IDENTIFY CANCER DRIVER GENES THROUGH SHARED MENDELIAN DISEASE PATHOGENIC VARIANTS AND CANCER SOMATIC MUTATIONS

MENG MA[1], CHANGCHANG WANG[2], BENJAMIN S. GLICKSBERG[1], ERIC E. SCHADT[1], SHUYU D. LI[1]*, RONG CHEN[1]*

[1]*Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, 1 Gustave L. Levy Pl. New York City, NY 10029, USA*

[2]*School of Computer Science, Anhui University, Anhui, P.R. China*
**Email: rong.chen@mssm.edu; shuyudan.li@mssm.edu.*

Genomic sequencing studies in the past several years have yielded a large number of cancer somatic mutations. There remains a major challenge in delineating a small fraction of somatic mutations that are oncogenic drivers from a background of predominantly passenger mutations. Although computational tools have been developed to predict the functional impact of mutations, their utility is limited. In this study, we applied an alternative approach to identify potentially novel cancer drivers as those somatic mutations that overlap with known pathogenic mutations in Mendelian diseases. We hypothesize that those shared mutations are more likely to be cancer drivers because they have the established molecular mechanisms to impact protein functions. We first show that the overlap between somatic mutations in COSMIC and pathogenic genetic variants in HGMD is associated with high mutation frequency in cancers and is enriched for known cancer genes. We then attempted to identify putative tumor suppressors based on the number of distinct HGMD/COSMIC overlapping mutations in a given gene, and our results suggest that ion channels, collagens and Marfan syndrome associated genes may represent new classes of tumor suppressors. To elucidate potentially novel oncogenes, we identified those HGMD/COSMIC overlapping mutations that are not only highly recurrent but also mutually exclusive from previously characterized oncogenic mutations in each specific cancer type. Taken together, our study represents a novel approach to discover new cancer genes from the vast amount of cancer genome sequencing data.

1. Introduction

Significant efforts in the past several years in cancer genomic sequencing by individual investigators and large consortium such as The Cancer Genome Atlas (TCGA) and The International Cancer Genome Consortium (ICGC) have uncovered a large number of novel oncogenic drivers. These studies not only advanced our understanding on the genetic basis of tumorigenesis and cancer progression, but also significantly enabled the development of personalized cancer therapeutics [1, 2]. Cancer genome or exome sequencing data have been generated from approximately 25,000 tumor samples covering more than 50 tumor types [3, 4], representing a comprehensive cancer genomic atlas. While data generation has been greatly facilitated by rapid technology development, interpretation of cancer sequence information still remains a major challenge. As most solid tumors harbor a median of 40-80 non-synonymous somatic mutations per tumor, only three to six of them are driver mutations [5]. The most commonly used approach to distinguish a small number of driver mutations from those background passenger mutations is to identify significantly mutated genes in a cohort study [6]. The underlying rationale is if a gene is mutated at significantly greater rate than the background mutation rate, it is more likely to be oncogenic, as the mutations conferring tumor growth advantage are evolutionarily selected during cancer development. To complement this approach, various computational tools have been developed to assess the effects of missense mutations on protein functions [7]. While such an approach has further characterized numerous novel cancer drivers and oncogenic pathways from cancer genomic sequencing data, it requires a large number of samples to uncover those drivers mutated at low population frequency in a given tumor type. This is particularly problematic for those cancers with high background mutation rates such as

melanomas and lung cancers. For example, it has been estimated that it would require approximately 4,000 melanoma patient samples to detect cancer genes mutated at 2% frequency, and more than 20,000 samples for genes mutated at 1% with 90% power for 90% of genes [8].

Many human genetic diseases are Mendelian disorders caused by one or more aberrations in the genome. These diseases are often heritable as the disease causing, pathogenic variants are passed on from parents' genome. To date, approximately 180,000 genetic variants in more than 7,000 genes have been identified as pathogenic for more than 4,000 Mendelian diseases [9]. Some of the first established cancer genes with frequent somatic mutations were originally identified from their associations with familial cancer syndromes. The first tumor suppressor RB1 was discovered by studying the familial form of retinoblastoma [10]. The most frequently mutated gene in cancers, p53, was also identified as a tumor suppressor inactivated in Li–Fraumeni syndrome, a rare cancer predisposition hereditary disorder. Other well-known cancer genes harboring high frequency somatic mutations and that are associated with Mendelian diseases include VHL in Von Hippel-Lindau syndrome, MLH1, MSH2, MSH6 in Lynch syndrome, TSC1, TSC2 in Tuberous sclerosis, and ATM in ataxia-telangiectasia [11]. Notably, a recent study has revealed potentially novel cancer-associated genes through analysis of comorbidity between cancers and Mendelian diseases [12].

