Grossman et al. Genome Biology 2010, 11(Suppl 1):I22
http://genomebiology.com/2010/11/S1/I22

Genome Biology

INVITED SPEAKER PRESENTATION Open Access

Mining data from 1000 genomes to identify the causal variant in regions under positive selection

Shari Grossman[1,2,3*†], Ilya Shlyakhter[1,2†], Elinor K Karlsson[1,2], Shervin Tabrizi[1,2], Kristian Andersen[1,2], John Rinn[2], Eric Lander[2], Steve Schaffner[2], Pardis C. Sabeti[1,2], The 1000 Genomes Project

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

The human genome contains hundreds of regions in which the patterns of genetic variation indicate recent positive natural selection, yet for most of these the underlying gene and the advantageous mutation remain unknown. We recently reported the development of a method, Composite of Multiple Signals (CMS), that combines tests for multiple signals of natural selection and increases resolution by up to 100-fold.

Applying CMS to candidate selected regions from the International Haplotype Map, we localized several hundred signals to ~50-100 kb, identifying individual gene and polymorphism targets of selection. These regions included genes involved in processes known to be targets of selection, such as infectious disease, skin pigment, metabolism, and hair and sweat. We further identified many candidates that are similar to regulatory elements. In several regions, we identified variants that are significantly associated with the expression of nearby genes in the selected population. Moreover nearly half of the ~200 regions we examined localized to regions with no genes. Thirty of the regions contain long non-coding RNAs that have been shown to often regulate nearby genes, suggesting that variation within the RNAs might have functional consequences.

With preliminary data now available from the 1000 Genomes Project, we are beginning to explore full sequence data, which should contains most if not all of the causal selected polymorphisms. We extended the CMS method to the preliminary data set, validating our previously identified candidates and identifying many new intriguing coding and regulatory variants.

† Contributed equally
[1]Center for Systems Biology and Department of Organismic and Evolutionary Biology, Cambridge, MA 02138, USA
Full list of author information is available at the end of the article

Author details
[1]Center for Systems Biology and Department of Organismic and Evolutionary Biology, Cambridge, MA 02138, USA. [2]Broad Institute of Harvard and MIT, Cambridge, MA 02139 USA. [3]Harvard Medical School, Boston, MA 0211, USA.

Published: 11 October 2010

doi:10.1186/gb-2010-11-S1-I22
Cite this article as: Grossman *et al.*: Mining data from 1000 genomes to identify the causal variant in regions under positive selection. *Genome Biology* 2010 **11**(Suppl 1):I22.