TRI-IVIS for Genome Knowledge Related to SNP Analyses

Johanna Pethke[1,*,†], Thomas Krause[2,†], Michael Kramer[3,†] and Matthias Hemmje[4,†]

[1]*Fernuniversität Hagen, Universitätsstr. 1, 58097 Hagen, Germany*
[2]*Fernuniversität Hagen, Universitätsstr. 1, 58097 Hagen, Germany*
[3]*ImmBioMed Business Consultants GmbH & Co. KG, Bergstr. 85, 64319 Pfungstadt*
[4]*Fernuniversität Hagen, Universitätsstr. 1, 58097 Hagen, Germany*

Abstract

The analysis and interpretation of Single Nucleotide Polymorphisms (SNPs) in regulated diagnostic environments requires the systematic integration of heterogeneous knowledge sources, ranging from structured genome databases and standardized formats to narrative scientific evidence and AI-generated predictions. Existing systems lack explicit epistemic differentiation of knowledge types within real-world diagnostic application contexts. This article presents an epistemically differentiated conceptual model for SNP-related knowledge that links structured, derived, and unstructured forms of genomic information with empirically derived diagnostic use cases. We define the Genome Knowledge Unit (GeKoU) as a formal object. This multidimensional modeling enables explicit differentiation between curated database knowledge, model-based predictions generated by Genomic Foundation Models, and narrative evidence extracted from the biomedical literature. The GeKoU is applied to a use case identified in a preliminary study and shows that SNPs do not exist as static data objects, but as context-dependent sets of epistemically heterogeneous knowledge units. By explicitly modeling representation changes and versioning levels the proposed framework supports transparency, regulatory traceability, and reproducibility.

The model forms the conceptual basis for the development of SnipFlow, a micro-service-based knowledge management component to be integrated into the GenDAI platform. By combining explicit knowledge modeling, AI-supported inference and interoperable standards, the approach contributes to the development of trustworthy, reproducible and intelligent information systems for genomic diagnostics and other regulated, data-intensive areas.

Keywords

TRI-IVIS, SNP Analysis, Genomic Foundation Models, Trustworthy AI, Reproducible Information Systems

1. Introduction

The analysis of ***S**ingle **N**ucleotide **P**olymorphisms (SNPs)* is a central part of modern genomic diagnostics and personalized medicine [1]. SNPs are defined by specific sequence deviations in DNA, which can be represented as complex knowledge artifacts. Their interpretation is based on the integration of heterogeneous information sources. In diagnostic practice, SNP-related knowledge is distributed across various forms of representation and knowledge domains. On the one hand, there are structured knowledge sources that include, for example, genomic databases. A distinction is made, among other things, between curated and non-curated databases, whereby in curated databases the content has been reviewed, evaluated, and approved by experts. Regardless of this, the knowledge in these sources is available in a formally structured, machine-readable form. For example, data formats such as ***V**ariant **C**alling **F**iles (VCF)* [3], which describe genetic variants in a standardized manner using information such as position in the genome, reference and alternative alleles, and quality metrics, can be integrated. Similarly, the ***H**uman **G**enome **V**ariation **S**ociety (HGVS)* nomenclature [4] defines a standardized syntax for the unambiguous description of genetic variants at various levels, such as genomic or proteomic. These standardizations enable consistent processing and integration of the data. The knowledge thus available forms a structured foundation for the analysis of genetic variants, particularly SNPs. Knowledge derived from such structured sources about the genome or specifically about SNPs is referred to below

Proceedings of TRI-IVIS 2026, the First International Workshop on Trustable, Reproducible, and Intelligent Information Visualization Systems, June 2026, Venice, Italy.

johanna.pethke@fernuni-hagen.de (J. Pethke); thomas.krause@fernuni-hagen.de (T. Krause); m.kramer@immbiomed.de (M. Kramer); matthias.hemmje@fernuni-hagen.de (M. Hemmje)

as ***Stru**ctured **Ge**nome **Kno**wledge (StuGeKo)*. On the other hand, unstructured knowledge sources such as scientific publications, clinical reports, or narrative evidence texts play a central role because they contain contextual, interpretative information that is not fully formalized. This unstructured knowledge can be used by experts for context-dependent interpretation. The knowledge included and extractable from these sources is referred to as ***Un**structured **Ge**nome **Kno**wledge (UnGeKo)*.