By definition, germline pathogenic variants impact the functions of key proteins involved in the developmental process and consequently cause heritable diseases. If the same germline pathogenic variants occur as somatic mutations in cancers, these mutations would also alter protein functions and may play a role in tumor initiation and progression, even though the same proteins can have very different functions during development than in adult tissues. Indeed in a recent report, several genes sharing identical mutations in Mendelian diseases and cancers were proposed as novel cancer genes [13]. Based on this underlying hypothesis, we carried out a systematic comparative analysis of the reported pathogenic variants in Mendelian diseases and cancer somatic mutations. There are several repositories for pathogenic variants. A comparison of four of the most comprehensive databases showed that HGMD is currently the largest collection of human disease variants, although each database has its own advantages in terms of the information collected as well as database infrastructure [9]. For cancer somatic mutations, COSMIC is recognized as the most comprehensive resource for somatic mutations in human cancers [14], with more than 1.4 million confirmed somatic mutations identified from 1.1 million tumor samples including genome-wide sequencing data from more than 20,000 tumors. In this study, we first identified overlapping mutations between pathogenic variants in HGMD [15] and cancer somatic mutations from the COSMIC database [14]. Further characterization of these mutations show that the mutation-harboring genes are significantly enriched for known cancer genes, supporting the above described hypothesis. We then examined those genes harboring the shared pathogenic variants and somatic mutations in cancers by applying additional filters such as the number of overlapping HGMD/COSMIC mutations in a given gene or the frequency of overlapping mutations in each tumor type. Moreover, those overlapping mutations with high recurrence in cancers were subjected to mutual exclusivity analysis with known oncogenes in each tumor type in order to identify novel oncogenic drivers. Taken together, our study represents a

novel approach to discover new cancer genes from the vast amount of cancer genome sequencing data.

2. Methods

COSMIC V73 was downloaded from sftp-cancer.sanger.ac.uk using GUI client WinSCP under protocol sftp and port 22. HGMD Professional can be accessed from https://www.qiagenbioinformatics.com/products/human-gene-mutation-database/ with an authorized license. 1000 Genome Phase3 was downloaded from ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20130502/. ExAC database was downloaded from ftp://ftp.broadinstitute.org/pub/ExAC_release/. All RefSeq Exons were downloaded from UCSC table refGene through UCSC Table Browser (clade: Mammal, genome: Human, assembly: Feb.2009 (GRCH37/hg19), group: Genes and Gene Predictions, track: RefSeq Genes, Table: refGene). Cancer Gene Census dataset was downloaded from http://cancer.sanger.ac.uk/census/.

All the analyses were performed using shell scripts, mysql scripts and R scripts. The mutual exclusivity heat map was generated using gitools (http://www.gitools.org/). The survival analysis was done through cBioPortal (http://www.cbioportal.org/). Several major scripts for database query and statistical analyses are available on github (https://github.com/CosmicHGMD/CancerMendelian).

3. Results

3.1. *Identification of overlapping pathogenic variants in HGMD and somatic mutations in COSMIC*

HGMD includes six classes of variants, and we only included disease-causing mutations (DM and DM?) in our analysis. The DM class variants have been demonstrated in literature to confer the associated clinical phenotype of the affected individuals. The DM? class variants have some degree of uncertainty, but nevertheless have strong evidence supporting their pathogenicity. At the time of this writing, there are a total of 153,593 DM/DM? class variants in HGMD database. 11,523 of these variants are present in the COSMIC database, representing 0.54% of the total mutations in COSMIC (Table 1). When we only include the confirmed somatic mutations in COSMIC, there are 8,582 mutations (0.6%) that overlap with HGMD DM/DM? variants. As the majority of the somatic mutation data in COSMIC are from cancer genomic sequencing studies, some of these mutations are likely false positives, particularly those from early whole genome/exome sequencing when computational methods for calling somatic mutation were less reliable or if the identified somatic mutations were not validated by a different sequencing platform. Therefore, we further restrict COSMIC data to include only those somatic mutations occurred in more than one tumor samples. Although the total number of overlapping mutations with HGMD DM/DM? variants is reduced to 3,470, using this limited but more reliable somatic mutation list, the percentage with respect to the total number of these recurrent mutations (215, 436) in COSMIC increases to 1.6% (Table 1), suggesting Mendelian disease pathogenic variants are over-represented in recurrent somatic mutations in cancers.

Then we randomly selected the same number of genetic variants (153,593) from 1000 genome (exonic region) or the ExAC database as control variant datasets, and performed the same analysis. The analysis of randomly selected, mostly non-pathogenic common genetic variants was repeated 1000 times, and the results indicated that percentages of common non-pathogenic variants overlapping with COSMIC mutations are lower than the HGMD pathogenic variants (Table 1). The statistical significance was assessed based on the distribution of results from 1000 simulations. This finding supports our initial hypothesis that overlapping pathogenic variants in HGMD with cancer somatic mutations could enable identification of novel cancer genes.

Table 1. Enrichment of HGMD pathogenic variants in cancer somatic mutations.

