

## Supplementary Data

### Frameworks

The *seqr* web application is built using Django, a Python web framework, to handle server-side functionality such as URL routing and database querying via its Object Relational Mapper (ORM). For the front end, *seqr* leverages React to provide an extensible, component-based user interface, integrated with Redux to handle state management. Semantic UI (via the Semantic UI React package) is used for Cascading Style Sheets (CSS) to ensure consistent styling throughout the application. While the *seqr* application is designed to be platform agnostic, it is currently hosted on Google Cloud Platform (GCP). Kubernetes is used to manage deployment and resource management within GCP, including using kubernetes secrets to handle sensitive deployment parameters.

### Databases

Data storage in *seqr* is separated into two different databases: a read-only database which includes all annotated variant information, and a read/write database which includes all project and user metadata, including user-generated data like variant tags and notes.

For the variant database, *seqr* leverages Elasticsearch (ES), a powerful no-SQL database optimized for flexible data queries. Data is loaded via an external pipeline, and *seqr* accesses the data in a read-only capacity. This ensures data integrity against malicious or accidental user manipulation. The data loading pipeline

(<https://github.com/broadinstitute/hail-elasticsearch-pipelines>) utilizes Hail to parse, annotate, and format a joint-called VCF with the metadata required for search and display in *seqr*. These annotations include population frequencies, in silico predictor scores, VEP consequences, and sample-specific VCF data such as quality metrics and zygosity. This data is then exported into project-specific ES indices, and indices are never modified after their creation. When new or updated data becomes available, data is re-exported to a new index and the old index is deleted.

All other application data is stored in postgres databases. The primary postgres database contains all user-provided application data. This includes user account information, metadata about projects, cases, and samples, user-generated gene lists and search criteria, and saved variant information. User-provided variant information includes tags, notes, and variant annotations such as transcript of interest. In addition, a copy of any variant saved by users in *seqr* is stored in postgres as well as Elasticsearch in order to improve application performance and consistency. Additionally, *seqr* stores external reference data in a separate postgres database. Such reference data includes gene level information from sources such as GENCODE, OMIM, and dbNSFP, as well as structured phenotype data from HPO. This reference data is updated periodically to ensure accuracy but is not editable by *seqr* application users. Additionally, *seqr* leverages redis for caching key request data in order to optimize application performance.

## Security

As a connected application, *seqr* legally resides within Terra's security boundary and all security controls around the data in Terra extend to *seqr* as well. In order to mitigate potential data breaches, *seqr* operates as a FISMA Moderate system (Federal Information Security Modernization Act; <https://www.cisa.gov/federal-information-security-modernization-act>) and is compliant with the NIST 800-53 Rev 4 information security standard at the Moderate baseline.

## Local installations

In addition to being available to all AnVIL users, the *seqr* platform is available as an open source project. Institutions or laboratories that prefer to operate their own *seqr* installation and have the technical resources to support it are invited to do so. GitHub is used to host the *seqr* code base as a public repository (<https://github.com/broadinstitute/seqr>) and is the same repository used by the Broad's *seqr* deployment. Copyleft protection for *seqr* is licensed under the GNU Affero General Public License v3.0 (<https://github.com/broadinstitute/seqr/blob/master/LICENSE.txt>). Docker and docker compose are used to manage pre-built *seqr* resources to streamline deployment. Instructions for installation onto any cloud platform or on-prem server are available in the GitHub repository ([https://github.com/broadinstitute/seqr/blob/master/deploy/LOCAL\\_INSTALL.md](https://github.com/broadinstitute/seqr/blob/master/deploy/LOCAL_INSTALL.md)).

seqr

Summary Data

Reports

Logged in as Broad Analyst

Log out

Family Variant Search

Project>> Neuromuscular Cohort>> Family: FAM\_100

1

2

Family Description:

Isolated case with LGMD

Analysis Status:

Solved, known gene for phenotype

Assigned Analyst

Broad Analyst

Analysed By

analyst@broadinstitute.org on 08/23/2021

Case Notes

Pending sanger confirmation

Analysis Notes

Completed standard de novo/dominant and recessive searches

Matchmaker Notes

Not submitted

Coded Phenotype:

limb-girdle muscular dystrophy

Post-discovery OMIM #

7

Variant Search

Add Manual Variant

Add Manual SV

MatchMaker Exchange

8

WES - SV LOADED 5/26/2021

WES LOADED 5/7/2020

WES LOADED 10/23/2017

4

FAM\_100\_1

Added 11/05/2020

Age:

