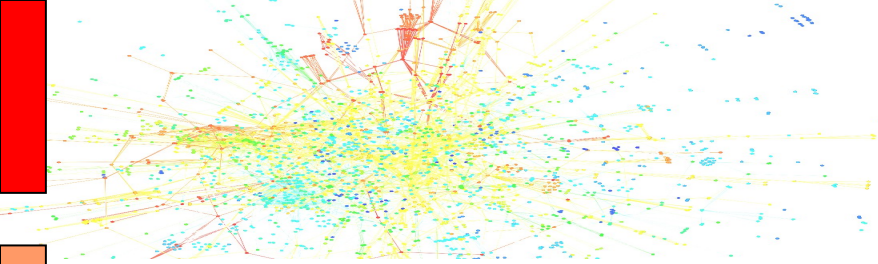




2D Gel Correlation Analysis & Perspectives On Systems Biology

Dr. Werner Van Belle
Medical Genetics
University Hospital Northern Norway
e-mail: werner@sigtrans.org

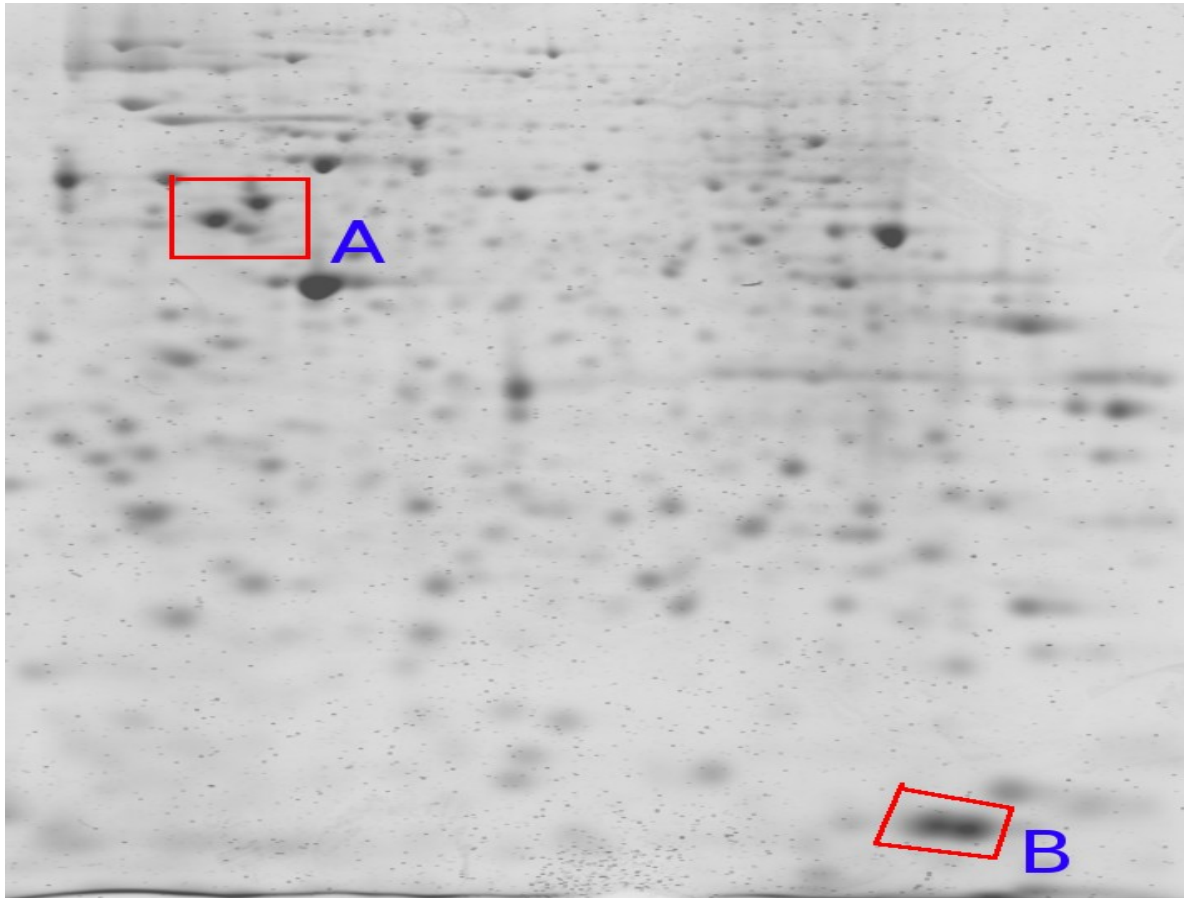


Part 1. 2DE Gel Analysis

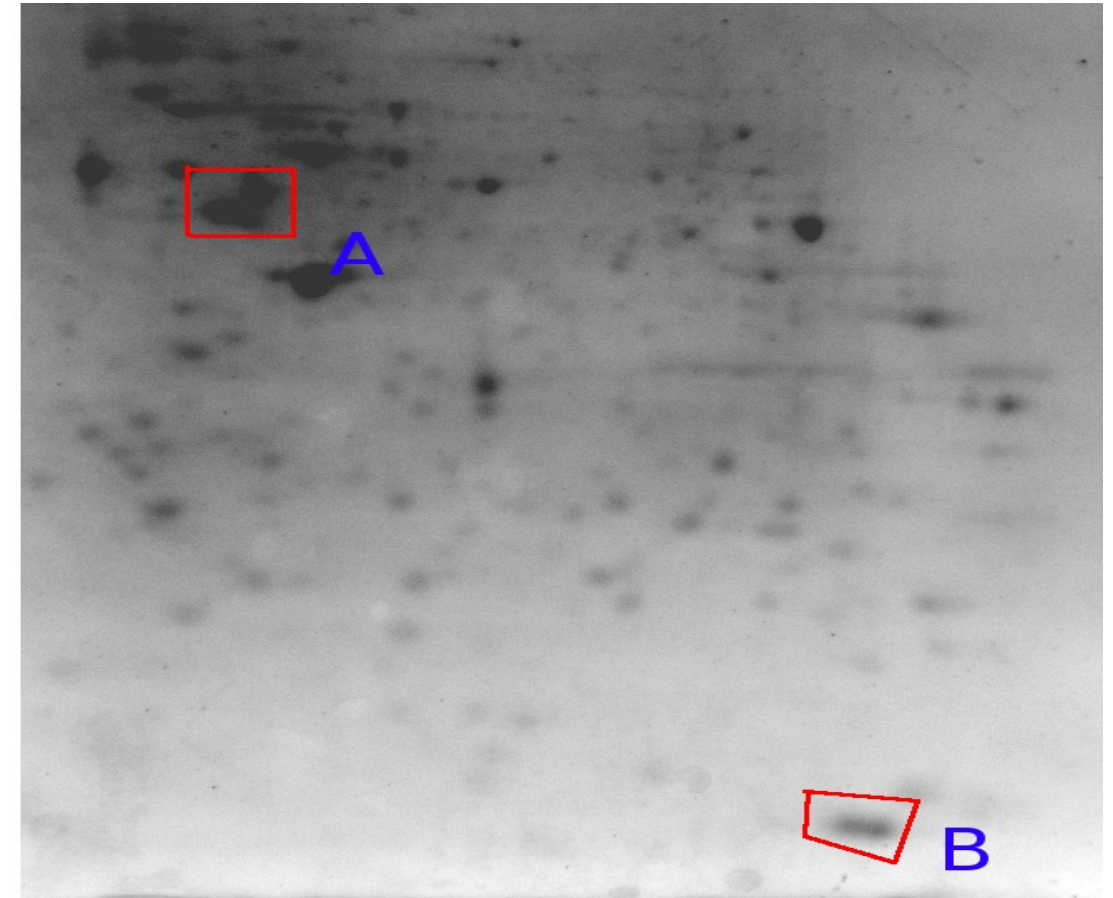
Werner Van Belle
werner @ sigtrans.org

In cooperation with: Bjørn Tore Gjertsen, Nina Ånensen
Ingvild Haaland, Gry Sjøholt, Kjell-Arild Høgda

2D Gels



Patient #1
Age: 57



Patient #2
Age: 46

Initial Problem

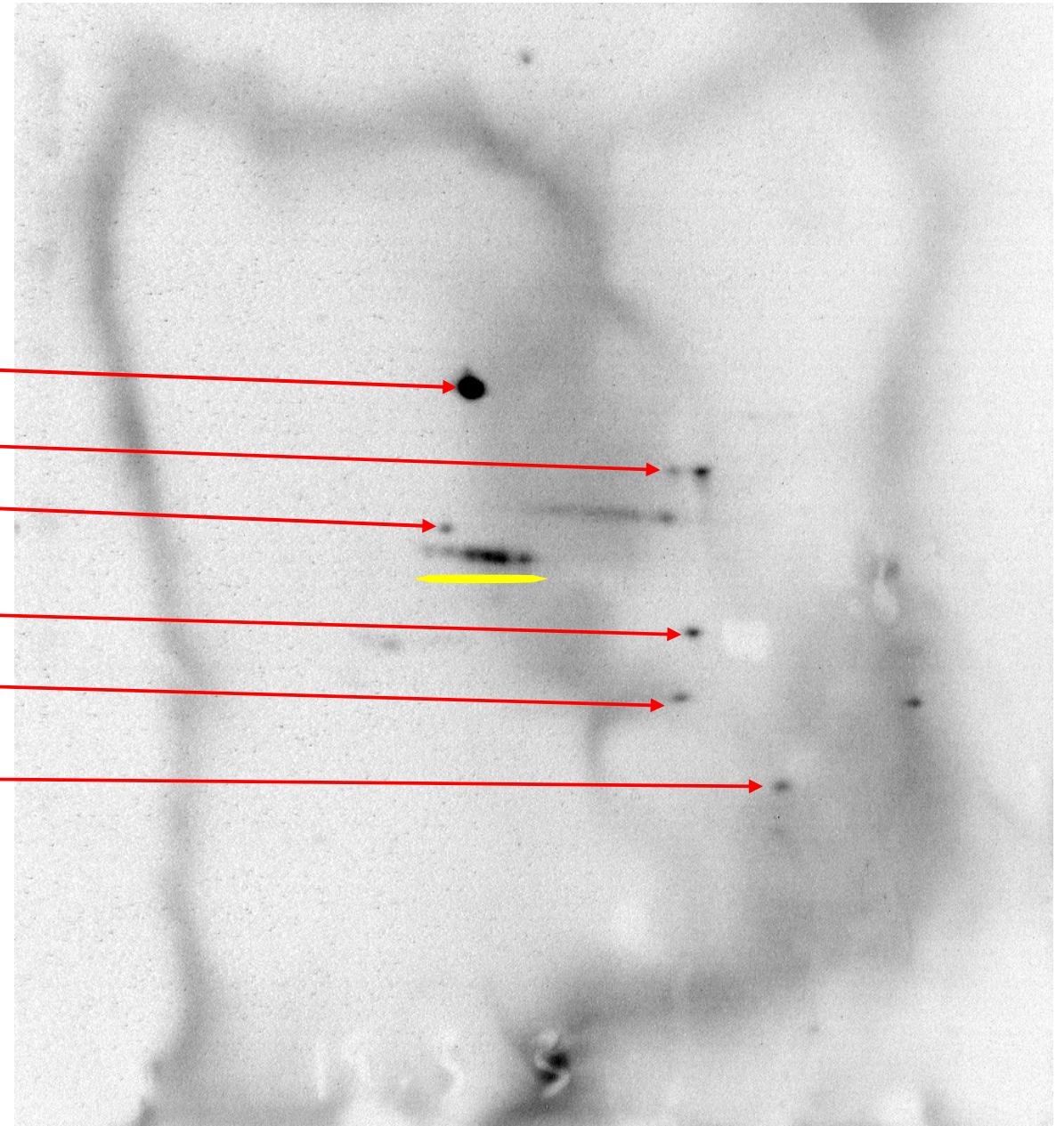
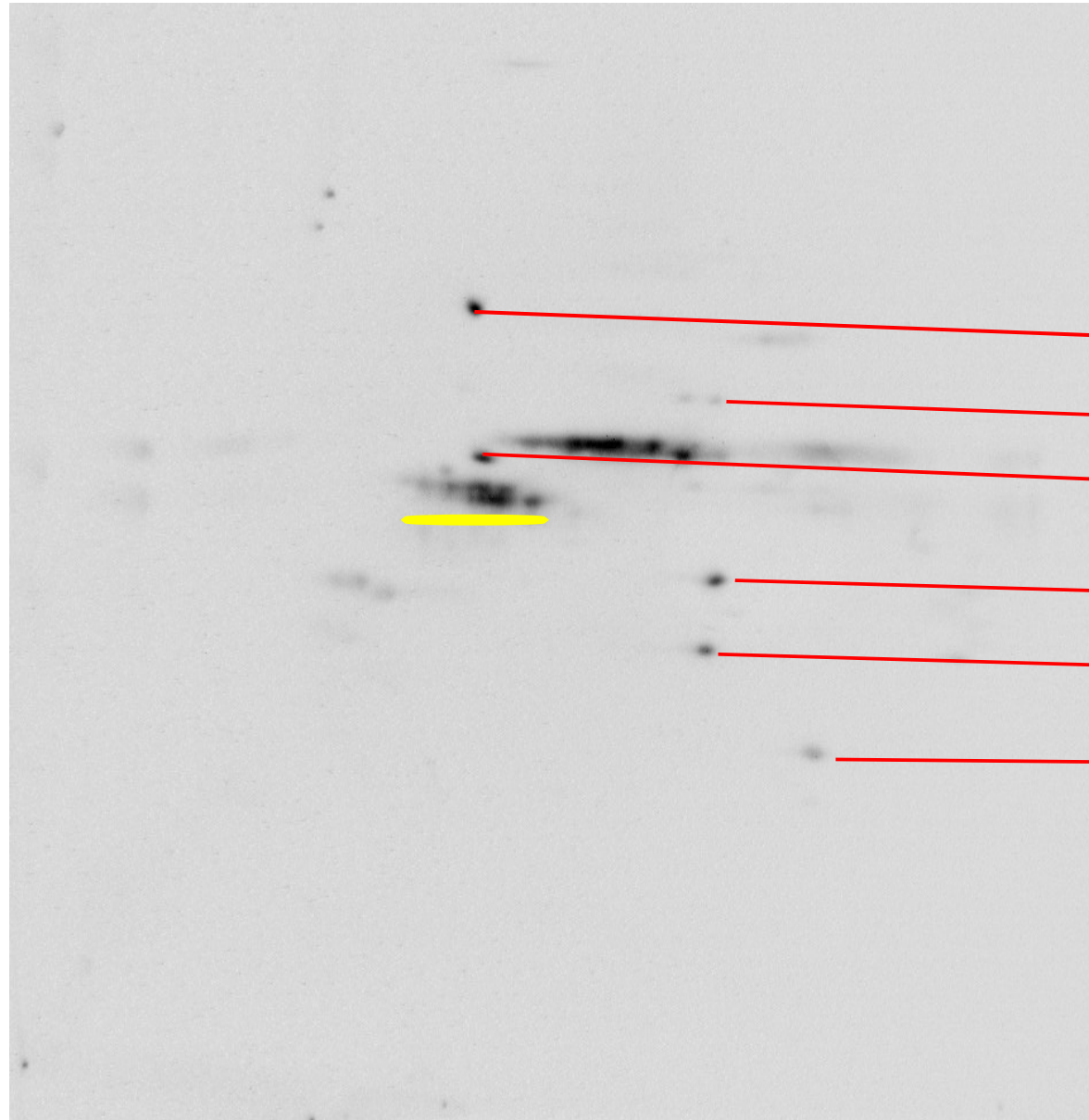
- The question we were asked
 - Is there a relation between various parameters of AML/ALL cancer patients and their P53 biosignatures / isoforms ?
- Gels: +/- 97 gel images of different patients
- Biological Parameters:
 - FAB Classification (AML/ALL), AML Class, Flt3 (WT/ITD)
 - Resistance AML, Resistance ALL, Survival AML, Survival ALL
 - BCL2, Stat5 GMCSF, Stat3 IL3, Stat1 Ifng, CD4, C34

Standard Solution

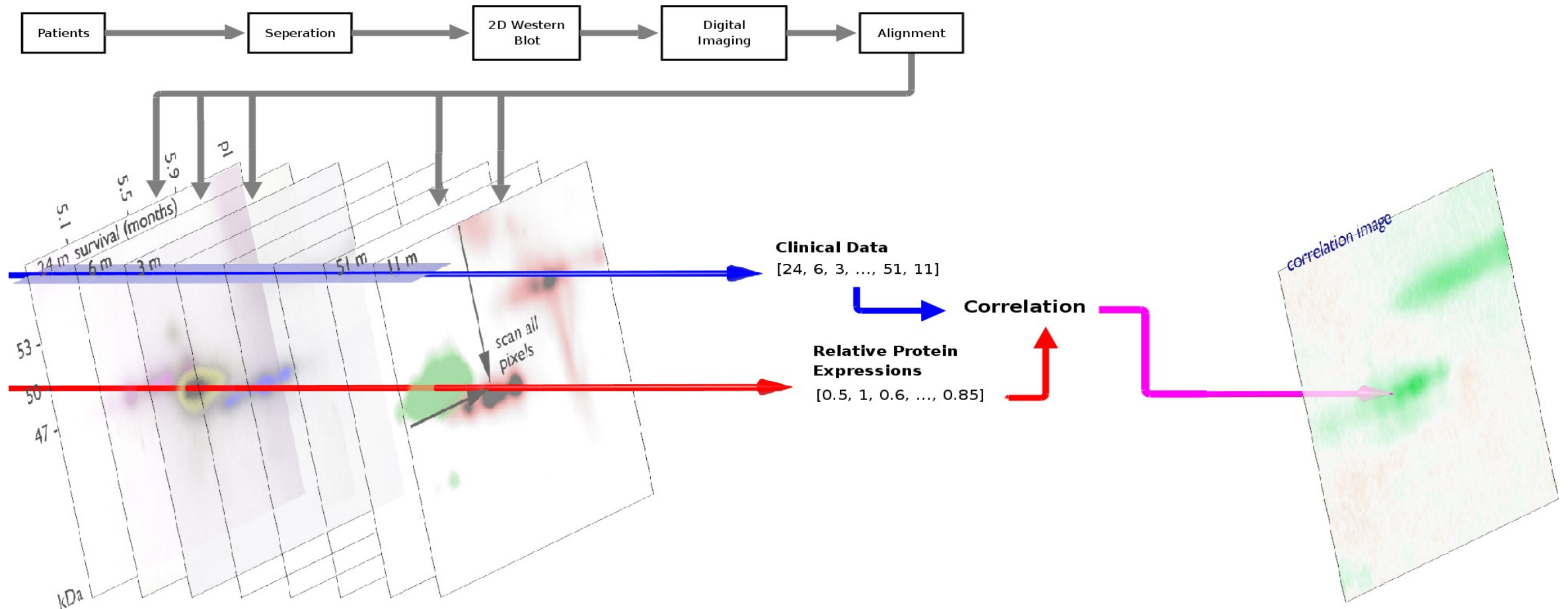
- Detect Spots, Measure Spot Volumes, Compare
- Non Trivial Solution
 - Spot identity unknown, often no calibration spots
 - Manual interpretation dangerous; shifts of spots are difficult to interpret
 - Some PTM influence spot positioning, complicating the matter

Complicated method
Tedious work
Less than optimal results

Manual Comparison



2D Gel Analysis





Step 1: Alignment & Registration

Step 1: Alignment and registration

The method requires proper direction and alignment of all gels. Presence of calibration spots facilitates this process, otherwise techniques such as Hough transformation [26, 52] for gel direction measurement and cross correlation [53] for multiple gel alignment can be used. Once the gels are aligned, further basic warping and registration [45] techniques are useful to account for small shifts between the different gels. The aligned images are denoted A'_z .



Alignment of Multiple Gels

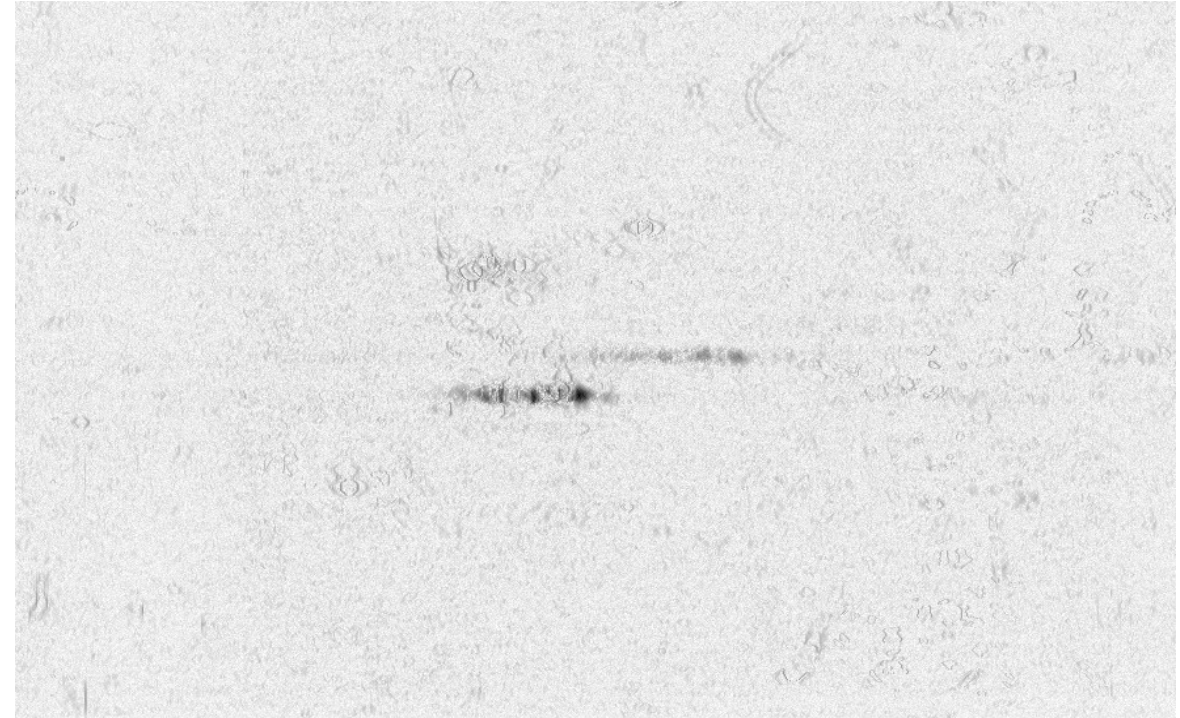
- Idea Cumulative Superposition
 - take first gel, superimpose second gel
 - take third gel, superimpose on projection of previous gels
 - repeat process for all gels

This does not work,
we merely find a suitable superposition
to reflect the first images.

