GenomeSwift

Precision medicine in Kispi through rapid and automated genome analysis.

1 Background

The integration of genomic technology into healthcare, especially for rare diseases, represents a crucial frontier in medical science [1]. Despite advancements, the fragmented nature of current genome analysis methodologies in Swiss healthcare notably delays diagnoses and treatment. The accumulation of extensive genomic datasets necessitates a shift towards more integrated and automated processes [2].

Our proposed GenomeSwift bioinformatic suite aims to address these **unmet needs**. It is designed to automate and amalgamate various genome sequence analysis processes, thus **feasibly** providing swift, precise, and expansive insights into genetic diseases. This suite adheres to FAIR data principles, ensuring advancement in research and conformity with best data management practices in **Kispi**.

The persistent bottleneck in translating genomic data into practical clinical insights underscores the project's necessity [3]. Even with tools like the Genome Analysis Toolkit setting standards for variant analysis, a comprehensive system integrating these tools is essential for a robust analysis process [4]. Our **previous developments** underpin GenomeSwift (**Figure 1**) [5], [6], [7]. We have additionally developed several tools that we are excited to integrate into the pipeline, including:

- ProteoMCLustR, a protocol for protein pathway clustering [8], [7].
- SkatRbrain, a statistical analysis pipeline effective in analysing rare variants [9], [10], [11].
- Archipelago, for unified genomic statistical analysis.
- ACMGuru, incorporating American College of Medical Genetics guidelines to translate genomic data into clinical recommendations [12].
- DeepInferR, for Bayesian analysis of genetic variant probabilities and their disease impacts.

By merging these tools into a unified pipeline, GenomeSwift aims to elevate the efficiency and output of genomic data analysis in Kispi, enabling real-time, high-throughput genomic data analysis for disease discovery.

2 Hypotheses and objective

The primary objective of this study is to develop and implement GenomeSwift, an automated, comprehensive software pipeline designed for rapid, precise genomic data analysis, with a particular focus on the diagnosis and treatment of genetic diseases in Kispi.

2.1 Hypotheses:

(i) Automation and integration efficiency: Automating the genome analysis process and integrating various existing tools into a single pipeline, GenomeSwift is expected to significantly expedite genomic data processing, thus enhancing the efficiency and speed of diagnosing and planning treatments for diseases. **(ii) Enhanced diagnostic accuracy**: By incorporating advanced tools for variant detection, statistical analysis, and clinical interpretation, GenomeSwift aims to improve diagnostic accuracy for rare diseases, leading to more precise and personalised treatment approaches. **(iii) Impact on disease research**: The use of GenomeSwift in rare disease contexts is anticipated to not only improve clinical outcomes but also to enrich genomic research, offering a workflow that provides traceable genetic evidence to better classify variants.

2.2 Objectives:

(i) Tool integration Integrate existing tools such as ProteoMCLustR, SkatRbrain, and ACMGuru into a unified and automated pipeline to streamline the process from data input to clinical interpretation. **(ii) Validation and refinement**: Validate GenomeSwift using both simulated [13] and real-world datasets to confirm its

reliability and accuracy across various clinical scenarios, refining the pipeline based on performance metrics and feedback [14], [15]. **(iii) User accessibility**: Ensure GenomeSwift is user-friendly for analysts while making evidence-based results accessible to healthcare professionals and researchers, facilitating broader adoption within Kispi and potentially other institutions. **(iv) Knowledge dissemination**: Disseminate the findings and capabilities of GenomeSwift through publications and collaborations, aiming to enhance the use of genomic data for healthcare improvement and scientific discovery.

