Introduction

EGASP: Introduction

Martin G Reese* and Roderic Guigó†

Addresses: *Omicia Inc., Christie Ave, Emeryville, CA 94608, USA. †Centre de Regulació Genòmica, Institut Municipal d'Investigació Mèdica-Universitat Pompeu Fabra, B08003 Barcelona, Catalonia, Spain.

Published: 7 August 2006

Genome ***Biology*** 2006, **7(Suppl 1):**S1

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2006/7/S1/S1

Science is about building causal relations between natural phenomena (for instance, between a mutation in a gene and a disease). The development of instruments to increase our capacity to observe natural phenomena has, therefore, played a crucial role in the development of science - the microscope being the paradigmatic example in biology. With the human genome, the natural world takes an unprecedented turn: it is better described as a sequence of symbols. Besides high-throughput machines such as sequencers and DNA chip readers, the computer and the associated software becomes the instrument to observe it, and the discipline of bioinformatics flourishes. However, as the separation between us (the observers) and the phenomena observed increases (from organism to cell to genome, for instance), instruments may capture phenomena only indirectly, through the footprints they leave. Instruments therefore need to be calibrated: the distance between the reality and the observation (through the instrument) needs to be accounted for. This issue of *Genome Biology* is about calibrating instruments to observe gene sequences; more specifically, computer programs to identify human genes in the sequence of the human genome.

After nearly 25 years of research in the area of computational gene finding, and genome annotation, and after the completion of the human genome sequence in 2003, it became important to assess the current state-of-the-art in this discipline because in the future the success of many genomic and systems biology projects will depend on the quality of genome annotations. In this endeavor we built on the efforts by the NIH initiated ENCODE (for ENCyclopedia of DNA Elements) project, the goal of which, in its first phase, is the development and assessment of methods to identify all functional elements in 1% of the human genome across 44 regions, so that these methods can later be applied to the entire human genome. Within this project, the GENCODE consortium has produced a high quality annotation of the protein coding content of the ENCODE regions. We have used this annotation as the 'golden standard' against which to measure the performance of the computational methods. Developing such a standard has been a difficult task and the paper by Harrow *et al.* in this issue is dedicated to describing the process by means of which the GENCODE standard annotation was obtained.

Scientists working in the field of computational genome annotation were asked to submit predictions on the ENCODE regions. Eighteen groups worldwide participated in the experiment - which we named EGASP (for ENCODE Genome Annotation Assessment Project), the second of its kind after GASP1 [1] - and submitted 30 prediction sets using state-of-the-art methods.

Predictions were compared to the golden standard in a workshop organized at the Sanger Center on May 6 and 7, 2005, and sponsored by the National Human Genome Research Institute (NHGRI) at the NIH. The gene finding evaluation experiment is described in detail in Guigó *et al.* in this issue and the promoter evaluation experiment is described in Bajic *et al.* Many of the computational gene finding methods applied are also described in this issue: Allen *et al.*, Arumugam and Brent, Carter and Durbin, Djebali *et al.*, Flicek and Brent, Solovyev *et al.*, Stanke *et al.* and Thierry-Mieg and Thierry-Mieg. The paper by Zheng and Gerstein describes the analysis of the pseudogenes.

The willingness of the scientists within the gene finding community to provide their computational annotation for public comparison and blind evaluation against one and the same standard annotation set allowed us to really identify pluses and minuses in the various methodical approaches. Too often superior performance is claimed for new datasets without a careful analysis of potential biases within them. Therefore, we hope that with this experiment we have again laid out a test bed for performance enhancement within the field of gene finding. The difficulty of a repeated experiment in the future will be that it is hard to distinguish whether performance has improved due to the more and better auxiliary data (for example better cDNA sequences) or due to algorithmic improvements. Therefore, we suggest repeating the experiment in the future with the same genomic sequence and using the same, 'frozen' auxiliary sequence databases. Furthermore, we expect that a future experiment would include a higher focus on multiple mRNA transcript evaluation, including 5' and 3' untranslated region transcript predictions besides the classic coding sequence evaluations.

