Scalable Methods and Algorithms for Detecting Specific DNA Changes in Clinical Patients and Offspring of Radar-Exposed Personnel

A transgenerational biomarker of paternal exposure to ionizing radiation

Doctoral thesis

to obtain a doctorate (PhD)

from the Faculty of Medicine

of the University of Bonn

Fabian Brand

from Schmalkalden, Germany

2024

Written with authorization of
the Faculty of Medicine of the University of Bonn

First reviewer: Prof. Dr. Peter Krawitz
Second reviewer: Prof. Dr. Christian Gilissen

Day of oral examination: 08.04.2025

From the Institute of Genomic Statistics and Bioinformatics (IGSB)

Table of Contents

List of abbreviations

AC	allele count	12
cDNM	clustered de novo mutation	6, 11–13, 69–73
CNN	convolutional neural network	10
CNV	copy number variant	7
DNA	deoxyribonucleic acid	7, 10, 11, 13, 72, 73
DNM	de novo mutation	6, 8–12, 69, 70, 72, 73
DSB	double-strand break	11, 73
IR	ionizing radiation	6, 10–13, 69–73
LET	linear energy transfer	73
NGS	next generation sequencing	8, 10, 12
PCR	polymerase chain reaction	7
PPV	positive predictive value	70
ROS	reactive oxygen species	11
SNP	single nucleotide polymorphism	72
SNV	single nucleotide variant	7, 8, 73

SV	structural variant	7, 8, 70–72
WES	whole exome sequencing	8
WGS	whole genome sequencing	6–8, 12, 13, 69–71, 73

Glossary

CRU	Cancer research ukraine cohort. Secondary case cohort of 130 offspring of exposed fathers and mothers accessed from Yeager, et al. (2021) via dbGAP phs001163.v1.p1	6, 12, 71, 72
Inova	Inova cohort. Control cohort comprised of 1,214 offspring of parents with no known exposure beyond naturally occurring ionizing radiation. Data was accessed from Inova Fairfax Medical Centre directly	6, 8, 13, 69, 71
Radar	Radar cohort. A newly recruited cohort of 110 offspring from fathers that were exposed to ionizing radiation as soldiers of the German army	6, 12, 13, 71

1 Abstract

The abundance of Whole genome sequencing (WGS) data now enables research into more complex phenotypes and into effects of Ionizing radiation (IR) on human DNA. We used WGS data to contribute to four advances: 1. Finding a novel disease-causing variant for Koolen-de Vries syndrome; 2. Detecting mutational signatures in error-prone sequencing data; 3. Establishing a transgenerational biomarker for paternal exposure to IR; and 4. assessing Covid-19 host genetics. To ascertain potential transgenerational effects of IR exposure, we recruited a cohort comprised of 110 offspring of radar personnel of both German armies and accessed two more cohorts, one cohort of 130 offspring of liquidators and inhabitants of the town of Pripyat that were exposed to IR following the nuclear accident in 1986 (CRU cohort), and a large control cohort featuring 1214 offspring of non-exposed parents (Inova). We analyzed all data for the Radarstudy and Covid-19 data using newly developed WGS data analysis pipelines. Previous works suggested clustered de novo mutations (cDNMs), which are defined as two or more de novo mutations (DNMs) within 20 bp as potential signature of paternal IR exposure. To optimize the detection accuracy of DNMs and cDNMs, we used data from validation experiments to create a custom DNM and cDNM calling algorithm based on DeepTrio. We showed that deep-learning approaches based on DeepTrio can be trained with low data requirements. Our DNM detection model achieved a sensitivity of 95.7 % and a precision of 89.6 %. More complex mutational signatures, like cDNMs, can be detected with a precision of 76.9 % at 100 % sensitivity. Using newly developed analysis pipelines for WGS data, we detected a 4.7k bp deletion in *KANSL1*, that was found to be causing the patient phenotype, and provided insights into Covid-19 host genetics by elucidating correlations of variation and disease progression. When analyzing the three large WGS cohorts, we found that cDNMs were increased in children born to parents that were irradiated prior to conception. We observed 2.65 cDNMs per offspring on average in the CRUcohort, 1.48 in the Radar cohort and 0.88 in the Inova cohort ($p < 0.005$). Further statistical models indicated that this increase in cDNMs scales with the paternal exposure to IR ($p < 0.001$). These results leave little doubt that cDNMs represent a transgenerational biomarker of paternal IR exposure.

2 Introduction

Since the first human genome was sequenced in 2003, the techniques to analyze DNA have continuously evolved. Sequencing devices have become more capable, datasets more numerous and questions driving research in human genetics have become more complicated. Nowadays, WGS has become commonplace in research and clinics, where it was still a significant challenge 5 years ago. The ubiquity of WGS data, and the machines to generate it enabled research into potentially complex mutational signatures for the first time.

The introduction of the Illumina HiSeq X 10 sequencer allowed researchers to perform WGS analyses for less than 1000 $ in 2014 (Check Hayden, 2014). Its flow cells sequenced up to 6 billion paired-end reads in a three-day cycle, enough for 10 genomes per run (Illumina, 2011) (Illumina, 2015). 3 years later, the machine was superseded by the Illumina NovaSeq, which increased the throughput to 48 WGS samples in 44 hours, reducing the costs to sequence large WGS cohorts (Illumina, 2022). While the overall quality of sequencing data on these platforms is exceptional, e.g. Q30 > 85 % on the NovaSeq 6000, specific error motifs, which can lead to enrichment of specific base exchanges, dinucleotide exchanges or larger errors, have been found in data from all generations of Illumina sequencers (Arora et al., 2019) (Stoler and Nekrutenko, 2021) (Ma et al., 2019). Errors might arise during laboratory sample preparation processes (e.g. DNA Extraction, Fragmentation), PCR amplification (if it is part of the protocol) or the actual sequencing process (e.g. Cluster Duplication). Exchanges like A > G and T > C have been shown to be more frequent than others on the NovaSeq 6000 sequencer, aggravated by factors like difficult to sequence and to call regions, e.g. tandem repeats, homopolymers and low-complexity regions of the genome (Ma et al., 2019) (Arora et al., 2019) (Lee and Schatz, 2012) (Hijikata et al., 2024). This extends to di- or trinucleotide exchanges, where TT > GG sequences have been found to be increased, among others (Arora et al., 2019) (Ma et al., 2019).

Analysis methods that keep up with the growing demand introduced by these sequencers, while incorporating ever more advanced methods of error detection and correction had been developed to analyze single nucleotide variants (SNVs), copy number variants (CNVs), struc-

tural variants (SVs), and more (Koboldt, 2020). Recent developments in acceleration technologies for the mapping and variant calling steps promised to speed up these computation on FPGA or GPU chips (Subramaniyan et al., 2021) (Vasimuddin et al., 2019). Productionized software in the form of toolkits like Illumina DRAGEN and NVIDIA Parabricks exist, and their acceleration of up to 65x compared to open-source implementations of the same algorithms now allows for large scale processing of WGS data in the cloud or on on-premises computing infrastructure (O'Connell et al., 2023). All these technologies enabled the sequencing and processing of large scale WGS cohorts, like the Inovacohort, and the analysis of complex traits (Wong et al., 2016). The Inova cohort consists of 1,214 parent-offspring trios, which were used to assess previously unknown maternal age effects, clustered mutations arising in the maternal germline and genetic effects of preterm births (Wong et al., 2016) (Goldmann et al., 2018) (Knijnenburg et al., 2019).

With advanced algorithms, SNVs and small insertions and deletions (Indels) can be called with great accuracy, but more complex types of variation, e.g. SVs or DNMs, are harder to call correctly using short-read data (Olson et al., 2022) (Wagner et al., 2022) (Gabrielaite et al., 2021). SVs are usually detected via recognizing statistical patterns in the NGS read data like increased or decreased coverage of a genomic region, inverted orientation of paired-end reads, or truncated reads (Gabrielaite et al., 2021). The fuzzy nature of the discovery techniques means that for structural variants, discovery is error-prone. Reports showed false discovery rates of up to 89 % in SV-calling experiments and only reached satisfactory accuracy (FDR $\leq$ 10%) in eight of 36 callsets of SVs (Mahmoud et al., 2019) (1000 Genomes Project et al., 2011). The accuracy of these calls is increased in WGS compared to WES or targeted panels, but pipelines for SV detection still have to deal with large amounts of false positive calls (Gabrielaite et al., 2021). Due to their often severe consequences however, these more difficult to identify variants are often incorporated into clinical analysis pipelines, where the sensitivity of calling is usually preferred over specificity (Demidov et al., 2024) (Schmidt et al., 2024). Similarly, DNMs are a class of substitutions that is hard to detect in common short read NGS data, but such variants are often implicated with the occurence of rare diseases. DNMs

are recognized in short-read data, if the offspring in a parent-child trio exhibits a particular heterozygous variant that is not present in read data of any parent, a so-called Mendelian error. DNMs naturally arise during cell division in spermatogonial stem cells, due to aging of the maternal oocytes, or due to post-zygotic mosaicism in early embryonic formation (Goldmann et al., 2019). In the former case, mutations arising due to aging of the parents, mutation rates for DNMs are $1.29 \cdot 10^{-8}$ (Kong et al., 2012) (Besenbacher et al., 2016). Further, it is known that paternal and maternal age contribute substantially to this number, with the paternal age effect being 1 - 2 mutations per year of age at conception of the offspring, and maternal age leading to an increase of 0.24 additional DNMs per year of age (Kong et al., 2012) (Jónsson et al., 2017) (Acuna-Hidalgo et al., 2016). Secondly, if mutations arise during the first few cell divisions after fertilization, they have been shown to cause different cell populations to exist within one individual (Acuna-Hidalgo et al., 2016) (Jonsson et al., 2021). A study recently concluded that 2.6 post-zygotic mutations arise per individual per generation, using data from three generational pedigrees to differentiate between germline and post-zygotic DNMs, where the middle generation featured monozygotic twins. Other works claimed that the number of post-zygotic DNMs is even higher, up to 10 % (Jonsson et al., 2021) (Sasani et al., 2019). The same studies also showed that the paternal age effect is variable across families, and that the overall number of DNMs is significantly influenced by the accumulation of post-zygotic mosaicism (Sasani et al., 2019). In sequencing experiments, two main factors combine to make the identification of DNMs in strict parent-offspring trios without further information difficult. By their nature DNMs are indistinguishable from sequencing errors, so much so that they have been used as quality assessment criteria for variant calling accuracy (the lower, the better) (Kothiyal et al., 2017) (Pilipenko et al., 2014) (Lin et al., 2018). Additionally, the high proportion of post-zygotic mosaicism can lead to a shift in the distribution of variant allele frequencies for these mutations, since they are only present in a single allele of a subpopulation of cells. Because of these facts, accurate calling of DNMs based on heuristics and statistical models remains a debated topic, with new methods and comparisons being released frequently (Shah et al., 2024) (Liang et al., 2019). Applications like “DeNovoGear”

or “GATK PhaseByTransmission” can call such variants, achieving F1-scores between 58 % and 84 % in different reports (Shah et al., 2024) (Khazeeva et al., 2022).

Deep learning had been introduced to the field of NGS variant calling with the first release of the DeepVariant framework in 2018 and its extension DeepTrio first recognized the potential for improvement of variant calls made by convolutional neural networks (CNNs) through incorporating parental data (Poplin et al., 2018) (Kolesnikov et al., 2021). These results showed that methods derived from common image analysis and image classification tasks can be applied to genetic data as well and improve overall variant calling accuracy by incorporating larger windows of reference data into every single variant call. DeepVariant and DeepTrio both constructed image-like tensors for 221 bp wide reference windows, which incorporate information from the reads at each reference position, including read base, quality scores, and hints at potential variant sites. DeepTrio extends upon DeepVariant by reserving the upper and lower third of the tensor for data of the father and the mother respectively, and both networks use a simple feed-forward CNN architecture to classify sites into one of three categories (hom-ref, het, hom-alt) (Poplin et al., 2018) (Kolesnikov et al., 2021). This provides a framework that can be extended by retraining or by changing specific properties of the input tensors or output neurons. Other Deep-Learning based DNM callers have been developed, utilizing custom encoding schemes to transform read data into a format suitable for CNNs (Khazeeva et al., 2022). These algorithms provided a marked improvement in DNM calling accuracy over traditional tools, but are limited to only DNMs (Khazeeva et al., 2022).

Ionizing radiation (IR) is a widely recognized mutagen that affects the human DNA. Mutational signatures rose to prominence in different fields of human genetics, e.g. in cancer diagnostics, and have been recognized as paternal or maternal age effect, but even though some candidates have been researched in mice, no definite mutational signature of IR exposure in humans has been found to date (Tate et al., 2019) (Wong et al., 2016) (Jónsson et al., 2017). Mutational signatures can be the number of a specific type of variant, e.g. number of DNMs informs the paternal age effect, but equally often they are formed by multiple variants, possibly even multiple specific base exchanges (Besenbacher et al., 2016) (Wong et al., 2016)

(Jónsson et al., 2017). In the case of IR, studies in mice and a single study in humans recognized clusters of DNMs (cDNMs) that are < 20 bp apart as potential signature (Adewoye et al., 2015) (Satoh et al., 2020) (Holtgrewe et al., 2018). Such clustered lesions may arise as a consequence of IR induced double-strand breaks (DSBs), and erroneous repair (Sage and Shikazono, 2017) (Georgakilas et al., 2013) (Frankenberg-Schwager, 1990). Most IR-induced DNA damage, e.g. DSBs, loss of bases or oxidized bases, is mediated through the generation of ROS in water molecules in the cells. ROS cause damage to the DNA double-strand in a range of up to 6 nm, roughly two helix turns of DNA (Georgakilas et al., 2013). Error-prone DNA repair mechanisms then turn lesions into mutations that can be inherited by following generations. Notably DNA repair is less efficient in spermatids and mature spermatozoa, which show the highest radio sensitivity of all stage of spermatogenesis (Jan et al., 2017) (Wdowiak et al., 2019).

The need to investigate mutational signatures of IR arises every time humans are accidentally or intentionally subjected to elevated doses, e.g. due to incidents at nuclear power plants (Chernobyl, 1986) or through the use of nuclear weapons (Bazyka et al., 2020) (Little et al., 2013). In Germany, many former radar personnel of both German armies suffer the consequences of improper shielding of radar devices that were in service from the 1950s to the 1980s. The soldiers, officers and operators of radar installations in both German armies, were exposed to higher doses of x-ray radiation throughout their service, until this issue was recognized and radiation-shielding measures were taken at the end of the 1980s (König et al., 2003). The German and other European governments already recognized certain diseases of radar operators as a consequence of their service, but the effects on their offspring remains to be elucidated (König et al., 2003) (Degrave et al., 2009). Effects on the genome of offspring of exposed radar personnel potentially includes cDNMs, but also balanced and unbalanced translocations (Holtgrewe et al., 2018). cDNMs are of particular interest, since the cluster formation mediated by exposure to IR is well understood and can lead to the introduction of DNMs in the paternal germline, but detection is difficult due to the compounding errors introduced by post-zygotic mosaicism and error-prone sequencers. Cohorts consisting

of mothers and fathers that were exposed to IR have recently been studied by Yeager, et al. and Moorhouse, et al. (Yeager et al., 2021) (Moorhouse et al., 2022). The former analyzed a cohort of 130 offspring of liquidators that were involved in the cleanup of the nuclear power plant in Chernobyl following its disastrous accident in 1986, and some people that lived there at the time. No increased rates of mutations of any kind was found, although no clusters smaller than 47k bp were analyzed (Yeager et al., 2021). The latter recruited soldiers from the British Army and Navy that were involved in nuclear tests. Moorhouse, et al. found no increased rate of standard mutations, DNM clusters up to 10 bp or 100 bp wide, or chromosomal aberrations (Moorhouse et al., 2022) (Lawrence et al., 2024). However, they detected a higher rate of SBS16 signatures in offspring of exposed nuclear test veterans (Tate et al., 2019) (Moorhouse et al., 2022).

This dissertation represents the culmination of work in: (1) Establishing the foundations of the analysis of WGS for clinical and statistical analysis; (2) The development of methods for detecting DNMs and mutational signatures like cDNMs in WGS data; and (3) Assessing and analyzing the rates of cDNMs in the general population and offspring of parents exposed to low-doses of IR. We worked together with the NGS sequencing facility to establish reliable and scalable workflows to process NGS data coming from the NovaSeq 6000 sequencers, first optimizing quality and accuracy of the results generated. When applied to a clinical patient with Koolen-de Vries syndrome, we were able to find a novel deletion, that was confirmed to be disease causing, and not detected in exome sequencing or panel data. The application of these methods to a large cohort of Covid-19 infected patients yielded insights into the host genetics of the disease.

Secondly, in parallel to our research into biomarkers for prolonged paternal exposure to ionizing radiation, we optimized different algorithms for accurate DNM calling. We developed a in-house DNM detection pipeline based on the `hail` Framework, and deployed it to recognize DNMs in large cohorts. These heuristic callers relied very much on information of all cohorts, such as the total allele count (AC), which served as one of the most stringent filters. To improve the accuracy of these calls further, we then invested in extending the DeepTrio

algorithm for the detection of DNMs and cDNMs, by retraining the algorithm using long-read sequencing data and validated DNMs and cDNMs from the Radar cohort as well as simulated data.

Thirdly, we collected variant calls, as well as DNMs and cDNMs from the Radar cohort, the CRU cohort from Yeager, et al. and the Inova cohort, a total of 4,337 WGS samples, into a single analysis on potential signatures of paternal exposure to ionizing radiation. Our focus was on the detection of cDNMs, since they were implicated with IR-induced DNA damages in earlier studies, and had been found with great frequency in data from our Pilotstudy (Holtgrewe et al., 2018). We recruited members of and sequenced data for the Radar cohort, performed a new alignment and joint-variant calling on all three cohorts, confirmed earlier results on the paternal age effect, and found increased rates of cDNMs in offspring of exposed individuals (Besenbacher et al., 2016) (Jónsson et al., 2017).

2.1 References

1000 Genomes Project, Mills RE, Walter K, Stewart C, Handsaker RE, Chen K, Alkan C, Abyzov A, Yoon SC, Ye K, Cheetham RK, Chinwalla A, Conrad DF, Fu Y, Grubert F, Hajirasouliha I, Hormozdiari F, Iakoucheva LM, Iqbal Z, Kang S, Kidd JM, Konkel MK, Korn J, Khurana E, Kural D, Lam HYK, Leng J, Li R, Li Y, Lin CY, Luo R, Mu XJ, Nemesh J, Peckham HE, Rausch T, Scally A, Shi X, Stromberg MP, Stütz AM, Urban AE, Walker JA, Wu J, Zhang Y, Zhang ZD, Batzer MA, Ding L, Marth GT, McVean G, Sebat J, Snyder M, Wang J, Ye K, Eichler EE, Gerstein MB, Hurles ME, Lee C, McCarroll SA, Korbel JO. Mapping Copy Number Variation by Population-Scale Genome Sequencing. In: Nature 2011; 470 (7332): 59–65. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature09708. URL: https://www.nature.com/articles/nature09708 (visited on 11/11/2024)

Acuna-Hidalgo R, Veltman JA, Hoischen A. New Insights into the Generation and Role of de Novo Mutations in Health and Disease. In: Genome Biology 2016; 17 (1): 241. ISSN: 1474-760X. DOI: 10.1186/s13059-016-1110-1. URL: http://genomebiology.biomedcentral.com/articles/10.1186/s13059-016-1110-1 (visited on 11/14/2024)

Adewoye AB, Lindsay SJ, Dubrova YE, Hurles ME. The Genome-Wide Effects of Ionizing Radiation on Mutation Induction in the Mammalian Germline. In: Nature communications 2015; 6: 6684

Arora K, Shah M, Johnson M, Sanghvi R, Shelton J, Nagulapalli K, Oschwald DM, Zody MC, Germer S, Jobanputra V, Carter J, Robine N. Deep Whole-Genome Sequencing of 3 Cancer Cell Lines on 2 Sequencing Platforms. In: Scientific Reports 2019; 9 (1): 19123. ISSN: 2045-2322. DOI: 10.1038/s41598-019-55636-3. URL: https://www.nature.com/articles/s41598-019-55636-3 (visited on 11/11/2024)

Bazyka D, Hatch M, Gudzenko N, Cahoon EK, Drozdovitch V, Little MP, Chumak V, Bakhanova E, Belyi D, Kryuchkov V, Golovanov I, Mabuchi K, Illienko I, Belayev Y, Bodelon C, Machiela MJ, Hutchinson A, Yeager M, De Gonzalez AB, Chanock SJ. Field Study of the Possible Effect of Parental Irradiation on the Germline of Children Born to Cleanup Workers and Evacuees of the Chornobyl Nuclear Accident. In: American Journal of Epidemiology 2020; 189 (12): 1451–1460. ISSN: 0002-9262, 1476-6256. DOI: 10.1093/aje/kwaa095. URL: https://academic.oup.com/aje/article/189/12/1451/5866142 (visited on 10/23/2024)

Besenbacher S, Sulem P, Helgason A, Helgason H, Kristjansson H, Jonasdottir A, Jonasdottir A, Magnusson OT, Thorsteinsdottir U, Masson G, Kong A, Gudbjartsson DF, Stefansson K. Multi-Nucleotide de Novo Mutations in Humans. In: PLOS Genetics 2016; 12 (11): ed. e1006315. ISSN: 1553-7404. DOI: 10.1371/journal.pgen.1006315. URL: https://dx.plos.org/10.1371/journal.pgen.1006315 (visited on 09/25/2023)

Check Hayden E. Is the $1,000 Genome for Real? In: Nature 2014: nature.2014.14530. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/nature.2014.14530. URL: https://www.nature.com/articles/nature.2014.14530 (visited on 11/11/2024)

Degrave E, Meeusen B, Grivegnée AR, Boniol M, Autier P. Causes of Death among Belgian Professional Military Radar Operators: A 37-Year Retrospective Cohort Study. In: International Journal of Cancer 2009; 124 (4): 945–951. DOI: 10.1002/ijc.23988

Demidov G, Laurie S, Torella A, Piluso G, Scala M, Morleo M, Nigro V, Graessner H, Banka S, Solve-RD consortium, Macaya A, Pérez-Dueñas B, Jackson A, Stevanin G, De Sainte

Agathe JM, Havlovicová M, Horvath R, Pinelli M, Van Os NJH, Van De Warrenburg BPC, Denommé-Pichon AS, Savarese M, Johari M, Dallapiccola B, Tartaglia M, Pauly MG, Sommer AK, Haack TB, Töpf A, Didier L, Fallerini C, Renieri A, Chinnery PF, Natera-de Benito D, Nascimento A, Trimouille A, Munell F, Marcé-Grau A, Rabah BY, Bonne G, Van De Vondel L, Lohmann K, Ossowski S. Structural Variant Calling and Clinical Interpretation in 6224 Unsolved Rare Disease Exomes. In: European Journal of Human Genetics 2024; 32 (8): 998–1004. ISSN: 1018-4813, 1476-5438. DOI: 10.1038/s41431-024-01637-4. URL: https://www.nature.com/articles/s41431-024-01637-4 (visited on 11/11/2024)

Frankenberg-Schwager M. Induction, Repair and Biological Relevance of Radiation-Induced DNA Lesions in Eukaryotic Cells. In: Radiation and environmental biophysics 1990; 29 (4): 273–292

Gabrielaite M, Torp MH, Rasmussen MS, Andreu-Sánchez S, Vieira FG, Pedersen CB, Kinalis S, Madsen MB, Kodama M, Demircan GS, Simonyan A, Yde CW, Olsen LR, Marvig RL, Østrup O, Rossing M, Nielsen FC, Winther O, Bagger FO. A Comparison of Tools for Copy-Number Variation Detection in Germline Whole Exome and Whole Genome Sequencing Data. In: Cancers 2021; 13 (24): 6283. ISSN: 2072-6694. DOI: 10.3390/cancers13246283. URL: https://www.mdpi.com/2072-6694/13/24/6283 (visited on 11/11/2024)

Georgakilas AG, O'Neill P, Stewart RD. Induction and Repair of Clustered DNA Lesions: What Do We Know so Far? In: Radiation research 2013; 180 (1): 100–109

Goldmann J, Veltman J, Gilissen C. De Novo Mutations Reflect Development and Aging of the Human Germline. In: Trends in Genetics 2019; 35 (11): 828–839. ISSN: 01689525. DOI: 10.1016/j.tig.2019.08.005. URL: https://linkinghub.elsevier.com/retrieve/pii/S0168952519301787 (visited on 11/14/2024)

Goldmann JM, Seplyarskiy VB, Wong WSW, Vilboux T, Neerincx PB, Bodian DL, Solomon BD, Veltman JA, Deeken JF, Gilissen C, Niederhuber JE. Germline de Novo Mutation Clusters Arise during Oocyte Aging in Genomic Regions with High Double-Strand-Break Incidence. In: Nature Genetics 2018; 50 (4): 487–492. ISSN: 1061-4036, 1546-1718. DOI:

10.1038/s41588-018-0071-6. URL: https://www.nature.com/articles/s41588-018-0071-6 (visited on 09/17/2024)

Hijikata A, Suyama M, Kikugawa S, Matoba R, Naruto T, Enomoto Y, Kurosawa K, Harada N, Yanagi K, Kaname T, Miyako K, Takazawa M, Sasai H, Hosokawa J, Itoga S, Yamaguchi T, Kosho T, Matsubara K, Kuroki Y, Fukami M, Adachi K, Nanba E, Tsuchida N, Uchiyama Y, Matsumoto N, Nishimura K, Ohara O. Exome-Wide Benchmark of Difficult-to-Sequence Regions Using Short-Read next-Generation DNA Sequencing. In: Nucleic Acids Research 2024; 52 (1): 114–124. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkad1140. URL: https://academic.oup.com/nar/article/52/1/114/7453260 (visited on 11/11/2024)

Holtgrewe M, Knaus A, Hildebrand G, Pantel JT, de Los Santos MR, Neveling K, Goldmann J, Schubach M, Jäger M, Coutelier M, et al. Multisite de Novo Mutations in Human Offspring after Paternal Exposure to Ionizing Radiation. In: Scientific reports 2018; 8 (1): 1–5

Illumina. HiSeq™ Sequencing Systems. HiSeq™ Sequencing Systems. URL: https://www.illumina.com/content/dam/illumina-support/documents//products/datasheets/datasheet_hiseq_systems.pdf

Illumina. HiSeq X™ Series of Sequencing Systems. URL: https://pdf.medicalexpo.com/pdf/illumina-inc/hiseq-x-series/83632-140610.html

Illumina. NovaSeq™ 6000 Sequencing System. NovaSeq™ 6000 Sequencing System. URL: https://emea.illumina.com/content/dam/illumina/gcs/assembled-assets/marketing-literature/novaseq-6000-spec-sheet-m-gl-00271/novaseq-6000-spec-sheet-m-gl-00271.pdf

Jan SZ, Vormer TL, Jongejan A, Röling MD, Silber SJ, De Rooij DG, Hamer G, Repping S, Van Pelt AMM. Unraveling Transcriptome Dynamics in Human Spermatogenesis. In: Development 2017; 144 (20): 3659–3673. ISSN: 1477-9129, 0950-1991. DOI: 10.1242/dev.152413. URL: https://journals.biologists.com/dev/article/144/20/3659/48147/Unraveling-transcriptome-dynamics-in-human (visited on 10/23/2024)

Jonsson H, Magnusdottir E, Eggertsson HP, Stefansson OA, Arnadottir GA, Eiriksson O, Zink F, Helgason EA, Jonsdottir I, Gylfason A, Jonasdottir A, Jonasdottir A, Beyter D, Ste-

ingrimsdottir T, Norddahl GL, Magnusson OT, Masson G, Halldorsson BV, Thorsteinsdottir U, Helgason A, Sulem P, Gudbjartsson DF, Stefansson K. Differences between Germline Genomes of Monozygotic Twins. In: Nature Genetics 2021; 53 (1): 27–34. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-020-00755-1. URL: https://www.nature.com/articles/s41588-020-00755-1 (visited on 11/14/2024)

Jónsson H, Sulem P, Kehr B, Kristmundsdottir S, Zink F, Hjartarson E, Hardarson MT, Hjorleifsson KE, Eggertsson HP, Gudjonsson SA, Ward LD, Arnadottir GA, Helgason EA, Helgason H, Gylfason A, Jonasdottir A, Jonasdottir A, Rafnar T, Frigge M, Stacey SN, Magnusson OT, Thorsteinsdottir U, Masson G, Kong A, Halldorsson BV, Helgason A, Gudbjartsson DF, Stefansson K. Parental Influence on Human Germline de Novo Mutations in 1,548 Trios from Iceland. In: Nature 2017; 549 (7673): 519–522. DOI: 10.1038/nature24018

Khazeeva G, Sablauskas K, van der Sanden B, Steyaert W, Kwint M, Rots D, Hinne M, van Gerven M, Yntema H, Vissers L, Gilissen C. DeNovoCNN: A Deep Learning Approach to *de Novo* Variant Calling in next Generation Sequencing Data. In: Nucleic Acids Research 2022; 50 (17): e97–e97. ISSN: 0305-1048, 1362-4962. DOI: 10.1093/nar/gkac511. URL: https://academic.oup.com/nar/article/50/17/e97/6609811 (visited on 11/11/2024)

Knijnenburg TA, Vockley JG, Chambwe N, Gibbs DL, Humphries C, Huddleston KC, Klein E, Kothiyal P, Tasseff R, Dhankani V. Genomic and Molecular Characterization of Preterm Birth. In: Proceedings of the National Academy of Sciences 2019; 116 (12): 5819–5827

Koboldt DC. Best Practices for Variant Calling in Clinical Sequencing. In: Genome Medicine 2020; 12 (1): 91. ISSN: 1756-994X. DOI: 10.1186/s13073-020-00791-w. URL: https://genomemedicine.biomedcentral.com/articles/10.1186/s13073-020-00791-w (visited on 11/11/2024)

Kolesnikov A, Goel S, Nattestad M, Yun T, Baid G, Yang H, McLean CY, Chang PC, Carroll A. DeepTrio: Variant Calling in Families Using Deep Learning. DOI: 10.1101/2021.04.05.438434. URL: http://biorxiv.org/lookup/doi/10.1101/2021.04.05.438434 (visited on 11/11/2024). Pre-published

Kong A, Frigge ML, Masson G, Besenbacher S, Sulem P, Magnusson G, Gudjonsson SA, Sigurdsson A, Jonasdottir A, Jonasdottir A, Wong WSW, Sigurdsson G, Walters GB, Steinberg S, Helgason H, Thorleifsson G, Gudbjartsson DF, Helgason A, Magnusson OT, Thorsteinsdottir U, Stefansson K. Rate of de Novo Mutations and the Importance of Father's Age to Disease Risk. In: Nature 2012; 488 (7412): 471–475. ISSN: 1476-4687. DOI: 10.1038/nature11396. URL: https://www.nature.com/articles/nature11396 (visited on 04/28/2022)

König W, Blettner M, Ambrosi P, Anger K, Beltz D, Brüggemeyer H, Eggert S, Franke B, Greiser E, Hille R, Käs G, Kiefer J, Kirchner G, Köhnlein W, List V, Paretzke HG, Schütz J, Gehrcke K, Thieme M. Bericht Der Expertenkommission Zur Frage Der Gefährdung Durch Strahlung in Früher Radareinrichtungen Der Bundeswehr Und Der NVA (Radarkommission)[Report of the Expert Group on the Issue of the Hazards of Radiation in Earlier Types of Radar Devices and the German Radar Commission]. Berlin, German Radar Commission. In: 2003: URL: https://upload.wikimedia.org/wikipedia/commons/4/44/Bericht_Radarkommission_Deutscher_Bundestag_Volltext.pdf

Kothiyal P, Wong WS, Bodian DL, Niederhuber JE. Mendelian Inheritance Errors in Whole Genome Sequenced Trios Are Enriched in Repeats and Cluster within Copy Number Losses. DOI: 10.1101/240424. URL: http://biorxiv.org/lookup/doi/10.1101/240424 (visited on 11/11/2024). Pre-published

Lawrence KJ, Scholze M, Seixo J, Daley F, Al-Haddad E, Craenen K, Gillham C, Rake C, Peto J, Anderson R. M-FISH Evaluation of Chromosome Aberrations to Examine for Historical Exposure to Ionising Radiation Due to Participation at British Nuclear Test Sites. In: Journal of Radiological Protection 2024; 44 (1): 011501. ISSN: 0952-4746, 1361-6498. DOI: 10.1088/1361-6498/ad1743. URL: https://iopscience.iop.org/article/10.1088/1361-6498/ad1743 (visited on 09/17/2024)

Lee H, Schatz MC. Genomic Dark Matter: The Reliability of Short Read Mapping Illustrated by the Genome Mappability Score. In: Bioinformatics 2012; 28 (16): 2097–2105. ISSN: 1367-4811, 1367-4803. DOI: 10.1093/bioinformatics/bts330. URL: https://academic.oup.com/bioinformatics/article/28/16/2097/323484 (visited on 11/11/2024)

Liang Y, He L, Zhao Y, Hao Y, Zhou Y, Li M, Li C, Pu X, Wen Z. Comparative Analysis for the Performance of Variant Calling Pipelines on Detecting the de Novo Mutations in Humans. In: Frontiers in Pharmacology 2019; 10: 358. ISSN: 1663-9812. DOI: 10.3389/fphar.2019.00358. URL: https://www.frontiersin.org/article/10.3389/fphar.2019.00358/full (visited on 11/11/2024)

Lin MF, Rodeh O, Penn J, Bai X, Reid JG, Krasheninina O, Salerno WJ. GLnexus: Joint Variant Calling for Large Cohort Sequencing. In: BioRxiv 2018: 343970

Little MP, Goodhead DT, Bridges BA, Bouffler SD. Evidence Relevant to Untargeted and Transgenerational Effects in the Offspring of Irradiated Parents. In: Mutation Research/Reviews in Mutation Research 2013; 753 (1): 50–67

Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, Liu Y, Chen X, Newman S, Nakitandwe J, Li Y, Li B, Shen S, Wang Z, Shurtleff S, Robison LL, Levy S, Easton J, Zhang J. Analysis of Error Profiles in Deep Next-Generation Sequencing Data. In: Genome Biology 2019; 20 (1): 50. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1659-6. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1659-6 (visited on 11/11/2024)

Mahmoud M, Gobet N, Cruz-Dávalos DI, Mounier N, Dessimoz C, Sedlazeck FJ. Structural Variant Calling: The Long and the Short of It. In: Genome Biology 2019; 20 (1): 246. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1828-7. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1828-7 (visited on 11/11/2024)

Moorhouse AJ, Scholze M, Sylvius N, Gillham C, Rake C, Peto J, Anderson R, Dubrova YE. No Evidence of Increased Mutations in the Germline of a Group of British Nuclear Test Veterans. In: Scientific Reports 2022; 12 (1): ISSN: 2045-2322. DOI: 10.1038/s41598-022-14999-w. URL: https://www.nature.com/articles/s41598-022-14999-w (visited on 09/17/2024)

O'Connell KA, Yosufzai ZB, Campbell RA, Lobb CJ, Engelken HT, Gorrell LM, Carlson TB, Catana JJ, Mikdadi D, Bonazzi VR, Klenk JA. Accelerating Genomic Workflows Using NVIDIA Parabricks. In: BMC Bioinformatics 2023; 24 (1): 221. ISSN: 1471-2105. DOI:

10.1186/s12859-023-05292-2. URL: https://bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-023-05292-2 (visited on 11/11/2024)

Olson ND, Wagner J, McDaniel J, Stephens SH, Westreich ST, Prasanna AG, Johanson E, Boja E, Maier EJ, Serang O, Jáspez D, Lorenzo-Salazar JM, Muñoz-Barrera A, Rubio-Rodríguez LA, Flores C, Kyriakidis K, Malousi A, Shafin K, Pesout T, Jain M, Paten B, Chang PC, Kolesnikov A, Nattestad M, Baid G, Goel S, Yang H, Carroll A, Eveleigh R, Bourgey M, Bourque G, Li G, Ma C, Tang L, Du Y, Zhang S, Morata J, Tonda R, Parra G, Trotta JR, Brueffer C, Demirkaya-Budak S, Kabakci-Zorlu D, Turgut D, Kalay Ö, Budak G, Narcı K, Arslan E, Brown R, Johnson IJ, Dolgoborodov A, Semenyuk V, Jain A, Tetikol HS, Jain V, Ruehle M, Lajoie B, Roddey C, Catreux S, Mehio R, Ahsan MU, Liu Q, Wang K, Ebrahim Sahraeian SM, Fang LT, Mohiyuddin M, Hung C, Jain C, Feng H, Li Z, Chen L, Sedlazeck FJ, Zook JM. PrecisionFDA Truth Challenge V2: Calling Variants from Short and Long Reads in Difficult-to-Map Regions. In: Cell Genomics 2022; 2 (5): 100129. ISSN: 2666979X. DOI: 10.1016/j.xgen.2022.100129. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666979X22000581 (visited on 11/11/2024)

Pilipenko VV, He H, Kurowski BG, Alexander ES, Zhang X, Ding L, Mersha TB, Kottyan L, Fardo DW, Martin LJ. Using Mendelian Inheritance Errors as Quality Control Criteria in Whole Genome Sequencing Data Set. In: BMC Proceedings 2014; 8 (S1): S21. ISSN: 1753-6561. DOI: 10.1186/1753-6561-8-S1-S21. URL: https://bmcproc.biomedcentral.com/articles/10.1186/1753-6561-8-S1-S21 (visited on 11/11/2024)

Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, Newburger D, Dijamco J, Nguyen N, Afshar PT, Gross SS, Dorfman L, McLean CY, DePristo MA. A Universal SNP and Small-Indel Variant Caller Using Deep Neural Networks. In: Nature Biotechnology 2018; 36 (10): 983–987. ISSN: 1087-0156, 1546-1696. DOI: 10.1038/nbt.4235. URL: https://www.nature.com/articles/nbt.4235 (visited on 11/11/2024)

Sage E, Shikazono N. Radiation-Induced Clustered DNA Lesions: Repair and Mutagenesis. In: Free Radical Biology and Medicine 2017; 107: 125–135

Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB, Quinlan AR. Large, Three-Generation Human Families Reveal Post-Zygotic Mosaicism and Variability in Germline Mutation Accumulation. In: eLife 2019; 8: e46922. ISSN: 2050-084X. DOI: 10.7554/eLife.46922. URL: https://elifesciences.org/articles/46922 (visited on 11/14/2024)

Satoh Y, Asakawa Ji, Nishimura M, Kuo T, Shinkai N, Cullings HM, Minakuchi Y, Sese J, Toyoda A, Shimada Y, et al. Characteristics of Induced Mutations in Offspring Derived from Irradiated Mouse Spermatogonia and Mature Oocytes. In: Scientific reports 2020; 10 (1): 1–13

Schmidt A, Danyel M, Grundmann K, Brunet T, Klinkhammer H, Hsieh TC, Engels H, Peters S, Knaus A, Moosa S, Averdunk L, Boschann F, Sczakiel HL, Schwartzmann S, Mensah MA, Pantel JT, Holtgrewe M, Bösch A, Weiß C, Weinhold N, Suter AA, Stoltenburg C, Neugebauer J, Kallinich T, Kaindl AM, Holzhauer S, Bührer C, Bufler P, Kornak U, Ott CE, Schülke M, Nguyen HHP, Hoffjan S, Grasemann C, Rothoeft T, Brinkmann F, Matar N, Sivalingam S, Perne C, Mangold E, Kreiss M, Cremer K, Betz RC, Mücke M, Grigull L, Klockgether T, Spier I, Heimbach A, Bender T, Brand F, Stieber C, Morawiec AM, Karakostas P, Schäfer VS, Bernsen S, Weydt P, Castro-Gomez S, Aziz A, Grobe-Einsler M, Kimmich O, Kobeleva X, Önder D, Lesmann H, Kumar S, Tacik P, Basin MA, Incardona P, Lee-Kirsch MA, Berner R, Schuetz C, Körholz J, Kretschmer T, Di Donato N, Schröck E, Heinen A, Reuner U, Hanßke AM, Kaiser FJ, Manka E, Munteanu M, Kuechler A, Cordula K, Hirtz R, Schlapakow E, Schlein C, Lisfeld J, Kubisch C, Herget T, Hempel M, Weiler-Normann C, Ullrich K, Schramm C, Rudolph C, Rillig F, Groffmann M, Muntau A, Tibelius A, Schwaibold EMC, Schaaf CP, Zawada M, Kaufmann L, Hinderhofer K, Okun PM, Kotzaeridou U, Hoffmann GF, Choukair D, Bettendorf M, Spielmann M, Ripke A, Pauly M, Münchau A, Lohmann K, Hüning I, Hanker B, Bäumer T, Herzog R, Hellenbroich Y, Westphal DS, Strom T, Kovacs R, Riedhammer KM, Mayerhanser K, Graf E, Brugger M, Hoefele J, Oexle K, Mirza-Schreiber N, Berutti R, Schatz U, Krenn M, Makowski C, Weigand H, Schröder S, Rohlfs M, Vill K, Hauck F, Borggraefe I, Müller-Felber W, Kurth I, Elbracht M, Knopp C, Begemann M, Kraft F, Lemke JR, Hentschel J, Platzer K, Strehlow

V, Abou Jamra R, Kehrer M, Demidov G, Beck-Wödl S, Graessner H, Sturm M, Zeltner L, Schöls LJ, Magg J, Bevot A, Kehrer C, Kaiser N, Turro E, Horn D, Grüters-Kieslich A, Klein C, Mundlos S, Nöthen M, Riess O, Meitinger T, Krude H, Krawitz PM, Haack T, Ehmke N, Wagner M. Next-Generation Phenotyping Integrated in a National Framework for Patients with Ultrarare Disorders Improves Genetic Diagnostics and Yields New Molecular Findings. In: Nature Genetics 2024; 56 (8): 1644–1653. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-024-01836-1. URL: https://www.nature.com/articles/s41588-024-01836-1 (visited on 11/11/2024)

Shah A, Monger S, Troup M, Ip EK, Giannoulatou E. Systematic Evaluation of *de Novo* Mutation Calling Tools Using Whole Genome Sequencing Data. DOI: 10.1101/2024.08.28.610208. URL: http://biorxiv.org/lookup/doi/10.1101/2024.08.28.610208 (visited on 11/11/2024). Pre-published

