Part A – Applicant

A.1 APPLICANT

Name student (initials, first name, last name, student number):	D., Demi Gommers, 0742570
Affiliation (university/institute + department):	Utrecht University
Name first examiner:	A.N. Spaan
Affiliation (university/institute + department):	UMC Utrecht, Department of Medical Microbiology
Name second examiner:	A. Schürch
Affiliation (university/institute + department):	UMC Utrecht, Department of Medical Microbiology

Part B – Scientific proposal

B.1 BASIC DETAILS

B.2.1 Title

Investigating effective filter criteria for functional variant discovery of inborn errors of immunity in whole exome sequencing data

B.2.2 Abstract

This proposal aims to enhance the discovery of disease-causing variants in whole exome sequencing data from undiagnosed patients, addressing existing analytical challenges and proposing a standardized pipeline for increased discovery of new inborn errors of immunity (IEIs). Currently, the International Union of Immunological Societies (IUIS) Expert Committee recognizes only 485 IEIs. These IEIs are utilized as gene panels for variant detection in new patients. However, only 46% of severe immune response cases are diagnosed through Next Generation Sequencing (NGS), indicating that there are still unknown IEIs. The diverse criteria used in current NGS analysis pipelines, coupled with the absence of a universal standard, underscores the need for a standardized approach. The proposed pipeline utilizes data from GTEx, gnomAD, and dbSNP, employing an incorporated and sequential filtering process at allele, gene, and protein levels. This prioritizes variants by allele-specific filtering based on quality, location, MAF, CADD score, mutation type, and coverage, followed by gene-specific criteria, such as expression and conservation, and concludes with variant effect prediction to assess the functionality of the protein with the given variant. By prioritizing variants according to predefined criteria, this pipeline offers the potential to uncover new IEIs, allowing the in-depth characterization of the mechanisms of immune diseases and facilitating accurate diagnosis and treatment for patients.

B.2.3 Layman's summary

This research proposal aims to make it easier to find out what is wrong in the genes of people who get sick a lot due to infection. Right now, scientists know about 485 mistakes in our genes that can cause immune

system problems. But, when they use a special way to read our DNA, they can only find the cause in about 46% of the cases. This means there might be more hidden gene mistakes that we do not know about yet. The way scientists are trying to find new mistakes is not universal. Scientists are producing their own method of finding new mistakes and sometimes they do not communicate what method that is. This project suggests creating a method of analysing the DNA data, so, that all scientists can use it to find new mistakes in genes that affect the immune system. The problem is, is that these mistakes are present in a really big pile of data. Imagine you need to find one gold rice kernel in ten bathtubs of white rice kernels. To filter out most of the white rice, it is important to use a systematic approach. Via this new method, scientists will look closely at our DNA to check for mistakes and use information from genetic databases to filter out irrelevant mistakes. Then, the method checks for specific genes linked to our immune system. Finally, scientists will use prediction tools to see if the mistakes make the protein become toxic to the body. This step is like making sure the rice kernels they find matter. If this whole method works well, it could help scientists find new gene mistakes and give doctors better information to help sick people get the right treatment.

B.2.4 Keywords

Whole exome sequencing; single nucleotide variant calling; inborn errors of immunity; functional variant detection; data analysis pipeline

B.2 SCIENTIFIC PROPOSAL

B.2.5 Research topic

Inborn Errors of Immunity

Inborn Errors of Immunity (IEIs) constitute a heterogeneous group of medical conditions caused by genetic mutations in a single gene, resulting in an increased susceptibility to severe infections, immune dysregulation, autoimmunity, and malignancy. With a prevalence ranging from 1 in 1000 to 5000 individuals, these encompass diseases categorized into Primary Immunodeficiency Diseases (PIDDs) and Primary Immune Regulatory Disorders (PIRDs) (Baloh & Chong, 2023). PIDD is diagnosed when the predominant feature is recurrent severe infections, whereas PIRD is diagnosed when there is an immune dysfunction. Within these disorders, there are multiple categories of diseases, all caused by different IEIs. Given this extensive genetic diversity, IEIs pose a significant challenge for accurate diagnosis and proper patient counselling.

The International Union of Immunological Societies (IUIS) Expert Committee has recently updated the catalogue up to 485 inborn errors of immunity, involved in over 400 distinct disorders (Tangye et al., 2022). This increase is due to the improved detection of new genetic variants. The introduction of next-generation sequencing (NGS) has revolutionized the field, leading to an increased identification of IEIs (Rawat et al., 2022). In 2019, the IUIS catalogue consisted of only 430 IEIs, increasing the discovery with almost sixty genes last four years.

Within NGS there are two subtypes currently used in research and diagnostics to discover new IEIs. Whole genome sequencing (WGS) sequences all DNA of the patients, including introns and exons of all chromosomes. Whole Exome Sequencing (WES) only sequences the exonic regions of the patient's DNA, focusing only on the protein-coding regions of the genome. In diagnostic settings, WES is applied as it is more cost-effective than WGS. Although these approaches have shown promising results in identifying new disease-causing genotypes, challenges in data analysis remain, as there is no universal golden standard for variant detection.

IEIs gene panel confines diagnosis of patients with immune disorders

Despite identifying 485 IEIs, a substantial number of patients remain undiagnosed following severe immune responses. These individuals evade clinical diagnosis as no mutations are detected within the established 485 IEIs. In the early days of the WES, only 30% of individuals could be diagnosed with an immunological disorder (Lye et al., 2019; Schwarze et al., 2018). However, this figure has increased over the years due to improved analysis methods. Despite these advancements, not all patients receive a molecular diagnosis. For example,

WES analyses on children with sepsis yielded diagnoses for only 38 out of 176 patients (Borghesi et al., 2020), and in individuals with symptoms of monogenic autoinflammatory disease (AID), only 26 out of 125 patients could be molecularly diagnosed (Poker et al., 2023). Even reanalysis of undiagnosed patients with the recently updated IEI gene panel resulted in only a modest increase of five per cent of the included 94 patients (Mørup et al., 2022).

