Pediatric Cancer Variant Pathogenicity Information Exchange (PeCanPIE): A Cloud-based Platform for Curating and Classifying Germline Variants

Michael N. Edmonson,[1,5] Aman N. Patel,[1,5] Dale J. Hedges,[1] Zhaoming Wang,[1] Evadnie Rampersaud,[1] Chimene A. Kesserwan,[2] Xin Zhou,[1] Yanling Liu,[1] Scott Newman,[1] Michael C. Rusch,[1] Clay L. McLeod,[1] Mark R. Wilkinson,[1] Stephen V. Rice,[1] Jared B. Becksfort,[1] Kim E. Nichols,[2] Leslie L. Robison,[3] James R. Downing,[4] and Jinghui Zhang[1]

[1]Department of Computational Biology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA; [2]Department of Oncology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA; [3]Department of Epidemiology & Cancer Control, St. Jude Children's Research Hospital, Memphis, TN 38105, USA; [4]Department of Pathology, St. Jude Children's Research Hospital, Memphis, TN 38105, USA

[5] These authors contributed equally to this work

* Corresponding author: Jinghui.Zhang@stjude.org

Running title: PeCanPIE: cloud-based variant classification

Keywords: germline, variant, cancer, pathogenicity, ACMG, classification, cloud

Abstract

Variant interpretation in the era of next-generation sequencing (NGS) is challenging. While many resources and guidelines are available to assist with this task, few integrated end-to-end tools exist. Here we present "PeCanPIE" – the Pediatric Cancer Variant Pathogenicity Information Exchange, a web- and cloud-based platform for annotation, identification, and classification of variations in known or putative disease genes. Starting from a set of variants in Variant Call Format (VCF), variants are annotated, ranked by putative pathogenicity, and presented for formal classification using a decision-support interface based on published guidelines from the American College of Medical Genetics and Genomics (ACMG). The system can accept files containing millions of variants and handle single-nucleotide variants (SNVs), simple insertions/deletions (indels), multiple-nucleotide variants (MNVs), and complex substitutions. PeCanPIE has been applied to classify variant pathogenicity in cancer predisposition genes in two large-scale investigations involving >4,000 pediatric cancer patients, and serves as a repository for the expert-reviewed results. While PeCanPIE's web-based interface was designed to be accessible to non-bioinformaticians, its back end pipelines may also be run independently on the cloud, facilitating direct integration and broader adoption. PeCanPIE is publicly available and free for research use.

Introduction

Next-generation sequencing (NGS) has quickly become a mainstay for genetic variation studies in many research and clinical genomics laboratories. However, the sheer abundance of data produced for a single individual means that complex and often tedious data processing and curation are required to identify potentially disease-causing mutations. The process is simultaneously burdened by the volume of novel variants, many of which have scarce information available, and the diverse, distributed nature of existing variant information resources. Variant annotation tools have been developed to assist with several aspects of this work, which can add coding and noncoding prediction annotations and population-specific allele frequencies, as well as provide filtering options for variant prioritization (Wang et al. 2010; Cingolani et al. 2012; Ng et al. 2009; McLaren et al. 2016). Likewise, variant curation tools supporting classification for clinical pathogenicity following the ACMG guidelines (Richards et al. 2015) have also been developed (Patel et al. 2017). While each resource offers valuable information to help researchers classify variant pathogenicity, integrated platforms are needed to provide support for all steps of the process, and streamline analysis of the thousands to millions of variants generated by NGS-based platforms. With these goals in mind, we developed “PeCanPIE” – the Pediatric Cancer Variant Pathogenicity Information Exchange – a cloud-based portal that provides an end-to-end workflow, beginning with a set of variants in VCF (Danecek et al. 2011) and ending with formal ACMG classification. PeCanPIE offers three key functions: 1) automated annotation, classification, and triage via our MedalCeremony pipeline (Zhang et al. 2015); 2) an interactive variant page and visualization tools to support expert curation and committee review; and 3) a reference database of expert-reviewed germline cancer-predisposing mutations.

