Discovering Mammalian Endocytic Pathways with High-Throughput Image-Based siRNA Screens and Network Analysis

Pauli Rämö

Institute of Molecular Systems biology

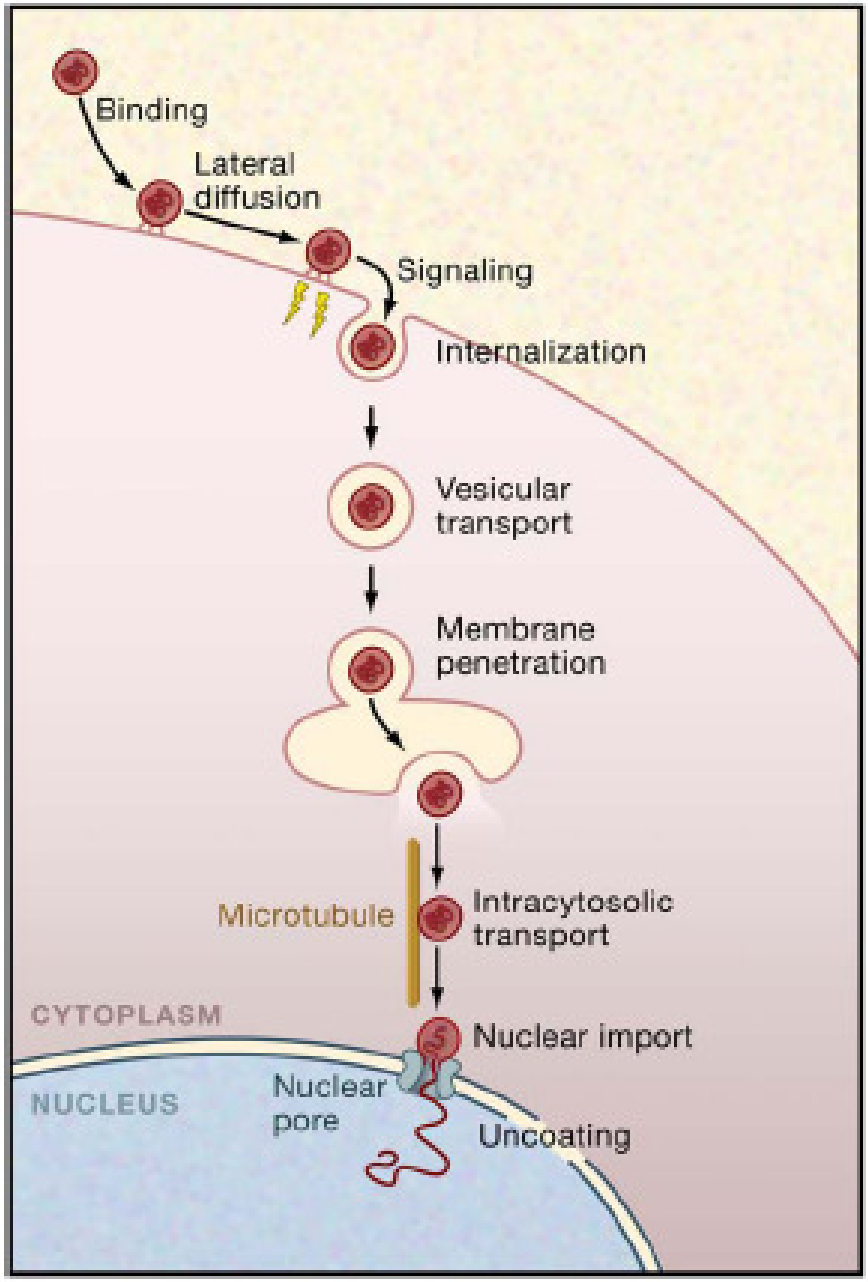
ETH Zürich, Switzerland

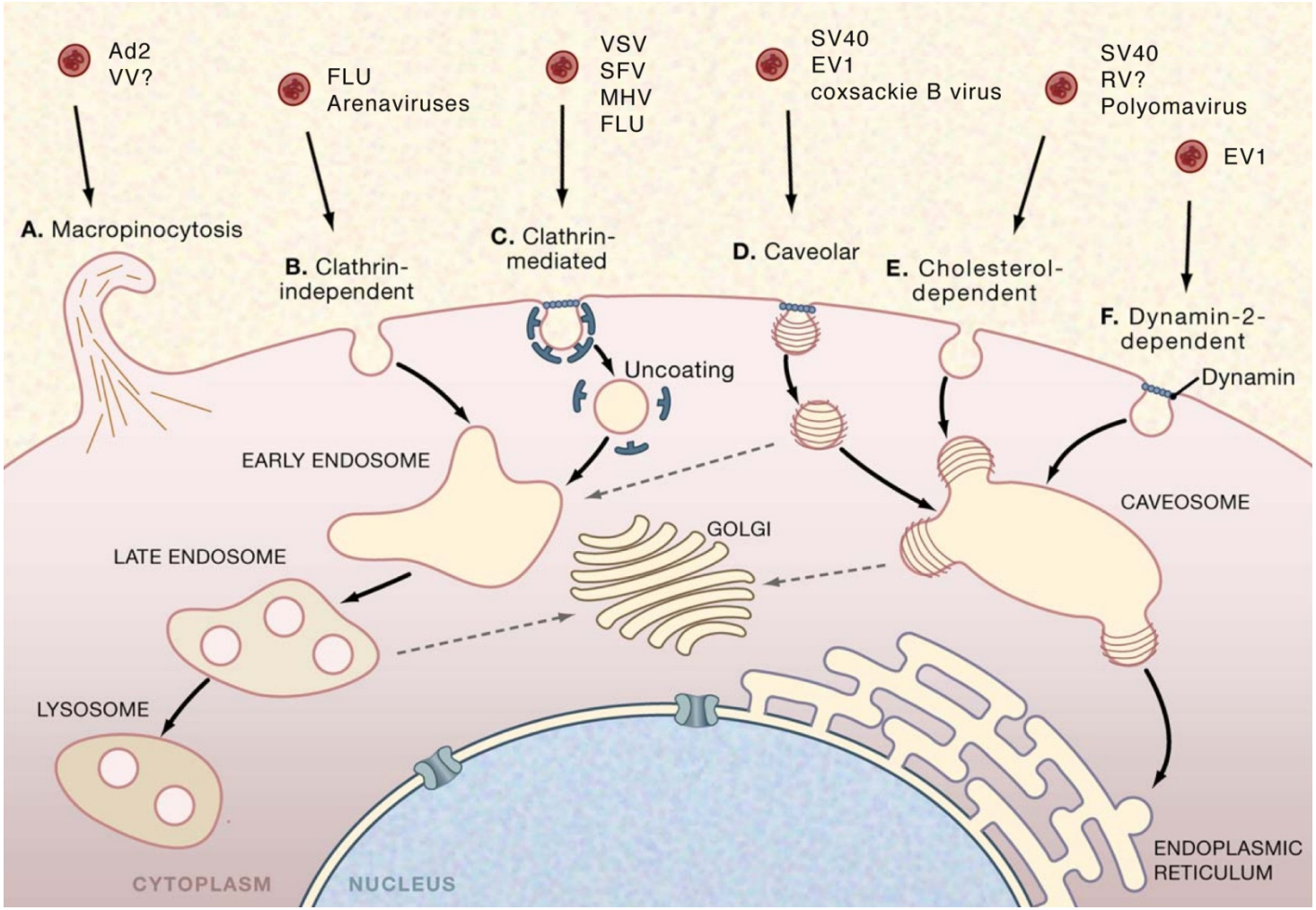
Outline

- Background in endocytosis and virus entry
- Biological assays
  - siRNA technology
  - Virus infection assays
- Data analysis
  - Phenotypical readouts
- Human protein interaction databases
- Network based methods
  - Data analysis, integration and visualization
- Summary

Background in endocytosis and virus entry

- Viruses enter the cell through endocytic pathways
  - Higher organisms such as mammals have many different endocytic pathways
  - Currently, the pathways and the genes involved in the steps of these pathways are only partially known


