Genome Biology

SELECTED ORAL PRESENTATION **Open Access**

# Defining the human reference protein-coding gene set

Suganthi Balasubramanian[1], Lukas Habegger[1], Adam Frankish[2], Daniel MacArthur[2], Rachel Harte[3], Chris Tyler-Smith[2], Jennifer Harrow[2], Mark Gerstein[1*]

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

The number of coding genes in the human genome is still under debate [1]. Here, we present a proposal to define the human reference gene set that takes into account the inter-individual differences in gene numbers arising from gene inactivation events, such as premature termination or aberrant splicing due to nonsense SNPs or SNPs at essential splice sites respectively. We have analyzed SNPs (specifically nonsense SNPs and SNPs affecting essential splice sites) from 23 personal genomes and exomes. We see a wide range in numbers of SNPs in each of the categories surveyed. A large fraction of these SNPs are singletons. Using a data set of high-confidence SNPs obtained by intersecting SNPs from dbSNP and the personal genomes, we identify a common set of 279 genes predicted to be pseudogenic (non-functional) in some individuals and functional in others.

We focused on two key questions arising from these considerations: (i) Which criteria should be used for inclusion and exclusion of genes from the reference set? (ii) What sequence should be used as the reference for genes that are non-functional in some humans? For the first question, we propose to include all genes that are functional even in one individual to produce a maximally-inclusive set of genes. For the second, we propose the use of the ancestral allele as the reference allele. This will provide a uniform basis for gene annotation and ensure that the reference gene set and sequence will be relatively stable as more individual genomes are sequenced. In the few cases where an ancestral state assignment is unavailable or ambiguous, we propose that genes be annotated as the functional allele.

[1] Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA
Full list of author information is available at the end of the article

**Author details**
[1]Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, CT 06520, USA. [2]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton CB10 1SA, UK. [3]Department of Biomolecular Engineering, University of California, Santa Cruz, 1156 High Street, Santa Cruz, California 95064, USA.

Published: 11 October 2010

**Reference**

1. Pertea M, Salzberg SL: **Between a chicken and a grape: estimating the number of human genes.** *Genome Biol* 2010, **11**:206.

doi:10.1186/gb-2010-11-S1-O5
**Cite this article as:** Balasubramanian *et al.*: **Defining the human reference protein-coding gene set.** *Genome Biology* 2010 **11**(Suppl 1):O5.