BIOGRAPHICAL SKETCH: C. Titus Brown

NAME: C. Titus Brown ****

eRA COMMONS USER NAME: BROWNCT

POSITION TITLE: Associate Professor, UC Davis

EDUCATION/TRAINING:

| INSTITUTION AND LOCATION | DEGREE  (if applicable) | Completion Date  MM/YYYY | FIELD OF STUDY |
| --- | --- | --- | --- |
|  |  |  |  |
| Reed College, Portland, OR | BA | 6/1997 | Mathematics |
| California Institute of Technology, Pasadena, CA | PhD | 6/2007 | Developmental Biology and Genomics |
| California Institute of Technology, Pasadena, CA | - | 5/2008 | Developmental Biology and Genomics |

# A. Personal Statement

I am focused on advancing biology through better, more flexible, and more advanced data analysis – what I call “data-intensive biology”. My primary research interest is **developing methods to generate and refine hypotheses using high-throughput sequencing data** from genomes, transcriptomes, and metagenomes; my group has created and maintains a number of open source software packages at **github.com/dib-lab/.** We use **open source and open science** methodologies to disseminate our methods as broadly as possible, and I write a widely-read blog on data-intensive methodologies. I am heavily invested in training research active scientists and run many **short-form intensive workshops** **on data analysis** for graduate students and postdocs; I work closely with the **Software and Data Carpentry** communities. I am increasingly interested in using training, user engagement, and community outreach to **grow communities of practice** in biomedical data science, which led to my role as co-coordinator for the **NIH Data Commons Pilot Phase Consortium** (2018) and Engagement Lead for the NIH Common Fund Data Ecosystem project (2019-present).

# B. Positions and Honors

# 2008-2014 Assistant Professor, Computer Science and Engineering / Microbiology and Molecular Genetics, Michigan State University, East Lansing, MI.

2015-present Associate Professor, School of Veterinary Medicine, University of California Davis.

2014-2019 Moore Foundation Data-Driven Discovery Investigator.

2018 co-coordinator, NIH Data Commons Pilot Phase Consortium.

2010-present Director, two-week intensive workshop in Analyzing Next-Generation Sequencing Data (except 2016)

2013- Scientific Advisory Board member, iPlant Collaborative/CyVerse, 2013-present.

2015-2018 ELIXIR-UK Scientific Advisory Board

2016-2017 Faculty, MBL Microbial Diversity course.

2012-present Lecturer, MBL STAMPS course.

2006-2008 TA and Lecturer, MBL Embryology course.

# C. Contribution to Science

*ORCID profile:* [*http://orcid.org/0000-0001-6001-2677*](http://orcid.org/0000-0001-6001-2677)

*ImpactStory:* [*https://profiles.impactstory.org/u/0000-0001-6001-2677*](https://profiles.impactstory.org/u/0000-0001-6001-2677)

*Google Scholar:* [*https://scholar.google.com/citations?hl=en&user=O4rYanMAAAAJ*](https://scholar.google.com/citations?hl=en&user=O4rYanMAAAAJ)

1. Scalable data structures and algorithms for data analysis and de novo assembly. (2008-present)

My lab is interested in scaling *de novo* metagenome and transcriptome assembly methods to the extremely large data sets generated by shotgun sequencing technologies. We have developed **novel data structures** (Pell et al., 2012), implemented advanced data structures in a fast and reusable **open library**, **built novel algorithms** (Brown et al., 2012; Brown et al., 2019), and applied these novel data structures and algorithms to **previously unresolvable data sets** such as immense and diverse soil metagenomes (Howe et al., 2014). Our khmer and sourmash software packages are widely used (1000-10,000 users), and have helped drive the field of metagenome assembly towards considerably more powerful techniques. More recently we have been exploring **mapping-free variant** calling (Standage et al, 2019) and **graph-based approaches** to metagenomics (Brown et al., 2020).

1. *Scaling metagenome sequence assembly with probabilistic de Bruijn graphs*. (2012)

Pell, Hintze, Canino-Koning, Howe, Tiedje, **Brown**. doi: 10.1073/pnas.1121464109

1. *Tackling soil diversity with the assembly of large, complex metagenomes* (2014).

