Gregg Thomas, Ph.D.

greggwct@gmail.com
https://gwct.github.io

CURRENT APPOINTMENT

Senior Bioinformatics Scientist
Informatics Group
Faculty of Arts and Sciences
Harvard University, Cambridge, MA

2021 – present

PROFESSIONAL APPOINTMENTS

Postdoctoral Research Associate
Division of Biological Sciences
University of Montana, Missoula, MT

2019 – 2021

EDUCATION

Distinguished Doctor of Philosophy Indiana University Bloomington, IN	Bioinformatics and Evolutionary Biology	2019
Master of Science Indiana University Bloomington, IN	Bioinformatics	2013
Bachelor of Science Purdue University West Lafayette, IN	Biology	2010

PUBLICATIONS

** = co-first authors, ** = undergraduate mentee*

1. **Thomas GWC**, Hughes JJ, Kumon T, Berv JS, Nordgren CE, Lampson M, Levine M, Searle JB, Good JM. 2025. The genomic landscape, causes, and consequences of extensive phylogenomic discordance in Murine rodents. ***Genome Biology and Evolution***. 17(2): evaf017. Link

2. Kopania EEK, **Thomas GWC**, Hutter CR, Mortimer SME, Callahan CM, Roycroft E, Achmadi AS, Breed WG, Clark NL, Esselstyn JA, Rowe KC, Good JM. 2025. Sperm competition intensity shapes divergence in both sperm morphology and Link

reproductive genes across murine rodents. ***Evolution***. 79(1):11-27. *Editor's Choice January 2025.*

3. Li Y, **Thomas GWC**, Richards S, Waterhouse RM, Zhou X, Pfrender ME. 2024. Rapid evolution of mitochondrion-related genes in haplodiploid arthropods. ***BMC Biology***. 22, 229. Link

4. **Thomas GWC**, Gemmell P, Shakya SB, Hu Z, Liu JS, Sackton TB, Edwards SV. 2024. Practical guidance and workflows for identifying fast evolving non-coding genomic elements using PhyloAcc. ***Integrative and Comparative Biology***. 64(5):1513-1525. Link

5. Mirchandani C, Shultz AJ, **Thomas GWC**, Smith SJ, Baylis M, Arnold B, Corbett-Detig R, Enbody E, Sackton TB. 2023. A fast, reproducible, high-throughput variant calling workflow for population genomics. ***Molecular Biology and Evolution***. 41(1): msad270. Link

6. Schiebelhut LM, …, **Thomas GWC**, …, Luikart G. 2023. Genomics and conservation: Guidance from training to analyses and applications. ***Molecular Ecology Resources***. 24. e13893. Link

7. Yan H*, Hu Z*, **Thomas GWC***, Edwards SV, Sackton TB, Liu JS. 2023. PhyloAcc-GT: A Bayesian method for inferring patterns of substitution rate shifts on targeted lineages accounting for gene tree discordance. ***Molecular Biology and Evolution***. 40(9):msad195. Link

8. Moore EC, **Thomas GWC**, Mortimer S, Kopania EEK, Hunnicutt KE, Clare-Salzler ZJ, Larson EL, Good JM. (2022). The evolution of widespread recombination suppression on the dwarf hamster (*Phodopus sungorus*) X chromosome. ***Genome Biology and Evolution***. 14(6):evac080. Link

9. **Thomas GWC**, Wang RJ, Nguyen J**, Harris RA, Raveendran M, Rogers J, Hahn MW. (2021). Origins and long-term patterns of copy-number variation in rhesus macaques. ***Molecular Biology and Evolution***. 38(4):1460-1471. Link

10. Sun C, …, **Thomas GWC**, …, Mueller RL. (46 co-authors). 2021. Genus-wide characterization of bumblebee genomes reveals variation associated with key ecological and behavioral traits of pollinators. ***Molecular Biology and Evolution***. 38(2):486-501. Link

11. Wang RJ, **Thomas GWC**, Raveendran M, Harris RA, Doddapaneni H, Muzny DM, Capitanio JP, Radivojac P, Rogers J, Hahn MW. 2020. Paternal age in rhesus macaques is positively associated with germline mutation accumulation but not with measures of offspring sociability. ***Genome Research***. 30:826-834. Link

12. **Thomas GWC**, Dohmen E, Hughes ST, Murali SC, Poelechau M, Glastad K, …, Chipman AD, Waterhouse RM, Bornberg-Bauer E, Hahn MW, Richards S. (79 co-authors). 2020. The genomic basis of Arthropod diversity. ***Genome Biology***. 21(15).

