IOWA STATE UNIVERSITY
College of Liberal Arts and Sciences – Genetics Major

Honors Capstone Project
December 6th, 2020

Haley Dostalik, Advised by Dr. Carolyn Lawrence-Dill

Functional Annotation of the Grape – All Genes in the Genome

Abstract:

The purpose of this project is to update the functional annotation for the grape genome. Previous work that created a functional annotation of the grape genome has been shown to be not very accurate as time has passed. The methods used to create the annotation are not reproducible, which takes away from the credibility of the annotations. As technology has improved, there is more accurate data available for functional annotations, and it is important to update old annotations to reflect new discoveries. Over the past couple of years, Dr. Carolyn Lawrence-Dill's lab has been working on creating updated functional genome annotations of many different species of plants using a software they developed called Gene Ontology Meta Annotator for Plants, abbreviated GOMAP (13). Using this software involves finding and manipulating a protein fasta file to the correct format, creating slurm files to run the software, and reviewing the GO annotations output by the software to check for accuracy. Gene Ontology (GO) is a collection of biological terms, each with a unique code, that are used when annotating genomes to help keep language consistent across research (14). It also contains evidence codes, that help scientists distinguish how the annotation was obtained (e.g. experimental, computer-based).This project has resulted in a new data set containing functional annotations for the grape with updated Gene Ontology terms, and discusses the accuracy of the generated annotations compared to other research.

Objectives:

- Create an updated functional annotation of the grape genome that is reproducible using GOMAP software
- Check for accuracy of new annotation by comparing the GO annotations to known pathways in the grape

Methods:

Locate and modify data set:

- Updated grape protein fasta file found on Genoscope (6)
- Ran python code to modify fasta file to be one transcript per gene and to remove extra characters

Prepare slurm and text files:

- GOMAP runs on a high-powered computer (HPC), and each step requires their own text file and slurm file to run (11)

Run GOMAP pipeline steps (Figure One)

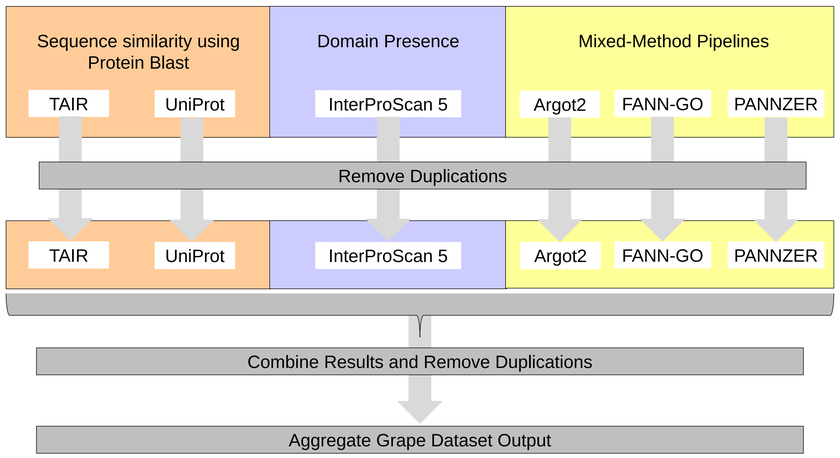
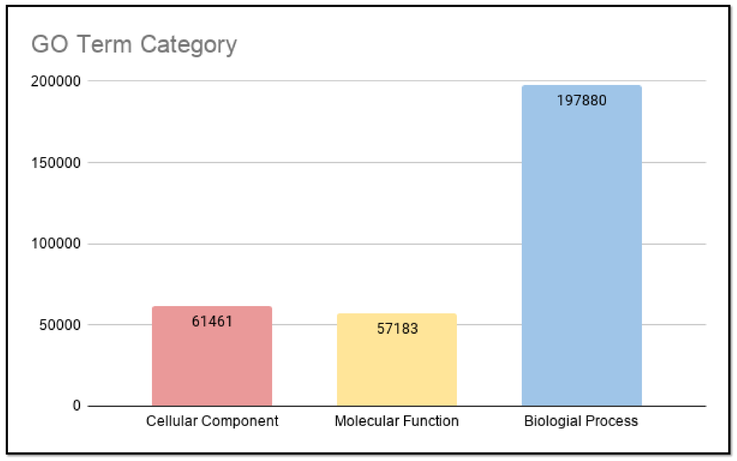


