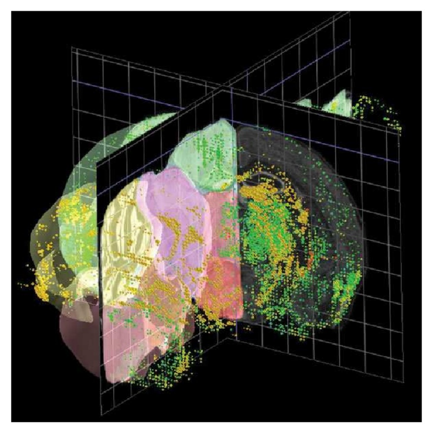
## Gene expression and cortical areas

Bayle Shanks and Charles Stevens

1. For each traditional cortical area, which genes mark this area?
2. Do the genes suggest non-traditional ways to carve up the cortex?

Adult mouse. Our tools will be open-sourced upon publication at `http://bshanks-stevens.nfshost.com/`

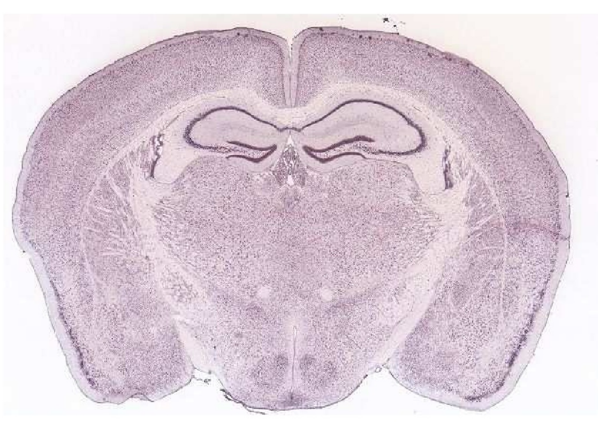
## Where the data came from:



The Allen (adult) Mouse Brain Atlas coronal dataset (about 4000 genes out of about 25000 total in mouse).

**in-situ hybridization**

Example: a slice of gene "0610007P14Rik"



Depth 6650 coronal slice of 0610007P14Rik from the ABA

**process images to detect gene expression**

Expression mask of depth 6650 coronal slice of 0610007P14Rik

**register (align) slices into 3D coordinate system**

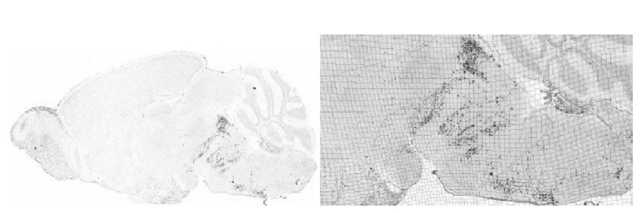
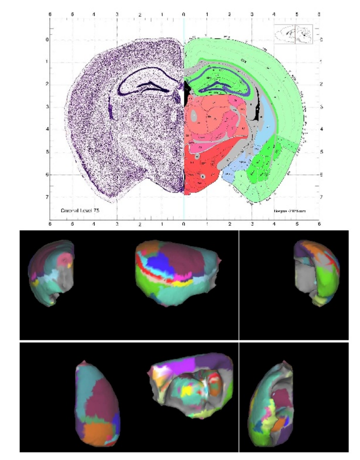


Figure 8 from `http://www.brain-map.org/pdf/InformaticsDataProcessing.pdf`
Voxels that are 200 microns on a side (why so large? registration error).

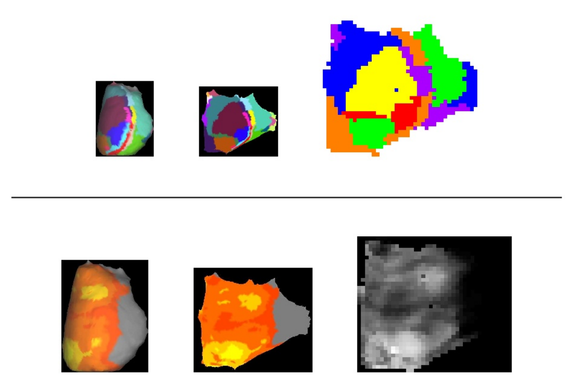
## The Allen Institute did all that. Now here's what we did:

**Select the cortex**



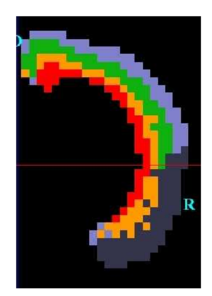
(and manually draw in areal boundaries)

**Map 3D expression data to 2D**



This part of the work was done with the assistance of the open-source program Caret. We also normalized.

