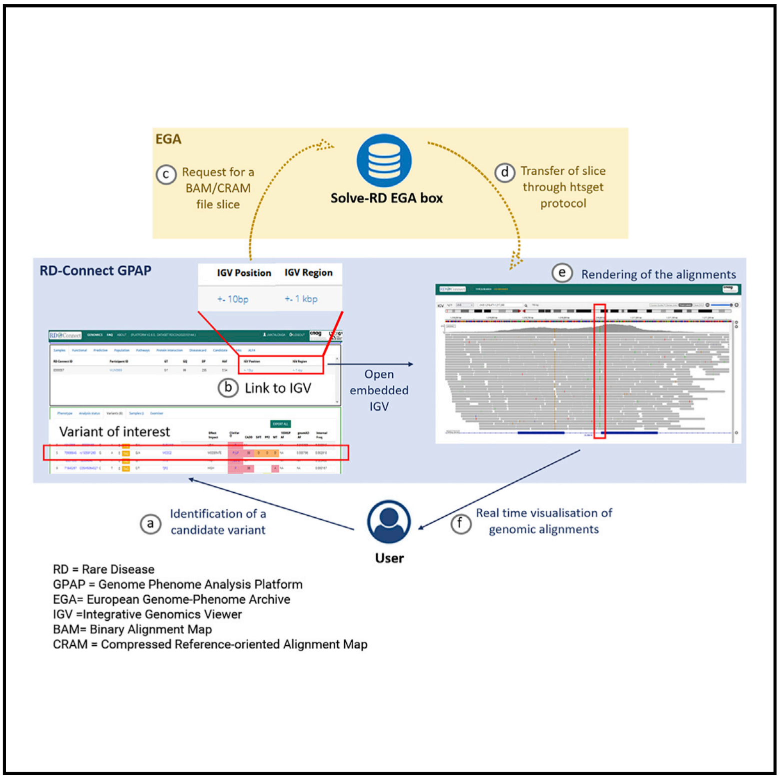
Cell Genomics

Technology

Remote visualization of large-scale genomic alignments for collaborative clinical research and diagnosis of rare diseases

Graphical abstract


EGA
c Request for a BAM/CRAM file slice
Solve-RD EGA box
d Transfer of slice through htsget protocol
RD-Connect GPAP
IGV Position
IGV Region
+- 10bp
+- 1 kbp
b Link to IGV
Open embedded IGV
e Rendering of the alignments
Variant of interest
a Identification of a candidate variant
User
f Real time visualisation of genomic alignments
RD = Rare Disease
GPAP = Genome Phenome Analysis Platform
EGA= European Genome-Phenome Archive
IGV =Integrative Genomics Viewer
BAM= Binary Alignment Map
CRAM = Compressed Reference-oriented Alignment Map

Highlights

- Visualization of sequence alignments enables the evaluation of genomic variants
- RD-Connect GPAP users can visualize alignments at the EGA through GA4GH htsget
- Remote visualization in real time avoids recurrent large data downloads
- Hundreds of rare disease-causing variants have been inspected in project Solve-RD

Authors

Alberto Corvò, Leslie Matalonga, Dylan Spalding, ..., Thomas Keane, Davide Piscia, Sergi Beltran

Correspondence

sergi.beltran@cnag.crg.eu

In brief

Corvò et al. present a GA4GH htsget implementation enabling RD-Connect GPAP users to remotely visualize genomic alignments at the EGA in real time, avoiding the need for large downloads. This has proven essential to evaluate potentially rare disease-causing genetic variants in hundreds of patients from the Solve-RD project.

Corvò et al., 2023, Cell Genomics 3, 100246
February 8, 2023 © 2023 The Authors.
https://doi.org/10.1016/j.xgen.2022.100246

CellPress

Cell Genomics

CellPress
OPEN ACCESS

Technology

Remote visualization of large-scale genomic alignments for collaborative clinical research and diagnosis of rare diseases

Alberto Corvò,[1,10] Leslie Matalonga,[1,10] Dylan Spalding,[2,3] Alexander Senf,[2,4] Steven Laurie,[1] Daniel Picó-Amador,[1] Marcos Fernandez-Callejo,[1] Ida Paramonov,[1] Anna Foix Romero,[2] Emilio Garcia-Rios,[2] Jorge Izquierdo Ciges,[2] Anand Mohan,[2] Coline Thomas,[2] Andres Felipe Silva Valencia,[2] Csaba Halmagyi,[2] Mallory Ann Freeberg,[2] Ana Töpf,[5] Rita Horvath,[6] Gary Saunders,[7] Ivo Gut,[1] Thomas Keane,[2] Davide Piscia,[1] and Sergi Beltran[1,8,9,11,*]

[1]CNAG-CRG, Centre for Genomic Regulation (CRG), The Barcelona Institute of Science and Technology, Baldiri Reixac 4, Barcelona 08028, Spain
[2]European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Hinxton, UK
[3]CSC, Espoo, Finland
[4]AI-Digital, Lincoln, UK
[5]John Walton Muscular Dystrophy Research Centre, Translational and Clinical Research Institute, Newcastle University and Newcastle Hospitals NHS Foundation Trust, Newcastle upon Tyne, UK
[6]Department of Clinical Neurosciences, John Van Geest Centre for Brain Repair, School of Clinical Medicine, University of Cambridge, Cambridge Biomedical Campus, Cambridge, UK
[7]ELIXIR Hub, Wellcome Genome Campus, Hinxton, Cambridge, UK
[8]Universitat Pompeu Fabra (UPF), Barcelona, Spain
[9]Departament de Genètica, Microbiologia i Estadística, Facultat de Biologia, Universitat de Barcelona (UB), 08028 Barcelona, Spain
[10]These authors contributed equally
[11]Lead contact
*Correspondence: sergi.beltran@cnag.crg.eu

