Gene Expression State Space Models and Cell Fate Transitions

John Quackenbush

Cancer Bioinformatics Workshop

2 September 2010

The Computational Biology and Functional Genomics Laboratory

at the Dana-Farber Cancer Institute and Harvard School of Public Health

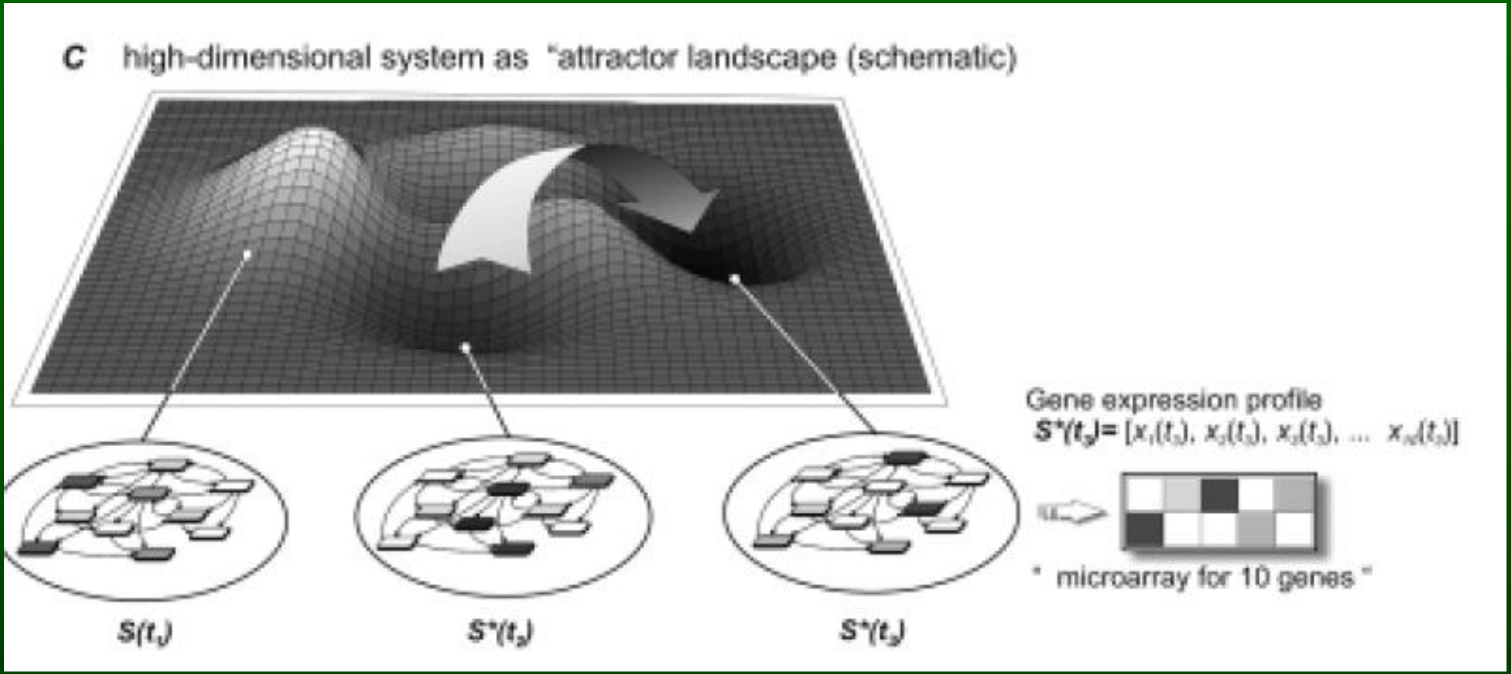
Phenomenology and Models

- Ultimately, we look to develop a theory that describes the interactions that drive biological systems
- The embodiment of the resulting theory should be a model describing the interactions we are seeking to understand
- Phenomenology, or phenomenological models, describe a body of knowledge that relates empirical observations of phenomena to each other, in a way which is consistent with fundamental theory, but is not directly derived from theory
- The question is not “Is this model right?” Rather, the question is “Is the model useful?”

State Space Models of Gene Expression

Jess Mar

Cells Converge to Attractive States



Stuart Kauffman presented the idea of a gene expression landscape with attractors

•~250 stable cell types each represent attractors

•Cells can be "pushed" or induced to converge to an attractor.

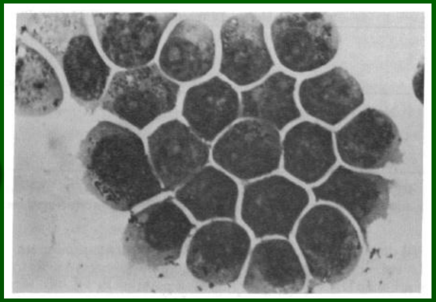
•Once in the attractor, a cell is robust to small perturbations.

Jess Mar

Differentiation of Promyelocytes into Neutrophil-Like Cells

Promyeloctyes

(HL-60 Cell Line)



RA used in differentiation therapy for acute promyelocytic leukemia.

Time 0

Affymetrix GeneChip

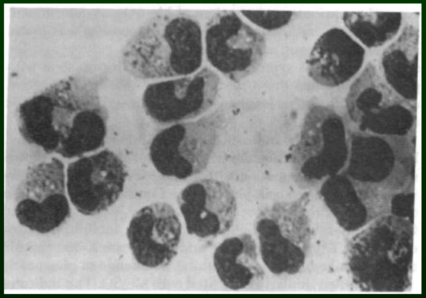
~6 days

Day 7

Dimethyl Sulfoxide (DMSO)

All-Trans Retinoic Acid (ATRA)

Neutrophil-like Cells



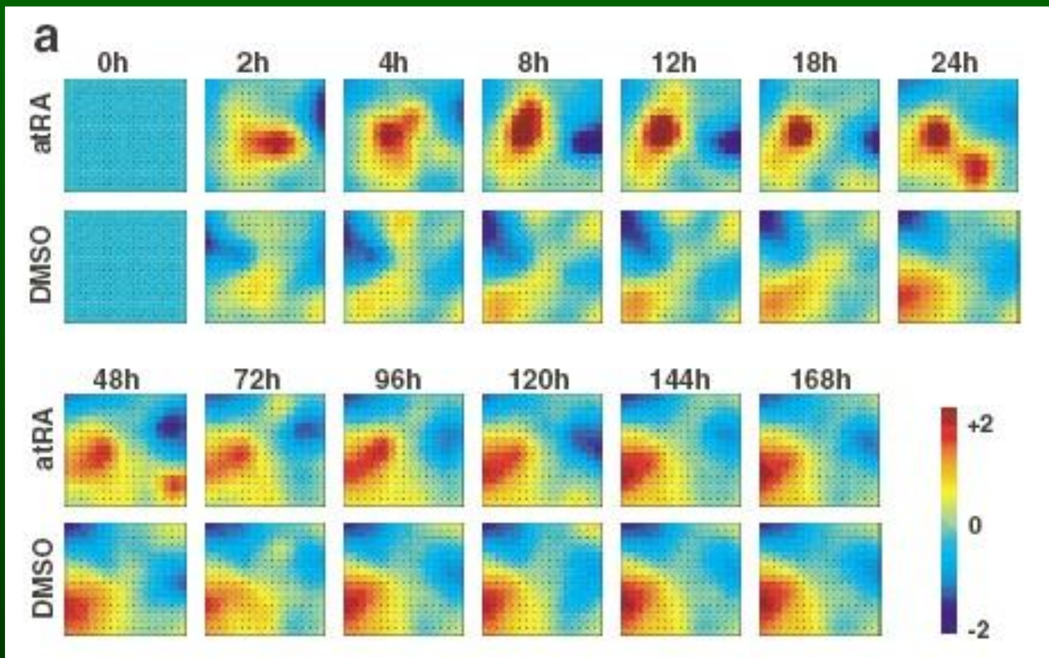
Combined with chemotherapy, complete remission rates as high as 90-95% can be achieved.

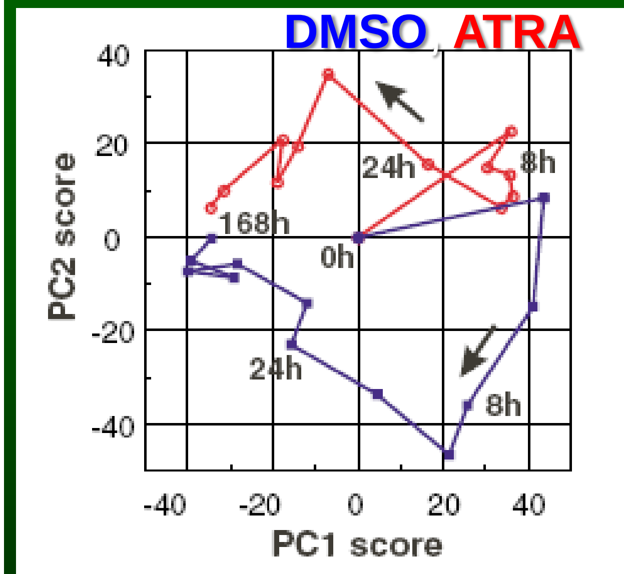
Collins et al. *PNAS* 1978

Jess Mar

Huang et al. *PRL* 2005

Cells Display Divergent Trajectories That Eventually Converge as they Differentiate





Graphical representation of the results from a Self-Organizing Map clustering.

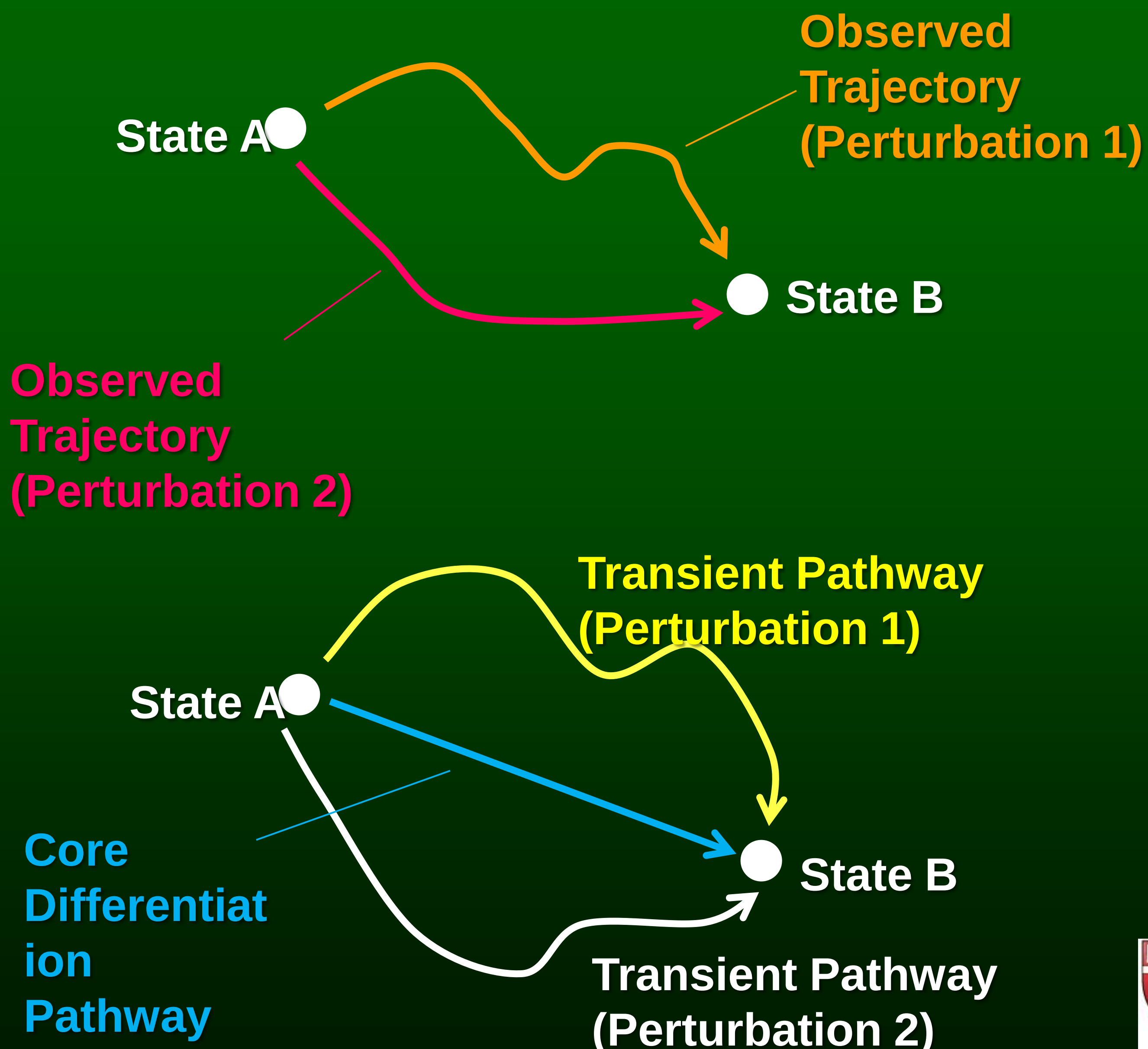
Expression data from a single sample (time point) clustered according to a grid.

What factors drive this divergent-then-convergent behavior?

Huang et al. *PRL* 2005

Jess Mar

Our Hypothesis



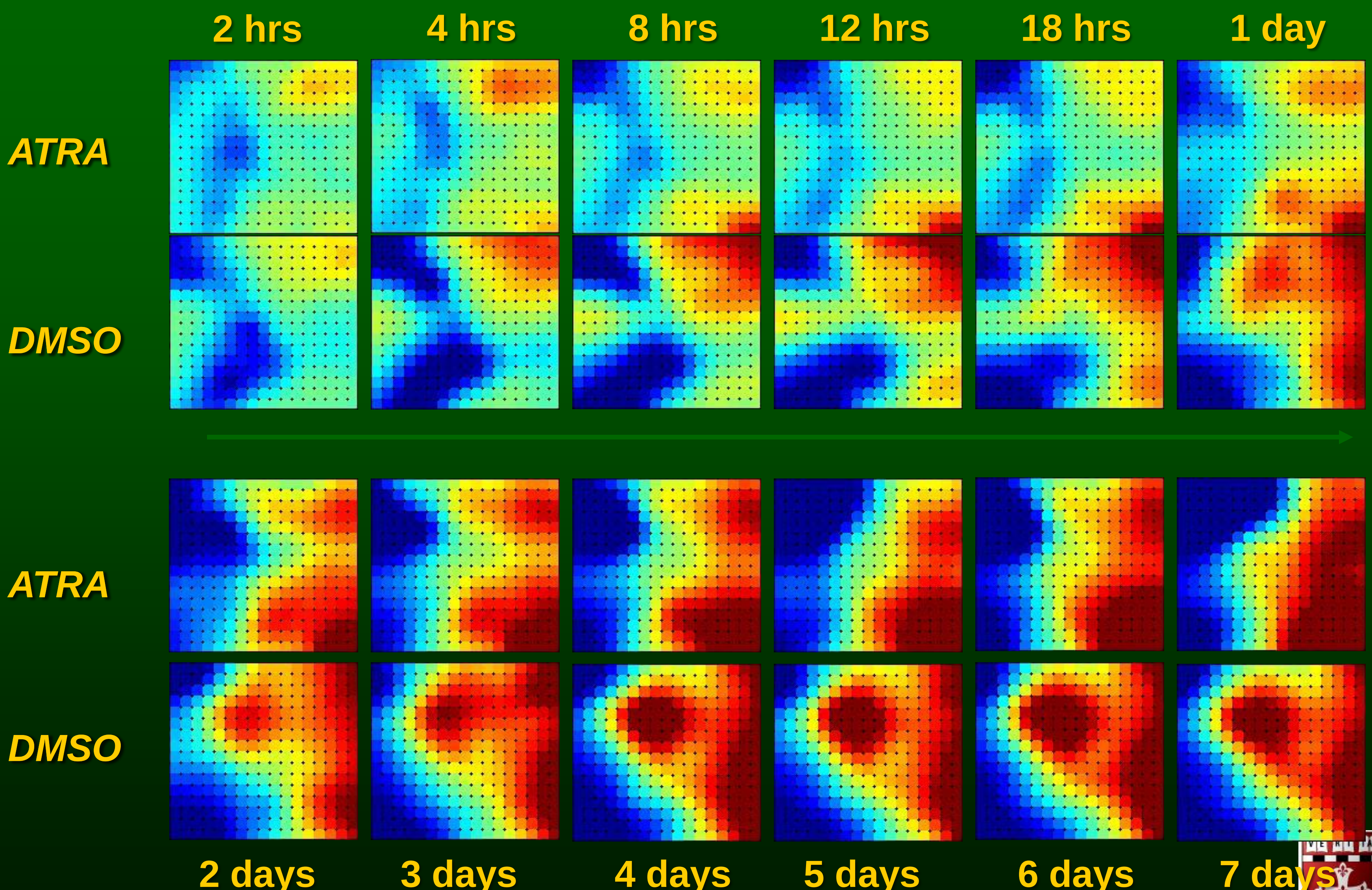
Jess Mar

Functional Enrichment Analysis

Enriched GO functional classes in each group.

