Genome **Biology**

SELECTED ORAL PRESENTATION **Open Access**

From identification to validation to gene count

Clara Amid[1*], Adam Frankish[1], HAVANA[1], Bronwen Aken[1], Iakes Ezkurdia[2], Felix Kokocinsk[1], James Gilbert[1], Simon White[1], Piero Carninci[3], Thomas Gingeras[4], Roderic Guigo[5], Steve Searle[1], Michael L Tress[2], Jennifer Harrow[1], Tim Hubbard[1]

From Beyond the Genome: The true gene count, human evolution and disease genomics
Boston, MA, USA. 11-13 October 2010

Background

The current GENCODE gene count of ~ 30,000, including 21,727 protein-coding and 8,483 RNA genes, is significantly lower than the 100,000 genes anticipated by early estimates. Accurate annotation of protein-coding and non-coding genes and pseudogenes is essential in calculating the true gene count and gaining insight into human evolution.

As part of the GENCODE Consortium, the HAVANA team produces high quality manual gene annotation, which forms the basis for the reference gene set being used by the ENCODE project and provides a rich annotation of alternative splice variants and assignment of functional potential. However, the protein-coding potential of some splice variants is uncertain and valid splice variants can remain unannotated if they are absent from current cDNA libraries. Recent technological developments in sequencing and mass spectrometry have created a vast amount of new transcript and protein data that facilitate the identification and validation of new and existing transcripts, while harboring their own limitations and problems.

Results

Historically, all gene models have been built based on support from mRNA, EST and protein evidence. The recent integration of RNA-seq data into our annotation pipeline has allowed us to identify new splice variants that were previously either unannotated or supported only by non-human transcript evidence. Owing to their short read length, however, mapping them to the genome is problematic as is their use in recapitulating full-length transcript models. In order to assess different computational methods to map, assemble and quantify human RNA-seq data and improve this pipeline, we have been involved in the RNA-seq Genome Annotation Assessment project (RGASP), which seeks to address these questions.

We will also present the use of CAGE and ditag data produced by the ENCODE transcriptome group to identify and verify the use of alternative transcription start and termination sites and describe their impact on the interpretation of coding potential. Finally, we will show how mass spectrometry data can validate annotated gene models, identify novel splice variants and lead us to change our interpretation of the functional potential of a locus or variant.

Conclusions

We believe that an understanding of complete gene sets (i.e. the total gene number and the number of alternative splice variants allied to accurate functional interpretation) is crucial for understanding the genome. We demonstrate the value of the integration of new data types into our annotation pipeline in helping to identify and validate loci and variants to reach this aim.

Author details
[1]The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. [2]Structural Biology and Biocomputing Programme, Spanish National Cancer Research Centre (CNIO), Madrid, Spain. [3]Omics Science Center, RIKEN Yokohama Institute, Kanagawa, Japan. [4]Cold Spring Harbor Laboratory, 1 Bungtown Road, Cold Spring Harbor, New York 1 1 724, USA. [5]Center for Genomic Regulation, Universitat Pompeu Fabra, Barcelona, Catalonia, Spain.

Published: 11 October 2010

doi:10.1186/gb-2010-11-S1-O1
Cite this article as: Amid *et al.*: From identification to validation to gene count. *Genome Biology* 2010 **11**(Suppl 1):O1.

[1]The Wellcome Trust Sanger Institute, The Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK
Full list of author information is available at the end of the article