The multi-facets of a data science project to answer:

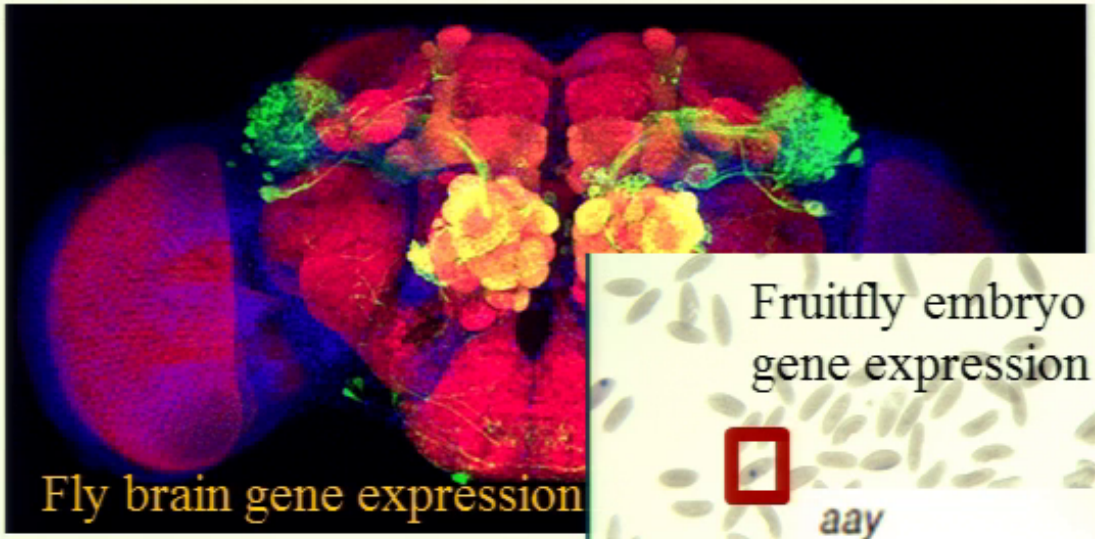
How are organs formed?

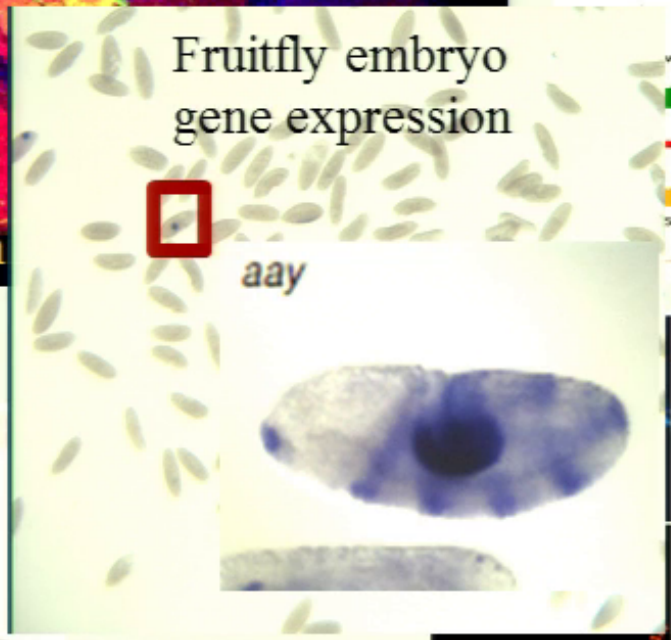
Bin Yu
Statistics and EECS, University of California-Berkeley

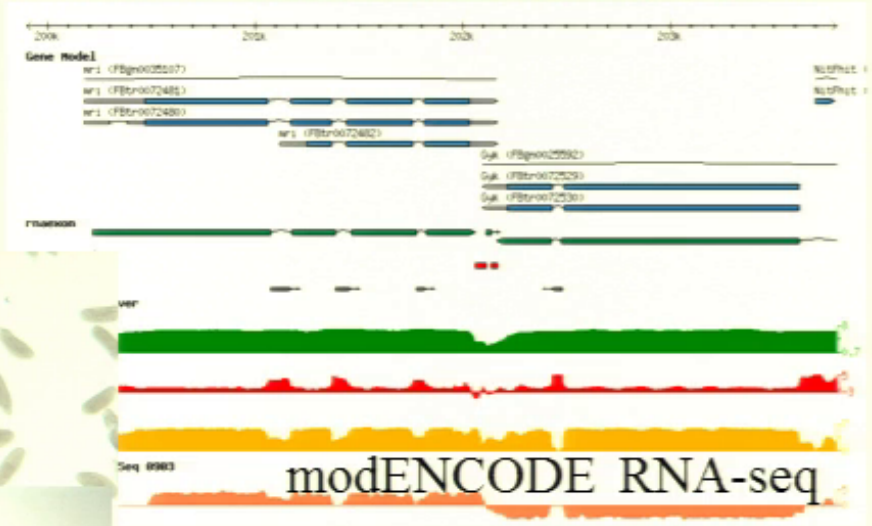
SIAM International Conference on Data Mining, Vancouver, 2015

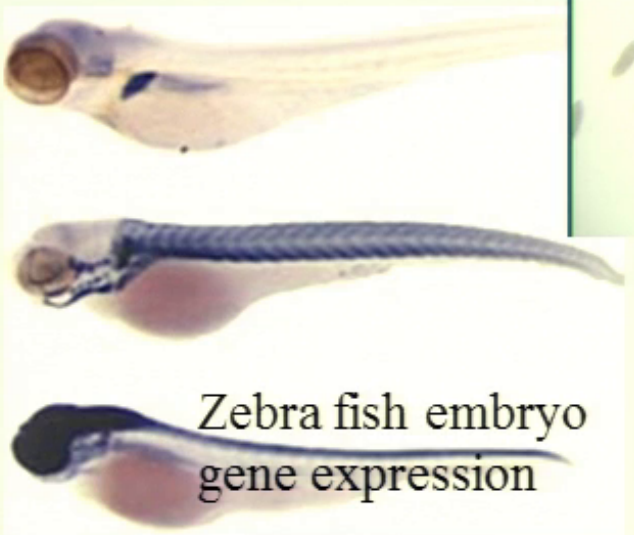
The abundance of systems biology data aims to answer:

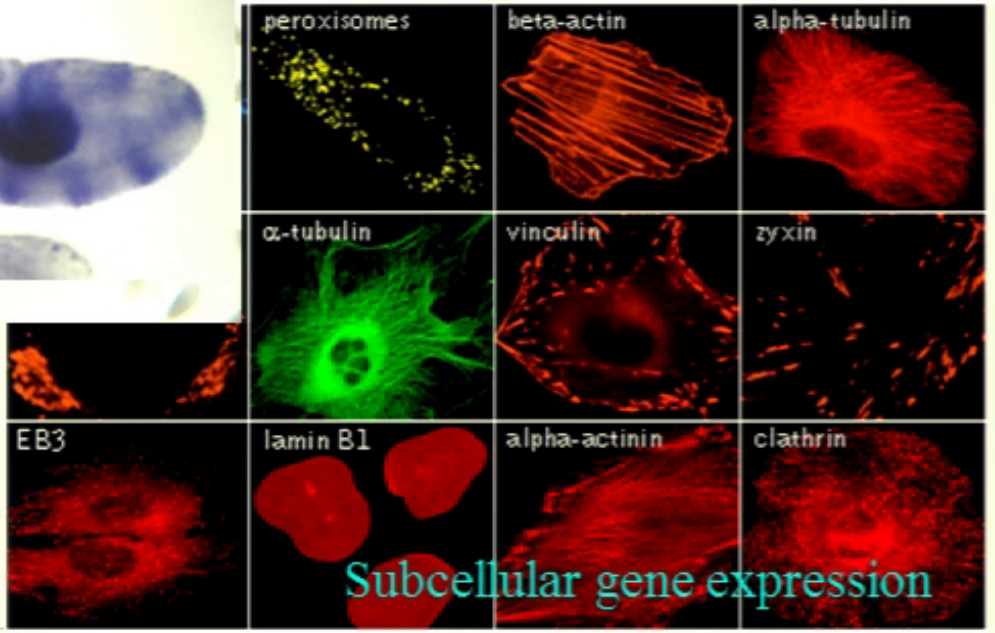
How do genes interact?









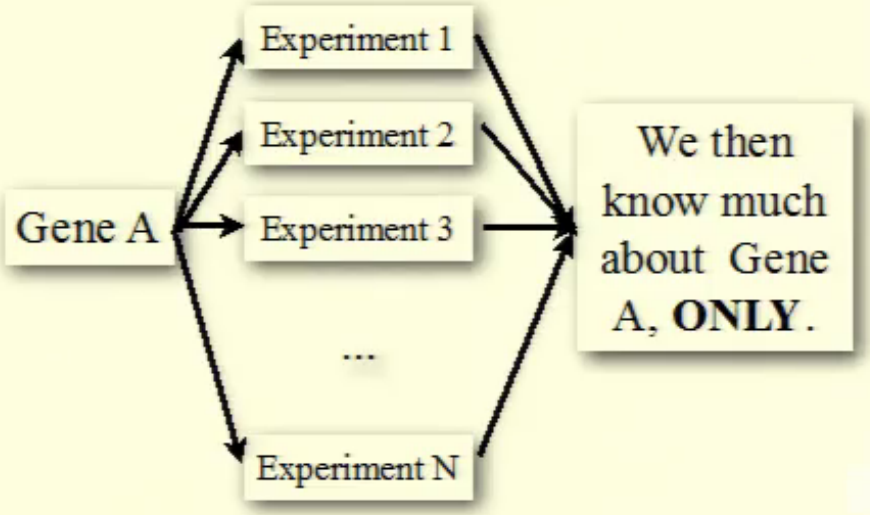


Big picture: Systems Biology

Systems Biology is the study of an organism as an **integrated and interacting network** of genes, proteins and biochemical reactions.

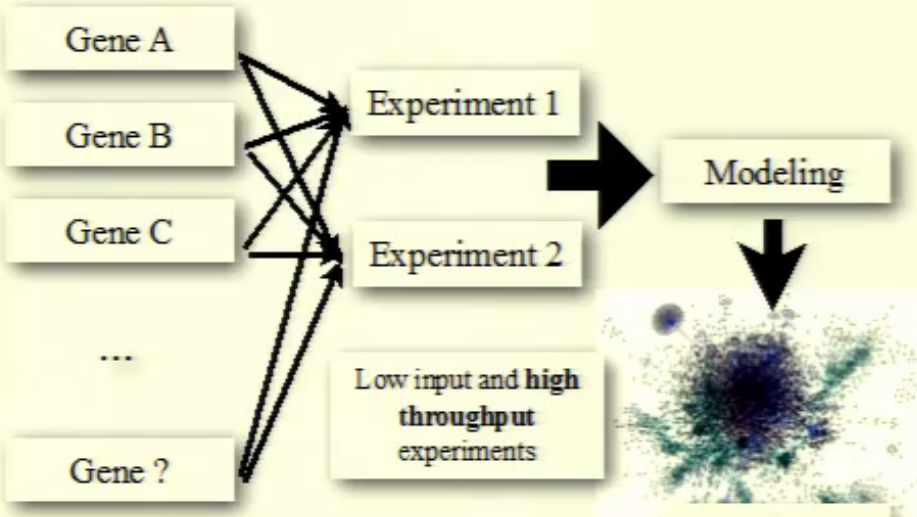
Traditional biology

Looking at **ONE** component at a time



Systems biology

Looking at **ALL** components at the same time



My guiding principles for data-intensive science

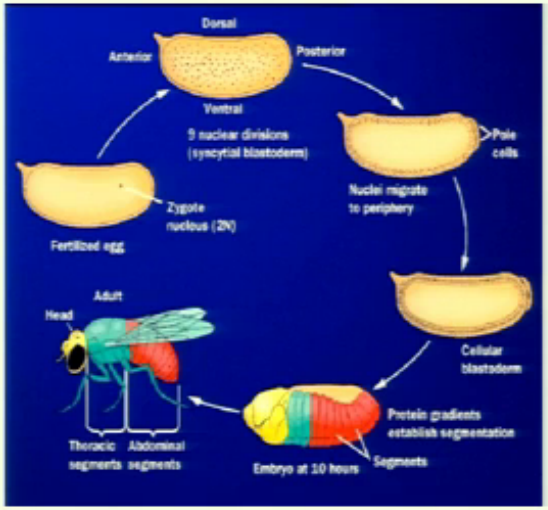
Seed scientific problem(s)

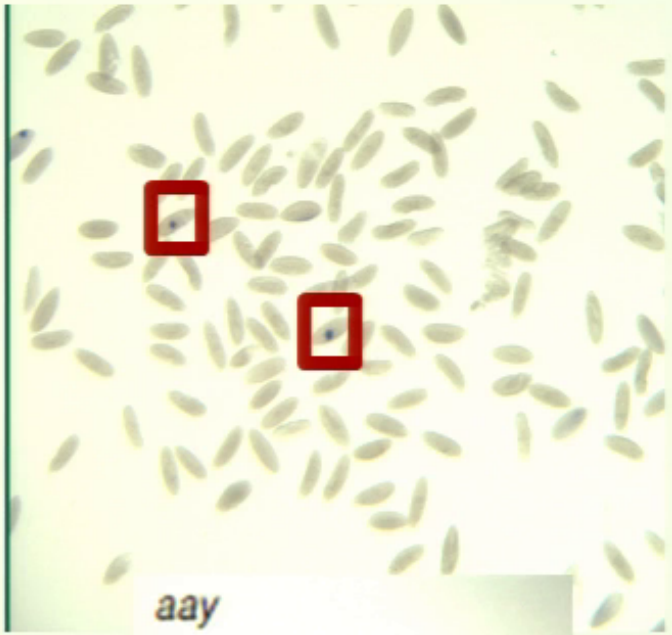
Generalization

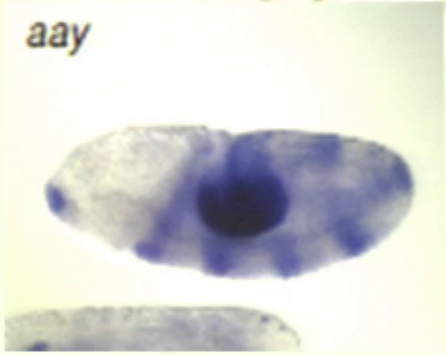
"Embedded" students/postdocs work on site,
in the wet lab

The Berkeley Drosophila Genome Project (BDGP)

(my biology collaborator Frise is in the BDGP Celniker Lab)





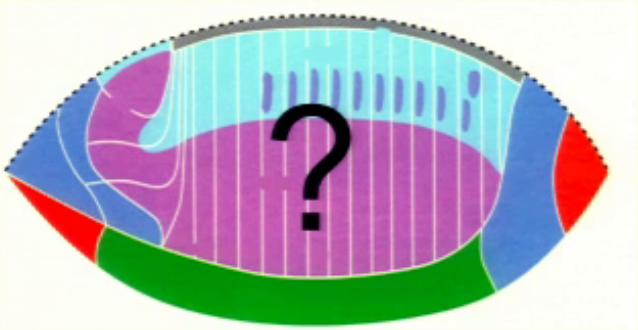


Genes interact in space:
spatial (image) gene expression data

The Berkeley Drosophila Genome Project (BDGP) (cont)

7K+ genes examined – about 1 TB data

We seek answers to the following questions:



Drosophila (fruitfly) embryo

How many functional regions are there?

