Title: Federated analysis of BRCA1 and BRCA2 variation in a Japanese cohort

Authors: James Casaletto[1], Michael Parsons[2], Charles Markello[1],Yusuke Iwasaki[3], Yukihide Momozawa[3], Amanda B. Spurdle[2], Melissa Cline[1]

Affiliations:

1. UC Santa Cruz Genomics Institute, Mail Stop: Genomics, University of California, 1156 High Street, Santa Cruz, CA 95064, USA.
2. QIMR Berghofer Medical Research Institute, 300 Herston Rd, Herston QLD 4006, Australia.
3. Laboratory for Genotyping Development, RIKEN Center for Integrative Medical Sciences, 1-7-22 Suehiro-cho, Tsurumi-ku, Yokohama City, Kanagawa 230-0045, Japan

Author List Footnotes:

Contact Information: jcasalet@ucsc.edu, mcline@ucsc.edu

Summary

More than 40% of the germline variants in ClinVar today are variants of uncertain significance (VUS). These variants remain unclassified in part because the patient-level data needed for their interpretation is siloed. Federated analysis can overcome this problem by "bringing the code to the data": analyzing the sensitive patient-level data computationally within its secure home institution and providing researchers with valuable insights from data that would not otherwise be accessible. We tested this principle with a federated analysis of breast cancer clinical data at RIKEN, derived from the BioBank Japan repository. We were able to analyze these data within RIKEN's secure computational framework, without the need to transfer the data, gathering evidence for the interpretation of several variants. This exercise represents an approach to help realize the core charter of the Global Alliance for Genomics and Health (GA4GH): to responsibly share genomic data for the benefit of human health.

Keywords

Variants of uncertain significance, federated computing, variant classification, benign, pathogenic, data privacy, co-occurrence analysis

Introduction

One obvious and well-studied example of how genetic variation can impact human health is the risk of cancer presented by pathogenic variation in the *BRCA1* and *BRCA2* genes. Pathogenic *BRCA1/2* variants greatly increase the risk of female breast and ovarian cancer (as reviewed) [1],

and they also confer significant risk of pancreatic, prostate, and male breast cancer (as reviewed) [1]. Genetic testing that identifies a pathogenic variant in these genes enables individuals and their families to better understand their heritable cancer risk, and to manage that risk through strategies such as increased screening, cascade testing of family members, and risk-reducing surgery and medication (as reviewed) [1]. However, these risk-reducing strategies are not available to an individual found to carry a variant of uncertain significance (VUS), a rare variant for which there is insufficient evidence to assess its clinical significance. While individually rare, these VUS are collectively abundant. As of May 2021, ClinVar [2], the world's leading resource on the clinical significance of genetic variants, reports that 8,592/25,028 (34.3%) of *BRCA1/2* variants therein are designated as VUS, while an additional 1,204 (4.8%) have conflicting interpretations. In other words, roughly 40% of *BRCA1/2* unique variants in ClinVar have no clear clinical interpretation. Meanwhile, there are many more variants that have been observed in individuals but are not yet in ClinVar: the Genome Aggregation Database (gnomAD) [3] includes an additional 35,635 *BRCA1/2* variants compiled from genomic sequencing research cohorts. Patients of non-European ancestry are significantly more likely to receive a VUS test report from *BRCA1/2* testing [4], a disparity that stems largely from historical biases in genetic studies [5,6].

The VUS problem persists in large part because VUS are rare variants; no single institution can readily gather a sufficient set of observations for robust variant classification. Data sharing would seem to be the natural solution, but it faces logistical challenges. Variant interpretation often requires some amount of case-derived information: clinical observations of the variant in patients and their families together with their cancer history. However, case-level data is sensitive and private, and can rarely be shared directly due to regulatory, legal, and ethical safeguards[7]. Yet sharing data on rare genetic variants is critical for the advancement of precision medicine, as advocated by organizations including the GA4GH [8], the American College of Molecular Geneticists [9] and the Wellcome Trust [10]. Fortunately, most variant interpretation does not require the case-level data *per se*, but rather variant-level summaries of information derived from those data. The ACMG/AMP Guidelines for variant interpretation [11], which specify forms of evidence for interpreting genetic variants, indicate use of variant-level summary evidence including population frequencies (*BA1, BS1, PM2)*, segregation of the variant and the disorder in patient families (*PP1, BS4),* case-control analysis (*PS4)*, and observations of the VUS *in cis* and *in trans* with known pathogenic variants (*PM3* and *BP2*, depending on the disorder). What is needed is an approach to derive this variant-level evidence from siloed case-level datasets without the need for direct access.

Federated analysis offers such an approach. Rather than an institution sharing its case-level data with external collaborators, those collaborators share an analysis workflow with the institution. The institution runs the workflow on their cohort, generating variant-level data that is less sensitive and can be shared more openly. This can yield valuable evidence for variant interpretation without the sensitive data leaving the home institution[12]. Container technologies support this approach by bundling the software and all its dependencies into a single module for straightforward installation and deployment on a collaborator's system[13]. These technologies include Docker [14], Singularity [15], and Jupyter [16]. Containers and workflows can be shared on the

Dockstore platform [17] so that multiple institutions can execute the exact same software, promoting reproducibility.

We developed analysis workflows to mine tumor pathology, allele frequency, and variant co-occurrence data for *BRCA1* and *BRCA2* from breast cancer patient cohorts at RIKEN, derived from BioBank Japan [18,27]. This analysis allowed the assessment of new variant interpretation knowledge from a cohort that would not otherwise be accessible. In addition to generating new knowledge on these genetic variants, this yielded new knowledge on the genetics of the Japanese population, which is underrepresented in most genetic knowledge bases. Moreover, we've generalized our container approach to work with any genotype-phenotype combinations of data.

Design

In principle, one could share access to a protected genomics dataset by transferring that data to a trusted third party, such as a secure cloud, but a dataset which contains personally identifiable information generally cannot or should not be moved from its secure source location. Indeed, the BioBank Japan data is prohibited from anonymous export. Federated analysis leaves the data securely in place and instead moves the analytic software (which tends to be many orders of magnitude smaller in size than a research cohort) to the data host institution. We designed our federated analysis software to be transparent, modular, and extensible. The analysis software creates multiple reports that capture data quality, associated phenotype, allele frequency, and variant co-occurrence.

Any researcher analyzing a dataset must first ensure that the data values are interpreted correctly; this is especially true when the researcher cannot interact with the data locally. The data quality report addresses that need by providing basic statistics (such as minimum, maximum, mean, mode, and median) and reporting any missing or unexpected data values. For this report, we provide a JavaScript Object Notation (JSON) configuration file which defines each of the fields of interest, here as exemplified for the content of the tumor pathology file. The report could be used to check data quality for any delimited file, with or without a header. This data quality report represents a general solution which can be reused for other data sets. The Supplemental Information section includes two full examples with a data quality report.

The variant frequency and co-occurrence report was written to summarize the variant counts stratified by patient group (affected vs. control) for estimating allele frequencies; and to report on variants of uncertain significance (VUS) which co-occur *in trans* either with known pathogenic variants in complex heterozygous genotypes, or with themselves as homozygous genotypes. The program takes as input a VCF file and outputs JSON files with the variant counts and the co-occurring variant information. If associated phenotype data is provided, then our software will intersect those phenotype data with the genotype data in the VUS reports. This requires using a tab-separated file with the string 'ID' as the primary key of this table whose values match those in the VCF file.

In addition to this generic, domain-agnostic data quality report, we provide the ability to integrate and call a custom, domain-specific report which can be leveraged to identify data anomalies in a known domain. This feature is optional. In our research, we leveraged this feature to implement a tumor pathology report in which we calculate the number and proportion of triple-negative breast cancers of all breast cancers for which ER, PR, and HER2 test results are available. This pathology report reads a tab-delimited file which is indexed by the sample identifier. Even though these sample identifiers are anonymized, we did not want to risk exposing any identifier in the results. Our tumor pathology report takes as input that same tumor pathology file, and for each pathology feature outputs a summary of the number and proportion of patients stratified by pathogenic variant status, with an odds ratio, confidence interval, and Fisher's exact p-value for the comparison. Additionally, the report includes a comparison of mean age at diagnosis (and entry) for the different patient groups. This can be extended to measure the statistics for any stratification of gene and pathology data. Importantly, this optional custom report can be independently used to validate that the researcher and the collaborator are reading and interpreting the data equivalently. In federated computing, the researcher never has direct access to the data, so any anomalies in the data could be identified if the researcher and collaborating institution agree to independently generate the same report and then compare the results. Indeed, we used this pathology report to validate our federated approach and to verify that there were no data anomalies that would preclude our analysis.

