



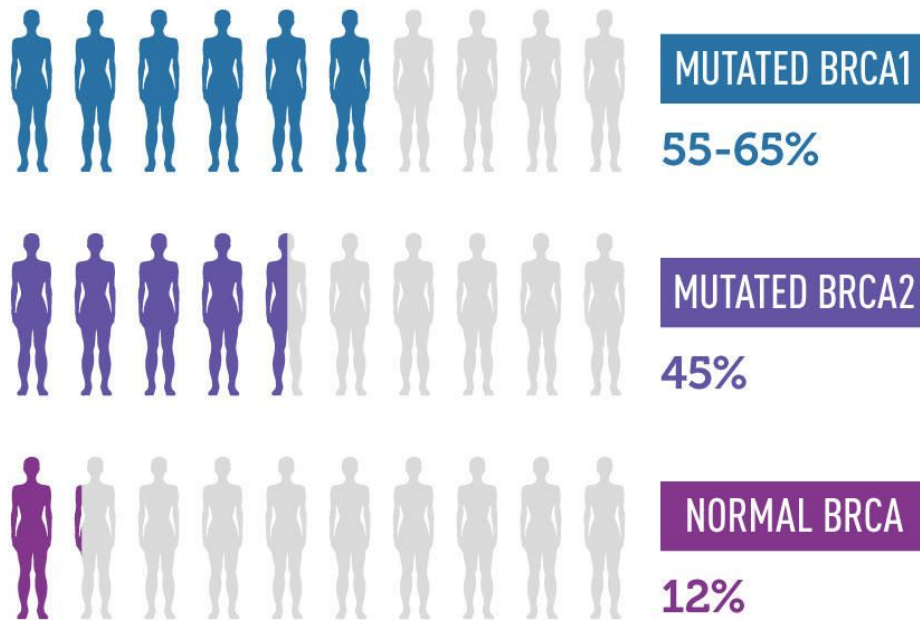
Bart Keagy

Federated Analysis for Cancer Variant Interpretation

Melissa Cline, Ph.D.



Genetics strongly influences the risk of breast cancer



NATIONAL CANCER INSTITUTE
CHANCES OF DEVELOPING
BREAST CANCER BY AGE 70

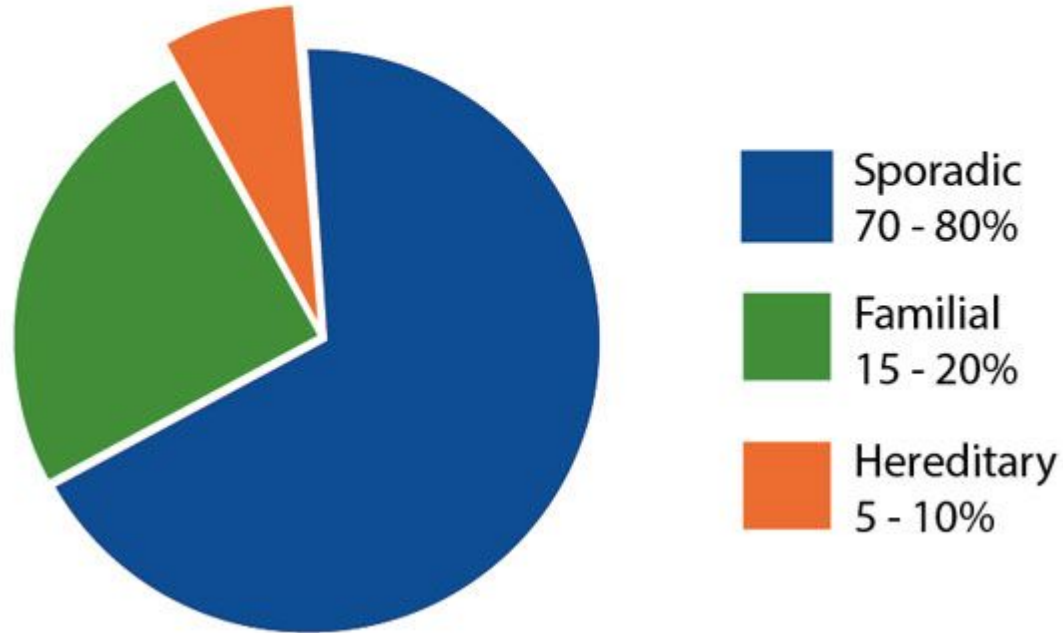
www.cancer.gov/brca-fact-sheet

Common Heritable Cancer Syndromes

Syndrome	Gene	Incidence	Cancers
Hereditary breast and ovarian cancer syndrome	BRCA1 BRCA2	1/300-800 Ashkenazi: 1/40	Breast, ovary, melanoma, prostate, pancreatic
Hereditary ovarian cancer syndrome	RAD51C RAD51D BRIP1	Unknown	Ovary
Lynch syndrome	MLH1 MSH2 MSH6 PMS2 EPCAM	1/660-2000	Uterine, colon, ovary, pancreatic, gastric, small intestine, central nervous system, renal, sebaceous
Cowden syndrome	PTEN	1/200,000	Breast, uterine, thyroid, colon, renal, sebaceous
Li-Fraumeni syndrome (LFS)	P53	Unknown	Sarcomas, breast, adrenal, brain, lung, endometrial
Peutz-Jeghers	STK11	1/25,000-300,00	Gastrointestinal, breast, ovarian, sex cord stromal, uterine, cervical (adenoma malignum)

How many cancers are heritable?