Variant dataset (total number of variants)	**Randomly selected variants**	**Overlap with COSMIC mutations (percentage)**		
		All mutations in COSMIC (2,132,117)	**Somatic mutations in COSMIC (1,425,978)**	**Recurrent somatic mutations in COSMIC (215,436)**
HGMD DM/DM? (153,593)	-	11,523 (0.54%)	8,582 (0.60%)	3,470 (1.6%)
1000 Genome exonic region (2,156,973)	153,593	8,092 (0.38%); p<0.001	5,983 (0.42%); p<0.001	1,975 (0.92%); p<0.001
ExAc (10,450,722)	153,593	6,919 (0.32%); p<0.001	4,841 (0.34%); p<0.001	1,325 (0.62%); p<0.001

Next, we tested if HGMD/COSMIC overlapping mutations are more likely to occur at high frequency in cancers than those somatic mutations non-overlapping with HGMD. We first divided the confirmed COSMIC somatic mutations into two groups. The first group includes those mutations overlapped with HGMD DM/DM? variants and the second group includes the rest of somatic mutations that are only present in the COSMIC database. Then, for a given recurrence frequency cutoff c, we computed the percentage of somatic mutations with recurrence frequency (f) greater than c in group 1 (denoted as $\%G1_{f>c}$) and those in group 2 (denoted as $\%G2_{f>c}$). This is followed by computing the ratio of $\%G1_{f>c}$ over $\%G2_{f>c}$ at various mutation frequencies. As illustrated in Figure 1A, as the recurrence frequency (x-axis) increases, this ratio (y-axis) also increases. For example, the ratio is approximately 25 for recurrence frequency 20, indicating that COSMIC mutations overlapping with HGMD pathogenic variants are 25 fold more likely to occur in more than 20 tumor samples than those not overlapping with HGMD variants. We also directly plotted $\%G1_{f>c}$ and $\%G2_{f>c}$ (Figure 1B), and it clearly shows the HGMD/COSMIC overlapping mutations have higher mutation frequencies than those mutations only in the COSMIC database with mean recurrence in 8.0 and 1.3 tumors respectively (p = 1.5E-5, one-sided t-test). Because the likelihood that a somatic mutation is a cancer driver increases with its mutation frequency in cancers, this result is consistent with the hypothesis that cancer mutations overlapping with germline disease pathogenic variants in HGMD are more likely to be oncogenic. We further examined the presence of known cancer genes in the two groups using cancer gene census annotation [16]. While only 4.3% of the COSMIC somatic mutations do not overlap with HGMD

are in the cancer gene census list, there are approximately 10% of the somatic mutations overlapping with HGMD occur in cancer census genes.

To determine if the combination of somatic mutation frequency and the presence of overlap with HGMD pathogenic mutations would facilitate cancer gene discovery, we computed the percentage of somatic mutations mapped to cancer census genes in all COSMIC confirmed somatic mutations or only in those overlapped with HGMD DM/DM? variants. This procedure was then repeated for mutations with increasing frequencies (Figure 2). Two observations were notable from the results. First, a somatic mutation is more likely to be in a cancer gene as its frequency increases, evidenced by increasing percentage of cancer census genes. Second, the probability that the mutation-harboring genes are cancer-related increases if there are overlapping with HGMD variants (Figure 2, red bars vs. blue bars).

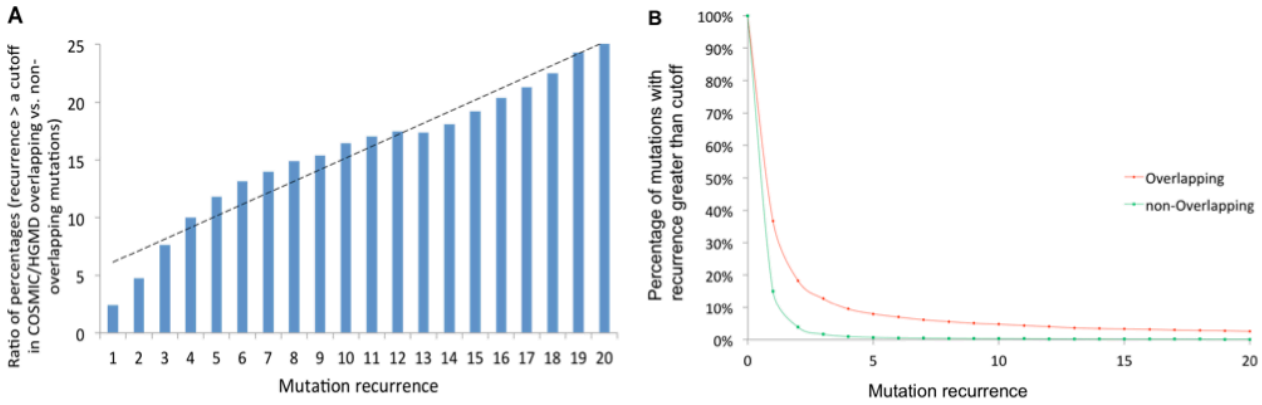

Figure 1. Overlap of HGMD variants with cancer somatic mutations is correlated with high mutation recurrence in cancers.

