phisms (SNPs) [a number comparable to what is already publicly available (*21*)] Venter *et al.* (*1*) show that these new opportunities—to paraphrase another milestone article—"have not escaped their notice" (*22*).

### References and Notes

1. C. Venter *et al.*, *Science* **291**, 1304 (2001).
2. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
3. C. K. Stover *et al.*, *Nature* **406**, 959 (2000).
4. I. Dunham *et al.*, *Nature* **402**, 489 (1999).
5. M. Hattori *et al.*, *Nature* **405**, 311 (2000).
6. B. Ewing, P. Green, *Nature Genet.* **25**, 232 (2000).
7. H. Roest Crollius *et al.*, *Nature Genet.* **25**, 235 (2000).
8. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
9. The observed 40,000-fold variation in eukaryote haploid DNA content ("*C* value") is unrelated to organismic complexity or to the numbers of protein-coding genes; see T. Cavalier-Smith, *J. Cell Sci.* **34**, 247 (1978).
10. L. Huang, R. J. Guan, A. B. Pardee, *Crit. Rev. Eukaryotic Gene Expr.* **9**, 175 (1999)
11. J. W. Fickett, W. W. Wasserman, *Curr. Opin. Biotechnol.* **11**, 19 (2000).
12. D. L. Wheeler *et al.*, *Nucleic Acids Res.* **29**, 11 (2001)
13. F. Liang *et al.*, *Nature Genet.* **25**, 239 (2000).
14. F. Liang *et al.*, *Nature Genet.* **26**, 501 (2000). The cited mRNA number does not take into account ESTs only sampled once and not overlapping with any others ("singletons"). There are about 300,000 singletons in the The Institute for Genomic Research human gene index (HGI release 6.0).
15. E. Beaudoing *et al. Genome Res.* **10**, 1001 (2000).
16. S. Audic, J.-M. Claverie, "The first draft of the human genome: An academic and industrial perspective," workshop at the Max-Planck-Institut für Molekulare Genetik, Berlin, 1 to 2 October 2000.
17. A. A. Mironov *et al.*, *Genome Res.* **9**, 1288 (1999).
18. H. H. McAdams, A. Arkin, *Proc. Natl. Acad. Sci. U.S.A.* **94**, 814 (1997).
19. See www.airbus.com/; thanks to P. Emrich, Airbus Industry Provisional Services Manager.
20. S. Brenner, *Science* **287**, 2173 (2000).
21. There are 2,558,564 SNPs as of 8 December 2000 in dbSNP (www.ncbi.nlm.nih.gov/SNP/), including 801,776 mapped SNPs generated by the SNP Consortium (http://snp.cshl.org/data/).
22. J. D. Watson, F. H. C. Crick, *Nature* **171**, 737 (1953).

FUTURE DIRECTIONS: SEQUENCE INTERPRETATION

# Making Sense of the Sequence

**David J. Galas**

In this issue of *Science* on page 1304 and this week's issue of *Nature* appear versions of the sequence of the human genome (*1, 2*) that signal the dawn of a new era. For the research biologist, it is easy to think about the advantages of having the sequence of every gene of potential interest, but another thing altogether to think about how to find all of them and to validate their identities and structures. The use of genome sequences to solve biological problems has even been afforded its own label; for better or worse, it's called "functional genomics." This new way of doing biology means some real changes, many of which are well under way in the community.

Since the publication of the *Saccharomyces cerevisiae* genome in 1996 (*3*), we have become familiar with the use of the full genome sequence in investigations of gene expression patterns and controls, protein-protein interaction networks, and other biological problems (*4–6*). These investigations are marked by a global point of view that was simply not possible before we had the sequence. Although we still do not know the function of about a third of the yeast genes, we do know that all possible protein and RNA participants in cellular function are encoded in the sequence we have.

As simple as it sounds, to know that there are no other unknown genetic components that can provide alternative explanations of experimental results is a fundamental shift of perspective. This shift is beginning to transform our approach to science, enabling researchers to face the challenge of identifying all the molecular components of the cell, as well as understanding how they are controlled, interact, and function. From a picture of the "software" of the single cell, we can look to the future when researchers will begin building, with as fine a degree of resolution, an integrated view of the universe of cell-cell interactions, differentiation, and development from single cell to organism. The availability of complete sequences of *Drosophila melanogaster*, *Caenorhabditis elegans*, and *Arabidopsis thaliana* (*7–9*) is already beginning to revolutionize such studies, and this list may soon include significant sequence from other biological models of metazoan development.