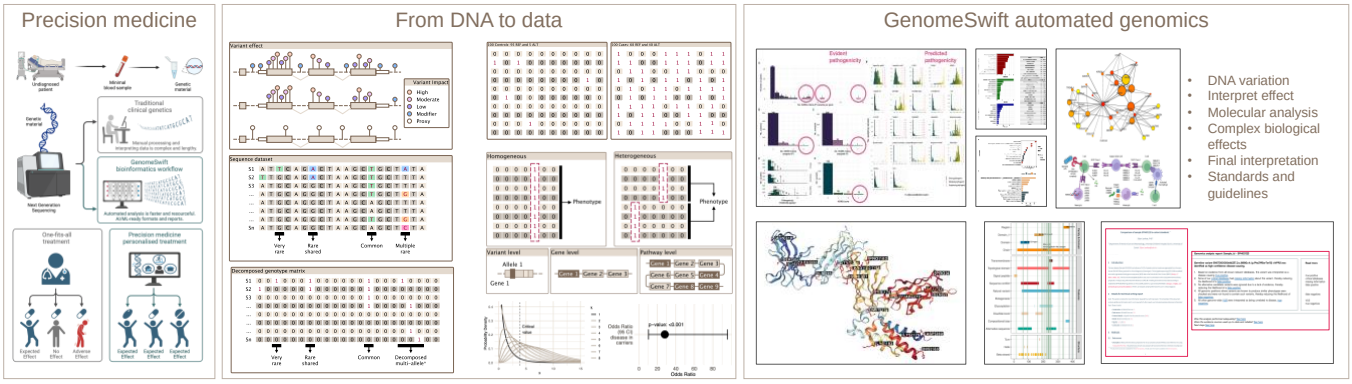

Figure 1. From DNA to diagnosis: Variant effect evidence is assessed based on standardised guidelines. Data is produced in an analysis-friendly formats for statistical or AI/ML reuse. GenomeSwift produces clinical genetics reports and database results using SPHN RDF schema concepts.

3 Detailed research plan

The GenomeSwift pipeline is designed to enhance genomic analysis through automation and integration of various analytical tools. Our detailed research plan outlines the methodology and statistical approaches we will employ to ensure the effectiveness and efficiency of GenomeSwift in processing and analysing genomic data.

(i) Integration of existing tools: The pipeline will integrate several tools that we have developed, including: ProteoMCLustR for protein pathway clustering [7], [8]; SkatRbrain for statistical analysis of genetic data [9], [10], [11]; Archipelago for a unified representation for genomic statistical analysis; ACMGuru for clinical genetic interpretation [12]; AutoConstructR for protein structure plotting, facilitating a comprehensive interpretation of genetic variations [16]; and other modular tools.

(ii) Data processing and analysis workflow: Data input: GenomeSwift will accept raw genomic data, applying preprocessing steps to ensure data quality and compatibility. **Variant detection**: Utilising the best practices from tools like GATK, the pipeline will perform variant calling, ensuring high-confidence identification of genetic variations [4]. **Statistical analysis**:Employing SkatRbrain, the pipeline will conduct robust statistical analyses to associate genomic variations with disease phenotypes, including rare variant analysis [9], [10], [11]. **Clinical interpretation**: ACMGuru will be used to interpret the clinical significance of detected variants, aligning with the American College of Medical Genetics guidelines [12].

(iii) Simulation and validation: The pipeline's efficacy will be validated using simulated datasets encompassing various disease scenarios (rare variant, common variant, polygenic risk) to ensure its robustness across different genetic contexts [3], [17], [18]. Validation will also include real-world data from Swiss hospitals to confirm the pipeline's practical applicability and accuracy.

(iv) Statistical methodologies: GenomeSwift will incorporate advanced statistical methods to analyse the association between genetic variants and diseases, ensuring the analyses are powered adequately to detect significant associations even in the context of rare diseases. The pipeline will employ a range of statistical tests suitable for different data types and study designs, ensuring the flexibility and comprehensiveness of the analysis. Specifically, optimised sequence kernel association tests (SKAT-O) will form the basis of statistical validation tests [10]. Successful outcomes will therefore demonstrate the ability to substitute

compatible drop-in methods; burden tests such as CMC [19] and WSS [20], variance component tests such as C-alpha [21] and SKAT [9], combined burden and variance component tests such as SKAT-O [10], other combination tests such as ACAT-RVAT [22], regression and generalised mixed models such as REGENIE [23] and SAGE-GENE+ [24], and others.

(v) Automation and user interface: The pipeline will feature containerisation to support development and use. Automation will be a key focus, with the pipeline designed to require minimal user intervention, streamlining the analysis process from start to finish. User output will include graphical interfaces and technical reporting documents.

