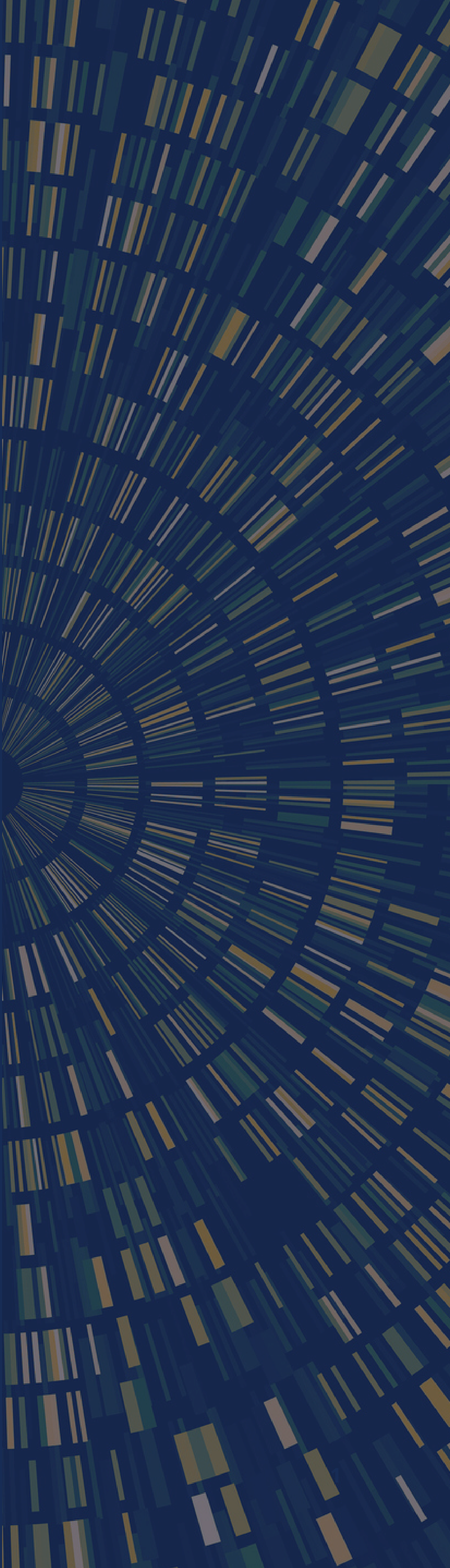
varsomeapi

WHITE PAPER

APIs for Reproducible Clinical Genomics



CONTENTS

Introduction

Genomics has entered an era of abundance. Next-generation sequencing (NGS) has become faster and more affordable, enabling laboratories and research programs worldwide to generate new datasets daily. **But this abundance comes with a cost.** Data is scattered across hospitals, research institutes, and national repositories; each using their own formats, pipelines, and access methods.

For clinical and translational scientists, this fragmentation creates real barriers. **Analysts spend time writing glue code just to move data between systems.** A variant interpreted as pathogenic in one laboratory may be reported as uncertain in another, not because of scientific disagreement, but because each group applied different databases, filters, or pipeline versions.

For regulators, reproducibility is a baseline requirement. Confidence in the system depends on being able to show that results are consistent and auditable, regardless of where or when they are produced. **For patients, reproducibility is less visible but no less important.** Most will never encounter pipelines themselves, but they feel the consequences when different laboratories return conflicting results or when inconsistent classifications prolong the diagnostic odyssey. A lack of reproducibility means more uncertainty, more referrals, and more delays in receiving clear answers. If genomic results are not reproducible, how can they be trusted to guide care?

The challenge of vastly distributed genomic datasets is one of the defining problems in modern bioinformatics. **Solving it requires more than faster sequencers or bigger, more diverse datasets.** It demands a common language for how data is accessed, workflows are run, and results are interpreted.

