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

Authors: James Casaletto[1], Michael Parsons[2], 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:

Corresponding Author Email: jcasalet@ucsc.edu, mcline@ucsc.edu

Summary

More than 40% of the germline variants in ClinVar today (May, 2021) are designated as Variants of Uncertain Significance (VUS). That is, there is insufficient evidence to determine the clinical impact of these variants, which confounds the clinical management of the individuals who carry them. These variants remain unclassified in part because the patient-level data needed for their interpretation is largely siloed, due to its sensitive nature. Federated analysis offers the potential to 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 patients and controls from clinical data at RIKEN, derived from the BioBank Japan repository. We used as exemplars variants in *BRCA1* and *BRCA2*, genes for which variants designated as pathogenic confer significant risk of breast, ovarian, and other cancers. By sharing analysis software workflows, 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 serves as a proof of concept, and 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. 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.

Introduction

One obvious 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 risk of female breast and ovarian cancer (Kuchenbaecker et al. 2017), and confer significant risk of pancreatic, prostate and male breast cancer (Castro et al. 2013;

Vietri et al. 2020; Mersch et al. 2015; Ruddy and Winer 2013). 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 (Pilié et al. 2019; Tuffaha et al. 2018). 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 (Landrum et al. 2018), 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) (Karczewski et al. 2020) 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 (Kurian 2010), a disparity that stems largely from historical biases in genetic studies (Landry et al. 2018; Sirugo, Williams, and Tishkoff 2019).

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 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 (Harris and Wyndham 2015). Yet sharing data on rare genetic variants is critical for the advancement of precision medicine, as advocated by organizations including the GA4GH (Siu et al. 2016), the American College of Molecular Geneticists (ACMG) (ACMG Board Of Directors 2017) and the Wellcome Trust (Wright et al. 2019). 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 (Richards et al. 2015), 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 (Suver et al. 2020). 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 (Schulz et al. 2016). These technologies include Docker (Turnbull 2014), Singularity (Kurtzer, Sochat, and Bauer 2017) and Jupyter (Toomey 2017). Containers and workflows can be shared on the Dockstore platform (O'Connor et al. 2017) so that multiple institutions can execute exactly the 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 (Momozawa et al. 2018). 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.

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 reusable. The analysis software creates multiple reports that capture data quality, tumor pathology, 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 the number of missing or unexpected data values. For this report, we expose a 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 represents a general solution which can be reused for other data sets.

In addition to this generic, domain-agnostic data quality report, we provide hooks to call a custom, domain-specific report which can be leveraged to identify data anomalies in a known domain. 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 contains sample identifiers. Even though these sample identifiers are anonymized, we did not want to risk exposing any identifier in the results. Asking our collaborators to review the results before making them available to us is prone to human error and does not account for more subtle and nefarious ways of exporting privacy-sensitive data. Accordingly, encoded our pathology report software in Structured Query Language (SQL) which

is the most prolific language used in data analysis. SQL, as a declarative language, omits low-level programming details, making it easier to read and interpret. SQL concisely declares which columns to use and what operations to perform on them, enabling readers to more readily detect privacy violations. The 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.

The variant frequency and co-occurrence report was written to report on 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. While our research focuses on VUS in *BRCA1* and *BRCA2* genes, the software was written to work with any genes. 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.

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". In order to ascertain whether the analysis is truly both secure and privacy-preserving, an auditor would need to inspect the Dockerfile definition of the container as well as all the software that runs in the container. We mitigated this risk by writing output to local text files which were 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, such as through 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.

Methods

The Dataset: Our analysis revolved around a case-control association studie of individuals of Japanese ancestry: 7,104 breast cancer patients and 23,731 controls (Momozawa et al. 2018). 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.

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 (Feng 2017; Tian et al. 2019).

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 (Richards et al. 2015). While gnomAD is commonly used as source of population frequencies, gnomAD 3.1 contains data from only 2,604 East Asian genomes (Tiao and and Goodrich 2020) while gnomAD 2.1 contains data from 9,977 exomes (Francioli et al. 2018); gnomAD 2.1 contained 76 Japanese exomes, while the number of Japanese genomes in gnomAD 3.1 is unknown. So 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 (Harrison, Biesecker, and Rehm 2019). 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 (Parsons, Tudini, and Spurdle 2020; Parsons and Spurdle 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 computation 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 (Feng 2017), a meta-predictor that has been shown to outperform most others (Tian et al. 2019). 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, Tudini, and Spurdle 2020; Parsons and Spurdle

2021).

