OMB No. 0925-0001 and 0925-0002 (Rev. 09/17 Approved Through 03/31/2020)

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: **Gregg Thomas**

eRA COMMONS USER NAME (credential, e.g., agency login): **GRTHOMAS6**

POSITION TITLE: **Bioinformatics Scientist**

EDUCATION/TRAINING (Begin with baccalaureate or other initial professional education, such as nursing, include postdoctoral training and residency training if applicable. Add/delete rows as necessary.)

| INSTITUTION AND LOCATION | DEGREE  (if applicable) | Start Date  MM/YYYY | Completion Date  MM/YYYY | FIELD OF STUDY |
| --- | --- | --- | --- | --- |
| Purdue University | B.S. | 08/2004 | 05/2010 | Biology |
| Indiana University | M.S. | 08/2011 | 05/2013 | Bioinformatics |
| Indiana University | Ph.D. | 08/2013 | 07/2019 | Evolutionary biology and Bioinformatics |

**A. Personal Statement**

**My long-term goal as a researcher is to use and develop computational methods to answer fundamental questions about species relationships and molecular and genome evolution, establishing an independent lab to perform this research and mentor others if the opportunity arises. My main line of inquiry centers around identifying the structure of the genome and the causes and drivers of molecular evolution. I aim to discover general patterns of molecular change between species and identify the underlying forces that cause these patterns. Ultimately, my goal is to develop an integrative genomics research program using sequence data of all types and practical computing and methods development to answer essential questions relating to phylogenetics and molecular evolution.**

**Under the advisement of Dr. Matthew Hahn at Indiana University, I was exposed to a broad variety of topics that allowed me to gain skills related to many aspects of computational sequence analysis. During my dissertation work on mutation rate variation in primates, I developed a theoretical model of mutation rate estimation, tested the predictive power of that model by using pedigree sequencing, and showed that different types of mutations follow different models. Besides my dissertation research, I was also able to study convergent evolution in depth, uncover patterns of molecular convergence, and refine the widely used methods for identifying excessive convergence. My work on convergent evolution proceeded largely due to my independent funding through the NIH supported Genetics, Cellular, and Molecular Sciences Training Grant at Indiana University. Thanks to Dr. Hahn’s vast expertise and collaborative network, I have led and been involved in several projects involving inference of species relationships in the presence of confounding phylogenetic factors, including the comparative phylogenomics portion of the i5K pilot project. I have also written and released software to more accurately infer gene gain and loss rates (part of CAFE version 3; 557 citations), accurately place whole genome duplications on a phylogeny (GRAMPA; 61 citations), and assign quality scores to genome assemblies (Referee; 15 citations). I also received extensive theoretical training in phylogenetics and machine learning thanks to research rotations with Drs. Elizabeth Housworth and Haixu Tang, respectively.**

**Since my post-doctoral work in Dr. Jeffrey Good’s lab at the University of Montana and subsequently starting as a Bioinformatics Scientist in the Informatics Group at Harvard University, I have become even more familiar with the details of phylogenomic analyses and methods development. As a post-doc, I spearheaded a large-scale comparative genomics project in murine rodents in which we infer the phylogeny of over 200 Old World mice and rats and quantified their molecular evolution. During that project, I developed methods to account for phylogenetic discordance in large-scale comparative datasets, which I’ve carried over in my current position as I’ve helped develop the new version of PhyloAcc (current pre-print). The new version now accounts for gene tree discordance when estimating shifts in substitution rates in non-coding genomic regions. My time in the Harvard Informatics group has also solidified my passion for developing user-friendly and well-documented bioinformatics tools and workflows, an area in which I find the field can sometimes be lacking. As genomics has progressed as a field, certain bioinformatics tasks have become common and demand succinct and fast programs to do them. However, as genomics is an experimental science even the most common tasks will need to be flexible when the experiment demands so. As such, I work to design my tools to perform core functions well, but to also be flexible to accommodate the experimental nature of the field.**

