GenoPhenoEnvo Governance and Operations Manual

This is a living document. Changes are expected throughout the life of the project.

Introduction

This Governance and Operations Manual (GOM) defines standard operating procedures and various policies created to clarify, support, and further the goals of the Genomics Phenotypes and Environment (GenoPhenoEnvo) Research Team.

Primary Office

To mitigate the effects of climate change on public health and conservation, we need to better understand the dynamic interplay between biological processes and environmental effects. Our team proposes to significantly advance technologies in Machine Learning (ML) and create a new interdisciplinary field, **computational ecogenomics** by (1) designing ML techniques for encoding heterogeneous genomic and environmental data, and mapping them to multi-level phenotypic traits, (2) reducing the amount of necessary training data, and (3) developing interactive visualizations to better interpret ML models and their outputs. These advances will responsibly and transparently inform policy to maximize resources during this crucial window for planetary health, while revealing underlying biological mechanisms of response to stress and evolutionary pressure.

Mission

Develop a machine learning framework capable of predicting phenotypes based on multi-scale data about genes and environments.

Vision

Foster innovative solutions using ethical data science approaches to solve grand challenges in a collaborative and inclusive institute.

Goals

- Predict phenotypes from agricultural data
- Test the phenotype model, scaling from a single species to ecological communities
- Prepare for sustainability - of our work, of the institute, resilience of the work products
- Break the bottleneck in linking genes, environments, and traits
- Demonstrate transferable applications of our work
- Commit to Open Science
- Produce software and trained models that can be used and contributed to by other groups

Participation and Roles

The NSF Ideas HDR grant was awarded to a distributed team from 4 institutions: Oregon State University, Michigan State University, University of Arizona, and Tufts University, and is strategically organized across two working groups: **internal** and **community**.

With a **technical focus**, members from both working groups will work together to realize the goals stated in the research strategy proposed to NSF. With an **operational focus**, some members from the internal working group will serve the overall vision and functioning of the members of this contract. A Program Manager (PM) will be accountable for managing the operationalization of the program, and for ensuring bidirectional communication between the awarded sites, the working groups, and the NSF. The PM is responsible for the following functions: Program Meeting Coordination, Communications, Project Management Infrastructure, and Reporting. Support and resources outside of this scope of work are at the discretion and responsibility of the Principal Investigators of the project.

We will coordinate all activities as follows:

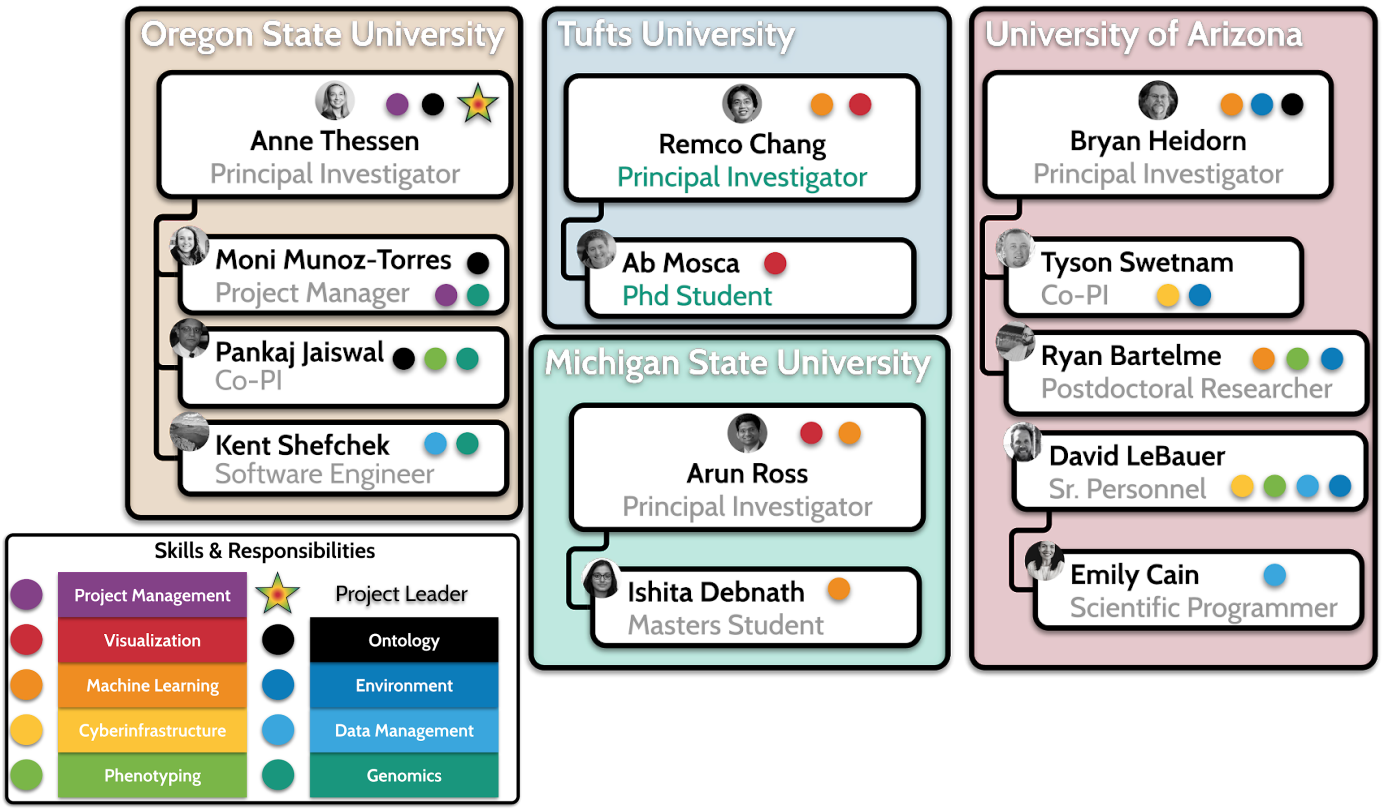
- **GenoPhenoEnvo Leads:** the leadership is formed by the project PIs and the Program Manager.
- **GenoPhenoEnvo Internal:** this team is formed by all team members listed in the table below.
- **GenoPhenoEnvo Community:** this is formed by all members of the scientific community who wish to volunteer their time with us for the purpose of advancing the new field of computational ecogenomics with us. This is an extended community of collaborators that we will likely incite to take part in the next proposal as part of a potential institute.