Or can we re-produce the fate map with our data?

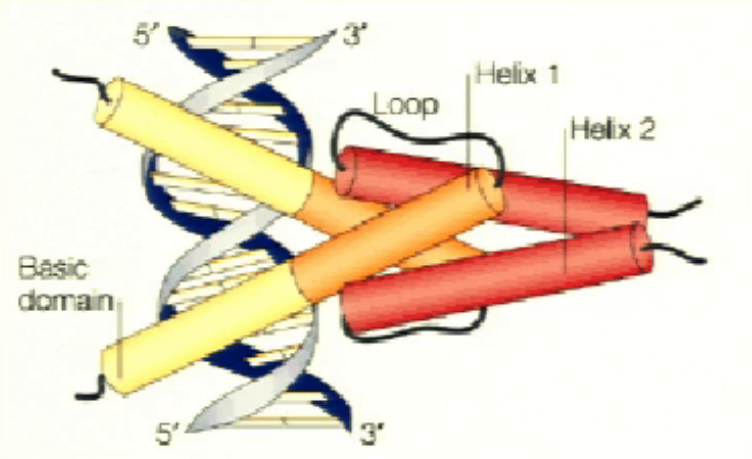
Gene A Gene B

What are the new gene functions and gene-gene interactions?

Transcription Factors (TFs):

trigger molecules, corresponding to genes

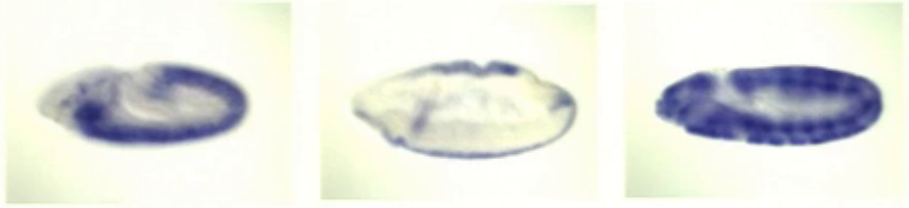
1. What are Transcription Factors(TFs)? DNA binding molecules. On-off switch – triggering other TFs/genes to express.



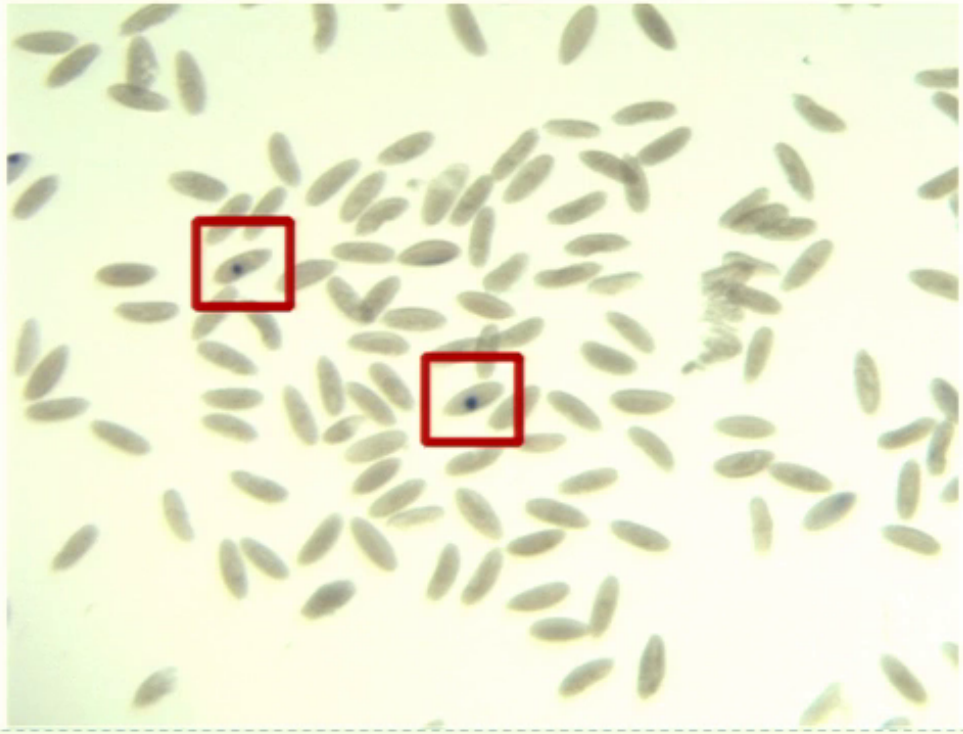
2. Why do we study TFs? They are drivers of gene function cascades and are believed to be almost all of the genes at work in early stages.

Data collection procedure (1/2)

* Using dye chemistry to visualize spatial gene patterns



* Imaging under a microscope



Data collection procedure (2/2)

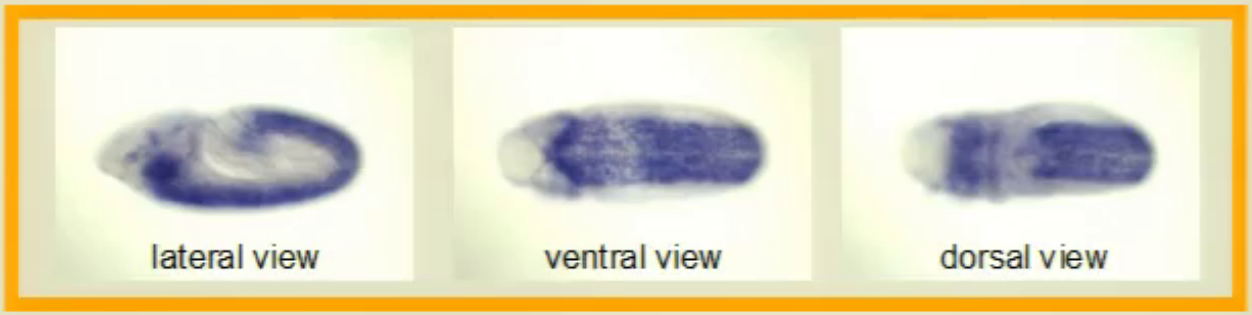
Find representatives for each stage bin.

e.g. CG6096

All these were done manually!

For each stage bin, find representatives for each of the three orientations: lateral, ventral and dorsal.

e.g. CG6096
stage 9-10



Data preprocessing (1/2)

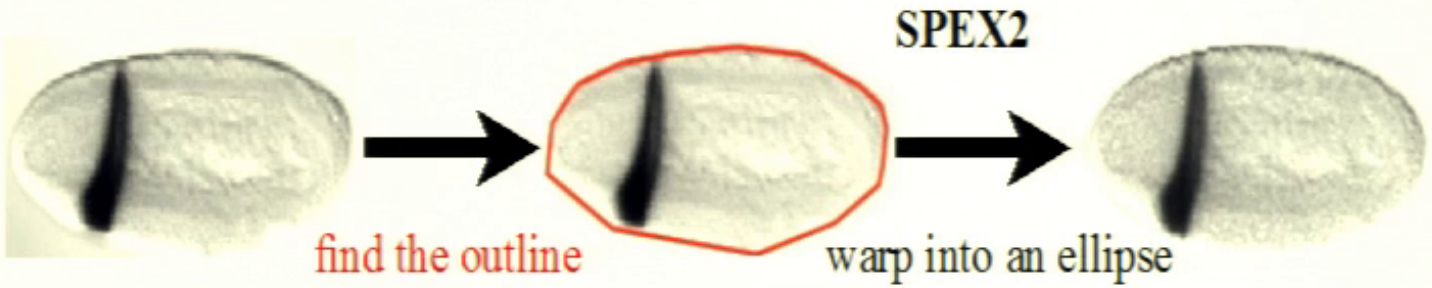
1. Image segmentation: Frise et.al. (2010).

2. Image registration on an ellipse: SPEX2 algorithm by Puniyani and Xing (2010).

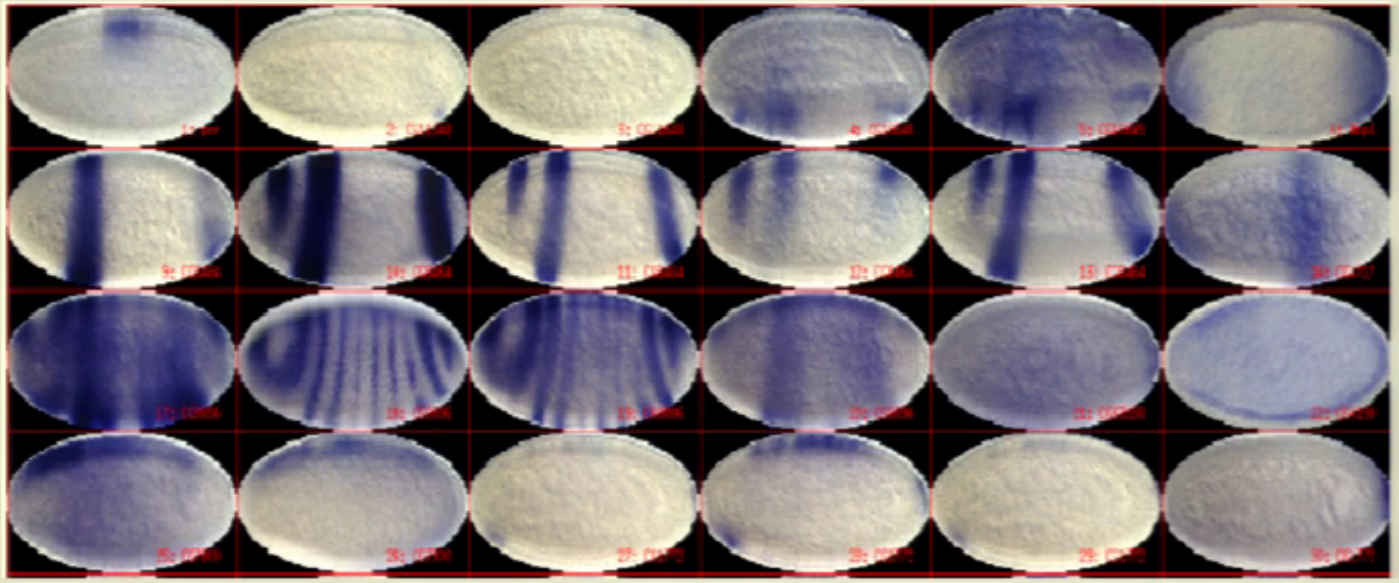
3. Stain extraction: remove shades and shadows (Wu)

4. FlyExpress – software toolkit by Kumar et. al (2011)
   MyFX – software toolkit by Muntiel et. al (2015)

Data preprocessing (2/2): much work



A sample of resulting images



• Stain extraction

Previous work on the same or similar expression data sets

1. RNA-seq data: measuring the expression level for the entire embryo – no spatial information.

2. Human annotation data: (Hammands et.al 2013) limited to a fixed set of control vocabulary; no intensity information.

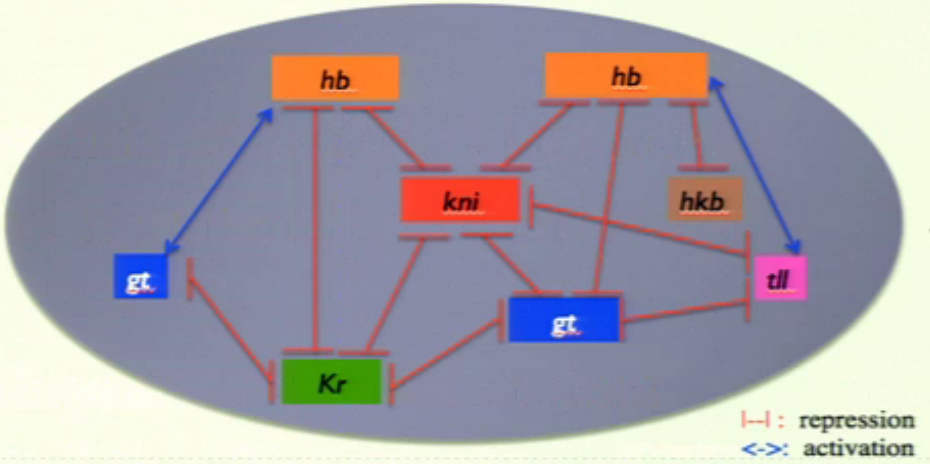
3. High resolution Drosophila embryonic expression data (Fowlkes et.al. 2008): only limited to a few number of TFs.