**Segment layers**



Custom clustering algorithm for the purpose of automatically identifying layer boundaries; based on the intuition that a voxel in a layer should have a gene expression profile that is more similar to other voxels in the same layer than it is similiar to other voxels in different layers. The algorithm is a greedy iterative hillclimber that iteratively modifies the layer boundaries to make them better.

## Which genes are markers?

**4 measures of the marker-ness of a gene**

- **Correlation**: find genes which express more strongly in the region than outside of it (or vice versa)
- **Logistic regression**: like correlation, but appropriate since target image has only 2 discrete classes
- **Gradient similarity**: find genes with sharp boundaries similiar in shape shape to the target image's boundaries. We invented this, and we think this measure works the best.
- **Information gain**: find genes whose expression values can be used in some way to give information (possibly in some complicated, non-monotonic way) about whether or not you are in the target region.

**Details of the measures**

**Correlation** $\frac{\sum_{pixels}(x-\mu_x)(y-\mu_y)}{\sigma_x \sigma_y \# pixels}$

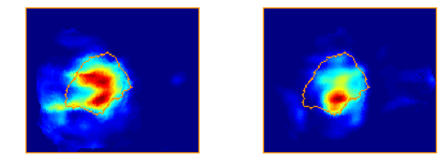
**Logisitic regression** fit $y = \frac{1}{1+e^{-(B_0+B_1 x)}}$, where x is the value of a pixel of gene expression, and y is the value of the corresponding target image pixel; genes such that this model assigns a high liklihood to the target image are good

**Gradient similarity** $\sum_{pixels} cos|(\angle\nabla_x - \angle\nabla_y)| \cdot \frac{|\nabla_x|+|\nabla_y|}{2} \cdot \frac{pixval_x + pixval_y}{2}$
where $\nabla_x$ and $\nabla_y$ are the gradient vectors of the two images at the current pixel; $pixval_i$ is the value of the current pixel in image $i$.

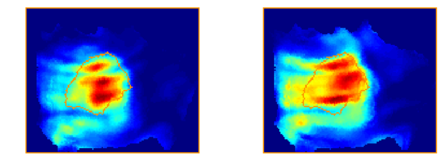
**Information gain** $H(Y) - H(Y|X)$
where X is the binary discretization of the normalized gene expression, with the discretization threshold chosen so as to maximize information gain, Y is the target image, H() is entropy.

**Correlation vs. logistic regression**

SS:
Correlation:



Logistic regression:



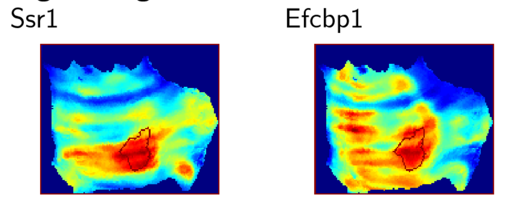
Top row: Genes *Nfic* and *A930001M12Rik* are the most correlated with area SS (somatosensory cortex). Bottom row: Genes *C130038G02Rik* and *Cacna1i* are those with the best fit using logistic regression.

**Logistic regression vs. Gradient similarity**

AUD:

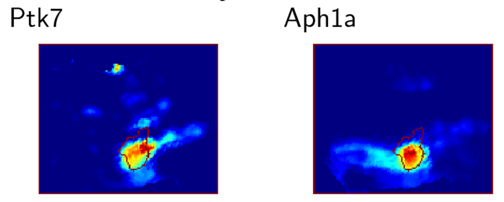
**Logistic regression**

Ssr1 Efcbp1



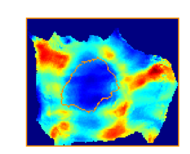
**Gradient similarity**

Ptk7 Aph1a



## Interesting phenomena

**Underexpression**

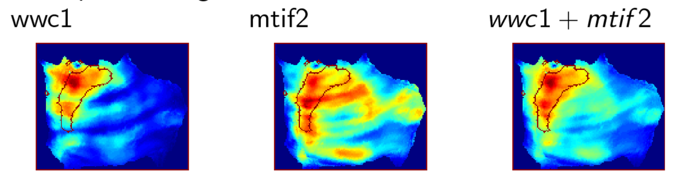


SS (Pitx2):

