The people of the 24th annual Pacific Symposium on Biocomputing in Hawaii

This manuscript (permalink) was automatically generated from dhimmel/psb-manuscript@29870e2 on January 5, 2019.

Authors

- **Weixuan Fu**
  0000-0002-6434-5468 · weixuanfu · weixuanfu
  Department of Biostatistics, Epidemiology and Informatics, Institute for Biomedical Informatics, University of Pennsylvania

- **Casey Greene**
  0000-0001-8713-9213 · cgreene · greenescientist
  Department of Systems Pharmacology & Translational Therapeutics, University of Pennsylvania; Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation

- **Daniel Himmelstein**
  0000-0002-3012-7446 · dhimmel · dhimmel
  Department of Systems Pharmacology & Translational Therapeutics, University of Pennsylvania

- **Trang Le**
  0000-0003-3737-6565 · trang1618 · trang1618
  University of Pennsylvania

- **Jaclyn Taroni**
  0000-0003-4734-4508 · jaclyn-taroni · jaclyn_taroni
  Childhood Cancer Data Lab, Alex's Lemonade Stand Foundation

Abstract

Manubot is an open source tool for writing manuscripts on GitHub in markdown format. Manubot applies the git-based software workflow to scholarly writing, enabling enhanced transparency, collaboration, automation, and reproducibility.

This manuscript is the result of a *special working group* at the 2019 Pacific Symposium on Biocomputing that will introduce attendees to collaborative writing with Manubot. Each conference attendee is invited to write a small blurb on themselves and their research, by submitting a pull request to the manuscript repository at https://github.com/dhimmel/psb-manuscript.

The working group also covers how to write your next manuscript using Manubot and what features of Manubot can help biomedical researchers document and publish their computational research. For example, Manubot enables citation by persistent identifier to automate bibliographic metadata retrieval and formatting as well as allowing templating so results can be directly inserted from the analyses that produced them.

Methods

In this section, PSB 2019 attendees are asked to contribute a bit about themselves and their research. As part of the special working group, we thought this would be a helpful activity to introduce biocomputational scientists to writing with Manubot. For inspiration, here are some prompts:

- Introduce yourself briefly.
- What do you research? Include any relevant links to your lab or personal website.
- What is your favorite study from your career or what study was your biggest discovery?
- What was your first scholarly publication?
- Add a code snippet or mathematical equation.
- Add a figure with a caption. This could be a picture of you in Hawaii or a figure from your previous work if the license is permissive enough to allow reuse.

Self-citations are explicitly encouraged, since one goal of this activity is to introduce attendees to the concept of citation by persistent identifier. By having attendees cite their existing works, we hope to show researchers how references can be created from just persistent identifiers, and how this can assist with collaborative and transparent authoring.

The markdown manuscript source has a section for each PSB 2019 attendee, generated from the online attendee list. Names are ordered alphabetically by last name. If you'd like to contribute, but are not already in the list, please insert your section at the appropriate alphabetical location.

For questions on how to use Manubot, see the usage documentation. More information on the tool and its inception is available in the project manuscript [1].

Attendees

Weixuan Fu

Aloha, I'm in the Institute for Biomedical Informatics (IBI) at the University of Pennsylvania and the developer of TPOT and PennAI [2].

My main interest of research is developing automated machine learning tools for the analysis of large scale biomedical/sequencing data. Besides that, I am working on optimizing analysis pipeline of predicting neoantigen specifically presented in tumor cells using DNA and RNA sequencing data, for designing personalized neoantigen vaccines in cancer immunotherapies.

Casey Greene

I run an integrative genomics research lab at the University of of Pennsylvania, and I direct the Childhood Cancer Data Lab for Alex's Lemonade Stand Foundation. The lab at Penn develops methods to integrate large-scale public datasets, primarily from transcriptomic assays, and applies these methods to a broad set of biological questions. We've studied numerous systems, and we currently have active research projects in the application areas of microbial systems [3,4], cancers [5,6,7,8], and rare diseases [9]. At this PSB, a postdoc from the group will present a paper describing Continental Breakfast Included (CBI) effect in the final talk of the final session of this year's meeting [10].

I'm also interested in technologies that enhance the future of scientific communication. Our lab has studied Sci-Hub [11]. We've led a large collaborative review of deep learning in biology and medicine [12]. Members of the lab have developed tools like manubot [1], which you are using now. More publications are available on our lab website.