A 2019 review estimates the diagnostic yield of NGS for patient diagnosis lies within a range of 15% to 46% within mixed PIDD groups. In specific subcategories, the data shows uneven distribution but more favourable yields, ranging from 30% to 79% (Yska et al., 2019). Noteworthy considerations for this low yield include sequencing biases, where certain bases experience inadequate coverage, and challenges arising from incomplete gene sequencing due to the presence of pseudogenes or high GC content. Nevertheless, the limited number of known IEIs suggests the existence of other yet undiscovered IEIs.

Variant analysis pipelines lack overarching filter criteria for proper functional variant detection

The analysis of genetic data includes a multitude of approaches, highlighting the absence of a standardized set of criteria. Within the WES pipeline, various methods are employed on the generated data by exome sequencing. The standard WES pipeline encompasses quality filtering, adapter removal, reference genome alignment, variant calling, copy number variation (CNV) detection, and variant effect prediction. Studies performing WES lack a clear description of their WES methodology or refer only to specific guidelines without noting any deviations (Yska et al., 2019). This lack of clarity poses challenges for the reproducibility of results and the establishment of explicit criteria for variant detection.

Interpreting Variant Call Format (VCF) files generated by WES poses significant challenges due to the large number of variants present in the human genome. The vast number of rare or novel variants is estimated at six hundred thousand per person, present across an individual genome (1000 Genomes Project Consortium et al., 2015). On average, a WES analysis yields between 20,000 and 23,000 variants per individual (Kremer et al., 2018). To manage this complexity, various metrics have been developed to filter out the majority of variants. In the following sections, the current metrics used for analysing WES data, including those applied at the variant, gene, and protein levels will be discussed.

Filtering on allele level

Currently, the found variants yield from a WES analysis are annotated with their minor allele frequency (MAF) from public databases, such as gnomAD (Karczewski et al., 2020), the Combined Annotation Dependent Depletion (CADD) score (Rentzsch et al., 2019), Gene Damage Index (GDI) (Itan et al., 2015) and protein function prediction tools such as PolyPhen-2 and SIFT. Although these annotations add value in filtering out functional variants, different filtering metrics are applied in research without proper clarification. An example is the MAF, which is the frequency at which the less common allele occurs in a specific population. Recent papers are filtering the MAF between 0.2 and 0.8 for heterozygous variants (Shaomei et al., 2022), whereas <0.01 for homozygous/hemizygous and <0.0001 for heterozygous variants are used as well (Borghesi et al., 2020). This discrepancy shows an additional reason to figure out the best filter criteria for discovering new IEIs, as it depends on the type of disease being investigated. A more prevalent disease has a higher MAF, compared to a rare disease. Since this proposal is aiming for discovery, the MAF should not be set in stone.

Additionally, the CADD framework is used to integrate diverse genome annotations and scores of known single nucleotide variants (SNV) or small insertions/deletions (indel). The CADD is a method that measures the deleteriousness of a certain variant, which is correlated with its functionality and pathogenicity. The score can be used to prioritize variants based on deleteriousness (Kircher et al., 2014; Rentzsch et al., 2019). Certain variants are highly penetrant contributors to the population or are the cause of severe Mendelian disorders. The CADD score makes a distinguishment between those two. The value of the CADD score is continuous that ranges from 1 to 99. A higher value indicates a more deleterious case (Niroula & Vihinen, 2019), whether the variant is more likely to be observed or simulated. However, this raw score does not have a unit of measure, but is merely a relative score, limiting comparison between variants.

In practice, the CADD score is applied as a scaled score, which ranks all found variants based on deleteriousness within a provided group, as a set baseline. Then, the variants are binned in order of magnitude, resulting in bins of 10%, 1%, and 0.1% representing CADD-10, CADD-20, and CADD-30 etc. The negligible difference in scaled CADD scores between variants ranking at the 25th percentile and those at the 75th percentile of the raw score indicates that these percentiles do not significantly impact the interpretation of variant deleteriousness. This means that the slight variation in CADD scores between these percentiles does not substantially change the assessment of variant deleteriousness. Consequently, analysts can focus their attention on more meaningful distinctions in scaled CADD scores, such as those between variants in the top 10% and 1%, which carry greater significance in the assessment of variant deleteriousness (Kircher et al., 2014; Rentzsch et al., 2019).

Filtering on the gene level

The Gene Damage Index (GDI) is a genome-wide, gene-specific metric that provides an efficient gene-level approach for filtering out false positive variants within genes highly affected in the general population. It correlates with evolutionary pressure, protein complexity, coding sequence length, and the number of paralogs, making it a reliable metric for prioritizing variants on a gene level (Itan et al., 2015). The GDI score is based on the comparison of the CADD score of each allele to the expected CADD score for variants with similar allele frequencies. Subsequently, the findings are standardized into a homogenized Phred I-score, offering a comparative ranking of each gene against others. A lower Phred score indicates a gene with a diminished likelihood of harbouring damaging variants, while a higher score suggests the opposite. Genes with high GDI scores typically experience reduced purifying selection pressure, potentially facilitating the retention of harmful mutations. Conversely, genes with lower GDIs tend to exhibit higher conservation across species, indicating their essential roles in fundamental cellular processes such as protein synthesis, immune response, protein degradation, and gene regulation. These genes undergo stronger purifying selection, minimizing the persistence of harmful mutations compared to the average human gene (Alyousfi et al., 2019; Itan et al., 2015).