4. Previous work on the same data set:
   Frise et.al. 2010 – global analysis
   Pruteanu-Malinici et.al 2011
   – use image data to predict human annotation
   Yuan et. al 2015 -- prediction of stage using human annotation

The gap gene network: genes interact locally in space

Segmentation (vertical "coordinate"): work of the gap gene network

Fruitfly embryo: segmentation

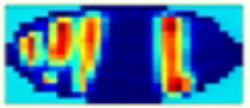


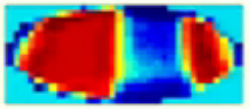
Human embryo: segmentation

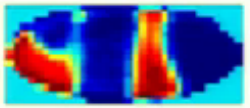
Nobel Prize in 1995 in Physiology or Medicine for work on gap gene network in the 70's.

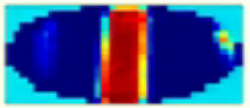
Christiane Nusslein-Volhard

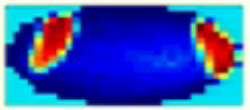
What are the gap genes?


Gt


Hb


Kni


Kr

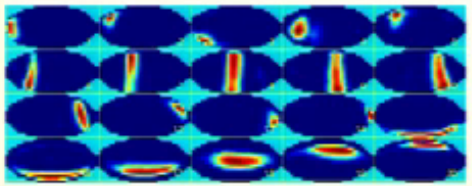

tll


hkb

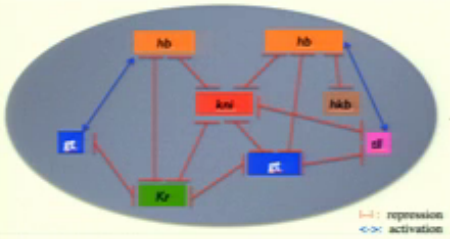
It is known that Hb represses Gt on the right of the embryo. In terms of expression patterns, Gt and Hb have complimentary patterns towards the right end of the embryo. So their local correlation is negative.

Similarly, Kr and Kni are mutual repressors in the middle part of the embryo.

Our framework overview for building local gene functional networks in systems biology



Interpretable representation (data-driven sparse coding) into principle patterns



"Spatially" local correlations via principle patterns

Recommended gene list

Gene Name	Relevance score
CG9571	1.0
CG10979	0.95
...	...

Recommendation on interactions (edges) via new statistics/machine learning

Wet-lab experiments for validation

Spark

"Recommendation system" for wet-lab actions

Statistics/ML methodology development with scalable algorithms

- Multiple hypothesis testing via decision theory and supervised learning (under dependence)
- Data perturbation-based methods (stability and uncertainty)
- Data fusion with "prior" information in PubMed via text summarization (CCS)

Computer Science development with open-source platforms:

CISBA

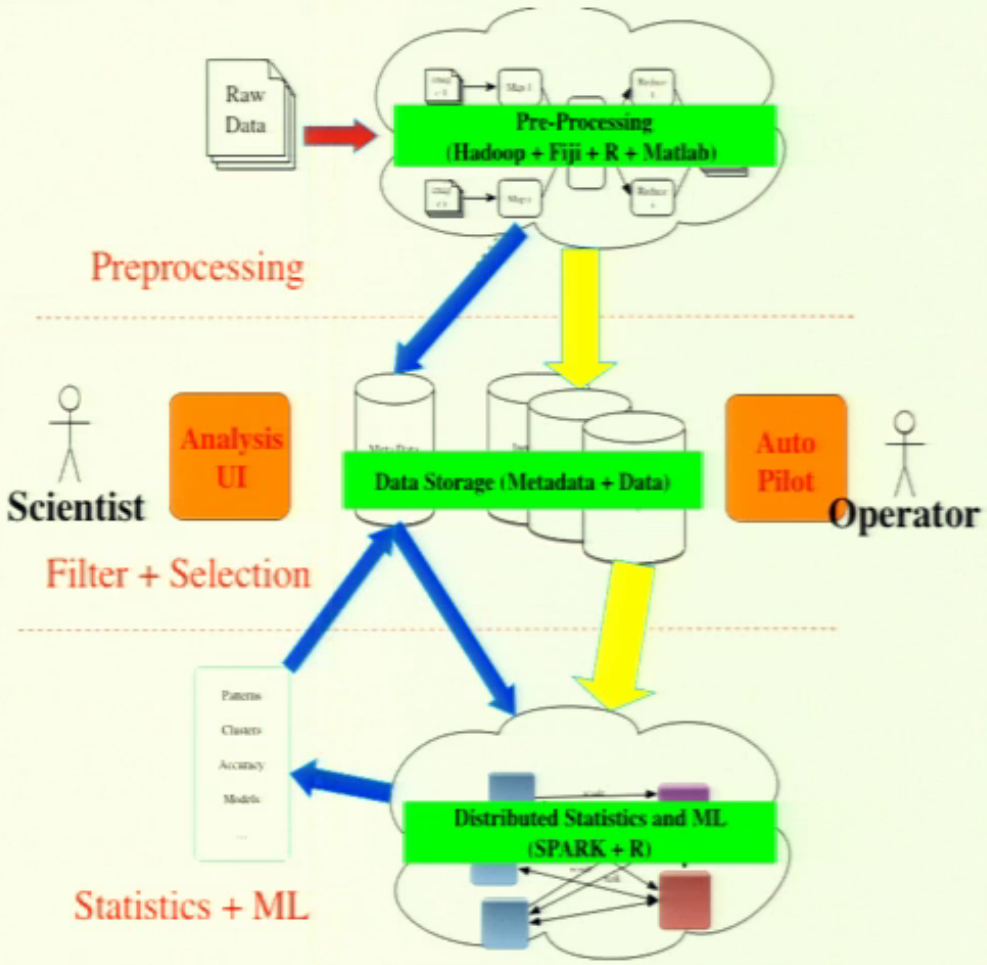
- Data and metadata management
- Data driven automation
- Open source

CISBA:
cloud-based infrastructure for systems biology analytics

For Scientists
- Intuitive UI
- Rich selection without coding
- Quick prototyping w/ Java or Python

For System Admin
- Easy module management
- Auto Monitoring and Recovery

Key CS Research Topics
- Separating data / metadata
- Data-driven automation
- Open source tools
  - Hadoop, SPARK
  - MongoDB, FIJI



How to find functional local regions?

Observation by biologists after examining a large number of expression images:

Complicated expression patterns can be decomposed into a set of simpler and connected **building block patterns**

Each building block pattern may correspond to a unique combination of gene functions or/and signaling pathways

Six building block patterns in Drosophila eggshell (follicle) cells. Yakoby et al 2008

Data-driven interpretable data representation vi Nonnegative Matrix Factorization (NMF)

$$\min_{\mathbf{D}\geq 0,\mathbf{A}\geq 0} \|\mathbf{X}-\mathbf{D}\mathbf{A}\|_F^2$$

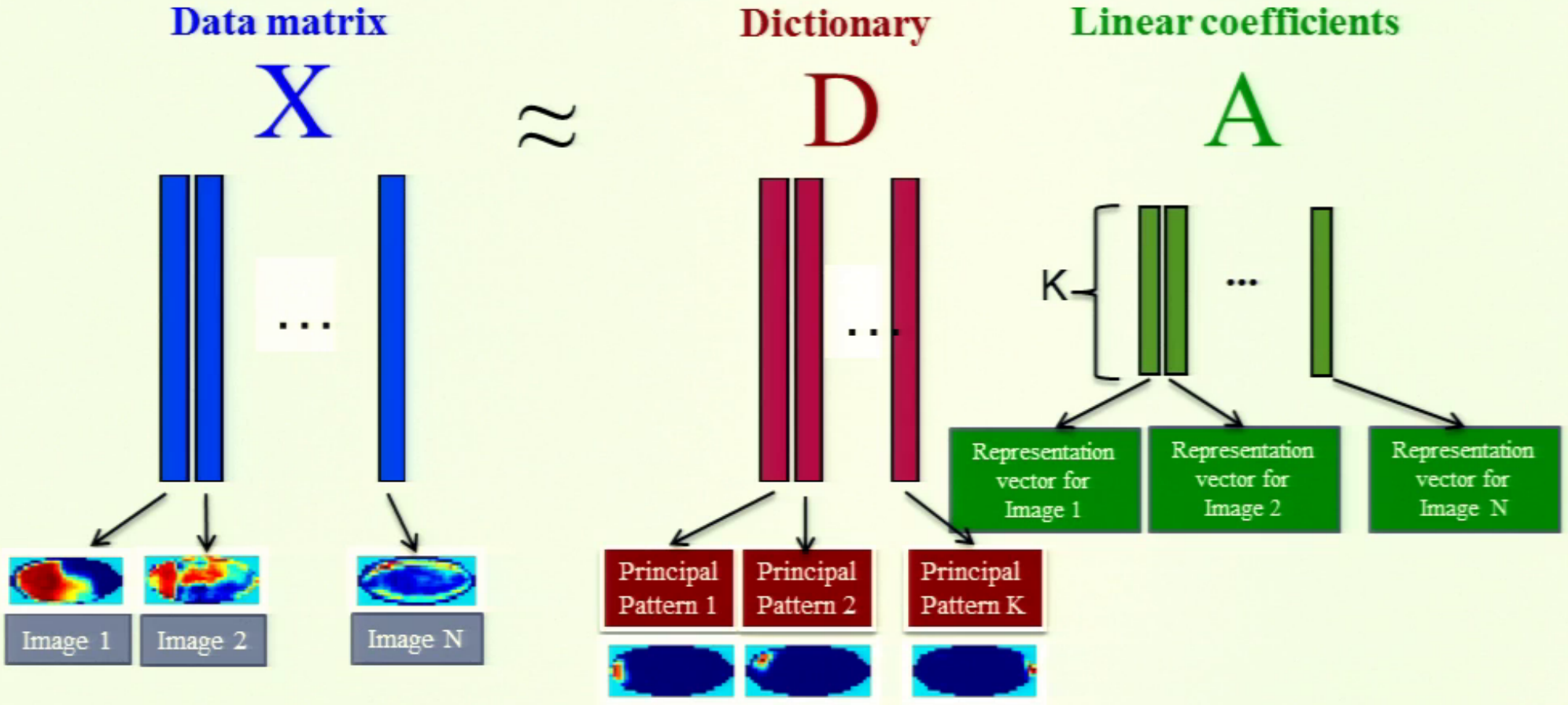
Data matrix **Dictionary** **Linear coefficients**

$$\mathbf{X} \approx \mathbf{D} \quad \mathbf{A}$$

References: … Lawton and Sylvestre (1971)… Lee and Seong (1999), … (SPAMS) by Marial et.al (2010).

Data-driven interpretable data representation vi Nonnegative Matrix Factorization (NMF)

$$\min_{\mathbf{D}\geq 0,\mathbf{A}\geq 0} \|\mathbf{X}-\mathbf{DA}\|_F^2$$

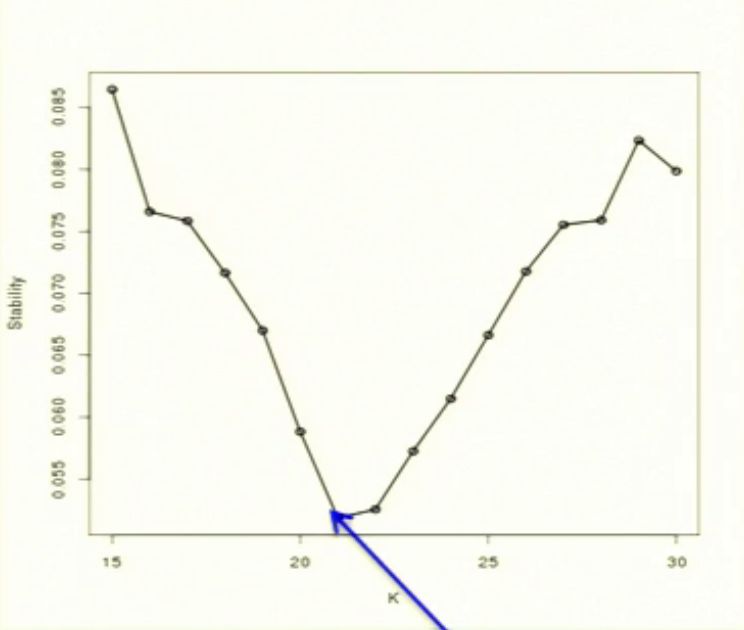


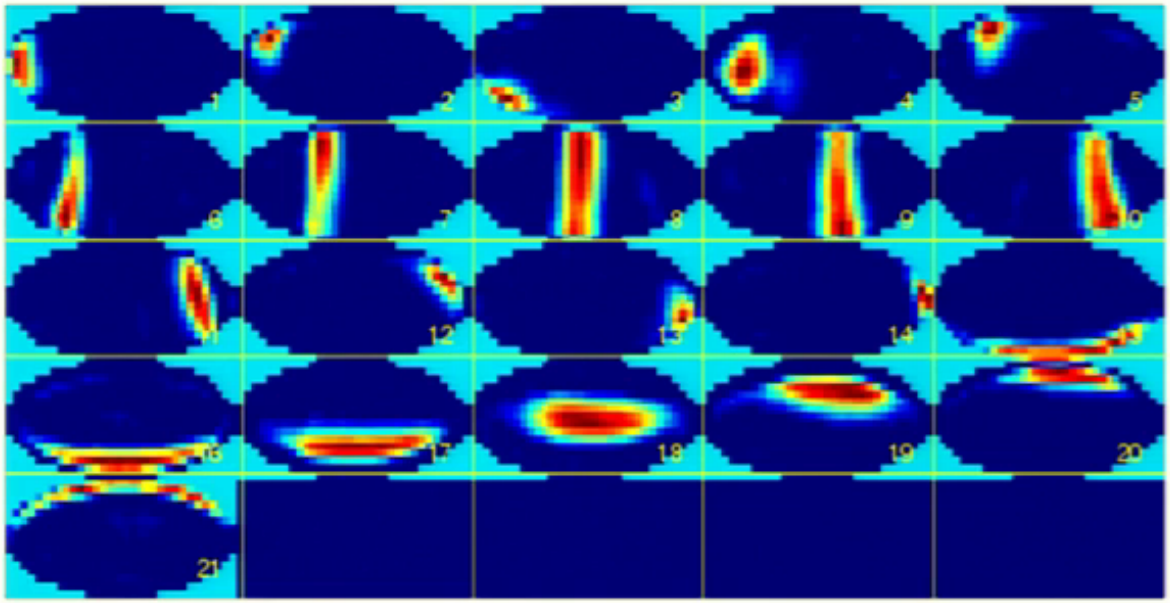
References: … Lawton and Sylvestre (1971)… Lee and Seong (1999), … (SPAMS) by Marial et.al (2010).

