BIOGRAPHICAL SKETCH

NAME: Johnson, William Evan

eRA COMMONS USER NAME (credential, e.g., agency login): WEJOHNSON

POSITION TITLE: Professor of Medicine

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
Southern Utah University, Cedar City, UT	B.S.	05/2002	Mathematics
Brigham Young University, Provo, UT	M.S.	08/2003	Statistics
Harvard University, Cambridge, MA	M.A.	06/2006	Biostatistics
Harvard University, Cambridge, MA	Ph.D.	06/2007	Biostatistics

A. Personal Statement

My research team has a longstanding history of developing and applying commonly used methods and software for biomedical data analysis, particularly for applying genetics, genomics, and metagenomics data to solve problems in precision cancer therapy, infectious diseases/host response, and addiction research. Our work includes a balance between statistical methods development, algorithm optimization, and clinical application. Statistical innovation in our group focuses on the development of clinically motivated tools that integrate linear modeling, Bayesian methods, factor analysis and structural equations models, Hidden Markov models, mixture models, dynamic programming, and high-performance parallel computing. This work has resulted in widely used tools and algorithms for profiling transcription factors (MAT, MA2C), preprocessing and integrating of genomic data (ComBat, BatchQC, SCAN-UPC), aligning sequencing reads (GNUMAP), developing multi-gene biomarker signatures (ASSIGN), and metagenomic profiling (PathoScope). We have successfully applied our tools and methods in the context of pathogen detection, pathway profiling of host and microbial communities, and studying host/pathogen interactions in various contexts including human nutrition, food-borne pathogens, human respiratory diseases, vector-borne pathogen biosurveillance, and in chronic conditions such as cancer and obesity. I am currently the Founding Director of the Center for Data Science and scientific director bioinformatics core at Rutgers New Jersey Medical School. From a training perspective, our laboratory is committed to promote diversity and enhance the training of scientists from all over the globe. Trainees in our lab in the past few years have come from China, South Korea, India, Uganda, and the United States. Our research includes explorations in the Black Women's Health Study and international research in the microbiome and host transcriptomics in developing countries such as Brazil, Uganda, Zambia, and India. Diversity in our team and research agenda is a priority for our research group and objectives.

Ongoing and recently completed projects that I would like to highlight include:

R01 GM127430-04 (Johnson) 05/01/2018-04/30/2027
Removing batch effects in genomic and epigenomic studies

CRDF Global DAA3-19-65672-1 (Ellner) 10/01/2018-9/30/2023
RePORT India Phase II
Biomarkers for Risk of Tuberculosis and for Tuberculosis Treatment Failure and Relapse

1U19AI162598-01 (Alland, Salgame, Ellner) 07/01/2021- 06/30/2026
Bacterial and Host Heterogeneity in TB latency, persistence and progression

B. Positions Scientific Appointments, and Honors

Positions and Employment

2022-Present	Professor of Medicine, Division of Infectious Disease, Director, Center for Data Science, Rutgers New Jersey Medical School
2015-2022	Associate Professor and Associate Chief, Division of Computational Biomedicine, Department of Medicine, Boston University School of Medicine
2011-2015	Assistant Professor, Division of Computational Biomedicine, Department of Medicine, Boston University School of Medicine
2008-2017	Adjunct Assistant Professor, Department of Oncological Sciences, University of Utah
2007-2011	Assistant Professor, Department of Statistics, Brigham Young University
2004-2007	Research Assistant, Biostatistics and Comp. Biology, Dana Farber Cancer Inst.
2003-2007	Teaching Assistant, Department of Biostatistics, Harvard University
2002-2003	Teaching Assistant, Department of Statistics, Brigham Young University

Other Experience and Professional Memberships

2022-Present	American Association of Immunologists
2009-Present	International Society for Computational Biology
2006-Present	International Biometrics Society, ENAR/WNAR
2006-Present	Institute of Mathematical Statistics
2004-Present	American Statistical Association, Biometrics Section

