Supplementary Appendix

This appendix has been provided by the authors to give readers additional information about their work.

Supplement to: Zhang J, Walsh MF, Wu G, et al. Germline mutations in predisposition genes in pediatric cancer. N Engl J Med 2015;373:2336-46. DOI: 10.1056/NEJMoa1508054

(Updated November 23, 2015.)

Supplementary Appendix 1

for

Germline Mutations in Predisposition Genes in Pediatric Cancer.

Jinghui Zhang[1,2*], Michael F. Walsh[2, 3*], Gang Wu[1,2*], Michael N. Edmonson[1,2], Tanja A. Gruber[2,3,4], John Easton[2], Dale Hedges[1], Xiaotu Ma[1,2], Xin Zhou[1], Donald A. Yergeau[2], Mark R. Wilkinson[1,4], Bhavin Vadodaria[2], Xiang Chen[1,2], Rose B. McGee[2,3], Stacy Hines-Dowell[2,3], Regina Nuccio[2,3], Emily Quinn [2,3], Sheila A. Shurtleff[2,4], Michael Rusch[1,2], Aman Patel[1,2], Jared B. Becksfort[1,2], Shuoguo Wang[1,2], Meaghann S. Weaver[2], Li Ding[5,6], Elaine R Mardis[5,6], Richard K Wilson[5,6], Amar Gajjar[2,3], David W. Ellison[2,4], APBerto S. Pappo[2,3], Ching-Hon Pui[2,3], Kim E. Nichols[2,3*] and James R. Downing[2,4]

Table of Contents

STUDY OVERSIGHT 5

SUPPLEMENTARY METHODS 5

Sample preparation and sequencing 5

Validation of germline variants 6

Analysis of tumor-in-normal contamination and germline mosaicism 7

Automated germline classification 8

Curation of somatic mutation databases 10

SUPPLEMENTARY RESULTS 12

Analysis of whole-exome sequencing data from 1000 Genome project 12

Analysis of NDAR autism genotype data 14

Significance of *BRCA1*, *BRCA2* and *PALB2* mutations in pediatric caner 16

Mutations in 29 genes associated with autosomal recessive cancer predisposition syndromes 17

Mutations in tumor suppressors, tyrosine kinases and other cancer genes 18

Assessment of enrichment of protein truncation mutations in PCGP patients 18

Estimation of germline mutation prevalence based on cancer subtype distribution in SEER 20

SUPPLEMENTARY FIGURES 22

Figure S1 Variant detection, analysis and classification pipeline 22

Figure S2 Germline copy number variations in *TP53*, *KRAS* and *PMS2* 23

Figure S3 Distribution of pathogenic or probably pathogenic germline mutations. 24

Figure S4. Compound heterozygous germline mutations of *ATM* causing loss of function of both copies of this gene in HGG027 25

Figure S5. A zoomed-in ProteinPaint view of *TP53* germline mutations on web portal. 26

Figure S6. Screenshot of ProteinPaint view of germline mutations in representative genes in each of the five categories. 27

Figure S7. Analyis of family history. A) 95 patients with pathogenic or probably pathogenic germline mutations. B) 100 patiens randomly selected from the remaining 1,025 patients. 33

Figure S8. Family history of two cases with *BRCA2* mutation 34

Figure S9. Genes sets assessed in this study. 35

SUPPLEMENTARY TABLES 36

Table S1. Clinical information on patients examined in this study 36

Table S2. Categories and information about the 565 cancer genes selected for germ line mutation analysis. 36

Table S3. Sequence coverage of the 565 genes analyzed. 36

Table S4. Germline mutations discovered in PCGP cohort. 36

Table S5. Germline mutation summary in PCGP cohort.....37

Table S6. Pathogenic (P) or probably pathogenic (PP) mutations identified in the 966 un-related individuals from 1000 Genome Project 38

Table S7. Pathogenic (P) or Probably Pathogenic (PP) mutations found in the NDAR autism cohort including 515 cases and 208 controls 40

Table S8. Twelve cases with multiple germline pathogenic/probably pathogenic or truncation mutations.....42

Table S9. Enrichment of truncation mutations in PCGP (n=1,120) compared to The Exome Aggregation Consortium (ExAC, n=60,706). 43

Table S10. Estimate of prevalence of germline cancer predisposition in SEER cohort.....46

SUPPLEMENTARY REFERENCES.....47

STUDY OVERSIGHT

The manuscript was written by JZ, MFW, GW, MNE, CHP, KEN and JRD; the final version of the manuscript incorporated changes from coauthors. All authors vouch for the accuracy and completeness of the data and analyses presented and for the adherence of the study to the protocol. No commercial support was involved in the planning or execution of the study

SUPPLEMENTARY METHODS

Sample preparation and sequencing

The cohort is comprised of 836 patients from St. Jude Children's Research Hospital, 127 from the Children's Oncology Group, 97 from Memorial Sloan Kettering Cancer Center, and 60 from collaborating centers in France, Brazil, Singapore, Taiwan, Toronto, Italy, Japan, and Australia. Nucleic acid extraction, library preparation, and sequencing have been previously described. [1-3] In brief, after tissue homogenization and cell lysis, DNA was isolated using phenol chloroform extraction followed by EtOH precipitation. Exome sequencing data was acquired using 1-2ug of genomic DNA with the Illumina Truseq Expanded Exome kit. Whole genome library preparation and sequencing from an input of 500ng-2ug of genomic DNA was performed. For transcriptome sequencing, 2ug of Trizol extracted total RNA was mRNA enriched using oligo dT beads. The resulting mRNA was used with the Superscript II Double Stranded cDNA synthesis kit to perform random primed cDNA and second strand DNA synthesis. After fragmentation of the DNA to approximately 200bp with advanced focused acoustics (Covaris), library construction and amplification was performed using the Illumina TRUSEQ RNA library

preparation Kit (v2). Paired end 101 nucleotide reads (209 cycles) were acquired for all library types on the Illumina HiSeq 2000 sequencing platform. The average coverage for whole-genome and whole-exome sequencing is 30-fold and 100-fold, respectively.

Whole-genome and whole-exome sequencing data has been uploaded to the European Bioinformatics Institute data portal (The European Genome-Phenome Archive, EGA; https://www.ebi.ac.uk/ega/) under the accessions of EGAD00001001432 for whole-genome sequencing and EGAD00001001433 for whole-exome sequencing. For patients with leukemia, germline samples were obtained once patients were in remission; for patients with solid tumors or brain tumors, a peripheral blood sample was used as the germline sample.

Validation of germline variants

All identified rare non-silent germline variants in the 60 autosomal dominant cancer susceptibility genes and *MUTYH*, the autosomal recessive gene on the ACMG list, regardless of pathogenic significance, were validated by MiSeq and/or Sanger sequencing in both germline and matched tumor samples. Primers were designed to flank the identified variant by using Primer3[4,5] with modifications. Optimal amplicon sizes ranged from approximately 400 bp to 800 bp for use in downstream library construction. PCR was performed using the 2X AmpliTaq Gold 360 master mix (Applied BioSystems), 400nM of each primer, and 20 ng of whole genome–amplified (WGA) DNA using the following parameters: 95°C for 10 min, 95 °C for 30 sec, 65 °C for 30 sec, 72 °C for 1 min for 35 cycles, 72 °C for 7 min, and storage at 4 °C. All amplicons were quality-checked on a 2% agarose E-gel (Invitrogen). Amplicons were pooled and purified using a Qiagen PCR cleanup kit. DNA libraries were created from pooled amplicons by using the

Nextera XT kit (Illumina), following the manufacturer's instructions. Libraries were normalized for sequencing on an Illumina MiSeq by using a 2 × 150 paired-end version 2 sequencing kit.

In total, validation was carried out for 623 germline variations that have available DNA. The overall validation rate was 98% (611/623). The 12 false positives were comprised of 7 wild-type genotype calls and 5 variants caused by tumor-in-normal contamination.

Analysis of tumor-in-normal contamination and germline mosaicism

A low mutant allele fraction (MAF) in the germline sample may arise from either germline mosaicism or tumor-in-normal contamination. The use of WGS/WES for paired tumor/normal specimens enables the distinction of tumor-in-normal contamination from mosaicism, as the former is expected to affect all somatic mutations identified in the tumor genome, while germline mosaicism will only affect a single variant. Potential tumor-in-normal contamination were assessed based on the presence of mutant allele in clonal somatic SNVs identified in tumor. These clonal somatic SNVs were identified based on the expected mutant allele fraction after adjusting for the copy number status of each variant in the following possible three categories: 1) diploid heterozygous, 2) copy number neutral LOH and 3) 1-copy loss LOH and subsequent assessment of contamination also taken into account the CNV status in tumor. We then used a binomial distribution model to assess whether the observed mutant allele count in a germline sample was significantly less than the count predicted by the tumor-in-normal contamination level. A significant P value ($P < 0.05$) resulted in rejection of the contamination hypothesis, and a high-depth verification sequencing was in turn carried out to validate the presence of mosaicism. Variants with significant P value but had a germline copy number gain were filtered as the

reduction of mutant allele of fraction from the expected 0.5 in a diploid genome could be attributed to bias in amplification of the chromosome that represents the reference allele.

Automated germline classification

With the exception of variants that matched those contained within the IARC *TP53* database (http://p53.iarc.fr/), only novel single nucleotide variants (SNVs) or SNVs with <0.1% population frequency in the NHLBI ESP database (http://evs.gs.washington.edu/EVS/) received an automatic classification. In-frame indels that matched dbSNP did not receive a classification. Indels present in 1000 Genomes or with multiple submissions in dbSNP were also excluded.

Automated germline classification was performed for the non-silent coding variants detected within the list of 565 cancer-associated genes (Figure S1). This classification was based on matches to multiple germline variant databases as well as evaluation of potential functional annotations. The curated mutation databases included International Agency for Research on Cancer *TP53*, Breast Cancer Information Core, NCBI ClinVar, Human Gene Mutation Database, Universal Mutation, Leiden Open Variation Database, locus-specific databases for *APC*, *MLH1*, *MSH2*, *MSH6*, *PMS2*, *RET* and *RB1*, recurrent somatic mutations identified by PCGP or deposited at Catalogue of Somatic Mutations in Cancer (COSMIC, http://cancer.sanger.ac.uk/cancergenome/projects/cosmic).

Classification consisted of three tiers, named gold, silver, and bronze. Gold is reserved for truncation mutations in tumor suppressor genes, matches to truncation mutations and hotspot mutations in somatic mutation database, and perfect matches to highly-curated locus-specific databases, including the IARC *TP53* database, NHGRI *BRCA1* and *BRCA2* database (http://research.nhgri.nih.gov/bic/) for records marked clinically important, ARUP *MEN2*

database for mutations in *RET* (http://www.arup.utah.edu/database/MEN2/MEN2_display.php), ASU database (http://telomerase.asu.edu/diseases.html) for *TERT* mutations, LOVD (http://chromium.lovd.nl/LOVD2/colon_cancer/home.php) for mutations in *APC* and *MSH2*, and the *RB1* mutation database (http://rb1-lsdb.d-lohmann.de/home.php). Silver is used for variants of less certain significance, and bronze for variants more likely to be benign. Additional databases that were utilized included HGMD (http://www.biobase-international.com/product/hgmd), ClinVar (http://www.ncbi.nlm.nih.gov/clinvar/), and UMD (http://www.umd.be/LSDB.html), matches to which generally received a silver classification. Truncation mutations located close to the C-terminus of the protein that do not map to a functional domain were examined manually and consulted with external gene experts for their pathogenicity.

Mutation databases were harmonized by the following approaches. Those containing transcript-oriented alleles were normalized to genomic orientation, where required (e.g. HGMD, COSMIC), and the genomic reference sequence was verified for all variants. Protein-change annotations were similarly standardized where possible, both to identify involved codons and to correct issues such as variations in stop codon formatting ("*", "X", "Ter", "Term", and "Stop"). The RefSeq accession used by each database for annotation was also extracted to ensure that matches to protein annotations are based on the same isoform. If an indication of clinical importance was provided, it was passed through and standardized into a common 5-tier classification of significance (pathogenic, probably pathogenic, variant of uncertain significance, probably benign or benign) where possible.

Matching was performed via genomic coordinates where available, otherwise amino acid change annotations (a matching RefSeq accession is required for the latter match type) were

used. The automated variant classifier identifies both perfect as well as imperfect matches, such as to the same genomic location or transcript codon number. Imperfect matches (e.g., change to the same codon but to a different amino acid) receive a less significant classification than perfect matches. For indels, variants are permitted +/- 3 nucleotides of mapping ambiguity, but are required to be of the same type and length. Additional evaluation is also performed based on gene class, variant functional annotation, and *in-silico* functional prediction algorithm calls. Truncation mutations including nonsense, frameshift and splice variants in tumor suppressor genes receive gold, otherwise silver. In-frame indels are classified as silver. If PolyPhen2 and MutationAssessor damage-prediction calls are available in dbNSFP[6], variants predicted damaging are classified as silver, otherwise they are bronze. Variants classified as gold by a separate somatic classifier receive silver. Useful annotations such as population information, pathological significance and PubMed IDs are extracted and reported where available.