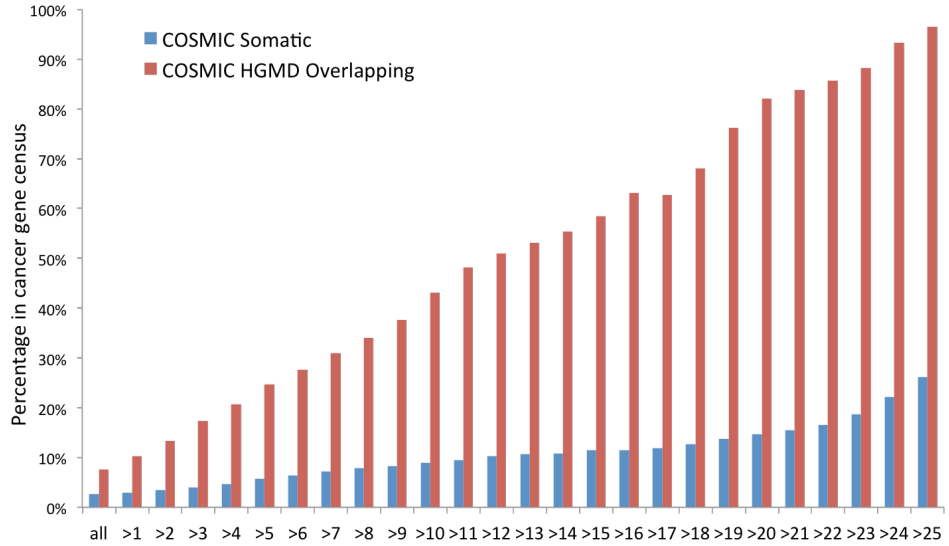

Figure 2. Novel cancer gene discovery through overlapping with HGMD and high mutation recurrence. X-axis represents mutation recurrence in COSMIC.

3.2. *Identification of potential tumor suppressors*

We examined whether the number of distinct overlapping HGMD/COSMIC mutations in a given gene is associated with the probability that the gene is a cancer gene. Figure 3 shows that as the number of distinct overlapping HGMD/COSMIC mutations in a given gene increases, the percentage of genes that belong to cancer gene census increases as well. Of those genes with more than six distinct overlapping mutations, approximately 25% are present in cancer gene census. It is generally recognized that while oncogenes are often mutated recurrently at certain positions (referred as hotspots), tumor suppressors tend to lack such mutational hotspots and are mutated at many positions across the gene sequences. Therefore, we reason that identifying genes with high number of distinct overlapping HGMD/COSMIC mutations would allow us to discover potentially novel tumor suppressors.

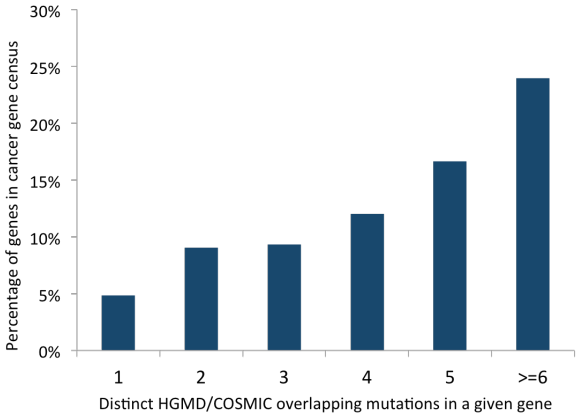

Figure 3. Identification of novel tumor suppressors based on the number of distinct HGMD/COSMIC overlapping mutations.

Accordingly, we ranked all the genes in COSMIC based on the number of distinct somatic mutations that overlap with HGMD DM/DM? variants and provided both cancer gene census annotations as well as oncogene/tumor suppressor classifications according to Vogelstein et al. [5] in Table 2. Almost half (23/48) of the genes with at least 20 overlapping HGMD/COSMIC mutations are in the cancer gene census list and/or annotated as an oncogene or a tumor suppressor, furthering the notion that HGMD pathogenic variant annotation may help distinguish driver oncogenic mutations from the passenger mutations in tumors. As expected, most of those genes with oncogene/tumor suppressor annotations are classified as tumor suppressors (19/21, 90%; Table 2). A literature search has provided support that some of the remaining genes are likely novel tumor suppressors. There are several genes that encode ion channels with many HGMD/COSMIC overlapping mutations, including SCN5A (67 overlapping mutations), SCN1A (52), CFTR (48), RYR1 (36) and RYR2 (30) (Table 2). While ion channels have not been recognized as a major class of cancer related genes, emerging evidence suggest at least some ion channels are involved in promoting malignancy. For example, CFTR, the cystic fibrosis (CF) gene, has been postulated to be a tumor suppressor because loss of CFTR enhanced tumor cell proliferation and epithelial-to-mesenchymal transition, and is associated with poor prognosis in several cancer types [17, 18, 19]. We also observed multiple collagen family genes with significant overlap between HGMD and COSMIC mutations, such as COL3A1 (29 overlapping mutations), COL7A1 (22) (Table 2), COL1A2 (19), COL4A5 (18), COL2A1 (14), COL6A3 (13), COL1A1 (13), and COL4A4 (11) (data not shown). Although collagens are considered as a barrier to suppress angiogenesis since they are key components of extracellular matrix in tumor microenvironment, only recent functional studies have shown a causal