Estimates from genes analyzed to date suggest that the average number of alternates spliced from the transcript of a single mammalian gene might be in the range of two to three or more. As the present sequence yields estimates of about 30,000 genes (*1, 2*), this would give us an estimated 90,000 or more distinct proteins encoded by the human genome, without considering proteolytic processing or posttranslational modifications. Thus, the complexity of the mammalian genome relative to that of yeast still presents formidable technical obstacles.

So how can the working biologist take advantage of all this new information and bring about the advances predicted? The first step is to understand that the present form of the available sequence information of the human genome is not a complete, fully annotated inventory of the human genes in each chromosome. Nor is the available sequence a single continuous and exact sequence for each chromosome. The reported genome sequence is represented by a set of sequences that cover the genome in a statistical sense but have a very large number of interruptions and gaps. Although the completeness and continuity will continue to improve, there are significant uncertainties when inferences are made from these data. The concept of the "contig" is essential to our understanding of this limitation. A contig is a contiguous piece of sequence information inferred by assembling sequence reads from single reactions (usually 400 to 800 bases in length). The number of contigs reported in the sequence data and their spectrum of sizes are important parameters in the analysis of genes. As of 12 December 2000, the public database at the U.S. National Center for Biotechnology Information (NCBI) reported that the largest contig in the entire available sequence was 28.5 megabase-pairs (Mb) in size; there were 43 contigs larger than 1 Mb, 566 contigs between 250 kb and 1 Mb, and 1628 contigs between 100 and 250 kb in size. This represented a total of approximately 600 Mb in contigs larger than 100 kb—less than 20% of the full sequence of the genome. As illustrated in figure 8 of IHGSC (*2*), half of the sequence lies in contigs 22 kb or smaller, though they can be joined to form larger contigs. We must distinguish here "initial sequence contigs," derived from sequenced clones, and "merged sequence contigs," derived by merging sequence contigs from overlapping sequenced clones [see figures 6 and 7 in (*2*)]. Because Venter *et al.* (*1*) assemble sequence contigs, not from sequenced clones, but from the entire collection of sequence reads, this distinction is not necessary in their report.

Because the average gene is of the same order of magnitude or larger than many of the contigs (a good estimate might be about 30,000 base pairs), this means that a significant fraction of human genes are unlikely to be represented on a single sequence contig in these data sets. The likelihood of finding one of the largest genes, such as Titin [~250 kb in size with >200 exons (*1*)] on a single contig is much smaller than for small, simple genes like the olfactory receptor genes, which average less than 2 kb (*2*). It will be a while before the gaps get filled in and the contigs are joined together.

Therefore, in the near future, many genes will have to be synthesized from an inferred organization of the contigs into a gapped mosaic of assemblies called "scaffolds." This

The author is at the Keck Graduate Institute of Applied Life Sciences, Claremont, CA 91711, USA. E-mail: David_Galas@kgi.edu

means that an even more important factor than continuity for using the sequence to construct models of genes is the uncertainty associated with positioning the contigs relative to each other. Ambiguities in order and orientation of the contigs will sharply increase the number of possible ways that the sequence can be fit together and will thereby obscure the actual gene structure.

The definition of a scaffold appears to be quite different in the two papers. Venter *et al.* (*1*) report that they built scaffolds by using the paired-end sequences of their plasmid clones to link together and orient sequence contigs. They could put together these chains of sequence contigs, in the right order and orientation and at known distances apart, because they used several, known sizes of plasmid clones for sequencing and always generated sequence pairs at known distances from each other. The advantages of relying on these kinds of sequence data were substantial in the assembly process. One of these advantages is that sequence contigs could be linked in the proper orientation and distance from each other even when they could not be merged into a single contig. Thus, the self-consistent assembly from these data would appear to have ensured a high level of order and orientation of contigs at every scale of length.

It may not be possible to fully assemble genes that fall into these scaffold segments if a gene segment falls into an unsequenced gap, but the picture of the gene that emerges should be fairly reliable. A gene would look something like the picture on a reconstructed Grecian urn (see figure, page 1259), with blank clay segments holding the places for the real, picture-completing fragments. A critical parameter for gene assembly and analysis for the Venter *et al.* approach is the size and coverage distribution of scaffolds [see figure 5 in (*1*)]. The average scaffold length reported was more than a megabase, with 25% of the genome in scaffolds of at least 10 Mb in size. As the average gap length between scaffolds was only 2 kb, this data set seems to represent a high level of coverage for gene analyzers, with a high level of consistent order and orientation.