Cancer Genetic Risk Assessment



Heritable cancer risk can be manageable

- **Screening** to detect any cancer in its early stages, when it is most treatable.
- **Medications** that lower cancer risk.
- **Risk-Reducing Surgery** to remove high-risk tissue before cancer can develop.

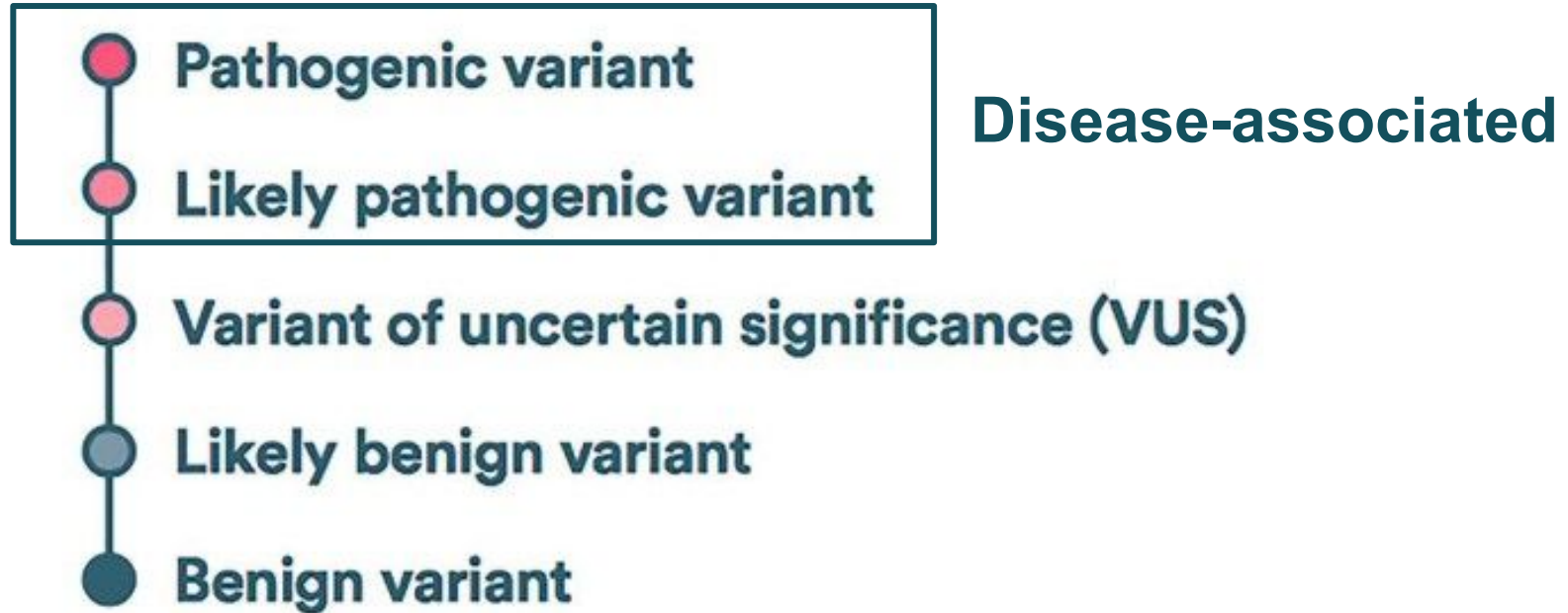
To take advantage of these strategies, you need to get tested, and the test needs to recognize your genetic cancer risk.

Genetic testing for hereditary cancer mutations can save lives.

ACMG/AMP Germline variant classification

- 
- Pathogenic variant**
 - Likely pathogenic variant**
 - Variant of uncertain significance (VUS)**
 - Likely benign variant**
 - Benign variant**

