CURRICULUM VITAE

Di Wu, PhD

2021

PERSONAL INFORMATION

Office:

4504 Koury Oral Health Sciences Building; CB #7455
The University of North Carolina at Chapel Hill
Chapel Hill, NC 27599
Phone: (919) 537-3277
Email: dwu@unc.edu

EDUCATION

Institution	Degree	Date Conferred	Major
University of Melbourne Institute of Medical Research Bioinformatics Division Australia	Ph.D.	Dec 2012	Statistical Bioinformatics
Case Western Reserve University School of Medicine Department of Epidemiology & Biostatistics Ohio	M.S.	May 2006	Biostatistics
Shanghai Jiao Tong University Department of Biotechnology China	B.S.	June 1998	Biotechnology

PROFESSIONAL EXPERIENCE

2021 – present	Associate Professor, Division of Oral and Craniofacial Health Sciences, Adams School of Dentistry, University of North Carolina at Chapel Hill, NC
2021 – present	Research Associate Professor, Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, NC

2019 – present	Full member, Cancer Genetics UNC Lineberger Comprehensive Cancer Center, University of North Carolina at Chapel Hill, NC
2016 – present	Core Member of Bioinformatics and computational Biology (BCB) in The Biological & Biomedical Sciences Program (BBSP) University of North Carolina at Chapel Hill, NC
2015 – present	Core Member, Carolina Health Informatics Program (CHIP), University of North Carolina at Chapel Hill, NC
2015 – 2021	Assistant Professor, Division of Oral and Craniofacial Health Sciences, Adams School of Dentistry, University of North Carolina at Chapel Hill, NC
2015 – 2021	Research Assistant Professor, Department of Biostatistics, Gillings School of Global Public Health, University of North Carolina at Chapel Hill, NC
2011 – 2015	Postdoctoral Fellow, Department of Statistics, Harvard University, Cambridge, MA
2011 – 2015	Postdoctoral Fellow, Biostatistics, Dana-Farber Cancer Institute, Boston, MA
2011 – 2015	Postdoctoral Fellow, Division of Genetics, Harvard Medical School and the Brigham and Women's Hospital, Boston, MA
2007 – 2011	Biostatistician (part-time), Center for Cancer Research, Monash Institute of Medical Research, Clayton, Victoria, Australia
2006 – 2007	Biostatistician, Center for Cancer Research, Monash Institute of Medical Research, Clayton, Victoria, Australia
2004 – 2006	Research Assistant III, Department of Pathology, Case Western Reserve University, Cleveland, OH
2003 – 2004	Research Assistant II, Case Comprehensive Cancer Center, Case Western Reserve University, Cleveland, OH
2001 – 2003	Research Assistant I & II, Department of Pathology, Case Western Reserve University, Cleveland, OH
1998 – 2001	Research Assistant, Shanghai Institute of Biochemistry,

Chinese Academy of Science, China

HONORS & AWARDS

2018	Junior Faculty Development Award, UNC-Chapel Hill
2018	Finalist, Joseph Lister Award for Young Investigators, American Association for Dental Research (AADR)
2010	Chinese Government Award for Outstanding Self-Financed Students Abroad
2010	Edith Moffat Overseas Travel Scholarship, Walter and Eliza Hall Institute, Australia
2007-2010	Australian Postgraduate Award, University of Melbourne, Australia
2009	Student Travel Bursary, AMATA Conference, Sydney, Australia
2007	Student Travel Bursary, MGED and AMATA Conference, Brisbane, Australia
2007	Travel Scholarships, ICE-EM Australian Graduate School in Mathematics
1995	Scholarship for Excellence, Shanghai Jiao Tong University, China

BIBLIOGRAPHY

Books & Book Chapters:

1. **Wu D**, Gantier MP. Normalization of Affymetrix miRNA Microarrays for the Analysis of Cancer Samples. *Microarray Data Analysis Methods and Applications* part of the *Methods in Molecular Biology Series* Ed. 2015; DOI:10.1007/7651_2015_239: 1-10.

2. Divaris K, Shungin D, Rodriguez-Cortes A, Basta PV, Roach J, Cho H, **Wu D**, Ferreira Zandona AG, Ginnis J, Ramamoorthy S, Kinchen JM, Kwintkiewicz J, Butz N, Ribeiro AA, Azcarate-Peril MA. The supragingival biofilm in early childhood caries: clinical and laboratory protocols and bioinformatics pipelines supporting metagenomics, metatranscriptomics, and metabolomics studies of the oral microbiome. *Methods Mol Biol*.2019;1992(40): 525-548

3. **Wu D**. Karhade D, Pillai M, Jiang M, Huang L, Li G, Cho H, Roach J, Li Y, Divaris K. Machine learning and deep learning in genetics and genomics . *Machine Learning in Dentistry*. Springer Nature Switzerland AG. 2021. doi.org/10.1007/978-3-030-71881-7_13. p p163-181

Submitted articles

1. Buren EV, Hu M, Cheng L, Wrobel J, Wilhelmsen K, Su L, Li Y, **Wu D**. TWO-SIGMA-G: A New Competitive Gene Set Testing Framework for scRNA-seq Data Accounting for Inter-Gene and Cell-Cell Correlation. Under Review. 29 page. https://www.biorxiv.org/content/10.1101/2021.01.24.427979v1

2. Cho H, Liu C, Tang B, Lin B, et al Love M, Divaris K, **Wu D** Distribution-based comprehensive evaluation of methods for differential expression analysis in metatranscriptomics. Under Review. 40 pages. https://www.biorxiv.org/content/10.1101/2021.07.14.452374v1

3. Cho H, Liu C, Preisser J, **Wu D**. A bivariate zero-inflated negative binomial model for identifying underlying dependence with application to single cell RNA sequencing data. Under Review. 34 pages. https://www.biorxiv.org/content/10.1101/2020.03.06.977728v1

4. Pillai M, **Wu D**. Validation approaches for computational drug repurposing: a review. Under Review. 19 pages

Refereed papers/journal articles

1. Xie J, Cho H, Lin B, et al **Wu D**. Improved Metabolite Prediction Using Microbiome Data-Based Elastic Net Models. Accepted. Frontiers in Cellular and Infection Microbiology. Sep 2021

2. Yang Y, Sun H, Zhang Y, Zhang T, Gong J, Wei Y, Duan Y, Sun M, Yang Y, **Wu D***, Yu D*. Dimensionality reduction by UMAP reinforces sample heterogeneity analysis in bulk transcriptomic data . ***Cell Report.*** July, 2021. DOI:https://doi.org/10.1016/j.celrep.2021.109442 *Co-corresponding author

3. Karhade DS, Roach J, Shrestha P, Simancas-Pallares MA, Ginnis J, Burk ZJS, Ribeiro AA, Cho H, **Wu D**, Divaris K. An Automated Machine Learning Classifier for Early Childhood Caries. ***Pediatr Dent***. 2021 May 15;43(3):191-197.

4. Cho S, Zhu Z, Li T, Baluyot K, Howell BR, Hazlett HC, Elison JT, Hauser J, S Norbert, **Wu D**, Lin W. (2021) Human milk 3'-Sialyllactose is positively associated with language development during infancy. ***The American Journal of Clinical Nutrition***. May, 2021, 4.doi: 10.1093/ajcn/nqab103

5. Rosa T, Neves A, Azcarate-Peril MA, Divaris K, **Wu D**, Cho H, Moss K, Paster J, Chen T, Freitas-Fernandes LB, Fidalgo TKS, Lopes R, Valente AP, Arnold RR, Ribeiro AA. (2021) The Bacterial Microbiome and Metabolome in Caries Progression and Arrest. ***The Journal of Oral Microbiology***. doi.org/10.1080/20002297.2021.1886748. June 16, 2021.

6. Zhang S, Philips KH, Moss K, **Wu D**, Adam H, Selvin E, Demmer RT, Pankow J, Norby F, Mustapha IZ, Beck JD. Periodontal and Risk of Incident Diabetes in the Atherosclerosis

Risk in Communities (ARIC) Study: A BMI-modified Association. ***J Clinical Endocrinology & Metabolism***, 2021; https://doi.org/10.1210/clinem/dgab337

7. Shallal-Ayzin M, Trinh T, Yeung W, Tawil PZ, Cl H, **Wu D**, Khan AA. (2021) A Prospective Analysis of the Correlation between Postoperative Pain and Vital Pulp Therapy. ***The Frontiers in Dental Medicine***. Feb 22, 2021.