Cumulative Superposition

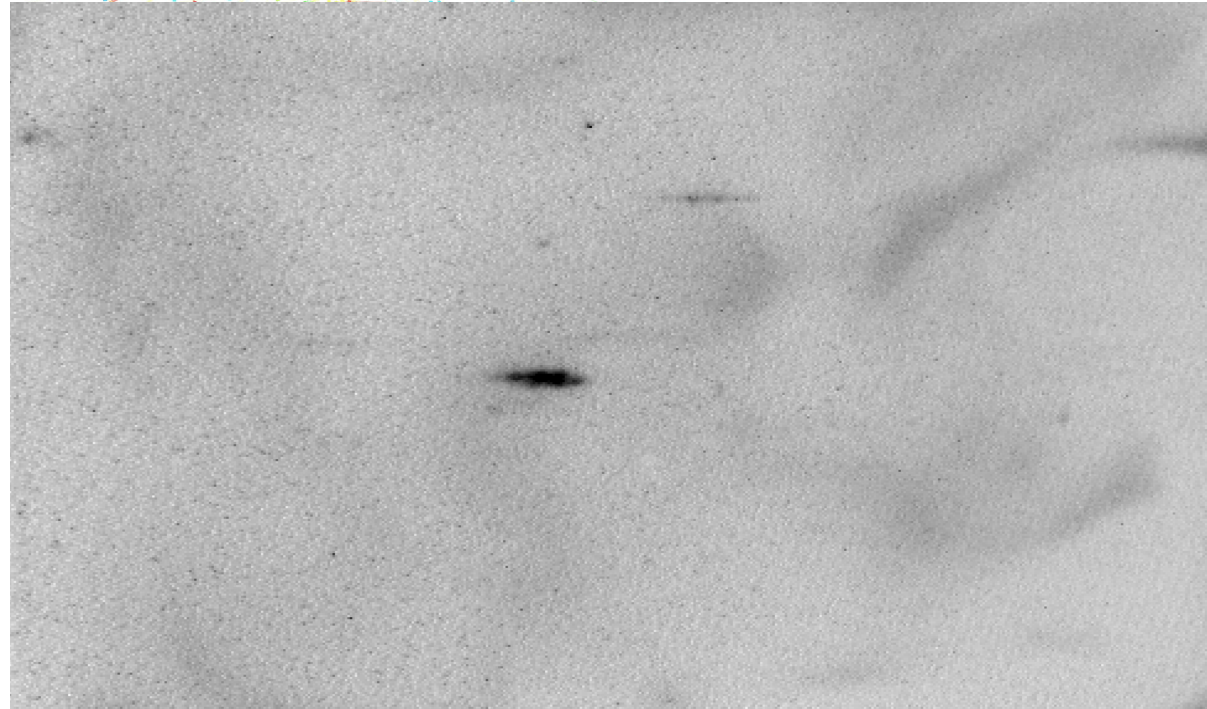


Initial 2DE Gel Image

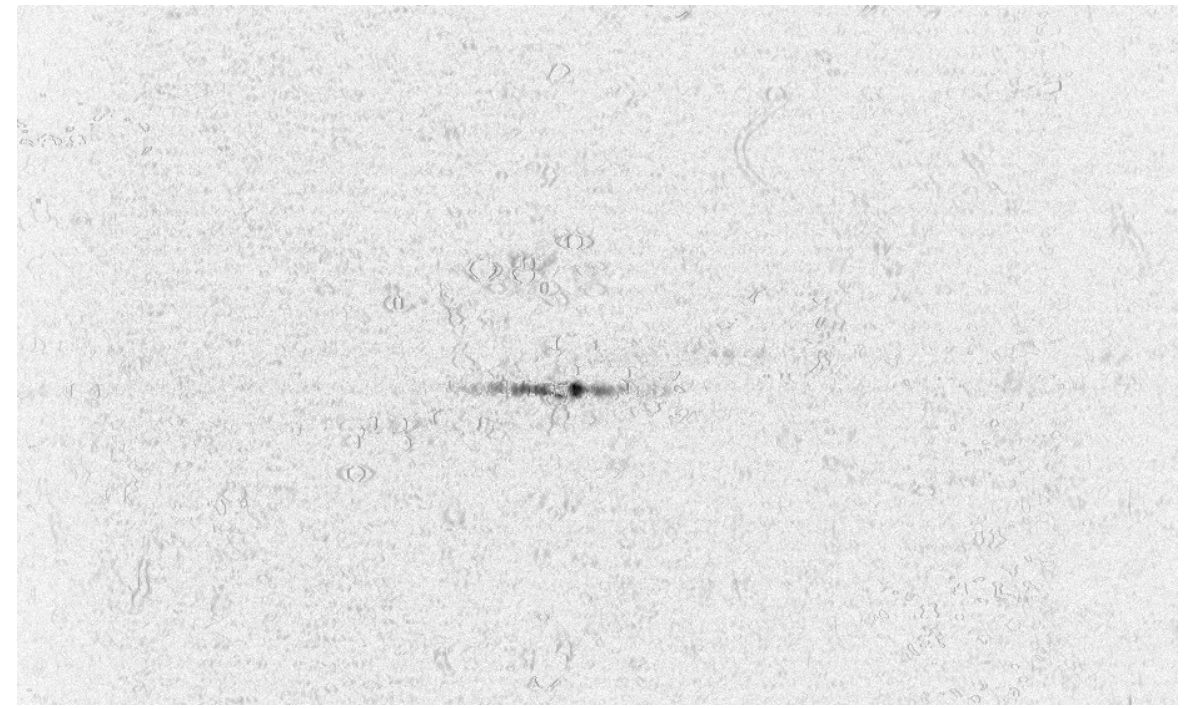


Final Overlay Image

Cumulative Superposition



Initial 2DE Gel Image



Final Overlay Image

Multi Gel Alignment

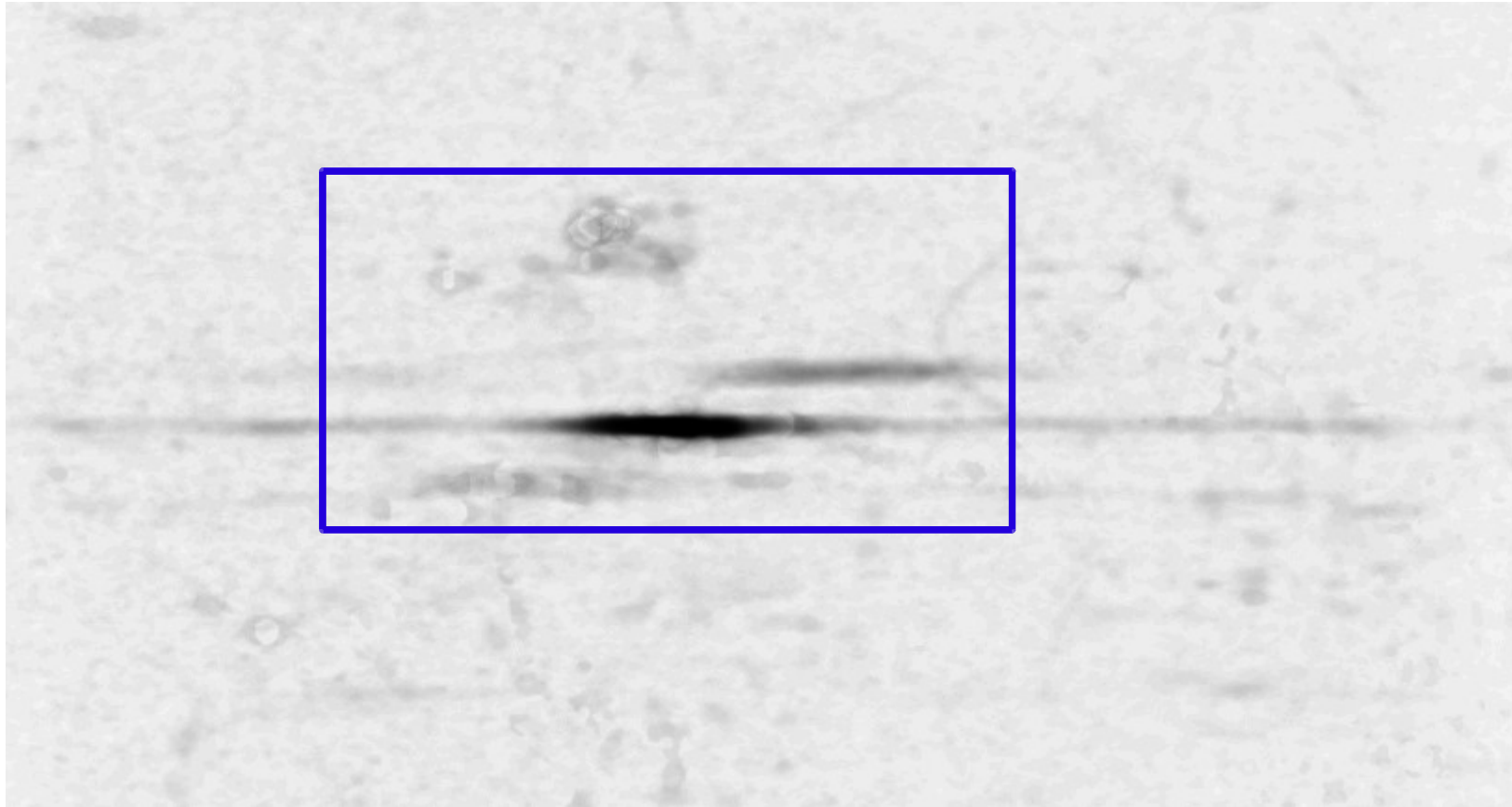
- 1- align all image pairs -> X.X alignments
- 2- find an optimal (x,y) position that minimizes the overall alignment error

	A	B	C	D	E
A		(50,80)	(0,-20)	(30,5)	
B	(2,45)		(-12,0)	(-12,70)	
C	(23,-156)	(15,-73)			
D					
E					

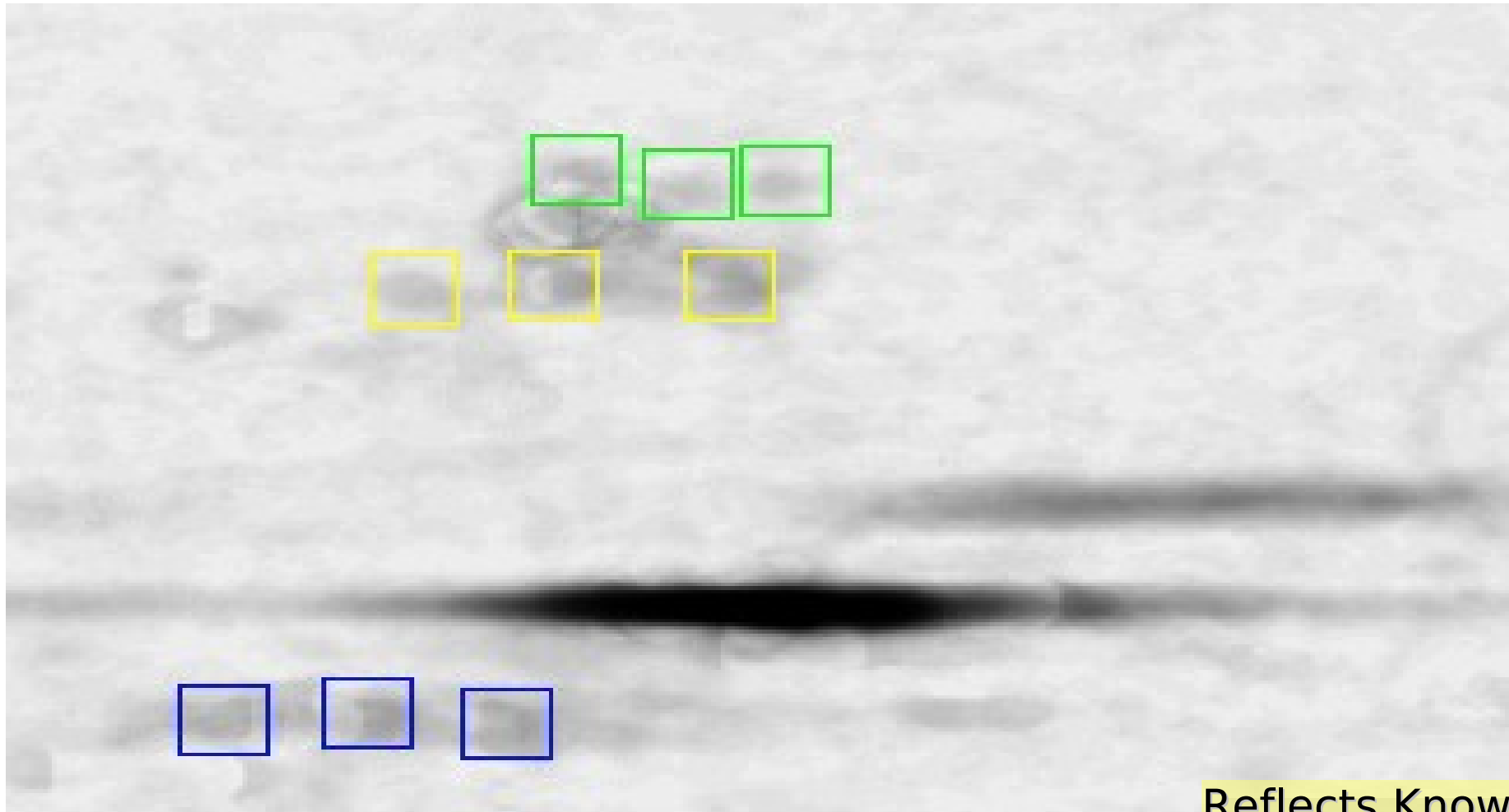
100 images at 1024 x 1024
65011712 operations per cross
correlation
5000 cross correlations
325058560000 operations in total
325.10⁹ FLOP
theoretical = 2.7 hours
practical = 3 days

2D Gel Overlays

Superposition of all images



2D Gel Overlays



Reflects Known Protein
Isoforms

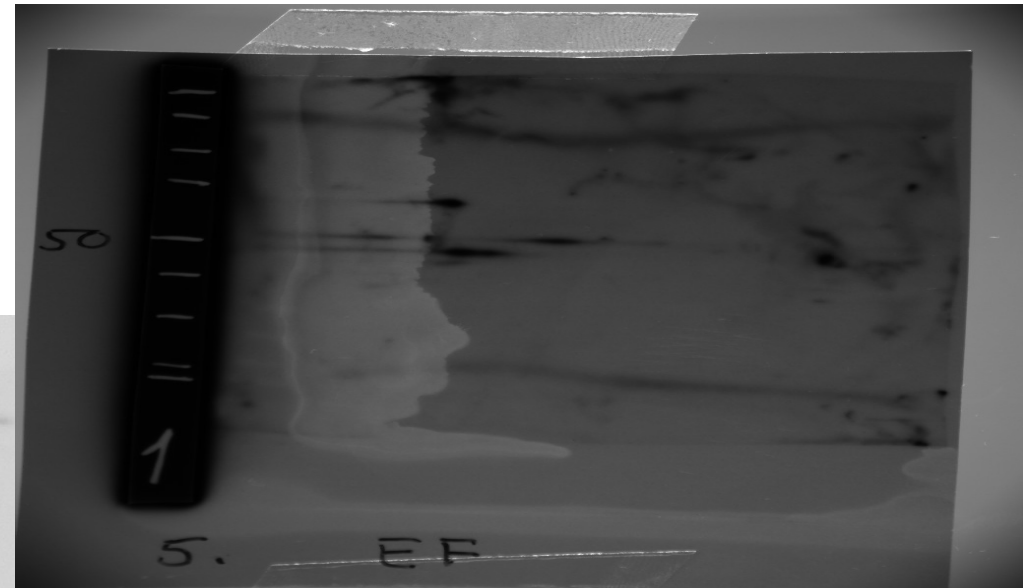
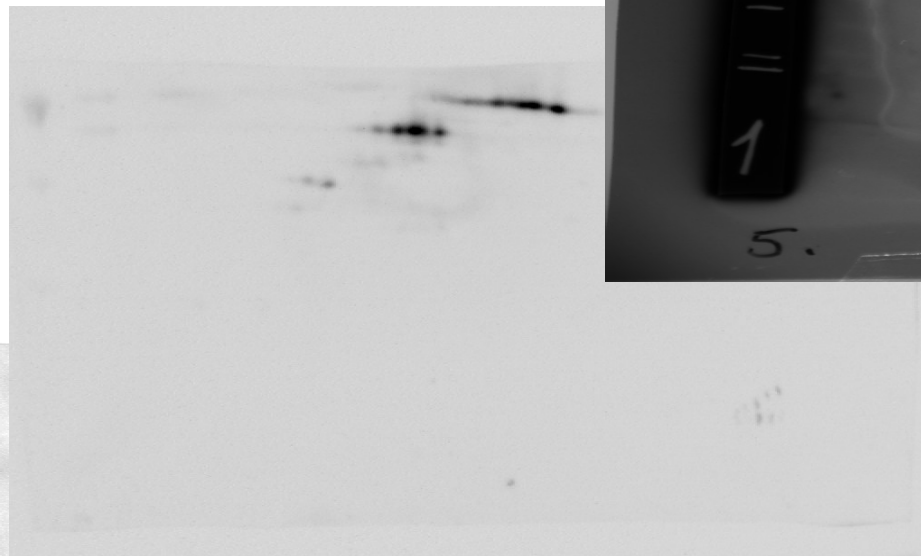
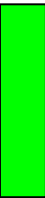
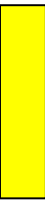


Step 2a: Background Intensity

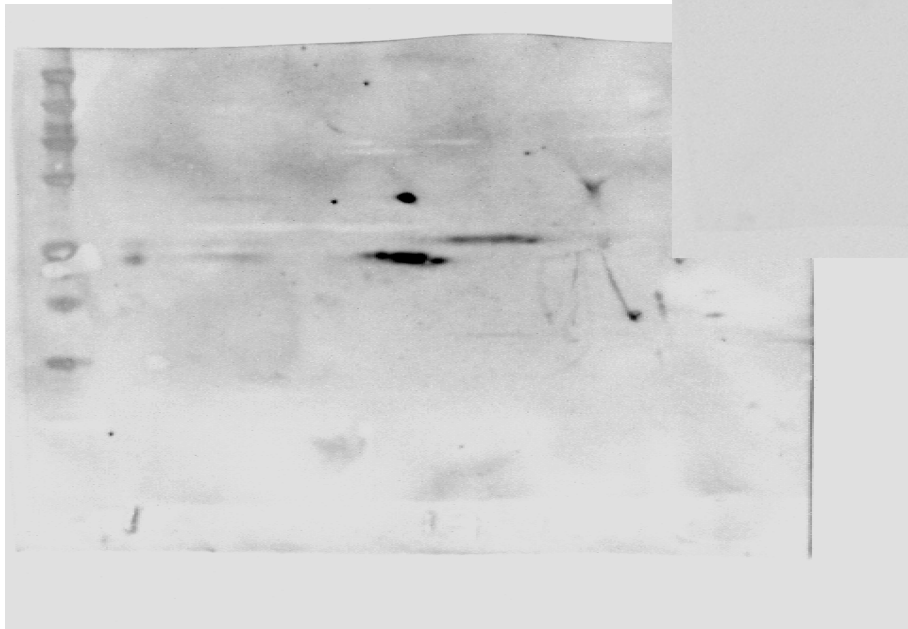
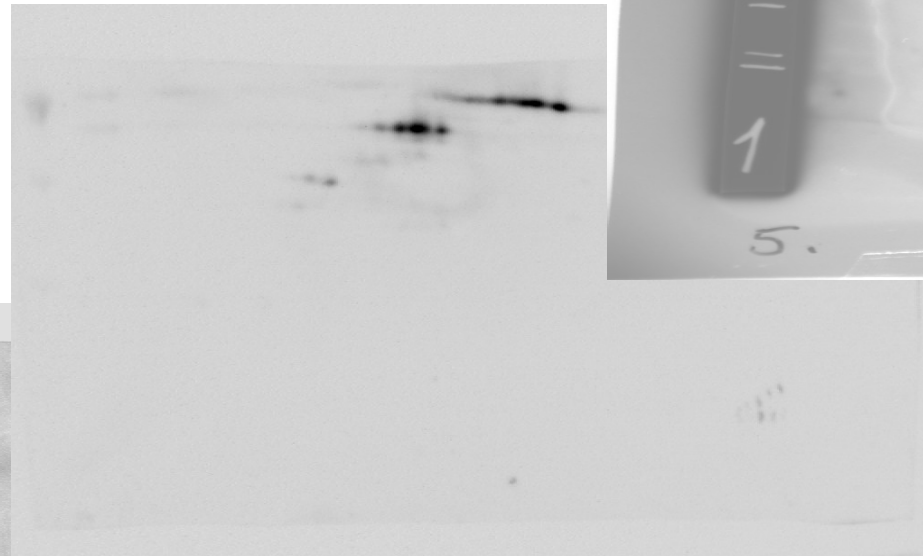
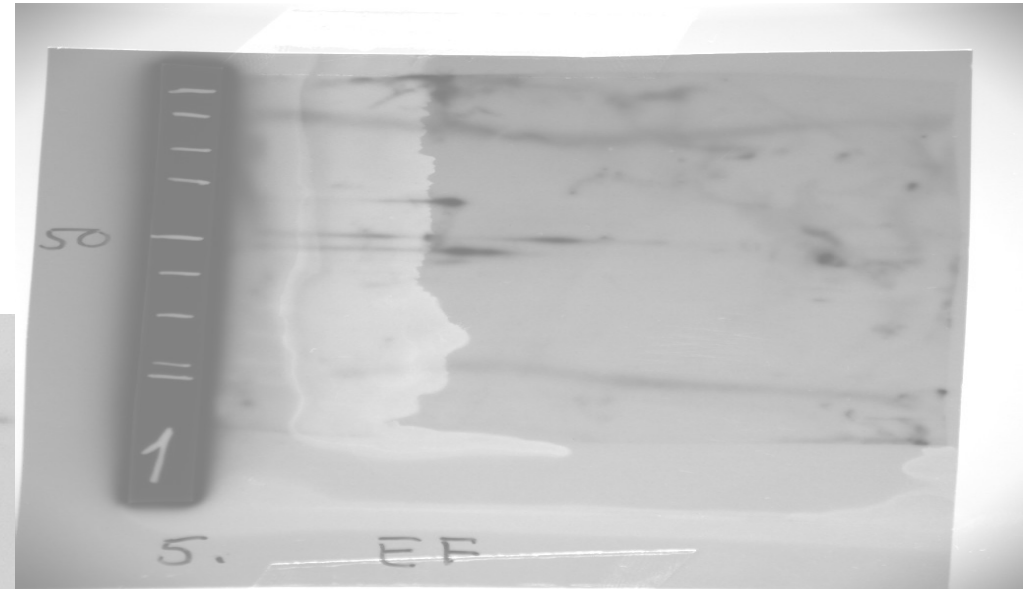
Step 2a: Background intensity

The background floor of a 2DE image refers to the brightness of empty gel areas. Different capture techniques produce different background floors. Background signal can be either added to all pixel values (additive background), or it can accumulate with a decaying signal (multiplicative background). As previously observed [44], most cameras introduce a mixture of additive and multiplicative backgrounds. Removal of additive noise can be done through subtracting the mean ($A_z'' := A_z' - \overline{A_z'}$) or median value ($A_z'' := A_z' - \text{median}(A_z')$). Removal of multiplicative noise can be done through $A_z'' := \frac{A_z'}{A_z'} - 1$. We would emphasize that whatever normalization scheme is used in this step, it should be performed on an individual gel basis.