Results

Process overview

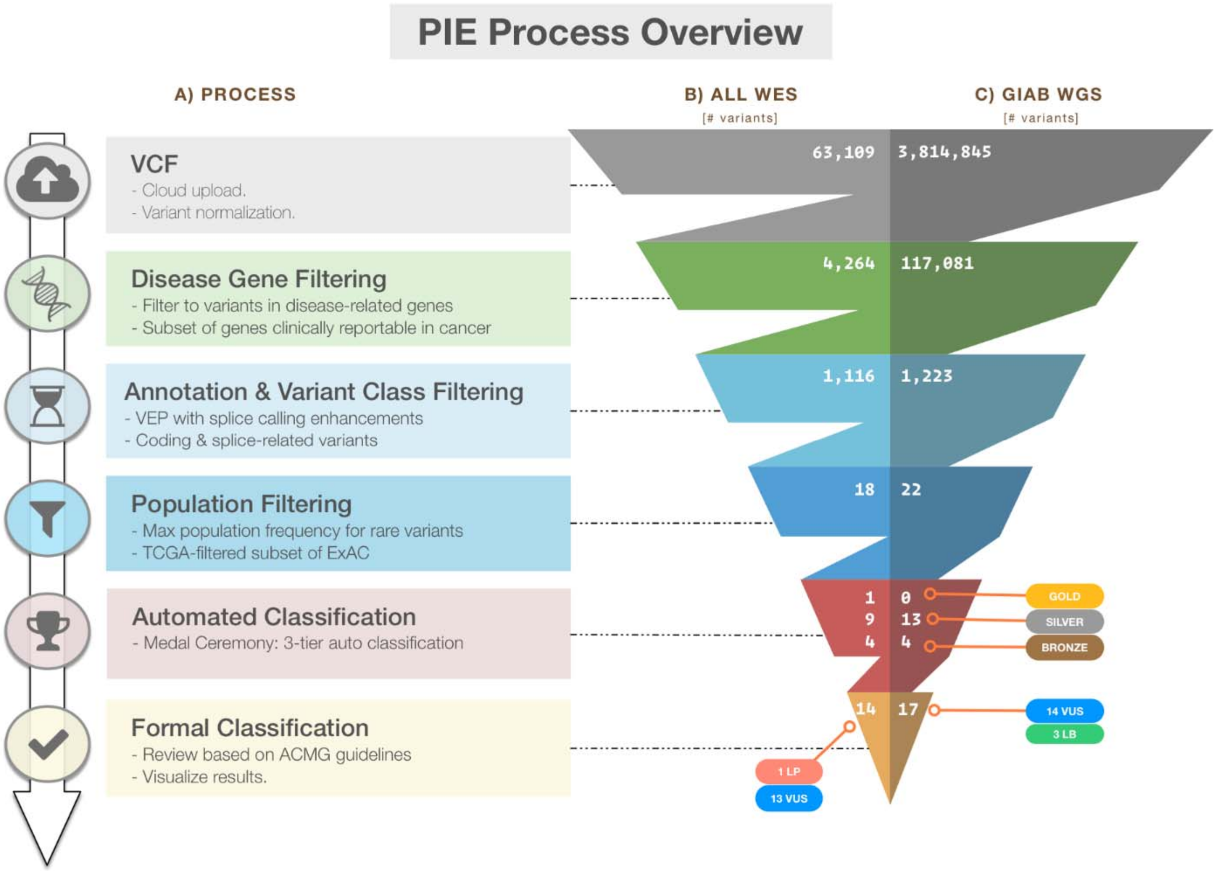

PIE Process Overview
A) PROCESS
B) ALL WES
[# variants]
C) GIAB WGS
[# variants]
VCF
- Cloud upload.
- Variant normalization.
Disease Gene Filtering
- Filter to variants in disease-related genes
- Subset of genes clinically reportable in cancer
Annotation & Variant Class Filtering
- VEP with splice calling enhancements
- Coding & splice-related variants
Population Filtering
- Max population frequency for rare variants
- TCGA-filtered subset of ExAC
Automated Classification
- Medal Ceremony: 3-tier auto classification
Formal Classification
- Review based on ACMG guidelines
- Visualize results.
63,109
3,814,845
4,264
117,081
1,116
1,223
18
22
1
9
4
0
13
4
GOLD
SILVER
BRONZE
14
17
14 VUS
3 LB
1 LP
13 VUS

Figure 1. Overview of variant classification using PeCanPIE. (A) Overview of processing steps from VCF through ACMG-based classification. Variant counts at each processing step for (B) whole-exome sequencing data generated from a germline sample of a patient with acute lymphoblastic leukemia (ALL), SJNORM015857_G1 (Methods) and (C) whole-genome sequencing data generated from Genome in a Bottle normal sample NA12878_HG001 (Methods).

As outlined in Fig. 1A, PeCanPIE launches with an interface for uploading a VCF file, which is then filtered to a set of disease-related genes (Methods, Table S1); users may alternatively specify their own list of genes of interest. Variants are next assigned gene and protein annotations and filtered by functional class and population frequency derived from the Exome Aggregation Consortium (ExAC) database (Lek et al. 2016). To ensure that pathogenic germline variants in cancer patients are retained, PeCanPIE uses the distribution of ExAC that excludes patient samples from The Cancer Genome Atlas (TCGA) (McLendon et al. 2008). The remaining variants are stratified into three tiers (gold, silver, and bronze) as an indication of potential pathogenicity computed by our MedalCeremony pipeline. Finally, each “medaled” variant is linked to a standalone page featuring an interface to support semi-automated pathogenicity classification using ACMG guidelines. Two examples in Fig. 1 demonstrate the classification process using VCF files generated from whole-exome sequencing (WES) of an acute lymphoblastic leukemia (ALL) patient (Moriyama et al. 2015) (Fig. 1B) and whole-genome sequencing (WGS) from the Genome in a Bottle (GiaB) project (Zook et al. 2014) (Fig. 1C), respectively. Only 14 of the 63,109 variants from the WES data and 17 of the approximately 4 million variants from the WGS data required expert review, which resulted in 1 and 0 pathogenic/likely pathogenic (P/LP) variants, respectively.

Automated classification by the MedalCeremony pipeline

Automated classification of variant pathogenicity implemented in the MedalCeremony pipeline classifies variants having a population frequency no higher than 0.001 (or a user-defined cutoff) in the ExAC database. Additional annotations are incorporated to aid with the classification process: 1) COSMIC (Forbes et al. 2008) hits; 2) functional annotations from dbNSFP (Liu et al. 2013) (protein domain and damage prediction algorithm calls); and 3) allele frequencies in the NHLBI GO Exome Sequencing Project (ESP), the Thousand Genomes Project (Auton et al. 2015), ExAC, and the Pediatric Cancer Genome Project (PCGP) (Downing et al. 2012).

An overview of the gold, silver, and bronze classification scheme implemented in MedalCeremony is shown in Fig. 2. Gold medals are assigned to truncating variants (including splice variants) in tumor suppressor genes (Zhao et al. 2016; Chakravarty et al. 2017), matches to highly-curated databases (IARC TP53 (Bouaoun et al. 2016), ClinVar expert-panel-reviewed pathogenic (P) or likely pathogenic (LP) variants, ASU TERT (Podlevsky et al. 2007), ARUP RET (Margraf et al. 2009), NHGRI Breast Cancer Information Core (Szabo et al. 2000), somatic mutation hotspots in COSMIC (observed in ≥10 tumors after removal of hypermutators) and PCGP, and St. Jude committee-reviewed germline P/LP variants. Silver medals are assigned to in-frame indels, truncation events in non-tumor-suppressor genes, variants predicted damaging by *in silico* algorithms, and matches to additional databases (ClinVar non-expert-panel P/LP, BRCA Share (Béroud et al. 2016), LOVD (Fokkema et al. 2011) locus-specific databases for APC and MSH2, and RB1 (Lohmann and Gallie 1993)). Unless otherwise medaled, variants predicted to be tolerated by *in silico* algorithms are assigned a bronze medal. Imperfect database matches (e.g., a different allele at the same genomic position or at the same codon but with a different amino acid change) are typically assigned a lower grade medal, e.g. silver rather than gold. Variants not meeting any of the previous criteria, e.g. most silent variants and those without any functional annotations, will not receive a medal. Amino acid and pathogenicity codes from the diverse variant databases used in this process are standardized to improve the reliability of annotations and utility of information (Methods). A summary of resources is shown in Table 1. MedalCeremony may also be run as a stand-alone pipeline on the St. Jude Cloud platform (Methods).

