Relevant references

Publications:

1. Lo, S.H. and Zheng, T. (2002) Backward haplotype transmission association (BHTA) algorithm-a fast multiple-marker screening method, Hum Hered, **53**, 197-215.

   Note: *This is the first paper of the partition retention framework that proposed the original haplotype-trait disequilibrium (HTD) score for case-parent trio data, which motivated the general form of the measure of influence based on partitions.*

2. Lo, S.H., Liu X. and Shao, Y. Z. (2003) A marginal likelihood model for family-based data. Ann. Hum. Genet., 67(4), 357-366.

3. Lo, S.H. and Zheng, T. (2004) A demonstration and findings of a statistical approach through reanalysis of inflammatory bowel disease data, Proc Natl Acad Sci U S A, **101**, 10386-10391.

   Note: *In this paper, BHTA (Lo and Zheng 2002) was applied to a dataset of 235 case-parent trios of inflammatory bowel disease (more specifically, Crhon's disease). 467 microsatellite markers were genotyped on these individuals. We confirmed 6 out of 7 IBD loci in the literature at the time of the analysis and reported 4 novel loci with the strongest signal by BHTA.*

   Recently Barrett et al (2008) conducted a genome-wide association study on 3,230 cases and 4,829 controls that identified 30 loci associated with the risk of Crohn's disease, which endorsed many of our 2004 findings. Most notably, the four novel loci reported in Lo and Zheng (2004), previously all undiscovered, were all confirmed by Barrett et al (2008). Considering that our 2004 study was with only hundreds of markers and 235 trios (only 7% of the size of the GWAs of Barrett et al), we believe our approach more efficiently extracts information in a genetic data set by dint of considering gene-gene interaction.

4. Ionita, I. and Lo, S.H. (2005) Multilocus linkage analysis of affected sib pairs. Human Heredity, 60, 227-240

   Note: *In this paper, the partition retention idea was extended to linkage studies by a smart use of multilocus inheritance vectors. The method extracts linkage information and is able to consider interactions among loci.*

5. Zheng, T., Wang, H. and Lo, S.H. (2006) Backward genotype-trait association (BGTA)-based dissection of complex traits in case-control designs, Hum Hered, **62**, 196-212.

   Note: *In this paper, the partition retention idea was extended to genotype data under case-control design.*

6. Ding, Y., Cong, L., Ionita-Laza, I., Lo, S.H. and Zheng, T. (2007) Constructing gene association network for rheumatoid arthritis using the backward genotype-trait association (BGTA) Algorithm. In "Genetic Analysis Workshop 15: Gene Expression Analysis and Approaches to Detecting Multiple Functional Loci." BMC Proceedings **1** S1:S13.

   Note: *In this paper, we applied BGTA to two data sets of Rheumatoid Arthritis. The first one was a genome scan with 5407 markers and the second was a candidate gene studies. Using different screening strategies, we were able to indentify gene-gene interactions associated with the risk of RA and construct association networks based on these findings.*

7. Zheng, T., Wang, S., Cong, L., Ding, Y., Ionita-Laza, I. and Lo, S.H. (2007) Joint study of genetic regulators for expression traits related to breast cancer. In "Genetic Analysis Workshop 15: Gene Expression Analysis and Approaches to Detecting Multiple Functional Loci." BMC Proceedings **1** S1:S10.

   Note: *In this paper, we used BGTA to study genetic regulators of the gene expression levels of breast cancer related genes.*

8. Lo, S.-H., Chernoff, H., Cong, L., Ding, Y. and Zheng, T. (2008) Discovering interactions among BRCA1 and other candidate genes involved in Sporadic Breast Cancer. Proc Natl Acad Sci U S A **105**, 12387-12392.