(vi) Output and reporting: GenomeSwift will generate comprehensive reports, detailing the analysis results, including variant identification, statistical associations, and clinical interpretations. The pipeline will ensure that outputs are presented in an easily interpretable format, facilitating clinical decision-making and further research. The key technical data will also be generated including formats for reporting with SPHN RDF schema concepts.

4 Statement of the relevance/significance of the project

GenomeSwift represents a pivotal development at the intersection of advanced genomics and clinical application, poised to influence diagnosis and treatment protocols. Its significance spans several critical aspects for Kispi:
- **Healthcare Impact:** GenomeSwift will enhance the diagnostic process for rare diseases by delivering rapid and precise genomic analysis. This capability enables more timely and **accurate treatment decisions**, crucial for improving patient outcomes. By providing detailed genetic insights, GenomeSwift supports the shift towards **personalised medicine**, where treatments are specifically tailored to individual genetic profiles.
- **Strategic alignment:** GenomeSwift aligns with the strategic objectives of the Children's Hospital Zurich, notably enhancing the institution's **capacity** for genetic diagnostics and personalised healthcare. The project also contributes to the efficiency and effectiveness of the **broader Swiss healthcare system**, underscoring innovation and leadership in medical genomics.
- **Scientific advancement:** This project marks a considerable advance by amalgamating multiple analytical tools into a unified pipeline, thereby establishing a new benchmark in genomic analysis. GenomeSwift will streamline genomic data processing, **accelerating** research into rare diseases and potentially revealing novel therapeutic targets.
- **Collaboration and Education:** GenomeSwift will enhance **collaboration** across researchers, clinicians, and institutions by offering a shared genomic analysis format, fostering an integrated healthcare and research approach. Additionally, it will serve as an **educational resource** to improve genomic literacy and train future experts, clarifying the application of clinical genetics guidelines in analysis for healthcare professionals.
- **Sustainability and accessibility**: As an open-source initiative, GenomeSwift is accessible to a diverse user base, promoting ongoing support. The project's focus on understanding and treating diseases has the potential to make a **lasting impact** on healthcare and research, leading to enhanced patient outcomes and broader scientific insights.

5 Involved persons

The GenomeSwift project is supported by a multidisciplinary team of experts, each bringing unique expertise and experience to ensure the project's success. **Prof. Luregn Schlapbach** will act as **mentor**: project management, expected milestones, and outcomes will be tracked with progress reports. **Prof. Jacques Fellay** will provide **advice** on the tool requirements, essential guidance on data privacy, consent, and ethical considerations, ensuring that GenomeSwift adheres to the highest standards of research ethics and data protection. **Dr. Dylan Lawless** will act as **project leader**: established in genomics and bioinformatics, with extensive experience in developing computational tools for genomic analysis. He has expertise in the understanding genetic variations and their implications in diseases, especially in the context of rare disease in children. **X** will act as bioinformatics specialist: he has a profound understanding of integrating and

analysing large-scale genomic datasets. His expertise is crucial in refining the data processing algorithms and ensuring that GenomeSwift can handle complex and voluminous datasets efficiently. **X** has a background in genetic analysis method development, bringing valuable insights into the clinical implications of genomic findings. His focus is on translating genomic data into actionable clinical knowledge, which is instrumental in designing the interpretative aspects of GenomeSwift. **X** will advise as genomic analysis statistics and software development. A **PhD student** will be employed to work on development of novel research features.

Additional collaborating bioinformatics specialists from UZH, ETHZ, EPFL, and CHUV will integrate and optimise the computational tools within GenomeSwift. They will test the flow of data in the format matching our data providers (1) (e.g., SwissMultiOmic Center) to (2) HPC clusters (e.g., BioMedIT) into (3) GenomeSwift. They will test the clinical applicability and relevance of the pipeline. Open-source development will be tested by our collaborators in swisspedhealth.ch, and EPFL. Software will be tested on multiple nodes including ETHZ SIS Leonhard Med and University of Basel sciCORE Med.

