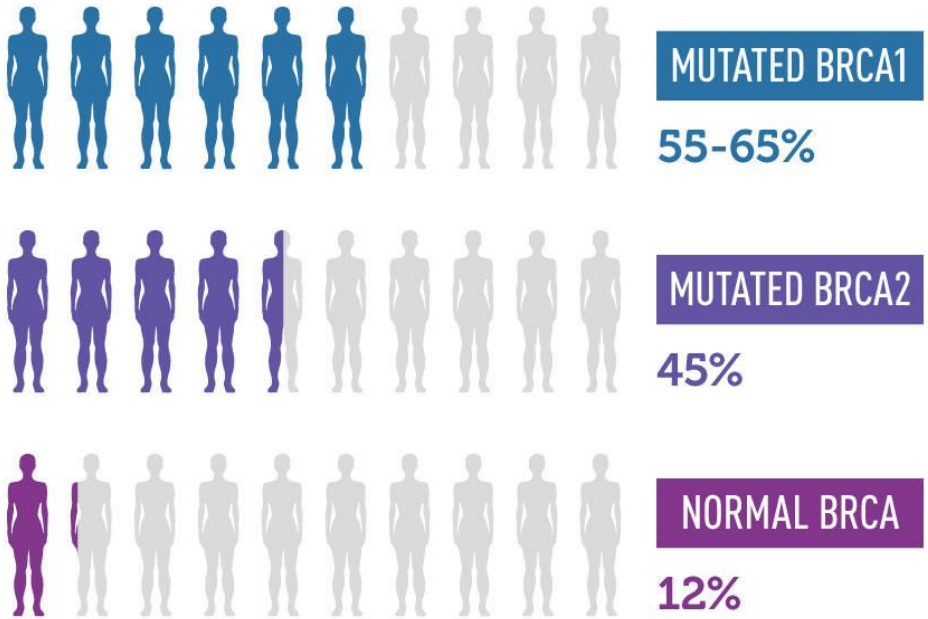



Federated Analysis for Cancer Variant Interpretation

Melissa Cline, Ph.D.

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

Genetics strongly influences the risk of breast cancer



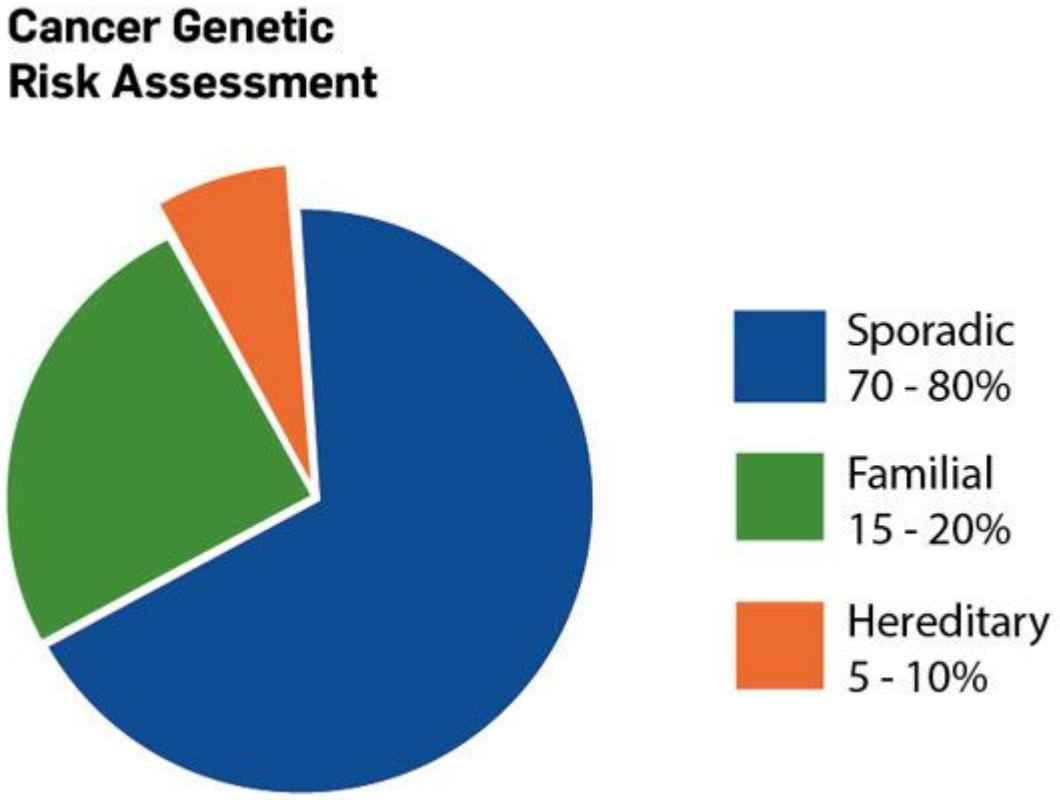
NATIONAL CANCER INSTITUTE
CHANCES OF DEVELOPING
BREAST CANCER BY AGE 70

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?



UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

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.

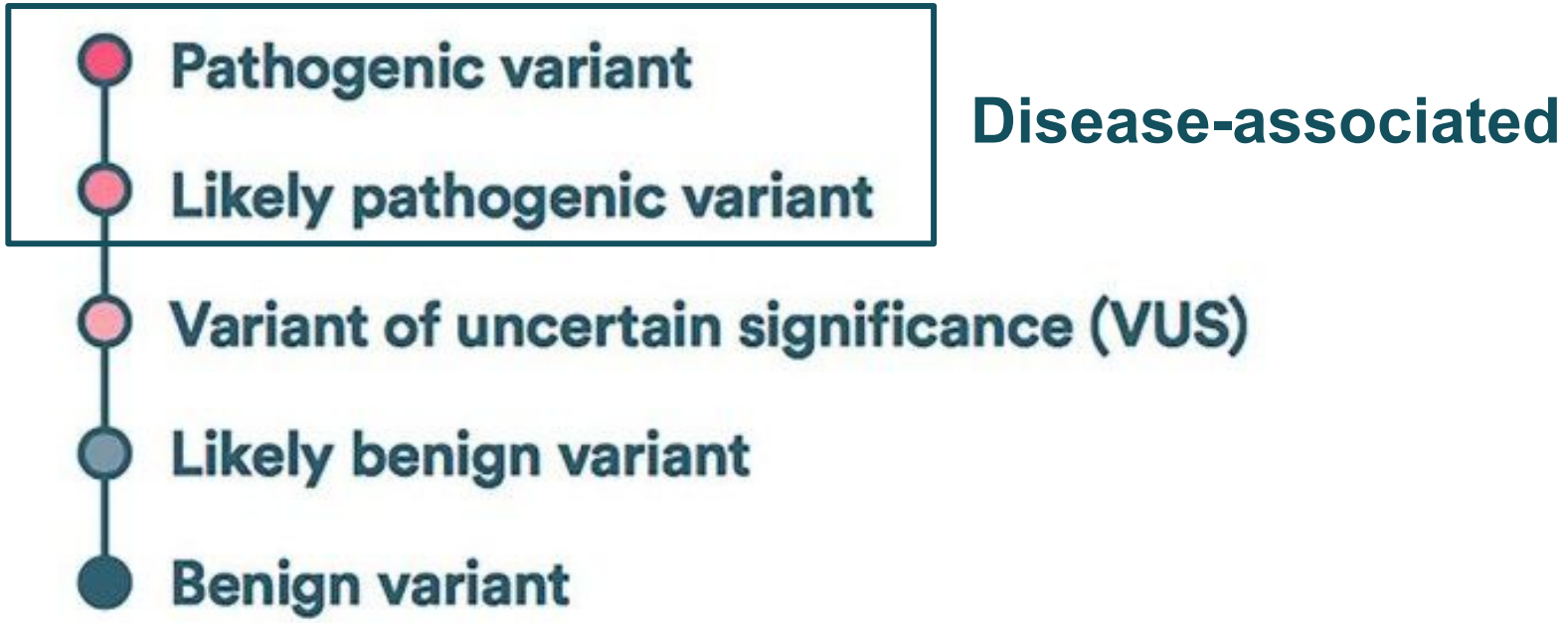
To take advantage of these strategies, you need to get tested, and the test needs to recognize your genetic cancer risk.

Genetic testing for hereditary cancer mutations can save lives.