**REPRESENTATIVE WEB RESOURCES FOR THE HUMAN GENOME: DATA ACCESS, TOOLS, ANALYSIS, AND ANNOTATION**

| | |
|---|---|
| **Public data and tools, NCBI, NHGRI** | **www.ncbi.nlm.nih.gov/Sitemap/index.html# Human Genome, www.nhgri.nih.gov/Data/, overview and guide to data and tools, additional links to a range of sites related to the genome project** |
| **Celera data, annotation, and tools** | **www.celera.com, central site for public access to data and tools, Celera Discovery System, and other resources** |
| **European Bioinformatics Institute (EBI/Sanger Centre)** | **www.ensembl.org/genome/central/, access to sequence information with automatic baseline annotation** |
| **University of California at Santa Cruz** | **http://genome.ucsc.edu, http://genome.cse.ucsc.edu, additional information on the nonredundant "working draft" of the Human Genome Sequence** |
| **RIKEN (The Institute of Physical and Chemical Research) and the University of Tokyo** | **www.rarf.riken.go.jp/ and http://hgrep.ims.u-tokyo.ac.jp/** |
| **Washington University** | **http://genome.wustl.edu/gsc/human/Mapping/, links to clone and accession maps of human genome** |
| **Genome Annotation Assessment Project (GASP1)** | **www.fruitfly.org/GASP1/, results of genome annotation tests, automated annotation of genomes, *Drosophila*** |
| **Baylor College of Medicine** | **www.hgsc.bcm.tmc.edu, another major genome sequencing center with useful data and tools** |
| **Oak Ridge National Laboratory (ORNL)** | **http://compbio.ornl.gov/tools/index.shtml, data and tools available include GRAIL and text-based searching http://compbio.ornl.gov/tools/channel/, a menu of tools for analysis and annotation** |
| **ORNL Genome Annotation Consortium and the Joint Genome Institute (JGI) (The University of California for the U.S. Department of Energy)** | **www.jgi.doe.gov/JGI_home.html, direct link to a collaborative annotation effort through a genome sequencing center** |
| **GENSCAN** | **http://genes.mit.edu/GENSCAN.html** |
| **GeneWise** | **www.sanger.ac.uk./Software/Wise2/** |

IHGSC (*2*) report that they built their scaffolds quite differently—largely by linking sequenced bacterial artificial chromosomes, BACs. This will still leave some sequence contigs within the BACs of the scaffold unordered or unoriented. The Grecian urn analogy does not fit here because the sizes and shapes of the gaps are not well known and, in some cases, the pieces may be in backwards or in the wrong order. The critical factor for the gene-analyzing biologist is the degree of ordering and orientation of contigs within the BACs that were linked to make the scaffolds, which is difficult to estimate from the report. Relevant measures include a reported overall estimate in the range of 10 to 15% for misordered or misoriented sequence contigs (*2*). This paper contains a useful new statistic to indicate sequence contig length or scaffold length that is systematically larger than the simple average length—the N50 length (the largest length such that 50% of all base-pairs are contained in contigs of this length or larger). The reported N50 length for sequence contigs was 82 kb (including data from the finished chromosomes 21 and 22), and the scaffold N50 was 270 kb. Direct comparison of these statistics with the averages from Venter *et al.* are not meaningful. To understand some of the subtleties, the interested reader will have to venture further into the data on distribution of lengths and other important complexities, keeping in mind the differences between the processes of assembly. It would appear, however, that the scaffold data reported in the *Science* paper, having 90% genome coverage with end-to-end, long scaffolds, is a powerful resource for the biologist that will steadily improve as new sequence fills in the contig gaps and resolves remaining ambiguities.