https://doi.org/10.1016/j.xgen.2022.100246

SUMMARY

The Solve-RD project objectives include solving undiagnosed rare diseases (RD) through collaborative research on shared genome-phenome datasets. The RD-Connect Genome-Phenome Analysis Platform (GPAP), for data collation and analysis, and the European Genome-Phenome Archive (EGA), for file storage, are two key components of the Solve-RD infrastructure. Clinical researchers can identify candidate genetic variants within the RD-Connect GPAP and, thanks to the developments presented here as part of joint ELIXIR activities, are able to remotely visualize the corresponding alignments stored at the EGA. The Global Alliance for Genomics and Health (GA4GH) htsget streaming application programming interface (API) is used to retrieve alignment slices, which are rendered by an integrated genome viewer (IGV) instance embedded in the GPAP. As a result, it is no longer necessary for over 11,000 datasets to download large alignment files to visualize them locally. This work highlights the advantages, from both the user and infrastructure perspectives, of implementing interoperability standards for establishing federated genomics data networks.

INTRODUCTION

Next-generation sequencing (NGS) permits cost-effective sequencing of exomes and genomes, facilitating clinical research and diagnostics[1] beyond what was possible a few years ago. Still, these technologies only enable a diagnostic yield for patients with a suspected monogenic rare disease (RD) of between 20% and 60%, depending on the underlying disease.[2–4] Further research and development, coupled with data sharing, offer additional opportunities to patients with inconclusive exomes or genomes.

Solve-RD (http://solve-rd.eu/) is a pan-European initiative that aims to reveal the molecular cause underlying undiagnosed RDs.[5] One of the main goals of the project is to comprehensively reanalyze more than 19,000 inconclusive exomes and genomes from undiagnosed patients submitted by partnering European Reference Networks (ERNs; https://ec.europa.eu/health/ern_en) and undiagnosed disease programs from Spain and Italy. Therefore, one of the main challenges facing Solve-RD is the ability to effectively collect, store, process, share, and interpret vast quantities of data, provided by more than 51 different centers across Europe, within a secure and collaborative environment.

To address this challenge, the Solve-RD infrastructure is built upon existing resources and technologies, such as the European Genome-Phenome Archive (EGA)[6] and the RD-Connect Genome-Phenome Analysis Platform (GPAP),[7] each funded by several European and national projects. The EGA and the RD-Connect GPAP are supported by ELIXIR (https://elixir-europe.org/), the intergovernmental organization that brings together life science resources from across Europe, and are also key for the European Joint Program of Rare Diseases (https://www.ejprarediseases.org/), which involves all ERNs. Solve-RD uses the RD-Connect GPAP as the phenotypic and genomic data entry point and the EGA for permanent data storage and re-use.

The EGA and the RD-Connect GPAP are both open to registration of clinical scientists outside the Solve-RD project. Registered users are able to submit and access data, providing clinical scientists with functionalities to share and analyze integrated genomic and phenotypic data from patients with an RD with the objective to identify causative genetic variants from undiagnosed patients or discover new gene-disease associations.

The EGA (https://ega-archive.org/) is a service for permanent archiving and sharing of personally identifiable genetic and phenotypic data resulting from biomedical research projects. The EGA provides access for approved researchers to foster data re-use, enable reproducibility, and accelerate biomedical and translational research in line with the Findable, Accessible, Interoperable, and Reusable (FAIR) principles.[8] The EGA has archived over 4,500 studies representing genetic, phenotypic, and clinical data from a variety of research fields, including over 900 studies across 200 RDs. RD data are submitted to the EGA by individual labs researching a particular disease, as well as by larger RD initiatives such as Solve-RD, the National Institute for Health and Care Research (NIHR), and the BioResource Rare Diseases BRIDGE consortium (https://bioresource.nihr.ac.uk/), a collaboration between 13 rare disease projects that aims to discover the genetic cause underlying unresolved inherited disorders. Authorized researchers can access and download data from the EGA to their computers or to secure data analysis platforms. Our work circumvents the need to download data by also enabling remote access to specific regions of genome alignment files.

The RD-Connect GPAP (https://platform.rd-connect.eu) is an International Rare Diseases Research Consortium (IRDiRC)-recognized resource and privacy-preserving environment for secure data analysis. The platform provides clinical scientists with a framework to process, analyze, and share integrated sequencing and phenotypic data from patients with an RD, and their relatives, through a powerful and user-friendly interface for diagnosis and gene discovery. The RD-Connect GPAP enables users to filter and prioritize genetic variants from exomes and genomes based on sequencing coverage and variant quality, inheritance model, known effects, predicted pathogenicity, population frequency, disease and phenotype associations, etc. Connection to multiple external web services and links to a large number of websites facilitate the interpretation of the filtered variants. The platform also enables researchers to look for additional patients with the same disease or candidate variants in the same genes, allowing the researchers to get in contact and share further details. The RD-Connect GPAP is also connected to the MatchMaker Exchange network,[9] which allows users registered in only one of the nodes to look for the availability of patients with similar phenotypes and candidate genes in other nodes of the network.