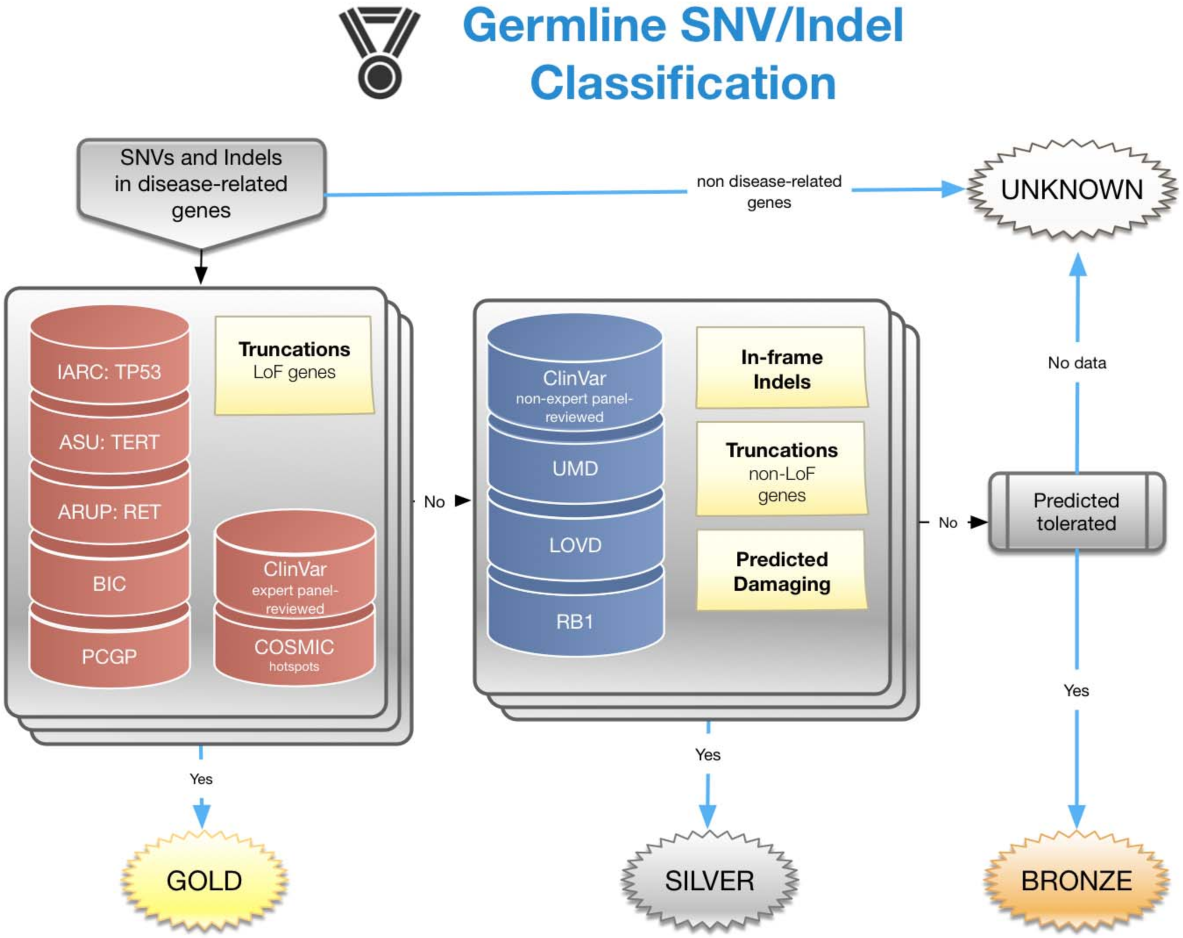

Germline SNV/Indel Classification
SNVs and Indels in disease-related genes
non disease-related genes
UNKNOWN
IARC: TP53
ASU: TERT
ARUP: RET
BIC
PCGP
Truncations LoF genes
ClinVar expert panel-reviewed
COSMIC hotspots
No
ClinVar non-expert panel-reviewed
UMD
LOVD
RB1
In-frame Indels
Truncations non-LoF genes
Predicted Damaging
No
Predicted tolerated
No data
Yes
Yes
Yes
GOLD
SILVER
BRONZE

Figure 2. Design of the MedalCeremony pipeline for automated germline variant classification. Truncating variants in loss-of-function genes (e.g. tumor suppressors) and those matching highly-curated databases receive gold medals. Truncations in non-loss-of-function genes, in-frame indels, predicted damaging variants, and matches to additional databases receive silver medals. Otherwise variants predicted to be tolerated by damage-prediction algorithms receive bronze. Imperfect database matches receive a lower-grade medal than exact matches. Variants not meeting any of the prior criteria receive a result of “unknown”.

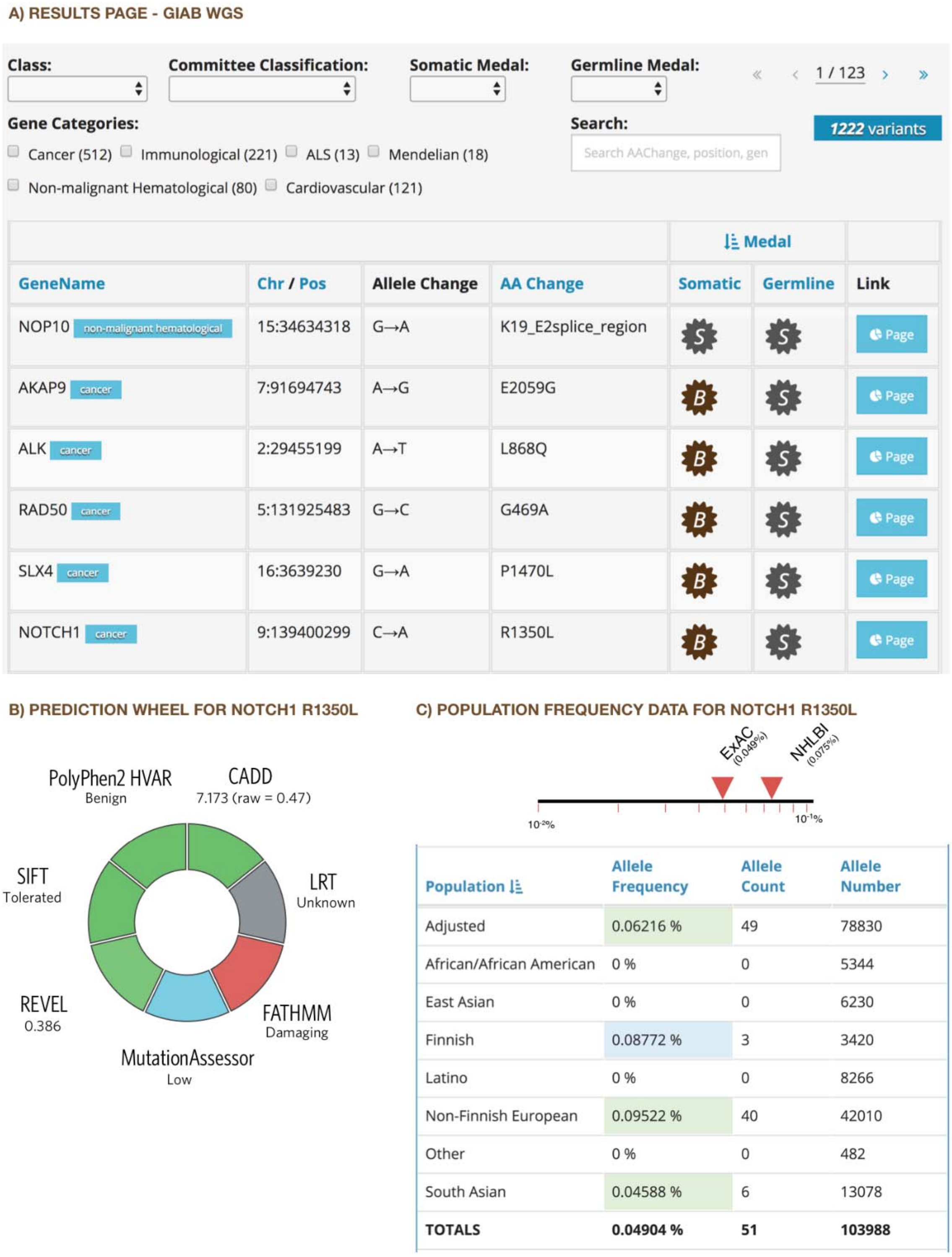
Table 1. Databases used in classification