New stability criterion for the number of principal patterns

Let $C \in \mathbb{R}^{K \times K}$ be the cross correlation matrix for two dictionaries D_1 and D_2 in $\mathbb{R}^{405 \times K}$. Define a dissimilarity measure between D_1 and D_2 as

$$diss(D_1, D_2) = \frac{1}{2}\left(\frac{1}{K}\sum_j (1 - \max_i C_{ij}) + \frac{1}{K}\sum_i (1 - \max_j C_{ij})\right).$$



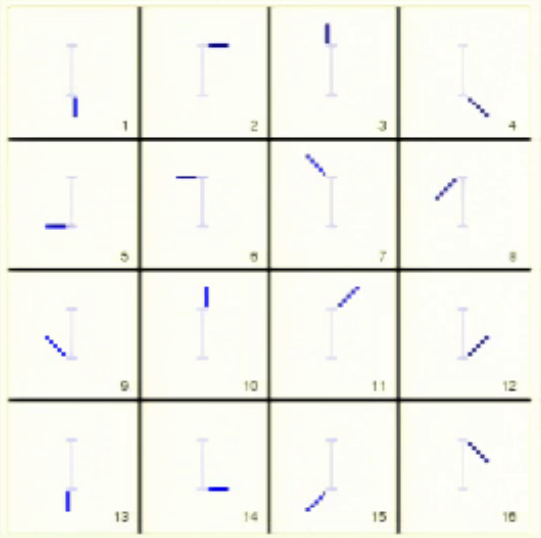


The data-driven dictionary with 21 principal patterns

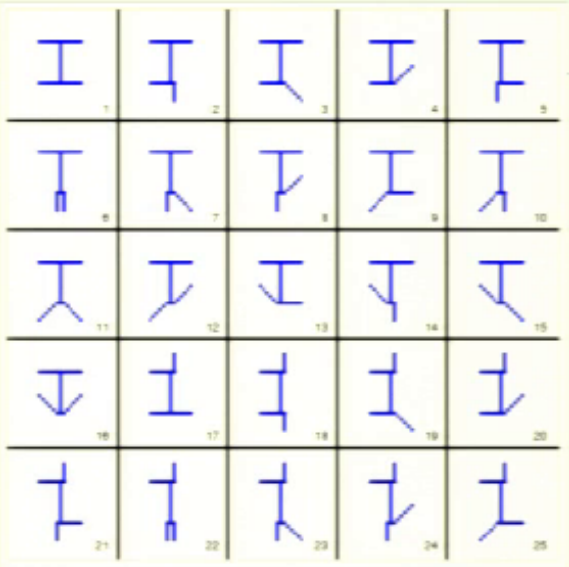
K = 21 is the number of patterns selected by our stability criterion

Stability heuristics on artificial data

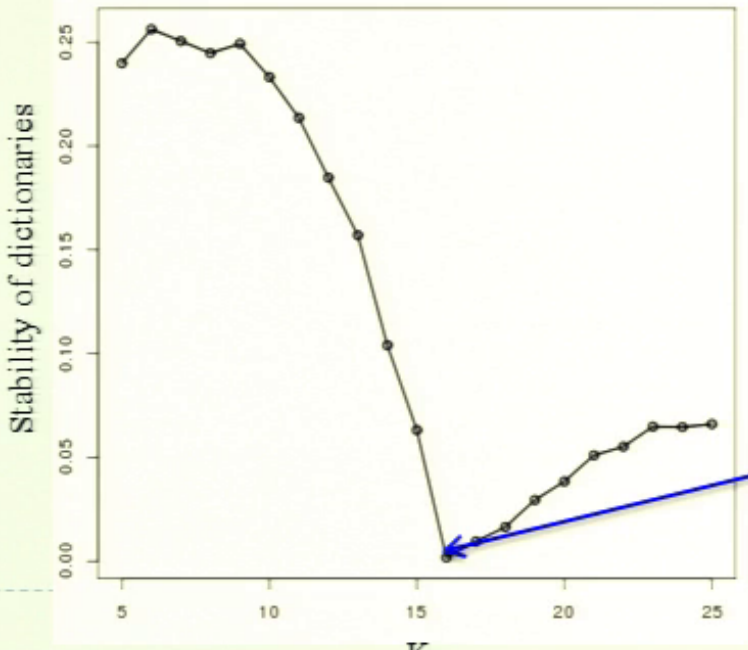
Swimmer data simulation (Donoho and Stodden 2001)