Stoler N, Nekrutenko A. Sequencing Error Profiles of Illumina Sequencing Instruments. In: NAR Genomics and Bioinformatics 2021; 3 (1): lqab019. ISSN: 2631-9268. DOI: 10.1093/nargab/lqab019. URL: https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqab019/6193612 (visited on 11/11/2024)

Subramaniyan A, Wadden J, Goliya K, Ozog N, Wu X, Narayanasamy S, Blaauw D, Das R. Accelerated Seeding for Genome Sequence Alignment with Enumerated Radix Trees. In: *2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA)*. 2021 ACM/IEEE 48th Annual International Symposium on Computer Architecture (ISCA). Valencia, Spain: IEEE: pp. 388–401. ISBN: 978-1-66543-333-4. DOI: 10.1109/ISCA52012.2021.00038. URL: https://ieeexplore.ieee.org/document/9499792/ (visited on 11/11/2024)

Tate JG, Bamford S, Jubb HC, Sondka Z, Beare DM, Bindal N, Boutselakis H, Cole CG, Creatore C, Dawson E, Fish P, Harsha B, Hathaway C, Jupe SC, Kok CY, Noble K, Ponting L, Ramshaw CC, Rye CE, Speedy HE, Stefancsik R, Thompson SL, Wang S, Ward S, Campbell PJ, Forbes SA. COSMIC: The Catalogue Of Somatic Mutations In Cancer. In: Nucleic Acids Research 2019; 47 (D1): D941–D947. ISSN: 0305-1048, 1362-4962. DOI:

10.1093/nar/gky1015. URL: https://academic.oup.com/nar/article/47/D1/D941/5146192 (visited on 11/11/2024)

Vasimuddin Md, Misra S, Li H, Aluru S. Efficient Architecture-Aware Acceleration of BWA-MEM for Multicore Systems. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. 2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS). Rio de Janeiro, Brazil: IEEE: pp. 314–324. ISBN: 978-1-72811-246-6. DOI: 10.1109/IPDPS.2019.00041. URL: https://ieeexplore.ieee.org/document/8820962/ (visited on 11/11/2024)

Wagner J, Olson ND, Harris L, Khan Z, Farek J, Mahmoud M, Stankovic A, Kovacevic V, Yoo B, Miller N, Rosenfeld JA, Ni B, Zarate S, Kirsche M, Aganezov S, Schatz MC, Narzisi G, Byrska-Bishop M, Clarke W, Evani US, Markello C, Shafin K, Zhou X, Sidow A, Bansal V, Ebert P, Marschall T, Lansdorp P, Hanlon V, Mattsson CA, Barrio AM, Fiddes IT, Xiao C, Fungtammasan A, Chin CS, Wenger AM, Rowell WJ, Sedlazeck FJ, Carroll A, Salit M, Zook JM. Benchmarking Challenging Small Variants with Linked and Long Reads. In: Cell Genomics 2022; 2 (5): 100128. ISSN: 2666979X. DOI: 10.1016/j.xgen.2022.100128. URL: https://linkinghub.elsevier.com/retrieve/pii/S2666979X2200057X (visited on 11/11/2024)

Wdowiak A, Skrzypek M, Stec M, Panasiuk L. Effect of Ionizing Radiation on the Male Reproductive System. In: Annals of Agricultural and Environmental Medicine 2019; 26 (2): 210–216. ISSN: 1232-1966, 1898-2263. DOI: 10.26444/aaem/106085. URL: http://www.journalssystem.com/aaem/EFFECT-OF-IONIZING-RADIATION-ON-MALE-REPRODUCTIVE-SYSTEM,106085,0,2.html (visited on 10/23/2024)

Wong WS, Solomon BD, Bodian DL, Kothiyal P, Eley G, Huddleston KC, Baker R, Thach DC, Iyer RK, Vockley JG, et al. New Observations on Maternal Age Effect on Germline de Novo Mutations. In: Nature communications 2016; 7: 10486

Yeager M, Machiela MJ, Kothiyal P, Dean M, Bodelon C, Suman S, Wang M, Mirabello L, Nelson CW, Zhou W, Palmer C, Ballew B, Colli LM, Freedman ND, Dagnall C, Hutchinson A, Vij V, Maruvka Y, Hatch M, Illienko I, Belayev Y, Nakamura N, Chumak V, Bakhanova E, Belyi D, Kryuchkov V, Golovanov I, Gudzenko N, Cahoon EK, Albert P, Drozdovitch V,

Little MP, Mabuchi K, Stewart C, Getz G, Bazyka D, Gonzalez AB de, Chanock SJ. Lack of Transgenerational Effects of Ionizing Radiation Exposure from the Chernobyl Accident. In: Science 2021; 372 (6543): 725–729. DOI: 10.1126/science.abg2365

3 Publications

3.1 Next-generation phenotyping contributing to the identification of a 4.7 kb deletion in KANSL1 causing Koolen-de Vries syndrome

Brand, Vijayananth, Hsieh, Schmidt, Peters, Mangold, Cremer, Bender, Sivalingam, Hundertmark, Knaus, Engels, Krawitz, and Perne, “Next-generation Phenotyping Contributing to the Identification of a 4.7 Kb Deletion in *KANSL1* Causing Koolen-de Vries Syndrome”

Year: 2022
Journal: Human Mutation
DOI: https://doi.org/10.1002/humu.24467

Received: 18 April 2022 | Revised: 6 September 2022 | Accepted: 8 September 2022

DOI: 10.1002/humu.24467

SPECIAL ARTICLE

Human Mutation HGVS HUMAN GENOME VARIATION SOCIETY WILEY

Next-generation phenotyping contributing to the identification of a 4.7 kb deletion in *KANSL1* causing Koolen-de Vries syndrome

Fabian Brand[1] | Aswinkumar Vijayananth[1] | Tzung-Chien Hsieh[1] | Axel Schmidt[2] | Sophia Peters[2] | Elisabeth Mangold[2] | Kirsten Cremer[2] | Tim Bender[3] | Sugirthan Sivalingam[4] | Hela Hundertmark[2] | Alexej Knaus[1] | Hartmut Engels[2] | Peter M. Krawitz[1] | Claudia Perne[2]

[1]Institute for Genomic Statistics and Bioinformatics, Bonn, Germany

[2]Institute of Human Genetics, School of Medicine, University Hospital Bonn, University of Bonn, Bonn, Germany

[3]Center for Rare Disease, Medical Faculty, University of Bonn, Bonn, Germany

[4]Core Unit for Bioinformatics Data Analysis, Medical Faculty, University of Bonn, Bonn, Germany

Correspondence
Peter M. Krawitz, Institute for Genomic Statistics and Bioinformatics, Bonn, Germany.
Email: pkrawitz@uni-bonn.de

Abstract

Next-generation phenotyping (NGP) is an application of advanced methods of computer vision on medical imaging data such as portrait photos of individuals with rare disorders. NGP on portraits results in gestalt scores that can be used for the selection of appropriate genetic tests, and for the interpretation of the molecular data. Here, we report on an exceptional case of a young girl that was presented at the age of 8 and 15 and enrolled in NGP diagnostics on the latter occasion. The girl had clinical features associated with Koolen-de Vries syndrome (KdVS) and a suggestive facial gestalt. However, chromosomal microarray (CMA), Sanger sequencing, multiplex ligation-dependent probe analysis (MLPA), and trio exome sequencing remained inconclusive. Based on the highly indicative gestalt score for KdVS, the decision was made to perform genome sequencing to also evaluate noncoding variants. This analysis revealed a 4.7 kb de novo deletion partially affecting intron 6 and exon 7 of the *KANSL1* gene. This is the smallest reported structural variant to date for this phenotype. The case illustrates how NGP can be integrated into the iterative diagnostic process of test selection and interpretation of sequencing results.

KEYWORDS
Koolen-de Vries syndrome, next-generation phenotyping, structural variant, WGS

1 | INTRODUCTION

Many genetic syndromes are associated with a distinctive facial gestalt which can be used to expedite the diagnostic process. Although high-throughput sequencing has helped to address the considerable heterogeneity of many syndromes in a single test, the rare expertise of dysmorphologists, which is still required for data interpretation, is often the bottleneck. In recent years, advances in machine learning have enabled the development of NGP tools, that can be used to analyze facial dysmorphology in patient portrait

photos (Dudding-Byth et al., 2017; Ferry et al., 2014; Gripp et al., 2016; Gurovich et al., 2019; Hadj-Rabia et al., 2017; Hsieh et al., 2022; Kuru et al., 2014; Liehr et al., 2018; Valentine et al., 2017; van der Donk et al., 2019; Wang & Luo, 2016). Amongst them is GestaltMatcher, which is a deep convolutional neural network that was trained on thousands of molecularly confirmed cases and achieves high accuracies in the identification of hundreds of syndromes (Hsieh et al., 2022). In this paper, we describe how the results of this artificial intelligence helped to solve a case with a typical phenotype of Koolen-de Vries syndrome (KdVS) but an unusual disease-causing mutation.

2 | RESULTS

We report a female patient who first presented to a syndromic consultation at the age of eight because of multiple phenotypic abnormalities. The girl had muscular hypotonia since early childhood. During infancy a developmental delay became noticeable and later she scored in the moderate range of intellectual disability. Brain MRI showed two heterotopic foci as well as symmetrically clumped hippocampi. Facial dysmorphism, which became more prominent as a teenager, included a long face, slightly upslanting palpebral fissures, ptosis of the left eye, a prominent, bulbous nasal tip, and low-hanging columella (Figure 1). Furthermore, she had pale skin with many moles, thick curly hair, and a missing left upper canine tooth. Her family described her as extremely friendly, but anxious in contact with other children. A chromosome analysis, a chromosomal microarray (CMA), and diagnostics for fragile X syndrome, which have been performed after the first consultation at the age of 8 years, were unremarkable.

At re-consultation 7 years later, the 15-year-old female was enrolled in a study protocol that involved NGP and trio exome sequencing (Krawitz, 2022). After analysis by GestaltMatcher, the computer-assisted assessment of portrait images yielded high gestalt scores for KdVS (Figure 1). Although some characteristic aspects of the facial gestalt, such as the elongation of the face and the pear-shaped nose, were more prominent at re-consultation, the gestalt score for the portrait at the age of 8 years was already comparably high (Figure 1).

With facial dysmorphism typical for KdVS and some matching phenotypic features such as a friendly personality, structural brain anomaly, anxiety, and the many moles, this diagnosis was suspected despite the inconclusive CMA results. According to the literature, approximately 95% of the cases with KdVS are due to 500–650 kb deletions in 17q21.31 and only approximately 5% are due to sequence variants in *KANSL1* (Koolen et al., 2006; Koolen et al., 2012; Koolen et al., 2016; Sharp et al., 2006; Shaw-Smith et al., 2006; Zollino et al., 2012; Zollino et al., 2015). Although a recent study

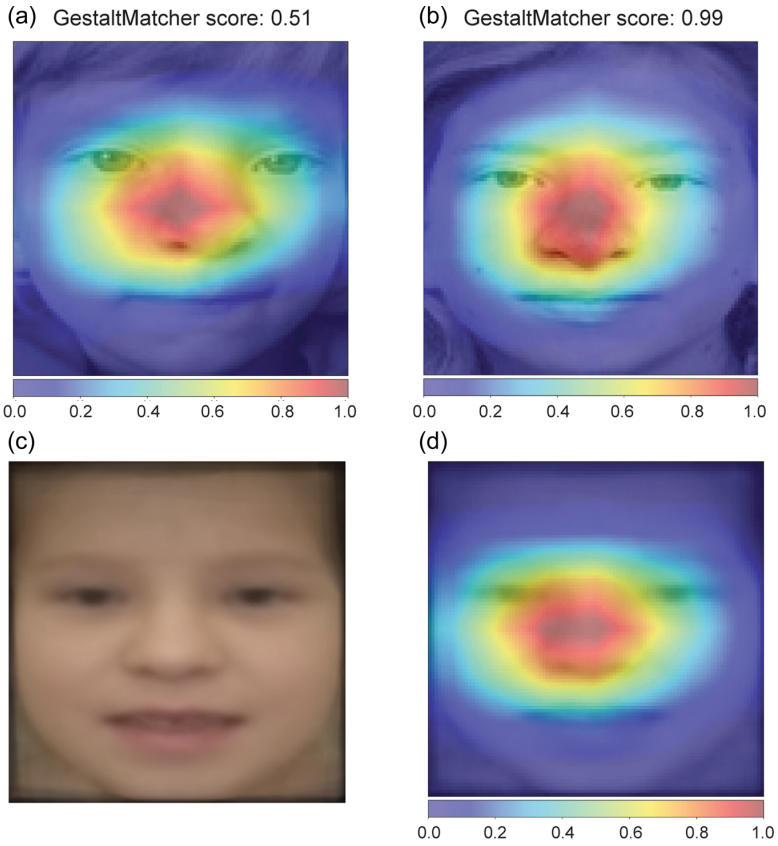


FIGURE 1 The pixel attribution maps for the KdVS class and the composite face of KdVS. Pixel attribution maps of (a) patient at the age of 8; (b) patient at the age of 15; (c) KdVS composite face. (d) the composite face of KdVS. Attribution maps show the prominence of the nose region for the classifier's prediction. Attribution maps of the patient's photos show high similarity with that of the composite face. The GestaltMatcher score ranges from 0 to 1. The value of one is the highest value indicating high similarity to the disorder.

indicates that the proportion of patients carrying a (likely) pathogenic sequence variant may be about 25%, recurrent deletion remains the predominant mutation type (Dingemans et al., 2021). Around the microdeletion in 17q21.31 large clusters of low complexity repeats at the breakpoints were described, suggesting an underlying mechanism of non-allelic homologous recombination (NAHR) (Dubourg et al., 2011; Stankiewicz & Lupski, 2002). Up to now, these deletions have been found by CMA. So far, only a few atypical deletions had been reported for individuals affected by KdVS, the smallest of these still 68 kb in size (Cooper et al., 2011; Dubourg et al., 2011; Koolen et al., 2012; Zollino et al., 2015). All of these deletions were also detected by CMA.

As the recurrent microdeletion in 17q21.31 was not supported by CMA we initiated Sanger sequencing and multiplex ligation-dependent probe amplification (MLPA) of *KANSL1*. Both analyses did not show any abnormal findings. We assume that allelic dropout is the reason for negative Sanger sequencing, whereas in MLPA the probe oligos did not bind to the relevant part of exon 7 and therefore missed the deletion. Next, in accordance with the study protocol, a trio exome analysis of the patient and her parents were performed. Data for the patient and her parents was generated using the NovaSeq platform (Illumina) and the SureSelect v6 exome capture kit (Agilent). Initial bioinformatics analysis was focused on relevant single nucleotide variants (SNVs) and indels using a local implementation of GATK best practices pipelines optimized for data from the NovaSeq sequencer. Copy number variants (CNVs) were initially generated using cn.MOPS (Klambauer et al., 2012). No variants that could explain the phenotype were detected in *KANSL1* nor any other gene. Following the inconclusive results of the trio exome analysis, genome sequencing was conducted to look for intronic sequence variants and structural variants missed by exome analysis and CMA. The bioinformatics analysis was performed using the NVIDIA Parabricks toolkit. This toolkit enables accelerated genome analysis by utilizing NVIDIA GPU resources. Several algorithms from this toolkit have been used to call SNVs and indels on the patient's genomic data. In particular, accelerated versions of BWA-mem and the HaplotypeCaller were crucial for fast processing and yielded variant calls of high quality. To determine candidates for structural variants (SVs) and CNVs, we used manta (Chen et al., 2016), delly (Rausch et al., 2012), and lumpy (Layer et al., 2014). Variant calls of all three tools were merged using a vote-based scheme to find candidates supported by all callers. A 4,708 bp deletion affecting the end of intron 6 and only the first 46 bp of exon 7 (NM_015443.4:c.1849-4611_1895del) was detected by all three tools. Furthermore, the deletion was also clearly visible by a drop of coverage and by split reads in the sequence alignment (Figure 2). In a careful reanalysis of the exome data, which was guided by the results from genome sequencing data, the deletion could also be detected using Pindel (Ye et al., 2009) (Figure 2). The deletion is supported by 16 split-read pairs, indicating a *de novo* origin of the variant. Adding further support for this hypothesis, the mean insert size of reads in the region of the deletion is 678.3(±1417.4) in the study participant and 215.3(±72.7) or 223.7(±79.9) in mother and father respectively. Changing some alignments preferences in Integrative Genome Viewer enabled the visualization of the deletion in *KANSL1* exome sequencing data (Figure 2). The deletion was also subsequently verified by qPCR.

To support our claims on the effectiveness of GestaltMatcher, we interpreted its predictions on the case's facial images using the Gradient-weighted Class Activation Mapping (Grad-CAM) (Selvaraju et al., 2017). This is a technique from the realm of Explainable AI, that can generate pixel attribution maps for GestaltMatcher's predictions on the case's facial photos, as shown in Figure 1. Pixel attribution maps highlight the pixels that were relevant for a certain image classification by a neural network (Molnar, 2020), such as a portrait to a syndrome by GestaltMatcher. The higher the attribution score of a pixel is, the more its contribution to a chosen target prediction of a class.

We generated attribution maps for the case's facial images (Figure 1a,b), with respect to the KdVS class. Following, we analyzed the maps in two different ways. In the first analysis, we looked for any association between the regions highlighted in the maps and phenotypic traits of KdVS. As it is evident from the maps, the classifier model focuses predominantly on the nose region, which can be related to the prominent bulbous nasal tip of KdVS. For the second analysis, we compared the attribution maps of the case with that of the composite face of KdVS (Figure 1d). The composite face (Figure 1c) provided a characteristic representation of the facial phenotype of the syndrome and was generated by combining the facial photos of KdVS patients in the GestaltMatcher Database (GMDB) (Hsieh et al., 2022). The similarity between attribution maps of the composite face and the case's photos showed the prominence of the syndrome's phenotype in the face of our patient. It also provided the rationale for the classifier's high confidence (GestaltMatcher score) in predicting the syndrome.

3 | DISCUSSION

Many SV and CNV tools for exome data rely on depth of coverage signals to identify likely candidates for structural changes in the genome in short read Illumina data. For both, exome and genome data, the effectiveness of this approach is limited by the availability of good normalized control data from other genomic regions in the same individual or other individuals of the same sequencing run. In case of the trio-exome sequencing experiment from our patient, this baseline was formed by other unrelated samples sequenced in parallel. Depth and variability of the coverage in certain genomic regions also has an influence on the ability of those callers to detect structural change to the genome. Other CNV detection methods rely on a mix of other factors (e.g., split-read pairs, insert- and fragment-sizes) to find likely candidates for variation. Pindel incorporates signals from split reads. These are read pairs in which one of the two reads cannot be aligned to the reference genome and is assumed to carry the precise breakpoint information of insertion or deletion events. Similar metrics are used also by other callers that were used for subsequent genome sequencing data analysis (e.g., manta, delly, and lumpy).

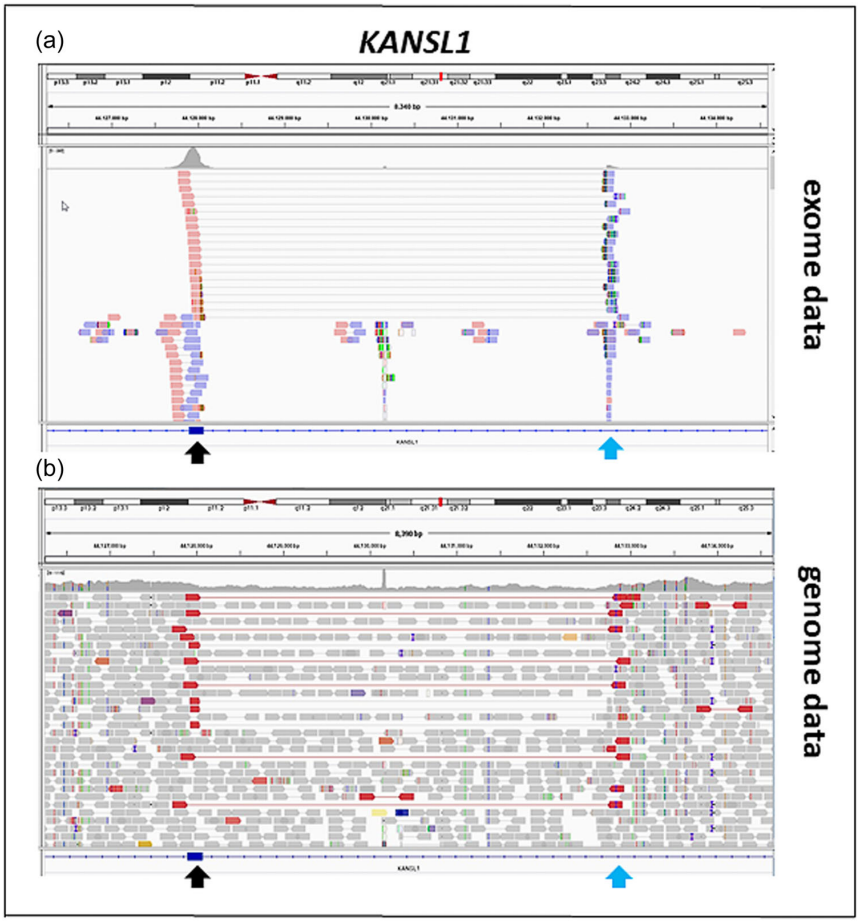


FIGURE 2 *KANSL1* whole exome and genome sequencing data. *KANSL1* sequencing data visualized in IGV. (a) A screenshot of the deletion in exome sequencing data. Reads are sorted by start location and grouped by read pairs. Soft clipped bases are included, making the exact breakpoint in DNA visible (NM_015443.4:c.1849-4611_1895del), even against the complete lack of other reads in the region. (b) A screenshot of genome sequencing data is shown. The deletion causes a noticeable drop in coverage. Additional support for the detected deletion is provided by split reads, marked in red. Black arrow = 3′- end of deletion (Exon 7), blue arrow = 5′-end of deletion (intron 6).

A combination of tools that feature different detection methods is necessary to increase the sensitivity for structural changes, especially around the edges of the target region of exome data, like in our case. Long-read technologies like PacBio or Oxford Nanopore sequencing can overcome these limitations of Illumina data. In particular, with the clinical evidence and NGP, both suggesting KdVS , Oxford Nanopore sequencing utilizing the ReadUntil protocol enables the screening for deletions in the size range currently missed by the standard protocols (Kovaka et al., 2021; Payne et al., 2021).

The initial negative result using other CNV calling methods is due to the suboptimal coverage distribution at some of the *KANSL1* exons and intronic regions and the fact that the deletion reaches only 46 bp into exon 7. The variant in question is mainly in the end of intron 6 making coverage-based detection of structural variants based on changes in coverage substantially more difficult than by incorporating split-read signals or using genome sequencing data. As a result, from sequence analysis, 130 pathogenic or likely pathogenic variants have been reported for *KANSL1* in the database ClinVar (Landrum et al., 2020). In contrast, the 4.7 kb deletion that we identified, is the first entry in ClinVar for a variant length in between 51 bp and 50 kb.

In conclusion, we reported a 4.7 kb deletion in *KANSL1* that is mainly noncoding and was therefore first detected by genome sequencing. However, retrospectively it could also be confirmed in exome sequencing data with fine-tuning of the filter settings. Since high accuracy in CMA analysis is limited to a resolution of 50 kb or higher, and in exome analysis to a resolution of 50 bp or lower, deletions in the order of few kilobases are not detected in the diagnostic tests most often used today. In genome sequencing data, on the other hand SV and CNVs in this size range can be identified more easily, but are usually more difficult to interpret, if they are noncoding.

Therefore, our case exemplifies, how computer-assisted analysis of the portrait can significantly contribute to the diagnostic process. First, NGP has the potential to speed up data analysis. If our Koolen-de Vries patient would have carried the recurrent microdeletion, a SNV, or indel, the high gestalt score would have made the molecular confirmation of the suspected clinical diagnosis straightforward using protocols such as the PEDIA workflow (Hsieh et al., 2019). Second, highly suggestive results of NGP can be used to request genome sequencing if exome or CMA analysis were inconclusive. Third, NGP can help with the classification of the pathogenicity of novel variants found in the genome. This is true for approximately 30%–40% of the more than 7000 rare syndromes involving dysmorphic craniofacial features (Hart & Hart, 2009). It is also noteworthy, that our patient with KdVS did not present with very specific clinical features apart from a characteristic gestalt, and therefore did not achieve a high feature score. In these cases, gene-prioritization algorithms that work only with HPO terms do not perform well (Robinson et al., 2008). In our patient none of the tools that we tested, yielded KdVS in the top-10 suggested differential diagnoses (Kohler et al., 2009; Peng et al., 2021; Zhao et al., 2020).

For the further development of NGP into medical products, certain requirements for software with AI have to be addressed. An important requirement is explaining the results of the AI so that physicians can reason about the decision of the model. This field of "Explainable AI" has recently evolved as an independent area of research. Generating pixel attribution maps using the methods like Grad-CAM, such as the ones displayed in Figure 1, is a common way to obtain some reasoning for the model's decision by backpropagating the selected syndrome class's score through the network and visualizing the differences with the input images. It helps us to understand why the given image is classified as a certain disorder, and we could further check whether this classification aligns with our clinical interpretation. In our case, we can see that the region highlighted in Figure 1 highly aligns with the clinical feature of a prominent bulbous nasal tip in KdVS. However, more techniques such as occlusion sensitivity mapping and counterfactual explanations should be explored for better interpretability in the future.

According to the guidelines from 2015 for the classification of sequence variants, a matching phenotype is only considered as supporting evidence for the pathogenicity of a sequence variant (PP4) (Richards et al., 2015). However, experienced dysmorphologists may attribute a higher level of evidence to the pathogenicity of a variant in a gene if the associated phenotype is highly specific (Zhang et al., 2020). Most clinicians that are confronted for the first time with such a specific diagnosis will be hesitant to apply these higher weights. Here, NGP could help, since syndromic distinctiveness can be measured and the similarity of a portrait to other molecularly confirmed cases can be quantified (Hsieh et al., 2022). By this means, NGP makes the visual inspection of a patient applicable to a Bayesian classification framework (Tavtigian et al., 2018). Interestingly, the distinctiveness of the facial gestalt of KdVS ranges only in the upper half of dysmorphic phenotypes and is exceeded for example by the distinctiveness of the facial gestalt of Baraitser–Winter syndrome or Seckel syndrome. This means if such syndromes score high gestalt scores in NGP and no pathogenic variant can be identified in an exome, genome sequencing might be indicated.

ACKNOWLEDGMENTS

The authors would like to thank Dr. Jessica Trautmann and Nuria Ortega Ibañez from the Institute of Human Genetics, University Hospital Bonn, Germany for performing chromosomal micro array and qPCR studies. The authors are grateful to the patient and her familiy for their cooperation. Open Access funding enabled and organized by Projekt DEAL.

CONFLICTS OF INTEREST

The authors declare no conflicts of interest.

ORCID

Peter M. Krawitz http://orcid.org/0000-0002-3194-8625

REFERENCES

Chen, X., Schulz-Trieglaff, O., Shaw, R., Barnes, B., Schlesinger, F., Kallberg, M., Cox, A. J., Kruglyak, S., & Saunders, C. T. (2016). Manta: Rapid detection of structural variants and indels for germline and cancer sequencing applications. *Bioinformatics*, *32*(8), 1220–1222. https://doi.org/10.1093/bioinformatics/btv710

Cooper, G. M., Coe, B. P., Girirajan, S., Rosenfeld, J. A., Vu, T. H., Baker, C., Williams, C., Stalker, H., Hamid, R., Hannig, V., Abdel-Hamid, H., Bader, P., McCracken, E., Niyazov, D., Leppig, K., Thiese, H., Hummel, M., Alexander, N., Gorski, J., … Eichler, E. E. (2011). A copy number variation morbidity map of developmental delay. *Nature Genetics*, *43*(9), 838–846. https://doi.org/10.1038/ng.909

Dingemans, A. J. M., Stremmelaar, D. E., van der Donk, R., Vissers, L., Koolen, D. A., Rump, P., Hehir-Kwa, J. Y., & de Vries, B. B. A. (2021). Quantitative facial phenotyping for koolen-de vries and 22q11.2 deletion syndrome. *European Journal of Human Genetics*, *29*(9), 1418–1423. https://doi.org/10.1038/s41431-021-00824-x

van der Donk, R., Jansen, S., Schuurs-Hoeijmakers, J. H. M., Koolen, D. A., Goltstein, L., Hoischen, A., Brunner, H. G., Kemmeren, P., Nellaker, C., Vissers, L., de Vries, B. B. A., & Hehir-Kwa, J. Y. (2019). Next-generation phenotyping using computer vision algorithms in rare genomic neurodevelopmental disorders. *Genetics in Medicine*, *21*(8), 1719–1725. https://doi.org/10.1038/s41436-018-0404-y

Dubourg, C., Sanlaville, D., Doco-Fenzy, M., Le Caignec, C., Missirian, C., Jaillard, S., Schluth-Bolard, C., Landais, E., Boute, O., Philip, N., Toutain, A., David, A., Edery, P., Moncla, A., Martin-Coignard, D., Vincent-Delorme, C., Mortemousque, I., Duban-Bedu, B., Drunat, S., … Andrieux, J. (2011). Clinical and molecular characterization of 17q21.31 microdeletion syndrome in 14 French patients with mental retardation. *European Journal of Medical Genetics*, *54*(2), 144–151. https://doi.org/10.1016/j.ejmg.2010.11.003

Dudding-Byth, T., Baxter, A., Holliday, E. G., Hackett, A., O'Donnell, S., White, S. M., Attia, J., Brunner, H., de Vries, B., Koolen, D., Kleefstra, T., Ratwatte, S., Riveros, C., Brain, S., & Lovell, B. C. (2017). Computer face-matching technology using two-dimensional photographs accurately matches the facial gestalt of unrelated individuals with the same syndromic form of intellectual disability. *BMC Biotechnology*, *17*(1), 90. https://doi.org/10.1186/s12896-017-0410-1

Ferry, Q., Steinberg, J., Webber, C., FitzPatrick, D. R., Ponting, C. P., Zisserman, A., & Nellaker, C. (2014). Diagnostically relevant facial

gestalt information from ordinary photos. *eLife*, *3*, e02020. https://doi.org/10.7554/eLife.02020

Gripp, K. W., Baker, L., Telegrafi, A., & Monaghan, K. G. (2016). Jul The role of objective facial analysis using FDNA in making diagnoses following whole exome analysis. Report of two patients with mutations in the BAF complex genes. *American Journal of Medical Genetics. Part A*, *170*(7), 1754–1762. https://doi.org/10.1002/ajmg.a.37672

Gurovich, Y., Hanani, Y., Bar, O., Nadav, G., Fleischer, N., Gelbman, D., Basel-Salmon, L., Krawitz, P. M., Kamphausen, S. B., Zenker, M., Bird, L. M., & Gripp, K. W. (2019). Identifying facial phenotypes of genetic disorders using deep learning. *Nature Medicine (New York, NY, United States)*, *25*(1), 60–64. https://doi.org/10.1038/s41591-018-0279-0

Hadj-Rabia, S., Schneider, H., Navarro, E., Klein, O., Kirby, N., Huttner, K., Wolf, L., Orin, M., Wohlfart, S., Bodemer, C., & Grange, D. K. (2017). Automatic recognition of the XLHED phenotype from facial images. *American Journal of Medical Genetics. Part A*, *173*(9), 2408–2414. https://doi.org/10.1002/ajmg.a.38343

Hart, T. C., & Hart, P. S. (2009). Genetic studies of craniofacial anomalies: Clinical implications and applications. *Orthodontics & Craniofacial Research*, *12*(3), 212–220. https://doi.org/10.1111/j.1601-6343.2009.01455.x

Hsieh, T. C., Bar-Haim, A., Moosa, S., Ehmke, N., Gripp, K. W., Pantel, J. T., Danyel, M., Mensah, M. A., Horn, D., Rosnev, S., Fleischer, N., Bonini, G., Hustinx, A., Schmid, A., Knaus, A., Javanmardi, B., Klinkhammer, H., Lesmann, H., Sivalingam, S., … Krawitz, P. M. (2022). GestaltMatcher facilitates rare disease matching using facial phenotype descriptors. *Nature Genetics*, *54*(3), 349–357. https://doi.org/10.1038/s41588-021-01010-x

Hsieh, T. C., Mensah, M. A., Pantel, J. T., Aguilar, D., Bar, O., Bayat, A., Becerra-Solano, L., Bentzen, H. B., Biskup, S., Borisov, O., Braaten, O., Ciaccio, C., Coutelier, M., Cremer, K., Danyel, M., Daschkey, S., Eden, H. D., Devriendt, K., Wilson, S., … Krawitz, P. M. (2019). PEDIA: Prioritization of exome data by image analysis. *Genetics in Medicine*, *21*(12), 2807–2814. https://doi.org/10.1038/s41436-019-0566-2

Klambauer, G., Schwarzbauer, K., Mayr, A., Clevert, D. A., Mitterecker, A., Bodenhofer, U., & Hochreiter, S. (2012). cn.MOPS: Mixture of Poissons for discovering copy number variations in next-generation sequencing data with a low false discovery rate. *Nucleic Acids Research*, *40*(9), e69. https://doi.org/10.1093/nar/gks003

Kohler, S., Schulz, M. H., Krawitz, P., Bauer, S., Dolken, S., Ott, C. E., Mundlos, C., Horn, D., Mundlos, S., & Robinson, P. N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. *American Journal of Human Genetics*, *85*(4), 457–464. https://doi.org/10.1016/j.ajhg.2009.09.003

Koolen, D. A., Kramer, J. M., Neveling, K., Nillesen, W. M., Moore-Barton, H. L., Elmslie, F. V., Toutain, A., Amiel, J., Malan, V., Tsai, A. C., Cheung, S. W., Gilissen, C., Verwiel, E. T., Martens, S., Feuth, T., Bongers, E. M., de Vries, P., Scheffer, H., Vissers, L. E., … de Vries, B. B. (2012). Mutations in the chromatin modifier gene KANSL1 cause the 17q21.31 microdeletion syndrome. *Nature Genetics*, *44*(6), 639–641. https://doi.org/10.1038/ng.2262

Koolen, D. A., Pfundt, R., Linda, K., Beunders, G., Veenstra-Knol, H. E., Conta, J. H., Fortuna, A. M., Gillessen-Kaesbach, G., Dugan, S., Halbach, S., Abdul-Rahman, O. A., Winesett, H. M., Chung, W. K., Dalton, M., Dimova, P. S., Mattina, T., Prescott, K., Zhang, H. Z., Saal, H. M., … de Vries, B. B. (2016). The Koolen-de Vries syndrome: A phenotypic comparison of patients with a 17q21.31 microdeletion versus a KANSL1 sequence variant, *European Journal of Human Genetics 24*, 5. 652–659. https://doi.org/10.1038/ejhg.2015.178

Koolen, D. A., Vissers, L. E., Pfundt, R., de Leeuw, N., Knight, S. J., Regan, R., Kooy, R. F., Reyniers, E., Romano, C., Fichera, M., Schinzel, A., Baumer, A., Anderlid, B. M., Schoumans, J., Knoers, N. V., van Kessel, A. G., Sistermans, E. A., Veltman, J. A., Brunner, H. G., & de Vries, B. B. (2006). A new chromosome 17q21.31 microdeletion syndrome associated with a common inversion polymorphism. *Nature Genetics*, *38*(9), 999–1001. https://doi.org/10.1038/ng1853

Kovaka, S., Fan, Y., Ni, B., Timp, W., & Schatz, M. C. (2021). Targeted nanopore sequencing by real-time mapping of raw electrical signal with UNCALLED. *Nature Biotechnology*, *39*, 4. 431–441. https://doi.org/10.1038/s41587-020-0731-9

Krawitz, P. (2022). A national diagnostic framework for patients with ultra-rare disorders: molecular genetic findings using phenotypic and sequencing data.

Kuru, K., Niranjan, M., Tunca, Y., Osvank, E., & Azim, T. (2014). Biomedical visual data analysis to build an intelligent diagnostic decision support system in medical genetics. *Artificial Intelligence in Medicine*, *62*(2), 105–118. https://doi.org/10.1016/j.artmed.2014.08.003

Landrum, M. J., Chitipiralla, S., Brown, G. R., Chen, C., Gu, B., Hart, J., Hoffman, D., Jang, W., Kaur, K., Liu, C., Lyoshin, V., Maddipatla, Z., Maiti, R., Mitchell, J., O'Leary, N., Riley, G. R., Shi, W., Zhou, G., Schneider, V., … Kattman, B. L. (2020). ClinVar: Improvements to accessing data. *Nucleic Acids Research*, *48*(D1), D835–D844. https://doi.org/10.1093/nar/gkz972

Layer, R. M., Chiang, C., Quinlan, A. R., & Hall, I. M. (2014). LUMPY: A probabilistic framework for structural variant discovery. *Genome Biology*, *15*(6), R84. https://doi.org/10.1186/gb-2014-15-6-r84

Liehr, T., Acquarola, N., Pyle, K., St-Pierre, S., Rinholm, M., Bar, O., Wilhelm, K., & Schreyer, I. (2018). Next generation phenotyping in Emanuel and Pallister-Killian syndrome using computer-aided facial dysmorphology analysis of 2D photos. *Clinical Genetics*, *93*(2), 378–381. https://doi.org/10.1111/cge.13087

Molnar, C. (2020). *Interpretable Machine Learning*. https://www.Lulu.com

Payne, A. Holmes, N.,Clarke, T.,Munro, R., Debebe, B. J., & Loose, M. (2021). Readfish enables targeted nanopore sequencing of gigabase-sized genomes. *Nature Biotechnology*, *39*(4), 442–450. https://doi.org/10.1038/s41587-020-00746-x

Peng, C., Dieck, S., Schmid, A., Ahmad, A., Knaus, A., Wenzel, M., Mehnert, L., Zirn, B., Haack, T., Ossowski, S., Wagner, M., Brunet, T., Ehmke, N., Danyel, M., Rosnev, S., Kamphans, T., Nadav, G., Fleischer, N., Frohlich, H., & Krawitz, P. (2021). CADA: Phenotype-driven gene prioritization based on a case-enriched knowledge graph. *NAR Genomics Bioinformatics*, *3*(3), lqab078. https://doi.org/10.1093/nargab/lqab078

Rausch, T., Zichner, T., Schlattl, A., Stutz, A. M., Benes, V., & Korbel, J. O. (2012). DELLY: Structural variant discovery by integrated paired-end and split-read analysis. *Bioinformatics*, *28*(18), i333–i339. https://doi.org/10.1093/bioinformatics/bts378

Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W. W., Hegde, M., Lyon, E., Spector, E., Voelkerding, K., Rehm, H. L., & Committee, A. L. Q. A. (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. *Genetics in Medicine*, *17*(5), 405–424. https://doi.org/10.1038/gim.2015.30

Robinson, P. N., Kohler, S., Bauer, S., Seelow, D., Horn, D., & Mundlos, S. (2008). The human phenotype ontology: A tool for annotating and analyzing human hereditary disease. *American Journal of Human Genetics*, *83*(5), 610–615. https://doi.org/10.1016/j.ajhg.2008.09.017

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization, 2017 IEEE International Conference on Computer Vision (ICCV), 618–626.

Sharp, A. J., Hansen, S., Selzer, R. R., Cheng, Z., Regan, R., Hurst, J. A., Stewart, H., Price, S. M., Blair, E., Hennekam, R. C., Fitzpatrick, C. A., Segraves, R., Richmond, T. A., Guiver, C., Albertson, D. G., Pinkel, D.,

Eis, P. S., Schwartz, S., Knight, S. J., & Eichler, E. E. (2006). Discovery of previously unidentified genomic disorders from the duplication architecture of the human genome. *Nature Genetics*, *38*(9), 1038–1042. https://doi.org/10.1038/ng1862

Shaw-Smith, C., Pittman, A. M., Willatt, L., Martin, H., Rickman, L., Gribble, S., Curley, R., Cumming, S., Dunn, C., Kalaitzopoulos, D., Porter, K., Prigmore, E., Krepischi-Santos, A. C., Varela, M. C., Koiffmann, C. P., Lees, A. J., Rosenberg, C., Firth, H. V., de Silva, R., & Carter, N. P. (2006). Microdeletion encompassing MAPT at chromosome 17q21.3 is associated with developmental delay and learning disability. *Nature Genetics*, *38*(9), 1032–1037. https://doi.org/10.1038/ng1858

Stankiewicz, P., & Lupski, J. R. (2002). Genome architecture, rearrangements and genomic disorders. *Trends in Genetics*, *18*(2), 74–82. https://doi.org/10.1016/s0168-9525(02)02592-1

Tavtigian, S. V., Greenblatt, M. S., Harrison, S. M., Nussbaum, R. L., Prabhu, S. A., Boucher, K. M., Biesecker, L. G., & ClinGen Sequence Variant Interpretation Working, G. (2018). Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. *Genetics in Medicine*, *20*(9), 1054–1060. https://doi.org/10.1038/gim.2017.210

Valentine, M., Bihm, D. C. J., Wolf, L., Hoyme, H. E., May, P. A., Buckley, D., Kalberg, W., & Abdul-Rahman, O. A. (2017). Computer-Aided recognition of facial attributes for fetal alcohol spectrum disorders. *Pediatrics*, *140*(6), e20162028. https://doi.org/10.1542/peds.2016-2028

Wang, K., & Luo, J. (2016). Detecting visually observable disease symptoms from faces. *EURASIP journal on bioinformatics & systems biology*, *2016*(1), 13. https://doi.org/10.1186/s13637-016-0048-7

Zhang, J., Yao, Y., He, H., & Shen, J. (2020). Jun Clinical interpretation of sequence variants. *Current Protocols in Human Genetics*, *106*(1), e98. https://doi.org/10.1002/cphg.98

Zhao, M., Havrilla, J. M., Fang, L., Chen, Y., Peng, J., Liu, C., Wu, C., Sarmady, M., Botas, P., Isla, J., Lyon, G. J., Weng, C., & Wang, K. (2020). Phen2Gene: rapid phenotype-driven gene prioritization for rare diseases. *NAR Genom Bioinform*, *2*(2), lqaa032. https://doi.org/10.1093/nargab/lqaa032

Zollino, M., Marangi, G., Ponzi, E., Orteschi, D., Ricciardi, S., Lattante, S., Murdolo, M., Battaglia, D., Contaldo, I., Mercuri, E., Stefanini, M. C., Caumes, R., Edery, P., Rossi, M., Piccione, M., Corsello, G., Della Monica, M., Scarano, F., Priolo, M., … Zackai, E. (2015). Intragenic KANSL1 mutations and chromosome 17q21.31 deletions: broadening the clinical spectrum and genotype-phenotype correlations in a large cohort of patients. *Journal of Medical Genetics*, *52*(12), 804–814. https://doi.org/10.1136/jmedgenet-2015-103184

Zollino, M., Orteschi, D., Murdolo, M., Lattante, S., Battaglia, D., Stefanini, C., Mercuri, E., Chiurazzi, P., Neri, G., & Marangi, G. (2012). Mutations in KANSL1 cause the 17q21.31 microdeletion syndrome phenotype. *Nature Genetics*, *44*(6), 636–638. https://doi.org/10.1038/ng.2257

How to cite this article: Brand, F., Vijayananth, A., Hsieh, T.-C., Schmidt, A., Peters, S., Mangold, E., Cremer, K., Bender, T., Sivalingam, S., Hundertmark, H., Knaus, A., Engels, H., Krawitz, P. M., & Perne, C. (2022). Next-generation phenotyping contributing to the identification of a 4.7 kb deletion in *KANSL1* causing Koolen-de Vries syndrome. *Human Mutation*, 43, 1659–1665. https://doi.org/10.1002/humu.24467

3.2 Extending DeepTrio for sensitive detection of complex de novo mutation patterns

Brand, Guski, and Krawitz, “Extending DeepTrio for Sensitive Detection of Complex *de Novo* Mutation Patterns”

Year: 2024
Journal: NAR Genomics and Bioinformatics
DOI: https://doi.org/10.1093/nargab/lqae013

NAR Genomics and Bioinformatics, 2024, **6**, lqae013
https://doi.org/10.1093/nargab/lqae013
Advance access publication date: 10 February 2024
Standard Article

OXFORD

Extending DeepTrio for sensitive detection of complex *de novo* mutation patterns

Fabian Brand[†], **Jannis Guski**[†] **and Peter Krawitz**[*]

Institut für Genomische Statistik und Bioinformatik (IGSB), University of Bonn, Bonn, Germany

[*]To whom correspondence should be addressed. Tel: +49 228 287 14733; Email: pkrawitz@uni-bonn.de
[†]The first two authors should be regarded as Joint First Authors.

