nature genetics

Integrating with integrity

Data worthy of integration with the results of other researchers need to be prepared to explicit export standards, linked to appropriate metadata and offered with field-specific caveats for use. Data exchange may need to be taught and discussed in handshaking workshops.

Data 'sharing' is a misnomer in that it is usually insufficient simply to make one's results available to the greater community. Not all data are fit for use by others. Each field's experts recognize and use a range of quality measures and caveats that may be more difficult to adapt for other applications than the data themselves. If different analytical approaches are demonstrated to be truly independent or 'orthogonal', then there may be value in combining them to increase power or generate new hypotheses. However, the independence of the datasets cannot protect against uncritical use of too much or too little data. Sample sizes, selection criteria, statistical significance, number of hypotheses tested, normalization and scaling procedures, read depth and sequence quality scores are all important considerations that can be misunderstood or missed in combining and reanalyzing data.

Whether integrative approaches are useful may depend upon whether integration preserves or destroys essential information. For example, disease-specific network and pathway models can be built that incorporate genes associated with a disease at various significance levels (for example, by selecting *P* values below 10^{-2}, 10^{-5} or 10^{-8}), but how sound is the evidence supporting a connection between the marker SNP tested in a GWAS and the transcription unit that was put into the model? Few SNPs are found within coding regions of genes, so most will, at best, be connected to nearby genes by perfect linkage disequilibrium in a particular population. There is less justification for adding to the model those genes situated within a window that is an arbitrary number of megabases from the tag SNP or those genes found together with the SNP in a published abstract. The same concerns apply if expression data are to be integrated with epidemiological results (for example, to prioritize a long list of SNPs that pass an initial significance cutoff). Results may be sensitive to whether the statistical significance of differential regulation is calculated on each gene independently or on a gene set or pathway.

Integration is of most value in two areas: bioinformatic modeling, to predict the effects of genetic and environmental perturbation, and clinical utility, to increase the speed and accuracy of the transfer of preclinical knowledge to clinical trial. Funding bodies hope that encouraging researchers to integrate their results will reduce duplication of effort. Trivially, researchers can agree to work on the same systems and samples or to use agreed standard control materials, but this can be problematic in practice.

One case where researchers are converging on an export standard is in reporting the cumulative evidence that supports the hypothesis that a gene variant differs from the wild-type form (often referred to as the 'evidence for the pathogenicity of a mutation'). The prior odds that any variant is associated with disease are determined by the genome size (total variants) and the population frequency of the disease. These prior odds can be promoted by likelihood ratio support from four sources of evidence: mendelian segregation in disease pedigrees, frequency information collected in association studies, evolutionary conservation determined among orthologs in multiple alignment and quantitative functional assays. Rather than regarding the latter as a qualitative trump card, it should be possible to express the performance of the variants in a functional assay as a likelihood ratio so that it can be combined with support obtained by other researchers. Regardless of the assay, if the wild type and its genetic variants produce mean and standard deviation values, it should be possible to convert assay results to quantitative statistical support.

Researchers can enable integrative studies by publishing their quality metrics and exchange standards in a timely way in regularly versioned, citable preprints (as with plans for data release, *Nat. Genet.* **41**, 1045 (2009)) and by holding integration workshops between data producers and data users from different fields. These exchanges should focus on honest assessment of what data are ready for use and explain the quality metrics used and where the pitfalls lie in using the data. In return, data producers can increase the citability of their datasets by better understanding the metadata needed by users. Requirements for open data deposition and integration that do not include mechanisms to agree on, publish and use data standards risk inflating inconsequential 'integrative' bubbles. ■