



Precision Medicine Bioinformatics of Childhood Cancer in Denmark

Otamendi , Adrian

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

[Link back to DTU Orbit](#)

Citation (APA):
Otamendi , A. (2023). *Precision Medicine Bioinformatics of Childhood Cancer in Denmark*. DTU Health Technology.

General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.



DTU Health Tech
Department of Health Technology

Technical University of Denmark

PhD Thesis

Precision Medicine Bioinformatics of Childhood Cancer in Denmark

Author: Adrian Otamendi

Main supervisor: Elena Papaleo, Institute of Health Technology, Technical University of Denmark

Co-Supervisor: Cornelis Jan Hendrik Pronk, Department of Pediatrics, Clinical Sciences Lund, Lund University

Co-Supervisor: Kjeld Schmiegelow, Department of Pediatrics and Adolescent Medicine, Copenhagen University Hospital

June 30, 2023

Preface

This thesis represents the culmination of my PhD research conducted at DTU Health Tech department during June 2020-2023 under the guidance of Associate Professor Ramneek Gupta (DTU Health Tech-Novonordisk) for the first year and with Associate Professor Elena Papaleo (DTU Health-Tech and Danish Cancer Society) as main supervisor during the last two years. Clinical Professor. Kjeld Schmiegelow from Julian Marie Center at Rigshospitalet and Associate Professor Cornelius-Jan Hendrik Pronk from Lund University acted as co-supervisors.

This PhD thesis consists of an introduction followed by four manuscripts, and a final epilogue discussing some key aspects of the thesis as well as a summary of the contribution of each manuscript.

Abstract

This PhD started under the umbrella of iCOPE (Interregional Childhood Oncology Precision Medicine Exploration) project, which aims to improve life of children with cancer in Denmark implementing a personalized medicine approach. This thesis contributed by establishing a set of Next Generation Sequencing (NGS) workflows in a High Performance Computer (HPC) environment that allowed to detect, annotate and store efficiently germline and somatic short variants (SNV) as well as structural variants (SV) from Danish children with cancer. With the developed tools, four manuscripts have been produced and included in this thesis: **Manuscript 1** where germline short variants identified by our first WGS variant calling pipeline were used to investigate how frequently children with cancer were likely to have a cancer predisposition syndrome (CPS); similarly, in the **Manuscript 4** our latest whole genome sequencing (WGS) short variant detection pipeline was used to report the genetic variation and investigate the genetic predisposition in children with molecularly classified ependymoma in Denmark over the past 20 years; for **Manuscript 2**, a cohort of 566 WGS samples from Danish children with cancer was analyzed and Variants of Uncertain Significance (VUS) from interested genes were evaluated using *RosettaDDG* framework for structure-based calculations of the free energy changes upon amino acid substitution ($\Delta\Delta G$ s). Finally, in **Manuscript 3** we performed germline WGS in siblings and parents as well as RNAseq and WGS of DNA from the tumours to identify short nucleotide variants (SNVs) as well as structural variants (SVs) and described mechanisms possibly underlying a previously unseen co-occurrence of Philadelphia + Acute Lymphoblastic Leukaemia (Ph+ ALL) in siblings from a Danish family.

Resumé

Denne ph.d. startede under paraplyen af iCOPE-projektet (Interregional Childhood Oncology Precision Medicine Exploration), som har til formål at forbedre livet for børn med kræft i Danmark ved at implementere en personlig medicintilgang. Denne afhandling bidrog ved at etablere et sæt af Next Generation Sequencing (NGS) arbejdsgange i et High Performance Computer (HPC) miljø, der gjorde det muligt at detektere, kommentere og gemme effektivt germline og somatiske korte varianter (SNV) samt strukturelle varianter (SV) fra danske børn med kræft. Med de udviklede værktøjer er fire manuskripter blevet produceret og inkluderet i denne afhandling: **Manuskript 1**, hvor kimlinjekorte varianter identificeret af vores første WGS-variant kaldende pipeline blev brugt til at undersøge, hvor ofte børn med kræft sandsynligvis havde et cancer prædisposition syndrom (CPS) ; på samme måde, i **Manuskript 4** blev vores seneste hele genomsekventering (WGS) korte variantdetektionspipeline brugt til at rapportere den genetiske variation og undersøge den genetiske disposition hos børn med molekylært klassificeret ependymom i Danmark over de sidste 20 år; til **Manuskript 2** blev en kohorte på 566 WGS prøver fra danske børn med cancer analyseret, og Variants of Uncertain Significance (VUS) fra interesserede gener blev evalueret ved hjælp af RosettaDDG framework til strukturbaserede beregninger af de frie energiændringer ved aminosyresubstitution ($\Delta\Delta G_s$) . Endelig udførte vi i **Manuskript 3** germline WGS i søskende og forældre samt RNAseq og WGS af DNA fra tumorerne for at identificere korte nukleotidvarianter (SNV'er) såvel som strukturelle varianter (SV'er) og beskrev mekanismer, der muligvis ligger til grund for en tidligere uset co- forekomst af Philadelphia + Akut Lymfoblastisk Leukæmi (Ph+ ALL) hos søskende fra en dansk familie.

Acknowledgements

I would like to express my sincere gratitude to all those who have contributed to the completion of this PhD thesis.

First and foremost, I am deeply thankful to my supervisor, Elena Papaleo, for her invaluable guidance, support, and mentorship throughout this challenging journey. Her expertise, patience, and unwavering belief in my abilities have been crucial in shaping this research work.

I am grateful to the Technical University of Denmark (DTU) for providing me with the necessary resources, facilities, and funding to conduct my research. The research environment and infrastructure have been conducive to my academic and intellectual growth.

I would like to extend my appreciation to the censors for taking the time to evaluate my work.

While it is not possible to individually name everyone who has contributed to my growth as a researcher and individual, I am sincerely grateful to each and every person who has played a role, however small, in shaping my academic and personal development

Publications

Manuscripts included in the thesis:

- I. Byrjalsen A, Hansen TVO, Stoltze UK, Mehrjouy MM, Barnkob NM, Hjalgrim LL, Mathiasen R, Lautrup CK, Gregersen PA, Hasle H, Wehner PS, Tuckuviene R, Sackett PW, **Laspiur AO**, Rossing M, Marvig RL, Tommerup N, Olsen TE, Scheie D, Gupta R, Gerdes AM, Schmiegelow K, Wadt K. Nationwide germline whole genome sequencing of 198 consecutive pediatric cancer patients reveals a high incidence of cancer prone syndromes. PLoS Genet. 2020 Dec 17;16(12):e1009231. doi: 10.1371/journal.pgen.1009231. PMID: 33332384; PMCID: PMC7787686.
- II. Sora V, **Laspiur AO**, Degn K, Arnaudi M, Utichi M, Beltrame L, De Menezes D, Orlandi M, Stoltze UK, Rigina O, Sackett PW, Wadt K, Schmiegelow K, Tiberti M, Papaleo E. RosettaDDGPrediction for high-throughput mutational scans: From stability to binding. Protein Sci. 2023 Jan;32(1):e4527. doi: 10.1002/pro.4527. PMID: 36461907; PMCID: PMC9795540
- III. **Adrian Otamendi Laspiur**[†], Ulrik Kristoffer Stoltze[†], Frederik Steensgard Gade, Olga Rigina, Peter Wadt Sackett, Rikke Linnemann Nielsen, Ramneek Gupta, Nikola Tom, Elena Papaleo[†], Kjeld Schmiegelow, Karin Wadt[†] “Genomic analysis of the first ever reported Childhood Philadelphia positive Acute Lymphoblastic Leukaemia non-twin siblings”
Submitted to Nature Leukemia on June 2023
- IV. Foss-Skiftesvik J, Stoltze UK, van Overeem Hansen T, Ahlborn LB, Sørensen E, Ostrowski SR, Kullegaard SMA, **Laspiur AO**, Melchior LC, Scheie D, Kristensen BW, Skjøth-Rasmussen J, Schmiegelow K, Wadt K, Mathiasen R. Redefining germline predisposition in children with molecularly characterized ependymoma: a population-based 20-year cohort. Acta Neuropathol Commun. 2022 Aug 25;10(1):123. doi: 10.1186/s40478-022-01429-1. PMID: 36008825; PMCID: PMC9404601

Manuscripts not included in the thesis:

- V. Ulrik Kristoffer Stoltze, Jon Foss-Skiftesvik, Thomas van Overeem Hansen, Anna Byrjalsen, **Adrian Otamendi Laspiur**, Kasper Amund Henriksen, Anne-Marie Gerdes, Sisse Rye Ostrowski,

Erik Sørensen, Mads Bak, Birgitte Diness, Karen Grønskov, Zeynep Tümer, Rene Mathiasen, Jesper Brok, Lisa Hjalgrim, Astrid Sehested, Bram Gorissen, Simon Rasmussen, Konrad J. Karczewski, Karin A. W. Wadt, Kjeld Schmiegelow. “Causal Germline Variants in Childhood Cancer”

Manuscript in preparation

- VI. Marianne Helenius, **Adrian Otamendi Laspiur**, Line Egerod Lund, Olga Riga, Anna Schrøder Lassen, Freja Dahl Hede, Nanna Møller Barnkob, Nanna Møller Barnkob, Frederik Steensgaard Gade, Bernadette Christiansen, Tobias Kragholm, Elena Papaleo, Karin Wadt, Kjeld Schmiegelow, Rikke Linnemann Nielsen, Ramneek Gupta “Exploring the role of germline and somatic variation in childhood acute lymphoblastic leukemia: A germline-somatic continuum?”

Manuscript in preparation

Abbreviations

ALL	Acute Lymphoblastic Leukemia
BAM	Binary Alignment Map
BQSR	Base Quality Score Recalibration
BWA-MEM	Burrows-Wheeler Aligner - Maximal Exact Match
CNV	Copy Number Variation
DTU	Technical University of Denmark
FASTQ	Output file from the sequencing experiment containing the read out of the sample
GATK	Genome Analysis toolkit
GMM	Gaussian Mixture Model
HPC	High Performance Computer
iCOPE	Interregional Childhood Oncology Precision Medicine Exploration
Indel	Insertion and Deletion
MRD	Minimal Residual Disease
NGS	Next Generation Sequencing
Ph+ ALL	Philadelphia + Acute Lymphoblastic Leukemia
PeCAN	Pediatric Cancer Repository (Saint Judes Children Hospital)
PCR	Polymerase Chain Reaction
SNV	Single Nucleotide Variant
STAGING	Sequencing Tumor And Germline DNA—Implications for National Guidelines
STAR	Spliced Transcripts Aligned to a Reference
SV	Structural Variant
VCF	Variant Calling Format
VQSLOD	Variant Quality Score Log Odds

VQSR Variant Quality Score Recalibration

WGS Whole Genome Sequencing

WES Whole Exome Sequencing

Table of Contents

Preface	<i>i</i>
Abstract	<i>iii</i>
Resumé	<i>iv</i>
Acknowledgements	<i>v</i>
Publications	<i>vi</i>
Abbreviations	<i>viii</i>
Table of Contents	<i>x</i>
Introduction	<i>1</i>
Importance of precision medicine in childhood cancer	<i>2</i>
Role of NGS bioinformatics in advancing precision medicine for childhood cancer patients	<i>3</i>
Purpose and scope of the thesis (iCOPE)	<i>5</i>
NGS Bioinformatics Pipelines for Childhood Cancer Research in iCOPE	<i>6</i>
The raw data	<i>6</i>
The HPC system: Computerome	<i>6</i>
Structuring the raw data in HPC.....	<i>7</i>
Quality Control.....	<i>9</i>
NGS pipelines in Computerome HPC.....	<i>10</i>
Variant storage and annotation: MySQL and Curatio	<i>16</i>
Manuscript 1	<i>19</i>
Manuscript 2	<i>24</i>
Manuscript 3	<i>26</i>
Manuscript 4:	<i>29</i>
Epilogue	<i>31</i>
Using Sentieon in Computerome for childhood cancer research	<i>31</i>

NGS Data management challenges in large-scale studies	33
Conclusion	34
<i>Bibliography</i>	35

Introduction

Childhood cancer is a major public health issue in Europe, affecting around 35,000 children every year (1). The incidence of childhood cancer has been steadily increasing by 5-10% per decade in Europe, and it remains a leading cause of death among children over one year old in western countries, accounting for 20% of deaths due to disease (Figure 1). Based on statistics provided by the Association of Nordic Cancer Registries (nordcan.iarc.fr), the age-standardized incidence rate of cancer in children under the age of 19 in Denmark is 20.4 for males and 17.8 for females per 100,000 inhabitants (Figure 1). In Denmark, approximately 280 children and adolescents are diagnosed with cancer annually, with acute lymphoblastic leukemia (ALL) being the most common cancer type with an age-standardized incidence rate of 3.8 in males and 2.9 in females per 100,000. (2)(3)

With remarkable advances in diagnostics and treatment (4), the five-year survival after childhood cancer has improved from 30% in the 1960s to more than 80% nowadays in most of Europe (5–7). As a result of the increasing survival and lack of primary preventive measures, the number of childhood cancer survivors in society is growing steadily. This growing population is at risk of long-term health consequences such as late effects induced by the cancer or the intensive treatment at a young age (8–10). Although many survivors are well after therapy, a wide spectrum of long-term adverse health consequences in childhood cancer survivors has been described, indicating higher risks of a broad range of somatic and mental late effects, including second cancers (5,10,11), higher overall mortality rates (10,11) and severe chronic health conditions (5,10,12,13).

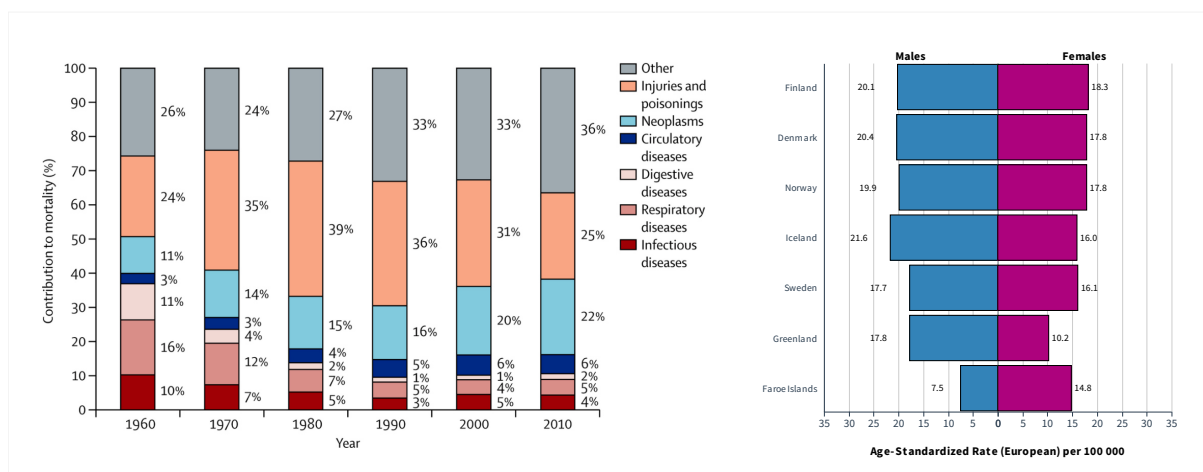


Figure 1. Cancer mortality in Europe from 1960-2010 (left, Wolfe et al. 2013). Age-standardized (0-19) incidence of cancer per 100 000 in Nordic countries from 2000-2020 (right).

Importance of precision medicine in childhood cancer

Precision medicine has emerged as an important approach for the treatment of cancer, and this is particularly true for childhood cancer patients (14). Childhood cancers are rare and have unique biological characteristics that differ from adult cancers (15). For instance, childhood cancers arise during critical periods of growth and development when cells undergo rapid proliferation and differentiation. The transformation of normal cells into cancerous cells during this dynamic developmental phase can result in unique tumor types and genetic alterations that differ from those seen in adult cancers. Children can inherit certain genetic mutations that increase their susceptibility to developing cancer. Germline mutations in specific genes, such as *TP53* (Li-Fraumeni syndrome) or *RB1* (retinoblastoma), contribute to the development of pediatric cancers (16,17). These inherited genetic predispositions play a more prominent role in childhood cancers compared to adult cancers. In that sense, childhood cancers often exhibit distinct biological characteristics and molecular profiles compared to adult cancers. Therefore, the treatment of childhood cancers requires a different approach, one that is tailored to the specific genetic and molecular characteristics of each patient's cancer (15). Precision medicine, which utilizes genetic and other types of information to tailor treatments to individual patients, has emerged as a promising approach for childhood cancer treatment (14).

One of the primary advantages of precision medicine for childhood cancer patients is the ability to minimize the side effects of treatment. Conventional chemotherapy and radiation therapy can have significant side effects, including damage to healthy tissues and organs (18). By tailoring treatment to the individual patient, precision medicine can reduce the risk of harmful side effects and improve the quality of life for childhood cancer patients. For example, some childhood cancers may respond better to chemotherapy than to radiation therapy, while others may require a combination of both. By using genetic information to identify the most effective treatment options for each patient, precision medicine can help to reduce the risk of side effects and improve treatment outcomes.

Another advantage of precision medicine for childhood cancer patients is the potential to identify new therapeutic targets (19). By using genomic and other molecular information to identify the specific drivers of childhood cancer, researchers and clinicians can develop new targeted therapies that specifically target these drivers. For example, researchers have used next generation sequencing (NGS) to identify specific mutations and gene fusions that drive the development of certain childhood cancers, such as pediatric ALL (14,20–23).

Precision medicine can also help to improve treatment outcomes for childhood cancer patients. By using genomic information to identify the most effective treatment options for each patient, precision medicine can improve the chances of success and reduce the risk of relapse (14,24). For example, researchers have used NGS to identify genetic markers that are associated with a poor prognosis in certain childhood cancers, such as high-risk neuroblastoma(25). By identifying these markers, clinicians can develop tailored treatment approaches that may improve outcomes for these patients.

Moreover, precision medicine has the potential to reduce healthcare costs associated with childhood cancer treatment. Conventional chemotherapy and radiation therapy can be expensive, and the costs associated with these treatments can be a significant burden for childhood cancer patients and their families or the administration (18). By tailoring treatment to the individual patient, precision medicine can reduce the need for unnecessary treatments and minimize the costs associated with side effects and complications.

In order to fully realize the potential of precision medicine for childhood cancer patients, it is important to continue to invest in research and development in this field. This includes the development of new genomic and molecular tools for the diagnosis and treatment of childhood cancer, as well as the creation of large-scale genomic databases for sharing information among researchers and clinicians.

[Role of NGS bioinformatics in advancing precision medicine for childhood cancer patients](#)

NGS is a powerful technology that enables the rapid and cost-effective sequencing of large amounts of DNA or RNA. It has revolutionized the field of genomics by providing a comprehensive view of the genetic landscape of diseases (14,26–29). NGS utilizes massively parallel sequencing, allowing for the simultaneous sequencing of millions of DNA fragments or RNA molecules. The process involves fragmenting the DNA or RNA, attaching adapters, amplifying the fragments, and sequencing them. Various NGS platforms, such as Illumina, Ion Torrent, and Pacific Biosciences, employ different sequencing chemistries and workflows, but the general principle remains the same. NGS enables the comprehensive profiling of the genomic alterations present in biological samples containing DNA or RNA. It can detect various types of genetic alterations, including single-nucleotide variants (SNVs), small insertions/deletions (indels), copy number variations (CNVs), and structural variations (SV, e.g., chromosomal rearrangements, gene fusions). By providing a comprehensive genomic picture, NGS allows for a more precise understanding of the genetic drivers and biology of childhood cancers. NGS

has transformed the landscape of precision medicine in childhood cancer. By identifying specific genetic alterations in tumors, such as driver mutations or gene fusions, NGS enables the selection of targeted therapies that specifically inhibit or modulate these alterations. This approach allows for personalized treatment strategies tailored to the unique genomic profile of each patient's cancer. Targeted therapies have shown promising results in certain childhood cancers such as Philadelphia + Acute Lymphoblastic Leukemia (Ph+ ALL), improving outcomes and minimizing treatment-related toxicities (30). NGS helps identify genetic alterations that serve as prognostic or predictive biomarkers in childhood cancers. These biomarkers can provide valuable information about a patient's prognosis, treatment response, and risk of disease recurrence. For example, specific genetic alterations detected by NGS may indicate a higher risk of relapse, allowing for more aggressive treatment or closer monitoring (14). Additionally, NGS can identify potential drug resistance mechanisms (31–33), enabling clinicians to make informed treatment decisions. Furthermore, NGS plays a crucial role in monitoring minimal residual disease (MRD), which refers to the small number of cancer cells that remain after treatment (34). MRD monitoring is particularly important in childhood cancers, as it helps assess treatment response, predict relapse, and guide therapeutic interventions. NGS-based approaches, such as targeted sequencing of specific cancer-associated mutations, can detect and quantify MRD with high sensitivity, facilitating timely interventions and personalized treatment adjustments.

NGS generates vast amounts of genomic data. By sharing and integrating these data through collaborative research efforts and data repositories, researchers and clinicians can gain deeper insights into any type of cancer. Large-scale initiatives, such as the Pediatric Cancer Genome Project from Saint Jude Children Hospital (PeCAN), or more region-oriented initiatives such as The Interregional Childhood Oncology Precision Medicine Exploration (iCOPE), leverage NGS data to accelerate discoveries, identify novel targets, and improve treatment strategies for pediatric cancers. This type of initiative is also being implemented in low and middle-income countries due to the reduction of the sequencing costs (35).

iCOPE project is a European initiative aimed at improving the diagnosis and treatment of childhood cancer with precision medicine. The project is funded by the European Regional Development Fund (ERDF) and involves partners from Denmark's Technical University (DTU) and Lund University in Sweden. The iCOPE project aims to improve diagnostics, treatment, cure rates, and the overall life situation of children with cancer by developing and implementing a precision medicine approach for the treatment of childhood cancer patients in the Øresund region (Denmark and Southern Sweden). This approach involves the use of genomic and other molecular and clinical information to tailor treatments to the specific genetic and molecular characteristics of each patient's cancer as well as

promote awareness of childhood cancer and the importance of precision medicine in improving treatment outcomes. The iCOPE project has several specific objectives, including the establishment of a regional network of healthcare providers and researchers who specialize in childhood cancer, the development of new diagnostic tools and therapies based on genomic and other molecular information, and the implementation of a platform for sharing data and information among partners:

- Build a database and biobank for childhood cancer research.
- Carry out extensive germline and tumor DNA and transcriptomics analyses to gain insights into cancer predisposition syndromes, somatic mutations, and tumor biology (NGS)
- Address ethical issues arising from genome sequencing.
- Investigate e-health as a tool to support healthcare at home.
- Develop telepresence robots to support children with staying connected to school during cancer treatment.

The iCOPE project is an important initiative that highlights the growing importance of precision medicine in the field of childhood cancer research and treatment. By bringing together researchers and healthcare providers from different regions and countries, the project aims to accelerate the development of new diagnostic and therapeutic approaches and improve outcomes for childhood cancer patients across Europe.

Purpose and scope of the thesis (iCOPE)

As part of the iCOPE project, this thesis will focus on establishing and applying a NGS workflow to detect germline and somatic short variants (SNV and indels) as well as SV from Danish children with cancer in a High-Performance Computer (HPC) setup. Furthermore, we worked on building a MySQL database to store and annotate the extracted variants that would serve as input data to other researchers in the consortium. From that perspective, this thesis can be categorized into two primary components. Firstly, the operational component involves setting up NGS pipelines to identify germline and somatic genetic variants in Computerome, and subsequently, store and annotate these variants in the MySQL database to facilitate research analysis (Chapter 3). Secondly, the research component involves analyzing the genetic variants identified by the NGS pipelines for specific research projects, which will be elaborated in Chapter 4.

NGS Bioinformatics Pipelines for Childhood Cancer Research in iCOPE

The raw data

Each patient in iCOPE underwent germline paired-end whole genome sequencing (WGS) analysis using DNA extracted from peripheral blood samples. In cases of hematologic malignancies, the blood samples were collected post-remission or skin biopsies in patients without remission. The genetic material was sent for sequencing to Beijing Genomic Institute (BGI, Hong Kong, China) using BGISEQ/HiSeqX (Illumina, San Diego, CA, USA) and resulting FASTQ files were subsequently sent back in hard drives to Copenhagen. This data was personally handed to Computerome-Support Team and transferred to our private folders in the HPC. Furthermore, we also had FASTQ files of Danish children with cancer available from the previous study Sequencing Tumor And Germline DNA—Implications for National Guidelines (STAGING) which offered WGS to all children and young people with cancer under the age of 18 in Denmark during 2016-2019. In this project, tumor RNA sequencing (RNAseq) was also performed on bone marrow or blood samples obtained before treatment in most patients. Details on the sequencing protocols are explained in Byrjalsen et al. 2020 (36). By the end of the project, we had 1023 FASTQ pairs: 730 germline WGS and 293 tumour RNAseq for 744 patients with multiple cancer types.

The HPC system: Computerome

Computerome is a high-performance computing (HPC) system and a Danish national supercomputer designed specifically for advanced scientific research and data-intensive computations. It is a state-of-the-art infrastructure located at DTU that provides researchers with the computational resources and storage capacity necessary for conducting large-scale and complex analyses. Computerome offers a wide range of computing resources, including 31,000 high-performance processor cores, 20PB of storage and a library of over 3000 different life science applications and reference datasets allowing researchers to perform computationally demanding tasks efficiently and effectively.

The system is specifically optimized for handling large-scale data and bioinformatics analyses and provides a secure and reliable environment for storing and processing sensitive genomic information, ensuring data integrity, privacy, and compliance with privacy regulations by implementing measures such as backups and regular validation procedures. Computerome is designed to meet high-security

standards by including measurements such as access controls, authentication mechanisms, data encryption and undergoes regular audits and assessments to maintain a secure computing environment for researchers. It offers advanced features such as parallel processing capabilities which enables researchers to handle the large volume of data generated by NGS technologies, accelerating the analysis pipelines, and facilitating the generation of results from large datafiles such as the alignment files generated in this project. One of the key advantages of Computerome is its scalability, allowing efficient handling of the substantial volume of data such as the data generated in this project as more children are included, thus accommodating the growing demands and ensuring efficient data management. In the same way, the software provided in Computerome is managed by modules that allow to update the version of the required software while keeping previous versions to maintain the reproducibility.

[Structuring the raw data in HPC](#)

Having a well-defined strategy for storing FASTQ files from a large consortium that has generated WGS and RNA-seq data for approximately 1000 samples of children with cancer is as important as the quality of the downstream analysis design and should not be underestimated. The strategy, which considers an effective storage strategy, facilitates data trackability and accessibility within the consortium and should also consider scalability to accommodate the growing volume of data and future expansion, ensuring that the consortium can handle incoming samples and evolving data analysis requirements. In STAGING-iCOPE we created a strategy that allowed us to keep track of every sequencing and analysis file of the project and facilitated the data management of such a large and constantly growing dataset by creating a naming system called the “iCOPE Naming System” (Figure 2). This system consists of encoding basic information of the sample such as patient identifier, sample type or type of analysis in the name of the files and folders with a specific hierarchy-based folder structure as described in Figure 2. With this system we were able to create a directory tree system where all the different folders/sub-folders contain information about the process of the sample sequencing.

NAMING CONVENTION:
 12345 A patient enters the system
 12345_123456.G A sample of tissue from that patient is collected
 12345_123456.G.T01.S1 The sample is sequenced on a platform
 12345_123456.G.T01.S1.PG001 The sequence is analysed using a bioinformatic pipeline

MMMMM_XXXXXX.B.TTT.SS.PPPPP
 Standard name structure (22 characters + 5 punctuations)

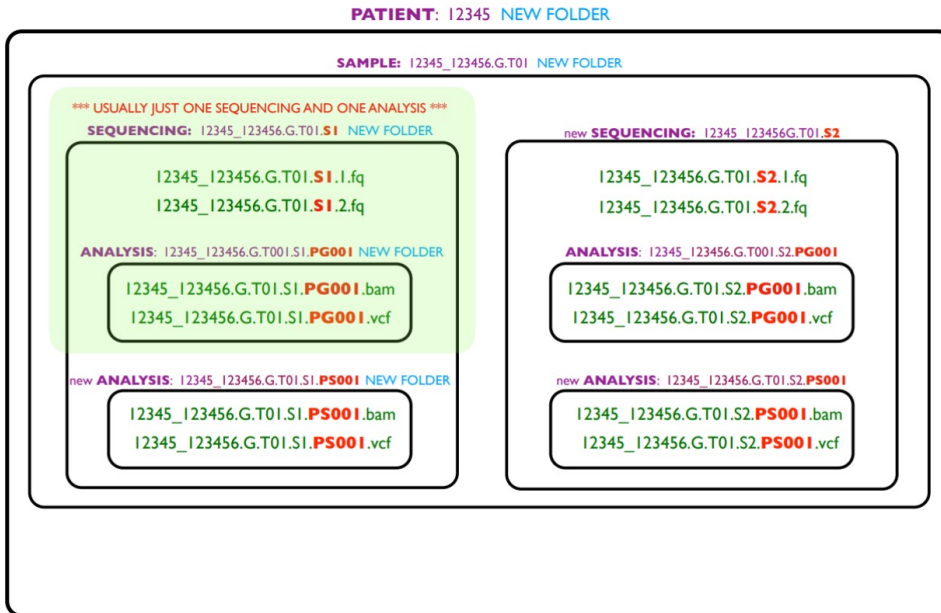


Figure 2. STAGING Dictionary, implemented also for iCOPE in Computerome.

Once we established a naming system to efficiently store the fastq files from the sequencing experiments of this project, we implemented a bash script (*create_tree*) that uses a list of FASTQ file names that follow the iCOPE naming system and creates a directory-tree based on the provided FASTQ file-name containing two symbolic links to the original FASTQ pair, which is stored in a secure non-writable folder. With this approach we prevent raw data corruption and ensure an independent area to analyze the raw FASTQ files that does not require any extra storage or data duplication.

Furthermore, we also created a bash script called "*sample_track*" which taking advantage of our naming system it reports a summary of the status of the folders and files for each of the patients (Figure 3). For instance, the script checks the presence and syntax of all the subfolders for each patient, so they follow the iCOPE naming system, as well as the presence of FASTQ symbolic links pointing to a healthy original FASTQ file. Furthermore, it also checks for the presence and integrity of the resulting alignment (BAM) and variant calling (VCF) files as well as the log file to report if an analysis was successful or not. After checking and displaying all the information about the patient folders

individually, it reports a summary of the general amount of FASTQ and analysis files such as BAM and VCF files we have in the cohort as well as information about the sequencing types or the biopsy types in the entire dataset.

MRD	SAMPLE	B	SEQTYPE	SEQ	FQ:R1/R2	ANALYSIS RESULTS
		G	T01	S3	FQ= ERROR (dangling)	PSG02:align= MISSING :vcf= OK
		G	T01	S3	FQ= ERROR (dangling)	PSG02:align= MISSING :vcf= OK
		G	T01	S3	FQ= ERROR (dangling)	PSG02:align= MISSING :vcf= OK

Figure 3. Example of the output produced by "sample_track" for three samples with different issues. Identifiers have been blurred for privacy. FASTQ symbolic links are broken (dangling) and BAM files cannot be detected for any of the samples.

Quality Control

Checking the integrity and quality of raw FASTQ files is of utmost importance in large-scale projects involving NGS data such as iCOPE or STAGING. These projects generate massive volumes of sequencing data from multiple samples, making it crucial to ensure the quality and integrity of the raw data before downstream analysis. Raw FASTQ files may contain errors introduced during sequencing, library preparation, or data transfer processes, which can significantly impact the reliability and accuracy of subsequent bioinformatics analyses.

In order to try to minimize these issues and due to the importance of the information contained in the FASTQ filenames in iCOPE, first we made sure the *sample_track* and *create_tree* scripts implemented checkpoints to ensure the FASTQ files were correctly named and placed in the correct folder in the directory tree with the correct filename. Furthermore, the *sample_track* script checks if the fastq files are stored in pairs or we are missing any of the files as well as the FASTQ pair file-size ratio, which is used to flag the potential corrupted FASTQ pairs.

Once we have the healthy FASTQ files placed in the correct folder we proceed to evaluate the quality of the samples with an in-house quality control script that combines software from *Picards*, *Samtools*, *Bedtools* and *GenomeCov* among others to evaluate potential errors such as errors introduced during the sequencing or library preparation. The output metrics of this script were combined and visualized using *MultiQC*. Germline FASTQ files passed all the filters in most of the cases and data-transfer errors were the cause of the files showing low quality metrics. Corrupted files were sent for transfer again and all germline FASTQ files were considered as good quality. However, the quality of the RNAseq samples was not as good as the germline FASTQ. In this case we observed multiple samples showing library preparation errors and the downstream analysis of these samples was performed very carefully.

Re-sequencing of some of the samples was proposed, however time and funding limitations did not allow to accomplish it.

[NGS pipelines in Computerome HPC](#)

In order to build the NGS pipeline for this project we used Sentieon software, a leading provider of bioinformatics software solutions for genomic data analysis built upon Genome Analysis Toolkit (GATK). The Sentieon software suite offers a comprehensive set of tools and pipelines that enable efficient and accurate analysis of NGS data. Sentieon offers dedicated pipelines for DNA analysis (*DNaseq*) and tumor-normal analysis (*TNseq*) for germline and somatic variant detection respectively. *DNaseq* is designed to process and analyze DNA sequencing data, encompassing best practices steps such as data preprocessing, read alignment, duplicate marking, base quality score recalibration, variant calling, variant filtering, annotation, and reporting. Furthermore, *DNAscope* from Sentieon allows to identify SV from germline BAM files. On the other hand, *TNseq* for somatic variant calling incorporates all the steps of the DNA analysis pipeline and includes additional features specifically tailored for paired tumor and normal samples. Additionally on the tumour-normal analysis, Sentieon offers the *TNscope* algorithm, a haplotype-based somatic variant caller with improved accuracy that also allows for somatic SV detection. Regarding RNAseq variant calling, Sentieon provides a pipeline based on *DNaseq* but specifically implemented for RNAseq data.

DNaseq:

As described in the Figure 4, the process begins with Sentieon-BWA (BWA-MEM, Burrows-Wheeler Aligner - Maximal Exact Match) which creates an index of the reference genome which builds an FM-index and auxiliary data structures for efficient sequence matching. When aligning paired-end reads, BWA-MEM treats each read of the pair independently. It starts by generating short substrings, known as seeds, from each read. These seeds act as anchor points and are used to initiate the alignment search in the reference genome. BWA-MEM employs an efficient algorithmic approach to extend these seeds in both directions, searching for potential alignments. During the extension process, it takes into account the presence of mismatches and gaps (indels) between the reads and the reference genome. This allows BWA-MEM to accurately align reads even in the presence of genetic variations or sequencing errors.

After aligning the paired-end reads independently, BWA-MEM analyzes the relative positions of their alignments to determine the proper orientation and distance between the read pairs. It considers both the mapping coordinates and orientation of the aligned reads to infer the actual insert size of the DNA fragments in the sample. By accurately determining the read pair orientations and distances, BWA-MEM enables the identification of genomic variations or structural rearrangements.

Once the alignment process with BWA-MEM for paired-end reads has ended, we sort the aligned read pairs based on their genomic coordinates. Sorting the read pairs ensures that both ends of the pairs are adjacent to each other in the alignment file, facilitating subsequent analyses that rely on proper read pairing information.

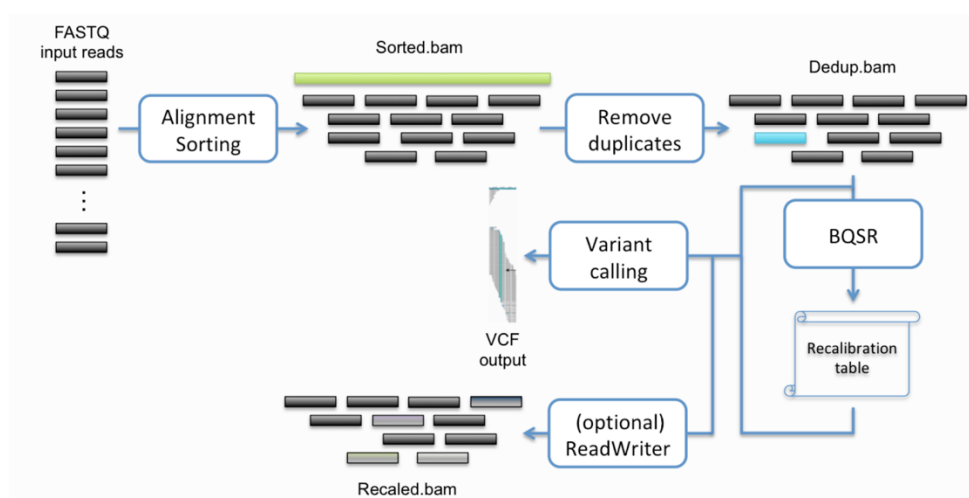


Figure 4. Sentieon DNaseq pipeline for variant calling (v202112.07)

Once the alignment file is ready, we evaluate the alignment looking into statistical summaries of the data quality and the pipeline data analysis results. For instance, the GC bias metrics includes a summary file providing an overview of biases in the proportion of guanine (G) and cytosine (C) bases in the sample DNA, as well as detailed metrics identifying specific regions with GC bias. Mean quality by cycle metrics calculates the average quality scores across different positions in the read, providing insights into sequencing quality variations throughout the read length. Quality distribution metrics analyzes the distribution of quality scores, helping identify problematic regions or artifacts. Insert size metrics assesses the distribution of library fragment sizes in paired-end sequencing data and alignment statistics metrics offers comprehensive information, including alignment rate, mapping quality, and duplicate rate, providing an overview of quality and potential artifacts in the read alignment. These metrics collectively provide insights into GC bias, quality scores, insert size distribution, and alignment quality, assisting in the evaluation and interpretation of the sequencing data.

Following sorting and alignment evaluation, the next step is to identify and remove PCR or optical duplicates in the paired-end data. Duplicates can arise during library preparation or sequencing and may lead to inflated variant calling statistics or biased downstream analyses. For this purpose, we use Sentieon *Dedup* and *LocusCollector* (GATK PicardMarkDuplicates) which consider both ends of the read pairs, considering their alignment positions and molecular indices. Duplicate read pairs are removed from the dataset, ensuring the removal of redundant or spurious data.

After removing duplicates, we perform an optional step, which is indel realignment by Sentieon *Realigner* algorithm (GATK Indel Realigner). During the alignment process, reads that span indel sites may not align perfectly due to the presence of insertions or deletions in the reference or sample genome. This can result in misalignments and mismatches in the aligned reads, which can be interpreted as mapping artifacts. To address this issue, the indel realignment step aims to improve the accuracy of the alignment by locally realigning the reads around the indel sites. The goal is to correctly align the reads to their most likely positions, considering the presence of indels. By performing this realignment, the artifacts caused by misalignments are minimized, and the accuracy of subsequent variant calling and downstream analyses is improved.

Once the reads are realigned, we apply Base Quality Score Recalibration (*BQSR*) which is a pivotal step in the analysis that by constructing a statistical model and recalibrating base quality scores to mitigate the impact of systematic errors, *BQSR* improves the accuracy and reliability of posterior variant calling. The algorithm operates through a multi-step process that involves constructing a statistical model and applying it to adjust base quality scores in aligned sequencing reads. The first step involves the construction of a statistical model using a training dataset that includes aligned reads and known variants obtained from high-quality reference databases. Various covariates are considered, such as sequence context, machine cycle, read group, sequencing platform, and other factors that can influence the accuracy of base calls.

Using Sentieon *QualCal* tool (GATK's BaseRecalibrator), the model analyzes the training dataset to estimate the systematic error rates associated with different combinations of covariate values. It identifies patterns and associations between covariates and the probability of base call errors. This statistical model provides insights into the relationship between specific covariates and the quality of base calls, allowing for the quantification of systematic errors.

Once the model is constructed, it is applied to adjust the base quality scores in the aligned reads using Sentieon *QualCal* (GATK's ApplyBQSR). This recalibration process involves calculating new base quality scores based on the estimated error rates obtained from the model. The recalibrated scores provide a more accurate representation of the true likelihoods of variant and non-variant base calls, effectively

correcting for systematic errors and improving the reliability of downstream variant calling and analysis.

For variant calling we used the *Haplotype* from Sentieon, built upon the principles of HaplotypeCaller from GATK. The *Haplotype* utilizes a sophisticated approach that considers the inherent complexity of the genome and incorporates local de novo assembly of haplotypes. The variant calling process begins by localizing regions of the genome with potential variants. The *Haplotype* segments the genome into smaller intervals and performs local de novo assembly for each interval, identifying possible haplotypes present in the data. By using a graph-based approach, the *Haplotype* constructs a haplotype assembly graph that represents the various potential haplotypes in the region. This approach is particularly useful for regions with complex variations, such as indels or multi-nucleotide polymorphisms.

Next, the *Haplotype* applies a likelihood-based model to assign probabilities to each possible haplotype and genotypes for each sample in the dataset. The model incorporates information from the sequencing data, including read alignments, base quality scores, and mapping qualities. It also takes into account potential sequencing errors, mapping biases, and various statistical factors to accurately determine the likelihoods of different genotypes. The *Haplotype* employs a local deconvolution algorithm to refine the likelihoods and genotypes, iteratively adjusting and recalibrating them based on the assembly graph and the read data. This iterative process improves the accuracy and robustness of the variant calling, particularly for regions with complex variations and low-frequency variants.

Finally, Sentieon *DNaseq* applies VQSR (Variant Quality Score Recalibration) using Sentieon *VarCal* and *ApplyVarCal* algorithms (GATK VariantRecalibrator and ApplyVQSR) which utilizes machine learning algorithms and statistical models to classify variants based on their likelihood of being true positives or false positives. VQSR begins with a training phase where a subset of high-quality variants, referred to as the truth set, is used to establish the relationship between various variant annotations and the truth status (known variants or novel mutations). These annotations may include attributes like base quality, mapping quality, strand bias, and more. During the training phase, a machine learning model, such as Gaussian mixture models (GMM), is trained using these variant annotations. The model learns the patterns and distributions of the variant annotations for true and false positive variants. It estimates the likelihood of a variant being true positive based on its annotations, allowing for classification of variants into different categories (e.g., "PASS" or "FAIL"). In the application phase of VQSR, the trained model is used to evaluate and recalibrate the variant quality scores of newly called variants. The model assigns a new quality score, known as the VQSLOD score (log-odds of being a true

variant), to each variant based on its annotations and the learned patterns from the training phase. Variants with higher VQSLOD scores are more likely to be true positive variants, while those with lower scores are more likely to be false positives. Based on the VQSLOD scores, a threshold is applied to filter out variants with low confidence, reducing the number of false positives and improving the accuracy of the final variant call set.

TNseq pipeline:

In this pipeline alignment, sorting, duplicate removal, indel realignment and BQSR are shared with DNaseq pipeline, however, variant calling is performed by *TNhaplotyper2* (Mutect2 from GATK4) which compares the genomic profiles of tumor and normal samples to identify somatic variants (Figure 5). The algorithm utilizes recalibrated BAM files for both samples obtained by independent analysis with DNaseq, along with optional parameters such as a panel of normals and a germline VCF file of known germline sites. *TNhaplotyper2* initiates the process by identifying "active regions" in the genome where variants are likely to be present by segmenting the genome into smaller intervals and focusing computational efforts on regions with potential variants. Within these active regions, *TNhaplotyper2* performs local assembly using a de Bruijn-like graph to reconstruct potential haplotypes and capture complex genomic variations, such as SNVs and small indels, even in regions with structural variations or copy number alterations. After local assembly, *TNhaplotyper2* applies a Bayesian classifier that estimates the probability of variants being present in the tumor sample but absent in the matched normal sample. By leveraging read evidence, quality scores, mapping qualities, and other features and by incorporating the contamination model and orientation bias, *TNhaplotyper2* improves the accuracy and reliability of somatic variant calling, ensuring robust analysis in the context of tumor-normal matched samples.

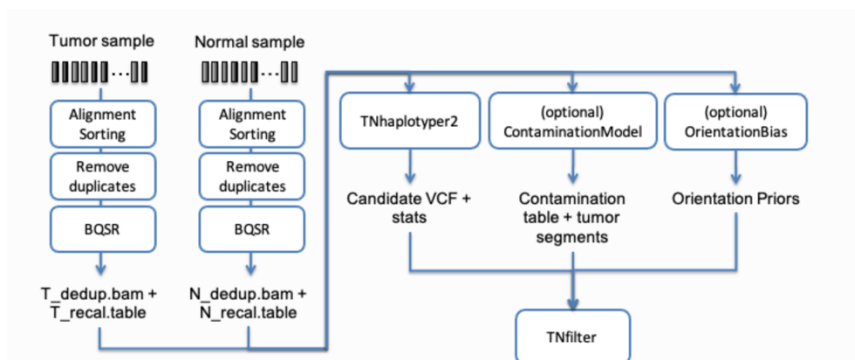


Figure 5. Recommended Tumour-Normal variant calling analysis (v202112.07)

Once the variant calling is completed, the resulting variant calls undergo a posterior filtering step using *TNfilter* from Sentieon (GATK's *FilterMutectCalls*). This step further refines the variant call set by applying additional filters based on quality metrics and specific thresholds in variables such as low tumor allele fraction, low quality scores, or high strand bias.

RNAseq pipeline:

While the overall pipeline for *RNAseq* is very similar to the *DNaseq*, we must use a different aligner (Spliced Transcripts Aligned to a Reference, *STAR*) to handle the specific requirements of the *RNAseq* experiment such as the presence of splice junctions, gene isoforms, and the possibility of transcript abundance estimation. *STAR* (Spliced Transcripts Alignment to a Reference) is designed to accurately align RNA-seq reads to a reference genome or transcriptome. It is known for its speed, sensitivity, and ability to handle spliced alignments, making it particularly suitable for detecting splice junctions and identifying alternative splicing events. *STAR* aligner utilizes a two-pass alignment strategy. In the first pass, the algorithm builds an index of the reference genome or transcriptome, identifying potential splice junctions based on annotated information. This indexing step enables efficient mapping of RNA-seq reads across exon-exon junctions. In the second pass, *STAR* aligns the reads to the indexed reference, considering the potential splice junctions detected in the first pass. This approach improves alignment accuracy and sensitivity, particularly for samples with complex splicing patterns or novel splice junctions not present in the reference annotation.

Similar to the *DNaseq* pipeline, once the reads are aligned and sorted, we calculate quality metrics to evaluate the alignment and we proceed to remove the duplicates.

During RNA-seq alignment, reads may span multiple exons due to alternative splicing or gene fusion events, so after alignment and sorting we proceed to split the reads at the splice junctions. With this step we ensure that each part of the read aligns to the appropriate genomic region, preserving the information about the splicing events. Splitting the reads at junctions allows for more accurate quantification of gene expression levels and identification of alternative splicing events. It enables the precise mapping of reads to specific exons, facilitating downstream analyses such as isoform quantification and detection of novel transcript variants.

After that, similar to the *DNaseq* pipeline, we apply *BQSR* and we call genetic variants by *Haplotype*. For this case, we need to use the option to trim the reads and apply a lower the Phred-scaled confidence threshold than *DNaseq* (emit and call confidence of 20) as suggested by Sentieon.

SV pipeline:

Sentieon provides two different options for SV calling: *DNAscope* (37) and *TNscope* (38). *DNAscope* includes preprocessing and assembly mathematics of the GATK's HaplotypeCaller with a machine-learned genotyping model and allows for efficient SNV, short variant and SV detection in germline WGS. *TNscope* is a haplotype-based variant caller developed for somatic SV from tumour and normal matching WGS. This haplotype-based variant caller incorporates the fundamental principles of mathematical models employed by GATK's MuTect2 by leveraging active region detection, de Bruijn-like graph-based haplotype assembly, pair-HMM for read-haplotype likelihood estimation, and subsequent genotype assignment. Notably, *TNscope* introduces a joint evaluation of haplotypes in tumor and normal samples, when available, which significantly enhances precision in somatic variant detection (Figure 6).

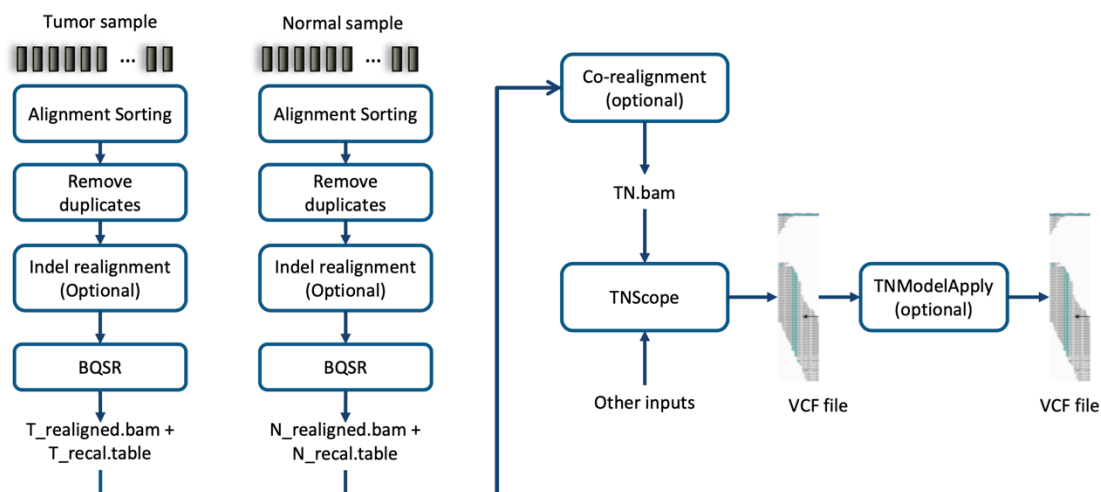


Figure 6. *TNscope* pipeline for Tumor-Normal somatic structural variant calling.

Variant storage and annotation: MySQL and Curatio

The variants identified by the different pipelines are stored in a private server called “Curatio” which is managed by the HPC specialist from iCOPE-DTU. This server has been designed to ensure data security and serves also as backup for the raw FASTQ files. In order to make the data available to other researchers in the consortium, this server hosts a MariaDB (MySQL) database that contains the variants identified by some of the pipelines such as the *DNaseq*, *RNAseq* or *TNseq* pipelines. Result VCF files from the mentioned pipelines were transferred from Computerome to Curatio and processed

by MySQL to parse the main fields of the VCF file such as genomic position, allele and information for the call set of each sample. This information by itself is almost meaningless, so we used different annotation sources such as Combined Annotation Dependent Depletion (CADD), REVEL or GnomAD to include relevant information about the genomic context and predicted pathogenetic impact of each variant. We used chromosomal position along with reference and alternate allele to combine all the annotations with the genetic variants from the samples as described in Figure 7.

