

SCIENTIFIC REPORTS



OPEN

Familial Cancer Variant Prioritization Pipeline version 2 (FCVPPv2) applied to a papillary thyroid cancer family

Abhishek Kumar¹, Obul Reddy Bandapalli¹, Nagarajan Paramasivam^{2,3}, Sara Giangioffe¹, Chiara Diquigiovanni⁴, Elena Bonora⁴, Roland Eils^{2,5}, Matthias Schlesner^{2,6}, Kari Hemminki^{1,7} & Asta Försti^{1,7}

Whole-genome sequencing methods in familial cancer are useful to unravel rare clinically important cancer predisposing variants. Here, we present improvements in our pedigree-based familial cancer variant prioritization pipeline referred as FCVPPv2, including 12 tools for evaluating deleteriousness and 5 intolerance scores for missense variants. This pipeline is also capable of assessing non-coding regions by combining FANTOM5 data with sets of tools like Bedtools, ChromHMM, Miranda, SNPnexus and TargetsCan. We tested this pipeline in a family with history of a papillary thyroid cancer. Only one variant causing an amino acid change G573R (dbSNP ID rs145736623, NM_019609.4:exon11:c.G1717A:p.G573R) in the carboxypeptidase gene *CPXM1* survived our pipeline. This variant is located in a highly conserved region across vertebrates in the peptidase_M14 domain (Pfam ID PF00246). The *CPXM1* gene may be involved in adipogenesis and extracellular matrix remodelling and it has been suggested to be a tumour suppressor in breast cancer. However, the presence of the variant in the ExAC database suggests it to be a rare polymorphism or a low-penetrance risk allele. Overall, our pipeline is a comprehensive approach for prediction of predisposing variants for high-risk cancer families, for which a functional characterization is a crucial step to confirm their role in cancer predisposition.

Oncogenomics has been boosted with rapid advancements in the next-generation sequencing (NGS) technologies in the last 10 years with large consortia describing several thousands of somatic variants. However, there has been far less success in the discovery of new cancer predisposing genes (CPGs) as only some new genes were identified using germline genome sequencing¹. The major fraction of CPGs were discovered in the late 1990s using the familial linkage analysis¹. Since then the interest in collecting cancer pedigrees diminished and the consequence has been that whole exome/genome sequencing (WES/WGS)-based family studies have been forced to resort either to small pedigrees or single cases from affected families². In cancer studies both types of approaches have been used^{3–6}. CPGs found include for example an *NTHL1* variant in colorectal cancer³ and an *RECQL* variant in breast cancer⁴.

As pedigree-based studies have a high discriminatory power if samples from many affected and unaffected members are available, we recently introduced the familial cancer variant prioritization pipeline (FCVPP⁵), which is a pipeline capable of detecting rare germline variants and their corresponding CPGs. In the current study, we describe FCVPPv2, an upgraded version of our FCVPP pipeline⁵. FCVPPv2 prioritizes rare deleterious and regulatory germline variants, both in the coding and non-coding region for cancer families. The advantages of

¹Division of Molecular Genetic Epidemiology, German Cancer Research Center (DKFZ), D69120, Heidelberg, Germany. ²Division of Theoretical Bioinformatics, German Cancer Research Center (DKFZ), D69120, Heidelberg, Germany. ³Medical Faculty Heidelberg, Heidelberg University, D69120, Heidelberg, Germany. ⁴Unit of Medical Genetics, S.Orsola-Malpighi Hospital, 40138, Bologna, Italy. ⁵Department of Bioinformatics and Functional Genomics, Institute of Pharmacy and Molecular Biotechnology (IPMB) and BioQuant, Heidelberg University, D69120, Heidelberg, Germany. ⁶Bioinformatics and Omics Data Analytics, German Cancer Research Center (DKFZ), D69120, Heidelberg, Germany. ⁷Center for Primary Health Care Research, Lund University, Malmö, Sweden. Abhishek Kumar and Obul Reddy Bandapalli contributed equally to this work. Correspondence and requests for materials should be addressed to A.K. (email: a.kumar@dkfz.de) or O.R.B. (email: o.bandapalli@dkfz.de)

this approach are several-fold such as (a) reducing the large number of variants through the pedigree segregation step; (b) assessing the deleterious nature of missense variants by a combination of 12 ranking tools and 5 intolerance scores; (c) analysis of non-coding variants by specialized tools such as Miranda⁷ and Targetscan 7.0⁸ for 3' UTR variants, SNPnexus⁹ for 5' UTR variants and FANTOM5 and SNPnexus⁹ for variants in enhancers¹⁰ and promoters^{11,12} and (d) this pipeline also takes advantages of the improvements of population frequencies in public databases, which assists inferring rarity of a particular variant. Herein, we present the improvements of FCVPPv2 as well as its application to a pedigree of papillary thyroid cancer (PTC).

Results

Familial cancer variant prioritization pipeline version 2 (FCVPPv2) comes with several improvements.

Overview of FCVPPv2 is provided in Figs 1 and 2 and description of this work involves several abbreviations of scientific words and tools as summarized in Supplementary Table S1. FCVPPv2 has implemented platypus tool¹³ for joint variant calling. It combines several resources for variant annotation like ANNOVAR¹⁴, exome aggregation consortium (ExAC¹⁵), exonic variant server with 6500 samples (EVS6500¹⁶), 1000 Genomes phase III¹⁷, dbNSFP v3.0¹⁸, and dbSNP¹⁹.

Variants are filtered with criteria using read quality parameters like coverage and quality scores (QUAL) must be >5 and >20, respectively. FCVPPv2 uses minor allele frequency (MAF) below 0.1% in the European populations in all used databases. The hallmark of FCVPPv2 is use of family-pedigree variant screening (Fig. 1). Post this step, CADD v1.3²⁰ based variant filtering is performed with PHRED CADD score of 10 as a cut-off. Subsequently, deleterious variants are segregated on their locations. The coding variants are considered deleterious based on the consensus from 5 intolerance scores, 3 conservation scores and 12 deleteriousness prediction tools (Fig. 1, Tables 1 and 2). FCVPPv2 characterizes the regulatory nature in the UTR regions by combining a set of tools like the Haploreg V4.1²¹, RegulomeDB²² and SNPnexus⁹ for variants in 5' UTR and miRNA binding sites using Miranda⁷ and Targetscan 7.0⁸ tools for variants in 3' UTRs (Fig. 1).

Above 98% of the human genome is non-coding and dealing with non-coding variants is a daunting task. No tools can accurately predict the regulatory nature of non-coding variants. To overcome this issue, FCVPPv2 uses a combination of the several standard tools like chromHMM, Segway, FunSeq2²³ and FANTOM5 data (Fig. 2). FCVPPv2 focuses the FANTOM5 data by matching variants within promoters and/or enhancers using Bedtools intersect function. FCVPPv2 makes assessments of transcription factor (TF) binding sites using SNPnexus⁹. This pipeline makes use of signals for chromatin binding using ChromHMM and genomic segmentation data from Segway via CADDv1.3. Additionally, FCVPPv2 checks if a putative non-coding regulatory variant is localized in the ultra-conserved non-coding elements (UCNEs) or ultra-conserved genomic regulatory blocks (UGRBs) with the help of UCNEbase²⁴, and in ultra-sensitive and sensitive regions (Ultrasen), defined by FunSeq2²³. FCVPPv2 uses Bedtools intersect function to assign variants in regions of super-enhancers using super-enhancer archive (SEA)²⁵ and dbSUPER²⁶ databases.

FCVPPv2 visualises top ranked variants for their regulatory natures using different genome browsers like Locuszoom²⁷, SNIpa²⁸, the UCSC²⁹ and ZENBU³⁰ (Fig. 1). The status of RNA and protein expression for genes carrying potential deleterious variants is examined with the help of FANTOM5, BioGPS³¹ and Human Protein Atlas³² (Fig. 1). This tool uses literature mining to check if these variants are found in known lists of cancer predisposing genes (CPGs¹). Towards end, FCVPPv2 provides a summary of potential variants by combining features from several databases like ClinVar³³, Online Mendelian Inheritance in Man (OMIM, <https://omim.org/>), CanVar Browser³⁴, cBioPortal³⁵, COSMIC³⁶, ICGC³⁷ and IntoGen³⁸ (Fig. 1).