The RD-Connect GPAP currently hosts phenotypic and processed genomic data (annotated genetic variants) from more than 26,500 patients and relatives submitted by authorized users. Genomic data are submitted in the most common file formats: binary alignment map (BAM), compressed reference-oriented alignment map (CRAM), or FASTQ, a text-based format. The RD-Connect GPAP does not store the genomic alignments (BAM/CRAM) on its online servers since these files are very large and other services, such as the EGA, are better suited for this purpose. Upon submitter authorization, the RD-Connect GPAP may transfer the corresponding raw alignment and genetic variant files (FASTQ and/or BAM/CRAM and variant call format [VCF]), the phenotypic information, and the files' metadata to the EGA for long-term controlled access and data re-use. However, it is essential for clinical scientists to have access to genomic alignments to visualize regions around candidate disease-causing variants. Although the algorithms identifying genetic variants are generally highly accurate, there are many regions of the genome where it remains challenging to detect variants from NGS data.[10] This is especially true in repetitive regions of the genome where properly identifying short insertions, deletions, and copy-number variants is more challenging; indeed, more than 400 medically relevant genes have been shown to be affected by such issues.[11] Despite this, even the untrained eye can instantly identify if those variants are unreliable upon the inspection of the region in the alignments themselves. However, having to download/transfer gigabytes of data from a remote source each time a researcher wants to visualize a specific genomic locus is neither efficient nor scalable. For this reason, the Global Alliance for Genomics and Health (GA4GH) developed the htsget streaming application programming interface (API), which allows users to request and receive genomic data in standardized, secure, and real-time streaming.[12]

DESIGN

To enable the users of the RD-Connect GPAP to visualize genomic alignments from data archived at the EGA, we have developed a scalable system in the context of ELIXIR, including the ELIXIR Rare Disease Community (https://elixir-europe.org/communities/rare-diseases), the ELIXIR Federated Human Data community (https://elixir-europe.org/communities/human-data), and Solve-RD. The system uses the GA4GH htsget streaming API[12] to request and retrieve genomic data archived at the EGA, and an instance of the integrative genomics viewer (IGV)[13] embedded in the RD-Connect GPAP, to render the alignments. The overall implementation currently enables RD-Connect users to remotely access and visualize 11,750 datasets from patients and relatives submitted through the Solve-RD project, which has identified causative variants in hundreds of cases with an RD.[5,14–18]

Data submission and storage at the EGA

Authorized users submit raw genomic data (FASTQ) to the RD-Connect GPAP (https://platform.rd-connect.eu/). As required by the RD-Connect GPAP code of conduct and the Solve-RD data sharing policy, the data submitter is responsible for checking that the data are suitable for submission to the RD-Connect GPAP and Solve-RD. Data are processed using a standardized analysis pipeline[9] and are made available for downstream analysis and interpretation within the RD-Connect GPAP. Upon user authorization, raw unmapped files (FASTQ and/or BAM/CRAM) together with their corresponding index file, analysis-ready alignments (BAM/CRAM), and corresponding indices (BAI and CRAI files), phenotypic information (Phenopackets, http://phenopackets.org/), genetic variant files (VCF), pedigree files (PED), and metadata (CSV) are encrypted using EGACryptor (https://ega-archive.org/submission/tools/egacryptor). Encrypted files are uploaded to the EGA through Aspera servers (https://ega-archive.org/submission/tools/ftp-aspera) into a specific RD-Connect GPAP box following standard protocols (https://ega-archive.org/submission/quickguide). The corresponding metadata are submitted through standard mechanisms using the EGA submitter portal (https://ega-archive.org/submission/tools/submitter-portal). Creation of the RD-Connect GPAP box was approved by the RD-Connect data access committee (DAC). Data submitted to the EGA are archived permanently for controlled access and re-use.

Secure access of data archived at the EGA

Registration to the RD-Connect GPAP is regulated by a DAC and a code of conduct (https://platform.rd-connect.eu/gpap_doc/). Data access is restricted to registered users, and authentication is managed by an open-source identity and access management solution that uses the OpenID Connect protocol (KeyCloak, https://www.keycloak.org/). The same permissions from the RD-Connect GPAP are applied to each user when requesting access to visualize data from an experiment archived at the EGA. A hypertext transfer protocol (HTTP) GET request from the RD-Connect GPAP is then sent to an internal Python-Flask micro service that verifies the credentials of the user and the permissions to access the requested file. If permission is granted, the system sends the requested information (file identifier and genomic interval) to the EGA server.

Slicing and transferring data archived at the EGA

Real-time genomic data slicing from files archived at the EGA is achieved via the GA4GH htsget API implementation (http://samtools.github.io/hts-specs/htsget.html) of the EGA data download service. This protocol uses industry security standard OAuth 2.0 tokens to securely authorize data requests (https://tools.ietf.org/html/rfc6749). The underlying API of the GA4GH htsget is a set of core micro services developed and maintained by the EGA, and the service has been in production since 2019. Through the htsget protocol, the requested data are returned in binary format, which are sent by the Python-Flask API to the RD-Connect GPAP client.

Genomic data visualization and exploration in the RD-Connect GPAP

The RD-Connect GPAP platform is built as a server-client architecture that users can access to submit clinical data or experiment and file metadata or to analyze participants' integrated data with a powerful user interface. The JavaScript library from the IGV (IGV.js, MIT, MA, USA) is embedded into the RD-Connect GPAP environment to enable visualization of genomic alignments (https://platform.rd-connect.eu/igvgpap/). IGV.js includes the main options to visualize, explore, and navigate genomic regions.[19] The IGV browser reads the binary BAM file. The read alignments are rendered in the dedicated module and mapped to genes according to the Reference Sequence (RefSeq) database (https://www.ncbi.nlm.nih.gov/refseq/). The system currently supports GRCh37 as the reference genome but could work with any version of the genome. Alignments can be explored by scrolling the view horizontally and vertically. Changing the pre-selected genomic interval triggers a new HTTP GET request, as described above.

