Deconvolution of isoforms from microarray data using non-negative matrix factorization

Charles Vaske and Josh Stuart

cvaske@soe.ucsc.edu jstuart@soe.ucsc.edu

UC SANTA CRUZ

Department of Biomolecular Engineering

Overview

Detecting the particular isoforms expressed in a cell is challenging. Oligomeric microarray platforms with junction probes provide genome wide assays of alternative splicing. Most approaches predict alt-splicing using data within one tissue and/or for local gene structural features. We present an approach that uses a matrix decomposition technique that learns which isoforms are expressed using the entire set of probes and conditions measured in a gene expression compendium. The method is therefore able to correlate information across the conditions and the probes to find a more reliable measure of alternative splicing.

Non-negative matrix factorization (NMF) is used in image processing to deconvolute linearly mixed signals. We describe the first application of NMF to learn both probe-isoform overlap and isoform expression from splice-junction microarrays. One advantage of the approach is that it can be used without any prior knowledge of a gene model.

Methods

NMF decoposes a matrix V into a product of H and W:

$$V = WH$$

In the splice array setting, V represents the probe-by-tissue intensity data measured on the microarray platform, H is the expression levels of the transcripts in each tissue, and W is the set of probes in each predicted transcript.

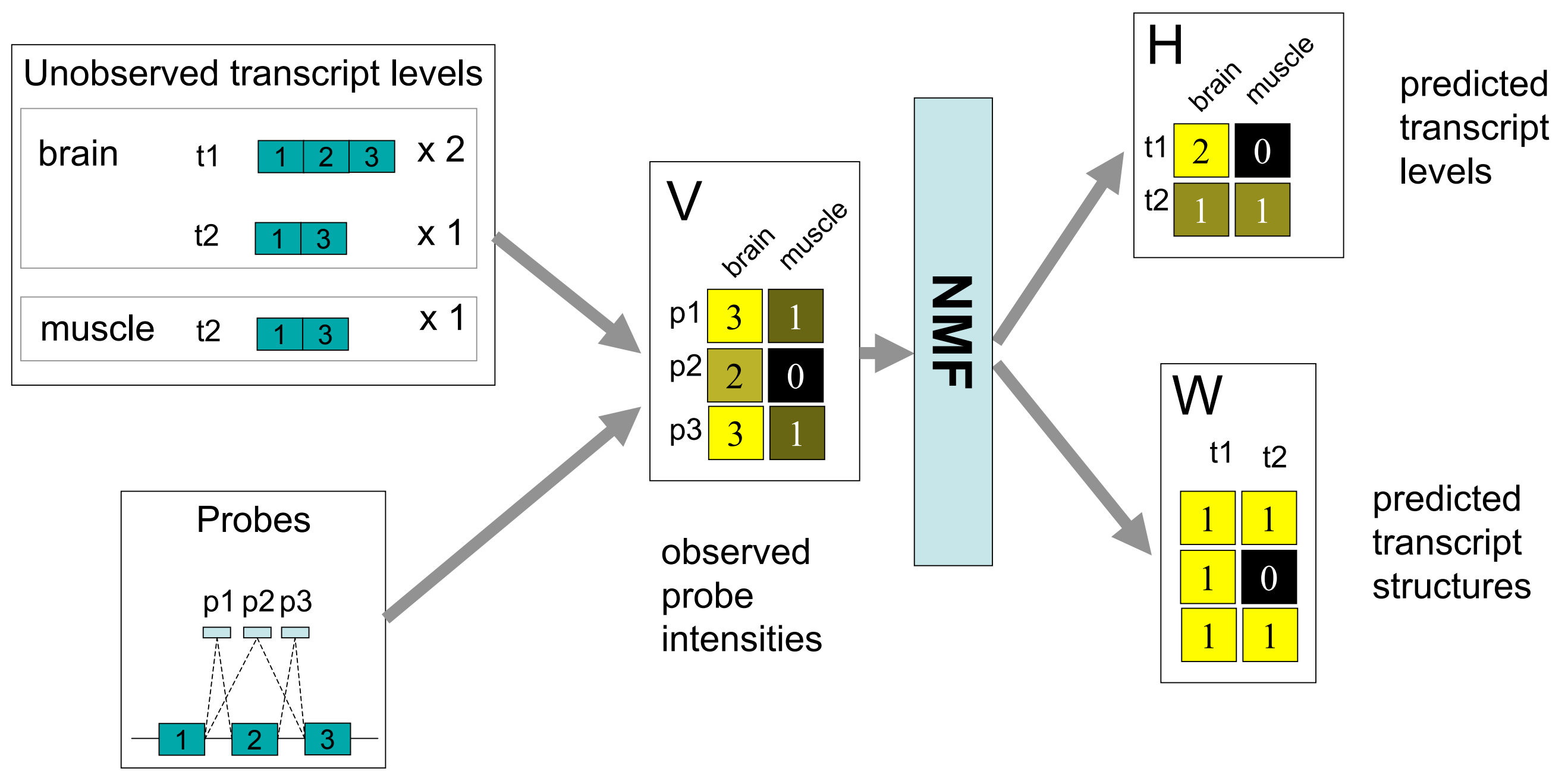