13. Bentz AB, **Thomas GWC**, Rusch DB, Rosvall KA. 2019. Tissue-specific expression profiles and positive selection analysis in the tree swallow (*Tachycineta bicolor*) using a *de novo* transcriptome assembly. ***Scientific Reports***. 9:15849. Link

14. **Thomas GWC** and Hahn MW. 2019. Referee: reference assembly quality scores. ***Genome Biology and Evolution***. 11(5):1483-1486. Link

15. Rogers J, …, **Thomas GWC**, …, Jolly CJ, Gibbs RA, Worley KC. (78 co-authors). 2019. The comparative genomics and complex population history of *Papio* baboons. ***Science Advances***. 5(1). Link

16. Da Lage J-L, **Thomas GWC**, Bonneau M, Courtier-Orgogozo V. 2019. Evolution of salivary glue genes in Drosophila species. ***BMC Evolutionary Biology***. 19(36). Link

17. Prost S, Armstrong EE, Nylander J, **Thomas GWC**, Suh A, Petersen B, Dalen L, Benz BW, Blom MPK, Palkopoulou E, Ericson PGP, Irestedt M. 2019. Comparative analyses identify genomic features potentially involved in the evolution of birds-of-paradise. ***GigaScience***. 8(5). Link

18. **Thomas GWC**, Wang RJ, Puri A**, Harris RA, Raveendran, 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. Link

19. Warren WC, García-Pérez R, …, **Thomas GWC**, …, Schartl M. (28 co-authors). 2018. Clonal polymorphism and high heterozygosity in the celibate genome of the Amazon molly. ***Nature Ecology and Evolution***. 2:669-679. Link

20. Schoville SD, Chen YH, …, **Thomas GWC**, …, Richards S. (59 co-authors). 2018. A model species for agricultural pest genomics: the genome of the Colorado potato beetle, *Leptinotarsa decemlineata* (Coleoptera: Chrysomelidae). ***Scientific Reports***. 8(1931). Link

21. Palesch D, Bosinger SE, …, **Thomas GWC**, …, Silvestri G. (30 co-authors). 2018. Sooty mangabey genome sequence provides insight into AIDS resistance in a natural SIV host. ***Nature***. 553:77-81. Link

22. **Thomas GWC**, Ather SA**, and Hahn MW. 2017. Gene-tree reconciliation with MUL-trees to resolve polyploidy events. ***Systematic Biology***. 66(6):1007-1018. Link

23. **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. Link

24. Warren WC, …, **Thomas GWC**, …, Freimer NB. (61 co-authors). 2015. The genome of the vervet (*Chlorocebus aethiops sabaeus*). **Genome *Research***. 25(12):1921-1933. Link

25. **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. Link

26. Foote AD*, Liu Y*, **Thomas GWC***, Vinař T*, …, Gibbs RA. (26 co-authors). 2015. Convergent evolution of the genomes of marine mammals. ***Nature Genetics***. 47(3):272-275. Link

27. Neafsey DE, Waterhouse RM, …, **Thomas GWC**, …, Besansky NJ. (120 co-authors). 2014. Highly evolvable malaria vectors: The genomes of 16 Anopheles mosquitoes. ***Science***. 347. Link

28. Montague MJ, …, **Thomas GWC**, … Warren WC. (25 co-authors). 2014. Comparative analysis of the domestic cat genome reveals genetic signatures underlying feline biology and domestication. ***Proc Natl Acad Sci USA***. 111(48):17230-17235. Link

29. Carbone L, … **Thomas GWC**, …, Gibbs RA. (93 co-authors). 2014. Gibbon genome and the fast karyotype evolution of small apes. ***Nature***. 513:195-201. Link

30. **Thomas GWC** and Hahn MW. 2014. The human mutation rate is increasing, even as it slows. ***Molecular Biology and Evolution***. 31(2):253-257. Link

31. 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. Link

PRE-PRINTS

1. **Thomas GWC**, McKibben MTW, Hahn MW, Barker MS. A comprehensive examination of Chelicerate genomes reveals no evidence for a whole genome duplication among spiders and scorpions. *bioRxiv*. Link

PRESENTATIONS & POSTERS

1. **Inferring substitution rate shifts in a phylogeny with PhyloAcc**
   The Allied Genetics Conference (TAGC), Washington, D.C. — 2024
   Poster

2. **Quantifying and mitigating reference bias in comparative genomics**
   Evolution Meeting, Albuquerque, NM — 2023
   Contributed talk

3. **Prioritizing loci for ILS-aware rate analyses using phylogenetic concordance factors**
   Society of Molecular Biology and Evolution Global Symposium 2 (Sustainability, Equity, and Efficiency in Computational Biology), virtual — 2022
   Contributed talk