Curation of somatic mutation databases

Somatic mutation data may facilitate the interpretation of germline data as familial studies have shown that germline mutations that confer cancer susceptibility may overlap with known somatic mutations[7,8]. Two sets of somatic mutation data were used to facilitate automated classification of germline mutation: a) somatic mutations identified from the PCGP; and b) Sanger's COSMIC database, the most comprehensive data repository for somatic mutations identified in cancer. For COSMIC, the mutation data as well as the gene list of COSMIC Cancer Gene Census (http://cancer.sanger.ac.uk/cancergenome/projects/census/) were downloaded. All gene symbols were corrected by using the mapping between Entrez Gene ID and HUGO gene symbol provided by National Institute of Biotechnological Information (NCBI, http://www.ncbi.nlm.nih.gov).

Using these input data, we aimed to generate a list of genes for reporting important sequence mutations. To identify these genes, we focused on validated non-silent coding somatic mutations that have over-representation of truncation mutations (i.e. nonsense, frameshift), a hallmark for the inactivation of tumor suppressor genes; or recurrent hotspot mutations which are commonly identified in oncogenes and functional important mutations in tumor suppressor genes. As a quality check, the 18 genes recently identified to have high mutation rate but with no apparent biological impact[9] were used as a negative control for this analysis. The mutation data from COSMIC used for this study included only those which were experimentally verified by using the tag "Confirmed somatic variant" or "Reported in another cancer sample as somatic". To remove false positive sites that were actually germline polymorphisms, we further filtered those that overlap with germline variants present in >10 of the 6,503 non-cancer individuals sequenced by NHLBI GO Exome Sequencing Project (ESP).

Hypermutable tumors may skew the global mutation profile of a gene due to their contribution of excessive number of somatic mutation; thus were excluded in our analysis. Hypermutable tumors were identified by the following process. If a tumor with >100 coding somatic mutations is among the top 10% highest mutated tumors of the published study, it is considered a hypermutable tumor. A total of 448 tumors in COSMIC were identified as hypermutable through this process.

SUPPLEMENTARY RESULTS

To determine the significance of our findings in pediatric cancer patients compared to control cohorts, we applied the same or similar analytical process (Fig. S1) to characterize pathogenic or probably pathogenic mutations in the 60 genes associated with autosomal dominant cancer predisposition syndromes (Fig. 2) in two non-cancer data sets. The first involves analyzing the raw high coverage depth whole exome sequencing data of 966 unidentified individuals from around the world (age >18 years) as part of the 1000 genome (1KG) project. The second involves analyzing genotype data of a pediatric cohort from the National Database for Autism Research (NDAR) consisting of 515 cases and 208 controls of autism. Details of data collection and analysis for each control cohort are described below.

Analysis of whole-exome sequencing data from 1000 Genome project

A total of 966 unrelated individuals selected from 970 subjects analyzed by whole exome sequencing by phase 3 of 1KG project were included in this analysis. The four excluded individuals either have their relatives (HG00124, HG00501, HG00635) or his/her parents (NA19240) included in the same cohort. Raw sequence data files in the format of binary sequence alignment (BAM) were downloaded by ftp (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase3). The sequence data were generated using Illumina platform and aligned to reference human genome assembly hg19 by bwa (v0.5.9-r16), both of which match the sequence data generation workflow used for PCGP cohort. To compare results derived from *de novo* analysis using raw BAM files with those from genotype data generated by 1KG project, files in Variant Call Format (VCF) from version 5 data release were

downloaded by ftp (ftp://ftp-trace.ncbi.nih.gov/1000genomes/ftp/release/20130502/supporting/vcf_with_sample_level_annotation/)

Using the raw BAM files, we carried out variant detection, variant filtering, manual review of sequence reads, mutation classification and panel review of mutation pathogenicity for the 60 autosomal dominant cancer susceptibility genes. A total of 11 pathogenic or probably pathogenic mutations were identified in *APC* (n=1), *BRCA1* (n=1), *BRCA2* (n=4), *MSH6* (n=1), *SDHA* (n=1), *SDHB* (n=1), and *TP53* (n=2), one in each of 11 individuals (Table S6A). Three of the 11 mutations were also detected in PCGP cohort.

Five out of the 11 mutations were classified as pathogenic (Table S6B) as they have strong evidence for cancer susceptibility based on published literature, information in locus-specific databases, and/or submitted case reports from molecular diagnostic labs in ClinVar, as detailed further below. Their presence within the 1KG cohort indicates uncharacterized cancer phenotypes in the 5 subjects enrolled, rather than refuting their significance in cancer susceptibility. One of the pathogenic mutations, p.R273H in *TP53*, is likely to be mosaic as the mutant allele fraction is below 0.3.

Six additional mutations were classified as probably pathogenic, five of which are truncation mutations in genes where loss of function is a known mechanism of disease. They were marked probably pathogenic because there is not enough corroborating evidence from other observations of the mutation in cancer to elevate them to the category of pathogenic under the ACMG (2015)-based guidelines. As with the pathogenic mutations described above, the presence of these probably pathogenic mutation in the 1KG individual is more likely indicative

of the unphenotyped cohort, as opposed to a lack of cancer relevance. Collectively, the presence of these mutations represents a frequency of 1.1% in 1KG, which is significantly lower than the 8.5% frequency discovered in our PCGP cohort (Fisher's exact test p value=5.9E-16, odds ratio=7.9, Table S6C).

To evaluate the importance of accessing the raw sequence data vs. relying upon pre-processed genotype data, we performed a parallel analysis using the VCF files for mutation classification. A total of 33 mutations were classified as pathogenic or probably pathogenic; however, manual review of sequence reads using the BAM files revealed that 25 of these variants were false positives due to mis-alignment in a polyA track or lack of sequence coverage. Furthermore, 4 pathogenic or probably pathogenic mutations identified by de novo analysis using the raw sequence files (e.g. the BAM files) were missing in the VCF files. The difference in pathogenic mutations identified by de novo analysis using BAM files compared to those derived by VCF files highlights the importance of accessing raw sequence data when establishing the prevalence of variant pathogenicity, as false positive or false negative errors can have significant impacts on final results.

Analysis of NDAR autism genotype data

The autism data set consists of 723 individuals including 515 cases and 208 controls. The median age is 6.0 years (range 1-37) which is highly comparable to the median of 6.9 years (range, 0 – 20 years) in PCGP. Genotype data for these individuals were from the VCF files downloaded via the link "Genotype calls BROAD hg19" on NADR web site (https://ndar.nih.gov/study.html?id=307) . We selected this data set because the genotype data were computed from whole exome sequencing data generated by Illumina HiSeq2000 or

Genome Analyzer IIx which is similar to those used for the analysis of PCGP cohort. The VCF files were used for classification of pathogenic mutations. Due to observed issues with increased false positive and negative rates with this approach as noted in the previous section, we removed mutations suspected to be caused by mis-alignment within polyA tracks or other known signatures of sequencing artifacts. *NF1* and *PTEN* were not included in the analysis as these two genes have been known to be associated with autism[10,11]

A total of 4 out of the 723 subjects (0.6%) have pathologic or probably pathologic mutations compared to the 8.0% in the PCGP cohort and the result is highly significant (Table S7A-C, Fisher's Exact test p=7.4e-16, odds ratio=15.7). None was found in PCGP cohort. All four variants are truncation mutations in genes where loss of function is a known mechanism of disease, two of which are classified pathogenic as they have strong evidence for disease association (Table S7B).

Our analysis of two cohorts not selected based on cancer phenotype (966 persons in 1KG and 723 persons in NDAR autism) shows approximately 1% prevalence of pathogenic mutations in the two cohorts, compared to the ~8% in PCGP; this difference is highly significant (p value <1E-10). Use of the raw sequence files enables us to deliver a reliable result for the analysis of the 1KG cohort, while the VCF file based approach used for the NDAR analysis can only be considered a general estimate of the trend. Interestingly, the number of individuals carrying pathogenic or probably pathogenic germline mutations in the two control data sets is not significantly different (Fisher's exact test, p value =0.16). The results provide sufficient support for the importance of our findings in pediatric cancer . The 1KG data set was analyzed with the same procedure as used for our PCGP cohort, thereby providing a control for evaluating the significance of the PCGP findings.

The presence of pathogenic mutations in 1KG and NDAR indicates uncharacterized cancer phenotypes in people enrolled these studies, rather than refuting the mutation pathogenicity in relation to cancer susceptibility. An estimated 35-40% of the population is diagnosed with cancer during the course their lifetime and 25% will succumb to their disease (http://seer.cancer.gov/statfacts/html/all.html) and previous population based studies have revealed 5-10% of cancer is heritable (http://www.cancer.org/cancer/cancercauses/geneticsandcancer/heredity-and-cancer). Therefore approximately 20-40 individuals in 1KG cohort are predicted to be predisposed to developing cancer. These estimates are consistent with our evaluation of the 1KG genomes whereby we detected 11 pathogenic or probably pathogenic mutations. The observation that we are on the lower end of this estimated range is also not surprising, given the incompleteness of our knowledge of cancer predisposition genes.

Significance of *BRCA1*, *BRCA2* and *PALB2* mutations in pediatric caner

We observed that eight children harbored heterozygous mutations in *BRCA1*, *BRCA2* and *PALB2*. These genes are not generally examined in children as they are considered to be predisposition genes for adult cancers.[12-17] Interestingly, Magnusson et al.[38] described a significantly higher prevalence of childhood cancer in families with germline *BRCA2* mutations, and compared to population-based control families. Brooks et al.[39] reported 20 cases of childhood cancer in 379 families with *BRCA1* or *BRCA2* mutations.[18] Although none of the children in either study were tested for the familial *BRCA1/2* mutations, they were assumed to be mutation carriers. The tumors described in these two reports strongly resemble those seen in our cohort, and include sarcomas, brain tumors and leukemias. Given the physical and functional interactions of *BRCA1*, *BRCA2* and *PALB2* during the DNA damage response,[19] it is perhaps not

surprising that the tumors observed in children with these mutations are similar to those seen in LFS.[20] Curiously, we observed that within the tumors from these patients, only one of our patients exhibited evidence for mutational inactivation or loss of the second allele. Several studies have looked for loss of heterozygosity (LOH) in the breast, ovarian, pancreatic and prostatic tumors of *BRCA1* and *BRCA2* mutation carriers.[21-27] Collectively, these studies reveal an LOH frequency of ~80% for *BRCA1* and ~70% for *BRCA2*.[28] Therefore, not all *BRCA1* or *BRCA2*-associated tumors exhibit LOH. Currently, we do not know whether the *BRCA1*, *BRCA2* or *PALB2* mutations reported here contributed to the formation of the tumors in this cohort. Nonetheless, it is by carrying out unbiased genetic studies such as ours, in conjunction with further biological investigations, that we will be able to elucidate whether there is a causal link between these mutations and the formation of tumors in children.

Mutations in 29 genes associated with autosomal recessive cancer predisposition syndromes

A total of 482 non-silent germline variants were identified within the 29 genes with autosomal recessive cancer predisposition syndromes. Thirty-nine variants were classified as pathogenic or probably pathogenic in 38 patients, and the majority (85%) were truncation mutations with the most commonly affected genes being *ATM* (n=5), *ERCC3* (n=5), and *WRN* (n=4). LOH was identified in the tumor tissues from 3 patients, including 1 each with mono-allelic germline mutations in *SH2B3*, *FANCA* or *FANCM*. A single patient with ataxia telangiectasia and a high-grade glioma harbored bi-allelic pathogenic mutations in *ATM* (Fig. S4 in Supplementary Appendix 1). Of note, 4 patients harbored mono-allelic pathogenic/probably pathogenic mutations in genes associated with autosomal dominant as well as an autosomal recessive cancer predisposition syndromes (Table S8 in Supplementary Appendix 1).

Mutations in tumor suppressors, tyrosine kinases and other cancer genes

Within the 58 other tumor suppressor genes examined, 730 non-silent mutations were identified (Table S5). These included 19 truncation mutations, 11 of which affect *CHEK2* (n=4), *PML* (n=4) and *BUB1B* (n=3). Within the 23 kinases, 315 non-silent germline mutations were identified. Four of these mutations occurred in the tyrosine kinase domain of *EGFR* (T790M (n=1), H773D (n=1) and P848L (n=2), and were recurrently targeted by somatic mutations, mostly in lung cancer. While T790 and P848 exhibited the same amino acid change in somatic and germline mutations, only T790M (rs121434569) was deemed probably pathogenic by ClinVar. This mutation was also previously found to confer inherited susceptibility in familial lung cancer and was associated with resistance to kinase inhibitors[29]. Finally, within the 395 other cancer genes examined, 3,239 non-silent mutations were observed, including 86 truncation mutations identified in 82 cases. Genes recurrently affected by truncation mutations included *USP6* (n=6), *RNF213* (n=5), *FCRL4* (n=4), *SHROOM2* (n=3), *LAMA5* (n=3), *GNAS* (n=3), *ECT2L* (n=3), *TSHR* (n=2), *TRIP11* (n=2), *POLD1* (n=2), *NUP214* (n=2), *KIF5B* (n=2), *HOXA9* (n=2), *ACSL6* (n=2). Excluding those with P/PP mutations identified in the 89 cancer predisposition genes, an additional 78 cases were found to have truncation mutations in tumor suppressor genes or other cancer genes.

Assessment of enrichment of protein truncation mutations in PCGP patients

Across the 565 genes analyzed in this study, a total of 190 truncation mutations were identified in 176 out of 1,120 children (Table S5). Truncation mutations were most enriched in *TP53*, *NOP10*, *FCRL4*, *PML*, *SHROOM2*, *USP6*, *WRN*, *ERCC3*, *HOXA9* and *RB1* when compared to control individuals based on the an analysis of enrichment of protein truncation mutations

described below.

