Genotype to Phenotype Mapping via Deep Learning

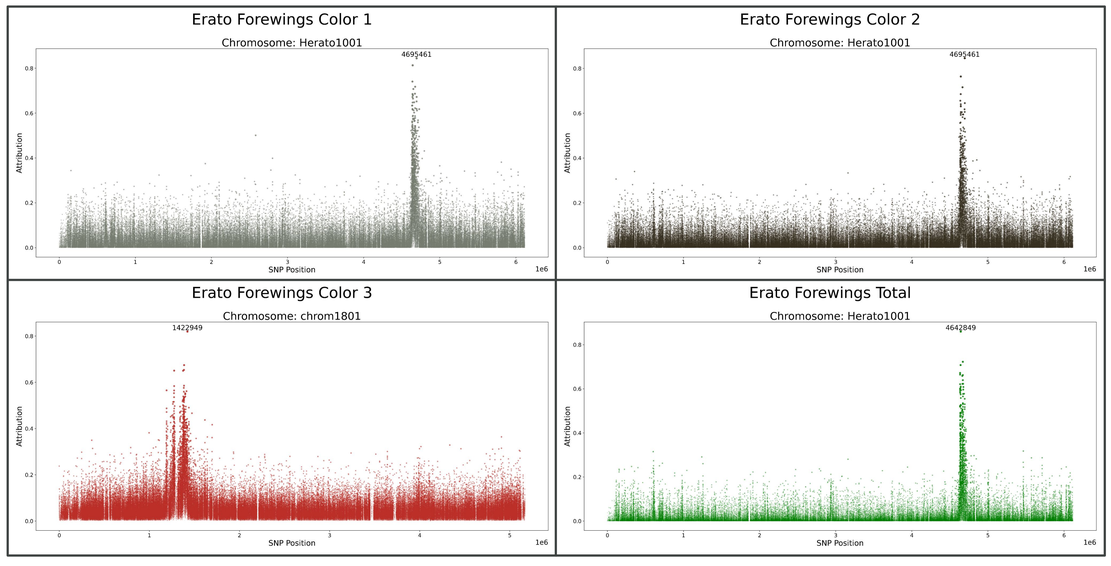
David Carlyn[1], Christopher Lawrence[2], Carlos Arias[3], Cyril Rauch[5], Owen McMillan[3], Daniel Rubenstein[2], Wei-Lun Chao[1], Tanya Berger-Wolf[1], Chuck Stewart[4]

[1]The Ohio State University, [2]Princeton University, [3]Smithsonian Tropical Research Institute, [4]Rensselaer Polytechnic Institute, [5]University of Nottingham

Abstract

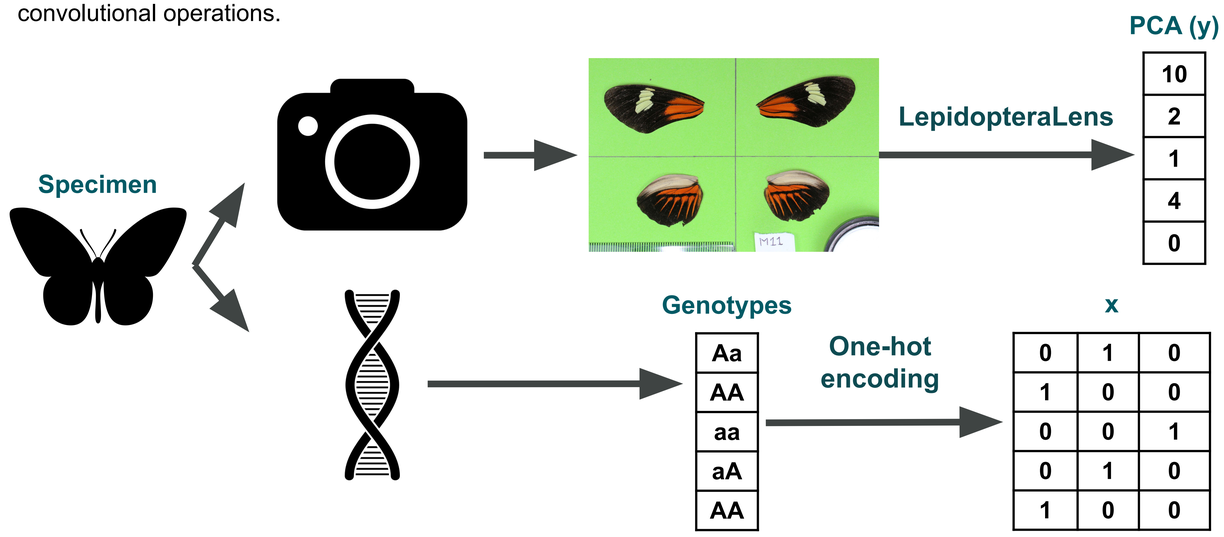
- We aim to **identify genes and mutations associated with phenotypic variation** measured in Heliconius butterfly images.
- We **train a convolutional neural network** to predict phenotypic variation from genotypes.
- Given the trained network **we apply saliency methods** to highlight which genes the neural network used to make its prediction.
- Our findings **identify existing genes** responsible for color pattern variations previously discovered by GWAS.
- Our future work aims to improve modeling, use deep learning features as phenotypes, address missing heritability, and uncover epistasis behavior.