Background Differences



Background Differences



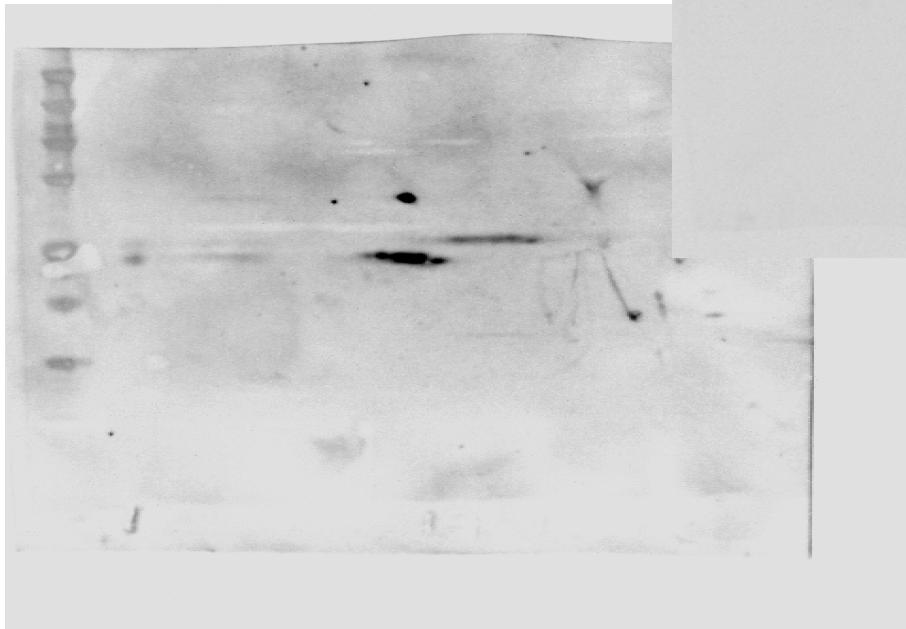
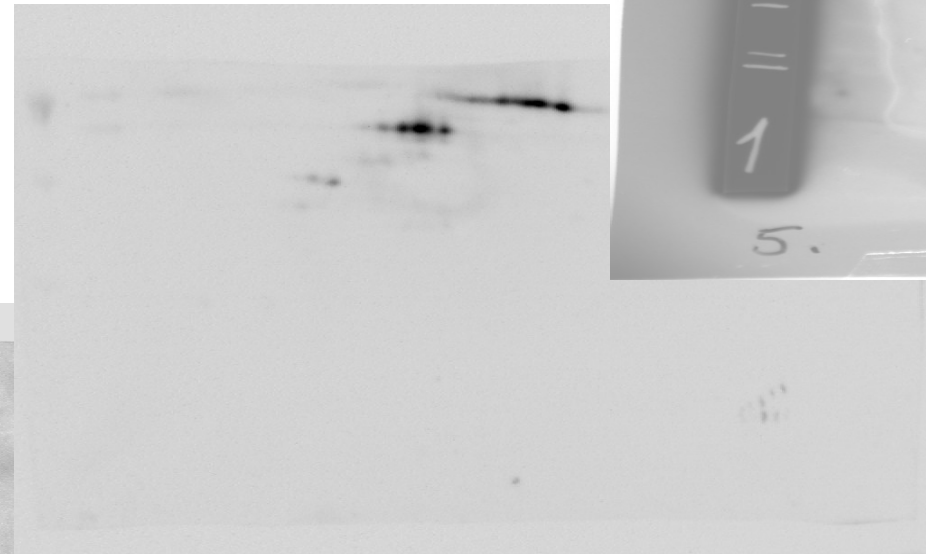
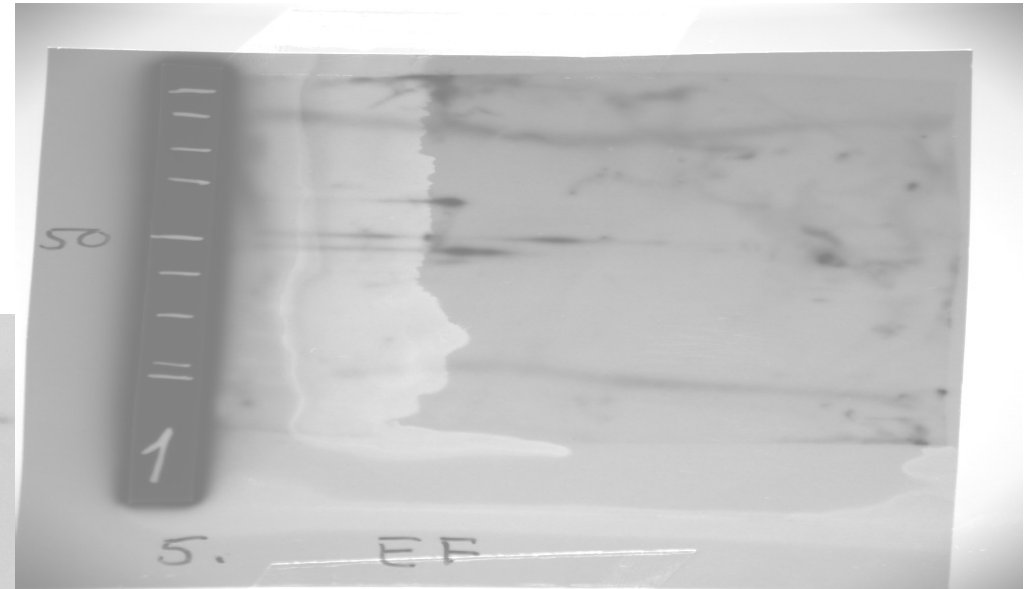


Step 2b: Intensity Normalization

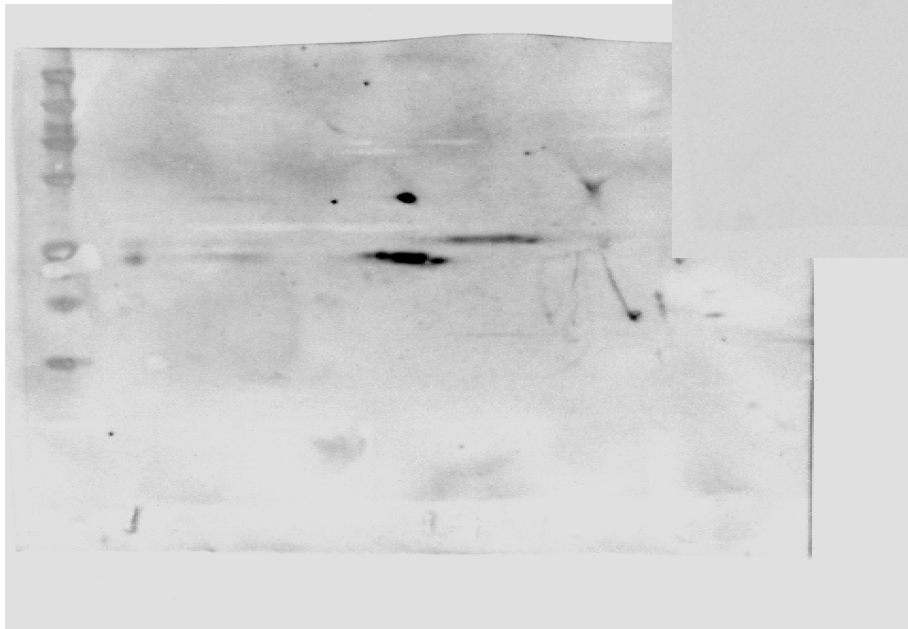
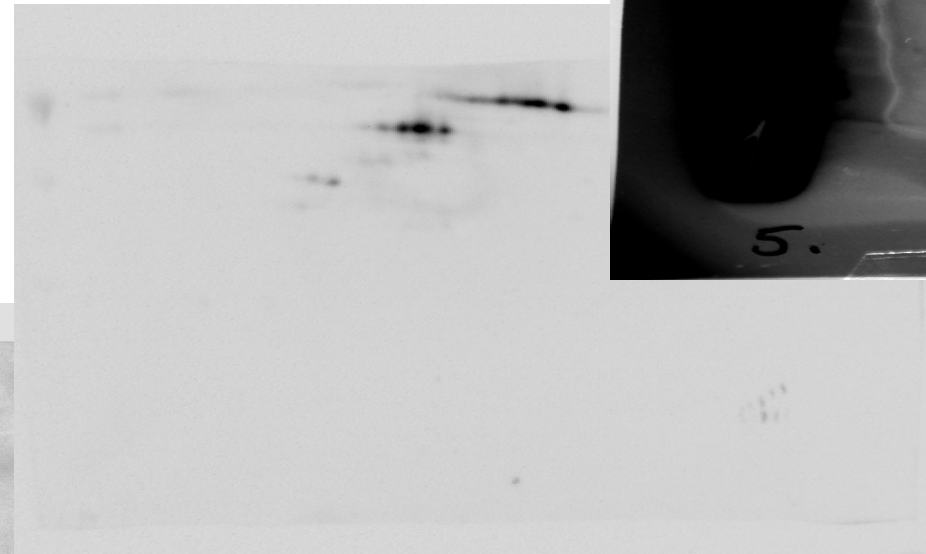
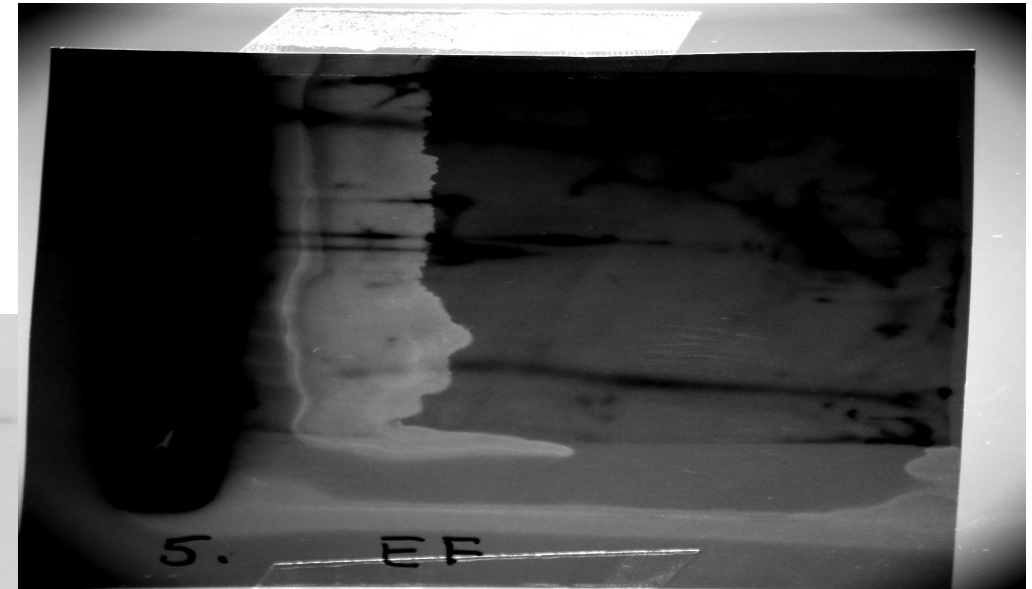
Step 2b: Scaling of gel intensity

After removal of the background floor, the dynamic range of the image is normalized through scaling of gel intensities. The presence of a calibration spot eases this process. If A' is the non-relative image and (x, y) is the calibration spot position, then the image $A'' := \frac{A'}{A'_{x,y}}$ defines the normalized image. Without calibration spot the total energy content (sum of all intensities or RMS value) forms a very reasonable scaling means: $A''_z = \frac{A'_z}{RMS(A'_z)}$

Contrast



Contrast





Step 3: Correlation

Step 3: Correlation image

The correlation

image is composed of pixels, each testing one position on the gel. The result of each test is a number between -1.0 (anti-correlation) and 1.0 (correlation), which, after appropriate scaling, defines the pixel color in the correlation image. The two vectors participating in the test are $A''_{x,y}$ and B . The first vector contains the gel expression levels at position (x, y) . Given 89 gel images, $A''_{x,y}$ will contain 89 different expression values; one for each gel. The second vector B contains 89 external values associated with every gel. Repeating this correlation test for every pixel results in the correlation image C (Eq. 1)

Step 3: Correlation

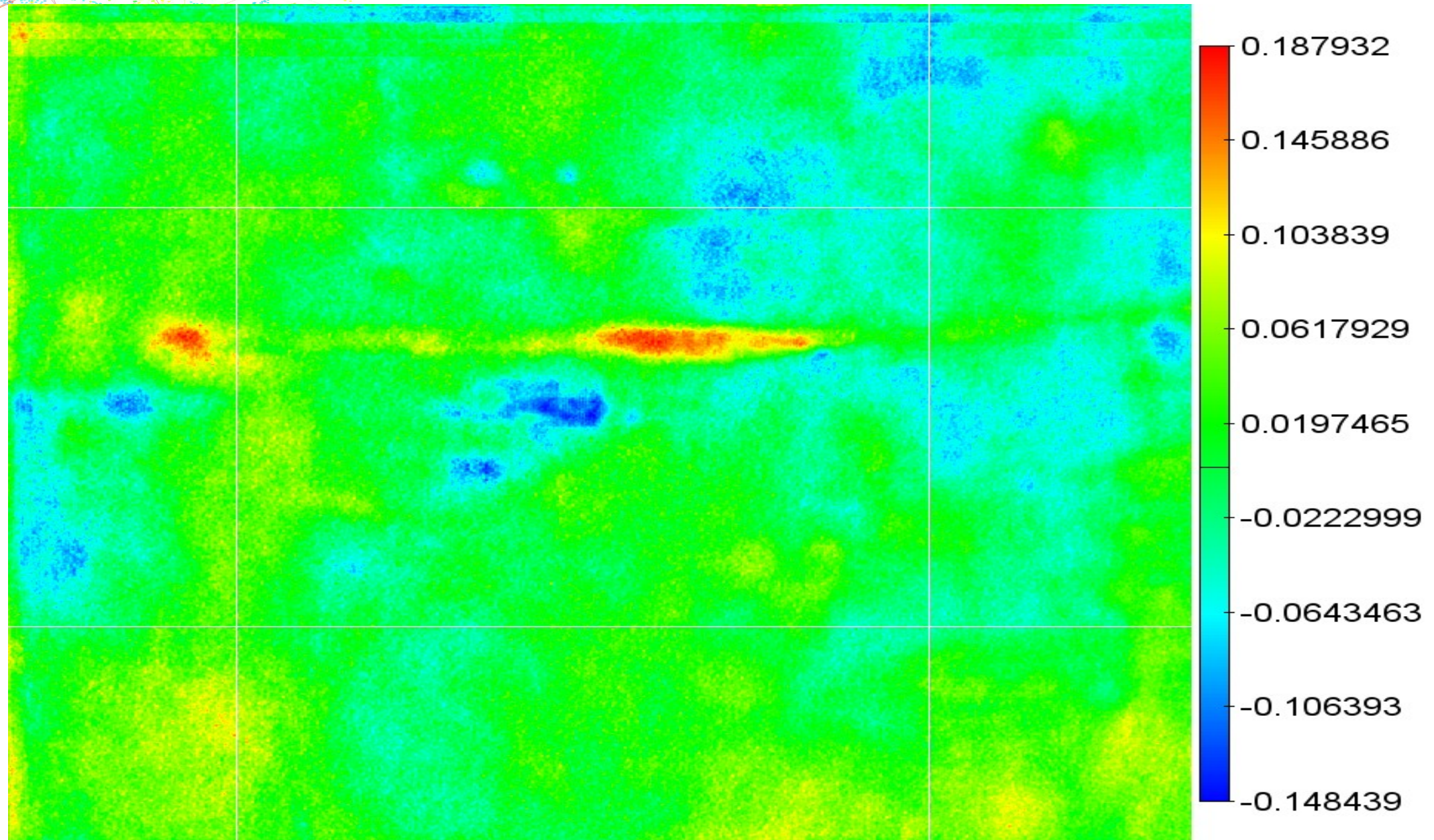
$$C_{x,y} = \rho(A''_{x,y}, T) \quad (1)$$

The correlation image can be visualized using different color schemes. In Fig. 1 green indicates positive correlations and brown negative correlations.

The preferred correlation is the robust Spearman rank order correlation (ρ -correlation) [27].

This non-parametric test allows us to ignore the specific distributions of gel intensity levels and external parameters. ρ -correlation requires a ranking of the two participating vectors and then relies on a standard linear Pearson correlation. The ranking process will replace every value in the input vector by its specific rank. When ties occur (the same value occurring more than once) their rank will by convention be the mean of their ranks as if they all would have had a slightly different value.

P53 Biosignature vs Age





Step 4: Masking

Step 4: Masking

Correlation does not necessarily imply a causal, significant, or useful relationship. To filter out some possibly useless relations, a number of masks limit the visible correlations. The first mask removes correlations that might be occurring by coincidence: some data sets easily correlate with any other data set (significance). The second mask removes correlations that offer little useful information (E.g: a data set containing all zero's).



Step 4a: Significance

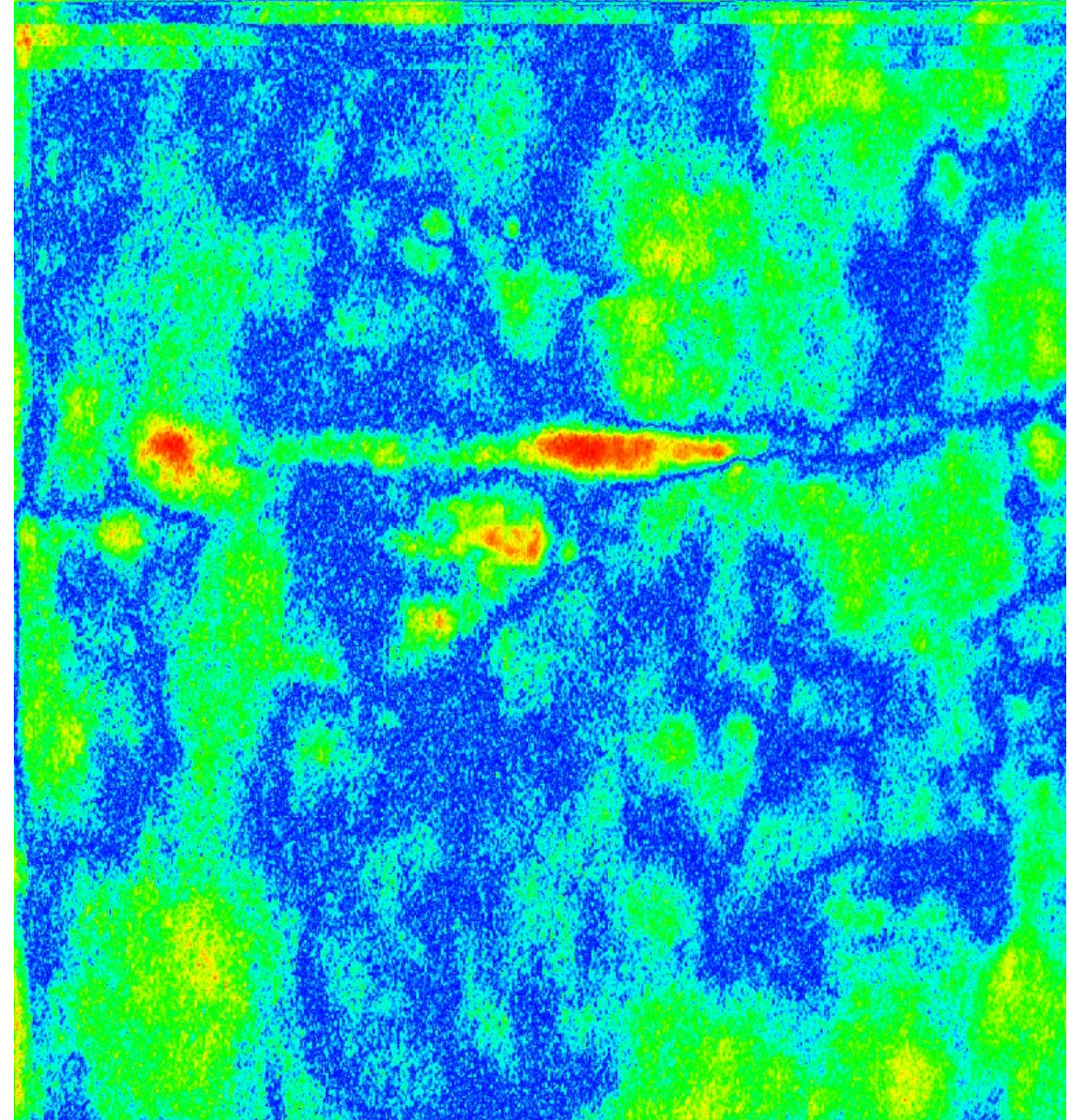
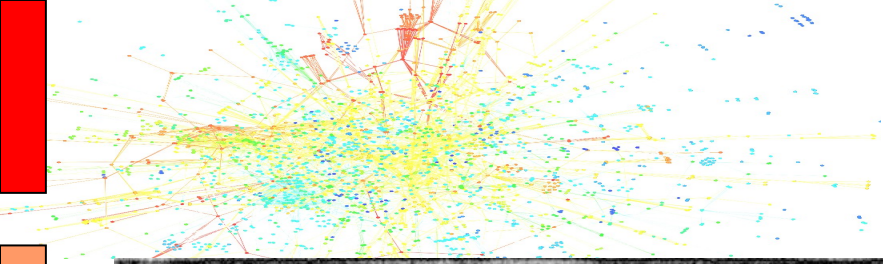
Step 4a: Significance

To remove correlations that have a high probability of occurring, the significance test typically associated with the Spearman correlation test was used. In this context, it is defined as

$$S_{x,y} = 1 - C_{x,y} \sqrt{\frac{n-2}{1-C_{x,y}^2}} \quad (2)$$

If this number is close to 1 then there exists a low probability that some random data would happen to correlate with the given result set. Likewise, if this number is 0 then there exists a high probability that the correlation is coincidental.