RESULTS

Genomic data workflow, storage, and permission access

The submission and archival of genomic alignments (BAM and/or CRAM files) was organized according to the EGA standard protocols (https://ega-archive.org/submission/quickguide). User authentication to access the alignments submitted by the RD-Connect GPAP to the EGA is done through a single EGA user profile. Therefore, the RD-Connect GPAP controls who can access the alignments at the EGA based on the code of conduct and the data access policies of the former. Users' authentication in the RD-Connect GPAP is managed by OpenID Connect. If permission is granted, access and retrieval of the data is managed by the EGA servers with the secure htsget API developed by the GA4GH Directory and Streaming API (Figure 1). EGA servers enable querying genomic slices in blocks of one Gigabyte.

Visualization of genomic alignments and coverage from data archived at the EGA from the RD-Connect GPAP user interface

To visualize alignments and coverage information in the RD-Connect GPAP, we developed a genome browser module, which includes an embedded IGV. This module is currently accessible to all registered users (Figure 1), who can request access to visualize genomic alignments in two different ways. The first is to click on one of the two IGV-labeled links available from each filtered genetic variant shown in the GPAP. These two links open the alignments in the embedded IGV at the variant position ±10 bp or ±1 Kbp, respectively. The second is to open the genome browser module PAGE and complete the form with an experiment code and chromosomal coordinates. These two options allow users to not only visualize variants prioritized in the RD-Connect GPAP, such as single-nucleotide variants (SNVs)/insertions or deletions (indels) and copy-number variations (CNVs), but also other types of candidate variants, including large re-arrangement breakpoints, which may have been identified outside of the GPAP. The full implementation can be viewed in Video S1.

The GA4GH htsget protocol and the IGV library used also support slicing and rendering of specific information from individual alignments including strand type, mapping quality, pair mapping, insert size, and base quality, which are available to the user when selecting one of the reads. This information is available to the user through the embedded IGV application, which also provides several options that enhance user experience. These enhancements include coloring by read strand, displaying soft clips and the three possible amino acid translations, tagging

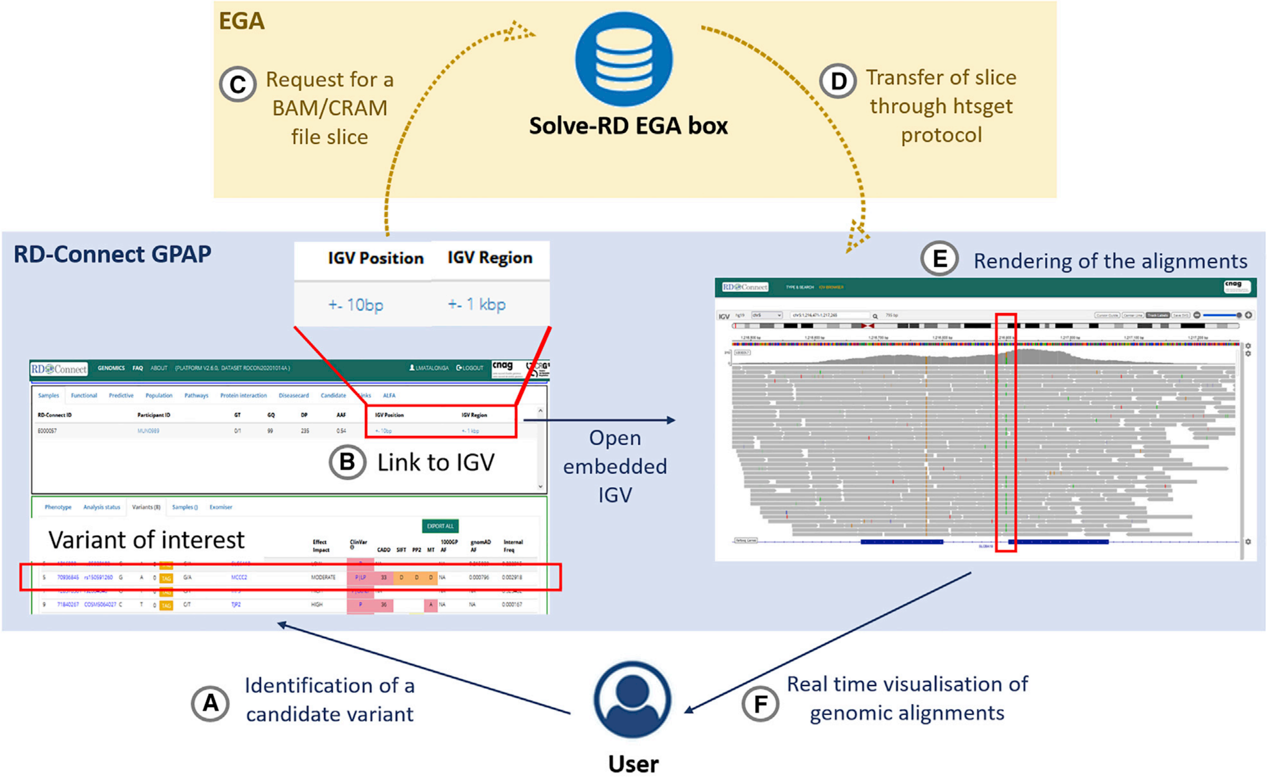

EGA
C Request for a BAM/CRAM file slice
Solve-RD EGA box
D Transfer of slice through htsget protocol
RD-Connect GPAP
IGV Position
IGV Region
+- 10bp
+- 1 kbp
B Link to IGV
Variant of interest
Open embedded IGV
E Rendering of the alignments
A Identification of a candidate variant
User
F Real time visualisation of genomic alignments

Figure 1. Remote real-time visualization workflow

(A and B) An RD-Connect GPAP authorized user (A) identifies a variant of interest and requests to visualize the associated genomic alignments by (B) clicking on the IGV links provided in the interface.

(C) The request is sent to the EGA, which directs access to the corresponding EGA box and alignment file (BAM/CRAM).