To evaluate whether there is significant enrichment of protein truncation mutations in pediatric cancer patients, we compared the truncation mutations detected in the 1,120 PCGP patients with those in the 60,706 unrelated individuals analyzed by the The Exome Aggregation Consortium (ExAC) on a gene by gene basis (Table S9). Truncation mutations included nonsense, frameshift and splice site variations affecting the +1, -1, -2 and -3 splice junctions. While detailed phenotype data are currently unavailable, ExAC has removed individuals with serious childhood disorders (http://exac.broadinstitute.org), making it an acceptable control for the PCGP cohort. ExAC data for each gene was downloaded by FTP (ftp://ftp.broadinstitute.org/pub/ExAC_release) as publicly available VCF files (release 0.3, January 2015). ExAC data were subsequently subjected to the following criteria to achieve consistency with the St. Jude germline variant filtering process: 1) variants present on a list consisting of known dbSNP variants were filtered unless thery were also present in a "rescue" list of known pathogenic mutations from ClinVar, HGMD, and COSMIC; and 2) all variants had to have an observed combined population frequency of <0.1%. While germline risk variants are expected to exceed this threshold in some instances, these criteria allowed us to focus on rare, highly penetrant alleles with very low population frequency. For each gene that showed enrichment of truncation variants in the PCGP cohort (odds ratio >1), we used Fisher's Exact test to assess the significance of enrichment compared to ExAc. This comparison was done for exploratory purposes only and multiple caveats can influence interpretation. These include differences in technical assays, sample demographics, and variant calling pipelines, any one of which can lead to systematic biases that adversely impact results. While a number of steps can be taken to mitigate these effects, most involve access to primary data for joint-variant calling,

and/or other sample meta-data which are not available in the public version of ExAC data set. Finally, ExAC include 7,601 TCGA patients out of 60,706 individuals (12.5%), and the comparison may inadvertently render genes that were mutated both in adult cancer and pediatric cancer insignficant. Given the above caveats, these results should be considered an exploratory comparison and interpreted with caution.

Estimation of germline mutation prevalence based on cancer subtype distribution in SEER

The patients examined in this study were from the Pediatric Cancer Genome Project, which was designed to include children with difficult-to-treat cancers for which there was limited understanding of the underlying mechanism driving tumor formation. The cohort includes an over-representation of patients with leukemia (particularly hypodiploid acute lymphoblastic leukemia, HYPO ALL) and adrenocortical tumors (ACT). ACT account for 1.7% and 3.5% of patients in SEER and PCGP, respectively; therefore, we reduced its contribution to overall germline mutation incidence by approximately 50%. HYPO accounts for 2% of the leukemia in SEER[30] but 8% in PCGP leukemia. The mutation frequency in HYPO is approximately 23.4% while non-hypo ALL is approximately 2.7%. We first re-calibrated the prevalence of germline mutations in leukemia by adjusting for the over-representation of hypodiploid ALL in our cohort; the resulting 3.1% is lower than the original 4.4% but is still substantial. PCGP cohort has an over-representation of leukemia (52.5%) compared to the SEER (25%). Adjusting for both the over-representation of leukemia and over-representation of HYPO in leukemia resulted an overall contribution of 0.78% germline incidence in SEER by leukemia. Similar process has been applied to all cancer subtypes that were analyzed by PCGP. For cancer subtypes that were not included in PCGP, we applied the lower and higher germline mutation frequencies as reported in the published literature. Summarizing the contribution from all cancer subtypes

resulted in an estimated 7.3%-9.8% germline prevalence for SEER cohort, which is comparable to the 8.5% in our study. The details are presented in Table S10.

SUPPLEMENTARY FIGURES

Figure S1 Variant detection, analysis and classification pipeline.
All coding variants were called from whole-genome sequencing (WGS) and whole-exome sequencing (WES). Transcriptome sequencing (RNAseq) were used for verifying aberrant splicing caused by splice variants. A series of QC checks was performed to ensure high-confidence variant calls. Only rare variants with a population frequency <0.1% were retained. Rare variants matching known pathogenic variants in selected databases were labeled as putative "pathogenic variants". If there were good matches in other databases that had weaker evidence of pathogenicity, or if there was *in-silico* evidence supporting possible pathogenicity, variants were classified as of uncertain significance. All pathogenic and probably pathogenic variants, and variants of uncertain significance affecting genes associated with autosomal dominant cancer predisposition syndromes, were validated using MiSeq or Sanger sequencing. A panel was formed to review each variant within the 89 cancer predisposition genes based on evidence from the literature, disease and locus specific databases, functional assay data, phenotypic data, disease inheritance patterns, and effects on RNA splicing. If the original mutant allele frequency (MAF) was lower than 0.3, we evaluated whether the variant was mosiac. P= Pathogenic, PP= Probably Pathogenic, U= Uncertain, PB=Probably Benign, and B= Benign.

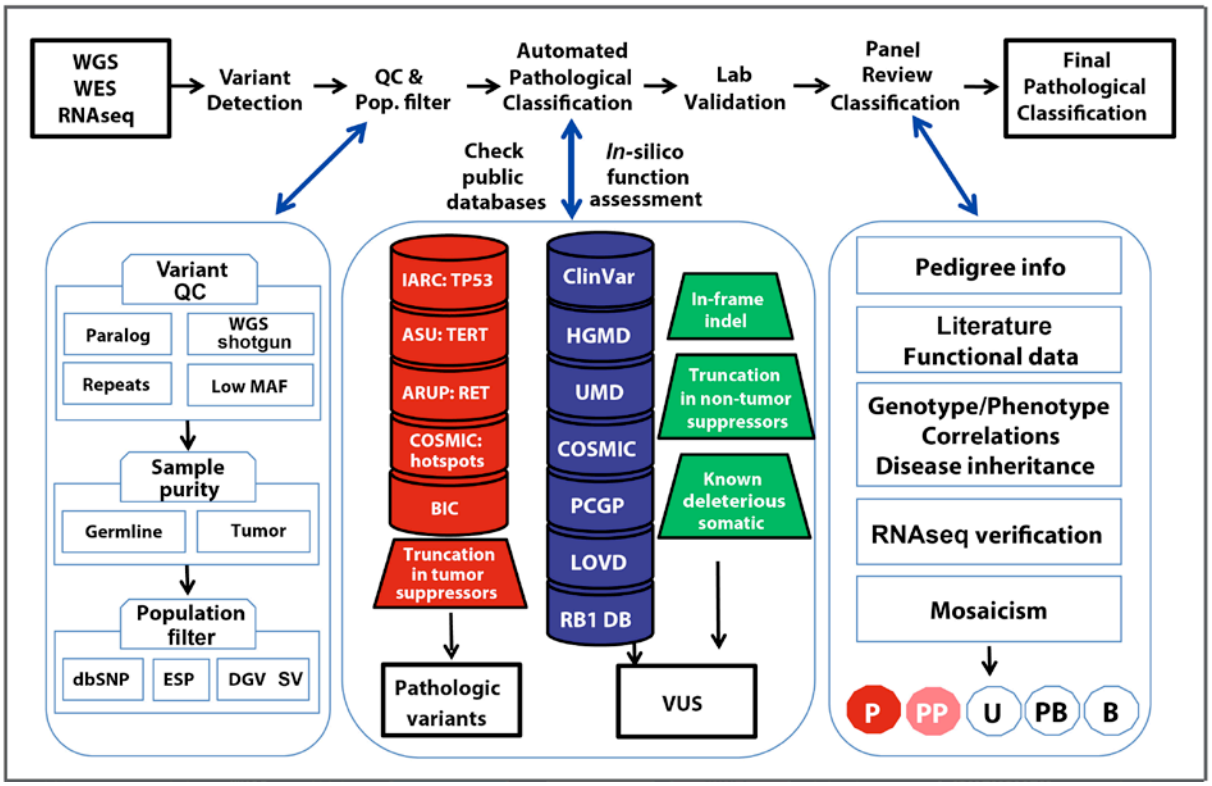


Figure S2 Germline copy number variations in *TP53*, *KRAS* and *PMS2*.
Wiggle plot of WGS read-depth spanning each CNV is shown on the UCSC genome browser. The germline genome is displayed in green color while the tumor genome is shown in brown color. A) In case HGG078, a 10.7-kb focal deletion removed the last two exons of *TP53*. B) In HYPO052, an 8.7-kb focal deletion removed exon 2 to exon 5 of *TP53*. The second allele of *TP53* is lost in the tumor of HYPO052. This finding is consistent with previously reported[31] aberrant splicing across exon 1-exon 6. C) In EWS001316, a 4.9-kb focal deletion removed exon 9 and exon 10 of germline *PMS2*. This CNV was also present in the patient's tumor genome. D) A 1.8-Mb germline amplification spanning the *KRAS* locus in ETV073. We verified that the germline specimen of this case did not have tumor-in-normal contamination as deletion of *CDKN2A*, a well-characterized somatic lesion in ALL, was observed only in the tumor but not in the germline genome (data not shown). E) *PMS2* germline deletion resulting from intra-chromosomal translocation in LGG040. A 33-kb focal deletion removed the first 11 exons in the *PMS2* locus. The gap was replaced with a 25.8-kb segment near the paralog *PMS2CL*. Therefore, in whole-genome sequencing, the *PMS2* locus shows a deletion, while the *PMS2CL* locus shows a pattern of focal amplification.

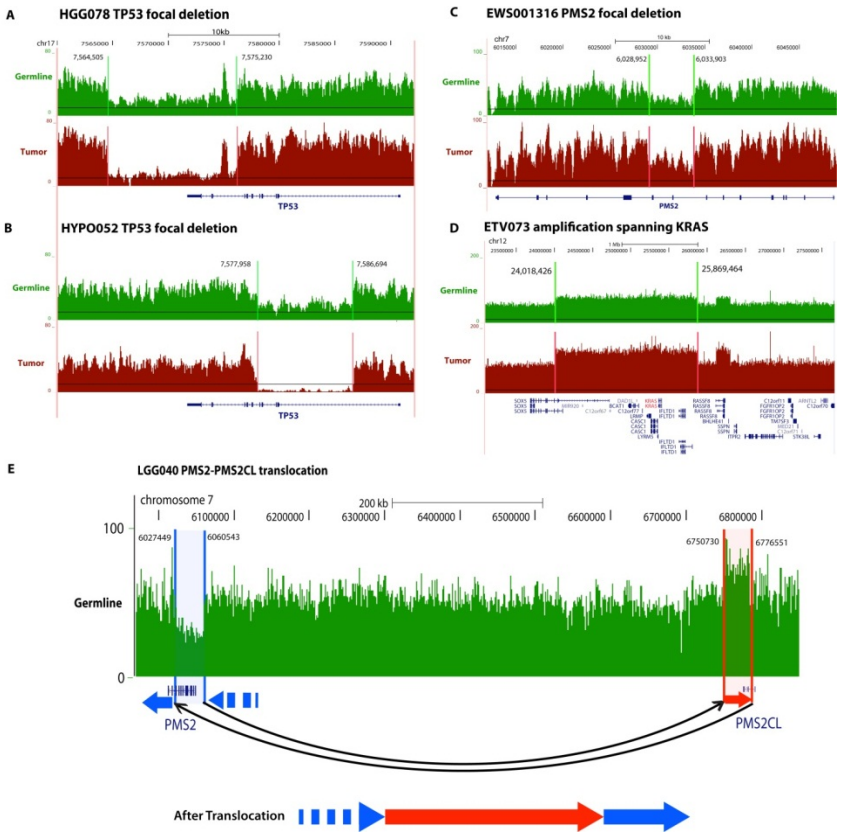


Figure S3 Distribution of pathogenic or probably pathogenic germline mutations.

Distribution of P/PP mutations within the 21 mutated genes associated with autosomal dominant cancer predisposition syndromes across patients with various cancers included in the PCGP cohort. The data is the same as Figure 3A.

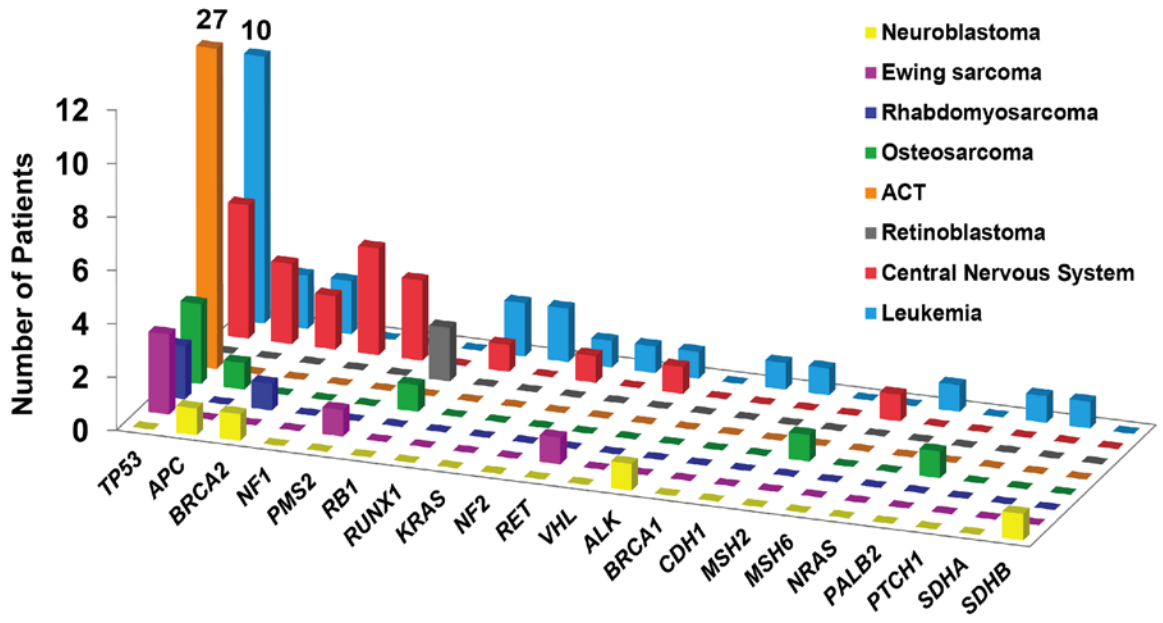


Figure S4. Compound heterozygous germline mutations of *ATM* causing loss of function of both copies of this gene in HGG027.