Source	**URL**

ClinVar	http://www.ncbi.nlm.nih.gov/clinvar/
dbNSFP	https://sites.google.com/site/jpopgen/dbNSFP
ExAC	http://exac.broadinstitute.org/
COSMIC	https://cancer.sanger.ac.uk/cosmic/
IARC TP53	http://tp53.iarc.fr/
St. Jude PCGP	https://pecan.stjude.cloud/pcgp-explore
NHGRI BIC	http://research.nhgri.nih.gov/bic/
RB1	http://rb1-lovd.d-lohmann.de/
BRCA Share	http://www.umd.be/BRCA1/
ASU TERT	http://telomerase.asu.edu/diseases.html#tert
University of Utah RET	http://www.arup.utah.edu/database/MEN2/MEN2_display.php
LOVD APC, MSH2	http://chromium.liacs.nl/LOVD2/colon_cancer/

Variant review interface

After MedalCeremony classification, the results are presented in a table that can be searched or filtered by gene, variant class, medal status, or classification by expert review (Fig. 3A). If a variant has been previously classified by the user or the St. Jude germline variant review committee, that information will be pre-populated. Each row links to a variant page containing extensive annotations, including gene information from NCBI and OMIM (Amberger et al. 2015), ClinVar match details, population frequency, and *in silico* predictions of deleteriousness (Fig. 3B). The page also includes an embedded ProteinPaint view (Zhou et al. 2015), which overlays the current variant with aggregated somatic mutations and expert-classified P/LP germline variants on the protein product. This enables visual inspection of variant recurrence, hotspots, and enrichment of loss-of-function mutations.



A) RESULTS PAGE - GIAB WGS

Class: | Committee Classification: | Somatic Medal: | Germline Medal: | « ‹ 1 / 123 › »

Gene Categories:

□ Cancer (512) □ Immunological (221) □ ALS (13) □ Mendelian (18)

□ Non-malignant Hematological (80) □ Cardiovascular (121)

Search: Search AAChange, position, gen

1222 variants

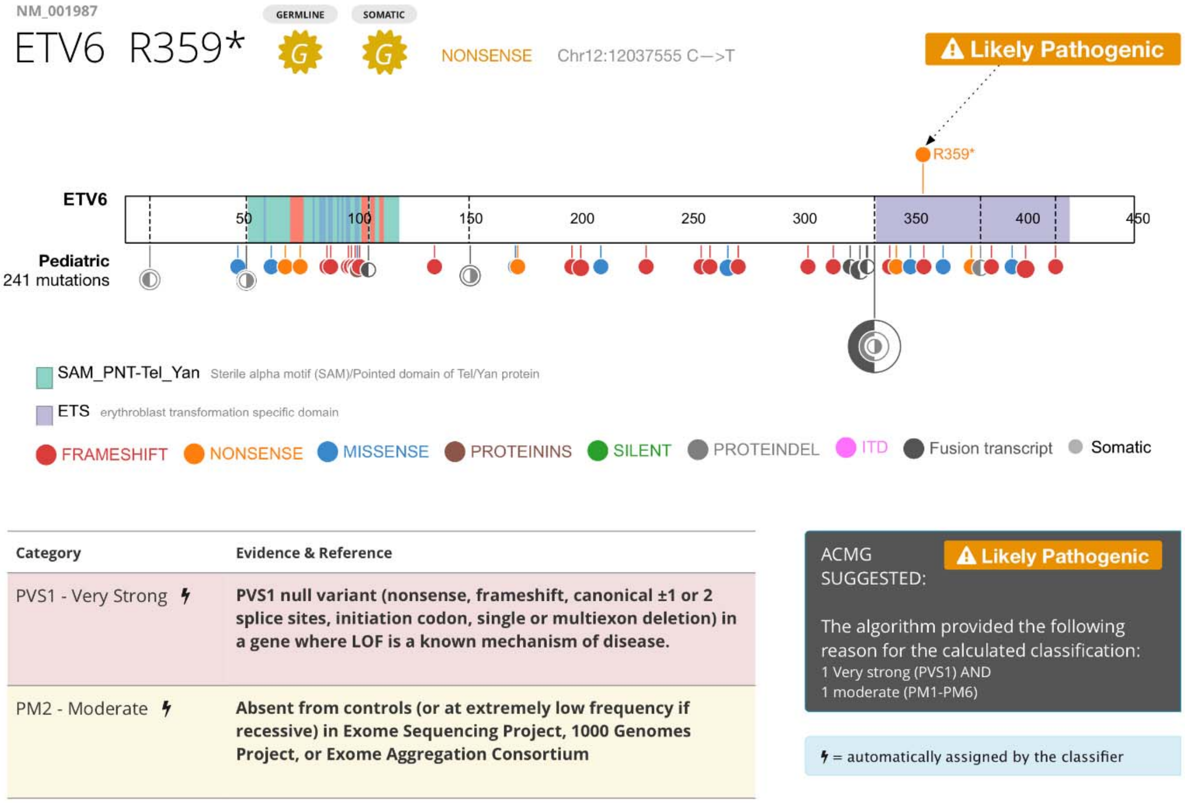
				Medal		
GeneName	Chr / Pos	Allele Change	AA Change	Somatic	Germline	Link
NOP10 non-malignant hematological	15:34634318	G→A	K19_E2splice_region	S	S	Page
AKAP9 cancer	7:91694743	A→G	E2059G	B	S	Page
ALK cancer	2:29455199	A→T	L868Q	B	S	Page
RAD50 cancer	5:131925483	G→C	G469A	B	S	Page
SLX4 cancer	16:3639230	G→A	P1470L	B	S	Page
NOTCH1 cancer	9:139400299	C→A	R1350L	B	S	Page

B) PREDICTION WHEEL FOR NOTCH1 R1350L

C) POPULATION FREQUENCY DATA FOR NOTCH1 R1350L

Population	Allele Frequency	Allele Count	Allele Number
Adjusted	0.06216 %	49	78830
African/African American	0 %	0	5344
East Asian	0 %	0	6230
Finnish	0.08772 %	3	3420
Latino	0 %	0	8266
Non-Finnish European	0.09522 %	40	42010
Other	0 %	0	482
South Asian	0.04588 %	6	13078
TOTALS	**0.04904 %**	**51**	**103988**

Figure 3. Annotation interface. Excerpts of PeCanPIE annotation interface. (A) Results for Genome in a Bottle WGS dataset. Variant page details for *NOTCH1* R1350L: (B) functional predictions, and (C) variant population frequency detail from ExAC ex-TCGA database.