Application of FCVPPv2 to a papillary thyroid cancer (PTC) family.

We used FCVPPv2 to a family with PTC with two distantly related cases (sample no. 2 and 3) and one unrelated case (sample no. 1, Fig. 3A). We ranked top variants for this family after WGS of the three samples. After variant annotation and removal of variants with MAF higher than 0.1% in at least one variant database, a total of 120,323 variants remained (Fig. 3B & Supplementary Table S2). We applied pedigree-filtering criteria imposing that the two related cases must have the variant, while the unrelated case should not have it. With this pedigree filtering approach we narrowed down the potential list of variants to 1970, which is about 1.6% of the initial variant set (Fig. 3B). Subsequently, we examined these variants based on their location in the genome and we found that only 28 of them were located in the coding region, which is 1.4% of all pedigree segregated variants, while 98.6% of them were located in the non-coding region with 1015 intronic, 901 intergenic and 26 up- or down-stream variants (Fig. 3B). Out of the 28 coding variants only 7 had a CADD PHRED score >10 and these included 5 exonic (4 non-synonymous and 1 non-frameshift insertion) and two UTR (1 in 3' UTR and 1 in ncRNA_UTR) variants (Fig. 3B and Table 3). The non-frameshift insertion and the UTR variants (Table 3) were indels localized in the repetitive regions in the human genome and hence these were not considered further. Of the non-coding variants, none was located within a predicted enhancer or promoter.

We focused on the 4 non-synonymous variants localized in four different genes-*C1orf27* (chromosome 1 open reading frame 27), *CPXMI* (carboxypeptidase X (M14 family), member 1), *FAM129A* (family with sequence similarity 129, member A) and *ZBTB41* (zinc finger and BTB domain containing 41) (Figs 3B and 4A). All these 4 variants had CADD score >20, which indicates top 1% deleterious variants in the human genome. The highest CADD score of 32 was reported for *CPXMI* ranking it to the list of top 0.1% deleterious variants. We prioritized the variant G573R in *CPXMI* as our top candidate, as it had very high conservation scores (GERP++ = 5.3, phastCons = 1.0, and phyloP = 7.6), and 9 out of the 12 deleterious prediction scores and 4 out of the 5 intolerance scores were favoring it. Also amino acid change from glycine to arginine is critical with a Grantham score of 125 (moderately radical, deduced from CADD annotation²⁰), while the remaining three variants had a low Grantham score (<30). Additionally, although the three other variants had a CADD score >20, mutations in *C1orf27* and *FAM129A* were classified as tolerated by all 5 intolerance scoring tools (listed in Table 1) and the *ZBTB41* variant

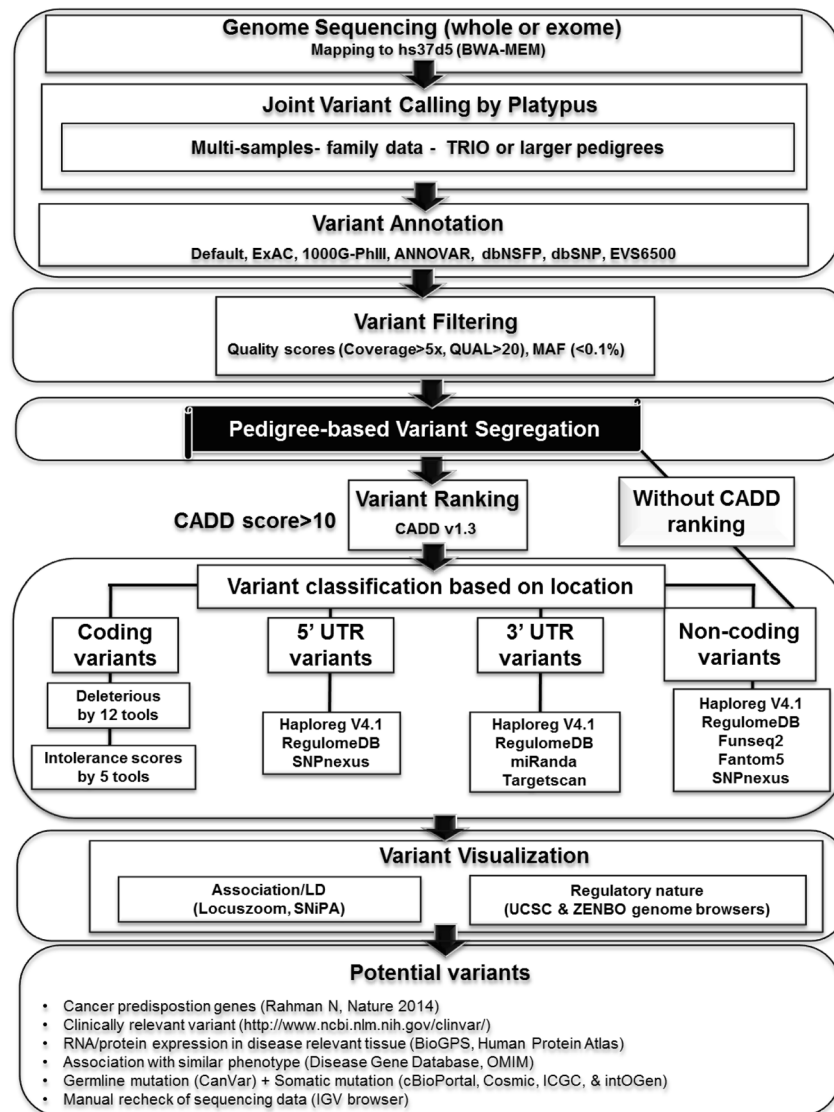


Figure 1. Summary of familial cancer variant prioritization pipeline version 2 (FCVPPv2). This pipeline uses platypus tool¹³ for joint variant calling after mapping of the sequencing reads from cases and controls. FCVPPv2 uses several external tools for variant annotation namely ExAC, 1000 Genomes phase III data, ANNOVAR and dbNSFPv3, dbSNP and EVS6500. For candidate variants the variants are filtered using read quality parameters like coverage and quality scores (QUAL) must be > 5 and > 20, respectively. Minor allele frequency (MAF) must be below 0.1% in the European populations in all used databases. Furthermore, these variants are screened with respect to family-pedigree and this is the most critical step in the germline genomics (shown in black shade). After this step, variants are ranked with the help of CADD v1.3²⁰ and any variants with CADD PHRED score of > 10 belongs to top 10% for probable functional and deleterious variants in the human genome. These deleterious variants are subsequently divided into 4 different categories based on their locations. The coding variants are considered deleterious based on the consensus from 12 deleteriousness prediction tools and 5 intolerance scores. Variants in the 5' UTR are considered regulatory based on the Haploreg V4.1²¹, RegulomeDB²² and SNPnexus⁹ while variants in the 3' UTR are regulatory if supported by the presence of miRNA binding site using Miranda⁷ and Targetscan 7.0⁸ tools and additional hints are received from Haploreg V4.1²¹ and RegulomeDB²². For variants in the non-coding segments we combined several state-of-the-art tools such as chromHMM, Segway, FunSeq2 and FANTOM5 data. Non-coding (intergenic and intronic) variants may not always have CADD > 10 even though they will have regulatory implications, so we analyzed all non-coding variants after pedigree segmentation, either with or without CADD > 10. Putative deleterious or regulatory variants are visualized using Locuszoom, SniPA and UCSC genome browser. Potential variants are also checked with sets of additional features, e.g. list of known CPGs¹ and clinically relevant variants (ClinVar), expression data and somatic mutations. We also checked the sequencing data of all cases and controls in a particular family for correctness using the IGV browser.

