Logo, company name

Description automatically generated

PENGQUIN - PErsoNal Genome QUery IN healthcare and clinical practice

Elias Crum

UGhent promotors: **Dr. Ruben Taelman**, Prof. Ruben Verborgh

VITO promotors: **Dr. Gökhan Ertaylan**, Dr. Bart Beulens

**Table of Contents**

[1. VITO research context 3](#_Toc145314657)

[1.1 Scientific environment and institutional collaboration 4](#_Toc145314658)

[2. Problem Identification 5](#_Toc145314659)

[3. Objectives 6](#_Toc145314660)

[4.1 Work Package 1: Review of Genome Data in Research and Healthcare practice (generation, storage, sharing, utilization) 8](#_Toc145314661)

[4.2 Work Package 2: Benchmark PGS Dataset Development 9](#_Toc145314662)

[4.3 Work Package 3: Storing and publishing personal genomic data in a decentralized environment 10](#_Toc145314663)

[4.4 Work Package 4: Querying over a decentralized single individual’s genomic data then a larger population of multiple individuals’ genomic data 11](#_Toc145314664)

[4.5 Work Package 5: Towards practice and Embedding of the suggested Personal Genome Sharing System in the Worldwide/European/Belgian Data Sharing Landscape 11](#_Toc145314665)

[5. Planning 13](#_Toc145314666)

[6. Risks and mitigation 14](#_Toc145314667)

[7. Application possibilities 14](#_Toc145314668)

[7.1 Overall Potential and impact 14](#_Toc145314669)

[Embedding in the VITO long term strategies 15](#_Toc145314670)

[References 16](#_Toc145314671)

[Management Summary 18](#_Toc145314672)

[Abbreviations Guide 20](#_Toc145314673)

# 1. VITO research context

As our understanding of genomics deepens, the role of Personal Genome Sequencing (*PGS*) in healthcare becomes increasingly vital. Current scientific literature highlights the potential of PGS in improving drug development processes by identifying individuals who are genetically predisposed to non-responsiveness or heightened risk of side effects, even before they participate in clinical trials1,2​. Moreover, PGS holds promise for cancer patients, as it can provide a detailed understanding of the diverse molecular processes involved in cancer development and metastasis3. This, in turn, can enhance tumour diagnosis and allow for the development of more effective treatments based on complete genome and transcriptome sequencing of each biopsy4​​.

Furthermore, PGS has the potential to help diagnose, better understand, and select optimal treatments for patients with undefined diseases. As a universal genetic test, it can be conducted once and used for life, amalgamating tests for various diseases and adverse reactions to drugs​5​. However, it is crucial to note that PGS is not devoid of challenges, including the potential overinterpretation of results due to a limited understanding of contextual information and the risk of genetic discrimination6.

The cost-effectiveness of PGS as a versatile universal test, along with the rapid advancements in medical genomics software, indicate that its broader implementation in medical practice will be feasible in the near future1​. Despite this promising outlook, there is a conspicuous gap in the current scientific discourse around implementing a citizen-centric, consent-driven personal genome data sharing framework.

This gap underscores the necessity of our proposed doctoral research project. Our aim is to develop Personal Genome Pods using Solid technology7, a set of specifications that enables decentralized data storage and sharing, which aligns well with the need for secure, consent-driven sharing of personal genome data. We envisage that this project will significantly contribute to the field of personalized medicine and genomic research by empowering individuals to utilize their personal genomes in healthcare and clinical practice in a manner that respects their autonomy and privacy.

While we continue to investigate the latest advancements and discussions in the field, we will particularly focus on the application of the Solid ecosystem in genomics and bioinformatics, as well as ethical, legal, and consent-driven sharing of personal genome data. We anticipate that our research will illuminate a path towards implementing a citizen-centric, consent-driven personal genome sharing framework in practice.

VITO Health has the technological capabilities and strategic collaborations with industrial and academic partners to play a crucial role in the development of Personal Genome Sharing. However, this key role also requires extended research into the collection, storage, and processing of large genomics data sets as it pertains to privacy and ethics.

This will be the focus of this project co-funded by VITO and Ghent University. VITO health aims to investigate the optimal method of sharing personal genomes in practice across different user scenarios and technologies together with other personal health data.

The project will further benefit from the “I am Frontier” study. The “I am Frontier” study is a small cohort of 30 healthy participants, 15 females and 15 males, who were followed up over 12 months. The 30 participants were asked to donate samples such as blood, urine, and hair monthly and collect information about health parameters and lifestyle through questionnaires and wearables. Hence, “I am Frontier” presents a unique opportunity to observe and address ethical questions regarding sensitive personal information and technical challenges associated with privacy and security on a small cohort before moving to a larger population level implementation. Clear technological, ethical and technical guidelines will be explored and set not only for VITO Health, but also on a national and international level. These guidelines and workflows can be used for future cohorts by VITO Health and other institutions.