(D and E) The htsget protocol generates a slice of the requested alignment as an answer (D), which is rendered by the IGV application implementation in RD-Connect GPAP (E).

(F) The user is able to visualize the alignment in the region of interest.

specific reads of interest, showing all bases, etc. In terms of performance, the mean overall time required to access an exome or genome region and render the corresponding alignments from an experiment within the RD-Connect GPAP for a ±20 bp query is 23.84 s (standard deviation [SD] = 8.54 s, min = 6.45 s, and max = 43.63 s) and for a ±1 Kbp query is 25.92 s (SD = 10.64 s, min = 6.44 s, and max = 102.07 s; Tables S1–S4). The system is reliable since we only had one failed request out of 864 tested (99.88% success rate).

Visualization of alignments to confirm quality of called genetic variants within the Solve-RD project

The described workflow was tested and scaled up to production with 11,750 datasets from the Solve-RD project. Solve-RD BAM files were submitted to a dedicated EGA box and made accessible for GPAP users to visualize selected subsections of the genome through the GA4GH htsget protocol. Clinical scientists from the project now have access to this functionality, allowing them to visually validate causative variants. Preliminary (re-)analysis of the first 4,400 Solve-RD cases resulted in 255 new diagnoses in mid-2021.[15,18] Figure 2 and Video S1 illustrate an example of the visualization of genomic alignments for a causative homozygous variant in *TRIP4* in a patient with cerebellar hypoplasia and spinal muscular atrophy.[18]

DISCUSSION

Real-time access to genomic alignments is necessary to allow clinical scientists to visualize regions and/or variants of interest when interpreting their data to confirm the quality of a candidate variant. Herein, we report an innovative and robust solution that enables visualization of a specific region from a genomic alignment archived remotely at the EGA directly within an IGV instance embedded into the RD-Connect GPAP interface. Similar initiatives to the RD-Connect GPAP, such as the Database of Genomic Variation and Phenotype in Humans using Ensembl Resources (DECIPHER), enable the visualization of causative variants in a genomic browser with pre-loaded tracks (e.g., Genome Aggregation Database [gnomAD], The Clinical Genome [ClinGen] Resource, other DECIPHER causative variants) but do not permit visualization of the specific genomic alignments from a particular individual. Our implementation enables both (1) the inspection of a causative variant from a solved case and comparing it with a current case under investigation and (2) the

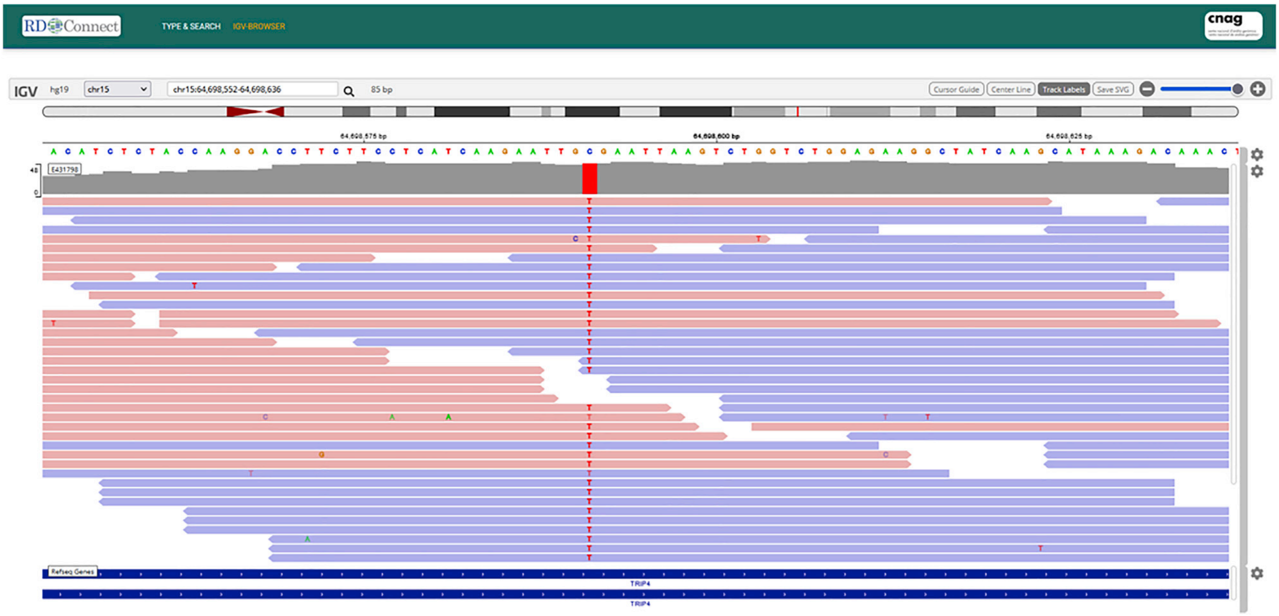


Figure 2. Visualization of slices of genomic alignments archived at the EGA

A screenshot of the GPAP's embedded IGV displaying a slice of a BAM file, archived at the EGA, from a patient with cerebellar hypoplasia and spinal muscular atrophy. The genomic data show a homozygous single-nucleotide change at position 15:64698591C>T (NM_016213.5:c.760C>T [pArg254Ter]). This variant has been reported as disease causing.[18]

investigation of genomic variants or regions of interest for interpretation purposes as shown in Figure 2.