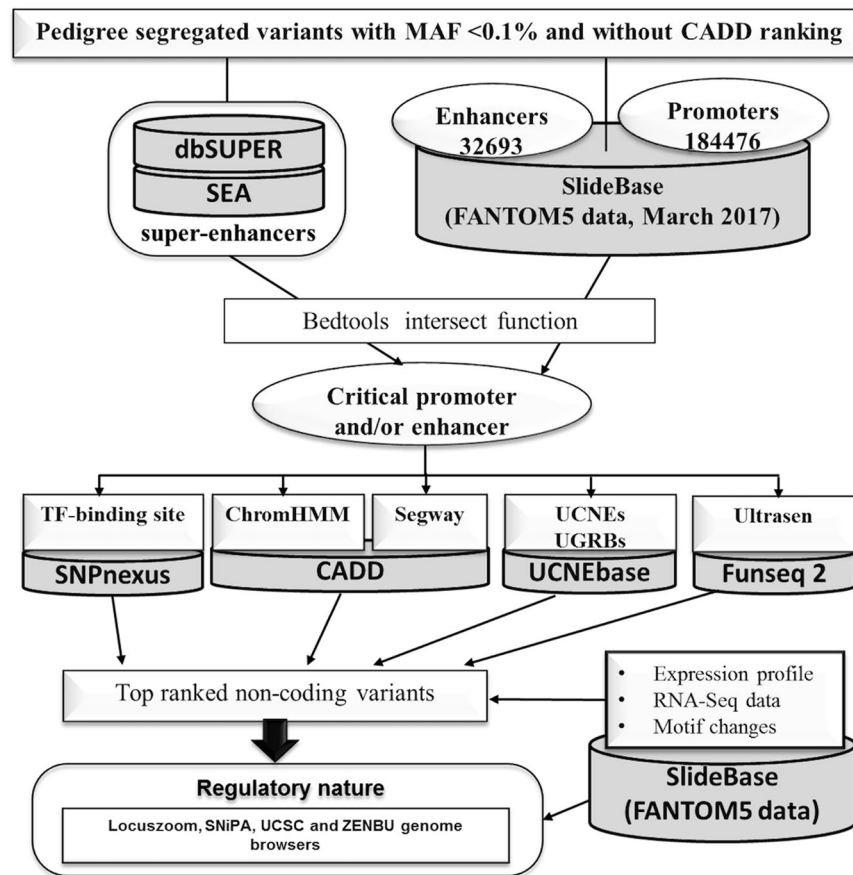


Figure 2. Overview of strategies for regulatory variant detection in the non-coding segments of the human genome. We utilized the FANTOM5 data using the SlideBase Tool (slidebase.binf.ku.dk) with 32693 enhancers and 184476 promoters (downloaded in March 2017). We matched our variants (pedigree segregated) with FANTOM5 data using Bedtools intersect to retrieve a list of potentially critical variants localized within promoters and/or enhancers, and we examined the status of transcription factor (TF) binding sites using SNPnexus⁹. We checked the signals for chromatin binding using ChromHMM and genomic segmentation data from Segway via CADDv1.3²⁰. In addition, we examined if the putative noncoding regulatory variants were localized in the ultra-conserved non-coding elements (UCNEs) and their clusters, also known as ultra-conserved genomic regulatory blocks (UGRBs) with the help of UCNEbase²⁴ and also if these variants were located in ultra-sensitive and sensitive regions (Ultrasen), defined by FunSeq2²³. The top-ranked variants were examined for their regulatory nature by using Locuszoom, SniPA, UCSC and ZENBU genome browsers. We also examined if the putative enhancer variant fall into the category of super-enhancers using super-enhancer archive (SEA)²⁵ and dbSUPER²⁶. Expression profile, RNA-seq data-based information and motif changes and disruptions were gathered with help from FANTOM5 data via the SlideBase.

was only predicted to be deleterious by 3 out of 12 tools (Fig. 4A). All in all, we found one predicted deleterious missense variant in the *CPXM1* gene, which was not found in any other 77 cancer families (including 4 PTC families), we have whole-genome sequenced by now. We only found two other missense variants in the entire *CPXM1* gene in three different families out of the 77 cancer families, each present either in only one case or only one control of the family (Supplementary Table S3). During the course of this study, the NM_019609.4:exon11:c.G1717A:p.G573R variant was identified in one colorectal cancer patient (Supplementary Table S4) as reported in the CanVar Browser³⁴ (a database of genetic variants of 1,006 early-onset familial colorectal cancer cases⁶). We also found a rare stop gained variant and 3 more frequent missense variants in this gene in the CanVar Browser³⁴ (Supplementary Table S4). Currently, the G573R variant (dbSNP ID - rs145736623) is also listed in the ExAC database with a frequency of 0.0004 in the total population (0.0006 in the European, non-Finnish population). Taken together, as we found the NM_019609.4:exon11:c.G1717A:p.G573R variant in 2 distantly related cases of our PTC family, it may be a low-penetrance allele predisposing to PTC, but it may also be a rare polymorphism.

Characterization of the *CPXM1* variant and potential roles of the *CPXM1* gene. Our familiar PTC candidate variant (NM_019609.4:exon11:c.G1717A:p.G573R) is located at the end of exon 11 of the *CPXM1* gene, which is composed of 14 exons (Fig. S1A). To confirm the accuracy of variant calling, we examined the genomic data of all sequenced samples using Integrative Genomics Viewer (IGV)³⁹, reassuring that this *CPXM1* variant is only present in the two related cases (sample no. 2 and 3, Fig. S1B), but not in the unrelated case (sample no. 1).

Tools	Details	Score Range	Significant score	Ref.
Residual Variation Intolerance Score (RVIS) RVIS - NHLBI-ESP6500 data set	based upon allele frequency data	Negative to Positive	RVIS < 0 - intolerant RVIS > 0 - tolerant	60
RVIS - ExAC data set				
RVIS - local data set				
pLI score	Developed by ExAC Consortium for Loss-of-Function (LoF) mutations		pLI ≥ 0.9 - highly LoF-intolerant pLI ≤ 0.1 - LoF tolerant	15
Z-score	Developed by ExAC Consortium for missense and synonymous variants		Positive Z scores -intolerant Negative Z scores -tolerant	15
Genomic Evolutionary Rate Profiling (GERP)		-12.3 to 6.17	>2	61
PhastCons		0 to 1	>0.3	62
Phylogenetic P-value (PhyloP)		-14 to +6	≥3.0	63

Table 1. Summary of intolerance scores and conservational scores.

The human *CPXM1* gene (also known as *CPX1*) encodes a zinc metallocarboxypeptidase. The human CPXM1 protein is 734 amino acids long (Fig. 4B). Upon Pfam domain scanning, we found that the CPXM1 protein is composed of three Pfam domains, namely F5_F8_type_C (PFam ID - PF00754), peptidase_M14 (PF00246) and carboxypepD_reg (PF13620) located at amino acid positions 132–271, 306–611 and 625–688, respectively. The CPXM1 variant (NM_019609.4:exon11:c.G1717A:p.G573R) is localized in the peptidase_M14 domain (Fig. 4B). Upon examination the 25 amino acids flanking the variant position, we found the G573 residue and the majority of the flanking residues to be highly conserved from human to zebrafish (Fig. 4C). The CPXM1 protein has no carboxypeptidase activity but it is a secreted N-glycoprotein that binds collagen⁴⁰. It has been reported to be involved in adipogenesis through extracellular matrix remodeling⁴¹. Peroxisome proliferator-activated receptor gamma (PPAR γ) is the master adipogenic regulator and it may promote growth and invasion of undifferentiated thyroid cancer (TC) cells⁴². Whether *CPXM1* may predispose to differentiated TC by acting as a complementary regulator to PPAR γ in adipogenesis or through extracellular matrix remodeling, remains to be discovered⁴¹. There is experimental evidence indicating that expression of *CPXM1* is epigenetically regulated in breast cancer and it may act as a tumor suppressor gene^{43,44}. Recently, an indirect role for *CPXM1* in PTC was also illustrated, as *CPXM1* was downregulated by a long non-coding RNA (lncRNA, Ensembl ID - ENSG00000273132.1)⁴⁵.

Discussion

The main theme for understanding the germline cancer genetics is the identification of pathogenic mutations and genes predisposing to cancer. Rapid improvements in scientific and technological aspects of genomics have contributed to revolutionary changes in cancer genetics in particular in cancer treatment but also in cancer risk assessment, cancer screening and prevention, thus setting up a milestone for approaching towards personalized medicine⁴⁶. With these advancements and decreasing costs, WES/WGS has become the state-of-the-art tool for identifying susceptibility loci in several types of Mendelian diseases². There are a handful of successful reports on disease gene identification for cancer syndromes such as *TERT* promoter mutation⁴⁷ and *POT1* mutations⁴⁸ in familial melanoma, *POLE*, *POLD1*⁴⁹ and *FAN1*⁵⁰ mutations in familial colorectal cancer and *KDR* mutation⁵¹ in familial Hodgkin lymphoma. However, delivering one out of several millions of human genetic variants as the main cause of hereditary cancer is a daunting computational task². Recently, we developed a pipeline for this purpose, which is known as the FCVPP⁵. We applied several types of improvements to this pipeline while working with different cancer families and now we are representing the second version of this pipeline as FCVPPv2. FCVPPv2 can deal with missense variants in a more sophisticated way by employing 12 deleteriousness assessment tools and 5 intolerance scores. Additionally, it has the capability of dealing with non-coding variants by the use of data from FANTOM5, super-enhancer databases, UCNEbase and FunSeq2, and without CADD filter as CADD may not pick up many non-coding positions as deleterious.