Significance Mask





Step 4b: Variance

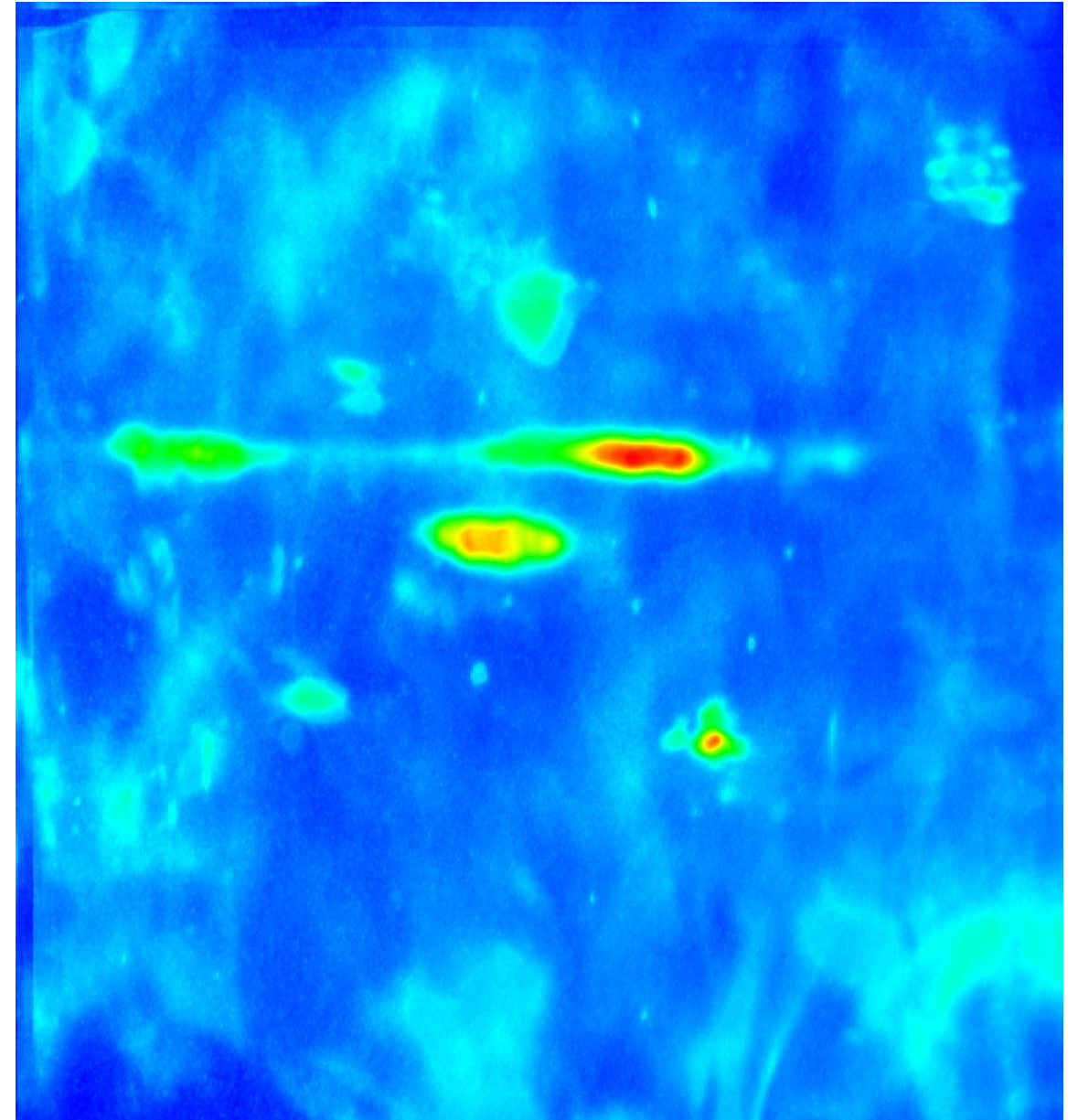
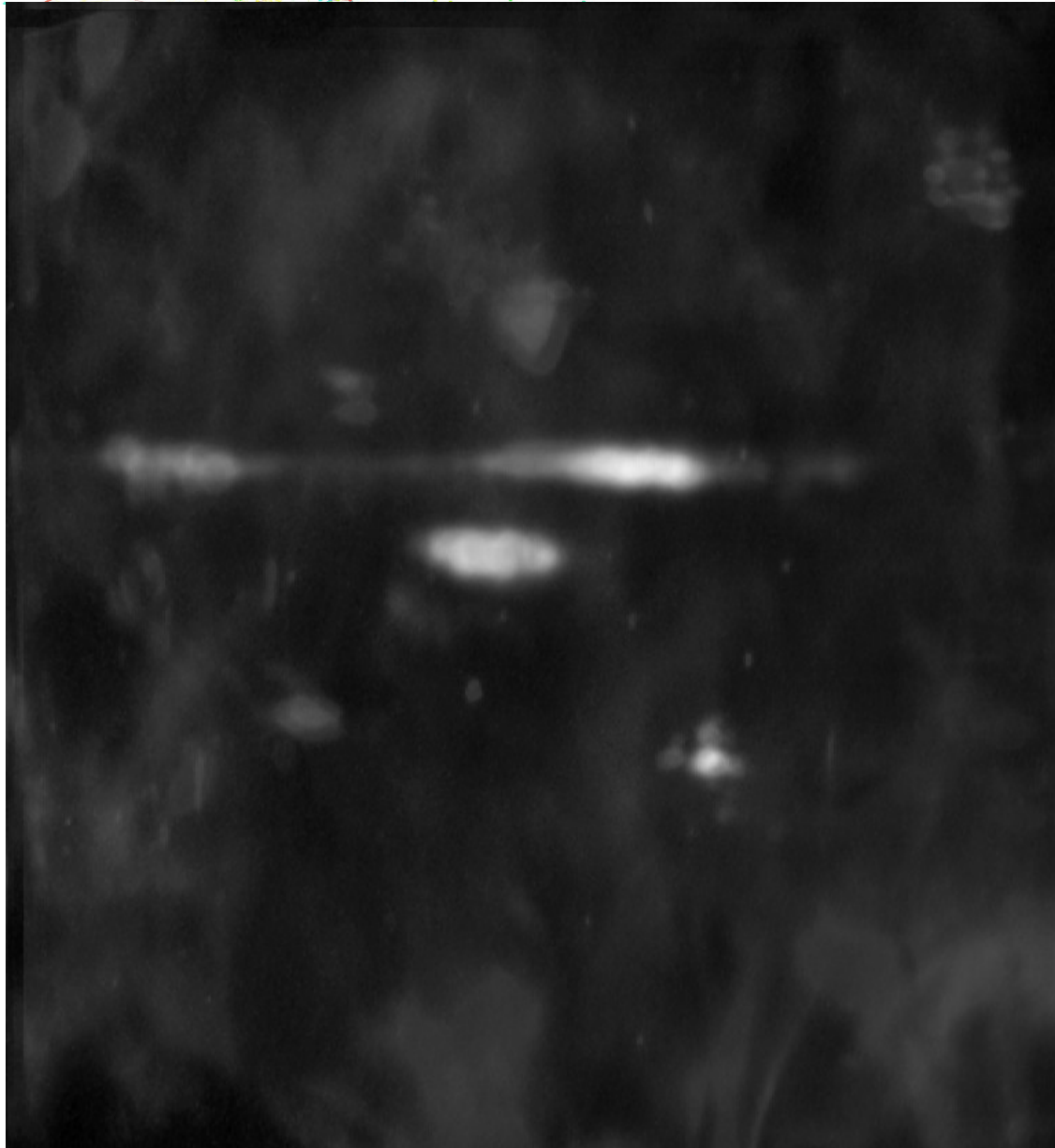
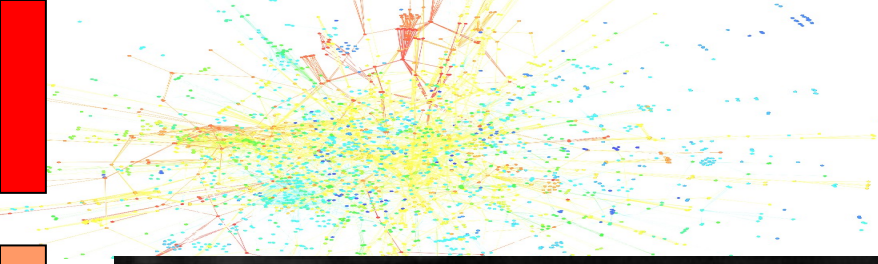
Step 4b: Variance

The second mask avoids strong and significant correlations that have a low biological significance because the gel intensities do not change enough. It relies on the standard deviation [54] measured on the relative, non-ranked, gel intensities

$$D_{x,y} = \frac{\sqrt{\sum_{z=0}^{n-1} \left(\frac{A''_{x,y,z}}{A''_{x,y,*}} - 1 \right)^2}}{N} \quad (3)$$

The standard variance (or RMS) of the mean divided gels will have a large value where there is a varying gel expression. At places where the gel expression is constant this value will be zero.

Variance Mask





Step 4c: Overall Mask

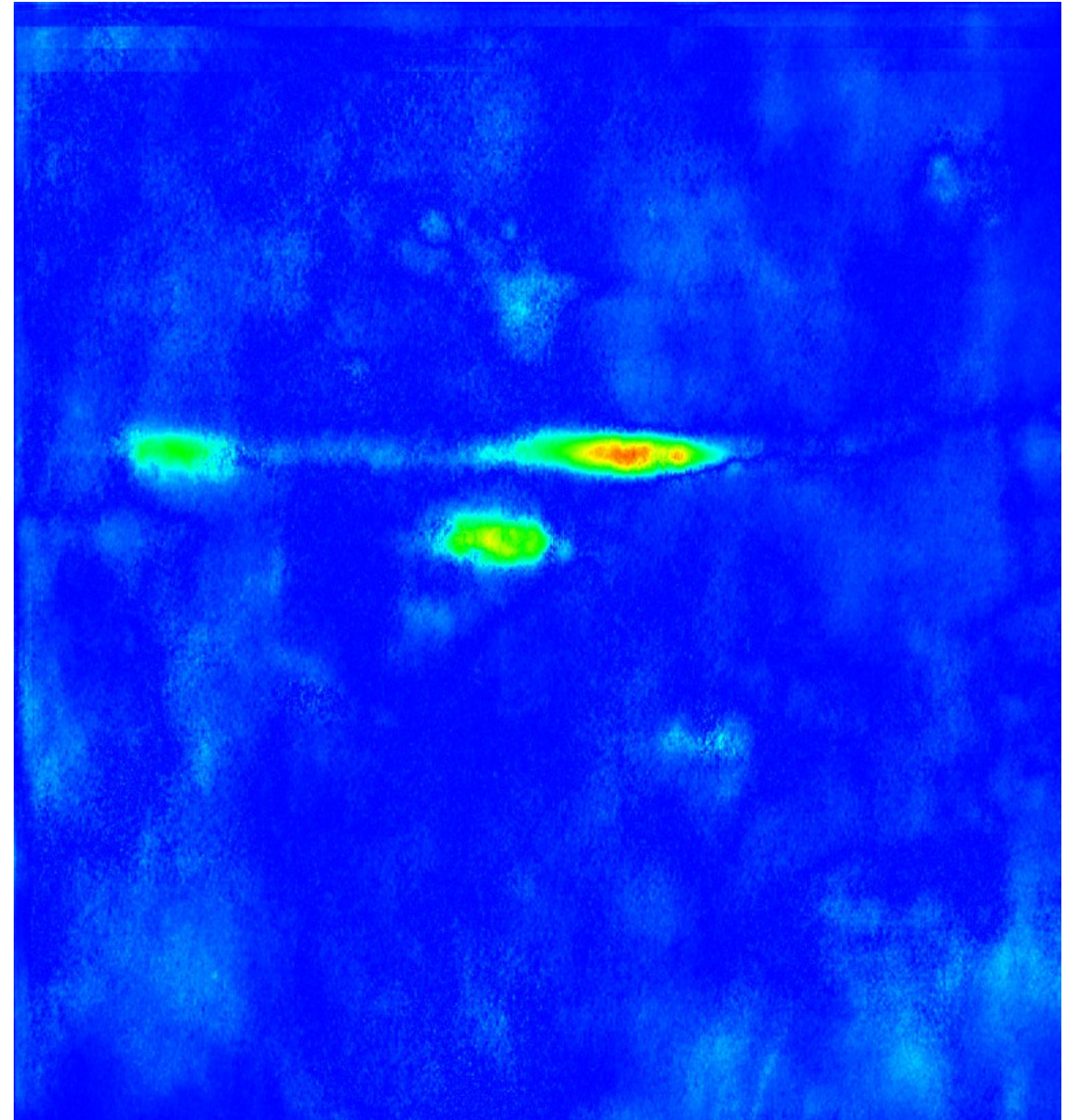
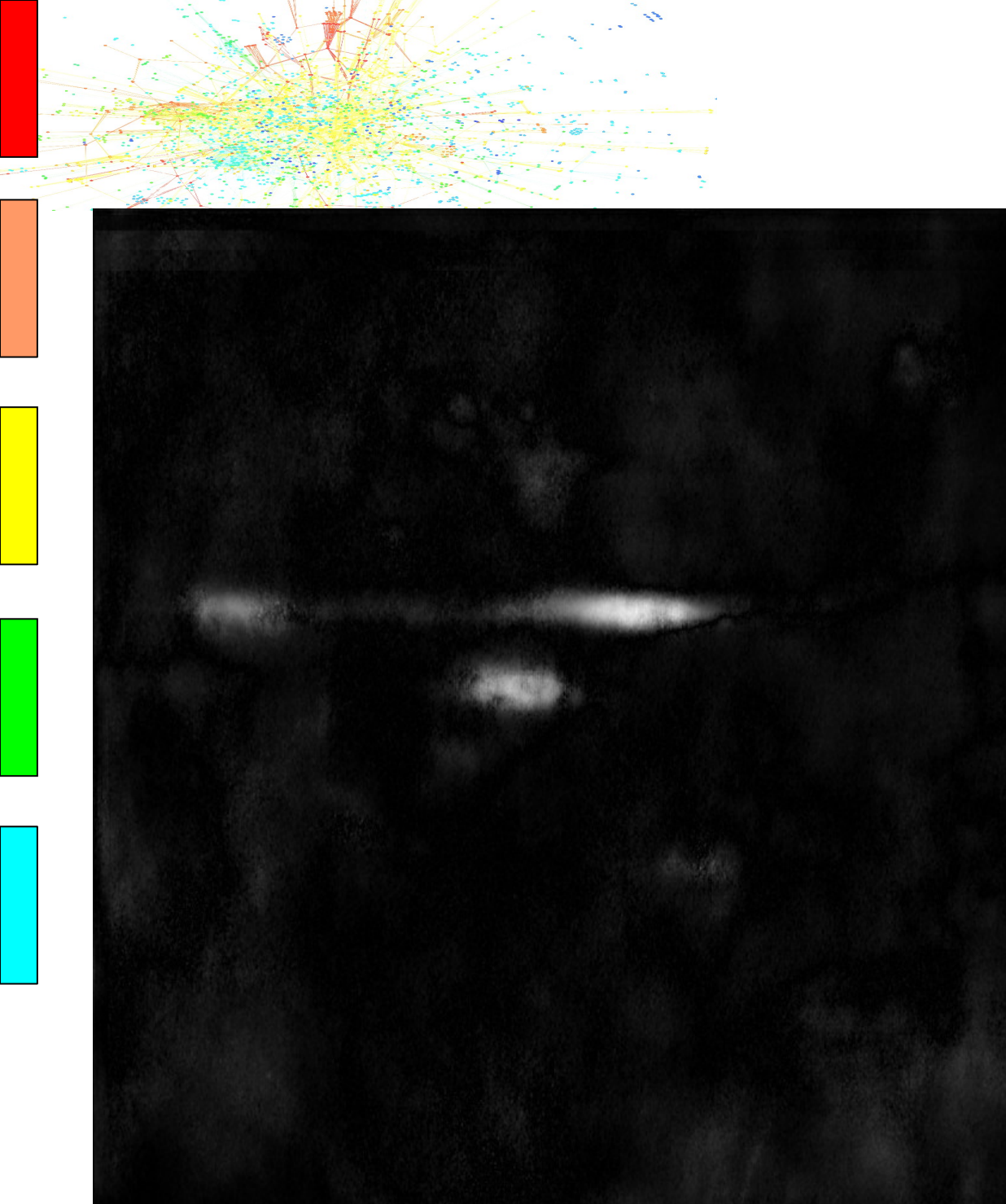
Step 4c: The masked correlation image

Multiplying the standard deviation mask (Eq. 3) with the significance mask (Eq. 2) gives a new mask that can be superimposed over the correlation image (Eq. 1).

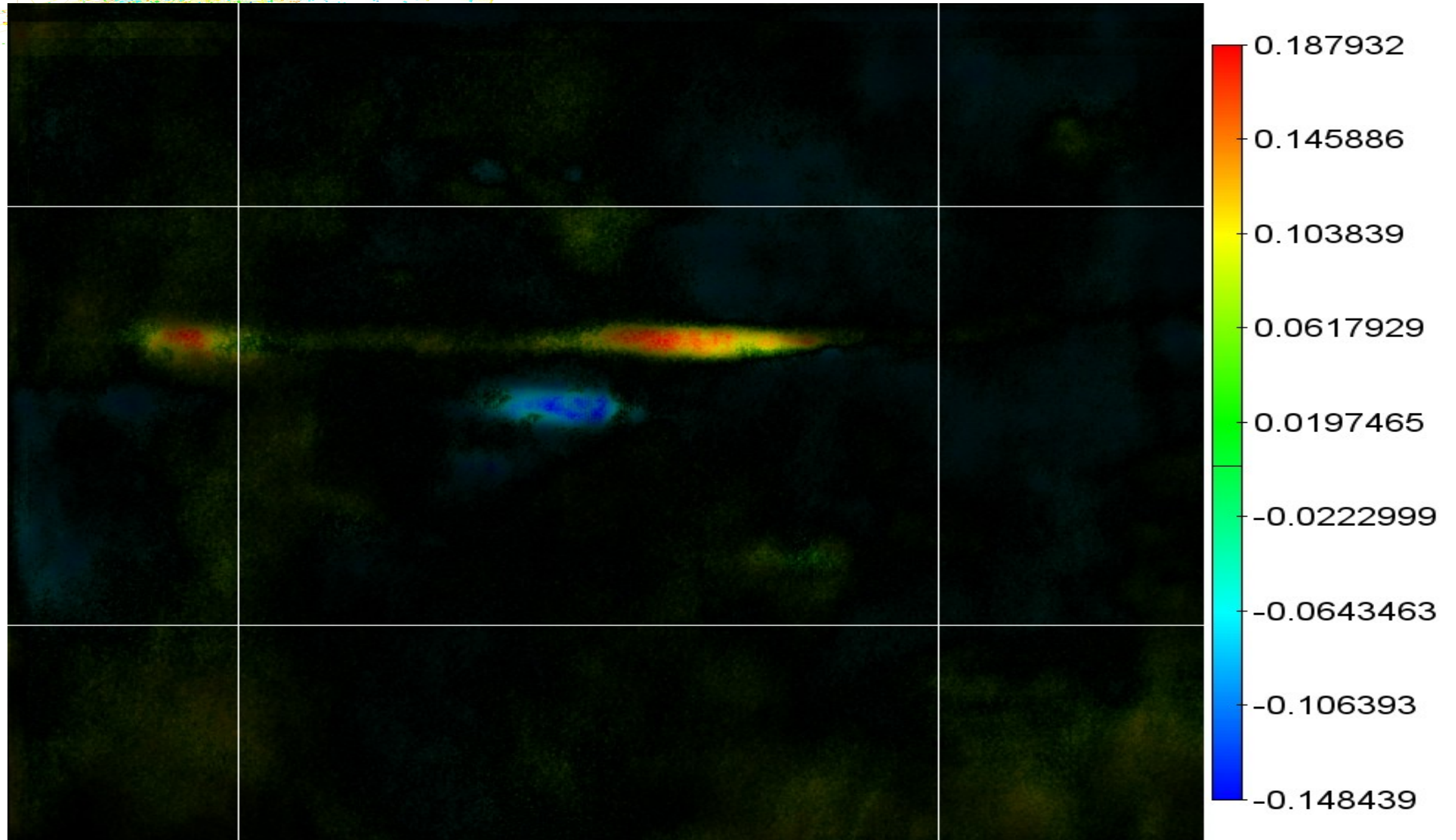
$$R = C \times S \times D$$

The pixel values of R no longer relates to the correct correlation measure. Therefore, R forms an indicator, showing position of possible interest.

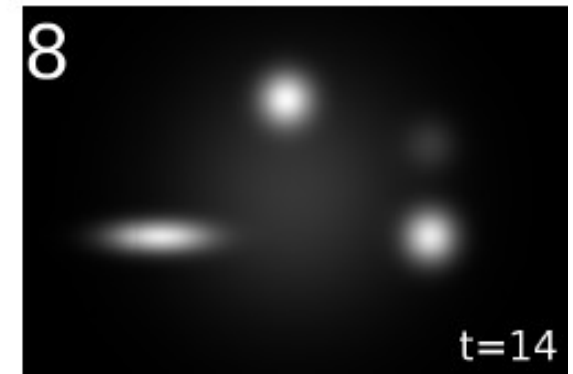
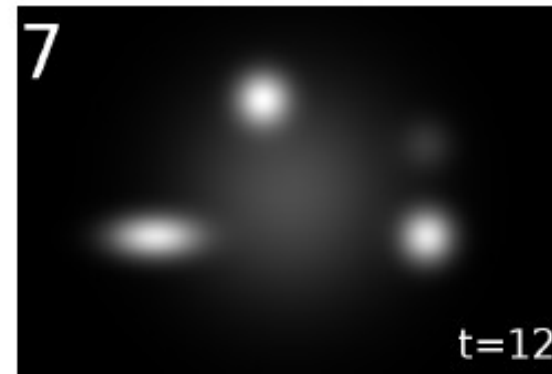
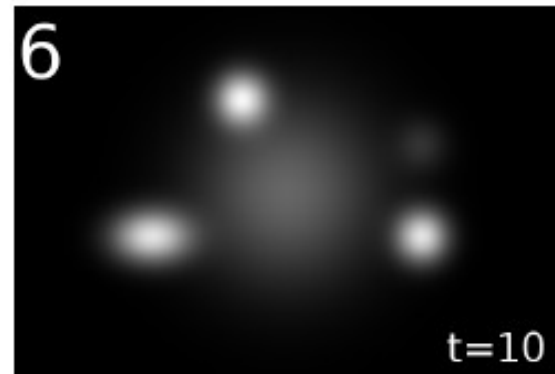
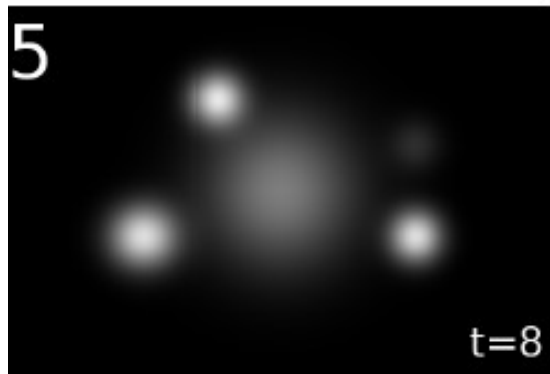
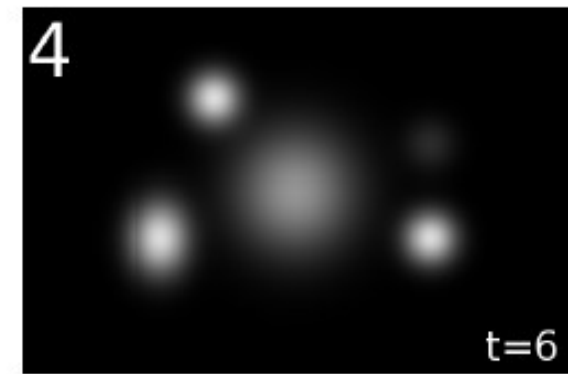
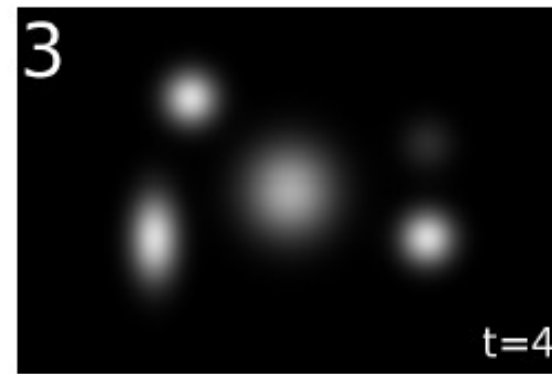
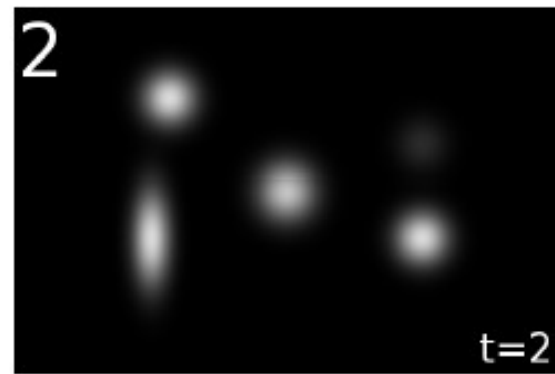
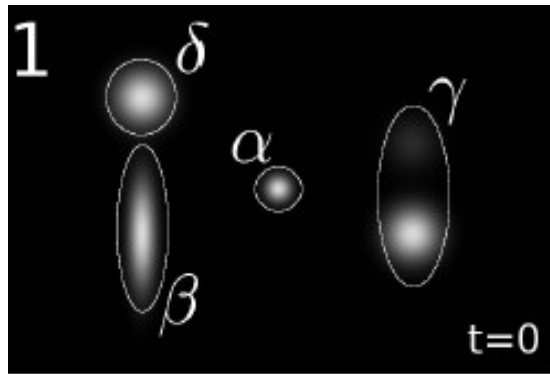
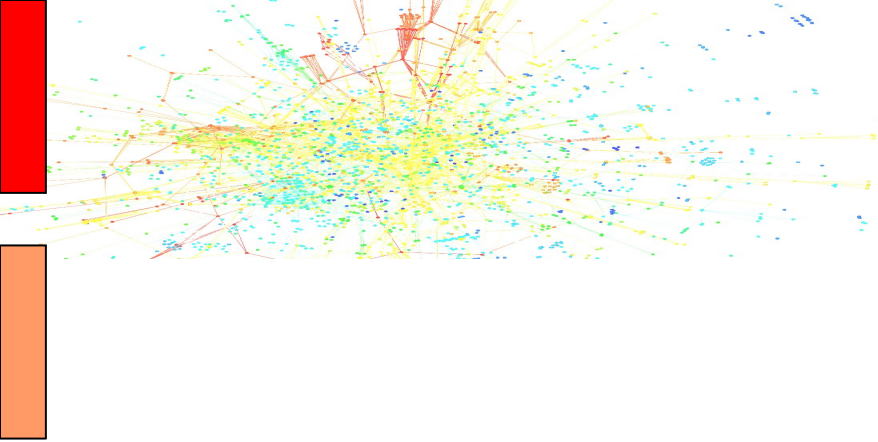
Overall Mask