While our research focuses on VUS in *BRCA1* and *BRCA2* genes and associated tumor pathologies, the software was written to work with any genotype-phenotype combinations of data. In the Supplemental Information section, we provide an illustration of how one might assess genetic variation in cardiomyopathy by evaluating VUS in the *MYH7* gene along with associated cardiac phenotype data. All the configuration is passed as command-line options to the program to define such parameters as gene name, whether the data are phased, and which human genome version to use as genomic coordinates. Moreover, all the Python libraries required to run this code are included in the Docker container. The Supplemental Information section contains two full examples with variant frequency and co-occurrence reports with intersecting phenotype data.

Methods

The Dataset

Our analysis revolved around case-control association study data of individuals of Japanese ancestry [18,27]. These data reside at RIKEN and cannot be accessed outside of RIKEN. The dataset reports the variants in coding regions of 11 genes associated with hereditary breast, ovarian, and pancreatic cancer syndrome, including *BRCA1* and *BRCA2.* Additionally, the dataset reports the tumor pathology of the breast cancer patients, including ER, PR and HER2 status. The controls within this cohort are individuals who were at least 60 years old when sequenced and who have neither personal nor family history of cancer. The variant data were stored in a Variant Call Format (VCF) file and the associated phenotype (pathology) data were

stored in a tab-delimited file. No other files were required for this analysis.

Variant Interpretation Evidence

We developed Docker containers to collect data for two forms of evidence (ACMG code/s designated in parenthesis): allele frequencies (*BA1, BS1)* and variant co-occurrences *(BS2).* In addition, we estimated *in silico* predictions of variant pathogenicity (*BP4, PP3)* using the BayesDel method for annotation of predicted missense substitutions and insertion-deletion changes [19].

Allele Frequencies

By the AMCG/AMP standards, the frequency of a variant in a large, outbred population can offer three different forms of evidence for variant interpretation. First, when the variant is observed at a far greater frequency than expected for the disorder in question, this is such a strong indicator of benign impact (*BA1)* that the variant can be considered benign without any further evidence. Second, when the variant's frequency does not meet the *BA1* threshold but is still greater than expected for the disorder, the frequency represents strong evidence (*BS1)* that can contribute to a benign interpretation. Third, when the variant is absent from controls or reference population datasets, its absence represents moderate evidence (*PM2)* that can contribute to a pathogenic interpretation [11]. While gnomAD is commonly used as source of population frequencies, gnomAD 3.1 contains data from only 2,604 East Asian genomes [3] while gnomAD 2.1 contains data from 9,977 exomes (Francioli et al. "gnomAD v2.1." *gnomAD Blog* (blog). October 17, 2018.); gnomAD 2.1 contained 76 Japanese exomes, while the number of Japanese genomes in gnomAD 3.1 is unknown. A Japanese biobank with tens of thousands of samples might plausibly contain additional evidence not available through gnomAD. When considering population frequencies, one must consider the source of the samples and whether individuals affected by the disorder are likely to be present in the dataset [20]. Accordingly, we evaluated the non-cancer subset of gnomAD and the control samples from BioBank Japan. Each ClinGen Variant Curation Expert Panel (VCEP) determines the precise rules for applying the ACMG/AMP standard to the genes and diseases under their purview, including the population frequency thresholds for *BA1* and *BS1* evidence. By the proposed rules of *BRCA* ClinGen Variant Curation Expert Panel (VCEP), the threshold for *BA1* evidence is an allele frequency of greater than 0.001 while the *BS1* frequency threshold is 0.0001 (A. Spurdle, M. Parsons, personal communication, March 12, 2021).

in silico Prediction

By ACMG/AMP standards, if multiple lines of computational evidence predict that a variant will impact either protein function or RNA splicing, that observation can contribute to a pathogenic interpretation (*PP3)*. Conversely, if multiple lines of computational evidence predict that the variant will have no functional impact, that observation can contribute to a benign interpretation (*BP2)*. We estimated the probability that the variant would impact protein function with BayesDel [19], a meta-predictor that has been shown to outperform most others [21]. By the proposed rules of the *BRCA* ClinGen VCEP, a BayesDel score of less than 0.3 predicts a

benign interpretation while a BayesDel score of greater than 0.3 predicts a pathogenic interpretation (Parsons, Michael, Emma Tudini, and Amanda Spurdle. 2020. “ENIGMA BRCA1 and BRCA2 Variant Curation Expert Panel.” December 1, 2020. https://clinicalgenome.org/affiliation/50087/).

in trans co-occurrence

In fully penetrant diseases with dominant patterns of inheritance, if one observes a VUS *in trans* (on the opposite copy of the gene) with a known pathogenic variant in the same gene in an individual without the disease phenotype, that observation represents evidence of a benign impact. For *BRCA2* (and more recently *BRCA1*), co-occurrences of two pathogenic variants in the same gene are associated with Fanconi Anemia, a rare debilitating disorder characterized by deficient homologous DNA repair activity, bone marrow failure, early cancer onset and a life expectancy that rarely extends past 40 [22]. Consequently, when an older individual is observed with a *BRCA1* or *BRCA2* VUS as either a homozygous genotype or a compound heterozygous genotype (*in trans* with a pathogenic variant in the same gene), that observation suggests a benign interpretation for the VUS. One caveat is that most clinical sequencing does not report phase; any single co-occurrence of two variants might be *in trans* or *in cis*. However, if a VUS co-occurs with two different pathogenic variants in two different patients, one can assume that at least one of those co-occurrences is *in trans* [23]. Based on these clinical observations, VUS homozygosity or compound heterozygosity with a known pathogenic variant in an individual known or inferred to be without Fanconi Anemia features provides strong evidence against pathogenicity (*BS2*) [21,22].

Collaboration Details

In advance of developing the containers, the authors communicated to determine which data were available and in which format the data were stored. In our research, the variant data were stored in a single VCF (Variant Call Format) file with anonymized sample identifiers, and the pathology data were stored in a single TSV (tab-separated values) file indexed by the same sample identifiers. The data were already prepared in these files as a product of the research that generated the data in the first place [18,26], so no additional data preparation steps were required. RIKEN provided a pair of files (one VCF file and one tumor pathology TSV file) with bogus data to preserve privacy but simultaneously allow the UCSC researchers to develop their containers. As previously mentioned, the UCSC team initially developed the container to generate a tumor pathology report that the RIKEN team already independently generated for a previous publication [18]. When the UCSC team finished preparing the container for that report, they notified the team at RIKEN to download the container code and run it against the data set. After a few iterations and email communications, the reports generated by each team were found to match exactly, thereby validating that accurate analysis could be performed on this data using a federated approach. Subsequently, the UCSC team developed the container to create the co-occurrence and allele frequency report along with the intersection and data quality report. Once those reports were generated, they were sent to the QIMR team to analyze for variant interpretation. The instructions for running the container are straightforward and are

well-documented in the software repository. There are no special skills required to run the container or collect the results. In all, the total amount of interaction required to collaborate was minimal, in part because the QIMR team had previously collaborated with the RIKEN team using some of the same data [18].

Analysis Approach

We created our Docker containers with Python 3.73 code which (a) collects observational statistics on tumor pathology, (b) gathers variant counts for estimating allele frequencies and (c) identifies VUS which either co-occur with a known pathogenic variant in the same gene, or which co-occur with themselves (homozygous VUS). When reporting co-occurrences, we also reported the age of the patient, to review data against expectations of age at presentation of Fanconi Anemia. To identify VUS, we checked the classifications provided by the ClinGen-approved ENIGMA expert panel in BRCA Exchange [25]. If the clinical significance was 'Unknown', or if the variant did not appear in BRCA Exchange, then we labeled the variant a VUS. We applied this container to the BioBank Japan samples. We identified *BRCA1* or *BRCA2* variants which appeared as homozygotes and/or co-occurred with a known pathogenic variant in the same gene. Sequencing data was not phased, but details on the co-occurring variant/s were provided to aid inference of whether a VUS was *in cis* or *in trans*.