ACMG/AMP Germline variant classification



ACMG/AMP Germline variant classification

● Pathogenic variant

● Likely pathogenic variant

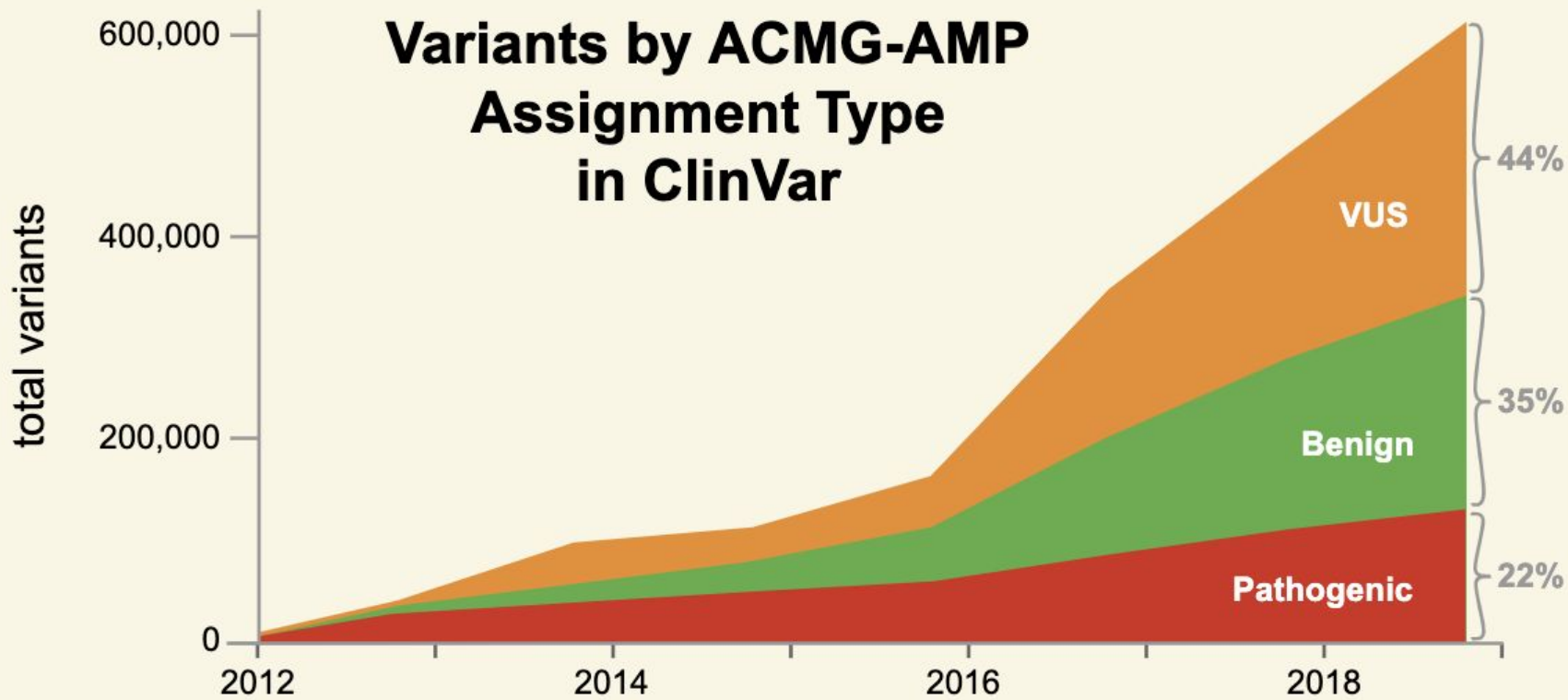
● Variant of uncertain significance (VUS)

Unknown!

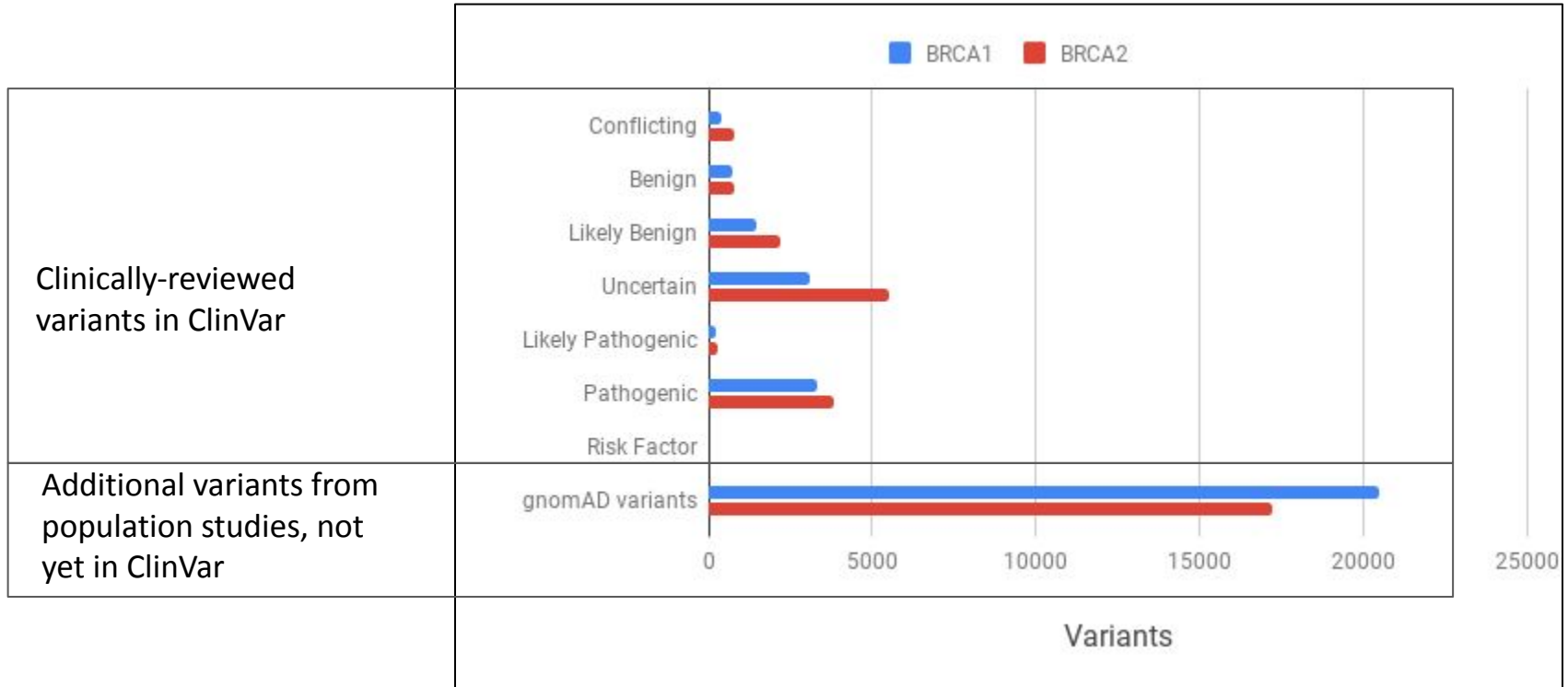
● Likely benign variant

● Benign variant

Prevalence of Variants of Uncertain Significance (VUS)



