

Collaborative Network

Our collaborative network includes several groups responding to this RFA. We shared proposals openly on GitHub and discussed their content via Slack. Selected interactions between groups are summarized in Fig 2 and below. We look forward to these and additional collaborations within this group and the broader HCA consortium.

Model interpretation. Assessing the performance of unsupervised techniques relies critically on interpretation relative to known covariates, gene function, and independent test datasets. Our collaborative network contains several investigators who are proposing new algorithms for unsupervised model learning. By developing ProjectoR as an efficient comparator, we will be able to interpret the function of our inferred transcriptional trajectories relative to gene signatures. Proposed model contributors include **Casey Greene** and Elena Fertig.

Visualization is also critical to such interpretation. Further collaboration with **Lana Garmire's** group on the development of Granatum will enable automated visualization of unsupervised patterns. We look forward to continued interactions with the consortia to optimize unsupervised model interpretation.

Efficient factorization methods. Several proposals include techniques for efficient factorization methods. We have an existing collaboration with **Elena Fertig** to develop and implement a parallelized version of GW-CoGAPs to identify patterns from single cell data. Much of our preliminary data arises from this fruitful and productive collaboration.

Furthermore, **Rob Patro** proposes these techniques to infer relevant features to quantify transcript abundance. We look forward to collaborating with the consortia to continue to develop state of the art factorization methods and establish a common framework for model validation and comparison.

Deep Learning. **Casey Greene's** deep learning techniques propose training on arbitrary patches are selected and arbitrary rotations of the data. This ensemble approach maps to the parallelization across subsets of samples in our proposal. We plan to work collaborate to evaluate methods for optimal feature selection across transcripts and samples for efficient and robust pattern inference. Dr. Greene's

Benchmark data. Assessing performance of the proposed algorithm relies on presence of matched, time-course data across bulk and single-cell RNA-sequencing measurement technologies. Benchmark data in the human cell atlas will be critical to optimal algorithm development. We will also benefit from datasets such as the retinal development data in bulk, smart-seq, and 10X platforms provided by **Loyal Goff**. Algorithm development will further benefit from datasets with parallel perturbations in multiple datasets from **Arjun Raj**.

Integration with imaging/spatial single cell data. In collaboration with **Arjun Raj**, we are working to interface their image analysis pipeline for single molecule FISH with a robust statistical test to permit multifactor analysis of these quantifications.