Unleashing the power of public gene expression data

Fei He (贺飞)
Research Associate
Kansas State University
plane83@gmail.com

Aug. 30, 2017

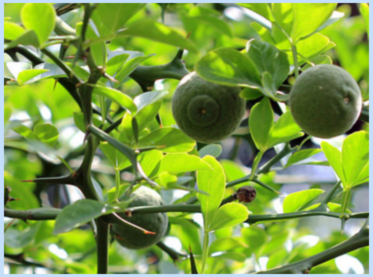
北美华人基因组学在线沙龙
Chinese Genomics Meet-up onllne

Outline

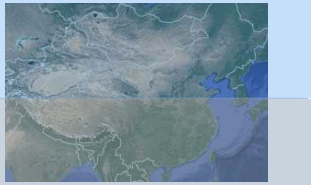
1. Background
2. The landscape of gene expression
3. Expression plasticity

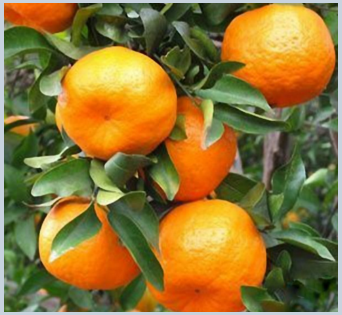
The same orange tree grows differently in different place

