Dante Automates Clinical Genetic Interpretation Reports in Genome Sequencing

Dylan Lawless*[1]

[1]Department of Intensive Care and Neonatology, University Children's Hospital Zürich, University of Zürich, Switzerland.

March 20, 2025

Word count: 1342

Abstract

Motivation: Accurate and efficient clinical genetic reporting from whole genome sequencing (WGS) data is imperative for modern genomic medicine. Current methods often involve labour-intensive processes with potential for inconsistency. Dante is designed to automate the transformation of raw WGS pipeline outputs into structured, standards-compliant clinical reports.

Results: Dante integrates a suite of R scripts and R Markdown templates to generate comprehensive clinical reports. The application incorporates variant annotation, ACMG guideline-based classification, and data visualisation—including principal component analysis (PCA) plots and detailed tabular summaries. Validation on test datasets demonstrates enhanced reproducibility and reduced processing time relative to conventional manual approaches.

Availability: The source code and documentation are accessible at https://github.com/DylanLawless/Dante. Dante is available under the MIT licence. The supporting datasets will be maintained for a minimum of two years following publication.

*Addresses for correspondence: Dylan.Lawless@kispi.uzh.ch

Acronyms

1 Introduction

Whole genome sequencing (WGS) has revolutionised clinical genetics by enabling comprehensive analysis of genomic variation. WGS analysis is typically performed using tools such as GATK4 and DeepVariant for variant detection, while annotation is achieved with specialised tools like Ensembl VEP, Annovar, Nirvana, and combined with a rich source of annotation databases such as dbNSFP and gnomAD, which compile detailed information on gene impact, variant consequences, and population frequencies. Variant classification and pathogenicity scoring are further refined using standards and tools such as the ACMG guidelines and GuRu, among others, to assess the deleterious potential of each variant.

Despite these advances, a critical gap remains in prioritising and delivering this complex genetic evidence in a clinically actionable format. The abundance of detailed clinical genetics data makes it challenging for bioinformaticians to effectively communicate results to clinicians, who often must rely on simplified metrics. Dante addresses this final step by automating the generation of clinical reports that synthesise and present rich underlying evidence in a clear, concise format, thereby facilitating informed decision-making at the patient bedside.

2 Materials and Methods

2.1 Implementation

Dante is implemented in R and employs a modular suite of functions and Markdown templates to process and visualise actionable WGS data results. The core functionality is distributed across multiple scripts, including `report_runner`, `pca`, and other dedicated modules for data cleaning, analysis, and report generation. The application employs key packages such as `knitr`, `kableExtra`, and various plotting libraries to create publication-ready PDF reports. The software architecture facilitates parameterised execution, enabling users to customise reports by supplying patient-specific identifiers and associated metadata.

2.2 Usage

To generate a report, users execute the `report.Rmd` script, which orchestrates the analysis and report compilation process. The script accepts parameters such as the

patient ID, data checksum, and a list of supplementary PDF files. Command-line examples and detailed instructions are provided in the documentation to ensure compatibility across diverse computing environments. The system automatically sources necessary R scripts (e.g. `import_priors`,`launch_report`) to integrate genetic variant data, ACMG scoring, and PCA visualisations into a cohesive report.

3 Results

Dante produces comprehensive clinical reports that summarise genetic variant data, including ACMG scores, variant impacts, and detailed annotations. A notable feature is the inclusion of PCA plots that contextualise individual patient data within global and local reference cohorts. Comparative analysis with manual reporting methods demonstrates that Dante not only improves consistency but also reduces the turnaround time for report generation. Figure 1 illustrates the performance metrics and visual output generated by Dante.

3.1 Integration of RAG-based Natural Language Interpretation

In Dante, candidate genetic variants—prioritised by tools such as GuRu based on diverse evidence sources—are filtered to fewer than ten top candidates. Given the complexity and variability of the supporting data, a retrieval-augmented generation (RAG) step is used to generate a concise, user-friendly interpretation summary. This summary, generated via the DeepSeek-R1 model through the `ollamar` R interface, translates quantitative evidence (e.g. ACMG scores, allele frequencies, gene annotations) into plain language. The process ensures that the narrative is directly supported by raw evidence, reducing the likelihood of hallucination and allowing users to rely on a quantifiable accuracy of the interpretation.

3.2 Benchmarking and Validation

Benchmarking was performed using clinical datasets of both known diagnosis and new unknown rare disease cases. Validation metrics focused on the reproducibility of ACMG-based classifications and the accuracy of candiadate variant prioritisation. Statistical measures confirmed that Dante delivers consistent performance across varying data sizes. The automated approach also minimised manual errors, thereby enhancing the overall reliability of the generated reports.

