LUND UNIVERSITY

Faculty of Engineering (LTH) | Master's Programme in Biotechnology

From Legacy Databases to Modern Standards:

A Bioinformatics Pipeline for Migrating Immunodeficiency Variant Data to LOVD 3.0 with GRCh38 Standardisation

Author: Khoi Nguyen Ngoc

Supervisor: Professor Mauno Vihinen

Examiner: Ghjuvan Micaelu Grimaud

Department of Experimental Medical Science

Lund University, Sweden

2026

Repository: github.com/knn1996/Migrating-BMC-IDBases-to-LOVD

Abstract

Primary immunodeficiency diseases (PIDs) are a heterogeneous group of inherited immune system disorders caused by pathogenic variants in over 400 genes, collectively affecting an estimated one in 10,000 individuals worldwide. The IDbases collection, maintained on the legacy MUTbase platform at Lund University, represents one of the most comprehensively curated repositories of such variants, cataloguing disease-causing variants across 136 immunodeficiency-related genes with 6,259 patient records accumulated over more than three decades of expert curation. Despite their scientific value, these databases remain in a format incompatible with current genomic standards: variant coordinates are anchored to obsolete partial ENA clone reference sequences, nomenclature does not conform to Human Genome Variation Society (HGVS) recommendations, and the data cannot be integrated with modern interoperable platforms.

This thesis presents a systematic seven-step bioinformatics pipeline for migrating the IDbases data into the Leiden Open Variation Database 3.0 (LOVD 3.0), the leading open-source platform for locus-specific variant databases. The pipeline extracted variants from all 136 gene databases, resolved reference sequence identities through an empirical offset algorithm, lifted over chromosomal coordinates to GRCh38, and normalised all variant descriptions to HGVS nomenclature through three parallel Mutalyzer 3 annotation tracks. Track A (NG_IDRefseq genomic route) achieved a 99.3% normalisation success rate (3,582/3,606 variants). Track B (NM_MANE) achieved 94.1% (4,496/4,775). Track C (NM_IDRefseq) served as a tertiary traceability fallback at 59.5% (2,862/4,811).

A central finding of this work is that IDbases IDRefSeq sequences are partial ENA clone fragments contained within full LRG NG_ reference sequences - a previously undocumented structural relationship that causes systematic out-of-bounds coordinate failures for variants added after the original database construction date. The empirical offset algorithm developed to resolve this achieved a match rate of ≥90% for 92 of 94 testable genes (genes where an IDRefSeq FASTA was available) in the genomic track. Following priority-based merging of the three tracks, 7,776 patient–variant records were resolved and, after deduplication, correspond to 2,240 distinct variants across 115 genes, linked to 2,729 patients (5,773 patient–variant pairs). A further 698 distinct variants remained unresolved, of which roughly three-quarters are automatically rescuable, giving an automated resolution rate of 76.2% with an estimated addressable ceiling near 94%. Systematic data quality issues uncovered during the migration - including strand orientation inconsistencies in 43 reverse-strand genes, missing

insertion sequences, legacy IVS intronic notation, and transcript version obsolescence - are catalogued as findings to inform future variant database design.

The pipeline is implemented as a series of documented, modular Python scripts and is designed for reuse in similar legacy database migration projects. The work makes a large, expertly curated immunodeficiency variant dataset available in a modern, interoperable format aligned with GRCh38 and MANE (Matched Annotation from the NCBI and EMBL-EBI) Select transcript standards.

Keywords: primary immunodeficiency, variant database migration, HGVS nomenclature, LOVD 3.0, GRCh38, Mutalyzer, IDbases, MUTbase, locus-specific database, bioinformatics pipeline, reference sequence standardisation

Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Definition
ACMG	American College of Medical Genetics and Genomics
BED	Browser Extensible Data (genomic coordinate file format)
BLAST	Basic Local Alignment Search Tool
CGD	Chronic Granulomatous Disease
CDS	Coding DNA Sequence
ENA	European Nucleotide Archive
GRCh37/38	Genome Reference Consortium Human Build 37/38
HGVS	Human Genome Variation Society
HPC	High-Performance Computing
HVP	Human Variome Project
LSDB	Locus-Specific DataBase
LOVD	Leiden Open Variation Database
LRG	Locus Reference Genomic sequence
MANE	Matched Annotation from NCBI and EMBL-EBI
MUTbase	Distributed mutation database management system (legacy)
NG_	NCBI RefSeqGene genomic accession prefix
NM_	NCBI RefSeq mRNA transcript accession prefix
OOB	Out-of-bounds (variant position exceeds reference span)
PID	Primary Immunodeficiency Disease
SCID	Severe Combined Immunodeficiency

Abbreviation	Definition
SNV	Single Nucleotide Variant
TAPBP	Tapasin (also TAPA-1 binding protein)
VCF	Variant Call Format
XLA	X-linked Agammaglobulinemia

1. Introduction

1.1 The Increasing of Primary Immunodeficiency Diseases

Primary immunodeficiency diseases are a clinically and genetically heterogeneous group of inherited disorders characterised by intrinsic defects of the immune system. Unlike secondary immunodeficiencies, which arise from external causes such as infection or chemotherapy, PIDs are caused by germline variants that impair the development, function, or regulation of immune system components. Patients typically present with increased susceptibility to recurrent and severe infections caused by bacterial, viral, f*UNG*al, or opportunistic pathogens that would ordinarily be contained by a normal immune system. Beyond infections, many PID patients suffer from autoimmune manifestations, inflammatory complications, and elevated cancer risk, particularly haematological malignancies.

The International Union of Immunological Societies (IUIS) Expert Committee issues updated classifications of PIDs approximately every two years. The most recent update [1,2] catalogues over 450 distinct PID entities distributed across ten functional categories, including combined immunodeficiencies, predominantly antibody deficiencies, congenital defects of phagocyte number and function, complement deficiencies, and disorders of immune dysregulation [3]. Causative variants have been identified in more than 400 genes distributed across all chromosomes, with many PID subtypes displaying genetic heterogeneity that ranges from monogenic to complex polygenic architectures [4]. The collective estimated prevalence of PIDs is approximately 1 in 10,000 live births, though this figure almost certainly underestimates the true burden, as many rare forms remain significantly underdiagnosed due to phenotypic overlap with common conditions and limited access to molecular diagnostics in lower-income settings [5]. The utility of advanced sequencing in resolving diagnostically elusive PID cases has been demonstrated even for novel entities: Bradshaw et al. [6] used whole exome sequencing to diagnose X-linked moesin-associated immunodeficiency in a patient undiagnosed for over three decades, illustrating how standardised variant curation and molecular tools together can close diagnostic gaps. Broader genomic approaches have been shown to substantially improve PID diagnostic yield across heterogeneous Mendelian presentations [7].

The molecular pathology of PIDs spans every component of the immune system. Severe combined immunodeficiencies (SCIDs) affect both T- and B-cell development and represent the most life-threatening forms, typically presenting in infancy. X-linked agammaglobulinemia (XLA), caused by variants in *BTK*, abolishes B-cell maturation and circulating

immunoglobulins. Chronic granulomatous disease (CGD), caused by variants in *CYBB* and *NCF1/2/CYBA*, impairs the neutrophil oxidative burst required to kill catalase-positive bacteria and f*UNG*i. Complement deficiencies render patients vulnerable to encapsulated bacterial pathogens. These examples illustrate the diversity of pathological mechanisms, each requiring specialised clinical management and each generating distinct variant data that must be carefully catalogued to enable genotype–phenotype correlations and guide therapeutic decisions.

1.2 The Critical Role of Variant Databases in PID Diagnosis and Research

Accurate molecular diagnosis is central to the clinical management of PIDs. For many PID subtypes, definitive diagnosis requires the identification of a causative variant in the appropriate gene, as clinical features and immunological parameters alone are often insufficient to distinguish between conditions with overlapping presentations. Variant databases - structured repositories of known disease-causing variants - play an indispensable role in this diagnostic process: a variant observed in a new patient can be compared against the curated database to assess its prior clinical association, functional impact, and population frequency.

Beyond diagnosis, variant databases enable research at a population level. Aggregated variant data from multiple patients allows the identification of mutational hotspots, correlation of genotype with clinical phenotype, assessment of founder effects in specific populations, and evaluation of therapeutic targets.[36] Analyses of variant databases at scale have shown that a substantial proportion of catalogued variants remain unresolved or underdiagnosed; Seidizadeh et al. [8] demonstrated through large-scale mining of LOVD and HGMD (commercial Human Gene Mutation Database) [37] that the true prevalence of disease-causing VWF alleles far exceeds clinical estimates, suggesting comprehensive databases are essential for capturing the full mutational landscape of inherited disease. The IDbases collection, with records spanning over three decades of PID genetics research, represents an irreplaceable repository of such information. However, its value has been increasingly constrained by the technical limitations of the legacy MUTbase platform on which it was built – essentially the lack of funding for a full-time curator – that prevent modern bioinformatics tools and clinical interpretation pipelines from accessing the data in a standardised, machine-readable format. LSDBs have become the primary vehicle for specialised variant curation. LOVD was developed as a platform-independent LSDB-in-a-Box framework aligned with HGVS standards [9], extended in v.2.0 with multi-gene patient support [10], and redesigned as LOVD 3.0, a genome-centred platform whose 56 registered public instances together hold around one billion variant observations from

1.5 million individuals, with 42 of them forming a federated network that shares some three million unique variants [11]. The value of this platform for variant classification has been demonstrated at scale: Yin et al. [12] applied gene-specific ACMG/AMP criteria to over 10,000 APC variants across ClinVar and LOVD, reducing variants of uncertain significance by 37%. The development of CASRdb further illustrates how disease-specific databases can substantially increase variant coverage beyond general repositories [13].

1.3 The IDbases Collection and Its Technical Debt

The IDbases collection was established by Professor Mauno Vihinen and colleagues at the University of Tampere (later Lund University) as a distributed network of gene-specific immunodeficiency databases [14]. Beginning with the BTKbase for X-linked agammaglobulinemia in 1994 [15], the collection expanded progressively to cover all major immunodeficiency genes, complemented by the broader ImmunoDeficiency Resource (IDR) knowledge framework [16,17]. Each database was managed through the MUTbase software system [18], which provided a distributed, web-based platform for recording patient-level variant entries comprising DNA, RNA, and protein level descriptions, along with clinical information and literature references. The comprehensive published description of the collection by Piirilä et al. [19] documented coverage of 107 genes with over 4,140 public patient entries and presented the first exhaustive statistical analysis of the mutation landscape across all known immunodeficiency genes.

Despite this valuable curation history, the IDbases databases have accumulated substantial technical debt over the intervening decades. The core problem is architectural: variant coordinates are anchored to IDRefSeq sequences, which are partial ENA clone sequences that were selected as local reference sequences at the time of individual database construction - in most cases during the late 1990s and early 2000s, prior to the widespread adoption of genome assembly-anchored variant annotation. These clone sequences are not versioned in any standard registry, are not aligned to any genome assembly, and in many cases represent only a portion of the gene-containing genomic interval. As a result, variants added to a database after its initial construction may extend beyond the boundaries of the original clone, making their coordinates unresolvable relative to that reference. Furthermore, the nomenclature used in IDbases predates full HGVS compliance, mixing legacy IVS intronic notation, protein-level descriptions, and partially compliant coding DNA descriptions in a manner that modern validation tools cannot directly process. Mutalyzer and VariantValidator were developed as nomenclature checkers to

address this class of problem, providing automated correction of variant descriptions against standard reference sequences, and had become an essential quality control step in LSDB curation and migration [20].

The practical consequence is that the IDbases data cannot be integrated with current clinical genomics platforms, cannot be queried programmatically through modern APIs, and cannot be cross-referenced with contemporary variant resources such as ClinVar, gnomAD, or the LOVD network. Migration to a standards-compliant platform is therefore not merely a technical convenience but a scientific necessity to preserve and extend the value of this uniquely curated dataset. Such migrations are not without precedent: Kahn et al. [21] described migrating a multi-institutional clinical research data warehouse to a public cloud, identifying eight categories of challenges including legacy system limitations and complex governance - issues broadly analogous to those encountered when migrating genomic variant databases between platforms with fundamentally different data models.

1.4 Objectives

The primary aim of this thesis is to develop and implement a systematic bioinformatics pipeline for migrating the IDbases immunodeficiency variant databases from the legacy MUTbase platform to LOVD 3.0, with all variant coordinates standardised to GRCh38 and all variant descriptions normalised to HGVS nomenclature. The specific objectives are:

1. To extract and catalogue all variant records from 136 IDbases gene databases.
2. To identify the IDRefSeq reference sequences, determine their provenance, and resolve their genomic relationship to current LRG/NG_ reference sequences through sequence alignment and an empirical offset algorithm.
3. To convert IDRefSeq-based variant coordinates to GRCh38 chromosomal coordinates via BLAST-based genome alignment, BED file generation, and the UCSC LiftOver tool [22].
4. To generate and validate HGVS-normalised variant descriptions at the genomic (g.), coding (c.), RNA (r.), and protein (p.) levels using the Mutalyzer 3 REST API across three parallel annotation tracks incorporating MANE Select transcripts.
5. To assess data quality systematically across all 136 databases and document unresolvable entries, strand orientation issues, missing sequence data, and nomenclature anomalies as findings for the variant database community.

6. To implement a priority-based track merge strategy that maximises MANE Select transcript compliance while preserving historical traceability.
7. To provide a documented, modular Python pipeline reusable by other research groups facing similar legacy LSDB migration challenges.

2. Background and Literature Review

2.1 Primary Immunodeficiencies: Classification and Molecular Mechanisms

The current IUIS classification of PIDs recognises ten major functional categories, each reflecting a distinct arm of the immune system that is compromised. Combined immunodeficiencies (CIDs), which affect both T and B lymphocytes, include SCID - the most severe PID - as well as less profound combined defects. SCID can arise from mutations in multiple genes encoding components of cytokine signalling (*JAK3, IL7R, IL2RG*), V(D)J recombination machinery (*RAG1, RAG2, DCLRE1C*), purine metabolism (ADA), or other essential lymphocyte developmental processes. In each case, the result is an absence of functional T cells and typically also B cells and NK cells, leaving affected infants without any adaptive immunity. Without treatment such as haematopoietic stem cell transplantation, SCID is universally fatal in early childhood [1].

Predominantly antibody deficiencies form the largest category of PIDs and include common variable immunodeficiency (CVID), X-linked agammaglobulinemia (XLA, caused by *BTK* variants), hyper-IgM syndromes (caused by *CD40L, CD40,* or *AID/AICDA* variants), and selective IgA deficiency. These conditions share imp*AIRE*d B-cell differentiation or immunoglobulin class switching, resulting in absent or severely reduced serum immunoglobulins and susceptibility to encapsulated bacterial pathogens.

Phagocyte defects - including CGD (*CYBB, CYBA, NCF1, NCF2*), severe congenital neutropenia (*HAX1, ELANE*), and leukocyte adhesion deficiencies (*ITGB2, SLC35C1, FERMT3)* - impair the innate immune response to bacterial and f*UNG*al pathogens. Complement deficiencies affect the complement cascade, with each component deficiency (*C1QA, C1QB, C1QC, C1S, C2, C3, C5, C6, C7, C8B, C9, CFD, CFH, CFI, CFP, MASP2*) producing characteristic infection susceptibilities. Disorders of immune dysregulation, including familial haemophagocytic lymphohistiocytosis (*PRF1, UNC13D, STX11, STXBP2, RAB27A*), involve variants in cytotoxic lymphocyte machinery leading to uncontrolled immune activation. This breadth of disease mechanisms explains why no single variant pattern

characterises all PIDs; each gene and disease subtype must be catalogued and analysed independently.

2.2 The History and Architecture of Locus-Specific Databases

The concept of the locus-specific database emerged in the early 1990s alongside the first systematic efforts to catalogue disease-causing variants at individual genetic loci. The Breast Cancer Mutation Data Base (BIC) for *BRCA1* and *BRCA2*, the Leiden muscular dystrophy pages, the KinMutBase registry of protein kinase variants [23], and the BTKbase for XLA were among the first LSDBs, each established by the research groups who had identified and characterised the relevant disease gene. These early databases were heterogeneous in format, software platform, and nomenclature convention, reflecting the absence of any cross-database standard; a contemporary directory catalogued hundreds of such LSDBs scattered across institutional websites [24].

The groundwork for standardised variant description was laid earlier by the Human Genome Variation Society (HGVS), whose nomenclature recommendations, initiated in the mid-1990s, first enabled consistent reference-anchored variant reporting. The Human Variome Project, launched in 2006, articulated the vision of a globally integrated LSDB network in which all human genetic variants would be collected, curated, and made accessible through standardised interfaces. This vision required two foundational elements: a standard nomenclature for variant description, and a standard software platform for data storage and sharing. The HGVS nomenclature system addressed the former; the LOVD platform addressed the latter. Cotton et al. [25] published formal recommendations and guideline for LSDB construction and curation that explicitly required the use of HGVS nomenclature, reference sequence versioning, and minimum data fields including clinical information and evidence for pathogenicity.