Figure 1. Illustration of how non-negative decomposition predicts both the transcripts and their expression levels. A hypothetical gene containing three exons is shown in the upper left along with three probes. The upper left panel illustrates the true expression levels of the "hidden". The hypothetical gene has two transcripts, t1 and t2.

Results - Synthetic Data

We estimated the accuracy of NMF by creating a synthetic a dataset in which both the isoforms and their expression profiles were known. We created synthetic data by generating random gene models with 1 to 5 isoforms. Noise was added to the observed probe intensities before running NMF. We measured the precision and recall as a function of experimental noise (Fig. 2). We also measured the ability of NMF to detect isoforms as a function of *k* (Fig. 2B). We find that NMF can reliably identify known transcripts at various strengths that were tested. When NMF is only allowed to predict a single transcript, it finds the transcript with the highest nearly every time. Recall remains high as the number of possible predictions is increased. For example, it can find all four transcripts that range that have a four-fold difference in abundance 40% of the time.

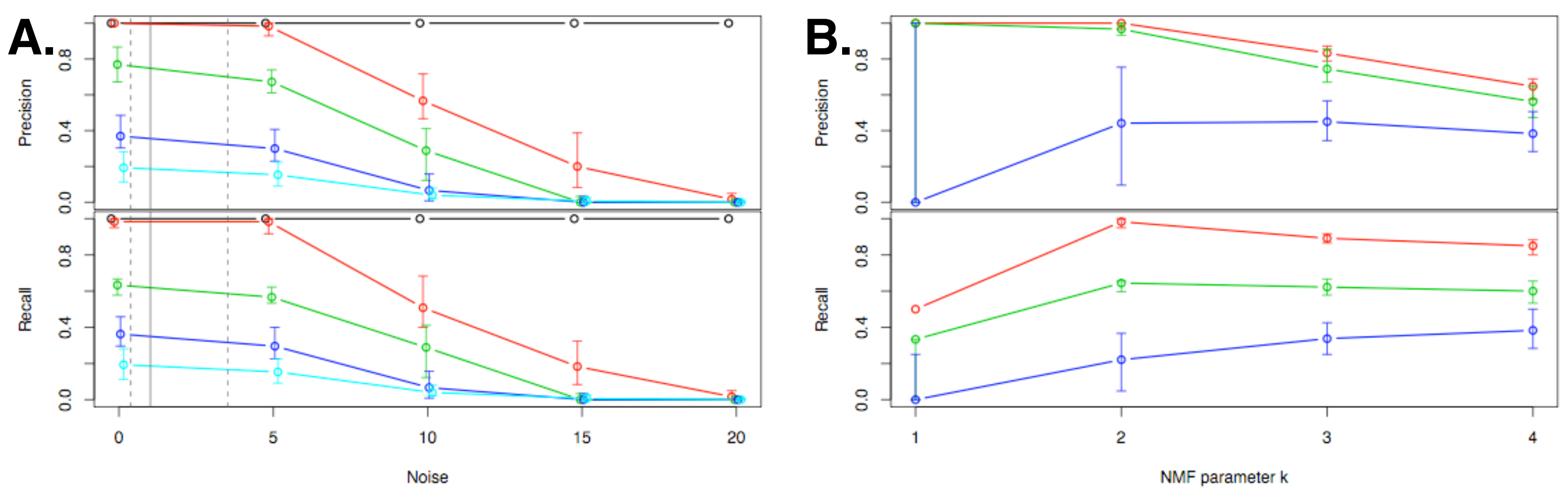


Figure 2. A. Accuracy of NMF in the presence of noise. Gray vertical lines indicate estimated 50% confidence interval on noise in the human microarray data. Black, red, green, blue, and cyan lines correspond to genes with 1, 2, 3, 4, and 5 transcripts respectively. B. NMF isoform recall when parameter k is varied from the true number of transcripts.

