(*as of 05/2019*)

Yuanyuan Li

Staff Scientist, Biostatistics and Computational Biology Branch
National Institute of Environmental Health Sciences / National Institutes of Health (NIEHS/NIH), Durham, NC
Email: yuanyuan.li@nih.gov
URL: https://yuanyuanli66.github.io/

Education:

2010	Ph.D., Electrical Engineering and Computer Science, University of Tennessee, Knoxville
2006	M.S., Electrical Engineering and Computer Science, UTK
2001	B.S., Computer Information Science, Minnesota State University, Mankato

Awards and Honors:

2017, 2014	NIH Fellows Award for Research Excellence (FARE)
2013	NIEHS Science Day Best Poster Presentation Award
2011-2014	NIH Intramural Research Training Award (IRTA) Postdoctoral Fellowship
2008	Upsilon Pi Epsilon (UPE) International Honor Society for the Computing Sciences
2007, 2004	Scholarly and Research Incentive Funds (SARIF) Scholarship, University of Tennessee, Knoxville
2006	Best Graduate Teaching Assistant Award, University of Tennessee, Knoxville
2001	Graduate with honors (cum laude), Minnesota State University, Mankato
1999-2001	International Student Endowment Fund Scholarship, Minnesota State University, Mankato
1998-2001	Dean's List, Minnesota State University, Mankato

Research Experience:

Staff Scientist, **Biostatistics & Computational Biology Branch**, NIH/NIEHS **10/2018 - present**

Method development and mining pan-cancer genomic data using machine learning (ML)

- Use bagging and boosting to identify biomarkers that associates with various clinical outcomes from The Cancer Genome Atlas (TCGA) RNA-seq and clinical data
- Develop in-house stochastic gradient boosted machine (GBM) packages using R and Java
- Extend logistic regression trees to handle pair-matched (case-control) data
- Develop RNA-seq pipelines to detect differential expressed genes from raw signals
- Key ML techniques include: GBMs with various loss functions (logloss, mean-squared-error and learning-to-rank), rank aggregation, permutation test, T-test and Wilcoxon rank-sum test with False Discovery Rate (FDR) correction, Kaplan–Meier estimator and Cox proportional-hazards model with left truncation

Collaboration: mining various -omics, diet, clinical chemistry/hormone data, and histopathology findings

- Use various supervised and unsupervised ML techniques to recognize patterns and identify features associated with outcomes of collaborators' interests
- Collaborators include: within BCBB, epidemiology branch, clinical research unit, signal transduction branch, genome integrity and structural biology, National Toxicology Program (NTP) and Centers for Disease Control and Prevention (CDC)
- Develop various supervised and unsupervised ML pipelines to detect biological signals
- Incorporate nested sampling in source data as weights into tree learning
- Apply developed T-KDE toolbox on ChIP-seq data (see below)
- Key ML techniques include: Classification and regression trees (CART), GBMs, boosted logistic regression tree stumps, cost-sensitive learning, various clustering analysis, topic models, Kernel Density Estimation (KDE), Principal Component Analysis (PCA), T-distributed Stochastic Neighbor Embedding (t-SNE), survival analysis, and various statistical hypothesis tests

Research Fellow, **Biostatistics & Computational Biology Branch**, NIH/NIEHS **05/2014 - 10/2018**

Pan-cancer classification using in-house Genetic Algorithm / K-nearest neighbors (GA/KNN)

- Developed a parallel version of GA/KNN using POSIX Threads (p-threads)
- Proposed to use Latent Semantic Indexing (LSI) to speedup searches in GA
- Proposed to use GBMs improve GA/KNN's performances
- Associated identified driver genes to their corresponding tumor types by various cluster analysis

IRTA Postdoctoral Fellow, **Biostatistics Branch**, NIH/NIEHS **05/2011 - 05/2014**

Method for analyzing genome-wide protein binding patterns from ChIP-seq data

- Proposed and implemented T-KDE toolbox, to identify the locations of constitutive protein binding sites using ChIP-seq data
- T-KDE, which combines a binary range tree with a kernel density estimator to quickly identify constitutive protein binding sites from multiple cell lines.
- T-KDE can also identify genomic “hot spots” where several different proteins bind and, conversely, cell-specific sites bound by a given protein