Results

We describe here an example of how federated analysis can add information of value for variant interpretation. We analyzed a case-control study of Japanese individuals whose case-level data resides at RIKEN [18,26]. Since these data are not accessible to external researchers, the UC Santa Cruz team developed analysis software, in the form of a Docker container, and shared it with the RIKEN team. The RIKEN team applied the container to analyze this cohort *in situ*, within their secure institutional environment, generating variant-level summary data that contained no personal information and can be shared more openly. The QIMR Berghofer team then applied these data to variant interpretation.

As an initial quality control exercise, we replicated the contents of Supplemental Table 4 from a previous publication on these data [18], using the tumor pathology data. This table contrasts the patients with or without pathogenic variants in terms of factors including family history of seven types of cancer; estrogen, progesterone and herceptin receptor status; and age at diagnosis. We were able to replicate this table precisely, indicating that we were able to process the data accurately. This exercise also demonstrated that our container can be used to generate scientifically meaningful results. While this step was not mandatory for our analysis, we recommend it for the reasons just stated.

Subsequently, we applied the Docker container to analyze the complete patient cohort. We observed 19 *BRCA* variants that have not yet been interpreted by the ClinGen *BRCA1/2* expert panel. For each VUS, we reported its allele frequency in the controls, and any observations of

the VUS co-occurring with a known pathogenic variant in the same gene (Table 1). We also annotated variants for single-submitter curations in ClinVar.

Eleven VUS met the standard for stand-alone evidence of benign impact (*BA1*) based on the allele frequencies in the BioBank Japan controls; all of these VUS were predicted bioinformatically to have benign impact (*BP4*). All eleven VUS will meet the standard of Benign interpretation based on their frequency evidence from the Japanese cohort. Additionally, two of these variants (*BRCA1* c.4729T>C; *BRCA2* c.964A>C) were observed to co-occur with at least two different pathogenic variants in the same gene, evidence sufficient to apply the *BS2* criterion. Of these eleven VUS, four have single-submitter classifications in ClinVar as Benign or Likely Benign, five have conflicting interpretations, and two are designated by ClinVar as VUS. Based on observations currently in gnomAD [3], seven of these variants would have met the *BA1* criterion, three would have met the *BS1* criterion, and one was absent (meeting the *PM2* criterion). For each of the variants present in gnomAD, East Asian was the continental population with the greatest allele frequency at the 95% confidence level (popmax) [30], a fact that itself adds confidence to the BioBank Japan observations. While seven of the variants could have been interpreted as benign on the basis of data in gnomAD, the federated analysis supported the interpretation of four additional variants. This greater sensitivity in the BioBank Japan results reflects the greater cohort size: while gnomAD contains 2,604 East Asian genomes and 9,977 East Asian exomes, the BioBank Japan control group contains 23,731 Japanese individuals.

Five VUS showed strong evidence of benign impact (*BS1*) based on their BioBank Japan allele frequencies, and evidence predictive of benign impact according to BayesDel (*BP4*). These five VUS meet the standard of Likely Benign interpretation based on their frequency and bioinformatic evidence combined. Additionally, two of these VUS had a single co-occurrence with a pathogenic variant in control individuals; while one should not put too much weight on any single homozygous observation, together with the *BS1* and *BP4* evidence, the data present a consistent picture of benign interpretation supported by multiple lines of evidence. One of these five variants is classified in ClinVar as Likely Benign, while the other four are classified as VUS. Four of these VUS would reach the *BS1* evidence standard based on their gnomAD population frequencies while a fifth is absent from gnomAD. The BioBank Japan analysis supports reclassifying five variants, only four of which could be reclassified on the basis of data in gnomAD.

Finally, three additional variants were each observed in a single heterozygous co-occurrence and have BayesDel scores predictive of benign impact (*BP4*). With one co-occurrence observation apiece, we cannot predict whether the co-occurrence is *in trans* or *in cis*, so these observations are not themselves sufficient for evidence of benign impact. However, these co-occurrences could contribute to benign evidence when and if the same VUS are observed to co-occur with other pathogenic variant(s) in another cohort. These VUS are rare variants absent from gnomAD and have either conflicting or VUS interpretations in ClinVar.

Discussion

With this demonstration of federated analysis, we analyzed a protected cohort that we would not have been able to access directly and gathered knowledge on Japanese genetics to further the interpretation of *BRCA1/2* variants. Of 19 variants currently tagged as VUS by the ClinGen BRCA expert panel, 12 were VUS or conflicting in ClinVar. The suggested interpretations based on bioinformatic and frequency analysis assign a Benign or Likely Benign classification for 16 variants and highlight the value of extending data capture to a subpopulation not yet well represented in gnomAD. We also demonstrated the federated collection of variant co-occurrences and age at presentation; these data together provided further evidence supporting the Benign and Likely Benign variant interpretations. This analysis would not be feasible with the existing population frequency resources. For example, gnomAD, the resource selected by ClinGen as its standard, does not yet have a large Japanese cohort, and now shares variant co-occurrences but without the patient age information that is needed for ruling out Fanconi Anemia under ENIGMA's variant interpretation rules. These samples had been analyzed previously by the RIKEN and ENIGMA teams [18,26], a fact that explains why an analysis of nearly 30,000 samples revealed only 19 VUS. This federated analysis allowed us to revisit these data with updated classification criteria, as well as collecting new evidence on variant co-occurrences. Further, by developing a tumour pathology report, we provide proof of principle that federated analysis can be designed to capture other clinical features relevant for variant interpretation. These additional data types are generally provided only in summary level data presentations from published cohorts, at best. Additionally, this method can be applied to any other phenotype-genotype relationship that could benefit from otherwise siloed datasets.

We have also demonstrated that there are international sequencing projects that contain valuable information that could be applied today to variant interpretation but are not yet represented in major population data repositories. This is illustrated by the number of Japanese samples analyzed in this study (7,104 cases plus 23,731 controls) versus the size of gnomAD's East Asian cohort (2,604 genomes plus 9,977 exomes). In principle, the gnomAD and the related population genomics resources will grow with time to comprehensively represent all global populations. In practice, due to the high cost of processing external sequence data, gnomAD mostly imports data from cohorts that were sequenced at the Broad, where sequencing data is processed to a common standard (H. Rehm, personal communication, October 4, 2021). For these reasons, capturing global genetic diversity can benefit from gathering evidence from international sources. Because traditional data sharing is blocked by barriers including laws that prohibit exporting genomic sequences, federated analysis can advance data sharing by limiting the scope of data to be shared to the information most needed.

In this instance, the data sharing was simplified by the fact that the RIKEN team had already assembled a case-control dataset on breast cancer, and in doing so, had already reduced the complex phenotypic data to a set of simplified terms. In a typical variant interpretation scenario, the situation is more involved. In genetic testing, the phenotypic data is often absent, or provided in unstructured text fields that must be curated manually prior to any analysis - traditional or federated. Where phenotypic data is available in a structured, electronic form,

federated analysis can be viable. The cancer diagnosis (or lack thereof) can be represented through Human Phenotype Ontology (HPO) terms [27], with Disease Ontology [28] terms representing the tumor pathology. For example, if the phenotype file had represented the disease phenotype with HPO terms rather than the simplified representation, one might distinguish between cases and controls in the genotype-phenotype report by recognizing breast cancer cases with the HPO term HP:0003002 (Breast Carcinoma), or potentially the less specific HPO term HP:0100013 (Neoplasm of the Breast). Similarly, if the phenotypic data were associated with cardiomyopathy, one could use the HPO term HP:0001639 to represent hypertrophic cardiomyopathy as a phenotype, or the more general HPO term HPO:0001639 to represent cardiomyopathy. Structured models for phenotypic and genomic data exchange, such as Phenopackets [29], increase the opportunity for federated approaches by improving the data interoperability. With the growth in standards developed by the GA4GH and other organizations, and increasing adoption of electronic data standards worldwide [29], this federated analysis model can be generalized and extended into more areas within genomics. Emerging GA4GH technologies including Beacon V2, Matchmaker Exchange and Data Connect can suggest the presence of samples of interest in remote, siloed cohorts, such as cases with rare monogenic disorders. This federated analysis approach complements such approaches by allowing further analysis of these samples while safeguarding patient privacy.