500 B. C.



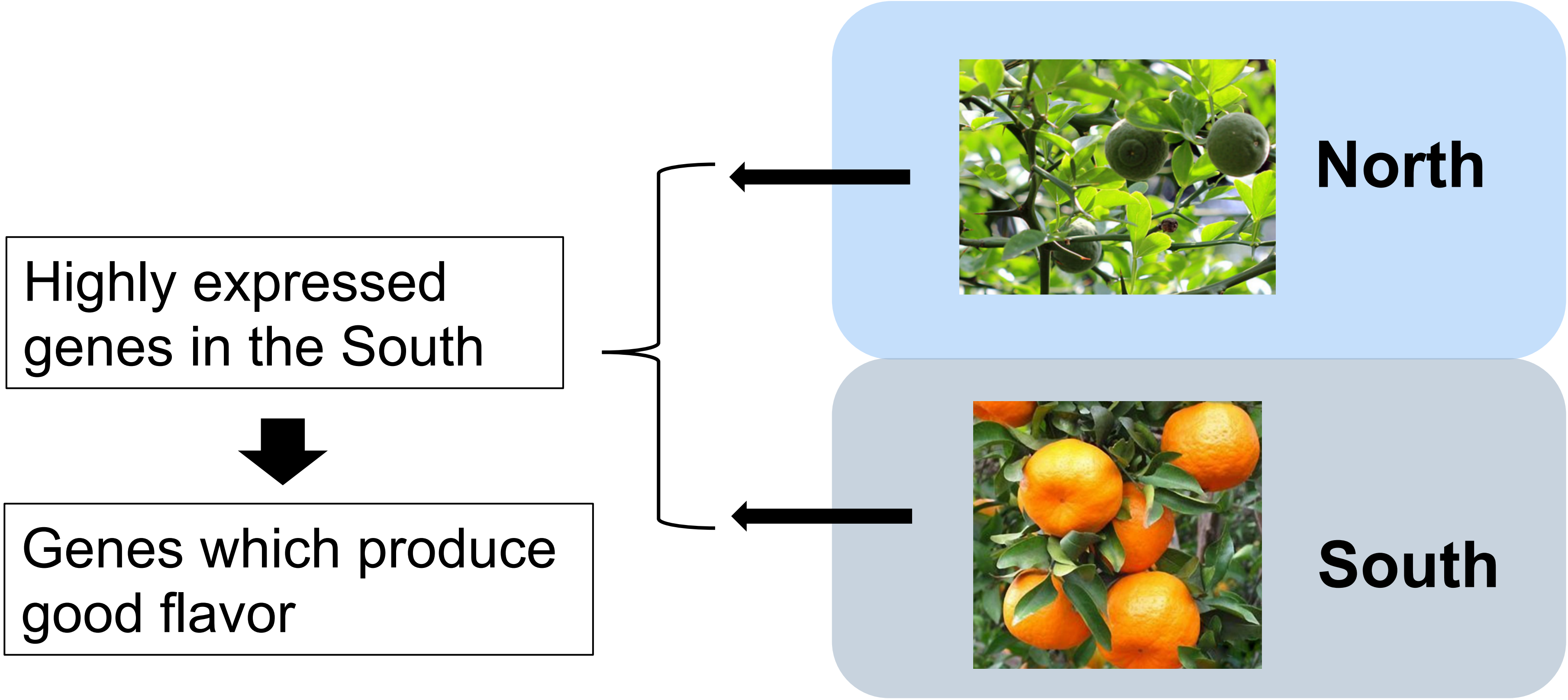
North





South

Molecular mechanism can be revealed by comparing gene expression



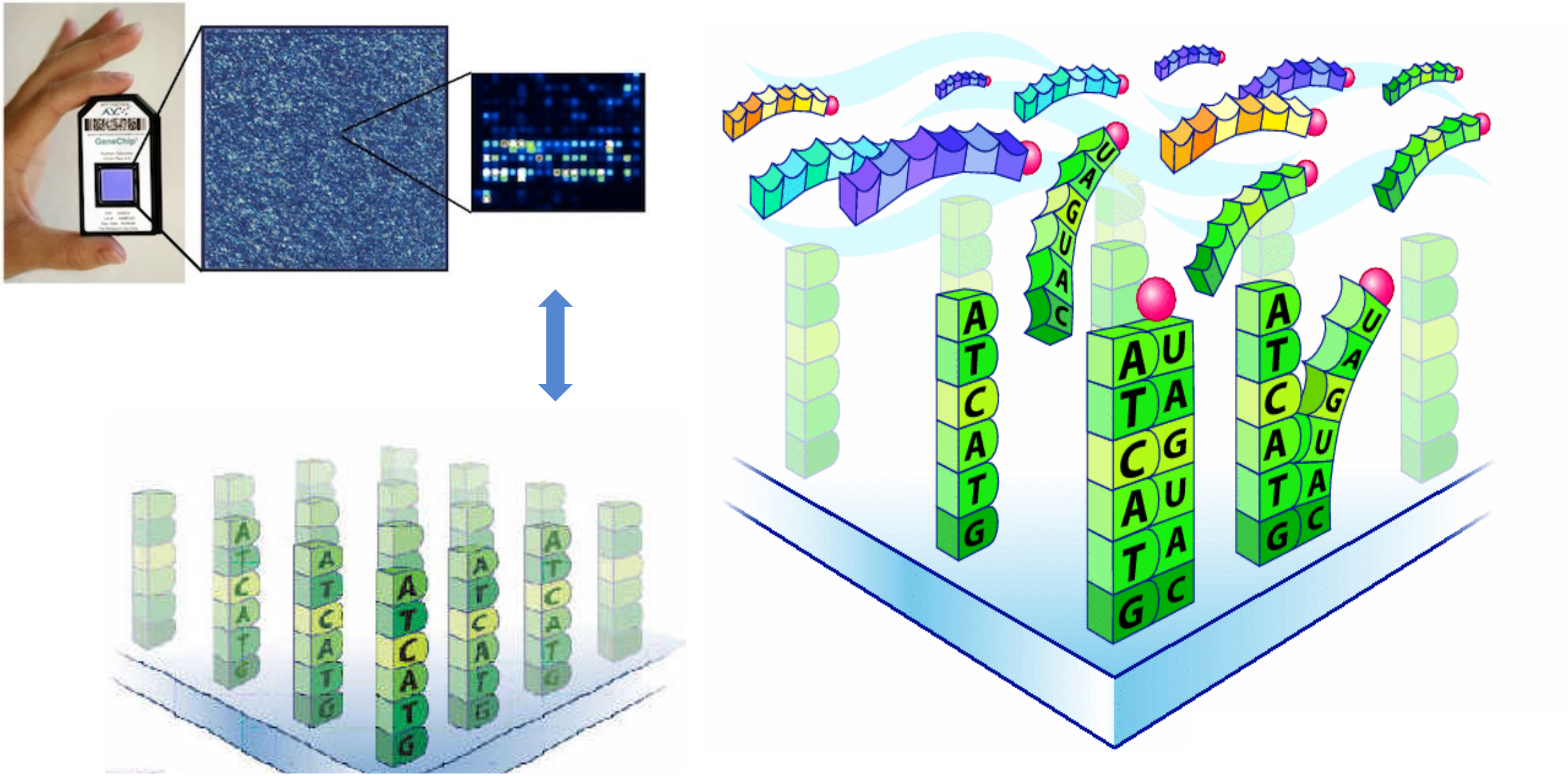
North

South

Highly expressed genes in the South

Genes which produce good flavor

Microarray can measure gene expression at a large-scale



A lot of microarray data are generated and stored in public databases

http://www.ncbi.nlm.nih.gov/geo/

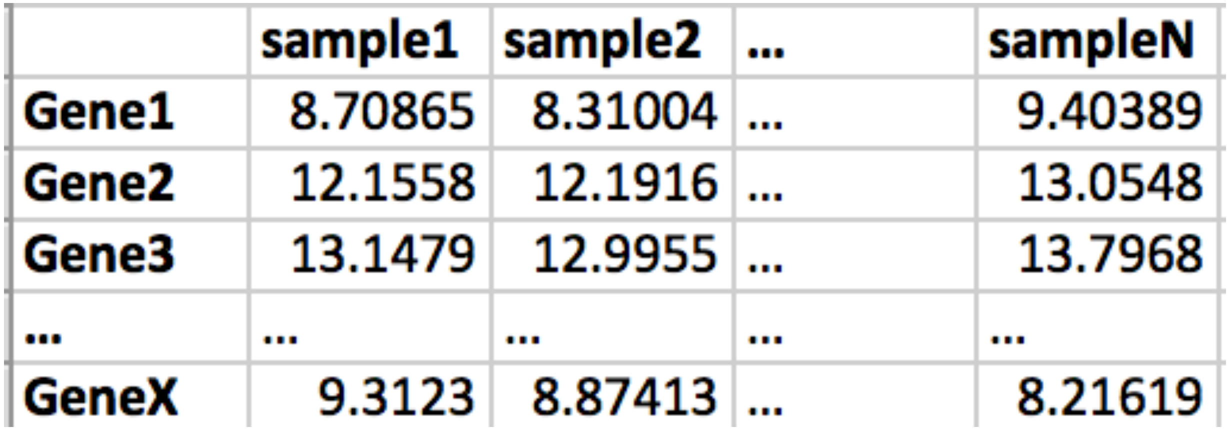
NCBI National Center for Biotechnology Information

GEO Gene Expression Omnibus

More than 1 million human microarray samples are stored in GEO.

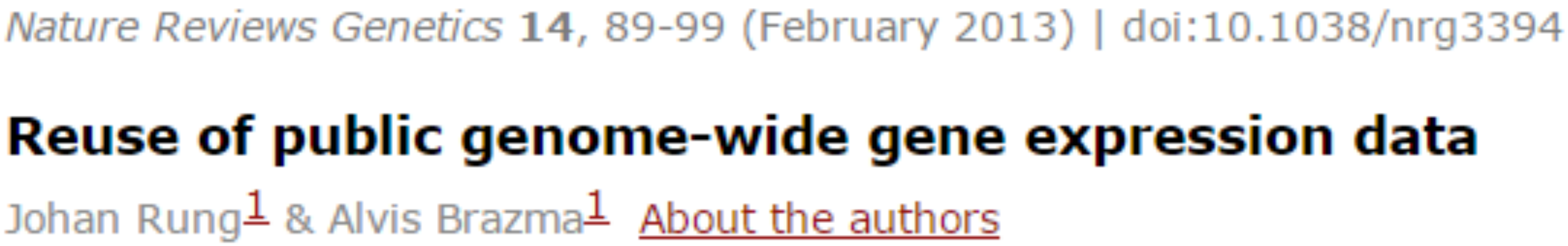
1,000,000 columns

20,000 rows



	sample1	sample2	...	sampleN
Gene1	8.70865	8.31004	...	9.40389
Gene2	12.1558	12.1916	...	13.0548
Gene3	13.1479	12.9955	...	13.7968
...	...	...	...	...
GeneX	9.3123	8.87413	...	8.21619

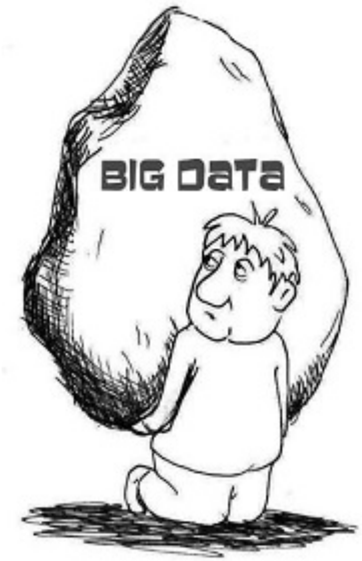
Making sense of public gene expression data is challenging