As an example we showed our experience in hunting predisposing genes for PTC in a high risk PTC family. Our approach involved WGS of germline DNA from several affected and unaffected family members. As a consequence, we found that only 1 out of 61 variants (1970 variants out of 120,323 rare variants remained after pedigree segregation) was able to pass the pedigree segregation filter (Fig. 3B). Hence, pedigree-based variant filtering is a highly effective way for filtering out non-causative variants; in this case only <1.6% of the rare variants survived this step.

We identified 28 variants out of 1970 (1.4%) as coding region variants, which is close to the proportion of coding regions of the whole genome (2%). A single coding variant, *CPXM1* (NM_019609.4:exon11:c.G1717A:p.G573R), passed our filtering criteria, CADD score >10, location in a highly conserved region, deleterious in >60% of the prediction tools and intolerance tools predicting it to be deleterious. We had three indel variants in the coding and untranslated regions but they were in repetitive genomic segments and hence were not considered for final prioritization. Furthermore, none of the 1942 non-coding variants were predicted to be located within a promoter or an enhancer. Literature search gave some evidence about a potential function of the *CPXM1* gene in cancer. *CPXM1* may serve as a tumour suppressor in breast cancer^{43,44}, potentially through involvement in adipogenesis or extracellular matrix remodelling^{40,41}, and it is reported to be down-regulated by a lncRNA in PTC⁴⁵. Our study identified one predicted pathogenic mutation located in the peptidase M14 domain of the CPXM1 protein in the PTC family, yet this domain is inactive in CPXM1 protein and lacks a typical carboxypeptidase

Tools	Methodology	Score ranges	Prediction	References
Sorting Intolerant from Tolerant (SIFT)	Position-specific scoring matrix (PSSM) with Dirichlet priors Sequence based. uses PSI-BLAST	0 to 1*	D - Damaging (<0.05) T - Tolerated (>0.05)	64
Polymorphism Phenotyping version-2 (PolyPhen-v2)	Naïve Bayes classifier trained using supervised machine-learning Sequence and structure based	0 to 1**	D - probably damaging (0.957-1) P - possibly damaging (0.453-0.956) B - benign (0.00-0.452)	65
PolyPhen2_HDIV (HumDiv [§])				
Polyphen2_HVAR (HumVar [¶])				
Log ratio test (LRT)	Uses log ratio test Sequence based	0 to 1***	D - Deleterious N - Neutral U - Unknown	69
MutationTaster	Naïve bayes model operated on the integrated data source Based on sequence and annotation.	0 to 1**	A - disease_causing_automatic D - disease_causing (>0.5) N - polymorphism (<0.5) P - polymorphism_automatic	70
MutationAssessor	Multiple sequence alignment (MSA) and conservation scores	-5.135 to 6.49**	H - High L - Low M - Medium N - Neutral	71
Functional Analysis Through Hidden Markov Models (FATHMM)	Hidden Markov models (HMM) Based on sequences and protein domains	-18.09 to 11.0*	D - Damaging (<= -1.5) T - Tolerated (>-1.5)	72
MetaSVM	Support vector machine (SVM) based score, derived by incorporating different scores [‡]	-2 to 3**	D - Damaging (>0) T - Tolerated (<0)	18
MetaLR	Logistic regression (LR) based score, derived by combining different scores [‡]	0 to 1**	D - Damaging (>0.5) T - Tolerated (<0.5)	18
Variant Effect Scoring Tool version 3 (VEST3)	Supervised machine learning-based method Combines conservalational and structural features	0 to 1**	NA	73
Protein Variation Effect Analyzer (PROVEAN)	Pair-wise alignment-based scoring method	-14 to 14*	D - Damaging (<= -2.5) N - Neutral (>-2.5)	74
Reliability index (RI)	SVM based Combines protein sequence and structural features	0 to 10**	D - Damaging (≥5) N - Neutral (<5)	75

Table 2. Summary of used tools for deleteriousness prediction for missense variants. *Lower scores indicate deleterious nature. **Higher scores indicate deleterious nature. ***Score cannot decide deleterious nature. [§]HumDiv - collection of mendelian disease variants (5564 deleterious + 7539 neutral in 978 human protein) against divergence from close mammalian homologs of human proteins ($\geq 95\%$ sequence identity). [¶]HumVar - compilation of all human variants (22196 deleterious + 21119 neutral) associated with some disease (non-cancer mutations) or loss of activity/function vs. common (MAF > 1%) human polymorphism with no reported association with a disease. [‡]10 scores from SIFT, PolyPhen-2 HDIV, PolyPhen-2 HVAR, GERP++, MutationTaster, Mutation Assessor, FATHMM, LRT, SiPhy and PhyloP and the maximum frequency observed in the 1000 G data.

function. Recent sequencing data on colorectal cancer and ExAC databases suggest that the identified variant may be a rare polymorphism. Alternatively it may be a low-penetrance PTC predisposing variant, as it was found in two distant relatives, with no other known PTCs in the family. Lack of functional data of the *CPXM1* gene complicates further evaluation of its function and the pathways it is involved. As tumour samples from the mutation carriers of the PTC family are not available, we cannot explore the tumour suppressor nature of *CPMX1* in PTC.

In general, our understanding of human genes and their roles in human diseases, including cancer, are still limited. Lack of proper annotation and unknown physiology hampers mechanistic groundwork for candidate variants. Genes involved in known pathways and with more information in the literature are more likely to be studied further than genes with little or no functional characterization. Without this knowledge or convincing segregation data there may be doubts to accept the detected genes as new bona fide tumor suppressor genes. Recent findings on *HABP2* gene in nonmedullary thyroid cancer⁵², deubiquitinating enzyme coding *BAP1* in multiple cancers⁵³, ovarian cancer gene 1 (*OVCA1*)⁵⁴, promyelocytic leukemia protein (*PML*)⁵⁵ and regucalcin (*RGN*)⁵⁶ are offered by variable levels of supporting functional evidence. From these, the *HABP2* mutation⁵² was later shown to be a common polymorphism⁵⁷. Another problem faced by the WES/WGS results in Mendelian diseases is also highlighted in our study: the potentially pathogenic mutation was found only in one family².

An estimate from 2012 suggested that WES/WGS studies in Mendelian diseases have a success rate of about 60–80%². However in cancer only handful of novel CPGs were found through WGS/WES⁵⁸. We think that cancer is more complex than other Mendelian diseases because carcinogenesis is the interplay of germline and somatic events in the form of tumour growth.