While gnomAD is considered the *de facto* source of ground truth for allele frequency data in genomic research [26] and the only one used in the FDA-approved variant curation interface of ClinGen, our federated solution does not, *per se*, require using it. Any database deemed more appropriate for a particular use case or cohort may be used as the source of allele frequencies provided that the data itself is formatted in a VCF sites file. Similarly, we used ClinVar as our source of ground truth for variant classification, and the ClinVar database may be substituted with another classification database if the data is formatted properly. These data formats are discussed in the Supplementary Information section.

Limitations of the Study

Federated computing is being widely adopted, but it does present its own challenges in data privacy and system security. Docker containers are, to an extent, “black boxes”. To ascertain whether the analysis is truly both secure and privacy-preserving, an auditor would need to carefully inspect the Dockerfile definition of the container as well as all the software that runs in the container. We mitigated this risk by writing our reports to local text files which could be examined by the RIKEN team before being shared externally. Additionally, we published the software as open source so it may be directly inspected by collaborators. A second, related problem is that one cannot readily determine whether software might damage or compromise the security of the system on which it runs. One promising solution to this problem is certification. Within the emerging field of applications security testing, there are software platforms that can dynamically assess the system accesses of the software under test; while the current platforms are commercial, there will likely be an open-source version in time. Eventually, this may become an element of the GA4GH Cloud Testbed, currently under development. This testbed infrastructure will initially serve as a platform for testing compliance

with GA4GH standards and will extend to encompass performance benchmarking. In the future, this platform could potentially also report activity that suggests a security risk, such as the details of outgoing network or disk traffic; and publishing these certification results could fit well within the framework of container libraries such as Dockstore. As an immediate solution to this problem, collaborating institutions should run such otherwise unsecured containers in a virtual machine sandbox environment which is completely isolated from their internal network.

Another limitation of our approach is that it requires data in the particular format that our software recognizes, namely tab-separated files and VCF files. In other words, the software is not agnostic of the file format. Moving forward, we will be able to generalize this approach by leveraging the data standards under development by the GA4GH, which will allow methods to compute over generalized data representation models rather than restricting their input to specific file formats. In particular, the standards of the GA4GH Cloud Workstream are already making it easier to leverage software methods across many different computing platforms. Further development will facilitate the streamlined execution of containerized workflows, the representation of phenotypic data, and the sharing of genetic knowledge.

Acknowledgements

We gratefully thank Gunnar Rätsch for instigating this project, and the members of the BRCA Challenge Evidence Gathering Group for discussion on the analytical design. JC is supported by NHGRI grant U54HG007990 and NHLBI grant U01HL137183.
ABS and MTP are supported by funding from the Australian National Health and Medical Research Council (APP177524). YM is supported by AMED under grant number JP19kk0305010 (to YM). MSC is supported by NCI grant U01CA242954 and BioData Catalyst fellowship OT3 HL147154 from the NHLBI through UNC-CH 5118777.

Author Contributions

MSC, ABS and YM planned the analysis. The docker container was developed by JC with input from YM and technical guidance from CM. YI and YM executed the container. MSC, ABS, JC and MTP analyzed the results, and prepared the manuscript.

Declaration of Interests

No conflict for MSC, JC, CM, ABS and MTP

Tables

Gene	BRCA2	BRCA2	BRCA2	BRCA1	BRCA2	BRCA2	BRCA2
Variant (cDNA HGVS)	c.6325G>A	c.7052C>G	c.943T>A	c.4729T>C	c.4365A>G	c.6131G>T	c.964A>C
Variant (Protein HGVS)	p.A2351G	p.A2351G	p.C315S	p.S1577P	p.A2351G	p.G2044V	p.K322Q
ClinVar Classification (May 1, 2021)	B/LB	B/LB	B/LB	B/LB	LB	Conflict	Conflict
gnomAD 2.1.1 Exome Frequency (EAS)	2.55E-03	1.87E-03	5.30E-03	2.65E-04	Absent	4.52E-04	4.31E-04
gnomAD 3.1.1 Genome Frequency (EAS)	2.39E-03	2.02E-03	5.03E-03	2.02E-04	2.01E-03	4.52E-03	2.41E-03
ACMG/AMP Code from gnomAD	BA1	BA1	BA1	BS1	BS1	BA1	BA1
Biobank Japan Frequency (Controls)	1.46E-02	3.16E-03	1.56E-03	1.14E-02	4.64E-04	3.29E-02	2.31E-03
ACMG/AMP Freq from BioBank Japan	BA1	BA1	BA1	BA1	BS1	BA1	BA1
BayesDel Score	-0.61	-0.24	-0.41	0.03	-0.52	-0.16	-0.08
Bioinformatic Code	BP4	BP4	BP4	BP4	BP4	BP4	BP4
ACMG/AMP Class based on Frequency and Bioinformatics	B	B	B	B	LB	B	B
Gene	BRCA1	BRCA1	BRCA2	BRCA2	BRCA2	BRCA2	
Variant (cDNA HGVS)	c.154C>T	c.811G>A	c.5969A>C	c.3395A>G	c.9733T>G	c.5660C>T	
Variant (Protein HGVS)	p.L52F	p.V271M	p.D1990A	p.K1132R	p.S3245A	p.T1887M	
ClinVar Classification (May 1, 2021)	Conflict	Conflict	Conflict	VUS	VUS	VUS	
gnomAD 2.1.1 Exome Frequency (EAS)	1.36E-03	1.32E-03	0	Absent	Absent	1.13E-04	
gnomAD 3.1.1 Genome Frequency (EAS)	4.03E-04	1.21E-03	4.03E-04	0.000201	Absent	Absent	
ACMG/AMP Code from gnomAD	BA1	BA1	BS1	BS1	PM2	BS1	
Biobank Japan Frequency (Controls)	6.78E-03	6.28E-03	2.61E-03	3.75E-03	1.01E-03	1.69E-04	
ACMG/AMP Freq from BioBank Japan	BA1	BA1	BA1	BA1	BA1	BS1	
BayesDel Score	0.14	0.06	-0.08	-0.2	-0.47	-0.29	
Bioinformatic Code	BP4	BP4	BP4	BP4	BP4	BP4	
ACMG/AMP Class based on Frequency and Bioinformatics	B	B	B	B	B	LB	

Gene	BRCA2	BRCA2	BRCA2	BRCA2	BRCA2	BRCA2	
Variant (cDNA HGVS)	c.2672T>A	c.587G>T	c.8040C>G	c.358G>A	c.3983G>A	c.6637T>C	
Variant (Protein HGVS)	p.V891D	p.S196I	p.D2680E	p.V120M	p.S1328N	p.S2213P	
ClinVar Classification (May 1, 2021)	VUS	VUS	VUS	Absent	Conflict	Conflict	
gnomAD 2.1.1 Exome Frequency (EAS)	Absent	1.78E-04	Absent	Absent	0	Absent	
gnomAD 3.1.1 Genome Frequency (EAS)	Absent	Absent	0.000202	Absent	Absent	Absent	
ACMG/AMP Code from gnomAD	PM2	BS1	BS1	PM2	PM2	PM2	
Biobank Japan Frequency (Controls)	9.69E-04	4.64E-04	9.69E-04	0	0	0	
ACMG/AMP Freq from BioBank Japan	BS1	BS1	BS1	PM2	PM2	PM2	
BayesDel Score	-0.05	-0.22	-0.05	-0.48	-0.57	-0.06	
Bioinformatic Code	BP4	BP4	BP4	BP4	BP4	BP4	
ACMG/AMP Class based on Frequency and Bioinformatics	LB	LB	LB	VUS	VUS	VUS	

Table 1: Summary of the variant data. The HGVS terms reflect the NM_007294.3 transcript for *BRCA1* and NM_000059.3 for *BRCA2*. Variants are designated as B (Benign), B/LB (Benign or Likely Benign), LB (Likely Benign), Conflict (Conflicting Interpretations), VUS (Uncertain Significance) or Absent (Not Found). All variants scored against the BayesDel *in silico* predictor with a score of less than 0.3, within the BP4 scoring range. Additionally, two variants were observed to co-occur with two more pathogenic variants in the same gene, indicating that at least one of these co-occurrences must be *in trans*, which meets the standards of BS2 evidence. In *BRCA1*, we observed co-occurrences of c.4729T>C with c.1518del and c.188T>A, and in *BRCA2,* we observed co-occurrences of c.964A>C with c.6952C>T, c.5645C>A and c.6244G>T. While these VUS had sufficient evidence for classification on allele frequencies only, these co-occurrences add further support to benign classification. We further observed co-occurrences of *BRCA2* c.5660C>T with c.1261C>T and c.4365A>G with c.7480C>T, evidence which could support a benign classification if these variants are observed in co-occurrences with different pathogenic variants in other patient cohorts.

