Genomes and phenotypes

Wolfgang Huber
EMBL
Genome Biology Unit (Heidelberg) & EBI (Cambridge)

What makes us different?

Genome-wide genotyping of individuals for $O(10^6)$ common variants, by microarray, is a commodity.

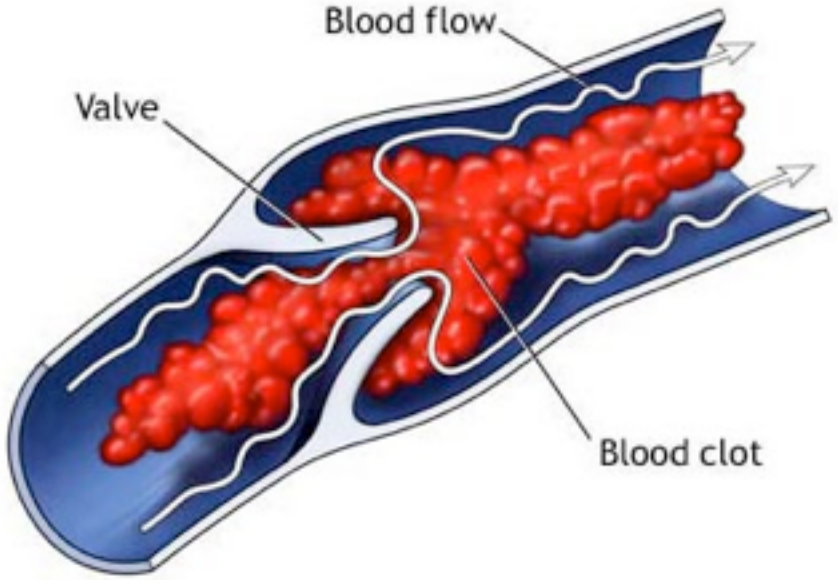
Genome sequencing also detects rare or private variants, and structural variants.

Why is that useful?

Warfarin

First use: rat and mice killer

Anticoagulant. Prevents embolism and thrombosis



IVAX
Warfarin
Tablets BP
28 tablets
1mg
IVAX
Warfarin
Tablets BP
28 tablets
3mg
IVAX
Warfarin
Tablets BP
28 tablets
5mg

Dose requirement ~ clinical & demographic variables;
VKORC1 (action)
CYP2C9 (metabolism)

Herceptin

Monoclonal antibody that interferes with the ERBB2 receptor.

Tyrosin Kinase Inhibitors

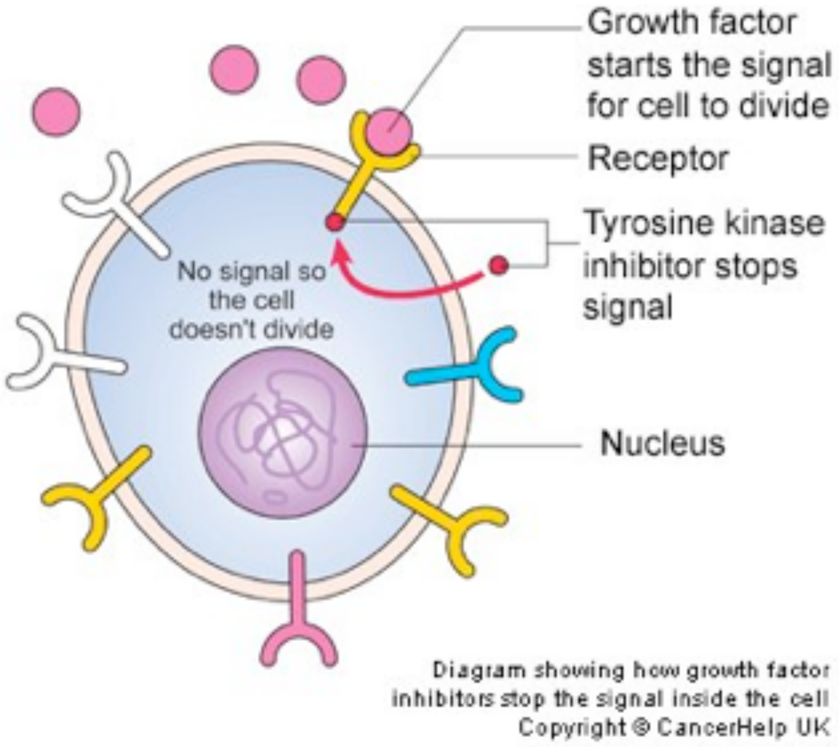

Growth factor starts the signal for cell to divide
Receptor
Tyrosine kinase inhibitor stops signal
No signal so the cell doesn't divide
Nucleus

Diagram showing how growth factor inhibitors stop the signal inside the cell
Copyright © CancerHelp UK

- Erlotinib (Tarceva)
- Imatinib (Glivec)
- Gefitinib (Iressa)
- Dasatinib (Sprycel)
- Sunitinib (Sutent)
- Nilotinib (Tasigna)
- Lapatinib (Tyverb)
- Sorafenib (Nexavr)
- Temsirolimus (Torisel)

NSCLC: resistance to TKI therapy ⇐ heterogeneity and mutational redundancy of the disease

Identify each patient’s specific ‘driver mutations’
E.g. Activation of EGFR by exon 19 deletion or exon 21 mutation ⇒ erlotinib and gefitinib

.... etc.

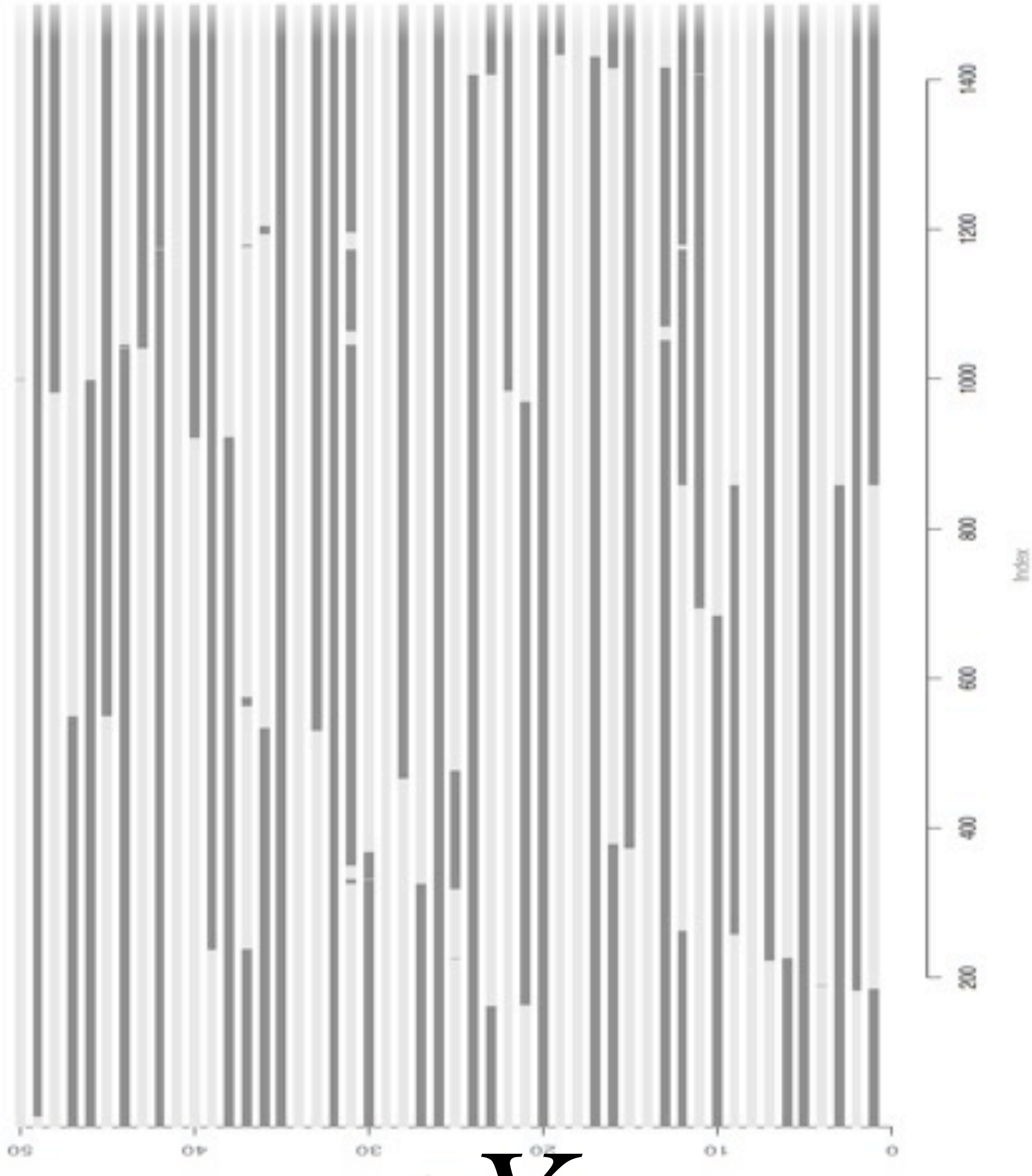
Genome-wide association studies

Y

$\sim$

← thousands of people →

↑ millions of genomic loci ↓



X

Genome wide association studies:

Identifiability - additive model with no interactions

Finding important variables (loci): impressive

Prediction performance, effect sizes: poor

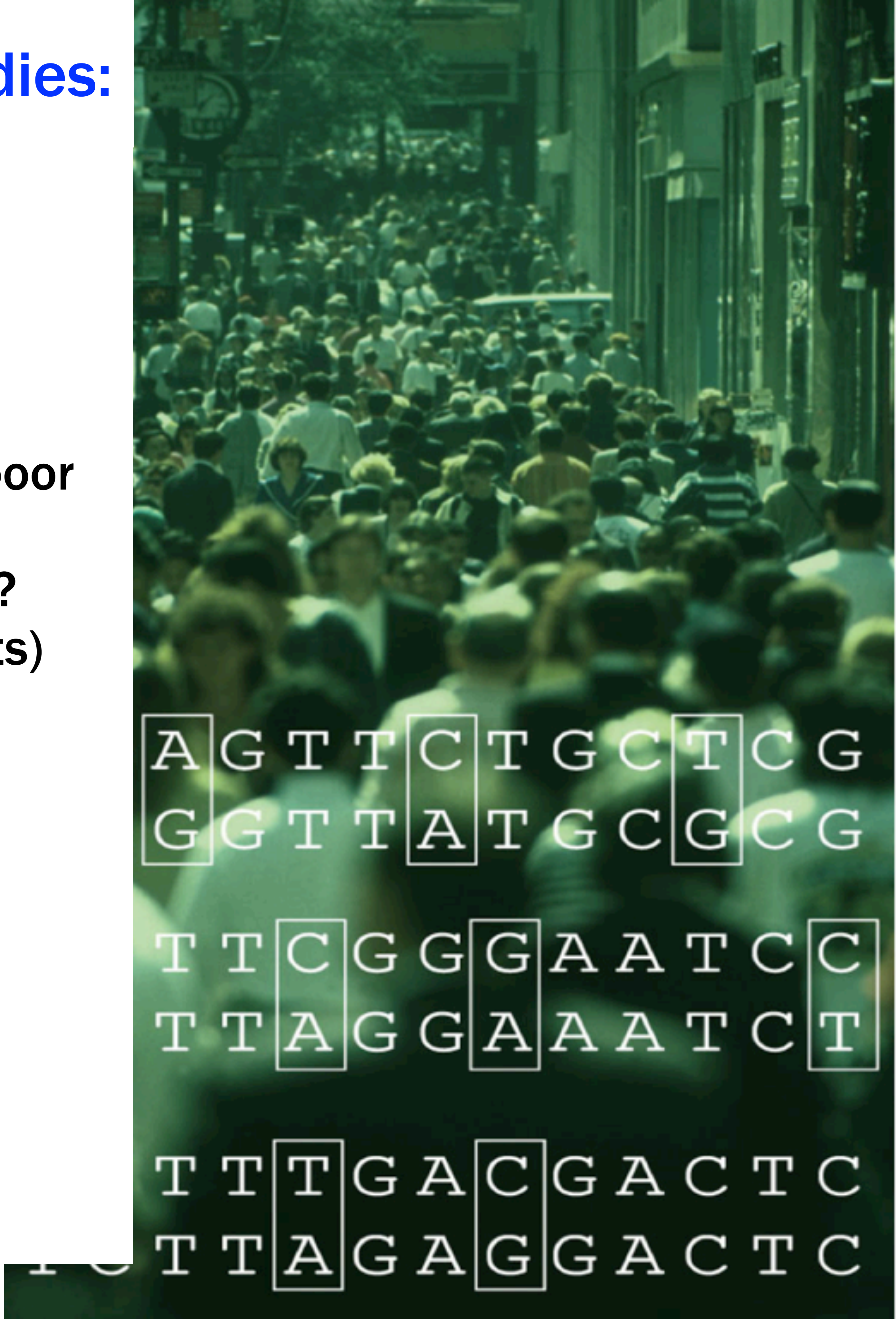
Genome wide association studies:

Identifiability - additive model with no interactions

Finding important variables (loci): impressive

Prediction performance, effect sizes: poor

- Have we missed important variables? (rare polymorphisms, structural variants)



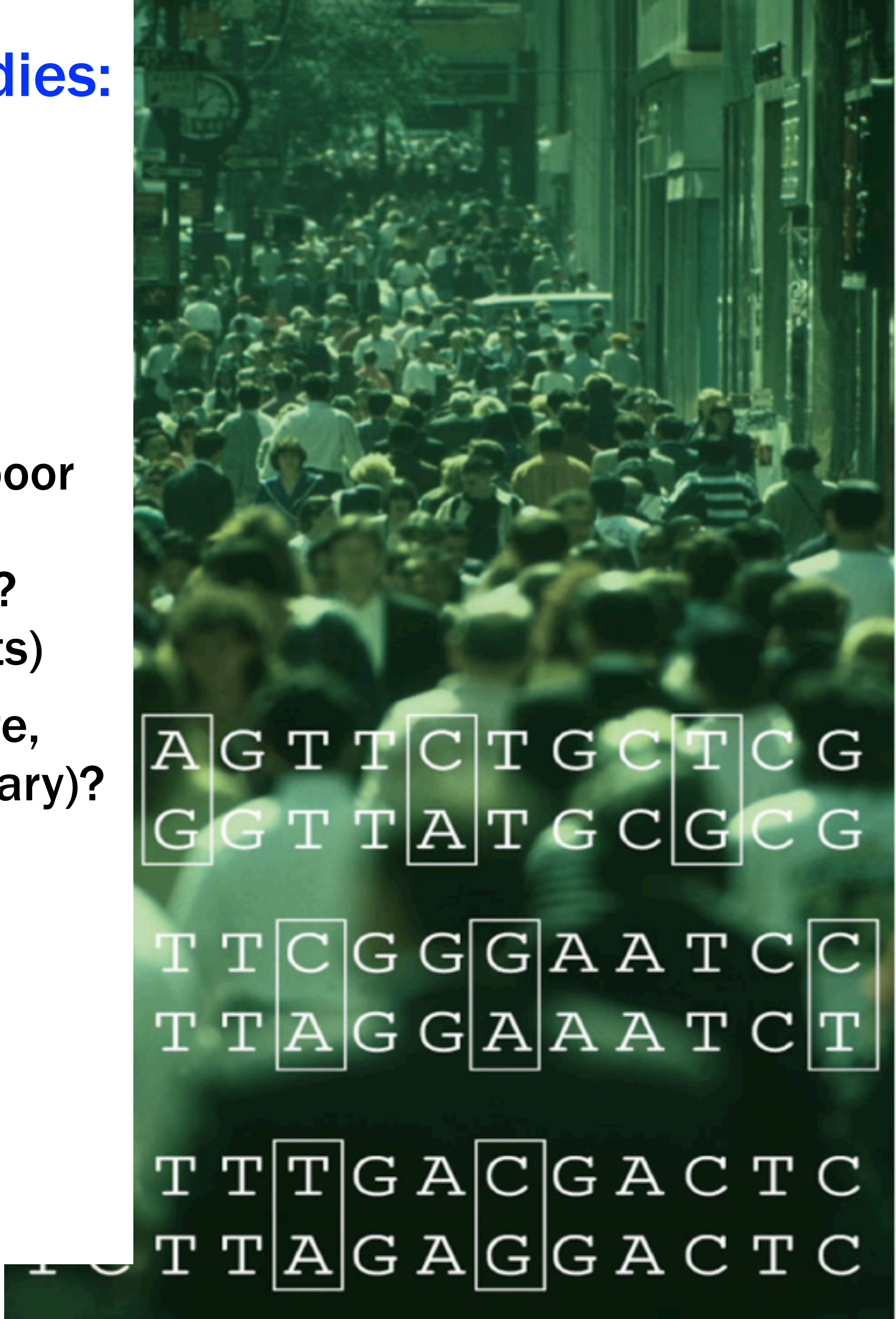
Genome wide association studies:

Identifiability - additive model with no interactions

Finding important variables (loci): impressive

Prediction performance, effect sizes: poor

- Have we missed important variables? (rare polymorphisms, structural variants)
- Are we overlooking variables with rare, strong effect (sufficient but not necessary)?



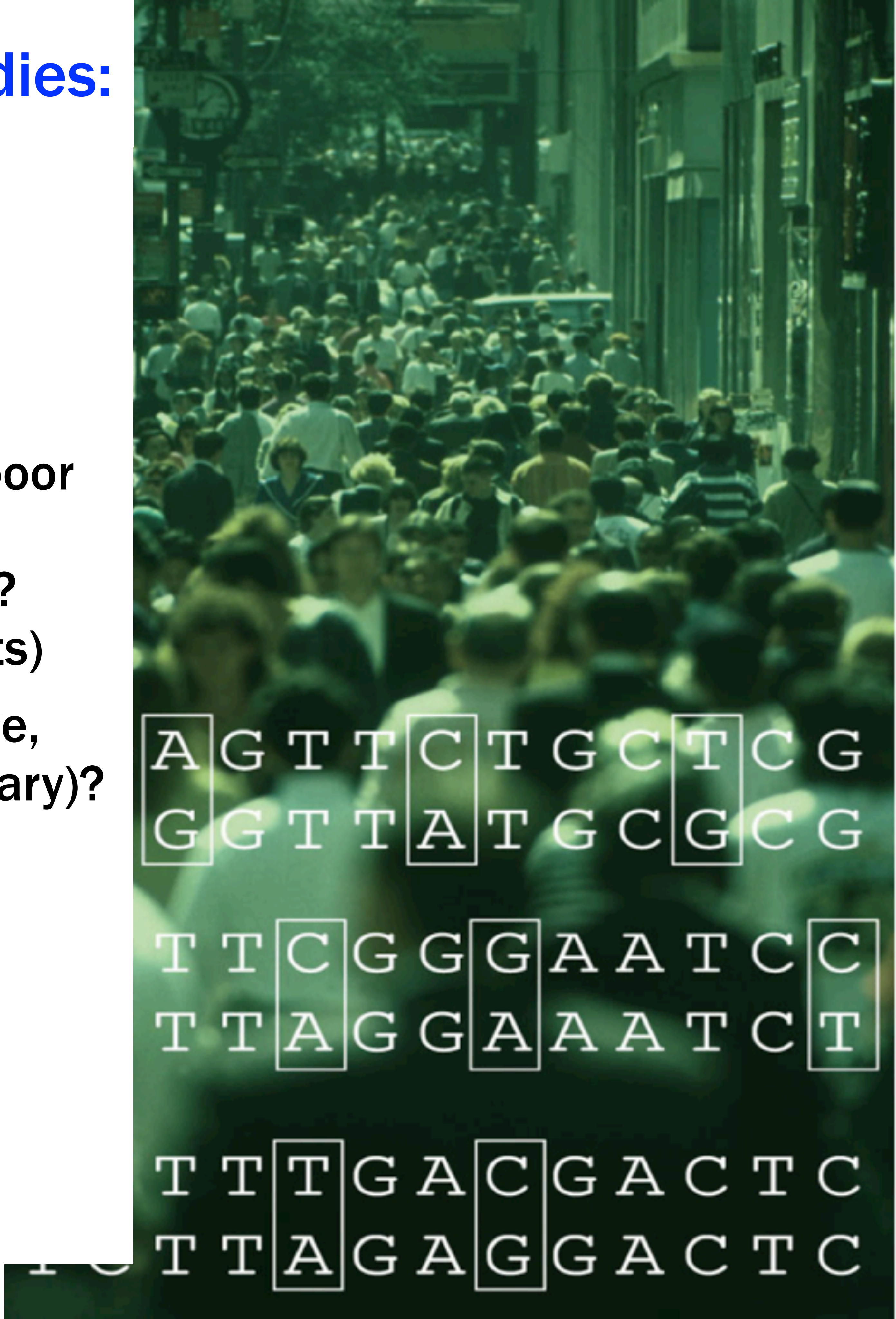
Genome wide association studies:

Identifiability - additive model with no interactions

Finding important variables (loci): impressive

Prediction performance, effect sizes: poor

- Have we missed important variables? (rare polymorphisms, structural variants)
- Are we overlooking variables with rare, strong effect (sufficient but not necessary)?
- Interactions (epistasis)



Genome wide association studies:

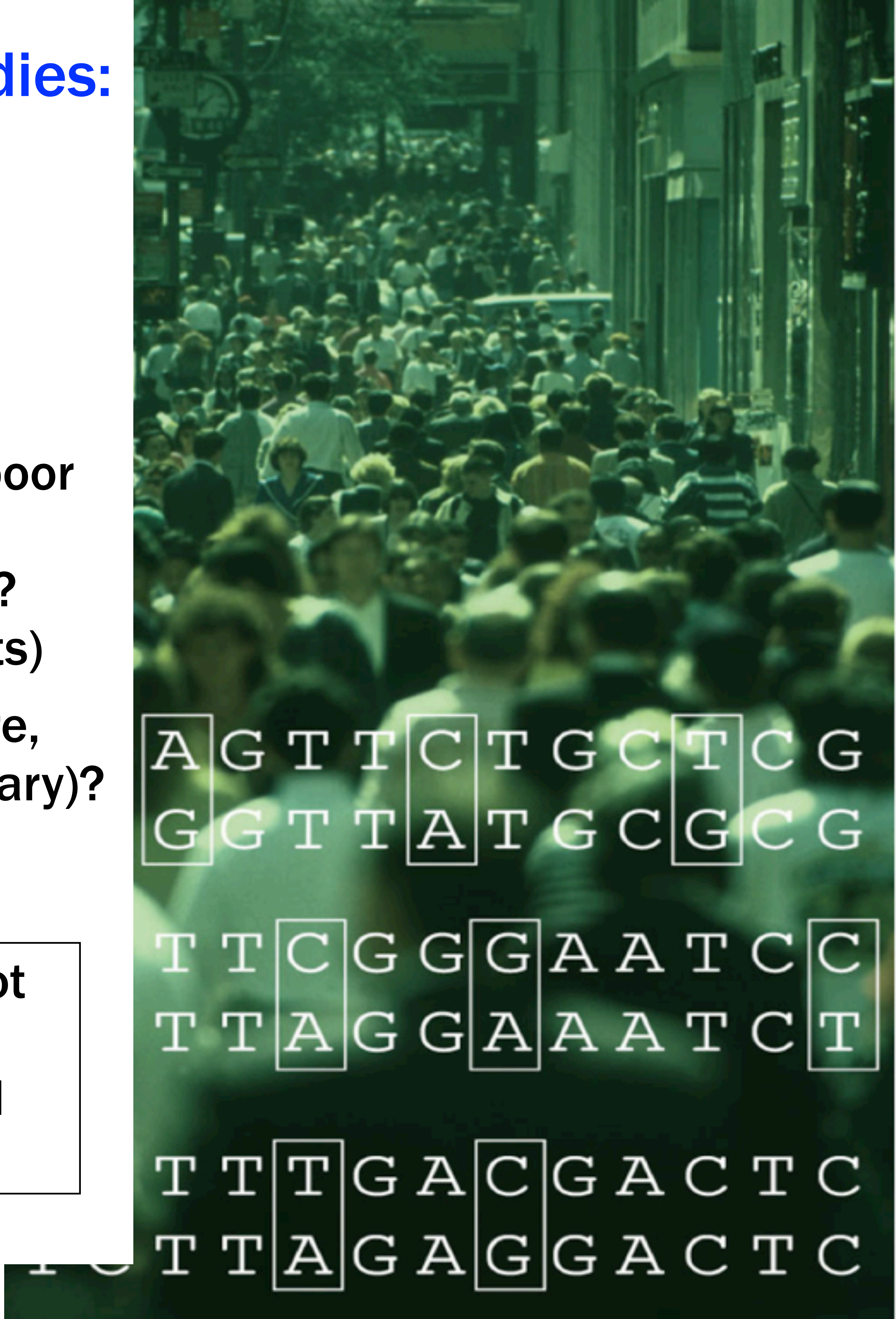
Identifiability - additive model with no interactions

Finding important variables (loci): impressive

Prediction performance, effect sizes: poor

- Have we missed important variables? (rare polymorphisms, structural variants)
- Are we overlooking variables with rare, strong effect (sufficient but not necessary)?
- Interactions (epistasis)

Association based approaches do not have enough power - we need perturbation experiments on model systems



