BIOGRAPHICAL SKETCH

Provide the following information for the Senior/key personnel and other significant contributors.
Follow this format for each person. **DO NOT EXCEED FIVE PAGES.**

NAME: Casey S. Greene

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

POSITION TITLE: Assistant Professor of Systems Pharmacology and Translational Therapeutics

EDUCATION/TRAINING

INSTITUTION AND LOCATION	DEGREE *(if applicable)*	Completion Date MM/YYYY	FIELD OF STUDY
Berry College	B.S.	05/05	Chemistry
Dartmouth College	Ph.D.	11/09	Computational Genetics
Princeton University	Postdoctoral	07/12	Bioinformatics

A. Personal Statement

My lab's overall research objective is to develop methods for unsupervised integration of large-scale public data compendia because we expect that patterns that are evident across such compendia are likely to be robust. We then seek to use these patterns to identify processes, cell types, or other factors that change at various stages of disease onset and progression, to characterize the nature of responses to therapeutics in clinical trials, to identify robust subgroupings of diseases, and for basic biology research.

The goals of the proposed research are to develop a suite of tools that use low-dimensional representations to produce a versioned catalog of processes and cell types and provide to search capabilities over the HCA. We recently showed that methods that create low-dimensional representations via neural networks are quite sensitive to hyperparameters which can hamper evaluation [1]. Our past work includes unsupervised methods that integrate heterogeneous data compendia into these low dimensional representations [2].

Our lab aims to develop robust and usable software. We perform code review for all code committed to a lab-wide repository, which ensures that software that we develop is both well documented and usable when it is released. We are developing approaches to automatically perform and update analyses whenever the underlying source code or data changes [3]. We disseminate our work under an open license to maximize opportunities for reuse, and we look forward to providing new capabilities to the Seed Network.

I recently led an effort to review the deep learning literature in biology and medicine. Our review, written in the open on GitHub, cites more than 500 papers. A PDF of the manuscript was downloaded more than 20,000 times in 2017 making it the most downloaded preprint posted in all of that year leading to it being listed in the article, "2017 in news: The science events that shaped the year" from the journal *Nature* [4]. We will bring this toolkit, which we used to author this CZI Seed Networks proposal in the open, to the network as well.

Four peer reviewed publications that highlight experience and qualifications for this project:

1. Hu Q, Greene CS. Parameter tuning is a key part of dimensionality reduction via deep variational autoencoders for single-cell transcriptomics. *Pac Symp Biocomput.* In Press (preprint doi: 10.1101/385534)
2. Tan J, Doing G, Lewis KA, Price CE, Chen KM, Cady KC, Perchuk B, Laub MT, Hogan DA, **Greene CS**. Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks. *Cell Systems*. 2017. 5:63-71. PMCID: 5532071
3. Beaulieu-Jones, B.K., **Greene, C.S.** Reproducibility of computational workflows is automated using continuous analysis. *Nat Biotech.* 2017. 35:342-346. PMID: 28288103
4. Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, Ferro E, Agapow P, Zietz M, Hoffman MM, Xie W, Rosen GL, Legenrich BJ, Israeli J, Lanchantin J, Woloszynek S, Carpenter AE, Shrikumar A, Xu J, Cofer EM, Lavender CA, Turaga SC, Alexandari AM, Lu Z, Harris DJ, DeCaprio D, Qi Y, Kundaje A, Peng Y, Wiley LK, Segler MHS, Boca SM, Swamidass SJ, Gitter A+, **Greene CS+.**

Opportunities and Obstacles for Deep Learning in Biology and Medicine. *Journal of the Royal Society Interface*. 2018. [Epub ahead of print – doi: 10.1098/rsif.2017.0387]