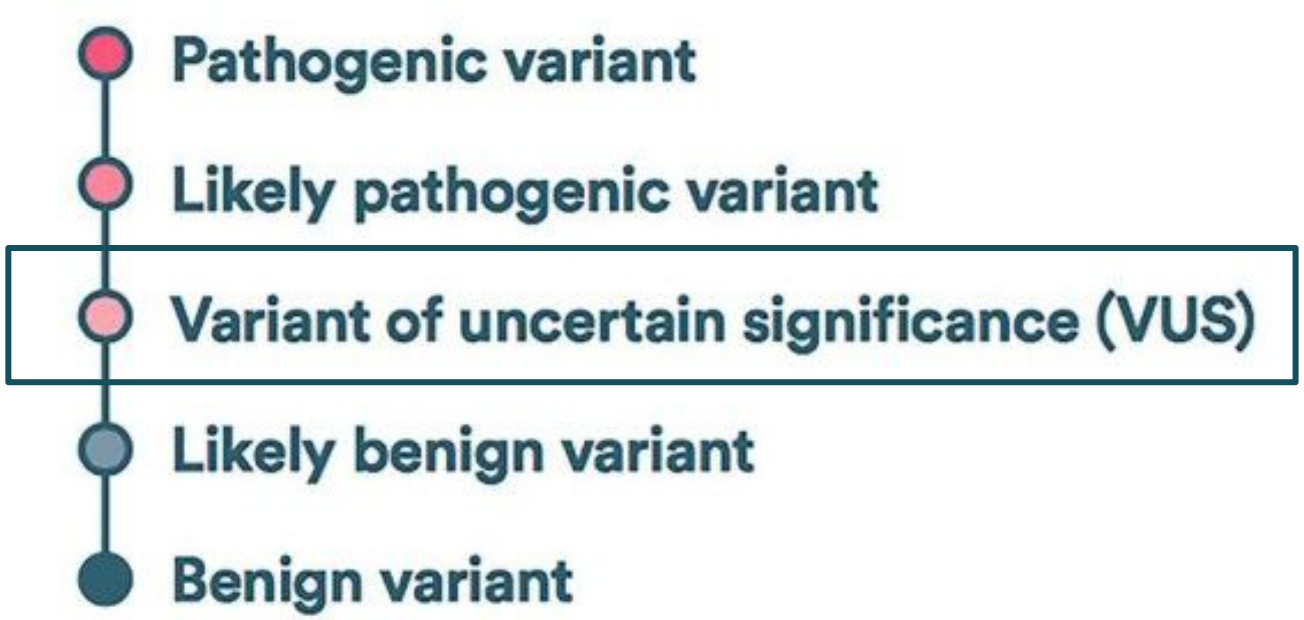
UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