**I also enjoy teaching and mentoring scientists at levels from undergrad to postdoc. For three successive years, I have developed and led workshops for a variety of venues including in my current position, and at the Conservation Genetics (ConGen) symposium at the University of Montana. My workshops, which typically address 30 early career researchers, introduce the fundamentals of bioinformatics, programming, and data visualization as well as advanced topics such as genome assembly. My goal is for participants to develop a broadly applicable understanding of the general procedures and objectives of bioinformatic workflows, so that they can work not only with current data tools, but with the next generation of cutting-edge bioinformatics software. Starting as a grad student, I also mentored undergraduate students one-on-one on research projects, a rewarding activity that I continue in my current position as I also routinely mentor grad students and post-doctoral researchers with their specific genomics projects.**

**Moving forward, I want to continue to identify patterns of molecular and structural evolution and to build open-source tools that can be used by the genomics community. I envision my** work **existing at the boundary of biological and computational/data science where I will be able to use my skills to develop novel and modern teaching and mentoring programs that integrate scientific lines of inquiry from traditional biology research with best practices from a computational standpoint. I strive to emphasize diversity in my mentorship and teaching and will continue to do so as I advance with the belief that the best research can only be done when all groups of people have equal opportunities to succeed. With these tenants of teaching and mentorship I hope to develop a question driven line of research, centered phylogenetic inference and identifying the forces and patterns of molecular evolution. I endeavor to learn skills that give me the ability to answer questions of interest in these fields based on existing data and methods, but also to be able to generate my own data and develop my own methods when the existing ones are insufficient.**

1. **Thomas GWC**, Wang RJ, Puri A, Harris RA, Raveendran M, Hughes DST, Murali SC,

Williams LE, Doddapaneni, Muzny DM, Gibbs RA, Abee CR, Galinski MR, Worley KC, Rogers J, Radivojac P, Hahn MW. 2018. Reproductive longevity predicts mutation rates in primates. *Current Biology*. 28(19):3193-3197.

1. **Thomas GWC**, Dohmen E, Hughes ST, Murali SC, Poelechau M, Glastad K, …, Chipman AD, Waterhouse RM, Bornberg-Bauer E, Hahn MW, Richards S.2020. Gene content evolution in the Arthropods. *Genome Biology*. 21(15).
2. **Thomas GWC** and Hahn MW. 2015. Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. *Molecular Biology and Evolution*. 32(5):1232-1236.
3. **Thomas GWC** and Hahn MW. 2019. Referee: reference genome quality scores. *Genome Biology and Evolution*. 11(5):1483-1486.

**B. Positions and Honors**

**Positions and Employment**

|  |  |
| --- | --- |
| 2011 - 2012 | Associate Instructor, I308 – Information Representation, Indiana University |
| 2014 | Associate Instructor, Z620/I590 – SNP Discovery and Population Genetics, Indiana University |
| 2015 - 2016 | Associate Instructor, I211 – Information Infrastructure Indiana University |
| 2017 - 2019 | Research Assistant, Dr. Matthew Hahn’s lab, Indiana University |
| 2019 - 2021 | Postdoctoral Researcher, Dr. Jeffrey Good’s lab, University of Montana |
| 2021 - | Bioinformatics Scientist, Informatics Group, Harvard University |

**Other Experience and Professional Memberships**

|  |  |
| --- | --- |
| 2012 - 2014 | Founding member, graduate student advisor, and treasurer (2012 only), Bioinformatics Club, Indiana University |
| 2014 | Mentor, Jim Holland Summer Research Program for high school students, Indiana University |
| 2016 - 2018 | Mentor, Computer Science Independent Study for undergraduate students, Indiana University |
| 2018 - 2019 | Mentor, CEWiT Research Experience for Undergraduate Women, Indiana University |
| 2013 - 2015 | Member, Society for Molecular Biology and Evolution |
| 2015 - 2017 | Member, Society for Systematic Biology |
| 2017 - 2019 | Member, Genetics Society of America |
| 2020 - 2021 | Member, American Association of Physical Anthropologists |
| 2022 - | Member, Society for Systematic Biology |

**Awards & Honors**

|  |  |
| --- | --- |
| 2009 | Ostroy Summer Internship, Purdue University |
| 2014 - 2015 | Genetics, Cellular, and Molecular Sciences Training Grant, Indiana University |
| 2020 | Indiana University Distinguished Ph.D. Dissertation Award |