Abstract

De novo mutations (DNMs), and among them clustered DNMs within 20 bp of each other (cDNMs) are known to be a potential cause of genetic disorders. However, identifying DNM in whole genome sequencing (WGS) data is a process that often suffers from low specificity. We propose a deep learning framework for DNM and cDNM detection in WGS data based on Google's DeepTrio software for variant calling, which considers regions of 110 bp up- and downstream from possible variants to take information from the surrounding region into account. We trained a model each for the DNM and cDNM detection tasks and tested it on data generated on the HiSeq and NovaSeq platforms. In total, the model was trained on 82 WGS trios generated on the NovaSeq and 16 on the HiSeq. For the DNM detection task, our model achieves a sensitivity of 95.7% and a precision of 89.6%. The extended model adds confidence information for cDNMs, in addition to standard variant classes and DNMs. While this causes a slight drop in DNM sensitivity (91.96%) and precision (90.5%), on HG002 cDNMs can be isolated from other variant classes in all cases (5 out of 5) with a precision of 76.9%. Since the model emits confidence probabilities for each variant class, it is possible to fine-tune cutoff thresholds to allow users to select a desired trade-off between sensitivity and specificity. These results show that DeepTrio can be retrained to identify complex mutational signatures with only little modification effort.

Introduction

All sequence variants can be attributed to exogenous and endogenous exposures to mutagens and errors in DNA replication that are not corrected by the cell's repair mechanisms (1). Such *de novo* mutations (DNMs) are not only a significant cause of cancer and genetic disorders but also the driver of evolution and thus of interest to medicine and our understanding of genome biology. The mutational processes and their influencing factors are best characterized for single nucleotide variants (SNVs) and INDELs that have been analyzed in cancer and germline cells by means of trio sequencing (2,3). However, for studying more complex mutational signatures, sequencing errors, uneven coverage and mapping artifacts still represent major challenges for accurate detection. Such artifacts can arise, for example, with varying chemistry or sequencing devices that are used to generate the sequence data. One particularly notable change occurs when switching from HiSeq X Ten devices with four-color chemistry to NovaSeq 6000 devices with two-color chemistry. It has been noted that the NovaSeq type of devices emits specific signatures in its artifacts (4). Such signatures also provide interesting targets for the analysis, since the detection of those can substantially improve variant calling accuracy in difficult scenarios.

Recently, complex mutational signatures have also been recognized to be involved in many processes, including the exposure of the human germline to mutagens that are known to cause cancer, such as ionizing radiation, familial mutation rates and hypermutation in the germline (3,5,6). Clustered *de novo* mutations (cDNMs, at least two *de novo* mutations within 20 bp) for example are a known signature for prolonged paternal exposure to ionizing radiation, which are hard to detect accurately with state-of-the-art statistical and heuristic methods (7). These and other complex types of mutations and rearrangements however are increasingly relevant for the assessment of environmental or otherwise smaller effects that are hard to distinguish in front of the background of Mendelian variation and sequencing device and chemistry-specific artifacts that might also vary from sequencing site to sequencing site (4,8) and need to be detected with great accuracy on large cohorts to enable statistical analysis (9,10). Since an extensive validation of mutational signatures is not possible in most settings due to cost and time constraints, algorithms are needed that can detect arbitrary complex mutational signatures accurately and with verifiable properties that allow users to fit the false discovery rates to the need of their analysis.

DeepVariant and its extension DeepTrio are convolutional neural network-based variant calling tools that are able to detect Mendelian sequence variation with high accuracy. DeepTrio extends the standard single-sample variant calling approach implemented by DeepVariant by incorporating parental information into the variant calling process. Using this additional information, the network is able to improve the detection accuracy for standard variants (11,12).

We developed a framework for retraining DeepVariant and DeepTrio on specific mutational patterns to enable the detection or suppression of such signatures with high accuracy at the same time as germline variant calling. We demonstrate this capability by retraining the networks for the detection of DNMs and cDNMs. In addition, we show how the

Received: July 3, 2023. **Revised:** January 16, 2024. **Editorial Decision:** January 22, 2024. **Accepted:** January 23, 2024

output of this *de novo* caller can be adjusted for high sensitivity or a high positive predictive value, depending on the specific application. Previous DNM calling approaches either relied on statistical and heuristic properties of DNMs for detection (e.g. DeNovoGear, our in-house pipeline) or used Deep Learning (e.g. DeNovoCNN). We show how the DeepTrio network can be retrained in a generic fashion to support the detection of any mutational signature with high accuracy even when trained with small amounts of trustworthy variants. This is necessary for many complex variants such as cDNMs or structural variants, as validation of all calls by complementary approaches such as Sanger sequencing is not feasible (13).

In this work, we illustrate our approach with the detection of cDNMs that are characterized by multiple lesions in close proximity of the DNA sequence and that can also be radiation-induced (14). Subsets of cDNM calls have been validated in cohorts of individuals that had been exposed or not exposed to radiation, resulting in guidelines that can help human experts to distinguish them from artifacts in alignments of trios. In the following, we will describe how the same data can be used to train the artificial intelligences and that they learn to pay attention to similar features as human experts.

Methods

DeepTrio by itself is an extension of the DeepVariant convolutional neural network architecture, aiming to improve variant calling accuracy by considering additional, parental information (11,12). Our first goal was to fine-tune the DeepTrio model to recognize DNMs in input data. Using DeepTrio as a base network allows us to draw on existing work, especially the efficient encoding of alignment data into tensors consumed by the network. Once satisfied with the accuracy of the model on DNMs, we retrained the resulting network to detect cDNMs. In accordance with the groundwork laid out in the DeepTrio algorithm, we create six-dimensional pile-up images (tensors) on the basis of alignment data for input into the convolutional neural network. The pile-up images span regions 110 bp up- and downstream from a putative variant site. Up to 100 reads are displayed for each trio member (300 total), in the order (from top to bottom) Parent 1, Child, Parent 2 and sorted by the reads' starting position. Each channel of the pile-up image encodes specific information for the BAM-file: read base, base- and mapping quality, read strand, read support for a variant and whether a base differs from the reference (Figure 1).

Network architecture. The standard output of DeepVariant is a probability mass function of the three possible genotypes (homozygous reference (HOM), heterozygous alternate (HET), homozygous alternate (ALT)), given input data generated from aligned reads (Figure 1). DeepTrio improves the genotyping quality by considering the parental data in addition to the child; however, it does not compute a likelihood for a heterozygous genotype being a *de novo* event. We added two more neurons to the output layer that acknowledge the existence of DNMs that can occur isolated or as a part of more complex patterns. Thus, the sum of the output of the first three neurons represents the probability of a genotype that is in agreement with a Mendelian pattern of inheritance, while the sum of neuron four and five represents the probability of any *de novo* event.

The fourth neuron was added to the standard Inception-V3 architecture of DeepTrio specifically to call DNMs. As a basis for the transfer learning, we used the network weights from the DeepTrio v1.3 Illumina WGS child model. To represent cDNMs in the output layer of the network, we followed the same method. We added a fifth output neuron that emits the likelihood of an event being a cDNM instead of an isolated *DNM*. We call the confidence emitted by the fourth output neuron the *de novo* score, and the confidence emitted by the fifth neuron is accordingly called cluster (cDNM) score.

The original DeepVariant and DeepTrio software is built based on Keras with Tensorflow as backend for the convolutional neural networks. We adopted the same toolchains for fine-tuning the Inception-V3 networks from DeepTrio. As DeepVariant constructs the output from three distinct steps ('make_examples', 'call_variants' and 'postprocess_variants'), there was no need to modify all steps to retrain the model for the desired effect. We retained the original 'make_examples' step that is used to generate the input tensors for the network. Using the built-in 'CustomizedClassesLabeller', we assigned the pileup images to one of the five classes, HOM (0), HET (1), ALT (2), DNM (3) or cDNM (4), for training and testing. Since the other two steps, 'call_variants' and 'postprocess_variants' are purpose-built for the standard variant calling methods, we discarded them and replaced them with our own implementations of the Inception-V3 training, variant calling and variant call file format (VCF) conversion code. In addition to all standard VCF fields, these scripts include the DNM as well as cDNM calls and their respective likelihood in the output files.

Training data and data augmentation

As a basic training data set, we used a set of whole genome sequencing trios. All trios were sequenced by the West German Genome Center (WGGC) on Illumina NovaSeq 6000 devices. In total, we used 92 trios from 62 distinct families for retraining the DeepTrio model and withheld 10 trios for later use as test data. Baseline small variant calls were computed using DRAGEN v3.6 against the GRCh37 reference genome. DRAGEN is an accelerated toolkit based on GATK best practice guidelines (15,16). To improve the accuracy of detection of more complicated variant classes like DNM and cDNM, we additionally called all samples using the base DeepVariant v1.3 Illumina WGS model. Quality of the raw data was checked extensively. Among other factors, we required minimum average genome coverage of 30X (mosdepth) and contamination of at most 3% as computed by VerifyBamID (17,18). Integrity of the families was checked prior to *de novo* calling using Peddy (19) and KING (20). We used the intersection of DRAGEN and DeepVariant callers on the whole genome sequencing data for training and testing. Variants were labeled as DNM if they lay in the overlap of both callsets and fulfilled a set of heuristic criteria that were derived from filter settings for small variants proposed by Adewoje *et al.* and Pedersen *et al.* (7,21). Among others, we required a depth of at least 10 reads in parents and 15 in the child and alternate allele frequency to be $0.35 \leqslant x \leqslant 0.65$ (Supplementary Table 1). Table 1 details the number of training examples that passed all of our quality criteria, which were less than 11% and 2% of all potential variants of each class for DNMs and cDNMs respectively. We used DNM candidates returned by only one of the two callers and reviewed them according to best practice guidelines for variant calling (22). If they fulfilled the criteria of an artifact, e.g. being in close proximity to INDEL variants, they were

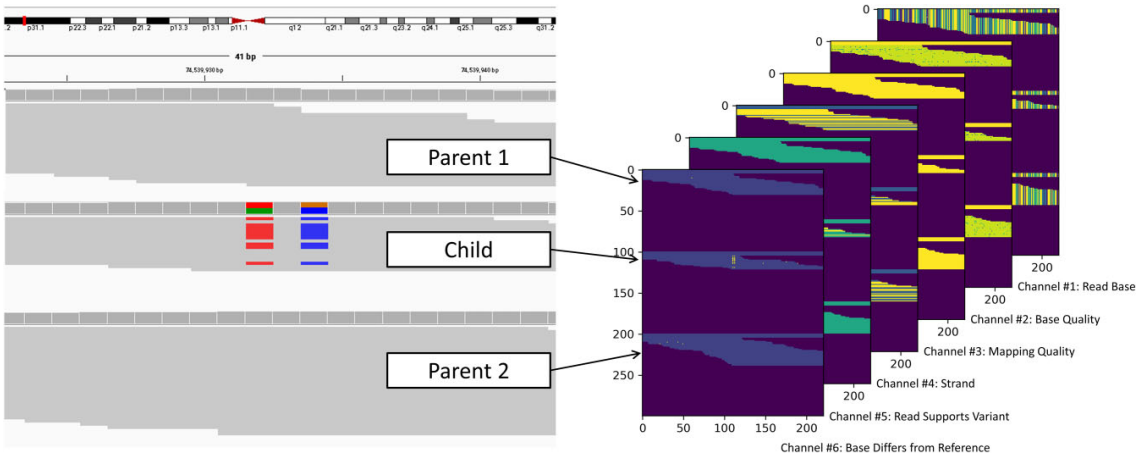


Figure 1. Clustered *de novo* mutations (cDNMs). Input data for DeepTrio for cDNM detection. A closeup of alignments in IGV around a clustered mutation is presented on the left. Two single nucleotide variants (SNVs) in close proximity to each other are only present in the child but not its parents. The right site features the generated Input data tensor of size 221 × 300 × 6 derived from read information at the given site. Input tensors encode information about read bases, base- and mapping quality scores, read strand and variant read support at each putative variant site discovered during the 'make_examples' step, which is the first step of variant calling.

Table 1. Training and Test Data for the DNM and cDNM model

		Training Data			Test data	
Model		DNM edge case model	DNM model	cDNM model	In-house Data	GIAB
Samples	Families	83	85	85	4	1
	Trios	113	117	117	10	1
Classes	HOM	501	6726	6726	50	N/A
	HET	10 000	11 184	11 184	1014	N/A
	ALT	4514	5412	5412	388	N/A
	DNM	4252	4410	4353	323	985
	cDNM Lesions	N/A	N/A	57	12	10
	HET (synthetic)	N/A	N/A	2802	N/A	N/A
	DNM (synthetic)	N/A	N/A	1300	N/A	N/A
	cDNM (synthetic)	N/A	N/A	5041	4891	N/A

This table details the number of mutations of each class that were present in the datasets used for fine tuning the original DeepTrio model and for testing the adjusted models. To retrain the model, first the standard variant classes together with a part of the detected DNMs were used to create the DNM edge case model. In preparation for the following training iterations, we first applied the augmentation methods to generate synthetic DNMs and identify edge cases to retrain the model on. This data was then used for the second training step for the DNM model and the final retraining step for the cDNM model. The number of examples for each variant class is listed in this table, for cDNMs it lists the total number of lesions that have to be identified as cDNM. Due to the fact that the original DeepVariant network was already trained on GIAB data, we refrained from including any examples from HG002 for the standard variant classes in our retraining effort. The overall number of calls on the GIAB data is detailed in Supplemental Table 1.

labeled as HET in the training set. Some artifacts that are representative for the classes that we excluded are presented in Supplemental Figure 5. This allows us to further increase the accuracy of the model for one of the common error cases of HaplotypeCaller-like variant callers, especially in genomic regions with low complexity (21). To balance the dataset, a number of loci called as HOM, HET or ALT by DRAGEN was randomly sampled to roughly match the number of variants labeled DNM. As non-de novo loci from this subset are sampled randomly, there is no guarantee that the training examples generated for the normal classes are particularly characteristic or hard to distinguish from true positive DNMs.

We generated training examples for cDNMs in the same manner, selecting clusters that are confirmed by both callers as positive examples for training. Note that this implies that the constituent DNMs are part of the overlap of both callers and would be selected as DNM for training, if not for the fact that they are within 20 bp of each other. Since these clusters are very rare events (less than one per trio on average), we decided to impute synthetic clusters into the alignments and variant calls. To make sure that these synthetic clusters are situated in genomic regions comparable to those of observed DNMs, each cluster was placed 150 to 200 bp up- or downstream from a true positive DNM. We made sure not to use any existing genomic variation, in particular DNM candidates, in the synthetic examples. The algorithm created for data augmentation was built to closely resemble characteristics observed in a large study, where cDNMs were extensively validated. We looked at 163 naturally occurring clusters with 411 lesions to define criteria for the generation of synthetic clusters. These criteria were used in the following algorithm to synthesize cDNMs: After selecting a location for the cluster that we wanted to create, we randomly decided on the number of lesions (two to four) and how far the lesions will be split apart (1–20 bp), with exponentially decreasing probabilities for more variants and larger distances. To create chal-

lenging examples for the network to train on (cDNM edge cases), we also generated clusters which are fully or partially inherited from one of the parents and thus do not fulfill the definition of a cDNM. This allows us to fine-tune the network on edge cases like DNMs within clusters of inherited variants. Further, we added clusters that were observed with heterozygous variants in the parents and homozygous in the children, since this pattern was found to be a common false positive call in pilot experiments resulting in a high cDNM score. Examples of the different classes of synthetic cDNMs are provided in Supplemental Figure 3. The synthetic data that was generated with the algorithm above was used for testing as well as training the model, with testing being performed on a set of specially generated clusters.

Training process. We retrained the basic DeepTrio Illumina WGS child model to recognize isolated and clustered DNMs in a three-step procedure (Figure 2). Initially, a *DNM edge case model* with four output neurons was trained on a set of labeled DNMs, as described in the prior section. After training, this model was used to identify difficult DNM candidates (DNM edge cases) by evaluating the model on four trios formerly withheld for validation. Then, a *DNM model*—also with four output neurons—learned to distinguish true positive DNMs from artifacts, with an explicit focus on the previously identified hard-to-distinguish edge cases. Afterwards, the DNM model was retrained as a *cDNM model* with an added fifth output neuron to additionally emit a confidence value for the occurrence of a cDNM. In total, there were 4410 DNMs and 57 cDNMs available for the network to learn the difference between inherited and *de novo* mutations (Table 1). By and large, the default hyperparameters from DeepTrio were used for the neural network models. Batch size was set to 64, the loss function to categorical cross-entropy and the initial learning rate to 0.0001 and then adapted using RMSprop and a momentum of 0.9, a decay of 0.9 and an epsilon value of 1. Additionally, the learning rate was decreased altogether by a factor of 0.995 every fifteen epochs. Approximately 80% of all triples were assigned to the training and 20% to the validation set in every run, and we made sure that triples from the same family are present in only one of the two sets. After each epoch, accuracy over all classes was calculated in the validation set and early stopping applied if there had not been an improvement over ten epochs in a row.

Based on calling with the DNM edge case model, we selected loci with a *de novo* score of $\geq$0.9 as DNM candidates and performed a visual check for systematic patterns of false positive *de novo* calls in the candidate set. In the absence of large, validated sets of DNMs, the visual investigation by an expert into the different variants remains the only feasible way of establishing a trusted test set. We identified false positives with a very low alternate allele frequency (AAF) and false positives in which other samples from the family harbored the same mutation with low allele frequency as DNM edge cases. In the first case, we are very likely dealing with sequencing artifacts or low frequency postzygotic mosaicism, while in the second case, the variant is most likely inherited from the parent with low read support for the variant or mixed mosaicism (23). To increase performance of the network on these specific classes of variants, we added examples of them to the training set for the training of the DNM model. In particular, we added false positive DNMs with an overall AAF of $<$0.3 as HOM calls and inherited variants (AAF $>$ 0.1 in at least one other sample from the family) as HET calls to the training set (Supplemental Figure 2).

The DNM model was then expanded as described above to introduce a fifth output neuron emitting the confidence that a given locus is part of a *de novo* mutation cluster. We retrained the model to detect cDNM using the same parameters as we did in previous training regimes using the augmented training data, now also accounting for cDNM with an additional label.

All iterations of the training loop were run on NVIDIA A100 (40GB, PCIE) GPUs. Each retraining task needed approximately 0.5 h of GPU time, for data drawn from a single whole genome sequencing trio. Variant calling using the model can be performed both on CPU and on NVIDIA GPUs. Performance is greatly improved when using GPUs for variant calling, requiring only 15 min on average on A100 GPUs, excluding the preparatory 'make_examples' step.

We evaluated the final models on ten trios withheld for testing and the Illumina WGS dataset of the Ashkenazim Trio published by the GIAB consortium (24). For the latter, we used the release version 4.2.1 of the data (25) and considered the curated set of single-nucleotide DNMs only from high-quality regions that was also used by Khazeeva *et al.* to benchmark DeNovoCNN. This provided a ground truth of 995 *de novo* mutations, including five *de novo* clusters of two lesions each (Supplemental Figure 4). To gauge the influence of the sequencing device on the quality and number of DNM calls observed, we made the data from six families available, that were sequenced in-house on both the NovaSeq and the HiSeq generation of Illumina sequencers. For both sequencers, we used the same ground-truth callset created based on the results of the NovaSeq sequencer.

Results

After training, the resulting models gained the capability to call DNMs and cDNMs in WGS data as separate classes compared to the original DeepTrio model. All trios that were marked for testing were called at whole genome level to compare the basic variant calling, DNM calling and cDNM calling accuracy. We estimate sensitivity as the portion of variants from the overlap that were called by the model. To establish the positive predictive value of the network for the complex variant classes DNM and cDNM, we visually validated the hits in the Integrative Genomics Viewer (IGV) and categorized them as true positive, false positive or ambiguous; that is, no unanimous vote by human experts, examples of these mutations are provided in Supplementary Figure 6.

Since the models emit confidence likelihood values for DNM and cDNM, we can tune the trade-off between the sensitivity and specificity of the detection of the variants by selecting a threshold for calling a variant a *de novo* or cluster event (Figure 3). As expected, by increasing this cutoff, the specificity of the calls increases at the expense of some sensitivity. For the detection of DNM, we experimentally established 0.985 as a suitable cutoff value (Figure 3A). For cDNM, the threshold was chosen in the same manner. Interestingly, the interplay of DNM and cDNM confidence values allows us to increase the sensitivity and specificity of the cluster detection (Figure 3B). We observe that *de novo* mutation clusters are highly likely to be true positive if the confidence values for *de novo* and cluster detection are close to the diagonal $f(x) = 1 - x, 0 \Leftarrow x \Leftarrow 1$.

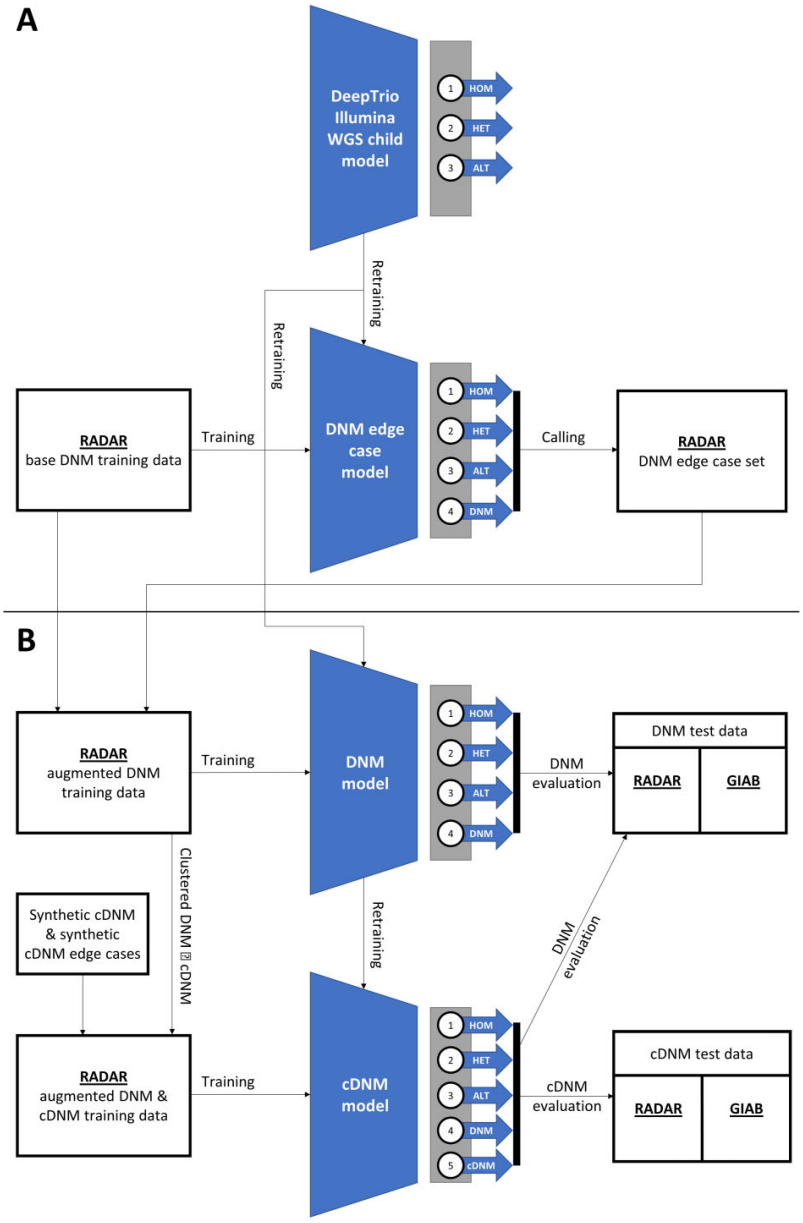


Figure 2. Architecture of DeepTrio. (**A**) identification of edge cases, (**B**) fine-tuning for cDNMs. (A) Overview of the training scheme employed to train the final DNM and cDNM detection models. Broadly, the two final models are created in a two-round process where the first round serves to generate challenging training examples for the subsequent rounds. The first model ('DNM edge case model') is trained based on the raw, partially validated DNM data. (B) After calling variants with this retrained model, we select challenging edge cases to include in the augmented DNM training set. Together with the base DNM training examples, we use these examples to retrain the DeepTrio Illumina WGS child model to the final DNM model ('DNM model'). This model is subsequently retrained again to include the cDNM output neuron. Here we use the cDNM training examples together with the synthetically generated ones to train the model. Note that the label for some examples might change from DNM to cDNM between the two training sets. Then, both networks are evaluated on the data from Illumina NovaSeq in-house sequencing and the GIAB HG002 Ashkenazim trio.

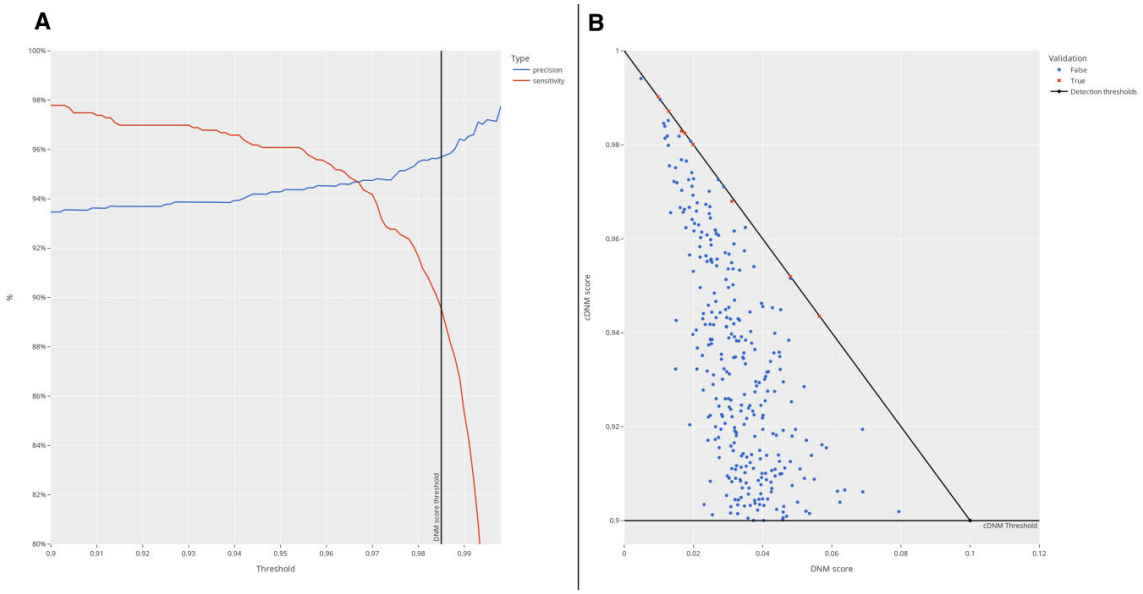


Figure 3. Accuracy of DNM calling. (**A**) Precision of DNM calling. (**B**) Precision of cDNM calling. (A) Performance of the retrained model for the DNM detection task on the HG002 GIAB trio. Precision and sensitivity are shown, depending on the DNM score cutoff chosen. Precision can be increased at the expense of sensitivity and the trade-off depends on the use case. We chose a cutoff value of 0.985, while a cutoff of 0.9 improves the sensitivity of detection significantly. At $p_{DNM} = 0.985$, the model achieves a sensitivity of 89.55% and precision of 95.70%. At $p_{DNM} = 0.9$ the precision drops to 93.47% but sensitivity increases to 97.79%. (B) Detection accuracy for cDNMs on the HG002 GIAB trio. With a simple cutoff of $p_{cDNM} = 0.9$ (horizontal line), the model achieves a sensitivity of 100.0% and precision of only 0.32%. Detection accuracy is significantly improved by the insight that true positive clusters are both *de novo* and clustered, so the sum p_{DNM} and p_{cDNM} should be close or equal to 1. With a cutoff of $1 - (p_{DNM} + p_{cDNM}) \Leftarrow 0.999$, the model detects all five cDNMs with a precision of 76.9%. These detection thresholds can be tuned similarly to the DNM score thresholds to achieve the desired accuracy both in terms of precision and sensitivity.

Overall, we achieve a precision of 95.7% and a sensitivity of 89.6% for DNM in the GIAB trio using the threshold of 0.985. At a cutoff of 0.9, we achieve a precision of 94.3% and sensitivity of 97.8% for the detection of DNMs (Table 2a). Inclusion of cDNMs leads to a slight drop in sensitivity and precision of DNM calls to 96.4% precision and 77.8% sensitivity ($t = 0.985$) and 90.5% precision and 91.96% sensitivity ($t = 0.9$), respectively. Choosing cDNMs along the linear boundary $f(x) = 1 - x$, $0.999 \Leftarrow x \Leftarrow 1$, we achieve a precision of 76.9% and detect five out of five variants. On unseen data from a different sequencing device and with different read length (100 bp; HiSeq X), the model achieves precision and sensitivity of 77.2% and 58.2% for DNMs (Table 2b). The model detected all six true positive cDNMs in test data from both sequencers individually, demonstrating an ability to generalize the learned features to recognize these variants. Testing on the augmented data revealed that the model is able to distinguish synthetic cDNMs from DNMs with high accuracy, featuring a lower sensitivity of 36.5% (HiSeq) and 61.7% (NovaSeq) compared to real data but a higher precision (HiSeq: 99.3%, NovaSeq: 98%, Table 2b) on a total of 4891 test cases.

To compare our method with existing, similar algorithms, we chose DeNovoCNN. We called DNMs on GIAB HG002 and six evaluation trios sequenced on both, the NovaSeq and HiSeq sequencers. The comparison of the two methods for the DNM detection task is detailed in Table 2. When searching for cDNM calls in the DeNovoCNN output by subsetting the *de novo* mutations to clusters of variants within 20 bp, we found that DeNovoCNN manages to detect all five clusters in GIAB data with a precision of 50%. On our in-house data it only finds three cDNMs out of the total of 6 cDNMs with a precision of 7% and 11% for HiSeq and NovaSeq data respectively.

Table 2a. Performance comparison of DeepTrio and DeNovoCNN for DNMs

HG002	TP	FP	FN	Precision	Sensitivity
Shared by both callers	946	47	49	N/A	N/A
DeepTrio	973	68	22	0.935	0.978
DeNovoCNN	965	182	30	0.841	0.970

Results of the comparison between DeNovoCNN and the modified DeepTrio network on GIAB HG002 data. We used DeNovoCNN v1 from the github page of the authors and compared it to the second round trained DNM detection model. We used 0.5 as the threshold for DeNovoCNN calling and left all other settings at their default values, as suggested by the authors in their performance comparison. We used 0.985 as the threshold for the DNM detection in our network. Both callers share the large majority of calls, only differing at 201 sites. As a validation set, we used the curated set from the DeNovoCNN authors, to allow the fairest comparison. In total, this set contains 995 *de-novo* mutation sites that can be used to assess the variant call quality. In total, 1188 sites were analyzed, including some sites where negative results were expected (e.g. both networks should not detect a DNM). Nevertheless, we refrain from assessing True Negatives (TN) here, as there is no clear definition of these cases, aside from all mutations that are not de novo, whose large number would make the interpretation of the statistical analysis impossible. DeepTrio also showed good results on a manually validated set of DNMs on in-house data, with comparable precision but lower sensitivity compared to the GIAB training set.

During the retraining, we made sure that the model retains its ability to accurately detect inherited germline variants in the input data. The model achieves an F1-score of 90.91% for the detection of small variants on GIAB HG002

Table 2b. Performance comparison of DeNovoCNN and DeepTrio on in-house sequencing data from different sequencers and simulated cDNM data. In the table below

Model	Sequencer	No. of trios	Sensitivity [95% CI]	Precision [95% CI]
cDNM	HiSeq	6	77.2% [71.7–82.6%]	58.2% [47.5–69.0%]
	NovaSeq	7	78.6% [70.7–86.5%]	58.7% [50.8–66.5%]
cDNM (synthetic)	HiSeq	6	36.5% [29.6–43.4%]	99.3% [97.4–100%]
	NovaSeq	7	61.7% [55.8–67.6%]	98.0% [95.7–100%]
DNM	HiSeq	6	66.8% [57.6–76.1%]	64.7% [58.1–71.2%]
	NovaSeq	7	89.4% [86.1–92.6%]	59.3% [52.3–66.3%]
DeNovoCNN	HiSeq	6	86.6% [71.0–100%]	53.2% [46.4–60.2%]
	NovaSeq	7	91.5% [85.2–95.6%]	48.7% [37.8–52.2%]

We show the sensitivity and precision of three models for the DNM detection task on the same families, both sequenced on the NovaSeq and HiSeq devices. For DeepTrio, we also analyzed the Sensitivity and Precision of calls made on simulated cDNM sites to assert that there is no overfitting against these data and to show an analysis that is more robust against outlier samples. All three models were tested and showed good agreement in their calls between different sequencing devices. On most trios, the modified DeepTrio model had considerably higher precision than DeNovoCNN, even though we change runtime parameters of DeNovoCNN to a threshold of 0.9 instead of the suggested 0.5 to reduce the amount of false positive calls. Supplemental Table 3 gives the detailed account of true and false positive DNM call counts per trio.

(Supplemental Table 2). This shows how we could generalize the knowledge of the model to other classes of variants (DNM and cDNM) without a prominent loss in the original functionality.

Discussion

The final, retrained models achieve good accuracy on the benchmark test sets. The models achieve competitive or superior performance to other, similar methods on the GIAB HG002 genome trio as well as Illumina HiSeq and NovaSeq data. Classes of variants that are very artifact-rich (e.g. cDNMs) can be detected with high sensitivity and precision. Such variant classes are especially difficult to call using traditional approaches, since these are inherently edge-cases with current sequencing technology, where even true positive clusters are hard to distinguish from artifacts. Furthermore, we have demonstrated a generic approach to retrain the DeepVariant or DeepTrio for the accurate detection of arbitrary complex variant classes. The code used for retraining the model and variant calling utilizing this model is available on github. To examine the model, we used attention maps (26–28). They can serve as a good reference to regions of an input tensor that are particularly informative for the algorithm's decision (29). Given that the input tensors for the model are very similar to visualizations that geneticists are used to from software like IGV or the UCSC Genome Browser, the attention maps can be interpreted very intuitively, in particular since the visual inspection of alignments is still the gold standard to verify germline mutations in many centers.

As humans have grown very experienced in evaluating short read data in pileup views over the years, it was also interesting to visualize the features and regions of the pileup examples that were important for the decision-making of the network. We generated SmoothGrad images for the classes present in the training data (Supplemental Figure 1). The activation maps computed by the SmoothGrad algorithm highlight areas of the input tensor, but collapses all 6 input dimensions into one to enable the output as 2D heatmap. For most cases, they confirm the intuition learned by human experts when considering the pileup views in short read sequencing data, but due to the nature of the SmoothGrad images it is impossible to discern which input dimension most significantly affects the variant call by the model. It is clear though, that the model effectively builds its decision based on the sequencing data from all three members of the trio, and is able to incorporate up to 110 bp up- and downstream from a putative variant into its decision, which allows for greater accuracy than many tools that only consider the data at a specific variant sites or only data present in a VCF file.

Underlining the hypothesis that the network recognizes DNM effectively is the fact that the detection of cDNM relies not only on the probability output for neuron five (cluster), but also the *de novo* probability. The network apparently recognized multiple facts from the definition of a cDNM and knows how to combine them. Namely, the network shows its understanding that the constituent mutations of a cluster are themselves DNMs. Based on the results for the detection of cDNMs, we conclude that the network is certain that each part of the cluster is *de novo* when a candidate cluster is encountered and scores high. We use this explanation of the model behavior to then improve the detection accuracy for clusters and allow for a further fine-tuning of the sensitivity and precision of the detection.

While the detection accuracy for DNMs and cDNMs can be increased through the training process, the accuracy for the standard variant classes HOM, HET and ALT decreases. This is likely due to the introduction of one or two further output neurons that also represent heterozygous mutations, which are labeled as DNMs or cDNMs, respectively. Since our training scheme relies on the exclusive labeling of each mutation with exactly one output mutation type, the retraining can lead to some slight confusion between the variant classes for the network. We tested converting the scheme to a multi-label classification problem, i.e. DNM examples were tagged with the two labels HET and DNM, and cDNM examples with the three labels HET, DNM and cDNM. This modification was discarded, though, because it led the model to call only the most frequent class (HET) instead of HET plus the *de novo* classes in almost all *de novo* examples. The performance of this alternative approach could also not be improved when the class imbalance was counteracted by reweighting the DNM class tenfold and the cDNM class ten- or hundredfold compared to the HET class. Another thing to bear in mind is that the assumption of all DNMs being heterozygous is not true in general. However, the case that a mutation occurs *de novo* in a homozygous state was deemed extremely unlikely and was therefore excluded from the retraining process. Overall, the accuracy of the general variant detection task for the modified network is still good, and can likely be improved with further filtering or GVCF joint genotyping steps. Additionally, since

the input is identical between the modified and unmodified DeepTrio networks, one can run multiple variant calling steps on the same examples. Since 'make_examples' constitutes a considerable share of the runtime of DeepTrio/DeepVariant, this combines advantages from both networks: Accurate detection of germline variants using DeepTrio and the detection of advanced mutational signatures using the modified network.

Our approach to the detection of *de novo* mutation clusters can also be applied to retrain the network for arbitrary mutational signatures, as long as they fall in a 221 bp window of the genome. Together with the algorithm for data augmentation proposed, it is feasible to retrain the network for the accurate detection of mutational signatures, even when no large validated set of ground truth data is available. Our results show that the generation of synthetic cDNM clusters helped the model to learn the basic structure of cDNMs, while not overfitting on these simplistic approximations of real data. Nevertheless, large sets of validated ground truth data (e.g. DNMs), help the network to learn to recognize these mutations more accurately than what would be feasible exclusively with augmented examples. The decrease in accuracy on data generated on the NovaSeq sequencer suggests that there are sequencing artifacts due to the two-color chemistry that have not been adequately learned as such with our training scheme (4). Given more training data containing labeled examples for false positives would likely yield a model that can distinguish even the NovaSeq-specific artifacts. In absence of more labeled training data, promising approaches to further increase the detection accuracy of complex mutational patterns is adding more information from reads to the input tensors (e.g. read orientation). Additionally, we showed that the network learned an understanding of the connection between the different variant types HET, DNM and cDNM. This information can be used to recognize complex mutational signatures derived from multiple events efficiently and accurately.

Data availability

All code and models are available at (DOI: 10.5281/zenodo.8079352) or https://github.com/jguski/deeptrio_dnm. No new data were generated or analysed in support of this study. The data used by this study can be accessed at EGA under Study Number EGAS00001007321.

Supplementary data

Supplementary Data are available at NARGAB Online.

Funding

No external funding.

Conflict of interest statement

None declared.