relationship between loss of collagens and tumor progression [20]. Our results suggest that collagens may represent another new class of tumor suppressors. Notably, two genes FBN1 and TGFBR2, associated with a genetic disorder of connective tissue known as Marfan syndrome [21, 22], had 43 and 22 HGMD/COSMIC overlapping mutations respectively. Upon further investigation, we found that the two genes are mutated frequently in lung squamous cell carcinomas (SCCs) with a combined mutation frequency 10% in the TCGA cohort [23]. Moreover, FBN1 and TGFBR2 mutations are associated with poor survival. As shown in Figure 4, FBN1 mutation-harboring lung SCCs had poor disease progression free survival (DFS) (Figure 4A), and those patients with TGFBR2 mutations had both poor DFS and overall survival (OS) (Figure 4B, 4E). The combined FBN1 and TFGBR2 mutations are associated with both poor DFS and OS (Figure 4C, 4F).

Table 2. Genes ranked by the number of distinct overlapping HGMD-COSMIC mutations. Only genes with at least 20 overlapping mutations are shown. CGC: cancer gene census. TSG: tumor suppressor gene.

Gene	Mutations	CGC	Oncogene/TSG	Gene	Mutations	CGC	Oncogene/TSG
TP53	198	Yes	TSG	F9	33		
APC	192	Yes	TSG	DMD	33		
VHL	173	Yes	TSG	PKHD1	31		
NF1	148	Yes	TSG	SMAD4	31	Yes	TSG
PTEN	145	Yes	TSG	PTPN11	31	Yes	Oncogene
RB1	91	Yes	TSG	RYR2	30		
SCN5A	67			COL3A1	29		
CDKN2A	66	Yes	TSG	MLH1	29	Yes	TSG
NF2	65	Yes	TSG	BRCA1	29	Yes	TSG
KMT2D	56	Yes		MSH2	28	Yes	TSG
F8	56			VWF	27		
MYH7	54			TSC2	27	Yes	
SCN1A	52			STK11	27	Yes	TSG
USH2A	50			PTCH1	27	Yes	TSG
ATM	50	Yes	TSG	ATP7B	25		
MEN1	49	Yes	TSG	WT1	24	Yes	TSG
CFTR	48			TGFBR2	22		
FBN1	43			PAH	22		
HNF1A	39	Yes	TSG	IRF6	22		
RET	39	Yes	Oncogene	COL7A1	22		
ABCA4	37			CASR	22		
RYR1	36			APOB	22		
BRCA2	36	Yes	TSG	GCK	21		
LDLR	35			MYBPC3	20		

3.3. *Identification of potential oncogenes*

To identify putative oncogenes from the overlapping HGMD/COSMIC mutations, we applied two criteria. First, as most well-known oncogenic, activating mutations are highly recurrent in a specific tumor type, we ranked HGMD/COSMIC overlapping mutations by their mutation frequency. This was done separately for each tumor type in COSMIC. Second, because different oncogenic mutations in a given tumor type are often mutually exclusive, we performed mutual exclusivity analysis to identify those HGMD/COSMIC overlapping mutations that are not only

highly recurrent but also mutually exclusive from mutations in known oncogenes based on oncogene classification by Vogelstein et al [5].

To achieve sufficient statistical power in mutual exclusivity analysis, we only analyzed 19 tumor types with at least 200 samples that had whole genome or exome sequencing data in COSMIC and focused on those HGMD/COSMIC overlapping mutations in non-cancer genes (oncogene or tumor suppressor according to Vogelstein et al.) that are mutated in at least 1% of the total samples in a specific tumor type. Interestingly, of the 19 tumor types we analyzed, only endometrium, large intestine, and upper aero-digestive tract (UADT) cancers had such mutations, indicating that while only a very small percentage of COSMIC somatic mutations overlap with HGMD pathogenic variants (Table 1), even fewer are mutated in cancers with high recurrence. Notably, the ACVR1 R206H mutation occurred in 3 endometrium cancer samples, and an additional endometrium tumor harbors the ACVR1 G356D mutation. Mutual exclusivity analysis revealed that 3 of these 4 samples are mutually exclusive from the most frequently mutated oncogene PIK3CA, CTNNB1 and KRAS in this tumor type (p-value = 0.078; Figure 5).