24

Age of Onset:

Childhood onset

Individual Notes

Consanguinity:

Other Affected Relatives:

Expected Mode of Inheritance:

Autosomal dominant inheritance, Autosomal recessive inheritance

Assisted Reproduction:

Maternal Ancestry:

French

Paternal Ancestry:

Italian

Imputed Population:

European (non-Finnish)

SV QC Flags:

Raw Calls: >100

Features:

Present

Limbs: Limb-girdle muscular dystrophy

Pre-discovery OMIM disorders:

253600

Previously Tested Genes:

(None)

Candidate Genes:

Genes associated with LGMD

5

FAM\_100\_2

Added 11/05/2020

Age:

65

Individual Notes

Maternal Ancestry:

Italian

Paternal Ancestry:

Italian

Imputed Population:

European (non-Finnish)

Features:

Unaffected

6

FAM\_100\_3

Added 11/05/2020

Age:

60

Individual Notes

Maternal Ancestry:

French

Paternal Ancestry:

French

Imputed Population:

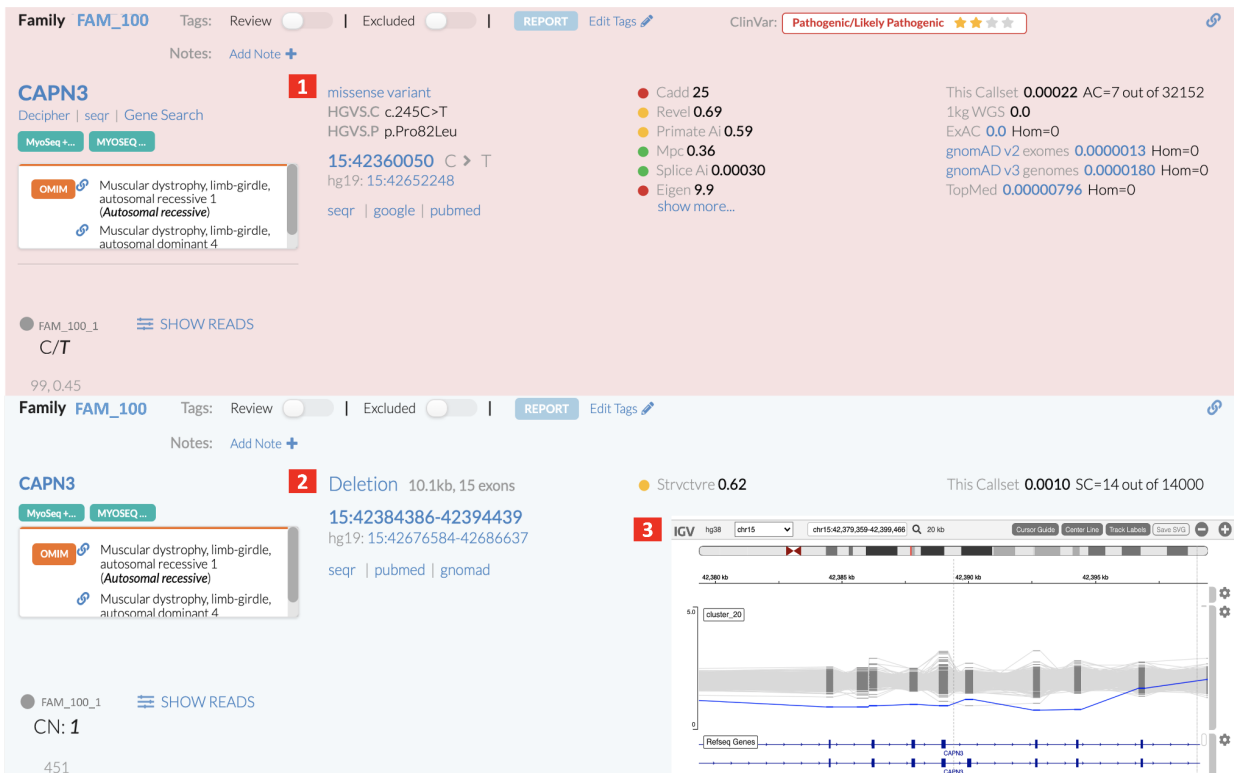
European (non-Finnish)

Features:

Unaffected

**Supplementary Figure S1:** The *seqr* Family Page displaying FAM\_100 as an example. (1) family pedigree; (2) general description of the case; (3) analysis and case details entered by the

analysis team; (4-6) individual level example data including age, ancestry, sample QC information, clinical information using HPO terms; (7) variant search link to begin analysis and an overview of tagged variants on the hover over of the saved variants box, if the case was previously analyzed; (8) type and date of data loaded into *seqr*.