By 2006, when the IDbases collection was comprehensively described by Hilkka Piirilä, Jouni Väliaho and Mauno Vihinen [19], the field was transitioning from legacy formats to these emerging standards. The MUTbase platform, which had been developed specifically for the IDbases collection, predated these standards and was not designed for interoperability. It used a gene-specific HTML-based format that, while effective for human-readable access, lacked machine-readable APIs, structured variant classification, and standardised reference sequence handling. The result is that the IDbases collection, despite representing one of the most carefully curated PID variant resources in existence, has been effectively isolated from the modern variant database ecosystem for the better part of a decade.

2.3 HGVS Nomenclature: Principles and Relevance to Legacy Migration

The Human Genome Variation Society nomenclature system is the internationally accepted standard for describing sequence variants in a reference-sequence-anchored, unambiguous manner. The standard, most recently updated in 2016 by den Dunnen et al. [26], distinguishes between four coordinate systems. Genomic (g.) descriptions reference chromosomal positions on a versioned assembly chromosome accession (e.g., NC_000021.9). Coding DNA (c.) descriptions reference positions relative to the coding sequence of a specific transcript version (e.g., NM_000383.4). RNA (r.) descriptions reference the transcript sequence. Protein (p.) descriptions reference the amino acid sequence.

A fundamental requirement of HGVS is that the reference sequence be explicitly identified with its accession and version number. This versioning requirement is what makes HGVS descriptions reproducible across time: a variant described as NM_000383.4:c.769C>T refers unambiguously to position 769 of a specific, immutable transcript sequence, regardless of what changes occur to genome annotations later. The IDbases systematic names do not meet this requirement: they reference the gene-specific IDRefSeq clone, which has no public accession versioning, is not tracked in any reference sequence registry, and in many cases is no longer accessible as a named entity.

A further HGVS requirement relevant to this migration is the 3′ rule for variant normalisation: when a variant can be described at multiple equivalent positions (e.g., a deletion within a repeat), the position closest to the 3′ end of the sequence must be chosen. Legacy databases frequently recorded variants without applying this normalisation, meaning that the same variant may appear under different positions in different databases. The Mutalyzer tool applies the 3′ rule automatically during normalisation, which means that some IDbases systematic names will produce normalised HGVS descriptions at different positions from those originally recorded. This normalisation is scientifically correct but must be handled carefully during migration to avoid mismatching records across databases.

Two specific categories of legacy notation in IDbases require special handling. The IVS (intervening sequence) notation was used to describe intronic variants before HGVS established the c. coordinate system: for example, IVS3+2A>G describes a transversion two nucleotides downstream of the donor splice site of intron 3. Mutalyzer can convert many such descriptions to standard c. format provided the transcript structure is unambiguously defined, but conversion fails for entries where the IVS position is ambiguous or where the referenced exon numbering

does not match the current transcript annotation. The second problematic category is compound entries, where a single database record describes a variant at the protein or RNA level without a corresponding DNA coordinate. Such entries cannot be directly processed by Mutalyzer and require either back-translation or manual curation.

2.4 Reference Sequence Infrastructure: Genome Assemblies, LRG, and RefSeqGene

The history of human genome assembly is directly relevant to this migration project because IDbases variant coordinates were anchored to reference sequences that were current at the time of database construction, primarily in the hg16 (NCBI Build 34), hg17 (Build 35), and hg18 (Build 36) eras. The subsequent transition to GRCh37 (hg19, 2009) and then GRCh38 (hg38, 2013) involved substantial changes to chromosomal coordinates across the genome, including complete reorientation of some chromosomal segments, addition of alternate loci and patch sequences, and corrections to assembly errors that shifted coordinates by hundreds of base pairs in some regions. Variants described on hg18 cannot be used for clinical reporting today without explicit coordinate conversion, typically via the UCSC LiftOver tool.

GRCh38 represents the current standard reference assembly for clinical genomics. It incorporates significant improvements over GRCh37, including better representation of centromeric and pericentromeric regions, reduced gap content, and more accurate annotation of repeat elements. All major clinical bioinformatics tools - ClinVar, gnomAD, Ensembl, LOVD 3.0, and Mutalyzer 3 - now operate primarily or exclusively on GRCh38. Coordinate conversion via LiftOver using pre-computed chain files is therefore a necessary step before any GRCh38-based analysis.

Locus Reference Genomic (LRG) sequences were developed specifically to address the instability problem that arises when variant databases reference versioned assembly chromosomes. LRG sequences are fixed, non-versioned genomic sequences covering a single gene locus, maintained as a collaborative effort between NCBI and EMBL-EBI. Each LRG record contains a fixed genomic sequence spanning the gene and its immediate flanking regions, together with one or more transcript annotation tracks. The immutability of LRG sequences makes them ideal reference anchors for long-term variant reporting: an LRG-anchored variant description will remain valid indefinitely regardless of changes to genome assembly or RefSeq annotations.

NCBI NG_ (RefSeqGene) accessions provide an alternative to LRG sequences for gene-level variant anchoring. Like LRG sequences, each NG_ accession corresponds to a specific genomic locus and is linked to one or more curated NM_ transcript accessions. In this project, the terms NG_ and LRG are used essentially interchangeably, as for the vast majority of IDbases genes with LRG records, the LRG and NG_ sequences are co-registered in NCBI's RefSeqGene database.

2.5 MANE Select: Standardising Transcript Selection for Clinical Reporting

The existence of multiple transcript isoforms per gene has historically been one of the most significant sources of variant description inconsistency in clinical genomics. The same variant may be described with different c. coordinates depending on which transcript is used as the reference, and different clinical laboratories, genome browsers, and databases have historically chosen different 'default' transcripts for the same gene. This inconsistency complicates variant sharing, database aggregation, and clinical decision support.

The Matched Annotation from NCBI and EMBL-EBI (MANE) project, a collaboration between the NCBI RefSeq and Ensembl/GENCODE annotation groups, was established to resolve this problem. The first comprehensive MANE transcript set was published in Nature in 2022 [27], defining MANE Select as one representative transcript per protein-coding gene that is: (1) perfectly aligned to GRCh38; (2) identical in sequence and splice structure between RefSeq and Ensembl/GENCODE; (3) representative of the predominant biological isoform based on expression data; and (4) curated to include all known pathogenic variants where possible. The MANE Select set now covers over 99% of human protein-coding genes, providing a universal default transcript for clinical variant reporting.

The significance of MANE Select for this migration project is twofold. First, MANE Select transcripts are already integrated into LOVD 3.0 and Mutalyzer 3, meaning that variants processed through the migration pipeline can be automatically annotated with MANE Select transcript identifiers. Second, the NM_ transcript versions historically used in IDbases frequently differ from the current MANE Select transcript for the same gene. This is particularly evident in Track C (NM_IDRefseq), where 1,726 ENOSELECTORFOUND errors reflect cases where the IDRefSeq NM_ accession is not recognised in the current Mutalyzer genome model. The recommended approach - preserving the original c. annotation on the historical NM_

accession while adding a MANE Select mapping as a secondary annotation - maintains backward traceability while aligning to current standards [28].

2.6 The LOVD Platform: Architecture, Import Mechanics, and Global Adoption

The Leiden Open Variation Database was first developed at Leiden University Medical Center in the late 1990s as an 'LSDB-in-a-Box' solution. LOVD 2.0 [10] introduced multi-gene per patient storage. LOVD 3.0 [11] represented a complete architectural redesign centred on three principles: genome-centricity, scalability, and enforced nomenclature compliance. The genome-centric architecture stores variants primarily by chromosomal position, enabling storage of genome-wide sequencing data alongside traditional gene-specific entries.

LOVD 3.0 import mechanics are relevant to this project. The platform accepts a tab-delimited import format with specific column headers for chromosomal variants (VariantOnGenome), transcript-level variants (VariantOnTranscript), individuals, and phenotypes. Each variant must carry at minimum a DNA description in HGVS format. LOVD validates all submitted HGVS descriptions in real time against the Mutalyzer service: any description that fails Mutalyzer validation is rejected at import time. This enforcement mechanism guarantees the nomenclature quality of all data in the LOVD installation.

Global adoption of LOVD has been extensive. As of the most comprehensive published assessment [11], its flagship installation — the Global Variome shared LOVD — was the largest single curated instance, hosting nearly 23,000 gene databases maintained by 365 volunteer curators and receiving over 5,000 new variant observations (nearly 3,000 new cases) every month. LOVD was endorsed as a recommended system by the Human Variome Project, an initiative that is no longer active. Migrating the IDbases data to LOVD 3.0 therefore places it within the most widely used LSDB infrastructure in existence.

2.7 Mutalyzer: Architecture and Validation Scope

Mutalyzer was developed at Leiden University Medical Center specifically to support LOVD's requirement for HGVS nomenclature validation. The first version, described by Wildeman et al. [20], provided a web interface for single-variant checking. Mutalyzer 2 [29] introduced a complete rewrite with a formal Extended Backus-Naur Form grammar for HGVS syntax parsing, enabling automated processing of complex and compound variant descriptions. The tool has processed over 133 million descriptions since its launch. Mutalyzer 3, the current version, operates exclusively on GRCh38, provides a REST API for programmatic batch

submission, and is the engine used throughout this migration project. VariantValidator offers an independent HGVS validation route and is well suited to resolving the residual transcript-selector failures, complementing the Mutalyzer-based pipeline used here.

The validation scope of Mutalyzer is an important consideration for interpreting migration results. For substitution variants, Mutalyzer validates both the position and the stated reference nucleotide (REF base) against the actual sequence of the reference at that position. An ESEQUENCEMISMATCH error indicates that the REF base stated in the variant description does not match the reference sequence at that chromosomal coordinate. For deletions, insertions, duplications, and indels variants, Mutalyzer validates only the position range. This asymmetry is significant: the 24 ESEQUENCEMISMATCH errors in Track A, all in forward-strand genes (*CD40L*, *RAG1*, *PTPRC*), represent genuine errors inherited from the original sources (the published literature or direct submissions) rather than from IDbases curation where the recorded REF base does not match the GRCh38 reference.

The Mutalyzer REST API returns a structured response for each submitted description. For successful submissions (status 'ok'), the response includes the normalised HGVS descriptions at all four levels (g., c., r., p.), the MANE Select transcript tag if applicable, protein position information, and the Mutalyzer-assigned gene name. The two-pass submission design used in this project - where g. HGVS is submitted first, and then the returned c. HGVS is re-submitted - is necessary because RNA and protein annotations are only returned when the i*NP*ut is in c. format with an explicit transcript selector.

2.8 Related Work: Variant Database Migration and Standardisation Efforts

The challenge of migrating legacy variant databases to modern standards is not unique to immunodeficiency genetics, though few published studies have addressed it systematically. The transition of BTKbase, the original XLA variant database, from its standalone format to the LOVD platform was achieved through manual curation rather than automated pipeline approaches, reflecting the state of available tooling at the time. The BRCA Exchange project aggregated variant data from multiple BRCA1/2 databases with different provenance and nomenclature conventions, demonstrating that cross-database harmonisation requires both automated normalisation and expert curation, and that the two cannot be fully separated [30].

The specific problem of ENA clone-anchored coordinate systems has not been addressed in any previously published migration workflow to our knowledge. The broader problem of coordinate

instability in legacy databases has been discussed in the context of ClinVar data quality [31], where a significant fraction of older submissions were found to use deprecated reference sequence versions or non-canonical coordinate descriptions. Within the immunodeficiency field specifically, there has been no previously published automated approach to migrating the IDbases collection. The seven-step pipeline presented in this thesis represents, to our knowledge, the first published automated workflow for migrating a large-scale, multi-gene disease-specific LSDB from a legacy ENA clone-anchored coordinate system to a GRCh38/HGVS/LOVD 3.0 standard.

3. Materials and Methods

3.1 Pipeline Architecture and Design Principles

The migration pipeline is organised as seven sequential steps, each implemented as one or more Python scripts. Steps 1 through 5 constitute the coordinate conversion chain for Track A; Step 6 encompasses all three Mutalyzer annotation tracks in parallel; Mutalyzer validates and corrects HGVS variant descriptions against reference sequences [20]; Step 7 merges the three tracks and prepares the LOVD import dataset. LOVD 3.0 is the target platform; it is a genome-centred open-source LSDB system supporting HGVS-compliant variant storage and federated data sharing [11]. The overall flow is summarised below.

Table 1. Summary of the seven-step IDbases-to-LOVD 3.0 migration pipeline.

Step	Name	Track(s)	Key Output
1a	Variant Extraction	All	all_mutations.tsv
1b	Reference Sequence Mapping	All	LRG_with_NM.csv
1c	Reference Sequence Download	All	NG_/NM_/ENA FASTA files
2	Offset Resolution & IDRefSeq Confirmation	All	lrg_offset_results.csv, idrefseq_pair_comparison.csv
3	BLAST Genome Alignment	A only	blast_first_hits_source_IDRefseq.tsv
4	BED File Generation & Filtering	A only	Filtered BED files (hg18)
5	LiftOver to GRCh38	A only	hg38 BED files with strand

Step	Name	Track(s)	Key Output
6	Mutalyzer HGVS Normalisation	A, B, C	mutalyzer_results*.tsv
7	Track Merge & LOVD Import Preparation	All	merged_variants.tsv, unresolved_variants.tsv

The three-track architecture was adopted because no single route from IDbases source data to HGVS-normalised annotation is sufficient for all genes and all variant types. Track A (NG_IDRefseq) uses chromosomal coordinates derived from BLAST alignment and LiftOver to produce genomic HGVS descriptions independent of transcript version. Track B (NM_MANE) uses the current MANE Select transcript for each gene to produce c. descriptions aligned to the current clinical standard. Track C (NM_IDRefseq) uses the original IDRefSeq NM_ accession to produce c. descriptions preserving backward compatibility with the literature. The priority merge in Step 7 assigns the best available annotation to each variant. All scripts are implemented in Python 3.10+. The pipeline code is available at github.com/knn1996/Migrating-BMC-IDBases-to-LOVD.

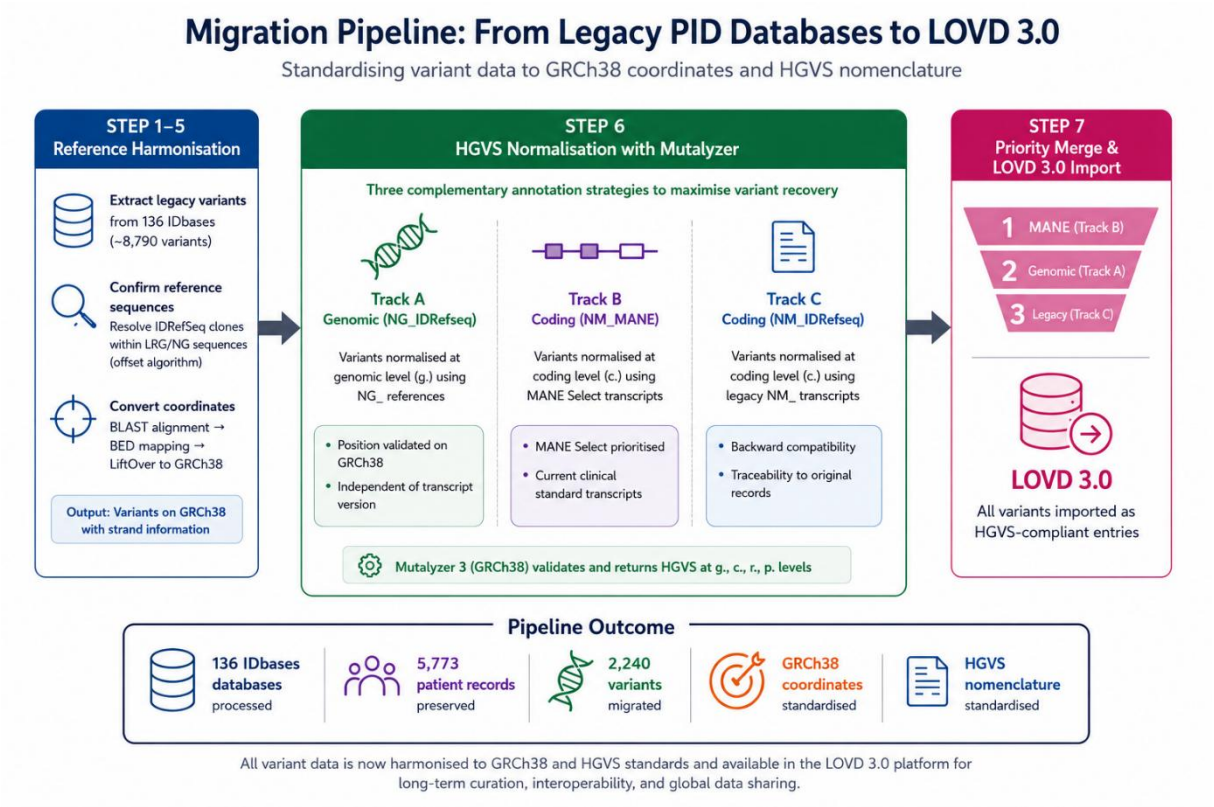

Figure 1. Overview of the IDbases-to-LOVD 3.0 variant migration pipeline. Schematic of the end-to-end pipeline for migrating legacy primary immunodeficiency (PID) variant data from the MUTbase-derived IDbases collection into LOVD 3.0,

standardising all entries to GRCh38 coordinates and HGVS nomenclature. The workflow comprises three phases. In Reference Harmonisation (Steps 1–5), legacy variants are extracted from the IDbases gene databases, each partial IDbases reference clone is located within its corresponding LRG/NG genomic sequence via an empirical offset algorithm, and coordinates are converted to GRCh38 (BLAST alignment → BED mapping → LiftOver), producing strand-aware genomic positions. In HGVS Normalisation (Step 6), variants are re-described and validated with Mutalyzer 3 against GRCh38 using three complementary annotation tracks: Track A at the genomic level (NG_ IDRefSeq), Track B at the coding level using MANE Select transcripts (NM_ MANE), and Track C at the coding level using the legacy IDbases transcripts (NM_ IDRefSeq); Mutalyzer returns validated descriptions at the g., c., r. and p. levels. In Priority Merge and Import (Step 7), per-variant descriptions are reconciled in priority order (Track B > Track A > Track C) and imported into LOVD 3.0 as HGVS-compliant entries. From 8,790 source variants across the in-scope IDbases genes, 2,240 distinct variants were resolved and 5,773 patient–variant links (2,730 patients) preserved.(Image generated by ChatGPT)

3.2 Step 1: Data Extraction and Reference Preparation

Step 1a extracted variant records from all 136 IDbases pub.html files using the extract_mutations.py parser. MUTbase stored each patient record as an HTML table within a *pub.html file, with variant information encoded in Feature/dna blocks. The parser identified /loc: IDRefSeq: fields to extract variant positions, Systematic name fields for existing HGVS-style descriptions, /name: fields for variant type annotation, and allele number designators. The output is all_mutations.tsv, with one row per variant-allele combination.

