Predicting Functional Relationships in Osteoblasts

Jacob M. Luber[1,2], Catherine Sharp[2], KB Choi[2], Cheryl Ackert-Bicknell[2,3], Matthew A. Hibbs[1,2]
[1]Department of Computer Science, Trinity University, San Antonio, Texas 78212 USA
[2]The Jackson Laboratory, Bar Harbor, Maine 04609 USA
[3]University of Rochester Medical Center, Rochester, New York 14642 USA

The Jackson Laboratory
Leading the search for tomorrow's cures

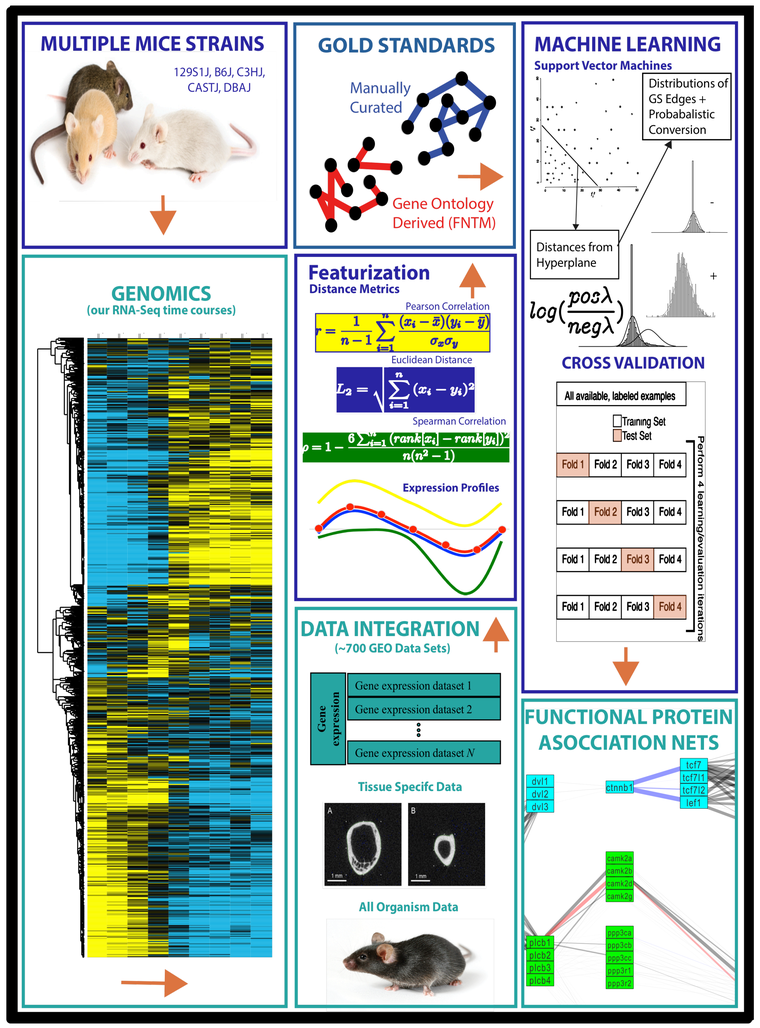
TRINITY UNIVERSITY
FOUNDED 1869

ABSTRACT

Predicting protein function by integrating diverse, high-throughput experimental data has proven useful for a range of applications. However, applying these methods to complex mammalian systems is complicated by the dramatically different roles individual proteins may play in different cellular contexts. Established methods for incorporating tissue context specificity automatically generate training sets from Gene Ontology (GO) annotations, combined with databases of tissue-specific gene expression, to make context-specific predictions without the need for excessive manual curation (e.g., Greene et. al.).

These methods perform tissue-specific heterogeneous data integration by using a training set containing positive edges only between genes that are both expressed in a given tissue and share GO annotations (Goya et. al.). The tissue specific versus non-tissue specific resulting networks underscore the importance of context-specificity in functional data integration. But the biological complexity of mammalian systems, especially given their large numbers of tissue and cell types, mandates that automatic training set generation is key to making large scale context-specific predictions tractable. However, automated context-specific training set generation is limited. Specifically, co-annotation to a GO term implies that two genes are functionally related in some context, but co-expression of those two genes in a specific context does not necessarily imply that the functional relationship occurs within that context. We show that thorough manual training set curation combined with context-specific input dataset selection potentially produces more biologically meaningful functional relationship predictions than automated approaches in the specific context of mouse osteoblast function and bone maintenance.