Category	Evidence & Reference
PVS1 - Very Strong ϟ	**PVS1 null variant (nonsense, frameshift, canonical ±1 or 2 splice sites, initiation codon, single or multiexon deletion) in a gene where LOF is a known mechanism of disease.**
PM2 - Moderate ϟ	**Absent from controls (or at extremely low frequency if recessive) in Exome Sequencing Project, 1000 Genomes Project, or Exome Aggregation Consortium**

ACMG SUGGESTED: ⚠ Likely Pathogenic

The algorithm provided the following reason for the calculated classification:
1 Very strong (PVS1) AND
1 moderate (PM1-PM6)

ϟ = automatically assigned by the classifier

Figure 4. ACMG classification on ETV6. Top, ProteinPaint display of somatic *ETV6* variants across 11 subtypes of pediatric leukemia, showing enrichment of loss-of-function mutations (frameshifts in red, nonsense variants in orange). Arrow indicates position of germline R359* variant. Bottom, detail of PeCanPIE ACMG classification interface for R359* variant.

ACMG classification interface

A powerful feature of the variant detail page is an interactive graphical interface that allows a reviewer to enter a series of pathogenicity criteria evidence tags (e.g., population frequency, segregation, functional significance, and *in silico* prediction), along with supporting information such as PubMed IDs, to automatically calculate a 5-tier classification: Pathogenic (P), Likely Pathogenic (LP), Unknown Significance (VUS), Likely Benign (LB), and Benign (B) based on the ACMG algorithm. MedalCeremony can automatically generate ACMG classification tags for variants, which are prepopulated into PeCanPIE's classification interface. The following automatic tags are implemented: PVS1 (truncating variant in a tumor suppressor or other loss-of-function gene), PM1 (somatic hotspot in COSMIC), PM2 (absent from ExAC or appearing at a frequency of no greater than 0.0001) and the companion BA1 tag (>5% population frequency in ExAC), PM4 (in-frame protein insertions and deletions), PS1 and PM5 (amino acid comparisons made vs. pathogenic variants in ClinVar or those identified by the St. Jude Germline Review Committee). Automatically-assigned tags may be removed by the analyst if desired. This automation provides improved support versus manual curation interfaces, while still retaining analyst control over the ultimate classification decisions. As shown on the variant page for *ETV6* Arg359Ter, the single gold-medal variant detected in the patient with ALL was expert-classified as likely pathogenic because the mutation is present in a disease-related gene (i.e., *ETV6* is a pediatric ALL driver gene), is a loss-of-function null variant, and is not present in the ExAC database (Fig. 4).

Comparison of a germline variant with aggregated somatic variants can help inform germline classification for cancer predisposition genes. For example, family studies have identified a *PAX5* G183S germline mutation conferring susceptibility to B-ALL, which corresponds to somatic mutations detected in pediatric B-ALL and lymphoma (Shah et al. 2013). A similar profile was observed in the example WES data from an ALL patient presented in Fig. 1B: MedalCeremony assigned a single gold medal—a novel *ETV6* nonsense variant within the ETS domain (NM_001987.4:c.1075C>T, NP_001978.1:p.Arg359Ter)—based on the criteria of truncation in a tumor suppressor gene. The ProteinPaint view embedded in the variant page confirmed that in *ETV6*, somatic mutations are dominated by loss-of-function mutations across pediatric leukemia (Fig. 4), consistent with the tumor-suppressor gene model. Reviewers may enter custom evidence such as this into the interface for use during final classification.

Pathogenicity classification of cancer predisposition genes in 4,000 pediatric cancer patients

PeCanPIE was designed in support of large-scale germline variation analysis projects, and was iteratively improved based on the feedback of an interdisciplinary group of researchers. Germline variants from the following studies have been analyzed thus far: 1) a study of germline variations in predisposition genes in 1,120 children with cancer (Zhang et al. 2015) classified 890 variants, identifying 109 as pathogenic (P) and 25 as likely-pathogenic (LP); 2) the St. Jude LIFE project, a follow-up study of 3,006 long-term survivors of pediatric cancer (Wang et al. 2018), classified 3,417 variants, including 188 P and 160 LP; and 3) Genomes for Kids (manuscript in preparation), a clinical research study of 310 pediatric cancer patients (https://clinicaltrials.gov/ct2/show/NCT02530658), clinically reported 25 P and 6 LP variants. PeCanPIE also serves as a repository for expert-curated decisions for the first two studies, whose resulting annotations are reapplied to incoming variant classification requests.

Discussion

Although PeCanPIE's features partially overlap those of other available tools (Li and Wang 2017; Masica et al. 2017), it provides several new capabilities. Specifically, variant classification is tightly integrated with the rich resource of somatic mutation data in pediatric cancer, which can be explored online via the embedded ProteinPaint view. Users can also analyze indels, MNVs, and complex substitutions, whereas web-based implementations of similar tools may be limited to SNVs alone (Li and Wang 2017). Another key feature is the cloud-based implementation of PeCanPIE, which obviates the need for complex software installation and command-line workflows. This design also allows back end analysis pipelines to be invoked independently from PeCanPIE, for users who prefer direct or programmatic access over a graphical interface. In comparison with web-based systems (Masica et al. 2017) which provide batch annotation of variants based on machine-learning scores (Carter et al. 2013, 2009), PeCanPIE provides more granular annotations and individual ACMG-recommended evidence tags to facilitate interpretation of pathogenicity classifications. Via dbNSFP, PeCanPIE also provides access to REVEL (Ioannidis et al. 2016) pathogenicity scores, which fared well in a recent comparison of algorithms for use with ACMG clinical variant interpretation guidelines (Ghosh et al. 2017). Lastly, PeCanPIE's workflow offers advantages over CIVIC's crowdsourced clinical interpretation of variants (Ta 2017), which relies on completely manual classification and data entry, i.e., VCF upload, annotation, and prioritization are not provided.

A limitation of the existing method is that damage-prediction algorithm scores are taken from the dbNSFP database, which only contains data for non-silent SNVs. While these annotations are unavailable for indels, because protein class annotations are taken into account by the scoring algorithm, high-impact events such as truncating variations will still be highly ranked. For variant population frequency filtering, we are currently using the TCGA-subtracted release of ExAC instead of gnomAD (Lek et al. 2016) because the gnomAD database contains TCGA samples; we plan to migrate to gnomAD once a TCGA-subtracted version becomes publicly available.

In conclusion, the PeCanPIE platform significantly accelerates the variant classification process by automating many prerequisite steps, helping to prioritize potentially pathogenic variants in NGS data, and providing a robust platform for investigating variant pathogenicity in disease-related genes. While PeCanPIE was developed and tested with pediatric cancer susceptibility as a primary focus, we are in the process of expanding its scope to other pediatric and adult diseases. Users are now able to specify custom gene lists to analyze appropriate to their diseases of interest, enabling disease-specific variant curation and facilitating gene discovery.