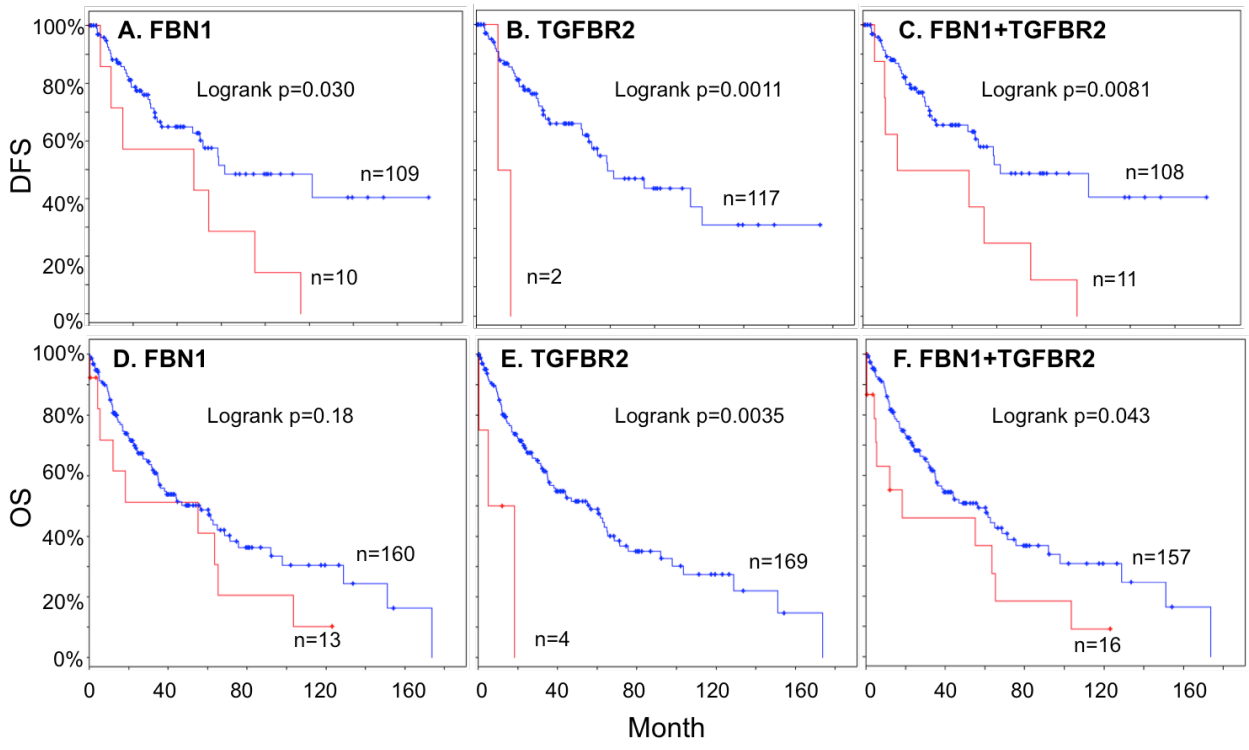

Figure 4. FBN1 and TGFBR2 mutations are associated with poor survival in lung squamous cell carcinomas. Disease free survival (DFS) are shown in panel A-C, and overall survival (OS) are shown in panel D-F. Red curves represent patients harboring somatic mutations for the indicated gene and blue curves represent patients with wild type gene. Sample size in red and blue curves, and logrank p-values in survival analysis are shown in each panel.

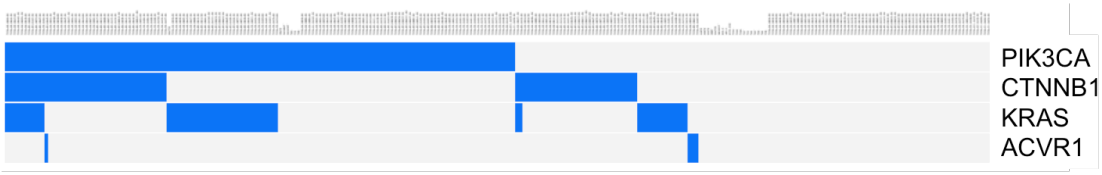

Figure 5. Mutual exclusivity of HGMD/COSMIC overlapping ACVR1 mutations from most frequently mutated oncogenes in endometrium cancers. Each column represents a tumor sample. The presence of a mutation in each gene in a given tumor sample is indicated by the blue color.

Since the above approach combining high mutation frequency and mutual exclusivity from known oncogenic drivers in each specific tumor type led to very few candidates as putative oncogenic mutations, we ranked somatic mutations only based on frequency across all cancer types in COSMIC without taking mutual exclusivity into consideration (Table 3). Many oncogenes have multiple mutational hotspots, and therefore for each gene we only show the mutation (at amino acid level) with the highest recurrence. Of genes with the most recurrent amino acid change occurring in at least 15 tumors, 18 had oncogene/tumor suppressor annotations (Table 3). While 50% (9/18) are classified as oncogenes, the presence of many tumor suppressors is not surprising because mutational hotspots (typically dominant negative mutations) are also observed in some tumor suppressors such TP53 [24]. The remaining genes without oncogene/tumor suppressor annotations provide possible candidate oncogenes due to the presence of mutational hotspots. It is noteworthy that there are 3 protein kinases that had a recurrent somatic mutation detected in more than 10 but less than 15 tumor samples: RAF1, p.S257L, 13 tumors; FGFR4, p.G388R, 12 tumors; TYK2, p.V362F, 12 tumors. Although the 3 kinases are not recognized as oncogenes, there are strong evidences that these recurrent mutations are activating and/or oncogenic [25, 26, 27], suggesting RAF1, FGFR4 and TYK2 are likely novel oncogenes.