This gap is further exacerbated by the increasing use of machine learning methods, particularly ***G**enomic **F**oundation **M**odels (GFMs)*. Such models generate knowledge artifacts that are formally structured, for example in the form of assessments or functional predictions, but their epistemic status differs from that of classical asserted knowledge. For regulated diagnostic contexts, it is therefore insufficient to classify knowledge solely based on its degree of formal structuring. It has to be clarified whether the predictions are model-based.
In this context, a distinction is made between two fundamental epistemic forms of knowledge: asserted knowledge refers to knowledge based on curated and validated, whereas inferred knowledge is derived through data-driven models or interpretive processes and thus has a probabilistic or hypothesis-based character. This distinction is particularly significant in regulated diagnostic contexts, as both forms of knowledge impose different requirements regarding traceability, validation, and regulatory assessment.

The processing and integration of these heterogeneous forms of knowledge takes place within the context of ***T**rustable **R**eproducible, and **I**ntelligent **I**nformation **VI**sualization **S**ystems **I**ntelligent (TRI-IVIS)*, a conceptual framework for designing modern information systems in regulated domains. TRI-IVIS aims to systematically combine transparency and traceability (Trustworthy), the repeatability of analysis and decision-making processes (Reproducible), and the intelligent processing and preparation of data and knowledge (Intelligent). In this context, the integration of structured and unstructured SNP-related knowledge poses a particular challenge, as different forms of knowledge, levels of evidence, and contexts have to be processed and made comprehensible and explainable.

Initially, the automated extraction of StuGeKo and UnGeKo has to be examined, addressing the different epistemic forms of asserted knowledge and inferred knowledge. The integration of asserted and inferred knowledge is particularly relevant, as both contribute differently to diagnostic decision-making and have to be handled in a transparent and traceable manner. Medical laboratories are also subject to strict regulatory requirements that mandate transparency, traceability, version control, and reproducibility for all steps of the analysis process. Therefore, an overarching goal is to design and develop a ***K**nowledge **M**anagement **S**ystem (KMS)* that meets these requirements by identifying, extracting, processing, and preparing knowledge artifacts in such a way that asserted and inferred knowledge are explicitly represented and made accessible for downstream visualization and analysis. These goals can be summarized in the following ***R**esearch **G**oals (RG)*:

RQ1 How can the extraction of StuGeKo be automatically supported by software?

RQ2 How can the identification and extraction of UnGeKo be automated by software?

RQ3 How can the combination and contextualization of UnGeKo and StuGeKo be automated by software?

RQ4 How can a KMS be modeled and developed that support medical diagnostic laboratories in accessing, processing and using SNP-Knowledge?

To support the answering of the research questions, the empirically collected, SNP-centered diagnostic application context is first introduced. This is based on a semi-structured expert interview in a medical diagnostic laboratory.

Building on this, a differentiated knowledge typology is developed that distinguishes between structured genome knowledge, unstructured genome knowledge as well as asserted genome knowledge and inferred genome knowledge. Subsequently, an epistemic metadata model is formalized in which each knowledge unit is described not only by its semantic content, but also by its degree of structuring, its evidence status, and its version. This formalization leads to the definition of the ***Ge**nome **Kno**wledge **U**nit (GeKoU)*, which is introduced in Section 3.

2. State of the Art in Science and Technology

This section examines current approaches to structured and unstructured representation of genome knowledge, AI-supported analysis, and interoperable clinical integration. Particular attention is given to how existing standards and platforms address interoperability, explainability, and reproducibility, and where conceptual gaps exist in terms of epistemic differentiation and context-dependent representation. A preliminary study is presented, along with the resulting use context and modeling requirements. Genome diagnostics are based on formally structured representation formats and curated databases that enable machine-readable storage and processing of variant information. Variant calling pipelines identify differences between the sample and reference genome and encode them in standardized formats such as VCF [3]. VCF presents genome coordinates, alleles, quality metrics, and annotations in a structured manner, thus forming the basis for automated and reproducible analysis processes.

