Topic 1:
Multidimensional data resources with gene prioritization system

簡介:
Major depressive disorder (MDD) is a complex and multi-factorial trait with the interplay between genetic and non-genetic risk factors. So far, there have been many datasets and individual studies with massive information from multiple resources of genetic findings. These include significant (or non-significant) results in association studies, linkage scans and gene expression studies for depression. This provides us an opportunity to conduct a prioritization system to utilize and combine multidimensional data to create an evidence-based gene set for depression.

問題描述:
Since we have 7 platforms, there are possible 8 different weights, and this forms an 8^7 = 2,097,152 weight matrix pool. Using this weight matrix pool and a pre-weight score for each platform, we can then calculate a combined score for each gene. An optimal weight scheme can be found from the weight matrix pool by the two-step approaches: (1) weight matrix selection by core genes; and (2) selection of weight matrix by GWAS dataset (p-values). All possible weight matrices were examined by the approach. Only those matrices could be selected if they could rank core genes in the top positions in ranked candidate gene list.

In first step, two parameters were introduced to assure that core genes being included in the top list. The setting could be changed freely. Five steps to select the weight matrices that met the threshold values of m and n: (1) a combined score of each gene is calculated with each weight matrix in the matrix pool; (2) the two lists of all genes in candidate genes and core genes are sorted by their combined scores; (3) the ranking positions of core genes in the ranked candidate gene list were recorded with a vector generated by program; (4) select the matrix basing on the threshold values of m and n, and record the position in the candidate gene list with a parameter; and (5) repeat the above processes until all weight matrices are analyzed.

In second step, we used the GWAS dataset to find which weight matrices

selected in step 1 can reach the optimization that top genes enrich markers with small p-value in GWAS. In this process, we used the smallest p-value of markers (SNPs) in each gene as the index. Program would select those top genes in number for each matrix and also randomly extracted 1000 subsets of genes from GWAS dataset with subset size j. for each random subset, we then compared whether p-value distribution is statistically different from the selected top ranked gene. The randomization was repeated 10 times to estimate the confidence of this approach (Table 1).

After the two-step approach, we randomly selected 10 matrices from the matrix pool that met different criteria. We then examined the distribution of GWAS p-values in the top ranked genes by these 10 weight matrices. We applied the best performance of weight matrix to calculating the combined scores for core genes and candidate genes (Figure 1). A cutoff value was set depending on the score distribution between the core and candidate genes for approximate maximization of combined scores for core genes (Figure 2).

Figure 1. A comparison of the GWAS p-values from the dataset of GAIN

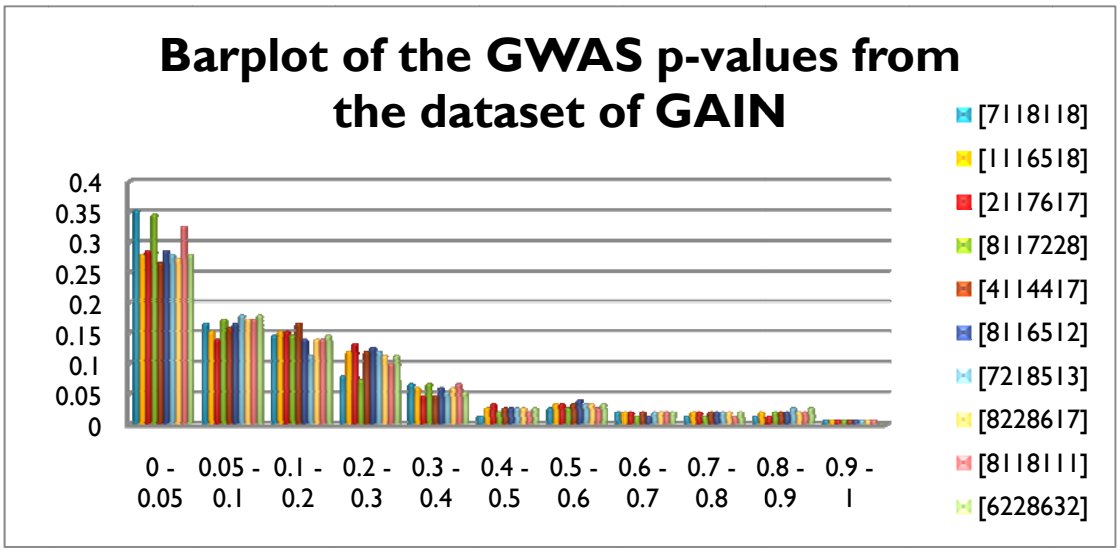
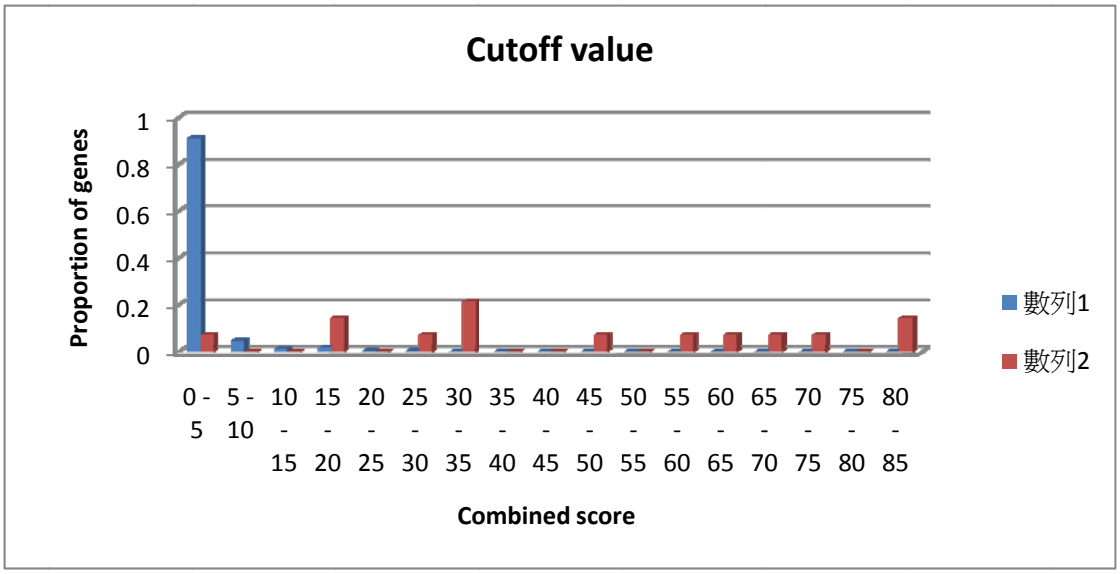


Figure 2. Selection of the cutoff value for candidate and core genes



We compared the gene expression pattern between the prioritized candidate genes and non-disease genes. We calculated the proportion of genes in a gene list expressed in a tissue by the count of genes expressed in the tissue divided by the total number of genes. Then we investigated the difference in gene expression distribution between the depression prioritized candidate genes and non-disease genes.

擬採用方法:

All programs are using statistical software R to perform. Two pre-weight settings were conducted, unequal weight and equal weight. Three sets of parameters were used for m and n, i.e. m = 0.8, 0.85, and 0.9; and n = 0.03, 0.04, and 0.05. The difficulty is that this program was computationally complicated. We have 18 settings. For this R program, we need about 8Gb memories for each setting. Time spent depends on the parameters setting. In first step, the time spent would need about 1-2 weeks for each setting. In second step, the time spent would need about 3-12 days for each setting.

預期成效:

This work was planned to be done by this October. We expected that our results would reveal that prioritized genes generated by such approach are promising for further biological experiment or replication.