Lessons learned from two decades of *BRCA1* and *BRCA2* genetic testing: the evolution of data sharing and variant classification

Amanda E. Toland, PhD[1], **BIC Steering Committee**[2], **Lawrence C. Brody, PhD**[3]

[1]Departments of Cancer Biology and Genetics and Internal Medicine, Comprehensive Cancer Center, The Ohio State University, Columbus, OH 43210 USA

[2]See Appendix for a full list of authors and affiliations.

[3]National Human Genome Research Institute, National Institutes of Health, Bethesda, MD 20892 USA

Nearly a generation (~24 years) has elapsed since the identification of the breast cancer susceptibility genes, *BRCA1*[1] and *BRCA2*[2]. Over that time the norms and policies surrounding the sharing of human genetic data have evolved. In this commentary, we examine the lessons learned about how data sharing can facilitate an understanding of the scope and consequences of genetic variation. Through this experience, we explore these lessons and their application to understanding human genomic variation.

The sharing of data among geneticists has waxed and waned through time. A notable nadir was reached during the race to identify the genes responsible for familial breast and ovarian cancer. The search for the *BRCA1* gene was characterized by intense competition and shifting alliances[3]. During the "gene hunt" phase, data sharing between (and even within) groups was minimal. After the *BRCA1* gene was identified in 1994[1], several of us called for a new, more open era to guide BRCA research in the future[4]. A tangible outcome of this call was the creation of an open access database, the Breast Cancer Information Core (BIC), in 1995[5]. The mission of the BIC was to accelerate research by gathering and freely sharing information related breast cancer genes. In particular, the BIC was established as a repository of germline variants in *BRCA1* and *BRCA2* (collectively, *BRCA*) in an effort to record all sequence variants and ensure that this information was freely available to the research community. The BIC has been in continuous operation for over two decades and has been cited in more than 2,700 publications. (https://research.nhgri.nih.gov/bic/).

Sharing Human Variant Data – The Early Days

From inception, the BIC used the then new World Wide Web to share data with anyone with an internet connection. The inspiration for using the web to distribute human genetic variant data came from the cystic fibrosis gene pathogenic variant database established by Lap Chi Tsui in Toronto[6]. Perhaps the most well-known single gene database at the time, this list

of *CFTR* variation was distributed by Dr. Tsui to subscribers each month via fax. One of us (LCB) sat near the fax machine and collected page after page as the CFTR "database" streamed onto the floor. In addition to saving paper, we thought that sharing information digitally would allow investigators to import and analyze the data directly.

The BIC website debuted in 1995. To place this event in context, the first widely used web browser, NCSA Mosaic, was introduced in the fall of 1993; Amazon Inc. was established in 1994; and Google would not debut for another three years. The BIC was sharing data a year before the Human Genome Project proposed the "Bermuda Principles," the plan that called for the prepublication release of genomic sequences (https://web.ornl.gov/sci/techresources/Human_Genome/research/bermuda.shtml).

The earliest *BRCA* data deposits were provided by researchers conducting sequence analyses of research participants. BIC was one of the first databases that provided free access to individual level, unpublished data, enabling the community to advance research and clinical studies[4]. Later, as testing moved from research to clinical labs throughout the world, the latter became the main sources of data. For more than a decade, the main United States testing lab, Myriad Genetics, freely shared their *BRCA* pathogenic variant data via the BIC. Although Myriad Genetics ceased contributing data to the BIC in 2006. Without Myriad, the volume of data being deposited decreased greatly and the main depositors were academic labs and non-US based clinical labs. Data volume changed again in 2013 (see below). In the last four years, more than fifty clinical testing laboratories have embraced an open access model and deposited tens of thousands of variants to public databases[7].

The collaborative relationship between the BIC, testing laboratories and researchers demonstrated the importance of capturing unpublished data directly from clinical labs, namely, it facilitates and expedites the classification of variants. For example, even in the absence of data on formal control samples, it quickly became clear that some missense variants, originally thought to be pathogenic, were actually benign population variants[8,9]. This practice of data sharing, pioneered by the BIC, has expanded to other loci as well, as clinical genetic testing laboratories recognize the value of data sharing in moving the field forward.

