From genotype and gene expression to phenotype: A brief review.

Wen Zhang*

Department of Psychiatry, Icahn School of Medicine at Mount Sinai, New York, USA

Accepted on September 10, 2017

Editorial

With developments of human genetics and computational biology as well as bioinformatics technology during the past two decades, people know more and more about genotypes, transcriptomes and phenotypes as well as nexus between them. The relation between them has been discovered a lot due to the fact that more cohorts of data are available nowadays for investigators to survey. For instance, Genotype-Tissue Expression (GTEx) data sets [1] cover genotypes and expressions of around 500 hundred samples across more than 40 tissues. Genetic European Variation in Health and Disease (GEUVADIS) data [2] includes RNA-sequencing as well as genotype data of 1000 Genomes samples. Depression Genes and Networks (DGN) [3] contains data for 922 whole-blood samples. Common Mind Consortium (CMC) generates genotypic and transcriptomic data of ~500 samples from several institutes including Mount Sinai, Pittsburg University and University of Pennsylvania [4]. To date, CMC is the largest cohort of data about human mind so far. Meanwhile, Stockholm-Tartu Atherosclerosis Reverse Network Engineering Task (STARNET) [5] is a human multi-tissue DNA and RNA dataset and biobank from 600 coronary artery disease patients. Certainly, there are many other public databases and data sets for special academic purposes such as The Cancer Genome Atlas (TCGA) [6], which stores expression data for variety of cancers. With the help of abovementioned data sets, people's knowledge of human genotypes, phenotypes as well as transcriptomes has been increasingly developed, which leads to a new era of evolution in this field. In the meantime, more and more still remain unknown in the area from perspective of either biotechnology or molecular genetics concerning the mechanisms behind phenotypes and genotypes as well as the relation between them. The roles of other so-called 'omics' during the process of different traits are not very clear also. Recently, the functions of epigenomics through the gene expression effects on phenotypes gain more and more attentions. This short editorial briefly reviews methods concerning connections between genotype, gene expression and phenotype.

As it is mentioned in Network-based approaches to study complex human diseases [7], gene expression research such as microarray and RNA-seq has been largely focused on the discovery of mechanisms for complex human diseases including tumor, heart defects, obesity, diabetes, schizophrenia, etc. In molecular pathway identification using biological network-regularized logistic models [8], a network-based logistic regression method was proposed to identify molecular pathways of breast cancer through the transcriptome and phenotype data. It is a representative network based approach to study human disease and a series of network studies are presented and reviewed in Zhang et al. study [7]. Phen-Gen [9] is another method to study rare disorders combing phenotype and genotypes. A pipeline of it is also available. In addition, the disease study is not only limited to complex diseases, as a good amount of methods generate pipelines for ranking or prioritizing candidate genes or SNPs in exomes or NGS panels with comprehensive coverage of human Mendelian disease information [10,11]. Both PhenIX and PredictSNP provide vcf format as input genotypes.

Genome-wide association studies (GWASs) have been widely used to successfully identify tons of genetic variants associated with complex traits and diseases [12]. When we study genotypes that affect phenotypes, GWAS as well as the resulting outcomes sometimes acts as an effective 'media' to link them [13]. Also, GWAS results provide reference statistics for association studies in the gene level. The so called *GWAS2* genes [13] provide a database of how common genetic variants affect gene expression. *GWAS2* gene also gives 57 GWAS results on different traits. MetaXcan [14], which is a well-developed pipeline, infers the association results of mapping traits via summary statistics from large-scale GWAS or meta-analyses. Actually, it is an extension of PrediXcan [15], which can impute transcriptomes through genotype and phenotypes. The main idea of PrediXcan is to use a cohort of genotype and gene expressions data as well as a reference genome to train and obtain a database of predictors, which is called PredictDB [15]. Then the predictors could be employed to predict or impute unknown gene expressions only using the corresponding cohort of genotype data. This method is promising since we could manage to calculate expression values *in silico* employing merely genotypes. For a large cohort of data, due to the high accuracy of the method [15], this approach will save a huge cost.