Nature Reviews Genetics **14**, 89-99 (February 2013) | doi:10.1038/nrg3394

Reuse of public genome-wide gene expression data

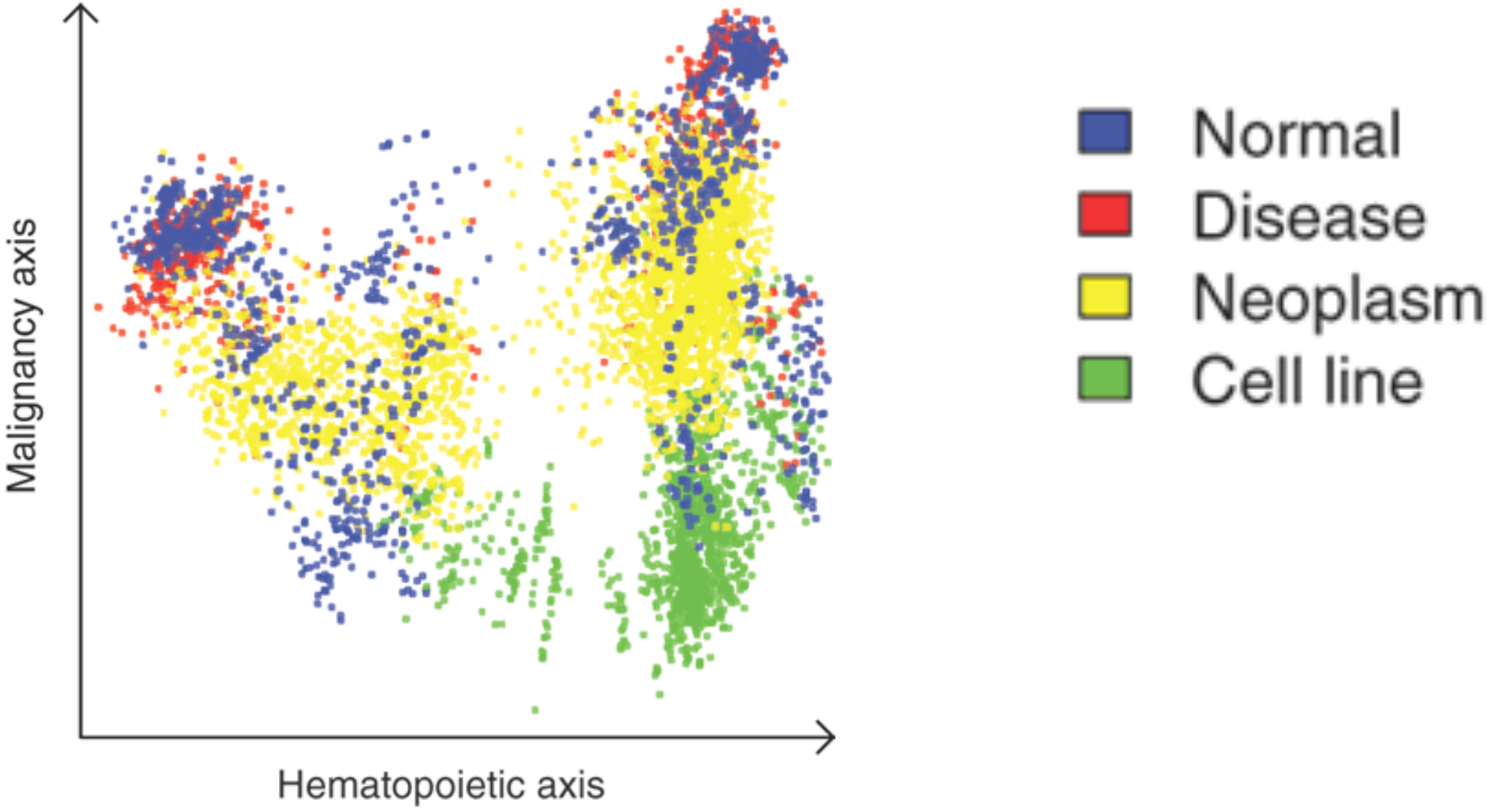
Johan Rung[1] & Alvis Brazma[1] About the authors



‘Reuse of public data can be very powerful, but there are many obstacles in data preparation and analysis and in the interpretation of the results.’