This is where standards, **Application Programming Interfaces (APIs),** and reproducible pipelines come in. The genomics community, led by the **Global Alliance for Genomics and Health (GA4GH)** and regional infrastructures such as ELIXIR, All of Us, and CanDIG, has converged on shared specifications for data access and workflow execution. Meanwhile, laboratories are embracing community pipelines like nf-core and Galaxy to make their analysis portable and auditable.

But even with these advances, one crucial bottleneck remains: variant interpretation. **VarSome API provides a practical, proven solution.** By delivering **a reproducible, programmatic interface to variant annotation and classification,** it ensures that distributed pipelines converge on the same evidence and interpretation framework, no matter where they are run.

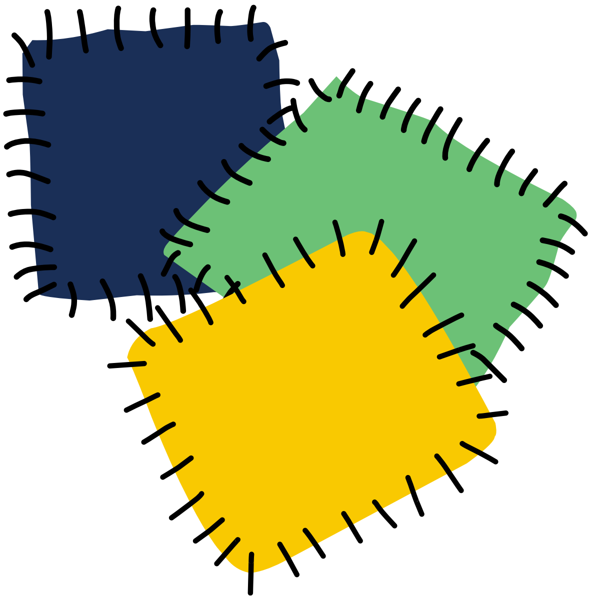
The Reproducibility Problem

Genomic data does not live in one place. Every hospital, national genome program, and research consortium generates its own datasets and stores them under its own rules. Some are held on institutional servers, others in national repositories, and an increasing number in commercial cloud environments. **Each system exposes its data differently.** Some with bespoke APIs, others with simple file downloads, and many with access restrictions tailored to local governance.

For bioinformaticians, **this patchwork creates a daily burden.** Analysts must write glue code to connect pipelines to each new repository. Scripts that work in one environment often break in another, because of small differences in file formats, access methods, or metadata conventions. For international collaborations, the problem compounds: two groups may analyze the same dataset but arrive at subtly different variant lists or classifications, simply because their pipelines interact with data differently from one another.

This creates **a reproducibility gap.** A variant deemed clinically significant in one lab may be flagged as uncertain in another, not due to biology but because of infrastructural inconsistency. For research consortia, this slows discovery and complicates data-sharing. For clinical laboratories, it increases turnaround times, introduces compliance risks, and **ultimately delays answers for patients.**

This fragmentation is now one of the central challenges in precision medicine. It's a drain on analyst time, drives variability between institutions, and raises barriers to scaling multi-center or population-scale genetics. Solving this problem requires **shared standards for data access** and pipelines that are both **portable and auditable.**



Standards to Address the Problem

Over the past decade, large-scale programs and international consortia have converged on a set of shared technical standards designed to make data and workflows interoperable. The GA4GH has been at the center of this effort, defining specifications that allow data to be discovered, accessed, and analyzed consistently.

These standards give datasets and workflows a common language. Instead of every repository exposing its own bespoke interface, a GA4GH-compliant service provides predictable endpoints: a **Data Repository Service (DRS)** identifier retrieves the right file regardless of where it is stored; a **Beacon query** asks whether a dataset contains a variant without exposing sensitive details; an **htsget request** streams only the relevant regions of a BAM or VCF rather than forcing full file transfers.

Standards also extend to analysis. The **Workflow Execution Service** defines how pipelines can be submitted consistently to different compute environments, while the **Tool Registry Service (TRS)** provides a place for laboratories to publish versioned workflows. Reproducibility is further strengthened by the use of portable workflow languages such as CWL, WDL, and Nextflow, together with container technologies like Docker or Singularity. Registries such as Dockstore and WorkflowHub make these workflows discoverable and versioned, so laboratories can adopt proven pipelines rather than starting from scratch. Together, this helps reduce the need for custom code, making it easier for a pipeline built in one setting to be reused in another.

