nature genetics

A treasury of exceptions

The recent publication of the finished euchromatic sequence of the human genome (*Nature* **431**, 931–945; 2004) makes it clear that our fundamental understanding of the human genome has changed since 2001, the year the draft sequence was published. As the problematic parts of the genome yield to investigation, we should consider the implications these changes have for future progress to define the genetic contributions to human phenotypic variation and disease susceptibility.

After the International Human Genome Sequencing Consortium (IHGSC) published the draft sequence, they directed their efforts towards generating a complete sequence of the euchromatic portion, leaving only gaps that were refractory to available techniques. They now report that the most recent genome sequence (build 35) covers 99% of the euchromatic genome (10% was omitted in the draft sequence) and contains only 341 gaps (there were 150,000 gaps in the draft sequence). Many of the problematic regions resolved with large-insert cloning efforts were in areas of segmental duplication. Some of these regions are still problematic, as 273 of the 341 remaining gaps are in segmental duplications. A comparison between the IHGSC clone-ordered assembly and the Celera whole-genome shotgun assembly (*Nature* **431**, 927–930; 2004) demonstrates the necessity of the clone-based approach; more than half of the segmental duplications are completely lost in the Celera sequence. This loss is due to the inability to resolve large, highly identical duplications when assembly is based solely on sequence overlap. Because of this, the Celera genome length is reduced and genes in segmental duplications are conspicuously absent from its assembly. Thus, the IHGSC assembly is a considerable achievement, even with the remaining refractory gaps.

As it turns out, the duplicated regions provide the seed for a substantial amount of normal variation in the form of large-scale copy-number variation (*Nat. Genet.* **36**, 949–951; 2004 and *Science* **305**, 525–528; 2004). Because these regions have a high sequence identity, they provide a substrate for nonallelic homologous recombination, leading to chromosome rearrangements. It is estimated that there are perhaps 200 large segments that vary in copy number in the human species, although it is possible that the frequency is greater, as the screens carried out so far have not reached saturation. Although there are cases in which segmental duplications leading to copy-number variation are responsible for human genetic disease through deletion or rearrangement (for example, Charcot-Marie-Tooth syndrome on 17p and DiGeorge syndrome on 22q), the involvement of common copy-number variation in common disease is not yet known.

Much effort is being put forth to track down the genetic basis of common diseases using genome-wide single-nucleotide polymorphism (SNP) variation. But some worry that there is little hope of finding the kind of structural changes with functional consequences that emerged from studies of mendelian disease loci. Perhaps a lesson to be learned from the genome-sequencing effort is that the pieces of the puzzle that didn't fit may provide a windfall of structure-function insights for those researchers alert to the fact that there is much more to human variation than SNPs.

With increased coverage, the estimated number of protein-coding genes has dropped from 30,000 to 20,000–25,000. This catalog was created using automated and manual gene annotation for almost all of the chromosomes. The unexpectedly low gene number raises two fundamental questions: how is functional complexity generated from the genome, and what is the noncoding sequence (~98% of the genome) doing? Again, exceptions to the rule will probably provide insights to these issues.

Although a large portion of the noncoding genome is composed of repetitive elements, there is a substantial quantity of nonrepetitive noncoding sequence ('gene deserts') about which we know little. Comparative genomics has identified many conserved noncoding sequences that could be *cis*-regulatory elements or noncoding RNAs. Genetically active noncoding RNAs, such as antisense RNAs and microRNAs, are not counted in the estimated gene number but have been proposed to provide an important regulatory component to the genetic programming of higher organisms (*Nat. Genet.* **36**, 19–25; 2004 and *Nat. Rev. Genet.* **5**, 316–323; 2004). It is now estimated that there are as many noncoding genes as protein-coding genes (*Cell* **116**, 449–509; 2004). Although noncoding RNAs are important for developmental processes in lower organisms and are the subject of intense investigation in humans and other mammals, the extent to which these noncoding RNAs contribute to rare and common diseases is not yet fully appreciated.

But the functional importance of conserved noncoding sequences, in general, is not a given, as deletion of two different large gene deserts in the mouse resulted in no detectable phenotypic consequences and very little changes in gene expression (*Nature* **431**, 988–992; 2004). These deletions removed sequences that are conserved in human and mouse and some that are conserved in human, mouse, chicken and frog, indicating that conservation may not be an indicator of readily detectable function.

On a genomic scale, the National Center for Biotechnology Information's ENCODE project aims to identify all the functional elements in the human genome sequence, including protein-coding and noncoding genes, transcriptional regulatory elements and sequences important for chromosome structure. ENCODE is currently in a pilot phase, carrying out a deep annotation of 1% of the genome. Large-scale discovery efforts will surely lead to valuable insights, and the development of tools to integrate these data for display in order to make it accessible to researchers will be essential.

Many recent discoveries indicate that our understanding of the human genome is far from complete, even if the sequencing of it is, theoretically, a finished project. As is nearly always the case, exceptions to the rule continue to provide a rich source of insight. ■

Corrigendum: A transactivation-deficient mouse model provides insight into p53 regulation and function

G S Jimenez, M Nister, M Beeche, J S Stommel, E Barcarse, S O'Gorman & G M Wahl
Nat. Genet. **26, 37–43 (2000).**

This article described the generation of a mouse in which homologous recombination was used to change codons 25 and 26 of the gene *Trp53* to glutamine and serine. The initial sequencing of the full-length cDNA from mouse embryonic fibroblasts expressing this allele was considered to be wild-type at all codons except for 25 and 26. A reanalysis of the sequence showed that this mutant also contains a valine at codon 135, which is referred to as a "provisional wild type codon" in the National Center for Biotechnology Information's LocusLink tool (accession numbers: *Trp53* mRNA, NM_011640; p53 protein, NP_035770). Sequence analysis showed that the lambda genomic clone containing an *Eco*RI fragment encompassing *Trp53* (*Nature* **356**, 215–221; 1992) that was used to prepare the targeting construct contained the Val135 codon, whereas *Trp53* in laboratory mice (*e.g.*, C57Bl6) encodes alanine at codon 135 (K. Krummel, F. Toledo, C. Lee & G.M.W., unpublished data). The properties of the two proteins have been investigated by M.N. *et al.* (*Oncogene*, in the press).

Erratum: A treasury of exceptions

Editorial
Nat. Genet. **36, 1239 (2004).**

The ENCODE project (http://www.genome.gov/ENCODE/) is funded and managed by the National Human Genome Research Institute at the US National Institutes of Health.