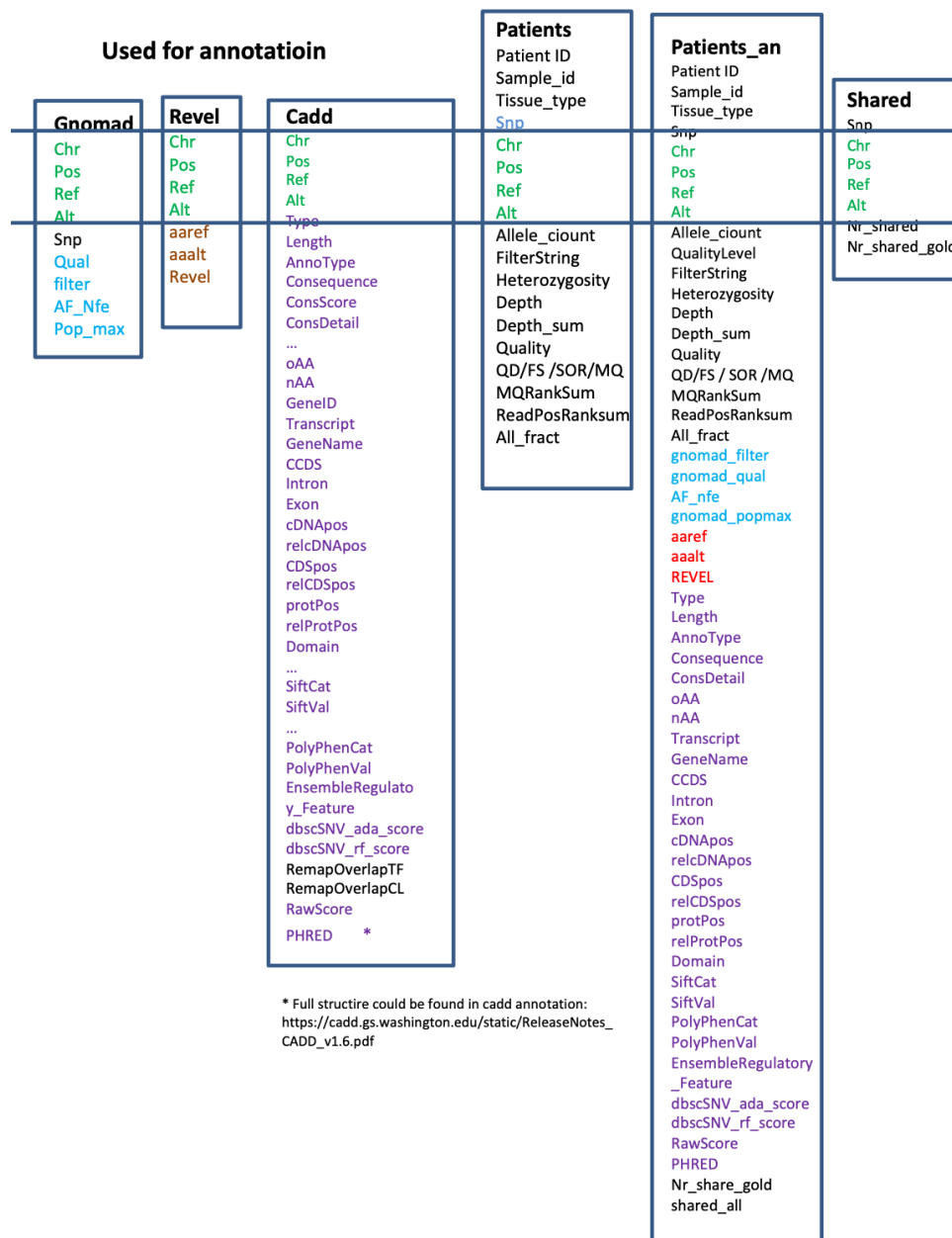


Figure 7. MySQL DB schema. Gnomad, Revel, CADD and Patients tables are combined into Patients_an for posterior querying.

The output table, *Patients_an* (Figure 7), contains records containing information about the specific call of a variant in a sample as well as the predicted genotype, genetic context, predicted impact scores

and population frequency information. We repeated this process for the pipelines *DNaseq* and *TNseq* with reference genome GCRh38; and for pipelines *DNaseq* and *RNAseq* with reference genome GCRh37 which were combined in one single table:

Table	Pipeline	Genome	Nº Variants	Size in disk (Gb)	Nº Samples
1	DNaseq and RNAseq	GCRh37	575659635	176,5	116
2	DNaseq	GCRh38	3596502564	959	566
3	TNseq	GCRh38	683103363	16,7	214

Figure 8. Summary of the main tables from the MySQL database.

Manuscript 1

Nationwide germline whole genome sequencing of 198 consecutive pediatric cancer patients reveals a high incidence of cancer prone syndromes

Byrjalsen A, Hansen TVO, Stoltze UK, Mehrjouy MM, Barnkob NM, Hjalgrim LL, Mathiasen R, Lautrup CK, Gregersen PA, Hasle H, Wehner PS, Tuckuviene R, Sackett PW, **Laspiur AO**, Rossing M, Marvig RL, Tommerup N, Olsen TE, Scheie D, Gupta R, Gerdes AM, Schmiegelow K, Wadt K.

PLoS Genetics December 2020

Contribution: In this manuscript I have contributed to the bioinformatics analysis of the NGS data. The first short variant detection pipeline I built for this project, based on Sentieon and human reference genome b37, was used to report process the FASTQ files and report genetic variants of the specific samples.

RESEARCH ARTICLE

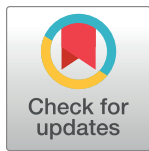
Nationwide germline whole genome sequencing of 198 consecutive pediatric cancer patients reveals a high incidence of cancer prone syndromes

Anna Byrjalsen^{1,2}, Thomas V. O. Hansen^{1,2}, Ulrik K. Stoltze¹, Mana M. Mehrjouy^{1,2}, Nanna Moeller Barnkob³, Lisa L. Hjalgrim², René Mathiasen², Charlotte K. Laurrup⁴, Pernille A. Gregersen⁵, Henrik Hasle⁶, Peder S. Wehner⁷, Ruta Tuckuviene⁸, Peter Wad Sackett³, Adrian O. Laspiur³, Maria Rossing⁹, Rasmus L. Marvig⁹, Niels Tommerup¹⁰, Tina Elisabeth Olsen¹¹, David Scheie¹¹, Ramneek Gupta³, Anne-Marie Gerdes^{1‡}, Kjeld Schmiegelow^{2,12‡}, Karin Wadt^{1‡*}

1 Department of Clinical Genetics, Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark, **2** Department of Paediatrics and Adolescent Medicine, Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark, **3** Department of Health Technology, Technical University of Denmark, Copenhagen, Denmark, **4** Department of Clinical Genetics, Aalborg University Hospital, Aalborg, Denmark, **5** Department of Clinical Genetics, Aarhus University Hospital, Aarhus, Denmark, **6** Department of Paediatrics and Adolescent Medicine, Aarhus University Hospital, Aarhus, Denmark, **7** Department of Paediatric Hematology and Oncology, H. C. Andersen Children's Hospital, Odense University Hospital, Odense, Denmark, **8** Department of Paediatrics and Adolescent Medicine, Aalborg University Hospital, Aalborg, Denmark, **9** Center for Genomic Medicine, Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark, **10** Department of Cellular and Molecular Medicine, University of Copenhagen, Copenhagen, Denmark, **11** Department of Pathology, Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark, **12** Institute of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark

‡ These authors are share senior authors on this work.

* karin.wadt@regionh.dk



OPEN ACCESS

Citation: Byrjalsen A, Hansen TVO, Stoltze UK, Mehrjouy MM, Barnkob NM, Hjalgrim LL, et al. (2020) Nationwide germline whole genome sequencing of 198 consecutive pediatric cancer patients reveals a high incidence of cancer prone syndromes. *PLoS Genet* 16(12): e1009231. <https://doi.org/10.1371/journal.pgen.1009231>

Editor: Charis Eng, Cleveland Clinic Genomic Medicine Institute, UNITED STATES

Received: August 24, 2020

Accepted: October 28, 2020

Published: December 17, 2020

Peer Review History: PLOS recognizes the benefits of transparency in the peer review process; therefore, we enable the publication of all of the content of peer review and author responses alongside final, published articles. The editorial history of this article is available here: <https://doi.org/10.1371/journal.pgen.1009231>

Copyright: © 2020 Byrjalsen et al. This is an open access article distributed under the terms of the [Creative Commons Attribution License](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Data Availability Statement: All relevant data are within the manuscript and its [Supporting Information](#) files.

Abstract

PURPOSE: Historically, cancer predisposition syndromes (CPSs) were rarely established for children with cancer. This nationwide, population-based study investigated how frequently children with cancer had or were likely to have a CPS. **METHODS:** Children (0–17 years) in Denmark with newly diagnosed cancer were invited to participate in whole-genome sequencing of germline DNA. Suspicion of CPS was assessed according to Jongmans'/McGill Interactive Pediatric OncoGenetic Guidelines (MIPOGG) criteria and familial cancer diagnoses were verified using population-based registries. **RESULTS:** 198 of 235 (84.3%) eligible patients participated, of whom 94/198 (47.5%) carried pathogenic variants (PVs) in a CPS gene or had clinical features indicating CPS. Twenty-nine of 198 (14.6%) patients harbored a CPS, of whom 21/198 (10.6%) harbored a childhood-onset and 9/198 (4.5%) an adult-onset CPS. In addition, 23/198 (11.6%) patients carried a PV associated with biallelic CPS. Seven of the 54 (12.9%) patients carried two or more variants in different CPS genes. Seventy of 198 (35.4%) patients fulfilled the Jongmans' and/or MIPOGG criteria indicating an underlying CPS, including two of the 9 (22.2%) patients with an adult-onset CPS versus 18 of the 21 (85.7%) patients with a childhood-onset CPS ($p = 0.0022$), eight of the

Funding: This study was financially supported by the Danish Childhood Cancer Foundation (KS, <https://boernecancerfonden.dk/>), Rigshospitalet (AB, KW, <https://www.rigshospitalet.dk/>), The Danish Cancer Society (KS), the Foundation for Health Research of the Capital Region of Denmark (AMG, <https://www.regionh.dk/til-fagfolk/Forskning-og-innovation/finansiering-og-fonde/s%C3%B8g-regionale-midler/Sider/Region-Hovedstadens-forskningsmidler.aspx>), The European Union's Interregional Öresund–Kattegat–Skagerrak grant (KS, <https://interreg-oks.eu/>), Aase and Ejnar Danielsen's Foundation (AMG, <https://danielsensfond.dk/>), and Engineer Otto Christensen's Foundation (AB, no URL), the Danish National Research Foundation (RLM, <https://dg.dk/>, grant number 126). No funding sources played a role in study design, data collection, analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

additional 23 (34.8%) patients with a heterozygous PV associated with biallelic CPS, and 42 patients without PVs. Children with a central nervous system (CNS) tumor had family members with CNS tumors more frequently than patients with other cancers (11/44, $p = 0.04$), but 42 of 44 (95.5%) cases did not have a PV in a CPS gene. **CONCLUSION:** These results demonstrate the value of systematically screening pediatric cancer patients for CPSs and indicate that a higher proportion of childhood cancers may be linked to predisposing germline variants than previously supposed.

Author summary

Traditionally pediatric cancer have been thought to be—mostly—caused by pure bad luck. In recent years, however, this notion has been challenged by novel findings as both maternal environmental exposure and genetic causes have been proven to increase the risk of certain pediatric cancers. With this study we have investigated a national cohort of pediatric cancer patients in Denmark. We have mapped family pedigree, made physical examination of the patients, and sequenced their genome, to get a 360-degree understanding of these patients. This revealed that a tenth of all patients carried a genetic variant causative of their cancer development. In addition, almost half of all patients were suspected of carrying a causative genetic variant based on tools that evaluate type of cancer, physical characteristics and family history. It also showed that tools to predict which patients carried a genetic variant did not identify all patients who in fact carried a genetic variant. Overall, roughly half of all patients were suspected of carrying an underlying genetic cause of their cancer, and a tenth had a verified underlying genetic variant predisposing to cancer pediatric cancer. This could suggest that the amount of pediatric cancer cases attributed to genetic factors may be even higher.

Introduction

In Europe, 15,000 children (1 in 300) are diagnosed with cancer each year.[1] Cancer can be attributed to genetic predisposition, exposure to carcinogens, and/or random mutations during cell division. Children are exposed to fewer carcinogens than adults.[2,3] Therefore, genetic predisposition and randomly acquired mutations are the major causes of most childhood cancers.

Cancer predisposition syndromes (CPSs) were previously considered rare among pediatric cancer patients, but increasing use of whole-exome sequencing (WES) and whole-genome sequencing (WGS) have identified up to 10% CPS among children, including several cases of CPS for adult-onset cancers not previously associated with childhood CPS. However, most studies investigated selected or single institution cohorts and included patients with specific diagnoses that were frequently associated with CPS.[4–6] Although some studies have included a broader range of pediatric cancer patients, [7–11] there have currently been no nationwide population-based studies. Moreover, most studies have focused on single nucleotide variants (SNVs) and few have included the effects of copy number variants (CNVs).[12]

Many clinical criteria have been developed to identify patients with CPS,[13–18] but these have not been validated in a national cohort.

Here we present the genetic SNV and CNV findings from the first 198 consecutive pediatric cancer patients included in the Danish, prospective, nationwide study Sequencing Tumor And Germline DNA—Implications for National Guidelines (STAGING).

Methods

Ethics statement

Ethical approval was obtained through the regional scientific ethical committee (the Ethical Scientific Committees for the Capital Region, H-15016782) and the Danish Data Protection Agency (RH-2016-219, I-Suite no: 04804). All parents/guardians and patients 15 years or older gave formal written consent to participation in this study.

Inclusion criteria and national setup

CPSs were defined as likely pathogenic or pathogenic variants (PVs) in a gene, predisposing the carrier to childhood- or adult-onset cancer. Between 1 July and 31 December 2016 we included 25 patients in the STAGING pilot study at Rigshospitalet (Copenhagen University Hospital, Denmark). On 1 January 2017, the study was expanded to all four pediatric oncology departments in Denmark. All patients were enrolled before June 2018.

Patients were eligible for inclusion if aged 0–17 years at diagnosis of a primary cancer including benign brain tumors, Langerhans cell histiocytosis (LCH), or myelodysplastic syndrome and parents spoke and read Danish.

Families were provided with written and oral information about the study by a research nurse or oncologist. A PhD student from STAGING (AB) or clinical geneticist provided genetic counselling to families interested in participating. Counselling sessions included pedigree construction (three generations), recording the child's clinical phenotypic features according to McGill Interactive Pediatric OncoGenetic Guidelines (MIPOGG)[18] and Jongmans' criteria[13] (Table 1), and explaining the potential consequences of genetic findings. These consequences included secondary findings, variants of unknown significance (VUS), implications of pathogenic findings associated with CPSs, and subsequent preventive and surveillance measures. Families choosing to enroll in the study were informed that PVs in 'actionable' genes listed by the American College of Medical Genetics and Genomics (ACMG)[19] would be disclosed to them. Families could select information regarding:

1: PVs in ACMG 'actionable' genes.

2: "1" and PVs in 314 known and putative cancer genes. Heterozygous variants in genes with solely recessive inheritance patterns were reported only if further familial genetic testing was warranted, in accordance with clinical guidelines.

3: In addition to "1" and "2", PVs in genes unrelated to CPSs (Table 2). Variants were only reported if clinical consequences were anticipated. Findings in these genes are not presented here.

Pedigrees covering 1st–3rd-generation family members were constructed for all patients. 1st-degree family members were parents and siblings, 2nd-degree family members were uncles/aunts and grandparents, and 3rd-degree family members were cousins, grandparents' siblings and great-grandparents. Cancer diagnoses were verified using unique civil registration numbers, which link family members to medical records, including pathological descriptions of cancer, in the Danish Pathology Data Bank. Living family members gave consent, whereas medical records of deceased family members could be retrieved without consent.

DNA sampling and sequencing

Genomic DNA was isolated from peripheral blood samples. For patients with hematologic malignancies, blood samples were drawn after remission, otherwise skin biopsies were obtained. Parental blood samples were collected to establish whether variants were paternally or maternally derived or occurred *de novo*.

Table 1. Tools to identify patients at risk of a cancer predisposition syndrome. Excerpt from the updated Jongmans' criteria[13] and McGill Interactive Pediatric OncoGenetic Guidelines (MIPOGG)[17].

JONGMANS' CRITERIA	
Criteria 1: Family history (3 generations)	Cancer history in the family: - 2 or more malignancies in family members <18 years of age - Any 1 st -degree relative with cancer <45 years of age - 2 or more 1 st - or 2 nd -degree relatives in same parental lineage with cancer <45 years of age Parents are consanguineous
Criteria 2: Neoplasm indicating underlying CPS	E.g., Hypodiploid ALL, botryoid rhabdomyosarcoma of the urogenital tract, gastrointestinal stromal tumors, retinoblastoma, schwannoma, subependymal giant cell astrocytoma
Criteria 3: Tumor analysis suggesting germline predisposition	E.g., Microsatellite instability in constitutional mismatch repair deficiency, loss of heterozygosity, other mutational signatures
Criteria 4: Patient with 2 or more malignancies	Secondary, bilateral, multifocal, or metachronous cancers
Criteria 5: Congenital or other phenotypic anomalies	Congenital anomalies (oral clefting, skeletal anomalies, facial dysmorphism) Developmental delay Growth anomalies Skin aberrations (café-au-lait spots, hypopigmentation, sun sensitivity) Immune deficiency
Criteria 6: Excessive toxicity related to cancer treatment	This criterion is not well defined and, based on an individual assessment by the pediatric oncologist/researcher, we have chosen not to include this criterion in this paper
MIPOGG CRITERIA	
Universal criteria	
Anamnestic criteria	- >1 primary tumor - Bilateral/multifocal primary tumors - Dysmorphic features/congenital abnormalities that the clinician deems to be related to cancer predisposition
Family anamnestic criteria	- Known cancer predisposition syndrome in the family - Close relative* with cancer <18 years OR a parent/sibling/half-sibling with cancer at <50 years - Close relative* with the same cancer type or same organ affected by cancer at any age - Close relative* with multiple primary tumors
Tumors for direct referral	
Tumors of the central nervous system and ocular tumors	Atypical teratoid rhabdoid tumor, choroid plexus carcinoma, dysplastic cerebellar gangliocytoma, endolymphatic sac tumor, hemangioblastoma, optic pathway glioma, pineoblastoma, pituitary adenoma, retinoblastoma, subependymal giant cell astrocytoma, vestibular schwannoma
Renal and neuroblastic tumors	Cystic nephroma, renal angiomyolipoma, renal cell carcinoma, renal rhabdoid tumor
Bone and soft-tissue tumors	Desmoid tumor, extrarenal rhabdoid tumor, Gardner fibroma, malignant periphery nerve sheath tumor, nasal chondromesenchymal hamartoma
Other tumors	Adrenocortical carcinoma, cardiac rhabdomyoma, colorectal carcinoma, gastrointestinal stromal tumor, hepatoblastoma, medullary thyroid cancer, ovarian Sertoli–Leydig cell tumor, parathyroid tumor, pheochromocytoma, paraganglioma, pleuropulmonary blastoma, trichilemmoma, small cell carcinoma of the ovary of hypercalcemic type, carcinoma of the breast, lung, cervix, uterus, or bladder

<https://doi.org/10.1371/journal.pgen.1009231.t001>

WGS was performed by the Norwegian Sequencing Center (Oslo, Norway) for the pilot study and by the Beijing Genomics Institute (Hong Kong, China) for the national study using the HiSeqX platform (Illumina, San Diego, CA, USA) with paired-end sequencing of

Table 2. Families could choose to receive one of the following levels of feedback from germline WGS of the affected child.

Level 1	Information regarding pathogenic or likely pathogenic variants in genes identified by the American College of Medical Genetics[19]. These genes are 'actionable' i.e., there are potential preventive, treatment, or surveillance modalities available. Half of these genes are related to CPS (primary findings); the others are related to cardiac disease, metabolic disorders, or familial hypercholesterolemia (secondary findings).
Level 2	In addition to the genes listed at level 1, information regarding pathogenic or likely pathogenic variants in other known or putative CPS genes (from the list of 314 CPS genes found in S1A Data). These were considered primary findings. However, if there was no known correlation between the clinical phenotype and the gene in question, the variant was considered a secondary finding.
Level 3	In addition to the genes listed at levels 1 and 2, families would also receive information regarding pathogenic or likely pathogenic variants in other genes not related to CPS (not presented in this paper). These were considered secondary findings.

<https://doi.org/10.1371/journal.pgen.1009231.t002>

150-bp reads and 30× average coverage. Reads were mapped to the hg19 reference genome sequence (GRCh37.p13; RefSeq assembly accession GCF_000001405.25) using BWA version 0.7.12,[20] and biobambam2 version 2.0.27[21] was used to sort and mark duplicate reads. Germline SNVs and indel variants were called with HaplotypeCaller using GATK version 3.8[22] or the DNaseq pipeline (Sentieon, San Jose, CA, USA). VarSeq software (version 2.2.0, Golden Helix, Bozeman, MT, USA) was used to annotate variants. Moreover, filtration was based on read depth ≥ 8 , genotype quality ≥ 20 , and variant allele frequency (VAF) ≥ 0.2 , and sequence ontology was used to exclude intronic and intergenic variants and variants located in the 3' and 5' untranslated region (UTR) prior to evaluation. Integrative Genomics Viewer (IGV, version 2.8.2, Boston, MA, USA) was used to visualize read alignments. Manta and CNVkit were applied for calling larger structural rearrangements.[23,24] All variants were reported according to HGVS nomenclature guidelines.[25] WGS data were filtered for PVs in the 59 ACMG 'actionable' genes and 314 cancer genes (S1 Data). The cancer gene panel was selected from Zhang *et al.*,[7] Rahman,[26] and novel genes recently linked to childhood or adult CPSs. All variants with a minor allele frequency of $< 1\%$ in any large population (gnomAD) were tabulated. For CPS genes with higher variant frequencies in the general population (e.g., *ATM*, *CHEK2*), a separate filter was used. We did not apply a specific variant filter to identify mosaicism. Variants were assessed by a team of clinical geneticists and molecular biologists based on variant type (e.g., frameshift, nonsense, missense), computational predictions of effect on protein and RNA function (e.g., Combined Annotation Dependent Depletion [CADD], PHRED quality score, ADA splice prediction score),[27] and database searches for published literature on each variant. Moreover, we used Alamut Visual 2.10 to evaluate variants effect on splicing (<https://www.interactive-biosoftware.com/alamut-visual/>). The effects of variants were considered significant if the scores of at least three programs were reduced $\geq 10\%$ or a strong cryptic acceptor or donor site was generated. Variants were classified as pathogenic (class 5), likely pathogenic (class 4), VUS (class 3), likely benign (class 2), and benign (class 1).[28] Class 4 and 5 variants were designated 'PVs'. Class 3 variants, especially those that potentially matched the child's diagnosis, were further investigated by segregation analysis and splice predictions, and tumor RNA sequencing was used to assess loss-of-heterozygosity (LOH) if tissue was available (Fig 1). In addition, we used the machine-learning tool ORVAL to predict whether combinations of genetic variants were likely to be pathogenic.[29] Variants were discussed at regular multidisciplinary meetings by pediatric oncologists, clinical geneticists, and bioinformaticians. PVs were verified by Sanger or next-generation sequencing before parents were informed.

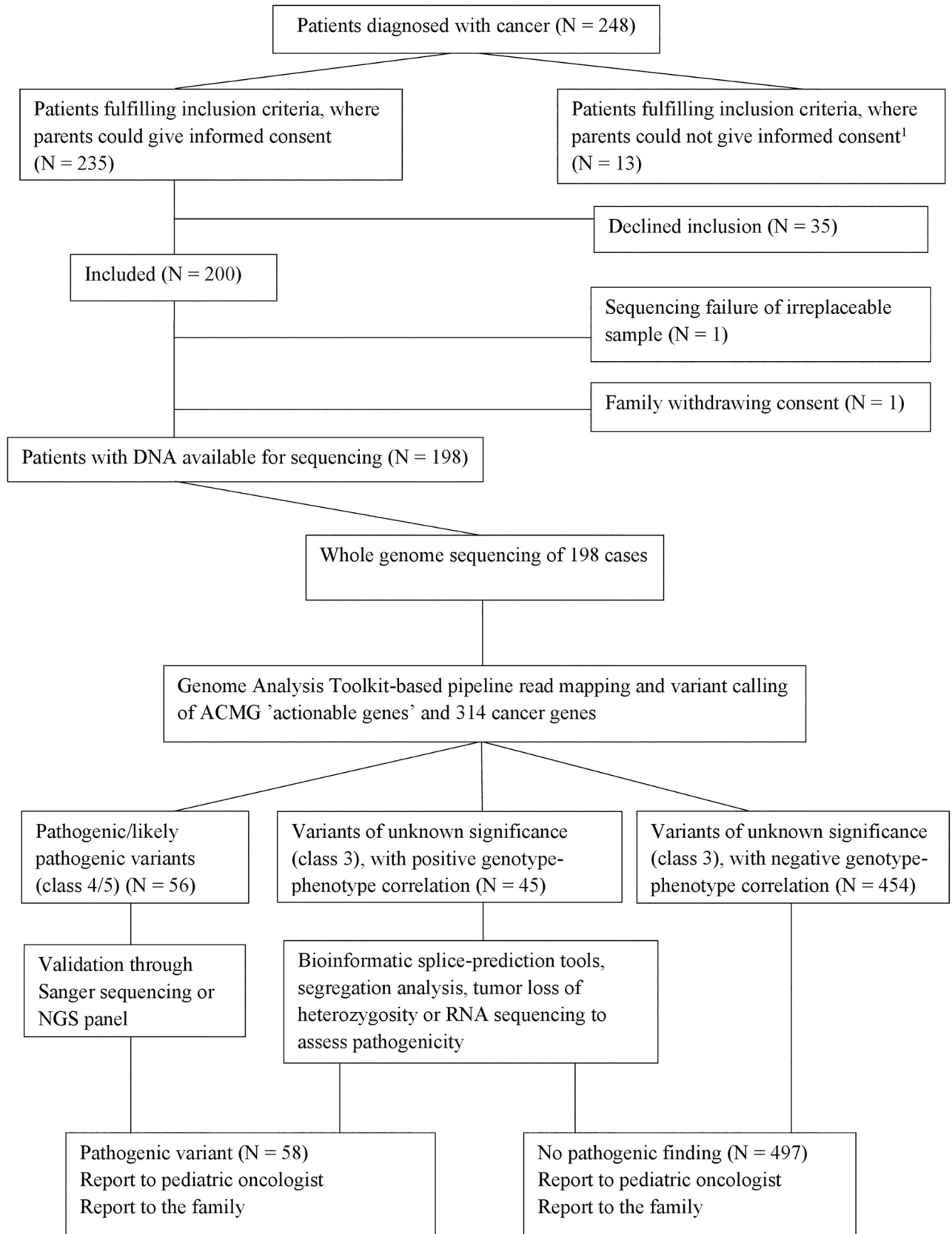


Fig 1. Inclusion and sequencing strategy. Variants presented are in genes associated with CPS. All variants of unknown significance have a CADD-PHRED score >20 and an allele frequency <1%. ¹Patients whose parents were not able to give informed consent due to language barriers or social issues (mainly parental psychiatric/severe somatic disease).

<https://doi.org/10.1371/journal.pgen.1009231.g001>

Statistical analysis

Patient/parental characteristics were compared using Pearson chi-square test. Two-sided *P*-values below 0.05 were considered statistically significant. Statistical calculations were carried out using the R software version 3.6.0.

Results

Patient characteristics

Of 248 consecutive pediatric cancer patients that fulfilled the inclusion criteria, 198 families consented to participation (Fig 1, Table 3). The included patients did not differ significantly ($p > 0.05$) from the 35 patients that declined to participate with regard to sex, age, or diagnoses. Of 198 families, 155/198 (78.3%) opted for feedback on all findings, 24/198 (12.1%) opted for feedback on PVs in 'actionable' genes and 314 cancer genes, and 19/198 (9.6%) opted for feedback on PVs in ACMG 'actionable' genes only. There were no significant differences ($p > 0.05$) in parental age, educational level, or income among families opting for the three levels of feedback.

Patients with known and suspected CPSs

Overall, 94/198 (47.5%) patients carried a PV in a CPS gene or were suspected of having an underlying CPS based on Jongmans'/MIPOGG criteria or family cancer history (Fig 2).

Twenty-nine of 198 (14.6%) patients carried PVs in at least one CPS gene, including four patients with trisomy 21. Of these, 21/198 (10.6%) had PVs in genes predisposing toward childhood-onset CPS, including *CDC73*, *DDX41* (biallelic, putative childhood-onset cancer gene), *LTZR1*, *NF1*, *RB1*, *SDHC*, *SMARCA4*, *TP53*, and *TSC2*, uniparental disomy (UPD) of chromosome 11p (clinical analysis), and trisomy 21.[30]

In addition, 9/198 (4.5%) patients had PVs in genes predisposing toward adult-onset CPS, one of whom also carried a deletion in a childhood-onset CPS gene. Adult-onset CPS variants were identified in *APC*, *ATM*, *AXIN2*, *BARD1*, *BRCA2*, *CHEK2*, *MUTYH*, and *PALB2*, some conferred a high risk of cancer[31–33] whereas others conferred a moderate risk[34–38] (Table 4). The specific variants in *APC* and *ATM* are only associated with an adult-onset CPS, had the specific variants been associated with childhood-onset CPS, these genes would have been listed above.

These 29 patients had 31 PVs in total, including seven frameshift, five nonsense, eight missense, and three splice-site variants. Three variants were larger deletions of at least one exon, one patient had UPD 11p, and four patients had trisomy 21.

Some CPSs occurred in several patients (Table 4). Two patients had PVs in two CPS genes: a patient with LCH had PVs in *BRCA2* and *AXIN2*; a patient with a malignant peripheral nerve sheath tumor had PVs in *NF1* and *PALB2*. All CPS PVs were monoallelic, except for one patient with biallelic PVs in *DDX41*. [39]

Of the 21 childhood-associated CPSs 18 (85.7%) had previously established links between genotype (e.g., trisomy 21 and leukemia) and cancers (Table 4). This study identified a PV in eight of the 21 (38.1%) pediatric CPS patients, whereas 13/21 (61.9%) patients had a previously established genetic predisposition syndrome (e.g., *NF1* or trisomy 21). Such a connection was established through clinical diagnosis and/or genetic testing. Among these 21 patients, only one had a family member diagnosed with cancer before 18 years of age (Table 5).

Table 3. Patients distributed according to sex, age at diagnosis, diagnosis, level of feedback and fulfillment of the Jongmans' and MIPOGGs criteria.

	N (%)
Sex	
Male	121 (61.1%)
Female	77 (38.9%)
Age at diagnosis	
0–5 years	104 (52.5%)
6–10 years	41 (20.7%)
11–15 years	39 (19.7%)
16–18 years	14 (7.0%)
Diagnosis	
<u>Hematologic cancer</u>	105 (53.0%)
Precursor B-ALL	45
Lymphoma	22
AML/CML/other myeloid leukemia	17
Precursor T-ALL	10
Langerhans cell histiocytosis	6
Myelodysplastic syndrome	3
Mixed lineage ALL	2
<u>Tumors of the central nervous system</u>	44 (22.2%)
Low-grade glioma, WHO grade I–II	17
High-grade glioma, WHO grade III–IV	6
Ependymoma	4
AT/RT	3
Medulloblastoma	2
Schwannoma	2
Other	10
<u>Solid tumors</u>	49 (25.0%)
Wilms tumor	8
Neuroblastoma	8
Rhabdomyosarcoma	6
Osteosarcoma	5
Retinoblastoma	5
Ewing's sarcoma	4
Malignant peripheral nerve sheath tumor	1
Other	12
Level of feedback	
Full feedback	155 (78.2%)
Limited feedback	24 (12.1%)
No feedback	19 (9.6%)
Fulfillment of Jongmans' criteria	
Fulfilled one or more criteria	56 (28.3%)
Did not fulfill any criteria	142 (71.7%)
Referral for genetic evaluation recommended by MIPOGG	
Referral recommended	64 (32.3%)
Referral not recommended	134 (67.7%)

<https://doi.org/10.1371/journal.pgen.1009231.t003>

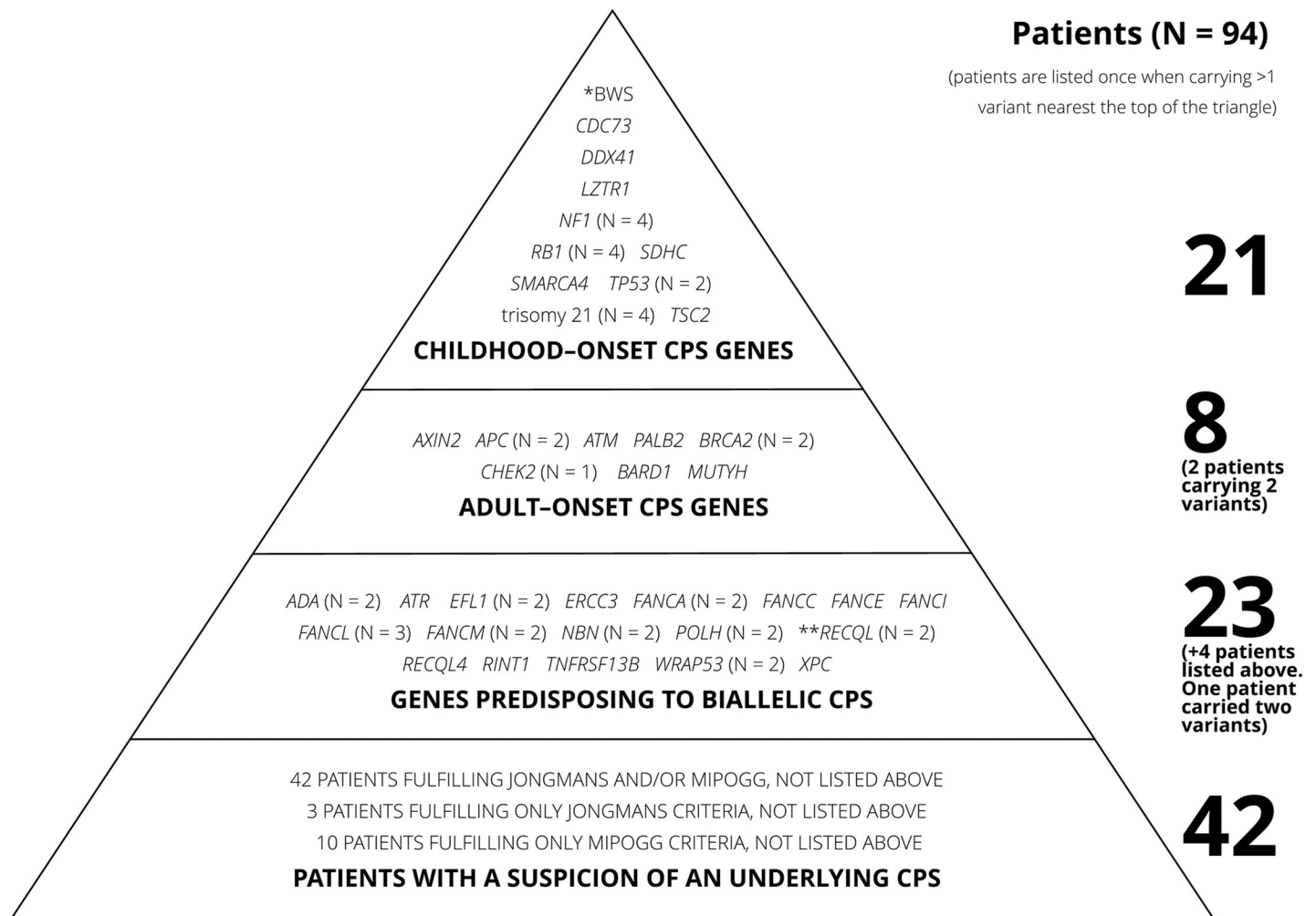


Fig 2. Triangle of patients with confirmed or suspected underlying cancer predisposition syndrome. Patients fulfilling criteria for more than one level of the triangle were only counted once, closest to the top of the triangle. The column on the left shows the number of pathogenic variants on each level. The column on the right shows the number of patients on each level.

<https://doi.org/10.1371/journal.pgen.1009231.g002>

PVs in biallelic CPS

Twenty-seven of 198 (13.6%) patients carried one ($n = 22$) or two ($n = 1$, *FANCM* and *ADA*) PVs predisposing toward CPS through biallelic inheritance, of whom four also had a PV in a monoallelic CPS gene (Table 6). Variants were found in *ADA*, *ATR*, *EFL1*, *ERCC3*, *FANCA*/*C*/*E*/*I*/*L*/*M*, *NBN*, *POLH*, *RECQL*, *RECQL4*, *RINT1*, *WRAP53*, and *XPC* (Fig 2).

Seven of the 198 (3.5%) patients carried two or more PVs/trisomy 21 (Table 7). Of these seven patients, one carried variants bioinformatically predicted to be oligogenically pathogenic. This patient carried biallelic variants in *DDX41* and a monoallelic variant in *NBN*. The digenic combination of a single *DDX41* variant and the *NBN* variant were also predicted to be pathogenic. No variant combinations in the six remaining patients were predicted to be oligogenically pathogenic.

Variants of unknown significance

All 198 patients carried VUS in one or more CPS genes. VUS with a frequency <1% and a CADD/PHRED score >20 are listed in S1 Data. Thirty-nine of 198 (19.7%) patients had a

Table 4. Pathogenic and likely pathogenic variants in cancer predisposition genes.

Diagnosis	Gene	Variant	Protein	Corresponding to phenotype of the patient	Inherited (M/P) ¹ / <i>de novo</i>	CADD-PHRED ² score
Pathogenic/likely pathogenic variants (childhood onset)						
Acute myeloid leukemia	<i>CDC73</i>	c.358C>T	p.(Arg120*)	-	-	37.0
Plasmacytoid dendritic cell leukemia	<i>DDX41</i> ³	c.962C>T	p.(Pro321Leu)	+	M	31.0
		c.937G>C	p.(Gly313Arg)	+	P	32.0
Osteosarcoma	<i>LZTR1</i>	c.955C>T	p.(Gln319*)	-	-	38.0
Optic nerve glioma	<i>NF1</i>	c.5242C>T	p.(Arg1748*)	+	M	38.0
Acute myeloid leukemia	<i>NF1</i>	c.288+1delG	p.(?)	+	-	34.0
Malignant peripheral nerve sheath tumor ⁴	<i>NF1</i>	c.61-22171_4110+2458del (deletion of exon 2-30)	p.(Leu21_Gln1370del)	+	M	-
Pilocytic astrocytoma	<i>NF1</i>	c.2033dupC	p.(Ile679Aspfs*21)	+	-	34.0
Retinoblastoma	<i>RB1</i>	c.1735C>T	p.(Arg579*)	+	-	36.0
Retinoblastoma	<i>RB1</i>	c.409_412delGAAA	p.(Glu137Leufs*15)	+	P	36.0
Retinoblastoma	<i>RB1</i>	c.219_220delAG	p.(Arg73Serfs*36)	+	<i>De novo</i>	31.0
Retinoblastoma	<i>RB1</i>	c.2663+2T>C	p.(?)	+	-	25.2
Acute promyelocytic leukemia	<i>SDHC</i>	c.148C>T	p.(Arg50Cys)	-	-	34.0
Small cell carcinoma ⁵	<i>SMARCA4</i> ⁶	c.2002-625_2124-1045del (deletion of exon 14)	p.(Glu669Cysfs*7)	+	-	
Precursor B-cell acute lymphoblastic leukemia	<i>TP53</i>	c.637C>G	p.(Arg213Gly)	+	P	28.4
Osteosarcoma	<i>TP53</i>	c.818G>A	p.(Arg273His)	+	<i>De novo</i>	27.3
Subependymal giant cell astrocytoma	<i>TSC2</i>	c.4141dupC	p.(Leu1382Profs*32)	+	<i>De novo</i>	23.5
Wilms tumor ⁷	Paternal uniparental disomy of chromosome 11, corresponding to Beckwith-Wiedemann syndrome			+	P	-
Acute megakaryoblastic leukemia ⁸	47,XY,+21			+	-	-
Acute megakaryoblastic leukemia ⁸	47,XY,+21			+	-	-
Hodgkins lymphoma	47,XX,+21			+	-	-
Acute lymphoblastic leukemia	47,XX,+21			+	-	-
Pathogenic/likely pathogenic variants (adult onset)						
Wilms tumor	<i>APC</i> ^{9, 10}	c.3920T>A	p.(Ile1307Lys)	-	-	5.3
Neuroblastoma	<i>APC</i> ^{9, 10}	c.3920T>A	p.(Ile1307Lys)	-	-	5.3
Rhabdomyosarcoma	<i>ATM</i> ^{9, 10}	c.9023G>C	p.(Arg3008Pro)	-	-	32.0
Wilms tumor	<i>BARD1</i> ¹⁰	g.215591264-215774591 (Deletion of exon 1-11)	p.(?)	-	-	-
Langerhans cell histiocytosis	<i>BRCA2</i>	c.5722_5723delCT	p.(Leu1908Argfs*2)	-	M	21.2
		c.815+1G>A	p.(?)	-	M	24.5
T-cell acute lymphoblastic leukemia	<i>BRCA2</i>	c.6486_6489delACAA	p.(Lys2162Asnfs*5)	-	-	29.1
T-cell acute lymphoblastic leukemia	<i>CHEK2</i> ¹⁰	c.1100del	p.(Thr367Metfs*15)	-	-	35.0
Precursor B-cell acute lymphoblastic leukemia	<i>MUTYH</i> ¹⁰	c.536A>G	p.(Tyr179Cys)	-	-	24.7

(Continued)

Table 4. (Continued)

Diagnosis	Gene	Variant	Protein	Corresponding to phenotype of the patient	Inherited (M/P) ¹ / <i>de novo</i>	CADD-PHRED ² score
Pathogenic/likely pathogenic variants (childhood onset)						
Malignant peripheral nerve sheath tumor ⁴	<i>PALB2</i>	c.2736G>A	p.(Trp912*)	–	P	41.0

– Value not known/applicable (e.g., parents not tested)

¹M: maternally inherited, P: paternally inherited

²CADD: combined annotation dependent depletion, PHRED: quality score

³Putative childhood CPS gene

⁴Same patient carrying the *NF1* deletion and *PALB2* mutation

⁵The initial diagnosis (synovial sarcoma) was revised later revised

⁶Validation in process

⁷Detected by clinical analysis

⁸These patients are twin brothers

⁹The specific variants in *APC* and *ATM* are only associated with an adult-onset cancer predisposition syndrome; had the variant been associated with childhood cancer predisposition syndrome, these genes would have been listed above.

¹⁰These variants confer a moderate risk of cancer in adulthood

<https://doi.org/10.1371/journal.pgen.1009231.t004>

VUS in DNA repair pathway genes (*ATM*, *BLM*, *NBN*, *MLH1*, *MSH2*, *MSH6*, *ERCC4*, *FANCA/E/F/G/I/L*, *BRCA1*, *BRCA2*, *RAD51C*, *RFWD3*, *SLX4*), 16/198 (8.1%) patients had a VUS in genes associated with bone marrow failure (*RPL5*, *RPS10*, *RPS19*, *CTC1*, *DKC1*, *NHP2*, *PARN*, *RTEL1*, *TERT*, *SBDS*), 12/198 (6.1%) patients had a VUS in the RAS pathway (*CBL*, *NF1*, *A2ML1*, *MAP2K2*, *PTPN11*, *RAF1*, *RRAS*, *SHOC2*), and nine of 198 (4.5%) patients had a VUS in genes associated with familial leukemia (*ETV6*, *GATA2*, *IKZF1*, *RUNX1*, *PAX5*). Thirty-eight of the 198 (19.2%) patients had a VUS in a gene previously associated with the cancer diagnosed in the child. When subgroups were separated by diagnoses, 30 of 105 (28.6%) patients with hematologic malignancies had a VUS in a relevant gene associated with childhood onset (*A2ML1*, *ADA*, *ATM*, *BLM*, *BRCA2*, *CREBBP*, *DDX41*, *DNAJC21*, *EFL1*, *EP300*, *ERCC6L2*, *FANCF*, *MAP2K2*, *PARN*, *PAX5*, *PTPN11*, *RPL5*, *RRAS*, *SH2D2A*, *SHOC2*, *SOS2*, *TERT*, *TNFRSF13B*). One of the 44 (2.3%) patients with central nervous system (CNS) tumors had a VUS in a gene predisposing to Fanconi Anemia in which childhood-onset brain tumors have been described (*FANCG*), and seven of the 49 (14.3%) patients with solid tumors had a VUS in a relevant gene associated with childhood onset (*BRCA2*, *CDH23*, *EP300*, *ERCC6L2*, *FANCI*, *RBI*, *RFWD3*, *XRCC2*). Most variants occurred in DNA repair pathway genes (S1 Data). Four patients had a VUS, but none had a PV, in a mismatch repair pathway gene.

Fulfillment of clinical criteria indicating an underlying CPS

All patients were evaluated using a phenotype checklist developed for this study (S1 Text), and 116/198 (58.6%) patients had one or more CPS-associated findings.

Overall, 70/198 (35.4%) patients fulfilled Jongmans' ($n = 56$, 28.3%) and/or MIPOGG criteria ($n = 64$, 32.3%) including 17 (81.0%) of the 21 patients with a childhood-onset CPS (Tables 8 and 9). Of the four patients with a PV in a childhood-onset CPS gene who did not fulfill Jongmans' criteria, none had excessive chemotherapy-induced toxicity. Patients not identified by either tool included two with PVs in *TP53*, and one with a pathogenic *SMARCA4* variant. The patient with a PV in *CDC73* was identified by MIPOGG and not by Jongmans. The *SMARCA4*-deletion patient was diagnosed with synovial sarcoma of the ovary, revised to

Table 5. Family history of patients with pathogenic and likely pathogenic variants in childhood-onset and adult-onset cancer genes.

Patient diagnosis	Gene	Variant	Protein	Family history
Patients with childhood-onset CPS				
Acute myeloid leukemia	<i>CDC73</i>	c.358C>T	p.(Arg120*)	2 nd -degree relative (maternal side): grandfather: chronic lymphoblastic leukemia, 62 years
Plasmacytoid dendritic cell leukemia	<i>DDX41</i> <i>DDX41</i>	c.962C>T c.937G>C	p.(Pro321Leu) p.(Gly313Arg)	No cancer cases in 1 st -2 nd -degree relatives Sister: intellectual disability corresponding to the phenotype of the syndrome identified in this patient
Osteosarcoma	<i>LZTR1</i>	c.955C>T	p.(Gln319*)	2 nd -degree relative (maternal side): grandfather: lymphoma, 68 years
Optic nerve glioma	<i>NF1</i>	c.5242C>T	p.(Arg1748*)	No cancer cases in 1 st -2 nd -degree relatives
Acute myeloid leukemia	<i>NF1</i>	c.288+1delG	p.(?)	No cancer cases in 1 st -2 nd -degree relatives
Malignant peripheral nerve sheath tumor ¹	<i>NF1</i>	c.61-22171_4110+2458del (deletion of exon 2–30)	p. (Leu21_Gln1370del)	No cancer cases in 1 st -2 nd -degree relatives
Pilocytic astrocytoma	<i>NF1</i>	c.2033dupC	p.(Ile679Aspfs*21)	2 nd -degree relative (maternal side): grandfather: colon cancer, 84 years
Retinoblastoma	<i>RB1</i>	c.1735C>T	p.(Arg579*)	No cancer cases in 1 st -2 nd -degree relatives
Retinoblastoma	<i>RB1</i>	c.409_412delGAAA	p.(Glu137Leufs*15)	1 st -degree relative: father, retinoblastoma, 0 years 2 nd -degree relatives (paternal side): father's sister: retinoblastoma (0 years), rhabdomyosarcoma (14 years), melanoma (20 years). Grandfather: melanoma (39 years), myelofibrosis (48 years)
Retinoblastoma	<i>RB1</i>	c.219_220delAG	p.(Arg73Serfs*36)	No cancer cases in 1 st -2 nd -degree relatives
Retinoblastoma	<i>RB1</i>	c.2663+2T>C	p.(?)	No cancer cases in 1 st -2 nd -degree relatives
Acute promyelocytic leukemia	<i>SDHC</i>	c.148C>T	p.(Arg50Cys)	No cancer cases in 1 st -2 nd -degree relatives
Small cell carcinoma	<i>SMARCA4</i>	c.2002-625_2124-1045del (Deletion of exon 14)	p.(Glu669Cysfs*7)	2 nd -degree relative (paternal side): grandfather: lung cancer, 71 years, (maternal side): grandfather: lung cancer, 79 years
Precursor B-cell acute lymphoblastic leukemia	<i>TP53</i>	c.637C>G	p.(Arg213Gly)	1 st -degree relative: father: pheochromocytoma, 34 years (not diagnosed at the time of the child's diagnosis) 2 nd -degree relative (paternal side): grandmother: hepatic cholangiocarcinoma, 36 years
Osteosarcoma	<i>TP53</i>	c.818G>A	p.(Arg273His)	2 nd -degree relative (paternal side): grandfather: colon cancer, 84 years
Subependymal giant cell astrocytoma	<i>TSC2</i>	c.4141dupC	p.(Leu1382Profs*32)	2 nd -degree relative (maternal side): grandfather: urothelial carcinoma, 71 years
Wilms tumor ²	Paternal uniparental disomy of chromosome 11, corresponding to Beckwith–Wiedemann syndrome			2 nd -degree relatives (maternal side): grandmother: lung cancer, 64 years, mother's brother: urothelial carcinoma, 41 years
Acute megakaryoblastic leukemia (two patients, twins)	47,XY,+21			2 nd -degree relative (paternal side): grandmother: lung cancer, 50 years
Hodgkin lymphoma	47,XX,+21			2 nd -degree relatives (maternal side): grandmother: ovarian cancer, 68 years, (paternal side): grandmother: gastrointestinal stromal tumor, 82 years
Precursor B-cell acute lymphoblastic leukemia	47,XX,+21			2 nd -degree relative: mother's sister: melanoma, 37 years
Patients with adult-onset CPS				
Wilms tumor	<i>APC</i>	c.3920T>A	p.(Ile1307Lys)	2 nd -degree relative (paternal side): grandmother: urothelial carcinoma, 74 years
Neuroblastoma	<i>APC</i>	c.3920T>A	p.(Ile1307Lys)	No cancer cases in 1 st -2 nd -degree relatives
Rhabdomyosarcoma	<i>ATM</i>	c.9023G>C	p.(Arg3008Pro)	2 nd -degree relative (maternal side): mother's brother: tumor on heart valve, 0 years
Wilms tumor	<i>BARD1</i>	g.215591264-215774591 (Deletion of exon 1–11)	p.(?)	No cancer cases in 1 st -2 nd -degree relatives
Langerhans cell histiocytosis	<i>BRCA2</i> <i>AXIN2</i>	c.5722_5723delCT c.815+1G>A	p.(Leu1908Argfs*2) p.(?)	2 nd -degree relative (paternal side): grandfather: esophageal cancer, 59 years

(Continued)

Table 5. (Continued)

Patient diagnosis	Gene	Variant	Protein	Family history
Patients with childhood-onset CPS				
T-cell acute lymphoblastic leukemia	<i>BRCA2</i>	c.6486_6489delACAA	p.(Lys2162Asnfs*5)	2 nd -degree relative (maternal side): grandmother: breast cancer, 63 years
T-cell acute lymphoblastic leukemia	<i>CHEK2</i>	c.1100del	p.(Thr367Metfs*15)	2 nd -degree relatives (paternal side): grandfather: prostate cancer, 65 years, grandmother: cervical cancer, 54 years
Precursor B-cell acute lymphoblastic leukemia	<i>MUTYH</i>	c.536A>G	p.(Tyr179Cys)	No cancer cases in 1 st -2 nd -degree relatives
Malignant peripheral nerve sheath tumor ¹	<i>PALB2</i>	c.2736G>A	p.(Trp912*)	No cancer cases in 1 st -2 nd -degree relatives

¹Same patient carrying these two variants.

