Principal Component Analysis – PCA

We applied principal component analysis (PCA) to the expression matrix, and produced visualizations in which each sample is represented by a point in the plane formed by two principal axes, and colors were assigned to each point according to the biological class.

We found that the first three principal components have biological interpretations; we named them the hematopoietic, malignancy and neurological axes.

Each dot represents one of the 5,372 samples in a multidimensional gene expression space projected on the principal plane formed by the first (hematopoietic) and second (malignancy) principal axes. The dots are colored semi-transparently according to the biological group the sample belongs to.
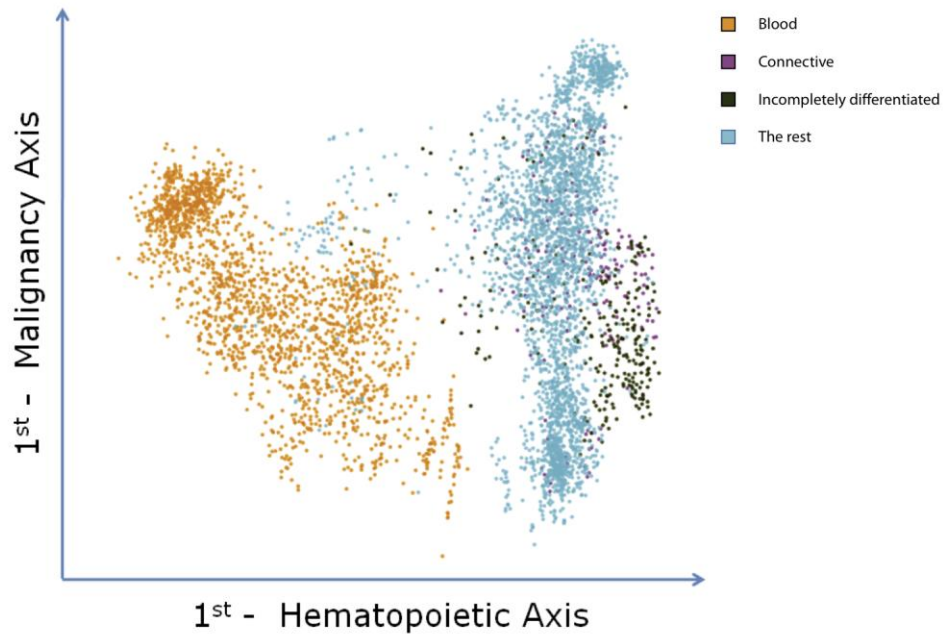
The first principal component separates hematopoietic system-derived samples from the rest of the samples, with connective tissues and incompletely differentiated cell–based samples forming a relatively