16 generator patterns

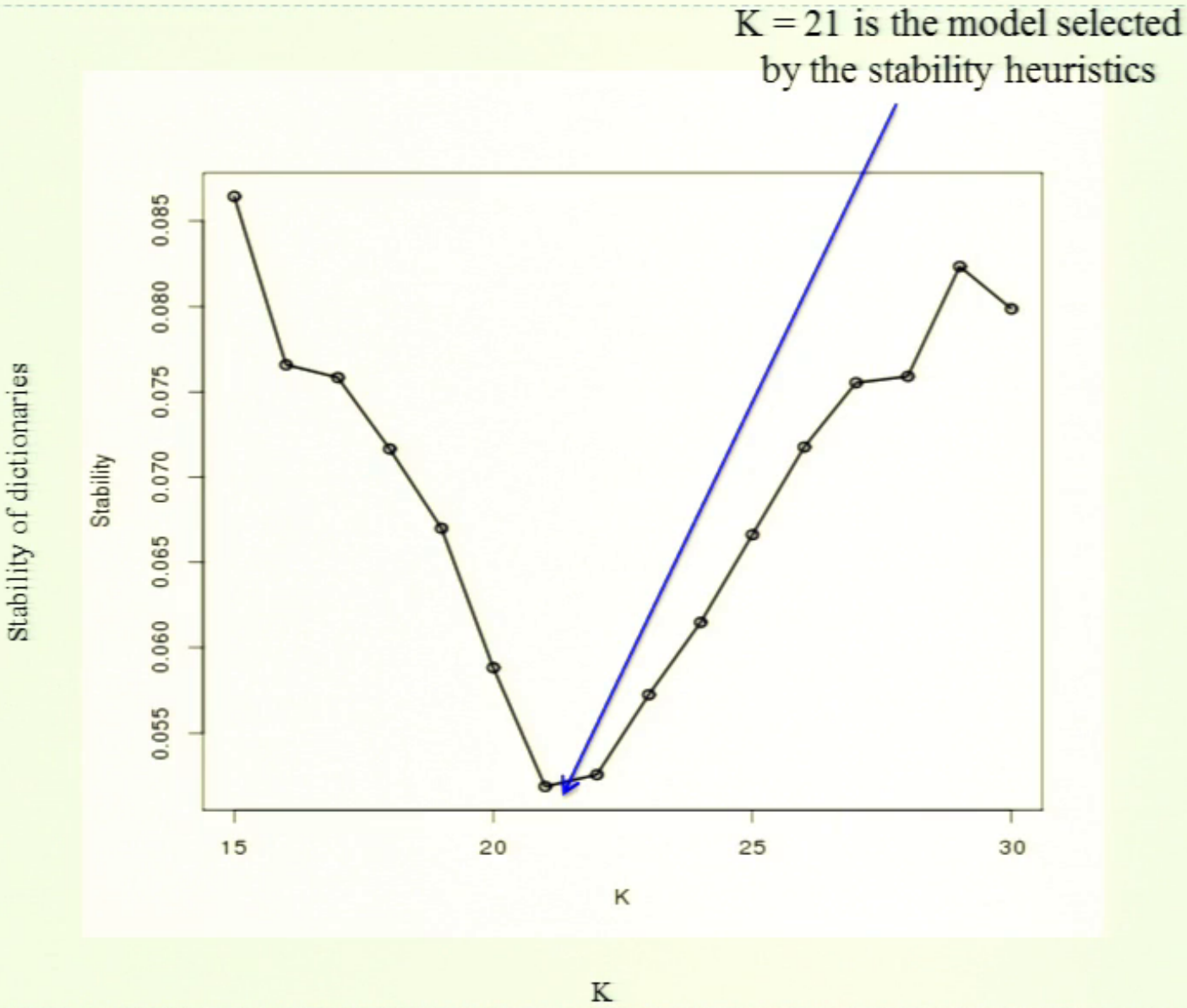


Sample images from the swimmer data

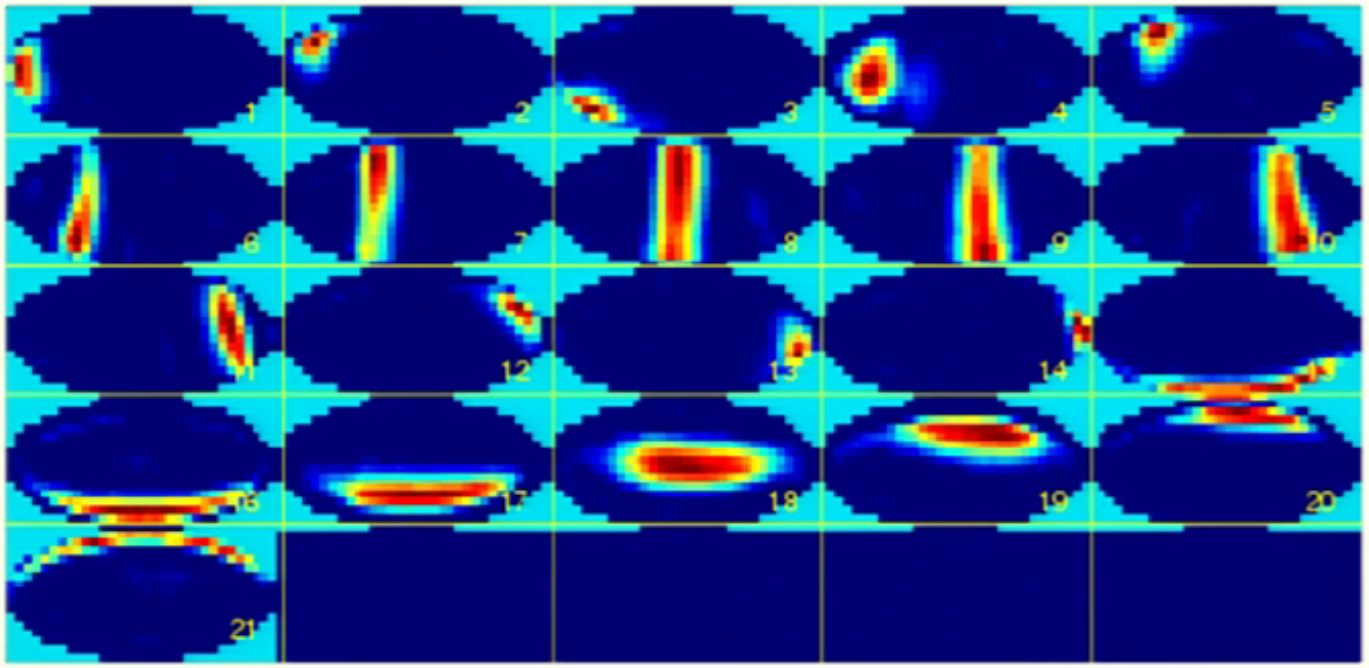


K = 16 is the model selected by the stability heuristics

Stability heuristics on the Drosophila expression data



Learned dictionary of 21 principal patterns via NMF

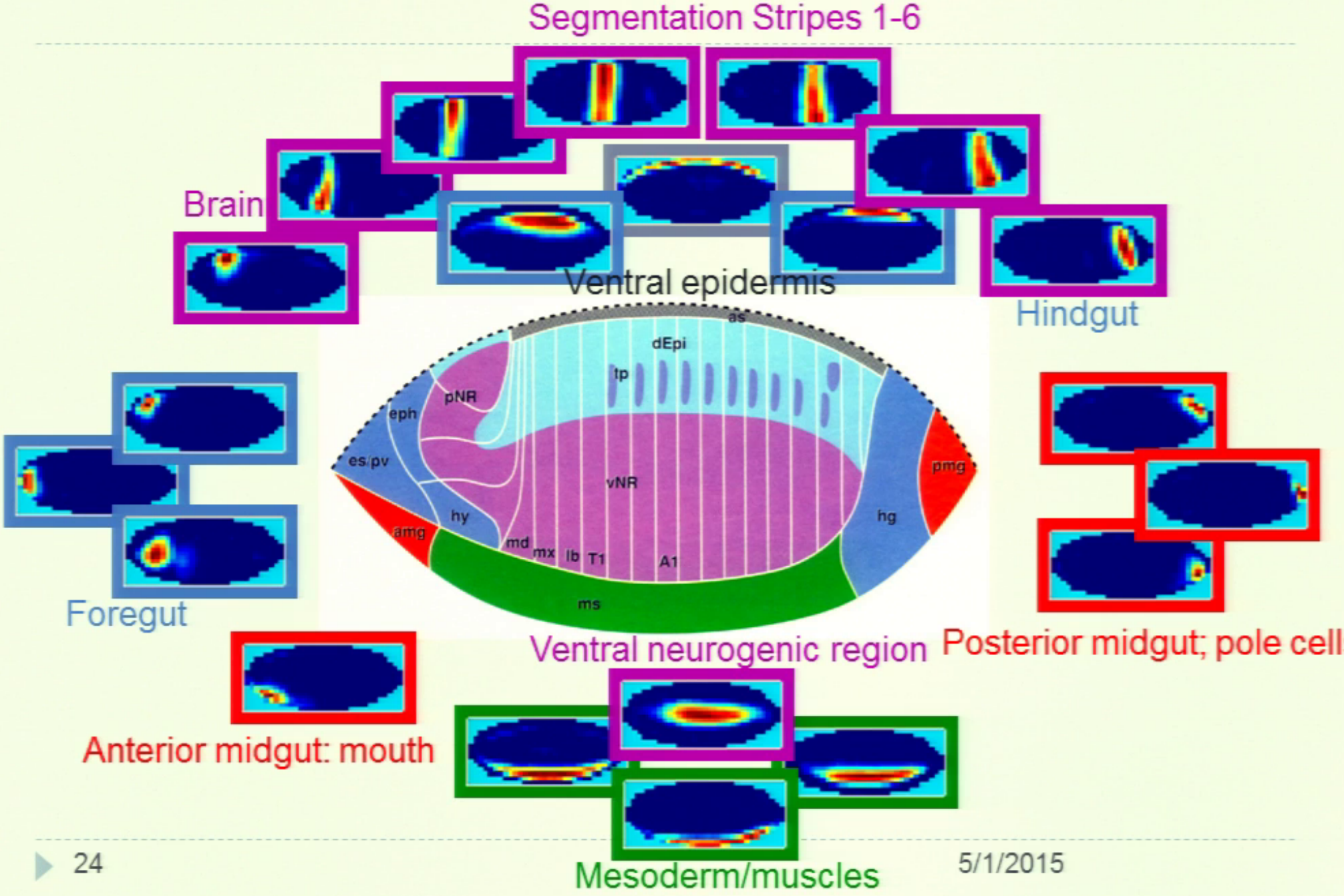


Four types of patterns:

Anterior, Vertical (stripes), Posterior, Horizontal

The four types of patterns work together to define a rough "coordinate" system in early Drosophila embryonic development.

Principal patterns are biologically meaningful

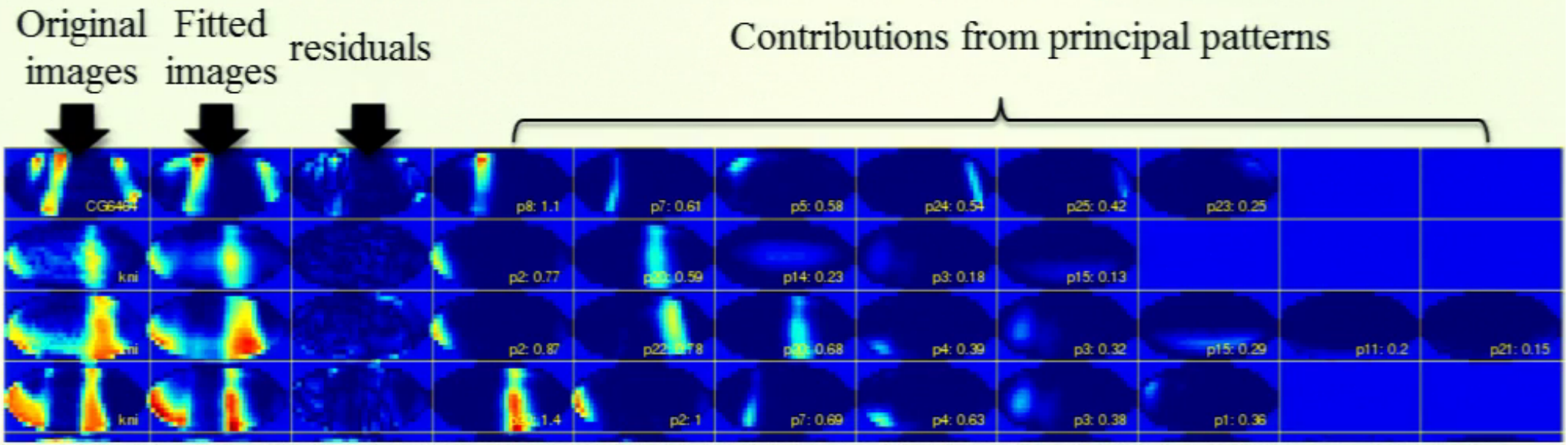


Local gene expression patterns via principal patterns

Lasso+Non-negative Least Squares (NLS)

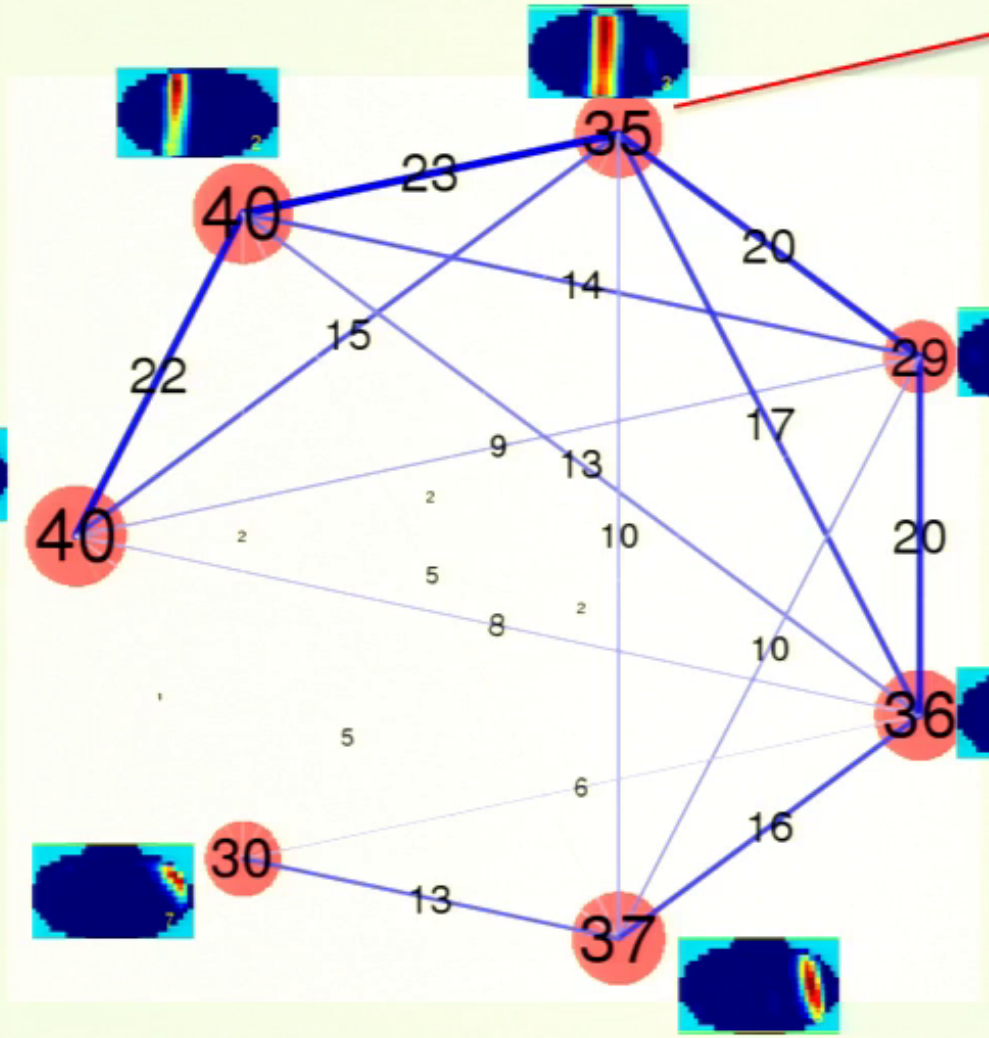
For each gene expression image,

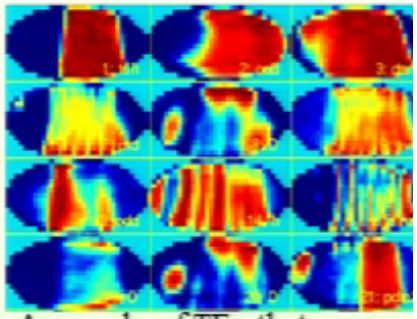
1. use the Lasso (with nonnegative constraint) to select the principal patterns, and
2. refit the image on the selected principal patterns with NLS



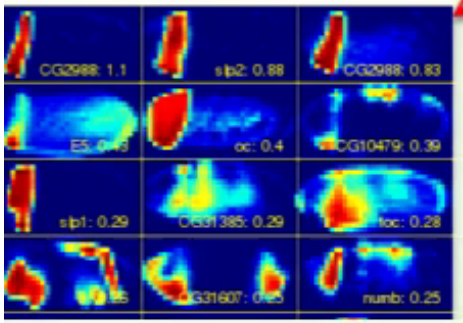
Gap genes: number of shared genes over neighboring stripes

There are 104 TFs (out of 156 TFs) that are expressed on at least one of the segmentation stripes.





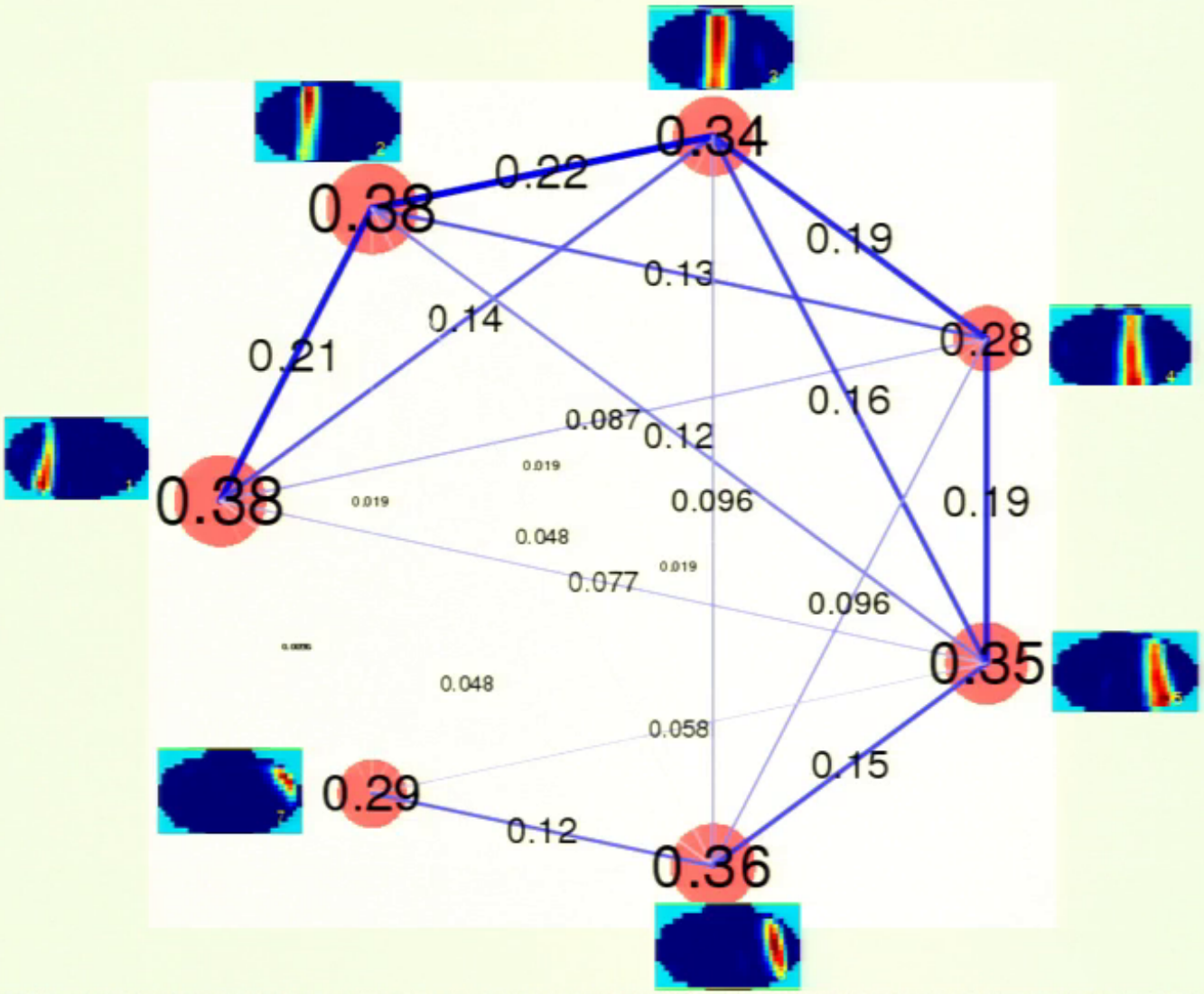
A sample of TFs that are expressed on Stripe 3 and Stripe 5



A sample of TFs that are expressed on Stripe 1

Proportions: previous numbers divided by 104

They are consistent with biologists' qualitative understanding.