Forward genetics

Nature Vol. 287 30 October 1980 **797**

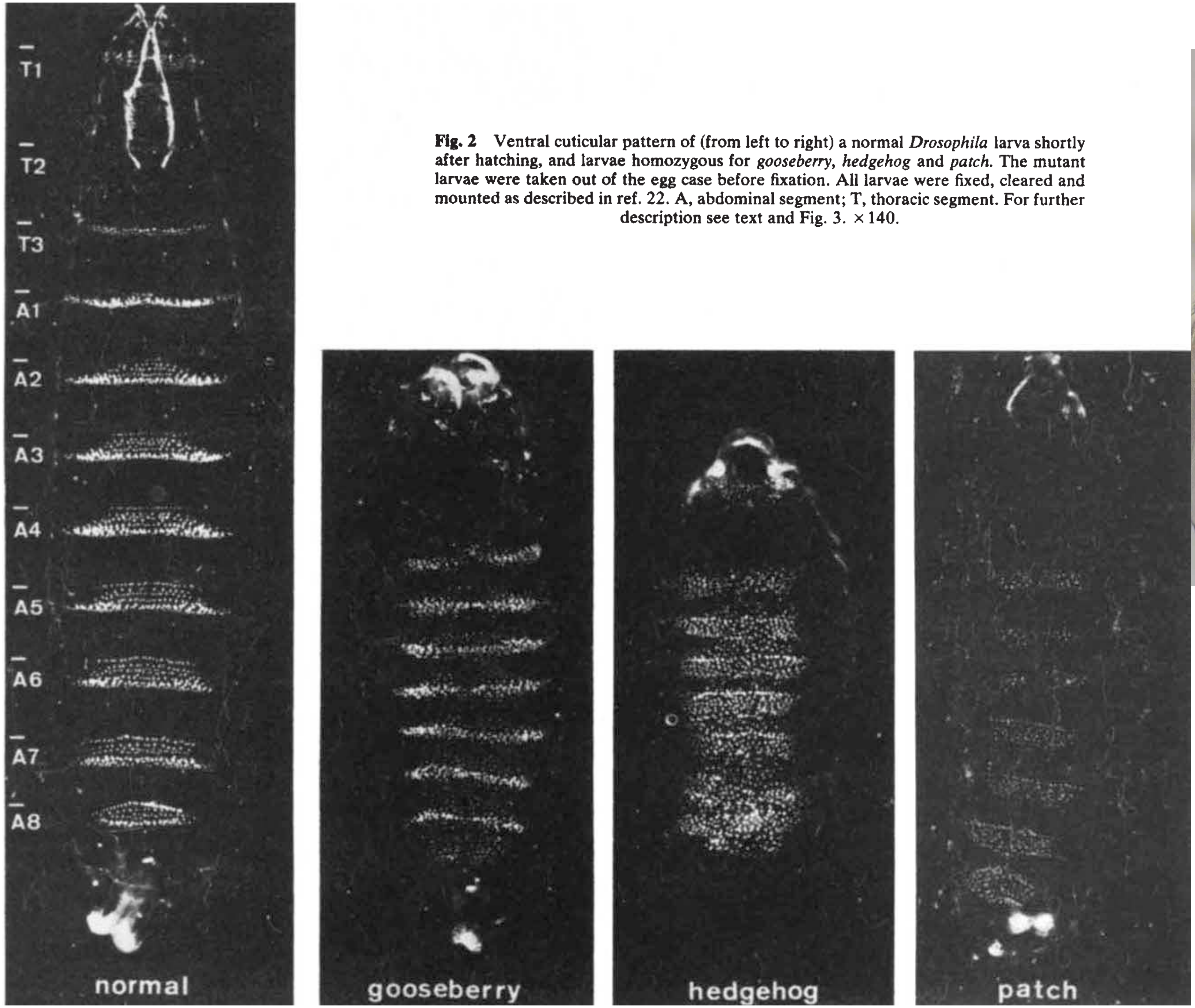
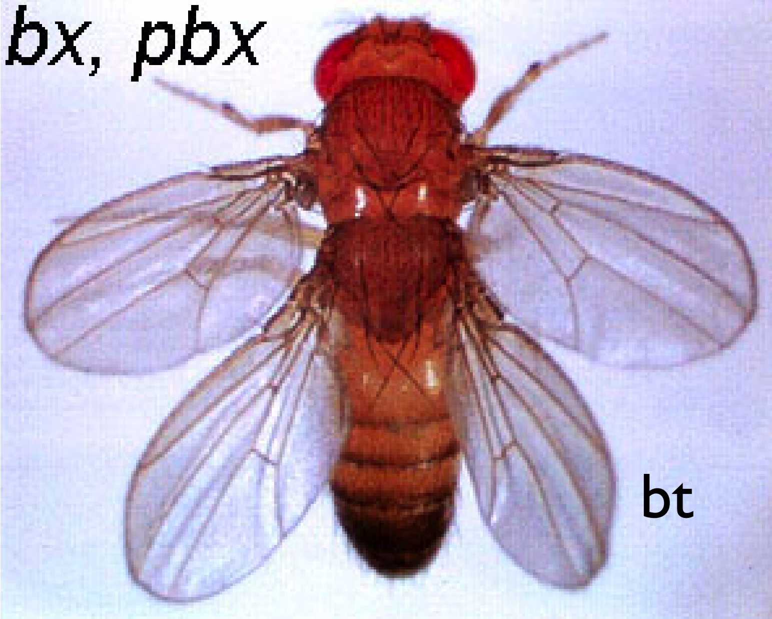
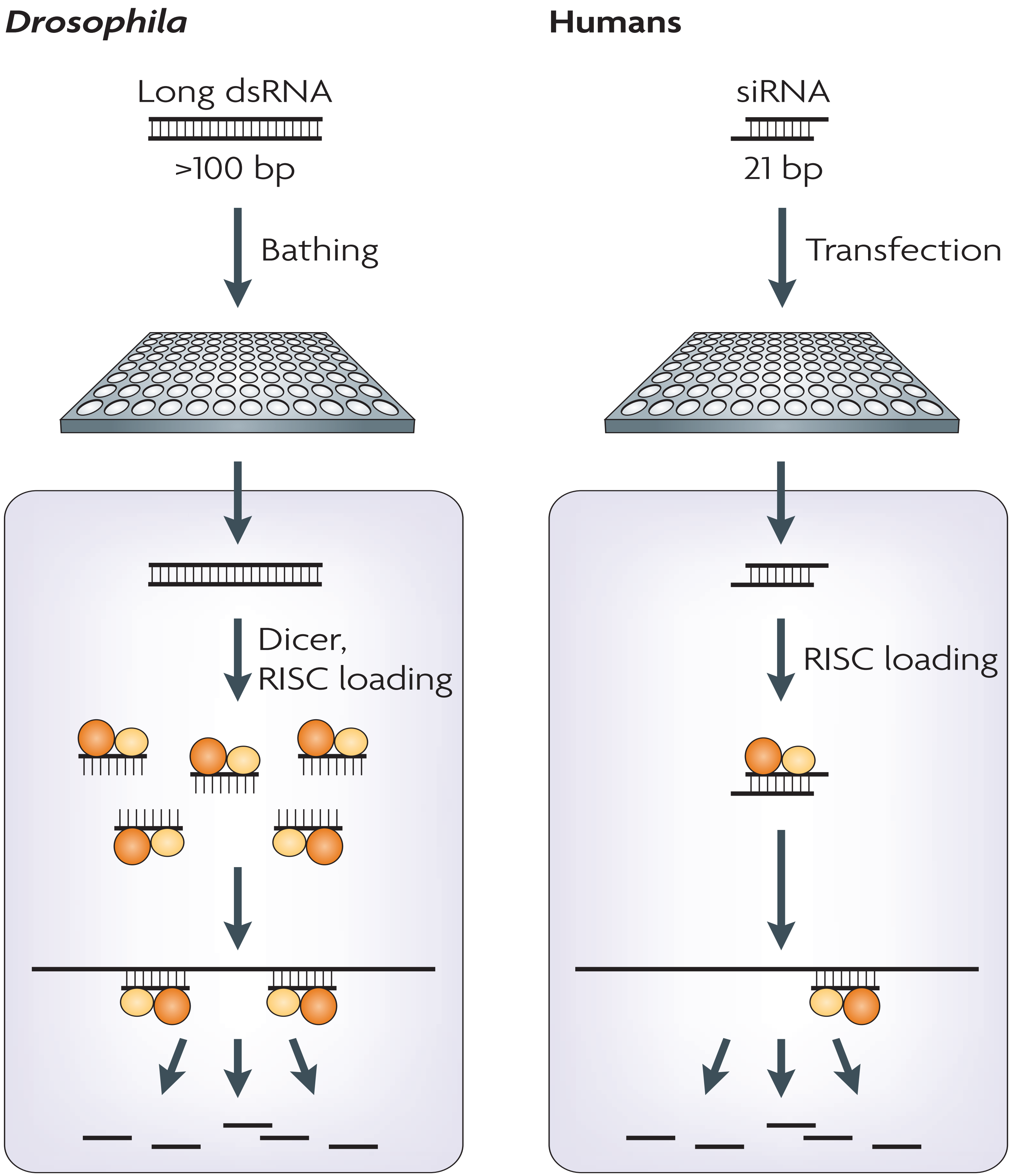


Fig. 2 Ventral cuticular pattern of (from left to right) a normal *Drosophila* larva shortly after hatching, and larvae homozygous for *gooseberry*, *hedgehog* and *patch*. The mutant larvae were taken out of the egg case before fixation. All larvae were fixed, cleared and mounted as described in ref. 22. A, abdominal segment; T, thoracic segment. For further description see text and Fig. 3. × 140.



RNAi: targeted depletion of a specific gene's products (mRNA)



Genome-wide "libraries"

Specificity

Efficiency

Reproducibility

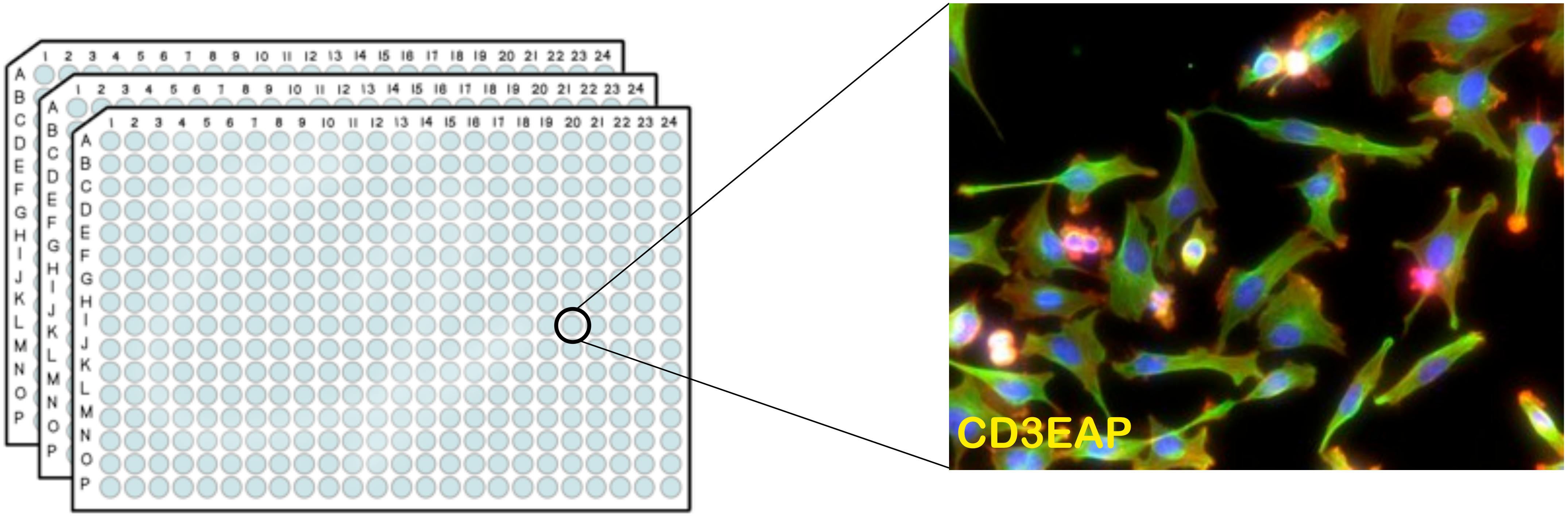
What do human cells do when you knock down each gene in turn?

with F. Fuchs, C. Budjan, Michael Boutros (DKFZ)

Genomewide RNAi library (Dharmacon, 22k siRNA-pools)

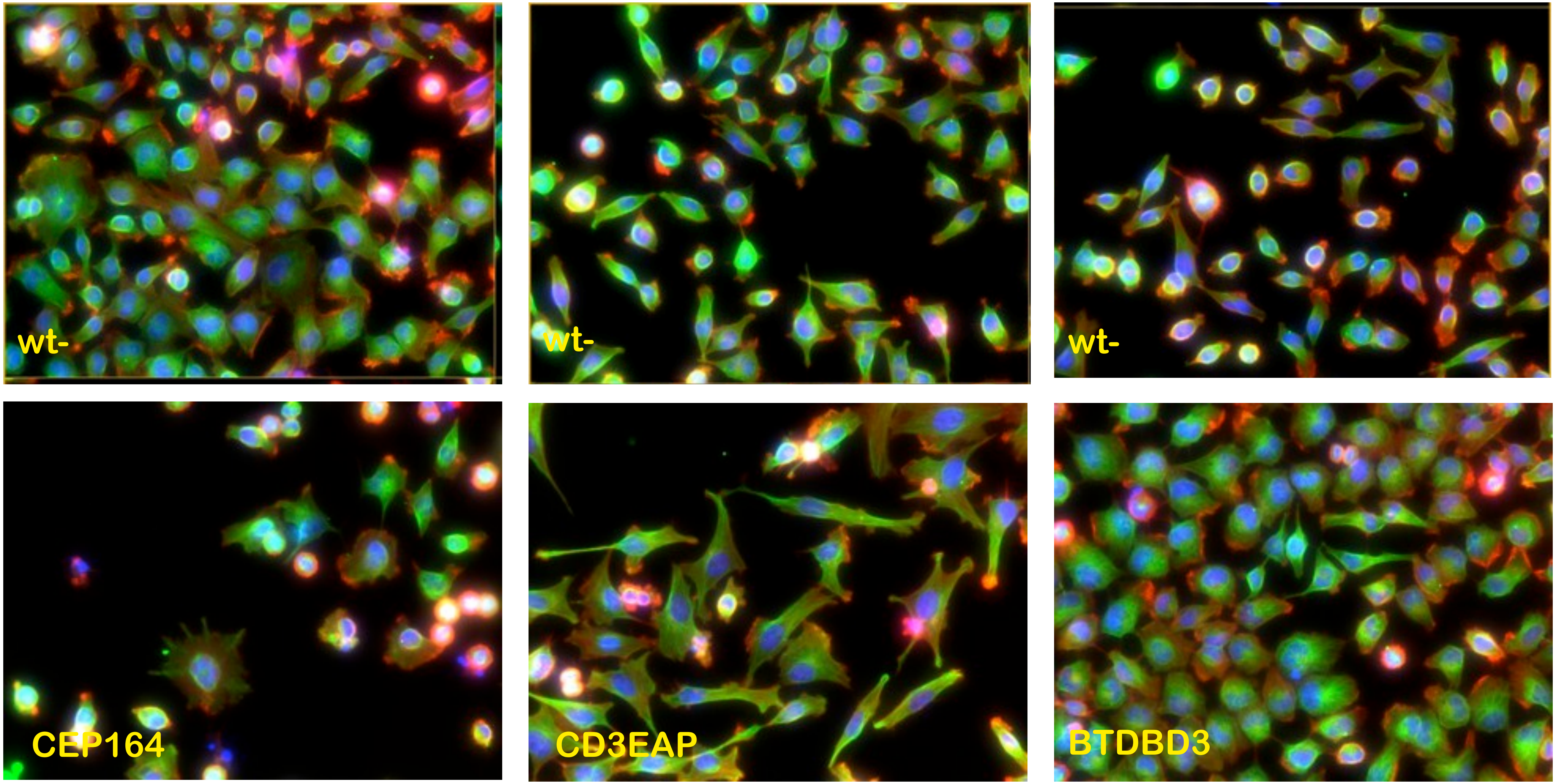
HeLa cells, incubated 48h, then fixed and stained

Microscopy readout: DNA (DAPI), tubulin (Alexa), actin (TRITC)



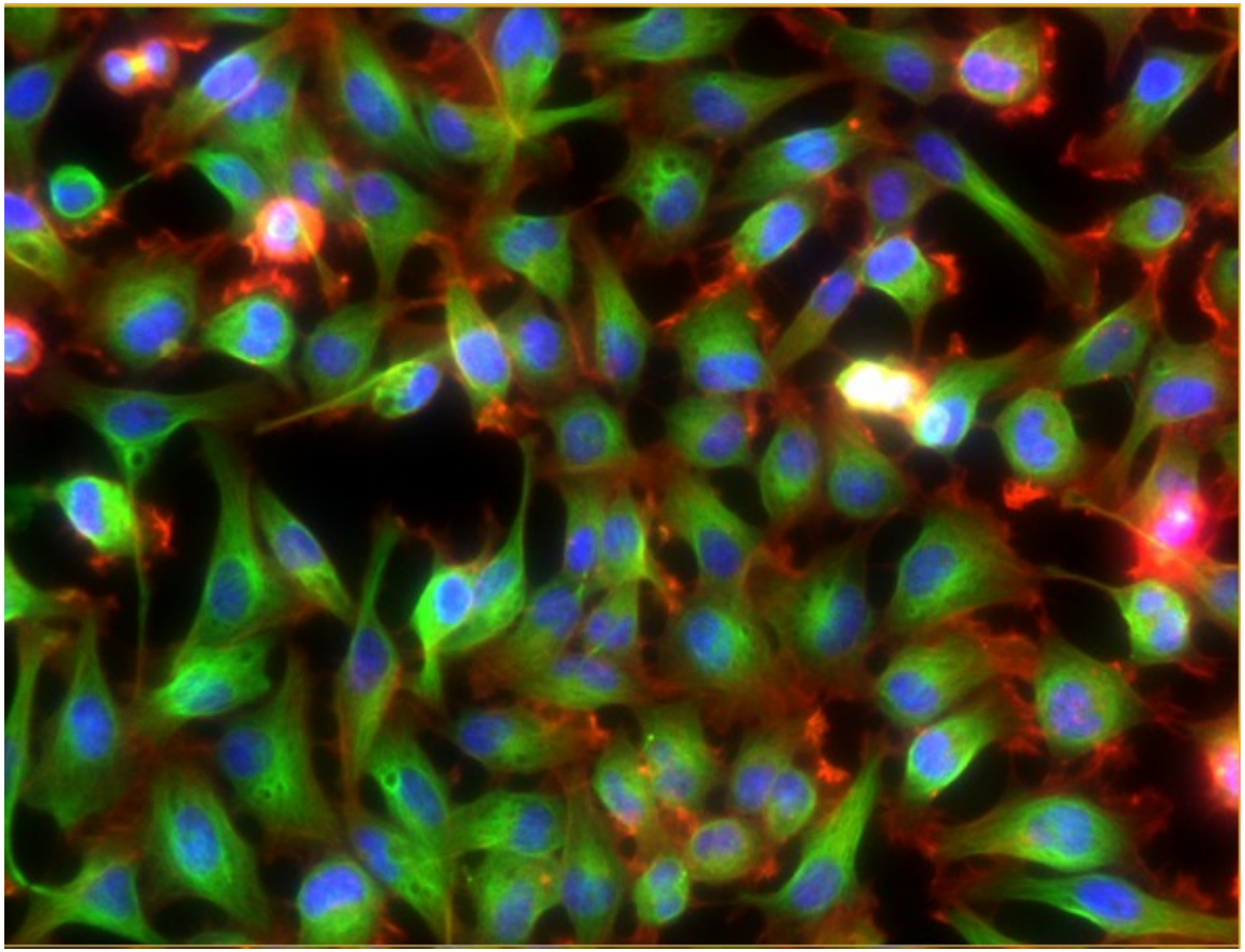
Molecular Systems Biology, 2010

RNAi perturbation phenotypes are observed by automated microscopy



22839 wells, 4 images per well
each with DNA, tubulin, actin (1344 x 1024 pixel at 3 x 12 bit)

Segmentation

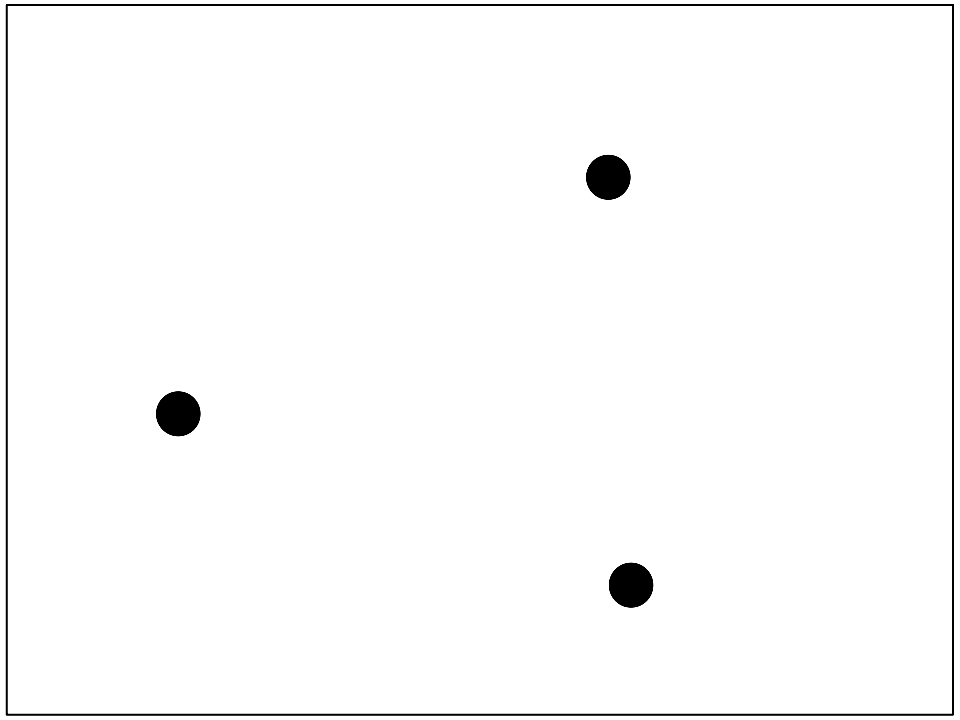


Nuclei are easy (e.g. locally adaptive threshold)

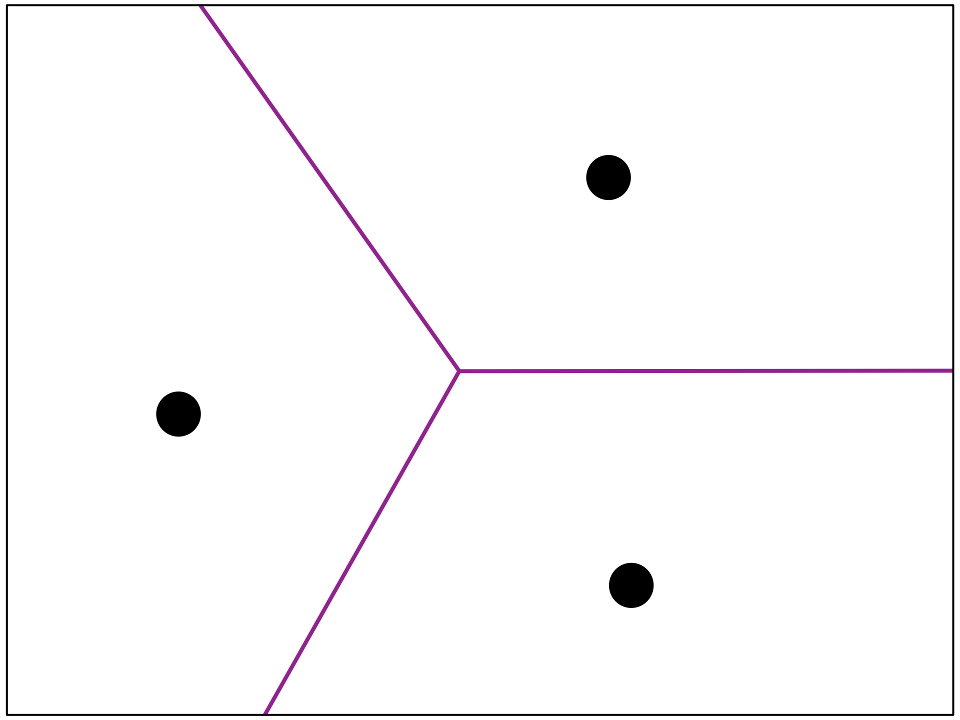
But cells touch.

How do you draw reasonable boundaries between cells?