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 (Auerbach 2009). 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 (Tavtigian et al. 2006)*. 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*) (Parsons, Tudini, and Spurdle 2020; Parsons and Spurdle 2021).

Analysis Approach: We created a Docker container 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 (i.e. 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. In order to identify VUS, we checked the classifications provided by the ClinGen-approved ENIGMA expert panel in BRCA Exchange (Cline et al. 2018), 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 (Momozawa et al. 2018). 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 (Momozawa et al. 2018), using the values from the tumor pathology report. 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.

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*) on the basis of 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 on the basis of 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 (Karczewski et al. 2020), 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) (Lek et al. 2016), a fact that itself adds confidence to the BioBank Japan observations. So 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*) on the basis of 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 on the basis of 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 on the basis of their gnomAD population frequencies while a fifth is absent from gnomAD. So 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 in 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

In summary, 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 show feasibility to capture relevant information on variant co-occurrence and age at presentation, results which aligned with the interpretations based on bioinformatics and frequency; these analyses would not be possible in gnomAD, which conveys neither *in trans* variant co-occurrences nor patient age. 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 not available or applicable to the gnomAD resource, and are generally provided only in summary level data presentations from published cohorts.

In principle, the gnomAD resource could grow with time to comprehensively represent all global populations. In practice, gnomAD mostly imports data from cohorts that were sequenced at the Broad, due to the high cost of reprocessing data that was sequenced elsewhere (Rehm 2020). Even if additional data could be shared directly with the gnomAD project, integrating those data into gnomAD might not be viable given the reprocessing cost. Consequently, there is currently valuable information in international sequencing projects that are not likely to be integrated into gnomAD, but could in principle be leveraged today for variant interpretation. 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). For capturing the genetic diversity of global populations that are currently underrepresented in genetic knowledgebases, collecting evidence directly from international sources is arguably more expedient than waiting for those populations to be characterized in sequencing studies at an American research institution. Where traditional data sharing is blocked by barriers including laws that prohibit exporting genomic sequences, federated analysis can advance data sharing. Additionally, federated analysis allows greater control in the data collection; for example the

ability to filter by participant age and phenotype, used here to infer absence of debilitating early onset disease.

One might be surprised that an analysis of roughly 30,000 Japanese individuals revealed only 19 VUS. This is explained by the fact that these samples were analyzed by the RIKEN and ENIGMA teams in previous studies, and most variants observed in those studies were interpreted (Momozawa et al. 2018, 2020). Previously, these variant interpretations were informed by case-control odds ratios and a previous, more-stringent frequency-based classification criteria (ENIGMA Consortium 2017). This federated analysis allowed us to revisit these data to apply updated frequency-based classification criteria, and to collect additional forms of variant co-occurrence data as an additional form of evidence.

The main limitation of our approach is that it requires getting data into the particular format that our software recognizes, namely a TSV file and a VCF file. 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 further 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, 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, 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 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

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.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 19 (7). https://doi.org/10.1038/gim.2016.196.

Auerbach, Arleen D. 2009. “Fanconi Anemia and Its Diagnosis.” *Mutation Research* 668 (1-2): 4–10.

Castro, Elena, Chee Goh, David Olmos, Ed Saunders, Daniel Leongamornlert, Malgorzata Tymrakiewicz, Nadiya Mahmud, et al. 2013. “Germline BRCA Mutations Are Associated with Higher Risk of Nodal Involvement, Distant Metastasis, and Poor Survival Outcomes in Prostate Cancer.” *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* 31 (14): 1748–57.

Cline, Melissa S., Rachel G. Liao, Michael T. Parsons, Benedict Paten, Faisal Alquaddoomi, Antonis Antoniou, Samantha Baxter, et al. 2018. “BRCA Challenge: BRCA Exchange as a Global Resource for Variants in BRCA1 and BRCA2.” *PLoS Genetics* 14 (12): e1007752.

ENIGMA Consortium. 2017. “ENIGMA BRCA1/2 Gene Variant Classification Criteria.” ENIGMA Consortium. June 29, 2017. https://enigmaconsortium.org/wp-content/uploads/2020/08/ENIGMA_Rules_2017-06-29-v2_5_1.pdf.