References

1. Moore,L., Cagan,A., Coorens,T.H.H., Neville,M.D.C., Sanghvi,R., Sanders,M.A., Oliver,T.R.W., Leongamornlert,D., Ellis,P., Noorani,A., *et al.* (2021) The mutational landscape of human somatic and germline cells. *Nature*, **597**, 381–386.
2. Jónsson,H., Sulem,P., Kehr,B., Kristmundsdottir,S., Zink,F., Hjartarson,E., Hardarson,M.T., Hjorleifsson,K.E., Eggertsson,H.P., Gudjonsson,S.A., *et al.* (2017) Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature*, **549**, 519–522.
3. Alexandrov,L.B., Nik-Zainal,S., Wedge,D.C., Aparicio,S.A.J.R., Behjati,S., Biankin,A.V., Bignell,G.R., Bolli,N., Borg,A., Børresen-Dale,A.-L., *et al.* (2013) Signatures of mutational processes in human cancer. *Nature*, **500**, 415.
4. Arora,K., Shah,M., Johnson,M., Sanghvi,R., Shelton,J., Nagulapalli,K., Oschwald,D.M., Zody,M.C., Germer,S., Jobanputra,V., *et al.* (2019) Deep whole-genome sequencing of 3 cancer cell lines on 2 sequencing platforms. *Sci. Rep.*, **9**, 19123.
5. Kaplanis,J., Ide,B., Sanghvi,R., Neville,M., Danecek,P., Coorens,T., Prigmore,E., Short,P., Gallone,G., McRae,J., *et al.* (2022) Genetic and chemotherapeutic influences on germline hypermutation. *Nature*, **605**, 503–508.
6. Sasani,T.A., Pedersen,B.S., Gao,Z., Baird,L., Przeworski,M., Jorde,L.B. and Quinlan,A.R. (2019) Large, three-generation human families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *eLife*, **8**, e46922.
7. Adewoye,A.B., Lindsay,S.J., Dubrova,Y.E. and Hurles,M.E. (2015) The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline. *Nat. Commun.*, **6**, 6684.
8. Zlobina,A., Farkhutdinov,I., Carvalho,F.P., Wang,N., Korotchenko,T., Baranovskaya,N. and Farkhutdinov,A. (2022) Impact of environmental radiation on the incidence of cancer and birth defects in regions with high natural radioactivity. *Int. J. Environ. Res. Public Health*, **19**, 8643.
9. Holtgrewe,M., Knaus,A., Hildebrand,G., Pantel,J.-T., Santos,M.R., de,L., Neveling,K., Goldmann,J., Schubach,M., Jäger,M., *et al.* (2018) Multisite de novo mutations in human offspring after paternal exposure to ionizing radiation. *Sci. Rep.*, **8**, 14611.
10. Brand,F., Klinkhammer,H., Knaus,A., Holtgrewe,M., Weinhold,L., Beule,D., Ludwig,K., Kothiyal,P., Maxwell,G., Noethen,M., *et al.* (2023) A transgenerational mutational signature from ionizing radiation exposure. medRxiv doi: https://doi.org/10.1101/2023.11.20.23298689, 20 November 2023, preprint: not peer reviewed.
11. Poplin,R., Chang,P.-C., Alexander,D., Schwartz,S., Colthurst,T., Ku,A., Newburger,D., Dijamco,J., Nguyen,N., Afshar,P.T., *et al.* (2018) A universal SNP and small-indel variant caller using deep neural networks. *Nat. Biotechnol.*, **36**, 983–987.
12. Kolesnikov,A., Goel,S., Nattestad,M., Yun,T., Baid,G., Yang,H., McLean,C., Chang,P.-C. and Carroll,A. (2021) DeepTrio: variant calling in families using deep learning. bioRxiv doi: https://doi.org/10.1101/2021.04.05.438434, 06 April 2021, preprint: not peer reviewed.
13. Sudmant,P.H., Rausch,T., Gardner,E.J., Handsaker,R.E., Abyzov,A., Huddleston,J., Zhang,Y., Ye,K., Jun,G., Fritz,M.H.-Y., *et al.* (2015) An integrated map of structural variation in 2,504 human genomes. *Nature*, **526**, 75–81.
14. Sage,E. and Shikazono,N. (2017) Radiation-induced clustered DNA lesions: repair and mutagenesis. *FreeRadic. Biol. Med.*, **107**, 125–135.
15. DePristo,M.A., Banks,E., Poplin,R., Garimella,K.V., Maguire,J.R., Hartl,C., Philippakis,A.A., del Angel,G., Rivas,M.A., Hanna,M., *et al.* (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. *Nat. Genet.*, **43**, 491–498.
16. Vasimuddin,M., Misra,S., Li,H. and Aluru,S. (2019) Efficient architecture-aware acceleration of BWA-MEM for multicore systems. In: *2019 IEEE International Parallel and Distributed Processing Symposium (IPDPS)*. pp. 314–324.
17. Pedersen,B.S. and Quinlan,A.R. (2018) Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics*, **34**, 867–868.
18. Jun,G., Flickinger,M., Hetrick,K.N., Romm,J.M., Doheny,K.F., Abecasis,G.R., Boehnke,M. and Kang,H.M. (2012) Detecting and

estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am.J. Hum. Genet.*, **91**, 839–848.
19. Pedersen,B.S. and Quinlan,A.R. (2017) Who's who? Detecting and resolving sample anomalies in human DNA sequencing studies with peddy. *Am. J. Hum. Genet.*, **100**, 406–413.
20. Manichaikul,A., Mychaleckyj,J.C., Rich,S.S., Daly,K., Sale,M. and Chen,W.-M. (2010) Robust relationship inference in genome-wide association studies. *Bioinformatics*, **26**, 2867–2873.
21. Pedersen,B.S., Brown,J.M., Dashnow,H., Wallace,A.D., Velinder,M., Tristani-Firouzi,M., Schiffman,J.D., Tvrdik,T., Mao,R., Best,D.H., *et al.* (2021) Effective variant filtering and expected candidate variant yield in studies of rare human disease. *NPJ Genom. Med.*, **6**, 60.
22. Koboldt,D.C. (2020) Best practices for variant calling in clinical sequencing. *Genome Med*, **12**, 91.
23. Bernkopf,M., Abdullah,U.B., Bush,S.J., Wood,K.A., Ghaffari,S., Giannoulatou,E., Koelling,N., Maher,G.J., Thibaut,L.M., Williams,J., *et al.* (2023) Personalized recurrence risk assessment following the birth of a child with a pathogenic de novo mutation. *Nat. Commun.*, **14**, 853.
24. Zook,J.M., Catoe,D., McDaniel,J., Vang,L., Spies,N., Sidow,A., Weng,Z., Liu,Y., Mason,C.E., Alexander,N., *et al.* (2016) Extensive sequencing of seven human genomes to characterize benchmark reference materials. *Sci. Data*, **3**, 160025.
25. Wagner,J., Olson,N.D., Harris,L., Khan,Z., Farek,J., Mahmoud,M., Stankovic,A., Kovacevic,V., Yoo,B., Miller,N., *et al.* (2022) Benchmarking challenging small variants with linked and long reads. *Cell Genom.*, **2**, 100128.
26. Smilkov,D., Thorat,N., Kim,B., Viégas,F. and Wattenberg,M. (2017) SmoothGrad: removing noise by adding noise. arXiv doi: https://arxiv.org/abs/1706.0382512 June 2017, preprint: not peer reviewed.
27. Chattopadhay,A., Sarkar,A., Howlader,P. and Balasubramanian,V.N. (2018) Grad-CAM++: Generalized Gradient-Based Visual Explanations for Deep Convolutional Networks. In: *2018 IEEE Winter Conference on Applications of Computer Vision (WACV)*. Lake TahoeN, V, USA, pp. 839–847.
28. Selvaraju,R.R., Cogswell,M., Das,A., Vedantam,R., Parikh,D. and Batra,D. (2016) Grad-CAM: visual explanations from deep networks via Gradient-based localization. arXiv doi: https://arxiv.org/abs/1610.02391, 07 October 2016, preprint: not peer reviewed.
29. Norgeot,B., Quer,G., Beaulieu-Jones,B.K., Torkamani,A., Dias,R., Gianfrancesco,M., Arnaout,R., Kohane,I.S., Saria,S., Topol,E., *et al.* (2020) Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nat. Med.*, **26**, 1320–1324.

Received: July 3, 2023. **Revised:** January 16, 2024. **Editorial Decision:** January 22, 2024. **Accepted:** January 23, 2024

3.3 A transgenerational mutational signature from ionizing radiation exposure.

Brand, Klinkhammer, Knaus, Holtgrewe, Weinhold, Beule, Ludwig, Kothiyal, Maxwell, Noethen, Schmid, Sperling, and Krawitz, “A Transgenerational Mutational Signature from Ionizing Radiation Exposure”

Year: 2023
Journal: Scientific Reports (*In Review*), medRxiv
DOI: https://doi.org/10.1101/2023.11.20.23298689

A transgenerational mutational signature from ionizing radiation exposure.

Fabian Brand[1], Hannah Klinkhammer[1,3], Alexej Knaus[1], Manuel Holtgrewe[2], Leonie Weinhold[3], Dieter Beule[2], Kerstin Ludwig[4], Prachi Kothiyal[5], George Maxwell[5], Markus Noethen[4], Matthias Schmid[3], Karl Sperling[6], Peter Krawitz[1]

[1] Institute of Genomics Statistics and Bioinformatics, School of Medicine, University Hospital Bonn & University of Bonn, Germany
[2] Core Unit Bioinformatics, Berlin Institute of Health, Berlin, Germany
[3] Institute of Medical Biometry, Informatics and Epidemiology, University Hopsital Bonn, Bonn, Germany
[4] Institute of Human Genetics, School of Medicine, University Hospital Bonn & University of Bonn, Germany
[5] Inova Translational Medicine Institute, Inova Health System, Falls Church, VA, USA
[6] Institute of Medical and Human Genetics, Charitè-Universitaetsmedizin Berlin, Germany

Abstract

The existence of transgenerational effects of accidental radiation exposure on the human germline remains controversial. Evidence for transgenerational biomarkers are of particular interest for populations, who have been exposed to higher than average levels of ionizing radiation (IR). This study investigated signatures of parental exposure to IR in offspring of former German radar operators and Chernobyl cleanup workers, focusing on clustered de novo mutations (cDNMs), defined as multiple de novo mutations (DNMs) within 20 bp. We recruited 110 offspring of former German radar operators, who were likely to have been exposed to IR (Radar cohort, exposure = 0-353 mGy), and reanalyzed sequencing data of 130 offspring of Chernobyl cleanup workers (CRU, exposure = 0-4,080 mGy) from Yeager, et al. In addition, we analyzed whole genome trio data of 1,275 offspring from unexposed families (Inova cohort). We observed on average 2.65 cDNMs (0.61 adjusted for the positive predictive value (PPV)) per offspring in the CRU cohort, 1.48 (0.34 PPV) in the Radar cohort and 0.88 (0.20 PPV) in the Inova cohort. This represented a significant increase ($p < 0.005$) of cDNMs counts, that scaled with paternal exposure to IR ($p < 0.001$). Our findings corroborate that cDNMs represent a transgenerational biomarker of paternal IR exposure.

Introduction

Transgenerational effects of ionizing radiation (IR) in the offspring of former radar operators are a stated concern of several European armed services [1]. Investigation of such effects is warranted in order to design effective preventive measures, and optimize health monitoring of military personnel and their offspring.

The potential of transmission of radiation-induced genetic alterations to the next generation is of particular concern for parents who may have been exposed to higher doses of IR and potentially for longer periods of time than considered safe. To address the concerns articulated

by former radar soldiers, the German Ministry of Defense initiated this study with the goal of investigating the transgenerational effects of IR in offspring of exposed parents. Some tasks, primarily mechanical calibration work on radar units during live operations, have been recognized as an occupational health hazard by the German government [1].

Previous studies into the transgenerational effects of IR on human DNA have investigated populations affected by nuclear weaponry, and the Chernobyl nuclear reactor incident of 1986 [1–8]. Recent studies in a small cohort of British nuclear test veterans found no enrichment in chromosomal aberrations, de novo mutations, or structural variation in the offspring of exposed soldiers [2,3]. However, a transgenerational effect could not be excluded, due the limited cohort size and the low and uncertain estimates of IR exposure [2,3]. Two studies of the impact of the Chernobyl incident found a twofold increase in mutation rates at minisatellites, microsatellites, and tandem repeat loci in the offspring of former clean-up workers [5,9]. However, the small cohort sizes, potential confounders, and the limited number of loci that were examined have led to ongoing debates concerning the statistical significance of these findings [10–12]. Yeager, et al. recruited a cohort of 105 individuals who had been exposed to IR following the 1986 Chernobyl nuclear accident and their 130 offspring (born 1987-2002). The parents had either been inhabitants of the town of Pripyat at the time of the accident, or had been employed as liquidators responsible for guarding or cleaning up the accident site [8]. No elevated mutation rates for isolated DNMs were detected [8,13]. Furthermore, the authors found no enrichment of C>T mutations within 47 kb intervals, which has been hypothesized as an indicator of hyper-mutability of single-strand intermediates during the repair of double-strand breaks, but no smaller mutation clusters were analyzed [14,15]. Therefore, to date, no definite transgenerational mutational signature of IR exposure has yet been identified.

Identification of a potential transgenerational mutational signature for IR exposure requires a detailed understanding of the impact of IR on human DNA in the germline. When IR interacts with DNA the energy transfer may directly cause. a variety of DNA lesions, such as strand breaks , oxidized bases, loss of bases [16–18]. However, the primary pathway for IR-induced DNA damage is often indirect; IR generates reactive oxygen species (ROS) through the ionization of nearby water molecules in the cell [16–18]. These ROS then induce a variety of DNA lesions, including oxidized bases, base losses, double strand breaks (DSBs), single strand breaks (SSBs), with DSBs being the most detrimental to DNA structure. The repair of DSBs involves two main mechanisms: 1) Homologous Recombination Repair (HRR), a process that involves a homologous template; and 2) Non-Homologous End Joining (NHEJ), a process in which broken DNA ends are ligated without a template [19]. In germline cells, particularly during spermatogenesis, HRR plays a critical role in maintaining genomic integrity, whereas NHEJ, despite being more common, is more likely to introduce errors [20]. Due to its error-prone nature, NHEJ in germline cells can result in complex, ROS-induced lesions, which are turned into mutations, within short genomic regions [17,21]. Consequently, these lesions may contribute to genomic instability and cause cell death or persist through cell division and, particularly in germline cells, be passed on to future generations, representing a potential signature of IR parental exposure [21,22]. Importantly, DNA repair is less efficient in spermatids and mature

spermatozoa which, consequently, show the highest radiosensitivity of all stages of spermatogenesis [23,24].

Recent research in mice has provided compelling evidence that clustered de novo mutations (cDNMs) within short DNA segments (<20 bp) can increase following paternal exposure to IR, with the magnitude of this effect being dose-dependent, particularly in hematopoietic stem cells [25–27]. To investigate this finding in humans, in 2018, our group performed a small WGS pilot study of 18 offspring of former radar operators from the German military [7]. The analyses identified an increased mean number of cDNMs (then called multi-site de novo mutations (MSDNs)), and cases with exceptionally high cDNM rates. The analyses also identified two translocations, which had resulted from neighboring mutations. The results of this pilot study suggest that cDNMs may represent a signature of IR-induced DNA damage in humans.

Accurate identification of DNMs and cDNMs in current generation WGS data has to account for natural and technical biases. Parental age at conception is a significant and known confounder for the number of *de novo* mutations in their offspring [14,28,29]. Paternal age is associated with an increase of isolated DNMs, averaging at roughly 1-2 DNMs per year of paternal age at conception of the child. In addition, the maternal age has primarily been implicated with rarer, more clustered mutations, showing a significant enrichment in 10k bp wide clusters [14,30]. WGS data enables the analysis of DNMs and cDNMs all over the human genome, but due to lower specificity in DNM calling many regions with repeats or regions with low genomic complexity have been excluded from earlier studies [7,25,28]. It is particularly hard to distinguish DNMs from sequencing errors in these regions, and the difficulty to capture these regions using PCR primers makes validation of potential DNM and cDNM calls challenging. The DNA source material and sequencing device that generates the data also have an influence on the quality and accuracy of DNM calls [31].

The aim of the present study was to determine whether the signature of IR-induced clustered DNA lesions was detectable in the offspring of fathers with a history of probable IR exposure. The analyses were conducted using a newly recruited cohort of former German-military radar operators, their wives, and offspring (Radar cohort), as well as WGS data accessed from Yeager et al. 2021 (CRU cohort), and from a previously reported cohort of individuals with no history of exposure to IR (Inova cohort). Herein, we describe how we tested whether cDNMs are a transgenerational biomarker of prolonged paternal exposure to IR. We sequenced data for the Radar cohort and performed a new joint variant calling analysis together with the CRU and Inova data in order to confirm the known paternal age effect in all three cohorts. Afterwards, we used negative binomial regression models to ascertain differences in the number of cDNMs per sample in each cohort and to associate them with the likely exposure of their fathers.

Materials and Methods

The Radar cohort

The study at hand was commissioned by the *"Bundesamt für Ausrüstung, Informationstechnik und Nutzung der Bundeswehr"* (Federal Office of Bundeswehr Equipment, Information Technology and In-Service Support, BAAINBw) to further improve the compensation for former radar personnel of both German armies (BT Drs. 18/9032). Recruitment for the present Radar cohort was conducted between 2019 and 2021. Former radar soldiers were approached by the study team via advertisements in relevant magazines and online forums, and through the "Bund zur Unterstützung Radargeschädigter e.V.", a support group for potentially IR-exposed former German radar operators. The primary inclusion criterion for the present study was a history of exposure to high dose IR when servicing radar installations during live operations, as judged by an independent expert on the basis of a self-report questionnaire sent to each potential participant. In total, 80 former radar operators from the West or East German armies (Bundeswehr and Nationale Volksarmee, NVA) were included in the present study, together with their wives (n = 80) and offspring (n = 110) (Supplemental Table S1) [1,7]. Even though the German government spent significant efforts in investigating health effects and occupational risks following long time service with radar units, and despite the fact that soldiers serving at unprotected radar installations have a higher risk to develop certain cancers, reliable data on the damage caused by the stray radiation from these devices is very limited [1,6].

Ethics Statement

Ethical approval for the present study was obtained from the ethics committee of the Medical Faculty of the University of Bonn (Ethikkommission der Medizinischen Fakultät Bonn). All participants from the Radar cohort, i.e. the former soldiers, their wives, and their offspring, provided written informed consent for the present analyses prior to inclusion. All participants of the CRU and Inova cohorts had provided written informed consent for the use of their data by other research groups within the context of the respective original investigation. The present analyses were performed within the guidelines specified in the informed consent documentation for all three cohorts, and within the limits of the approval granted by the ethics committee of the Medical Faculty of Bonn. All study procedures were conducted in accordance with the principles of the Declaration of Helsinki.

Estimation of IR dose in the Radar cohort

Retrospective estimations of IR dose for the Radar cohort were made at the Radiation Measurement Facility (Strahlenmessstelle) of the German Federal Armed Forces (Bundeswehr) [32]. For this purpose, the military service record of each Radar participant was accessed. In estimating IR dose, two factors were considered. The first factor was the role and duties of the given Radar cohort participant, and the period of time for which they had served in a radar unit of the West or East German army (Supplemental Material 1.2). The second factor was the radar device that had been in active service at the time of the participant's military

service. Dose estimations were based on: 1) historical measurements, which had been taken from common radar devices by the military at the time the device had been in active service; or 2) measurements of the emissions of out-of-service radar devices that were reconstructed for the purpose of retrospective assessment. For each radar device, potential sources of stray radiation were determined in order to establish a realistic base rate of IR emission during service. Even though the retrospective dose assessment was carried out with great care, it is likely that there are errors introduced by both aforementioned factors that formed the basis for the dose estimation. Section 1.2 of the Supplemental Materials gives a detailed account of the procedures for retrospective dose estimation.

Whole Genome Sequencing Data

For all participants of the Radar cohort, sequencing was performed at the NGS core facility of the West German Genome Center (WGGC) in Bonn to a minimum whole genome coverage of 30X. WGS was performed according to the standard protocols on an Illumina NovaSeq device (Supplemental Material 1.3). To ensure that data generated on HiSeq devices (i.e. Inova and a subset of CRU) and sequences that were generated by the newer generation NovaSeq devices (i.e. Radar and the remainder of CRU) were comparable, and that the different read lengths did not induce confounding errors, three families from the Radar cohort were sequenced on both devices.

WGS data from the cohort of Yeager, et al. were accessed under dbGAP accession number phs001163.v1.p1, and all parent-offspring trios were downloaded from dbGAP. All offspring in the CRU cohort were older than 18 years and apparently healthy, only 13 of them were conceived at the time of the reactor accident, most of them many years later [33].

In addition to the two case cohorts described above, we accessed trio WGS control data from the Inova cohort of Wong et al. [28]. The Inova cohort comprises 1,214 familial trios (Inova, Supplemental Material 1.1.2, Supplemental Table S1), with no recorded history of exposure to non-naturally occurring IR [28,30,34]. This WGS data was also used to analyze the effect of parental age on isolated DNMs in the germline and genetic effects on preterm births [28,30,34]. We downloaded the Inova WGS data from an AWS S3 bucket.

Variant Calling and Quality Control

For the Radar and Inova cohorts, variant calling was performed using Illumina DRAGEN v3.6.3 in the Amazon web services cloud (Region: Ireland). For the CRU cohort, data were processed on an on-premises computing cluster using NVIDIA Parabricks (Supplemental Material 1.4.1). For variant analysis, all data were aligned to the GRCh37 reference genome, and joint variant calling was performed using GLnexus v1.3.1 on a total of 4,337 whole genome sequences [35]. For all samples and cohorts, we subsequently performed exhaustive quality control checks (Supplemental Material 1.4.3, Supplementary Figure S5, S6). These included sex and ancestry controls using Peddy, as well as assessments of contamination, sequencing depth and variants (e.g. transition-transversion ratio), in order to control for any technical bias that may have arisen secondary to differences in sequencing technology, chemistry, or other

forms of error [36–40]. Further detailed information about the variant calling efforts is present in the Supplemental Material, Section 1.4.1.

Detection of de novo and clustered de novo mutations in all parent offspring trios. To detect DNMs first and later cDNMs, a set of filters was applied to all three cohorts in parallel. For each of the three cohorts, the output of the variant calling pipeline formed the basis for the detection of DNMs and cDNMs. For each parent-offspring trio Python3 and Hail v0.2.89 were used to find potential DNM candidates and to refine this call set to clusters, as based on a window size of 20 bp [41]. The detection of DNMs was based on heuristics, including the score calculated by "hl.de_novo" (> 0.85); sequencing depth at the site (>10); parental genotypes and read data (< 2 reads featuring the variant allele); the allele count in all samples (AC <= 1); and other criteria (Supplemental Material 1.4.2). Supp. Table S2 shows the filtering settings that were used to detect DNMs. The most stringent filter employed in the detection of DNMs was the AC = 1 filter, whereby any variant with an $AC > 1$ in the combined cohort (Radar, CRU, Inova) was discarded. This filter removed all familial variants as well as DNMs and cDNMs with a high population allele frequency whose origin was thus unlikely to have been radiation [42]. To ascertain the quality of all *de novo* calls made in the Radar cohort, replication of each DNM call was attempted using Graphtyper. A concordance rate for DNM calls in each sample was then calculated [43–46].

Phasing. To determine the parental gamete of origin, read-based phasing of DNMs with informative variants was used (Supplemental Material, Section 1.4.4) [47]. Since clustered DNMs can extend over several base pairs, not all lesions in a cluster can necessarily be phased. If read-based phasing suggested the paternal or maternal germline based on the information from at least one lesion we assumed this origin for the whole cluster. Clusters where multiple DNMs showed differing evidence for paternal or maternal origin were called contradictory cDNMs (Supplemental Material, Section 1.4.4). In addition to read-based phasing, for a subset of cDNMs in the Radar cohort, the parental origin was also determined using Sanger and PacBio long-read sequencing. We resequenced a phase informative single nucleotide polymorphism (SNP) alongside each cluster, which uniquely identified the parental origin of each cluster.

cDNM Window Size. cDNMs are defined as genomic regions where at least two *de novo* mutations occur within 20 bp distance of each other. A window size of 20 bp was selected as the cutoff for cDNMs, since we hypothesized that the ROS-induced DSBs are the primary driver of radiation induced cDNMs in the human germline. ROS affects human DNA in a range of only 4-6nm, and 20 bp is the interval size used for these clusters in previous investigations [4]. Clusters of this size have also earlier been implicated in gonadal exposure to IR in humans and mice [7,8,18]. We also assessed different cluster window sizes that have been used in the literature in a sensitivity analysis (10 bp, 30 bp, 10k bp, 47k bp) by recalling all cDNMs with the given window size [2,8,14,48]. Cluster sizes in the range of 10k bp to 100k bp have previously been connected with the maternal age effect by some studies, but no association with ionizing radiation has been shown thus far [8,14,30].

cDNM Calling. Using the selected DNMs, which passed all filtering criteria, *de novo* mutation clusters were assembled using a trivial algorithm, whereby DNMs were added to a single cluster, if the distance to their direct predecessor on the same chromosome was < 20 bp. A cluster with > 2 lesions can span a total size larger than 20 bp.

Validation of cDNMs in the Radar Cohort.

Since cDNMs have a higher false positive rate than germline mutations and isolated de novo events, all cDNMs in the Radar cohort were validated by at least one of three methods using a three-step iterative approach (Supplemental Material 1.5, Supplemental Figure S7). This involved the use of Sanger and PacBio sequencing data to derive criteria for identifying true and false positive clusters, as based on the IGV Browser visualization. For both Sanger and PacBio sequencing, primers were first designed using Primer3, followed by manual optimization based on available short-read sequence data [49,50]. However, due to the complex nature of the genomic regions in which the candidate cDNMs were located, many were not validated in subsequent experiments.

Sanger sequencing was conducted on a subset of 71 potential cDNMs. These spanned the following categories: tandems (n=14); GG>TT tandems (n=15); indels (n=21); large cDNMs (involving more than three lesions; n=5); and cDNMs located within repetitive regions (n=16). Of the 71 analyzed clusters, interpretable results were obtained for 44. Specifically, 5 clusters were true positives, 39 clusters were false positives, and the status of 27 clusters remained undetermined due to sequencing challenges. Notably, none of the GG>TT tandem clusters achieved validated true positive status, emphasizing the difficulty in accurately sequencing certain mutation types in these regions. From this dataset, a set of guidelines were established for ascertaining which potential cDNMs were true or false positive calls (Supplemental Material 1.5, Supplementary Figure S7). The data from the final validation callset was used to establish the positive predictive value (PPV) of cDNM calls on the Radar cohort. In the statistical analysis, the PPV was used exclusively for a downsampling simulation (Supplemental Material 1.6.6).

Statistical Analysis

To ascertain potential statistical differences in the count data in this study (e.g. number of DNMs or cDNMs per offspring), generalized linear models were used (Supplemental Material 1.6.2). Under Bonferroni correction, the significance level was set at $\alpha = \frac{0.05}{9} = 0.00556$ for nominal p-values p_{nom} , or p-values were adjusted to $p_{adj} = 9p_{nom}$.

Since parental age at conception is a known and significant confounder for all analyses including DNMs, correction for the paternal and maternal age was performed in all analyses, unless otherwise stated. Since paternal and maternal age are highly correlated (Pearson-R: 0.71, $p < 5 \cdot 10^{-100}$), the age of the father at conception was used as a proxy for the effect of parental age, which includes both maternal and paternal age effects (Supplemental Material 1.1.4, Supplemental Figure S2). For models that did not incorporate paternal age directly, age

matching was used to control for this confounder (Supplemental Material 1.6.1), by selecting subcohorts of Radar, Inova and CRU with homogenous age distributions. The age matching procedure downsampled Inova and CRU to the same or n-times ($n \geq 1$) the size of the smallest cohort (Radar, n = 110). This was achieved by computing the minimum-weight bipartite matching between node sets representing the individual samples of any two cohorts. Two nodes in the graph, representing offspring in either cohort, were connected by an edge, whose weights were set to the sum of the age differences of the mothers and fathers. The minimum weight bipartite matching in this graph is then the subcohort of Inova and CRU respectively, that minimizes the age difference indicated by the edge weights in the graph (Supplemental Material 1.6.2).

Results

Estimated IR dose

Because some soldiers served in military roles that probably did not result in elevated levels of exposure, and due to the challenging retrospective dose estimations, the dose estimations remained inconclusive for the majority of soldiers ($n_{exposed} = 22$, $n_{no\ exposure} = 55$, $n_{no\ documents} = 3$, Supplementary Figure S3, Supplementary Table S10, Supplementary Material "Bericht S209/20"). We call the offspring of soldiers with an estimated dose of > 0 mGy the Exposed subcohort, while the Unexposed subcohort was comprised of all children of fathers that were deemed unlikely to be exposed.

Dose estimations for the Radar cohort were performed after the recruitment phase ended and the average estimated IR dose in the total Radar cohort was 9.21 ($\pm$ 53.33) mGy (median = 0 mGy). In the subgroup of radar technicians with a dose estimation of > 0 mGy (n = 22), the average was $\mu = 34.35$ ($\pm$99.77) mGy ($median = 0.0021$ mGy) (Figure 1a) [32].

For the CRU cohort, dose estimations were accessed from the data published by Yeager, et al. [8]. On average, fathers in the CRU cohort were exposed to $\mu = 365.42$ ($\pm$684.55) mGy ($median = 29$) of IR (Figure 1a, Supplementary Material, Section 1.1.3, Supplementary Figure S1).

Analysis of de novo mutations

After obtaining all necessary data, we processed all samples using the equivalent bioinformatics pipelines, as detailed in the Materials and Methods section. Before computing the set of DNMs, we confirmed that all quality control checks passed, in particular that there was no substantial difference in the whole genome coverage between the cohorts, and that the pedigree in all cohorts matches the reported family structure (Figure 1b, Supplementary Figure S4). In accordance with the literature, the number of isolated DNMs increased by 2% per year of paternal age in all three cohorts (Figure 2, Supplementary Table S5), which translates to an accumulation of 1-2 mutations per year of age of the father [28,29,51,52]. In the age-matched analyses, the rate of isolated DNMs per generation was: (i) Inova, 72.67 (18.15, median = 79);

(ii) CRU, 65.43 (13.57, median = 65); (iii) Radar, 67.95 (17.25, median=64) (Supplemental Tables S1, S3). For Inova and CRU, these rates are comparable to the values reported in the original studies. No significant difference in the rate of isolated DNMs per generation was found between the three cohorts (Supplemental Material, Section 1.6.3, Supplementary Table S4) [8,28]. None of the datasets showed a bias towards specific nucleotide exchanges, as has been reported previously for certain generations of sequencing devices (Supplemental Figure S8) [31]. Our bioinformatic replication using the Graphtyper algorithm yielded a concordance of ≥ 88 % for DNMs in the radar cohort. The sequencing replicates of three families yielded a PPV of 90.2% for the DNM calls made on the NovaSeq, assuming calls from the HiSeq as ground truth. (Supplementary Table S14).

Analysis of clustered de novo mutations

We continued our analysis by filtering for loci with multiple lesions (clustered DNMs, cDNMs). In total, 1,989 cDNMs were detected in 1,515 offspring ($n_{Inova} = 1275, n_{Radar} = 110, n_{CRU} = 130$). These mutations were enriched in offspring of irradiated fathers in the Radar cohort ($\mu = 1.48 \pm 1.72, median = 1$) and in the CRU cohort ($\mu = 2.65 \pm 2.65, median = 2$) compared to the Inova cohort ($\mu = 0.88 \pm 0.98, median = 1$) (Supplemental Table S6). In offspring from the CRU cohort, the median number of clustered DNMs was two, which was twice as many as that detected in the age-matched subset of the Inova cohort. A negative binomial regression model confirmed that the estimated number of cDNMs in the Inova cohort was less than in the Radar or CRU cohort ($n = 110, p_{adj}^{Radar} = 0.045, p_{adj}^{CRU} < 1 \cdot 10^{-3}$; Figure 3, Supplemental Table S7). These differences were more prominent when the Radar cohort was divided into the Exposed and Unexposed subcohorts, where the children of exposed parents showed a higher number (Exposed = 1.72, Unexposed = 1.39) of cDNMs on average (Supplemental Material, Section 1.6.8, Supplementary Figure S10).

We also found cDNMs to be significantly increased in offspring of irradiated fathers for 10 bp and 30 bp windows in our sensitivity analysis. In contrast, the larger window sizes led to a reduction in the difference in the number of cDNMs between the cohorts and a substantial drop in p_{nom} (Supplemental Material, Section 1.6.12, Supplemental Figures S13, S14).

cDNM Validation in the Radar cohort

In general, the false positive rate is higher for clustered DNMs [48]. Our visual inspection criteria were applied to 163 cDNMs in the Radar cohort. Of these, 37 were found to be true positives, 17 of which were also confirmed by PacBio and/or Sanger sequencing data. Therefore, the final PPV for cDNM detection in the Radar cohort was $\frac{37}{163} = 0.23$ (95% Clopper-Pearson confidence interval 0.17 - 0.30). Notably, in some individuals, none of the detected cDNMs could be validated, including the outlier with 14 cDNMs shown in Figure 3. However, this had no substantial effect on the negative binomial regression model, since this models the median count of cDNMs per offspring, which is robust against outliers. Additionally, simulations

accounting for this PPV did not affect the significance of the test results (Supplemental Material, Section 1.6.6, Supplementary Table S8).

All true positive cDNMs identified in the Radar cohort were analyzed with respect to their likely relevance to disease states or their impact on the coding region in general. None was found to have any implications in terms of genetic conditions reported by the study participants (Supplemental Material, Section 1.5.1, Supplementary Data "Clinical Data").

Phasing of DNMs and cDNMs

For technical and stochastic reasons, the proportion of DNMs that could be phased varied between the three cohorts. The influencing factors were the distance between DNM and the phase-informative SNP, the coverage in the respective region, the length of the sequencing reads (100bp in the control cohort vs 150bp in both case cohorts), and the distribution of fragment sizes. No inter-cohort differences were observed in the number of isolated DNMs that were attributable to the paternal or maternal alleles (Chi-Squared-Test, Supplemental Material, Section 1.6.7, Supplementary Figure S9, Supplementary Table S9). We did not observe any contradictory cDNM clusters. The parental origin of 26 clusters in the Radar cohort was validated, with 17 clusters of paternal and nine of maternal origin being present. Due to the shorter read length of 100bp in the Inova cohort, no reliable estimate for this ratio in the population could be computed.

Analysis of radiation exposure

In addition to an increase in the number of cDNMs per sample in the Radar and CRU cohorts, a positive correlation was found between the estimated dose and the number of cDNMs per sample. Using a negative binomial regression model, a significant ($p_{adj} < 0.009$) increase in cDNMs was observed per mGy of paternal radiation exposure. The regression model estimated the increase of cDNMs as $f(n) = 1.55 \cdot e^{0.0005n}$ mutations per n mGy, when combining the Radar and CRU cohorts and $\beta_{CRU} = 0.0005$ and $\beta_{Radar} = 0.0007$ when analyzing each cohort separately (Figure 4, Supplemental Material 1.6.10, Supplementary Table S12, Supplementary Figure S12). However, it was not possible to assert statistical significance for this model in either the Radar cohort alone, or for the inverse model, by inferring the paternal IR dose from the cDNM count of the respective offspring (Supplementary Section 1.6.11, Supplementary Table S13). The highest number of DNMs per cluster observed in the three cohorts was eight (Inova), nine (Radar), and 11 (CRU) mutations respectively (Supplemental Figure S11, Supplementary Table S11). An analysis of the distribution of cluster sizes across the three cohorts yielded no statistically significant shift (Supplemental Material, Section 1.6.9).

Discussion

The question of whether IR confers transgenerational health effects on the human genome has been a topic of research for over 70 years, i.e. since epidemiological studies first investigated the offspring of atomic bomb survivors [53]. However, the disadvantage of epidemiological

analyses is that they require a readout on the phenotypic-level such as malformations [53]. More recent studies indicate that larger quantities of environmental radiation lead to a higher incidence of cancer and birth defects, emphasizing that there are subtle effects of ionizing radiation on the human germline that have not been captured by earlier studies [54]. If an analysis is based on phenotypic data alone, an increased rate of *de novo* mutations or other subtle changes may remain unobserved, since most *de novo* mutations are rare and occur in the non-coding part of the DNA. Therefore, the unique capabilities of whole genome sequencing can yield much deeper insights into the consequences of prolonged ionizing radiation exposure on the human genome than possible with earlier sequencing technologies. Yeager, et al. already studied isolated *de novo* mutations and clusters on the scale of kb, which are associated with the repair of double-strand breaks, and did not find a significant increase in mutation rates [8,28]. Lawrence, et al. and Moorhouse, et al. studied DNMs, structural variants and chromosomal aberrations in offspring of British nuclear test veterans with unclear total radiation exposure. Similar to Yeager, et al., they did not find increased mutation rates in any of their target mutation types, which included DNM clusters with 10 bp or 100 bp size [2,3]. In mice and somatic cells however, clusters on the single-bp scale were previously implied as a consequence of the secondary effect of high-energy particles interacting with DNA [25–27]. The present study investigated the presence of cDNMs in WGS data from the offspring of parents who had been exposed to IR either during past military service (the Radar cohort) or following the Chernobyl nuclear accident (CRU cohort of Yeager, et al., 2021). A significant increase in de novo mutation clusters was found in offspring of irradiated parents compared to controls. Furthermore, our statistical models indicated that the number cDNMs detected in offspring increased with the estimated dose of paternal IR exposure.

Most of the present statistical results indicate an influence of paternal radiation exposure on the number of cDNMs per offspring, even when only the smaller and less exposed Radar cohort is considered. Our sensitivity analysis showed that small window sizes have larger effect sizes, i.e. the difference in cDNM count between the case and control cohorts were the largest in the smaller cluster sizes (10bp - 30 bp). These observations demonstrate the lack of bias in the choice of cluster sizes and further support the hypothesis that the ROS-induced DSBs primarily result in clusters on the single- or double-digit scale. These findings are further supported by the increase of effect size and statistical significance observed in the Exposed subgroup of the Radar cohort.

In the offspring of the Radar and CRU cohort, we observed that the number of cDNMs increased by one to two per genome. To derive a clinical interpretation of these statistical results, the number and impact of cDNMs was compared with the disease burden due to all DNMs. The total number of cDNMs exceeds that of the general population by 0.6 for the Radar and 1.77 for the CRU cohort. In addition to the increase in the total number of DNMs per sample secondary to radiation exposure, it is plausible that the functional impact of cDNMs is larger compared to isolated DNMs, if they fall within coding regions of the human genome. This increased impact could lead to pathogenicity, or even embryonic lethality, in cases where cDNMs affect important parts of the coding region. However, the present authors are of the opinion that given the low overall increase in cDNMs following paternal exposure to ionizing

radiation and the low proportion of the genome that is protein coding, the likelihood that a disease occurring in the offspring of exposed parents is triggered by a cDNM is minimal. Therefore, with a paternal age effect of approximately 1 additional DNM per year of paternal age, and an expected average of 60 to 80 DNMs per generation, we conclude that paternal exposure to low dose IR contributes less to an individual's risk for genetic diseases than age. Thus, based on the current state of knowledge, the excess risk attributable to cDNMs that arose after paternal exposure to IR is negligible compared to the base risk for genetic diseases. These findings are consistent with the reports made on British nuclear test veterans, whereby no contribution to genetic disease in the offspring of former soldiers was found for potentially radiation-induced mutational patterns [2,3]. While the expected clinical consequences of a clustered or isolated DNM is of comparable order, the consequences of a DSB that is incorrectly repaired, is usually more severe. Translocations are most likely to represent an indirect consequence of DSBs, and have been observed with increased frequency in irradiated mice as well as in the offspring of Radar soldiers [7,25,26]. However, in contrast to cDNMs, comparing mutation rates for structural variants is more prone to errors when the respective cohorts were sequenced using different short read lengths and we have therefore refrained from assessing these statistically.

The present study had three main limitations. These concern the issues of IR dose estimation, the calling accuracy of cDNMs, and recruitment bias. Dose estimations, i.e. data on the level of exposure to IR for each soldier were retrospective and limited. In practice, that meant that radar devices that had been in active service more than 50 years ago had to be rendered operational again in order to measure scattered radiation dose profiles (Supplemental Material "Bericht S209/20") [32].In addition, while the service hours and proximity to the radar device during operation and maintenance were derived from a generalized service manual of the German armies based on rank, position and mission of the soldier, in many cases the recollections of study participants differed, suggesting that this approach introduced a potential source of errors. For example, anecdotal evidence from the Radar cohort participants suggests that in contrast to official records, higher ranking soldiers participated in radar maintenance work. Thus, some of the individuals who were classified as unexposed in the present analysis, might actually have been irradiated (Figure 3 and 4), and dose values given for members of the Radar cohort are likely to have been underestimated. Additionally, since these estimations are based largely on measurements taken in a laboratory setting years after these devices have been removed from service, they should be considered inaccurate. Similar errors might be present in the CRU cohort, due to the large delay between radiation exposure and conception of most childs in their cohort, and potential inaccuracies in dose assessment [33,55]. Despite the discussed inaccuracies of dose estimation, we proceeded with dose-effect estimations, and excluded offspring of allegedly not exposed fathers from the negative binomial regression models, including the outlier with 14 cDNMs (Figure 4). This is likely to have introduced Berksonian and classical errors into our models [56].

The second limitation is that a comprehensive validation of cDNM calls in all three cohorts was infeasible. We lacked DNA material to perform any validation experiments for the CRU

or Inova cohorts, rendering us unable to assess the PPV of cDNM calls on this cohort independently from the Radar cohort.

A third limitation of the study was the presence of several potential ascertainment biases during recruitment. First, individuals who were under the subjective impression that they had been exposed to IR during their term of military service were more likely to participate (volunteer bias). Second, former radar soldiers who had operated devices emitting the highest quantities of stray radiation, were in their eighties at the time of recruitment. Furthermore, radar soldiers had a high personal risk for diseases following their service (survivorship bias) [1]. Additional biases may also have arisen due to geographical effects. While all three investigated cohorts shared a common genetic ancestry, the possibility that environmental effects contributed to the observed differences, can not be ruled out [54]. To our knowledge, the geographical origins of the three cohorts, i.e. Germany, Ukraine, and the East Coast of the USA, do not have higher than average levels of background radiation. In the present analysis, the background radiation dose was therefore assumed to be similar for all three cohorts.

