Knowledge multiplies when shared — when calling things by their right name: improving the validation and exchange of genetic data in research and diagnostics

Fokkema, I.F.A.C.

CHAPTER 1

General introduction

Ivo F.A.C. Fokkema

1.1 Introduction

Genetic disorders form significant health challenges, and the relative impact on the global disease landscape has increased because the prevention and treatment of communicable diseases and malnutrition have improved. Genetic conditions affect an estimated 5.3% of newborns by the age of 25 and contribute to both prenatal and postnatal mortality.[1] They can also manifest later in adulthood, impacting lifespan and quality of life, and increasing the likelihood of transferring the condition or causative DNA variants to offspring. Moreover, genetic predisposition plays a role in many common diseases of later life, influencing up to 40% of the global population.[2] In some countries, the prevalence of certain genetic disorders is increased due to factors such as consanguineous marriages and historical exposure to other diseases. A common example is malaria, which affects healthy individuals more than healthy carriers of hemoglobinopathies, driving selection towards a higher incidence of these inherited blood disorders.[3]

Thanks to advances in genomic analysis technologies, we can now determine the sequence of a human individual within a day. Compared to the standard reference sequence, an average human genome of roughly 6.2 billion base pairs (~3.1 billion inherited from each parent) contains about 4 million variants.[4] When genome sequencing is used to diagnose a genetic disorder, and a likely genetic cause is identified, only one or two of these variants are usually associated with the presented condition. Where conventional single-gene screenings often used to identify only one or two variants in a selected candidate gene, and classifying the results could be handled by an expert on the gene screened, specialists are now expected to classify a large number of variants for genes genome-wide. Even for a specific disease, the number of genes included in the disorder's gene panel may go up to several hundred. A broader analysis for, e.g., developmental disorders, requires the analysis of 1700 genes.[5] The most helpful information to classify a variant as disease-causing or not is having direct access to all available data for the variant and phenotype, including previous observations and the opinion of an expert in the field. To this end, expert-curated gene variant databases are an essential resource for any diagnostic or research lab, saving precious time and resources to get to a valid conclusion.

This thesis describes our work on improving data sharing in genetics, focused on gene variant databases. Chapters 2–3 introduce the most important topics underlining this thesis' subject, genetic variation and databases, respectively. Chapters 4–9 each describe an important building block involved in data sharing of genetic variants. Together, this thesis describes over 20 years of work in the field of gene variant databases.

1.2 Thesis outline

Before we can discuss the concept of data sharing in genetics, we must first understand the main subject that will be discussed throughout this thesis — what is genetic variation? How do variants occur? Is every change in our DNA relevant? What does this data look like? **Chapter 2** answers these questions, introduces different standards relevant to describing the data, and highlights the main challenges that diagnostics face when analyzing patient data.

While traditionally, information about patients with genetic variants of interest was only shared in the scientific literature, ever since the internet became commonplace, researchers were able to publish their data online either through simple web pages, like the Leiden Muscular Dystrophy pages started as early as 1996, or more complex systems. As such, many of these systems exist today, often specialized towards a specific type of data (e.g., disease data versus genetic variants) or a specific depth (e.g., summary data collected for all genes versus detailed case-level data on a specific gene or disease). **Chapter 3** lists the most important databases available today and explains the benefits that each system has over others.

This thesis focuses mostly on the work around the gene variant databases. These databases store detailed case-level data and usually focus on a specific gene or disease, curated by an expert in that field. While gene variant databases are, obviously, a much better way to share data broadly than having data hidden in journals that require expensive subscriptions, many different such systems initially existed, lacking cross-platform queries, causing enormous fragmentation of the data. **Chapter 4** discusses the third major release of the Leiden Open Variation Database (LOVD) software, the first free and open-source software package that allowed researchers and clinical labs to build their own gene variant databases. The LOVD software now powers the largest gene variant database network, allowing queries across different LOVD instances and making it much easier to find detailed information on genetic and genomic variants.

To efficiently check whether a certain data source contains information matching a patient currently being diagnosed in the clinic, it is paramount that standards are used to name and describe data. While this is true for information on disease (consider "DMD" versus "Duchenne Muscular Dystrophy" versus "Muscular Dystrophy, Duchenne type" versus "OMIM:310200", etc.), genetic variant data seems most affected by ambiguity as variants can be described on the genomic DNA level, the gene transcript DNA level, the RNA level, or the protein level. Not being able to match a patient to the available data can lead to incorrect diagnoses, which in turn can lead to unfounded medical decisions. It is, therefore, vital for genetic variant data that it is described using one single standard. Over the last two decades, the Human Genome Variation Society (HGVS) Nomenclature has become the global standard for

describing variants in DNA, RNA, and protein sequences. **Chapter 5** describes recent updates to the HGVS Nomenclature and the various improvements implemented by the HGVS Variant Nomenclature Committee (HVNC), the committee maintaining the standard, to the HGVS Nomenclature website and the governance of the standard.

Although the HGVS Nomenclature is considered the universal standard in describing and reporting variants in DNA, RNA, and protein sequences, its application is not without problems. **Chapter 6** shows that the majority of variant descriptions submitted by authors to journals contain errors, and correction is needed to allow proper identification of what variants were actually identified. To catch problems early in the submission process, we discuss integrating VariantValidator, software to validate and annotate DNA variants, into journal publishing and database submission systems, and we introduce the collaboration between the LOVD and VariantValidator teams aimed at improving the recognition and correction of as many invalid variant descriptions as possible.