The Variant Interpretation Bottleneck

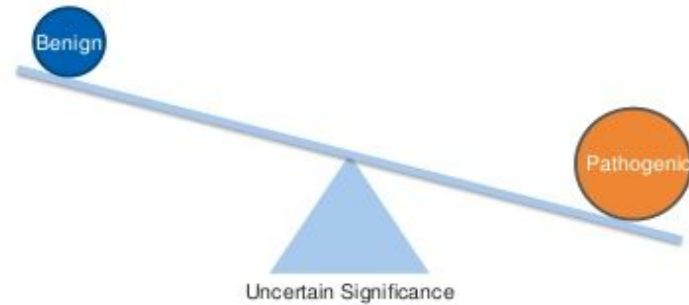


Data as of April 2021

Variant interpretation involves assessing many forms of evidence together

Types of Variant Evidence

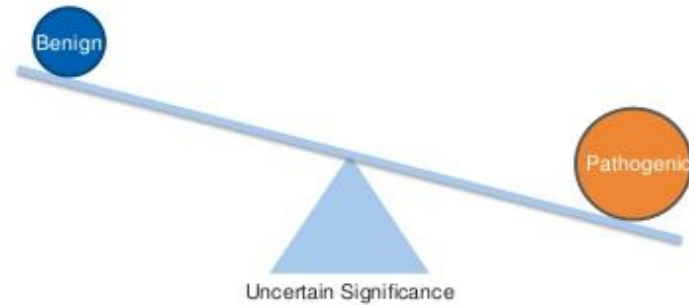
- ❑ Clinical Family Studies
- ❑ Clinical Patient Observations
- ❑ Functional Assays
- ❑ Population Frequencies
- ❑ Computational Prediction



Variant interpretation involves assessing many forms of evidence together

Types of Variant Evidence

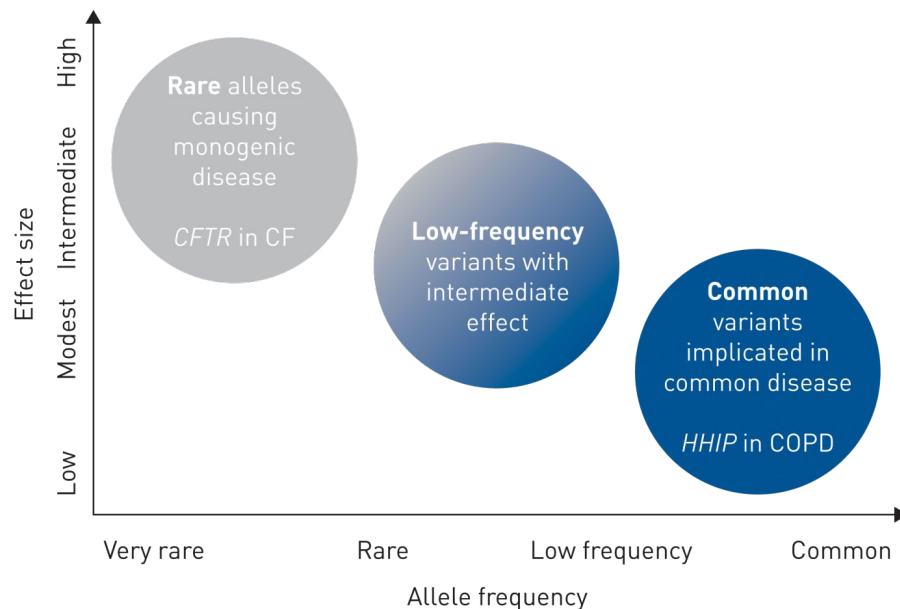
- ❑ Clinical Family Studies
- ❑ Clinical Patient Observations
- ❑ Functional Assays
- ❑ Population Frequencies
- ❑ Computational Prediction



Most variant interpretation requires some clinical evidence

Most cancer-associated variants are highly rare

- No single institution is likely to have enough patient data for robust variant interpretation
- Patient-level data is impossible to share for privacy reasons
- How can one share knowledge about genetic variants while safeguarding patient privacy?



BRCA Challenge

- Global consortium launched by the GA4GH.
- Vision: assemble team of experts to pioneer data sharing for BRCA and cancer, as an exemplar for other genes and disorders.



