EGASP: collaboration through competition to find human genes

Roderic Guigó & Martin G Reese

According to the most recent estimates, the number of human genes is possibly—but not certainly—between 20,000 and 25,000. To contribute strategies to reduce this uncertainty, several groups working on computational gene prediction met recently at the Welcome Trust Sanger Institute with the goal to test and compare predictive methods of genome annotation.

From molecular biology textbooks we learn that the genetic code was deciphered in the early 1960s. Yet even four years after the first drafts of the human genome sequence became available[1,2], and more than two years after the announced completion of the sequencing[3], there is still great uncertainty in the set of protein sequences encoded in the human genome[4].

Finding human genes is a complex task because of the peculiar anatomy of the eukaryotic genome. Eukaryotic genes lie within long stretches of intergenic DNA, and within the genes only a few short fragments—the exons—are spliced together, often in alternative configurations, to form the mRNAs. Sequence signals in the genome are degenerate, and computational programs using them are able to identify the exons and link them into genes with relative success (see ref. 5 for a recent review). But only through the sequencing of the corresponding mRNA molecule can a gene be unequivocally identified. It is unclear, however, what fraction of genes can be ascertained through mRNA sequencing. In addition, genes are only one type of functional elements. It is likely that most of the functionality of the human genome sequence remains largely unexplored.

To address this limitation, the National Human Genome Research Institute (NHGRI) launched the ENCyclopedia Of DNA Elements (ENCODE) project two years ago[6]. The aim of ENCODE is to identify all functional elements in the genome sequence through the collaborative effort of computational and laboratory-based scientists. The pilot phase of the project is focused on a selected 30 megabases in 44 regions across the human genome (about 1% of the genome sequence). Within ENCODE, the GENCODE consortium was created with the goal to identify all protein

Table 1 | EGASP'05 participant groups and affiliations

AceScan	Salk Institute
Aceview	National Center for Biotechnology Information
ASPic	Università degli Studi di Milano
CSTminer	Università degli Studi di Milano
Augustus	Georg-August-Universität Göttingen
DOGFISH	The Wellcome Trust Sanger Institute
EnsEMBL	The Wellcome Trust Sanger Institute European Bioinformatics Institute
Exogean	Ecole Normale Superieure, Paris
ExonHunter	University of Waterloo
FGenesh++	Softwerry Inc.
Fprom	Softwerry Inc.
Softberry_pseudogenes	Softwerry Inc.
GeneID_U12	Institut Municipal d'Investigació Mèdica, Barcelona
SGP_U12	Institut Municipal d'Investigació Mèdica, Barcelona
GeneMark	Georgia Institute of Technology
GeneZilla	The Institute for Genomic Research
JigSaw	The Institute for Genomic Research
McPromoter	University of Virginia
Uncover	University of Virginia
N-Scan	Washington University
Paraigon	Washington University
SAGA	University of California at Berkeley
SPIDAI	European Bioinformatics Institute
Twinscan MARS	Washington University European Bioinformatics Institute

Roderic Guigó is at the Municipal Institute of Medical Research and Center for Genomic Regulation, University Pompeu Fabra, C/ Dr. Aiguader 80, 08003 Barcelona, Catalonia, Spain. Martin G. Reese is at Omicia Inc., 5980 Horton Street, Suite 235, Emeryville, California 94608, USA.
e-mail: rguigo@imim.es