Classification of Variants of Uncertain Significance

During its first decade, the BIC's main user base were scientists who found value in having easy access to *BRCA* variant data. Importantly, scientists were comfortable classifying variants as clinically significant, benign, or unknown. The BIC operating principles were to share data and have the scientific community determine the functional significance of each allele. This approach worked well until large numbers of clinicians, diagnostic laboratory staff, and even patients themselves registered to use BIC data. Of particular interest were variants of unknown significance (VUS), i.e., variants whose functional consequences were unknown. Such a clinical test result can be difficult to explain to patients and many clinicians are inexperienced in understanding the inherent uncertainty in genetic testing. The BIC Steering Committee recognized the VUS problem created by declaring a variant "uncertain" and developed a more consistent classification process managed by the steering

committee. Classifications of clinical significance were made following discussions that weighed all available data and relied on member expertise and experience. This process was successful but resource-limited; therefore, a more robust and scalable approach was required[10-12].

The **E**vidence-based **N**etwork for the **I**nterpretation of **G**ermline **M**utant **A**lleles (ENIGMA)[13] (https://enigmaconsortium.org) grew out of the BIC SC in 2009 to promote large-scale collaborative studies and standardized approaches to assess the clinical significance of *BRCA1* and *BRCA2* variants and other breast cancer susceptibility genes. The defining feature of the ENIGMA approach is the integration of multiple types of data[14]. ENIGMA developed a set of likelihood-based rules for BRCA variant classification. These rules derive quantitative and qualitative measures by comparing the behavior of known pathogenic and non-pathogenic alleles with regard to multiple phenotypes, e.g., segregation in families, tumor pathology, associated cancers, phylogenetic analysis. Conceptually, these are similar to the classification criteria for mismatch repair genes developed for inherited colon cancer[15] and formalized by the International Society for Gastrointestinal Hereditary Tumors (InSIGHT)[16] (http://www.insight-database.org/classifications/). A uniform structured classification criteria should result in objective variant classification. In this way, the hereditary breast and ovarian cancer and hereditary colon cancer research communities have been able to move beyond "expert opinion" as the main mode of variant classification. Open and transparent classification methods also creates a community of professionals who initiate inter-laboratory discssions when discordant classifications are reported. National organizations, such as the American College of Medical Genetics and Genomics (ACMG) /Association of Molecular Pathology (AMP) have developed their own guidelines to serve as a more generic framework for variant classification of Mendelian diseases. These recommendations, are based on a structured review of different types of qualitative evidence with pre-assigned weights[17,18].

Shifting Landscapes

In the late spring of 2013, one technological advance and one judicial ruling irreversibly changed the landscape of genetic testing for susceptibility to inherited cancer. Technical progress came in the form of massively parallel sequencing technologies, which led to multiplexed DNA sequence-based testing. Tests could now easily include five to fifty putative cancer susceptibility genes for a lower cost than single gene tests. The second event occurred in June of 2013 when the US Supreme Court unanimously invalidated Myriad Genetics' patents on the *BRCA* genes. In the US, immediately after this ruling new clinical labs entered the *BRCA1* and *BRCA2* test market. In this competitive environment, the cost of a combined *BRCA1* and *BRCA2* test dropped from ~$4,000 USD to less than $400 USD.

These changes in the testing landscape greatly increased the amount of *BRCA* sequence data being generated[19]. Multiple commercial laboratories began sharing *BRCA1* and *BRCA2* variants from all patients with the BIC. The BIC curation pipeline could not process this volume. In response, the BIC began processing these new data in conjunction with the National Center for Biotechnology Information (NCBI). This represented a break from the past where locus specific databases (LSDBs) were curated by small groups of

collaborators. Using the BIC as a model, NCBI created a new aggregation of LSDBs, dubbed ClinVar (https://www.ncbi.nlm.nih.gov/clinvar/). ClinVar now contains variant data for many clinically-relevant genes, and includes all historical BIC data as well as newly sequenced variants for *BRCA1* and *BRCA2*. Transferring the data acquisition, archiving, and display from the BIC to ClinVar has two advantages. ClinVar employs dedicated staff to process, curate and display large datasets. In addition, as an integral part of the NCBI, ClinVar has a commitment to archive data permanently.

The Need for Expert Panels

For patients undergoing clinical *BRCA* testing, the VUS rate ranges from 2% to 15% depending on the testing laboratory and patients' ethnic background[20-22]. While the proportion of VUS results has substantially decreased since the early 2000s (due to research and classification efforts), a significant number of individuals are informed that they carry a VUS. Widespread data sharing can help to decrease the rate of VUS test results, as increased knowledge about both phenotypes and allele frequencies contribute to variant classification.

