ARTICLE

Towards automation of germline variant curation in clinical cancer genetics

Vignesh Ravichandran [1, 2], Zarina Shameer [1, 2], Yelena Kemel [1, 2], Michael Walsh[2], Karen Cadoo[2], Steven Lipkin[3], Diana Mandelker[4], Liying Zhang[4], Zsofia Stadler[2], Mark Robson[1,2,3,5], Kenneth Offit [1, 2, 3, 6] and Joseph Vijai [1, 2,✉]

[1]Niehaus Center For Inherited Cancer Genomics, Memorial Sloan Kettering Cancer Center, [2]Clinical Genetics Service, Department of Medicine, Memorial Sloan Kettering Cancer Center; [3]Weill Cornell Medical College, New York, N.Y., 10065; [4] Diagnostic Molecular Pathology, Memorial Sloan Kettering Cancer Center, [5]Breast Medicine Service, Department of Medicine, Memorial Sloan Kettering Cancer Center, [6]Cancer Biology and Genetics Program, Memorial Sloan Kettering Cancer Center, New York, N.Y., 10065

* corresponding author: josephv@mskcc.org

Cancer care professionals are confronted with interpreting results from multiplexed gene sequencing of patients at hereditary risk for cancer. Assessments for variant classification now require orthogonal data searches, requiring aggregation of multiple lines of evidence from diverse resources. The burden of evidence for each variant to meet thresholds for pathogenicity or actionability now poses a growing challenge for those seeking to counsel patients and families following germline genetic testing. A computational algorithm that automates, provides uniformity and significantly accelerates this interpretive process is needed. The tool described here, **Patho**genicity of **M**utation **An**alyzer (PathoMAN) automates germline genomic variant curation from clinical sequencing based on ACMG guidelines. PathoMAN aggregates multiple tracks of genomic, protein and disease specific information from public sources. We compared expert manually curated variant data from studies on (i) prostate cancer (ii) breast cancer and (iii) ClinVar to assess performance. PathoMAN achieves high concordance (83.1% pathogenic, 75.5% benign) and negligible discordance (0.04% pathogenic, 0.9% benign) when contrasted against expert curation. Some loss of resolution (8.6% pathogenic, 23.64% benign) and gain of resolution (6.6% pathogenic, 1.6% benign) was also observed. We highlight the advantages and weaknesses related to the programmable automation of variant classification. We also propose a new nosology for the five ACMG classes to facilitate more accurate reporting to ClinVar. The proposed refinements will enhance utility of ClinVar to allow further automation in cancer genetics. PathoMAN will reduce the manual workload of domain level experts. It provides a substantial advance in rapid classification of genetic variants by generating robust models using a knowledge-base of diverse genetic data. *https://pathoman.mskcc.org*

Introduction

The uptake of genetic testing and targeted resequencing of cancer susceptibility genes to facilitate precision cancer prevention and early diagnosis has grown exponentially in concert with the decreasing costs of next generation sequencing (NGS)[1; 2]. A major challenge is the interpretation of sequence variants which result from clinical sequencing. American College of Medical Genetics and Genomics (ACMG) and the Association of Molecular Pathology (AMP) have published guidelines on interpretation of germline variants taking into account not only their pathogenicity but also their clinical actionability[3]. Nonetheless, germline variant classification continues to pose an immense burden on the time and resources of diagnostic molecular labs and cancer care professionals. The ACMG classification schema requires manually exploring multiple lines of public data, other orthogonal data sources and from the literature; then aggregation and scoring to provide evidence for classifying variants[4]. There is currently no automated computational framework for classifying genetic variants that is widely available. We developed PathoMAN, a computational resource that automates this classification with uniformity, transparency and speed, in order to facilitate variant curation for the cancer genetics community.

PathoMAN is a Python based variant curation algorithm that classifies germline genomic variants that are detected in the context of clinical cancer genetics sequencing. The schema is inspired by ACMG/AMP classification[3]. It aggregates multiple tracks of genetic and molecular evidences using variant annotators and from public repositories containing evidence for pathogenicity assertion. The compiled data is then used in 28 distinct categories addressing the type of the mutation, its biological impact, in silico predictions, presence in the control cohort as well as the familial information and inheritance pattern of the mutation. The aggregate score resulting from these categories is used in classifying the variant as Pathogenic (P), Likely-Pathogenic (LP), Benign (B), Likely-Benign (LB) or Variant of Uncertain Significance (VUS). The performance of PathoMAN was measured by re-evaluating expertly curated germline cancer variants in cancer genes compiled from two large published studies on breast and prostate cancer. In this study, we also assess the frequency of clinically actionable mutations present in general population using ExAC data in cancer susceptible and cancer predisposing genes as well as in ClinVar. We test the application of ACMG criteria for germline cancer variants and address the bottleneck of variant curation in using automated algorithms in variant classification.

Materials and Methods

Test datasets

We tested 11,196 germline variants in 76 genes of highest cancer relevance to describe and compare performance of the algorithm in cancer related genomic variant data. Clinically actionable variants (P/LP) in these 76 genes are reported back to patients at MSKCC as part of the IMPACT® assay using the appropriate IRB approved protocol[5; 6]. Several of the genes in this set are strong cancer predisposition genes, while others are putative candidate genes across one or more syndromic cancers. We have selected this gene list, referred as IMPACT-76[7],(**Suppl Table 1**) uniformly throughout this manuscript. We chose exonic and essential splice site (+/- 1,2) variants in the IMPACT-76 genes from three manually curated datasets - (i) prostate cancer study[8]; (ii) breast cancer study[9] and (iii) ClinVar[10; 11].

The Exome Aggregation Consortium[12] (ExAC) is a joint effort to aggregate exome sequencing data from fourteen large sequencing projects to provide summary data such as ethnicity specific allele frequency for a wider scientific community. The ExAC-noTCGA data is a subset of 53,105 samples and it doesn't include 7601 The Cancer Genome Atlas (TCGA) cancer germline samples. The variants in test data were compared against ExAC-noTCGA Non-Finnish European population and were

considered not to have *de novo* evidence and co-segregation, as this information was unavailable for these datasets. However, PathoMAN can use such information, if available to assign ACMG classes. The results were compared against the manual curation. They are reported here in four categories: concordance, discordance, loss of resolution and gain of resolution.

When the reported P/LP and B/LB variants are re-classified as P/LP and B/LB respectively by PathoMAN, then the results are considered concordant. Similarly when reported P/LP and B/LB variants are re-classified as B/LB and P/LP by PathoMAN respectively, then the variants are considered discordant. When reported P/LP or B/LB variants are re-classified as VUS by PathoMAN, then they are placed in the loss of resolution (LOR) category and when the reported VUS are re-classified as P/LP or B/LB, then are considered as gain of resolution (GOR) category **(Table 1)**. As we describe below, these evaluations aid in understanding the real-world usage of the eight ACMG categories of evidence in cancer genomics, and in their ability to discriminate between P/LP, B/LB and VUS.

Rating	Reported	PathoMAN
Concordance	P/LP; B/LB	P/LP; B/LB
Discordance	P/LP; B/LB	B/LB; P/LP
Loss of Resolution (LOR)	P/LP; B/LB	VUS
Gain of Resolution (GOR)	VUS	P/LP; B/LB

P-pathogenic, LP–likely pathogenic; B–benign; LB–likely benign; VUS–Variants of Uncertain Significance.

Table 1: Rating the comparison of results between PathoMAN variant curation and domain expert-curation of germline variants.

ExAC subset of IMPACT-76

The ExAC[12] is a public resource often utilized as convenience controls for several human traits in case-control studies. We wanted to estimate the burden of variants in ExAC-noTCGA as classified by PathoMAN and contrast against known information in ClinVar. We selected 55,566 variants from IMPACT-76 genes which were in exonic or essential splice site regions. This is not considered part of the test datasets descibed earlier, as ExAC data is used as part of the ACMG criteria PS4, PM2, BA1, BS1 and BS2.

ACMG/AMP guidelines

The 2015 ACMG/AMP guidelines consist of 16 criteria that aid in classifiying pathogenicity and 12 criteria that aid for benignity. Pathogenicity criteria were broadly grouped as: very strong evidence (PVS1), strong evidence (PS1-PS4), moderate evidence (PM1 – PM5) and supporting evidence (PP1-PP5). Benign criteria are broadly grouped as: standalone evidence (BA1), strong evidence (BS1-BS4) and supporting evidence (BP1-BP7). This classification system resolved a variant as pathogenic or benign based on eight components – population frequency data, genomic annotation and computational predictive data, functional data, segregation data, *de novo* data, allelic/genotypic data, public databases and literature and other data **(Table 2).** The variant classification criteria used by PathoMAN were inspired by these ACMG/AMP guidelines, although they do not precisely adhere to these published norms. We describe below the criteria and their modifications for variant classification by PathoMAN for cancer genetics.

Variant Annotation

PathoMAN makes use of CAVA[13] for genomic annotation, dbNSFP[14-16] using Annovar[17] for in silico predictions, ExAC-noTCGA and gnomAD[12] for public control frequencies, ClinVar[10,11] for public evidences, and curated list of variants from the literature for functional evidences.

Determination of PVS1: null variants

Curated lists of cancer-causing genes from the literature, various genetic testing panels,[18-22] and OMIM genes that caused autosomal dominant disease[23] were aggregated. If the variant in a gene from this list was a Tier 1 mutation (frameshift, truncating, essential splice variant and initiation codon), and not present in thelast exon, then PVS1 was scored 1. A gene with a functional domain encoded by the last exon, such as, *ATM* was an exception to the last exon criteria[24]. PVS1 was not scored for *BRCA2* mutations observed after the polymorphic stop rs11571833 (K3326X).

