Bioinformatics, 31(11), 2015, 1848–1850
doi: 10.1093/bioinformatics/btv050
Advance Access Publication Date: 30 January 2015
Applications Note

OXFORD

Gene expression

GenoExp: a web tool for predicting gene expression levels from single nucleotide polymorphisms

Ohad Manor[1,†] and Eran Segal[1,2,*]

[1]Department of Computer Science and Applied Mathematics and [2]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot 76100, Israel

*To whom correspondence should be addressed.
[†]Present address: Department of Genome Sciences, University of Washington, Seattle WA 98102, USA
Associate Editor: Ziv Bar-Joseph

Received on May 14, 2014; revised on January 7, 2015; accepted on January 26, 2015

Abstract

Summary: Understanding the effect of single nucleotide polymorphisms (SNPs) on the expression level of genes is an important goal. We recently published a study in which we devised a multi-SNP predictive model for gene expression in Lymphoblastoid cell lines (LCL), and showed that it can robustly predict the expression of a small number of genes in test individuals. Here, we validate the generality of our models by predicting expression profiles for genes in LCL in an independent study, and extend the pool of predictable genes for which we are able to explain more than 25% of their expression variability to 232 genes across 14 different cell types. As the number of people who obtained their SNP profiles through companies such as 23andMe is rising rapidly, we developed *GenoExp*, a web-based tool in which users can upload their individual SNP data and obtain predicted expression levels for the set of predictable genes across the 14 different cell types. Our tool thus allows users with biological knowledge to study the possible effects that their set of SNPs might have on these genes and predict their cell-specific expression levels relative to the population average.

Availability and implementation: *GenoExp* is freely available at http://genie.weizmann.ac.il/pubs/GenoExp/.

Contact: eran.segal@weizmann.ac.il

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In recent years, individual genotype [i.e. single nucleotide polymorphisms (SNPs)] data have become readily accessible via direct-to-consumer genetic testing (DTCGT) companies such as 23andMe, deCODEme and Navigenics. These companies allow users to acquire their allele calls at ~0.5 M SNPs across the genome, as well as a report of higher and lower risk for certain diseases, based on their genotyped SNPs. Since we have recently developed methods to robustly predict gene expression solely from genotype (Manor and Segal, 2013), we decided to enable users to predict their own gene expression across multiple cell types. To this end, we developed *GenoExp*, a web-based tool where users can upload their raw genotype files obtained from any DTCGT company, and view their predicted gene expression (for a set of predictable genes) across 14 different cell types.

To demonstrate the application of *GenoExp*, we used 23andMe genotype data provided willingly by one individual. We show the resulting predictions of gene expression and highlight examples such as a gene known to play a role in learning and memory that is predicted to be expressed above population-mean in brain cells of that individual.

2 Results

Our recently published method (Manor and Segal, 2013) was developed using genotype and gene expression data taken from Lymphoblastoid cell lines (LCLs) of 715 individuals. Therefore, we wished to test whether we could apply it to other tissues or cell types as well. For this purpose, we obtained two additional independent datasets that contained both genotype and gene expression measurements for different individuals: Genotype-Tissue Expression (*GTEx*; Lonsdale *et al.*, 2013), containing data from 91–166 individuals across 9 different cell types, and Abundant Quantitative Trait Loci Exist for DNA Methylation and Gene Expression in Human Brain (*Brain-eQTL*; Gibbs *et al.*, 2010), containing data from 146 individuals across 4 different brain cell types. Our full dataset thus included gene expression data from a total of 18 124 unique genes across 1039 different individuals in 14 cell types (Supplementary Table S1). Next, we ran our predictive pipeline separately for the data in each of the 14 cell types (Supplementary Figs. S1–S4). Specifically, for each cell type we used a strict 5-fold cross-validation scheme to learn a predictive model for the expression of each gene from the genotype of its *cis*-SNPs across individuals, and evaluated the performance of our models on held-out test individuals from the same gene and cell type. We found that across all cell types, a total of 232 unique genes exceeded an R^2 of 0.25 on test data (i.e. more than 25% of expression variability is explained on held-out individuals) in at least one of the cell types. These genes were therefore defined as predictable genes (Supplementary Table S2). Next, to test the generality of our models, we further obtained additional genotype data coupled with gene expression measurements of 381 twin-pairs (i.e. a total of 762 individuals) from the Multiple Tissue Human Expression Resource (MuTHER; Grundberg *et al.*, 2012). When applying the models learned from LCLs to the MuTHER data, we found that in 79% of cases, we predicted the correct direction of over- or under-expression (see Supplementary text for extended details), strengthening the validity of our predictions.

