# Subha Guha

Department of Biostatistics
University of Florida

*Phone:* 352.294.5770
*E-mail:* s.guha@ufl.edu

## Research Interests

Bayesian modeling, statistical computing, nonparametric Bayesian methods, high-dimensional inference, cancer genomics, connectomics, neuroscience, independent component analysis, spatial statistics, hidden Markov models, survival analysis, MCMC simulation, generalized linear models, computational methods for Big Data

## Education

**Ohio State University**, Columbus, OH
Ph.D. in Statistics, June 2004

**Indian Institute of Technology**, Kanpur, India
M.Sc. in Statistics, June 1997

## Professional and Academic Experience

**University of Florida**, Department of Biostatistics, Gainesville, FL
*Associate Professor* **2018 - present**

**University of Missouri**, Department of Statistics, Columbia, MO
*Associate Professor* **2013 - 2017**
*Assistant Professor* **2007 - 2013**

**Harvard School of Public Health**, Department of Biostatistics, Boston, MA
*Postdoctoral Research Fellow* **2004 - 2007**

**Ohio State University**, Department of Statistics, Columbus, OH
*Graduate Teaching Assistant* **1999 - 2004**

## Research Grants

**National Science Foundation/National Institutes of Health**, *Collaborative Research: New Bayesian Nonparametric Paradigms of Personalized Medicine for Lung Cancer*, $**1,600,000**, 9/1/2015 to 8/31/2019.

This project aims to develop versatile and flexible nonparametric statistical techniques for identifying differential genomic signatures in cancer, invent integrative probabilistic frameworks for massive multiple-domain data, foster massively parallel algorithms and high-performance computational and inferential tools, and train the next generation of quantitative scientists to meet the challenges of interdisciplinary research involving high-dimensional data.

*Principal Investigators*: Subha Guha of University of Florida; Veera Baladandayuthapani and Ignacio Wistuba of M.D. Anderson Cancer Center. This body of work leverages the parallel computing expertise of Marc Suchard of UCLA.

## Research Grants, Continued

**Department of Health and Human Services**, *Statistical Informatics for Cancer Research*, one month's summer salary, 2009 to 2013 (Co-investigator).

**University of Missouri Research Council Committee**, $\$$**7,000**, Summer Research Fellowship 2008.

**National Science Foundation**, *Bayesian Mixture Models: Unified Theoretical Frameworks and MCMC Methods*, $\$$**150,000**, 7/1/2009 to 6/30/2013 (Principal Investigator).

## Honors and Awards

Top Faculty Achiever of the University of Missouri in February, 2016.

Faculty Fellowship Award of the Department of Statistics, University of Missouri in August, 2014.

Travel support awarded by organizers of the Seminar on Bayesian Inference in Econometrics and Statistics held in St. Louis, Missouri in 2005 and 2009; Valencia/ISBA World Meeting on Bayesian Statistics held in Alicante, Spain in 2006; 2nd IMS-ISBA Joint Meeting held in Bormio, Italy in 2005; Fourth International Workshop on Objective Prior Methodology held in Aussois, France in 2003.

Travel support awarded by Department of Biostatistics, Harvard School of Public Health, for attending ENAR Spring Meeting 2005, Joint Statistical Meetings 2005 and Joint Statistical Meetings 2006; ASA Section on Bayesian Statistical Science for attending Joint Statistical Meetings 2004; Graduate School and Department of Statistics, Ohio State University for attending Joint Statistical Meetings 2003.

Craig Cooley Memorial Prize for scholarly excellence and leadership qualities, 2004, Department of Statistics, Ohio State University.

Thomas and Jean Powers Teaching Award, 2003, Department of Statistics, Ohio State University.

Distinguished teaching assistant, Department of Statistics, Ohio State University, for four consecutive quarters.

Winner of the Student Paper Competition 2002 organized by ASA Statistical Computing and Graphics Sections.

## Academic Service

**Conference Organization**

Organized a session of invited talks at the conference titled, *Ordered Data Analysis, Models and Health Research Methods: An International Conference in Honor of H.N. Nagaraja for His 60th Birthday*, held at University of Texas at Dallas, March 2014.

