefficient (k) is above 0.75, being highly connected if k is between 0.5 and 0.75, being Moderately connected if k is between 0.25 and 0.5, while being poor connected if k is less than 0.25.

Table 3 The number of agreements and disagreements for the observations on cancer *a* and *b*

		Cancer b		
		1	*0*	*Row total*
Cancer a	1	C(1, 1) = 3	C(1, 0) = 0	C(1,*) = 3
	0	C(0, 1) = 1	C(0, 0) = 1	C(0, *) = 2
Column total		C(*, 1) = 4	C(*, 0) = 1	C(*, *) = 5

3 Results and discussion

3.1 A human cancer map (HCM) based on cancer-linked genes and their associated mutation classes

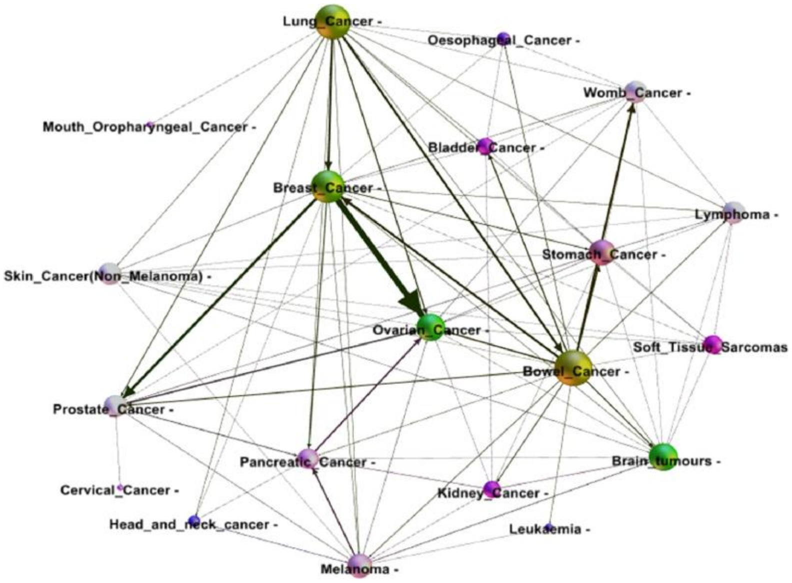
The HCM map showing the associations between cancers was interrogated by using multiple packages in R, including Reshape, Match, Plyr, as well as in-house developed software to process the 8879 mutation records (Figure 2). Two cancer disorders are connected only if they involve the same mutation class affecting the same gene. The HCM displayed connections between nodes of primary cancers, with each connection representing an implicated gene(s) and its contributing mutation class. Of the 26 primary cancers involved in this study, 20 cancers showed at least one link to other cancers in the map. 69 unique genes (16.2% of the total in the gene set) underpinned these inter-cancer links. This observation suggests that both gene and associated mutation class plays a central role in associating cancer nodes (Li et al., 1997).

Next, the degree of connectivity of the primary cancer nodes distributed in the map was assessed. It is shown that 7 primary cancers, representing one particular hub in the map, were found to be connected to a large number of other cancers as indicated by their high connectivity (c value). For example, Bowel cancer (c = 18), Lung cancer (c = 17), Breast cancer (c = 16), Brain tumours (c = 13), Ovarian cancer (c = 13), Stomach cancers (c = 12), and Melanoma (c = 11). This finding indicates that each of these primary cancers is associated with several other primary cancers via the same gene and mutation classes (Figure 3).

3.2 Mapping genes involved in the human cancer map to their genome-wide chromosomal position (GWD map)

The next step of this study was to investigate the distribution of the mutations of the 69 identified genes involved in 20 primary cancers on their respective chromosome (Figure 2). 4964 mutation records are involved in these 69 genes. Based on the 4964 mutation records the association of genes was investigated. Two or more genes were associated if they are both involved with the same cancer via the same mutation class. For example the genes *XPC*, *ERCC6* are considered to be associated with each other because both tended to be causative genes for Bladder cancer at the same Missense/Nonsense mutation (García-Closas et al., 2006). The gene *MLH1* was excluded from such an association with Bladder cancer because it is associated by a different type of mutation class (i.e., small deletions). Conversely, Bowel cancer is associated three genes (*XPC*, *ERCC6* and *MLH1*) based on the same missense/nonsense mutation. As a result of this mapping process, a total of 1158 associations/connections between 69 cancer genes were detected. These connections were further explored in CIRCOS (Krzywinski et al., 2009) to construct a genome-wide distribution map of human cancers. The 1158 associations between cancer genes and their positions in the human genome are present in Figure 4.