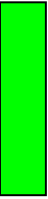
P53 Biosignature vs Age



Simulated Gel Stack



Correlation Images

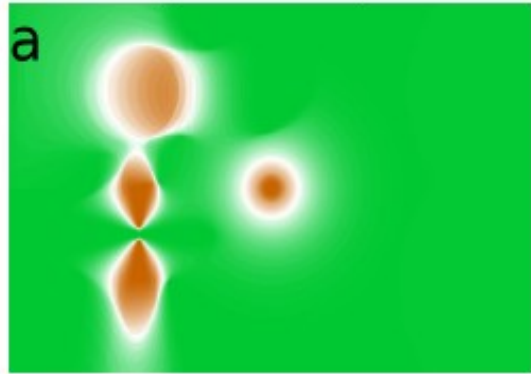


B

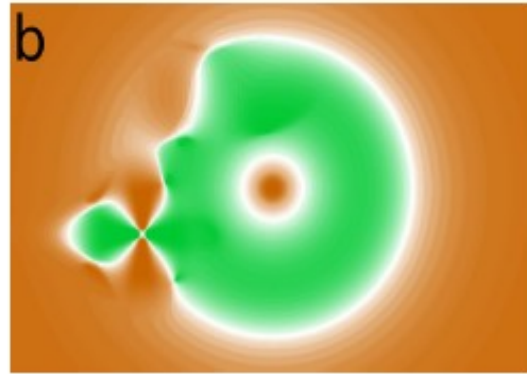
Correlation

Masked
Correlation

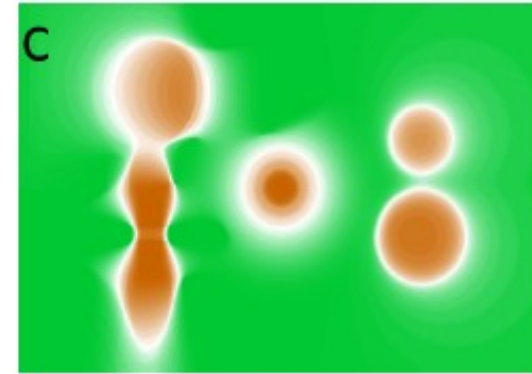
Background: As is



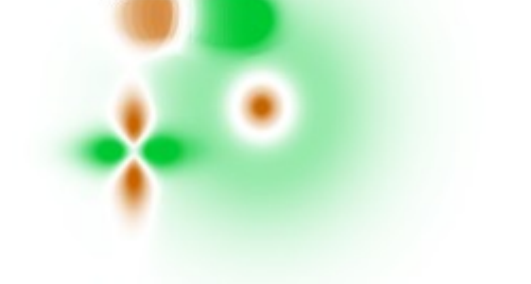
Subtracted



Divided



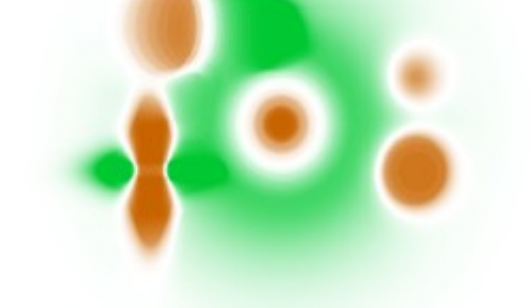
a' δ -smear



b' δ -smear



c' δ -smear



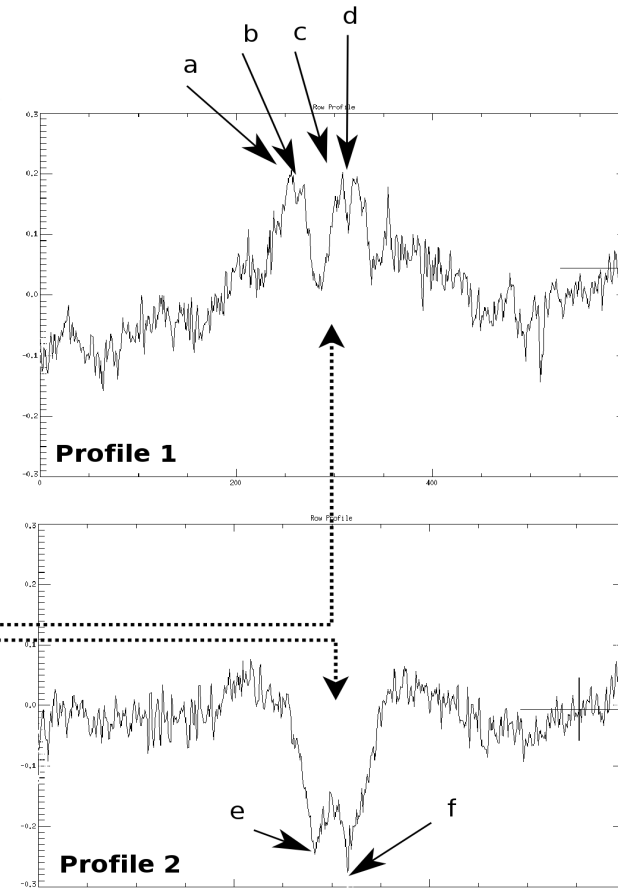
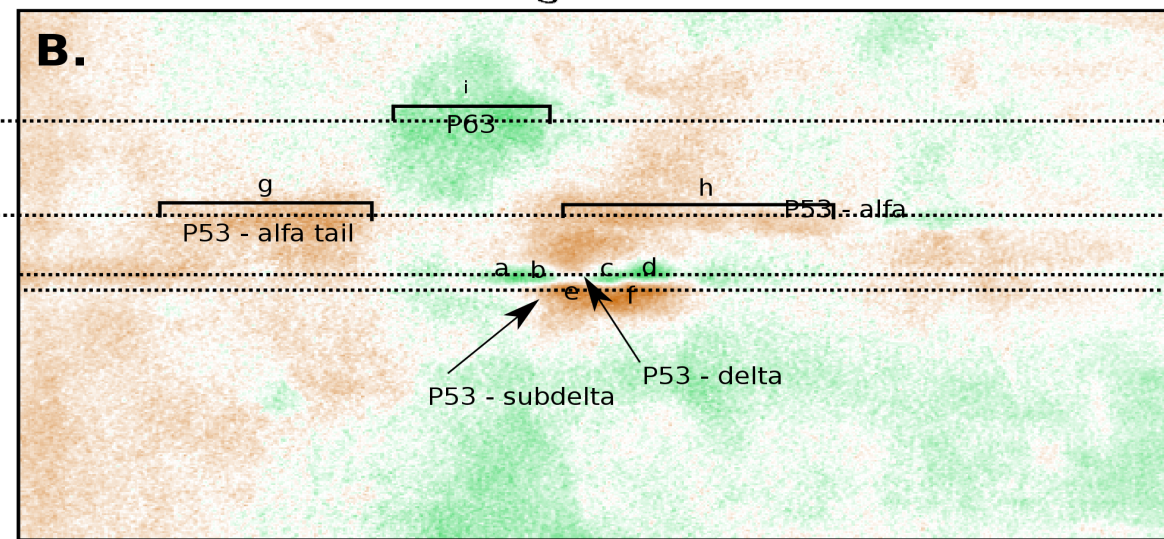
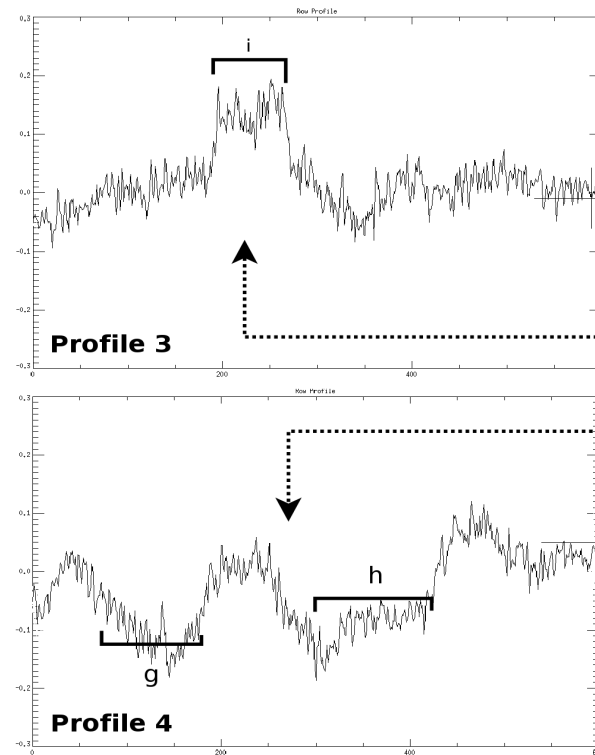
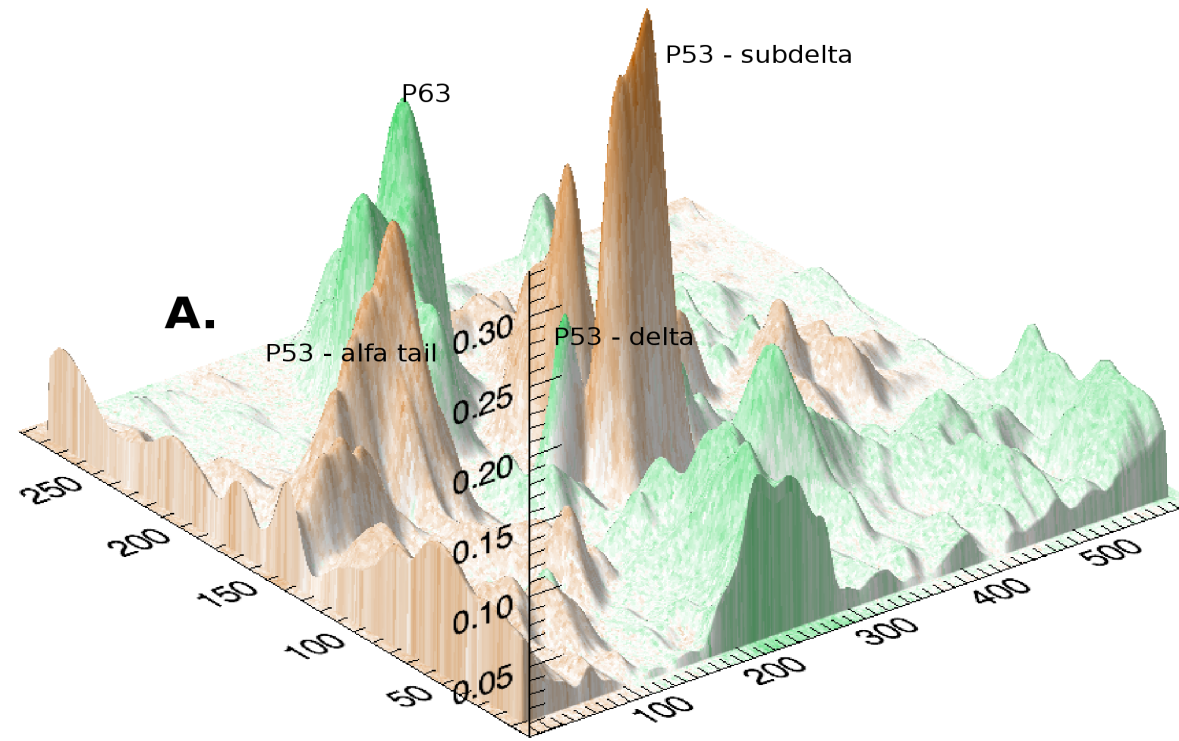
Maximal Correlation
No Correlation
Minimal Correlation

Step 5: 3D Visualization

Positive Correlation

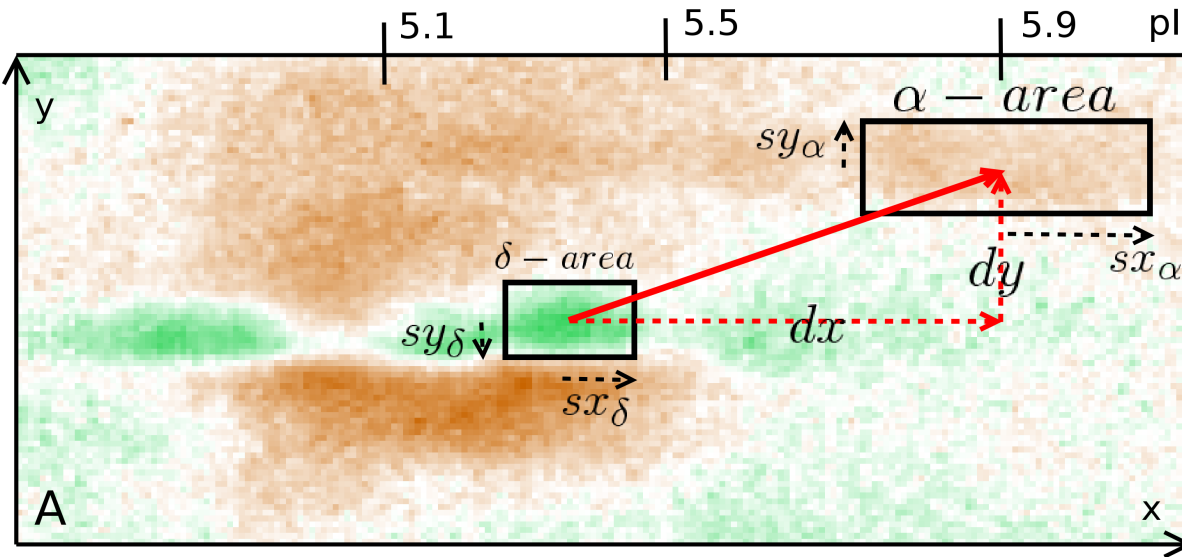


Negative Correlation

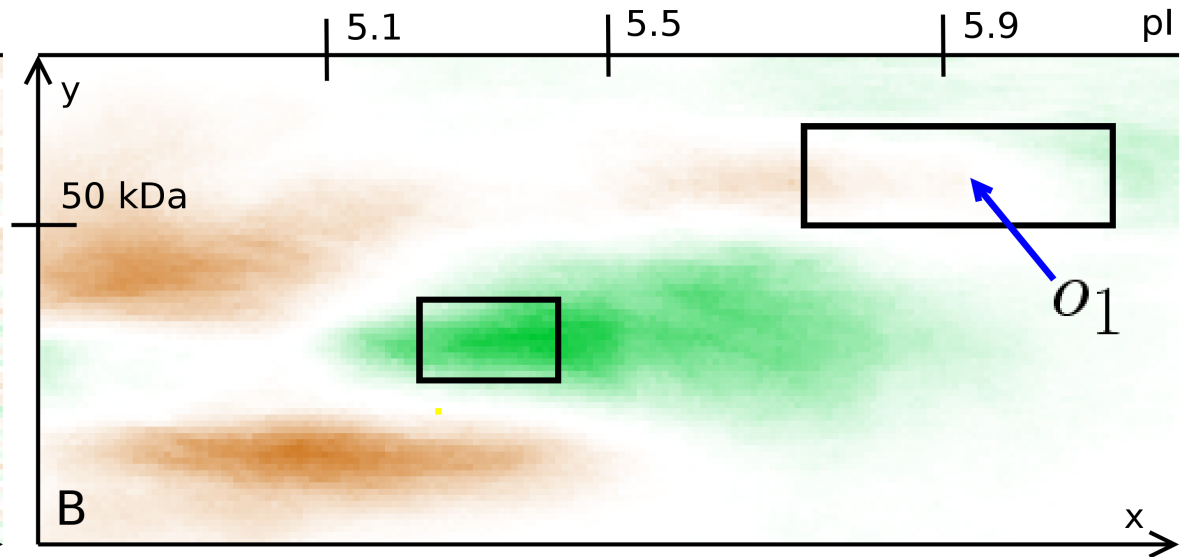


Intra Gel Relation Correlations

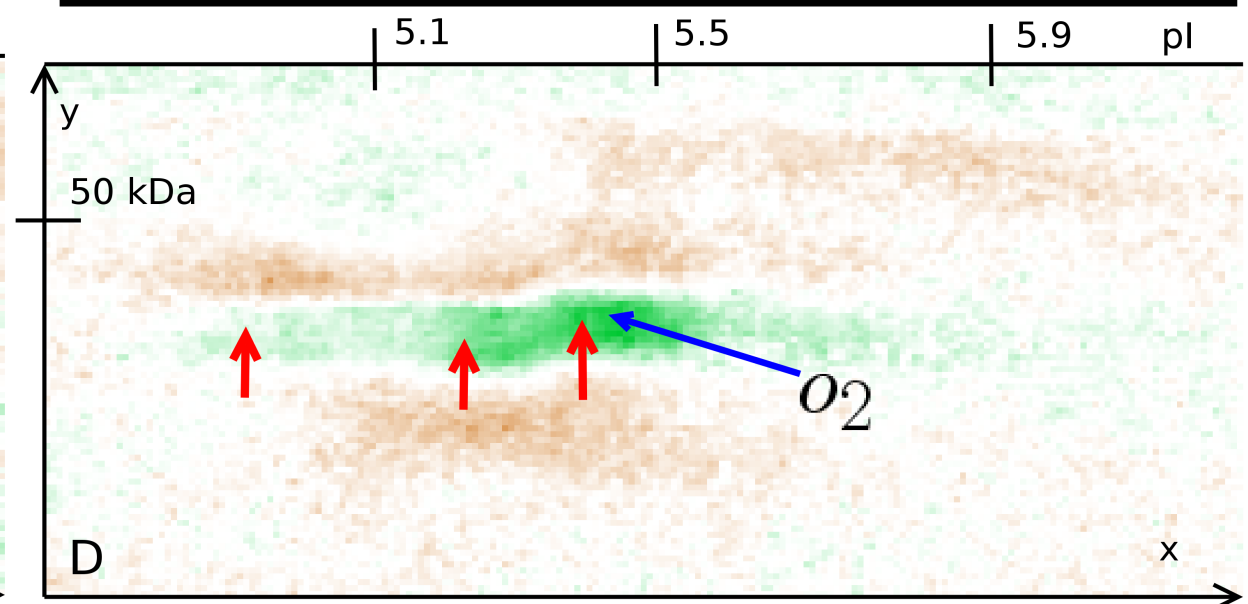
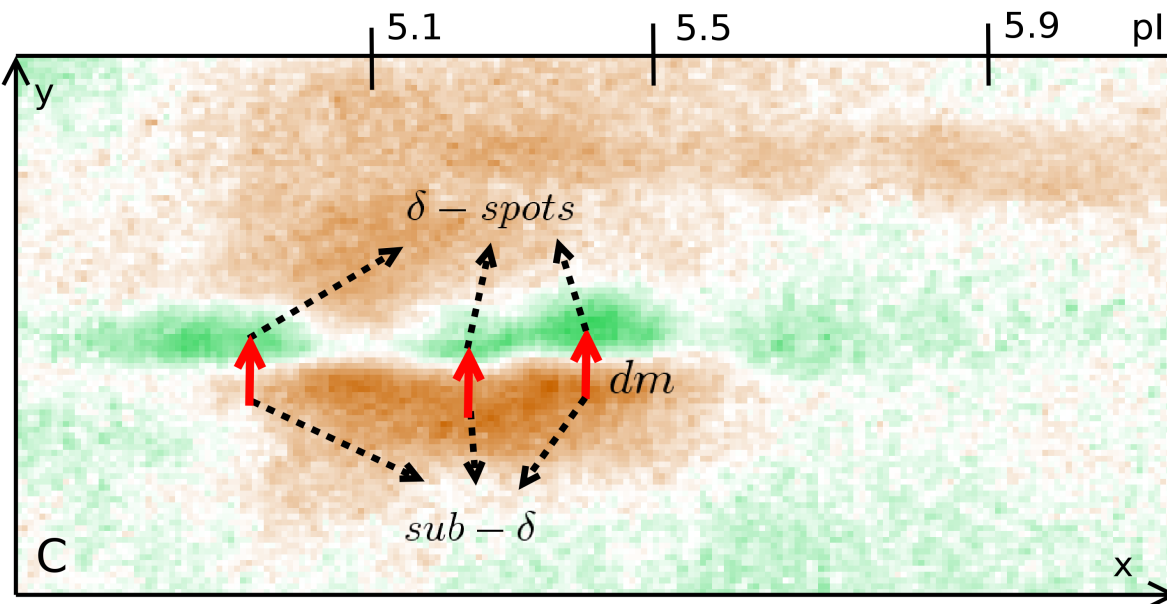
AML Differentiation vs Gel Intensity



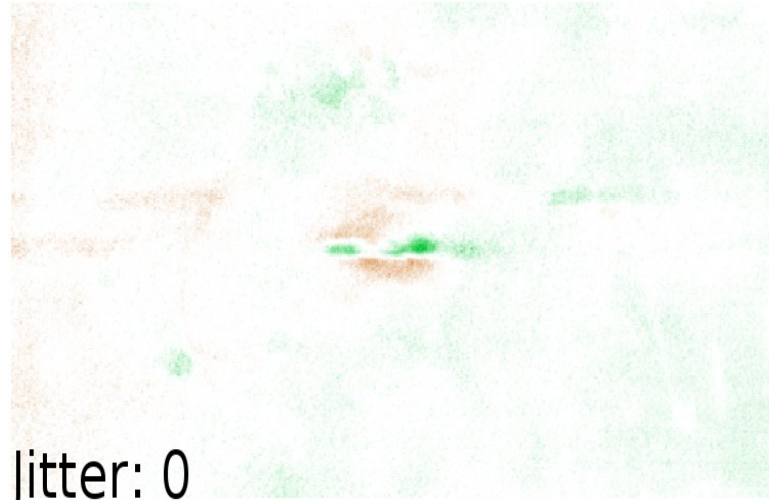
AML Diff. vs Difference between Alfa & Delta-Area Intensity



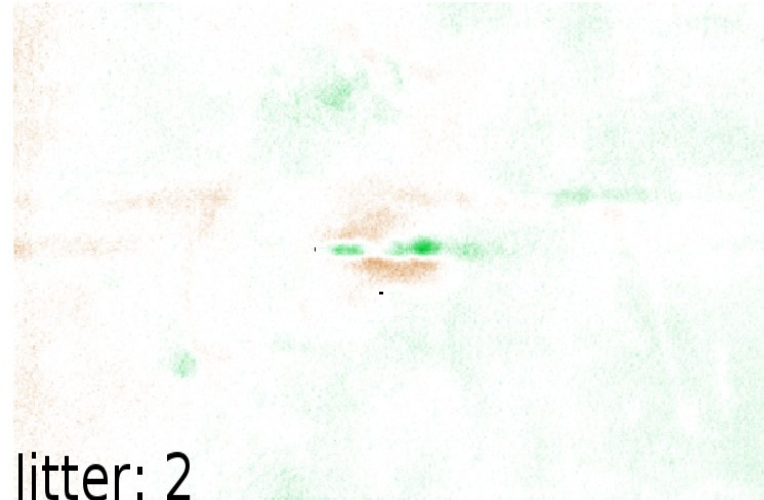
AML Differentiation vs Perceived Mass Difference



