G&I Genomics & Informatics

pISSN 1598-866X eISSN 2234-0742
Genomics & Informatics 2012;10(2):123-127
http://dx.doi.org/10.5808/GI.2012.10.2.123

ARTICLE

Performance Comparison of Two Gene Set Analysis Methods for Genome-wide Association Study Results: GSA-SNP vs i-GSEA4GWAS

Ji-sun Kwon[1], Jihye Kim[1], Dougu Nam[2], Sangsoo Kim[1]*

[1]Department of Bioinformatics and Life Science, Soongsil University, Seoul 156-743, Korea, [2]School of Nano-Bioscience and Chemical Engineering, Ulsan National Institute of Science and Technology, Ulsan 689-798, Korea

Gene set analysis (GSA) is useful in interpreting a genome-wide association study (GWAS) result in terms of biological mechanism. We compared the performance of two different GSA implementations that accept GWAS p-values of single nucleotide polymorphisms (SNPs) or gene-by-gene summaries thereof, GSA-SNP and i-GSEA4GWAS, under the same settings of inputs and parameters. GSA runs were made with two sets of p-values from a Korean type 2 diabetes mellitus GWAS study: 259,188 and 1,152,947 SNPs of the original and imputed genotype datasets, respectively. When Gene Ontology terms were used as gene sets, i-GSEA4GWAS produced 283 and 1,070 hits for the unimputed and imputed datasets, respectively. On the other hand, GSA-SNP reported 94 and 38 hits, respectively, for both datasets. Similar, but to a lesser degree, trends were observed with Kyoto Encyclopedia of Genes and Genomes (KEGG) gene sets as well. The huge number of hits by i-GSEA4GWAS for the imputed dataset was probably an artifact due to the scaling step in the algorithm. The decrease in hits by GSA-SNP for the imputed dataset may be due to the fact that it relies on Z-statistics, which is sensitive to variations in the background level of associations. Judicious evaluation of the GSA outcomes, perhaps based on multiple programs, is recommended.

Keywords: gene set analysis, genome-wide association study, GSA-SNP, i-GSEA4GWAS, imputation

Introduction

Gene set analysis (GSA) is useful in understanding the biological mechanisms underneath a phenotype by assessing the overall evidence of association of variations in an entire set of genes with a disease or a quantitative trait. GSA has recently been used to investigate many common diseases as an approach for the secondary analysis of a genome-wide association study (GWAS) result [1-3]. Currently, several software tools for GSA are available: GSA-SNP [4], i-GSEA4GWAS [5], GSEA-SNP [6], GeSBAP [7], and so on.

GSA-SNP is useful only when p-values of the single nucleotide polymorphism (SNP) markers are available. While GSA-SNP has implemented several options for estimating the significance of a gene set, its implementation of Z-statistics may be the most convenient. Other methods require permuted p-values that are obtained from sample permutation trials; this requires lengthy computation runs. The Z-statistics method accepts only one set of unpermuted original p-values and compares the score of a gene set against the background distribution made by all the genes; these p-values should be readily available for a typical GWAS. Similarly, i-GSEA4GWAS also uses only the original set of p-values and thus is as convenient as GSA-SNP. Instead of sample permutation, it estimates the significance of a gene set via SNP permutation [8]. One of its unique features is a scaling step that emphasizes the gene sets that are enriched with strongly associated genes.

Often, these approaches give different results in terms of the number of gene set hits. Hence, we compared these two methods using the same dataset while controlling the input parameters as much as possible. Here, we reanalyzed the

Received May 4, 2012; Revised May 21, 2012; Accepted May 22, 2012

***Corresponding author:** Tel: +82-2-820-0457, Fax: +82-2-820-0816, E-mail: sskimb@ssu.ac.kr

Copyright © 2012 by The Korea Genome Organization
It is identical to the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0/).

type 2 diabetes mellitus (T2DM) GWAS results for the Korean population against the Gene Ontology (GO) and Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway databases. The GWAS has been done with both the original unimputed and imputed genotypes. A large discrepancy in the number of significant gene set hits was observed between the two programs as well as the gene set databases. We also observed that the results were strongly affected by imputation.