Identify functional relevance of CCCTC-binding factor (CTCF) protein

- Analyzed CTCF’s genomic distributions, transcriptional environment, and epigenomic environment

Postdoctoral Researcher, **Biomedical Engineering**, UTK **09/2010 - 05/2011**

Machine learning-based approach for immune system and drug design

- Developed ML approach to model immune system and drug interaction using Fuzzy-clustering combined with variable length Markov model implemented in forms of Probabilistic Suffix Tree (PST)

Immune-inspired computational model

- Developed immune-inspired game theory for irregular warfare

Plant-based sensor network for nanoparticles toxicity study

- Developed and submitted a NSF proposal
- Developed a plant-based sensor-network for characterizing, monitoring, and understanding the environmental impact of both naturally occurring and man-made nanoparticles

Graduate Research Assistant, Distributed Intelligence Laboratory, **EECS**, UTK **05/2004 - 08/2010**

Anomaly detection in unknown environments using wireless sensor networks (WSNs) and a mobile robot (Partly funded by Oak Ridge National Laboratory)

- Designed and implemented a variety of distributed machine learning algorithms on a hierarchical resource-constrained sensor network (MICA2 and MICA2dots)
- **Video demonstration**: (Physical robots) Sensor network detects abnormal situation, with mobile robot (Pioneer 3) responding to location of anomaly (2007)
- Proposed a novel multiple missing data imputation technique that uses KD-tree with Mahalanobis distance for WSNs
- Key ML techniques include: PSTs, KD-trees, Fuzzy Adaptive Resonance Theory (Fuzzy-ART) neural network, Lempel–Ziv–Welch (LZW) algorithm for compression, likelihood-ratio test, autoregressive model and R^2

Indoor wireless localization for mobile robots

- Designed and implemented wireless indoor positioning system to locate mobile robots using triangulation and fingerprinting

Graduate Research Assistant, Computer Science, MNSU **01/2002 - 08/2003**

Bluetooth network simulator (Individual study research project)

- Developed Bluetooth network simulation software that simulated the behavior of a Bluetooth PicoNet with 1 to 7 slaves by using JavaSimulation package (a Java package for process-based discrete event simulation)

Text-to-Speech Synthesis for Mandarin Chinese

- Researched text-to-speech synthesis for Mandarin Chinese

Professional Experience:

Midwest Wireless Corporation (now Alltel Corporation), Mankato, MN **05/2003 - 08/2003**

Software engineer co-op

- Developed a framework monitoring the SMSC (short messaging) server including short messages from phone-to-phone (NOKIA7160), phone-to-PC and PC-to-phone; and service messages push to the phone

SpeechGear Inc (U.S. Naval Research funded project), Northfield, MN **07/2002 - 05/2003**

Software engineer co-op

- Designed and developed multiple interfaces for voice-enabled dictionary running on PDAs (Windows CE) using Java and eMbedded Visual Basic/C++

DataPlanIT Consulting, Mankato, MN **01/2002 - 05/2002**

Software engineer (part-time)

- Web design and development for surrounding businesses in Mankato using Active Server Pages

J.D. Edwards Company (now Oracle Corporation), Denver, CO **06/2001 - 08/2001**

Software engineer intern

- Wrote testing cases and suites for the MetaData software using JUnite

Hairs Supply Company, Chicago, IL **09/2000 - 12/2000**

Software engineer part-time

- Designed and developed an e-commerce website that sells hair supplies using Active Server Pages

Visible Edge Company, Mankato, MN **06/2000 - 08/2000**

Software engineer intern

- Upgraded and debugged the Performance Look Up System (PLUS) for Minnesota high schools using Visual Basic

Teaching Experience:

Instructor, Winter/Summer Biostatistics and Bioinformatics Short Courses, NIEHS **12/2017 - present**

DNA microarray data analysis

- Overview of microarray technology, experimental design, data preprocessing, statistical hypothesis testing, clustering, and classification

Teaching Assistant, Electrical Engineering and Computer Science, UTK **08/2003 - 09/2010**

CS100: Introduction to Computer Science (for non-majors)

- Programming: HTML, JavaScript and basic algorithms

CS102: Introduction to Computer Science (for majors)

- Programming: C++

CS302: Fundamental Algorithms

CS365: Programming Languages and Systems

- Programming: Java, Python and Perl

CS530: Computer Systems Organization

CS594: Data Mining Practices and Principles

Research Facilitator, **The Oak Ridge Associated Universities (ORAU)**, TN **07/2008 - 08/2008**

- Mentored a team of high school students on how to solve challenging navigation problems using a Vex Robotics Kit; sponsored by Appalachian Regional Commission and Oak Ridge Associated Universities.

