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 genomic data are available in public domain and these datasets provide unprecedented opportunities to reveal disease 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. In terms of statistical methodology, **I am particularly interested in modeling and optimization for high-dimensional data, Bayesian methods, graphical models and statistical computing,** which are capable of accommodating the high-dimensional nature of genomic data. In terms of genomics and bioinformatics application, I have collaborated with biologists in the fields of **cancer** and **psychiatry** to analyze a broad range of genomic data (e.g. microarray data to sequencing data), which motivates me to develop practical methodology and user-friendly softwares. However, genomics and bioinformatics are fast-evolving field and I am willing to expand my research towards new powerful data types (e.g. imaging data, **single cell data**). My long term goal is to bridge bioinformatics and statistics/machine learning and eliminate the gap between them.

Below is a highlight of my **past and going research, as well as future research plan**.

1 Statistical methodology

I am particularly interested in modeling and optimization for high-dimensional data, Bayesian methods, graphical models and statistical computing. Note that some of these areas can potentially overlap. I am not restricted to these areas I have explored. If I encounter other meaningful and challenging problems, I am definitely eager to learn or collaborate with other people.

1.1 Data integration

- **Past and on-going research:**

  Due to rapid development of high-throughput experimental techniques and dropping prices, many transcriptomic datasets have been generated and accumulated in the public domain (e.g. TCGA, GEO, SRA). Single cohort/data type may suffer from small sample size issue or reproducibility issue. A natural question is how to combine or integrate these complex data and increase statistical power, interpretation and reproducibility. This includes two direction: **horizontal meta-analysis** and **vertical integrative analysis**. Horizontal meta-analysis aims to combine genomic data of same type from multiple cohorts. I have worked on several meta-analytical methodologies including **disease subtype discovery**[1], **candidate marker detection**[2], **differential co-expression network detection**[4] and **dimensional reduction**[5]. Vertical integrative analysis combines multi-omics data (e.g. gene expression, CNV, genotyping, methylation, somatic mutation, miRNA) of the same cohort. I have developed a disease subtype discovery algorithm **integrating multi-level omics data with prior biological information**[3]. These methods will help better characterize the disease and deliver personalized medicine.

- **Future direction:**

  1. I have worked on horizontal meta-analysis and vertical integrative analysis respectively. A natural extension is **two-way integration** - combing horizontal meta-analysis and vertical integrative analysis.
  2. Newly innovated **epigenetic data** and **neuroimaging data** also allow people to understand diseases with new insight. Integrating these data types with genomic data is a potential future direction.

1.2 Modeling and optimization for high dimensional data

- **Past and on-going research:**

High-throughput data (e.g. genomic data) has more than 20,000 genes in human being but only tens or hundreds of samples. This large p and small n problem brings statistical challenges to reveal disease mechanism behind the big data. I am particularly interested in modeling high-dimensional data and solving related optimization problem with various forms of **regularizations**. I have used a **lasso penalty** on clustering problems to formalize a statistical objective and perform optimization to solve the problem[1]. In another clustering problem, which required incorporating prior knowledge, I proposed **sparse overlapping group lasso** and used **ADMM** to solve the challenging optimization problem[3]. By these techniques, we can discover the intrinsic information behind the high dimensional data.

- **Future direction:**

  1. I have worked 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.
  2. High dimensional problems are challenging in perspectives of both optimization and theory. I have worked on high-dimensional optimization problem quite a lot. **High-dimensional theory** is also interesting and challenging, which is essential for a sound methodology with theoretical guarantee.

1.3 Bayesian methods and graphical models

- **Past and on-going research:**

  Bayesian methods and graphical models are very convenient to model complex data and its dependent structure. I have worked on a **Bayesian non-parametric** approach to combine summary statistics from multiple cohorts to perform meta-analysis[2]. I am working on **Bayesian variable selection** problems[6] with multi-layer overlapping groups. Bayesian approaches and graphical models are also growing field themselves and I expect these technique will play an important role genomics and bioinformatics.

- **Future direction:**

  1. I have achieved success on high-dimensional clustering problem using frequentist approaches. I will tackle the same problem in Bayesian perspective.
  2. Part of my thesis is using **conditional random field** for a fast single cell imputation.

2 Bioinformatics application

As a biostatistician, an important job is to work with local biologists on their data. This is exciting for me not only I can help biologists towards innovating scientific findings, but also their data can motivate me to develop relevant statistical methodology.

- **Past and on-going research:**

  I have been mainly involved in **cancer** and **psychiatry** diseases research. For cancer research, I have worked on disease subtypes of breast cancer[11], DNA methylation of parity-induced mice[8], copy number variation of prostate cancer[9], fusion genes[7]. For psychiatry diseases, I have worked on schizophrenia, bipolar disorder and major depressive disorder in human pyramidal neurons[10] and parvalbumin neurons[12]. Currently, I am working on other aspects of in psychiatry (e.g. Induced pluripotent stem cell, sex related depression effect and circadian pattern).

- **Future direction:**

  1. Collaboration is always important part for statistician/biostatistician. I will definitely seek for opportunities to work with local biologists, which will in turn motivate statistical methodology development.
  2. I have worked extensively on microarray data and sequencing data. But I understand that this is a fast moving field. I am ready for any newly developed data types, as technique advances. At this stage, I am exposed to single cell data (single cell methylation and single cell gene expression) as part of my thesis.

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] 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. *Bioinformatics* (accepted).

[5] SungHwan Kim, Dongwan Kang, **Zhiguang Huo**, Yongseok Park, George C. Tseng. (2016) Meta-analytic principal component analysis. Submitted.

[6] Li Zhu, **Zhiguang Huo**, Tianzhou Ma and George Tseng. Bayesian indicator variable selection model with multi-layer overlapping groups. (in preparation).

[7] 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.

[8] 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 That Persists Long after Parturition. *Cancer Prevention Research* 8, no. 10 (2015): 1000-1009.

[9] 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.

[10] 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).

[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] 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).