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.

Samenvatting

Genetische aandoeningen en genetische predispositie voor ziekten vormen wereldwijd een aanzienlijke belasting voor de gezondheidszorg. Toch wordt, meer dan twintig jaar na de voltooiing van het Human Genome Project, slechts een fractie van de wereldwijd gegenereerde genomische data actief gedeeld. Wanneer gegevens wel worden gedeeld, gebeurt dit vaak via inefficiënte methoden of met foutieve variantbeschrijvingen, die in het ergste geval de gegevens onbruikbaar maken. Het werk dat in dit proefschrift wordt beschreven, richt zich op het verbeteren van datadeling in de genetica, in het bijzonder via genvariantendatabases en door te zorgen dat gegevens worden weergegeven met standaarden die unieke en eenduidige variantbeschrijvingen mogelijk maken.

De belangrijkste onderwerpen die ten grondslag liggen aan dit proefschrift worden geïntroduceerd in de eerste drie hoofdstukken. **Hoofdstuk 1** schetst de strekking van het proefschrift en geeft een overzicht van alle hoofdstukken, waarbij de onderlinge verbanden worden toegelicht. **Hoofdstuk 2** introduceert genetische variatie, de mogelijke gevolgen van varianten en de standaarden die essentieel zijn voor het beschrijven, interpreteren en rapporteren van genetische varianten en fenotypen. **Hoofdstuk 3** bespreekt algemene en "gerichte" databases en licht hun belangrijkste doelstellingen en voordelen toe.

Genvariantendatabases vormen de ruggengraat van DNA-gebaseerde diagnostiek. Zij zijn de meest betrouwbare bron van informatie over varianten en hun gevolgen, aangezien zij doorgaans voldoen aan strikte datastandaarden en richtlijnen. Deze databases slaan gedetailleerde casusdata op en richten zich meestal op een specifiek gen of een bepaalde ziekte, beheerd door experts in het veld. Om onderzoekers en klinische laboratoria in staat te stellen hun eigen genvariantendatabases op te bouwen, hebben wij de gratis, open-source Leiden Open Variation Database (LOVD) software ontwikkeld, die al snel de meest gebruikte tool voor dit doel werd. Na twee succesvolle grote releases hebben wij de software volledig herontworpen en LOVD3 uitgebracht (beschreven in **Hoofdstuk 4**). LOVD3 stelde grotere en meer diverse groepen curatoren in staat samen te werken binnen één instantie. Tegelijkertijd verschoof de focus in het veld, door nieuwe sequencingtechnieken, van gen-gebaseerde naar genoom-gebaseerde variantbeschrijvingen, wat een grondige herziening van het datamodel en van alle views en formulieren vereiste. LOVD3 is genoomgecentreerd en ondersteunt zowel samenvattende variantdata als gedetailleerde casusdata, inclusief informatie over individuen, fenotypen, screenings en varianten. Deze flexibiliteit, gecombineerd met Application Programming Interfaces (API's) die directe interactie met de database mogelijk maken, heeft geleid tot het grootste netwerk van genvariantendatabases ter wereld. Dit netwerk ondersteunt zoekopdrachten over verschillende LOVD-instanties en maakt het voor gebruikers mogelijk snel te identificeren waar relevante informatie over een

variant is opgeslagen.

Het combineren van informatie uit verschillende systemen voor de diagnose van een patiënt in de klinische diagnostiek is alleen mogelijk wanneer deze systemen dezelfde standaarden gebruiken om DNA-varianten en fenotypische gegevens te beschrijven. In de afgelopen twee decennia is de Human Genome Variation Society (HGVS) Nomenclature uitgegroeid tot de wereldwijde standaard voor het beschrijven van varianten in DNA-, RNA- en eiwitsequenties. Toch worden ook nu nog de meeste varianten die in de literatuur worden gerapporteerd onjuist of onvolledig beschreven. Hoewel dit de rol van databases als meest betrouwbare bron voor variantinformatie versterkt, belemmert het de integratie van literatuurdata in databases en vertraagt het uiteindelijk de diagnostiek. In de afgelopen jaren heeft de HGVS Variant Nomenclature Committee (HVNC) verschillende updates uitgebracht (beschreven in **Hoofdstuk 5**) om de correcte toepassing van de nomenclatuur in de literatuur en in klinische rapportages te verbeteren. De website werd herontworpen om de leesbaarheid van de documentatie te verbeteren, fouten en ambiguïteiten werden verwijderd, alle wijzigingen werden duidelijk gedocumenteerd, en er werd een aparte pagina toegevoegd met software die variantbeschrijvingen kan valideren en corrigeren. Ook werd de betrokkenheid van de gemeenschap versterkt door extra kanalen te bieden voor gebruikersvragen en door methoden te faciliteren waarmee wijzigingen in de nomenclatuur kunnen worden voorgesteld.

Databases zijn afhankelijk van gespecialiseerde software om ingediende variantbeschrijvingen te valideren. Een directe manier om beschrijvingen in de literatuur te verbeteren is om deze validatiesoftware te integreren in het manuscript-indieningsproces. Dit vereist samenwerking tussen uitgevers en softwareontwikkelaars, waardoor manuscripten grondig worden gecontroleerd en zo nodig gecorrigeerd voordat zij worden gepubliceerd. Bovendien zou dergelijke software varianten die in een artikel worden beschreven zelfs automatisch kunnen indienen bij relevante databases. Het LOVD-team werkt samen met de ontwikkelaars van VariantValidator en heeft hun software voorbereid op integratie in uitgeversworkflows (zie **Hoofdstuk 6**).