B. Positions and Honors

Employment

2009-2012 Postdoctoral Research Associate, Lewis-Sigler Institute for Integrative Genomics, Princeton University, Princeton, NJ.
2010 Lecturer, Department of Computer Science, Princeton University, Princeton, NJ.
2012-2015 Assistant Professor, Department of Genetics, Geisel School of Medicine, Hanover, NH.
2015- Assistant Professor, Department of Systems Pharmacology and Translational Therapeutics, University of Pennsylvania, Philadelphia, PA

Honors

2014 Moore Investigator in Data-Driven Discovery, Gordon and Betty Moore Foundation.

Other Experience and Professional Memberships

2011- Member International Society for Computational Biology
2013- Program Committee for Intelligent Systems for Molecular Biology (ISMB)
2013-2014 Co-chair of the "Text and Data Mining for Biomedical Discovery" session at PSB.
2015 Co-chair of the "Computational Approaches to Study Microbes and Microbiomes" workshop at PSB.
2016- Co-chair. Posters. Intelligent Systems for Molecular Biology (ISMB) Meeting

C. Contributions to Science

*** indicates co-first author.**

Neural Network Methods for Unsupervised Analysis of Gene Expression

Gene expression data provide a broad lens for generating hypotheses about biological systems. Because these data are generated in a genome-wide manner, algorithms that mine these data are less biased toward well studied features of biology. Analyses that apply unsupervised machine learning algorithms to large compendia that include experiments covering diverse processes can be particularly well suited to discovering general properties of the measured biological systems. This work was performed in my own lab. During this work I developed and evaluated new algorithms for mining gene expression data, I drafted and reviewed manuscripts, and I supervised a PhD student, Jie Tan, and undergraduate student, Kathleen Chen, in my own lab. Programming support was carried out in part by Matt Huyck, Dongbo Hu, and Rene Zelaya who are programmers in my research group. Molecular experiments to validate computational predictions were performed in Dr. Deb Hogan's lab.

1. Tan, J., Ung, M., Cheng, C., and **Greene, C.S.** Unsupervised feature construction and knowledge extraction from genome-wide assays of breast cancer with denoising autoencoders. *Pac Symp Biocomput*. 2015; 20:132-43. PMCID: PMC4299935
2. Tan, J., Hammond, J.H, Hogan, D.A., **Greene, C.S.** ADAGE-based integration of publicly available pseudomonas aeruginosa gene expression data with denoising autoencoders illuminates microbe-host interactions. *mSystems*. 1(1):e00025-15. PMCID: PMC5069748
3. Tan J, Doing G, Lewis KA, Price CE, Chen KM, Cady KC, Perchuk B, Laub MT, Hogan DA, **Greene CS**. Unsupervised extraction of stable expression signatures from public compendia with an ensemble of neural networks. *Cell Systems.* 2017. 5:63-71. PMCID: 5532071
4. Tan, J., Huyck, M., Hu, D., Zelaya, RA., Hogan, DA., **Greene, CS.** ADAGE signature analysis: differential expression analysis with data-defined gene sets. *BMC Bioinformatics*. 2017. 18(1):512. PMCID: 5700673

Context-specific Functional Relationship Networks Extracted from Large-Scale Data

Proteins act in concert to carry out the biological processes required to develop and sustain living organisms. Because of the limits of annotation specificity and experimental coverage, pathway databases and protein-protein networks frequently treat an interaction at any measured time as a universal interaction. We showed that these context-specific networks were useful for generating hypotheses related to the development of

asymmetry in zebrafish [1,2], phase-specific interactions in the human cell cycle [3], and the tissue-specific response to pro-inflammatory cytokines [4]. In our work, we have experimentally validated such discoveries. For these contributions I developed new algorithms, created the software infrastructure to enable the integration of large-scale databases, evaluated the resulting predictions, and wrote the manuscripts.