In addition, the HGVS nomenclature defines a standardized syntax for describing sequence variants at the genomic, transcriptomic, and proteomic levels [4]. Together, they form a syntactic-semantic basis for structured genome knowledge.

Curated databases such as dbSNP and ClinVar extend this formal structure by integrating frequencies, clinical assessments, and referenced evidence [5, 6]. ClinVar aggregates laboratory contributions and assigns variants to standardized classifications that are documented, versioned, and referenced. The term ***a**sserted **Ge**nome **Kn**owledge (aGeKo)* refers to verified genome knowledge. This includes knowledge verified by experts and extracted by automated methods from unstructured sources.

Beyond file-based formats, interoperability standards such as ***H**ealth **L**evel Seven **F**ast **H**ealthcare **I**nteroperability **R**esources (HL7 FHIR)* enable the embedding of genomic findings in clinical information systems [7]. FHIR defines resources for the structured, versioned, and traceable representation of findings, annotations, and contextual information. However, these standards primarily address syntactic and structural aspects. A systematic epistemic typology that distinguishes between curated statements, model-based predictions, and context-dependent interpretations is not provided.
Curated statements are understood to be documented units of knowledge based on verified, validated, and often expert-reviewed sources, and thus exhibit a high degree of reliability and traceability. Model-based predictions, on the other hand, refer to units of knowledge derived through data-driven methods, particularly machine learning, and therefore have a probabilistic and hypothetical nature. Context-dependent interpretations encompass units of knowledge that are generated within the context of specific application scenarios, such as by experts in the diagnostic process, taking additional contextual information into account, and are therefore not fully standardizable.

In regulated diagnostic contexts, provenance and traceability are just as relevant as structural formalization. Provenance models such as W3C-PROV describe entities, activities, and agents and enable the formal reconstruction of chains of origin [8]. FHIR also contains a provenance resource for documenting creation and modification processes [7]. These approaches increase transparency and regulatory traceability, but primarily address origin and strength of evidence. They do not explicitly model the epistemic nature of a knowledge unit and do not capture systematic status transitions within diagnostic workflows.

The use of data-driven models, in particular GFMs, broadens the spectrum of structured knowledge generation [10]. These models derive functional effects or pathogenicity probabilities directly from sequence contexts. The resulting scores and predictions are formally structured but are not necessarily based on curated or experimentally confirmed evidence. They are referred to here as ***i**nferred **Ge**nome **Kn**owledge (iGeKo).*

While aGeKo is based on documented and curated statements, iGeKo is generated through data-driven inference processes. Both forms of knowledge differ fundamentally in their epistemic status. This difference is not systematically modeled by existing standards. Narrative knowledge sources such as publications, case reports, or expert interpretations contain context-rich information on functional effects or disease associations [11, 12, 13]. Despite such processing steps, their epistemic origin remains text-based and context-dependent.

Approaches such as knowledge graphs, ontologies, or hybrid AI systems integrate heterogeneous data sources into common semantic structures [15, 16, 17]. They improve interoperability and relational modeling, but typically do not systematically distinguish between StuGeKo, UnGeKo, aGeKo and iGeKo. Platforms such as ***Gen**omics **D**iagnostics and **A**rtificial **I**ntelligence (GenDAI)* address the infrastructural integration of AI-supported analyses [18], while reference models such as ***A**rtificial **I**ntelligence and **I**nformation **Vis**ualization for **Big Data** (AI2VIS4BigData)* emphasize aspects such as explainability and reproducibility [19]. However, SNP-centered modeling that explicitly differentiates between structured and unstructured forms of knowledge epistemically and systematically links them along real diagnostic application contexts is still lacking.

An exploratory preliminary study was conducted to systematically derive the diagnostic usage context. The data collection occurred at the ImmBioMed diagnostic laboratory in Heidelberg as part of a qualitative, exploratory field study designed to capture the practical diagnostic context. ImmBioMed Business Consultants GmbH & Co. KG is a specialized diagnostic laboratory that focuses on genomic and molecular genetic analyses in a clinical context, operating under strict regulatory requirements regarding quality assurance, documentation, and traceability.

The selected methodological approach was a combination of semi-structured expert interviews and participant observation. The interviewers were Prof. Dr. Hemmje and Johanna Pethke. Prof. Dr. Kramer and Fernando Leonardi participated in the roles of clinical pathologist, laboratory data analyst, medical data analyst, and compliance officer.