Relating TFs by their principal patterns via correlation

1. **Filter Step via the learned patterns**: For each stripe principal pattern, filter out (by the LASSO+OLS procedure) the TFs that are expressed in the region or its neighbor stripe patterns.

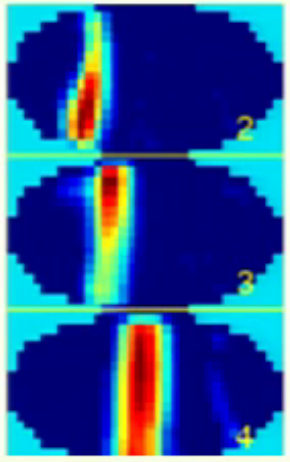
2. Compute the **locally weighted correlation** between the selected TFs.

e.g. Weights for Stripe 2

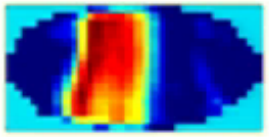
Neighbor stripe: Stripe 1

Stripe 2

Neighbor stripe: Stripe 3

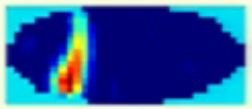


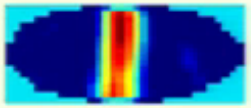
sum =

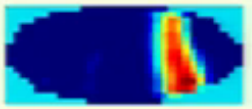


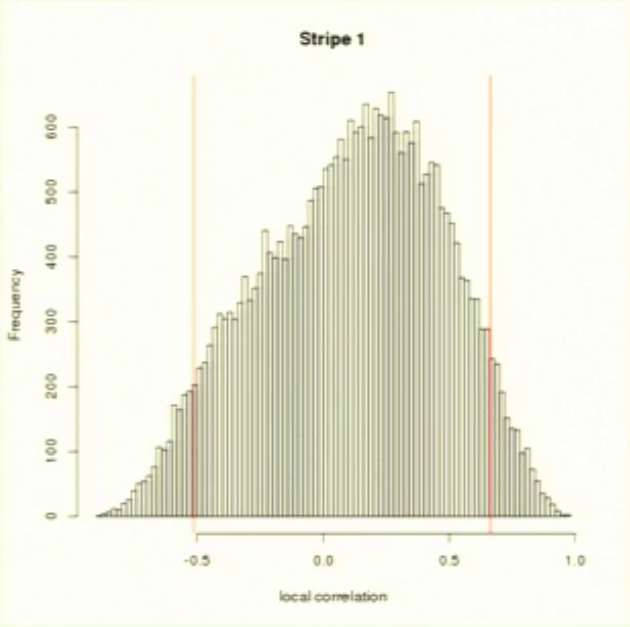
Then normalize to have sum = 1

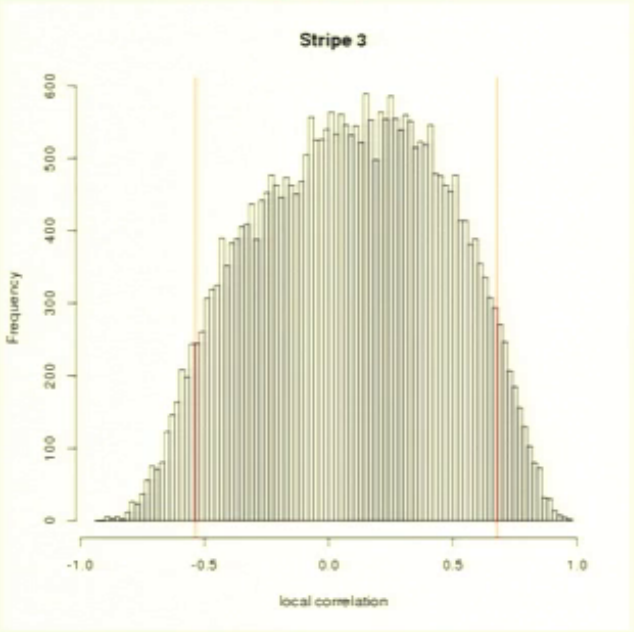
Distribution of correlations for three networks

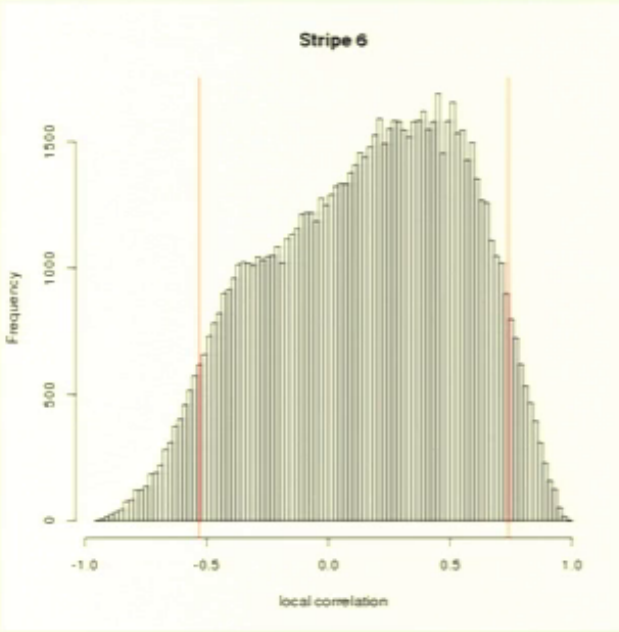






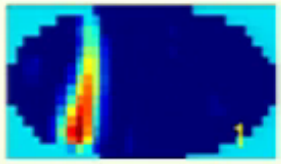






Note the asymmetry of the distribution – tends to skew towards the positive direction. The mean is greater than zero.

Spatially local TF network constructed based on Stripe 1



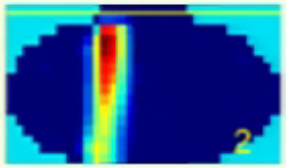
Stripe 1 in the set of learned principal patterns

Edges with correlations beyond the 5% of the lower and upper tails of the correlation distribution are displayed.

Blue edges indicate positive correlations.

Red edges indicate negative correlations.

Spatially local TF network constructed based on Stripe 2

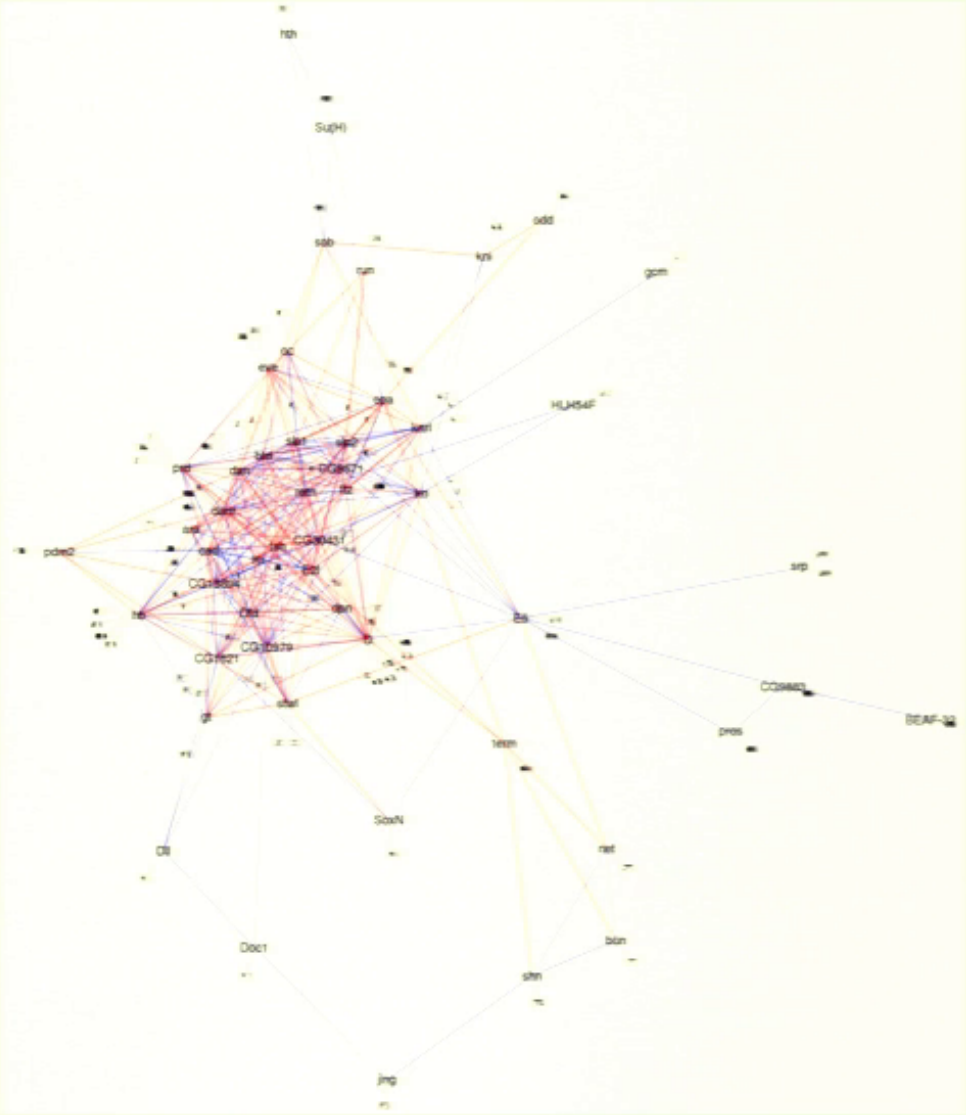


Stripe 2 in the set of learned principal patterns

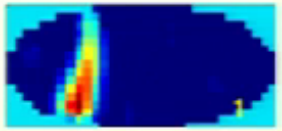
Edges with correlations beyond the 5% of the lower and upper tails of the correlation distribution are displayed.

Blue edges indicate positive correlations.

Red edges indicate negative correlations.



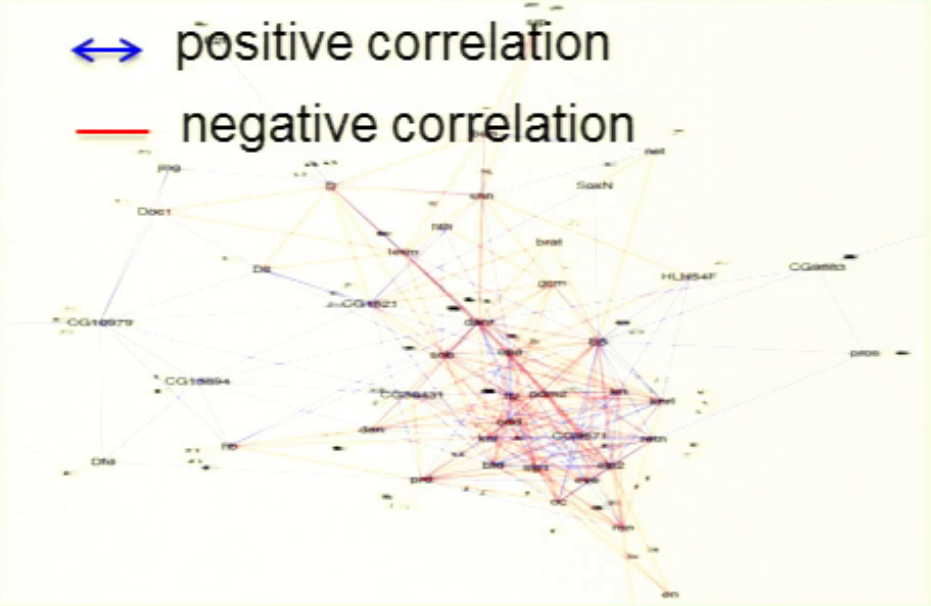
Two spatially local gene networks



Stripe 1 in the set of principal patterns

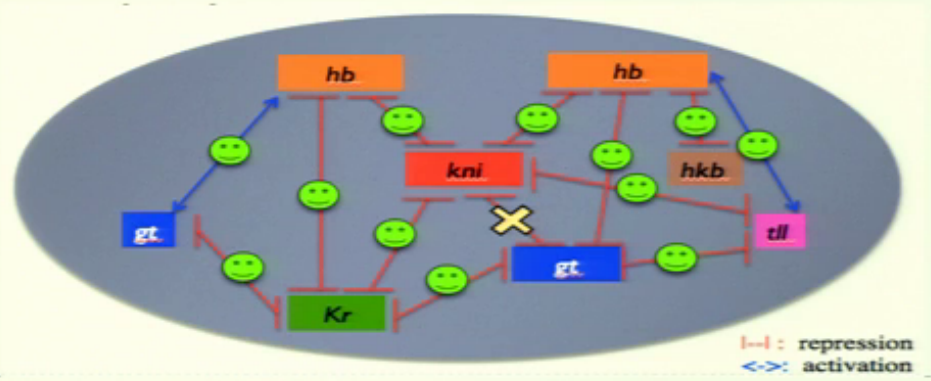
Simple cut-off rule:

keep edges with correlations outside the 5th and 95th percentiles of the empirical correlation distribution.

