AmbryShare: A new model to fostering a global effort for sharing genomic data to better understand human disease

Layla Shahmirzadi[1], Aaron Elliott[1], Hsiao-Mei Lu[1], Shuwei Li[1], Jill Dolinsky[1], Carin Espenschied[1], Charles Dunlop[1]

[1]Ambry Genetics, Aliso Viejo, CA, United States

In the era of genomic testing, large amounts of data are available per individual genome sequenced. This data has proven to be valuable at the individual level to determine pathogenicity of variants. It has also proven useful at the cohort level for identification of rare disease-causing variants, discovery of genes implicated in a disease, and identification of protective genes. A few large general population datasets are available for use and are typically used as reference data, however some merge exome data from multiple sources using various platforms and techniques, presenting significant challenges in data analysis. Some datasets also exist on large disease cohorts, which are very valuable for studying genetic etiology, new drug targets and other treatment applications of genetic information, but most are not available to the research community at large. This data is often kept private or made proprietary to support internal research efforts and novel discoveries or sold for profit to pharmaceutical companies who purchase the data and race to develop new targets. Here we provide an example of a paradigm shift for the use of genomic data to promote a global effort that freely shares data to promote a collaborative effort to better understand the human genome.

In March of 2016, we launched AmbryShare. AmbryShare is a public database comprised of exome data from 10,000 individuals with a focus on breast and/or ovarian cancer. All samples were ascertained from individuals undergoing clinical testing at our laboratory. Selected anonymized samples underwent research grade whole exome sequencing at our laboratory, ensuring data integrity and following the same procedure of library construction, sequencing and variant calling. Sequence data was aggregated and entered into a publicly available database for use by the scientific community. The AmbryShare database may be easily searched online and researchers may request a full download of the aggregate data. Additional disease datasets will be similarly made available in the future. We hope that this project will be the first step toward better understanding the human genome.