## Academic Service (Continued)

**Grant Review**

National Institutes of Health, Biostatistical Methods and Research Design (BMRD) Study Section (temporary member, June 2017)

National Security Agency, Mathematical Sciences Grant Program

National Science Foundation (December 2015)

**American Statistical Association**

Chair, Contributed Session on *Modern Statistical Methods for Multi-Scale and Time Series Data*, Joint Statistical Meetings 2017

Committee Member, Student Paper Award of the Statistics in Genetics and Genomics Section, Joint Statistical Meetings 2016

Committee Member, Student Paper Award of the Section on Bayesian Statistical Science, Joint Statistical Meetings 2011 and Joint Statistical Meetings 2013

**International Society for Bayesian Analysis**

Member of Savage Award Committee, 2016

**University of Missouri**

| | |
|---|---|
| 2007–2017 | Member of 15 Ph.D. and 4 Master's Thesis Committees |
| 2016–2017 | Chair, Qualifying Exam Committee |
| 2016–2017 | Chair, Computing Committee |
| 2012–2013 | University of Missouri Research Board |
| 2011–2015 | Graduate Admissions Committee |
| 2011 | Chair, Bayesian Discussion Group |
| 2011–2014 | Faculty Search Committee |
| 2010–2012 | Advisor, Minor in Psychological Statistics and Methods |
| 2009–2014 | Member, Qualifying/Preliminary Exam Committee |
| 2008–2011 | Seminar Series Chair |
| 2008 | Co-organizer, Winemiller Conference |

## Professional Memberships

American Statistical Association
International Indian Statistical Association
International Society for Bayesian Analysis

## Journal Reviewer for

Australian and New Zealand Journal of Statistics, Bayesian Analysis, Bioinformatics, Biometrics, Biostatistics, Computational Statistics and Data Analysis, Journal of the American Statistical Association, Journal of Applied Statistics, Journal of Computational and Graphical Statistics, Journal of the Royal Statistical Society Series B, Journal of Statistical Computation and Simulation, Quality Technology and Quantitative Management, Statistical Methods and Applications, Statistica Sinica, Statistics in Medicine

## Publications

1. Jha, C., Li, Y. and **Guha, S.** (2017). Semiparametric Bayesian Analysis of High-Dimensional Censored Outcome Data. *Statistical Theory and Related Fields*, in press.

2. **Guha, S.** and Baladandayuthapani, V. (2016). A Nonparametric Bayesian Technique for High-Dimensional Regression. *Electronic Journal of Statistics*, 10, 3374–3424.

3. **Guha, S.**, Banerjee, S., Gu, C. and Baladandayuthapani, V. (2015). Nonparametric Variable Selection, Clustering and Prediction for Large Biological Datasets. *Nonparametric Bayesian Inference in Biostatistics* (eds. Mitra, R. and Müller, P.), Springer International Publishing.

4. Cui, S., **Guha, S.**, Ferreira, M. A. R. and Tegge, A. N. (2015). A Hidden Markov Model for Detecting Differentially Expressed Genes from RNA-Seq Data. *Annals of Applied Statistics*, 9, 901–925.

5. **Guha, S.**, Ji, Y., and Baladandayuthapani, V. (2014). Bayesian Disease Classification using Copy Number Data. *Cancer Informatics*, 13(S2), 83–91.

6. **Guha, S.** (2011). Discussion of *Sampling schemes for generalized linear Dirichlet process random effects models* by Kyung, Gill and Casella. *Statistical Methods & Applications*, 20, 291–293.

7. MacEachern, S. N. and **Guha, S.** (2010). Parametric and Semiparametric Hypotheses in the Linear Model. *The Canadian Journal of Statistics*, 39, 165–180.

8. **Guha, S.** (2010). Posterior Simulation in Countable Mixture Models for Large Datasets. *Journal of the American Statistical Association*, 105, 775–786.