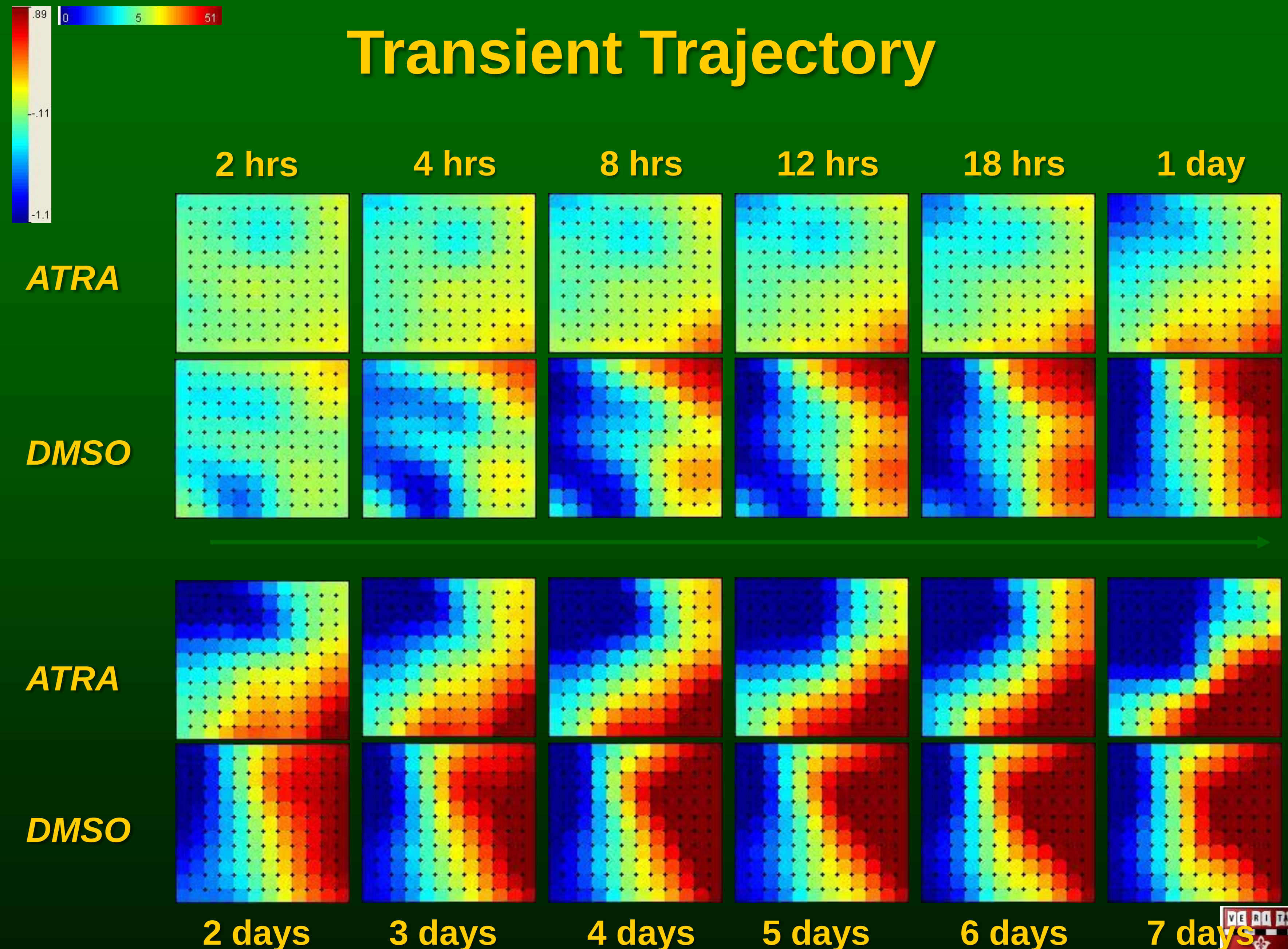
Core Gene Group	RNA metabolic process Transcription RNA biosynthetic process Steroid biosynthetic process Transcription, DNA-dependent Regulation of transcription, DNA-dependent Regulation of transcription Nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
Transient Gene Group	Defense response Response to external stimulus Response to wounding Inflammatory response Signal transduction Response to stimulus Cell communication

Observed Trajectory

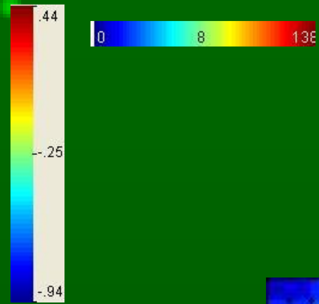


Jess Mar

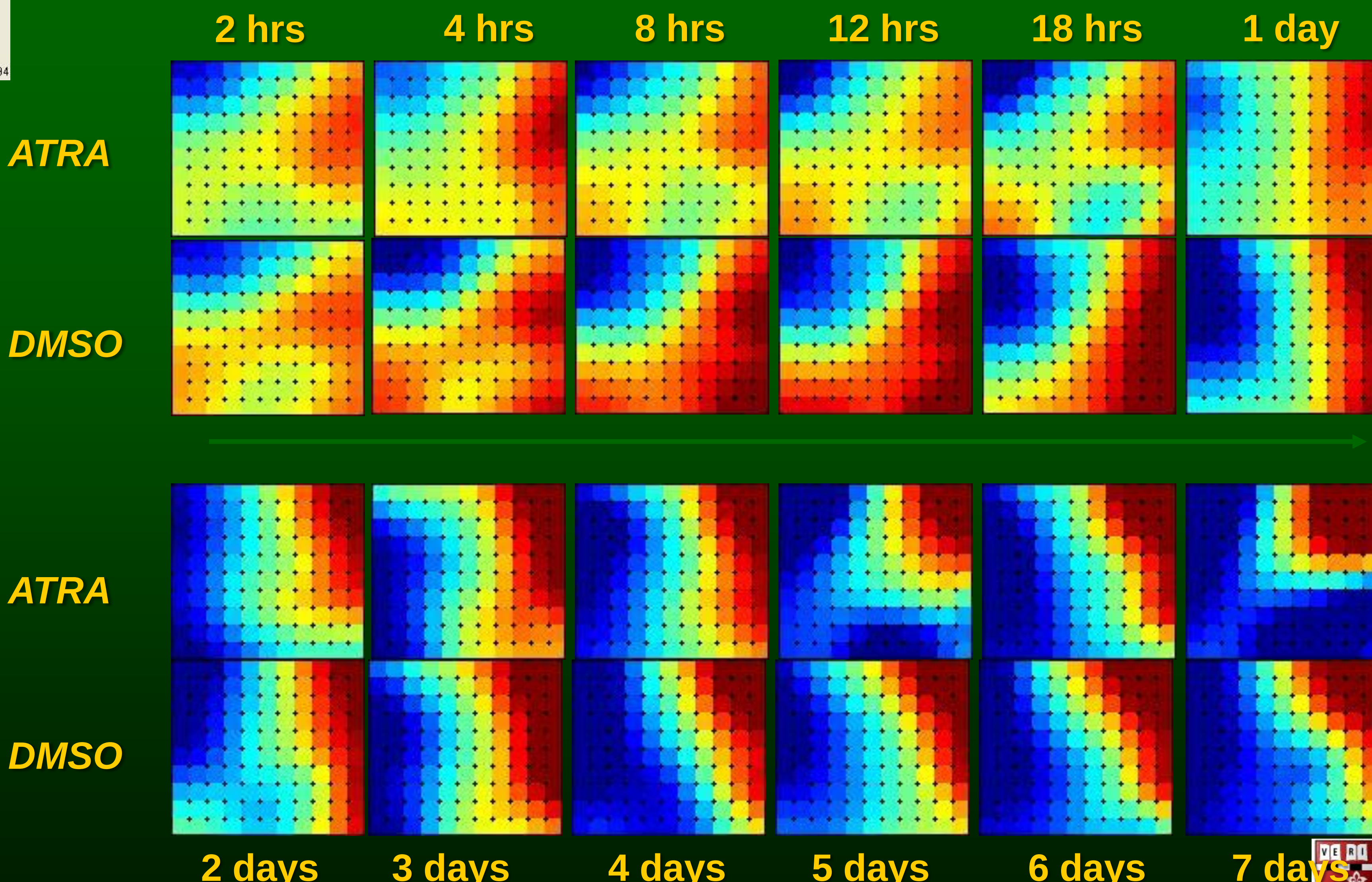
Transient Trajectory



Jess Mar



Core Trajectory

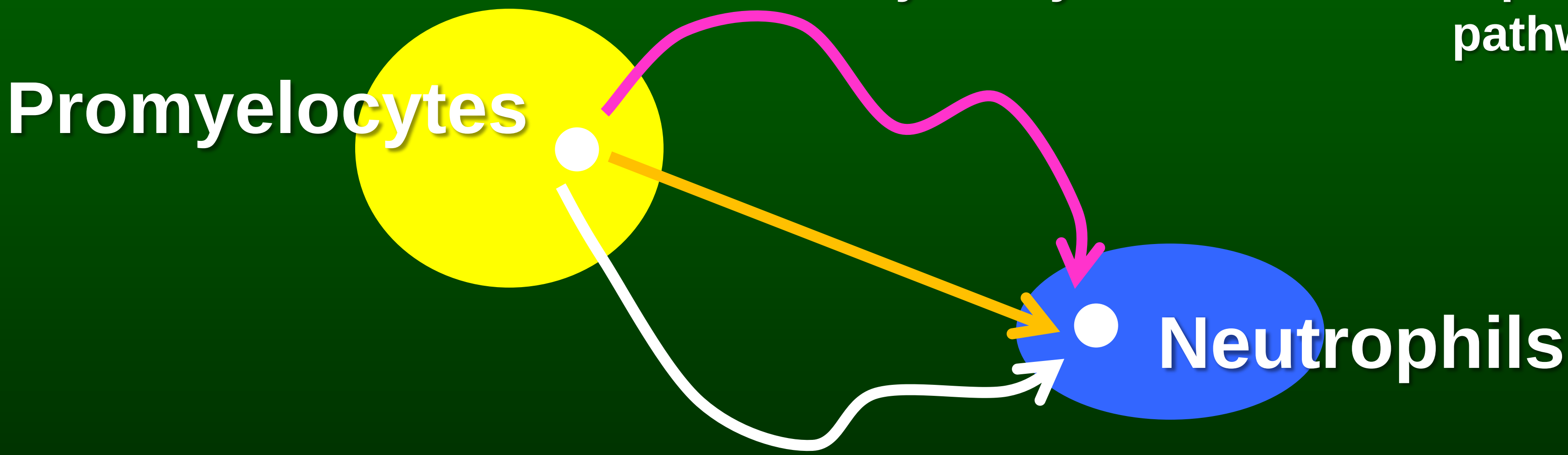


Jess Mar

What Have We Learned?

Transition from one state to another is driven by two classes of genes:

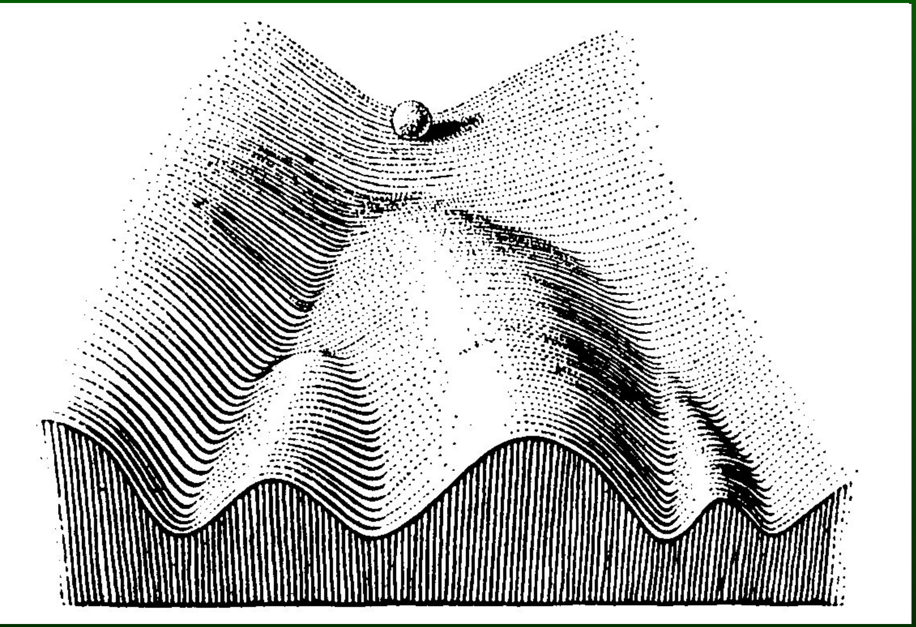
Core genes whose sustained expression carry the system down developmental pathways.



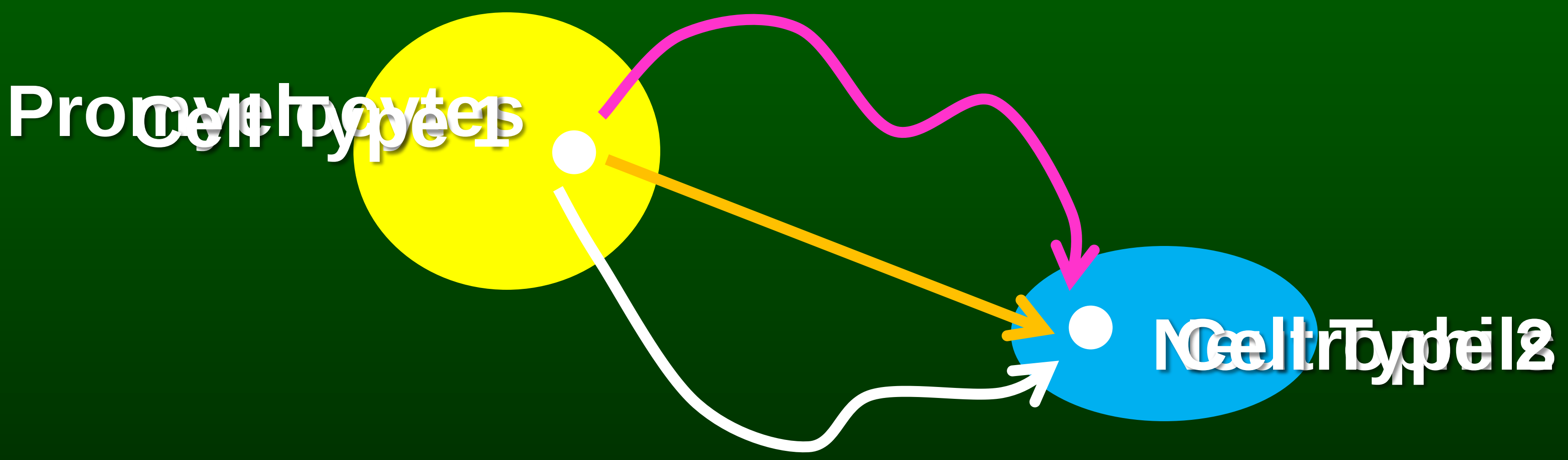
Transient genes that fire initially in response to a stimulus, but whose expression decays over time. These are instrumental in kicking the system into the transition.