ACMG/AMP Germline variant classification

ACMG/AMP Germline variant classification

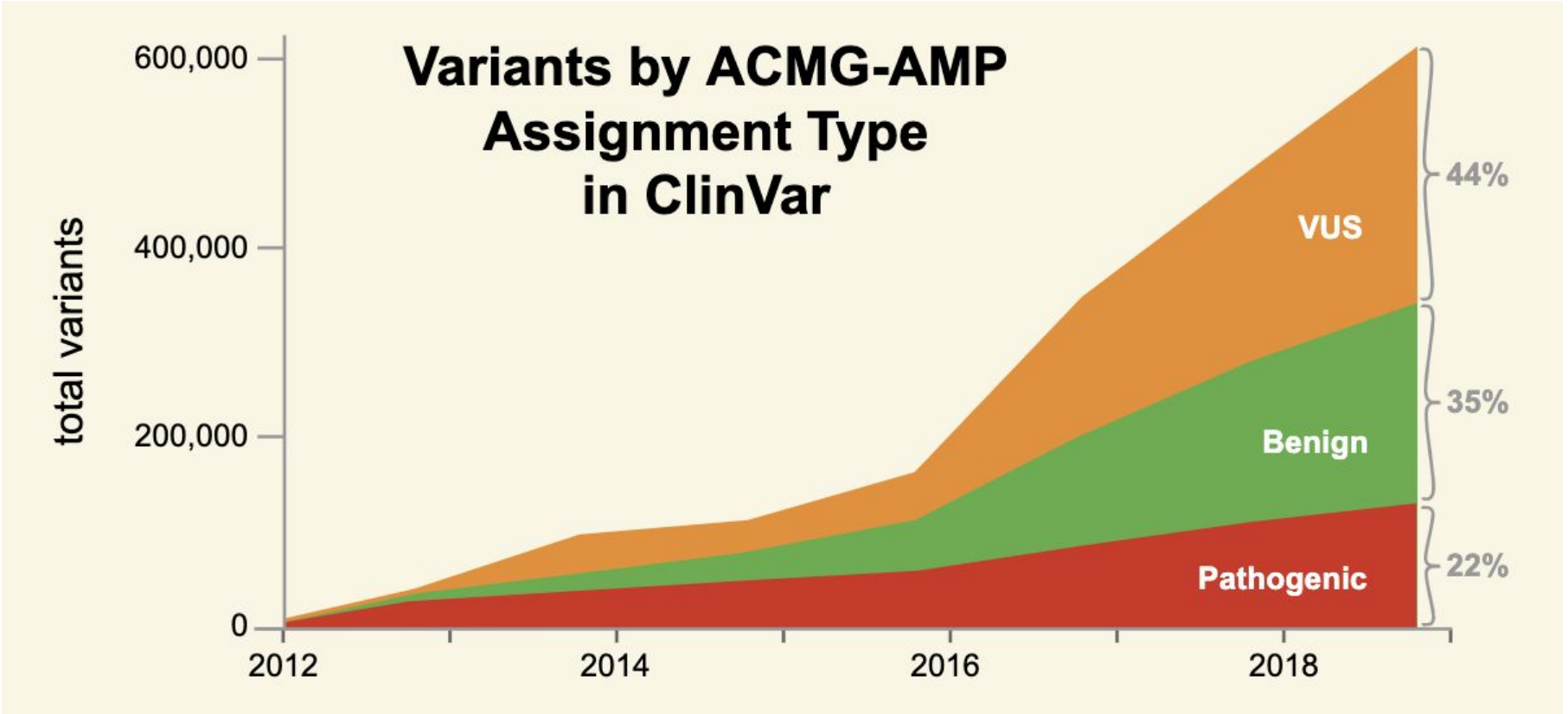


ACMG/AMP Germline variant classification



Unknown!

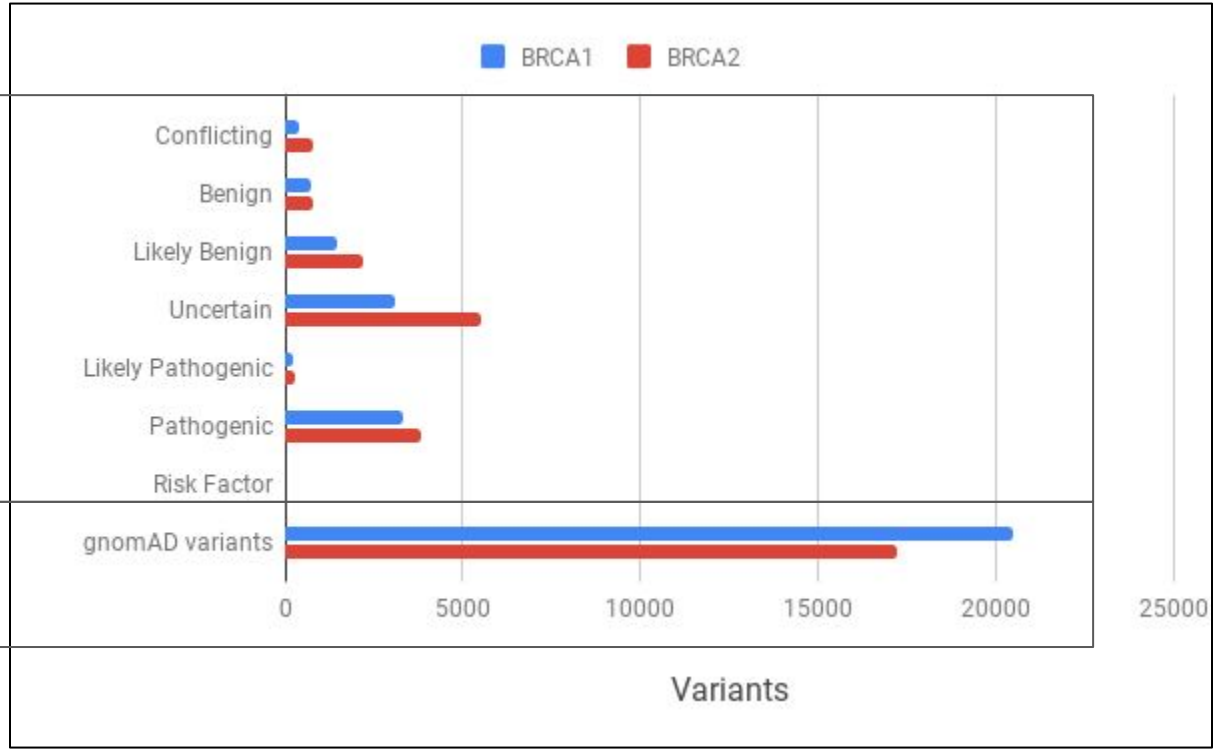
Prevalence of Variants of Uncertain Signifcance (VUS)