**C. Contributions to Science**

1. **Early Career: During the latter part of my undergraduate degree, I became interested in bioinformatics and was able to work on protein function prediction as an undergraduate researcher. My interests broadened during my Master’s degree when I was exposed to more topics relating to genome sequencing and evolutionary biology. During this time, I was able to work on the software CAFE, which uses maximum likelihood to analyze gene family evolution by inferring gene gain and loss rates and ancestral gene counts among a set of related species. I was able to use probability distributions on observed gene counts to estimate the amount of error in the input gene count data by implementing a pseudo-likelihood gradient search within CAFE. This error estimate is then used by CAFE to achieve more accurate estimates of gene gain and loss rates.**
2. Han MV, **Thomas GWC**, Lugo-Martinez J, and Hahn MW. 2013. Estimating gene gain and loss rates in the presence of error in genome assembly and annotation using CAFE 3. *Molecular Biology and Evolution*. 30(8):1987-1997.
3. **Mutation rate variation in primates: Mutations are the source of all genetic novelty, so I was very interested in studying them as a doctoral student. I found it particularly interesting that the *rate* at which mutations arise varies between species. Since mutations arise primarily during DNA replication, a proposed explanation for this variation involves varying selective forces acting on the DNA replication and repair proteins. However, mutation rates per generation also depend on the age of the father, so I wondered if the varying ages at which we measure mutation rates contribute to their variation. I developed a model for mutation rate estimation that incorporates both changes in the underlying mutational machinery of the cell as well as life history traits. I tested this model by sequencing parent-offspring trios of owl monkeys (*Aotus nancymaae*). I found that the mutation rate in owl monkeys is much lower than in humans, but that this difference is predicted by my model simply by shifting the ages of puberty and reproduction between these two species, rather than requiring some selected changes on the DNA replication and repair proteins themselves. I went on to test whether longer mutations, such as deletions and duplications of hundreds to thousands of base pairs (CNVs), follow a similar model by sequencing parent-offspring trios of rhesus macaques (*Macaca mulatta*). I hypothesized that since the mechanism for these mutations is dependent on meiotic cell divisions that occur only once per generation that there should be no age effects for new CNVs. This hypothesis was confirmed with both the macaque data and human data from another study.**
4. **Thomas GWC** and Hahn MW. 2014. The human mutation rate is increasing, even as it slows. *Molecular Biology and Evolution*. 31(2):253-257.
5. **Thomas GWC**, Wang RJ, Puri A, Harris RA, Raveendran M, Hughes DST, Murali SC,

Williams LE, Doddapaneni, Muzny DM, Gibbs RA, Abee CR, Galinski MR, Worley KC, Rogers J, Radivojac P, Hahn MW. 2018. Reproductive longevity predicts mutation rates in primates. *Current Biology*. 28(19):3193-3197.

1. Wang RJ, **Thomas GWC**, Raveendran M, Harris RA, Doddapaneni H, Muzny DM, Capitanio JP, Radivojac P, Rogers J, Hahn MW. 2019. Paternal age in rhesus macaques is positively associated with germline mutation accumulation but not with measures of offspring sociability. *Genome Research.* 30:826-834
2. **Thomas GWC, Wang RJ, Nguyen J, Harris RA, Raveendran M, Rogers J, Hahn MW. 2019. Origins and long-term patterns of copy-number variation in rhesus macaques.** *Molecular Biology & Evolution*. 38(4):1460-1471.
3. **Comparative genomics: C**omparing genomes between related species in the context of their phylogeny provides us the opportunity to ask and answer questions about how these species evolved at the molecular level. Using comparative genomics, we can study patterns such as substitution rate variation, convergent molecular evolution, gene family evolution, gene expression and much more. I aim to uncover these patterns across the tree of life. I have **worked on detecting and refining methods for molecular convergence with the aim of being able to trace convergent phenotypes to their molecular underpinnings. With three independent transitions to an aquatic lifestyle, marine mammals provide a great opportunity to study broad convergent phenotypes. I was part of a collaboration that sequenced the genomes of several marine mammals to detect convergent substitutions among them. While we found a few promising examples of adaptive molecular convergence, the striking pattern we found that was, even among non-phenotypically convergent species, convergent substitutions are common. This non-adaptive, or background, convergence makes it difficult to detect adaptive convergence at the single amino acid level.**

