Institution: University of Leicester

Unit of Assessment: UoA 5

Title of case study: VariantValidator: a novel, open-access and user-friendly software to improve genetic diagnosis

Period when the underpinning research was undertaken: 2012-2020

Details of staff conducting the underpinning research from the submitting unit:

Name(s):	Role(s) (e.g. job title):	Period(s) employed by submitting HEI:
(1) Prof Raymond Dalgleish	(1) Professor of Human Genetics	(1) 1984 - present
(2) Dr Peter Freeman	(2) Post-Doctoral Research Associate	(2) 2007- 2019
(3) Prof Anthony Brookes	(3) Professor of Bioinformatics and Genomics	(3) 2004 - present
(4) Mr Liam Gretton	(4) Systems Specialist	(4) 2009 - present

Period when the claimed impact occurred: 2013-2020

Is this case study continued from a case study submitted in 2014? N

1. Summary of the impact

Your DNA sequence is unique. It is over 3 billion letters long and found in almost every cell in your body. Each individual's DNA sequence has millions of differences (variants); most are harmless but some cause disease. For healthcare professionals worldwide it is critical that variants be reported accurately to ensure correct decision-making; something that was problematic with existing software. VariantValidator software developed at the University of Leicester (UoL) has solved this issue and is established by key partners as the gold-standard tool for variant description validation. It is used routinely by thousands of professionals globally, including; NHS training programmes, leading biomedical journals and databases and healthcare and research organisations

2. Underpinning research

Based on a long-standing track record of research, UoL has been developing accessible and freely available software solutions to allow healthcare professionals to obtain accurate information on genetic variants in individual's genome.

The 100,000 Genomes Project and the current NHS Genomic Medicine Service place the UK at the forefront of the global genomic-medicine revolution. Genomics England Ltd (GEL), who run these projects, are developing computational models to represent DNA sequence variations in genes that cause a range of health conditions, from cancers to rare developmental disorders. GEL's data-analysis models aim to simplify the process that allows healthcare professionals to determine whether a particular sequence variation is responsible for a human disease, and thereby to establish accurate diagnoses and appropriate treatment options. However, progress has been hampered by inconsistent application of genetic data guidelines and lack of data sharing, combined with 'black-boxed' analytical software resulting in variable interpretation between laboratories, which can significantly affect patient outcomes **[R2]**.

Sequence variants presented in scientific papers, clinical reports and databases must be reported using the HGVS Nomenclature, co-authored by Professor Dalgleish **[R3]**. However, evidence shows that adherence to the nomenclature standard is variable, with examples of incorrect and sometimes misleading descriptions being presented **[R1]**. At best, such errors cause confusion, at worst they may lead to patients receiving inappropriate medical treatments or advice, or even missing out on potentially lifesaving treatments.

Errors arise because some aspects of the HGVS nomenclature are complex, making it difficult to comprehend and use for both experts and no-experts resulting in inaccuracies in communication of variant data. Consequently, high-quality software tools are required to ensure that variant descriptions are valid, complete, and consistent with the nomenclature guidelines. In 2015, there was only one purpose-built tool for validating HGVS descriptions, this tool was limited in three key areas:

a) It was not fully standards-compliant
b) It was unable to validate variation between corresponding reference genomic and mRNA sequences if slight sequence mismatches existed between the two
c) It was unable to detect, and automatically correct, many common nomenclature errors **[R1]**

VariantValidator is the result of extensive research and long-standing expertise by Dalgleish's team at Leicester **[R4, R5, R6]** and was developed to address the recognised deficiencies of existing tools **[R1]**. VariantValidator provides rigorous checking of sequence variant descriptions with respect to syntax and data inconsistencies. This was achieved through a methodical development process tailored to user requirements whilst ensuring strict adherence to standards **[R2]**. VariantValidator is widely regarded as the most functional and accurate sequence variant description tool in the world and remains the only platform capable of accurately validating sequence variants in a fully standards-compliant fashion **[R2]**.

