Whole genome sequencing and inheritance-based variant filtering as a tool for unraveling missing heritability in pediatric cancer

Charlotte Derpoorter[a,b,c,*], Ruben Van Paemel[a,c], Katrien Vandemeulebroecke[a,b,c], Jolien Vanhooren[a,b,c], Bram De Wilde[a,b,c], Geneviève Laureys[a,b,c,†] and Tim Lammens[a,b,c,†]

[a]*Department of Pediatric Hematology-Oncology and Stem Cell Transplantation, Ghent University Hospital, Ghent, Belgium*

[b]*Department of Internal Medicine and Pediatrics, Ghent University, Ghent, Belgium*

[c]*Cancer Research Institute Ghent, Ghent, Belgium*

***Correspondence:** *Charlotte Derpoorter*

3K12D, C. Heymanslaan 10

9000 Ghent, Belgium

charlotte.derpoorter@ugent.be

Tel. +32-9-332-2456

[†]*These authors contributed equally as last authors*

Word count: 3383

Figures: 4

Tables: 1

Supplementary Tables: 5

Abstract

Survival rates for pediatric cancer have significantly increased the past decades, now exceeding 70-80% for most cancer types. The cause of cancer in children and adolescents remains largely unknown and a genetic susceptibility is considered in up to 10% of the cases, but most likely this is an underestimation. Families with multiple pediatric cancer patients are rare and strongly suggestive for an underlying predisposition to cancer. The absence of identifiable mutations in known cancer predisposing genes in such families could indicate undiscovered heritability. To discover candidate susceptibility variants, whole genome sequencing was performed on germline DNA of a family with two children affected by Burkitt lymphoma. Using an inheritance-based filtering approach, 18 correctly segregating coding variants were prioritized without a biased focus on specific genes or variants. Two variants in FAT4 and DCHS2 were highlighted, both involved in the Hippo signaling pathway, which controls tissue growth and stem cell activity. Similarly, a set of nine non-coding variants was prioritized, which might contribute, in differing degrees, to the increased cancer risk within this family. In conclusion, inheritance-based whole genome sequencing in selected families or cases is a valuable approach to prioritize variants and, thus, to further unravel genetic predisposition in childhood cancer.

Keywords

whole genome sequencing, germline predisposition, pediatric cancer, familial, missing heritability

Introduction

The causes of cancer in children and adolescents are largely unknown and an underlying genetic susceptibility is considered in up to 10% of the cases. This estimation, however, only includes germline mutations in known cancer predisposing genes.[1] Nevertheless, as lifestyle factors only play a minor part in childhood cancer development, a higher contribution of genetic variation is expected. The absence of identifiable germline mutations in patients with an obvious familial history and/or with clinical signs of predisposition syndromes is an indication for undiscovered genetic heritability.[2] Most cancer predisposition genes were identified through genome-wide linkage analysis or candidate gene approaches. Recent advances in next-generation sequencing (NGS) technology and associated bioinformatics offer the possibility to identify additional cancer predisposition genes.[3] However, analysis and interpretation of the enormous amount of sequencing data remains a key challenge. Several studies using whole exome/genome sequencing to identify genetic susceptibility to cancer included family cases. Only a minority uses unselected cases in case-control or population-based studies.[4] Efforts are undertaken to offer sequencing to all pediatric cancer patients, as predisposition is thought to play a major role in childhood carcinogenesis.[5,6] Family-based studies thus far often focus on variants shared between patients and perform segregation analysis only for the candidate variants. Other studies use unaffected relatives as controls to perform variant filtering.[7–9] Additionally, variants are frequently filtered based on a predefined gene list, affecting protein sequence (non-synonymous variants) and/or having high impact consequences, limiting further elucidation of cancer susceptibility.

Part of the missing heritability in childhood cancer might reside in other variant categories, such as non-coding regions or epigenetic modifications, or in other cancer associated genes yet to be identified. Importantly, most research has been focused on high-penetrant lesions with large effect sizes and low population frequencies. Some variants have smaller or intermediate

effect sizes, increasing the risk or susceptibility, without showing clear Mendelian segregation. Common variants within this category are typically covered by genome-wide association studies, but rare variants with intermediate or smaller effect sizes are not sufficiently captured. These variants could, individually or combined, have substantial effects and explain a large proportion of inherited cancer risk. Discovery of novel susceptibility genes provides improved diagnostic and prognostic information and enables early detection and modified treatment strategies. Screening and genetic counselling for patients and at-risk relatives becomes a possibility. Importantly, discovery of novel susceptibility genes might provide more insights into cancer biology and eventually lead to development of targeted therapies.[3,10]

Childhood malignancies are considered rare diseases and a combination of genetic and environmental factors, such as infection exposure, are thought to play a major role in the disease etiology. Nevertheless, as reported by Jongmans and colleagues, well-defined characteristics suggest inherited cancer susceptibility and warrant further investigation.[10] Families with multiple pediatric patients are rare and provide an unique opportunity for the investigation of undiscovered genetic variability to improve our understanding in cancer biology and contribute to the diagnosis, treatment and prevention of cancer.

This study describes an identification approach for candidate cancer predisposition genes or variants in a family with two Burkitt lymphoma (BL) patients using whole genome sequencing (WGS) and rare variant filtering based on pedigree information.

Materials and Methods

Sample Recruitment and Mononuclear Cell Isolation

To characterize an underlying genetic cancer susceptibility in this family, WGS was performed on germline DNA from patients (n = 2), parents (n = 4) and grandparents (n = 4). Constitutional

blood samples were collected after obtaining informed consent from the participants and/or their legal representatives. The study for identification of (candidate) susceptibility variants in this family was approved by the Ghent University Hospital Ethics Committee (2016/0385) and performed in accordance with the Declaration of Helsinki. Mononuclear cells (MNCs) were isolated using density gradient centrifugation in Leucosep tubes pre-filled with separation medium (Greiner Bio-One, Vilvoorde). In brief, blood was diluted in half with RPMI and added into the Leucosep tube following manufacturer's instructions. After centrifugation (12 min at 980 g, brake off), MNCs can be found as an interphase between the plasma and separation medium. This layer was collected, washed with RPMI and subsequently centrifuged at 650 g for 10 min. The washing step was repeated twice (5 min at 450 g) and MNCs were resuspended in RPMI with 10% FCS.