Daniel Himmelstein

Greetings, I'm in the Greene Lab at the University of Pennsylvania and am the lead developer of the Manubot project. 2019 is my first PSB and I'm exciting to backpack around the Big Island following the conference.

My main area of research is integrating biomedical knowledge using hetnets [13,14]. However, I've also studied Sci-Hub, finding that it provides access to nearly all paywalled scholarly literature [15]. Perhaps my biggest discovery was observing an epidemiological association that higher elevation counties have lower rates of lung cancer, suggesting that oxygen is an inhaled carcinogen (Figure 1) [16,17].

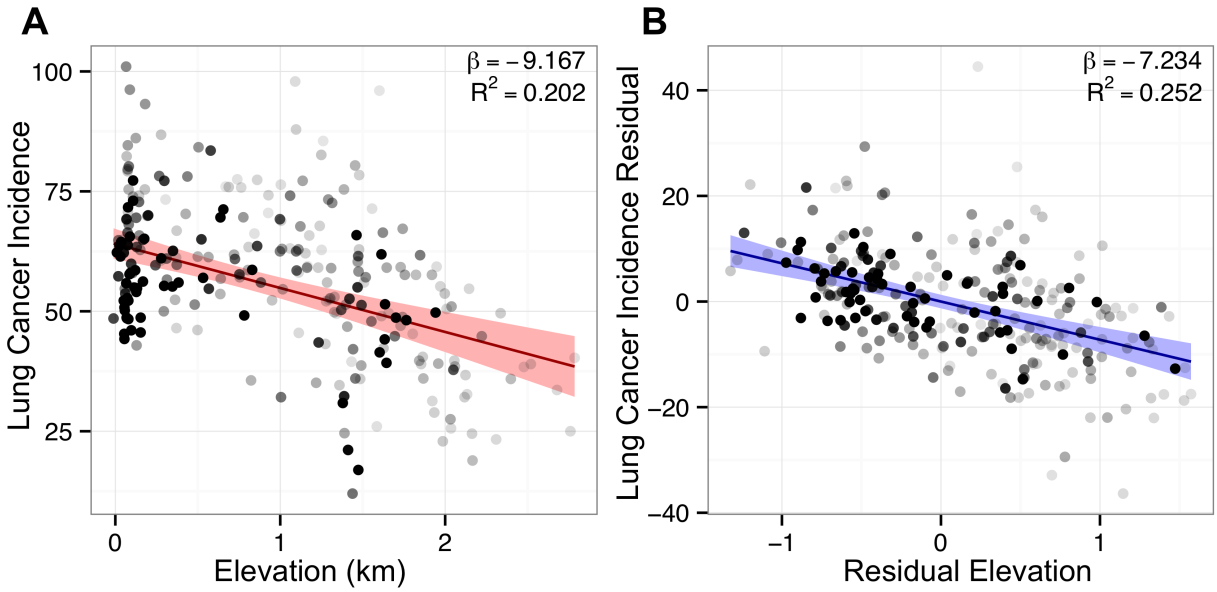


Figure 1: The association between elevation and lung cancer across Western U.S. counties. This figure is reused from here under its CC BY 4.0 License.

I haven't done much text mining, but I did enjoy extracting attendee names for PSB from the online PDF. Converting the PDF to text in Python was as easy as:

```
# https://stackoverflow.com/a/48673754
import tika.parser
parsed = tika.parser.from_file('attendees.pdf')
text = parsed["content"]
```

Trang Le

Hello from the Moore lab at the University of Pennsylvania! I'm a mathematician who's currently excited about automated machine learning.

Here goes the self-citations:

- My own favorite study: Generalization of the Fermi Pseudopotential [18] - a piece of mathematical physics work I got to do when procrastinating writing my dissertation.
- My first (first-author) scholarly publication: Differential privacy-based evaporative cooling feature selection and classification with relief-F and random forests [19]. Check out the Github repo for this study here.
- Code snippet I'm most proud of:

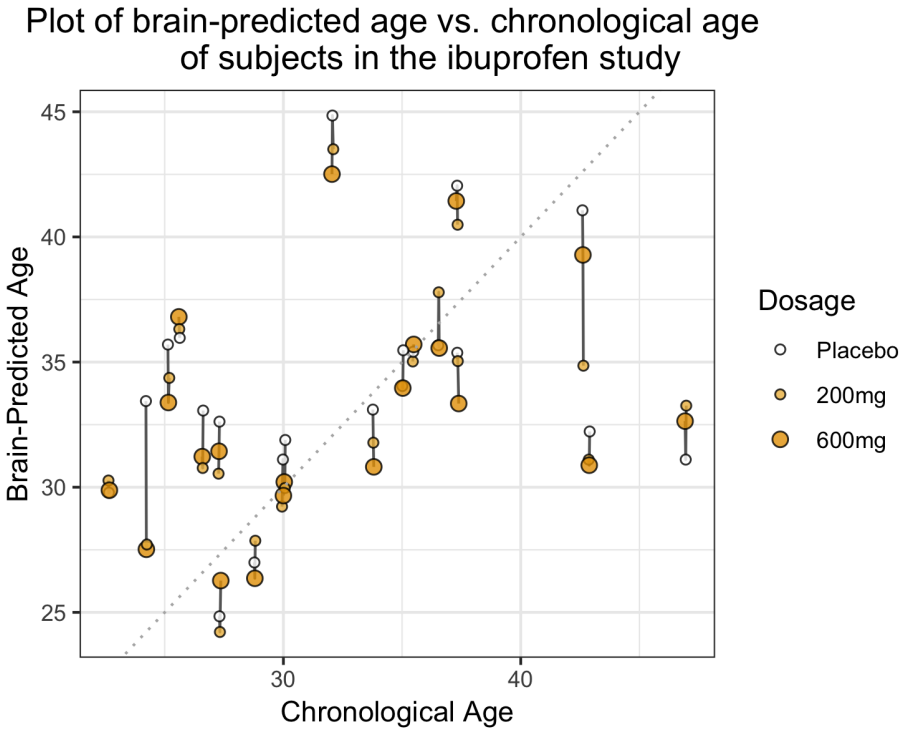
  ```
  M = dec2bin(0:2^(n*n)-1,n*n)
  ```

  I will be impressed if you could tell what the language is. This is my answer to a question on Math StackExchange.

- I have too many favorite mathematical equations, but here's one:

  $a^p \equiv a \mod p$

  Anyone recognize this theorem?
- And a figure with a caption:



This is an improved version of my main figure in this interesting study [20].

Aloah from the Pinello Lab!

I am a computational biologist studying the role of chromatin structure/dynamics and non-coding regions including enhancers, promoters, insulators and their role in gene regulation. The mission of my lab is the integration of omics data to explore and better understand the functional mechanisms of the non-coding genome and to provide accessible tools for the community to accelerate discovery in this field. The long-term goal of my research is to develop innovative computational approaches and to use cutting-edge experimental assays, such as single cell and genome editing, to systematically analyze sources of genetic and epigenetic variation that affect gene regulation in different human traits and diseases. I believe this will further our understanding of disease etiology involving these poorly characterized regions and will provide a foundation for the development of new drugs and more targeted treatments.

I am excited to share during the workshop Reading between the genes: Interpreting noncoding DNA in high throughput a new computational methods we have recently developed to analyze CRISPR tiling screen called CRISPR-SURF. You can read more on the manuscript that was recently published in Nature Methods [21].

Jaclyn Taroni

I'm a data scientist at the Childhood Cancer Data Lab (CCDL), an initiative of Alex's Lemonade Stand Foundation. I'm interested in how diverse collections of publicly available transcriptomic data can help us learn about the biology of rare diseases. As a graduate student, I studied systemic sclerosis [22]. In the PSB 2019 Text Mining and Machine Learning for Precision Medicine Workshop, I'll present our MultiPLIER project [9]. With the CCDL, I've been working on refine.bio, a project for uniformly processing transcriptomic data from multiple species.