The GA4GH htsget API is a key component of the implementation, enabling the streaming of specific alignment regions and thus negating the need for clinical scientists to have to download or transfer full alignment files between centers, systems, and/or partners with existing solutions such as file transfer protocol (FTP), Aspera, or Globus.[12] The implementation of the htsget API at the EGA had a couple of relevant challenges. One of them was the incompatibility of the htsget API with older versions of BAM files and with compressed alignments (CRAM files). This has been solved in newer versions of the htsget API by incorporating backwards compatibility with older versions of BAM files and adding a functionality to access CRAM files directly. Another challenge was to incorporate decryption of the specific alignment region, as the data in the EGA are stored encrypted at rest and must be decrypted to be of any use to clinical scientists. This was solved by implementing a decryption step in the process of serving the data to the user; the fact that the data are encrypted using a block cipher meant that it was possible to retrieve and decrypt only the blocks that were relevant to the requested alignment region instead of needing to decrypt the entire file.

The described implementation has been tested with 11,750 datasets from Solve-RD, enabling more than 120 users from the project to access and visualize this information in real time and contributing to solving hundreds of previously undiagnosed patients.[5,15] It is the first to demonstrate how third-party systems such as the RD-Connect GPAP can access and render (upon authorization) data archived at the EGA without needing to download the corresponding files in full. The same approach will be used in Solve-RD for other types of omics data such as transcriptomics to visualize RNA sequencing (RNA-seq) alignments and the corresponding variants, sashimi plots, and splice junctions. The htsget functionality is one component of many that make up the EGA data download API, along with components that handle authentication, authorization, etc. The stability of the entire API, like any production service, depends on many aspects such as the stability of the underlying infrastructure, shared resources of the hosting institution, activity patterns of users, and acts of nature. Similarly, there are many approaches to mitigate against instability issues including having failover infrastructure, monitoring and auditing service usage, and developing robust services that, when they do encounter issues, deal with them elegantly and transparently for users. As one example, the EGA data download API implements re-tries automatically for users, and it can also re-start data transfers from where they left off if the user's connection is interrupted.

The GA4GH htsget protocol has been previously implemented by a set of research and commercial providers of human data for demonstration purposes using a public trio dataset[12] and by Genomics England to serve all of its genomic data from the 100,000 Genomes Program in their local clusters. Our approach goes beyond local environments and demonstrates the possibility of connecting and federating systems installed in different countries and under different institutions. The scalability of this model toward federated approaches (federated EGA and/or federated RD-Connect GPAP instances) is feasible for large-scale projects and initiatives such as the European 1+ Million Genomes initiative (https://digital-strategy.ec.europa.eu/en/policies/1-million-genomes) but would require the implementation of a more complex user authentication and authorization process. Other GA4GH standards such as GA4GH Passports (http://bit.ly/ga4gh-passport-v1#passport) and the Authentication & Authorization Infrastructure (AAI) specifications (http://bit.ly/ga4gh-aai-profile) will likely be useful for such purposes since these standards work in conjunction in order to reliably authenticate researchers' digital identities and automate their access to a requested genomic dataset.

In conclusion, we have developed a robust and scalable solution to enable clinical scientists to visualize remotely stored genomic alignment data for clinical research and diagnosis. This implementation has already proven very useful to confirm the quality of candidate variants for patients with a previously undiagnosed RD, and many other applications could be enabled with the same or similar approaches. Our work highlights the impact of developing and implementing interoperability standards, which will be essential for the establishment of large, federated genomics data networks.

Limitations

Our study design faces two main limitations while reporting on rendering performance. The first one is that the performance when rendering alignments from files stored at different EGA boxes has not been tested, and the second is that our approach focuses on the visualization of genomic alignments and does not assess performance on the use of similar formats from other types of omics data such as transcriptomics.

STAR★METHODS

Detailed methods are provided in the online version of this paper and include the following:

- KEY RESOURCES TABLE
- RESOURCE AVAILABILITY
  - ○ Lead contact
  - ○ Materials availability
  - ○ Data and code availability
- METHOD DETAILS
- QUANTIFICATION AND STATISTICAL ANALYSIS

SUPPLEMENTAL INFORMATION

Supplemental information can be found online at https://doi.org/10.1016/j.xgen.2022.100246.

ACKNOWLEDGMENTS

The authors would like to thank the ELIXIR Rare Disease Community, the ELIXIR Federated Human Data Community, and the GA4GH htsget API working group for support and insightful discussions. The authors would also like to acknowledge the Solve-RD ERNs (ITHACA, RND, EURO-NMD, and GENTURIS), the Solve-RD Data Analysis Task Force, and the Solve-RD Infrastructure Work Package members for suggestions on the use cases, testing the system, and providing feedback. This study was partially funded by ELIXIR Implementation Studies "Remote real-time visualization of human RDs genomics data (RD-Connect) stored at the EGA ELIXIR" (2017–2018), "Integration of ELIXIR-IIB in ELIXIR Rare Disease activities (2017–2018)," and "Rare Disease Infrastructure ELIXIR (2019–2020)" and the Solve-RD project, which received funding from the European Union's Horizon 2020 research and innovation program under grant agreement no. 779257. Data were analyzed using the RD-Connect Genome-Phenome Analysis Platform, which has received funding from EU projects RD-Connect, Solve-RD, and EJP-RD (grant nos. FP7 305444, H2020 779257, and H2020 825575), Instituto de Salud Carlos III (grant nos. PT13/0001/0044 and PT17/0009/0019; Instituto Nacional de Bioinformática, INB), and ELIXIR Implementation Studies. We acknowledge the support of the Spanish Ministry of Economy, Industry and Competitiveness (MEIC) to the EMBL partnership, the Centro de Excelencia Severo Ochoa, and the CERCA Programme/Generalitat de Catalunya. We also acknowledge the support of the Generalitat de Catalunya through Departament de Salut and Departament d'Empresa i Coneixement and co-financing by the Spanish Ministry of Economy, Industry, and Competitiveness (MEIC) with funds from the European Regional Development Fund (ERDF) corresponding to the 2014–2020 Smart Growth Operating Program.