Since release in 2016, VariantValidator has been adopted widely and, at the request of field-leading partners, is undergoing continuous development including:

- Creation of VariantFormatter to enable Genomics England Ltd to validate their databases against HGVS nomenclature for the first time
- Development of a new reference sequence transcript archive allowing the 100,000 Genomes Project and COSMIC to validate and warrant their data
- Full integration with flagship software packages from SOPHiA Genetics, GA4GH and The Jackson Laboratory
- Developments enabling integration into diagnostic pipelines in NHS genomic Laboratory Hubs (https://www.england.nhs.uk/genomics/genomic-laboratory-hubs/) and diagnostic hospitals in Europe and the USA
- Collaborative development to integrate VariantValidator into the Leiden Open Variant Database (LOVD), the largest federated network of Open Source and Open Access variant databases globally

3. References to the research

R1. **Freeman PJ**, Hart RK, **Gretton LJ**, **Brookes AJ**, **Dalgleish R** (2018). VariantValidator: Accurate validation, mapping, and formatting of variation descriptions. Human Mutation 39:61-68. doi: 10.1002/humu.23348

R2. Wang M, Callenberg KM, **Dalgleish R**, Fedtsov A, Fox NK, **Freeman PJ**, Jacobs KB, Kaleta P, McMurry AJ, Prlić A, Rajaraman V, Hart RK (2018). hgvs: A Python package for manipulating sequence variants using HGVS nomenclature: 2018 Update. Human Mutation 39:1803-1813. doi: 10.1002/humu.23615

R3. den Dunnen JT, **Dalgleish, R,** Maglott DR, Hart RK, Greenblatt MS, McGowan-Jordan J, Roux A, Smith T, Antonarakis SE and Taschner PEM (2016) HGVS recommendations for the description of sequence variants: 2016 update. Human Mutation, 37: 564-569. doi: 10.1002/humu.22981

R4. Lancaster O, Beck T, Atlan D, Swertz M, Thangavelu D, Veal C, **Dalgleish R**, **Brookes AJ**. (2015) Cafe Variome: general-purpose software designed to make genotype-phenotype data easily and appropriately discoverable in restricted or open access contexts. Human Mutation, 36: 957-964. doi: 10.1002/humu.22841

R5. Gaspar P, Lopes P, Oliveira J, Santos R, **Dalgleish R** and Oliveira J L (2014) Variobox: Automatic detection and annotation of human genetic variants. Human Mutation, 35: 202-207. doi:10.1002/humu.22474

R6. Vihinen M, den Dunnen JT, **Dalgleish R**, Cotton RGH (2012) Guidelines for establishing locus specific databases. Human Mutation, 33: 298-305. doi: 10.1002/humu.21646

4. Details of the impact

VariantValidator is the cumulative result of decades of research, expertise, and development of human genome sequence variation data, by Dalgleish and his team, providing solutions to major obstacles in treatment-programme development using genetic information. The beneficiaries range from individual parties (patients and healthcare professionals) to multinational organisations and research programmes. VariantValidator has defined the future direction for interpretation of genetic sequencing by simultaneously ensuring future NHS Clinical Scientists are trained to far higher standards, and by significantly reducing potentially catastrophic errors in scientific research outputs being made, disseminated or repeated. VariantValidator is the global gold-standard HGVS variant nomenclature validation software.

Freely available, the open access VariantValidator software supports data handling and validation in field-leading organisations worldwide. Between launch in 2016 and 2020, VariantValidator had been used ~55,000 times by 40,451 unique users from 118 countries, with total user numbers increasing rapidly each year **[E1]**.
Example organisations who have embraced VariantValidator include:

- *GEL*: Responsible for delivering 100,000 Genomes Project **[E2a]**
- *ClinVar*: US National Institutes for Health database used globally for sequence variation interpretation **[E3]**
- *Catalogue Of Somatic Mutations In Cancer (COSMIC*): The world's largest and most comprehensive resource for exploring the impact of somatic mutations in cancer **[E2b]**
- Laboratoire de Génétique Moléculaire, CHU de Montpellier, France: have developed a variant annotation and interpretation software platform (MobiDetails) for which the variant validation engine of VariantValidator is a central embedded component **[E4]**
- *European Bioinformatics Institute*: International Non-Governmental Organisation providing bioinformatics services, training, and resources to the science community globally **[E5]**

In each case, and countless others, VariantValidator has consistently demonstrated its superiority over competitors "*similar systems known to us do not provide the modularity, code libraries or lack some key functionality*" GEL **[E2a]** and is now built into workflows facilitating high-impact biomedical research **[E6]** through substantially improved accuracy, efficiency, and usability. VariantValidator reduces genome coordinate mapping time from ~60 minutes per case to ~3 minutes **[E2c]**. In many leading organisations VariantValidator has replaced existing tools or is the preferred sequence variation curation tool including:

- Global Alliance for Genomics and Health (GA4GH)
- Centre for Medical Genetics, Brussels, Belgium **[E2d]**
- Leiden Open Variation Database (LOVD) **[E2e]**
- London South Genomics Laboratory Hub, North West Genomic Laboratory Hub, and other NHS laboratories
- The Manchester Academy for Healthcare Scientist Education (MAHSE) training programmes in medical genomics and genomics bioinformatics
- Global PGCert, and MSc courses delivered by the University of Manchester **[E8]**
- Center for Genomics Interpretation, Utah, USA **[E2f]**

The widespread response to the software has been immensely positive with many echoing Professor Robinson of the Jackson Laboratory, *"We have also used what is probably VariantValidator's main competitor, Mutalyzer, but have chosen to use VariantValidator for our work because of the substantially greater amount of information presented to the user, the ease of use and the flexibility of the API and paramaterized URLs and the much better graphical user interface, all of which improves the efficiency of our work"* **[E2c]**. At the Centre for Medical Genetics, Brussels, "*VariantValidator has been indispensable*" and they "*are now validating thousands of variants every week with VariantValidator*" **[E2d]**.

New HGVS Variant Nomenclature recommendations have recently been developed by PF and RD to accommodate the use of Ensembl transcript reference sequences for variant reporting **[E9]**. This enables NHS laboratories, for the first time, to utilise genome sequencing variant data,

generated by GEL. This novel development is being incorporated into a new reference sequence transcript archive that is, in direct collaboration with GEL and LOVD, being integrated into VariantValidator. The purpose of this development is to produce a tool enabling GEL processing to provide NHS laboratories with correctly formatted variant descriptions in the context of Ensembl reference sequences and enabling LOVD to accommodate Ensembl transcript data alongside existing RefSeq data. *"At present, we are unable to do this for the RefSeq datasets which is an unmet requirement for reporting in the clinical setting. This is a common requirement among clinicians to facilitate interpretation of genomic variation*." GEL **[E2a].** Immediate beneficiaries of this development include the GEL 100,000 Genomes Project **[E2a]**, COSMIC **[E3]** and a wider set of NHS Genomics Laboratory Hubs who will soon be able to validate and safely share their variant data outputs.

Collaboration between the University of Leicester and GEL has also resulted in the development of VariantFormatter software libraries enabling GEL to correctly present data using HGVS nomenclature. This system rapidly validates HGVS descriptions in the context of genomic reference sequences projecting them rapidly onto all relevant transcripts. In addition, the VariantValidator tool enables improved accuracy in variant descriptions for use in clinical reports and databases, lift-overs between Ensembl and pre-existing reference sequences, and interconversions between transcript and genomic-coordinate-based descriptions **[E2a]**. In addition, a novel version of the VariantFormatter tool has been created to produce accurately formatted data for the LOVD databases, and alongside VariantValidator, is being embedded directly into the LOVD analysis pipelines. This will have a huge impact on genomic medicine in Europe where LOVD is rapidly becoming a staple platform, replacing commercial solutions in diagnostic settings **[E10]**.

The proven superiority of VariantValidator has changed industry direction. Since 2017, Alamut Visual, the flagship software product from SOPHiA Genetics, has incorporated links to VariantValidator to allow users of their software to ensure accurate validation of sequence variant descriptions and have recently installed the VariantValidator code on their servers for internal use **[E8].** VariantValidator has also been adopted as the variant description validation program and interconversion tool of choice for the Global Alliance for Genomics and Health (GA4GH) 'PhenoPackets' and 'Solving the Unsolved Rare Diseases' (SOLVE-RD) projects **[E6]**, is a key component of the Jackson Laboratory's new HPO-based phenotype annotation software; HPOCaseAnnotator **[E11]**, and is used as standard practice by the Center for Genomic Interpretation in Utah to improve their clinical reporting **[E2f]**.