6 National / international collaborations

GenomeSwift is enhanced by an extensive network of collaborations both nationally and internationally, significantly impacting healthcare and research communities. These partnerships facilitate knowledge sharing and resource exchange, crucial for the project's success. **SwissMultiOmic Center:** Serving as our primary data provider, this key national partner offers access to state-of-the-art technologies and datasets, enabling GenomeSwift to utilise high-quality genomic data and analytical resources. Regular communication between our groups aids in refining and advancing the pipeline. **Centre Hospitalier Universitaire Vaudois (CHUV):** Collaboration with CHUV researchers to review GenomeSwift's adaptability across different healthcare settings, ensuring its effectiveness and versatility. CHUV's commitment to medical genetics research provides a solid basis for collaborative enhancements of the pipeline. **Ecole Polytechnique Fédérale de Lausanne (EPFL) and ETH Zurich:** These partnerships grant GenomeSwift access to extensive expertise in bioinformatics, computational biology, and genomics, fostering a multidisciplinary integration of varied user needs and methodologies. **Global Alliance for Genomics and Health (GA4GH):** GenomeSwift aligns with GA4GH standards to enhance data interoperability and security worldwide (https://www.ga4gh.org). This partnership ensures GenomeSwift's integration into global genomic databases, adhering to international best practices in data privacy and ethics, thereby contributing to the global 'internet of genomics'. **Open-source software community:** As a collaborative open-source project, GenomeSwift benefits from the collective insights of a diverse group of developers and researchers, promoting continuous innovation and ensuring the platform's ongoing availability and improvement.

7 Timetable analysis

Our project timetable was designed and assessed using the critical path method (CPM) and program evaluation and review technique (PERT). In **Figure 2** (A) we present the network of project activities, highlighting critical paths crucial for project completion. **Figure 2** (B) illustrates the Gantt chart, detailing activity timelines and critical paths, which we tested for effective project scheduling. Lastly, **Figure 2** (C) shows the probabilistic distribution of project completion time, where we tested the risk and probability of meeting deadlines. Our analysis using CPM and PERT, provides the optimal plan to assist in timely delivery of outcomes. With a calculated probability of completion within 53 weeks at 0.97, the project is highly likely to finish on time according to the PERT analysis.[25], [26], [27], [28], [29]

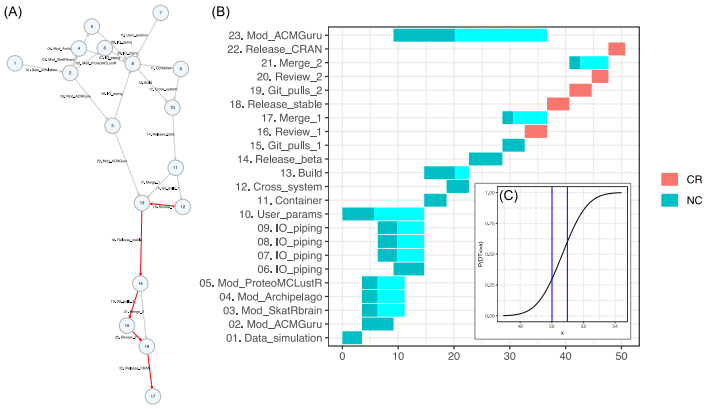

Figure 2. Timetable with PERT network and Gantt chart. (A) Network diagram depicting project activities and critical paths. Critical paths with time dependencies highlighted by color. (B) Gantt

chart illustrating project timelines and critical paths. (C) Probability distribution of project completion time for risk assessment and deadline management. CR, time-critical; NC, non-critical.

8 Budget

8.1 Available resources

The GenomeSwift project will use existing resources and infrastructure available through our collaborations and institutional support, ensuring a cost-effective approach to its development and implementation. **Institutional support**: Access to computational resources and infrastructure provided by our collaborating institutions, including the SwissMultiOmic Center, ETHZ SIS Leonhard Med, and University of Basel sciCORE Med. **Existing grants**: We will also performed development as part of current funding from related projects within our group which can be found at their respective research project pages: swisspedhealth.ch, and EPFL, CHUV.