Results - Human Tissue Compendium

We applied NMF on a human tissue compendium of Johnson et al. (2003). As a positive control we ran NMF on ADD3's expression levels and found that NMF accurately recovers the known splicing pattern for this gene (Fig3 3).

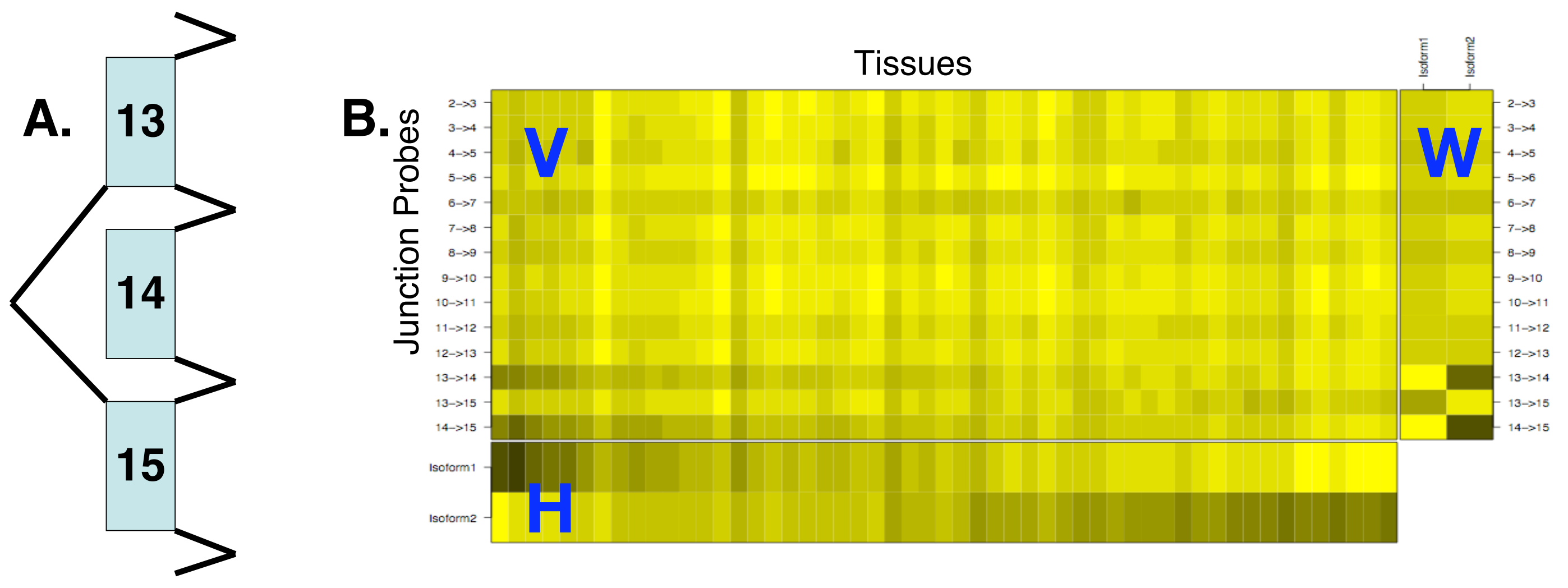


Figure 3. Decomposition of gene ADD3 into isoform-probe overlap matrix and isoform-tissue expression matrices.

We next applied NMF to the entire set of genes from the Johnson et al study. We predicted 15710 transcripts (Y transcripts per gene on average). To visualize the entire expression program and identify splicing programs among the genes, we used hierarchical clustering to cluster both the predicted transcripts and the tissues (Fig 4).