Variant effect prediction

After these metrics are applied and filtered, there are still variants left that may be disease-causing variants. With variant effect prediction (VEP) methods, the variated amino acid sequence can be assessed whether the variant causes a different effect in the protein. There are multiple methods recently reviewed (Horne & Shukla, 2022), where they presented the SIFT (sorting intolerant from tolerant) algorithm as one of the universally used VEPs. However, SIFT is outperformed in metrics of deleteriousness, pathogenicity, and molecular functionality. Meyts et al. (2016) theorized that stop mutations either up- or downstream in the gene can still result in a functional protein. They elaborate that the function of a protein can still be persevered when the stop mutation is sufficiently downstream. When this stop mutation occurs in the upstream region, reinitiation of the translation may overrule the new stop codon, or via alternative splicing the variant can be bypassed, which may result in a functional isoform (Meyts et al., 2016). However, using the CADD score alone would leave the analysis pipeline biased.

Recent developments of AlphaFold's AlphaMissense can add additional validation of the found variants. AlphaMissense is an adaptation of AlphaFold to predict missense variant pathogenicity on human and primate variants in protein sequences, based on population frequency databases. In comparison to AlphaFold, AlphaMissense does not predict the structure of molecules but instead predicts pathogenicity as scalar values. Additionally, AlphaMissense outperforms both SIFT and CADD in distinguishing likely benign or pathogenic (Cheng et al., 2023). Other methods are known that are not discussed here. The lack of ultimate methods for VEP adds a reason why a standardized discovery pipeline is necessary.

Proposing a standardized discovery pipeline for inborn errors of immunity

In summary, the lack of standardization of filtering criteria raises the risk of overlooking potentially disease-causing variants. In the dynamic landscape of IEIs, the list of identified IEIs is expanding annually (Bousfiha et al., 2020; Tangye et al., 2020, 2022). Still, more than half of the patients with severe immune responses

remain undiagnosed. The challenge for these patients lies in the fact that the detected genetic variants in patients' DNA do not align with the existing catalogue of IEIs of IUIS, emphasizing the need for a more nuanced and comprehensive approach to variant detection and classification. This proposal aims to investigate and design an improved functional variant discovery pipeline of WES data to discover new inborn errors of immunity and aid the vast majority of unexplained immune responses in patients.

Aim

Developing an innovative approach for functional variant detection in whole exome sequencing data to discover previously unknown inborn errors of immunity.

Objectives:

(1) Evaluate and select methods for functional variant detection in whole exome sequencing data of patients with IEI.
(2) Develop a pipeline for functional variant detection in whole exome sequencing data of patients with IEI.
(3) Phased validation of the developed pipeline with the diagnosed disease group, the control group, and the undiagnosed disease group.

B.2.6 Approach

Objective 1. Literary review

Goal: Evaluate and select methods for functional variant detection to discover new inborn errors of immunities in whole exome sequencing data

The primary goal is to evaluate and select methodologies for functional variant detection in whole exome sequencing data by means of a review. This review will be written focussing on WES analysis in IEIs while keeping flagship papers about current WES filter criteria in mind. These papers cover CADD (Rentzsch et al., 2019), MAF, and GDI. Moreover, standard protocols for WES analysis should be taken into account, such as the STAR Protocols standard WES analysis pipeline (Verrou et al., 2022). Additional papers will be sourced on PubMed by key search terms that could include: "whole exome sequencing," "variant calling," "variant effect prediction," "variant filtering," "inborn errors of immunity," and "variant annotation." The recent review on IEIs can be used as a guide to present the state-of-the-art findings of IEIs (Baloh & Chong, 2023). By using these search parameters, the aim is to identify the current approaches, methodologies, and filtering criteria that have emerged in recent years, ensuring that the review captures the latest advancements in the field. Apart from recent publications, the review will also draw insights from various sources used in WES analysis pipelines, such as gnomAD, dbSNP, and 1000G databases. Objective 1 will deliver the most useful metrics and their accompanying filtering criteria to be applied in the discovery of yet unknown IEIs.

Objective 2. Pipeline development

Goal: Develop a new pipeline for functional variant detection in WES data of patients with IEIs using variant prioritizing in a three-step approach.

The second objective is to develop an efficient pipeline that culminates in an interactive application, requiring only a Variant Call Format (VCF) file as input. This discovery pipeline will be designed with default filtering criteria aimed at discovering novel IEIs while allowing for flexibility to be adjusted based on specific research needs. The evaluated filter criteria from Objective 1 will be used in the pipeline. The filtering in the pipeline is based on three key steps and is implemented subsequently.

Comprising of several key steps, the pipeline refines the analysis process and prioritizes potential disease-causing variants. Phase I involves standard in-house WES with adapter removal and read quality control, followed by variant calling using GATK Haplotype Caller. Phase II employs the systematic filtering existing out of three main filtering steps. In the first step, variants undergo filtering on allele level, incorporating annotations such as MAF, CADD scores, genotype quality, and coverage on identified variants. This step

targets the elimination of standard SNPs prevalent in the population. Additionally, mutation type indications are assessed, and only the relevant variants progress to step 2: gene-level filtering.

Step 2 of the pipeline focuses on gene-level filtering, where it evaluates the mutation's location within the gene, with a specific emphasis on coding regions and splice sites. During this stage, the GDI is assessed for the genes in which the variants are present. The variants in genes with a high GDI level are then filtered out. This assessment provides valuable insights into the potential damage that the gene can do with the given variants. Simultaneously, the expression of the gene in provided tissues is examined based on the GTEx database (GTEx Portal), adding a layer of information to enhance the filtering process.