Finally, we wanted to allow users to predict their own gene expression across multiple cell types using our predictive models. To this end, we developed *GenoExp*, a web-based tool where users can upload their raw genotype files obtained from any DTCGT company (e.g. 23andMe), and view their predicted gene expression (for the predictable genes) across the different 14 cell types (Supplementary Fig. S5). To demonstrate the application of *GenoExp*, we used 23andME genotype data willingly provided by one individual (Supplementary Fig. S6). We found that for this individual, some genes are predicted as expressed above population-mean in all cell types where prediction is possible (e.g. HLA-DRB5, SNPs used for prediction in skin shown in Supplementary Fig. S7), whereas other genes are predicted as expressed below population-mean (e.g. KCTD10). Clearly, this over- and under-expression does not necessarily imply functional consequences, yet they could be suggestive. For example, the gene CHURC1, which was predicted to be expressed above population-mean in all 4 cell types in the brain, has a homolog that was shown to play a key role in the development of neurons (Sheng *et al.*, 2003), and a deletion of the genomic region containing it has been linked to autism (Griswold *et al.*, 2011). In addition, the AMFR gene that was shown to be involved in the process of learning and memory in the central nervous system of mice (Yang *et al.*, 2012) was also predicted to be expressed above population-mean in the brain cells of the tested individual.

3 Discussion

One of the major goals of human genomics and medicine today is to reach the stage of personalized medicine. That is, tailoring the diagnosis, prognosis and treatment of diseases to a specific individual. The connections between gene expression and various human diseases have long been made, and over- and under-expression of specific genes have been shown to promote or protect from certain diseases. In addition, studies have shown that many SNPs are significantly associated with the expression of various genes, although no predictive models are usually offered in these studies, making it impossible for an individual to predict their own gene expression from SNPs. This drawback has led us to recently publish a study where we learn a predictive model of gene expression from multiple SNPs. This modeling framework allows both the combination of different SNPs in the model, and the ability to predict the expression of a gene in unseen individuals given their entire SNP profile. Here, we extend our predictive models to 232 genes where we can explain 25% or more of the expression variability in held-out test individuals in at least one of 14 cell types.

As acquiring one's genetic data (i.e. SNP values) to learn more about his or her own disease risks and biology is becoming easier and more prevalent through companies such as 23andMe, we set out to build a web-based tool, that would allow users to input their measured SNP values, and obtain predictions of gene expression across the various cell types in our study. We acknowledge that currently, only individuals that have biological knowledge could potentially benefit from our tool. However, we believe that as more data of genotype and phenotype becomes available, models such as ours can become more sophisticated, allowing more accurate predictions of gene expression profiles and ultimately of organismal phenotypes such as disease risks or possible adverse effects of drugs, thereby enabling users to take more knowledge-based actions when it comes to their well-being.

4 Methods

4.1 Data and pre-processing

For the HapMap LCL dataset, we downloaded the SNP genotypes of phase 3 (Altshuler *et al.*, 2010) and obtained gene expression measurements for genes of these individuals from (Stranger *et al.*, 2012). For the GTEx and Brain-eQTL datasets, we downloaded the SNP genotype and gene expression measurements from (Gibbs *et al.*, 2010; Lonsdale *et al.*, 2013) after obtaining approval. For the MuTHER dataset, we downloaded SNP genotype from (Grundberg *et al.*, 2012) and corresponding gene expression measurements from the ArrayExpress website (http://www.ebi.ac.uk/arrayexpress/). Except for normalization of expression data discussed below, we did not process the data further than the processing done by the original publications. For each gene, we extracted all *cis*-SNPs located inside the gene or within 100 kb from the transcription start or end sites (gene locations were downloaded from the UCSC genome browser and SNP locations from dbSNP). We transformed each SNP to a discrete variable with values of 0, 1 or 2, corresponding to the number of minor alleles that each individual carries for the SNP. Therefore, an individual with 0 minor alleles will have a value of 0, etc.