The present results suggest several avenues for future research. First, studies with longer read lengths, ideally larger cohorts, and more accurate radio-dosimetry are required to improve the characterization of the dose-response relationships and disease risk of transgenerational signatures of prolonged paternal exposure to low dose IR, such as cDNMs. Second, to determine the paternal to maternal cDNM ratio in the general population, deep sequencing of an appropriate cohort using a greater read length than what was possible in the present study is required. Accurate measurement of the paternal to maternal cDNM ratio in the general population is necessary in order to enable an accurate assessment of the influence of IR exposure, since the number of clusters of paternal and maternal origin is expected to differ due to the accumulation of repair errors In exposed cohorts, a further shift towards more paternally inherited clusters would provide additional evidence for the correlation between IR exposure in the fathers and cDNM rates in the offspring. Third, modeling the gonadal dose of fathers based on the basis of cDNMs in the respective offspring could provide further interesting avenues for analysis, if the positive predictive value for cDNM detection were to improve. Currently, these models are impacted by the low sample size and the low number of true positive cDNMs per sample. However, a plausible hypothesis is that a more specific analysis would pinpoint this relationship. Fourth, subjecting samples to long read sequencing would render targeted statistical analysis of structural variants and translocations in the general population compared to exposed cohorts feasible. Fifth, further investigation of the potential impact of the linear energy transfer (LET) on the cluster size would be of interest. When individuals are subjected to IR with higher LET, the damage would be expected to increase in direct proportion to the LET level, leading to larger clusters, or an increased number of structural variants. This effect could also explain differences in the number and nature of the clusters observed between the two exposed cohorts in the present study, since the gamma ray spectrum of the IR to which the two cohorts were exposed differed, with a difference in the LET being one of the consequences thereof [57]. Moreover, the significant difference in radiosensitivity of mature sperm and spermatogonial stem cells should be taken into account.

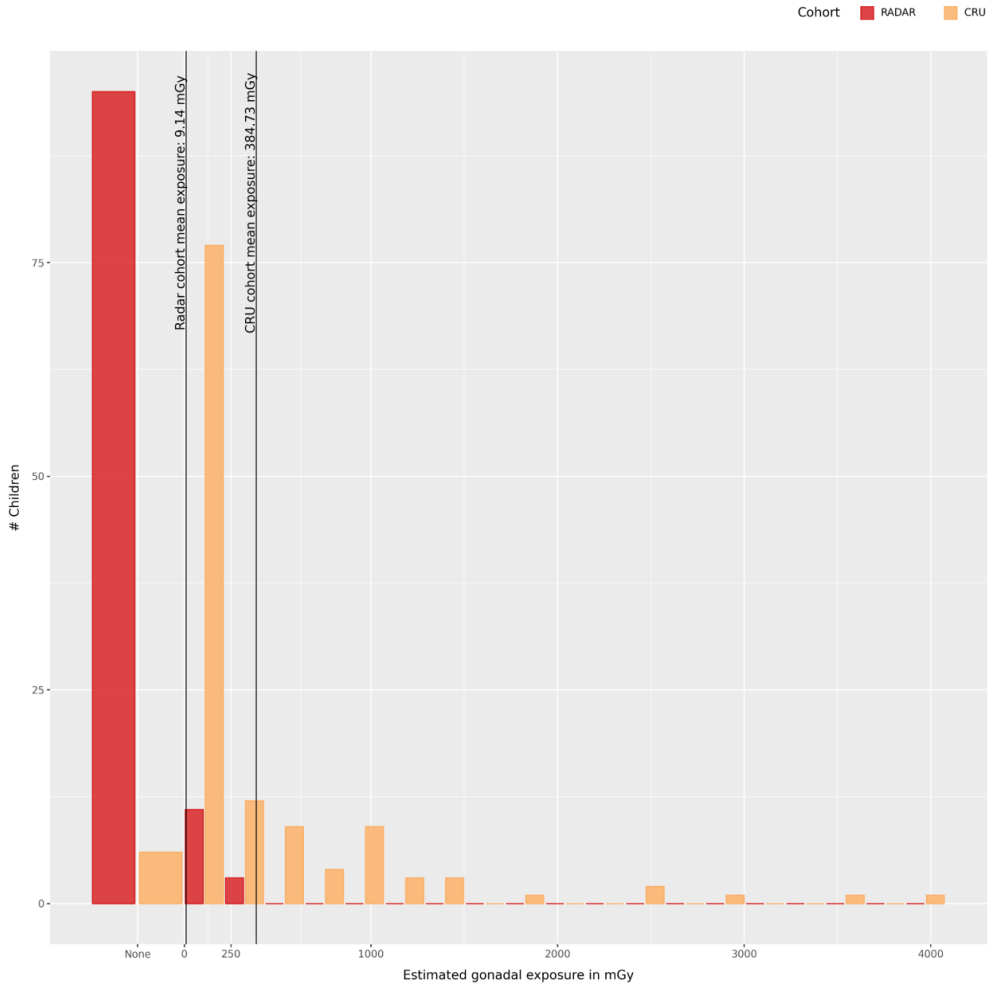
In conclusion, we found a significant increase in the cDNM count in offspring of irradiated parents, and a potential association between the dose estimations and the number of cDNMs in the respective offspring. Despite uncertainty concerning the precise nature and quantity of the IR involved, the present study is the first to provide evidence for the existence of a transgenerational effect of prolonged paternal exposure to low-dose IR on the human genome. The additional risk due to IR induced cDNMs on the scale of single base-pairs was very low. The present findings suggest several further promising research avenues for characterizing further transgenerational signatures of the effect of IR on the human genome, including the analysis of structural changes such as translocations, which are more complicated to detect than cDNMs.

Figures

Figure 1: Study Cohorts

- **a)** Distribution of paternal exposure for the Radar and CRU cohorts. The maximum exposure observed in the Radar cohort is 353 mGy, and 4,079 mGy in the CRU cohort. In the Radar cohort, 77 out of 110 children were born to soldiers with a dose estimation of 0 mGy, 30 to soldiers with a valid exposure estimation > 0 mGy and for the father of 3 offspring, the dose estimation could not be made.
- **b)** Age distribution of the three study cohorts. Due to the large differences in cohort size, the y-Axis indicates the percentage of the total cohort size. Values in the bottom half of the y-Axis show the distribution of maternal age, and values in the top half show the distribution of paternal age. On average, fathers in the Inova cohort were >5 years older compared to fathers in the Radar and CRU cohorts, and mothers were >5.5 years older on average.

a)


Cohort
RADAR
CRU
Radar cohort mean exposure: 9.14 mGy
CRU cohort mean exposure: 384.73 mGy
Children
75
50
25
0
None
0
250
1000
2000
3000
4000
Estimated gonadal exposure in mGy

b)

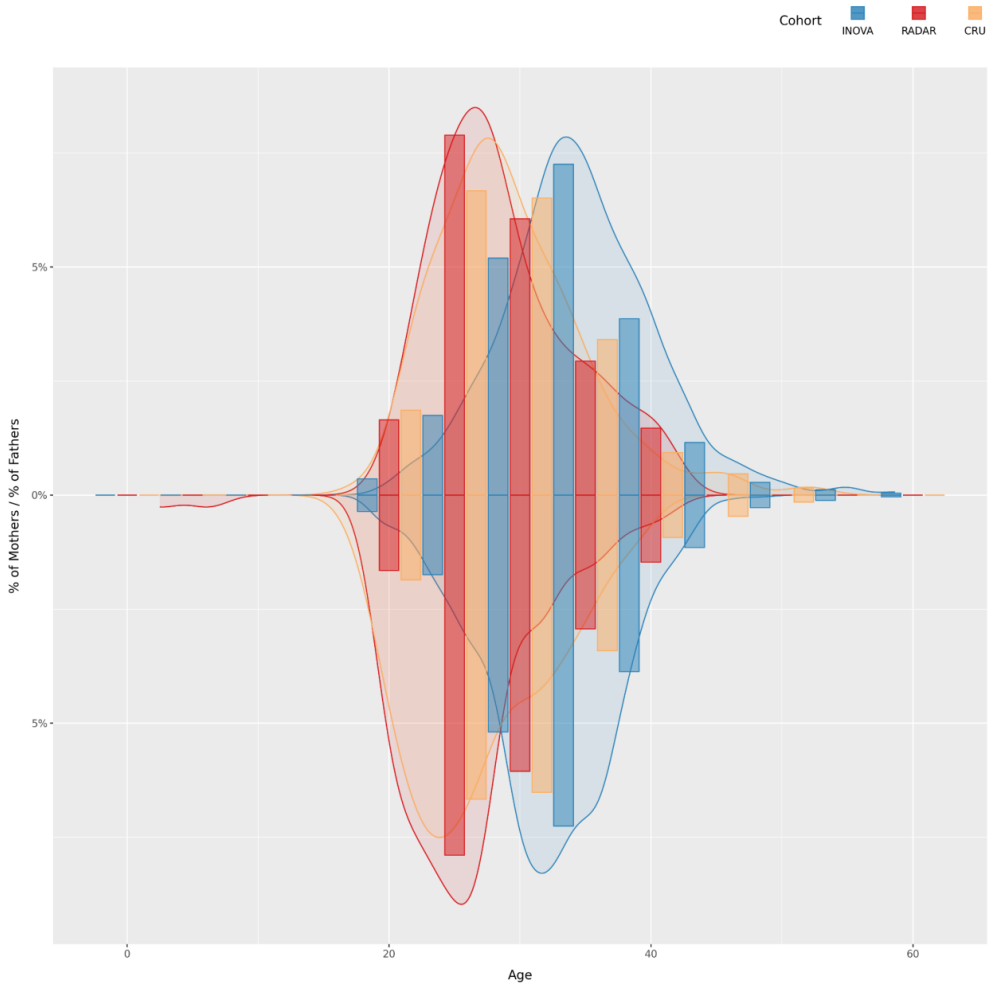

Cohort
INOVA
RADAR
CRU
5%
0%
5%
% of Mothers / % of Fathers
0
20
40
60
Age

Figure 2: Paternal Age Effect

Paternal age effects computed by a negative binomial regression model to estimate the number of DNMs according to the paternal age at conception for the offspring in each cohort. When fitting this model, no age matching was applied to the data. Therefore, on average, the parents are older in the Inova cohort. Nevertheless, the paternal age effect for each of the cohorts is approximately 2%, which results in an increase of ~1 DNM per year of paternal age.

Figure 3: Number of cDNMs per sample

Violin plot of the number of clustered de novo mutations (cDNMs) per sample, as grouped according to cohort. The width of the violin at each integer value of the y-axis indicates the number of samples and their respective number of cDNM clusters, without correcting for the PPV of 0.23. Our simulation experiments controlling for the effects of this PPV on the statistical tests are presented in Supplemental Table S8. The box plot for each cohort is included inside the respective violin to display the quartile ranges and median number of cDNMs per sample in the respective cohort. On average, the age matched analysis detected 0.88 ($\pm$0.98; $median = 1$; $ppv - adjusted = 0.20$) cDNMs in the Inova cohort, 1.48 ($\pm$1.72; $median = 1$; $ppv - adjusted = 0.34$) in the Radar cohort, and 2.65 ($\pm$2.19; $median = 2$; $ppv - adjusted = 0.61$) in the CRU cohort.

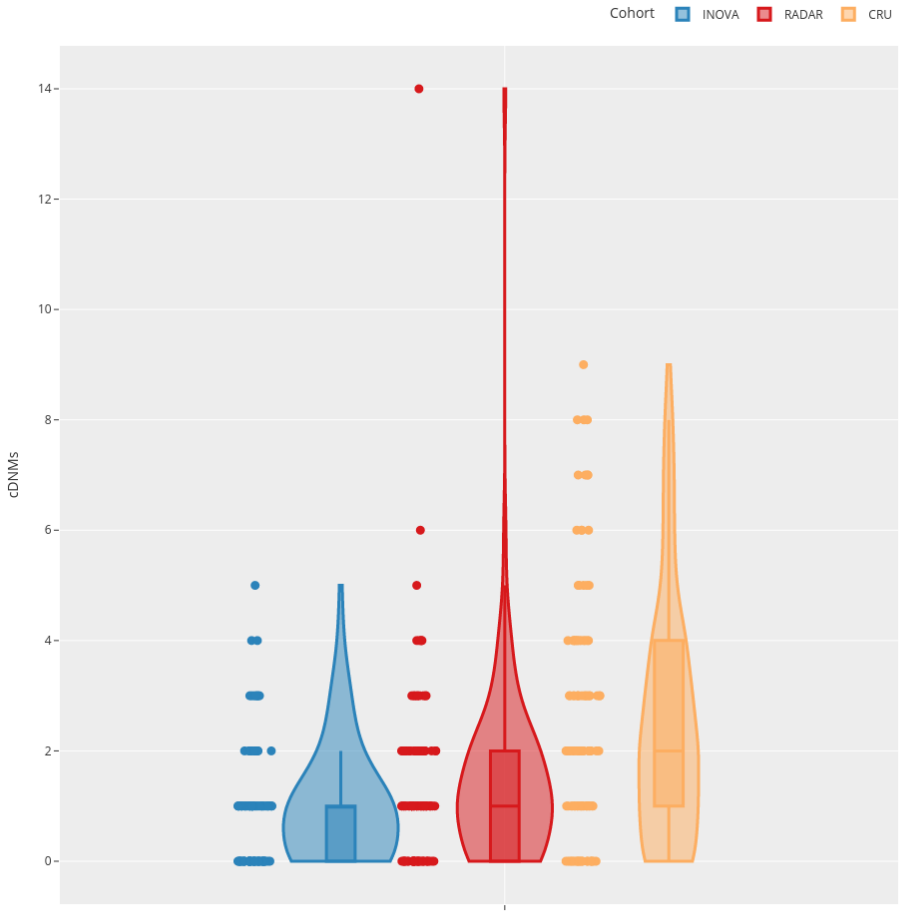
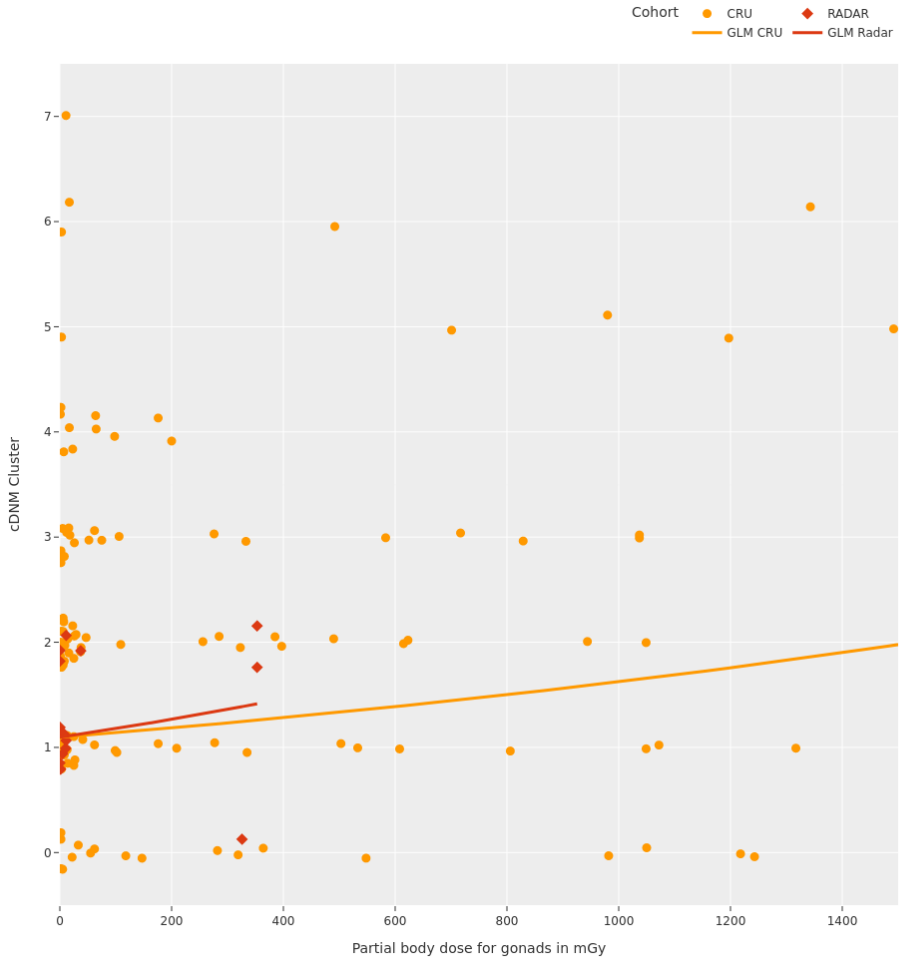


Figure 4: Number of cDNMs per mGy of paternal exposure
Estimation of the number of cDNMs according to the paternal exposure in mGy as computed by a negative binomial regression model with logarithmic link function. Since the accuracy of the negative binomial regression model deteriorates rapidly with larger exposure estimates, the x-axis has been cut off at 1.5k mGy, which means that five samples from the CRU cohort, for which the estimation by the model is very inaccurate, are hidden. The fit shown in this image is conditional on the number of mGy of paternal exposure and the cohort. A model integrating both cohorts into one exposed supercohort is shown in Supplemental Figure S12. This was restricted to the samples from the Exposed subgroup of the Radar cohort, i.e. all offspring of fathers with an estimated exposure of >0 mGy. This restriction in the cohort leads to the exclusion of some offspring with many cDNMs, e.g. the outlier in the Radar cohort in Figure 3.



Data Availability

- Sequencing data can be accessed at ega:
    - EGA: Study Accession EGAS00001007321
- Code can be accessed at github and zenodo:
    - https://github.com/brand-fabian/radarstudy
    - https://doi.org/10.5281/zenodo.8431077

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgements

The Radar study was funded by the German Ministry of Defense and conducted by Deutsches Zentrum für Luft und Raumfahrt (DLR) under the supervision of an expert committee consisting of Michael Krawczak, Hajo Zeeb, Ivo Gut, André Reis and experts from the Bundeswehr (Reinhard Ullmann, Matthias Port). Ionizing radiation doses for the present study were estimated by Andreas Schirmer from the "*Strahlenmessstelle der Bundeswehr"* (Radiation measurement facility of the German Federal Armed Services). The authors thank Dietmar Glaner, Josef Wiesner, and the "Bund zur Unterstützung Radargeschädigter (BzUR e.V.)" for their assistance with the recruitment of the Radar cohort. The authors also thank Hákon Jónsson from DeCode for helpful discussions on the de novo mutation rates in unexposed controls. Data was generated by the NGS core facility at Universitätsklinikum Bonn, which is part of the West German Genome Center (WGGC) and processed by the Core Unit for Bioinformatics in Bonn. Epigenetic analysis was performed by Sascha Tierling and Jörn Walter from the University Saarland. The authors thank Christine Schmäl for proofreading and language editing of the manuscript.

Author contributions

Conceptualization: FB, HK, AK, MH, LW, DB, MN, MS, KS, PMK
Data curation: FB, AK
Formal analysis: FB, HK, LW
Funding acquisition: PMK, DB
Investigation: FB, HK, AK
Methodology: FB, HK, AK, LW, MS, PMK
Project administration: PMK
Resources: FB, AK, MH, KUL, PK, GM, DB, PMK

Supervision: DB, MN, MS, KS, PMK
Software: FB
Visualization: FB
Writing – original draft: FB, HK, AK, PMK
Writing – review & editing: FB, HK, AK, PMK

Bibliography

1. König, W. *et al.* Bericht der Expertenkommission zur Frage der Gefährdung durch Strahlung in früher Radareinrichtungen der Bundeswehr und der NVA (Radarkommission)[Report of the expert group on the issue of the hazards of radiation in earlier types of radar devices and the German Radar Commission]. Berlin, German Radar Commission. (2003).
2. Moorhouse, A. J. *et al.* No evidence of increased mutations in the germline of a group of British nuclear test veterans. *Sci. Rep.* **12**, (2022).
3. Lawrence, K. J. *et al.* M-FISH evaluation of chromosome aberrations to examine for historical exposure to ionising radiation due to participation at British nuclear test sites. *J. Radiol. Prot.* **44**, 011501 (2024).
4. Dubrova, Y. E. *et al.* Human minisatellite mutation rate after the Chernobyl accident. *Nature* **380**, 683–686 (1996).
5. Slebos, R. J. *et al.* Mini-and microsatellite mutations in children from Chernobyl accident cleanup workers. *Mutat. Res. Toxicol. Environ. Mutagen.* **559**, 143–151 (2004).
6. Degrave, E., Meeusen, B., Grivegnée, A.-R., Boniol, M. & Autier, P. Causes of death among Belgian professional military radar operators: A 37-year retrospective cohort study. *Int. J. Cancer* **124**, 945–951 (2009).
7. Holtgrewe, M. *et al.* Multisite de novo mutations in human offspring after paternal exposure to ionizing radiation. *Sci. Rep.* **8**, 1–5 (2018).
8. Yeager, M. *et al.* Lack of transgenerational effects of ionizing radiation exposure from the Chernobyl accident. *Science* **372**, 725–729 (2021).

9. Dubrova, Y. E. *et al.* Human minisatellite mutation rate after the Chernobyl accident. *Nature* **380**, 683–686 (1996).
10. Little, M. P., Goodhead, D. T., Bridges, B. A. & Bouffler, S. D. Evidence relevant to untargeted and transgenerational effects in the offspring of irradiated parents. *Mutat. Res. Mutat. Res.* **753**, 50–67 (2013).
11. Stephens, J. *et al.* A systematic review of human evidence for the intergenerational effects of exposure to ionizing radiation. *Int. J. Radiat. Biol.* **100**, 1330–1363 (2024).
12. Amrenova, A. *et al.* Intergenerational effects of ionizing radiation: review of recent studies from human data (2018–2021). *Int. J. Radiat. Biol.* **100**, 1253–1263 (2024).
13. Chumak, V. *et al.* Estimation of radiation gonadal doses for the American–Ukrainian trio study of parental irradiation in Chornobyl cleanup workers and evacuees and germline mutations in their offspring. *J. Radiol. Prot.* **41**, 764–791 (2021).
14. Jónsson, H. *et al.* Parental influence on human germline de novo mutations in 1,548 trios from Iceland. *Nature* **549**, 519–522 (2017).
15. Roberts, S. A. & Gordenin, D. A. Hypermutation in human cancer genomes: footprints and mechanisms. *Nat. Rev. Cancer* **14**, 786–800 (2014).
16. Sage, E. & Shikazono, N. Radiation-induced clustered DNA lesions: Repair and mutagenesis. *Free Radic. Biol. Med.* **107**, 125–135 (2017).
17. Georgakilas, A. G., O'Neill, P. & Stewart, R. D. Induction and repair of clustered DNA lesions: what do we know so far? *Radiat. Res.* **180**, 100–109 (2013).
18. Frankenberg-Schwager, M. Induction, repair and biological relevance of radiation-induced DNA lesions in eukaryotic cells. *Radiat. Environ. Biophys.* **29**, 273–292 (1990).
19. Scully, R., Panday, A., Elango, R. & Willis, N. A. DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat. Rev. Mol. Cell Biol.* **20**, 698–714 (2019).

20. Wang, S., Meyer, D. H. & Schumacher, B. Inheritance of paternal DNA damage by histone-mediated repair restriction. *Nature* **613**, 365–374 (2023).
21. Sage, E. & Harrison, L. Clustered DNA lesion repair in eukaryotes: relevance to mutagenesis and cell survival. *Mutat. Res. Mol. Mech. Mutagen.* **711**, 123–133 (2011).
22. Pagès, V. & Fuchs, R. P. How DNA lesions are turned into mutations within cells? *Oncogene* **21**, 8957–8966 (2002).
23. Jan, S. Z. *et al.* Unraveling transcriptome dynamics in human spermatogenesis. *Development* **144**, 3659–3673 (2017).
24. Wdowiak, A., Skrzypek, M., Stec, M. & Panasiuk, L. Effect of ionizing radiation on the male reproductive system. *Ann. Agric. Environ. Med.* **26**, 210–216 (2019).
25. Adewoye, A. B., Lindsay, S. J., Dubrova, Y. E. & Hurles, M. E. The genome-wide effects of ionizing radiation on mutation induction in the mammalian germline. *Nat. Commun.* **6**, 6684 (2015).
26. Satoh, Y. *et al.* Characteristics of induced mutations in offspring derived from irradiated mouse spermatogonia and mature oocytes. *Sci. Rep.* **10**, 1–13 (2020).
27. Matsuda, Y. *et al.* Spectra and characteristics of somatic mutations induced by ionizing radiation in hematopoietic stem cells. *Proc. Natl. Acad. Sci.* **120**, e2216550120 (2023).
28. Wong, W. S. *et al.* New observations on maternal age effect on germline de novo mutations. *Nat. Commun.* **7**, 10486 (2016).
29. Kong, A. *et al.* Rate of de novo mutations and the importance of father's age to disease risk. *Nature* **488**, 471–475 (2012).
30. Goldmann, J. M. *et al.* Germline de novo mutation clusters arise during oocyte aging in genomic regions with high double-strand-break incidence. *Nat. Genet.* **53**, 1270–1270 (2018).

31. Arora, K. *et al.* Deep sequencing of 3 cancer cell lines on 2 sequencing platforms. *bioRxiv* 623702 (2019).

32. Schirmer, A. *Bericht S209/20: Retrospektive Dosisberechnung Röntgenstörstrahlung*. (2021).

33. Bazyka, D. *et al.* Field Study of the Possible Effect of Parental Irradiation on the Germline of Children Born to Cleanup Workers and Evacuees of the Chornobyl Nuclear Accident. *Am. J. Epidemiol.* **189**, 1451–1460 (2020).

34. Knijnenburg, T. A. *et al.* Genomic and molecular characterization of preterm birth. *Proc. Natl. Acad. Sci.* **116**, 5819–5827 (2019).

35. Lin, M. F. *et al.* GLnexus: joint variant calling for large cohort sequencing. *BioRxiv* 343970 (2018).

36. Pedersen, B. S. & Quinlan, A. R. Who's who? Detecting and resolving sample anomalies in human DNA sequencing studies with peddy. *Am. J. Hum. Genet.* **100**, 406–413 (2017).

37. Jun, G. *et al.* Detecting and estimating contamination of human DNA samples in sequencing and array-based genotype data. *Am. J. Hum. Genet.* **91**, 839–848 (2012).

38. Pedersen, B. S. & Quinlan, A. R. Mosdepth: quick coverage calculation for genomes and exomes. *Bioinformatics* **34**, 867–868 (2018).

39. Li, H. *et al.* The sequence alignment/map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

40. Li, H. A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).

41. Hail Team. Hail 0.2.89. (2023).

42. Wang, Q. *et al.* Landscape of multi-nucleotide variants in 125,748 human exomes and 15,708 genomes. *Nat. Commun.* **11**, 1–13 (2020).

43. Eggertsson, H. P. *et al.* Graphtyper enables population-scale genotyping using pangenome graphs. *Nat. Genet.* **49**, 1654–1660 (2017).

44. Eggertsson, H. P. *et al.* GraphTyper2 enables population-scale genotyping of structural variation using pangenome graphs. *Nat. Commun.* **10**, 5402 (2019).

45. Lin, Y.-L. *et al.* Comparison of GATK and DeepVariant by trio sequencing. *Sci. Rep.* **12**, 1809 (2022).

46. Hanssen, F. *et al.* NCBench: providing an open, reproducible, transparent, adaptable, and continuous benchmark approach for DNA-sequencing-based variant calling. *F1000Research* **12**, 1125 (2024).

47. Martin, M. *et al.* WhatsHap: fast and accurate read-based phasing. *BioRxiv* 085050 (2016).

48. Besenbacher, S. *et al.* Multi-nucleotide de novo Mutations in Humans. *PLOS Genet.* **12**, e1006315 (2016).

49. Koressaar, T. & Remm, M. Enhancements and modifications of primer design program Primer3. *Bioinformatics* **23**, 1289–1291 (2007).

50. Untergasser, A. *et al.* Primer3—new capabilities and interfaces. *Nucleic Acids Res.* **40**, e115–e115 (2012).

51. Sasani, T. A. *et al.* Large, three-generation CEPH families reveal post-zygotic mosaicism and variability in germline mutation accumulation. *bioRxiv* 552117 (2019).

52. Wood, K. A. & Goriely, A. The impact of paternal age on new mutations and disease in the next generation. *Fertil. Steril.* **118**, 1001–1012 (2022).

53. Yamada, M. *et al.* Congenital malformations and perinatal deaths among the children of atomic bomb survivors: a reappraisal. *Am. J. Epidemiol.* **190**, 2323–2333 (2021).

54. Zlobina, A. *et al.* Impact of Environmental Radiation on the Incidence of Cancer and Birth Defects in Regions with High Natural Radioactivity. *Int. J. Environ. Res. Public. Health* **19**, 8643 (2022).
55. Schmitz-Feuerhake, I. RE: Comment to Yeager M et al. Lack of transgenerational effects of ionizing radiation exposure from the Chernobyl accident. Science 2021 Apr 22. (2021).
56. Bellamy, M. B. *et al.* Recommendations on statistical approaches to account for dose uncertainties in radiation epidemiologic risk models. *Int. J. Radiat. Biol.* **100**, 1393–1404 (2024).
57. Semenenko, V. & Stewart, R. A fast Monte Carlo algorithm to simulate the spectrum of DNA damages formed by ionizing radiation. *Radiat. Res.* **161**, 451–457 (2004).

3.4 Systematic assessment of COVID-19 host genetics using whole genome sequencing data

Schmidt, Casadei, Brand, Demidov, Vojgani, Abolhassani, Aldisi, Butler-Laporte, group, Alawathurage, Augustin, Bals, Bellinghausen, Berger, Bitzer, Bode, Boos, Brenner, Cornely, Eggermann, Erber, Feldt, Fuchsberger, Gagneur, Göpel, Haack, Häberle, Hanses, Heggemann, Hehr, Hellmuth, Herr, Hinney, Hoffmann, Illig, Jensen, Keitel, Kim-Hellmuth, Koehler, Kurth, Lanz, Latz, Lehmann, Luedde, Maj, Mian, Miller, Muenchhoff, Pink, Protzer, Rohn, Rybniker, Scaggiante, Schaffeldt, Scherer, Schieck, Schmidt, Schommers, Spinner, Vehreschild, Velavan, Volland, Wilfling, Winter, Richards, DeCOI, Heimbach, Becker, Ossowski, Schultze, Nürnberg, Nöthen, Motameny, Nothnagel, Riess, Schulte, and Ludwig “Systematic assessment of COVID-19 host genetics using whole genome sequencing data”

Year: 2024
Journal: PLOS Pathogens
DOI: https://doi.org/10.1371/journal.ppat.1012786

PLOS PATHOGENS

RESEARCH ARTICLE

Systematic assessment of COVID-19 host genetics using whole genome sequencing data

Axel Schmidt[1,2], Nicolas Casadei[3,4], Fabian Brand[5], German Demidov[4,6], Elaheh Vojgani[7], Ayda Abolhassani[8], Rana Aldisi[5], Guillaume Butler-Laporte[9,10], DeCOI host genetics group[¶], T. Madhusankha Alawathurage[1], Max Augustin[11,12,13], Robert Bals[14,15], Carla Bellinghausen[16], Marc Moritz Berger[17], Michael Bitzer[18,19], Christian Bode[20], Jannik Boos[1], Thorsten Brenner[17], Oliver A. Cornely[11,12,13,21,22], Thomas Eggermann[23], Johanna Erber[24], Torsten Feldt[25], Christian Fuchsberger[26], Julien Gagneur[27,28,29], Siri Göpel[19,30], Tobias Haack[4], Helene Häberle[31], Frank Hanses[32,33], Julia Heggemann[1], Ute Hehr[34], Johannes C. Hellmuth[35,36], Christian Herr[14], Anke Hinney[37], Per Hoffmann[1], Thomas Illig[38], Björn-Erik Ole Jensen[25], Verena Keitel[25], Sarah Kim-Hellmuth[39,40], Philipp Koehler[11,12,22], Ingo Kurth[23], Anna-Lisa Lanz[39], Eicke Latz[41], Clara Lehmann[11,12,13], Tom Luedde[25], Carlo Maj[42], Michael Mian[43], Abigail Miller[1], Maximilian Muenchhoff[35,44], Isabell Pink[45], Ulrike Protzer[46,47], Hana Rohn[48], Jan Rybniker[11,12,13], Federica Scaggiante[49], Anna Schaffeldt[23], Clemens Scherer[35,50], Maximilian Schieck[38], Susanne V. Schmidt[41], Philipp Schommers[11,12,13], Christoph D. Spinner[24,46], Maria J. G. T. Vehreschild[51], Thirumalaisamy P. Velavan[52,53], Sonja Volland[38], Sibylle Wilfling[34,54], Christof Winter[55,56,57,58], J. Brent Richards[9,59,60,61,62,63], DeCOI[¶], André Heimbach[1,64], Kerstin Becker[7,65], Stephan Ossowski[4,6], Joachim L. Schultze[66,67,68], Peter Nürnberg[7], Markus M. Nöthen[1], Susanne Motameny[7,65], Michael Nothnagel[7], Olaf Riess[3,4], Eva C. Schulte[1,8,47,69,70☯], Kerstin U. Ludwig[1☯*]

1 Institute of Human Genetics, School of Medicine, University Bonn & University Hospital Bonn, Bonn, Germany, **2** Department of Pediatric Neurology, School of Medicine, University Bonn & University Hospital Bonn, Bonn, Germany, **3** DFG NGS Competence Center Tübingen (NCCT), University of Tübingen, Tübingen, Germany, **4** Institute of Medical Genetics and Applied Genomics, University of Tübingen, Tübingen, Germany, **5** Institute of Genomic Statistics and Bioinformatics, School of Medicine, University Bonn & University Hospital Bonn, Bonn, Germany, **6** Institute for Bioinformatics and Medical Informatics (IBMI), University of Tübingen, Tübingen, Germany, **7** Cologne Center for Genomics (CCG), University of Cologne, Cologne, Germany, **8** Department of Psychiatry and Psychotherapy, School of Medicine, University Bonn & University Hospital Bonn, Bonn, Germany, **9** Lady Davis Institute, Jewish General Hospital, McGill University, Montréal, Québec, Canada, **10** Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom, **11** Center for Molecular Medicine Cologne (CMMC), University of Cologne, Cologne, Germany, **12** Department I of Internal Medicine, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, **13** German Center for Infection Research (DZIF), Partner Site Bonn-Cologne, Cologne, Germany, **14** Department of Internal Medicine V, Saarland University, Homburg, Germany, **15** Helmholtz Institute for Pharmaceutical Research Saarland (HIPS), Saarbrücken, Germany, **16** Department of Internal Medicine, Pneumology, University Hospital Frankfurt, Goethe University, Frankfurt am Main, Germany, **17** Department of Anesthesiology and Intensive Care Medicine, University Hospital Essen, University Duisburg-Essen, Essen, Germany, **18** Center for Personalized Medicine, University Hospital Tübingen, Tübingen, Germany, **19** Department of Internal Medicine I, University Hospital Tübingen, Tübingen, Germany, **20** Department of Anesthesiology and Intensive Care Medicine, University Hospital Bonn, Bonn, Germany, **21** Clinical Trials Center Cologne, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, **22** Institute of Translational Research, Cologne Excellence Cluster on Cellular Stress Responses in Aging-Associated Diseases (CECAD), Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, **23** Institute for Human Genetics and Genomic Medicine, Medical Faculty, RWTH Aachen University, Aachen, Germany, **24** Department of Internal Medicine II, University Hospital rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany, **25** Department of Gastroenterology, Hepatology and Infectious Diseases, University Hospital Duesseldorf, Medical Faculty, Düsseldorf, Germany, **26** Eurac Research, Institute for Biomedicine, Bolzano, Italy, **27** Computational Health Center, Helmholtz Zentrum

OPEN ACCESS

Citation: Schmidt A, Casadei N, Brand F, Demidov G, Vojgani E, Abolhassani A, et al. (2024) Systematic assessment of COVID-19 host genetics using whole genome sequencing data. PLoS Pathog 20(12): e1012786. https://doi.org/10.1371/journal.ppat.1012786

Editor: Helen Su, NIAID: National Institute of Allergy and Infectious Diseases, UNITED STATES OF AMERICA

Received: February 17, 2024

Accepted: November 27, 2024

Published: December 23, 2024

Data Availability Statement: All data supporting the findings of this study are available either within the article, supplementary data files, or have been deposited in public resources. Specifically, single variant association studies summary statistics were deposited in the GWAS catalog (https://www.ebi.ac.uk/gwas/; accession numbers: GCST90435134 - GCST90435139), and gene-burden results are available in zenodo (https://doi.org/10.5281/zenodo.12625864). Participant level data from each participating cohort may be

accessed according to the rules of each cohort's data sharing policy. Requests for access to participant level data can be directed to dac_decoi_hostgenetics@listen.uni-bonn.de. Additionally, contact information of each study is provided in S1 Table. Code availability: Code used for the analysis of the WGS data is available at GitHub (https://github.com/Ax-Sch/DeCOI_WGS).

Funding: Genome sequencing was supported by institutional grants from the German Research Foundation (Deutsche Forschungsgemeinschaft - DFG) (286/2020B01 - 428994620), and was performed by the DFG-funded NGS Competence Center Tübingen (NCCT; INST 37/1049-1) and West German Genome Center (WGGC; INST 216/981-1). The study was supported in part by an unrestricted grant from Illumina, Berlin. The following investigators were financially supported: ASchm (BONFOR program of the Medical Faculty of the University of Bonn (O-149.0134); RA (DFG (428902522)), SKH (Emmy Noether Programme of the DFG (KI 2091/2-1), other DFG grants (459153572), SFB/TRR237-B29 (369799452), ERC Starting Grant (101076303)), AH (Stiftung Universitätsmedizin Essen); ECS (DFG (Schu 2914/2-1). ECS was further supported by the Munich Clinician Scientist Program (MCSP). JLS, MMN, and KUL are members of the DFG-funded Cluster of Excellence ImmunoSensation - EXC2151 – 390873048. PK reports grants or contracts from German Federal Ministry of Research and Education (BMBF), B-FAST (Bundesweites Forschungsnetz Angewandte Surveillance und Testung) and German National Pandemic Cohort Network (Nationales Pandemie Kohorten Netz - NAPKON) of the Network University Medicine (Netzwerk-Universitätsmedizin - NUM) and the State of North Rhine-Westphalia. Recruitment of participating cohorts was funded by institutional support of: the Technical University of Munich (COMRI cohort), the Institute of Human Genetics, University Hospital Bonn (BoSCO cohort), the Federal Ministry of Education and Research (NUM - COVIM: 01KX2021, ReCOV cohort), the Rolf M. Schweite Stiftung and the State of Saarland (2020-013; both CORSAAR cohort), the Uniklinik RWTH Aachen and the Institute for Human Genetics and Genomic Medicine at the University Hospital Aachen (COVAS cohort), the UME and the Stiftung Universitätsmedizin Essen (COVES cohort), Hessisches Ministerium für Wissenschaft und Kunst (CCHROMO cohort), the Healthcare System of the Autonomous Province of Bolzano/Bozen (Val Gardena cohort), the University Hospital Düsseldorf (COVID-19 UKD Biobank study), the Ministry of Science and Culture (MWK) of Lower-Saxony (COVID-19 MWK Biobank study), the Free

München, Neuherberg, Germany, **28** Institute of Human Genetics, School of Medicine, Technical University of Munich, Munich, Germany, **29** School of Computation, Information and Technology, Technical University of Munich, Garching, Germany, **30** German Center for Infection Research (DZIF), Partner Site Tübingen, Tübingen, Germany, **31** Department of Anesthesiology and Intensive Care Medicine, University Hospital Tübingen, Tübingen, Germany, **32** Department for Infection Control and Infectious Diseases, University Hospital Regensburg, Regensburg, Germany, **33** Emergency Department, University Hospital Regensburg, Regensburg, Germany, **34** Center for Human Genetics Regensburg, Regensburg, Germany, **35** COVID-19 Registry of the LMU Munich (CORKUM), University Hospital, LMU Munich, Munich, Germany, **36** Department of Medicine III, University Hospital, LMU Munich, Munich, Germany, **37** Department of Child and Adolescent Psychiatry and Psychotherapy, University Hospital Essen, University of Duisburg-Essen, Essen, Germany, **38** Hannover Unified Biobank, Hannover Medical School, Hannover, Germany, **39** Department of Pediatrics, Dr. von Hauner Children's Hospital, University Hospital LMU Munich, Munich, Germany, **40** Institute of Translational Genomics, Helmholtz Munich, Neuherberg, Germany, **41** Institute of Innate Immunity, University Hospital Bonn, Bonn, Germany, **42** Center for Human Genetics, Philipps University of Marburg, Marburg, Germany, **43** Service for Innovation, Research and Teaching, (SABES-ASDAA), Bolzano-Bozen, Italy; Teaching Hospital of Paracelsus Medical University, **44** Max von Pettenkofer Institute and Gene Center, Virology, National Reference Center for Retroviruses, LMU Munich, Munich, Germany, **45** Department of Pneumology, Hannover Medical School, Hannover, Germany, **46** German Center for Infection research (DZIF), Partner Site Munich, Munich, Germany, **47** Institute of Virology, Technical University Munich/Helmholtz Munich, Munich, Germany, **48** Department of Infectious Diseases, West German Centre of Infectious Diseases, University Hospital Essen, University Duisburg-Essen, Essen, Germany, **49** Laboratorio di Patologia Clinica di Bressanone, Hospital of Bressanone (SABES-ASDAA), Bressanone-Brixen, Italy; Teaching Hospital of Paracelsus Medical University, **50** Department of Medicine I, University Hospital, LMU Munich, Munich, Germany, **51** Department of Internal Medicine, Infectious Diseases, University Hospital Frankfurt, Goethe University, Frankfurt am Main, Germany, **52** Institute of Tropical Medicine, Universitätsklinikum Tübingen, Tübingen, Germany, **53** Vietnamese-German Center for Medical Research (VG-CARE), Hanoi, Vietnam, **54** Department of Neurology, Bezirksklinikum Regensburg, University of Regensburg, Regensburg, Germany, **55** German Cancer Consortium (DKTK), Partner Site Munich, Munich, Germany, **56** German Cancer Research Center (DKFZ), Heidelberg, Germany, **57** Institute of Clinical Chemistry and Pathobiochemistry, Klinikum Rechts der Isar, School of Medicine, Technical University of Munich, Munich, Germany, **58** TranslaTUM, Center for Translational Cancer Research, Technical University of Munich, Munich, Germany, **59** 5 Prime Sciences Inc, Montréal, Québec, Canada, **60** Department of Epidemiology, Biostatistics and Occupational Health, McGill University, Montréal, Québec, Canada, **61** Department of Human Genetics, McGill University, Montréal, Québec, Canada, **62** Department of Twin Research, King's College London, London, United Kingdom, **63** Infectious Diseases and Immunity in Global Health Program, Research Institute of the McGill University Health Centre, Montréal, Québec, Canada, **64** NGS Core Facility Bonn, University of Bonn, School of Medicine & University Hospital Bonn, Bonn, Germany, **65** West German Genome Center - Cologne, University of Cologne, Cologne, Germany, **66** Genomics and Immunoregulation, Life & Medical Sciences (LIMES) Institute, University of Bonn, Bonn, Germany, **67** PRECISE Platform for Genomics and Epigenomics, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE) e.V. and University of Bonn, Bonn, Germany, **68** Systems Medicine, Deutsches Zentrum für Neurodegenerative Erkrankungen (DZNE) e.V., Bonn, Germany, **69** Department of Psychiatry & Psychotherapy, University of Munich, Munich, Germany, **70** Institute of Psychiatric Phenomics and Genomics, University of Munich, Munich, Germany

☯ These authors contributed equally to this work.
¶ Membership of DeCOI host genetics group is provided in Supporting Information file S15 Table.
¶ Membership of DeCOI is provided in Supporting Information file S16 Table.
* kerstin.ludwig@uni-bonn.de

Abstract

Courses of SARS-CoV-2 infections are highly variable, ranging from asymptomatic to lethal COVID-19. Though research has shown that host genetic factors contribute to this variability, cohort-based joint analyses of variants from the entire allelic spectrum in individuals with confirmed SARS-CoV-2 infections are still lacking. Here, we present the results of whole genome sequencing in 1,220 mainly vaccine-naïve individuals with confirmed SARS-CoV-2 infection, including 827 hospitalized COVID-19 cases. We observed the presence of

State of Bavaria under the Excellence Strategy of the Federal State Government (LMUExcellent, CORKUM cohort), the Care-for-Rare Foundation (Ped-COVID-19 cohort), the BMBF (01KI20343; COVIMMUNE cohort), and the Bavarian State Ministry for Science and Art (COVUR cohort). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: JBR is CEO and founder of the company 5 Prime Sciences Inc, which offers analyses of human genetic data for the selection of drug targets. PK received grants or personal fees from Ambu GmbH, Gilead Sciences, Mundipharma Resarch Limited, Noxxon N.V. and Pfizer Pharma; he received honoraria for lectures from Akademie für Infektionsmedizin e.V., Ambu GmbH, Astellas Pharma, BioRad Laboratories Inc., Datamed GmbH, European Confederation of Medical Mycology, Gilead Sciences, GPR Academy Ruesselsheim, HELIOS Kliniken GmbH, Lahn-Dill-Kliniken GmbH, medupdate GmbH, MedMedia GmbH, MSD Sharp & Dohme GmbH, Pfizer Pharma GmbH, Scilink Comunicación Científica SC, streamedup! GmbH and University Hospital and LMU Munich and he participates in an Advisory Board from Ambu GmbH, Gilead Sciences, Mundipharma Resarch Limited and Pfizer Pharma. CDS received grants or personal fees from AstraZeneca, BBraun Melsungen, BioNtech, Gilead Sciences, Janssen-Cilag, Eli Lilly, Formycon, Pfizer, Roche, Apeiron, MSD, Cepheid, GSK, Molecular partners, SOBI, AbbVie, Synairgen, Shionogi and ViiV Healthcare. None of the funders did have any influence on study design or execution.

autosomal-recessive or likely compound heterozygous monogenic disorders in six individuals, all of which were hospitalized and significantly younger than the rest of the cohort. We did not observe any suggestive causal variants in or around the established risk gene *TLR7*. Burden testing in the largest population subgroup (i.e., Europeans) suggested nominal enrichments of rare variants in coding and non-coding regions of interferon immune response genes in the overall analysis and male subgroup. Case-control analyses of more common variants confirmed associations with previously reported risk loci, with the key locus at 3p21 reaching genome-wide significance. Polygenic scores accurately captured risk in an age-dependent manner. By enabling joint analyses of different types of variation across the entire frequency spectrum, this data will continue to contribute to the elucidation of COVID-19 etiology.