## 1.1 Scientific environment and institutional collaboration

The Ph.D. candidate will work on the “Query Execution over Personal Genomes in Practice” project. This project is an institutional collaboration between VITO Health, and UGhent creating a strong interdisciplinary team. The Ph.D. candidate will be supported by his supervisors with strong track records in the field of interest.

The data analysis parts of the Ph.D. candidate will be co-supervised by Dr. Ruben Taelman and Dr. Gökhan Ertaylan. Dr. Ruben Taelman and Dr. Gökhan Ertaylan will be the co-promotors of the Ph.D. candidate with additional oversight provided by Dr. Ruben Verborgh. Dr. Ruben Verborgh is a professor of Semantic Web technology at Ghent University – IMEC and will provide high level feedback regarding the aims and structure of the project. Dr. Ruben Taelman currently has a faculty position at the IDLab – IMEC at UGhent, renowned for their expertise in education, and research in decentralization, query processing, and Semantic Web technologies.

Dr. Gökhan Ertaylan is the Research Lead at Precision Health Group within VITO and the project lead in the system-level data analysis within the “I am Frontier '' study. He will supervise, support, and coordinate the study. He further developed genomics applications including polygenic risk scores for disease risk prediction and pharmacogenomics passports application for prescription of drugs in a personalized manner taking into account personal biological sensitivities and potential side effects.

The proposed project will allow the Ph.D. candidate to develop expertise as a data scientist and bioinformatician with a proficiency in areas of decentralization, query processing and software development. The Ph.D. candidate will expand his network at a national and international level by attending and presenting in research meetings, workshops, and conferences. He will register in the Data Science Hub of VITO (lead by Dr. Bart Beulens) where he will actively participate and learn from other related projects. Through FLAMES and the doctoral school of UGhent, he will develop a further expertise in data sciences and bioinformatics by following trainings and workshops for broader skill development.

Lastly, there are two synergistic projects running in parallel with PENGQUIN at VITO Health. First one is the “Beyond the Genome: Ethical Aspects of Large Cohort Studies,” which has 3 Ph.D. candidates in total. The candidate will work synergistically with one of the PhD candidates with a background in bioethics. This ensures that both the technical and ethical aspects of genome data sharing are firmly covered while working together. Secondly, the real-life application of Pharmacogenomics Passports is being developed as part of a collaborative project with JESSA and UHasselt. This will ensure that the genome sharing functionality proposed to be developed in this project will actively collaborate and keep in mind the real-life applications during the project. Lastly, Data Science Hub is focusing on developing querying for other sorts of Health Data under the supervision of Dr. Bart Beulens which is going to run synergistically with the proposed PhD Project. Furthermore, this project is aligned with the SolidLab project at UGhent, which is investigating various research aspects of the Solid ecosystem but is not tied to a specific application domain. Concretely, the Ph.D. candidate will work closely together with 3 other Ph.D. students working in the SolidLab projects on storage and query processing challenges.

# 2. Problem Identification

The shift towards increasingly personalized, precise, and preventative healthcare has been catalysed by and revolves around human genome sequencing. The human genome’s growing role in healthcare, particularly at the level of the distinct, individual patient, for both medical diagnosis and treatment necessitates the collection, storage, and analysis of large, highly personal, multi-omics datasets [Aday et al.]. This Ph.D. project aims to address three main challenges presented by PGS usage in healthcare: (1) ***the privacy concerns presented by PGS data storage***, (2) ***the size of human genomes***, and (3) ***the complexity of fetching medically-relevant information from stored PGSs in real time***.

(1) **Privacy**. PGS data, like other personal data, is by nature private and its sharing should be controlled by the individual. This stance has been codified in the European Union’s General Data Protection Regulation (GDPR) law since 20188. PGS data is categorized as sensitive data in the GDPR and the PENGQUIN project has been designed with an emphasis placed on data storage strategies that maximize personal control over sensitive genomic data.

(2) **Storage**. The average human genome is slightly over 3 billion base pairs in length (3 Gbp)9 and those nucleotide files are analyzed by bioinformaticians to reveal differences between and PGS and a reference. The specific differences are recorded in a standardized variable call format (VCF) file10. VCF files are typically between 1-4 GBs in size when compressed and represent ~106 nucleotide positions of a sample human genome and associated metadata. Despite being orders of magnitude smaller than the whole genome sequence, such datasets still present storage and querying problems due to their size. Additionally, the complexity introduced by storing such data in ways conducive to protecting personal privacy further compound the challenge presented.

(3) **Querying**. This project aims to store genomic data in Solid data pods, where data may be spread across multiple files and other kinds of views allowing for both control over and the security of PGS data. These multiple files or views will be represented by a collection of nodes that are interlinked directly or indirectly in a knowledge graph contained within the user’s personal Solid data pod. A major challenge presented by this approach is that there is a need for efficient query processing algorithms that can handle large scale decentralization. We hypothesize that novel algorithms will be necessary to handle data at this size, and that we will need to introduce precomputed summaries11 into the decentralized network to achieve query execution at acceptable levels of performance.