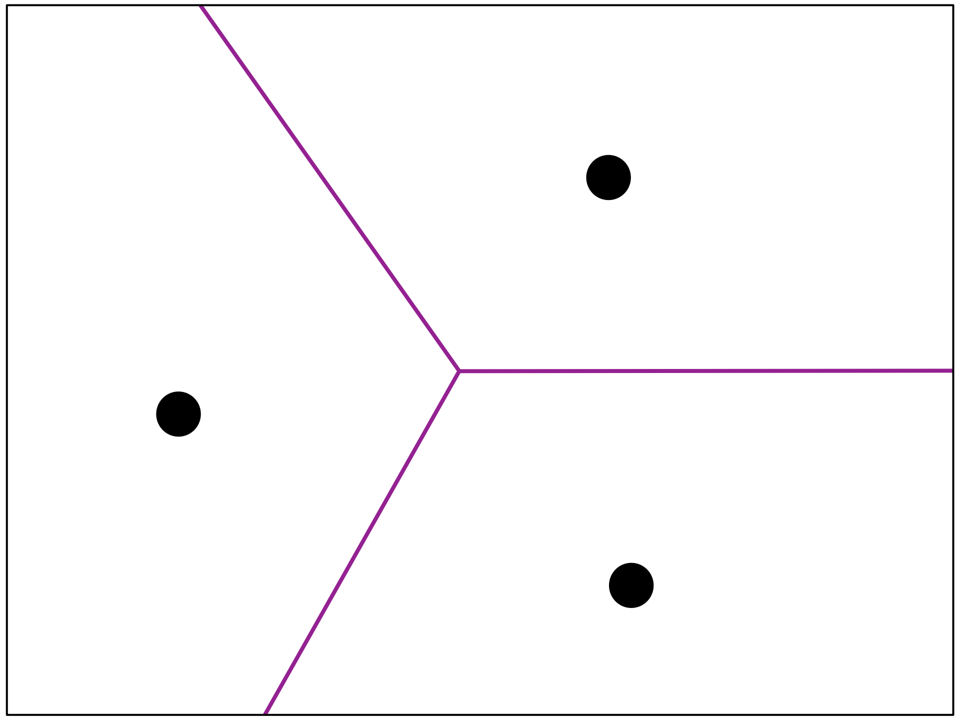
Voronoi segmentation

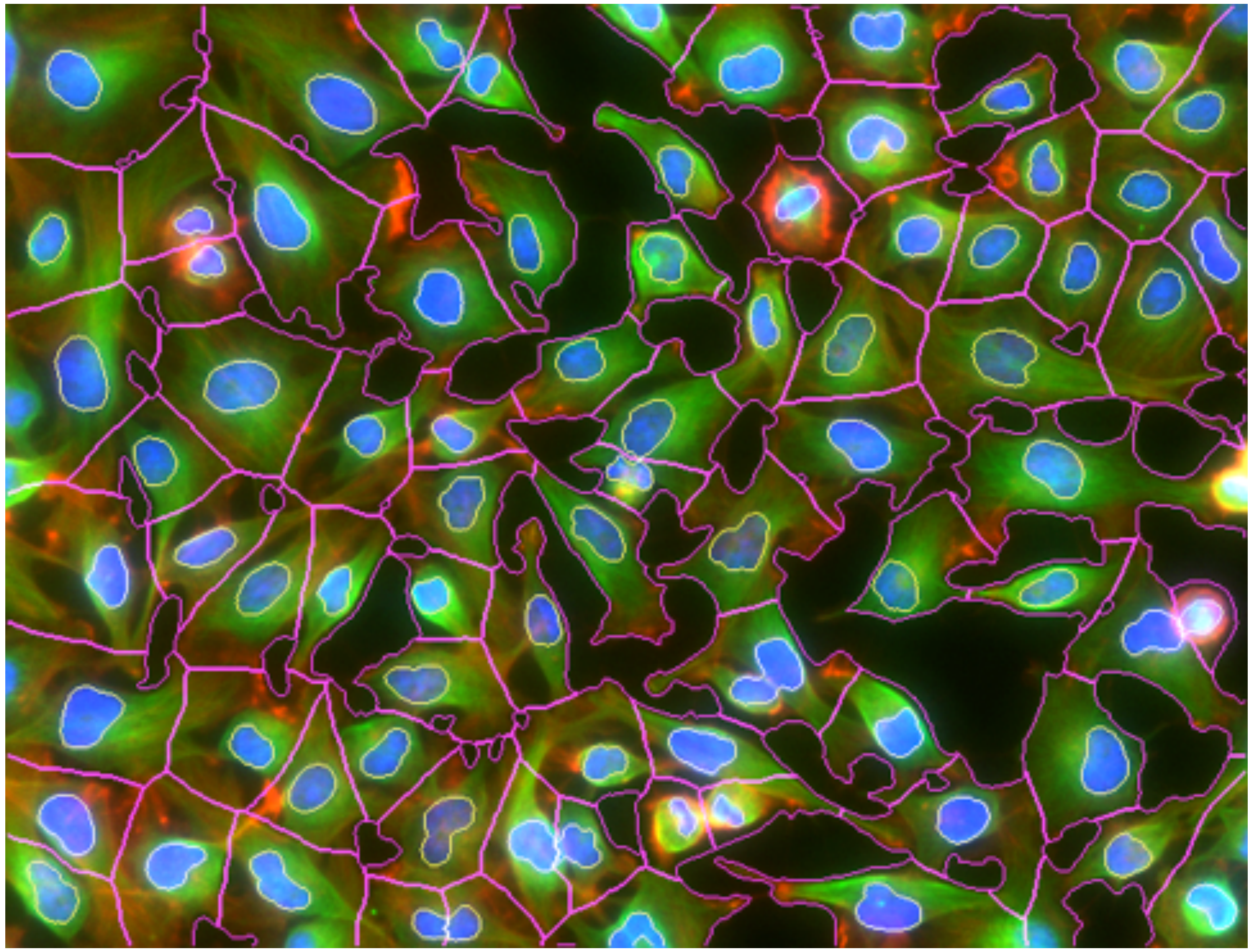


Voronoi segmentation

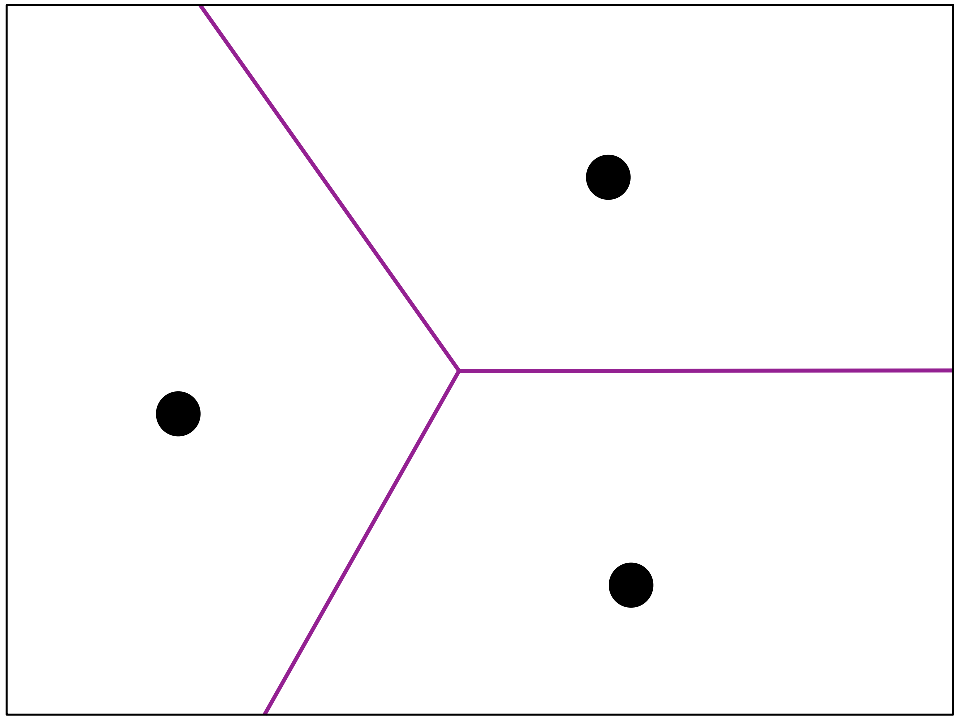


Voronoi segmentation

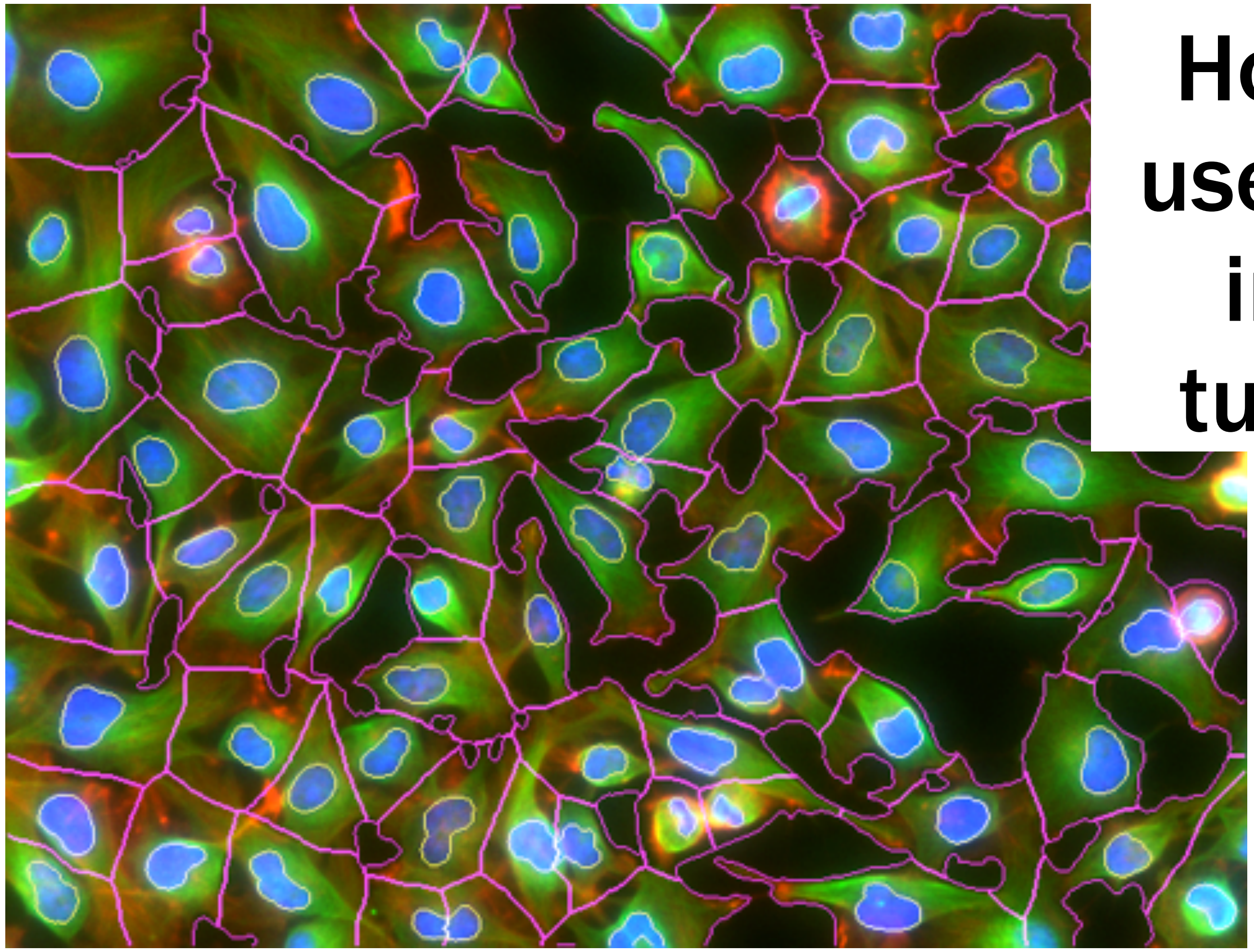




Voronoi segmentation



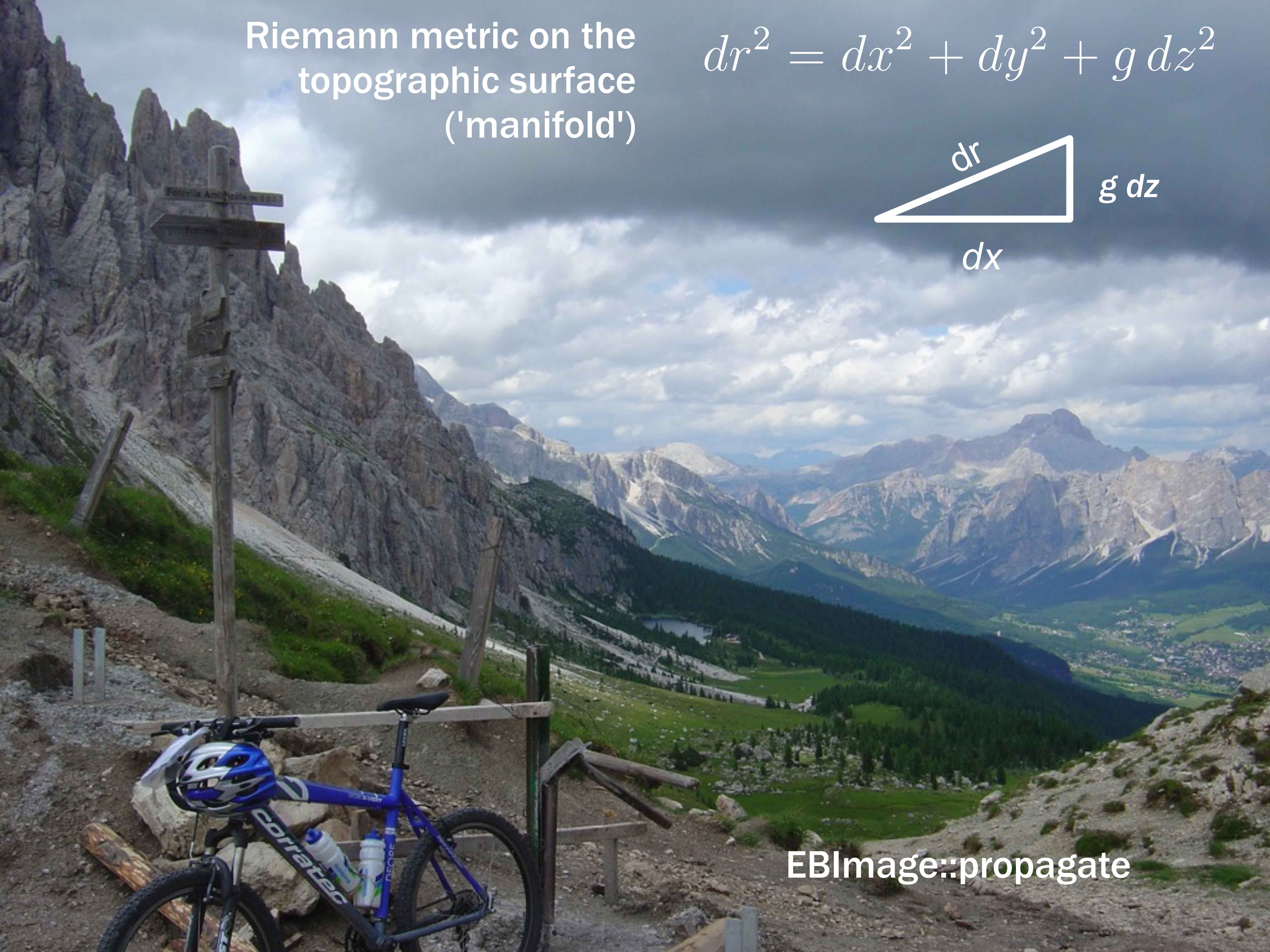
But we only used the nuclei. The boundaries are artificially straight.



How can we better use the information in the actin and tubulin channels?

Riemann metric on the topographic surface ('manifold')

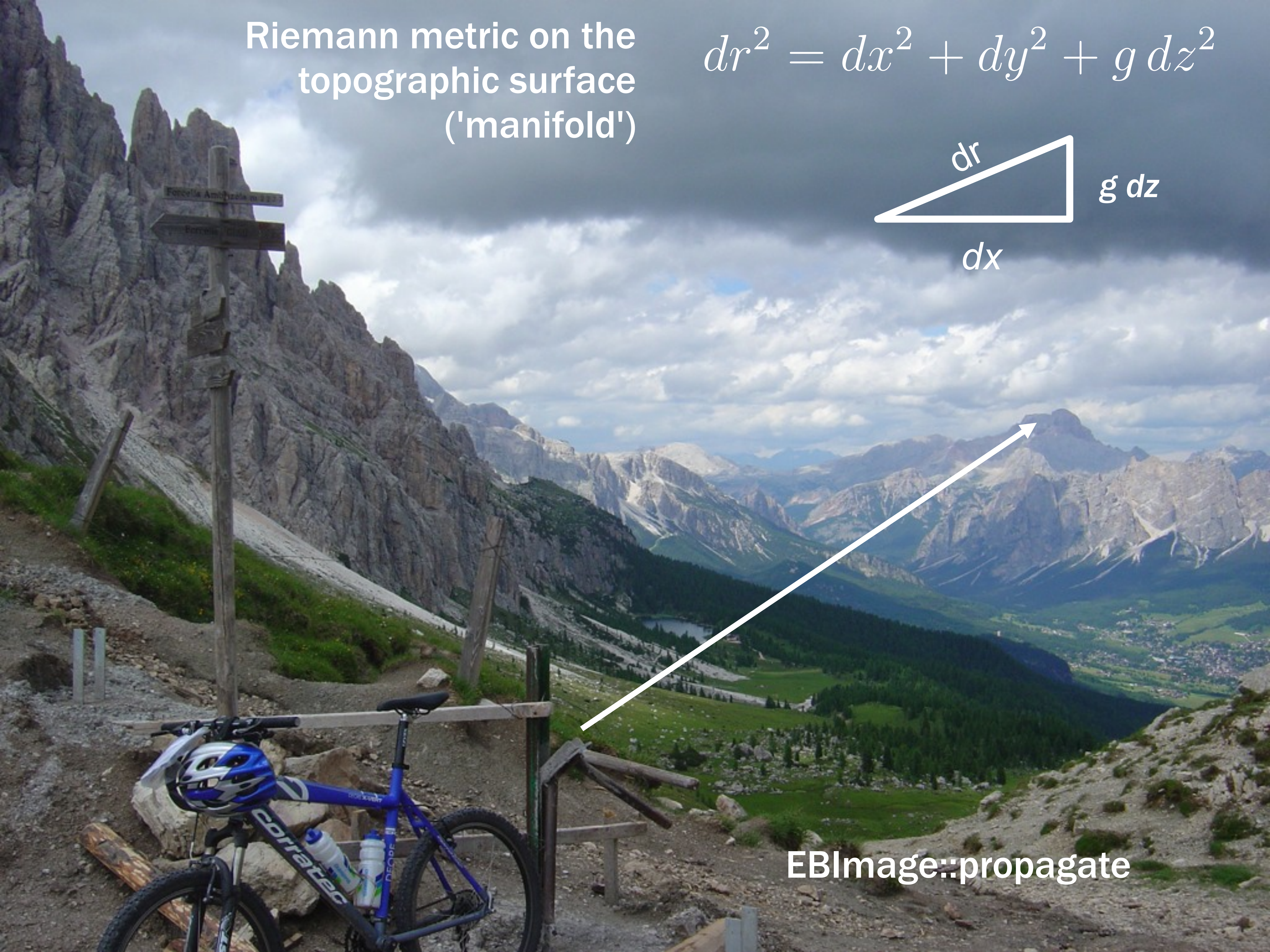
$$dr^2 = dx^2 + dy^2 + g\,dz^2$$



EBImage::propagate

Riemann metric on the topographic surface ('manifold')

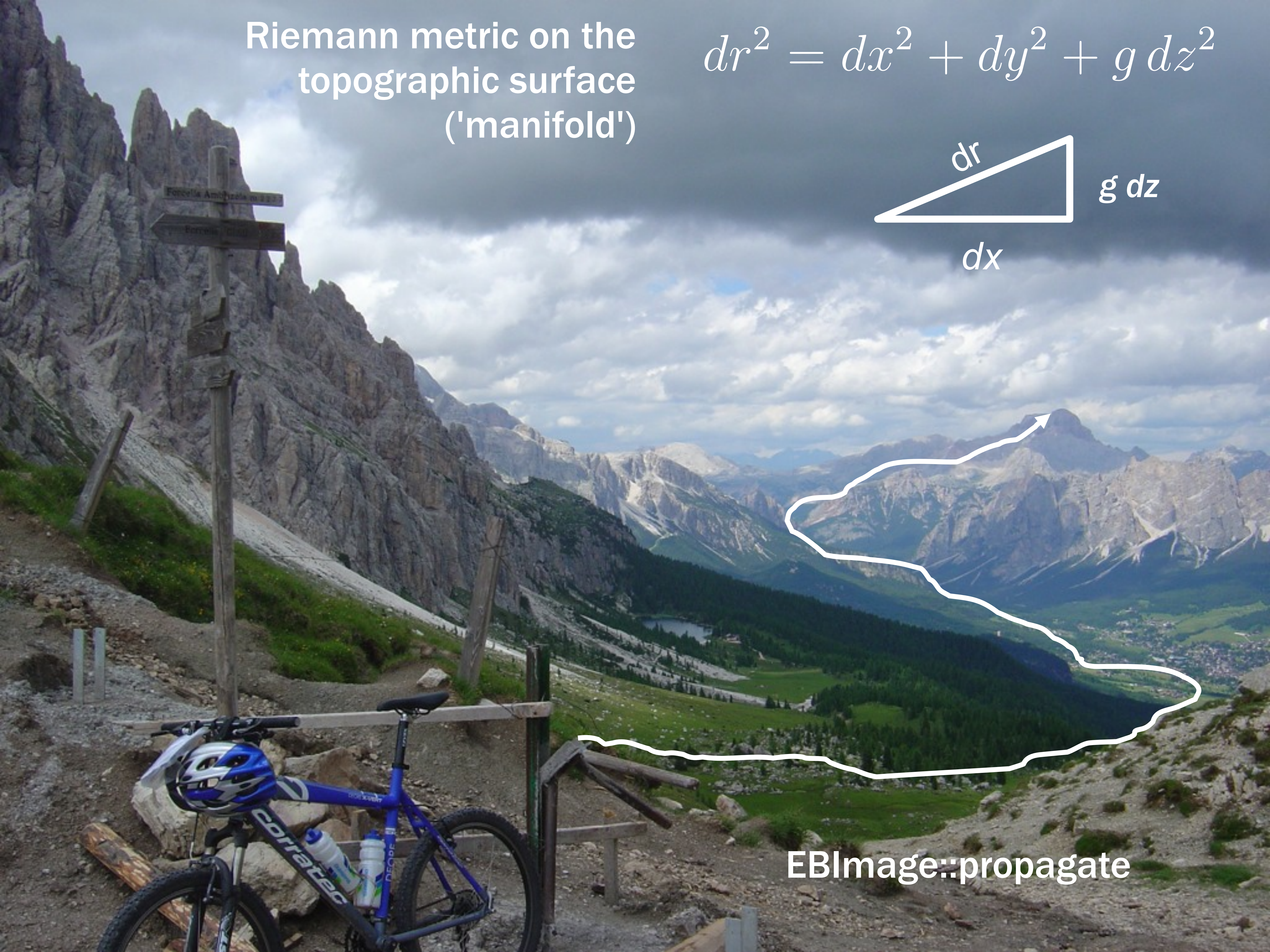
$$dr^2 = dx^2 + dy^2 + g\,dz^2$$



EBImage::propagate

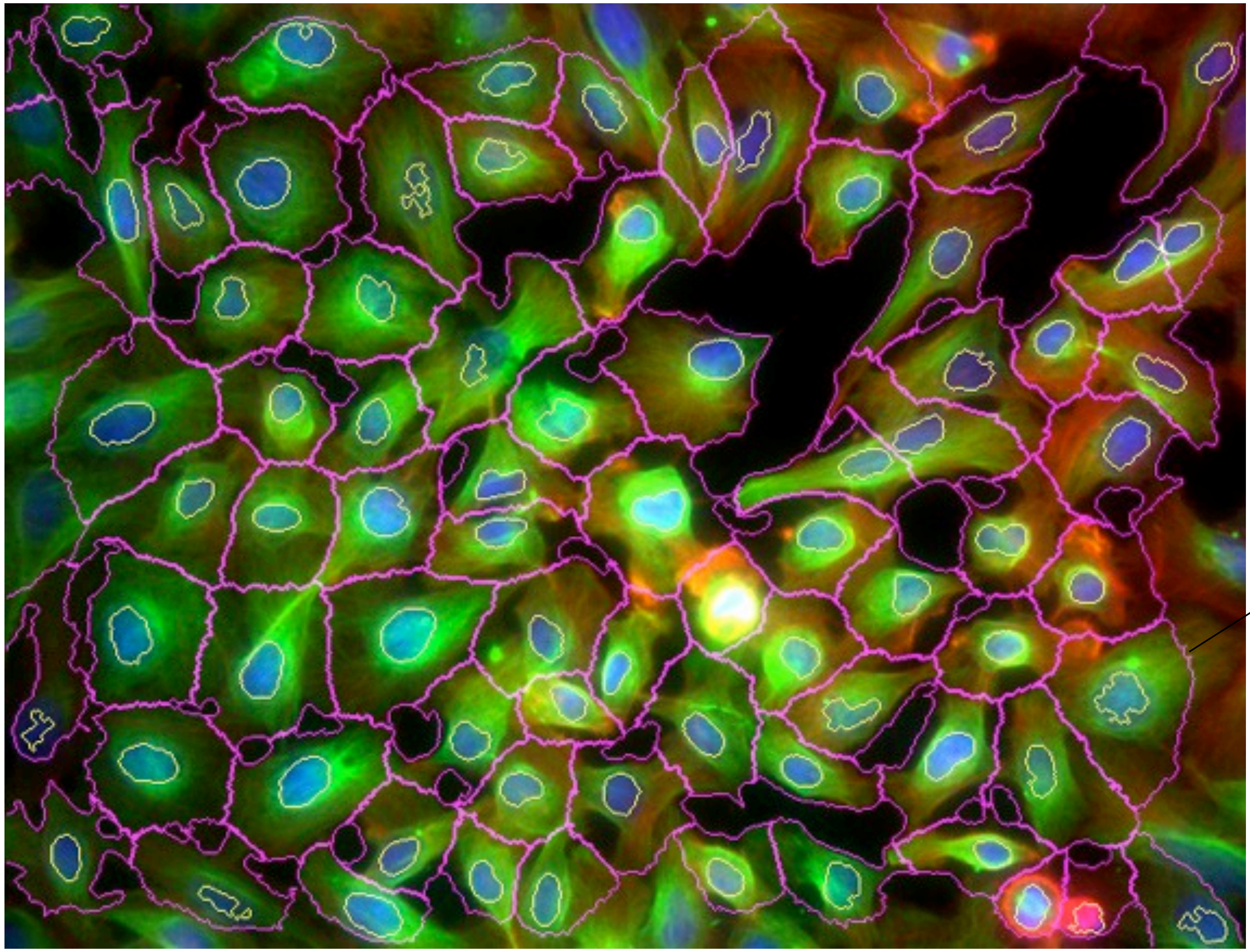
Riemann metric on the topographic surface ('manifold')

$$dr^2 = dx^2 + dy^2 + g\,dz^2$$



EBImage::propagate

Converting images into quantitative features

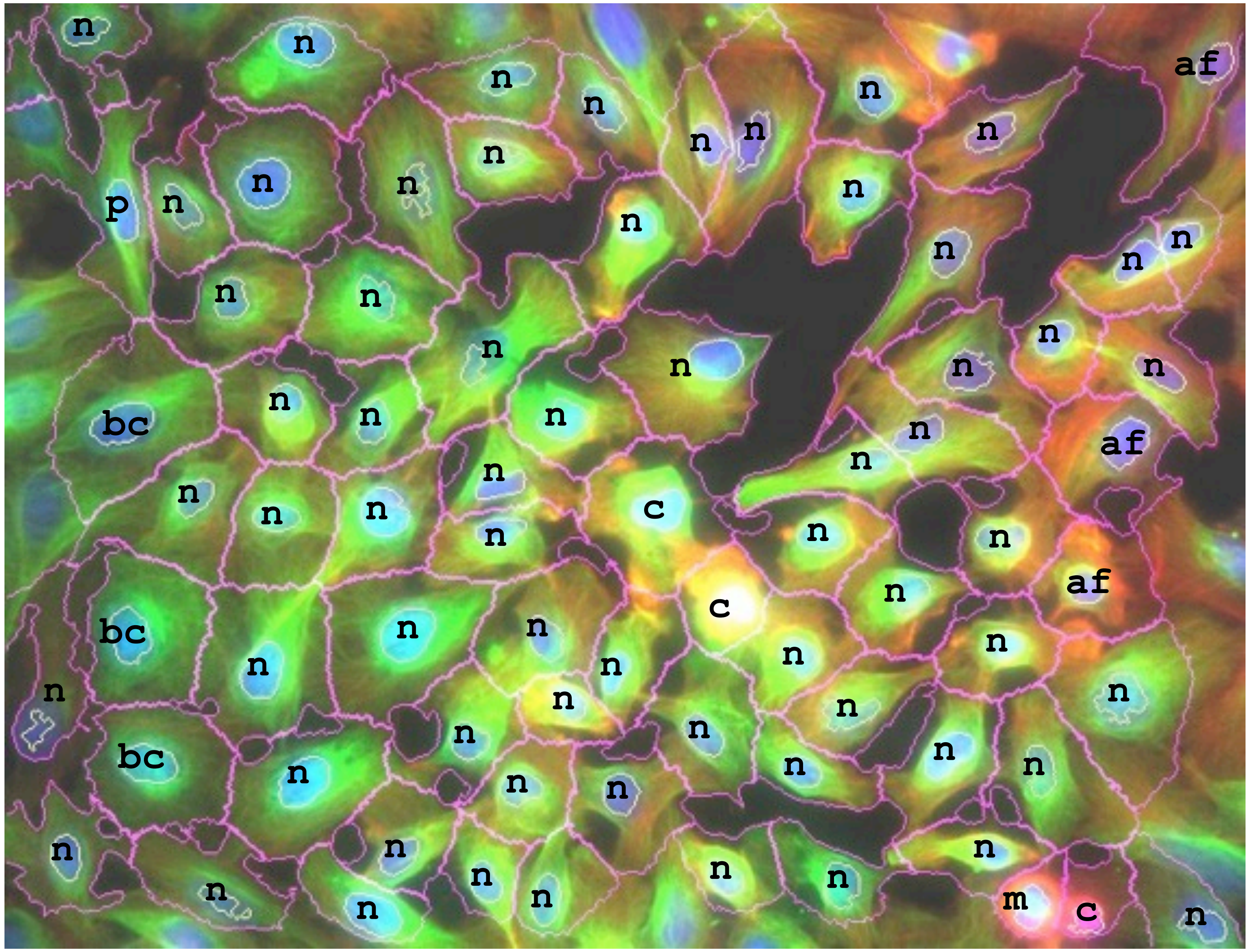


cell size	289
cell intensity	34.33118
eccentricity	0.472934
nucleus size	2857.356
DNA content	485.2710
actin content	0.828876
tubulin content	0.098647
actin F11	0.049594
actin F12	0.081746
actin F21	0.158817
actin F22	0.179339
tubulin F11	0.009249
tubulin F12	0.219697
...	...