Waddington's Hypothesis

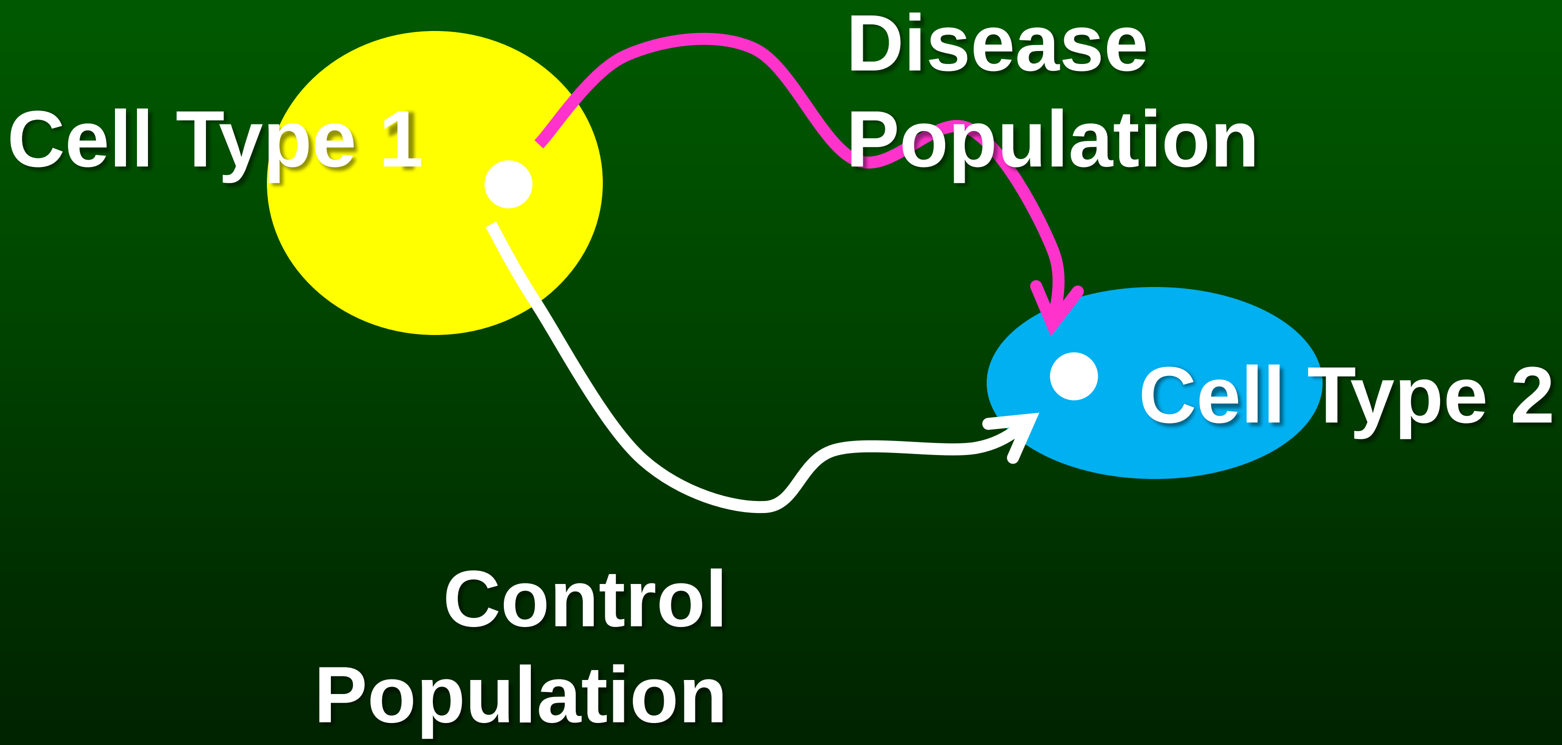
- **Can we model 'attractor states' if we accept that a cell has multiple phenotypes, many of which are shared with other cells types?**
- ***Can we define Competency if we don't first understand the cellular state of play?***
- **Evocation is more than just the external signal - it must define essential aspects of a canalised network**
- ***Canalisation: An evolutionarily conserved process that has specialised as organisms become more complex. The means to model complexity at a genetic, epigenetic or transcriptome level*.**
- **Individuation: What is the range of normal, and can we use this to predict disease states.**



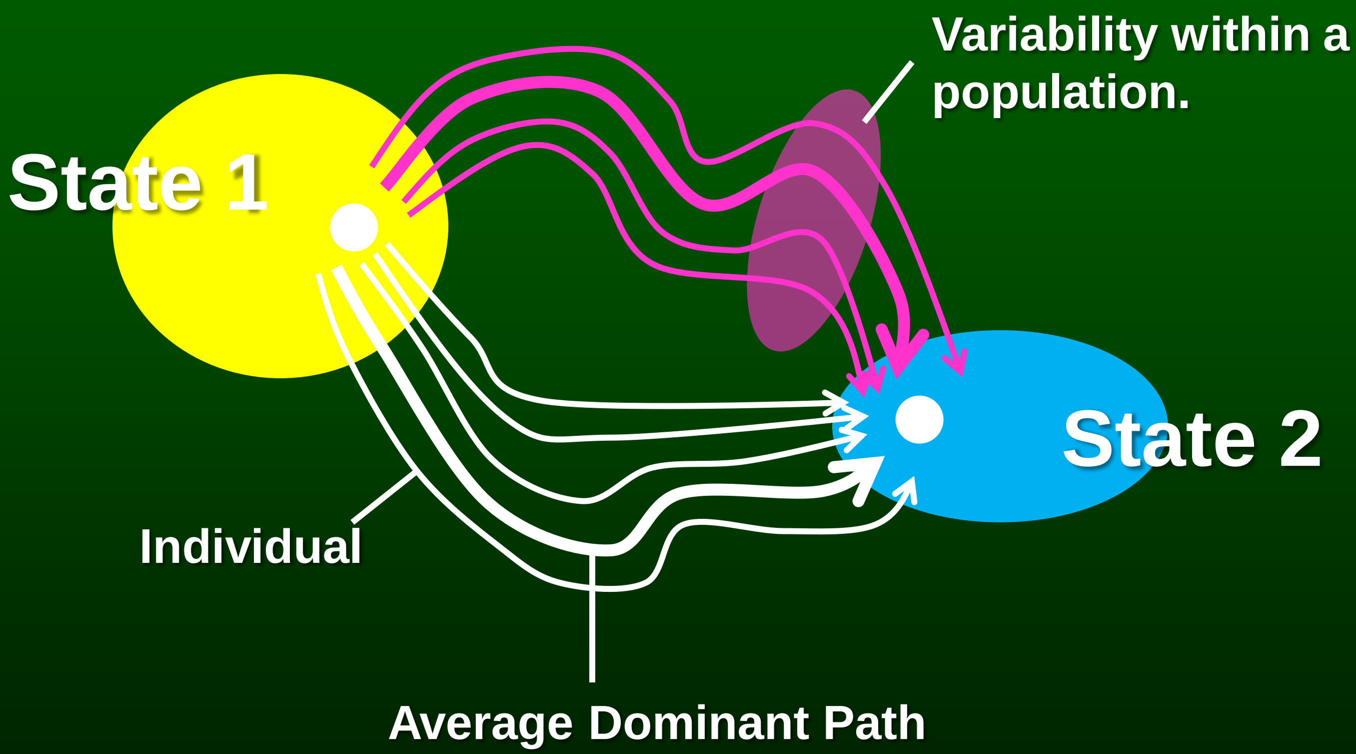
More generally, we can think about other transitions between states.



In the presence of disease…



Within any one population of individuals, we can think of individuals each having their unique trajectory.

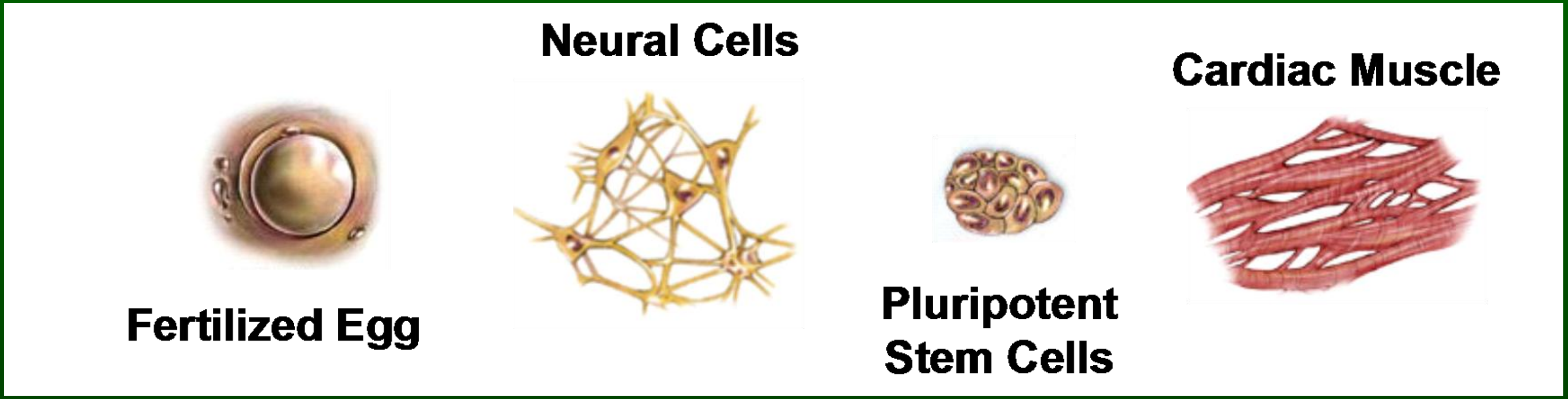


Attract: a method for identifying core pathways that underlie cell fate transitions

Jessica C. Mar, DFCI
Christine Wells, Griffith University and Eskitis Institute

Cell Diversity

A mammalian organism consists of ~250 highly-specialized cell types.



Most cell types share the same genome.

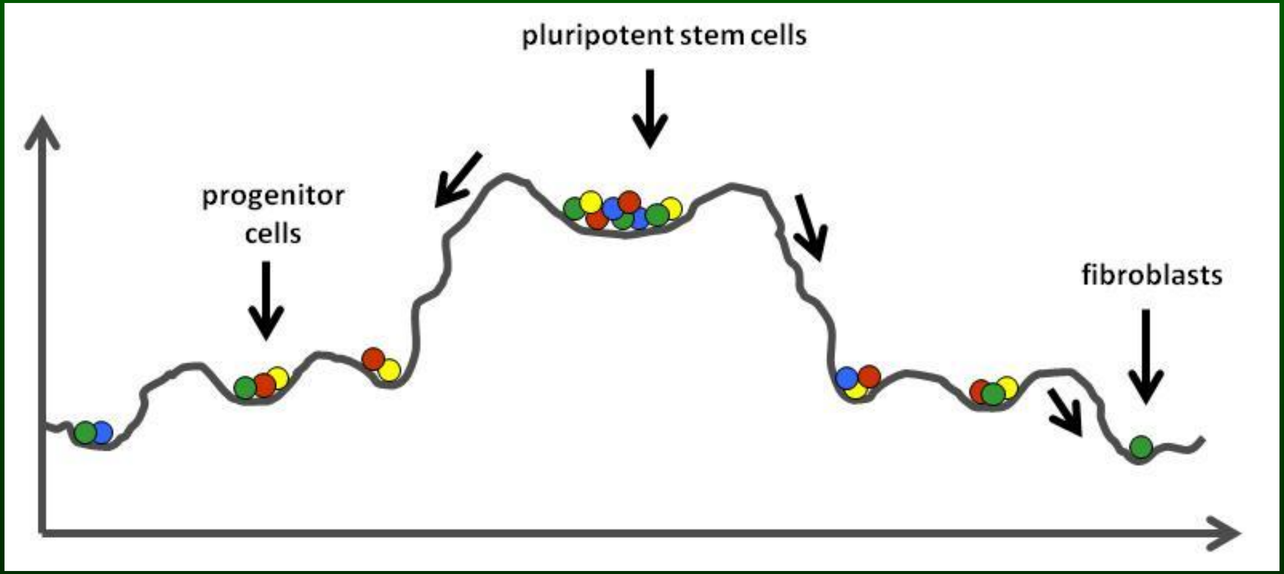
Epigenetic modification and transcription factor networks generate the mechanism for cell type-specific diversity.

A cell type's unique program is manifested by its transcriptional profile.

Deconstructing a Cell's Gene Expression Program

Isolating the active biological pathways that are specific to a cell type allows us to begin to model the transcriptional landscape of cellular states.

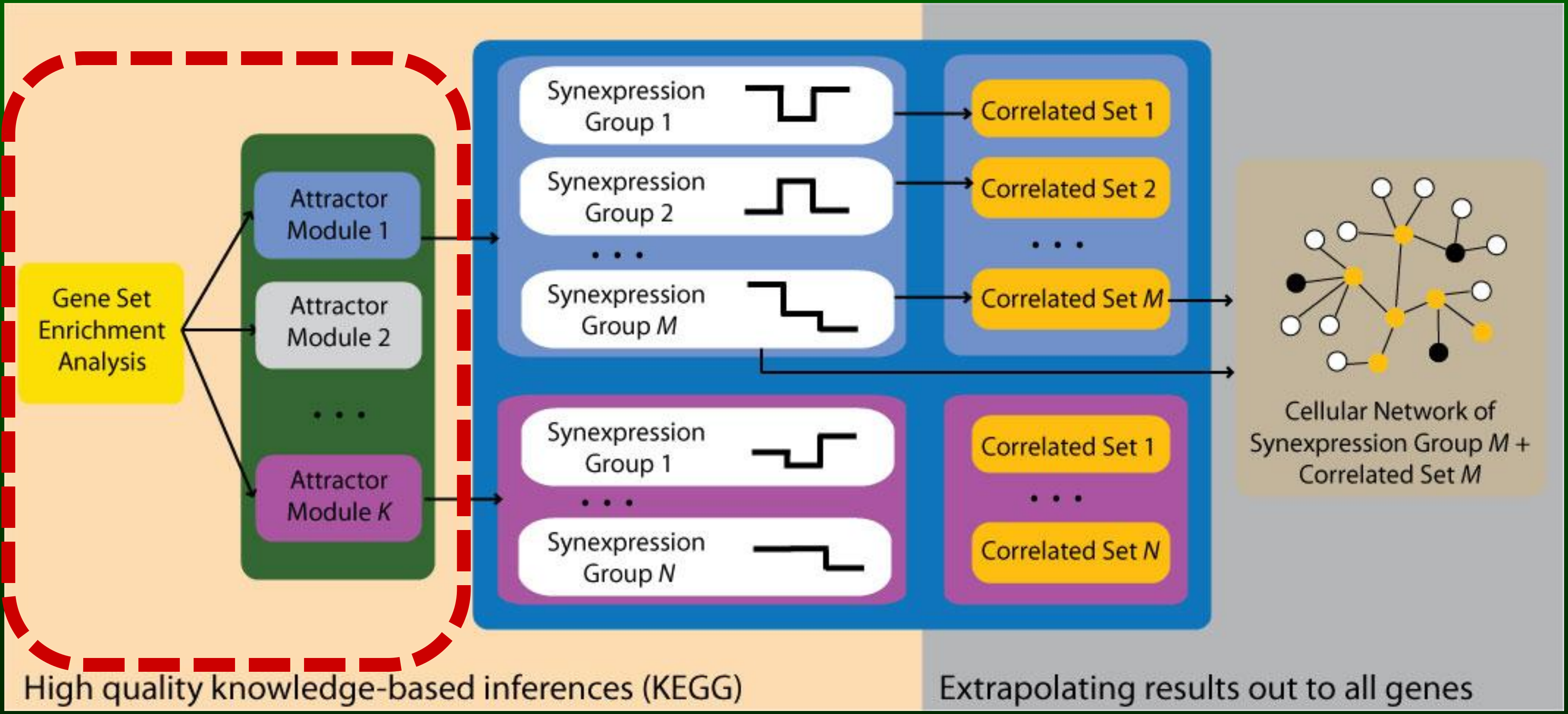
Linking gene signatures to cell lines is a start.



Our goal is to go further, and (eventually) model cell fate transitions.

Adapted from Sui Huang, Bioessays 31:546, 2009

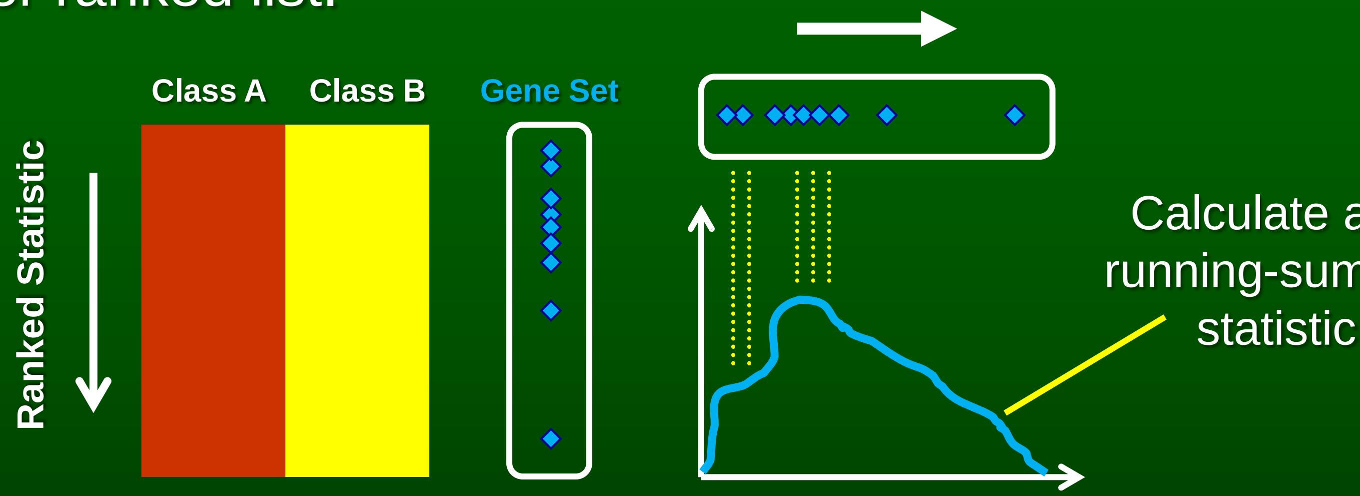
Finding Core Pathways that Underlie Cell Fate Transition



R package **`attract`** available from Bioconductor

GSEA + Linear Model

GSEA tests if members of the gene set are randomly distributed in the larger ranked list.