The current innovations taking place in personalized medicine have been aided by the decrease in cost of genome sequencing in the past decades and other breakthroughs in how PGS data is used in healthcare. Thus, now is the opportune time for the establishment and optimization of a PGS storage and querying framework that can contribute to the further improvement and accessibility of personalized healthcare.

# 3. Objectives

The overarching goal is to create a system that protects the privacy of personal genome data while allowing for highly performing and widely applicable genome data storage and querying. To achieve this goal, the Ph.D. candidate will investigate the use of Solid data pods to store personal genome data in a decentralized and privacy-oriented way while addressing technical challenges that come with it. The following objectives will determine the success of the PENGQUIN project.

(1) *Investigate the existing landscape of PGS storage approaches along with the storage principles and infrastructure of the Solid ecosystem to evaluate how Solid pod storage of genomic data can be achieved in a scalable, efficient manner.* [Findings will be detailed within a review paper.]

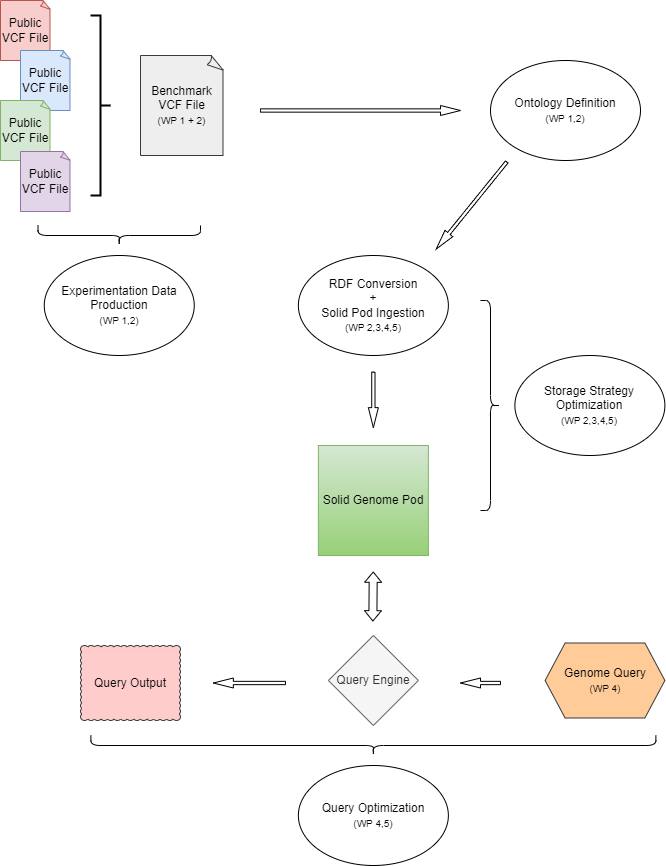
(2) *Create a simulated, practically inspired, PGS dataset that can be used for experimentation and benchmarking of storage and querying aspects of the project.* [This test dataset will include both a single PGS as well as a set of PGSs to simulate real-world situations in which PGS storage and querying is used.]

(3) *Formulate a framework for the storage, accessing, and displaying of personal genomic data using Solid that addresses the privacy concerns presented by personal genomic data for future application in advancing Personalized Healthcare and preventative medicine.* [This aim involves genomic data ingestion into a personal Solid pod as Linked Data, definition of a publicly available genomic data ontology, and collaborate in the creation of an intuitive display of genomic data within a user’s personal Solid pod.]

(4) *Investigate query processing techniques for handling the large scale of personal genomic data stored within and across Solid data vaults. Specifically, by researching efficient Linked Data querying algorithms, and strategies for improving the performance and run-time of Linked Data traversal querying algorithms for efficient querying of PGS data for clinical and personal use.*

**4. Project description**

The workflow will proceed in a relatively linear fashion. The first tasks will be the assessment of if Solid technology can accommodate efficient and effective genome storage. Simultaneously, benchmark VCF genome data will be produced to be representative of real-world VCF data for experimentation in later steps of framework development. The project will then require the definition of a genomic ontology language that will be published and made publicly available. Once a VCF ontology is established, then VCF to resource description framework (RDF) format conversion and Solid pod data ingestion will begin being experimented with. During data ingestion established data formatting standards will be integrated to allow for greater data availability for software development and scalability of this framework. Following data ingestion, query algorithms will be tested, adjusted, and created. Iterative development is likely for data ingestion methods and querying approaches due to size and complexity of VCF genomic data. Querying is planned to be possible over a single Solid pod (one VCF file) as well as over multiple Solid pods (>1 VCF files), thus iterative querying methods will also be developed.