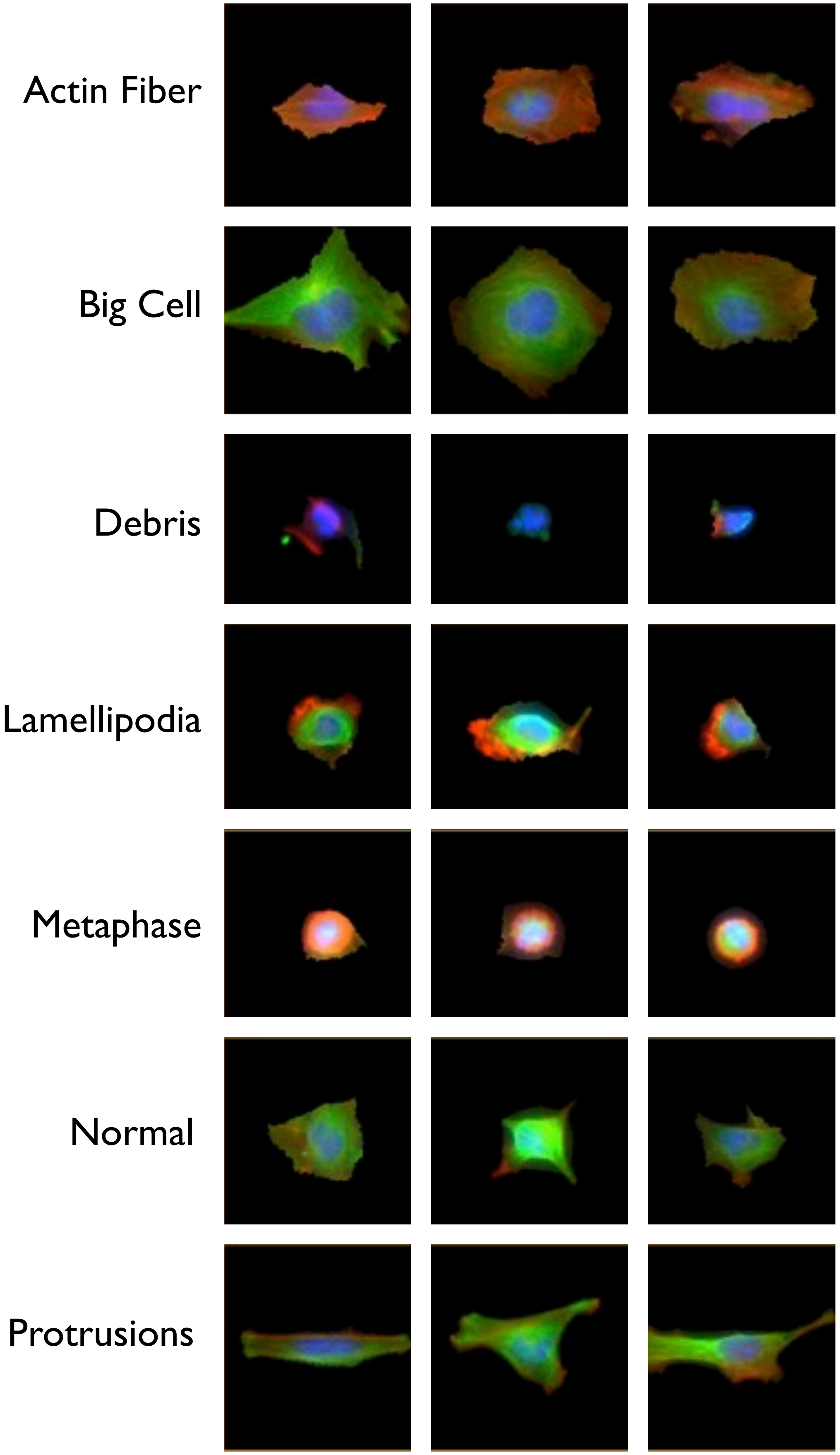
178 features per cell

EBImage::computeFeatures

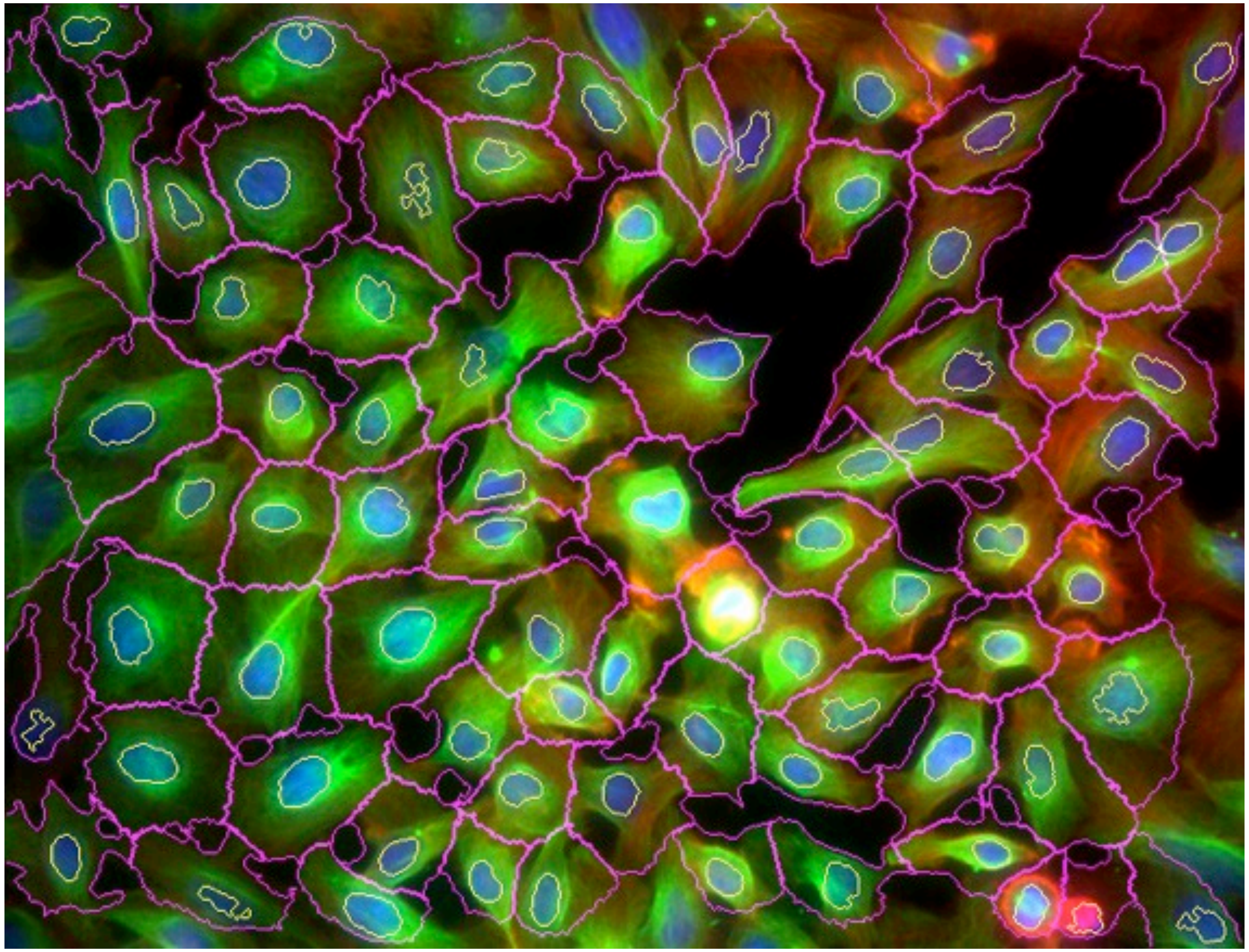
Cells are classified into predefined classes



178 features per cell
Radial-kernel SVM
Manually annotated training set of ~3000 cells
Accuracy: ~ 90 %



The image is now represented by a 13-dim vector: “phenotypic profile”



n	289
ext	34.33118
ecc	0.472934
Next	2857.356
Nint	485.2710
AtoTint	0.828876
NtoATsz	0.098647
AF %	0.049594
BC %	0.081746
C %	0.158817
M %	0.179339
LA %	0.009249
P %	0.219697

How do you measure distance and similarity in a 13-dimensional phenotypic profile space?

Similarity depends on the choice and weighting of descriptors

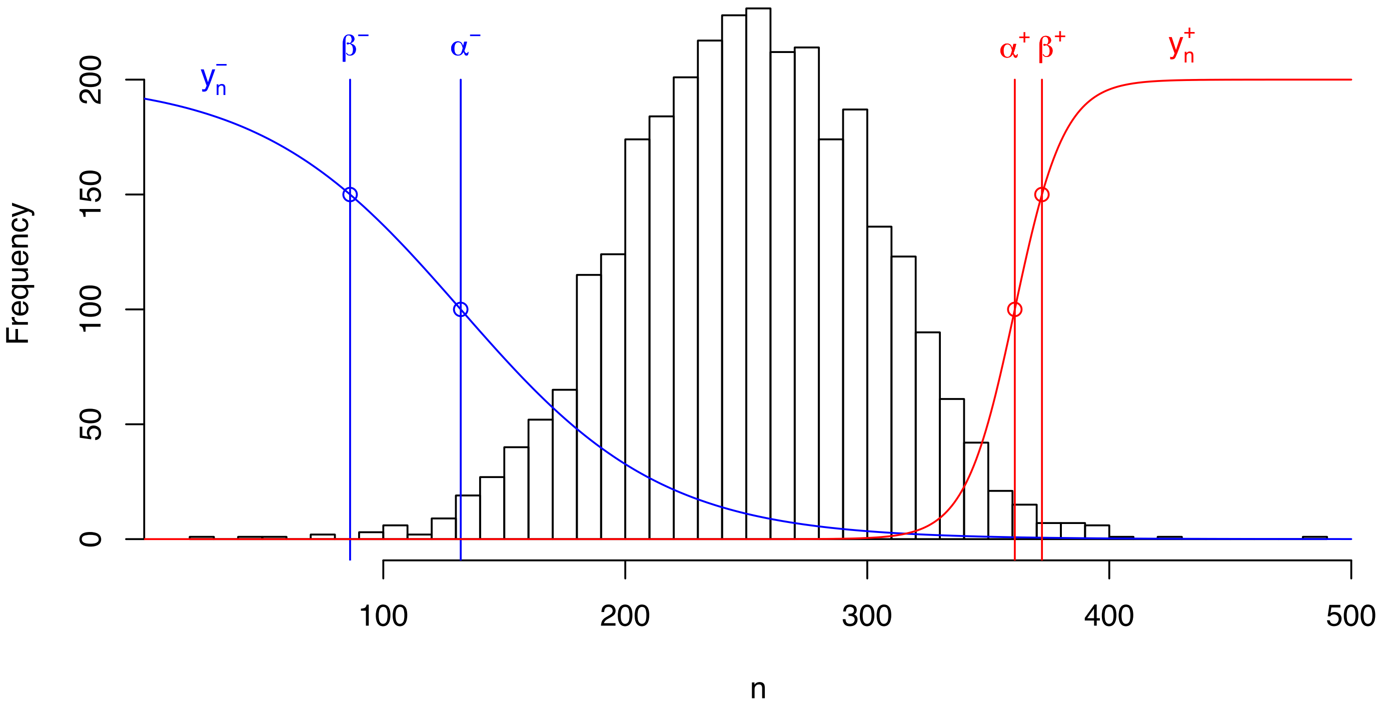
Distance metric learning

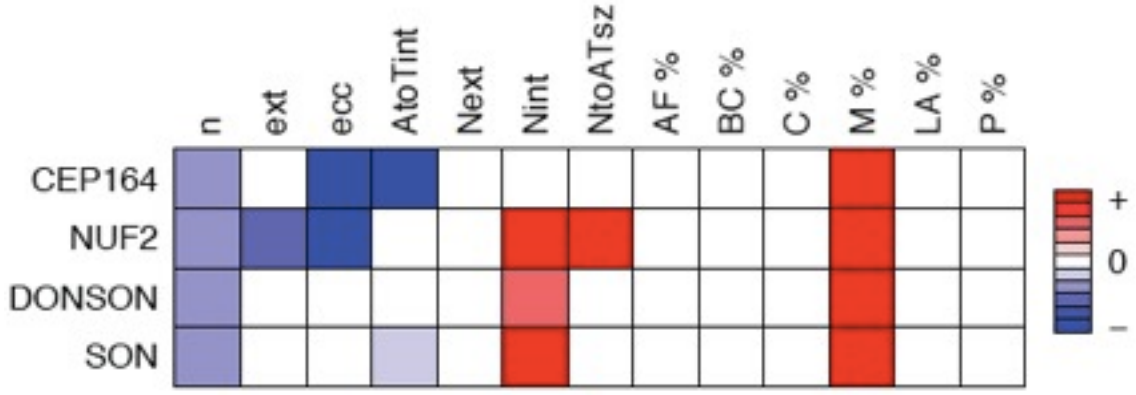
$$d(x,y) = \sum_k |f_k(x_k) - f_k(y_k)|$$

$$f_k(x) = \frac{1}{1 + \exp\left(-\eta_k(x-\alpha_k)\right)}$$

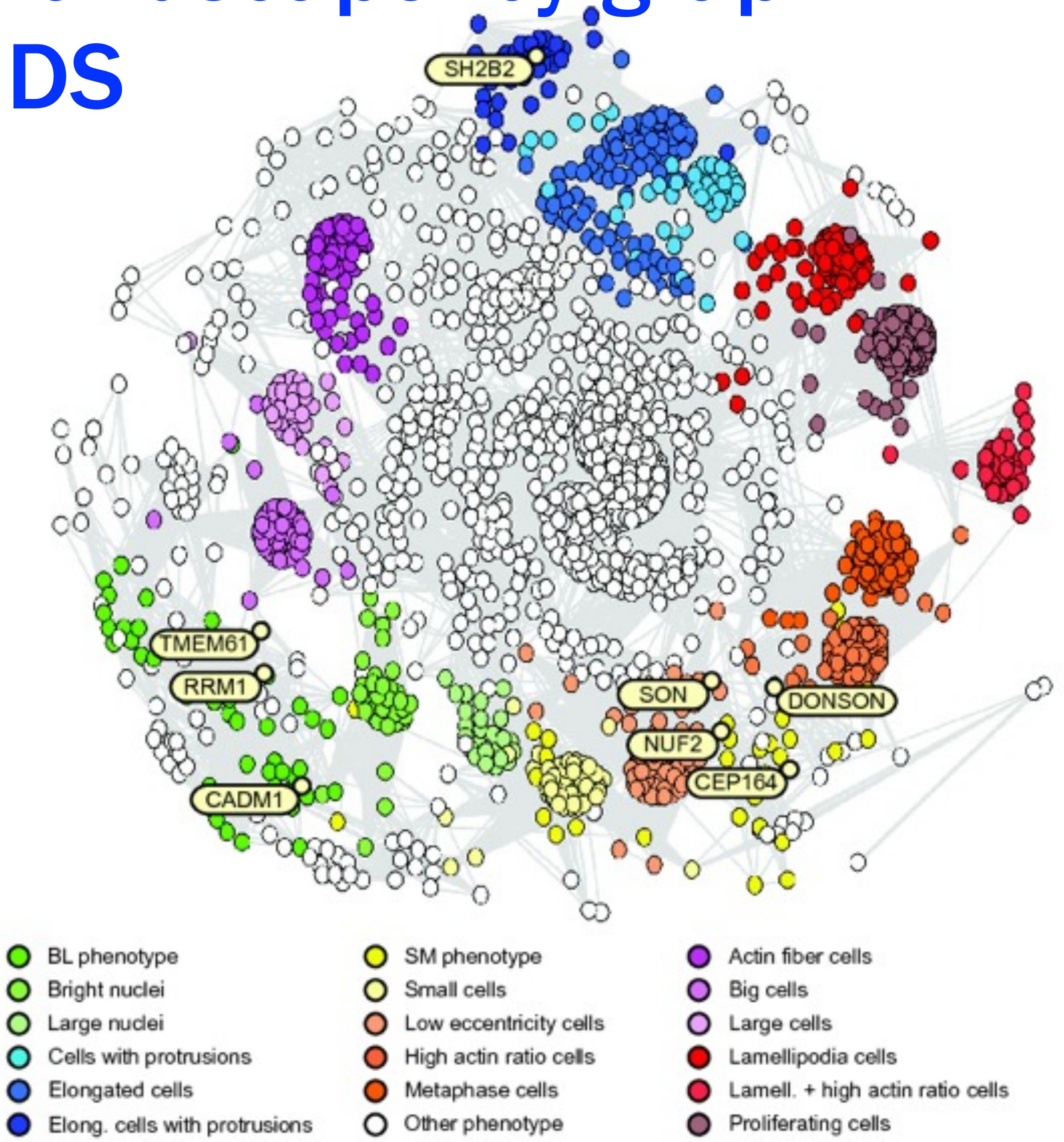
$$x = \left[\begin{array}{ll} \text{n} & 289 \\ \text{ext} & 34.33118 \\ \text{ecc} & 0.472934 \\ \text{Next} & 2857.356 \\ \text{Nint} & 485.2710 \\ \text{a2i} & 0.828876 \\ \text{Next2} & 0.098647 \\ \text{AF \%} & 0.049594 \\ \text{BC \%} & 0.081746 \\ \text{C \%} & 0.158817 \\ \text{M \%} & 0.179339 \\ \text{LA \%} & 0.009249 \\ \text{P \%} & 0.219697 \end{array}\right.$$

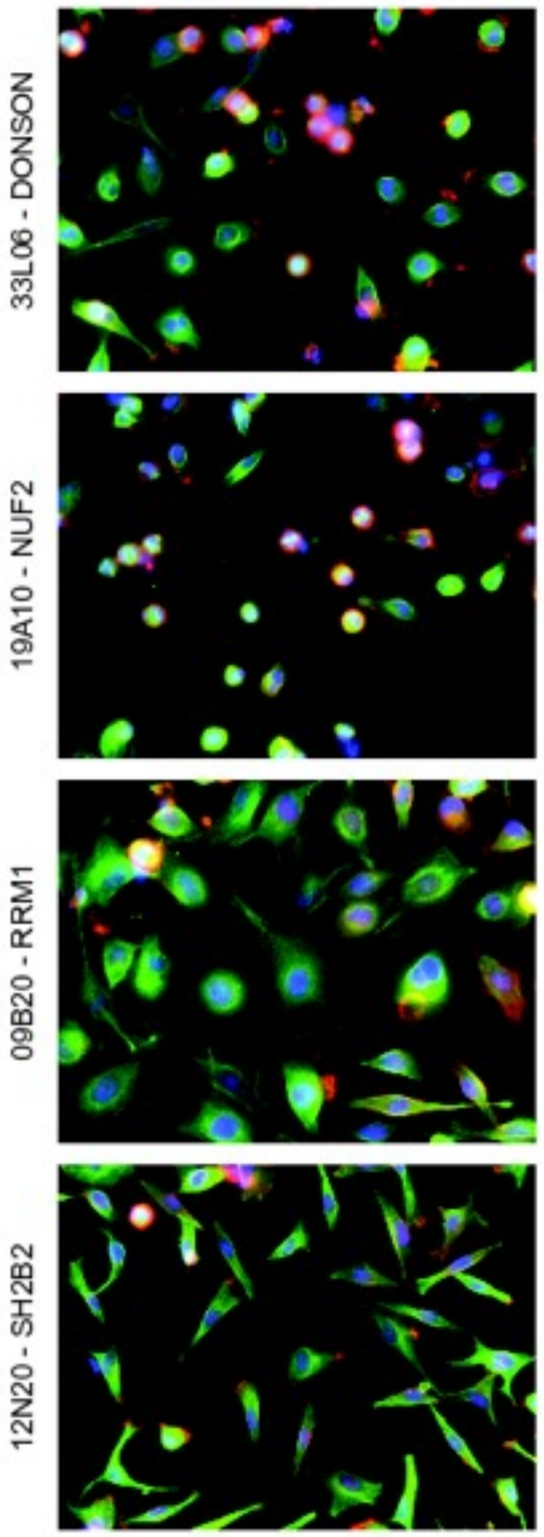
Training set: pairs of genes that are somehow 'related': EMBL STRING
Get (η, α) by minimizing average distance between training set genes, keeping average distance of all genes fixed.

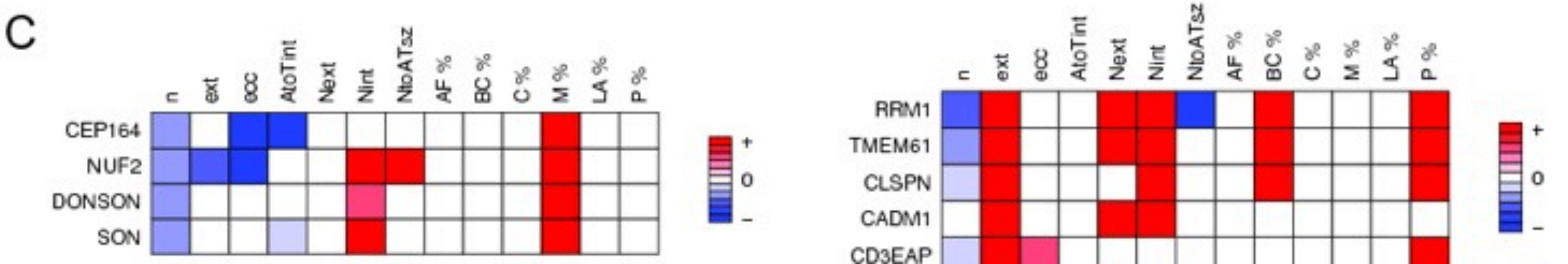




Phenotype landscape: by graph layout or MDS



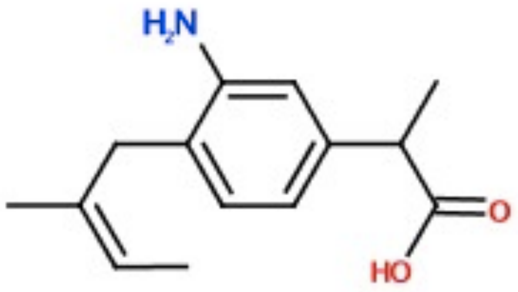




Summary

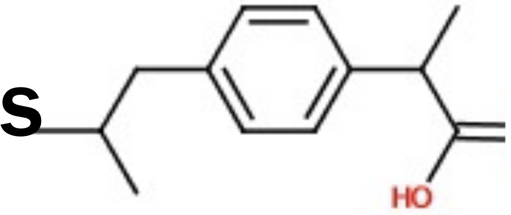
Automated phenotyping of cells upon genetic perturbations by microscopy and image analysis

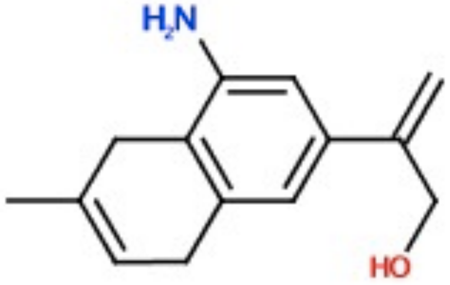
Segmentation, feature extraction, classification, distance metric learning, multi-dimensional scaling, clustering.



“Phenotypic map” is useful to biologists

Method is also being applied to drugs





Collaboration with Michael Boutros, German Cancer Research Centre (Heidelberg)

Fuchs, Pau et al., Molecular Systems Biology (2010)

All data and software available at
http://www.cellmorph.org
packages `EBImage` and `imageHTS`

Gregoire Pau

BIOCONDUCTOR

Focus on the analysis of genomic data

Based on R and CRAN

Six-monthly release cycle, in sync with R

Releases:

- 1.0 in March 2003 (15 packages), ...,
- 2.8 in April 2011 (466 software packages)

BIOCONDUCTOR

What's the added value?

Complex data containers (S4 classes) for new experimental technologies (microarrays, sequencing) shared between packages - even from different authors.

metadata packages: gene annotation, pathways, genomes

experiment data packages: landmark datasets

stronger emphasis on vignette-style documentation

stricter submission review (much more could be done)

more package interdependence → releases

training courses

mailing list is amenable to software and domain (bio) questions

Push new technologies: S4, vignettes, string handling, computations with ranges, out-of-RAM objects

Interactive Reports

Distinguish

- interactive exploration by data analyst
- reports (presentation graphics)

Interactive Reports

Distinguish

- interactive exploration by data analyst
- reports (presentation graphics)

Everybody has a PDF reader.

Interactive Reports

Distinguish

- interactive exploration by data analyst
- reports (presentation graphics)

Everybody has a PDF reader.

Everbody has a web browser.