Category	Knowledge-base
Population frequency data	ExAC-noTCGA and gnomAD
Genomic annotation and computational predictive data	CAVA, SnpEFF, dbNSFP using Annovar (CADD, FATHMM, LRT, MutationAssessor, MutationTaster, PROVEAN, Polyphen2-HDIV, Polyphen2-HVAR, RadialSVM, SIFT, VEST3, MCAP) and dbscSNV using Annovar
Functional data	In-house curated list of variants with functional evidence from literature
Segregation data	User-defined information
Denovo data	User-defined information
Allelic/genotypic data	NA
Public database and literature	ClinVar, Pubmed, curated list of autosomal dominant genes
Other data	Family history to specific disease with single gene etiology

Table 2: List of data sources and annotators used in building PathoMAN's knowledge-base

Determination of PS1 and PM5: known pathogenic missense

If a missense variant was reported as pathogenic by multiple submitters with no conflicts and had a gold star of 2 or more in ClinVar[25,10,11], irrespective of the alternative allele but leading to the same amino acid change, then PS1 was scored 1. If a missense variant was not seen in ClinVar but had another pathogenic missense variant at the same amino acid with a different amino acid change, then PM5 was coded 1.

Determination of PS3 or BS3: strong prior evidence of pathogenic or benign

An aggregated select list of reported pathogenic and benign variants from the literature[2; 26-29] were used as a knowledge-base for PS3/BS3. If the variants were in the curated list (missense variants in BRCA1/2 reported by ENIGMA), or if it was a truncating variant and ClinVar had reported it as pathogenic or benign with a gold star 2 or more, then PS3 or BS3 was coded 1. Our selective use of ClinVar was a workaround to establish prior knowledge for ascribing higher confidence for the truncating variants and select missense variants reported by domain experts that are either pathogenic or benign.

Determination of PS4 PM2 BA1 BS1 and BS2: rarity and enrichment of variant in cases

If a variant was present in aggregated public controls such as the ExAC-noTCGA dataset and gnomAD[12] with an allele frequency greater than 5%, then the variant was coded 1 for BA1. If the variant had an allele frequency in public controls between 1% and 5% then BS1 was coded 1. If the variant was also present

in a homozygous form in the public controls, then BS2 was coded 1. If the variant was absent from ExAC-noTCGA data or gnomAD general population data, and then PM2 was coded 1. For variants, not scored as BA1, BS1, BS2 or PM2, Fishers Exact test was performed against the user defined population (ExAC-noTCGA or gnomAD). The population included all the major groups (NFE, FIN, SAS, AMR, AFR, EAS) in ExAC and ASJ population in the gnomAD database. If the odds ratio was greater than 3 and p-value less than 0.05, then the variant was given a score of 1 for PS4. This was a robust measure for weighting pathogenicity in uncommon variants and non-singletons.

Determination of PM1: membership in a protein domain of functional significance

If the amino acid that was being altered by the mutation was present in a protein domain, or a residue involved in signalling, binding with other proteins, or in an active site, then PM1 was coded 1. Currently, we use Uniprot for annotation of protein features[30]. We acknowledge the incremental value of a curated somatic hotspot list (http://cancerhotspots.org) to aid in this classification.

Determination of PM4 BP3: genomic complexity and context of the variant

If the mutation was an in-frame insertion/deletion or a stop loss in a non-repetitive region, then the variant was coded 1 for PM4. Instead, if it was an in-frame insertion/deletion in a repetitive region, then BP3 was coded 1. The repeat masker track from UCSC genome browser was used for this[31-34] criteria.

Determination of PP3 BP4: In silico prediction of deleteriousness

We used Annovar[17] to annotate the variants with dbNSFP[14-16,35] track to get results of deleteriousness predictions from 12 in silico algorithms – CADD[36], FATHMM[37], LRT[38], MutationAssessor[39], MutationTaster[40], PROVEAN[41], Polyphen2-HDIV[42], Polyphen2-HVAR[42], RadialSVM[35], SIFT[43], VEST3[44] and M-CAP[45]. Use of an ensemble of in silico prediction algorithms improves prediction across a wide range of genes and cancer types[46]. Hence if more than 7 (>50%) algorithms call a variant deleterious, then the variant was coded 1 for PP3. Otherwise, BP4 was coded 1. In contrast, many of the old ClinVar records relied on only SIFT or Polyphen.

Determination of PP5 BP6: Known variant with insufficient details.

If the variant was in ClinVar[10,11] with gold star less than 2 and was pathogenic, then PP5 was scored 1. If it's benign, BP6 was scored 1. Thus, a variant reported once in ClinVar does not command high value in the pathogenicity determination, but can be upgraded depending on other ancillary information tagged to it.

Determination of BP7: synonymous variants

For synonymous silent mutations, we used adaptive boosting and random forest scores from dbscSNV[47], which if it was less than 0.6, then BP7 was scored 1. dbscSNV is a database of precomputed prediction scores for SNVs, that may occur in splice consensus regions. Higher scores reflect the variants effect in splicing[47].

Determination of PP2 and BP1: missense driven disease genes

We used ClinVar[10; 11] to collect all reported missense variants per gene. We then selected the confident (gold star 2 or more) pathogenic and benign calls. The list of genes with higher ratio of pathogenic to benign variants called missense-driven pathogenic genes and the list of genes with lower ratio of pathogenic to benign variants were called missense-driven benign genes. Any missense variant in the missense-driven pathogenic gene list was scored PP2 and any missense variant in the missense -driven benign gene list was scored BP1. Classic example of such genes are *PTEN*[49] and *TP53*[50]. Almost all *PTEN* missense mutations were pathogenic **(Supp Figure 1)**.

Determination of PS2, PM6, PP1 and BS4: denovo and co-segregation

ACMG criteria require both paternity and maternity confirmed for *de novo* variants. PathoMAN requires user input for *de novo* status and segregation information for classification. Three options for *de novo* evidence includes, *de novo* with both paternity and maternity confirmed, *de novo* without paternity or maternity confirmed and no *de novo* evidence at al. Similarly, for co-segregation evidence, the options provided were co-segregation with the disease, lack of co-segregation with the disease and no co-segregation. For larger trio studies, in the future, we expect to include a module that looks for *de novo* variants computationally from a mutisample VCF and pedigree file information.

Determination of PM3, BP2: recessive inheritance

These two categories apply to variants with a recessive disease. Germline cancer variants are generally associated with cancer predisposition syndromes in an autosomal dominant inheritance pattern. Hence, PM3 and BP2 doesn't apply for current PathoMAN variant classification. They were scored 0. We expect to add compound heterozygosity to the next version of the algorithm. We also will incorporate select gene variants in mismatch repair genes that can be classified by this category into the knowledge-base.

Determination of PP4, BP5: disease specific conditions

PP4 and BP5, per ACMG were two criteria to consider in a patient's predisposition to a specific disease with single gene aetiology. Cancer is a disease with multi-gene aetiology although certain genes such as *RB1* may be strong candidates for PP4. Once we incorporate a pedigree file, and cancer phenotype variables, we should be able to apply these criteria. In the current iteration, these were scored 0.

The final classification schema was based on the original ACMG scoring and pathogenicity was predicted for the datasets described.

Usage of ClinVar (Release 12/28/2017) in PathoMAN

ClinVar is the most popular database that processes submissions of human DNA variants and assertions made regarding their clinical significance. One of its stated goals is to support computational re-evaluation of genotypes and assertions. ClinVar currently does not provide detailed information on pathogenicity assertions, unlike another initiative ClinGen[51; 52], which aims to build a more discriminatory and accurate assessment of variant pathogenicity. However, access to the ClinGen interface is at the moment restricted to a smaller niche group. PathoMAN uses ClinVar as one of its knowledge-base. We use the ClinVar data parser tool[53] to extract important fields that will aid in variant classification. PathoMAN takes advantage of the gold star system awarded to variants with highest evidences supporting the assertion of clinical significance. PathoMAN up-weights variants with gold star 2 or more in the knowledge-base. We selected variants with gold star < 2 and with the term "cancer" in traits field and used them as another test dataset. This test data set was filtered for only variants in the exonic and essential splice regions in IMPACT-76[54] genes (n=9164) variants (Supp Table 1).

