nature immunology

Sequencing is believing?

An unprecedented international effort over the past few years has yielded one of the most unusual and heralded resources ever put together for humankind: a map of the human genome. Like all orchestrations of such magnitude, a few lines in the score are unavoidably still works in progress. Syncopating the sometimes competing rhythms of multiple methods, centers, assays and individuals is an intricate task, so it is a testament to the leaders of the project that a working product is already in hand. But what should those of us outside of the sequencing symphonia do when we find inconsistencies, possible inaccuracies or confusion?

Even though most hard-core immunologists are not intimately involved with the genome projects, they have been eagerly tracking the progress of the genomic and bioinformatic communities. The human and mouse genome projects already comprise a most useful compendium of data that is accessible to anyone with the patience to master the interfaces and probe the amassed data. However, reports are surfacing that reveal an uncertainty about the veracity of some annotations. A student searching the web found that the order of mouse immunoglobulin heavy-chain constant regions online was different from that in the textbooks, which reflect the order originally deduced by Tasuku Honjo and reported in a classic *Nature* paper in 1981. Which to believe? His professor advised him to stick with the Honjo data. In other cases, investigators working on different proteins have noted divergences in the order of their favorite gene segments or family members from the gene order originally established from traditional cloning and walking methods. And then there is the occasional 'folk wisdom', heard only by those close enough to the geneticists of their field, that a particular part of the genome is annotated incorrectly or is missing a segment of sequence. The proteins that most interest immunologists are often members of huge gene families that result from multiple duplications. These are some of the most difficult regions to map accurately. On top of that, many of these regions are so polymorphic that determining whether the present genome model has errors or whether new polymorphisms are being uncovered becomes a major difficulty. Thus the genomic regions in which we as immunologists are most interested are those associated with some of the highest uncertainties. Reliance on the human genome map as the final arbiter of truth may be premature at this time.

The various groups that have shared the mapping of the human genome have divided the chore of annotating and curating by chromosome. A list of the individuals and centers that are taking responsibility for each can be found at www.genome.wustl.edu/projects/human/index.php?coordinators=1 and www.ncbi.nlm.nih.gov/genome/seq/HsCenters.html, respectively. Last fall the major centers agreed to pool the progress they are making on their individual chromosomes. Jim Ostell of the US National Center for Biotechnology Information (NCBI) points out to *Nature Immunology* that the NCBI produces regularly updated 'builds' (build 35 is expected shortly) that incorporate validated changes. The NCBI, the Sanger Centre, The University of Santa Cruz (UCSC) and many other centers are involved in verifying sequence and annotations and are moving collaboratively toward an error-free fully annotated genome. According to Jim Kent at UCSC, the next version of the HLA region will include multiple haplotypes (two haplotypes, A3-B7-DR15 and A1-B8-DR3, are already available on the Sanger Institute's Vertebrate Genome Annotation (VEGA) database at http://vega.sanger.ac.uk/), which should help iron out discrepancies in that region.

Assembly of the genome and its annotation are both complex projects, and many checks have been built into the system to minimize mistakes. However, such a vast array of multilevel data needs the help of those interested in particular regions. What should you do if you suspect that a region of the genome has errors that are unlikely to be due to polymorphisms? How do you know if the annotation or the older data is correct? The answers may be elusive. Andy Mungall of the Sanger Institute recommends that errors found with annotations in the chromosomes now in VEGA should be directed to the VEGA help desk (http://vega.sanger.ac.uk/helpdesk/index.html) and also to those individuals listed as being responsible for that chromosome. For other chromosomes, go directly to the individuals listed at www.genome.wustl.edu/projects/human/index.php?coordinators=1. Notifying the VEGA help desk also helps to keep track of error reports. If the problem persists, Adam Felsenfeld of the US National Human Genome Research Institute (http://www.nhgri.nih.gov/) informed *Nature Immunology* that he can be contacted to help resolve the matter.

Detecting discrepancies assumes a basic level of competence in navigating the genomic websites. For those immunologists just wandering into genomics, the free *User's Guide II* to basic bioinformatics tools available on the web, published by *Nature Genetics* in a September 2003 supplement, remains an excellent resource. Along these lines, *Nature Immunology* now hosts an online tutorial created by Anjana Rao and colleagues from Harvard University that leads the 'bench biologist' step by step through the many comparative genomics tools that can be used to find physiologically relevant regulatory regions, an area not covered in the *User's Guide*. Given the importance of control regions to proper differentiation and activation, many immunologists are investing much effort into the precise identification of noncoding regulatory regions of the genome. In their accompanying Commentary in this issue, the Rao group explains the tools used in the tutorial and how to validate the database findings with 'wet biology'. So in this 'bio-Information Age', immunologists can both help maintain an accurate annotation of the genome and begin intensive exploitation of the masterpiece unfolding before us.