DNA Extraction

DNA of germline samples was extracted using the DNeasy Blood and Tissue Kit (Qiagen, Valencia, CA, USA) according to the manufacturer's instructions. For each sample, the concentration and DNA quality were measured using a NanoDrop ND-1000 spectrometer (NanoDrop Technologies, Wilmington, DE, USA).

Whole Genome Sequencing

To identify (candidate) predisposing variants, WGS was performed for both patients, their parents and grandparents. Library preparation and paired-end sequencing was done by BGI (Shenzhen, China) using the Illumina HiSeq X Ten. On average, per individual, the sequencing depth on the whole genome was a 34.51-fold coverage, with 98.63% bases having at least 10X coverage. Data was pre-processed by removing of adapter sequences, discarding low quality

reads (< 99.4% avg.) and mapping to the human reference genome (GRCh37/Hg19) using Burrows-Wheeler Aligner (BWA v0.7.12). Subsequent data-analysis, including variant discovery and call-set refinement, was undertaken according to the genome analysis toolkit (GATK v3.7.0) best practices, with duplicate reads removed by Picard tools (v2.8.1).[11] In brief, HaplotypeCaller was used to call SNPs per-sample to generate an intermediate genomic gVCF, which were merged into one combined VCF file using GenotypeGVCF. Subsequently, in the variant quality score recalibration (VQSR) step, the raw call-set was filtered based on machine learning to get high-confident variant calls using the GATK tools VariantRecalibrator and ApplyRecalibration. VQSR used high-quality known variant sets as training and truth resources (HapMap, Omni, 1000 Genomes, dbSNP) and built a predictive model to filter out false positive variants.[11] Finally, Ensembl Variant Effect Predictor was used to annotate variants.[12]

Candidate Variant Identification

Identification of candidate variants was performed using an inheritance-based filtering approach. This strategy filters variants according to genotype requirements that meet an autosomal dominant, recessive or compound heterozygous inheritance pattern in this family. More specifically, in an autosomal dominant model, patients and obligate carriers must be heterozygous and unrelated spouses must be homozygous reference. In a recessive model, both patients must be homozygous alternative and their parents should be heterozygous. Other family members cannot be homozygous alternative. In a compound heterozygous inheritance, the phenotype is caused by two heterozygous recessive alleles at different loci in a particular gene and variants were filtered accordingly. In addition, exonic and non-exonic variants were compared respectively to the databases ExAC (non-Finnish European subset, ExAC-NFE)[13]

and 1000 Genomes (European subset, 1000G-EUR)[14] for their absence or rarity in the general population (allele frequency (AF) ≤ 0.01).

Candidate Variant Prioritization

Variants were further analyzed depending on their genomic location, with coding variants ranked using the combine annotation dependent depletion (CADD) tool.[15] A scaled Phred CADD score > 10 represents the top 10% of probable functional variants and are considered deleterious, while the top 1% and 0.1% variants will have CADD scores of > 20 and > 30 respectively. Protein-coding variants were further prioritized based on evolutionary conservation using the Genomic Evolutionary Rate Profiling (GERP),[16] PhastCons[17] and PhyloP tools.[18] Methods to identify evolutionary conservation are used to identify elements that evolve faster or slower than expected under neutral drift and have emerged as powerful tools to discover potential functional elements in genomic sequences. A highly conserved sequence indicates functional importance and is used to predict variant deleteriousness. In addition, multiple functional prediction tools were used to assess for deleteriousness, including SIFT, PolyPhen, LRT, MutationTaster, FATHMM, PROVEAN, VEST4, FATHMM-MKL, BayesDel.[12,19] These different methods were chosen in such manner that they are not based on too similar principles, not have high missing rates and based on performance scores.[19] Similarly, for non-protein-coding variants, CADD, FATHMM, Eigen, ReMM, FunSeq, GWAVA and fitcons were used for prioritization.[20] Furthermore, the overlap of variants with regulatory elements (transcription factor binding sites, enhancers, promotors, insulators and silencers) in blood was evaluated using HaploReg,[21] Regulome DB[22] and The Ensembl Regulatory Build[23], encompassing data from the ENCyclopedia of DNA Elements (ENCODE) and Roadmap Epigenomics projects.

Sanger Sequencing

To confirm variants identified by WGS, DNA was amplified using primers designed by pxlence (http://www.pxlence.com)[24] or using the following primers:

rs115264616 F: 5'-TATAATCCACGAACCCTTTCC-3'

R: 5'-CCACGTTAAGACGACTGG-3'

rs536599551 F: 5'-CAGTTGGGAGATAGACCAGAG-3'

R: 5'-AACAGGATAATGAGGGAGTGG-3'

rs551826795 F: 5'-AAGCTCAAGATTGGAGACCAT-3'

R: 5'-AATGAATCACAAGCAGCTCTC-3'

rs192262761 F: 5'-CTGTTGTTTCTGTCGCCTC-3'

R: 5'-CTAACGGGTTTCGAAGGCAT-3'

rs1281405205 F: 5'-CAAGCCGACTCTTTGTCACC-3'

R: 5'-GGTAGGATGAGCAGCCAATG-3'

rs1034532353 F: 5'-AGAGTTATTACCTGCCAGCC-3'

R: 5'-CTCTATCACGCCTACCCATC-3'

rs181697534 F: 5'-GAGCAAGGGACTACGGATTT-3'

R: 5'-TTAGGCAGGCCAAGTGTTAT-3'

PCR was performed using the KAPA2G Robust HotStart ReadyMix (Sigma-Aldrich, Diegem) according to the manufactures' guidelines. Primer annealing occurred at 60 °C for 10 s and extension at 72 °C for 15 s. PCR amplification products were sequenced bidirectionally (Genewiz, Takeley, Essex, UK) and sequences were manually reviewed using MEGA6.[25]