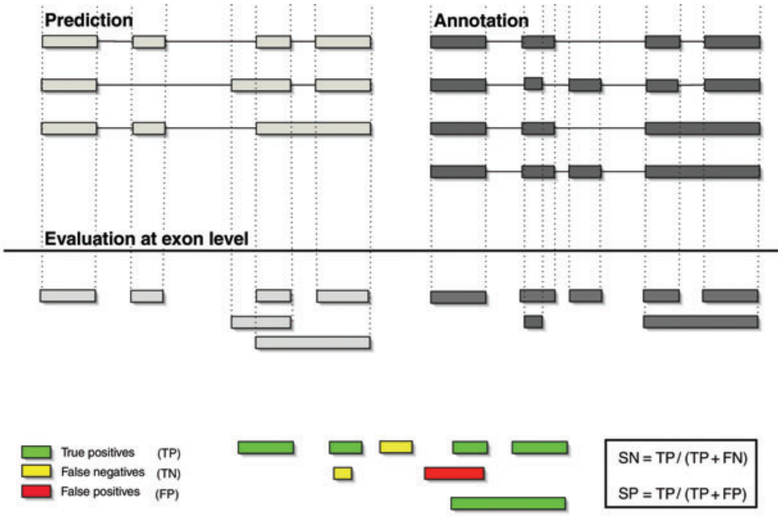

Figure 1 | Measuring accuracy of gene prediction tools against expert annotations. The figure schematizes how accuracy is computed at the exon level. The sets of unique exons from both the predicted and the annotated isoforms are identified—ignoring the exon links into isoforms. The sets are matched against each other, and the number of annotated exons correctly predicted (the true positives) over the total number of annotated exons (the true positives plus the false negatives, the annotated exons not in the prediction) is taken as a measure of sensitivity. Conversely, the number of exons correctly predicted over the total number of predicted exons (the true positives plus the false positives, the predictions which do not correspond to annotated exons) is taken as a measure of specificity. Other measures are used to assess how well the exons are linked into gene structures.

coding genes, and a first map of the gene content of ENCODE has just been released (http://genome.ucsc.edu/encode). To build this map—likely to be one of the most detailed ever built over large regions of the human genome—the ENCODE regions were subjected to a detailed manual curation by an annotation expert team at the Sanger Institute (the Havana team). The annotators weighted evidence based on alignments of known mRNA and protein sequences to the human genome to infer the location of potential genes (see ref. 7 for a review). The initial gene map delineated in this way was then experimentally refined by several techniques aimed at verifying the existence of the predicted genes.

One of the goals of the ENCODE project is to evaluate alternative high-throughput strategies to identify those that can be most efficiently scaled up to analyze the entire human genome. It is unclear how efficiently the approach used to produce the GENCODE map would be scaled up, as it requires substantial human intervention, which is costly both in terms of resources and time. In addition, the GENCODE annotation is still based on available evidence, and it cannot be completely ruled out that it misses an important fraction of protein coding genes.

To address these two issues, we organized EGASP'05, a community experiment (http://genome.imim.es/gencode/workshop2005.html). EGASP stands for ENCODE GASP, as it was inspired by the Genome Annotation Assessment Project (GASP1; http://www.fruitfly.org/GASP1), an experiment in 1999 to explore the accuracy of automated gene annotation prediction tools before the sequencing of the *Drosophila melanogaster* genome[8]. Specifically, the goals of EGASP'05 were first, to evaluate how well automatic methods are able to reproduce the GENCODE protein coding gene map and second, to assess the completeness and the correctness of the map. EGASP'05 was organized in two phases. When the complete map for 13 of the 44 ENCODE regions was released in January 2005, gene prediction groups worldwide were asked to submit predictions for the remaining 31 regions. Eighteen groups participated, submitting 30 prediction sets within four months (Table 1). When the annotation of the entire set of 44 ENCODE regions was released in May, participants, organizers and a committee of external assessors met at the Sanger Institute in a workshop sponsored by the NHGRI to compare the GENCODE gene map with the gene maps predicted by the participating groups.