Author summary

After infection with SARS-CoV-2, symptoms vary widely. On average, individuals who are older, males and those with certain comorbidities tend to be more severely affected by COVID-19. Additionally, genetics of the infected individuals (host genetics) modulate the severity of symptoms, but so far, most studies on COVID-19 host genetics have focused either on common or on rare variants, but not both. In this study, we analyzed genetic variants comprehensively by whole genome sequencing of 1,220 SARS-CoV-2 positive individuals with varying degrees of COVID-19 severity. In our cohort, we replicate several associations between common variants and COVID-19 severity, with a region on chromosome 3 showing the largest effect size. We additionally show that common variants, taken together, can help to predict COVID-19 severity, particularly in individuals younger than 60 years. We also identified six individuals with moderate or severe COVID-19 who had underlying rare genetic diseases, which creates interesting new hypotheses. Finally, we observed an enrichment of rare variants in immune pathways in severe or moderate COVID-19. This study provides comprehensive novel insights into COVID-19 host genetics.

Introduction

Since late 2019, severe acute respiratory syndrome coronavirus type 2 (SARS-CoV-2) has infected hundreds of millions of people worldwide. SARS-CoV-2 infections are clinically heterogeneous and can remain asymptomatic or become symptomatic, the latter being referred to as Coronavirus Disease 2019 (COVID-19). COVID-19 mainly affects the respiratory tract and can lead to severe pneumonia, but other organ systems may also be affected. Research has shown that the clinical heterogeneity of COVID-19 can be explained in part by demographic factors (e.g., advanced age and male sex [1]), and the presence of predisposing medical conditions [2] or auto-antibodies [3]. In addition, epidemiological data have implicated host genetic factors [4].

Through the work of large global consortia, such as the COVID-19 Host Genetics Initiative (COVID-19 HGI) [5], the analyses of data from biobanks, and individual clinical studies, multiple host genetic loci that contribute to an individual's risk for severe disease secondary to SARS-CoV-2 infection have now been identified [6]. Specifically, genome-wide association

studies (GWAS) have highlighted at least 71 loci at which common variants contribute to infection susceptibility or COVID-19 severity [7–10]. These efforts have been complemented by whole exome sequencing (WES) studies of severely affected individuals, which have led to the identification of rare loss-of-function (LoF) variants in genes involved in the innate immune response [11,12], some of which are known inborn errors of immunity or have subsequently been classified as such [13]. At the time of writing, the COVID-19 risk gene with the most compelling evidence in terms of rare variants is the X-chromosomal toll-like receptor 7 gene (*TLR7*), for which LoF variants were initially detected in two pairs of previously healthy young (aged 21–32 years) brothers with severe to fatal disease [14]. Subsequent candidate gene-, machine learning-, and WES-based rare variant association approaches have generated independent support for the role of *TLR7* in severe COVID-19 in males [15–18], with recent estimates suggesting the presence of a TLR7 deficiency in around 1–2% of male cases [15,19]. Besides *TLR7*, additional candidate genes have been suggested, e.g. 13 genes of the type I interferon (IFN) immunity [11,12,20].

To date, most investigations of host genetic factors in SARS-CoV-2 infections have analyzed either common variants (mainly through genome-wide array-based genotyping followed by imputation) [7,8,10,21–27] or rare variants in protein-coding regions (mostly through WES in either clinical cohorts [11,15,17,18,20,28,29]; or families [14,30,31]). However, these approaches fail to cover a substantial fraction of the total genetic variability (such as rare variants in non-coding regions), and are rarely combined on the same individual genomes, thereby precluding joint analyses of variants along the entire allelic spectrum. These issues can be resolved via whole genome sequencing (WGS). To date, however, WGS has rarely been applied in this field because of its relatively high costs and its full potential in COVID-19 has yet to be explored [22,29].

By building on the German COVID-19 Omics Initiative (DeCOI) [32], we established a national consortium to investigate the host genetics of COVID-19 (S1 Table). WGS data of 1,220 individuals with reported SARS-CoV-2 infection and variable disease outcomes were used to characterize genetic risk factors related to COVID-19 severity. We investigated the presence of: (i) potentially causal rare variants within the *TLR7* locus, including adjacent non-coding regions, and in additional 13 candidate genes; (ii) monogenic conditions that might increase the risk for severe COVID-19; and (iii) immune-relevant gene sets (in both coding and non-coding regions) that are enriched for functionally-relevant rare variation. Furthermore, we investigated the polygenic architecture of severe COVID-19 in age-stratified groups. These analyses comprehensively characterize the joint contribution of variants of the entire allelic spectrum to severe COVID-19.

Results

The DeCOI cohort

Following quality control (see Methods), the DeCOI cohort comprised 1,220 individuals from across the entire phenotypic spectrum of SARS-CoV-2 infections (Figs 1A–1C and S1). The average age of the cohort was 56.2 years (range: 1–100 years), and 490 participants were female (40.2%). Based on the available phenotypic information, 393 individuals were classified as having had mild SARS-CoV-2 infections ("ambulatory mild", World Health Organization ordinal scale for COVID-19 severity (WHO score, [33]) 1–3), 482 individuals were classified as having been hospitalized without the need for high-flow oxygen or mechanical ventilation ("hospitalized moderate", WHO 4–5), and 345 individuals were classified as having either required at least high-flow oxygen or mechanical ventilation, or having had lethal COVID-19 ("hospitalized severe", WHO 6–10). Consistent with available epidemiological evidence, both the

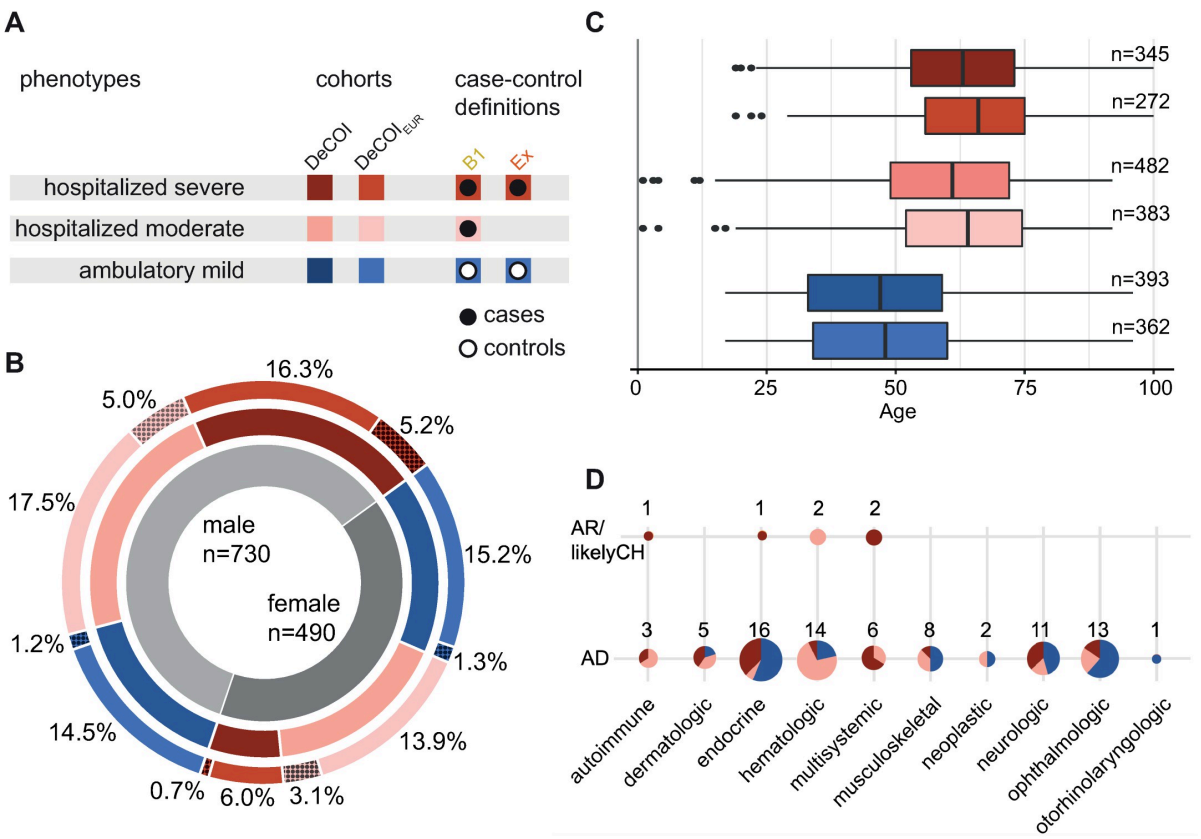


Fig 1. The DeCOI and the DeCOI$_{EUR}$ cohort. (A) Individuals in the DeCOI cohort are classified into three phenotypes based on WHO definition. In addition, the cohort was subsetted to an unrelated cohort of the European population (DeCOI$_{EUR}$) for association analyses. Based on the phenotypes, case-control definitions were established within DeCOI$_{EUR}$. (B) Composition of the DeCOI cohort according to sex (inner circle), phenotype (color coded, middle circle), and population (outer circle). Shaded intervals in the outer circle represent non-European individuals. (C) Age distribution of individuals from the DeCOI cohort (n = 1,220) and the European subcohort (DeCOI$_{EUR}$; n = 1,017), as stratified according to severity (color coded). In both subcohorts, the average age increases with disease course severity. Numbers indicate individuals in the respective group. (D) Phenotype distribution of individuals harboring ClinVar-annotated variants, as grouped according to disorder class. Autosomal recessive patterns of inheritance (AR/likely compound-heterozygous (CH), n = 6 diseases in six individuals) are displayed in the upper panel, and autosomal dominant inheritance patterns (AD, n = 79 diseases in 77 individuals) are displayed in the lower panel.

https://doi.org/10.1371/journal.ppat.1012786.g001

average age and the proportion of male individuals increased with increasing COVID-19 severity (Fig 1B and 1C and S2 Table).

The European subcohort, DeCOI$_{EUR}$, comprised 1,017 individuals (WHO 1–3: n = 362; WHO 4–5: n = 383; WHO 6–10: n = 272, S2 Fig). Again, the average age and proportion of male individuals increased with COVID-19 severity (Fig 1B and 1C and S2 Table). For association analyses in DeCOI$_{EUR}$, we created two case-control definitions: (i) "extreme" (Ex / cases: hospitalized severe, n = 272 / controls: ambulatory mild, n = 362), and (ii) "all_hospitalized" (B1 / cases: hospitalized moderate and hospitalized severe, n = 655 / controls: ambulatory mild, n = 362), with B1 being in accordance with the case control definition of the COVID-19 HGI and Ex representing the analysis along the phenotypic extremes.

Targeted analysis of variants at the *TLR7* locus

Given that some monogenic disorders are likely to impact the course of COVID-19 disease [34], the multi-ethnic DeCOI cohort was analyzed for the presence of known monogenic diseases. We first queried for variants that may cause *TLR7* deficiency, since at the time of writing, this represents the most robustly established monogenic cause of severe COVID-19, particularly in young men [14,15,19]. Within the coding sequence of *TLR7*, three known variants were identified (S5 Table). Each of these variants had low REVEL/CADD scores. Carriers were observed in all phenotypic categories, which is consistent with the normal functional characteristics of these three variants, as described elsewhere [15]. Within non-coding regions with evidence for regulatory function (see Methods), 23 variants with an MAF $< 1\%$ were identified across all phenotypic groups (S5 Table). The most notable variant was rs192357402, which was observed in 3/199 severely affected males of European-ancestry but was not detected in 391 males of European-ancestry with non-severe disease ($p = 0.038$, Fisher's exact test). This finding was not replicated in 672 males of European ancestry in an independent dataset from the Biobank Quebec COVID-19 Cohort (Methods, 1/113 severe vs. 2/559 non-severe; $p = 0.42$, S6 Table). Based on coverage data in the DeCOI cohort VCF, a search was also conducted in males for evidence of deletions within a region spanning approximately 200kb centered around *TLR7*. While 57 individuals were found to have short stretches of missing coverage, visual inspection provided no evidence that these were true deletions.

Analysis of 13 genes previously implicated in severe COVID-19

Previously, deleterious variants in 13 genes of the type I interferon (IFN) immunity were implicated in life-threatening COVID-19 pneumonia [11]. We queried these genes for variants predicted to be loss-of-function (pLoF), as well as for missense variants previously demonstrated to be LoF or strongly hypomorphic (see Methods). Six heterozygous pLoF variants in the genes *UNC93B1*, *IRF7*, *IRF3*, *IFNAR1* and *IFNAR2* and two heterozygous missense variants in *IRF3* and *IRF7* (S7 Table) were identified. Interestingly, one moderately affected male aged 25–34 years carried two of these variants (IFNAR2/pLoF and IRF3/missense). The carriers of these variants were 46.1±15.8 years old on average ($p = 0.13$, Student's t-test, comparison against the remainder of the DeCOI cohort), three of the seven individuals were female. Only one individual was severely affected, three were moderately and three were mildly affected, which indicates that the phenotype of these individuals is not more severe than expected by chance (expected number of individuals by random chance: 2.0 severe, 2.8 moderate and 2.2 mild). No homozygous or potentially compound heterozygous variants that passed our filter criteria were identified. Systematic testing for joint association of variants within the 13 genes of the type I IFN immunity can be found below.

Targeted analysis of monogenic disorders

Next, the DeCOI cohort was queried for the presence of established causes of monogenic diseases, as based on variants reported in ClinVar. Autosomal-recessive (AR), autosomal-dominant (AD) and X-linked (XL) patterns of inheritance were considered (see Methods). Established homozygous variants causing monogenic disorders were found in 4 out of 1,220 individuals, and likely compound-heterozygous variants were identified in two individuals (jointly 0.5%, Table 1). All six individuals were male and hospitalized (3/6 with a fatal course). Notably, the six individuals were significantly younger on average than the remainder of the DeCOI cohort (mean±SD = 38±14.5yrs; $p = 0.027$, Student's t-test; S3 Fig). Heterozygous variants with established associations to dominantly inherited monogenic diseases, and that are annotated as "pathogenic" or "likely pathogenic" in ClinVar, were present in 77 out of 1,220

Table 1. Characteristics of carriers of homozygous or likely compound heterozygous disease variants in the DeCOI cohort.

Gene	Variant / Genotype	Monogenic disease	Sex, age range	COVID-19 severity	Monogenic disease previously reported	Additional information	Population background
BBS1	Homozygous splice variant: chr11_66523577_G_A; c.951+1G>A;p.?	Bardet-Biedl syndrome 1	Male, 35–44 years	Fatal (WHO 10)	no	Clinically intellectual development disorder, blindness, and seizures	AMR
AGXT	Homozygous frameshift variant: chr2_240868890_A_AC; p.Lys12GlnfsTer156	Primary Hyperoxaluria Type 1	Male, 25–34 years	Fatal (WHO 10)	yes	Post renal and liver transplant status (no details available concerning immunosuppressive therapy)	EUR
SERPIN1C	Homozygous missense variant: chr1_173914743_G_A; p.Pro73Leu	Antithrombin Budapest 3	Male, 35–44 years	Moderate (WHO 4)	not available	-	SAS
AIRE	Homozygous nonsense variant: chr21_44289773_C_T; p.Arg257Ter	Polyglandular autoimmune syndrome	Male, 15–24 years	Severe (WHO 6)	yes	-	AMR
HBB	Likely compound heterozygous variants: Intron variant chr11_5225832_G_C; NM_000518.5:c.316-106C>G Nonsense variant chr11_5226774_G_A; p.Gln40Ter	Beta-thalassemia major	Male, 35–44 years	Moderate (WHO 4)	not available	-	EUR
PAH	Likely compound heterozygous variants: Missense variant chr12_102843676_T_C; p.Glu390Gly Missense variant chr12_102855313_C_G; p.Val177Leu	Mild Phenylketonuria (PKU)	Male, 55–64 years	Fatal (WHO 10)	no	Heart disease and diabetes mellitus type 2	AMR

Abbreviations: AMR: Admixed American; EUR: European; SAS: South Asian. Note that the genomic position is given in GRCh38 coordinates.

https://doi.org/10.1371/journal.ppat.1012786.t001

DeCOI individuals (6.4%). The associated diseases covered a broad range of categories, with endocrine, hematologic, and ophthalmologic disorders being the most commonly represented (Fig 1D). Overall, carriers of heterozygous (likely) pathogenic ClinVar variants did not differ significantly from the rest of the DeCOI cohort with respect to sex, age, or severity of COVID-19 (S3 Fig). No hemizygous or homozygous variants on the X-chromosome were identified that are annotated as "pathogenic" or "likely pathogenic" in ClinVar.

Gene- and gene-set-based collapsing analyses

Next, the analyses were expanded to study joint effects of rare variants across: (i) single genes, and (ii) sets of genes with presumed importance to COVID-19 (see Methods, S3 and S4 Tables). For this purpose, variants were selected on the basis of allele frequency and predicted functional effect, and all variants were collapsed across a gene or a gene-set. Association testing was then performed with logistic regression, including polygenic score based on common variants as one covariate in addition to principal components (PC) and age-/sex-derived measures (see Methods for more details). Results of the gene-based collapsing analyses are shown in S8 Table for analysis Ex, and S9 Table for analysis B1. Some nominally significant results were observed. However, these did not withstand correction for multiple testing, and their number was not larger than would be expected by chance (S4 Fig).

The gene-set analyses were performed on the case-control definitions Ex and B1 overall, and then as stratified according to sex (male/female) and age (younger than 60 years/older or equal 60 years). In total, 14 nominally significant phenotype / gene-set / mask combinations were identified, all of which were observed in either the overall phenotypes (Ex_all/B1_all) or the male subcohort (Ex_male; B1_male; Fig 2 and S10 Table). None of the other stratifications (female or age-stratified) yielded any significant enrichment. Nominally, the most significant enrichment was found among severe COVID-19 patients in genes of the innate immune system, for the functional masks (FM) that included predicted loss-of-function (pLoF) (B1_all: $p = 5.85x10^{-03}$; beta = 0.27, SE = 0.099) and pLoF+missense (Ex_all: $p = 7.04x10^{-03}$; beta = 0.11, SE = 0.042). Among the non-coding variants, a nominally significant depletion of 3'UTR variants with high CADD scores (CADD≥10) was observed in both gene sets related to IFN-response (Ex_male/IFN_response_COVID-19/UTR3_CADD: p = 0.019; n = 31 genes), and the subset of 13 genes with *a priori* evidence for an involvement in severe COVID-19 (Ex_all/Zhang et al./UTR3_CADD: p = 0.029). In the gene-based analyses that did not include individual PRS as a covariate, highly correlated results were generated (S4 Fig).

Single variant association analyses

After analyzing lower frequency variants, we next investigated more common variants. Using WGS genotype calls, GWAS were performed for phenotypes Ex and B1, respectively (Figs 3 and S5). Interestingly, despite the relatively low sample size of the Ex case-control definition, association reached genome-wide significance for variants at the established key risk locus

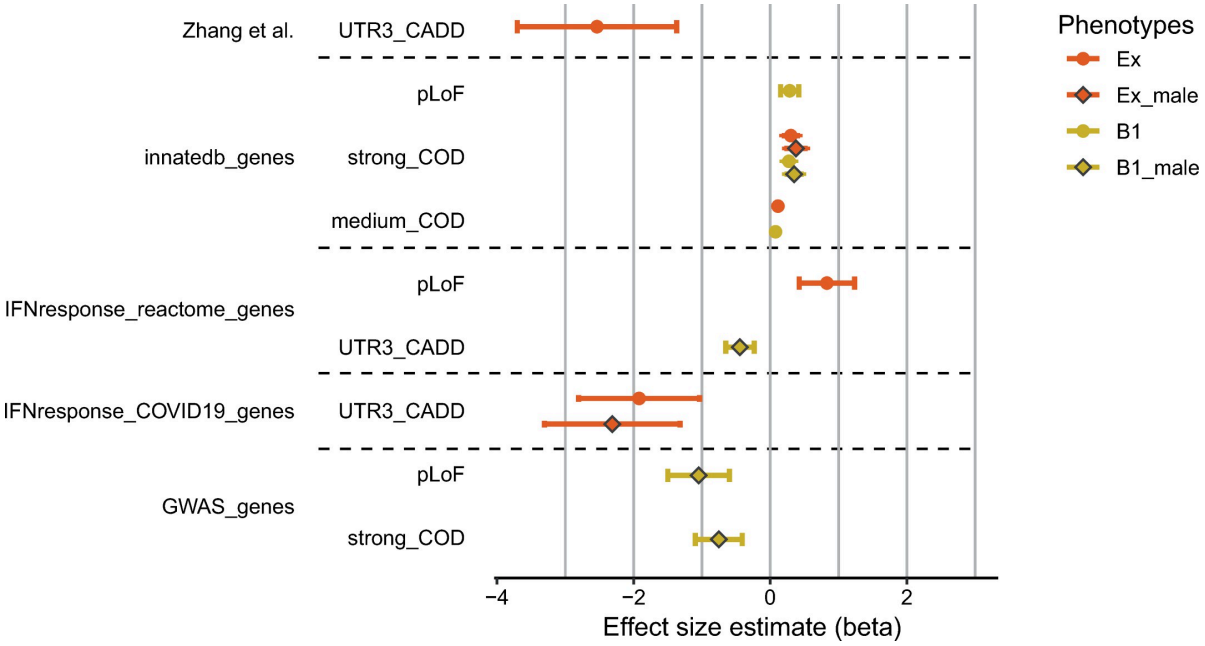


Fig 2. Effect sizes of nominally significant gene-set based tests in the $DeCOI_{EUR}$ cohort. Gene-sets and the corresponding functional masks (S4 Table) that were tested are given on the y-axis. On the x-axis, effect size estimates (betas) are shown as markers with error bars indicating the standard errors of betas. Note that phenotypes are color-coded, and the markers outlined in black indicate analyses that only included males. Nominally significant findings were only obtained in the overall analyses and male sub-stratification. None was observed in female-only or age-stratified analyses. A list of genes that were included in each gene-set can be found in S3 Table.

https://doi.org/10.1371/journal.ppat.1012786.g002

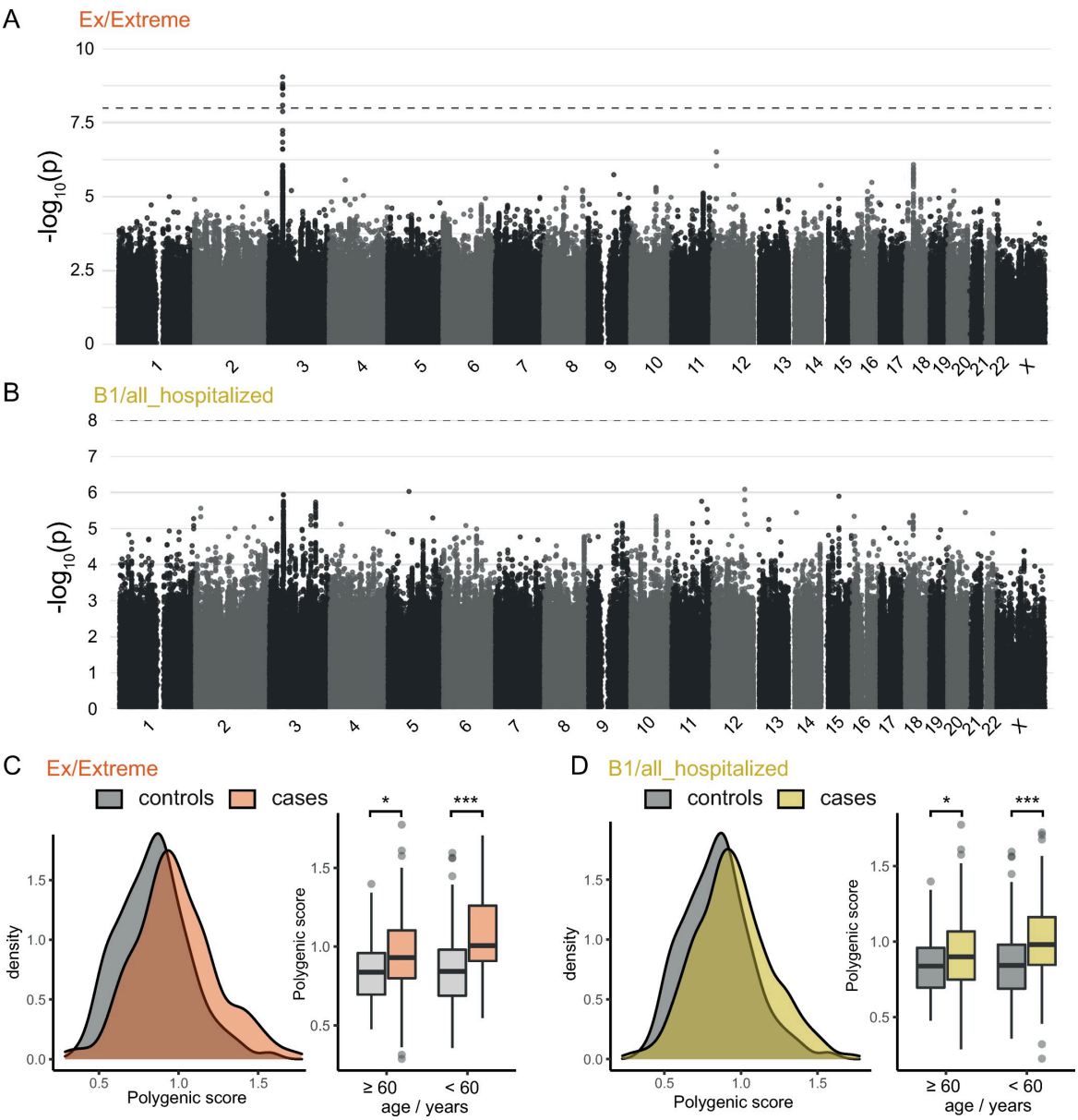


Fig 3. Analysis of common variants within the DeCOI$_{EUR}$ cohort. (A) and (B): Manhattan plots of association analyses of single variants (MAF>0.5%) in DeCOI$_{EUR}$ (n = 1,017 individuals), for phenotype Ex (272 severely affected individuals vs. 362 mild controls) and B1 (655 hospitalized individuals vs. 362 non-hospitalized controls), respectively. Genomic inflation factors were 1.04 (Ex) and 1.00 (B1). Among the strongest associations is the well-established risk locus at 3p21.31. Panels (C) and (D) show the distribution of individual polygenic risk scores (PRS) among cases (orange or yellow) and controls (gray) of Ex (C) or B1 (D) overall (density plots in the left parts) or when stratified according to age below or above 60 years (box plots in the right parts). The elements of the box plots correspond to the following values: thick line: median, box: 25th and 75th percentile, whiskers: largest / smallest value not further away from the box than 1.5 times the interquartile range, points: values outside of the range of the whiskers. *: $p<0.05$, ***: $p<0.001$; Wald test followed by Bonferroni correction. MAF: Minor Allele Frequency.

https://doi.org/10.1371/journal.ppat.1012786.g003

3p21.31. In analysis Ex, 177 variants with $p<1x10^{-05}$ were observed at 19 loci, the majority of which (n = 128) mapped to the 3p21.31 region (S11 Table). The variant with the strongest evidence of association was rs73064425 (chr3:45859597:C:T, $p = 9.00x10^{-10}$; beta = 1.44, SE = 0.23). In Europeans, this variant is in perfect LD with all previously reported lead variants (i.e., rs11385942 [9], rs10490770 [35], and rs35044562 [36]). No additional support for any of the 49 variants outside 3p21.31 was found in data from the WGS-based summary statistics from GenOMICC [22] or the array-based GWAS of the COVID-19 HGI (release 7, without GenOMICC [10], S11 Table). At established risk loci for SARS-CoV-2 related traits (n = 71) [7–10], nominal significance was observed for the reported lead variants at 11 loci (Tables 2 and S12), whereby a minor overlap of samples between COVID-HGI and DeCOI (<0.04%) must be kept in mind. No significant association was found for two variants that were reported to be associated with severe COVID-19 in previous independent German cohorts (i.e., rs5443 (p = 0.72 (Ex) and p = 0.14 (B1)); and rs5010528 (p = 0.41 (Ex) and p = 0.77 (B1))) [37,38]. Finally, the $DeCOI_{EUR}$ cohort was stratified according to age or sex, and the better-powered Ex analysis was repeated for different substrata. No variants in any of the stratified analyses reached genome-wide significance (S6 Fig and S13 Table).

Autozygosity

To investigate a possible effect of autozygosity on disease severity, inbreeding coefficients were calculated as a measure for autozygosity within the $DeCOI_{EUR}$ cohort, with no prior filtering of variant frequency. For phenotype Ex, no significant differences in autozygosity levels were observed. Significantly increased inbreeding coefficients were observed in cases of phenotype B1 (cases: mean±sd: 0.002±0.01; controls: 0.001±0.005; p = 0.023, one-sided Wilcoxon test; S7 Fig). This result was mainly driven by a small subset of individuals with inbreeding coefficients above 0.02 (FI>0.02: 3.51% in cases, 0.83% in controls; FI>0.05: 0.76% vs. 0.15%; FI>0.1: 0.55% vs. 0.0%), who largely overlapped with samples that were located outside of the central European-ancestry cluster on the PC plot (S8 Fig). When the first 10 PCs were added

Table 2. Previously reported risk loci for COVID-19 with nominal significance in $DeCOI_{EUR}$.

Chr	Pos	ID	Ref/Alt	Extreme (272 cases[a], 362 controls)[1]			All_Hospitalized COVID-19 (655 cases[b], 362 controls)			Candidate gene(s)	Ref (PMID)
				beta	SE	P-value	beta	SE	P-value		
1	9067157	rs2478868	A/C	**0.34**	**0.15**	**0.025**	**0.36**	**0.12**	**0.0021**	SLC2A5	37198478
1	77501822	rs71658797	T/A	0.51	0.26	0.050	**0.59**	**0.20**	**0.0041**	AK5	37198478
1	155197995	rs41264915	A/G	**-0.78**	**0.25**	**0.0015**	**-0.75**	**0.19**	**0.00011**	THBS3, MUC1	35922517
3	45818159	rs17713054	G/A	**1.39**	**0.23**	**0.0000000021**	**0.84**	**0.19**	**0.0000091**	LZTFL1, CXCR6	32558485
4	25312372	rs16877005	A/G	**0.74**	**0.37**	**0.048**	0.49	0.27	0.075	PI4K2B	37674002
4	167824478	rs1073165	A/G	**0.29**	**0.14**	**0.0361**	0.075	0.11	0.51	DDX60	37198478
10	112972548	rs7897438	C/A	-0.33	0.18	0.061	**-0.28**	**0.14**	**0.044**	TCF7L2	37674002
11	34482745	rs61882275	G/A	**-0.37**	**0.15**	**0.012**	**-0.39**	**0.12**	**0.00079**	ELF5	35255492
19	10305768	rs73510898	G/A	**0.71**	**0.27**	**0.010**	**0.55**	**0.21**	**0.0092**	ZGLP1, RAVER1, ICAM5	3525549233307546
19	10414696	rs142770866	G/A	0.53	0.27	0.051	**0.47**	**0.21**	**0.028**	PDE4A	37198478
19	48867352	rs4801778	G/T	**-0.41**	**0.19**	**0.032**	**-0.35**	**0.15**	**0.020**	PLEKHA4, TULP2	34237774

Bold if nominally significant in the respective analysis.

[a]WHO-scores 6–10.

[b]WHO-scores 4–10, corresponding to the B1 phenotype definition of COVID-19 HGI. Abbreviations: Chr: Chromosome; Pos: Position in GRCh38 coordinates; ID: rs-ID of the SNP; Ref: Reference allele; Alt: Alternative allele;SE: Standard error; Ref (PMID): Reference given as PubMed ID.

https://doi.org/10.1371/journal.ppat.1012786.t002

as covariates to a logistic regression in order to capture population substructure, the above results became non-significant (p = 0.55, Wald test). Prior filtering of variants with MAF <1% rendered the difference between cases and controls non-significant (p = 0.068, one-sided Wilcoxon test).

Polygenic risk scoring

Next, analyses were performed to investigate whether the aggregated effect of common variants in PRS was significantly increased in cases compared to controls in Ex and B1, and whether the effect differed across age groups. Using PRS generated for individuals within the $DeCOI_{EUR}$ cohort on the basis of the GenOMICC study [22], a significantly larger PRS was observed in cases compared to controls for both phenotypes ($p<0.001$, Wald test followed by Bonferroni correction of p-values). Upon age stratification (younger than 60 years/older or equal 60 years), this result became even more pronounced, with higher mean PRS values being observed in younger cases than in older cases (Fig 3 and S14 Table; <60 years: p(Ex)<0.001, p(B1)<0.001; ≥60 years: p(Ex) = 0.009, p(B1) = 0.035, Wald test followed by Bonferroni).

Analyses were then performed to determine whether the inclusion of PRS improved the approximation of the present data by logistic regression models. For this purpose, two logistic regression models were fitted: 1) with covariates only (namely sex, age, age^2, age*sex and the first 10 PCs derived from common variants); and 2) with the same covariates and PRS. When PRS were added, a significant increase in Nagelkerke's R^2 was observed (Ex: from 0.466 to 0.504; $p = 1.34x10^{-7}$; B1: from 0.403 to 0.424, $p = 1.85x10^{-6,}$ likelihood-ratio test). Analyses were then performed to test whether the addition of PRS to the covariates improved the prediction of hospitalization or a severe disease course. The dataset was split at random 1,000 times into test and training sets, and logistic regression models were fitted to the training set (see Methods). Areas Under the Curves of the Receiver Operating Characteristic curves (AUROCs) were then determined on the test sets. In 1,000 splits, AUROCs were higher (on average) for the model that included PRS, and the median increase of AUROCs was 0.022 (minimum: -0.200, maximum: 0.263) for the hospitalization (B1) and 0.056 (minimum: 0.033, maximum: 0.078) for the extreme (Ex) case-control definition.

Discussion

The present report introduces the DeCOI cohort as one of only a few WGS datasets of 1,000 or more SARS-CoV-2 positive individuals worldwide. While we did not detect any causal variant in or around the established risk gene *TLR7*, the analyses identified carrier status for six autosomal-recessive monogenic disorders in young males who had been hospitalized due to COVID-19. In the European subset ($DeCOI_{EUR}$), burden testing revealed nominal enrichments of rare variants in coding and non-coding regions of genes that are implicated in the interferon immune response both in the cohort overall and in the male-only subgroup. The present analyses also confirmed associations between previously reported common risk loci and COVID-19 severity, including a genome-wide significant association for the risk locus at 3p21.31, and showed that their aggregation into PRS accurately captured risk in an age-dependent manner. Besides complementing ongoing, systematic COVID-19 host genetic efforts to study common [7–10] or rare variants [11,12,14,17,18], our study can be used to jointly analyze variation across the entire frequency spectrum as part of larger, multi-study efforts.

The largest WGS study on severe COVID-19 to date was performed by GenOMICC, and focused on critically-ill patients from intensive care units [22]. This study included more than 7,400 individuals with severe COVID-19, and rare variant associations were analyzed using standard gene-based approaches [22]. Here, the DeCOI WGS data were explored in additional

dimensions, including analyses performed from a clinical genetics perspective. Although our sample size was limited, two characteristics of the DeCOI cohort rendered it suitable for the present analyses. First, the cohort included SARS-CoV-2 infected individuals with mild disease who could be used as controls. The presence of rare causal risk variants among these controls was unlikely, thereby increasing confidence in the rare variant results. Second, the vast majority of participants were recruited during the first 12 months of the pandemic, when: (i) most individuals were not vaccinated against SARS-CoV-2; (ii) re-infections were uncommon; and (iii) SARS-CoV-2 diversity was still low. On the other hand, use of the WHO classification system as a proxy phenotype for severity likely increased classification heterogeneity - this might have limited our statistical power. We envision that the robust identification of low-frequency and rare risk variants will require large cohorts, which is supported by the fact that the GenOMICC consortium failed to identify rare individual genetic factors at the level of genome-wide significance, despite their relatively large sample size and homogenous phenotype definition. Further, additional factors such as prior stimulation of the immune system through viral infections [39] and/or vaccination [40], or the presence of type-I-interferon autoantibodies [3,41], also shape the immune response of each individual, and contribute to the clinical outcomes of SARS-CoV-2 infections. Therefore, future approaches involving the integration of genetic data with clinical information on immune related traits and multi-omics data could facilitate elucidation of the etiological landscape of COVID-19. Notably, such information (e.g., single-cell transcriptomics [42,43]) is already available to some extent for the DeCOI cohort and will be used for subsequent integrative analyses.

Studies that identified TLR7 deficiency as a monogenic form of severe COVID-19 [14,15,19] were limited to the *TLR7* coding region, and thus did not consider potential causal variants in adjacent regions with evidence of regulatory function (including structural variants). Despite comprehensive analyses, no causal SNVs or small indels were detected in the DeCOI cohort, neither in coding nor non-coding regions. This included a lack of any potential causal deletion at the *TLR7* locus in males, which we investigated using coverage data. Nevertheless, the analysis suggested the overrepresentation of a low-frequency variant, located in a constitutive enhancer element that was identified by ENCODE, in severely affected men. However, this result could not be replicated in a small independent WGS dataset, and thus remains inconclusive. We also investigated the association between variants in additional 13 genes of type I interferon (IFN) immunity, for which a recent study estimated a joint odds ratio of 3.11 [95% confidence interval (CI) 1.4–8.6] for having life-threatening COVID-19 when carrying heterozygous pLoF variants in these 13 genes [20] (reported allele frequency of pLoF variants within the 13 genes: 0.004). In our cohort, we identified 7 carriers of at least one heterozygous variant in 5 of these genes but the mutation carriers did not show more severe disease courses than expected by random chance, in line with the absence of replication in other clinically heterogeneous cohorts [17,18,22,44]. Interestingly, in our study we observed an odds ratio of 4.03 for the common lead variant at 3p21.31 (Ex, rs17713054, 95% CI: 2.56–6.37, MAF: 0.08). We speculate that in our cohort, the relevance of monoallelic (i.e. heterozygous) deleterious variants in the 13 genes of the type I IFN immunity is limited. However, this does not exclude the possibility that biallelic variants resulting in rare autosomal recessive inborn errors of immunity within these genes could underlie unexpectedly severe cases, such as severe COVID-19 in children, in the German population, for which our dataset was underpowered.