Teaching Assistant, Computer Science, MNSU **08/2001 - 2002**

CS100: Introduction to Computer and Computing

- Programming: Microsoft Office 2000, HTML and JavaScript

Publications:

Refereed journal papers

1. **Y. Li**, M. Li, I. Shats, J. M. Krahn, G. P. Flake, D. M. Umbach, X. Li, and L. Li, “Glypican 6 is a putative biomarker for metastatic progression of cutaneous melanoma”, *PLoS One*, to appear, 2019

2. T.T. Nguyen, S. A. Grimm, P. R. Bushel, J. Li, **Y. Li**, B.D. Bennett, D. C. Fargo, C. W. Anderson, L. Li, M. A. Resnick, and D. Menendez, "Revealing the human p53 universe", *Nucleic acids research*, 46 (16), 8153-8167, 2018
3. **Y. Li**, D. M. Umbach, and L. Li, "Putative genomic characteristics of BRAF V600K versus V600E cutaneous melanoma", *Melanoma Research*, 27 (6), 527-535, 2017
4. **Y. Li**, J. M. Krahn, N. Croutwater, K. Lee, D. M. Umbach, and L. Li, "A comprehensive genomic pan-cancer analysis using The Cancer Genome Atlas gene expression data", *BMC genomics*, 18 (1), 508, 2017 (Cited by 19, source: Google)
5. **Y. Li**, J. M. Krahn, G. P. Flake, D. M. Umbach and L. Li, "Toward predicting metastatic progression of melanoma based on gene expression data", *Pigment Cell & Melanoma Research*, 28 (4), 453-463, 2015
6. **Y. Li**, D. M. Umbach, and L. Li, "T-KDE: A method for analyzing genome-wide protein binding patterns from ChIP-seq data sets", *BMC Genomics*, 15 (1), 27, 2014
7. **Y. Li**, M. Thomason, and L. E. Parker, "Sequential anomaly detection using wireless sensor networks in unknown environments", *Human behavior understanding in networked sensing - Theory and Applications of Networks of Sensors*, 99-123, 2014
8. **Y. Li**, and L. E. Parker, "Nearest neighbor imputation using spatial-temporal correlations in wireless sensor networks", *Information fusion*, 15, 64-79, 2014 (Cited by 44, source: Google)
9. **Y. Li**, W. Huang, L. Niu, S. Covo, D. M. Umbach, and L. Li, "Characterization of constitutive CTCF/Cohesin loci: a possible role in establishing topological domains in mammalian genomes", *BMC Genomics*, 14 (1), 553, 2013 (Cited by 61, source: Google)
10. S. Lenaghan, **Y. Li** (co-first authors), H. Zhang, J. Burris, C. Stewart, L. E. Parker, and M. Zhang, "Monitoring the environmental impact of TiO2 nanoparticles using a Plant-based sensor-network", *IEEE Transactions on Nanotechnology*, 2 (2), 182-189, 2013
11. **Y. Li**, S. Lenaghan, and M. Zhang, "A data-driven predictive approach for drug delivery using machine learning techniques", *PLoS one*, 7(2): e31724, 2012