²Not identified by whole-genome sequencing.

<https://doi.org/10.1371/journal.pgen.1009231.t005>

small-cell carcinoma based on this study, and would have fulfilled both criteria if the initial diagnosis had been correct. Of the patients with adult-onset CPS, 2/9 (22.2%) fulfilled Jongmans' ($n = 2$) and MIPOGG ($n = 1$) criteria, which is significantly fewer than for childhood-onset CPS ($p = 0.0022$). Of the additional 23 patients with a heterozygous PV predisposing to biallelic CPS, eight (34.8%) fulfilled Jongmans' ($n = 5$) and MIPOGG ($n = 7$) criteria.

The number of VUS identified were higher among patients without a CPS and with adult-onset CPS compared to patients with a childhood-onset CPS, in the first group patients on average carried 2.5 VUS compared to 1.6 VUS in the latter group. The same was the case when comparing patients with a childhood-onset CPS to patients who solely fulfilled Jongmans/MIPOGGs criteria, patients carrying a childhood-onset CPS carried an average of 1.6 VUS compared to 2.5 VUS among patients fulfilling Jongmans/MIPOGGs criteria alone.

Family histories of cancer

Parents reported cancer diagnoses for 704 family members, 106 of whom resided outside Denmark, precluding further verification. Cancer diagnoses were verified for 328 (54.8%) of the remaining 598 family members, whereas the others did not consent to retrieval of medical records ($n = 45$) or their diagnoses could not be verified ($n = 225$) due to difficulties identifying distant/deceased family members or cancer occurrence prior to registration in Danish registries (before 1943). For 1st-, 2nd-, and 3rd-degree relatives 16 (100.0%), 133 (84.1%), and 179 (42.2%) cases were verified, respectively. The following is based on verified diagnoses and family recollection (for 1st to 3rd generation family members) when verification was impossible.

In total, 191/198 (96.5%) participants had a family history of cancer. Seven of 198 (3.5%) participants had a family member diagnosed with cancer before 18 years of age (two had a CPS). Fifty-six of 198 (28.3%) participants had at least one relative diagnosed with cancer between the ages of 18 and 45. Three of 198 (1.5%) participants had two or more relatives under the age of 45 diagnosed with cancer.

Forty-three of 198 (21.7%) participants had a relative with a cancer of the same organ system as the patient. Patients with hematologic malignancies and solid tumors did not have more family members with cancers of the same organ system than the other two patient groups. In contrast, patients with a CNS tumor had a family member with a malignancy in the CNS more frequently than patients with either solid tumors or hematologic malignancies ($p = 0.04$). This association also held ($p = 0.04$) when patients with a CPS were eliminated (Table 10). Family history for the patients with a confirmed CPS can be found in Table 5.

Table 6. Patients with pathogenic or likely pathogenic variants in biallelic cancer predisposition genes.

Diagnosis	Gene	Variant	Protein	CADD-PHRED score
Acute myeloid leukemia ¹	<i>ADA</i>	c.646G>A	p.(Gly216Arg)	28.0
Precursor B-cell acute lymphoblastic leukemia	<i>ADA</i>	c.1078+2T>A	p.(?)	22.9
Chronic myeloid leukemia	<i>ATR</i>	c.2320dupA	p.(Ile774Asnfs*3)	20.1
Rhabdomyosarcoma	<i>EFL1</i>	c.159+3A>G	p.(?)	14.84
Precursor T-cell acute lymphoblastic leukemia	<i>EFL1</i>	c.2430_2431delCC	p.(Leu811Asnfs*10)	15.9
Wilms tumor	<i>ERCC3</i>	c.1115_1120dupAGCAGT	p.(Trp374*)	37.0
Astrocytoma	<i>FANCA</i>	c.3482C>T	p.(Thr1161Met)	17.3
Precursor B-cell acute lymphoblastic leukemia	<i>FANCA</i>	c.3391A>G	p.(Thr1131Ala)	23.5
Acute myeloid leukemia	<i>FANCC</i>	c.535C>T	p.(Arg179*)	35.0
Yolk sac tumor	<i>FANCE</i>	c.108delG	p.(Pro37Leufs*47)	16.7
Precursor B-cell acute lymphoblastic leukemia	<i>FANCI</i>	c.158-5A>G	p.(?)	12.7
Precursor T-cell acute lymphoblastic leukemia	<i>FANCL</i>	c.540+1G>A	p.(?)	23.3
Precursor B-cell acute lymphoblastic leukemia	<i>FANCL</i>	c.1007_1009delTAT	p.(Ile336_Cys337delinsSer)	23.5
Lymphoma	<i>FANCL</i>	c.1096_1099dupATTA	p.(Thr367Asnfs*13)	35.0
Acute myeloid leukemia ¹	<i>FANCM</i>	c.681+1G>C	p.(?)	25.9
Rhabdomyosarcoma	<i>FANCM</i>	c.2156_2160delAACCA	p.(Lys719Serfs*15)	36.0
Plasmacytoid dendritic cell leukemia	<i>NBN</i>	c.156_157delTT	p.(Ser53Cysfs*8)	25.6
Wilms tumor	<i>NBN</i>	c.834dupA	p.(Gln279Serfs*6)	37.0
Precursor B-cell acute lymphoblastic leukemia	<i>POLH</i>	c.1600_1610delCA	p.(Gln534Glufs*11)	40.0
Precursor B-cell acute lymphoblastic leukemia	<i>POLH</i>	c.491-69_660+30del (deletion of exon 5)	p.(Glu164Glyfs*37)	–
Ganglioglioma	<i>RECQL</i>	c.1859C>G	p.(Ser620*)	38.0
Craniopharyngioma	<i>RECQL</i>	c.1859C>G	p.(Ser620*)	38.0
Precursor B-cell acute lymphoblastic leukemia	<i>RECQL4</i>	c.3072delA	p.(Val1026Cysfs*18)	21.7
Lymphoma	<i>RINT1</i>	c.88+3A>G	p.(?)	15.4
Glioma ²	<i>TNFRSF13B</i>	c.431C>G	p.(Ser144*)	35.0
Craniopharyngioma	<i>WRAP53</i>	c.1192C>T	p.(Arg398Trp)	34.0
Precursor B-cell acute lymphoblastic leukemia	<i>WRAP53</i>	c.1192C>T	p.(Arg398Trp)	34.0
Neurofibroma	<i>XPC</i>	c.1934delC	p.(Pro645Leufs*5)	26.6

Three of the patients listed above also had a monoallelic pathogenic germline mutation (fam no. 13, 25 and 51).

Monoallelic variants in *RECQL* are pathogenic; however, their relationship to cancer is uncertain. Therefore, they are listed here.

¹Same patient carrying the *ADA* and *FANCM* variants.

²Pathogenic variants (Romberg *et al.*, 2013) may be inherited via an autosomal dominant or autosomal recessive pattern. Based on the patient's phenotype, this variant was considered inherited recessively.

<https://doi.org/10.1371/journal.pgen.1009231.t006>

Secondary findings

Two patients had PVs in genes associated with familial hypercholesterolemia (*APOB* and *LDLR*), three had PVs in genes associated with arrhythmic right ventricular cardiomyopathy (*DSC2*, *DSG2*, and *PKP2*), and one patient had a PV in *KCNQ1*, which is associated with long QT syndrome. These six variants are associated with increased risk of disease and families were informed. Overall, 18 patients (9.1%) had an ACMG 'actionable' PV, including 12 PVs in CPS genes and six PVs in genes associated with other non-malignant diseases (Table 11).

Discussion

In this first nationwide unselected cohort of consecutive pediatric cancer patients, half had a likely or validated underlying CPS, based on WGS, clinical examination, and pedigree

Table 7. Patients carrying more than one pathogenic/likely pathogenic variant.

Diagnosis of the patient	Gene	Variant	Protein	Gene	Variant	Protein	Gene	Variant	Protein
Malignant peripheral nerve sheath tumor	<i>NF1</i>	c.61-22171_4110+2458del (deletion of exon 2–30)	p.(Leu21_Gln1370del)	<i>PALB2</i>	c.2736G>A	p.(Trp912*)	–	–	–
Optic nerve glioma	<i>NF1</i>	c.5242C>T	p.(Arg1748*)	<i>XPC</i>	c.1934delC	p.(Pro645Leufs*5)	–	–	–
B-cell acute lymphoblastic leukemia	47,XX,+21			<i>POLH</i>	c.1600_1610delCA	p.(Gln534Glufs*11)	–	–	–
Plasmacytoid dendritic cell leukemia	<i>DDX41</i>	c.962C>T	p.(Pro321Leu)	<i>DDX41</i>	c.937G>C	p.(Gly313Arg)	<i>NBN</i>	c.834dupA	p.(Gln279Serfs*6)
Wilms tumor	<i>BARD1</i>	g.215591264-215774591 (deletion of exon 1–11)	p.(?)	<i>ERCC3</i>	c.1115_1120dupAGCAGT	p.(Trp374*)	–	–	–
Langerhans cell histiocytosis	<i>BRCA2</i>	c.5722_5723delCT	p.(Leu1908Argfs*2)	<i>AXIN2</i>	c.815+1G>A	p.(?)	–	–	–
Acute myeloid leukemia	<i>ADA</i>	c.646G>A	p.(Gly216Arg)	<i>FANCM</i>	c.681+1G>C	p.(?)	–	–	–

<https://doi.org/10.1371/journal.pgen.1009231.t007>

mapping, and 14.6% had a genetically verified CPS. These findings strongly indicate that genetic predisposition to childhood cancer may be far more common than previously supposed. Furthermore, other modes of inheritance (di-, oligo- and polygenic risk) may play significant roles in pathogenesis.

Our childhood-onset CPS results are consistent with previous studies that found a CPS in 7–10% of pediatric cancer patients.[8–11]

The frequency of PVs in our study was significantly higher than that observed in control cohorts wherein 0.6–1.1% of adult patients in a Genomics England cohort and a cohort of pediatric and adult patients with autism had PVs in CPS genes.[7] This study, however, excluded patients with known CPS and included different genes (including genes with frequent somatic variants) and is thus not directly comparable. A study of osteosarcoma patients found a remarkably high frequency of CPSs in control cohorts (12.1% and 9.3%), probably because many of the genes included (e.g., *PMS1*, and *COL7A1*) were not definitely linked to cancer.[6] Other studies have included adult-onset CPS genes. For example, Zhang *et al.* found only 0.7% of their patients carried an adult-onset CPS variant (*BRCA1/2* and *PALB2*).[7] We found that 1.5% of our patients carried PVs in *BRCA2* and *PALB2*. Interestingly, another Scandinavian study reported a significantly higher prevalence of childhood cancer in families with PVs in *BRCA2*. [40] Similarly, Wilson *et al.* showed that *BRCA2* was one of the most frequently mutated genes among childhood cancer survivors.[11] Therefore, *BRCA2* variants may be important in childhood cancer etiology,[40] potentially influencing treatment options if deficiencies in homologous repair promote some tumors. Even though a higher frequency of PVs in *BRCA2* than *BRCA1* have been found in the general population[41], this does not explain why multiple studies have identified so few PVs in *BRCA1*, which, in the Danish population, are more frequently associated with breast cancer than PVs in *BRCA2*. [42,43] Overall, WGS data from ethnically comparable children are lacking making comparisons of genetic findings in patients difficult. As most children survive cancer; therefore, identifying adult-onset CPSs is important for future surveillance and counselling.

Table 8. Patients with an underlying cancer predisposition syndrome according to Jongmans' and MIPOGG criteria.

Diagnosis (patient ID)	CPS (gene)	Jongmans	MIPOGG
Precursor B-cell acute lymphoblastic leukemia (A1)	Li-Fraumeni syndrome (<i>TP53</i>)	–	–
Osteosarcoma (A2)	Li-Fraumeni syndrome (<i>TP53</i>)	–	–
Acute promyelocyte leukemia (B1)	Familial paraganglioma and pheochromocytoma syndrome (<i>SDHC</i>)	+	+
Acute myeloid leukemia (C1)	Hyperparathyroidism-Jaw tumor syndrome (<i>CDC73</i>)	–	+
Acute myeloid leukemia (E1)	Neurofibromatosis type 1 (<i>NF1</i>)	+	+
Malignant peripheral nerve sheath tumor ¹ (E2)	Neurofibromatosis type 1 (<i>NF1</i>)	+	+
Optic nerve glioma (E3)	Neurofibromatosis type 1 (<i>NF1</i>)	+	+
Pilocytic astrocytoma (E4)	Neurofibromatosis type 1 (<i>NF1</i>)	+	+
Osteosarcoma (F1)	Schwannomatosis/Noonan syndrome (<i>LZTR1</i>)	+	+
Plasmacytoid dendritic cell leukemia (G1)	Novel putative childhood leukemia cancer predisposition syndrome (biallelic <i>DDX41</i>)	+	+
Retinoblastoma (H1)	Familial retinoblastoma syndrome (<i>RB1</i>)	+	+
Retinoblastoma (H2)	Familial retinoblastoma syndrome (<i>RB1</i>)	+	+
Retinoblastoma (H3)	Familial retinoblastoma syndrome (<i>RB1</i>)	+	+
Retinoblastoma (H4)	Familial retinoblastoma syndrome (<i>RB1</i>)	+	+
Wilms tumor (I1)	Beckwith-Wiedemann syndrome (pUPD chr 11)	+	+
Subependymal giant cell astrocytoma (J1)	Tuberous sclerosis complex (<i>TSC2</i>)	+	+
Small cell carcinoma of the ovary (K1)	Rhabdoid tumor predisposition syndrome (<i>SMARCA4</i>)	–	–
Acute myeloid leukemia (L1)	Down syndrome (46,XY+21)	+	+
Acute myeloid leukemia (L2)	Down syndrome (46,XY+21)	+	+
Hodgkin lymphoma (L3)	Down syndrome (46,XX+21)	+	+
Precursor B-cell acute lymphoblastic leukemia (L4)	Down syndrome (46,XX+21)	+	+
Total number of patients with childhood cancer predisposition syndrome		17/21 = 80.9%	18/21 = 85.7%
Precursor B-cell acute lymphoblastic leukemia (M1)	MUTYH-associated polyposis (<i>MUTYH</i>)	–	–
T-cell acute lymphoblastic leukemia (N1)	Familial breast and ovarian cancer (<i>BRCA2</i>)	–	–
T-cell acute lymphoblastic leukemia (O1)	Familial breast cancer (<i>CHEK2</i>)	–	–
Langerhans cell histiocytosis (N2)	Familial breast and ovarian cancer (<i>BRCA2</i>), oligodontia-colorectal cancer syndrome (<i>AXIN2</i>)	–	–
Malignant peripheral nerve sheath tumor ¹ (P1)	Familial breast and ovarian cancer (<i>PALB2</i>)	+	+
Neuroblastoma (Q1)	Familial adenomatous polyposis (<i>APC</i>)	+	–
Wilms tumor (Q2)	Familial adenomatous polyposis (<i>APC</i>)	+	+
Rhabdomyosarcoma (R1)	Ataxia telangiectasia (<i>ATM</i>)	–	–
Wilms tumor (S1)	Familial breast and ovarian cancer and familial neuroblastoma (<i>BARD1</i>)	–	–
Total number of patients with adult-onset cancer predisposition syndrome		2/9 = 22.2%	1/9 = 11.1%

+ fulfills the criteria, – does not fulfill the criteria.

¹same patient carrying these variants, patient not counted in the adult-onset cancer predisposition syndrome.

Red: childhood cancer predispositions syndrome, Blue: adult-onset cancer predisposition syndrome.

<https://doi.org/10.1371/journal.pgen.1009231.t008>

Table 9. Phenotypes identified using Jongmans'/MIPOGG criteria.

Cancer diagnosis	Non-cancer diagnosis	Phenotypic finding	Genetic findings
Diffuse intrinsic pontine glioma	Autism spectrum disorder	developmental delay, speech delay, learning difficulties	Klinefelter syndrome
Neuroblastoma	Autism spectrum disorder	developmental delay, learning difficulties	–
Acute myeloid leukemia	Autism spectrum disorder	developmental delay (does not speak until age four), learning difficulties, strabismus	–
Glioblastoma	Autism spectrum disorder	developmental delay (speech delay), learning difficulties	–
Acute myeloid leukemia: monozygotic twin of patient below	Down syndrome	intellectual disability, epicanthus, strabismus, developmental delay, single palmar crease	47,XY,+21
Acute myeloid leukemia: monozygotic twin of patient above	Down syndrome	intellectual disability, epicanthus, strabismus, developmental delay, single palmar crease	47,XY,+21
Hodgkin lymphoma	Down syndrome	intellectual disability, epicanthus, strabismus, developmental delay, single palmar crease	47,XX,+21
Precursor B-cell acute lymphoblastic leukemia	Down syndrome	intellectual disability, epicanthus, strabismus, flat-footed, hearing deficit	47,XX,+21 <i>POLH</i> , c.1600_1610delCA, p.(Gln534Gluufs*11)
Acute myeloid leukemia	Neurofibromatosis type 1	multiple café-au-lait spots	<i>NFI</i> , c.288+1delG, p.(?)
Neurofibroma	Neurofibromatosis type 1	multiple café-au-lait spots	<i>NFI</i> , c.5242C>T, p.(Arg1748*) <i>XPC</i> , c.1934delC, p.(Pro645Leufs*5)
Malignant peripheral nerve sheath tumor	Neurofibromatosis type 1	multiple café-au-lait spots hearing deficits left ear	<i>NFI</i> , c.61-22171_4110+2458del (deletion of exon 2–30), p.(Leu21_Gln1370del) <i>PALB2</i> , c.2736G>A, p.(Trp912*)
Pilocytic astrocytoma	Neurofibromatosis type 1	multiple café-au-lait spots near-sightedness	<i>NFI</i> , c.2033dupC, p.(Ile679Aspfs21)
Subependymal giant cell astrocytoma	Tuberous sclerosis	intellectual disability, hypomelanotic macules, seizures	<i>TSC2</i> , c.4141dupC, p.(Leu1382Profs*32)
T-cell acute lymphoblastic leukemia	Goldenhar syndrome, craniofacial microsomia	mild phenotype with skintags in front of both ears, lack of iris coloring in part of the right eye	<i>EFL1</i> , c.2430_2431delCC, p.(Leu811Asnfs*10) (no genes are known to cause Goldenhar syndrome)
Hodgkin lymphoma	Behcet's disease	frequent mucocutaneous ulcerations, hyperpigmentation of the lower back	– (tissue type: HLA-B7)
Precursor B-cell acute lymphoblastic leukemia	Suspicion of Charcot-Marie-Tooth, but no definite molecular genetic diagnosis	strabismus, delayed motor development, hypotonia, hyperpigmentation of the head and legs, severe vincristine toxicity. Mother has a form of skeletal dysplasia with short fingers and arms, extensions defect in the elbow joint, flat feet	<i>SH3TC2</i> , c.279G>A (AR) <i>KIF1B</i> , c.3401CT, p.(Pro1134Leu) (VUS, parental testing planned)
Wilms tumor	Beckwith-Wiedemann syndrome	epicanthus, down slanting palpebrae, hypertelorism, macroglossia, overgrowth	Paternal uniparental disomy of chromosome 11, corresponding to Beckwith-Wiedemann syndrome phenotype
Plasmacytoid dendritic cell leukemia	Novel childhood predisposition syndrome associated with leukemia and intellectual disability	macroglossia, poor mouth motor skills, small milk teeth, deformed fingers and toes, hypotonia	<i>DDX41</i> , c.962C>T, p.(Pro321Leu) <i>DDX41</i> , c.937G>C, p.(Gly313Arg) <i>NBN</i> , c.156_157delTT, p.(Ser53Cysfs*8)
T-cell acute lymphoblastic leukemia	No genetic diagnosis	deeply set eyes, hypertelorism, large café-au-lait spots, right leg	–
Hodgkin lymphoma	No genetic diagnosis	severe speech delay–speech not understandable age 5 years, macrocephaly	–

– No relevant genetic variants.

<https://doi.org/10.1371/journal.pgen.1009231.t009>

Many of our patients carried PVs in genes involved in DNA repair. Children have high cell-division rates, and deficiencies in DNA repair may result in accumulation of DNA damage and ultimately cancer. Fanconi anemia (FA) is associated with many large genes, and the

Table 10. Family pedigree findings for 1st–3rd degree relatives.

	All patients, <i>n</i> (%)	Patients with pathogenic/likely pathogenic variant in cancer predisposition gene, <i>n</i> (%) ¹
Malignancies in family members <18 years	7 (3.5%)	2 (6.9%)
Relatives with cancer aged 18–45 years	55 (27.8%)	9 (31.0%)
Two or more 1 st - or 2 nd -degree relatives in the same parental lineage with cancer <45 years	4 (2.0%)	1 (3.4%)
Any cancer history in the family	189 (95.5%)	29 (100.0%)
More than one family member with cancer	174 (87.9%)	28 (96.6%)
Any family member with cancer of the same organ as the patient	33 (16.7%)	7 (24.1%)
Any family member with a hematologic malignancy	43 (21.7%)	9 (31.0%)
- To a child with a hematologic tumor	25 (23.8%)	
- To a child with a CNS tumor	8 (18.2%)	
- To a child with a solid tumor	10 (20.4%)	
Any family member with a CNS tumor	25 (12.6%)	2 (6.9%)
- a child with a hematologic tumor	9 (8.6%)	
- To a child with a CNS tumor	11 (25.0%)	
- To a child with a solid tumor	5 (10.2%)	
Any family member with a solid tumor (defined as any kidney tumor, retinoblastoma, bone tumor, neuroendocrine tumors, gastrointestinal stromal tumor, or rhabdomyosarcoma)	27 (13.6%)	7 (24.1%)
- To a child with a hematologic tumor	14 (13.3%)	
- To a child with a CNS tumor	6 (13.6%)	
- To a child with a solid tumor	7 (14.3%)	
Chi2 test for differences between the groups	<i>p</i> = 0.04	
Any family member with breast cancer	76 (38.4%)	11 (37.9%)
Two or more family members with breast cancer	29 (14.6%)	2 (6.9%)
Any family member with any gastrointestinal cancer	74 (37.4%)	12 (41.4%)
Two or more family members with any gastrointestinal cancer	20 (10.1%)	4 (13.8%)

¹Percentages are fractions of the 29 patients with a cancer predisposition syndrome.

<https://doi.org/10.1371/journal.pgen.1009231.t010>

frequency of PVs in FA genes was 4.3% in an adult population of 7,578 patients from the Exome Sequencing Project and the 1000 Genomes Project.[44] This is consistent with our results, which showed that 13 (6.6%) patients carried a PV in a FA gene. Pathogenic FA variants are associated with a small increase in lifetime adult-onset cancer risk,[45–47] and this may also be true for childhood-onset CPS.

Table 11. Pathogenic/likely pathogenic variants in genes deemed ‘actionable’ by the American College of Medical Genetics.

Phenotype	Diagnosis	Gene/chromosomal alteration	Variant	Protein	Events corresponding to carrier status
Familial hypercholesterolemia	Burkitt lymphoma	<i>APOB</i>	c.1013delC	p. (Gln3378Hisfs*4)	–
Familial hypercholesterolemia	Glioma	<i>LDLR</i>	c.409G>A	p.(Gly137Ser)	–
Arrhythmogenic right ventricular cardiomyopathy	Diffuse intrinsic pontine glioma	<i>PKP2</i>	c.1643delG	p. (Gly548Valfs*15)	–
Arrhythmogenic right ventricular cardiomyopathy	Precursor B-cell acute lymphoblastic leukemia	<i>DSC2</i>	c.2508 +5G>A	p.(?)	–
Arrhythmogenic right ventricular cardiomyopathy	Glioma	<i>DSG2</i>	c.918G>A	p.(Trp306*)	–
Romano–Ward long QT syndrome	Anaplastic large-cell lymphoma	<i>KCNQ1</i>	c.905C>T	p.(Ala302Val)	Cardiac arrest during treatment, attributed to large tumor in thorax and subsequent mechanical obstruction

– No known clinical events

<https://doi.org/10.1371/journal.pgen.1009231.t011>

Interestingly, we observed one *CDC73* VUS and one PV in two patients with hematologic malignancies. PVs in *CDC73* cause ‘hyperparathyroidism-jaw tumor syndrome’ and parathyroid carcinoma,[48] and have been linked with hematologic cancer in mouse models.[49] RNA sequencing of leukemic cells from these patients showed no LOH, making a causal association less likely but not impossible.[50] Furthermore, we identified two patients with heterozygous deleterious variants in *ERCC6L2*, a gene linked to a bone marrow failure syndrome.[51] These patients were diagnosed with T-lineage acute lymphoblastic leukemia (ALL) and rhabdomyosarcoma, respectively. Tumor tissue was not available for further investigation.

Seven patients had more than one PV in CPS genes, suggesting that di-, oligo-, and polygenic inheritance can cause predisposition to childhood cancer. Four of these patients exhibited the phenotype associated with the childhood-onset PV. Bioinformatic predictions suggested that pathogenicity was highly likely in one of the seven patients. Other studies have also found more than one PV in the same patient, suggesting that digenic/polygenic inheritance may play a role in childhood cancer etiology,[8,52,53] as *MUTYH* and *OGG1* do in colorectal cancer etiology.[54] Kuhlen *et al.*[55] proposed a model of concomitant digenic inheritance involving two PVs within the same pathway combining to increase the likelihood of disease development. Generally, the observations from this and other studies suggest that the risk of disease development may increase by having more than one PV, even if the corresponding genes function in different pathways.

PVs in genes not previously associated with cancer development were identified. Some of these genes were ‘actionable’ and their identities were disclosed to the children’s families, in accordance with ACMG recommendations. It is important to identify genes associated with increased risk of cardiac disease in pediatric cancer patients due to the increased risk of both cardiomyopathy[56] and symptoms in patients with long QT syndrome[57] undergoing anti-cancer treatment. PVs in these genes may also have clinical implications for family members.

We applied Jongmans’ and MIPOGG criteria to assess the risk of underlying childhood CPSs. The majority of patients (85.7%) with a childhood-onset CPS were identified using these criteria. However, among the three unidentified patients were the two with Li-Fraumeni syndrome (one with ALL and one with osteosarcoma). Li-Fraumeni syndrome is associated with a high lifetime-risk of cancer, and the risk of a secondary cancer is further increased when the first cancer occurs during childhood.[58,59] Data suggest that surveillance programs for Li-Fraumeni patients increase their survival rates.[60] However, in contrast to other studies, patients with Li-Fraumeni syndrome were not identified here.[8,16] A possible explanation is that one of our patients carried a *de novo* *TP53* variant that could not be identified from a family history of cancer. Additionally, CPSs that are not associated with syndromic features may not fulfill relevant criteria and will be difficult to identify if the cancer is not pathognomonic of the CPS. Family history was only rarely the cause of fulfillment of Jongmans’/MIPOGG criteria in both patients with childhood- and adult-onset CPS. The primary causes of fulfillment of Jongmans’/MIPOGG criteria were the patient’s diagnosis and clinical characteristics. This is interesting as family history is believed to be highly indicative of adult-onset CPS like hereditary breast- and ovarian cancer and Lynch syndrome. A possible explanation for this could be that more variants are *de novo* in pediatric cancer patients and that the age of pediatric cancer patient’s parents is lower than parents of adult cancer patients.

We found that a family history of CNS tumors was associated with the case of childhood CNS tumors. However, only two of 44 patients with a CNS tumor carried a germline variant (*TSC2*, *NF1*), and none of these two patients had a 1st-3rd-degree family member with a CNS tumor. Therefore, there may be unidentified predisposition genes among CNS tumor patients. However, recall bias cannot be excluded, because CNS tumors among family members might be more memorable, especially if a child is diagnosed with a CNS tumor.

One limitation of this study is the lack of a comparison cohort, because population-based WGS data from ethnically comparable children are not publicly available. Another problem is whether PVs in cancer genes of children with cancers that are unassociated with that particular gene have occurred randomly. Other studies have found similar cases, in which the genotype and phenotype were not previously reported,[4,8] and it remains uncertain whether PVs in adult CPSs are driver or passenger mutations.[5,61] Thus, a large international collection of cases should be investigated to describe the phenotypic spectrum associated with each CPS variant.

Strengths of our study include the national setup with a consecutive cohort of unselected pediatric cancer patients, in-depth clinical examinations of children with cancer, and use of national databases to verify cancer diagnoses in family members. Additionally, we performed WGS instead of WES or gene panel analyses so that large structural rearrangements and CNVs could be identified, if present. Moreover, WGS will facilitate future analyses of deep intronic variants that affect splicing, variants within putative regulatory areas, and novel CPS genes.

These results demonstrate the value of systematically screening pediatric cancer patients for CPSs and strongly indicate that a higher proportion of childhood cancers may be linked to predisposing germline variants than previously supposed.

Supporting information

S1 Data. Gene panel and List of Variants of Unknown Significance (VUS). Gene panel outlining the 314 genes examined in each patient included in this study. List outlining all the VUS (CADD-PHRED score >20 and allele frequency <1%) found in the patients included in this study.

(XLSX)

S1 Text. Clinical checklist used in the clinical examination of all patient.

(DOCX)

Acknowledgments

We should like to acknowledge the research nurses in all Pediatric Oncology Departments in Denmark, they were instrumental in making this study possible.

Author Contributions

Conceptualization: Kjeld Schmiegelow, Karin Wadt.

Data curation: Anna Byrjalsen, Thomas V. O. Hansen, Anne-Marie Gerdes, Kjeld Schmiegelow, Karin Wadt.

Formal analysis: Anna Byrjalsen, Thomas V. O. Hansen, Mana M. Mehrjouy, Karin Wadt.

Funding acquisition: Anna Byrjalsen, Ulrik K. Stoltze, Rasmus L. Marvig, Anne-Marie Gerdes, Kjeld Schmiegelow, Karin Wadt.

Investigation: Anna Byrjalsen, Lisa L. Hjalgrim, René Mathiasen, Charlotte K. Lautrup, Pernille A. Gregersen, Henrik Hasle, Peder S. Wehner, Ruta Tuckuviene, Maria Rossing, Rasmus L. Marvig, Niels Tommerup, Tina Elisabeth Olsen, David Scheie, Kjeld Schmiegelow, Karin Wadt.

Project administration: Kjeld Schmiegelow, Karin Wadt.

Software: Ulrik K. Stoltze, Nanna Moeller Barnkob, Peter Wad Sackett, Adrian O. Laspiur, Ramneek Gupta, Karin Wadt.

Supervision: Anne-Marie Gerdes, Kjeld Schmiegelow, Karin Wadt.

Writing – original draft: Anna Byrjalsen.

Writing – review & editing: Anna Byrjalsen.

References

1. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. *Int J Cancer*. 2010; 127:2893–2917. <https://doi.org/10.1002/ijc.25516> PMID: 21351269
2. Volk J, Heck JE, Schmiegelow K, Hansen J. Parental occupational organic dust exposure and selected childhood cancers in Denmark 1968–2016. *Cancer Epidemiol*. 2020; 65:101667. <https://doi.org/10.1016/j.canep.2020.101667> PMID: 31955038
3. Kirkeleit J, Riise T, Bjørge T, Christiani DC, Bråtveit M, Baccarelli A, et al. Maternal exposure to gasoline and exhaust increases the risk of childhood leukaemia in offspring—a prospective study in the Norwegian Mother and Child Cohort Study. *Br J Cancer*. 2018; 119:1028–1035. <https://doi.org/10.1038/s41416-018-0295-3> PMID: 30318517
4. Parsons DW, Roy A, Yang Y, Wang T, Scollon S, Bergstrom K, et al. Diagnostic yield of clinical tumor and germline whole-exome sequencing for children with solid tumors. *JAMA Oncol*. 2016; 2:616–624. <https://doi.org/10.1001/jamaoncol.2015.5699> PMID: 26822237
5. Waszak SM, Northcott PA, Buchhalter I, Robinson GW, Sutter C, Groebner S, et al. Spectrum and prevalence of genetic predisposition in medulloblastoma: a retrospective genetic study and prospective validation in a clinical trial cohort. 2019; 19:785–798. [https://doi.org/10.1016/S1470-2045\(18\)30242-0](https://doi.org/10.1016/S1470-2045(18)30242-0) PMID: 29753700
6. Mirabello L, Zhu B, Koster R, Karlins E, Dean M, Yeager M, et al. Frequency of Pathogenic Germline Variants in Cancer-Susceptibility Genes in Patients With Osteosarcoma. *JAMA Oncol*. 2020. <https://doi.org/10.1001/jamaoncol.2020.0197> PMID: 32191290
7. Zhang J, Walsh MF, Wu G, Edmonson MN, Gruber TA, Easton J, et al. Germline Mutations in Predisposition Genes in Pediatric Cancer. *N Engl J Med*. 2015; 373:2336–2346. <https://doi.org/10.1056/NEJMoa1508054> PMID: 26580448
8. Chan SH, Chew W, Ishak NDB, Lim WK, Li ST, Tan SH, et al. Clinical relevance of screening checklists for detecting cancer predisposition syndromes in Asian childhood tumours. *npj Genomic Med*. 2018; 3:30. <https://doi.org/10.1038/s41525-018-0070-7> PMID: 30455982
9. Mody RJ, Wu YM, Lonigro RJ, Cao X, Roychowdhury S, Vats P, et al. Integrative clinical sequencing in the management of refractory or relapsed cancer in youth. *JAMA—J Am Med Assoc*. 2015; 314:913–925. <https://doi.org/10.1001/jama.2015.10080> PMID: 26325560
10. Gröbner SN, Worst BC, Weischenfeldt J, Buchhalter I, Kleinheinz K, Rudneva VA, et al. The landscape of genomic alterations across childhood cancers. *Nature*. 2018; 555:321–327. <https://doi.org/10.1038/nature25480> PMID: 29489754
11. Wilson CL, Wang Z, Liu Q, Ehrhardt MJ, Mostafavi R, Easton J, et al. Estimated number of adult survivors of childhood cancer in United States with cancer-predisposing germline variants. *Pediatr Blood Cancer*. 2020; 67:e28047. <https://doi.org/10.1002/pbc.28047> PMID: 31736278
12. Gambale A, Russo R, Andolfo I, Quaglietta L, De Rosa G, Contestabile V, et al. Germline mutations and new copy number variants among 40 pediatric cancer patients suspected for genetic predisposition. *Clin Genet*. 2019; 96:359–365. <https://doi.org/10.1111/cge.13600> PMID: 31278746
13. Jongmans MCJ, Loeffen JLCM, Waanders E, Hoogerbrugge PM, Ligtenberg MJL, Kuiper RP, et al. Recognition of genetic predisposition in pediatric cancer patients: An easy-to-use selection tool. *Eur J Med Genet*. 2016; 59:116–125. <https://doi.org/10.1016/j.ejmg.2016.01.008> PMID: 26825391
14. Ripperger T, Bielack SS, Borkhardt A, Brecht IB, Burkhardt B, Calaminus G, et al. Childhood cancer predisposition syndromes—A concise review and recommendations by the Cancer Predisposition Working Group of the Society for Pediatric Oncology and Hematology. *Am J Med Genet Part A*. 2017; 173:1017–1037. <https://doi.org/10.1002/ajmg.a.38142> PMID: 28168833
15. Postema FA, Hopman SM, De Borgie CAJM, Hammond P, Hennekam RC, Merks JH, et al. Validation of a clinical screening instrument for tumour predisposition syndromes in patients with childhood cancer (TuPS): Protocol for a prospective, observational, multicentre study. *BMJ Open*. 2017; 7:e013237. <https://doi.org/10.1136/bmjopen-2016-013237> PMID: 28110285

16. Brozou T, Tæubner J, Velleuer E, Dugas M, Wieczorek D, Borkhardt A, et al. Genetic predisposition in children with cancer—affected families' acceptance of Trio-WES. *Eur J Pediatr*. 2018; 177:53–60. <https://doi.org/10.1007/s00431-017-2997-6> PMID: 28929227
17. Goudie C, Coltin H, Witkowski L, Mourad S, Malkin D, Foulkes WD. The McGill Interactive Pediatric OncoGenetic Guidelines: An approach to identifying pediatric oncology patients most likely to benefit from a genetic evaluation. *Pediatr Blood Cancer*. 2017; 64. <https://doi.org/10.1002/pbc.26441> PMID: 28097779
18. Goudie C, Cullinan N, Villani A, Mathews N, van Engelen K, Malkin D, et al. Retrospective evaluation of a decision-support algorithm (MIPOGG) for genetic referrals for children with neuroblastic tumors. *Pediatr Blood Cancer*. 2018; 65:e27390. <https://doi.org/10.1002/pbc.27390> PMID: 30117275
19. Kalia SS, Adelman K, Bale SJ, Chung WK, Eng C, Evans JP, et al. Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): A policy statement of the American College of Medical Genetics and Genomics. *Genet Med*. 2017; 19:249–255. <https://doi.org/10.1038/gim.2016.190> PMID: 27854360
20. Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*. 2009; 25:1754–1760. <https://doi.org/10.1093/bioinformatics/btp324> PMID: 19451168
21. Tischler G, Leonard S. Biobambam: Tools for read pair collation based algorithms on BAM files. Source Code for Biology and Medicine. BioMed Central Ltd.; 2014. p. 13. <https://doi.org/10.1186/1751-0473-9-13>
22. Van der Auwera GA, Carneiro MO, Hartl C, Poplin R, del Angel G, Levy-Moonshine A, et al. From fastQ data to high-confidence variant calls: The genome analysis toolkit best practices pipeline. *Curr Protoc Bioinforma*. 2013. <https://doi.org/10.1002/0471250953.bi1110s43> PMID: 25431634
23. Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, et al. Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*. 2016; 32:1220–1222. <https://doi.org/10.1093/bioinformatics/btv710> PMID: 26647377
24. Talevich E, Shain AH, Botton T, Bastian BC. CNVkit: Genome-Wide Copy Number Detection and Visualization from Targeted DNA Sequencing. *PLoS Comput Biol*. 2016; 12:e1004873. <https://doi.org/10.1371/journal.pcbi.1004873> PMID: 27100738
25. den Dunnen JT, Dalgleish R, Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, et al. HGVS Recommendations for the Description of Sequence Variants: 2016 Update. *Hum Mutat*. 2016; 37:564–569. <https://doi.org/10.1002/humu.22981> PMID: 26931183
26. Rahman N. Realizing the promise of cancer predisposition genes. *Nature*. Nature Publishing Group; 2014. pp. 302–308. <https://doi.org/10.1038/nature12981> PMID: 24429628
27. 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; 46:310–315. <https://doi.org/10.1038/ng.2892> PMID: 24487276
28. Richards S, Aziz N, Bale S, Bick D, Das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. 2015; 17:405–424. <https://doi.org/10.1038/gim.2015.30> PMID: 25741868
29. Renaux A, Papadimitriou S, Versbraegen N, Nachtegaal C, Boutry S, Nowé A, et al. ORVAL: a novel platform for the prediction and exploration of disease-causing oligogenic variant combinations. *Nucleic Acids Res*. 2019; 47:W93–W98. <https://doi.org/10.1093/nar/gkz437> PMID: 31147699
30. Bache I, Wadt K, Mehrjouy MM, Rossing M, Østrup O, Byrjalsen A, et al. A shared somatic translocation involving CUX1 in monozygotic twins as an early driver of AMKL in Down syndrome. *Blood Cancer Journal*. Springer Nature; 2020. p. 27. <https://doi.org/10.1038/s41408-020-0293-6> PMID: 32127516
31. Mavaddat N, Peock S, Frost D, Ellis S, Platte R, Fineberg E, et al. Cancer risks for BRCA1 and BRCA2 mutation carriers: Results from prospective analysis of EMBRACE. *J Natl Cancer Inst*. 2013; 105:812–822. <https://doi.org/10.1093/jnci/djt095> PMID: 23628597
32. Yang X, Leslie G, Doroszuk A, Schneider S, Allen J, Decker B, et al. Cancer risks associated with germline PALB2 pathogenic variants: An international study of 524 families. *J Clin Oncol*. 2020; 38:674–685. <https://doi.org/10.1200/JCO.19.01907> PMID: 31841383
33. Mazzoni SM, Fearon ER. AXIN1 and AXIN2 variants in gastrointestinal cancers. *Cancer Letters*. Elsevier Ireland Ltd; 2014. pp. 1–8. <https://doi.org/10.1016/j.canlet.2014.09.018> PMID: 25236910
34. Leshno A, Shapira S, Liberman E, Kraus S, Srour M, Harlap-Gat A, et al. The APC I1307K allele conveys a significant increased risk for cancer. *Int J Cancer*. 2016; 138:1361–1367. <https://doi.org/10.1002/ijc.29876> PMID: 26421687

35. Goldgar DE, Healey S, Dowty JG, Da Silva L, Chen X, Spurdle AB, et al. Rare variants in the ATM gene and risk of breast cancer. *Breast Cancer Res.* 2011; 13. <https://doi.org/10.1186/bcr2919> PMID: 21787400
36. Couch FJ, Shimelis H, Hu C, Hart SN, Polley EC, Na J, et al. Associations between cancer predisposition testing panel genes and breast cancer. *JAMA Oncol.* 2017; 3:1190–1196. <https://doi.org/10.1001/jamaoncol.2017.0424> PMID: 28418444
37. Schmidt MK, Hogervorst F, Van Hien R, Cornelissen S, Broeks A, Adank MA, et al. Age-And tumor sub-type-specific breast cancer risk estimates for CHEK2*1100delC Carriers. *J Clin Oncol.* 2016; 34:2750–2760. <https://doi.org/10.1200/JCO.2016.66.5844> PMID: 27269948
38. Jones N, Vogt S, Nielsen M, Christian D, Wark PA, Eccles D, et al. Increased Colorectal Cancer Incidence in Obligate Carriers of Heterozygous Mutations in MUTYH. *Gastroenterology.* 2009; 137. <https://doi.org/10.1053/j.gastro.2009.04.047> PMID: 19394335
39. Diness BR, Risom L, Frandsen TL, Hansen B, Andersen MK, Schmiegelow K, et al. Putative new childhood leukemia cancer predisposition syndrome caused by germline bi-allelic missense mutations in DDX41. *Genes Chromosom Cancer.* 2018; 57:670–674. <https://doi.org/10.1002/gcc.22680> PMID: 30144193
40. Magnusson S, Borg Å, Kristoffersson U, Nilbert M, Wiebe T, Olsson H. Higher occurrence of childhood cancer in families with germline mutations in BRCA2, MMR and CDKN2A genes. *Fam Cancer.* 2008; 7:331–337. <https://doi.org/10.1007/s10689-008-9195-7> PMID: 18481196
41. Ferla R, Calò V, Cascio S, Rinaldi G, Badalamenti G, Carreca I, et al. Founder mutations in BRCA1 and BRCA2 genes. *Annals of Oncology.* 2007. p. vi93–8. <https://doi.org/10.1093/annonc/mdm234> PMID: 17591843
42. Terkelsen T, Christensen LL, Fenton DC, Jensen UB, Sunde L, Thomassen M, et al. Population frequencies of pathogenic alleles of BRCA1 and BRCA2: analysis of 173 Danish breast cancer pedigrees using the BOADICEA model. *Fam Cancer.* 2019; 18:381–388. <https://doi.org/10.1007/s10689-019-00141-9> PMID: 31435815
43. Soenderstrup IMH, Laenkholm A V, Jensen MB, Eriksen JO, Gerdes AM, Hansen TVO, et al. Clinical and molecular characterization of BRCA-associated breast cancer: results from the DBCG. *Acta Oncol (Madr).* 2018; 57:95–101. <https://doi.org/10.1080/0284186X.2017.1398415> PMID: 29164974
44. Rogers KJ, Fu W, Akey JM, Monnat RJ. Global and disease-associated genetic variation in the human Fanconi anemia gene family. *Hum Mol Genet.* 2014; 23:6815–6825. <https://doi.org/10.1093/hmg/ddu400> PMID: 25104853
45. Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, et al. Truncating mutations in the Fanconi anemia J gene BRIP1 are low-penetrance breast cancer susceptibility alleles. *Nat Genet.* 2006; 38:1239–1241. <https://doi.org/10.1038/ng1902> PMID: 17033622
46. Berwick M, Satagopan JM, Ben-Porat L, Carlson A, Mah K, Henry R, et al. Genetic heterogeneity among fanconi anemia heterozygotes and risk of cancer. *Cancer Res.* 2007; 67:9591–9596. <https://doi.org/10.1158/0008-5472.CAN-07-1501> PMID: 17909071
47. Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, et al. PALB2, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat Genet.* 2007; 39:165–167. <https://doi.org/10.1038/ng1959> PMID: 17200668
48. Hyde SM, Rich TA, Waguespack SG, Perrier ND, Hu MI. CDC73-Related Disorders. *GeneReviews*®. 1993. Available from: <http://www.ncbi.nlm.nih.gov/pubmed/20301744>. PMID: 20301744
49. Saha N, Ropa J, Chen L, Hu H, Mysliwski M, Friedman A, et al. The PAF1c Subunit CDC73 Is Required for Mouse Hematopoietic Stem Cell Maintenance but Displays Leukemia-Specific Gene Regulation. *Stem Cell Reports.* 2019; 12:1069–1083. <https://doi.org/10.1016/j.stemcr.2019.03.010> PMID: 31031188
50. Ciuffi S, Cianferotti L, Nesi G, Luzi E, Marini F, Giusti F, et al. Characterization of a novel CDC73 gene mutation in a hyperparathyroidism-jaw tumor patient affected by parathyroid carcinoma in the absence of somatic loss of heterozygosity. *Endocr J.* 2019; 66:319–327. <https://doi.org/10.1507/endocrj.EJ18-0387> PMID: 30799315
51. Bluteau O, Sebert M, Leblanc T, Peffault de Latour R, Quentin S, Lainey E, et al. A landscape of germline mutations in a cohort of inherited bone marrow failure patients. *Blood.* 2018; 131:717–732. <https://doi.org/10.1182/blood-2017-09-806489> PMID: 29146883
52. Diets IJ, Waanders E, Ligtenberg MJ, van Bladel DAG, Kamping EJ, Hoogerbrugge PM, 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; 24:1594–1603. <https://doi.org/10.1158/1078-0432.CCR-17-1725> PMID: 29351919
53. Whitworth J, Smith PS, Martin JE, West H, Luchetti A, Rodger F, et al. Comprehensive Cancer-Predisposition Gene Testing in an Adult Multiple Primary Tumor Series Shows a Broad Range of Deleterious

- Variants and Atypical Tumor Phenotypes. *Am J Hum Genet.* 2018; 103:3–18. <https://doi.org/10.1016/j.ajhg.2018.04.013> PMID: 29909963
54. Morak M, Massdorf T, Sykora H, Kerscher M, Holinski-Feder E. First evidence for digenic inheritance in hereditary colorectal cancer by mutations in the base excision repair genes. *Eur J Cancer.* 2011; 47:1046–1055. <https://doi.org/10.1016/j.ejca.2010.11.016> PMID: 21195604
 55. Kuhlen M, Borkhardt A. Trio sequencing in pediatric cancer and clinical implications. *EMBO Mol Med.* 2018; 10. <https://doi.org/10.15252/emmm.201708641> PMID: 29507082
 56. Chang VY, Wang JJ. Pharmacogenetics of Chemotherapy-Induced Cardiotoxicity. *Current Oncology Reports. Current Medicine Group LLC* 1; 2018. <https://doi.org/10.1007/s11912-018-0696-8> PMID: 29713898
 57. Cubeddu L. Drug-induced Inhibition and Trafficking Disruption of ion Channels: Pathogenesis of QT Abnormalities and Drug-induced Fatal Arrhythmias. *Curr Cardiol Rev.* 2016; 12:141–154. <https://doi.org/10.2174/1573403x12666160301120217> PMID: 26926294
 58. Chompret A. The Li-Fraumeni syndrome. *Biochimie.* 2002; 84:75–82. [https://doi.org/10.1016/s0300-9084\(01\)01361-x](https://doi.org/10.1016/s0300-9084(01)01361-x) PMID: 11900879
 59. Hisada M, Garber JE, Fung CY, Fraumeni JF, Li FP. Multiple primary cancers in families with Li-Fraumeni syndrome. *J Natl Cancer Inst.* 1998; 90:606–611. <https://doi.org/10.1093/jnci/90.8.606> PMID: 9554443
 60. Villani A, Shore A, Wasserman JD, Stephens D, Kim RH, Druker H, et al. Biochemical and imaging surveillance in germline TP53 mutation carriers with Li-Fraumeni syndrome: 11 year follow-up of a prospective observational study. *Lancet Oncol.* 2016; 17:1295–305. [https://doi.org/10.1016/S1470-2045\(16\)30249-2](https://doi.org/10.1016/S1470-2045(16)30249-2) PMID: 27501770
 61. Walsh MF, Kennedy J, Harlan M, Kentsis A, Shukla N, Musinsky J, et al. Germline BRCA2 mutations detected in pediatric sequencing studies impact parents' evaluation and care. *Cold Spring Harb Mol case Stud.* 2017; 3. <https://doi.org/10.1101/mcs.a001925> PMID: 28655807

Supplementary Material

S1 Data._Gene panel and List of Variants of Unknown Significance (VUS).

Gene panel outlining the 314 genes examined in each patient included in this study. List outlining all the VUS (CADD-PHRED score >20 and allele frequency <1%) found in the patients included in this study.

<https://doi.org/10.1371/journal.pgen.1009231.s001>

S1 Text._Clinical checklist used in the clinical examination of all patient.

<https://doi.org/10.1371/journal.pgen.1009231.s002>

Manuscript 2

Comprehensive quartet genomic analyses of fraternal siblings with childhood Philadelphia positive acute lymphoblastic leukaemia.

Authors: Adrian Otamendi Laspiur, Ulrik Kristoffer Stoltze, Frederik Steensgard Gade, Olga Rigina, Peter Wad Sacket, Rikke Linnemann Nielsen, Ramneek Gupta, Nikola Tom, Elena Papaleo, Kjeld Schmiegelow, Karin A. W. Wadt

* These authors contributed equally to this work.

Contribution: In this manuscript I have contributed to the design of the study, bioinformatic analysis of the samples, interpretation of the data and writing the manuscript. The analysis of the structural variants was carried in collaboration with Frederik Steensgard Gade and the manuscript has been developed in collaboration with Elena Papaleo, Karin Wadt and Ulrik Stoltze who provided crucial guidance.