Epidemiological evidence suggests that pre-existing conditions are a major risk factor for severe COVID-19 [2,34]. The present analyses identified six recessive monogenic disorders in male individuals, who had presented with severe or moderate COVID-19. While this does not imply any causality, it is of note that these six individuals had an age that was below the average age of the DeCOI cohort overall. In several of these individuals, a modification of the COVID-

19 phenotype by the underlying monogenic disease is biologically plausible (see S1 Text). For example, biallelic variants within *AIRE* can cause autoimmune polyendocrinopathy syndrome type 1 (APS-1). In individuals with APS-1, antibodies against IFN-α and IFN-ω are frequently present, and moderate or severe COVID-19 has been described in SARS-CoV-2 infected APS-1 patients [45–47]. Additionally, some of the recessive diseases identified lead to an impairment of important organ systems and could therefore indirectly predispose to more severe COVID-19 disease outcomes (see S1 Text), e.g. Bardet-Biedl Syndrome 1 probably caused the intellectual developmental disorder, and primary hyperoxaluria might have been responsible for the kidney and liver transplant in the two study participants who died of COVID-19, respectively.

In contrast to individuals with putative autosomal-recessive disorders, individuals with putative autosomal-dominant disorders did not differ from the remainder of the DeCOI cohort regarding age or COVID-19 severity. This could be due to a lack of power, which might be attributable to factors such as reduced penetrance, which is more common in dominantly inherited disorders [48]. Overall, it needs to be kept in mind that the results for both autosomal recessive and autosomal dominant monogenic disorders are from a non-representative sample and insufficient to establish any causality.

At the single-gene level, no significant enrichment of rare variants was observed beyond that which would have been expected based on chance alone. Furthermore, the gene-set based analysis of rare variants across candidate genes only yielded nominally significant results. The lowest p-values in our gene-set based analysis were generated for genes that are implicated in the innate immune system, specifically the IFN pathways. Here, pLoF variants, either alone or in combination with missense variants, were enriched in hospitalized or severely affected individuals. Surprisingly, we also observed nominally significant enrichments in mild COVID-19, of variants in the 3'UTRs of genes from the interferon pathway and at GWAS loci. While these results do not withstand statistical correction and warrant independent replication, they are complementing a recent study which identified a highly significant depletion of 3'UTR variants in the gene *IL18RAP* in amyotrophic lateral sclerosis (ALS) patients [49]. Specifically, for IFN genes, we speculate that 3'UTR variants might contribute to an increased stability or abundance of gene product, e.g. through abolishment of miRNA binding sites, as recently suggested for a 3'UTR variant in *TRIM14*, a gene also implicated in the type I IFN pathway [50]. In the gene-/gene-set based collapsing analyses, the availability of the individual's common genotypes was leveraged in order to weigh down individuals with higher PRS, as it has been suggested that integration of PRS into rare-variant burden analyses might be beneficial in terms of their statistical power [51]. It is important to note that most of the rare variant burden signals in the present study were driven by male individuals, which suggests the presence of sex-differences in terms of the extent to which rare variants contribute to severe COVID-19 risk. This finding requires replication in independent cohorts. Also, in the future, novel statistical models that include variants spanning the entire frequency spectrum may enhance the power for rare variant and/or gene identification in cohorts such as DeCOI. A subsample of the present DeCOI cohort already contributed to one such effort [28].

Interestingly, despite our relatively small cohort size, in the association analysis of more frequent variants, our analysis found a comparably large effect size for the contribution of the known risk locus at 3p21.31 to COVID-19 severity, resulting in genome-wide significance. This indicates that this locus is relevant to our cohort of mainly German individuals which might also be true to the German population. Additionally, previously reported GWAS signals were replicated at nominal level, despite a sample size that was substantially lower than those of the discovery cohorts (i.e., GenOMICC or COVID-HGI) [10,22]. When common variants were aggregated into PRS and applied to overall and age-stratified groups, a larger genetic

contribution of common genetic variation to COVID-19 severity was observed in younger individuals. While this has been described previously for candidate lead variants at individual major risk loci [35,52], the present study expanded this analysis to the genome-wide scale. In older individuals, the addition of PRS for COVID-19 severity only moderately improved predictive models, as shown in data from the UK Biobank alone [53] or in the UK Biobank plus three additional US-American cohorts [27]. Since neither of these two studies performed age-stratified analyses, our data suggest that the addition of genetic factors to predictive models could prove particularly helpful in younger individuals, and highlight the translatory potential of PRS. Importantly we constructed the PRS on the basis of WGS data from the GenOMICC cohort, thus reducing the impact of technical variation on score construction.

In conclusion, while the performance of WGS studies continue to be hampered by considerations of cost and sample size, this flagship analysis of the DeCOI cohort highlights the potential of WGS in terms of both investigating variants that are inaccessible to other methods, and performing combined analyses of variants from the entire allelic spectrum, respectively. A more complete understanding of the underlying genetic architecture will be of paramount importance to the clinical (risk) management of individuals with COVID-19 and its post-acute sequelae, which are likely to play important roles in quotidian clinical practice for years to come.

Methods

Ethics statement

Written informed consent for host genetics analyses was obtained from each participant or their legal representative in case of minors. The study received ethical approval by the Ethical Review Board (ERB) of each participating center: Faculty of Medicine at Technical University Munich (TUM 217/20, TUM 221/20S, TUM 440/20S); Medical Faculty of the University Bonn (Approval Nr. 171/20 and 468/20); University of Cologne (20–1295); University Hospital Cologne (160054 and 2001187); Landesärztekammer des Saarlandes (62/20); Medical Faculty of the University Hospital Tübingen (Approval Nr. 286/2020B01); University Hospital RWTH Aachen (EK 080–20); University Hospital Essen (UME: 21-9900-BO); Medical Faculty of Goethe University Frankfurt am Main (20–748); Healthcare System of the Autonomous Province of Bolzano; Medical Faculty of Heinrich-Heine-University Düsseldorf (5350 - amendment for COVID19); Hannover Medical School (9001_BO_K); LMU University Hospital Munich (20–245); Medical Faculty of the LMU Munich (20–263); and Medical Faculty of the University of Regensburg (20-1785-101). Additional details on ERBs are provided in S1 Table.

Recruitment of participants

DeCOI was founded in the spring of 2020, with the aim of advancing next-generation sequencing (NGS)-based COVID-19 research in the areas of viral epidemiology, functional genomics, and host genetics [32]. For the host genetic analyses participants were recruited at 16 different sites, 15 of which were situated in Germany, and one in the German-speaking region of Italy (South Tyrol), from individual COVID-19 studies that were being conducted at the respective institutions. The inclusion criteria for the host genetics analyses were: (i) available DNA; (ii) a test-confirmed SARS-CoV-2 infection; and (iii) explicit consent for WGS analysis. Notably, the type of test used for confirmation of a SARS-CoV-2 infection (self-reports based on rapid antigen tests and/or qPCR) varied across the 16 recruitment sites. Descriptions of the individual studies are provided in S1 Table.

We included 1,275 individuals for WGS analysis. The minimum phenotypic dataset for each individual that was available to the research team comprised sex, age, and information on COVID-19 disease course in accordance with the World Health Organization (WHO) ordinal

scale [33]. The majority of individuals (n = 1,204; 94.4%) were infected in 2020 (n = 1,136/1,275; 89.1%) or early 2021 (January to April 2021, n = 68; 5.3%) and therefore were naive for any COVID-19 vaccination at the time of reported infection. For 71 individuals, no information on vaccination status was available. However, given the limited population-wide availability of COVID-19 vaccination during 2021, and the fact that the latest time point of reported infection in these cases was December of 2021, these individuals are unlikely to have been vaccinated at the time of recruitment.

WGS data generation

Library preparation and sequencing was performed using consolidated workflows at three different sites of the German NGS Competence Centers, i.e., the Cologne and Bonn sites of the West Germany Genome Center (WGGC), and the NGS Competence Center Tübingen (NCCT). In brief, genomic DNA was quantified using the Qubit dsDNA HS assay kit and a Qubit fluorometer (ThermoFisher). DNA library preparation was performed using the TruSeq DNA PCR-Free kit (Illumina), in accordance with the manufacturer's instructions. Up to 1.2 μg of genomic DNA was fragmented to 350 bp using ultrasonication on the LE220 focused-ultrasonicator (Covaris). The resulting libraries were sequenced as paired-end 150 bp reads on an Illumina NovaSeq6000, with a sequencing output of approximately 120 Gb per sample.

At each sequencing site, demultiplexing and FastQ file generation was performed using bcl2fastq2 version 2.20.0.422, and quality control (QC) statistics were generated using FastQC v0.11.9. Subsequently, sequencing reads were aligned to the human reference genome (GRCh38), duplicates were removed, and single nucleotide variants (SNVs) as well as short indels were called using the Illumina DRAGEN platform (software version 3.5.7 or 3.6.3). The resulting gVCF files were transferred to the study analysis hub (WGGC_Bonn), and joint variant calling of all samples was performed using a slightly modified version of GLnexus v1.3.1 (setting: "gatk") in order to yield a raw cohort VCF ("raw-cVCF"). Modifications to the standard GLnexus pipeline included community changes that optimize the caller for haploid regions, which are reported differently in GATK and DRAGEN.

WGS data analysis

The raw-cVCF was modified in order to retain biallelic variants with high-quality individual genotypes only. For this purpose, individual genotypes were set to "missing" if they had low coverage (sequencing depth (DP) < 4 reads) or a genotype quality (GQ) < 20. Furthermore, genotypes were only retained if the fraction of reads with alternative alleles was $<10\%$ or $>90\%$ for homozygous or hemizygous positions, or between 25% and 75% for heterozygous positions. Based on this list of high-quality variants ("cVCF"), two variant sets were established by applying additional filters. The first variant set was termed *"Common variants for QC"* (n = 452,867). Here, the variant set was restricted to variant calls with a minimum DP of 8, a minimum variant call rate (vCR) of 95%, and a minor allele frequency (MAF) $>1\%$. Variants were then limited to those outside of regions with high linkage disequilibrium[32] (LD; see URL section), and were pruned (r^2: 0.2, window size: 1Mb). The second variant set was termed *"Generic variant set"* (n = 53,195,313). Here, after removing samples that did not pass sample QC (see below), calls with DP<8 were set to missing in all genomic regions of females and in autosomal/pseudoautosomal (PAR) genomic regions of males. In addition, heterozygous calls in non-PAR regions of males were set to missing, and only variants with a vCR above 95% were retained.

Functional annotation of variants *in silico* was performed using: (i) the command line version of Variant Effect Predictor (VEP; version 101) with the plugin TSSDistance; (ii) the

external annotation sources gnomAD (version 2.1.1 as well as 3.1.2), ClinVar (version 20221008), dbNSFP (version 4.1a), CADD (version 1.6), SpliceAI and core regions of DNAse I hypersensitive sites (see URLs). The option "pick_allele_gene" was used to ensure that only one consequence per gene was reported for each variant allele.

Sample QC and population subcohorts

Of the 1,275 samples, 35 had an average coverage of <20x and/or a call rate of <90% (based on the "*common variants for QC*" set and autosomal regions, S1 Fig), and were therefore excluded. Next, a subset of the "*common variants for QC*" (Hardy-Weinberg p-values above 0.001 in presumed females) was used to determine genetic sex via the check-sex function of PLINK (version 1.9). Here, 20 individuals were excluded due to divergent genotypic and phenotypic sex. This resulted in a final set of 1,220 individuals ("DeCOI cohort"; Fig 1A and 1B and S2 Table) with diverse population backgrounds.

For the formal statistical analyses, a homogeneous subset of unrelated individuals from one major population background was generated using the "*common variants for QC*" variant set and data from the 1000 genomes project [54]. Principal component (PC) analysis was conducted on variants that were common to both datasets using PLINK (version 1.9). Based on the obtained PCs and the population annotations within the 1000 genomes project, individuals in the DeCOI cohort were then assigned to continental populations. To determine relatedness, kinship coefficients were calculated using the KING software (version 2.2.7). Individuals were defined as related when they had kinship coefficients > 0.04, which indicates third-degree relatedness or closer. From each pair of related individuals, the least severely affected individual was excluded. This approach resulted in a cohort of 1,017 unrelated individuals from the European population ("$DeCOI_{EUR}$"; Fig 1A and 1B and S2 Table). Due to the low number of individuals of non-European ancestry, no other population subcohort was suitable for association testing.

Case/control definitions for association analyses

On the basis of the available phenotypic information, the study participants were classified as having one of three phenotypes: "ambulatory mild" (WHO 1–3), "hospitalized moderate" (WHO 4–5), or "hospitalized severe" (WHO 6–10). For association analyses, these classes were used to assign case/control status to 1,017 individuals of the $DeCOI_{EUR}$ cohort, for two separate case/control definitions (Fig 1A and 1B): (i) "extreme" (Ex / cases: hospitalized severe, n = 272 / controls: ambulatory mild, n = 362), and (ii) "all_hospitalized" (B1 / cases: hospitalized moderate and hospitalized severe, n = 655 / controls: ambulatory mild, n = 362). The phenotype B1 is in accordance with the definition by the COVID-19 HGI [8].

Targeted analysis of variants at the TLR7 locus

The following SNVs were retrieved from the raw-cVCF: (i) those located within *TLR7* protein-coding regions; and (ii) those located in the promoter, 3'/5' untranslated regions (UTRs) and regions annotated as SCREEN enhancers by the ENCODE project (accessed November 30, 2022; 13 elements within the gene body and 50 kb upstream of the transcription start site (TSS)). For the protein-coding regions, the following were selected: (i) all putative loss of function (pLoF) and non-synonymous variants (VEP impact "high" or "moderate"); and (ii) variants with potential effects on splicing (defined as "any spliceAI delta score above 0.5"), independent of MAF. For the non-coding regions, variants were included if they had a maximum allele frequency of 1% according to gnomAD v3.1.2 (popmax value). To identify

potential deletions at the *TLR7* locus, the cohort VCF (region: chrX:12760551–12980636) was queried for stretches of 3 or more variant positions with missing coverage in male individuals.

Filtering for rare variants with strong effects according to variant effect predictions or ClinVar

To identify rare variants with strong effects in DeCOI, we selected variants with an allele count of <5 within the cVCF (n = 1,220 individuals), and excluded variants that had more than one homozygous report in any population from either gnomAD exomes (version 2.1.1) or gnomAD genomes (version 3.1.2). For variants in genes linked to dominant Mendelian disorders, an allele count of 50 or below in gnomAD exomes or genomes was required (sum across all population backgrounds, respectively).

For homozygous or hemizygous variants, a ratio between alternative and total reads (allelic balance) of higher than 95% was required. For heterozygous variants an allelic balance between 25% and 75% was required, as well as a read count of at least 4 for both the reference and the alternative allele. To identify potential compound heterozygous variant carriers, we first filtered for individuals with $\geq$ 2 variants in the same gene. Subsequently, variant co-occurrence (gnomAD version 2; [55]) and/or review of the literature was used to determine if the variants are likely affecting one allele (in *cis*) or both alleles (in *trans*, i.e. compound heterozygous). Based on this strategy, the following analyses were performed:

a. For the "*Analysis of 13 genes previously implicated in severe COVID-19*" we only considered variants that were predicted to be LoF (VEP impact "high") or that were previously shown to result in functional alterations [11].

b. For the "*Targeted analysis of monogenic disorders*", we only retained variants reported as being pathogenic or likely pathogenic in ClinVar by multiple submitters or by expert panels (version 20221008, n = 40,189) [56]. Variants within genes from the American College of Medical Genetics and Genomics (ACMG) secondary findings list [57] were excluded, and variants were only retained if they affected a gene annotated with an Online Mendelian Inheritance in Man (OMIM) phenotype (data downloaded: November 18, 2021). Modes of inheritance were determined using OMIM-data. Genes annotated as being dominant were only retained if they were not annotated with any recessive phenotype in OMIM. The zygosity of the variants identified in the DeCOI cohort had to match the zygosity expected based on the mode of inheritance of the gene, respectively.

To reduce the risk of re-identification for the participants, identified dominant Mendelian diseases are grouped as broad categories and age ranges are reported rather than exact ages.

Gene- and gene-set-based collapsing analyses

Next, gene- and gene-set-based collapsing analyses were conducted to study joint effects of rare variants across single genes and sets of genes with presumed importance to COVID-19. The gene- and gene-set-based collapsing analyses involved three stages.

First, the *definition of genes and gene-sets*: Variants were assigned to one of 19,630 protein-coding genes, as based on position (VEP's annotation; column "SYMBOL"). Furthermore, five gene-sets were curated based on *a priori* evidence or biological plausibility for an involvement in COVID-19 etiology: (a) "GWAS_genes" (94 genes, closest to lead SNV and/or reported as a candidate gene at 71 risk loci identified in prior GWAS for SARS-CoV-2 related traits, including susceptibility and severity, S3 Table); (b) "IFNresponse_COVID-19_genes" (31 genes of the interferon signaling pathway, based on a recent review [30]); (c)

"IFNresponse_reactome_genes" (185 genes of the interferon signaling pathway, based on reactome [58]); (d) "innate_db" (1,037 genes involved in the innate immunity pathway according to the InnateDB platform [59]; and (e) "Zhang_et_al" (13 genes involved in immune response to viral infection with a reported prior enrichment of LoF variants [11]). The major histocompatibility complex (MHC) region was excluded from all lists. Notably, a small overlap was present between individuals from $DeCOI_{EUR}$ and several of the studies from which the "GWAS_genes" list was derived. However, given that this represented less than 0.04% of the entire sample used in the GWAS, the sample overlap was not expected to drive any associations.

Second, *the definition of functional masks for collapsing analyses*: Eleven functional masks (FM) were defined, as based on the predicted consequences of variants (see S4 Table). Briefly, coding variants were classified into categories analogous to those applied in previous studies [18,19]. These categories comprised: (a) predicted loss-of-function (pLoF) variants; and (b) four missense deleteriousness categories, as based on REVEL scores [60]. For non-coding variants, categories of promoters, 5' and 3' UTRs, as well as regulatory elements were defined, and CADD scores [61] were included as a proxy measure of deleteriousness. Variants located in the core regions of DNAse I hypersensitive sites (Altius index) [62] and within 1 kb to 50 kb upstream of the respective TSS were defined as variants in regulatory elements.

Third, *the statistical analyses*. Gene- and gene-set-based collapsing analyses were performed with regenie (version 3.1 [63]), using the $DeCOI_{EUR}$ cohort and the generic variant set (see above). For each analysis, 11 FMs (see above) and two phenotypes (Ex, B1) were tested for association using the default additive model and the '—build-mask sum' option. Based on prior evidence of varying heritability estimates for different age and sex categories [24], gene-set-analyses were also stratified for age (age lower than 60 years / greater or equal to 60 years), and for sex (male / female). For age and sex, stratification applied to both cases and controls. The covariates and options described for the GWAS were used (see section "Single-variant association analyses" below; settings "firth" and "ignore-pred"), with individual polygenic risk score (PRS) being added as a covariate (see section "Polygenic risk scoring" below). The same analysis was also run without PRS. The included variants had an MAF below 0.1%. Allele frequency was determined based on the maximum allele frequency in either the present cohort or gnomAD (version 3.1.2; all populations). Conservative Bonferroni-based thresholds for multiple corrections were alpha = $1.16x10^{-07}$ (19,630 genes, 11 FM, 2 phenotypes) for the single gene analyses, and alpha = $9.1x10^{-05}$ for the gene-set-analysis (5 sets, 11 FM, 2 phenotypes, 5 stratifications). Statistical analyses were only performed if the category contained at least one variant.

Single-variant association analyses

For single variant analyses, two GWAS were performed in the $DeCOI_{EUR}$ cohort using the case/control definitions Ex and B1. For each of the two GWAS, variants were removed from the generic variant list if they met any of the following criteria: MAF < 0.5%, vCR < 98%, missingness-difference between cases and controls above 2%, Hardy-Weinberg $p<10^{-6}$ (among autosomal variants in respective controls), $p<10^{-10}$ (among autosomal variants in cases), $p<10^{-6}$ (among X-chromosomal variants in females). These GWAS variant sets (n = 15,708,109 variants (Ex), n = 15,742,368 (B1)) were pruned ("indep-pairwise 50 5 0.05" command, autosomal variants only, performed in PLINK, n = 548,183 (Ex) and n = 549,436 variants (B1) remaining) and used for calculation of PCs in order to capture the population structure within each GWAS. Together with age, sex, age*age, and age*sex, these 10 PCs were used as covariates in a logistic regression, which was conducted using regenie (version 3.1; options "firth" and "ignore-pred").

For the Ex case-control definition, analysis was re-run in phenotypic substrata (i.e., male/female and younger than 60 years/older or equal 60 years; see above).

Replication cohorts/data

For selected analyses, *in silico* replication was attempted using previously generated summary statistics from the COVID-19 HGI release 7 (array-based data, without GenOMICC and 23andMe) [10] and GenOMICC (WGS data) [22]. For low frequency candidate variants, or when individual genotype data were required, WGS data from the BQC-19 project (Quebec Biobank) [64] were re-analyzed.

Autozygosity

For each individual in the $DeCOI_{EUR}$ cohort, the inbreeding coefficient (F_I) was estimated in accordance with the definition proposed by Wright [65,66], and as implemented in PLINK v1.9 with the—ibc command (Fhat3). F_I was first calculated on the basis of all variants, and then on the basis of those with a MAF $\geq$ 1% (PLINK, option—maf 0.01) to evaluate the robustness of the analysis. Using the Ex and B1 case-control definitions respectively, F_I values between cases and controls were compared using: (i) a one-sided Wilcoxon-test; and (ii) logistic regression with 10 PCs as covariates, as described in the section "Single-variant association analyses". The autozygosity definition follows the standard approach used by Cruz et al. [24] for their "F_{GRM}" analysis. Their "F_{ROH}" analysis approach, which is an ad-hoc assessment of the autozygous proportions in the human genome but not a direct autozygosity measure, was not pursued.

Polygenic risk scoring

WGS-based GWAS data from the GenOMICC study [22], which has no known sample overlap with the DeCOI cohort, were used to generate a PRS for severe COVID-19. The program PRS-CS (version 1.0.0) [67] was applied to the summary statistics of European-ancestry individuals from GenOMICC, using the UK Biobank-based LD reference panel, as provided by PRS-CS. The resulting predictor contained 967,463 variants. PRS for individuals from the $DeCOI_{EUR}$ cohort were then obtained using the '—score' option within PLINK (version 1.9) for variants with MAF>1% of the generic variant set (required: vCR > 98%). These individual scores were included as covariates in the collapsing-analyses (described above).

P-values for the predictor PRS were determined using logistic regression (function glm within R using the parameter family = binomial(link = "logit")), which included PRS as well as the same covariates as those used in the GWAS (see above). To determine whether the PRS improved prediction, two logistic regression models were fitted: (i) with the covariates only; and (ii) with the covariates and the PRS, as described above. Subsequently, the Nakelkerke R^2 was calculated for both models (NagelkerkeR2 function of the R package fmsb). The significance of the differences between the two models were then determined using the likelihood ratio test (lrtest function of the R package rms).

Since logistic regression models can be biased towards the sample used (overfitting), glmnet was also employed, since this provides a combination of ridge and lasso regressions, and is more suitable for the prediction on unknown data. To determine whether PRS added value over random noise, 100 predictors from a normal distribution were simulated, and these were used to train glmnet. To estimate the effect size using independent test data, multiple (1,000) subsampling of our dataset was performed using a random proportion of individuals from 75% to 95% for training, and the remaining dataset for testing. The unequal size of the training set was necessary in order to address the discrete nature of the data and the lack of variability

on comparatively small samples. As a training procedure, cross-validation was used for choosing the optimal parameter, and glmnet was used for the model. Instead of an absolute optimum, lambda plus one standard error was chosen as a more conservative estimate. Statistical analyses were performed as implemented in glmnet (see URL).

Supporting information

S1 Fig. Schematic representation of the quality control (QC) process. After alignment and joint calling of SNVs and Indels, 1,275 individuals with appropriate phenotype data underwent sample quality control to yield a final dataset consisting of 1,017 unrelated individuals of European ancestry ($DeCOI_{EUR}$).
(JPG)

S2 Fig. Principal component analysis. For each individual, principal components were calculated based on the "common variants for QC" variant set. (A) The first two principal components (PC1 and PC2) are plotted for all individuals of DeCOI (empty forms) together with individuals from the 1000 genomes project (1 KG reference cohort, grey circles). Individuals assigned to the European subcohort of DeCOI ($DeCOI_{EUR}$) are plotted in blue circles, while all others are indicated in black triangles. The region marked by the dashed box is enlarged in panels B-D. (B) and (C): The individuals of $DeCOI_{EUR}$ are plotted within the PC-space, colored by their case-control definitions in analyses Ex and B1. In (D), all individuals of $DeCOI_{EUR}$ are plotted with colors indicating their respective site of sequencing.
(JPG)

S3 Fig. Characteristics of carriers of pathogenic variants with established links to monogenic diseases. (A) Box plot indicating the age distribution of individuals in which a heterozygous (filled with checkerboard pattern) or biallelic (blue data points, includes compound heterozygous) variant with an established link to a monogenic disease was or was not found (filled in white). The elements of the box plot correspond to the following values: thick line: median, box: 25th and 75th percentile, whiskers: largest / smallest value not further away from the box than 1.5 times the interquartile range, points: values outside of the range of the whiskers. Panels (B) to (D) show the proportion of heterozygous variant carriers according to cohort membership (B), severity (C) or sex (D). The numbers above the bars indicate the total number of individuals in each stratum. Note that statistical testing was performed using student's t-test for age (A) or fisher's exact test (B-D). Except for nominally significant differences in age, no statistically significant different proportions between strata were detected (lowest nominal p-value: 0.13). p_{nom}: uncorrected p-value.
(JPG)

S4 Fig. Gene-based collapsing analyses in $DeCOI_{EUR}$. (A-B) Quantile-quantile plots for phenotypes Ex (A) and B1 (B). (C-D) Scatter plots showing the negative decadic logarithm of the p-values for gene / functional mask combinations when PRS was included (x-axis) or not included (y-axis) as a covariate. The p-values were calculated using the phenotype definitions, as indicated in the left upper corner of the scatter plots. Pearson correlation coefficients between negative decadic logarithms of the p-values calculated with or without PRS as covariate were 0.92 for Ex and 0.96 for both B1.
(JPG)

S5 Fig. Quantile-quantile (QQ) plots of GWAS. Phenotypes and corresponding genomic inflation factors (lambda) are indicated within the respective panels.
(JPG)

S6 Fig. Results of stratified analyses within Ex. Manhattan plots (left panel) and quantile-quantile plots (right panel) are represented for analyses including individuals which were of female (Ex_female) or male (Ex_male) sex, and younger than 60 years (Ex_LT60) or 60 years or older (Ex_GE60). Details on all variants with $P<10^{-05}$ in any of the four substrata are listed in S13 Table.
(JPG)

S7 Fig. Distribution of autozygosity in samples of the $DeCOI_{EUR}$ cohort. Distribution of inbreeding coefficients in cases and controls according to the B1 and Ex classifications. The dashed horizontal lines represent thresholds of 0.02 (green), 0.05 (blue) and 0.1 (red), respectively.
(JPG)

S8 Fig. Comparison of PCs in samples of $DeCOI_{EUR}$ cohort. Values of principal component 1 and 2 for individuals of the $DeCOI_{EUR}$ cohort are shown for different ranges of the inbreeding coefficient (FI). Case / control status for B1 (left) or Ex (right) is color coded only, if individuals were within the specified range of FI, otherwise individuals are colored in grey.
(JPG)

S1 Table. Description of individual cohorts.
(XLSX)

S2 Table. Characteristics of the overall DeCOI cohort (left) and the European subcohort ($DeCOI_{EUR}$, right).
(XLSX)

S3 Table. Overview of genes used in five different gene-sets.
(XLSX)

S4 Table. Definition of functional masks for gene collapsing analyses.
(XLSX)

S5 Table. Rare variants within coding and non-coding regions of *TLR7*.
(XLSX)

S6 Table. Replication results for rs192357402 in the Quebec Biobank.
(XLSX)

S7 Table. Variants in 13 genes previously implicated in severe COVID-19 and characteristics of carriers.
(XLSX)

S8 Table. Results of gene collapsing analysis in Ex. This table contains the 5000 most significant results, for a full list please refer to the Data Availability section.
(XLSX)

S9 Table. Results of gene collapsing analysis in B1. This table contains the 5000 most significant results, for a full list please refer to the Data Availability section.
(XLSX)

S10 Table. Results of gene-set analyses.
(XLSX)

S11 Table. Results of most significant variants in B1 analysis of $DeCOI_{EUR}$.
(XLSX)

S12 Table. Association results for known risk loci.
(XLSX)

S13 Table. Results of age- and sex-stratified single variant association analysis in B1. All variants that have P<10–05 in at least one subcategory are shown.
(XLSX)

S14 Table. Mean PRS values in cases and controls.
(XLSX)

S15 Table. Names and affiliations of members of the DeCOI host genetics group.
(XLSX)

S16 Table. Names and affiliations of members of the DeCOI group.
(XLSX)

S1 Text. This file contains additional information and references on the four autosomal-recessive genes.
(PDF)

Acknowledgments

The authors thank the Next Generation Sequencing Competence Network (NGS-CN) for their continuous and invaluable input in terms of study organization and logistics. We also thank the following individuals for supporting the work in the laboratory: Michèle Hochstein, Matthias Potschka, Julia Fazaal, Laura Müller, Wenke Barkey, Norma Koch, Sophie Hinreiner, Antje Schulze Selting, and Natascha Demovski. We thank David Ellinghaus, Lea Nikolai, and Ersoy Kocak for their support with data transfer. Finally, we thank Martina Seibert for her support of the required clinical work.

URLs

TSSDistance plugin of VEP:
https://github.com/Ensembl/VEP_plugins/blob/release/101/TSSDistance.pm
ENCODE / SCREEN: https://screen.wenglab.org
glmnet: https://glmnet.stanford.edu/articles/glmnet.html
gnomAD version 3.1.2: https://gnomad.broadinstitute.org/downloads
ClinVar: https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/
dbNSFP4.1: https://sites.google.com/site/jpopgen/dbNSFP
CADD (version 1.6): https://cadd.gs.washington.edu/download
SpliceAI: Pre-computed scores were downloaded from Illumina Basespace after registration (https://basespace.illumina.com/s/otSPW8hnhaZR)
DNAse I hypersensitive sites: https://doi.org/10.1101/822510
Regions of strong LD:
https://genome.sph.umich.edu/wiki/Regions_of_high_linkage_disequilibrium_(LD))

Author Contributions

Conceptualization: Axel Schmidt, Stephan Ossowski, Olaf Riess, Eva C. Schulte, Kerstin U. Ludwig.

Data curation: Nicolas Casadei, Max Augustin, Robert Bals, Carla Bellinghausen, Marc Moritz Berger, Michael Bitzer, Christian Bode, Jannik Boos, Thorsten Brenner, Oliver A. Cornely, Thomas Eggermann, Johanna Erber, Torsten Feldt, Christian Fuchsberger, Julien Gagneur, Siri Göpel, Tobias Haack, Helene Häberle, Frank Hanses, Julia Heggemann, Ute Hehr,

Johannes C. Hellmuth, Christian Herr, Anke Hinney, Thomas Illig, Björn-Erik Ole Jensen, Verena Keitel, Sarah Kim-Hellmuth, Philipp Koehler, Ingo Kurth, Anna-Lisa Lanz, Eicke Latz, Clara Lehmann, Tom Luedde, Michael Mian, Abigail Miller, Maximilian Muenchhoff, Isabell Pink, Ulrike Protzer, Hana Rohn, Jan Rybniker, Federica Scaggiante, Anna Schaffeldt, Clemens Scherer, Maximilian Schieck, Susanne V. Schmidt, Philipp Schommers, Christoph D. Spinner, Maria J. G. T. Vehreschild, Thirumalaisamy P. Velavan, Sonja Volland, Sibylle Wilfling, Christof Winter, J. Brent Richards, André Heimbach, Kerstin Becker, Stephan Ossowski, Susanne Motameny, Michael Nothnagel, Olaf Riess, Eva C. Schulte, Kerstin U. Ludwig.

Formal analysis: Axel Schmidt, Nicolas Casadei, Fabian Brand, German Demidov, Elaheh Vojgani, Ayda Abolhassani, Rana Aldisi, Guillaume Butler-Laporte, T. Madhusankha Alawathurage, Carlo Maj, J. Brent Richards, Stephan Ossowski, Susanne Motameny, Michael Nothnagel, Eva C. Schulte, Kerstin U. Ludwig.

Funding acquisition: Per Hoffmann, Joachim L. Schultze, Peter Nürnberg, Markus M. Nöthen, Olaf Riess, Eva C. Schulte, Kerstin U. Ludwig.

Investigation: Axel Schmidt, Nicolas Casadei, Fabian Brand, German Demidov, Elaheh Vojgani, Ayda Abolhassani, Rana Aldisi, Guillaume Butler-Laporte, T. Madhusankha Alawathurage, Carlo Maj, J. Brent Richards, Stephan Ossowski, Susanne Motameny, Michael Nothnagel, Eva C. Schulte, Kerstin U. Ludwig.

Resources: Nicolas Casadei, Max Augustin, Robert Bals, Carla Bellinghausen, Marc Moritz Berger, Michael Bitzer, Christian Bode, Jannik Boos, Thorsten Brenner, Oliver A. Cornely, Thomas Eggermann, Johanna Erber, Torsten Feldt, Christian Fuchsberger, Julien Gagneur, Siri Göpel, Tobias Haack, Helene Häberle, Frank Hanses, Julia Heggemann, Ute Hehr, Johannes C. Hellmuth, Christian Herr, Anke Hinney, Thomas Illig, Björn-Erik Ole Jensen, Verena Keitel, Sarah Kim-Hellmuth, Philipp Koehler, Ingo Kurth, Anna-Lisa Lanz, Eicke Latz, Clara Lehmann, Tom Luedde, Michael Mian, Abigail Miller, Maximilian Muenchhoff, Isabell Pink, Ulrike Protzer, Hana Rohn, Jan Rybniker, Federica Scaggiante, Anna Schaffeldt, Clemens Scherer, Maximilian Schieck, Susanne V. Schmidt, Philipp Schommers, Christoph D. Spinner, Maria J. G. T. Vehreschild, Thirumalaisamy P. Velavan, Sonja Volland, Sibylle Wilfling, Christof Winter, J. Brent Richards, André Heimbach, Kerstin Becker, Stephan Ossowski, Susanne Motameny, Michael Nothnagel, Olaf Riess, Eva C. Schulte, Kerstin U. Ludwig.

Supervision: Axel Schmidt, Eva C. Schulte, Kerstin U. Ludwig.

Visualization: Eva C. Schulte.

Writing – original draft: Axel Schmidt, Nicolas Casadei, Fabian Brand, German Demidov, Elaheh Vojgani, Ayda Abolhassani, Rana Aldisi, Guillaume Butler-Laporte, T. Madhusankha Alawathurage, Susanne Motameny, Michael Nothnagel, Eva C. Schulte, Kerstin U. Ludwig.

Writing – review & editing: Axel Schmidt, Nicolas Casadei, Fabian Brand, German Demidov, Elaheh Vojgani, Ayda Abolhassani, Rana Aldisi, Guillaume Butler-Laporte, T. Madhusankha Alawathurage, Max Augustin, Robert Bals, Carla Bellinghausen, Marc Moritz Berger, Michael Bitzer, Christian Bode, Jannik Boos, Thorsten Brenner, Oliver A. Cornely, Thomas Eggermann, Johanna Erber, Torsten Feldt, Christian Fuchsberger, Julien Gagneur, Siri Göpel, Tobias Haack, Helene Häberle, Frank Hanses, Julia Heggemann, Ute Hehr, Johannes C. Hellmuth, Christian Herr,

Anke Hinney, Per Hoffmann, Thomas Illig, Björn-Erik Ole Jensen, Verena Keitel, Sarah Kim-Hellmuth, Philipp Koehler, Ingo Kurth, Anna-Lisa Lanz, Eicke Latz, Clara Lehmann, Tom Luedde, Carlo Maj, Michael Mian, Abigail Miller, Maximilian Muenchhoff, Isabell Pink, Ulrike Protzer, Hana Rohn, Jan Rybniker, Federica Scaggiante, Anna Schaffeldt, Clemens Scherer, Maximilian Schieck, Susanne V. Schmidt, Philipp Schommers, Christoph D. Spinner, Maria J. G. T. Vehreschild, Thirumalaisamy P. Velavan, Sonja Volland, Sibylle Wilfling, Christof Winter, J. Brent Richards, André Heimbach, Kerstin Becker, Stephan Ossowski, Joachim L. Schultze, Peter Nürnberg, Markus M. Nöthen, Olaf Riess.

References

1. O'Driscoll M, Ribeiro Dos Santos G, Wang L, Cummings DAT, Azman AS, Paireau J, et al. Age-specific mortality and immunity patterns of SARS-CoV-2. Nature. 2021 Feb; 590(7844):140–5. https://doi.org/10.1038/s41586-020-2918-0 PMID: 33137809
2. Williamson EJ, Walker AJ, Bhaskaran K, Bacon S, Bates C, Morton CE, et al. Factors associated with COVID-19-related death using OpenSAFELY. Nature. 2020 Aug; 584(7821):430–6. https://doi.org/10.1038/s41586-020-2521-4 PMID: 32640463
3. Bastard P, Rosen LB, Zhang Q, Michailidis E, Hoffmann HH, Zhang Y, et al. Autoantibodies against type I IFNs in patients with life-threatening COVID-19. Science. 2020 Oct 23; 370(6515):eabd4585. https://doi.org/10.1126/science.abd4585 PMID: 32972996
4. Williams FMK, Freidin MB, Mangino M, Couvreur S, Visconti A, Bowyer RCE, et al. Self-Reported Symptoms of COVID-19, Including Symptoms Most Predictive of SARS-CoV-2 Infection, Are Heritable. Twin Res Hum Genet Off J Int Soc Twin Stud. 2020 Dec; 23(6):316–21. https://doi.org/10.1017/thg.2020.85 PMID: 33558003
5. COVID-19 Host Genetics Initiative. The COVID-19 Host Genetics Initiative, a global initiative to elucidate the role of host genetic factors in susceptibility and severity of the SARS-CoV-2 virus pandemic. Eur J Hum Genet EJHG. 2020 Jun;28(6):715–8.
6. Niemi MEK, Daly MJ, Ganna A. The human genetic epidemiology of COVID-19. Nat Rev Genet. 2022 Sep; 23(9):533–46. https://doi.org/10.1038/s41576-022-00478-5 PMID: 35501396
7. Pairo-Castineira E, Clohisey S, Klaric L, Bretherick AD, Rawlik K, Pasko D, et al. Genetic mechanisms of critical illness in COVID-19. Nature. 2021 Mar; 591(7848):92–8. https://doi.org/10.1038/s41586-020-03065-y PMID: 33307546
8. COVID-19 Host Genetics Initiative. Mapping the human genetic architecture of COVID-19. Nature. 2021 Dec;600(7889):472–7.
9. Severe Covid-19 GWAS Group, Ellinghaus D, Degenhardt F, Bujanda L, Buti M, Albillos A, et al. Genomewide Association Study of Severe Covid-19 with Respiratory Failure. N Engl J Med. 2020 Oct 15;383(16):1522–34.
10. COVID-19 Host Genetics Initiative. A second update on mapping the human genetic architecture of COVID-19. Nature. 2023 Sep;621(7977):E7–26.
11. Zhang Q, Bastard P, Liu Z, Le Pen J, Moncada-Velez M, Chen J, et al. Inborn errors of type I IFN immunity in patients with life-threatening COVID-19. Science. 2020 Oct 23; 370(6515):eabd4570.
12. Schmidt A, Peters S, Knaus A, Sabir H, Hamsen F, Maj C, et al. TBK1 and TNFRSF13B mutations and an autoinflammatory disease in a child with lethal COVID-19. NPJ Genomic Med. 2021 Jul 1; 6(1):55.
13. Tangye SG, Al-Herz W, Bousfiha A, Cunningham-Rundles C, Franco JL, Holland SM, et al. Human Inborn Errors of Immunity: 2022 Update on the Classification from the International Union of Immunological Societies Expert Committee. J Clin Immunol. 2022 Oct; 42(7):1473–507. https://doi.org/10.1007/s10875-022-01289-3 PMID: 35748970
14. van der Made CI, Simons A, Schuurs-Hoeijmakers J, van den Heuvel G, Mantere T, Kersten S, et al. Presence of Genetic Variants Among Young Men With Severe COVID-19. JAMA. 2020 Aug 18; 324(7):663–73. https://doi.org/10.1001/jama.2020.13719 PMID: 32706371
15. Asano T, Boisson B, Onodi F, Matuozzo D, Moncada-Velez M, Maglorius Renkilaraj MRL, et al. X-linked recessive TLR7 deficiency in ~1% of men under 60 years old with life-threatening COVID-19. Sci Immunol. 2021 Aug 19; 6(62):eabl4348. https://doi.org/10.1126/sciimmunol.abl4348 PMID: 34413140
16. Fallerini C, Daga S, Mantovani S, Benetti E, Picchiotti N, Francisci D, et al. Association of Toll-like receptor 7 variants with life-threatening COVID-19 disease in males: findings from a nested case-control study. eLife. 2021 Mar 2; 10:e67569. https://doi.org/10.7554/eLife.67569 PMID: 33650967

17. Kosmicki JA, Horowitz JE, Banerjee N, Lanche R, Marcketta A, Maxwell E, et al. Pan-ancestry exome-wide association analyses of COVID-19 outcomes in 586,157 individuals. Am J Hum Genet. 2021 Jul 1; 108(7):1350–5. https://doi.org/10.1016/j.ajhg.2021.05.017 PMID: 34115965

18. Butler-Laporte G, Povysil G, Kosmicki JA, Cirulli ET, Drivas T, Furini S, et al. Exome-wide association study to identify rare variants influencing COVID-19 outcomes: Results from the Host Genetics Initiative. PLoS Genet. 2022 Nov; 18(11):e1010367. https://doi.org/10.1371/journal.pgen.1010367 PMID: 36327219

19. Boos J, van der Made CI, Ramakrishnan G, Coughlan E, Asselta R, Löscher BS, et al. Stratified analyses refine association between TLR7 rare variants and severe COVID-19. HGG Adv. 2024 Jun 28;100323. https://doi.org/10.1016/j.xhgg.2024.100323 PMID: 38944683

20. Matuozzo D, Talouarn E, Marchal A, Zhang P, Manry J, Seeleuthner Y, et al. Rare predicted loss-of-function variants of type I IFN immunity genes are associated with life-threatening COVID-19. Genome Med. 2023 Apr 5; 15(1):22. https://doi.org/10.1186/s13073-023-01173-8 PMID: 37020259

21. Namkoong H, Edahiro R, Takano T, Nishihara H, Shirai Y, Sonehara K, et al. DOCK2 is involved in the host genetics and biology of severe COVID-19. Nature. 2022 Sep; 609(7928):754–60. https://doi.org/10.1038/s41586-022-05163-5 PMID: 35940203

22. Kousathanas A, Pairo-Castineira E, Rawlik K, Stuckey A, Odhams CA, Walker S, et al. Whole-genome sequencing reveals host factors underlying critical COVID-19. Nature. 2022 Jul; 607(7917):97–103. https://doi.org/10.1038/s41586-022-04576-6 PMID: 35255492

23. COVID-19 Host Genetics Initiative. A first update on mapping the human genetic architecture of COVID-19. Nature. 2022 Aug;608(7921):E1–10.