Howe, Jansson, Malfatti, Tringe, Tiedje, **Brown**. doi: 10.1073/pnas.1402564111.

1. *A Reference-Free Algorithm for Computational Normalization of Shotgun Sequencing Data* (2012).

**Brown,** Howe, Zhang, Pyrkosz, Brom. preprint arXiv:1203.4802

1. *Kevlar: a mapping-free framework for accurate discovery of de novo variants* (2019).

Standage, **Brown**, Hormozdiari. doi: 10.1101/549154.

1. *Exploring neighborhoods in large metagenome assembly graphs using spacegraphcats reveals hidden sequence diversity* (2029).

**Brown**, Moritz, O’Brien, Reidl, Reiter, Sullivan. Accepted, Genome Biology, May 2020. preprint at doi: 10.1101/462788.

2. Applied genomics and bioinformatics. (2008-present)

Sequencing is increasingly making the genomics of animals available for detailed study, and I have participated extensively in **genomic and transcriptomic analysis** of a wide variety of metazoa -- sea urchin, lamprey, ascidians, coral, chick, and single-celled eukaryotes. With my move to the UC Davis School of Veterinary Medicine in 2015 and my membership in the UC Davis Genome Center, I am increasingly involved in genome- and transcriptome-wide analysis of complex genomes, including the Arabidopsis, horse, and dog genomes. I am also involved in several collaborations around GWAS for dog genetic diseases as well as dog tumor genome and transcriptome analysis.

1. *Tissue resolved, gene structure refined equine transcriptome* (2017).

Mansour et al., **Brown.** doi: 10.1186/s12864-016-3451-2.

1. *Centromere location in Arabidopsis is unaltered by extreme divergence in CENH3 protein sequence.* (2017).

Maheshwari, Ishii, **Brown,** Houben, Comai. doi: 10.1101/gr.214619.116.

1. *Whole genome variant association across 100 dogs identifies a frame shift mutation in DISHEVELLED 2 which contributes to Robinow-like syndrome in Bulldogs and related screw tail dog breeds* (2018).

Mansour et al., **Brown,** Bannasch. doi: 10.1371/journal.pgen.1007850.

1. *Re-assembly, quality evaluation, and annotation of 678 microbial eukaryotic reference transcriptomes* (2019).

Johnson, Alexander, **Brown**. doi: 10.1093/gigascience/giy158.

1. *Draft genome assemblies using sequencing reads from Oxford Nanopore Technology and Illumina platforms for four species of North American Fundulus killifish.* Johnson, Sahasrabudhe, Gill, Roach, Froenicke, **Brown**, Whitehead. doi: 10.1093/gigascience/giaa067

3. Robust and open scientific software development. (2000-present)

For 20 years, I have championed a central role for **good, open, and sustainable software development** **practices in science**. As software becomes ever more important for data analysis and modeling in science, robust, reusable, and remixable software can accelerate the practice of both computational and experimental science. As a participant in the Carpentries training program, as a blogger, and as a reviewer of bioinformatics software papers, I have helped develop and drive better software practice. Our paper on “Best Practices for Scientific Computing” (Wilson et al., 2014) was the most downloaded paper at PLOS Biology that year, and has received over 200 citations. I routinely engage with other software developers online and in workshops to help advance the methods, process, and culture of scientific computing in biology.

1. *Best Practices for Scientific Computing* (2014).

Wilson et al. doi: 10.1371/journal.pbio.1001745.

1. *The khmer software package: enabling efficient nucleotide sequence analysis* (2015).

Crusoe et al., **Brown.** doi: 10.12688/f1000research.6924.1.

1. *Walking the Talk: Adopting and Adapting Sustainable Scientific Software Development processes in a Small Biology Lab* (2016).

Crusoe and **Brown**. doi: 10.5334/jors.35.

1. *Channeling Community Contributions to Scientific Software: A Sprint Experience* (2016).

Crusoe and **Brown**. doi: 10.5334/jors.96

1. *sourmash: a library for MinHash sketching of DNA.*

Irber and **Brown**. doi: 10.21105/joss.00027.