Step 1b constructed the reference sequence mapping table using NCBI E-utilities (esearch, elink, esummary) and the LRG FTP resource, producing LRG_with_NM.csv. This maps each gene to its LRG accession, NG_ accession, and MANE Select NM_ accession. All of 136 IDbases genes were included in LRG_with_NM.csv; the immunoglobulin genes *IGHG2* and IGHM lack a dedicated LRG record and were mapped through NG_ constant-region accessions instead. BTK and SH2 were excluded from the migration scope as they had already been migrated separately, leaving 134 databases in active scope.

Step 1c downloaded three sets of reference sequences: (1) ENA EMBL clone FASTAs for each gene (the original IDRefSeq sequences, Source 1); (2) NCBI genomic clone FASTAs (Source 2); (3) LRG NG_ FASTAs, MANE Select NG_ FASTAs, and MANE Select NM_ FASTAs (Source 3). All 136 genes have corresponding FASTAs in the Mane_Select_NG directory.

3.3 Step 2: Reference Sequence Confirmation and Offset Resolution

3.3.1 IDRefSeq Containment Analysis

For each gene where both a Source 1 (IDRefSeq ENA clone) and Source 3 (LRG/NG_) FASTA were available, compare_idrefseq_pairs.py performed a pairwise containment test: the script searched for the complete Source 1 sequence as a substring of the Source 3 sequence and recorded whether containment was confirmed (s1_inside_s3) or could not be established (no_containment).

3.3.2 Offset Algorithm

The offset algorithm (find_lrg_offset.py) resolves the starting position of each IDRefSeq clone within the corresponding LRG/NG_ sequence. It filters all_mutations.tsv for rows where variant_type == 'genomic' AND mut_class == 'substitution', assembles a test set of (IDRefSeq_position, REF_base) pairs, and scans all possible offsets within the NG_ sequence to find the position that maximises the proportion of correctly matched reference nucleotides. A match fraction of ≥90% is required for 'ok' status.

Three rescue passes are applied sequentially: (1) a forward linear scan; (2) a reverse complement pass for minus-strand genes (NG_ route only); (3) a ±1 position shift pass for indexing inconsistencies. Genes are absent from lrg_offset_results.csv if they have no genomic substitution rows in all_mutations.tsv - this affects 11 genes (*CD3E*, *CD59*, ICOS, *IL12B*, LRRC8A, *PRKDC*, *RNF168*, *SP110*, TAPBP, TCIRG1, *TYK2*) which are handled exclusively through the NM-based tracks.

3.3.3 Reference Sequence Validation by Offset Anchoring

The offset resolved in Section 3.3.2 can be verified directly only for variant classes that reference one or more nucleotides of the original sequence. Substitutions, deletions, duplications, and deletion–insertions all carry such reference content and are checked against the NG_ sequence during normalisation. Insertions are the exception: they specify only the inserted bases and the position between which they fall, carry no reference nucleotide of their own, and therefore cannot be confirmed by direct base comparison. A statistical validation of the offset itself was therefore performed so that the confidence established on checkable anchors transfers, by proxy, to insertions placed by the same offset.

For a gene g with resolved offset δg, each IDRefSeq anchor position p_i is translated to its NG_ coordinate q_i by

$$q_i = p_i + \delta_g \ (forward\ strand) \quad (3.1)$$

with the analogous reverse-complement mapping applied for the confirmed reverse-strand genes (Section 3.3.2). For each gene resolved through the NG_ route (Track A), the reference base r_i of every genomic substitution – the same anchors used to resolve the offset – was compared with the nucleotide $s(q_i)$ present in the NG_ sequence at the offset-translated position, and a gene-level match rate over the ng anchors was computed as

$$m_g = \left(\frac{1}{n_g}\right) \cdot \Sigma_{\{i=1\}}^{\{n_g\}} \mathbb{1}[\, s(q_i) = r_i \,] \quad (3.2)$$

where $\mathbb{1}[\cdot]$ is the indicator function. This test applies only to genes anchored through an NG_ reference; genes that found no genomic sequence in LRG (IDRefseq Project) were recorded in lrg_offset_results.csv as no_fasta and are resolved instead through the MANE Select and NM-based tracks (Section 3.7), where Mutalyzer performs its own reference checking. Their absence from the offset table is therefore a subsequent of Track A reference preparation rather than a gap in reference-sequence coverage.

The null hypothesis was that anchors align to the NG_ sequence no better than chance. Because the genome comprises four nucleotides, a randomly placed anchor matches with probability 0.25, so the null match probability was set at $p_0 = 0.25$. (This uniform-composition assumption is mildly conservative: under the actual genomic base composition the random-match probability is marginally higher, ≈ 0.26, which would only raise the bar for significance and does not affect the conclusions.) Letting $x_g = m_g \cdot n_g$ denote the number of matching anchors, each gene was evaluated against this null using a one-sided exact binomial test, with the alternative hypothesis that the true match probability exceeds p_0:

$$pg = \Sigma_{\{k = x_g\}}^{\{n_g\}} C(n_g, k) \cdot p^{0k} \cdot (1 - p^0)^{\{n_g - k\}} \quad (3.3)$$

An aggregate test pooling anchors across all tested genes, with $X = \Sigma_g x_g$ matches over $N = \Sigma_g n_g$ anchors, was computed identically:

$$p_{agg} = \Sigma_{\{k = X\}}^{\{N\}} C(N, k) \cdot p^{0k} \cdot (1 - p^0)^{\{N - k\}} \quad (3.4)$$

A match rate significantly above p_0 indicates that the offset places anchors at biologically correct positions, confirming the underlying reference assignment and, by extension, the placement of insertions that cannot be checked individually. Because 94 genes were tested, per-gene significance was additionally assessed against a Bonferroni-corrected threshold ($\alpha / 94 \approx 5.3 \times 10^{-4}$); the pooled aggregate test (Equation 3.4) served as the primary inference and is unaffected by multiplicity.

3.4 Step 3: Genome Alignment via BLAST

IDRefSeq clone FASTA sequences were aligned against the hg16, hg17 and hg18 human genome assemblies using blastn (BLAST+ 2.16.0), retaining only full-identity hits. Most genes were aligned locally, but the largest clone fragments (>100 kb FASTA), which exceeded the memory available on a standard workstation, were run on the LUNARC COSMOS high-performance computing cluster at Lund University (project lu2025-2-88, partition lu48) under the SLURM workload manager. These were submitted as a SLURM job array, one task per large clone, each querying the three pre-built assembly databases (hg16, hg17, hg18) with tabular output (outfmt 7), an E-value threshold of 1e-10 and up to 100 target hits. After an initial 8-core/32 GB allocation failed for the largest inputs, the jobs were run with 16 cores and 64 GB of memory. The first hit per gene and assembly was retained, and the BLAST sstart/send orientation provided the authoritative strand assignment: when sstart > send, the gene lies on the reverse strand. Results were compiled into blast_first_hits_source_IDRefseq.tsv.

3.5 Step 4: BED File Generation and Filtering

BED files encoding the chromosomal position of each variant were generated by generate_bed.py. IDRefSeq positions (1-based) were converted to chromosomal coordinates using linear interpolation from BLAST alignment endpoints, with strand direction recorded in BED column 6. For five genes (*ZAP70*, *STX11*, *UNC13D*, ADA, *NP*) where IDRefSeq positions extend beyond the BLAST alignment end, linear extrapolation was applied. The filter_bed.py script removed non-matching accessions, selecting the best source by highest match percentage.

3.6 Step 5: Coordinate LiftOver to GRCh38

The UCSC LiftOver command-line tool converted chromosomal coordinates from hg16/17/18 to GRCh38 (hg38) using the appropriate pre-computed chain files. LiftOver is a position-only conversion and does not validate the reference nucleotide; nucleotide validation is deferred to Mutalyzer for substitution variants. Variants failing LiftOver were written to per-gene unlifted.bed files and not carried forward.

3.7 Step 6: Mutalyzer HGVS Normalisation

3.7.1 Track A - NG_IDRefseq

GRCh38 chromosomal coordinates were formatted as HGVS genomic notation and submitted to the Mutalyzer 3 /api/normalize REST endpoint via run_mutalyzer.py. A two-pass design was used: Pass 1 submits the g. HGVS and retrieves c_hgvs; Pass 2 re-submits c. HGVS to obtain r., p. descriptions, MANE Select tag, and protein positions. All API responses are cached

incrementally to JSONL files for resumability. For 43 confirmed reverse-strand genes (BLAST sstart > send), the complement() transformation is applied to REF and ALT before constructing the HGVS string in bed_to_mutalyzer.py.

3.7.2 Tracks B and C - NM_MANE and NM_IDRefseq

For Tracks B and C, variant records from all_mutations.tsv were matched to transcript accessions from LRG_with_NM.csv at query time. Track B used MANE Select NM_ accessions; Track C used the IDRefSeq NM_ accessions. HGVS c. strings were constructed as NC_XXXXXXX.XX(NM_XXXXX.X):c.SYSNAME and submitted via run_mutalyzer_NM.py. Tracks B and C do not pass through Steps 3, 4, or 5. A gene-specific override was implemented for *ORAI1*, which is submitted in NM-only format (NM_032790.4:c.XXX without the NC_ prefix) due to its absence from the Mutalyzer genome model.

3.7.3 Mutalyzer Error Classification

Table 2. Mutalyzer error classification scheme and remediation categories.

Error code	Description	Track A	Track B	Track C
ESEQUENCEMISMATCH	Reference base mismatch at position	24	211	153
EINTRONIC	Intronic position in c. notation	0	60	60
ENOSELECTORFOUND	Transcript not in Mutalyzer genome model	0	0*	1,726
ESYNTAXUEOF	U*NP*arseable HGVS string	0	6	6
OTHER	Other errors	0	9	24

** NM_MANE ENOSELECTORFOUND was 7 (ORAI1); resolved by NM-only submission fix.*

3.8 Step 7: Priority-Based Track Merge

The merge_tracks.py script combines the three Mutalyzer output tracks into a single canonical record per variant using (gene, accession, allele_num) as the match key. Phase 1 (cross-check) compares HGVS strings between tracks after stripping transcript version numbers; disagreements are written to crosscheck_disagreements.tsv. Phase 2 (priority merge) applies the order NM_MANE → NG_IDRefseq → NM_IDRefseq. Resolved variants are written to merged_variants.tsv; variants failing all three tracks go to unresolved_variants.tsv.

4. Results

4.1 Variant Extraction and Database Scope

Variant extraction from 136 IDbases *pub.html files produced a total of approximately 7,341 variant rows in all_mutations.tsv, representing genomic-coordinate and coding-coordinate entries across all genes. The largest individual database was *CYBB* (gp91^{3hx}, X-linked CGD) with 1,256 patient records in the coding tracks, followed by *CTSC* (248), *PTPN11* (247), *PRF1* (236), and *SBDS* (234). The smallest databases contain only one to three records (e.g., *CD3E*, *CD3G*, *RAC2*). This span of two orders of magnitude in database size reflects both the clinical prevalence of the corresponding diseases and the research focus of the contributing groups.

From IDBases_Summary.csv, 97 genes had complete reference sequence information and were carried forward through the full Track A pipeline (liftover_to_hg38 = yes). The remaining 39 genes are handled exclusively through Tracks B and C.

- Genes with BED generated and successfully lifted to GRCh38: 97
- Genes with no Track A processing (liftover_to_hg38 = no): 37 - ADA, AICDA, BLNK, *C7*, *CARD9*, *CD40*, *CD79A*, *CFD*, CFH, DCLRE1C, *DNMT3B*, ELA2, *FASLG*, *FOXN1*, *FOXP3*, ICOS, *IFNGR1*, IGHM, IKBKG, LRRC8A, MPO, *NRAS*, *PIK3R1*, *PTPN11*, *RASA1*, *RASGRP2*, *RFXAP*, SERPING1, SH2D1A, *STAT2*, *TAP1*, *TAP2*, TAPBP, *TAZ*, TCIRG1, *TCN2*, WAS.

4.2 Reference Sequence Confirmation: IDRefSeq as LRG Fragments

The containment analysis in compare_idrefseq_pairs.py tested 47 gene pairs where both a Source 1 (IDRefSeq ENA clone) and a Source 3 (LRG/NG_) FASTA were available. Of these, 31 (66%) showed confirmed containment (s1_inside_s3), and 16 (34%) showed no_containment.

Table 3. IDRefSeq sequences confirmed as partial LRG/NG clone fragments.

Result	Count	Example genes
s1_inside_s3	30	*AP3B1*, *BLNK*, *C1QA*, *C1QB*, *CASP10*, *CASP8*, *CD40L*, *CD55*, *CD59*, *CYBB*, *FCGR3A*, *GFI1*, *IFNGR2*, *IL12RB1*, *IL2RA*, *LIG4*, *MLPH*, *MRE11A*, *NFKBIA*, *RAC2*, *RAG2*, *RFXANK*, *SBDS*, *SMARCAL1*, *SPINK5*, *STAT1*, *STAT5B*, *STX11*, *TNFRSF13B*, *UNG*

Result	Count	Example genes
no_containment	16	*C1S*, *C2*, *C3*, *C6*, *C9*, *CTSC*, *DKC1*, *FCGR1A*, *IGLL1*, *LIG1*, *MASP2*, *MYO5A*, *PRF1*, *PTPRC*, *RAG1*, *SLC35C1*

The s1_inside_s3 cases directly confirm the hypothesis that IDbases IDRefSeq sequences are sub-sequences of the corresponding LRG/NG_ sequences. For example, the *CD40L* IDRefSeq clone (NG_007280, 14,196 bp) begins at position 0 within NG_007280.1 (15,197 bp). In contrast, the *C1QA* IDRefSeq clone begins at position 4,821 within NG_007282.1, meaning variants with IDRefSeq positions less than 4,821 bp from the LRG start are out of bounds relative to the IDRefSeq reference.

4.3 Offset Algorithm Results

The offset algorithm (find_lrg_offset.py --source IDRefseq) processed 125 genes in lrg_offset_results.csv. The 11 genes absent from this file (*CD3E*, *CD59*, ICOS, *IL12B*, LRRC8A, *PRKDC*, *RNF168*, *SP110*, TAPBP, TCIRG1, *TYK2*) were not processed because the algorithm requires genomic substitution rows in all_mutations.tsv as anchors, and these genes have none - either their IDbases entries are coding-only or they had no Track A pipeline processing. All 11 are handled by Tracks B and C only.