Results

PathoMAN versus manual curation for test datasets

PathoMAN is a classification algorithm in python,

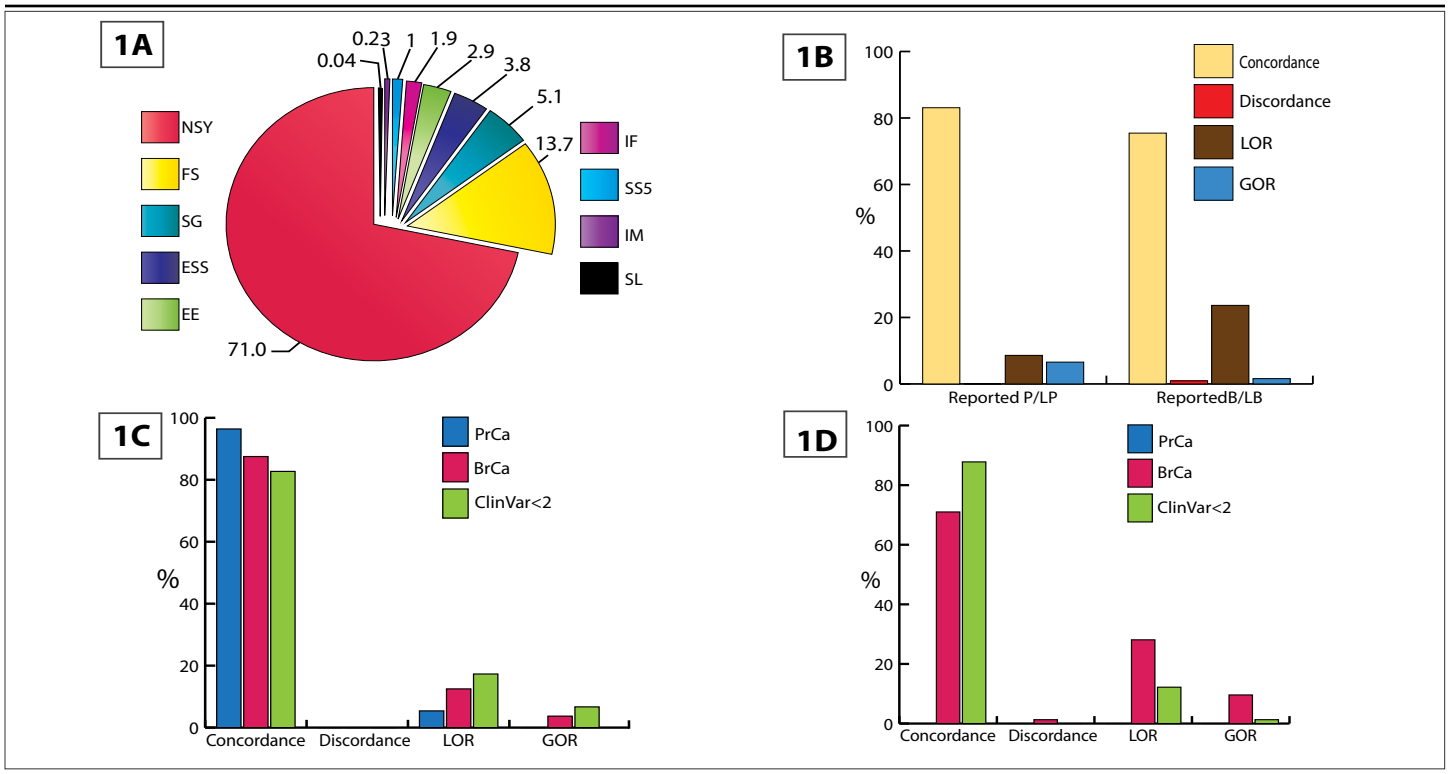


Figure 1A: Distribution of mutations in test data by variant class. (Stop loss - SL; Start codon alteration - IM; Splice variant that alters +5 splice site - SS5; Inframe ins/del - IF; Alters first 3 bases of codon - EE; Essential splice site +/- 1, 2 - ESS; Stop gain - SG; Frameshift - FS; Non-synonymous - NSY). **Figure 1B: Overall comparison results of PathoMAN variant curation and expert curation for germline variants from test dataset. Figure 1C: PathoMAN results for reported pathogenic and likely pathogenic germline variants in test dataset. Figure 1D: PathoMAN results for reported benign and likely benign germline variants in test dataset.**

provided as a web based service, that either allows user to query single variants or upload a file with a batch of variants. The web-tool is created using Flask web framework. Single variant query works on chromosome, position, reference allele, alternative allele, allele count, allele number, *de novo* status, co-segregation status and preferred control population sub-group. Batch upload requires six columns that follow VCF 4.2 format [*chr, pos, ref, alt, ac, an*]. User can select *de novo* status, co-segregation status and preferred control population sub group. Our pipeline is currently equipped to annotate a minimal VCF and prepare it for PathoMAN variant classification. The result of a single variant query is displayed back on to the web-page while the batch upload results will be mailed back to the registered user. For an annotated VCF file containing 1000 variants, PathoMAN takes 85 minutes, which is approximately 5 seconds per variant. This provides a massive advantage in terms of speed, uniformity, efficiency and service assurance than manual curation.

Overall the test dataset contained 11,196 variants in 76 genes with reported expert curation from three groups – (i) prostate cancer study[8]; (ii) breast cancer study[9] and (iii) ClinVar[10,11]. Missense mutations account for 71% of the variants, Frameshift variants 14%, Stop gain variants 5% and the rest distributed among splice variants, in-frame insertion/deletion and stop loss in the test dataset (**Figure 1A, Table 3**). We annotated the variants with CAVA, Annovar, ExAC noTCGA, gnomAD and ClinVar and prepared the VCFs for PathoMAN.

PathoMAN achieves an overall concordance of 83.1% for P/LP variants and 75.5% for B/LB variants. A minimal discordance of 0.04% (P/LP) and 0.9% (B/LB) and a LOR of 8.6% (P/LP) and 23.64% (B/LB) were observed. PathoMAN achieves GOR by resolving 6.6% of the VUS as P/LP and 1.59% as B/LB **(Figure 1B-D).** We estimated reliability of recall using the Cohen's kappa coefficient (K) between the manual curation and PathoMAN classification for the test dataset. The number of observed agreements was 89.21% (9988 unique variants) (K=0.74, CI 95% 0.73-0.75).

Out of the 2,535 P/LP variants reported by manual curation, PathoMAN showed very high concordance for frameshift, essential splice sites and truncating variants. PathoMAN failed to resolve reported P/LP for 233 missense variants (**Table 4A**). All of these were low confidence variants in the ClinVar dataset with no assertion criteria provided (gold star=0) or variants with a single submitter (gold star=1). Similarly for reported benign variants (n=440), we observe high concordance with the exception of missense variant class. LOR is observed for 90 missense variants with conflicting reports in ClinVar (**Table 4B**).

Variant Class	ClinVar	Prostate	Breast	Total
SL	4	0	0	4
IM	24	0	2	26
SS5	101	0	0	101
IF	199	0	12	211
EE	313	0	16	329
ESS	422	4	3	429
SG	543	13	13	569
FS	1482	36	13	1531
NSY	7425	3	568	7996
Grand Total	**10513**	**56**	**627**	**11196**

Table 3: Distribution of variants by variant class in the test dataset. Stop loss (SL),Start codon alteration (IM), Splice variant that alters +5 splice site (SS5), In-frame ins/del (IF), Alters first 3 bases of codon (EE), Essential splice site (+1, +2, -1, -1) (ESS), Stop gain (SG), Frameshift (FS),Non-synonymous (NSY).

Many rare variants that are reported from clinical sequencing/ studies have no public records documenting their pathogenicity classification and there may be insufficient evidence to call them either pathogenic or benign. These variants have been called VUS by the manual curators (n=8221 variants) at the time. PathoMAN is

able to resolve 6.7% of the VUS as LP and 1.6% as LB (**Table 4C**).

PathoMAN results for ExAC (no TCGA) dataset

We asked if we could predict the different classes of ACMG mutations in this public resource using PathoMAN. We selected 55566 exonic and essential splice variants from IMPACT-76 genes and classified them. Overall, PathoMAN calls <1% of the heterozygous genotypes in ExAC noTCGA dataset, from 53,105 samples, as P/LP. Further we tabulated pathogenic variant burden by genes and compared them against ClinVar (**Table 5**). The results show, PathoMAN calls a similar number of variants reported in ClinVar for the bona_fide cancer genes like *BRCA1* and *BRCA2*. PathoMAN also predicts a few rare variants as P/LP in these genes which have not been reported previously in ClinVar. Investigators who intend to use the ExAC noTCGA dataset as controls in cancer sequencing studies can use PathoMAN to get a rapid count of variants across genes above and beyond those reported in ClinVar.

Table 4A

Class	B/LB	P/LP	VUS	Total
EE	0	7	24	31
ESS	0	389	2	391
FS	0	1165	74	1239
IF	1	0	26	27
IM	0	0	15	15
NSY	0	87	233	320
SG	0	457	42	499
SL	0	1	0	1
SS5	0	0	12	12

Table 4B

Class	B/LB	P/LP	VUS	Total
EE	30	0	8	38
FS	1	0	3	4
IF	3	1	1	5
IM	0	0	1	1
NSY	296	3	90	389
SG	1	0	0	1
SS5	1	0	1	2

Table 4C

Class	B/LB	P/LP	VUS	Total
EE	2	10	248	260
ESS	0	35	3	38
FS	0	261	27	288
IF	1	1	177	179
IM	0	0	10	10
NSY	128	182	6977	7287
SG	0	51	18	69
SL	0	0	3	3
SS5	0	0	87	87

(Stop loss - SL; Start codon alteration - IM; Splice variant that alters +5 splice site - SS5; Inframe ins/del - IF; Alters first 3 bases of codon - EE; Essential splice site +/- 1, 2 - ESS; Stop gain - SG; Frameshift - FS; Non-synonymous - NSY).

Table 4: PathoMAN re-classification of expertly curated germline cancer variant reported in the test datasets. All tables show distribution by variant classes. **Table A** shows only reported P/LP variants**; Table B** shows only reported B/LB**;** and **Table C** shows only Reported VUS.