Feng, Bing-Jian. 2017. “PERCH: A Unified Framework for Disease Gene Prioritization.” *Human Mutation* 38 (3): 243–51.

Francioli, Laurent, Grace Tiao, Konrad Karczewski, Matthew Solomonson, and Nick Watts. 2018. “gnomAD v2.1.” *gnomAD Blog* (blog). October 17, 2018. https://gnomad.broadinstitute.org/blog/2018-10-gnomad-v2-1/.

Harrison, Steven M., Leslie G. Biesecker, and Heidi L. Rehm. 2019. “Overview of Specifications to the ACMG/AMP Variant Interpretation Guidelines.” *Current Protocols in Human Genetics / Editorial Board, Jonathan L. Haines ... [et Al.]* 103 (1): e93.

Harris, T. L., and J. M. Wyndham. 2015. “Data Rights and Responsibilities: A Human Rights Perspective on Data Sharing.” *Journal of Empirical Research on Human Research Ethics: JERHRE* 10 (3). https://doi.org/10.1177/1556264615591558.

Karczewski, Konrad J., Laurent C. Francioli, Grace Tiao, Beryl B. Cummings, Jessica Alföldi, Qingbo Wang, Ryan L. Collins, et al. 2020. “The Mutational Constraint Spectrum Quantified from Variation in 141,456 Humans.” *Nature* 581 (7809): 434–43.

Kuchenbaecker, Karoline B., John L. Hopper, Daniel R. Barnes, Kelly-Anne Phillips, Thea M. Mooij, Marie-José Roos-Blom, Sarah Jervis, et al. 2017. “Risks of Breast, Ovarian, and Contralateral Breast Cancer for BRCA1 and BRCA2 Mutation Carriers.” *JAMA: The Journal of the American Medical Association* 317 (23): 2402–16.

Kurian, Allison W. 2010. “BRCA1 and BRCA2 Mutations across Race and Ethnicity: Distribution and Clinical Implications.” *Current Opinion in Obstetrics & Gynecology* 22 (1): 72–78.

Kurtzer, Gregory M., Vanessa Sochat, and Michael W. Bauer. 2017. “Singularity: Scientific Containers for Mobility of Compute.” *PloS One* 12 (5): e0177459.

Landrum, Melissa J., Jennifer M. Lee, Mark Benson, Garth R. Brown, Chen Chao, Shanmuga Chitipiralla, Baoshan Gu, et al. 2018. “ClinVar: Improving Access to Variant Interpretations and Supporting Evidence.” *Nucleic Acids Research* 46 (D1): D1062–67.

Landry, Latrice G., Nadya Ali, David R. Williams, Heidi L. Rehm, and Vence L. Bonham. 2018. “Lack Of Diversity In Genomic Databases Is A Barrier To Translating Precision Medicine Research Into Practice.” *Health Affairs* 37 (5): 780–85.

Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O’Donnell-Luria, et al. 2016. “Analysis of Protein-Coding Genetic Variation in 60,706 Humans.” *Nature* 536 (7616): 285–91.

Mersch, Jacqueline, Michelle A. Jackson, Minjeong Park, Denise Nebgen, Susan K. Peterson, Claire Singletary, Banu K. Arun, and Jennifer K. Litton. 2015. “Cancers Associated with BRCA1 and BRCA2 Mutations Other than Breast and Ovarian.” *Cancer* 121 (2): 269–75.

Momozawa, Yukihide, Yusuke Iwasaki, Makoto Hirata, Xiaoxi Liu, Yoichiro Kamatani, Atsushi Takahashi, Kokichi Sugano, et al. 2020. “Germline Pathogenic Variants in 7636 Japanese Patients With Prostate Cancer and 12 366 Controls.” *JNCI: Journal of the National Cancer Institute*. https://doi.org/10.1093/jnci/djz124.

Momozawa, Yukihide, Yusuke Iwasaki, Michael T. Parsons, Yoichiro Kamatani, Atsushi Takahashi, Chieko Tamura, Toyomasa Katagiri, et al. 2018. “Germline Pathogenic Variants of 11 Breast Cancer Genes in 7,051 Japanese Patients and 11,241 Controls.” *Nature Communications* 9 (1): 4083.