This collaboration is further described in **Chapter 7**. Over the years, during curation, the LOVD team collected a large amount of incorrect DNA variant descriptions and developed the LOVD HGVS syntax checker tool that attempts to recognize what the submitter most likely meant, providing informative feedback and automatically correcting the description when possible. This way, the tool fills a gap in the validation of DNA variant descriptions, where existing tools focus on sequence-level validation only. As our tool validates on the syntax level, it supports more variant types than any other tool available.

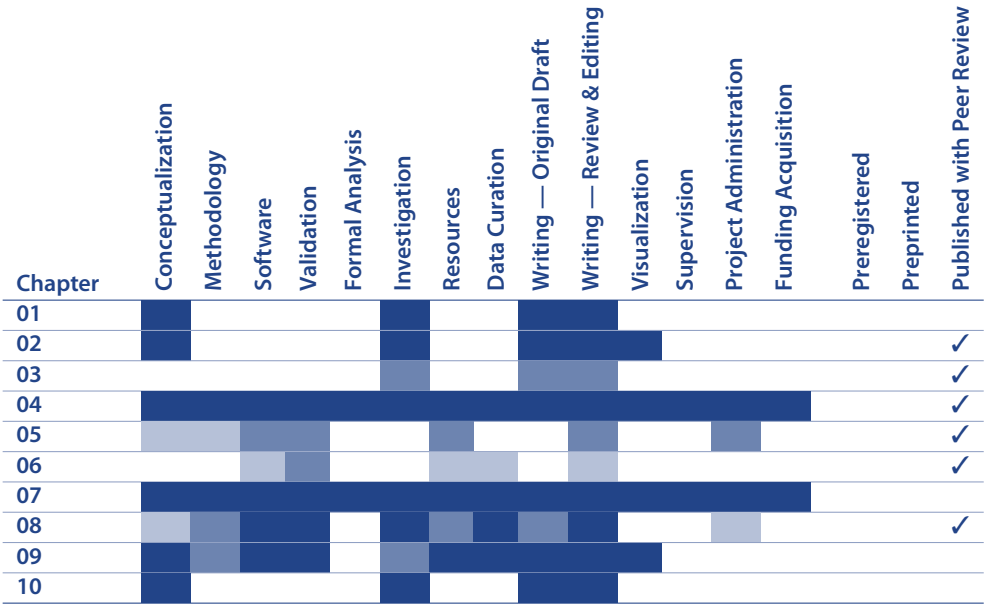
Public data sharing and correcting DNA variant descriptions also play a prominent role in **Chapter 8**. In this chapter, we introduce the DNA variant-sharing procedure set up for the Dutch genome diagnostic labs. These labs produce a large amount of variant data each year, and we describe how we unlocked this data by processing it centrally, correcting all variant descriptions, and then sharing it back nationally in a dedicated database and internationally using the LOVD platform. We discuss the challenges we encountered and the solutions we implemented. We hope to inspire other labs and countries to do the same, showcasing the improved data quality and solving additional clinical cases thanks to additional exposure of the variants identified.

While the data-sharing process outlined in Chapter 8 involved only summary-level data, gene variant databases can contain very complex datasets spread over multiple database tables. Such datasets can not be expressed in simple tab-delimited text files but require structured data formats. **Chapter 9** showcases recent improvements to a data format capable of storing entire gene variant databases, with all their complexity. This incredibly powerful and flexible format, allowing a patient-centered as well as a variant-centered implementation, is built into

LOVD to allow both downloads and fully automated submission of case-level datasets.

Finally, **Chapter 10** summarizes the results of this thesis and discusses the overall relevance and our vision for the future. Furthermore, we look back at what we have learned in the past two decades working on gene variant databases.

1.3 CRediT table for this thesis



1.4 References

[1] I.C. Verma and R.D. Puri. *Global burden of genetic disease and the role of genetic screening*. Seminars in Fetal and Neonatal Medicine 2015; 20 (5) 354–363.

[2] Victor Boulyjenkov. *WHO Human Genetics Programme: a brief overview*. Community Genetics 1998; 1 (2) 57–60.

[3] Philip W. Hedrick. *Resistance to malaria in humans: the impact of strong, recent selection*. Malaria Journal 2012; 11.

[4] Samuel Levy, Granger Sutton, Pauline C. Ng, Lars Feuk, Aaron L. Halpern, Brian P. Walenz, Nelson Axelrod, Jiaqi Huang, Ewen F. Kirkness, Gennady Denisov, Yuan Lin, Jeffrey R. MacDonald, Andy Wing Chun Pang, Mary Shago, Timothy B. Stockwell, Alexia Tsiamouri, Vineet Bafna, Vikas Bansal, Saul A. Kravitz, Dana A. Busam, Karen Y. Beeson, Tina C. McIntosh, Karin A. Remington, Josep F. Abril, John Gill, et al. *The diploid genome sequence of an individual human*. PLoS biology 2007; 5 (10) 2113–2144.

[5] Robin N. Beaumont and Caroline F. Wright. *Estimating diagnostic noise in panel-based genomic analysis*. Genetics in Medicine 2022; 24 (10) 2042–2050.