NAISHA SHAH

NSHAH@JCVI.ORG
301-633-3562

SUMMARY

Accomplished data scientist with 10+ years experience in computer science and genomics. Research focus in integrative analyses of genomic, transcriptomic, immune cell frequency, phenotypic and clinical data. Extensive experience in analyzing large amounts of whole genome sequencing (>30X) data, variant annotation, and cross-disciplinary analysis of multi-modal data using machine learning algorithms. Experience in both industry and academia.

EDUCATION

Bioinformatics, Doctor of Philosophy

Nov 2007 – Dec 2011

University College Dublin, Ireland
Thesis: Genetics of Autism Spectrum Disorder: A Bioinformatic Perspective
Supervisors: Prof. Denis Shields, Dr. Sean Ennis & Dr. Louise Gallagher

Computing Science and Molecular Biology and Biochemistry, Bachelor of Science

June 2007

Simon Fraser University, B.C., Canada

SKILLS

Cloud comput. (AWS)
Databricks

GitHub

R
Python
Spark
C/C++
Java
SQL
AWK
BASH scripting

Data Analysis
Statistical Modeling
Machine Learning

Bioinfo Tools

Public data analyses (e.g.: GEO, ClinVar, dbGaP)

RESEARCH EXPERIENCE

Associate Professor

March 2019 – Present

J. Craig Venter Institute, La Jolla, USA

- Leading data analytics for the upcoming Human Health and Performance Center at the JCVI.

Research Data Scientist

Oct 2018 – March 2019

Amazon, USA

Resigned as I was unable to relocate due to family commitment.

Head of Data Science & Analytics, Health Nucleus

Jan 2018 – Oct 2018

Human Longevity Inc., San Diego, USA

- Led genome and phenome analysis of Health Nucleus clients with data from several data modalities including WGS, MRI, metabolome and clinical data.
- Created an integrative framework for analyses of phenotypic data collected from several clinical instruments such as DEXA, MRI, Echocardiogram and ECG.
- Created multi-modal integrated disease risk prediction models using machine learning on large, longitudinal datasets such as UKB, ADNI and MESA.

Bioinformatics Research Scientist

August 2015 – Dec 2017

Human Longevity Inc., San Diego, USA

- Created estimated absolute disease risk model for the age-related chronic diseases using epidemiology-based risk score (Global Burden of Disease) and predicted genetic risk score.

- Led efforts to rank, categorize and annotate pathogenic variants for automated clinical reporting of the genome sequence. This includes integrating data from several public and commercial databases, and identification of misclassified variant interpretations.
- Analyzed terabytes of genomic data from more than 10,000 individuals to better characterize the human genome and to understand sequencing limitations.

Post-Doctoral Visiting Fellow April 2012 – July 2015

John Tsang's Lab, NIAID, NIH, Maryland, USA

- Created a community-based web platform for reusing publicly available gene expression data by creating, annotating and analyzing data compendium across studies and technology platforms. https://omicc.niaid.nih.gov/. The tools is used by dozens of national and international researchers.
- Developed a cell sorting-free approach to cellular composition deconvolution. Machine-learning models were trained on cellular and blood transcriptomic data from a healthy cohort by leveraging inter-individual variation. These models were then applied to blood transcriptomic data from 100+ case-control groups to derive relative frequencies of immune cell subsets and to find subsets associated with diverse conditions including autoimmune diseases.
- Collaborated on several projects including
  - Analysis of protein-expression heterogeneity in immune cells and finding genetic markers associated with the heterogeneity
  - Analysis of memory B cells transcriptome in HIV-infected individuals
  - RNA-seq analysis of sorted CD5^{hi} and CD5^{lo} T cells in mice
  - RNA-seq analysis of samples from patients with and without Monocytopenia and mycobacterial infection
- Mentored two summer students in the lab.