Interactive Reports

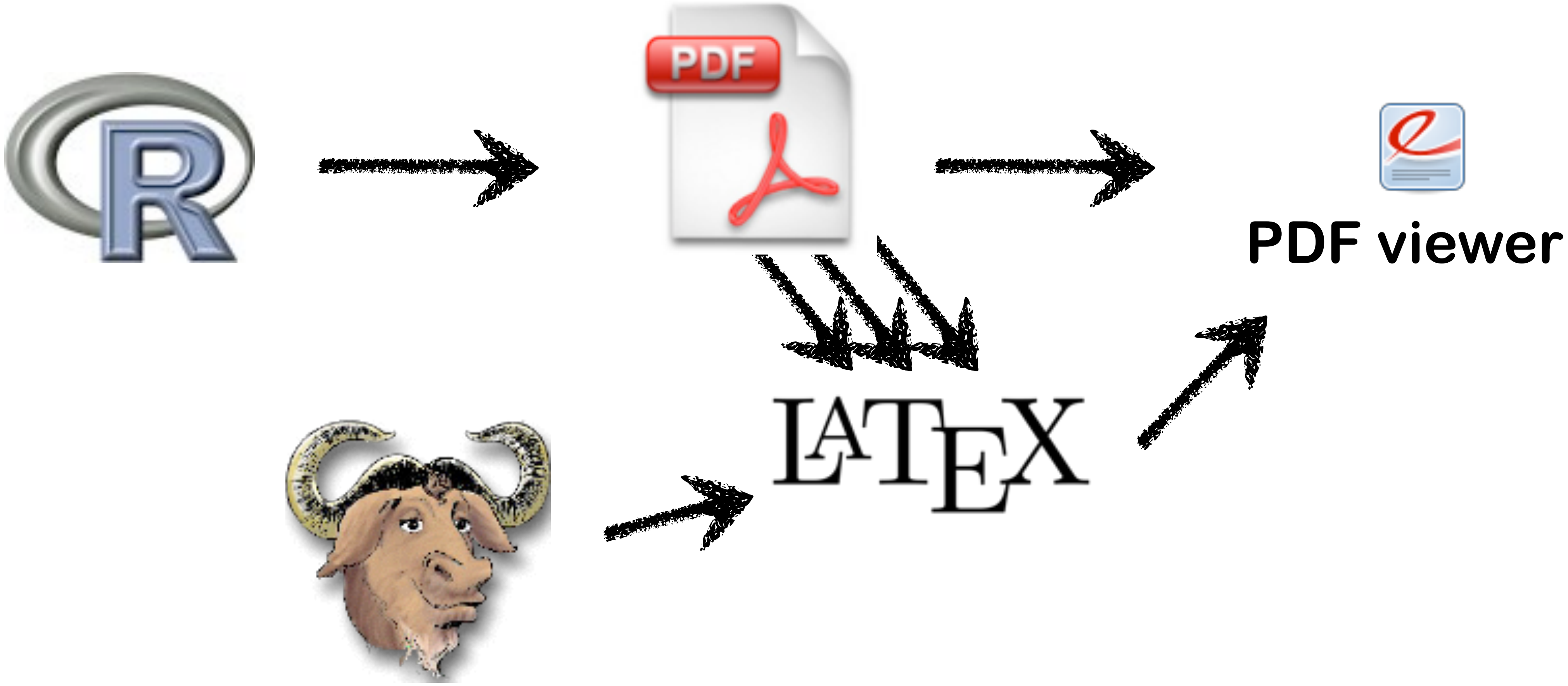
Distinguish

- interactive exploration by data analyst
- reports (presentation graphics)

Everybody has a PDF reader.

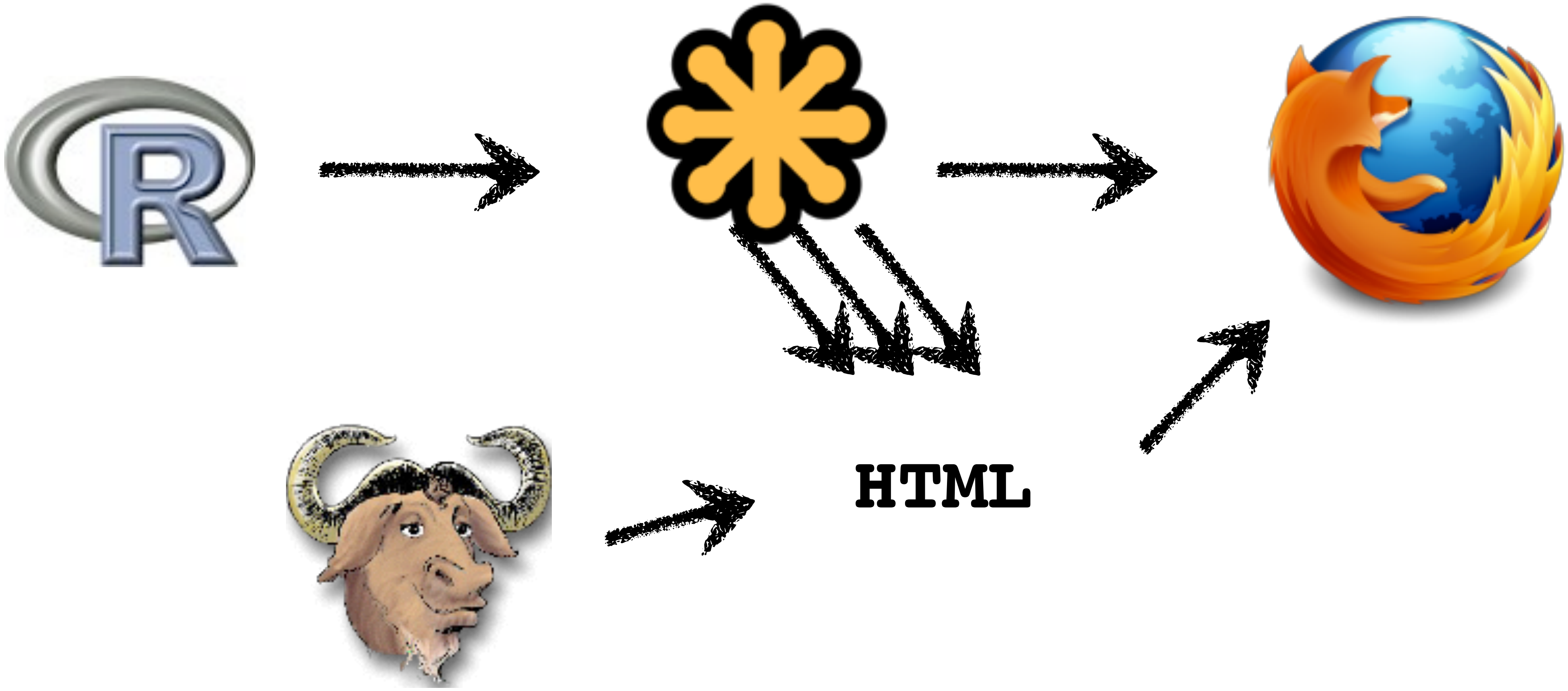
Everbody has a web browser.

Web browsers are turning into an operating system.

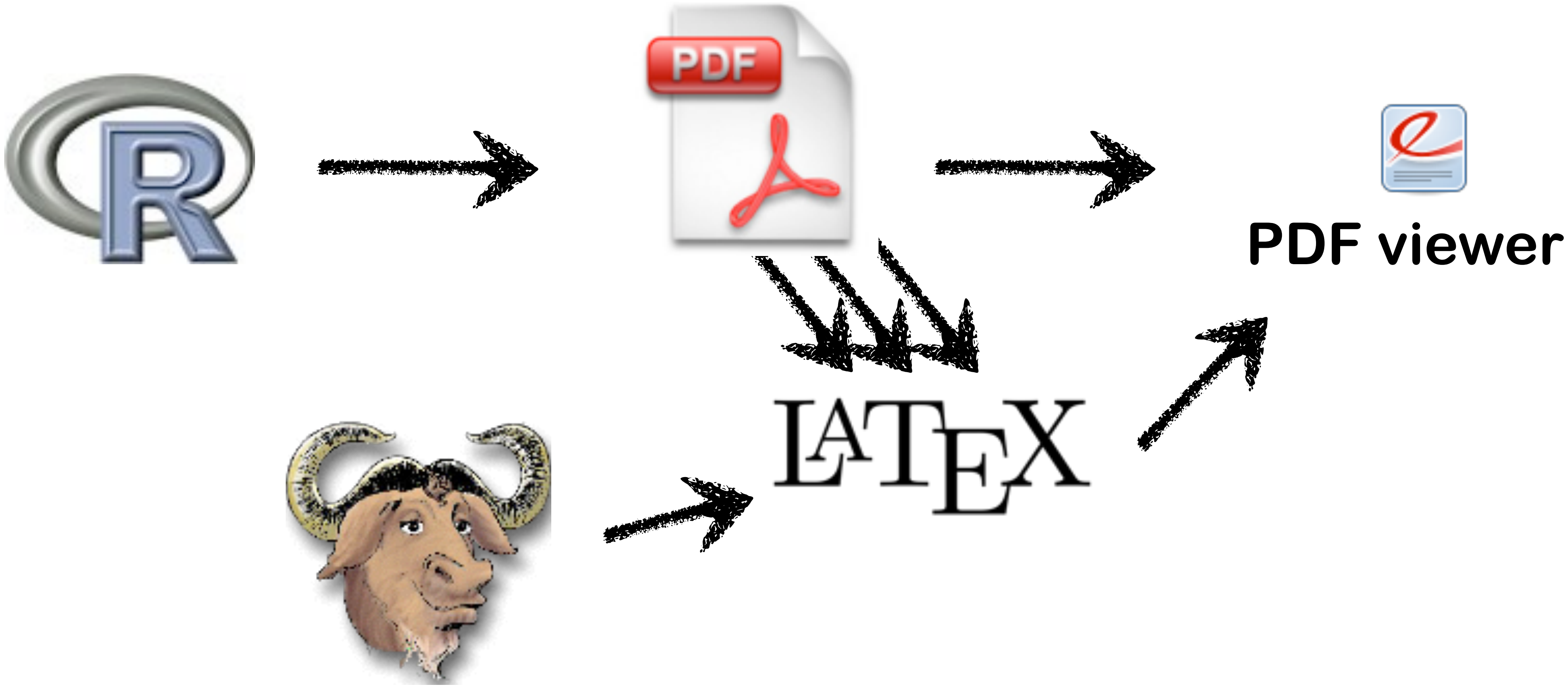


LaTeX

PDF viewer



HTML



LaTeX

PDF viewer

arrayQualityMetrics

Reports on Quality of Microarray Datasets

effort to collect all extant, useful quality metrics for microarrays

funding by EU FP7 and by Genentech

used by public databases (EBI::ArrayExpress) to annotate their data offerings for users

Example report

arrayQualityMetrics

Reports

effort to

funding

used by

offering

- mouseover → tooltip (rendered as an HTML table next to the plot)
- click → select & highlight (propagated to several plots, tables)
- expand, collapse sections
- use HTML elements (checkboxes) to control plots

Comments and outlook

SVG is part of HTML 5:

- linked plots and brushing
- HTML widgets as controllers (checkboxes, wheels)

SVG/HTML post-processing via the XML package

Callback processing currently in JavaScript.
Use R? On server: googleVis talk by Markus Gesmann, Diego de Castillo; locally: browser plugin

Duncan Temple Lang's SVGAnnotation package: works for any R graphic (incl. base), but depends on undocumented / changeable behavior of cairo.

Paul Murrell's gridSVG package: cleaner and more durable approach, based on grid graphics.

Generalisation?

arrayQualityMetrics is for microarrays

Software sees:

- a set of items (arrays)
- a set of modules that compute the sections of the report (PCA, boxplots, scatterplots)

This could be generalised to reports on very different types of subject matter - I will be happy to discuss this.

What makes us different?

From Genome Wide Association Studies, ~400 variants that contribute to common traits and diseases are known

Individual and the cumulative effects are disappointingly small

A G A G T T C T G C T C G
A G G G T T A T G C G C G
C G T T C G G G A A T C C
C G T T A G G A A A T C T
T C T T T G A C G A C T C
T C T T A G A G G A C T C

What makes us different?

From Genome Wide Association Studies, ~400 variants that contribute to common traits and diseases are known

Individual and the cumulative effects are disappointingly small

Epistasis, interactions

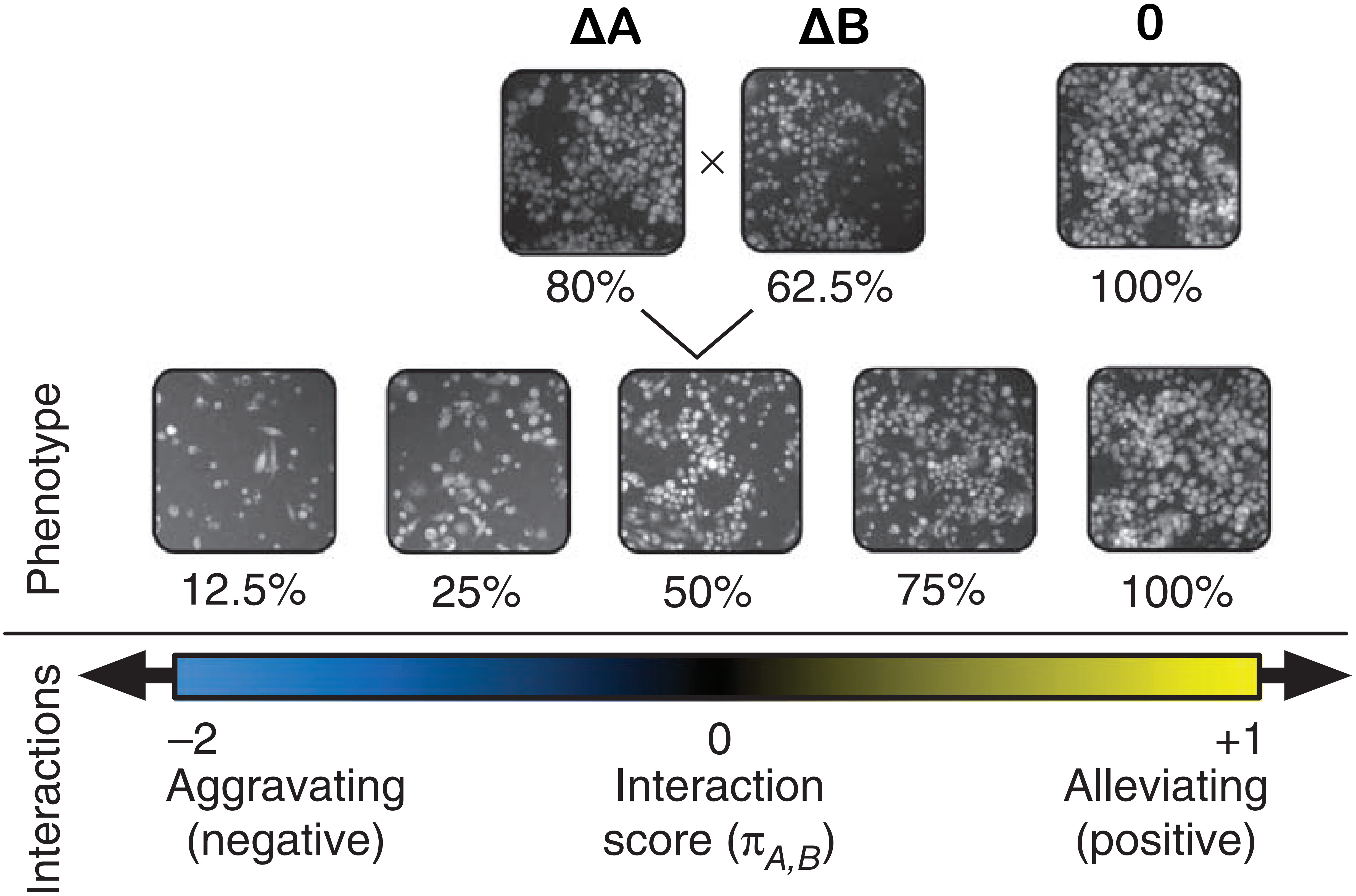
$$\phi = \phi_0 + \sum_{i=1}^{5}\phi_i x_i + \sum_{i,j=1}^{5}\phi_{ij} x_i x_j + \sum_{i,j,k=1}^{5}\phi_{ijk} x_i x_j x_k + \ldots$$

Take a step back...

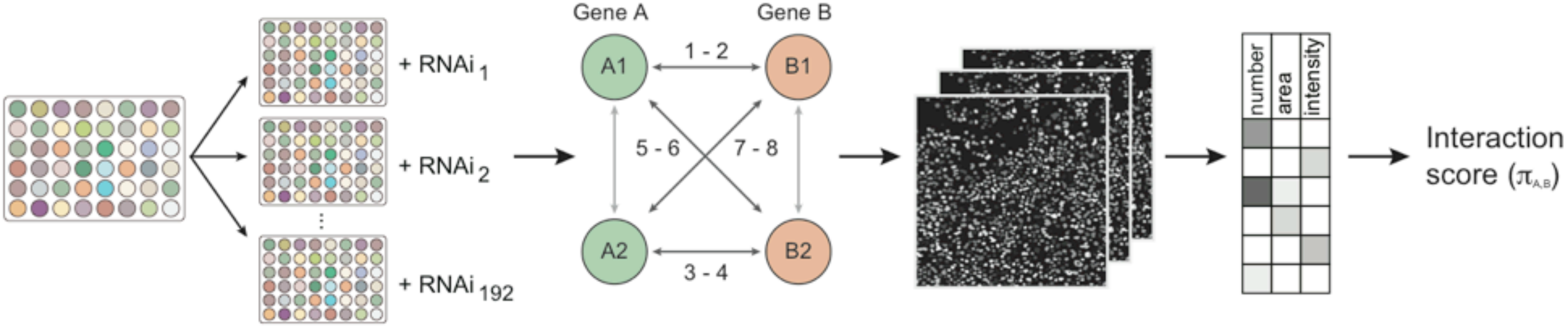
Genetic interactions

- only pairwise
- for a simple phenotype
- in a simple model system

Simplest "model system": pairwise gene knock-down interactions and a scalar phenotype



A combinatorial RNAi screen



Combinatorial RNAi

Imaging and image analysis

Modelling of genetic interactions

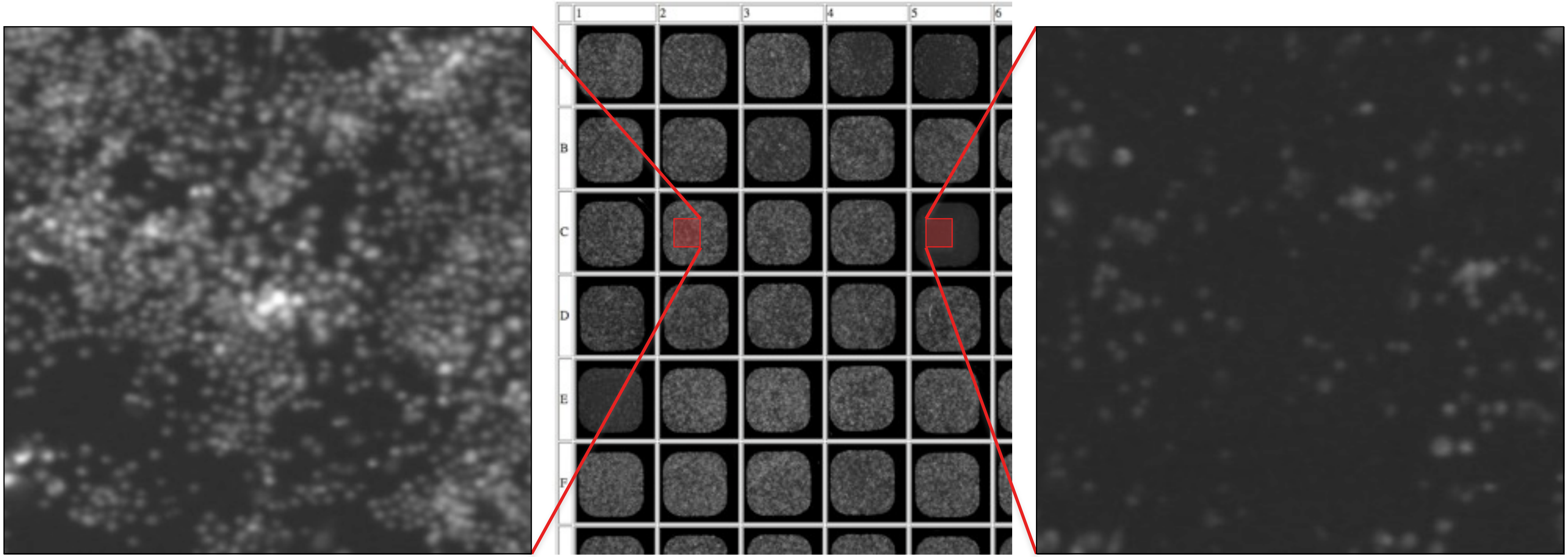
- 93 Dm kinases and phosphatases
- Each targeted by two independent dsRNA designs
- Validation of knock-down by qPCR
- 96 plates (~37.000 wells)
- 4.600 distinct gene pairs

with Bernd Fischer (EMBL) and
M. Boutros, Thomas Sandmann,
Thomas Horn (DKFZ)

Nature Methods 4/2011

Image analysis and feature extraction
(version of 2010)

- number of cells
- DAPI intensity for each cell
- DAPI area for each cell



Modelling Genetic Interactions

For many phenotypes, the perturbation effects combine multiplicatively for non-interacting genes i, j:

$$d_{ij} = \omega\, \mu_i\, \mu_j$$

... i.e. additive on a logarithmic scale

$$\log d_{ijk} = w + m_i + m'_{j} + g_{ij} + \varepsilon_{ijk}$$

measurement
(nr cells, growth rate, ...)

baseline

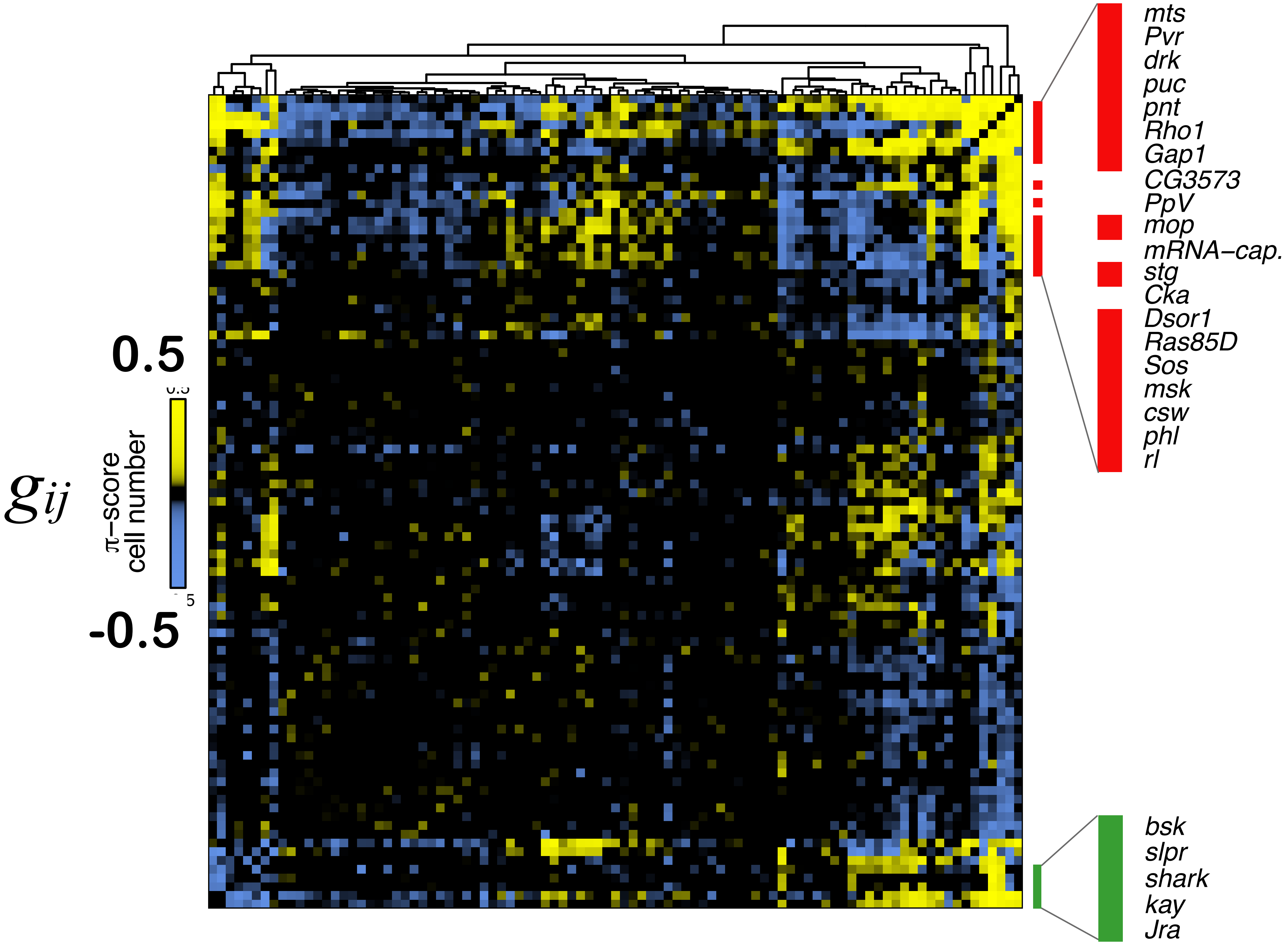
main effect
of dsRNA i

main effect
of dsRNA j

interaction

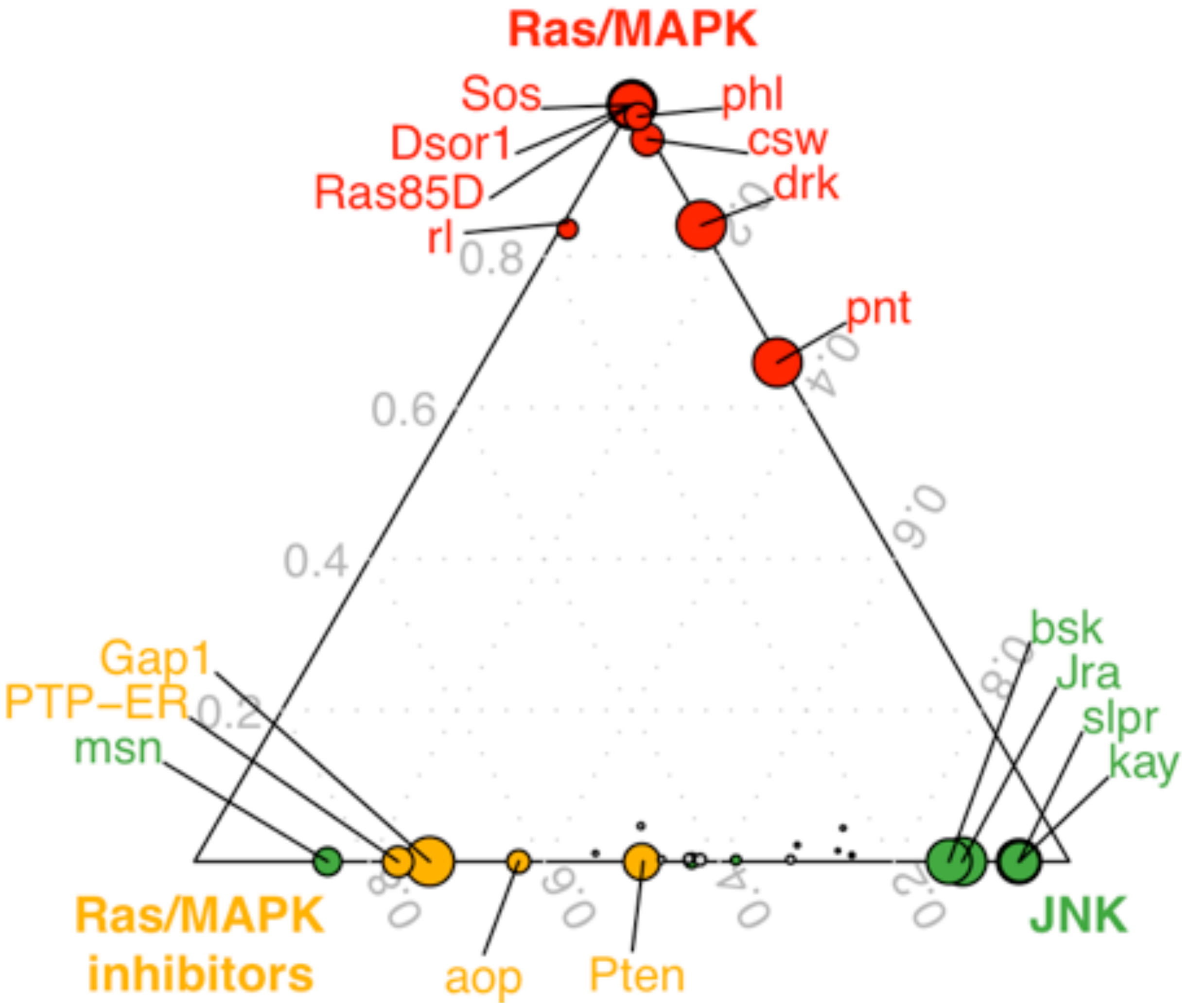
measurement error

Thus we get a matrix of interaction parameters: profile clustering reflects functional modules