Honors

2013	Outstanding Poster Award, Intelligent Systems for Molecular Biology, Berlin, Germany
2006	Prior Alumni Fellowship, Alpha Chi Honor Society
2006	Award for Achievement in Instructional Technology, Harvard University
2004-2006	Certificate of Distinction in Teaching, Harvard Biostatistics Department
2003-2007	NIH Pre-doctoral Traineeship in Cancer Research

C. Contribution to Science

The focus of our research is to develop computational and statistical tools to investigate core components that contribute to disease prognosis and etiology, and for the accurate determination of optimal diagnostic, prognostic, and therapeutic regimens for individual patients. Research in our group spans the fields of statistics, computer science, biology, and biomedicine in the following areas:

1. *Combining genomic data from multiple studies:* One of the most impactful contributions from our research stems from the development of statistical methods and software for combining genomic data from multiple studies, batches, or platforms. We developed our ComBat and ComBat-Seq software packages, which rely on parametric and non-parametric hierarchical Empirical Bayesian regression models and is a flexible and straightforward approach to remove technical artifacts due processing facility and data batch. ComBat has now been established as the most common approach for combining genomic data across experiments, labs, and platforms, and has been shown to be useful for data from a broad range of types and biological systems. We have also released and revised our BatchQC interface, that allows for interactive exploration and evaluation of batch effects. Our published manuscripts on this topic have been cited more than eight thousand times, and over the past year our software was downloaded more than 100,000 times (top 3% of all Bioconductor packages), indicating that it is actively used by the community. This work has been integrated in many academic and commercial software pipelines, and several researchers have developed modified versions of this approach to fit their own analysis needs. This is still an active area of research for our group; we have several manuscripts currently submitted or in preparation that build on this method.

    a. Johnson WE, Li C, Rabinovic A. Adjusting batch effects in microarray expression data using empirical Bayes methods. Biostatistics. 2007 Jan;8(1):118-27. PubMed PMID: 16632515.

b. Leek JT, Johnson WE, Parker HS, Jaffe AE, Storey JD. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. Bioinformatics. 2012 Mar 15;28(6):882-3. PubMed PMID: 22257669; PubMed Central PMCID: PMC3307112.
c. Manimaran S, Selby HM, Okrah K, Ruberman C, Leek JT, Quackenbush J, Haibe-Kains B, Bravo HC, Johnson WE. BatchQC: interactive software for evaluating sample and batch effects in genomic data. Bioinformatics. 2016 Aug 18. pii: btw538. [Epub ahead of print] PubMed PMID: 27540268.
d. Zhang Y, Parmigiani G, Johnson WE. ComBat-seq: batch effect adjustment for RNA-seq count data. NAR Genom Bioinform. 2020 Sep;2(3):lqaa078. doi: 10.1093/nargab/lqaa078. Epub 2020 Sep 21. PubMed PMID: 33015620; PubMed Central PMCID: PMC7518324.

2. *Preprocessing and standardizing genomic data:* An important facet our work is focused on the development or methods for the preprocessing and analysis of genome-wide profiling data, particularly for profiling protein occupancy, epigenetic marks, and global gene expression. A novel contribution of this work includes intrinsic statistical models for modeling feature background that can remove as much as 50% of the noise in the data. Our work also contributes novel methods for single sample normalization, and natural data harmonization across profiling technologies, including microarrays and sequencing platforms (including our UPC barcoding approach). Work in this area has had a substantial impact on my lab's research, laying the groundwork for methodological extensions and multiple collaborative manuscripts enabled by this method. Furthermore, our work served as a basis for highly impactful methods developed by other groups, and our software has facilitated data analysis for multiple recent high-profile manuscripts.