9. **Guha, S.** (2010). Bayesian Hidden Markov Modeling of Array CGH Data. *Bayesian Modeling in Bioinformatics* (eds. Dey, D. K., Ghosh, S. and Mallick, B.), Chapman & Hall/CRC.

10. **Guha, S.**, Ryan, L. and Morara, M. (2009). Gauss-Seidel Estimation of Generalized Linear Mixed Models with Application to Poisson Modeling of Spatially Varying Disease Rates. *Journal of Computational and Graphical Statistics*, 18, 818–837.

11. **Guha, S.** (2008). Posterior Simulation in the Generalized Linear Mixed Model with Semiparametric Random Effects. *Journal of Computational and Graphical Statistics*, 17, 410–425.

12. **Guha, S.**, Li, Y. and Neuberg, D. (2008). Bayesian Hidden Markov Modeling of Array CGH Data. *Journal of the American Statistical Association*, 103, 485–497.

13. **Guha, S.** and MacEachern, S. N. (2006). Generalized Post-stratification and Importance Sampling for Subsampled Markov Chain Monte Carlo Estimation. *Journal of the American Statistical Association*, 101, 1175–1184.

14. Li, Y., Tiwari R., and **Guha, S.** (2006). Mixture Cure Survival Models with Dependent Censoring. *Journal of the Royal Statistical Society - Series B*, 69, 285–306.

15. Burden, S., **Guha, S.**, Morgan, G., Ryan, L. Sparks, G. and Young, L. (2005). Spatio-temporal Analysis of Ischemic Heart Disease in NSW, Australia. *Environmental and Ecological Statistics*, 12, 427–448.

16. **Guha, S.**, MacEachern, S. N. and Peruggia, M. (2004). Benchmark Estimation for Markov Chain Monte Carlo Samples. *Journal of Computational and Graphical Statistics*, 13, 683–701.

17. MacEachern, S. N., Peruggia, M. and **Guha, S.** (2003). Discussion of *A theory of statistical models for Monte Carlo integration* by Kong, McCullagh, Nicolae, Tan and Meng. *Journal of the Royal Statistical Society - Series B*, 65, 612.

## Work in Progress

1. **Guha, S.**, Jung, R. and Dunson, D. Predicting Phenotypes from Brain Connection Structure.

2. Yan, D., Baladandayuthapani, V. and **Guha, S.** Flexible Prediction and Clustering Models for Integrating Information in Multi-Domain Genomic Data.

3. Som, A., **Guha, S.** and Dunson, D. Bayesian Collapsed Regression for High-Dimensional Correlated Predictors.

4. **Guha, S.** and Ghosh, S. Bayesian Estimation of Conic Sections from Noisy Data.

5. Gu, C., Baladandayuthapani, V., Morris, J. and **Guha, S.** Bayesian Nonparametric Differential Analysis for Dependent Multigroup Data with Applications to Lung Cancer.

6. **Guha, S.** and Ghosh, S. A Nonparametric Bayesian Approach to Nonlinear Independent Component Analysis.

7. Gu, C., Baladandayuthapani, V. and **Guha, S.** Nonparametric Bayesian Comparison of Cancer Cell Lines and Tumor Samples by their Genomic Profiles.

8. Gu, C., Baladandayuthapani, V., **Guha, S.** and Suchard, M. Scalable Bayesian Nonparametric Learning for High-Dimensional Lung Cancer Genomics Data.