8.2 Requested resources

8.2.1 Personal costs (including social security contributions)

A senior staff scientist and a PhD student will each be employed 50% FTE to work on development, including social security contributions. Additional funding is not required for the following: the main applicant who will oversee the project, collaborating bioinformatics specialists from UZH, ETHZ, EPFL, and CHUV who will integrate and optimise the computational tools within GenomeSwift under their current roles within their respective institutions; our collaborating clinical geneticists who will review the clinical applicability and relevance of the pipeline. Total personal costs are 90’448 CHF.

8.2.2 Material costs

Costs associated with high-performance computing (HPC) resources for data processing and analysis are listed in Table 1, including discounts covered by institutional support. Total costs are 7’628.

8.3 Summary budget table

1 year project costs	
Senior staff (107'729) 50% FTE	53865
Social security contributions (estimated 16%)	8618
Postdocs (i.e. 91'280)	0
Salary for doctoral student (i.e. 48'216) 50% FTE	24108
Social security contributions (estimated 16%)	3857
Other	
Total Direct Costs for Personnel	90448
Material costs	
HPC GPU time	1520
HPC compute time	2008
HPC storage 30 TB	2100
HPC configuration and support	1000
HPC access services	1000
Total material costs	7628
Total	98076

9 References

[1] J.-L. Casanova and L. Abel, "The human genetic determinism of life-threatening infectious diseases: genetic heterogeneity and physiological homogeneity?," *Hum. Genet.*, vol. 139, pp. 681–694, 2020.

[2] G. Povysil, S. Petrovski, J. Hostyk, V. Aggarwal, A. S. Allen, and D. B. Goldstein, "Rare-variant collapsing analyses for complex traits: guidelines and applications," *Nat. Rev. Genet.*, vol. 20, no. 12, Art. no. 12, 2019, doi: 10.1038/s41576-019-0177-4.

[3] B. S. Pedersen *et al.*, "Effective variant filtering and expected candidate variant yield in studies of rare human disease," *NPJ Genomic Med.*, vol. 6, no. 1, pp. 1–8, 2021.

[4] G. A. Van der Auwera *et al.*, "From FastQ Data to High-Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline," *Curr. Protoc. Bioinforma.*, vol. 43, no. 1, p. 11.10.1-11.10.33, 2013, doi: 10.1002/0471250953.bi1110s43.

[5] D. Lawless *et al.*, "Prevalence of CFTR variants in primary immunodeficiency patients with bronchiectasis is an important modifying cofactor," *J. Allergy Clin. Immunol.*, vol. 152, no. 1, pp. 257–265, Jul. 2023, doi: 10.1016/j.jaci.2023.01.035.

[6] D. Lawless *et al.*, "Viral Genetic Determinants of Prolonged Respiratory Syncytial Virus Infection Among Infants in a Healthy Term Birth Cohort," *J. Infect. Dis.*, vol. 227, no. 10, pp. 1194–1202, May 2023, doi: 10.1093/infdis/jiac442.

[7] Z. M. Xu *et al.*, "Genome-to-genome analysis reveals associations between human and mycobacterial genetic variation in tuberculosis patients from Tanzania." May 11, 2023. doi: 10.1101/2023.05.11.23289848.

[8] D. Szklarczyk *et al.*, "The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D362–D368, 2016, doi: 10.1093/nar/gkw937.

[9] M. C. Wu, S. Lee, T. Cai, Y. Li, M. Boehnke, and X. Lin, "Rare-Variant Association Testing for Sequencing Data with the Sequence Kernel Association Test," *Am. J. Hum. Genet.*, vol. 89, no. 1, pp. 82–93, 2011, doi: https://doi.org/10.1016/j.ajhg.2011.05.029.

[10] S. Lee *et al.*, "Optimal Unified Approach for Rare-Variant Association Testing with Application to Small-Sample Case-Control Whole-Exome Sequencing Studies," *Am. J. Hum. Genet.*, vol. 91, no. 2, pp. 224–237, 2012, doi: https://doi.org/10.1016/j.ajhg.2012.06.007.

[11] S. Lee, M. C. Wu, and X. Lin, "Optimal tests for rare variant effects in sequencing association studies," *Biostatistics*, vol. 13, no. 4, pp. 762–775, Jun. 2012, doi: 10.1093/biostatistics/kxs014.