**Supplementary Figure S2:** Variant filtration of SNVs/indels and SVs in tandem. In this example, a recessive restrictive search identified (1) a missense variant and (2) deletion in *CAPN3*. Variants in the raw read data can be viewed using (3) IGV within *seqr*. Both variants were externally validated and reported as the diagnosis for this research participant with limb-girdle muscular dystrophy.

✕

[OMIM](#)
[PubMed](#)
[GeneCards](#)
[Protein Atlas](#)
[NCBI Gene](#)
[GTEx Portal](#)
[Monarch](#)
[Decipher](#)
[UniProt](#)
[gnomAD](#)
[MGI](#)
[IMPC](#)
[ClinVar](#)
[HGMD](#)

<b>Symbol</b>	GNAI1
<b>Ensembl ID</b>	ENSG00000127955
<b>Description</b>	Guanine nucleotide-binding proteins (G proteins) function as transducers downstream of G protein-coupled receptors (GPCRs) in numerous signaling cascades. The alpha chain contains the guanine nucleotide binding site and alternates between an active, GTP-bound state and an inactive, GDP-bound state. Signaling by an activated GPCR promotes GDP release and GTP binding. The alpha subunit has a low GTPase activity that converts bound GTP to GDP, thereby terminating the signal. Both GDP release and GTP hydrolysis are modulated by numerous regulatory proteins (PubMed: 8774883, PubMed: 18434541). Signaling is mediated via effector proteins, such as adenylate cyclase. Inhibits adenylate cyclase activity, leading to decreased intracellular cAMP levels (By similarity). The inactive GDP-bound form prevents the association of RGS14 with centrosomes and is required for the translocation of RGS14 from the cytoplasm to the plasma membrane. Required for normal cytokinesis during mitosis (PubMed: 17635935). Required for cortical dynein-dynactin complex recruitment during metaphase (PubMed: 22327364). [ECO: 0000250] UniProtKB:P10824, ECO: 0000269 [PubMed: 17635935, ECO: 0000269] PubMed: 18434541, ECO: 0000269 [PubMed: 22327364, ECO: 0000269] PubMed: 8774883].
<b>Coordinates</b>	chr7:79768028-80226181 (hg19: chr7:79397344-79855497)
<b>Gene Type</b>	protein_coding

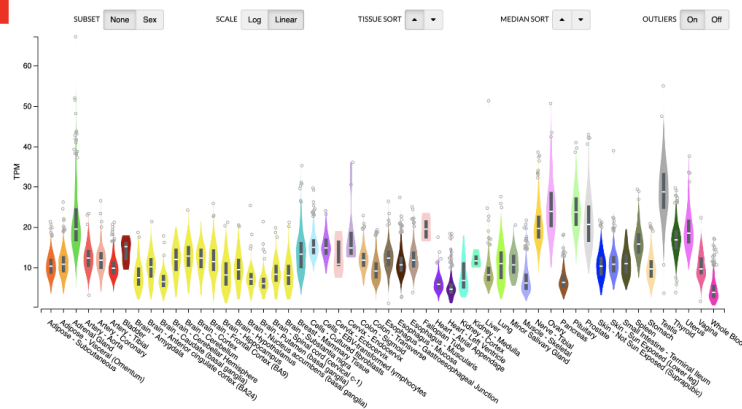
Coding Size	1.1kb
Missense Constraint	z-score: 2.628 (ranked 1563 most constrained out of 19704 genes under study) NOTE: Missense constraint is a measure of the degree to which the number of missense variants found in this gene in ExAC v0.3 is higher or lower than expected according to the statistical model described in [K. Samocha 2014]. In general this metric is most useful for genes that act via a dominant mechanism, and where a large proportion of the protein is heavily functionally constrained. For more details see this <a href="#">README</a> .
LoF Constraint	loofc: 0.3790 (ranked 3327 most intolerant of LoF mutations out of 19704 genes under study) pLI-score: 0.9113 (ranked 2989 most intolerant of LoF mutations out of 19704 genes under study) NOTE: These metrics are based on the amount of expected variation observed in the gnomAD data and is a measure of how likely the gene is to be intolerant of loss-of-function mutations.
Haploinsufficient	pHI-score: 0.9535 NOTE: These are a score under development by the Talkowski lab that predict whether a gene is haploinsufficient based on large chromosomal microarray data set analysis. Scores >0.84 are considered to have high likelihood to be haploinsufficient.
Triplosensitive	pTS-score: 0.7469 NOTE: These are a score under development by the Talkowski lab that predict whether a gene is triplosensitive based on large chromosomal microarray dataset analysis. Scores >0.993 are considered to have high likelihood to be triplosensitive.