1. Wong, A.K.*, Park, C.Y.*, **Greene, C.S.***, Bongo, L.A., Guan, Y., and Troyanskaya, O.G. IMP: A multi-species functional genomics portal for integration, visualization and prediction of protein functions and networks. *Nucleic Acids Res*. 2012 Jul;40(Web Server issue):W484-490. PMCID: PMC3394282
2. Park, C.Y.*, Wong, A.K.*, **Greene, C.S.***, Rowland, J., Guan, Y., Bongo, L.A., Burdine, R.D., and Troyanskaya, O.G. Functional knowledge transfer for high-accuracy prediction of under-studied biological processes. *PLoS Comput Biol*. 2013;9(3):e1002957. PMCID: PMC3597527
3. Tan, J., Grant, G.D., Whitfield, M.L., and **Greene, C.S**. 2013. Time-Point Specific Weighting Improves Coexpression Networks from Time-Course Experiments. *Evolutionary Computation, Machine Learning, and Data Mining in Bioinformatics*. 7833:11-22.
4. **Greene, C.S.***, Krishnan, A.*, Wong, A.K.*, Riccioti, E., Zelaya, R.A., Himmelstein, D.S., Zhang, R., Hartmann, B.M., Zaslavsky, E., Sealfon, S.C., Chasman, D.I., FitzGerald, G.A., Dolinski, K., Grosser, T., Troyanskaya, O.G. Understanding multicellular function and disease with human tissue-specific networks. *Nat Genetics*. 2015. PMCID: PMC4828725

New Computing Approaches to Identify gene-gene Interactions in Genome-wide Association Studies

Gene-gene interactions represent a potential source of the missing heritability associated with common human diseases. As a graduate student, I developed new computational approaches to identify gene-gene interactions in genetic association data. These algorithms were focused on the particularly challenging problem of identifying gene-gene associations in which no SNP had a main effect. I developed methods based on evolutionary computation [1], heuristic approaches [3], and statistical approaches [4] to test for interactions in the context of main effects. I also developed a simulation approach to evaluate the implications of gene-gene interactions on standard genome-wide association study designs [2]. For these papers I developed and implemented new algorithms, evaluated these algorithms in the context of both simulation studies and real data analysis, and wrote the manuscripts.

1. **Greene, C.S.**, White, B.C., and Moore, J.H. Ant colony optimization for genome-wide genetic analysis. *Ant Colony Optimization and Swarm Intelligence*. 2008. 5217:37-47.
2. **Greene, C.S.**, Penrod, N.M., Williams, S.M., and Moore, J.H. Failure to replicate a genetic association may provide important clues about genetic architecture. *PLoS ONE*. 2009 June2;4(6):e5639. PMCID: PMC2685469
3. **Greene, C.S.**, Penrod, N.M., Kiralis, J., and Moore, J.H. Spatially Uniform ReliefF (SURF) for computationally-efficient filtering of gene-gene interactions. *BioData Min*. 2009 Sep 22;2(1):5. PMCID: PMC2761303
4. **Greene, C.S.**, Himmelstein, D.S., Nelson, H.H., Kelsey, K.T., Williams, S.M., Andrew, A.S., Karagas, M.R., and Moore, J.H. Enabling personal genomics with an explicit test of epistasis. *Pac Symp Biocomput*. 2010:327-36. PMCID: PMC2916690

Discovery of Gene-Phenotype and Drug-Phenotype Associations from Genetic Association Data

Algorithms that we developed and discussed above have been used by others to identify gene-gene interactions associated with human phenotypes. I have also participated in studies to identify gene-gene interactions associated with disease using our methods and those developed by others. In part, this required the development of new computing approaches. In Beretta et al. [1], I performed an analysis of gene-gene interactions from a candidate SNP dataset. In Greene et al. [2], we developed a new computing framework, graphics cards, to perform a genome-wide analysis of epistasis because the problem was not tractable on traditional computing platforms. In Mahoney et al. [3], I guided the analysis of students and postdocs using functional networks to analyze and interpret genetic association results. In Cordell et al. [4], I used an integrative bioinformatics approach to prioritize small molecules based on GWAS-associated pathways. For these projects, I participated in the development and evaluation of methods, the analysis of genetic association data, the guidance of undergraduate, graduate students, and postdocs, and the drafting of manuscripts.