Figure One:

GOMAP pipeline steps (11,12). Each colored box represents a separate type of annotation assignment. Sequence similarity (orange) checks for sequence similarity between the grape input file and the TAIR dataset (2) and the UniProt dataset (10) using the software protein BLAST(1). Annotations for sequences that match from TAIR and UniProt are assigned to the grape dataset. InterProScan5(8) assigns annotations based on domain presence (blue). The three mixed-method pipelines (yellow) (Argot2(4), FANN-GO(5), PANNZER(9)) use a combination of sequence similarity and domain presence to assign annotations. Any duplicate annotations created by each software are removed before combining all the results together, then any duplications after combining results are removed. A final grape dataset full of annotations is then created.

Figure Two

GOMAP Term Predictions for Grape. Bars indicate GOMAP term counts for the whole grape genome across the three GO term types. Cellular component terms (red) relate to cellular anatomy (e.g. Mitochondria). Molecular function terms (yellow) relate to molecular activities within cells, such as the activity of a cellular component (e.g. ATP synthase activity). Biological processes (blue) are processes that require multiple molecular functions to happen (e.g. DNA repair).



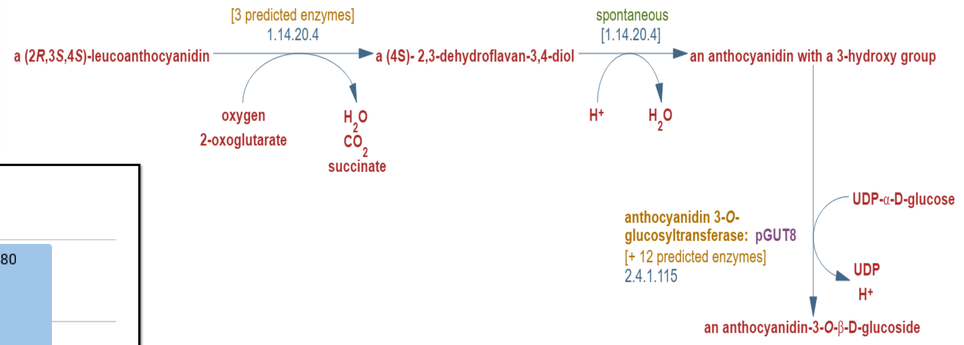


Figure Three:

Anthocyanin biosynthesis pathway in grape (3). This figure shown demonstrates the compounds and mechanisms involved in anthocyanin biosynthesis in grapes.

The scientists behind this research currently have determined anthocyanidin 3-*O*-glucosyltransferase is involved with the pathway, but there are still 15 predicted enzymes. These enzymes are predicted by computational methods, but have not been experimentally determined yet. Currently, I am working on comparing the genes annotated by GOMAP related to anthocyanin biosynthesis to these predicted enzymes to either help support or disprove if they are part of the anthocyanin pathway.

Results:

- 316,524 GO terms assigned (breakout of categories in Figure Two above)
- Distribution of a higher number of annotations assigned in the biological process GO term category matches trend of previous GOMAP annotations on other plant species (12)

Anthocyanin is a compound responsible for a red-purple pigment in many kinds of fruit, including the grape. It makes sense there would be many genes related to the biosynthesis of anthocyanin in the grape, because many grapes are red-purple in color (7). After the grape gene transcripts were put through GOMAP, 104 transcripts had the GO term correlating to anthocyanin-containing compound biosynthetic process assigned to them (GO:0009718).

