

# Nan Xiao

---

Genomic Data Scientist  
Seven Bridges Genomics  
Cambridge, Massachusetts

[me@nanx.me](mailto:me@nanx.me)  
<https://nanx.me>  
[GitHub](#) | [LinkedIn](#)

## Employment

2016 – Now Genomic Data Scientist. Seven Bridges Genomics, Inc. Cambridge, MA, USA

## Education

2015 – 2016 Ph.D. Student (Human Genetics). The University of Chicago. Chicago, IL, USA  
Advisor: Prof. Matthew Stephens

2012 – 2018 Ph.D. Candidate (Statistics). Central South University. Changsha, China  
Advisor: Prof. Qing-Song Xu

2008 – 2012 Bachelor of Science (Statistics). Central South University. Changsha, China

## Research Experience

2016 Matthew Stephens Lab Chicago, IL, USA  
Department of Human Genetics, The University of Chicago.  
Graduate student rotation

2016 Yoav Gilad Lab Chicago, IL, USA  
Department of Human Genetics, The University of Chicago.  
Graduate student rotation

2013 – 2015 Computational Biology and Drug Design Group Changsha, China  
Xiangya School of Medicine, Central South University.  
Graduate student research

## Publications

### Preprints

2016 [Nan Xiao](#), Q.-S. Xu, and M.-Z. Li (2016). hdnom: Building nomograms for penalized Cox models with high-dimensional survival data. *bioRxiv*. doi: [10.1101/065524](https://doi.org/10.1101/065524).

### Journal articles

2017 L.-L. Wang, Y.-W. Lin, X.-F. Wang, [Nan Xiao](#), Y.-D. Xu, H.-D. Li, Q.-S. Xu (2017). A selective review and comparison for interval variable selection in spectroscopic modeling. *Chemometrics and Intelligent Laboratory Systems*. doi: [10.1016/j.chemolab.2017.11.008](https://doi.org/10.1016/j.chemolab.2017.11.008).

- 2017 Y.-W. Lin, **Nan Xiao**, L.-L. Wang, C.-Q. Li, Q.-S. Xu (2017). Ordered homogeneity pursuit lasso for group variable selection with applications to spectroscopic data. *Chemometrics and Intelligent Laboratory Systems*. doi: [10.1016/j.chemolab.2017.07.004](https://doi.org/10.1016/j.chemolab.2017.07.004).
- 2016 L. Shen, D.-S. Cao, Q.-S. Xu, X. Huang, **Nan Xiao**, Y.-Z. Liang (2015). A novel local manifold-ranking based k-NN for modeling the regression between bioactivity and molecular descriptors. *Chemometrics and Intelligent Laboratory Systems*. doi: [10.1016/j.chemolab.2015.12.005](https://doi.org/10.1016/j.chemolab.2015.12.005).
- 2015 **Nan Xiao** and Q.-S. Xu (2015). Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection. *Journal of Statistical Computation and Simulation*. doi: [10.1080/00949655.2015.1016944](https://doi.org/10.1080/00949655.2015.1016944).
- 2015 **Nan Xiao**, D.-S. Cao, M.-F. Zhu, and Q.-S. Xu (2015). protr/ProtrWeb: R package and web server for generating various numerical representation schemes of protein sequence. *Bioinformatics*. doi: [10.1093/bioinformatics/btv042](https://doi.org/10.1093/bioinformatics/btv042).
- 2015 D.-S. Cao\*, **Nan Xiao**\*, Q.-S. Xu and A. F. Chen (2015). Rcp: R/Bioconductor package to generate various descriptors of proteins, compounds, and their interactions, *Bioinformatics*. \*Joint first authors. doi: [10.1093/bioinformatics/btu624](https://doi.org/10.1093/bioinformatics/btu624).
- 2015 D.-S. Cao, **Nan Xiao**, Y.-J. Li, W.-B. Zeng, Y.-Z. Liang, A.-P. Lu, Q.-S. Xu, A. F. Chen (2015). Integrating multiple evidence sources to predict adverse drug reactions based on systems pharmacology model. *CPT: Pharmacometrics & Systems Pharmacology*. doi: [10.1002/psp4.12002](https://doi.org/10.1002/psp4.12002).
- 2015 J.-B. Wang, D.-S. Cao, M.-F. Zhu, Y.-H. Yun, **Nan Xiao**, Y.-Z. Liang (2015). *In silico* evaluation of logD<sub>7.4</sub> and comparison with other prediction methods. *Journal of Chemometrics*. doi: [10.1002/cem.2718](https://doi.org/10.1002/cem.2718).

### **Book translations**

- 2016 Max Kuhn and Kjell Johnson (2016). *Applied Predictive Modeling*. (Hui Lin, Yi-Xuan Qiu, En-Chi Ma, **Nan Xiao**, & Vivian Zhang, Trans.). China Machine Press (Original work published in 2013). ISBN: [978-7-1115-3342-9](https://www.isbn-international.org/product/978-7-1115-3342-9).
- 2014 Winston Chang (2014). *R Graphics Cookbook*. (**Nan Xiao**, Yi-Shuo Deng, Tai-Yun Wei, & Yi-Xuan Qiu, Trans.). Posts and Telecom Press (Original work published in 2013). ISBN: [978-7-115-34227-0](https://www.isbn-international.org/product/978-7-115-34227-0).
- 2013 Hadley Wickham (2013). *ggplot2: Elegant Graphics for Data Analysis*. (Tai-Yun Wei, Yi-Xuan Qiu, **Nan Xiao**, Tao Gao, & Wei-Cheng Zhu, Trans.). Xi'an Jiaotong University Press (Original work published in 2010). ISBN: [978-7-5605-4969-9](https://www.isbn-international.org/product/978-7-5605-4969-9).
- 2013 Robert Kabacoff (2013). *R in Action: Data Analysis and Graphics with R*. (Tao Gao, **Nan Xiao**, & Gang Chen, Trans.). Posts and Telecom Press (Original work published in 2011). ISBN: [978-7-115-29990-1](https://www.isbn-international.org/product/978-7-115-29990-1).