Our Team

Researcher	Background and experience	Project role / leadership
	At the Translational and Integrative Sciences Lab in the Environmental & Molecular Toxicology Department, **Dr. Anne Thessen** creates knowledge from data using semantic technology and informatics. She also leads the development of semantic environmental data representations and data harmonization workflows.	• PI (Overall project lead) • Knowledge graph engineering • Management of PM and the OSU technical lead • Biology domain expert
	**Dr. Arun Ross** develops ML techniques that can be used to learn complex relationships in medical, biological, and biometric data.	• PI • ML models for phenotype prediction • Advise a PhD student
	**Dr. Bryan Heidorn** integrates heterogeneous data and data science methods across biological disciplines, particularly data from natural environments.	• PI • Environmental data capture • Liaising with NPN, PhenoCam and NEON • Advise PostDoc
	**Dr. Remco Chang** creates visualization systems and interactive visual machine learning systems for everyday users; the most notable of these include a financial fraud detection system for Bank of America [14], a visual bridge management system for the Department of Transportation [60], and a visualization of biomechanical motion for researchers at Brown University.	• PI • Human-computer interaction (Chang and Ross teams will collaboratively develop tools for exploring and explaining trained ML models) • Advise PhD student (Ab Mosca)
	As the spatial data infrastructure lead for CyVerse, **Dr. Tyson Swetnam** manages multiple environmental data streams using advanced cyberinfrastructure; he provides geoinformatics expertise to numerous research groups including NEON and LTAR.	• CoPI • Environmental data transformation • co-advise PostDoc • CyVerse technical lead
	**Dr. Pankaj Jaiswal** investigates comparative genomics and phenomics in crop species using ontologies and data standards as a part of the Gramene and Planteome projects.	• CoPI • Agricultural genomics and phenomics data & analysis • Supervision of a postdoc (Curation of plant data using ontologies)
	**Dr. David LeBauer** is the director of Data Science for the Agricultural Experiment Station at the University of Arizona. His research is focused on using science to engineer more sustainable and productive crops and agricultural systems, and developing open software and data to integrate data and knowledge across disciplines.	• Senior Personnel • Supervise data scientist / programmer • Identify and curate TERRA-REF data.
	**Dr. Monica Munoz-Torres** is an Assistant Professor in the Environmental & Molecular Toxicology Department at Oregon State University. She is a Program Manager for the Translational and Integrative Science Lab and the Monarch Initiative, with extensive experience as Project Lead for scientific software development both in public & private sectors.	• Program Manager • Genomics and ontologies domain expert.
	**Emily Cain** is a scientific programmer with the Data Science and Infrastructure for Agriculture Group (DIAG) at the University of Arizona with a focus on curating data collected from the Agricultural Experiment Station for machine learning research.	• TERRA-REF data curation and annotation • Facilitating ML model building