## Conference Abstracts

1. **Guha, S.** and Baladandayuthapani, V. (2017). A Nonparametric Bayesian Technique For High-Dimensional Regression. *NBER-NSF Seminar on Bayesian Inference in Econometrics and Statistics, Washington University in St. Louis, Missouri.*
2. Gu, C., **Guha, S.** and Baladandayuthapani, V. (2017). Scalable Bayesian Nonparametric Learning for High-Dimensional Lung Cancer Genomics Data. *ENAR Spring meeting held in Washington, DC.*
3. Jha, C., Li, Y. and **Guha, S.** (2016). Semiparametric Bayesian Analysis of High-Dimensional Censored Outcome Data: Discovering Spatial Variation of Breast Cancer Mortality Rates in New Mexico. *Joint Statistical Meetings held in Chicago, Illinois.*
4. Gu, C., **Guha, S.** and Baladandayuthapani, V. (2015). Scalable Bayesian Nonparametric Learning for High-Dimensional Lung Cancer Genomics Data. *International Chinese Statistical Association (ICSA) Applied Statistical Symposium held in Atlanta, GA.*
5. Gu, C., **Guha, S.** and Baladandayuthapani, V. (2015). Scalable Bayesian Nonparametric Learning for High-Dimensional Lung Cancer Genomics Data. *Joint Statistical Meetings held in Seattle, Washington.*
6. **Guha, S.** and Baladandayuthapani, V. (2015). Nonparametric Variable Selection, Clustering and Prediction for High-Dimensional Regression. *10th Conference on Bayesian Nonparametrics held in Raleigh, North Carolina.*
7. Gu, C., **Guha, S.** and Baladandayuthapani, V. (2015). Scalable Bayesian Nonparametric Learning for High-Dimensional Lung Cancer Genomics Data. *ENAR Spring Meeting held in Miami, Florida.*
8. **Guha, S.** and Baladandayuthapani, V. (2014). Variable selection in Bayesian high-dimensional regression of survival outcomes. *Joint Statistical Meetings held in Boston, Massachusetts.*

9. **Guha, S.** and Baladandayuthapani, V. (2014). Variable selection in Bayesian high-dimensional regression of survival outcomes. *International Conference in Honor of H.N. Nagaraja for His 60th Birthday.*
10. **Guha, S.** and Baladandayuthapani, V. (2013). Survival Predictor Detection by Dirichlet Processes. *Tenth Annual Conference on Frontiers in Applied and Computational Mathematics held at the New Jersey Institute of Technology in Newark, New Jersey.*
11. **Guha, S.** (2011). Posterior Simulation in Countable Mixture Models for Large Datasets. *Eighth Workshop on Bayesian Nonparametrics held at Veracruz, Mexico.*
12. **Guha, S.** (2010). Posterior Simulation in Countable Mixture Models for Large Datasets. *SAMSI Program on Semiparametric Bayesian Inference: Applications in Pharmacokinetics and Pharmacodynamics held at Research Triangle Park, North Carolina.*
13. **Guha, S.** (2010). Markov Chain Monte Carlo Methods for Mixture Models. *Ninth Valencia International Meeting on Bayesian Statistics / ISBA 2010 World Meeting held in Alicante, Spain.*
14. **Guha, S.**, Li Y. and Neuberg D. (2009). Bayesian Hidden Markov Modeling of Array CGH Data. *Seminar on Bayesian Inference in Econometrics and Statistics, held in St. Louis, Missouri.*
15. **Guha, S.** (2009). Markov Chain Monte Carlo Methods for Mixture Models. *Joint Statistical Meetings held in Washington D.C.*
16. **Guha, S.**, Li Y. and Neuberg D. (2008). Bayesian Hidden Markov Modeling of Array CGH Data. *Southern Regional Council on Statistics held in Charleston, South Carolina.*
17. **Guha, S.** (2008). Posterior Simulation in Countable Mixture Models for Large Datasets. *Joint Statistical Meetings held in Denver, Colorado.*
18. **Guha, S.**, Li Y. and Neuberg D. (2008). Bayesian Hidden Markov Modeling of Array CGH Data. *Conference on Statistical Paradigms: Recent Advances and Reconciliations, held in Kolkata, India.*
19. **Guha, S.**, Li Y. and Neuberg D. (2008). Bayesian Hidden Markov Modeling of Array CGH Data. *Session on Bayesian Computational Methods for Biomedical Applications, Joint Statistical Meetings, held in Seattle, Washington.*
20. **Guha, S.** and MacEachern, S. N. (2005). Generalized Post-stratification and Importance Sampling for Subsampled Markov Chain Monte Carlo Estimation. *Seminar on Bayesian Inference in Econometrics and Statistics, held in St. Louis, Missouri.*
21. **Guha, S.**, MacEachern, S. N. and Peruggia, M. (2002). Benchmark Estimation for Markov Chain Monte Carlo Samples. *Joint Statistical Meetings held in New York City.*

## Invited Academic Presentations

Seminar on Bayesian Inference in Econometrics and Statistics (SBIES) conference held in St. Louis, Missouri; Seminar talks at the Department of Statistics, North Carolina State University; Department of Statistics and Applied Mathematics, University of California at Santa Cruz; Statistical and Applied Mathematical Sciences Institute (SAMSI) in RTP, North Carolina, from October 2015 - present.

