RESEARCH STATEMENT

Zhiguang (Caleb) Huo (zhh18@pitt.edu)

My research interest lies in both statistical **methodology and application on genomics and bioinformatics**. Nowadays, large amount of genomics data are available and required to develop rigorous statistical method to understand the disease and pursue better treatment. These datasets provided unprecedented opportunities to reveal cancer mechanisms via combining multiple cohorts or multiple-level omics data types. I have worked on **horizontal omics meta-analysis** (combine multiple cohorts of the same type of omics data) and **vertical omics integrative analysis** (combine multi-level omics data of the same patient cohort), which will help increase statistical power, interpretability and reproducibility. **I am particular interested in Bayesian approach, regularizations (frequentist perspective), graphical models and effective statistical computing,** which are capable of accommodating the high-dimensional nature of genomics data. In terms of genomics and bioinformatics application, I have worked extensively on various **cancer** types as well as **psychiatry** data with local collaborators. These genomic data ranges from microarray to sequencing, gene expression, copy number variation to methylation. These collaborations not only allowed me to apply statistics methods towards innovating biological findings, but also motivated me to develop most appropriate statistical methods for their data. In return, I have worked on user friendly **software development** and efficient computing algorithm, in order to facilitate these biological application. However, genomics and bioinformatics are fast evolving field and new types of data are gradually replacing old types of data (e.g. NGS data are replacing microarray data). Therefore, I am willing to expand my research towards new powerful data types (e.g. imaging data, **single cell data**) and I have actively exposed to single cell data.

Below is a summary of my **past and going research, as well as future research plan** in terms of statistical methodology and bioinformatics application.

1 Statistical methodology

1.1 Data integration

Due to rapid development of high-throughput experimental techniques and fast dropping prices, many transcriptomic datasets have been generated and accumulated in the public domain (e.g. TCGA, GEO, SRA). Local research institute/hospitals also generate tons of data. Single cohort/data type may suffer from small sample size issue or reproducibility issue. There are multiple cohorts generated from different sources, also given the same cohort, there are multiple data types (gene expression, methylation, CNV, mutation). A natural question is how to combine or integrate these complex data and increase statistical power and interpretation. There are two directions for data integration: horizontal meta-analysis and vertical integration.

1.1.1 Horizontal meta-analysis

Meta-analysis aims to combine multiple same-type genomic data to increase statistical power, accuracy and validated conclusion. I have been extensively working in this area and developed several statistical approaches: in terms of disease subtype discovery; differential expression gene detection; differential expression meta-pattern detection; differential co-expression network and dimension reduction by combining same type of omics data of multiple cohorts.

- **A meta-analytic framework of sparse clustering:** Disease phenotyping by omics data has become a popular approach that potentially can lead to better personalized treatment. Identifying disease subtypes via unsupervised machine learning is the first step towards this goal. I have developed **Meta Sparse K-means** algorithm[1] to discovery disease subtype from multiple genomics cohort. The algorithm will achieve feature selection and guarantee consistent clustering pattern across cohorts simultaneously. The algorithm is important in that it will enhance the reproducibility and stability of disease subtype discovery.
- **A meta-analytic framework of bio-marker detection of differential expression pattern characterization:** Meta-analysis combining multiple transcriptomic studies can increase statistical power to detect disease related biomarkers. I have developed a novel Bayesian latent hierarchical model (**BayesMP**)[2],

which is capable of detecting genes that are differentially expressed (DE) in only a subset of the combined studies, and the latent variables help quantify homogeneous and heterogeneous differential expression signals across studies. This is helpful to increase disease biomarkers detection power and facilitate biologist to formulate hypothesis from the data.

- **A meta-analysis framework for differential co-expression network detection.** This is a project that I was partially involved in. Gene co-expression network analysis from large transcriptomic studies is often used to elucidate potential gene-gene interactions and regulatory mechanisms. Co-expression networks are first constructed in cases and controls separately in each study. Differential co-expression seed modules are detected by optimizing an energy function via simulated annealing. The result sheds light on the underlying disease mechanisms in a systems manner.

- **Meta analytical principal component analysis.** This is a project that I was partially involved in. Due to high-dimensional nature of the data, methods such as principal component analysis (PCA) have been widely applied, aiming at effective dimension reduction and exploratory visualization. In this paper, we combine multiple omics datasets of identical or similar biological hypothesis and introduce two variations of meta-analytic framework of PCA, namely MetaPCA.

1.1.2 Genomic data integration

Genomic data integration combines multi-omics data (e.g. gene expression, CNV, genotyping, methylation, somatic mutation, miRNA) of the same cohort. We could gain statistical power and have a better understanding of the inner omics relationships.

- **An integration framework of sparse clustering:** With the accumulation of massive multi-omics datasets and established biological knowledge databases, omics data integration with incorporation of rich existing biological knowledge is essential for deciphering biological mechanism behind the complex diseases. I proposed an integrative sparse K-means (**IS-Kmeans**)[3] approach to combine multi-omics datasets to discover disease subtypes with the guidance of prior biological knowledge via sparse overlapping group lasso. This computation efficient algorithm helps reveal consistent disease subtype patterns and inner omics relationships.

