I found my passion for evolutionary genetics, a field that culminates all my past interests and consumes my present thoughts, after initially exploring a few other fields. My past experiences have inspired me to help others find their home in STEM fields, by continually disseminating the knowledge I have gained and continue to gain during my research.

**Beginnings:** I entered the field of evolutionary genetics from an atypical background: my undergraduate work was in economics, political science, and statistics. Following a period of uncertainty and war in the US and abroad, I was captivated by using statistics to model rare and dire political phenomena such as interstate conflict and governmental failure. I grew especially interested in the methodology involved in modeling events with observational data, and I learned as much probability theory, Bayesian statistics, and computational statistics as I could. Having been interested in computer programming since I was a young teenager, I was captivated by the idea that advanced statistics and computing could be combined to understand complex data sets.

The impetus of my transition from modeling rare political events to studying evolutionary genetics was an early love of plants and evolution. I have always been enchanted by biology and plants, especially orchids. As a teenager, I collected and grew a variety of orchids with a particular interest in species in the genuses *Angraecum* and *Aeranthes* of Madagascar. In high school I begged our lab manager to allow me to come after hours to aseptically germinate *Piperia elegans*, a terrestrial orchid native to my hometown of San Francisco. During my final year as an undergraduate a college friend worked a summer at Pacific Biosciences and returned talking about genomics. Immediately, I was intrigued at the wealth of genomic data being produced and saw an opportunity to unite my love of biology with my passion for statistical modeling and programming to understand complex phenomenon.

**Bioinformatics Core:** After graduation I made a decision that changed my life trajectory: I delayed applying to graduate programs in statistics and political methodology and sought a job in the field of genomics that would combine my statistical skills and passion for the life sciences. It was nerve-wracking leaving a field I was comfortable in to work alongside PhD biologists in a subject where I lacked formal training, but I soon began a position at the Bioinformatics Core at the UC Davis Genome Center. I was enthralled by working with genomic data and became the group's primary statistical programmer. I was spending all day working on projects, and nights and weekends reading everything I could about molecular and cellular biology. Each project I worked on was different, and each presented an opportunity for me to learn about a new system, new algorithms, and new statistical techniques.

In addition to solidifying my love of biology and genomics, my three years at the Bioinformatics Core gave me an exceptional training in bioinformatics. The UC Davis Genome Center was churning out masses of data, and our group was in charge of creating, evaluating, and applying analysis techniques, including short read quality control, genome assembly, read mapping, SNP calling, and differential expression analysis. I was the only one in my group to have training in statistics and extensive programming skills, so I was routinely given the most difficult problems for which I came up with innovative solutions. I conducted bioinformatics analyses and coauthored a number of papers ranging from the characterization of disease response in citrus with RNA-seq data using a novel application of unsupervised machine learning approaches (Martinelli et al, 2012), to designing an algorithm to identify and classify chemotherapy-induced chromosomal translocations arising in the human MLL gene (Shih, et al 2012), and to assessing the quality of genome assemblies as part of the Assemblathon project (Earl et al 2012). During this time I also developed open source software

(http://github.com/vsbuffalo) and contributed code to large-scale open source bioinformatics projects like Samtools, seqtk, and bioawk. My software for quality control of next-generation sequencing data, including Scythe (http://github.com/vsbuffalo/scythe) and QRQC (http://bit.ly/bioc-qrqc), has been widely adopted in genomics data processing pipelines.

**Recent Research Work:** After working closely with biologists at the Bioinformatics Core, I knew that I wanted to pursue a PhD in biology. However, through working closely with graduate students to help analyze their genomic data I had seen firsthand the level of commitment a PhD involves. I wanted firsthand experience in a variety of labs working on a variety of topics before choosing a specific field to pursue. Since I have had a lifelong love of plants, I accepted a bioinformatics position in professor Jorge Dubcovsky's wheat genetics and breeding lab. There, I worked on inventive computational methods to approach the tricky problem of transcriptome assembly in polyploid wheat (*Triticum turgidum*), leading to the development of an algorithm that allowed us to separate the homeologs in wheat transcriptome assemblies. This work culminated in a paper (Krasileva, Buffalo, et al, 2013) and the open source program readphaser (http://github.com/vsbuffalo/readphaser), a program which is still used for a variety of polyploid transcriptome assembly tasks in other species. Through my position working with Jorge, I cultivated a strong passion for genetics. At one point, a postdoc colleague recommend that given how much I enjoy statistics, I should look into population genetics. After religiously attending professor Jeffrey Ross-Ibarra's plant population genetics journal club, I became enthralled with the subject. After years of searching for my discipline in biology, I had found that my passion was population and evolutionary genetics — and soon joined professor Ross-Ibarra's lab working on maize and teosinte population and statistical genetics.

