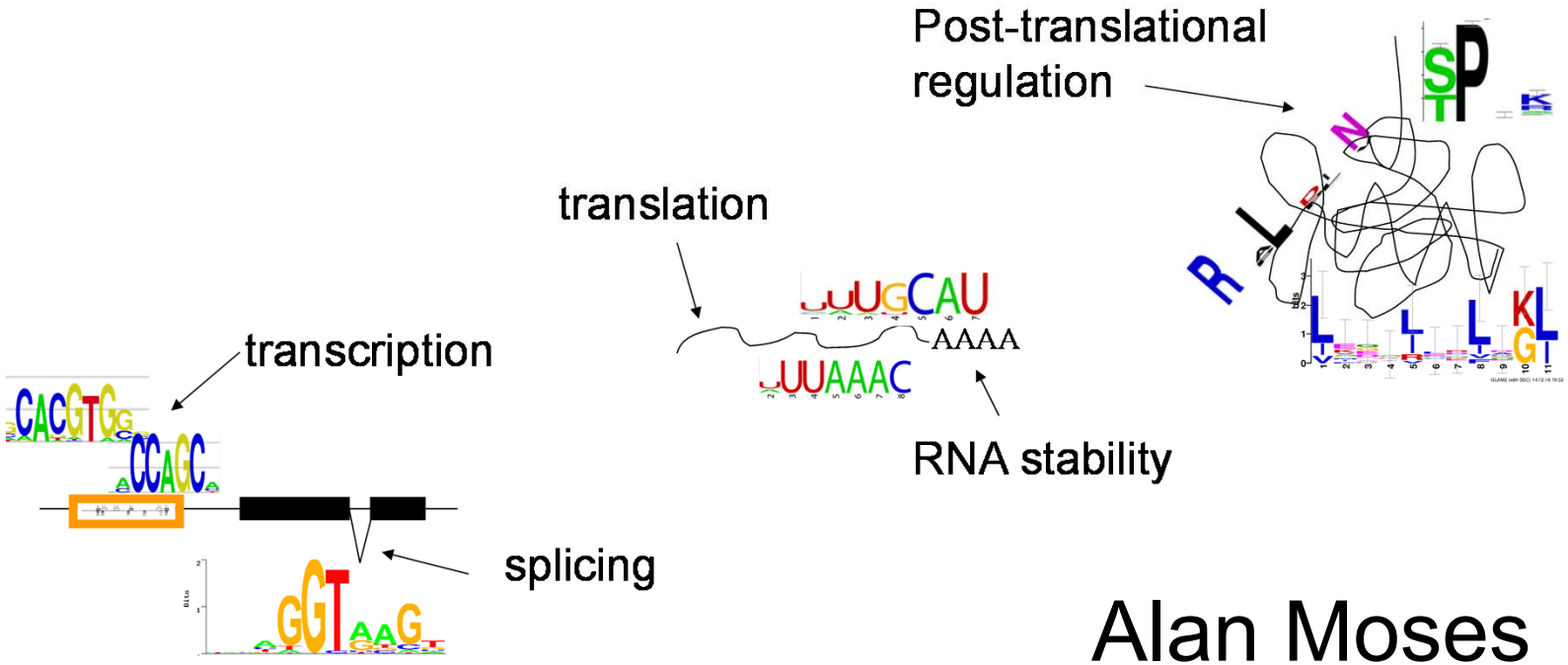
Unsupervised data analysis for the regulated proteome



Alan Moses

MOSES LAB

GBB

CSB Cell & Systems Biology

Computational Biology Lab

CAGEF

UNIVERSITY OF TORONTO

Outline

- Introduction: regulation of proteins
- Automatic identification of protein localization changes in microscopy images
- Unsupervised classification of intrinsically disordered protein regions

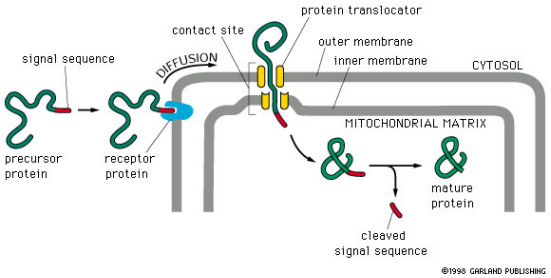
@alexijielu

@taraneh_z

What controls protein subcellular localization and stability?

- "Signals" in the primary amino acid sequence

  - Now often called motifs
  - controlled by postranslational modifications (often phosphorylation)

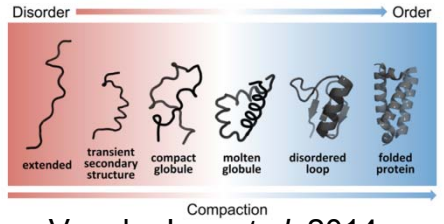


Molecular Cell
Review

A Million Peptide Motifs for the Molecular Biologist

http://elm.eu.org/

- Regulatory parts of proteins are often intrinsically disordered (IDRs)



Van der Lee *et al.* 2014

- Determinants of localization to non-membrane bound organelles are not currently understood

Intrinsically disordered proteins in cellular signalling and regulation

Peter E. Wright and H. Jane Dyson

Trends in Cell Biology

CellPress
REVIEWS

Review

Protein Phase Separation: A New Phase in Cell Biology

Regulation of p27 (aka KIP1, CDKN1B)

p27

```
  1 MSNVRVSNGSPSLERMDARQAEHPKPSACRNLFGPVDHEELTRDLEKHCRDMEEASQRKW

 61 NFDFQNHKPLEGKYEWQEVEKGSLPEFYYRPPRPPKGACKVPAQESQDVSGSRPAAPLIG

121 APANSEDTHLVDPKTDPSDSQTGLAEQCAGIRKRPATDDSSTQNKRANRTEENVSDGSPN

181 AGSVEQTPKKPGLRRRQT
```

Residues 153-166 are an NLS

Importin p27

Zeng et al. 2002

Regulation of p27 (aka KIP1, CDKN1B)

p27

```
  1 MSNVRVSNGSPSLERMDARQAEHPKPSACRNLFGPVDHEELTRDLEKHCRDMEEASQRKW

 61 NFDFQNHKPLEGKYEWQEVEKGSLPEFYYRPPRPPKGACKVPAQESQDVSGSRPAAPLIG

121 APANSEDTHLVDPKTDPSDSQTGLAEQCAGIRKRPATDDSSTQNKRANRTEENVSDGSPN

181 AGSVEQTPKKPGLRRRQT
```

Residues 153-166 are an NLS

To nucleus

Zeng et al. 2002

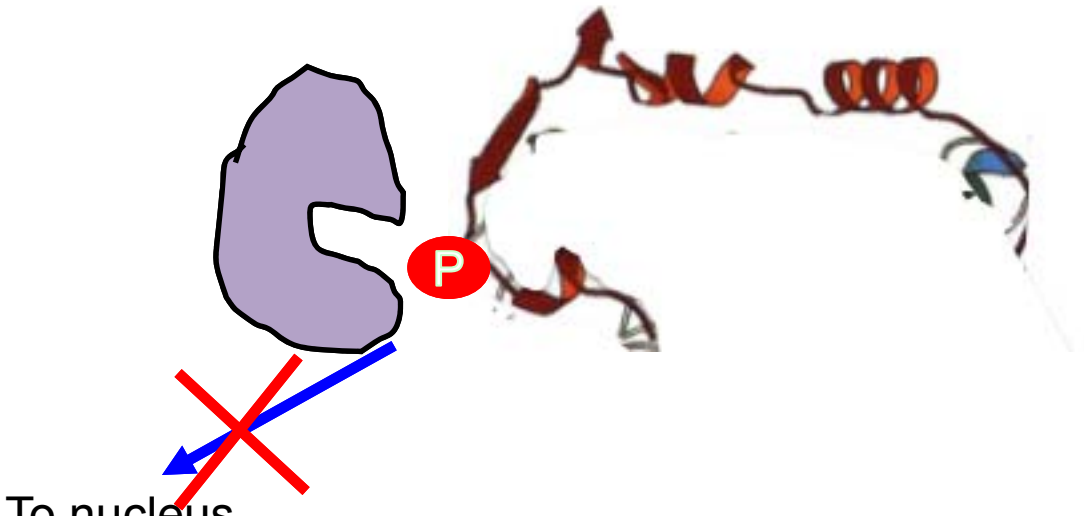
Regulation of p27 (aka KIP1, CDKN1B)

p27

```
  1 MSNVRVSNGSPSLERMDARQAEHPKPSACRNLFGPVDHEELTRDLEKHCRDMEEASQRKW

 61 NFDFQNHKPLEGKYEWQEVEKGSLPEFYYRPPRPPKGACKVPAQESQDVSGSRPAAPLIG

121 APANSEDTHLVDPKTDPSDSQTGLAEQCAGIRKRPATDESSTQNKRANRTEENVSDGSPN

181 AGSVEQTPKKPGLRRRQT
```