4. **Molecular and morphological evolution across the most species-rich radiation in mammals**
   Evolution Meeting, Cleveland, OH — 2022
   Poster

5. **Prioritizing loci for ILS-aware rate analyses using phylogenetic concordance factors**
   Evolution Meeting, Cleveland, OH — 2022
   Contributed talk

6. **Speciation and introgression across the most species-rich radiation in mammals**
   Population, Evolutionary, and Quantitative Genetics Conference, Pacific Grove, CA — 2022
   Platform talk

7. **Patterns of genomic variation across the tree of life**
   University of Massachusetts Lowell Department of Biology Seminar, Lowell, MA — 2022
   Invited talk

8. **Patterns of genomic variation across the tree of life**
   Harvard Museum of Comparative Zoology Seminar, Cambridge, MA — 2022
   Invited talk

9. **Pedigree sequencing and mutation rate variation in primates**
   Meeting of the American Association of Biological Anthropologists, virtual — 2021
   Invited talk

10. **Origins and long-term patterns of genomic variation across the tree of life** — 2021
    Harvard University, Bioinformatics group, virtual
    Invited talk

11. **Origins and long-term patterns of genomic variation across the tree of life** — 2021
    Binghamton University, Department of Biological Sciences, virtual
    Invited talk

12. **The origins and patterns of genomic variation across the tree of life** — 2021
    Utah Valley University, Department of Biology, virtual
    Invited talk

13. **Patterns of molecular evolution in Arthropods** — 2020
    Arthropod Genomics Symposium, virtual
    Invited talk

14. **Causes and consequences of structural variation in the *Macaca mulatta* genome** — 2019
    First author Nguyen J**
    Center of Excellence for Women & Technology Research Experience for Undergraduates Symposium, Bloomington, IN
    Poster

15. **Reproductive longevity predicts mutation rates in primates** — 2018
    Population, Evolutionary, and Quantitative Genetics Conference, Madison, WI
    Platform talk

16. **The evolution of the genes and genomes of 76 arthropod species** — 2017
    Evolution Meeting, Portland, OR
    Regular talk

17. **The evolution of the genes and genomes of 76 arthropod species** — 2017
    Arthropod Genomics Symposium, Notre Dame University, South Bend, IN
    Invited talk

18. **Gene-tree reconciliation with MUL-trees for polyploidy analysis** — 2016
    Evolution Meeting, Austin, TX
    Regular talk

19. **Accounting for sequencing error in phylogenetics** — 2015
    Society of Systematic Biologists, University of Michigan, Ann Arbor, MI

Lightning Talk

20. **Inferring molecular convergence from genomic data**
    Midwest Ecology and Evolution Conference, Indiana University, Bloomington, IN — 2015
    Contributed talk

21. **Convergent evolution of the genomes of marine mammals**
    Society for Molecular Biology and Evolution, San Juan, Puerto Rico — 2014
    Contributed talk

22. **Convergent evolution of the genomes of marine mammals**
    Genetics, Cellular, and Molecular Sciences Symposium, Bloomington, IN — 2014
    Poster

RESEARCH EXPERIENCE

Senior Bioinformatics Scientist
FAS Informatics & Scientific Applications Group — 2021 – present
Harvard University, Cambridge, MA

- Develop and maintain methods to account for phylogenetic discordance during Bayesian estimation of substitution rates (PhyloAcc).
- Collaborate with research groups at the university on phylogenetic projects related to medicinal plants and coat color change in hamsters.
- Develop genomic and phylogenetic software with a main goal for reproducible and accurate inference, including Snakemake pipelines for whole genome alignment (cactus-snakemake) and annotation of degeneracy of coding transcripts (degenotate).
- Design and teach workshops on bioinformatics, programming, and AI use for graduate students at the university.
- Assist and consult with researchers from across the university with genomic and phylogenetic analyses.
- Teach workshops on how to use Bayesian substitution rate estimation software (PhyloAcc).
- Present work to researchers outside the university at conferences regarding evolutionary genomics and bioinformatics.
- Develop and maintain the group's website.

Postdoctoral Research Associate
Laboratory of Jeffrey Good
Division of Biological Sciences — 2019 – 2021
University of Montana, Missoula, MT

- Lead a comparative project to study molecular evolution and phylogenetics in murine rodents using whole exome sequences from 210 species and whole genome sequences from dozens of species.

