## Fig. 1



$$D_{ata} = A_{mplitudes} \times P_{atterns}$$

A (gene weights for each biological process)

P (biological process in each sample)

genes

$$= \begin{array}{cc} 0 & 1 \\ 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{array} \times$$

genes

samples

subjects

The algorithm looks for patterns that define the biological activity in a set of samples (e.g., a collection of tumors) and the amplitude at which each gene contributes to that biological process.

## Fig. 3



HCA Training Data (Goff / Raj)

Feature quantification (Patro)

Image features (Raj)

Reproducible Workflows ( Garmire / Greene / Patro / Piccolo )

Efficient factorization

Sensitivity to input features and variance

Denoising / Filtering (Krisnaswamy / Piccolo)

Deep learning

Sub/re-sampling evaluation

Trajectory / temporal patterns (Fertig)

Gene signatures and gene regulation (Greene / Krisnaswamy)

Model interpretation (Goff)

Visualization (Garmire / Krishnaswamy)
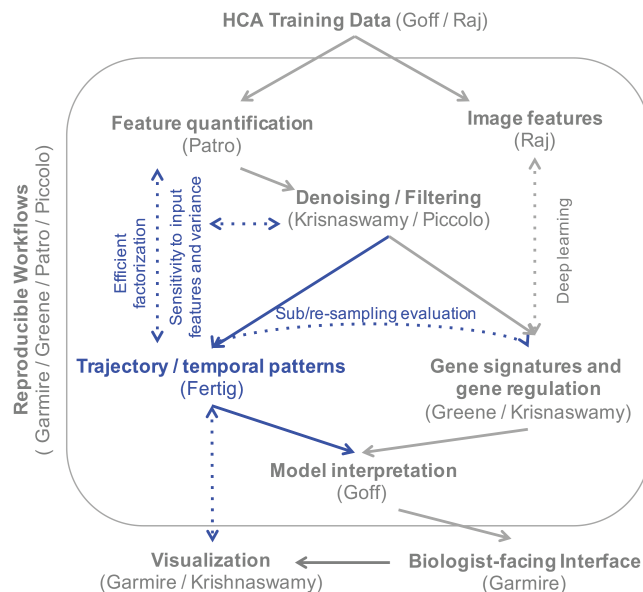
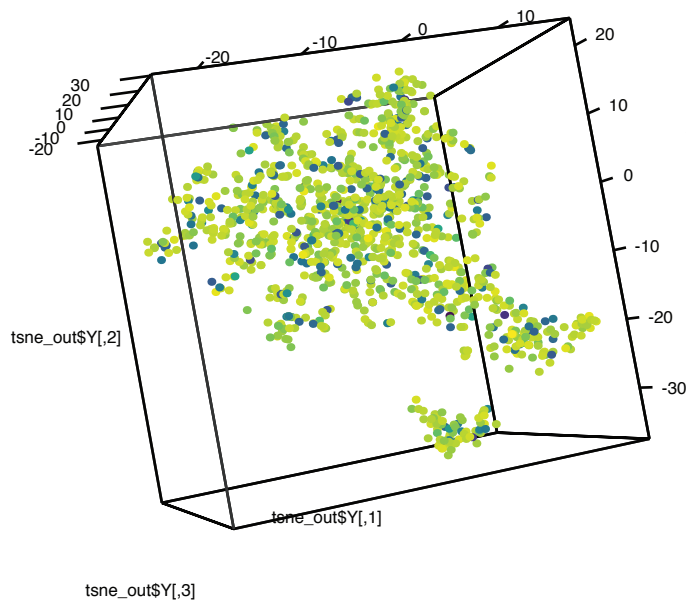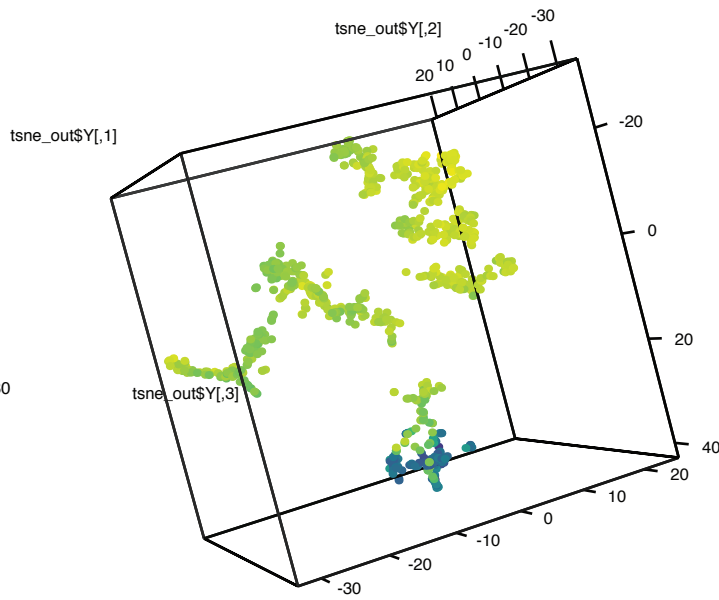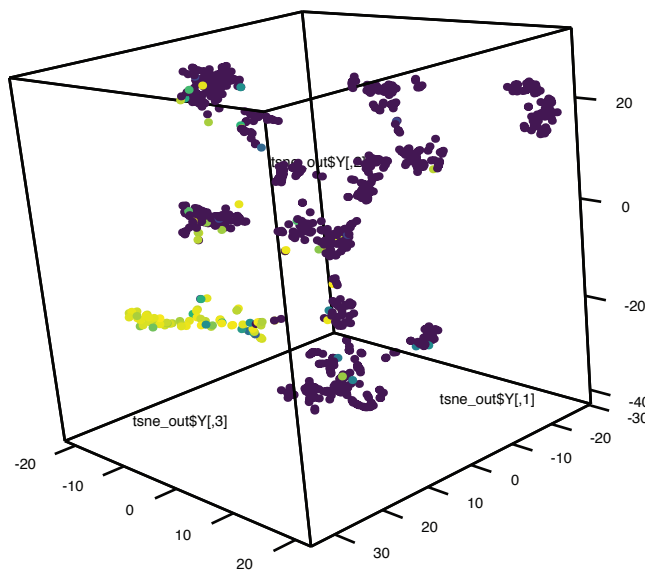Biologist-facing Interface (Garmire)

## Fig. 2

### A. T-SNE of raw scRNAseq data colored by # genes expressed



### B. T-SNE of CoGAPS patterns for scRNAseq colored by # genes expressed



### C. T-SNE of CoGAPS patterns colored by Crx expression (Crx is a marker of photoreceptors)



### D. T-SNE of CoGAPS patterns colored by Ccnd1 expression (Ccnd1 is a marker of cell cycle)