## R Packages Authored

### [sevenbridges-r](#)

2016 – Seven Bridges API client, CWL schema, metadata schema, and SDK helper in R.

### [msaenet](#)

2016 – Multi-step adaptive elastic-net algorithm for high-dimensional feature selection. Integrated by Max Kuhn's *caret* package for streamlined machine learning modeling.

### [ggsci](#)

2016 – Scientific journal and sci-fi themed color palettes for ggplot2. Downloaded 10k+/month. Top 2% of 11,000+ R packages on CRAN.

### [liftr](#)

2015 – Containerize R Markdown documents with Docker. DockerCon 2017 [talk invited](#) by Docker, Inc.

### [enpls](#)

2014 – Ensemble partial least squares algorithm for feature screening and outlier detection. Integrated by Max Kuhn's *caret* package for streamlined machine learning modeling.

### [OHPL](#)

2017 – Ordered homogeneity pursuit lasso algorithm for group feature selection.

### [hdnom](#)

2015 – Benchmarking and visualization toolkit for high-dimensional survival modeling.

### [protr](#)

2012 – Efficient protein sequence feature extraction for machine learning modeling.

### [Rcpi](#)

2013 – Integrative molecular feature extraction for computational drug discovery.

### [RECA](#)

2012 – Relevant component analysis algorithm for supervised distance metric learning.

### [grex](#)

2016 – Gene ID mapping for Genotype-Tissue Expression (GTEx) data.

## R Packages Contributed

### [mxnet-r](#)

2015 – [Contributor](#) to the R binding for Amazon-backed deep learning framework MXNet.

### [ashr](#)

2015 – Empirical Bayes approach for large-scale hypothesis testing and FDR estimation.

### [dml](#)

2015 – Distance metric learning toolkit for dimensionality reduction in computer vision.

## Web Applications Authored

### [DockFlow](#)

2017 – Bioconductor workflow containerization and orchestration using Docker and liftr.

### [hdnom.io](#)

2015 – Shiny app for benchmarking and visualizing high-dimensional survival models. Selected as RStudio [Shiny user showcase](#) app.

### [ImgSVD](#)

2014 – Shiny app for image compression via singular value decomposition. Joint work with *Yihui Xie*, *Yixuan Qiu*, and *Tong He*.

### [TargetNet](#)

2014 – Shiny app for drug target identification by learning from binding affinities data.

### [ProtrWeb](#)

2013 – Shiny app for efficient protein sequence feature extraction.

### [Signify](#)

2015 – Shiny app for making your ( $>0.05$ )  $p$ -values sound significant.

## Talks

2017 – Reproducible Dynamic Report Generation with Docker and R  
Invited talk. DockerCon 2017, Austin, TX. April 2017.

2017 – Persistent Reproducible Reporting with Docker and R  
Invited talk. The 10<sup>th</sup> China R Conference, Tsinghua University, China. May 2017.

2016 – hdnom.io: High-Dimensional Survival Modeling with Shiny  
Invited talk. RStudio Shiny Developer Conference, Stanford University. January 2016.

2015 – Introduction to Reproducible Research in Bioinformatics  
Invited talk. CRI Annual Bioinformatics Workshop, Center for Research Informatics, The University of Chicago. December 2015.

2015 – liftr & sbgr kickstart  
Invited workshop. Joint with Dan Tenenbaum and Tengfei Yin.  
BioC 2015, Fred Hutchinson Cancer Research Center, Seattle, WA. July 2015.

2014 – Integrated Pipeline for Systems Pharmacology in R/Bioconductor  
Invited talk. The 7<sup>th</sup> China R Conference, Renmin University of China. May 2014.

2013 – Web Scraping with R  
Invited talk. The 6<sup>th</sup> China R Conference, Renmin University of China. May 2013.

## **Posters**

- 2017 DockFlow: Bioconductor Workflow Containerization and Orchestration with liftr  
**Nan Xiao**, Tengfei Yin, and Miao Zhu Li.  
BioC 2017, Dana-Farber Cancer Institute, Boston, MA. July 2017.
- 2017 The Deep Connection between Drugs and Side Effects  
ISCB Art in Science Competition. ISMB/ECCB 2017, Prague, Czech Republic. July 2017.
- 2015 liftr: Reproducible Bioinformatics and Statistical Data Analysis with Docker, Rabix, and knitr  
**Nan Xiao**, Tengfei Yin, and Miao Zhu Li.  
BioC 2015, Fred Hutchinson Cancer Research Center, Seattle, WA. July 2015.

## **Journal Referee**

*Journal of Statistical Computation and Simulation*  
*Chemometrics and Intelligent Laboratory Systems*  
*Genetic Epidemiology*  
*The R Journal*  
*Bioinformatics*

## **Services**

- 2017 Session Chair on Software Tools. The 10<sup>th</sup> China R Conference (Beijing)  
2014 Organizing Committee Member. The 7<sup>th</sup> China R Conference (Beijing)  
2013 Co-Organizer. The 6<sup>th</sup> China R Conference (Beijing)

## **Honors and awards**

- 2013 National scholarship for graduate students (5%).  
Highest award the government established for graduate students in China.
- 2011 1<sup>st</sup> Prize (top 5%) in Hunan Contest District in China Undergraduate Mathematical Contest in Modeling (CUMCM). With *Guan-Chun Liu* and *Tian-Yu Zhao*.

Last revision: December 2017.