Results

A workflow for familial whole genome sequencing and inheritance-based variant filtering was applied to a family with two pediatric BL patients for the identification and prioritization of candidate predisposing variants. In this family, a 14 year old male and his five year old third cousin were diagnosed at the department of Pediatric Hematology-Oncology and Stem Cell Transplantation, Ghent University Hospital (Belgium) (Figure 1). Both patients presented with an abdominal tumor localization and negative serology for EBV. Histological examination of the biopsy specimen of both cases revealed a 'starry sky' pattern and a MYC amplification with immunohistochemistry. FISH analysis on the specimen revealed a MYC rearrangement for both patients. After surgical biopsy and histological confirmation, they were treated according to the Inter-B-NHL 2010 protocols (European Inter-Group for Childhood Non-Hodgkin Lymphoma (EICNHL), ClinicalTrials.gov NCT01516580) for Murphy Stage III patients, since no bone or bone marrow invasion and central nervous system involvement were present. The 14 year old male was registered in the Inter-B-NHL ritux 2010 protocol, Group B because of high lactate dehydrogenase (LDH) level, and randomized without Rituximab. The five year old female was treated according to therapeutic recommendations for low/intermediate risk B-cell NHL because LDH was less than two times the upper limit of normal ("inter-B-NHL 2010 low/intermediate"). Both patients fare well six years after stop of treatment. An environmental factor or epidemiological association might play a role in the

tumorigenesis in this family. However, both children live 50 km apart and have never seen each other before diagnosis. Although the incidence of BL in the context of known cancer predisposition syndromes is rare, familial clustering has been described in case reports, suggesting a genetic predisposition as possible etiological factor for BL.[26–32]

To characterize a potential underlying genetic susceptibility in this family with two pediatric BL patients, WGS was performed for both patients, their parents and grandparents. No known pathogenic variants were detected in any of the previously recognized 565 cancer predisposing and other cancer related genes.[1] Inheritance-based variant filtering narrowed down the call-set to 1.771 and 341 rare single nucleotide variants (SNVs) and indels (AF $\leq$ 0.01) following an autosomal dominant and autosomal recessive model respectively. In addition, 15 variant combinations were identified following a compound heterozygous model with one variant segregating in all related family members and the other present in the unaffected parent in the same gene (Figure 2, Table S1).

The majority of the identified variants are non-coding and could impact on genomic elements that regulate gene expression. Further analysis was done using CADD, FATHMM, Eigen, ReMM, FunSeq, GWAVA and fitcons. In addition, non-coding variants with deleterious prediction scores were manually reviewed in HaploReg, RegulomeDB and The Ensembl Regulatory Build for their annotation in blood cells, as gene regulatory elements can be tissue specific (Table S2). These prioritization steps highlighted nine variants having three or more deleterious prediction scores in JADE1, ARAP2, ZNF827, MAML3, ANAPC4 and TMEM156. These might contribute, in differing degrees, to the increased cancer risk within this family (Table S2, Table S3).

Only 18 protein-coding variants remained after inheritance-based filtering and were further ranked using CADD (Table 1). Sanger sequencing was done for all autosomal dominant coding

variants and compound heterozygous variants (n = 26) for both patients, their parents and grandparents (n = 10). These results could validate 247 of 260 SNPs, leaving 21 correctly segregating candidate variants (Table S4). Non-coding variants (n = 7) having three or more deleterious prediction scores with a gene regulatory annotation in blood could all be validated (Table S4).

Non-synonymous variants were analyzed using functional prediction tools and evolutionary conservation (Table S5). Only three variants have multiple scores predicting variant deleteriousness, including a stop-gained variant with CADD > 30 (rs146298768, NM_001358235.2:c.7520T>G). This variant was considered lower priority based on the low conservation scores and location in the last protein-coding exon. In addition, two missense variants were prioritized, one in FAT4 (rs72914988, NM_001291303.3:c.7358G>T) and another in DCHS2 (rs79535970, NM_001358235.2:c.4816G>A), with a Phred CADD score of 24 and 18.53 respectively, which indicates a top 1% and 10% most deleterious variants. Both variants are located in a conserved region, as illustrated by high evolutionary conservation scores (Table S5). FAT4 and DCHS2 are large atypical cadherins and have both been implicated in planar cell polarity.[33] In D. melanogaster, the homologues Fat and Dachsous are well established members of Hippo signaling, which controls tissue growth and stem cell activity.[34] Recent evidence suggests that FAT4 can play a role in Wnt/b-catenin signaling and this pathway is essential for B-cell development.[35] As part of the Integrative OncoGenomics (IntOGen) framework, FAT4 has been described as a somatic mutational cancer driver in several cancer types, including diffuse large B-cell lymphoma, chronic lymphoblastic leukemia, lymphoma and acute lymphoblastic leukemia, with a predicted tumor suppressor role in cancer.[36,37] IntOGen is a systematic approach combining seven complementary methods for cancer driver identification from patient mutational data obtained from large sequencing efforts and smaller sequencing studies.[36]

The candidate variant in FAT4 (rs72914988, NM_001291303.3:c.7358G>T) is located in exon 9/18 (NM_001291303.3) and was confirmed by Sanger sequencing (Figure 3A). Both patients and their related family members (obligate carriers) are heterozygous for the variant, while it is absent in the unrelated parents and grandparents. The population allele frequency is very rare with 0.1170% in ExAC-NFE. Biallelic mutations in FAT4 have been associated with Hennekam lymphangiectasia-lymphedema syndrome 2 (HKLLS2, MIM 616006) and Van Maldergem syndrome 2 (VMLDS2, MIM 615546). The human FAT4 protocadherin is a large protein consisting of 4983 amino acids and composed of a 32 Cadherin repeat, two Calcium-binding EGF domains, a hEGF domain and two Laminin G domains. The candidate variant is located in a cadherin domain (Figure 3B) and was found highly conserved (Figure 3C). The variant introduces an Isoleucine, a bigger and more hydrophobic amino acid as compared to the wild-type Serine (p.Ser2453Ile), which can alter the cell-cell interaction function of the domain.

