Bari J. Ballew, Ph.D.

https://github.com/bballew

(973) 713-6549
bari.ballew@gmail.com

3708 Massachusetts Avenue
Washington, DC 20016

Experienced human genomics scientist with in-depth knowledge of DNA sequencing data analysis. Proven track record leading teams in developing scalable, production-quality, reproducible bioinformatic pipelines by incorporating software development best practices. Dedicated practitioner of servant leadership.

PROFESSIONAL EXPERIENCE

Senior Bioinformatician
54gene, Washington, DC 01/2021-present

- Served as AWS expert for genomics use cases. Configured, deployed, and administered an on-demand virtual HPC via AWS ParallelCluster for the Genomics and Data Science team. Maintained S3 buckets, including implementation of lifecycle rules for cleanup and object archival, and configuration of bucket permissions to maintain secure access during data sharing collaborations.
- Developed a WGS germline variant calling pipeline that encompasses the GATK best practices guidelines, incorporates QC metrics including FastQC metrics before/after trimming and adapter removal, estimated sample contamination, inferred/reported sex concordance, unexpected duplicate samples, approximate relatedness assessment, and automated sample exclusion based on coverage, het/hom ratio, and contamination. The pipeline is scalable, with user-configurable memory settings for all dependencies and special consideration for steps that require high filesystem throughput, high memory requirements, high file I/O, etc. The entire workflow is portable and self-contained, using conda to create per-rule environments where dependencies are automatically installed when the pipeline is run.
- Oversaw the hiring process that resulted in doubling our team's size (including one direct report). Wrote up technical best practices guidance, code of conduct, and onboarding instructions for the team. Set group priorities and implemented quarterly reviews/goal setting. Created a collaboration-friendly working environment, including pair programming, asynchronous code review, and academic-style large group code review.
- Co-wrote an R package for non-destructively and reproducibly cleaning extremely messy questionnaire data, enforcing data types, identifying outlier samples/variables, performing initial exploratory analysis, and creating derived variables (e.g. applying inverse normal transform). The package includes over 110 unit tests to ensure robustness.
- Co-developed a variety of other pipelines for reproducible analysis, including downstream VCF filtering, benchmarking against NA12878 truth datasets, applying annotation via VEP, exploring CYP2D6 star alleles, detecting potentially pathogenic repeat expansions, and performing genotyping array QC.

Manager, Bioinformatics Development and Analysis
Cancer Genomics Research Laboratory, NCI/Leidos Biomedical Research, Inc. 07/2020-01/2021

- Led a team of seven senior bioinformaticians responsible for all pipeline development for CGR and for advanced analysis support beyond pipelined deliverables
- Worked closely with PIs to guide experimental design and selection of appropriate controls, to ensure a correct match between biological question and experimental technique, and to define and prioritize analysis, development, and data management requirements
- Enhanced professionalism and reproducibility across the team by using style guides and linters, standardized directory structure using cookiecutter, environment management via conda, and continuing our previous efforts with version control and testing
- Conscientiously shaped a team culture of support, mentorship, and safety to take risks, despite being fully remote during a pandemic, by curating both formal opportunities for interaction (lab meeting, code review, journal club, lunch-and-learn) and informal ones (asynchronous discussion of pertinent blog posts, challenging errors/troubleshooting, design decisions, etc.)
- Formalized policies and processes for conference attendance, one-on-one meetings, skip-levels, career development support, and quarterly reviews/goals

Bioinformatics Scientist
Cancer Genomics Research Laboratory, NCI/Leidos Biomedical Research, Inc. 01/2017-06/2020

- Led the four-person microbiome pipeline development team in constructing a QIIME-based pipeline coordinated through Snakemake; instituted best practices including version control, code review, and unit, regression, black box, and acceptance testing; optimized pipeline using mock communities as benchmarks
- Wrote a modular, platform-agnostic pipeline to coordinate structural variant calling using multiple containerized callers, then annotate and compare results across callers; awarded $20,000 for development and deployment to AWS and Google Cloud platforms
- Designed an ensemble germline variant calling pipeline using GATK HaplotypeCaller and Google's DeepVariant, using GLnexus to merge gVCFs over large cohorts; combined existing VCF merging tools (bcftools) with custom Python scripts to generate an output VCF that maintains all tags from the original two callers; benchmarked with GiaB and SynDip datasets to determine precision/recall for variants called by both callers vs. only one caller
- Developed a pipeline to implement non-negative matrix factorization to detect mutational signatures in somatic sequencing data; discovered evidence for a novel risk factor in esophageal cancer via the signatures detected
- Taught workshops on pipeline development with Snakemake and on data analysis with Python's pandas package (see GitHub repositories for interactive tutorials)

