
Xiuwen Zheng

CONTACT INFORMATION

Address: University of Washington, Department of Biostatistics
HSB F-646, Box 357232
Seattle, WA 98195-7232

Phone: 206-661-6949

E-mail: zhengx@u.washington.edu

Webpage: <http://students.washington.edu/zhengx/>

EDUCATION

- **PhD, Biostatistics,** Expected 8/12
Dept. of Biostatistics, University of Washington (UW), Seattle, WA
Dissertation: Analysis of Relatedness and Statistical Prediction of HLA Alleles in Genome-Wide Association Studies
Advisor: Dr. Bruce Weir, Chair and Professor
- **MS, Statistics,** 5/07
Dept. of Mathematical Sciences, University of Texas at Dallas (UTD), TX
- **BA, Finance and Statistics,** 7/05
Dept. of Statistics and Finance, University of Science and Technology of China (USTC), Hefei

RESEARCH INTERESTS

- Statistical genetics – such as research in methodology for relatedness, population structure and imputation, and applications in Genome-Wide Association Studies (GWAS)
- Machine learning, data mining
- Bioinformatics, high-dimensional data analysis
- Parallel / high-performance computing

REPRESENTATIVE PROJECTS

Gene Environment Association Studies project (GENEVA)

- GENEVA is a NIH-funded consortium of 16 genome-wide association studies from 12 universities and research institutes, which aims to accelerate understanding of genetic and environmental contributions to health and disease with thousands of samples and millions of SNPs
- Performed data cleaning and analysis on large-scale genotypic data, and involved in preparation of manuscripts

Human Leukocyte Antigen (HLA) prediction project

- Collaborated with GlaxoSmithKline (GSK) for a study of statistical prediction of HLA alleles
- Applied and developed machine learning algorithms (random forest and attribute bagging), and prepared manuscripts

CoreArray high-performance computing project

- Developed parallel computing algorithms using C/C++ for relatedness and principal component analysis in GWAS, and prepared manuscripts
- My algorithms achieve up to a 300-fold speedup over the original serial implementations

The electronic Medical Records and Genomics (eMERGE) network project

- The aim is to identify genetic variants associated white blood cell count differential leukocyte types in 13,923 subjects in the eMERGE network
- Performed data analysis and involved in preparation of manuscripts

SNP microarray project

- Mosaics for large chromosomal anomalies were detected using SNP microarray data from over 50,000 subjects of GENEVA
- Performed data analysis and involved in preparation of manuscripts

EXPERIENCE

- Research Assistant, Genetics Coordinating Center, Dept. of Biostatistics, UW 9/07 – 6/12
- Performed independent and collaborative research on major projects: GENEVA, HLA, CoreArray, eMERGE and SNP microarray
- Research Assistant, Bioinformatics Lab, Dept. of Computer Science, UTD 2/06 – 12/06
- Participated in Microarray Quality Control (MAQC) Project
 - Conducted microarray data cleaning and analysis, involved in preparation of manuscripts
- Teaching Assistant, Dept. of Biostatistics, UW 10/11 – 12/11
- Teaching Assistant, Dept. of Mathematical Sciences, UTD 8/05 – 5/07
- Courses: Applied Calculus, Probability and Statistics for Management and Economics, Medical Biometry (graduate course)
 - Led small discussion sections, held office hours and graded papers
- Statistical Consulting, Dept. of Biostatistics and Statistics, UW 9/10 – 12/10
- Primary projects:
- Power simulation in logistic regressions with confounders for oral cancer
 - Correlated data analysis on the effects of the King County menu labeling regulation

PUBLICATIONS

- David R. Crosslin, Andrew McDavid, Noah Weston, Sarah C. Nelson, **Xiuwen Zheng**, Eugene Hart, Mariza de Andrade, Iftikhar J. Kullo, Catherine A. McCarty, Kimberly F. Doheny, Elizabeth Pugh, Abel Kho, M. Geoffrey Hayes, Stephanie Pretel, Alexander Saip, Marylyn D. Ritchie, Dana C. Crawford, Paul K. Crane, Katherine Newton, Rongling Li, Daniel Mirel, Andrew Crenshaw, Eric B. Larson, Chris Carlson, Gail P. Jarvik, The electronic Medical Records and Genomics (eMERGE) Network. Genetic variants associated with the white blood cell count in 13,923 subjects in the eMERGE Network. 2011; accepted by *Human Genetics*.
- Laurie CC, Doheny KF, Mirel DB, Pugh EW, Bierut LJ, Bhangale T, Boehm F, Caporaso NE, Cornelis MC, Edenberg HJ, Gabriel SB, Harris EL, Hu FB, Jacobs KB, Kraft P, Landi MT, Lumley

T, Manolio TA, McHugh C, Painter I, Paschall J, Rice JP, Rice KM, **Zheng X**, Weir BS; GENEVA Investigators. Quality control and quality assurance in genotypic data for genome-wide association studies. *Genet Epidemiol*. 2010 Sep; 34(6): 591-602.