Table 4. Offset algorithm anchoring outcomes across genes.

Status	Count	Notes
ok	92	Match fraction ≥90%; offset resolved successfully (97.9% of 94 tested)
below_threshold	2	*RAG1*, *RAG2*: best match <90% after all rescue done
no_fasta	31	No IDRefSeq FASTA available; includes 11 genes with no genomic substitution rows

Among the below_threshold results, *RAG1* achieved 89.5% match (51/57) and *RAG2* achieved 78.7% (37/47) after all rescue passes; these persistent failures likely reflect minor sequence differences between the ENA clone deposited at the time of database construction and the current LRG sequence; *RAG1*/*RAG2* hypomorphic variants have been linked to clinically heterogeneous presentations that may compound these curation challenges [33].

4.3.1 Statistical Validation of Offset Anchoring

Of the 125 genes in lrg_offset_results.csv, 94 carried genomic substitution anchors and were testable; the remaining 31 lacked an NG_ reference at this step (no_fasta) and were resolved through the NM-based tracks rather than representing uncovered genes. Across the 94 tested genes, 3,799 of 3,826 anchors matched the NG_ reference base at the offset-translated position - an aggregate match rate of 99.29%. Against the chance expectation of 25%, this is overwhelmingly significant, with a one-sided exact binomial p-value indistinguishable from zero. The offset conversion therefore reproduces the reference sequence essentially perfectly at the aggregate level.

At the individual gene level, 77 of the 94 genes reached significance at $\alpha = 0.05$ and 67 reached $p < 10^{-3}$. The 17 genes that did not reach significance were almost entirely those with very few anchors: all seventeen had only one or two substitutions to anchor on (nine with a single anchor, eight with two), and every one matched perfectly. With a single anchor a perfect match yields $p = 0.25$, and with two anchors a perfect match yields only $p = 0.0625$, so significance is mathematically unreachable regardless of how convincingly the bases agree. Their non-significance reflects a lack of statistical power for sparsely sampled genes, not a failure of the offset.

Two genes fell below the 90% match threshold and were flagged for manual review rather than accepted automatically (Section 4.3); both nonetheless exceeded the chance expectation by a wide margin. Because pure insertions carry no reference nucleotide and cannot be position-checked individually, the strength of the aggregate anchor agreement is the key evidence supporting their placement: the 99.29% concordance demonstrates that the gene-specific offsets position anchors correctly, so insertions mapped by the same offset inherit that correctness. This validation closes the one gap that direct base comparison could not address.

Reference-sequence coverage for the dataset as a whole should not be inferred from the offset table, whose no_fasta entries are Track A preparation artefacts; genes without an NG_ reference are recovered through the MANE Select and NM tracks. Overall coverage is instead reflected in the final merged output, which retained variants for 115 of the 134 in-scope genes (85.8%) - corresponding to 5,773 unique patient–variant pairs - leaving 19 in-scope genes without any resolved variant.

The wide confidence intervals (Figure 2) seen for many genes are a direct consequence of low anchor counts, not of poor offset accuracy. The width of the Wilson interval is governed by the number of substitution anchors n: when only one or two substitutions are available, even a

perfect match leaves the true match rate poorly constrained, so the one-sided 95% lower bound falls far below the observed value (≈27% at n = 1 and ≈43% at n = 2, despite 100% observed agreement). The same low-n limitation is why these genes do not reach significance against the 25% chance null, since a perfect match yields only p = 0.25 at n = 1 and p = 0.0625 at n = 2. Both effects reflect limited statistical power for sparsely sampled genes rather than any failure of the offset, as confirmed at the aggregate level where pooling all 3,826 anchors gives a tight 99.29% estimate and a binomial p-value indistinguishable from zero. RAG1 and RAG2 are the distinct exception: with 57 and 47 anchors they have narrow intervals and tiny p-values, yet their point estimates fall below the 90% acceptance threshold, which is why they are flagged and routed to Track B.

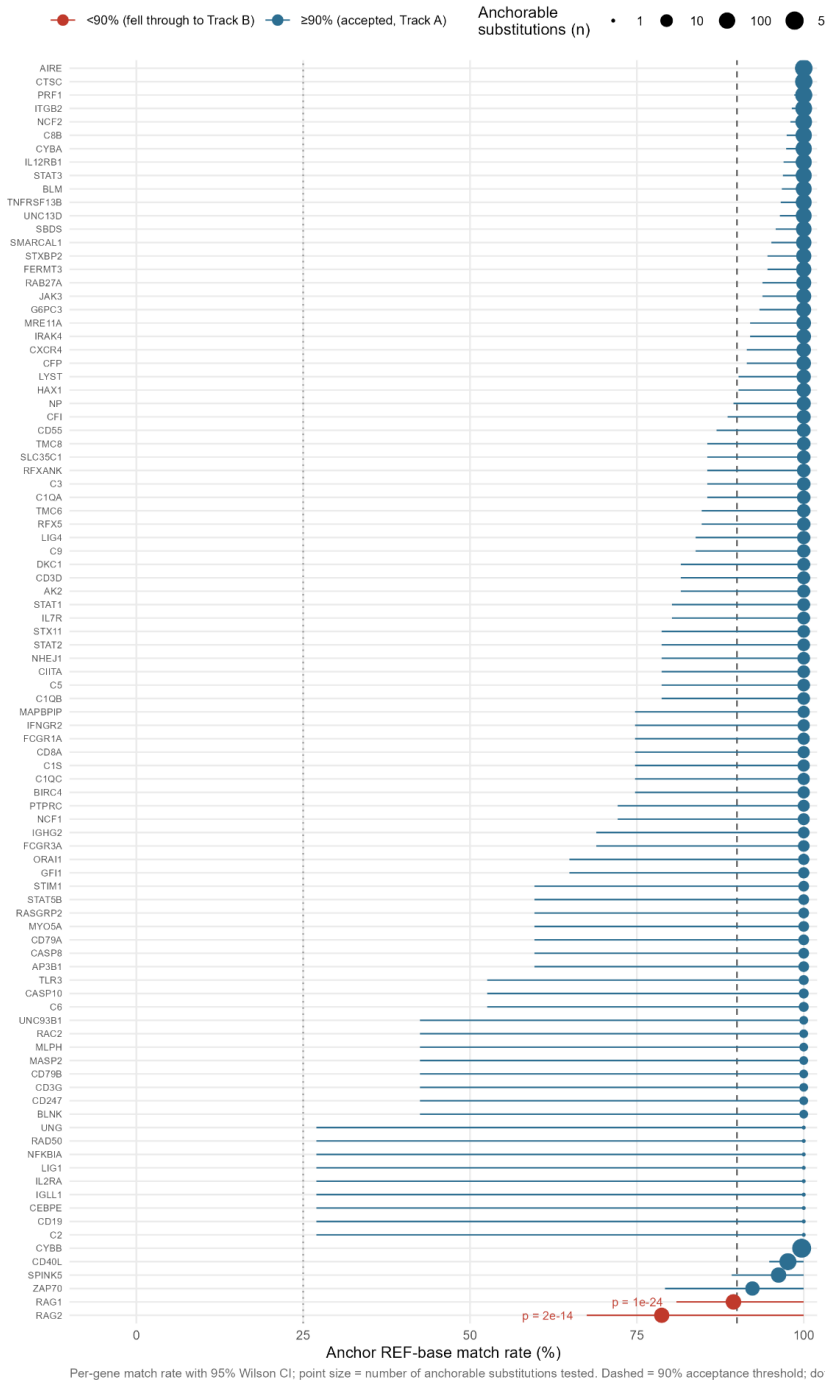

Figure 2. Per-gene anchor match rates for IDRefSeq offset confirmation. For each of the 94 genes with a testable NG_ reference, the empirical offset algorithm was used to align IDRefSeq coordinates within the full LRG/NG_ sequence and the proportion of correctly matching reference bases was computed. Points show the observed match rate, sized by the number of substitutions tested per gene; horizontal lines give the one-sided 95% Wilson lower confidence bound (z = 1.645, upper bound fixed at 1.0). The dashed line marks the 90% acceptance threshold and the dotted line the 25% chance level expected under random alignment. RAG1 and RAG2 (red) fall below the acceptance threshold but remain highly significant against the chance null (binomial $p \approx 10^{-24}$ and 10^{-14} respectively), and are correctly routed to Track B.

4.4 Genome Alignment and LiftOver

BLAST alignment against hg16, hg17, and hg18 was successful for all 104 genes with available IDRefSeq FASTA and confirmed mapping in reference_summary.csv. UCSC LiftOver

converted 1,191 variant positions from hg18 to GRCh38. The completeness of variant capture remains a wider challenge: Charoenngam et al. [13] demonstrated that over 58% of CASR disease-causing variants identified in the literature were absent from ClinVar and LOVD, a pattern likely mirrored in the IDbases collection. The four variants in mutalyzer_skipped.tsv (*C3*, *CD40L*, *CYBB*, *IL7R*) contained either compound descriptions or bare insertional notation without specifying the inserted sequence and were unresolvable from the pipeline.

4.5 Mutalyzer Normalisation: Track-Level Results

Table 5. Mutalyzer normalisation results by resolution track.

Track	Total submitted	OK	% OK	ESEQMISMATCH	EINTRONIC	ENOSELECTOR	ESYNTAXUEOF	OTHER
A - NG_IDRefseq	3,606	3,582	99.3%	24	0	0	0	0
B - NM_MANE	4,775	4,496	94.1%	211	60	0*	6	9
C - NM_IDRefseq	4,811	2,862	59.5%	153	60	1,726	6	24

** NM_MANE ENOSELECTORFOUND was 7 (ORAI1); resolved by NM-only submission fix.*

The high overall resolution rate contrasts with findings from other disease-specific database audits: Seidizadeh et al. [8] found that large-scale mining of LOVD and HGMD revealed substantially more pathogenic VWF alleles than clinical registries had captured, underscoring the importance of systematic database coverage. Track A's 24 ESEQUENCEMISMATCH errors are confirmed genuine errors inherited from the original sources rather than introduced during IDbases curation: the recorded REF base does not match GRCh38 at the corresponding chromosomal position, and the mismatch is not the complement relationship that would indicate a strand error. Track B's 211 ESEQUENCEMISMATCH errors are primarily due to NM_ version mismatch between IDbases and the current MANE Select coding sequence. Track C's 1,726 ENOSELECTORFOUND errors represent systematic NM_ version obsolescence.

4.6 Per-Gene Mutalyzer Results: Notable Cases

The *NCF1* gene (p47^{3hx}, autosomal recessive CGD) shows a distinctive pattern: Track A achieves 100% ok (31/31), while both NM tracks achieve only 7.5% (4/53). *NCF1* resides in a segmentally duplicated region of chromosome 7q11 containing pseudogene paralogues *NCF1*B

and *NCF1*C. The NM track failures likely reflect variants described relative to a haplotype sequence differing from the GRCh38 primary assembly annotation at the *NCF1* locus. Track A's success confirms that the genomic coordinates are correct; the problem is specific to the c. coordinate description and requires specialist review.

Table 6. Notable per-gene Mutalyzer resolution cases.

Gene	Track A ok%	Track B ok%	Track C ok%	Notable feature
CYBB	100%	99.4%	99.4%	Largest database (1,051/1,256); all tracks >99%
AIRE	100%	N/A	N/A	Track A only; 268/268 ok
CTSC	99.2%	98.8%	0%	Track C: 248 ENOSELECTORFOUND - NM_ version obsolete
PRF1	100%	90.3%	0%	Track B: 23 ESEQMISMATCH; Track C: 236 ENOSELECTORFOUND
SBDS	100%	100%	0%	Track C: 234 ENOSELECTORFOUND - NM_ fully obsolete
NCF1	100%	7.5%	7.5%	Both NM tracks: 49 ESEQMISMATCH - complex duplicated genomic region
RAG1	89.5%	87.8%	87.8%	6 confirmed curation errors in Track A; 10 in NM tracks
C1QA	100%	29.4%	0%	Track B: 12 ESEQMISMATCH - NM_ version mismatch
IL7R	100%	18.2%	0%	Track B: 9 ESEQMISMATCH - significant NM version divergence
C9	100%	21.4%	0%	Track B: 11 ESEQMISMATCH - complement gene cluster complexity
TAZ	N/A	75.6%	75.6%	21/86 EINTRONIC errors - high proportion of intronic variants
SERPING1	N/A (below_threshold)	N/A	N/A	Offset below threshold; NM tracks also complex

4.7 Track Merge Results

Priority-based merging produced 7,776 resolved patient–variant records in merged_variants.tsv. Because each biological variant may appear as separate genomic and coding rows across tracks and is recorded once per contributing patient, this record count overstates the biological scope. After deduplication (dedup_merged_variants.tsv) the dataset resolves to 2,240 distinct variants across 115 genes, linked to 2,729 patients through 5,773 patient–variant pairs, with 698 distinct variants left unresolved and characterised in Section 4.12.

Table 7. Priority-based track merge outcomes.

Outcome	Count	Notes
Resolved - source: NM_MANE (priority 1)	3,798	MANE Select-aligned c. annotation
Resolved - source: NG_IDRefseq (priority 2)	3,582	Genomic route; second-largest contributor to the merged set
Resolved - source: NM_IDRefseq (priority 3)	396	Both higher-priority tracks failed
Unresolved - all tracks invalid	698	Distinct variants unresolved across all tracks (see Section 4.12); ~75% automatically rescuable
TOTAL	7,776	Resolved patient–variant records (3,798 + 3,582 + 396). After deduplication: 2,240 distinct variants, 2,729 patients, 5,773 patient–variant pairs.
Unique variants after deduplication	2,240	Distinct variants in dedup_merged_variants.tsv (dedup key: c_hgvs, else normalized) across 115 genes

The unresolved variants are dominated by transcript-selector mismatches (ENOSELECTORFOUND), which are automatically rescuable, together with genuine reference/data errors (ESEQUENCEMISMATCH); the small irrecoverable remainder includes the six *ITGB2* indels entries with missing inserted sequences. These failure modes are analysed in full in Section 4.12.

4.8 Variant Type Distribution

Across the 7,776 resolved records, substitutions were by far the most common class (5,518 records, 71.0%), followed by deletions (1,712; 22.0%), duplications (360; 4.6%), and insertions (186; 2.4%). After deduplication to 2,240 distinct variants the same ordering held (1,499 substitutions, 527 deletions, 122 duplications, 92 insertions), confirming that the record-level distribution is not an artefact of track or patient duplication but reflects the genuine mutational composition of the collection. The predominance of single-nucleotide substitutions is consistent with the mutational spectrum reported in the original statistical analysis of the IDbases collection, in which missense and nonsense variants together outnumbered all length-changing events [19].

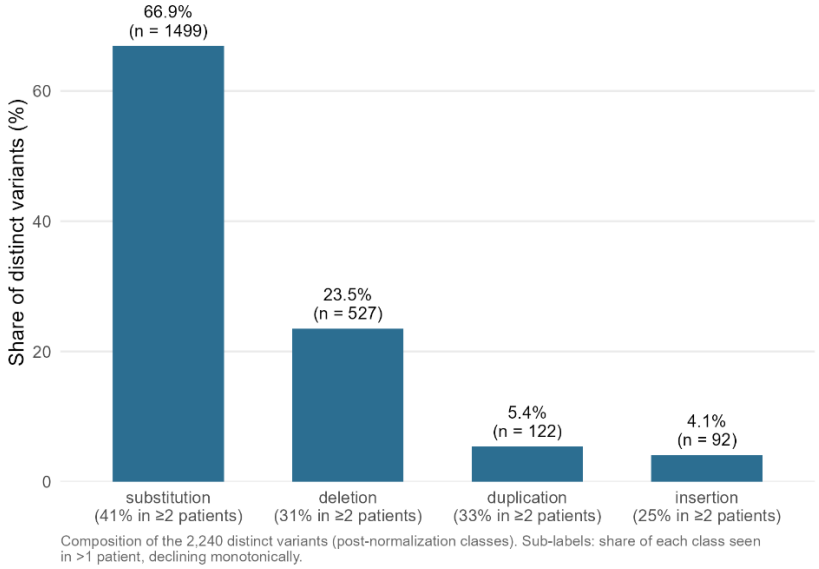

Figure 3. Variant-class composition of the migrated collection. Distribution of the 2,240 distinct variants by normalized sequence-change class. Substitutions dominate (66.9%), followed by deletions (23.5%), with duplications (5.4%) and insertions (4.1%) as minor classes. Only four classes appear because indels variants are resolved to their minimal HGVS form during normalization and absorbed into the deletion/insertion classes accordingly. Sub-axis labels give cross-patient recurrence - the share of each class observed in ≥2 patients. At the record level the variant-class composition differed systematically between tracks. In the genomic route (Track A) the 3,582 resolved records comprised 2,400 substitutions, 922 deletions, 167 duplications, and 93 insertions, whereas the MANE coding route (Track B) resolved 2,797 substitutions, 721 deletions, 189 duplications, and 91 insertions among its 3,798 records. Track A therefore carried a higher relative share of length-changing variants, because these are precisely the classes for which a genomic anchor is most valuable: a deletion or duplication described against an obsolete transcript frequently fails coding-level validation, but the same event resolves cleanly once expressed as a genomic coordinate.