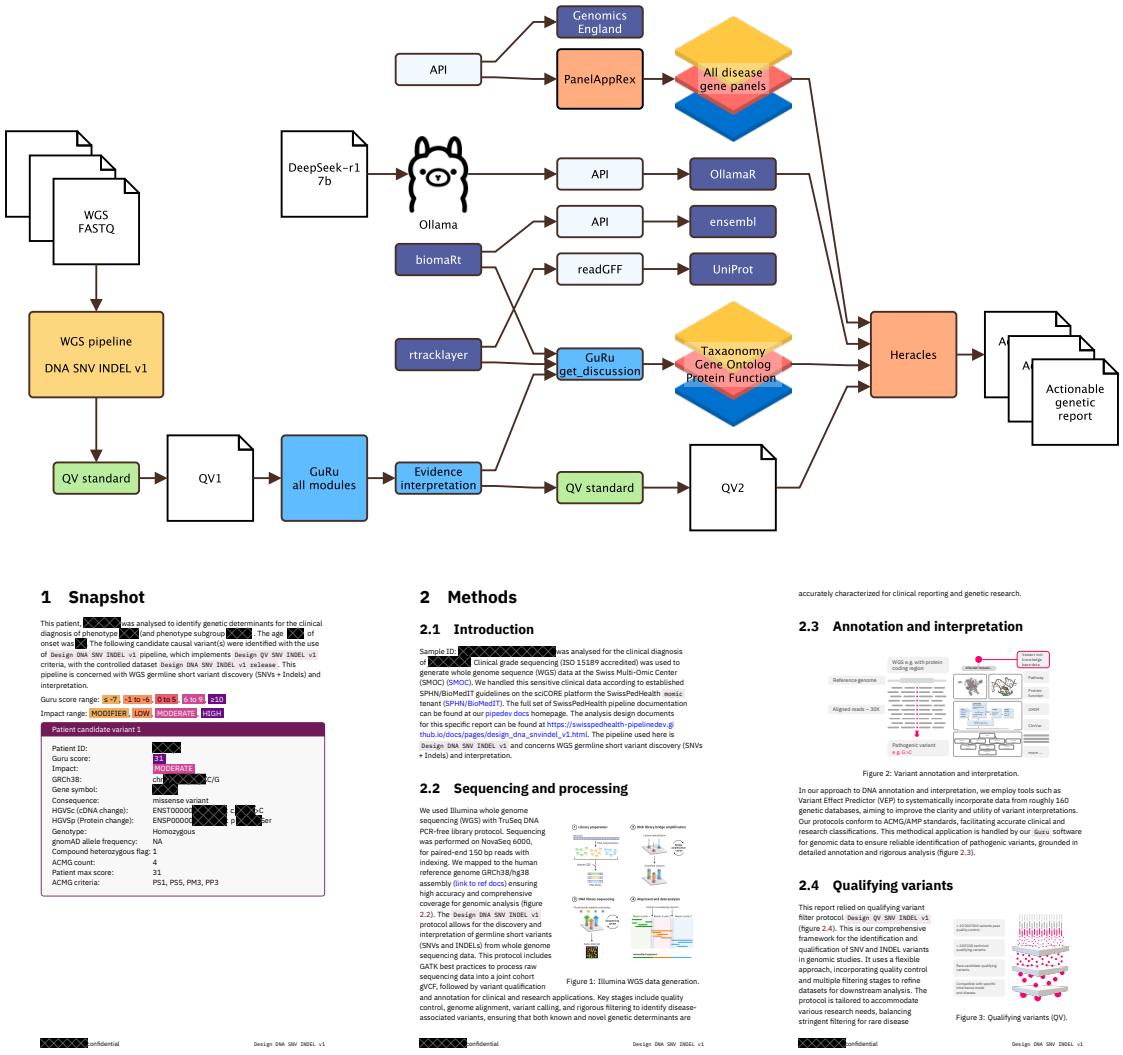

Figure 1: Model and visual outputs generated by Dante, demonstrating efficient integration of variant annotation, ACMG scoring, other metrics.

4 Summary

Dante represents a significant advancement in the field of clinical genetic report generation. By automating the transformation of raw WGS data into structured, standards-compliant reports, it streamlines the workflow for clinical laboratories and genomic research facilities. The integration of detailed variant annotation, ACMG guideline adherence, and comprehensive visualisation techniques ensures that reports are both technically robust and clinically informative. Future developments will focus on integrating additional external databases and expanding the tool's capabilities to support a broader range of genomic analyses.

Acknowledgements

The development of Dante was supported by colleagues at the University Children's Hospital Zürich and the University of Zürich. We thank the Swiss Multi-Omic Center for providing access to critical test datasets.

Author Contributions

Dylan Lawless conceived and developed the Dante pipeline, designed the study, and drafted the manuscript.

Funding

This work was funded by internal research grants at the University Children's Hospital Zürich and the University of Zürich.

Data Availability

The Dante source code, user documentation, and sample datasets are available at https://github.com/DylanLawless/Dante. Supplementary materials and further documentation can be accessed through the project repository.

Competing interest

We declare no competing interest.