

Compete, collaborate, compel

Procedures for microattribution need to be established by journals and databases so that data producers have an overwhelming incentive to deposit their results in public databases and thereby to receive quantitative credit for the use of every published data accession.

The excellent work of database designers at (for example) the National Center for Biotechnology Information (NCBI) and the European Bioinformatics Institute (EBI) offers researchers a range of stably funded and user-friendly data repositories in which data producers and annotators can be uniquely identified along with their data across the ubiquitous Web. Accession numbers to database entries are routinely used for data retrieval. They should now also be used to accrue quantitative credit for their authors in a systematic process of micro-attribution.

Every field of scientific discovery goes through three main stages. First, competitive discoveries result from innovations. Second, competitive collaborations are ventured in an important process of negotiated pooling of resources. Third, standards are agreed upon and data begin to be shared by agreed protocols and databases.

Public and not-for-profit funding has generated resources in genomics and genome-wide tools for genetics such as genome sequences and the HapMap. Genotyping and sequencing costs have fallen dramatically, largely due to profit-driven industrial research, but also greatly enabled by the efforts of academic researchers and not-for-profit funders like the US National Human Genome Research Institute and the UK's Wellcome Trust. These funders are, therefore, in an excellent position to maximize the utility of the research they fund by insisting that it be rapidly deposited in public databases—in the case of genome sequences, in advance of the publications that bring credit to the data producers. When is the right time to insist upon data release of genome-wide associations (GWASs) and other large genetic data sets? And what is the best way to demonstrate the benefit to data producers?

The experience of the microarray field has been that mandatory data release has helped to regularize standards of proof and enable replication, which have been achieved in the GWAS field without widespread data release. With the stellar exception of the CGEMS project (<http://cgems.cancer.gov/>), the value to other researchers of publicly deposited data remains to be demonstrated. Full data release may be necessary, but it may not be sufficient for a researcher's needs. Even the best-designed database is no substitute for the knowledge and cooperation of the data producers themselves.

In genomics, the 2003 Fort Lauderdale statement on data release (<http://www.wellcome.ac.uk/assets/wtd003207.pdf>) is an example of inspired policy but incomplete implementation. Its guidelines recommend contacting data producers and respecting their published intentions for the data they produce. If the data producers agree to collaborate, reconciliation of opposed interests and conflicting observations can lead to a synthesis with powerful potential for clarification

and discovery. If both groups have generated data, the collaboration can lead to discoveries with the power that more data confers.

If no collaboration is possible, it is difficult for the data user to give full credit except by thanking producers for depositing their data, citing their published intentions, listing exhaustively all the sequence traces and assembly by accession number and then detailing exactly how the user made use of this data. To achieve this consistently—for accessions of many different kinds of data—requires a heroic structural solution at the journal level. Each paper might append a table of accessions and uses. However, authors would fill this in only if there were tools to help them connect the databases and journals and if there were some further incentive for giving credit to accession codes. That could be achieved if the journals would produce league tables of most-cited data items—those that are currently most useful to other researchers and hence of most interest to funders and appointment committees.

A study of papers that were published in this journal in 2006 (http://blogs.nature.com/ng/freeassociation/2007/02/duke_of_url.html) shows that citation of the first deposition of the authors' own data is achieved in accordance with journal policy, but that reporting of other accessions used in the study is neither extensive nor systematic. We neither compel it nor currently reward it, and we are reluctant to implement compulsion without appropriate incentives.

Database bioinformaticians can also help the process of microattribution by providing the tools to quantitate citations to their entries in publications and annotations in other databases. High-throughput data sets (particularly those without validated biological information) should be counted, but the greatest value in giving credit to high-input efforts comes in the area of genotype-phenotype correlations. Curated and peer-reviewed phenotypic annotations (trait measurements and diagnostic categories) attached to gene variants (SNPs, mutations and structural alterations) and indexed to genome sequences could be further highlighted with quantitative counts of annotation activity and publication. In this way, the intensity of effort and interest in specific areas of genotype-phenotype investigation would be evident to the genetics community, locus by locus (see also the editorial in the April issue: *Nat. Genet.* **39**, 423 (2007)).

When requiring authors to deposit data in public databases, journals, databases and funders should ensure that quantitative credit for the use of every data entry will accrue to the relevant members of the data-producing and annotating teams. In an era in which consortia are producing more (and more useful) papers than individuals and small groups, the careers of individuals are as much in need of specific credit as those of the scientific visionaries and wranglers who hold the consortia together. ■