Joint Statistical Meetings held in Seattle, Washington; 10th Conference on Bayesian Nonparametrics, held in Raleigh, North Carolina; ENAR Spring Meeting held in Miami, Florida; Joint Statistical Meetings held in Boston, Massachusetts; Tenth Annual Conference on Frontiers in Applied and Computational Mathematics held at the New Jersey Institute of Technology in Newark, New Jersey; Eighth Workshop on Bayesian Nonparametrics held at Veracruz, Mexico; SAMSI Program on Semiparametric Bayesian Inference: Applications in Pharmacokinetics and Pharmacodynamics held at Research Triangle Park, North Carolina, from July 2010 - September 2015.

Environmental Statistics Retreat 2010 organized by the Department of Biostatistics, Harvard School of Public Health in Rockport, Massachusetts. Seminar talks at the Department of Mathematics and Statistics, University of Missouri - Kansas City; Division of Biostatistics, Washington University in St. Louis; Department of Biostatistics, Harvard School of Public Health; Department of Statistics, Western Michigan University; Department of Biostatistics, MD Anderson Cancer Center, from October 2009 - June 2010.

Seminar talks at the Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health; Department of Statistics, University of Virginia; Department of Statistics, Iowa State University; Department of Epidemiology & Biostatistics, University of Florida; Department of Biostatistics & Medical Informatics, University of Wisconsin-Madison; Department of Statistics, University of California, Davis; Department of Applied Mathematics and Statistics, University of California, Santa Cruz, from January - March 2007.

Seminar talks at the Department of Ambulatory Care and Prevention, Harvard Medical School; Department of Statistics, Temple University; Department of Statistics, University of California, Irvine; Department of Biostatistics, University of California, Los Angeles; Department of Epidemiology and Biostatistics, University of South Carolina; Department of Health Care Policy, Harvard Medical School; Statistics Department, University of Pennsylvania, from January - March 2007.

Racebrook Environmental Statistics Retreat, held in Sheffield, Massachusetts. Seminar talks at the Department of Statistics, Texas A&M University; Statistics Department, Harvard University; Department of Statistics, Texas A&M University; Department of Mathematical Sciences, University of Arkansas; Department of Statistics, Oklahoma State University, from January 2004 - November 2006.

## Software

2017 R package *NPCluster* implementing VariScan, a nonparametric Bayesian technique for clustering, variable selection, and prediction in high-throughput regression settings. The methodology is developed in Guha and Baladandayuthapani (EJS, 2016). Available at https://github.com/suchard-group/NPCluster.

We are currently working on implementing the package's MCMC procedure in a parallel computing framework using graphical processing units. Co-developed with Veera Baladandayuthapani, Chiyu Gu, and Marc Suchard.

2017 R package *BayesConics* implementing a Bayesian hierarchical approach for flexibly fitting different conic sections (hyperbola, parabola, ellipse, and circle) to noisy data and for estimating the conics parameters. The methodology is proposed in Guha and Ghosh (2017, work in progress). Co-developed with Sujit Ghosh.

2015 R package *hmmSeq* implementing the Bayesian technique for analyzing RNA-Seq data proposed in Cui, Guha, Ferreira and Tegge (AOAS, 2015). Joint work with the paper's co-authors.

2011 R package *glmmGS* for fitting generalized linear mixed models to massive datasets. Publicly available from CRAN at http://r-forge.r-project.org/projects/glmmgs. Co-developed with Michele Morara, Louise Ryan, and Christopher Paciorek.