Figure 3 Human cancer map (HCM). Each node corresponds to a primary cancer and a link between two nodes represents the presence of shared cancer genes and mutation class. The size of each node is proportional to the number of interconnections connected to that node i.e., number of distinct primary cancers connecting to it. The width of the link-lines reflects the number of genes linking each cancer nodes (see online version for colours)



Thereafter the distributions of the 69 associated genes in each chromosome were calculated (Figure 5) to show the number of implicated genes on each chromosome. It is shown that Chr17, Chr1, Chr15, Chr2 and Chr3 tended to contribute to cancer genes disproportionally compared with other chromosome, whereas Chr21 and Chr-y did not show any contribution to any of the cancer-associated genes. Chromosome 17 carried the highest number of cancer-related genes (14% of total), with chromosome X carrying the fewest (1%). The Missense/Nonsense Mutation type was the most common of all mutation classes, and was detected in 22 chromosomes. Moreover, a mutation of the Regulatory Mutation type was detected in 11 of 22 chromosomes. Together these mutation types were most commonly associated with cancer-pathway genes (Kanehisa et al., 2006).

Figure 4 Genome-wide mutations, where the blocks in the outer circle indicate the particular chromosomes and the position of 69 genes on each chromosome. The inner interconnections represent the associations between two or more genes connected if they contribute to same cancer via the same mutation class. The colours of the interconnections represent the corresponding chromosome (see online version for colours)

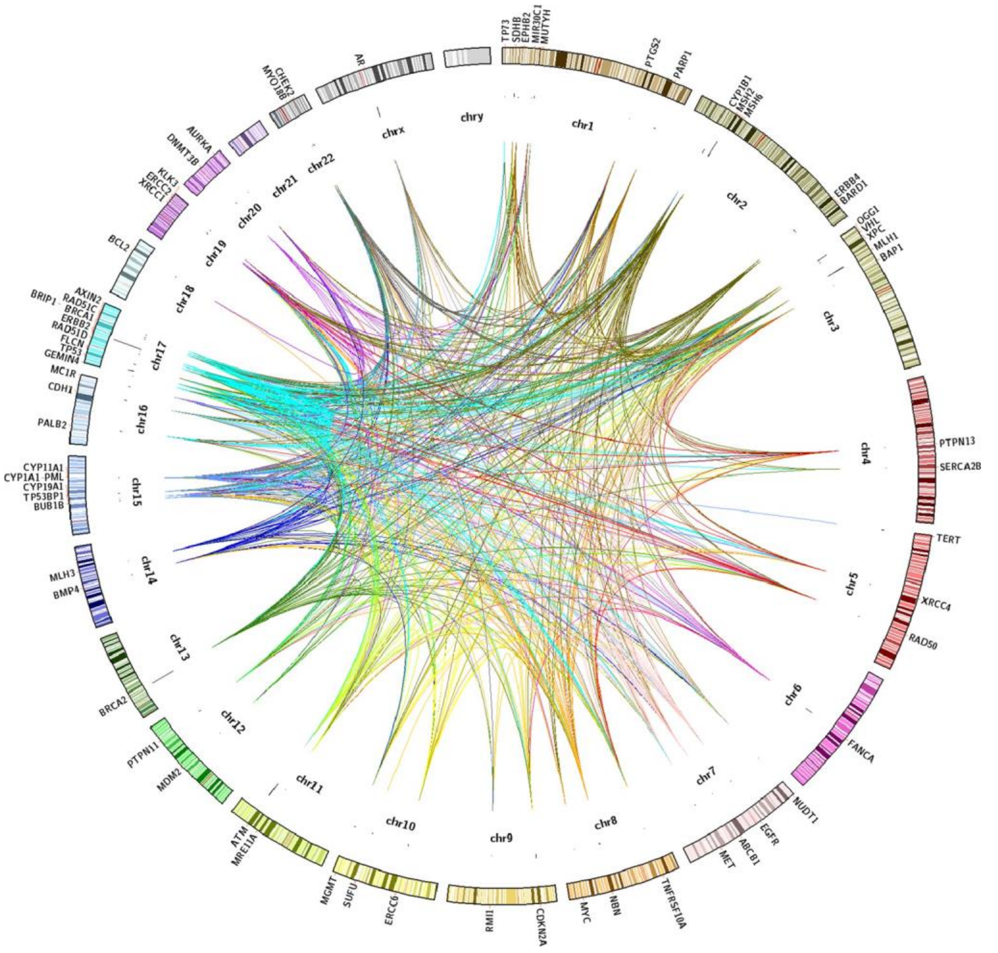
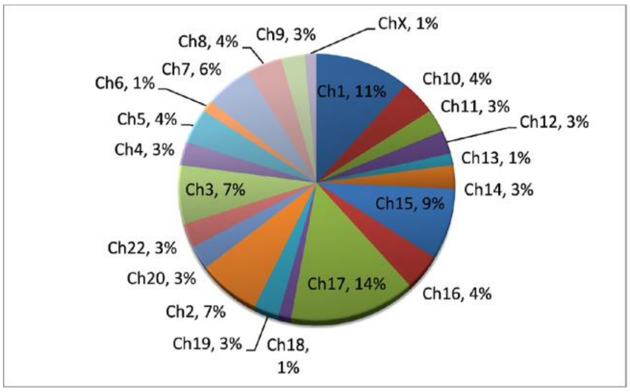