Usage of ACMG/AMP categories in PathoMAN

We analyzed the real-world usage of the eight categories of evidence (population frequency, genomic annotation and computational prediction, functional evidence, co-segregation, *de novo* status, allelic/genotypic data, public databases, scientific literature and other data) used in the ACMG/AMP guidelines. Interestingly, we find that the categories: population frequency data, genomic annotation and computational predictions, databases and scientific literature (**Figure 2**) are the most used. These are available due to generous data and tool-kit sharing policies in the genomics field. The categories that are rarely if ever used are familial co-segregation data or *de novo* status, allelic data, and functional data. The co-segregation data and *de novo* status data are limited to familial studies, and are mostly unavailable in sporadic case-control settings since these are collected by investigators, doctors, genetic counsellors and commercial labs based on patient input. For a variant to be classified as pathogenic or likely pathogenic by ACMG criteria, one needs a maximum of 1 PVS1 or 2 PSs or 3 PMs or 4 PPs for which, the knowledge-base and resources used by PathoMAN were demonstrably sufficient. We describe

GENE	ClinVar P/LP	PathoMAN P/LP	Difference
BLM	1	9	8
TP53	15	7	8
ATM	80	74	6
BRCA2	101	96	5
BARD1	10	15	5
SDHA	5	0	5
MLH1	6	11	5
EGFR	0	5	5
RAD51B	0	5	5
PALB2	21	26	5
RAD50	21	17	4
PMS2	15	11	4
CDH1	3	7	4
MRE11A	11	7	4
EPCAM	0	4	4
BRCA1	68	65	3
NF1	3	6	3
PTEN	2	5	3
STK11	0	3	3
KRAS	2	0	2
MUTYH	26	24	2
BRIP1	21	23	2
RAD51C	17	15	2
FH	6	4	2
RET	2	4	2
BMPR1A	1	3	2
FAM175A	1	0	1
RAD51	1	0	1
MSH6	13	12	1
NBN	10	9	1
RAD51D	6	7	1
APC	4	5	1
CDKN2A	5	4	1
DICER1	3	2	1
BAP1	0	1	1

Table 5:Comparison of pathogenic gene burden in ExACnoTCGA between ClinVar and PathoMAN. Columns contain variant counts**.**

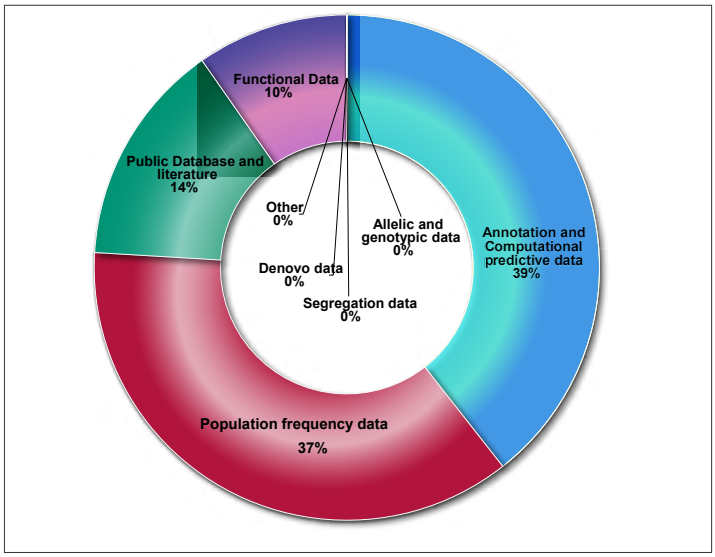


Figure 2: Utilization of knowledgebase components by PathoMAN during variant curation of the test datasets.

below the bottlenecks in sharing this information and propose a novel framework to circumvent and ameliorate these issues.

Discussion

PathoMAN as a tool to aid variant curation

Traditionally, genetic variant curation has been performed manually by expert groups of individuals. However this is a time intensive task that requires aggregation and interpretation of information from multiple sources. In the cancer realm, this was relatively easy at times when only a single gene such as *BRCA1/2* was under investigation. In contemporary testing scenarios which routinely rely on multiplex gene-panels, this task is onerous. Large gene discovery efforts, as well as clinical reporting, could use a simplified, automated, method for prioritizing variants for a closer look or in the best case, be useful as the classification tool of choice. PathoMAN addresses this critical unmet need for an unbiased algorithmic approach towards classifying genetic variants of clinical interest in cancer predisposition. PathoMAN can be easily accessed through a web browser and results for individual variants are almost immediately available, while batch uploads of variant VCFs may take a few hours.

Genetic testing labs have started utilizing the ACMG/AMP classification rules to classify variants for pathogenicity within cancer predisposition genes. However, results vary depending on availability of accessible data and interpretational differences[55]. Efforts are being made to resolve interpretational differences through initiatives underway such as ClinGen. In a recent report[56], 13 variants in ClinVar were re-analyzed, and were found to be unresolved, underscoring the difficulties even for expert curator groups. For manual or automated curation the minimal set of information required to classify a variant as likely pathogenic or likely benign are: population frequency, computational predictors and evidences from public databases[56]. PathoMAN compiles this information uniformly in a machine accessible format which is used as a knowledge-base for variant classification. An advantage of using PathoMAN is that it can easily tag benign variants based on public allele frequency and the genomic context information from annotators. This reduces the variant pool of interest to a manageable subset. In a typical multiplexed gene-panel variant list, after filtering for only rare high or moderate impact variants, PathoMAN will classify about one third of the variants as B/LB with high precision. This saves time and effort for the variant curators and helps them to focus on curating the remaining potentially actionable variants. PathoMAN can also tag founder mutations.

Cancer is a complex disease with multi-gene aetiology. Some cancer genes confer high risk whereas some only moderately affect the carrier's risk. Panel testing is currently used for active surveillance and intervention to lower disease risk. Large sequencing and genotyping efforts to discover new cancer predisposition genes are being carried out by several consortia like BCAC[57], SIMPLEXO[58]; COMPLEXO[59], CIMBA[60], etc. Several commercial and academic labs also now offer multiplexed panels. As the cost for sequencing these panels decreases, the number of genes tested in panels is increasing. Automation allows for rapid processing, service assurance and reproducibility of results for these large panels. Gold standard sets of curation pioneered by ClinGen[51; 52] would aid in refining these pathogenicity classifications further, while efforts such as the PROMPT[29] registry enable accurate penetrance estimates of mutations in susceptibility genes. The PROMPT registry has identified a 26% discordance rate among laboratories and an 11% rate with conflicting interpretations, a discrepancy that has implications for altering medical management.

In the three datasets we used to test PathoMAN, we demonstrate that, when contrasted against an expert curated set of variants in the IMPACT-76 genes, there is a high concordance rate for both pathogenic and benign variants. The concordance for P/LP variants is excellent when limited to truncating variants; frameshift variants and essential splice variants such as those reported in the prostate cancer study[8]. When missense variants are also considered such as in the breast cancer study[9], we see more VUS, but the discordance is still minimal.

Much of the discordance can be traced back to the ClinVar submissions with conflicting interpretations; e.g. *BRCA1*:c.5348T>C (p.Met1783Thr). Currently, PathoMAN doesn't have access to proprietary databases (such as HGMD)[62,63], which may have additional evidence for pathogenicity or benignity. For splice variants, our source is limited to −3-to+8 at the 5' splice site and −12-to+2 at the 3' splice site. Hence, we may be missing out on certain extended regions.

Many labs and certain programs such as cardio classifier[64] and InterVar[65] use prior knowledge of disease-gene pair association. This is advantageous to reduce classifications leading to P/LP for those genes that are not in a disease-gene pair. However, it also suffers from the disadvantage that it cannot be used for lesser known genes-disease pairs or for novel gene hunting. In a recent report, we showed that, half of the cases, in a series consisting of selected advanced cancers at a single institution, were non-syndromic associations[5]. Probands or their close relatives had clinically actionable variants in cancer genes not directly associated with the specific cancers for which there were known syndromic associations. PathoMAN does not use the contextual syndromic association in deciphering pathogenicity of variants. However, with the applications envisaged for novel gene discovery, this is a distinct advantage. For clinical sequencing which is more focused on specific sets of genes, is limiting to disease-gene pairs to identify pathogenic variants.

The variants that could not be classified by PathoMAN and are called VUS are due to lack of accessible, supporting evidence for the clinical assertion by using ACMG guidelines. These are classic examples of rare variants absent in ClinVar and ExAC datasets. For these LOR variants, we believe that the expert curators may have had additional evidence form literature, in-house functional evidence[66] or familial co-segregation information[67] that helped classify these variants as P/LP or B/LB. The upgrade for VUS to either LP or LB by PathoMAN is based on the three categories - lines

of available evidence in public databases, population frequency and computational and in silico prediction on deleteriousness. These variants can be re-classified as either pathogenic or benign if additional functional or co-segregation data become available through literature or initiatives such as ClinGen[56].

Commercial testing laboratories have proprietary versions of interpretation pipelines such as Sherloc[68] (Invitae Corporation) and MyVISION (Myriad Genetics). However, these are unavailable to the community at large. PathoMAN is designed to provide an optimized platform for clinical variant calling utilizing publically available data resources.

Using ACMG for variant classification in Cancer

Mutations in tumor suppressors and oncogenes lead to tumorigenesis, and the Knudson two-hit hypothesis[69] is seen to operate in many common cancers. Common examples include *APC*, *TP53*, *BRCA1/2* genes etc. However, several of these genes, especially those that are part of the Fanconi complex (*FANCS-BRCA1, FANCD1-BRCA2, FANCJ-BRIP1, FANCN-PALB2, FANCP-SLX4, RAD51C*), neurofibromatosis (*NF1*), Ataxia-telangiectasia (*ATM*), Bloom syndrome (*BLM*), Niemegen breakage syndrome (*NBN*), dyskeratosis congenita (*TERT*) that lead to autosomal recessive rare Mendelian disorders, are also found to be risk genes for autosomal dominant cancer predisposition. Heterozygous carriers of these gene mutations are reported to have increased risks for syndromic cancers[70]. Occasionally, gene disrupting heterozygous mutations in these genes that are rare, absent in public controls such as ExAC and gNOMAD may be observed in sequenced cancer cohorts. Their ClinVar record for pathogenicity is usually based on their Mendelian recessive syndrome and not to the cancer phenotypes. Hence, applying the ACMG rules to genes without membership in the ACMG list may be fraught with misclassification. However, we believe that continuing data streams for variants in these genes will lead to better classifications, especially when coupled with familial co-segregation and functional validations. While PathoMAN classifications for such genes are a useful starting point for identifying variants that may be pathogenic, and discarding benign ; we emphasise on expert manual curation to disentangle these issue.

Limitations of automation

Automating variant classification based on publically available information has some pitfalls. Supporting evidences provided in ClinVar for variants are not computation friendly and requires manual curation to interpret free text. In several instances, the citations are not relevant to the specific records. Technologies such as natural language processing and tagging will eventually help to build a knowledge-base that can further be used for deep learning.