Methods

Genotyping, imputation, and GWAS

Korean samples used in the East Asian T2DM meta-analysis have been described by Cho *et al*. [9]. The samples included 1,042 cases with T2DM and 2,943 controls from the Korea Association Resource (KARE) project that were recruited from Ansan and Ansung population-based cohorts, aged 40 to 69. We used the p-values of both the original unimputed (259,188 SNPs) and imputed (1,152,947 SNPs) genotypes, based on Affymetrix Genome-Wide Human SNP array 5.0 (Affymetrix, Santa Clara, CA, USA), that were made by removing samples and markers that failed a quality control test [10]. SNP imputation has been described by Cho *et al*. [9]. Briefly, IMPUTE, MACH, and BEAGLE were used, together with haplotype reference panels from the Japanese (JPT) and Han Chinese (CHB) samples that are available in the HapMap database on the basis of HapMap build 36.

GSA-SNP and i-GSEA4GWAS

The GO database was downloaded from the Gene Ontology Consortium (12,902 terms) [11], and the KEGG pathway database was downloaded from the KEGG (311 terms) [12].

For both databases, only the gene sets having 10-200 member genes were used (3534 GO and 211 KEGG terms). For both GSA-SNP and i-GSEA4GWAS, 20-kb padding was added to both ends of each gene. Usually, these methods, like i-GSEA4GWAS, pick up the best p-value and assign it to the encompassing gene. On the contrary, GSA-SNP allows one to choose different schemes for assigning the SNP p-values to each gene: either the best or the second best p-value within the gene boundary. Choosing the second best p-value has been recommended, as it may reduce the false positive associations with little loss of sensitivity [4].

For GSA-SNP, we downloaded the standalone program (as of Jan. 2011) and executed it locally. For i-GSEA4GWAS, we used the web server version by uploading the SNP p-values. GSA-SNP allows several approaches of evaluating gene set significance. While other approaches require p-values from permutation tests, Z-standardization requires no permuted p-values. The score of a gene is defined as -log of the p-value assigned to the gene. For each gene set, the scores of its member genes are averaged, and the Z-statistics of these scores are used to estimate the significance under the assumption of a normal distribution. The effect of multiple testing is corrected by the false discovery rate (FDR) method [13]. On the other hand, i-GSEA4GWAS compares the distribution of the member gene scores of a gene set to all the genes using K-S statistics and corrects the multiple testing effect using FDR that is based on SNP permutation tests. Variation in the number of member genes among gene sets is taken care of by multiplying a ratio of 'highly significant' genes in a gene set relative to those among all genes. Here, the 'highly significant' genes are defined as the genes that map with at least one of the top 5% of all SNPs in the dataset.

Results

The performance of GSA-SNP and i-GSEA4GWAS looked quite different in terms of the number of hits (Table 1). GSA-SNP detected 27-94 hits for either the imputed or unimputed genotype dataset with GO gene sets, while 9-20 hits were detected with KEGG gene sets. Consistently, more gene sets were hit with the scheme of using the second-best p-value of a gene than the best one. Previously, it has been recommended that one assign the second-best, not the best,

Table 1. The number of gene set hits identified by gene set analyses

Software	Gene score[a]	GO		KEGG	
		Unimputed	Imputed	Unimputed	Imputed
i-GSEA4GWAS	Best	283	1,070	12	78
GSA-SNP	Best	61	27	14	9
	Second best	94	38	20	19

GO, Gene Ontology; KEGG, Kyoto Encyclopedia of Genes and Genomes; GSA, gene set analysis; SNP, single nucleotide polymorphism.
[a]Either the best or second-best p-value of SNPs residing inside or within 20 kb of the gene boundary was assigned to each gene as the score. Unlike i-GSEA4GWAS, which assigns the best p-value, GSA-SNP has an option to assign the second-best p-value.