1 **Working title: Comprehensive quartet genomic analyses of**
2 **siblings with childhood Philadelphia positive acute**
3 **lymphoblastic leukaemia.**

4
5
6 **Authors:** Adrian Otamendi Laspiur^{1*} (adrota@dtu.dk), Ulrik
7 Kristoffer Stoltze^{2,3*} (ulrik.kristoffer.stoltze@regionh.dk),
8 Frederik Steensgard Gade^{1,4} (fstga@dtu.dk), Olga Rigina⁴
9 (orig@dtu.dk), Peter Wad Sacket⁵ (pwsa@dtu.dk), Rikke
10 Linnemann Nielsen^{2,6} (rlni@dtu.dk), Ramneek Gupta^{4,6}
11 (ramg@dtu.dk), Nikola Tom¹ (niktom@dtu.dk), Elena Papaleo^{1*}
12 (elpap@dtu.dk), Kjeld Schmiegelow^{2,7}
13 (Kjeld.Schmiegelow@regionh.dk), Karin A. W. Wadt^{2,7*}
14 (karin.wadt@regionh.dk)

15 * These authors contributed equally to this work.

16
17 **Affiliations:** 1 Cancer Systems Biology, Section for
18 Bioinformatics, Department of Health and Technology,
19 Technical University of Denmark, 2800, Lyngby, Denmark;
20 2 Department of Paediatrics and Adolescent Medicine,
21 Copenhagen University Hospital Rigshospitalet, Copenhagen,
22 Denmark; 3 Department of Clinical Genetics, Copenhagen
23 University Hospital Rigshospitalet, Copenhagen, Denmark; 4
24 Disease Data Intelligence, Section for Bioinformatics,
25 Department of Health Technology, Technical University of
26 Denmark, 2800 Kongens Lyngby, Denmark. 5 Section for
27 Bioinformatics, Department of Health and Technology,

28 Technical University of Denmark, 2800, Lyngby, Denmark; 6
29 Department of Computational Biology, Novo Nordisk Research
30 Centre Oxford, Oxford OX3 7FZ, UK. 7 Department of Clinical
31 Medicine, Faculty of Health and Medical Sciences, University of
32 Copenhagen, Copenhagen, Denmark.

33

34 **Corresponding author:** Karin Wadt (karin.wadt@regionh.dk),
35 Department of Clinical Genetics, Copenhagen University
36 Hospital Rigshospitalet, Copenhagen, Denmark.

37 Elena Papaleo^{3*} (elpap@dtu.dk), Cancer Systems Biology,
38 Section for Bioinformatics, Department of Health and
39 Technology, Technical University of Denmark, 2800, Lyngby,
40 Denmark

41

42 **Competing interests:** The authors declare no competing
43 interests

44 **Letter**

45 Dear editor:

46 While acute lymphoblastic leukaemia (ALL) is the most frequent
47 malignant disease in children and adolescent, the rare subtype,
48 Philadelphia chromosome (Ph+, (t(9;22) (q34;q11)) ALL,
49 accounts for only 3-5% of paediatric ALL cases (1) (2) (SF1).

50 Generally, the incidence of ALL in Danish children under five
51 years of age is 7/100 000 (<https://nordcan.iarc.fr/>)

52 corresponding to just one or two cases of paediatric Ph+ ALL
53 per year. Considering this rarity of Ph+ ALL in childhood, clinical
54 suspicion of genetic predisposition was raised when two siblings

55 were consecutively diagnosed with Ph+ ALL at age 60 and 16
56 months, respectively. Heritability studies of paediatric

57 leukaemias are hampered by co-occurrence precipitated not by
58 two independent tumorigeneses, but rather by placental cross-
59 circulation of a single (pre)malignant leukemic clone, particularly

60 in monozygotic twins. However, studies of ALL have shown
61 significantly increased standardized incidence ratio of 3.2 in
62 siblings (3). Known genetic disposition to ALL may be divided

63 into syndromic and non-syndromic disposition. The syndromic
64 disposition consists of well-defined syndromes where
65 phenotypic traits and risk of other manifestations than ALL are

66 more prevalent, i.e., Neurofibromatosis type 1, Noonan
67 syndrome, and Bloom syndrome, while non-syndromic germline
68 disposition genes such as *PAX5*, *ETV6*, *IKZF1*, *RUNX1* and

69 *TP53*, have no non-cancer phenotype. Of note, none of these
70 well-defined monogenic ALL predispositions have a particular

71 increased risk of Ph+ ALL. While a single nucleotide
72 polymorphism in *GATA3* (rs3824662) has been associated
73 specifically with Ph-like ALL phenotype, no SNPs have been
74 associated with Ph+ ALL. Here, for the first time, we explore the
75 genomic landscape of a quartet of siblings with paediatric Ph+
76 ALL and their parents. We performed germline whole genome
77 sequencing (WGS) in siblings and parents as well as RNAseq
78 and WGS of the tumours including mutational signature
79 analysis. Structural variants (SVs), single nucleotide variants
80 (SNVs) and short indels were annotated and stored at an in-
81 house mySQL database for posterior analysis (SF3). Overview
82 of the genetic alterations found in each sister are represented in
83 the Figure 1b and 1c.

84

85 Firstly, we investigated any possibility of metachronous
86 transplacental transfer of leukemic clones during each
87 pregnancy. Leukemic clone transmission from foetus to the
88 mother and vice versa has been previously reported (4),
89 however an absence of a clinical impact on maternal health
90 indicates that infiltrating malignant foetal cells are
91 immunological cleared. Transplacental transmission between
92 non-twin siblings have to our knowledge not been reported in
93 the literature. To assess the origin of the tumoral clone present
94 in each sister we compared the somatic structural alterations in
95 the sisters' tumours obtained by GRIDSS and TNscope (SI2 and
96 SF3). The major driving translocation forming the Philadelphia
97 chromosome had disparate breakpoints in the two sisters,

98 indicating two independent somatic translocation events. In both
99 cases translocation occurred within the minor breakpoint cluster
100 (m-BCR) of chromosome 22, coding for the p190^{Bcr-Abl} subtype
101 (SF4a and SF4b, ST1). Furthermore, we observed other
102 structural alterations not shared in the sisters such as *ETV6-*
103 *RUNX1* only in sister1 (SF4a and 4c) and t(9:22:17) only in
104 sister2 (SF4b), where chromosome 17 has been reported to be
105 a frequent fusion partner for t(9:22). Interestingly for the sister2,
106 the variant allele frequency (VAF) of the t(9:17) (VAF=0,38-0,68)
107 is predicted to be significantly higher than t(9:22) (VAF=0,15-
108 0,26) suggesting the rearrangement between chromosomes 9
109 and 17 as an earlier event (ST1). Somaticly, we also
110 investigated *CDKN2A/B*, *IKZF1* and *PAX5* loci which are
111 frequently altered in ALL (1), observing deletions in both tumour
112 samples in *CDKN2A/B* but not in *IKZF1* and *PAX5*. Deletions of
113 the *CDKN2A* locus span different regions in the two sisters, a
114 471kb deletion covering *CDKN2A* entirely in sister1, and two
115 deletions of 34kb and 146kb affecting the first exon in sister2,
116 respectively (ST1). We observed gene expression of *PAX5* and
117 *BLM* similar to non-Ph+ ALL samples, however, expression of
118 *GATA3*, *IKZF1* and *CDKN2A* is lower in the Ph+ ALL samples
119 and the sisters than non-Ph+ ALL samples from our ALL cohort
120 (N=94, Ph+=5 and non-Ph+=89) (Figure2c). We identified 18
121 smaller somatic variants, such as SNV or indels, that were
122 shared between both sisters (none were exonic, ST1) and
123 another 18 non-synonymous SNVs not shared by the sisters
124 (ST1).

125

126 In order to evaluate if both sisters have been affected by similar
127 mutagenesis processes, we investigated COSMIC-v3
128 Mutational Signatures (SI3). The highest exposures in leukemic
129 cells from both sisters are found in signatures SBS1, SBS8,
130 SBS10a, SBS39, SBS43, SBS55, SBS89 as observed in the
131 Figure 2. SBS1 is found in most cancers and SBS89 appears to
132 be most active in the first decade of life in concordance with our
133 data. Interestingly, the leukemic cells from both sisters had high
134 exposure for SBS10a which is associated with polymerase
135 epsilon exonuclease domain mutations generating large
136 numbers of somatic mutations. SBS 8, SBS39, SBS43 and
137 SBS55 have unknown aetiology or are possible sequencing
138 artifacts.

139 For leukemic cells from sister1 we observe a significantly
140 increased signal for SBS3 which is associated with defective
141 homologous recombination-based DNA damage repair as well
142 as SBS26 which is associated with defective DNA mismatch
143 repair. In leukemic cells from sister2 we observed relatively high
144 exposures for SBS30 which has been associated with
145 deficiency in base excision repair due to inactivating mutations
146 in *NTHL1*.

147

148 Epidemiological studies have shown that familial cases of ALL
149 tend to be the same subtype in higher rate than would be
150 expected by chance, suggesting a genetic susceptibility to the
151 same subtype of ALL (5). To evaluate the potential genetic

152 predisposition in the family we explored germline SV, SNV and
153 indels shared by the sisters and parents. We identified 67
154 potentially deleterious shared germline SNV (ST1) between the
155 sisters and parents passing population-frequency filters (SI2).
156 No clear association with cancer was found and non-
157 synonymous rare variants will be further studied in the identified
158 genes, with no known cancer predisposition to ALL. We
159 identified 19 *de novo* SNVs (SI4), of which none were exonic
160 (ST1).

161

162 The sisters shared 59 (37% of 153 total) structural deletions
163 covering 98 genes, with 47 deletions covering 92 genes
164 showing identical start and end breakpoints. All but one deletion
165 were seen in an in-house database of WGS data (non-cancer
166 patients), in WGS data on 593 other children with cancer, or
167 both. The remaining deletion, a Variant of Uncertain
168 Significance (VUS) spanning 10,159bp in the first exon of *BLM*
169 (chr15:90 721 650-90 731 809, c.-5+4210_c.-5+14369del,
170 GRCh38) (ST1), was not detected in any other samples nor
171 reported or deciphered in gnomAD. The same deletion was
172 detected in the maternal germline WGS data (Figure 1a). No
173 pathogenic SNV alterations were otherwise identified in the *BLM*
174 gene, and neither of the sisters harboured a Bloom syndrome
175 phenotype.

176

177 The major structural somatic differences in the two samples
178 (also summarized in Figure 1b and 1c) suggest the origins of the
179 tumours were independent, and any hypothesis of

180 transplacental clonal transfer may be rejected. Germline and
181 somatic SNV analysis did not show any clear evidence of shared
182 germline Ph⁺ ALL drivers. However, structural deletion
183 including the first exon of the *BLM* gene in both sisters could
184 potentially explain a higher risk of cancer development in this
185 family. Biallelic mutation in the *BLM* gene is associated with
186 Bloom Syndrome (BS), a syndrome known to cause significantly
187 increased risk of cancer, particularly childhood and early-onset
188 acute leukemia of myeloid, lymphocytic, or mixed lineages.
189 Based on the Bloom Syndrome Registry there are only 40
190 occurrences of leukemia from which 11 cases are ALL.
191 However, information on the subtype is not provided (6). Higher
192 rate of chromosomal rearrangements in cells lacking *BLM* (7)
193 and BS patients with haematological malignancies (8)(9) has
194 been reported, explained by the role of *BLM* as 3'-5' ATP-
195 dependent RecQ DNA helicase, one of the most important
196 genome stabilizers that controls DNA replication,
197 recombination, and both homologous and non-homologous
198 processes of double-strand break repair. Furthermore, the *BCR-*
199 *ABL* fusion protein has been shown to affect *BLM*
200 expression. Heterozygous carriers of pathogenic *BLM* variants
201 retain an increased level of sister chromatid exchange (10), and
202 heterozygous germline *BLM* pathogenic variants are linked to
203 increased risk of adult cancers (11). Several studies point to
204 monoallelic *BLM* inactivation playing a role in tumorigenesis
205 (12)(13) not necessarily involving somatic inactivation of the
206 wild-type allele (14). This highlights that haploinsufficiency

207 of *BLM* does retain some of the cancer risk phenotype
208 associated with the fulminant autosomal recessive syndrome
209 such as deficiencies in the DNA repair mechanisms and higher
210 rate of chromosomal rearrangements.

211

212 This is the first report of non-twin siblings with Ph+ ALL.
213 Transplacental transfer of tumour cells between the sisters was
214 disproved and significant somatic alterations, including *BCR-*
215 *ABL1*, *ETV6-RUNX1*, t(9:22:17), and deletions in the
216 *CDKN2A/B* locus have been reported. Germline and somatic
217 SNV did not show any clear evidence of a common germline
218 genetic driver in the siblings. However, we hypothesise the
219 heterozygous VUS deletion of the first exon of the gene as a
220 potential risk factor, but as 1 out of 900 in the population are
221 heterozygous pathogenic variant *BLM* carriers (15), likely yet
222 unidentified additional genetic or environmental factors are
223 involved.

224

225 **Acknowledgements:** This work is part of Interregional
226 Childhood Oncology Precision Medicine Exploration (iCOPE), a
227 cross-Oresund collaboration between University Hospital
228 Copenhagen, Rigshospitalet, Lund University, Region Skåne
229 and Technical University Denmark (DTU), supported by the
230 European Regional Development Fund.

231

232 **Author contribution:**

233 Conception and design of the study: RG, UKS, KW, AOL, EP,
234 KS
235 Bioinformatic Analysis of the data: AOL, FSG, UKS
236 Database managers: PW, OR
237 Interpretation of the data: AOL, EP, KW, UKS, FSG, NT
238 Drafted the manuscript: AOL, UKS, KW, EP
239 Supervision: KW, EP
240 Revision and approval of final manuscript: All authors.

241

242 **Data Availability Statement:** The datasets generated during
243 and/or analysed during the current study are not publicly
244 available due to individual privacy compromise but are available
245 from the corresponding author on reasonable request.

246 **References:**

- 247 1. Bernt KM, Hunger SP. Current concepts in pediatric
248 Philadelphia chromosome-positive acute lymphoblastic
249 leukemia. *Front Oncol.* 2014;4 MAR(March):1–21.
- 250 2. Iacobucci I, Mullighan CG. Genetic basis of acute
251 lymphoblastic leukemia. *Journal of Clinical Oncology.*
252 2017.
- 253 3. Kharazmi E, da Silva Filho MI, Pukkala E, Sundquist K,
254 Thomsen H, Hemminki K. Familial risks for childhood
255 acute lymphocytic leukaemia in Sweden and Finland:
256 Far exceeding the effects of known germline variants. *Br*
257 *J Haematol.* 2012 Dec;159(5):585–8.
- 258 4. Fries C, Noronha SA, Metlay L, Zhang B. Transplacental
259 transfer of congenital B-cell acute lymphoblastic
260 leukemia to the maternal vasculature. *Pediatr Blood*
261 *Cancer.* 2021;68(9):2–3.
- 262 5. Schmiegelow K, Lausten Thomsen U, Baruchel A,
263 Pacheco CE, Pieters R, Pombo-De-Oliveira MS, et al.
264 High concordance of subtypes of childhood acute
265 lymphoblastic leukemia within families: Lessons from
266 sibships with multiple cases of leukemia. *Leukemia.*
267 2012;26(4):675–81.
- 268 6. Cunniff C, Djavaid AR, Carrubba S, Cohen B, Ellis NA,
269 Levy CF, et al. Health supervision for people with Bloom
270 syndrome. *Am J Med Genet A.* 2018;176(9):1872–81.

- 271 7. Gaymes TJ, North PS, Brady N, Hickson ID, Mufti GJ,
272 Rassool F V. Increased error-prone non homologous
273 DNA end-joining - A proposed mechanism of
274 chromosomal instability in Bloom's syndrome.
275 *Oncogene*. 2002;21(16):2525–33.
276 8. Kaneko H, Inoue R, Yamada Y, Sukegawa K, Fukao T,
277 Tashita H, et al. Microsatellite instability in B-cell
278 lymphoma originating from Bloom syndrome. *Int J*
279 *Cancer*. 1996;69(6):480–3.
280 9. Schuetz JM, MacArthur AC, Leach S, Lai AS, Gallagher
281 RP, Connors JM, et al. Genetic variation in the NBS1,
282 MRE11, RAD50 and BLM genes and susceptibility to
283 non-Hodgkin lymphoma. *BMC Med Genet*. 2009;10:1–
284 10.
285 10. Ben Salah G, Hadj Salem I, Masmoudi A, Kallabi F,
286 Turki H, Fakhfakh F, et al. A novel frameshift mutation in
287 BLM gene associated with high sister chromatid
288 exchanges (SCE) in heterozygous family members. *Mol*
289 *Biol Rep*. 2014;41(11):7373–80.
290 11. Gruber SB, Ellis NA, Rennert G, Offit K. Heterozygosity
291 and the Risk of Colorectal Cancer. 2013;297(September
292 2002).
293 12. Kunz JB, Rausch T, Bandapalli OR, Eilers J, Pechanska
294 P, Schuessle S, et al. Pediatric T-cell lymphoblastic
295 leukemia evolves into relapse by clonal selection,
296 acquisition of mutations and promoter hypomethylation.
297 *Haematologica*. 2015;100(11):1442–50.
298 13. Davari P, Hebert JL, Albertson DG, Huey B, Roy R,
299 Mancianti ML, et al. Loss of Blm enhances basal cell
300 carcinoma and rhabdomyosarcoma tumorigenesis in
301 *Ptch1*^{+/-}-mice. *Carcinogenesis*. 2009;31(6):968–73.
302 14. Suspitsin EN, Yanus GA, Sokolenko AP, Yatsuk OS,
303 Zaitseva OA, Bessonov AA, et al. Development of
304 breast tumors in CHEK2, NBN/NBS1 and BLM mutation
305 carriers does not commonly involve somatic inactivation
306 of the wild-type allele. *Med Oncol*. 2014;31(2).
307 15. De Voer RM, Hahn MM, Mensenkamp AR, Hoischen A,
308 Gilissen C, Henkes A, et al. Deleterious Germline BLM
309 mutations and the risk for early-onset colorectal cancer.
310 *Sci Rep*. 2015;5:1–7.
311

312 **Figure Legends:**

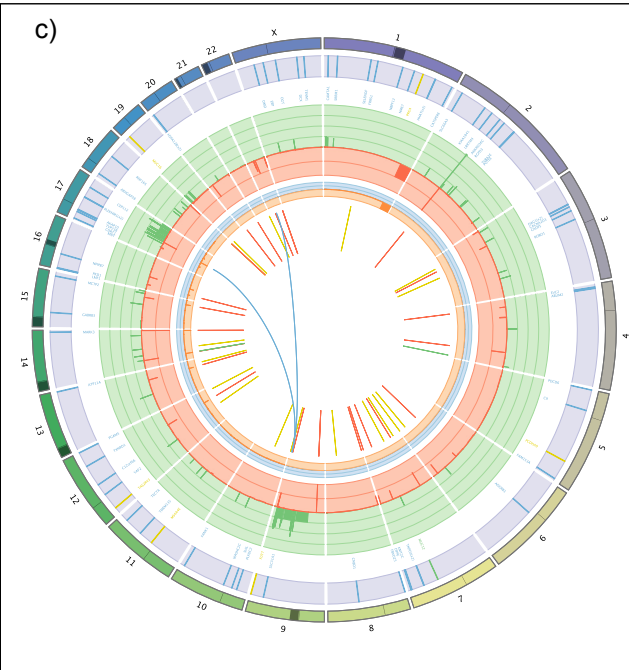
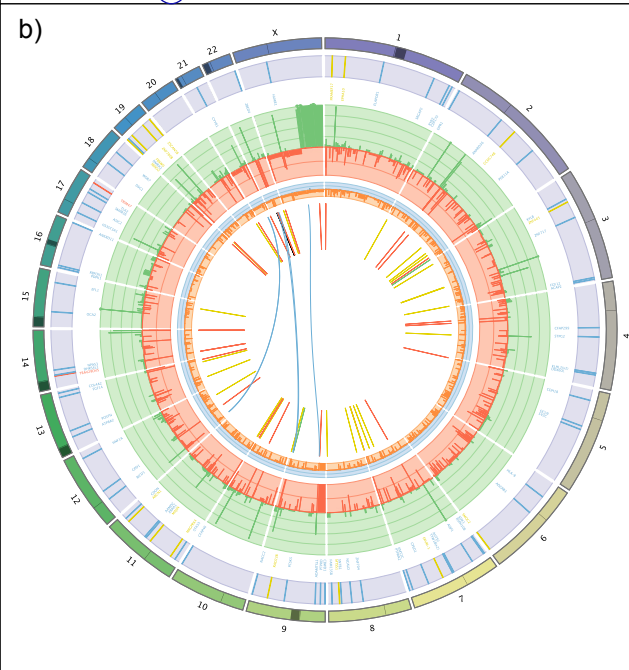
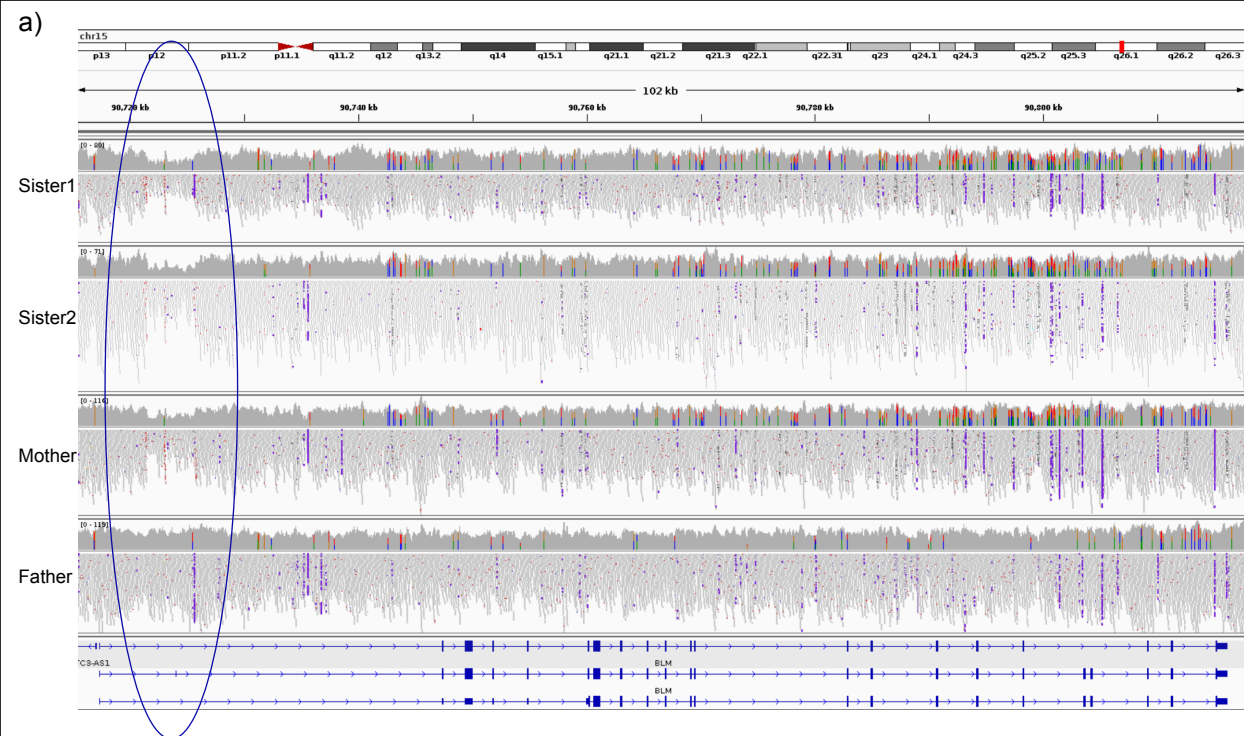
313 Figure 1a: Integrative Genome Viewer (IGV) alignment
314 representation of the germline WGS of the sisters and parents
315 in the BLM region where the deletion was observed. First 4
316 tracks represent the alignment for sister1, sister2, mother and
317 father respectively. BLM gene is represented in the bottom

318 track. Lower coverage of the deleted region can be observed in
319 both sisters and the mother within the marked area.

320 Figure 1b and 1c: Overview circos plot representing each sister
321 genomic landscape. From inner ring to outer ring: Observed SV
322 within or between chromosomes (Fusions), Minor Alleles Copy
323 Number (0-3, expected 1), Purity adjusted copy number
324 changes (0-6, expected 2), somatic SNV coloured by
325 substitution-type similar to Alexandrov et al. 2013,
326 chromosomes and non-accessible regions (grey).

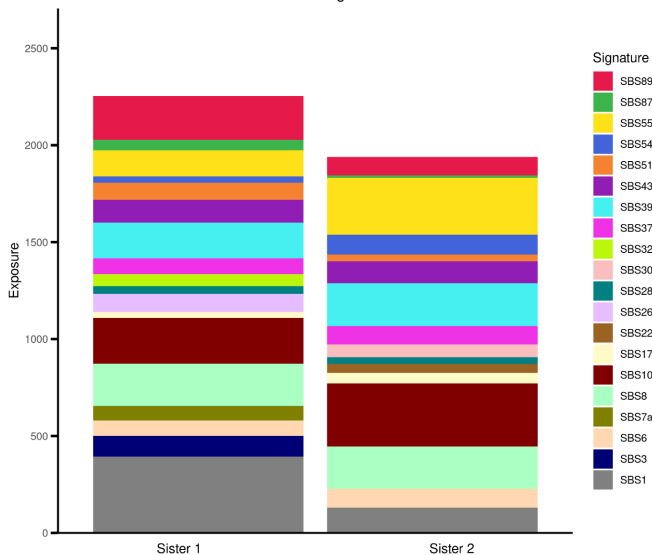
327 Figure 2a and 2b: Exposure of the COSMIC v3 Mutational
328 signatures in the sisters.

329 Figure 2c: Gene expression on Philadelphia+ ALL vs non-
330 Philadelphia+ ALL in *CDKN2A*, *PAX5*, *IKZF1*, *GATA3* and *BLM*.



a)

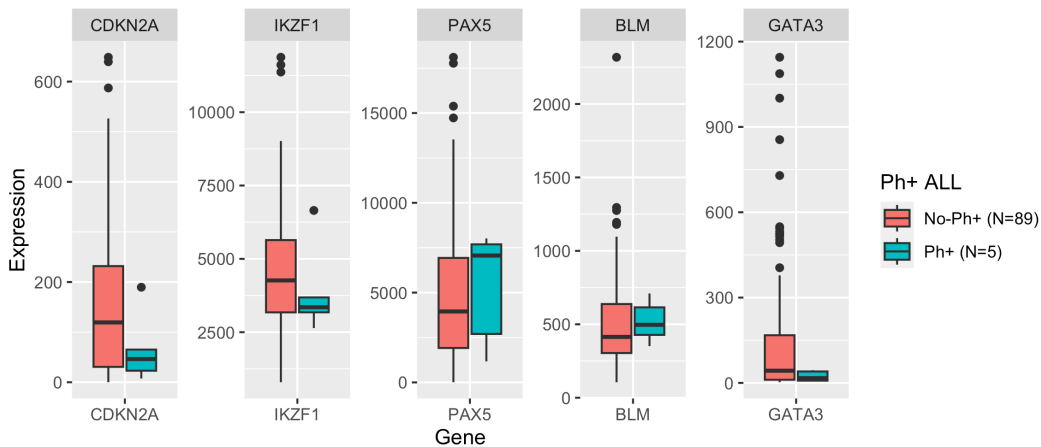
COSMIC v3 Signatures



b)

COSMIC ID	Sister 1	Sister 2
SBS1	394,437	130,832
SBS3	106,260	0,000
SBS6	80,078	96,842
SBS7a	74,389	0,000
SBS8	217,734	217,939
SBS10a	236,754	325,452
SBS17b	29,666	54,073
SBS22	0,000	46,454
SBS26	93,639	0,000
SBS28	39,829	34,296
SBS30	0,000	66,438
SBS32	62,765	0,000
SBS37	80,120	94,852
SBS39	185,903	221,827
SBS43	116,624	111,619
SBS51	88,266	35,784
SBS54	31,595	101,615
SBS55	136,678	292,681
SBS87	54,118	14,535
SBS89	225,544	94,168

c)




Manuscript 3

RosettaDDGPrediction for high-throughput mutational scans: From stability to binding.

Authors: Valentina Sora, **Adrian Otamendi Laspiur**, Kristine Degn, Matteo Arnaudi, Mattia Utichi, Ludovica Beltrame, Dayana De Menezes, Matteo Orlandi, Ulrik Kristoffer Stoltze, Olga Rigina, Peter Wad Sackett, Karin Wadt, Kjeld Schmiegelow, Matteo Tiberti, Elena Papaleo

Contribution: For this paper I contributed to the case study 4 where I analyzed 566 childhood cancer WGS samples with *DNAseq* pipeline and collected the VUS in the genes of interest to apply RosettaDDG framework and investigate the changes in folding free energy upon amino acid substitutions ($\Delta\Delta G$ s).

RosettaDDGPrediction for high-throughput mutational scans: From stability to binding

Valentina Sora^{1,2} | Adrian Otamendi Laspiur² | Kristine Degn² |
 Matteo Arnaudi^{1,2} | Mattia Utichi^{1,2} | Ludovica Beltrame^{1,2} |
 Dayana De Menezes² | Matteo Orlandi² | Ulrik Kristoffer Stoltze^{3,4,5} |
 Olga Rigina² | Peter Wad Sackett² | Karin Wadt^{3,5} | Kjeld Schmiegelow^{4,5} |
 Matteo Tiberti¹ | Elena Papaleo^{1,2} 

¹Cancer Structural Biology, Danish Cancer Society Research Center, Copenhagen, Denmark

²Cancer Systems Biology, Section for Bioinformatics, Department of Health and Technology, Technical University of Denmark, Lyngby, Denmark

³Department of Clinical Genetics, Copenhagen University Hospital Rigshospitalet, Copenhagen, Denmark

⁴Department of Pediatrics and Adolescent Medicine, University Hospital Rigshospitalet, Copenhagen, Denmark

⁵Institute of Clinical Medicine, Faculty of Medicine, University of Copenhagen, Copenhagen, Denmark

Correspondence

Elena Papaleo, Cancer Structural Biology, Danish Cancer Society Research Center, 2100 Copenhagen, Denmark.
 Email: elpap@dtu.dk; elenap@cancer.dk

Funding information

Carlsberg Foundation Distinguished Fellowship, Grant/Award Number: CF18-0314; Danmarks Grundforskningsfond, Grant/Award Number: DNR125; Hartmanns Fond, Grant/Award Number: R241-A33877; LEO Foundation, Grant/Award Number: LF17006; NovoNordisk Fonden Bioscience and Basic Biomedicine, Grant/Award Number: NNF20OC0065262; European Regional Development Fund; Danish Cancer Society, Grant/Award Number: R-257-A14720; Danish Childhood Cancer Foundation, Grant/Award Numbers: 2020-5769, 2019-5934

Review Editor: Nir Ben-Tal

Abstract

Reliable prediction of free energy changes upon amino acid substitutions ($\Delta\Delta G$ s) is crucial to investigate their impact on protein stability and protein-protein interaction. Advances in experimental mutational scans allow high-throughput studies thanks to multiplex techniques. On the other hand, genomics initiatives provide a large amount of data on disease-related variants that can benefit from analyses with structure-based methods. Therefore, the computational field should keep the same pace and provide new tools for fast and accurate high-throughput $\Delta\Delta G$ calculations. In this context, the Rosetta modeling suite implements effective approaches to predict folding/unfolding $\Delta\Delta G$ s in a protein monomer upon amino acid substitutions and calculate the changes in binding free energy in protein complexes. However, their application can be challenging to users without extensive experience with Rosetta. Furthermore, Rosetta protocols for $\Delta\Delta G$ prediction are designed considering one variant at a time, making the setup of high-throughput screenings cumbersome. For these reasons, we devised RosettaDDGPrediction, a customizable Python wrapper designed to run free energy calculations on a set of amino acid substitutions using Rosetta protocols with little intervention from the user. Moreover, RosettaDDGPrediction assists with checking completed runs and aggregates raw data for multiple variants, as well as generates publication-

Valentina Sora and Adrian Otamendi Laspiur equally contributed to this study.

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial](https://creativecommons.org/licenses/by-nc/4.0/) License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.

© 2022 The Authors. *Protein Science* published by Wiley Periodicals LLC on behalf of The Protein Society.

ready graphics. We showed the potential of the tool in four case studies, including variants of uncertain significance in childhood cancer, proteins with known experimental unfolding $\Delta\Delta G$ s values, interactions between target proteins and disordered motifs, and phosphomimetics. RosettaDDGPrediction is available, free of charge and under GNU General Public License v3.0, at <https://github.com/ELELAB/RosettaDDGPrediction>.

KEYWORDS

binding free energy, folding free energy, free energy calculations, Rosetta

1 | INTRODUCTION

Predicting the impact of amino acid substitutions in a protein or at a protein–protein interface is becoming more and more relevant as high-throughput sequencing data reveal a high rate of sequence polymorphisms of uncertain functional significance in protein-coding regions (Federici & Soddu, 2020). In this context, multiplex-based assays provide massive data that can be complemented by structural studies on the effects of protein variants (Anderson et al., 2022; Cagiada et al., 2021; Gasperini et al., 2016; Ollodart et al., 2021; Weile & Roth, 2018). Furthermore, saturation mutagenesis is experimentally accessible thanks to the advances in multiplex technologies. Therefore, molecular modeling approaches must keep the same pace and continue developing toward high-throughput applications.

A convenient and quantitative manner for assessing the impact of amino acid substitutions related to coding variants is based on estimating the changes in Gibbs free energy of folding/unfolding or binding. In this context, several computational approaches based on the analysis of protein structures are available to predict free energy changes upon mutation ($\Delta\Delta G$ s) in protein structures (Barlow et al., 2018; Delgado et al., 2019; Frenz et al., 2020; Geng et al., 2019; Kortemme & Baker, 2002; Kumari et al., 2014; Park et al., 2016; Schymkowitz et al., 2005; Seeliger & de Groot, 2010; Smith & Kortemme, 2008). These measurements can be used to classify the effect of disease-related variants on protein structural stability. As a consequence, they provide predictions of potential alterations in the protein cellular level, propensity to aggregation or proteasomal degradation (Gerasimavicius et al., 2020; Stein et al., 2019). In addition, they can also pinpoint functional effects due to local changes in the interactions with other proteins or biomolecules (Degn et al., 2022; Fas et al., 2020; Jepsen et al., 2020).

Rosetta provides a variety of protocols to estimate changes in free energy in terms of binding and folding/unfolding (Barlow et al., 2018; Frenz et al., 2020; Kellogg et al., 2011; Kortemme & Baker, 2002; Park et al., 2016). Most of these protocols estimate the change in free energy as an average over the free energy changes calculated in an ensemble of paired wild-type/mutated structures.

Three features generally characterize Rosetta protocols for the prediction of free energy changes upon mutation: (i) the sampling method employed to generate the structural ensemble, (ii) the energy function used to quantify the free energy associated with each structure, and (iii) the degree of flexibility allowed in the structure to accommodate the mutation.

Currently, three state-of-the-art strategies are available in Rosetta to estimate the change in either folding or binding free energy upon mutation. The first one, presented by Park and coworkers (Park et al., 2016) and referred to as *cartddg*, is designed to work on monomeric proteins. In this protocol, a sampling in the Cartesian space (as opposed to internal dihedrals sampling) is carried out, allowing small local backbone movements in a three-residue window around the mutation site, together with side-chains movements within a 6 Å radius from the mutation site. The second protocol, *cartddg2020*, represents an updated variant of *cartddg* (Frenz et al., 2020). The third protocol, developed by Barlow and coworkers (Barlow et al., 2018) and named here *flexddg*, deals with estimating the changes in binding free energy upon mutation in a protein complex. It applies the “backrub” sampling method (Smith & Kortemme, 2008) to recapitulate local backbone motions observed in crystal lattices. The *flexddg* protocol seems to perform better with the *talaris2014* energy function (Barlow et al., 2018). This protocol for binding free energies relies on a local sampling of backbone and side chains for residues within an 8 Å radius from the mutation, followed by global optimization of the side chains.

Rosetta is a feature-rich software suite under active development, backed by a sizable community of users, and built over roughly 20 years. Running the protocols mentioned above directly with Rosetta requires an extensive computational background and prior exposure to several Rosetta features. These requirements may discourage users with a more biology-oriented skillset, despite the benefit that accurate predictions of free energy changes upon mutations may bring to their research. Furthermore, Rosetta protocols for $\Delta\Delta G$ prediction are designed to be run considering one mutation at a time exclusively, making high-throughput screenings cumbersome to set up. We recently faced a similar challenge with implementing high-throughput scans based on the FoldX free energy function to make them parallelizable, more easily approachable, and applicable to structural ensembles. This led to the development of MutateX (Tiberti et al., 2022). FoldX, however, is known to suffer from limitations due to backbone stiffness during the sampling (Usmanova et al., 2018) and often low accuracy in predicting mutations with stabilizing effects, even though most prediction methods are biased toward destabilizing mutations (Buß et al., 2018; Usmanova et al., 2018). Rosetta-based calculations could offer a valuable complement to the $\Delta\Delta G$ estimates currently accessible with MutateX. Thus, we developed RosettaDDGPrediction, a Python wrapper to perform Rosetta-based protocols for $\Delta\Delta G$ prediction. RosettaDDGPrediction's outputs can also be converted to a format compatible with the MutateX plotting system, allowing for an expanded visualization toolkit. Here, we illustrate the applications and limits of the approach to four different cases of study, covering both methodological and biological applications. We focused on the comparison with experimentally determined unfolding $\Delta\Delta G$ values (Case Study 1). We showed an example of the application of RosettaDDGPrediction to the study of protein-protein interactions and posttranslational modifications (PTMs) (Case Study 2). We then evaluated the influence of using AlphaFold2 models as starting structures for the calculations (Case Study 3). We then used models from AlphaFold2 to assess the functional impact of mutations identified by whole genome sequencing to address cancer predisposition (Case Study 4).

2 | RESULTS

2.1 | Overview of the package

RosettaDDGPrediction is a pure Python package providing a uniform and easily accessible command-line interface to *flexddg*, *cartddg*, and *cartddg2020* protocols for

calculating free energy changes upon mutation. It is devised to help users unfamiliar with the Rosetta suite perform mutational scans and collect, aggregate, and visualize data from those scans in an intuitive fashion. In RosettaDDGPrediction, a “protocol” is intended as a set of Rosetta runs and Python-based processing steps. Each protocol takes as inputs the three-dimensional (3D) structure of the protein of interest and a list of mutations to be performed, finally returning the predicted free energy changes associated with each input mutation. The *flexddg* protocol consists of only one call to the *rosetta_scripts* executable for each mutation, which performs all the necessary calculations as defined by Barlow and coworkers (Barlow et al., 2018). On the other hand, the *cartddg* protocol first energetically relaxes the input structure by using the Rosetta *relax* program to generate an ensemble of relaxed conformations, followed by the selection of the most suitable one. Finally, it uses the *cartesian_ddg* application to relax the structure further and perform the free energy calculations. The *cartddg2020* protocol represents an updated version of the original *cartddg* protocol. Here, the relaxation is performed by a Rosetta script passed to the *rosetta_scripts* executable, and then *cartesian_ddg* is run on the lowest energy structure produced by the relaxation. It is worth noting that the relaxation procedure produces only one structure, as per the original files provided with the work first describing the *cartddg2020* protocol (Frenz et al., 2020). However, if the user decides to produce several relaxed structures, the most suitable one (according to user-selected criteria) will then be passed to *cartesian_ddg*. The standard protocols are described in specific YAML files provided with the package. With these files, expert users can still tap into the full potential of the Rosetta interface by providing virtually any Rosetta-compatible option to the executables used by each protocol.

RosettaDDGPrediction consists of four main executables (*rosetta_ddg_run*, *rosetta_ddg_check_run*, *rosetta_ddg_aggregate*, *rosetta_ddg_plot*) performing different tasks (Figure 1). Their behavior is controlled by a set of configuration files, which can be fully customized to fine-tune the parameters of each protocol, aggregation options, and plot aesthetics.

rosetta_ddg_run is the executable responsible for running a Rosetta protocol to predict free energy changes upon mutation over a set of selected mutations. Given a protein structure in PDB format and a set of mutations, it generates all the data structures and configuration files to perform several runs in parallel, making them straightforward to perform and making the most of modern many-cores computing infrastructures.

In *rosetta_ddg_run*, the user can specify the amino acid substitutions to be performed in two different ways.

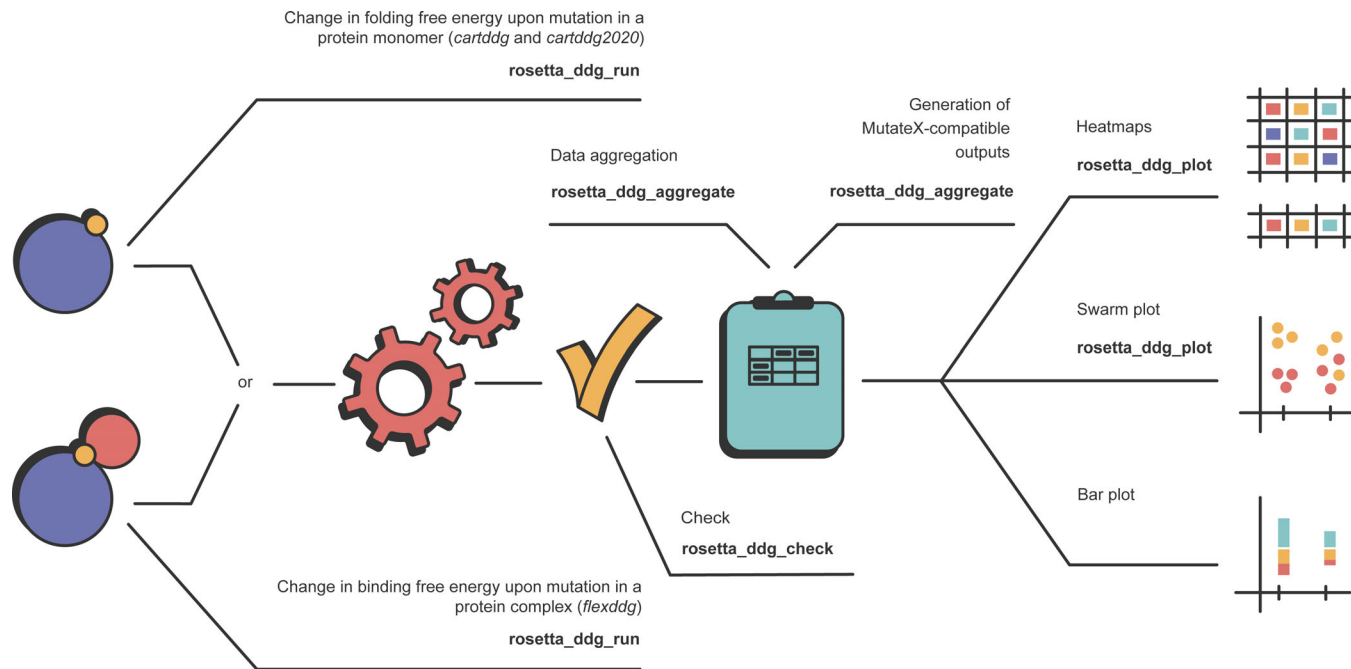


FIGURE 1 The RosettaDDGPrediction workflow and schematized plot types. The first step consists in running the *rosetta_ddg_run* executable to obtain the predicted $\Delta\Delta G$ values for the changes in folding free energy (for monomeric proteins) or binding free energy (for protein complexes). Then, *rosetta_ddg_check* can be used to ensure that all runs have been completed successfully. Data aggregation can then be performed with *rosetta_ddg_aggregate*, and aggregate data can finally be visualized in different ways (heatmaps, bar plots, swarm plots) using *rosetta_ddg_plot*.

In details, it is possible to provide a list of mutations containing both the positions and the residues to which such positions should be mutated (i.e., every single mutation is uniquely identified by the protein chain, the wild-type residue, the position of the residue in the chain, and the mutated residue). As an alternative, the user can specify multiple mutations to be performed simultaneously on different protein residues. On the other hand, the user can also pass a list of residues (each one identified by the protein chain, the wild-type residue, and the position of the residue in the chain) and a list of residue types. In the latter case, all the residues specified in the first list will be mutated, one at a time, to each residue type specified in the second list. This allows RosettaDDGPrediction to implement saturation mutagenesis scans alongside scans of specific mutations.

The *rosetta_ddg_run* executable can optimize the workload distribution over the available resources to ensure efficient scheduling of the runs, thanks to the Dask Python package operating under the hood. *rosetta_ddg_run* easily handles multistep protocols, requiring sequential Rosetta calls and possibly processing the output data between the steps. For example, for the aforementioned *cartddg* and *cartddg2020* protocols, *rosetta_ddg_run* takes care of both Rosetta calls and the processing steps.

Once the runs are completed, users can perform a sanity check on the calculations using *rosetta_ddg_check_run*, which identifies problematic runs by scraping the Rosetta output files. Finally, *rosetta_ddg_aggregate* can aggregate raw data from the large numbers of collected mutation runs into easily readable CSV table files for successfully completed runs. These aggregate files contain, together with the calculated differences in free energy, additional information about each mutation, the Rosetta energy function used, and the number of structures generated. *rosetta_ddg_aggregate* also allows generating aggregate outputs compatible with the MutateX plotting system. Indeed, MutateX offers additional visualization tools, including density plots, logo plots, distribution plots, and summary tables that can be easily navigated (Tiberti et al., 2022).

Finally, *rosetta_ddg_plot* provides plotting utilities to explore the aggregated data through several visualization types, such as one-dimensional or two-dimensional heatmaps. The latter is particularly convenient when a saturation mutagenesis scan is run on a set of positions. In addition, the contribution of each term of the energy function to the final $\Delta\Delta G$ values can be visualized as stacked bar plots, where positive and negative contributions add up on the corresponding semiaxes. Finally, since all protocols implemented so far in

TABLE 1 Examples of performances of RosettaDDGPrediction for different protein sizes, the number of cores, and protocols applied

Protein size	Number of cores	Protocol	Time (h)
120	16	cartddg (ref2015)	33
250	24	cartddg (ref2015)	67
250	8	cartddg (talaris2014)	89
340	16	cartddg (ref2015)	160
340	16	cartddg (talaris2014)	70
340	1	Relax	16
600	1	Relax	40
900	1	Relax	65
120; 17 ^a	40	flexddg (talaris2014)	67

Note: In the case of complexes, the “protein size” column includes two values, that is, one value for each protein/peptide in the complex. Calculations were run on servers equipped with either dual Xeon 6142 processors or dual Xeon 6242 processors. Each processor features 32 cores. The estimate refers to calculations done with Rosetta version 3.12.

^aThe one for which the saturation mutational scan was carried out.

RosettaDDGPrediction determine the $\Delta\Delta G$ value associated with a mutation by averaging over the values produced by an ensemble of structures, the user may want to visualize the distribution of such values to investigate the source of potential outliers that may bias the average. In this case, a swarm plot displaying such values as separate data points constitutes a very insightful overview provided by *rosetta_ddg_plot*.

To guide the user on the number of cores and time required for calculation, depending on the RosettaDDGPrediction protocol, energy function, and protein size, we reported the results for different saturation scans in Table 1.

2.2 | Case Study 1: Prediction of changes in folding free energy upon mutations and comparison with experimental values from the ThermoMut database

To illustrate the performance of the *ref2015* energy function, we performed folding free energy calculations with both the *cartddg* (Figure 2) and the *cartddg2020* (Figure S1) protocols and compared them to experimentally determined unfolding $\Delta\Delta G$ values. The following section illustrates, as an example, our findings when using the *cartddg* protocol. We downloaded the ThermoMut database (ThermoMutDB) (Xavier et al., 2021) and selected four proteins as detailed in Section 4. In particular, we selected two bacterial enzymes with 117 and 597 mutations, respectively: Enterobacteria phage T4 Endolysin, ENLYS (UniProt ID: P00720), and *Staphylococcus aureus* Thermo-nuclease, NUC (P00644). In addition, we performed the calculations on two human proteins of interest in health and disease, that is, TP53

(P04637) and FKBP1A (P62942), with 45 and 68 mutations with structural coverage, respectively. We applied the secondary structure definition of PDBe (Varadi, Anyango, Armstrong, et al., 2022) and annotated each position as either α -helix, β -sheet, or loop in the wild-type structures. This case study investigates the relationship between experimental and predicted values per-mutation when the data from all four proteins are pooled, allowing us to achieve better statistical power than considering each protein separately.

We performed a preliminary data exploration to understand the agreement between the experimentally determined and the predicted stability. Interestingly, data points from the experimental and prediction dataset were similarly distributed (Figure 2a), as corroborated by the Kolmogorov–Smirnov test ($p = 0.21$).

We then investigated the relationship between predicted and experimental data using a simple linear regression model (SLM), assuming that a perfect agreement between the experimental and predicted values would have an intercept of 0 and a coefficient of 1. The SLM regression line had an intercept of 0.81 and a slope of 0.719 (Figure 2b). The variance of the linear model (σ^2) is 3.95, and the model produced an R^2 of 0.44, a Pearson correlation coefficient (PCC) of 0.66, and a mean absolute error (MAE) between the predicted and experimental $\Delta\Delta G$ s (MAE) of 1.39. The residuals plot for this model showed how the poor R^2 value was at least partially due to systematic bias (Figure S2). This illustrates that a linear model does not entirely explain the variance in the data.