A human gene expression map

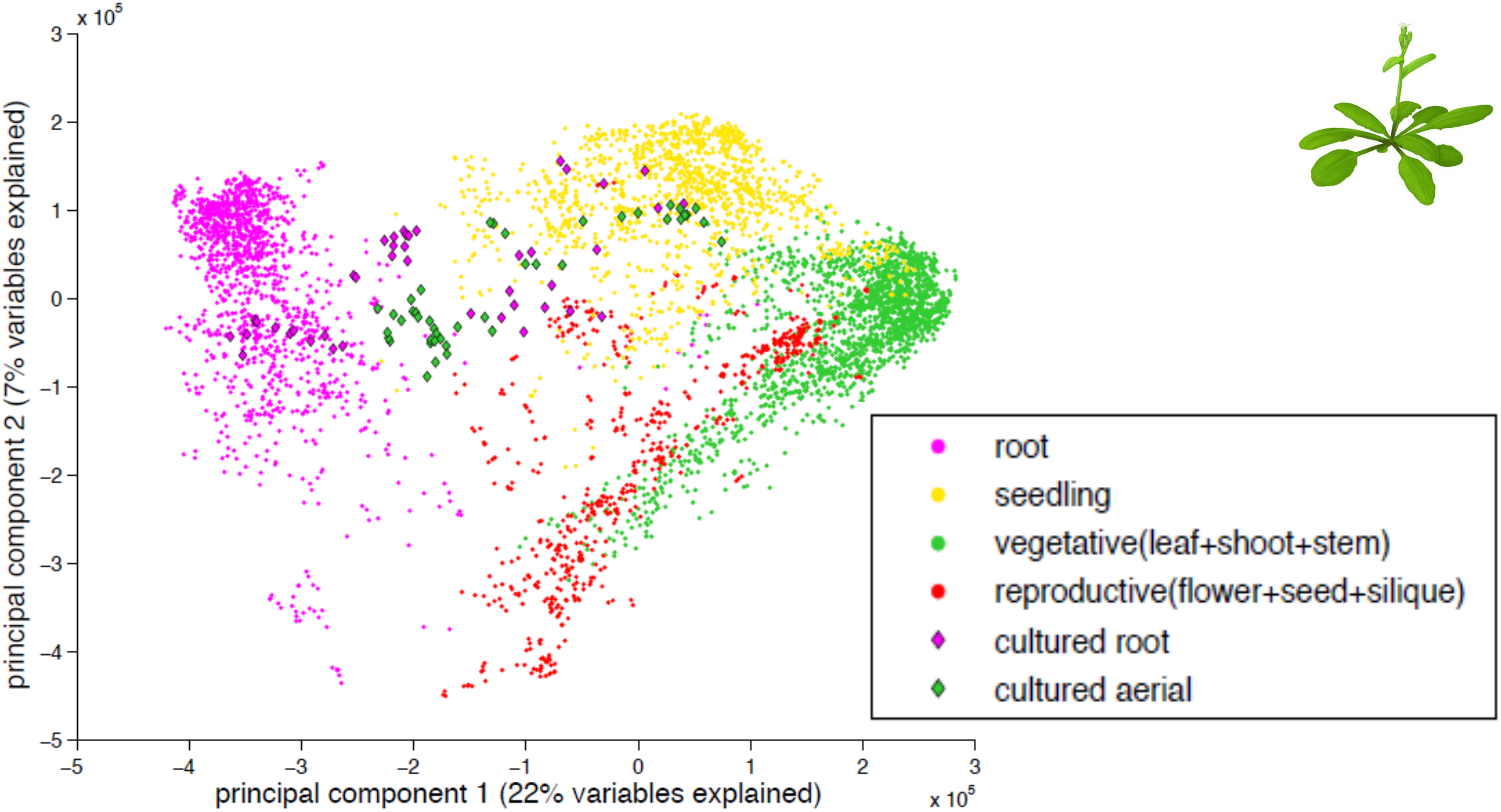
Lukk et al., 2010 Nature Biotechnology



Outline

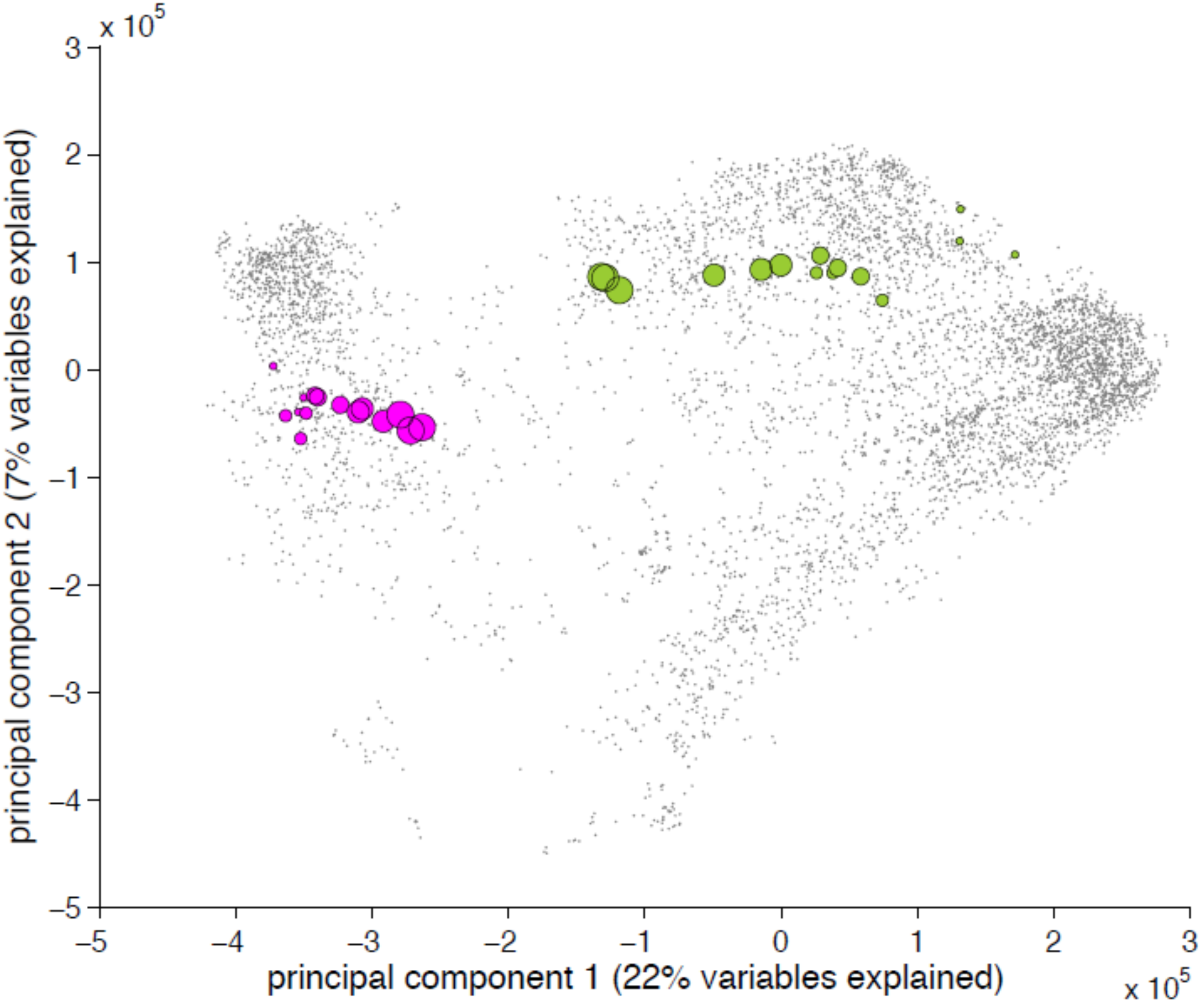
1. Background
2. The landscape of gene expression
3. Expression plasticity

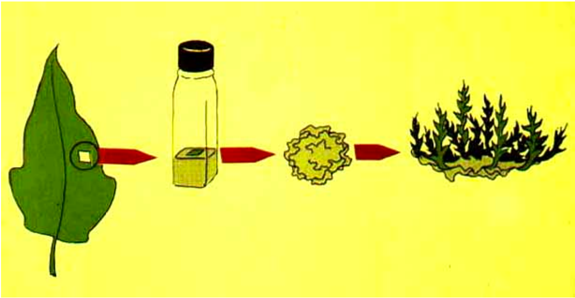
A gene expression landscape for Arabidopsis was reveled by meta-analysis of more than 6000 microarray samples