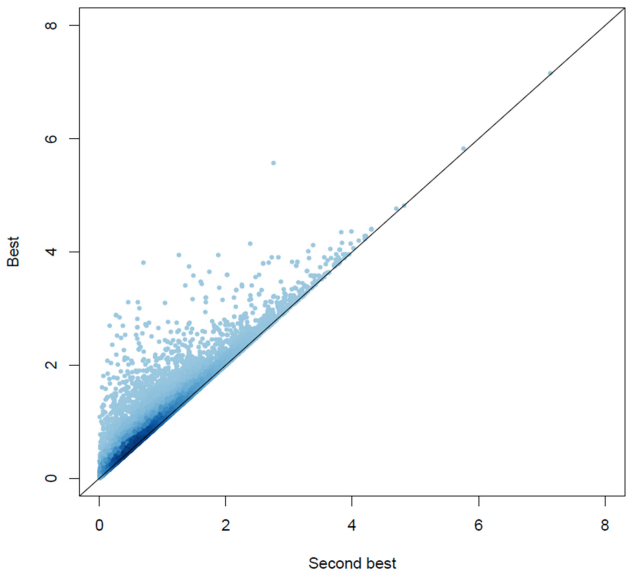

Fig. 1. Comparison of the gene scores calculated by two different schemes. All 15,829 genes that were mapped by at least two single nucleotide polymorphisms (SNPs) (either genotyped or imputed) were included in the high-volume scatter plot that displays the local density of points by a false color representation. For a given gene, the p-values of SNPs located inside or within 20 kb of the gene boundary were surveyed. The best (or second-best) of their -log-transformed values was assigned as the gene score. The X and Y axes represent the second-best and best values, respectively.

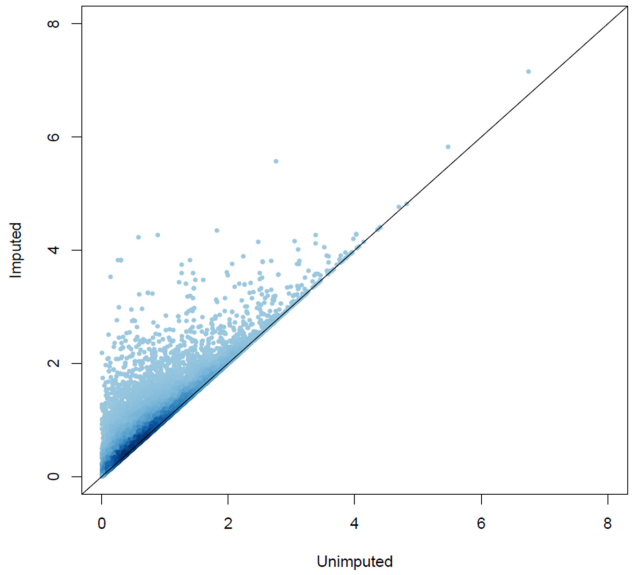

Fig. 2. Comparison of the gene scores from the unimputed and imputed datasets. All 15,829 genes that were mapped by at least two single nucleotide polymorphisms (SNPs) (either genotyped or imputed) were included in the high-volume scatter plot that displays the local density of points by a false color representation. For a given gene, the p-values of SNPs located inside or within 20 kb of the gene boundary were surveyed. The best of their -log-transformed values was assigned as the gene score. The X and Y axes represent the gene scores from the unimputed and imputed datasets, respectively.

p-value to a gene due to a concern that the use of the best p-values of a gene may produce more false positives than the second-best one [4]. We compared the distribution of gene scores that were calculated based on the best and the second-best p-values using a high-volume scatter plot that represented the local density of points by a false color representation (Fig. 1). One may notice the densely populated points along the diagonal axis, meaning that the differences in gene scores were small for the majority of genes. Nevertheless, there were many genes off-diagonal; for these genes, the gene scores that were calculated with the best p-values were larger than those calculated with the second-best p-values. While this effect was negligible for the genes having high scores ($p < 10^{-4}$), many genes having low scores displayed large differences. The latter were genes that had only one SNP within its boundary plus 20 kb of padding standing out in terms of significance and the rest falling short. If the best SNP of a gene had been located within a strong linkage disequilibrium (LD) block, the second-best SNP would have been chosen from this block with a p-value close to the best one. On the other hand, if the best SNP is in weak LD with the second-best one, they would differ from each other considerably. Considering that Fig. 1 was based on the highly densely imputed genotypes, those genes that showed a large difference may have been located within narrow LD blocks or recombination hot spots where imputation may be invalid. Indeed, the genes that were not located within haplotype blocks showed larger differences in gene scores than those located within haplotype blocks (Supplementary Fig. 1). For those genes located within haplotype blocks, there were inverse relationships between the block size and the difference in gene scores (Supplementary Fig. 2). It is also possible that the apparent associations seen with the best p-values may have been due to random errors; assigning the second-best p-value to a gene may then reduce false associations. One may argue that the large difference between the best and second-best scores for some genes may be due to the small number of SNPs for those genes. If this is the case, there should be an inverse relationship between these two quantities. As shown in Supplementary Fig. 3, there is no such evidence.