Methods

Disease-related gene list

The disease-related gene list comprises both cancer-related and non-cancer genes (Table S1). The cancer gene list was compiled from public resources and cancer genetic studies including: 1) studies of germline mutations in predisposition genes in cancer patients (Zhang et al. 2015; Huang et al. 2018; Wang et al. 2018); 2) cancer predisposition genes compiled by Rahman (Rahman 2014); 3) the Cancer Gene Census (Futreal et al. 2004); and 4) driver genes identified in pediatric and adult pan-cancer studies (Ma et al. 2018; Gröbner et al. 2018). Publications were reviewed to confirm the presence of either loss-of-function or gain-of-function mutations in cancer driver genes, excluding those previously identified as having elevated mutation rates (e.g. *LRP1B* (Lawrence et al. 2013)) and those reported only as fusion partners. Other disease-related genes include non-malignant hematological, immunodeficiency, and amyotrophic lateral sclerosis (ALS)-related genes (Taylor et al. 2016), and genes from ACMG and Ambry Genetics incidental finding gene lists (Kalia et al. 2017). Filtering the variants to disease-related genes helps focus on areas with relevant research interest and reduce the downstream processing burden, which is especially helpful for WGS data which may contain 4-5 million variants per sample. A user may choose to focus on one or more of these pre-defined disease categories for expert review or provide their own gene lists for custom analysis.

Gene annotation and splice calling enhancement

Gene annotations are performed using the Ensembl Variant Effect Predictor (VEP) pipeline (McLaren et al. 2016), which provides information on a variant basis for the affected gene and transcript, functional class (e.g., silent, missense, and nonsense), and effect on protein coding. We enhanced splice variant annotation by reclassifying silent or missense variants at exon boundaries, which may impact splicing (e.g., *TP53* NM_000546.5:c.375G>A, NP_000537.3:p.Thr125Thr (Soudon et al. 1991)). While certainly not all of these variants will ultimately prove to be splice-related, these adjustments ensure additional scrutiny during expert review. A subsequent filtering step retains only variants in coding and splice-related regions. Silent variants are also kept because, in rare cases, they may cause aberrant splicing and thus be pathogenic. For example, ClinVar (Landrum et al. 2018) ID 90407 is a "silent" variant in the colon cancer predisposition gene *MLH1* (NM_000249.3:c.882C>T, NP_000240.1:p.Leu294=) that has been determined by an expert panel to be a pathogenic splice variant (Auclair et al. 2006). We refer to this enhanced pipeline as VEP+, which may also be run separately on the St. Jude Cloud platform.

St. Jude Cloud platform

While PeCanPIE was designed as a web portal to maximize ease of use for non-bioinformaticians, two component pipelines are also publicly accessible. On its back end, St. Jude Cloud (https://stjude.cloud) uses DNAnexus (https://www.dnanexus.com/), a platform where user-created software pipelines can be installed and run on cloud computing instances. A DNAnexus account is required to use PeCanPIE for secure storage and to send notifications when submitted jobs are complete. Once a pipeline has been installed on DNAnexus, it is straightforward for non-expert users to run it, either from a standardized web interface or a command-line client. We have created two DNAnexus pipelines that are used by PeCanPIE, VEP+ for variant annotation (app-stjude_vep_plus) and MedalCeremony for automated classification (app-stjude_medal_ceremony). The availability of these component pipelines on the cloud provides users and institutions straightforward, scalable access to the software, and our centralized maintenance allows all users to immediately benefit from updates and new features as they become available. PeCanPIE is free for non-commercial use.

Nomenclature standardization

We have observed that various variant databases which form the foundation of annotations for PeCanPIE vary in the structure and quality of variant specification. For example, databases may provide only protein-level annotations, only genomic annotations, or both. Likewise, there are many variations on the HGVS-like protein annotation nomenclature in circulation. The PeCanPIE code attempts to be flexible in parsing, standardizing, and formatting where possible, e.g. protein annotations may use either 3-character or 1-character protein codes (e.g. “Ser” or “S”), and a number of variations on stop codon formatting have been observed (“Ter”, “Term”, “*”, “X”, and “Stop”). In some cases partial information such as codon numbers were extracted from an otherwise incomplete annotation. Some databases also provide variations on the 5-tier ACMG pathogenicity calls which PeCanPIE attempts to standardize into B/LB/VUS/LP/P for easier comparison. We believe these standardizations further improve the reliability of annotations and utility of information provided by the PeCanPIE platform.

Example data

The ALL variants in Figure 1b were called from St. Jude sample SJNORM015857_G1. Variant calling was performed with Bambino using the “high 20” profile which consists of the following command-line parameters: “-min-quality 20 -min-flanking-quality 20 -min-alt-allele-count 3 -min-minor-frequency 0 -broad-min-quality 10 -mmf-max-hq-mismatches 4 -mmf-max-hq-mismatches-xt-u 10 -mmf-min-quality 15 -mmf-max-any-mismatches 6 -unique-filter-coverage 2 -no-strand-skew-filter”. The results were subsequently filtered to variants having a variant allele frequency of at least 20%, an average mapping quality of 20 for variant reads, at least 5 reads of coverage for the variant allele, bi-directional confirmation of the variant allele, and at least 20 reads of total coverage. The results were converted to VCF by an in-house script and uploaded to PeCanPIE. The Genome-in-a-Bottle VCF used for Figure 1c is available from ftp://ftp-trace.ncbi.nlm.nih.gov/giab/ftp/release/NA12878_HG001/NISTv3.3.2/GRCh37/HG001_GRCh37_GIAB_highconf_CG-IllFB-IllGATKHC-Ion-10X-SOLID_CHROM1-X_v.3.3.2_highconf_PGandRTGphasetransfer.vcf.gz. This bgzip-compressed VCF file may be used directly with PeCanPIE.

Software Availability

PeCanPIE is available at https://platform.stjude.cloud/tools/pecan_pie and is one component of the St. Jude Cloud platform (https://stjude.cloud/).

Acknowledgements

This project was supported by the American Lebanese Syrian Associated Charities (ALSAC) of St. Jude Children's Research Hospital, by a Cancer Center Support (Core) grant (CA21765) and a grant to JZ (CA21635) from the National Cancer Institute. We thank Yiyang Wu for discussions of *in silico* algorithms.

Author contributions: Analysis pipeline design and development (M.N.E.), web software design and development (A.N.P., X.Z., J.B.B.), cloud pipeline development (M.N.E., C.L.M.), tool development (M.N.E., S.V.R., M.C.R.), genomic data analysis (E.R., D.J.H., Y.L., C.A.K., J.Z., S.N., Z.W., L.L.R., A.N.P., M.N.E.), manuscript text (M.N.E., J.Z., E.R., C.A.K.), figure preparation (A.N.P., M.N.E., J.Z.), database support (M.R.W.), project direction and supervision (J.Z., J.R.D., K.E.N.)