References

1. Collins, J.M., and Isaacs, C. (2020). Management of breast cancer risk in *BRCA1/2* mutation carriers who are unaffected with cancer. Breast J *26*, 1520–1527.
2. Landrum, M.J., Lee, J.M., Benson, M., Brown, G.R., Chao, C., Chitipiralla, S., Gu, B., Hart, J., Hoffman, D., Jang, W., et al. (2018). ClinVar: improving access to variant interpretations and supporting evidence. Nucleic Acids Research *46*, D1062–D1067.
3. Karczewski, K.J., Francioli, L.C., Tiao, G., Cummings, B.B., Alföldi, J., Wang, Q., Collins, R.L., Laricchia, K.M., Ganna, A., Birnbaum, D.P., et al. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature *581*, 434–443.
4. Kurian, A.W. (2010). BRCA1 and BRCA2 mutations across race and ethnicity: distribution and clinical implications. Current Opinion in Obstetrics & Gynecology *22*, 72–78.
5. Landry, L.G., Ali, N., Williams, D.R., Rehm, H.L., and Bonham, V.L. (2018). Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice. Health Affairs *37*, 780–785.
6. Sirugo, G., Williams, S.M., and Tishkoff, S.A. (2019). The Missing Diversity in Human Genetic Studies. Cell *177*, 26–31.
7. Harris, T.L., and Wyndham, J.M. (2015). Data Rights and Responsibilities: A Human Rights Perspective on Data Sharing. Journal of Empirical Research on Human Research Ethics *10*, 334–337.
8. Siu, L.L., Lawler, M., Haussler, D., Knoppers, B.M., Lewin, J., Vis, D.J., Liao, R.G., Andre, F., Banks, I., Barrett, J.C., et al. (2016). Facilitating a culture of responsible and effective sharing of cancer genome data. Nat Med *22*, 464–471.
9. ACMG Board of Directors (2017). Laboratory and clinical genomic data sharing is crucial to improving genetic health care: a position statement of the American College of Medical Genetics and Genomics. Genet Med *19*, 721–722.
10. Wright, C.F., Ware, J.S., Lucassen, A.M., Hall, A., Middleton, A., Rahman, N., Ellard, S., and Firth, H.V. (2019). Genomic variant sharing: a position statement. Wellcome Open Res *4*, 22.
11. on behalf of the ACMG Laboratory Quality Assurance Committee, Richards, S., Aziz, N., Bale, S., Bick, D., Das, S., Gastier-Foster, J., Grody, W.W., Hegde, M., Lyon, E., et al. (2015). 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*, 405–423.
12. Suver, C., Thorogood, A., Doerr, M., Wilbanks, J., and Knoppers, B. (2020). Bringing Code to Data: Do Not Forget Governance. J Med Internet Res *22*, e18087.
13. Schulz, W., Durant, T., Siddon, A., and Torres, R. (2016). Use of application containers

and workflows for genomic data analysis. J Pathol Inform *7*, 53.
14. Matthias, K., and Kane, S.P. (2015). Docker: up and running (Sebastopol, CA: O'Reilly Media, Inc).
15. Kurtzer, G.M., Sochat, V., and Bauer, M.W. (2017). Singularity: Scientific containers for mobility of compute. PLoS ONE *12*, e0177459.
16. Toomey, D. (2017). Jupyter for Data Science: exploratory analysis, statistical modeling, machine learning, and data visualization with Jupyter (Birmingham Mumbai: Packt).
17. O'Connor, B.D., Yuen, D., Chung, V., Duncan, A.G., Liu, X.K., Patricia, J., Paten, B., Stein, L., and Ferretti, V. (2017). The Dockstore: enabling modular, community-focused sharing of Docker-based genomics tools and workflows. F1000Res *6*, 52.
18. Momozawa, Y., Iwasaki, Y., Parsons, M.T., Kamatani, Y., Takahashi, A., Tamura, C., Katagiri, T., Yoshida, T., Nakamura, S., Sugano, K., et al. (2018). Germline pathogenic variants of 11 breast cancer genes in 7,051 Japanese patients and 11,241 controls. Nat Commun *9*, 4083.
19. Feng, B.-J. (2017). PERCH: A Unified Framework for Disease Gene Prioritization: HUMAN MUTATION. Human Mutation *38*, 243–251.
20. Harrison, S.M., Biesecker, L.G., and Rehm, H.L. (2019). Overview of Specifications to the ACMG/AMP Variant Interpretation Guidelines. Current Protocols in Human Genetics *103*.
21. Tian, Y., Pesaran, T., Chamberlin, A., Fenwick, R.B., Li, S., Gau, C.-L., Chao, E.C., Lu, H.-M., Black, M.H., and Qian, D. (2019). REVEL and BayesDel outperform other in silico meta-predictors for clinical variant classification. Sci Rep *9*, 12752.
22. Auerbach, A.D. (2009). Fanconi anemia and its diagnosis. Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis *668*, 4–10.
23. Tavtigian, S.V. (2005). Comprehensive statistical study of 452 BRCA1 missense substitutions with classification of eight recurrent substitutions as neutral. Journal of Medical Genetics *43*, 295–305.
24. Cline, M.S., Liao, R.G., Parsons, M.T., Paten, B., Alquaddoomi, F., Antoniou, A., Baxter, S., Brody, L., Cook-Deegan, R., Coffin, A., et al. (2018). BRCA Challenge: BRCA Exchange as a global resource for variants in BRCA1 and BRCA2. PLoS Genet *14*, e1007752.
25. Exome Aggregation Consortium, Lek, M., Karczewski, K.J., Minikel, E.V., Samocha, K.E., Banks, E., Fennell, T., O'Donnell-Luria, A.H., Ware, J.S., Hill, A.J., et al. (2016). Analysis of protein-coding genetic variation in 60,706 humans. Nature *536*, 285–291.
26. MMomozawa, Y., Iwasaki, Y., Hirata, M., Liu, X., Kamatani, Y., Takahashi, A., Sugano, K., Yoshida, T., Murakami, Y., Matsuda, K., et al. (2020). Germline Pathogenic Variants in 7636 Japanese Patients With Prostate Cancer and 12 366 Controls. JNCI: Journal of the National Cancer Institute *112*, 369–376.
27. Robinson, P.N., Köhler, S., Bauer, S., Seelow, D., Horn, D., and Mundlos, S. (2008).

The Human Phenotype Ontology: A Tool for Annotating and Analyzing Human Hereditary Disease. The American Journal of Human Genetics *83*, 610–615.

28. Schriml, L.M., and Mitraka, E. (2015). The Disease Ontology: fostering interoperability between biological and clinical human disease-related data. Mamm Genome *26*, 584–589.
29. Shefchek, K.A., Harris, N.L., Gargano, M., Matentzoglu, N., Unni, D., Brush, M., Keith, D., Conlin, T., Vasilevsky, N., Zhang, X.A., et al. (2020). The Monarch Initiative in 2019: an integrative data and analytic platform connecting phenotypes to genotypes across species. Nucleic Acids Research *48*, D704–D715.
30. Gaff, C.L., M. Winship, I., M. Forrest, S., P. Hansen, D., Clark, J., M. Waring, P., South, M., and H. Sinclair, A. (2017). Preparing for genomic medicine: a real world demonstration of health system change. Npj Genomic Med *2*, 16.