To better understand this behavior, we tried to fit the data using a generalized additive model (GAM) (Figure 2d). The resulting model had a roughly linear behavior in the ~ 0 –5 kcal/mol range but becomes less so

at lower or larger $\Delta\Delta G$ values. Similarly, the confidence interval is very narrow in the linear regime interval, and it is wider for large and small $\Delta\Delta G$ values, for which we have fewer data points. This observation is in alignment

with Høie et al. (Høie et al., 2022), who found that $\Delta\Delta G$ predictions made with *ref2015* and the *cartesian2020* protocol in 29 proteins correlated with altered protein functions for $\Delta\Delta G > 4.5$ kcal/mol, but the severity of the

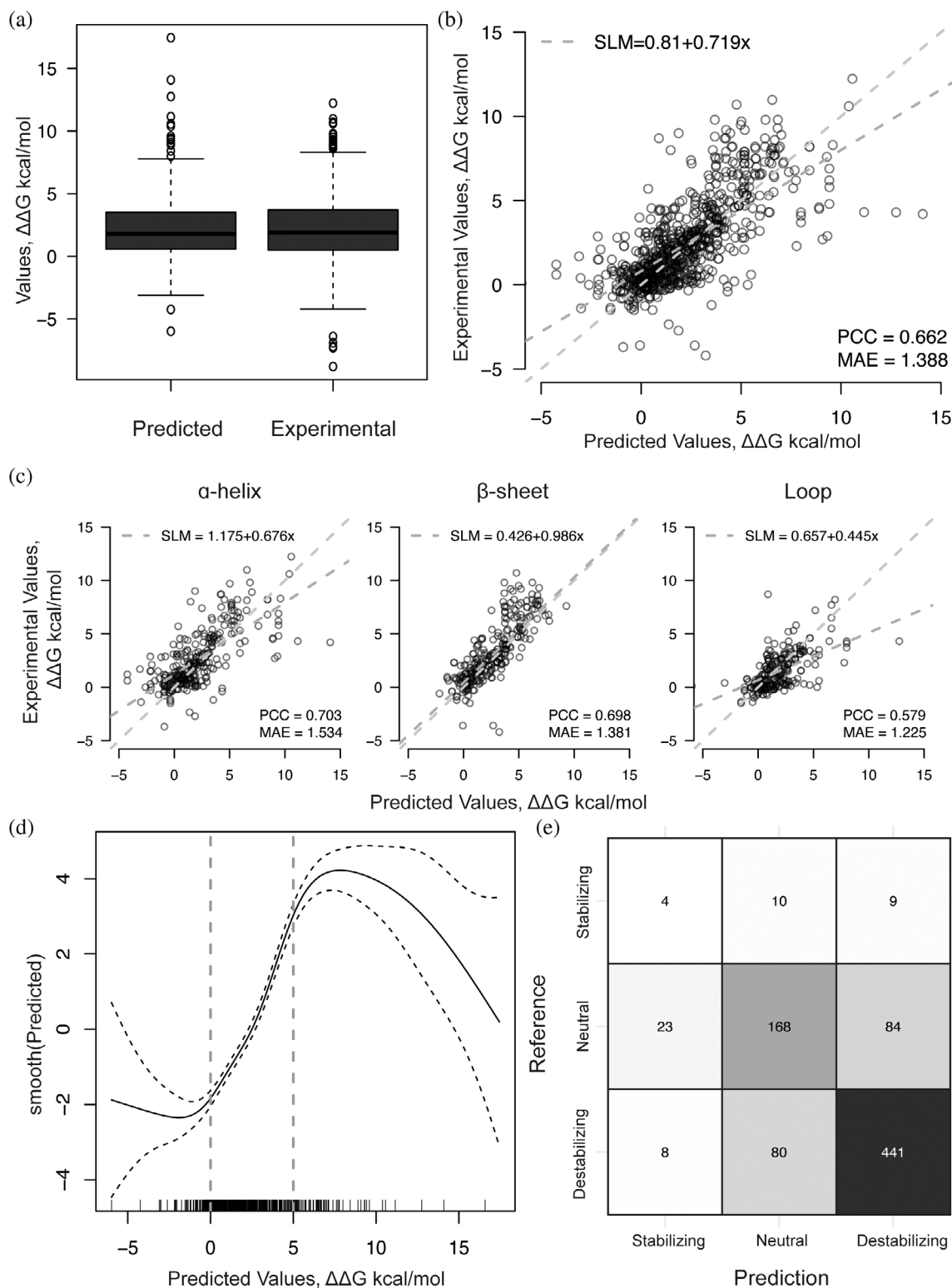


FIGURE 2 Legend on next page.

impact did not increase remarkably beyond this point. We then assessed the impact of the secondary structure on the performance of the prediction by building a SLM for each of the secondary structure groups, divided into α -helices, β -sheets, and loops (Figure 2c). Residues involved in structured regions are more likely to be part of the protein core, less flexible, and more sensitive to mutation with respect to solvent-accessible unstructured loops. According to the PCC, MAE slope, σ^2 , and R^2 values, the prediction was less consistent for the unstructured regions. Indeed, the loop subset featured a low PCC score (0.58), indicating a poor linear relationship. On the other side, the loop subset also resulted in the lowest MAE value (0.33) between predicted and experimental data, which could imply a better fit to the model. The low MAE is, however, an artifact of the comparatively low $\Delta\Delta G$ values observed in the loop regions. Loop regions are often flexible and less sensitive to changes in folding/unfolding $\Delta\Delta G$ upon mutation. This caused all the predicted or experimentally determined values to be grouped close together (Figure 2c), which caused lower MAE values than what observed for the α -helix or β -sheet subsets. In the loop dataset, we noticed several outliers in which amino acid substitutions are predicted to have a large destabilizing impact, whereas the experiments found the same variants to be neutral or mildly destabilizing. The experimental findings mostly align with the expectation that substitutions in flexible loops have mild effects on stability. However, some loop substitutions may extend or create secondary structure elements, for example, as a result of substitutions from proline (Pires et al., 2019). The difference witnessed in this dataset was likely due to Rosetta allowing local main chain flexibility, which might not be enough to represent the conformational heterogeneity that disordered regions experience in solution. We noticed similar behavior in applying

FoldX, which we could mitigate using ensembles of structures generated, for example, by molecular dynamics simulations (Fas et al., 2020; Nygaard et al., 2016; Tiberti et al., 2022). It should be noted that the α -helix mutations dataset also contains outliers. This dataset, however, had an overall better correlation with the experimental dataset, and the coefficient of its regression line is closer to 1. This suggests that changes in loops are more difficult to predict.

We then evaluated the performance of the predictions in classifying mutations into destabilizing, neutral, and stabilizing classes. We did so by classifying all mutations causing stability changes above 1 kcal/mol as destabilizing, all mutations causing stability changes between 1 and -1 kcal/mol as neutral, and all mutations causing stability changes below -1 kcal/mol as stabilizing (Frenz et al., 2020; Park et al., 2016) and constructing a confusion matrix (Figure 2e). This confusion matrix yielded an accuracy of 0.74. The prediction accuracy was best for the destabilizing class (0.76), with high sensitivity (0.83), while the accuracy of the stabilizing class was only 0.56, with a sensitivity of 0.17, indicating that the destabilizing class is more likely to be correctly identified as compared to the stabilizing class. The low accuracy could also be expected due to the imbalanced dataset available for the study, where a low number of stabilizing mutations is available. While this dataset is not balanced, this may not explain the bias in full, as the methodology itself could have been developed on a biased dataset (Bæk & Kepp, 2022; Pancotti et al., 2022). The neutral class performed similarly to the destabilizing class (Table S1).

We performed the same analyses on the dataset obtained using the *cartesian2020* protocol, which showed similar trends overall (Figure S1). Additionally, for comparison, we used a recent deep learning tool, which aims at simulating the *cartddg* protocol, that is, RaSP

FIGURE 2 Comparison of changes in structural stability predicted with the *ref2015 cartddg* protocol and experiments. (a) Distribution of the predicted and experimental stability changes in kilocalorie per mole. (b) Scatterplot of the $\Delta\Delta G$ values predicted by the *ref2015 cartddg* protocol and experimental values for the corresponding mutations. The blue line indicates a perfect correspondence between the variables. The green line is the fitted simple linear model. The model has an intercept of 0.81, a slope of 0.72, a variance (σ^2) of 3.95, and a R^2 of 0.44, a Pearson's correlation coefficient (PCC) of 0.66, and a mean absolute error between the predicted and experimental $\Delta\Delta G$ s (MAE) of 1.39. (c) Scatterplots dividing the data by the wild-type secondary structure of the mutated position. The blue line indicates a perfect correspondence between the variables for each plot. The green line is the fitted simple linear model. Here, it is evident how the structured sections have a better correlation when compared to the loops. This is likely due to the flexibility of the unstructured sections. α -helices: PCC = 0.70, slope = 0.68, σ^2 = 3.63, R^2 = 0.49, MAE = 1.53. β -sheets: PCC = 0.70, slope = 0.99, σ^2 = 4.98, R^2 = 0.49, MAE = 1.38. Loop: PCC = 0.58, slope = 0.45, σ^2 = 1.99, R^2 = 0.33, MAE = 1.22. (d) Generalized additive model (GAM) modeling the response variable, the experimental $\Delta\Delta G$ value, to a predictive variable, the predicted $\Delta\Delta G$ value, by estimating a smooth function, smooth (predict). The smooth function has an effective degree of freedom of 6.5, quantifying the complexity of the line. The dotted black lines indicate the confidence interval, which is sufficiently narrow in the $\Delta\Delta G$ interval 0–5 kcal/mol (indicated with red dotted lines) to indicate that a linear relationship is present in this interval. (e) Confusion matrix where the experimental values are annotated as the reference values. The threshold used to define the classes is a $\Delta\Delta G$ of < -1 kcal/mol for stabilizing mutations, $-1 < \Delta\Delta G < 1$ kcal/mol for neutral mutations and $\Delta\Delta G > 1$ kcal/mol for destabilizing mutations. The resulting accuracy is 0.74.

(Blaabjerg et al., 2022) (Figure S3). Here, we also noticed a remarkable performance both compared to both the experimental values and the Rosetta predictions.

In conclusion, the Case Study 1 showed a good linear correlation between predicted and experimental values, especially in the range of 0–5 kcal/mol, where the trend is generally conserved. Outside this range, the relationship between the predicted and the experimental values is less direct. We also show how predictions are more reliable for structured regions of the protein while correlation values are lower for unstructured regions.

2.3 | Case Study 2: Prediction of changes in binding free energy for protein-short linear motifs interactions

Within the protein–protein interaction landscape, intrinsically disordered proteins (IDPs) or regions (IDRs) have been proven to play an essential role in different biological events. IDPs and IDRs include functional motifs known as short linear motifs (SLiMs) that are important for the binding between IDPs and their target proteins (Davey, 2019; Davey et al., 2011; van Roey et al., 2014). An example is the LC3 interacting region (LIR), that is, a class of SLiMs involved in selective autophagy (Sora et al., 2020). One of the main features for regulating LIR binding to proteins of the LC3 family is through PTMs, especially through phosphorylation (Sora et al., 2020). Here, we aim to show an application of the *flexddg* protocol to capture the changes in binding free energy upon phosphorylation or mutations in the core region of LIR-containing proteins.

First, we selected two examples of experimentally characterized phospho-regulated LIRs for which the structures were available on the Protein Data Bank, that is, FUNDC1 in complex with LC3B (PDB entry 2N9X; Kuang et al., 2016) and PIK3C3 in complex with GABARAP (PDB entry 6HOG; Birgisdottir et al., 2019). Experimental data from isothermal titration calorimetry (ITC) or peptide arrays are available for these two complexes and include phosphorylations or phospho-mimetics (Birgisdottir et al., 2019; Kuang et al., 2016; Lv et al., 2017). We applied the *flexddg* protocol with the *talaris2014* Rosetta energy function to investigate the effects of single and multiple phospho-mimetic mutations at the known phosphosites (see Section 4). Indeed, Rosetta does not provide parameters for phosphorylated residues. In addition, we included a comparison with the estimates provided by FoldX using the binding free energy protocol implemented in MutateX. The results are described in detail below and reported in Figure 3.

FUNDC1 is a mitophagy receptor that mediates the selective removal of damaged mitochondria. It contains a canonical LIR (core region, 18-YEVL-21), which is necessary for interacting with LC3 and its role in mitophagy (Kuang et al., 2016). FUNDC1 presents three experimentally validated phosphosites in the surroundings of its LIR motif: S13, S17, and Y18 (Figure 3a). ITC experiments with different FUNDC1 LIR peptides and LC3B reported a K_d of $0.40 \pm 0.06 \mu\text{M}$ for the wild-type variant. Phosphorylation at the S13 site resulted only in a slight decrease of the LC3B affinity ($K_d = 0.60 \pm 0.05 \mu\text{M}$) with respect to the wild type. On the other hand, Y18 phosphorylation caused a five-fold K_d increase ($K_d = 1.72 \pm 0.30 \mu\text{M}$). This increase is slightly augmented if both phosphorylations are combined ($K_d = 2.00 \pm 0.37 \mu\text{M}$) (Kuang et al., 2016). Additionally, another work reported that S17 phosphorylation has an opposite effect and increases the binding affinity for LC3B by three folds (Lv et al., 2017). The *flexddg* protocol predicted the S13D and S13E substitutions to have neutral effects on the binding, in agreement with experiments (i.e., average $\Delta\Delta G < 0.25$ kcal/mol). However, the average $\Delta\Delta G$ s for the S17E and S17D mutations are also low, suggesting that, in this other case, the prediction cannot capture the changes in the binding affinity observed experimentally (Figure 3a). In the case of the single phospho-mimetic mutations at Y18 and S13, the predicted $\Delta\Delta G$ sign was in overall agreement with the effect measured experimentally. Although, we noticed that, in this case, to use tryptophan as a phospho-mimetic residue for phosphorylated tyrosine does not efficiently capture the destabilizing effects of the PTM.

Surprisingly, the combination of phospho-mimetic mutations at S13 and Y18 sites (i.e., S13E_Y18E and S13E_Y18W) resulted in negative $\Delta\Delta G$ values, suggesting a stabilizing effect in disagreement with what observed experimentally (Figure 3b). Nevertheless, we observed that the associated standard deviations are very high not allowing for quantitative conclusions.

We then studied PIK3C3, a class III phosphoinositide 3-kinase enzyme of the PtdIns3K complexes, involved in autophagy initiation. PIK3C3 presents a canonical F-type LIR (250-FELV-253) required for the interaction with GABARAP and GABARAPL1 (Birgisdottir, et al., 2019). The effect of a double phosphorylation at S244 and S249 was studied with ITC. In these experiments, the substitution of both the phosphosites with glutamate caused a 17-fold increase in GABARAP binding ($K_d = 2.9 \pm 0.1 \mu\text{M}$) compared to the wild-type variant ($K_d = 49.5 \pm 3.9 \mu\text{M}$). Moreover, peptide array experiments showed an increase in the binding affinity of the LIR peptide with all the LC3 family members for the S249E variant (Birgisdottir et al., 2019).

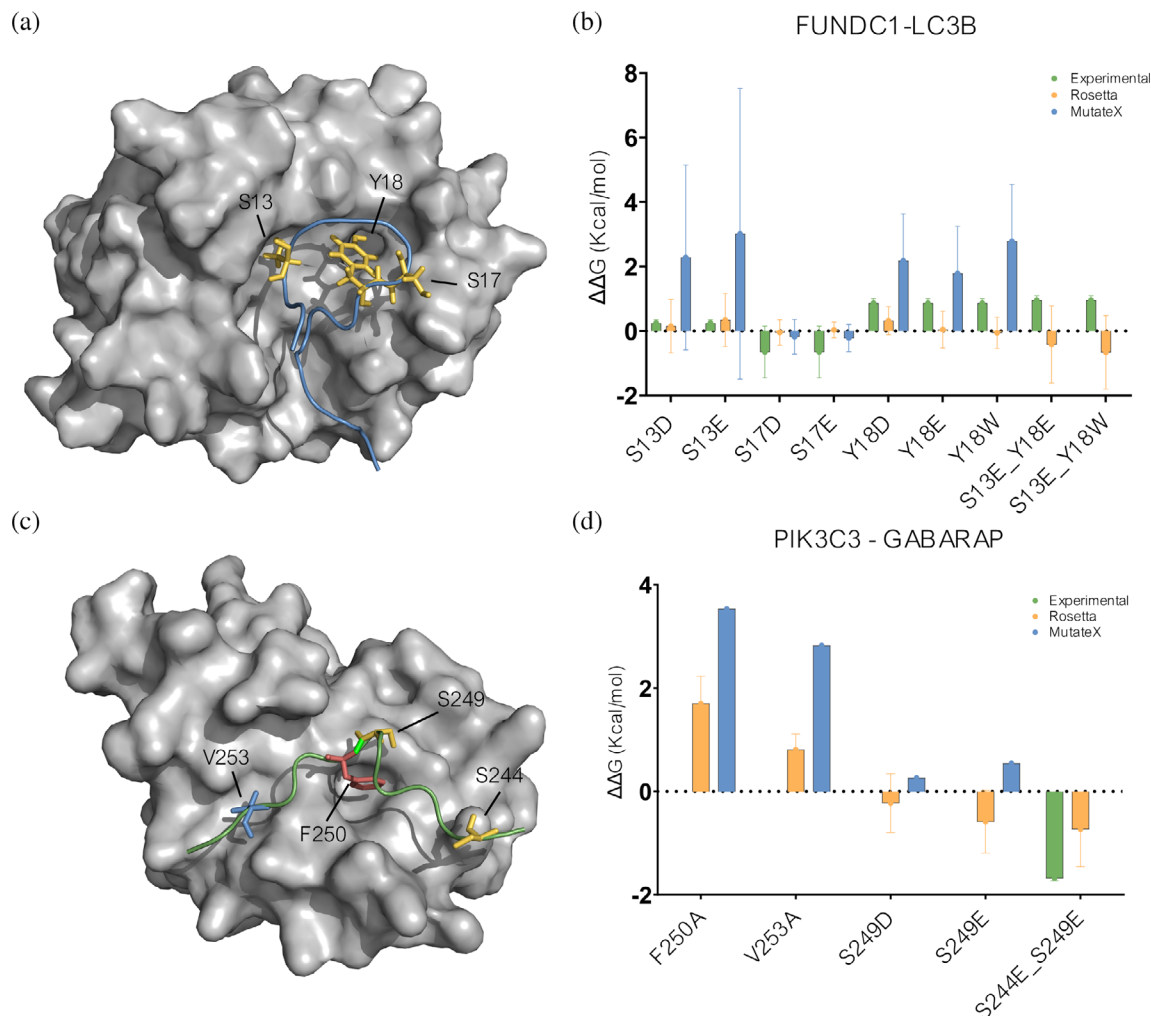


FIGURE 3 Prediction of changes in binding free energy using the *flexddg* protocol for protein interactions mediated by short linear motifs. (a) FUNDC1 LIR peptide (blue) in complex with LC3B (gray) in the structure associated with the PDB entry 2N9X. The S13, Y18, and S17 phosphosites are shown as sticks and colored in yellow. (b) We report the predicted binding $\Delta\Delta G$ s for the single and double phosphomimetic mutations for the FUNDC1 LIR phosphosites for which experimental data are available for comparison, along with the same single mutations predicted with MutateX. For every variant, we also included the $\Delta\Delta G$ values obtained from experimental K_d of the phosphorylated variants (pS13, pS17, pY18, and pS13_pY18). (c) PIK3C3 LIR peptide (blue) in complex with GABARAP (gray) in the structure associated with the PDB entry 6HOG. The S244 and S249 phosphosites are shown as sticks and colored in yellow, while the residues for binding to the GABARAP HP1 and HP2 pockets are shown as red and blue sticks, respectively. (d) We report the predicted binding $\Delta\Delta G$ s for single and double phosphomimetic mutations, along with mutations to alanine in the core motif of the PIK3C3 LIR, for which experimental data are available for comparison, along with the same single mutations predicted with MutateX. We also included the $\Delta\Delta G$ value obtained from the experimental K_d of a S244E_S249E variant.

To assess the potential of the *flexddg* protocol in capturing the effects induced by the phosphorylations of the PIK3C3 LIR, we modeled the S249E variant and a variant including phosphomimetic mutations at both the S244 and S249 sites (i.e., S244E_S249E; Figure 3c). We also tested the effect of the S249D substitution as a possible phosphomimetic, even if no experimental data are available for this mutation. We noticed that using S249D as a phosphomimetic provides different result than introducing a glutamate (Figure 3d). This supports the notion that aspartate and glutamate cannot always be used as

phosphomimetics in an interchangeable manner. The S249E variant had a slightly stabilizing effect on the binding (average $\Delta\Delta G = -0.59$ kcal/mol) and, in general, values of $\Delta\Delta G$ lower than 0 across the 35 independent runs (Figure 3d). This is in partial agreement with the peptide array results mentioned above. The results for the double mutant variant S244E_S249E are in agreement with the expected increase in binding affinity observed experimentally, even if with large deviations across measurements, suggesting that the *flexddg* protocol could provide, in some cases, a qualitative

understanding of the effects of multiple amino acid substitution if they are located at structural proximity. We noticed that FoldX failed in identifying the stabilizing effect of the mutation on the binding and predicted slightly destabilizing effects despite the possibility to model the phosphorylated variant of the residue.

Furthermore, we evaluated whether the *flexddg* protocol could provide insights into the effects of mutations in SLiMs where PTMs are not involved. In the case of LIRs, the interaction between an LIR-containing protein and an LC3 family member is mainly driven by two residues of the LIR motif, which bind to the hydrophobic pocket 1 (HP1) and the hydrophobic pocket 2 (HP2) residues of the LC3 protein, respectively (Sora et al., 2020). Thus, we tested the capability of the *flexddg* protocol with the *talaris2014* energy function to predict the impact of the known detrimental mutations F250A (residue for interaction with HP1 pocket) and V253A (HP2 pocket) of PIK3C in complex with GABARAP (PDB entry, 6HOG; Birgisdottir et al., 2019). We observed a good agreement between Rosetta- and FoldX-based calculations in identifying these mutations as detrimental for binding to GABARAP (Figure 3d) (Sora et al., 2020).

Overall, Case Study 2 illustrates the potential and limitations of the RosettaDDGPrediction workflow. We identified as main challenges the prediction of increased binding affinity (i.e., stabilizing mutations), to study combined mutations and the usage of phosphomimetic mutations instead of phosphorylated residues. Moreover, we noticed that the results from RosettaDDGPrediction are more consistent with the $\Delta\Delta G$ s from the ITC experiments with respect to the ones obtained from the MutateX protocol with the FoldX free energy function despite FoldX allows to include phosphorylated residues. The estimates provided by FoldX seem to capture variants destabilizing for the binding to the target protein but with highest $\Delta\Delta G$ than experimentally measured.

2.4 | Case Study 3: Influence of the source of initial structures for the calculations

Using structural models to perform prediction of $\Delta\Delta G$ s is a tantalizing perspective because of intrinsic limitations in the availability of experimental structures. This has been shown to be reliable to a good extent—for instance, using homology models with Rosetta allowed to achieve similar performance when comparing predictions with experimental $\Delta\Delta G$ s, as long as the sequence identity of the template to the target protein was at least of 40% (Valanciute et al., n.d.) and results obtained using Rosetta are relatively robust to the use of models (Blaabjerg

et al., 2022; Valanciute et al., n.d.). The advent of AlphaFold has revolutionized molecular modeling and structural biology (Jumper et al., 2021), resulting in models of 3D structures of proteins with quality comparable to that achievable with experimental approaches and useful in the context of computational biology, including the prediction of changes of free energy (Akdel et al., 2022). The current version (release 4) of the AlphaFold Protein Structure Database contains over 214 million predicted protein structures, corresponding to most proteins in Uniprot 2021_4 and including 48 complete proteomes (Varadi, Anyango, Deshpande, et al., 2022), providing a rich source of structures for in silico mutational scans.

Here, we evaluated the influence of using a model based on AlphaFold2 with respect to an x-ray structure of the same protein with good resolution. For this goal, we used as a case study the DNA binding domain (DBD) of p53, for which experimental data are also available on 31 mutant variants from ThermoMutDB (Xavier et al., 2021). We evaluated the agreement between our calculated and experimentally available data using the same parameters and energy functions, either the *cartddg* or *cartddg2020* protocol, and the two different starting structures. We also included a variant of *cartddg* in which we increased the number of runs per mutation up to 10 to determine whether it would improve our results. As the final $\Delta\Delta G$ depends on the values obtained by the single runs, we expect that increasing the number of samples might lead to better converged final $\Delta\Delta G$ values. We measured the agreement through several metrics, such as the Pearson correlation coefficient, MAE, and a ROC curve. We performed most of our comparison considering runs performed with the *cartddg* protocol. Therefore, in this section, we will refer to the *cartddg* protocol unless stated otherwise.

We obtained a similar pattern when comparing predictions and experiments using the experimental structure and the AlphaFold2 model (Figure 4a) with a positive linear correlation, as quantified by the Pearson correlation coefficient (Figure 4b). The highest Pearson correlation coefficient obtained was 0.79 using the scoring function *talaris2014* with the AlphaFold2 model and 10 runs (Figure 4b). However, all runs, including the ones using the *cartddg2020* protocol, achieved a correlation in the range of 0.57–0.79. Values ranging from 0.74 to 0.79 were obtained by all runs using the x-ray structure and by *talaris2014* with AlphaFold2 using 3 or 10 runs. Using *ref2015* with the AlphaFold2 model led to a slightly worse correlation of 0.57 for 3 runs and 0.68 for 10 runs. The runs with *ref2015* energy function and the *cartddg* protocol (x-ray structure) using 10 runs had the smallest MAE of 0.90 kcal/mol (Figure 4c). This result suggests that this combination featured the lowest average

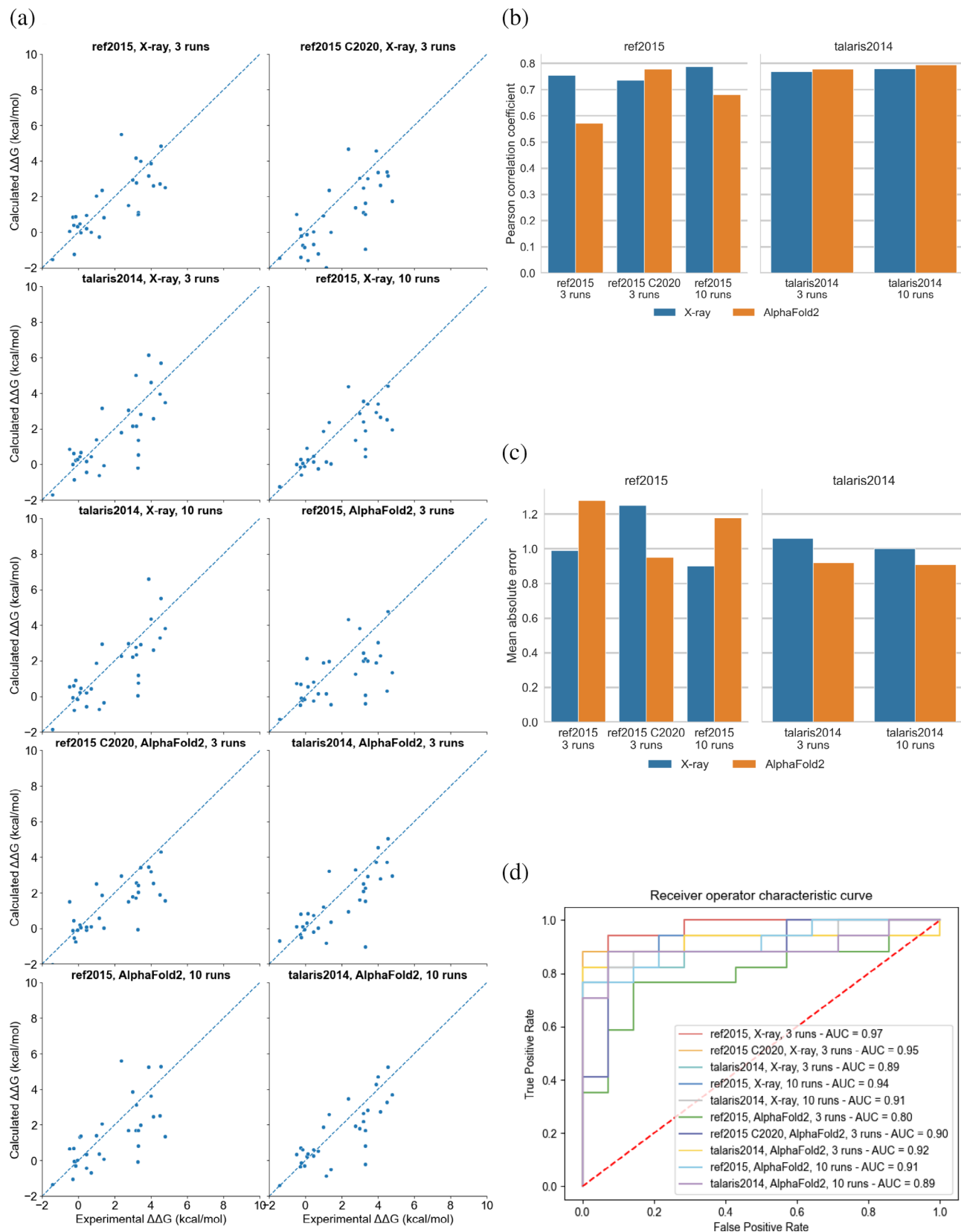


FIGURE 4 Legend on next page.

distance between predicted and target values among all the different tested methods. It was closely followed by *talaris2014* with the AlphaFold2 structure using 3 and 10 cycles with 0.92 and 0.91, respectively. The rest of the combinations had a MAE between 0.95 and 1.28.

Considering the ROC curve, we used the experimental free energy changes from ThermoMutDB as ground truth. We partitioned our dataset into destabilizing and nondestabilizing mutations depending on whether our prediction or ground truth had $\Delta\Delta G > = 1.2$ kcal/mol (Degn et al., 2022). The best area under the curve (AUC) was achieved by using the scoring function *ref2015* using the *cartdgg* protocol and the x-ray structure, yielding a value of 0.97 (Figure 4d). In general, the different scoring functions and structures behaved similarly.

The usage of the experimental x-ray structure or the AlphaFold2 model did not affect the prediction performance. The only exception was the combination of the *ref2015* energy function and the *cartddg* protocol with the AlphaFold2 model, which had a lower correlation and ROC AUC with respect to the other cases. Increasing the number of runs also slightly improved the performance, but with the trade-off of a considerably increased computing time. Finally, we obtained mixed results when comparing the *ref2015* x-ray three-cycles with *cartddg* with the corresponding *cartddg2020* run. We did not see any appreciable improvement when using *cartddg2020* on the x-ray structure, as the *cartddg2020* run has a slightly lower correlation (0.74 vs. 0.76), higher MAE (1.25 vs. 0.99 kcal/mol), and lower AUC (0.95 vs. 0.97) considering the experimental data. Nonetheless, using *cartddg2020* with the AlphaFold2 model rescued the subpar performance of *ref2015*, as all its performance measures are more similar to those of the other cases.

It should be noted that the DNA-binding domain in the p53 AlphaFold2 model, ranging from residues 91 to 289, features a good per-residue confidence score (pLDDT) score, mostly above 70. This implies that more tests on models or regions with lower quality should be carried out in the future to determine whether our findings can be generalized. It has been shown that protein regions predicted with a confidence score lower than 50 are less in agreement with experimental data, possibly

due to the low-confidence regions including more often disordered regions with a higher tolerance to mutations affecting stability (Akdel et al., 2022).

2.5 | Case Study 4: Variants predisposing to childhood cancer

In a recent study, 198 samples from different childhood cancer types were analyzed in terms of germline variation and cancer predisposition (Byrjalsen et al., 2020). Among these, different variants of uncertain significance (VUS) have been found with a frequency of <1% in the healthy population. Approximately, 20% of the patients investigated had VUS in DNA repair pathway genes. In addition, we carried out new analyses on a larger dataset accounting for more than 550 germline samples from Danish children. The selection criteria for the proteins and the variants included in the study are described in detail in Section 4 and in Figures S4 and S5. We retained 14 proteins, that is, ERCC4, BLM, FANCA, FANCE, FANCF, FANCG, FANCI, FANCL, MLH1, MSH2, MSH6, NBN, RAD51C, and RFD3 for structure-based calculations of the changes in folding $\Delta\Delta G$ s for the VUS. All these genes are classified as tumor suppressor genes in the COSMIC Cancer Gene Census v96 (Sondka et al., 2018) or from the literature, in the case of FANCI (Zhang et al., 2016) and RAD51C (Somyajit et al., 2010).

Since mutations in tumor suppressor genes are generally causing loss-of-function in cancer (Wang et al., 2018), we were interested in identifying VUS that could destabilize the protein structure and result in positive predicted $\Delta\Delta G$ values upon mutation. These variants could be relevant to investigate further in terms of genomic alterations predisposing to cancer. To this aim, we retained the variants with structural coverage in AlphaFold2 and high confidence scores for a total of 126 variants analyzed (Figures 5–7 and Table S2). According to searches in ClinVar (Landrum et al., 2014, 2020), some of the variants were annotated as benign or likely benign but not related to childhood cancer. On the other hand, only T1131A in FANCA was found as pathogenic. The remaining were not deposited in ClinVar or annotated as

FIGURE 4 Comparison of experimental and predicted $\Delta\Delta G$ s using p53 as a case study. $\Delta\Delta G$ values were predicted using Rosetta version 3.12 with the *ref2015* and *talaris2014* scoring functions, and the *cartddg* and *cartddg2020* protocols (referred to as “C2020” in the figure). We used the x-ray structure (PDB entry 2XWR) and a model from the AlphaFold2 database for the residues 91–289 of p53 as initial structures, using our default number of runs (3) or 10 runs. (a) Experimental versus predicted $\Delta\Delta G$ values. (b) Pearson's correlation coefficient between experimental and predicted values. (c) Mean absolute error (MAE) between experimental and predicted values. (d) Receiver operator characteristic (ROC) curve. The classification for this curve was done by considering the changes of free energy values reported in ThermoMutDB as ground truth, using 1.2 kcal/mol as $\Delta\Delta G$ cut-off to distinguish between destabilizing and nondestabilizing mutations (see Section 4). The same criterion was used for the predicted mutations.

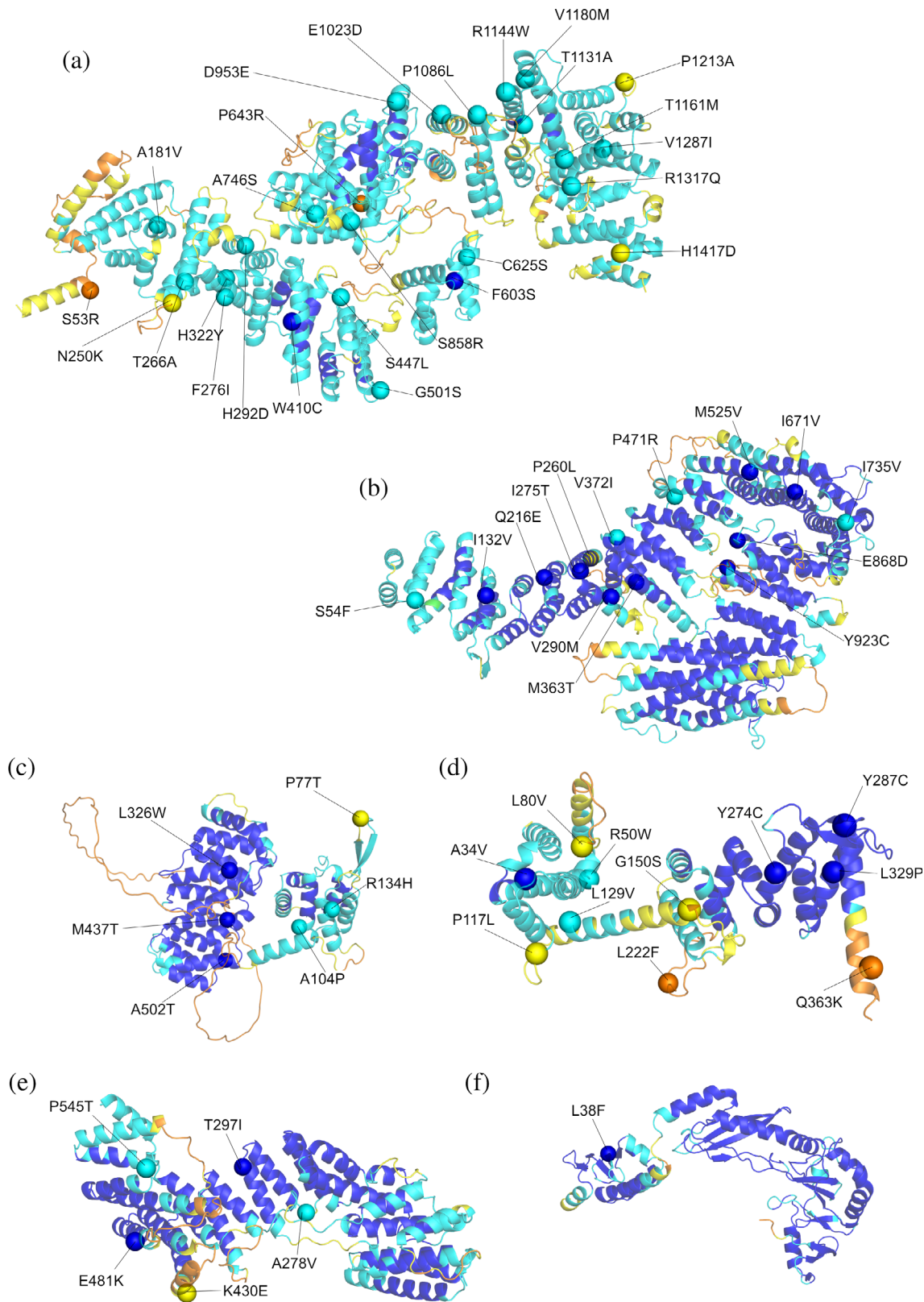


FIGURE 5 Trimmed AlphaFold structures of the FA (Fanconi Anemia) proteins selected for the Case Study 4. Cartoon representation of (a) FANCA₃₇₋₁₄₄₁, (b) FANCI₁₋₁₂₇₉, (c) FANCE₁₂₋₅₃₄, (d) FANCF₂₋₃₆₉, (e) FANCG₁₂₋₆₁₆, and (f) FANCL₁₋₃₇₅. The proteins are colored according to the AlphaFold2 pLDDT score: very low (orange, pLDDT < 50), low (yellow, 50 < pLDDT < 70), confident (light blue, 70 < pLDDT < 90), and very high (blue, pLDDT > 90). The Ca of the residues found mutated in pediatric cancer patients are shown as spheres and labeled.

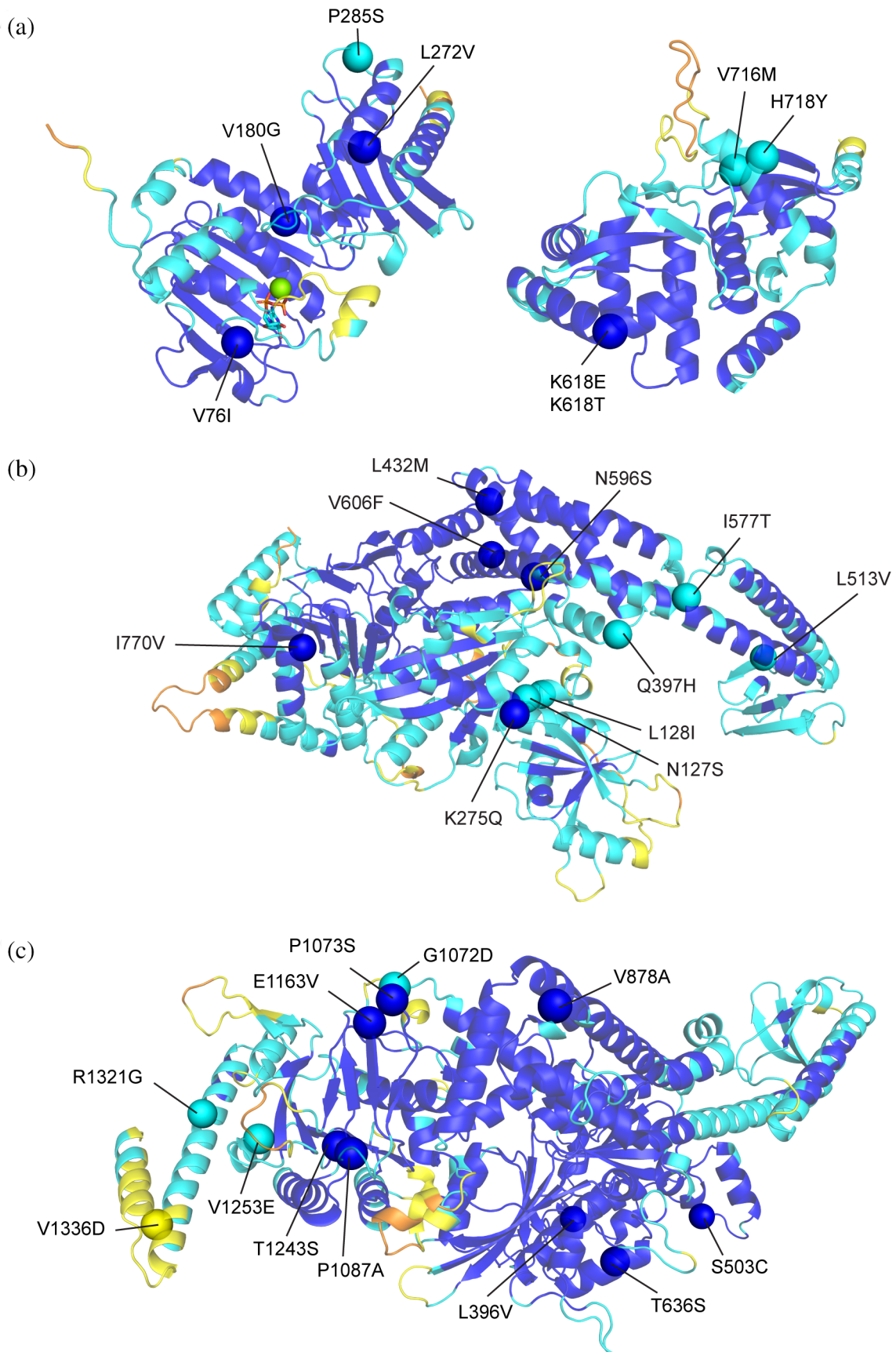


FIGURE 6 Legend on next page.

uncertain significance or with conflicting evidence, emphasizing the importance of additional analyses to understand the effects at the protein level.

In this example, we applied the *cartddg2020* protocol, which considers the $\Delta\Delta G$ value referring to the mutant structure with the lowest total energy. At first, we retained, as predicted destabilizing, the variants with $\Delta\Delta G$ values >1 kcal/mol (see Section 4) and confirmed destabilizing by calculations with MutateX (Table 2). Indeed, the *foldX5* energy function, which is applied in the MutateX protocol, is effective in capturing loss-of-function mutations (Gerasimavicius et al., 2022; Nielsen et al., 2017; Scheller et al., 2019). Of note, the pathogenic variant T1131A is not predicted to destabilize the structure of FANCA by both Rosetta and FoldX calculations. We hypothesize that the detrimental effects triggered by this variant could be due to other properties such as impaired activity, interactions, or PTMs at the cellular level. Experimental studies at the cellular level confirm that the T1131A substitution does not affect the protein levels, in agreement with a neutral effect on the folding $\Delta\Delta G$ s (Wilkes et al., 2017) and that the phenotype reflects a functional impairment that has a mild impact on drug sensitivity and the monoubiquitination of another protein (Adachi et al., 2002; Wilkes et al., 2017). T1131A could be further investigated using our recently proposed multilayered structural framework for variant annotations in proteins, that is, Multilayered Assessment of Variants by Structure for proteins (MAVISp) (Arnaudi et al., 2022).

We also observed that one variant annotated as benign in ClinVar (i.e., L605F FANCI) has predicted changes in folding $\Delta\Delta G$ higher than 3 kcal/mol and is, therefore, classified as destabilizing for the structural stability by our analysis. The variant has been characterized with cellular assays, showing decreased protein levels when compared to the wild type, which confirms our prediction (Fierheller et al., 2021). On the other hand, the variant P55L (predicted folding $\Delta\Delta G < 2.0$ kcal/mol) was expressed at the same level as the wild-type variant. In addition, other benign variants in ClinVar resulted in changes in free energy in the range of 1–2 kcal/mol (Table 2). This observation suggests that variants for which the predicted changes in stability are within 1–3 kcal/mol should be further investigated to evaluate whether they could result in neutral effects in terms of protein levels in the cell or propensity for degradation. In

the case of MSH2 and MLH1, for example, it has been shown that a predicted destabilization of more than 3 kcal/mol is sufficient to cause cellular degradation of the proteins (Abildgaard et al., 2019; Nielsen et al., 2017). Similar observations have been recently done in another recent work on different proteins with benign variants featuring predicted changes in stability in the range of 0.9–2.7 kcal/mol (Blaabjerg et al., 2022).

According to the results in Table 2 and the observation above, if we consider folding $\Delta\Delta G$ values higher than 3 kcal/mol, our analyses suggest a number of VUS that could predispose to loss-of-function through destabilization of the protein structure and have a high REVEL score which further support their possible pathogenic impact (i.e., A797T in BLM, I706T in ERCC4, W410C and F603S in FANCA, L329P in FANCF, V180G in MLH1, V606F in MSH2, and G1072D in MSH6). Of note, L329P in FANCE has been suggested to disrupt the stability of the catalytic module of the protein in a previous structural study (Shakeel et al., 2019).

3 | DISCUSSION

We developed RosettaDDGPrediction moved by the need to provide easy and scalable access to Rosetta-based approaches to predict free energy changes in proteins upon mutations. The possibility to perform mutational scans in an efficient and scalable manner allows to have a new systematic and large-scale approach at such data. The fact that other implementations of this process have been released in recent years (e.g., https://github.com/KULL-Centre/PRISM/tree/main/software/rosetta_ddg_pipeline) is a testament to its utility.

RosettaDDGPrediction takes care of the whole process by performing a large number of $\Delta\Delta G$ predictions in an efficient and scalable manner, making a high-throughput calculations with Rosetta accessible, which is helpful for both extensive mutational scans and structured benchmarks.

RosettaDDGPrediction is, to our knowledge, the first wrapper devised to integrate state-of-the-art Rosetta-based protocols for the predictions of free energy changes upon mutation on binding and stability under a uniform framework.

Furthermore, the software checks the success of the runs, aggregates the data in CSV tables that are easy to

FIGURE 6 Trimmed AlphaFold structures of the of the DNA mismatch repair proteins selected for the Case Study 4. Cartoon representation of (a) MLH1_{1–341} and MLH1_{501–756}, (b) MSH2_{1–934} and (c) MSH6_{362–1360}. The proteins are colored according to the AlphaFold2 pLDDT score: very low (orange, pLDDT < 50), low (yellow, $50 < \text{pLDDT} < 70$), confident (light blue, $70 < \text{pLDDT} < 90$), and very high (blue, pLDDT > 90). The C α of the residues found mutated in pediatric cancer patients are shown as spheres and labeled.

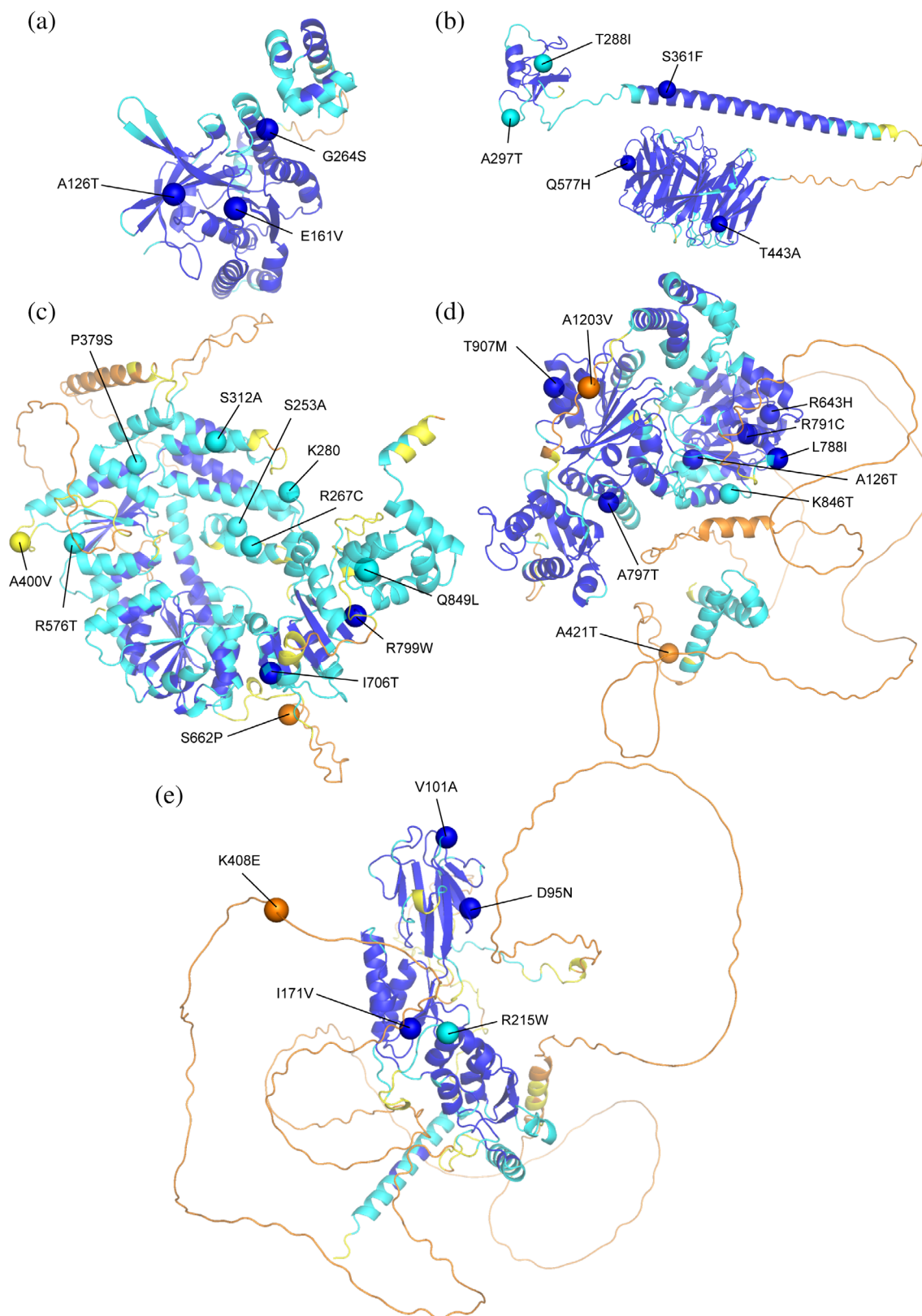


FIGURE 7 Trimmed AlphaFold structures of the proteins promoting the double-strand break (DSB) repair (RAD51C, RFWD3, ERCC4, and NBN) and RECQ helicase (BLM) selected for proteins used for the Case Study 4. Cartoon representation of (a) RAD51C₁₃₋₃₅₀, (b) RFWD3₂₈₄₋₇₇₄, (c) ERCC4₁₂₋₉₁₄, (d) BLM₃₆₈₋₁₂₉₀, and (e) NBN₁₋₇₄₉. The proteins are colored according to the AlphaFold2 pLDDT score: very low (orange, pLDDT < 50), low (yellow, 50 < pLDDT < 70), confident (light blue, 70 < pLDDT < 90), and very high (blue, pLDDT > 90). The Ca of the residues found mutated in pediatric cancer patients are shown as spheres and labeled.