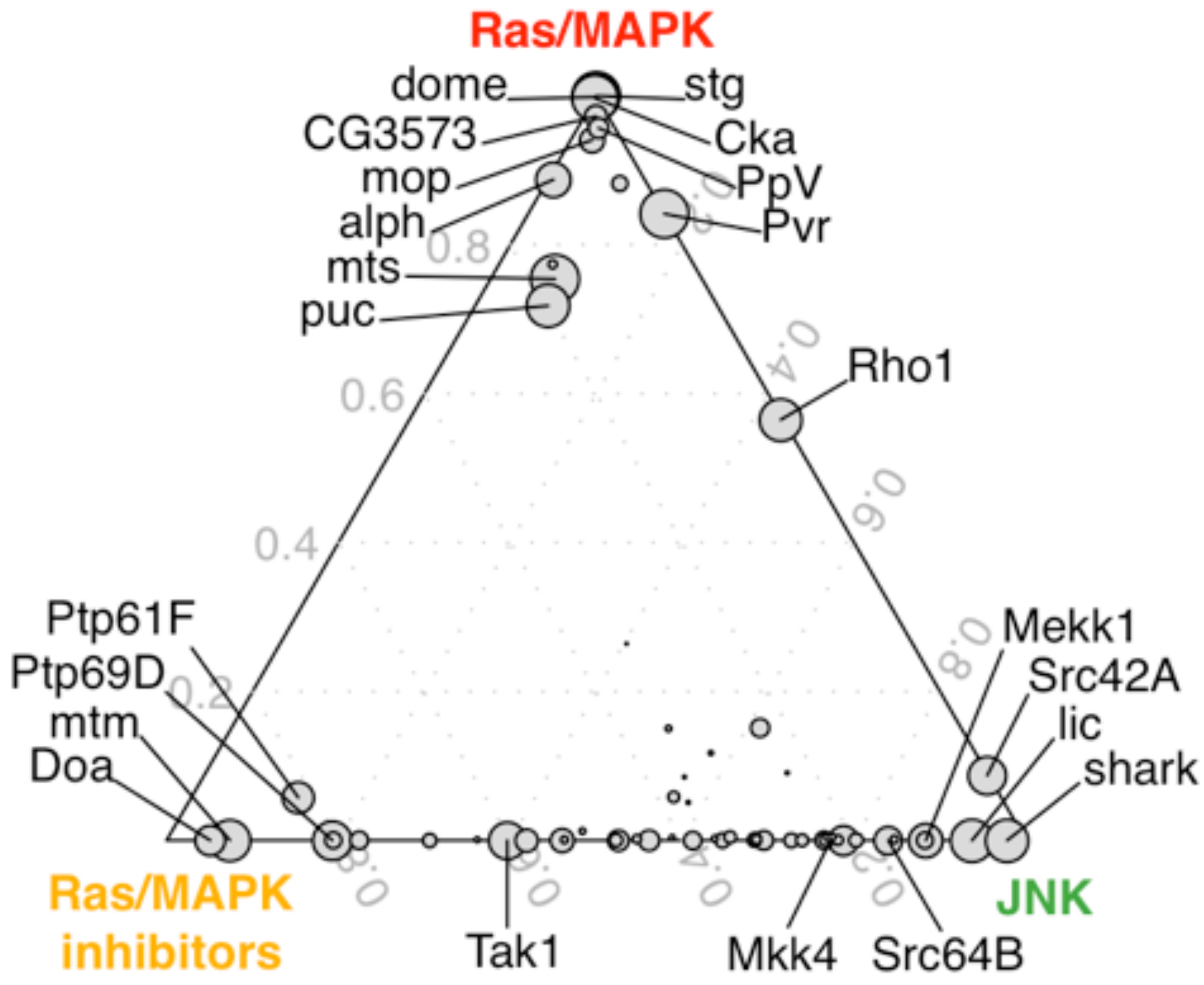


Classification of genes by function through their interaction profiles

cross-validated performance on training set



functional prediction applied to new genes

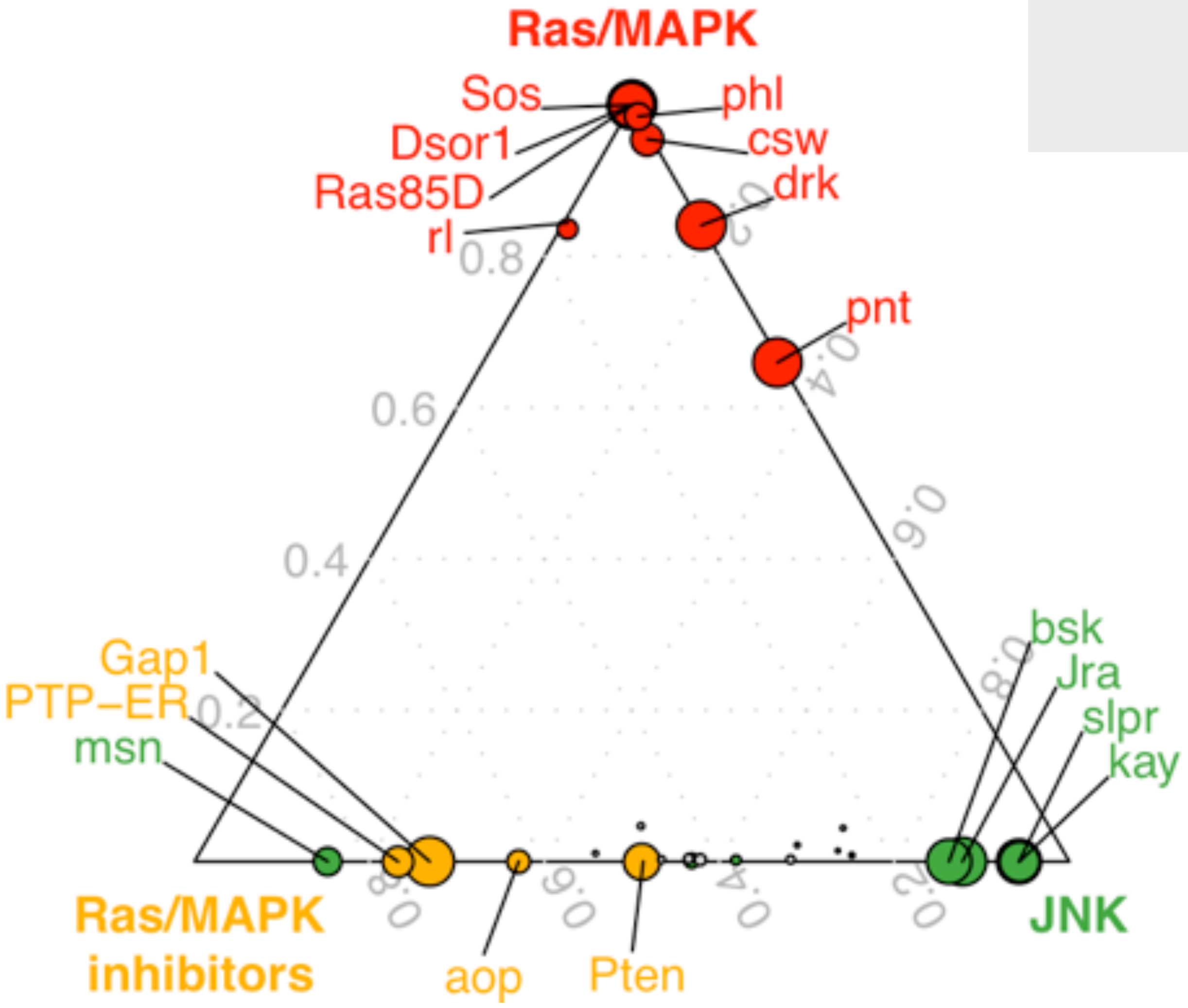


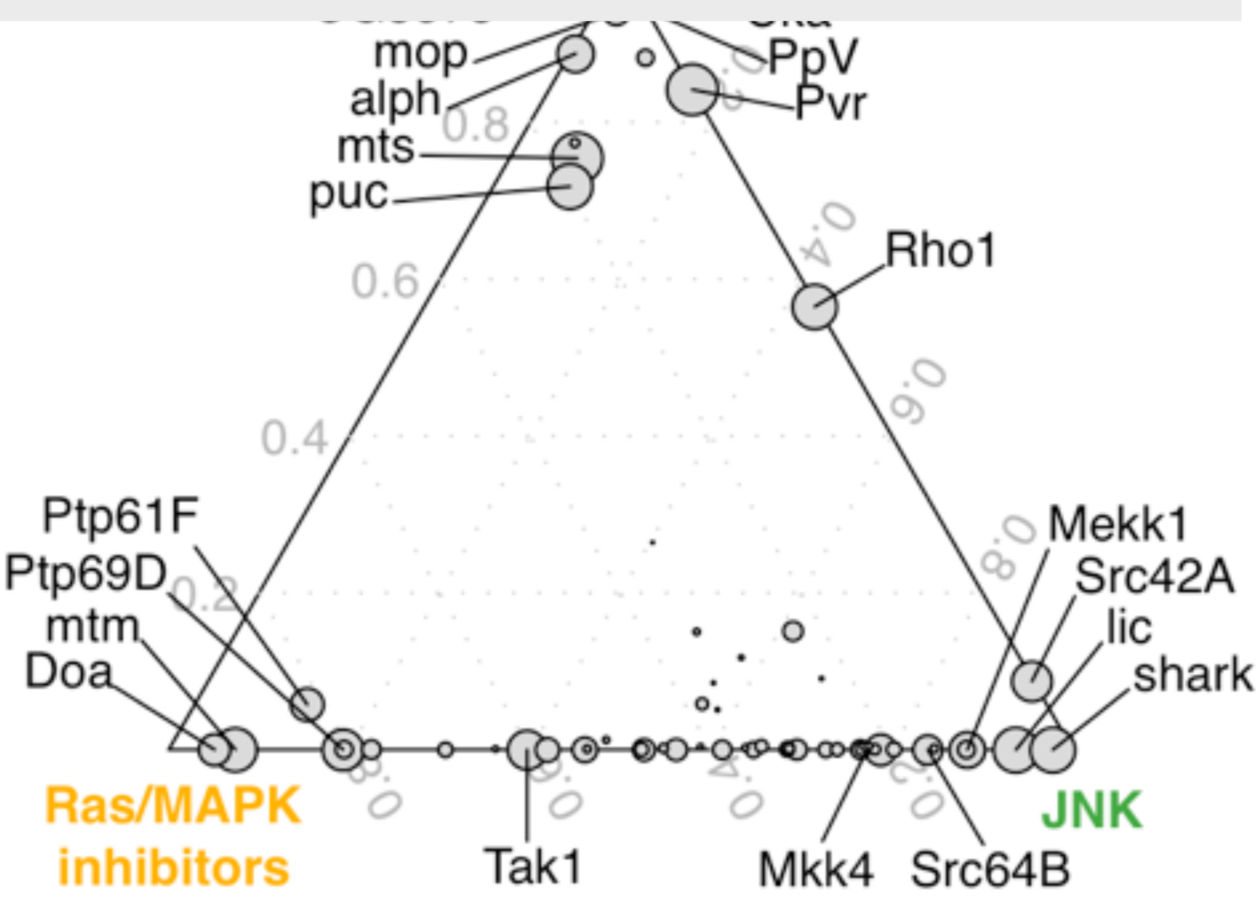
circle sizes ~ cross-validated posterior probabilities of the classifier

Classification of g
their int

Show Me Your Friends and
I'll Tell You Who You Are

cross-validated performance
on training set



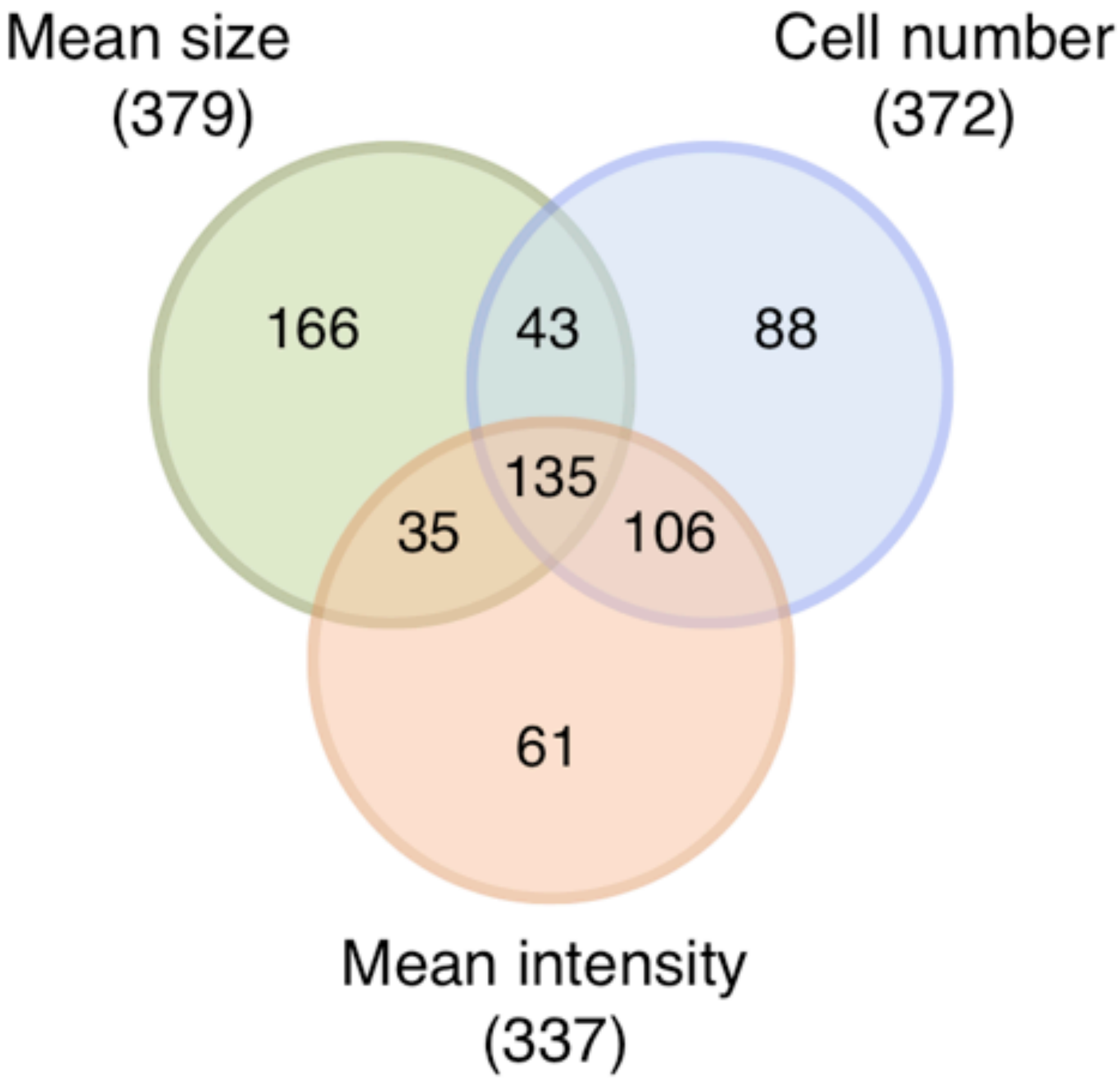


circle sizes ~ cross-validated posterior
probabilities of the classifier

Genetic interactions in 3 dimensions

Different phenotypes produce different sets of interactions

For each set, significant overlap with known genetic interactions and with human interologs