8. Yip J, Liu C, **Wu D**, Fouad A. (2021) The association of apical periodontitis and type 2 diabetes mellitus: A large hospital network cross-sectional case-controlled study.. ***The Journal of the American Dental Association***. Jan 11, 2021. doi.org/10.1016/j.adaj.2021.01.005

9. Zou M, Jiang D, Wu T, ZHang X, ZHao Y, **Wu D**, Sun W, Cui J, Moreland L, Li G. (2021) Post-GWAS functional studies reveal an RA-associated CD40 induced NF-kB signal transduction and transcriptional regulation network targeted by class II HDAC inhibitors. ***Human Molecular Genetics***. 2021 Jan doi.org/10.1093/hmg/ddab032

10. Lou JT, Yang Y, Gu QS, Price BA, Qiu YH, Fedoriw Y, Desai S, Mose LE, Chen B, Tateishi S, Parker JS, Vaziri C, **Wu D**. (2021) Rad18 mediates specific mutational signatures and shapes the genomic landscape of carcinogen-induced tumors in vivo. ***Nucleic Acid Research Cancer***, 2021 Jan 06; 3(1) zcaa037, doi.org/10.1093/narcan/zcaa037

11. Heimisdóttir L, Lin B, Cho H, Orlenko A; Ribeiro A; Simon-Soro A, Roach J, Shungin D, Ginnis J, Simancas-Pallares Miguel, Spangler Hudson, Zandona A, Wright JT, Ramamoorthy S, Moore J, Koo HM, **Wu D**, Divaris K. (2021) Metabolomics insights in early childhood caries. ***J Dent Res***. 2021 Jan 9;22034520982963. doi: 10.1177/0022034520982963

12. Wang N, Du N, PengY, Yang K,Shu Z, CHang K, **Wu D**, Yu J,Zhou Y, Li X, Liu B, Gao Z, Zhnnag R, Zhou X. (2021) Network Patterns of Herbal Combinations in Traditional Chinese Clinical Prescriptions. ***Frontiers in pharmacology.*** 2020; 11: 590824. doi: 10.3389/fphar.2020.590824

13. Crowley C , Yang YC , Qiu YJ, Hu BX, Abnousi A, Lipiński J, Plewczynski D , **Wu D**, Won HJ , Ren B , Hu M, Yun Li. (2020) FIREcaller: Detecting frequently interacting regions from Hi-C data. ***Computational and Structural Biotechnology Journal*** 2020 Dec 29; 19(2021): 355-362. doi: 10.1016/j.csbj.2020.12.026

14. Zou MJ, ZhangXY, Jiang DL, Zhao YH, Wu T, Gong QK, Su H, **Wu D**, Moreland L, Li G. (2020) Transcriptional Regulation of CD40 Expression by 4 Ribosomal Proteins via a Functional SNP on a Disease-Associated CD40 Locus. ***Genes (Basel)*** 2020 Dec 21; 11(12):1526. doi: 10.3390/genes11121526.

15. Buren EV, Hu M, Weng C, Jin FL, Li Y, **Wu D***, Li Y* . (2020) TWO-SIGMA: A novel two-component single cell model-based association method for single-cell RNA-seq data. ***Genet Epidemiol*** 2020 Sep 29. doi: 10.1002/gepi.22361. *Co-corresponding author

16. Guo ZL, Wang G, Wu B, Chou WC, Cheng L, Zhou CL, Lou JT, **Wu D**, Su LS, Zheng JN, Ting J, Wan YY. (2020) DCAF1 regulates Treg senescence via the ROS axis during immunological ageing. ***J Clin Invest***, 2020 Jul 30;136466. doi: 10.1172/JCI136466.

17. Zhao Y, **Wu D**, Jiang D, Zhang X, Wei S, Wu T, Cui J, Qian M, Zhao J, Oesterreich S, Finkel T, Li G. (2020) A sequential methodology for the rapid identification and characterization of breast cancer-associated functional SNPs. ***Nature Communications***. 2020, July 11 (1), 1-11. DOI 10.1038/s41467-020-17159-8.

18. Zhong W, Dong L, Poston TB, Spracklen CN, **Wu D**, Darville T, Mohlke K, Li Y, Li Q and Zheng X. (2020) Inferring Causal Networks from Mixed Observational Data Using Directed Acyclic Graphs. ***Frontiers in Genetics***. 2020 Feb 7;11:8. doi: 10.3389/fgene.2020.00008. eCollection 2020.

19. Divaris K, Slade GD, Ferreira Zandona AG, Preisser JS, Ginnis J, Simancas-Pallares MA, Agler CS, Shrestha P, Karhade DS, Ribeiro AA, Cho HY, Gu Y, Meyer BD, Joshi AR, Azcarate-Peril MA, Basta PV, **Wu D**, North KE. (2020) Cohort profile: ZOE 2.0—a community-based genetic epidemiologic study of early childhood oral health. ***Int. J. Environ. Res. Public Health*** 2019, Nov 1; 17(21):8056; https://doi.org/10.3390/ijerph17218056

20. Tanoue Y, Toyoda T, Sun J, Mustofa MK, Tateishi C, Endo S, Motoyama N, Araki K, **Wu D**, Okuno Y, Tsukamoto T, Takeya M, Ihn H, Vaziri C, Tateishi S. (2018) Differential Roles of Rad18 and Chk2 in Genome Maintenance and Skin Carcinogenesis Following Uv Exposure. ***J Invest Dermatol*** 2018 Dec;138(12):2550-2557. doi: 10.1016/j.jid.2018.05.015. Epub 2018 May 31.

21. Wu B, Zhang S, Guo Z, Wang G, Zhang G, Xie L, Lou J, Chen X, **Wu D**, Bergmeier W, Zheng J, Wan YY. (2018) Ras P21 Protein Activator 3 (Rasa3) Specifically Promotes Pathogenic T Helper 17 Cell Generation by Repressing T-Helper-2-Cell-Biased Programs. ***Immunity***. 2018 Nov 20;49(5):886-898.e5. doi: 10.1016/j.immuni.2018.09.004. Epub 2018 Nov 13.

22. Li G, Martinez-Bonet M, **Wu D**, Yang Y, Cui J, Nguyen HN, Cunin P, Levescot A, Bai M, Westra HJ, Okada Y, Brenner MB, Raychaudhuri S, Hendrickson EA, Maas RL, Nigrovic PA. (2018) High-Throughput Identification of Noncoding Functional Snps Via Type Iis Enzyme Restriction. ***Nat Genet.*** 2018 Aug;50(8):1180-1188. doi: 10.1038/s41588-018-0159-z. Epub 2018 Jul 16.

23. Duan Q, Xu Z, Raffield LM, Chang S, **Wu D**, Lange EM, Reiner AP, Li Y. (2018) A Robust and Powerful Two-Step Testing Procedure for Local Ancestry Adjusted Allelic Association Analysis in Admixed Populations. ***Genet Epidemiol.*** Apr;42(3):288-302. doi: 10.1002/gepi.22104. Epub 2017 Dec 10.

24. Yu L, **Wu D**, Gao H, Balic JJ, Tsykin A, Han TS, Liu YD, Kennedy CL, Li JK, Mao JQ, Tan P, Oshima M, Goodall GJ, Jenkins BJ. (2018) Clinical Utility of a Stat3-Regulated Mirna-200 Family Signature with Prognostic Potential in Early Gastric Cancer. ***Clin Cancer Res.*** 2018 Mar 15;24(6):1459-1472. doi: 10.1158/1078-0432.CCR-17-2485. Epub 2018 Jan 12.

25. Morelli T, Moss KL, Preisser JS, Beck JD, Divaris K, **Wu D**, Offenbacher S. (2018) Periodontal Profile Classes Predict Periodontal Disease Progression and Tooth Loss. ***J Periodontol*** 2018 Feb;89(2):148-156. doi: 10.1002/JPER.17-0427. Epub 2018 Feb 22.

26. Miao MZ, Wang B, **Wu D**, Zhang S, Wong S, Shi O, Hu A, Mao L, Fang B. (2018) Temporomandibular Joint Positional Change Accompanies Post-Surgical Mandibular Relapse-a Long-Term Retrospective Study among Patients Who Underwent Mandibular Advancement. ***Orthod Craniofac Res.*** 2018 Feb;21(1):33-40. doi: 10.1111/ocr.12209. Epub 2017 Dec 7.

27. Ferrand J, Croft NP, Pepin G, Diener KR, **Wu D**, Mangan NE, Pedersen J, Behlke MA, Hayball JD, Purcell AW, Ferrero RL, Gantier MP. (2018) The Use of Crispr/Cas9 Gene Editing to Confirm Congenic Contaminations in Host-Pathogen Interaction Studies. ***Front Cell Infect Microbiol.*** 2018 Mar 19;8:87. doi: 10.3389/fcimb.2018.00087. eCollection 2018.

28. Taggar T, **Wu D**, Khan AA. (2017) A Randomized Clinical Trial Comparing 2 Ibuprofen Formulations in Patients with Acute Odontogenic Pain. ***J Endod***. 2017 May;43(5):674-678. doi: 10.1016/j.joen.2016.12.017. Epub 2017 Mar 18.

29. Morelli T, Moss KL, Beck J, Preisser JS, **Wu D**, Divaris K, Offenbacher S. (2017) Derivation and Validation of the Periodontal and Tooth Profile Classification System for Patient Stratification. ***J Periodontol***. 2017 Feb;88(2):153-165. doi: 10.1902/jop.2016.160379. Epub 2016 Sep 13.

30. Sun XX, Dalpiaz D, **Wu D**, Liu JS, Zhong W, Ma P. (2016) Statistical Inference for Time Course Rna-Seq Data Using a Negative Binomial Mixed-Effect Model. ***BMC Bioinformatics.*** 2016 Aug 26;17(1):324. doi: 10.1186/s12859-016-1180-9.

31. Leong YA, Chen Y, Ong HS, **Wu D**, Man K, Deleage C, Minnich M, Meckiff BJ, Wei Y, Hou Z, Zotos D, Fenix KA, Atnerkar A, Preston S, Chipman JG, Beilman GJ, Allison CC, Sun L, Wang P, Xu J, Toe JG, Lu HK, Tao Y, Palendira U, Dent AL, Landay AL, Pellegrini M, Comerford I, McColl SR, Schacker TW, Long HM, Estes JD, Busslinger M, Belz GT, Lewin SR, Kallies A, Yu D. (2016) Cxcr5(+) Follicular Cytotoxic T Cells Control Viral Infection in B Cell Follicles. ***Nat Immunol.*** 2016 Oct;17(10):1187-96. doi: 10.1038/ni.3543. Epub 2016 Aug 3.

