Disambiguation And Normalization Of Genomic Variants

Exposé for the Bachelor thesis

Alexander Harrisson — **alexander.harrisson@hu.berlin.de**

March 12, 2021

Supervisor: Ulf Leser

1 Introduction

In recent years sequencing the human genome has become relatively easy and affordable when compared to just two decades ago. When the first draft of the human genome has been sequenced in 1999 over a period of more than a year it cost approximately $300 million [1] while it is less than $100 today and only takes a couple of hours.

Gene-based therapy has become a widely used tool in medical fields such as oncology. It is based on the fact that a patient's genomic variants (alterations on genomic code) can influence tumor development [2] and druggability. Every human being or living creature has a unique set of these variants, lots of them can result in a change of the phenotype and some of them even result in non-benign cancerous tumors. Some patients may share similar or even the same key symptoms and yet they show different reactions to certain treatments or drugs [4]. In precision medicine a patients genetic information is an important part of clinical decision making. People get treated with a drug that is effective for their specific allelic variants.

Biomedical information, including information on gene variants is organised in numerous knowledge bases (KBs), such as ClinVar [15] or CiViC [14]. This helps researches to get quick and easy access to information without reading through all of the literature. These databases are mostly manually constructed and maintained which leads to incomplete and fragmented information in all of these KBs.
Before new information is added to one of these knowledge bases it has to be found in the vast number of publications. A large amount of biomedical publications are indexed on PubMed every day. Search queries on this data can be difficult because of the various notations for genomic variants in the research community [3]. Despite the existence of

standard nomenclatures for describing these variants, researchers and authors often use their own abbreviations and names. Although the use of these standardized notations has increased in recent years, even today studies report that more than 76% [3] of all genomic variants are described in non-normalized form when published.

2 Describing Genomic Variants in Texts

Before published biomedical information will get indexed at a KB, two problems must be solved. While Named-entity recognition (NER) is the process of identifying the mention of entities in a text using regular expressions or machine learning techniques, Named-entity normalization (NEN) is the process of mapping these found entities to a certain concept or ID so that same entities with different notations are recognised to be the same concept [3]. Normalization removes the ambiguity and allows people to find information on the correct variant in several different databases. One established format to reference or describe these variants was first introduced by Den Dunnen and the HGVS (Human Genome Variation Society) in 2003 [12].

Variants can be described by using different reference sequences, for example genomic, protein or coding DNA sequences. They can also be described on different assemblies and in various notations. One way to do this is by using the HGVS nomenclature.
The *Human Genome Variation Society* (HGVS) nomenclature is a recommended standard for the description of sequence variants especially used in clinical diagnostics [5]. The general structure of this format is **"reference:description"**, for example **"NM_004006.2:c.4375C>T"**. In this case **NM_004006.2** is the reference sequence, based on a protein coding RNA (mRNA) and **c.4375C>T** is a description of a mutation in which C (cytosine) has been replaced with T (thymine) at position 4375 on coding DNA (**c**) [6].
The HGVS package [11] is publicly available for the python programming language and it is the basis for validation of sequence variation nomenclature tools such as Mutalyzer [9].

A variant can have multiple distinct descriptions, even in the HGVS standard. A substitution for example can also be refered to as a deletion followed by an insertion at a specific location. Variants can be described on different genomic reference sequences (e.g. **NC_000023.9:g.32290917C>T** or **NC_000023.10:g.32380996C>T**) or on different coding DNA reference sequences (e.g. **LRG_199t1:c.5234G>A** or **NM_004006.3:c.5234G>A**). This ambiguity and confusion can lead to errors when databases are queried for specific information [6]. Therefore lots of databases uses their own unique identifiers to refer to gene variants.
Sometimes variants are provided in pseudo-HGVS form, which means that only parts of the recommended variant description is given. For example a lot of times it can happen that the only thing the reader knows is that there is a mutation on a certain Gene (e.g. **MUT KRAS:**) without an actual description of it.
In other cases, authors often do not use the HGVS format at all, but describe the variant

using an identifier of other databases. Some of these IDs are from the dbSNP, OMIM or ClinVar databases [6].