****

***Figure 1 PENGQUIN project workflow.*** *White circles represent locations within the workflow where objectives will be achieved. For each of the steps in the workflow, the work packages affecting this step are also shown in parentheses (if applicable). This workflow is slightly oversimplified for ease of understanding the project.*

## 4.1 Work Package 1: Review of Genome Data in Research and Healthcare practice (generation, storage, sharing, utilization)

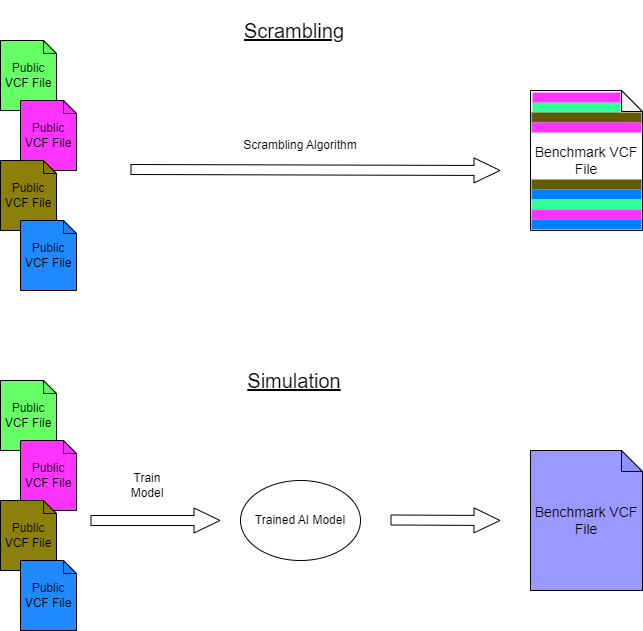
The Ph.D. candidate will conduct a literature review of: the Solid decentralized data storage framework, secure/confidential data storage in decentralized environments, existing human genome single nucleotide polymorphism summary file (.vcf files) parsing approaches, methods of ingesting non-rdf data into rdf data for storage in Solid pods, the establishment of an ontology that can handle such ingestion of genomic data, rdf and non-rdf data fragmentation and fragment organization strategies, link-traversal-based query processing paradigm, and associated linked-data querying algorithms. Along with these subjects, the candidate will also investigate strategies pertaining to improving performance of query algorithms using pre-computed summaries and/or linked data short-cut summaries. This will be coupled with the representation of genomic data as Linked Data, which will either be materialized beforehand or will be mapped on-the-fly at query-time via a virtualization approach.

While performing this literature study, the Ph.D. candidate will compose a review paper on the field of storage and query processing within decentralized environments. There has not yet been published a similar research paper, and giving an in-depth overview of the current state-of-the-art techniques will help the research community work towards making personal genomics data sharing more standardized for clinical purposes.

## 4.2 Work Package 2: Benchmark PGS Dataset Development

In this work package, the Ph.D. candidate will set up the foundations for experimentation in later work packages. Concretely, various benchmark datasets will be developed using two approaches, genome scrambling and simulation using AI [Cipriani et al.]. Genome scrambling refers to the composition of a single benchmark VCF file from the combination of parts of multiple publicly available VCF files in a way that produces a single complete VCF file from different portions from publicly available VCF files. The produced benchmark VCF file will not represent any one VCF file but will be a hybrid of all VCF files used to create it, thus preserving anonymity of publicly available VCF files. Simulation will be performed by training an AI model on various publicly available VCF files, then using the model to produce a simulated VCF file from a template publicly available VCF file. The simulation method is useful for focusing on the production of VCF files that contain rare genotypic information that may not be represented commonly in publicly available VCF files.

The primary goal of these methods of creating benchmark VCF data is to anonymize publicly available VCF data for use in development of genomic storage and querying algorithms. This approach will be used to provide benchmark VCF files for testing the storage and querying approaches within a single Solid data vault and across multiple Solid data vaults. Also importantly, this benchmark data will offer a set of realistic query workload data that mimics real-world genomic data and offers useful feedback during algorithm development.



***Figure 2 Benchmark VCF Data Production.*** *Figure illustrates the different methods of benchmark VCF data production -- scrambling versus simulation. Scrambling, shown on the top, takes publicly available VCF files as inputs and feeds them through a scrambling algorithm to return a Benchmark VCf file made of parts of the original. Simulation, shown on the bottom, takes publicly available VCF files as inputs to train an AI model that then produces a simulated benchmark VCF file that is similar to but contains no explicit parts of the input VCF files.*

## 4.3 Work Package 3: Storing and publishing personal genomic data in a decentralized environment

Today, personal genomic data is commonly represented in VCF format. To input this data into the decentralized Solid ecosystem, it needs to be translated into RDF format12 according to the Linked Data principles13. For this, there is a need for a bidirectional mapping process, that allows genomic data to be converted into RDF, and the other way around for enabling data-updates. This mapping process could either happen as a pre-processing step during the ingestion of data into Solid data pods (materialization), or during the retrieval process if genomic data will be stored in their raw non-RDF representation in Solid data pods (virtualization).

The goal of this work package is to investigate such a mapping process while comparing the impact of materialization and virtualization in terms of ingestion speed, storage size, and query performance. Furthermore, since data inside Solid data pods can be represented in a variety of different views, such as documents, the Ph.D. candidate will investigate different fragmentation strategies for publishing genomic data represented in RDF. These fragmentation strategies will be compared in terms of storage size, query performance, and privacy-compliance.