32. Li G, Cunin P, **Wu D**, Diogo D, Yang Y, Okada Y, Plenge RM, Nigrovic PA. (2016) The Rheumatoid Arthritis Risk Variant Ccr6dnp Regulates Ccr6 Via Parp-1. ***PLoS Genet***. 2016 Sep 14;12(9):e1006292. doi: 10.1371/journal.pgen.1006292. eCollection 2016 Sep.

33. Zhang S, Divaris K, Moss K, Yu N, Barros S, Marchesan J, Morelli T, Agler C, Kim SJ, **Wu D**, North KE, Beck J, Offenbacher S. (2016) The Novel Asic2 Locus Is Associated with Severe Gingival Inflammation. ***JDR Clin Trans Res***. 2016 Jul;1(2):163-170. doi: 10.1177/2380084416645290. Epub 2016 Apr 20.

34. **Wu D**, Gantier MP. (2016) Normalization of Affymetrix Mirna Microarrays for the Analysis of Cancer Samples. ***Methods Mol Biol.*** 2016;1375:1-10. doi: 10.1007/7651_2015_239.

35. Worley MJ, Jr., Liu S, Hua Y, Kwok JS, Samuel A, Hou L, Shoni M, Lu S, Sandberg EM, Keryan A, **Wu D**, Ng SK, Kuo WP, Parra-Herran CE, Tsui SK, Welch W, Crum C, Berkowitz RS, Ng SW. (2015) Molecular Changes in Endometriosis-Associated Ovarian Clear Cell Carcinoma. ***Eur J Cancer***. 2015 Sep;51(13):1831-42. doi: 10.1016/j.ejca.2015.05.011. Epub 2015 Jun 6.

36. Xu H, Xiao T, Chen CH, Li W, Meyer CA, Wu Q, **Wu D**, Cong L, Zhang F, Liu JS, Brown M, Liu XS. (2015) Sequence determinants of Improved CRISPR sgRNA Design. ***Genome Res.*** 2015 Aug;25(8):1147-57. doi: 10.1101/gr.191452.115. Epub 2015 Jun 10.

37. Ng AP, Hu Y, Metcalf D, Hyland CD, Ierino H, Phipson B, **Wu D**, Baldwin TM, Kauppi M, Kiu H, Di Rago L, Hilton DJ, Smyth GK, Alexander WS. (2015) Early Lineage Priming by Trisomy of Erg Leads to Myeloproliferation in a Down Syndrome Model. ***PLoS Genet***. 2015 May 14;11(5):e1005211. doi: 10.1371/journal.pgen.1005211.

38. Ritchie ME, Phipson B, **Wu D**, Hu Y, Law CW, Shi W, Smyth GK. (2015) Limma Powers Differential Expression Analyses for Rna-Sequencing and Microarray Studies. ***Nucleic Acids Res***. 2015 Apr 20;43(7):e47. doi: 10.1093/nar/gkv007. Epub 2015 Jan 20.

39. Okada Y, **Wu D**, Trynka G, Raj T, Terao C, Ikari K, Kochi Y, Ohmura K, Suzuki A, Yoshida S, Graham RR, Manoharan A, Ortmann W, Bhangale T, Denny JC, Carroll RJ, Eyler AE, Greenberg JD, Kremer JM, Pappas DA, Jiang L, Yin J, Ye L, Su DF, Yang J, Xie G, Keystone E, Westra HJ, Esko T, Metspalu A, Zhou X, Gupta N, Mirel D, Stahl EA, Diogo D, Cui J, Liao K, Guo MH, Myouzen K, Kawaguchi T, Coenen MJ, van Riel PL, van de Laar MA, Guchelaar HJ, Huizinga TW, Dieude P, Mariette X, Bridges SL, Jr., Zhernakova A, Toes RE, Tak PP, Miceli-Richard C, Bang SY, Lee HS, Martin J, Gonzalez-Gay MA, Rodriguez-Rodriguez L, Rantapaa-Dahlqvist S, Arlestig L, Choi HK, Kamatani Y, Galan P, Lathrop M, consortium R, consortium G, Eyre S, Bowes J, Barton A, de Vries N, Moreland LW, Criswell LA, Karlson EW, Taniguchi A, Yamada R, Kubo M, Liu JS, Bae SC, Worthington J, Padyukov L, Klareskog L, Gregersen PK, Raychaudhuri S, Stranger BE, De Jager PL, Franke L, Visscher PM, Brown MA, Yamanaka H, Mimori T, Takahashi A, Xu H, Behrens TW, Siminovitch KA, Momohara S, Matsuda F, Yamamoto K, Plenge RM. (2014) Genetics of Rheumatoid Arthritis Contributes to Biology and Drug Discovery. ***Nature***. 2014 Feb 20;506(7488):376-81. doi: 10.1038/nature12873.

40. Tartakoff AM, **Wu D**. (2014) The Axis of Progression of Disease. ***Cancer Inform***. 2014; 13(Suppl 6): 7–13. Published online 2014 Oct 19. doi: 10.4137/CIN.S17683

41. **Wu D**, Pang Y, Wilkerson MD, Wang D, Hammerman PS, Liu JS. (2013) Gene-Expression Data Integration to Squamous Cell Lung Cancer Subtypes Reveals Drug Sensitivity. ***Br J Cancer***. 2013 Sep 17;109(6):1599-608. doi: 10.1038/bjc.2013.452. Epub 2013 Sep 3.

42. **Wu D**, Hu Y, Tong S, Williams BR, Smyth GK, Gantier MP. (2013) The Use of Mirna Microarrays for the Analysis of Cancer Samples with Global Mirna Decrease. ***RNA***. 2013 Jul;19(7):876-88. doi: 10.1261/rna.035055.112. Epub 2013 May 24.

43. Li G, Diogo D, **Wu D**, Spoonamore J, Dancik V, Franke L, Kurreeman F, Rossin EJ, Duclos G, Hartland C, Zhou X, Li K, Liu J, De Jager PL, Siminovitch KA, Zhernakova A,

Raychaudhuri S, Bowes J, Eyre S, Padyukov L, Gregersen PK, Worthington J, Rheumatoid Arthritis Consortium I, Gupta N, Clemons PA, Stahl E, Tolliday N, Plenge RM. (2013) Human Genetics in Rheumatoid Arthritis Guides a High-Throughput Drug Screen of the Cd40 Signaling Pathway. ***PLoS Genet***. 2013 May;9(5):e1003487. doi: 10.1371/journal.pgen.1003487. Epub 2013 May 16.

44. Woollard DJ, Opeskin K, Coso S, **Wu D**, Baldwin ME, Williams ED. (2013) Differential Expression of Vegf Ligands and Receptors in Prostate Cancer. ***Prostate***. 2013 May;73(6):563-72. doi: 10.1002/pros.22596. Epub 2012 Oct 4.

45. **Wu D**, Smyth GK. (2012) Camera: A Competitive Gene Set Test Accounting for Inter-Gene Correlation. ***Nucleic Acids Res***. 2012 Sep 1;40(17):e133. doi: 10.1093/nar/gks461. Epub 2012 May 25.

46. **Wu D**, Wang D. (2012) Application of Advanced Gene Set Tests in Breast Cancer Research. ***China Journal of Bioinformatics***. 10(2): 92-95. (in Chinese)

47. Asselin-Labat ML, Sutherland KD, Vaillant F, Gyorki DE, **Wu D**, Holroyd S, Breslin K, Ward T, Shi W, Bath ML, Deb S, Fox SB, Smyth GK, Lindeman GJ, Visvader JE. (2011) Gata-3 Negatively Regulates the Tumor-Initiating Capacity of Mammary Luminal Progenitor Cells and Targets the Putative Tumor Suppressor Caspase-14. ***Mol Cell Biol***. 2011 Nov;31(22):4609-22. doi: 10.1128/MCB.05766-11. Epub 2011 Sep 19.

48. **Wu D**, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK. (2010) Roast: Rotation Gene Set Tests for Complex Microarray Experiments. ***Bioinformatics***. 2010 Sep 1;26(17):2176-82. doi: 10.1093/bioinformatics/btq401. Epub 2010 Jul 7.

49. Asselin-Labat ML, Vaillant F, Sheridan JM, Pal B, **Wu D**, Simpson ER, Yasuda H, Smyth GK, Martin TJ, Lindeman GJ, Visvader JE. (2010) Control of Mammary Stem Cell Function by Steroid Hormone Signalling. ***Nature***. 2010 Jun 10;465(7299):798-802. doi: 10.1038/nature09027. Epub 2010 Apr 11.

50. Lim E, **Wu D***, Pal B, Bouras T, Asselin-Labat ML, Vaillant F, Yagita H, Lindeman GJ, Smyth GK, Visvader JE. (2010) Transcriptome Analyses of Mouse and Human Mammary Cell Subpopulations Reveal Multiple Conserved Genes and Pathways. ***Breast Cancer Res***. 2010;12(2):R21. doi: 10.1186/bcr2560. Epub 2010 Mar 26. *Co-first Authors

51. Lim E, Vaillant F, **Wu D**, Forrest NC, Pal B, Hart AH, Asselin-Labat ML, Gyorki DE, Ward T, Partanen A, Feleppa F, Huschtscha LI, Thorne HJ, kConFab, Fox SB, Yan M, French JD, Brown MA, Smyth GK, Visvader JE, Lindeman GJ. (2009) Aberrant Luminal Progenitors as the Candidate Target Population for Basal Tumor Development in Brca1 Mutation Carriers. ***Nat Med***. 2009 Aug;15(8):907-13. doi: 10.1038/nm.2000. Epub 2009 Aug 2.