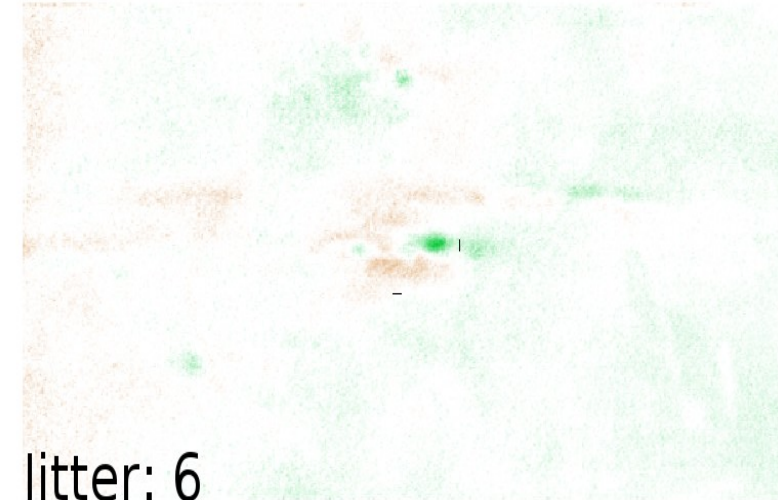
Alignment Jitter



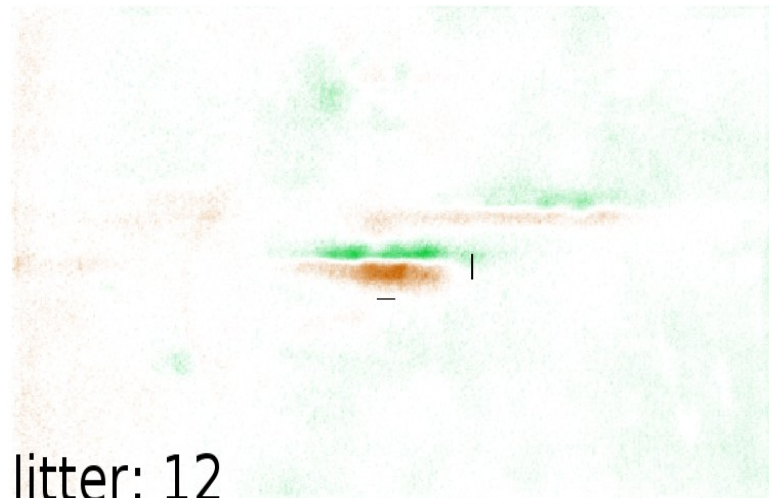
Jitter: 0



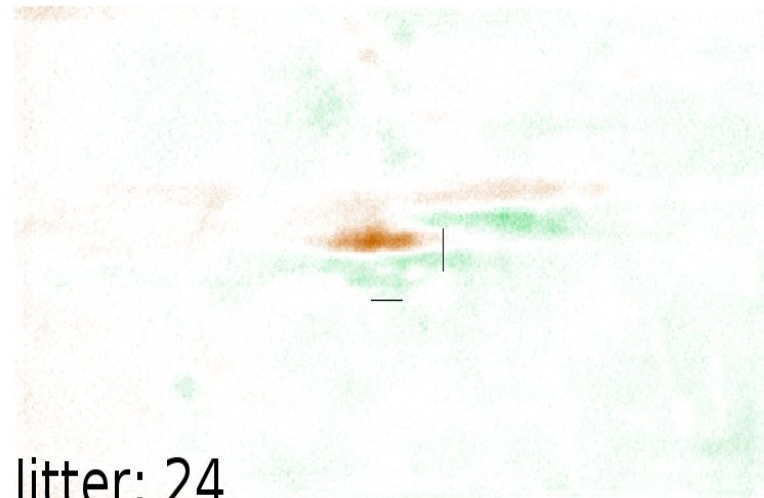
Jitter: 2



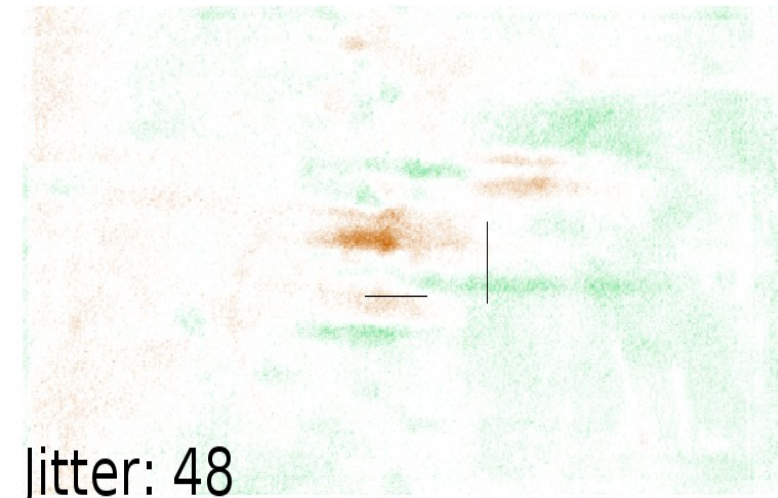
Jitter: 6



Jitter: 12

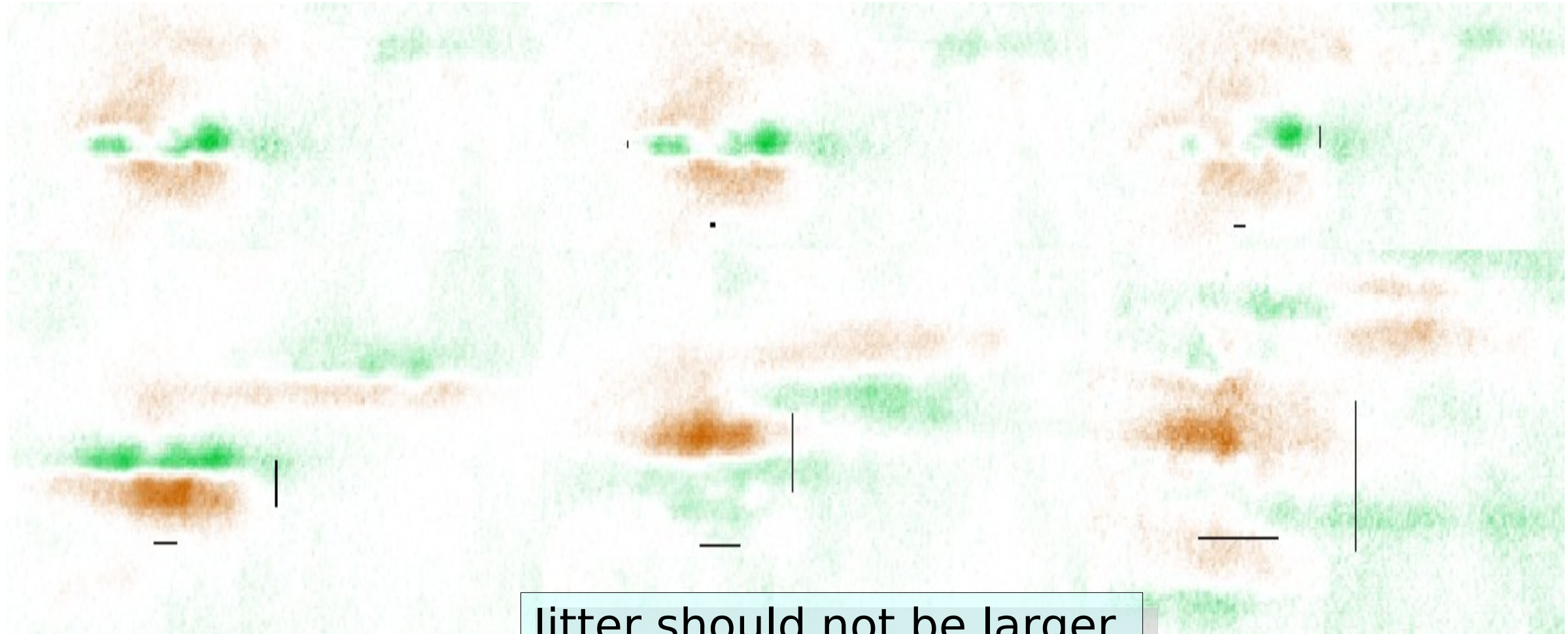
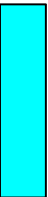
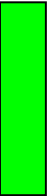


Jitter: 24



Jitter: 48

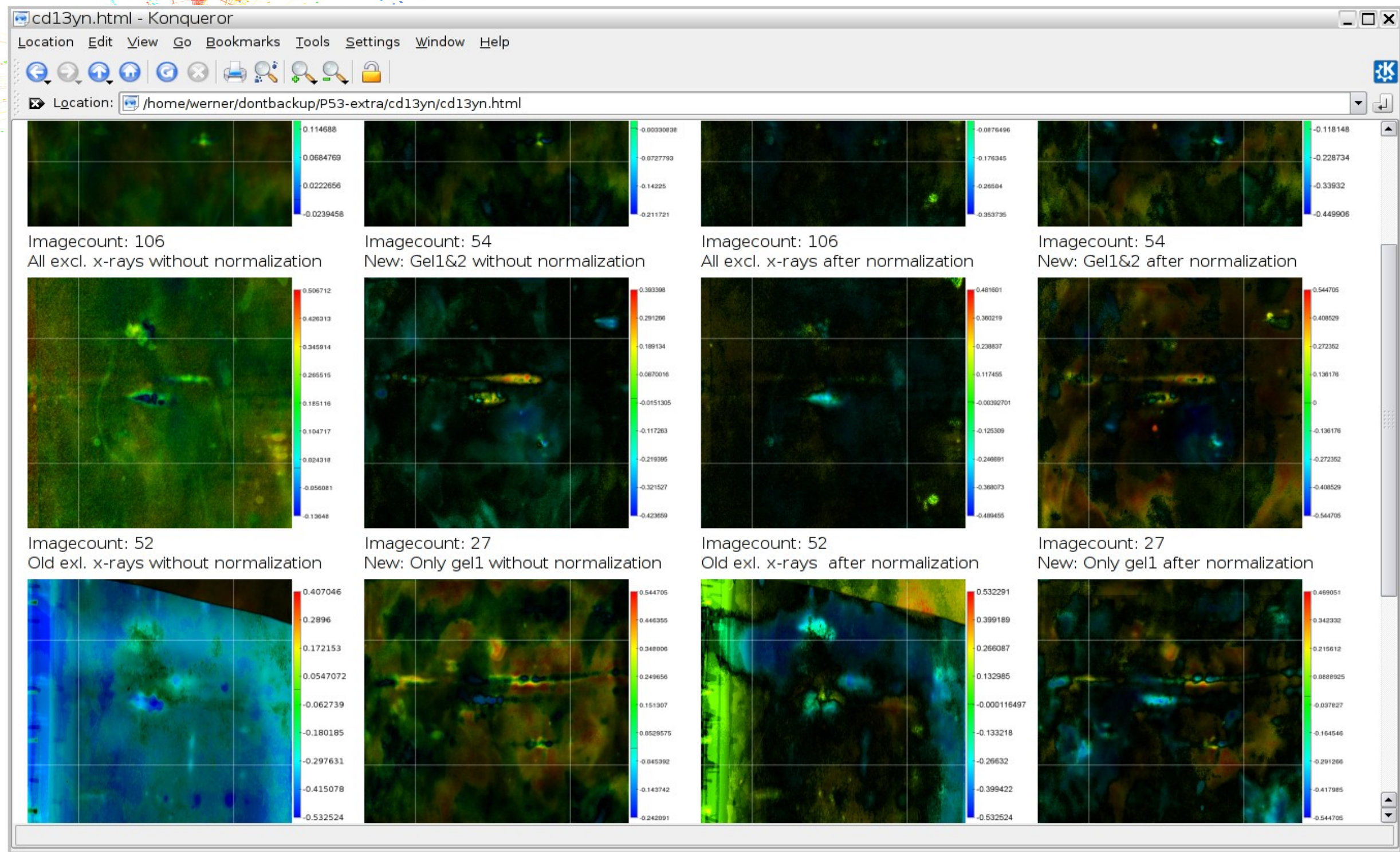
Alignment Jitter



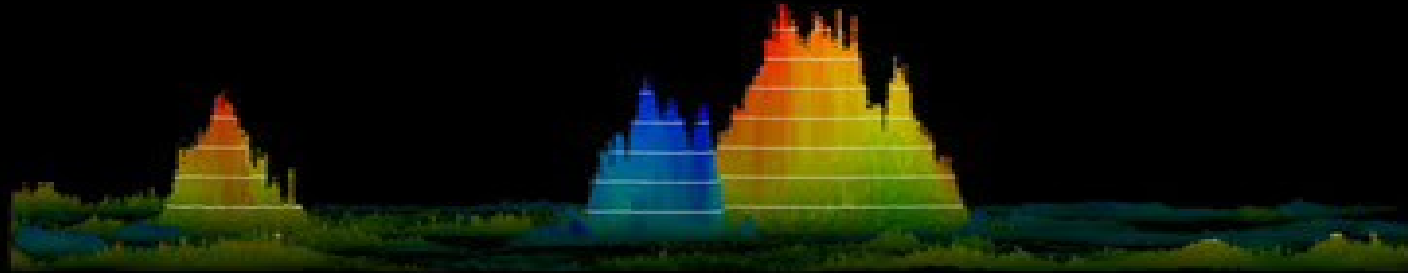
Jitter should not be larger
than the mean spot size

Resource Usage

- 132 Parameters, 13 correlation sets, 128 images
- Creating the fine-tuned overlay alignment: 72h
- Computing all the correlations: 85.55h, which produced 5.8 Gb of raw data.
- Rendering of the movies: 5 hours per movie, with 1416 images: 7080h -> 93 Gb



Step 5: 3D Visualization





Part 2. Systems Biology

Science is built up with facts, as a house is with stones. But a collection of facts is no more science than a heap of stones is a house - Henri Poincaré

Dr. Werner Van Belle

Medical Genetics

University Hospital Northern Norway

e-mail: werner@sigtrans.org



Biological Networks in computers

- Interpretation
 - Visualization can help guide the interpretation process
 - Clustering can aggregate seemingly incoherent measurements
- Model building
 - infer general properties that are supported by experiments and explain the results coherently
- Prediction
 - how will the network react in hypothetical situations (E.g: suppose we would knock out this gene)



Biological Networks in computers

- Coupled differential equations
- Boolean networks
- Symbolic Approaches (KEGG).
- Continuous networks
- Stochastic

Why not include protein interactions ?



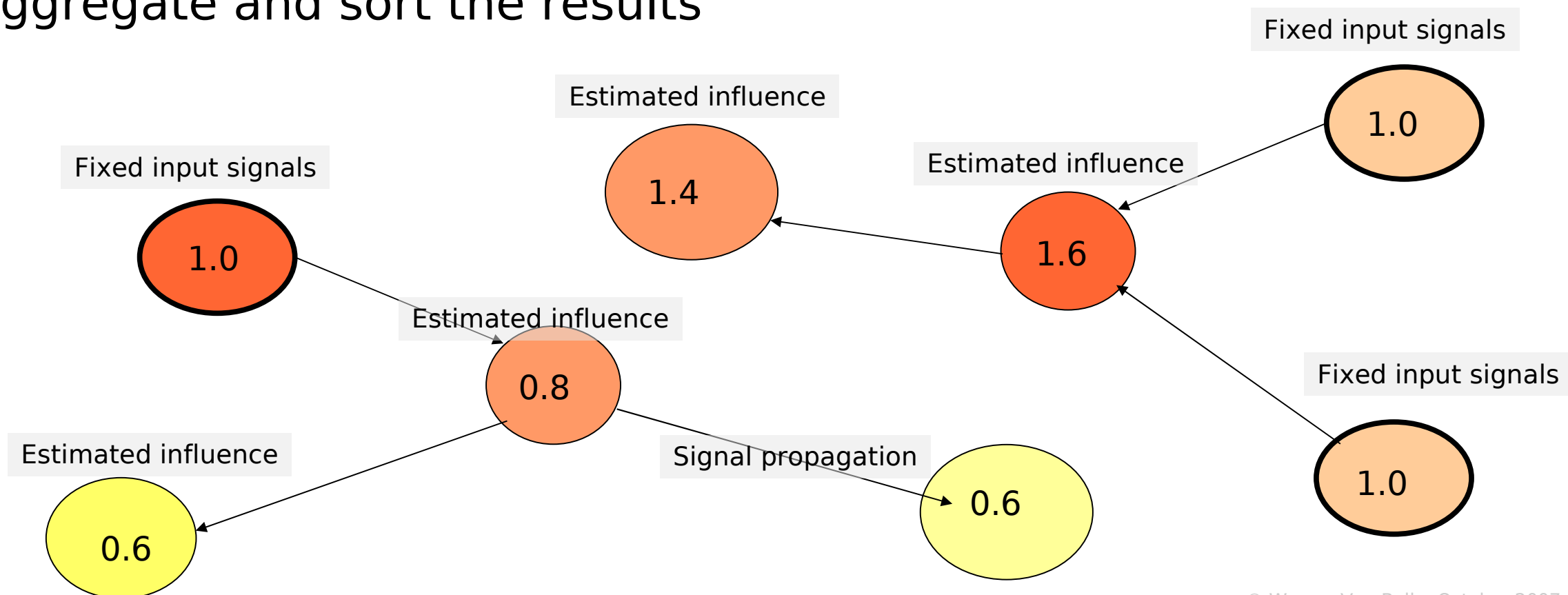
Influenced by/Influences

- MK5 leads to multiple changes in gene expression
- 27000 gene expressions measured
- Those that change will very likely influence other proteins

Which proteins are likely influenced by our measured up/down regulations ?

Influence Propagation

- Create the graph using a protein interaction map
- Initialize graph with micro array measurements
- Propagate the influence to the neighbors
- [Normalize the network]
- Repeat
- Aggregate and sort the results



Involved Proteins by Rank

PROTEIN CGI-126 (PROTEIN HSPC155)

RAD50-INTERACTING PROTEIN 1

RHO-RELATED BTB DOMAIN-CONTAINING PROTEIN 2 (DELETED IN BREAST CANCER 2 GENE PROTEIN

NADH-UBIQUINONE OXIDOREDUCTASE 18 KDA SUBUNIT, MITOCHONDRIAL PRECURSOR (EC 1.6.5.3)

CHROMATIN ACCESSIBILITY COMPLEX PROTEIN 1 (CHRAC-1) (CHRAC-15) (HUCHRAC15) (DNA POLYM

ADIPONECTIN RECEPTOR 2

ODD-SKIPPED RELATED 1; ODZ (ODD OZ/TEN-M) RELATED 1.

DNA POLYMERASE EPSILON P12 SUBUNIT (DNA POLYMERASE EPSILON SUBUNIT 4)

PROTEIN X 0004

XPA BINDING PROTEIN 1; MBD2 INTERACTOR PROTEIN; PUTATIVE ATP(GTP)-BINDING PROTEIN

HBS1-LIKE

HOMEODOMAIN PROTEIN HLX1 (HOMEODOMAIN PROTEIN HB24).

NUCLEAR TRANSCRIPTION FACTOR Y SUBUNIT BETA (NF-Y PROTEIN CHAIN B) (NF-YB) (CCAAT-BINDIN

GROWTH FACTOR RECEPTOR-BOUND PROTEIN 2 (GRB2 ADAPTER PROTEIN) (SH2/SH3 ADAPTER GRB

SERINE/THREONINE-PROTEIN KINASE NEK2 (EC 2.7.1.37) (NIMA-RELATED PROTEIN KINASE 2) (NIMA-I

E2A-PBX1-ASSOCIATED PROTEIN; PUTATIVE 47 KDA PROTEIN.

NEURON NAVIGATOR 1; NEURON NAVIGATOR-1; PORE MEMBRANE AND/OR FILAMENT INTERACTING

NEURON NAVIGATOR 3; PORE MEMBRANE AND/OR FILAMENT INTERACTING LIKE PROTEIN 1; STEERING

Involved Proteins Network

- 
- Red = Highest involvement; Blue = Lowest Involvement
 - Based on our lowest estimates for up/down regulation
 - Based on the high confidence set of protein interactions
 - Measured gene expressions are not listed

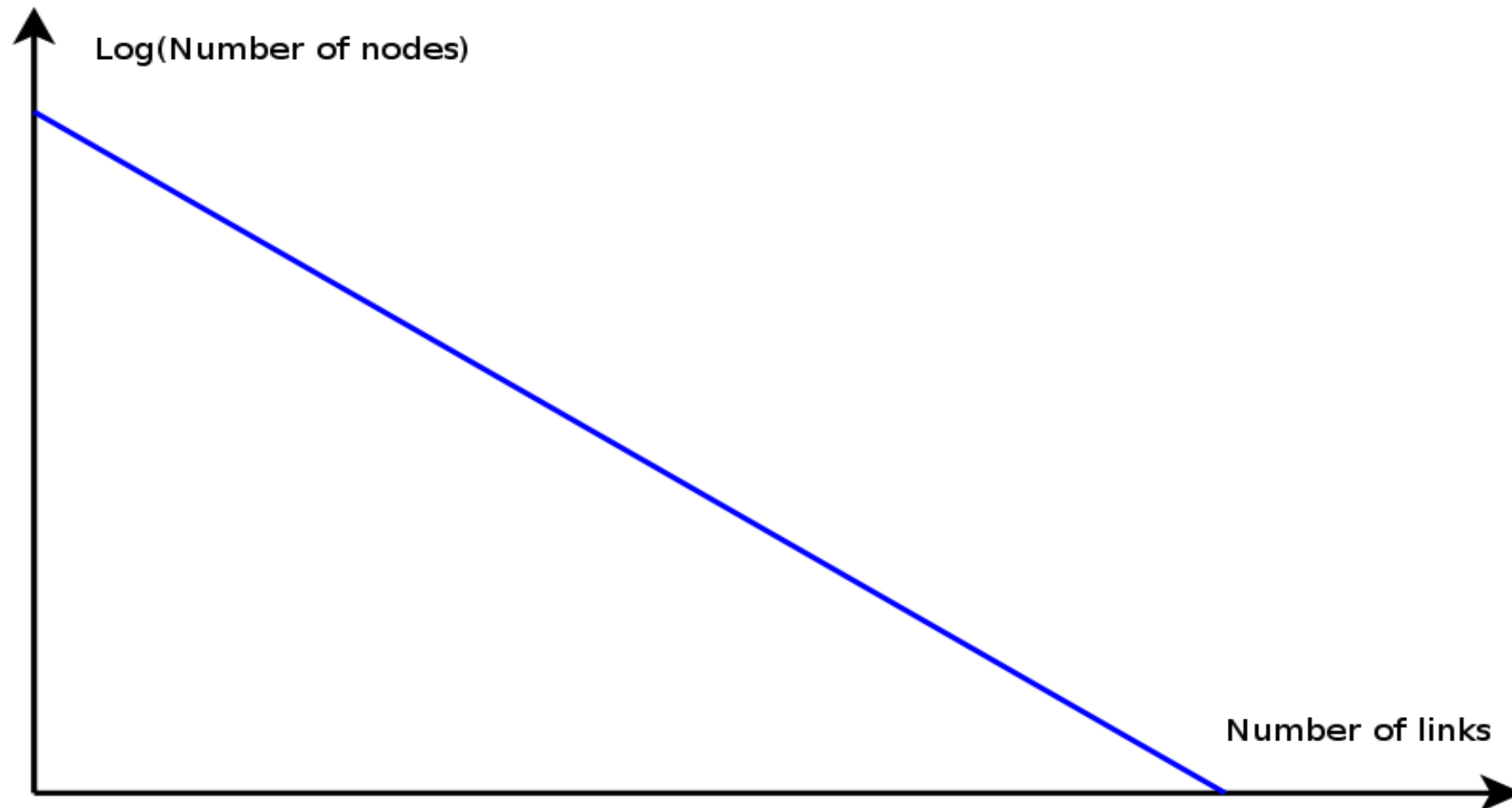
Jean François Rual *et al.* Towards a Proteome Scale Map of the Human Protein Protein Interaction Network – Nature 2005 – vol 437, p. 1173-1178

Model Variabilities

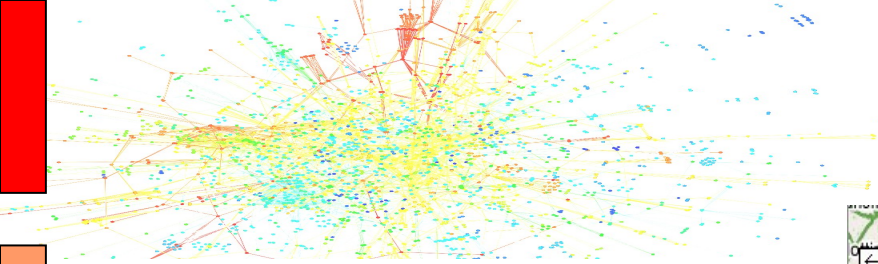
- What does the signal represent ?
 - signal in each node is the regulation ratio
 - signal in each node is the abs regulation ratio
 - signal in each node is the log abs regulation ratio
 - signal is one of the micro-array measurements
 - signal is the log of the micro-array measurement
- How to propagate ?
 - based on the protein interaction strength
 - based on the inverse of the protein interaction strength
 - unweighed