- Devised best-practices for assembly and annotation of a large sample of exomes.
- Update and maintain software released by the lab to automate reference-guided genome assembly through iterative mapping (pseudo-it).
- Analyze the phylogenetic relationships of newly sequenced rodent genomes using the reference genomes and genetic resources from the well annotated mouse and rat systems to build an empirical landscape of phylogenetic discordance across chromosomes.
- Applied for grants (NIH NRSA) to expand whole genome sampling of murine rodents to study phylogenetic discordance, patterns of molecular evolution, convergent evolution, and sex chromosome rearrangements.
- Administrator of lab's computational resources, including two 32 core 200GB servers and one 192TB NAS, and the lab github account.
- Mentor and guide graduate and undergraduate students in the lab regarding computational methods in genomics, phylogenetics, and molecular evolution.

Research Assistant
Laboratory of Matthew Hahn
School of Informatics, Computing, and Engineering 2012 – 2019
Department of Biology
Indiana University, Bloomington, IN

- Developed a method to estimate genome assembly and annotation error from gene count data using CAFE's error model function (caferror).
- Studied patterns of convergent evolution in marine mammals and echolocating mammals and devised best practices for identifying molecular convergence.
- Devised a method to infer the presence and mode of polyploidy from gene tree topologies (GRAMPA).
- Modeled and observed mutation rate patterns in primates, including single nucleotide mutations and structural variants, by sequencing families of owl monkeys and macaques.
- Led the comparative phylogenetic portion of the i5K pilot project which involved analyzing the genomes of 76 arthropods.
- Wrote software to annotate genomes with quality scores (Referee).
- Participated in several collaborations by performing comparative analyses, such as phylogeny reconstruction and assessment, gene family analysis, and positive selection scans.

TEACHING EXPERIENCE

Instructor
Informatics Group Workshops 2023-present
Harvard University

- Developed and taught workshops to biology grad students and postdoctoral fellows.
- Designed hands-on activities in R markdown to demonstrate basic R syntax, data manipulation with the tidyverse, and plotting with ggplot.

- Designed hands-on activities in R markdown to teach about command line tools like awk, grep, samtools, bedtools, and bcftools.
- Introduced basic bioinformatics file formats to students, such as fasta, fastq, bed, bam, gff, and vcf files.
- Demonstrated best practices in bioinformatics.
- Revamped curriculum for AI use in a scientific context, including planning a University-wide collaboration between data science groups to propose consistent and critical use of AI.

Guest Instructor
OEB 275R: Comparative Genomics: Phylogenetic Approaches to Linking Genomes and Phenotypes — 2022
Prof. Scott Edwards
Harvard University

- Led graduate students in discussions on comparative genomics and bioinformatics
- Designed hands-on activities to demonstrate how to infer accelerated substitution rates on a phylogeny using tools such as PhyloAcc, git, and RStudio and R markdown.
- Guided the students through a web-based workshop on how to use Bayesian substitution rate estimation software (PhyloAcc).
- Demonstrated best practices in bioinformatics.

Instructor
Conservation Genetics and Population Genomics course (ConGen) — 2020 – 2023
Virtual course
University of Montana

- Gave keynote lecture on genome sequencing and assembly.
- Designed hands-on activities for a 2-hour workshop on genome assembly and read mapping.
- Designed and presented a workshop to teach introductory bioinformatics skills including project organization, common bioinformatics file formats, and examples of basic bioinformatics tasks
- Met one-on-one with students during office hours to discuss and give advice about their data.

Student Mentor
School of Informatics, Computing, and Engineering, Department of Biology — 2014 – 2019
Indiana University, Bloomington, IN

Provided guidance to high school and undergraduate students in conceptualizing evolution by involving them in various computational projects, providing a basis in programming, data analysis, and scholarship.

- CEWiT Research Experience for Undergraduate Women — 2018 – 2019
- Computer Science Independent Study — 2017 – 2018
- Computer Science Independent Study — 2016 – 2017
- Jim Holland Summer Science Research Program — 2014

Teaching Assistant
School of Informatics, Computing, and Engineering,
Department of Biology
Indiana University, Bloomington, IN
2011 – 2016

Taught lab sessions, led class discussions, graded assignments, and met with students individually to assist them.

- INFO-I211: Information Infrastructure 2014, 2016
- BIOL-Z620/INFO-I590: SNP Discovery and Population Genetics 2014
- INFO-I308: Information Representation 2011 – 2012

PROFESSIONAL SERVICE

Graduate Student Advisor
Indiana University Bioinformatics Club
Indiana University, Bloomington, IN
2012 – 2014

Served as a co-founding member and treasurer (2012 only) to raise awareness of bioinformatics and associated opportunities for undergraduate and graduate students by facilitating group projects and discussions, tours, and social events.