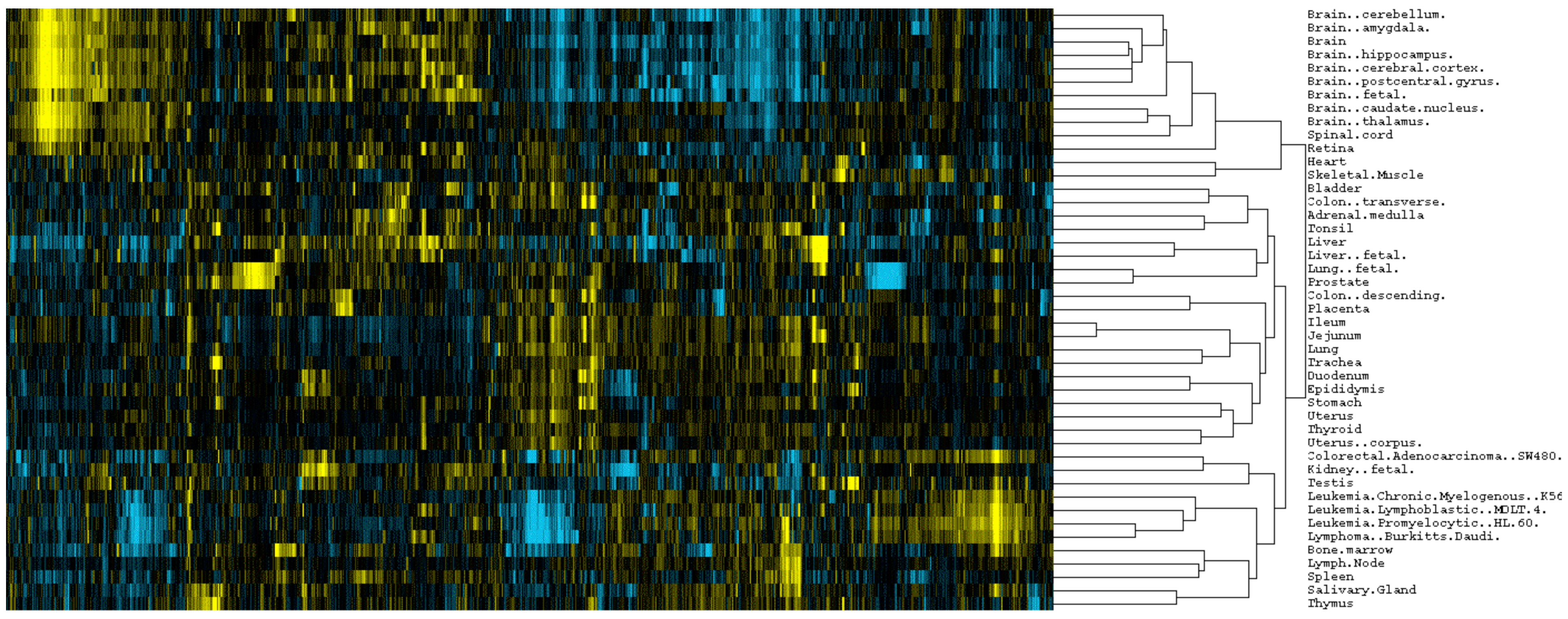


Figure 4. Clustering of predicted transcript expression levels. There are a large number of brain specific transcripts in the far left, and a cluster of cancer-related transcripts on the right.

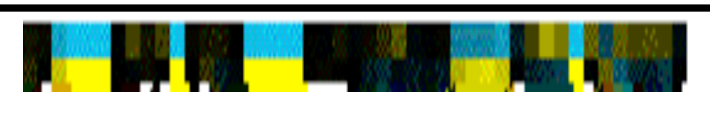
Name		Predicted transcripts	Correlation between transcripts	Supporting Evidence
POLK		2	-0.5883	2nd transcript recently pub.
SACM2L		2	-0.5886	Two transcripts in RefSeq
LRCH3		2	-0.5921	
UBE4A		2	-0.5909	
NR1H2		2	-0.5886	

Table 1. Genes with most distant predicted transcript expression profiles.

Future Directions

- Model probe hybridization affinity. Currently, genes with heterogenous probe hybridization affinities will degrade NMF performance.
- Filter predicted transcripts that have unlikely or improbable probe overlap vectors
- Use sequence around predicted splice variations to find splicing *cis*-elements within clusters

References

- Lee, D.D. and H.S. Seung, Learning the parts of objects by non-negative matrix factorization. Nature, 1999. 401(6755): p. 788-91.
- Wang, H., et al., Gene structure-based splice variant deconvolution using a microarray platform. Bioinformatics, 2003. 19 Suppl 1: p. 315-22.