Binding
Lateral diffusion
Signaling
Internalization
Vesicular transport
Membrane penetration
Microtubule
Intracytosolic transport
CYTOPLASM
Nuclear import
NUCLEUS
Nuclear pore
Uncoating


Ad2
VV?
FLU
Arenaviruses
VSV
SFV
MHV
FLU
SV40
EV1
coxsackie B virus
SV40
RV?
Polyomavirus
EV1
A. Macropinocytosis
B. Clathrin-independent
C. Clathrin-mediated
D. Caveolar
E. Cholesterol-dependent
F. Dynamin-2-dependent
Uncoating
Dynamin
EARLY ENDOSOME
CAVEOSOME
GOLGI
LATE ENDOSOME
LYSOSOME
CYTOPLASM
NUCLEUS
ENDOPLASMIC RETICULUM

siRNA technology

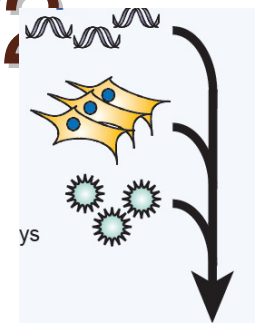
- Small interfering RNAs (siRNA) are double-stranded RNA molecules that can be artificially created to knock-down specific genes
- Companies have generated libraries of siRNAs that target most or a subset of human genes
- Problems with the siRNA technology
  - Individual siRNAs may have a considerable off-target phenotypes or varying silencing efficiencies
  - In practice, several siRNAs are used to silence each target gene

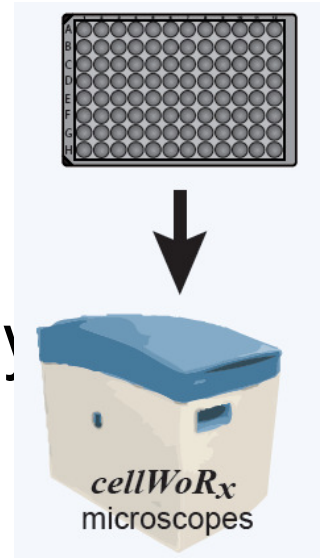
Virus infection assays 1/2

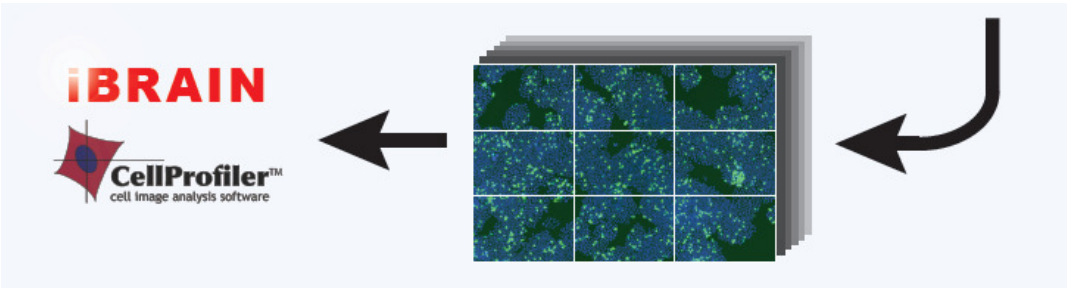
- Our goal is to find **host** gene knock-downs that cause changes in the level of infection of various viruses
  - Integrating the data will help us to discover which pathways different viruses use and which genes constitute the main components of these pathways
- In our assays we use:
  - Human culture cells (e.g. HeLa cells)
  - 7000 genes each targeted with 3 individual siRNAs
    - Results in more than 200 384-well plates and 1TB of measurement data per virus
  - So far 6 different viruses

Virus infection assays 2/2

- Virus infection assays
  - siRNA transfection
  - Cell seeding
  - Virus infection
  - Cell fixation and staining
- Automated low resolution microscopy
- Automated image analysis