Figure 5 The genome-wide distributions for cancer genes showing the distribution of the 69 genes over human chromosomes (see online version for colours)

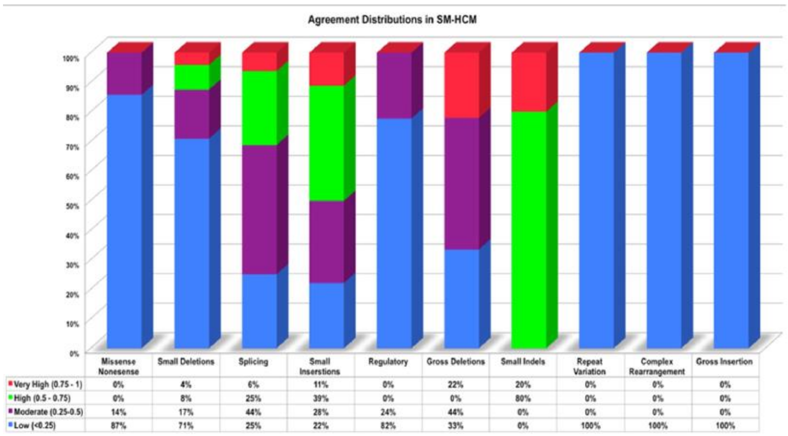


3.3 *Extent of agreement between primary cancer nodes on the basis of shared gene and mutation class in HCM*

The method of Choen's Kappa was employed to quantify the agreement between two primary cancer nodes in the HCM. The agreement revealed a number of interesting relationships between cancer nodes across the HCM (Figure 6). These include:

- missense/nonsense mutations (44% of the HCM) and Regulatory mutation (10% of the HCM) showed only low to moderate agreement between their interconnected nodes
- small deletions, splicing, small insertions, gross deletions and small indels (3–14% of the HCM) exhibited low to very high agreements
- repeat variation, complex rearrangement and cross insertion mutations (1–2% of the HCM) showed very low agreements between their interconnected cancer nodes.

Figure 6 Agreement distributions in SM-HCM. Blue shows low agreements (<0.25), purple represents moderate agreements ((0.25–0.50), green reflects all high agreement (0.50–0.75), while red refers to the very high agreement (0.75–1). The mutations subclasses names are placed under each bar (see online version for colours)



	Missense Nonesense	Small Deletions	Splicing	Small Inserstions	Regulatory	Gross Deletions	Small Indels	Repeat Variation	Complex Rearrangement	Gross Insertion
🟥 Very High (0.75 - 1)	0%	4%	6%	11%	0%	22%	20%	0%	0%	0%
🟩 High (0.5 - 0.75)	0%	8%	25%	39%	0%	0%	80%	0%	0%	0%
🟪 Moderate (0.25-0.5)	14%	17%	44%	28%	24%	44%	0%	0%	0%	0%
🟦 Low (<0.25)	87%	71%	25%	22%	82%	33%	0%	100%	100%	100%

While Missense/nonsense mutations dominate the network, in terms of type of interconnections, if removed could cause the network to collapse, their interconnection strength (*k* score) is low to moderate. By contract therefore, interconnections derived from Small deletions, splicing, small insertions, gross deletions and small indels can influence the network disproportionately due to their higher *k*-scores.