Following this gene-level evaluation, the remaining variants proceed to the final filtering step. Here, the protein function will be assessed and the potential impact of implicated variants on the protein will be evaluated. To achieve this, we will employ the innovative tool AlphaMissense, known for its capability to distinguish between benign and pathogenic missense variants. Moreover, we enhance the depth and accuracy of this final filtering process through the integration of SIFT. These additional tools contribute further layers of analysis, providing a comprehensive evaluation of the variants in terms of their potential impact on protein function.

The output of the pipeline is formatted into a tabular .txt file, ensuring clarity and simplicity in presentation. Each column in this file corresponds to a specific filtering step within the pipeline, clearly indicating whether a variant has passed the filtering step or not. This systematic arrangement allows researchers to easily navigate and comprehend the results, streamlining the identification and prioritization of novel IEIs. By presenting the data in a transparent and structured manner, we aim to empower researchers with a pipeline that not only excels in functional variant detection but also enhances the interpretability of the decision-making process on the provided variants. This approach promotes the reliability and reproducibility of genomic analyses, fostering a more accessible and collaborative research environment and improving the discovery of novel IEIs.

This objective requires are HPC environment, with a CentOS operation system, a minimal 16 physical cores, 128GB of RAM and 1TB memory. The tools that will be used in the pipeline are singularity, FastQC, MultiQC, Trim Galore!, bwa, GATK, DeepVariant, bcftools, samtools, htslib, BEDTools, UCSC tools, R, python3. All tools will be installed as image, which is a container for each tool. This installation ensures version control and makes sure that the installation of other tools is not interfering with each other. Additionally, databases need to be integrated with the HPC environment. Required databases for this tool are gnomAD v4.0, dbSNP 151, GTEx, CADD v1.7, and the human reference genome GRCh38, which can be retrieved with *wget*. The advised files and required versions are presented in **Error! Reference source not found.**.

Pipeline overview

The flowing section will highlight the pipeline's steps with additional explanation. A visual representation can be observed in *Figure 1.*

Phase I: WES Analysis

Step 1: Variant Calling of SNVs and Indels

The initial step of the pipeline involves variant calling with GATK Haplotype Caller, identifying SNVs, and small insertions/deletions (Indels) within the WES data. This process is crucial for pinpointing genetic variations that could contribute to inborn errors of immunity (IEIs).

Step 2: Quality Control

Quality control measures are implemented to ensure the reliability and accuracy of the subsequent analyses. This includes assessing the mapping quality, ensuring high-quality alignment of sequenced data to the reference genome, and evaluating the depth-of-coverage, ensuring sufficient sequencing depth for accurate variant calling.

Step 3: Incorporate Databases

The pipeline will incorporate databases from various sources, including gnomAD for MAF, dbSNP for known variants, and GTEx for gene expression per tissue, and CADD scores. The databases will be integrated in this pipeline, by saving each copy locally.

The GTEx data retrieval offers flexibility through both direct download and API access and is freely accessible for research purposes. Specifically, via GTEx Portal API (2.0.0) the expression data will be sourced through their Data Endpoints, with in specific the Median Exon Expression. Furthermore, the recently updated CADD v1.7 (Schubach et al., 2024) will be integrated to annotate the deleteriousness of identified variants, for which a local version of CADD and the scoring scripts are stored. The dbSNPs are available as a zipped .txt file including all the known variants for easy comparison. In **Error! Reference source not found.** an overview of the databases/files used for this pipeline is presented.

Table 1: Overview of the used databases and datasets in the discovery pipeline.

Database	Link
gnomAD v4.0	https://gnomad.broadinstitute.org/downloads#v4
dbSNP 151	https://hgdownload.soe.ucsc.edu/goldenPath/hg38/database/snp151.txt.gz
GTEx v8	https://storage.googleapis.com/adult-gtex/bulk-gex/v8/rna-seq/GTEx_Analysis_2017-06-05_v8_RNASeQCv1.1.9_gene_median_tpm.gct.gz
CADD v1.7	https://kircherlab.bihealth.org/download/CADD/v1.7/GRCh38/whole_genome_SNVs_inclAnno.tsv.gz
AlphaMissense	https://console.cloud.google.com/storage/browser/_details/dm_alphamissense/AlphaMissense_gene_hg38.tsv.gz

Phase II: Prioritizing and Filtering of Variants

Phase II starts after the WES analysis and the annotation of the VCF file per individual with the data presented earlier. The next part will highlight each filtering step per allele, gene and protein level. This prioritization and filtering is a guide and will be adjusted based on the outcomes of *Objective 1.* Literary review

Step 1: Allele Level Filtering

The pipeline employs filtering criteria to refine the variant selection process based on allelic variants. This step mainly focuses on excluding the noise from the VCF file. The filtering starts with excluding variants with a low genotype quality, where the cut-off is determined based on the distribution across all variants' genotype quality. Additionally, the MAF < 0.01 (1%) filtering step will focus only on the rare variants. The CADD scaled score will be applied and only the variants included in bin CADD-30 and CADD-40 will remain in the pipeline. Finally, the type of mutation will be selected, with a focus on missense and predicted loss-of-function (pLOF) variants, such as nonsense or frameshift mutations.

Step 2: Gene Level Filtering

The remaining variants will continue in the gene prioritization with a focus on the location of the variant, the GDI and the gene expression values of the GTEx database. Categorizing variants based on their location within genes, being either coding sequences (CDS) or splice sites will aid in distinguishing relevant variants. The gene expression of each gene in which a variant is present will be retrieved from the GTEx Portal. With the immune domain in mind, the gene expression tissue will elucidate whether the gene and its variant are involved in the immune system. Finally, the GDI will be determined and exclude variants in highly conserved genes – variants with a high GDI score.

Step 3: Variant Effect Prediction

The pipeline leverages protein prediction tools, including AlphaMissense and SIFT, to assess the potential impact of identified variants on protein function. These tools enhance the understanding of how genetic variations may influence the functional aspects of proteins associated with IEIs. The VEP score of both tools will be included, and when both tools comply, the variant will continue to the last phase.