In this study, we provide details of a variant prioritization pipeline FCVPPv2 for gene identification in high-risk cancer families applying pedigree segregation-based variant filtering and variant prioritization using state-of-the-art bioinformatics tools and databases. This pipeline detected a predicted deleterious variant in the

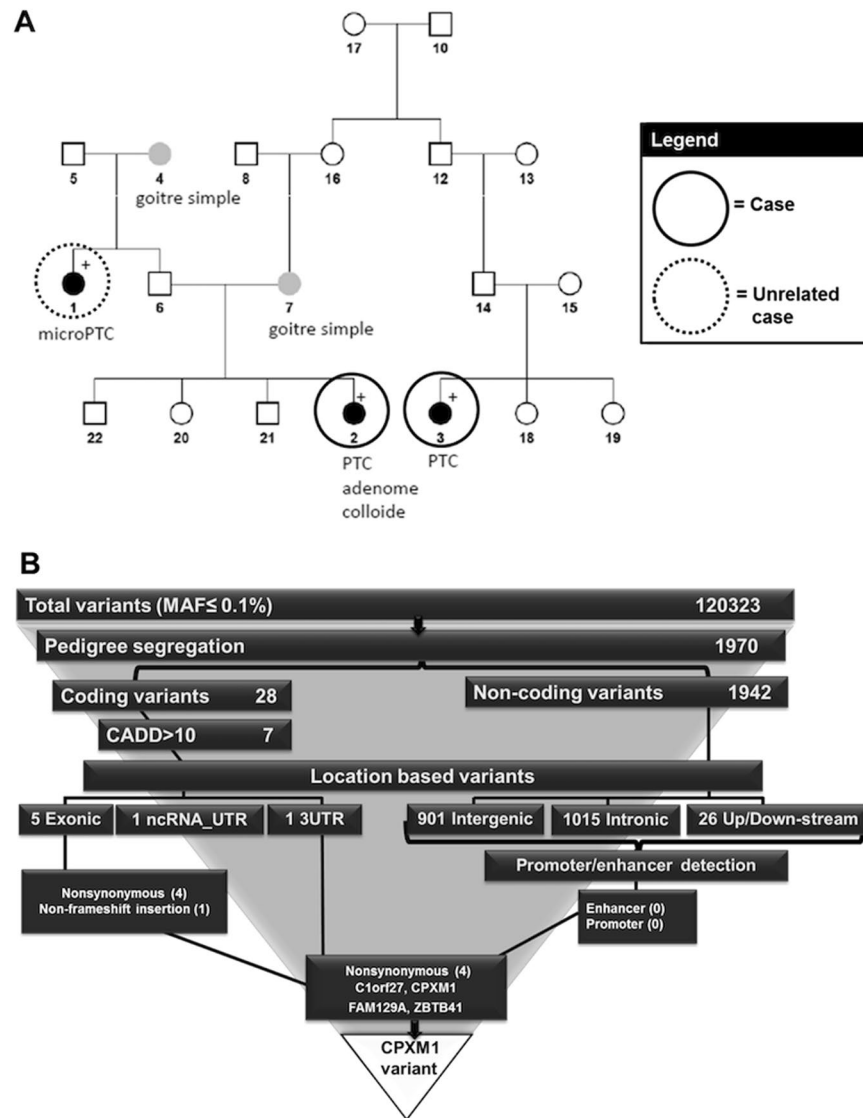


Figure 3. Summary of the papillary thyroid cancer (PTC) family and variant ranking within this family. (A) Pedigree of the PTC family. (B) Variant ranking for the PTC family and selection of CPXM1 variant as the top deleterious variant.

CPMX1 gene in a PTC family. However, as recent sequencing efforts have found the variant in both the ExAC population and a colorectal cancer family member, experimental validation of the identified variant and functional characterization of the gene are necessary for mechanistic understanding and evaluation of the potential cancer predisposing nature of CPMX1.

Material and Methods

Ethic permissions for the experiments from two committees. All experimental protocols were approved by two ethical committees namely “Comitato Etico Indipendente dell’Azienda Ospedaliero-Universitaria di Bologna, Policlinico S. Orsola-Malpighi (Bologna, Italy)” and “comité consultatif de protection des personnes dans la recherche biomédicale, Le centre de lutte contre le cancer Léon-Bérard (Lyon, France)”. Sample collection was carried out in accordance with relevant guidelines and regulations of these two committees. As per guidelines of these two committees, we obtained informed consents of patients for use of their blood samples and the DNA extraction from bloods.

Whole-exome/genome sequencing, mapping, variant calling, filtering and annotation. WES/WGS of the cases and controls from different families considered into the current study was performed in the Illumina X10 platform using DNA extracted from the blood samples. WES/WGS was performed as a paired-end sequencing with a read length of 150 bp. Mapping of reads to reference human genome (assembly version hs37d5) was performed using BWA-0.7.8.r2.05 mem (convey and alignment parameter as $-T 0$)⁵⁹ and duplicates were removed using bammarkduplicates-0-0.148. Variants were called by using Platypus-0.8.1¹³ (with details as

Gene Name	Gene Description	Variant	Variant nomenclature ^s	Variant type	No. of cases	No. of unknown cases	ANNOVAR Annotation	Exonic Classification	CADD score
C1orf27	chromosome 1 open reading frame 27	1_186355211_G_A	NM_017847.5:exon4:c.G326A:p.R109H	SNVs	2	0	exonic	nonsynonymous SNV	25.1
FAM129A	family with sequence similarity 129, member A	1_184792402_T_C	NM_052966.3:exon8:c.A884G:p.K295R	SNVs	2	0	exonic	nonsynonymous SNV	23.9
ZBTB41	zinc finger and BTB domain containing 41	1_197128680_C_T	NM_194314.2:exon10:c.G2539A:p.D847N	SNVs	2	0	exonic	nonsynonymous SNV	23.1
CPXM1	carboxypeptidase X (M14 family), member 1	20_2776248_C_T	NM_019609.4:exon11:c.G1717A:p.G573R	SNVs	2	0	exonic	nonsynonymous SNV	32
KCNE3	potassium voltage-gated channel, Isk-related family, member 3	11_74167200_AATAT_A	NM_005472.4:c.*1097-1097delATAT	Indel	2	0	ncRNA_UTR3	.	11
AR	androgen receptor	X_66765158_T_TGCAGCAGCA	NM_000044.3:c.239_240insGCAGCAGCA	Indel	2	0	exonic	nonframeshift insertion	12.8
NLK	glucose-6-phosphate isomerase	17_26522009_T_TCACA	NM_016231.4:c.*347_*348insCACA	Indel	2	0	UTR3	.	11.7

Table 3. Overview of the 7 top-ranked germline variants detected in the PTC family. \$ - as per guidelines of the Human Genome Variation Society (HGVS, website <http://www.hgvs.org/>).

Platypos-0.8.1.py callVariants —genIndels = 1 —genSNPs = 1 —minFlank = 0). Variants were annotated using ANNOVAR¹⁴, dbNSFP v3.0¹⁸, 1000 Genomes phase III¹⁷, dbSNP¹⁹ and ExAC¹⁵. Variant filtering was performed with considering the quality score >20 and coverage of minimum 5 reads. Minor allele frequency (MAF) of 0.1% was used with respect to population databases (the 1000 Genomes phase III¹⁷, non-TCGA exome aggregation (ExAC) data¹⁵, and local data sets). A pairwise comparison of variants among the cohort was performed to check for sample swaps and family relatedness.

Segregation in Pedigrees. The variants were filtered separately in each family based on the pedigree data by considering cancer patients as cases and unaffected persons as controls, and by applying to each individual a probability consideration for being a Mendelian case or a true control. However, a rule of thumb was that a maximum number of cases and a minimum number of controls in each family must carry the variant.

Variant ranking using combined annotation dependent depletion (CADD). After pedigree segregation, variants were prioritized using the CADD tool v1.3²⁰ with the scaled PHRED-like CADD score greater than 10, which accounts for top 10% of probable deleterious variants in the human genome. Similarly the scores >20 and >30 are indicative of the top 1% and top 0.1% of deleterious variants, respectively²⁰. All variants with CADD score >10 were taken into further consideration.

Screening genic variants using intolerance score. Intolerance score ranks genes based on their capabilities to possess more or less common functional genetic variation relative to the genome wide expectation. Hence, genes, which have less common functional variation, are referred to as “intolerant” genes while genes prone to have more variations are called as “tolerant” genes. Tools were developed for calculating intolerance scores as reported in Table 1. Initially, residual variation intolerance score (RVIS) was developed based on allele frequency data as represented in whole exome sequence data from the NHLBI-ESP6500 data set⁶⁰. We used this scoring system with ExAC¹⁵ and with our local datasets to create two additional intolerance scores, which were included into the prioritization process (Table 1).

Based on large-scale exome sequencing data, the ExAC consortium has developed two scoring systems which are known as loss-of-function (LoF) intolerance score (pLI score) and Z-score for missense and synonymous variants¹⁵. The pLI score is the probability that a gene is intolerant to a LoF mutation. There are three major classes for LoF mutations: (i) null, where LoF variation is completely tolerated; (ii) recessive, where heterozygous LoFs are tolerated; and (iii) haploinsufficient, where heterozygous LoFs are not tolerated. The closer the pLI score is to 1, the less tolerant the gene is to LoF, with pLI >= 0.9 reflecting an extremely LoF-intolerant set of genes. Similarly, the Z-score was developed by the ExAC consortium for missense and synonymous variants. The Z-score is based on the deviation of the observed from the expected number (Table 1). Positive Z-scores indicate that the gene has less variants than expected and hence is intolerant to variation, while genes with more variants will have negative Z scores¹⁵.