Om de herkenning van variantbeschrijvingen in de literatuur echter volledig te automatiseren, zijn aanvullende tools nodig. Bestaande validatiesoftware gaat ervan uit dat beschrijvingen grotendeels al conform de HGVS Nomenclature zijn; veel publicaties gebruiken echter verouderde namen of onvolledige beschrijvingen, die door de huidige tools niet goed worden herkend of gecorrigeerd. Bovendien worden beschrijvingen van zeldzame varianten of beschrijvingen met een zekere mate van onzekerheid zelden ondersteund. Om dit aan te pakken hebben wij de LOVD HGVS syntax checker ontwikkeld (zie **Hoofdstuk 7**), ontworpen om zowel geldige als veelvoorkomende ongeldige variantbeschrijvingen te herkennen. Voor ongeldige invoer genereert de software automatisch correctievoorstellen en kent een

zekerheidswaarde toe aan elke suggestie. Door zich te richten op de syntaxis bereikt de LOVD HGVS syntax checker een ongeëvenaarde dekking van de HGVS Nomenclature. Omdat volledige sequentieniveau-validatie nog steeds noodzakelijk is, zijn LOVD en VariantValidator geïntegreerd in één platform, dat een uitgebreide oplossing biedt. Momenteel voeren wij actieve gesprekken met uitgevers om deze tools in hun indieningspijplijnen te integreren, wat de kwaliteit van variantbeschrijvingen in de wetenschappelijke literatuur ingrijpend zou verbeteren.

Hoewel onderzoeksgegevens vaak in wetenschappelijke tijdschriften worden gepubliceerd, is het merendeel van de genetische data afkomstig uit diagnostische laboratoria. Alleen al in Nederland worden jaarlijks duizenden individuen gescreend op genetische aandoeningen. Hoewel diagnostische laboratoria zelden gedetailleerde patiëntgegevens delen, delen Nederlandse laboratoria routinematig de varianten die zij identificeren en de daarbij behorende classificaties (pathogeen of niet), wat waardevolle informatie oplevert. Deze gegevens worden verwerkt via een gecentraliseerde pijplijn (beschreven in **Hoofdstuk 8**), waardoor laboratoria kunnen samenwerken en hun bevindingen met elkaar kunnen delen. Daarnaast worden de gegevens verwerkt en vrijgegeven via LOVD, waardoor zij wereldwijd beschikbaar zijn. Omdat de gegevens volledig worden gevalideerd voordat zij worden gepubliceerd, is de kwaliteit aanzienlijk verbeterd en worden zowel interne conflicten als discrepanties tussen laboratoria automatisch gedetecteerd en gerapporteerd.

Het delen van meer dan alleen varianten en classificaties vereist echter geavanceerdere benaderingen. Databases zoals LOVD kunnen complexe casusdata opslaan, waaronder fenotypen, laboratoriumresultaten, meerdere varianten en gedetailleerde informatie over de effecten op één of meer genen. Dergelijke datasets kunnen niet worden weergegeven in eenvoudige tab-gescheiden tekstbestanden, maar vereisen gestructureerde formaten. Om dit te ondersteunen hebben wij het VarioML-formaat, oorspronkelijk gepubliceerd in 2012, volledig geüpdatet en gemoderniseerd naar een native JSON-implementatie (zie **Hoofdstuk 9**). Het nieuwe formaat is eenvoudiger te integreren in moderne systemen, ondersteunt zowel patiënt- als variantgerichte weergaven en is geïmplementeerd in LOVD3 om downloads en geautomatiseerde indieningen mogelijk te maken. Door een ontologie-gedreven benadering toe te passen, brengt VarioML LOVD bovendien dichter bij naleving van de FAIR-principes (Findable, Accessible, Interoperable, Reusable).

Ondanks hun transformerende rol in onderzoek en diagnostiek, worden DNA-variantendatabases zoals LOVD geconfronteerd met aanhoudende uitdagingen (zie **Hoofdstuk 10**). Ten eerste blijft financiering schaars en zijn al meerdere grote databases verloren gegaan door financiële- of personeelsproblemen. Het onderhoud wordt verder bemoeilijkt door toenemende interesse van grote commerciële partijen, die buitensporig veel verkeer

kunnen genereren en servers tijdelijk onbruikbaar kunnen maken. Ten tweede wordt slechts een fractie van de beschikbare data ingediend in databases, en zijn er meer inspanningen nodig om extra gegevens te ontsluiten en te delen. Ten derde is de naleving van standaarden in bronnen zoals de literatuur en klinische rapportages beperkt, wat betekent dat grote hoeveelheden niet-gestandaardiseerde beschrijvingen herkend en gecorrigeerd moeten worden, bij voorkeur op geautomatiseerde wijze. In de nabije toekomst zullen onze inspanningen om deze problemen op te lossen samenkomen in samenwerkingen met uitgevers om datavalidatie te automatiseren in de pre-publicatiefase, gevolgd door datasubmissie na acceptatie van het manuscript. Tegelijkertijd zullen wij nieuwe methoden blijven ontwikkelen voor geautomatiseerde interactie met databases, waaronder gefedereerde zoekopdrachten over FAIR-conforme systemen.

Concluderend heeft het in dit proefschrift gepresenteerde werk bijgedragen aan de vooruitgang van de vakgebieden van genvariantendatabases, datavalidatie en datadeling, wereldwijd het genetisch onderzoek en de klinische diagnostiek ondersteunend.