O’Connor, Brian D., Denis Yuen, Vincent Chung, Andrew G. Duncan, Xiang Kun Liu, Janice Patricia, Benedict Paten, Lincoln Stein, and Vincent Ferretti. 2017. “The Dockstore: Enabling Modular, Community-Focused Sharing of Docker-Based Genomics Tools and Workflows.” *F1000Research* 6. https://doi.org/10.12688/f1000research.10137.1.

Parsons, Michael, and Amanda Spurdle. 2021. “Summary of the Draft Rules of the ENIGMA BRCA1 and BRCA2 VCEP,” March 12, 2021.

Parsons, Michael, Emma Tudini, and Amanda Spurdle. 2020. “ENIGMA BRCA1 and BRCA2 Variant Curation Expert Panel.” December 1, 2020. https://clinicalgenome.org/affiliation/50087/.

Pilié, Patrick G., Chad Tang, Gordon B. Mills, and Timothy A. Yap. 2019. “State-of-the-Art Strategies for Targeting the DNA Damage Response in Cancer.” *Nature Reviews. Clinical Oncology* 16 (2): 81–104.

Rehm, Heidi. 2020. “Personal Communication on gnomAD Data Integration,” 2020.

Richards, Sue, Nazneen Aziz, Sherri Bale, David Bick, Soma Das, Julie Gastier-Foster, Wayne W. Grody, 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.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 17 (5): 405–24.

Ruddy, K. J., and E. P. Winer. 2013. “Male Breast Cancer: Risk Factors, Biology, Diagnosis, Treatment, and Survivorship.” *Annals of Oncology: Official Journal of the European Society for Medical Oncology / ESMO* 24 (6): 1434–43.

Schulz, Wade L., Thomas J. S. Durant, Alexa J. Siddon, and Richard Torres. 2016. “Use of Application Containers and Workflows for Genomic Data Analysis.” *Journal of Pathology Informatics* 7. https://doi.org/10.4103/2153-3539.197197.

Sirugo, Giorgio, Scott M. Williams, and Sarah A. Tishkoff. 2019. “The Missing Diversity in Human Genetic Studies.” *Cell* 177 (4): 1080.

Siu, L. L., M. Lawler, D. Haussler, B. M. Knoppers, J. Lewin, D. J. Vis, R. G. Liao, et al. 2016. “Facilitating a Culture of Responsible and Effective Sharing of Cancer Genome Data.” *Nature Medicine* 22 (5). https://doi.org/10.1038/nm.4089.

Suver, C., A. Thorogood, M. Doerr, J. Wilbanks, and B. Knoppers. 2020. “Bringing Code to Data: Do Not Forget Governance.” *Journal of Medical Internet Research* 22 (7). https://doi.org/10.2196/18087.

Tavtigian, S. V., A. M. Deffenbaugh, L. Yin, T. Judkins, T. Scholl, P. B. Samollow, D. de Silva, A. Zharkikh, and A. Thomas. 2006. “Comprehensive Statistical Study of 452 BRCA1 Missense Substitutions with Classification of Eight Recurrent Substitutions as Neutral.” *Journal of Medical Genetics* 43 (4): 295–305.

Tian, Yuan, Tina Pesaran, Adam Chamberlin, R. Bryn Fenwick, Shuwei Li, Chia-Ling Gau, Elizabeth C. Chao, Hsiao-Mei Lu, Mary Helen Black, and Dajun Qian. 2019. “REVEL and BayesDel Outperform Other in Silico Meta-Predictors for Clinical Variant Classification.” *Scientific Reports* 9 (1): 12752.

Tiao, Grace, and Julia and Goodrich. 2020. “gnomAD v3.1 New Content, Methods, Annotations, and Data Availability.” *gnomAD Blog* (blog). October 29, 2020. https://gnomad.broadinstitute.org/blog/2020-10-gnomad-v3-1-new-content-methods-annotations-and-data-availability/.

Toomey, Dan. 2017. *Jupyter for Data Science: Exploratory Analysis, Statistical Modeling, Machine Learning, and Data Visualization with Jupyter*. Packt Publishing Ltd.