A) Bi-allelic loss of function germline mutations in *ATM* detected in HGG027 includes a frameshift mutation (p.L2312fs) and a splice site SNV at the splice junction of exon 52. B) The coverage tracks showed the expression of *ATM* gene in HGG027 tumor sample by RNA-seq. C) The germline splice variant leads to a novel exon junction at 11bp downsteam of the canonical exon junction of exon 52, producing a truncated ATM protein

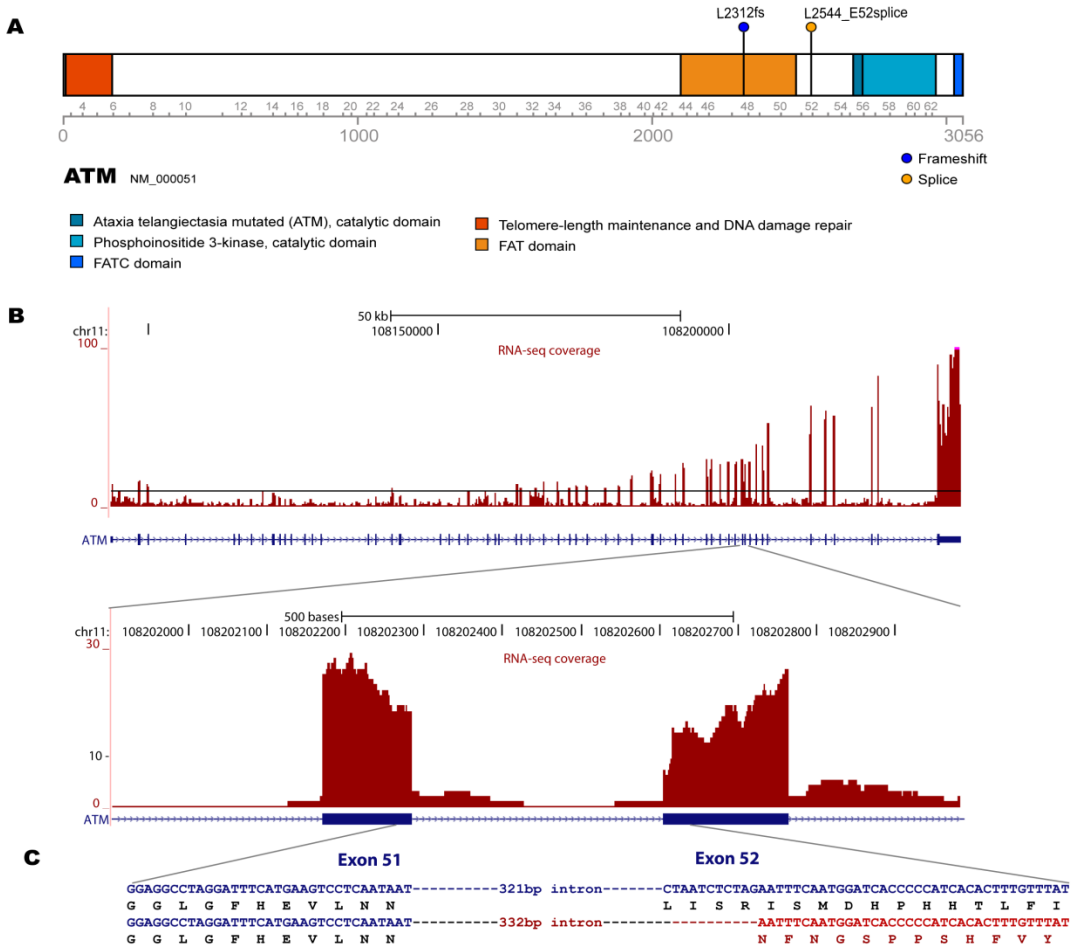


Figure S5. A zoomed-in ProteinPaint view of *TP53* germline mutations on web portal.

The URL for ProteinPaint is http://pecan.stjude.org. The default ProteinPaint view integrates somatic and germline mutations at each amino acid position with a solid arc marking the proportion of germline mutations at each site. This feature is illustrated in the legend "Origin". The legend "Germline" itself also enables a germline-only view, indicated by striking through the "Somatic" and "Relapse" labels. The pull-down menu "Cancer subtype" summarizes the number and the classes of germline mutations in each cancer subtypes in a tabular form and serves as a "menu" for a subtype-specific proteinpaint view. Each subtype can be selected by a user to view mutations in a cancer subtype. The "Sunburst" option provides an alternative summary of the relative contribution of different cancer subtypes to *TP53* germline mutations detected in the PCGP cohort. The user can either view the entire protein or zoom in to a specific amino acid position (T125 in TP53) to view the coding sequence. The dotted line inside the protein demarcates the splice junction. The "silent" mutation T125T is classified as a splice variant (shown in color purple) because this mutation which is located at the -1 donor site causes aberrant splicing of *TP53* based on our evaluation of RNASeq.

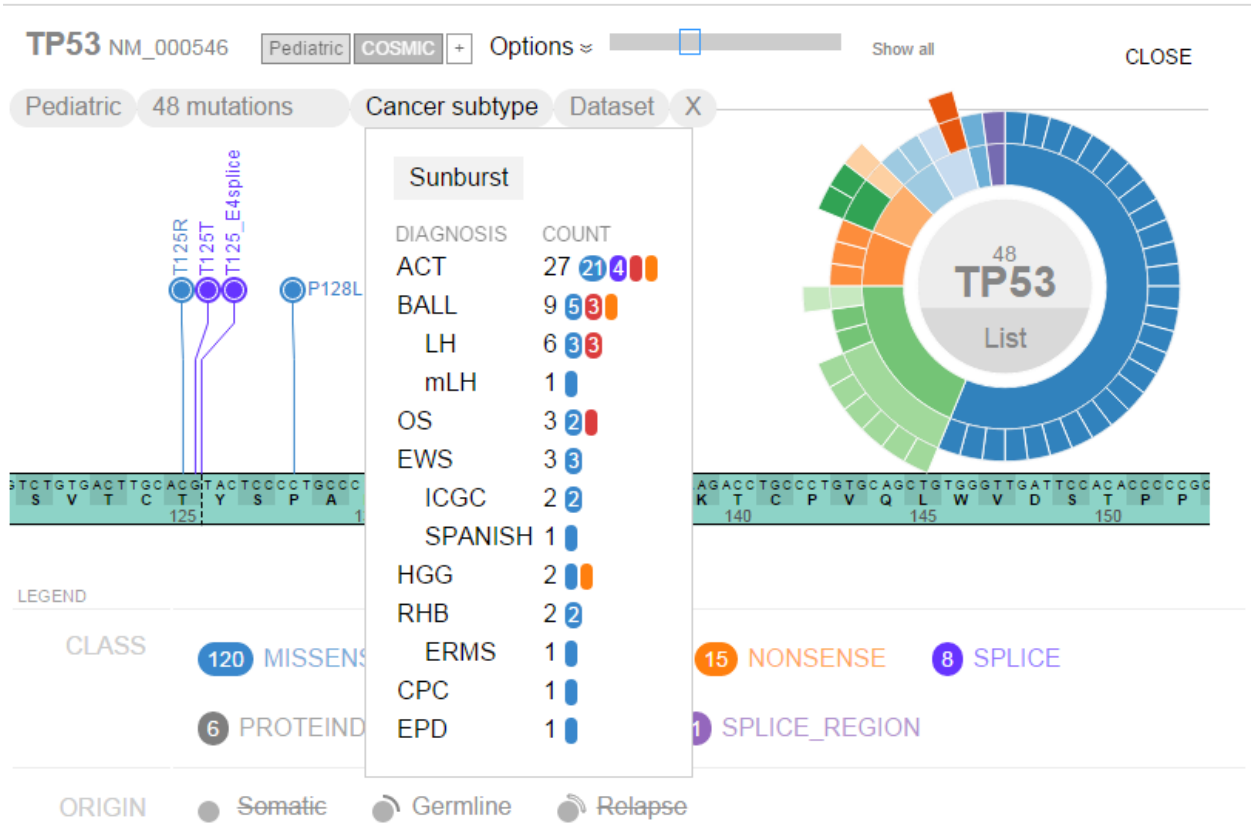


Figure S6. Screenshot of ProteinPaint view of germline mutations in representative genes in each of the five categories.

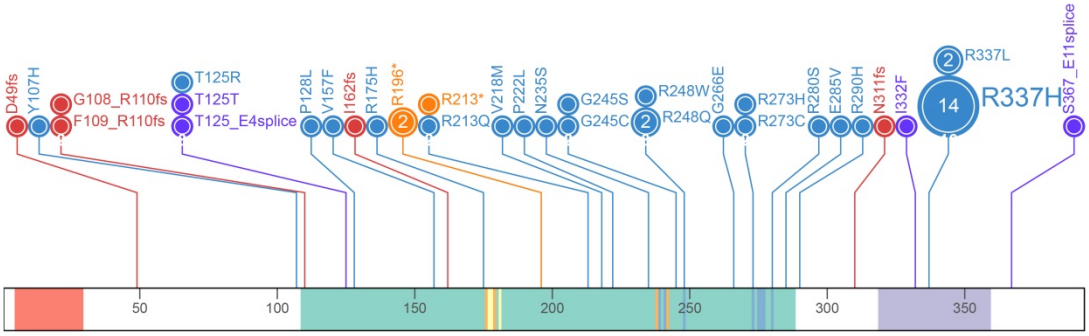
Panel A, B and C shows pathogenic and probably pathogenic mutations in TP53, APC and BRCA2, respectively, from the category of autosomal dominant cancer susceptibility genes. The color code representing mutation class in TP53 applies to all genes.

Panel D, E and F shows pathogenic and probably pathogenic mutations in ATM, ERCC3 and WRN, respectively, from the category of autosomal recessive cancer susceptibility genes.

Panel G, H, I shows protein truncation mutations in CHEK2, PML and BUB1B, respectively, from the category of tumor suppressor genes. Panel J, K shows protein truncation mutations in USP6, FCRL4, respectively, from the category of other cancer genes.

Panel L, M, N shows the three germline mutations T790M, H773D, P848L, respectively, in the tyrosine kinase domain of EGFR that match somatic mutation hotspots within the COSMIC database. The COSMIC data are shown at the bottom, indicating that there are a total of 353, 12 tumors (mostly of lung cancer) are mutated at T790M, P848L, respectively. P848L is the only variant present at low frequency (0.09%) in NHLBI and its status as a valid somatic mutation was only verified by 1 tumor, raising the possibility that it could be a rare germline variation masquerading as a somatic mutation due to the fact that earlier study may have relied on the absence in dbSNP to determine the somatic origin of a genetic variation[32]. P848L is classified as benign in ClinVar, further raising doubts about it being a valid somatic mutation hotspot. By contrast, T790M is absent from NHLBI and is a confirmed somatic variants in >90 tumors and is classified as probably pathogenic by ClinVar. In COSMIC, seven SNVs and 11 in-frame insertions affect H773, although none of them has the alternative allele resulting in a Histidine (H) to Aspartate (D) change. No germline variation at H773 was found at NHLBI or dbSNP.