The candidate variant in DCHS2 (rs79535970, NM_001358235.2:c.4816G>A) is located in exon 9/20 (NM_001358235.2) and was also validated by Sanger sequencing (Figure 4A). The variant is compound heterozygous in the patients and segregates in their related family members. An alternate allele at a different loci within the same gene is inherited from the unrelated parent (rs4696593, NM_001358235.2:c.1686C>T). The population allele frequency is more common (5.8248% in ExAC-NFE), however in the CH model, both variants need to be present implying a rare combined allele frequency. The suggested role in Hippo signaling and planar cell polarity of both genes, however, is of high interest considering a polygenic risk for cancer development. No gene-phenotype relationships are catalogued for DCHS2, although homozygous mutations in DCHS1 have been associated with Van Maldergem syndrome 1 (VMLDS1, MIM 601390). DCHS2 is a protocadherin consisting of 3371 amino acids and composed of a 24 Cadherin repeat (Figure 4B). As for FAT4, the segregating variant is highly

conserved and introduces a bigger and more hydrophobic amino acid in a Cadherin domain (p.Ala1606Thr) (Figure 4C).

Discussion

Studies have reported that up to 30-35% of pediatric oncology patients fulfill the Jongmans and/or MIPOGG criteria, indicating a higher risk for an underlying genetic predisposition based on family history, type of malignancy, presence of multiple primary malignancies, specific clinical features and excessive toxicity.[5,38,39] However, in only 20-50% of those children at risk a pathogenic mutation in one of the known cancer predisposition is found.[5,6,39,40] This discrepancy may, at least in part, be caused by the lack in understanding non-coding mutations and difficulties in interpreting whole genome sequencing data. In addition, it recently became clear that synonymous mutations, that have often been discarded in analyses pipelines until now, can also affect gene expression levels and thus impact disease mechanisms.[41] Families with multiple pediatric cancer patients are strongly suggestive for an underlying genetic susceptibility and the absence of identifiable germline mutations could indicate undiscovered heritability.

This study aims to combine the advantages of NGS for the identification of rare variants with familial pedigree information in an inheritance-based filtering approach. This strategy reduces substantially the number of candidate variants, without predefined assumptions on gene function and/or mutation type. In a family with two pediatric BL patients, only 18 protein-coding variants remained after filtering, enabling further genetic and functional exploration for all these genes and variants. Additional refinement is possible using functional prediction software, but caution is needed as those tools are only able to model part of the true biological complexity. Therefore, multiple prediction scores were combined and variants were prioritized

accordingly, highlighting two missense variants in FAT4 and DCHS2 with a CADD score > 10, multiple deleterious prediction scores and located in a conserved region. Although knowledge about the exact biological function of FAT4 or DCHS2 remains limited, tissue specific roles in embryonic development and epithelial-to-mesenchymal transition have been described.[33,35,42–44] This is especially relevant as several lines of evidence indicate a developmental origin of childhood cancer.[45] For example, a fetal origin of tumors for hepatoblastoma and childhood lymphoblastic leukemia has been shown.[45] By contrast, most adult cancers arise in aging cells as a consequence of accumulated DNA damage. Interestingly, both genes are involved in planar cell polarity and Hippo signaling, which could indicate a combined signaling effect as possible mechanism of cancer predisposition. FAT4 has been shown to interact with DCHS1, a paralog of DCHS2.[43] Mice deficient for FAT4 or DCHS1 show developmental abnormalities and die shortly after birth.[43] DCHS2 null mice are viable and fertile, probably related to functional redundancy between DCHS2 and DCHS1.[46] Biallelic mutations in FAT4 have been identified in Hennekam syndrome, a disorder characterized by lymphedema, lymphangiectasia and cognitive impairment and uncovered an important role for FAT4 in the lymphatic vasculature.[47,48] FAT4 is recurrently mutated in several cancer types and gene suppression by mutagenesis or silencing induces tumorigenesis in breast and gastric cancers.[49,50] Interestingly, TEAD transcription factors act downstream of the Hippo signaling pathway and were shown to form a regulatory feedback mechanism with MYC to coordinate gene expression required for cell proliferation.[51] In addition, FAT4 downregulation has been associated with increased MYC expression in gastric cancer, ovarian cancer and oral squamous cells carcinoma.[35,52,53]

It is becoming more apparent that non-coding variants play a major role in cancer development through up- or downregulating expression, altering splice-sites and gene silencing. The deleterious impact of non-coding variants, however, remains difficult to predict, related to the

complex or unknown function of non-coding regions and the cellular specific context. Current functional prediction tools perform reasonably well on variants close to the protein-coding regions, but their variant prioritization in other regions is not consistent.

The majority of cancer predisposition syndromes are inherited autosomal dominant or autosomal recessive, however, many do not follow Mendelian genetics and show incomplete penetrance. This can explain the presence of unaffected family members, as noted for this family, but also why clinically healthy individuals can carry (potentially) pathogenic variants. Variants with smaller or intermediate effect sizes could individually or combined underlie a significant proportion of pediatric cancer patients, without showing clear Mendelian segregation. Polygenic inheritance is one factor that adds complexity to cancer development, although other factors, such as environmental exposures, can also play a major role in disease penetrance. Importantly, the putative role of variants identified within this family needs to be established and the cancer risk, clinical phenotype, genotype-phenotype associations and modifying factors need to be clarified. The variants should be confirmed and evaluated in other families, BL tumors and germline DNA of sporadic cases. In addition, the impact on RNA and protein expression should be analyzed and the functional impact on cell growth, apoptosis, response to genotoxic insult and differentiation should be evaluated. Evaluating this clinical impact, however, is exceedingly difficult due to the rarity of cancer predisposing mutations. Functional assays can provide evidence for pathogenicity and help unravel the molecular mechanism, nevertheless, clinical interpretation still requires robust genetic evidence. The latter would be facilitated by the implementation of NGS into routine clinical practice and improved availability of international networks and large-scale databases. Next to evaluating (likely) pathogenic variants in known cancer predisposing genes, novel and systematic prioritization strategies are needed to further unravel genetic predisposition to childhood cancer. Currently, multiple laboratories are implementing routine screening for identification

of variants in cancer predisposing genes using NGS approaches.[5,6] The strategy presented in this report will be applied for those patients in which routine screening strategies have not identified any known pathogenic predisposing variant.

In conclusion, this work highlights the value of using index families with multiple children affected by cancer as an unbiased approach to prioritize candidate cancer predisposition variants. To this end, variant identification should not solely focus on filtering one single top-hit mutation in known cancer genes and could further explore on rare, possibly polygenic, variants as part of missing heritability in pediatric oncology. This approach can be expanded to other pediatric cancer entities to improve understanding of the pathophysiology and molecular biology, offer adequate genetic counseling and most importantly, to contribute to novel diagnostic applications and therapeutic options.