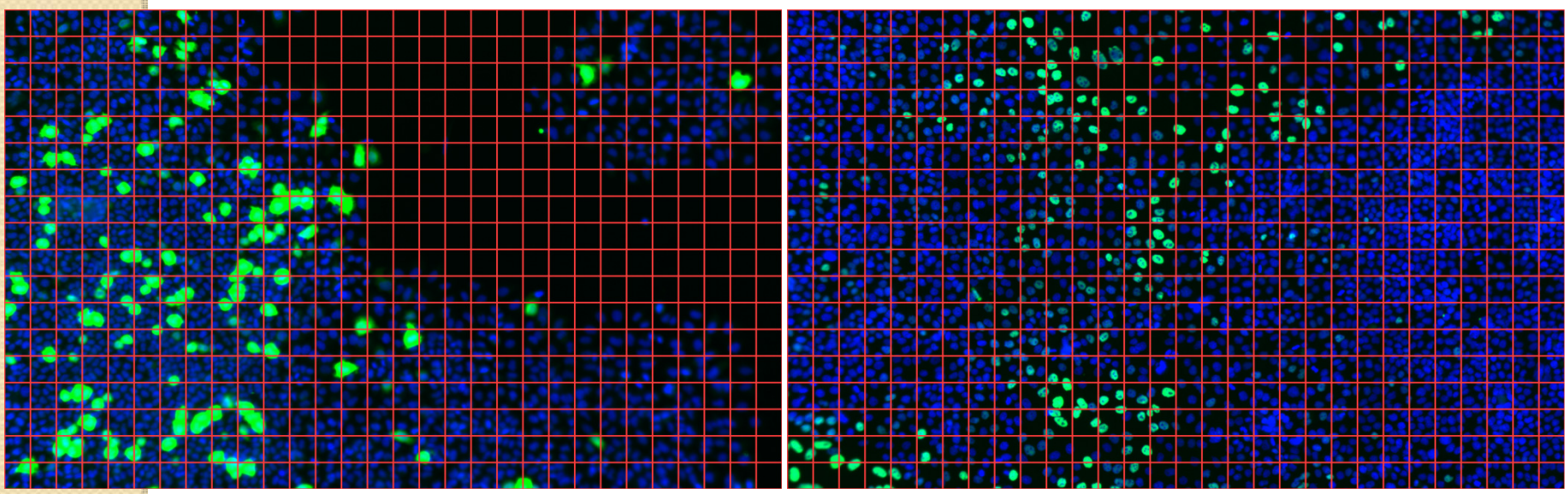






SFV

SV40



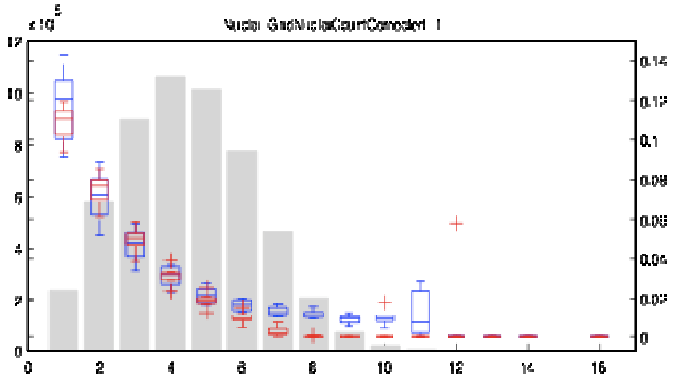
Data analysis

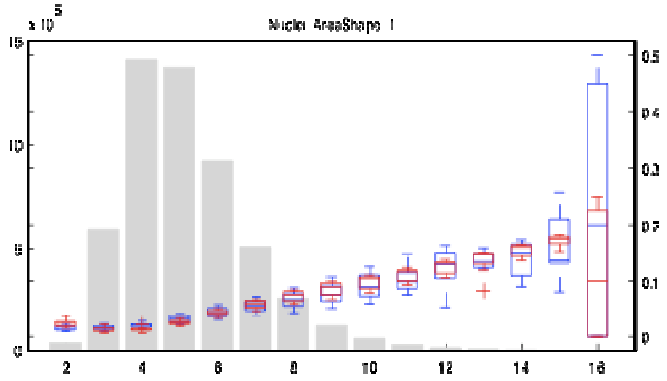
- Image analysis
  - We use CellProfiler toolbox to perform our image analysis on a computer cluster
    - Cell segmentation
    - Feature extraction
    - Detection of infected cells
- Data coming from large-scale screens must first be checked for quality and normalized
- We have developed statistical hit estimation methods for large-scale siRNA screens
  - We measure different phenotype “hits” for each knock-down