AUTHOR CONTRIBUTIONS

A.C. contributed to development and manuscript writing; L.M. contributed to study design and manuscript writing; D.S. contributed to study design; A.S., D.P.-A., and D.P. contributed to technical implementation; S.L. contributed to manuscript writing; M.F.-C. and I.P. contributed to Solve-RD data processing and submission; A.F.R., E.G.-R., A.M., and A.F.S.V. contributed to software implementation and testing; C.H. and J.I.C. contributed to development and design; C.T. contributed to data submission; M.A.F., T.K., and S.B. contributed to coordination, study design, and manuscript writing; A.T. and R.H. contributed to use case; and G.S. and I.G. contributed to coordination. All authors reviewed the manuscript.

DECLARATION OF INTERESTS

The authors declare that they have no conflict of interest.

Received: November 29, 2021
Revised: April 4, 2022
Accepted: December 14, 2022
Published: January 11, 2023

REFERENCES

1. Boycott, K.M., Hartley, T., Biesecker, L.G., Gibbs, R.A., Innes, A.M., Riess, O., Belmont, J., Dunwoodie, S.L., Jojic, N., Lassmann, T., et al. (2019). A diagnosis for all rare genetic diseases: the Horizon and the Next frontiers. Cell *177*, 32–37. https://doi.org/10.1016/j.cell.2019.02.040.
2. Farwell, K.D., Shahmirzadi, L., El-Khechen, D., Powis, Z., Chao, E.C., Tippin Davis, B., Baxter, R.M., Zeng, W., Mroske, C., Parra, M.C., et al. (2015). Enhanced utility of family-centered diagnostic exome sequencing with inheritance model-based analysis: results from 500 unselected families with undiagnosed genetic conditions. Genet. Med. *17*, 578–586. https://doi.org/10.1038/gim.2014.154.
3. Stark, Z., Tan, T.Y., Chong, B., Brett, G.R., Yap, P., Walsh, M., Yeung, A., Peters, H., Mordaunt, D., Cowie, S., et al. (2016). A prospective evaluation of whole-exome sequencing as a first-tier molecular test in infants with suspected monogenic disorders. Genet. Med. *18*, 1090–1096. https://doi.org/10.1038/gim.2016.1.
4. Wright, C.F., Fitzgerald, T.W., Jones, W.D., Clayton, S., McRae, J.F., van Kogelenberg, M., King, D.A., Ambridge, K., Barrett, D.M., Bayzetinova, T., et al. (2015). Genetic diagnosis of developmental disorders in the DDD study: a scalable analysis of genome-wide research data. Lancet *385*, 1305–1314. https://doi.org/10.1016/S0140-6736(14)61705-0.
5. Zurek, B., Ellwanger, K., Vissers, L.E.L.M., Schüle, R., Synofzik, M., Töpf, A., de Voer, R.M., Laurie, S., Matalonga, L., Gilissen, C., et al.; Solve-RD consortium (2021). Solve-RD: systematic pan-European data sharing and collaborative analysis to solve rare diseases. Eur. J. Hum. Genet. *29*, 1325–1331. https://doi.org/10.1038/s41431-021-00859-0.
6. Freeberg, M.A., Fromont, L.A., D'Altri, T., Romero, A.F., Ciges, J.I., Jene, A., Kerry, G., Moldes, M., Ariosa, R., Bahena, S., et al. (2022). The European genome-phenome archive in 2021. Nucleic Acids Res. *50*, D980–D987. https://doi.org/10.1093/nar/gkab1059.

7. Laurie, S., Piscia, D., Matalonga, L., Corvó, A., Fernández-Callejo, M., Garcia-Linares, C., Hernandez-Ferrer, C., Luengo, C., Martínez, I., Papakonstantinou, A., et al. (2022). The RD-Connect Genome-Phenome Analysis Platform: accelerating diagnosis, research, and gene discovery for rare diseases. Hum. Mutat. *43*, 717–733. https://doi.org/10.1002/humu.24353.

8. Wilkinson, M.D., Dumontier, M., Aalbersberg, I.J.J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.W., da Silva Santos, L.B., Bourne, P.E., et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data *3*, 160018. https://doi.org/10.1038/sdata.2016.18.

9. Philippakis, A.A., Azzariti, D.R., Beltran, S., Brookes, A.J., Brownstein, C.A., Brudno, M., Brunner, H.G., Buske, O.J., Carey, K., Doll, C., et al. (2015). The Matchmaker Exchange: a platform for rare disease gene discovery. Hum. Mutat. *36*, 915–921. https://doi.org/10.1002/humu.22858.

10. Laurie, S., Fernandez-Callejo, M., Marco-Sola, S., Trotta, J.R., Camps, J., Chacón, A., Espinosa, A., Gut, M., Gut, I., Heath, S., and Beltran, S. (2016). From wet-lab to Variations: concordance and speed of bioinformatics pipelines for whole genome and whole exome sequencing. Hum. Mutat. *37*, 1263–1271. https://doi.org/10.1002/humu.23114.

11. Wagner, J., Olson, N.D., Harris, L., McDaniel, J., Cheng, H., Fungtammasan, A., Hwang, Y.C., Gupta, R., Wenger, A.M., Rowell, W.J., et al. (2022). Curated variation benchmarks for challenging medically relevant autosomal genes. Nat. Biotechnol. *40*, 672–680. https://doi.org/10.1038/s41587-021-01158-1.

12. Kelleher, J., Lin, M., Albach, C.H., Birney, E., Davies, R., Gourtovaia, M., Glazer, D., Gonzalez, C.Y., Jackson, D.K., Kemp, A., et al.; GA4GH Streaming Task Team (2019). htsget: a protocol for securely streaming genomic data. Bioinformatics *35*, 119–121. https://doi.org/10.1093/bioinformatics/bty492.