He et al., 2016

The position of an expression sample indicates its cellular state



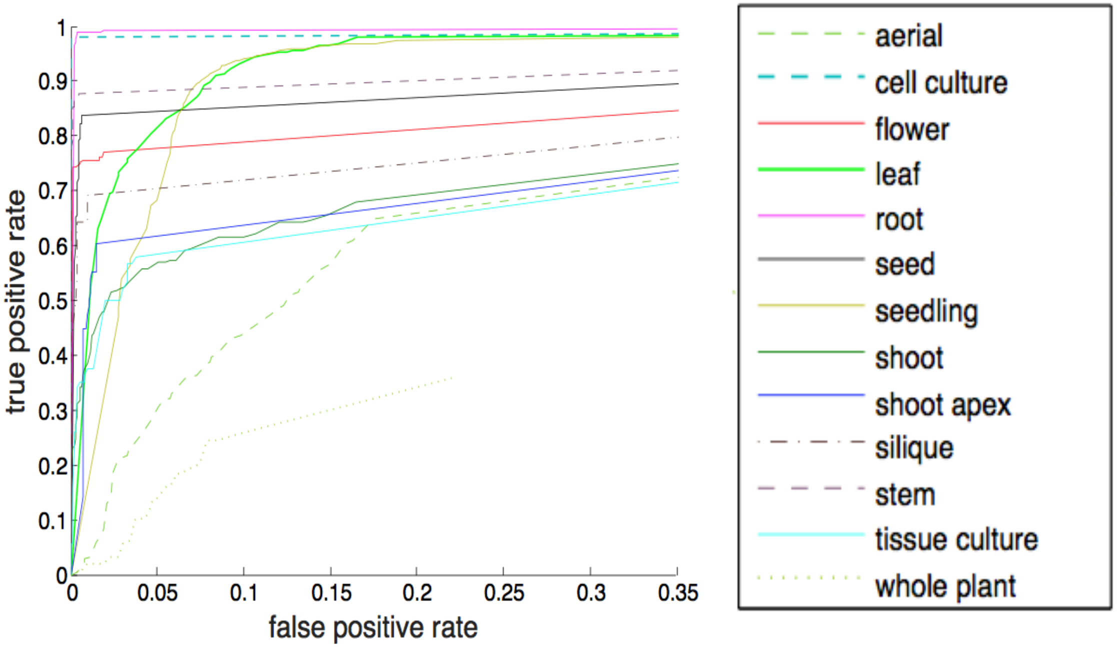


He et al., 2016

Automatic annotation of tissue type for Arabidopsis transcriptome data

We can actually predict mistakenly annotated samples in the NCBI GEO. For example, samples from GSE6826 were described as shoot, however those samples were predicted as root in our results.

Shinjae Yoo



Potential use cases:
Quality control for public data

He et al., 2016

Outline

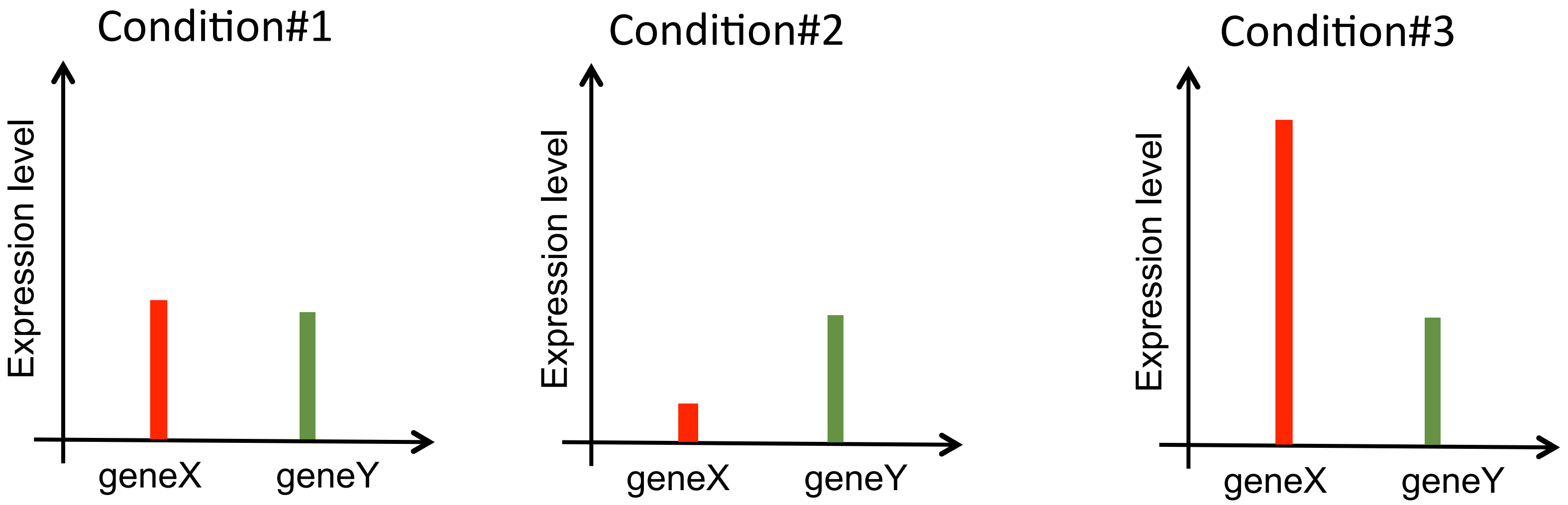
1. Background
2. The landscape of gene expression
3. Expression plasticity

Plasticity: gene's ability to response to different signals

My hypothesis: plasticity is an internal trait for genes.

Some genes tend to response to signals frequently.

Some genes tend to not response to any signals.

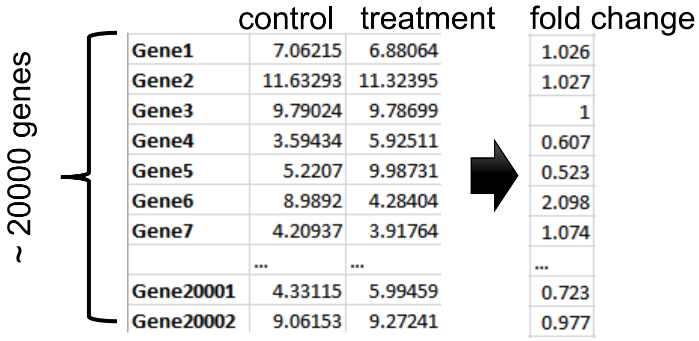


GeneX has higher plasticity than geneY

Infer gene expression plasticity using public expression data

$$\boxed{\text{plasticity} = \log_2 \frac{1}{k} \sum_{i=0}^{k} (\log_2 r_i)^2}$$

k is the number of comparisons.
r is the fold change for a gene under a certain perturbation.