Genome Scientist
Personal Genome Diagnostics, Baltimore, MD 01/2016-12/2016

- Developed, implemented, and maintained automated report generation tool, eliminating sources of manual error and reducing turnaround time for reporting NGS results from hours to seconds
- Wrote Perl scripts to facilitate extraction and formatting of required data from raw pipeline output to concise, human-readable summary reports
- Collaborated closely with software engineers to establish requirements, regression testing strategy, and user acceptance testing plans for CLIA reports and related software
- Evaluated the effect of switching from SnpEff to VEP annotation databases on our pipeline with regard to improving sensitivity and accuracy
- Instituted and led weekly Genome Sciences team training sessions covering use of UNIX and NGS analysis tools to facilitate data analysis

Research Fellow, Dr. Sharon A. Savage Laboratory
Division of Cancer Epidemiology and Genetics, NCI, NIH, Rockville, MD 03/2012-12/2015

- Wrote a set of custom Perl and Bash tools to facilitate whole exome sequencing data analysis of both pedigrees and large populations on the NIH compute cluster
- Identified novel causal mutations in patients with inherited cancer predisposition syndromes, resulting in numerous collaborations and nine publications (five first or co-first author)
- Authored proposals that were awarded over $60,000 in funding to leverage the DCEG biospecimen collection via innovative NGS analyses

Doctoral Candidate, Dr. Vicki Lundblad Laboratory
Salk Institute for Biological Studies/UC San Diego, La Jolla, CA 09/2005-12/2011

- Led two major projects to explore mechanisms by which telomeres are protected from being misinterpreted by the cell as DNA breaks, resulting in first- and middle-author publications in peer-reviewed journals

TECHNICAL PROFICIENCIES

- NGS data analysis: bedtools, vcftools, bcftools, GATK, IGV, QIIME, PLINK, etc.
- Workflow management languages: Snakemake
- Containerization: Singularity, Docker
- Jupyter notebooks
- Version control: git, GitHub, GitLab
- Cloud computing platform: AWS
- Cloud environments for collaboration: Google Colab, Binder
- Package/environment managers: anaconda/mamba, pip, poetry, homebrew, yum
- Virtual machines: Vagrant
- Documentation: Sphinx, Roxygen

- Programming/scripting languages (in order of decreasing proficiency): Python, bash, R, Perl, VBA, MATLAB, and C++
- Testing frameworks: pytest, testthat
- Variant annotation tools: ANNOVAR, SnpEff, VEP, dbNSFP
- *NIX operating systems and Grid Engine- and Slurm-based HPCs
- Genomic/cancer biology data standards: HGVS, SNOMED, GO ontology
- Public genomic databases: gnomad, ESP, ExAC, 1000 Genomes, dbSNP

EDUCATION

Ph.D., Biology — **University of California, San Diego**, San Diego, CA, 2011
B.S., Molecular and Cellular Biology — **Johns Hopkins University**, Baltimore, MD, 2005

SELECTED PUBLICATIONS

Yeager M, Machiela MJ, Kothiyal P, Dean M, Bodelon C, Suman S, Wang M, Mirabello L, Nelson CW, Zhou W, Palmer C, **Ballew B**, Colli LM, Freedman ND, Dagnall C, Hutchinson A, Vij V, Maruvka Y, Hatch M, Illienko I, Belayev Y, Nakamura N, Chumak V, Bakhanova E, Belyi D, Kryuchkov V, Golovanov I, Gudzenko N, Cahoon EK, Albert P, Drozdovitch V, Little MP, Mabuchi K, Stewart C, Getz G, Bazyka D, Berrington de Gonzalez A, Chanock SJ. Lack of transgenerational effects of ionizing radiation exposure from the Chernobyl accident. ***Science.*** doi: 10.1126/science.abg2365 (2021).