Data and Code Availability

The workflows are available at Dockstore at https://dockstore.org/workflows/github.com/BRCAChallenge/federated-analysis/cooccurrence:master, and the source code is available on GitHub at https://github.com/BRCAChallenge/federated-analysis.

Supplemental Information

There are 3 independent reports generated from our software solution: the variant co-occurrence and allele frequency report, the optional data quality report, and the genotype-phenotype intersection report. This section defines the detailed configuration for and output of each of these reports with an associated example.

Requirements

To run our container, Docker must be installed in the runtime environment at the institution where the data are stored. We tested our container on Docker versions 18.03 and 19.03. The container also requires the appropriate ClinVar VCF file (for GRCh37 or GRCh38) which can be downloaded from their HTTP or FTP site (`https://ftp.ncbi.nlm.nih.gov/pub/clinvar/`). We used the `bcftools` command to reduce the size of this file to include only the gene(s) of interest. Last, the container requires the gnomAD sites VCF file which can be downloaded from their HTTP site (`https://gnomad.broadinstitute.org/downloads`). Again, we used the `bcftools` command to reduce the size of this file to include only the genes of interest.

Variant co-occurrence and allele frequency report

We use the variant co-occurrence and allele frequency report for *BRCA1* and *BRCA2*, but it has been generalized to find co-occurrences on other genes. Users can specify which version of the human genome (37 or 38), the chromosome and the gene on which to find VUS co-occurring *in trans* with themselves or with known pathogenic variants. The software runs on both phased and un-phased data, though inferring the genotype phase from un-phased data requires VCEP expertise. To determine variant classification, users must provide a delimited file with the following fields: `Clinical_significance` and `Genomic_Coordinate_hg37` (or `Genomic_Coordinate_hg38`). The possible strings for the `Clinical_significance` field are shown in the table below. We leveraged ClinVar variant classifications and Gnomad for allele frequencies.

pathogenicity	`Clinical_significance` values
Pathogenic	`'Pathogenic'`, `'Likely pathogenic'`, `'Likely_pathogenic'`, `'Pathogenic/Likely_pathogenic'`
Benign	`'Benign'`, `'Likely benign'`, `'Likely_benign'`, `'Benign/Likely_benign'`

Uncertain significance	`'Uncertain significance',`
	`'Uncertain_significance',`
	`'Conflicting_interpretations_of_pathogenicity',`
	`'-'`

Variants with pathogenicity values which are not defined in the tab-delimited file are considered VUS. Genomic coordinates must have the form of this example variant: "chr13:g.32314514:C>T", where this represents the variant on chromosome 13, position 32314514 which changes a C nucleotide to a T nucleotide.

Data quality report

The data quality report supports multiple types of field values, including categorical, numerical, and "free-form". The report also validates field values if defined in the `fieldValues` list. Any values found in the pathology data file not in this list are flagged in the report. If this list is left empty, then the report doesn't perform this validation step. In our research, we used the data quality report to make sure that the data values fell within an acceptable range.

Genotype-phenotype intersection report

If the optional phenotype data file is provided as a tab-separated file whose 'ID' fields match those used in the VCF file, then our software will intersect the VUS with the phenotype data.

Examples

In this section, we show 3 examples of using our software: one with associated phenotype data on BRCA2, one without associated phenotype data on BRCA1, and one with associated phenotype data on MYH7.

Example 1: BRCA2 variants with associated phenotype (pathology) data

The following is a basic example which we will use to generate a co-occurrence and allele frequency report, a data quality report, and a phenotype-genotype intersection report.

1. Examine the `brca2.vcf` VCF file in the `examples/BRCA2/data` directory.

```
##fileformat=VCFv4.2
#CHROM  POS     ID      REF     ALT     QUAL    FILTER  INFO    FORMAT  01      02      03      04
chr13   32355250        .       T       C       .       .       .       GT      1|0     0|0     0|0     0|0
chr13   32316508        .       GAC     G       .       .       .       GT      0|1     0|0     0|0     0|0
chr13   32353470        .       A       C       .       .       .       GT      0|0     1|0     0|0     0|0
chr13   32340836        .       GACAA   G       .       .       .       GT      0|0     0|1     0|0     0|0
chr13   32353519        .       A       G       .       .       .       GT      0|0     0|0     1|0     0|0
chr13   32338749        .       AATTAC  A       .       .       .       GT      0|0     0|0     0|1     0|0
chr13   32355250        .       T       C       .       .       .       GT      0|0     0|0     0|0     1|1
```

2. Examine the associated phenotype file `brca2-pathology.tsv` in the `examples/BRCA2/data` directory.

```
ID      Family history / breast cancer  Age at onset    ER      PgR     HER2    CarrierGene
01      1       57      Positive        Positive        3+      BRCA2
02      0       51      Negative        Positive        1+      BRCA2
03      0       66      Negative        Negative        Negative        BRCA2
04      0       0       NA      NA      NA      NonCarrier
```

3. Examine the BRCA2 variant pathogenicity file `clinvar_brca2.tsv` in the `examples/BRCA2/data` directory.

```
Clinical_significance   Genomic_Coordinate_hg37 Genomic_Coordinate_hg38
Pathogenic      -       chr13:32316508:GAC>G
Pathogenic      -       chr13:32340836:GACAA>G
Pathogenic      -       chr13:32338749:AATTAC>A
-       -       chr13:32355250:T>C
-       -       chr13:32355250:T>C
-       -       chr13:32353470:A>C
-       -       chr13:32353519:A>G
```

4. Examine the data quality report `examples/BRCA2/config/brca2-report-config.json` JSON configuration file.

```
{
  "qualityReport": "data/BRCA2-data-quality-report.txt",
  "pathologyReport": "/dev/null",
  "fileName": "data/brca2-pathology.tsv",
  "fileHeader": "True",
  "fieldDelimiter": "\t",
  "printConfigFileInfo": "False",
  "printBadValues": "True",
  "suppressAllOutput": "False",
  "RScriptPath": "/usr/bin/Rscript",
  "fieldFilters": [
    {"fieldName": "ER", "fieldType": "categorical",
        "fieldValues": ["Positive", "Negative", "NA"], "printFieldCount": "True"},
    {"fieldName": "PgR", "fieldType": "categorical",
        "fieldValues": ["Positive", "Negative", "NA"], "printFieldCount": "True"},
    {"fieldName": "HER2", "fieldType": "categorical",
        "fieldValues": ["0", "1+", "2+", "3+", "NA"], "printFieldCount": "True"},
    {"fieldName": "Age at onset", "fieldType": "numerical",
        "fieldValues": {"low": 10, "high": 150}, "printFieldCount": "True", "printStats": "True"},
    {"fieldName": "Family history / breast cancer", "fieldType": "numerical",
        "fieldValues": [0, 1], "printFieldCount": "True", "printStats": "True"},
    {"fieldName": "CarrierGene", "fieldType": "categorical",
        "fieldValues": [], "printFieldCount": "True"}]
}
```

5. Examine the `gnomad_chr13_brca2.tsv` file in the `examples/BRCA2/data` directory. Note the screenshot was truncated as the contents of the file are too large to display here.

```
chr13   32316508        rs80359483      GAC     G       .       PASS    AC=1;AN=152032;AF=6.57756e-06;popma
x=nfe;faf95_popmax=0
```

6. Run the report workflow.

```
./runMe.sh \
-c 13 \
-p True \
-dd $(pwd)/examples/BRCA2/data \
-cd $(pwd)/examples/BRCA2/config \
-vf brca2.vcf \
-vpf clinvar_brca2.tsv \
-gf gnomad_chr13_brca2.vcf \
-rc brca2-report-config.json \
-g BRCA2 \
-spf brca2-pathology.tsv \
-hg 38
```

Here `-c` is the chromosome number, `-p` is whether the data is phased, `-dd` is the absolute path to the data directory, `-cd` is the absolute path to the configuration directory, `-vf` is the name of the VCF file, `-gf` is the name of the pathogenicity file in the `data` directory, `-rc` is the name of the report configuration file in the `config` directory, `-g` is the name of the gene, `-spf` is the name of the sample phenotype file in the `data` directory, `-gf` is the name of the gnomAD allele frequency file in the `data` directory, and `-hg` is the version of the human genome reference.