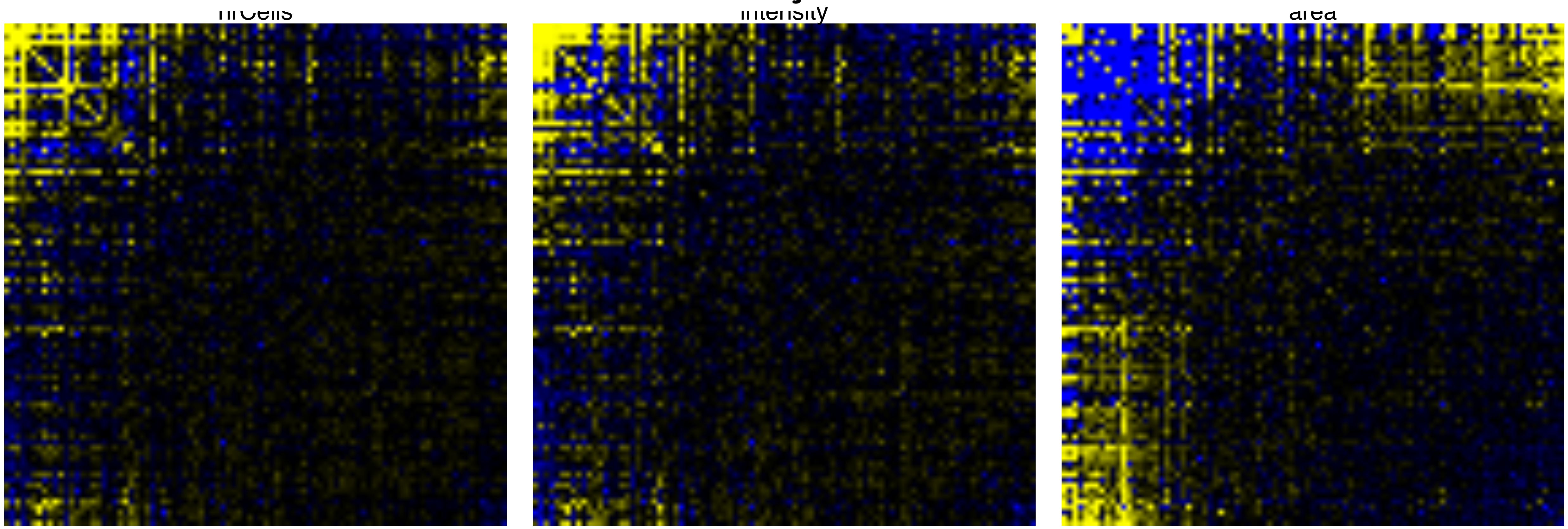


Interaction matrices

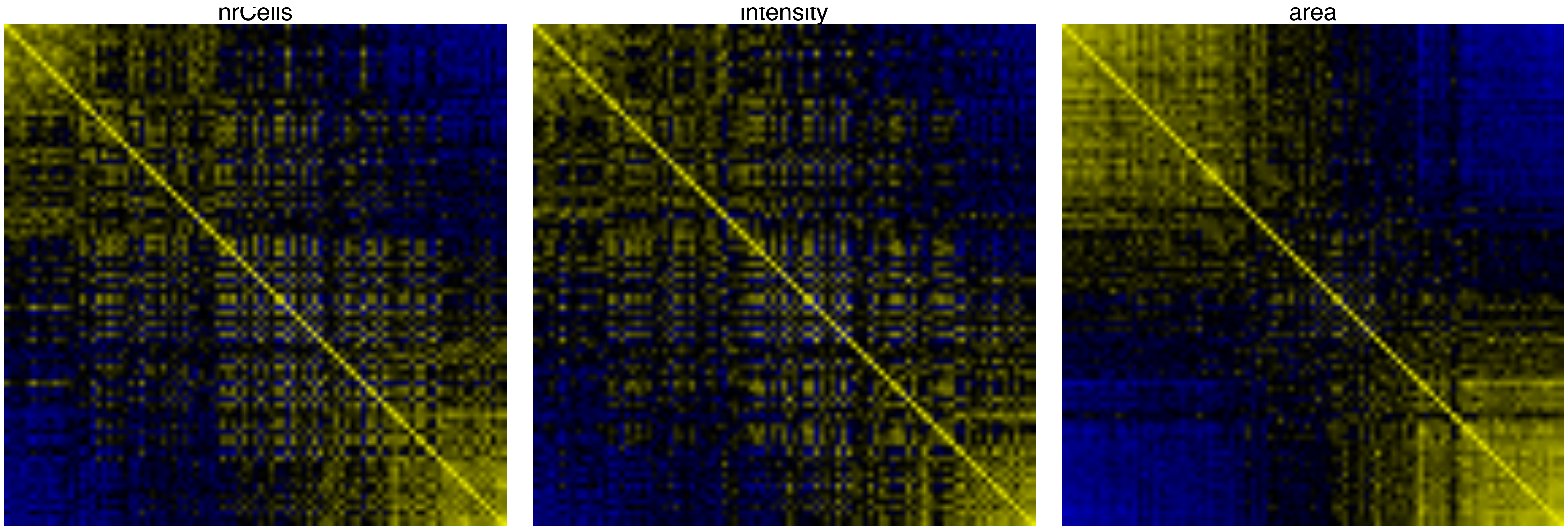
number of cells

intensity

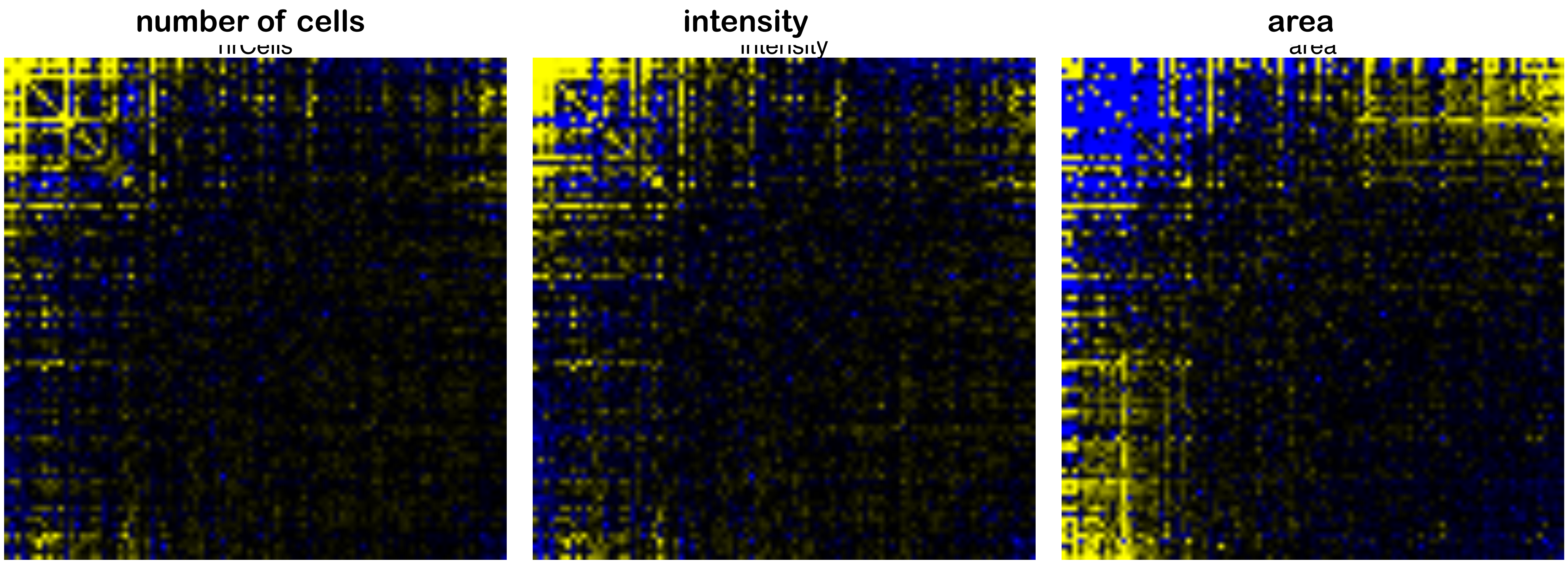
area



Correlation matrices



Interaction matrices

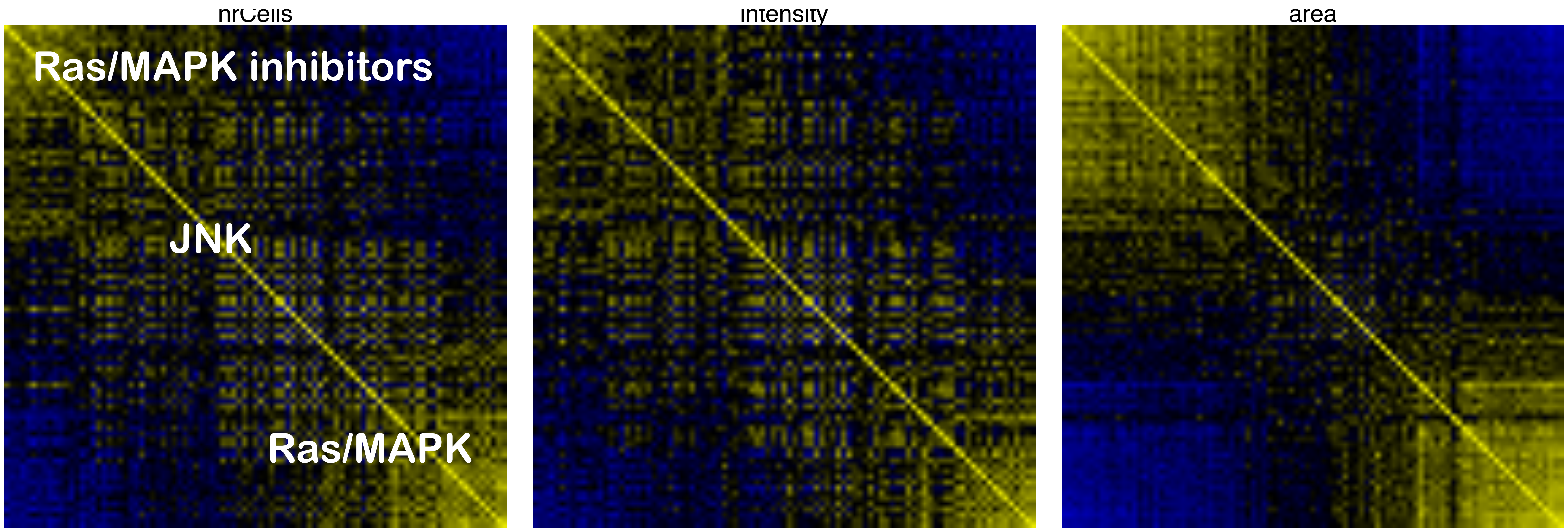


number of cells

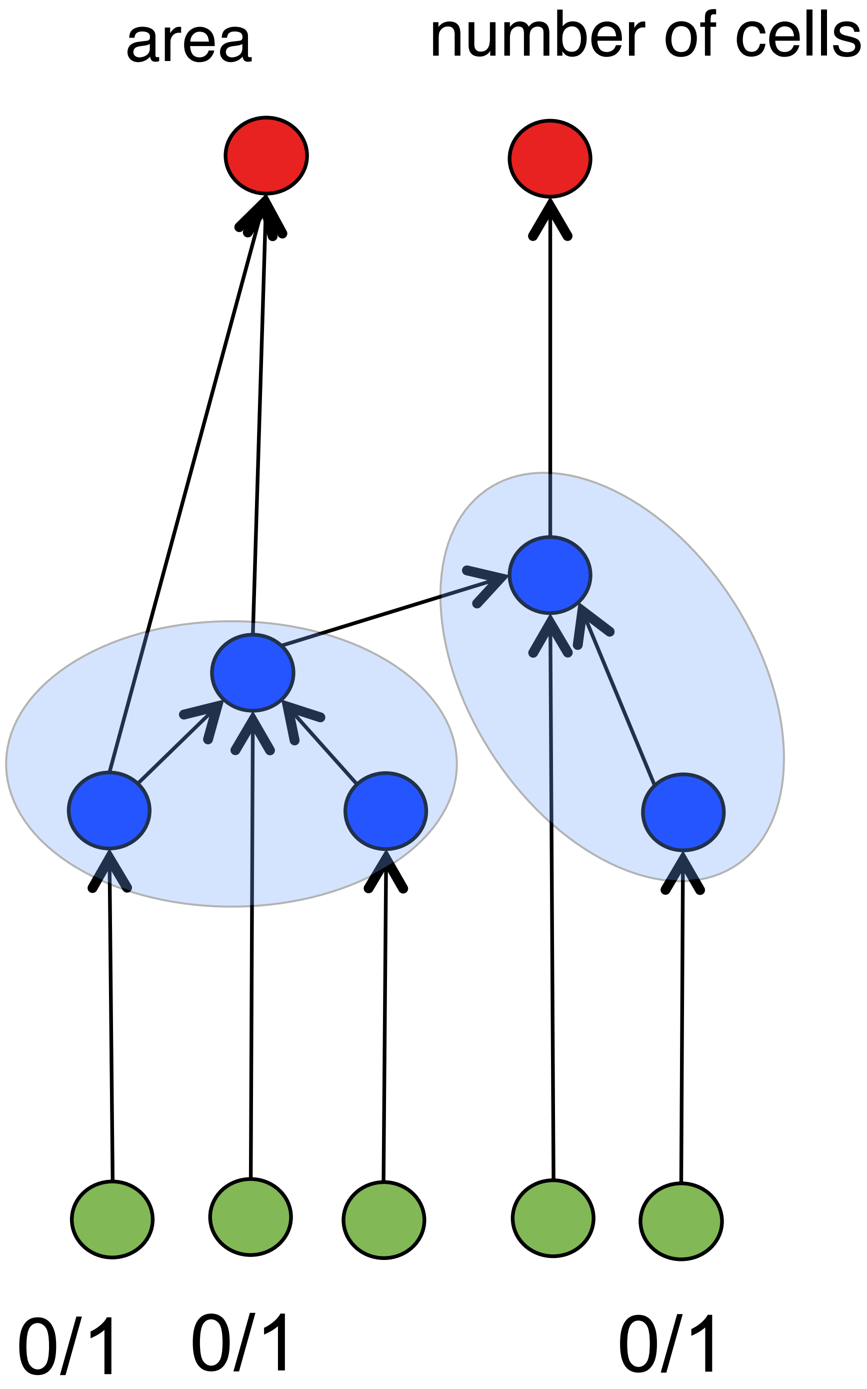
intensity

area

Correlation matrices



Network learning - identify the underlying molecular modules



phenotypes (p)
(observed)

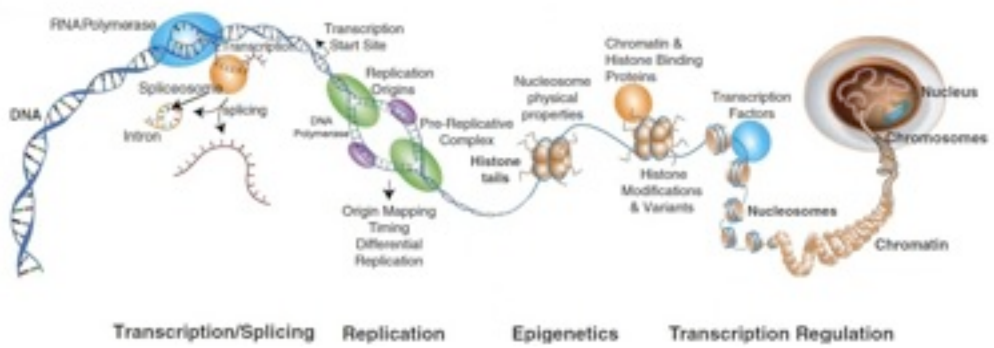
activity (a) of core modules (e.g. complexes, ‘path-ways’)
(hidden)

binary **genetic** perturbation (g)
(observed)

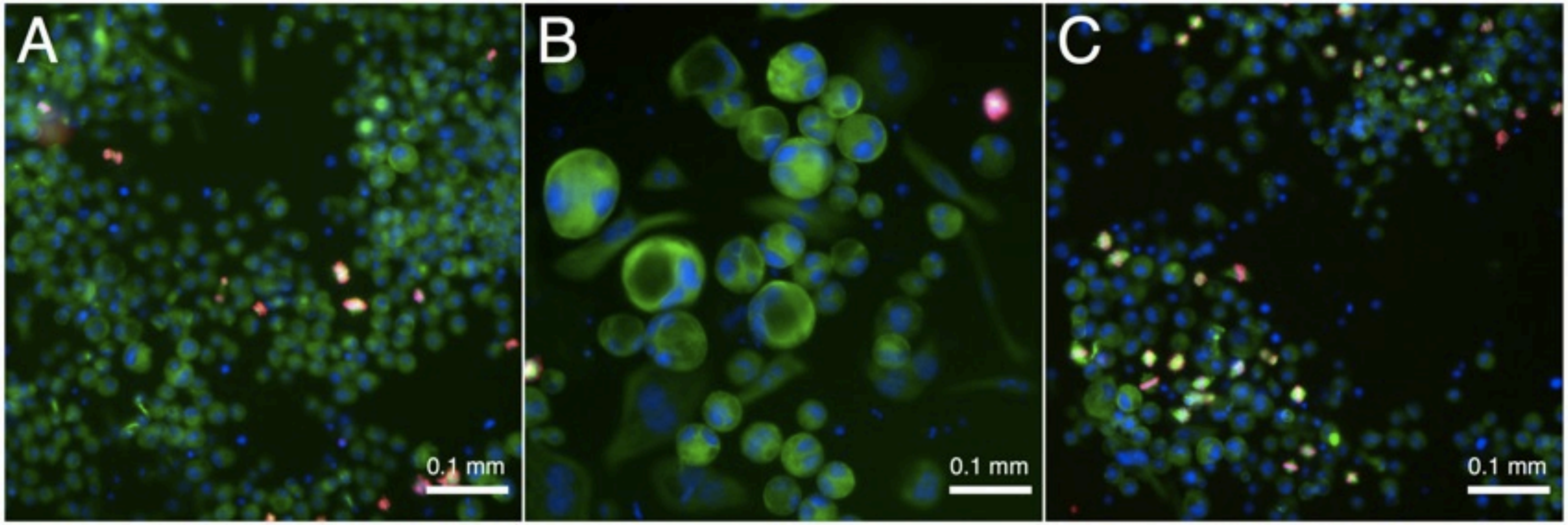
$$P\left(p \mid g; \alpha, \beta, \gamma\right) = \sum_{a} P\left(p \mid a; \alpha\right) \prod_{i=1}^{N} P\left(a_i \mid a_{pa(i)}, g_{pa(i)}; \beta, \gamma\right)$$

Ongoing: a much bigger matrix

- Larger matrix, again Dmel2 cells
- ~1500 chromatin-related genes x 100 query genes
- full microscopic readout (4x and 20x), 3 channels:
  - DAPI
  - phospho-His3 (mitosis marker)
  - aTubulin (for spindle phenotypes)



- 1600 384-well plates, ~ 300.000 measurements

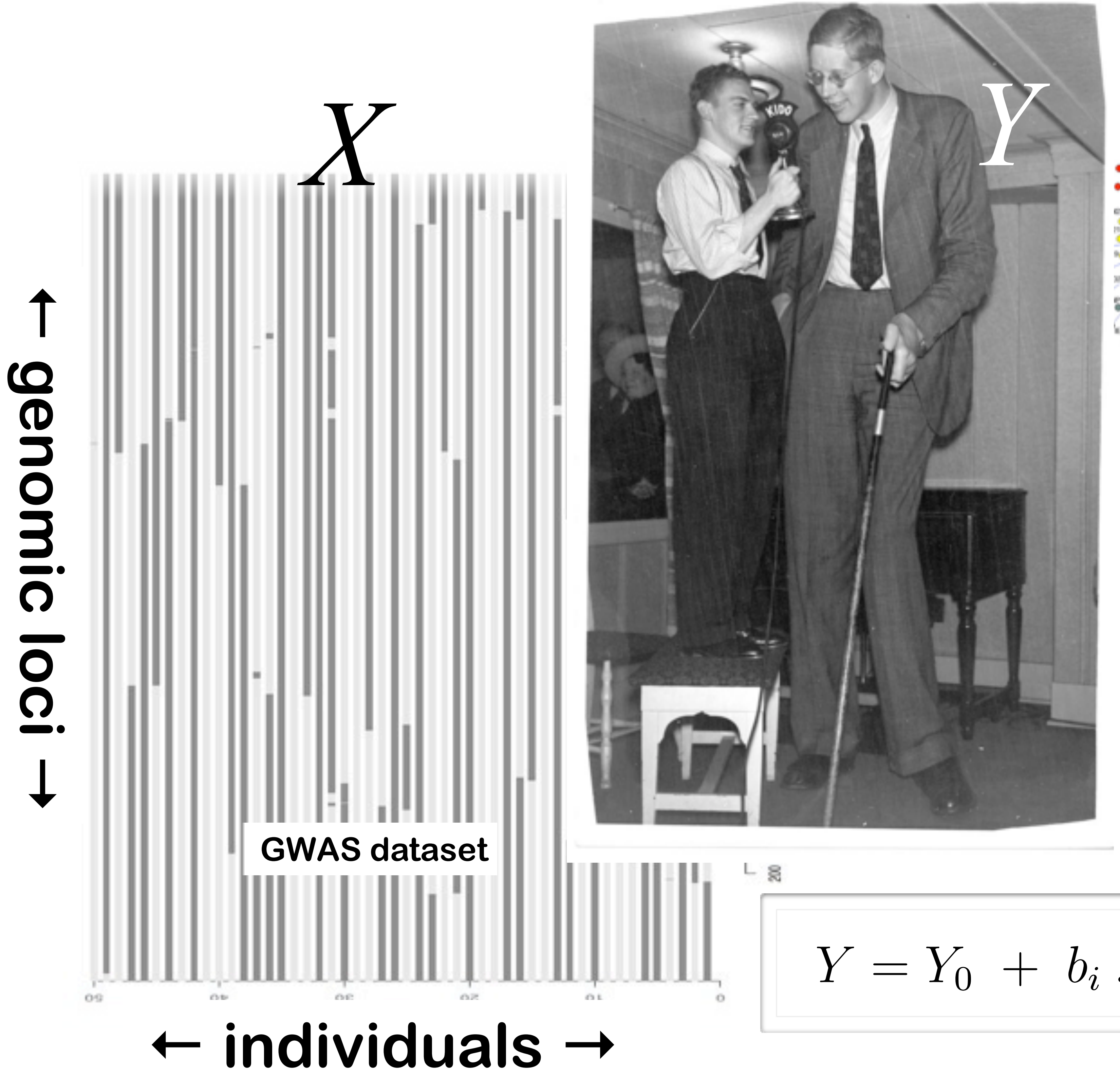


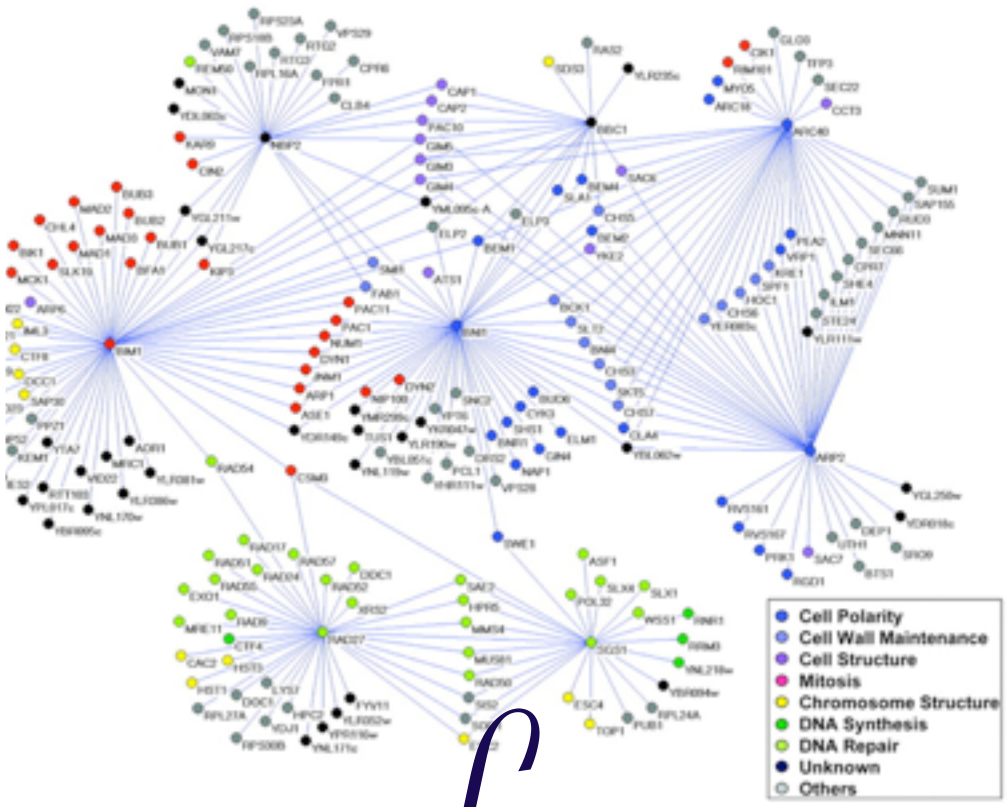
ctrl dsRNA

Rho1 dsRNA

Dynein light chain dsRNA

Outlook: genetic interactions from model system experiments as regularisation/priors for the identification of genetic interactions in observational studies





$\mathcal{L}$

sparsity constraints
on b

$$Y = Y_0 \ + \ b_i \ x_i \ + \ b_{ij} \ x_i x_j \ + \ b_{ijk} \ x_i x_j x_k$$

Summary

Quantitative, combinatorial RNAi works in metazoan cells. Technological tour de force; data exploration, QA/QC, normalisation and transformation....

Individual genetic interactions *vs* interaction profiles.

Data are high-dimensional and complicated:

- dose effects,
- different / multivariate phenotypes
- relative timing

reveal non-redundant interactions.

All data & code available from BIOCONDUCTOR

Bernd Fischer,
Thomas Horn, Thomas Sandmann, Michael Boutros
Nature Methods 2011(4)

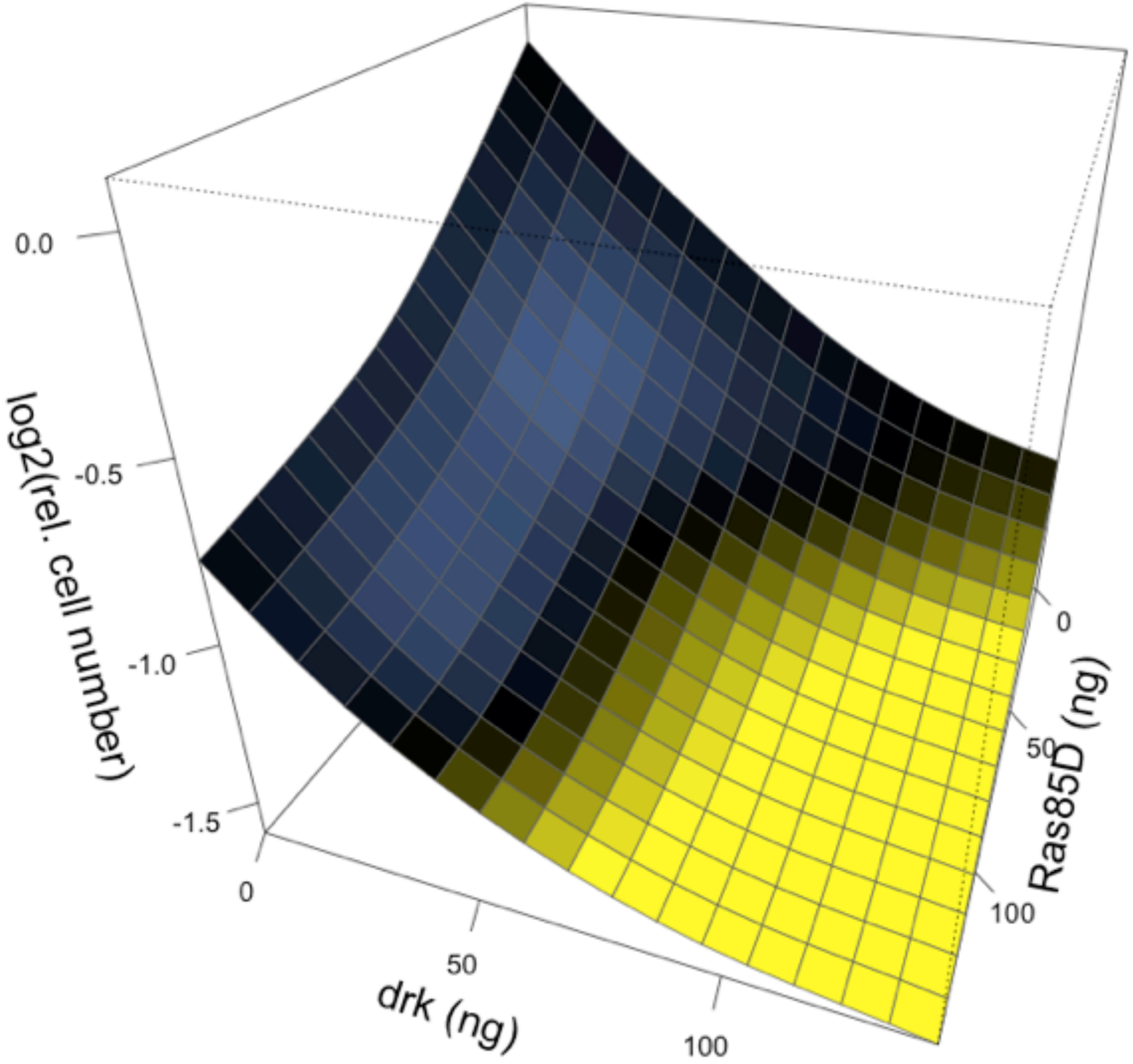
Simon Anders
Joseph Barry
Bernd Fischer
Ishaan Gupta
Felix Klein
Gregoire Pau
Aleksandra Pekowska
Paul-Theodor Pyl
Alejandro Reyes

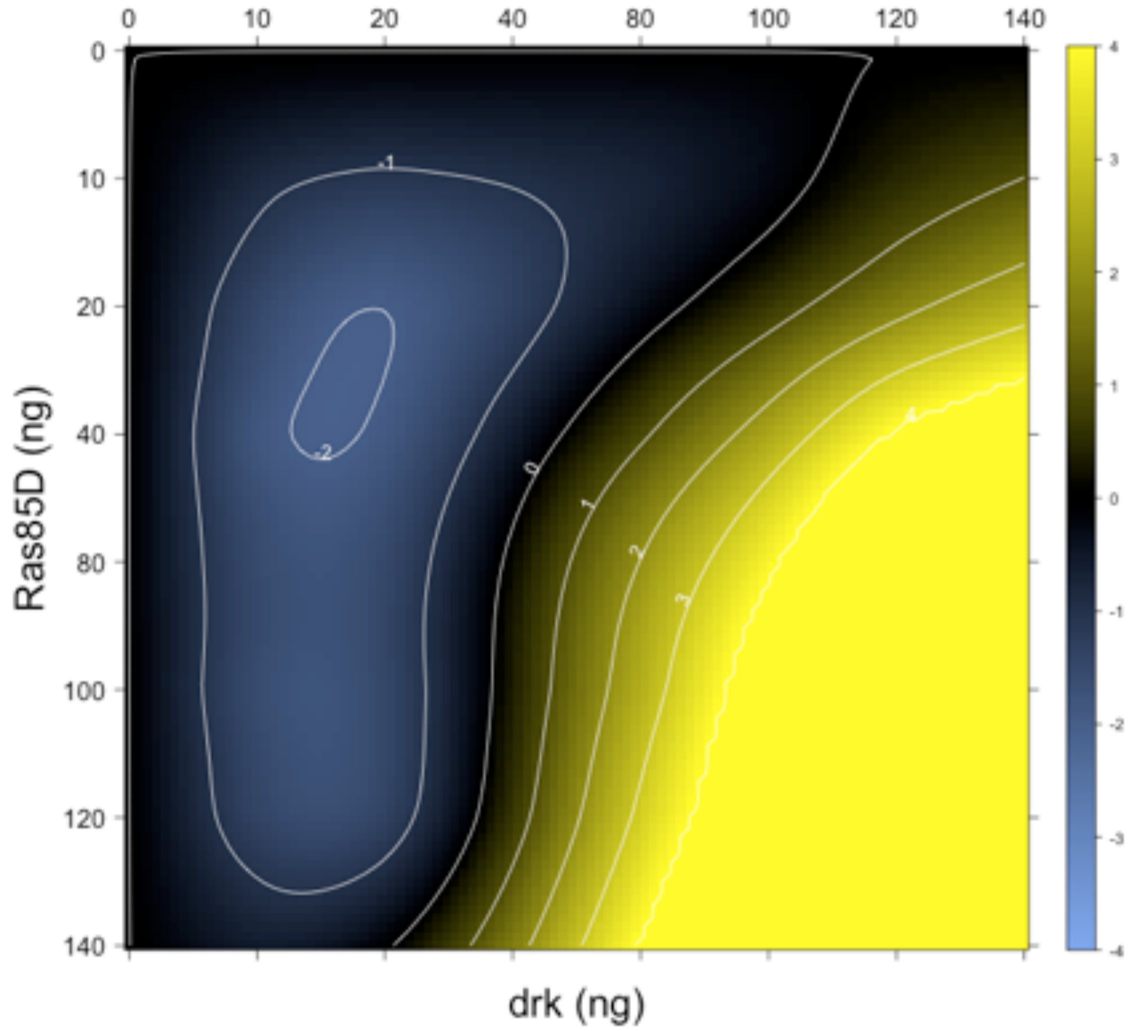
Collaborators
Lars Steinmetz
Michael Boutros (DKFZ)
Robert Gentleman (Genentech)
Jan Korbel

Michael Knop (Uni HD)
Jan Ellenberg
Kathryn Lilley (Cambridge)
Anne-Claude Gavin
Alvis Brazma (EBI)
Paul Bertone (EBI)
Ewan Birney (EBI)