Conclusion:

- Generated a new data set containing 316,524 GO terms assigned to gene transcripts from the grape genome
- Published data set will contain intermediate files generated and steps on how they were created – allows others to reproduce increasing credibility
- GO term for anthocyanin biosynthesis assigned to 104 different transcripts – expected due to large role in how grapes are colored

Analysis work on this project is still ongoing, and current analysis work on this involves comparing other researcher's predictions for enzyme function (Figure 3) in the anthocyanin biosynthesis pathway to the functions GOMAP assigned to those genes. Once this is done, it can provide extra support to the credibility of the GOMAP-generated functional annotations of the grape genome.

References:

1. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Basic local alignment search tool. Journal of Molecular Biology, 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
2. Berardini, T. Z., Reiser, L., Li, D., Mezheritsky, Y., Muller, R., Strait, E., & Huala, E. (2015). The Arabidopsis information resource: Making and mining the "gold standard" annotated reference plant genome. Genesis, 53, 474–485. https://doi.org/10.1002/dvg.22877
3. Carnegie Institute of Science. (2020). Vitis vinifera Pathway: anthocyanin biosynthesis. Retrieved from https://pmn.plantcyc.org/GRAPE/NEW-IMAGE?type=PATHWAY&object=PWY-5125
4. Clark, W. T., & Radivojac, P. (2011). Analysis of protein function and its prediction from amino acid sequence. Proteins, 79, 2086–2096. https://doi.org/10.1002/prot.23029
5. Falda, M., Toppo, S., Pescarolo, A., Lavezzo, E., Di Camillo, B., Facchinetti, A., . . . Fontana, P. (2012). Argot2: a large scale function prediction tool relying on semantic similarity of weighted Gene Ontology terms. BMC Bioinformatics, 13(Suppl 4), S14. https://doi.org/10.1186/1471-2105-13-S4-S14
6. Genoscope, INRA, Institute of Applied Genomics, et al., 2010, assembly, Retrieved from http://www.genoscope.cns.fr/externe/Download/Projets/Projet_ML/data/12X/.
7. He F, Mu L, Yan GL, Liang NN, Pan QH, Wang J, Reeves MJ, Duan CQ. Biosynthesis of anthocyanins and their regulation in colored grapes. Molecules. 2010 Dec 9;15(12):9057-91. doi: 10.3390/molecules15129057. PMID: 21150825; PMCID: PMC6259108.
8. Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., . . .Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. Bioinformatics, 30, 1236–1240. https://doi.org/10.1093/bioinformatics/btu031
9. Koskinen, P., Teoreonen, P., Nokso-Koivisto, J., & Holm, L. (2015). PANNZER: High-throughput functional annotation of uncharacterized proteins in an error-prone environment. Bioinformatics, 31, 1544–1552. https://doi.org/10.1093/bioinformatics/btu851
10. UniProt Consortium (2015). UniProt: A hub for protein information. Nucleic Acids Research, 43, D204–D212.
11. Wimalanathan, K. 2018, Welcome to GOMAP-singularity's documentation!, Retrieved from https://gomap-singularity.readthedocs.io/en/latest/index.html
12. Wimalanathan, K., Friedberg, I., Andorf, C., Lawrence-Dill, C.J., 2018, Maize GO Annotation-Methods, Evaluation, and Review (maize-GAMER), Wiley Online Library, https://onlinelibrary.wiley.com/doi/full/10.1002/pld3.52 (February 16, 2020)
13. Wimalanathan, K., Lawrence-Dill, C.J., Gene Ontology MetaAnnotator for Plants, bioRxiv The Preprint Server for Biology, https://www.biorxiv.org/content/10.1101/809988v1
14. 2020, The Gene Ontology Resource. Retrieved from http://geneontology.org/

Acknowledgements

Special thanks to Leila Fattel, Colleen Yanarella, and the rest of the Dill-Picl lab for all their support throughout this project!