52. Gargett CE, Schwab KE, Zillwood RM, Nguyen HP, **Wu D**. Isolation and Culture of Epithelial Progenitors and Mesenchymal Stem Cells from Human Endometrium. ***Biol Reprod***. 2009 Jun;80(6):1136-45. doi: 10.1095/biolreprod.108.075226. Epub 2009 Feb 18.

53. **Wu D**, Townsley E, Tartakoff AM. (2007) Covert Genetic Selections to Optimize Phenotypes. ***PLoS One***. 2007 Nov 21;2(11):e1200. doi: 10.1371/journal.pone.0001200.

54. Huang S, Liang J, Zheng M, Li X, Wang M, Wang P, Vanegas D, **Wu D**, Chakraborty B, Hays AP, Chen K, Chen SG, Booth S, Cohen M, Gambetti P, Kong Q. (2007) Inducible Overexpression of Wild-Type Prion Protein in the Muscles Leads to a Primary Myopathy in Transgenic Mice. ***Proc Natl Acad Sci USA***. 2007 Apr 17;104(16):6800-5. doi: 10.1073/pnas.0608885104. Epub 2007 Apr 9.

55. Kong Q, Huang S, Zou W, Vanegas D, Wang M, **Wu D**, Yuan J, Zheng M, Bai H, Deng H, Chen K, Jenny AL, O'Rourke K, Belay ED, Schonberger LB, Petersen RB, Sy MS, Chen SG, Gambetti P. (2005) Chronic Wasting Disease of Elk: Transmissibility to Humans Examined by Transgenic Mouse Models. ***J Neurosci***. 2005 Aug 31;25(35):7944-9. doi: 10.1523/JNEUROSCI.2467-05.2005.

56. Kong Q, Huang S, **Wu D**, Wang M, Vanegas D, Bai H, Deng H, Chen K, Zou W, Jenny AL, O'Rourke K, Sy M., Chen SG and Gambetti P. (2004) Transmissibility of chronic wasting disease of elk and deer to humans. ***J Neuropathology and Experimental Neurology***. 63:515.

Published Refereed Abstracts

1. **Wu D**, Cho H, Simancas-Pallares MA, Ginnis J, Ferreira Zandona AG, Divaris K. Understanding the Early Childhood Caries Microbiome: Integrative Metatranscriptomics-Metagenomics Analyses. J Dent Res 2019;98 (Spec Iss A): 3072291 (IADR/AADR/CADR).

2. Shungin D, Simancas-Pallares MA, Ginnis J, Ferreira Zandona AG, **Wu D**, Divaris K. Characterizing Supragingival Biofilm Metatranscriptome And Metagenome In Early Childhood Caries. J Dent Res 2019;98 (Spec Iss A): 3181468 (IADR/AADR/CADR).

3. Ribeiro AA, Rosa T, Neves AA, Azcarate-Peril MA, Divaris K, **Wu D**, Cho H, Paster BJ, Chen T, Fidalgo T, Freitas-Fernandes LB, Valente A, Arnold R. The Oral Microbiome and Metabolome in Caries Initiation and Arrestment. J Dent Res 2019;98 (Spec Iss A): 3072291 (IADR/AADR/CADR).

4. Heimisdottir LH, Cho H, Ginnis J, Simancas-Pallares MA, Ferreira Zandona MA, Shungin D, **Wu D**, Divaris K. Metabolomics Insights in Early Childhood Caries. J Dent Res 2019;98 (Spec Iss A): 3072291 (IADR/AADR/CADR).

5. Li G, Chen M, Li G, **Wu D**, Lian C, Sun Q, Shen D. A longitudinal MRI study of amygdala and hippocampal subfields for infants with risk of Autism, Annual meeting of Organization for Human Brain Mapping. Rome Italy. GLMI 2019. Lecture Notes in Computer Science, Vol 11849. Springer, Cham.

6. **Wu D**, Gupta A, Moss K, Morelli T, Beck J, Offenbacher S. Novel patient level and teeth level classification on PAVE study to compare community and periodontal treatment

outcomes. International Association for Dental Research (IADR). London, UK. J Dent Res 2018;97 (Spec Iss B): 2959972.

7. Divaris K, Cho H, **Wu D,** Roach J, Rodríguez Cortés A, Basta PV, Ferreira Zandoná AG, Ginnis J, Meyer BD, Hu S, Simancas-Pallares MA, Butz N, Azcarate-Peril MA. Supragingival Biofilm Transcriptomics in Early Childhood Oral Health. J Dent Res 2018;97 (Spec Iss B): (IADR/CED).

8. **Wu D**, Wang R, Teles F, Integrative Statistical Analysis of the Microbiome, Metabolome and Inflammation Data in Fanconi Anemia to Understand Oral Cancer Causality. American Association for Dental Research (AADR), Florida. J Dent Res 2018;97 (Spec Iss A): 2865782.

9. **Wu D**, Moss K, Morelli T, Beck J, Offenbacher S. Extracting dental school periodontal records to measure treatment outcomes and risk modification. American Association for Dental Research (AADR)/IADR, San Francisco, CA. J Dent Res 2017;96 (Spec Iss A):2640139.

10. Morelli T, Moss KL, Beck J, Preisser JS, **Wu D**, Divaris K, Offenbacher S. Periodontal Profile Classes Predict Periodontitis Progression and Tooth Loss. J Dent Res 2017;96 (Spec Iss A): 0129 (IADR/AADR/CADR).

Invited Oral Presentations and Unpublished Abstracts

Local

2019 Transcriptional gene set tests and microbial omics data analysis. Biostatistics Department Seminar, UNC-Chapel Hill, Chapel Hill, NC.

2017 Extracting dental school periodontal records to measure treatment outcomes for precision medicine. UNC Perio Expo. UNC-Chapel Hill, Chapel Hill, NC.

2016 Biostatistics Seminar Series - NC TraCS, Colloquium Seminars - Bioinformatics and Computational Biology, and Center for Image Analysis and Informatics. UNC-Chapel Hill, Chapel Hill, NC.

2015 Oral cancer susceptibility and causality: integrative analysis of the microbiome, metabolome and inflammation in Fanconi Anemia. Dental Research Day. UNC School of Dentistry. UNC-Chapel Hill, Chapel Hill, NC.

2015 A mixture model for contamination detection in target DNAseq. UNC Lineberger Comprehensive Cancer Center. UNC-Chapel Hill, Chapel Hill, NC.

2015 Cancer personalized medicine. UNC Health Informatics Program. UNC-Chapel Hill, Chapel Hill, NC.

2015 RNAseq data analysis. UNC Genetics. UNC-Chapel Hill, Chapel Hill, NC.

2014 Genomic data integration to discover/repurpose drugs for complex diseases. School of Dentistry. UNC-Chapel Hill, Chapel Hill, NC.

2014 Genomic data integration to discover/repurpose drugs for complex diseases. Biostatistics Department. UNC-Chapel Hill, Chapel Hill, NC.

National

2020 Gene set testing methods for single cell RNAseq (scRNAseq) data, Joint Statistical Meetings (JSM), Philadelphia, PA. (online)

2019 A bivariate zero inflated negative binomial model for gene-gene dependence in single cell RNAseq, Joint Statistical Meetings (JSM), Denver, CO.

2018 Integrative causal analysis of microbiome, metabolome and inflammation data to understand oral cancer mechanism. Symposium-International Chinese Statistical Association (ICSA), New Brunswick, NJ.

2017 Classification method of dental electronic health record (EHR) data potentially improves precision treatment. American Medical Informatics Association (AMIA) Annual Meeting, Washington, DC.

2014 ROMER, ranked based rotation gene set test. Joint Statistical Meetings (JSM), Boston, MA.

2014 Introduction to novel gene set tests. Bioconductor Annual Meeting, Boston, MA.

2014 Data integration for drug repurposing. Symposium-International Chinese Statistical Association (ICSA), Portland, OR.

2014 Genomic data integration to discover/repurpose drugs for complex diseases. Cleveland Clinic Foundation, Cleveland, OH.

2014 Genomic data integration to discover/repurpose drugs for complex diseases. Massachusetts General Hospital, Harvard Medical School, Cambridge, MA.

2014 Genomic data-based drug discovery/repurposing. Biostatistics Department, MD Anderson Cancer Center, Houston, TX.

2013 The Genotype-Tissue Expression (GTEx) retreat, Cambridge, MA.

2013 Integrating GWAS data with drug information for drug repurposing. Program in Quantitative Genomics, Harvard School of Public Health, Boston, MA.

2013 Gene set tests in breast cancer, stem cell and drug repurposing. Institute for Stem Cell

Biology and Regenerative Medicine, Stanford University, Polo Alto, CA.

2012 Novel gene set tests in breast cancer and stem cell research. Symposium-International Chinese Statistical Association (ICSA). Cambridge, MA.

2012 Bayesian Gene Set Test. Eastern North American Region/International Biometric Society Meeting (ENAR), Washington DC.