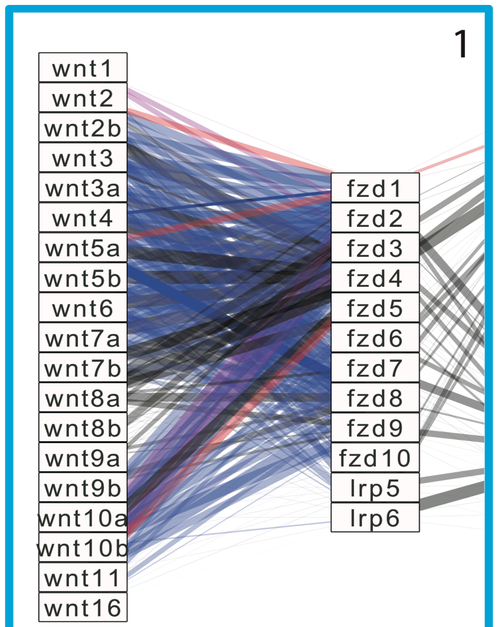
OVERVIEW OF ALGORITHMIC PIPELINE



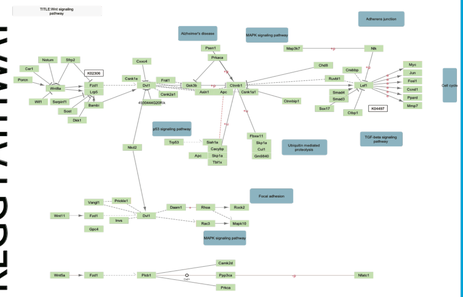
REFERENCES

Ackert-Bicknell et. al., Aging Research Using Mouse, *Current protocols in mouse biology*, 5(2):95-133,2015.
Cain et. al., Increased Gs Signaling in Osteoblasts Reduces Bone Marrow and Whole Body Adiposity in Male Mice, *Endocrinology*, 1867-1873, 2016.
Dowell et. al., Novel insights into embryonic stem cell self-renewal revealed through comparative human and mouse systems biology networks, *Stem cells*, 32(5):1161-1172, 2014.
Greene et. al., Understanding multicellular function and disease with human tissue-specific networks, *Nature Genetics*, 47(6):569-576, 2015.
Goya et. al., FNTM: a server for predicting functional networks of tissues in mouse, *Nucleic Acids Research*, 43(W1):W182-7,2015.