## 4.4 Work Package 4: Querying over a decentralized single individual’s genomic data then a larger population of multiple individuals’ genomic data

This work package will build upon the storage of personal genomic data in Solid data pods from the previous work package. The goal of this work package is to enable queries to be executed across personal genomic data that are contained within a single data pod or are spread over multiple data pods. Concretely, we will build upon the link traversal query processing (LTQP) paradigm, which has been shown to be an effective method for querying within a decentralized environment such as Solid14. However, LTQP is known to currently perform sub-optimally for larger dataset sizes and complex queries, which are properties present in the genomic data space. As such, there is a need for new LTQP algorithms that can achieve the required levels of performance in terms of query execution time. Concretely, the Ph.D. candidate will design new LTQP algorithms that will exploit intrinsic properties of genomic data that can help speed up query performance. Furthermore, the impact of pre-generated summaries11 will be investigated when introducing them to the decentralized environment. The working approach will then be applied to a population of personal genomic data in which new strategies of organizing computational architecture and utilizing metadata between genomes will be investigated to optimize storage size and query performance when querying more than one personal genome.

## 

## 4.5 Work Package 5: Towards practice and Embedding of the suggested Personal Genome Sharing System in the Worldwide/European/Belgian Data Sharing Landscape

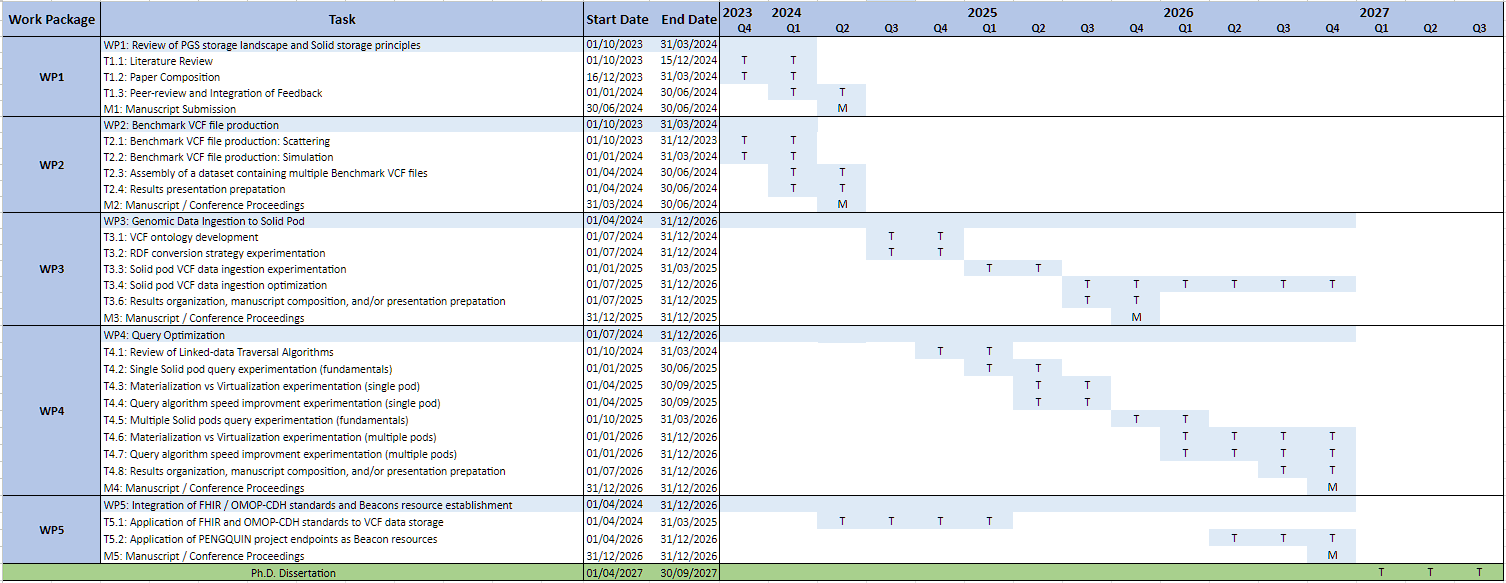
The final work package be applicable to all other work packages to make the suggested work relevant in real life practice as well as allow the current project to be both improved and applied to other applications in the future.

To avoid privacy issues and improve the communicability of sensitive medical data findings, the Observational Medical Observations Partnership (*OMOP*)15 initiative has established methods and data sharing standards that allow for the distribution of medical data findings without the fear of infringing on the privacy of sensitive medical data. Within the OMOP initiative, the genomic common data model (G-CDM)16 has been proposed as a method of improving the circulation and discoverability of genomic research results for researchers. Similarly, the FHIR Genomics Implementation initiative17 attempts to solve similar issues with genomic data sharing and discoverability without infringing on medical data privacy through the establishment of a modular data ontology. We aim to integrate one or both of these data structure and discoverability standards in our PENGQUIN project architecture to encourage improved data and result sharing among the greater research community.

