nature genetics
volume 33 no. 4 april 2003

Genetical implications

By happy accident (and not a little human effort), we mark this month the fiftieth anniversary of the publication of the proposed double-helical structure of DNA and the completion of the sequence of the human genome (more or less). That the two achievements are part of the same remarkable story is clear enough, although of course the paper of 25 April 1953 by Watson and Crick said nothing explicit about the impact the structure might have on biology. It was not until their paper of 30 May ("Genetical implications of the structure of deoxyribonucleic acid") that they had enough confidence to say in print some things they surely had been thinking all along, such as "…it therefore seems likely that the precise sequence of the bases is the code which carries the genetical information." Fifty years later, these and other implications of molecular biology's Big Bang have given us a new universe to think about, and beginning on page 449, we're pleased to present a set of three commentaries on the impact of the double helix.

How do you account for a universe? By counting, measuring, describing, isolating and annotating. The discovery of the double helix not only implied that whole genomes would ultimately be sequenced and analyzed, but also kick-started a new, molecular approach to cell biology, development, cancer, physiology and evolution—in short, a new way of approaching almost all of biology. The immense amount of information that has been generated is, in most ways, a blessing. At the same time, it is, if not quite a curse, then at least a problem to be faced. If genomes, cells and organisms have become open books, how is everything we've learned going to be managed?

The US National Academy of Sciences recently released a report entitled *Sharing Publication-Related Data and Materials: Responsibilities of Authorship in the Life Sciences*. The report's authors note that high-throughput methods often make it difficult to include all of the relevant data in a publication. This is nothing new, of course. It's easy to forget that, even in the low-throughput world of 1953, Watson and Crick concluded their landmark paper by saying "Full details of the structure…will be published elsewhere."

Such a statement probably wouldn't pass peer review these days, and in any case, supplementary information is routinely accessible on journal websites, making large volumes of data available to online readers. Much of the need to document and present every last detail of these large data sets is driven by the interconnectedness of modern molecular biology, where every sequence or expression pattern is potentially useful to another researcher in another context. The Gene Expression Omnibus and ArrayExpress microarray databases were established precisely because array data can

(and should) be mined repeatedly to generate new knowledge. The growth in the number of independent online databases covering a bewildering variety of genomic information is remarkable. The latest *Nucleic Acids Research* compilation includes descriptions of 130, and it is by no means the last word. Many of these community websites are superb resources (see *Nat. Genet.* **31**, 327–328 (2002)).

The need for complete information is no less true for clinical genetics. In a News and Views article on page 440 of this issue, Judith Hall explains the importance of providing clinical descriptions of individuals with a particular disease or syndrome who do not have a mutation in the gene under study. Although negative data don't always see the light of day, here again it can be extremely valuable to clinicians in generating hypotheses and in making appropriate decisions as to who should be offered a particular genetic test and who should not.

This data-rich era in biology thus places demands on authors to declare all of their results; on informaticians to develop central repositories that are easy to navigate; and on funding agencies to support the dissemination of large data sets, even when the payoff is not immediately obvious. Of course, there is no less of a challenge to editors and publishers to ensure that the published record is as complete as possible and to continually consolidate and summarize this unwieldy mass of data in useful and innovative ways. Three recent supplements to *Nature Genetics*—*A User's Guide to the Human Genome* (soon to be updated on the web), *The Chipping Forecast II* and *A Ten-Year Retrospective*—have been part of an effort to bring some semblance of order to certain areas of genetics that are of intense interest and have grown perhaps too rapidly for non-specialists to keep up with.

Our colleagues in the reference division of the Nature Publishing Group have also been laboring to produce comprehensive sources of information that cover essentially all of biology. *The Encyclopedia of Life Sciences* (ELS), a text and online reference containing thousands of entries on everything from *Acanthamoeba* to *Zygomycota*, is a unique compendium that has been produced with the collaboration of many of our own readers and referees. In June, *The Encyclopedia of the Human Genome* will be launched, and we urge the many users of ELS to explore this exciting new resource.

The generation of data shows no signs of slowing, and there are clear indications that some cellular processes are just beginning to be appreciated. The role of noncoding RNA—the genome's 'dark matter,' in Gary Ruvkun's words—implies a new mode of gene regulation. The possibility of a 'histone code' adds another—layers upon layers.

Is every last bit of these data essential for the field to advance? Perhaps not. Is much of it of use to someone? Almost certainly. Peter Medawar once said this about the role of new facts in science: "The factual burden of a science varies inversely with its degree of maturity... In all sciences we are being progressively relieved of the burden of singular instances, the tyranny of the particular. We need no longer record the fall of every apple." The distinguishing characteristic of the program of molecular biology and genetics as inaugurated by Watson and Crick 50 years ago is that, however impressive its explanatory power, maturity (at least in this sense) always seems to elude its grasp. It won't forever, of course, but for now we're still counting apples.