Acknowledgments

This research was funded by the Belgian Foundation against Cancer, grant number 2016–126 (grant to GL), and vzw Kinderkankerfonds, a non-profit childhood cancer foundation under Belgian law (grant to TL).

Declaration of interest

The authors report there are no competing interests to declare.

Data availability statement

The data presented in this study are available in Supplementary materials. Raw WGS call-sets are not publicly available due to privacy concerns.

References

1. Zhang J, Walsh MF, Wu G, et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. *N Engl J Med*. 2015;373(24):2336-2346. doi:10.1056/NEJMoa1508054

2. Brodeur GM, Nichols KE, Plon SE, Schiffman JD, Malkin D. Pediatric Cancer Predisposition and Surveillance: An Overview, and a Tribute to Alfred G. Knudson Jr. *Clin Cancer Res*. 2017;23(11):e1-e5. doi:10.1158/1078-0432.CCR-17-0702

3. Rahman N. Realizing the promise of cancer predisposition genes. *Nature*. 2014;505(7483):302-308. doi:10.1038/nature12981

4. Rotunno M, Barajas R, Clyne M, et al. A Systematic Literature Review of Whole Exome and Genome Sequencing Population Studies of Genetic Susceptibility to Cancer. *Cancer Epidemiol Biomarkers Prev*. 2020;29(8):1519-1534. doi:10.1158/1055-9965.EPI-19-1551

5. Byrjalsen A, Hansen TVO, Stoltze UK, et al. Nationwide germline whole genome sequencing of 198 consecutive pediatric cancer patients reveals a high incidence of cancer prone syndromes. *PLoS Genet*. 2020;16(12):e1009231. doi:10.1371/journal.pgen.1009231

6. Wagener R, Taeubner J, Walter C, et al. Comprehensive germline-genomic and clinical profiling in 160 unselected children and adolescents with cancer. *Eur J Hum Genet*. 2021;29(8):1301-1311. doi:10.1038/s41431-021-00878-x

7. Zhang L, Jin Y, Zheng K, et al. Whole-Genome Sequencing Identifies a Novel Variation of WAS Gene Coordinating With Heterozygous Germline Mutation of APC to Enhance Hepatoblastoma Oncogenesis. *Front Genet*. 2018;9:668. doi:10.3389/fgene.2018.00668

8. Srivastava A, Giangiobbe S, Skopelitou D, et al. Whole Genome Sequencing Prioritizes CHEK2, EWSR1, and TIAM1 as Possible Predisposition Genes for Familial Non-Medullary Thyroid Cancer. *Front Endocrinol (Lausanne)*. 2021;12:600682. doi:10.3389/fendo.2021.600682

9. Fewings E, Larionov A, Redman J, et al. Germline pathogenic variants in PALB2 and other cancer-predisposing genes in families with hereditary diffuse gastric cancer without CDH1 mutation: a whole-exome sequencing study. *Lancet Gastroenterol Hepatol*. 2018;3(7):489-498. doi:10.1016/S2468-1253(18)30079-7

10. Jongmans MCJ, Loeffen JLCM, Waanders E, et al. Recognition of genetic predisposition in pediatric cancer patients: An easy-to-use selection tool. *Eur J Med Genet*. 2016;59(3):116-125. doi:10.1016/j.ejmg.2016.01.008

11. Van der Auwera GA, Carneiro MO, Hartl C, et al. From FastQ data to high confidence variant calls: the Genome Analysis Toolkit best practices pipeline. *Curr Protoc Bioinformatics*. 2013;43:11.10.1-11.10.33. doi:10.1002/0471250953.bi1110s43

12. McLaren W, Gil L, Hunt SE, et al. The Ensembl Variant Effect Predictor. *Genome Biol*. 2016;17(1):122. doi:10.1186/s13059-016-0974-4

13. Karczewski KJ, Weisburd B, Thomas B, et al. The ExAC browser: displaying reference data information from over 60 000 exomes. *Nucleic Acids Res*. 2017;45(D1):D840-D845. doi:10.1093/nar/gkw971

14. 1000 Genomes Project Consortium, Auton A, Brooks LD, et al. A global reference for human genetic variation. *Nature*. 2015;526(7571):68-74. doi:10.1038/nature15393

15. Kircher M, Witten DM, Jain P, O'Roak BJ, Cooper GM, Shendure J. A general framework for estimating the relative pathogenicity of human genetic variants. *Nat Genet*. 2014;46(3):310-315. doi:10.1038/ng.2892

16. Cooper GM, Stone EA, Asimenos G, et al. Distribution and intensity of constraint in mammalian genomic sequence. *Genome Res*. 2005;15(7):901-913. doi:10.1101/gr.3577405

17. Siepel A, Bejerano G, Pedersen JS, et al. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res*. 2005;15(8):1034-1050. doi:10.1101/gr.3715005

18. Pollard KS, Hubisz MJ, Rosenbloom KR, Siepel A. Detection of nonneutral substitution rates on mammalian phylogenies. *Genome Res*. 2010;20(1):110-121. doi:10.1101/gr.097857.109

19. Liu X, Li C, Mou C, Dong Y, Tu Y. dbNSFP v4: a comprehensive database of transcript-specific functional predictions and annotations for human nonsynonymous and splice-site SNVs. *Genome Med*. 2020;12(1):103. doi:10.1186/s13073-020-00803-9

20. Oscanoa J, Sivapalan L, Gadaleta E, Dayem Ullah AZ, Lemoine NR, Chelala C. SNPnexus: a web server for functional annotation of human genome sequence variation (2020 update). *Nucleic Acids Res*. 2020;48(W1):W185-W192. doi:10.1093/nar/gkaa420

21. Ward LD, Kellis M. HaploReg: a resource for exploring chromatin states, conservation, and regulatory motif alterations within sets of genetically linked variants. *Nucleic Acids Res*. 2012;40(Database issue):D930-934. doi:10.1093/nar/gkr917

22. Boyle AP, Hong EL, Hariharan M, et al. Annotation of functional variation in personal genomes using RegulomeDB. *Genome Res*. 2012;22(9):1790-1797. doi:10.1101/gr.137323.112