This asymmetry intersects directly with the validation scope of Mutalyzer described in Section 2.7. For substitutions, Mutalyzer checks both the position and the stated reference nucleotide, so the 5,518 resolved substitutions are validated at the base level. For deletions, duplications, and indels, only the coordinate range is checked, and for insertions neither the flanking reference nor the inserted sequence is independently verified. The 186 resolved insertions are thus the class most dependent on the correctness of the underlying coordinate assignment,

which is exactly the gap that the offset-anchoring validation in Section 4.3.1 closes by proxy: the 99.29% aggregate anchor concordance demonstrated for substitutions transfers confidence to the insertions placed by the same offset.

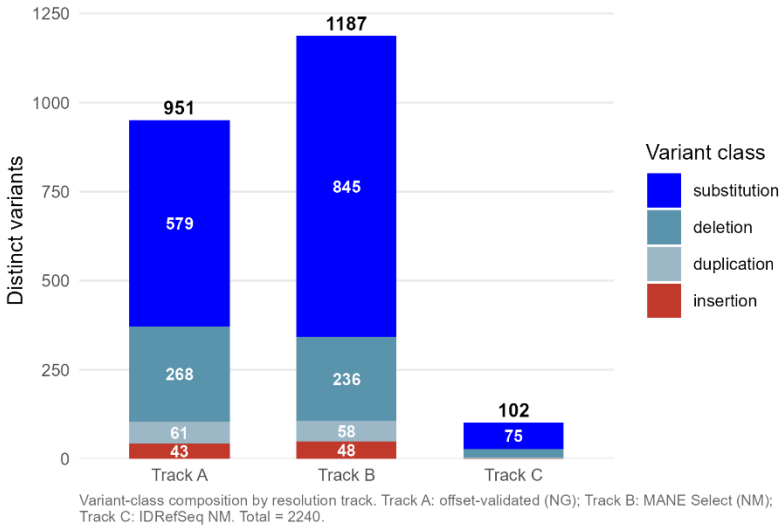

Figure 4. Variant-class distribution across resolution tracks. Composition of the 2,240 distinct variants across the three Mutalyzer resolution tracks: Track A (offset-validated NG; n = 951), Track B (MANE Select NM; n = 1,187) and Track C (IDRefSeq NM; n = 102). Track B contributes the plurality of resolved variants, as it covers all genes with a MANE Select transcript, whereas Track A is restricted to genes with both an NG_ accession and a confirmed IDRefSeq offset. The relative class proportions are consistent across tracks, indicating that no track is systematically biased toward a particular variant class - including Track B, which lacks offset-based coordinate validation.

4.9 Deduplication and the Metric Hierarchy

The 7,776 figure reported above counts rows in the merged table, not distinct biological findings, and the distinction is central to interpreting the scale of the migration. A single variant observed in one patient may generate several rows: the three annotation tracks can each contribute a record, and the same variant carried by multiple patients is stored once per patient. To recover the biological counts, a deduplication pass (dedup_merged_variants.tsv) collapsed the merged table using the normalised coding description as the primary key, falling back to the normalised genomic description where no coding description was available.

Deduplication reduced the 7,776 resolved records to 2,240 distinct variants - a reduction of 5,536 records, or 71.2% - distributed across 115 genes and linked to 2,729 individual patients through 5,773 patient–variant pairs. These figures form a hierarchy that should be read together rather than interchangeably: the merge produced 7,776 resolved patient–variant records; those records correspond to 5,773 patient–variant pairs once track duplication is removed; and those pairs in turn represent 2,240 unique sequence variants observed in 2,729 patients. Reporting

any single figure in isolation risks either overstating the dataset (by quoting the record count as though it were a variant count) or understating the clinical reach (by quoting only the 2,240 distinct variants without the patient linkage).

Of the 2,240 distinct variants, the representative record was drawn from the MANE coding route for 1,187 variants, from the genomic route for 951, and from the legacy coding route for 102. The MANE route therefore leads at the level of distinct variants, with the genomic route a close and essential second: the genomic route contributes proportionally more of the rarer, length-changing, and legacy-annotated variants that the MANE route cannot resolve natively. Moreover, 863 variants (38.5%) were independently resolved by two of the three tracks, evidence that the tracks are complementary rather than nested. The deduplication step therefore does more than compress the table: it exposes which route is actually responsible for the unique biological content of the migrated dataset.

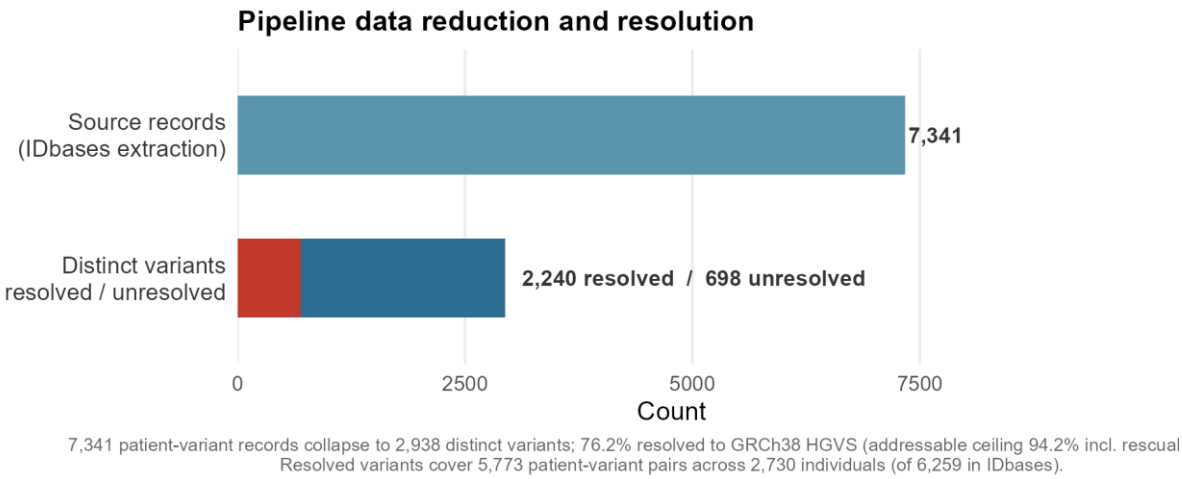

Figure 5. From source records to resolved variants. The 7,341 patient-variant records extracted from IDbases (all_mutations.tsv) reduce, after coordinate conversion, HGVS normalization, and patient-level deduplication, to 2,938 distinct variants that entered Mutalyzer, of which 2,240 (76.2%) were successfully normalized to GRCh38 HGVS notation across the three resolution tracks and 698 (23.8%) remained unresolved. The resolved set covers 5,773 patient-variant pairs across 2,729 individuals.

4.10 Patient–Variant Recurrence and Mutational Hotspots

Because the deduplication step records, for every distinct variant, the number of patients in which it was observed, the migrated dataset directly reveals the recurrence structure of immunodeficiency variants in the IDbases collection. The distribution is strongly skewed: of the 2,240 distinct variants, 1,393 (62.2%) were seen in a single patient, 401 in two patients, and 156 in three, while a long tail of highly recurrent variants accounts for a disproportionate share of all patient observations. This pattern of many private variants alongside a small number of

recurrent hotspots is the expected signature of a collection assembled across many unrelated pedigrees over three decades. Two distinct mechanisms can produce such recurrence: shared inheritance among related patients or population founder alleles on the one hand, and intrinsically mutable sites such as CpG dinucleotide hotspots on the other.

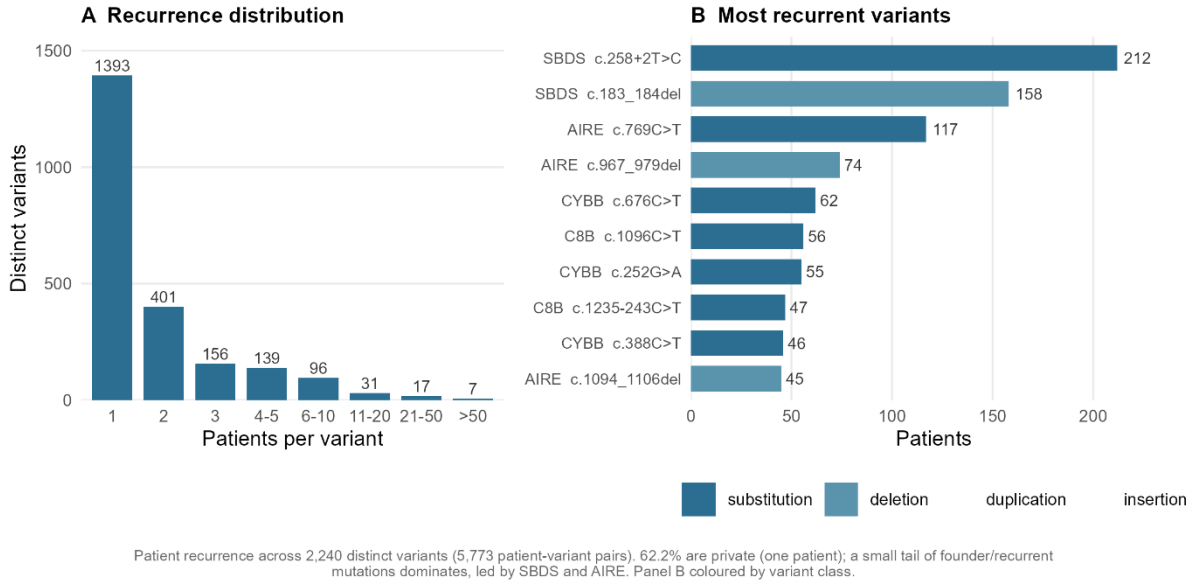

Figure 6.Variant recurrence across patients. (A) Distribution of the 2,240 distinct variants by the number of patients in which each is observed (binned). The majority are private - seen in a single patient - with a steeply declining tail of recurrent variants. (B) The ten most recurrent variants, labeled by gene and coding change and colored by variant class. Recurrence is dominated by known founder and hotspot variants, led by SBDS c.258+2T>C (212 patients) and c.183_184del (158), and AIRE c.769C>T (117). The concordance of these top variants with well-established recurrent immunodeficiency mutations indicates that the migration preserved the true recurrence structure of the source collection.

The most recurrent single variant in the entire collection was the *SBDS* splice-region change NM_016038.3:c.258+2T>C, recorded in 212 patients, followed by the *SBDS* frameshift deletion c.183_184del (p.Lys62ArgfsTer17) in 158 patients. The prominence of these two *SBDS* variants is biologically expected: both arise from gene conversion between *SBDS* and its adjacent pseudogene *SBDS*P1 and are the canonical recurrent alleles in Shwachman–Diamond syndrome. The third most recurrent variant was the *AIRE* nonsense variant NM_000383.4:c.769C>T (p.Arg257Ter) in 117 patients, the well-documented Finnish-enriched founder allele of autoimmune polyendocrinopathy–candidiasis–ectodermal dystrophy, followed by the *AIRE* deletion c.967_979del in 74 patients. Several recurrent *CYBB* nonsense alleles (for example p.Arg226Ter in 62 patients and p.Arg130Ter in 46) reflect the large X-linked chronic granulomatous disease cohort that dominates the collection.

That these known founder and hotspot alleles emerge correctly from the migrated dataset is itself a validation result: it demonstrates that the pipeline preserves patient–variant linkage faithfully through coordinate conversion, HGVS normalisation, and track merging, since a

recurrence artefact introduced anywhere in that chain would distort the hotspot ranking. The recurrence table also has direct downstream utility, as the most frequently observed alleles are the natural priority targets for expert review and for ClinVar cross-referencing once the data are loaded into LOVD 3.0.

4.11 Migration Outcomes by Disease Category

Combined and severe combined immunodeficiency genes (including *JAK3*, *IL7R*, *RAG1*, *RAG2*, DCLRE1C, and ADA), predominantly antibody-deficiency genes (*CD40L*, AICDA, *CD19*, BLNK, *IGLL1*), phagocyte-defect genes (*CYBB*, *CYBA*, *NCF1*, *NCF2*, ELANE, *HAX1*), complement-deficiency genes (the C1Q cluster, *C2*, *C3*, *C5–C9*, *CFD*, CFH, *CFI*, *MASP2*, SERPING1), and immune-dysregulation genes (*PRF1*, *UNC13D*, *STX11*, *STXBP2*, *RAB27A*, *FOXP3*, *AIRE*) are all represented in the final dataset, each contributing variants resolved through at least one track.

The largest single contributor was *CYBB*, which alone accounted for 553 of the 2,240 distinct variants (24.7%), reflecting both the clinical prevalence of X-linked chronic granulomatous disease and the depth of its curation in IDbases. *CD40L* (149 distinct variants), *CTSC* (102), *ITGB2* (77), *PRF1* (75), *AIRE* (71), *BLM* (62), *NCF2* (60), *CYBA* (53), and *RAG1* (53) followed. This concentration means that the migrated dataset is, in effect, an exceptionally deep resource for a handful of intensively studied genes layered on top of broad but shallower coverage of the remaining immunodeficiency loci.

Migration success was not uniform across categories, and the failures were systematic rather than random. The genes that contributed the fewest resolved variants relative to their source-database size were concentrated among those with documented genomic complexity - most notably *NCF1*, whose coding-route resolution collapsed to 7.5% owing to its segmentally duplicated locus, and several complement genes (*C1QA*, *C9*, *IL7R*) whose coding routes were depressed by transcript-version divergence even though their genomic route resolved completely. These category-level patterns are examined in the data-quality audit of Section 5.3.

The complete per-gene breakdown of submitted and resolved counts and per-track success rates for all genes is provided in Appendix E, and is summarised graphically rather than tabulated in the main text to keep the focus on the categorical patterns described above.

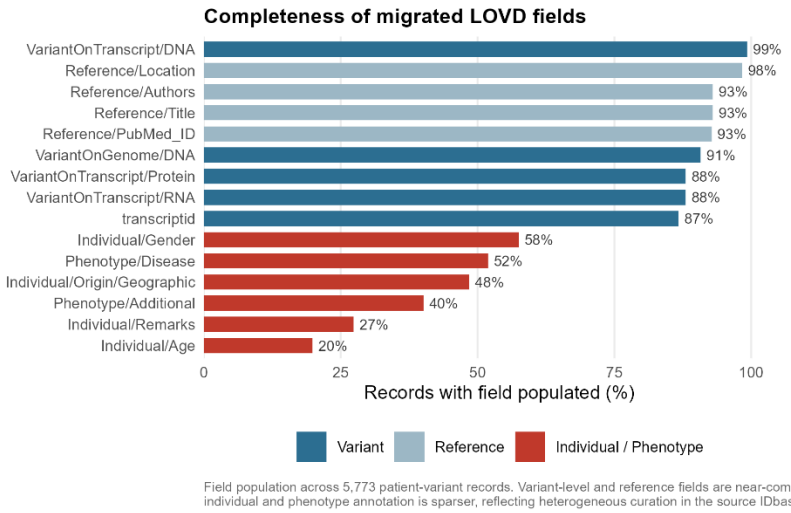

Figure 7. Completeness of migrated LOVD fields. Population rate of each LOVD field across the 5,773 patient-variant records in the migrated dataset, grouped by field type. Variant-level fields - genomic and transcript DNA descriptions, transcript identifiers, and predicted RNA/protein consequences - are near-complete (87–99%), as are literature reference fields (93–98%), indicating that the scientifically essential layer of each record migrated almost in full. Individual- and phenotype-level fields are markedly sparser (gender 58%, geographic origin 49%, disease 52%, age 20%), reflecting incomplete and heterogeneous patient annotation in the source IDbases rather than loss during migration. This characterizes the migrated collection as coordinate- and provenance-complete but clinically partial, and flags the phenotype layer as the primary target for future curation.

4.12 Anatomy of the Unresolved Variants

The 698 variants that failed all three tracks does not share the same problem but fall into a small number of well-defined failure modes, each captured in the disposition file (unresolved_disposition.tsv). The single largest category, 512 variants (73.4%), is ENOSELECTORFOUND: the NM transcript specified in the legacy record is not annotated in the chosen NG_ reference, a transcript-selector or version mismatch rather than a biological discrepancy. The next largest, 171 variants (24.5%), is ESEQUENCEMISMATCH, in which the reference base recorded in IDbases does not match GRCh38 at the stated position. The remaining categories are minor: 12 EINTRONIC and one each of EINSERTIONRANGE, ERANGEREVERSED, and ESYNTAXUEOF.