Here is the rationale for the more hits made by GSA-SNP with the use of the second-best p-values than the best p-values in assigning them to a gene. GSA-SNP assigns a score to a gene set by averaging gene scores (-logP) of its member genes and calculates Z-statistic by subtracting the mean of all gene scores from the gene set score. Assigning the second-best p-value to a gene produces lower gene scores than assigning the best one. While this effect is more

pronounced with low-scoring genes, the high-scoring genes suffer little. Since the mean of all gene scores would also decrease by using the second-best p-values compared to the use of the best p-values, the resulting Z-statistic for a gene set that is composed of high-scoring genes would then increase, yielding more hits.

The GSA-SNP runs with p-values of the imputed genotypes consistently produced fewer hits than with those of the unimputed genotypes. Imputation guesses the genotype of an untyped marker and fills it in. The GWAS p-value of an imputed SNP can be either larger or smaller than those of the neighboring genotyped, unimputed ones. For the former cases, it would not change the best p-value that is assigned to a gene. On the other hand, for the latter cases, the best p-value that is assigned to a gene can get smaller. As shown in Fig. 2, the imputation increased the significance of many genes. This has the effect of reducing the number of hits for the GSA-SNP runs with the imputed dataset compared with the unimputed one, similar to the argument given above for the schemes of assigning the best or second-best p-values to a gene.

With the p-values of the imputed SNPs, i-GSEA4GWAS claimed 1,070 GO terms as significant, almost 40-fold more than GSA-SNP, which yielded only 27 terms. Even with the p-values of the original unimputed SNPs, i-GSEA4GWAS produced about 4.6 times more terms than GSA-SNP (283 vs. 61). It appears that i-GSEA4GWAS unrealistically produced too many hits (about 1/3 of the input GO terms), hampered by the high proportion of false positives. This trend was much more pronounced with GO than KEGG, probably due to the more redundant nature of GO than KEGG.

Unlike GSA-SNP, which reported fewer hits with the imputed dataset than the unimputed one, i-GSEA4GWAS produced about 5 times more hits with the former than the latter. Why did i-GSEA4GWAS perform even more poorly with the imputed dataset that should be more ideal in terms of marker density than the unimputed one? i-GSEA4GWAS assigns the best p-value to a gene and evaluates the gene set score by comparing the distribution of gene scores (-logP) of a gene set to that of all gene scores. The significance of the gene set score is estimated by the FDR, which compares the unpermuted distribution of the gene set scores with that of the gene set scores generated from a number of SNP-permuted datasets. The gene set score is multiplied by a factor, k/K, where k represents the proportion of so-called 'significant' genes within the gene set and K represents the proportion of 'significant' genes of all genes. i-GSEA4GWAS defines the 'significant' genes as those that are mapped with at least one 'top 5%' SNP. This step is a unique feature of i-GSEA4GWAS, which has the prefix 'i' in its name. Generally, an imputed dataset has much higher marker density than the corresponding unimputed one (4.45-fold difference in our case). The so-called 'top 5%' SNPs will be more with the imputed dataset than the unimputed one (again, a 4.45-fold difference in our case). This corresponded to p-value cutoff of 0.047 in our case. The number of so-called 'significant' genes does not increase as much as the increase in the marker density: from 2,967 genes for the unimputed dataset to 4,156 genes for the imputed one in our case. The concept of augmenting gene set scores by the proportion of 'significant' genes may be useful, as demonstrated previously in comparison with GSEA [5, 14]. However, the 'top 5%' threshold used by i-GSEA4GWAS may be too high, inflating the false positive rates. This inflation may be more pronounced with an imputed dataset - 1,000 more genes were treated as 'significant' with our imputed dataset than with the unimputed one. Probably, this is one of the main reasons for the particularly poor performance of i-GSEA4GWAS with the imputed dataset.