Conservational screening of variants. Since high evolutionary conservation indicates functional importance of a position, it can be used to predict if a variant is deleterious or not. Based on this approach, evolutionary conservation-based parameters were developed as summarized in Table 1. Genomic Evolutionary Rate Profiling (GERP)⁶¹ and the PhastCons⁶² were utilized for the assessments of variant conservation with the GERP score of >2.0 and the PhastCons score of >0.3 indicating a high level of conservation of the variant position and were used as threshold in the screening of variants. PhyloP is a module of the PHAST package⁶³, which calculates p-values for conservation using a defined multiple alignment⁶³. PhyloP scores range from -14 to +6 where a higher score indicates a higher level of conservation. During the variant ranking process, a PhyloP score ≥3.0 was used as a criterion for a high level of conservation (Table 1).

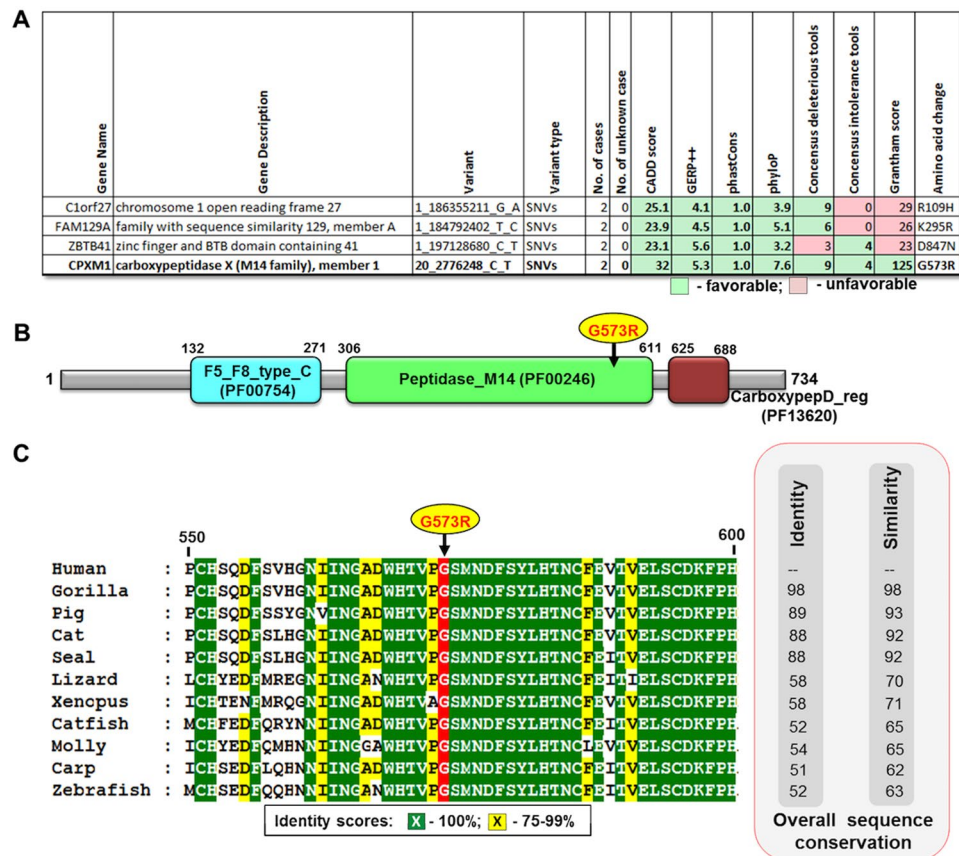


Figure 4. Overview of the top missense variants in the PTC family. (A) The 4 top ranked variants are shown with their favorable and unfavorable features. Grantham scores - 0–50 - conservative, 51–100 - moderately conservative, 101–150 - moderately radical and ≥ 151 - radical. (B) Location of the G573R variant in the peptidase M14 domain of CXPM1. (C) The G573R variant is localized in a highly conserved region. CXPM1 protein sequences were downloaded from GenBank as human (GenBank ID - NP_062555.1), gorilla (XP_004061758.1), cat (XP_003983774.1), pig (XP_003134381.1), seal (XP_021544821.1), lizard (XP_008120663.1), *Xenopus* (XP_002936314.1), catfish (XP_017320329.1), carp (XP_018934262.1), molly (XP_014844715.1) and zebrafish (XP_693256.4).

Screening missense variants using 12 deleterious ranking tools. All missense variants were assessed for deleteriousness using 12 tools as summarized in Table 2. These tools were developed using information based on (a) sequence conservation, (b) structure, (c) combination of sequence and structure information and (d) meta-prediction using already known tools. SIFT is based on sequence data⁶⁴, while PolyPhen uses both sequence and structure features⁶⁵, and MetaSVM and MetaLR are combining pre-existing tools and hence these are examples of meta prediction tools¹⁸. Scores from these tools were gathered using dbNSFP¹⁸. Variants predicted to be deleterious by at least 60% of these tools were analysed further.

Prediction of regulatory nature of the non-coding variants. Putative miRNA targets among the 3' UTR variants were detected using the miRanda suite; a mirSVR score lower than -0.1 is indicative of a “good” miRNA target⁷. Furthermore, we used entire dataset of the human miRNA target atlas from targetscan 7.0⁸ and scanned it with help of the intersect function of the bedtools^{66,67}. The 5' UTRs were scanned for transcription factor binding sites using SNPnexus⁹. For regulatory variants, we merged enhancer¹⁰ and promoter^{11,12} data from the FANTOM5 consortium using the intersect function of bedtools. We employed a similar strategy for variants potentially localized in the super enhancer regions using super-enhancer archive (SEA)²⁵ and dbSUPER²⁶. Furthermore, the regulatory nature and impact of non-coding variants were assessed using CADD v1.3²⁰, HaploReg V4²¹ and RegulomeDB²², which are based mainly on the ENCODE data⁶⁸. SNPnexus⁹ was used to evaluate changes in transcription factor binding sites. Additionally, epigenomic data and marks from 127 cell lines from the NIH Roadmap Epigenomics Mapping Consortium were accessed via CADD v1.3²⁰ for regulatory variants. We also tested if our variants were located within the ultra-conserved non-coding elements (UCNEs) and their clusters also known as ultra-conserved genomic regulatory blocks (UGRBs) with the help of UCNEbase²⁴ and also if these variants were located in ultra-sensitive regions (Ultrasen), defined by FunSeq2²³.

Visualization of the variants. Variants were visualized in the human genome (version hg19) using the Locuszoom²⁷, SNIpa²⁸, the UCSC²⁹ and ZENBU³⁰ genome browsers.

Ranked deleterious variants were examined for additional features. Potential variants were examined carefully for several additional features like if these variants were found in known list of CPGs¹. It was also examined whether clinical data and associated phenotypic data from ClinVar³³, Online Mendelian Inheritance in Man (OMIM, <https://omim.org/>) and other disease gene databases were available for the concerned variants. The status of RNA and protein expression for genes carrying potential deleterious variants was examined with the help of BioGPS³¹ and Human Protein Atlas³², respectively. We also checked if the concerned germline variant was already reported in known germline variant databases like CanVar Browser³⁴ or in the somatic mutation cancer databases like cBioPortal³⁵, COSMIC³⁶, ICGC³⁷ and IntOGen³⁸. The sequencing data for the concerned variants were rechecked manually using Integrative Genomics Viewer (IGV)³⁹ and validated using Sanger sequencing.

Other data analysis by FCVPPv2. We can assist with data analysis using FCVPPv2, please send us a personal communication to either AK (a.kumar@dkfz.de) or AF (a.foersti@dkfz.de).