Residues 153-166 are an NLS

Residue 157 is a phosphorylation site for Akt

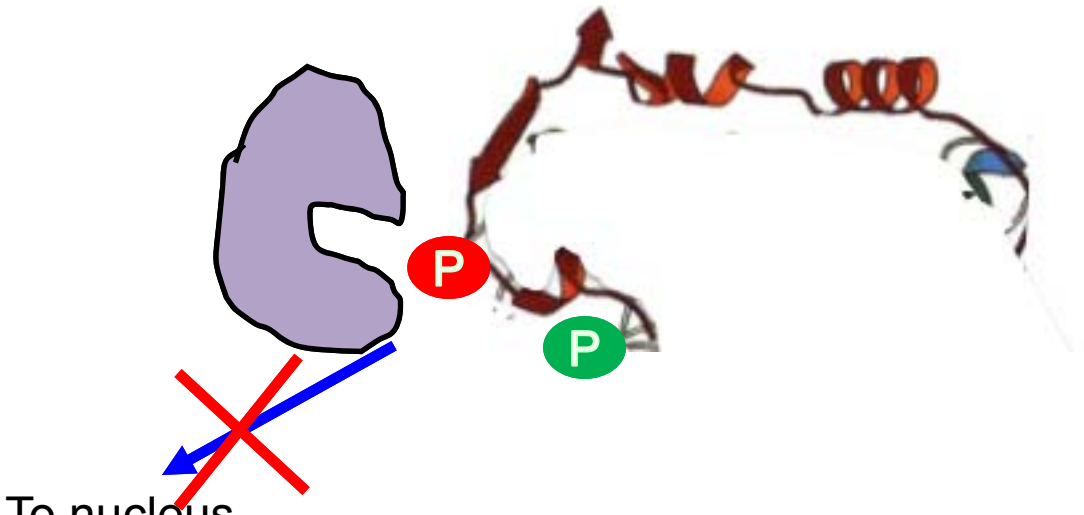


Shin et al. 2002
Zeng et al. 2002

Regulation of p27 (aka KIP1, CDKN1B)

p27

```
  1 MSNVRVSNGSPSLERMDARQAEHPKPSACRNLFGPVDHEELTRDLEKHCRDMEEASQRKW

 61 NFDFQNHKPLEGKYEWQEVEKGSLPEFYYRPPRPPKGACKVPAQESQDVSGSRPAAPLIG

121 APANSEDTHLVDPKTDPSDSQTGLAEQCAGIRKRPATDDSSTQNKRANRTEENVSDGSPN

181 AGSVEQTPKKPGLRRRQT
```



Residues 153-166 are an NLS

Residue 157 is a phosphorylation site for Akt

Residue 187 is a phosphorylation site for CDK2/cycE

Shin et al. 2002
Zeng et al. 2002

Regulation of p27 (aka KIP1, CDKN1B)

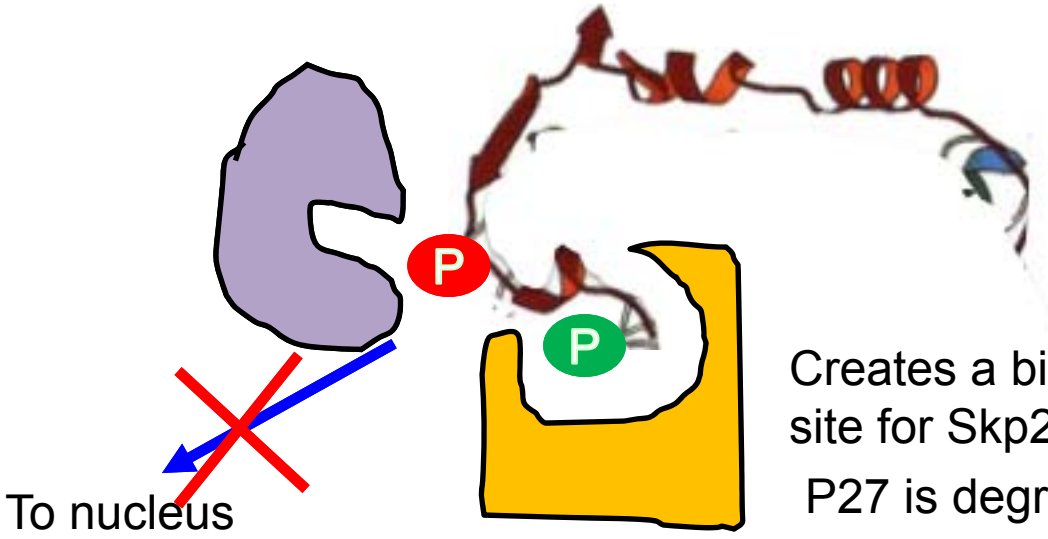
p27

```
  1 MSNVRVSNGSPSLERMDARQAEHPKPSACRNLFGPVDHEELTRDLEKHCRDMEEASQRKW

 61 NFDFQNHKPLEGKYEWQEVEKGSLPEFYYRPPRPPKGACKVPAQESQDVSGSRPAAPLIG

121 APANSEDTHLVDPKTDPSDSQTGLAEQCAGIRKRPATDESSTQNKRANRTEENVSDGSPN

181 AGSVEQTPKKPGLRRRQT
```

Residues 153-166 are an NLS

Residue 157 is a phosphorylation site for Akt

Residue 187 is a phosphorylation site for CDK2/cycE

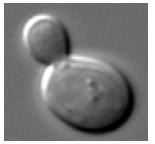


Creates a binding site for Skp2

P27 is degraded

Shin et al. 2002
Zeng et al. 2002

High-throughput cell biology



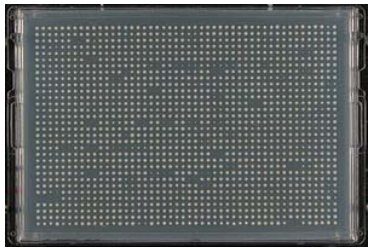
- GFP-tagging of all yeast proteins in 2003

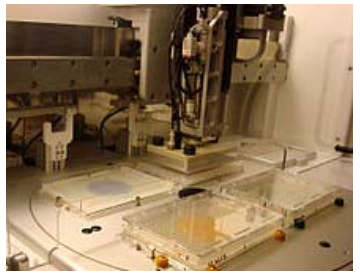
articles

Global analysis of protein localization in budding yeast

Won-Ki Huh[1]*, James V. Falvo[1]*, Luke C. Gerke[1], Adam S. Carroll[1], Russell W. Howson[1], Jonathan S. Weissman[1,2] & Erin K. O'Shea[1]

- Recent advances in automated microscopy and genetics



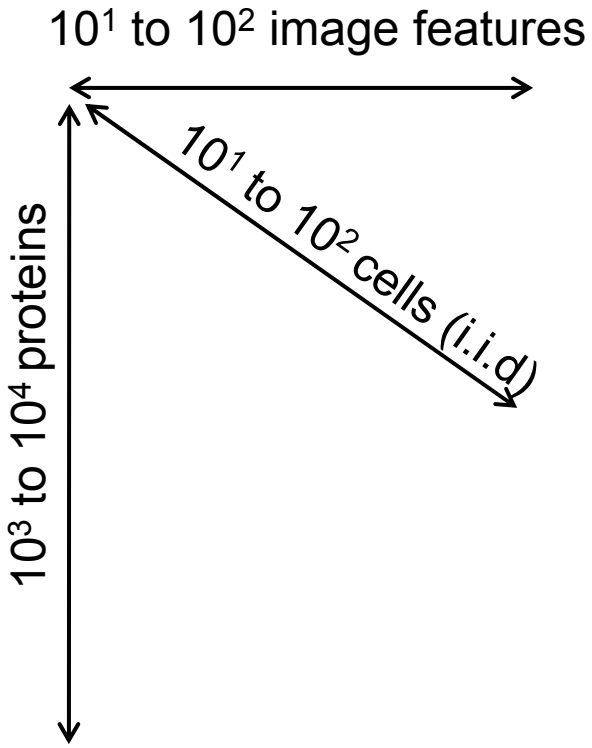


Proteome-scale microscopy data

Per experiment:

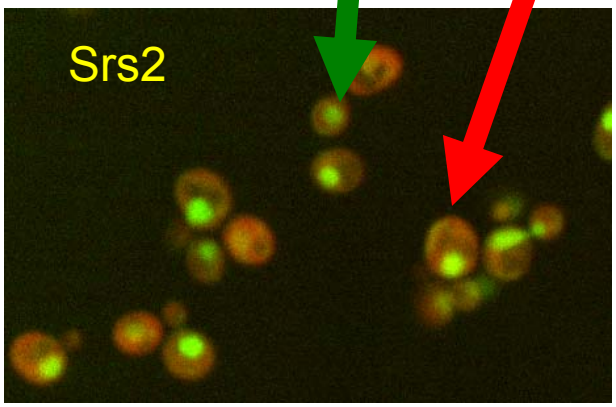
Raw: 10^5 images, 50GB

Processed: 10^6 to 10^7 data points



In each cell, Srs2 appears green

The cytoplasm appears read



For each strain, ~200 cells are imaged at high resolution

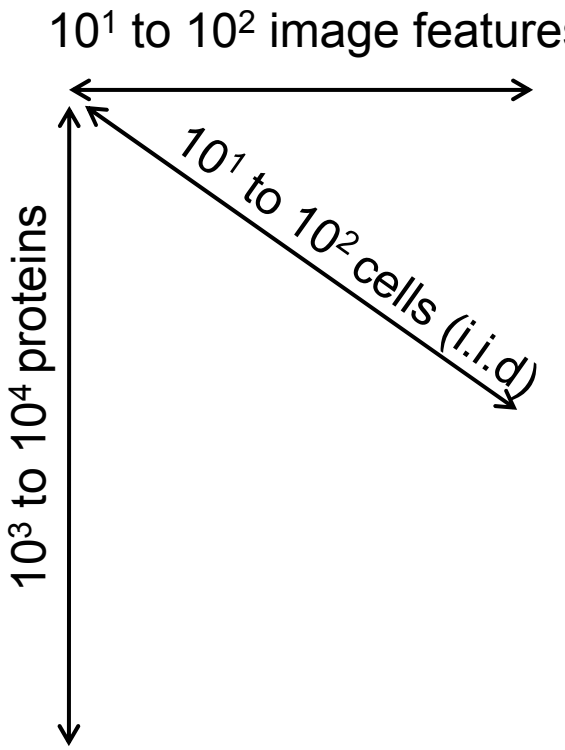
Handfield et al. *PLoS Comp. Biol.* 2013

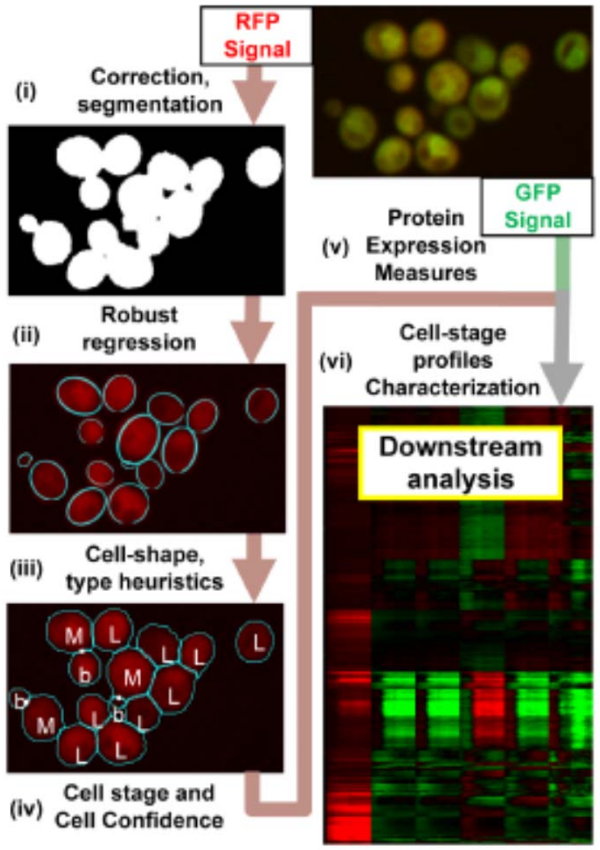
Proteome-scale microscopy data

Per experiment:

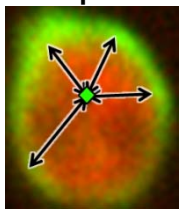
Raw: 10^5 images, 50GB

Processed: 10^6 to 10^7 data points





Interpretable features



Centre of mass of GFP

$$\vec{m} = \sum_{\vec{z} \in A_c} \vec{z} f_c(\vec{z})$$

$$X_c(\vec{z}) = \sum_{\vec{z} \in A_c} ||\vec{z} - \vec{m}|| f_c(\vec{z}),$$

"expected distance to centre of mass"

Measures "compactness" or "spread" of the GFP subcellular localization pattern.

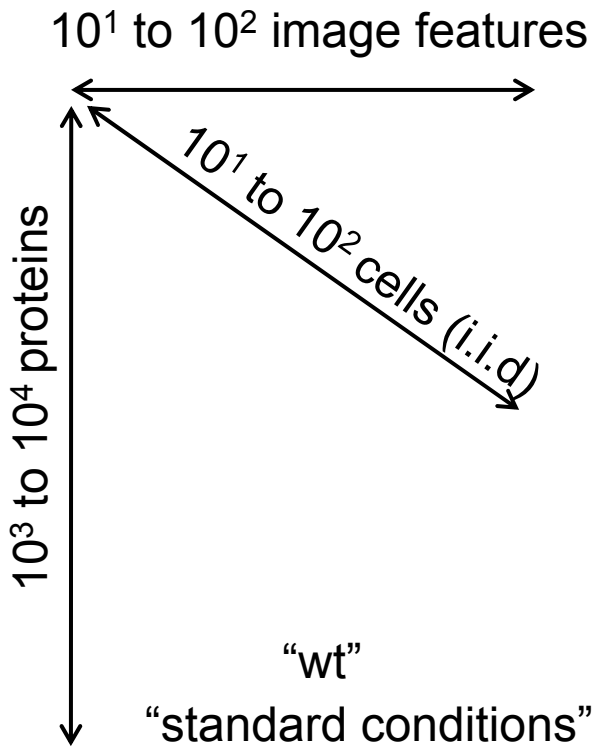
Handfield et al. *PLoS Comp. Biol.* 2013

Proteome-scale microscopy data

Per experiment:

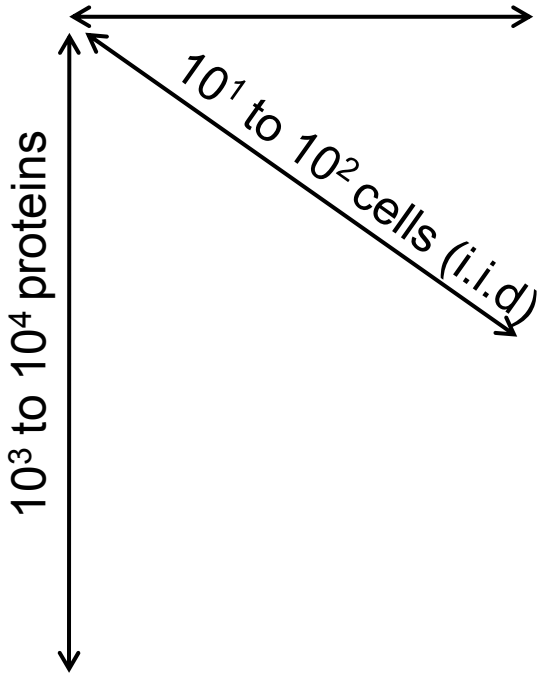
Raw: 10^5 images, 50GB

Processed: 10^6 to 10^7 data points



"wt"
"standard conditions"
"control"

Repeat the experiment !



