Xuewei Cao

Center for Statistical Genetics, The Gertrude H. Sergievsky Center, Columbia University, New York, NY 10032
Email : xc2270@cumc.columbia.edu https://xueweic.github.io

PROFESSIONAL APPOINTMENT

Postdoctoral Research Scientist May 2023 - present
Center for Statistical Genetics, The Gertrude H. Sergievsky Center, Columbia University, NY, USA
Research direction: Statistical functional genomics methodology, molecular quantitative trait loci analysis

EDUCATION

Doctor of Philosophy in Statistics August 2018 - April 2023
Department of Mathematical Science, Michigan Technological University, MI, USA
Research direction: Statistical genetics methodology, genome-wide association studies

Master of Science in System Theory August 2015 - June 2018
School of Systems Science, Beijing Normal University, Beijing, China
Research direction: Applied complex network analysis

Bachelor of Science in Statistics August 2011 - June 2015
Department of Mathematical Science, Heilongjiang University, Harbin, China

RESEARCH INTERESTS

Statistical and functional genomics, Biological network analysis, Bayesian modeling, Machine learning.

RESEARCH PUBLICATIONS

In Revision and Submitted

1. **Cao, X.**, Zhang, S. & Sha, Q. (2023+) A novel method for multiple phenotype association studies based on genotype and phenotype network. *In revision at PLoS Genetics*. https://doi.org/10.1101/2023.02.23.529687
2. **Cao, X.**†, Zhu, L.†, Liang, X., Zhang, S. & Sha, Q. (2023+). Constructing genotype and phenotype network helps reveal disease heritability and phenome-wide association studies. *Submitted in October 2023*. https://doi.org/10.1101/2023.11.14.23297400
3. Zhao, C., Liu, A., Zhang, X., **Cao, X.**, Ding, Z., Sha, Q., Shen, H., Deng, H.W. & Zhou, W. CLCLSA: Cross-omics Linked embedding with Contrastive Learning and Self Attention for multi-omics integration with incomplete multi-omics data. (2023+). *Submitted in June 2023*. https://doi.org/10.48550/arXiv.2304.05542
4. Zhao, C.†, Su, K.†, Wu, C., **Cao, X.**, Sha, Q., Li, W., Luo, Z., Qin, T., Qiu, C., Liu, A., Jiang, L., Zhang, X., Shen, H., Deng, H.W. & Zhou, W. (2023+). Multi-View Variational Autoencoder for Missing Value Imputation in Untargeted Metabolomics. *Submitted in October 2023*. https://pubmed.ncbi.nlm.nih.gov/37873011/
5. Qi, Z., Pelletier, A., Willwerscheid, J. **Cao, X.**, Wen, X., Cruchaga, C., De Jager, P.L., TCW, J. & Wang, G., (2023+). Novel Missing Data Imputation Approaches Enhance Quantitative Trait Loci Discovery in Multi-Omics Analysis. *Submitted in November 2023*. https://doi.org/10.1101/2023.11.29.23299181

Published

1. **Cao, X.**†, Keyak, J.H.†, Sigurdsson, S., Zhao, C., Zhou, W., Liu, A., Lang, T.F., Deng, H.W., Gudnason, V. & Sha, Q. (2024). A New Hip Fracture Risk Index Derived from FEA-Computed Proximal Femur Fracture Loads and Energies-to-Failure. *Osteoporosis International*. Accepted. https://doi.org/10.48550/arXiv.2210.01032
2. **Cao, X.**†, Zhang, L.†, Islam, M.K., Zhao, M., He, C., Zhang, K., Liu, S., Sha, Q. & Wei, H., (2023). TGPred: efficient methods for predicting target genes of a transcription factor by integrating statistics, machine learning and optimization. *NAR Genomics and Bioinformatics*, 5(3), p.lqad083. https://doi.org/10.1093/nargab/lqad083
3. **Cao, X.**, Liang, X., Zhang, S. & Sha, Q. (2022). Gene selection by incorporating genetic networks into case-control association studies. *European Journal of Human Genetics*. https://doi.org/10.1038/s41431-022-01264-x
4. **Cao, X.**, Wang, X., Zhang, S. & Sha, Q. (2022). Gene-based association tests using GWAS summary statistics and incorporating eQTL. *Scientific Reports*. 12(1):3553. https://doi.org/10.1038/s41598-022-07465-0