The Variant Interpretation Bottleneck

Clinically-reviewed variants in ClinVar

Additional variants from population studies, not yet in ClinVar

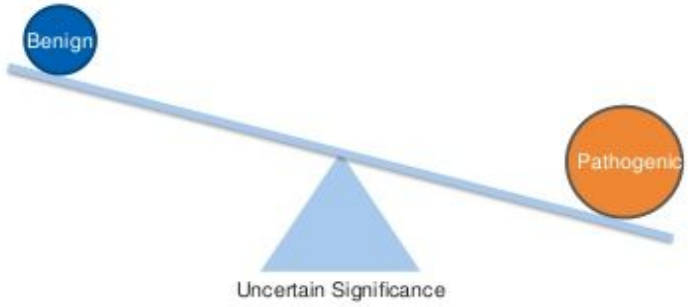


Data as of April 2021

Variant interpretation involves assessing many forms of evidence together

Types of Variant Evidence

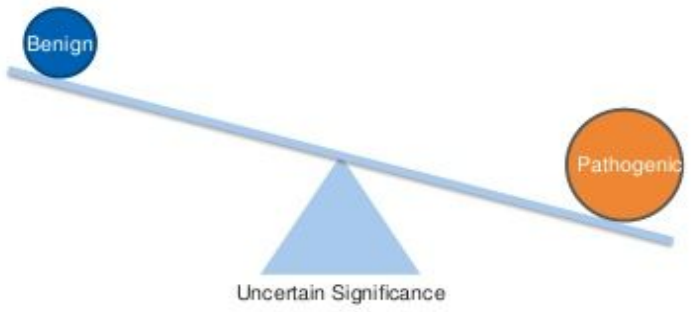
- ❑ Clinical Family Studies
- ❑ Clinical Patient Observations
- ❑ Functional Assays
- ❑ Population Frequencies
- ❑ Computational Prediction



Variant interpretation involves assessing many forms of evidence together

Types of Variant Evidence

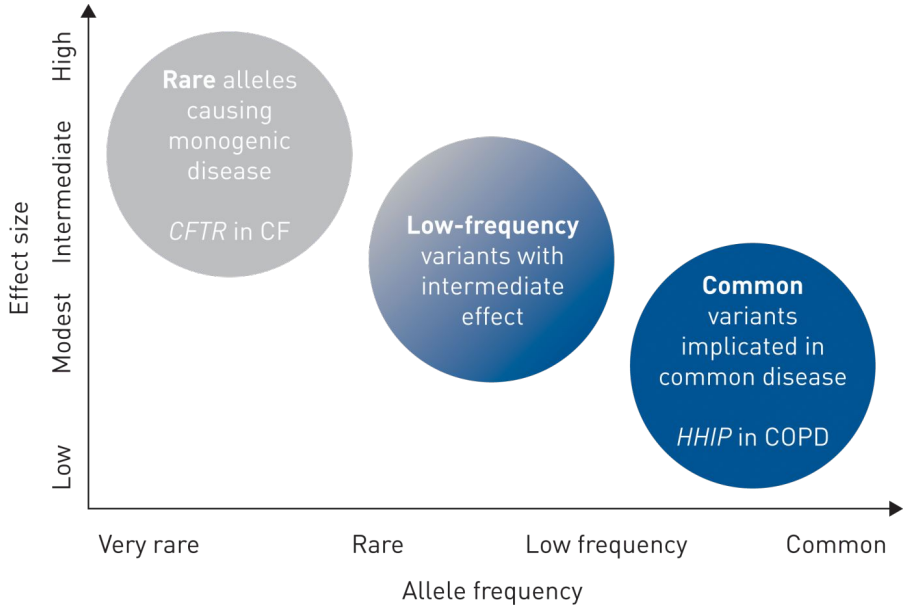
- ❑ **Clinical Family Studies**
- ❑ **Clinical Patient Observations**
- ❑ Functional Assays
- ❑ Population Frequencies
- ❑ Computational Prediction



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

THE HUMAN VARIOME PROJECT

Global Alliance for Genomics & Health

Collaborate. Innovate. Accelerate.