Peer Review

- *G3*
- *PLoS ONE*
- *Molecular Biology and Evolution*
- *New Phytologist*
- *Pacific Symposium of Biocomputing*
- *Genes*
- *Nature Communications*
- *Systematic Biology*
- *Genomics*
- *GigaScience*
- *Genome Biology*
- *Journal of Molecular Evolution*
- *Evolution Letters*
- *Science Advances*
- *Genome Biology and Evolution*
- *Society of Systematic Biologists Graduate Student Research Award*
- *Molecular Ecology Resources*
- *Molecular Ecology*
- *BMC Ecology and Evolution*
- *PeerJ*
- *BMC Biology*
- *Insect Systematics and Diversity*

GRANTS

Genetics, Cellular, and Molecular Sciences Training Grant
Department of Biology
Indiana University, Bloomington, IN
2014 – 2015

Sandy Ostroy Summer Research Award for Undergraduates
Department of Biology
Purdue University, West Lafayette, IN
2008

AWARDS

Distinguished Ph.D. Dissertation Award
The University Graduate School
Indiana University, Bloomington, IN
2020

SOFTWARE & RESOURCES

Cactus-snakemake
https://github.com/harvardinformatics/cactus-snakemake
- Snakemake workflows to run the Cactus whole genome alignment software efficiently on an HPC cluster, allowing for GPU use. Facilitates creation of new alignments and editing previously generated alignments, as well as creation of pan-genome graphs.

Harvard FAS Informatics website
https://informatics.fas.harvard.edu/
- I develop and maintain the Harvard FAS Informatics Material for Mkdocs website.

Harvard FAS Informatics workshops
https://informatics.fas.harvard.edu/events-workshops/
- Workshops in our group that I have contributed to developing and teaching include Bioinformatics Tips & Tricks (now Genomics on the Command Line), Healthy Habits for Data Science, Introduction to R, Python Intensive, Introduction to Snakemake, Command Line 101, and Population Genomics: Getting started with SNPArcher

PhyloAcc: Bayesian estimation of substitution rates while accounting for phylogenetic discordance
https://phyloacc.github.io/
- Contributed to version 2.0 by developing methods to use concordance factors to mitigate the effects of phylogenetic discordance on inferences and optimized and improved the usability of software.

degenotate: Annotation of coding sites with codon degeneracy and MK tables
https://github.com/harvardinformatics/degenotate
- This software outputs a bed file containing information about the degeneracy of every coding site in a genome as well as counts of polymorphisms for MK tests.

bonsai: Tree pruning with concordance factors
https://github.com/gwct/bonsai
- Prunes large phylogenetic trees to maximize the concordance of the underlying alignments.

pseudo-it: Pseudo-genome assembly with iterative mapping
https://github.com/goodest-goodlab/pseudo-it

- This software iteratively maps reads to generate a pseudo-assembly to reduce reference bias. I re-wrote this software to modularize it and speed it up.

ConGen workshops
https://gwct.github.io/congen/

- I built these websites as a workshop resource for students during the Conservation Genomics Course.

Referee: Reference genome quality scores
https://gwct.github.io/referee

- This software uses genotype likelihoods from reads mapped back to their assembly to calculate a quality score for every position in the assembled genome.

***Drosophila* 25 species phylogeny**
http://dx.doi.org/10.6084/m9.figshare.5450602

- As part of a larger project, I inferred the phylogeny of 25 *Drosophila* species and published it standalone on FigShare as a resource for others to use.

GRAMPA: Gene-tree Reconciliation Algorithm with MUL-trees for Polyploid Analysis
https://gwct.github.io/grampa.html

- Given a singly-labeled species topology and a set of corresponding gene-trees, this software can infer if any whole genome duplications have occurred and, if so, infer the mode of polyploidization and the placement on the phylogeny.

i5K Phylogenomics Website
https://arthrofam.org

- With the vast amount of data involved in the i5K pilot project, I developed this website to organize and share the phylogenetic and comparative results with colleagues.

caferror
https://github.com/hahnlab/CAFE5

- Part of CAFE version 3, I wrote this program to use CAFE's error modeling function to estimate genome assembly and annotation error. The algorithm is now integrated in CAFE version 5.

REFERENCES

Dr. Matthew Hahn
Ph.D. Advisor
Departments of Biology and Computer Science
Indiana University
mwh@iu.edu

Dr. Jeffrey Good
Postdoctoral Advisor
Division of Biological Sciences
University of Montana
jeffrey.good@umontana.edu

Dr. Timothy Sackton
Current Supervisor
FAS Informatics
Harvard University
tsackton@g.harvard.edu