Note: *In this paper, using high-density genomewide association study data, we carried out gene-wise analysis of gene-gene interaction among 18 breast cancer candidate genes. We adopted the partition retention influence measure and devised new measures of gene-gene interaction.*

9. Chernoff, H., Lo, S. H. and Zheng, T. (2009) Discovering influential variables: a method of partitions. Annals of Applied Statistics. Vol. 3, No. 4, pp. 1335-1369

Note: *In this paper, we introduced a general partition-based framework called Partition Retention (PR) that included all previously developed methods as special cases. The preliminary results suggested that PR is capable of explaining a number of important puzzles: why the methods remain sensitive and effective even when the data are high-dimensional; and why the methods are much more powerful than other popular methods such as the Random Forests (RF) and Multifactor Dimensionality Reduction (MDR).*

Also proposed and studied preliminarily in Chernoff, Lo and Zheng (2008) was a screening strategy guided by the partition-based information measures, which was called the "resuscitation" method. It is a systematic screening scheme for large-scale genomic studies such as current genome-wide association studies (GWAs). Using a genome scan data on Rheumatoid Arthritis (RA), we compared this strategy with other methods including the Random Forests and the multi-factor dimensionality reduction (MDR) and concluded that our resuscitation strategy rediscovered many more genetic loci that were previously reported to be associated with RA. With supplement funding, we can further our research along this very important direction and provide a more timely solution to efficiently identify disease-associated high-dimensional gene-gene interactions using current GWAs data.

10. Zheng, T., Chernoff, H., Hu, I., Ionita-Laza, I., & Lo, S. H. (2011). Discovering influential variables: a general computer intensive method for common genetic disorders. In *Handbook of Statistical Bioinformatics* (pp. 87-107). Springer Berlin Heidelberg. [url]

Note: *This is a review paper that discussed several extensions of BHTA/BGTA and unified them under the same framework of partition retention (PR).*

11. Liu, Y., Huang, C. H., Hu, I., Lo, S. H., & Zheng, T. (2011, November). Association screening for genes with multiple potentially rare variants: an inverse-probability weighted clustering approach. In *BMC proceedings* (Vol. 5, No. 9, p. 1).

Note: *In this paper, to address the need of rare variants in genetic studies, we used a novel clustering algorithm to create partitions before applying the PR method.*

12. Wang, H., Lo, S. H., Zheng, T., & Hu, I. (2012). Interaction-based feature selection and classification for high-dimensional biological data. *Bioinformatics*, *28*(21), 2834-2842.

Note: *This was the first paper where the PR's method was applied to improve prediction in a supervised learning setting.*

13. Fan, R., & Lo, S. H. (2013). A robust model-free approach for rare variants association studies incorporating gene-gene and gene-environmental interactions. *PloS one*, *8*(12), e83057.

Note: *In this paper, modification to I-score to accommodate rare variants was proposed and compared with other methods on rare variants.*

14. Fan, R., Huang, C. H., Hu, I., Wang, H., Zheng, T., & Lo, S. H. (2014, June). A partition-based approach to identify gene-environment interactions in genome wide association studies. In *BMC proceedings* (Vol. 8, No. 1, p. 1).

Note: *We proposed a special method for measuring influence on Y from GxE interactions.*

15. Agne, M., Huang, C. H., Hu, I., Wang, H., Zheng, T., & Lo, S. H. (2014, June). Considering interactive effects in the identification of influential regions with extremely rare variants via fixed bin approach. In *BMC proceedings* (Vol. 8, No. 1, p. 1).

Note: *In this paper, the partition was constructed by presence and counts of rare variants in a genetic region.*

16. Lo, A., Chernoff, H., Zheng, T., & Lo, S. H. (2015). Why significant variables aren't automatically good predictors. Proceedings of the National Academy of Sciences, 112(45), 13892-13897.

Note: *In this paper, we provide a general discussion on why we need new measures and methods for selecting variables or variable sets for predictions. We provide simulation studies to show that significance variables and predictive variables do not always overlap.*

17. Lo, A., Chernoff, H., Zheng, T., & Lo, S. H. (2016). Making Good Prediction: A Theoretical Framework.

Note: *In this paper, we put forward the theoretical framework of evaluating prediction rate of variable set based on PR's I-score.*