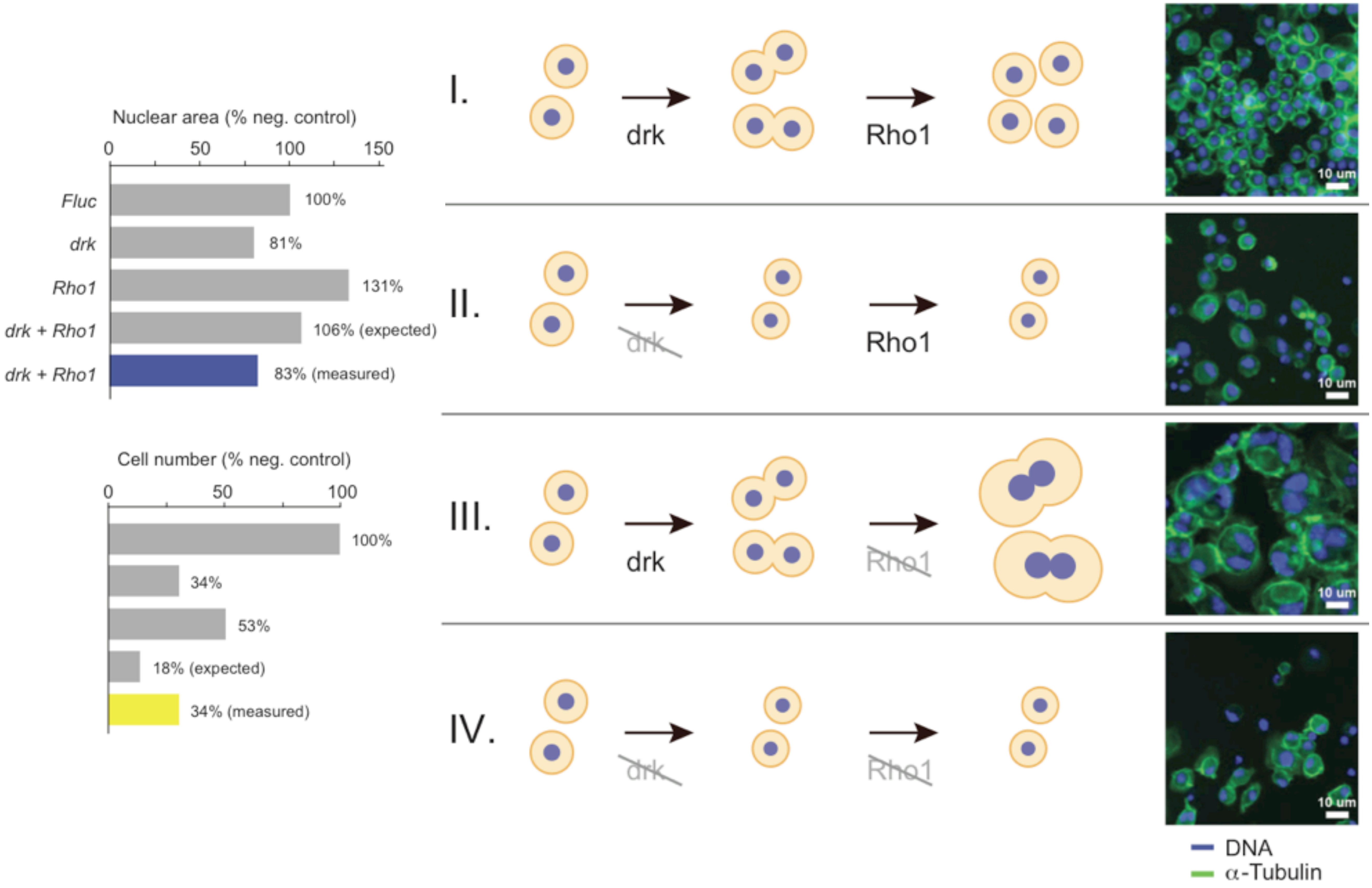
Ras85D and drk: concentration dependence

strength, presence and direction of an interaction can depend on reagent concentration (cf. drug-drug interactions)



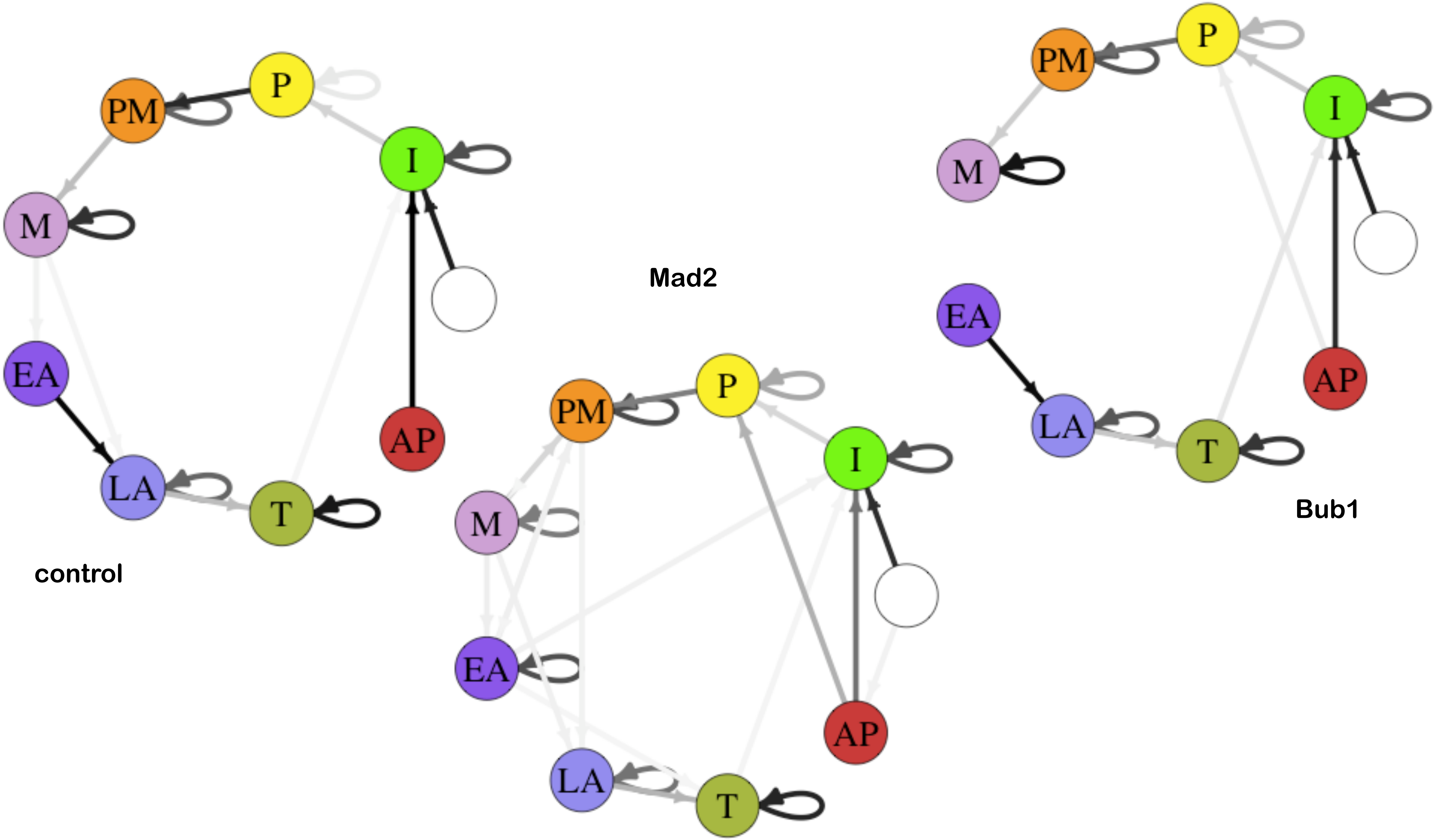


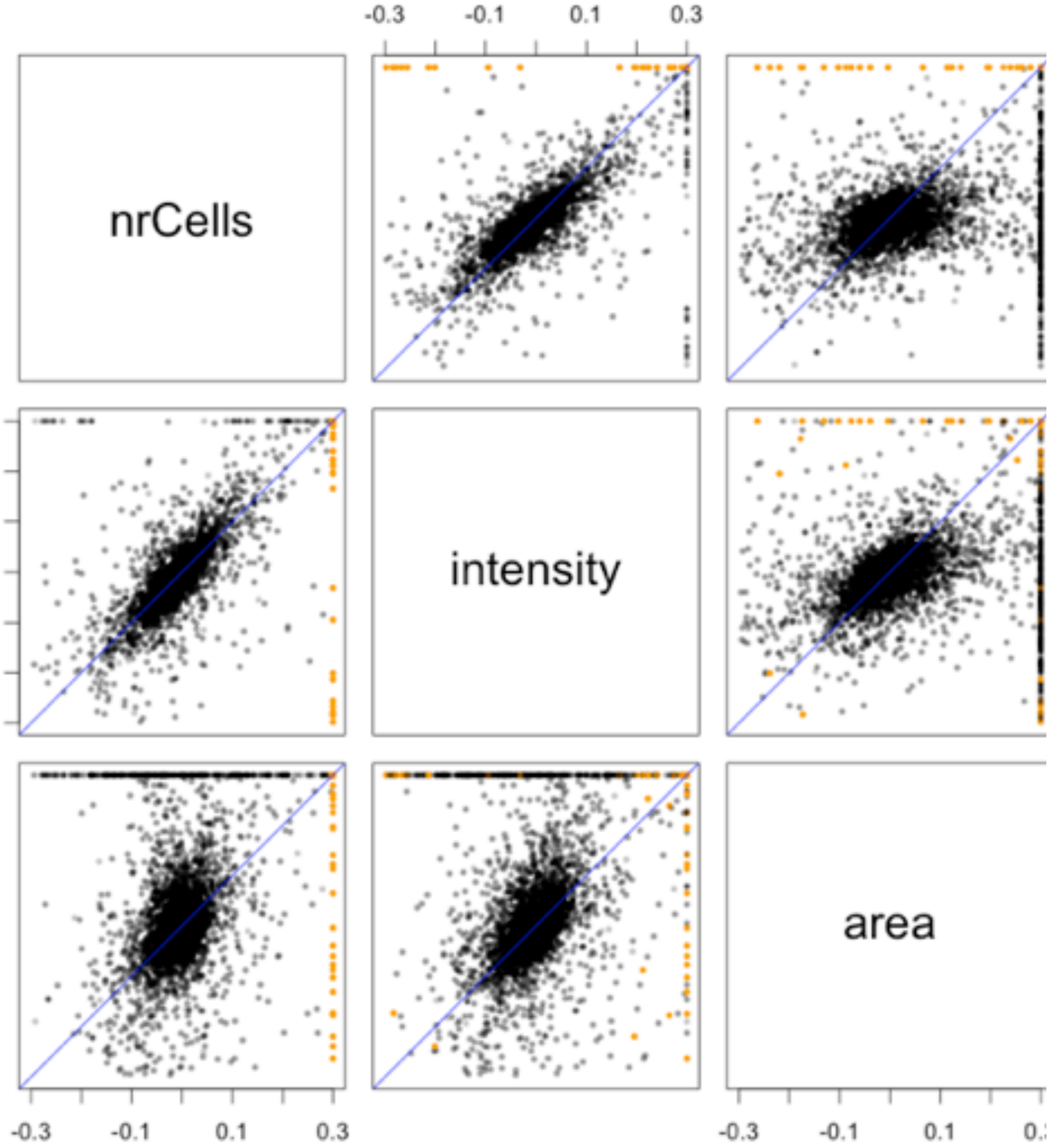
Sign inversion for different phenotypes



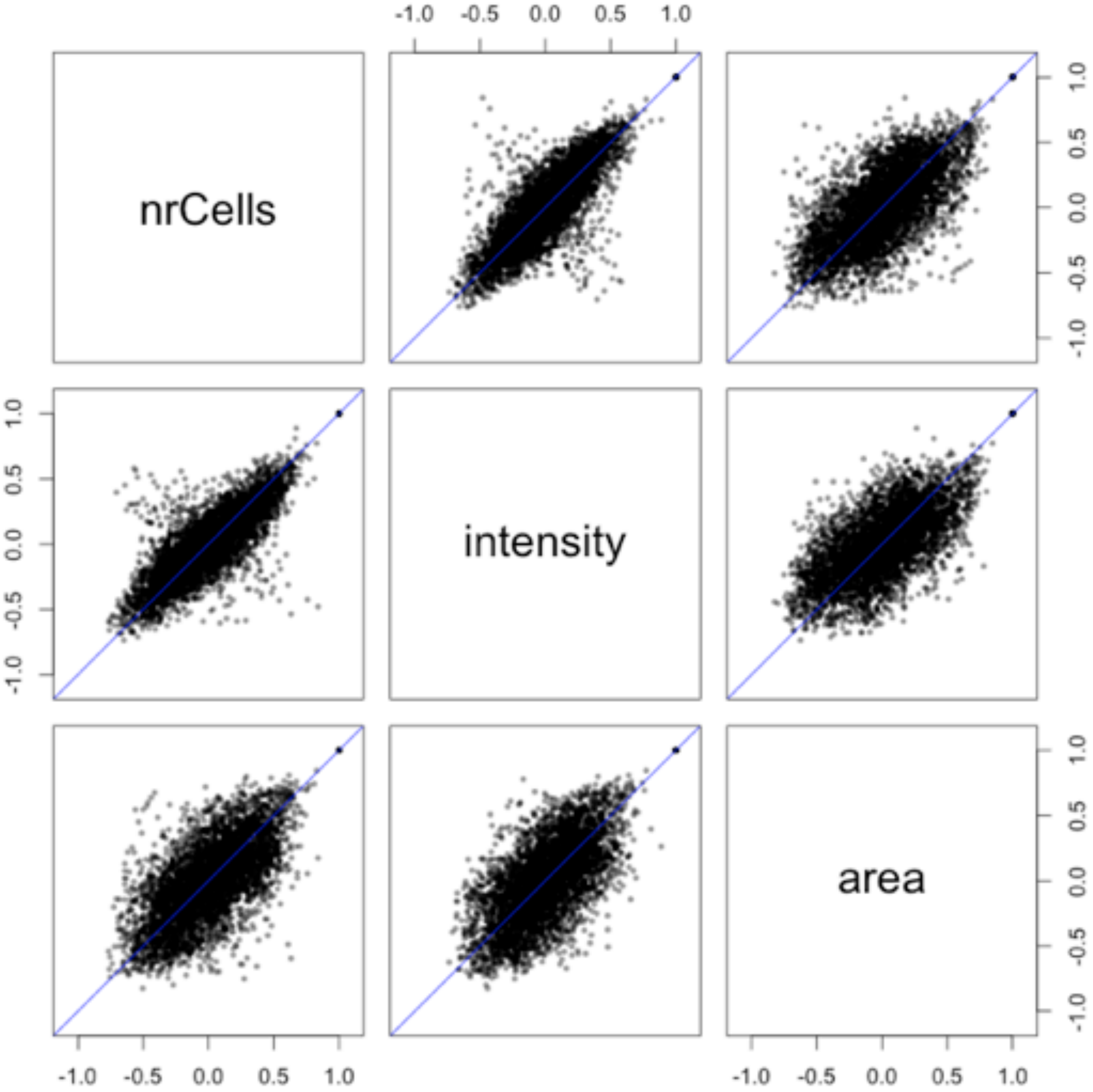
Hidden Markov Model on class labels: parameters summarise the data

Learn HMM on class labels



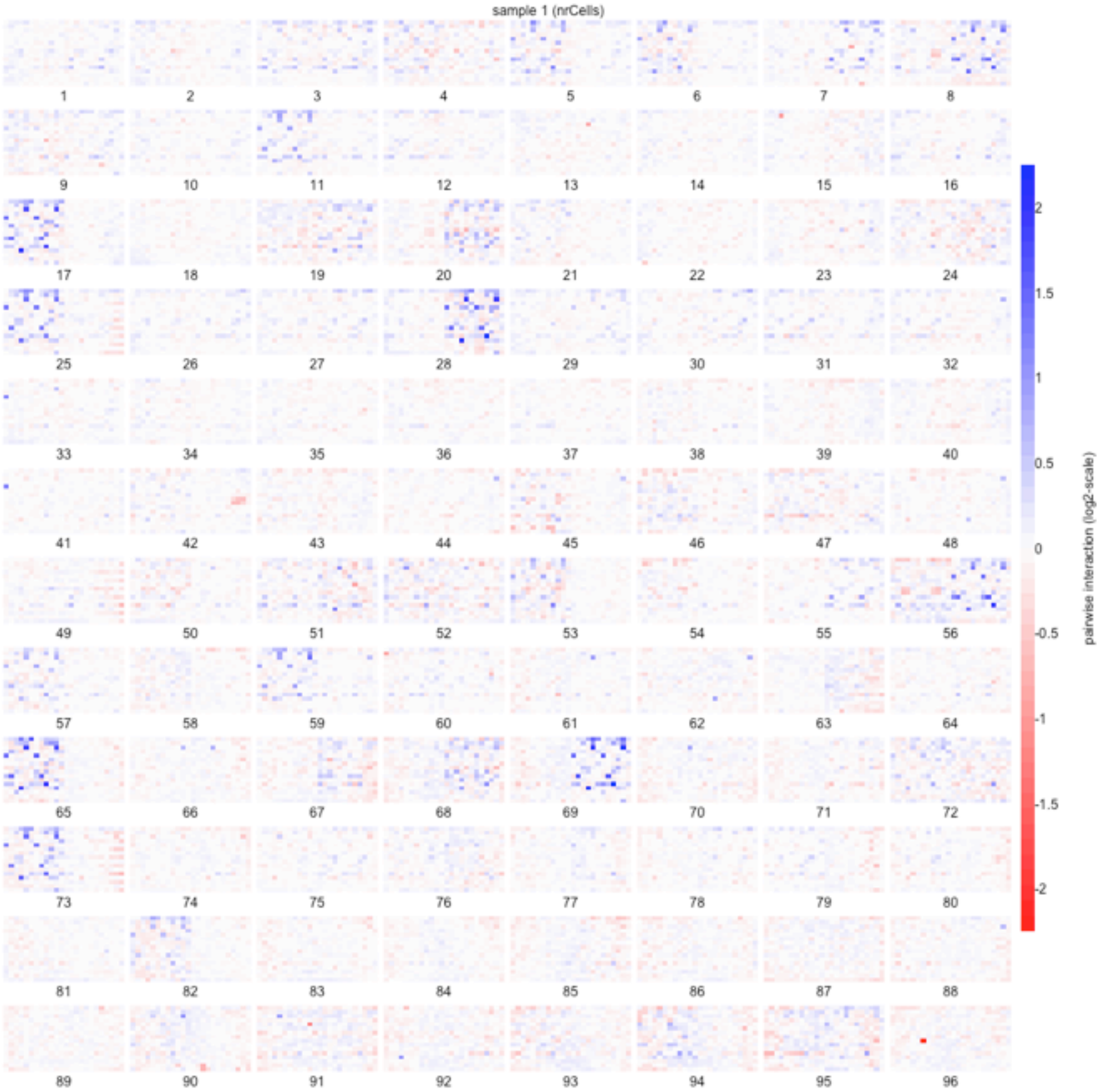


Interaction scores

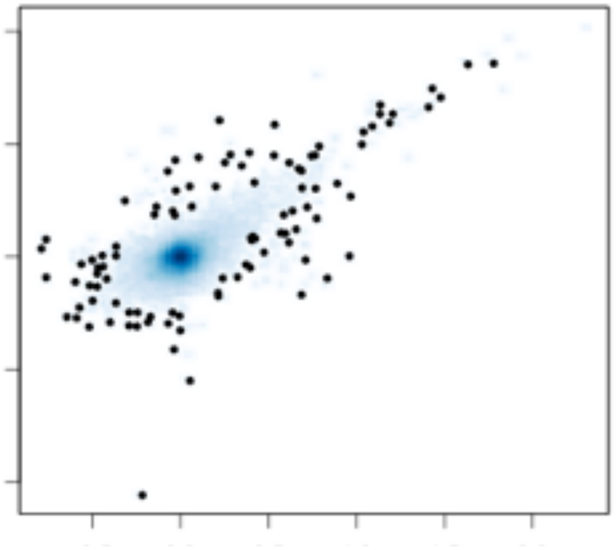


Correlations

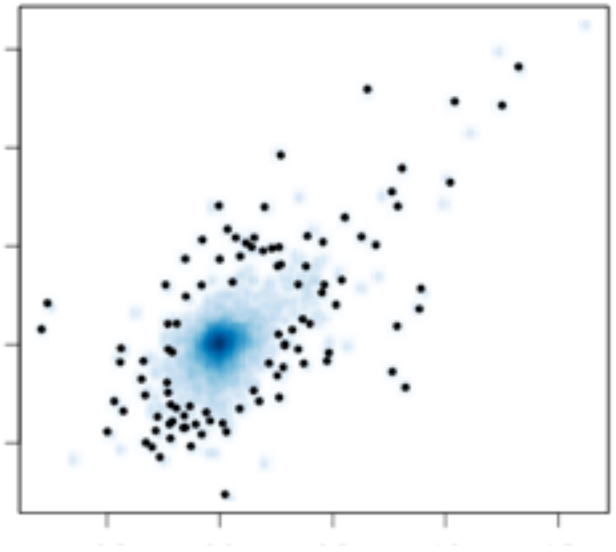
Screen Plot of Interaction Score (#cells)



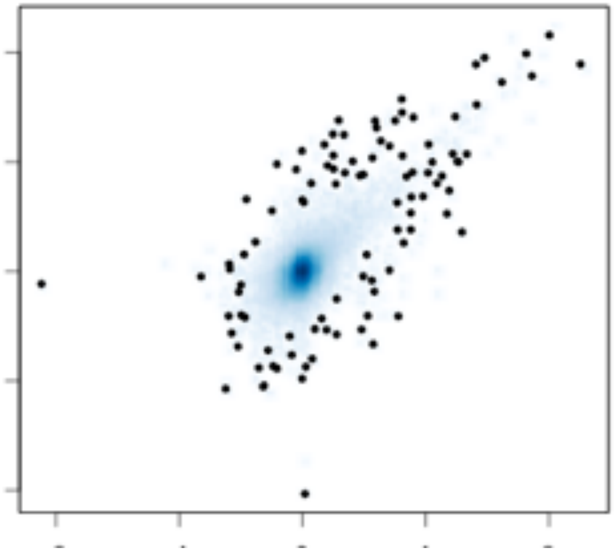
within screen replicates (cor=**0.968**)



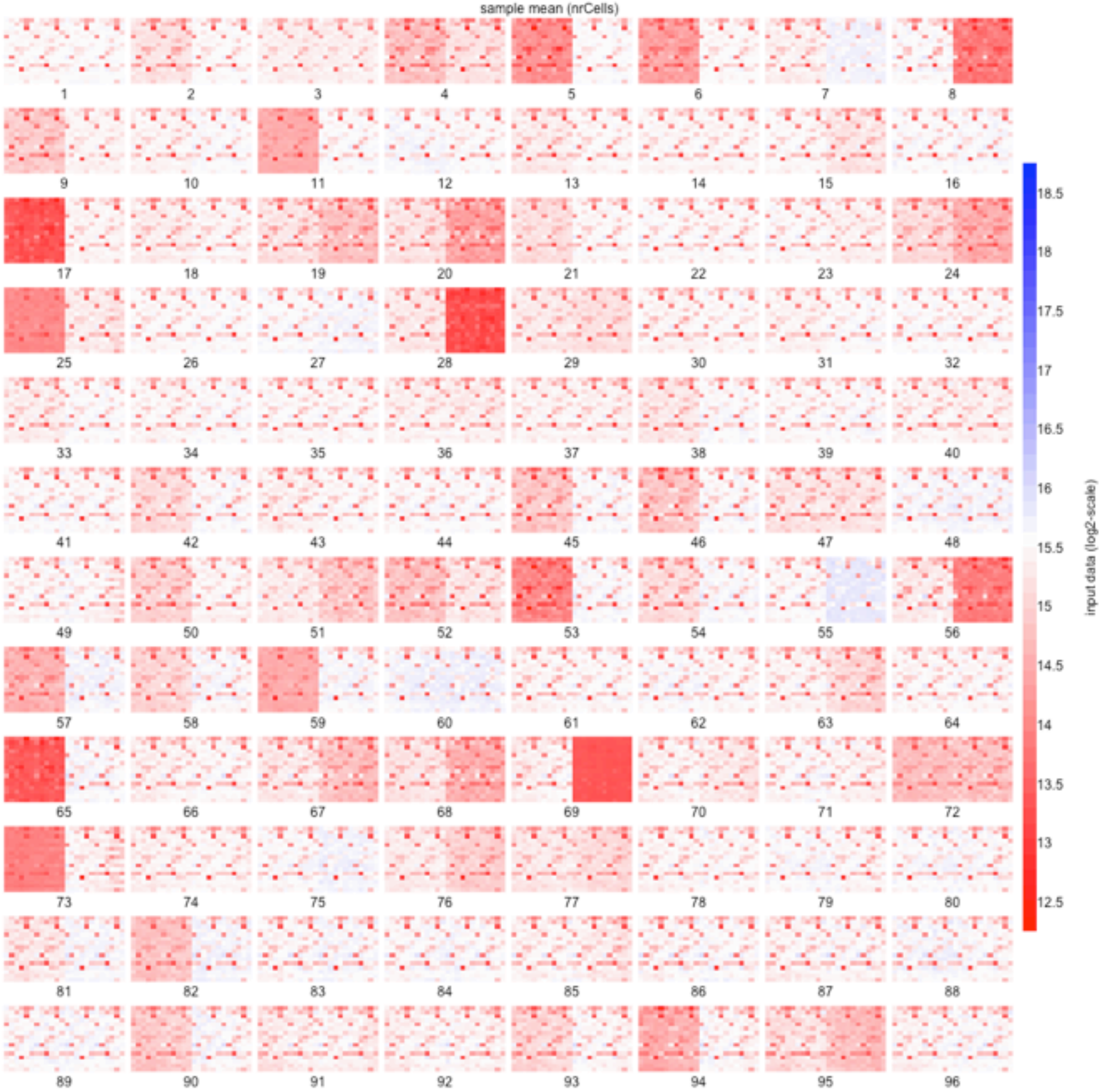
independent daRNA designs (cor=**0.902**)



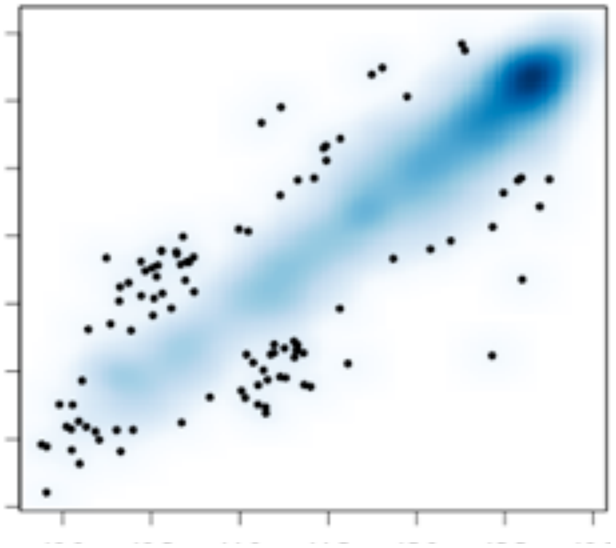
between screen replicates (cor=**0.948**)



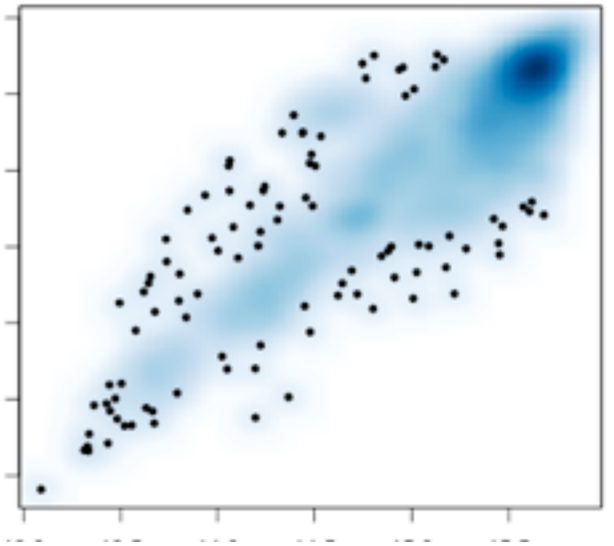
Screen Plot of Read-out (Number of Cells)



within screen replicates (cor=**0.97**)



independent dsRNA designs (cor=**0.90**)



between screen replicates (cor=**0.95**)