Bioinformatics PhD Candidate Nov 2007 – Dec 2011

Denis Shields' Lab, University College Dublin, Ireland

- Involved with the Autism Genome Project consortium, a large-scale collaborative genetics research project that aimed to identify the genes contributing to Autism Spectrum Disorder (ASD).
- Thesis work included detection and analysis of Copy Number Variations (CNVs) in trio families with ASD, parental origin of *de novo* CNVs, maternal genetic effects, and pathway enrichment analysis of genotypic data from ASD cohort integrated with eQTL data from cerebellum brain tissue.
- Other collaborative work included homozygous haplotype sharing in autism and CNV detection and analysis in the following diseases: agenesis of the corpus callosum with midline lipoma, Ewing sarcoma, Rhabdomyosarcoma and Vesicoureteral reflux.
- Supervised a stage 2 medical student with research project on "Copy number variations in Autism". Student's presentation of the research project was published in Irish Journal of Medical Science, Volume 179 Supplement 6.

PhD Industrial Placement July 2010 – Sept 2010

Merck Research Lab, Boston, USA

- Developed a module that takes into account the linkage disequilibrium structure in detecting expression Quantitative Trait Locus (eQTL). This was applied to eQTL detection in gene expression and genotype data of Cerebellum brain tissue.
- Worked on a project that focused on identifying coherent sets of genes underlying cellular processes by integrative analysis of several expression and genotype datasets.

Bioinformatics Research Assistant Jan 2007 – Oct 2007

Hancock Lab, University of British Columbia, Canada

- Designed and developed a curator tool for InnateDB (http://www.innatedb.ca/), a database for genes, proteins, interactions and pathways involved in innate immunity.
- Assisted in development of the front-end of InnateDB.