7. Examine the co-occurrence and allele frequency report.


```
{
    "cooccurring vus": {
        "(13, 32355250, 'T', 'C')": {
            "likelihood data": { "p1": 0.375, "p2": 0.001, "n": 2, "k": 1,
                "likelihood": 0.0042624 },
            "allele frequencies": { "maxPop": null, "maxPopFreq": 0.0,
                "minPop": null, "minPopFreq": 1.0, "cohortFreq": 0.5 },
            "pathogenic variants": [ [ 13, 32316508, "GAC", "G" ] ]
        },
        "(13, 32353470, 'A', 'C')": {
            "likelihood data": { "p1": 0.375, "p2": 0.001, "n": 1, "k": 1,
                "likelihood": 0.0026666666666666666 },
            "allele frequencies": { "maxPop": null, "maxPopFreq": 0.0,
                "minPop": null, "minPopFreq": 1.0, "cohortFreq": 0.25 },
            "pathogenic variants": [ [ 13, 32340836, "GACAA", "G" ] ]
        },
        "(13, 32353519, 'A', 'G')": {
            "likelihood data": { "p1": 0.375, "p2": 0.001, "n": 1, "k": 1,
                "likelihood": 0.0026666666666666666 },
            "allele frequencies": { "maxPop": null, "maxPopFreq": 0.0,
                "minPop": null, "minPopFreq": 1.0, "cohortFreq": 0.25 },
            "pathogenic variants": [ [ 13, 32338749, "AATTAC", "A" ] ]
        }
    },
    "homozygous vus": {
        "(13, 32355250, 'T', 'C')": { "count": 1, "maxPop": null, "maxPopFreq": 0.0,
                "minPop": null, "minPopFreq": 1.0, "cohortFreq": 0.25 }
    }
}
```

8. Examine the data quality report.

```
=============================================
total records read from data file: 4
=============================================
column: ER / type: categorical
{
    "fieldCount": { "NA": 1, "Negative": 2, "Positive": 1 }
}
=============================================
column: PgR / type: categorical
{
    "fieldCount": { "NA": 1, "Negative": 1, "Positive": 2 }
}
=============================================
column: HER2 / type: categorical
{
    "fieldCount": { "1+": 1, "3+": 1, "NA": 1 }
}
=============================================
column: Age at onset / type: numerical
{
    "fieldCount": { "51": 1, "57": 1, "66": 1 }
}
min = 51, max = 66, mean = 58 median = 57 stdev = 7.54983443527075
=============================================
column: Family history / breast cancer / type: numerical
{
    "fieldCount": { "0": 3, "1": 1 }
}
min = 0, max = 1, mean = 0.25 median = 0.0 stdev = 0.5
=============================================
column: CarrierGene / type: categorical
{
    "fieldCount": { "BRCA2": 3, "NonCarrier": 1 }
}
=============================================
bad values: {2: [('HER2', 'Negative')], 3: [('Age at onset', 0)]}
=============================================
missing values: {}
```

9. Examine the phenotype-genotype intersection report.

```
{
    "cooccurring": {
        "(13, 32353470, 'A', 'C')": {
            "phenotype": [ { "Age at onset": [ 51 ], "CarrierGene": [ "BRCA2" ],
                        "ER": [ "Negative" ], "Family history / breast cancer": [ 0 ],
                        "HER2": [ "1+" ], "ID": [ 2 ], "PgR": [ "Positive" ] } ]
        },
        "(13, 32353519, 'A', 'G')": {
            "phenotype": [ { "Age at onset": [ 66 ], "CarrierGene": [ "BRCA2" ],
                        "ER": [ "Negative" ], "Family history / breast cancer": [ 0 ],
                        "HER2": [ "Negative" ], "ID": [ 3 ], "PgR": [ "Negative" ] } ]
        },
        "(13, 32355250, 'T', 'C')": {
            "phenotype": [ { "Age at onset": [ 57 ], "CarrierGene": [ "BRCA2" ],
                        "ER": [ "Positive" ], "Family history / breast cancer": [ 1 ],
                        "HER2": [ "3+" ], "ID": [ 1 ], "PgR": [ "Positive" ] } ]
        }
    },
    "homozygous": {
        "(13, 32355250, 'T', 'C')": {
            "phenotype": [ { "Age at onset": [ 0 ], "CarrierGene": [ "NonCarrier" ],
                        "ER": [ NaN ], "Family history / breast cancer": [ 0 ],
                        "HER2": [ NaN ], "ID": [ 4 ], "PgR": [ NaN ] } ] }
    },
    "numMissing": 0
}
```

Example 2: BRCA1 variants without associated phenotype data

The following is a basic example which we will use to generate a co-occurrence and allele frequency report.

1. Examine the `brca1.vcf` VCF file in the `examples/BRCA1/data` directory.

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO	FORMAT	01	02	03	
17	43074403		.	C	A	.	.	.	GT	0|0	0|0	1|0
17	43088049		.	G	A	.	.	.	GT	1|0	0|1	1|1
17	43110245		.	T	G	.	.	.	GT	0|0	1|1	0|0
17	43124027		.	ACT	A	.	.	.	GT	0|1	1|0	0|0

2. Examine the BRCA1 variant pathogenicity file `brca1-pathogenicity.tsv` in the `examples/BRCA1/data` directory.

Clinical_significance		Genomic_Coordinate_hg37	Genomic_Coordinate_hg38
Pathogenic	-	chr17:43074403:C>A	
Pathogenic	-	chr17:43124027:ACT>A	
-	-	chr17:43088049:G>A	
-	-	chr17:43110245:T>G	

3. Run the report workflow.

```
./runMe.sh \
-c 17 \
-p True \
-dd $(pwd)/examples/BRCA1/data \
-cd $(pwd)/examples/BRCA1/config \
-vf brca1.vcf \
-vpf clinvar_brca1.tsv \
-gf gnomad_chr17_brca1.vcf \
-g BRCA1 \
-hg 38
```

4. Examine the co-occurrence and allele frequency report in the `data` directory.

```
{
    "cooccurring vus": {
        "(14, 23413860, 'G', 'A')": {
            "likelihood data": { "p1": 0.5, "p2": 0.001, "n": 2, "k": 2, "likelihood": 4e-06
            },
            "allele frequencies": { "maxPop": "nfe", "maxPopFreq": "4.88e-06", "cohortFreq": 0.5
            },
            "pathogenic variants": [ [ 14, 23413847, "T", "A" ] ]
        },
        "(14, 23413889, 'T', 'C')": {
            "likelihood data": { "p1": 0.5, "p2": 0.001, "n": 2, "k": 2, "likelihood": 4e-06
            },
            "allele frequencies": { "maxPop": "afr", "maxPopFreq": "1.919e-05", "cohortFreq": 0.5
            },
            "pathogenic variants": [ [ 14, 23413847, "T", "A" ], [ 14, 23414002, "C", "G" ] ]
        },
        "(14, 23413831, 'T', 'G')": {
            "likelihood data": { "p1": 0.5, "p2": 0.001, "n": 3, "k": 3, "likelihood": 8e-09
            },
            "allele frequencies": { "maxPop": "nfe", "maxPopFreq": "0.00018197", "cohortFreq": 0.75
            },
            "pathogenic variants": [ [ 14, 23414002, "C", "G" ], [ 14, 23413847, "T", "A" ] ]
        }
    },
    "homozygous vus": {
        "(14, 23413889, 'T', 'C')": { "count": 1, "maxPop": "afr", "maxPopFreq": "1.919e-05", "cohortFreq"
: 0.25
        },
        "(14, 23413831, 'T', 'G')": { "count": 1, "maxPop": "nfe", "maxPopFreq": "0.00018197", "cohortFreq
": 0.25
        }
    }
}
```

Example 3: MYH7 variants with associated phenotype (cardiac) data

The following is a basic example which we will use to generate a data quality report, a co-occurrence and allele frequency report, and a cardiac report. Note that since we have an associated phenotype file, we will also generate a fourth report which intersects VUS with cardiac data.