23. Zerbino DR, Wilder SP, Johnson N, Juettemann T, Flicek PR. The ensembl regulatory build. *Genome Biol*. 2015;16:56. doi:10.1186/s13059-015-0621-5

24. Coppieters F, Verniers K, De Leeneer K, Vandesompele J, Lefever S. Targeted resequencing and variant validation using pxlence PCR assays. *Biomol Detect Quantif*. 2016;6:22-26. doi:10.1016/j.bdq.2015.09.001

25. Tamura K, Stecher G, Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Mol Biol Evol*. 2013;30(12):2725-2729. doi:10.1093/molbev/mst197

26. Anderson KC, Jamison DS, Peters WP, Li FP. Familial Burkitt's lymphoma. Association with altered lymphocyte subsets in family members. *Am J Med*. 1986;81(1):158-162. doi:10.1016/0002-9343(86)90202-0

27. Joncas JH, Rioux E, Wastiaux JP, Leyritz M, Robillard L, Menezes J. Nasopharyngeal carcinoma and Burkitt's lymphoma in a Canadian family. I. HLA typing, EBV antibodies and serum immunoglobulins. *Can Med Assoc J*. 1976;115(9):858-860.

28. Brown TM, Heath CW. Time-space clustering among cases of Burkitt's tumor. *Cancer Res*. 1974;34(5):1216-1218.

29. Salawu L, Fatusi OA, Kemi-Rotimi F, Adeodu OO, Durosinmi MA. Familial Burkitt's lymphoma in Nigerians. *Ann Trop Paediatr*. 1997;17(4):375-379. doi:10.1080/02724936.1997.11747913

30. Hopman S, Merks J, Eussen H, et al. Structural genome variations in individuals with childhood cancer and tumour predisposition syndromes. *Eur J Cancer*. 2013;49(9):2170-2178. doi:10.1016/j.ejca.2013.02.002

31. Okabe M, Morishita T, Yasuda T, et al. Targeted deep next generation sequencing identifies potential somatic and germline variants for predisposition to familial Burkitt lymphoma. *Eur J Haematol*. 2021;107(1):166-169. doi:10.1111/ejh.13629

32. Winnett A, Thomas SJ, Brabin BJ, Bain C, Alpers MA, Moss DJ. Familial Burkitt's lymphoma in Papua New Guinea. *Br J Cancer*. 1997;75(5):757-761. doi:10.1038/bjc.1997.134

33. Saburi S, Hester I, Fischer E, et al. Loss of Fat4 disrupts PCP signaling and oriented cell division and leads to cystic kidney disease. *Nat Genet*. 2008;40(8):1010-1015. doi:10.1038/ng.179

34. Blair S, McNeill H. Big roles for Fat cadherins. *Curr Opin Cell Biol*. 2018;51:73-80. doi:10.1016/j.ceb.2017.11.006

35. Cai J, Feng D, Hu L, et al. FAT4 functions as a tumour suppressor in gastric cancer by modulating Wnt/β-catenin signalling. *Br J Cancer*. 2015;113(12):1720-1729. doi:10.1038/bjc.2015.367

36. Martínez-Jiménez F, Muiños F, Sentís I, et al. A compendium of mutational cancer driver genes. *Nat Rev Cancer*. 2020;20(10):555-572. doi:10.1038/s41568-020-0290-x

37. Tate JG, Bamford S, Jubb HC, et al. COSMIC: the Catalogue Of Somatic Mutations In Cancer. *Nucleic Acids Res*. 2019;47(D1):D941-D947. doi:10.1093/nar/gky1015

38. Knapke S, Nagarajan R, Correll J, Kent D, Burns K. Hereditary cancer risk assessment in a pediatric oncology follow-up clinic. *Pediatr Blood Cancer*. 2012;58(1):85-89. doi:10.1002/pbc.23283

39. Goudie C, Witkowski L, Cullinan N, et al. Performance of the McGill Interactive Pediatric OncoGenetic Guidelines for Identifying Cancer Predisposition Syndromes. *JAMA Oncol*. 2021;7(12):1806-1814. doi:10.1001/jamaoncol.2021.4536

40. Diets IJ, Waanders E, Ligtenberg MJ, et al. High Yield of Pathogenic Germline Mutations Causative or Likely Causative of the Cancer Phenotype in Selected Children with Cancer. *Clin Cancer Res*. 2018;24(7):1594-1603. doi:10.1158/1078-0432.CCR-17-1725

41. Shen X, Song S, Li C, Zhang J. Synonymous mutations in representative yeast genes are mostly strongly non-neutral. *Nature*. Published online June 8, 2022. doi:10.1038/s41586-022-04823-w

42. Van Hateren NJ, Das RM, Hautbergue GM, Borycki AG, Placzek M, Wilson SA. FatJ acts via the Hippo mediator Yap1 to restrict the size of neural progenitor cell pools. *Development*. 2011;138(10):1893-1902. doi:10.1242/dev.064204

43. Mao Y, Mulvaney J, Zakaria S, et al. Characterization of a Dchs1 mutant mouse reveals requirements for Dchs1-Fat4 signaling during mammalian development. *Development*. 2011;138(5):947-957. doi:10.1242/dev.057166

44. Ishiuchi T, Misaki K, Yonemura S, Takeichi M, Tanoue T. Mammalian Fat and Dachsous cadherins regulate apical membrane organization in the embryonic cerebral cortex. *J Cell Biol*. 2009;185(6):959-967. doi:10.1083/jcb.200811030

45. Behjati S, Gilbertson RJ, Pfister SM. Maturation Block in Childhood Cancer. *Cancer Discov*. 2021;11(3):542-544. doi:10.1158/2159-8290.CD-20-0926

46. Bagherie-Lachidan M, Reginensi A, Pan Q, et al. Stromal Fat4 acts non-autonomously with Dchs1/2 to restrict the nephron progenitor pool. *Development*. 2015;142(15):2564-2573. doi:10.1242/dev.122648

47. Betterman KL, Sutton DL, Secker GA, et al. Atypical cadherin FAT4 orchestrates lymphatic endothelial cell polarity in response to flow. *J Clin Invest*. 2020;130(6):3315-3328. doi:10.1172/JCI99027

