

Mercredi 4 Février 2004

JEAN WEISSENBACH

Directeur

Genoscope - CNS

What is still insufficient in the human genome sequence ?

The announcement of an almost complete sequence of the human genome in April 2003 marked a milestone in the history of biology and medicine. Although such announcements need to be taken cautiously, the present version of the human genome sequence represents a major improvement for the daily users in human genetics, compared to the initial draft of June 2000. The present state of the sequence with regard to quality and completeness will be briefly reviewed: what the human genome has taught us in the recent past and what improvements could still provide a better utilization of what should be considered as the infrastructure on which molecular medicine will be established.

With about one error event for 100,000 base pairs, the accuracy substantially exceeds the threshold fixed at one error for 10,000 base pairs. About 400 gaps in the assembly corresponding essentially to clones missing from the existing libraries remain to be filled. They correspond mainly to regions difficult to clone or to map such as segmental duplications. Although an impressive number of SNP sequence variants have been identified, many more will be needed to dissect the complexity of multifactorial traits. It has been suggested that these haplotypes may be organized in blocks that are partly shared by the major human populations. A haplotype map consisting of 600,000 SNPs genotyped in three human populations (West African, Caucasian and East Asian) is in progress. It should be available by the end of 2004 and will give us a much deeper insight about haplotype structure and occurrence in human populations.

Annotation of the human genome has improved in parallel with the sequence. Surprisingly we observe a decrease in the present estimates of the number of protein-encoding genes, which will probably remain below 30,000. It is nevertheless

possible that a number of small open reading frames sometimes encoded by single exon genes have been largely overlooked in the automated annotation procedures used. The new additional vertebrate sequence resources that are progressively becoming available should be of great help in this issue. Comparative studies also highlight the conservation of non-coding sequence elements. The potential role(s) of such elements as well as of non-conserved non-coding transcripts remains a matter of speculation and discussion.