4.2 Data normalization

Since the absolute values of gene expression in each dataset could be different given the measurement platforms, to remove dataset specific effects, we normalized the expression of every gene across all individuals in each dataset to have a mean equal to 0 and standard deviation equal to 1.

4.3 Learning a predictive model of gene expression

For each gene, we learned a separate predictive model for each of the 14 cell types in which the gene was expressed. We used the set

of *cis*-SNPs as covariates in the prediction problem, and learned the predictive model as described in Manor and Segal (2013). However, due to the fact that the K-Nearest-Neighbor model is both memory and computationally intensive, in the web-tool we restricted ourselves to the regularized linear version of the model (Supplementary Fig. S1).

4.4 Selecting the set of predictable genes

To select the set of predictable genes, we used the coefficient of determination (i.e. R^2) as out threshold. R^2 is a measure of the proportion of variance explained that is widely used when assessing the accuracy of models aiming to predict a continuous variable such as gene expression. The threshold of $R^2 \geq 0.25$ is chosen since explaining over 25% of individual gene expression variability on held-out test data is considered by us to be a useful prediction.

4.5 Building a web-based tool to enable users to predict their expression

For the web tool, we use a combination of Perl and Matlab scripts to do the following: (i) Upload the user's SNP genotypes file (i.e. from 23andMe) from the web form to the server. (ii) Impute the missing SNPs needed by the gene specific models using the IMPUTE2 (Howie *et al.*, 2011, 2009) software, and using the 1000 Genomes Project together with HapMap data as the genome reference for imputation. (iii) Predict the expression of available genes given the user's SNPs (iv) Create a table that summarizes the results and links out to various resources regarding the predicted genes (Supplementary Figs. S5–S7).

Funding

This work was supported by grants from the European Research Council (ERC) grant 614504, the U.S. National Institutes of Health (NIH) grant R01 CA119176-06, and the EU SYSCOL project grant 258236 to E.S. E.S. is the incumbent of the Soretta and Henry Shapiro career development chair. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Conflict of Interest: none declared.

References

Altshuler,D.M. *et al.* (2010) Integrating common and rare genetic variation in diverse human populations. *Nature*, **467**, 52–58.

Gibbs,J.R. *et al.* (2010) Abundant quantitative trait loci exist for DNA methylation and gene expression in human brain. *PLoS Genet.*, **6**, e1000952.

Griswold,A.J. *et al.* (2011) A de novo 1.5 Mb microdeletion on chromosome 14q23.2-23.3 in a patient with autism and spherocytosis. *Autism Res.*, **4**, 221–227.

Grundberg,E. *et al.* (2012) Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nat. Genet.*, **44**, 1084–1089.

Howie,B. *et al.* (2011) Genotype imputation with thousands of genomes. *G3*, **1**, 457–470.

Howie,B.N. *et al.* (2009) A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. *PLoS Genet.*, **5**, e1000529.

Lonsdale,J. *et al.* (2013) The Genotype-Tissue Expression (GTEx) project. *Nat. Genet.*, **45**, 580–585.

Manor,O. and Segal,E. (2013) Robust prediction of expression differences among human individuals using only genotype information. *PLoS Genet.*, **9**, e1003396.

Sheng,G. *et al.* (2003) Churchill, a zinc finger transcriptional activator, regulates the transition between gastrulation and neurulation. *Cell*, **115**, 603–613.

Stranger,B.E. *et al.* (2012) Patterns of cis regulatory variation in diverse human populations. *PLoS Genet.*, **8**, e1002639.

Yang,Y. *et al.* (2012) Autocrine motility factor receptor is involved in the process of learning and memory in the central nervous system. *Behav. Brain Res.*, **229**, 412–418.