[12] S. Richards *et al.*, "Standards and guidelines for the interpretation of sequence variants: a joint consensus recommendation of the American College of Medical Genetics and Genomics and the Association for Molecular Pathology," *Genet Med*, vol. 17, no. 5. pp. 405–24, 2015. doi: 10.1038/gim.2015.30.

[13] O. Bocher, G. Marenne, E. Génin, and H. Perdry, "Ravages: An R package for the simulation and analysis of rare variants in multicategory phenotypes," *Genet. Epidemiol.*, vol. 47, no. 6, pp. 450–460, Sep. 2023, doi: 10.1002/gepi.22529.

[14] R. H. Duerr *et al.*, "A Genome-Wide Association Study Identifies *IL23R* as an Inflammatory Bowel Disease Gene," *Science*, vol. 314, no. 5804, pp. 1461–1463, Dec. 2006, doi: 10.1126/science.1135245.

[15] Y. Liu and J. Xie, "Cauchy Combination Test: A Powerful Test With Analytic *p* -Value Calculation Under Arbitrary Dependency Structures," *J. Am. Stat. Assoc.*, vol. 115, no. 529, pp. 393–402, Jan. 2020, doi: 10.1080/01621459.2018.1554485.

[16] T. U. Consortium, "UniProt: the universal protein knowledgebase," *Nucleic Acids Res.*, vol. 45, no. D1, pp. D158–D169, Nov. 2016, doi: 10.1093/nar/gkw1099.

[17] X. Li *et al.*, "Dynamic incorporation of multiple in silico functional annotations empowers rare variant association analysis of large whole-genome sequencing studies at scale," *Nat. Genet.*, vol. 52, no. 9, pp. 969–983, Sep. 2020, doi: 10.1038/s41588-020-0676-4.

[18] Z. Li *et al.*, "A framework for detecting noncoding rare-variant associations of large-scale whole-genome sequencing studies," *Nat. Methods*, vol. 19, no. 12, pp. 1599–1611, Dec. 2022, doi: 10.1038/s41592-022-01640-x.

[19] B. Li and S. M. Leal, “Methods for detecting associations with rare variants for common diseases: application to analysis of sequence data,” *Am. J. Hum. Genet.*, vol. 83, no. 3, pp. 311–321, 2008.
[20] B. E. Madsen and S. R. Browning, “A groupwise association test for rare mutations using a weighted sum statistic,” *PLoS Genet*, vol. 5, no. 2. p. e1000384, 2009. doi: 10.1371/journal.pgen.1000384.
[21] B. M. Neale *et al.*, “Testing for an unusual distribution of rare variants,” *PLoS Genet.*, vol. 7, no. 3, p. e1001322, 2011.
[22] Y. Liu, S. Chen, Z. Li, A. C. Morrison, E. Boerwinkle, and X. Lin, “ACAT: a fast and powerful p value combination method for rare-variant analysis in sequencing studies,” *Am. J. Hum. Genet.*, vol. 104, no. 3, pp. 410–421, 2019.
[23] J. Mbatchou *et al.*, “Computationally efficient whole-genome regression for quantitative and binary traits,” *Nat. Genet.*, vol. 53, no. 7, pp. 1097–1103, Jul. 2021, doi: 10.1038/s41588-021-00870-7.
[24] W. Zhou *et al.*, “SAIGE-GENE+ improves the efficiency and accuracy of set-based rare variant association tests,” *Nat. Genet.*, vol. 54, no. 10, pp. 1466–1469, Oct. 2022, doi: 10.1038/s41588-022-01178-w.
[25] D. Miszczyńska and M. Miszczyński, *Wybrane metody badań operacyjnych*. 2002.
[26] P. R. Murthy, *Operations research (linear programming)*. bohem press, 2005.
[27] Y. Cohen and A. Sadeh, “A new approach for constructing and generating AOA networks,” *J. Eng. Comput. Archit.*, vol. 1, no. 1, pp. 1–13, 2007.
[28] H. A. Taha, “Operations research an introduction,” 2007.
[29] I. Konarzewska, M. Jewczak, and A. Kucharski, *Optymalizacja w logistyce, tom 1. Modelowanie logistycznych procesów decyzyjnych*. Wydawnictwo Uniwersytetu Łódzkiego, 2020.