2012 shRNA data analysis using limma. Broad Institute, Project Achilles Group.

2012 shRNA data analysis using limma. Dana Farber Institute, Matthew Meyerson group, Boston, MA.

2011 Gene set testing, stem cells and breast cancer. The high dimension data seminar in Biostatistics Department, Harvard University, Boston, MA.

2010 Analysis of transcriptional signatures reveals the cell of origin for breast cancer subtypes. Harvard University, Broad Institute, Dana Farber Cancer Institute, Boston, MA. John Hopkins University, MD Anderson Cancer Center, Stanford University.

2010 ROAST: rotation gene set tests for complex microarray experiments. 18th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB). Boston, MA.

<u>International</u>

2019 Evaluation of statistical methods for differential expression analysis in microbiome metatranscriptomics data. Symposium-International Chinese Statistical Association (ICSA), Hangzhou, China.

2019 Understanding early child cavities by analysis of metatranscriptomics and metagenomics, American Association for Dental Research (AADR)/IADR, Vancouver, Canada.

2019 Microbial association with the age-adjusted mullen score in a longitudinal study of children's early life – Baby Connectome Project (BCP), 6th International Conference on Nutrition & Growth (N&G 2019), Valencia, Spain.

2018 Comparison of statistical methods to analyze metatranscriptome data for differential abundance in an early childhood oral health study, International Association for Dental Research (IADR), London, UK.

2018 Statistical methods to access gene-gene associations in single cell RNA seq data, ISCA-China, Qingdao, China.

2016 A mixture model for contamination detection in target DNA seq, the 3rd Taihu International Statistics Forum, Shanghai, China.

2013 Genomic data-based drug discovery/repurposing. Ontario Institute for Cancer Research.

University of Toronto, Canada.

2012 CD40 pathway analysis, Shandong Academy of Sciences, Jinan, China.

2011 Investigation of gene-gene interaction in GWAS data and eQTL data. WEHI. Melbourne, Australia.

2010 Analysis of transcriptional signatures reveals the cell of origin for breast cancer subtypes. National Information and Communications Technology Australia (NICTA), Melbourne, Australia.

2010 Analysis of transcriptional signatures reveals the cell of origin for breast cancer subtypes. The Eskitis Institute for Cell and Molecular Therapies, Brisbane, Australia.

2009 Patterns across data sets: finding the cell origin of basal-like breast tumours. Bioinformatics Australia. Melbourne, Australia.

2009 Patterns across data sets: finding the cell origin of basal-like breast tumours. The 9th Annual Australian Microarray and Associated Technologies Association (AMATA-9). Sydney, Australia.

2008 Roast, a gene set testing method for laboratory generated microarray data. The 8th Annual Australian Microarray and Associated Technologies Association (AMATA-8). Dunedin, New Zealand.

2008 Residual space permutation for gene set testing in designed microarray experiments. Australian Statistical Conference, Melbourne, Australia.

2007 An empirical Bayes approach used in microarray data analysis. Graduate Statistics Course of Australian Mathematical Sciences Institute (AMSI), Brisbane, Australia.

2007 Gene set enrichment tests in microarray experiments with small sample sizes. The 10th annual Microarray and Gene Expression Data Society meeting (MGED-10) and the 7th Annual Australian Microarray and Associated Technologies Association (AMATA-7). Brisbane, Australia.

Patents:

2019 Compositions comprising human milk oligosaccharides for use in a subject to support language development, Submitted application. Dec 2019. UNC. Weili Lin, **Di Wu**, Tengfei Li, Ziliang Zhu and Seoyoon Cho. Docket No. 3712036-03567.

2009 'Gene expression profiles and uses therefor' WEHI, Geoffrey LINDEMAN, Jane Visvader, Gordon Smyth, **Di Wu**. PCT number pending. Priority Date: 6th May 2009. Provisional patent Application No. AU2009901989. Human breast stem and luminal

progenitor cells.

Digital and other novel forms of scholarship: Software Developed

2019 R package BZINB available in CRAN https://cran.r-project.org/web/packages/bzinb/index.html and gitHub https://github.com/Hunyong/BZINB . 'bzinb: Bivariate Zero-Inflated Negative Binomial Model Estimator', by Hunyong Cho, Chuwen Liu, Jinyoung Park, **Di Wu**. It provides a maximum likelihood estimation of Bivariate Zero-Inflated Negative Binomial (BZINB) model or the nested model parameters. Also estimates the underlying correlation of a pair of count data. The paper draft is submitted to Biometrics and is currently under review. The draft can be found in BioRxiv www.biorxiv.org.

2019 R package twosigma available in gitHub https://github.com/edvanburen/twosigma . It's a TWO-component SInGle cell Model-based Association method for differential expression (DE) analyses in single-cell RNA-seq (scRNA-seq) data. The first component models the probability of "drop-out" with a mixed-effects logistic regression model and the second component models the (conditional) mean expression with a mixed-effects negative binomial regression model. It also allows random effects due to the correlation among cells with-in a sample for better Type I error control. The paper draft is submitted and currently under review. It can be found in BioRxiv https://www.biorxiv.org/content/10.1101/709238v2.

2007-15 R package LIMMA. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, Smyth GK (2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies." *Nucleic Acids Research*, 43(7), e47. doi: 10.1093/nar/gkv007.

2010 R function *roast* in LIMMA Package. Wu D, Lim E, Vaillant F, Asselin-Labat ML, Visvader JE, Smyth GK. 2010. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics*. 26(17):2176-82

2007 R function *camera* function for the correlation adjusted mean rank gene set analysis in LIMMA Package to estimate variance-inflation factor for means of correlated genes. See https://rdrr.io/bioc/limma/src/R/geneset-camera.R. **Wu D**, Smyth GK. Camera: A Competitive Gene Set Test Accounting for Inter-Gene Correlation. Nucleic Acids Res 2012;4017: e133,epub date: 2012/05/29 PMC3458527

TEACHING ACTIVITIES

Major Teaching and Administrative Responsibilities
UNC Adams School of Dentistry
Course Participation
2017-Present Lecturer, DENG 703: Applied Research Methods
1 hour – Spring/Fall ~28 MS students

UNC Biostatistics

Course Participation

2021 Lecturer, BIOS 784: Introduction to Computational Biology
1 hour – Spring ~30 PhD students

2020 Lecturer, BIOS 785: Statistical Methods for Gene Expression Analysis
1 hour – Fall ~30 PhD students

2019 Lecturer, BIOS 785: Statistical Methods for Gene Expression Analysis
1 hour – Spring ~29 PhD students

2017 Lecturer, BIOS 784: Introduction to Computational Biology
1 hour – Fall ~30 PhD students

UNC Information and Library Sciences

Course Participation

UNC Information and Library Sciences

2017 Lecturer, INLS 890: Health Informatics-Advanced Special Topics
1 hour – Fall ~30 PhD students

UNC Bioinformatics and Computational Biology Curriculum

Course Participation

UNC Bioinformatics and Computational Biology Curriculum

2020 Fall-2021Spring
co-mentor, BBSP First Year PhD Groups
1 hour – Fall ~30 PhD students

2016 Lecturer, BCB 720: Introduction to Statistical Modeling
1 hour – Fall ~30 PhD students

MENTORING ACTIVITIES

Undergraduate research projects mentored/supervised

2017-2021 Clustering analysis of single cell RNAseq data
Brian Chen, Undergraduates – Research Intern
University of North Carolina at Chapel Hill, Primary Mentor: Di Wu

2018-2019 Weighted inference of gene expression variability in single cell RNAseq data for gene set tests: a replication study
Jie He, Undergraduate, honor student
University of North Carolina at Chapel Hill, Primary Mentor: Di Wu
The defense was approved with highest honors

2012-2013 Lung cancer subtype classification using gene set tests.
Yushu Pang, Summer Intern in Statistics
Harvard University.Primary Mentor: Jun Liu, co-mentor: Di Wu

MS degree thesis mentoring

2020- present	Childhood Obesity, Early Childhood Caries, and Oral Microbiome: A secondary Biostatistical Analysis of ZOE 2.0. Seung-Hyun Lee. MS in Public Health, University of North Carolina at Chapel Hill, Primary Mentor: Di Wu
2020	ChIP-seq data analysis in causal pathways. Ty Darnell. MS in Biostatistics, University of North Carolina at Chapel Hill, Primary Mentor: Di Wu
2020	Multi-mediator analysis in causal pathways. Arryn Panagos. MS in Biostatistics, University of North Carolina at Chapel Hill, Primary Mentor: Di Wu
2020	Analysis of functional genomics and electronic dental/medical records of oral diseases. Amrita Tembhe. MS in Public health, University of North Carolina at Chapel Hill, Primary Mentor: Di Wu
2019	Evaluation of library size effect on normalization and differential abundance testing methods for microbiome sequencing data. Karen Chen. MS in Biostatistics, University of North Carolina at Chapel Hill, Primary Mentor: Di Wu
2019	Cell population analysis in single cell RNAseq data. Liyong. MS in Biostatistics, University of North Carolina at Chapel Hill, Primary Mentor: Di Wu
2018	Exploration of gene expression variability in single cell RNAseq data for gene set tests. Yiling Liu. MS in Biostatistics, University of North Carolina at Chapel Hill, Primary Mentor: Di Wu
2017	Validation of periodontal and tooth profile classification system for determining periodontitis treatment outcomes: aimed at precision medicine. Ayushi Gupta. MS in Health Informatics, University of North Carolina at Chapel Hill, Carolina Health Informatics Program (CHIP) Primary Mentor: Di Wu
2015	The analgesic effect of a new ibuprofen formulation on odontogenic pain. Tanjit Taggar. MS in School of Dentistry. University of North Carolina at Chapel Hill, Primary advisor: Asma Khan. Co-Advisor: Di Wu