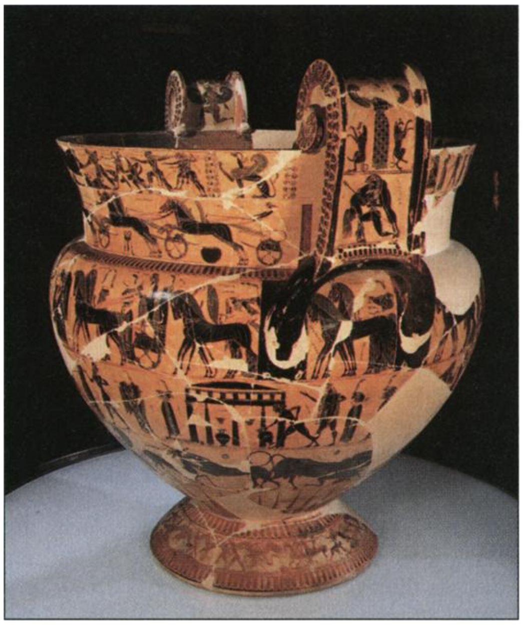
The effectiveness of finding genes by similarity to a given sequence segment is determined by a much simpler statistic, the total coverage of the genome by the collective set of sequence contigs. As the overall coverage of the genome is virtually complete (≥90%), there is a strong likelihood that every gene is represented, at least in part, in the data. Thus, finding any gene by sequence similarity searches using sufficient sequence to ensure significance is almost always possible using the data published this week. Caution must be exercised, however, as the identification of the gene may still be ambiguous. This is because a highly similar sequence from a receptor gene from *Drosophila*, for example, could be found in several different, homologous genes, which may have similar or entirely different functions or are nonfunctioning pseudogenes. In other words, common domains or motifs can be present in many different genes. The use of the approximate similarity search tool BLAST is probably still the best way to find similar sequences. The excellent primer at the NCBI site (*10*) should be used to understand the nature of the growing armament of BLAST-based tools, as well as the sometimes subtle issue of statistical significance and the limitations of this kind of approximate algorithm. For most purposes, the approximation used by the BLAST algorithm is irrelevant, but the user should be aware of the specific kinds of similarities that may be missed by each available form of the algorithm. For example, since certain kinds of interrupted similarities are ignored, the more widely separated two similar sequences are, the less reliable will be the assessments of statistical significance. Newer methods attempting to use the structural cues inherent in the coding sequence to detect similarities are pushing back the detection limits for significant similarity (*11*).

Although enormous progress has been made in automating the identification of genes in genomic sequence, building accurate models of genes from the sequence still requires a lot of human, "hands-on" effort. The best models are built of genes whose full-length mRNA sequences are available. The RNA sequence [in the form of complementary DNA (cDNA)] can be used to thread together the exon structure of the gene from genomic sequence no matter where the pieces may reside—continuity, order, and orientation of the fragments are not essential to this process. Of course, the presence of pseudogenes and highly similar repeats can defeat even this strategy. Nonetheless, this represents a strong argument for gathering much more full-length cDNA sequence data.

There are two general approaches to gene finding. The homology-based methods include the use of known mRNA sequences as well as gene families and interspecific sequence comparisons. The ab initio methods include detection of exons and other sequence signals, like splice sites, by various computational methods within the sequence being analyzed.

In every gene model, the location and structure of the sequences involved in regulation and control stands as one of the most difficult annotation problems. Finding and dissecting these important sequence regions can be done in some cases by means of motifs known to be conserved in transcription factor–binding regions (*12*), but our ability to define and predict control regions is currently rather poor and unreliable. Interspecific genome comparisons are one of the ways of getting at these regions, under the assumption that the regions will stand out as being conserved (*13*). New experimental methods like an array-based technique to locate genome-wide sites of action of transcription factors (*4*), will also make significant contributions to sorting out the cis-regulatory signals in the genome.

A number of tools are currently available for automated annotation, but a discussion of their advantages and limitations in specific circumstances is beyond the scope of this viewpoint. Approaches that use a combination of statistical and heuristic methods to recognize genes and gene features are prevalent (hidden Markov models, neural nets, and Bayesian networks are among the methods used). They are most effective, however, in finding genes, rather than modeling them accurately, and are usually used in concert with homology-based methods. Factors that can have strong effects on the effectiveness of such algorithms include errors in sequencing and statistical biases like base composition. Noise in the data can sharply degrade performance, so draft sequence, in which the error rate is higher, can be markedly inferior to finished sequence for ab initio prediction.



**Visualizing gene assembly.**

CREDIT: SCALA/ART RESOURCE

GENSCAN is a widely used piece of software for gene finding and prediction, but newer developments like Genie also look promising (*14*, *15*). Genie is a hidden-Markov–model system that allows for the integration of information from different sources such as signal sensors (splice sites, start codon, etc.); sensors of introns and exons; and alignments of mRNA expressed sequence tag (EST), and peptide sequences. Other software tools,. like GENEBUILDER, GLIMMERM, FGENES, GRAIL, and others, have also been reviewed recently (*16*, *17*). There is no one simple way to compare them, as they appear to perform differently in different tests. Using the *Drosophila* genome as a primary example, the Genome Annotation Assessment Project (GASP1—see table, page 1258) provides a very useful analysis of progress and problems in eukaryotic genome annotation (*18*). A similar comparison has been done using the *Arabidopsis* genome (*19*).