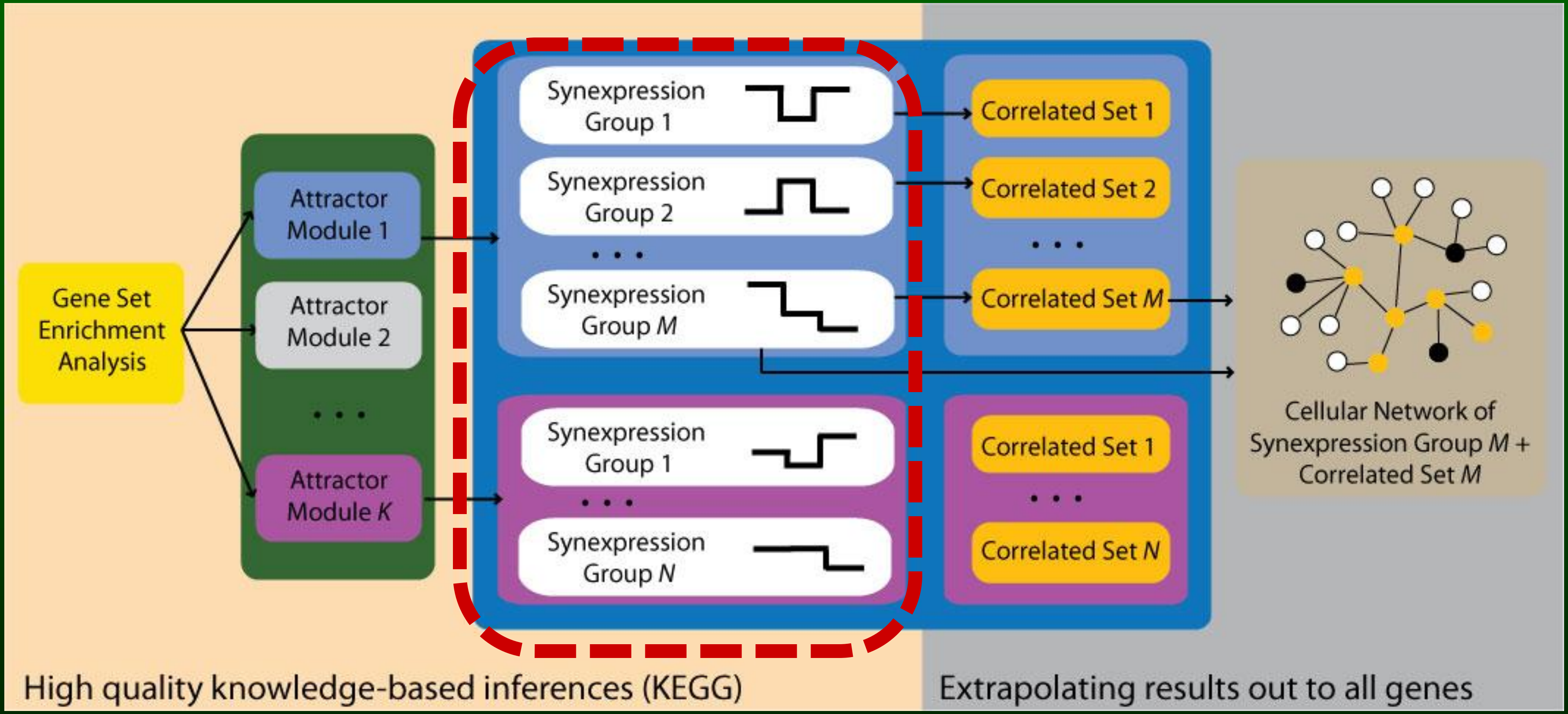
Jiang and Gentleman extended the original implementation by Subramanian.

They generalized the ranking statistic using a generic linear model:

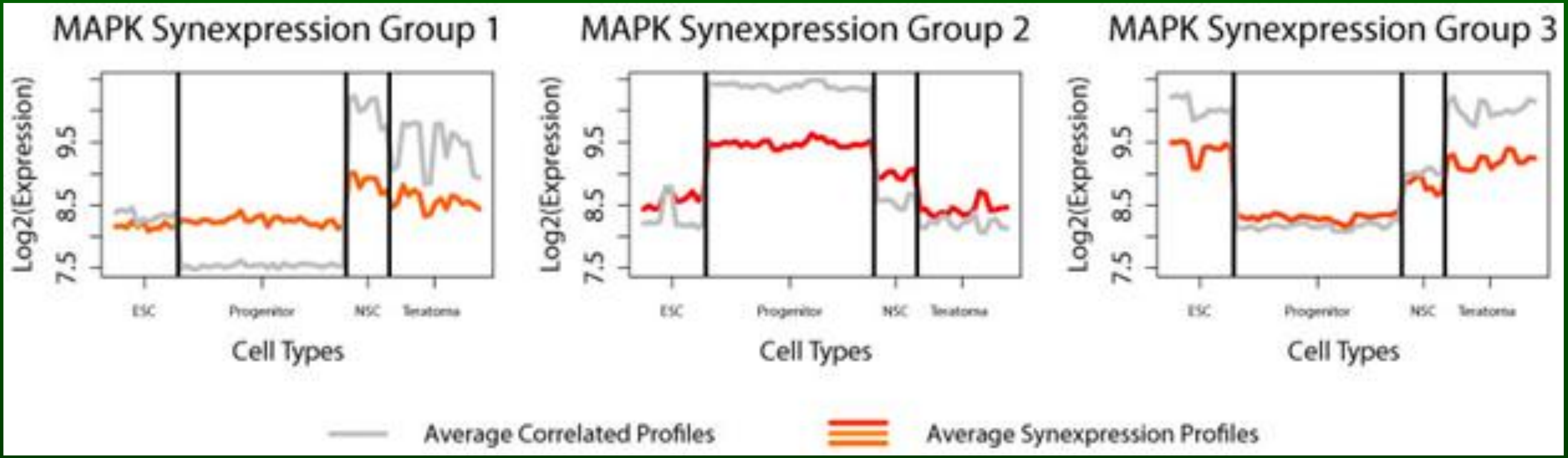
$$y_{gi} = \beta_{g0} + \sum_{j=1}^{p} X_{ij}\beta_{gj} + \epsilon_{gi},$$

for gene $\boldsymbol{g}$, sample $\boldsymbol{i}$ and $\boldsymbol{p}$ covariates.

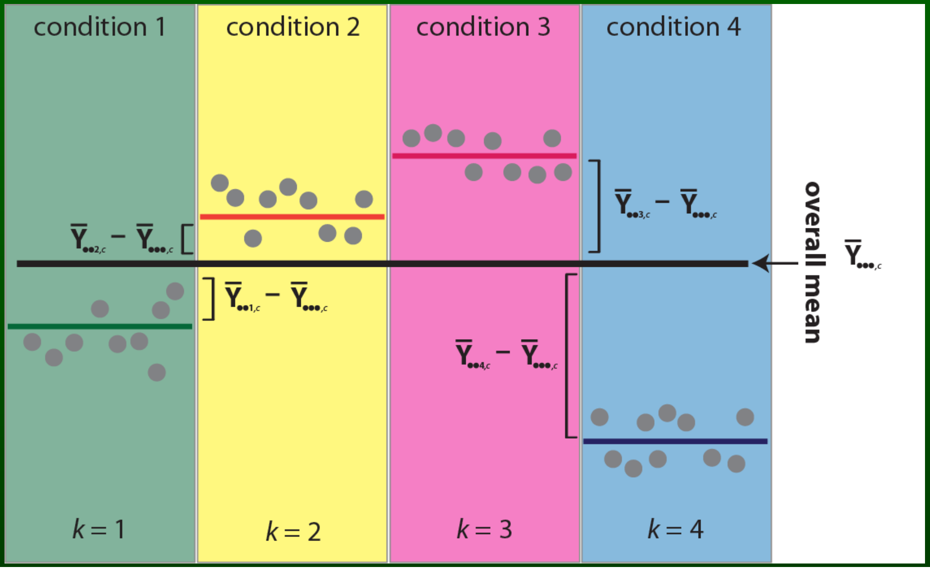
Finding Core Pathways that Underlie Cell Fate Transition

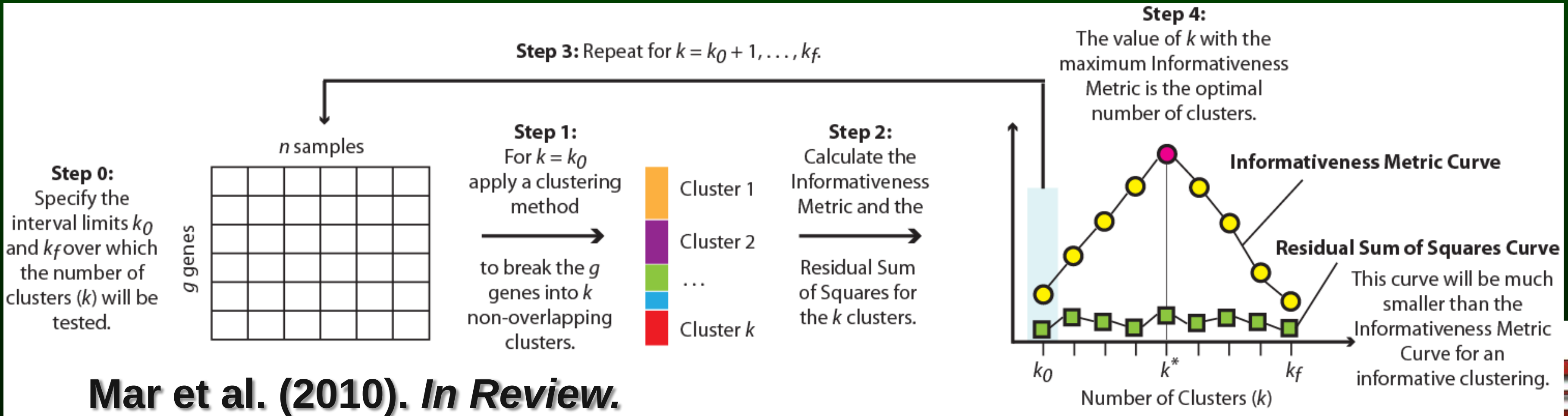


R package `attract` available from Bioconductor


MAPK Synexpression Group 1
MAPK Synexpression Group 2
MAPK Synexpression Group 3
Log2(Expression)
7.5
8.5
9.5
ESC
Progenitor
NSC
Teratoma
Cell Types
Average Correlated Profiles
Average Synexpression Profiles

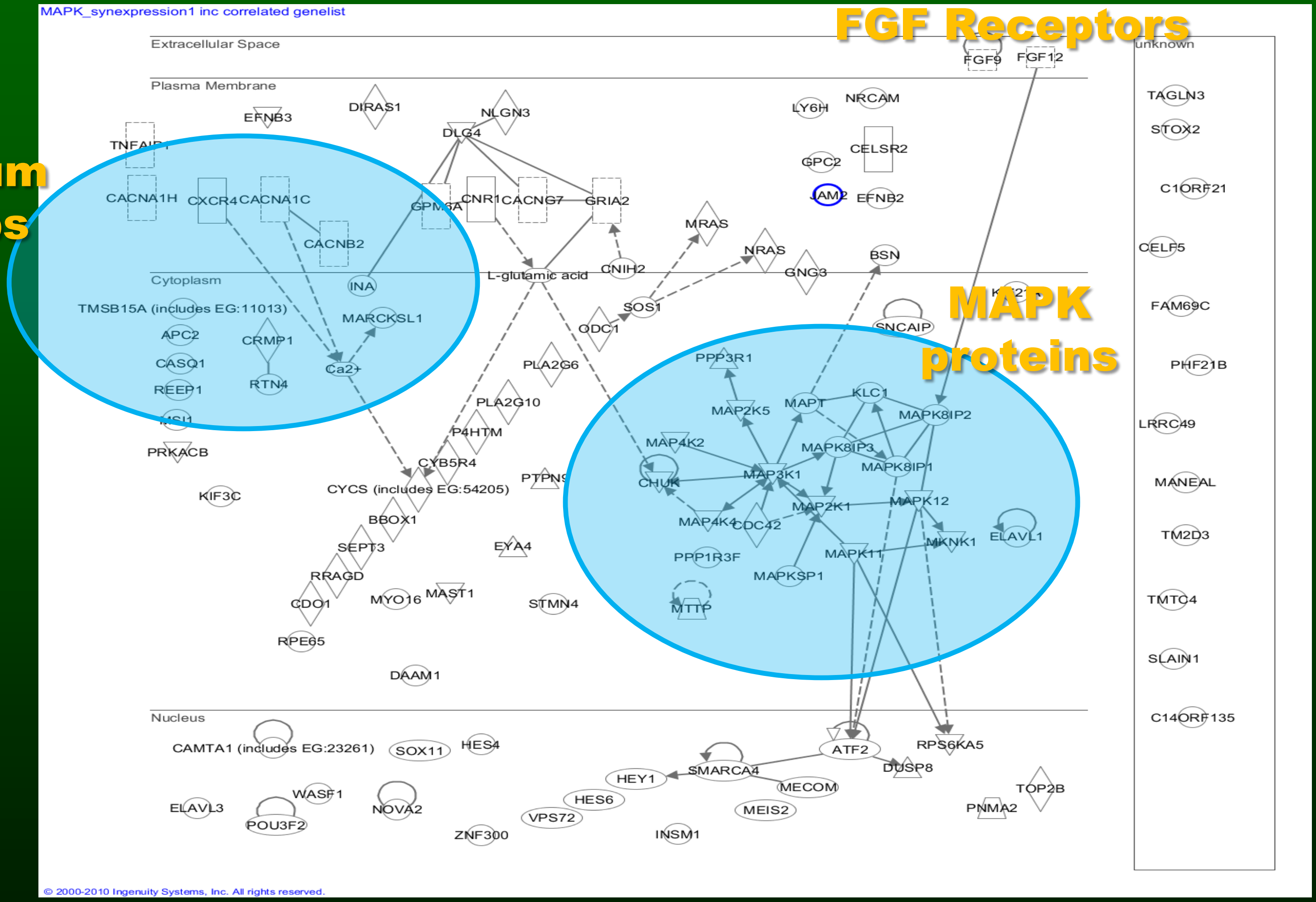
Defining an Informativeness Metric





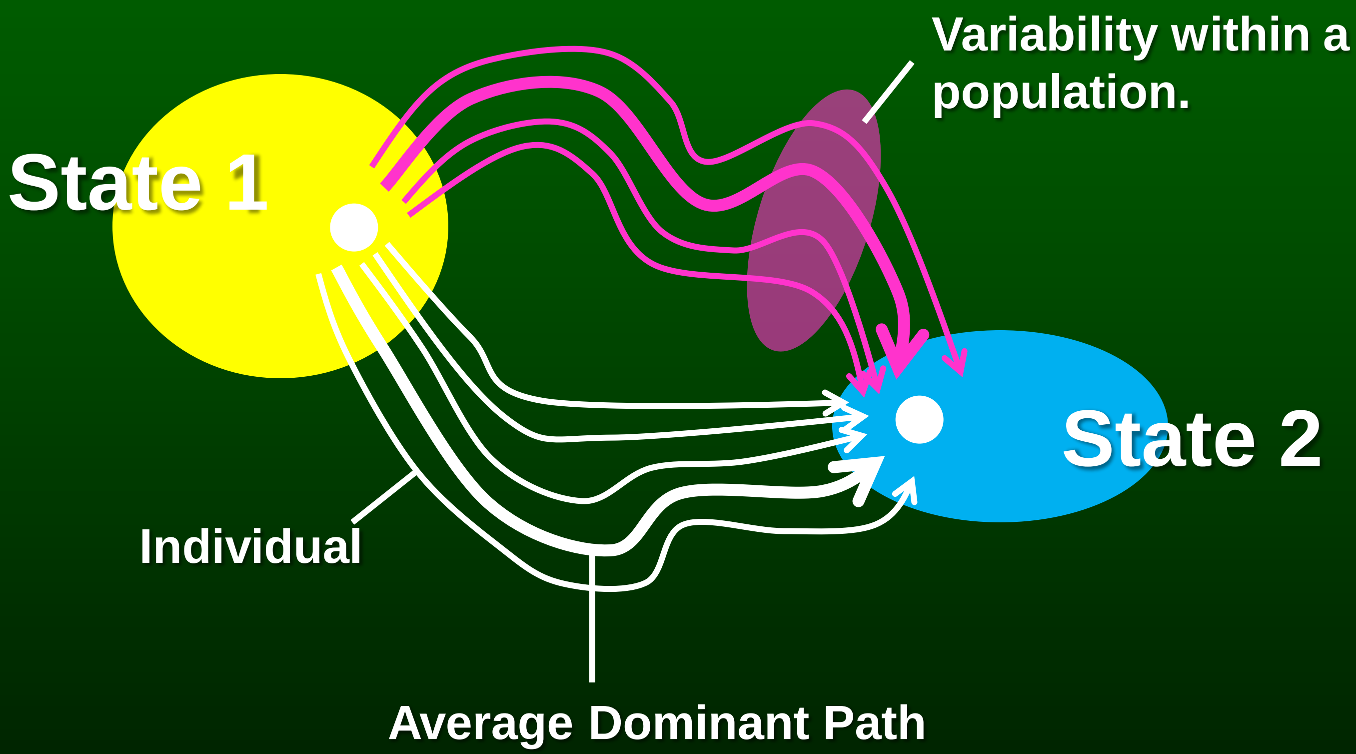
Mar et al. (2010). *In Review.*

Interpreting Synexpression Groups through Biological Networks



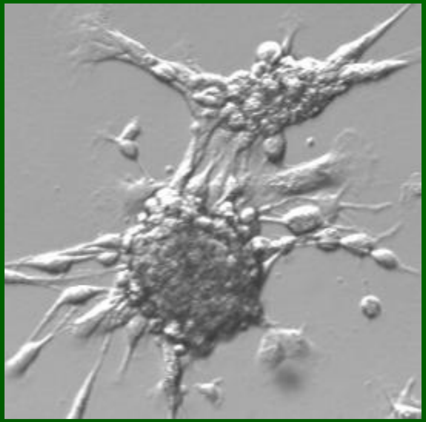
Example of MAPK Synexpression Group

Within any one population of individuals, we can think of individuals each having their unique trajectory.