    a. Johnson WE, Li W, Meyer CA, Gottardo R, Carroll JS, Brown M, Liu XS. Model-based analysis of tiling-arrays for ChIP-chip. Proc Natl Acad Sci U S A. 2006 Aug 15;103(33):12457-62. PubMed PMID: 16895995; PubMed Central PMCID: PMC1567901.
    b. Piccolo SR, Sun Y, Campbell JD, Lenburg ME, Bild AH, Johnson WE. A single-sample microarray normalization method to facilitate personalized-medicine workflows. Genomics. 2012 Dec;100(6):337-44. PubMed PMID: 22959562; NIHMSID: NIHMS401888; PubMed Central PMCID: PMC3508193.
    c. Piccolo SR, Withers MR, Francis OE, Bild AH, Johnson WE. Multiplatform single-sample estimates of transcriptional activation. Proc Natl Acad Sci U S A. 2013 Oct 29;110(44):17778-83. PubMed PMID: 24128763; PubMed Central PMCID: PMC3816418.

3. *Methods and applications in precision medicine:* We are actively developing novel computational tools for processing and integrating individual genomic samples for use in personalized medicine settings. To do this, we have developed adaptive Bayesian models that integrate genome signatures pathway perturbations and drug screens with individual patient profiles to identify optimal therapeutic regimens or to reposition drugs. We have applied these methods in clinical and biomedical applications including breast cancer, pulmonary disease, genomic clinical trials, epigenetic drug efficacy and repositioning, and chemical carcinogenicity. We have published a software package through the R/Bioconductor project implementing this approach, which is currently being used extensively in our lab and in the labs of our colleagues.

    a. Cohen AL, Soldi R, Zhang H, Gustafson AM, Wilcox R, Welm BE, Chang JT, Johnson WE, Spira A, Jeffrey SS, Bild AH. A pharmacogenomic method for individualized prediction of drug sensitivity. Mol Syst Biol. 2011 Jul 19;7:513. PubMed PMID: 21772261; PubMed Central PMCID: PMC3159972.
    b. Cohen AL, Piccolo SR, Cheng L, Soldi R, Han B, Johnson WE, Bild AH. Genomic pathway analysis reveals that EZH2 and HDAC4 represent mutually exclusive epigenetic pathways across human cancers. BMC Med Genomics. 2013 Sep 30;6:35. PubMed PMID: 24079712; PubMed Central PMCID: PMC3850967.
    c. Shen Y, Rahman M, Piccolo SR, Gusenleitner D, El-Chaar NN, Cheng L, Monti S, Bild AH, Johnson WE. ASSIGN: context-specific genomic profiling of multiple heterogeneous biological pathways. Bioinformatics. 2015 Jan 22;PubMed PMID: 25617415.
    d. Piccolo SR, Andrulis IL, Cohen AL, Conner T, Moos PJ, Spira AE, Buys SS, Johnson WE, Bild AH. Gene-expression patterns in peripheral blood classify familial breast cancer susceptibility. BMC Med Genomics. 2015 Nov 4;8:72. doi: 10.1186/s12920-015-0145-6. PubMed PMID: 26538066; PubMed Central PMCID: PMC4634735.

4. *Biomarkers and models for TB diagnostics and immunology:* A recent focus in our group has been to evaluate and develop biomarkers for TB outcomes. We developed a 29-gene biomarker (PREDICT29), that can accurately identify which individuals with latent TB infection will progress to active disease, in some

cases 4-5 years prior to active disease. We have also developed our TBSignatureProfiler, an R-based platform for profiling more than 50 signatures of TB disease, risk, and response to treatment. In an extension of this work, we developed a 107-gene NanoString platform that can cost-effectively profile more than 15 TB signatures. We are also engaged in curating more than 100 gene expression studies, totaling more than 10,000 samples in our ongoing curatedTBData project. We are actively using these platforms, signatures, and data in multiple international studies and consortia and studies to profile clinical and molecular TB outcomes, especially in patients with co-morbid conditions such as malnutrition, HIV, and diabetes.