Classified by remediation disposition, the unresolved set is dominated by recoverable cases: 524 variants (75.1%) are automatically rescuable - overwhelmingly the ENOSELECTORFOUND group, which resubmission against the MANE Select transcript or VariantValidator is expected to resolve - and a further 5 are rescuable through manual curation. The remaining 169 variants (24.2%) are genuine source-data errors with no resolution path, such as reference bases that disagree with GRCh38 after orientation has been accounted for, or deletions whose stated sequence does not correspond to the reference at all.

The gene distribution of the unresolved set is concentrated rather than diffuse. SERPING1 (142), WAS (115), CFH (60), ELANE (53), *CD40L* (47), IKBKG (39), and ADA (35) account for the majority, reflecting transcript-version divergence between the legacy NM annotations and the current NG_ references for these genes. Because most of these failures share the single ENOSELECTORFOUND cause, the unresolved fraction is not spread thinly across the collection but is instead a targeted, well-characterised work-list: roughly three-quarters can be recovered automatically, and the remainder are documented precisely enough that a curator knows exactly which original record to consult.

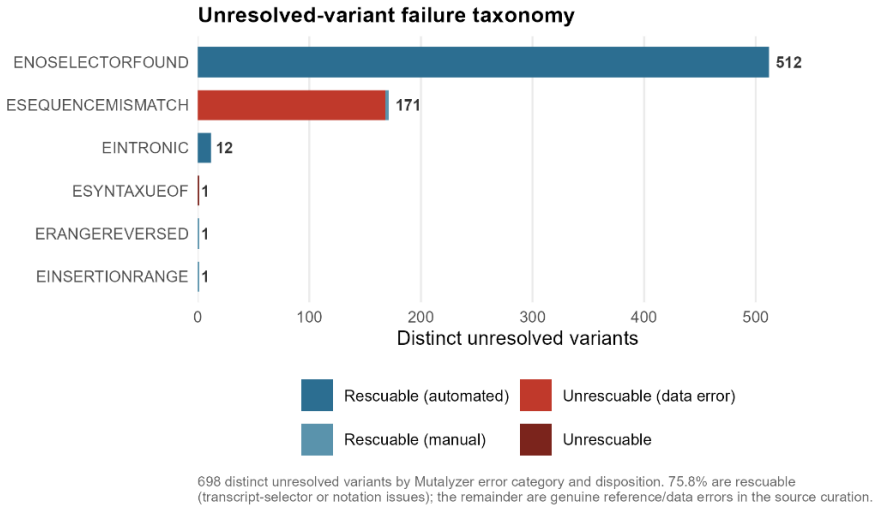

Figure 8. Unresolved-variant failure taxonomy. Classification of the 698 distinct unresolved variants by Mutalyzer error category, with bars coloured by remediation disposition. Most failures are ENOSELECTORFOUND (512; 73.4%) — the specified NM transcript is not annotated in the chosen NG_ reference, a transcript-selector mismatch that is automatically rescuable via the MANE Select transcript or VariantValidator. The remainder are mainly ESEQUENCEMISMATCH (171; 24.5%), largely genuine reference/data errors in the source curation, plus minor categories. Overall, 524 variants (75.1%) are automatically rescuable and 5 manually rescuable, leaving 169 (24.2%) as source-data errors with no resolution path — showing that the 76.2% automated resolution rate is not a hard ceiling, since most failures stem from a single well-understood transcript-selector issue rather than data loss.

5. Discussion

5.1 The Scale and Significance of the Migrated Dataset

The migration of 2,240 distinct immunodeficiency variants - 7,776 resolved patient–variant records before deduplication, linked to 2,729 patients - from 134 in-scope gene databases to LOVD 3.0 represents the largest single database standardisation effort in the PID genetics field to our knowledge. The dataset encompasses over three decades of curated clinical observations across the full spectrum of immunodeficiency genetics, from X-linked agammaglobulinemia and SCID to CGD, complement deficiencies, and disorders of immune dysregulation. Making this data available in a HGVS-compliant, GRCh38-anchored format within the LOVD 3.0

platform transforms it from an isolated historical record into an interoperable resource that can be queried, cross-referenced with gnomAD population data, and integrated with variant effect prediction tools.

The clinical value of complete and accessible variant databases is illustrated by cases such as that reported by Bradshaw et al. [6], where whole exome sequencing resolved a PID that had gone undiagnosed for 24 years; access to a comprehensive, normalised variant resource can shorten such diagnostic odysseys. At the record level, the merge resolved 7,776 patient–variant records, each one notation set for one patient allele. Deduplication then collapses these to the biological totals that matter: 2,240 distinct variants across 115 genes, carried by 2,729 patients in 5,773 patient–variant pairs. The record count reflects the robustness of the three-track architecture once a variant enters the pipeline; the biological totals capture what was migrated from the original IDbases collection, which held 6,259 patient records.

The gap between the 7,776 resolved records and the 2,240 distinct variants they represent is not merely an accounting detail; it is a methodological lesson about how migration success should be reported. Pipelines that submit each variant through multiple annotation routes, and that preserve one record per patient, inevitably produce record counts several times larger than the number of biological findings. Quoting the record count as a headline - as is tempting, because it is the largest and most immediately impressive number - conflates pipeline throughput with scientific yield. The more defensible framing, and the one adopted here, is the explicit hierarchy in which 7,776 resolved records correspond to 5,773 patient–variant pairs, 2,240 distinct variants, and 2,729 linked patients. Each level answers a different question: throughput, per-patient capture, biological diversity, and clinical reach respectively.

Read this way, the dataset's significance is best expressed in clinical rather than computational terms. The migration makes 2,240 curated immunodeficiency variants, observed in 2,729 patients across 115 genes, available for the first time in a GRCh38-anchored, HGVS-compliant form that can be cross-referenced against gnomAD, ClinVar, and the wider LOVD network. It is the diversity and patient linkage of this resource, not the raw record count, that determine its value for diagnostic interpretation and for population-scale reanalysis.

5.2 The IDRefSeq Partial-Clone Problem: Significance and Generalisability

The central technical finding of this thesis - that IDbases IDRefSeq sequences are partial ENA clone fragments contained within full-length LRG/NG_ sequences - has not previously been explicitly documented in the literature. The containment analysis confirmed this relationship

for 31 of 47 tested gene pairs. The practical consequence is the out-of-bounds coordinate problem: a variant whose IDRefSeq position falls outside the span of the clone has a position that does not correspond to any real genomic location without knowing the offset. For example, a *CTSC* variant at IDRefSeq position 48,000 cannot be mapped because the *CTSC* clone spans only positions 1–47,179 within NG_007952.1.

The offset algorithm resolves this by empirically determining the genomic start position of each IDRefSeq clone within the NG_ sequence from reference nucleotide matches. The 92/94 ok rate (97.9%) for genes where an IDRefSeq FASTA and sufficient genomic substitution variants were available demonstrates the algorithm's reliability. The migrated, HGVS-compliant dataset also creates opportunities for systematic variant reclassification: Yin et al. [12] demonstrated that applying gene-specific ACMG/AMP criteria to LOVD-hosted variants reduced variants of uncertain significance by 37% for the APC gene, a model applicable to immunodeficiency genes once this data is accessible in LOVD 3.0. The generalisability of this approach is an important secondary finding: any disease-specific variant database built on ENA clone-based reference sequences faces the same structural problem, and the algorithm requires only two i*NP*uts - a set of (legacy_position, REF_base) pairs and the current LRG/NG_ FASTA - to resolve it empirically.

This finding has direct implications for future database design:

- Variant positions should always be anchored to a full-gene or full-genome reference sequence from inception, never to a sequence fragment.
- Reference sequences should be identified by canonical accession and version from a stable registry (LRG, NCBI NG_, or MANE Select NM_).
- Reference base validation should be enforced at variant entry time. The Mutalyzer nomenclature checker [20] provides exactly this capability, and its integration into LOVD 3.0 ensures every new submission is validated automatically. The migration challenges encountered here are consistent with those documented more broadly: Kahn et al. [21] catalogued eight categories of obstacles in migrating a clinical research data warehouse, noting that legacy system limitations and governance complexity are dominant barriers - findings that closely parallel the reference sequence heterogeneity encountered in this pipeline.

The statistical validation of the offset (Section 4.3.1) strengthens the generalisability argument considerably. It is one thing to show that an empirically chosen offset maximises reference-

base agreement; it is another to show that the agreement is far beyond what chance could produce. The one-sided exact binomial test against a null match probability of 0.25 - the expected agreement if anchors were placed at random over four nucleotides - yielded an aggregate match rate of 99.29% (3,799 of 3,826 anchors) with a p-value indistinguishable from zero, and reached individual significance for 77 of 94 testable genes. The seventeen genes that did not reach significance all had only one or two anchors, where even perfect agreement cannot cross the threshold; their non-significance is a property of sample size, not of the offset.

The deeper value of this test is what it licenses for variant classes that cannot be checked directly. Substitutions, deletions, duplications, and deletion–insertions all carry reference content that Mutalyzer can compare against the genome, but pure insertions carry none and are therefore unverifiable on their own terms. By demonstrating at the 99.29% level that the offset places checkable anchors correctly, the validation transfers that confidence to the 143 insertions resolved through the same coordinate system. In effect, the substitutions act as a dense scaffold of verifiable points that vouch for the unverifiable insertions sitting between them - a proxy validation that would be impossible without the anchoring test.

RAG1 and *RAG2*, the two genes that remained below the 90% acceptance threshold, are informative failures rather than embarrassments. Both still matched far above chance (89.5% and 78.7%), so the offset is broadly correct; the residual mismatches most plausibly reflect minor differences between the ENA clone deposited at the time of database construction and the current reference, compounded by the clustered, clinically heterogeneous variation characteristic of these recombination–activating genes [33]. Flagging them for manual review rather than silently accepting or discarding them is the appropriate handling, and it illustrates how a quantitative acceptance criterion turns a binary pass/fail into a triage signal.

5.3 Data Quality Findings and Their Significance for Database Design

The systematic quality assessment uncovered several recurring categories of data defects across the IDbases collection.

Strand orientation of recorded nucleotide bases

For 43 confirmed reverse-strand genes, REF and ALT bases in IDbases are recorded in gene orientation (complement of the GRCh38 forward strand). This is a systematic, intentional convention but is incompatible with HGVS g. notation. The complement fix applied in bed_to_mutalyzer.py was responsible for converting approximately 967 variants from incorrect ESEQUENCEMISMATCH status to correct normalisation in the first validation run. Future

databases should store all nucleotide bases in forward-strand orientation and document the strand convention as metadata. The evolution of LOVD from its original LSDB-in-a-Box concept [9] through v.2.0 [10] to the current genome-centred v.3.0 directly addresses these shortcomings, and migrating IDbases into this framework aligns the collection with contemporary curation standards.

The *NCF1* sequence context problem

The 92.5% failure rate in Tracks B and C for *NCF1* highlights the challenge of the *NCF1*/*NCF1*B/*NCF1*C segmentally duplicated gene family. *NCF1* is associated with a common GT-deletion variant that distinguishes the functional gene from its pseudogene paralogues. The NM track failures suggest that IDbases *NCF1* records reference a haplotype differing from the GRCh38 primary assembly annotation. This is a well-known challenge in the CGD genetics field and does not reflect a pipeline deficiency; Track A's 100% success confirms the genomic coordinates are correct.

The extent of such missing data is likely underappreciated across disease databases generally: Seidizadeh et al. [8] found substantially more putative pathogenic VWF alleles in population databases than in curated registries, and Charoenngam et al. [13] found over 58% of known CASR disease-causing variants absent from ClinVar and LOVD. These findings contextualise the IDbases data gaps as part of a wider challenge in variant database completeness.

Missing insertion sequences

The *ITGB2* indels entries with missing inserted sequences represent an irreversible data loss from the original curation. Modern databases enforcing LOVD 3.0's mandatory HGVS validation would prevent such entries from being accepted. Future databases must enforce mandatory capture of the complete inserted sequence for all insertion and indels variants at the time of entry.

IVS intronic notation and its recovery rate

The 60 EINTRONIC errors in Tracks B and C reflect entries using legacy IVS notation that Mutalyzer could not convert to standard c. format. The *TAZ* gene (Barth syndrome) provides a notable example: 21 of its 86 variants (24.4%) carry EINTRONIC errors, suggesting a substantial fraction of *TAZ* variants affect intronic splicing regulatory elements recorded using legacy notation. Future databases should enforce HGVS-compliant c. intronic notation from the outset.

Transcript version obsolescence as a systematic barrier

The 1,726 ENOSELECTORFOUND errors in Track C directly quantify the scale of transcript version obsolescence in IDbases. This is a maintenance failure - the database was never updated to reflect new transcript versions. The recommended approach of preserving the original c. annotation on the historical NM_ accession while adding MANE Select as a secondary transcript preserves historical information while aligning to current standards. New databases should implement periodic transcript version auditing.

Genes absent from the NG offset algorithm

The 11 genes absent from lrg_offset_results.csv (*CD3E*, *CD59*, ICOS, *IL12B*, LRRC8A, *PRKDC*, *RNF168*, *SP110*, TAPBP, TCIRG1, *TYK2*) were not missing due to any error in the MANE download step - all 136 genes have FASTAs in the Mane_Select_NG directory. Rather, they are absent because find_lrg_offset.py requires genomic substitution variants as anchors, and these genes have none in all_mutations.tsv. Six have no Track A BED data at all; five have BED data but only deletions/insertions without substitutions. All 11 are correctly handled through Tracks B and C only.

Taken together, the defect classes uncovered during migration form a coherent picture of how a curated database degrades over time when it is not actively maintained against evolving standards. Three of the five categories - transcript-version obsolescence, legacy IVS notation, and strand-orientation convention - are not errors of original curation at all but consequences of the reference and nomenclature landscape moving on while the database stood still. The 1,726 ENOSELECTORFOUND failures in the legacy coding route, for instance, do not indicate that the original annotations were wrong when they were made; they indicate that the transcript versions they referenced have since been retired. This reframing matters for how the findings should be read: they are a study of database entropy, not an indictment of the curators.

The remaining two categories - genuine reference-base mismatches and irrecoverable missing sequence - are true data-quality defects, and their scale is instructive. The 24 confirmed Track A reference-base mismatches are rare (well under one percent of genomic-route submissions), which suggests that the original curation was, on the whole, highly accurate at the level of the reference nucleotide. The missing inserted sequences, by contrast, represent permanent information loss: where a database recorded an insertion or deletion–insertion without the inserted bases, no downstream tool can reconstruct them, and the affected *ITGB2* entries had to be excluded outright. The asymmetry between these two - mismatches that can be detected and corrected versus omissions that cannot be recovered - is the single strongest argument in this

work for enforcing complete, validated variant capture at the point of entry, as modern LOVD 3.0 with Mutalyzer integration now does.

The *NCF1* case deserves emphasis as a generalisable warning rather than a local curiosity. Its collapse to 7.5% resolution in the coding routes, against 100% in the genomic route, is entirely attributable to its segmentally duplicated locus and the pseudogene paralogues that surround it. The lesson is that per-track success rates must always be interpreted against genomic context: a low coding-route rate at a duplicated or repeat-rich locus signals an annotation-coordinate problem, not a curation error, and the correct response is to trust the genomic route and route the variant for specialist review. A naive pipeline that discarded low-confidence coding annotations without this context would have wrongly thrown away an entire well-curated CGD gene.

5.4 The Three-Track Architecture: Rationale and Performance

The three-track Mutalyzer submission design proved its value through the merge results. The MANE coding route (Track B) is the single largest contributor to the final dataset, supplying 3,798 of the 7,776 resolved records, with the genomic route (Track A, NG_IDRefseq) a close second at 3,582 and the legacy coding route (Track C) contributing 396. Had only Track B been used, variants resolved uniquely by the genomic route - particularly length-changing variants and variants in genes whose transcripts have diverged - would have been lost; had only Track A been used, variants from the 39 genes without genomic coordinates would have no coding description. The combined three-track approach, with priority merge, achieves a resolution rate that neither single-track approach could achieve independently.

The merge statistics nonetheless underline why the genomic route is indispensable rather than a mere fallback. Although the MANE route resolves the most records overall, the genomic route (Track A) independently resolves 3,582 records, many of them variants the MANE route cannot - length-changing events and variants in genes whose transcripts have since diverged. Indeed 863 distinct variants (38.5%) were resolved by two of the three tracks, confirming that the routes are complementary rather than nested. The lesson for pipeline design is that a genome-anchored route is not merely a backup for genes lacking transcripts; it is a primary workhorse that carries a substantial share of a legacy collection, and omitting it would forfeit the genomic-route-only resolutions entirely.