4. Discussion

Owing to technological advancement and cost reduction, genomic sequencing is a new paradigm in cancer research and personalized cancer therapeutics. A large number of cancer somatic mutations have been described from whole genome/exome sequencing studies. As only a small percentage of somatic mutations are cancer drivers, it is of paramount importance to distinguish those driver mutations from a background of predominantly passenger mutations. Although many computational methods have been developed to predict the functional consequences of mutations [7], it has been indicated that their utility is limited [28]. In this study, we applied an alternative approach to discover cancer drivers from genomic sequencing data. By overlapping cancer somatic mutations and well-defined pathogenic disease-causing germline variants in Mendelian diseases, we identified putative tumor suppressors and oncogenes, which warrant follow-up functional studies. Our analyses suggested that ion channels, collagens and Marfan syndrome-related genes may represent new classes of tumor suppressors. More significantly, mutations in two Marfan syndrome-related genes FBN1 and TGFBR2 are associated with poor prognosis in lung squamous cell carcinomas, providing novel biomarkers with potential clinical relevance in areas of prevention, diagnosis and treatment [29]. Although the previous report

by Zhao and Pritchard [13] also interrogated overlapping pathogenic mutations in inherited diseases and cancer somatic mutations, we applied a novel approach to identify candidate tumor suppressors and oncogenes separately based on different criteria. Our approach is particularly useful in identifying the above highlighted putative tumor suppressors.

Table 3. Genes ranked by mutation frequency of the most recurrent HGMD-COSMIC overlapping mutation (at amino acid level) for each gene. Tumor samples with genome-wide sequencing data were used in the analysis. Only genes with the most recurrent amino acid change in at least 15 tumors are shown. TSG: tumor suppressor gene.

Gene	Mutation	Tumors	Oncogene/TSG	Gene	Mutation	Tumors	Oncogene/TSG
KRAS	p.G12D	524	Oncogene	TMEM106B	p.T185S	19	
IDH1	p.R132H	293	Oncogene	TAS2R43	p.H212R	19	
PIK3CA	p.H1047R	274	Oncogene	ROCK2	p.T431N	18	
TP53	p.R175H	226	TSG	PRNP	p.M129V	18	
APC	p.R1450*	66	TSG	GZMB	p.P94A	18	
PTEN	p.R130Q	42	TSG	PON2	p.S311C	17	
CDKN2A	p.R80*	40	TSG	KRT14	p.A94T	17	
CHEK2	p.Y390C	37		HNF1A	p.I27L	17	TSG
SMAD4	p.R361H	31	TSG	FGFR2	p.S252W	17	Oncogene
ABCD1	p.S606P	29		NRAS	p.G13D	16	Oncogene
KMT2C	p.T316S	26		IL1A	p.A114S	16	
OPRD1	p.C27F	25		HLA-DPB1	p.M105V	16	
PRDM9	p.T681S	24		EME1	p.I350T	16	
IDH2	p.R140Q	24	Oncogene	ALK	p.R1275Q	16	Oncogene
ARID1A	p.R1989*	24	TSG	ABCA1	p.R219K	16	
AR	p.Q58L	24	Oncogene	POU5F1B	p.E238Q	15	
UGT2A1	p.R75K	23		LTF	p.K47R	15	
PRSS1	p.K170E	23		IFIH1	p.A946T	15	
UGT1A7	p.N129K	22		HLA-A	p.L180*	15	
USH2A	p.C3416G	21		GRIN3B	p.T577M	15	
TGFB1	p.P10L	20		FGFR3	p.Y373C	15	Oncogene
RAD21L1	p.C90R	20		BRCA2	p.N372H	15	TSG
HRG	p.P204S	20		ATM	p.R337C	15	TSG

From our analyses, we rediscovered genes with cancer predisposing mutations, including TP53, APC, VHL, RB1 and many others (Table 2), which enhanced our confidence in the approach. However, as these genes have been well studied with respect to both germline mutations in familial cancer syndromes and somatic mutations in cancers, our focus lies on those genes with unknown connections between Mendelian diseases and specific cancers associated with the identical mutations. As genes often function differently in development versus in adult tissues, it is critical to further investigate the molecular pathways modulated by those genes in order to understand the mechanisms by which the same mutations can cause Mendelian diseases during development and drive tumor growth in adult tissues. This is best illustrated by an example in our oncogene discovery that revealed 5 HGMD/COSMIC overlapping mutations in ACVR1 gene cumulatively occurred in 19 central nervous system (CNS) cancers (data not shown). While the 5 ACVR1 mutations in germline cause fibrodysplasia ossificans progressiva (FOP), an autosomal dominant disorder of skeletal malformation and disabling heterotopic ossification [30], the same mutations are somatic oncogenic drivers in a subtype of CNS cancers, specifically

diffuse intrinsic pontine glioma (DIPG) [31]. Functional studies have demonstrated the ACVR1 mutations in germline activate the canonical bone morphogenic protein (BMP) pathway to promote osteogenic differentiation and endochondral bone formation resulting in FOP, and the same BMP pathway activated by these mutations in astrocyte cells in the brain accelerates cell proliferation ultimately leading to malignancy [32]. Therefore, these seemingly unrelated two diseases involving different tissue and cell types might be connected by the same molecular pathway activated by identical mutations in germline or in somatic cells. As described in the results section, two of these five mutations are also present in endometrium cancers, and they are largely mutual exclusive from the most frequently mutated oncogenes (Figure 5), suggesting that deregulated activation of the BMP pathway in uterus epithelial cells is likely a key oncogenic mechanism in at least some cases of endometrium cancers. Interestingly, the ACVR1 mutations and their potential oncogenic roles in endometrium cancers were also discussed in a recent study [13].