13. Robinson, J.T., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E.S., Getz, G., and Mesirov, J.P. (2011). Integrative genomics viewer. Nat. Biotechnol. *29*, 24–26. https://doi.org/10.1038/nbt.1754.

14. de Boer, E., Ockeloen, C.W., Matalonga, L., Horvath, R., Solve-RD SNV-indel working group; Rodenburg, R.J., Coenen, M.J.H., Janssen, M., Henssen, D., Gilissen, C., Steyaert, W., Paramonov, I., Solve-RD-DITF-ITHACA; Trimouille, A., Kleefstra, T., Verloes, A., and Vissers, L.E.L.M. (2021). A MT-TL1 variant identified by whole exome sequencing in an individual with intellectual disability, epilepsy, and spastic tetraparesis. Eur. J. Hum. Genet. *29*, 1470–1471. https://doi.org/10.1038/s41431-021-00937-3.

15. Matalonga, L., Hernández-Ferrer, C., Piscia, D., Solve-RD SNV-indel working group; Schüle, R., Synofzik, M., Töpf, A., Vissers, L.E.L.M., de Voer, R., Solve-RD DITF-GENTURIS; Solve-RD DITF-ITHACA; Solve-RD DITF-euroNMD; Solve-RD DITF-RND; Tonda, R., Laurie, S., et al.; Solve-RD Consortia (2021). Solving patients with rare diseases through programmatic reanalysis of genome-phenome data. Eur. J. Hum. Genet. *29*, 1337–1347. https://doi.org/10.1038/s41431-021-00852-7.

16. Schüle, R., Timmann, D., Erasmus, C.E., Reichbauer, J., Wayand, M., Solve-RD-DITF-RND; van de Warrenburg, B., Schöls, L., Wilke, C., Bevot, A., Zuchner, S., et al.; Solve-RD Consortium (2021). Solving unsolved rare neurological diseases-a Solve-RD viewpoint. Eur. J. Hum. Genet. *29*, 1332–1336. https://doi.org/10.1038/s41431-021-00901-1.

17. Te Paske, I.B.A.W., Garcia-Pelaez, J., Sommer, A.K., Matalonga, L., Starzynska, T., Jakubowska, A., Solve-RD-GENTURIS group; van der Post, R.S., Lubinski, J., Oliveira, C., Hoogerbrugge, N., and de Voer, R.M. (2021). A mosaic PIK3CA variant in a young adult with diffuse gastric cancer: case report. Eur. J. Hum. Genet. *29*, 1354–1358. https://doi.org/10.1038/s41431-021-00853-6.

18. Töpf, A., Pyle, A., Griffin, H., Matalonga, L., Schon, K., Solve-RD SNV-indel working group; Solve-RD DITF-euroNMD; Sickmann, A., Schara-Schmidt, U., Hentschel, A., Chinnery, P.F., Kölbel, H., et al. (2021). Exome reanalysis and proteomic profiling identified TRIP4 as a novel cause of cerebellar hypoplasia and spinal muscular atrophy (PCH1). Eur. J. Hum. Genet. *29*, 1348–1353. https://doi.org/10.1038/s41431-021-00851-8.

19. Robinson James, T., Helga, T., Turner, D., and Mesirov Jill, P. (2020). igv.js: an embeddable JavaScript implementation of the Integrative Genomics Viewer (IGV). Preprint at bioRxiv. https://doi.org/10.1101/2020.05.03.075499.

STAR★METHODS

KEY RESOURCES TABLE

REAGENT or RESOURCE	SOURCE	IDENTIFIER
Software and algorithms		
Htsget protocol	Kelleher et al.[12]	http://samtools.github.io/hts-specs/htsget.html
RD-Connect Genome-Phenome Analysis Platform (GPAP)	Laurie et al.[7]	https://platform.rd-connect.eu/
RD-Connect GPAP Genomics Browser	This paper	https://doi.org/10.5281/zenodo.7386672

RESOURCE AVAILABILITY

Lead contact

Further information and requests for resources should be directed to and will be fulfilled by the lead contact, Sergi Beltran (sergi.beltran@cnag.crg.eu).

Materials availability

This study did not generate new unique reagents.

Data and code availability

This study did not generate any data. The source code of the EGA Data API is available at https://github.com/EGA-archive/ega-data-api. The RD-Connect GPAP Genomics Browser (web interface and server pseudocode using the PyEGA package) is available at: https://github.com/bag-cnag/gpap-genomics-browser, Zenodo: https://doi.org/10.5281/zenodo.7386672. Any additional information required to reanalyze the data reported in this work paper is available from the lead contact upon request.

METHOD DETAILS

To explore and access the herein described application a demo platform is available at the address: https://playground.rd-connect.eu/. All scripts required to run this protocol are deposited in the following GitHub repository: https://github.com/bag-cnag/gpap-genomics-browser (DOI: https://doi.org/10.5281/zenodo.7386672).

QUANTIFICATION AND STATISTICAL ANALYSIS

We used a python script to assess the speed and robustness of the htsget implementation across two different file types, Whole Exome Sequencing (WES) and Whole Genome Sequencing (WGS) and two different range lengths (+/− 20 basepairs, bp and +/− 1 kilobasepairs, kbp). For each of these combinations we queried 24 regions (one in each autosomal chromosome, one in chromosome X and one in the mitochondrial genome) 3 times in a given time window. We repeated the tests in 3 different time windows (Tables S1–S4). The steps of the process tested include user authentication and authorization, transfer of request, data identification, data slicing and transfer of alignment slice.