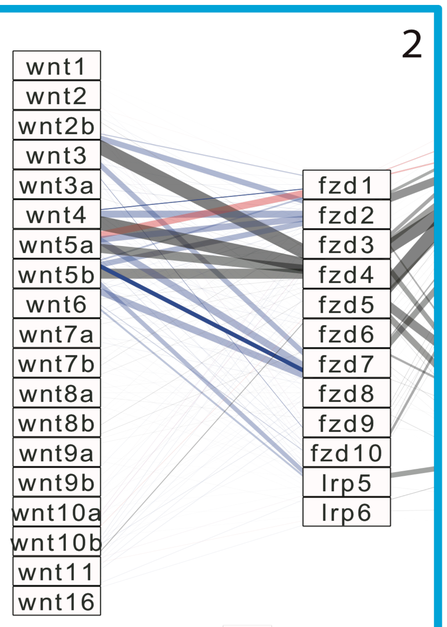
ANALYSIS OF WNT SIGNALLING PATHWAY

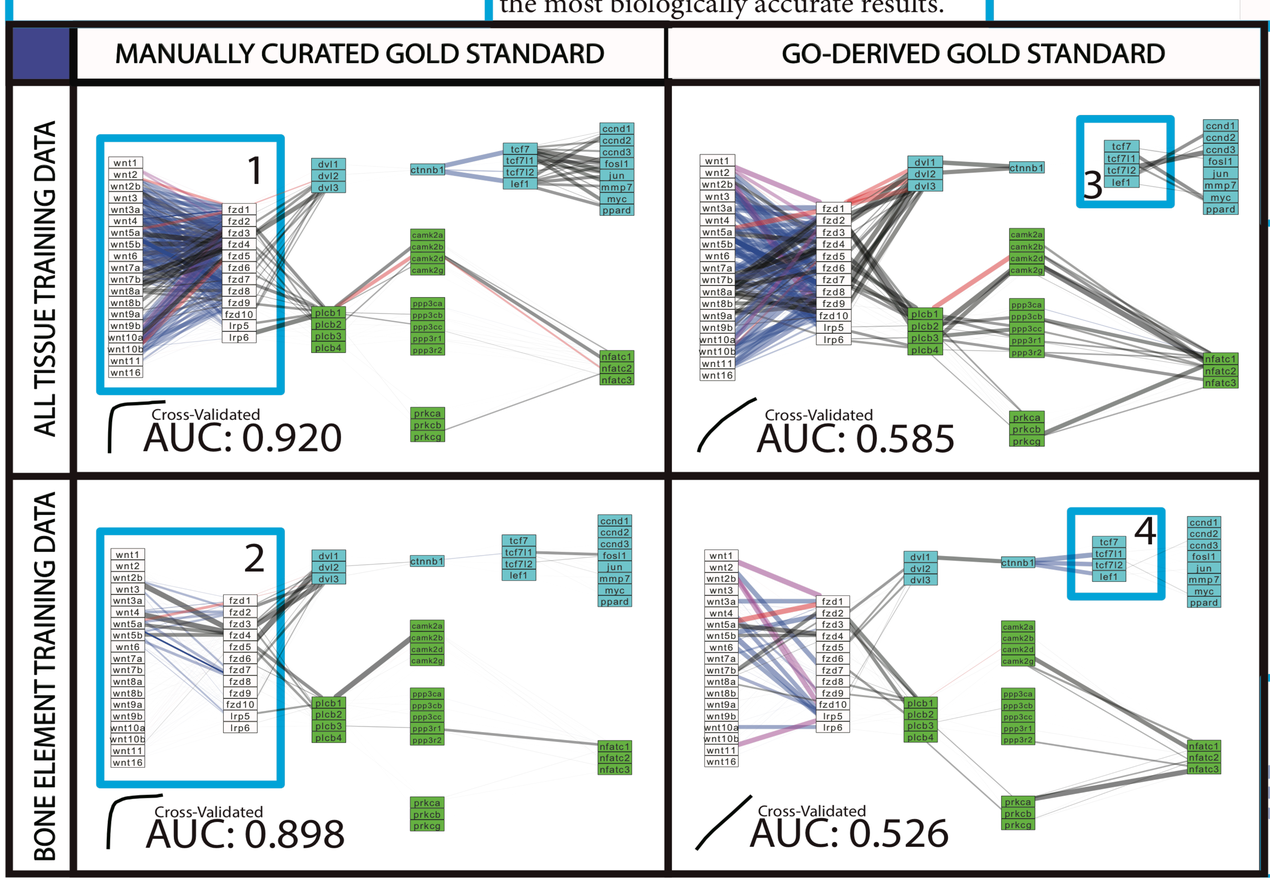


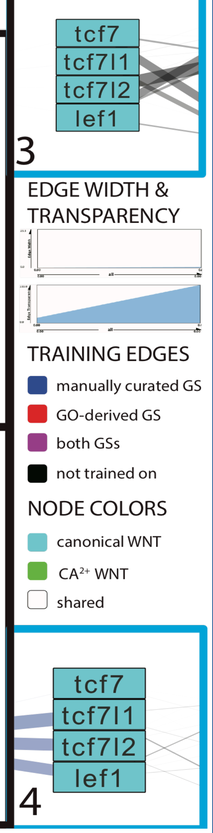
KEGG PATHWAY



Our analysis focused on both the canonical WNT Signaling Pathway as well as the Ca^{2+} non-canonical pathway. Comparing both Machine Learning performance and the subset of these pathways predicted accross models suggests that training SVMs with a manually curated gold standard on bone element specific datasets yields the most biologically accurate results.







DISCUSSION

Our results suggest that tissue specificity of functional relationships is a significant issue and must be accounted for in related machine learning approaches that aim to predict function. While we have proved this issue exists only for one tissue it is likely that such a pattern will hold over a broader range of contexts. Our training set and input feature curation methods, featurization methods, iterative dataset pruning, and likelihood calculations all contribute to properly handling context specificity issues. This is currently problematic, as our methods do not generalize easily to any tissue as there are large time demands for high quality manual curation. This suggests that there is room for further development of computational and statistical methods that could mimic the performance of our method without the manual labor required for dataset and training set curation.

KEY TAKEAWAYS

- Predictions made by the four classifiers are very dissimilar
- Likely that some of the highly predicted edges may not actually be related within the context of bone biology for all tissue classifiers
- Literature evidence linking classifier trained on manually curated data and applied to only bone element data provides most accurate picture of bone biology (Cain et. al.)
- Curated gold standard contains edges not supported by bone only data---we have preliminarily found that only a subset of FRs in the literature are actually supported by co-expression