24. Cruz R, Diz-de Almeida S, López de Heredia M, Quintela I, Ceballos FC, Pita G, et al. Novel genes and sex differences in COVID-19 severity. Hum Mol Genet. 2022 Nov 10; 31(22):3789–806. https://doi.org/10.1093/hmg/ddac132 PMID: 35708486

25. Shelton JF, Shastri AJ, Fletez-Brant K, 23andMe COVID-19 Team, Aslibekyan S, Auton A. The UGT2A1/UGT2A2 locus is associated with COVID-19-related loss of smell or taste. Nat Genet. 2022 Feb; 54(2):121–4.

26. Shelton JF, Shastri AJ, Ye C, Weldon CH, Filshtein-Sonmez T, Coker D, et al. Trans-ancestry analysis reveals genetic and nongenetic associations with COVID-19 susceptibility and severity. Nat Genet. 2021 Jun; 53(6):801–8. https://doi.org/10.1038/s41588-021-00854-7 PMID: 33888907

27. Horowitz JE, Kosmicki JA, Damask A, Sharma D, Roberts GHL, Justice AE, et al. Genome-wide analysis provides genetic evidence that ACE2 influences COVID-19 risk and yields risk scores associated with severe disease. Nat Genet. 2022 Apr; 54(4):382–92. https://doi.org/10.1038/s41588-021-01006-7 PMID: 35241825

28. Fallerini C, Picchiotti N, Baldassarri M, Zguro K, Daga S, Fava F, et al. Common, low-frequency, rare, and ultra-rare coding variants contribute to COVID-19 severity. Hum Genet. 2022 Jan; 141(1):147–73. https://doi.org/10.1007/s00439-021-02397-7 PMID: 34889978

29. Wang F, Huang S, Gao R, Zhou Y, Lai C, Li Z, et al. Initial whole-genome sequencing and analysis of the host genetic contribution to COVID-19 severity and susceptibility. Cell Discov. 2020 Nov 10; 6 (1):83. https://doi.org/10.1038/s41421-020-00231-4 PMID: 33298875

30. van der Made CI, Netea MG, van der Veerdonk FL, Hoischen A. Clinical implications of host genetic variation and susceptibility to severe or critical COVID-19. Genome Med. 2022 Aug 19; 14(1):96. https://doi.org/10.1186/s13073-022-01100-3 PMID: 35986347

31. Abolhassani H, Landegren N, Bastard P, Materna M, Modaresi M, Du L, et al. Inherited IFNAR1 Deficiency in a Child with Both Critical COVID-19 Pneumonia and Multisystem Inflammatory Syndrome. J Clin Immunol. 2022 Apr; 42(3):471–83. https://doi.org/10.1007/s10875-022-01215-7 PMID: 35091979

32. Schultze JL. Deutsche COVID-19 Omics Initiative (DeCOI). Biospektrum Z Ges Biol Chem GBCH Ver Allg Angew Mikrobiol VAAM. 2021; 27(3):227.

33. WHO Working Group on the Clinical Characterisation and Management of COVID-19 infection. A minimal common outcome measure set for COVID-19 clinical research. Lancet Infect Dis. 2020 Aug;20(8): e192–7.

34. Zhang H, Thygesen JH, Shi T, Gkoutos GV, Hemingway H, Guthrie B, et al. Increased COVID-19 mortality rate in rare disease patients: a retrospective cohort study in participants of the Genomics England 100,000 Genomes project. Orphanet J Rare Dis. 2022 Apr 12; 17(1):166. https://doi.org/10.1186/s13023-022-02312-x PMID: 35414031

35. Nakanishi T, Pigazzini S, Degenhardt F, Cordioli M, Butler-Laporte G, Maya-Miles D, et al. Age-dependent impact of the major common genetic risk factor for COVID-19 on severity and mortality. J Clin Invest. 2021 Dec 1; 131(23):e152386. https://doi.org/10.1172/JCI152386 PMID: 34597274

36. Zeberg H, Pääbo S. The major genetic risk factor for severe COVID-19 is inherited from Neanderthals. Nature. 2020 Nov; 587(7835):610–2. https://doi.org/10.1038/s41586-020-2818-3 PMID: 32998156

37. Möhlendick B, Schönfelder K, Zacher C, Elsner C, Rohn H, Konik MJ, et al. The GNB3 c.825C>T (rs5443) polymorphism and protection against fatal outcome of corona virus disease 2019 (COVID-19). Front Genet. 2022; 13:960731.

38. Weiner J, Suwalski P, Holtgrewe M, Rakitko A, Thibeault C, Müller M, et al. Increased risk of severe clinical course of COVID-19 in carriers of HLA-C*04:01. EClinicalMedicine. 2021 Oct; 40:101099.

39. Sagar M, Reifler K, Rossi M, Miller NS, Sinha P, White LF, et al. Recent endemic coronavirus infection is associated with less-severe COVID-19. J Clin Invest. 2021 Jan 4; 131(1):e143380, 143380. https://doi.org/10.1172/JCI143380 PMID: 32997649

40. Becker M, Dulovic A, Junker D, Ruetalo N, Kaiser PD, Pinilla YT, et al. Immune response to SARS-CoV-2 variants of concern in vaccinated individuals. Nat Commun. 2021 May 25; 12(1):3109. https://doi.org/10.1038/s41467-021-23473-6 PMID: 34035301

41. Wang EY, Mao T, Klein J, Dai Y, Huck JD, Jaycox JR, et al. Diverse functional autoantibodies in patients with COVID-19. Nature. 2021 Jul 8; 595(7866):283–8. https://doi.org/10.1038/s41586-021-03631-y PMID: 34010947

42. Schulte-Schrepping J, Reusch N, Paclik D, Baßler K, Schlickeiser S, Zhang B, et al. Severe COVID-19 Is Marked by a Dysregulated Myeloid Cell Compartment. Cell. 2020 Sep 17; 182(6):1419–1440.e23. https://doi.org/10.1016/j.cell.2020.08.001 PMID: 32810438

43. Aschenbrenner AC, Mouktaroudi M, Krämer B, Oestreich M, Antonakos N, Nuesch-Germano M, et al. Disease severity-specific neutrophil signatures in blood transcriptomes stratify COVID-19 patients. Genome Med. 2021 Jan 13; 13(1):7. https://doi.org/10.1186/s13073-020-00823-5 PMID: 33441124

44. Povysil G, Butler-Laporte G, Shang N, Wang C, Khan A, Alaamery M, et al. Rare loss-of-function variants in type I IFN immunity genes are not associated with severe COVID-19. J Clin Invest. 2021 Jul 15; 131(14):e147834. https://doi.org/10.1172/JCI147834 PMID: 34043590

45. Beccuti G, Ghizzoni L, Cambria V, Codullo V, Sacchi P, Lovati E, et al. A COVID-19 pneumonia case report of autoimmune polyendocrine syndrome type 1 in Lombardy, Italy: letter to the editor. J Endocrinol Invest. 2020 Aug; 43(8):1175–7. https://doi.org/10.1007/s40618-020-01323-4 PMID: 32519200

46. Lemarquis A, Campbell T, Aranda-Guillén M, Hennings V, Brodin P, Kämpe O, et al. Severe COVID-19 in an APS1 patient with interferon autoantibodies treated with plasmapheresis. J Allergy Clin Immunol. 2021 Jul; 148(1):96–8. https://doi.org/10.1016/j.jaci.2021.03.034 PMID: 33892926

47. Le Voyer T, Parent AV, Liu X, Cederholm A, Gervais A, Rosain J, et al. Autoantibodies against type I IFNs in humans with alternative NF-κB pathway deficiency. Nature. 2023 Nov; 623(7988):803–13.

48. Cooper DN, Krawczak M, Polychronakos C, Tyler-Smith C, Kehrer-Sawatzki H. Where genotype is not predictive of phenotype: towards an understanding of the molecular basis of reduced penetrance in human inherited disease. Hum Genet. 2013 Oct; 132(10):1077–130. https://doi.org/10.1007/s00439-013-1331-2 PMID: 23820649

49. Eitan C, Siany A, Barkan E, Olender T, van Eijk KR, Moisse M, et al. Whole-genome sequencing reveals that variants in the Interleukin 18 Receptor Accessory Protein 3'UTR protect against ALS. Nat Neurosci. 2022 Apr; 25(4):433–45. https://doi.org/10.1038/s41593-022-01040-6 PMID: 35361972

50. Griesemer D, Xue JR, Reilly SK, Ulirsch JC, Kukreja K, Davis JR, et al. Genome-wide functional screen of 3'UTR variants uncovers causal variants for human disease and evolution. Cell. 2021 Sep 30; 184(20):5247–5260.e19. https://doi.org/10.1016/j.cell.2021.08.025 PMID: 34534445

51. Zhou D, Yu D, Scharf JM, Mathews CA, McGrath L, Cook E, et al. Contextualizing genetic risk score for disease screening and rare variant discovery. Nat Commun. 2021 Jul 20; 12(1):4418. https://doi.org/10.1038/s41467-021-24387-z PMID: 34285202

52. Degenhardt F, Ellinghaus D, Juzenas S, Lerga-Jaso J, Wendorff M, Maya-Miles D, et al. Detailed stratified GWAS analysis for severe COVID-19 in four European populations. Hum Mol Genet. 2022 Nov 28; 31(23):3945–66. https://doi.org/10.1093/hmg/ddac158 PMID: 35848942

53. Zhu D, Zhao R, Yuan H, Xie Y, Jiang Y, Xu K, et al. Host Genetic Factors, Comorbidities and the Risk of Severe COVID-19. J Epidemiol Glob Health. 2023 May 9; 13(2):279–91. https://doi.org/10.1007/s44197-023-00106-3 PMID: 37160831

54. The 1000 Genomes Project Consortium, Corresponding authors, Auton A, Abecasis GR, Steering committee, Altshuler DM, et al. A global reference for human genetic variation. Nature. 2015 Oct 1;526(7571):68–74.

55. Guo MH, Francioli LC, Stenton SL, Goodrich JK, Watts NA, Singer-Berk M, et al. Inferring compound heterozygosity from large-scale exome sequencing data. Nat Genet. 2024 Jan; 56(1):152–61. https://doi.org/10.1038/s41588-023-01608-3 PMID: 38057443

56. Landrum MJ, Lee JM, Benson M, Brown GR, Chao C, Chitipiralla S, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018 Jan 4; 46(D1):D1062–7. https://doi.org/10.1093/nar/gkx1153 PMID: 29165669

57. Miller DT, Lee K, Chung WK, Gordon AS, Herman GE, Klein TE, et al. ACMG SF v3.0 list for reporting of secondary findings in clinical exome and genome sequencing: a policy statement of the American College of Medical Genetics and Genomics (ACMG). Genet Med Off J Am Coll Med Genet. 2021 Aug; 23(8):1381–90.

58. Gillespie M, Jassal B, Stephan R, Milacic M, Rothfels K, Senff-Ribeiro A, et al. The reactome pathway knowledgebase 2022. Nucleic Acids Res. 2022 Jan 7; 50(D1):D687–92. https://doi.org/10.1093/nar/gkab1028 PMID: 34788843

59. Breuer K, Foroushani AK, Laird MR, Chen C, Sribnaia A, Lo R, et al. InnateDB: systems biology of innate immunity and beyond—recent updates and continuing curation. Nucleic Acids Res. 2013 Jan; 41:D1228–1233. https://doi.org/10.1093/nar/gks1147 PMID: 23180781

60. 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 Oct 6; 99 (4):877–85. https://doi.org/10.1016/j.ajhg.2016.08.016 PMID: 27666373

61. Rentzsch P, Schubach M, Shendure J, Kircher M. CADD-Splice-improving genome-wide variant effect prediction using deep learning-derived splice scores. Genome Med. 2021 Feb 22; 13(1):31. https://doi.org/10.1186/s13073-021-00835-9 PMID: 33618777

62. Meuleman W, Muratov A, Rynes E, Halow J, Lee K, Bates D, et al. Index and biological spectrum of human DNase I hypersensitive sites. Nature. 2020 Aug; 584(7820):244–51. https://doi.org/10.1038/s41586-020-2559-3 PMID: 32728217

63. Mbatchou J, Barnard L, Backman J, Marcketta A, Kosmicki JA, Ziyatdinov A, et al. Computationally efficient whole-genome regression for quantitative and binary traits. Nat Genet. 2021 Jul; 53(7):1097–103. https://doi.org/10.1038/s41588-021-00870-7 PMID: 34017140

64. Tremblay K, Rousseau S, Zawati MH, Auld D, Chassé M, Coderre D, et al. The Biobanque québécoise de la COVID-19 (BQC19)-A cohort to prospectively study the clinical and biological determinants of COVID-19 clinical trajectories. PloS One. 2021; 16(5):e0245031.

65. Wright S. SYSTEMS OF MATING. I. THE BIOMETRIC RELATIONS BETWEEN PARENT AND OFFSPRING. Genetics. 1921 Mar 1; 6(2):111–23. https://doi.org/10.1093/genetics/6.2.111 PMID: 17245958

66. Wright S. Coefficients of Inbreeding and Relationship. Am Nat. 1922 Jul; 56(645):330–8.

67. Ge T, Chen CY, Ni Y, Feng YCA, Smoller JW. Polygenic prediction via Bayesian regression and continuous shrinkage priors. Nat Commun. 2019 Apr 16; 10(1):1776. https://doi.org/10.1038/s41467-019-09718-5 PMID: 30992449

4 Discussion

WGS will continue to be the bedrock of many studies in human genetics for the foreseeable future. While the sequencing machines continue to evolve, some best practice workflows, which ought to be considered very stable, have emerged. In this dissertation we presented how we implemented these workflows for analyzing sequencing data, and how we used the data to derive clinical insights and establish a biomarker for paternal exposure to IR. WGS data from a single individual allowed us to detect a 4.7k bp deletion in the *KANSL1* gene, which was found to be causative for Koolen-de Vries syndrome. WGS data from multiple large cohorts was analyzed and used for two studies. Firstly, we used validated DNM and cDNM calls as well as data from ancillary sequencing techniques to build a deep-learning model for the detection of DNMs and cDNMs from raw-read data. We showed that this model achieves state-of-the-art accuracy in all tested scenarios, generalizing well to data from other sequencers or with other read lengths. Secondly, we analyzed the data from 4,337 WGS cases for DNMs and cDNMs to find potential signatures of paternal exposure to ionizing radiation. We found an increased number of cDNMs in the offspring of former radar personnel of both German armies and in offspring of Liquidators and inhabitants of the town of Pripyat at the time of the nuclear accident. We could associate this increase in the number of cDNMs with the increase in received dose of the parents of each offspring.

The reports contained herein showed, that deep-learning can improve the detection accuracy of complex mutational signatures in short-read data over traditional statistical and heuristic methods, and that, contrary to other claims, there likely are transgenerational biomarkers for the paternal exposure to IR (Yeager et al., 2021). The latter result highlights the advantages of utilizing large WGS cohorts for the detection of signals in environments with unfavorable signal-to-noise ratios, as exemplified by the third and fourth report contained herein. (cf. subsection 3.3, subsection 3.4). The advancements made in the development of scalable pipelines for the analysis of raw-read data, and for statistical analysis of variant data second enabled the analysis of our large control cohort (Inova), which informs many of the statistical results. We utilized information from thousands of WGS cases to find DNMs and cDNMs, the

latter being rare mutations with a frequency of roughly 1 cDNM per offspring, and to correlate these mutations with the paternal exposure to IR. Earlier studies were smaller or had to rely on inaccurate phenotypical readouts, such as malformations, limiting their statistical power (Little et al., 2013) (Yamada et al., 2021). The results leave little doubt that cDNMs are a biomarker of paternal exposure to IR and that their rate increases proportional to the dose a father was exposed to prior to conception of the child.

The main limitations, some of which shared between them, of the studies in this dissertation are that despite multiple attempts, the detection of cDNMs and other complex biomarkers shows low specificity (e.g. 23 % in the Radar study); that de novo SVs have not been analyzed in the large cohort studies; and that dose assessments and with them the connection of cDNMs as IR response are lacking in accuracy. The low detection accuracy of cDNMs is a crucial caveat when interpreting the cDNM rates called during the Radar study. Potentially, the positive predictive value (PPV), which is only 23 %, could lead to reduced significance or the wrongful comparisons of the cDNM rates. We conducted simulation analyses that showed that the statistical tests are not sensitive to the total number of clusters, but only to the ratio found between the cohorts, under the assumption that the PPV of cDNM detection is equal in all three cohorts. This assumes that there is no difference due to the variant calling pipelines or sequencer. Our work extending DeepTrio supports this hypothesis, suggesting that the detection accuracy of cDNMs and DNMs is very high on both the HiSeq X 10 and NovaSeq 6000 sequencers, but the error signatures in the heuristic calls might differ from the deep learning approach. As part of the Radar study we validated all DNM calls bioinformatically, using Graphtyper, and studied the concordance of DNM calls made on the HiSeq and NovaSeq sequencers, and found both to be very high: (1) a PPV of $\geq$ 90% when comparing Graphtyper calls and DNM calls made from the DRAGEN pipeline; (2) 90.2% concordance of DNM calls on HiSeq and NovaSeq data (Eggertsson et al., 2017) (Eggertsson et al., 2019). Crucially, these results show that our DNM calling pipelines have high accuracy, but, even if the sequencing errors would not confound the creation of erroneous clusters, the PPV of cDNM detection under optimal circumstances is unlikely to exceed 81 %, given the aforemen-

tioned results. Earlier studies already showed that some multi-nucleotide contexts are more prone to be read falsely by the current generation of Illumina sequencers and chemistry, a fact which we confirmed. For example, no GG > TT tandem de novo mutation (two de novo mutation in direct succession) could be validated (Arora et al., 2019). Unfortunately, due to our selection of the target region (all autosomes), we could not generate reliable primers for resequencing of many target loci, since putative cDNMs often fall in low-complexity regions of the genome, which are also known regions where sequencing accuracy is degraded, further reducing cDNM calling accuracy. (Stoler and Nekrutenko, 2021) (Ma et al., 2019). We also observed true positive cDNM clusters, validated by Sanger sequencing, with very high population allele frequency in the control cohort and in gnomAD, which are explained by post-zygotic mosaicism, or mutational mechanisms other than IR (Wang et al., 2020) (Jonsson et al., 2021).

Secondly, due to technical constraints, no de novo SV calling could be performed. While published algorithms to find CNVs and SVs in WGS data exist, and were used to find the deletion in *KANSL1* in the first report (subsection 3.1), a statistical assessment is much more difficult (Rausch et al., 2012) (Layer et al., 2014) (Chen et al., 2016). Two of the WGS cohorts (Radar and CRU) were sequenced with paired-end reads of 150 bp on the NovaSeq, whereas the Inova cohort was sequenced with a read length of 100 bp. This difference and the precision of short-read SV callers prohibits any comparison between SV calls made on the control cohort and on the case cohorts, especially when considering that, in order to connect any potential finding to the IR of the fathers, we would have to call de novo SVs on the paternal allele, a task which is error-prone, even under optimal circumstances (Wang et al., 2024) (Cameron et al., 2019). Therefore, we did not assess SVs statistically in the reports contained herein.

Outside of technical and bioinformatic limitations, some doubts remain concerning the dose assessments on both exposed cohorts. For retrospective dose estimations, radar devices that have been in storage for decades had to be made operational again to measure scattered radiation dose profiles. Together with inconsistencies between the task description in German army manuals and the reports made by the radar soldiers themselves, it is plausible

that a considerable error is associated with the estimations by Dr. Schirmer, et al. (Schirmer, 2021). Concerns have also been voiced for the estimations on the CRU cohort by Yeager, et al., in particular that many of the children were born decades after the accident and potential exposure (Bazyka et al., 2020) (Chumak et al., 2021). This could lead to a skew in the statistical models which is hard to estimate. Considering the difference between estimated doses for this study and in court documents created for the soldiers, which generally assume exposures orders of magnitude larger than our report, it is likely that the dose assessment is a lower limit of the received dose in reality. Nonetheless, lacking or nonexistent dosimetry has been shown to be a hindrance for studies of the consequences of IR exposure on offspring in humans (Yeager et al., 2021) (Moorhouse et al., 2022).

Accurate long-read sequencing data or generating data of a second generation of offspring of radar soldiers could alleviate some of the aforementioned errors. Germline DNMs can be confirmed by checking the inheritance pattern of these mutations in the second generation of offspring (Sasani et al., 2019) (Jonsson et al., 2021). Of course, this extends to the phasing and detection of cDNMs as well. The increasing adoption of Oxford Nanopore and PacBio sequencing data will allow for SV detection with greater accuracy, which was not possible in this report. While SNP arrays could be used to detect very large insertion or deletion events, long-reads from PacBio or Oxford Nanopore sequencers will allow for the accurate detection of all types of SVs and potential methylation signatures of IR exposure (Glessner et al., 2021) (Wang et al., 2024) (Cameron et al., 2019). A further advantage of these techniques is that they would allow phasing a greater percentage of cDNMs, allowing for the differentiation between paternal and maternal cDNM ratios in the general population and in offspring of exposed fathers, where we would expect an increase compared to the general population. Other confounders, which our studies were not able to address were due to properties of the IR. Repair mechanisms, while one of the driver of the mutations that we do see, are working continuously to prevent damage to the germline. To our knowledge, the effect of prolonged exposure to low-dose IR on the regulation of DNA repair mechanisms has yet to be studied. DNA repair likely prevents some damage from entering the germline entirely, while large

changes overwhelming these protective systems might lead to cell death. Quantifying potential doses or the time it takes for the damaged systems to recover, a fact which has been discussed in reference to the inclusion criteria for offspring in the Radar study, could lead to improved cohort selection and allow for the creation of better statistical models for biomarkers of IR in human offspring. In our statistical models, we were also unable to observe a change of the size of clusters, a fact which could have hinted at the linear energy transfer (LET) of the underlying radiation. It is interesting to hypothesize that an increase in LET does not only lead to an increase in DSBs, but following that to an increase in the number of DNMs per cluster (Sage and Shikazono, 2017).

The techniques described herein lend themselves to extensions. The WGS analysis can be extended to other scenarios and cohorts, informing studies on other complex phenotypes, similar to the approach in the study on Covid-19 (cf. subsection 3.4). Future studies could also take this work as a basis and aggregate SNV frequencies across the whole genome (Karczewski et al., 2020) (Wang et al., 2020) (Chen et al., 2024). Our work on DeepTrio shows how any mutational signature can be adapted into this framework to be detected with good accuracy, even if only small sets of variants are available for training.

Over the course of the work on this dissertation, we established scalable pipelines for the analysis of short-read WGS and variant data, the accurate detection of DNMs and cDNMs. We used this data to associate the paternal exposure to IR with an increase in cDNMs. Despite the uncertainties in cDNM detection and dose estimation, we provided compelling evidence for the existence of transgenerational effects of IR on human DNA for the first time.

4.1 References

Arora K, Shah M, Johnson M, Sanghvi R, Shelton J, Nagulapalli K, Oschwald DM, Zody MC, Germer S, Jobanputra V, Carter J, Robine N. Deep Whole-Genome Sequencing of 3 Cancer Cell Lines on 2 Sequencing Platforms. In: Scientific Reports 2019; 9 (1): 19123. ISSN: 2045-2322. DOI: 10.1038/s41598-019-55636-3. URL: https://www.nature.com/articles/s41598-019-55636-3 (visited on 11/11/2024)

Bazyka D, Hatch M, Gudzenko N, Cahoon EK, Drozdovitch V, Little MP, Chumak V, Bakhanova E, Belyi D, Kryuchkov V, Golovanov I, Mabuchi K, Illienko I, Belayev Y, Bodelon C, Machiela MJ, Hutchinson A, Yeager M, De Gonzalez AB, Chanock SJ. Field Study of the Possible Effect of Parental Irradiation on the Germline of Children Born to Cleanup Workers and Evacuees of the Chornobyl Nuclear Accident. In: American Journal of Epidemiology 2020; 189 (12): 1451–1460. ISSN: 0002-9262, 1476-6256. DOI: 10.1093/aje/kwaa095. URL: https://academic.oup.com/aje/article/189/12/1451/5866142 (visited on 10/23/2024)

Cameron DL, Di Stefano L, Papenfuss AT. Comprehensive Evaluation and Characterisation of Short Read General-Purpose Structural Variant Calling Software. In: Nature Communications 2019; 10 (1): 3240. ISSN: 2041-1723. DOI: 10.1038/s41467-019-11146-4. URL: https://www.nature.com/articles/s41467-019-11146-4 (visited on 11/11/2024)

Chen S, Francioli LC, Goodrich JK, Collins RL, Kanai M, Wang Q, Alföldi J, Watts NA, Vittal C, Gauthier LD, Poterba T, Wilson MW, Tarasova Y, Phu W, Grant R, Yohannes MT, Koenig Z, Farjoun Y, Banks E, Donnelly S, Gabriel S, Gupta N, Ferriera S, Tolonen C, Novod S, Bergelson L, Roazen D, Ruano-Rubio V, Covarrubias M, Llanwarne C, Petrillo N, Wade G, Jeandet T, Munshi R, Tibbetts K, Genome Aggregation Database Consortium, Abreu M, Aguilar Salinas CA, Ahmad T, Albert CM, Ardissino D, Armean IM, Atkinson EG, Atzmon G, Barnard J, Baxter SM, Beaugerie L, Benjamin EJ, Benjamin D, Boehnke M, Bonnycastle LL, Bottinger EP, Bowden DW, Bown MJ, Brand H, Brant S, Brookings T, Bryant S, Calvo SE, Campos H, Chambers JC, Chan JC, Chao KR, Chapman S, Chasman DI, Chisholm R, Cho J, Chowdhury R, Chung MK, Chung WK, Cibulskis K, Cohen B, Connolly KM, Correa A, Cummings BB, Dabelea D, Danesh J, Darbar D, Darnowsky P, Denny J, Duggirala R, Dupuis J, Ellinor PT, Elosua R, Emery J, England E, Erdmann J, Esko T, Evangelista E, Fatkin D, Florez J, Franke A, Fu J, Färkkilä M, Garimella K, Gentry J, Getz G, Glahn DC, Glaser B, Glatt SJ, Goldstein D, Gonzalez C, Groop L, Gudmundsson S, Haessly A, Haiman C, Hall I, Hanis CL, Harms M, Hiltunen M, Holi MM, Hultman CM, Jalas C, Kallela M, Kaplan D, Kaprio J, Kathiresan S, Kenny EE, Kim BJ, Kim YJ, King D, Kirov G, Kooner J, Koskinen S, Krumholz HM, Kugathasan S, Kwak SH, Laakso M, Lake N, Langsford T,

Laricchia KM, Lehtimäki T, Lek M, Lipscomb E, Loos RJF, Lu W, Lubitz SA, Luna TT, Ma RCW, Marcus GM, Marrugat J, Mattila KM, McCarroll S, McCarthy MI, McCauley JL, McGovern D, McPherson R, Meigs JB, Melander O, Metspalu A, Meyers D, Minikel EV, Mitchell BD, Mootha VK, Naheed A, Nazarian S, Nilsson PM, O'Donovan MC, Okada Y, Ongur D, Orozco L, Owen MJ, Palmer C, Palmer ND, Palotie A, Park KS, Pato C, Pulver AE, Rader D, Rahman N, Reiner A, Remes AM, Rhodes D, Rich S, Rioux JD, Ripatti S, Roden DM, Rotter JI, Sahakian N, Saleheen D, Salomaa V, Saltzman A, Samani NJ, Samocha KE, Sanchis-Juan A, Scharf J, Schleicher M, Schunkert H, Schönherr S, Seaby EG, Shah SH, Shand M, Sharpe T, Shoemaker MB, Shyong T, Silverman EK, Singer-Berk M, Sklar P, Smith JT, Smith JG, Soininen H, Sokol H, Son RG, Soto J, Spector T, Stevens C, Stitziel NO, Sullivan PF, Suvisaari J, Tai ES, Taylor KD, Teo YY, Tsuang M, Tuomi T, Turner D, Tusie-Luna T, Vartiainen E, Vawter M, Wang L, Wang A, Ware JS, Watkins H, Weersma RK, Weisburd B, Wessman M, Whiffin N, Wilson JG, Xavier RJ, O'Donnell-Luria A, Solomonson M, Seed C, Martin AR, Talkowski ME, Rehm HL, Daly MJ, Tiao G, Neale BM, MacArthur DG, Karczewski KJ. A Genomic Mutational Constraint Map Using Variation in 76,156 Human Genomes. In: Nature 2024; 625 (7993): 92–100. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-023-06045-0. URL: https://www.nature.com/articles/s41586-023-06045-0 (visited on 11/11/2024)

Chen X, Schulz-Trieglaff O, Shaw R, Barnes B, Schlesinger F, Källberg M, Cox AJ, Kruglyak S, Saunders CT. Manta: Rapid Detection of Structural Variants and Indels for Germline and Cancer Sequencing Applications. In: Bioinformatics 2016; 32 (8): 1220–1222. ISSN: 1367-4803, 1367-4811. DOI: 10.1093/bioinformatics/btv710. URL: https://academic.oup.com/bioinformatics/article/32/8/1220/1743909 (visited on 11/11/2024)

Chumak V, Bakhanova E, Kryuchkov V, Golovanov I, Chizhov K, Bazyka D, Gudzenko N, Trotsuk N, Mabuchi K, Hatch M, Cahoon EK, Little MP, Kukhta T, Gonzalez AB de, Chanock SJ, Drozdovitch V. Estimation of Radiation Gonadal Doses for the American–Ukrainian Trio Study of Parental Irradiation in Chornobyl Cleanup Workers and Evacuees and Germline

Mutations in Their Offspring. In: Journal of Radiological Protection 2021; 41 (4): 764–791. DOI: 10.1088/1361-6498/abf0f4

Eggertsson HP, Jonsson H, Kristmundsdottir S, Hjartarson E, Kehr B, Masson G, Zink F, Hjorleifsson KE, Jonasdottir A, Jonasdottir A, Jonsdottir I, Gudbjartsson DF, Melsted P, Stefansson K, Halldorsson BV. Graphtyper Enables Population-Scale Genotyping Using Pangenome Graphs. In: Nature Genetics 2017; 49 (11): 1654–1660. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/ng.3964. URL: https://www.nature.com/articles/ng.3964 (visited on 10/11/2024)

Eggertsson HP, Kristmundsdottir S, Beyter D, Jonsson H, Skuladottir A, Hardarson MT, Gudbjartsson DF, Stefansson K, Halldorsson BV, Melsted P. GraphTyper2 Enables Population-Scale Genotyping of Structural Variation Using Pangenome Graphs. In: Nature Communications 2019; 10 (1): 5402. ISSN: 2041-1723. DOI: 10.1038/s41467-019-13341-9. URL: https://www.nature.com/articles/s41467-019-13341-9 (visited on 10/11/2024)

Glessner JT, Hou X, Zhong C, Zhang J, Khan M, Brand F, Krawitz P, Sleiman PMA, Hakonarson H, Wei Z. DeepCNV: A Deep Learning Approach for Authenticating Copy Number Variations. In: Briefings in Bioinformatics 2021; 22 (5): bbaa381. ISSN: 1467-5463, 1477-4054. DOI: 10.1093/bib/bbaa381. URL: https://academic.oup.com/bib/article/doi/10.1093/bib/bbaa381/6082822 (visited on 11/11/2024)

Jonsson H, Magnusdottir E, Eggertsson HP, Stefansson OA, Arnadottir GA, Eiriksson O, Zink F, Helgason EA, Jonsdottir I, Gylfason A, Jonasdottir A, Jonasdottir A, Beyter D, Steingrimsdottir T, Norddahl GL, Magnusson OT, Masson G, Halldorsson BV, Thorsteinsdottir U, Helgason A, Sulem P, Gudbjartsson DF, Stefansson K. Differences between Germline Genomes of Monozygotic Twins. In: Nature Genetics 2021; 53 (1): 27–34. ISSN: 1061-4036, 1546-1718. DOI: 10.1038/s41588-020-00755-1. URL: https://www.nature.com/articles/s41588-020-00755-1 (visited on 11/14/2024)

Karczewski KJ, Francioli LC, Tiao G, Cummings BB, Alföldi J, Wang Q, Collins RL, Laricchia KM, Ganna A, Birnbaum DP, Gauthier LD, Brand H, Solomonson M, Watts NA, Rhodes D, Singer-Berk M, England EM, Seaby EG, Kosmicki JA, Walters RK, Tashman K, Farjoun

Y, Banks E, Poterba T, Wang A, Seed C, Whiffin N, Chong JX, Samocha KE, Pierce-Hoffman E, Zappala Z, O'Donnell-Luria AH, Minikel EV, Weisburd B, Lek M, Ware JS, Vittal C, Armean IM, Bergelson L, Cibulskis K, Connolly KM, Covarrubias M, Donnelly S, Ferriera S, Gabriel S, Gentry J, Gupta N, Jeandet T, Kaplan D, Llanwarne C, Munshi R, Novod S, Petrillo N, Roazen D, Ruano-Rubio V, Saltzman A, Schleicher M, Soto J, Tibbetts K, Tolonen C, Wade G, Talkowski ME, Genome Aggregation Database Consortium, Aguilar Salinas CA, Ahmad T, Albert CM, Ardissino D, Atzmon G, Barnard J, Beaugerie L, Benjamin EJ, Boehnke M, Bonnycastle LL, Bottinger EP, Bowden DW, Bown MJ, Chambers JC, Chan JC, Chasman D, Cho J, Chung MK, Cohen B, Correa A, Dabelea D, Daly MJ, Darbar D, Duggirala R, Dupuis J, Ellinor PT, Elosua R, Erdmann J, Esko T, Färkkilä M, Florez J, Franke A, Getz G, Glaser B, Glatt SJ, Goldstein D, Gonzalez C, Groop L, Haiman C, Hanis C, Harms M, Hiltunen M, Holi MM, Hultman CM, Kallela M, Kaprio J, Kathiresan S, Kim BJ, Kim YJ, Kirov G, Kooner J, Koskinen S, Krumholz HM, Kugathasan S, Kwak SH, Laakso M, Lehtimäki T, Loos RJF, Lubitz SA, Ma RCW, MacArthur DG, Marrugat J, Mattila KM, McCarroll S, McCarthy MI, McGovern D, McPherson R, Meigs JB, Melander O, Metspalu A, Neale BM, Nilsson PM, O'Donovan MC, Ongur D, Orozco L, Owen MJ, Palmer CNA, Palotie A, Park KS, Pato C, Pulver AE, Rahman N, Remes AM, Rioux JD, Ripatti S, Roden DM, Saleheen D, Salomaa V, Samani NJ, Scharf J, Schunkert H, Shoemaker MB, Sklar P, Soininen H, Sokol H, Spector T, Sullivan PF, Suvisaari J, Tai ES, Teo YY, Tiinamaija T, Tsuang M, Turner D, Tusie-Luna T, Vartiainen E, Vawter MP, Ware JS, Watkins H, Weersma RK, Wessman M, Wilson JG, Xavier RJ, Neale BM, Daly MJ, MacArthur DG. The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans. In: Nature 2020; 581 (7809): 434–443. ISSN: 0028-0836, 1476-4687. DOI: 10.1038/s41586-020-2308-7. URL: https://www.nature.com/articles/s41586-020-2308-7 (visited on 11/11/2024)

Layer RM, Chiang C, Quinlan AR, Hall IM. LUMPY: A Probabilistic Framework for Structural Variant Discovery. In: Genome Biology 2014; 15 (6): R84. ISSN: 1474-760X. DOI: 10.1186/

gb-2014-15-6-r84. URL: https://genomebiology.biomedcentral.com/articles/10.1186/gb-2014-15-6-r84 (visited on 11/11/2024)

Little MP, Goodhead DT, Bridges BA, Bouffler SD. Evidence Relevant to Untargeted and Transgenerational Effects in the Offspring of Irradiated Parents. In: Mutation Research/Reviews in Mutation Research 2013; 753 (1): 50–67

Ma X, Shao Y, Tian L, Flasch DA, Mulder HL, Edmonson MN, Liu Y, Chen X, Newman S, Nakitandwe J, Li Y, Li B, Shen S, Wang Z, Shurtleff S, Robison LL, Levy S, Easton J, Zhang J. Analysis of Error Profiles in Deep Next-Generation Sequencing Data. In: Genome Biology 2019; 20 (1): 50. ISSN: 1474-760X. DOI: 10.1186/s13059-019-1659-6. URL: https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1659-6 (visited on 11/11/2024)

Moorhouse AJ, Scholze M, Sylvius N, Gillham C, Rake C, Peto J, Anderson R, Dubrova YE. No Evidence of Increased Mutations in the Germline of a Group of British Nuclear Test Veterans. In: Scientific Reports 2022; 12 (1): ISSN: 2045-2322. DOI: 10.1038/s41598-022-14999-w. URL: https://www.nature.com/articles/s41598-022-14999-w (visited on 09/17/2024)

Rausch T, Zichner T, Schlattl A, Stütz AM, Benes V, Korbel JO. DELLY: Structural Variant Discovery by Integrated Paired-End and Split-Read Analysis. In: Bioinformatics 2012; 28 (18): i333–i339. ISSN: 1367-4811, 1367-4803. DOI: 10.1093/bioinformatics/bts378. URL: https://academic.oup.com/bioinformatics/article/28/18/i333/245403 (visited on 11/11/2024)

Sage E, Shikazono N. Radiation-Induced Clustered DNA Lesions: Repair and Mutagenesis. In: Free Radical Biology and Medicine 2017; 107: 125–135

Sasani TA, Pedersen BS, Gao Z, Baird L, Przeworski M, Jorde LB, Quinlan AR. Large, Three-Generation Human Families Reveal Post-Zygotic Mosaicism and Variability in Germline Mutation Accumulation. In: eLife 2019; 8: e46922. ISSN: 2050-084X. DOI: 10.7554/eLife.46922. URL: https://elifesciences.org/articles/46922 (visited on 11/14/2024)

Schirmer A. Bericht S209/20: Retrospektive Dosisberechnung Röntgenstörstrahlung. Bericht S209/20. Bundesamt für Infrastruktur, Umweltschutz und Dienstleistungen der Bundeswehr (BAIUDBw)

Stoler N, Nekrutenko A. Sequencing Error Profiles of Illumina Sequencing Instruments. In: NAR Genomics and Bioinformatics 2021; 3 (1): lqab019. ISSN: 2631-9268. DOI: 10.1093/nargab/lqab019. URL: https://academic.oup.com/nargab/article/doi/10.1093/nargab/lqab019/6193612 (visited on 11/11/2024)

Wang Q, Pierce-Hoffman E, Cummings BB, Alföldi J, Francioli LC, Gauthier LD, Hill AJ, O'Donnell-Luria AH, Karczewski KJ, MacArthur DG. Landscape of Multi-Nucleotide Variants in 125,748 Human Exomes and 15,708 Genomes. In: Nature communications 2020; 11 (1): 1–13

Wang S, Lin J, Jia P, Xu T, Li X, Liu Y, Xu D, Bush SJ, Meng D, Ye K. De Novo and Somatic Structural Variant Discovery with SVision-pro. In: Nature Biotechnology 2024: ISSN: 1087-0156, 1546-1696. DOI: 10.1038/s41587-024-02190-7. URL: https://www.nature.com/articles/s41587-024-02190-7 (visited on 11/11/2024)

Yamada M, Furukawa K, Tatsukawa Y, Marumo K, Funamoto S, Sakata R, Ozasa K, Cullings HM, Preston DL, Kurttio P. Congenital Malformations and Perinatal Deaths among the Children of Atomic Bomb Survivors: A Reappraisal. In: American Journal of Epidemiology 2021; 190 (11): 2323–2333

Yeager M, Machiela MJ, Kothiyal P, Dean M, Bodelon C, Suman S, Wang M, Mirabello L, Nelson CW, Zhou W, Palmer C, Ballew B, Colli LM, Freedman ND, Dagnall C, Hutchinson A, Vij V, Maruvka Y, Hatch M, Illienko I, Belayev Y, Nakamura N, Chumak V, Bakhanova E, Belyi D, Kryuchkov V, Golovanov I, Gudzenko N, Cahoon EK, Albert P, Drozdovitch V, Little MP, Mabuchi K, Stewart C, Getz G, Bazyka D, Gonzalez AB de, Chanock SJ. Lack of Transgenerational Effects of Ionizing Radiation Exposure from the Chernobyl Accident. In: Science 2021; 372 (6543): 725–729. DOI: 10.1126/science.abg2365

5 Acknowledgment

First and foremost I want to thank my supervisor Peter Krawitz and the members of the promotion committee Christian Gilissen, Markus Nöthen and Matthias Schmid for their invaluable guidance and input during the research. I also want to thank the families that participated in the Radarstudy, and Dietmar Glaner from the BzUR who enabled the Radarstudy and its recruitment, without which many of the results herein would not have been achievable. Karl Sperling provided not only great feedback for the study, but also an unforgettable introduction to human genetics in the context of IR.

I also want to thank the colleagues at IGSB, IMBIE and at the HPC/A Lab for their support and help, in particular Alexej Knaus, Hannah Klinkhammer and Leonie Weinhold who helped me further my understanding of WGS and statistical methods to analyze that data, respectively.

Finally, I owe great thanks to Moka, Ingo and my whole family for supporting me through all of this time.