- **Bayesian omics data integration**

1.1.3 Future direction

Integrating epi-genetic data, neuroimaging data.

1.2 High dimensional data optimization and statistical computing

Genomics data has more than 20,000 genes in human being and normally there are only tens or hundreds of samples. Not all genes are related to the disease of interest. How to effectively select the intrinsic subset of genes from the high dimensional data are very important. There are two main folds for high dimensional data: 1, optimization; 2, theory. I have worked on high dimensional data optimization problems exclusively.

1.2.1 High dimensional data optimization

I am quite familiar with optimization algorithms, especially varies lasso related problems. My research projects are highly involved with optimization (ref).

- KKT, convert the complex problem to a optimization problem with lasso.

- The second example ADMM to solve the overlapping group lasso problem.

1.2.2 statistical computing

- Adaptively weighted Fisher's method (AW) is an powerful approach to combine p-values from K independent studies and provide better biological interpretation by characterizing which studies contribute to meta-analysis. However, the original AW method suffers from slow computing because of permutation. By using importance sampling and spline, we obtain a fast computing for AW, which is weighted fisher. This is an on-going work.
- github. Github is originally design for programmers. It is also very powerful for statistician. This is an easy way to host my code, program and packages. In the future, I will also make my teaching materials on gitbub.

1.2.3 Future work

- I have worked extensively on lasso, group lasso or overlapping group lasso problems. Other regularization techniques such as penalization on inverse covariance matrix or low rank penalty are also appealing to genomic applications.
- High dimensional problems are challenging in perspective of optimization and theory. I have been working of high-dimensional optimization problem quite a lot. High-dimensional theory is also interesting and challenging, and it can provide theoritical guarantee for the sound methodology.

1.3 Bayesian inference and graphical model

Bayesian is very convenient to model complex data structure, get soft inference on result. Graphical model is very powerful to model the dependent structure of data. It has been widely used in the field of computational biology, computer vision, natural language processing.

1.3.1 Bayesian nonparametric

- For meta-analysis combining p-value methods, frequentists' approach may suffer from in composite null and alternative hypothesis, which will make the inference and false positive rate control not accurate. Bayesian is a natural way to get around this problem and I developed BayesMP. In order to preserve the uncertainty of the alternative distribution and avoid empirical Bayes approach, we utilized the Bayesian non-parametric model (Dirichlet process) to characterize the alternative distribution.

1.3.2 single cell methylation imputation using conditional random field

- Epigenetic plays an important role in gene regulation. Recent development of single cell methylation makes it feasible to look at methylation at single cell level. However, due to inadequate of bisulfate conversion and sequencing error, the observed methylation level is not interpretable. I proposed a fast single cell imputation method using conditional random field. This is an on-going project and part of my thesis as well.

1.3.3 Feature selection using Bayesian approach

- Variable selection is a pervasive question in modern high-dimensional data analysis where the number of features often exceeds the sample size (a.k.a. small-n-large-p problem). Incorporation of group structure knowledge to improve variable selection has been widely studied. In this paper, we consider prior knowledge of a multi-layer overlapping group structure to improve variable selection in regression setting.

1.3.4 Future work

- Single cell mixture problem.
- Single cell expression clustering.

2 Bioinformatics application

As a biostatistician, an important job is to work with local biologist on their data. This is exciting for me not only I can learn interesting and well-designed experiment, but also their data can motivate me to develop use statistical methodology. Genomics and bioinformatics are fast moving fields and I have been exposed to multiple types of data. Though data are different and processing pipeline may differnet, but the enssence of statistical approach remain the same.

2.1 Genomics data application

- Cancer research. Disease subtype of breast cancer, DNA methylation via sequencing, copy number variation in prostate cancer,
- Psychiatry. Cell specific Psychiatry diseases, circadian patterns, sex related depression effect, human iPSC data.
- At this stage, I am also exposured in single cell data (single cell methylation and single cell gene expression). I am trying to develop methods to the newly developed data type. This will be part of my thesis.

2.2 Software development

There are often gaps between biology and statistics and bioinformaticians should be the bridge to bring them together. I have developed several R packages for other user to use. However, these are far less for biologists. User-friendly software is the key component for biologists who are not familiar with statistical computing. Previously I was involved in Meta Omics R package. I spent a summer to improve the Meta Omics R package into a user-friendly JAVA software. Now, we are developing a Meta Omics suit implemented using R shiny. The purpose is to design user-friendly software to facilitate biologist to fully utilize these statistical algorithms.

3 Summary

To summarize, I am interested in both genomics/bioinformatics methodology and application.

For methodology, I am particular interested in meta-analysis, data integration, high-dimensional data analysis, Bayesian approach, optimization, statistical computing, and software development. For high-dimensional data, theory is equally important as optimization. I may pursue high-dimensional data theory in the future.

For application, I am interested in all types of genomics data (microarray, sequencing...). But the field is moving very fast and data are evolving so quickly. I am also interested in imaging data, if there is an opportunity for me in the future.