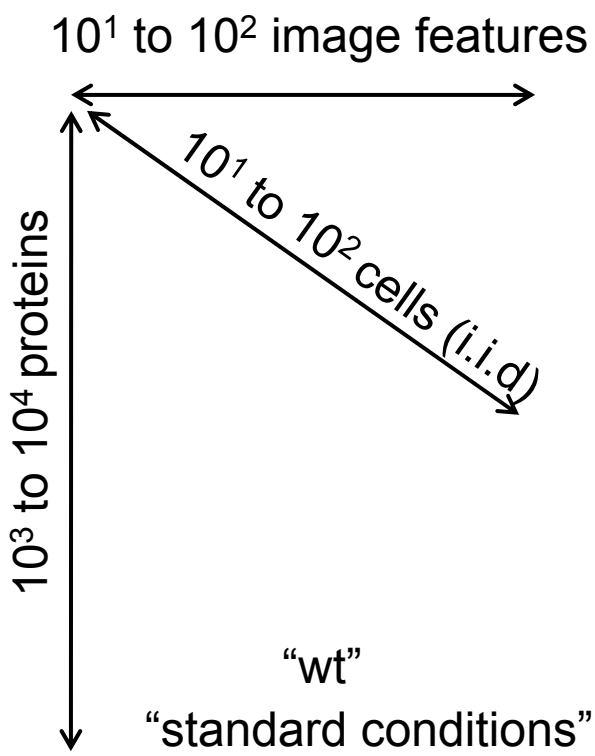
"ΔGeneX"

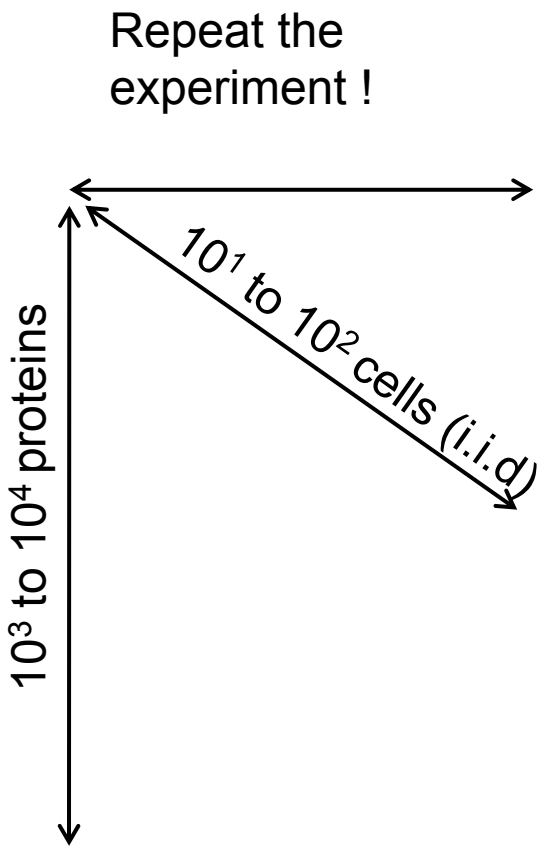
Proteome-scale microscopy data

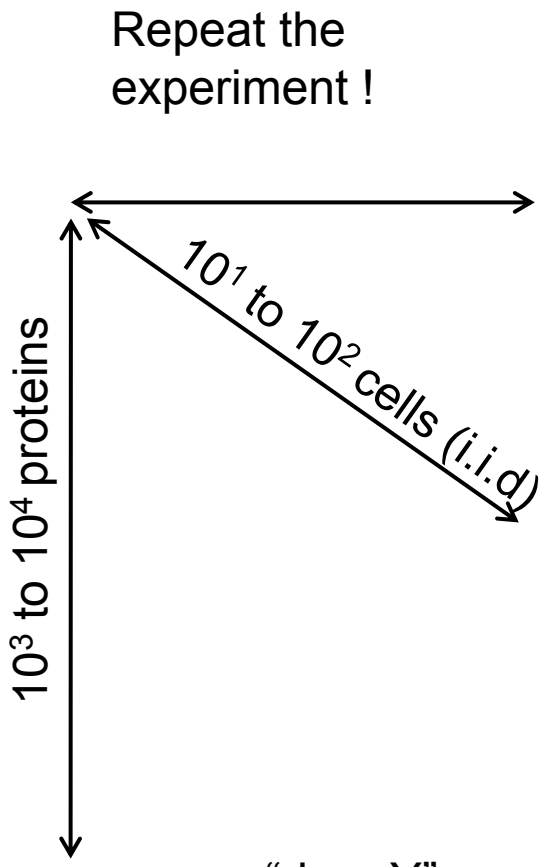
Per experiment:

Raw: 10^5 images, 50GB

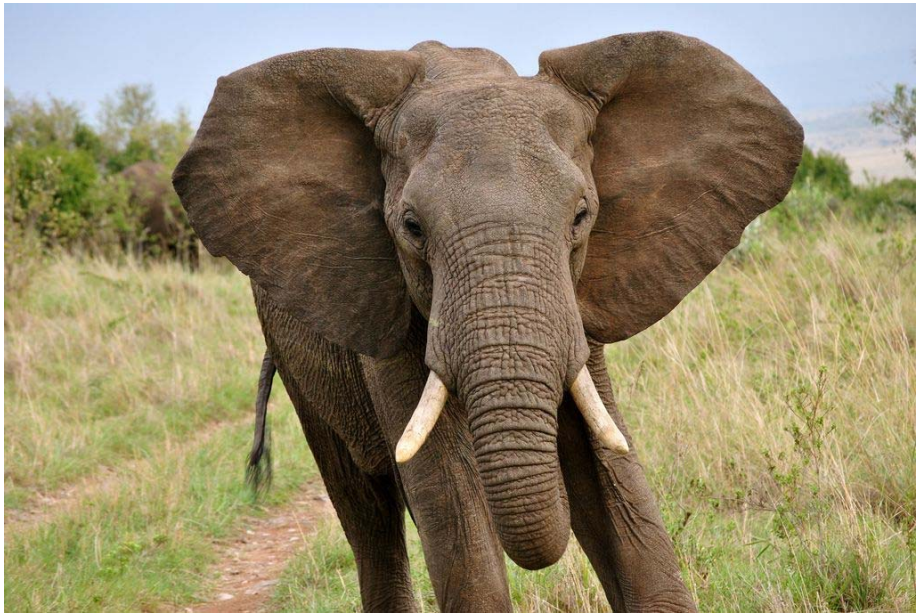
Processed: 10^6 to 10^7 data points







Chong et al. *Cell* 2015
Cyclops database (Koh et al. *G3* 2015)



change detection

nature cell biology

2012

Dissecting DNA damage response pathways by analysing protein localization and abundance changes during DNA replication stress

Johnny M. Tkach[1,2], Askar Yimit[1,2], Anna Y. Lee[2,3], Michael Riffle[4], Michael Costanzo[2,3], Daniel Jaschob[4], Jason A. Hendry[1,2], Jiongwen Ou[1,2], Jason Moffat[2,3], Charles Boone[2,3], Trisha N. Davis[4], Corey Nislow[2,3] and Grant W. Brown[1,2,5]

A novel single-cell screening platform reveals proteome plasticity during yeast stress responses

JCB 2013

Michal Breker,[1] Melissa Gymrek,[2,3] and Maya Schuldiner[1]