References

- Rahman, N. Realizing the promise of cancer predisposition genes. *Nature* **505**, 302–308 (2014).
- Gilissen, C., Hoischen, A., Brunner, H. G. & Veltman, J. A. Disease gene identification strategies for exome sequencing. *European journal of human genetics: EJHG* **20**, 490–497, <https://doi.org/10.1038/ejhg.2011.258> (2012).
- Weren, R. D. *et al.* A germline homozygous mutation in the base-excision repair gene NTHL1 causes adenomatous polyposis and colorectal cancer. *Nat Genet* **47**, 668–671, <https://doi.org/10.1038/ng.3287> (2015).
- Cybulski, C. *et al.* Germline RECQL mutations are associated with breast cancer susceptibility. *Nat Genet* **47**, 643–646, <https://doi.org/10.1038/ng.3284> (2015).
- Forsti, A. *et al.* Pedigree based DNA sequencing pipeline for germline genomes of cancer families. *Hereditary cancer in clinical practice* **14**, 16, <https://doi.org/10.1186/s13053-016-0058-1> (2016).
- Chubb, D. *et al.* Rare disruptive mutations and their contribution to the heritable risk of colorectal cancer. *Nature communications* **7**, 11883, <https://doi.org/10.1038/ncomms11883> (2016).
- Betel, D., Koppal, A., Agius, P., Sander, C. & Leslie, C. Comprehensive modeling of microRNA targets predicts functional non-conserved and non-canonical sites. *Genome Biol* **11**, R90, <https://doi.org/10.1186/gb-2010-11-8-r90> (2010).
- Agarwal, V., Bell, G. W., Nam, J. W. & Bartel, D. P. Predicting effective microRNA target sites in mammalian mRNAs. *Elife* **4**, <https://doi.org/10.7554/eLife.05005> (2015).
- Dayem Ullah, A. Z., Lemoine, N. R. & Chelala, C. A practical guide for the functional annotation of genetic variations using SNPnexus. *Briefings in bioinformatics* **14**, 437–447, <https://doi.org/10.1093/bib/bbt004> (2013).
- Andersson, R. *et al.* An atlas of active enhancers across human cell types and tissues. *Nature* **507**, 455–461, <https://doi.org/10.1038/nature12787> (2014).
- Lizio, M. *et al.* Gateways to the FANTOM5 promoter level mammalian expression atlas. *Genome Biol* **16**, 22, <https://doi.org/10.1186/s13059-014-0560-6> (2015).
- Consortium, F. *et al.* A promoter-level mammalian expression atlas. *Nature* **507**, 462–470, <https://doi.org/10.1038/nature13182> (2014).
- Rimmer, A. *et al.* Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nature genetics* **46**, 912–918, <https://doi.org/10.1038/ng.3036> (2014).
- Wang, K., Li, M. & Hakonarson, H. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic acids research* **38**, e164, <https://doi.org/10.1093/nar/gkq603> (2010).
- Lek, M. *et al.* Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291, <https://doi.org/10.1038/nature19057> (2016).
- Fu, W. *et al.* Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. *Nature* **493**, 216–220, <https://doi.org/10.1038/nature11690> (2013).
- Genomes Project, C. *et al.* A global reference for human genetic variation. *Nature* **526**, 68–74, <https://doi.org/10.1038/nature15393> (2015).
- Liu, X., Wu, C., Li, C. & Boerwinkle, E. dbNSFPv3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs. *Human mutation* **37**, 235–241, <https://doi.org/10.1002/humu.22932> (2016).
- Smigielski, E. M., Sirotnik, K., Ward, M. & Sherry, S. T. dbSNP: a database of single nucleotide polymorphisms. *Nucleic acids research* **28**, 352–355 (2000).
- Kircher, M., Witten, D. M., Jain, P., O’Roak, B. J. & Cooper, G. M. A general framework for estimating the relative pathogenicity of human genetic variants. **46**, 310–315 <https://doi.org/10.1038/ng.2892> (2014).
- Ward, L. D. & Kellis, M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic acids research* **40**, D930–934, <https://doi.org/10.1093/nar/gkr917> (2012).
- Boyle, A. P. *et al.* Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res* **22**, 1790–1797, <https://doi.org/10.1101/gr.137323.112> (2012).
- Fu, Y. *et al.* FunSeq. 2: a framework for prioritizing noncoding regulatory variants in cancer. *Genome Biol* **15**, 480, <https://doi.org/10.1186/s13059-014-0480-5> (2014).
- Dimitrieva, S. & Bucher, P. UCNEbase—a database of ultraconserved non-coding elements and genomic regulatory blocks. *Nucleic acids research* **41**, D101–109, <https://doi.org/10.1093/nar/gks1092> (2013).
- Wei, Y. *et al.* SEA: a super-enhancer archive. *Nucleic acids research* **44**, D172–179, <https://doi.org/10.1093/nar/gkv1243> (2016).
- Khan, A. & Zhang, X. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic acids research* **44**, D164–171, <https://doi.org/10.1093/nar/gkv1002> (2016).
- Pruim, R. J. *et al.* LocusZoom: regional visualization of genome-wide association scan results. *Bioinformatics* **26**, 2336–2337, <https://doi.org/10.1093/bioinformatics/btq419> (2010).
- Arnold, M., Raffler, J., Pfeufer, A., Suhre, K. & Kastenmuller, G. SNIIPA: an interactive, genetic variant-centered annotation browser. *Bioinformatics* **31**, 1334–1336, <https://doi.org/10.1093/bioinformatics/btu779> (2015).
- Rosenbloom, K. R. *et al.* The UCSC Genome Browser database: 2015 update. *Nucleic acids research* **43**, D670–681, <https://doi.org/10.1093/nar/gku1177> (2015).
- Severin, J. *et al.* Interactive visualization and analysis of large-scale sequencing datasets using ZENBU. *Nature biotechnology* **32**, 217–219, <https://doi.org/10.1038/nbt.2840> (2014).
- Wu, C., Jin, X., Tsueng, G., Afrasiabi, C. & Su, A. I. BioGPS: building your own mash-up of gene annotations and expression profiles. *Nucleic acids research* **44**, D313–316, <https://doi.org/10.1093/nar/gkv1104> (2016).
- Uhlen, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419, <https://doi.org/10.1126/science.1260419> (2015).
- Landrum, M. J. *et al.* ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic acids research* **42**, D980–985, <https://doi.org/10.1093/nar/gkt1113> (2014).