Results



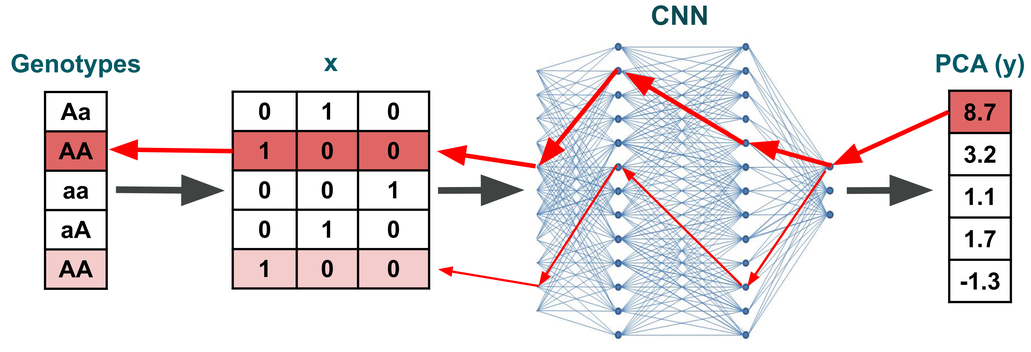
Methods (Data)

- Each butterfly specimen is **imaged and sequenced** into a format compatible for machine learning training.
- **Images** are sent through **Patternize** to obtain PCA values associated with a targeted phenotype (e.g. red color variation).
- The **DNA sequence** of each butterfly is processed to obtain genotypes and then **one-hot encoded** to allow for convolutional operations.



Methods (Modeling)

- A **convolutional neural network** based on **SoybeanNet** is trained. Early stopping with a held out validation set is incorporated to prevent overfitting.
- We then apply **Guided Grad-CAM** to the predicted output across a held out testing set and visualize the aggregated importance of each gene.



References

Liu, Y., Wang, D., He, F., Wang, J., Joshi, T., & Xu, D. (2019). Phenotype prediction and genome-wide association study using deep convolutional neural network of soybean. Frontiers in genetics, 10, 486384.

Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., & Batra, D. (2017). Grad-cam: Visual explanations from deep networks via gradient-based localization. In Proceedings of the IEEE international conference on computer vision (pp. 618-626).

Patricio A. Salazar, Nicola Nadeau, Gabriela Montejo-Kovacevich, & Chris Jiggins. (2020). Sheffield butterfly wing collection - Patricio Salazar, Nicola Nadeau, Ikiam broods batch 1 and 2. Zenodo. https://doi.org/10.5281/zenodo.4288311

This material is based upon work supported by the National Science Foundation under Award No. 2118240.