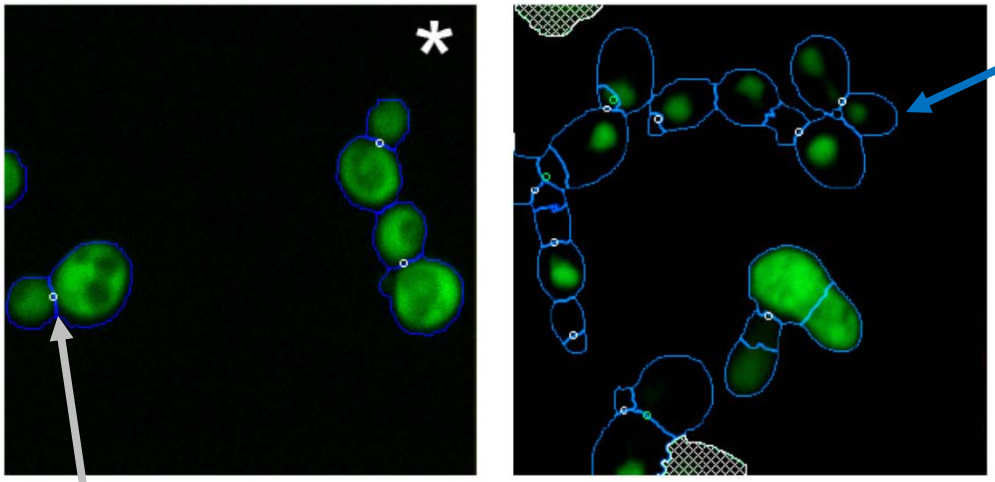
[1]Department of Molecular Genetics, Weizmann Institute of Science, Rehovot, Israel 76100
[2]Harvard-MIT Division of Health Sciences and Technology, Massachusetts Institute of Technology, Cambridge, MA 02139
[3]Whitehead Institute for Biomedical Research, Cambridge, MA 02142

No automated detection of changes in subcellular localization patterns reported… images were examined by *eye*

Why is automated change detection in microscopy images so hard?

wt

*ELM*1 Δ



The inferred cell boundary indicated in blue

Gre3 (example of a real change)

The inferred bud neck indicated in white

- Mutation causes cells to grow in long chains
- Cells are smaller and not as round
- Segmentation algorithm (trained on wt cells) is totally confused
- Some cells are unaffected

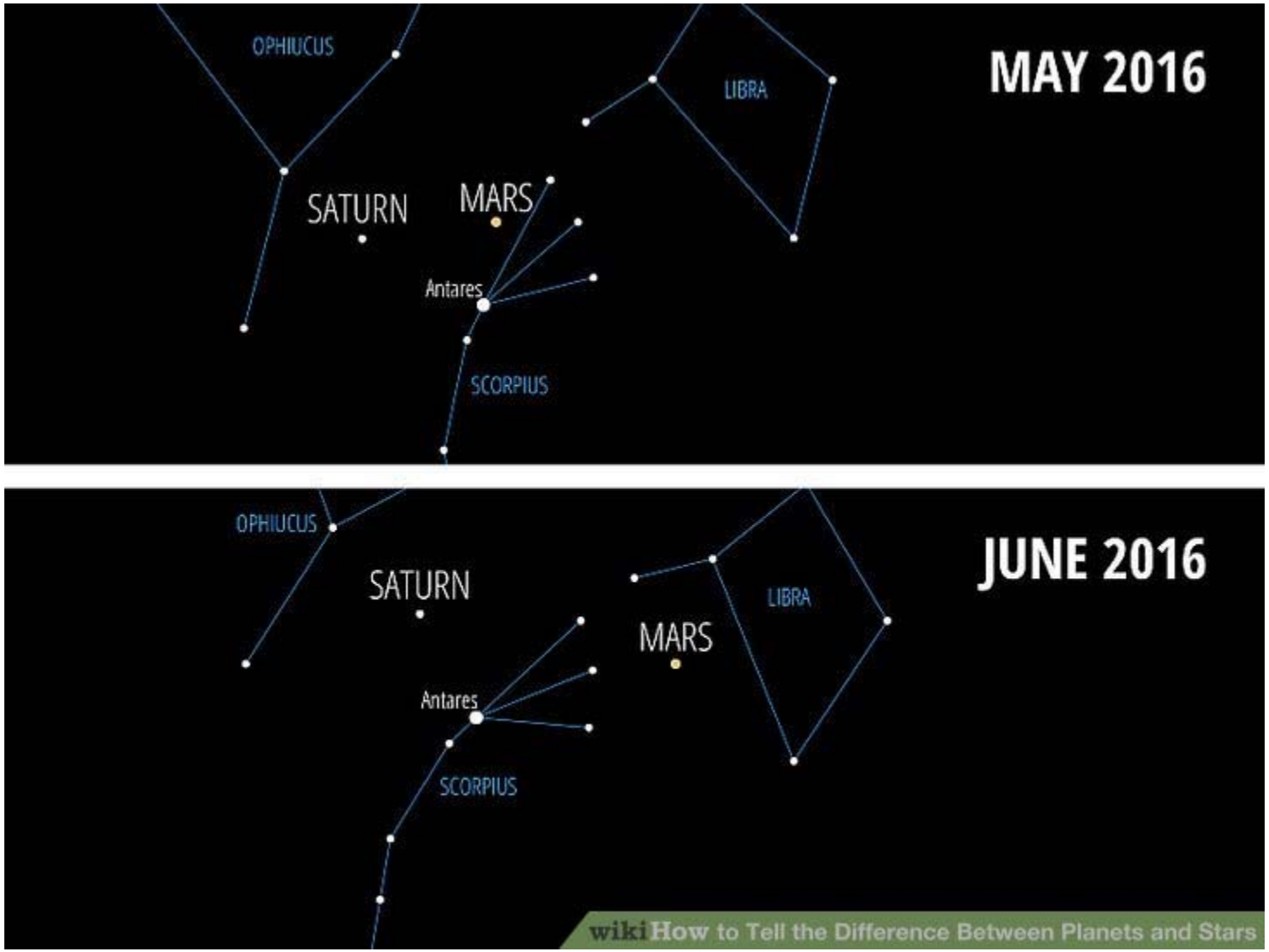
Naïve statistics in the feature space identify 1000s of changes that are biased by the localization class

Lu & Moses *PLoS One 2016*

Why is automated change detection in microscopy images so hard?

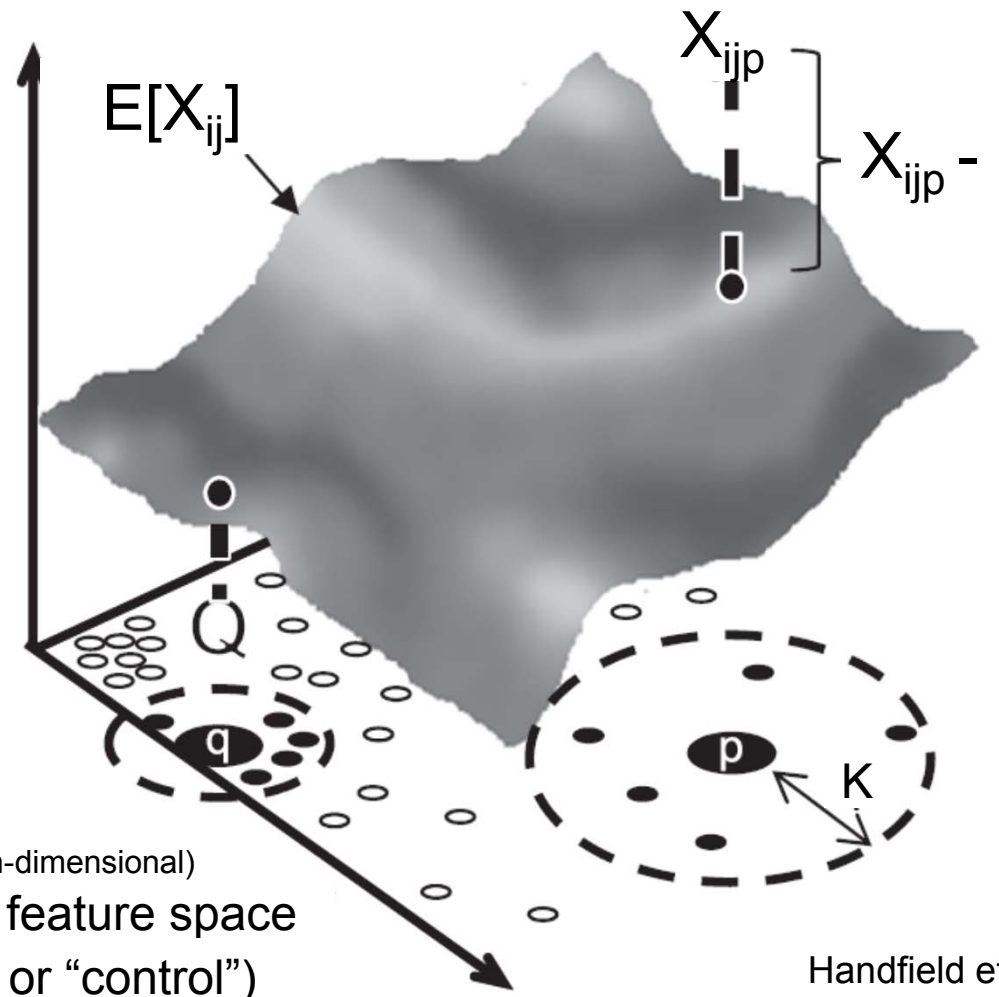
- “Global effects”, a type of covariate shift such that all localization patterns look a little bit different
  - Mutation or drug causes overall change in cell shape
  - Cell growth, nutritional differences, technical differences in microscopes, laser age etc.
  - Effect on feature space is heterogeneous
    - E.g., Nuclear proteins may be affected differently than cell membrane
- Cell to cell variability (incomplete penetrance)
  - Not all cells in the image display the same type of change