48. Pujol F, Hodgson T, Martinez-Corral I, et al. Dachsous1-Fat4 Signaling Controls Endothelial Cell Polarization During Lymphatic Valve Morphogenesis-Brief Report. *Arterioscler Thromb Vasc Biol*. 2017;37(9):1732-1735. doi:10.1161/ATVBAHA.117.309818

49. Qi C, Zhu YT, Hu L, Zhu YJ. Identification of Fat4 as a candidate tumor suppressor gene in breast cancers. *Int J Cancer*. 2009;124(4):793-798. doi:10.1002/ijc.23775

50. Zang ZJ, Cutcutache I, Poon SL, et al. Exome sequencing of gastric adenocarcinoma identifies recurrent somatic mutations in cell adhesion and chromatin remodeling genes. *Nat Genet*. 2012;44(5):570-574. doi:10.1038/ng.2246

51. Huh HD, Kim DH, Jeong HS, Park HW. Regulation of TEAD Transcription Factors in Cancer Biology. *Cells*. 2019;8(6):E600. doi:10.3390/cells8060600

52. Malgundkar SH, Burney I, Al Moundhri M, et al. FAT4 silencing promotes epithelial-to-mesenchymal transition and invasion via regulation of YAP and β-catenin activity in ovarian cancer. *BMC Cancer*. 2020;20(1):374. doi:10.1186/s12885-020-06900-7

53. Ma L, Cui J, Xi H, Bian S, Wei B, Chen L. Fat4 suppression induces Yap translocation accounting for the promoted proliferation and migration of gastric cancer cells. *Cancer Biol Ther*. 2016;17(1):36-47. doi:10.1080/15384047.2015.1108488

Supporting information

The following supporting information is provided: Table S1: Subset of original data including rare variants following an autosomal dominant, autosomal recessive and compound heterozygous inheritance model. Table S2: Overview of the nine top-ranked germline non-coding variants. Table S3: Gene description for non-protein-coding variants. Table S4: Validation of the identified candidate variants using Sanger sequencing. Table S5: Overview of non-synonymous variants with functional prediction scores.

Tables

Table 1. Overview of protein-coding germline variants after inheritance-based filtering.

Gene Name	Variant (HGVS c.)	Variant (HGVS p.)	Variant Consequence	Inheritance	CADD score
DCHS2	NM_001358235.2:c.7520T>G	NP_001345164.1:p.Leu2507Ter	Stop-gained	AD	34
FAT4	NM_001291303.3:c.7358G>T	NP_001278232.1:p.Ser2453Ile	Missense variant	AD	24
C4orf33	NM_001099783.2:c.311C>T	NP_001093253.1:p.Ser104Leu	Missense variant	CH	22.1
FAT4	NM_001291303.3:c.5987A>G	NP_001278232.1:p.Lys1996Arg	Missense variant	AD	20.3
DCHS2	NM_001358235.2:c.4816G>A	NP_001345164.1:p.Ala1606Thr	Missense variant	CH	18.53
FAT4	NM_001291303.3:c.12506C>T	NP_001278232.1:p.Thr4169Ile	Missense variant	AD	18.04
PLCE1	*NM_001165979.2:c.246G>A	*NP_001159451.1:p.Lys82=	Synonymous variant	AD	14.97
DCHS2	NM_001358235.2:c.5917A>C	NP_001345164.1:p.Thr1973Pro	Missense variant	CH	10.16
DCHS2	NM_001358235.2:c.1686C>T	NP_001345164.1:p.Ser562=	Synonymous variant	CH	7.97
SCLT1	NM_144643.4:c.1032C>T	NP_653244.2:p.Asn344=	Synonymous variant	AD	7.27
DCHS2	NM_001358235.2:c.5926G>T	NP_001345164.1:p.Ala1976Ser	Missense variant	AD	3.52
DCHS2	NM_001358235.2:c.6084C>G	NP_001345164.1:p.Val2028=	Synonymous variant	CH	2.589
SCLT1	*NM_001300898.2:c.375C>T	*NP_001287827.1:p.Leu125=	Synonymous variant	AD	2.452
FAT4	NM_001291303.3:c.7935C>T	NP_001278232.1:p.Asp2645=	Synonymous variant	AD	2.275
SETD7	†NR_171655.1:n.381G>A	-	Non-coding variant	AD	0.66
SCLT1	NM_144643.4:c.663C>T	NP_653244.2:p.Ile221=	Synonymous variant	CH	0.527
DCHS2	*NM_001142552.1:c.4038G>A	*NP_001136024.1:p.Lys1346=	Synonymous variant	CH	0.324
PCDH18	NM_019035.5:c.1047C>T	NP_061908.1:p.Asp349=	Synonymous variant	AD	-

*Only the most severe of all observed consequence types is shown. †Most severe consequence for EMBL-EBI transcripts is synonymous, but the consequence is non-coding for NCBI transcripts.

Figure captions

Figure 1. Pedigree of the family. Stars indicate sequenced individuals.