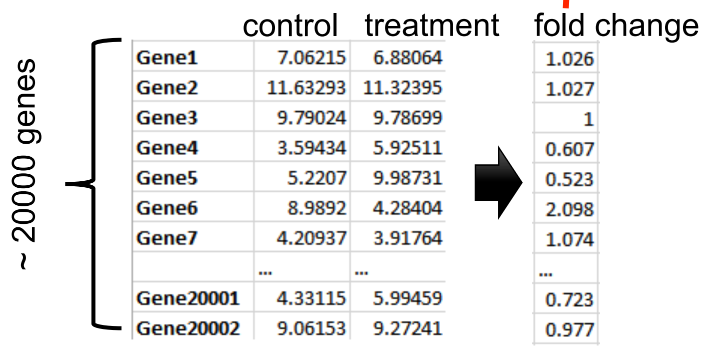
~ 20000 genes

	control	treatment	fold change
Gene1	7.06215	6.88064	1.026
Gene2	11.63293	11.32395	1.027
Gene3	9.79024	9.78699	1
Gene4	3.59434	5.92511	0.607
Gene5	5.2207	9.98731	0.523
Gene6	8.9892	4.28404	2.098
Gene7	4.20937	3.91764	1.074
	...	...	...
Gene20001	4.33115	5.99459	0.723
Gene20002	9.06153	9.27241	0.977

Infer gene expression plasticity using public expression data

$$\text{plasticity} = \log_2 \frac{1}{k}\sum_{i=0}^{k}(\log_2 \boxed{r_i})^2$$

k is the number of comparisons.
r is the fold change for a gene under a certain perturbation.

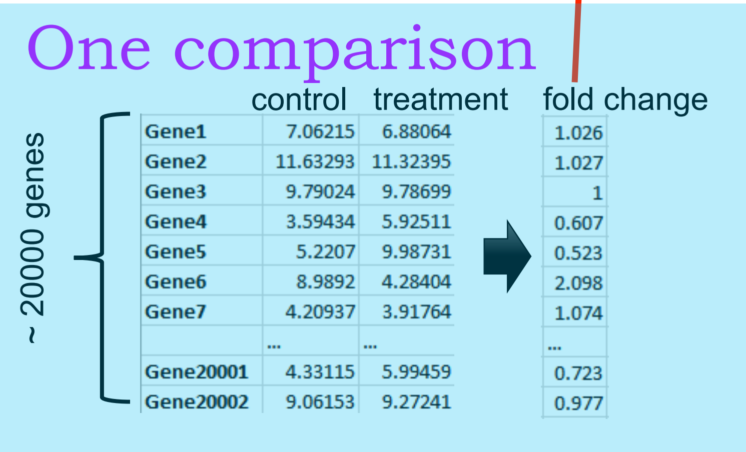

~ 20000 genes

	control	treatment	fold change
Gene1	7.06215	6.88064	1.026
Gene2	11.63293	11.32395	1.027
Gene3	9.79024	9.78699	1
Gene4	3.59434	5.92511	0.607
Gene5	5.2207	9.98731	0.523
Gene6	8.9892	4.28404	2.098
Gene7	4.20937	3.91764	1.074
	...	...	...
Gene20001	4.33115	5.99459	0.723
Gene20002	9.06153	9.27241	0.977

Infer gene expression plasticity using public expression data

$$\text{plasticity} = \log_2 \frac{1}{k} \sum_{i=0}^{k} (\log_2 r_i)^2$$

k is the number of comparisons.
r is the fold change for a gene under a certain perturbation.



One comparison

~ 20000 genes

	control	treatment	fold change
Gene1	7.06215	6.88064	1.026
Gene2	11.63293	11.32395	1.027
Gene3	9.79024	9.78699	1
Gene4	3.59434	5.92511	0.607
Gene5	5.2207	9.98731	0.523
Gene6	8.9892	4.28404	2.098
Gene7	4.20937	3.91764	1.074
	...	...	...
Gene20001	4.33115	5.99459	0.723
Gene20002	9.06153	9.27241	0.977

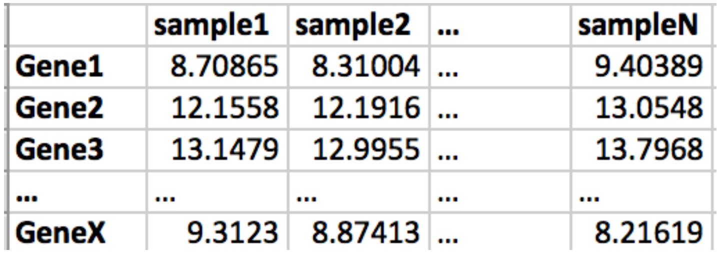
We collected >1000 such comparisons from public repositories for Arabidopsis

Now, we calculated a number to represent the plasticity of a gene The larger this value, the more likely for a gene to response to environmental perturbations.

‘Cutting Big Data Down to a Usable Size’

>10,000 columns

20000 rows

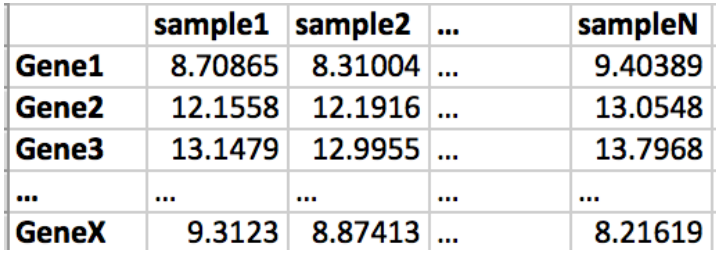


	sample1	sample2	...	sampleN
Gene1	8.70865	8.31004	...	9.40389
Gene2	12.1558	12.1916	...	13.0548
Gene3	13.1479	12.9955	...	13.7968
...	...	...	...	...
GeneX	9.3123	8.87413	...	8.21619

‘Cutting Big Data Down to a Usable Size’