Tuffaha, Haitham W., Andrew Mitchell, Robyn L. Ward, Luke Connelly, James R. G. Butler, Sarah Norris, and Paul A. Scuffham. 2018. “Cost-Effectiveness Analysis of Germ-Line BRCA Testing in Women with Breast Cancer and Cascade Testing in Family Members of Mutation Carriers.” *Genetics in Medicine: Official Journal of the American College of Medical Genetics* 20 (9): 985–94.

Turnbull, James. 2014. *The Docker Book: Containerization Is the New Virtualization*. James Turnbull.

Vietri, Maria Teresa, Gemma Caliendo, Giovanna D’Elia, Marianna Resse, Amelia Casamassimi, Pellegrino Biagio Minucci, Michele Cioffi, and Anna Maria Molinari. 2020. “BRCA and PALB2 Mutations in a Cohort of Male Breast Cancer with One Bilateral Case.” *European Journal of Medical Genetics* 63 (6): 103883.

Wright, C. F., J. S. Ware, A. M. Lucassen, A. Hall, A. Middleton, N. Rahman, S. Ellard, and H. V. Firth. 2019. “Genomic Variant Sharing: A Position Statement.” *Wellcome Open Research* 4 (February). https://doi.org/10.12688/wellcomeopenres.15090.2.

Supplemental Material

There are 3 independent reports generated from our software solution: data quality report, tumor pathology report, and variant co-occurrence and allele frequency report. This section defines the detailed configuration for and output of each of these reports.

Data quality report

Here is an example of the JSON configuration file for the data quality report:

```
{
  "qualityReport": "data/data-quality-report.txt",
  "pathologyReport": "data/tumor-pathology-report.txt",
  "fileName": "data/shuffle.tsv", "fileHeader": "True",
  "fieldDelimiter": "\t", "printConfigFileInfo": "False",
  "printBadValues": "False", "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": [], "printFieldCount": "True", "printStats": "True"},
    {"fieldName": "Ovarian cancer history", "fieldType": "numerical",
        "fieldValues": [], "printFieldCount": "True", "printStats": "True"},
    {"fieldName": "Family history / breast cancer", "fieldType": "numerical",
        "fieldValues": [0, 1], "printFieldCount": "True", "printStats": "True"},
    {"fieldName": "Family history / pancreatic cancer", "fieldType": "numerical",
        "fieldValues": [0, 1], "printFieldCount": "True", "printStats": "True"},
    {"fieldName": "Family history / ovarian cancer", "fieldType": "numerical",
        "fieldValues": [0, 1], "printFieldCount": "True", "printStats": "True"},
    {"fieldName": "Family history / stomach cancer", "fieldType": "numerical",
        "fieldValues": [0, 1], "printFieldCount": "True", "printStats": "True"},
    {"fieldName": "Family history / liver cancer", "fieldType": "numerical",
        "fieldValues": [0, 1], "printFieldCount": "True", "printStats": "True"},
    {"fieldName": "Family history / bone tumor", "fieldType": "numerical",
        "fieldValues": [0, 1], "printFieldCount": "True", "printStats": "True"},
    {"fieldName": "Family history / bladder cancer", "fieldType": "numerical",
        "fieldValues": [], "printFieldCount": "True", "printStats": "True"},
    {"fieldName": "TNM classification / N", "fieldType": "categorical",
        "fieldValues": ["0", "1", "2", "3", "NA"], "printFieldCount": "True",
        "printStats": "True"},
    {"fieldName": "TNM classification / M", "fieldType": "categorical",
        "fieldValues": ["0", "1", "NA"], "printFieldCount": "True", "printStats": "True"},
    {"fieldName": "Tissue type (3 groups)", "fieldType": "free-form",
        "fieldValues": ["ascii", "utf-8"], "printFieldCount": "True"},
    {"fieldName": "CarrierGene", "fieldType": "categorical",
        "fieldValues": [], "printFieldCount": "True"}]
}
```