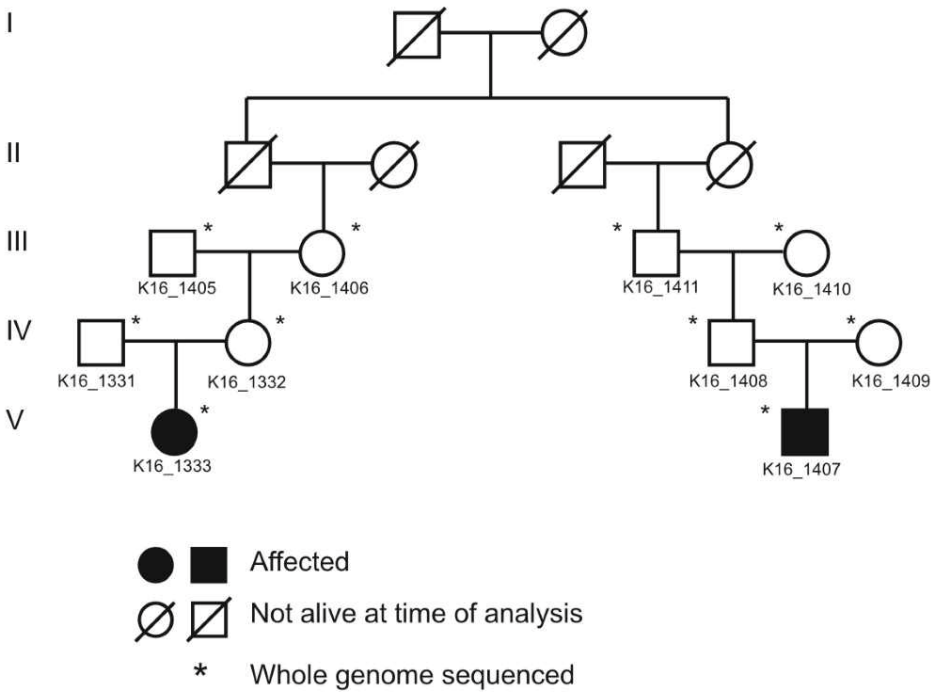
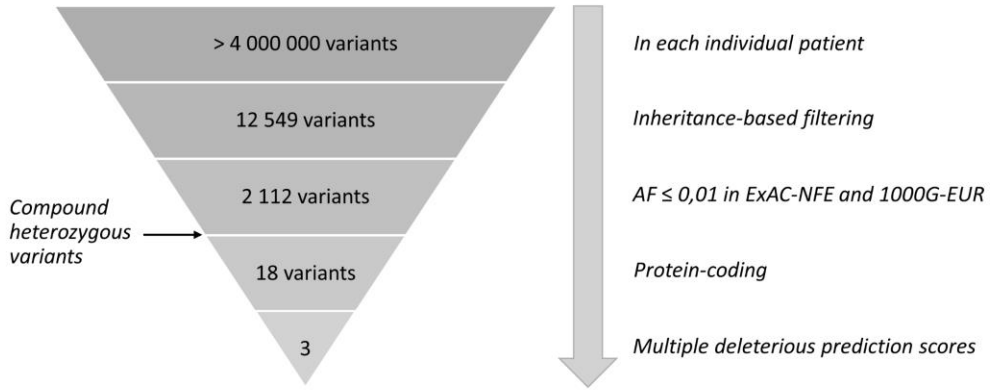


Figure 2. Inheritance-based filtering approach. (A) Germline variants were filtered according to different inheritance models and absence or rarity in the general population (AF ≤ 0.01). (B) Number of rare candidate variants obtained after inheritance-based filtering.

(**A**)



(**B**)

Filtering	**Protein-coding**		**Non-protein-coding**	
	SNVs	*Indels*	*SNVs*	*Indels*
Inheritance-based filtering approach				
▪ *Autosomal recessive*	0	0	22	319
▪ *Autosomal dominant*	11	0	1 069	691
▪ *Compound heterozygous*	15 variant combinations			

Figure 3. Germline FAT4 missense variant in the index family. (A) Sequencing chromatograms showing the variant sequence (c.7358G>T) in both patients. (B) Structure of the FAT4 protein. The p.S2453I variant is indicated. (C) Conservation of FAT4 at the S2453 residue across various species.

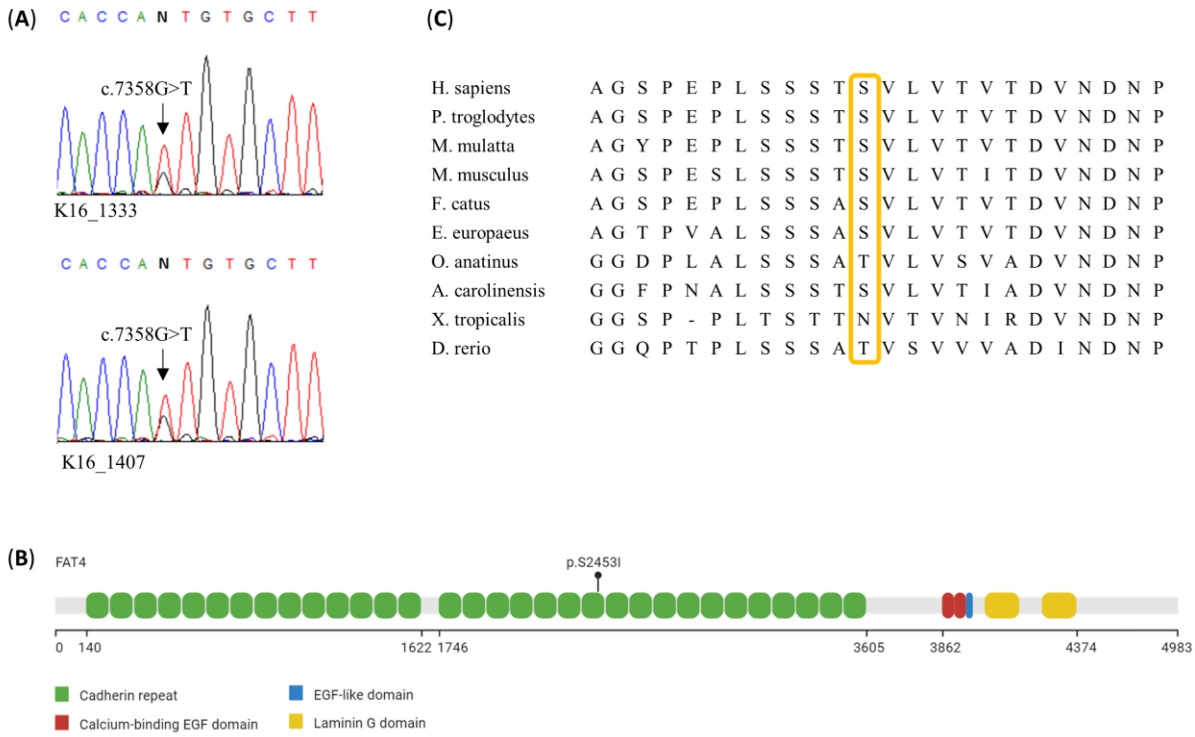


Figure 4. Germline DCHS2 missense variant in the index family. (A) Sequencing chromatograms showing the variant sequence (c.4816G>A) in both patients. (B) Structure of the DCHS2 protein. The p.A1606T variant is indicated. (C) Conservation of DCHS2 at the A1606 residue across various species.