A variational approach to expression analysis in human disease

Jessica C. Mar, DFCI
Christine Wells, Griffith University and Eskitis Institute



Data Set: Studying Adult Stem Cell Populations

Nasal biopsies from a control group of **related donors** from a larger study on *Parkinson's disease* and *schizophrenia*.

Mesenchymal stem cells from a group of **unrelated donors** from three sources: human placenta, chord blood and bone marrow.

Control Lines

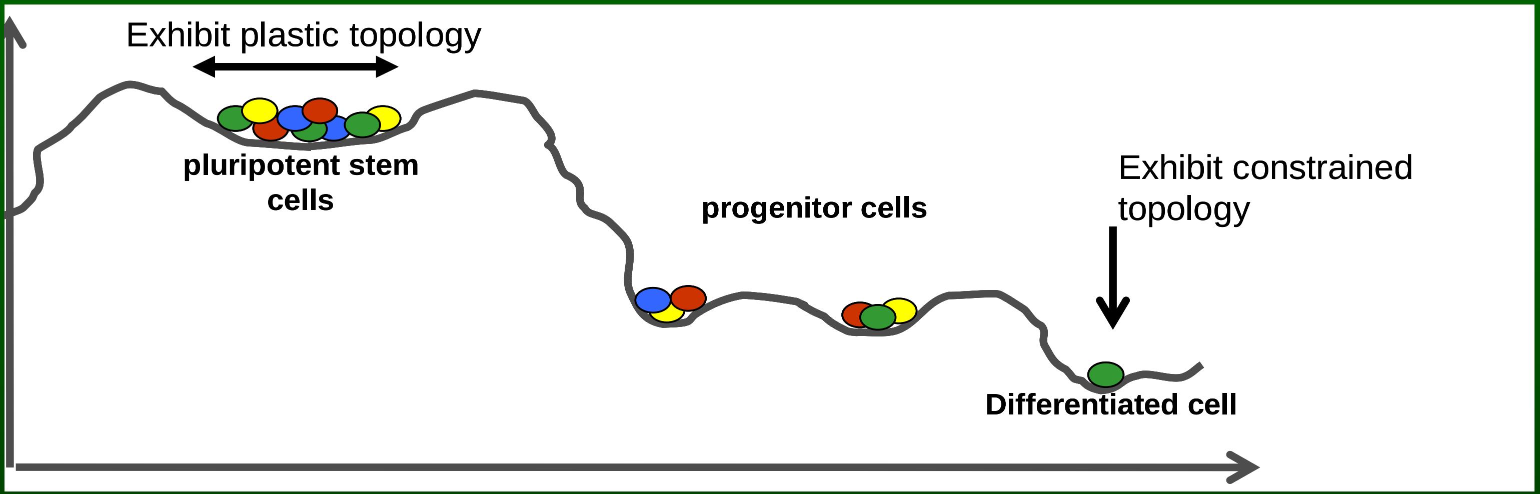
9 Fibroblasts
9 OPBs Primary Olfactory Biopsies
15 ONCs Expanded Olfactory neurosphere-derived Cells
12 MSCs Mesenchymal stem cells

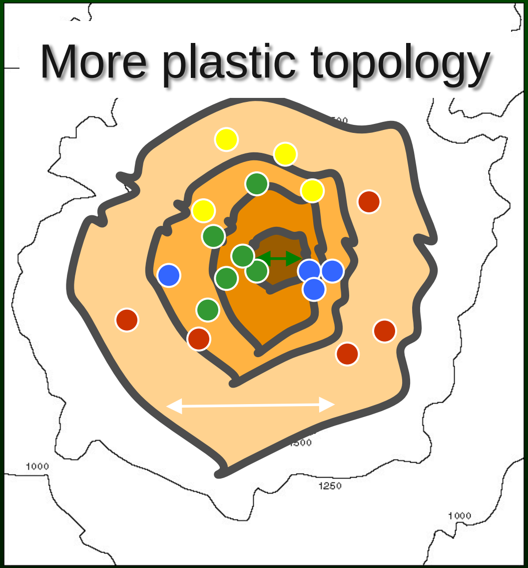
Disease Grops

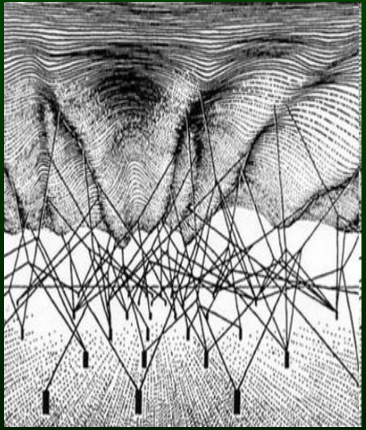
9 Fibroblasts
9 Schizophrenia ONCs
15 Parkinson's Disease ONCs

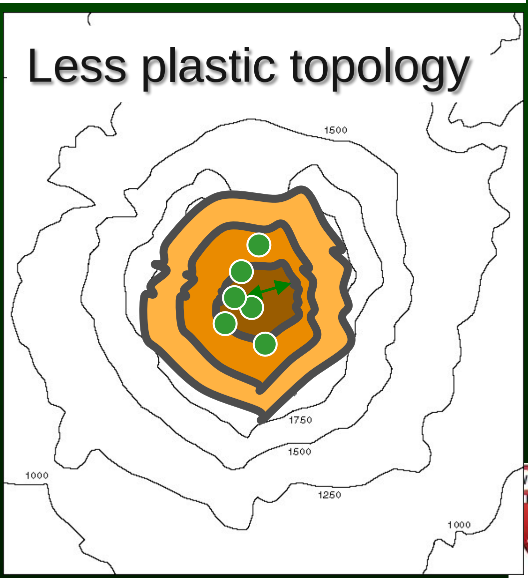
Olfactory Stem Cells Have More Plasticity Across Attractor Modules

Indicative of competency to respond to external signals

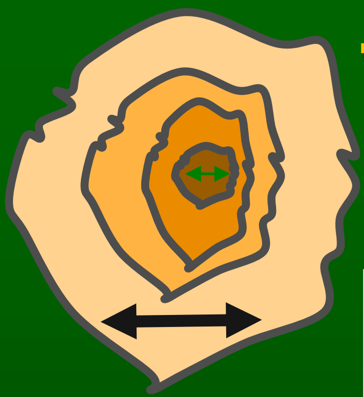




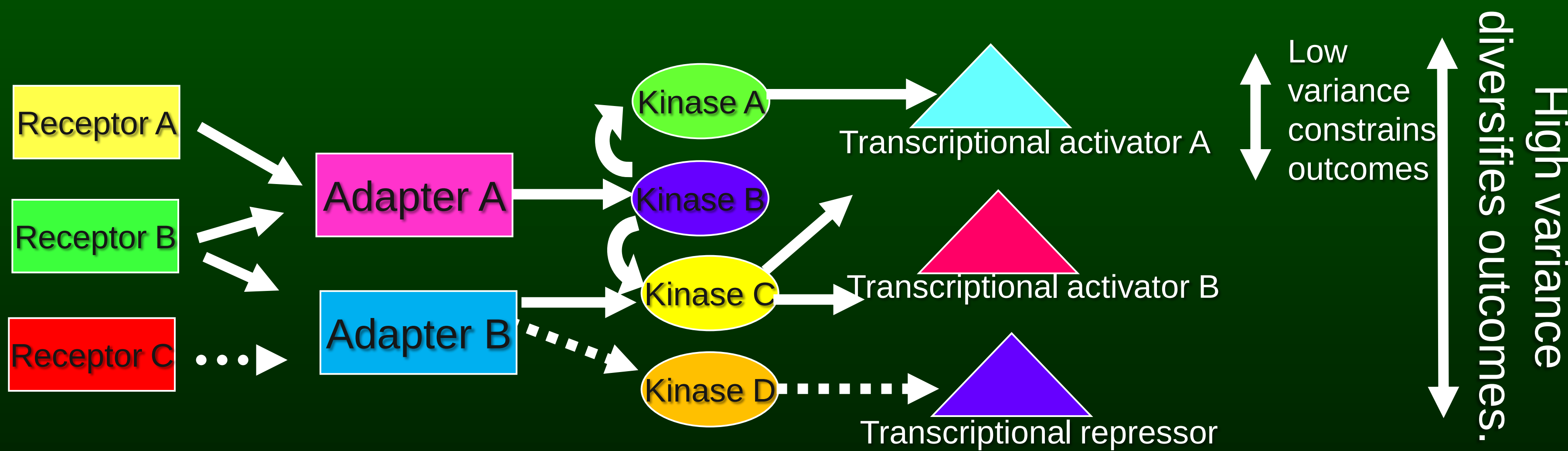




Variance of expression imposes topology on the network



Low variance indicates tighter regulatory constraints
High variance indicates more functional plasticity



Identifying the Core Attractor State Pathway Modules

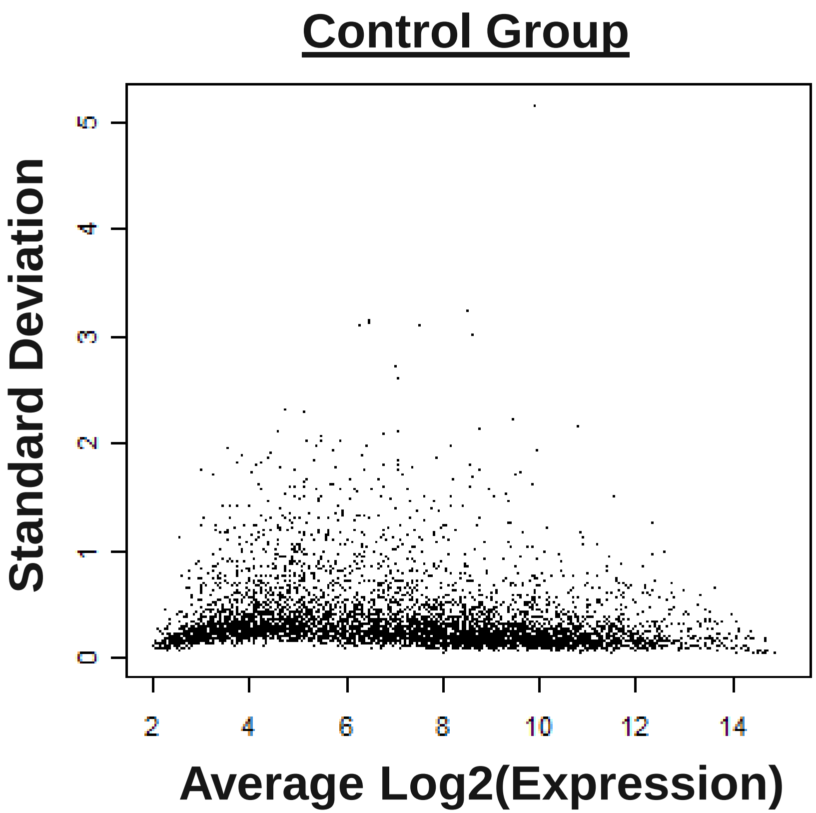
For the Control Group only, we used the data set on 4 cell lines:

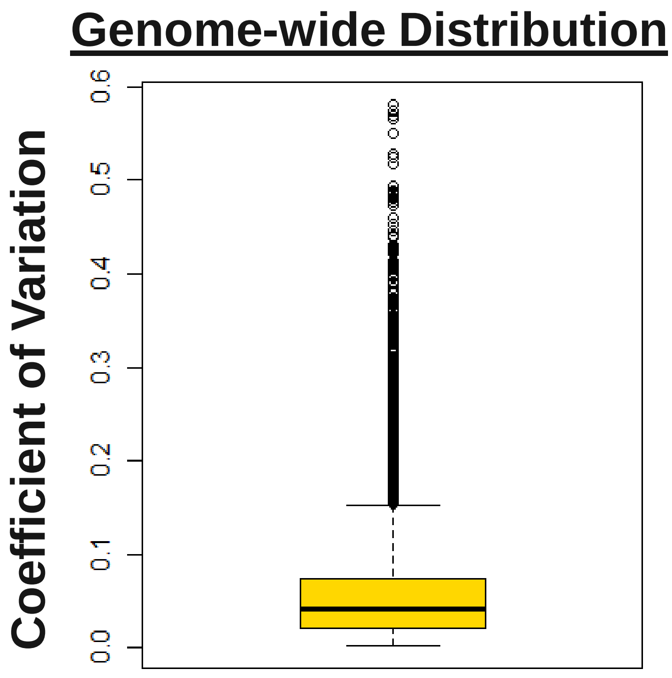
Rank	KEGG Pathway ID	KEGG Pathway Name	P-value	Number of Illumina IDs
1	4010	MAPK signaling pathway	0	238
2	4810	Regulation of actin cytoskeleton	0	196
3	4510	Focal adhesion	0	194
4	4120	Ubiquitin mediated proteolysis	0	141
5	4910	Insulin signaling pathway	0	132
6	4310	Wnt signaling pathway	0	131
7	4020	Calcium signaling pathway	0	129
8	4530	Tight junction	0	115
9	4670	Leukocyte transendothelial migration	0	97
10	4650	Natural killer cell mediated cytotoxicity	0	96

Measuring Variability

Assess standard deviation of probe fluorescent intensity across all of the donors.