Phenotypical readouts

- Image based siRNA assays allow wide possibilities for phenotypical readouts
  - Level of infection
  - Support vector machine (SVM) classification of cellular phenotypes
    - Interphase, mitotic, or apoptotic nuclei
    - Different staining patterns
  - We discovered that infection patterns depend on cell population properties
    - Local cell density
    - Cell size
    - Total cell number
    - Cells on colony edges
    - We developed a model to correct (and use) these correlations

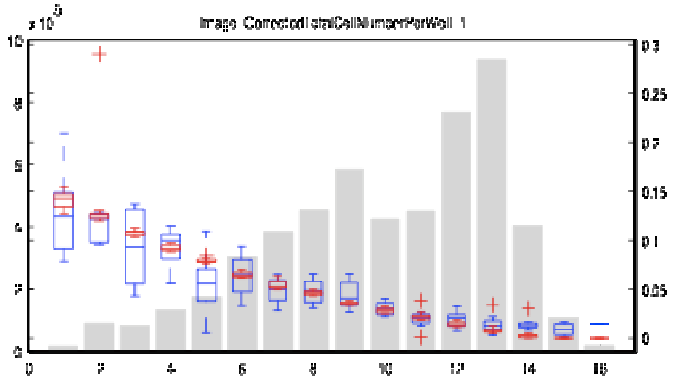
RV_KY : Linear model

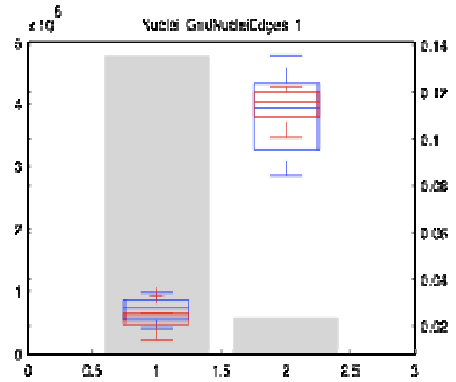


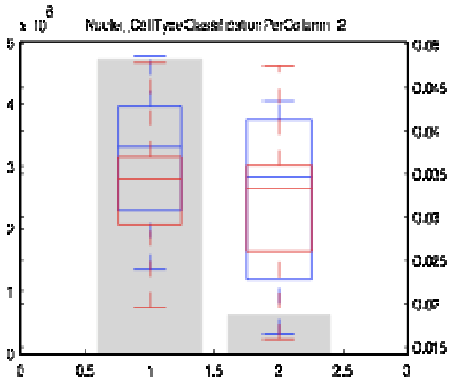


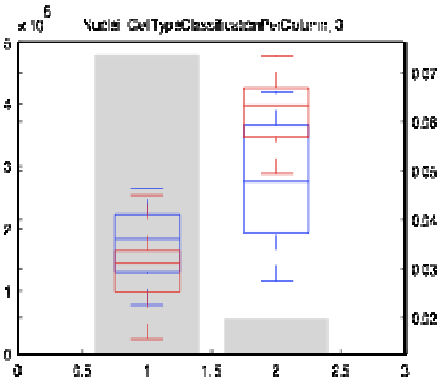
Model dimensions	values
Constant	0.0419
Nuclei, GridNucleiCountCorrected, 1	-0.0058
Nuclei, AreaShape, 1	0.0078
Nuclei, GridNucleiEdges, 1	0.0424
Image, CorrectedTotalCellNumberPerWell, 1	-0.0040
Nuclei, CellTypeClassificationPerColumn, 2	0.0245
Nuclei, CellTypeClassificationPerColumn, 3	0.0519

LinearRegression: First order polynomial
NumOfWeightedIndices: 7695
W: 1CN^1
X: Binned
Y: Infection index (pooled over all plates)