EGASP highlighted the recent progress in computational gene finding. Computer programs are increasingly sophisticated, efficient and accurate in mapping cDNA and protein sequences onto the genome sequence, as well as in using genome comparisons to other organisms. Despite the progress, however, programs are still not able to replace the insight of human annotators. Difficulties arise not only from the quality of the source data, but also because of the complexities of biology and the complex structure of human genes: the bulk of cDNA sequences are partial, and contain many sequence errors, and genome sequences are often incomplete and errors may exist in the assemblies; mapping of cDNA sequences onto the genome is compounded by the presence of pseudogenes and recent duplicates, which occasionally makes it very difficult to identify the exact genomic locus for a given cDNA sequence; and alternative splicing, for instance, is widespread, and involves, more often than until very recently expected, exons from apparently different loci. The diversity of the human proteome may be much higher than that derived simply from the total number of genes.

Computational methods at EGASP also predicted many exons and genes that were not included in the standard GENCODE annotation. While predictions mapped within annotated loci could correspond to novel alternative splice forms of known genes, predictions in intergenic regions might reveal novel genes. However, only a handful of such predictions could be verified by RT-PCR experiments using 24 human tissue libraries. This certainly seems to suggest that the standard GENCODE annotation is quite complete, and that, in general, not many novel gene loci remain to be discovered in the sequence of the human genome. This is, however, in contrast with results from recent large-scale surveys of the transcriptional activity of the human genome using high-throughput sequencing and hybridization based technologies [2].These reveal, also in the ENCODE regions, a wealth of sites of transcription that are neither included in the standard annotation or predicted by the programs. Whether these correspond to real, novel, protein coding genes or to non-coding RNAs could not be answered within the project described here.

In summary, from the EGASP project we have learnt that the current human genome annotation is almost complete in terms of novel protein coding loci. Nevertheless, the annotation of the exact structure and the transcriptional organization of a gene is still nowhere near completed. Three years after the completion of the human genome sequence and after all the human chromosomes have been published it seems that the gene locus annotations are still in flux. Therefore, we believe that efforts towards annotating the human genome should be extended. Almost correct gene annotations are simply not good enough as the blueprint of human biology. These errors can mislead many follow-on projects such as genetic variation experiments, mRNA expression profiling as well as proteomic experiments. Efforts to systematically and continuously sequence high-quality cDNA libraries to obtain full-length cDNA sequences, such as those at the Mammalian Gene Collection [3] need to be continued, although increasingly aggressive sequencing of cDNA libraries appears to have reached a plateau and is yielding only a fraction (which could be small) of lowly or rarely expressed transcripts. Hybridization based techniques, such as high density genome tiling micro-arrays, could constitute, in this regard, a complementary approach.

The ultimate goal in human genome annotation should be to map onto the genome sequence all primary and processed RNA molecules that exist in a given cell type at a given time - and ideally measure their relative abundance. This is a task that will likely take at least a decade to achieve completely. While the biological roles of non-coding RNAs are increasingly appreciated, EGASP focused on protein coding genes. In this regard, EGASP has shown that computational methods do not provide evidence for many additional, still un-annotated protein coding genes in the human genome, and, therefore, there is no need to drastically re-evaluate the current estimations of the total number of human genes. Current annotation efforts within the ENCODE project, in which RACE reactions are hybridized into genome tiling arrays, have, however, uncovered a wealth of additional transcripts mapping onto annotated protein coding loci. Often these transcripts reach upstream genes and include exons from intervening loci. These transcriptional continuums, in which boundaries between loci seem to fade away, challenge our very concept of what a gene is, and makes estimating the total number of human genes almost a futile task.

Acknowledgements

We thank the NHGRI for providing funding for the meeting, the Sanger Institute (Wellcome Trust) for help in its organization.

This article has been published as part of *Genome Biology* Volume 7, Supplement 1, 2006: EGASP '05. The full contents of the supplement are available online at http://genomebiology.com/supplements/7/S1.

References

1. Reese MG, Hartzell G, Harris NL, Ohler U, Abril JF, Lewis SE: **Genome annotation assessment in *Drosophila melanogaster*.** *Genome Res* 2000, **10:**483-501.
2. Cheng J, Kapranov P, Drenkow J, Dike S, Brubaker S, Patel S, Long J, Stern D, Tammana H, Helt G, *et al.*: **Transcriptional maps of 10 human chromosomes at 5-nucleotide resolution.** *Science* 2005, **308:**1149-1154.
3. Gerhard DS, Wagner L, Feingold EA, Shenmen CM, Grouse LH, Schuler G, Klein SL, Old S, Rasooly R, Good P, *et al.*: **The status, quality, and expansion of the NIH full-length cDNA project: the Mammalian Gene Collection (MGC).** *Genome Res* 2004, **14:**2121-2127.