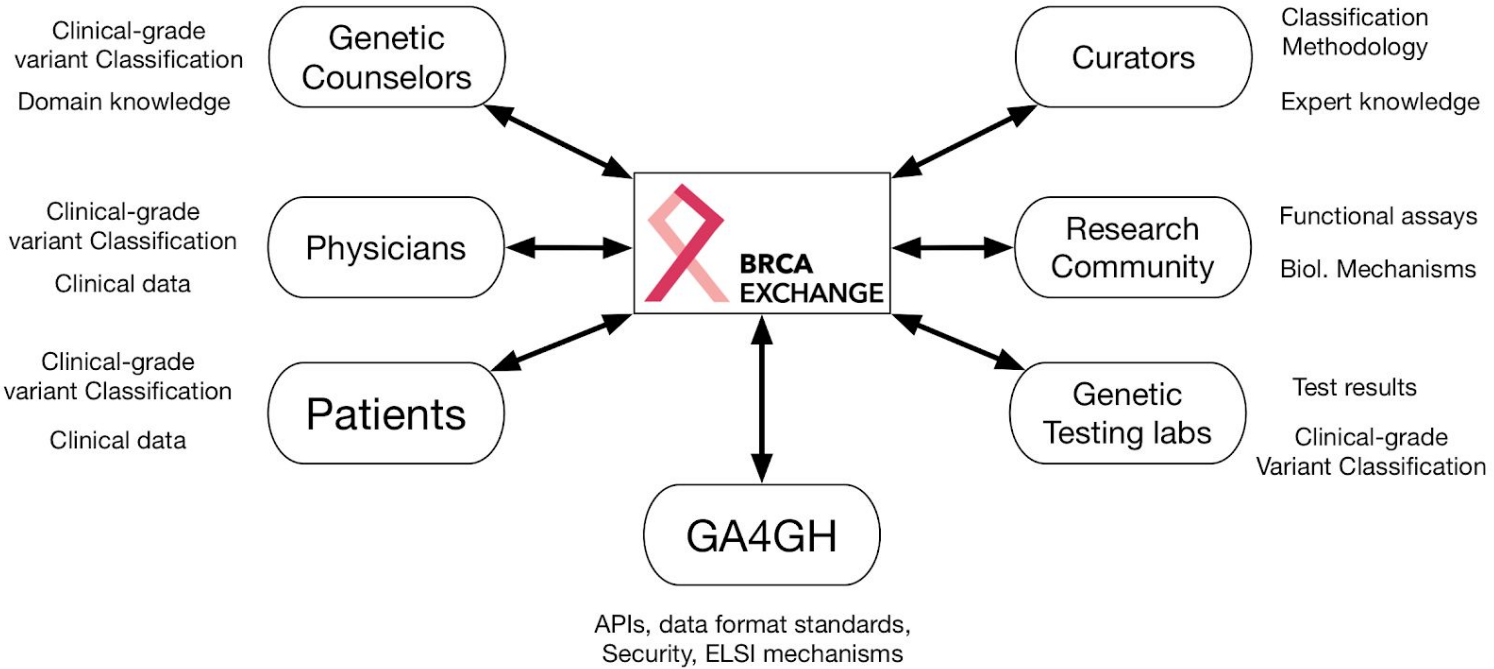
BRCA CHALLENGE

UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute

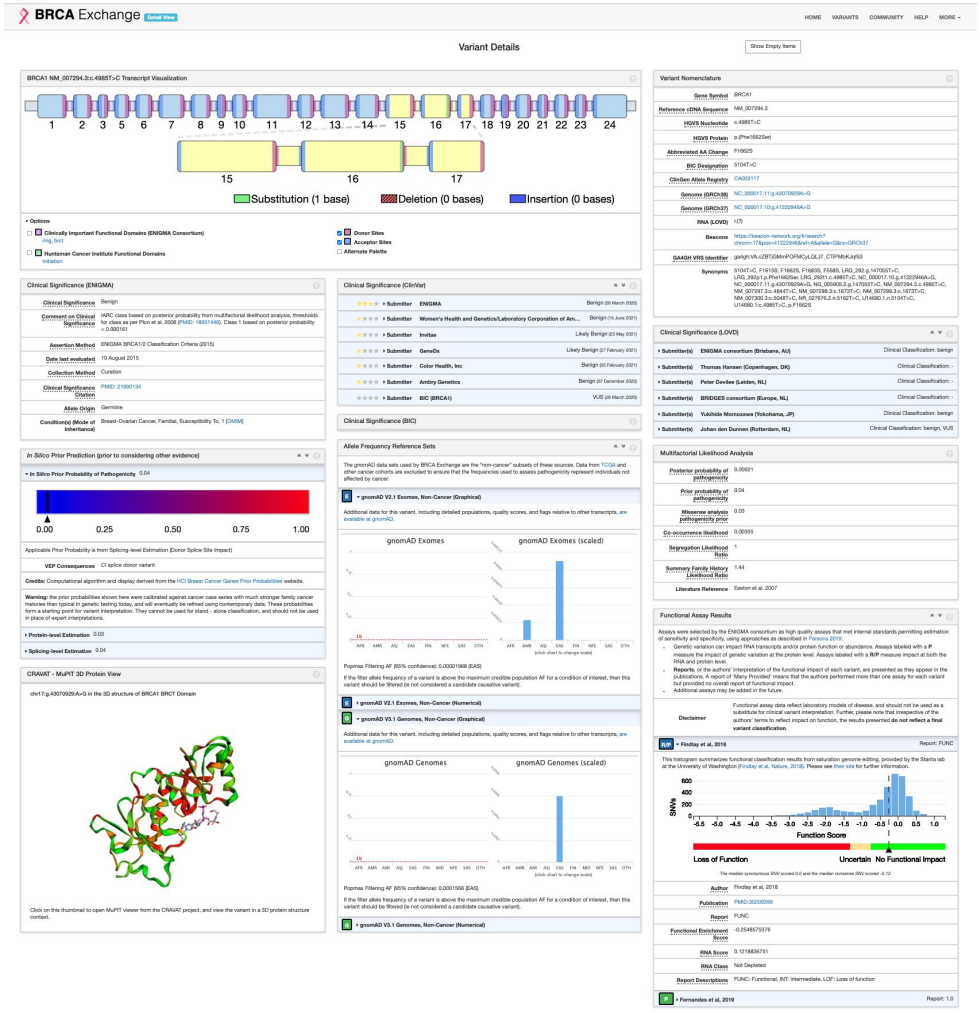
Solution: BRCA Exchange

Benedict Paten, UCSC

Gunnar Rätsch, ETH Zurich



UNIVERSITY OF CALIFORNIA SANTA CRUZ Genomics Institute



Variant identification

Clinical Assertions

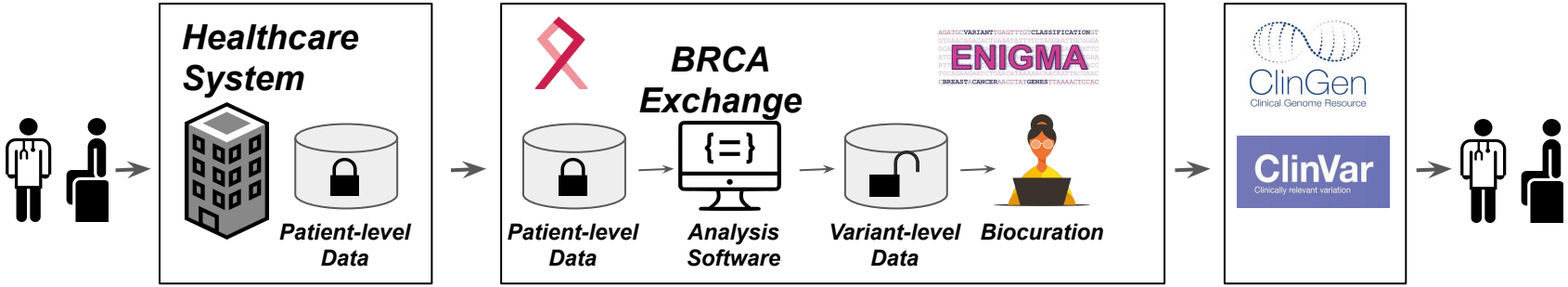
Evidence of Variant Pathogenicity