The issue that first surfaced at the workshop was that of the most appropriate

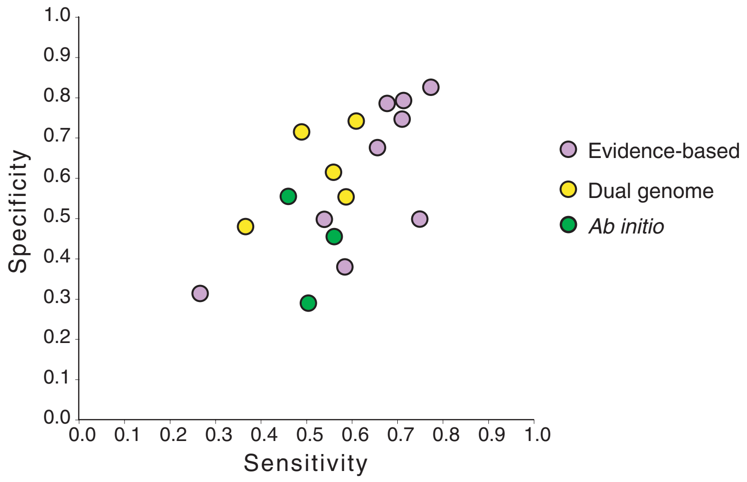

Figure 2 | Sensitivity and specificity at the exon level of several predictions submitted to EGASP'05. Sensitivity is the fraction of annotated exons that have been correctly predicted. For instance, a sensitivity value of 0.4 means that 40% of the annotated exons have been correctly predicted (including the exact exon boundaries). Specificty is the fraction of predicted exons that are correct (that is, that are annotated). For instance, a specificity value of 0.4 means that 40% of the predicted exons are annotated (including the exact exon boundaries). Each dot corresponds to a program, and the color indicates the approach-based category to which the program belongs.

metric to compare predictions with annotations and to measure in this way the accuracy of the predictions. This issue is far from trivial and emerges often as controversial when evaluating predictive methodologies. Within the field of computational gene prediction, the community has accepted several measures as the *de facto* standards (for instance, see refs. 9,10 for a discussion). Essentially, two numbers are computed (Fig. 1): the proportion of annotated features that have been predicted (the so-called sensitivity) and the proportion of predicted features that are annotated (the so-called specificity). To avoid trivial misevaluations—for instance, labeling as incorrect a prediction in which only one exon boundary has been mispredicted—the features to be evaluated are considered at three different levels: nucleotide, exon and gene[11]. At EGASP'05, however, we faced an additional difficulty. As the GENCODE annotation revealed, alternative splicing appears to occur in most human genes. Thus, to evaluate predictions, many predicted isoforms for a given gene need to be compared with many annotated isoforms. At EGASP'05, we also discussed how to address this difficulty by generalizing on existing measures (Fig. 1) and introducing new ones.

Figure 2 summarizes exon level accuracy for several of the submitted predictions These were categorized into three main classes based on the underlying methodological approach. Not surprisingly, programs using evidence from alignments to mRNA and protein sequences ('evidence-based') reproduced the GENCODE map the best, with values of sensitivity and specificity of about 0.8. These were followed by dual-genome or multiple-genome predictors, programs that rely on multispecies genome sequence comparisons to detect the genes. Whereas the best of these programs maintained specificity close to 0.8, the sensitivity dropped to about 0.6. Finally, the so-called '*ab initio*' methods, which rely only on sequence statistical patterns, achieved values of sensitivity and specificity of around 0.5. Whereas exon level accuracy is reasonably high, linking exons together into gene structures still appears to be a challenge for automatic methods, with the best of the programs being able to resolve about 40% of the complete gene structures inferred by the human annotators. To close this gap between manual and computational annotation, different groups are taking different approximations: improving the quality of the sequence alignments on which predictions are based, developing intelligent agents that simulate the behavior of human annotators and incorporating into the predictive frameworks realistic models of the biological processes (transcription, splicing and translation) that mediate the mapping from DNA to protein sequences, are some of the strategies being explored.