A

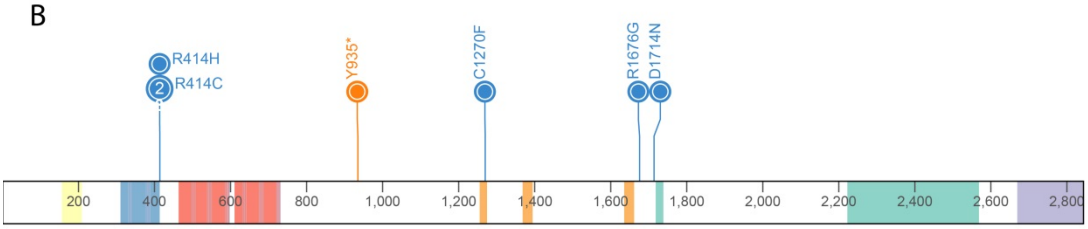


TP53 NM_000546

- **P53** P53 DNA-binding domain; cd08367
- **P53_tetramer** P53 tetramerisation motif; pfam07710
- **P53_TAD** P53 transactivation motif; pfam08563
- **DNA binding** DNA binding site [nucleotide binding]
- **other** dimerization site [polypeptide binding]
- **other** zinc binding site [ion binding]

- MISSENSE
- FRAMESHIFT
- NONSENSE
- SPLICE
- PROTEINDEL
- PROTEININS
- SPLICE_REGION

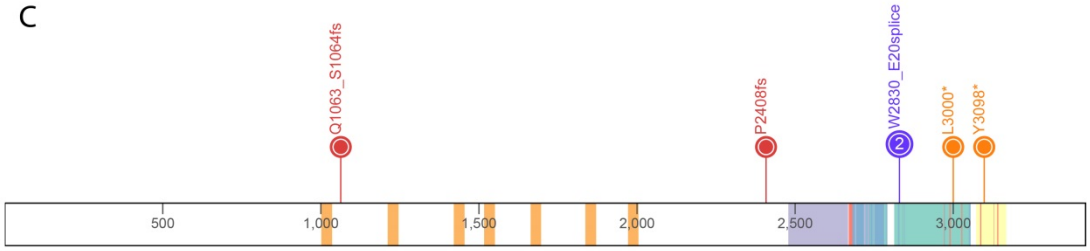
B



APC NM_000038

- **APC_basic** APC basic domain; pfam05956
- **ARM** Armadillo/beta-catenin-like repeats. An ...
- **EB1_binding** EB-1 Binding Domain; pfam05937
- **ARM** Armadillo/beta-catenin-like repeats. An ...
- **APC_crr** APC cysteine-rich region; pfam05923
- **Suppressor_APC** Adenomatous polyposis coli tumour suppre...
- **other** protein binding surface [polypeptide binding]
- **SAMP** SAMP Motif; pfam05924

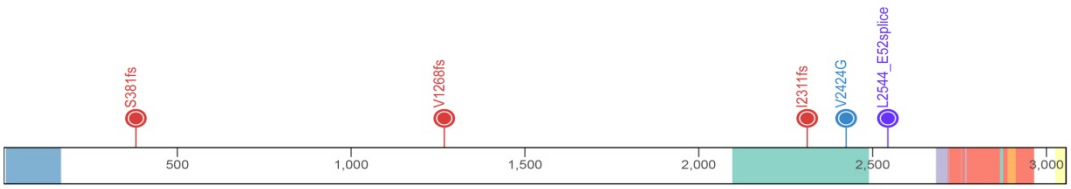
C



BRCA2 NM_000059

- **BRCA2DBD_OB2** BRCA2DBD_OB2: A subfamily of OB folds co...
- **BRCA2** BRCA2 repeat; pfam00634
- **BRCA-2_helical** BRCA2, helical; pfam09169
- **BRCA-2_OB1** BRCA2, oligonucleotide/oligosaccharide-b...
- **BRCA2DBD_OB1** BRCA2DBD_OB1: A subfamily of OB folds co...
- **BRCA2DBD_OB3** BRCA2DBD_OB3: A subfamily of OB folds co...
- **other** ssDNA binding site [nucleotide binding]
- **other** generic binding surface I
- **binding** binding surface II
- **other** OB2/OB3 interface [polypeptide binding]
- **other** generic binding surface II

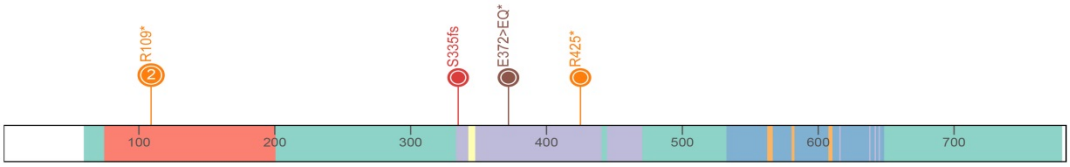
D



ATM NM_000051

- **FAT** FAT domain; pfam02259
- **PIKKc_ATM** Ataxia telangiectasia mutated (ATM), cat...
- **PI3Kc** Phosphoinositide 3-kinase, catalytic dom...
- **TAN** Telomere-length maintenance and DNA dama...
- **FATC** FATC domain; pfam02260
- **other** activation loop (A-loop)
- **other** ATP binding site [chemical binding]
- **active** catalytic loop [active]

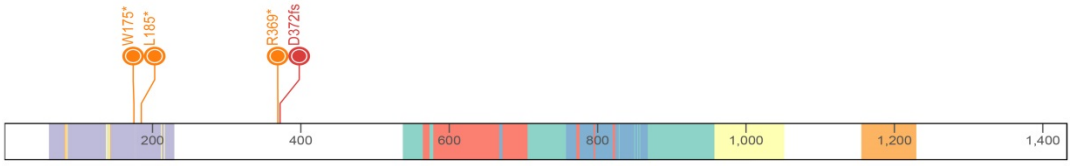
E



ERCC3 NM_000122

- **rad25** DNA repair helicase rad25; TIGR00603
- **DEXDc** DEAD-like helicases superfamily. A diver...
- **Helicase_C_3** Helicase conserved C-terminal domain; pf...
- **HELICc** Helicase superfamily c-terminal domain; ...
- **other** nucleotide binding region [chemical binding]
- **other** ATP binding site [chemical binding]
- **other** putative Mg++ binding site [ion binding]
- **other** ATP-binding site [chemical binding]

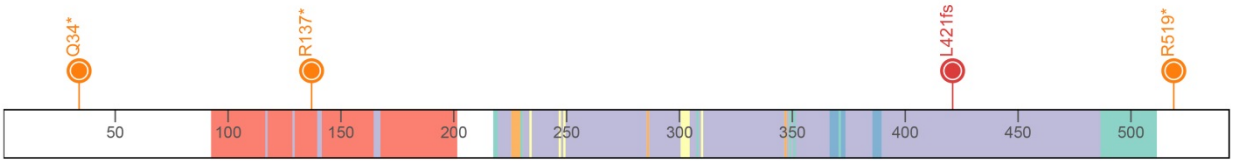
F



WRN NM_000553

- **recQ_fam** ATP-dependent DNA helicase, RecQ family;...
- **RNaseD_like** DEDDy 3'-5' exonuclease domain of RNase ...
- **DEXDc** DEAD-like helicases superfamily. A diver...
- **HELICc** Helicase superfamily c-terminal domain; ...
- **RQC** RQC domain; cl09632
- **HRDC** HRDC domain; cl02578
- **active** putative active site [active]
- **other** putative substrate binding site [chemica...
- **other** nucleotide binding region [chemical binding]
- **active** catalytic site [active]
- **other** ATP binding site [chemical binding]
- **other** putative Mg++ binding site [ion binding]
- **other** ATP-binding site [chemical binding]

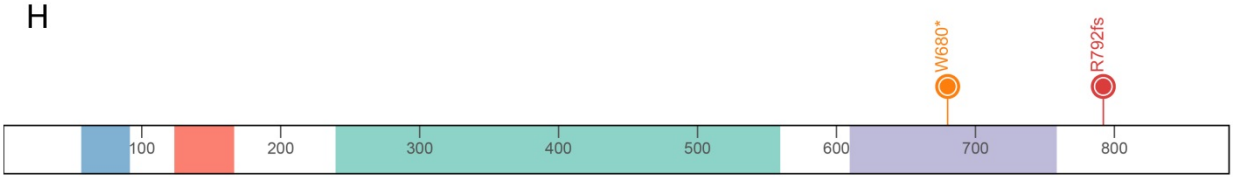
G



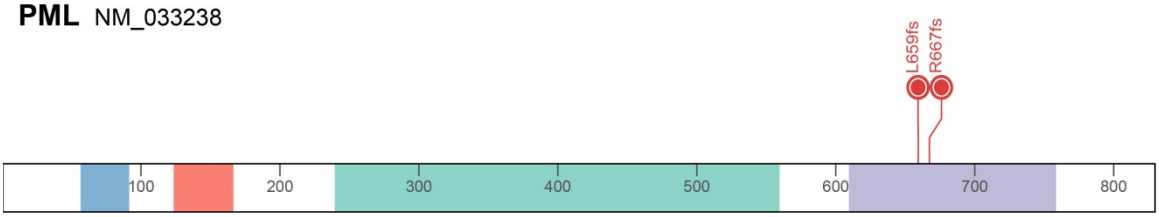
CHEK2 NM_007194

- **PKc_like** Protein Kinases, catalytic domain; cl09925
- **S_TKc** Serine/Threonine protein kinases, cataly...
- **FHA** Forkhead associated domain (FHA); found ...
- **active**
- **other** ATP binding site [chemical binding]
- **other** activation loop (A-loop)
- **other** substrate binding site [chemical binding]
- **other** phosphopeptide binding site

H



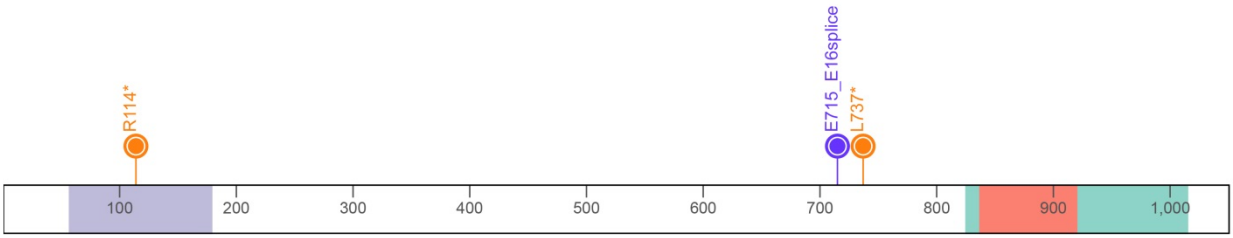
PML NM_033238



PML NM_033239

- **DUF3583** Protein of unknown function (DUF3583); p...
- **DnaQ_like_exo** DnaQ-like (or DEDD) 3'-5' exonuclease do...
- **BBOX** B-Box-type zinc finger; zinc binding dom...
- **RING** RING-finger (Really Interesting New Gene...

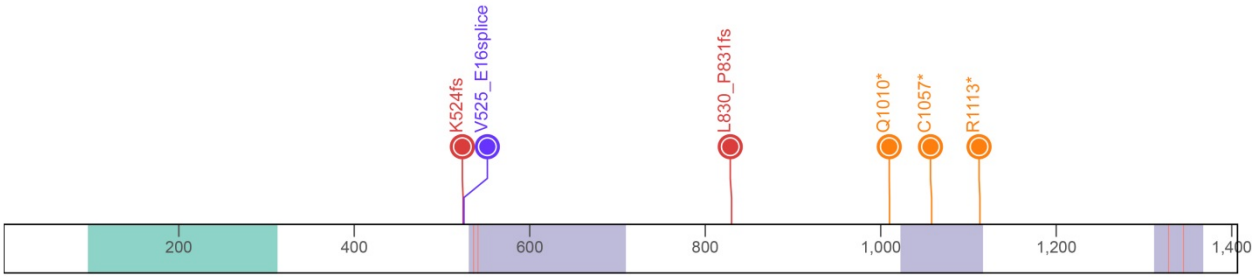
I



BUB1B NM_001211

- **SPS1** Serine/threonine protein kinase [General...
- **Mad3_BUB1_I** Mad3/BUB1 homology region 1; cl07066
- **PKc_like** Protein Kinases, catalytic domain; cl09925

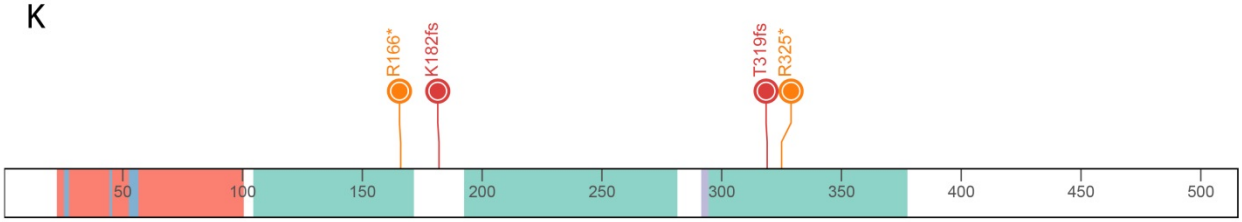
J



USP6 NM_004505

- **Peptidase_C19** Peptidase C19 contains ubiquitinyl hydro...
- **RabGAP-TBC** Rab-GTPase-TBC domain; cl02495
- **active**

K



FCRL4 NM_031282

- **Ig** Immunoglobulin domain; cl11960
- **I-set** Immunoglobulin I-set domain; pfam07679
- **Ig1_FcgammaR_like** Frst immunoglobulin (Ig)-like domain of ...
- **other** pentraxin binding site