Sharing clinical data for variant interpretation

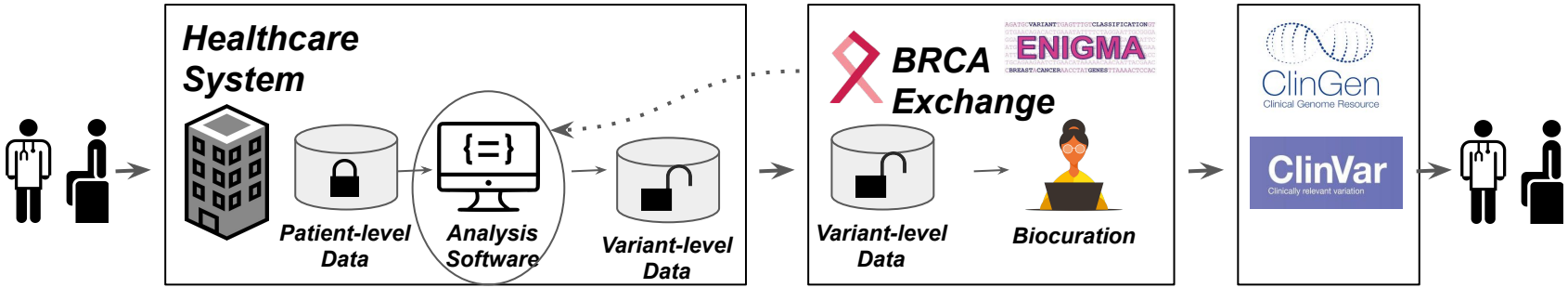
- Variant interpretation relies on sensitive patient-level data
- But, the information actually needed are variant-level summaries, *which are much less sensitive!*
    - Case-control ratios
    - Allele Frequencies
    - Variant co-occurrences
- **Vision:** instead of sharing the patient-level data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