The second goal of EGASP'05 is to assess the completeness and correctness of the GENCODE annotations. Indeed, one of the deliverables we expected from EGASP'05 was a set of putative exons consistently predicted by many programs, but absent from the GENCODE annotations, that would be subjected to experimental verification. In particular, EGASP'05 predictions contained more than 7,000 unique exons in regions annotated as intergenic—and that, therefore, could correspond to novel, yet undetected genes. Experiments are under way to test a few hundred of them that have been predicted by more than one program. The correctness of the GENCODE map is similarly being assessed by testing the annotated gene models therein. The experimental results will be available in a few months, marking the completion of EGASP'05. After this exercise, we expect to have a better understanding of the relative merits of manual versus computational annotation of the human genome sequence, and of the level of experimental verification required to exhaustively locate human genes. This understanding will help in delineating strategies that can be efficiently scaled up to the entire human genome, so that the catalog of all human (protein-coding) genes can be completed within the next very few years.

Community experiments such as EGASP'05 are critical to get an accurate assessment of the state of the art within certain fields, and conceptually similar examples can be found in other areas of computational biology: CASP (Critical Assessment of Techniques for Protein Structure Prediction), BioCreAtIve (for evaluating text mining techniques), the GAWs (Genetic Analysis Workshops) and CAMDA (Critical Assessment of Microarray Data Analysis) are examples. They are excellent exercises to focus a whole community on a certain problem task and motivate groups and individuals to participate and submit their best possible solutions. External assessment of the results is critical and standards and rules have to be laid out clearly at the beginning of the experiment. Furthermore, such focused workshops that allow submitters to present their results on a clearly characterized test set make it very easy to compare methods and identify successful new developments in the field. Although it is hard work for the predictors to submit the predictions in the right format in time, for the annotators to produce accurate gene maps, for the evaluators to analyze the results in time for the assessor meeting, and for the assessors to evaluate the results to make sure they are accurate and consistent, there has been great enthusiasm at the workshop, and we believe it will have a great impact in tool development within computational gene prediction—an impact that will extend beyond the human genome to the many genomes to be sequenced in the coming years, and for which the amount of manual and experimental resources devoted to the human genome will certainly not be available.

ACKNOWLEDGMENTS

The GENCODE annotation is mostly the effort of A. Frankish, D. Swarbreck, J. Gilbert, J. Ashurst and T. Hubbard from the Havana team at the Wellcome Trust Sanger Institute; of C. Ucla, A. Reymond and S. Antonarakis from the Faculté de Médecine Université de Genève, and of F. Denoeud, J. Lagarde and R. Guigó from the Institut Municipal d'Investigació Mèdica (IMIM) in Barcelona. P. Flicek from the European Bioinformatics Insitute and J.F. Abril from the IMIM were responsible for computing most of the accuracy measures. J. Ashurst, E. Birney, P. Good, R. Guigó and T. Hubbard participated in the organizing committee. M. Ashburner, V.B. Bajic, T. Gingeras, S. Lewis and M. Reese were members of the advisory committee. The workshop was part of the ENCODE project that is funded by the US National Institutes of Health NHGRI.

1. Venter, J.C. *et al. Science* **291**, 1304–1351 (2001).
2. Lander, E.S. *et al. Nature* **409**, 860–921 (2001).
3. Rogers, J. *Cold Spring Harb. Symp. Quant. Biol.* **68**, 1–11 (2003).
4. International Human Genome Sequencing Consortium. *Nature* **431**, 931–945 (2004).
5. Brent, M.R. & Guigo, R. *Curr. Opin. Struct. Biol.* **14**, 264–272 (2004).
6. ENCODE Project Consortium. *Science* **306**, 636–340 (2004).
7. Ashurst, J.L. & Collins, J.E. *Annu. Rev. Genomics Hum. Genet.* **4**, 69–88 (2003).
8. Reese, M.G. *et al. Genome Res.* **10**, 483–501 (2000).
9. Bajic, V.B. *Brief Bioinform.* **1**, 214228 (2000).
10. Baldi, P., Brunak, S., Chauvin, Y., Andersen, C.A. & Nielsen, H. *Bioinformatics* **16**, 412–424 (2000).
11. Burset, M. & Guigo, R. *Genomics* **34**, 353–367 (1996).