	**Ab Mosca** is a PhD student at Tufts University working under Remco Chang. Their research focuses on using visualization to make advanced analytics more accessible to non-technical people.	• Help develop tools for exploring and explaining trained ML models
	**Ishita Debnath** is a Master's student working under Dr. Arun Ross at Michigan State University.	• Develop ML models for phenotype prediction
	**Kent Shefchek** is a Scientific Software Engineer at Oregon State University..	• Data science and software development
	**Dr. Ryan Bartelme** is a postdoctoral researcher who specializes in ecology, genomics, microbiology and data science.	• Develop Bayesian Network Model • Assist in CyVerse integration • NEON data analysis

Organizational Structure



Operations

We are ensuring that the load for oversight, project management, and coordination is distributed among a group of experts who can make day-to-day decisions on their respective components, while putting in place a communication strategy to keep each other informed on key strategic changes and decisions. The plan outlined here will efficiently and effectively distribute decision-making authority across our highly diverse team to maximize our effectiveness and efficiency.

Communications

Our team is highly interdisciplinary and geographically distributed. As such, communication will be challenging, but very important for a successful collaboration.

- **Mailing lists**: The mailing lists will be the primary routes of communications, and teleconferences and in-person meetings will be announced using these.
    - For the internal group: We request members to send communications about all topics concerning only the internal team using the following address: genophenoenvo-internal@googlegroups.com
    - For the community: All members of the community interested in learning and discussing about the project, and all other questions can be directed to the user mailing list at the following address: genophenoenvo@googlegroups.com
- **Technical contributions**: Communication regarding technical contributions and questions are expected in the Github repository (see *GitHub Repositories* below), with relevant technical discussion can happen on Slack. If using Slack, please use the "@" symbol to raise a notification for a team member of interest. If everyone needs to see the message, please use @channel. Our team strives to respond to communication within 48 hours during regular business hours:
    - Slack workspace:
        - Workspace name: GenomePhenome
        - Channels: #general, #predicting-phenotype
        - URL: genomephenome.slack.com
- **Style Guide:** For all outward facing, written communication, please see our style guide.
- **Manuscripts:** Proposed workflows for manuscript preparation include:
    - ***Google Documents*** - for manuscripts with "traditional" content that includes text, tables, and figures, and does not include workflows or scripts for community use.
    - ***R markdown*** - a more widely accepted platform when including analytics workflows.
- **Requests for added clarification**: During meetings, subgroup participants should feel empowered to politely interrupt a colleague to get a definition of an acronym or jargon. If needed, a 10 min "seminar" can be requested to be scheduled at a team meeting.

Work Procedures

In general, work such as program code, schemas, and documentation should be done in the corresponding GenoPhenoEnvo GitHub repository. Changes, additions, and revisions to all code should be requested on the GitHub trackers. Blockers will be addressed in a timely fashion by either being solved directly or finding an alternate path to success.

All changes to the codebase require a pull request, which shall require approval from two members of the community. Proposed changes should be on GitHub for at least 72 hours prior to discussion on a technical meeting for final decisions, in order to enable sufficient community comment. Decisions should be recorded on the GitHub tracker. Links to the issue should be made in the relevant commit message(s) and the pull request. If no objections are forthcoming within 72 hours, then the pull request can be merged into the respective repository.

- **GitHub Repositories**:
    - GenoPhenoEnvo GH organization: https://github.com/genophenoenvo
    - List specific repos within our org:
        - *docker*: This is where docker images will reside before CyVerse Integration
        - *documentation*: Markdown formatted tutorials on how to use scripts, or CyVerse VICE apps
        - *genophenoevo.github.io*: Organization website development repository
        - *terraref-datasets*: Code and small datasets from TERRA-REF project; this could be test code before integration into CyVerse infrastructure
        - *genomic_data*: code repository for analysis, wrangling, reformatting, etc., of genomic data.

- **CyVerse:**
    - CyVerse will provide a data science workbench for analysis (VICE). For this project, CyVerse will host data, Jupyter Notebooks, containers, analysis running etc. Tyson Swetnam will help other members of the team with preparing and running as needed.