Small Worlds

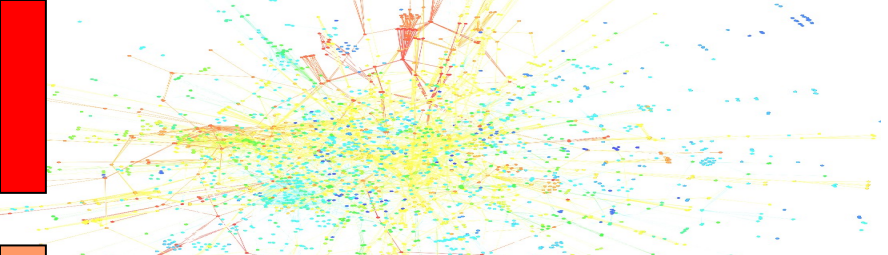
- Number of nodes that have a specific number of links: $\log(\#nodes) \sim -\#links$



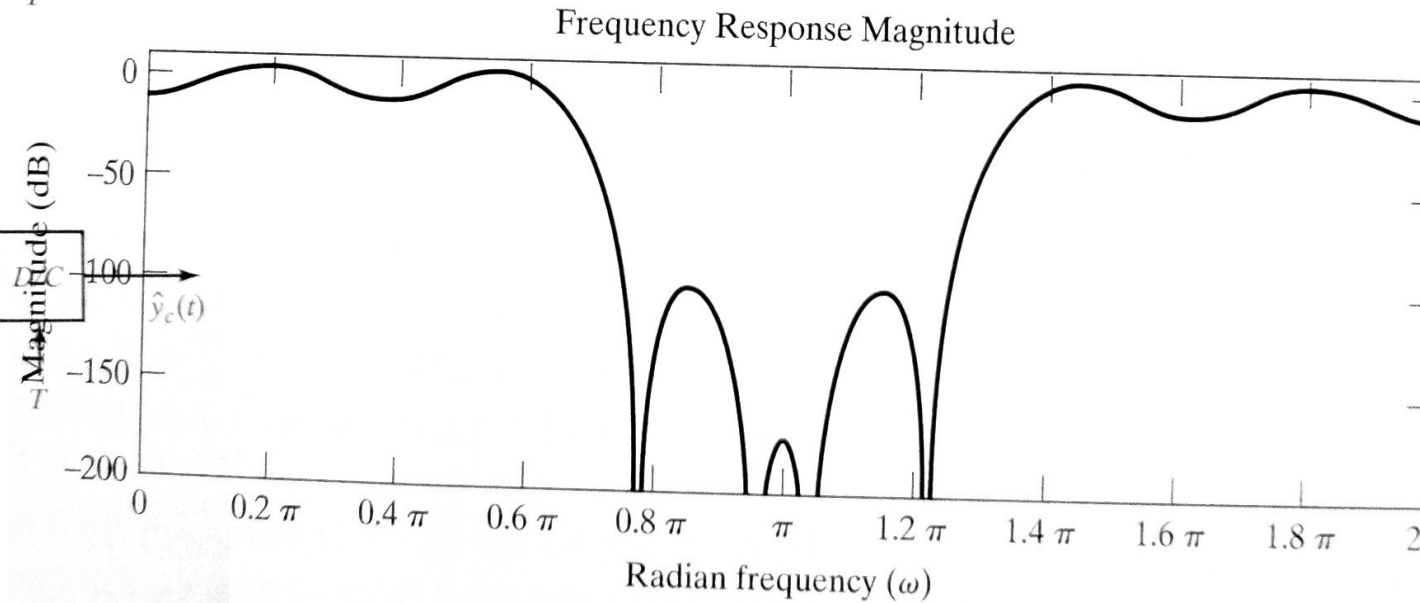
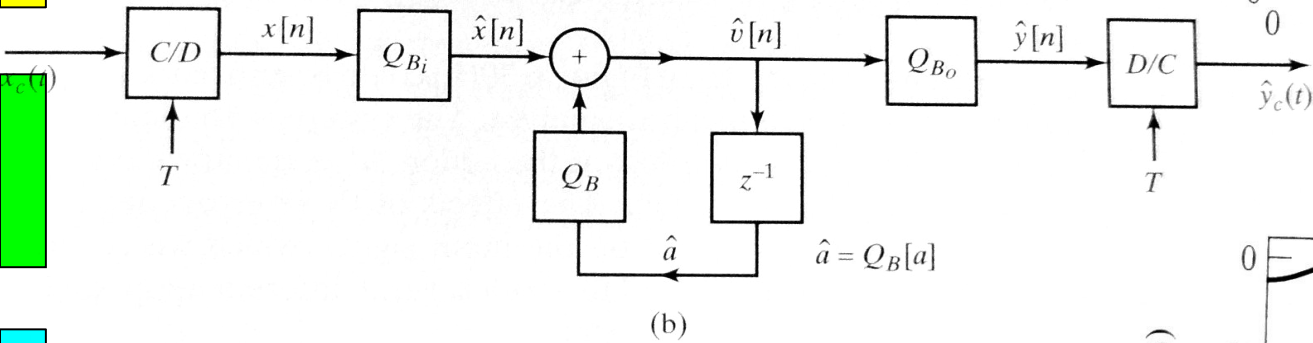
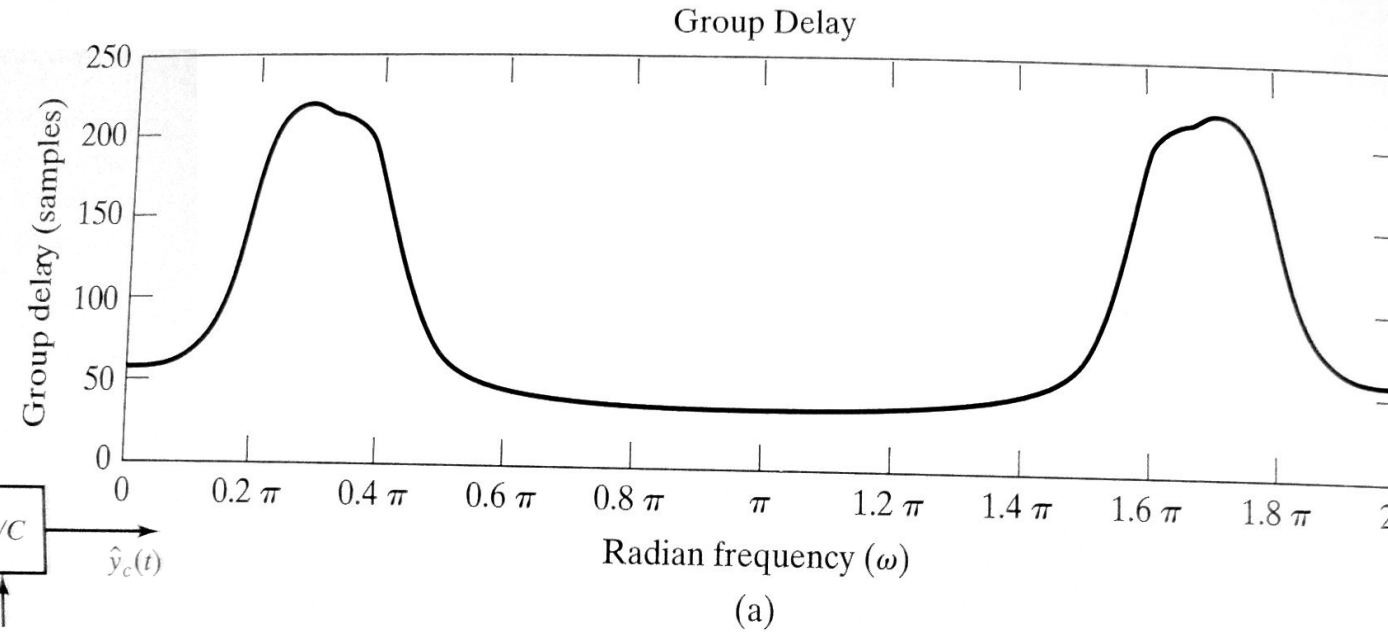
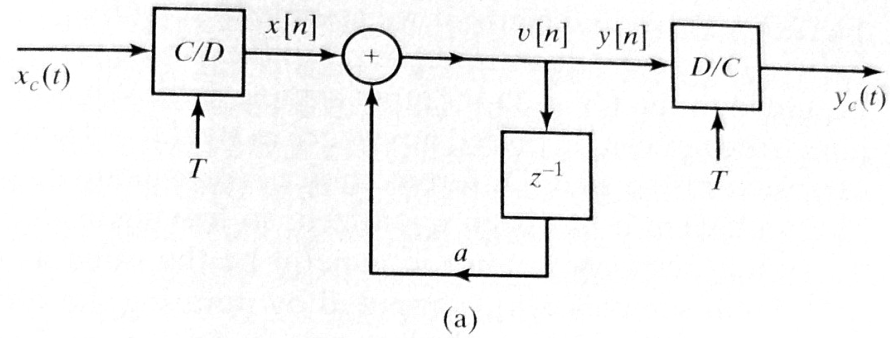
Small World



Network Structure

- 
- Relevance
 - Is a protein its function defined by its position in the network ?
 - Is the network dependent on a protein its proper functioning ?
 - What [useful] general properties of cell systems are available ?

Digital Filter Systems

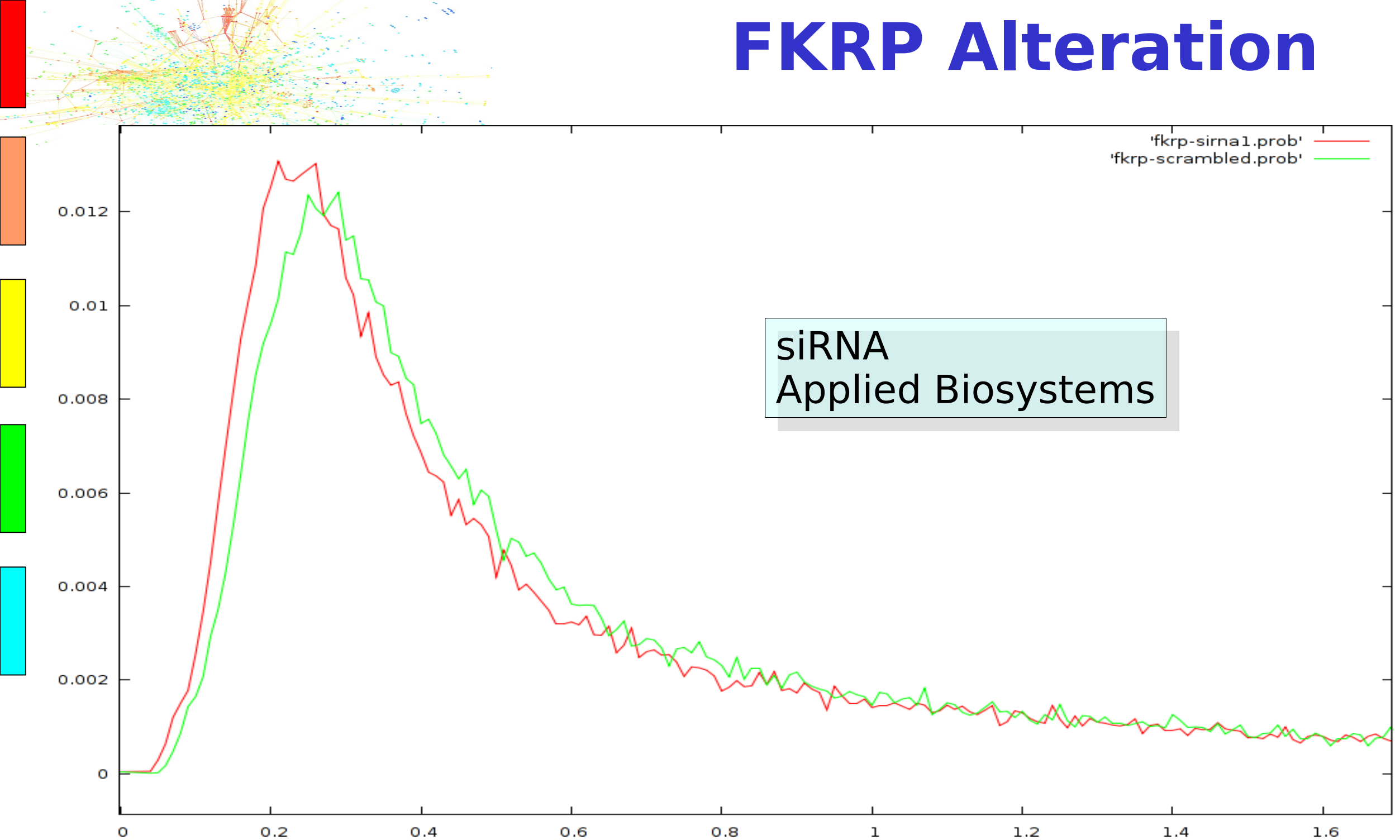


Network Structure

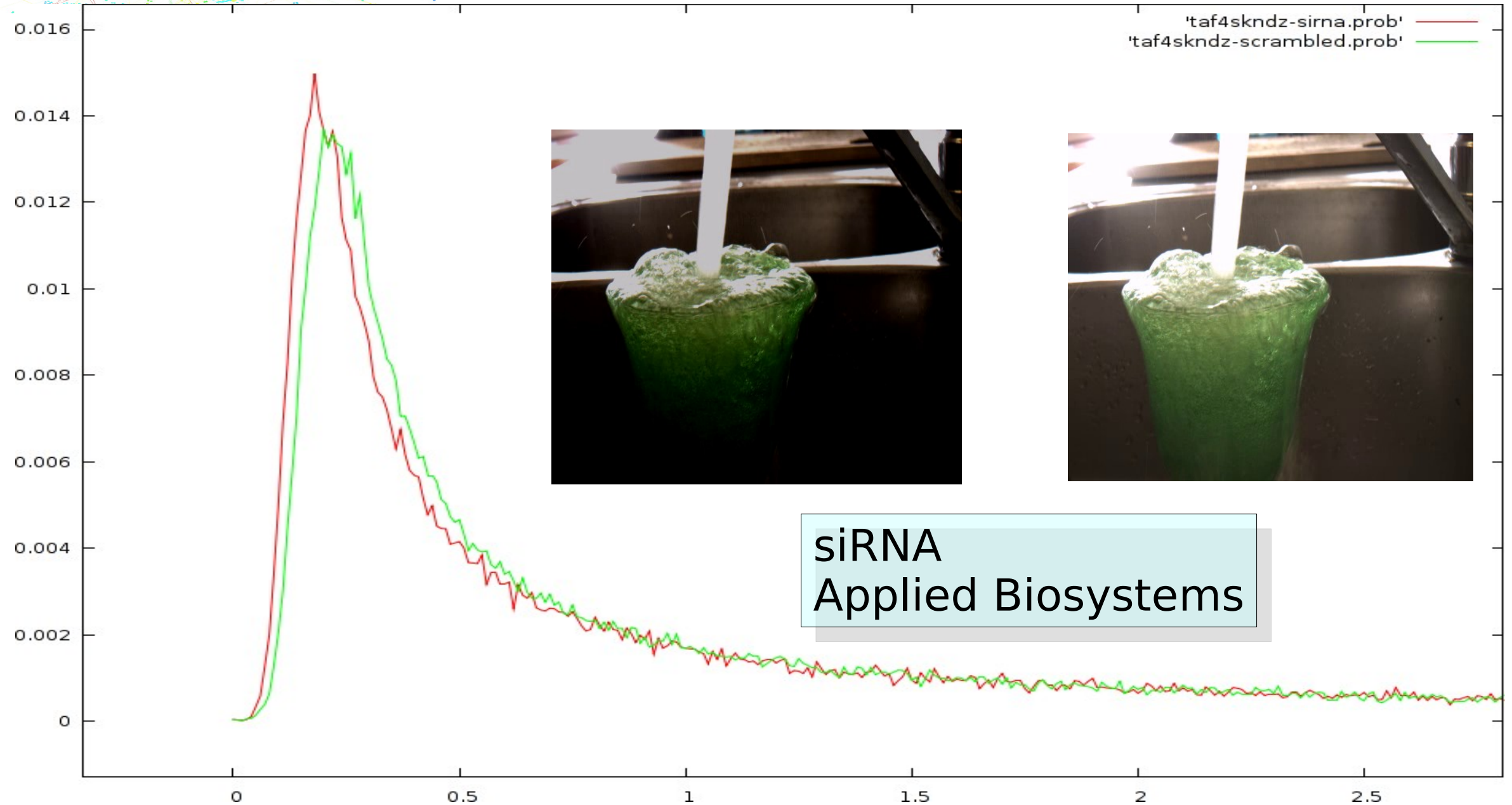
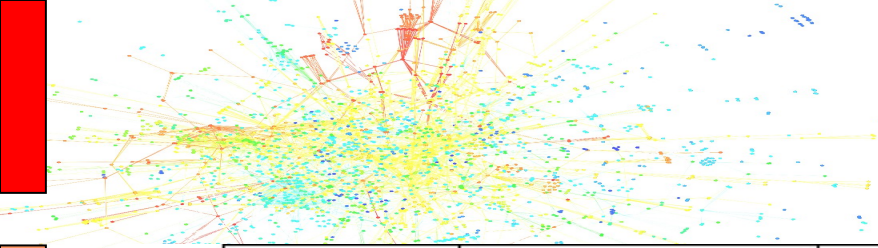
- What [useful] general properties of cell systems are available ?
 - throughput, capacity, delay, synchronization behavior, frequency response, phase response etc...

=> Micro-array distributions

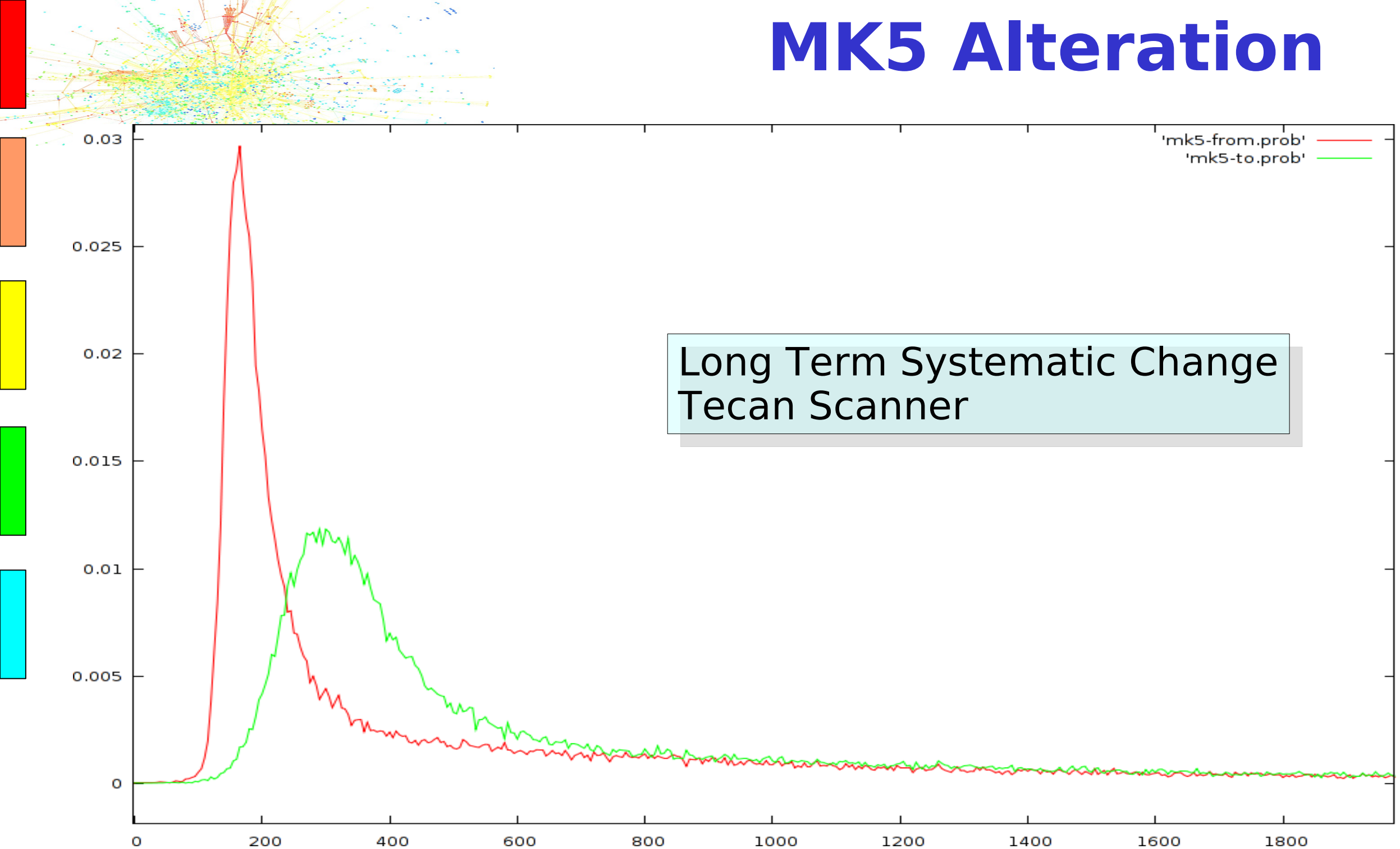
FKRP Alteration



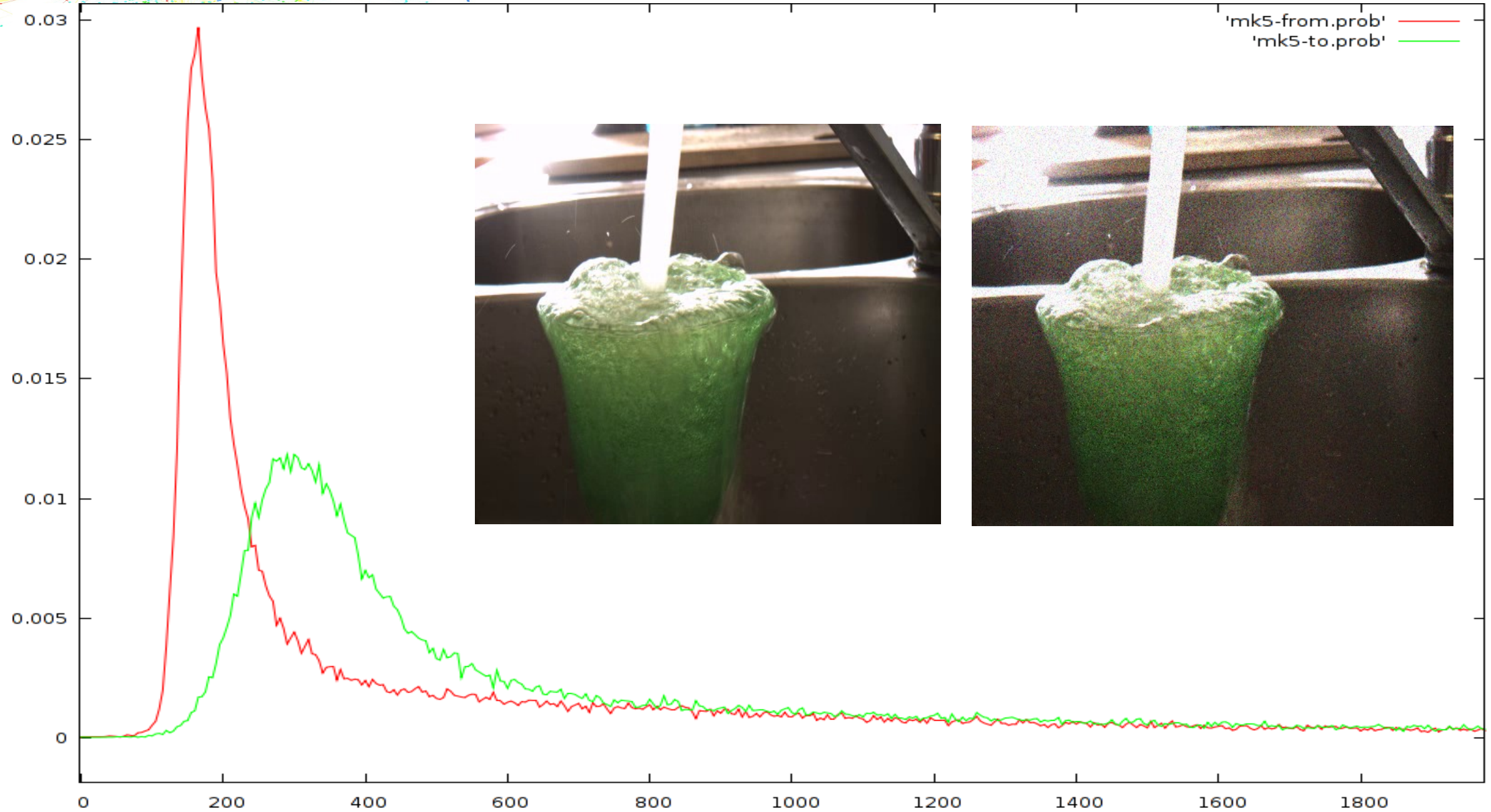
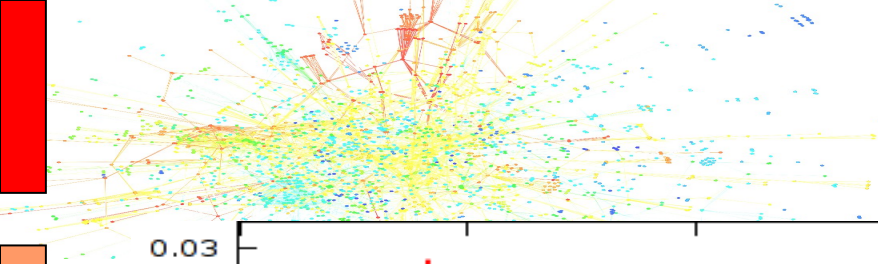
TAF4 Alteration