3 Research Questions

A Variant Information System (VIS) collects and aggregates data from multiple different public sources and makes their joint information available to its users. [2]. I-VIS (Integrated-VIS) is one implementation of a Variant Information System and will be the basis for our research. It is part of the PREDICT project and is not yet published.
"The central aim of the PREDICT project is to develop a software system that enables clinicians to use the large body of data on the relationships between genetic/epigenetic alterations and treatment options/success in cancer, to support (a) the rapid development of new, targeted studies whose design essentially is based on genomic features, and to (b) enable a maximally informed and structured clinical decision process. " [24]. I-VIS collects and provides various data points, including variant information through its API from 12 public knowledge bases.

Table 1: I-VIS integrated KBs [24]

Biomarkers DB [13]	https://www.cancergenomeinterpreter.org/biomarkers
CIViC [14]	https://civicdb.org
ClinVar [15]	https://www.ncbi.nlm.nih.gov/clinvar
Cosmic [16]	https://cancer.sanger.ac.uk/cosmic
dbnsfp DB [17]	https://sites.google.com/site/jpopgen/dbNSFP
docm[18]	http://www.docm.info
exac [19]	https://gnomad.broadinstitute.org/
oncoKB [19]	https://oncokb.org
pmkb [20]	https://pmkb.weill.cornell.edu
target[21]	https://software.broadinstitute.org/cancer/cga/target
ctg [22]	https://aact.ctti-clinicaltrials.org
1000gp [23]	https://www.internationalgenome.org

The first research question of this student project is the following:

1. How to normalize these different notations in the I-VIS database to a standardized identifier such as a Reference SNP cluster ID (RSID)?

We will start by analyzing the entries in the dataset and quantify how many different notation types, such as WNM, RSID or HGVS like formats are being used. Based on the results of the first step we will use several tools, such as SETH [7] or tmVar 2.0 [8] that help normalize these notations. There will be many entries where this will not be possible because there is not enough information stored on the variant (e.g. **MUT**

KRAS:). In this case the second part of this project would be:

2. Trying to quantify and categorize the results and the original data.

We will be analyzing the different types of notations and error sources and quantify how many entries can not be normalized because of missing or incorrect information. We will also analyze the overlap between these KBs and determine how many variants appear in multiple KBs and how many are unique to a single source.

4 Methodology

The *first* step would be to parse and analyze the entire I-VIS database with its millions of entries and check how many different notation types are being used. This can be done all at once or in multiple consecutive runs with smaller batches of data at a time. Based on this information we can statistically determine which tools could later be used for normalization by figuring out the most frequently used formats and categorize them. Not every tool works well with all notations, so we will use a tool that is particularly good at normalizing this type of notation.
Secondly we would systematically check for mistakes, for example incorrect descriptions or missing information in the dataset.
Thirdly we will try to convert and validate these variants in the HGVS nomenclature using the HGVS python package which makes it easier to find a corresponding standardized identifier such as an RSID. This could be accomplished by using tools such as SETH [7] or tmVar 2.0 [8]. We would run these NER/NEN tools on our raw data and use their normalization techniques to get a dbSNP identifier in return if one exists. SETH's NEN component can be used in Java or directly from the command line.
The last step would be to analyze and quantify how many variants overlap and how many could not be normalized because of missing or incorrect data. We will compare the results of the used normalization tools and show which KBs contain how many unique variants and how many overlap.

The HGVS package is publicly available for Python and can easily be installed via pip or directly from source. It provides many features such as "[...] parsing, formatting, validating, and normalizing variants on genome, transcript, and protein sequences, projecting variants between aligned sequences, including those with gapped alignments [...]" [11]. For the actual normalization to a unique identifier we will be needing access to the dbSNP database and an API which can find and retrieve NCBI data. This is necessary because some variants might need additional information, that is not provided by I-VIS before they can be normalized.
One candidate for this would be *EntrezPy* [10]. It is a public API to the NCBI Entrez system and allows access to the Entrez databases including PubMed and is also available for Python ($\geq$3.6) which means that it can easily be combined with HGVS.

References

[1] National Human Genome Research Institute, "The Cost of Sequencing a Human Genome", https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost, 2020

[2] J Starlinger, S Pallarz et al, (2018). "Variant information systems for precision oncology.", BMC Medical Informatics and Decision Making. 18. 10.1186/s12911-018-0665-z.

[3] Lee K, Wei CH, Lu Z. "Recent advances of automated methods for searching and extracting genomic variant information from biomedical literature", Brief Bioinform. 2020 Aug 7:bbaa142. doi: 10.1093/bib/bbaa142. Epub ahead of print. PMID: 32770181.

[4] Aronson SJ, Rehm HL. "Building the foundation for genomics in precision medicine.", Nature. 2015 Oct 15;526(7573):336-42. doi: 10.1038/nature15816. PMID: 26469044; PMCID: PMC5669797.