References

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM®), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**: D789–D798. http://www.ncbi.nlm.nih.gov/pubmed/25428349 (Accessed May 25, 2018).

Auclair J, Busine MP, Navarro C, Ruano E, Montmain G, Desseigne F, Saurin JC, Lasset C, Bonadona V, Giraud S, et al. 2006. Systematic mRNA analysis for the effect ofMLH1 andMSH2 missense and silent mutations on aberrant splicing. *Hum Mutat* **27**: 145–154. http://www.ncbi.nlm.nih.gov/pubmed/16395668 (Accessed April 3, 2018).

Auton A, Abecasis GR, Altshuler DM, Durbin RM, Abecasis GR, Bentley DR, Chakravarti A, Clark AG, Donnelly P, Eichler EE, et al. 2015. A global reference for human genetic variation. *Nature* **526**: 68–74. http://www.ncbi.nlm.nih.gov/pubmed/26432245 (Accessed April 16, 2018).

Béroud C, Letovsky SI, Braastad CD, Caputo SM, Beaudoux O, Bignon YJ, Bressac-De Paillerets B, Bronner M, Buell CM, Collod-Béroud G, et al. 2016. BRCA Share: A Collection of Clinical BRCA Gene Variants. *Hum Mutat* **37**: 1318–1328. http://www.ncbi.nlm.nih.gov/pubmed/27633797 (Accessed May 23, 2018).

Bouaoun L, Sonkin D, Ardin M, Hollstein M, Byrnes G, Zavadil J, Olivier M. 2016. *TP53* Variations in Human Cancers: New Lessons from the IARC TP53 Database and Genomics Data. *Hum Mutat* **37**: 865–876. http://www.ncbi.nlm.nih.gov/pubmed/27328919 (Accessed April 3, 2018).

Carter H, Chen S, Isik L, Tyekucheva S, Velculescu VE, Kinzler KW, Vogelstein B, Karchin R. 2009. Cancer-Specific High-Throughput Annotation of Somatic Mutations: Computational

Prediction of Driver Missense Mutations. *Cancer Res* **69**: 6660–6667. http://www.ncbi.nlm.nih.gov/pubmed/19654296 (Accessed April 2, 2018).

Carter H, Douville C, Stenson PD, Cooper DN, Karchin R. 2013. Identifying Mendelian disease genes with the Variant Effect Scoring Tool. *BMC Genomics* **14**: S3. http://www.ncbi.nlm.nih.gov/pubmed/23819870 (Accessed April 2, 2018).

Chakravarty D, Gao J, Phillips SM, Kundra R, Zhang H, Wang J, Rudolph JE, Yaeger R, Soumerai T, Nissan MH, et al. 2017. OncoKB: A Precision Oncology Knowledge Base. *JCO Precis Oncol* **2017**. http://www.ncbi.nlm.nih.gov/pubmed/28890946 (Accessed May 23, 2018).

Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, Land SJ, Lu X, Ruden DM. 2012. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly (Austin)* **6**: 80–92. http://www.ncbi.nlm.nih.gov/pubmed/22728672 (Accessed March 30, 2018).

Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, Handsaker RE, Lunter G, Marth GT, Sherry ST, et al. 2011. The variant call format and VCFtools. *Bioinformatics* **27**: 2156–2158. http://www.ncbi.nlm.nih.gov/pubmed/21653522 (Accessed May 17, 2018).

Downing JR, Wilson RK, Zhang J, Mardis ER, Pui C-H, Ding L, Ley TJ, Evans WE. 2012. The Pediatric Cancer Genome Project. *Nat Genet* **44**: 619–622. http://www.ncbi.nlm.nih.gov/pubmed/22641210 (Accessed March 30, 2018).

Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT. 2011. LOVD v.2.0: the next generation in gene variant databases. *Hum Mutat* **32**: 557–563. http://www.ncbi.nlm.nih.gov/pubmed/21520333 (Accessed May 23, 2018).

Forbes SA, Bhamra G, Bamford S, Dawson E, Kok C, Clements J, Menzies A, Teague JW, Futreal PA, Stratton MR. 2008. The Catalogue of Somatic Mutations in Cancer (COSMIC). In *Current Protocols in Human Genetics*, Vol. Chapter 10 of, p. Unit 10.11, John Wiley & Sons, Inc., Hoboken, NJ, USA http://www.ncbi.nlm.nih.gov/pubmed/18428421 (Accessed April 11, 2018).

Futreal PA, Coin L, Marshall M, Down T, Hubbard T, Wooster R, Rahman N, Stratton MR. 2004. A census of human cancer genes. *Nat Rev Cancer* **4**: 177–183. http://www.nature.com/articles/nrc1299 (Accessed April 19, 2018).

Ghosh R, Oak N, Plon SE. 2017. Evaluation of in silico algorithms for use with ACMG/AMP clinical variant interpretation guidelines. *Genome Biol* **18**: 225. http://www.ncbi.nlm.nih.gov/pubmed/29179779 (Accessed March 27, 2018).

Gröbner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, Johann PD, Balasubramanian GP, Segura-Wang M, Brabetz S, et al. 2018. The landscape of genomic alterations across childhood cancers. *Nature* **555**: 321–327. http://www.ncbi.nlm.nih.gov/pubmed/29489754 (Accessed April 23, 2018).

Huang K, Mashl RJ, Wu Y, Ritter DI, Wang J, Oh C, Paczkowska M, Reynolds S, Wyczalkowski MA, Oak N, et al. 2018. Pathogenic Germline Variants in 10,389 Adult Cancers. *Cell* **173**: 355–370.e14. http://linkinghub.elsevier.com/retrieve/pii/S0092867418303635 (Accessed April 19, 2018).

Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, Musolf A, Li Q, Holzinger E, Karyadi D, et al. 2016. REVEL: An Ensemble Method for Predicting the Pathogenicity of Rare Missense Variants. *Am J Hum Genet* **99**: 877–885. http://www.ncbi.nlm.nih.gov/pubmed/27666373 (Accessed April 2, 2018).

Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, Herman GE, Hufnagel SB, Klein TE, Korf BR, et al. 2017. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet Med* **19**: 249–255. http://www.ncbi.nlm.nih.gov/pubmed/27854360 (Accessed May 7, 2018).

Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, Gu B, Hart J, Hoffman D, Jang W, et al. 2018. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**: D1062–D1067. http://www.ncbi.nlm.nih.gov/pubmed/29165669 (Accessed April 16, 2018).

Lawrence MS, Stojanov P, Polak P, Kryukov G V., Cibulskis K, Sivachenko A, Carter SL, Stewart C, Mermel CH, Roberts SA, et al. 2013. Mutational heterogeneity in cancer and the search for new cancer-associated genes. *Nature* **499**: 214–218.