Bai J, Shi J, Li C, Wang S, Zhang T, Hua X, Zhu B, Koka H, Wu HH, Song L, Wang D, Wang M, Zhou W, **Ballew BJ**, Zhu B, Hicks B, Mirabello L, Parry DM, Zhai Y, Li M, Du J, Wang J, Zhang S, Liu Q, Zhao P, Gui S, Goldstein AM, Zhang Y, Yang XR. Whole genome sequencing of skull-base chordoma reveals genomic alterations associated with recurrence and chordoma-specific survival. ***Nat Commun.*** doi: 10.1038/s41467-021-21026-5 (2021).

McReynolds LJ, Wang Y, Thompson AS, **Ballew BJ**, Kim J, Alter BP, Hicks B, Zhu B, Jones K, Spellman SR, Wang T, Lee SJ, Savage SA, Gadalla SM. Population frequency of Fanconi pathway gene variants and their association with survival after hematopoietic cell transplantation for severe aplastic anemia. ***Biol Blood Marrow Transplant.*** doi: 10.1016/j.bbmt.2020.01.011 (2020).

Wang Y, McReynolds LJ, **Ballew BJ**, Katki HA, Dagnall CL, Haagenson MD, Spellman SR, Wang T, Hicks B, Zhu B, Freedman ND, Jones K, Wang D, Lee SJ, Savage SA, Gadalla SM. Pre-HCT Telomere Abnormalities and Mortality after Unrelated Donor Hematopoietic Cell Transplant for Severe Aplastic Anemia. ***Biol Blood Marrow Transplant.*** 25(3):S417 (2019)

Gadalla SH*, **Ballew BJ***, Haagenson M, Spellman S, Hicks B, Alter BP, Zhu B, Zhou W, Yeager M, Wang T, Fleischhauer K, Hsu K, Verneris M, Freedman N, Lee SJ, Savage SA. Germline mutations in marrow failure predisposition genes in patients receiving unrelated donor hematopoietic cell transplant for severe aplastic anemia. ***J Clin Invest.*** 128(22):68 (2016). *authors contributed equally

Burris AM*, **Ballew BJ***, Kentosh JB*, Turner CE, Norton SA, NCI DCEG Cancer Genomics Research Laboratory, NCI DCEG Cancer Sequencing Working Group, Giri N, Alter BP, Nellan A, Gamper C, Hartman KR, Savage SA. Hoyeraal-Hreidarsson syndrome due to *PARN* mutations: fourteen years of followup. ***Pediatr Neurol***. doi: 10.1016/j.pediatrneurol.2015.12.005 (2016). *authors contributed equally

Kocak H*, **Ballew BJ***, Bisht K*, Eggebeen R, Hicks BD, Suman S, O'Neil A, Giri N, NCI DCEG Cancer Genomics Research Laboratory, NCI DCEG Cancer Sequencing Working Group, Maillard I, Alter BP, Keegan CE, Nandakumar J, Savage SA. Hoyeraal-Hreidarsson syndrome caused by a germline mutation in the TEL patch of the telomere protein TPP1. ***Genes Dev.*** 28(19):2090-102 (2014). *authors contributed equally

Ballew BJ*, Joseph V*, De S*, Sarek G, Vannier JB, Stracker T, Schrader KA, Small TN, O'Reilly R, Manschreck C, Harlan Fleischut MM, Zhang L, Sullivan J, Stratton K, Yeager M, Jacobs K, Giri N, Alter BP, Boland J, Burdett L, Offit K, Boulton SJ, Savage SA, Petrini JH. A recessive founder mutation in regulator of telomere elongation helicase 1, *RTEL1*, underlies severe immunodeficiency and features of Hoyeraal Hreidarsson syndrome. ***PLoS Genet*** 9(8):e1003695 (2013). *authors contributed equally

Ballew BJ, Yeager M, Jacobs K, Giri N, Boland J, Burdett L, Alter BP, Savage SA. Germline mutations of regulator of telomere elongation helicase 1, *RTEL1*, in Dyskeratosis congenita. ***Human Genetics*** 132(4):473-80 (2013). **most frequently cited article of 2013