Privacy-preserving data sharing through federated analysis

Traditional Data Sharing



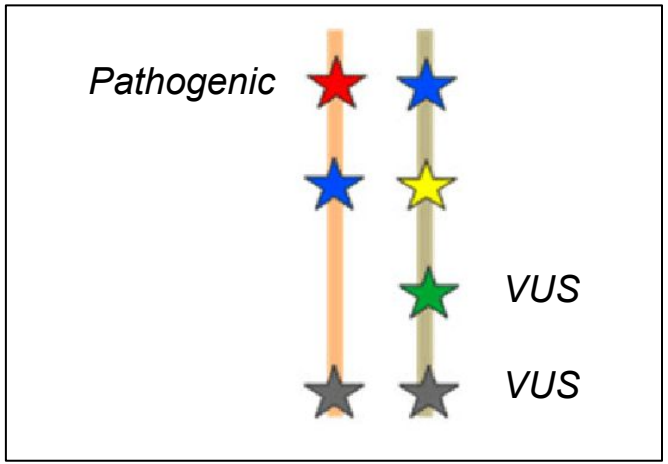
Federated Analysis



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



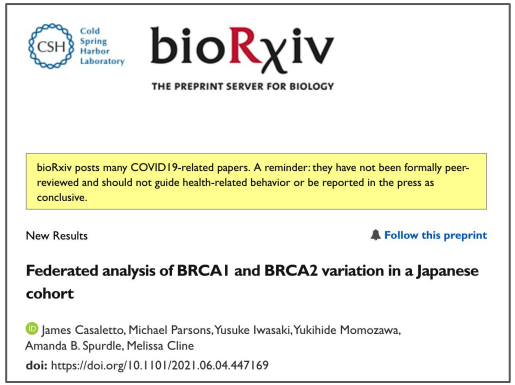
Federated analysis of BioBank Japan Data

- 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!

BIOBANK JAPAN

Yukihide Momozawa, RIKEN

RIKEN



Cold Spring Harbor Laboratory

bioRxiv

THE PREPRINT SERVER FOR BIOLOGY

bioRxiv posts many COVID19-related papers. A reminder: they have not been formally peer-reviewed and should not guide health-related behavior or be reported in the press as conclusive.

New Results

Follow this preprint

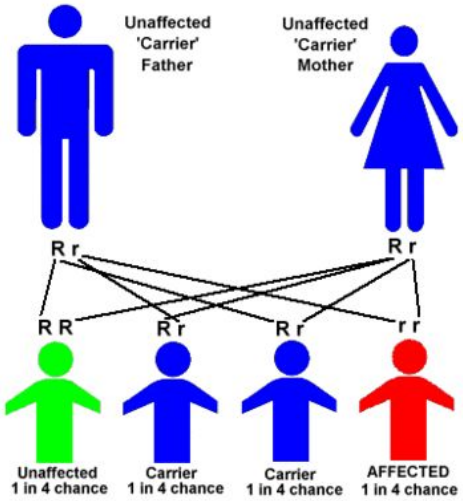
Federated analysis of BRCA1 and BRCA2 variation in a Japanese cohort

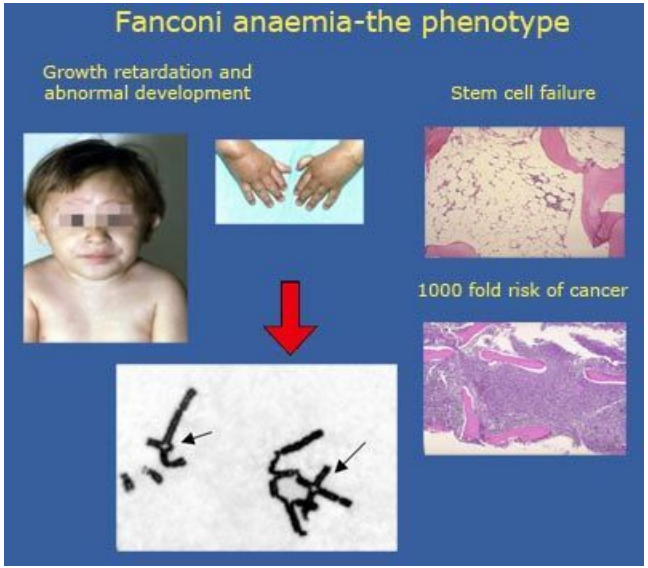
James Casaletto, Michael Parsons, Yusuke Iwasaki, Yukihide Momozawa, Amanda B. Spurdle, Melissa Cline

doi: https://doi.org/10.1101/2021.06.04.447169

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.





Gabriella Miller Kids First
PEDIATRIC RESEARCH PROGRAM
Data Resource Center

CAVATICA

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

NIH *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)

NIH *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

BioCompute Objects

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 Rätsch
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
NIH Big Data to Knowledge (BD2K)

NATIONAL CANCER INSTITUTE

NIH
National Heart, Lung, and Blood Institute

BioData CATALYST

NIH NATIONAL CANCER INSTITUTE
Informatics Technology for Cancer Research (ITCR)