An overarching aim for this project is to help drive improvement and accessibility of personalized healthcare through improved PGS privacy and storage and querying efficiency. These aims align well with the international Beacon initiative18,19 to increase the availability and ease of access of genomic data for researchers globally. Thus, we aim to make the PENGQUIN Solid pod VCF data storage endpoints beacon resources that can be discoverable via the Beacon API. Importantly, this option to contribute personal genomic data to the greater Beacon data network, thereby making it discoverable to researchers, will be controllable on an individual basis as an opt-in option. Another advantage of the Solid ecosystem in this context is the ability of any individual to change their beacon sharing preference from their personal Solid pod at any point in time. Increasing the amount of genomic data available to researchers will help contribute to the further progression of personalized healthcare efficacy for larger, and more diverse populations of people, aligning closely with our goals for this Ph.D. project.

Within VITO, the PENGQUIN framework is to be designed to be integrated into ongoing research projects related to pharmacogenomics. More specifically, ongoing research investigating an improved system of drug prescription based on PGS factors for individual patients is currently underway at VITO. The aim for the PENGQUIN project is to provide a more complete, efficient, and reliable framework for patient genome storage and querying to improve the efficiency and efficacy of how pharmaceuticals are prescribed to patients.

# 5. Planning

The Ph.D. project consists of 5 work packages denoted as “WP”. This will be bundled into a Ph.D. dissertation for a final thesis defense. *Figure 3* shows a Gantt-chart of the Ph.D. project on a quarterly basis. Each work package is split up into different tasks with a dedicated amount of time allocated to it. This will allow for a good time and project management.

***Figure 3******Gantt chart of the Ph.D. project timeline on a quarterly basis****. T denotes different tasks of a work package. M denotes different milestones, such as finishing and submitting a manuscript to a journal. Green denotes working on the thesis dissertation and defence. This is allocated as the final 6-9 months of the Ph.D. project to make all preparations for the defence.*

# 6. Risks and mitigation

* **Dataset size:** If personal genome datasets are too large to store in a decentralized manner with acceptable query performance. If this occurs, a shift towards more centralization will be investigated in the form of summaries [3], and more expressive query interfaces will be added to Solid data vaults [7]. The range of pods over which summaries and query interfaces are defined will be configurable, which means that the trade-off between centralization and decentralization can be tuned.
* **Data availability, ethics and legislation (GDPR guidelines):** VITO Health and UGhent have an institutional agreement which regulates the legal and ethical aspect of the project. This includes confidentiality, ownership, and privacy issues. Genomics data from other sources is freely accessible online as the data is currently not legislated under GDPR guidelines.
* **Computational Load:** Various software packages will be used for brute force SAP detection. Both VITO and UGhent have access to the Flemish Supercomputing Centre for grid computing in case the Ph.D. lacks computational resources.
* **Evolving PGS Technological Advances**: Long read sequencing has been decreasing in price in recent years and offers many advantages over the currently most common form of short-read shotgun genome sequencing. With a widespread transition to long read sequencing, the way downstream genomic files, such as VCF files, are created and formatted may change. It is unlikely that such a drastic change will take place during this Ph.D. project, but if such a breakthrough occurs, the Ph.D. candidate and advisors will assess altering the input genomic files to align with the project’s goal of creating a framework useful for advancing personalized healthcare.
* **Software Dependencies:** Various software will be used for this project. Generally, software is known to exhibit issues concerning quality, versioning, and documentation. If such problems are encountered troubleshooting will include utilizing other, comparable software to accomplish the task or contacting the producer of said software for troubleshooting.

# 7. Application possibilities

## 7.1 Overall Potential and impact

The project supports the desire to move towards a healthcare system increasingly focused on personalized genomics and preventive medicine. This is highly desired as personalized medicine offers an improved medication selection, targeted therapy, reduced adverse effects, increases patient compliance and confidence, and improves cost effectiveness [Mathur et al.]. Importantly, personalized healthcare, particularly its reliance on PGS, introduces technical and privacy challenges. Additionally, with increased use of personal genomics within healthcare, highly performing systems must be developed to store and access such data in an efficient, secure, and reliable way. Such a system presents a challenge due to the large size, complexity, and sensitive nature of personal genomic data.

This project will provide an important step toward creating such a system while also increasing knowledge of how biologically and medically relevant Linked Data can be stored and queried while preserving the private nature of genomic data for Ghent University and VITO.

# Embedding in the VITO long term strategies

Precision Health Group in VITO Health has the ambition to develop applications towards making personalized healthcare a reality through data-driven research. This is facilitated through having joined PhDs with academic institutes to maintain strong and relevant research. The Ph.D. candidate will actively make connections with other institutes contributing to the research field of interest. Algorithms, methodologies, and results from the Ph.D. project may be implemented in different institutions partnered with VITO, JESSA and Ghent University such as Domus Medica or Hospitals to contribute towards unveiling the personalized Genome Pods’ power in practice.

PENGQUIN, by facilitating the storing and consent driven dynamic sharing of PGS, will be poised to play a central role in developing Digital Twins of citizens/patients in healthcare and clinical practice.