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. Here is a sample snippet of output

from the data quality report that shows counts for fields in the pathology file.

```
================================================
total records read from data file: 7051
================================================
column: ER / type: categorical
{
    "fieldCount": {
        "NA": 2200,
        "Negative": 1313,
        "Positive": 3538
    }
}
================================================
column: Age at onset / type: numerical
{
    "fieldCount": {
        "19.0": 1,
        "22.0": 1,
        "23.0": 1,
        "24.0": 3,
        "25.0": 2,
        "26.0": 7,
        ...
        "96.0": 1
    }
}
min = 19.0, max = 96.0, mean = 55.83 median = 55.0 stdev = 11.98
================================================
column: Family history / ovarian cancer / type: numerical
{
    "fieldCount": {
        "0": 6968,
        "1": 83
    }
}
min = 0, max = 1, mean = 0.01 median = 0 stdev = 0.10
================================================
column: Tissue type (3 groups) / type: free-form
{
    "fieldCount": {
        "Invasive carcinoma": 3792,
        "NA": 2368,
        "Noinvasive carcinoma": 416,
        "Others": 465,
        "Paget's disease": 10
    }
}
================================================
missing values: {}
================================================
```

In this second snippet from an example data quality report, we show the prevalence of receptor statuses for each carrier gene.


```
==========================================
gene: BRCA1
{
    "counts": {
        "('ER', 'Negative')": 23,
        "('ER', 'Negative', 'age<50')": 9,
        "('ER', 'Negative', 'age>=50')": 14,
        "('ER', 'Positive')": 54,
        "('ER', 'Positive', 'age<50')": 17,
        "('ER', 'Positive', 'age>=50')": 37,
        "('HER2', '0')": 11,
        "('HER2', '0', 'age<50')": 3,
        "('HER2', '0', 'age>=50')": 8,
        "('HER2', '1+')": 14,
        "('HER2', '1+', 'age<50')": 5,
        "('HER2', '1+', 'age>=50')": 9,
        "('HER2', '2+')": 9,
        "('HER2', '2+', 'age<50')": 2,
        "('HER2', '2+', 'age>=50')": 7,
        "('HER2', '3+')": 2,
        "('HER2', '3+', 'age>=50')": 2,
        "('PgR', 'Negative')": 23,
        "('PgR', 'Negative', 'age<50')": 5,
        "('PgR', 'Negative', 'age>=50')": 18,
        "('PgR', 'Positive')": 50,
        "('PgR', 'Positive', 'age<50')": 13,
        "('PgR', 'Positive', 'age>=50')": 37
    }
}
gene: BRCA2
{
    "counts": {
```


Tumor pathology report

The tumor pathology report code is implemented as a custom data report and is called by the data quality report code. Here is a snippet from the Python code that implements the tumor

pathology report. It shows selecting those samples with and without a family history of breast cancer using SQL.

```
results['Triple negative breast cancer'] = {
    'with': {
        'Yes': psql.sqldf("select * from dfWithPath where (`PgR` = 'Negative' and `ER` = 'Negative') and \
        (`HER2` = '0'  or `HER2` = '1+')", locals()).shape[0],
        'No': psql.sqldf("select * from (select * from dfWithPath where  `PgR` != 'NA' and  `ER` != 'NA' \
        and  `HER2` != 'NA') T where  T.`PgR` != 'Negative' or  T.`ER` != 'Negative' or  (T.`HER2` != '0' and \
        T.`HER2` != '1+')", locals()).shape[0]},
    'without': {
        'Yes': psql.sqldf("select * from dfWithoutPath where (`PgR` = 'Negative' and `ER` = 'Negative') and \
        (`HER2` = '0'  or `HER2` = '1+')", locals()).shape[0],
        'No': psql.sqldf("select * from (select * from dfWithoutPath where  `PgR` != 'NA' and  `ER` != 'NA' \
        and  `HER2` != 'NA') T where  T.`PgR` != 'Negative' or  T.`ER` != 'Negative' or  (T.`HER2` != '0' and \
        T.`HER2` != '1+')", locals()).shape[0]}}

getPercentage(results, 'Triple negative breast cancer', 'Yes', ['Yes', 'No'])
getFisherExact(results, 'Triple negative breast cancer', ['Yes', 'No'])
```

The `getPercentage()` method takes the result set from the SQL query and just calculates a percentage of how many samples have a history of breast cancer. The `getFisherExact()` method takes the result set and calculates the odds ratio, p-value, and a 95% confidence interval for the statistics. Here is a sample output of the pathology report.