Stephen Chanock



Sir John Burn



Rachel Liao



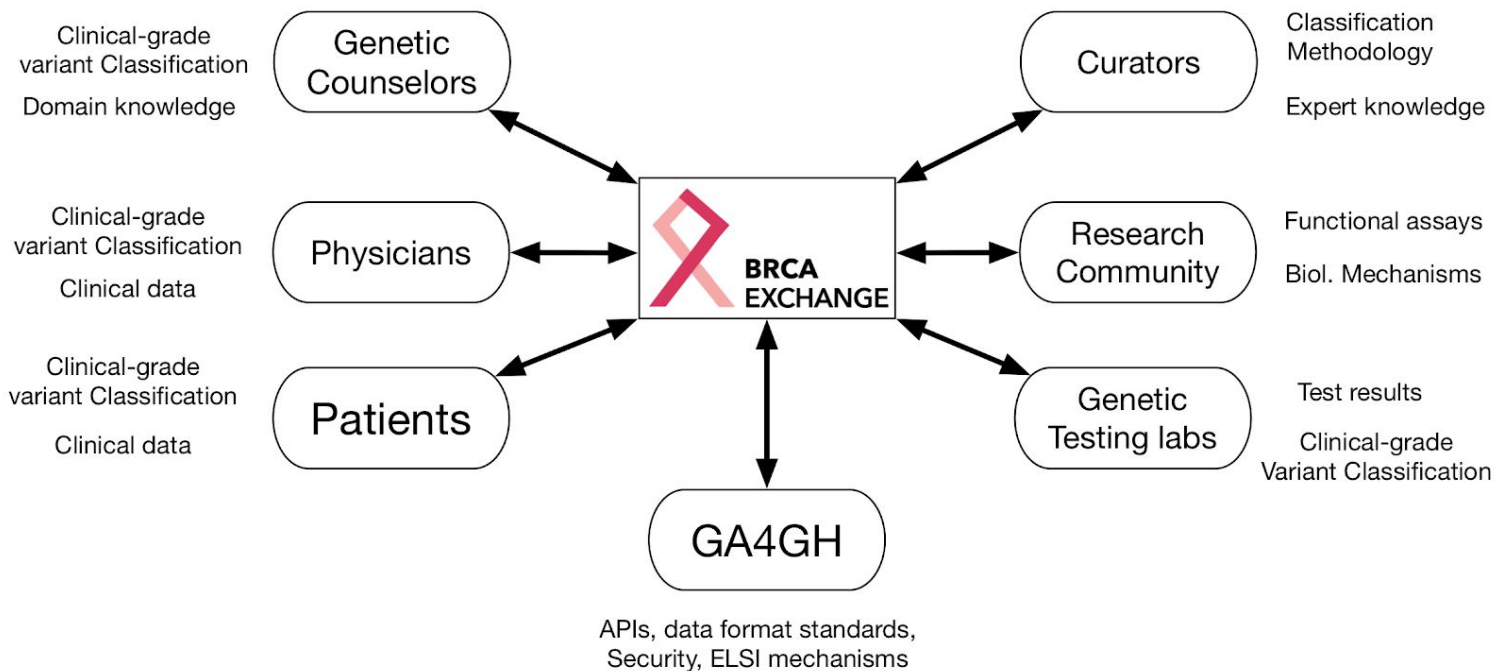
Solution: BRCA Exchange



Benedict Paten, UCSC

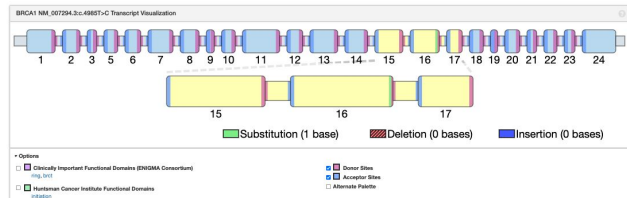


Gunnar Rättsch, ETH Zurich



Variant Details

Show Empty Items



Clinical Significance (ENIGMA)

Clinical Significance benign

Comment on Clinical Significance HVC class based on posterior probability from multifactorial likelihood analysis, thresholds for class are: Poor (P < 0.001), Class 1 (based on posterior probability < 0.0001)

Assertion Method ENIGMA BRCA1/2 Classification Criteria (2015)

Date last evaluated 10 August 2015

Collection Method Curator

Clinical Significance Criteria PAF0: 21999134

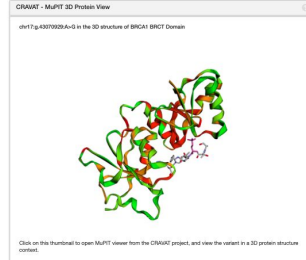
Allele Origin Germline

Conditions (Mode of Inheritance) Breast-Ovarian Cancer, Familial, Susceptibility To, 1 [ORPHA]



CHUNK - MUYT 3D Protein View

chr17:g.4307028A>G is in 3D structure of BRCA1 BRCT Domain



Clinical Significance (ClinVar)

★★★★☆ Submitter: ENIGMA Benign (10 June 2005)

★★★★☆ Submitter: Women's Health and Genetics/Laboratory Corporation of Am., Benign (19 June 2011)

★★★★☆ Submitter: Invitae Likely Benign (03 May 2021)

★★★★☆ Submitter: GeneDx Benign (01 February 2021)

★★★★☆ Submitter: Color Health, Inc Benign (01 February 2021)

★★★★☆ Submitter: Ambry Genetics Benign (02 December 2008)

★★★★☆ Submitter: BRCA (BRCA1) VUS (01 March 2008)

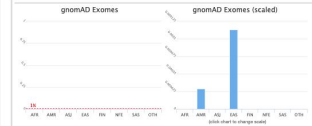
Clinical Significance (BIC)

Allele Frequency Reference Sets

The gnomAD data sets used by BRCA Exchange are the "non-cancer" subsets of their sources. Data from T2D and other cancer cohorts are excluded to ensure that the frequencies used to assess pathogenicity represent individuals not affected by cancer.

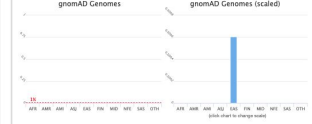
*** gnomAD V0.1 Exomes, Non-Cancer (gnomAD)**

Additional data for this variant, including detailed populations, quality scores, and flags relative to other transcripts, are available at gnomAD.



*** gnomAD V0.1 Exomes, Non-Cancer (Hemnet)**

Additional data for this variant, including detailed populations, quality scores, and flags relative to other transcripts, are available at gnomAD.