References

1. **Open collaborative writing with Manubot**
Daniel S. Himmelstein, David R. Slochower, Venkat S. Malladi, Casey S. Greene, Anthony Gitter
Manubot Preprint (2018-12-31) https://greenelab.github.io/meta-review/

2. **A System for Accessible Artificial Intelligence**
Randal S. Olson, Moshe Sipper, William La Cava, Sharon Tartarone, Steven Vitale, Weixuan Fu, Patryk Orzechowski, Ryan J. Urbanowicz, John H. Holmes, Jason H. Moore
Genetic Programming Theory and Practice XV (2018) https://doi.org/gfsptm
DOI: 10.1007/978-3-319-90512-9_8

3. **ADAGE-Based Integration of Publicly Available Pseudomonas aeruginosa Gene Expression Data with Denoising Autoencoders Illuminates Microbe-Host Interactions**
Jie Tan, John H. Hammond, Deborah A. Hogan, Casey S. Greene
mSystems (2016-01-19) https://doi.org/gcgmbq
DOI: 10.1128/msystems.00025-15 · PMID: 27822512 · PMCID: PMC5069748

4. **Unsupervised Extraction of Stable Expression Signatures from Public Compendia with an Ensemble of Neural Networks**
Jie Tan, Georgia Doing, Kimberley A. Lewis, Courtney E. Price, Kathleen M. Chen, Kyle C. Cady, Barret Perchuk, Michael T. Laub, Deborah A. Hogan, Casey S. Greene
Cell Systems (2017-07) https://doi.org/gcw9f4
DOI: 10.1016/j.cels.2017.06.003 · PMID: 28711280 · PMCID: PMC5532071

5. **Extracting a biologically relevant latent space from cancer transcriptomes with variational autoencoders**
Gregory P. Way, Casey S. Greene
Biocomputing 2018 (2017-11-17) https://doi.org/gfspsd
DOI: 10.1142/9789813235533_0008

6. **Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas**
Gregory P. Way, Francisco Sanchez-Vega, Konnor La, Joshua Armenia, Walid K. Chatila, Augustin Luna, Chris Sander, Andrew D. Cherniack, Marco Mina, Giovanni Ciriello, … Armaz Mariamidze
Cell Reports (2018-04) https://doi.org/gfspsb
DOI: 10.1016/j.celrep.2018.03.046 · PMID: 29617658 · PMCID: PMC5918694

7. **Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas**
Theo A. Knijnenburg, Linghua Wang, Michael T. Zimmermann, Nyasha Chambwe, Galen F. Gao,

Andrew D. Cherniack, Huihui Fan, Hui Shen, Gregory P. Way, Casey S. Greene, … Armaz Mariamidze
Cell Reports (2018-04) https://doi.org/gfspsc
DOI: 10.1016/j.celrep.2018.03.076 · PMID: 29617664 · PMCID: PMC5961503

8. **Oncogenic Signaling Pathways in The Cancer Genome Atlas**
Francisco Sanchez-Vega, Marco Mina, Joshua Armenia, Walid K. Chatila, Augustin Luna, Konnor C. La, Sofia Dimitriadoy, David L. Liu, Havish S. Kantheti, Sadegh Saghafinia, … Armaz Mariamidze
Cell (2018-04) https://doi.org/gc7r9b
DOI: 10.1016/j.cell.2018.03.035 · PMID: 29625050 · PMCID: PMC6070353

9. **MultiPLIER: a transfer learning framework reveals systemic features of rare autoimmune disease**
Jaclyn N Taroni, Peter C Grayson, Qiwen Hu, Sean Eddy, Matthias Kretzler, Peter A Merkel, Casey S Greene
Cold Spring Harbor Laboratory (2018-08-20) https://doi.org/gfc9bb
DOI: 10.1101/395947

10. **Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single cell RNA transcriptomics**
Qiwen Hu, Casey S Greene
Cold Spring Harbor Laboratory (2018-08-05) https://doi.org/gdxxjf
DOI: 10.1101/385534

11. **Sci-Hub provides access to nearly all scholarly literature**
Daniel S Himmelstein, Ariel Rodriguez Romero, Jacob G Levernier, Thomas Anthony Munro, Stephen Reid McLaughlin, Bastian Greshake Tzovaras, Casey S Greene
eLife (2018-03-01) https://doi.org/ckcj
DOI: 10.7554/elife.32822 · PMID: 29424689 · PMCID: PMC5832410

12. **Opportunities and obstacles for deep learning in biology and medicine**
Travers Ching, Daniel S. Himmelstein, Brett K. Beaulieu-Jones, Alexandr A. Kalinin, Brian T. Do, Gregory P. Way, Enrico Ferrero, Paul-Michael Agapow, Michael Zietz, Michael M. Hoffman, … Casey S. Greene
Journal of The Royal Society Interface (2018-04) https://doi.org/gddkhn
DOI: 10.1098/rsif.2017.0387 · PMID: 29618526 · PMCID: PMC5938574