We recognize the limitations in our study. Since the percentage of cancer somatic mutations overlapping with germline pathogenic variants is small (0.6% of somatic mutations, 1.6% of recurrent somatic mutations in COSMIC; Table 1), our approach will not be applicable to the majority of the somatic mutation data from cancer genomic sequencing. Furthermore, identification of putative oncogenes based on high recurrence and mutual exclusivity from known oncogenes yielded few candidates. This is partly due to the fact that very few HGMD/COSMIC overlapping mutations have high recurrence in cancers. In addition, lack of mutual exclusivity with known oncogenes does not necessarily preclude the mutations as cancer drivers. Our goal was only to identify potentially novel cancer genes with high confidence. As more cancer genomic sequencing data become available in COSMIC, our approach will likely lead to the identification of additional putative oncogenes. Another limitation is that most of the candidate cancer genes from our analysis lack apparent functional connection to cancer development. This is somewhat expected due the inherent nature of our approach using Mendelian diseases pathogenic variants to aid novel cancer gene discovery. Accordingly, our study demonstrates a powerful technique for hypothesis generation to identify associations that warrant further experimental validation.

Acknowledgments

We thank Robert Maki from Icahn School of Medicine at Mount Sinai for valuable discussions. This work was supported in part through the computational resources and staff expertise provided by the Department of Scientific Computing at the Icahn School of Medicine at Mount Sinai.

References

1. Chmielecki J, Meyerson M. *Annual review of medicine* **65**, 63-79 (2014).
2. Garraway LA, Lander ES. *Cell* **153**, 17-37 (2013).
3. Hudson TJ*, et al. Nature* **464**, 993-998 (2010).
4. Tomczak K, Czerwinska P, Wiznerowicz M. *Contemporary oncology (Poznan, Poland)* **19**, A68-77 (2015).
5. Vogelstein B, Papadopoulos N, Velculescu VE, Zhou S, Diaz LA, Jr., Kinzler KW. *Science (New York, NY)* **339**, 1546-1558 (2013).
6. Lawrence MS*, et al. Nature* **499**, 214-218 (2013).

7. Gnad F, Baucom A, Mukhyala K, Manning G, Zhang Z. *BMC genomics* **14 Suppl 3**, S7 (2013).
8. Lawrence MS*, et al. Nature* **505**, 495-501 (2014).
9. Peterson TA, Doughty E, Kann MG. *Journal of molecular biology* **425**, 4047-4063 (2013).
10. Friend SH*, et al. Nature* **323**, 643-646 (1986).
11. Nagy R, Sweet K, Eng C. *Oncogene* **23**, 6445-6470 (2004).
12. Melamed RD, Emmett KJ, Madubata C, Rzhetsky A, Rabadan R. *Nature communications* **6**, 7033 (2015).
13. Zhao B, Pritchard JR. *PLoS genetics* **12**, e1006081 (2016).
14. Forbes SA*, et al. Nucleic acids research* **43**, D805-811 (2015).
15. Stenson PD, Mort M, Ball EV, Shaw K, Phillips A, Cooper DN. *Human genetics* **133**, 1-9 (2014).
16. Futreal PA*, et al. Nature reviews Cancer* **4**, 177-183 (2004).
17. Than BL*, et al. Oncogene*, (2016).
18. Xie C*, et al. Oncogene* **32**, 2282-2291, 2291.e2281-2287 (2013).
19. Zhang JT*, et al. Biochimica et biophysica acta* **1833**, 2961-2969 (2013).
20. Martins VL*, et al. Journal of the National Cancer Institute* **108**, (2016).
21. Loeys B*, et al. Human mutation* **24**, 140-146 (2004).
22. Mizuguchi T*, et al. Nature genetics* **36**, 855-860 (2004).
23. The Cancer Genome Atlas. *Nature* **489**, 519-525 (2012).
24. Stracquadanio G*, et al. Nature reviews Cancer* **16**, 251-265 (2016).
25. Imielinski M*, et al. The Journal of clinical investigation* **124**, 1582-1586 (2014).
26. Tomasson MH*, et al. Blood* **111**, 4797-4808 (2008).
27. Ulaganathan VK, Sperl B, Rapp UR, Ullrich A. *Nature* **528**, 570-574 (2015).
28. Miosge LA*, et al. Proceedings of the National Academy of Sciences of the United States of America* **112**, E5189-5198 (2015).
29. Iyengar P, Tsao MS. *Surgical oncology* **11**, 167-179 (2002).
30. Kaplan FS*, et al. Human mutation* **30**, 379-390 (2009).
31. Zadeh G, Aldape K. *Nature genetics* **46**, 421-422 (2014).
32. Taylor KR, Vinci M, Bullock AN, Jones C. *Cancer research* **74**, 4565-4570 (2014).