These efforts also align closely with the FAIR principles; **data should be *Findable, Accessible, Interoperable*, and *Reusable*.** FAIR provides the conceptual framework for responsible data stewardship, while GA4GH standards supply the technical means to deliver it in genomics. Beacon makes datasets findable, DRS and htsget provide controlled access, the Workflow Execution Service and TRS support interoperability, and registries such as Dockstore and WorkflowHub ensure reusability through versioning and provenance. Together, FAIR and GA4GH reinforce each other, strengthening the foundation for reproducible and collaborative genomic analysis.

Importantly, adoption is global. In Europe, ELIXIR and the European Genome-phenome Archive have implemented multiple GA4GH standards, including APIs such as htsget, Beacon, and TRS, to support cross-border studies. In the USA, the All of Us program makes use of the same tools to enable secure access to its cohort data. Canada’s CanDIG project, Australian Genomics, and many others around the world are also adopting GA4GH standards to support federated data sharing.

These initiatives provide the basis for a more connected ecosystem. They create **a foundation on which reproducible, collaborative, and secure analysis can be built.** But even with this shared infrastructure, one critical piece remains stubbornly inconsistent.

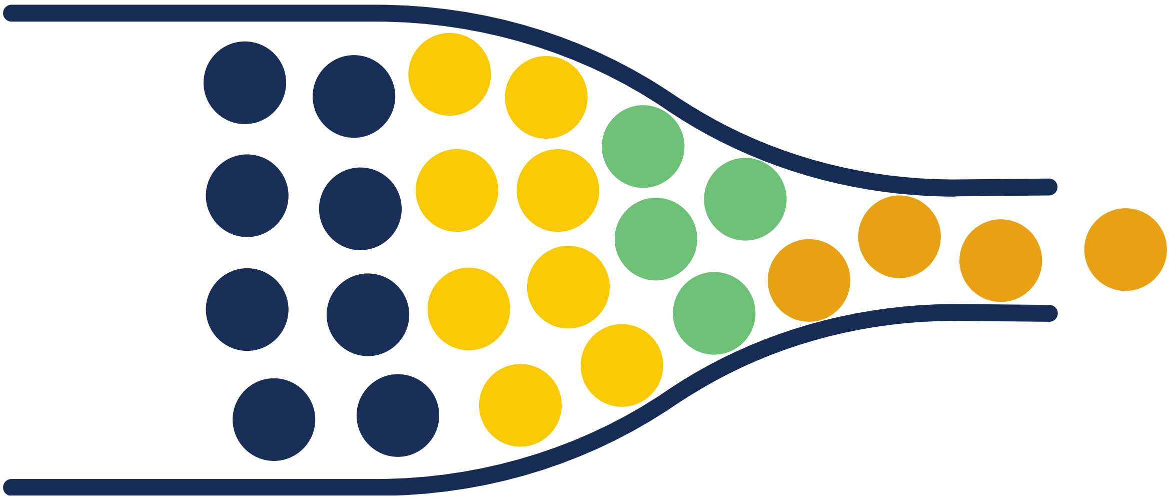
The Interpretation Bottleneck

Variant interpretation, at its simplest, is pulling together evidence from population databases, functional annotations, *in silico* predictions, literature, and established guidelines to determine the clinical significance of a genetic finding. However, no two laboratories will draw on all this evidence in exactly the same way. Some maintain local annotation databases, others rely on a patchwork of public sources, and many adapt ACMG criteria differently.

Working with public databases also brings technical challenges. Updates are released on different schedules, and data structures sometimes change without or with very little warning. This can break annotations and interpretation pipelines, particularly in laboratories with limited technical resources, where errors may only be detected late or after results have been reported.