In professor Ross-Ibarra's lab, I further solidified my love of population genetics. I was particularly amazed how genomic data combined with population genetics theory and methods could be used to reconstruct the histories of important species such as crop plants. Domesticated crops provided a fascinating system to think about the interplay of selection and demography (e.g. bottlenecks during domestication). During this position, I mentored undergraduates working in the lab and taught them important computational skills and statistics. I also held a weekly Python programming course for a diverse group of plant biology graduate students. In this position, I also developed novel statistical genetics methods and software that phases and imputes the genotypes of reduced-sequencing data (e.g. genotype by sequencing) of large progeny arrays. Our data comprise over 1,000 plants genotyped with an 80% heterozygous error rate and 40% missing data, and my method reconstructs their genotypes and phases using pedigree structures. My methods are open source (https://github.com/vsbuffalo/ProgenyArray) and I believe will be widely applicable in the ecological and evolutionary genetics communities.

**Teaching Computational Skills to STEM Researchers.** My career at the Bioinformatics Core had a profound impact during my development as a scientist, driving me to adopt a cautious attitude and seek the most robust methods to apply to genomic data. I saw numerous firsthand examples of how poor bioinformatics methods could bias biological conclusions. I realized that I was only able to detect these problems because I had programming and statistical skills at my disposal, whereas my biologist friends with less experience may not spot these subtle complications with analysis methods. I realized I was in a unique outreach position: I could teach others the skills I had learned by analyzing dozens of species' data at the Bioinformatics Core.

I became an active teacher both in my local scientific community and a wider internet community. Locally, I helped teach two week-long bioinformatics courses that emphasized

methods for non-model organisms, and designed and taught a data analysis and visualization workshop (which is now open source, here: http://bit.ly/ucddav2012). Recently I helped teach a Software Carpentry course (an organization that teaches software skills to scientists), assist a diverse group of students in my local R users' group, and I am consulted on bioinformatics projects. On the Internet, I am active in the bioinformatics community on Twitter, and I maintain a notebook of bioinformatics and programming-related essays (http://vincebuffalo.com).

**My Book - *Bioinformatics Data Skills*.** While these are important teaching contributions, I wanted to reach a wider audience to teach bioinformatics as I approach it — with a heavy emphasis on robust methods and reproducibility. I am now finishing a book published by O'Reilly Media entitled *Bioinformatics Data Skills: Robust Research with Open Source Tools* which is currently available for purchase as a digital early release (http://oreil.ly/10IS4IB), and will be in print in early 2015. My book does not try to teach readers workflows in genome assembly or mapping. Rather, because bioinformatics tools and data change perpetually, bioinformaticians instead rely on data skills more than specific workflows. The goal of my book is to teach bioinformatics through learning core data skills, such that biologists have the computational agility to explore data and open black-box methods to robustly extract biological information. I hope my book will have a wide impact on the genomics community and improve STEM education. Already, the early release version of my book is currently being used as the primary course material for a discussion group at the University of Minnesota of more than 20 biologists (http://bit.ly/0-compute), and other instructors have notified me that my book will be used in their bioinformatics cirricula.

**Future Goals:** I joined professor Graham Coop's lab this fall, to pursue my interests in using advanced statistical methods to better understand evolutionary biology through genetic data. Dr. Coop has extensive experience developing methods and modeling the evolutionary phenomenon I'm most interested in, including local adaptation and polygenic adaptation. My extensive past research experience has accelerated my drive to do more research; thus, my future goal is to become a professor of evolutionary biology and population genetics.

**Future Outreach: Engaging Young Women in STEM.** Although my past and current outreach efforts have predominantly been geared towards teaching and creating education material in computational skills to STEM researchers, I hope to tune more material towards teaching younger audience, especially young women. I have two young sisters (ages 12 and 14) and have seen firsthand how biases against young women in math and sciences classes affect them — even at a very young age. Both of my sisters have been told that they are not naturally good at math and science and should pursue other interests, which had drastically discouraged them from studying these fields. I also did poorly in these subjects as a child, but that lack of success was never characterized as a lack of aptitude. I have re-engaged my sisters' interests in the sciences through interesting and relevant examples, including genetics topics like calculating how much DNA we all share in common. However, such biases happen on a larger scale and I strongly believe both male and female scientists need to actively work to counteract the numerous biases that affect young women. Along these lines, I have started to develop motivating material and exercises to create a fun interactive blog. I've learned an extraordinary amount about effectively communicating examples while writing my book, and with funding from the NSF I will enhance a better educational infrastructure for a more diverse research community in STEM sciences. In conclusion, through outreach I hope to help others develop their skills and find the field they're most passionate about — just as I found my passion in evolutionary genetics.