TABLE 2 Summary of predicted $\Delta\Delta G$ s for predicted destabilizing variants in childhood cancer

Variant	ClinVar	Predicted folding $\Delta\Delta G$ (kcal/mol)	REVEL
BLM - L788I	Conflicting interpretations of pathogenicity	2.233	0.503
BLM - A797T	Entry N.A.	3.045	0.893
BLM - K846T	Uncertain significance	1.227	0.136
BLM - Y1024C	Uncertain significance	1.776	0.590
ERCC4 - R267C	Uncertain significance	1.495	0.470
ERCC4 - P379S	Conflicting interpretations of pathogenicity	2.051	0.526
ERCC4 - R576T	Uncertain significance	1.012	0.274
ERCC4 - I706T	Conflicting interpretations of pathogenicity	3.242	0.609
FANCA - F276I	Entry N.A.	1.922	0.160
FANCA - W410C	Entry N.A.	4.169	0.622
FANCA - F603S	Uncertain significance	6.111	0.631
FANCA - A746S	Benign/likely benign	1.696	0.374
FANCA - P1086L	Entry N.A.	1.012	0.775
FANCE - A104P	Entry N.A.	5.691	0.319
FANCE - L326W	Uncertain significance	1.620	0.154
FANCE - M437T	Conflicting interpretations of pathogenicity	1.930	0.134
FANCF - L329P	Uncertain significance	8.170	0.417
FANCF - Y287C	Uncertain significance	2.732	0.195
FANCF - Y274C	Uncertain significance	3.526	0.193
FANCF - L129V	Uncertain significance	1.086	0.069
FANCF - L80V	Entry N.A.	2.001	0.104
FANCG - P545T	Entry N.A.	1.918	0.465
FANCI - I275T	Uncertain significance	3.041	0.22
FANCI - M363T	Entry N.A.	2.593	0.242
FANCI - P471R	Uncertain significance	1.730	0.837
FANCI - M525V	Conflicting interpretations of pathogenicity	2.137	0.487
FANCI - L605F	Benign/Likely benign	3.773	0.238
FANCI - C742S	Benign	1.084	0.075
FANCI - Y923C	Uncertain significance	3.806	0.391
MLH1 - P285S	Uncertain significance	1.373	0.838
MLH1 - K618E	Benign/Likely benign	1.296	0.874
MLH1 - V180G	Uncertain significance	3.847	0.91
MSH2 - N127S	Benign	2.172	0.741
MSH2 - L128V	Conflicting interpretations of pathogenicity	2.309	0.613
MSH2 - L513V	Uncertain significance	2.595	0.829
MSH2 - I577T	Likely benign	2.067	0.928
MSH2 - V606F	Entry N.A.	5.096	0.889
MSH2 - I770V	Conflicting interpretations of pathogenicity	1.043	0.417
MSH6 - L396V	Benign	1.369	0.322
MSH6 - S503C	Entry N.A.	1.288	0.413
MSH6 - V878A	Benign	2.073	0.155
MSH6 - G1072D	Uncertain significance	6.349	0.623
MSH6 - V1253E	Uncertain significance	2.986	0.952

(Continues)

TABLE 2 (Continued)

Variant	ClinVar	Predicted folding $\Delta\Delta G$ (kcal/mol)	REVEL
NBN - D95N	Conflicting interpretations of pathogenicity	1.448	0.583
NBN - I171V	Conflicting interpretations of pathogenicity	1.133	0.398
RFWD3 - Q577H	Entry N.A.	1.469	0.162

Note: We did not report RAD51C and FANCL in the table since all the variants analyzed here for these proteins were predicted with neutral effects for stability. We reported the full list of VUS in Table S2. Here, we included those variants that are predicted destabilizing by both Rosetta- and FoldX-based estimates, using a threshold of folding $\Delta\Delta G$ of 1 kcal/mol. This is a threshold often used to discuss the effects of mutations on structural stability applying Rosetta- or FoldX-based methods. We observed that there are variants that resulted in changes of $\Delta\Delta G$ above the threshold but with a benign Clinvar classification. This suggests that a $\Delta\Delta G$ threshold of approximately 2–3 kcal/mol could be more suited to pinpoint pathogenic variants. NA indicates “not available.”

mine, and generates visual reports. As these steps are independent, the aggregation and visualization tools can be used on different datasets. In addition, we support additional output formats compatible with the MutateX plotting scheme (Tiberti et al., 2022). At the same time, raw or aggregated data can be easily manipulated externally. RosettaDDGPrediction also devotes particular attention to ensuring technical reproducibility by being controlled through configuration files. Further developments of RosettaDDGPrediction will focus on integrating its functionalities within MutateX, to provide a method-agnostic container to perform and collect high-throughput mutational scans in a reproducible, automatized, and sustainable manner.

In this context, the performances of RosettaDDGPrediction and MutateX are only as good as those of the Rosetta- and FoldX-based methods that they incorporate. Indeed, Rosetta-based protocols implemented so far rely on different sampling methods to obtain models of the mutant variant structures and on scoring the resulting structures via knowledge-based energy functions to predict changes in the folding and binding free energy upon mutation (O'Meara et al., 2015; Park et al., 2016). However, more rigorous strategies are available to predict both the effect of mutations on the folding free energy and the binding free energy (Benedix et al., 2009; Kumari et al., 2014b; Seeliger & de Groot, 2010; Siebenmorgen & Zacharias, 2020). For example, approaches leveraging enhanced sampling along reaction coordinates designed to study binding and unbinding events are available (Bertazzo et al., 2021; Raniolo & Limongelli, 2020; Wingbermühle & Schäfer, 2020).

The time and computational resources needed by these methods still prevent their usage for investigations going beyond a few mutations. In these contexts, which include, for instance, saturation mutagenesis scans, Rosetta- and FoldX-based protocols represent a good trade-off between accuracy and speed.

Nevertheless, Rosetta still presents a challenge when noncanonical residue types are considered. Indeed, while

most noncanonical amino acids are supported, mutations to phosphorylated residues cannot be performed in either protocol to predict free energy changes. For this reason, including strategies circumventing this issue would greatly expand the application of RosettaDDGPrediction.

Furthermore, a milestone in structural bioinformatics has been reached lately, with the release of AlphaFold2 and its outstanding performance in the CASP14 challenge (Jumper et al., 2021). Originally developed to solve the long-standing protein folding problem, AlphaFold2 has already seen many spin-off studies to assess its potential (Evans et al., 2022; Porta-Pardo et al., 2022; Robertson et al., 2021; Ruff & Pappu, 2021; Tsuban et al., 2022). So far, evidence suggests that AlphaFold2 cannot effectively predict changes in folding free energy upon mutation (Buel & Walters, 2022; McBride et al., 2022; Pak et al., 2021). However, more studies are needed to explore this possibility fully.

Our wrappers have been devised to be inherently extensible. As stated earlier, a long-term perspective may include transforming them into a more general platform for structure-based methods to predict free energy changes upon mutation based on freely accessible, open-source software. This will also allow us to support other energy functions or schemes for free energy calculations, as well as to include support to transmembrane proteins including protocols as the one developed by Tiemann et al. (Tiemann et al., n.d.). Moreover, the results obtained here and in the original publication (Blaabjerg et al., 2022) with the deep-learning method RaSP pinpoints this approach as an additional candidate to include in a unified framework together with the support to FoldX- and Rosetta-based calculations.

The efforts of centralizing the development of software for in silico deep mutational scans using free energy functions will help to move a step forward toward a unified framework for high-throughput structure-based calculations of free energy changes upon mutation.

4 | METHODS

The data and documentation on the case studies are reported in the OSF repository, <https://osf.io/84kwe/> (10.17605/OSF.IO/84KWE).

4.1 | Case Study 1

The ThermoMut database (Xavier et al., 2021) was downloaded on April 22, 2022, as a JSON file. We processed the database following four main steps: (i) For each reported protein, we retained only the entries including single mutations with an experimental value of $\Delta\Delta G$ discarding entries with multiple mutations with a combined $\Delta\Delta G$; (ii) we reversed the sign of all the $\Delta\Delta G$ values to fit the sign provided by the outputs of RosettaDDGPrediction, (iii) we retained information on pH values and experimental methods as metadata, and (iv) we removed protein entries for which <10 mutations were reported. Upon processing, we identified 133 proteins. We then searched for 3D structures available for each protein in the Protein Data Bank. In this step, we retained matches that covered at least one mutation of interest. We retained only protein structures in their free state (i.e., not in a complex with other interactors) for a total of 121 target proteins, effectively removing 12 proteins where no structure or free state was found. We selected two enzymes that included a large number of amino acid substitutions with structural coverage (i.e., ENLYS and NUC as represented by the PDB structures 1P7S; Mooers et al., 2003 and 1EY0; Chen et al., 2000, and two human proteins of interest in health and disease; p53 and FKBP1A as represented by the PDB structures 2XWR; Natan et al., 2011 and 2PPN; Szep et al., 2009 as case studies for this work). All are used as simplistic monomeric structures and chosen based on the coverage, quality, and lack of interactors. The experimental values obtained in an acidic or alkaline experimental setting (pH < 6 and pH > 8) were excluded, as the *ref2015* Rosetta energy function (Cartesian space version) is simulating an environment at pH 7.

This leaves 845 observations across the four proteins for pH values 6, 7, and 8 and three methodologies, two chemically denaturant-induced protein unfolding experimental protocols, guanidine hydrochloride (GdnHCl), Urea Denaturation (Urea), and one thermal denaturation protocol (Thermal). We modeled the experimental and predicted values using a simple linear model, analyzed the contribution of secondary structures, and built a generalized additive model, thereby defining the limitations of the model. Furthermore, we constructed a confusion

matrix based on the thresholds $\Delta\Delta G < -1$ kcal/mol for the stabilizing group, -1 kcal/mol $> \Delta\Delta G < 1$ kcal/mol for the neutral group, and $\Delta\Delta G > 1$ kcal/mol for the destabilizing group. Calculations were carried out with Rosetta 3.12.

In addition, we applied the rapid protein stability prediction (RaSP) (Blaabjerg et al., 2022) for comparison. We used the Colab version of the software (<https://colab.research.google.com/github/KULL-Centre/papers/blob/main/2022/ML-ddG-Blaabjerg-et-al/RaSPLab.ipynb#scrollTo=Z8nUmHI5rgjy>). We used the resulting *score_ml* as a proxy for the $\Delta\Delta G$ values. We applied a simple linear model to compare RaSP predicted $\Delta\Delta G$ values to the experimentally derived $\Delta\Delta G$ s and to the $\Delta\Delta G$ values predicted by the Rosetta-based protocols implemented in RosettaDDGPrediction. Additionally, to explore the model limitations, we built a generalized additive model.

4.2 | Case Study 2

We started from the phospho-regulated LIRs reported in our previous review article (Sora et al., 2020) and other literature search, and, for each of them, we verified if a complex with one of the LC3/GABARAP family members was available to use as starting structure for the mutational scan. We retained for the analyses the following complexes: LC3B:FUNCD1 (PDB entry 2N9X; Kuang et al., 2016) and GABARAP: PIK3C3 (PDB entry 6HOG; Birgisdottir et al., 2019).

We reconstructed missing coordinates in the structures using MODELER version 10.1 (Webb & Sali, 2016).

We used the *flexddg* protocol, as implemented in RosettaDDGPrediction, with the *talaris2014* energy function and Rosetta 3.12. Rosetta energy units (REUs) were converted to kilocalorie per mole with the conversion factors provided for this energy function (Park et al., 2016). We modeled the phosphorylated residues using phosphomimetic mutations to aspartic acid and glutamic acid for each phosphosite and included also tryptophan for phospho-tyrosine to identify possible effects due to steric hindrance. In the calculations, we used 35,000 backrub trials and an absolute score threshold for minimization convergence of 1 REUs. We generated an ensemble of 35 structures for each mutant variant and calculated the average $\Delta\Delta G$ s and the standard deviation among the individual binding free energies.

For the MutateX runs, we calculated changes in binding free energy using the Build Model and Analyze Complex functions of FoldX5 suite and averaging over five runs. The standard deviation for the $\Delta\Delta G$ values predicted with RosettaDDGPrediction and MutateX have been calculated using the GraphPad Prism 9 software.

We derived the $\Delta\Delta G$ values for each experimental K_d by using the following Gibbs free energy and constant equilibrium equations:

$$\Delta\Delta G = -RT\ln K_{\text{eq}}$$

$$K_{\text{eq}} = \frac{1}{K_d}$$

We combined the equations in order to compute the ΔG for the mutant and the WT as follows:

$$\Delta G = RT\ln K_d$$

The $\Delta\Delta G$ has been calculated subtracting the ΔG of the WT to the ΔG of the mutant:

$$\Delta\Delta G = RT\ln K_{d_mutant} - RT\ln K_{d_WT}$$

$$\Delta\Delta G = RT\ln \frac{K_{d_mutant}}{K_{d_WT}}$$

The standard deviations associated with the K_d measurement have been propagated for the $\Delta\Delta G$ calculations by using Uncertainty Calculator (<https://uncertaintycalculator.com>) and Propagation-of-Uncertainty-Calculator (<https://nicoco007.github.io/Propagation-of-Uncertainty-Calculator/>).

4.3 | Case Study 3

We retrieved experimental $\Delta\Delta G$ values from point mutations of the p53 DNA-binding domain from ThermoMutDB. Since ThermoMutDB stores $\Delta\Delta G_u$ values, they were converted to $\Delta\Delta G_f$ by changing the sign to make them easily comparable with Rosetta output values. A total of 31 mutations were selected, and when multiple experimental values were reported for the same variant, the average of their $\Delta\Delta G_f$ was used.

We used two different structures. The first one consists of the x-ray crystallography of the PDB entry 2XWR, with a resolution of 1.68 Å, which covers the DNA-binding domain from residues 91 to 289 and includes the zinc ion. The water molecules were removed using PyMOL (<http://www.pymol.org/pymol>). We also used the model from the AlphaFold2 database, which was trimmed to cover the same residues as the experimental x-ray structure, from 91 to 289. The missing zinc ion was added using PyMOL, identifying its coordinates by rigid body superimposition with the original structure. Before, we verified that the residues, which coordinate the zinc ion (C176, H179, C238, and C242), had a good alignment

and similar rotamer conformations between the two structures.

For the $\Delta\Delta G$ predictions, we mostly used the *cartddg* protocol with the *ref2015* and *talaris2014* scoring functions, each with 3 and 10 sampling runs and Rosetta 3.12. We also used the *cartddg2020* protocol on both structures, but only using the *ref2015* scoring function and three runs.

The performance was measured by Pearson correlation coefficient, MAE, and area under the curve (AUC) of the ROC curve. For the ROC curve, we used a threshold of 1.2 kcal/mol for the ThermoMutDB averaged values, meaning that mutations associated with a free energy change higher than 1.2 kcal/mol were considered destabilizing, according to the threshold selection proposed for p53 in our previous study (Degn et al., 2022). In comparison, mutations associated with a free energy change lower than the threshold were classified as nondestabilizing. The MAE was calculated using the following equation:

$$\xi = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$$

Here, y_i is the experimental values and \hat{y}_i is the predicted values.

4.4 | Case Study 4

We retrieved relevant VUSs from germline WGS data on 566 children with cancer included in an in-house dataset and published, in part, in a previous study (Byrjalsen et al., 2020). In addition, we analyzed a dataset including 566 samples from Danish children with different cancer types and whole genome sequencing data. An illustration of the workflows for analyzing the sequencing data and annotating the called variants is provided in Figures S4 and S5. The sequencing data have been processed with a pipeline based on Sentieon using the reference genome reference genome Grch38/hg38 from GATK resource bundle (ELELAB/sentieon_wgs_pipeline). Reads were aligned with BWA-MEM to the reference genome, and duplicate reads were removed. Reads were realigned around indels, and we applied Base Quality Score Recalibration together with the Haplotype algorithm for variant calling (equivalent to the GATK Haplotype; van der Auwera & O'Connor, n.d.). Then, as suggested by GATK best practices, we used Variant Quality Score Recalibration, which is an advanced filtering technique used on the variant call set that models the technical profile of variants in a training set using machine learning and filters

out potential artifacts from the callset. The filtered variants were uploaded to an in-house MySQL database where we linked them with information about genomic context (ENSEMBL v95), ENSEMBL consequences, deleteriousness-scores (CADD 1.6, REVEL, SIFT, PolyPhen) and variant frequency in the healthy population (gnomAD v3; Karczewski et al., 2020) based on their genomic position and alternate allele (Figure S3). To this purpose, we annotated our variants with the gnomAD popmax allele frequency, that is, the maximum allele frequency in all the continental populations (Gudmundsson et al., 2022), along with the allele frequency of the non-Finnish European population.

In particular, we annotated the REVEL score (Ioannidis et al., 2016) associated with the genomic mutation by using the publicly available dataset of pre-computed scores, by matching genomic coordinates, annotated transcript for the mutation and alternate nucleotide. We could not annotate a REVEL score for four of the identified variants, most likely as they were not missense. Two of them caused early translation termination by introducing a stop codon in the reference BRCA2 transcript (13:g.32337185A>T and 13:g.32398489A>T, corresponding to p.Lys944* and p.Lys3326* at protein level). The other two (2:g.47607407G>A and 2:g.47607446G>A, corresponding to p.Arg923Gln and p.Gly936Asp in MSH2 at protein level for the reference transcript) were annotated as both missense and nonsense-mediated decay in our dataset, meaning they are annotated as nonsense-mediated decay for at least some of the MSH2 transcripts, and this is probably the reason they were not available in the REVEL database.

We retained, as VUS to investigate, those variants located in the coding regions and found with an allele frequency in the non-Finnish European population lower than 1% in gnomAD v3 (build 38) as a proxy for a healthy population. This threshold has been selected according to the guidelines for clinical VUS studies (Richards et al., 2015). An illustration of the workflow for analyzing the sequencing data is provided in Figure S4.

We searched each variant in the selected 14 genes for the study in ClinVar (Landrum et al., 2014; Landrum et al., 2020) and retrieved annotations on them to verify if they are VUS, variants with conflicting evidence, or not reported yet in the database. To select the proteins and variants that can be investigated with RosettaDDGPrediction, we then searched in the AlphaFold2 database (Varadi, Anyango, Deshpande, et al., 2022) for the corresponding protein structures and retained those that had structural coverage for the variants in regions with high confidence (pLDDT > 70) trimming the N-terminal or C-terminal tails. For MLH1, we used the structure of the

two protein domains, for the other proteins, we retained cases in which the pLDDT score was low but located in loops that connect structured regions of folded domains. These regions are often very flexible in a protein structure, and it is thus expected that they could have a lower pLDDT score. We analyzed 14 proteins and 126 variants in total.

We excluded mutations either not covered by our trimmed models or derived from an isoform different from the one available in the AlphaFold2 database. Concerning MSH2, we did not analyze G936D since our isoform had 934 residues, while R293Q refers to the A0A2R8Y-G02_HUMAN isoform (Hillier et al., 2005). In the case of MSH6, T1125M was removed since derived from the A0A494C0M1_HUMAN transcript (Hillier et al., 2005). Furthermore, the following seven variants found in FANCL were also disregarded: S356N, S356N, G322V, F257C, T229A, I199V, and V181I. These variants were generated from FANCL isoform 2 (ENST00000402135.8, Q9NW38-2, 380aa), which did not match the AlphaFold model for FANCL (ENST00000233741.9, Q9NW38, 375aa).

AUTHOR CONTRIBUTIONS

Valentina Sora: Conceptualization (equal); methodology (equal); software (equal); visualization (supporting); writing – original draft (equal); writing – review and editing (equal). **Adrian Otamendi Laspiur:** Conceptualization (equal); data curation (equal); formal analysis (equal); investigation (supporting); methodology (equal); software (supporting); validation (supporting); visualization (equal); writing – original draft (supporting); writing – review and editing (supporting). **Kristine Degn:** Conceptualization (supporting); data curation (supporting); formal analysis (supporting); investigation (supporting); methodology (supporting); supervision (supporting); visualization (equal); writing – original draft (supporting). **Matteo Arnaudi:** Conceptualization (supporting); formal analysis (supporting); investigation (supporting); visualization (equal); writing – original draft (supporting). **Mattia Utichi:** Conceptualization (supporting); formal analysis (supporting); investigation (supporting); visualization (supporting); writing – original draft (supporting). **Ludovica Beltrame:** Formal analysis (supporting); investigation (supporting); validation (supporting); visualization (supporting); writing – original draft (supporting). **Dayana De Menezes:** Formal analysis (supporting); investigation (supporting); visualization (supporting); writing – original draft (supporting). **Matteo Orlandi:** Data curation (supporting); formal analysis (supporting). **Ulrik Kristoffer Stoltze:** Data curation (supporting); validation (supporting); writing – review and editing (supporting). **Olga Rigina:** Data curation

(supporting); software (supporting). **Peter Wad Sackett:** Resources (supporting); software (supporting). **Karin Wadt:** Data curation (supporting); funding acquisition (supporting); supervision (supporting); writing – review and editing (supporting). **Kjeld Schmiegelow:** Funding acquisition (equal); supervision (supporting); writing – review and editing (supporting). **Matteo Tiberti:** Conceptualization (supporting); investigation (supporting); methodology (supporting); supervision (supporting); validation (equal); visualization (supporting); writing – original draft (supporting); writing – review and editing (supporting). **Elena Papaleo:** Conceptualization (lead); data curation (equal); formal analysis (supporting); funding acquisition (lead); investigation (equal); methodology (supporting); project administration (lead); resources (lead); supervision (lead); validation (equal); visualization (supporting); writing – original draft (lead); writing – review and editing (lead).

ACKNOWLEDGMENTS

Our research has been supported by Carlsberg Foundation Distinguished Fellowship (CF18-0314), Danmarks Grundforskningsfond (DNRF125), Hartmanns Fond (R241-A33877), LEO Foundation (LF17006), and Novo-Nordisk Fonden Bioscience and Basic Biomedicine (NNF20OC0065262). This work is part of Interregional Childhood Oncology Precision Medicine Exploration (iCOPE), a cross-Oresund collaboration between University Hospital Copenhagen, Rigshospitalet, Lund University, Region Skåne, and Technical University Denmark (DTU), supported by the European Regional Development Fund. This work is also part of the Danish nationwide research program Childhood Oncology Network Targeting Research, Organisation & Life expectancy (CONTROL) and supported by the Danish Cancer Society (R-257-A14720) and the Danish Childhood Cancer Foundation (2019-5934 and 2020-5769).

DATA AVAILABILITY STATEMENT

The software and data presented in this work are available at <https://github.com/ELELAB/RosettaDDGPrediction>, <https://osf.io/84kwe/> and https://github.com/ELELAB/sentieon_wgs_pipeline.

ORCID

Elena Papaleo  <https://orcid.org/0000-0002-7376-5894>

REFERENCES

- Abildgaard AB, Stein A, Nielsen SV, Schultz-Knudsen K, Papaleo E, Shrikhande A, et al. Computational and cellular studies reveal structural destabilization and degradation of MLH1 variants in Lynch syndrome. *Elife*. 2019;8:e49138. <https://doi.org/10.7554/eLife.49138.001>
- Adachi D, Oda T, Yagasaki H, Nakasato K, Taniguchi T, D'Andrea AD, et al. Heterogeneous activation of the Fanconi anemia pathway by patient-derived FANCA mutants. *Hum Mol Genet*. 2002;11:3125–34.
- Akdel M, Pires DEV, Pardo EP, Jänes J, Zalevsky AO, Mészáros B, et al. A structural biology community assessment of AlphaFold2 applications. *Nat Struct Mol Biol*. 2022;29:1056–67. <https://doi.org/10.1038/s41594-022-00849-w>
- Anderson CL, Munawar S, Reilly L, Kamp TJ, January CT, Delisle BP, et al. How functional genomics can keep pace with VUS identification. *Front Cardiovasc Med*. 2022;9:90043.
- Arnaudi M, Beltrame L, Degn K, Utichi M, Pettenella A, Scrima S, et al. MAVISp: multi-layered assessment of Variants by structure for proteins. *bioRxiv*. 2022. <https://doi.org/10.1101/2022.10.22.513328>
- Bæk KT, Kepp KP. Data set and fitting dependencies when estimating protein mutant stability: toward simple, balanced, and interpretable models. *J Comput Chem*. 2022;43:504–18.
- Barlow KA, Ó Conchúir S, Thompson S, Suresh P, Lucas JE, Heinonen M, et al. Flex ddG: Rosetta Ensemble-based estimation of changes in protein–protein binding affinity upon mutation. *J Phys Chem B*. 2018;122:5389–99.
- Benedix A, Becker CM, de Groot BL, Cafilisch A, Böckmann RA. Predicting free energy changes using structural ensembles. *Nat Methods*. 2009;6:3–4.
- Bertazzo M, Gobbo D, Decherchi S, Cavalli A. Machine learning and enhanced sampling simulations for computing the potential of mean force and standard binding free energy. *J Chem Theory Comput*. 2021;17:5287–300.
- Birgisdottir ÁB, Mouilleron S, Bhujabal Z, Wirth M, Sjøttem E, Evjen G, et al. Members of the autophagy class III phosphatidylinositol 3-kinase complex I interact with GABARAP and GABARAPL1 via LIR motifs. *Autophagy*. 2019;15:1333–55.
- Blaabjerg LM, Kassem MM, Good LL, Jonsson N, Cagiada M, Johansson KE, et al. Rapid protein stability prediction using deep learning representations. *bioRxiv*. 2022. <https://doi.org/10.1101/2022.07.14.500157>
- Buel GR, Walters KJ. Can AlphaFold2 predict the impact of missense mutations on structure? *Nat Struct Mol Biol*. 2022;29(1):1–2.
- Buß O, Rudat J, Ochsenreither K. FoldX as protein engineering tool: better than random based approaches? *Comput Struct Biotechnol J*. 2018;16:25–33.
- Byrjalsen A, Hansen TVO, Stoltze UK, Mehrjouy MM, Barnkob NM, Hjalgrim LL, 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:e1009231.
- Cagiada M, Johansson KE, Valanciute A, Nielsen SV, Hartmann-Petersen R, Yang JJ, et al. Understanding the origins of loss of protein function by analyzing the effects of thousands of variants on activity and abundance. *Mol Biol Evol*. 2021;38:3235–46.
- Chen J, Lu Z, Sakon J, Stites WE. Increasing the thermostability of staphylococcal nuclease: implications for the origin of protein thermostability. *J Mol Biol*. 2000;303:125–30.

- Davey NE. The functional importance of structure in unstructured protein regions. *Curr Opin Struct Biol.* 2019;56:155–63.
- Davey NE, van Roey K, Weatheritt RJ, Toedt G, Uyar B, Altenberg B, et al. Attributes of short linear motifs. *Mol Biosyst.* 2011;8:268–81.
- Degn K, Beltrame L, Dahl Hede F, Sora V, Nicolaci V, Vabistsevits M, et al. Cancer-related mutations with local or long-range effects on an allosteric loop of p53. *J Mol Biol.* 2022; 434:167663.
- Delgado J, Radusky LG, Cianferoni D, Serrano L. FoldX 5.0: working with RNA, small molecules and a new graphical interface. *Bioinformatics.* 2019;35:1–2. <https://doi.org/10.1093/bioinformatics/btz184>
- Evans R, O'Neill M, Pritzel A, Senior NAA, Green T, Židek A, et al. Protein complex prediction with AlphaFold-Multimer. *bioRxiv.* 2022. <https://doi.org/10.1101/2021.10.04.463034>
- Fas BA, Maiani E, Sora V, Kumar M, Mashkooor M, Lambrugh M, et al. The conformational and mutational landscape of the ubiquitin-like marker for autophagosome formation in cancer. *Autophagy.* 2020;17:1–24.
- Federici G, Soddu S. Variants of uncertain significance in the era of high-throughput genome sequencing: a lesson from breast and ovary cancers. *J Exp Clin Cancer Res.* 2020;39:1–12.
- Fierheller CT, Guitton-Sert L, Alenezi WM, Revil T, Oros KK, Gao Y, et al. A functionally impaired missense variant identified in French Canadian families implicates FANCI as a candidate ovarian cancer-predisposing gene. *Genome Med.* 2021;13: 1–26.
- Frenz B, Lewis SM, King I, DiMaio F, Park H, Song Y. Prediction of protein mutational free energy: benchmark and sampling improvements increase classification accuracy. *Front Bioeng Biotechnol.* 2020;8:558247.
- Gasparini M, Starita L, Shendure J. The power of multiplexed functional analysis of genetic variants. *Nat Protoc.* 2016;11(10): 1782–7.
- Geng C, Xue LC, Roel-Touris J, Bonvin AMJJ. Finding the $\Delta\Delta G$ spot: are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it? *Wiley Interdiscip Rev Comput Mol Sci.* 2019;9:e1410.
- Gerasimavicius L, Liu X, Marsh JA. Identification of pathogenic missense mutations using protein stability predictors. *Sci Rep.* 2020;10(1):1–10.
- Gerasimavicius L, Livesey BJ, Marsh JA. Loss-of-function, gain-of-function and dominant-negative mutations have profoundly different effects on protein structure. *Nature Commun.* 2022; 13(1):1–15.
- Gudmundsson S, Singer-Berk M, Watts NA, Phu W, Goodrich JK, Solomonson M, et al. Variant interpretation using population databases: lessons from gnomAD. *Hum Mutat.* 2022;43:1012–30. <https://doi.org/10.1002/humu.24309>
- Hillier LDW, Graves TA, Fulton RS, Fulton LA, Pepin KH, Minx P, et al. Generation and annotation of the DNA sequences of human chromosomes 2 and 4. *Nature.* 2005;434: 724–31.
- Hoie MH, Cagiada M, Beck Frederiksen AH, Stein A, Lindorff-Larsen K. Predicting and interpreting large-scale mutagenesis data using analyses of protein stability and conservation. *Cell Rep.* 2022;38:110207.
- Ioannidis NM, Rothstein JH, Pejaver V, Middha S, McDonnell SK, Baheti S, et al. REVEL: an ensemble method for predicting the pathogenicity of rare missense variants. *Am J Hum Genet.* 2016;99:877–85.
- Jepsen MM, Fowler DM, Hartmann-Petersen R, Stein A, Lindorff-Larsen K. Classifying disease-associated variants using measures of protein activity and stability. *Protein Homeostasis Diseases;* 2020. p. 91–107. <https://doi.org/10.1016/B978-0-12-819132-3.00005-1>
- Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature.* 2021;596:583–9.
- Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature.* 2020;581:434–43.
- Katarina Sooe Tiemann J, Zschach H, Lindorff-Larsen K, Stein A. Interpreting the molecular mechanisms of disease variants in human membrane proteins. *bioRxiv.* <https://doi.org/10.1101/2022.07.12.499731>
- Kellogg EH, Leaver-Fay A, Baker D. Role of conformational sampling in computing mutation-induced changes in protein structure and stability. *Proteins.* 2011;79:830–8.
- Kortemme T, Baker D. A simple physical model for binding energy hot spots in protein–protein complexes. *Proc Natl Acad Sci U S A.* 2002;99(14):116–14121.
- Kuang Y, Ma K, Zhou C, Ding P, Zhu Y, Chen Q, et al. Structural basis for the phosphorylation of FUNDC1 LIR as a molecular switch of mitophagy. *Autophagy.* 2016;12:2363–73.
- Kumari R, Kumar R, Lynn A. G-mmpbsa – a GROMACS tool for high-throughput MM-PBSA calculations. *J Chem Inf Model.* 2014;54:1951–62.
- Landrum MJ, Chitipiralla S, Brown GR, Chen C, Gu B, Hart J, et al. ClinVar: improvements to accessing data. *Nucleic Acids Res.* 2020;48:D835–44.
- Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* 2014;42:D980–5.
- Lv M, Wang C, Li F, Peng J, Wen B, Gong Q, et al. Structural insights into the recognition of phosphorylated FUNDC1 by LC3B in mitophagy. *Protein Cell.* 2017;8:25–38.
- McBride JM, Polev K, Reinharz V, Grzybowski BA, Tlustý T. AlphaFold2 can predict structural and phenotypic effects of single mutations. *bioRxiv.* 2022. <https://doi.org/10.1101/2022.04.14.488301>
- Mooers BHM, Datta D, Baase WA, Zollars ES, Mayo SL, Matthews BW. Repacking the core of T4 lysozyme by automated design. *J Mol Biol.* 2003;332:741–56.
- Natan E, Baloglu C, Pagel K, Freund SMV, Morgner N, Robinson CV, et al. Interaction of the p53 DNA-binding domain with its n-terminal extension modulates the stability of the p53 tetramer. *J Mol Biol.* 2011;409:358–68.
- Nielsen SV, Stein A, Dinitzen AB, Papaleo E, Tatham MH, Poulsen EG, et al. Predicting the impact of Lynch syndrome-causing missense mutations from structural calculations. *PLoS Genet.* 2017;13:e1006739.
- Nygaard M, Terkelsen T, Vidas Olsen A, Sora V, Salamanca Vitoria J, Rizza F, et al. The mutational landscape of the

- oncogenic MZF1 SCAN domain in cancer. *Front Mol Biosci.* 2016;3:1–18.
- Ollodart AR, Yeh CLC, Miller AW, Shirts BH, Gordon AS, Dunham MJ. Multiplexing mutation rate assessment: determining pathogenicity of Msh2 variants in *Saccharomyces cerevisiae*. *Genetics.* 2021;218:iyab058.
- O'Meara MJ, Leaver-Fay A, Tyka MD, Stein A, Houlihan K, DiMaio F, et al. Combined covalent-electrostatic model of hydrogen bonding improves structure prediction with Rosetta. *J Chem Theory Comput.* 2015;11:609–22.
- Pak MA, Markhieva KA, Novikova MS, Petrov DS, Vorobyev IS, Maksimova ES, et al. Using AlphaFold to predict the impact of single mutations on protein stability and function. *bioRxiv.* 2021. <https://doi.org/10.1101/2021.09.19.460937>
- Pancotti C, Benevenuto S, Birolo G, Alberini V, Repetto V, Sanavia T, et al. Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. *Brief Bioinform.* 2022;23:bbab555.
- Park H, Bradley P, Greisen P Jr, Liu Y, Mulligan VK, Kim DE, et al. Simultaneous optimization of biomolecular energy functions on features from small molecules and macromolecules. *J Chem Theory Comput.* 2016;12:6201–12.
- Pires DEV, Rodrigues CHM, Albanaz ATS, Myung MKY, Xavier J, Michanetz E-M, et al. Exploring protein supersecondary structure through changes in protein folding, stability, and flexibility. *Methods Mol Biol.* 2019;1958:173–85.
- Porta-Pardo E, Ruiz-Serra V, Valentini S, Valencia A. The structural coverage of the human proteome before and after AlphaFold. *PLoS Comput Biol.* 2022;18:e1009818.
- Raniolo S, Limongelli V. Ligand binding free-energy calculations with funnel metadynamics. *Nat Protoc.* 2020;15(9):2837–66.
- Richards S, Aziz N, Bale S, Bick D, das S, Gastier-Foster J, et al. Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med.* 2015;17:405–24.
- Robertson AJ, Courtney JM, Shen Y, Ying J, Bax A. Concordance of X-ray and AlphaFold2 models of SARS-CoV-2 Main protease with residual dipolar couplings measured in solution. *J Am Chem Soc.* 2021;143(19):306–19310.
- Ruff KM, Pappu R v. AlphaFold and implications for intrinsically disordered proteins. *J Mol Biol.* 2021;433:167208.
- Scheller R, Stein A, Nielsen SV, Marin FI, Gerdes AM, Marco M, et al. Toward mechanistic models for genotype–phenotype correlations in phenylketonuria using protein stability calculations. *Hum Mutat.* 2019;40:444–57. <https://doi.org/10.1002/humu.23707>
- Schymkowitz J, Borg J, Stricher F, Nys R, Rousseau F, Serrano L. The FoldX web server: an online force field. *Nucleic Acids Res.* 2005;33:W382–8.
- Seeliger D, de Groot BL. Protein thermostability calculations using alchemical free energy simulations. *Biophys J.* 2010;98:2309–16.
- Shakeel S, Rajendra E, Alcón P, O'Reilly F, Chorev DS, Maslen S, et al. Structure of the Fanconi anaemia monoubiquitin ligase complex. *Nature.* 2019;575(7781):234–7.
- Siebenmorgen T, Zacharias M. Computational prediction of protein–protein binding affinities. *Wiley Interdiscip Rev Comput Mol Sci.* 2020;10:e1448.
- Smith CA, Kortemme T. Backrub-like backbone simulation recapitulates natural protein conformational variability and improves mutant side-chain prediction. *J Mol Biol.* 2008;380:742–56.
- Somyajit K, Subramanya S, Nagaraju G. RAD51C: a novel cancer susceptibility gene is linked to Fanconi anemia and breast cancer. *Carcinogenesis.* 2010;31:2031–8.
- Sondka Z, Bamford S, Cole CG, Ward SA, Dunham I, Forbes SA. The COSMIC cancer gene census: describing genetic dysfunction across all human cancers. *Nature Rev Cancer.* 2018;18(11):696–705.
- Sora V, Kumar M, Maiani E, Lambrugh M, Tiberti M, Papaleo E. Structure and dynamics in the ATG8 family from experimental to computational techniques. *Front Cell Dev Biol.* 2020;8:420.
- Stein A, Fowler DM, Hartmann-Petersen R, Lindorff-Larsen K. Biophysical and mechanistic models for disease-causing protein variants. *Trends Biochem Sci.* 2019;44:575–88.
- Szep S, Park S, Boder ET, van Duyne GD, Saven JG. Structural coupling between FKBP12 and buried water. *Proteins.* 2009;74:603–11.
- Tiberti M, Terkelsen T, Degn K, Beltrame L, Cremers TC, da Piedade I, et al. MutateX: an automated pipeline for in silico saturation mutagenesis of protein structures and structural ensembles. *Brief Bioinform.* 2022;23:bbac074.
- Tsaban T, Varga JK, Avraham O, Ben-Aharon Z, Khramushin A, Schueler-Furman O. Harnessing protein folding neural networks for peptide-protein docking. *Nat Commun.* 2022;13:176.
- Usmanova DR, Bogatyreva NS, Ariño Bernad J, Eremina AA, Gorshkova AA, Kanevskiy GM, et al. Self-consistency test reveals systematic bias in programs for prediction change of stability upon mutation. *Bioinformatics.* 2018;34:3653–8.
- Valanciute A, Nygaard L, Zschach H, Jepsen MM, Lindorff-Larsen K, Stein A. Accurate protein stability predictions from homology models. *Comput Struct Biotech J.* 2022;21:66–73.
- van der Auwera G, O'Connor B. *Genomics in the Cloud.* O'Reilly Media, Inc. 2020, p. 300.
- van Roey K, Uyar B, Weatheritt RJ, Dinkel H, Seiler M, Budd A, et al. Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem Rev.* 2014;114:6733–78.
- Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* 2022;50:D439–44.
- Varadi M, Anyango S, Armstrong D, Berrisford J, Choudhary P, Deshpande M, et al. PDBE-KB: collaboratively defining the biological context of structural data. *Nucleic Acids Res.* 2022;50:D534–42.
- Wang LH, Wu CF, Rajasekaran N, Shin YK. Loss of tumor suppressor gene function in human cancer: an overview. *Cell Physiol Biochem.* 2018;51:2647–93.
- Webb B, Sali A. Comparative protein structure modeling using MODELLER. *Curr Protoc Bioinformatics.* 2016;54:5.6.1–5.6.37.
- Weile J, Roth FP. Multiplexed assays of variant effects contribute to a growing genotype–phenotype atlas. *Hum Genet.* 2018;137(9):665–78.

- Wilkes DC, Sailer V, Xue H, Cheng H, Collins CC, Gleave M, et al. A germline FANCA alteration that is associated with increased sensitivity to DNA damaging agents. *Cold Spring Harb Mol Case Stud.* 2017;3:a001487.
- Wingbermhle S, Schäfer L v. Capturing the flexibility of a protein-ligand complex: binding free energies from different enhanced sampling techniques. *J Chem Theory Comput.* 2020;16:4615–30.
- Xavier JS, Nguyen TB, Karmarkar M, Portelli S, Rezende PM, Velloso JPL, et al. ThermoMutDB: a thermodynamic database for missense mutations. *Nucleic Acids Res.* 2021;49:D475–9.
- Zhang X, Lu X, Akhter S, Georgescu MM, Legerski RJ. FANCI is a negative regulator of Akt activation. *Cell Cycle.* 2016;15:1134–43.

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Sora V, Laspiur AO, Degn K, Arnaudi M, Utichi M, Beltrame L, et al. RosettaDDGPrediction for high-throughput mutational scans: From stability to binding. *Protein Science.* 2023;32(1):e4527. <https://doi.org/10.1002/pro.4527>

Protocol	Variable	Class: destabilizing	Class: neutral	Class: stabilizing
ref2015, cartesian	Sensitivity	0.8336	0.6109	0.173913
	Specificity	0.6879	0.8370	0.961443
	Pos Pred Value	0.8258	0.6512	0.114286
	Neg Pred Value	0.6997	0.8120	0.976010
	Prevalence	0.6397	0.3325	0.027811
	Detection Rate	0.5333	0.2031	0.004837
	Detection Prevalence	0.6457	0.3120	0.042322
	Balanced Accuracy	0.7608	0.7239	0.567678
ref2015, cartesian202	Sensitivity	0.7316	0.5782	0.304348
	Specificity	0.7987	0.7645	0.895522
	Pos Pred Value	0.8658	0.5502	0.076923
	Neg Pred Value	0.6263	0.7844	0.978261
	Prevalence	0.6397	0.3325	0.027811
	Detection Rate	0.4680	0.1923	0.008464
	Detection Prevalence	0.5405	0.3495	0.110036
	Balanced Accuracy	0.7651	0.6713	0.599935

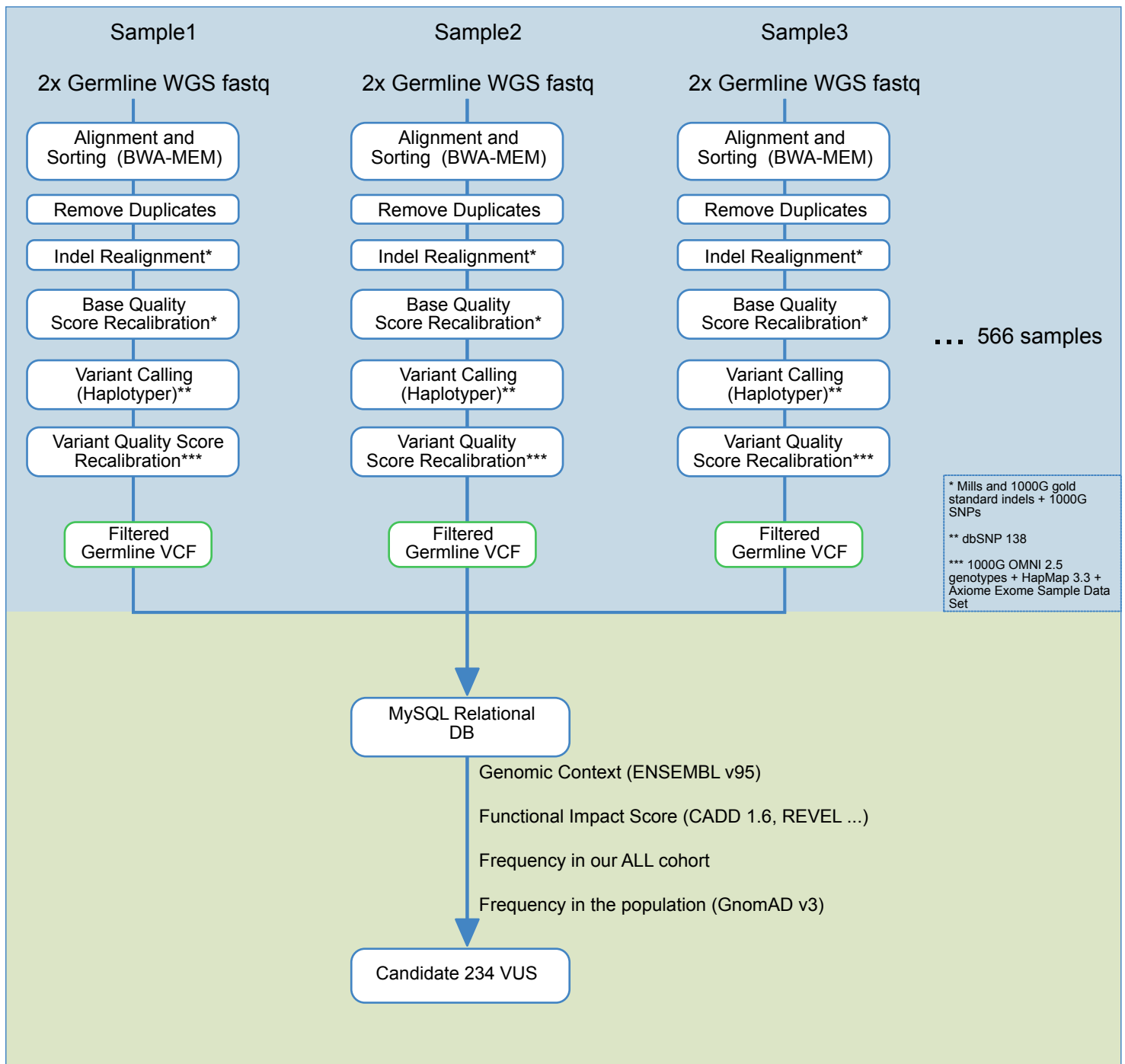
Table S2. Summary of the variants used in the case study 4 with structural coverage. The variants analyzed with Rosetta or FoldX are reported in the table, along with the corresponding REVEL scores and ClinVar classification. 'N.A.' indicates that the entry is not available in Clinvar. In bold we highlighted the variants with the consensus prediction between FoldX and Rosetta and changes in folding free energy > 1 kcal/mol.

gene name	variant	uniprot ID	CLINVAR	CLINVAR ACCESSION ID	FoldX (kcal/mol)	Rosetta (kcal/mol)	REVEL
BLM	A421T	P54132	Uncertain significance	VCV000127474.12	0,615	0,23	0,017
BLM	R643H	P54132	Conflicting interpretations of pathogenicity	VCV000127480.42	0,71	1,073	0,145
BLM	L788I	P54132	Conflicting interpretations of pathogenicity	VCV000127485.17	2,197	2,233	0,503
BLM	R791C	P54132	Uncertain significance	VCV000133698.13	1,166	0,8	0,411
BLM	A797T	P54132	Entry N.A.	Entry N.A.	1,635	3,045	0,893
BLM	K846T	P54132	Uncertain significance	VCV001327067.1	1,933	1,227	0,136
BLM	T907M	P54132	Uncertain significance	VCV000127492.13	-0,667	0,766	0,28
BLM	Y1024C	P54132	Uncertain significance	VCV000317408.9	2,809	1,776	0,59
BLM	A1203V	P54132	Uncertain significance	VCV000485332.10	0,254	-0,0673	0,141
ERCC4	S253A	Q92889	Entry N.A.	Entry N.A.	1,18	0,133	0,16
ERCC4	R267C	Q92889	Uncertain significance	VCV001060514.4	1,079	1,495	0,47
ERCC4	K280M	Q92889	Entry N.A.	Entry N.A.	-0,922	-0,355	0,557
ERCC4	S312A	Q92889	Uncertain significance	VCV000864133.5	1,18	-0,236	0,143
ERCC4	P379S	Q92889	Conflicting interpretations of pathogenicity	VCV000134148.18	3,62	2,051	0,526
ERCC4	A400V	Q92889	Entry N.A.	Entry N.A.	0,111	0,274	0,041
ERCC4	R576T	Q92889	Uncertain significance	VCV000134158.19	1,366	1,012	0,274
ERCC4	S662P	Q92889	Benign	VCV000134134.10	-0,327	3,666	0,015
ERCC4	I706T	Q92889	Conflicting interpretations of pathogenicity	VCV000134142.22	3,607	3,242	0,609
ERCC4	R799W	Q92889	Conflicting interpretations of pathogenicity	VCV000016580.39	-0,514	-0,573	0,51
ERCC4	Q849L	Q92889	Uncertain significance	VCV000943396.5	-1,306	-0,712	0,221
RFWD3	T288I	Q6PCD5	Entry N.A.	Entry N.A.	-0,24	-0,064	0,153
RFWD3	A297T	Q6PCD5	Entry N.A.	Entry N.A.	-0,299	-0,019	0,175
RFWD3	S361F	Q6PCD5	Benign	VCV001050685.1	-0,035	-0,46	0,151
RFWD3	T443A	Q6PCD5	Entry N.A.	Entry N.A.	0,688	0,301	0,06
RFWD3	Q577H	Q6PCD5	Entry N.A.	Entry N.A.	1,438	1,469	0,162
FANCA	S53R	O15360	Uncertain significance	VCV000237036.20	0,164	-0,356	0,022
FANCA	A181V	O15360	Benign/Likely benign; Uncertain Significance	VCV000134287.16	-0,34	0,034	0,173
FANCA	N250K	O15360	Entry N.A.	Entry N.A.	-0,19	1,097	0,02
FANCA	T266A	O15360	Likely benign; Benign	VCV000134294.21	1,228	-0,332	0,084
FANCA	F276I	O15360	Entry N.A.	Entry N.A.	2,917	1,922	0,16
FANCA	H292D	O15360	Uncertain significance	VCV000456139.9	1,09	0,969	0,092
FANCA	H322Y	O15360	Uncertain significance	VCV000456146.3	-1,603	-0,34	0,157