The data was collected using two complementary methods: First, semi-structured interviews were conducted to systematically capture explicit knowledge, decision-making logic, and requirements. Second, participant observation was carried out on-site in the form of a focused ethnographic field study. During a walk-through of the laboratory, process steps, systems in use, documentation practices, and media breaks that occurred were directly observed and documented in structured field notes.

The collected data were analyzed using a structured qualitative content analysis with a combined deductive-inductive approach to category formation. Deductive categories were derived from the research questions and theoretical considerations, while inductive categories were developed directly from the empirical data. The aim of the analysis was to reconstruct the actual diagnostic context, including implicit knowledge practices, informal documentation solutions, and regulatory frameworks. The choice of a qualitative research approach is justified by the fact that diagnostic processes are highly context-dependent and characterized by implicit expert knowledge, informal practices, and complex interactions that cannot be adequately captured by quantitative methods.

Based on this empirical reconstruction, the use context model shown in Figure 1 was developed. This model abstracts the identified roles, artifacts, knowledge units, transformation steps, and provenance requirements into a structured representation of the usage context and forms the methodologically sound basis for the subsequent conceptual development of GeKoU.

The results of the preliminary study conducted within the TRI-IVIS framework show that current diagnostic practice is characterized by several structural and epistemic features. In particular, it emerged that information is distributed across various systems and knowledge sources, including databases, scientific literature, model outputs, and laboratory artifacts. Furthermore, the interpretation of sequence data depends highly on implicit expert knowledge, particularly in the context of visual analyses. Concurrently, there is a lack of suitable approaches to systematically and transparently represent the versions and provenance of knowledge artifacts.

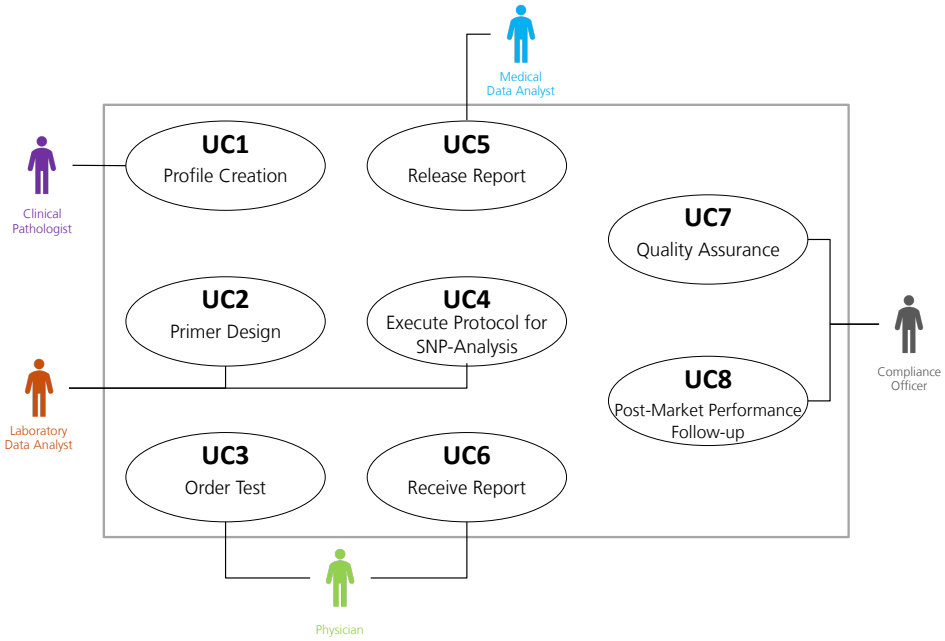

Figure 1: Use Context

These findings present ***R**emaining **C**hallenges (RCs)* for the design of information systems in genomic diagnostics:

- **RC1: Integration of heterogeneous knowledge sources**,
  to consistently consolidate fragmented information from different systems and make it usable.
- **RC2: Explicit modeling of knowledge forms**,
  to make interpretive decision-making processes more transparent, traceable, and reproducible.
- **RC3: Systematic mapping of versions and provenance**,
  to ensure the regulatory requirements for traceability and reproducibility of analysis processes.

