Updating a three-decade database of inherited immune-disease variants

Popular science summary · Khoi Nguyen Ngoc

Inherited immune diseases are rare, but for the people affected they can be severe: infections that are normally manageable may become repeated, dangerous, or hard to explain. In many cases, the clue is a small change in DNA. My thesis was about rescuing an old but valuable database of these changes, so that it can still be used with the tools researchers and clinicians rely on today.

For many years, researchers working with the IDbases collection recorded variants linked to primary immunodeficiency diseases. The collection covers 136 genes and thousands of patient records, which represents a large amount of careful expert work. The problem was not the biology; it was the format. The database had been built over a long period, using reference sequences and naming conventions that made sense at the time but no longer match current standards. Without a dedicated curator and regular updates, the data gradually became harder for modern software to read.

A variant database works a little like a map. To see where a DNA change is, you need to know which reference sequence the position is measured from. During the project, I found that many of the old IDbases references were not full gene references. They were shorter clone fragments sitting inside longer modern references. This matters because a position that is correct on a small fragment can become misleading when compared directly with a full-length reference. In some cases, variants recorded later in the history of the database were effectively being described outside the piece of sequence that had originally been used.

To solve this, I built a pipeline that finds where each old fragment belongs within the longer reference sequence. Once that right starting point is determined, the old coordinates can be shifted to that correct positions and then converted to GRCh38, the current human genome reference. Then Mutalyzer was used to rewrite and check the variants in HGVS nomenclature, the standard way of describing genetic variants.

The final output is a cleaned and standardised dataset ready for import into LOVD 3.0, an open variant database platform. After merging the different annotation routes and removing duplicates, the project resolved about 2,240 distinct variants across 115 genes, linked to roughly 2,700 patients.

The practical value is simple: information that had become difficult to use is now much closer to being searchable, shareable, and reusable. The project also leaves behind a reusable Python workflow for other legacy databases with similar reference-sequence problems. Not every record was rescued automatically, and some will still need manual checking, but the main structure is now in place.

Master's thesis, Lund University (LTH), 2026. Supervisor: Prof. Mauno Vihinen. Examiner: Ghjuvan Micaelu Grimaud
Original title: From Legacy Databases to Modern Standards - A Bioinformatics Pipeline for Migrating Immunodeficiency Variant Data to LOVD 3.0 with GRCh38 Standardisation.
Full project: github.com/knn1996/Migrating-BMC-IDBases-to-LOVD