nature genetics

Capture and release

Fostering scientific progress and ensuring that the community has access to human exome data can be difficult to do when faced with the divergent interests of patients, data generators, data funders and potential data users. We support the archiving of sensitive datasets in secure repositories with appropriate mechanisms in place to control access.

The publication of papers containing human exome datasets has increased dramatically in the past year. In the January 2010 issue, *Nature Genetics* published the first application of exome sequencing to identify the gene underlying a rare Mendelian disease (*Nat. Genet.* **42**, 30–35, 2010). In 2010, we published 6 studies that included human exome sequencing, and so far in 2011 we have published 18 studies that have used this approach.

It is well known that it is possible to detect individuals within pooled, de-identified genomic data (*Nat. Genet.* **41**, 965–967, 2009; *Nat. Genet.* **41**, 1253–1257, 2009), raising the possibility that the release of genomic data into the public domain could compromise the privacy of human research subjects. Whole-exome data also raises the possibility of incidental findings: that is, such data may be generated with the intent of discovering new mutations that cause a specified disease, but they carry the possibility of finding that a sequenced individual harbors known mutations associated with other diseases, including fatal diseases with no known treatments. Together, these concerns raise serious ethical issues in regard to the collection, publication and release of whole-exome datasets.

The long tradition of data sharing between colleagues fosters the advance of science, as scientific discoveries build upon themselves and each other. However, the commercial and translational potential of genomic discoveries (and the associated high cost of making such discoveries) can inhibit data sharing because those making the discoveries may seek to protect potential benefits that they, their funders and their institutes may reap. Despite the different interests that may exist among patients, data generators, data funders and publishers, all those who participate in the collection and publication of human exome data ought to agree that scientific progress should be promoted and not hampered. The cost of archiving secure data is small compared to the cost of collecting human samples and obtaining consent agreements from patients.

One of the first questions a researcher should be able to answer after sequencing an exome is, “Has this gene variant ever been seen before?” Tools that would make it possible to answer this question easily are not available, although projects such as the NHLBI Grand Opportunity Exome Sequencing Project (ESP) are disseminating such data generated with funding from the NHLBI through the Exome Variant Server (http://snp.gs.washington.edu/EVS/). Diseasome is another useful database that integrates information on genes, genetic variation and disease (http://diseasome.kobic.re.kr/index.jsp). The universal archiving of human exome data in secure repositories is another step toward this end, and there are two such repositories available in the community.

The purpose of the European Genome-phenome Archive (EGA) repository (http://www.ebi.ac.uk/ega/) is to provide a secure database for research data that can be shared but cannot be released openly. The EGA is equipped to store human exomes, genomes, cancer exomes and genomes, case-control genotype sets, RNA sequence transcriptome data and proteome data. EGA can also store phenotype data associated with samples that do or do not have accompanying molecular data. The EGA works directly with local Data Access Committees that limit access to the data to approved users. The EGA facilitates the applications of registered users for data access, and once the local Data Access Committee makes a decision, the EGA implements the decision.

The database of Genotypes and Phenotypes (dbGAP) is another repository (http://www.ncbi.nlm.nih.gov/gap) developed to archive and distribute datasets that link genotypes and phenotypes. dbGAP holds both open-access and controlled datasets; similar to the mechanisms in place at EGA, access to controlled data is limited to those who have been approved by an NIH Data Access Committee. Potential users can apply for access by submitting a Data Use Certification and describing their intended research goals. Potential users must declare that they will only use the controlled data for approved research, they will not attempt to identify individuals within datasets, data will not be shared with non-approved parties, the original data generators will be cited in any subsequent publications, and all conclusions derived from analysis of controlled data will remain in the public domain and not be subject to licensing requirements.

Archiving of human exome datasets in such repositories facilitates data-sharing among relevant users without broad release of sensitive information. We support these efforts and insist that authors explain their data management plan as a condition of peer review. We recommend that authors of papers with human exome datasets include a dbGAP or EGA accession code at manuscript submission. Upon publication, accession codes will be linked to the appropriate database, allowing potential users to easily locate the data. The advent of exome-sequencing technologies brings new responsibilities to data generators, publishers and potential data users. EGA and dbGAP are two secure repositories available to the scientific community at large that enable ethical and responsible use of human exome data, a resource that should be both safeguarded and shared. ■