This inconsistency has direct consequences. For patients, it **risks confusion and delay.** For laboratories, it can **undermine efficiency and compliance,** and for collaborative research, it creates noise that complicates efforts to aggregate results across centers.

An API-based approach offers a solution. By centralizing annotation and classification logic, it ensures that all users draw on the same evidence base and rule set. **VarSome API's MolecularDB is powered by over 140 data sources**, covering population frequencies, clinical databases, functional predictions, and scientific literature. **Users can decide how much or how little data they receive,** tailoring outputs to their workflow needs.



The VarSome API provides programmatic access to germline and somatic classifiers, based on American College of Medical Genetics and Genomics (ACMG) and Association for Molecular Pathology (AMP) guidelines. These classifiers are **designed to support both rare disease diagnostics and precision oncology applications.** The API also provides access to additional tools – gene- and region-level context, *in silico* predictions, and transparent rule explanations – so results are reproducible and ready for expert review.

To meet different operational needs, VarSome offers **Stable, Live, and Staging API environments.**

- **Stable:** A frozen, production-grade release, updated quarterly. It ensures reproducibility for regulatory use but may lag behind Live data.
- **Live:** The continuously updated environment with the most current data and algorithms, including support for community contributions.
- **Staging:** A throttled test environment reflecting the next Stable release. Used for validation, it may include only partial data.

All results are returned in structured JSON, making them easy to integrate into bioinformatic pipelines at scale.

In this way, the VarSome API helps address the last major gap in reproducibility. **Data can be accessed consistently, workflows can be run portably, and with VarSome, the results can be interpreted against a consistent, transparent framework**, whether the analysis takes place in a hospital, a research institute, or a national genome program.

varsomeapi

APIs can be powerful enablers of consistency, but like any tool, they need careful governance to deliver their full value.

Case Studies From Literature

Independent groups have integrated the VarSome API into their pipelines and published the results, showing how a centralized interpretation service can improve consistency and performance across different contexts and sequencing technologies.

Semi-automated ACMG Interpretation in Diagnostic Pipelines

Sorrentino *et al.* (2021) incorporated the VarSome API into a validated diagnostic NGS workflow. By combining automated ACMG rule evaluation with expert review, the pipeline achieved 100% sensitivity, specificity, and accuracy for variant classification in a clinical setting. This paper illustrates how the VarSome API can provide reproducible interpretation data while still allowing space for human expertise.

Long-read Sequencing with PacMAGI

Sorrentino *et al.* (2022) extended this work to long-read sequencing, integrating the VarSome API into the PacMAGI pipeline for variant detection in the *RPE65* gene. The pipeline successfully identified both known and novel pathogenic variants, highlighting the API's flexibility. Here, VarSome API helped ensure that the same evidence sources and ACMG-based framework could be applied consistently, regardless of sequencing technology.

Modified ACMG Classifier for Diagnostic Consistency

Cristofoli *et al.* (2021) used the VarSome API to create a modified classifier that adapted ACMG rules to their local diagnostic practices. By blending automation with manual curation, the team improved classification consistency across their workflow while retaining the transparency needed for clinical decision-making.

Together, these studies show that the VarSome API can be applied in a variety of research and diagnostic settings. They highlight its use not only with short-read sequencing, but also with long-read technologies and customized workflows that balance automation with expert review.

Challenges & Best Practices

For laboratories integrating APIs into their analysis pipelines, there are several considerations that shape both reproducibility and long-term sustainability.

Versioning & Stability

Annotation sources are constantly updated, making careful version control crucial. Best practice is to fix workflows to a specific database version or use an environment such as the Stable VarSome API, which provides the consistency required for regulatory use.

Managing Updates

While stability is essential, access to the latest knowledge is equally valuable, especially in research contexts. Organizations may choose to use a Live environment for exploratory use and a Stable environment for validated pipelines.

Cost & Performance

Annotation can involve large volumes of variants. Batching queries, caching results, and filtering input VCFs before annotation help control cost and reduce latency. These practices also minimize redundant calls when the same variants appear across multiple samples.