Human protein interaction databases

- Human protein-protein and genetic interaction networks are much worse known than for other model organisms
- Some protein-protein interaction databases are promising
  - HPRD (human protein reference database), interactions from literature
  - STRING, protein interactions for example from literature, species orthologs, and co-expression assays
  - Ingenuity, a commercial literature based interaction database and data visualization tool

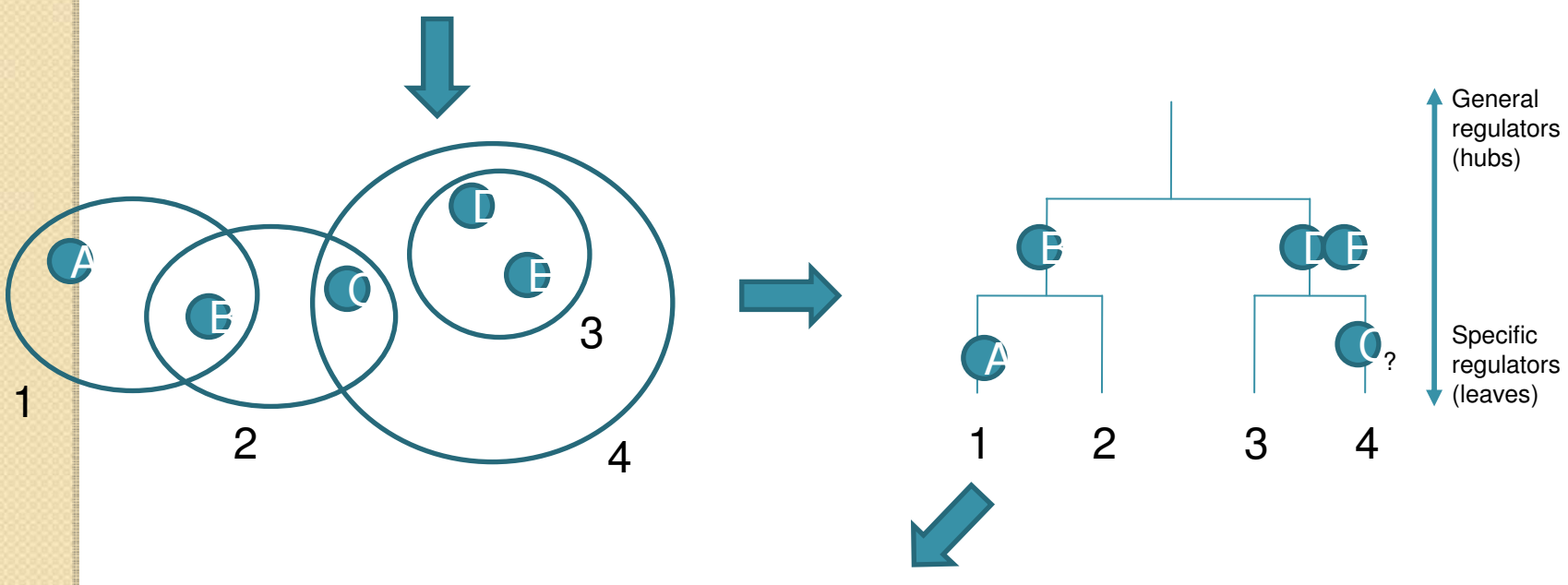
Network based methods

- Our network based approaches include
  - Using information from existing protein-protein interactions as a-priori knowledge for hit detection
    - Care should be taken since the low quality of the network data
  - Finding enriched subnetworks
  - Integrate networks with our own phenotypical data
  - Finding correlations between the data and network properties (e.g. hit probability vs. node degree)
  - Data visualization