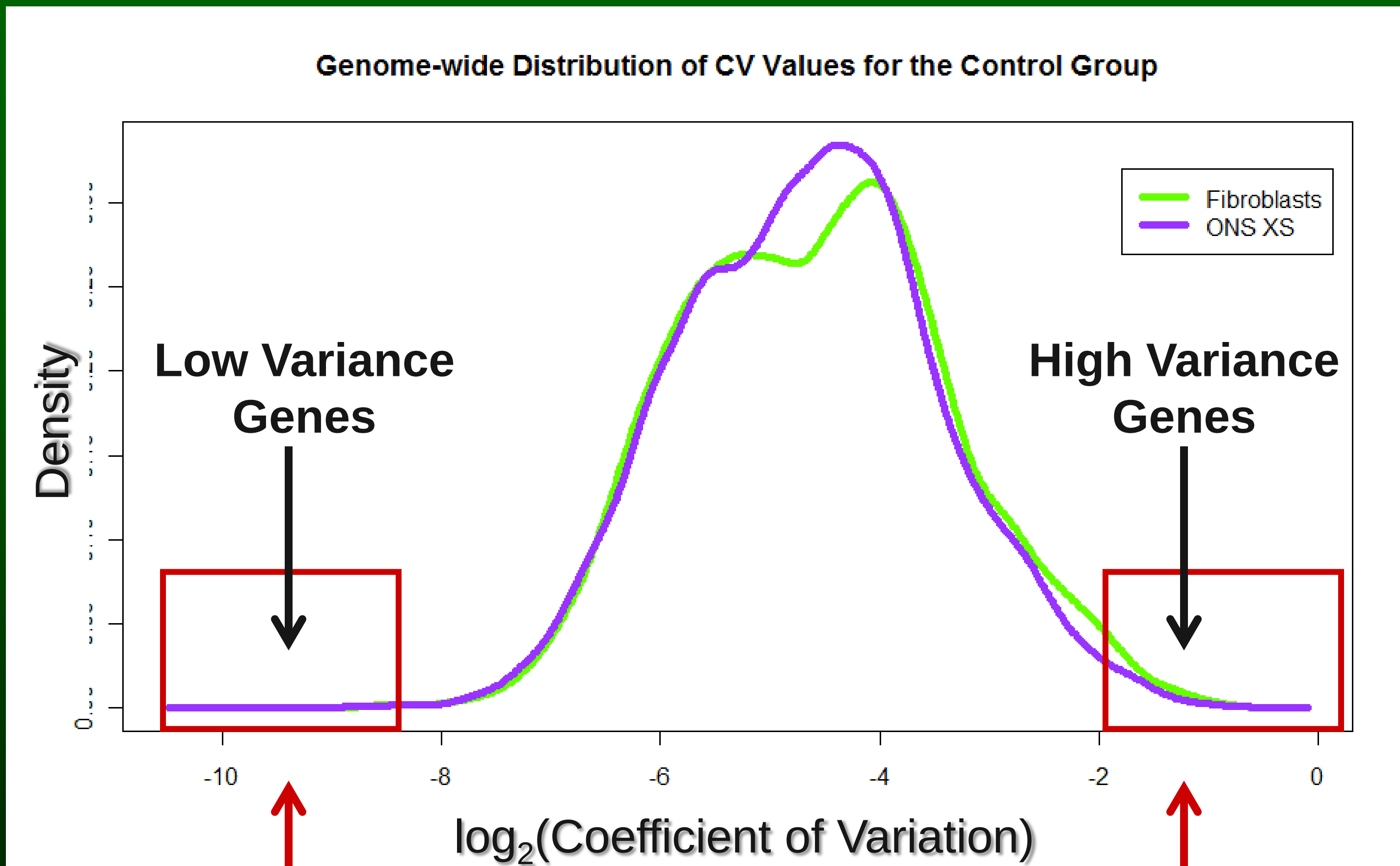
Coefficient of Variation = StandardDeviation:Mean





Illumina Ref8v2 chips, 13709 probes met detection threshold criteria.

Fibroblasts and Stem Cells Have Similar Genome-wide CV

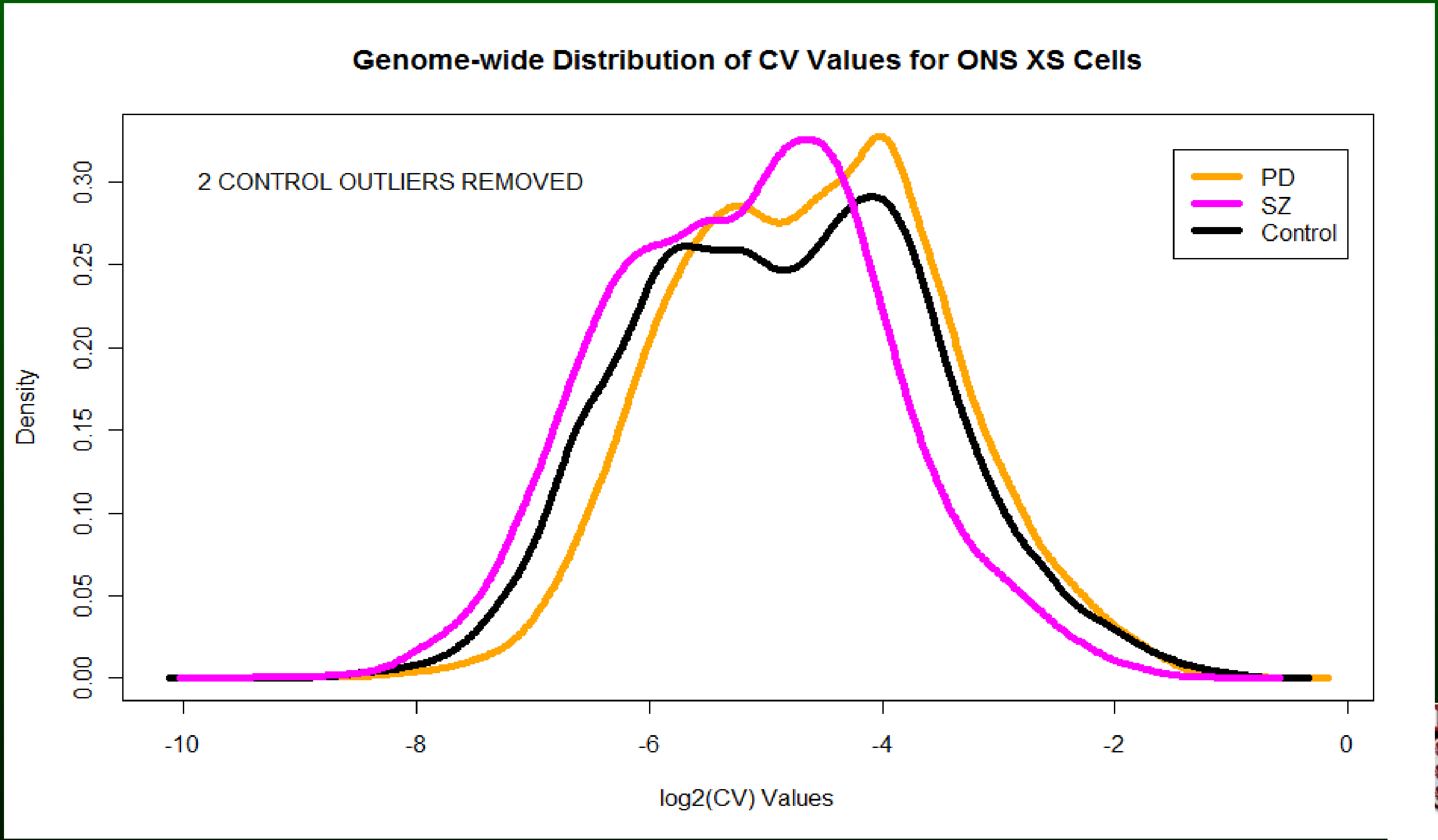


Highly Constrained Genes

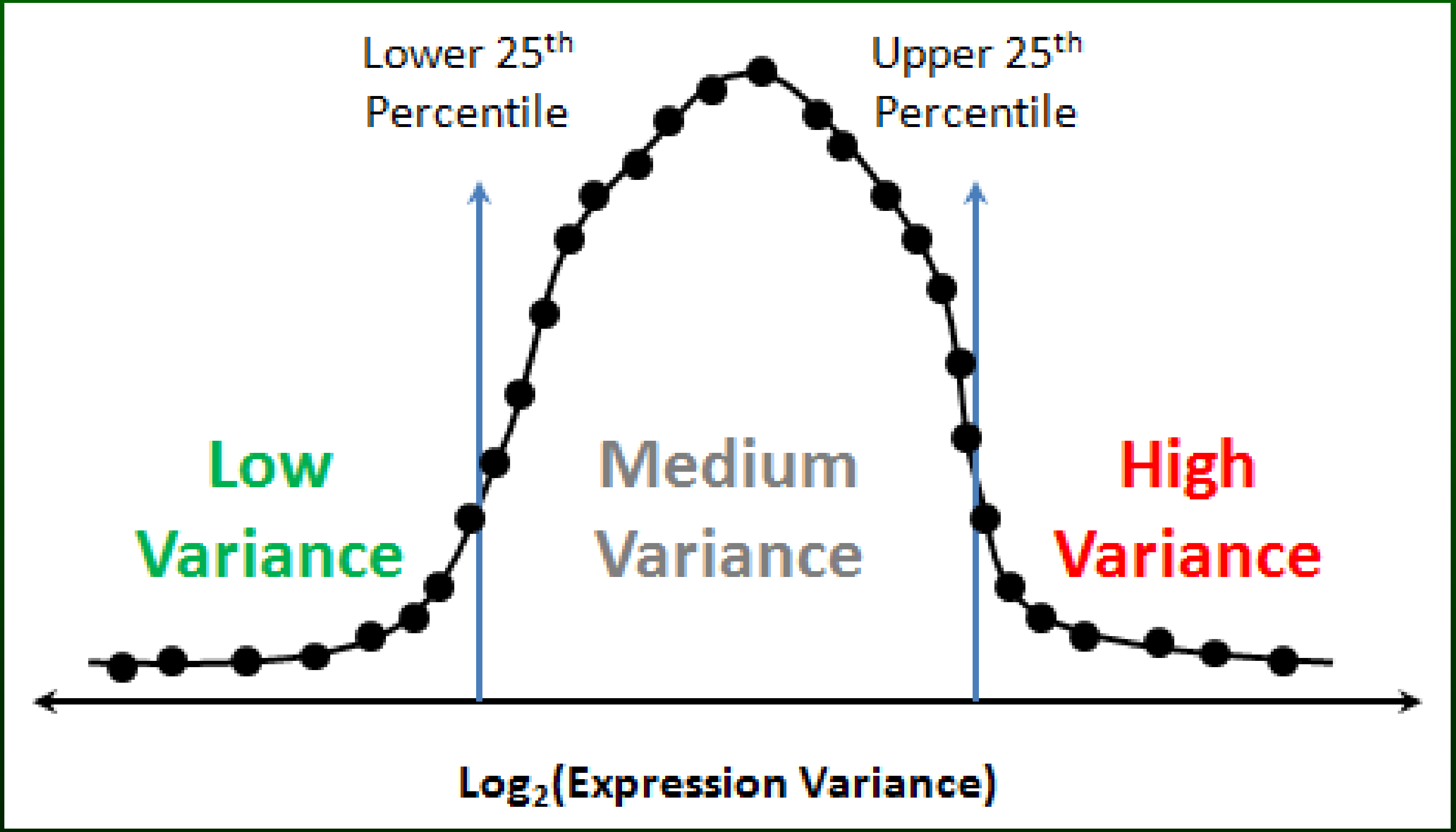
Lowly Constrained Genes

Genome-wide Donor Variability Distributions Are Similar Between Disease Groups

For the ONS cells: 9 SZ patients, 11 controls, 13 PD patients.



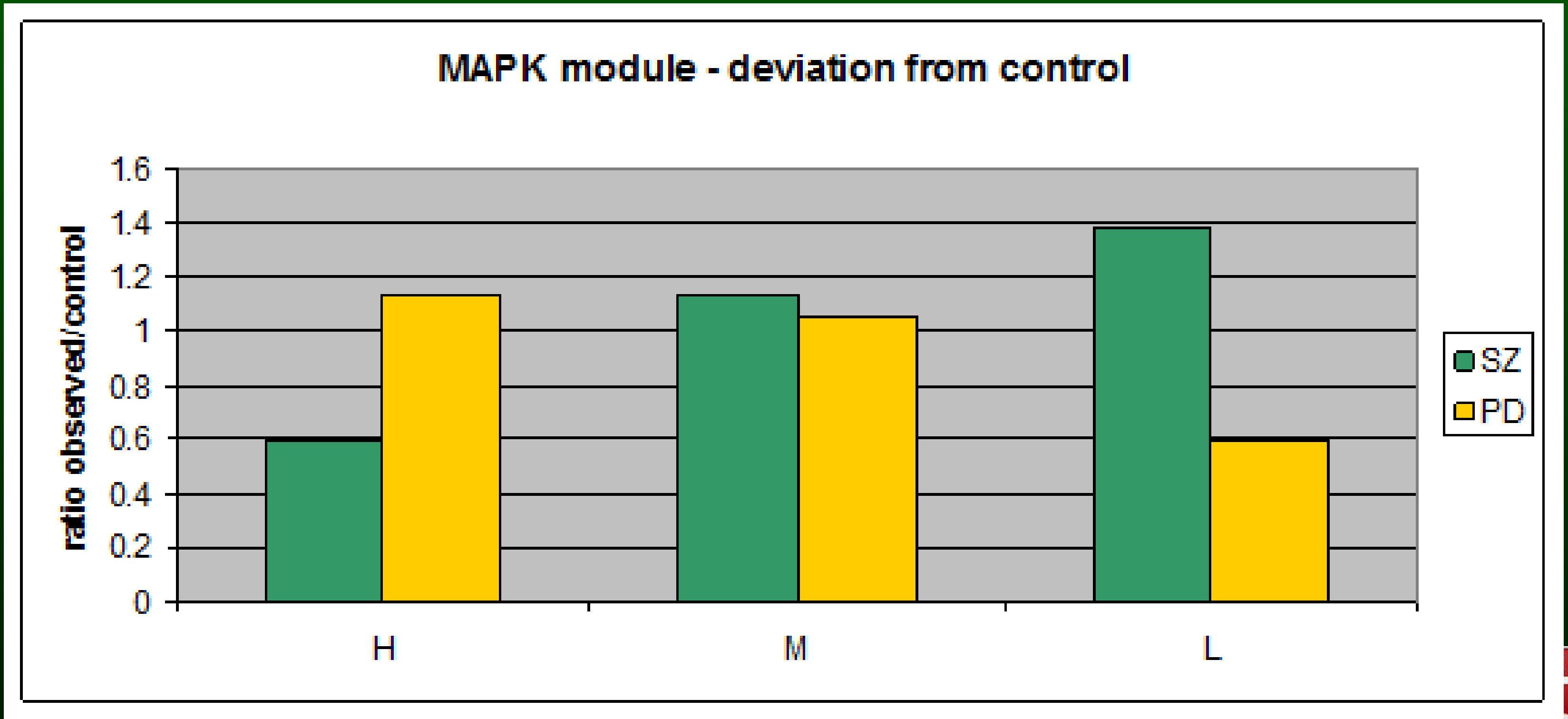
Characterizing Variability in a Disease Group

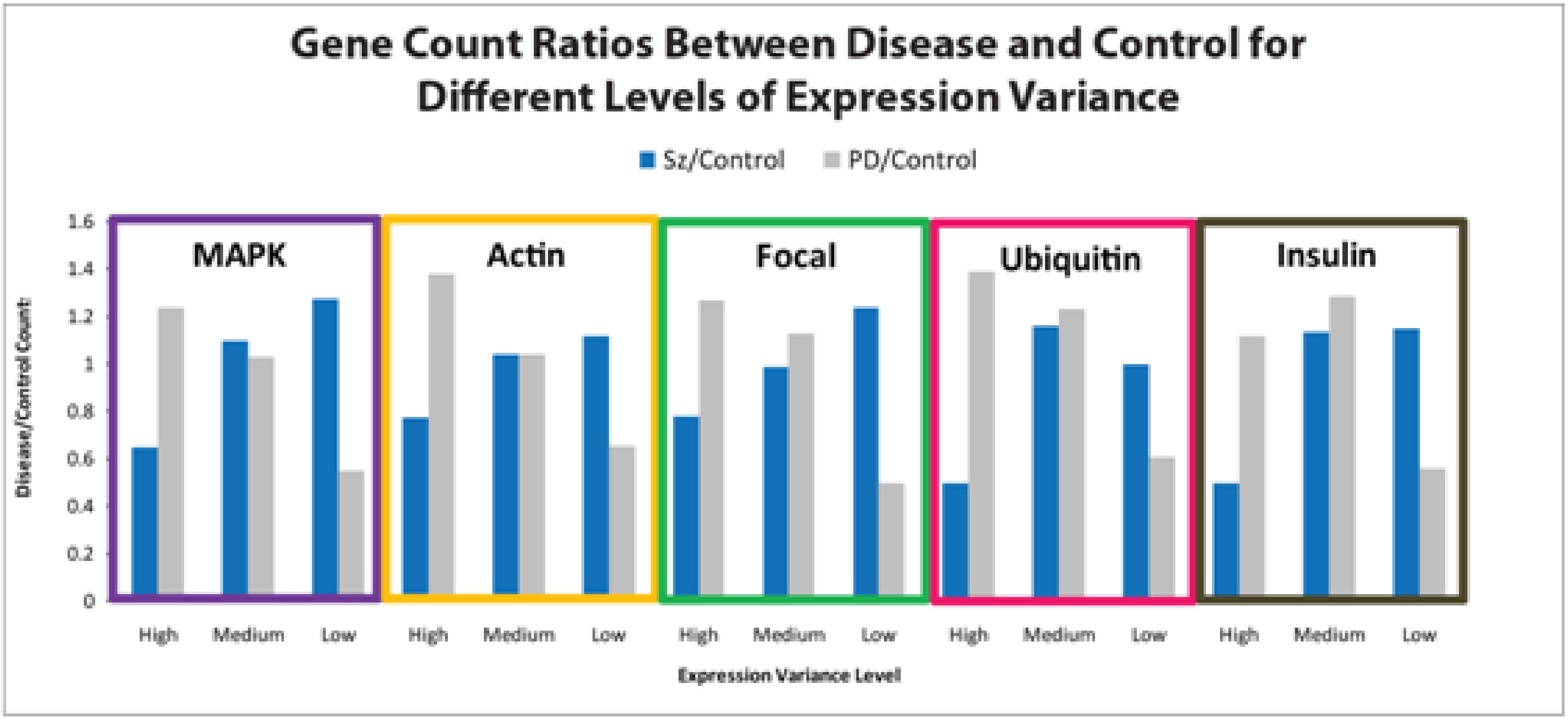


SZ and PD Show Strikingly Different Variability Profiles

We count the number of highly constrained and lowly constrained genes in each patient group.