Lek M, Karczewski KJ, Minikel E V., Samocha KE, Banks E, Fennell T, O'Donnell-Luria AH, Ware JS, Hill AJ, Cummings BB, et al. 2016. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**: 285–291. http://www.nature.com/articles/nature19057 (Accessed March 27, 2018).

Li Q, Wang K. 2017. InterVar: Clinical Interpretation of Genetic Variants by the 2015 ACMG-AMP Guidelines. *Am J Hum Genet* **100**: 267–280. http://www.ncbi.nlm.nih.gov/pubmed/28132688 (Accessed April 2, 2018).

Liu X, Jian X, Boerwinkle E. 2013. dbNSFP v2.0: A Database of Human Non-synonymous SNVs and Their Functional Predictions and Annotations. *Hum Mutat* **34**: E2393–E2402. http://doi.wiley.com/10.1002/humu.22376 (Accessed March 27, 2018).

Lohmann DR, Gallie BL. 1993. *Retinoblastoma*. http://www.ncbi.nlm.nih.gov/pubmed/20301625 (Accessed May 21, 2018).

Ma X, Liu Y, Liu Y, Alexandrov LB, Edmonson MN, Gawad C, Zhou X, Li Y, Rusch MC, Easton

J, et al. 2018. Pan-cancer genome and transcriptome analyses of 1,699 paediatric leukaemias and solid tumours. *Nature* **555**: 371–376. http://www.nature.com/doifinder/10.1038/nature25795 (Accessed March 27, 2018).

Margraf RL, Crockett DK, Krautscheid PMF, Seamons R, Calderon FRO, Wittwer CT, Mao R. 2009. Multiple endocrine neoplasia type 2 *RET* protooncogene database: Repository of MEN2-associated *RET* sequence variation and reference for genotype/phenotype correlations. *Hum Mutat* **30**: 548–556. http://www.ncbi.nlm.nih.gov/pubmed/19177457 (Accessed May 18, 2018).

Masica DL, Douville C, Tokheim C, Bhattacharya R, Kim R, Moad K, Ryan MC, Karchin R. 2017. CRAVAT 4: Cancer-Related Analysis of Variants Toolkit. *Cancer Res* **77**: e35–e38. http://www.ncbi.nlm.nih.gov/pubmed/29092935 (Accessed April 2, 2018).

McLaren W, Gil L, Hunt SE, Riat HS, Ritchie GRS, Thormann A, Flicek P, Cunningham F. 2016. The Ensembl Variant Effect Predictor. *Genome Biol* **17**: 122. http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-0974-4 (Accessed March 27, 2018).

McLendon R, Friedman A, Bigner D, Van Meir EG, Brat DJ, M. Mastrogianakis G, Olson JJ, Mikkelsen T, Lehman N, Aldape K, et al. 2008. Comprehensive genomic characterization defines human glioblastoma genes and core pathways. *Nature* **455**: 1061–1068. http://www.ncbi.nlm.nih.gov/pubmed/18772890 (Accessed May 18, 2018).

Moriyama T, Metzger ML, Wu G, Nishii R, Qian M, Devidas M, Yang W, Cheng C, Cao X, Quinn E, et al. 2015. Germline genetic variation in ETV6 and risk of childhood acute lymphoblastic leukaemia: a systematic genetic study. *Lancet Oncol* **16**: 1659–1666. http://www.ncbi.nlm.nih.gov/pubmed/26522332 (Accessed March 27, 2018).

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, et al. 2009. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**: 272–276. http://www.ncbi.nlm.nih.gov/pubmed/19684571 (Accessed March 30, 2018).

Patel RY, Shah N, Jackson AR, 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*.

Podlevsky JD, Bley CJ, Omana R V., Qi X, Chen JJ-L. 2007. The Telomerase Database. *Nucleic Acids Res* **36**: D339–D343. http://www.ncbi.nlm.nih.gov/pubmed/18073191 (Accessed May 18, 2018).

Rahman N. 2014. Realizing the promise of cancer predisposition genes. *Nature*.

Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, Grody WW, 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*.

Shah S, Schrader KA, Waanders E, Timms AE, Vijai J, Miething C, Wechsler J, Yang J, Hayes J, Klein RJ, et al. 2013. A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. *Nat Genet* **45**: 1226–1231. http://www.ncbi.nlm.nih.gov/pubmed/24013638 (Accessed May 6, 2018).

Soudon J, Caron de Fromentel C, Bernard O, Larsen CJ. 1991. Inactivation of the p53 gene expression by a splice donor site mutation in a human T-cell leukemia cell line. *Leukemia* **5**: 917–20. http://www.ncbi.nlm.nih.gov/pubmed/1961027 (Accessed March 27, 2018).

Szabo C, Masiello A, Ryan JF, Brody LC. 2000. The Breast Cancer Information Core: Database design, structure, and scope. *Hum Mutat* **16**: 123–131.

http://www.ncbi.nlm.nih.gov/pubmed/10923033 (Accessed April 4, 2018).

Ta EN. 2017. CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer. *Nat Publ Gr* **49**.

Taylor JP, Brown RH, Cleveland DW. 2016. Decoding ALS: from genes to mechanism. *Nature* **539**: 197–206. http://www.ncbi.nlm.nih.gov/pubmed/27830784 (Accessed May 3, 2018).

Wang K, Li M, Hakonarson H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38**: e164. http://www.ncbi.nlm.nih.gov/pubmed/20601685 (Accessed March 27, 2018).

Wang Z, Wilson CL, Easton J, Thrasher A, Mulder H, Liu Q, Hedges D, Wang S, Rusch M, Edmonson M, et al. 2018. Genetic Risk for Subsequent Neoplasms among Long-term Survivors of Childhood Cancer. *J Clin Oncol*.

Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, Hedges D, Ma X, Zhou X, Yergeau DA, et al. 2015. Germline Mutations in Predisposition Genes in Pediatric Cancer. *N Engl J Med*.

Zhao M, Kim P, Mitra R, Zhao J, Zhao Z. 2016. TSGene 2.0: an updated literature-based knowledgebase for tumor suppressor genes. *Nucleic Acids Res* **44**: D1023–D1031. http://www.ncbi.nlm.nih.gov/pubmed/26590405 (Accessed May 23, 2018).

Zhou X, Edmonson MN, Wilkinson MR, Patel A, Wu G, Liu Y, Li Y, Zhang Z, Rusch MC, Parker M, et al. 2015. Exploring genomic alteration in pediatric cancer using ProteinPaint. *Nat Genet* **48**.

Zook JM, Chapman B, Wang J, Mittelman D, Hofmann O, Hide W, Salit M. 2014. Integrating human sequence data sets provides a resource of benchmark SNP and indel genotype calls. *Nat Biotechnol* **32**: 246–251. http://www.nature.com/articles/nbt.2835 (Accessed

March 27, 2018).