The two genome papers have used systems consisting of multiple tools to create their initial gene inventories. IHGSC (*2*) used a system called *Ensembl* that follows ab initio predictions by GENSCAN with mRNA, EST, and protein motif information comparisons for the initial predictions (*19*). It then uses a program called GeneWise, which has been used on the *Drosophila* genome (*20*), to extend protein matches. In contrast, Venter *et al.* (*1*) report the development of a rule-based expert system for annotation they call "Otto" that attempts to embed some human curatorial functions in software.

All these annotation efforts in the community are also being linked to new ways of visualizing genomes with their annotation (*18*, *21*, *22*). Genome browsers that enable the reader to navigate through many levels of genome information are now available that take the first steps in this direction. These tools can be accessed at several sites (see table, page 1258). Commercial firms are also beginning to market similar kinds of software and are likely to continue to develop sophisticated, user-friendly packages for these purposes.

In the future, when the annotation of the genome is complete, the information from the sequence will be indicated in agreed-upon terms that can be searched directly by the text of the annotation—for example, a gene will be found by its name, by its family, by the protein domains it codes for, etc. Clearly, a combination of sequence similarity searching tools, ab initio methods, and annotation-based searches must be used by researchers for the foreseeable future. The next stage of annotation will also require the integration of independent experimental information into the gene annotation. To fully explore the properties of complex, highly interactive systems, databases will need to have pointers that link a gene to other genes by a variety of causal interactions, such as gene product *X* has a binding partner *Y*, exerts control on the expression of gene *Y* through a cis-regulatory site, produces a metabolic product that interacts with the product of gene *Y*, or participates in the same (or linked) signaling pathway with the product of gene *Y*. The way to this future has been opened by the availability of sequence information. Now we have to learn to use it to understand the biology of the organism.

**References and Notes**

1. C. Venter *et al.*, *Science* **291**, 1304 (2001).
2. International Human Genome Sequencing Consortium (IHGSC), *Nature* **409**, 860 (2001).
3. A. Goffeau *et al.*, *Science* **274**, 546 (1996).
4. B. Ren *et al.*, *Science* **290**, 2306 (2000).
5. M. T. Laub *et al.*, *Science* **290**, 2144 (2000).
6. J. L. DeRisi, V. R. Iyer, P. O. Brown, *Science* **278**, 680 (1997).
7. M. D. Adams *et al.*, *Science* **287**, 2185 (2000).
8. The *C. elegans* Sequencing Consortium, *Science* **282**, 2012 (1998).
9. The *Arabidopsis* Genome Initiative, *Nature*, **408**, 796 (2000).
10. www.ncbi.nlm.nih.gov/BLAST/
11. S. F. Altschul, M. S. Boguski, W. Gish, J. C. Wootton, *Nature Genet.* **6**, 119 (1994).
12. K. Struhl, *Annu. Rev. Genet.* **29**, 651 (1995).
13. W. W. Wasserman *et al.*, *Nature Genet.* **26**, 225 (2000).
14. C. Burge, S. Karlin, *J. Mol. Biol.* **268**, 78 (1997)
15. M. G. Reese, D. Kulp, H. Tammana, D. Haussler, *Genome Res.* **10**, 529 (2000).
16. D. Haussler, *Trends Biochem. Sci.* **23** (suppl.), 12 (1998).
17. J. M. Claverie, *Hum. Mol. Genet.* **6**, 1735 (1997).
18. S. Lewis, M. Ashburner, M. G. Reese, *Curr. Opin. Struct. Biol.* **10**, 349 (2000).
19. N. Pavy *et al.*, *Bioinformatics* **15**, 887 (1999).
20. T. Hubbard, E. Birney, *Nature* **403**, 825 (2000).
21. E. Birney, R. Durbin, *Genome Res.* **10**, 547 (2000).
22. *Genome Res.* **10** (no. 4) (2000) [an entire issue devoted to genome annotation].