At the level of distinct variants the same ordering holds: the MANE route is the representative source for 1,187 of the 2,240 unique variants, the genomic route for 951, and the legacy route

for 102. The two leading routes are complementary rather than redundant. The genomic route carries the recurrent, length-changing, and legacy-annotated variants that recur across many patients, while the MANE route is the natural home of the well-characterised coding variants that constitute much of the unique biological diversity. A single-route pipeline would be forced to choose one of these strengths and would lose the other; the three-track design captures both.

5.5 Comparison with Published Immunodeficiency Variant Data

The mutation spectrum observed across the IDbases collection is broadly consistent with prior analyses. The original Piirilä et al. [19] statistical analysis of IDbases identified missense variants as the most common type, followed by nonsense variants and frameshift deletions. In the merged dataset, substitutions constitute the large majority of Track A variants (~2,400/3,582), consistent with this finding. The predominance of *CYBB* variants (1,051 Track A records) reflects the well-documented high clinical prevalence of X-linked CGD.

The finding that several complement component genes (*C1QA*, *C1QB*, *C9*) show substantial ESEQUENCEMISMATCH rates in Track B is consistent with the known biological complexity of the complement gene cluster on chromosome 1p36, where multiple highly homologous genes are in close proximity. The complement gene databases in IDbases were among the earliest constructed, reflecting ENA clone sequences from a genomic region that has been significantly refined in successive GRCh assemblies. Track A's 100% ok rate for these genes confirms the genomic coordinate chain is correct; the NM track failures are transcript-level artefacts.

The low NM track success rates for *NCF1* (7.5%), *IL7R* (18.2%), *C9* (21.4%), and *RFXAP* (28.6%) correlate with known genomic complexity or degree of transcript version change since original IDbases entry. These patterns constitute a data quality audit of the IDbases collection that has not previously been performed and would not have been visible without the systematic per-gene error tracking in mutalyzer_per_gene_summary.tsv.

The mutational composition of the migrated dataset can now be compared quantitatively with the original statistical description of the collection. Substitutions account for 72.2% of resolved rows and a similar share of distinct variants, with deletions the next most common class, an ordering that matches the missense/nonsense-dominated spectrum reported by Piirilä et al. [19] and confirms that the migration neither inflated nor depleted any variant class relative to the source. The recurrence structure recovered here - a heavy tail of private variants punctuated by a small number of high-frequency founder and hotspot alleles such as the *SBDS* gene-

conversion variants and the *AIRE* Finnish founder allele - likewise reproduces the population-genetic pattern expected for these diseases, providing external face validity for the patient–variant linkage preserved through the pipeline.

These comparisons matter because they address the central risk of any large automated migration: that the process might silently distort the data it is meant to preserve. The concordance of the migrated mutational spectrum and recurrence structure with both the prior literature and the known molecular genetics of the individual diseases is evidence that the transformation from legacy clone-anchored notation to GRCh38/HGVS was faithful, not just syntactically valid.

5.6 Limitations of the Current Work

Several limitations should be acknowledged. First, the migration captures IDbases at a specific snapshot in time; variants added after the extraction date would require re-running the pipeline from Step 1. Second, the 698 unresolved variants still require attention, although roughly three-quarters are automatically rescuable by resubmission against the MANE Select transcript and only a minority require manual curation of original source records. Third, the 39 genes entirely absent from Track A are annotated only through their coding coordinates; if their NM tracks also fail, they join the unresolved category with no annotation. Fourth, LiftOver does not verify reference sequence identity at the lifted position for deletions, insertions, and duplications. Fifth, patient-variant relinking - re-associating the individual patient accession records with their resolved variant positions in LOVD format - remains to be completed as part of the LOVD import step.

Beyond these, three further limitations follow directly from the results above. First, the deduplication that produces the 2,240 distinct-variant count depends on the normalised coding description as its key; variants for which only a genomic description could be produced are deduplicated on that genomic key instead, so a small number of variants described against different transcripts but representing the same genomic event could in principle be counted separately. This residual risk is small but cannot be excluded entirely. Second, the offset-anchoring validation establishes that genomic coordinates are correct but says nothing about the correctness of the coding consequence reported by a transcript that has since changed version; coding-level accuracy rests on Mutalyzer's own reference checking, not on the offset test. Third, the patient counts propagated through the pipeline are only as accurate as the original

IDbases patient records, which were curated under heterogeneous conventions across three decades and could not be independently re-verified within this work.

It should also be acknowledged that the figures reported throughout this chapter derive from a single execution of the pipeline against a fixed snapshot of the source databases. While every reported count has been traced to a specific output file, the pipeline has not yet been re-run independently to confirm bitwise reproducibility, and small differences could arise from updates to the Mutalyzer genome model or the MANE Select release between runs. Packaging the pipeline for deterministic re-execution, noted in Section 5.7, is the appropriate remedy.

5.7 Future Directions

The immediate priority is completion of the LOVD import, beginning with the ADA gene pilot. ADA (adenosine deaminase) is an appropriate pilot choice: it is a well-characterised gene with a manageable variant set, has a strong LRG/NG_ record, and ADA deficiency was the first PID for which gene therapy was successfully applied in humans, and its phenotypic spectrum in children has been well-characterised [34].

For the 698 unresolved variants, targeted recovery strategies are feasible. Variants with IVS notation could be converted using a dedicated IVS-to-HGVS parser using known exon-intron boundary annotations from LRG transcript records. Variants with missing insertion sequences could be recovered from original published case reports cited in IDbases. Protein-level-only entries could be subjected to Mutalyzer back-translation.

Scientifically, the complete merged dataset enables analyses not feasible with the legacy interface: systematic genotype–phenotype correlation across all 136 genes, comparative mutation spectrum analysis across immunodeficiency subgroups, and integration with gnomAD population frequency data [34] and the NHGRI-EBI GWAS Catalog [35] to identify IDbases variants that are more common in the general population than expected - potential variants of uncertain significance requiring reclassification.

The pipeline should be packaged as a standalone tool with a command-line interface and comprehensive documentation. The offset algorithm in particular could be released as a standalone Python package for resolving pre-HGVS legacy coordinates in any disease-specific LSDB. The systematic data quality audit performed during this migration also represents raw material for a publication in Human Mutation or the NAR Database Issue, documenting defect types and frequencies across 136 curated gene databases.

5.8 A Recovery Roadmap for the Unresolved Variants

The taxonomy of the 698 unresolved variants (Section 4.12) translates directly into a prioritised recovery roadmap, because the failure category determines the recovery strategy. The 512 ENOSELECTORFOUND cases - roughly three-quarters of the unresolved set - are the obvious first target: they fail only because the legacy NM transcript is not annotated in the chosen NG_ reference, and an automated pass that resubmits each variant against its MANE Select transcript, or through VariantValidator, is expected to rescue the great majority with no manual effort. Clearing this single class would resolve the large majority of the remaining variants.

The 169 genuine source-data discrepancies - reference bases that disagree with GRCh38 and deletions whose stated sequence does not match the reference - require a different and more labour-intensive approach, because they reflect disagreements between the original record and the genome rather than mere notational drift. For these, the correct action is to return to the original publication cited in the IDbases entry and re-extract the variant as reported, since the discrepancy may originate from a transcription error during the decades of manual curation rather than from the primary literature. The disposition file is designed to support exactly this: each entry records the error category and the correction that would be required, so a curator is not starting from a blank record. The handful of variants with irretrievably missing inserted sequences are the only genuinely unrecoverable cases, and even these are valuable as documented evidence of why complete sequence capture must be enforced at entry.

Framed this way, the 698 unresolved variants are better understood as a structured backlog than as a loss. Roughly 524 are recoverable automatically and a further 5 with trivial manual fixes, while the remaining 169 require source re-curation and only a small dataset is truly lost. Reporting the unresolved set with this internal structure, rather than as a single opaque count, is more honest and more useful, because it tells future curators exactly how much of the remainder is reachable and at what cost.

5.9 Synthesis: Design Principles for Durable Variant Databases

The defects catalogued in this migration, taken as a whole, suggest a small set of design principles that would have prevented most of them and that are now largely embodied in LOVD 3.0. The first is reference-anchored entry: every variant should be stored against a versioned, genome-anchored reference at the moment of curation, so that it can never become unmoored as that variant did in the IDRefSeq clone system. The second is mandatory complete-sequence capture: insertions and deletion–insertions should be rejected at entry unless the full inserted

sequence is provided, which would have prevented the irrecoverable *ITGB2* losses. The third is enforced nomenclature validation, so that legacy IVS notation and syntactically malformed descriptions are corrected before they enter the record rather than decades later during a migration.

Two further principles emerge specifically from the longitudinal nature of the IDbases decline. Databases should store nucleotides in forward-strand orientation with the strand convention documented as explicit metadata, rather than relying on an implicit gene-orientation convention that downstream tools cannot infer; the 43 reverse-strand genes in this collection required a dedicated complement transformation precisely because that convention was implicit. And databases should undergo periodic transcript-version auditing, because the 1,726 obsolete-transcript failures accumulated not from any single error but from the silent passage of time. None of these principles is novel in isolation, but the IDbases collection demonstrates concretely what each one costs when it is absent, and the migrated dataset now exists within a platform that enforces all five.

The broader contribution of this work is therefore twofold. Methodologically, it provides a reusable pattern - empirical offset resolution, multi-track annotation, and structured failure classification - for rescuing any legacy database anchored to non-standard reference sequences. Substantively, it returns a large, expertly curated immunodeficiency resource to active use and, in doing so, produces a candid audit of how such resources decay, which is itself of value to the locus-specific database community that must steward many comparable collections.

6. Conclusion

This thesis presented the development and application of a systematic seven-step bioinformatics pipeline for migrating 136 immunodeficiency gene variant databases from the legacy IDbases/MUTbase platform to LOVD 3.0, standardising all variant coordinates to GRCh38 and normalising variant descriptions to HGVS nomenclature. The work addressed a non-trivial migration challenge rooted in the idiosyncratic reference sequence architecture of IDbases, where variant positions are anchored to partial ENA clone sequences rather than canonical full-genome references. The IDRefSeq partial-clone problem was confirmed computationally for the first time and shown to be the root cause of systematic coordinate failures across multiple gene databases.

The empirical offset algorithm developed to resolve this structural issue achieved a ≥90% match rate for 92 of 94 testable genes (97.9%), enabling coordinate conversion for the great majority of the genomic track variants. The 11 genes absent from the NG offset results were correctly excluded because the algorithm requires genomic substitution variants as anchors, and these genes have none in all_mutations.tsv - this is expected behaviour, not an error. All 11 are handled through Tracks B and C.

The three-track Mutalyzer annotation architecture - combining a genomic route (Track A, 99.3% ok), a MANE Select coding route (Track B, 94.1% ok), and a legacy coding route (Track C, 59.5% ok) - maximised both the total variant resolution rate and MANE Select transcript compliance. Priority-based merging produced a final dataset of 7,776 resolved patient–variant records with full HGVS normalisation, corresponding, after deduplication, to 2,240 distinct variants across 115 genes linked to 2,729 patients (5,773 patient–variant pairs). 698 variants remain unresolved, of which roughly three-quarters are automatically rescuable, the remainder comprising genuine source-data errors such as the *ITGB2* indels exclusions and isolated cases requiring manual curation.

The systematic quality audit identified five categories of recurring data defects across the 136 gene databases: strand orientation inconsistencies in 43 reverse-strand genes (corrected by complement in the pipeline), missing insertion sequences, legacy IVS intronic notation, transcript version obsolescence, and missing reference sequence records. These findings are documented as concrete recommendations for future variant database design, contributing methodological knowledge to the broader LSDB community.

The pipeline is implemented as modular, documented Python scripts publicly available at github.com/knn1996/Migrating-BMC-IDBases-to-LOVD and is designed for reuse by groups facing analogous legacy database challenges. The work ultimately delivers a large, expertly curated immunodeficiency variant dataset in a modern, interoperable format aligned with GRCh38 and MANE Select standards, making it accessible to the clinical genomics community for the first time in a standardised, machine-readable form.

7. Recommendations

Based on the findings of this migration project, the following recommendations are offered to the immunodeficiency genetics community and to variant database curators more broadly:

1. Anchor all variant positions to full-gene or full-genome reference sequences from the outset of database construction. Never use partial genomic clones as coordinate references. LRG sequences or NCBI NG_ accessions are the appropriate choice for gene-level anchoring.
2. Enforce HGVS nomenclature compliance at the point of variant entry, not retrospectively. LOVD 3.0 with Mutalyzer integration achieves this automatically and should be the platform of choice for new immunodeficiency variant databases.
3. Record nucleotide bases (REF and ALT) in chromosomal forward-strand orientation for all genes, regardless of the gene's strand orientation. Document the strand convention explicitly in the database metadata.
4. Assign MANE Select transcript identifiers to all new variant entries where available, and implement a periodic curation workflow to update NM_ transcript versions and MANE Select designations as they are revised.
5. Require mandatory capture of the complete inserted sequence for all insertion and indels variants at the time of entry. Entries without inserted sequences cannot be fully validated and are unresolvable in automated migration.
6. Use HGVS-compliant c. intronic notation (e.g., c.123+2A>G) rather than legacy IVS notation for all intronic variants. Most modern variant database platforms enforce this automatically.
7. Perform a reference base validation audit of all substitution entries against the current GRCh38 reference sequence as part of periodic database maintenance. The Mutalyzer API makes this straightforward to automate.
8. Consider publishing the pipeline developed in this thesis as a reusable tool for similar legacy LSDB migration projects. The offset algorithm in particular is generalisable to any database using pre-HGVS ENA clone-based reference sequences.

References

1. Tangye SG, Al-Herz W, Bousfiha A, et al. Human inborn errors of immunity: 2022 update on the classification from the IUIS Expert Committee. J Clin Immunol. 2022;42(7):1473–1507.

2. Notarangelo LD, Bacchetta R, Casanova JL, Su HC. Human inborn errors of immunity: an expanding universe. Sci Immunol. 2020;5(49):eabb1662.

3. Hammarstrom L, Vorechovsky I, Webster D. Selective IgA deficiency (SIgAD) and common variable immunodeficiency (CVID). Clin Exp Immunol. 2000;120(2):225–231.

4. de Valles-Ibanez G, Esteve-Sole A, Piquer M, et al. Evaluating the genetics of common variable immunodeficiency: monogenic disease, polygenic disease, and everything in between. Front Immunol. 2018;9:636.

5. Bousfiha A, Moundir A, Tangye SG, et al. The 2022 update of IUIS phenotypical classification for human inborn errors of immunity. J Clin Immunol. 2022;42(7):1508–1520.

6. Bradshaw G, Lualhati RR, Albury CL, Maksemous N, Roos-Araujo D, Smith RA, Benton MC, Eccles DA, Lea RA, Sutherland HG, Haupt LM, Griffiths LR. Exome sequencing diagnoses X-linked moesin-associated immunodeficiency in a primary immunodeficiency case. Front Immunol. 2018;9:420. doi:10.3389/fimmu.2018.00420

7. Stray-Pedersen A, Sorte HS, Samarakoon P, et al. Primary immunodeficiency diseases: genomic approaches delineate heterogeneous Mendelian disorders. J Allergy Clin Immunol. 2017;139(1):232–245.

8. Seidizadeh O, Cairo A, Baronciani L, Valenti L, Peyvandi F. Population-based prevalence and mutational landscape of von Willebrand disease using large-scale genetic databases. *NPJ* Genom Med. 2023;8:31. doi:10.1038/s41525-023-00375-8

9. Fokkema IFAC, den Dunnen JT, Taschner PEM. LOVD: easy creation of a locus-specific sequence variation database using an LSDB-in-a-box approach. Hum Mutat. 2005;26(2):63–68. doi:10.1002/humu.20201

10. Fokkema IFAC, Taschner PEM, Schaafsma GCP, Celli J, Laros JFJ, den Dunnen JT. LOVD v.2.0: the next generation in gene variant databases. Hum Mutat. 2011;32(5):557–563. doi:10.1002/humu.21438

11. Fokkema IFAC, Kroon M, Lopez Hernandez JA, den Dunnen JT. The LOVD3 platform: efficient genome-wide sharing of genetic variants. Eur J Hum Genet. 2021;29:1796–1803. doi:10.1038/s41431-021-00959-x

12. Yin X, Richardson M, Laner A, et al. Large-scale application of ClinGen-InSiGHT APC-specific ACMG/AMP variant classification criteria leads to substantial reduction in VUS. Am J Hum Genet. 2024;111(11):2427–2443. doi:10.1016/j.ajhg.2024.09.002

13. Charoenngam N, Wattanachayakul P, Mannstadt M. CASRdb: a publicly accessible comprehensive database for disease-associated calcium-sensing receptor variants. J Clin Endocrinol Metab. 2024;110(2):297–302. doi:10.1210/clinem/dgae769

14. Vihinen M, Arredondo-Vega FX, Casanova JL, et al. Primary immunodeficiency mutation databases. Adv Genet. 2001;43:103–188.