- Wenyuan Li, **Xiuwen Zheng**, and Ying Liu. Gene Selection by Matrix Reordering and Replicator Dynamics. *7th International Workshop on Data Mining in Bioinformatics (BIOKDD '07) in ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Jose, CA, USA, August 2007.
- **Xiuwen Zheng**, Hung-Chung Huang, Wenyuan Li, Peng Liu, Quan-Zhen Li, Ying Liu (2007) Modeling Nonlinearity in Dilution Design Microarray Data. *Bioinformatics*. 23(11): 1339-1347, June 2007.

Works In progress:

- Cathy C. Laurie, Cecelia A. Laurie, ..., **Xiuwen Zheng**, ..., Bruce Weir (73 coauthors in total). Somatic mosaicism for large chromosomal anomalies from birth to old age and its relationship to cancer (submitted to *Nature Genetics*, under review).
- **Xiuwen Zheng**, David Levine, Jess Shen, Stephanie M. Gogarten, Cathy Laurie, Bruce Weir. A High-performance Computing Toolset for Relatedness and Principal Component Analysis in GWAS. (in preparation for submission to *Bioinformatics*).
- **Xiuwen Zheng**, Judong Shen, Margaret G. Ehm, Matthew R. Nelson, Jon Wakefield, Bruce Weir. Attribute Bagging – A Statistical Method for Imputing HLA Types with SNP Data. (in preparation for submission to *American Journal of Human Genetics*).
- **Xiuwen Zheng**, Bruce Weir. Bias and Robustness of IBD Estimators with Miss-specified Allele Frequencies (in preparation).

PRESENTATIONS

- Poster, International Congress of Human Genetics (ICHG), Montreal, QC Canada, Oct 11 – 15, 2011; Title: “A High-Performance Computing Package for Relatedness and Principal Component Analysis in GWAS”. **Presenting author**.
- Talk, International Congress of Human Genetics (ICHG), Montreal, QC Canada, Oct 11 – 15, 2011; Title: “Somatic mosaicism of large chromosomal anomalies in blood cells of normal adults”. **Contributing author**.
- Poster, International Congress of Human Genetics (ICHG), Montreal, QC Canada, Oct 11 – 15, 2011; Title: “Genetic variation that predicts white blood cell count differential leukocyte types in the eMERGE Network”. **Contributing author**.
- Poster, 8th International Conference on Forensic Inference and Statistics, Seattle, WA, Jul 18 – 21, 2011; Title: “How Many SNPs Does It Take To Establish Relatedness?”. **Presenting author**.
- Poster, American Society of Human Genetics annual meeting (ASHG), Washington DC, Nov 2 – 6, 2010; Title: “Statistical Prediction of Classical HLA Typing Using Unphased SNP Data”. **Presenting author**.
- Poster, American Society of Human Genetics annual meeting (ASHG), Hawaii, Oct 20 – 24, 2009; Title: “Quality assurance of genotypic data for genome-wide association studies”; **Contributing author**.

PROFESSIONAL ACTIVITIES AND AFFILIATIONS

Referee for Genetics Research

Member of American Statistical Association, American Society of Human Genetics,
ENAR International Biometric Society

HONORS AND AWARDS

- 2007 – Present: Graduate Study Scholarship at Univ. of Washington – Seattle
- 2008: Department of Biostatistics Pfizer Award at Univ. of Washington – Seattle
- 2007: Department of Biostatistics Pfizer Award at Univ. of Washington – Seattle
- 2005 – 2007: Graduate Study Scholarship at Univ. of Texas – Dallas

SOFTWARE PROJECTS

CoreArray C/C++ library project, developed portable and scalable storage technologies for bioinformatic data, allowing parallel computing at the multicore and cluster levels. <http://corearray.sourceforge.net/>
Two R packages are available online (gdsfmt and SNPRelate) for high-performance computing on relatedness and principal component analysis in GWAS

COMPUTATIONAL SKILLS

- Programming Languages: Proficient in C/C++, Delphi (Object Pascal), working knowledge of JAVA, PERL and PYTHON
- Software: R, SAS, MATLAB, SQL and MPI