The research has shown that genomic diagnostics has a mature ecosystem for the structural representation, semantic integration, and procedural traceability of genomic data. Empirical research into the diagnostic application context has made it clear that these established approaches do not fully reflect the actual use of knowledge in the laboratory. SNP-related knowledge does not appear in the diagnostic workflow as a static, uniformly structured object, but rather in different forms of representation, each of which is associated with a specific epistemic status.

In addition, it was found that the same SNP undergoes several representation states in the course of the analysis workflow. These transitions are epistemically significant, but are not usually modeled as such in existing systems. What is missing is a formal entity that integrates semantic content, structural form, epistemic status, and versioning, while at the same time explicitly mapping the transformations between different states of knowledge.

An approach is needed that regards SNP-related knowledge not exclusively as a data object, but as an epistemically typed knowledge unit, and that can systematically capture its context dependency and transformation dynamics.

3. Conceptual Modeling

Previous research has revealed that existing standards ensure structural interoperability and procedural traceability, but do not explicitly model the epistemic nature of knowledge units in a diagnostic context. To address this gap, this section introduces the ***Ge**nome **Kno**wledge **U**nit (GeKoU)* as a conceptual

modeling unit. The GeKoU regards SNP-related knowledge as an integrated unit that combines semantic content, structural representation, epistemic status, and provenance and version information.

The aim is not to treat knowledge elements in isolation as data fields or annotations, but to model them as epistemically characterized units that can assume different states within a diagnostic workflow.

The Genome Knowledge Unit is introduced as a formally defined, structured, and versioned knowledge unit that models SNP-related knowledge in diagnostic environments. A GeKoU integrates the semantic content of a statement about a variant, its concrete structural representation form, metadata on origin and versioning, and an explicitly designated epistemic status.

A fundamental distinction is made between StuGeKo and UnGeKo. StuGeKo comprises formally representable knowledge. UnGeKo encompasses narrative, text-based knowledge sources such as scientific publications or clinical reports. This differentiation is particularly essential in a regulatory context, as inferred hypotheses do not have the same epistemic status as asserted findings.
Genome knowledge can be further divided into aGeKo and iGeKo. aGeKo comprises curated and verified information, such as that found in databases, standardized formats, or rule-based classification systems. Even knowledge from unstructured sources can become aGeKo after being reviewed by experts. iGeKo refers to knowledge that is derived by machines, whether from text-based sources or data-driven models.

Based on this typology of knowledge, an epistemic metadata model is proposed that describes each unit of knowledge not only by its semantic content, but also by additional constitutive properties.
A GeKoU is formally defined as an object K:

$$K = (c, s, e, v) \tag{1}$$

Here, c denotes the semantic content of the statement, s the degree of structure, e the evidence status and v denotes the version of its originating source. The degree of structure $s \in$ structured, unstructured describes the formal representability of the knowledge unit. The evidence status $e \in$ asserted, inferred differentiates between asserted and model-based inferred origins. The version component v enables temporal and systemic classification and belongs to a source from which the knowledge or data originates.

This explicit modeling makes it possible to systematically distinguish between two units of knowledge with identical content but different epistemic status. This metadata model is fundamentally compatible with interoperable standards such as HL7 FHIR. While FHIR provides resources for the structured mapping of clinical information, the model proposed here supplements these structures with explicit epistemic differentiation.

In this context, the four GeKoU dimensions directly support the principles of TRI-IVIS: semantic content and structural representation enable intelligent processing, evidence status contributes to trustworthiness, and versioning ensures reproducibility.

The preliminary study showed that visualization plays a central role in everyday diagnostics, especially in the interpretation of chromatograms, the verification of genotype calls, and the traceability of validation steps. However, existing visualizations focus primarily on signal or sequence representations and do not explicitly map epistemic properties.

The model proposed here enables classic SNP visualizations to be expanded to include epistemic meta-dimensions. Specifically, this results in the following visualization requirements:

- **R1** Asserted and inferred knowledge should be visually distinguishable.
- **R2** Transitions in evidence status, particularly from inferred to asserted knowledge, should be modeled as explicit state changes.
- **R3** Each GeKoU should interactively display its version and source.