*** gnomAD V0.1 Genomes, Non-Cancer (Hemnet)**

Variant Nomenclature

Gene Symbol: BRCA1

Reference cDNA Sequence: NM_007294.3

HGVN Nucleotide: c.4887T>C

HGVN Protein: p.Phe163Leu

Abbreviated AA Change: P163L

BIC Designation: VUS (T-C)

Clinical Allele Registry: VUS (000117)

Biocompare (BioC): NC_000017.1:g.4307028A>G

Genome (BioC): NC_000017.10:g.4307028A>G

RNA SLOVD: c175

Recom: 1633/1633con-network-ugm/1633p7

dbSNP: rs1041102108/1633con-network-ugm-020037

gnAD V0.1 Identifier: gnomAD_V0.1:GEMF0NCA0L0L1_C179NKA0L0

Aliases:

15047T>C: P163L; P163L; P163L; P163L; URG_292_p.147057T>C; URG_292_p.Phe163Leu; URG_292_p.c.4887T>C; NC_000017.10:g.4307028A>G; NC_000017.1:g.4307028A>G; NC_000017.10:g.4307028A>G; NC_000017.10:g.4307028A>G; NC_007294.3:c.4887T>C; NC_007294.3:c.4887T>C; NC_007294.3:c.4887T>C; U16860:10:49857T>C:p.P163L

Clinical Significance (SLOVD)

Submitter: ENIGMA consortium (Bethesda, MD) Clinical Classification: benign

Submitter: Thomas Hansen (Copenhagen, DNK) Clinical Classification: -

Submitter: Peter Decker (Linden, NJ) Clinical Classification: -

Submitter: BRISQES consortium (Barnes, NJ) Clinical Classification: -

Submitter: Yuhaihua Momen (Prohokema, JP) Clinical Classification: benign

Submitter: Johan den Dunnen (Pittersden, NL) Clinical Classification: benign; VUS

Multifactorial Likelihood Analysis

Posterior probability of pathogenicity: 0.0001

Prior probability of pathogenicity: 0.04

Mutation analysis: 0.03

Co-occurrence likelihood: 0.00085

Segregation likelihood: 1

Summary Family History Likelihood Ratio: 1.44

Literature Reference: Gaston et al. 2007

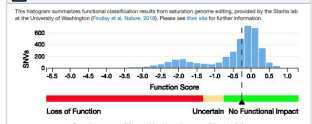
Functional Assay Results

Assays were selected by the ENIGMA consortium as high-quality assays that met internal standards permitting estimation of sensitivity and specificity using quantitative or absolute (Yes/No) data.

- Genetic variation can impact RNA transcripts and/or protein function or abundance. Assays labeled with a * measure the impact of genetic variation on the protein level. Assays labeled with a *BP measure impact on both the RNA and protein level.
- Results in the "after" interpretation of the functional impact of each variant, are considered to fit across the sub-labels. A report of "No Effect/Protein" means that the authors performed more than one assay for each variant but provided no overall report of functional impact.
- Additional assays may be added in the future.

Disclaimer

Functional assay data reflect laboratory models of disease, and should not be used as a substitute for clinical expert interpretation. Further, please note that interpretation of the author's terms to affect impact on function, the results presented do not reflect a final variant classification.