>10,000 columns

20000 rows



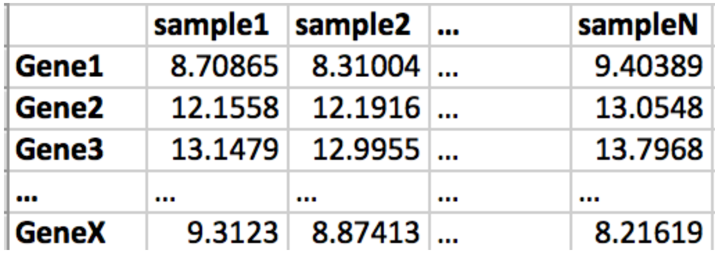
	sample1	sample2	...	sampleN
Gene1	8.70865	8.31004	...	9.40389
Gene2	12.1558	12.1916	...	13.0548
Gene3	13.1479	12.9955	...	13.7968
...	...	...	...	...
GeneX	9.3123	8.87413	...	8.21619

Gene ID	Plasticity
Gene1	0.02
Gene2	0.09
Gene3	0.87
…	…
GeneX	0.31

‘Cutting Big Data Down to a Usable Size’

>10,000 columns

20000 rows



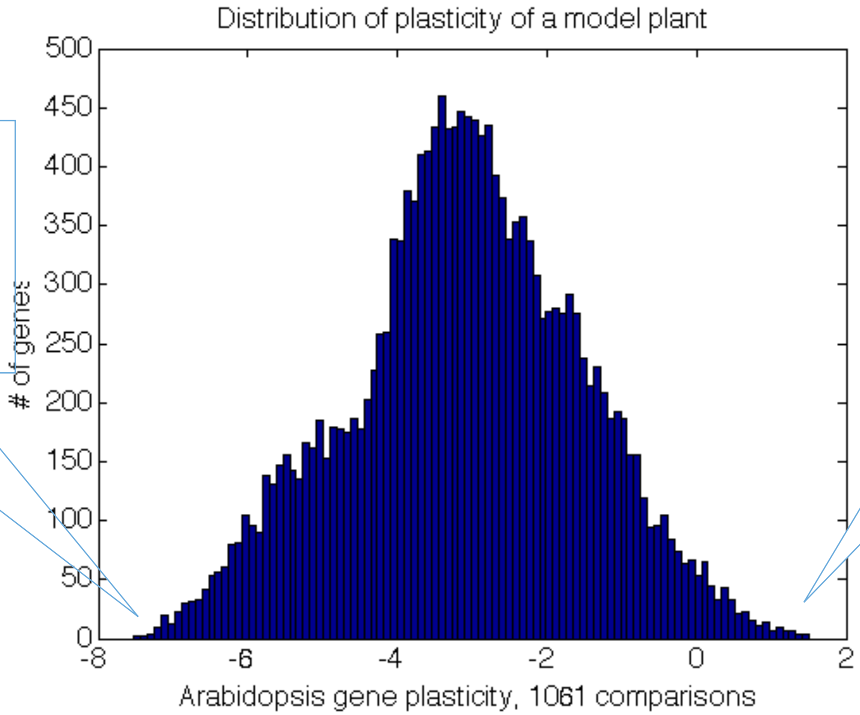
	sample1	sample2	...	sampleN
Gene1	8.70865	8.31004	...	9.40389
Gene2	12.1558	12.1916	...	13.0548
Gene3	13.1479	12.9955	...	13.7968
...	...	...	...	...
GeneX	9.3123	8.87413	...	8.21619

Gene ID	Plasticity
Gene1	0.02
Gene2	0.09
Gene3	0.87
…	…
GeneX	0.31

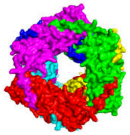
What kind of genes have the largest expression plasticity?

Expression plasticity might be an attribute for genes in Arabidopsis

Homeodomain protein



Heat shock protein



Protect other proteins under stress environment

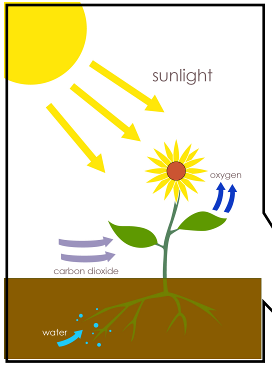
normal

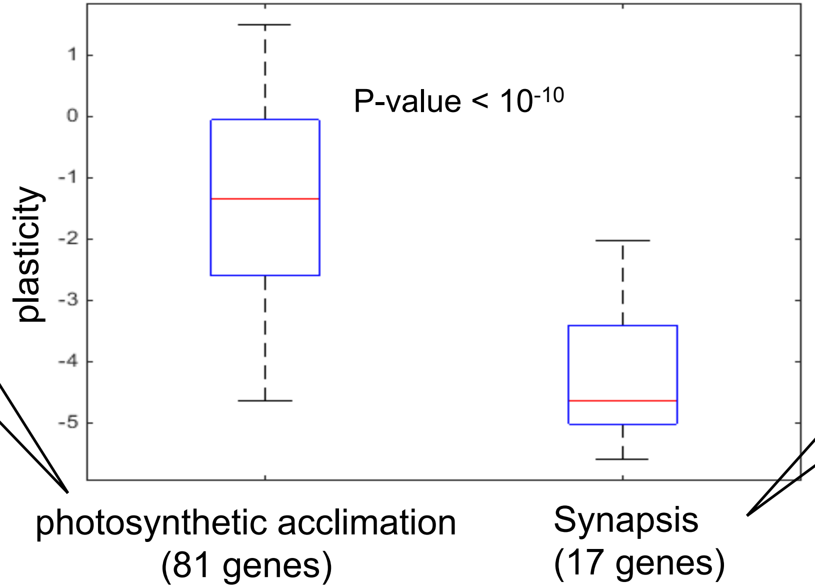
Hox mutant

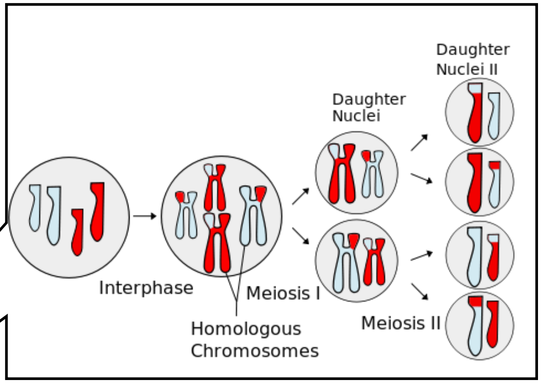
It looks like expression plasticity is corresponding to gene function.