Phase III: Output Format

The last step of the pipeline involves presenting the findings based on the given filtering cut-offs in a tabular, human-readable format, encompassing essential information such as chromosome (CHROM), position (POS), reference (REF), alternate (ALT) alleles, gene, gene ID, exon, allele frequency (AF), CADD score, GTEx tissue, GTEx RNAseq expression, and function prediction. This organized output streamlines the interpretation of identified variants. We propose to add additional validation tags to each filtering step to see which variant passes which filtering step. An example could be that one variant passes the filter to be potentially disease-causing by CADD, but not by AlphaMissense, then the analyst may decide based on the phenotypic characteristics of the patients whether the variant can be further investigated.

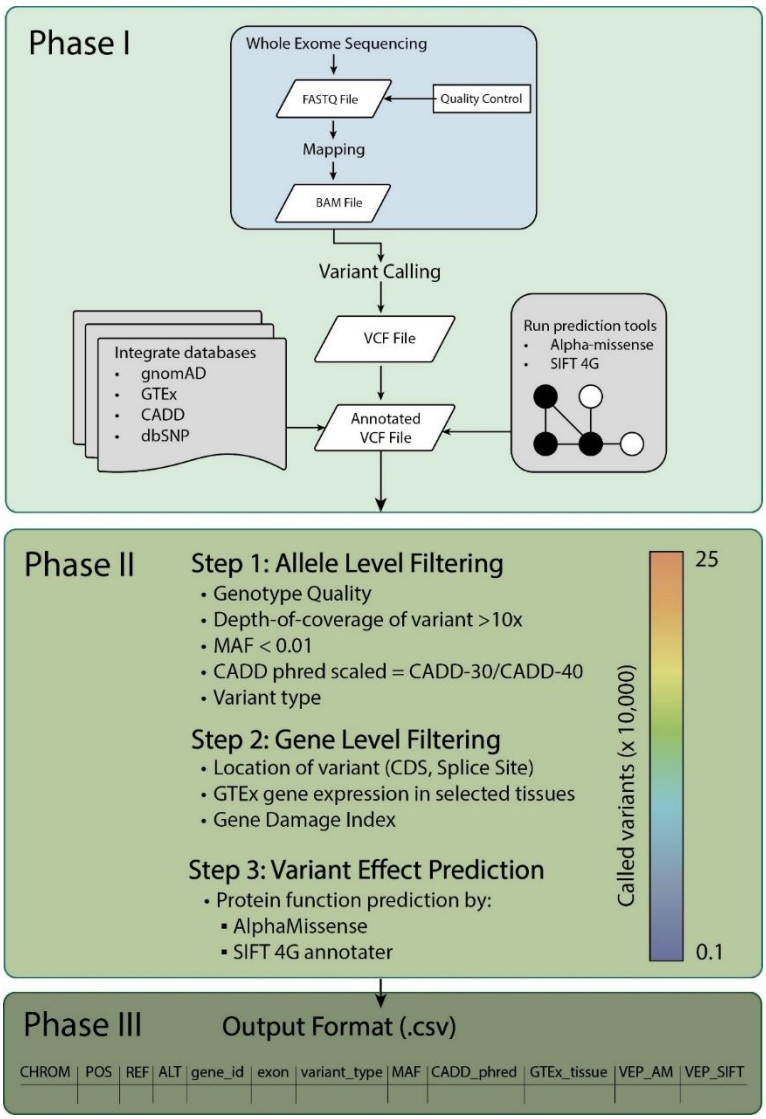

Figure 1: Proposed novel pipeline for discovery of novel functional variants in Inborn Errors of Immunity. Phase I employs standard in-house Whole Exome Sequence analysis, followed by variant calling performed by GATK Haplotype Caller. Databases will be incorporated and variant effect prediction will be performed on the available variants. Phase II entails the prioritization and filtering steps existing out of three main steps, per allele, gene and protein level. Phase III presents the aimed human-readable output format for downstream analysis.

Objective 3. Validate pipeline with positive and negative controls

Goal: To validate the efficacy and accuracy of the developed pipeline using a combination of internal and external datasets.

The validation process of the discovery pipeline will be divided into three distinct phases, each dedicated to validating the detection of genuine variants, the absence of variants, and the discovery of novel variants (*Table 2*). Phase I focuses on examining VCFs originating from WES analysis of ten in-house patients diagnosed with IEIs, aiming to detect the same ten IEIs as positive control samples. This phase serves to identify real IEIs and establish a benchmark for detection accuracy. Moving to Phase II, the validation extends to 1000G WGS data with 30X coverage, devoid of any IEI history (Byrska-Bishop et al., 2022), providing robust negative controls for accurate differentiation between affected and unaffected individuals. Phase III introduces a shift towards exploration, involving 78 VCFs from undiagnosed patients exhibiting severe immune responses, thereby expanding the spectrum beyond the 485 IEIs of the IUIS gene panel. This step underscores the pipeline's capacity for uncovering potentially novel genetic markers or variants. Similarly, Phase IV delves deeper, by analysing 15 in-house whole blood samples of undiagnosed patients. These samples will go through the entire pipeline, including WES analysis, to enhance the discovery phase before practical application. Successful outcomes in each phase are indispensable for progressing through subsequent validation stages, culminating in the pipeline's deployment in research settings.