I will keep update with the data and develop cutting edge methodology to help biologists with the data interpretation.

References

[1] **Zhiguang Huo**, Ying Ding, Silvia Liu, Steffi Oesterreich, and George Tseng. Meta-Analytic Framework for Sparse K-Means to Identify Disease Subtypes in Multiple Transcriptomic Studies. *Journal of the American Statistical Association*, 111, no. 513 (2016): 27-42.

[2] **Zhiguang Huo**, Chi Song, George C. Tseng. (2016) Bayesian latent hierarchical model for transcriptomic meta-analysis to detect biomarkers with clustered meta-patterns of differential expression signals. Submitted to *Annals of Applied Statistics* (under second round of review).

[3] **Zhiguang Huo**, George C. Tseng. (2016) Integrative Sparse K-means for disease subtype discovery using multi-level omics data. Submitted to *Annals of Applied Statistics* (under second round of review).

[4] Silvia Liu, Wei-Hsiang Tsai, Ying Ding, Rui Chen, Zhou Fang, **Zhiguang Huo**, SungHwan Kim, Tianzhou Ma, Ting-Yu Chang, Nolan Michael Priedigkeit, Adrian V. Lee, Jianhua Luo, Hsei-Wei Wang, I-Fang Chung, George C. Tseng. (2015). Comprehensive evaluation of fusion transcript detection algorithms and a meta-caller to combine top performing methods in paired-end RNA-seq data. *Nucleic Acids Research*, 10.1093/nar/gkv1234.

[5] Tiffany A. Katz, Serena G. Liao, Vincent J. Palmieri, Robert K. Dearth, Thushangi Pathiraja, **Zhiguang Huo**, Patricia Shaw, Sarah Small, Nancy E. Davidson, David G. Peters, George C. Tseng, Steffi Oesterreich, Adrian V. Lee. (2015) Targeted DNA methylation screen in the mouse mammary genome reveals a parity-induced hypermethylation of igf1r which persists long after parturition. *Cancer Prevention Research*, pages canprevres-0178.

[6] Yan P. Yu, Silvia Liu, **Zhiguang Huo**, Amantha Martin, Joel B. Nelson, George C. Tseng and Jian-Hua Luo. (2015) Genomic copy number variations in the genomes of leukocytes predict prostate cancer clinical outcomes. *PloS one*, 10(8):e0135982.

[7] Xingbin Wang, Dongwan Kang, Kui Shen, Chi Song, Shuya Lu, Lunching Chang, Serena G. Liao, **Zhiguang Huo**, Naftali Kaminski, Etienne Sibille, Yan Lin, Jia Li and George C. Tseng. (2012) A Suite of R Packages for Quality Control, Differentially Expressed Gene and Enriched Pathway Detection in Microarray Meta-analysis. *Bioinformatics*, 28:2534-2536.

[8] Dominique Arion, **Zhiguang Huo**, John F. Enwright, John P. Corradi, George Tseng and David A. Lewis. Transcriptome alterations in prefrontal pyramidal neurons distinguish schizophrenia from bipolar and major depressive disorders. Submitted to *Biological Psychiatry*, (under second round of review).

[9] SungHwan Kim, Dongwan Kang, **Zhiguang Huo**, Yongseok Park, George C. Tseng. (2016) Meta-analytic principal component analysis. Submitted to *Annals of Applied Statistics* (under revision).

[10] Li Zhu, Ying Ding, Cho-Yi Chen, Lin Wang, **Zhiguang Huo**, SungHwan Kim, Christos Sotiriou, Steffi Oesterreich and George C. Tseng. (2016) MetaDCN: meta-analysis framework for differential coexpression network detection with an application in breast cancer Submitted to *Bioinformatics* (under revision).

[11] Oesterreich, S., Katz, T.A., Logan, G., Levine, K., Nagle, A., **Huo, Z.**, Tseng, G.C., Rui, H., Lee, A.V. and Butler, L.M., 2016. Abstract PD2-08: Potential role of prolactin signaling in development and growth of the lobular subtype of breast cancer. *Cancer Research*, 76(4 Supplement), pp.PD2-08.

[12] Enwright, John, Dominique Arion, John Corradi, Aiqing He, **Zhiguang Huo**, George Tseng, and David Lewis. (2015) Transcriptome Profiling of Layer 3 Parvalbumin Neurons from the Dorsolateral Prefrontal Cortex of Schizophrenia Subjects. *NEUROPSYCHOPHARMACOLOGY*, vol. 40, pp. S400-S401.

[13] John Enwright, Dominique Arion, **Zhiguang Huo**, George Tseng and David A. Lewis. Transcriptome alterations in layer 3 parvalbumin neurons in the dorsolateral prefrontal cortex in schizophrenia differ from those in layer 3 pyramidal cells. (in preparation).

[14] **Zhiguang Huo**, Shaowu Tang, YongSeok Park and George Tseng. Biomarker categorization and fast computing of adaptively weighted Fisher's method for meta-analysis in omics applications. (in preparation).