Current ACMG guidelines do not directly link ClinVar functional evidence provided as supporting observations, which leads to loss of information that could be used in variant classification. Due to this lack of data structure, the variants in ClinVar are scored only PP5 or BP6 and not PS3 or BS3. We employed the gold star 2 or more status as a proxy for functional evidence. Not all clinical sequencing projects are equipped or do independent analyses to assess functional evidences for their clinical assertion. If the ClinVar evidence is coded with proper tags, it would be helpful for molecular geneticists and clinical curators to use this information for their pathogenicity estimation. For example *TP53* (R273H), *BRCA1* (Y105C) and *BRCA1* (V1688del) variants have overwhelming literature evidences (**Supp Figure 2**); however the evidence present in the description of the submissions within ClinVar, are computationally un-derivable. Similarly there are many variants reported in the literature which may have some level of supporting evidence for pathogenicity or benignity in ClinVar. Currently all of these data integration is done by manual curators on a case-by-case basis.

We propose a framework to report ClinVar data that can be structured and parsable for an automated algorithm in the context of cancer. This format consists of 6 important fields that compress the vast information that is present in literature or clinical reports.

1. Population/Ethnicity (NFE, AFR, SAS, AMR, ASJ, FIN, OTH, EAS, others)
2. Inheritance model (AD,AR, *de novo*, X-linked)
3. Allelic status (Hom, Het)
4. Family history/Co-segregation information (Yes-1; No-0)
5. Disease association (TCGA code/ Oncotree code[71])
6. Functional Evidence (Experiment type: NMC, LOH, etc.)

For example, *ERCC3* (R109X) variant[72] can be depicted as ASJ-AD-Het-1:1-*BRCA,BLCA*-NMD. This variant was seen in Ashkenazi Jewish individuals with an autosomal dominant inheritance for the heterozygous allele. This variant co-segregated in one family with cancer history. The variant was found in Breast cancer and Bladder cancer individuals and the functional evidence for pathogenicity was carried out by testing for non-sense mediated decay and other experiments.

Large sequencing studies and gene specific functional studies give curated list of variants with their pathogenic impacts like TP53 database[28] and a functional study on *PALB2* variants[26; 27]. As a primer, we have collated a list of *PALB2, TP53* variants from the literature as supporting the knowledge-base for PathoMAN but there is a real need to create a publically available well curated list of variants from the literature that is amenable to programmatic interpretation. Similarly, as standards evolve for the incorporation of somatic mutations into germline interpretation, we expect an integration of such events for atleast some tumor suppressor and oncogenes. The roles played by the ENIGMA Consortium[73; 74], G4GH[75], *BRCA*-Share[76] in this regard are meritorious. Though Clinical laboratories collaborate to resolve the differences in variant interpretations submitted to ClinVar[56], the fact remains however, that a unified framework for incorporation of supporting machine readable evidences in any variant database including ClinVar remains a critical bottleneck.

Functional data is rarely available for most genes. Exceptions are *BRCA1/2* due to the concerted efforts of the ENIGMA consortium[73; 74]. In single variant reports, data is usually buried within scientific jargon that is not compatible with genomic variant information. In many instances, functional data is dependent on the models used, e.g. overexpression of a mutant construct, deletion of a region using a CRISPR endonuclease and sometimes, introduction of the specific nucleotide through homology directed DNA repair. It is also likely, that the results from these three methods do not agree. Novel methods to understand deleteriousness using saturation mutagenesis are also starting to emerge[77-79]. We hope these will add a uniform layer of functional data that can be used in determining pathogenicity in the coming years.

In conclusion, we performed pathogenicity assessment of 66,762 variants in germline cancer genes (IMPACT-76), the first and largest uniform classification using an unbiased computational tool. We demonstrate the high concordance and low discordance when compared with manual curation as a harbinger of how such programs will in the near future, be able to work as well as domain experts and manual curators. PathoMAN is a first step towards our goal of automating the complex process of variant classification and interpretation. A beta version of the web app is available at https://pathoman.mskcc.org/

Web –resources

1. CAVA – https://github.com/RahmanTeam/CAVA
2. SNPEff - http://snpeff.sourceforge.net/SnpEff_manual.html
3. Annovar - https:// annovar.openbioinformatics.org
4. ExACnoTCGA - http://exac.broadinstitute.org
5. gnomAD - http://gnomad.broadinstitute.org/
6. ClinVar - https://www.ncbi.nlm.nih.gov/clinvar/
7. IARC database - http://p53.iarc.fr/
8. ClinVar parser tool - https://github.com/macarthur-lab/clinvar
9. dbNSfP and dbscSNV - https://sites.google.com/site/jpopgen/dbNSFP
10. Gene List - https://github.com/macarthur-lab/gene_lists
11. Repeat masker - http://www.repeatmasker.org/
12. UCSC Genome Browser - https://genome.ucsc.edu
13. Cardio classifier - https://www.cardioclassifier.org/
14. InterVar - https://github.com/WGLab/InterVar
15. ACMG - https://www.acmg.net/
16. PathoMAN- http://pathoman.mskcc.org/

Acknowledgements

We thank Sabine Topka PhD and Zoe Steinsynder BS for helpful suggestions to improve the manuscript.

Funding/Support: Research reported in this pre-print was supported by National Cancer Insittute of the National Institutes of Health under award number R21CA029533 and R21CA178800 as well as Cycle for Survival, the Breast Cancer Research Foundation and The V Foundation for Cancer Research. It is also supported by the Cancer Center core grant P30CA008748 and The Robert and Kate Niehaus Center for Inherited Cancer Genomics. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health or other funding agencies.

References

1. Offit, K. (2017). Multigene Testing for Hereditary Cancer: When, Why, and How. J Natl Compr Canc Netw 15, 741-743.

2. Desmond, A., Kurian, A.W., Gabree, M., Mills, M.A., Anderson, M.J., Kobayashi, Y., Horick, N., Yang, S., Shannon, K.M., Tung, N., et al. (2015). Clinical Actionability of Multigene Panel Testing for Hereditary Breast and Ovarian Cancer Risk Assessment. JAMA Oncol 1, 943-951.

3. Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., Spector, E., et al. (2015). Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. Genet Med 17, 405-424.

4. Pandey, K.R., Maden, N., Poudel, B., Pradhananga, S., and Sharma, A.K. (2012). The curation of genetic variants: difficulties and possible solutions. Genomics Proteomics Bioinformatics 10, 317-325.

5. Mandelker, D., Zhang, L., Kemel, Y., Stadler, Z.K., Joseph, V., Zehir, A., Pradhan, N., Arnold, A., Walsh, M.F., Li, Y., et al. (2017). Mutation Detection in Patients With Advanced Cancer by Universal Sequencing of Cancer-Related Genes in Tumor and Normal DNA vs Guideline-Based Germline Testing. JAMA 318, 825-835.

6. Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., et al. (2017). Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. Nat Med 23, 703-713.

7. https://www.mskcc.org/press-releases/msk-impact-first-tumor-profiling-multiplex-panel-authorized-fda-setting-new-pathway-market-future-oncopanels

8. Pritchard, C.C., Mateo, J., Walsh, M.F., De Sarkar, N., Abida, W., Beltran, H., Garofalo, A., Gulati, R., Carreira, S., Eeles, R., et al. (2016). Inherited DNA-Repair Gene Mutations in Men with Metastatic Prostate Cancer. N Engl J Med 375, 443-453.

9. Maxwell, K.N., Hart, S.N., Vijai, J., Schrader, K.A., Slavin, T.P., Thomas, T., Wubbenhorst, B., Ravichandran, V., Moore, R.M., Hu, C., et al. (2016). Evaluation of ACMG-Guideline-Based Variant Classification of Cancer Susceptibility and Non-Cancer-Associated Genes in Families Affected by Breast Cancer. Am J Hum Genet 98, 801-817.

10. Landrum, M.J., Lee, J.M., Benson, M., Brown, G., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Hoover, J., et al. (2016). ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 44, D862-868.

11. Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res 42, D980-985.

12. Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., Cummings, B.B., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature 536, 285-291.

13. Munz, M., Ruark, E., Renwick, A., Ramsay, E., Clarke, M., Mahamdallie, S., Cloke, V., Seal, S., Strydom, A., Lunter, G., et al. (2015). CSN and CAVA: variant annotation tools for rapid, robust next-generation sequencing analysis in the clinical setting. Genome Med 7, 76.

14. Liu, X., Wu, C., Li, C., and Boerwinkle, E. (2016). dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. Hum Mutat 37, 235-241.

15. Liu, X., Jian, X., and Boerwinkle, E. (2013). dbNSFP v2.0: a database of human non-synonymous SNVs and their functional predictions and annotations. Hum Mutat 34, E2393-2402.

16. Liu, X., Jian, X., and Boerwinkle, E. (2011). dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Hum Mutat 32, 894-899.

17. Wang, K., Li, M., and Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. Nucleic Acids Res 38, e164.

18. Blekhman, R., Man, O., Herrmann, L., Boyko, A.R., Indap, A., Kosiol, C., Bustamante, C.D., Teshima, K.M., and Przeworski, M. (2008). Natural selection on genes that underlie human disease susceptibility. Curr Biol 18, 883-889.

19. Berg, J.S., Adams, M., Nassar, N., Bizon, C., Lee, K., Schmitt, C.P., Wilhelmsen, K.C., and Evans, J.P. (2013). An informatics approach to analyzing the incidentalome. Genet Med 15, 36-44.