compact group on the right. The cyan dots among the blood samples on the right side represent samples from bronchoalveolar lavage cells (a possible sample contamination with blood) and kidney. The dark green dots at the center include embryonic stem cells.
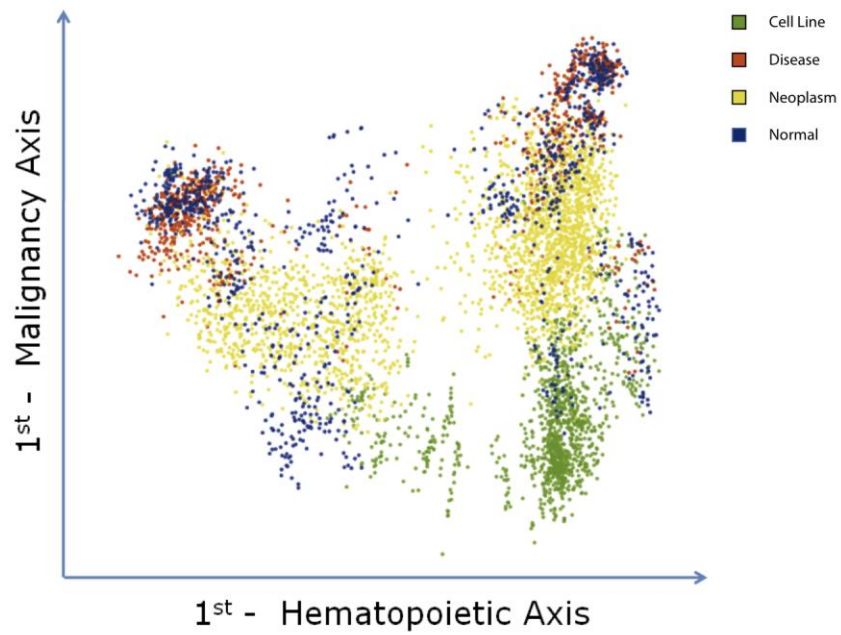
The second principal axis predominantly arranges cell line samples at the bottom, neoplasm samples in the middle and a mixture of non neoplastic disease and normal samples at the top.
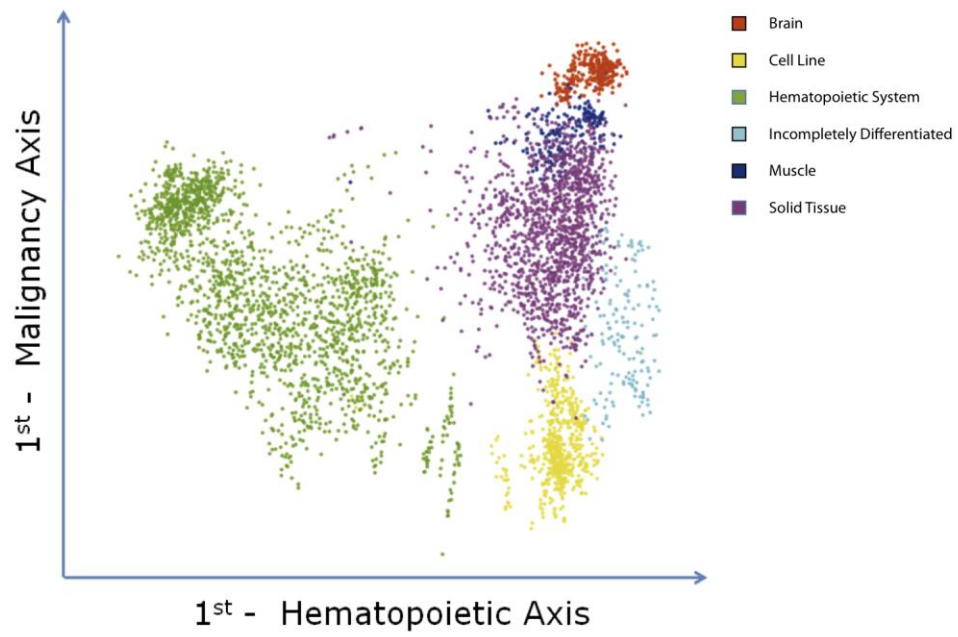
Samples colored by 6 clusters (see legend) identified on self-vs-self heat-map of 96 biological groups. The samples are visualized on the 1st and 2nd principal component plane.
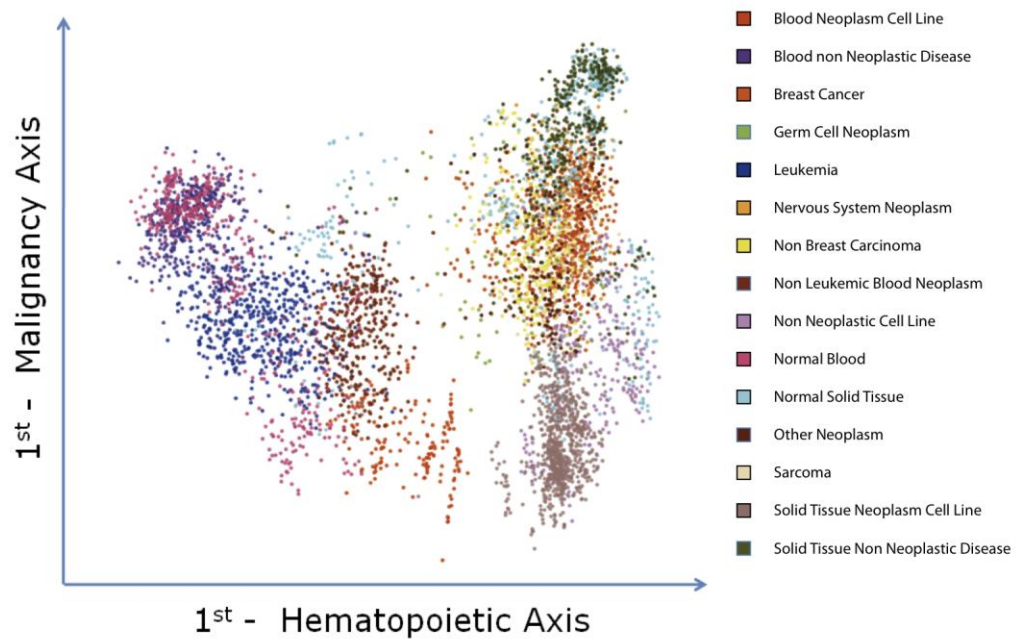
Samples colored by 15 meta-groups (see legend) on the 1st and 2nd principal component plane.
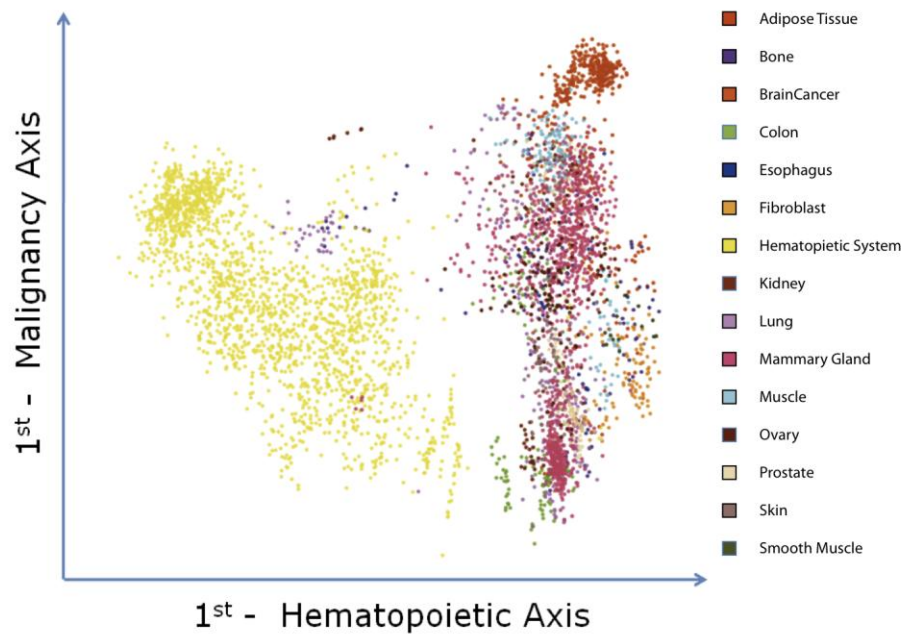
Samples colored by 15 groups of tissues of origin (see legend) on the 1st and 2nd principal component plane.