The methods to dissect genetics of complex traits using summary association statistics are reviewed in Pasaniuc et al. [16] Now-a-days; meta-analysis develops significantly fast to discover mechanisms between transcriptome, genotype and phenotype. Coloc makes it possible to perform systematic meta-analysis type comparisons across multiple GWAS datasets [17]. As an extension of coloc, moloc can analyze jointly more than two traits [18]. The online database of moloc is provided (http://icahn.mssm.edu/moloc) for investigators to research summary statistics from GWAS, eQTLs and mQTLs. The importance of this rigorous approach was demonstrated by applying the method to the largest GWAS in schizophrenia, linking risk loci with QTLs affecting gene expression and DNA methylation in human brain tissue [18].

Admittedly, if we talk about methods relating genotype, gene expression and phenotype, at least several classes of methodology could be grouped. For example, there are methods to study genotype affecting phenotype, and methods to research gene expression influencing phenotypes and so on. However, integrative methods combing several approaches are more encouraging due to the fact that studies in each field always across each other. A more detailed classification of different kinds of methods concerning relations and mechanisms between genotypes, transcriptomes and even epigenomes is out of the scope of this mini-review. The aim of this article is to encourage more opinions, ideas and discussions about the approaches as well as developments that stimulate the study to a terrace of even more extensive level.

References

1. GTEx Consortium. The genotype-tissue expression (GTEx) project. Nat Genet. 2013; 45: 580-095.
2. Lappalainen T, Sammeth M, Friedlander MR, et al. Transcriptome and genome sequencing uncovers functional variation in humans. Nature. 2013;501-11.
3. Battle A, Mostafavi S, Zhu X, et al. Characterizing the genetic basis of transcriptomes diversity through RNA-sequencing of 922 individuals. Genome Res. 2014;24: 14-24.
4. Fromer M, Roussos P, Sieberts SK, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat Neurosci. 2016;19:1442-53.
5. Franzén O, Ermel R, Cohain A, et al. Cardiometabolic risk loci share downstream cis and trans-gene regulation across tissues and diseases. Science. 2016;353:827-30.
6. Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer Analysis Project. L Nat Genet. 2013; 45:1113–20.
7. Zhang Q, Zhang W, Nogales-Cadenas R, et al. From gene expression to disease phenotypes: Network-based approaches to study complex human diseases. Transcr Gene Regul. Springer; 2016: 115-140.
8. Zhang W, Wan YW, Allen GI, et al. Molecular pathway identification using biological network-regularized logistic models. BMC Genomics. 2013;14 Suppl 8:S7.
9. Javed A, Agrawal S, Ng PC, et al. Combining phenotype and genotype to analyze rare disorders. Nat Method. 2014;11:935-7.
10. Smedley D, Jacobsen JOB, Jager M, et al. Next-generation diagnostics and disease-gene discovery with the Exomiser. Nat Protoc. 2015;10: 2004–15.
11. Bendl J, Stourac J, Salanda O, et al. Predict SNP: Robust and accurate consensus classifier for prediction of disease-related mutations. PLoS Comput Biol. 2014;10: e1003440.
12. Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48: 245–52.
13. Hauberg ME, Zhang W, Giambartolomei C, et al. Large-scale identification of common trait and disease variants affecting gene expression. Am J Hum Genet. 2017;100:885–94.
14. Barbeira A, Shah KP, Torres JM, et al. MetaXcan: Summary statistics based gene-level association method infers accurate prediXcan results. BioRxiv. 2016;45260.
15. Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47: 1091-98.
16. Pasaniuc B, Price AL. Dissecting the genetics of complex traits using summary association statistics. Nat Rev. Genet. 2017;18:117–27.
17. Giambartolomei C, Vukcevic D, Schadt EE, et al. Bayesian test for co-localization between pairs of genetic association studies using summary statistics. PLOS Genet. 2014;10: e1004383.
18. Giambartolomei C, Liu JZ, Zhang W, et al. A Bayesian framework for multiple trait co-localization from summary association statistics. BioRxiv. 2017: 155481.

***Correspondence to:**

Wen Zhang
Department of Psychiatry
Icahn School of Medicine at Mount Sinai
New York, NY 10029,
USA
Tel: +1 212-241-6696
E-mail: zhang.wen81@gmail.com