20. Rahner, N., and Steinke, V. (2008). Hereditary cancer syndromes. Dtsch Arztebl Int 105, 706-714.

21. Cheng, D.T., Mitchell, T.N., Zehir, A., Shah, R.H., Benayed, R., Syed, A., Chandramohan, R., Liu, Z.Y., Won, H.H., Scott, S.N., et al. (2015). Memorial Sloan Kettering-Integrated Mutation Profiling of Actionable Cancer Targets (MSK-IMPACT): A Hybridization Capture-Based Next-Generation Sequencing Clinical Assay for Solid Tumor Molecular Oncology. J Mol Diagn 17, 251-264.

22. Cheng, D.T., Prasad, M., Chekaluk, Y., Benayed, R., Sadowska, J., Zehir, A., Syed, A., Wang, Y.E., Somar, J., Li, Y., et al. (2017). Comprehensive detection of germline variants by MSK-IMPACT, a clinical diagnostic platform for solid tumor molecular oncology and concurrent cancer predisposition testing. BMC Med Genomics 10, 33.

23. https://github.com/macarthur-lab/gene_lists

24. Thorstenson, Y.R., Shen, P., Tusher, V.G., Wayne, T.L., Davis, R.W., Chu, G., and Oefner, P.J. (2001). Global analysis of ATM polymorphism reveals significant functional constraint. Am J Hum Genet 69, 396-412.

25. https://www.ncbi.nlm.nih.gov/clinvar/docs/details/

26. Janatova, M., Kleibl, Z., Stribrna, J., Panczak, A., Vesela, K., Zimovjanova, M., Kleiblova, P., Dundr, P., Soukupova, J., and Pohlreich, P. (2013). The PALB2 gene is a strong candidate for clinical testing in BRCA1- and BRCA2-negative hereditary breast cancer. Cancer Epidemiol Biomarkers Prev 22, 2323-2332.

27. Southey, M.C., Goldgar, D.E., Winqvist, R., Pylkas, K., Couch, F., Tischkowitz, M., Foulkes, W.D., Dennis, J., Michailidou, K., van Rensburg, E.J., et al. (2016). PALB2, CHEK2 and ATM rare variants and cancer risk: data from COGS. J Med Genet 53, 800-811.

28. Bouaoun, L., Sonkin, D., Ardin, M., Hollstein, M., Byrnes, G., Zavadil, J., and Olivier, M. (2016). TP53 Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. Hum Mutat 37, 865-876.

29. Balmana, J., Digiovanni, L., Gaddam, P., Walsh, M.F., Joseph, V., Stadler, Z.K., Nathanson, K.L., Garber, J.E., Couch, F.J., Offit, K., et al. (2016). Conflicting Interpretation of Genetic Variants and Cancer Risk by Commercial Laboratories as Assessed by the Prospective Registry of Multiplex Testing. J Clin Oncol 34, 4071-4078.

30. Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., et al. (2004). UniProt: the Universal Protein knowledgebase. Nucleic Acids Res 32, D115-119.

31. Speir, M.L., Zweig, A.S., Rosenbloom, K.R., Raney, B.J., Paten, B., Nejad, P., Lee, B.T., Learned, K., Karolchik, D., Hinrichs, A.S., et al. (2016). The UCSC Genome Browser database: 2016 update. Nucleic Acids Res 44, D717-725.

32. Tarailo-Graovac, M., and Chen, N. (2009). Using RepeatMasker to identify repetitive elements in genomic sequences. Curr Protoc Bioinformatics Chapter 4, Unit 4 10.

33. Casper, J., Zweig, A.S., Villarreal, C., Tyner, C., Speir, M.L., Rosenbloom, K.R., Raney, B.J., Lee, C.M., Lee, B.T., Karolchik, D., et al. (2017). The UCSC Genome Browser database: 2018 update. Nucleic Acids Res.

34. Kent, W.J., Sugnet, C.W., Furey, T.S., Roskin, K.M., Pringle, T.H., Zahler, A.M., and Haussler, D. (2002). The human genome browser at UCSC. Genome Res 12, 996-1006.

35. Dong, C., Wei, P., Jian, X., Gibbs, R., Boerwinkle, E., Wang, K., and Liu, X. (2015). Comparison and integration of deleteriousness prediction methods for nonsynonymous SNVs in whole exome sequencing studies. Hum Mol Genet 24, 2125-2137.

36. Kircher, M., Witten, D.M., Jain, P., O'Roak, B.J., Cooper, G.M., and Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. Nat Genet 46, 310-315.

37. Shihab, H.A., Gough, J., Cooper, D.N., Day, I.N., and Gaunt, T.R. (2013). Predicting the functional consequences of cancer-associated amino acid substitutions. Bioinformatics 29, 1504-1510.

38. Chun, S., and Fay, J.C. (2009). Identification of deleterious mutations within three human genomes. Genome Res 19, 1553-1561.

39. Reva, B., Antipin, Y., and Sander, C. (2011). Predicting the functional impact of protein mutations: application to cancer genomics. Nucleic Acids Res 39, e118.

40. Schwarz, J.M., Rodelsperger, C., Schuelke, M., and Seelow, D. (2010). MutationTaster evaluates disease-causing potential of sequence alterations. Nat Methods 7, 575-576.

41. Choi, Y., and Chan, A.P. (2015). PROVEAN web server: a tool to predict the functional effect of amino acid substitutions and indels. Bioinformatics 31, 2745-2747.

42. Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Methods 7, 248-249.

43. Kumar, P., Henikoff, S., and Ng, P.C. (2009). Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. Nat Protoc 4, 1073-1081.

44. Carter, H., Douville, C., Stenson, P.D., Cooper, D.N., and Karchin, R. (2013). Identifying Mendelian disease genes with the variant effect scoring tool. BMC Genomics 14 Suppl 3, S3.

45. Jagadeesh, K.A., Wenger, A.M., Berger, M.J., Guturu, H., Stenson, P.D., Cooper, D.N., Bernstein, J.A., and Bejerano, G. (2016). M-CAP eliminates a majority of variants of uncertain significance in clinical exomes at high sensitivity. Nat Genet 48, 1581-1586.

46. Ghosh, R., Oak, N., and Plon, S.E. (2017). Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. Genome Biol 18, 225.

47. Jian, X., Boerwinkle, E., and Liu, X. (2014). In silico prediction of splice-altering single nucleotide variants in the human genome. Nucleic Acids Res 42, 13534-13544.

48. Jian, X., and Liu, X. (2017). In Silico Prediction of Deleteriousness for Nonsynonymous and Splice-Altering Single Nucleotide Variants in the Human Genome. Methods Mol Biol 1498, 191-197.

49. Waite, K.A., and Eng, C. (2002). Protean PTEN: form and function. Am J Hum Genet 70, 829-844.

50. Zerdoumi, Y., Aury-Landas, J., Bonaiti-Pellie, C., Derambure, C., Sesboue, R., Renaux-Petel, M., Frebourg, T., Bougeard, G., and Flaman, J.M. (2013). Drastic effect of germline TP53 missense mutations in Li-Fraumeni patients. Hum Mutat 34, 453-461.

51. Patel, R.Y., Shah, N., Jackson, A.R., Ghosh, R., Pawliczek, P., Paithankar, S., Baker, A., Riehle, K., Chen, H., Milosavljevic, S., et al. (2017). ClinGen Pathogenicity Calculator: a configurable system for assessing pathogenicity of genetic variants. Genome Med 9, 3.

52. Rehm, H.L., Berg, J.S., Brooks, L.D., Bustamante, C.D., Evans, J.P., Landrum, M.J., Ledbetter, D.H., Maglott, D.R., Martin, C.L., Nussbaum, R.L., et al. (2015). ClinGen--the Clinical Genome Resource. N Engl J Med 372, 2235-2242.

53. Zhang, X., Minikel, E.V., O'Donnell-Luria, A.H., MacArthur, D.G., Ware, J.S., and Weisburd, B. (2017). ClinVar data parsing. Wellcome Open Res 2, 33.

54. https://www.mskcc.org/blog/new-tumor-sequencing-test-will-bring-personalized-treatment-options-more-patients

55. Amendola, L.M., Jarvik, G.P., Leo, M.C., McLaughlin, H.M., Akkari, Y., Amaral, M.D., Berg, J.S., Biswas, S., Bowling, K.M., Conlin, L.K., et al. (2016). Performance of ACMG-AMP Variant-Interpretation Guidelines among Nine Laboratories in the Clinical Sequencing Exploratory Research Consortium. Am J Hum Genet 99, 247.

56. Harrison, S.M., Dolinsky, J.S., Knight Johnson, A.E., Pesaran, T., Azzariti, D.R., Bale, S., Chao, E.C., Das, S., Vincent, L., and Rehm, H.L. (2017). Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. Genet Med 19, 1096-1104.

57. Breast Cancer Association, C. (2006). Commonly studied single-nucleotide polymorphisms and breast cancer: results from the Breast Cancer Association Consortium. J Natl Cancer Inst 98, 1382-1396.

58. Hart, S.N., Maxwell, K.N., Thomas, T., Ravichandran, V., Wubberhorst, B., Klein, R.J., Schrader, K., Szabo, C., Weitzel, J.N., Neuhausen, S.L., et al. (2016). Collaborative science in the next-generation sequencing era: a viewpoint on how to combine exome sequencing data across sites to identify novel disease susceptibility genes. Brief Bioinform 17, 672-677.

59. Complexo, Southey, M.C., Park, D.J., Nguyen-Dumont, T., Campbell, I., Thompson, E., Trainer, A.H., Chenevix-Trench, G., Simard, J., Dumont, M., et al. (2013). COMPLEXO: identifying the missing heritability of breast cancer via next generation collaboration. Breast Cancer Res 15, 402.

60. Chenevix-Trench, G., Milne, R.L., Antoniou, A.C., Couch, F.J., Easton, D.F., Goldgar, D.E., and Cimba. (2007). An international initiative to identify genetic modifiers of cancer risk in BRCA1 and BRCA2 mutation carriers: the Consortium of Investigators of Modifiers of BRCA1 and BRCA2 (CIMBA). Breast Cancer Res 9, 104.