† **denote co-first author.**

5. Boby, N.†, **Cao, X.†**, Williams, K., Gadila, S.K.G., Shroyer, M.N., Didier, P.J., Srivastav, S.K., Das, A., Baker, K., Sha, Q. & Pahar, B. (2022). Simian Immunodeficiency Virus Infection Mediated Changes in Jejunum and Peripheral SARS-CoV-2 Receptor ACE2 and Associated Proteins or Genes in Rhesus Macaques. *Frontiers in Immunology*. 13, p.835686. https://doi.org/10.3389/fimmu.2022.835686
6. Boby, N.†, **Cao, X.†**, Ransom, A., Pace, B.T., Mabee, C., Shroyer, M.N., Das, A., Didier, P.J., Srivastav, S.K., Porter, E., Sha, Q. & Pahar, B. (2021) Identification, Characterization, and Transcriptional Reprogramming of Epithelial Stem Cells and Intestinal Enteroids in Simian Immunodeficiency Virus Infected Rhesus Macaques. *Frontiers in Immunology*. 12, p.769990. https://doi.org/10.3389/fimmu.2021.769990
7. Zhu, L., Yan, S., **Cao, X.**, Sha, Q. & Zhang, S., (2024). Integrating External Controls by Regression Calibration for Genome-Wide Association Study. *Genes*. 15(1), 67. https://doi.org/10.3390/genes15010067
8. Xie, H., **Cao, X.**, Zhang, S. & Sha, Q. (2023). Joint analysis of multiple phenotypes for extremely unbalanced case-control association studies using multi-layer network. *Bioinformatics*. 39(12), btad707. https://doi.org/10.1093/bioinformatics/btad707
9. Wang, M., **Cao, X.**, Zhang, S. & Sha, Q. (2023). A clustering linear combination method for multiple phenotype association studies based on GWAS summary statistics. *Scientific Reports*. 13(1), p.3389. https://doi.org/10.1038/s41598-023-30415-3
10. Xie, H., **Cao, X.**, Zhang, S. & Sha, Q. (2023). Joint analysis of multiple phenotypes for extremely unbalanced case-control association studies. *Genetic Epidemiology*. 47(2), pp.185-197. https://doi.org/10.1002/gepi.22513
11. Liang, X., **Cao, X.**, Sha, Q. & Zhang, S. (2022). HCLC-FC: a Novel Statistical Method for Phenome-Wide Association Studies. *PLOS ONE*. 17(11): e0276646. https://doi.org/10.1371/journal.pone.0276646
12. Zhao, C.†, Keyak, J.H.†, **Cao, X.**, Sha, Q., Wu, L., Luo, Z., Zhao, L., Tian, Q., Qiu, C., Su, R., Shen, H., Deng, H.W. & Zhou, W. (2023). Multi-view information fusion using multi-view variational autoencoders to predict proximal femoral strength. *Frontiers in Endocrinology, section Bone Research*. https://doi.org/10.48550/arXiv.2210.00674

RESEARCH EXPERIENCE

Columbia University May. 2023 - present
Postdoctoral Research Scientist, with Prof. Gao Wang and Prof. Kushal K. Dey.

- Develop a new gradient boosting informed multi-omics colocalization method to improve the discovery of molecular quantitative trait loci for complex diseases.
- Develop an efficient genome-wide fine-mapping method integrating a large number of functional annotations based on the "sum of single effect" model.
- Develop a new method for identification of *trans* regulatory networks for Alzheimer's Disease from multi-tissue, multi-omics quantitative trait loci in human brains.

Michigan Technological University Aug. 2018 - Apr. 2023
Research Assistant, with Prof. Qiuying Sha.

- Develop novel statistical methods and efficient bioinformatical tools to address problems from genome-wide association studies and phenome-wide association studies by incorporating genotype and phenotype network.

Michigan Technological University Summer 2021
Research Intern, with Prof. Hairong Wei.

- Dap-Seq and R-loop genome data analysis using bowtie2, samtools, MACS2 etc;
- Reconstruction of gene regulatory network by incorporating the prior biological information.

Arizona State University Jan. 2018 - Feb. 2018
Research Visiting Scholar, the Complex Systems Program

- Workshops: Introduction to Complex Networks; The role of info processing in Sustainability; Analysis of Human Behavior using Social Networks.