Visualization is thus understood not only as a form of data representation, but also as an epistemic interface for evaluating evidence status and regulatory traceability. To operationalize the GeKoU model within diagnostic workflows, we propose an interactive interface concept that integrates epistemic layers directly into the visual representation of SNP-related knowledge units.

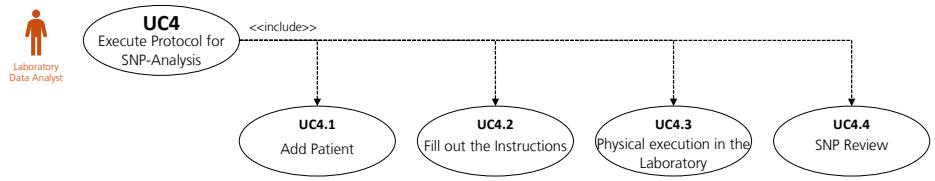

Figure 2: Use Case 4: Execute Protocol for SNP-Analysis

UC4 is the execution of the lab protocol for SNP analyses. This use case can be divided into the Sub-UCs outlined in Figure 2.

First, the patients' data has to be entered into the laboratory information system as part of UC4.1. Several SNP analyses can be requested for a single patient. The patient is assigned an identifier, which enables pseudonymized handling of the data during the SNP analysis. UC4.2 describes the following test planning and documentation. The reaction mixtures, concentrations and PCR programs created for each profile are now supplemented with the identifier, enabling the steps to be documented. UC4.3 covers the practical execution of DNA purification, measurement, amplification by PCR and sequencing. The result is an .ab1 file containing the raw signal from the sequencing. This represents a structured unit of knowledge. The subsequent review of the sequenced data within the framework of UC4.4 is carried out using a representation such as that provided by Chromas. The target position is identified, annotated manually and checked according to the dual control principle. The date, abbreviation and signature document the review and its documentation status.

The GeKoU of UC4.4 (SNP Review) exhibit epistemic heterogeneity. While the raw sequence data (.ab1) and profile identifiers are structured and asserted, the genotype call and review record represent inferred and context-dependent knowledge units. Here, epistemic differentiation becomes particularly relevant, as structured signal data coexists with interpretative elements that are inferred within the laboratory process.

4. Implementation

The GeKoU model is based on an empirical preliminary study of diagnostic practice and specifically addresses the challenges RC1, RC2 and RC3 identified in Section 2. The goal of the implementation is to systematically and verifiably address these challenges in a real-world system environment.

Figure 3 illustrates the prototypical implementation of the GeKoU model in the context of UC4.4 (SNP Review). The sequenced DNA is visualized in the form of a chromatogram, which emerges as structured raw data from a sequencer. By providing the corresponding Genome Knowledge Unit, expert output can be classified and utilized more transparently.

The implementation addresses the identified challenges as follows:
RC1 is addressed by integrating heterogeneous knowledge sources into a unified GeKoU structure, which enables data from different systems (e.g., patient databases, genome databases, literature, model outputs, and laboratory artifacts) to be consistently consolidated and presented together. This remaining challenge can be extended beyond this use case.
RC2 is implemented through the explicit modeling of epistemic properties, which also allows implicit, interpretive knowledge components to be formally captured and made traceable. In particular, a distinction is made between asserted and inferred knowledge, and these are labeled accordingly.
RC3 is realized through the systematic recording and visualization of versions and provenance, so that changes to knowledge artifacts and their origins can be transparently presented and documented in a manner that is traceable for regulatory purposes.

The implementation can be evaluated across several dimensions. It must be examined whether the explicit labeling of epistemic status increases the transparency of diagnostic decision-making processes. The formal distinction between curated, validated knowledge and model-based or interpretatively derived knowledge makes decision-making processes in review or audit scenarios reconstructible and