MK5 Alteration



MK5 Alteration



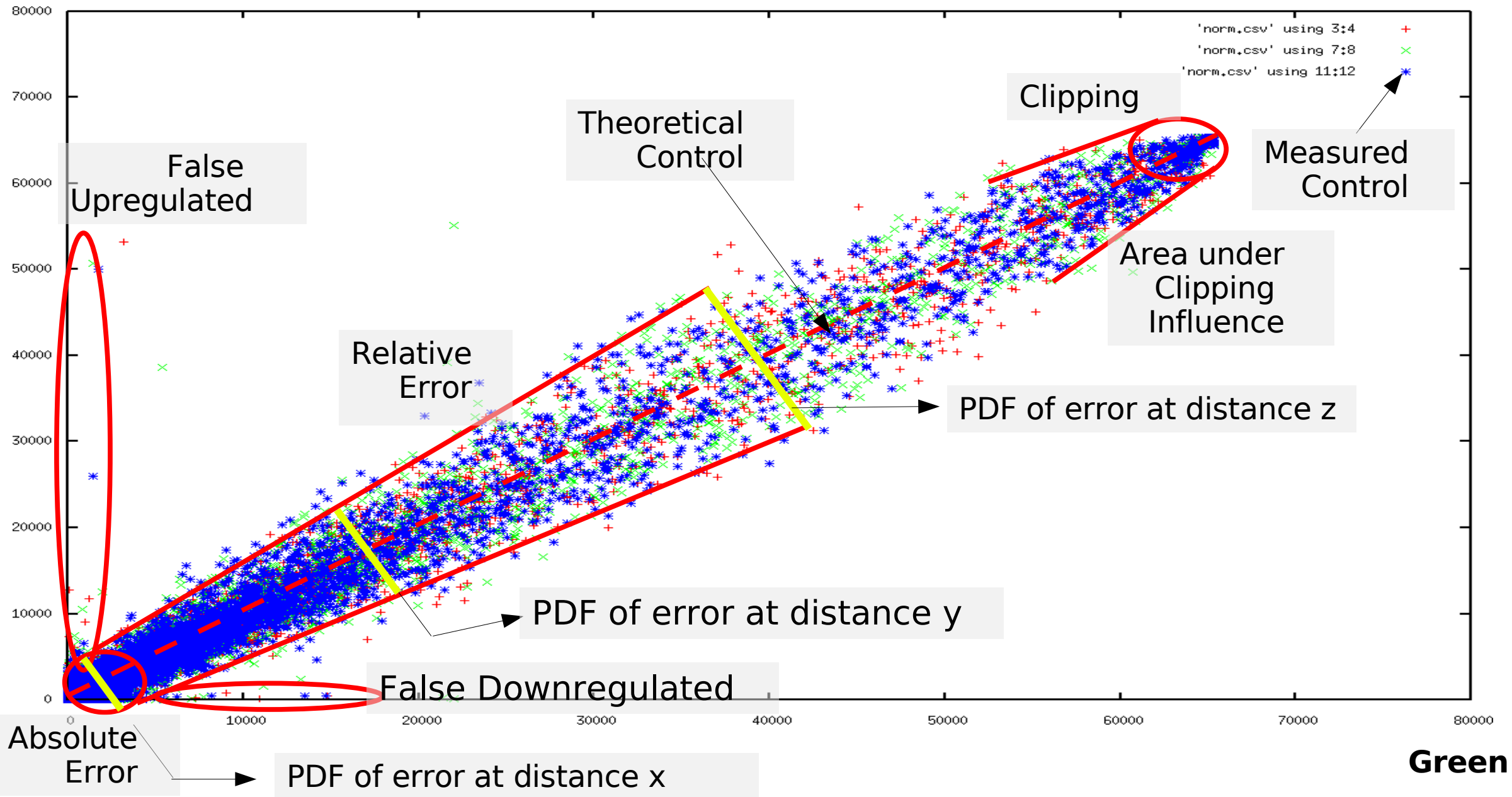


Sources of Errors

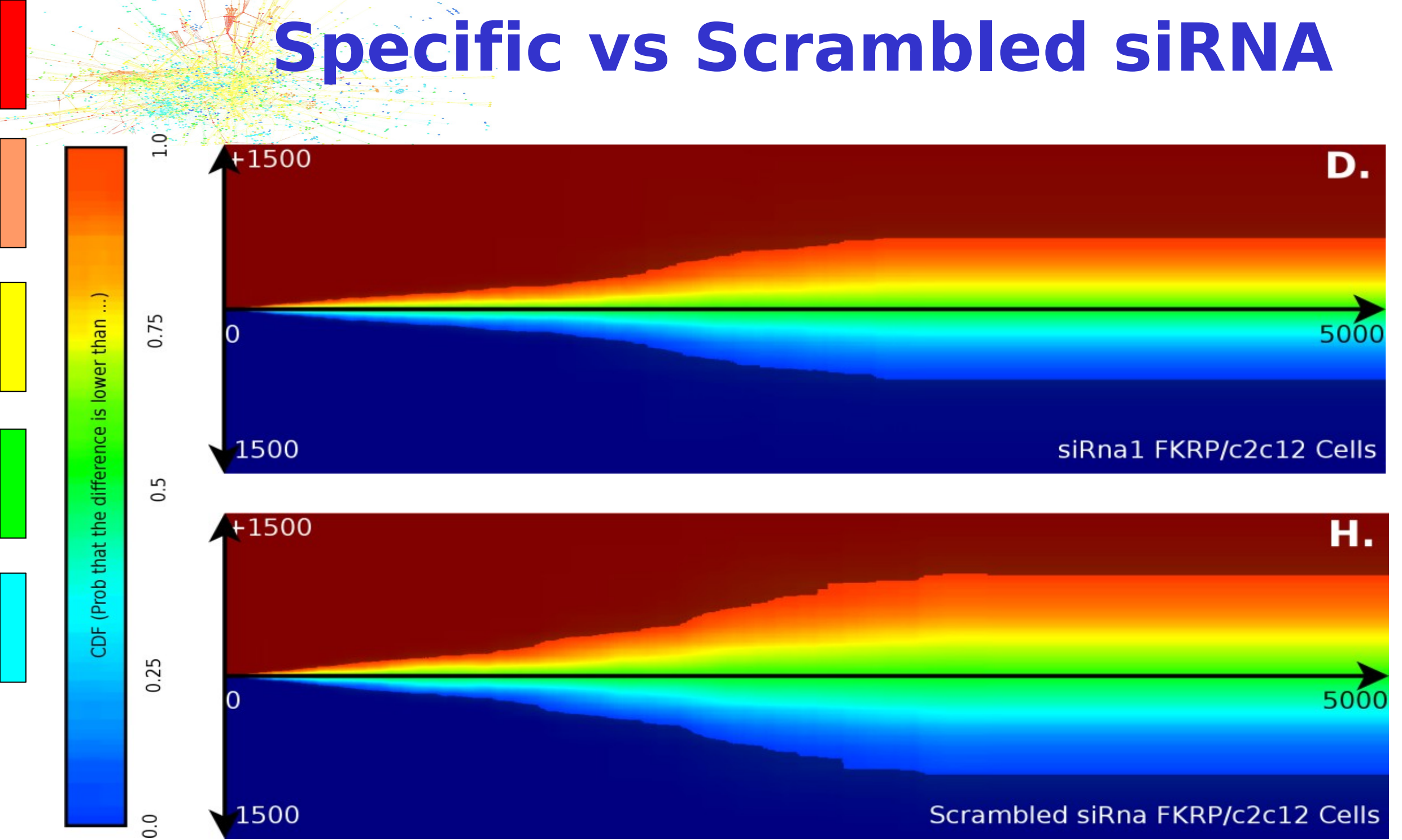
- Chemical/Physical
 - Hybridization
 - Quenching
 - Probe efficiency
 - Age of the plates
- Experimental
 - Laboratory setup
 - Sample handling
- Machine related
 - Measurement sensitivity
 - Dynamic range
- Biological Amplification process

A Control Slide

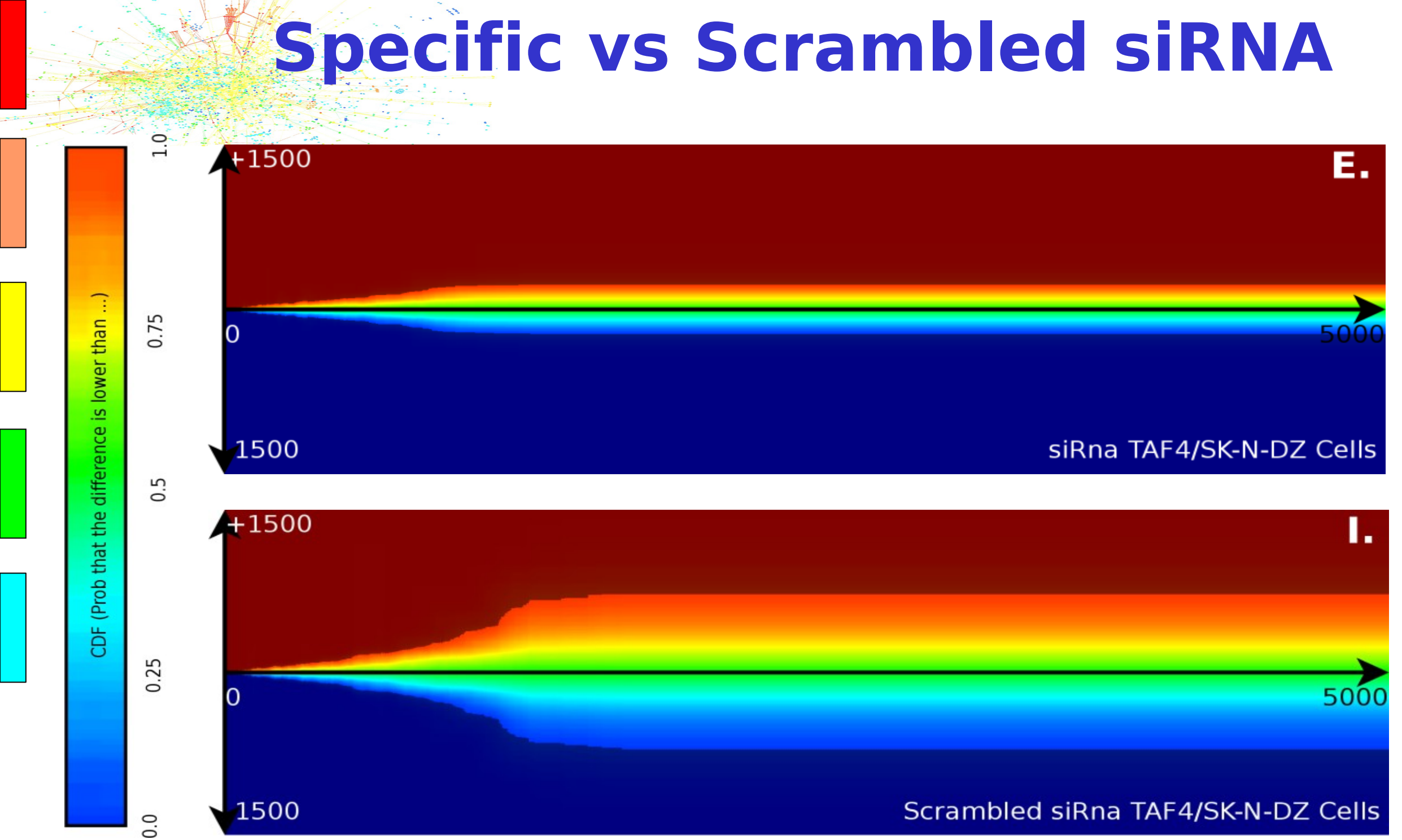
Red



Specific vs Scrambled siRNA



Specific vs Scrambled siRNA

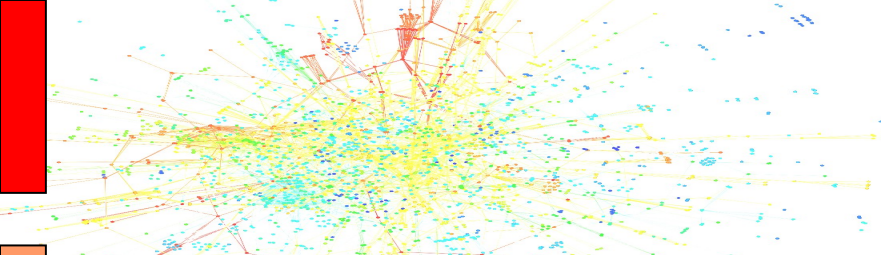


Taken Together

- Information is propagated throughout networks
- Multiplicative errors
- Widening of the probability distribution

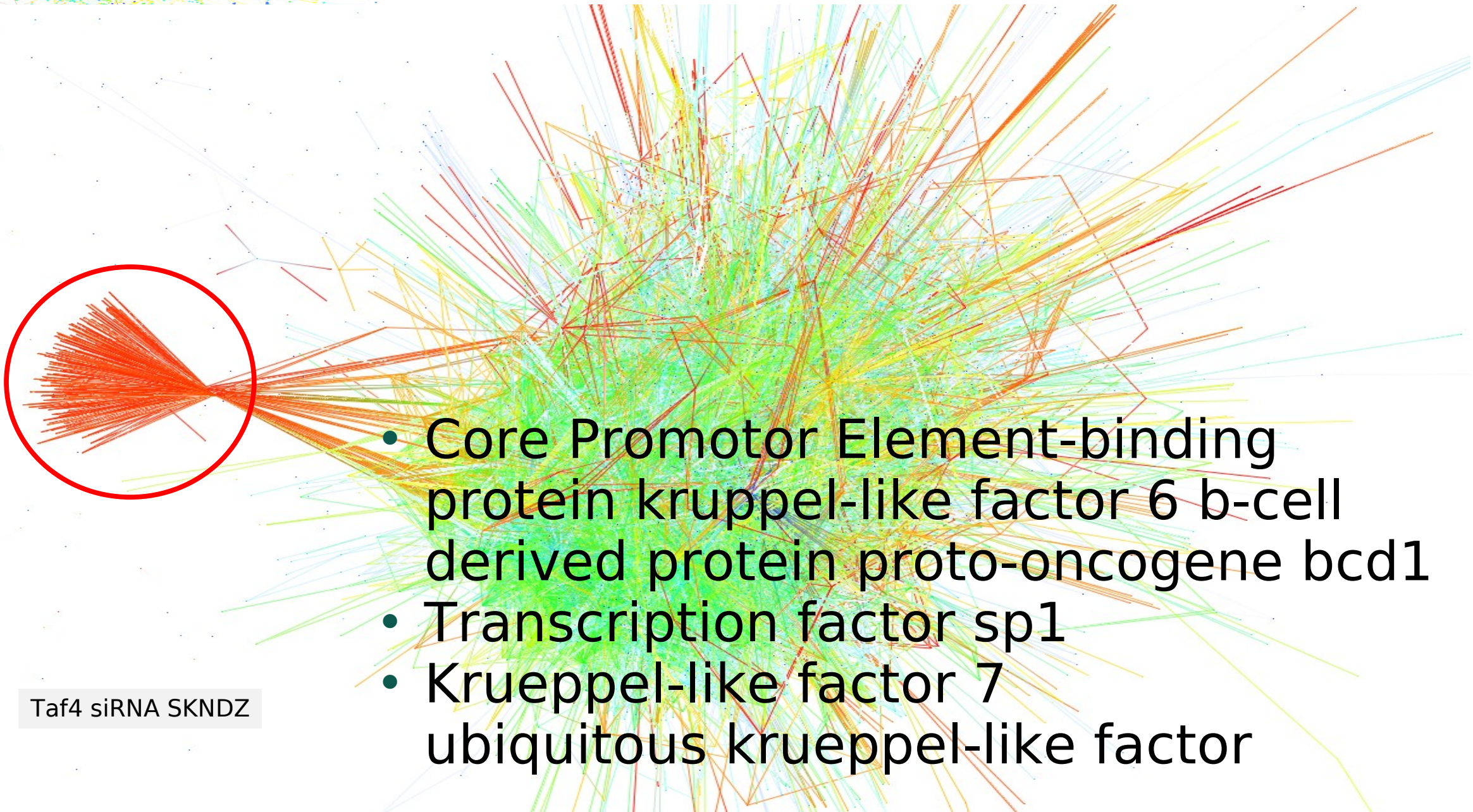
Presence of a Systematic Factor with
most gene alterations
-> some form of noise

Questions

- 
- Is the variability real noise or an oscillatory phenomenon or an occurrence of random events ?
 - What impact has synchronization of cells on the measurement/wideness ?
 - How does the overall distribution affect the cell behavior
 - How does the protein distribution affect the working of proteins for which its function is well understood
 - Can we sharpen, widen the distribution
 - Is the distribution related to the energy output/input of the cell ?

How does this relate to networks ?

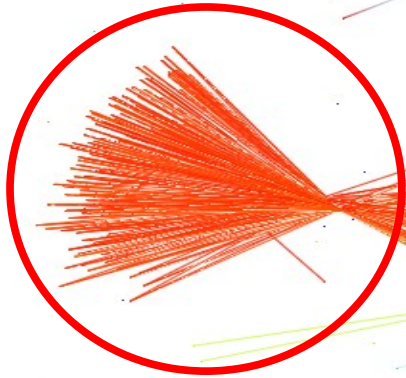
Network Position



Taf4 siRNA SKNDZ

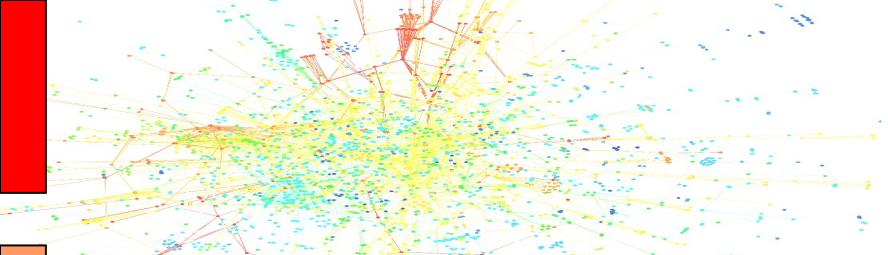
Network Position

- Will highly connected proteins
 - become more stable/unstable
 - drive noise into/away from other pathways
 - provide a noise background for the cell system ?

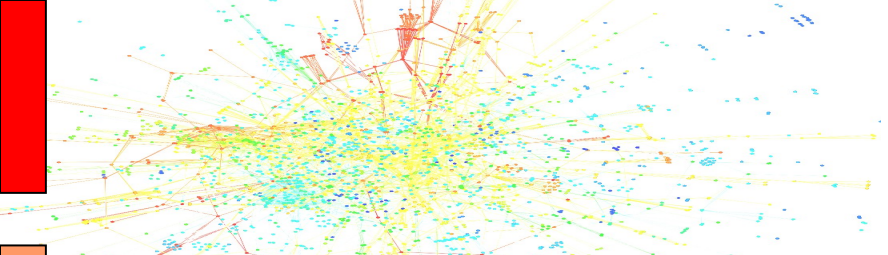


Taf4 siRNA SKNDZ

Questions

- 
- How does 1 node influence the overall 'noise' output
 - How does the overall noise affect each node ?
 - Does one protein increase or decreases the noise level of another protein without altering its expression
 - Can we relate the noise level to the distance of the alteration ?

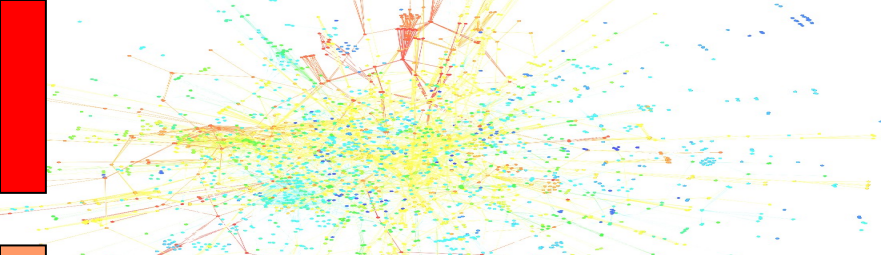
Acknowledgments

- 
- MK5
 - Nancy Gerits
 - Ugo Moens
 - TAF4
 - Kirsti Jakobsen
 - Marijke Van Ghelue
 - Ugo Moens
 - FKRP
 - Vigdis Brox
 - Marijke Van Ghelue
 - P53
 - Bjørn Tore Gjertsen
 - Nina Ånensen
 - Gry Sjøholt
 - Øystein Bruserud
 - Ingvild Haaland
 - 2DCOR
 - Kjell Arild Høgda

References

- Werner Van Belle, Nina Ånensen, Ingvild Haaland, Øystein Bruserud, Kjell-Arild Høgda, Bjørn Tore Gjertsen; *Correlation Analysis of 2Dimensional Gel Electrophoretic Protein Patterns and Biological Variables*; BMC Bioinformatics volume 7; nr 198; April 2006
- Nina Ånensen, Ingvild Haaland, live D'Santos, Werner Van Belle, Bjørn Tore Gjertsen; *Proteomics of p53 in Diagnostics and Therapy of Acute Myeloid Leukemia*; Current Pharmaceutical Biotechnology; Bentham Science Publishers Ltd; Volume 7; nr 3; July 2006
- Werner Van Belle, Nancy Gerits, Kirsti Jakobsen, Vigdis Brox, Marijke Van Ghelue, Ugo Moens; *Confidence Intervals on Microarray Measurements of Differentially Expressed Genes: A Case study on the effects of MK5, TAF4 and FKRP on the Transcriptome*; Gene Regulation and Systems Biology, Libertas Academus Press; nr 1; pages 52-72: May 2007

References

- 
- Mark Buchanan; Small World: Uncovering nature's hidden networks; ISBN 0 75381 689 X
 - Jean François Rual *et al.* Towards a Proteome Scale Map of the Human Protein Protein Interaction Network – Nature 2005 – vol 437, p. 1173-1178
 - Tulip - <http://tulip-software.org/>