OMIM	No disease associations
------	-------------------------

## Shared Notes

Information saved here will be shared across seqr. Please consider using this space to share gene-specific information you learn while researching candidates.

[Variants in GNAI1 cause a syndrome associated with variable features including developmental delay, seizures, and hypotonia - PMC8107131](#)
 By Lynn Pais (6/2/2021)

Add Note 

**Supplementary Figure S3:** Gene Page. Includes (1) links to external resources; (2) basic gene information; (3) gene constraint and dosage sensitivity details; (4) gene-disease relationships based on OMIM; (5) user notes shared across *seqr*; (6) GTEx tissue expression data.

## FAM\_008

**1** Submitted Genotypes: *GNAI1*  
7:80212805 G>C (hg38)

Submitted Phenotypes: Cerebral visual impairment (HP:0100704) • Generalized hypotonia (HP:0001290) • Global developmental delay (HP:0001263)  
• Infantile spasms (HP:0012469) • Seizure (HP:0001250)

[Search for New Matches](#) [Update Submission](#) [Delete Submission](#)

2 Match	First Seen	Contact	Genes	Phenotypes	Follow Up Status	<a href="#">Download Table</a>
0214	5/18/2021	Ally Grater, Swampsea University	<i>GNAI1</i>	DD, ID, seizures	<div>We Contacted Host</div> <div>3 Contact Host</div>	
		Contact Notes				
7802	11/18/2019	G. Mendel, St. Thomas Labs	<i>GNAI1</i>	congenital contractures	<div>We Contacted Host</div> <div>Deemed Irrelevant</div>	<a href="#">Contact Host</a>
		Contact Notes				

### 4 Send Contact Email for FAM\_008

Send To:

Subject:

Dear Dr. Mendel,

We recently matched with one of your patients in Matchmaker Exchange harboring variants in *GNAI1*. Our patient has a de novo missense variant 7:80212805 G>C (hg38) (c.810G>C/p.Lys270Asn), and presents with cerebral visual impairment, generalized hypotonia, global developmental delay and seizures. Would you be willing to share whether your patient's phenotype and genotype match with ours? We are very grateful for your help and look forward to hearing more.

Best wishes,

Analyst  
Broad Institute

Cancel

Send

**Supplementary Figure S4:** The Matchmaker Exchange (MME) node in *seqr*. (1) variant and HPO terms listed in *seqr* that were submitted to MME; (2) examples of MME matches with host's contact details, gene ID, and phenotype, if included in the submission; (3) contact feature to communicate with hosts and track the status of matches within *seqr*; (4) autogenerated email with sample variant and phenotype information from *seqr*.

## TRAPPC4

Decipher | seqr | Gene Search

OMIM

Neurodevelopmental disorder with epilepsy, spasticity, and brain atrophy (*Autosomal recessive*)

splice region variant  
HGVS.C c.454+3A>G

11:119020256 A > G

hg19: 11:118890966

seqr | google | pubmed

● Cadd 22  
● Splice Ai 0.55  
● Eigen 4.8

This Callset 0.00080 AC=25 out of 31448  
1kg WGS 0.0010  
ExAC 0.00027 Hom=0  
gnomAD v2 exomes 0.00019 Hom=0  
gnomAD v3 genomes 0.000198 Hom=0  
TopMed 0.000231 Hom=0

● FAM\_009\_3

G/G

≡ SHOW READS

99,1.0

**Supplementary Figure S5:** Allele count for a splice region variant in the gene *TRAPPC4* showing 25 alleles in the CMG callset. A review of the variants revealed three other cases with similar phenotypes that were this homozygous for this variant across distinct projects and research groups.