Refereed conference papers

1. **Y. Li**, S. Lenaghan, J. Burris, C. N. Stewart, L. E. Parker, and M. Zhang, "Detecting the environmental impact of nanoparticles using plant-based biosensors", *The* 11^{th} *IEEE Conference on Nanotechnology (IEEE-NANO)*, pages 48-52, doi:10.1109/NANO.2011.6144505, August, 2011
2. **Y. Li**, M. Thomason, and L. E. Parker, "Detecting time-related changes in wireless sensor networks using symbol compression and probabilistic suffix trees", *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pages 2946-2951, doi:10.1109/IROS.2010.5649660, October, 2010
3. **Y. Li**, and L. E. Parker, "Detecting and monitoring time-related abnormal events using a wireless sensor network and mobile robot", *IROS*, pages 3292-3298, doi:10.1109/IROS.2008.4651031, October, 2008 (Cited by 30, source: Google)
4. **Y. Li**, and L. E. Parker, "A spatial-temporal imputation technique for classification with missing data in a wireless sensor network", *IROS*, pages 3272-3279, doi:10.1109/IROS.2008.4650774, October, 2008 (Cited by 37, source: Google)
5. **Y. Li**, and L. E. Parker, "Intruder detection using a wireless sensor network with an intelligent mobile robot response", *IEEE Southeastcon*, pages 37-42, doi:10.1109/SECON.2008.4494250, April, 2008, (Cited by 56, source: Google)
6. **Y. Li**, and L. E. Parker, "Classification with missing data in a wireless sensor network", *IEEE Southeastcon*, pages 533-538, doi:10.1109/SECON.2008.4494352, April, 2008 (Cited by 23, source: Google)
7. **Y. Li**, and S. Case, "Text-to-Speech synthesis for Mandarin Chinese", *In Proceedings of the Midwest Instruction and Computing Symposium (MICS)*, April, 2003

Papers under review:

1. **Y. Li**, A. Bingham, D. M. Umbach, and L. Li, "Putative biomarkers for predicting tumor sample purity based on gene expression data", *Scientific Reports*, 2019
2. K. Kang, Q. Meng, I. Shats, D. M. Umbach, B. Papas, **Y. Li**, X. Li and L. Li, "A novel complete computational deconvolution method using RNA-seq data", *Bioinformatics*, 2019

3. C. M. Clinton, J. R. Bain, M. J. Muehlbauer, **Y. Li**, L. Li, S. K. O'Neal, B. L. Hughes, D. E. Cantonwine, T. F. McElrath, K. K. Ferguson, "Non-targeted urinary metabolomics in pregnancy and associations with fetal growth restriction", 2019

Papers in preparation:

1. **Y. Li**, D. M. Umbach, and L. Li "Classification of breast cancer and sub-classification of triple-negative breast cancer samples based on TCGA gene and protein expression data"
2. **Y. Li**, and L. Li "Gene expression profiles to predict melanoma's target distant organs"
3. M. Shi, L. Li, A. Wise, D.M. Umbach, J. Krahn, **Y. Li**, C. R. Weinberg, "GA-KNN algorithm for detecting epistasis effects in case-parents triads"

Dissertation:

1. "Anomaly detection in unknown environments using wireless sensor networks", Distributed Intelligence Laboratory, EECS, UTK, May 2010

Technical Reports:

1. **Y. Li**, "Indoor positioning using 802.11b for mobile robots", DiLab, EECS, UTK, December 2005

Posters:

1. **Y. Li**, D. M. Umbach, and L. Li, "Putative biomarkers for tumor sample purity prediction based on gene expression data" *American Association for Cancer Research (AACR) Annual Meeting*, March, 2019
2. C. M. Clinton, J. R. Bain, M. J. Muehlbauer, **Y. Li**, L. Li, S. K. O'Neal, B. L. Hughes, D. E. Cantonwine, T. F. McElrath, and K. K. Ferguson, "Urinary metabolimic profiles in pregnancy and feta growth restriction", *The Pregnancy Meeting*, February, 2019
3. K. Kang, Q. Meng, I. Shats, D. M. Umbach, M. Li, **Y. Li**, X. Li, and L. Li, "A novel computational complete deconvolution method using RNA-seq data", *The* 17^{th} *European Conference on Computational Biology (ECCB)*, September, 2018
4. **Y. Li**, A. Bingham, Q. Li, Y. Zhuang, D. M. Umbach and L. Li, "Using tumor sample gene expression data to infer tumor purity levels with stochastic gradient boosting machines", *AACR Annual Meeting*, March, 2018
5. Q. Xu, I. Shats, **Y. Li**, L. Li, and X. Li, "1HNF4A-mediated methionine metabolism confers sensitivity of human hepatocellular carcinoma to methionine restriction", *DIR Board of Scientific Counselors (BSC) Review*, NIEHS, July, 2018
6. **Y. Li**, A. Bingham, D. M. Umbach, and L. Li, "Using tumor sample gene expression to learn about tumor purity and the tumor microenvironment" *NIEHS Science Day*, 2017
7. A. Bingham, **Y. Li**, D. M. Umbach, and L. Li, "Using tumor sample gene expression data to learn about tumor purity levels and the tumor microenvironment" *NIEHS summer intern poster competition*, **Best Poster Presentation Award**, June, 2017
8. **Y. Li**, J. Krahn, N. Croutwater, K. Lee, D. M. Umbach, and L. Li, "A comprehensive genomic pan-cancer analysis using The Cancer Genome Atlas gene expression data" *DIR BSC Review*, NIEHS, November, 2016
9. D. M. Umbach, M. Shi, A. Wise, J. Krahn, **Y. Li**, C. R. Weinberg, and L. Li, "A stochastic search algorithm for finding multi-SNP effects using nuclear families", *Joint Statistical Meeting (JSM)*, July, 2016
10. N. Croutwater, L. Li, and, **Y. Li**, "A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data", *NIEHS summer intern poster competition*, June, 2016
11. **Y. Li**, D. M. Umbach, L. Li, "A comprehensive genomic pan-cancer analysis comparing males and females using The Cancer Genome Atlas gene expression data" *AACR Precision Medicine Series: Targeting the Vulnerabilities of Cancer Conference*, May, 2016
12. **Y. Li**, J. M. Krahn, G. P. Flake, D. M. Umbach, and L. Li, "Glypican 6 is a putative biomarker for metastatic progression of cutaneous melanoma", *NIEHS Science Day*, 2015
13. C. R. Weinberg, M. Shi, A. Wise, D. M. Umbach, J. Krahn, **Y. Li**, and L. Li, "A stochastic search algorithm for finding multi-SNP effects using nuclear families", *International Genetic Epidemiology Society Conference (IGES)*, October, 2015

14. **Y. Li**, J. M. Krahn, and L. Li, “Putative biomarkers indicative of metastatic progression of skin cutaneous melanoma”, *AACR Melanoma: From Biology to Target Conference*, 2014
15. A. Mateja, **Y. Li**, and L. Li, “Using T-KDE to discover novel loci that may be implicated in X-inactivation”, *NIEHS summer intern poster competition*, June, 2014
16. **Y. Li**, D. M. Umbach, and L. Li, “T-KDE: A method for analyzing genome-wide protein binding patterns from ChIP-seq data”, *NIEHS Science Day*, **Best Poster Presentation Award**, 2013
17. **Y. Li**, D. M. Umbach, and L. Li, “Analysis of genome-wide protein binding patterns using kernel density estimators”, *the Biology of Genomes Conference*, May, 2013
18. **Y. Li**, W. Huang, D. M. Umbach, S. Covo, and L. Li, “Constitutive CTCF/Cohesin loci in a transcriptionally complex environment”, *NIEHS Science Day*, 2012
19. **Y. Li**, J. Wu, S. C. Lenaghan, and M. Zhang, “An Immuno-Inspired Game Theoretic Computational Framework for Irregular Warfare”, *Naval Science & Technology Partnership Conference*, 2010

Conference and Research Presentations:

- “Learning about tumor microenvironment using tumor sample gene expression and purity data” *North Carolina Biotechnology Seminar Series (invited talk)*, RTP, NC, 2019
- “Tree learing for big omics data”, *NIEHS seminar*, 2018
- “Learning with eXtreme Gradient Boosting - a gradient boosting approach”, *BCBB seminar*, NIEHS, 2017
- “A comprehensive genomic pan-cancer classification using The Cancer Genome Atlas gene expression data”, *BCBB retreat*, NIEHS, 2016
- “Identifying constitutive binding sites using kernel approach”, *BCBB retreat*, NIEHS, 2012
- “Detecting the environmental impact of nanoparticles using plant-based biosensors”, *IEEE-NANO conference*, August, 2011
- “Study complex biological systems: network modeling and AI-based analysis”, *Guest lecture: Systems Biology and Complex System Theory*, BME, UTK, October, 2010
- “Detecting time-related changes in wireless sensor networks using symbol compression and probabilistic suffix trees”, *IROS conference*, October, 2010
- “Environment monitoring using Wireless Sensor Networks”, *Guest lecture: Artificial Intelligence*, EECS, UTK, November, 2009
- “Detecting and monitoring time-related abnormal events using a wireless sensor network and mobile robot”, *IROS conference*, October, 2008
- “A spatial-temporal imputation technique for classification with missing data in a wireless sensor network”, *IROS conference*, October, 2008
- “Intruder detection using a wireless sensor network with an intelligent mobile robot response”, *the IEEE Southeast conference*, April, 2008
- “Classification with missing data in a wireless sensor network”, *the IEEE Southeast conference*, April, 2008, Alabama, Huntsville, USA
- “Exploring the impact of mobility in wireless sensor network”, *Oak Ridge National Laboratory*, 2006
- “Text-to-Speech synthesis for Mandarin Chinese”, *MICS conference*, April, 2003

Technical Skills:

- **Programming languages:** Java, Javascript, C/C++, NesC, eMbeddedVB/C++, shell scripts, Visual Basic, JSP, ASP, SQL, PL/SQL, XML, and WML
- **Applications:** Player/Stage, Matlab, R, WEKA, Oracle, MySQL, Orion server, LaTeX, BibTeX, Microsoft Office, and other common productivity packages for Windows and Linux platforms
- **Operating Systems:** Microsoft Windows, Unix/Linux, Mac OS, and TinyOS

Services:

Paper reviewing:

- NIH internal paper review, 2012 - present
- PLoS One Journal, 2011 - present
- Program committee of Association for the Advancement of Artificial Intelligence (AAAI), 2018
- Program committee of International Joint Conference on Artificial Intelligence (IJCAI), 2016 - 2018

- Scientific Reports - Nature, 2017 - 2018
- International Journal of Wireless Information Networks (IJWI), 2013 - 2017
- Bioinformatics Journal, 2017
- Soft Computing Journal, 2017
- NIH FARES award, 2017
- IEEE Sensors Journal, 2014 - 2017
- Neurocomputing Journal, 2016
- IEEE Systems Journal, 2014 - 2016
- International Journal of Computer Systems Science and Engineering (IJCSSE), 2013
- Sensors (ISSN 1424-8220), 2010
- IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS), 2006 - 2010, 2018
- IEEE International Conference on Robotics and Automation (ICRA), 2006 - 2007
- IEEE Intelligent Systems, 2006
- Journals of Robotics and Autonomous Systems (JRAS), 2005

Mentoring:

- Special volunteer within the group, BCBB, NIEHS, 2016 - present
- Summer undergraduate and graduate students within the group, BCBB, NIEHS, 2014 - present
  - Summer student own the 2017 best graduate student presentation award

Robotics activities:

- Coach for the First LEGO League (FLL) robotics competition, 2015 - 2016, 2018 - present
  - Two groups own robot design awards (2015 and 2019)
- Software judge for the FLL robotics competition in Tennessee and North Carolina, 2004 - 2014
- Tour guide for the Tennessee Junior Science & Humanities Symposium, 2010
- Robotic demonstration for the National Science Foundation (NSF) campus tour, UTK, 2008
- Robotic demonstration for local middle school and high school students, UTK, 2005 - 2010
- Maintained operating systems, software and hardware for Pioneer autonomous robots and Crossbow sensors for the Distributed Intelligence Laboratory, UTK, 2005 - 2010

Others:

- Biostatistics branch liaison for the NIEHS Trainees Assembly (NTA) steering committee, 2012 - 2014
- Judge for Science Fair for Triad Math and Science Academy, North Carolina, 2014
- Volunteer for the United Way fund allocation committee, Tennessee, 2005 to 2008
- President of the Chinese Student and Scholar Association, Minnesota State University, 2001
- Director of Activities of the International Student Association, Minnesota State University, 2000
- Peer Leader for the Intentional Student Orientation, Minnesota State University, 2000
- Vice President of the Taiwan Student Association, Minnesota State University, 1999