1. Beretta, L., Cappiello, F., Moore, J.H., Barili, M., **Greene, C.S.**, Scorza, R. Ability of epistatic interac-

tions of cytokine single-nucleotide polymorphisms to predict susceptibility to disease subsets in systemic sclerosis patients. *Arthritis Rheum*. 2008 Jul 15;59(7):974-83. PMID:18576303
2. **Greene, C.S.***, Sinnott-Armstrong N.A.*, Himmelstein, D.S., Park, P.J., Moore, J.H., and Harris, B.T. Multifactor dimensionality reduction for graphics processing units enables genome-wide testing of epistasis in sporadic ALS. *Bioinformatics*. 2010 Mar 1;26(5):694-5. PMCID: PMC2828117
3. Mahoney, J.M., Taroni, J., Martyanov, V., Wood, T.A., **Greene, C.S.**, Pioli, P.A., Hinchcliff, M.E., Whitfield, M.L. Systems level analysis of systemic sclerosis shows a network of immune and profibrotic pathways connected with genetic polymorphisms. *PLoS Comput Biol*. 2015. Jan 8;11(1):e1004005. PMCID: PMC4288710
4. Cordell, H.J., Han, Y., Li, Mells, G.F., Hirshfield, G.M., **Greene, C.S.**, Xie, G., Juran, B.D., Zhu, D., Qian, D.C., Floyd, J.A.B., Morley, K.I., Prati, D., Lleo, A., Cusi, D., Canadian-US PBC Consortium, Italian PBC Genetics Study Group, UK-PBC Consortium, Gershwin, M.E., Anderson, C.A., Lazaridis, K.N., Invernizzi, P., Seldin, M.F., Sandford, R.N., Amos, C.I., and Siminovitch, K.A. An international genome-wide meta-analysis identifies new primary biliary cirrhosis risk loci and highlights pathogenic pathways for drug targeting. *Nat Commun.* PMCID: PMC4580981

Machine-learning approaches to construct single- and pan-cancer pathway activity signatures

We developed approaches to detect the activity levels of pathways from gene expression data by starting from the combination of sequencing and expression data. We have applied these approaches to NF1 in glioblastoma [1] as well as TP53's role in the DNA damage repair pathway across cancers [2] and Ras signaling across cancer types [3] through the PanCancerAtlas Pathways effort [4]. I supervised a student within my lab for each of these projects, and I led the PanCancerAtlas Ras Pathway characterization efforts [3].