MD/PhD degree mentoring

2021-present	Immune, HIV and Leukemia

Sophie Maharry. MD/PhD in School of Medicine, University of North Carolina at Chapel Hill, Co- Mentor: Di Wu and George Fedoriw

PhD degree mentoring

2020-present	Cell trajectory estimation using single cell RNAseq data with biological replicates. Ji-Eun Park. PhD in Biostatistics, University of North Carolina at Chapel Hill, Co- Mentor: Di Wu and Michael Love
2020-present	Estimation of matrix of Poisson parameters in single cell RNAseq data. Yue Pan. PhD in Biostatistics, University of North Carolina at Chapel Hill, Co-Mentor: Di Wu and Steve Marron
2019-present	Data integration for pathway analysis across Metagenome, Metatranscriptome, and Metabolome. Bridget Lin. PhD in Biostatistics, University of North Carolina at Chapel Hill, Primary Mentor: Di Wu
2017-2020 Oct	Improved statistical methods to analyze single cell RNAseq data: association method and gene set tests. Eric Van Buren. PhD in Biostatistics, University of North Carolina at Chapel Hill, Co-Mentor: Di Wu and Yun Li (Now a postdoc fellow in the Department of Biostatistics at Harvard University)
2017-2021 June	Precision Medicine Methodology development with application to survival and Genomics data. Hunvong Cho. PhD in Biostatistics, University of North Carolina at Chapel Hill, Co-Mentor: Di Wu and Michael Kosorok
2017-present	Computational drug repurposing validation strategies. Malvika Pillai. PhD in Public Health Informatics, University of North Carolina at Chapel Hill, Primary Mentor: Di Wu

PhD thesis committee membership

2020-2021	Methods for characterizing Chromatin Interactions Taylor Lagler. PhD in Biostatistics, University of North Carolina at Chapel Hill, Primary Mentor: Yun Li
2020-2021	Transcriptome-wide association studies: challenges, solutions, abd future directions. Amanda Tapia, PhD in Biostatistics, University of North Carolina at Chapel Hill, Primary Mentor: Yun Li
2020-2021 Aug	Deconvolution and Network construction by single cell RNA sequencing

	Data. Meichen Dong. PhD in Biostatistics, University of North Carolina at Chapel Hill, Primary Mentor: Fei Zou and Yuchao Jiang
2020-present	Developing Statistical Methods in Genetics. Jonathan Rosen. PhD in Biostatistics, University of North Carolina at Chapel Hill, Primary Mentor: Yun Li
2019-present	Defining novel genome maintenance mechanisms in Head and neck squamous cell carcinoma Deepika Jayaprakash. PhD in Oral and Craniofacial Biomedicine, University of North Carolina at Chapel Hill, Primary Mentor: Cyrus Vaziri
2019-2020	Evaluation and Incorporation of uncertainty quantification in differential transcript usage in bulk RNAseq and single cell RNAseq. Scott Van Buren. PhD in Biostatistics, University of North Carolina at Chapel Hill, Primary Mentor: Naim Rashid and Mike Love (Now Bioinformatics Scientist - Harvard University)
2019-2020	Clustering of Bulk RNA-Seq and Missing Data Methods in Deep Learning. David Lim. PhD in Biostatistics, University of North Carolina at Chapel Hill, Primary Mentor: Naim U. Rashid and Joseph G. Ibrahim
2019-2021	Statistical Methods to analyze Hi-C data. Cheynna Crowley. PhD in Biostatistics, University of North Carolina at Chapel Hill, Primary Mentor: Yun Li
2018-2021	Microbiol 16s and RNA-Seq data integration in plants. Isai Salas-Gonzalez. PhD in Genetics, University of North Carolina at Chapel Hill, Primary Mentor: Jeff Dangl
2017-2018	Cluster Ensemble Methods for Single Cell RNA-Seq Data and Deconvolution of Bulk Hi-C Data. Ruth Huh. PhD in Biostatistics, University of North Carolina at Chapel Hill, Primary Mentor: Yun Li
2016	Statistical Methods in Cell Type Abundance Estimation and eQTL Mapping. Doug Wilson. PhD in Biostatistics. University of North Carolina at Chapel Hill, Primary Mentor: Wei Sun

MS thesis research committees

2019-present	Metabolomics Insights in Early Childhood Caries. Lara Heimisdottir. MS in Adams School of Dentistry, University of North

Carolina at Chapel Hill, Primary Mentor: Kimon Divaris

2019-present Influences on dental caries and overweight/obesity among preschool-age children in North Carolina.
Meredith Davis. MS in Adams School of Dentistry, University of North Carolina at Chapel Hill, Primary Mentor: Kimon Divaris

2018-2019 Changes and oral microbiome shifts in HIV and patients following periodontal therapy.
Karin Schey. MS in Adams School of Dentistry, University of North Carolina at Chapel Hill, Primary Mentor: Jennifer Webster-Cyriaque

2016-2018 Oral microbiome changes associated with fixed prosthodontic restoration
Sarah Lee. MS in Adams School of Dentistry, University of North Carolina at Chapel Hill, Primary Mentor: Kimon Divaris

GRANTS AND CONTRACTS

Active

2021/07-22 UNC Idea Grant by ORD: "A Pilot Study to Use Early Oral Biopsy Samples for Early Detection of Oral Cancer"
Total Award: $18,000 (one year)
Principal Investigator: Di Wu -- 0% effort

2021-25 NIH/NIDCR R01: "Development of a cavitation enhancement technology to access archived tissues for epigenetic-based biomedical research"
Principal Investigator: Pattenden -- 3% effort

2019-21 NIH/NIDCR R03: "Investigating the Microbial Basis of Early Childhood Caries via Metagenomics and Metatranscriptomics Analyses"
Total Award: $200,000
Principal Investigator: Di Wu (August 2019-July 2021) -- 25% effort
(with no-cost extension)

2019-24 NIH-NCI R01: "Establishing MAGE-A4/RAD18 as a novel cancer-specific chemotherapeutic target"
Total Award: $413,559
Principal Investigator: Cyrus Vaziri (April 2019-March 2024) -- 5% effort

2018-23 NIH/NIEHS R01: "Pathological Reprogramming of DNA Damage Signaling in Neoplastic Cells"
Total Award: $302,167
Co-Investigator: Cyrus Vaziri (December 2018-November 2023) -- 9% effort

2018-23 NIH-NCI R01: "Defining Mechanisms of Pathological Trans-Lesion Synthesis During Carcinogenesis."
Total Award: $334,841
Co-Investigator: Cyrus Vaziri (February 2018-January 2023) – 16.6% effort

2020 Pilot Award, UNC computational Medicine: "A novel computational pipeline for single cell RNA-seq analysis to reveal the role of type-I IFN in HIV-induced immune dysfunction and viral persistence"
Total Award: $50,000
Principal Investigator: Di Wu (January 2020-December 2020) -- 0% effort (with extension)

Completed

2020 Pilot Award, UNC NC TraCS: "determination of genital mucosal T cell responses involved in protection from Chlamydia trachomatis in women using single cell RNA sequencing analysis"
Total Award: $50,000
Co-Investigator: Di Wu (January 2020-December 2020) -- 0% effort

2019-20 UNC Lineberger Developmental Award: "Validating Trans-Lesion Synthesis as a Novel Therapeutic Target in Glioblastoma"
Total Award: $200,000
Co-Investigator: Di Wu (January 2019-December 2020) -- 0% effort

2018-19 UNC Lineberger Developmental Award: "Quantitative Imaging Data in a Community-Based Mammography Registry: A Feasibility Study"
Total Award: $140,000
Co-Investigator: Di Wu (January 2018-December 2019) 0% effort

2018-19 Pilot Award, UNC Center for Environmental Health and Susceptibility (CEHS): "Defining novel Chk2 functions in suppression of UV-induced skin carcinogenesis"
Total Award: $25,000
Principal Investigator: Di Wu (January 2018-March 2019) 0% effort

2016 NIH-NINDS R21: "Using High Throughput Approach to Identify/Characterize Functional Variants on Multiple Sclerosis"
Total Award: $273,045
Sub Principal Investigator: Di Wu -- 5% effort

2011-2015 Australian National Health and Medical Research Council (NHMRC) Early Career Overseas Fellowship: "Epistatic and cross-tissue analysis for human gene expression traits"

Total Award: $340,000
Principal Investigator: Di Wu -- 100% effort

Pending

2020-21 UNC Lineberger Developmental Award: "Using early oral biopsy samples and electronic medical records to explore early detection of oral cancer, a pilot study"
Total Award: $100,000
Principal Investigator: Di Wu (July 2020-June 2021) -- 0% effort

PROFESSIONAL SERVICE & SOCIETY MEMBERSHIPS

National and International

2020 Symposium Organizer: American/International Association of Dental Research (AADR/IADR) Washington, DC

2019 Session Organizer: International Chinese Statistical Association (ICSA), Hangzhou, China

2018 Symposium Organizer: International Association of Dental Research (IADR), London, UK

2018 Poster Reviewer: American Medical Informatics Association (AMIA), San Francisco, CA

2018 Oral Session Chair: American Association of Dental Research (AADR), Fort Lauderdale, FL

2018 Poster Session Chair: American Association of Dental Research (AADR), Fort Lauderdale, FL

2012-2015 Mentor: Harvard Graduate Women in Science and Engineering, Cambridge, MA