VariantValidator is having a profound impact on education in genomic bioinformatics, primarily in the UK, but also to a global audience. VariantValidator's demonstrable success convinced the Manchester Academy for Healthcare Scientist Training (MAHSE) to incorporate the software into NHS Scientist Training Programmes (STP) that contain genomics-focused modules, in 2019. VariantValidator is now taught to NHS STP MSc Students in Clinical Genetics, Clinical Bioinformatics, Health Informatics, Cancer Genomics, Medical Physics, and Genomic Counselling. Over 150 students are trained annually using the software leading to a wider deployment in NHS Clinical Genomics Centres and Genetics Diagnostics Laboratories. As well as learning to use VariantValidator, the course also enables students to produce additional resources for the system thereby enabling continuous resource growth and improvement alongside the internationally praised training scheme **[E8]**. This training not only provides VariantValidator with a direct pathway into clinical practice, and establishes a community of practice who not only use the software, but also steer future iterations of the software to meet the evolving needs of the genomic medicine community.

VariantValidator is embedded into healthcare professional CPD courses 'Fundamentals in Human Genetics and Genomics' funded by Health Education England and taught at the University of Manchester **[E12]**. VariantValidator is also embedded into the 'first of its kind' distance-learning Postgraduate Certificate in Clinical Bioinformatics **[E13]**. The current run of the course includes 37 students from around the world from a broad range of professions, including clinicians in non-UK diagnostics laboratories and the Qatari genome project. The success of the

2019 PGCert was disseminated to audiences in international educational symposia as evidenced in a University of Manchester blog article **[E8]**. VariantValidator is also embedded into the MSc in Genomic Medicine at the University of Manchester.

In 2020, following a successful pilot published by *Human Mutation* (Wiley) and *Genetics in Medicine* (Nature), several leading biomedical journals formally adopted VariantValidator as one of two approved validation tools to ensure correctness of variant descriptions in published manuscripts and to minimise errors long-term **[E14]**. This critical change in professional practice has already yielded benefits to patients, including successfully identifying and correcting multiple invalid and incorrect sequence descriptions previously used to define treatment for cystic fibrosis patients: "*These efforts will undoubtedly lead to tangible improvements in patient care*" **[E2g]**.

5. Sources to corroborate the impact

E1: Variant Validator usage statistics

E2: Supporting Testimonials

(a) Genomics England Limited (GEL), UK

(b) COSMIC, Sanger Institute, Hinxton, UK

(c) Jackson Laboratory, Farmington, CT, USA

(d) Centre for Medical Genetics, Brussels, Belgium

(e) LOVD, LUMC, Leiden, Netherlands

(f) Center for Genomic Interpretation, Sandy, UT, USA

(g) Johns Hopkins University, Baltimore, MD, USA

E3: 'Instructions for ClinVar Submission Spreadsheets', November 2020

E4: Credits on the MobiDetails web page

E5: Locus Reference Genomic (LRG) information page, EMBL-EBI, Hinxton, UK

E6: Selection of articles in high-impact biomedical journals confirming use of VariantValidator in underpinning sequence variation analyses

E7: 'International Recognition for New Digital Teaching Methods' University of Manchester, July 2020

E8: Release Notes for Alamut Visual Version 2.8.0 confirming VariantValidator integration

E9: HGVS Sequence Variant Nomenclature web page describing valid reference sequences: https://varnomen.hgvs.org/bg-material/refseq/

E10: Fokkema, Ivo FAC, et al. "Dutch genome diagnostic laboratories accelerated and improved variant interpretation and increased accuracy by sharing data." *Human mutation* 40.12 (2019): 2230-2238.

E11: Carmody, Leigh C., et al. "Significantly different clinical phenotypes associated with mutations in synthesis and transamidase+ remodeling glycosylphosphatidylinositol (GPI)-anchor biosynthesis genes." *Orphanet journal of rare diseases* 15.1 (2020): 40.

E12: NHS Health Education England Genomics Education Programme web page describing the taught course "Fundamentals in Human Genetics and Genomics": https://www.genomicseducation.hee.nhs.uk/education/taught-courses/fundamentals-in-human-genetics-and-genomics/

E13: University of Manchester web page describing the online course PGCert Clinical Bioinformatics: https://www.manchester.ac.uk/study/masters/courses/list/12099/pgcert-clinical-bioinformatics/course-details/

E14: Higgins, Jan, et al. "Verifying Nomenclature of DNA Variants in Submitted Manuscripts: Guidance for Journals." *Human Mutation* (2020).