**Combinatorial coding**

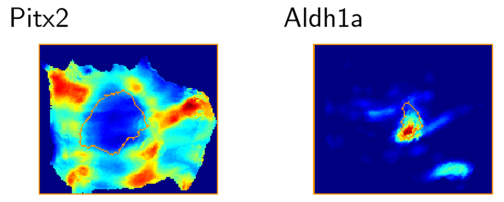
MO requires two genes:

wwc1 mtif2 *wwc1* + *mtif2*

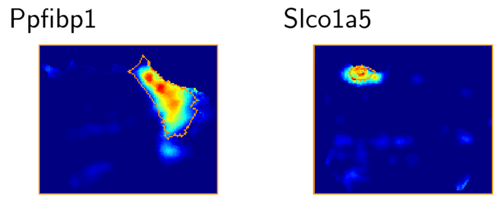


## Areas defined by single genes

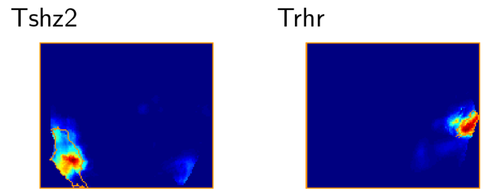
SS Pitx2 SSs Aldh1a



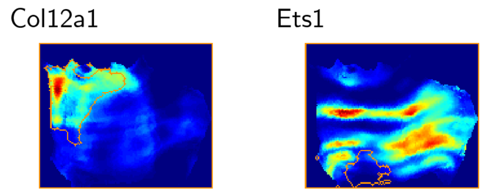
PIR Ppfibp1 FRP Slco1a5

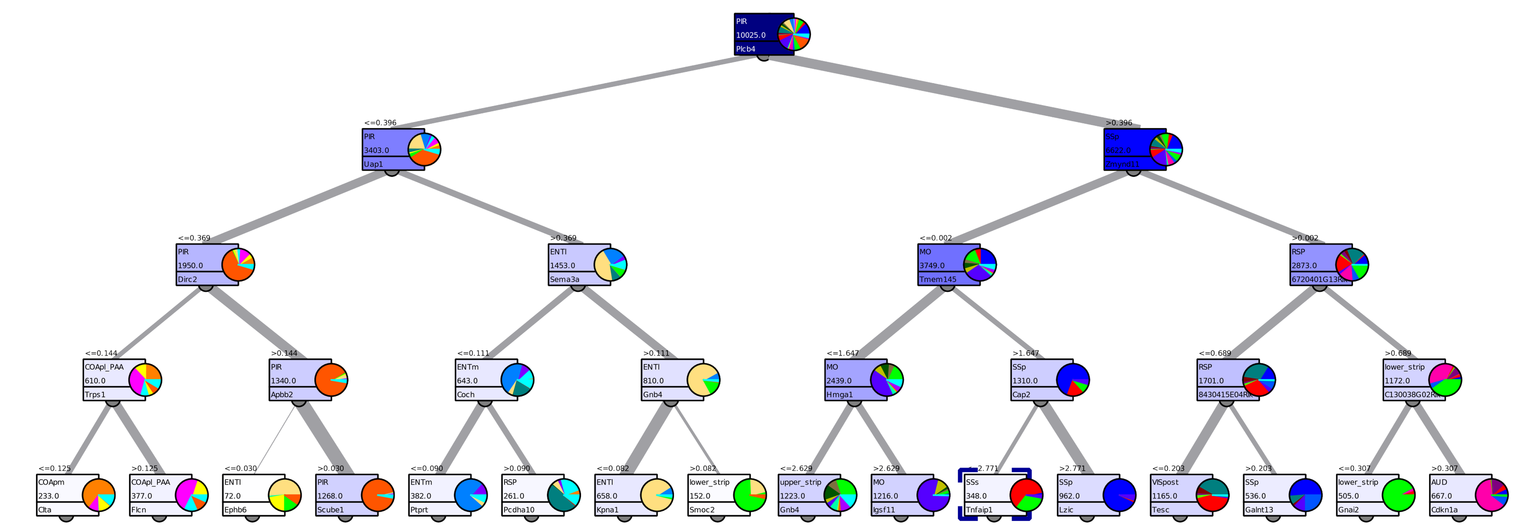


RSP Tshz2 COApm Trhr



ACA+PL+ILA+DP+ORB+MO Col12a1 posterior and lateral visual Ets1

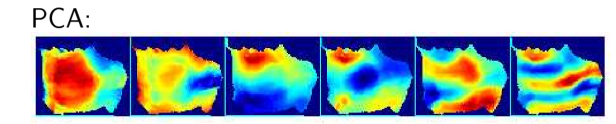




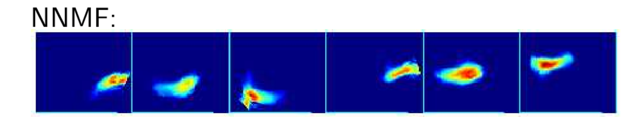

Best 9 fitting genes for area VIS; top: according to logistic regression; bottom: according to gradient similarity. I have 52 other figures like this, for a variety of other areas; ask me if you want to see them.

