Editorial
TIME TO DEFEND WHAT WE HAVE WON

The publication of the sequence of the human genome (Venter et al., 2001; The International Human Genome Sequencing Consortium, 2001) stands as a seminal accomplishment of the scientific community. That there was even a debate about whether this most fundamental of all biomedical science data sets would be available to the academic community is an indication of how fragile the consensus on data sharing really is. The field of bioinformatics depends on free and open access to data. As a community, computational biologists must stand in support of open data.

It is tempting to look back on the origins of EMBL, GenBank, the Protein Databank and Medline as halcyon days, but this ignores many hard fought battles and the dedicated work of leaders who forged a consensus of open data in biomedical databases. There will always be a temptation for scientists to want to control the fruits of their labor, but publishing conclusions without providing access to the data that supports those conclusions is not science. Deposition of data in public repositories guarantees an archival record, allows broad searches across multiple data sets and frees the investigator of the need to maintain repositories themselves. For science to proceed, scientists must have access to the results of their peers. The Human Genome Project, including the Celera project, could not have succeeded without free sharing of data.

The challenges to open data release are many, and the scientific community must stand vigil to ensure that the one exception to data release policies made in accepting a truly seminal paper (Venter et al., 2001) for publication in an academic journal does not become the breach that destroys the levee. In the US, universities are required by the Bayh-Dole act to attempt to commercialize the results of scientific research, and many have interpreted this as a requirement to prevent researchers from freely disseminating software and data. Publication, including dissemination through web sites, is a part of the scientific process and should not be restricted.

The field of global gene expression analysis began with encouraging examples of data release set by the pioneering labs, but this early lead is in danger of slipping away. Progress in populating the public repositories such as the EBI ArrayExpress and NCBI Gene Expression Omnibus has been notably slow. Global gene expression studies can be very useful in assessing whether experimental conditions have been reproduced, and even if different labs are using the same cell lines. Journals accepting publications based on the use of global expression analysis and funding agencies supporting this work should require that the full expression data sets be deposited in public repositories just as is the case with sequence and structure data.

Not all is bleak. We now enjoy access to a remarkable range and depth of biomedical information. In a different light, the Celera publication (Venter et al., 2001) can be viewed as a remarkably generous gift of data produced at private expense to the scientific community. We certainly hope that other private concerns involved in science will follow this lead.

REFERENCES

David States
Associate Editor
states@ccb.wustl.edu