Ratios of gene counts between disease:control



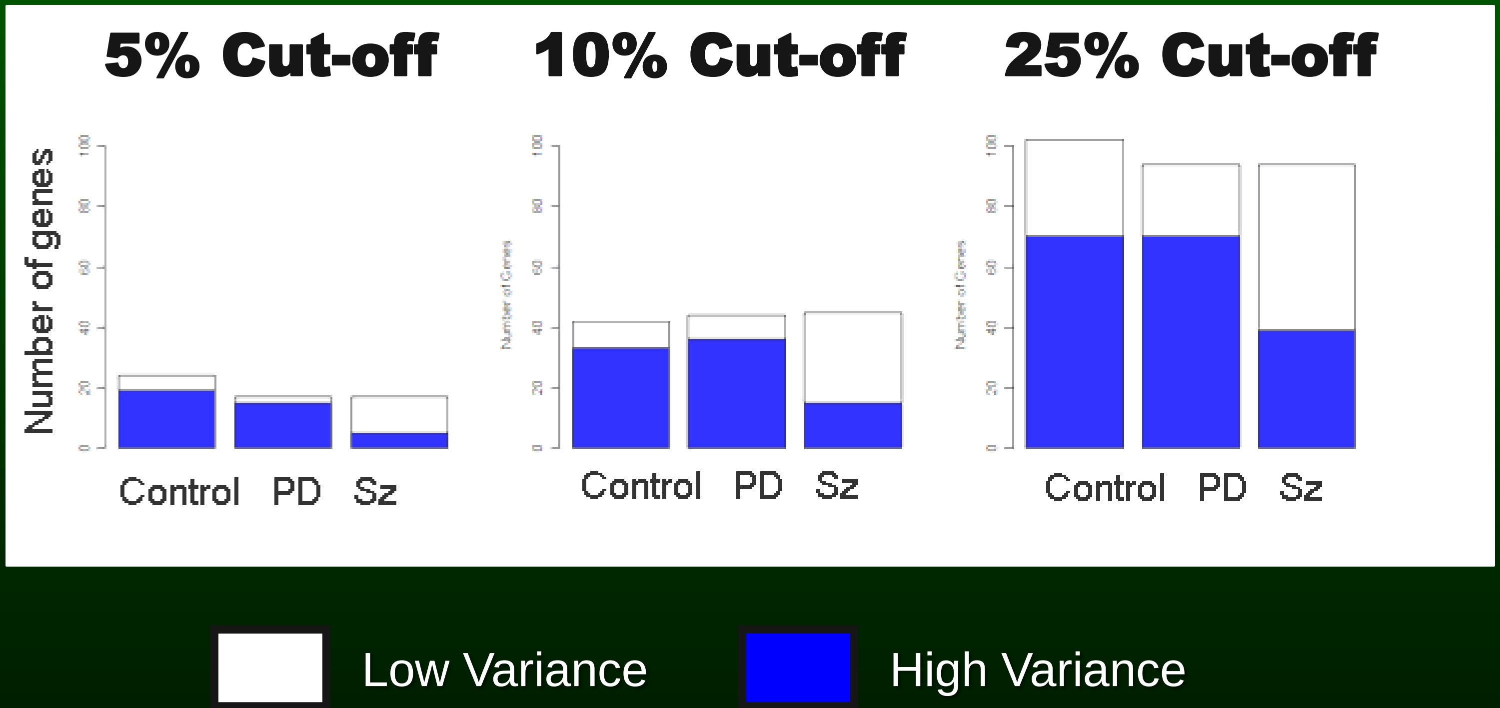


P-values	MAPK	Actin	Focal	Ubiquitin	Insulin
SZ versus Control	0.002769	0.243118	0.087842	0.051139	0.015744
PD versus Control	0.002807	0.00252	0.000315	0.001123	0.001936

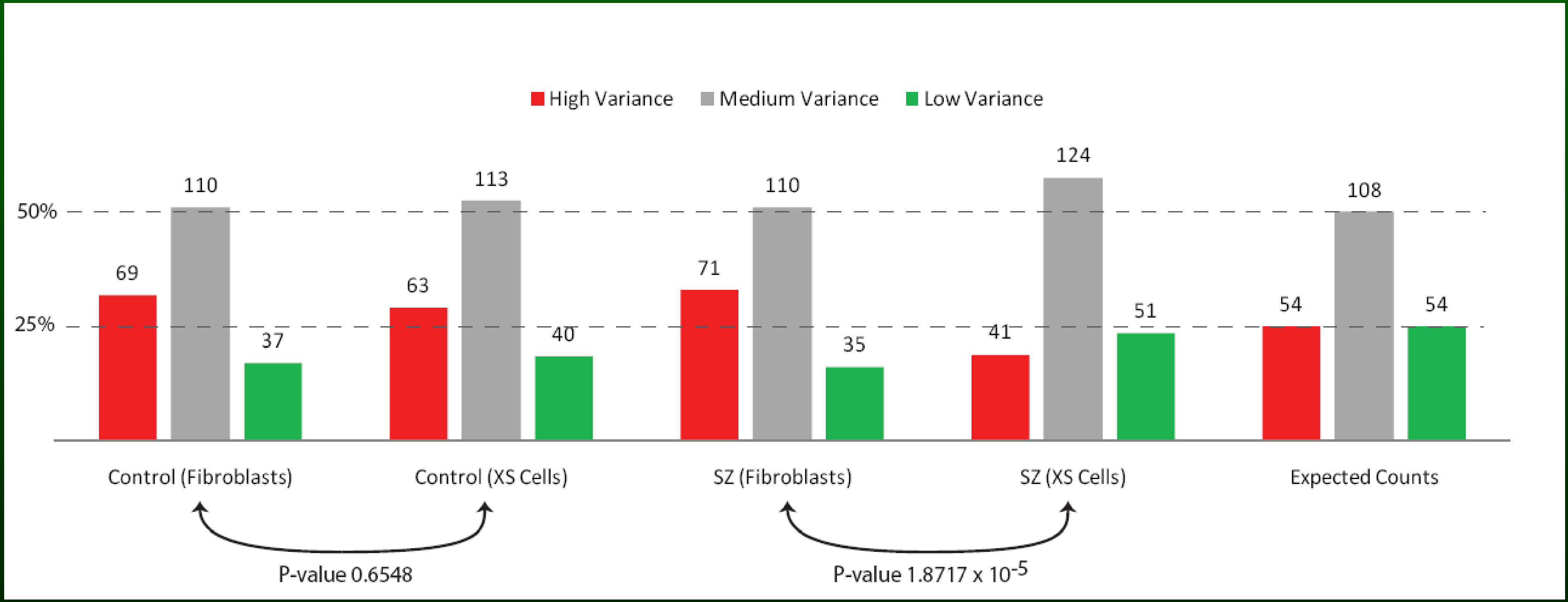
SZ Group Shows Increased Variance for the MAPK Pathway

Definition of high and low variance is based on our 25% cut-off imposed on the pooled distribution.

Patterns of variability are still retained even after increasing the stringency of this cut-off.



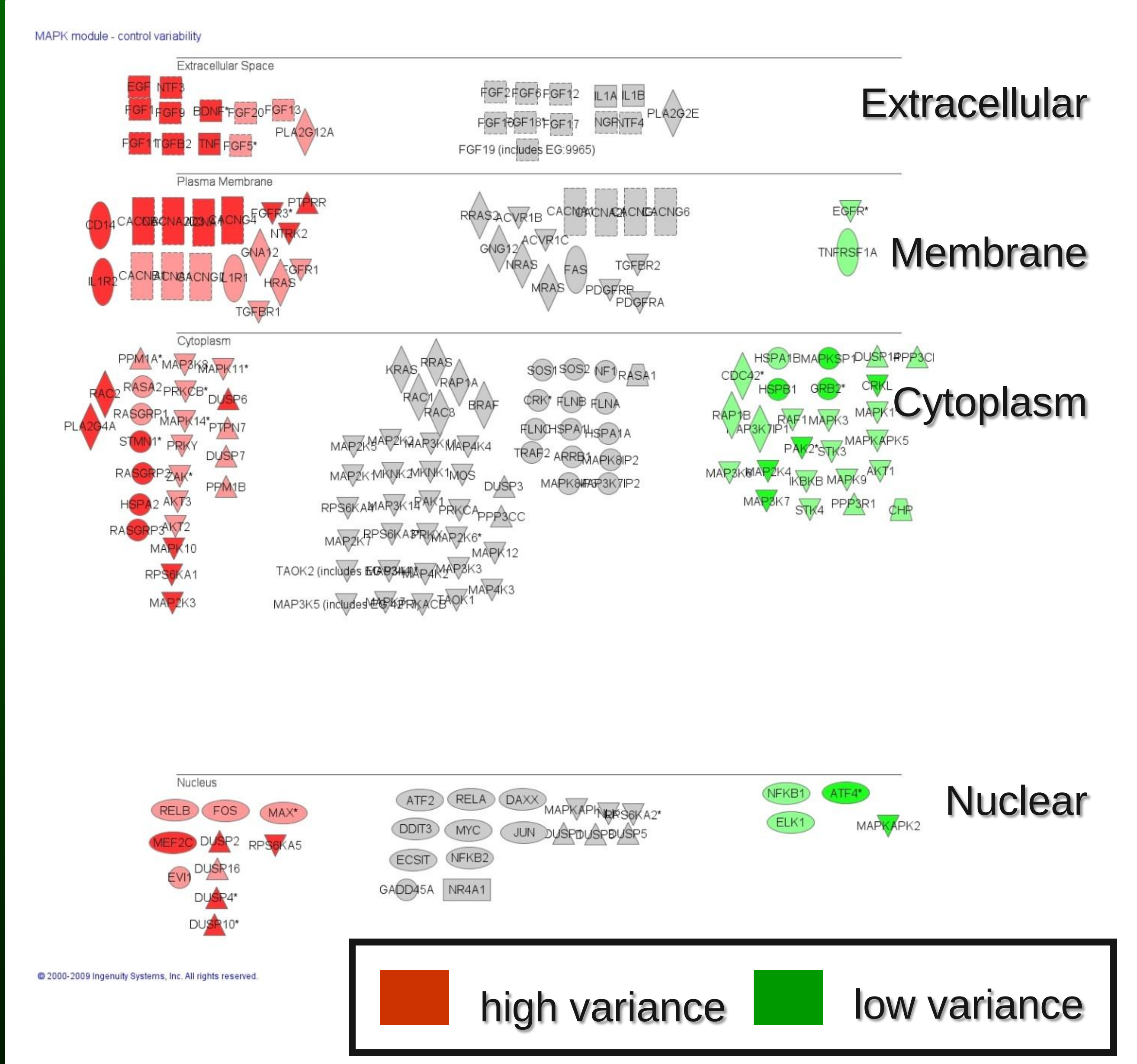
SZ Stem Cells are Different from Fibroblasts



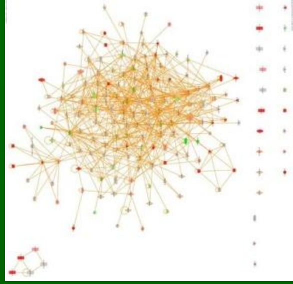
Functional Roles Are Associated with Constraint

High-variance genes tend to function as cell surface receptors.

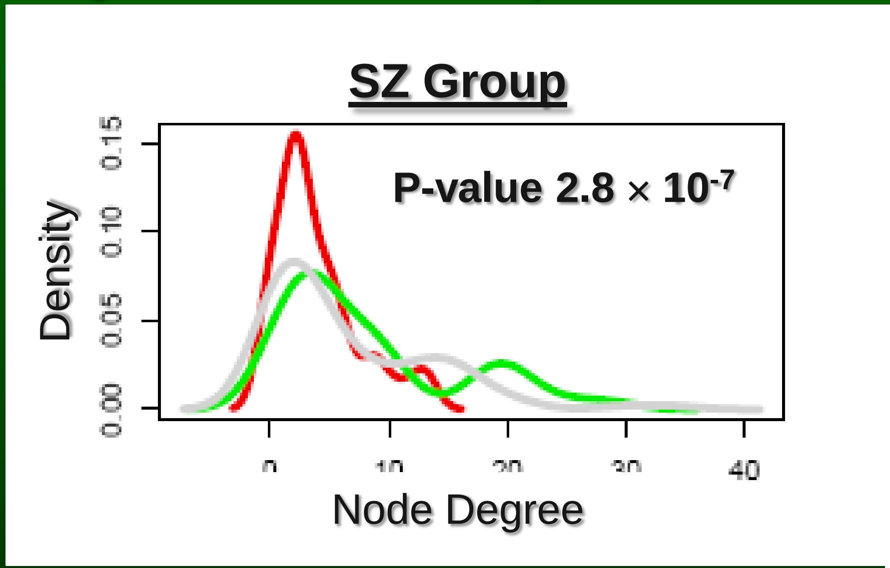
Low-variance genes function as kinases and transferases.

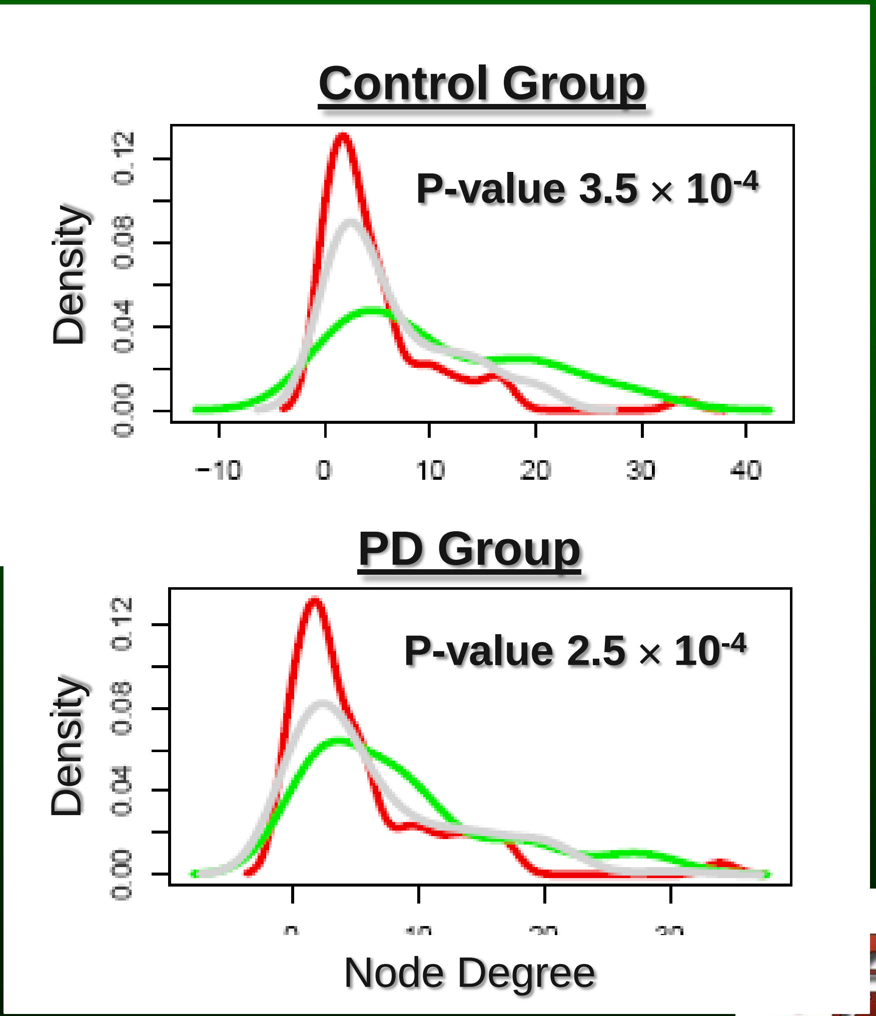


Variance Constraints Alter Network Topology



Degree distributions for the MAPK module are significantly different (Kolmogorov-Smirnov test).