Global effects



Local statistics

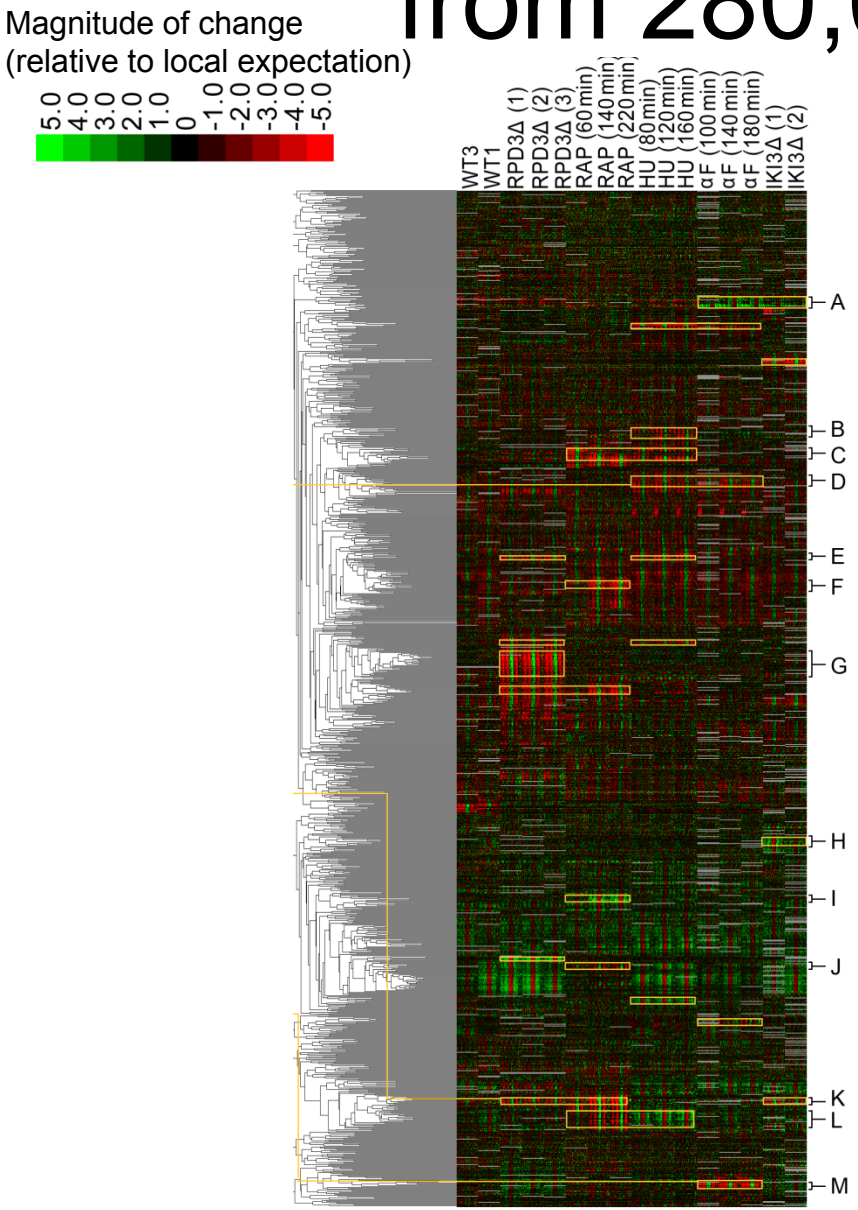
K nearest neighbours (or adaptive bandwidth kernel regression) can be used to compute local expectation



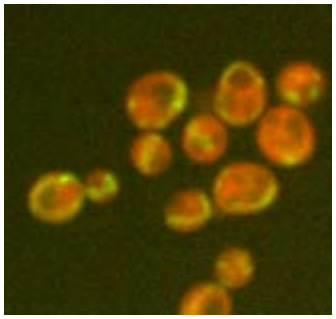
In practice, we compare the observations for protein p to 50 most similar proteins

Handfield et al. *Bioinformatics 2014*
Lu & Moses *PLoS One 2016*

Patterns of localization change from 280,000 images



4143 yeast GFP-fusion proteins
RFP labelled cytoplasm
281,724 images
15.5 million cells