1. Examine the `myh7.vcf` VCF file in the `examples/MYH7/data` directory.

```
##fileformat=VCFv4.2
#CHROM  POS        ID  REF  ALT  QUAL  FILTER  INFO  FORMAT  01   02   03   04
14      23413831   .   T    G    .     .       .     GT      0|0  1|0  0|1  1|1
14      23413860   .   G    A    .     .       .     GT      1|0  0|0  0|0  0|1
14      23413889   .   T    C    .     .       .     GT      1|1  0|0  1|0  0|0
14      23413847   .   T    A    .     .       .     GT      0|1  0|0  0|0  1|0
14      23414002   .   C    G    .     .       .     GT      1|0  0|1  1|0  0|0
```

2. Examine the associated phenotype file `myh7-cardiac.tsv` in the `examples/MYH7/data` directory.

ID	Age at onset	Native T1 (msec)	Technique	ECV
01	57	976	MOLLI	25.9
02	51	1159	ShMOLLI	26.9
03	66	1122	MOLLI	26.4
04	0	994	MOLLI	25.4

3. Examine the ClinVar pathogenicity file `gnomad_myh7.tsv` in the `examples/MYH7/data` directory.

```
Clinical_significance   Genomic_Coordinate_hg37 Genomic_Coordinate_hg38
Pathogenic      -       chr14:23413847:T>A
Pathogenic      -       chr14:23414002:C>G
Conflicting interpretations of pathogenicity    -       chr14:23413831:T>G
Uncertain significance  -       chr14:23413860:G>A
Uncertain significance  -       chr14:23413889:T>C
Likely pathogenic       -       chr14:23413890:CC>C
```

4. Examine the data quality report `examples/MYH7/config/myh7-report-config.json` JSON configuration file.

```
{
  "qualityReport": "data/MYH7-data-quality-report.txt",
  "pathologyReport": "data/MYH7-cardiac-report.txt",
  "fileName": "data/myh7-cardiac.tsv",
  "fileHeader": "True",
  "fieldDelimiter": "\t",
  "printConfigFileInfo": "False",
  "printBadValues": "True",
  "suppressAllOutput": "False",
  "RScriptPath": "/usr/bin/Rscript",
  "fieldFilters": [
    {"fieldName": "Technique", "fieldType": "categorical",
        "fieldValues": ["MOLLI", "ShMOLLI", "NA"], "printFieldCount": "True"},
    {"fieldName": "Age at onset", "fieldType": "numerical",
        "fieldValues": {"low": 10, "high": 120}, "printFieldCount": "True", "printStats": "True"
},
    {"fieldName": "Native T1 (msec)", "fieldType": "numerical",
        "fieldValues": [], "printFieldCount": "True", "printStats": "True"},
    {"fieldName": "ECV", "fieldType": "numerical",
        "fieldValues": [], "printFieldCount": "True", "printStats": "True"}]
}
```

5. Examine the `gnomad_chr14_myh7.tsv` file in the `examples/MYH7/data` directory. Note the screenshot was truncated as the contents of the file are too large to display here.

```
chr14   23413831        rs45523233      T       G       .       PASS    AC=23;AN=152258;AF=0.000151059;popm
ax=nfe;faf95_popmax=0.00018197;
```

6. Run the report workflow.

```
./runMe.sh \
-c 14 \
-p True \
-dd $(pwd)/examples/MYH7/data \
-cd $(pwd)/examples/MYH7/config \
-vf myh7.vcf \
-vpf clinvar_myh7.tsv \
-gf gnomad_chr14_myh7.vcf \
-rc myh7-report-config.json \
-g MYH7 \
-spf myh7-cardiac.tsv \
-hg 38
```

7. Examine the data quality report.

```
==============================================
total records read from data file: 4
==============================================
column: Technique / type: categorical
{
    "fieldCount": { "MOLLI": 3, "ShMOLLI": 1 }
}
==============================================
column: Age at onset / type: numerical
{
    "fieldCount": { "51": 1, "57": 1, "66": 1 }
}
min = 51, max = 66, mean = 58 median = 57 stdev = 7.54983443527075
==============================================
column: Native T1 (msec) / type: numerical
{
    "fieldCount": { "976": 1, "994": 1, "1122": 1, "1159": 1 }
}
min = 976, max = 1159, mean = 1062.75 median = 1058.0 stdev = 91.33591845489921
==============================================
column: ECV / type: numerical
{
    "fieldCount": { "25.4": 1, "25.9": 1, "26.4": 1, "26.9": 1 }
}
min = 25.4, max = 26.9, mean = 26.15 median = 26.15 stdev = 0.6454972243679028
==============================================
bad values: {3: [('Age at onset', 0)]}
==============================================
missing values: {}
==============================================
```

8. Examine the co-occurrence and allele frequency report.


```
{
    "cooccurring vus": {
        "(14, 23413860, 'G', 'A')": {
            "likelihood data": { "p1": 0.5, "p2": 0.001, "n": 2, "k": 2, "likelihood": 4e-06
            },
            "allele frequencies": { "maxPop": "nfe", "maxPopFreq": "4.88e-06", "cohortFreq": 0.5
            },
            "pathogenic variants": [ [ 14, 23413847, "T", "A" ] ]
        },
        "(14, 23413889, 'T', 'C')": {
            "likelihood data": { "p1": 0.5, "p2": 0.001, "n": 2, "k": 2, "likelihood": 4e-06
            },
            "allele frequencies": { "maxPop": "afr", "maxPopFreq": "1.919e-05", "cohortFreq": 0.5
            },
            "pathogenic variants": [ [ 14, 23413847, "T", "A" ], [ 14, 23414002, "C", "G" ] ]
        },
        "(14, 23413831, 'T', 'G')": {
            "likelihood data": { "p1": 0.5, "p2": 0.001, "n": 3, "k": 3, "likelihood": 8e-09
            },
            "allele frequencies": { "maxPop": "nfe", "maxPopFreq": "0.00018197", "cohortFreq": 0.75
            },
            "pathogenic variants": [ [ 14, 23414002, "C", "G" ], [ 14, 23413847, "T", "A" ] ]
        }
    },
    "homozygous vus": {
        "(14, 23413889, 'T', 'C')": { "count": 1, "maxPop": "afr", "maxPopFreq": "1.919e-05", "cohortFreq"
: 0.25
        },
        "(14, 23413831, 'T', 'G')": { "count": 1, "maxPop": "nfe", "maxPopFreq": "0.00018197", "cohortFreq
": 0.25
        }
    }
}
```

9. Examine the phenotype-genotype intersection report.


```
{
    "cooccurring": {
        "(14, 23413831, 'T', 'G')": {
            "phenotype": [
                { "Age at onset": [ 66 ], "ECV": [ 26.4 ], "Native T1 (msec)": [ 1122 ],
                    "Technique": [ "MOLLI" ] },
                { "Age at onset": [ 51 ], "ECV": [ 26.9 ], "Native T1 (msec)": [ 1159 ],
                    "Technique": [ "ShMOLLI" ] } ] },
        "(14, 23413860, 'G', 'A')": {
            "phenotype": [
                { "Age at onset": [ 57 ], "ECV": [ 25.9 ], "Native T1 (msec)": [ 976 ],
                    "Technique": [ "MOLLI" ] },
                { "Age at onset": [ 0 ], "ECV": [ 25.4 ], "Native T1 (msec)": [ 994 ],
                    "Technique": [ "MOLLI" ] } ] },
        "(14, 23413889, 'T', 'C')": {
            "phenotype": [
                { "Age at onset": [ 66 ], "ECV": [ 26.4 ], "Native T1 (msec)": [ 1122 ],
                    "Technique": [ "MOLLI" ] } ] } },
    "homozygous": {
        "(14, 23413831, 'T', 'G')": {
            "phenotype": [
                { "Age at onset": [ 0 ], "ECV": [ 25.4 ], "Native T1 (msec)": [ 994 ],
                    "Technique": [ "MOLLI" ] } ] },
        "(14, 23413889, 'T', 'C')": {
            "phenotype": [
                { "Age at onset": [ 57 ], "ECV": [ 25.9 ], "Native T1 (msec)": [ 976 ],
                    "Technique": [ "MOLLI" ] } ] } },
    "numMissing": 0
}
```