Furthermore, a comprehensive statistical analysis will enhance the validation process. Receiver Operating Characteristic (ROC) curves will be employed to evaluate the sensitivity and specificity of the pipeline using the provided control samples, offering insights into its overall performance. The calculation of the Area Under the Curve (AUC) with a 95% confidence interval will provide a quantitative measure of the pipeline's discriminative ability during Phase I and II. Additionally, the assessment of false positive and false negative rates will offer valuable information on the precision and recall of the pipeline's predictions, ensuring a thorough evaluation of its effectiveness. This statistical analysis will complement the validation efforts, offering a comprehensive assessment of the pipeline's reliability and accuracy.

Table 2: Sample validation overview

Phase	Samples	Required outcome to continue
Phase I	WES analysis, VCFs of 10 patients, diagnosed with IEIs	Detecting 10 IEIs
Phase II	1000G WGS 30X VCFs (exome variants only)	No IEIs are found
Phase III	78 VCFs of undiagnosed patients with severe immune response	Discovery
Phase IV	15 Whole Blood Samples of undiagnosed patients with severe immune response	Discovery

B.2.7 Feasibility / Risk assessment

Risk assessment

The implementation of the proposed research pipeline introduces concrete risks that demand attention for successful execution and reliable analysis. Firstly, the OS version on which the pipeline runs poses a tangible risk of compatibility issues, potentially impacting performance and reliability. This risk is actively mitigated by selecting an OS version widely supported by computational tools and ensuring consistent use across the research team to enhance overall compatibility. Secondly, the risk of software compatibility and version control arises, where variations in software versions across bioinformatic tools can introduce inconsistencies. These risks pose a threat to the overall consistency and reproducibility of the results of the pipeline. By

regular updates and via version control of the filtering pipeline, the aim is to systematically track changes in the pipeline, mitigating these risks and ensuring reproducibility. Thirdly, this pipeline will rely on the public genetic variant databases, which poses a risk when these databases are no longer updated, or curated. When these databases are not available, or updated, the pipeline might no longer work in future settings. To mitigate this risk, it is essential to only include curated databases from well-known institutes, such as gnomAD, GTEx, and dbSNP (NCBI). Furthermore, the databases will be stored locally to prevent any unforeseen updates to online databases, making it able to run the pipeline when the databases are stored locally at any time. However, this mitigation requires a large amount of memory, which is something to keep in mind. At last, from the perspective of IEI's heterogeneity, the genetic heterogeneity of IEIs may lead to challenges in establishing universally applicable filter criteria. This challenge can be solved by actively changing filter criteria and making the pipeline include state-of-the-art knowledge on the best criteria to use. Furthermore, this pipeline can anticipate the heterogeneity of the sample, as the filter criteria can be adjusted to specific cases.

Feasibility

The project requires in total of one and a half years to fully investigate the best filter criteria, develop the new pipeline, and validate it according to the presented process. A rough time sketch can be observed to indicate the projects timeline (*Table 3*).

Concerning feasibility, this project requires an experienced researcher, i.e. Post-doc or analyst with prior knowledge of bioinformatic databases such as gnomAD, AlphaMissense, and GTEx. Experience in genetic data handling and experience in programming languages such as bash, python, R, Java and experience in working in HPC environment is required. Additionally, a medium understanding of the immune system and its pathways would aid the researcher in understanding the scope of the project. It is important that the discovery pipeline is a standalone tool/pipeline and can be used by other researchers after proper validation, where the filter criteria are set in a default setting but can be adjusted when needed. With these criteria in mind, the feasibility of this project is sound.

Table 3: Timeline overview of the project.

	Year 1				Year 2 (half)	
Objectives	Q1	Q2	Q3	Q4	Q1	Q2
Objective 1: Literary Review						
Objective 2: Pipeline development						
Objective 3: Pipeline validation				Phase I	Phase II + III	Phase III + IV

B.2.8 Scientific (a) and societal (b) impact

Scientific Impact: A standard pipeline for functional variant detection may discover novel IEIs

The presented pipeline presents an opportunity to establish a new standard in genomic variant analysis, revolutionizing the field of IEIs. By streamlining the discovery of novel IEIs, the pipeline holds potential to contribute significantly to the scientific community's understanding of IEIs. Moreover, the pipeline's capacity to potentially extend the list of involved genes in IEIs adds a valuable dimension to the understanding of immunological diseases. By increasing the list of IEIs, additional research can be conducted, aiding our understanding of immunological pathways, and fostering new avenues for targeted research and therapeutic interventions.

Societal Impact: The discovery of more IEIs may improve patient diagnostics and aid patients further in genetic counselling, and access to appropriate therapies.

The discovery of additional IEIs through our new pipeline can be significant for patient diagnostics and care. The pipeline aims to discover new IEIs, and with that, it increases the list of IEIs and their gene panels. This can lead to improved diagnostics and can lead to earlier intervention, positively influencing patient prognoses

and enhancing their overall quality of life. In the medical domain, the ability to screen patients earlier for IEIs can potentially identify conditions before they manifest, facilitating timely and targeted therapeutic interventions. Furthermore, the genetic diagnosis provided by the pipeline can significantly aid patients in genetic counselling, enabling informed family planning and prenatal diagnosis. The pipeline's societal impact extends to ensuring patients have access to appropriate therapeutic options, marking a crucial step forward in personalized medicine and patient-centred care.

B.2.9 Ethical considerations

Ethical Considerations and Informed Consent

Overseeing genetic data raises ethical concerns related to privacy, consent, and potential implications for patients, especially in cases of undiagnosed conditions. Adherence to rigorous ethical guidelines is imperative to address these concerns responsibly. The project will be held within the University Medical Centre Utrecht, which has a department of Bioethics & Health Humanities, which engages in the Research Ethics Committee in the UMC Utrecht (NedMec), whom we can approach for ethical consultation. Additionally, guidelines for ethical practices, informed consent, and data privacy will be designed in collaboration with the Research Ethics Committee.