He et al., unpublished

Different functional groups have different expression plasticity



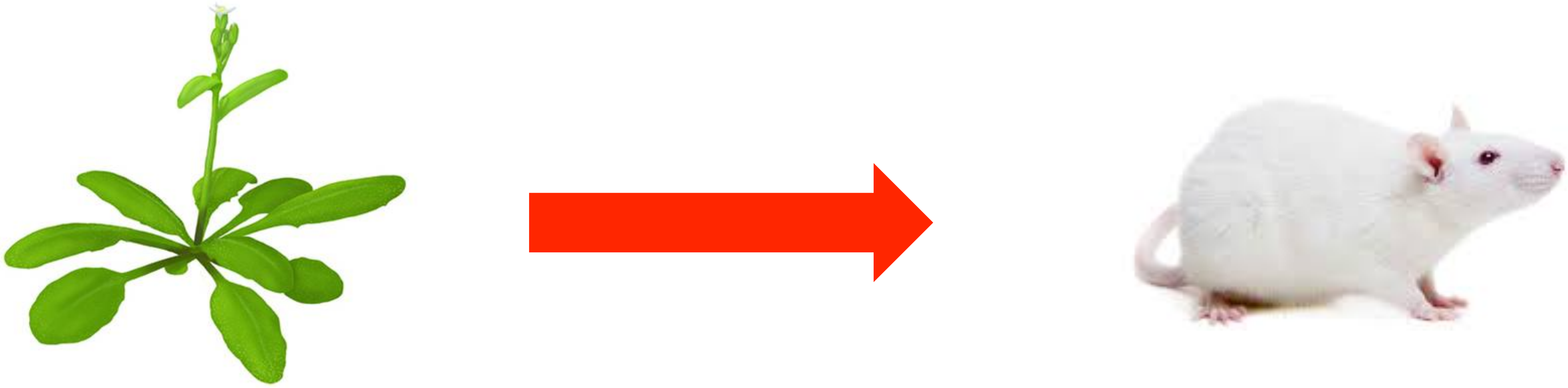




- Reproductive systems may need to be very stable under different environments in order to maintain the genetic stability.
- Photosynthetic system may need to be highly responsive to environmental signals in order to generate energy

He et al., unpublished

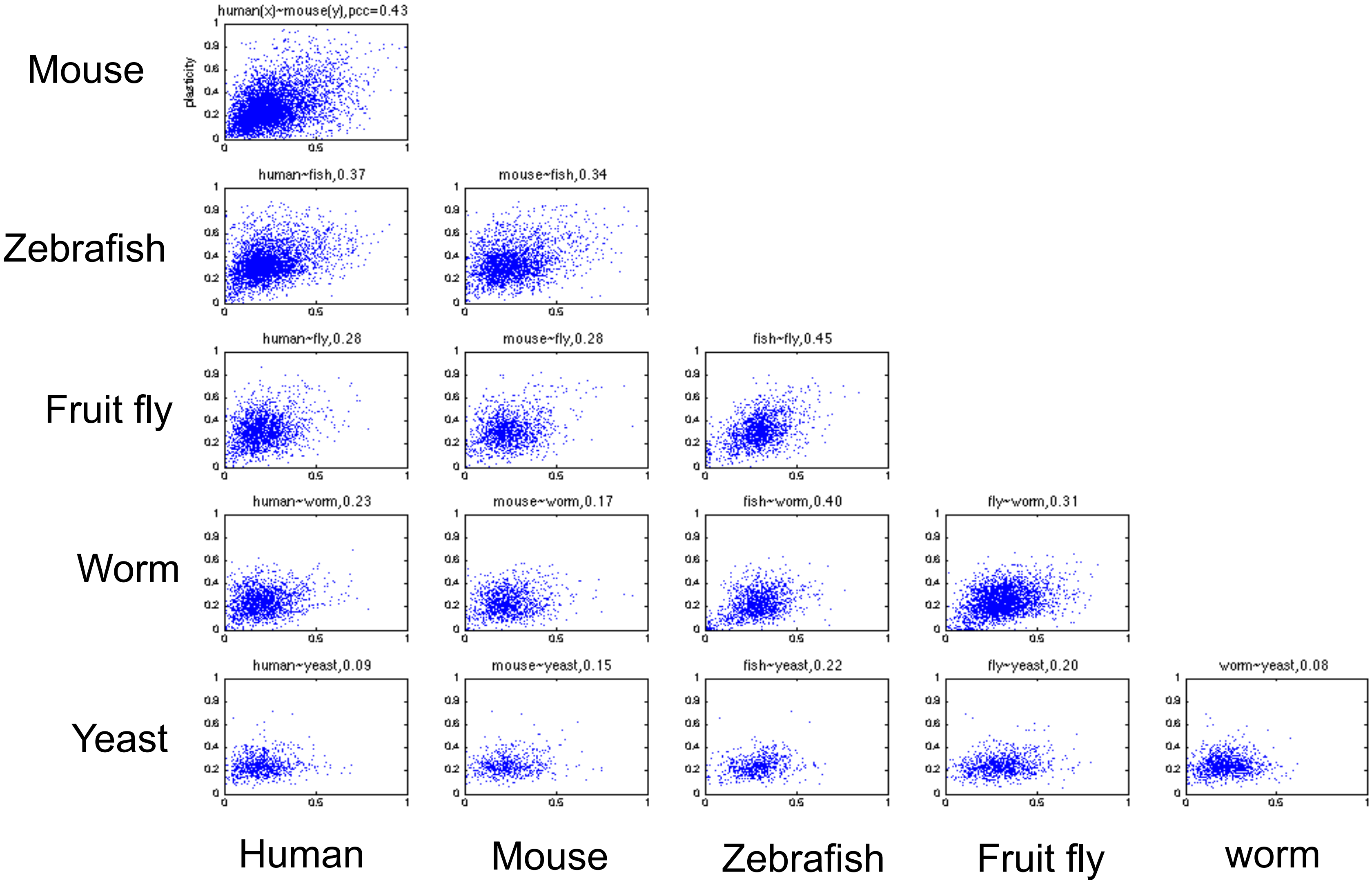
Why not calculating the expression plasticity for other model organisms?



Expression plasticity is an evolvable trait

Dot represents 1vs1 ortholog

Each species contains 1000~6000 samples.



He et al., unpublished

Acknowledgements

KBase
PREDICTIVE BIOLOGY

BROOKHAVEN NATIONAL LABORATORY
a passion for discovery

U.S. DEPARTMENT OF ENERGY | Office of Science

ILLINOIS

Biological and Environmental Research (BER)

Sergei Maslov@University of Illinois Urbana-Champaign

Zhuo Du@Institute of Genetics and Developmental Biology, Beijing, China