ClinVar is now the largest source of directly deposited *BRCA* variant data. ClinVar staff do not evaluate the biological or clinical impact of variants. Instead, ClinVar compiles and shares variant classifications performed both by labs submitting variants and by "Expert Panels" that evaluate variants deposited by others using as many resources as possible. ENIGMA serves as an expert panel for the *BRCA1* and *BRCA2* genes in ClinVar. Even for well-curated genes such as *BRCA1* and *BRCA2*, the interpretation of variants is one of the largest hurdles in dealing with the massive amounts of data generated through gene panels, exome, and genome sequencing. Successful VUS classification relies heavily on open access, transparent data. Open access data also allows other groups to download and redistribute data with significant enhancements. An example of this is the newly created BRCA Exchange (http://brcaexchange.org), which is striving to facilitate collection of variants and associated clinical data from around the world and display this information using a clinician and patient accessible interface.

BRCA Testing Evolves and Expands

Twenty years ago, genetic testing for *BRCA* was offered in a limited number of academic clinical centers, and only to those who had a high prior probability of carrying a clinically significant variant. Today, hundreds of thousands of genetic tests are ordered annually in a variety of settings. Exome and genome sequencing are used clinically, particularly for undiagnosed pediatric patients and rare mendelian disorders. Exome sequencing and gene panel testing is being used to find somatic pathogenic variants in tumors. Genetic testing of *BRCA* to guide treatment options such as PARP inhibitors is currently recommended for ovarian cancer and metastatic breast cancer and may become the standard of care for other cancers[23]. There have also been calls for population-based screening of *BRCA*[24,25], but testing of unselected individuals is controversial. Undoubtedly, the increased screening for *BRCA* variants, both directly and as a secondary finding, will increase the number of VUSs reported. Ongoing deposition of these new variants and associated clinical data into public databases will be vital if expert panels are to continue their classification and resolve

VUSs[26]. While great progress has been made in this area, the sharing of variant data is not yet universal. Complete ascertainment of data will require changes in culture, polices, and business models, some of which hold that the patient data they generate transforms into proprietary information.

The Path Forward

For the last two decades, LSDBs were the main way gene-specific data were collected, stored, curated, and distributed to the community. There are several reasons for this: historically, individual scientists were experts on single genes or gene families; in the early days of sequence data acquisition there was no standardization of database architecture; and sequencing of large numbers of genes across individuals was not yet feasible. Computationally, LSDBs represented a "tower of Babel" as each database custodian collected data in an organic way and developed their own data fields, codes, and methods of presenting data. This heterogeneity inhibited centralization. In 2013, it was estimated that there were over 2000 databases on genes and diseases worldwide[27]. Because of these issues, national centers such as NCBI, EMBL-EBI, and other groups operating central databases were not interested in absorbing LSDBs. The separation of LSDBs from central sequence data narrowed with the widespread acceptance of the Leiden Open Variation database (LOVD). The goal of LOVD is to provide a "flexible, freely available tool for *Gene-centered collection and display of DNA variations*" (http://www.lovd.nl/). As a large number of LSDBs adopted this format, it became easier for centralized databases, such as ClinVar, to import the locus specific information. It also enabled functional and other data to be integrated according to standardized guidelines applicable to any gene or genomic locus.

Difficult issues relating to clinical data collection on a genome-wide scale remain. One of the largest is securing sufficient and stable funding to cover the personnel and computational infrastructure required to coordinate data collection and distribution and variant curation and classification. Those depositing data also require resources to collect and prepare the data for submission. It is difficult for academics to secure grant funding for these activities, and commercial entities must use their own funds to support data sharing. When financial support for submission is no longer available, data flow stops. Curtailing either submission or curation leads to a database quickly becoming outdated. In theory, computational methods could make the entire process less labor intensive. However, the availability of large amounts of clinical sequencing data has revealed that "one size fits all" *in silico* based variant classification tools perform very poorly unless thay are used in conjunction with additional data such as functional assays or multifactorial models. For genes associated with very rare diseases, there may only be a small number of individuals with the expertise to appropriately assess the data. Gene-specific knowledge of elements, such as key functional domains, disease-associated functions and types of variants that are causal of phenotype, remains important and are the basis for the ACMG/AMP classification scheme. Thus, the long-term need for locus-specific experts will continue.