## Discovering how to carve up cortex

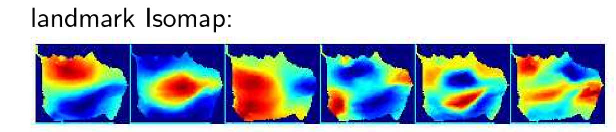
PCA:



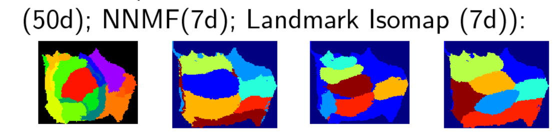
NNMF:

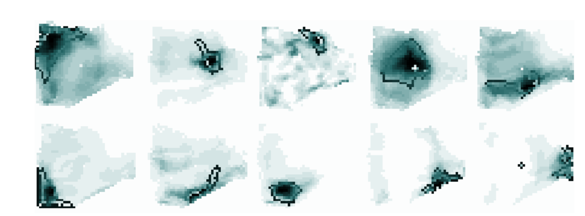


landmark Isomap:

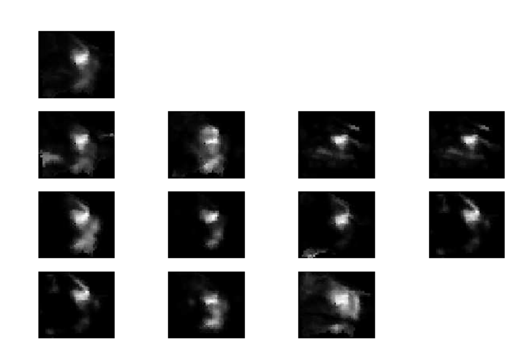


K-means (7-cluster traditional anatomy; PCA (50d); NNMF(7d); Landmark Isomap (7d)):





Prototypes corresponding to sample gene clusters, clustered by gradient similarity. Region boundaries for the region that most matches each prototype are overlaid. This suggests that clusters of genes may be meaningful.



A cluster of genes with an intriguing expression pattern. The top left is the average. I have ~15 other clusters if you'd like to see them.

## Caveats

- $N = 1$: probable overfitting
- Need to doublecheck for histological artifacts
- Slice artifacts
- Allen Reference Atlas parcellation
- Registration error in the ABA
- Specificity of the in situs
- Layer-finding algorithm

Therefore these results should not be taken as definitive, but rather should be taken as a starting point for experimental confirmation. The lists of genes produced by the algorithm should be regarded merely as a list of potential genetic markers.

## Conclusions

We have:

- identified genetic markers for a variety of cortical areas.
- created some useful datasets, including a machine-readable annotation of cortical areas onto the Allen Mouse Brain Atlas, and a flatmap Allen Mouse Brain Atlas cortex
- created an open-source toolbox for interaction with the Allen Brain Atlas, and for automated discovery of marker genes.

Please let us know if you'd like a copy of any of this.

## Future work

We are working on:

- testing our automated cortical layer segmentation algorithm.
- comparing the results of various algorithms for the purpose of automatically discovering new ways to carve up the cortex based on gene expression.

Thanks to the Allen Brain Institute for the data, the Van Essen lab for Caret, the Ljubljana lab for Orange, which we used to make the decision tree, and the Octave, Python, NumPy, and SciPy open source teams.
Thanks also to Chuck Stevens, Yongling Zhu, Mary Anne Pilla, Richard Jacobs, Sunhwa Lee, Ying Zhang, Ed Han, Vitaly Klyachko, Jian Xu, Maxim Bazhenov, Corinne Teeter, Doug Rubino, Nikoosh Carlo, Trygve Bakken, Ruadhan O'Flanagan, Will Barkis, Neville Sanjana, Dana Dahlstrom, Quentin Gaudry, Adam Stocker, Martyn Goulding, Tak Mori, Erik Flister, Kevin Briggman, Andrew Hires, and Ed Boyden, the rest of my committee and all of my other lab mates, and the Coates chair endowment.