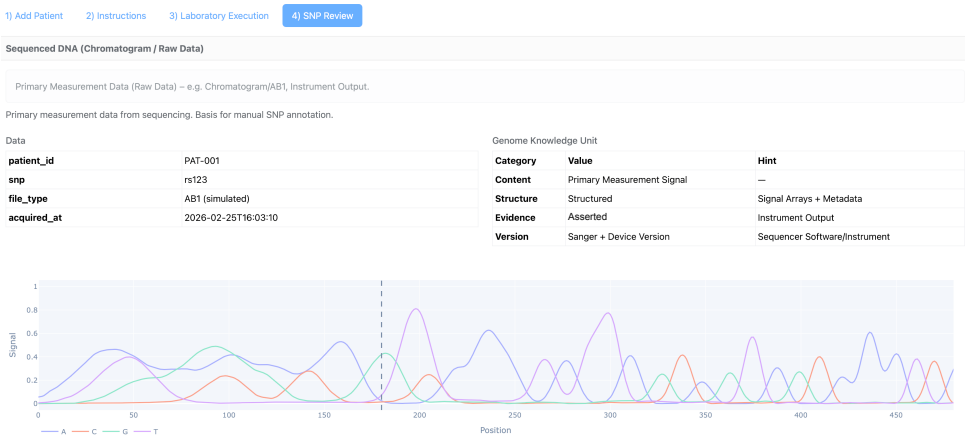

Figure 3: Mock-Up of the GeKoU "Raw Sequence" of UC4.4

systematically traceable.
In regulated diagnostic environments, the seamless traceability of decision-making bases is essential. The explicit modeling and versioning of epistemic transitions can enhance the quality of documentation as well as the reproducibility of diagnostic decisions. The GeKoU model is not intended as a replacement for existing standards such as FHIR or W3C PROV, but rather as a complementary epistemic modeling layer.

The technical implementation is carried out as part of the development of SnipFlow as a microservice architecture within the GenDAI platform. SnipFlow complements this infrastructure with an explicit representation of knowledge by modeling, storing, and versioning analysis results as GeKoU. In accordance with TRI-IVIS, the integration of SnipFlow into GenDAI operationalizes the principles of trustworthy, reproducible, and intelligent information systems.

5. Conclusion

This paper has introduced a conceptual model for the epistemically differentiated representation of SNP-related knowledge, systematically linking diagnostic use cases with structured and inferred knowledge. The focus lies on the formal modeling of the Genome Knowledge Unit (GeKoU), which captures semantic content together with degree of structure, evidence status, and version. By explicitly distinguishing between asserted and inferred knowledge, the model enables transparent declaration, differentiation, and versioning of heterogeneous knowledge units.

The analysis UC4 demonstrates that an SNP does not exist as a static data object within diagnostic workflows, but as a context-dependent constellation of epistemically heterogeneous knowledge units. Making these epistemic differences explicit establishes the formal foundation for regulatory traceability and reproducible decision-making processes.

Beyond conceptual modeling, the framework creates the conditions for epistemically enriched visualization concepts that display not only data, but also evidence status and version histories. The approach is transferable to other regulated fields in which heterogeneous knowledge forms must be transparently integrated and represented context-dependently.

With regard to the research questions, the model contributes to RQ1 and RQ2 by providing a structured framework within which extracted structured and unstructured knowledge can be formally declared and epistemically differentiated. Concerning RQ3, it enables systematic examination of how heterogeneous knowledge units can be combined context-dependently within diagnostic workflows. Regarding RQ4, it offers the conceptual basis for embedding such an epistemically differentiated knowledge model into a knowledge management system such as SnipFlow within GenDAI.

Declaration on Generative AI

The authors used AI-assisted tools, including ChatGPT 5 (OpenAI), DeepL's built-in language suggestions, solely for language polishing, such as improving grammar, rephrasing, and summarization. The authors reviewed and edited all suggestions and take full responsibility for the final content of the manuscript.

References

[1] C. L. Sawyers, L. J. van 't Veer, Reliable and effective diagnostics are keys to accelerating personalized cancer medicine and transforming cancer care: A policy statement from the american association for cancer research, Clinical Cancer Research 20 (2014) 4978–4981. URL: http://dx.doi.org/10.1158/1078-0432.ccr-14-2295. doi:10.1158/1078-0432.ccr-14-2295.

[2] J. Pethke, T. Krause, B. Andrade, M. Kramer, M. Hemmje, Genomic Foundation Models for SNP Analysis , in: 2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2025, pp. 7214–7218.