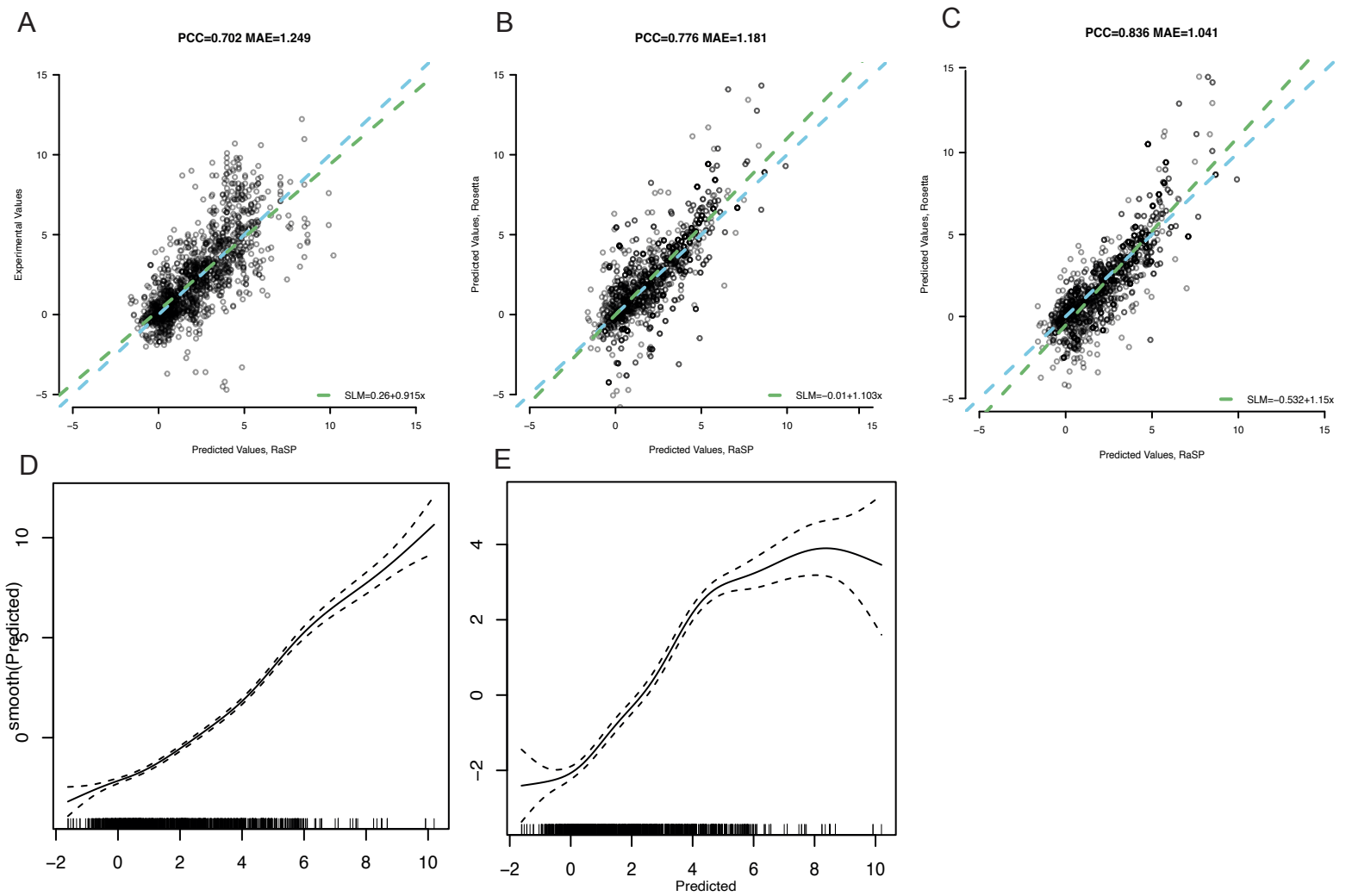
FANCA	W410C	O15360	Entry N.A.	Entry N.A.	5,549	4,169	0,622
FANCA	S447L	O15360	Uncertain significance	VCV000237034.10	1,833	-0,432	0,519
FANCA	G501S	O15360	Benign	VCV000134244.24	-0,015	-0,674	0,261
FANCA	F603S	O15360	Uncertain significance	VCV000654516.10	5,668	5,804	0,631
FANCA	C625S	O15360	Conflicting interpretations of pathogenicity	VCV000265136.31	0,8	1,272	0,806
FANCA	P643R	O15360	Benign	VCV000134248.16	0,292	1,196	0,179
FANCA	A746S	O15360	Benign/Likely benign	VCV000321349.13	2,007	1,696	0,374
FANCA	S858R	O15360	Benign/Likely benign	VCV000134256.26	-0,246	0,0761	0,134
FANCA	D953E	O15360	Uncertain significance	VCV000376974.21	-0,327	-0,487	0,156
FANCA	E1023D	O15360	Uncertain significance	VCV000550170.6	0,473	0,934	0,459
FANCA	P1086L	O15360	Entry N.A.	Entry N.A.	1,695	1,011	0,775
FANCA	T1131A	O15360	Pathogenic/Likely pathogenic	VCV000237048.29	0,52	0,405	0,842
FANCA	R1144W	O15360	Uncertain significance	VCV000408171.17	1,214	0,189	0,458
FANCA	T1161M	O15360	Conflicting interpretations of pathogenicity	VCV000435126.8	0,472	0,2812	0,427
FANCA	V1180M	O15360	Likely benign	VCV000134270.13	-2,514	-1,426	0,225
FANCA	P1213A	O15360	Uncertain significance	VCV000526331.10	0,797	0,553	0,167
FANCA	V1287I	O15360	Benign/Likely benign	VCV000134275.16	-0,93	-0,634	0,179
FANCA	R1317Q	O15360	Uncertain significance	VCV000321329.9	-0,255	0,0309	0,155
FANCA	H1417D	O15360	Conflicting interpretations of pathogenicity	VCV000134282.26	-0,285	1,335	0,221
RAD51C	A126T	O43502	Benign/Likely benign	VCV000132721.45	1,393	0,232	0,102
RAD51C	E161V	O43502	Uncertain significance	VCV000496504.5	-0,043	-0,412	0,892
RAD51C	G264S	O43502	Conflicting interpretations of pathogenicity	VCV000128211.49	-1,038	-0,481	0,202
MSH2	N127S	P43246	Benign	VCV000036577.32	1,579	2,172	0,741
MSH2	L128V	P43246	Conflicting interpretations of pathogenicity	VCV000127644.23	1,776	2,309	0,613
MSH2	K275Q	P43246	Uncertain significance	VCV000838497.5	0,077	1,261	0,726
MSH2	Q397H	P43246	Conflicting interpretations of pathogenicity	VCV000224575.15	-0,001	0,623	0,51
MSH2	L432M	P43246	Uncertain significance	VCV000663232.8	0,397	2,306	0,644
MSH2	L513V	P43246	Uncertain significance	VCV000479825.5	2,968	2,595	0,829
MSH2	I577T	P43246	Likely benign	VCV000041644.45	1,074	2,067	0,928
MSH2	N596S	P43246	Conflicting interpretations of pathogenicity	VCV000041646.44	0,304	0,685	0,385
MSH2	V606F	P43246	Entry N.A.	Entry N.A.	4,898	5,095	0,889
MSH2	I770V	P43246	Conflicting interpretations of pathogenicity	VCV000090955.21	1,253	1,043	0,417
MSH6	L396V	P52701	Benign	VCV000036582.33	3,764	1,369	0,322
MSH6	S503C	P52701	Entry N.A.	VCV000089198.40	1,077	1,288	0,413
MSH6	T636S	P52701	Uncertain significance	VCV000483839.6	0,639	0,681	0,182
MSH6	V878A	P52701	Benign	VCV000008931.49	2,192	2,0738	0,155
MSH6	G1072D	P52701	Uncertain significance	VCV000234031.15	4,278	6,349	0,623
MSH6	P1073S	P52701	Likely benign	VCV000041593.44	0,706	1,116	0,312
MSH6	P1087A	P52701	Conflicting interpretations of pathogenicity	VCV000135841.25	1,355	0,962	0,597
MSH6	E1163V	P52701	Likely benign	VCV000089400.22	0,392	-0,247	0,84
MSH6	T1243S	P52701	Conflicting interpretations of pathogenicity	VCV000127589.37	3,332	0,866	0,587
MSH6	V1253E	P52701	Uncertain significance	VCV000127591.22	2,855	2,986	0,952

MSH6	R1321G	P52701	Conflicting interpretations of pathogenicity	VCV000089490.41	0,827	1,34	0,672
MSH6	V1336D	P52701	Entry N.A.	Entry N.A.	0,495	1,608	0,584
MLH1	V76I	P40692	Uncertain significance	VCV000237335.17	-0,252	0,852	0,616
MLH1	V180G	P40692	Uncertain significance	VCV000090257.30	4,582	3,847	0,91
MLH1	L272V	P40692	Uncertain significance	VCV000090385.2	2,134	0,897	0,622
MLH1	P285S	P40692	Uncertain significance	VCV000420749.11	2,104	1,373	0,838
MLH1	K618E	P40692	Benign/Likely benign	VCV000089903.36	1,536	1,296	0,874
MLH1	K618T	P40692	Benign	VCV000089906.37	-0,596	0,425	0,963
MLH1	V716M	P40692	Entry N.A.	VCV000041639.42	1,034	0,401	0,523
MLH1	H718Y	P40692	Benign	VCV000036547.28	-0,352	-0,456	0,898
NBN	D95N	O60934	Conflicting interpretations of pathogenicity	VCV000127869.57	2,324	1,448	0,583
NBN	V101A	O60934	Uncertain significance	VCV000185055.26	1,38	0,718	0,298
NBN	I171V	O60934	Conflicting interpretations of pathogenicity	VCV000006946.49	1,345	1,133	0,398
NBN	R215W	O60934	Conflicting interpretations of pathogenicity	VCV000006948.52	-0,433	2,541	0,343
NBN	K408E	O60934	Benign/Likely benign	VCV000127856.29	-0,245	2,816	0,081
FANCF	A34V	Q9NPI8	Uncertain significance	VCV000304208.6	0,215	2,499	0,241
FANCF	R50W	Q9NPI8	Conflicting interpretations of pathogenicity	VCV000697625.17	1,037	0,917	0,024
FANCF	L80V	Q9NPI8	Entry N.A.	Entry N.A.	2,681	2,001	0,104
FANCF	P117L	Q9NPI8	Conflicting interpretations of pathogenicity	VCV000304204.12	1,303	0,908	0,139
FANCF	L129V	Q9NPI8	Uncertain significance	VCV000134348.13	2,112	1,0863	0,069
FANCF	G150S	Q9NPI8	Entry N.A.	Entry N.A.	0,25	-0,0768	0,004
FANCF	L222F	Q9NPI8	Entry N.A.	Entry N.A.	-0,168	0,64	0,046
FANCF	Y274C	Q9NPI8	Uncertain significance	VCV001375139.3	4,051	3,526	0,193
FANCF	Y287C	Q9NPI8	Uncertain significance	VCV000304201.9	1,599	2,732	0,195
FANCF	L329P	Q9NPI8	Uncertain significance	VCV000859048.5	6,591	8,17	0,417
FANCF	Q363K	Q9NPI8	Entry N.A.	Entry N.A.	-0,302	0,24	0,03
FANCI	S54F	Q9NVI1	Entry N.A.	Entry N.A.	6,178	0,289	0,195
FANCI	I132V	Q9NVI1	Uncertain significance	VCV000456222.2	0,342	0,0034	0,042
FANCI	Q216E	Q9NVI1	Entry N.A.	Entry N.A.	-0,026	0,0748	0,587
FANCI	P260L	Q9NVI1	Entry N.A.	Entry N.A.	0,456	0,7884	0,39
FANCI	I275T	Q9NVI1	Uncertain significance	VCV000449021.11	3,656	3,0411	0,22
FANCI	V290M	Q9NVI1	Benign	VCV000317267.11	-1,566	0,47	0,017
FANCI	M363T	Q9NVI1	Entry N.A.	Entry N.A.	2,013	2,593	0,242
FANCI	V372I	Q9NVI1	Benign	VCV000317271.12	-0,403	-0,0904	0,195
FANCI	P471R	Q9NVI1	Uncertain significance	VCV000408243.6	1,527	1,7306	0,837
FANCI	M525V	Q9NVI1	Conflicting interpretations of pathogenicity	VCV000238309.20	3,345	2,137	0,487
FANCI	I671V	Q9NVI1	Benign/Likely benign	VCV000238312.13	1,255	0,924	0,023
FANCI	I735V	Q9NVI1	Uncertain significance	VCV000456205.5	0,544	-0,0255	0,092
FANCI	E868D	Q9NVI1	Benign/Likely benign	VCV000238317.13	0,413	0,0333	0,032
FANCI	Y923C	Q9NVI1	Uncertain significance	VCV000317288.7	3,832	3,806	0,391
FANCL	L38F	Q9NW38	Conflicting interpretations of pathogenicity	VCV000221092.23	0,678	3,57	0,24

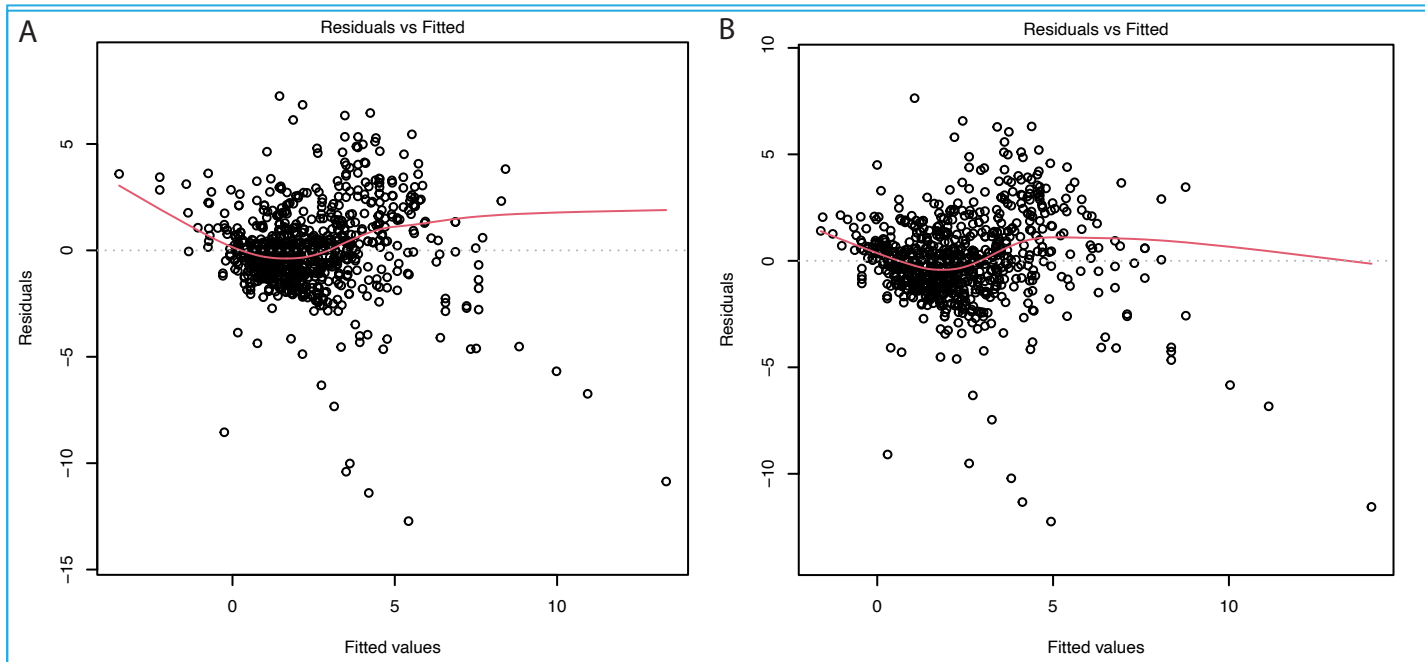
FANCE	P77T	Q9HB96	Uncertain significance	VCV000134330.12	1,064	0,855	0,21
FANCE	A104P	Q9HB96	Entry N.A.	Entry N.A.	1,914	5,691	0,319
FANCE	R134H	Q9HB96	Uncertain significance	VCV001027036.2	0,424	0,0619	0,025
FANCE	L326W	Q9HB96	Uncertain significance	VCV000356449.9	1,019	1,62	0,154
FANCE	M437T	Q9HB96	Conflicting interpretations of pathogenicity	VCV000414821.13	4,411	1,93	0,134
FANCE	A502T	Q9HB96	Benign	VCV000134345.17	0,648	1,329	0,05
FANCG	A278V	O15287	Uncertain significance	VCV000847372.3	0,749	4,068	0,079
FANCG	T297I	O15287	Benign/Likely benign	VCV000134367.20	-0,257	0,069	0,043
FANCG	K430E	O15287	Entry N.A.	Entry N.A.	0,017	0,198	0,048
FANCG	E481K	O15287	Entry N.A.	Entry N.A.	0,346	1,208	0,132
FANCG	P545T	O15287	Entry N.A.	Entry N.A.	2,408	1,918	0,465



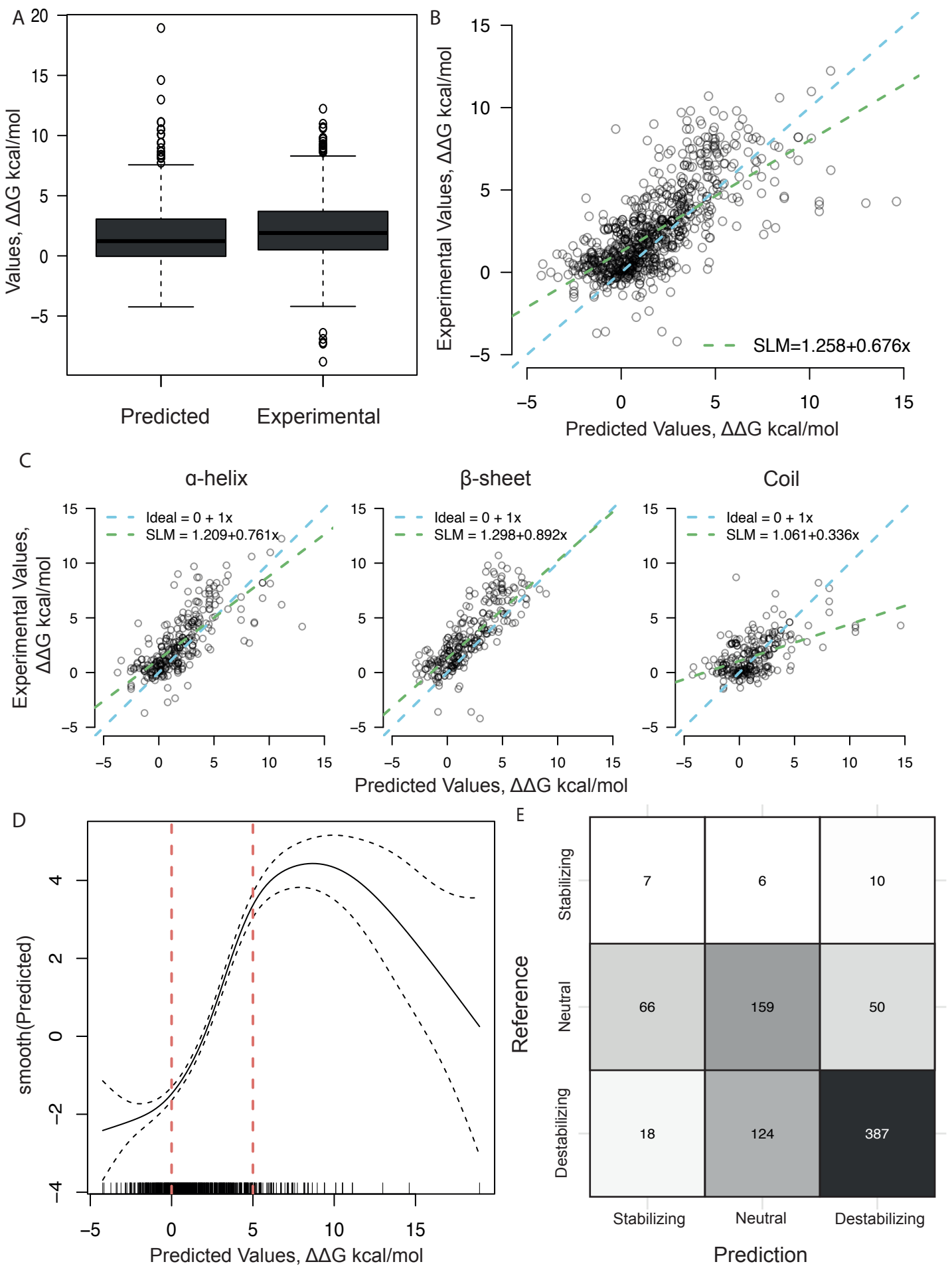
Supplementary Figure S4. Schematic representation of NGS bioinformatics pipeline for candidate VUS extraction and annotation. Fastq pairs for each sample (normal blood) are processed independently (n=566) with Sentieon® software based on GATK best practices. Reads are aligned to Human Genome build 38 (GATK resource bundle for hg38) with BWA-MEM algorithm. Duplicate reads are removed and reads are realigned around indels. After that, we perform Base Quality Score Recalibration to detect systematic errors produced by the sequencing machine. These last two steps are performed using as reference known SNPs and indels from Mills and 1000G projects (*). Then, variants are called with Haplotype algorithm (equivalent to GATK Haplotype Caller (<https://doi.org/10.1101/201178>)) using SNPs from dbSNP 138 as known resource (**). Finally, Variant Quality Score Recalibration is performed which uses 1000G OMNI 2.5 genotypes, HapMap 3.3 genotypes and Axiom Exome Sample Data Set (***) to model the technical profile of the variants and filter our potential artifacts from our call set. Per sample call sets are independently uploaded into an in-house MySQL (MariaDB) database. For each variant, annotation for genomic context from Ensembl v95 is added, as well as Functional Impact scores (CADD 1.6, REVEL...) and frequency in the cohort and healthy population (GnomAD v3). Variants affecting the coding region with frequency in the healthy population of <1% were selected as candidate VUS (n=234).



Supplementary figure S3. Comparison of stability prediction performance of RaSP. (A) SLM of the RaSP predicted values and the experimental values. (B) SLM of the RaSP predicted values and the values predicted by Rosetta with the Cartesian protocol. (C) SLM of the RaSP predicted values and the values predicted by Rosetta with the Cartesian2020 protocol. The best performance is seen in panel C, the protocol RaSP is build to simulate. (D) a GAM of the predicted RaSP values towards the predicted Rosetta Cartesian 2020 values, illustrating how there is a linear relationship between the two. (E) a GAM of the predicted RaSP values towards the experimental values illustrating how the linear relationship is limited to a similar area, 0-5 as the Rosetta pedictions.



Supplementary Figure S2. Residual plots. (A) The difference between the observed and the fitted response values for the simple linear model from the ref2015 cartesian protocol. (B) The difference between the observed and the fitted response values for the simple linear model from the ref2015 cartesian2020 protocol. Both plots illustrate a biased fit, as they are not randomly scattered around the identity line $y=0$.



Supplementary Figure S1. Comparison of stability changes from the ref2015 cartesian2020 protocol and experimentally found values. (A) The distribution of the predicted and experimental stability changes in kcal/mol. (B) A scatterplot of the $\Delta\Delta G$ values predicted by the ref2015 cartesian2020 protocol and experimentally found values for the corresponding mutations. The green line is the fitted simple linear model. The model has an intercept=1.26, slope=0.68, $\sigma^2=4.06$, $R^2=0.43$, Pearson Correlation Coefficient=0.65, (C) Three scatterplots to illustrate the data as divided by the wild-type secondary structure of the mutated position. The green line is the fitted simple linear model. Here it is evident how the structured sections have a better correspondence as compared to the coils. α -helices: Pearson correlation coefficient=0.75, β -sheets: Pearson correlation coefficient=0.69, Coil: Pearson correlation coefficient=0.52, (D) A generalized additive model (GAM) modeling the response variable, experimental $\Delta\Delta G$, to a predictive variable, predicted $\Delta\Delta G$ by estimating a smooth function, smooth(Predict). The smooth function has an effective degree of freedom of 5.5, quantifying the complexity of the line. The confidence interval is sufficiently narrow in the $\Delta\Delta G$ interval 0-5 kcal/mol to indicate that a linear relationship is present in this interval. (E) A confusion matrix where the experimental values are annotated as the reference values. The threshold used to define the classes is a $\Delta\Delta G$ of < -1 kcal/mol for stabilizing mutations, $-1 < \Delta\Delta G < 1$ kcal/mol for neutral mutations and $\Delta\Delta G > 1$ kcal/mol for destabilizing mutations. The resulting accuracy is 0.67.

Manuscript 4:

Redefining germline predisposition in children with molecularly characterized ependymoma: a population-based 20-year cohort.

Authors: Jon Foss-Skiftesvik*† , Ulrik Kristoffer Stoltze†, Thomas van Overeem Hansen, Lise Barlebo Ahlborn, Erik Sørensen, Sisse Rye Ostrowski, Solvej Margrete Aldringer Kullegaard, Adrian Otamendi Laspiur, Linea Cecilie Melchior, David Scheie, Bjarne Winther Kristensen, Jane Skjøth-Rasmussen, Kjeld Schmiegelow, Karin Wadt† and René Mathiasen†

†Jon Foss-Skiftesvik and Ulrik Kristoffer Stoltze contributed equally as co-first authors

†Karin Wadt and René Mathiasen contributed equally as co-last authors

*Correspondence: jon.foss-skiftesvik@regionh.dk

Contribution: I contributed to this project with the bioinformatics analysis of the NGS samples. I used the updated version of Sentieon *DNaseq* pipeline based on GCRh38 for short variants detection and reported the genetic variants identified on the specific samples which were further studied by Jon and Ulrik among others.

RESEARCH

Open Access



Redefining germline predisposition in children with molecularly characterized ependymoma: a population-based 20-year cohort

Jon Foss-Skiftesvik^{1,2,4,10,11*} , Ulrik Kristoffer Stoltze^{1,3,4†}, Thomas van Overeem Hansen^{3,4}, Lise Barlebo Ahlborn⁵, Erik Sørensen⁶, Sisse Rye Ostrowski^{4,6}, Solvej Margrete Aldringer Kullegaard¹, Adrian Otamendi Laspiur⁹, Linea Cecilie Melchior⁷, David Scheie⁷, Bjarne Winther Kristensen^{4,7,8}, Jane Skjøth-Rasmussen^{2,4}, Kjeld Schmiegelow^{1,4}, Karin Wadt^{3†} and René Mathiasen^{1†}

Abstract

Ependymoma is the second most common malignant brain tumor in children. The etiology is largely unknown and germline DNA sequencing studies focusing on childhood ependymoma are limited. We therefore performed germline whole-genome sequencing on a population-based cohort of children diagnosed with ependymoma in Denmark over the past 20 years ($n = 43$). Single nucleotide and structural germline variants in 457 cancer related genes and 2986 highly evolutionarily constrained genes were assessed in 37 children with normal tissue available for sequencing. Molecular ependymoma classification was performed using DNA methylation profiling for 39 children with available tumor tissue. Pathogenic germline variants in known cancer predisposition genes were detected in 11% (4/37; *NF2*, *LZTR1*, *NF1* & *TP53*). However, DNA methylation profiling resulted in revision of the histopathological ependymoma diagnosis to non-ependymoma tumor types in 8% (3/39). This included the two children with pathogenic germline variants in *TP53* and *NF1* whose tumors were reclassified to a diffuse midline glioma and a rosette-forming glioneuronal tumor, respectively. Consequently, 50% (2/4) of children with pathogenic germline variants in fact had other tumor types. A meta-analysis combining our findings with pediatric pan-cancer germline sequencing studies showed an overall frequency of pathogenic germline variants of 3.4% (7/207) in children with ependymoma. In summary, less than 4% of childhood ependymoma is explained by genetic predisposition, virtually restricted to pathogenic variants in *NF2* and *NF1*. For children with other cancer predisposition syndromes, diagnostic reconsideration is recommended for ependymomas without molecular classification. Additionally, *LZTR1* is suggested as a novel putative ependymoma predisposition gene.

Keywords: DNA methylation profiling, Molecular classification, Genomics, Genetic susceptibility, Pediatrics

[†]Jon Foss-Skiftesvik and Ulrik Kristoffer Stoltze contributed equally as co-first authors

[†]Karin Wadt and René Mathiasen contributed equally as co-last authors

*Correspondence: jon.foss-skiftesvik@regionh.dk

¹⁰ Department of Neurosurgery, Section 6031, Rigshospitalet University Hospital, Inge Lehmanns Vej 6, 2100 Copenhagen, Denmark
Full list of author information is available at the end of the article

Introduction

Ependymoma is the second most common malignant central nervous system (CNS) tumor in children and is associated with poor long-term survival [1, 2]. Apart from a very limited number of children with neurofibromatosis type-2 associated spinal ependymoma, the underlying causes of ependymoma remain unknown [3,



4]. Several factors indicate that genetic predisposition plays a role including increased population-based familial risk [5], reports of familial intracranial ependymoma [6, 7], genetic ancestry-based risk differences [8] and an absence of known environmental risk factors [9].

No systematic germline sequencing investigation of genetic predisposition specific to childhood ependymoma has been reported to date. Over the last decade, several large pediatric pan-cancer germline sequencing studies have been performed, with childhood ependymoma accounting for less than 5% (191/4833) of the combined sample size [10–19]. Taken together, these whole-exome/-genome sequencing (WES/WGS) studies report rare pathogenic germline variants in 4.7% (9/191) of children with ependymoma, although individual study estimates range from 0 to 21%. Lack of molecular tumor classification [10, 11, 13–19], small ependymoma sample sizes [11, 12, 14–19], restriction to gene panels [10–19] and lack of population-based study designs [10–14, 16–19] further complicate the delineation of the nature and extent of genetic predisposition in childhood ependymoma.

The aim of this population-based study was to investigate genetic predisposition in children with molecularly classified ependymoma due to rare pathogenic germline variants both in and outside known cancer genes. Moreover, we assessed the feasibility of performing germline WGS and tumor DNA methylation profiling in a combined retro-/prospective nationwide cohort spanning more than 20 years.

Material and methods

Retrospective cohort

Children (<18 years) diagnosed with ependymoma from 2000 to 2016 in Denmark were identified through the Danish Childhood Cancer Registry (DCCR) [20]. Registry data on date of birth, gender, histopathology and tumor location was validated by cross-linkage with the National Pathology Registry. Living patients aged >18 years at the time of the study were informed and offered inclusion both in writing and by telephone. For minors (<18 years at the time of the study) and for deceased patients, parents or legal guardians were contacted. Detailed clinical and four-generational pedigrees were retrieved through patient health record review for included patients.

Prospective cohort

Since 2016, all children (<18 years) diagnosed with cancer in Denmark have been offered germline WGS through the STAGING study, described in detail elsewhere [15, 21]. The prospective cohort consists of children with ependymoma included in STAGING from 2016 to 2021. Similarly to the retrospective cohort, data

variables were retrieved through patient health record and histopathology report review.

Collection of tissue for germline DNA sequencing

Leukocyte DNA was isolated from peripheral blood samples drawn in parallel with clinical sampling when possible. For deceased patients, archived blood samples were collected from the Copenhagen Hospital Biobank (CHB) [22]. For those without obtainable blood samples, dissection of normal brain tissue was performed on formalin-fixed paraffin embedded (FFPE) tumor tissue samples.

Germline whole-genome and -exome sequencing

Germline WGS was performed on leukocyte DNA using the HiSeqX platform (Illumina, USA) with paired-end sequencing of 150 bp reads and target 30X average coverage. Germline WES of healthy brain tissue was performed using Novaseq 6000 (Illumina, USA). Exomes were sequenced as 2 × 150 bp paired-end reads to an average median coverage of 60X. Tissue handling, sequencing and bioinformatics procedures including variant filtering are further detailed in the Additional file 1: Methods.

Cancer gene panel analysis

For the gene panel analysis, WGS/WES data was limited to filtered SNVs and SV deletions identified in a predefined set of 457 genes. This panel consisted of 390 cancer related genes supplemented by 67 genes with either established or suggested roles in ependymoma tumorigenesis selected based on the scientific literature (Additional file 2: Tables S1 and S2). Variants were reviewed by a multidisciplinary team specialized in pediatric cancer predisposition. Variants were classified as either “benign”, “likely benign”, “likely pathogenic”, “pathogenic”, or as “variants of unknown significance” (VUS) in accordance with international standards [23]. In the context of this study, “likely pathogenic” and “pathogenic” variants are referred to simply as “pathogenic”.

Constrained gene analysis

For the constrained gene analysis, all rare, coding SNVs and SV deletions, were subsetted to variants predicted to cause loss-of-function (pLoF) in 2986 highly constrained genes. Based on metadata from 141,456 humans without serious childhood disease, evolutionarily constrained genes were defined by a LoF observed/expected upper bound fractions (LOEUF) score of ≤ 0.35 which is indicative of depletion of pLoF variation and in line with recent recommendations [24, 25]. Curation of resulting variants, including use of 586 in-house whole genome sequences from children with cancers other than ependymoma, is

detailed in the Additional file 1: Methods and Additional file 2: Table S3.

Tumor DNA methylation profiling and molecular classification

Molecular tumor classification was performed using retrospectively collected iDAT files for patients with existing clinical DNA methylation profiles. For all others, archived FFPE or freshly frozen (FF) tumor samples were collected and underwent DNA methylation profiling using the Infinium MethylationEPIC BeadChip Kit (Illumina, USA). Archived tumor DNA was restored using the Infinium HD FFPE DNA Restore Kit (Illumina, USA) prior to methylation profiling. Tumor methylation class and subclass were predicted using a publicly available classifier tool [26]. The classifier version and employed cut-off scores are further detailed in Additional file 1: Methods. For an illustrative overview of the cohort and methods used, please see Fig. 1).

Statistical analyses

Statistical analyses were performed using R v.3.6.1 and IBM SPSS Statistics v.25.

Ethical approvals

This study was approved by the Capital Region Scientific Ethical Committee (H-15016782, prospective cohort) and the Danish National Committee on Health Research

Ethics (2000407). All patients and/or parents/legal guardians provided informed consent.

Results

Patient characteristics

A total of 43 children registered with an ependymoma diagnosis were included. Median age at diagnosis (5.3, SD 4.7), gender distribution (females 44.2%), histopathology diagnosis, and tumor location (Table 1) were in line with existing population-based reports [27–29]. The overall inclusion rate was 77% (43/56). For the retrospective cohort, in which both living and deceased patients were eligible for inclusion, a higher rate of inclusion was seen for deceased patients compared to living patients (91% vs. 66%, Fisher’s exact test, $p=0.067$). The inclusion process, including main reasons for exclusion, is illustrated in Additional file 1: Fig. S3.

Molecular tumor classification

Molecular tumor (re-)classification based on DNA methylation profiling was possible for 90% (39/43) of patients. Distribution of original histopathological diagnosis and resulting tumor methylation class and subclass is listed in Table 1 and illustrated in Fig. 2, respectively.

Ultimately, the reclassification rate for patients histopathologically diagnosed with ependymoma and with available tumor tissue was 7.7% (3/39). Initially, tumor methylation class prediction mandated amendment of the registered diagnosis to a non-ependymoma entity for four patients (Figs. 2 and 3). Two patients with

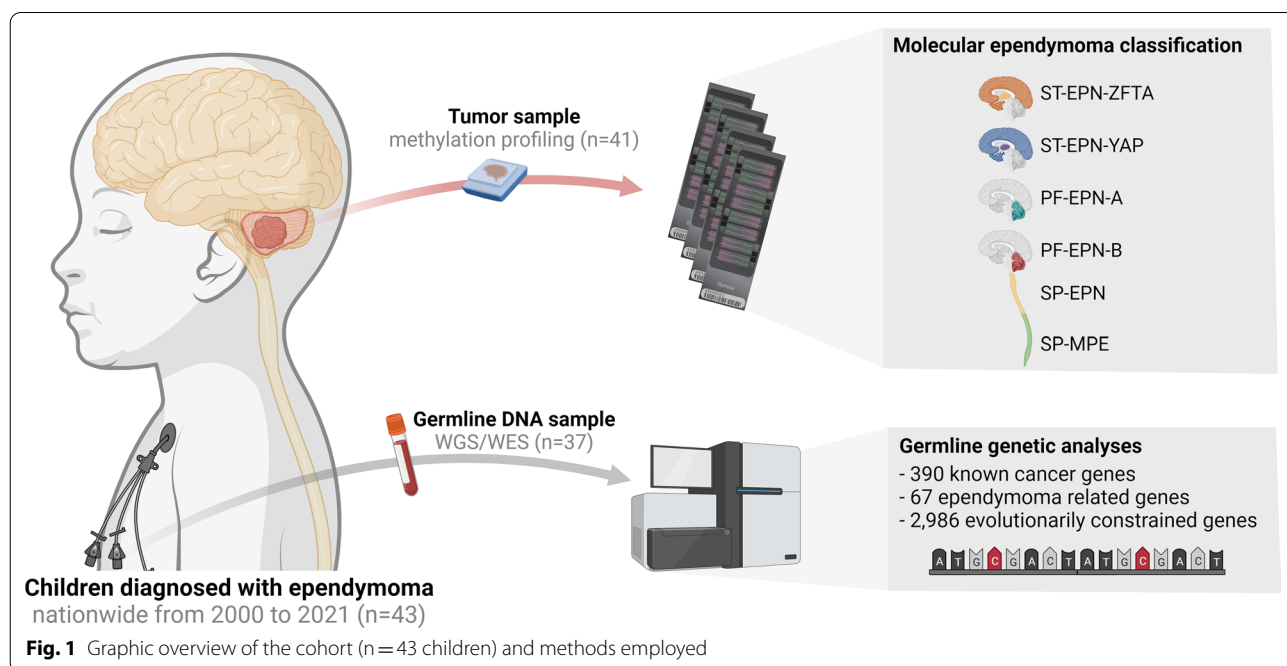


Table 1 Patient clinical characteristics

Patient characteristics	n (% of total)
Total	43 (100%)
Median age at diagnosis, y (SD)	5.3 (4.7)
<i>Status</i>	
Alive	27 (62.8%)
Deceased	16 (37.2%)
<i>Gender</i>	
Female	19 (44.2%)
Male	24 (55.8%)
<i>Cohort</i>	
Retrospective	34 (79.1%)
Prospective	9 (20.9%)
<i>Histopathological diagnosis</i>	
Myxopapillary ependymoma, WHO 2	1 (2.3%)
Ependymoma, WHO 2	14 (32.6%)
Ependymoma, WHO 3	26 (60.5%)
Other*	2 (4.7%)
<i>Tumor location</i>	
Supratentorial	7 (16.3%)
Posterior fossa	30 (69.8%)
Spinal	5 (11.6%)
Multifocal**	1 (2.3%)

* Includes one patient initially diagnosed with atypical glioblastoma for whom subsequent clinical tumor methylation profiling resulted in an ependymoma diagnosis and one patient with ependymoblastoma incorrectly registered as ependymoma

** Includes one patient with disseminated ependymoma at diagnosis with tumor tissue located adherent to the insular cortex, the ventral surface of the brainstem and the caudal spinal cord

SD, standard deviation; y, years; WHO, the World Health Organization histological grade

histopathologically diagnosed WHO grade 3 ependymomas located in the pons and thalamus, respectively, were both reclassified as *H3K27*-mutant or *EZH1P* expressing diffuse midline gliomas (DMG_H3K27). Another tumor, extending through the aqueduct from the fourth ventricle also registered as ependymoma in the DCCR, was reclassified as a rosette-forming glioneuronal tumor (RGNT) based on DNA methylation profiling. Of note, the original histopathology report of this tumor revealed a discussion of several differential diagnoses. Finally, an ependymoblastoma incorrectly coded as ependymoma in the registry was specified as a C19mc-altered embryonal tumor with multilayered rosettes (ETMR). All reclassifications were supported by subsequent review of the original histopathology reports by a senior pediatric neuropathologist. For one of the molecularly classified ependymoma patients, a chart review revealed a previous alteration of the initial histopathological diagnosis of atypical glioblastoma to ependymoma based on clinical DNA methylation profiling (Fig. 2).

Germline DNA sequencing

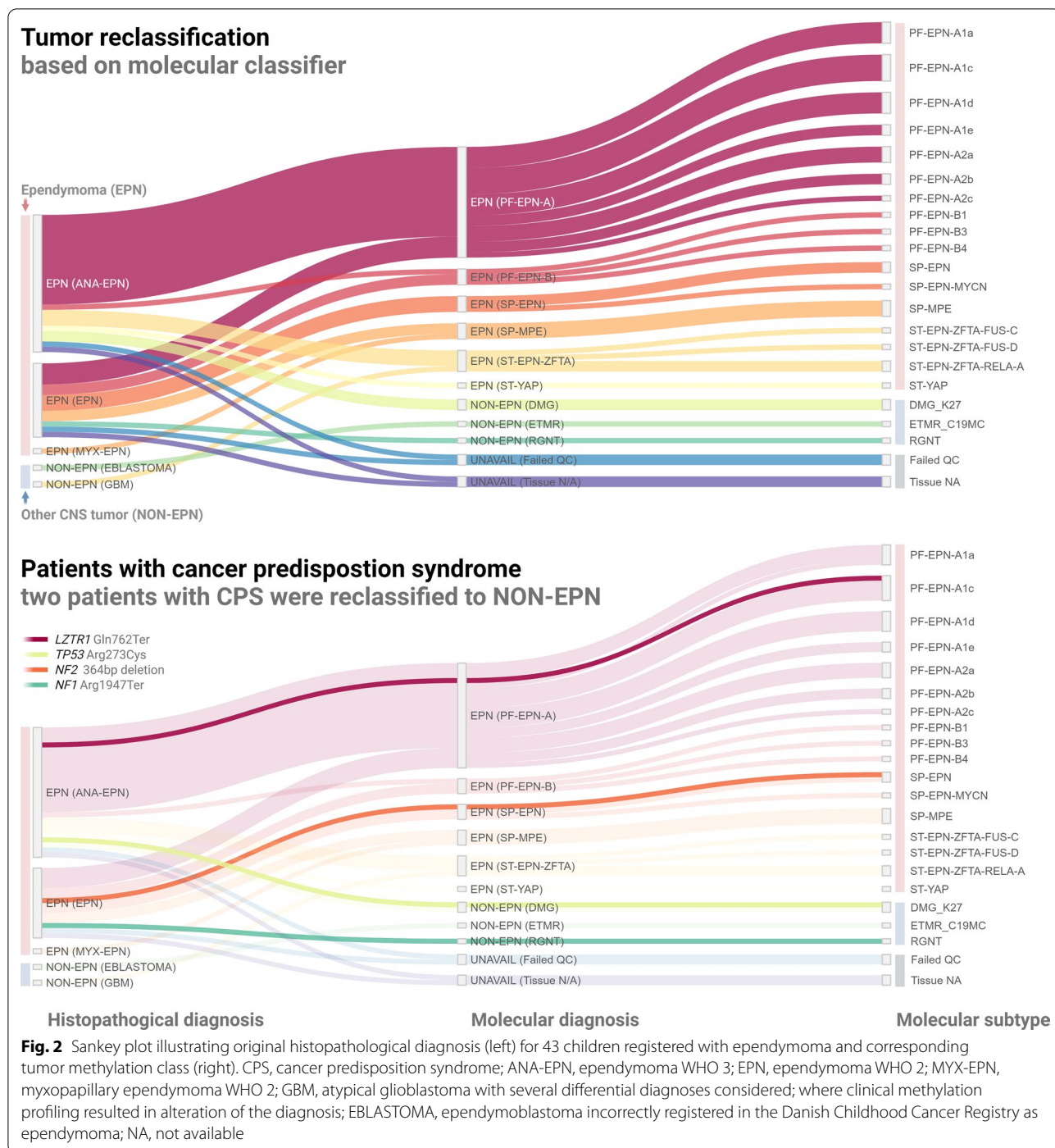
Tissue for germline DNA sequencing was available for 86% (37/43; new or prospectively collected blood samples $n=28$, archived blood samples $n=6$, normal brain tissue $n=3$) of which 34 retained ependymoma status following molecular tumor classification. The six patients not undergoing germline sequencing were all deceased, part of the retrospective cohort, and without available archived blood samples or dissectible healthy brain tissue in the FFPE tumor samples.

Cancer panel analysis findings

Nine pathogenic variants (eight SNVs, one SV) in nine patients were detected across the 457 cancer panel genes. Five heterozygous loss-of-function variants in the recessive genes *FANCM*, *ERCC3*, and *SBDS*, along with relatively common risk allele variants in *CHEK2* and *BRIP1*, were considered unrelated to ependymoma, but are further detailed in Additional file 2: Table S6.

Two of the four pathogenic variants at first assumed to be related to ependymoma were found in children where the histopathological diagnosis was subsequently altered following tumor DNA methylation profiling (Fig. 3). This included an *NF1* nonsense variant (p.Arg1947Ter [c.5839C > T]) in a patient with a molecularly confirmed RGNT and a *TP53* missense variant (p.Arg273Cys [c.817C > T]) in a child with a thalamic DMG_H3K27. Thus, the likelihood of diagnostic reclassification by DNA methylation profiling to a non-ependymoma tumor entity was significantly higher for children with detected pathogenic germline variants (2/4 vs. 0/29, Fisher's exact test, $p=0.011$, analysis limited to patients with ependymoma confirmed as initial histopathology diagnosis and both available tumor and germline tissue, $n=33$, Additional file 2: Table S4).

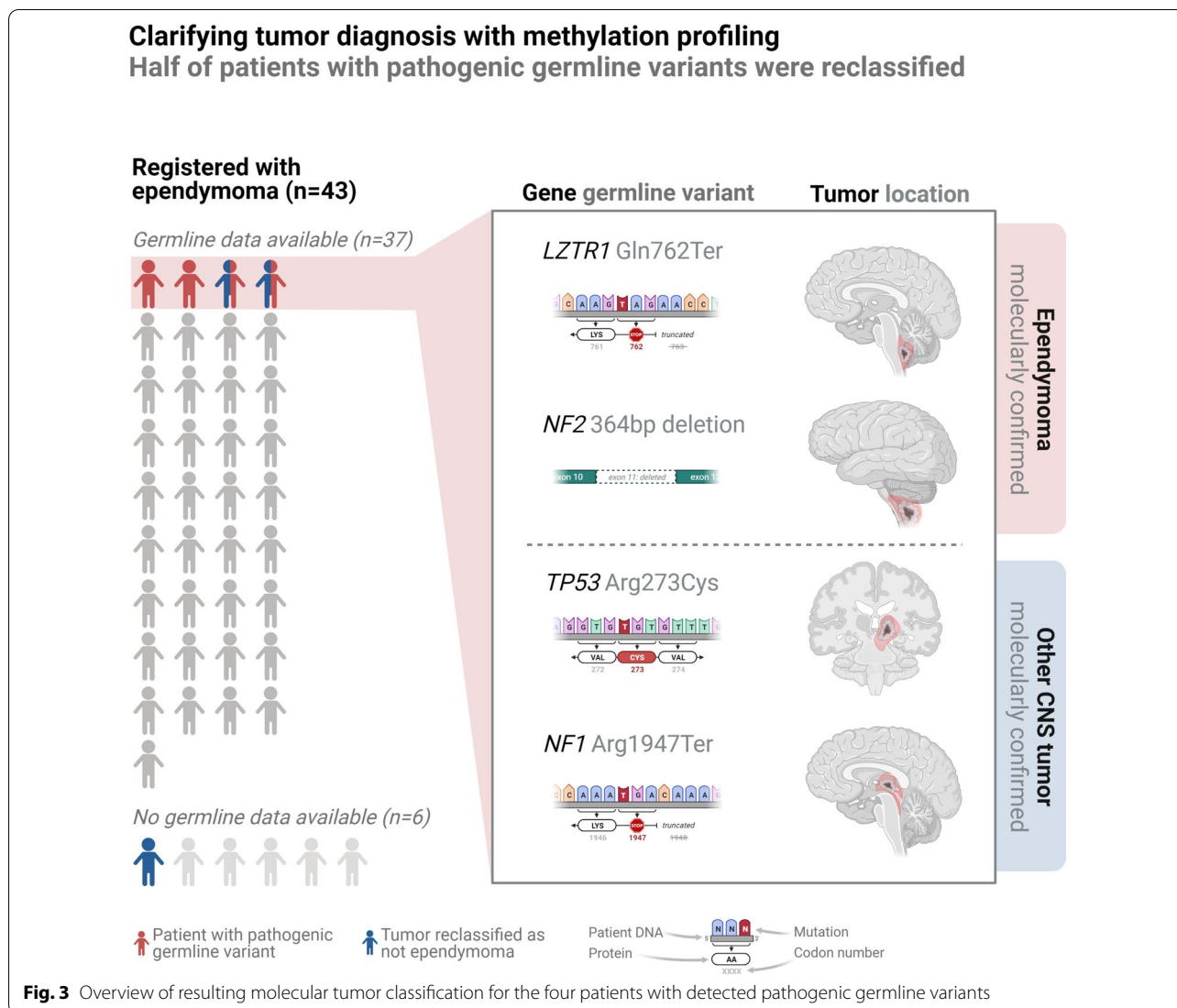
A causative 364 bp *NF2* deletion (chr22:30067648–30068012; p.Met334_Leu374del [c.1000-167_1122+75del]) was detected in a young child diagnosed with a WHO grade 2 ependymoma (methylation class spinal ependymoma (SP-EPN)) located at the cervicomedullary junction. The patient was initially treated with partial surgical resection followed by focal radiation and adjuvant chemotherapy, after which a minor contrast-enhancing tumor remnant has remained stable for more than 10 years. During follow-up, the patient developed bilateral vestibular schwannomas. Despite a family history with one third generation and several fourth generation relatives with clinically diagnosed neurofibromatosis type-2, the diagnosis had not been suspected until the patient debuted with ependymoma.



Finally, a pathogenic nonsense variant in *LZTR1* (p.Gln762Ter [c.2284C>T]), a gene not formerly linked with ependymoma, was detected in an otherwise healthy child diagnosed with a WHO grade 3 ependymoma (methylation class posterior fossa group A (PF-EPN-A) ependymoma, subclass 1c), located in

and around the foramen of Luschka. Of note, the only other *LZTR1* variant observed in our cohort was a VUS (p.Asp703Asn [c.2107G>A]) in another child diagnosed with the same molecular ependymoma subclass (PF-EPN-A1c) in the same location.

No pathogenic variants were detected in the supplementary panel of 67 ependymoma related genes.



Constrained gene analysis findings

Sixteen pLoF variants (11 SNVs and five SVs) were observed in the same number of constrained genes in 12 patients. Both pLoF variants already known to cause ependymoma (in *NF2* and *NF1*), were rediscovered. However, the nonsense *NF1* variant was found in a patient for whom DNA methylation profiling amended the ependymoma diagnosis to RGNT (Additional file 2: Tables S3 and S5).

Following molecular reclassification, 14 constrained gene pLoF variants remained, and were located in the following genes (ordered according to rising LOEUF scores); *CHD6*, *NF2*, *COL1A1*, *FGD5*, *BRWD1*, *UHRF2*, *ZNF1*, *FOXO3*, *CDC42BPA*, *DHX37*, *DNAJC2*, *TRIM67*, *ZMYM2*, *VPS4A*. No significant enrichments were detected using the String Database v.11 [30]. However, all but one (*COL1A1*) are expressed in normal brain tissue

[31]. Interestingly, 6/7 of the constrained genes in which pLoF variants were found in patients with posterior fossa ependymoma show particularly high expression levels in cerebellar tissue (*UHRF2*, *FOXO3*, *CDC42BPA*, *ZMYM2*, *CHD6* and *DNAJC2*) [32].

Other than a 3.4-fold enrichment of non-membrane-bounded organelles (false discovery rate 3.56e-3) the GO PANTHER Cellular Component Overrepresentation Test [33] did not reveal any other significant enrichments for the detected constrained genes when compared to all other genes.

Discussion

In this combined retro- and prospective study, we performed germline WGS/WES and tumor DNA methylation profiling of a population-based cohort diagnosed nationwide over a timespan of 21 years to determine the

role of genetic predisposition in childhood ependymoma. Both known cancer genes and genes somatically or epigenetically associated with ependymoma were analyzed for pathogenic germline variants, as were evolutionarily constrained genes. Our findings establish ependymoma as a disease where germline pathogenic variants in known cancer genes only rarely play an underlying role, especially when precise molecular (re)classification is available. We also identify new putative ependymoma predisposition genes. Lastly, we highlight the essential role of including molecular tumor classification in ependymoma studies and the feasibility of using archived tumor samples for this purpose.

Pathogenic variants detected in known cancer genes

Of the 37 patients undergoing germline WGS, 11% (4/37) were found to harbor pathogenic variants in the cancer panel genes (*NF1*, *NF2*, *TP53*, and *LZTR1*). By comparison, both the carrier frequency and the genes involved were similar to the findings of Zhang et al. from their pediatric pan-cancer germline study (n=1120) which included 67 ependymoma patients (4/67 (6%), *NF1*, *NF2*, *TP53*) [10]. However, in our cohort, tumor DNA methylation profiling reclassified two of the patients with

pathogenic germline variants in *NF1* and *TP53* to tumor types other than ependymoma. Consequently, only two pathogenic germline variants were detected among children with molecularly confirmed ependymoma (2/34, 5.9%). In this context, it is worth noting that the reclassification rate in our study (7.7%) is comparable to that reported by Capper et al. [26]. In their prospective cohort of 101 histopathologically diagnosed ependymoma samples, 6.0% (6/101) were reclassified based on tumor DNA methylation profiling to a non-ependymoma entity, including neuroepithelial tumors and two DMGs, similarly to our cohort.