62. Stenson, P.D., Mort, M., Ball, E.V., Evans, K., Hayden, M., Heywood, S., Hussain, M., Phillips, A.D., and Cooper, D.N. (2017). The Human Gene Mutation Database: towards a comprehensive repository of inherited mutation data for medical research, genetic diagnosis and next-generation sequencing studies. Hum Genet 136, 665-677.

63. Cooper, D.N., and Krawczak, M. (1996). Human Gene Mutation Database. Hum Genet 98, 629.

64. Whiffin, N., Walsh, R., Govind, R., Edwards, M., Ahmad, M., Zhang, X., Tayal, U., Buchan, R., Midwinter, W., Wilk, A., et al. (2017). CardioClassifier: demonstrating the power of disease- and gene-specific computational decision support for clinical genome interpretation. bioRxiv.

65. Li, Q., and Wang, K. (2017). InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. Am J Hum Genet 100, 267-280.

66. Kelly, M.A., Caleshu, C., Morales, A., Buchan, J., Wolf, Z., Harrison, S.M., Cook, S., Dillon, M.W., Garcia, J., Haverfield, E., et al. (2018). Adaptation and validation of the ACMG/AMP variant classification framework for MYH7-associated inherited cardiomyopathies: recommendations by ClinGen's Inherited Cardiomyopathy Expert Panel. Genet Med.

67. Jarvik, G.P., and Browning, B.L. (2016). Consideration of Cosegregation in the Pathogenicity Classification of Genomic Variants. Am J Hum Genet 98, 1077-1081.

68. Nykamp, K., Anderson, M., Powers, M., Garcia, J., Herrera, B., Ho, Y.Y., Kobayashi, Y., Patil, N., Thusberg, J., Westbrook, M., et al. (2017). Sherloc: a comprehensive refinement of the ACMG-AMP variant classification criteria. Genet Med 19, 1105-1117.

69. Knudson, A.G., Jr. (1971). Mutation and cancer: statistical study of retinoblastoma. Proc Natl Acad Sci U S A 68, 820-823.

70. Olsen, J.H., Hahnemann, J.M., Borresen-Dale, A.L., Brondum-Nielsen, K., Hammarstrom, L., Kleinerman, R., Kaariainen, H., Lonnqvist, T., Sankila, R., Seersholm, N., et al. (2001). Cancer in patients with ataxia-telangiectasia and in their relatives in the nordic countries. J Natl Cancer Inst 93, 121-127.

71. http://oncotree.mskcc.org/oncotree/

72. Vijai, J., Topka, S., Villano, D., Ravichandran, V., Maxwell, K.N., Maria, A., Thomas, T., Gaddam, P., Lincoln, A., Kazzaz, S., et al. (2016). A Recurrent ERCC3 Truncating Mutation Confers Moderate Risk for Breast Cancer. Cancer Discov 6, 1267-1275.

73. Guidugli, L., Carreira, A., Caputo, S.M., Ehlen, A., Galli, A., Monteiro, A.N., Neuhausen, S.L., Hansen, T.V., Couch, F.J., Vreeswijk, M.P., et al. (2014). Functional assays for analysis of variants of uncertain significance in BRCA2. Hum Mutat 35, 151-164.

74. Spurdle, A.B., Healey, S., Devereau, A., Hogervorst, F.B., Monteiro, A.N., Nathanson, K.L., Radice, P., Stoppa-Lyonnet, D., Tavtigian, S., Wappenschmidt, B., et al. (2012). ENIGMA--evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. Hum Mutat 33, 2-7.

75. https://www.ga4gh.org/

76. Beroud, C., Letovsky, S.I., Braastad, C.D., Caputo, S.M., Beaudoux, O., Bignon, Y.J., Bressac-De Paillerets, B., Bronner, M., Buell, C.M., Collod-Beroud, G., et al. (2016). BRCA Share: A Collection of Clinical BRCA Gene Variants. Hum Mutat 37, 1318-1328.

77. Findlay, G.M., Boyle, E.A., Hause, R.J., Klein, J.C., and Shendure, J. (2014). Saturation editing of genomic regions by multiplex homology-directed repair. Nature 513, 120-123.

78. Starita, L.M., Young, D.L., Islam, M., Kitzman, J.O., Gullingsrud, J., Hause, R.J., Fowler, D.M., Parvin, J.D., Shendure, J., and Fields, S. (2015). Massively Parallel Functional Analysis of BRCA1 RING Domain Variants. Genetics 200, 413-422.

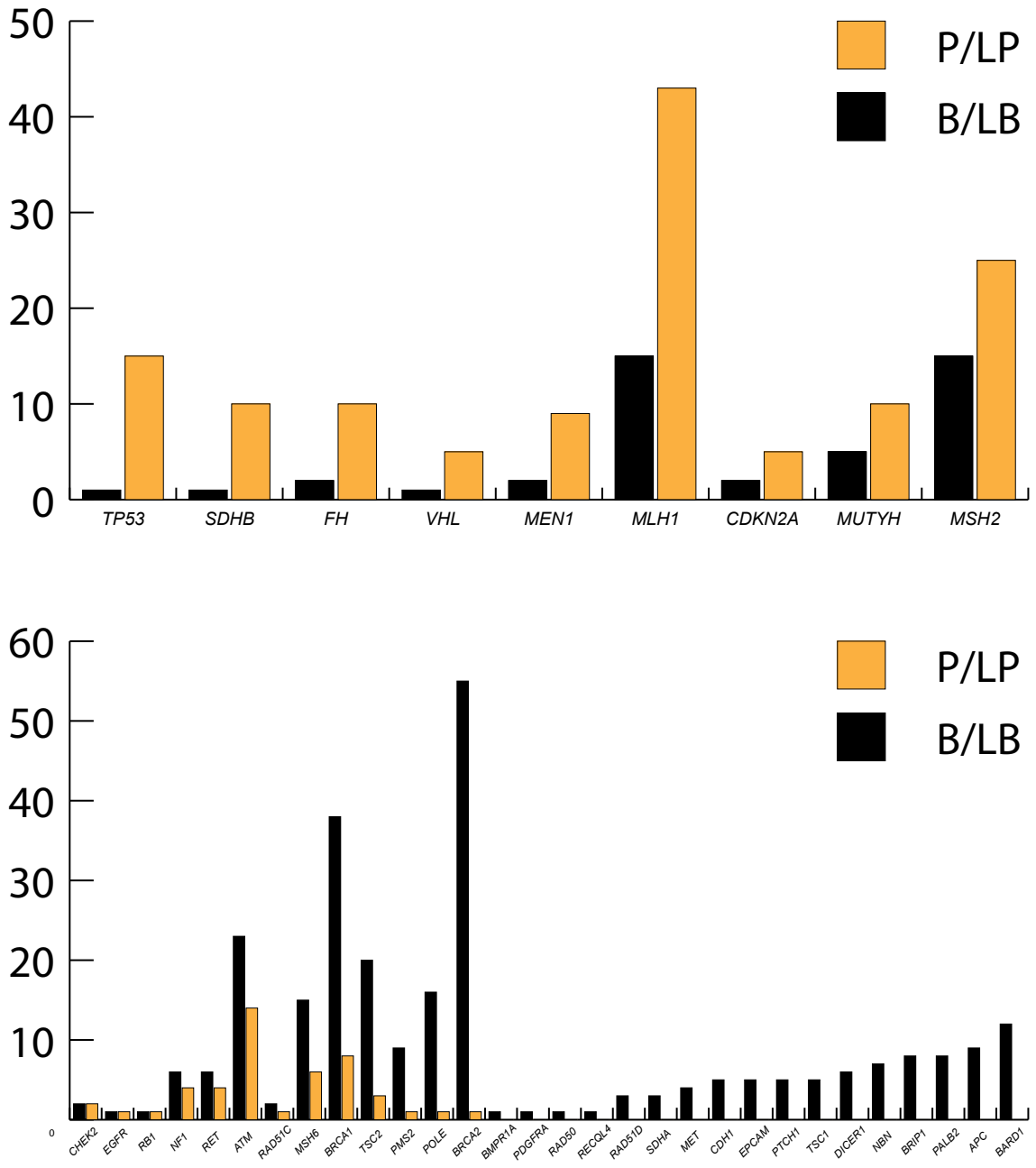
79. Starita, L.M., Ahituv, N., Dunham, M.J., Kitzman, J.O., Roth, F.P., Seelig, G., Shendure, J., and Fowler, D.M. (2017). Variant Interpretation: Functional Assays to the Rescue. Am J Hum Genet 101, 315-325.

Supplementary Information

GeneName	GeneName	GeneName	GeneName
ALK	*FLCN*	*NRAS*	*SDHAF2*
APC	*GATA2*	*PALB2*	*SDHB*
ATM	*GREM1*	*PAX5*	*SDHC*
BAP1	*HRAS*	*PDGFRA*	*SDHD*
BARD1	*JAK2*	*PHOX2B*	*SMAD3*
BLM	*KIT*	*PMS2*	*SMAD4*
BMPR1A	*KRAS*	*POLE*	*SMARCA4*
BRCA1	*MAX*	*PTCH1*	*SMARCB1*
BRCA2	*MEN1*	*PTEN*	*STK11*
BRIP1	*MET*	*RAD50*	*SUFU*
CDH1	*MITF*	*RAD51*	*TERT*
CDK4	*MLH1*	*RAD51B*	*TGFBR1*
CDKN2A	*MRE11A*	*RAD51C*	*TGFBR2*
CHEK2	*MSH2*	*RAD51D*	*TMEM127*
DICER1	*MSH6*	*RB1*	*TP53*
EGFR	*MUTYH*	*RECQL4*	*TSC1*
EPCAM	*NBN*	*RET*	*TSC2*
FAM175A	*NF1*	*RUNX1*	*VHL*
FH	*NF2*	*SDHA*	*WT1*

IMPACT-76 gene list: These are genes included in the IMPACT-76 gene list, where clinically actionable return of results is practiced at MSKCC

Supplementary Figure 1:



Supplementary Figure 1 (Top): Histogram of cancer related genes with high ratio of pathogenic missense variants from ClinVar; (Bottom): Histogram of cancer related genes with high ratio of benign missense variants from ClinVar.