Balancing Automation with Expertise

Automated classifiers accelerate workflows, but human expertise is vital, particularly for variants of uncertain significance or borderline cases. Combining reproducible API calls with structured manual review ensures consistency without losing clinical judgment.

Security & Privacy

Genomic data is sensitive and any integration must respect patient confidentiality. Secure authentication, encryption in transit, and role-based access controls are now baseline requirements. Aligning with GA4GH’s passport and authentication frameworks helps ensure compatibility with emerging infrastructures.

Importantly, adoption is global. ELIXER, All of Us, CanDIG and many others around the world are adopting GA4GH standards to support federated data sharing.

Practical Outcomes

When data access, workflows, and interpretation are aligned, the benefits extend across research, clinical, and collaborative settings.

Efficiency & Turnaround Time

Automated annotation via API reduces the manual effort needed for each case. Analysts can focus their time where expert judgment is most valuable, rather than on routine analysis. This translates to fast reporting for patients and quick iterations in research.

Scalability

As sample volumes grow, pipelines that rely on reproducible workflows and programmatic annotation can scale without being redesigned. From individual hospitals to population-scale initiatives, the model scales without structural changes.

Compliance Readiness

Reproducibility and auditability are regulatory expectations. Versioned workflows and stable annotation environments support accreditation under frameworks such as IVDR, CLIA, and ISO 15189.

Global Interoperability

By aligning with GA4GH standards, these pipelines can function across borders. A workflow developed in one country can be reused in another with minimal adaptation, and results can be compared on common terms.

These outcomes help move genomics closer to a world where results are both clinically meaningful and consistently reproducible, regardless of where or how the data was generated.

Future Directions

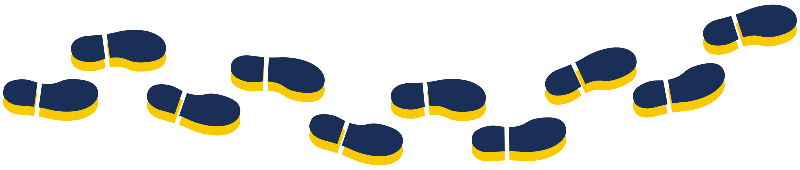
One of the most significant shifts on the horizon is the move to federated analysis. As genomic datasets grow in size and are subject to stricter governance, the traditional approach of copying files between institutions becomes increasingly impractical. Instead, the emerging model is to run the analysis where the data already resides, and only share back the results. This approach, often described as **bringing compute to the data**, means laboratories can send queries or workflows to a remote repository, which executes the analysis locally. APIs such as Beacon v2 and DRS make this possible, allowing researchers to query datasets and run analyses without moving sensitive information across borders. This helps reduce costs, respects data sovereignty, and enables collaboration on a scale that would have been unmanageable only a few years ago.

Efficiency will increasingly depend on streaming access to data rather than wholesale downloads. With technologies such as htsget, workflows can request only the genomic regions they need, **minimizing bandwidth and storage requirements**. As projects grow to population scale, this selective access will become important to keeping pipelines practical and responsive.

The rapid advance of artificial intelligence (AI) and machine learning is another important factor. Models are already being developed to **improve variant prioritization and predict functional impact.** For these approaches to be trusted in clinical contexts, however, they must be embedded within **reproducible frameworks and accompanied by transparent audit trails.** APIs provide a way to get AI-derived insights in a controlled, versioned manner, ensuring that machine learning augments interpretation without undermining traceability.

Finally, regulatory alignment will continue to shape how laboratories develop their pipelines. Frameworks such as IVDR, CLIA, and ISO 15189 are all **tightening expectations around reproducibility and auditability.** Stable, versioned workflows and annotation environments are seen more and more not only as best practices, but as regulatory prerequisites.

These trends point to a future in which reproducibility is treated as an integral part of genomic analysis. Standards for access, workflows, and interpretation will continue to mature, enabling results that are **portable, audible, and trusted** across institutions and national boundaries.