Moreover, two interconnected cancer nodes (breast and ovarian cancers) had high *k*-scores for gene mutations of the splicing, small insertions and small indels type, when examining *BRCA1, BRCA2, BRIP1, RAD51C, RAD51D, TP53*. This finding, agreed with several studies in the literature (Welcsh and King, 2001, Grzybowska et al., 2002), show that germline mutation of *BRCA2* and *BRCA1* predispose to breast and ovarian cancer. Similarly, we found that brain tumours tend to be highly linked to melanoma, womb and bowel cancer in terms of the mutation of *MLH1, CDKN2A* and *BRAC2,* including splicing, small insertion, gross deletion and small indels mutations. We also found that melanoma and pancreatic cancer nodes were highly connected via their similar splicing and small insertion mutations for *CDKN2A, BRAC2,* which is in agreement with previous reports (Whelan et al., 1995, Goldstein, 2004, de Snoo et al., 2008). Melanoma also exhibited very high agreement with the head & neck cancer by virtue of small indels mutations in the *CDKN2A* gene, which agreed, interestingly, with a finding reported by Cabanillas et al. (2011). Therefore, this suggests that *CDKN2A* acts as a key gene in human cancer that its mutations are found in many distinct primary cancers (Cabanillas et al., 2013; Whelan et al., 1995). A similar case may be made for the *BRAC2* gene, where shared small insertion mutation mutations are found in prostate, melanoma, pancreatic cancer and lung cancer (Vasen et al., 2000; Lynch et al., 2002).

4 Conclusion

From a biological perspective, our data suggests that Missense/Nonsense and Regulatory mutation has a disproportionately large impact on associating cancer nodes with each other in a HCM and GWD-MAP. Without the involvement of these two mutation classes the HCM would fragment, and almost 50% of cancer nodes would be disconnected. This then may suggest that mutations of these 2 classes may contain driver mutations in the associated cancers. An alternative but not mutually exclusive way of interpreting our finding is the significant role of Small Deletions, Splicing, Small Insertions, Gross Deletions, and Small Indels) in developing cancers, given their high to very high kappa values which are indicative of their involvement in how strongly interconnected the cancer nodes are. Thus, mutations of these types may be particularly deleterious as they would have greater impact on cancer gene connectivity in the HCM and GWD-MAP.

Acknowledgements

We thank HGMD, Online Mendelian Inheritance in Man (OMIM), genetic association database (GAD) and Sanger database. This research was presented based on the biological data provided by their online sources.

References

Amberger, J., Bocchini, C.A., Scott, A.F. and Hamosh, A. (2009) 'McKusick's online Mendelian inheritance in man (OMIM®)', *Nucleic Acids Research*, Vol. 37, p.D793.

Becker, K.G., Barnes, K.C., Bright, T.J. and Wang, S.A. (2004) 'The genetic association database', *Nature Genetics*, Vol. 36, pp.431–432.

Cabanillas, R., Astudillo, A., Valle, M., De La Rosa, J., Álvarez, R., Durán, N.S. and Cadiñanos, J. (2011) 'Novel germline CDKN2A mutation associated with head and neck squamous cell carcinomas and melanomas', *Head & Neck*, Vol. 35, pp.80–84.

Cabanillas, R., Astudillo, A., Valle, M., De La Rosa, J., Álvarez, R., Durán, N.S. and Cadiñanos, J. (2013) 'Novel germline CDKN2A mutation associated with head and neck squamous cell carcinomas and melanomas', *Head & Neck*, Vol. 35, pp.E80–E84.

Carletta, J. (1996) 'Assessing agreement on classification tasks: the kappa statistic', *Computational Linguistics*, Vol. 22, pp.249–254.

Chaudhuri, S. and Dayal, U. (1997) 'An overview of data warehousing and OLAP technology', *ACM Sigmod Record*, Vol. 26, pp.65–74.

Cohen, J. (1960) 'A coefficient of agreement for nominal scales', *Educational and Psychological Measurement*, Vol. 20, pp.37–46.

de Snoo, F.A., Bishop, D.T., Bergman, W., Van Leeuwen, I., Van der Drift, C., Van Nieuwpoort, F.A., Out-Luiting, C.J., Vasen, H.F., Ter Huurne, J.A. and Frants, R.R. (2008) 'Increased risk of cancer other than melanoma in CDKN2A founder mutation (p16-Leiden)-positive melanoma families', *Clinical Cancer Research*, Vol. 14, pp.7151–7157.

Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R. and Palsson, B.Ø. (2007) 'Global reconstruction of the human metabolic network based on genomic and bibliomic data', *Proceedings of the National Academy of Sciences*, Vol. 104, pp.1777–1782.

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', *Nature Reviews Cancer*, Vol. 4, pp.177–183.

García-Closas, M., Malats, N., Real, F.X., Welch, R., Kogevinas, M., Chatterjee, N., Pfeiffer, R., Silverman, D., Dosemeci, M. and Tardón, A. (2006) 'Genetic variation in the nucleotide excision repair pathway and bladder cancer risk', *Cancer Epidemiology Biomarkers & Prevention*, Vol. 15, pp.536–542.

Goldstein, A.M. (2004) 'Familial melanoma, pancreatic cancer and germline CDKN2A mutations', *Human Mutation*, Vol. 23, pp.630–630.

Grzybowska, E., Siemiñska, M., Zientek, H., Kalinowska, E., Michalska, J., Utracka-hutka, B., Rogoziñska-Szczepka, J. and Kazmierczak-Maciejewska, M. (2002) 'Germline mutations in the BRCA1 gene predisposing to breast and ovarian cancers in Upper Silesia population', *Acta Biochimica Polonica-English Edition*, Vol. 49, pp.351–356.

Halevy, A.Y. (2001) 'Answering queries using views: a survey', *The VLDB Journal—The International Journal on Very Large Data Bases*, Vol. 10, pp.270–294.

Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. (2006) 'From genomics to chemical genomics: new developments in KEGG', *Nucleic Acids Research*, Vol. 34, pp.D354–D357.

Krzywinski, M., Schein, J., Birol, İ., Connors, J., Gascoyne, R., Horsman, D., Jones, S.J. and Marra, M.A. (2009) 'Circos: an information aesthetic for comparative genomics', *Genome Research*, Vol. 19, pp.1639–1645.

Li, J., Yen, C., Liaw, D., Podsypanina, K., Bose, S., Wang, S.I., Puc, J., Miliaresis, C., Rodgers, L. and Mccombie, R. (1997) 'PTEN, a putative protein tyrosine phosphatase gene mutated in human brain, breast, and prostate cancer', *Science*, Vol. 275, pp.1943–1947.

Lynch, H.T., Brand, R.E., Hogg, D., Deters, C.A., Fusaro, R.M., Lynch, J.F., Liu, L., Knezetic, J., Lassam, N.J. and Goggins, M. (2002) 'Phenotypic variation in eight extended CDKN2A germline mutation familial atypical multiple mole melanoma–pancreatic carcinoma–prone families', *Cancer*, Vol. 94, pp.84–96.

Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C. and Lempicki, R.A. (2007) 'DAVID bioinformatics resources: expanded annotation database and novel algorithms to better extract biology from large gene lists', *Nucleic Acids Research*, Vol. 35, pp.W169–W175.

Stark, C., Breitkreutz, B-J., Chatr-Aryamontri, A., Boucher, L., Oughtred, R., Livstone, M.S., Nixon, J., Van Auken, K., Wang, X. and Shi, X. (2011) 'The BioGRID interaction database: 2011 update', *Nucleic Acids Research*, Vol. 39, pp.D698–D704.

Stenson, P.D., Mort, M., Ball, E.V., Howells, K., Phillips, A.D., Thomas, N. and Cooper, D.N. (2009) 'The human gene mutation database: 2008 update', *Genome Med.*, Vol. 1, p.13.

Stratton, M.R., Campbell, P.J. and Futreal, P.A. (2009) 'The cancer genome'. *Nature*, Vol. 458, pp.719–724.

Vasen, H., Gruis, N., Frants, R., Van der Velden, P., Hille, E. and Bergman, W. (2000) 'Risk of developing pancreatic cancer in families with familial atypical multiple mole melanoma associated with a specific 19 deletion of p16 (p16-Leiden)', *International Journal of Cancer*, Vol. 87, pp.809–811.

Welcsh, P.L. and King, M-C. (2001) 'BRCA1 and BRCA2 and the genetics of breast and ovarian cancer', *Human Molecular Genetics*, Vol. 10, pp.705–713.

Whelan, A.J., Bartsch, D. and Goodfellow, P.J. (1995) 'A familial syndrome of pancreatic cancer and melanoma with a mutation in the CDKN2 tumor-suppressor gene', *New England Journal of Medicine*, Vol. 333, pp.975–977.