[3] P. Danecek, A. Auton, G. Abecasis, C. A. Albers, E. Banks, M. A. DePristo, R. E. Handsaker, G. Lunter, G. T. Marth, S. T. Sherry, G. McVean, R. Durbin, 1000 Genomes Project Analysis Group, The variant call format and VCFtools, Bioinformatics 27 (2011) 2156–2158.

[4] R. K. Hart, I. F. A. C. Fokkema, M. DiStefano, R. Hastings, J. F. J. Laros, R. Taylor, A. H. Wagner, J. T. den Dunnen, HGVS nomenclature 2024: improvements to community engagement, usability, and computability, Genome Med. 16 (2024) 149.

[5] L. Phan, H. Zhang, Q. Wang, R. Villamarin, T. Hefferon, A. Ramanathan, B. Kattman, The evolution of dbsnp: 25 years of impact in genomic research., Nucleic Acids Res 53 (2025) D925–D931.

[6] M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J. B. Holmes, B. L. Kattman, D. R. Maglott, Clinvar: improving access to variant interpretations and supporting evidence, Nucleic Acids Research 46 (2017) D1062–D1067.

[7] HL7 International, Hl7 fhir standard, https://hl7.org/fhir/, 2023. Accessed: 2026-02-21.

[8] P. Missier, K. Belhajjame, J. Cheney, The w3c prov family of specifications for modelling provenance metadata, Association for Computing Machinery, New York, NY, USA, 2013, p. 773–776. URL: https://doi.org/10.1145/2452376.2452478. doi:10.1145/2452376.2452478.

[9] S. Richards, N. Aziz, S. Bale, D. Bick, S. Das, J. Gastier-Foster, W. W. Grody, M. Hegde, E. Lyon, E. Spector, K. Voelkerding, H. L. Rehm, 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 17 (2015) 405–424.

[10] Z. Li, V. Subasri, Y. Shen, D. Li, Y. Zhao, G.-B. Stan, C. Shan, Omni-dna: A unified genomic foundation model for cross-modal and multi-task learning, 2025. URL: https://arxiv.org/abs/2502.03499. arXiv:2502.03499.

[11] C. Ferguson, D. Araújo, L. Faulk, Y. Gou, A. Hamelers, Z. Huang, M. Ide-Smith, M. Levchenko, N. Marinos, R. Nambiar, M. Nassar, M. Parkin, X. Pi, F. Rahman, F. Rogers, Y. Roochun, S. Saha, M. Selim, Z. Shafique, S. Sharma, D. Stephenson, F. Talo', A. Thouvenin, S. Tirunagari, V. Vartak, A. Venkatesan, X. Yang, J. McEntyre, Europe PMC in 2020, Nucleic Acids Res. 49 (2021).

[12] A. Allot, C.-H. Wei, L. Phan, T. Hefferon, M. Landrum, H. L. Rehm, Z. Lu, Tracking genetic variants in the biomedical literature using litvar 2.0, Nature Genetics 55 (2023) 901–903.

[13] N. I. of Health (NIH), About Gene RIF - Gene - NCBI — ncbi.nlm.nih.gov, https://www.ncbi.nlm.nih.gov/gene/about-generif, 2025. [Accessed 07-12-2025].

[14] C.-H. Wei, A. Allot, P.-T. Lai, R. Leaman, S. Tian, L. Luo, Q. Jin, Z. Wang, Q. Chen, Z. Lu, Pubtator 3.0: an ai-powered literature resource for unlocking biomedical knowledge, Nucleic Acids Research 52 (2024).

[15] C. Peng, F. Xia, M. Naseriparsa, F. Osborne, Knowledge graphs: Opportunities and challenges (2023).
[16] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, G. Sherlock, Gene ontology: tool for the unification of biology. the gene ontology consortium, Nat. Genet. 25 (2000) 25–29.
[17] M. A. Gargano, N. Matentzoglu, B. Coleman, E. B. Addo-Lartey, et al., The human phenotype ontology in 2024: phenotypes around the world, Nucleic Acids Res. 52 (2024) D1333–D1346.
[18] T. Krause, Empowering medical laboratories with automated ai and big data approaches to genomics-based diagnostics, 2025.
[19] T. Reis, Empowering users in visual big data analysis by applying artificial intelligence and machine learning, 2024.