As we move from single genes to genome sequences, we will need to determine what features of variant classification can apply to many genes and what needs to be considered on a gene-by-gene basis. The newly enacted regulations covering, and the emerging

awareness of, data privacy may further complicate the sharing of individual multilocus data. Finally, even with these frameworks in place and extant expert panels for all genes, there is a need to acknowledge the importance of quality control, analytical validity, and data interpretation. Higher throughput sequencing technology has its own weak spots in terms of analytical validity, read depth, coverage of specific regions, pseudogenes, and large rearrangements. The use of national oversight on clinical sequencing data from organizations such as College of American Pathologists, US Clinical Laboratory Improvement Amendments, Euro QC network (and others) is essential.

Conclusions

One of the critical questions moving forward is how to scale variant curation and interpretation to cover the thousands of genes associated with Mendelian disorders. Errors in classification or annotation can have clinical consequences. For example, several *BRCA* variants have been down-graded from pathogenic to VUS, a situation particularly likely when such variants have been identified in understudied populations, where control data might not have been available at the time of original classification[28, 29]. For individuals who had prophylactic mastectomies based on inaccurate classification or misinterpretation, this impact is real[30]. This underscores the importance of obtaining genetic variation data from populations of diverse ancestry. This can be achieved by infusing the culture of data sharing into genetic testing labs across the globe and ensuring broad access to genetic testing services to underrepresented populations. The large numbers of clinical tests being performed, the increasing wllingness of academic and commercial interests to share data and the existence of expert panels to provide ongoing classification created a virtuous cycle. The actions of the inherited cancer susceptibility research community can serve as a model for scaling of variant curation.

One lesson we can take from the classification of variants in *BRCA* and *BRCA2* and other cancer-predisposition genes is that there is not a universal approach to variant classification. For each gene/syndrome, classification of variants using integrated multifactorial models may require creating gene-specific tools and collecting disease-specific phenotypic data. It is critical not to lower our standards on what evidence is required for variant classification. Over twenty years of *BRCA* research and extensive testing data were required to arrive at our current depth of knowledge. Moving forward, we expect that the pace of variant classification and integration of genetic data into clinical settings will increase, led not only by technological innovations but also by our evolving understanding of the data required for each gene.

The history of variant classification for inherited breast and ovarian cancer has produced a set of best practices for the *BRCA* genes. This history can inform the field as we endeavor to understand variation in other genes. Generating such knowledge takes energy, time, and funding to generate and disseminate. In the short term, we need to be honest, comfortable, and transparent with the elements of uncertainty currently present when evaluating the clinical impact of genetic variation. The sharing of sequence and phenotypic data by researchers and clinical testing labs from around the world, serving multiple diverse populations, is essential to the classification process. We need to be aware of what has been

done before so as not to "re-invent the wheel" but rather to leverage the strides that have been made in understanding the phenotypic implications of genetic variation.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

References

1. Miki Y et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 266, 66–71 (1994). [PubMed: 7545954]

2. Wooster R et al. Identification of the breast cancer susceptibility gene BRCA2. Nature 378, 789–92 (1995). [PubMed: 8524414]

3. Davies K & White M Breakthrough : the race to find the breast cancer gene, ix, 310 p. (J. Wiley, New York, 1996).

4. Friend S et al. Breast cancer information on the web. Nat Genet 11, 238–9 (1995). [PubMed: 7581445]

5. Szabo C, Masiello A, Ryan JF & Brody LC The breast cancer information core: database design, structure, and scope. Hum Mutat 16, 123–31 (2000). [PubMed: 10923033]

6. Tsui LC & Dorfman R The cystic fibrosis gene: a molecular genetic perspective. Cold Spring Harb Perspect Med 3, a009472 (2013). [PubMed: 23378595]

7. Landrum MJ et al. ClinVar: public archive of interpretations of clinically relevant variants. Nucleic Acids Res 44, D862–8 (2016). [PubMed: 26582918]

8. Mazoyer S et al. A polymorphic stop codon in BRCA2. Nat Genet. 14, 253–4 (1996). [PubMed: 8896551]

9. Wagner TM et al. Global sequence diversity of BRCA2: analysis of 71 breast cancer families and 95 control individuals of worldwide populations. Hum Mol Genet 8, 413–23 (1999). [PubMed: 9971877]