Discussion

GSA is useful method in interpreting the result from a GWAS. A systematic evaluation of its performance is of paramount interest to the GWAS community, as the method is getting popular. Here, we compared the performance of two such methods using the common datasets and gene set databases. While GSA-SNP behaved predictably, i-GSEA4GWAS produced too many hits for most of the test settings. For example, i-GSEA4GWAS reported 3.8- and 6.5-fold more hits, respectively, for GO and KEGG, with the imputed dataset than with the unimputed one. Imputation is such a useful practice that augments the power of a genotype dataset, and ideally, gene set analyses can benefit from it. Our study warns that one must be cautious in applying i-GSEA4GWAS to an imputed dataset. As we pinpointed above concerning the 'top 5%' threshold as the potential cause of the high hit rates of i-GSEA4GWAS, it would be interesting to re-evaluate its performance with lower thresholds. Currently, the threshold is not available for the users to change it. It would have been better if the user could choose it at will. For GSA-SNP, we recommend using an imputed dataset if at all possible. GSA-SNP allows the user to choose k in assigning the k-th best p-value to a gene. We recommend using $k = 2$ instead of $k = 1$, as the latter inflates the scores for some genes, diminishing the power of GSA.

Supplementary materials

Supplemetary data including three figures can be found with this article online at http://www.genominfo.org/src/sm/gni-10-123-s001.pdf.

Acknowledgments

This research was performed within the Consortium for Large Scale Genome Wide Association Study III (2011E7300400), which was supported by the genotyping data (the Korean Genome Analysis Project, 4845-301) and the phenotypic data (the Korean Genome Epidemiology Study, 4851-302) from the Korea Center for Disease Control. This work was supported financially by the Basic Science Research Program through the National Research Foundation of Korea (NRF), funded by the Ministry of Education, Science, and Technology (NRF-2010-0021811).

References

1. Lambert JC, Grenier-Boley B, Chouraki V, Heath S, Zelenika D, Fievet N, *et al*. Implication of the immune system in Alzheimer's disease: evidence from genome-wide pathway analysis. *J Alzheimers Dis* 2010;20:1107-1118.
2. Baranzini SE, Galwey NW, Wang J, Khankhanian P, Lindberg R, Pelletier D, *et al*. Pathway and network-based analysis of genome-wide association studies in multiple sclerosis. *Hum Mol Genet* 2009;18:2078-2090.
3. Torkamani A, Topol EJ, Schork NJ. Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* 2008;92:265-272.
4. Nam D, Kim J, Kim SY, Kim S. GSA-SNP: a general approach for gene set analysis of polymorphisms. *Nucleic Acids Res* 2010;38:W749-W754.
5. Zhang K, Cui S, Chang S, Zhang L, Wang J. i-GSEA4GWAS: a web server for identification of pathways/gene sets associated with traits by applying an improved gene set enrichment analysis to genome-wide association study. *Nucleic Acids Res* 2010;38:W90-W95.
6. Holden M, Deng S, Wojnowski L, Kulle B. GSEA-SNP: applying gene set enrichment analysis to SNP data from genome-wide association studies. *Bioinformatics* 2008;24:2784-2785.
7. Medina I, Montaner D, Bonifaci N, Pujana MA, Carbonell J, Tarraga J, *et al*. Gene set-based analysis of polymorphisms: finding pathways or biological processes associated to traits in genome-wide association studies. *Nucleic Acids Res* 2009; 37:W340-W344.
8. Guo YF, Li J, Chen Y, Zhang LS, Deng HW. A new permutation strategy of pathway-based approach for genome-wide association study. *BMC Bioinformatics* 2009;10:429.
9. Cho YS, Chen CH, Hu C, Long J, Ong RT, Sim X, *et al*. Meta-analysis of genome-wide association studies identifies eight new loci for type 2 diabetes in east Asians. *Nat Genet* 2012;44:67-72.
10. Cho YS, Go MJ, Kim YJ, Heo JY, Oh JH, Ban HJ, *et al*. A large-scale genome-wide association study of Asian populations uncovers genetic factors influencing eight quantitative traits. *Nat Genet* 2009;41:527-534.
11. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, *et al*. Gene ontology: tool for the unification of biology. *Nat Genet* 2000;25:25-29.
12. Kanehisa M, Goto S. KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000;28:27-30.
13. Benjamini Y, Hochberg Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Series B Methodol* 1995;57:289-300.
14. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA, *et al*. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proc Natl Acad Sci U S A* 2005;102:15545-15550.