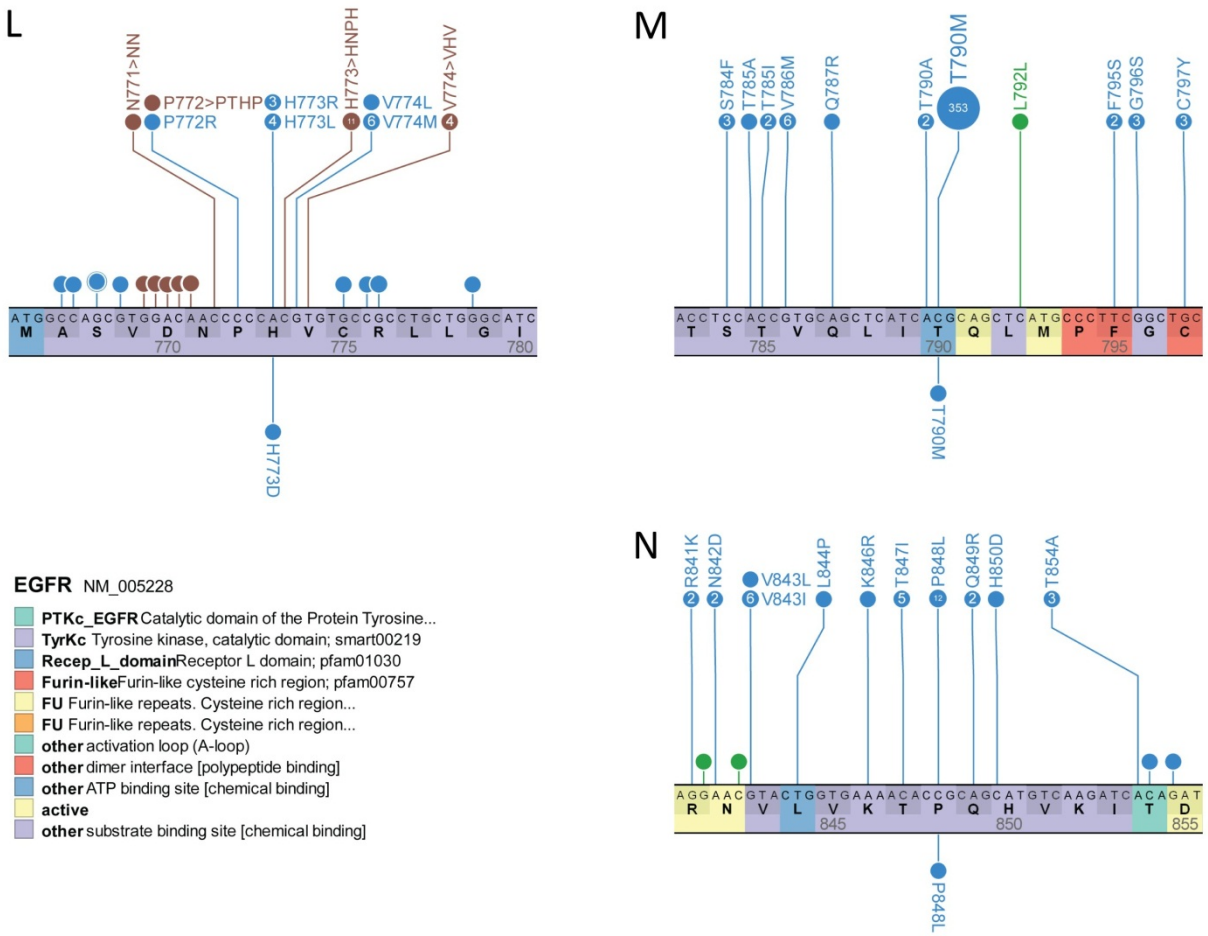

L
N771>NN
P772>PTHP
P772R
H773R
H773L
H773>HNPH
V774L
V774M
V774>VHV
ATGGCCAGCGTGGACAACCCCCACGTGTGCCGCCTGCTGGGCATC
M A S V D N P H V C R L L G I
770
775
780
H773D
M
S784F
T785A
T785I
V786M
Q787R
T790A
T790M
353
L792L
F795S
G796S
C797Y
ACCTCCACCGTGCAGCTCATCACGCAGCTCATGCCCTTCGGCTGC
T S T V Q L I T Q L M P F G C
785
790
795
T790M
N
R841K
N842D
V843L
V843I
L844P
K846R
T847I
P848L
Q849R
H850D
T854A
AGGAACGTACTGGTGAAAACACCGCAGCATGTCAAGATCACAGAT
R N V L V K T P Q H V K I T D
845
850
855
P848L
EGFR NM_005228
PTKc_EGFR Catalytic domain of the Protein Tyrosine...
TyrKc Tyrosine kinase, catalytic domain; smart00219
Recep_L_domain Receptor L domain; pfam01030
Furin-like Furin-like cysteine rich region; pfam00757
FU Furin-like repeats. Cysteine rich region...
FU Furin-like repeats. Cysteine rich region...
other activation loop (A-loop)
other dimer interface [polypeptide binding]
other ATP binding site [chemical binding]
active
other substrate binding site [chemical binding]

Figure S7. Analyis of family history. A) 95 patients with pathogenic or probably pathogenic germline mutations. B) 100 patiens randomly selected from the remaining 1,025 patients.

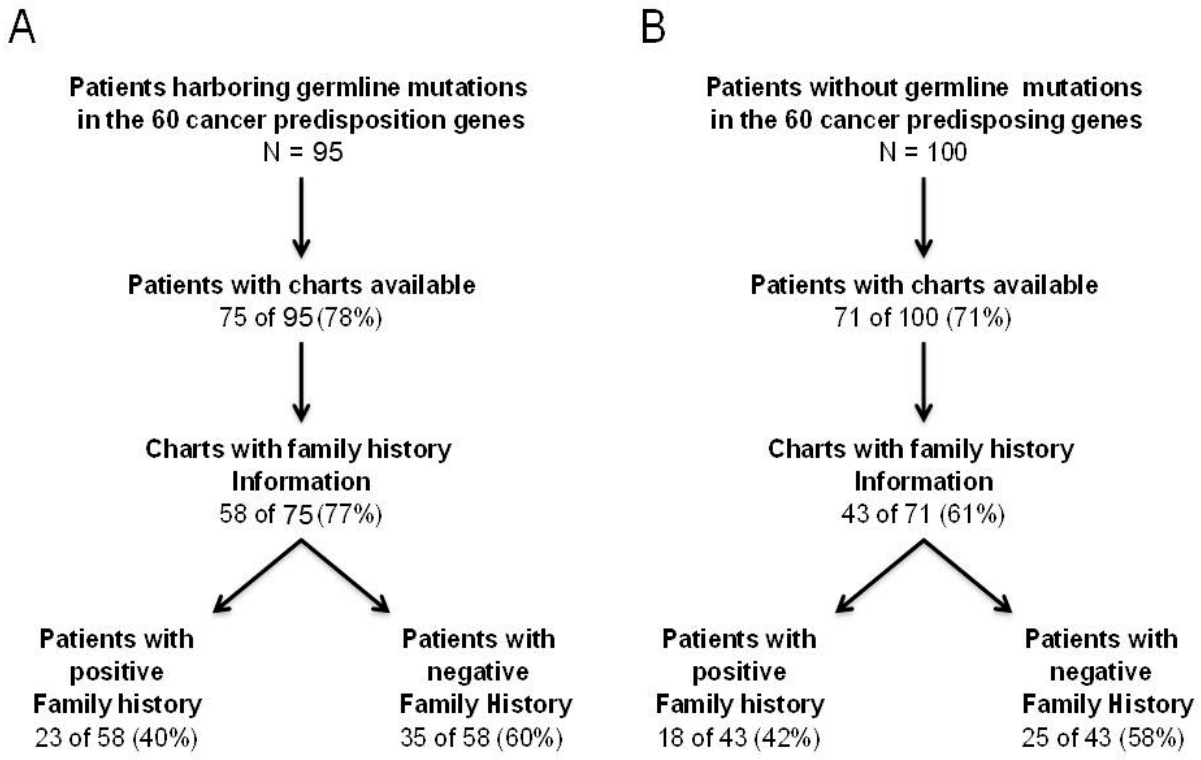


Figure S8. Family history of two cases with *BRCA2* mutation.

A) Family history of MB039 who harbors a heterozygous germline L3000X mutation affecting *BRCA2*. The affected cases from three generations of this family were all mutation-positive. B) Family history of case RHB010 who harbors a heterozygous frameshift mutation in *BRCA2*. The family members were sequenced by Myriad.

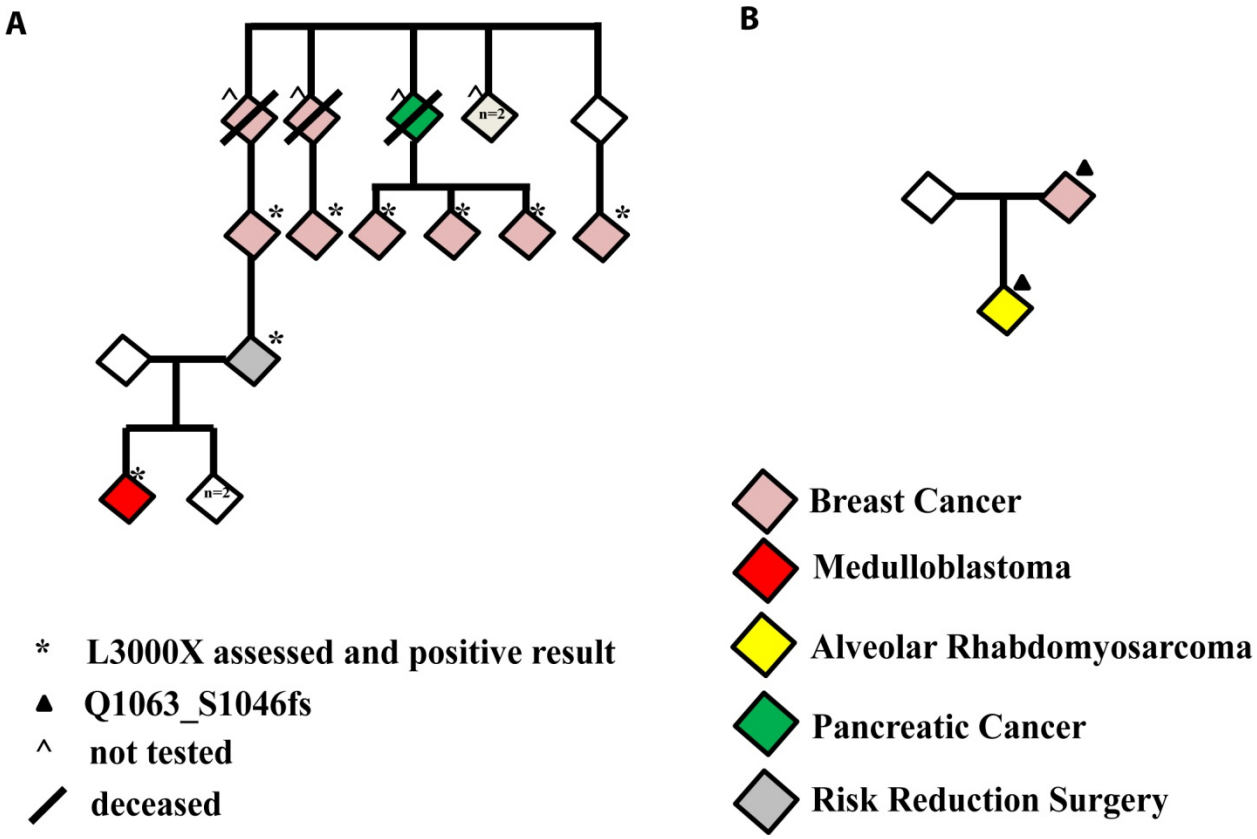
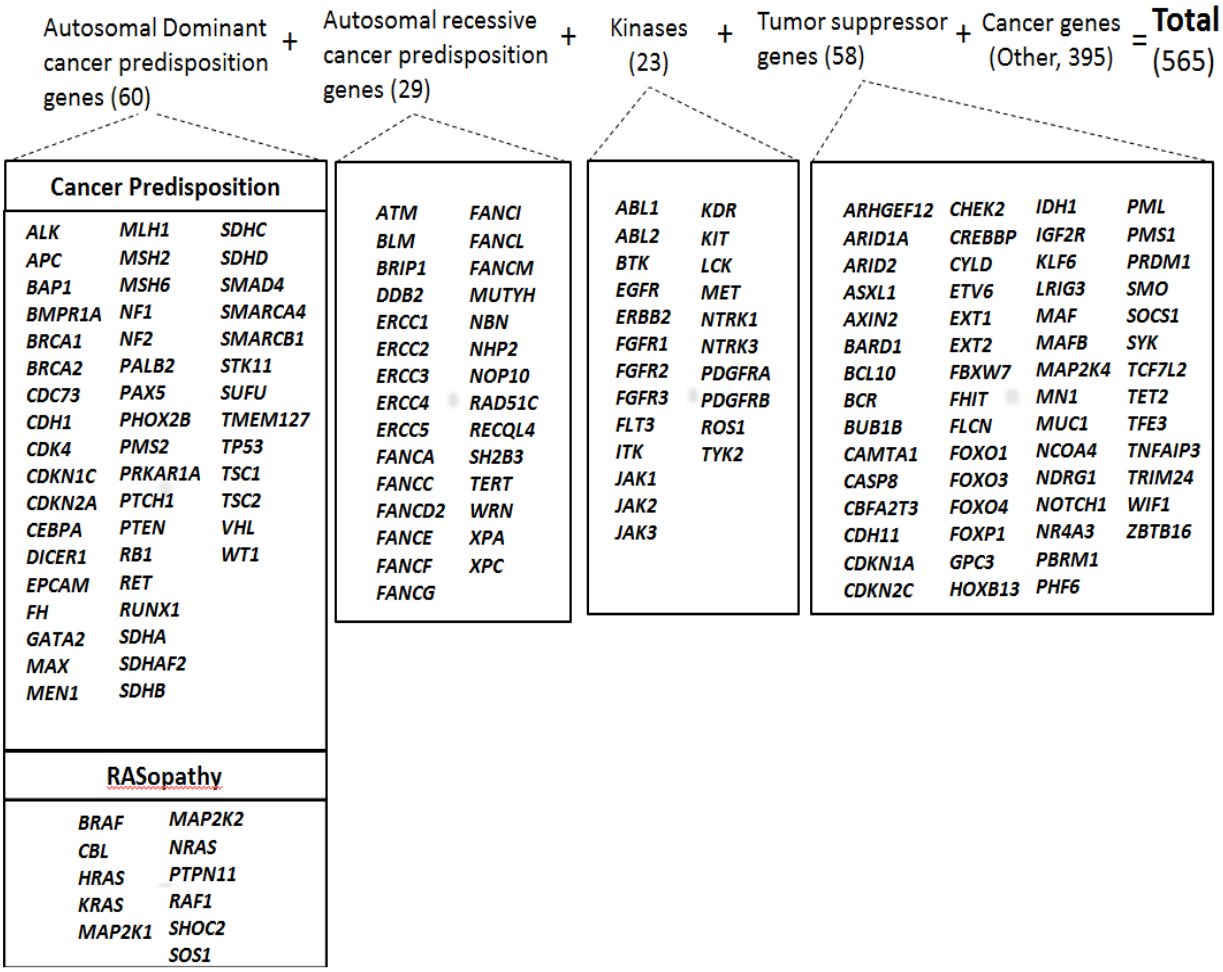


Figure S9. Genes sets assessed in this study.