Author Finlay et al. 2018

Publication PAF0:000000

Report PUNC

Functional Evidence Score 0.254873276

RNA SLOVD 0.1218884761

RNA Class Not Determined

Report Descriptions PUNC: Functional, RPT: Intermediate, LOF: Loss of function

*** Fernandes et al. 2019** Report: 1.0

Variant identification

Clinical Assertions

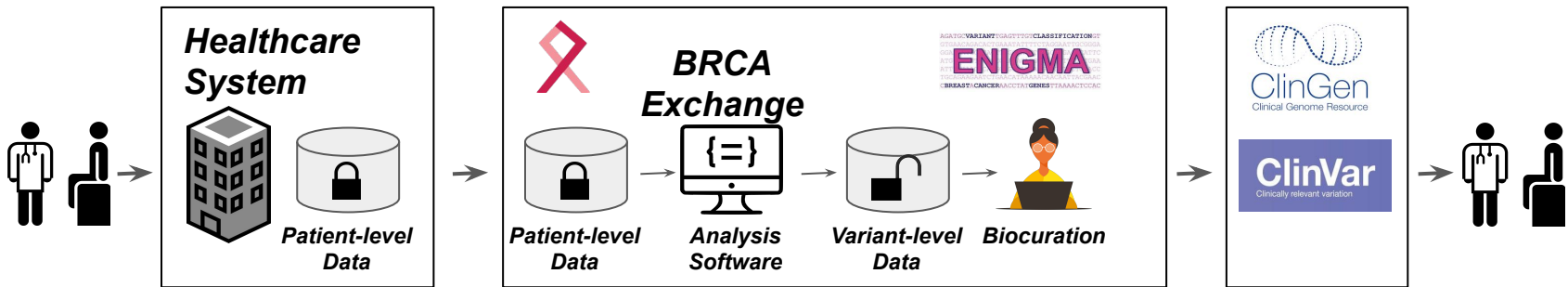
Evidence of Variant Pathogenicity

Sharing clinical data for variant interpretation

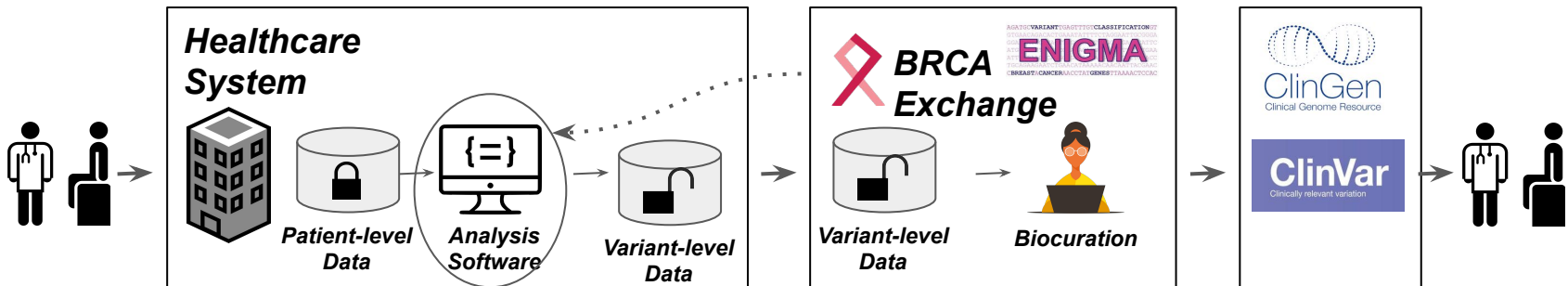
- Variant interpretation relies on sensitive patient-level data
- But, the information actually needed are variant-level summaries, *which are much less sensitive!*
 - Case-control ratios
 - Allele Frequencies
 - Variant co-occurrences
- **Vision:** instead of sharing the patient-level data directly, share a container to compute the variant-level summary data

Privacy-preserving data sharing through federated analysis

Traditional Data Sharing



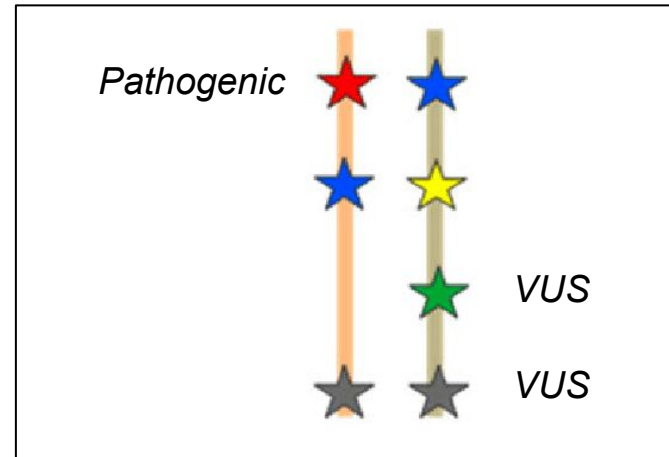
Federated Analysis



Variant Co-occurrence Estimation

- **Rationale:** BRCA1/2 are “essential” genes. The cell needs at least one working copy.
- Two pathogenic BRCA variants *in trans* is usually embryonic lethal, but can lead to a rare disease with a top life expectancy of 40
- If a VUS is observed in an unaffected individual of age 40 or greater and the VUS is either:
 - *in trans* with a known pathogenic variant (*green*),
 - or in a homozygous genotype (*grey*)

Then this supports a benign interpretation

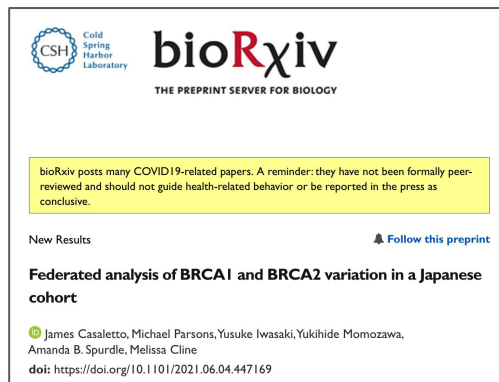


Federated analysis of BioBank Japan Data

- RIKEN holds a large cohort of cancer patients and controls from BioBank Japan, which they cannot share directly.
- We shared a Docker container with them to analyze their patient cohort for variant co-occurrences and allele frequencies
- The container generated variant-level data, which we are now using together with the ENIGMA Consortium to interpret BRCA variants!

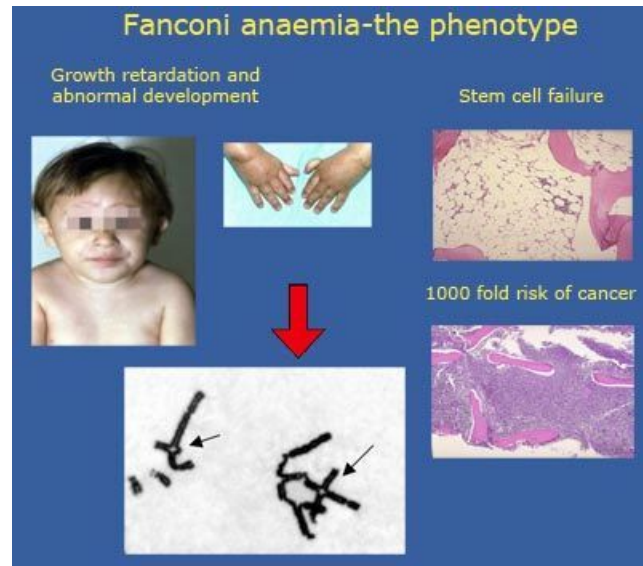
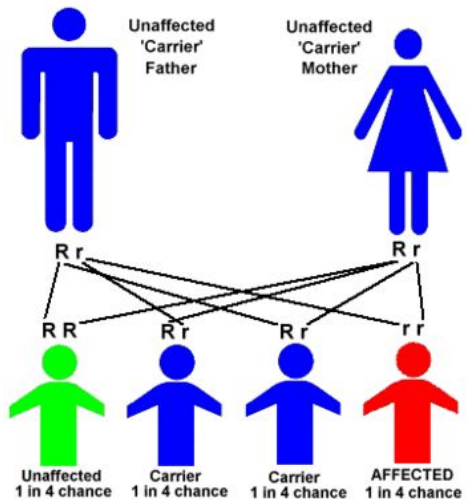