**I also led the comparative portion of the i5K pilot project, in which we sequenced the genomes of 28 new arthropod species and combined them with a number of previously sequenced arthropods to form a final set of 76 species. From these species I was able to infer an arthropod species level phylogeny and trace genomic changes on the phylogeny. Using this phylogeny, I reconstructed gene family evolution throughout the history of arthropods. This allowed us to map genomic changes onto particular lineages in the phylogeny. For instance, we found several gene families regarding wing development that emerged on the lineage that gave rise to insects. We also found that rates of substitution and gene gain and loss are both consistently low among arthropods, however they are not correlated with one another. This study is the largest comparative study in arthropods to date that uses whole genome sequences and will form the basis of much further research.**

**I now lead a project to dissect the molecular evolution across murine rodents, the largest adaptive radiation of mammals. In this work I have also become more focused on studying how phylogenetic discordance affects inferences across the tree of life.**

1. Foote AD, Liu Y, **Thomas GWC**, Vinař T, …, Gibbs RA. 2015. Convergent evolution of the genomes of marine mammals. *Nature Genetics*. 47(3):272-275.
2. **Thomas GWC** and Hahn MW. 2015. Determining the null model for detecting adaptive convergence from genomic data: a case study using echolocating mammals. *Molecular Biology and Evolution*. 32(5):1232-1236.
3. **Thomas GWC**, Hahn MW, and Hahn Y. 2017. The effects of increasing the number of taxa on inferences of molecular convergence. *Genome Biology and Evolution*. 9(1):213-221.
4. **Thomas GWC**, Dohmen E, Hughes ST, Murali SC, Poelechau M, Glastad K, …, Chipman AD, Waterhouse RM, Bornberg-Bauer E, Hahn MW, Richards S. 2020. Gene content evolution in the Arthropods. *Genome Biology*. 21(15).
5. **Software and methods development: With the advent of high-throughput sequencing technologies, the field of biology has become intertwined with data science. I strive to develop and employ practical best practices from a data science and programming perspective as I develop methods and software to handle the deluge of sequence data available. While these new technologies allow researchers to ask and answer questions about biology that were never before possible, they are also error-prone, meaning methods must be developed to account for those errors. I strive to devise methods that can utilize sequence data in novel ways to answer biological questions. One example of this is the software GRAMPA (Gene-tree Reconciliation Algorithm with MUL-trees for Polyploid Analysis), which uses gene-tree topologies along with multiply-labeled species trees to identify when whole genome duplications occurred in the history of a set of species. GRAMPA is one of the only methods available that can distinguish between the two modes of whole genome duplication (auto- vs. allopolyploidy), and it uses a parsimony approach to accurately count gene gains and losses in the presence of whole genome duplications. I also aim to develop software that can mitigate the effects of technical error on downstream analyses and conclusions. In addition to my early-career efforts with error mitigation in the CAFE software, I have also developed Referee, a tool that annotates a genome assembly with a simple and intuitive quality score, and I have worked on pseudo-it, a tool that eliminates reference bias in comparative studies with iterative mapping. I am also developing the next version of PhyloAcc to account for phylogenetic discordance when estimating substitution rates in a Bayesian framework.**
6. **Thomas GWC, Ather SA, Hahn MW. 2017. Gene-tree reconciliation with MUL-trees to resolve polyploid events. *Systematic Biology*. 66(6):1007-1018.**
7. **Thomas GWC, Hahn MW. 2019. Referee: reference genome quality scores. *Genome Biology and Evolution*. 11(5): 1483-1486.**
8. **Yan H, Hu Z, Thomas GWC, Edwards SV, Sackton TB, Liu JS. 2022. PhyloAcc-GT: A Bayesian method for inferring patterns of substitution rate shifts and associations with binary traits under gene tree discordance.** *bioRxiv* **[Preprint]**. doi: 10.1101/2022.12.23.521765

**I have no other current or pending support.**