Supplementary Figure 2
TP53: c.818G>A (p.Arg273His)

Variation ID: 12366
Review status: ★★☆☆ criteria provided, multiple submitters, no conf

Clinical significance: Pathogenic/Likely pathogenic
Last evaluated: Jul 14, 2017
Number of submission(s): 37
Condition(s):
- Adrenocortical carcinoma, hereditary [MedGen - OMIM]
- Familial cancer of breast [MedGen - Orphanet - OMIM]
- Glioma susceptibility 1 [MedGen - OMIM]
- Liver cancer
- Chronic lymphocytic leukemia [MeSH - MedGen - OMIM - Human Phenotype Ontology]
- Medulloblastoma [MeSH - MedGen - Orphanet - OMIM - Human Phenotype Ontology]
- Malignant melanoma of skin [MeSH - MedGen - OMIM]
- Multiple myeloma [MeSH - MedGen - Orphanet - OMIM - Human Phenotype Ontology]
- Osteosarcoma [MeSH - MedGen - Orphanet - OMIM - Human Phenotype Ontology]
- Squamous cell carcinoma of the head and neck [MeSH - MedGen - Orphane OMIM]
- Li-Fraumeni syndrome 1 [Gene - MedGen - OMIM]
- Nasopharyngeal carcinoma [Gene - MedGen - OMIM]
- Carcinoma of pancreas [MeSH - MedGen - Orphanet - OMIM]
- Choroid plexus papilloma [MedGen - Orphanet - OMIM - Human Phenotype Ontology]
- Small cell lung cancer [Gene - MeSH - MedGen - Orphanet - OMIM - Human Phenotype Ontology]
- Adenocarcinoma of lung [MeSH - MedGen - Human Phenotype Ontology]
- Thyroid cancer, anaplastic [MedGen - Orphanet - Human Phenotype Ontoloc
- Carcinoma of colon [MedGen - OMIM]
- Li-Fraumeni syndrome [MedGen - Orphanet - OMIM]

Submitter	Families	Individuals	Allele origin	Ethnicity	Geographic origin	Citations and Databases	Description
Total for all submitters	not provided	1	germline, somatic, unknown	not provided	not provided		
ARUP Laboratories, Molecular Genetics and Genomics	not provided	not provided	germline	not provided	not provided	not provided	not provided
Ambry Genetics	not provided	1	germline	not provided	not provided	PubMed	Lines of evidence used in supp…Full description
Counsyl	not provided	not provided	unknown	not provided	not provided	PubMed	not provided
Database of Curated Mutations (DoCM)	not provided	not provided	somatic	not provided	not provided	PubMed Other citation	not provided
Fulgent Genetics,Fulgent Genetics	not provided	not provided	unknown	not provided	not provided	not provided	not provided
GeneDx	not provided	not provided	germline	not provided	not provided	not provided	TP53 Arg273His has been report…Full description
Invitae	not provided	not provided	germline	not provided	not provided	not provided	This sequence change replaces …Full description

Full description for GeneDx

Germline

TP53 Arg273His has been reported in individuals with various types of Li Fraumeni-associated cancers, including adrenocortical carcinoma, choroid plexus carcinoma, sarcomas, gastric carcinoma, breast cancer, uterine serous cancer, and leukemia (Malkin 1992, Bemis 2007, Curry 2011, Masciari 2011, Melhem-Bertrandt 2012, Pennington 2013, Wasserman 2015, Schlegelberger 2015). This variant is reported as having non-functional transactivation in the International Agency for Research on Cancer TP53 database based on functional assays by Kato et al. (2003). Additionally, other in vitro-based functional assays have demonstrated that TP53 Arg273His results in severely deficient transactivation activity and exerted a dominant negative effect over wild-type p53 (Dong 2007, Malcikova 2010, Monti 2011, Wang 2013, Wasserman 2015). Since Arginine and Histidine share similar properties, this is considered a conservative amino acid substitution. TP53 Arg273His occurs at a position that is conserved across species and is located in the DNA binding domain (Bode 2004). Based on the current evidence, we consider this variant to be pathogenic.

Full description for Invitae

Germline

This sequence change replaces arginine with histidine at codon 273 of the TP53 protein (p.Arg273His). The arginine residue is highly conserved and there is a small physicochemical difference between arginine and histidine. This variant is present in population databases (rs28934576, ExAC 0.05%). This variant has been reported in numerous individuals and families affected with Li-Fraumeni syndrome (LFS) or Li-Fraumeni-associated cancers, and has been shown to segregate with disease in affected families (PMID: 9242456, 21484931, 17540308, 1565144, 20693561, 21552135). ClinVar contains an entry for this variant (Variation ID: 12366). This variant is located in the DNA-binding domain of the TP53 protein and is defined as a contact mutation that eliminates an essential DNA contact (PMID: 20516128). A mutant mouse model for this variant develops a variety of tumors and carcinomas by recapitulating LFS (PMID: 15607980). In addition, experimental studies have shown that this variant disrupts transcriptional activity in yeast-based assays (PMID: 12826609) and enhances cell proliferation, invasion, migration, and drug resistance in vitro (PMID: 17636407, 24677579). For these reasons, this variant has been classified as Pathogenic.

Germline

Lines of evidence used in support of classification: Detected in individual satisfying established diagnostic critera for classic disease without a clear mutation,Deficient protein function in appropriate functional assay(s),Well-characterized mutation at same position

Supplementary Figure 2 contiued..

BRCA1:c.5062_5064delGTT (p.Val1688del)

Variation ID: 55368

Review status: ★ criteria provide

Interpretation

Clinical significance: Conflicting interpretations of pathogenicity
Pathogenic(6);Uncertain significance(2)

Last evaluated: Aug 18, 2017

Number of submission(s): 8

Condition(s):
- Breast-ovarian cancer, familial 1 [MedGen - OMIM]
- Hereditary breast and ovarian cancer syndrome [MedGen - Orphanet - OMIM]
- Hereditary cancer-predisposing syndrome [MedGen]

Submitter	Families	Individuals	Allele origin	Ethnicity	Geographic origin	Citations and Databases	Description
Total for all submitters	34	9	germline, unknown	Caucasian; Western European	Italy		
Ambry Genetics	not provided	1	germline	not provided	not provided	• PubMed	Deficient protein function in ...Full description
Breast Cancer Information Core (BIC) (BRCA1)	not provided	8	germline	Caucasian; Western European	Italy	not provided	not provided
Color Genomics, Inc.	not provided	not provided	germline	not provided	not provided	not provided	not provided
Consortium of Investigators of Modifiers of BRCA1/2 (CIMBA), c/o University of Cambridge	34	not provided	germline	not provided	not provided	not provided	not provided

Full description for Ambry Genetics

Germline

Deficient protein function in appropriate functional assay(s),Good segregation with disease (lod 1.5-3 = 5-9 meioses),Detected in individual satisfying established diagnostic critera for classic disease without a clear mutation,Other strong data supporting pathogenic classification,Structural Evidence

Full description for Invitae

Germline

This sequence change deletes 3 nucleotides from exon 16 of the BRCA1 mRNA (c.5062_5064delGTT). This leads to the deletion of 1 amino acid residues in the BRCA1 protein (p.Val1688del) but otherwise preserves the integrity of the reading frame. This variant is not present in population databases (ExAC no frequency). This variant has been reported in multiple individuals and families affected with breast and/or ovarian cancer, with evidence of segregation with disease (PMID: 18165637, 19706752, 18821011, 8968102, 18703817, 23697973, 25814778, 26306726). ClinVar contains an entry for this variant (Variation ID: 55368). Experimental studies have shown that this in-frame deletion disrupts many aspects of the BRCA1 protein function, leading to defective DNA damage response and repair (PMID: 19706752, 23867111, 11157798). Based on a multifactorial likelihood algorithm using genetic, in silico, and statistical data, this variant has been determined to have a high probability of being pathogenic (PMID: 17924331, 21990134). For these reasons, this variant has been classified as Pathogenic.

Full description for GeneDx

Germline

This in-frame deletion of three nucleotides in BRCA1 is denoted c.5062_5064delGTT at the cDNA level and p.Val1688del (V1688del) at the protein level. Using alternate nomenclature, this variant would be defined as BRCA1 5181_5183delGTT or c.5062_5064del. The normal sequence, with the bases that are deleted in braces, is TGTT[GTT]ATGA. This deletion of a single Valine residue occurs at a position where amino acids with properties similar to Valine are tolerated across species and is located within the BRCT1 domain (Narod 2004). BRCA1 Val1688del has been observed in familial breast and/or ovarian cancer cases (Montagna 1996, Malacrida 2008, Tazzite 2012) and was strongly predicted by Lindor et al. (2012) to be pathogenic based on tumor pathology, clinical histories, family studies, and co-occurrence with deleterious variants. Functional studies by Bouwman et al. (2013) have suggested that BRCA1 Val1688del is pathogenic based on its inability to rescue the proliferation defect in BRCA1 deficient mouse embryonic stem cells and a statistically significant increase in sensitivity to cisplatin compared to controls. Additional functional assessments have demonstrated that BRCA1 Val1688del disrupts interactions with known partner proteins (De Nicolo 2009) and causes deficient transactivation activity (Vallon-Christersson 2001). We consider this variant to be pathogenic.