2014 Oral Session Chair: International Chinese Statistical Association (ICSA), Portland, OR

2014 Judge: American Society of Human Genetics (ASHG), DNA Day Essay Contest, San Diego, CA

2013 Member: Harvard University Postdoc Advisory Board, Cambridge, MA

2012 Session Chair: International Chinese Statistical Association (ICSA), Boston, MA

2012 Coordinator: Sino-American Pharmaceutical Professionals Association New England (SAPA-NE) Annual Conference, Boston, MA

2011 Coordinator: Australia-China Biomedical Research Conference (ACABS), Melbourne, Australia

University of North Carolina-Chapel Hill

2020 Reviewer, COVID-19 Gillings Innovation Laboratory (GIL) proposals, UNC Gillings School of Public Health

2020 Member: Faculty Search Committee, Biomedical Research Imaging Center, Department of Radiology, UNC School of Medicine

2020 Poster Judge: UNC Adams School of Dentistry Dental Research Day

2018, 2019 Member: UNC Bioinformatics and Computational Biology (BCB) PhD Admissions Committee

2018 Member: Tenure Track Faculty Search Committee, jointly in UNC Adams School of Dentistry, Division of Oral and Craniofacial Health Sciences (OCHS), and in the UNC School of Medicine, Department of Orthodontics and Biomedical Research Imaging Center

2018 Poster Judge: Annual Student Research Day, UNC School of Medicine

2018 Speaker: Faculty Research Showcase. UNC Gillings School of Global Public Health

2017 Member: UNC Admissions Committee Biological and Biomedical Sciences Program (BBSP) Foreign Admissions Committee

2017 Faculty: UNC Health Informatics Program (CHIP) Training Program T15

2016 Member: UNC Biostatistics Computing Committee

2016 Member: UNC Admissions Committee for the Master of Professional Science in Biomedical and Health Informatics program, Carolina Health Informatics Program (CHIP)

Editorial boards and peer review activity

2020/2021 Topic Editor for Research Topic "Genetics and Molecular Mechanisms of Oral and Esophageal Squamous Cell Carcinoma" in Journal of *Frontier in Oncology*

2020 Ad hoc Reviewer -- Briefings in Bioinformatics, Bioinformatics (twice), Bioscience Reports, Nucleic Acids Research Cancer

2019 Ad hoc Reviewer -- Proceedings of the National Academy of Sciences

2018 Ad hoc Reviewer – Nature Method, Genetics in Medicine, Genome Biology, Bioinformatics

2017 Ad hoc Reviewer – PLoS Computational Biology, PLoS One, Gene, Nature Scientific Report, Molecules (from MDPI), International Journal of Gynecological Cancer, OncoImmunology, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Genetics in Medicine, Bioinformatics

2016 Ad hoc Reviewer – Statistics in Medicine, PLoS One, Statistics in Biosciences, Journal of the American Statistical Association (JASA), Bioinformatics

2015 Ad hoc Reviewer – Nucleic Acids Research, BMC Bioinformatics, PLoS One, International Journal of Dentistry, PLoS Computational Biology, Bioinformatics, Biometrics

2015 Scientific peer Reviewer -- Australian National Health and Medical Research Council

2014 Ad hoc Reviewer – Biometrics, Nucleid Acids Research, PLoS One, Statistics in Medicine, IEEE/ACM Transactions on Computational Biology and Bioinformatics, Bioinformatics, Briefings in Bioinformatics

2014 Lead Guest Editor -- Cancer Informatics journal supplement (Libertas Academica)

2013 Ad hoc Reviewer – Biomarkers in Cancer, PLoS One, Bionformatics

2012 Ad hoc Reviewer – Nucleic Acids Research, Genomics, Proteomics and Bioinformatics-Elsevier, Bioinformatics

Society Memberships

2016-present Member: American/International Association of Dental Research (AADR/IADR)

2015-present Member: American Statistical Association

2017-2018 Member: American Medical Informatics Association (AMIA)

RESEARCH STATEMENT

The ultimate goal of my academic career is, via genomic data analysis and integration, to better understand disease mechanism and to improve people's health. I am trained as a biostatistician working in the bioinformatics field. My research interest and contributions are in both the statistical methodology development and the application of these methods (among other methods) in biomedical research for scientific discovery and translational medicine. In my statistical bioinformatics group located primarily in the Adams School of Dentistry at UNC,

jointly with the Department of Biostatistics at the Gilings School of Global Public Health, the three main focuses are microbiome metagenomics/metatranscriptomics, single cell data, and cancer genomics (with so far a total of >50 publications).
https://www.ncbi.nlm.nih.gov/myncbi/di.wu.1/bibliography/public/
The data integration pipeline of the electronic medical records from UNC health system, for precision medicine and drug reposition in rare diseases and cancer subtypes, has also been under development.

Before joining UNC, I developed novel statistical methods for genomics data analysis that include gene set analysis (ROAST, CAMERA, method for time course data) and a miRNA array normalization method. These methods have been highly cited in transcriptome studies for pathway analysis and relating datasets by similar expression patterns. My PhD work about breast cancer was to find the cell of origin of different breast cancer molecular subtypes by pathway analysis and gene sets tests across multiple datasets. As a postdoc, I developed a drug repurposing framework for autoimmune diseases based on GWAS risk SNPs and public drug target database, and a bioinformatics data integration framework of drug discovery for lung squamous cell carcinoma (SqCC). At UNC, for the main focuses currently in my statistical bioinformatics group, below are brief yet more detailed descriptions of each:

Focus 1, Microbial sequencing data analysis

1.1 Normalization

There are discussions whether the raw count data or the compositional microbiome data should be used for downstream differential abundance/expression analysis. We compare the performance of seven normalization and differential abundance testing methods used for the analysis of microbiome data to access library size effects. They are evaluated in type I error rates and power using data simulated from a zero-inflated negative binomial distribution. Previous studies have used existing microbiome datasets to compare methods in the presence of batch effects but have not thoroughly addressed the library size issue.

1.2 Differential abundance (DA) analysis and differential expression (DE) analysis

Microbial association testing between groups is not straightforward due to the sparsity (i.e., zero inflation) and high dimensionality of these data, as well as their inherent hierarchical compositional structure. Importantly, no statistical methods or mature pipelines have been specifically developed for differential expression analysis for metatranscriptomics. We posit that metagenomics and metatranscriptomics count data likely follow the same types of distributions with different parameters. To evaluate the computational methods for differential abundance (DA) in metagenomics and differential expression (DE) analysis in metatranscriptomics, counts will be simulated through a novel strategy, flexible enough to capture both zero inflation and over-dispersion, using parameters obtained from the real data. The optimized pipeline for metagenomics/metatranscriptomics will be applied to identify ECC associated bacteria, bacterial genes and pathways.

1.3 microbiome data integration (cross-sectional omics data)

Zero inflation compounds put the challenges of joint and integrative analyses of matched metagenomics and metatranscriptomics data. This hampers the detection of differential relative biological activity (i.e., gene expression over abundance) of bacterial taxa and genes between phenotypes. Importantly, no statistical methods have been developed specifically for this context

to allow for valid inferences. We propose the conduct of joint metagenomics and metatranscriptomics analyses to characterize the ECC related supragingival biofilm dysbiosis among a community based sample of 170 children, enrolled in a large-scale investigation of early childhood oral health in North Carolina (parent study: ZOE 2.0; NIH/NIDCR U01DE025046; PI: Divaris at Adam School of Dentistry at UNC). To our knowledge, this pilot dataset is the largest with matched metagenomics and metatranscriptomics data available in the oral health domain, offering improved power to detect putative associations of microbial composition and transcriptome with ECC. We use RNA/DNA ratio to represent relative activity and to accommodate the matched data structure. As modeling zero-inflated compositional data ratios (or normalized counts) is challenging, we are developing a two-step procedure to decompose the mixture of distributions. Upon completion, we anticipate that the study will provide novel insights into the microbial basis of ECC. 1.2 and 1.3 are supported with an R03.
Our group also participates the cohort study of (Baby Connectome Program) BCP-enriched through collaboration with Dr. Weili Lin group at Biomedical Research Imaging Center (BRIC), UNC, to integrate longitudinal fecal microbiome data with brain imaging, behavior measurements and nutrition for babies of 3month-5 years.

Focus 2, Single cell RNAseq

We are also developing various statistical methods in scRNAseq clustering analysis, differential expression analysis and pathway analysis (applied for understanding cancer initiation/progression, and characterizing HIV infected mouse).

2.1 Differential expression analysis: Two-Sigma

This method accommodates a correlation structure between cells from the same biological sample via random effect terms. It also allows for overdispersed and zero-inflated counts. Simulations studies show that TWO-SIGMA outperforms alternative regression-based approaches in both type-I error control and power enhancement when the data contains even moderate within-sample correlation. Dr. Yun Li and I supervised our PhD Eric Van Buren to work on this and published it. An R package TwoSigma is available on GitHub. We further extend TWO-SIGMA competitive gene set testing as TWO-SIGMA-G, where gene-level statistics are collected from the TWO-SIGMA model. It not only holds type I errors as gene set tests developed for bulk RNAseq, also achieves improved power by fitting the proper distribution, with draft available at www.biorxiv.org. An R package TwoSigma and Two-Sigma-G is available on GitHub.

