Available online at http://jurnal.ahmar.id/index.php/asci
Journal of Applied Science, Engineering, Technology, and Education
ISSN 2685-0591 (Online)

Journal of Applied Science, Engineering, Technology, and Education Vol. 1 No. 1 (2019)
https://doi.org/10.35877/454RI.asci1116

Computational drug repositioning using big data from genetic studies

Wen Zhang*

Icahn School of Medicine at Mount Sinai, 1470 Madison Ave., New York, NY 10029, USA

Abstract

This mini-review gives the development of computational drug repositioning using big data from perspective of genetic study. The reverse profile principle is utilized to reposition drug hits by investigating gene expression, genotyping and GWAS data. Several big data sets are introduced, which are remarkable references that utilized for the genetic studies. Relevant computational genetics methods and the developments are briefly described as well. This review aims to give insights into this area so as to raise more ideas as well as to attract more attentions for this ascendant field. With the development of information technology and engineering applications, prosperous results are highly expected.

Keywords: Computational drug repositioning, big data, GWAS, genetic study.

1. Introduction

In the past two decades, computational genetics and bioinformatics technology developed super fast. Investigators could study gene expression, genotypes, and the correlations between genotyping and transcriptome through big data so as to reveal potential pharmacological targets. To date, there are many public data sets that researches could apply, for instance, CMap [1] is the largest genetic medicine database that researchers could refer to. CMap was initiated and organized by Broad Institute, which utilized L1000 technology to generate RNA-seq data. There were more than 3000 samples that involved, and the data included more than 70 cell lines as well as information about more than 12000 genes. Figure 1 gives the flowchart of CMap database, briefly introducing the data in different levels. In the genotyping and gene expression perspective, the updated GTEx data [2] contained information about more than 630 samples. The CommonMind Consortium (CMC) [3], which was proposed by Mount Sinai, Pittsburgh University and University of Pennsylvania, generated genetic data for more than 570 samples. Based upon CMC and GTEx brain tissue data, PsychENCODE consortium integrated one of the biggest human brain data sets so far [4], which could benefit the community for neuropsychiatry disease studies. STARNET [5] is another human DNA and RNA-seq data set, which contains information about more than 600 coronary artery disease patients. In addition, TCGA database [6] has genetic data of multiple cancers such as breast cancer, lung cancer etc. However, from perspective of biological technology and computational genetics or bioinformatics, people could not fully explain the relationship between genotypes and gene expressions as well as the correlation between them, which could potentially reveal medical

* Corresponding author.
E-mail address: zhang.wen81@gmail.com (Wen Zhang)

mechanisms for human traits and medication treatments. This mini-review aims to include gene expression, genotyping and the drug repositioning that related to the big data so as to look forward to the frontier issues in this field from the perspectives of computational genetics and bioinformatics.

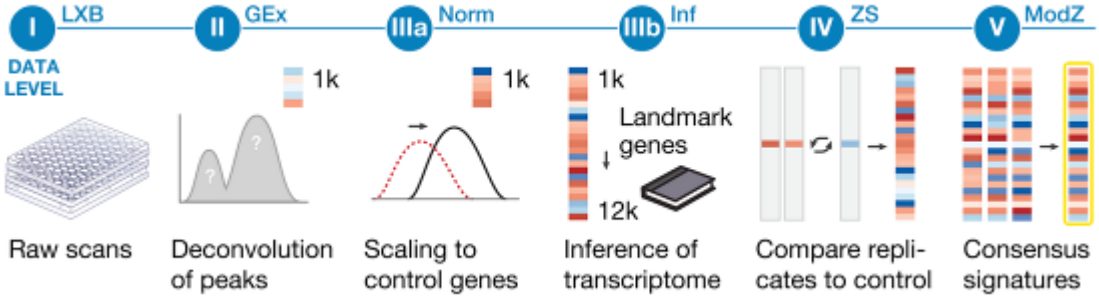

Figure 1. CMap data in each level (figure resource：https://clue.io/). A profile (also referred to as an experiment or an example) represents a data point for the generation of interference for a particular cell type at a particular therapeutic dose and for a particular duration of treatment. The numbers in the profile represent raw fluorescence intensity values (level 1 or raw data) or deconvolution (level 2 data) or post normalization (e.g., quantile normalization) to generate level 3 data. Finally, the configuration file is compared to the appropriate controls to generate a list of features that are differentially expressed (level 4 data). Each experiment is usually repeated for three times, which is then averaged into a vector of differential expression to create a signature (5 levels of data).