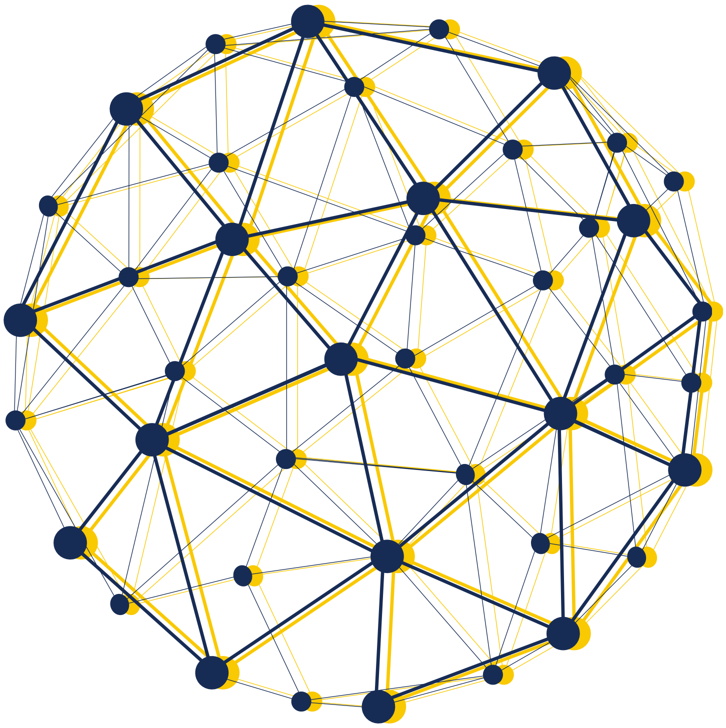


Conclusions

Genomics has advanced rapidly, but the challenge of distributed data and fragmented workflows continues to shape how results are generated and shared. Over the past decade, the community has responded by establishing a shared foundation of standards for data access and workflow execution. These efforts have helped to bring the field closer to **a connected and reproducible ecosystem.** But one stage of the pipeline remains resistant to standardization. Without a consistent approach to variant interpretation, laboratories risk reaching different conclusions.

By centralizing access to **over 140 data sources, applying ACMG/AMP-based germline and somatic classifiers, and offering reproducible environments tailored to different needs**, VarSome API helps to make variant interpretation consistent and auditable. The publications highlighted here show that it is not just a theoretical solution but one already being applied in diagnostics and research.

As genomic medicine scales to larger cohorts and tighter regulatory frameworks, **reproducibility will only grow in importance.** Standards and workflows continue to mature, but interpretation methods must keep pace. By integrating the VarSome API, laboratories can ensure that their pipelines are efficient, scalable, and aligned with the highest expectations of consistency, transparency, and clinical reliability.



References

1. Sorrentino E, Cristofoli F, Modena C, Paolacci S, Bertelli M, Marceddu G. Integration of VarSome API in an existing bioinformatic pipeline for automated ACMG interpretation of clinical variants. *Eur Rev Med Pharmacol Sci*. 2021;25(1 Suppl):1-6. doi:10.26355/eurrev_202112_27325

2. Sorrentino E, Albion E, Modena C, et al. PacMAGI: A pipeline including accurate indel detection for the analysis of PacBio sequencing data applied to RPE65. *Gene*. 2022;832:146554. doi:10.1016/j.gene.2022.146554

3. Cristofoli F, Sorrentino E, Guerri G, et al. Variant Selection and Interpretation: An Example of Modified VarSome Classifier of ACMG Guidelines in the Diagnostic Setting. *Genes*. 2021;12(12):1885. doi:10.3390/genes12121885

varsomeapi

If you'd like to learn more about the VarSome API, or any of our other products or services, visit our website or reach out to us at sales@varsome.com.

varsomeclinical varsomepremium varsomeinsights

VSA-WP-001-V01

Built with Swiss precision by

SAPHETOR

EPFL Innovation Park - C
1015 Lausanne, Switzerland