The members of the primary gene sets examined in this study are depicted. The autosomal dominant cancer predisposition genes (N=60) represent the primary focus of our analyses, followed by autosomal recessive (N=29), kinases (N=23), and tumor suppressors (N=58). We note that the categorization here is simplified for the purpose of display, and some genes fall within multiple categories. MSH6 and TERT, for example, can display both autosomal dominant and autosomal recessive inheritance patterns depending on the biological context. Another example is CHEK2, which is both a tumor suppressor and a kinase. These distinctions are better captured in Supplementary Appendix 2, Table S2. There, the reader will also find additional genes implicated in cancer (N=395) that were surveyed in this study.



SUPPLEMENTARY TABLES

Table S1. Clinical information on patients examined in this study

Disease	Age at Diagnosis						Sex		Subtotal
	0-1 y	1-5 y	5-10 y	10-15 y	15-20 y	Unknown	Female	Male	
Leukemia	34	103	91	107	74	179	260	328	588
Central Nervous System	12	84	85	43	12	9	115	130	245
ACT	3	27	4	3	1	1	25	14	39
Ewing sarcoma	0	0	0	0	0	46	17	29	46
Melanoma	1	1	0	2	1	0	2	3	5
Neuroblastoma	6	25	17	9	2	41	47	53	100
Osteosarcoma	0	0	5	8	3	23	19	20	39
Retinoblastoma	3	1	0	0	0	11	8	7	15
Rhabdomyosarcoma	3	14	16	4	6	0	21	22	43
Subtotal	62	255	218	176	99	310	514	606	1120

Table S2. Categories and information about the 565 cancer genes selected for germ line mutation analysis.

See Supplementary Appendix 2, Table S2.

Table S3. Sequence coverage of the 565 genes analyzed.

Percentage of coding exonic bases at >=10x, 20x and 30x across 565 genes.

NGS	Specimen	% coding exonic bases covered		
		>=10x	>=20x	>=30x
WGS	Germline	96	85	55
	Tumor	97	91	79
WES	Germline	95	91	84
	Tumor	95	91	85

Table S4. Germline mutations discovered in PCGP cohort.

See Supplementary Appendix 2, Table S4a-e for mutations in autosomal dominant, autosomal recessive, tumor suppressor, kinase and other cancer genes.

Table S5. Germline mutation summary in PCGP cohort.

A. Summary of autosomal dominant and autosomal recessive cancer susceptibility genes

Category	#Gene	All Mutations			Pathogenic/Probably Pathogenic		
		#Mut	#Cases	#Genes	#Mut	#Cases	#Genes
Autosomal Dominant	60	633	478	51	95	94	21
Autosomal Recessive	29	486	400	28	39	38	20

B. Summary of tumor suppressor, kinase and other cancer genes

Category	#Gene	All Mutations			Protein Truncation Mutations		
		#Mut	#Cases	#Genes	#Mut	#Cases	#Genes
Tumor Suppressor	58	730	520	53	19	19	9
Kinase	23	315	270	21	9	9	7
Other Cancer Gene	395	3,297	1,070	309	86	82	59

Table S6. Pathogenic (P) or probably pathogenic (PP) mutations identified in the 966 unrelated individuals from 1000 Genome Project

A. Pathogenic or probably pathogenic mutations

^Subject	Gene	Mutation	Type	Pathogenicity
HG00365	*BRCA1*	splice	A1453_E12splice	P
NA20517	*SDHA*	nonsense	S384*	P
HG00127	*SDHB*	nonsense	R90*	P
^NA11994	*TP53*	missense	R273H	P
^NA12003	*TP53*	missense	R290H	P
HG00611	*BRCA2*	frameshift	D2566fs	PP
HG01377	*BRCA2*	nonsense	E2355*	PP
NA19130	*BRCA2*	nonsense	N1747>N*	PP
NA19397	*BRCA2*	frameshift	L760_Y761fs	PP
NA20585	*MSH6*	frameshift	E226fs	PP
^HG00332	*APC*	missense	R414C	PP

^Subjects harbor mutations that were also found in PCGP cohort.

The five loss-of-function mutations highlighted in the middle have strong evidence for pathogenicity and would have been classified as "P" if they had corroborating evidence from other observations of the allele in cancer cases which is required for the Pathogenic status defined by ACMG

B. Evidence supporting classification of five pathogenic mutations in 1KG

BRCA1 A1453_E12splice

- Very rare variant [1 in 120k chromosomes in ExAc]
- Truncating splice [at exon 12/22] in known haploinsufficient gene [ACMG:VeryStrong]
- Previously reported in 3 breast cancer patients, suggesting enrichment [ClinVar]
- 3 previous ratings of pathogenic of familial breast/ovarian cancer [ClinVar]
- Observed in 21 NHGRI *BRCA1* patients, indicating enrichment

SDHA S384*

- Very rare variant [2 in 120k chromosomes in ExAc]
- Truncating allele within gene where loss of function is a known mechanism of disease
- Observed in a homozygous state in an individual with phenotype and no additional known causal alleles (PMID: 21505157)

SDHB R90*

- Very rare variant [1 in 120k chromosomes in ExAc]
- Truncating variant in gene where heterozygous loss-of-function alleles known to result in paraganglioma/pheochromocytoma
- Identified in 5 individuals with paraganglioma/pheochromocytoma [OMIM]
- Segregated with disease in 5 affected family members from 3 families (Astuti 2001, Amar 2005, Sivapackianathan 2007)
- 4 previous ratings of pathogenic in Hereditary Paraganglioma-Pheochromocytoma Syndromes [ClinVar]

TP53 R273H

- Very rare variant [3 in 114k chromosomes in ExAc]
- Germline mutation hotspot in cancer patients [54 instances in IARC-TP53 database]
- Segregated with disease in a family with second malignant neoplasms (PMID: 1565144)
- 4 previous ratings of pathogenic including 2 Li-Fraumeni Syndrome and 2 hereditary cancer predisposition syndrome[ClinVar]
- Somatic mutation hotspot [679 tumors in COSMIC]

TP53 R290H

- Very rare variant [19 in 121k chromosomes in ExAc]
- Germline mutation in 24 cancer patients [IARC-TP53 database]
- 3 Li-Fraumeni Syndrome & hereditary cancer predisposition syndrome[ClinVar]
- Recurrent somatic mutation [20 tumors in COSMIC]

C. Comparison of mutation frequency in PCGP with 1000 Genome

Comparison	**PCGP (mut/all)**	**1KG (mut/all)**	**Fisher's exact test**	
			p-value	**Odds ratio**
All PCGP	94/1120	11/966	5.94×10^{-16}	7.9
PCGP exclude ACT, HYPO	56/1034	11/966	3.04×10^{-8}	5

^: excluding hypodiploid-ALL (HYPO, n=47) and adrenocortical carcinoma (ACT, n=39).

Table S7. Pathogenic (P) or Probably Pathogenic (PP) mutations found in the NDAR autism cohort including 515 cases and 208 controls

A. Pathologenic or probably pathogenic mutations

Subject	Gene	Mutation	Type	Pathogenicity
98HI0605A	*PALB2*	R753*	nonsense	P
03HI2710A	*SDHA*	R31*	nonsense	P
04HI2908A	*APC*	E74_E4splice	splice	PP
04HI3508B	*STK11*	Q8*	nonsense	PP

NF1 and *PTEN* were not included in the analysis as both genes are known to be associated with autism. Suspected false positive calls removed from this analysis include *PTCH1* Y1013* which was found to be false positive in the 1KG data and was also filtered by ExAC. Two splice site variants, *APC* R216_E7splice and *RB1* R445_E14splice, were detected in 3 and 5 individuals with 3 individuals harboring both variants, suggesting a potential systematic variant calling error.

B. Evidence for mutation pathogenicity in Autism Project

PALB2 R753*

- Very rare variant [4 in 120k chromosomes in ExAc]
- Truncating mutation in a gene where LoF is a known mechanism of disease (PMID: 17200671)
- Reported as shared between two women with hereditary breast cancer (PMID: 19763884)
- Recently rated pathogenic of hereditary cancer-predisposing syndrome[ClinVar]

SDHA R31*

- Very rare variant [20 in 121k chromosomes in ExAc]
- Truncating mutation in a gene where LoF is a known mechanism of disease [ACMG:VeryStrong]
- Observed in 5 independent paragliomas [ACMG:Strong]
- Recently rated pathogenic of hereditary cancer-predisposing syndrome[ClinVar]
- Observed in GST patient as part of compound heterozygote (PMID 21505157)

C. Comparison of mutation frequency in PCGP with Autism

Comparison	PCGP§ (mut/all)	Autism§ (mut/all)	Fisher's exact test	
			p-value	Odds ratio
All PCGP	90/1120	4/723	7.36×10^{-16}	15.7
^PCGP exclude ACT, HYPO	52/1034	4/723	8.13×10^{-9}	9.5

^: excluding hypodiploid-ALL (HYPO, n=47) and adrenocortical carcinoma (ACT, n=39).

§: NF1 and PTEN were excluded for summary.

Table S8. Twelve cases with multiple germline pathogenic/probably pathogenic or truncation mutations.

	[1]Pathogenic or Probably Pathogenic		Truncation		
Cases	**Autosomal Dominant**	**Autosomal Recessive**	**Tumor Suppressor**	**Kinase**	**Other**
HGG111	*PMS2* p.Q643* *PMS2* p.K413fs				*PDCD1LG2* p.W221fs *KTN1* p.R1206*
HGG027		*ATM* p.L2312fs p.L2544_E52splice			
EWS010375	*TP53* p.P222L	*MUTYH* p.A379fs			
EPD023	*NF2* p.K149_E4splice	*NBN* p.S418_N419fs			
NBL135	*ALK* p.R1275Q	*FANCA* E1023*			*ECT2L* p.E351*
E2A036	*BRCA2* p.Y3098*	*FANCD2* p.W1268G			
ACT061	*TP53* p.R337H		*FOXO1* p.R269*		
BALL021324	*BRCA1* p.I783fs				*GNAS* p.L166fs
E2A028		*MUTYH* p.V309_E11splice		*FLT3* p.Q494fs	
EPD008	*TP53* p.Y107H				*CBLC* p.V399fs *HOXA9* p.E238*
MLL014					*RNF213* p.R4601fs, *MRE11A* p.L57*
PHALL021					*POLD3* p.S371fs *FIP1L1* p.R298*

[1]Pathogenic or probably pathogenic mutations were classified only for the 60 autosomal dominant genes and 29 autosomal recessive genes

Table S9. Enrichment of truncation mutations in PCGP (n=1,120) compared to The Exome Aggregation Consortium (ExAC, n=60,706).

Genes with enrichment in truncation mutations are shown by descending order of statistical significance based on Fisher's exact test. The level of enrichment for truncation mutations is shown, ranked by descending order of statistical significance based on Fisher's exact test on allele counts. Only genes with odds ratio >1 are included. Correction for multiple testing was performed using FDR, as described in Benjamini, Y. and Hochberg (1995)[33] and implemented in the R adjust function. These enrichment statistics and rankings should be considered exploratory in nature, as per the caveats indicated in Supplemental Methods.