PUBLICATIONS

1. Shomorony I, Cirulli E, Huang L, Napier L, Heister R,... Venter JC, Karow D, Kirkness EF, Shah N. **(2020)**. An unsupervised learning approach to identify novel signatures of health and disease from multimodal data. **Genome Medicine**, 12(1), 1-14
2. Shah N, Hou YCC, Yu HC, Sainger R, Dec E, Perkins B, Caskey CT, Venter JC, Telenti A. **(2018)**. Identification of misclassified ClinVar variants via disease population prevalence. **Am J Hum Genet.** 102(4):609-619.
3. Cirulli E, Guo L, Swisher CL, Shah N, Huang L, Napier LA, Kirkness EF, Spector TD, Caskey CT, Thorens B, Venter JC, Telenti A. **(2018).** Profound perturbation of the human metabolome by obesity. **Cell Metabolism.**
4. Perkins BA, Caskey CT, Brar P, Dec E, Karow DS, Kahn AM, Hou YC, Shah N, *et al.* **(2018)**. Precision medicine screening using whole-genome sequencing and advanced imaging to identify disease risk in adults. **Proc Natl Acad Sci.** 115(14):3686-3691.
5. di Iulio J, Bartha I, Wong EHM, Yu HC, Lavrenko V, Yang D, Jung I, Hicks MA, Shah N, *et al.* **(2018)**. The human noncoding genome defined by genetic diversity. **Nat Genetics.** 50(3):333-337.
6. Cohen IV, Cirulli ET, Mitchell MW, Jonsson TJ, Yu J, Shah N, Spector TD, Guo L, Venter JC, Telenti A. **(2018)**. Acetaminophen (Paracetamol) Use Modifies the Sulfation of Sex Hormones. **EBioMedicine.** 28:316-323.
7. Telenti A, Pierce LCT, Biggs WH, di Iulio J, Wong EHM, Fabani MM, Kirkness EF, Moustafa A, Shah N, *et al.* **(2016)**. Deep sequencing of 10,000 human genomes. **Proc Natl Acad Sci.** 113(42):11901-11906.
8. Shah N, Guo Y, Wendelsdroff KV, Lu Y, Sparks R, Tsang JS. **(2016).** A crowdsourcing approach for reusing and meta-analyzing gene expression data. **Nat. Biotech.** 34, 803–806
9. Lu Y, Biancotto A, Cheung F, Remmers E, Shah N, McCoy JP, Tsang JS. **(2016).** Systematic analysis of cell-to-cell expression variation of T lymphocytes in a human cohort identifies aging and genetic associations. **Immunity**. 45(5):1162-1175
10. Busby B, Dillman A, Simpson CL, Fingerman I, Yun S, Kristensen DM, Shah N, *et al.* Building genomic analysis pipelines in a hackathon setting with bioinformatician teams: DNA-seq, Epigenomics, Metagenomics and RNA-seq. (**Pre-print available on bioRxiv: doi: http://dx.doi.org/10.1101/018085)**
11. Wagschal A, Najafi-Shoushtari SH, Wang L, Goedeke L, Sinha S, deLemos AS, Black JC, Ramírez CM, Li Y, Tewhey R, Hatoum I, Shah N, et al. **(2015)**. Genome-wide identification of microRNAs regulating cholesterol and triglyceride homeostasis. **Nat Med.** 21(11):1290-7
12. Fitzpatrick DJ, Ryan CJ, Shah N, Greene D, Molony C, Shields DC **(2015).** Genome-wide epistatic expression quantitative trait loci discovery in four human tissues reveals the importance of local chromosomal interactions governing gene expression. **BMC Genomics**.16:109.
13. Kardava L, Moir S, Shah N, *et al.* **(2014**). Abnormal B cell memory subsets dominated HIV-specific responses in infected individuals. **J Clin Invest.** 124(7):3252-62
14. Conroy J, McGettigan PA, McCreary D, Shah N, *et al.* **(2014).** Towards the identification of a genetic basis for Landau-Kleffner syndrome. **Epilepsia.** 55(6):858-65
15. Lynn M, Shah N, Conroy J, Ennis S, Morris T, Betts D, O'Sullivan M. **(2014).** A study of alveolar rhabdomyosarcoma copy number alterations by single nucleotide polymorphism analysis. **Appl Immunohistochem Mol Morphol.** 22(3):213-21
16. Lynn M, Wang Y, Slater J, Shah N, *et al.* **(2013).** High-resolution genome-wide copy-number analyses identify localized copy-number alterations in Ewing sarcoma. **Diagn Mol Pathol.** 22(2):76-84.
17. Magalhaes TR, Casey JP, Conroy J, Regan R, Fitzpatrick DJ, Shah N, *et al.* (**2012**). HGDP and HapMap analysis by Ancestry Mapper reveals local and global population relationships. **PLoS One.** 7(11):e49438
18. Anney R, *et al.* (co-author: Shah N) **(2012**). Individual common variants exert weak effects on the risk for autism spectrum disorders. **Hum Mol Genet.** 21(21):4781-92
19. Casey JP, Magalhaes T, Conroy JM, Regan R, Shah N, *et al.* **(2011).** A novel approach of homozygous haplotype sharing identifies candidate genes in autism spectrum disorder. **Hum Genet.** 131(4):565-79

20. Anney R, *et al.* (co-author: Shah N) **(2011).** Gene-ontology enrichment analysis in two independent family-based samples highlights biologically plausible processes for autism spectrum disorders. **Eur J Hum Genet.** 19(10):1082-9.
21. Baker LB, Conroy J, Donoghue V, Mullarkey M, Shah N, Murphy N, Murphy J, Ennis S, Lynch SA **(2011).** Agenesis of the corpus callosum with midline lipoma associated with an Xp22.31-Xp22.12 deletion. **Clin Dysmorphol**. 20(1):21-5.
22. Anney R, *et al.* (co-author: Shah N) **(2010).** A genome-wide scan for common alleles affecting risk for autism. **Hum Mol Genet**. 19(20):4072-82.
23. Pinto D, *et al.* (co-author: Shah N) **(2010).** Functional impact of global rare copy number variation in autism spectrum disorders. **Nature**. 466(7304):368-72.
24. Lynn DJ, Winsor GL, Chan C, Richard N, Laird MR, Barsky A, Gardy JL, Roche FM, Chan TH, Shah N, *et al.* **(2008)**. InnateDB: facilitating systems-level analyses of the mammalian innate immune response. **Mol Syst Biol.** 4:218.