1. *Large-scale sequence comparisons with sourmash*

Pierce, Irber, Reiter, Brooks, Brown. doi: 10.12688/f1000research.19675.1

4. Training and teaching in data-intensive biology. (2008-present)

I believe that one of the biggest challenges facing biology today is the lack of basic computational expertise amongst early career researchers in biology. To help address this problem, I’ve developed over 20 different peer-instruction workshops and short courses, and taught over 40 workshops in the last 6 years. This includes my flagship two week NIH-funded summer course, “Analyzing Next-Generation Sequencing Data” (2010-present), as well as several short courses on sequence analysis, data management, and over a dozen Software Carpentry-like short courses. All of my course materials and presentations are posted online and available under an open Creative Commons license (e.g. angus.rtfd.io/), and receive well over 100,000 “hits” a year. At my new position at UC Davis, I have been specifically tasked with significantly expanding my training program in data intensive biology, and in 2017 (and onwards) we expanded our summer training program significantly to run a ~100-150 person summer institute, [ivory.idyll.org/dibsi/.](http://ivory.idyll.org/dibsi/)

5. Open science (2008-present)

I have worked for over a decade to increase the amount and effectiveness of openness in scientific practice, in part by practicing it myself (posting preprints, making source code available, and implementing open and reproducible paper pipelines). My experiences in open practice have led me to join with others in advocating for specific research policies and practices.

1. *Point of View: How open science helps researchers succeed* (2016).

McKiernan et al. doi: doi: 10.7554/eLife.16800.

1. *Sustainable computational science: the ReScience initiative* (2017).

Rougier et al. doi: 10.7717/peerj-cs.142.

1. *The principles of tomorrow's university* (2018).

Katz et al. doi: 10.12688/f1000research.17425.1.

# D. Research Support

Ongoing:

1. Moore Data Driven Discovery Investigator Award (2014-2021). This five-year career award is aimed at the development of new techniques and approaches for harnessing large biological data sets. It explicitly supports the development of robust, general software for sharing both raw data and analyses, as well as field-specific software development. I am the sole PI.
2. NIH Common Fund Data Ecosystem OT Awardee (2019-Dec 2020). I am a subawardee (to Owen White at UMB) and act as Engagement Team Co-lead on an effort to build a FAIR ecosystem for Common Fund data sets.
3. NSF DBI-2018911 Aug 2020-July 2023. “BBSRC-NSF/BIO:Collaborative Research: genomeRxiv: a microbial whole-genome database and diagnostic marker design resource for classification, identification, and data sharing.” PI.
4. 19. USDA 2020-67015-31675 July 2020-June 2023. “Tools and resources for cattle pangenomics.” PI.

Completed:

5. NSF OCI Supplement to NSF BEACON STC, “Materials and Workshops for Cyberinfrastructure Education in Biology” (2013-2016). This supplement to the BEACON STC supports collaboration between several NSF BIO centers (including iPlant and C-MORE) to develop workshops for data-intensive biology. The Data Carpentry workshop series ([datacarpentry.org](http://www.datacarpentry.org/)) emerged partly from this funding. I am the sole PI.

6. NSF ABI III: III: Small: Collaborative Research: Supporting Efficient Discrete Box Queries for Sequence Analysis on Large Scale Genome Databases (2013-2016). Collaborative work with computer scientists on kmer databases. Co-PIs: Qiang Zhu (U. Michigan) and Sakti Pramanik (Michigan State University).

7. USDA NIFA research award, “Improvement of the Chicken Reference Genome” (2013-2016). Supports improvement of the chicken reference genome. Primary PI: Wes Warren. I was involved in sequencing data integration and analysis.

8. NIH BIGDATA R01, “Low-memory Streaming Prefilter for Biological Sequencing Data” (2013-2016). This supports our software and algorithm development for sequence analysis, with the goal of enabling very scalable analysis of large sequencing data sets. I am the sole PI.

9. NIH Data Commons Pilot Phase Consortium OT Awardee (2017-2019). This “Other Transactions” award supports a leadership role in the definition and creation of the NIH Data Commons.