	with	without	P val	OR	95% CI
No. of subjects	404.0	6647.0	-	-	-
Age at onset	55.86	55.83	-	-	-
History of ovarian cancer	{Yes: 0.99%}	{Yes: 0.65%}	0.344	1.53	(0.4, 4.26)
TNM clinical classification N	{0:8%, 1:1%, 2:3.9%, 3:0%}	{0:74.8%, 1:2%, 2:2.5%, 3:1.7%}	-	-	-
TNM clinical classification M	{0: 98.1%, 1: 1.9%}	{0: 97.4%, 1: 2.5%}	0.817	1.30	(0.48, 4.95)
Estrogen-receptor status	{Positive: 73.14%}	{Positive: 72.92%}	1	1.01	(0.77, 1.34)
Progesterone-receptor status	{Positive: 67.04%}	{Positive: 60.64%}	0.04	1.32	(1.01, 1.73)
Triple negative breast cancer	{Yes: 5.95%}	{Yess: 8.7%}	0.545	0.66	(0.21, 1.66)
Family history of breast cancer	{Yes: 10.64%}	{Yes: 11.9%}	0.476	0.88	(0.62, 1.22)
Family history of ovarian cancer	{Yes: 1.73%}	{Yes: 1.14%}	0.334	1.52	(0.59, 3.33)
Family history of pancreas cancer	{Yes: 3.71%}	{Yes: 3.45%}	0.778	1.08	(0.59, 1.84)
Family history of stomach cancer	{Yes: 19.8%}	{Yes: 20.67%}	0.704	0.94	(0.73, 1.22)
Family history of liver cancer	{Yes: 8.17%}	{Yes: 6.36%}	0.174	1.30	(0.88, 1.9)
Family history of bone tumor	{Yes: 0.25%}	{Yes: 0.24%}	1	1.02	(0.02, 6.65)
Family history of bladder cancer	{Yes: 1.49%}	{Yes: 1.62%}	1	0.91	(0.33, 2.07)

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. In order to identify pathogenic variants, users must provide a delimited file with the following fields: `Clinical_significance_ENIGMA` and `Genomic_Coordinate_hg37` (or `Genomic_Coordinate_hg38`). The possible strings for the `Clinical_significance_ENIGMA` field include “Pathogenic”, “Likely pathogenic”,

“Benign”, “Likely benign”, “Uncertain significance”, and “-”, following the conventions defined by the BRCA Exchange. Variants that 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. Here is a snippet of the output showing data for a co-occurring VUS and a homozygous VUS.


```
{
    "cooccurring vus": {
        "(13, 32317399, 'T', 'G')": {
            "likelihood data": {
                "p1": 0.5,
                "p2": 0.001,
                "n": 2,
                "k": 2,
                "likelihood": 4e-06
            },
            "allele frequencies": {
                "maxPop": "Allele_frequency_genome_AMR_GnomAD",
                "maxPopFreq": 0.001179,
                "minPop": "Allele_frequency_genome_AFR_GnomAD",
                "minPopFreq": 0.0,
                "cohortFreq": 1.0
            },
            "pathogenic variants": [
                [
                    13,
                    32338277,
                    "G",
                    "T"
                ]
            ]
        },
...
    },
    "homozygous vus": {
        "(13, 32399624, 'T', 'C')": {
            "count": 1,
            "maxPop": "Allele_frequency_genome_AMR_GnomAD",
            "maxPopFreq": 0.001179,
            "minPop": "Allele_frequency_genome_AFR_GnomAD",
            "minPopFreq": 0.0,
            "cohortFreq": 0.5
        },
...
        }
    }
}
```

Example

The following is a basic example which we will use to generate our 3 reports.

1. Create a VCF file with 4 samples and 6 variants and put it in the `/tmp/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. Create a tumor pathology file for those 4 samples and put it in the `/tmp/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. Create a variant pathogenicity file for those 6 variants and put it in the `/tmp/data` directory.

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

4. Configure the report workflow.

```
{
  "qualityReport": "data/data-quality-report.txt",
  "pathologyReport": "data/tumor-pathology-report.txt",
  "fileName": "data/mypf.tsv",
  "fileHeader": "True",
  "fieldDelimiter": "\t",
  "printConfigFileInfo": "False",
  "printBadValues": "False",
  "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": [], "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. Run the report workflow.

```
$ ./runMe.sh -rc config/report-config.json -vf my.vcf -hg 38 -er 99 \
-c 13 -p True --p2 0.001 -g BRCA2 -pf mybrca.tsv -dd /tmp/data \
-st True -sp mypf.tsv
```

6. All the reports will be located in the `/tmp/data` directory.