Yukihide Momozawa, RIKEN



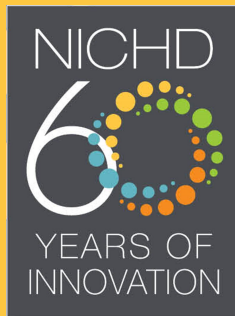
Next Federated Analysis: Fanconi Anemia (FA)

- Rare childhood-onset disease that arises from two pathogenic BRCA (and other) variants *in trans*.
- Diverse phenotype that includes early onset cancer
- Ongoing research on which germline variants are implicated in FA.
- Vision: Mine the Kids First data on Cavatica for new examples of FA.



Challenge #1: not all clouds are created equal

- Had previously been running the co-occurrence Docker container on Terra on NHLBI BioData Catalyst
- Cavatica is based on Seven Bridges
- Porting the container from Terra (WDL) to Seven Bridges (CWL) required work
- **Bigger issue: making containers portable across cloud platforms**



Focus on the Science with a few Elements of Style

January 14, 2022

Anne Deslattes Mays, PhD

National Scholar

@adeslat

anne.deslattesmays@nih.gov



Eunice Kennedy Shriver National Institute
of Child Health and Human Development



Researcher challenges to data and compute



Current State



Future State

LapTop or On Premise



Cloud Environment
CAVATICA, TERRA, CloudOS
(Platforms as a Service)

Shell Based Workflow



Community based
Standard Workflow Languages

Downloading
Data



Bring the Compute to the Data

Not sustainable



Sustainable



Focus on the science...

- Using platforms as a service lets the researcher focus on the science
 - Platform assists in managing compute costs (automatic logging out)
 - Platforms facilitate collaboration
 - Platforms abstract out the technical machine details, allowing work at a higher level
- Good science provides repeatable steps for others to follow
 - Well documented methods allows the next scientist to repeat what you have done
 - Document those steps using markdown
 - Give attribution to those who wrote those processes or workflows you have expanded upon or modified
- Using these practices of containerization abstracts away the platform-specific layer making your work more portable and accessible for use on other platforms

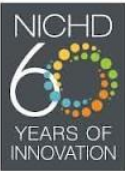


Elements of Style in Reproducible Workflow Creation

- All workflows share the same basic steps - differences are in the technical details
 - Input
 - Process
 - Output
- Containerize at the process level
 - Dockerfile to capture the details
 - Conda to control the environment
 - GitHub to build the Dockerfile and Create the Image
 - Automate and keep up-to-date with GitHub actions
 - Document the objective
- Stitch together with the standard workflow languages dictated by the platform
- Use GitHub to store and share the workflow and markdown to document your steps

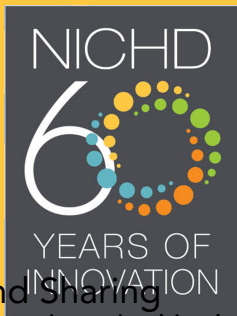


Elements of Style in Reproducible Workflow Creation and Analysis: An INCLUDE & Kids First Training Event



Coming soon!





Office of Data Science and Sharing
<https://www.nichd.nih.gov/about/org/od/odss>



Rebecca Rosen
Director
Office of Data Science and Sharing (ODSS)
Eunice Kennedy Shriver National Institute of
Child Health and Human Development
(NICHD)



Valerie Cotten
Deputy Director
Office of Data Science and Sharing (ODSS)
Eunice Kennedy Shriver National Institute of
Child Health and Human Development (NICHD)



Anne Deslattes Mays
Senior DATA Strategist
Office of Data Science and Sharing (ODSS)
Eunice Kennedy Shriver National Institute of
Child Health and Human Development (NICHD)



Eunice Kennedy Shriver National Institute
of Child Health and Human Development



Challenge #2: Container certification

- Goal of the BRCA Challenge: gather data from diverse global repositories
 - On-going work in integration with the GA4GH Starter Kit
- Challenge: how can you demonstrate to external collaborators that your container is trustworthy?

Ongoing work related to container certification

- NHLBI BioData Catalyst and Dockstore
 - Emerging best practices including: no root dependencies, no data egress
- GA4GH Cloud Workstream Test Bed
 - Quantifies basic system demands such a CPU usage
- BioCompute Objects
 - Community-driven initiative to build a framework for characterizing and sharing computational workflows
 - Has FDA support



Summary

- Container methods and federated analysis show promise for sharing information from patient data for variant interpretation
 - Publication to appear in Cell Genomics
- Ongoing challenges include:
 - Cloud portability
 - Workflow certification
- There are promising developments in both areas

Acknowledgements

Zack Fischmann
James Casaletto
Benedict Paten
Charlie Markello
Mary Goldman
Gunnar Rättsch
Marc Zimmerman
Faisal Alquaddoomi

Amanda Spurdle
Michael Parsons
Yukihide Momozawa

Anne Deslattes Mays
Ian Fore
Jeremy Adams

BRCA Challenge Steering Committee
BRCA Challenge Evidence Gathering Group



Global Alliance
for Genomics & Health



*NIH Big Data to
Knowledge (BD2K)*



 **NATIONAL CANCER INSTITUTE**

**Informatics Technology for
Cancer Research (ITCR)**