As of this writing, ten large (n > 100) pediatric pan-cancer germline sequencing studies including children with ependymoma have been published (Table 2). Combined, these investigations report pathogenic germline variants in 4.7% (9/191) of children with histopathologically diagnosed ependymoma. Following exclusion of a likely benign *TP53* variant (detailed below), three variants likely unrelated to ependymoma (incidental findings from the ACMG v2.0 [34]) and duplicate patients (detailed in Table 2), just 2.9% (5/173) of children with ependymoma are reported to harbor pathogenic germline variants in known cancer genes. Of these, all were

Table 2 Overview of large (> 100 cases) pan-childhood cancer germline sequencing studies with reported findings for ependymoma

Author, jr	Year	Patients w/pathogenic CPS gene variants (n/total (%))			Comments
		Full childhood cancer cohort	CNS subcohort	Ependymoma subcohort	
Zhang, J (NEJM)	2015	95/1120 (8.5%)	21/245 (8.6%)	4/67 (6.0%)	<i>NF1</i> (n = 2), <i>NF2</i> (n = 1) and <i>TP53</i> (n = 1). The latter has later been assessed as likely benign. Limited to intracranial ependymoma
Parsons, DW (JAMA Onc)	2016	13/150 (8.7%)	2/56 (3.6%)	0/9 (0.0%)	
Oberg, JA (Genome Med)	2016	18/101 (17.8%)	5/17 (29.4%)	2/3 (66.7%)	ACMG secondary findings in <i>BRCA1</i> (n = 1) and <i>VHL</i> (n = 1)
Gröbner, SN (Nature)	2018	69/914 (7.6%)	39/542 (7.2%)	0/59 (0.0%)	14 cases are overlapping with Zhang et al. (incl. the patient w the reported <i>TP53</i> variant). Limited to intracranial ependymoma
Wong, N (Nature Med)	2020	40/247 (16.2%)	17/92 (18.5%)	0/8 (0.0%)	
Byrjalsen, A (PLoS Gen)	2020	29/198 (14.7%)	3/44 (6.8%)	0/4 (0.0%)	Ependymoma cases (n = 4) overlap with the current study
Fiala, EM (Nature Can)	2021	138/751 (18.4%)	30/143 (21.0%)	3/14 (21.4%)	<i>NF1</i> (n = 1), <i>NF2</i> (n = 1) and an ACMG secondary finding in <i>FANCA</i> (n = 1)
Newmann, S (Cancer Discovery)	2021	55/300 (18.3%)	19/97 (19.6%)	0/11 (0.0%)	
Stedingk, KV (Sci rep)	2021	30/790 (3.8%)	8/149 (5.4%)	0/14 (0.0%)	Limited to SNV analysis
Wagener, R (EJHG)	2021	11/160 (6.9%)	3/32 (9.4%)	0/2 (0.0%)	
Total		509/4833 (10.5%)	147/1425 (10.3%)	9/191 (4.7%)	
<i>Adjusted total for ependymoma</i>				5/173 (2.9%)	Excl. ACMG secondary findings, 14 cases overlapping in Zhang et al./Gröbner et al. and the four cases reported by Byrjalsen et al. also in the current cohort
<i>Our study</i>				2/34 (5.9%)	Restricted to molecularly confirmed ependymoma
Current best estimate				7/207 (3.4%)	

in *NF1* (n=3) or *NF2* (n=2). This estimate is strikingly similar to our observations, especially when taking into consideration the low frequencies and sample size and the fact that the gene panels used in the majority of the previous studies did not include *LZTR1*.

Neurofibromatosis type-2 predisposes both to intraspinal and -cranial childhood ependymoma

The association between neurofibromatosis type-2 and spinal ependymoma is well established [35] and somatic *NF2* variants are recurrently altered in ependymomas with intraspinal location [36]. Yet, several cases of intracranial ependymoma (especially located to the cervicomedullary junction) have been reported in children and young adults with neurofibromatosis type-2 [37–41]. Combined with our findings of a cervicomedullary located ependymoma in a child with a pathogenic germline *NF2* variant, there is mounting evidence that germline *NF2*-related ependymomas may be located intracranially, as well as intraspinally. While the former will often represent SP-EPN located in or around the cervicomedullary junction, cases of PF-EPN-B ependymoma have also been reported [37].

Still, pathogenic germline *NF2* variants are relatively rare in the overall pediatric ependymoma population and thus explain only a minority of cases: Among the 173 children with ependymoma included in the reviewed pan-childhood cancer germline sequencing studies [10–19], only two patients (1.2%) were reported to harbor pathogenic *NF2* alterations [10, 16] (Table 2), for whom neither tumor location nor molecular subclass were described.

Questioning Li-Fraumeni Syndrome's association with (molecularly classified) ependymoma

Both somatic and germline *TP53* variants have been reported in other pediatric CNS tumors, yet such alterations are extremely rare in ependymoma tumor tissue [42]. Of all the children with ependymoma included in the aforementioned germline predisposition investigations, only one patient (0.6%, 1/173) was found to carry a *TP53* variant characterized as pathogenic [10]. The variant (NM_000546:p.Tyr107His, c.319T>C), which was detected in a 10-year-old girl with an infratentorial ependymoma, has later been classified as benign in ClinVar [43] and was not reported as pathogenic by Gröbner et al., who included the same patient in their subsequent study [13]. Furthermore, the variant has been found in 0.1% of healthy adults that self-identified as African/African American [24]. Apart from the 173 children with ependymoma reviewed above, five cases of children with ependymoma and pathogenic germline *TP53* variants have been reported in the literature [44–46]. Of

note, molecular tumor classification was not performed in any of these cases. Were it not for DNA methylation profiling-based reclassification to DMG, the erroneous ependymoma phenotype in our cohort would have been reported as associated with the germline *TP53* variant. This underscores the importance of molecular classification of ependymal tumors.

Pathogenic *NF1* germline variants also appear to play a role in childhood ependymoma

Pathogenic *NF1* germline variants are extremely rare among children with ependymoma. No such variants were detected among the 34 children with molecularly classified ependymoma following diagnostic revision to RGNT for the child with a nonsense variant in *NF1*. In comparison, three of the reported 173 germline sequenced children with ependymoma (1.7%) have been found to carry pathogenic *NF1* variants (Table 2). These include two children with intracranial ependymoma reported by Zhang et al. [10] and one 6-year-old child with synchronous schwannoma and CNS ependymoma reported by Fiala et al. [16]. Only two additional cases of children with (clinically) diagnosed neurofibromatosis type-1 and intracranial ependymoma have been reported in the literature [47]. Diagnostic confirmation and tumor molecular subtyping by DNA methylation profiling was not reported for any of these patients. This may have inflated the reported *NF1* carrier rate in patients with ependymoma. This phenomenon is illustrated by the diagnostic revision in both our cohort and others, where histopathologically diagnosed ependymomas were reclassified to pilocytic astrocytomas and neuroepithelial tumors based on DNA methylation profiling [26]. Importantly, both of these tumor types have a much higher rate of germline *NF1* alterations [48, 49].

LZTR1 might represent a novel putative ependymoma predisposition gene

A likely pathogenic *LZTR1* variant (p.Gln762Ter [c.2284C>T]), undetected among >125,000 healthy adult in gnomAD [24], was found in a child diagnosed with a fourth ventricle PF-EPN-A1c ependymoma. Pathogenic germline variants in *LZTR1* have not previously been reported in patients with ependymoma. The gene, which is centromeric to *NF2* and *SMARCB1* on chromosome 22q11.21, was recently uncovered as a germline predisposition gene in schwannomatosis [50]. Pathogenic *LZTR1* germline variants have been reported in children with different cancer types, including high-grade glioma [13], but have not been evaluated in the majority of the existing large pan-childhood cancer germline sequencing studies [10, 11, 16–18]. Although monozygosity of 22q has been reported in ~40% of *RELA*-fusion positive

supratentorial ependymoma (ST-EPN-RELA) [52], the rarity of pathogenic somatic *NF2* variants in the majority of intracranial ependymoma suggests a different tumor suppressor gene to be located on chromosome 22 [51, 53, 54]. We therefore speculate that pathogenic germline *LZTR1* variants may play a role in tumorigenesis for a limited subset of children with ependymoma, perhaps restricted to the PF-EPN-A1c molecular subtype.

Upon review of *LZTR1* findings in our childhood (non-ependymoma) cancer control cohort, the *LZTR1* missense VUS (p.Asp703Asn [c.2107G>A]) detected in another patient with PF-EPN-A1c was observed in a child with acute myeloid leukemia. Moreover, this variant has been reported in 5/26,128 (0.02%) Swedish individuals reported without serious childhood disease in gnomAD [24].

Less than 4% of childhood ependymoma is explained by pathogenic variants in known cancer genes

Based on the described meta-analysis, the current best estimate of germline predisposition in childhood ependymoma suggests that 3.4% (7/207) carry a causative pathogenic germline variant, mainly located in *NF2* and *NFI* (Fig. 2). This estimate indicates that germline predisposition is significantly less frequent than what is reported for pediatric brain and spinal cord tumors in general [3] (Fisher's exact test, 7/207 vs. 147/1425, OR=0.30 [0.11–0.66], $p < 0.001$). Consequently, one may question the need to perform extensive genetic testing in newly diagnosed children with ependymoma if no family history or other signs or symptoms of neurofibromatosis are present. Of course, the lack of germline findings in the majority of children with ependymoma may reflect limitations in our current knowledge of genetics. Or, perhaps more likely, other biological mechanisms including epigenetic dysregulation, which has been suggested as the main driver for the largest molecular subgroup, PF-EPN-A [55, 56].

Several factors may, however, have influenced the validity of the combined risk estimate; Opposite to our study, all but one of the germline investigations listed in Table 2 did not report molecular tumor classification [12]. Moreover, their lack of population-based study design may have introduced selection bias. As illustrated by the pathogenic *NF2* deletion detected in our cohort, limiting bioinformatic analyses solely to SNVs, as done in one of the reviewed sequencing studies [18], may miss pathogenic alterations. Also affecting the generalizability of the combined estimate is the fact that the two cohorts contributing 65% (112/173) of the total ependymoma sample size were limited to intracranial ependymoma, likely resulting in underreporting of *NF2*-associated cases [10, 13].

Constrained gene analysis may explain additional genetic risk

Focusing on genes exhibiting evolutionary intolerance of inactivating alterations has recently emerged as a novel approach of investigating genetic predisposition to any state that limits reproduction, such as fatal childhood diseases [24]. We have previously detailed how a constrained gene approach may be useful in investigations of genetic predisposition to childhood (CNS) malignancies [21].

Constrained gene analysis of children with molecularly confirmed ependymoma rediscovered the *NF2* deletion detected in our cancer gene panel analysis. Apart from *NF2*, none of the 14 constrained genes in which pLoF variants were detected have previously been linked with ependymoma. Interestingly, several are suggested to have tumor suppressor roles (*FOXO3* [57], *TRIM67* [58], *UHRF2* [59, 60], *CHD6* [61]).

As no single gene was found to harbor pLoF variants in more than one patient, further research of the concept is needed before a common or broader role for constrained genes in ependymoma predisposition can be ascertained. In our cohort, the lack of consistent constrained gene findings likely reflects the limited sample size and its subtype heterogeneity, or, alternatively, the growing notion that PF-EPN-A is an epigenetically driven disease. In this context, it is worth mentioning that two of the constrained genes, in which pLoF variants were detected in children with PF-EPN-A, affect epigenetic gene expression control (*UHRF2* [60, 62] and *DNAJC2* [63]). As neither the detected constrained genes nor *LZTR1* have been analyzed in the majority of the aforementioned pediatric pan-cancer germline sequencing studies, their inclusion in future larger ependymoma cohorts will be important to confidently suggest any disease-related roles and indication for further study.

Strengths and limitations

Key strengths of this study include its population-based design, high inclusion rates and molecular tumor classification based on DNA methylation profiling. Moreover, our germline WGS-based SNV and SV and WES SNV analysis included not only 390 known cancer genes, but also 67 other genes with implied roles in ependymoma tumorigenesis and constrained gene analysis. The comprehensive literature review-based meta-analysis further strengthens the value of our investigation.

However, even with a nationwide inclusion period of more than 20 years, our sample size limits generalizability of the observed carrier frequencies. Tumor and germline tissue were unavailable for four and six patients, respectively. Finally, the use of a non-ependymoma childhood

cancer control cohort in the filtering of germline variants might have affected variant filtration in a conservative direction. Optimally, an equal or larger control cohort of representative and ethnically comparable whole-genome sequenced children would have been available.

In summary

This population-based germline sequencing study of childhood ependymoma, including constrained gene analysis, establishes that genetic predisposition plays a role for less than 4% of patients. This is significantly lower than for pediatric CNS tumors in general. Moreover, we show that pathogenic germline variants in children with ependymoma are virtually restricted to *NF2* and *NF1*. Our results emphasize the importance of molecular tumor classification, as the likelihood of diagnostic reclassification to a non-ependymoma tumor was significantly higher for children with detected pathogenic germline variants. We therefore advocate diagnostic reconsideration in children with non-molecularly classified ependymoma with cancer predisposition syndromes other than neurofibromatosis type-2. In addition, we present *LZTR1* as a novel putative ependymoma predisposition gene.

Supplementary Information

The online version contains supplementary material available at <https://doi.org/10.1186/s40478-022-01429-1>.

Additional file 1. Redefining germline predisposition in children with molecularly characterized ependymoma: a population-based 20-year cohort. **Figure S1.** Flowchart illustrating the filtering of single nucleotide germline variants in 37 children with histopathologically diagnosed ependymoma. **Figure S2.** Flowchart illustrating the filtering of structural germline variants in 37 children with histopathologically diagnosed ependymoma. **Figure S3.** Overview of the inclusion process and germline tissue availability.

Additional file 2: Table S1. Overview of the 390 genes associated with cancer included in the panel analysis. **Table S2.** Genes reported with potential germline/somatic role in ependymoma. **Table S3.** Constrained genes manual curation results. **Table S4.** Cohort overview: clinical, histopathological and molecular data. **Table S5.** Overview of pathogenic variants in known cancer associated genes and pLoF variants in evolutionarily constrained genes by molecular tumor type. **Table S6.** Five heterozygous loss-of-function variants in recessive genes considered unrelated to ependymoma.

Acknowledgements

Jeanette Krogh Petersen, Marianne Schmidt, Henning Boldt and Benedicte Parm Ullhøi for retrieval of archived tumor tissue and iDAT files.

Author contributions

JF-S, UKS, RM, KW, JS-R & KS: conception, design. JF-S, UKS, SMAK, RM, JS-R, ES, SRY, AOL, LCM, DS, KW: patient inclusion, sample and/or data acquisition. JF-S, UKS, ES, SRO, TOvH, KW, KS, RM, LBA, LCM, DS, BWK, AOL: data interpretation. JF-S, UKS: drafting of manuscript. All authors: manuscript revision and approval. All authors have made substantial contributions to the manuscript and approved its final version.

Funding

This work is part of Childhood Oncology Network Targeting Research, Organization & Life expectancy (CONTROL) and supported by the Danish Cancer Society (R-257-A14720), the Danish Childhood Cancer Foundation (2019-5934), the Danish Childhood Brain Tumor Foundation and the European Union's Interregional Oresund–Kattegat–Skagerrak Grant.

Availability of data and materials

All data produced in the present work are contained in the manuscript with the exception of genetic sequencing data. Danish legal regulation does not permit uploading of raw sequencing data. Selected data may be made available upon reasonable request (dependent on required approvals from relevant scientific ethic boards) to the authors.

Declarations

Ethics approval and consent to participate

This study was approved by the Capital Region Scientific Ethical Committee (H-15016782, prospective cohort) and the Danish National Committee on Health Research Ethics (2000407). All patients and/or parents/legal guardians provided informed consent.

Competing interests

The authors declare no competing interests.

Author details

¹Department of Pediatrics and Adolescent Medicine, Rigshospitalet University Hospital, Copenhagen, Denmark. ²Department of Neurosurgery, Rigshospitalet University Hospital, Copenhagen, Denmark. ³Department of Clinical Genetics, University of Copenhagen, Copenhagen, Denmark. ⁴Department of Clinical Medicine, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. ⁵Department of Genomic Medicine, Rigshospitalet University Hospital, Copenhagen, Denmark. ⁶Department of Clinical Immunology, Rigshospitalet University Hospital, Copenhagen, Denmark. ⁷Department of Pathology, Rigshospitalet University Hospital, Copenhagen, Denmark. ⁸Biotech Research and Innovation Center, University of Copenhagen, Copenhagen, Denmark. ⁹Department of Health Technology, Cancer Systems Biology and Bioinformatics, Technical University of Denmark, Lyngby, Denmark. ¹⁰Department of Neurosurgery, Section 6031, Rigshospitalet University Hospital, Inge Lehmanns Vej 6, 2100 Copenhagen, Denmark. ¹¹The Pediatric Oncology Research Laboratory, Section 5704, Department of Pediatrics and Adolescent Medicine, Rigshospitalet University Hospital, Henrik Harpestrengs Vej 6A, 2100 Copenhagen, Denmark.

Received: 2 June 2022 Accepted: 11 August 2022

Published online: 25 August 2022

References

- Ostrom QT, De Blank PM, Kruchko C et al (2014) Alex's Lemonade stand foundation infant and childhood primary brain and central nervous system tumors diagnosed in the United States in 2007–2011. *Neuro-Oncology* 16:x1–x35. <https://doi.org/10.1093/neuonc/nou327>
- Marinoff AE, Ma C, Guo D et al (2017) Rethinking childhood ependymoma: a retrospective, multi-center analysis reveals poor long-term overall survival. *J Neurooncol* 135(1):201–211. <https://doi.org/10.1007/s11060-017-2568-8>
- Muskens IS, Zhang C, de Smith AJ, Biegel JA, Walsh KM, Wiemels JL (2019) Germline genetic landscape of pediatric central nervous system tumors. *Neuro-Oncology* 21(11):1376–1388. <https://doi.org/10.1093/neuonc/noz108>
- Zhang C, Ostrom QT, Semmes EC et al (2020) Genetic predisposition to longer telomere length and risk of childhood, adolescent and adult-onset ependymoma. *Acta Neuropathol Commun* 8(1):173. <https://doi.org/10.1186/s40478-020-01038-w>
- Hemminki K, Tretli S, Sundquist J, Johannesen TB, Granström C (2009) Familial risks in nervous-system tumours: a histology-specific analysis from Sweden and Norway. *Lancet Oncol* 10(5):481–488. [https://doi.org/10.1016/S1470-2045\(09\)70076-2](https://doi.org/10.1016/S1470-2045(09)70076-2)

6. Chen BY, Praeger A, Christie M, Yuen T (2020) Familial intracranial ependymoma mimicking an extra-lesion: a case report and review of the literature. *J Clin Neurosci* 74:250–253. <https://doi.org/10.1016/j.jocn.2020.01.051>
7. Dimopoulos VG, Fountas KN, Robinson JS (2006) Familial intracranial ependymomas. Report of three cases in a family and review of the literature. *Neurosurg Focus* 20(1):E8. <https://doi.org/10.3171/foc.2006.20.1.9>
8. Zhang C, Ostrom QT, Hansen HM et al (2020) European genetic ancestry associated with risk of childhood ependymoma. *Neuro-Oncology* 22(11):1637–1646. <https://doi.org/10.1093/neuonc/naaa130>
9. Johnson KJ, Cullen J, Barnholtz-Sloan JS et al (2014) Childhood brain tumor epidemiology: a brain tumor epidemiology consortium review. *Cancer Epidemiol Prev Biomark* 23(12):2716–2736. <https://doi.org/10.1158/1055-9965.EPI-14-0207>
10. Zhang J, Walsh MF, Wu G et al (2015) Germline mutations in predisposition genes in pediatric cancer. *N Engl J Med* 373(24):2336–2346. <https://doi.org/10.1056/NEJMoa1508054>
11. Parsons DW, Roy A, Yang Y et al (2016) Diagnostic yield of clinical tumor and germline whole-exome sequencing for children with solid tumors. *JAMA Oncol* 2(5):616–624. <https://doi.org/10.1001/jamaoncol.2015.5699>
12. Oberg JA, Glade Bender JL, Sulis ML et al (2016) Implementation of next generation sequencing into pediatric hematology-oncology practice: moving beyond actionable alterations. *Genome Med* 8:133. <https://doi.org/10.1186/s13073-016-0389-6>
13. Gröbner SN, Worst BC, Weischenfeldt J et al (2018) The landscape of genomic alterations across childhood cancers. *Nature* 555(7696):321–327. <https://doi.org/10.1038/nature25480>
14. Wong M, Mayoh C, Lau LMS et al (2020) Whole genome, transcriptome and methylome profiling enhances actionable target discovery in high-risk pediatric cancer. *Nat Med* 26(11):1742–1753. <https://doi.org/10.1038/s41591-020-1072-4>
15. Byrjalsen A, Hansen TVO, Stoltze UK et al (2020) Nationwide germline whole genome sequencing of 198 consecutive pediatric cancer patients reveals a high frequency of cancer prone syndromes. *PLoS Genet*. <https://doi.org/10.1371/JOURNAL.PGEN.1009231>
16. Fiala EM, Jayakumaran G, Mauguén A et al (2021) Prospective pan-cancer germline testing using MSK-IMPACT informs clinical translation in 751 patients with pediatric solid tumors. *Nat Cancer* 2(3):357–365. <https://doi.org/10.1038/s43018-021-00172-1>
17. Newman S, Nakitandwe J, Kesserwan CA et al (2021) Genomes for kids: the scope of pathogenic mutations in pediatric cancer revealed by comprehensive DNA and RNA sequencing. *Cancer Discov*. <https://doi.org/10.1158/2159-8290.CD-20-1631>
18. von Stedingk K, Stjernfeldt KJ, Kvist A et al (2021) Prevalence of germline pathogenic variants in 22 cancer susceptibility genes in Swedish pediatric cancer patients. *Sci Rep* 11(1):5307. <https://doi.org/10.1038/s41598-021-84502-4>
19. Wagener R, Tæubner J, Walter C et al (2021) Comprehensive germline-genomic and clinical profiling in 160 unselected children and adolescents with cancer. *Eur J Hum Genet* 29(8):1301–1311. <https://doi.org/10.1038/s41431-021-00878-x>
20. Schrøder H, Rechner C, Wehner PS et al (2016) Danish childhood cancer registry. *Clin Epidemiol* 8:461–464. <https://doi.org/10.2147/CLEP.S99508>
21. Stoltze UK, Foss-Skiftesvik J, van Overeem Hansen T et al (2022) Genetic predisposition & evolutionary traces of pediatric cancer risk: a prospective 5-year population-based genome sequencing study of children with CNS tumors. *Neuro-Oncology*. <https://doi.org/10.1093/neuonc/naoc187>
22. Sørensen E, Christiansen L, Wilkowski B et al (2021) Data resource profile: the Copenhagen Hospital Biobank (CHB). *Int J Epidemiol* 50(3):719–720e. <https://doi.org/10.1093/ije/dyaa157>
23. Richards S, Aziz N, Bale S et al (2015) Standards and guidelines for the interpretation of sequence variants: A joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology. *Genet Med*. <https://doi.org/10.1038/gim.2015.30>
24. Karczewski KJ, Francioli LC, Tiao G et al (2020) The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* 581(7809):434–443. <https://doi.org/10.1038/s41586-020-2308-7>
25. Abramovs N, Brass A, Tassabehji M (2020) GeVIR is a continuous gene-level metric that uses variant distribution patterns to prioritize disease candidate genes. *Nat Genet* 52(1):35–39. <https://doi.org/10.1038/s41588-019-0560-2>
26. Capper D, Jones DTW, Sill M et al (2018) DNA methylation-based classification of central nervous system tumours. *Nature*. <https://doi.org/10.1038/nature26000>
27. Elsamadicy AA, Koo AB, David WB et al (2020) Comparison of epidemiology, treatments, and outcomes in pediatric versus adult ependymoma. *Neuro-Oncol Adv* 2(1):vdaa019. <https://doi.org/10.1093/oaajnl/vdaa019>
28. Villano JL, Parker CK, Dolecek TA (2013) Descriptive epidemiology of ependymal tumours in the United States. *Br J Cancer* 108(11):2367–2371. <https://doi.org/10.1038/bjc.2013.221>
29. McGuire CS, Sainani KL, Fisher PG (2009) Incidence patterns for ependymoma: a surveillance, epidemiology, and end results study: clinical article. *J Neurosurg* 110(4):725–729. <https://doi.org/10.3171/2008.9.JNS08117>
30. Szklarczyk D, Gable AL, Lyon D et al (2019) STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets. *Nucleic Acids Res* 47(D1):D607–D613. <https://doi.org/10.1093/nar/gky1131>
31. Fagerberg L, Hallström BM, Oksvold P et al (2014) Analysis of the human tissue-specific expression by genome-wide integration of transcriptomics and antibody-based proteomics. *Mol Cell Proteomics* 13(2):397–406. <https://doi.org/10.1074/mcp.M113.035600>
32. Duff MO, Olson S, Wei X et al (2015) Genome-wide identification of zero nucleotide recursive splicing in drosophila. *Nature* 521(7552):376–379. <https://doi.org/10.1038/nature14475>
33. Gene Ontology Consortium (2021) The gene ontology resource: enriching a GOLD mine. *Nucleic Acids Res* 49(D1):D325–D334. <https://doi.org/10.1093/nar/gkaa1113>
34. Kalia SS, Adelman K, Bale SJ et al (2017) Recommendations for reporting of secondary findings in clinical exome and genome sequencing, 2016 update (ACMG SF v2.0): a policy statement of the American College of Medical Genetics and Genomics. *Genet Med* 19(2):249–255. <https://doi.org/10.1038/gim.2016.190>
35. Coy S, Rashid R, Stemmer-Rachamimov A, Santagata S (2020) An update on the CNS manifestations of neurofibromatosis type 2. *Acta Neuropathol (Berl)* 139(4):643–665. <https://doi.org/10.1007/s00401-019-02029-5>
36. Ebert C, von Haken M, Meyer-Puttlitz B et al (1999) Molecular genetic analysis of ependymal tumors. NF2 mutations and chromosome 22q loss occur preferentially in intramedullary spinal ependymomas. *Am J Pathol* 155(2):627–632. [https://doi.org/10.1016/S0002-9440\(10\)65158-9](https://doi.org/10.1016/S0002-9440(10)65158-9)
37. Kresbach C, Dorostkar MM, Suwala AK et al (2021) Neurofibromatosis type 2 predisposes to ependymomas of various localization, histology, and molecular subtype. *Acta Neuropathol (Berl)* 141(6):971–974. <https://doi.org/10.1007/s00401-021-02304-4>
38. Ruggieri M, Praticò AD, Serra A et al (2016) Childhood neurofibromatosis type 2 (NF2) and related disorders: from bench to bedside and biologically targeted therapies. *Acta Otorhinolaryngol Ital* 36(5):345–367. <https://doi.org/10.14639/0392-100X-1093>
39. Wheeler A, Metrock K, Li R, Singh S (2022) Cystic meningoangiomas and cerebellar ependymoma in a child with neurofibromatosis type 2. *Radiol Case Rep* 17(4):1082–1087. <https://doi.org/10.1016/j.radcr.2022.01.050>
40. Halliday D, Emmanouil B, Vassallo G et al (2019) Trends in phenotype in the English paediatric neurofibromatosis type 2 cohort stratified by genetic severity. *Clin Genet* 96(2):151–162. <https://doi.org/10.1111/cge.13551>
41. Essayed WI, Bernard A, Kalamirides M (2015) Clinical response associated with radiographic regression of a cervicomedullary ependymoma in a NF2 patient treated by bevacizumab. *J Neurooncol* 125(2):445–446. <https://doi.org/10.1007/s11060-015-1925-8>
42. AACR Project GENIE (2017) Powering precision medicine through an International Consortium. *Cancer Discov* 7(8):818–831. <https://doi.org/10.1158/2159-8290.CD-17-0151>
43. Landrum MJ, Lee JM, Benson M et al (2018) ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* 46(D1):D1062–D1067. <https://doi.org/10.1093/nar/gkx1153>

44. Bougeard G, Renaux-Petel M, Flaman JM et al (2015) Revisiting Li-Fraumeni Syndrome from TP53 mutation carriers. *J Clin Oncol* 33(21):2345–2352. <https://doi.org/10.1200/JCO.2014.59.5728>
45. Metzger AK, Sheffield VC, Duyk G, Daneshvar L, Edwards MS, Cogen PH (1991) Identification of a germ-line mutation in the p53 gene in a patient with an intracranial ependymoma. *Proc Natl Acad Sci U S A* 88(17):7825–7829. <https://doi.org/10.1073/pnas.88.17.7825>
46. Hosoya T, Kambe A, Nishimura Y, Sakamoto M, Maegaki Y, Kurosaki M (2018) Pediatric case of Li-Fraumeni syndrome complicated with supratentorial anaplastic ependymoma. *World Neurosurg* 120:125–128. <https://doi.org/10.1016/j.wneu.2018.08.203>
47. Riffaud L, Vinchon M, Ragragui O, Delestret I, Ruchoux MM, Dhellemmes P (2002) Hemispheric cerebral gliomas in children with NF1: arguments for a long-term follow-up. *Childs Nerv Syst* 18(1–2):43–47. <https://doi.org/10.1007/s00381-001-0534-3>
48. Fisher MJ, Jones DTW, Li Y et al (2021) Integrated molecular and clinical analysis of low-grade gliomas in children with neurofibromatosis type 1 (NF1). *Acta Neuropathol (Berl)* 141(4):605–617. <https://doi.org/10.1007/s00401-021-02276-5>
49. Sievers P, Appay R, Schrimpf D et al (2019) Rosette-forming glioneuronal tumors share a distinct DNA methylation profile and mutations in FGFR1, with recurrent co-mutation of PIK3CA and NF1. *Acta Neuropathol (Berl)* 138(3):497–504. <https://doi.org/10.1007/s00401-019-02038-4>
50. Piotrowski A, Xie J, Liu YF et al (2014) Germline loss-of-function mutations in LZTR1 predispose to an inherited disorder of multiple schwannomas. *Nat Genet* 46(2):182–187. <https://doi.org/10.1038/ng.2855>
51. Ebert C, von Haken M, Meyer-Puttitz B et al (1999) Molecular genetic analysis of ependymal tumors. *Am J Pathol* 155(2):627–632
52. Pajtler KW, Witt H, Sill M et al (2015) Molecular classification of ependymal tumors across all CNS compartments, histopathological grades, and age groups. *Cancer Cell* 27(5):728–743. <https://doi.org/10.1016/j.ccell.2015.04.002>
53. Singh PK, Gutmann DH, Fuller CE, Newsham IF, Perry A (2002) Differential involvement of protein 4.1 family members DAL-1 and NF2 in intracranial and intraspinal ependymomas. *Mod Pathol Off J U S Can Acad Pathol Inc* 15(5):526–531. <https://doi.org/10.1038/modpathol.3880558>
54. Slavc I, MacCollin MM, Dunn M et al (1995) Exon scanning for mutations of the NF2 gene in pediatric ependymomas, rhabdoid tumors and meningiomas. *Int J Cancer* 64(4):243–247. <https://doi.org/10.1002/ijc.2910640406>
55. Mack SC, Witt H, Piro RM et al (2014) Epigenomic alterations define lethal CLMP-positive ependymomas of infancy. *Nature* 506(7489):445–450. <https://doi.org/10.1038/nature13108>
56. Stuckert A, Bertrand KC, Wang P, Smith A, Mack SC (2020) Weighing ependymoma as an epigenetic disease. *J Neurooncol* 150(1):57–61. <https://doi.org/10.1007/s11060-020-03562-0>
57. Jiramongkol Y, Lam EWF (2020) FOXO transcription factor family in cancer and metastasis. *Cancer Metastasis Rev* 39(3):681–709. <https://doi.org/10.1007/s10555-020-09883-w>
58. Wang S, Zhang Y, Huang J et al (2019) TRIM67 activates p53 to suppress colorectal cancer initiation and progression. *Cancer Res* 79(16):4086–4098. <https://doi.org/10.1158/0008-5472.CAN-18-3614>
59. Lu H, Bhoopatiraju S, Wang H et al (2016) Loss of UHRF2 expression is associated with human neoplasia, promoter hypermethylation, decreased 5-hydroxymethylcytosine, and high proliferative activity. *Oncotarget* 7(46):76047–76061. <https://doi.org/10.18632/oncotarget.12583>
60. Mori T, Ikeda DD, Yamaguchi Y, Unoki M, NIRF Project (2012) NIRF/UHRF2 occupies a central position in the cell cycle network and allows coupling with the epigenetic landscape. *FEBS Lett* 586(11):1570–1583. <https://doi.org/10.1016/j.febslet.2012.04.038>
61. Egan CM, Nyman U, Skotte J et al (2013) CHD5 is required for neurogenesis and has a dual role in facilitating gene expression and polycomb gene repression. *Dev Cell* 26(3):223–236. <https://doi.org/10.1016/j.devcel.2013.07.008>
62. Pichler G, Wolf P, Schmidt CS et al (2011) Cooperative DNA and histone binding by UHRF2 links the two major repressive epigenetic pathways. *J Cell Biochem* 112(9):2585–2593. <https://doi.org/10.1002/jcb.23185>
63. Richly H, Rocha-Viegas L, Ribeiro JD et al (2010) Transcriptional activation of polycomb-repressed genes by ZRF1. *Nature* 468(7327):1124–1128. <https://doi.org/10.1038/nature09574>

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Ready to submit your research? Choose BMC and benefit from:

- fast, convenient online submission
- thorough peer review by experienced researchers in your field
- rapid publication on acceptance
- support for research data, including large and complex data types
- gold Open Access which fosters wider collaboration and increased citations
- maximum visibility for your research: over 100M website views per year

At BMC, research is always in progress.

Learn more biomedcentral.com/submissions



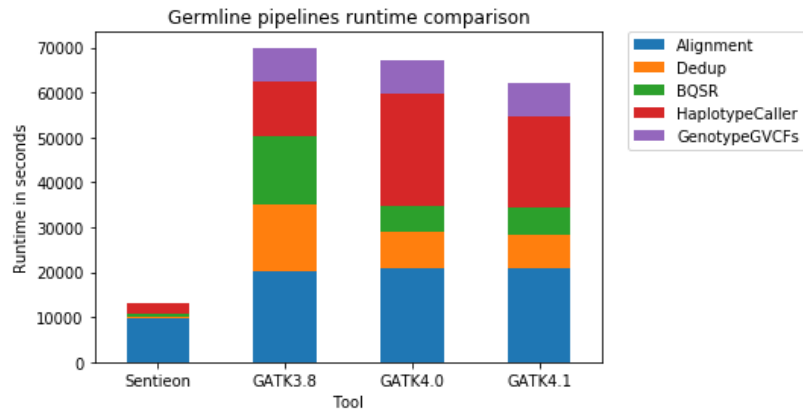
Epilogue

Due to the nature of the project, where large volumes of data need to be processed, stored, and interpreted, a combination of very efficient software and hardware along with a meticulous data management plan is required.

[Using Sentieon in Computerome for childhood cancer research](#)

As mentioned previously, in this project we used Sentieon software which is built upon GATK's foundation to create a more efficient and accelerated solution while achieving the same accuracy scores as the GATK Best Practices (Genome Analysis Toolkit) which is considered "Best Practices" because it represents a set of well-established guidelines and methodologies for analyzing NGS data. It has undergone extensive method development, validation, and refinement by a team of experts at the Broad Institute, a renowned research institution, and is widely recognized and adopted within the genomics community. The methods and algorithms implemented in GATK have been thoroughly tested and validated using various datasets and benchmarking approaches.

Sentieon's optimization of GATK involves several key aspects such as speed, scalability, memory efficiency and compatibility (Figure 9). Sentieon's software is designed to efficiently utilize computational resources, including multi-core processors and HPC environments. This scalability ensures that the analysis can be seamlessly adapted to handle large-scale genomic datasets. Furthermore, it has made optimizations to reduce memory usage during data processing, making it more memory-efficient compared to the standard GATK pipeline. This optimization allows for the analysis of larger datasets without excessive memory requirements. Sentieon has also implemented highly optimized algorithms and parallel computing techniques to significantly accelerate the processing and analysis of NGS data.



Stage	Sentieon	GATK3.8	GATK4.0	GATK4.1
Alignment	2:42:44	5:38:35	5:49:39	5:45:39
Dedup	0:06:16	4:04:25	2:11:43	2:06:32
BQSR	0:10:10	4:17:09	1:39:57	1:40:06
HaplotypeCaller	0:41:02	3:21:37	6:56:53	5:37:52
GenotypeGVCFs	0:00:55	2:04:08	2:02:55	2:05:22
Total	3:41:07	19:25:54	18:41:07	17:15:31
Sentieon SpeedUp	--	5.3X	5.1X	4.7X

Type	TRUTH			QUERY		METRIC		
	TOTAL	TP	FN	TOTAL	FP	Recall	Precision	F1_Score
INDEL	855716	850790	4926	894426	10869	0.994243	0.987848	0.991035
SNP	3999272	3990379	8893	4006624	11826	0.997776	0.997048	0.997412

Figure 9. Sentieon vs GATK performance summary (github.com/Sentieon/sentieon-dnaseq). Specs of the test: Google Compute Engine with n1-standard-32 (32 vCPUs, 120 GB memory), Local SSD Scratch Disk 2x375G and centos-7-v20190619.

All this optimization allows for faster execution times, enabling quicker turnaround in data analysis achieving the same precision and accuracy as GATK “Best Practices”. The optimized performance of

Sentieon software enables researchers and clinicians to analyze genomic data more rapidly and efficiently, making it particularly valuable in large-scale projects and time-sensitive clinical setting.

Performing genomic analysis with Sentieon in Computerome offers distinct advantages due to the cluster's robust infrastructure, which includes a high number of core processors and high-speed memory. Sentieon, known for its optimized and accelerated algorithms, capitalizes on Computerome's capabilities to deliver efficient and rapid analysis of large-scale genomic datasets. The abundance of core processors enables parallelization of computational tasks, allowing for concurrent processing of multiple samples and reducing analysis time significantly. Additionally, the high-speed memory ensures quick data access and processing, facilitating seamless execution of Sentieon's bioinformatics workflows. The synergy between Sentieon and Computerome empowers researchers to leverage the full potential of genomic analysis, enabling insights into genetic variants, somatic mutations, and structural variations with enhanced precision and efficiency.

For instance, *TNscope* encompasses several advancements in its mathematical model to augment recall and precision for somatic variations. The utilization of Sentieon tools allows for lower active region triggering thresholds, facilitating a comprehensive assessment of potential variation sites. Moreover, the use of statistical criteria to initiate active regions rather than employing rigid cutoffs results in higher-quality active region identification. Local assembly methods are refined to more frequently identify the correct variant haplotype. Genotyping accuracy is improved through the introduction of a novel quality score and modified nonparametric statistical tests to filter out false-positive variant candidates. Furthermore, *DNAscope* uniquely combines the well-established methods from haplotype-based variant callers with machine learning to achieve improved accuracy. As a successor to GATK HaplotypeCaller, *DNAscope* uses a similar logical architecture, but introduces improvements to active region detection and local assembly for improved sensitivity and robustness, especially across high-complexity regions. When a machine learning model is applied, *DNAscope* outputs candidate variants with additional informative annotations. Annotated variant candidates are then passed to a machine learning model for variant genotyping, resulting in improvements in both calling and genotyping accuracy.

[NGS Data management challenges in large-scale studies](#)

The volume of data generated by NGS technologies is substantial, emphasizing the need for meticulous organization to reduce storage costs while ensuring data scalability, accessibility, complete traceability, and consistency. This aspect has been crucial in this project. Due to the nature of this type of studies where new data is generated constantly, the data management strategy should allow the data to grow

while keeping consistency with older data. All the data needs to be stored in an accessible manner and other researchers should be able to use it. Furthermore, ensuring complete traceability of the analysis performed enables researchers to track and validate the results, supporting scientific rigor and data integrity.

Conclusion

In summary, this PhD thesis has contributed to the main goal of iCOPE, improving the clinical life of the children with cancer, by providing a set of NGS pipelines matching best practices which allowed to generate datasets of genetic variation in Danish children with cancer that served as input for different research projects such as the manuscripts presented in this thesis. **Manuscript 1** made use of the pipelines and the data generated by the pipelines created during this PhD thesis and demonstrated the significance of implementing systematic screening for Cancer Predisposing Syndromes (CPSs) in pediatric cancer patients, revealing a notable association between a larger proportion of childhood cancers and underlying predisposing germline variants. **Manuscript 2** includes the genomic analysis of the first clinical case of a family with a rare subtype of leukemia (Ph+ALL). No clear genomic driver of Ph+ ALL was identified, however, due to the singularity of the case the reported variants and findings could provide valuable insights for clinicians and researchers, offering potential avenues for further investigation on this rare subtype of leukemia. **Manuscript 3** utilized germline variants of unknown significance in *ERCC4*, *BLM*, *FANCA*, *FANCE*, *FANCF*, *FANCG*, *FANCI*, *FANCL*, *MLH1*, *MSH2*, *MSH6*, *NBN*, *RAD51C*, and *RFWD3* identified by our pipeline as a case study to prove the validity of RosettaDDG tool for prediction of free energy changes upon amino acid substitutions. Lastly, genetic variants obtained from samples diagnosed with ependymoma were used to identify germline SNV and help redefine the germline predisposition of molecularly characterized ependymoma (**Manuscript 4**).

Bibliography

1. Vassal G, Schrappe M, Pritchard-Jones K, Arnold F, Basset L, Biondi A, et al. The SIOPE strategic plan: A European cancer plan for children and adolescents. *J Cancer Policy*. 2016;8:17–32.
2. Wolfe I, Thompson M, Gill P, Tamburlini G, Blair M, Van Den Bruel A, et al. Health services for children in western Europe. *The Lancet*. 2013;381(9873):1224–34.
3. Grabas MR, Kjaer SK, Frederiksen MH, Winther JF, Erdmann F, Dehlendorff C, et al. Incidence and time trends of childhood cancer in Denmark, 1943–2014. *Acta Oncol (Madr)*. 2020;59(5):588–95.
4. Toft N, Birgens H, Abrahamsson J, Griškevičius L, Hallböök H, Heyman M, et al. Results of NOPHO ALL2008 treatment for patients aged 1-45 years with acute lymphoblastic leukemia. *Leukemia*. 2018;32(3):606–15.
5. Erdmann F, Frederiksen LE, Bonaventure A, Mader L, Hasle H, Robison LL, et al. Childhood cancer: Survival, treatment modalities, late effects and improvements over time. *Cancer Epidemiol [Internet]*. 2021;71(PB):101733. Available from: <https://doi.org/10.1016/j.canep.2020.101733>
6. Allemani C, Matsuda T, Di Carlo V, Harewood R, Matz M, Nikšić M, et al. Global surveillance of trends in cancer survival 2000–14 (CONCORD-3): analysis of individual records for 37 513 025 patients diagnosed with one of 18 cancers from 322 population-based registries in 71 countries. *The Lancet*. 2018;391(10125):1023–75.
7. Gatta G, Botta L, Rossi S, Aareleid T, Bielska-Lasota M, Clavel J, et al. Childhood cancer survival in Europe 1999-2007: Results of EURO CARE-5-a population-based study. *Lancet Oncol*. 2014;15(1):35–47.
8. Bhatia S, Armenian SH, Armstrong GT, Van Dulmen-Den Broeder E, Hawkins MM, Kremer LCM, et al. Collaborative research in childhood cancer survivorship: The current landscape. *Journal of Clinical Oncology*. 2015;33(27):3055–64.

9. Hudson MM, Armenian SH, Armstrong GT, Chow EJ, Henderson TO. Optimization of health and extension of lifespan through childhood cancer survivorship research. *Journal of Clinical Oncology*. 2018;36(21):2133–4.
10. Norsker FN, Pedersen C, Armstrong GT, Robison LL, McBride ML, Hawkins M, et al. Late Effects in Childhood Cancer Survivors: Early Studies, Survivor Cohorts, and Significant Contributions to the Field of Late Effects. *Pediatr Clin North Am*. 2020;67(6):1033–49.
11. Olsen JH, Möller T, Anderson H, Langmark F, Sankila R, Tryggvadóttir L, et al. Lifelong cancer incidence in 47 697 patients treated for childhood cancer in the nordic countries. *J Natl Cancer Inst*. 2009;101(11):806–13.
12. Geenen MM, Cardous-Ubbink MC, Kremer LCM, Van Den Bos C, Van Der Pal HJH, Heinen RC, et al. Medical assessment of adverse health outcomes in long-term survivors of childhood cancer. *JAMA*. 2007;297(24):2705–15.
13. de Fine Licht S, Rugbjerg K, Gudmundsdottir T, Bonnesen TG, Asdahl PH, Holmqvist AS, et al. Long-term inpatient disease burden in the Adult Life after Childhood Cancer in Scandinavia (ALiCCS) study: A cohort study of 21,297 childhood cancer survivors. *PLoS Med*. 2017;14(5).
14. Tran TH, Hunger SP. The genomic landscape of pediatric acute lymphoblastic leukemia and precision medicine opportunities. *Semin Cancer Biol [Internet]*. 2022;84(October 2020):144–52. Available from: <https://doi.org/10.1016/j.semcancer.2020.10.013>
15. Kattner P, Strobel H, Khoshnevis N, Grunert M, Bartholomae S, Pruss M, et al. Compare and contrast: pediatric cancer versus adult malignancies. *Cancer and Metastasis Reviews*. 2019;38(4):673–82.
16. Frebourg T, Bajalica Lagercrantz S, Oliveira C, Magenheimer R, Evans DG, Hoogerbrugge N, et al. Guidelines for the Li–Fraumeni and heritable TP53-related cancer syndromes. *European Journal of Human Genetics*. 2020;28(10):1379–86.
17. Chinnam M, Goodrich DW. RB1, Development, and Cancer. Vol. 94, *Current Topics in Developmental Biology*. 2011. 129–169 p.
18. Berkey FJ. Managing the adverse effects of radiation therapy. *Am Fam Physician*. 2010;82(4):381–8.
19. Wang X. New strategies of clinical precision medicine. *Clin Transl Med*. 2022;12(2):11–3.

20. Nishii R, Baskin-Doerfler R, Yang W, Oak N, Zhao X, Yang W, et al. Molecular basis of ETV6-mediated predisposition to childhood acute lymphoblastic leukemia. *Blood*. 2021;137(3):364–73.
21. Coccaro N, Anelli L, Zagaria A, Specchia G, Albano F. Next-generation sequencing in acute lymphoblastic Leukemia. *Int J Mol Sci*. 2019;20(12).
22. Li Y, Yang W, Devidas M, Winter SS, Kesserwan C, Yang W, et al. Germline RUNX1 variation and predisposition to childhood acute lymphoblastic leukemia. *Journal of Clinical Investigation*. 2021;131(17):1–14.
23. Mullighan CG, Zhang J, Harvey RC, Collins-Underwood JR, Schulman BA, Phillips LA, et al. JAK mutations in high-risk childhood acute lymphoblastic leukemia. *Proc Natl Acad Sci U S A*. 2009;106(23):9414–8.
24. Pui CH. Precision medicine in acute lymphoblastic leukemia. *Front Med*. 2020;14(6):689–700.
25. Irwin MS, Naranjo A, Zhang FF, Cohn SL, London WB, Gastier-Foster JM, et al. Revised Neuroblastoma Risk Classification System: A Report From the Children’s Oncology Group. *Journal of Clinical Oncology*. 2021;39(29):3229–41.
26. Garofalo A, Sholl L, Reardon B, Taylor-Weiner A, Amin-Mansour A, Miao D, et al. The impact of tumor profiling approaches and genomic data strategies for cancer precision medicine. *Genome Med [Internet]*. 2016;8(1):1–10. Available from: <http://dx.doi.org/10.1186/s13073-016-0333-9>
27. Roychowdhury S, Iyer MK, Robinson DR, Lonigro RJ, Wu YM, Cao X, et al. Personalized oncology through integrative high-throughput sequencing: A pilot study. *Sci Transl Med*. 2011;3(111).
28. Mody RJ, Wu YM, Lonigro RJ, Cao X, Roychowdhury S, Vats P, et al. Integrative clinical sequencing in the management of refractory or relapsed cancer in youth. *JAMA - Journal of the American Medical Association*. 2015;314(9):913–25.
29. Schwaederle M, Zhao M, Lee JJ, Eggermont AM, Schilsky RL, Mendelsohn J, et al. Impact of precision medicine in diverse cancers: A meta-analysis of phase II clinical trials. *Journal of Clinical Oncology*. 2015;33(32):3817–25.
30. Roskoski R. Targeting BCR-Abl in the treatment of Philadelphia-chromosome positive chronic myelogenous leukemia. *Pharmacol Res*. 2022;178:1–14.

31. Li X, Li M, Huang M, Lin Q, Fang Q, Liu J, et al. The multi-molecular mechanisms of tumor-targeted drug resistance in precision medicine. *Biomedicine and Pharmacotherapy* [Internet]. 2022;150(March):113064. Available from: <https://doi.org/10.1016/j.biopha.2022.113064>
32. Tong Z, Yan C, Dong YA, Yao M, Zhang H, Liu L, et al. Whole-exome sequencing reveals potential mechanisms of drug resistance to FGFR3-TACC3 targeted therapy and subsequent drug selection: Towards a personalized medicine. *BMC Med Genomics*. 2020;13(1):1–15.
33. Haferlach T. Advancing leukemia diagnostics: Role of next generation sequencing (ngs) in acute myeloid leukemia. *Hematol Rep*. 2020;12(S1):1–12.
34. Sánchez R, Ayala R, Martínez-López J. Minimal residual disease monitoring with next-generation sequencing methodologies in hematological malignancies. *Int J Mol Sci*. 2019;20(11).
35. Patrinos GP, Pasparakis E, Koiliari E, Pereira AC, Hünemeier T, Pereira L V., et al. Roadmap for Establishing Large-Scale Genomic Medicine Initiatives in Low- and Middle-Income Countries. *Am J Hum Genet*. 2020;107(4):589–95.
36. Byrjalsen A, Hansen TVO, Stoltze UK, Mehrjouy MM, Barnkob NM, Hjalgrim LL, et al. Nationwide germline whole genome sequencing of 198 consecutive pediatric cancer patients reveals a high frequency of cancer prone syndromes. *PLoS Genet*. 2020;16(12):1–24.
37. Freed D, Pan R, Chen H, Li Z, Hu J, Aldana R. DNAscope: High accuracy small variant calling using machine learning. *bioRxiv* [Internet]. 2022;2022.05.20.492556. Available from: <https://doi.org/10.1101/2022.05.20.492556>
38. Freed D, Pan R, Aldana R. TNscope : Accurate Detection of Somatic Mutations with Haplotype-based Variant Candidate Detection and Machine Learning Filtering. *bioRxiv*. 2018;