13. **Heterogeneous Network Edge Prediction: A Data Integration Approach to Prioritize Disease-Associated Genes.**
Daniel S Himmelstein, Sergio E Baranzini
PLoS computational biology (2015-07-09) https://www.ncbi.nlm.nih.gov/pubmed/26158728
DOI: 10.1371/journal.pcbi.1004259 · PMID: 26158728 · PMCID: PMC4497619

14. **Systematic integration of biomedical knowledge prioritizes drugs for repurposing**
Daniel Scott Himmelstein, Antoine Lizee, Christine Hessler, Leo Brueggeman, Sabrina L Chen, Dexter Hadley, Ari Green, Pouya Khankhanian, Sergio E Baranzini
eLife (2017-09-22) https://doi.org/cdfk
DOI: 10.7554/elife.26726 · PMID: 28936969 · PMCID: PMC5640425

15. **Sci-Hub provides access to nearly all scholarly literature**
Daniel S Himmelstein, Ariel Rodriguez Romero, Jacob G Levernier, Thomas Anthony Munro, Stephen Reid McLaughlin, Bastian Greshake Tzovaras, Casey S Greene
eLife (2018-03-01) https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5832410/
DOI: 10.7554/elife.32822 · PMID: 29424689 · PMCID: PMC5832410

16. **Lung cancer incidence decreases with elevation: evidence for oxygen as an inhaled carcinogen**
Kamen P. Simeonov, Daniel S. Himmelstein
PeerJ (2015-01-13) https://doi.org/98p
DOI: 10.7717/peerj.705 · PMID: 25648772 · PMCID: PMC4304851

17. **Unraveling the Ties of Altitude, Oxygen and Lung Cancer**
George Johnson
The New York Times (2016-01-25) https://www.nytimes.com/2016/01/26/science/unraveling-the-ties-of-altitude-oxygen-and-lung-cancer.html

18. **Generalization of the Fermi Pseudopotential**
Trang T. Le, Zach Osman, D. K. Watson, Martin Dunn, B. A. McKinney
arXiv (2018-06-14) https://arxiv.org/abs/1806.05726v1

19. **Differential privacy-based evaporative cooling feature selection and classification with relief-F and random forests**
Trang T Le, W Kyle Simmons, Masaya Misaki, Jerzy Bodurka, Bill C White, Jonathan Savitz, Brett A McKinney
Bioinformatics (2017-05-04) https://doi.org/f96b8d
DOI: 10.1093/bioinformatics/btx298 · PMID: 28472232 · PMCID: PMC5870708

20. **Effect of Ibuprofen on BrainAGE: A Randomized, Placebo-Controlled, Dose-Response Exploratory Study**
Trang T. Le, Rayus Kuplicki, Hung-Wen Yeh, Robin L. Aupperle, Sahib S. Khalsa, W. Kyle Simmons, Martin P. Paulus
Biological Psychiatry: Cognitive Neuroscience and Neuroimaging (2018-10) https://doi.org/gfsprv
DOI: 10.1016/j.bpsc.2018.05.002 · PMID: 29941380

21. **CRISPR-SURF: discovering regulatory elements by deconvolution of CRISPR tiling screen data.**
Jonathan Y Hsu, Charles P Fulco, Mitchel A Cole, Matthew C Canver, Danilo Pellin, Falak Sher, Rick Farouni, Kendell Clement, Jimmy A Guo, Luca Biasco, … Luca Pinello
Nature methods (2018-12) https://www.ncbi.nlm.nih.gov/pubmed/30504875
DOI: 10.1038/s41592-018-0225-6 · PMID: 30504875

22. **A novel multi-network approach reveals tissue-specific cellular modulators of fibrosis in systemic sclerosis**
Jaclyn N. Taroni, Casey S. Greene, Viktor Martyanov, Tammara A. Wood, Romy B. Christmann, Harrison W. Farber, Robert A. Lafyatis, Christopher P. Denton, Monique E. Hinchcliff, Patricia A. Pioli, … Michael L. Whitfield
Genome Medicine (2017-03-23) https://doi.org/gfsptx
DOI: 10.1186/s13073-017-0417-1 · PMID: 28330499 · PMCID: PMC5363043