10. Goldgar DE et al. Integrated evaluation of DNA sequence variants of unknown clinical significance: application to BRCA1 and BRCA2. Am J Hum Genet 75, 535–44 (2004). [PubMed: 15290653]

11. Easton DF et al. A systematic genetic assessment of 1,433 sequence variants of unknown clinical significance in the BRCA1 and BRCA2 breast cancer-predisposition genes. Am J Hum Genet 81, 873–83 (2007). [PubMed: 17924331]

12. Greenblatt MS et al. Locus-specific databases and recommendations to strengthen their contribution to the classification of variants in cancer susceptibility genes. Hum Mutat 29, 1273–81 (2008). [PubMed: 18951438]

13. Spurdle AB et al. ENIGMA--evidence-based network for the interpretation of germline mutant alleles: an international initiative to evaluate risk and clinical significance associated with sequence variation in BRCA1 and BRCA2 genes. Hum Mutat 33, 2–7 (2012). [PubMed: 21990146]

14. Whiley PJ et al. Multifactorial likelihood assessment of BRCA1 and BRCA2 missense variants confirms that BRCA1:c.122A>G(p.His41Arg) is a pathogenic mutation. PLoS One 9, e86836 (2014). [PubMed: 24489791]

15. Thompson BA et al. A multifactorial likelihood model for MMR gene variant classification incorporating probabilities based on sequence bioinformatics and tumor characteristics: a report from the Colon Cancer Family Registry. Hum Mutat 34, 200–9 (2013). [PubMed: 22949379]

16. Thompson BA et al. Application of a 5-tiered scheme for standardized classification of 2,360 unique mismatch repair gene variants in the InSiGHT locus-specific database. Nat Genet 46, 107–115 (2014). [PubMed: 24362816]

17. Richards S et al. 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, 405–24 (2015). [PubMed: 25741868]

18. Tavtigian SV et al. Modeling the ACMG/AMP variant classification guidelines as a Bayesian classification framework. Genet Med (2018).

19. Chen Z et al. Trends in utilization and costs of BRCA testing among women aged 18-64 years in the United States, 2003-2014. Genet Med 20, 428–434 (2018). [PubMed: 28933789]

20. Eccles DM et al. BRCA1 and BRCA2 genetic testing-pitfalls and recommendations for managing variants of uncertain clinical significance. Ann Oncol 26, 2057–65 (2015). [PubMed: 26153499]

21. Harrison SM et al. Clinical laboratories collaborate to resolve differences in variant interpretations submitted to ClinVar. Genet Med 19, 1096–1104 (2017). [PubMed: 28301460]

22. Lincoln SE et al. Consistency of BRCA1 and BRCA2 Variant Classifications Among Clinical Diagnostic Laboratories. JCO Precis Oncol 1(2017).

23. Buchtel KM et al. FDA Approval of PARP Inhibitors and the Impact on Genetic Counseling and Genetic Testing Practices. J Genet Couns 27, 131–139 (2018). [PubMed: 28780754]

24. Levy-Lahad E, Lahad A & King MC Precision medicine meets public health: population screening for BRCA1 and BRCA2. J Natl Cancer Inst 107, 420 (2015). [PubMed: 25550384]

25. Foulkes WD, Knoppers BM & Turnbull C Population genetic testing for cancer susceptibility: founder mutations to genomes. Nat Rev Clin Oncol 13, 41–54 (2016). [PubMed: 26483301]

26. Kurian AW et al. Gaps in Incorporating Germline Genetic Testing Into Treatment Decision-Making for Early-Stage Breast Cancer. J Clin Oncol 35, 2232–2239 (2017). [PubMed: 28402748]

27. Marshall E Biomedicine. NIH seeks better database for genetic diagnosis. Science 342, 27 (2013). [PubMed: 24092711]

28. Slavin TP et al. Prospective Study of Cancer Genetic Variants: Variation in Rate of Reclassification by Ancestry. J Natl Cancer Inst.. doi: 10.1093/jnci/djy027. [Epub ahead of print] (2018)

29. Mersch J, et al. revalence of Variant Reclassification Following Hereditary Cancer Genetic Testing. JAMA 320, 1266–1274 (2018). [PubMed: 30264118]

30. Bever L 'Damaged for the rest of my life': Woman says surgeons mistakenly removed her breasts and uterus. in The Washington Post (Washington, D.C., 2017).