a. Leong S, Zhao Y, Joseph NM, Hochberg NS, Sarkar S, Pleskunas J, Hom D, Lakshminarayanan S, Horsburgh CR Jr, Roy G, Ellner JJ, Johnson WE, Salgame P. Existing blood transcriptional classifiers accurately discriminate active tuberculosis from latent infection in individuals from south India. Tuberculosis (Edinb). 2018 Mar;109:41-51. Epub 2018 Jan 31. PubMed PMID: 29559120.
b. Manabe YC, Andrade BB, Gupte N, Leong S, Kintali M, Matoga M, Riviere C, Sameneka W, Lama JR, Naidoo K, Zhao Y, Johnson WE, Ellner JJ, Hosseinipour MC, Bisson GP, Salgame P, Gupta A. A Parsimonious Host Inflammatory Biomarker Signature Predicts Incident TB and Mortality in Advanced HIV. Clin Infect Dis. 2019 Nov 25;. PubMed PMID: 31761933.
c. Leong S, Zhao Y, Ribeiro-Rodrigues R, Jones-López EC, Acuña-Villaorduña C, Rodrigues PM, Palaci M, Alland D, Dietze R, Ellner JJ, Johnson WE, Salgame P. Cross-validation of existing signatures and derivation of a novel 29-gene transcriptomic signature predictive of progression to TB in a Brazilian cohort of household contacts of pulmonary TB. Tuberculosis (Edinb). 2020 Jan;120:101898. Epub 2020 Jan 7. PubMed PMID: 32090859; PubMed Central PMCID: PMC7066850.
d. Johnson WE, Odom A, Cintron C, Muthaiah M, Knudsen S, Joseph N, Babu S, Lakshminarayanan S, Jenkins DF, Zhao Y, Nankya E, Horsburgh CR, Roy G, Ellner J, Sarkar S, Salgame P, Hochberg NS. Comparing tuberculosis gene signatures in malnourished individuals using the TBSignatureProfiler. BMC Infect Dis. 2021 Jan 22;21(1):106. doi: 10.1186/s12879-020-05598-z. PubMed PMID: 33482742; PubMed Central PMCID: PMC7821401.

5. *Statistical methods and applications for metagenomic profiling:* Recent efforts in our lab has involved the development of analytical tools and methods in burgeoning field of metagenomics. We developed PathoScope, which utilizes a Bayesian statistical framework for the rapid species/strain attribution of sequencing reads that can accurately discriminate between very closely related strains of the same species with very little coverage of the genome. In the past few years, we have published five methodological manuscripts and multiple collaborative manuscripts. PathoScope has been used by a number of different groups, most notably being used for data analysis in two recent publications in *Nature*. This approach also has commercial potential, as we know of at least one start-up company that is using this approach as a basis for their proprietary analysis software toolkit.
   a. Francis OE, Bendall M, Manimaran S, Hong C, Clement NL, Castro-Nallar E, Snell Q, Schaalje GB, Clement MJ, Crandall KA, Johnson WE. Pathoscope: species identification and strain attribution with unassembled sequencing data. Genome Res. 2013 Oct;23(10):1721-9. PubMed PMID: 23843222; PubMed Central PMCID: PMC3787268.
   b. Hong C, Manimaran S, Shen Y, Perez-Rogers JF, Byrd AL, Castro-Nallar E, Crandall KA, Johnson WE. PathoScope 2.0: a complete computational framework for strain identification in environmental or clinical sequencing samples. Microbiome. 2014;2:33. PubMed PMID: 25225611; PubMed Central PMCID: PMC4164323.
   c. Castro-Nallar E, Shen Y, Freishtat RJ, Pérez-Losada M, Manimaran S, Liu G, Johnson WE, Crandall KA. Integrating microbial and host transcriptomics to characterize asthma-associated microbial communities. BMC Med Genomics. 2015 Aug 16;8:50. doi: 10.1186/s12920-015-0121-1. PubMed PMID: 26277095; PubMed Central PMCID: PMC4537781.

Complete List of Published Work in MyBibliography:
http://www.ncbi.nlm.nih.gov/sites/myncbi/william.johnson.1/bibliography/41689720/public/?sort=date&direction=descending