Data Management Plan

The project will prioritize the development of a clear data management plan, acknowledging the importance of ethical and legal considerations. This plan will be crafted to align with the FAIR principles, emphasizing the data's Findability, Accessibility, Interoperability, and Reusability. The latter is of most importance, as the aim of this project is to develop a universal pipeline for the discovery of IEIs. Moreover, the proposal will be designed to comply with the regulations set by the General Data Protection Regulation (GDPR) of the European Union, ensuring the secure and responsible handling of personal and sensitive information. Standardized metadata practices, access controls, and comprehensive documentation will be elaborated on to elevate the project's transparency and collaborative potential and with that improve further discovery of new IEIs.

B.2.10 Literature/references

1000 Genomes Project Consortium, Auton, A., Brooks, L. D., Durbin, R. M., Garrison, E. P., Kang, H. M., Korbel, J. O., Marchini, J. L., McCarthy, S., McVean, G. A., & Abecasis, G. R. (2015). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Alyousfi, D., Baralle, D., & Collins, A. (2019). Gene-specific metrics to facilitate identification of disease genes for molecular diagnosis in patient genomes: A systematic review. *Briefings in Functional Genomics*, *18*(1), 23–29. https://doi.org/10.1093/bfgp/ely033

Baloh, C. H., & Chong, H. (2023). Inborn Errors of Immunity. *Primary Care*, *50*(2), 253–268. https://doi.org/10.1016/j.pop.2022.12.001

Borghesi, A., Trück, J., Asgari, S., Sancho-Shimizu, V., Agyeman, P. K. A., Bellos, E., Giannoni, E., Stocker, M., Posfay-Barbe, K. M., Heininger, U., Bernhard-Stirnemann, S., Niederer-Loher, A., Kahlert, C. R., Natalucci, G., Relly, C., Riedel, T., Kuehni, C. E., Thorball, C. W., Chaturvedi, N., … Schlapbach, L. J. (2020). Whole-exome Sequencing

for the Identification of Rare Variants in Primary Immunodeficiency Genes in Children With Sepsis: A Prospective, Population-based Cohort Study. *Clinical Infectious Diseases*, *71*(10), e614–e623. https://doi.org/10.1093/cid/ciaa290

Bousfiha, A., Jeddane, L., Picard, C., Al-Herz, W., Ailal, F., Chatila, T., Cunningham-Rundles, C., Etzioni, A., Franco, J. L., Holland, S. M., Klein, C., Morio, T., Ochs, H. D., Oksenhendler, E., Puck, J., Torgerson, T. R., Casanova, J.-L., Sullivan, K. E., & Tangye, S. G. (2020). Human Inborn Errors of Immunity: 2019 Update of the IUIS Phenotypical Classification. *Journal of Clinical Immunology*, *40*(1), 66–81. https://doi.org/10.1007/s10875-020-00758-x

Byrska-Bishop, M., Evani, U. S., Zhao, X., Basile, A. O., Abel, H. J., Regier, A. A., Corvelo, A., Clarke, W. E., Musunuri, R., Nagulapalli, K., Fairley, S., Runnels, A., Winterkorn, L., Lowy, E., Human Genome Structural Variation Consortium, Paul Flicek, null, Germer, S., Brand, H., Hall, I. M., … Zody, M. C. (2022). High-coverage whole-genome sequencing of the expanded 1000 Genomes Project cohort including 602 trios. *Cell*, *185*(18), 3426-3440.e19. https://doi.org/10.1016/j.cell.2022.08.004

Cheng, J., Novati, G., Pan, J., Bycroft, C., Žemgulytė, A., Applebaum, T., Pritzel, A., Wong, L. H., Zielinski, M., Sargeant, T., Schneider, R. G., Senior, A. W., Jumper, J., Hassabis, D., Kohli, P., & Avsec, Ž. (2023). Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science (New York, N.Y.)*, *381*(6664), eadg7492. https://doi.org/10.1126/science.adg7492

Horne, J., & Shukla, D. (2022). Recent Advances in Machine Learning Variant Effect Prediction Tools for Protein Engineering. *Industrial & Engineering Chemistry Research*, *61*(19), 6235–6245. https://doi.org/10.1021/acs.iecr.1c04943

Itan, Y., Shang, L., Boisson, B., Patin, E., Bolze, A., Moncada-Vélez, M., Scott, E., Ciancanelli, M. J., Lafaille, F. G., Markle, J. G., Martinez-Barricarte, R., de Jong, S. J., Kong, X.-F., Nitschke, P., Belkadi, A., Bustamante, J., Puel, A., Boisson-Dupuis, S., Stenson, P. D., … Casanova, J.-L. (2015). The human gene damage index as a gene-level approach to prioritizing exome variants. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(44), 13615–13620. https://doi.org/10.1073/pnas.1518646112

Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., Collins, R. L., Laricchia, K. M., Ganna, A., Birnbaum, D. P., Gauthier, L. D., Brand, H., Solomonson, M., Watts, N. A., Rhodes, D., Singer-Berk, M., England,

E. M., Seaby, E. G., Kosmicki, J. A., … MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature*, *581*(7809), 434–443. https://doi.org/10.1038/s41586-020-2308-7

Kircher, M., Witten, D. M., Jain, P., O'Roak, B. J., Cooper, G. M., & Shendure, J. (2014). A general framework for estimating the relative pathogenicity of human genetic variants. *Nature Genetics*, *46*(3), 310–315. https://doi.org/10.1038/ng.2892

Kremer, L. S., Wortmann, S. B., & Prokisch, H. (2018). "Transcriptomics": Molecular diagnosis of inborn errors of metabolism via RNA-sequencing. *Journal of Inherited Metabolic Disease*, *41*(3), 525–532. https://doi.org/10.1007/s10545-017-0133-4