FUTURE DIRECTIONS: COMPUTATIONAL BIOLOGY

# Bioinformatics—Trying to Swim in a Sea of Data

**David S. Roos**

Advances in many areas of genomics research are heavily rooted in engineering technology, from the capillary electrophoresis units used in large-scale DNA sequencing projects, to the photolithography and robotics technology used in chip manufacture, to the confocal imaging systems used to read those chips, to the beam and detector technology driving high-throughput mass spectroscopy. Further advances in (for example) materials science and nanotechnology promise to improve the sensitivity and cost of these technologies greatly in the near future. Genomic research makes it possible to look at biological phenomena on a scale not previously possible: all genes in a genome, all transcripts in a cell, all metabolic processes in a tissue.

One feature that all of these approaches share is the production of massive quantities of data. GenBank, for example, now accommodates $>10^{10}$ nucleotides of nucleic acid sequence data and continues to more than double in size every year. New technologies for assaying gene expression patterns, protein structure, protein-protein interactions, etc., will provide even more data. How to handle these data, make sense of them, and render them accessible to biologists working on a wide variety of problems is the challenge facing bioinformatics—an emerging field that seeks to integrate computer science with applications derived from molecular biology. We are swimming in a rapidly rising sea of data...how do we keep from drowning?

Bioinformatics faces its share of growing pains, many of which presage problems that all biologists will soon encounter as we focus on large-scale science projects. For starters, few scientists can claim a strong background on both sides of the divide separating computer science from biomedical research. This shortage means a lack of mentors who might train the next generation of "bioinformaticians." Lack of familiarity with the intellectual questions that motivate each side can also lead to misunderstandings. For example, writing a computer program that assembles overlapping expressed sequence tag (EST) sequences may be of great importance to the biologist without breaking any new ground in computer science. Similarly, proving that it is impossible to determine a globally optimal phylogenetic tree under certain conditions may constitute a significant finding in computer science, while being of little practical use to the biologist. Identifying problems of intellectual value to all concerned is an important goal for the maturation of computational biology as a distinct discipline. "Real" biology is increasingly carried out in front of a computer, while an increasing number of projects in computer science will be driven by biological problems.

Further difficulties stem from the fact that bioinformatics is an inherently integrative discipline, requiring access to data from a wide range of sources. Without the underlying data, and the ability to combine these data in new and interesting ways, the field of bioinformatics would be very much limited in scope. For example, the widespread utility of BLAST for the identification of gene similarity (*1*) is attributable not only to the algorithm itself (and its implementation), but also to the availability of databases such as GenBank, the European Molecular Biology Laboratory (EMBL), and the DNA Data Bank of Japan (DDBJ), which pool genomic data from a variety of sources. BLAST would be of limited utility without a broad-based database to query.

One core aspect of research in computational biology focuses on database development: how to integrate and optimally query data from (for example) genomic DNA sequence, spatial and temporal patterns of mRNA expression, protein structure, immunological reactivity, clinical outcomes, publication records, and other sources. A second focus involves pattern recognition algorithms for such areas as nucleic acid or protein sequence assembly, sequence alignment for similarity comparisons or phylogeny reconstruction, motif recognition in linear sequences or higher-order structure, and common patterns of gene expression. Both database integration and pattern recognition depend absolutely on accessing data from diverse sources, and being able to integrate, transform, and reproduce these data in new formats.

As noted above, computational biology is a fundamentally collaborative discipline, owing its very existence to the availability of rich and extensive data sets for analysis, integration, and manipulation. Data accessibility and usability are therefore critical, raising concerns about data release policies—what constitutes primary data, who owns this resource, when and how data should be released, and what restrictions may be placed on further use. Two challenges have emerged that could potentially restrict the advancement of bioinformatics research: (i) questions related to the appropriate use of data released before publication and (ii) restrictions on the reposting of published data.

The first challenge to bioinformatics research relates to the analysis of data posted on the Web in advance of publication. Recognizing the value of early data release for a wide range of studies, the Human Genome Project adopted a policy of prepublication data release (*2*), and many genome projects (and the funding agencies that support them) now adhere to similar rules. Because bioinformatics depends absolutely on the ability to integrate data from a wide variety of sources, it is to be hoped that other projects that generate genomic-scale data (including expression analysis and proteomics research) will follow a similar policy (*3*), because immensely valuable results can emerge from large-scale comparative studies of genome structure, microarray data, protein interactions, and so on (*4–6*). The success of such altruistic data

The author is at the Department of Biology and Genomics Institute, University of Pennsylvania, Philadelphia, PA 19104, USA. E-mail: droos@sas.upenn.edu