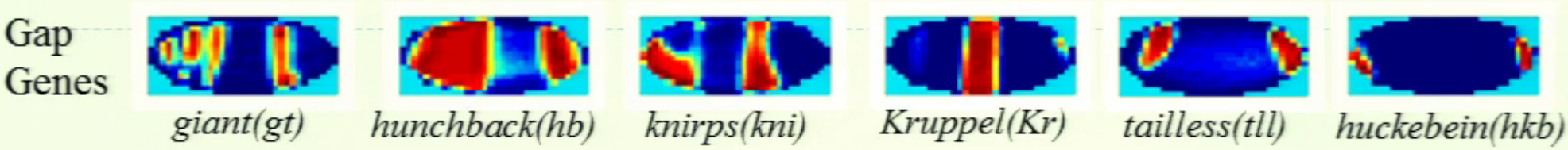


"Reproducing" gap gene network:

12 out of 13 known edges correctly predicted by our simple cut-off rule.

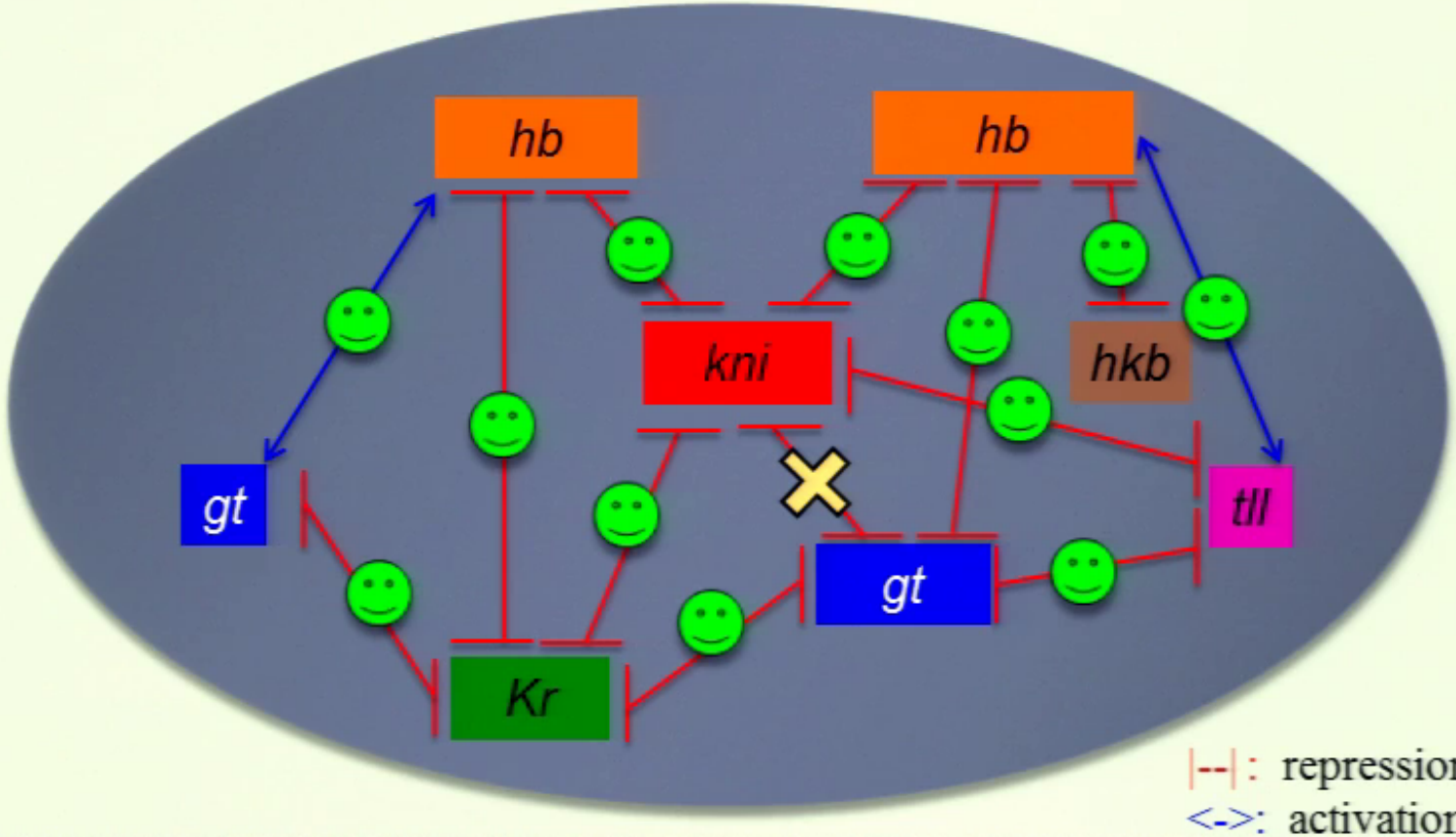


Our analysis agrees well with the known gap gene network



: Interactions correctly predicted

: Interactions that we missed

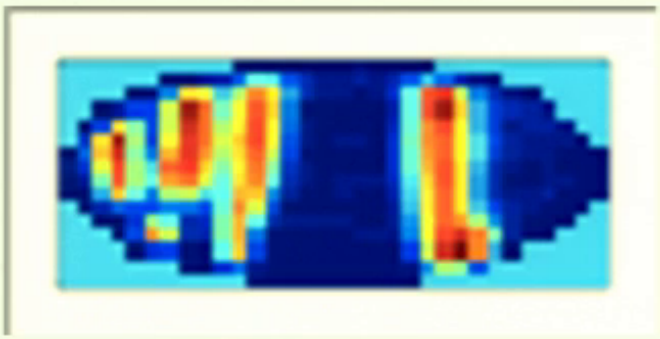


Anterior of the embryo

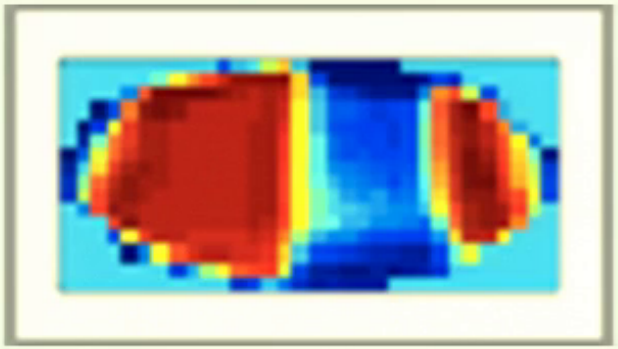
Posterio the emb

Gt – Hb relationship

Our data

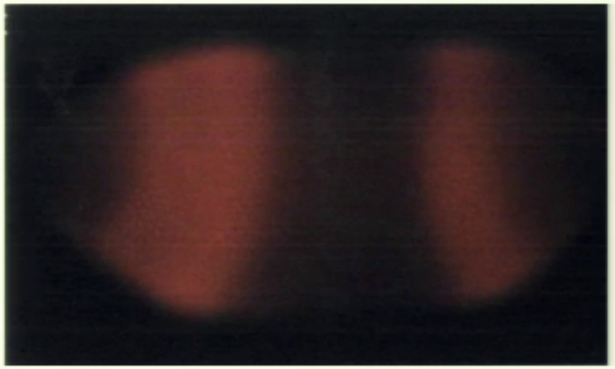


Gt

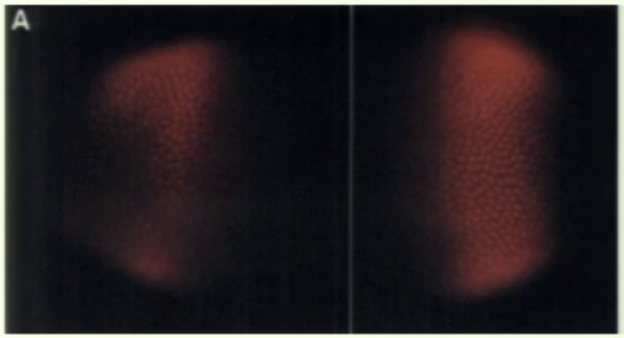


Hb

Biological validation

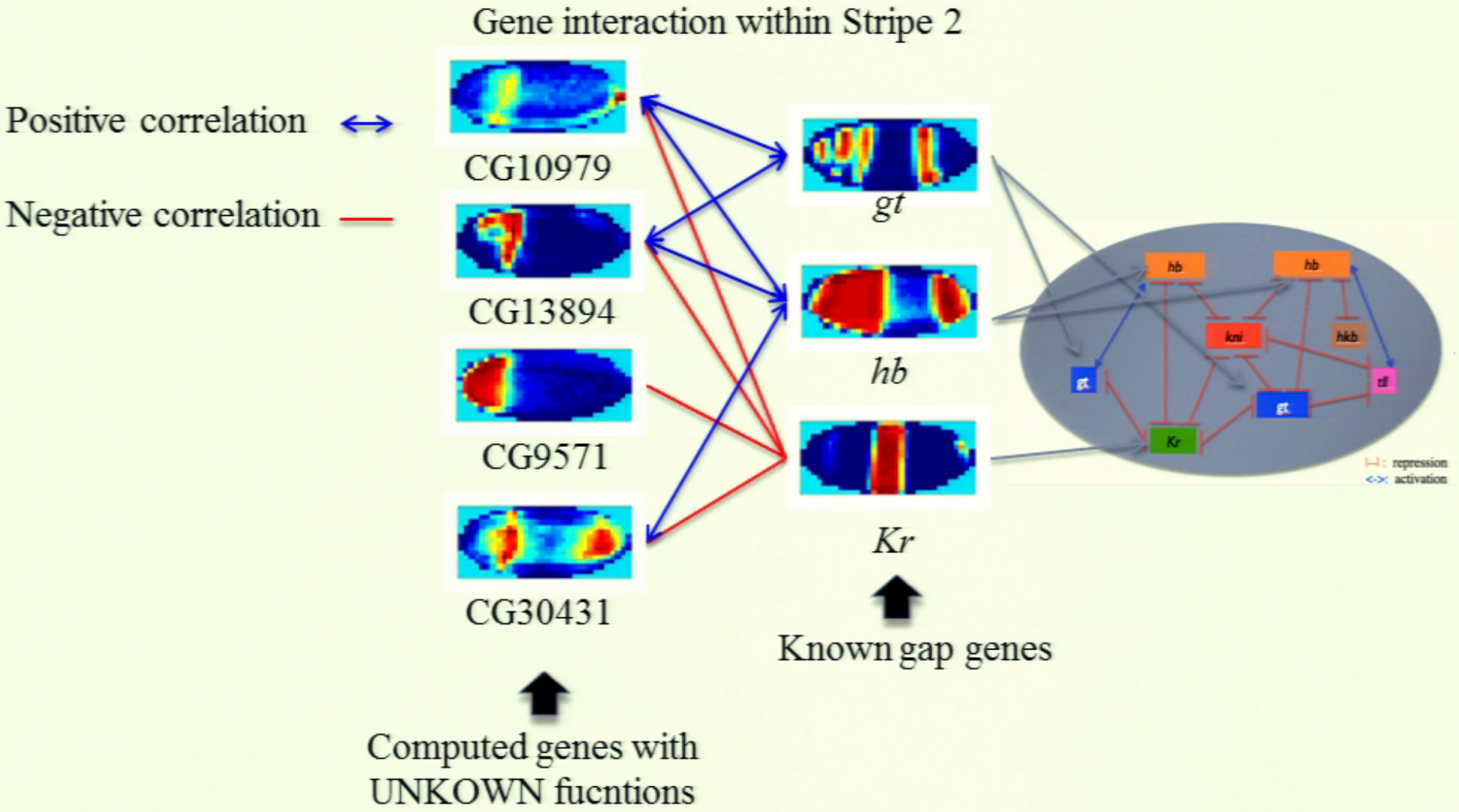


Gt in Hb wild-type



Gt in Hb mutant

New gene and new gene interactions in the gap network?



IV: On –going knock-out experiments using the CRISPR-Cas9 system

- 4 computed genes (CGs) as potential new discoveries
- 5 well-studied TFs, 3 in the gap gene network and 2 pair-ruled genes
- 2 hypothesized directions: upstream/downstream
- Total 4*5*2 = 40 experiments
- For each set-up, examine the spatial pattern of the downstream gene:

Positive correlation predicts weakened pattern downstream;
Negative correlation predicts expanded pattern downstream.

Theoretical study: when does dictionary learning algorithm work?

$\mathbf{D} \in \mathbb{R}^{d \times K}$: a dictionary with K atoms.

$\mathbf{x}_i \in \mathbb{R}^d$: an observation, e.g. an image.