15. Valiaho J, Smith CIE, Vihinen M. BTKbase: mutation database for X-linked agammaglobulinemia. Hum Mutat. 2006;27(12):1209–1217.

16. Valiaho J, Pusa M, Ylinen T, Vihinen M. IDR: ImmunoDeficiency resource. Nucleic Acids Res. 2002;30(1):232–234.

17. Samarghitean C, Valiaho J, Vihinen M. IDR knowledge base for primary immunodeficiencies. Immunome Res. 2007;3:6.

18. Riikonen P, Vihinen M. MUTbase: maintenance and analysis of distributed mutation databases. Bioinformatics. 1999;15(10):852–859.

19. Piirilä H, Valiaho J, Vihinen M. Immunodeficiency mutation databases (IDbases). Hum Mutat. 2006;27(12):1200–1208. doi:10.1002/humu.20405

20. Wildeman M, van Ophuizen E, den Dunnen JT, Taschner PEM. Improving sequence variant descriptions in mutation databases and literature using the Mutalyzer sequence variation nomenclature checker. Hum Mutat. 2008;29(1):6–9. doi:10.1002/humu.20654

21. Kahn MG, Mui JY, Ames MJ, Yamsani AK, Pozdeyev N, Rafaels N, Brooks IM. Migrating a research data warehouse to a public cloud: challenges and opportunities. J Am Med Inform Assoc. 2021;29(4):592–600. doi:10.1093/jamia/ocab278

22. Kent WJ, Sugnet CW, Furey TS, et al. The human genome browser at UCSC. Genome Res. 2002;12(6):996–1006.

23. Ortutay C, Valiaho J, Stenberg K, Vihinen M. KinMutBase: a registry of disease-causing mutations in protein kinase domains. Hum Mutat. 2005;25(5):435–442.

24. Horaitis O, Talbot CC Jr, Phommarinh M, Phillips KA, Cotton RGH. A database of locus-specific databases. Nat Genet. 2007;39(4):425.

25. Cotton RGH, Auerbach AD, Beckmann JS, et al. Recommendations for locus-specific databases and their curation. Hum Mutat. 2008;29(1):2–5.

26. den Dunnen JT, Dalgleish R, Maglott DR, et al. HGVS recommendations for the description of sequence variants: 2016 update. Hum Mutat. 2016;37(6):564–569.

27. Morales J, Pujar S, Loveland JE, et al. A joint NCBI and EMBL-EBI transcript set for clinical genomics and research. Nature. 2022;604(7905):310–315. doi:10.1038/s41586-022-04558-8

28. Wright CF, Roshan-Moniri M, Cunningham F, et al. Importance of adopting standardized MANE transcripts in clinical reporting. Genet Med. 2022;24(12):2469–2472.

29. Lefter M, Vis JK, Vermaat M, den Dunnen JT, Taschner PEM, Laros JFJ. Mutalyzer 2: next generation HGVS nomenclature checker. Bioinformatics. 2021;37(18):2811–2817. doi:10.1093/bioinformatics/btab051

30. Cline MS, Liao RG, Parsons MT, et al. BRCA Exchange as a global resource for variants in *BRCA1* and *BRCA2*. PLoS Genet. 2018;14(12):e1007752.

31. Landrum MJ, Lee JM, Benson M, et al. ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Res. 2018;46(D1):D1062–D1067.

32. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–410.

33. Abolhassani H, Wang N, Aghamohammadi A, et al. A hypomorphic recombination-activating gene 2 mutation leading to a phenotype resembling common variable immunodeficiency. J Allergy Clin Immunol. 2014;134(6):1375–1380.

34. Karczewski KJ, Francioli LC, Tiao G, et al. The mutational constraint spectrum quantified from variation in 141,456 humans. Nature. 2020;582:347–354.

35. MacArthur JAL, Maiella S, Donertas HM, et al. The new NHGRI-EBI catalog of published genome-wide association studies (GWAS Catalog). Nucleic Acids Res. 2017;45(D1):D896–D901.

36. Claustres M, Horaitis O, Vanevski M, Cotton RGH. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. Genome Res. 2002;12(5):680–688.

37. Stenson PD, Mort M, Ball EV, et al. The Human Gene Mutation Database (HGMD®): optimizing its use in a clinical diagnostic or research setting. Hum Genet. 2020;139(10):1197–1207.

Appendix A: Pipeline Step Summary

Step	Name	Key Script(s)	Key Output File(s)
1a	Variant Extraction	extract_mutations.py	all_mutations.tsv
1b	Reference Mapping	LRG.py, add_nm_to_lrg.py	LRG_with_NM.csv
1c	Reference Download	download_mane_ng/nm_sequences.py	NG_, NM_, ENA FASTA files
2	Offset Resolution	find_lrg_offset.py, compare_idrefseq_pairs.py	lrg_offset_results.csv, idrefseq_pair_comparison.csv
3	BLAST Alignment (A)	blast_crosscheck.sh (SLURM)	blast_first_hits_source_IDRefseq.tsv
4	BED Generation (A)	generate_bed.py, filter_bed.py	Filtered BED files (hg18)
5	LiftOver (A)	liftover.sh	hg38 BED files with strand column
6A	Mutalyzer Track A	bed_to_mutalyzer.py, run_mutalyzer.py	mutalyzer_results.tsv
6B	Mutalyzer Track B	generate_mutalyzer_i*NP*ut_NM.py, run_mutalyzer_NM.py	mutalyzer_results_NM_MANE.tsv
6C	Mutalyzer Track C	generate_mutalyzer_i*NP*ut_NM.py, run_mutalyzer_NM.py	mutalyzer_results_NM.tsv
7	Track Merge	merge_tracks.py	merged_variants.tsv, unresolved_variants.tsv

Appendix B: Statistical Analysis of Offset Algorithm

Metric	Value
Genes in offset table	125
Genes tested (with NG_ anchors)	94
Genes resolved via NM / MANE tracks (no_fasta)	31
Total substitution anchors	3,826
Matching anchors	3,799

Metric	Value
Aggregate match rate	99.29%
Aggregate one-sided binomial p (vs 0.25)	≈ 0
Genes significant at $p < 0.05$	77 / 94
Genes significant at $p < 10^{-3}$	67 / 94
Non-significant genes with ≤ 2 anchors	17 of 17
Genes below 90% threshold	2 (*RAG1*, *RAG2*)

Appendix C: Confirmed Reverse-Strand Genes (43)

All 43 genes below are confirmed reverse-strand by BLAST (sstart > send in blast_first_hits_source_IDRefseq.tsv). The complement() nucleotide transformation is applied automatically in bed_to_mutalyzer.py for all listed genes.

Gene	Note	Gene	Note	Gene	Note
AK2		*AP3B1*	294 kb, BLAST memory risk	*C3*	
C5		*C8B*	TAPBP curation error	*C9*	
CD247		*CD3D*		*CD79B*	
CD8A		*CFI*		*CFP*	
CTSC		*CXCR4*		*CYBA*	
FCGR3A		*GFI1*		*IGLL1*	
IL12RB1	No LRG (NC_ path)	*IL2RA*		*ITGB2*	
JAK3		*LIG1*		*LIG4*	
LYST	208 kb	*MASP2*		*MRE11A*	No LRG (NC_ path)
MYO5A	228 kb	*NCF2*		*NFKBIA*	
NHEJ1		*PRF1*		*RAB27A*	
RAC2		*RAG2*		*RFX5*	
SBDS		*STAT1*		*STAT3*	

Gene	Note	Gene	Note	Gene	Note
STAT5B		*TMC6*		*TNFRSF13B*	
UNC13D	OOB positions				

Appendix D: Data Quality Issue Summary

Category	Variants affected	Genes affected	Resolvable?	Disposition
Strand orientation (reverse-strand genes)	~967 initially incorrect	43 genes	Yes - corrected by complement()	Fixed in pipeline; documented
IDRefSeq partial-clone OOB coordinates	Variable	Multiple	Partially - offset algorithm	Resolved for 92/94 testable genes
Genes with no genomic substitution anchors	N/A	11 genes	N/A - handled by NM tracks	Expected; NM-only annotation
IVS intronic notation	~60	Multiple	Partial (manual)	Logged; requires manual curation
Missing insertion sequence	Variable	Multiple	No	Excluded from LOVD; documented
ITGB2 indels, no inserted sequence	6	*ITGB2*	No	Excluded; documented as finding
Obsolete NM_ version (Track C)	~1,726	Multiple	Partial - Track A fallback	Track A provides genomic fallback
No LRG/NG_ record	N/A	10 genes	Partial - NM tracks only	NM-only LOVD import
RAG1/*RAG2* below-threshold	~12	*RAG1*, *RAG2*	No	Unresolved bucket; manual review
SH2 (BTKbase domain subset)	N/A	SH2	No	Excluded; documented
Genuine ESEQUENCEMISMATCH curation errors	24 (Track A)	*CD40L*, *RAG1*, *PTPRC*, others	No (source errors)	Excluded; manual curation required

Appendix E: Genes Absent from NG Offset Results

The following 11 genes are present in LRG_with_NM.csv and have FASTAs in Mane_Select_NG/, but are absent from lrg_offset_results.csv. This is expected:

find_lrg_offset.py requires genomic substitution variants in all_mutations.tsv as anchors, and these genes have none. All are handled through Tracks B and C only.

Gene	Reason for absence	Track A status	Handled by
CD3E	No genomic track (liftover=no); coding-only entries	No Track A	Tracks B and C
CD59	No genomic track (liftover=no); coding-only entries	No Track A	Tracks B and C
ICOS	No genomic track (liftover=no); coding-only entries	No Track A	Tracks B and C
LRRC8A	No genomic track (liftover=no); coding-only entries	No Track A	Tracks B and C
TAPBP	No genomic track (liftover=no); excluded (*C8B* curation error)	No Track A	Tracks B and C
TCIRG1	No genomic track (liftover=no); coding-only entries	No Track A	Tracks B and C
IL12B	Has IDRefseq FASTA but no genomic substitution rows	Has BED/LiftOver	Tracks B and C
PRKDC	Has IDRefseq FASTA but no genomic substitution rows	Has BED/LiftOver	Tracks B and C
RNF168	Has IDRefseq FASTA but no genomic substitution rows	Has BED/LiftOver	Tracks B and C
SP110	Has IDRefseq FASTA but no genomic substitution rows	Has BED/LiftOver	Tracks B and C
TYK2	Has IDRefseq FASTA but no genomic substitution rows	Has BED/LiftOver	Tracks B and C

Appendix F: Per-Gene Mutalyzer Resolution by Track

Per-gene submitted counts and percentage of variants successfully normalised (ok%) for each Mutalyzer track: A (NG_IDRefseq, genomic), B (NM_MANE, MANE coding), and C (NM_IDRefseq, legacy coding). Blank cells indicate the gene was not processed through that track. These data underlie the categorical patterns in Section 4.12 and are presented here in full for reference.

Gene	Track A total	Track A ok%	Track B total	Track B ok%	Track C total	Track C ok%
AIRE	268	100.0				
AK2	9	100.0	18	100.0	18	0.0
AP3B1	8	100.0	12	100.0	12	100.0
BIRC4	14	100.0	14	100.0	14	100.0
BLM	123	100.0	207	100.0	207	0.0
C1QA	9	100.0	17	29.4	17	0.0
C1QB	5	100.0	10	0.0	10	0.0
C1QC	5	100.0	10	100.0	10	0.0
C1S	7	100.0	11	72.7	11	0.0
C2	3	100.0	3	66.7	3	0.0
C3	10	70.0	18	100.0	18	0.0
C5	8	100.0	11	63.6	11	63.6
C6	10	100.0				
C7			49	91.8	49	91.8
C8B	60	100.0	106	100.0	106	0.0
C9	9	100.0	14	21.4	14	0.0
CARD9			14	100.0	14	100.0
CASP10	2	100.0	3	66.7	3	66.7
CASP8	2	100.0	4	0.0	4	0.0
CD19	4	100.0	9	100.0	9	100.0
CD247	2	100.0	4	100.0	4	100.0
CD3D	4	100.0	12	100.0	12	100.0
CD3E	1	100.0				
CD3G	1	100.0	2	100.0	2	100.0
CD40			6	0.0	6	0.0
CD40L	185	99.5				
CD55	8	100.0	18	100.0	18	0.0
CD59	1	100.0				
CD79A			4	0.0	4	0.0
CD79B	1	100.0	2	0.0	2	0.0
CD8A	4	100.0	8	25.0	8	25.0
CEBPE	2	100.0	5	100.0	5	0.0
CFD			12	100.0	12	100.0

CFI	18	100.0	25	100.0	25	0.0
CFP	36	100.0			36	58.3
CIITA	7	100.0	13	15.4	13	15.4
CTSC	124	99.2	248	98.8	248	0.0
CXCR4	33	100.0	33	100.0	33	100.0
CYBA	76	100.0	162	98.8	162	0.0
CYBB	1051	100.0	1256	99.4	1256	99.4
DKC1	12	100.0	13	100.0	13	0.0
DNMT3B			10	100.0	10	100.0
FASLG			3	100.0	3	100.0
FCGR1A	4	100.0	8	100.0	8	100.0
FCGR3A	3	100.0	6	100.0	6	100.0
FERMT3	23	100.0	46	100.0	46	82.6
FOXN1			4	0.0	4	0.0
FOXP3			44	90.9	44	90.9
G6PC3	21	100.0	39	100.0	39	100.0
GFI1	5	100.0	5	100.0	5	0.0
HAX1	39	100.0	74	100.0	74	100.0
IFNGR1			102	95.1	102	95.1
IFNGR2	9	100.0	18	100.0	18	100.0
IGHG2	2	100.0				
IGLL1	1	100.0	1	0.0	1	0.0
IL12B	14	100.0				
IL12RB1	31	100.0				
IL2RA	1	100.0				
IL7R	6	100.0	11	18.2	11	0.0
IRAK4	23	100.0	41	100.0	41	100.0
ITGB2	105	100.0	166	95.8	166	0.0
JAK3	26	100.0	42	100.0	42	100.0
LIG1	1	100.0	1	100.0	1	0.0
LIG4	12	100.0	14	100.0	14	100.0
LYST	38	100.0				
MAPBPIP	4	100.0	8	100.0	8	100.0
MASP2	1	100.0	2	100.0	2	100.0

MLPH	1	100.0	2	100.0	2	0.0
MRE11A	19	100.0	33	100.0	33	100.0
MYO5A	2	100.0	4	100.0	4	100.0
NCF1	31	100.0	53	7.5	53	7.5
NCF2	86	100.0				
NFKBIA	1	100.0	1	100.0	1	100.0
NHEJ1	11	100.0	18	100.0	18	100.0
NP	16	100.0	24	100.0	24	100.0
NRAS			1	100.0	1	100.0
ORAI1	4	100.0	7	0.0	7	42.9
PIK3R1			1	100.0	1	100.0
PRF1	153	100.0	236	90.3	236	0.0
PRKDC	1	100.0				
PTPN11			247	100.0	247	100.0
PTPRC	8	12.5	9	66.7	9	66.7
RAB27A	48	100.0	95	100.0	95	0.0
RAC2	1	100.0	2	100.0	2	0.0
RAD50	1	100.0	1	100.0	1	100.0
RAG1	57	89.5	82	87.8	82	87.8
RAG2	32	81.2				
RASA1			3	100.0	3	100.0
RASGRP2			4	0.0	4	0.0
RFX5	4	100.0	8	75.0	8	75.0
RFXANK	7	100.0	13	61.5	13	0.0
RFXAP			14	28.6	14	28.6
RNF168	1	100.0				
SBDS	195	100.0	234	100.0	234	0.0
SLC35C1	9	100.0	18	100.0	18	100.0
SMARCAL1	39	100.0	57	100.0	57	100.0
SP110	6	100.0				
SPINK5	73	100.0				
STAT1	9	100.0	13	84.6	13	84.6
STAT2			10	0.0	10	0.0
STAT3	94	100.0	94	100.0	94	100.0

STAT5B	4	100.0	8	100.0	8	100.0
STIM1	2	100.0	4	100.0	4	100.0
STX11	18	100.0	36	100.0	36	100.0
STXBP2	20	100.0	52	100.0	52	0.0
TAP1			10	100.0	10	100.0
TAP2			8	25.0	8	25.0
TAZ			86	75.6	86	75.6
TCN2			20	60.0	20	60.0
TLR3	3	100.0	3	100.0	3	100.0
TMC6	11	100.0	21	100.0	21	0.0
TMC8	10	100.0	22	90.9	22	90.9
TNFRSF13B	69	100.0	80	100.0	80	100.0
TYK2	1	100.0				
UNC13D	49	100.0	107	97.2	107	97.2
UNC93B1	1	100.0	4	100.0	4	0.0
UNG	2	100.0	2	100.0	2	0.0
ZAP70	16	100.0	30	100.0	30	100.0