2009 The Bayesian hidden Markov model strategy proposed by Guha, Li and Neuberg (JASA, 2008) for array CGH genomics data is implemented in Bioinformatics Toolbox 3.2. It is available at http://www.mathworks.com

2007 Created software for generating the sample paths of a number of common stochastic processes, verifying their theoretical properties by simulation and visualizing abstract results like the martingale central limit theorem. Used in the Department of Biostatistics, Harvard School of Public Health to teach the graduate level courses, *Analysis of Failure Time Data* and *Probability Theory and Applications I*.

## Teaching Experience at University of Missouri

**STAT 9710 – Mathematical Statistics I**. *Fall 2011–2014.*
Level: Graduate. Theory of estimation and tests of hypotheses including sufficiency, completeness and exponential families. Neyman-Pearson lemma, most powerful tests, similarity and invariance. Bayes and minimum variance unbiased estimates. Confidence intervals and ellipsoids.

**STAT 9720 – Mathematical Statistics II**. *Spring 2012–2017.*
Level: Graduate. Asymptotic distributions of maximum likelihood estimators, chi-square and likelihood ratio test statistics. EM algorithm, bootstrap, and introduction to generalized linear models.

**STAT 4710/7710 – Introduction to Mathematical Statistics**. *Spring 2008–2010, Spring 2012, Spring 2014, Spring 2017, Fall 2010, Fall 2013, Fall 2016, Fall 2017.*
Level: Graduate/Undergraduate. Introduction to theory of probability and statistics using concepts and methods of calculus.

**STAT 4510/7510 – Applied Statistical Models I**. *Spring 2011, Spring 2013, Spring 2015, Fall 2011–2012, Fall 2014, Fall 2016.*
Level: Graduate/Advanced Undergraduate. Introduction to applied linear models including regression (simple and multiple, subset selection, estimation and testing) and analysis of variance (fixed and random effects, multifactor models, contrasts, multiple testing).

**STAT 4750/7750 – Introduction to Probability Theory**. *Spring 2009–2011, Fall 2009–2010, Fall 2017.*
Level: Graduate/Advanced Undergraduate. Probability spaces; random variables and their distributions; repeated trials; probability limit theorems.

**STAT 4410/7410 – Biostatistics**. *Fall 2009.*
Level: Graduate/Advanced Undergraduate. Study of statistical techniques for the design and analysis of clinical trials, laboratory studies and epidemiology. Topics include randomization, power and sample size calculation, sequential monitoring, Carcinogenicity bioassay and case-cohort designs.

**STAT 4830/7830 – Categorical Data Analysis**. *Fall 2008.*
Level: Graduate/Advanced Undergraduate. Discrete distributions, frequency data, multinomial data, chi-square and likelihood ratio tests, logistic regression, log linear models, rates, relative risks, random effects, case studies.

## Other Teaching Experience

**Harvard School of Public Health**

BIO503 – Programming and Statistical Modeling in R (co-taught with Christopher Paciorek). *Spring 2007.*

BIO283 – Spatial Statistics for Health Research (co-taught with Louise Ryan, Yi Li, Christopher Paciorek). *Fall 2004.*

**Ohio State University**

STAT 428 – Introduction to Probability and Statistics for Engineering and the Sciences II. *Fall 2002, Winter 2003, Spring 2003.*

## Current and Past Ph.D. Students

Shiqi Cui. Dissertation topic: *Bayesian Mixture Models for High-Throughput Bioinformatics Applications.* Co-directed with Marco Ferreira. Graduated in December, 2014. Currently working at Google Inc.

Chiyu Gu. Dissertation topic: *Scalable Bayesian Nonparametric Learning for Biomedical Big Data.* Expected to graduate in June 2018.

Chetkar Jha. Dissertation topic: *Bayesian Nonparametric Analysis of Multivariate Unordered Categorical data.*

Dongyan Yan. Dissertation topic: *Scalable Computational Algorithms and Massively Parallel Computing for Bayesian Mixture Models.*

He Yuan. Dissertation topic: *Fast Markov Chain Monte Carlo Algorithms for Big Data.*