2. Method review

The genome-wide association studies (GWASs) have been widely leveraged to successfully identify and indicate a good amount of genetic variants or genes that associated with complex human diseases [7]. When investigators study genetic or bioinformatics problems, GWAS and relevant summary statistics could provide references for gene-level associations. S-PrediXcan [8] is one efficient computational method that based on GWAS analysis to predict gene expression profiles. This method benefits the community a lot since researchers could computationally obtain gene expression profiles from GWAS genotyping data using the available reference panels. The large-scale reference datasets as well as the related predictors are the fundamental basis of this method [9]. In high-throughput data, this approach could save a lot of costs, which is broadly employed. One specific application is in drug repositioning [10]. In [10], authors calculated gene expression profile from GWAS data, and repositioned candidate drug hits for diseases through comparisons with genetic medicine databases. The method belonged to the reverse profile principle category for drug repositioning. A series of efficient statistics were employed for the drug repositioning. That method could help reduce drug developing expenses and offer reference values for real clinical drug repositioning.

Nevertheless, the method has somehow limitations. First of all, the method depends on the predicted or imputed gene expressions. S-PrediXcan was embedded in the computations to predict the transcriptome profile. Therefore the method requires very high accuracy. Though the reference panels and the predictive method are of high efficiency [8-10], there is a long way to go to make the method better. For instance, the prediction process employs the elastic net method [9], which treats each SNP equally. From epigenomic study, we know that SNPs actually have different priorities, i.e., the association between every SNP and the trait varies. If we take this aspect into consideration, the prediction precision could be furthermore improved. Second, the powers of GWASs are different. The more samples that involved in the GWAS, the more powerful the GWAS. And the results of more powerful GWASs make more sense. Certainly, improved statistical methods are demanded in detecting drug-induced gene expressions. Traditional statistical methods are Spearman, Pearson or KS approaches. In [10], the combined statistics were utilized and different cut-offs were determined to provide good accuracy. Last but not the least, drug repositioning results need experimental or clinical validations. For some diseases such as psychiatry traits, the mouse model culture and verification are extremely time-consuming. There is no standard for validation of computational drug repositioning, which makes the approaches lack of real practice.

3. Discussions

Researches of computational drug repositioning are ascendant in recent decades. There are many issues that need to be addressed to further study how genetic variants affect gene expressions. Downstream analyses in this field are urgent to push forward the study. The purpose of this mini-review is to attract more attentions in this area so as to raise more ideas. With the development of information technology, experimental techniques and their engineering applications, prosperous results are highly expected.

References

[1] Subramanian A, Narayan R, Corsello SM, et al. A Next Generation Connectivity Map: L1000 Platform and the First 1,000,000 Profiles. Cell. 2017; 171: 1473-1452.
[2] GTEx Consortium. The Genotype-Tissue Expression (GTEx) project. Nat Genet. 2013;45:580–5.
[3] Fromer M, Roussos P, Sieberts SK, et al. Gene expression elucidates functional impact of polygenic risk for schizophrenia. Nat Neurosci. 2016;19:1442–53.
[4] Wang D, Liu S, Warrell J, et al. Comprehensive functional genomic resource and integrative model for the human brain. Science. 2018; 14,362(6420). pii: eaat8464. doi: 10.1126/science.aat8464.
[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] Cancer Genome Atlas Research Network, Weinstein JN, Collisson EA, Mills GB, et al. The Cancer Genome Atlas Pan-Cancer analysis project. Nat Genet. 2013;45:1113–20.
[7] Gusev A, Ko A, Shi H, et al. Integrative approaches for large-scale transcriptome-wide association studies. Nat Genet. 2016;48:245–52.
[8] Barbeira AN, Dickinson SP, Bonazzola R, et al. Exploring the phenotypic consequences of tissue specific gene expression variation inferred from GWAS summary statistics. Nat Commun. 2018; 9:1825.
[9] Gamazon ER, Wheeler HE, Shah KP, et al. A gene-based association method for mapping traits using reference transcriptome data. Nat Genet. 2015;47:1091–8.
[10] So HC, Chau C KL, Chiu WT, et al. Analysis of genome-wide association data highlights candidates for drug repositioning in psychiatry. Nat Neurosci. 2017; 20:1342-1349.