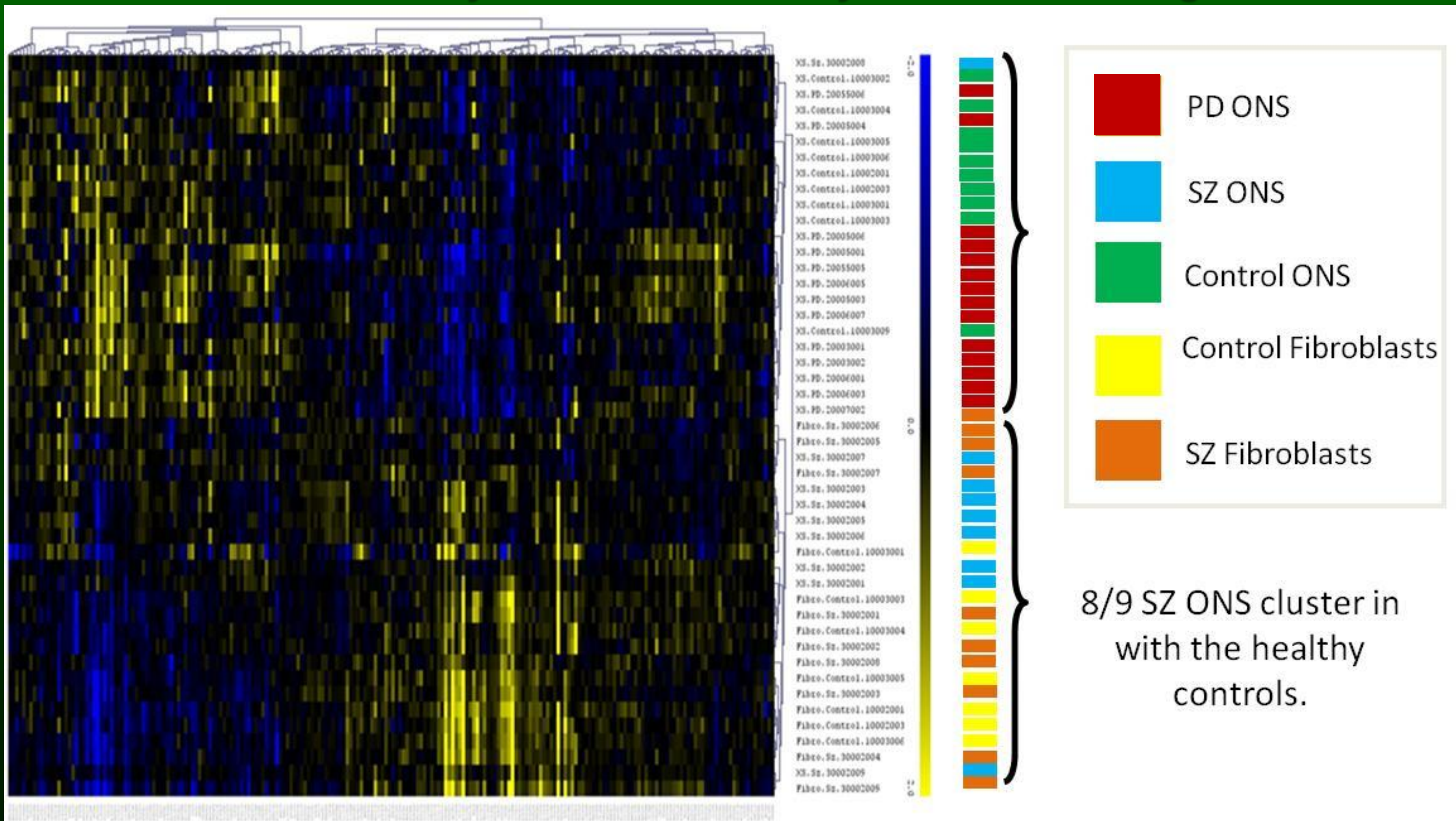


Severity of statistical significance is altered by disease status.

high variance **low variance**

SZ Stem Cells Are More Similar to Healthy Fibroblasts

The transcriptional profiles of ONS XS cells from SZ patients more closely resemble those of healthy fibroblasts than any other stem cell signature.



Disease Variational Analysis

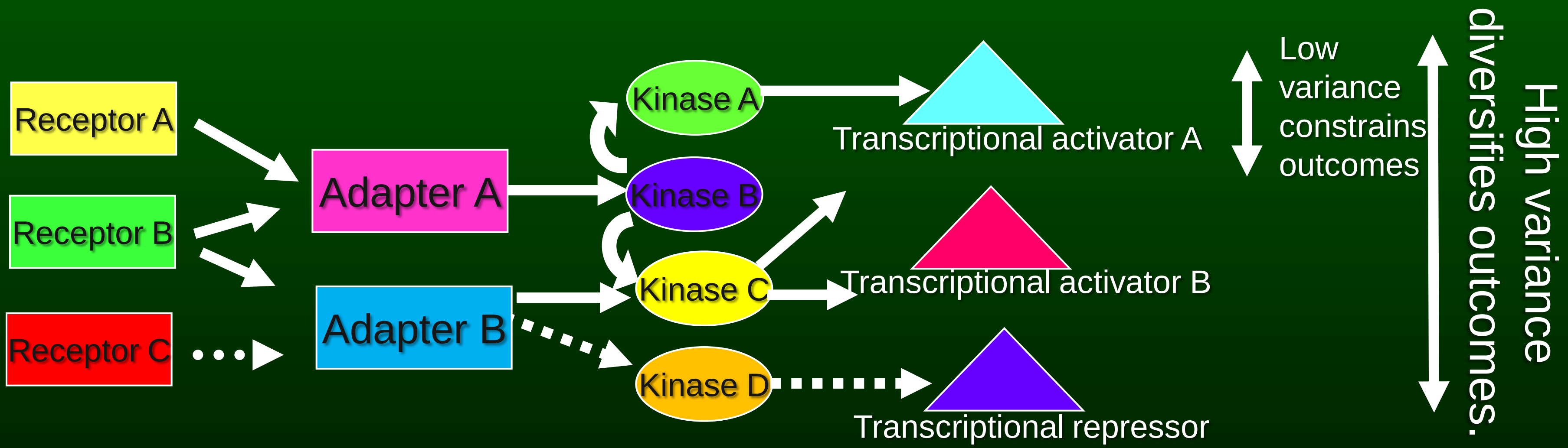
- SZ and PD sit at opposite ends of the expression variance spectrum for core pathway modules.
- A marked decrease in variance was observed for the SZ patients; this raises the possibility that neural stems (and the individuals they were derived from) may be less able to respond to disturbances in the environment.
- This is supported by the observation that SZ stem cells have expression profiles that are more similar to healthy fibroblasts.
- PD was associated with an increase in variance; this may be a result common to other diseases of aging.
- What are the underlying genetic effects that give rise to this variation in expression?

Extrapolating to Individuals

Derive a probabilistic model that determines the most likely path of interactions in a network/pathway.

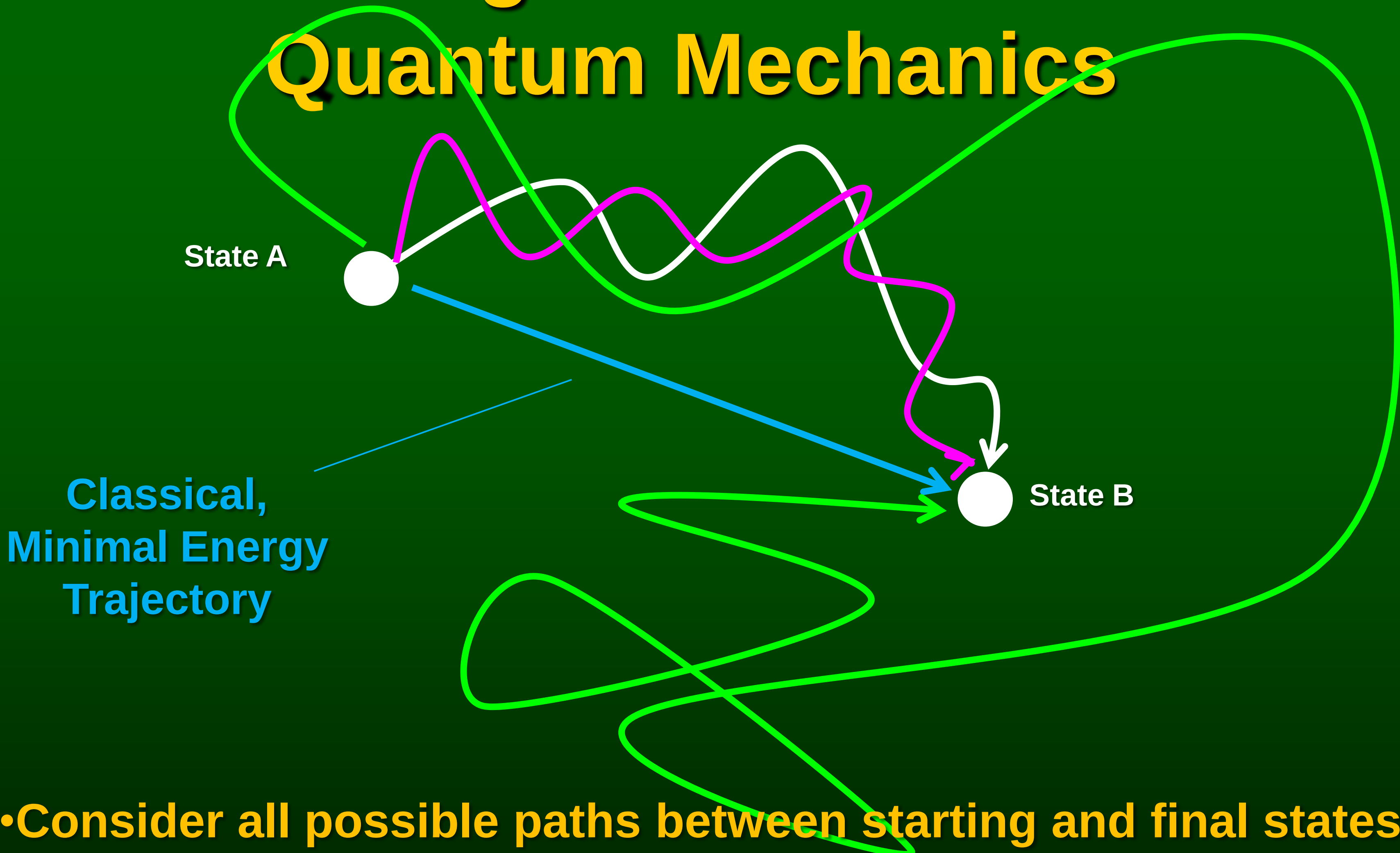
Variance seems like an intuitively appealing starting point:

low variance suggests high probability of an interaction.



Provide a means to rank individuals and predict paths for an individual.

Path Integral Formulation of Quantum Mechanics



- Consider all possible paths between starting and final states
- Weight each by a complex phase factor ~exp(i*Energy)
- Sum over all possible paths

Where are we going?

- There is still a role for biology!
- We are approaching a time in which we can begin to look at cells and organisms holistically.
- We also need to begin to think about integrating diverse data types in an intelligent way.
- This must include cross-species comparisons and inclusion of environmental effects.
- We may soon be in a position to begin development of a theoretical biology.
- Theoretical biology will require a transition from a Deterministic to a Stochastic approach.

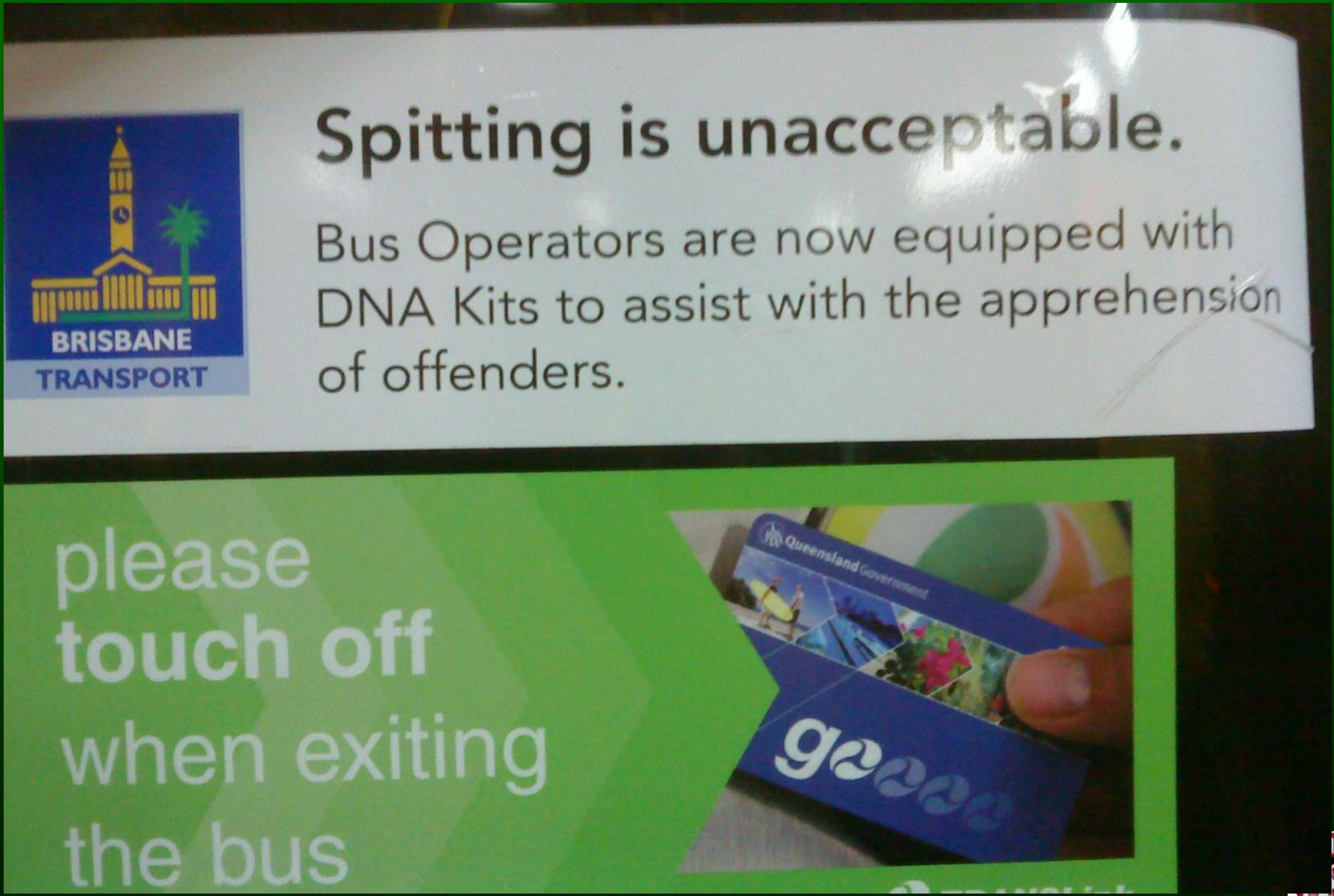
Essentially, all models are wrong, but some are useful.

– George E. Box

Before I came here I was confused about this subject.
After listening to your lecture,
I am still confused but at a higher level.

- Enrico Fermi, (1901-1954)

Genomics is here to stay



Acknowledgments

<johnq@jimmy.harvard.edu>

The Gene Index Team
Corina Antonescu
Valentin Antonescu
Fenglong Liu
Geo Pertea
Razvan Sultana
John Quackenbush

Array Software Hit Team
Katie Franklin
Eleanor Howe
Sarita Nair
Jerry Papenhausen
John Quackenbush
Dan Schlauch
Raktim Sinha
Joseph White

Eskitis Institute
Christine Wells
Alan Mackay-Sim

Center for Cancer Computational Biology
Mick Correll
Howie Goodell
Kristina Holton
Jerry Papenhausen
Patricia Papastamos
John Quackenbush
http://cccb.dfci.harvard.edu

Microarray Expression Team
Stefan Bentink
Thomas Chittenden
Aedin Culhane
Kristina Holton
Jane Pak
Renee Rubio

(Former) Stellar Students
Martin Aryee
Kaveh Maghsoudi
Jess Mar

Systems Support
Stas Alekseev, Sys Admin

Assistant
Patricia Papastamos

http://compbio.dfci.harvard.edu