nature genetics

A feat worth replicating

Research is undertaken on the assumption that the experimental results will stand up to independent replication by other scientists. The reality check of replication should keep published work useful as the foundation for future experiments. Here we argue, on grounds of utility, that there is a place for competition in generating community resources. The data generators are in pole position to coordinate resources built on their data and should be contacted even by those intending to compete. Replication by complementary techniques is preferable to duplication of effort.

Last month, *Nature* published a landmark resource for the genetics community, the finished sequence of the human X chromosome (*Nature* **434**, 325–337, 2005; see also page 343 of this issue). In this work, Ross *et al.* used independent data to assess the coverage and quality of the sequence: the deCODE linkage map, RefSeq transcription unit annotations and the end sequences of more than 17,000 fosmid clones. They also provide a complete annotation of the finished sequence in accordance with their previously published protocol, using a number of tools for gene prediction, similarity to cDNAs and ESTs and evolutionary conservation of nucleotide or predicted protein sequence. The annotation website (http://vega.sanger.ac.uk/Homo_sapiens/) includes updates to ensure a resource of lasting utility.

The annotation of the finished sequence of the human X chromosome marks a first of another kind. With a view to providing a guide to the protein-coding genes of the X chromosome for their own proteomic studies, Akhilesh Pandey and colleagues report on page 331 of this issue an independent public annotation of the human X chromosome. Despite using similar methods, Pandey and colleagues report a smaller number of genetic elements than Ross *et al.* This may reflect the facts that they worked on a smaller data set or applied more conservative criteria in annotating transcription units. But the real utility of their annotation may come from comparing the two groups' representations of the 699 (or 696) genes already known on the X. According to Pandey and colleagues, 45% of these are alternatively spliced, 22% include new exons and 35% exhibit exon skipping. They also report that 64 of 142 previously reported 'noncoding RNAs' are fragments of conventional protein-coding genes.

Because curation is distinct from sequencing and the tools and data were already in the public domain, Pandey and colleagues did not need to wait for the sequence generators to begin their own annotation. Likewise, in accepting funding to provide a community resource, the sequencers undertook to make the finished genome sequence publicly available as soon as it was of verified quality for release. These facts do not automatically entail conflict, because the Fort Lauderdale agreement (http://www.wellcome.ac.uk/assets/wtd003207.pdf) leaves room for competing groups to annotate genome sequence, as the following compromises and exhortations show:

Resource producers should

recognize that even if the resource is occasionally used in ways that violate normal standards of scientific etiquette, this is a necessary risk set against the considerable benefits of immediate data release.

Resource users should

respect the producer's legitimate interests as set out, e.g., in a Project Description, while being free to use the data in any creative way. There should be no restrictions on the use of the data but the best interests of the community are served when all act responsibly to promote the highest standards of respect for the scientific contribution of others. In some cases, this might best be done by discussion or coordination with the resource producers.

It does not require a Project Description to point out that annotation is an obvious use of the data. Respecting the wishes of sequencers to make a useful annotation from their own data is certainly in the interest of most researchers who use genome sequences. But not all users will have the same interests: competitive annotation of publicly available sequence data would be an excellent way to test new algorithms for gene prediction, or might be initiated by a group that had amassed a large collection of ESTs derived from rare tissue-specific transcripts.

Because some users will be looking for the accurate representation of a single genetic element rather than the most complete representation of the genome, there is no utilitarian reason to suppose the sequence generators' version should be used as the template and all competitive annotations as updates. Were a third party to generate a database or browser incorporating the annotation efforts of all previous groups, this would serve users just as well and indicate areas where the structure of genetic elements requires further experiment. To be useful, such a third-party tool would have to give correct attribution to each piece of annotation. Theft or 'rebranding' of data would lead to disuse.

Reputations in the resource field are made by priority and utility. We recommend that the best way to protect these is to circulate or publish project descriptions and communicate with data generators, who are in the best position to advise researchers on how to coordinate their efforts. Coordination can lead to more robust results through partially overlapping replication using complementary techniques, for example, whole genome shotgun sequencing for speed and library-based systematic sequencing for comprehensive coverage.

Resources generated with a large proportion of a funding body's investment might be deemed too expensive to be subjected to truly independent verification, but wasn't the intention always there? If the argument is economics against science, Pandey and colleagues have the answer: Baltimore and Bangalore offer comparative advantages for different parts of the project. ■