- Some ideas on the following slides…

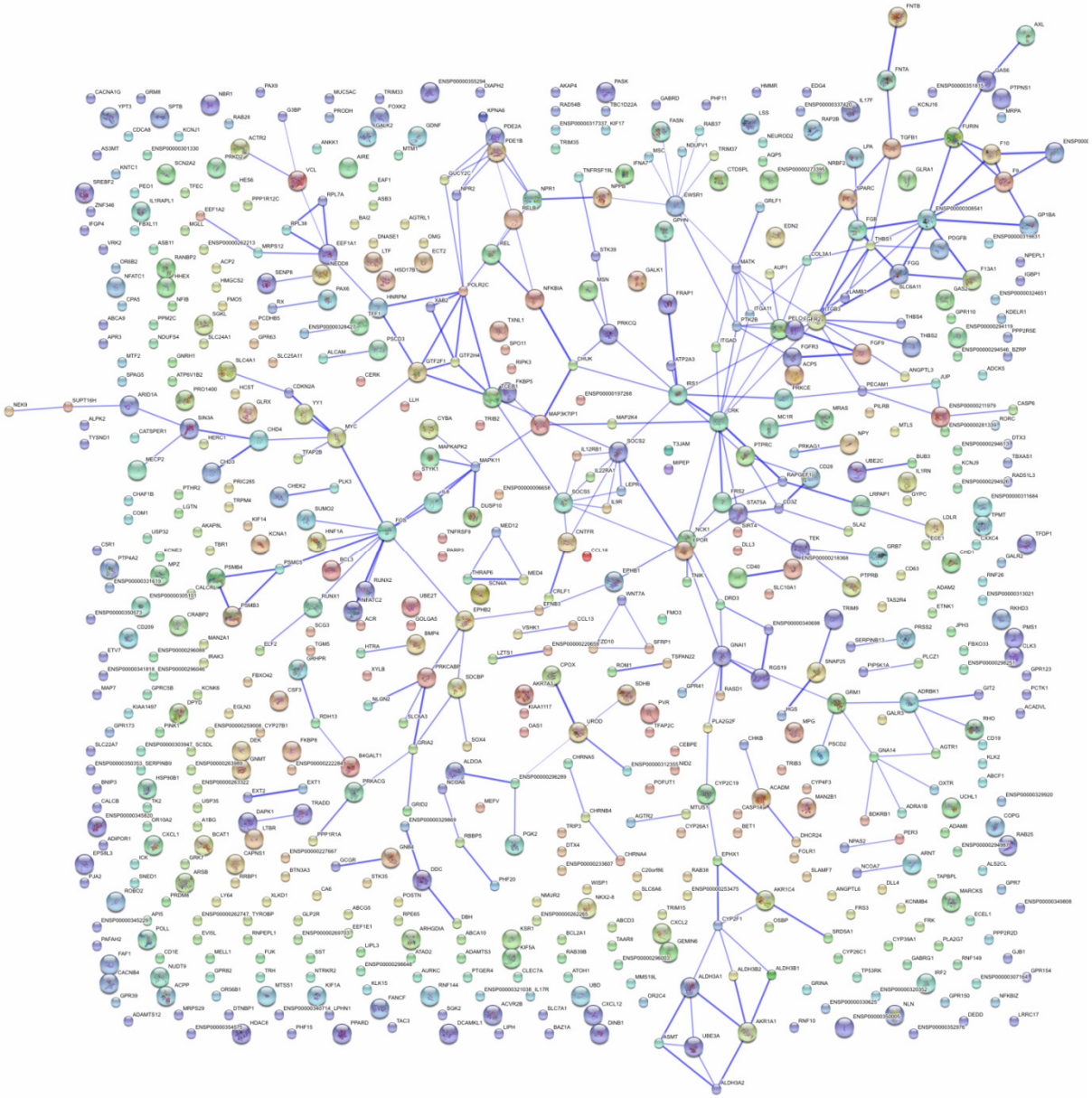
Hierarchical clustering of genes

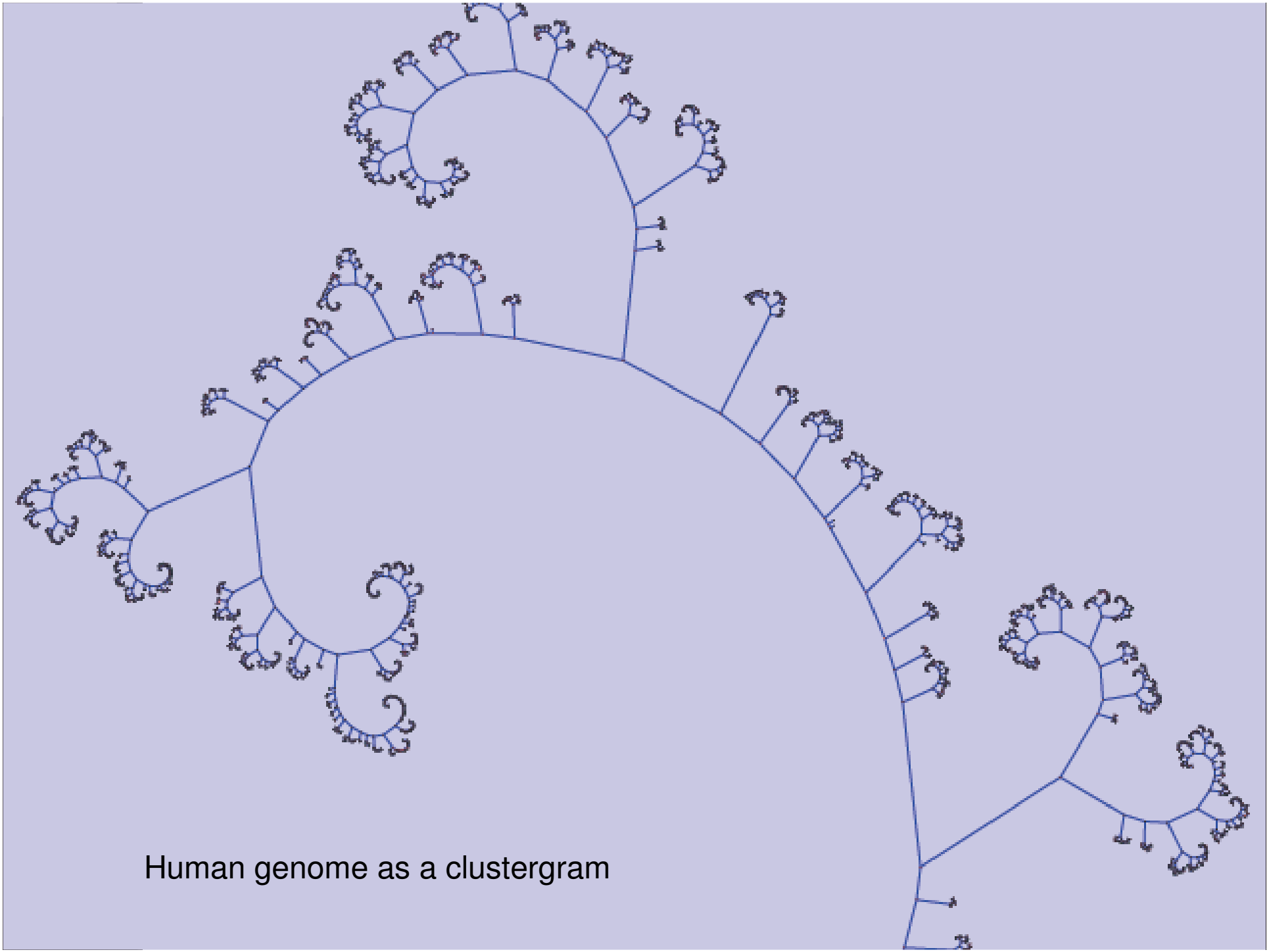
	Gene A	Gene B	Gene C	Gene D	Gene E
Phenotype 1	1	1	0	0	0
Phenotype 2	0	1	1	0	0
Phenotype 3	0	0	0	1	1
Phenotype 4	0	0	1	1	1



- Data integration (interactions, annotations) and ***understanding***

Interactions between hits





Human genome as a clustergram

Summary

- Image based siRNA screens are experimentally, computationally and statistically difficult
  - Provide huge potential for computational biologists
- Network based data analysis approaches are interesting in this context as
  - Natural framework for data integration
  - Finding enriched subnetworks
  - Data visualization

Acknowledgements

- Our lab:
  - Lucas Pelkmans
  - Berend Snijder
  - Raphael Sacher
  - David Lamparter
  - Lilli Stergiou
  - Eva-Maria Damm
  - Mirko Birbaumer
  - Manuel Bauer
  - Herbert Polzhofer
  - Karin Mench
- EMBO
- Human Frontier Science Program