Bo Sun

Data Scientist

Email: bos@usc.edu | **Phone**: 213-618-6617 | **Website:** https://bsunfullsite.github.io/

Blog: https://bsun0802.github.io/ **Github:**https://github.com/bsun0802

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

Education

University of Southern California, Los Angeles, CA, USA

2016 – Present Ph.D in Quantitative and Computational Biology. (Expected Grad. 2020 Spring) 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.

Miscellaneous: Deep learning(PyTorch), Ruby, Ruby on Rails, SQL, MySQL, MATLAB, C++, SQL

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 gene variability. With this method, real injury-triggered gene expression changes were separated from noise(internal inter-individual variability, sex related genes, etc.) **(*manuscript under review)**
- Created and deployed a R *Shiny* dashboard web application to which could interactively visualize the gene expression changes over time.[link]
- Novelly applied a machine learning algorithm *Monocle* to this type of bulk analysis.

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 by basically decompose the residuals in linear model by gene effect and isoforms effect to estimate the similarity between cells.

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

- Performed extensive simulations to assess different linear regression models when the *Gaussian* residuals assumption failed. Tested on different type of residual distributions, outliers, and noises. Plotted ROC curves to compare the model performances.
- 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

- Calculated codon usage at single-nucleotide resolution. Assorted hypothesis testing was used to justify the significance of biased codon usage in different situations
- Led a team of three members, presented in Tongji University Innovation Wall event.

Publications

Two article manuscripts are submitted and under review. Working on another one article.