PRESENTATIONS

- Shah N, *et al.* (2018). Towards Personalized Health: Multimodal Risk Prediction Models for Dementia. **ASHG 2018. Reviewer's Choice.**
- Shah N, *et al.* (2014). Integrative and Comparative Analyses of Blood Transcriptomic Signatures in Disease: A cell-centric view. **ISMB 2014.**
- Shah N, *et al.* (2011). The Search of Maternal Genetic Effects in Autism. **HGV 2011**.
- Shah N, *et al.* (2010). Genetics of Autism Spectrum Disorder. **Computational Biology and Innovation PhD Symposium 2010**.
- Shah N, *et al.* (2009). Parental Origin Bias in *De novo* CNVs Detected in Autism Probands. **ISHG 2009** (published in ***Ulster Med J*** 2010;79(1):33-42) & **HGV 2009**.
- Shah N, *et al.* (2009). Parental Origin of *De novo* CNVs & Autism Spectrum Disorder. **ASHG 2009**.
- Shah N, *et al.* (2008). Evaluating Illumina's Infinium Human 1M SNP Data and Existing CNV Prediction Algorithms. **ISHG 2008** & **HGV 2008**.

SCHOLARSHIPS AND AWARDS

- IRTA postdoctoral fellowship, National Institutes of Health, USA (2012)
- Supplemental computational fellowship, National Institutes of Health, USA (2012)
- Won second place ($3000), Health 2.0 Code-a-thon: Preventing obesity, Washington DC (2012)
- Best poster award, Human Genome Variation (HGV) conference, Toronto, Canada (2008)
- IRCSET postgraduate scholarship, University College Dublin, Ireland (2007)
- NSERC undergraduate student research award, Simon Fraser University, Canada (2005)
- Golden Key Honour's society member, Simon Fraser University, Canada (2004)
- Phi Theta Kappa scholarship, Simon Fraser University, Canada (2004)

TEACHING EXPERIENCE

Instructor

Jan 2015 – June 2015

Graduate School at the Foundation for Advanced Education in the Sciences, NIH, Maryland, USA

- Designed the curriculum and taught a graduate level course on using computer programming language Python for advanced bioinformatics analysis (BIOF309)

Demonstrator (Tutor)

Jan 2009 – Dec 2010

University College Dublin, Ireland

- Courses: COMP10060 - First Engineering Computer Science, COMP30100 - Principles of Programming Languages, and MDSA30240 - Bioinformatics.

OTHER EXPERIENCE AND COURSES

Course on Neural Networks and Deep Learning (by Andrew Ng) 2017

Coursera

Peer-reviewer *Ad-hoc*

- Reviewed articles for several journals including Bioinformatics, Nature Scientific Reports and BMC Biology.

Graduate Student Research Symposium Judge Jan 2015

NIH, Maryland, USA

- Judged graduate student's research poster presentation at the 11th annual NIH graduate student research symposium.

PhD Symposium Committee Member Dec 2010

University College Dublin, Ireland

- Organized the first-ever and highly successful PhD Symposium in Computational Biology and Innovation (http://bioinfo-casl.ucd.ie/phdsymposium/) where students got an opportunity to showcase their research to a diverse international audience. The event paved the way for new partnership and collaboration with academia and industry. The success of this event led to its establishment as a key annual academic international event in the UCD Bioinformatics and Systems Biology PhD Programme.
- Selected speakers to invite based on range of international researchers.
- Judged and awarded prizes to speakers at the symposium.