34. Chubb, D., Broderick, P., Dobbins, S. E. & Houlston, R. S. CanVar: A resource for sharing germline variation in cancer patients. *F1000Research* **5**, 2813, <https://doi.org/10.12688/f1000research.10058.1> (2016).
35. Gao, J. *et al.* Integrative analysis of complex cancer genomics and clinical profiles using the cBioPortal. *Science signaling* **6**, p11, <https://doi.org/10.1126/scisignal.2004088> (2013).
36. Forbes, S. A. *et al.* The Catalogue of Somatic Mutations in Cancer (COSMIC). *Current protocols in human genetics* Chapter 10(Unit10), 11, <https://doi.org/10.1002/0471142905.hg1011s57> (2008).
37. Zhang, J. *et al.* International Cancer Genome Consortium Data Portal—a one-stop shop for cancer genomics data. *Database: the journal of biological databases and curation* **2011**, bar026, <https://doi.org/10.1093/database/bar026> (2011).
38. Gonzalez-Perez, A. *et al.* IntOGen-mutations identifies cancer drivers across tumor types. *Nature methods* **10**, 1081–1082, <https://doi.org/10.1038/nmeth.2642> (2013).
39. Thorvaldsdottir, H., Robinson, J. T. & Mesirov, J. P. Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Briefings in bioinformatics* **14**, 178–192, <https://doi.org/10.1093/bib/bbs017> (2013).
40. Kim, Y. H., O'Neill, H. M. & Whitehead, J. P. Carboxypeptidase X-1 (CPX-1) is a secreted collagen-binding glycoprotein. *Biochem Biophys Res Commun* **468**, 894–899, <https://doi.org/10.1016/j.bbrc.2015.11.053> (2015).
41. Kim, Y. H. *et al.* Identification of carboxypeptidase X (CPX)-1 as a positive regulator of adipogenesis. *FASEB journal: official publication of the Federation of American Societies for Experimental Biology* **30**, 2528–2540, <https://doi.org/10.1096/fj.201500107R> (2016).
42. Wood, W. M. *et al.* PPARgamma Promotes Growth and Invasion of ThyroidCancer Cells. *PPAR research* **2011**, 171765, <https://doi.org/10.1155/2011/171765> (2011).
43. Uehiro, N. *et al.* Circulating cell-free DNA-based epigenetic assay can detect early breast cancer. *Breast cancer research: BCR* **18**, 129, <https://doi.org/10.1186/s13058-016-0788-z> (2016).
44. Lindqvist, B. M., Wingren, S., Motlagh, P. B. & Nilsson, T. K. Whole genome DNA methylation signature of HER2-positive breast cancer. *Epigenetics* **9**, 1149–1162, <https://doi.org/10.4161/epi.29632> (2014).
45. Goedert, L. *et al.* Identification of Long Noncoding RNAs Deregulated in Papillary Thyroid Cancer and Correlated with BRAFV600E Mutation by Bioinformatics Integrative Analysis. *Sci Rep* **7**, 1662, <https://doi.org/10.1038/s41598-017-01957-0> (2017).
46. Weitzel, J. N., Blazer, K. R., MacDonald, D. J., Culver, J. O. & Offit, K. Genetics, genomics, and cancer risk assessment: State of the Art and Future Directions in the Era of Personalized Medicine. *CA: a cancer journal for clinicians* **61**, 327–359, <https://doi.org/10.3322/caac.20128> (2011).
47. Horn, S. *et al.* TERT promoter mutations in familial and sporadic melanoma. *Science* **339**, 959–961, <https://doi.org/10.1126/science.1230062> (2013).
48. Shi, J. *et al.* Rare missense variants in POT1 predispose to familial cutaneous malignant melanoma. *Nature genetics* **46**, 482–486, <https://doi.org/10.1038/ng.2941> (2014).
49. Palles, C. *et al.* Germline mutations affecting the proofreading domains of POLE and POLD1 predispose to colorectal adenomas and carcinomas. *Nature genetics* **45**, 136–144, <https://doi.org/10.1038/ng.2503> (2013).
50. Segui, N. *et al.* Germline Mutations in FAN1 Cause Hereditary Colorectal Cancer by Impairing DNA Repair. *Gastroenterology* **149**, 563–566, <https://doi.org/10.1053/j.gastro.2015.05.056> (2015).
51. Rotunno, M. *et al.* Whole exome sequencing in families at high risk for Hodgkin lymphoma: identification of a predisposing mutation in the KDR gene. *Haematologica* **101**, 853–860, <https://doi.org/10.3324/haematol.2015.135475> (2016).
52. Gara, S. K. *et al.* Germline HABP2 Mutation Causing Familial Nonmedullary Thyroid Cancer. *The New England journal of medicine* **373**, 448–455, <https://doi.org/10.1056/NEJMoa1502449> (2015).
53. White, A. E. & Harper, J. W. Cancer. Emerging anatomy of the BAP1 tumor suppressor system. *Science* **337**, 1463–1464, <https://doi.org/10.1126/science.1228463> (2012).
54. Jensen, M. R. & Helin, K. OVCA1: emerging as a bona fide tumor suppressor. *Genes & development* **18**, 245–248, <https://doi.org/10.1101/gad.1184004> (2004).
55. Reineke, E. L. & Kao, H. Y. PML: An emerging tumor suppressor and a target with therapeutic potential. *Cancer therapy* **7**, 219–226 (2009).
56. Vaz, C. V. *et al.* The Emerging Role of Regucalcin as a Tumor Suppressor: Facts and Views. *Current molecular medicine* **16**, 607–619 (2016).
57. Zhao, X., Li, X. & Zhang, X. HABP2 Mutation and Nonmedullary Thyroid Cancer. *The New England journal of medicine* **373**, 2084, <https://doi.org/10.1056/NEJMc1511631#SA1> (2015).
58. Offit, K. The future of clinical cancer genomics. *Seminars in oncology* **43**, 615–622, <https://doi.org/10.1053/j.seminoncol.2016.10.002> (2016).
59. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* **25**, 1754–1760, <https://doi.org/10.1093/bioinformatics/btp324> (2009).
60. Petrovski, S., Wang, Q., Heinzen, E. L., Allen, A. S. & Goldstein, D. B. Genic intolerance to functional variation and the interpretation of personal genomes. *PLoS Genet* **9**, e1003709, <https://doi.org/10.1371/journal.pgen.1003709> (2013).
61. Cooper, G. M. *et al.* Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res* **15**, 901–913, <https://doi.org/10.1101/gr.3577405> (2005).
62. Siepel, A. *et al.* Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res* **15**, 1034–1050, <https://doi.org/10.1101/gr.3715005> (2005).
63. Pollard, K. S., Hubisz, M. J., Rosenbloom, K. R. & Siepel, A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res* **20**, 110–121, <https://doi.org/10.1101/gr.097857.109> (2010).
64. Kumar, P., Henikoff, S. & Ng, P. C. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature protocols* **4**, 1073–1081, <https://doi.org/10.1038/nprot.2009.86> (2009).
65. Adzhubei, I., Jordan, D. M. & Sunyaev, S. R. Predicting functional effect of human missense mutations using PolyPhen-2. *Current protocols in human genetics* Chapter 7 (Unit7), 20, <https://doi.org/10.1002/0471142905.hg0720s76> (2013).
66. Quinlan, A. R. & Hall, I. M. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842, <https://doi.org/10.1093/bioinformatics/btq033> (2010).
67. Quinlan, A. R. BEDTools: The Swiss-Army Tool for Genome Feature Analysis. *Current protocols in bioinformatics/editorial board, Andreas D. Baxevanis... [et al.]* **47**, 11 12 11-34 <https://doi.org/10.1002/0471250953.bil112s47> (2014).
68. Consortium, E. P. *et al.* Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature* **447**, 799–816, <https://doi.org/10.1038/nature05874> (2007).
69. Chun, S. & Fay, J. C. Identification of deleterious mutations within three human genomes. *Genome Res* **19**, 1553–1561, <https://doi.org/10.1101/gr.092619.109> (2009).
70. Schwarz, J. M., Rodelsperger, C., Schuelke, M. & Seelow, D. MutationTaster evaluates disease-causing potential of sequence alterations. *Nature methods* **7**, 575–576, <https://doi.org/10.1038/nmeth0810-575> (2010).
71. Reva, B., Antipin, Y. & Sander, C. Predicting the functional impact of protein mutations: application to cancer genomics. *Nucleic acids research* **39**, e118, <https://doi.org/10.1093/nar/gkr407> (2011).
72. Shihab, H. A. *et al.* Predicting the functional, molecular, and phenotypic consequences of amino acid substitutions using hidden Markov models. *Human mutation* **34**, 57–65, <https://doi.org/10.1002/humu.22225> (2013).

73. Carter, H., Douville, C., Stenson, P. D., Cooper, D. N. & Karchin, R. Identifying Mendelian disease genes with the variant effect scoring tool. *BMC Genomics* **14**(Suppl 3), S3, <https://doi.org/10.1186/1471-2164-14-S3-S3> (2013).
74. Choi, Y., Sims, G. E., Murphy, S., Miller, J. R. & Chan, A. P. Predicting the functional effect of amino acid substitutions and indels. *PLoS One* **7**, e46688, <https://doi.org/10.1371/journal.pone.0046688> (2012).
75. Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P. L. & Casadio, R. Functional annotations improve the predictive score of human disease-related mutations in proteins. *Human mutation* **30**, 1237–1244, <https://doi.org/10.1002/humu.21047> (2009).

Acknowledgements

We thank the DKFZ Genomics and Proteomics Core Facility for Illumina Sequencing Services and Data Management Group (DMG) in the Division of Theoretical Bioinformatics, DKFZ for managing the NGS data. We thank Transcan ERA-NET funding from the German Federal Ministry of Education and Research (BMBF).

Author Contributions

Conceived and designed the experiments: A.K., O.R.B., K.H. and A.F. Performed the experiments: A.K. O.R.B., N.P., A.F. Analyzed the data: A.K. O.R.B., N.P. S.G., C.D., E.B., R.E., M.S., K.H. and A.F. Contributed to the writing of the manuscript: A.K., K.H. and A.F. All authors have approved final version of the manuscript.

Additional Information

Supplementary information accompanies this paper at <https://doi.org/10.1038/s41598-018-29952-z>.

Competing Interests: The authors declare no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit <http://creativecommons.org/licenses/by/4.0/>.

© The Author(s) 2018