SCIENTIFIC EXPERIENCE

- **Principal investigator**. Constructing genotype and phenotype network helps reveal phenome-wide association studies in biobanks. The UK Biobank Student Project (Application ID 102999) (April 2023 - April 2026).
- **Session organizer**. Invited session, entitled *Statistical Functional Genomics: advances and emerging challenges to unravel molecular effects of genetic variation at population scale*, at Conference on Statistics in Genomics and Genetics (STATGEN 2024) (May 2024)
- **Consortium member**. The Alzheimer's Disease Functional Genomics Consortium. (May, 2023 - present)
- **Consortium member**. Impact of Genomic Variation on Function (IGVF) Consortium. (Oct., 2023 - present)
- **Referee**. Frontiers in Genetics (Oct., 2023 - present)
- **Referee**. Genetic Epidemiology (Nov., 2023 - present)
- **Referee**. PLOS ONE (Jan., 2024 - present)

SOFTWARE

- **R/GPN**. A novel tool for genome-wide association studies by constructing a bipartite signed network, linking phenotypes and genotypes into a Genotype and Phenotype Network (GPN). https://github.com/xueweic/GPN
- **R/TGPred & Python/TGPred**. R and Python software including six efficient methods for predicting target genes of a transcription factor by integrating statistics, machine learning, and optimization. https://github.com/xueweic/TGPred
- **Python/CLCLSA**. A deep learning method for multi-omics integration with incomplete data by Cross-omics Linked unified embedding with Contrastive Learning and Self Attention (CLCLSA). https://github.com/xueweic/CLCLSA

CONFERENCES AND PRESENTATIONS

- American Society of Human Genetics 2023 (ASHG 2023), A new gradient boosting informed colocalization algorithm improves the identification of functional disease causal variants. **Poster**, Washington, DC, Nov.3 , 2023
- The Joint Statistical Meetings (JSM 2023). Constructing genotype and phenotype network helps reveal disease heritability and phenome-wide association studies. **Poster**, Toronto, Ontario, Canada, Aug. 9, 2023
- ICSA 2023 Applied Statistics Symposium (ICSA 2023). Constructing genotype and phenotype network helps reveal disease heritability and phenome-wide association studies. **Talk**, Ann Arbor, Michigan, Jun. 14, 2023
- Institute of Computing and Cybersystems (ICC) Joint Center for Biocomputing and Digital Health. Gene selection and genetic association test by incorporating the genetic networks. **Talk**, Hougthon, Michigan, Dec. 9, 2022
- American Society of Human Genetics 2022 (ASHG 2022), A novel method for multiple phenotype association studies based on genotype and phenotype network. **Poster**, Los Angeles, California, Oct. 25 - Oct. 29, 2022
- The American Society for Bone and Mineral Research Annual Meeting 2022 (ASBMR 2022), Hip Fracture Prediction using the First Principal Component Derived from FEA-Computed Fracture Loads. **Poster**, Austin, Texas, Sep. 9 - Sep. 12, 2022
- Upper Peninsula Medical Conference, The Impact of Medication Adherence on Health Care Costs for People with Diabetes. **Poster**, Houghton, Michigan, Aug. 26 - Aug. 28, 2022
- American Society of Human Genetics 2021 (ASHG 2021), Gene selection by incorporating genetic networks into case-control association studies. **Poster**, Virtual, Oct. 18 - Oct. 22, 2021
- International Genetic Epidemiology Society 2021 (IGES 2021), Gene-based association tests using GWAS summary statistics and incorporating eQTL. **Poster**, Virtual, Oct. 13 - Oct. 16, 2021
- IEEE 3rd International Conference on Big Data Analysis (ICBDA), The evolution of network topology structure of Chinese stock market. **Talk**, Shanghai, China, Mar. 9 - Mar. 12, 2018
- The 1st Chinese System Sciences Conference (CSSC2017), The structure of network topology of Chinese stock market. **Talk**, Beijing, China, May 13 - May 14, 2017

AWARDS AND HONORS

- Outstanding Scholarship Award — Spring 2023
- Doctoral Finishing Fellowship — Spring 2023

- Portage Health Foundation (PHF) Graduate Assistantship — Fall 2022
- Health Research Institute (HRI) Graduate Fellowship — Summer 2022
- Outstanding research award graduate student at Department of Mathematical Sciences — Spring 2022
- Outstanding academic achievement in Advanced Topics in Statistics (MA6700), Mathematical Statistics II (MA5712) — Spring 2019
- Outstanding academic achievement in Linear Algebra (MA4330), Linear Models (MA5731), Mathematical Statistics I (MA5711) — Fall 2018
- Outstanding graduates of Beijing City and Beijing Normal University — June 2018
- Outstanding graduates of Heilongjiang University — June 2015