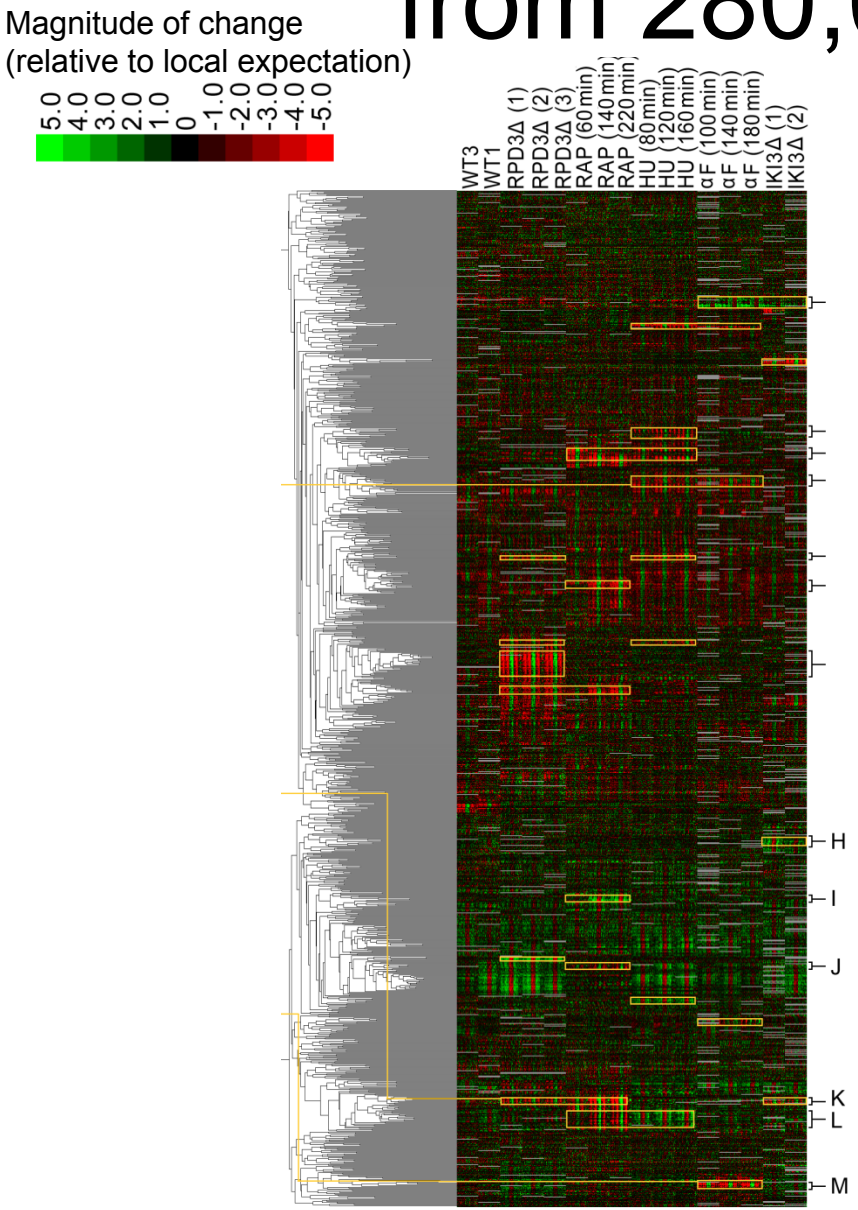


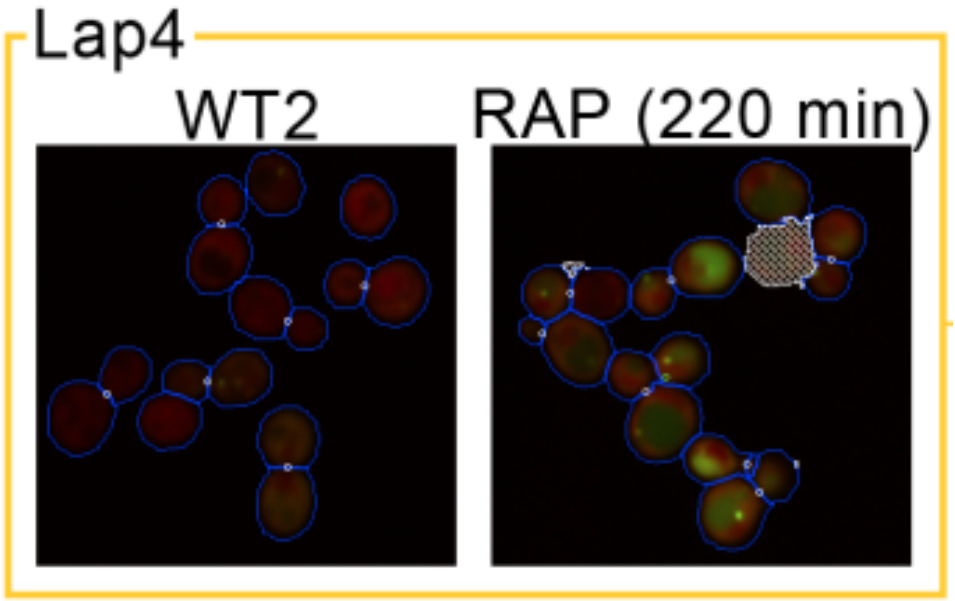
Cyclops database
(Koh et al. *G3* 2015)