We study the l_1 norm minimization dictionary learning (Wu and Y., 2015):

$$\min_{\mathbf{D} \in \mathcal{D}} L_N(\mathbf{D}) = \min_{\mathbf{D} \in \mathcal{D}} \frac{1}{N} \sum_{i=1}^{N} l(\mathbf{x}_i, \mathbf{D}).$$

$$\text{where } l(\mathbf{x}, \mathbf{D}) = \min_{\boldsymbol{\alpha} \in \mathbb{R}^K} \{\|\boldsymbol{\alpha}\|_1, \text{ subject to } \mathbf{x} = \mathbf{D}\boldsymbol{\alpha}\}.$$

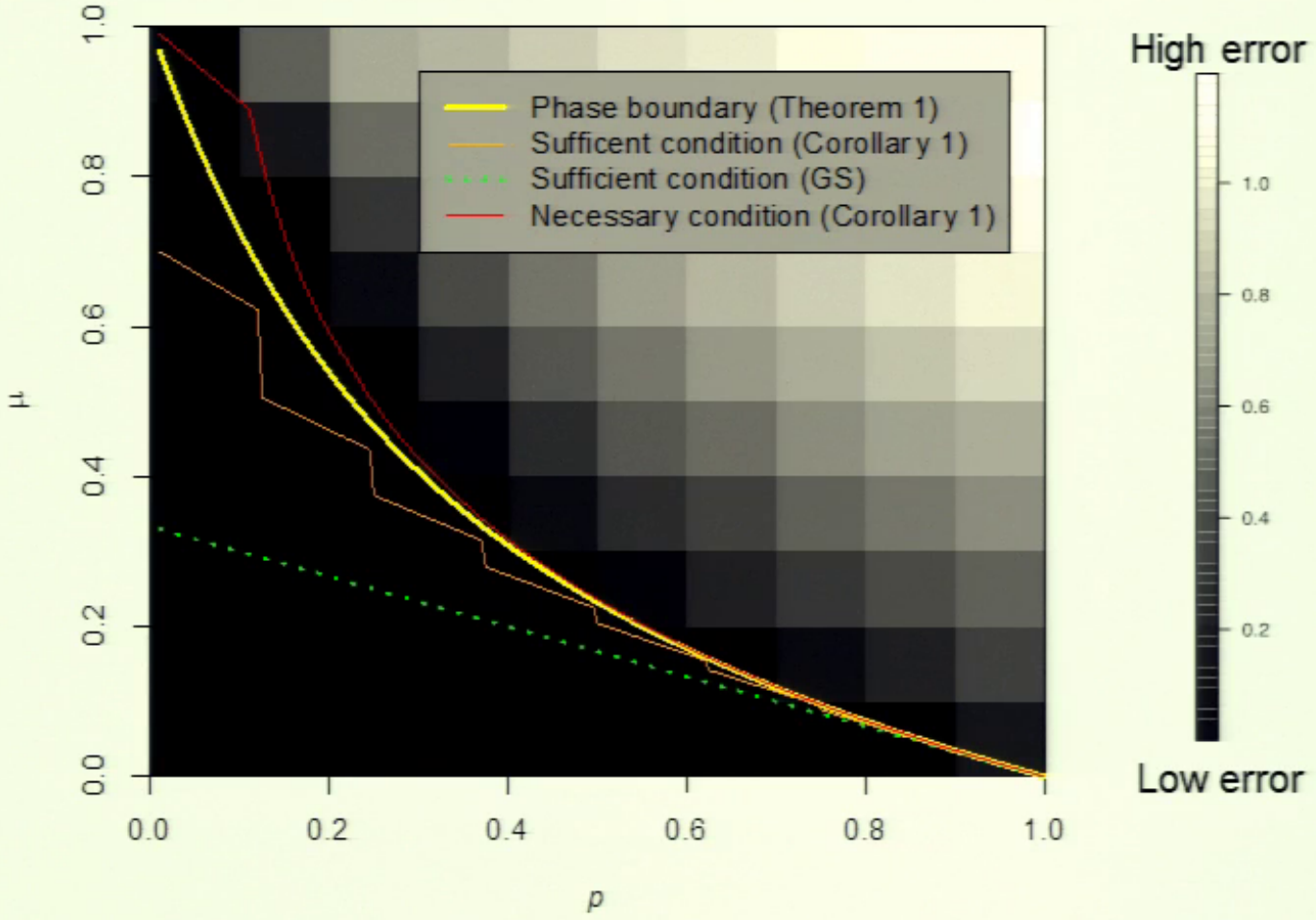
Local identifiability: If the data $\mathbf{x}_i$ is generated according to $\mathbf{x}_i = \mathbf{D}_0 \boldsymbol{\alpha}_i$, where $\boldsymbol{\alpha}_i$ is a random vector, under what conditions is $\mathbf{D}_0$ a local minimum of L_N?

Gribonval and Schnass 2010, Geng et al. 2011, Gribonval et al. 2014: **sufficient conditions** for local idenfitiability.

We find **a sufficient and almost necessary condition** for the case of square dictionary and noiseless observations.

Sufficient and almost necessary condition: an example

Constant inner product dictionary: $\mathbf{D}_0^T\mathbf{D}_0 = \mu\mathbf{1}\mathbf{1}^T + (1-\mu)\mathbf{I}$ for $\mu \in [0, 1]$.
Bernoulli-Gaussian model with sparsity level $p \in [0, 1]$.



Summary

1. NMF + Stability for localization of gene expressions

2. Preliminary correlation gene networks and on-going biological validation in known gap gene network.

3. Theoretical results on local identifiability of dictionary learning.

On-going and future work

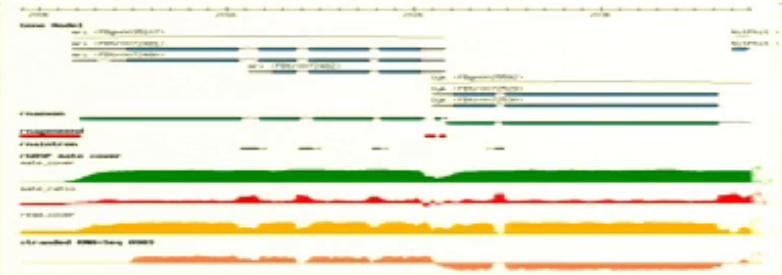
1. Further method developments for our data:
   other link prediction methods, later stage data, time-dynamics…

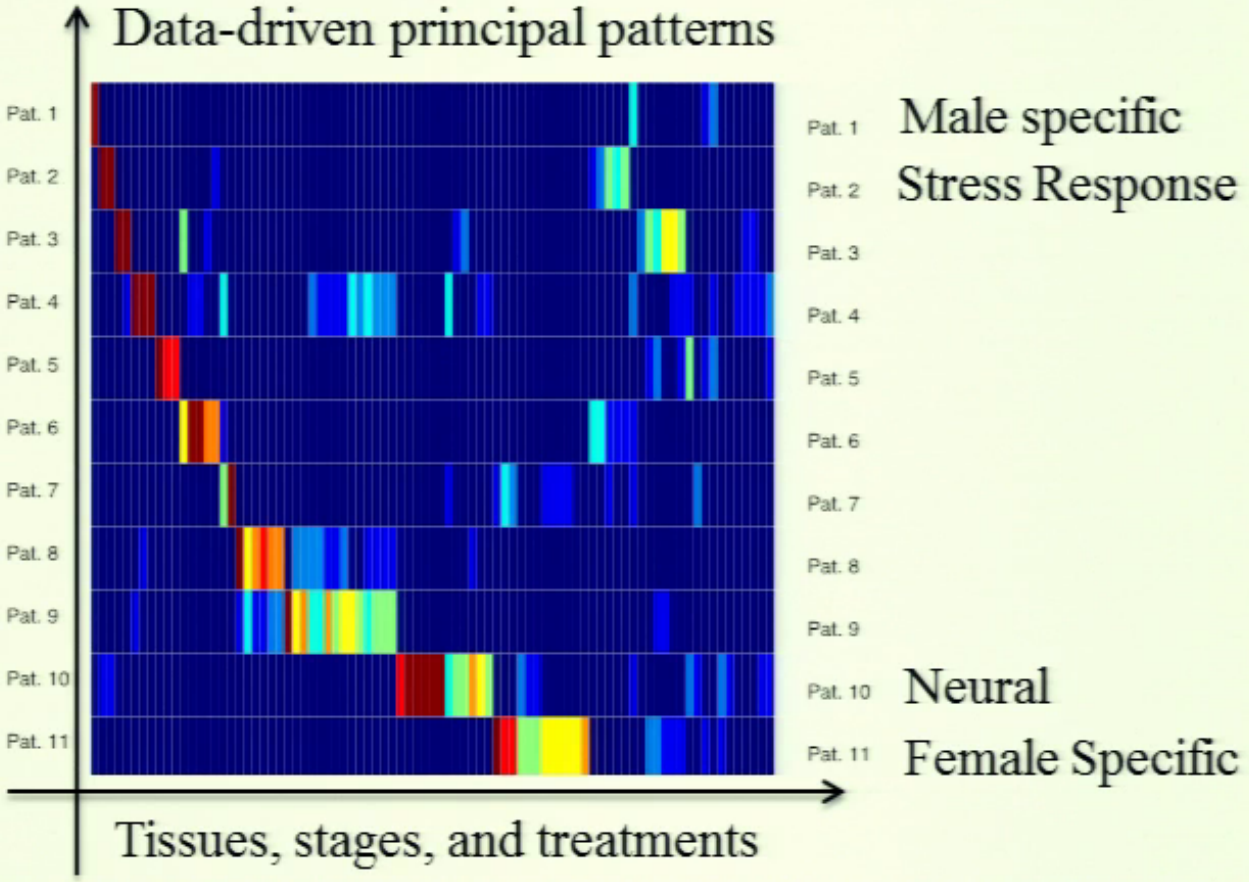
2. Other spatial genomics data

3. Software development: CISBA

One generalization: RNA-seq data (modENCODE)

Fruitfly RNA-seq data for different tissues, stages, and treatments



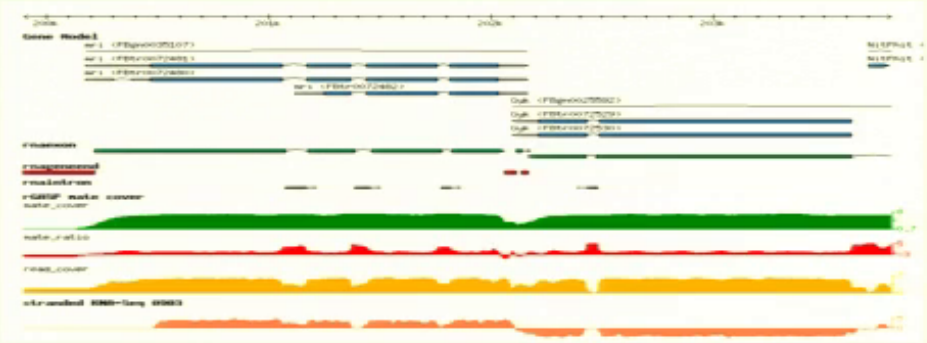


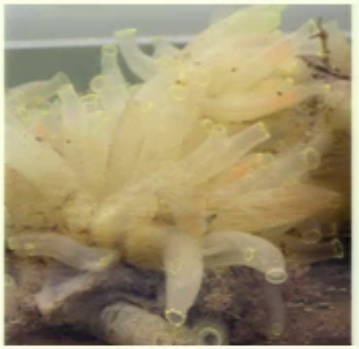
Software: CISBA and R-codes

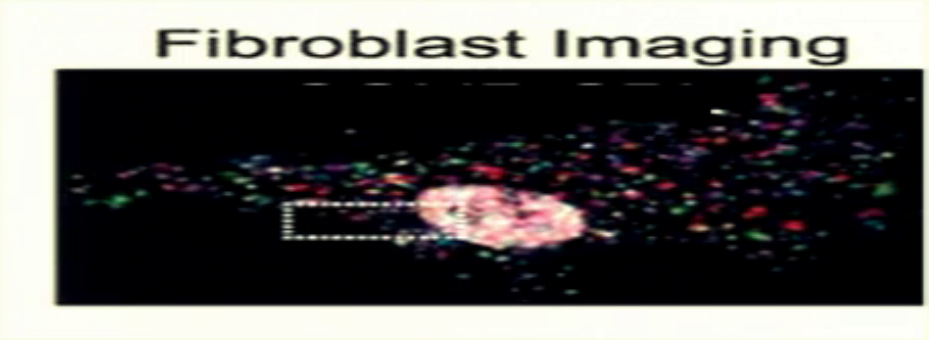
CISBA (cloud-based infrastructure for systems biology analytics)

R-codes for inferring correlation-based networks and other methods

Other data sources for CISBA and R-codes:







Advances expected through CISBA

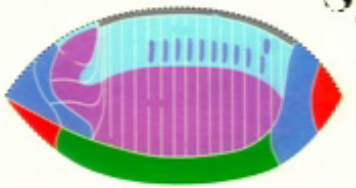
Statistics and Machine Learning

- Data perturbation methods
- Link prediction based on corr., partial corr. prior information (e.g. known gene networks)
- Theoretical analysis

Computer Science

- Data driven automation
- Data/metadata separation
- Open source tools

Systems Biology



Functional regions

Gene A ⟷ Gene B

Gene function and interactions

Advances expected through CISBA

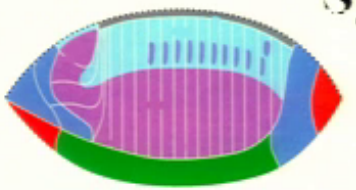
Statistics and Machine Learning

- Data perturbation methods
- Link prediction based on corr., partial corr. prior information (e.g. known gene networks)
- Theoretical analysis

Computer Science

- Data driven automation
- Data/metadata separation
- Open source tools

Systems Biology



Functional regions

Gene A Gene B

Gene function and interactions

Human genetics

Advances expected

Statistics and Machine Learning

Data perturbation methods

Decision theory for dependent hypotheses

"Prior" estimation based on text mining of literature

Computer Science

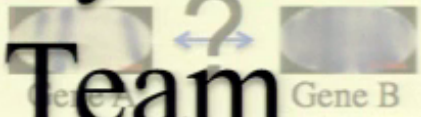
Data driven automation

Data/metadata separation

Open source tools

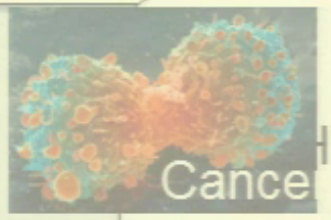
Systems Biology

Functional regions



Gene function and interactions

Data Science by Multidisciplinary and International Team



Human diseases