Bo Sun

Data Scientist

Email: bos@usc.edu | Mobile: 213-618-6617 | **Personal Website:** https://bsun0802.github.io

Shiny App(a dash board): https://lianglabusc.shinyapps.io/shinyapp/

My Links

Gitbook: https://htmlpreview.github.io/?https://raw.githubusercontent.com/Yaphets-Bo/kidney-project/master/_book/index.html **Github:** https://github.com/bsun0802

Education

University of Southern California, Los Angeles, CA, USA

2016 – Present Ph.D in Quantitative and Computational Biology. (Expected Grad. 2019 Fall) GPA: 3.89/4.0

2018 – 2019 M.S. in Computer Science Data Science. (Graduation: 2019 spring) GPA: 3.87/4.0

Additional: 6 statistical & numerical classes taken from M.S of Applied Math.

2012 – 2016 B.S. in Bioinformatics, Tongji University, Shanghai, CHINA. GPA: 4.67/5.0

Skills

Key competency: Years of quantitative projects experience. **Advanced** in Python and R. Strong mathematics and statistics background. Solid understanding of **machine learning** algorithms and experienced in applying them. Experienced with database and **SQL** query.

Capability: Agile ad-hoc dataset exploratory analysis, when a quick diagnosis and solution is needed. In-depth data modeling analysis for larger scale problems when a justifiable data-driven argument is needed. Identify valuable questions given certain data, and able to implement and test it. Data integration and manipulation from different sources, tidy data, make appealing data visualization in ggplot2. Good oral and written communication skills.

Other Languages: MATLAB, C++, SQL, MySQL, Ruby, Ruby on Rails.

Clinical Data

Quantifying the progression from acute to chronic kidney injury. 2017 – 2018

- Collaborated with USC medical school on the **largest** clinical dataset of human kidney transplantation at that time, 42 kidney allografts over 4 time points, 163 transcriptoms (humen whole-genome sequencing data) in total.
- Developed a Linear Mixed Model which decompose the variance among two variables and can quantify the explanation power of them to the dependent variable. With this method, real injury-triggered gene expression changes were separated from noise(internal inter-individual variability, sex related genes, etc.)
- Created and deployed a R shiny dashboard to interactively visualize the gene expression changes over time. (https://lianglabusc.shinyapps.io/shinyapp/)

Researches

Imbalanced clustering of 3 x 10^3 cells and revealing subclasses. 2018 – Present

- Developed and trained a non-linear noise model to reduce data dimension from over 100,000 to 30,000 while kept most informative features. Speed was boosted 4 times in pair-wise cell whole genome comparisons, and 10^3 times overall.
- Designed an innovative distance metric based on least-square residuals which was robust in measuring similarity between cells.

Applying robust regression on heavy-tailed RNA-seq data. 2017 – Present

- Performed extensive simulations comparing performance of linear regression models when the Gaussian residuals assumption of least square failed. Tested on different residual distributions and noises and judged by ROC curves.
- On real data set, fitted a Beta-distribution estimation for original p-values, reduced the numbers of permutation needed down to below 10^2 while maintained low False Discovery Rate and high recall.

Profiling codon usage on alternative splicing sites. 2014 – 2016

- Led a team of three member, presented in Tongji University Innovation Wall event.
- Calculated codon usage at single-nucleotide resolution. And justify the significance of biased codon usage with hypothesis testing.

Award

China National Undergraduate Science and Technology Innovation project platform.

- 2016 Shanghai provincial outstanding project. (the last project in Projects section above).

Tongji University Academic Scholarship. - Three times, year 2013, 2014, 2015.

- Three times, year 2013, 2014 & 2015.

Publication Full list of publications could be found in my personal website