Moreover, the Ph.D. project will establish a link to both the healthcare industry and computer science research by developing computational tools and methodologies which will aid in the confident storage and querying of genomic data for use in existing research aims within pharmacogenomics at VITO.

Lastly, the algorithms and methodologies developed will be made available for future research. This way, the research will continuously contribute to academia and industry.

# 

# References

1. Wang, L., McLeod, H. L. & Weinshilboum, R. M. Genomics and Drug Response. *N. Engl. J. Med.* **364**, 1144–1153 (2011).

2. Ko, Y. K. & Gim, J.-A. New Drug Development and Clinical Trial Design by Applying Genomic Information Management. *Pharmaceutics* **14**, 1539 (2022).

3. Ibrahim, J., Peeters, M., Van Camp, G. & Op De Beeck, K. Methylation biomarkers for early cancer detection and diagnosis: Current and future perspectives. *Eur. J. Cancer* **178**, 91–113 (2023).

4. McLeod, H. L. Cancer Pharmacogenomics: Early Promise, But Concerted Effort Needed. *Science* **339**, 1563–1566 (2013).

5. Hens, K., Nys, H., Cassiman, J.-J. & Dierickx, K. The return of individual research findings in paediatric genetic research. *J. Med. Ethics* **37**, 179–183 (2011).

6. Alzu’bi, A., Zhou, L. & Watzlaf, V. Personal genomic information management and personalized medicine: challenges, current solutions, and roles of HIM professionals. *Perspect. Health Inf. Manag.* **11**, 1c (2014).

7. Mansour, E. *et al.* A Demonstration of the Solid Platform for Social Web Applications. in *Proceedings of the 25th International Conference Companion on World Wide Web - WWW ’16 Companion* 223–226 (ACM Press, 2016). doi:10.1145/2872518.2890529.

8. *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)*. *OJ* 1–88 (2016).

9. Piovesan, A. *et al.* On the length, weight and GC content of the human genome. *BMC Res. Notes* **12**, 106 (2019).

10. Danecek, P. *et al.* The variant call format and VCFtools. *Bioinformatics* **27**, 2156–2158 (2011).

11. Umbrich, J., Hose, K., Karnstedt, M., Harth, A. & Polleres, A. Comparing data summaries for processing live queries over Linked Data. *World Wide Web* **14**, 495–544 (2011).

12. Cyganiak, R., Wood, D. & Lanthaler, M. Concepts and Abstract Syntax. *W3C* (2014).

13. Berners-Lee, T. Linked Data. (2009).

14. Taelman, R. & Verborgh, R. Evaluation of Link Traversal Query Execution over Decentralized Environments with Structural Assumptions. (2023) doi:10.48550/ARXIV.2302.06933.

15. FitzHenry, F. *et al.* Creating a Common Data Model for Comparative Effectiveness with the Observational Medical Outcomes Partnership. *Appl. Clin. Inform.* **06**, 536–547 (2015).

16. Shin, S. J. *et al.* Genomic Common Data Model for Seamless Interoperation of Biomedical Data in Clinical Practice: Retrospective Study. *J. Med. Internet Res.* **21**, e13249 (2019).

17. Ayaz, M., Pasha, M. F., Alzahrani, M. Y., Budiarto, R. & Stiawan, D. The Fast Health Interoperability Resources (FHIR) Standard: Systematic Literature Review of Implementations, Applications, Challenges and Opportunities. *JMIR Med. Inform.* **9**, e21929 (2021).

18. Fiume, M. *et al.* Federated discovery and sharing of genomic data using Beacons. *Nat. Biotechnol.* **37**, 220–224 (2019).

19. Rambla, J. *et al.* Beacon v2 and Beacon networks: A ‘lingua franca’ for federated data discovery in biomedical genomics, and beyond. *Hum. Mutat.* **43**, 791–799 (2022).

# Management Summary

Human genome sequencing has heavily contributed to advancements in personalized medicine, such as in drug development and prescription as well as in cancer diagnosis and treatment. As the cost of human genome sequencing has decreased, personalized medicine has become more accessible to the general population, which has, and will continue to, increase the importance of personal genome sequencing (PGS) and hopefully improve healthcare outcomes. Similar to other sensitive personal data, these PGS data will necessitate scalable, efficient, and secure storage, sharing, and querying technologies. Thus, the investigation of adapting existing methods of secure, decentralized data storage and querying techniques will help drive further improvement and availability of personalized medicine. Here we present *PENGQUIN - PErsoNal Genome QUery IN healthcare and clinical practice*, a Ph.D. project designed to address the challenges mentioned above while providing a **scalable framework for PGS storage and querying** to be used by clinical and healthcare professionals for improving the availability and efficiency of personalized medicine.

The PENGQUIN project aims to use **decentralized web technology to store PGS data**, decreasing the risk of PGS data being leaked in data breaches, while also giving consumers more direct control over exactly who can view their genome. Currently, most PGS storage is handled centrally, essentially preventing consumers control over their personal genome data. Another challenge posed by storing PGS data is its size. PGS data is very large and, thus, difficult to search through quickly and efficiently. The PENGQUIN project aims to create a framework that safeguards the privacy of personal genome data while simultaneously offering high-performance and versatile genome data storage and querying capabilities **by leveraging the Solid ecosystem and Linked-data querying algorithms**.