Gene	Category	Rank	Num. in PCGP	Num. in ExAc	original p-value	corrected p-value
TP53	Autosomal Dominant	1	11	11	1.23E-14	1.353E-12
NOP10	Autosomal Recessive	2	2	1	0.0010	0.0394
FCRL4	Other CancerGene	3	4	27	0.0011	0.0394
PML	Tumor Suppressor	4	4	30	0.0028	0.0634
SHROOM2	Other CancerGene	5	3	16	0.0035	0.0634
USP6	Other CancerGene	6	6	91	0.0037	0.0634
WRN	Autosomal Recessive	7	5	72	0.0040	0.0634
ERCC3	Autosomal Recessive	8	4	39	0.0057	0.0782
HOXA9	Other CancerGene	9	2	6	0.0067	0.0820
RB1	Autosomal Dominant	10	3	23	0.0095	0.1045
GNAS	Other CancerGene	11	3	25	0.0132	0.1313
BUB1B	Tumor Suppressor	12	3	32	0.0146	0.1313
CHEK2	Tumor Suppressor	13	4	63	0.0155	0.1313
RECQL4	Autosomal Recessive	14	4	75	0.0199	0.1562
FLT3	Other CancerGene	15	2	17	0.0267	0.1836
KIF5B	Other CancerGene	16	2	19	0.0267	0.1836
PMS2	Autosomal Dominant	17	3	55	0.0350	0.2115
ELANE	Other CancerGene	18	1	1	0.0365	0.2115
FOXO1	Tumor Suppressor	19	1	1	0.0365	0.2115
ACSL6	Other CancerGene	20	2	23	0.0385	0.2117
BRCA2	Autosomal Dominant	21	6	141	0.0410	0.2150
MUTYH	Autosomal Recessive	22	3	51	0.0448	0.2240
ABL1	Kinase	23	1	3	0.0543	0.2597
RNF213	Other CancerGene	24	5	141	0.0662	0.2874
IGF2R	Tumor Suppressor	25	2	27	0.0666	0.2874
SH3GL1	Other CancerGene	26	1	7	0.0717	0.2874
TTL	Other CancerGene	27	1	3	0.0717	0.2874
TSHR	Other CancerGene	28	2	27	0.0771	0.2874
ABL2	Other CancerGene	29	2	31	0.0881	0.2874
GPHN	Other CancerGene	30	1	5	0.0888	0.2874
NF2	Autosomal Dominant	31	1	5	0.0888	0.2874

RUNX1	Autosomal Dominant	32	1	8	0.0888	0.2874
SEPT6	Other CancerGene	33	1	4	0.0888	0.2874
TGFBR1	Other CancerGene	34	1	5	0.0888	0.2874
POLD1	Other CancerGene	35	2	36	0.0938	0.2917
ECT2L	Other CancerGene	36	3	74	0.0955	0.2917
ARID1A	Tumor Suppressor	37	1	7	0.1056	0.2936
MED12	Other CancerGene	38	1	7	0.1056	0.2936
SYK	Tumor Suppressor	39	1	5	0.1056	0.2936
LAMA5	Other CancerGene	40	3	76	0.1068	0.2936
CRTC3	Other CancerGene	41	1	12	0.1221	0.3169
EZR	Other CancerGene	42	1	10	0.1221	0.3169
DDB2	Autosomal Recessive	43	1	9	0.1383	0.3169
ELN	Other CancerGene	44	1	27	0.1383	0.3169
MLLT1	Other CancerGene	45	1	7	0.1383	0.3169
MSH2	Autosomal Dominant	46	1	12	0.1383	0.3169
NSD1	Other CancerGene	47	1	7	0.1383	0.3169
POLD3	Other CancerGene	48	1	11	0.1383	0.3169
ATM	Autosomal Recessive	49	4	137	0.1464	0.3200
NF1	Autosomal Dominant	50	3	84	0.1478	0.3200
CARD11	Other CancerGene	51	1	11	0.1542	0.3200
FIP1L1	Other CancerGene	52	1	18	0.1542	0.3200
KDSR	Other CancerGene	53	1	8	0.1542	0.3200
FANCL	Autosomal Recessive	54	2	48	0.1615	0.3290
EP300	Other CancerGene	55	1	11	0.1698	0.3396
NUP214	Other CancerGene	56	2	52	0.1746	0.3430
PDCD1LG2	Other CancerGene	57	1	13	0.1851	0.3510
TCF12	Other CancerGene	58	1	14	0.1851	0.3510
TERT	Autosomal Recessive	59	1	17	0.2001	0.3731
FANCI	Autosomal Recessive	60	2	68	0.2284	0.3942
CIITA	Other CancerGene	61	1	23	0.2293	0.3942
G6PC3	Other CancerGene	62	1	16	0.2293	0.3942
PPP	Other CancerGene	63	1	15	0.2293	0.3942
NUMA1	Other CancerGene	64	1	17	0.2293	0.3942
OMD	Other CancerGene	65	1	14	0.2435	0.4121
APC	Autosomal Dominant	66	1	21	0.2575	0.4165
CBLC	Other CancerGene	67	1	21	0.2575	0.4165
PTCH1	Autosomal Dominant	68	1	17	0.2575	0.4165
ALDH2	Other CancerGene	69	1	21	0.2712	0.4201
DIP2B	Other CancerGene	70	1	23	0.2712	0.4201
TPR	Other CancerGene	71	1	21	0.2712	0.4201
TRIP11	Other CancerGene	72	2	58	0.2762	0.4220
ETV4	Other CancerGene	73	1	23	0.2846	0.4220
FANCF	Autosomal Recessive	74	1	18	0.2846	0.4220

MRE11A	Other CancerGene	75	1	23	0.2978	0.4220
SH2B3	Autosomal Recessive	76	1	20	0.2978	0.4220
TCF3	Other CancerGene	77	1	26	0.2978	0.4220
ERBB2	Other CancerGene	78	1	27	0.3107	0.4220
FGFR3	Other CancerGene	79	1	30	0.3107	0.4220
FLCN	Tumor Suppressor	80	1	22	0.3107	0.4220
PCSK7	Other CancerGene	81	1	24	0.3107	0.4220
BRIP1	Autosomal Recessive	82	2	68	0.3171	0.4242
IL7R	Other CancerGene	83	1	23	0.3234	0.4242
FANCM	Autosomal Recessive	84	2	68	0.3239	0.4242
HAX1	Other CancerGene	85	1	27	0.3359	0.4247
NCKIPSD	Other CancerGene	86	1	26	0.3359	0.4247
SETD2	Other CancerGene	87	1	23	0.3359	0.4247
KLK2	Other CancerGene	88	1	26	0.3481	0.4303
SDHA	Autosomal Dominant	89	1	28	0.3481	0.4303
PCM1	Other CancerGene	90	1	38	0.3602	0.4402
FANCG	Autosomal Recessive	91	1	35	0.3720	0.4496
KTN1	Other CancerGene	92	1	35	0.3949	0.4671
LIFR	Other CancerGene	93	1	32	0.3949	0.4671
FANCC	Autosomal Recessive	94	1	31	0.4060	0.4702
PINK1	Other CancerGene	95	1	31	0.4060	0.4702
BARD1	Tumor Suppressor	96	1	34	0.4170	0.4778
NBN	Autosomal Recessive	97	1	46	0.4688	0.5316
SSX1	Other CancerGene	98	1	51	0.4786	0.5372
ATIC	Other CancerGene	99	1	49	0.5069	0.5632
ERCC5	Autosomal Recessive	100	1	44	0.5336	0.5870
MSH6	Autosomal Dominant	101	1	47	0.5671	0.6176
PAPB2	Autosomal Dominant	102	1	53	0.5829	0.6286
ERCC2	Autosomal Recessive	103	1	71	0.5906	0.6307

Table S10. Estimate of prevalence of germline cancer predisposition in SEER cohort.
See Supplementary Appendix 2, Table S10.

SUPPLEMENTARY REFERENCES

1. Gruber TA, Larson Gedman A, Zhang J, et al. An Inv(16)(p13.3q24.3)-Encoded CBFA2T3-GLIS2 Fusion Protein Defines an Aggressive Subtype of Pediatric Acute Megakaryoblastic Leukemia. Cancer Cell 2012;22:683-97.
2. Zhang J, Ding L, Holmfeldt L, et al. The genetic basis of early T-cell precursor acute lymphoblastic leukaemia. Nature 2012;481:157-63.
3. Zhang J, Benavente CA, McEvoy J, et al. A novel retinoblastoma therapy from genomic and epigenetic analyses. Nature 2012;481:329-34.
4. Untergasser A, Cutcutache I, Koressaar T, et al. Primer3--new capabilities and interfaces. Nucleic acids research 2012;40:e115.
5. Koressaar T, Remm M. Enhancements and modifications of primer design program Primer3. Bioinformatics 2007;23:1289-91.
6. Liu X, Jian X, Boerwinkle E. dbNSFP: a lightweight database of human nonsynonymous SNPs and their functional predictions. Human mutation 2011;32:894-9.
7. Shah S, Schrader KA, Waanders E, et al. A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. Nature genetics 2013;45:1226-31.
8. Zhang MY, Churpek JE, Keel SB, et al. Germline ETV6 mutations in familial thrombocytopenia and hematologic malignancy. Nature genetics 2015;47:180-5.
9. Lawrence MS, Stojanov P, Polak P, et al. Mutational heterogeneity in cancer and the search for new cancer-associated genes. Nature 2013;499:214-8.
10. Plasschaert E, Descheemaeker MJ, Van Eylen L, Noens I, Steyaert J, Legius E. Prevalence of Autism Spectrum Disorder symptoms in children with neurofibromatosis type 1. American journal of medical genetics Part B, Neuropsychiatric genetics : the official publication of the International Society of Psychiatric Genetics 2015;168B:72-80.
11. Varga EA, Pastore M, Prior T, Herman GE, McBride KL. The prevalence of PTEN mutations in a clinical pediatric cohort with autism spectrum disorders, developmental delay, and macrocephaly. Genetics in medicine : official journal of the American College of Medical Genetics 2009;11:111-7.
12. Antoniou AC, Casadei S, Heikkinen T, et al. Breast-cancer risk in families with mutations inPALB2. The New England journal of medicine 2014;371:497-506.
13. Jones S, Anagnostou V, Lytle K, et al. Personalized genomic analyses for cancer mutation discovery and interpretation. Science translational medicine 2015;7:283ra53.
14. Mavaddat N, Peock S, Frost D, et al. Cancer risks for BRCA1 and BRCA2 mutation carriers: results from prospective analysis of EMBRACE. Journal of the National Cancer Institute 2013;105:812-22.
15. Miki Y, Swensen J, Shattuck-Eidens D, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 1994;266:66-71.
16. Wooster R, Bignell G, Lancaster J, et al. Identification of the breast cancer susceptibility gene BRCA2. Nature 1995;378:789-92.
17. Johnston JJ, Rubinstein WS, Facio FM, et al. Secondary variants in individuals undergoing exome sequencing: screening of 572 individuals identifies high-penetrance mutations in cancer-susceptibility genes. American journal of human genetics 2012;91:97-108.
18. Brooks GA, Stopfer JE, Erlichman J, Davidson R, Nathanson KL, Domchek SM. Childhood cancer in families with and without BRCA1 or BRCA2 mutations ascertained at a high-risk breast cancer clinic. Cancer biology & therapy 2006;5:1098-102.
19. Zhang F, Ma J, Wu J, et al.PALB2 links BRCA1 and BRCA2 in the DNA-damage response. Current biology : CB 2009;19:524-9.
20. Silva AG, Ewald IP, Sapienza M, et al. Li-Fraumeni-like syndrome associated with a large BRCA1 intragenic deletion. BMC cancer 2012;12:237.

21. Armes JE, Trute L, White D, et al. Distinct molecular pathogeneses of early-onset breast cancers in BRCA1 and BRCA2 mutation carriers: a population-based study. Cancer research 1999;59:2011-7.
22. Staff S, Isola JJ, Johannsson O, Borg A, Tanner MM. Frequent somatic loss of BRCA1 in breast tumours from BRCA2 germ-line mutation carriers and vice versa. British journal of cancer 2001;85:1201-5.
23. Osorio A, de la Hoya M, Rodriguez-Lopez R, et al. Loss of heterozygosity analysis at the BRCA loci in tumor samples from patients with familial breast cancer. International journal of cancer Journal international du cancer 2002;99:305-9.
24. Kanchi KL, Johnson KJ, Lu C, et al. Integrated analysis of germline and somatic variants in ovarian cancer. Nature communications 2014;5:3156.
25. Lucas AL, Shakya R, Lipsyc MD, et al. High prevalence of BRCA1 and BRCA2 germline mutations with loss of heterozygosity in a series of resected pancreatic adenocarcinoma and other neoplastic lesions. Clinical cancer research : an official journal of the American Association for Cancer Research 2013;19:3396-403.
26. Willems AJ, Dawson SJ, Samaratunga H, et al. Loss of heterozygosity at the BRCA2 locus detected by multiplex ligation-dependent probe amplification is common in prostate cancers from men with a germline BRCA2 mutation. Clinical cancer research : an official journal of the American Association for Cancer Research 2008;14:2953-61.
27. Cavalli LR, Singh B, Isaacs C, Dickson RB, Haddad BR. Loss of heterozygosity in normal breast epithelial tissue and benign breast lesions in BRCA1/2 carriers with breast cancer. Cancer genetics and cytogenetics 2004;149:38-43.
28. Dworkin AM, Spearman AD, Tseng SY, Sweet K, Toland AE. Methylation not a frequent "second hit" in tumors with germline BRCA mutations. Familial cancer 2009;8:339-46.
29. Bell DW, Gore I, Okimoto RA, et al. Inherited susceptibility to lung cancer may be associated with the T790M drug resistance mutation in EGFR. Nature genetics 2005;37:1315-6.
30. Pui CH, Carroll WL, Meshinchi S, Arceci RJ. Biology, risk stratification, and therapy of pediatric acute leukemias: an update. Journal of clinical oncology : official journal of the American Society of Clinical Oncology 2011;29:551-65.
31. Holmfeldt L, Wei L, Diaz-Flores E, et al. The genomic landscape of hypodiploid acute lymphoblastic leukemia. Nature genetics 2013;45:242-52.
32. Chang YL, Wu CT, Shih JY, Lee YC. Comparison of p53 and epidermal growth factor receptor gene status between primary tumors and lymph node metastases in non-small cell lung cancers. Annals of surgical oncology 2011;18:543-50.
33. Benjamini YaH, Y. . Controlling the False Discovery Rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society B 1995;57:289-300.