Yale University

Bass Building, Rm 432A
260/266 Whitney Avenue
PO Box 208114
New Haven, CT 06520-8114

Telephone:
203 432 6105
360 838 7861 (fax)
mark.gerstein@yale.edu
http://bioinfo.mbb.yale.edu

Genome Biology
Biomed Central
236 Gray's Inn Road
London WC1X 8HB
United Kingdom

Dear Editor,

We would like to submit our manuscript entitled "**Allele-specific binding and expression: a uniform survey over many individuals and assays**" for publication in *Genome Biology*.

The recent 1000 Genomes Project and exome sequencing projects have uncovered a preponderance of rare variants within the human population. The accumulating number and diversity of personal genomes being sequenced will continue to contribute to a growing catalog of variation, with most of their functional effects unknown. One way to functionally annotate these variants is to overlap the personal genomes with signals from their corresponding 'personal' functional assays, such as ChIP-seq and RNA-seq datasets.

In our study, we focus on interpreting single nucleotide variants (SNVs), including the rare ones, and also genomic regions associated with allele-specific binding (ASB) and expression (ASE). By integrating existing DNA sequences, ChIP-seq and RNA-seq datasets, we assess ASB and ASE SNVs based on allelic imbalance observed in the readouts of the functional assays and then compute the enrichment of allele-specific SNVs in a population-aware fashion for each genomic annotation or region.

Our study introduces a general pipeline to use existing datasets from disparate studies in the allele-specific annotation of personal genomes. These datasets are typically heterogeneous, and allele-specific behavior detection is extremely sensitive to the heterogeneity stemming from overdispersion and technical issues, arising from using inconsistent tools and parameters in processing the datasets. For example, many studies align ChIP-seq and RNA-seq reads to the human reference genome; this introduces reference bias. Also, studies might not remove SNVs found in copy number variants but copy number changes can easily masquerade as allelic imbalance, thus causing a higher rate of false positives in allele-specific SNVs detection within copy number variants. Hence, to alleviate these issues, we uniformly process the datasets and

account for overdispersion. In all, we constructed 382 personal genomes and reprocessed 287 ChIP-seq and 993 RNA-seq datasets from various studies, notably from the ENCODE and gEUVADIS projects. The endeavor took about 600 days in CPU time (1.6 years), but the pipeline is highly parallelizable, thereby streamlining the process. We consolidate the results in a database, AlleleDB. Subsequently, we are able to investigate the heritability and selection pressure of allele-specific behavior. We also provide a large-scale comprehensive survey of allele-specific behavior in the human genome, delving into 708 non-coding genomic annotations, 19,257 autosomal protein-coding genes, and several categories of genes, gene elements and enhancer regions. The survey allows us to identify genomic annotations and regions that might be sensitive to allelic changes.

As more diverse personal genomes, tissue types and cell lines, with corresponding functional assays become available, we expect the resource and framework to be of high value to researchers involved not only in allele-specific regulation or gene expression, but to the scientific community at large. Thus, we believe our work will be of considerable interest to your readership.

Yours sincerely,

Mark Gerstein

Albert L. Williams Professor of Biomedical Informatics,
Molecular Biophysics & Biochemistry,
and Computer Science,
Co-director of the Yale Program in Computational Biology
and Bioinformatics

We list a number of suitable reviewers for the paper:

Professor Aleksandar Milosavljevic
Baylor College of Medicine, Texas, USA
amilosav@bcm.edu

Professor Tom Gingeras
Cold Spring Harbor Laboratory, New York, USA
gingeras@cshl.edu

Professor Roderic Guigo
Centre for Genomic Regulation, Barcelona, Spain
roderic.guigo@crg.cat

Professor Zhiping Weng
University of Massachusetts Medical School, Massachusetts, USA
zhiping.weng@umassmed.edu

Dr. Paul Bertone
EMBL-EBI, Cambridge, United Kingdoms
bertone@ebi.ac.uk

Due to conflict of interests, we would like to request that our manuscript not be reviewed by:

Professor Tuuli Lappalainen
New York Genome Center, New York, USA
tlappalainen@nygenome.org

Professor Emmanouil Dermitzakis
University of Geneva, Geneva, Switzerland
emmanouil.dermitzakis@unige.ch

Professor Jonathan Pritchard
Stanford University, California, USA
pritch@stanford.edu

Professor Lior Pachter
University of California at Berkeley, California, USA
lpachter@math.berkeley.edu