The Ph.D. project’s primary objectives aim to address 4 main scientific questions. (1) Can a decentralized data storage framework such as *Solid handle large genomic data storage*? (2) Can existing and/or novel *linked-data traversal query algorithms efficiently and quickly query genomic-sized data*? (3) Can existing and/or novel q*uerying algorithms handle queries pertaining to multiple PGSs*? (4) Can a *scalable, intuitive product be created from these components* that improves healthcare for both consumers and providers?

The PENGQUIN project will be the **first published attempt to use the Solid framework for the storage of genomic data**. Additionally, the querying of genomic data will doubtless require the modification of existing linked-data traversal algorithms for querying, with it being likely that novel algorithmic approaches be developed as well.

VITO Health, in collaboration with Ghent University, aims to play a crucial role in **advancing personal genome sharing**. The PENGQUIN project's primary focus is to be used to investigate practical scenarios and technologies for optimal personal genome sharing, and potentially be expanded to other health data. Additionally, the VITO "I am Frontier'' study, involving 30 participants providing various health-related data, serves as a valuable testing ground to implement the PENGQUIN project framework and assess its efficacy at genome sharing efficiency and privacy preservation. Ultimately, the project aims to establish clear guidelines and workflows applicable not only to VITO Health but also on a broader national and international scale for the continued improvement and accessibility of personalized medicine.

We expect to build a system for the storage, sharing, and querying of PGS data while placing the consumer in control of privacy decisions regarding their data. This result will **encourage more accessibility and public support for PGS and personalized medicine** while also being **built on a standardized data accessing system** to be universally used by healthcare professionals and software developers alike. Such a standardized system can be scaled and allow for increased ease of PSG storing, sharing, and querying.

The objectives of the project will be accomplished through *5 main work packages* over the course of the 4-year Ph.D. and are as follows: (1) *Review of Genome Data in Research and Healthcare practice*: The first work package focuses on in-depth investigation of the existing landscape of PGS approaches and the principles of Solid data pods. This exploration will culminate in a comprehensive review paper that synthesizes findings and insights of the current state of the field. (2) *Benchmark PGS Dataset Development*: To facilitate experimentation and benchmarking, a simulated, practically inspired PGS dataset will be created. This dataset will encompass both a single personal genome sequence and a set of such sequences to emulate real-world scenarios. These synthetic data sets will be invaluable for evaluating the performance of the storage and querying components of the PENGQUIN system. (3) *Storing and publishing personal genomic data in a decentralized environment*: This objective focuses on formulating a robust framework for the storage, retrieval, and display of personal genomic data using the Solid decentralized web framework. This work package will encompass the ingestion of genomic data into a personal Solid pod as Linked Data, the establishment of a publicly available genomic data ontology, and the creation of an intuitive interface for users to access and interpret their genomic information securely. (4) *Querying over a decentralized single individual’s genomic data then a larger population of multiple individuals’ genomic data*: Centres on exploring innovative query processing techniques for handling the vast scale of personal genomic data stored within and across Solid data vaults. Specifically, the research will delve into the development of efficient Linked Data querying algorithms and strategies to establish the performance and runtime efficacy of Linked Data traversal querying algorithms for PENGQUIN. These advances will enable efficient querying of personal genome data for both clinical and personal purposes. (5) *Towards practice and Embedding of the suggested Personal Genome Sharing System in the Worldwide/European/Belgian Data Sharing Landscape*: allow the option of personal genome data and query results to be discoverable by researchers using established, privacy retention focused medical data sharing initiatives.

The packages will be worked on in a somewhat sequential manner, with work packages 1 and 2 being the primary focus of the first two quarters of the project. Work packages 3 and 4 will be investigated next and will be repeated many times throughout the remainder of the Ph.D. to optimize performance and efficiency of storage and querying. Work package 5 will be applied in the beginning planning phases as well as at the end nearer to deployment. For specific details about the planned timing and milestones within the Ph.D. project, please consult the Doctoral Plan.

The PENGQUIN project aims to propel personalized medicine forward by fostering the development of a secure, efficient, and widely applicable system for personal genome storage and querying within the Solid ecosystem. The PENGQUIN project endeavours to unlock the full potential of personal genomic data storage and querying to help advance personalized healthcare and preventive medicine, ultimately benefiting individuals and the broader healthcare community.

# Abbreviations Guide

FHIR - Fast Health Interoperability Resources

FLAMES - Flanders' Training Network for Methodology and Statistics

GDPR - European Union’s General Data Protection Regulation

G-CDM - genomic common data model

LTQP - Linked-traversal-based query processing

OPOM - Observational Medical Observations Partnership

PENGQUIN - PErsoNal Genome QUery IN healthcare and clinical practice

PGS - Personal genome sequence

RDF - Resource description format

VCF - Variable call format