- **Documentation**:
    - The main source of information for our project is our website: https://genophenoenvo.github.io/
    - Our documentation will be prepared using ReadTheDocs, based on our GitHub Repositories.

Scientific and Technical Decision making

The team will default to the decisions made in the awarded proposal. Should difficulties in realizing those plans arise, and the decisions in the awarded proposal are no longer valid, new decisions and plans will be made.

- Decision-making will be undertaken in a consensus manner after open discussion with all team members.
- Each PI is encouraged and entitled to make local decisions. the Project Leader, will make the final call.
- If there are any changes to be made in the direction of the project, all PIs will work together to discuss these changes.
- All decisions will be fully documented, including the process and the logic behind the decision. These documents will be reviewed by all team members and kept available in the project Google Folder. If changes merit it, we will notify NSF.
- **Funds**: The PI's will disburse the award funds in accordance with the submitted budgets, and each PI will have oversight of their portion of the budget.

Community Practices

Our Commitment to Open Science

We are committed to open science, as such practices increase the informational value and impact of our research. The GenoPhenoEnvo team strives for open licensing whenever possible, and in following with the principles outlined in the *Voluntary commitment to research transparency* http://www.researchtransparency.org/. Open science is not necessarily in conflict with activities related to developing revenue-generating tools. Our focus should be on building community, both for support and adoption. Maintaining open practices also includes striving for reproducibility for all our analyses.
Our **analysis code, scripts, containers, software, and documentation, and original data** will be licensed BSD (or another compatible Open Software Initiative approved license) and **ingested data** will either follow the requirements of the provider or have a CC-BY or CC-0 license. We will strive for long-term accessibility and preservation of data and models as we are able.

Need to add a note about heterogeneity of data sources and the variety of licenses (@moni remember Licensing page for Monarch).

Diversity Statement

We encourage everyone to participate and are committed to building a community for all. Although we will fail at times, we seek to treat everyone as fairly and equally as possible. Whenever a participant has made a mistake, we expect them to take responsibility for it. If someone has been harmed or offended, if you are a witness to an event of this nature, it is your responsibility to both listen carefully and respectfully to the victims, as well as to take action to bring attention to the behaviour, and to make every effort to stop the behaviour. It is also our responsibility to do our best to right the wrong.

Although this list cannot be exhaustive, we explicitly honor diversity in age, gender, gender identity or expression, culture, ethnicity, language, national origin, political beliefs, profession, race, religion, sexual orientation, socioeconomic status, and technical ability. We will not tolerate discrimination based on any of the protected characteristics above, including participants with disabilities.

Code of Conduct

This code of conduct outlines our expectations for the GenoPhenoEnvo community, which includes faculty, staff, users, data providers, students, etc., as well as describes how to report unacceptable behavior. We are committed to providing a welcoming and inspiring environment for all and expect this code of conduct to be honored. All members of GenoPhenoEnvo are expected to abide by this code of conduct in agreement with the standards for professional behavior outlined and enforced by the Office of Diversity and Inclusion (ODI), Office of the Director (OD) of the National Science Foundation (NSF), at https://www.nsf.gov/od/odi/index.jsp. Our GenoPhenoEnvo community strives to:

- **Be friendly and patient**.
- **Be welcoming**: We expect cooperation from all members to help ensure a safe environment for everybody. We strive to be a community that provides a harassment-free experience for everyone, welcoming and supporting people of all backgrounds and identities. This includes, but is not limited to members of any race, ethnicity, culture, national origin, colour, immigration status, social and economic class, educational level, sex, sexual orientation, gender identity and expression, age, size, family status, political belief, religion, and mental and physical ability.
- **Be considerate**: Your work will be used by other people, and you in turn will depend on the work of others. Any decision you take will affect users and colleagues, and you should take those consequences into account when making decisions. Remember that we're a large community, so you might not be communicating in someone else's primary language.
- **Be respectful:** Not all of us will agree all the time, but disagreement is no excuse for poor behavior and poor manners. We might all experience some frustration now and then, but we cannot allow that frustration to turn into a personal attack. It's important to remember that a community where people feel uncomfortable or threatened is not a productive one.
- **Be careful in the words that we choose:** We are a community of professionals, and we conduct ourselves professionally. Be kind to others. Do not insult or put down other participants. Harassment and other exclusionary behavior aren't acceptable. This includes, but is not limited to: Violent threats or language directed against another person, Discriminatory jokes and language, Posting sexually explicit or violent material, Posting (or threatening to post) other people's personally identifying information ("doxing"), photography or recordings, Personal insults, especially those using racist or sexist terms, Inappropriate physical contact, Unwelcome sexual attention, Advocating for, or encouraging, any of the above behavior, Repeated harassment of others. Harassment also includes offensive verbal comments related to gender, gender identity and expression, age, sexual orientation, disability, physical appearance, body size, race, religion, deliberate intimidation, stalking, following, sustained disruption of talks or other events. In general, if someone asks you to stop, then stop.
- **Try to understand why we disagree:** Disagreements, both social and technical, happen all the time. It is important that we resolve disagreements and differing views constructively. Remember that we're different. Diversity contributes to the strength of our community, which is composed of people from a wide range of backgrounds. Different people have different perspectives on issues. Being unable to understand why someone holds a viewpoint doesn't mean that they're wrong. Don't forget that it is human to err and blaming each other doesn't get us anywhere. Instead, focus on helping to resolve issues and learning from mistakes.
- All members of GenoPhenoEnvo will be expected to abide by the highest research ethical standards in accordance with the Online Ethics Center for Engineering and Science (https://www.onlineethics.org/), Moore (2011), and the AI code of ethics (https://futureoflife.org/ai-principles/). Plagiarism and data fabrication will not be allowed. These and other ethics violations will be reported.

Conflict resolution

The PIs have an established track record of successful collaboration, and expect to reach common agreement on management issues by thoroughly discussing and carefully considering the pros and cons of specific actions. They do not foresee any disagreements that would negatively affect the proposed research. It is thus expected that any scientific challenges or differences of opinion will be resolved through constructive discussion among all involved individuals. If they fail to resolve the dispute within 7 days, the conflict shall be referred to an arbitration committee consisting of one impartial senior executive from each PI's institution and a fifth impartial

senior executive mutually agreed upon by the four PIs. No members of the arbitration committee will be directly involved in the research grant or disagreement. Input will be sought from the NSF Program Officer in resolving the conflict.

Reporting Issues

If you experience or witness unacceptable behavior, or have any other concerns, please report it by contacting Dr. Monica Munoz-Torres (munoztmo [at] oregonstate.edu) or Dr. David LeBauer (dlebauer [at] email.arizona.edu). All reports will be handled with discretion. In your report please include:

- Your contact information.
- Names (real, nicknames, or pseudonyms) of any individuals involved. If there are additional witnesses, please include them as well. Your account of what occurred, and if you believe the incident is ongoing. If there is a publicly available record (e.g. a mailing list archive or a public IRC logger), please include a link.
- Any additional information that may be helpful.

If you file a report, Dr. Monica Munoz-Torres or Dr. David LeBauer will contact you personally, review the incident, follow up with any additional questions, and make a decision as to how to respond. If the person who is harassing you is one of the persons designated to receive these reports, please contact the main PI, Dr. Anne Thessen (thessena [at] oregonstate.edu) instead. We will respect confidentiality requests for the purpose of protecting victims of abuse.

Attribution, Authorship, and Ownership

Publications that include results that are the work of the GenoPhenoEnvo Team should be announced to the entire subgroup *ideally at least one month in advance*. Co-authorships should reflect a common-sense assessment, but in general should reflect substantial work or intellectual contributions to the topic of the publication. Any team member can request to be included as an author on any project output. In general, we choose to take an inclusive approach to authorship, with contributions including any type or size of intellectual contribution. Authors are expected to do at least one of the following: edit or write manuscript; provide data, figure, or table; provide critical ideas; write code. Author order will be decided *a priori* following conventions of the target community. Every manuscript will include an “Author Contributions” section that specifies the role of each team member. We will strive to publish in open access. Students and postdocs are encouraged to participate in these discussions.

Conference abstracts and presentations should be shared with a minimum of 48 hours notice before submission deadlines expire, and before presenting the work. The acknowledgements section will include recognition of funding and will give credit for all other contributions.

Our **analysis code, scripts, containers, software, and documentation, and original data** will be licensed BSD (or another compatible Open Software Initiative approved license) and **ingested data** will either follow the requirements of the provider or have a CC-BY or CC-0 license. We will strive for long-term accessibility and preservation of data and models as we are able.

Acknowledgements

This code of conduct is based on the Open Code of Conduct from the TODO Group and on the Code of Conduct from the Galaxy Community Conference 2019.

END OF GENOPHENOENVO GOVERNANCE DOCUMENT

==========