2.2 Pathway analysis: BZINB

Measuring gene-gene dependence in single cell RNA sequencing (scRNA-seq) count data is often of interest and remains challenging, because an unidentified portion of the zero counts represent non-detected RNA due to technical reasons. Conventional statistical methods that fail to account for technical zeros incorrectly measure the dependence among genes. To address this problem, we propose a bivariate zero-inflated negative binomial (BZINB) model constructed using a bivariate Poisson-gamma mixture with dropout indicators for the technical (excess) zeros. Compared to existing models, the proposed BZINB model is specifically designed for estimating dependence and is more flexible, while preserving the marginal zero-inflated negative binomial distributions. An R package bzinb is available on CRAN and GitHub. I supervised PhD student Hunyong Cho to work on this and just have the draft submitted. The draft is also available https://www.biorxiv.org/content/10.1101/2020.03.06.977728v1 2.3 A novel computational pipeline for single cell RNA-seq analysis in HIV related study with close

collaboration with Dr. Lishan Su at UNC, the proposed study aimed to understand the effects IFN-I on HIV-induce persistent inflammation, immune dysfunction and viral persistence.

We hypothesize these effects can be detected at the single cell level using bioinformatics tools. We propose to detect how HIV infection and persistent IFN-I signaling change the composition of cell populations, gene expression patterns, cell-cell interactions and pathways. Besides the regular computational tools for standard questions answered via scRNASeq, we will apply the computational methods (Two-Sigma for DE analysis, Two-Sigma-G and BZINB for pathway analysis) we have developed to analysis the data to achieve a full spectrum of understanding the effects of HIV and IFN-I signaling on chronic inflammation, T cell dysfunction and to ultimately help clinical decision. The data analysis methods specifically developed for cell populations in immune system can further reveal the change of trajectory and cell-cell talk affected by HIV-1 infection and IFN-I signaling between and within cell populations. We will integrate these methods as a novel pipeline to analyze scRNAseq data to answer HIV related immune questions, so other researchers or clinician scientists can follow it easily. This work is funded by UNC computational Medicine program. Beyond single cell data analysis, we also developed statistical causal pathway method for omics data using the framework of directed acyclic graph (DAG), that may potentially be used in scRNAseq omics data.

Focus 3, Cancer data integration for precision medicine

At UNC, I have collaborated with Professor Cyrus Vaziri in SOM on the relation between DNA repair pathways and cancer mechanism. We have successfully obtained three R01s (co-I).
For details, we analyzed local mouse WES data and integrated TCGA data (Draft available, corresponding author). Rad18 gene promotes a damage-tolerant and error-prone mode of DNA replication termed Trans-Lesion Synthesis (TLS) that is pathologically activated in neoplastic cells and therefore represents a potential source of mutations in cancer. However, the impact of vertebrate Rad18 on mutagenesis has never been tested in a chromosomal setting and the extent to which Rad18 impacts cancer genomes is not known. We have performed whole exome sequencing (WES) to analyze carcinogen DMBA-induced skin tumors generated in Rad18+/+ and Rad18-/- mice. We show that Rad18 mediates specific mutational signatures characterized by high levels of AT>TA Single Nucleotide Variations (SNVs). Comparison of mouse tumor genomes with human mutational signatures shows that COSMIC signature 22 predominates in the mutational portrait of Rad18+/+ tumors. In contrast, Rad18-/- tumors are characterized by increased contribution of COSMIC signature 3 (a hallmark of BRCA-mutant tumors). Analysis of TCGA data shows that RAD18 expression is strongly associated with high SNV burdens in human Lung Adenocarcinoma, Lung Squamous Cell Carcinoma and Bladder Cancer, suggesting that RAD18 also promotes mutagenesis in human cancers. Taken together our results show that Rad18 promotes mutagenesis in vivo, modulates DNA repair pathway choice in neoplastic cells and impacts the levels of specific mutational signatures that are relevant to human tumors.

We are now extending the strategy to HPV-negative oral cancer studies. Recently, we also connect between our local dental Electronic Patient Records (EPR) and their Electronic Medical Records (EMR) to define the cohorts with early oral pathology samples and high risk of oral cancer, potentially for risk prediction, drug repurposing and biomarker identification. In addition to my three main primary research focuses, my broad and extensive collaboration brought me collaborators who generate real data for our data inspired methodology development, and allowed me to be involved in many cutting edges biological discovery including genetics,

immunity and translational medicine with being co-authors in high profiling journals.

TEACHING STATEMENT

I have been a student of many excellent educators and I remember how I felt when I attended their lectures and discussed subject matter with them. This inspires me to be one of them. I have always enjoyed working with students, in the classroom or for projects.

In the classroom at UNC, I have been invited to give lectures in the Adams School of Dentistry (ASOD), School of Medicine, Gillings School of Global Public Health, School of Information and Library Science, and I present annual lectures at the ASOD on genomic data analysis. I teach this subject as well in the Biostatistics Department and the Bioinformatics and Computational Biology (BCB) program. I teach "gene set testing", and "genomics data based drug repurposing" in the Carolina Health Informatics Program (CHIP).

In the following paragraphs, I will mainly focus on classroom teaching. The goal of teaching is to pass knowledge from teachers to students, and to bridge between knowledge and creative thinking.

I hope that what my students learn from me is useful for their future problem solving and will enhance their critical thinking. Teaching is also bi-directional, as students can inspire the teacher. The teaching process often causes teachers to think in new ways about familiar concepts.

Students may have quite different goals, e.g., for curiosity, for pleasure of knowing, for achieving a satisfactory grade, or just for fulfilling degree requirements. Spontaneous interest is the best teacher; however, being encouraged can help students to initiate interest. Sometimes discussion of the history of a subject, its broader utility, and its relation to other disciplines can give a student an appreciation beyond what they get directly from the subject itself.

I prepare teaching material using different approaches. For example, I talk with previous students about their opinions, communicate with colleagues who teach other related classes, or even practice in front of my friends or family. I also take advantage of available websites for a subject or a relevant topic to find the best way to explain and the stories/usage to impress students. Of course, the accuracy of the information from websites needs to be confirmed.

I find that it is helpful to discuss a simple example when a new difficult concept is introduced. So I give students time to ask questions in the class, and encourage them to discuss. The teacher can help with a few leading questions. For example, shall the algorithms used in one question be used in any other situation? How does a new problem relate to other problems? Are there different ways to solve the new problem and what are their advantages or disadvantages? How do biostatistics problems connect with real life questions?

I find solving problems in a small group is very helpful, as the students can learn from each other in a more flexible way. Interestingly, I find it important to remember the students' names, their majors and their projects. Students often come from different educational backgrounds. It is critical to have a good idea of their general education, what they have learned, why they come to class and how I can help them to achieve their goals. To some extent, the class and lecture need to be customized for different audiences. A pre-class survey can help to understand what customization is appropriate.

I find that the first time to teach a new class can be especially difficult. Also, large classes can be particularly hard to handle since the students tend to lose focus more easily than in a small class. If one considers that a class is like a piece of music, rhythm is very important. You have to keep

up the pace. Don't let anyone get bored. Yet, sometimes, we need to slow down a little to ask students questions to help them focus.

When I teach, I have the perspective of a (nearly) full-time researcher. What I think most important is to motivate the students to think independently and to give them support generously. A mentor should let the student know the key point in the project and care that each student – and the group as a whole - achieve understanding.

In summary, there are three reasons that qualify me to teach. First, I already have intensive experience working with medical students, and abundant experience supervising both undergraduate and graduate students. Second, I come from a multi-disciplinary background so I can communicate and understand students with different knowledge structure. Third, I am patient, very responsible in preparing lectures and I am enthusiastic about teaching, both in the classroom and in a computational biology laboratory.

At UNC, I have also spent much time supervising graduate students in their research and thesis/dissertations. From 2015-2016, I mainly worked with graduate research assistant GRA MS and PhD students. In 2017, I started mentoring PhD students with their dissertation work. In my current group of 14 students, I am mentoring five PhD candidates for their dissertations, one PhD GRA (1st author paper published in 2021 NAR Cancer, and co-authored in Immunity), two first year PhD students, four MS students and one undergraduate student. Three PhD candidates have passed their preliminary exams. For the earliest two PhD candidates, both have published 1st-author methodology papers; one is now a postdoc fellow at Harvard Biostatistics; another decides to stay in academia and has a good chance to get a faculty position next year. For the graduated MS students, three of them work at Massachusetts General Hospital, Washington University and Medical College of Wisconsin as data analysts; One is enrolled in the PhD program. The undergraduate student I have supervised also became a PhD at UNC Biostatistics. To supervise these students, I schedule regular weekly individual meetings and lab meetings. I encourage them to gradually think and work more independently and explore different way to solve a problem. It's a challenging journey to complete a PhD dissertation for both the student and mentor, in terms of choosing the right topics and how to transit what students learn from classroom to a research topic. During a survey of teaching styles in the department, I realized I am a combined coach and listener. Knowing this is helpful to guide what's to be improved. The challenge to supervise MS students/honor-undergraduate is the limited time frame. I keep adjusting feasibility and what they can learn.

Finally, a task that we as scientists share is to communicate with and educate the general community regarding our scientific goals, our degree of commitment, and our achievements. I am passionate to do so.