1. Way GP, Allaway RJ, Bouley SJ, Fadul CE, Sanchez Y, **Greene CS.** A machine learning classifier trained on cancer transcriptomes detects NF1 inactivation signal in glioblastoma. *BMC Genomics*. 2017 Feb 6;18(1):127. PMCID: PMC5292791
2. Knijnenburg TA, Wang L, Zimmermann MT, Chambwe N, Gao GF, Cherniack AD, Fan H, Shen H, Way GP, **Greene CS**, Liu Y, Akbani R, Feng B, Donehower LA, Miller C, Shen Y, Karimi M, Chen H, Kim P, Jia P, Shinbrot E, Zhang S, Liu J, Hu H, Bailey MH, Yau C, Wolf D, Zhao Z, Weinstein JN, Li L, Ding L, Mills GB, Laird PW, Wheeler DA, Shmulevich I; Cancer Genome Atlas Research Network, Monnat RJ Jr., Xiao Y, Wang C. Genomic and Molecular Landscape of DNA Damage Repair Deficiency across The Cancer Genome Atlas. *Cell Rep*. 2018 Apr 3;23(1):239-254.e6. PMID: 29617664
3. Way GP, Sanchez-Vega F, La K, Armenia J, Chatila WK, Luna A, Sander C, Cherniack AD, Mina M, Ciriello G, Schultz N; Cancer Genome Atlas Research Network, Sanchez Y, **Greene CS**. Machine Learning Detects Pan-cancer Ras Pathway Activation in The Cancer Genome Atlas. *Cell Rep*. 2018 Apr 3;23(1):172-180.e3. PMID: 29617658
4. Sanchez-Vega F, Mina M, Armenia J, Chatila WK, Luna A, La KC, Dimitriadoy S, Liu DL, Kantheti HS, Saghafinia S, Chakravarty D, Daian F, Gao Q, Bailey MH, Liang W, Foltz SM, Shmulevich I, Ding L, Heins Z, Ochoa A, Gross B, Gao J, Zhang H, Kundra R, Kandoth C, Bahceci I, Deversh L, Dogrusoz U, Zhou W, Shen H, Laird PW, Way GP, **Greene CS,** Liang H, Xiao Y, Wang C, Iavarone A, Berger AH, Bivona TG, Lazar AJ, Hammer GD, Giordano T, Kwong LN, McArthur G, Huang C, Tward AD, Frederick MJ, McCormic F, Meyerson M, Cancer Genome Analysis Research Network, Van Allen EM, Cherniack AD, Ciriello G, Sander C, Schultz N. Oncogenic Signaling Pathways in The Cancer Genome Atlas. *Cell*. 2018. 173(2):321-337.e10.

PubMed Indexed work on My Bibliography:
http://www.ncbi.nlm.nih.gov/myncbi/browse/collection/40332249/?sort=date&direction=descending
Full list of published work (includes computer science publications) on Google Scholar:
https://scholar.google.com/citations?user=ETJoidYAAAAJ

D. Research Support

Ongoing Research Support

Moore Investigator — Greene (PI) — 12/01/2014-11/30/2019
Gordon and Betty Moore Foundation, Moore Investigators in Data-Driven Discovery
Learning the context of publicly available genome-wide data

We are developing algorithms that summarize publicly available assays of gene expression into models that represent important contextual information, e.g. the environment of the assayed cells and their responses to it.
Role: Principal Investigator

NIH R01 CA200854 Doherty, Schildkraut (PIs) 12/01/2015-11/30/2020
NIH/NCI
Characterizing Molecular Subtypes of Ovarian Cancer in African-American Women
We are characterizing the distribution of high-grade serous ovarian cancer subtypes in African-American women.
Role: Co-Investigator

NIH R01 NS095411 Sanchez, Ratner, Hoopes (PIs) 09/30/2015-07/31/2020
NIH/NINDS
Targeting tumors with NF1 loss.
We are developing a classifier that identifies cancers with genetic alterations that render them susceptible to treatment with drugs that are synthetic lethal with the loss of NF1.
Role: Co-Investigator

ALSF CCDL Greene (PI) 07/01/2017-06/30/2019
Alex's Lemonade Stand Foundation
Childhood Cancer Data Lab
The major goal of this project is to establish a Childhood Cancer Data Lab within Alex's Lemonade Stand Foundation to build software infrastructure and enabling technologies for researchers studying childhood cancers.
Role: Principal Investigator

Digital Innovation Grant Greene (PI) 12/01/2017-12/30/2018
Pfizer
"Digital Data Innovation Project"
The goal of this project is to develop software to construct hetnets that can be incorporated into Pfizer's research workflows.
Role: Principal Investigator

Human Cell Atlas Computational Methods Greene (PI) 03/01/2018-02/28/2019
Chan-Zuckerberg Initiative
Genome-wide hypothesis generation for single-cell expression via latent spaces of deep neural networks
The goal of this project is to develop software to construct biologically meaningful low-dimensional representations from single cell data generated via the Human Cell Atlas project.
Role: Principal Investigator