[5] JTD Dunnen, "Sequence Variant nomenclature", https://varnomen.hgvs.org/bg-material/basics/, 2020

[6] JTD Dunnen, "Sequence Variant nomenclature", https://varnomen.hgvs.org/bg-material/simple/, 2020

[7] Thomas, Philippe Hakenberg, Jörg et al "SETH detects and normalizes genetic variants in text", 2016, Bioinformatics. 32. 10.1093/bioinformatics/btw234.

[8] CH Wei, L Phan et al "tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine", 2017, Bioinformatics. 34. 10.1093/bioinformatics/btx541

[9] Thomas, P.E., Klinger, R., Furlong, L.I. et al. "Challenges in the association of human single nucleotide polymorphism mentions with unique database identifiers.", BMC Bioinformatics 12, S4 (2011). https://doi.org/10.1186/1471-2105-12-S4-S4

[10] Jan P Buchmann, Edward C Holmes "Entrezpy: a Python library to dynamically interact with the NCBI Entrez databases", Bioinformatics, Volume 35, Issue 21, 1 November 2019, Pages 4511–4514, https://doi.org/10.1093/bioinformatics/btz385

[11] Wang M, Callenberg KM, Dalgleish R, Fedtsov A et al "hgvs: A Python package for manipulating sequence variants using HGVS nomenclature", 2018 Update. Hum Mutat. 2018 Dec;39(12):1803-1813. doi: 10.1002/humu.23615. Epub 2018 Sep 5. PMID: 30129167; PMCID: PMC6282708.

[12] den Dunnen JT, Paalman MH. "Standardizing mutation nomenclature: why bother?", Hum Mutat. 2003 Sep;22(3):181-2. doi: 10.1002/humu.10262. PMID: 12938082.

[13] Tamborero, David ,Rubio-Perez, Carlota ,Deu-Pons, Jordi ,Schroeder et al (2018) "Cancer Genome Interpreter annotates the biological and clinical relevance of tumor alterations.", Genome medicine. 10. 25. 10.1186/s13073-018-0531-8.

[14] Griffith M, Spies NC, Krysiak K, McMichael JF, Coffman AC et al "CIViC is a community knowledgebase for expert crowdsourcing the clinical interpretation of variants in cancer." , Nat Genet. 2017 Jan 31;49(2):170-174. doi: 10.1038/ng.3774. PMID: 28138153; PMCID: PMC5367263.

[15] M J Landrum, JM Lee, M Benson et al "ClinVar: improving access to variant interpretations and supporting evidence" Nucleic Acids Research, Volume 46, Issue D1, 4 January 2018, Pages D1062–D1067, https://doi.org/10.1093/nar/gkx1153

[16] Bamford, S., Dawson, E., Forbes, S. et al "The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website" Br J Cancer 91, 355–358 (2004). https://doi.org/10.1038/sj.bjc.6601894

[17] X Liu, C Wu, C Li, E Boerwinkle "dbNSFP v3.0: A One-Stop Database of Functional Predictions and Annotations for Human Nonsynonymous and Splice-Site SNVs." , Hum Mutat. 2016 Mar;37(3):235-41. doi: 10.1002/humu.22932. Epub 2016 Jan 5. PMID: 26555599; PMCID: PMC4752381.

[18] Ainscough, B., Griffith, M., Coffman, A. et al. "DoCM: a database of curated mutations in cancer" Nat Methods 13, 806–807 (2016). https://doi.org/10.1038/nmeth.4000

[19] Chakravarty, Debyani, Gao, Jianjiong et al "OncoKB: A Precision Oncology Knowledge Base" JCO precision oncology. 2017. 10.1200/PO.17.00011..

[20] Weill Cornell Medicine `https://pmkb.weill.cornell.edu/`, 2021

[21] Cancer Genome Analysis `https://software.broadinstitute.org/cancer/cga/target`, 2021

[22] Clinicial Trials Transformation Initiative `https://www.ctti-clinicaltrials.org/`, 2021

[23] IGSR: The International Genome Sample Resource `https://www.internationalgenome.org/`, 2021

[24] Charite Universitätsmedizin Berlin Charite, Berlin Institute of Health BIH, Humboldt-Universität zu Berlin, "comPREhensive Data Integration for Cancer Treatment", `https://predict.informatik.hu-berlin.de/`, 2021