Lye, J. J., Williams, A., & Baralle, D. (2019). Exploring the RNA Gap for Improving Diagnostic Yield in Primary Immunodeficiencies. *Frontiers in Genetics*, *10*, 1204. https://doi.org/10.3389/fgene.2019.01204

Meyts, I., Bosch, B., Bolze, A., Boisson, B., Itan, Y., Belkadi, A., Pedergnana, V., Moens, L., Picard, C., Cobat, A., Bossuyt, X., Abel, L., & Casanova, J.-L. (2016). Exome and genome sequencing for inborn errors of immunity. *The Journal of Allergy and Clinical Immunology*, *138*(4), 957–969. https://doi.org/10.1016/j.jaci.2016.08.003

Mørup, S. B., Nazaryan-Petersen, L., Gabrielaite, M., Reekie, J., Marquart, H. V., Hartling, H. J., Marvig, R. L., Katzenstein, T. L., Masmas, T. N., Lundgren, J., Murray, D. D., Helleberg, M., & Borgwardt, L. (2022). Added Value of Reanalysis of Whole Exome- and Whole Genome Sequencing Data From Patients Suspected of Primary Immune Deficiency Using an Extended Gene Panel and Structural Variation Calling. *Frontiers in Immunology*, *13*, 906328. https://doi.org/10.3389/fimmu.2022.906328

Niroula, A., & Vihinen, M. (2019). How good are pathogenicity predictors in detecting benign variants? *PLoS Computational Biology*, *15*(2), e1006481. https://doi.org/10.1371/journal.pcbi.1006481

Poker, Y., von Hardenberg, S., Hofmann, W., Tang, M., Baumann, U., Schwerk, N., Wetzke, M., Lindenthal, V., Auber, B., Schlegelberger, B., Ott, H., von Bismarck, P., Viemann, D., Dressler, F., Klemann, C., & Bergmann, A. K. (2023). Systematic genetic analysis of pediatric patients with autoinflammatory diseases. *Frontiers in Genetics*, *14*, 1065907. https://doi.org/10.3389/fgene.2023.1065907

Rawat, A., Sharma, M., Vignesh, P., Jindal, A. K., Suri, D., Das, J., Joshi, V., Tyagi, R., Sharma, J., Kaur, G., Lau, Y.-L., Imai, K., Nonoyama, S., Lenardo, M., & Singh, S. (2022). Utility of targeted next generation sequencing for inborn

errors of immunity at a tertiary care centre in North India. *Scientific Reports*, *12*(1), 10416. https://doi.org/10.1038/s41598-022-14522-1

Rentzsch, P., Witten, D., Cooper, G. M., Shendure, J., & Kircher, M. (2019). CADD: Predicting the deleteriousness of variants throughout the human genome. *Nucleic Acids Research*, *47*(D1), D886–D894. https://doi.org/10.1093/nar/gky1016

Schubach, M., Maass, T., Nazaretyan, L., Röner, S., & Kircher, M. (2024). CADD v1.7: Using protein language models, regulatory CNNs and other nucleotide-level scores to improve genome-wide variant predictions. *Nucleic Acids Research*, *52*(D1), D1143–D1154. https://doi.org/10.1093/nar/gkad989

Schwarze, K., Buchanan, J., Taylor, J. C., & Wordsworth, S. (2018). Are whole-exome and whole-genome sequencing approaches cost-effective? A systematic review of the literature. *Genetics in Medicine*, *20*(10), 1122–1130. https://doi.org/10.1038/gim.2017.247

Shaomei, W., Yongbin, P., Daiyue, Y., Zhaorong, H., Huirong, Y., Nan, L., Huanbin, L., Yuzhu, L., & Kai, W. (2022). Whole exome sequencing applied to 42 Han Chinese patients with posterior hypospadias. *Steroids*, *184*, 109041. https://doi.org/10.1016/j.steroids.2022.109041

Tangye, S. G., Al-Herz, W., Bousfiha, A., Chatila, T., Cunningham-Rundles, C., Etzioni, A., Franco, J. L., Holland, S. M., Klein, C., Morio, T., Ochs, H. D., Oksenhendler, E., Picard, C., Puck, J., Torgerson, T. R., Casanova, J.-L., & Sullivan, K. E. (2020). Human Inborn Errors of Immunity: 2019 Update on the Classification from the International Union of Immunological Societies Expert Committee. *Journal of Clinical Immunology*, *40*(1), 24–64. https://doi.org/10.1007/s10875-019-00737-x

Tangye, S. G., Al-Herz, W., Bousfiha, A., Cunningham-Rundles, C., Franco, J. L., Holland, S. M., Klein, C., Morio, T., Oksenhendler, E., Picard, C., Puel, A., Puck, J., Seppänen, M. R. J., Somech, R., Su, H. C., Sullivan, K. E., Torgerson, T. R., & Meyts, I. (2022). Human Inborn Errors of Immunity: 2022 Update on the Classification from the International Union of Immunological Societies Expert Committee. *Journal of Clinical Immunology*, *42*(7), 1473–1507. https://doi.org/10.1007/s10875-022-01289-3

Verrou, K.-M., Pavlopoulos, G. A., & Moulos, P. (2022). Protocol for unbiased, consolidated variant calling from whole exome sequencing data. *STAR Protocols*, *3*(2), 101418. https://doi.org/10.1016/j.xpro.2022.101418

Yska, H. A. F., Elsink, K., Kuijpers, T. W., Frederix, G. W. J., Van Gijn, M. E., & Van Montfrans, J. M. (2019). Diagnostic Yield of Next Generation Sequencing in Genetically Undiagnosed Patients with Primary Immunodeficiencies: A Systematic Review. *Journal of Clinical Immunology*, *39*(6), 577–591. https://doi.org/10.1007/s10875-019-00656-x