Global view of localization changes across many experiments

Lu et al. *eLife 2018*

Patterns of localization change from 280,000 images

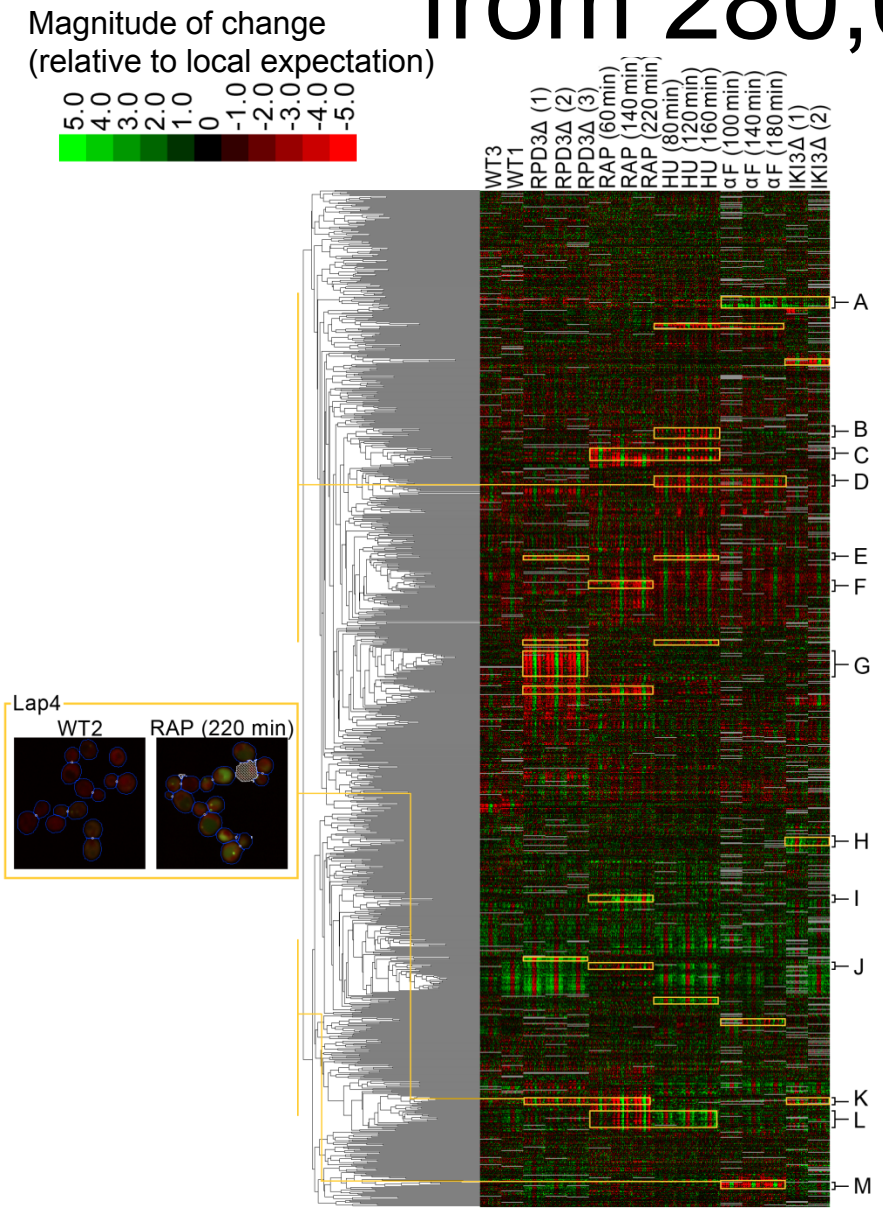


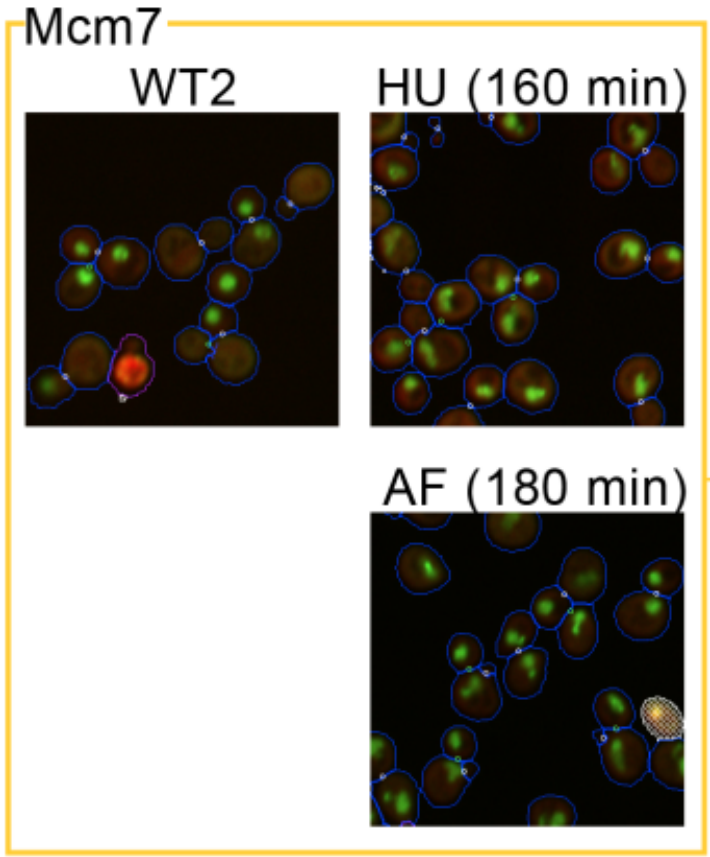


Changes for arbitrary localization classes

Lu et al. *eLife 2018*

Patterns of localization change from 280,000 images

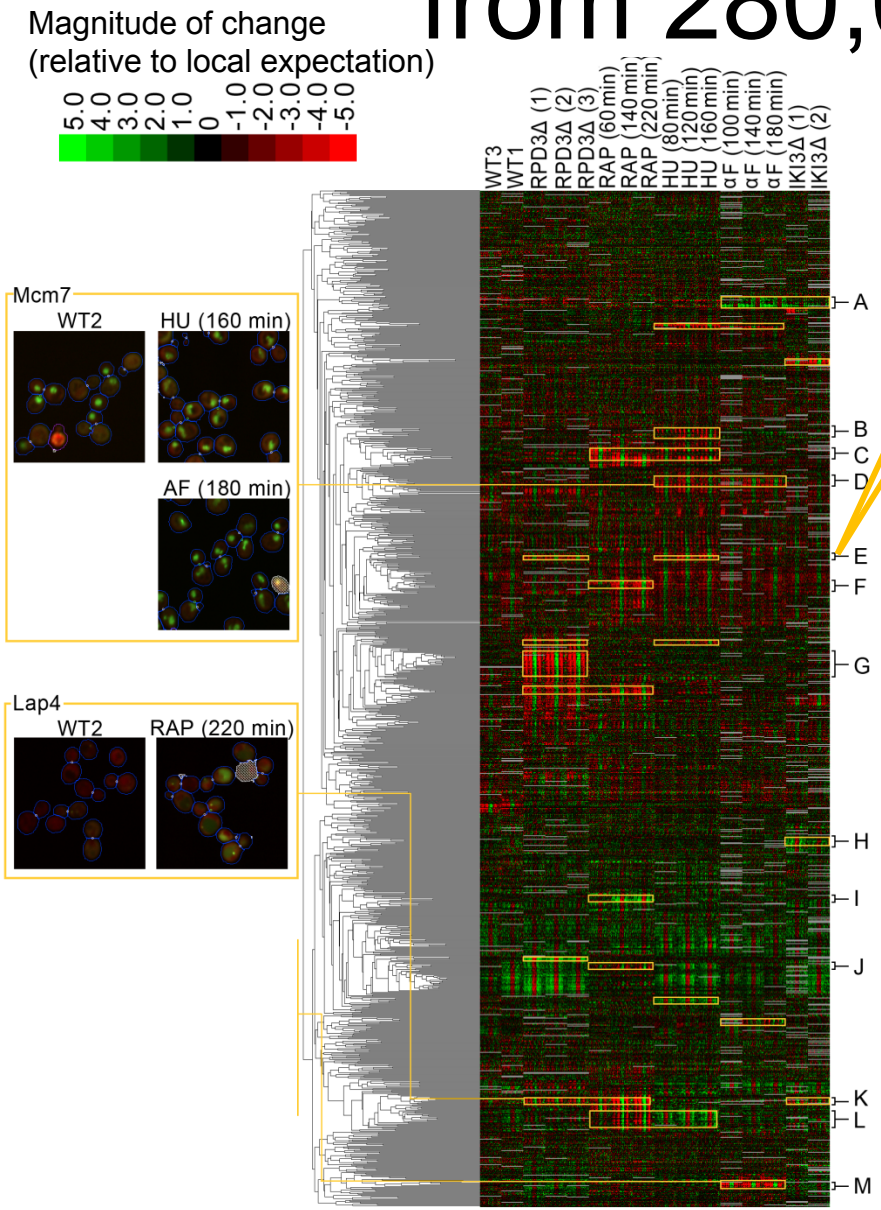


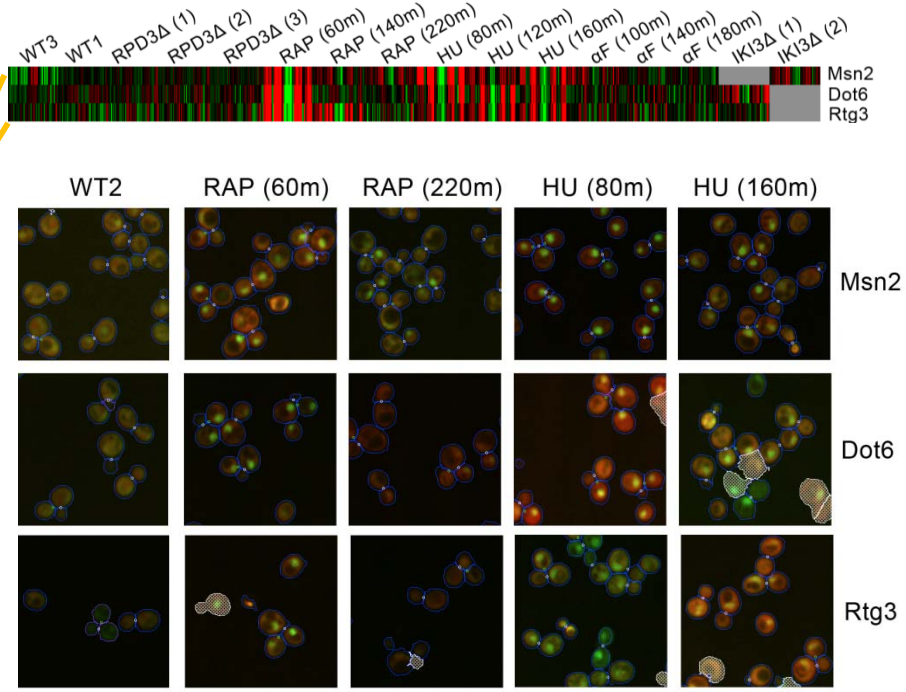


Identify shared changes

Lu et al. *eLife 2018*

Patterns of localization change from 280,000 images





Guilt by association identifies Rtg3 as a "pulsing" stress transcription factor

Lu et al. *eLife 2018*

How to generalize unsupervised automated image analysis?

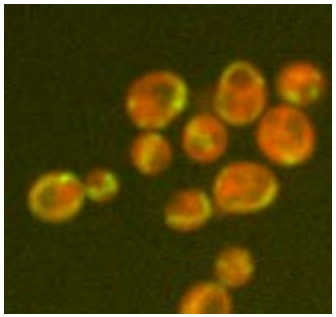
- Data integration across image collections seems to work even with changes in cell morphology

  Lu et al. *eLife 2018*

- But segmentation and image features used were designed for RFP labelled yeast cells (bud and mother, distance to budneck, etc.)
- Can we apply this to other datasets?
- Use generalized segmentation and image features

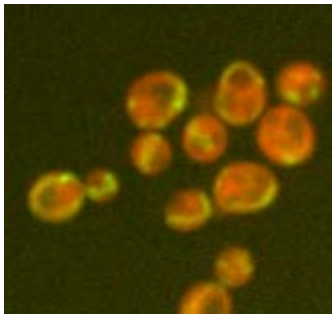
Neural networks

4143 yeast GFP-fusion proteins
RFP labelled cytoplasm
281,724 images
15.5 million cells



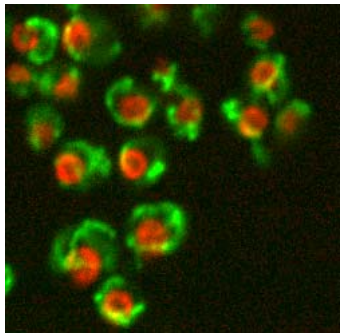
Lu et al. *eLife* 2018
Cyclops database
(Koh et al. *G3* 2015)

4143 yeast GFP-fusion proteins
RFP labelled cytoplasm
281,724 images
15.5 million cells



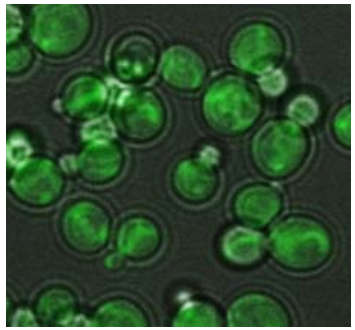
Lu et al. *eLife* 2018
Cyclops database
(Koh et al. *G3* 2015)

4143 yeast GFP-fusion proteins
RFP labelled nuclear pore
~35,000 images
~4 million cells.

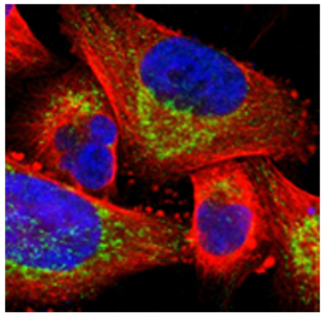


Tkatch et al. *Nat. Cell Bio.* 2012

~4000 yeast GFP-fusion proteins
Bright field background
11895 images
0.566 million cells.



Weill et al. *J. Mol. Biol.* 2018
LoQate database
(Breker et al. *NAR* 2014)



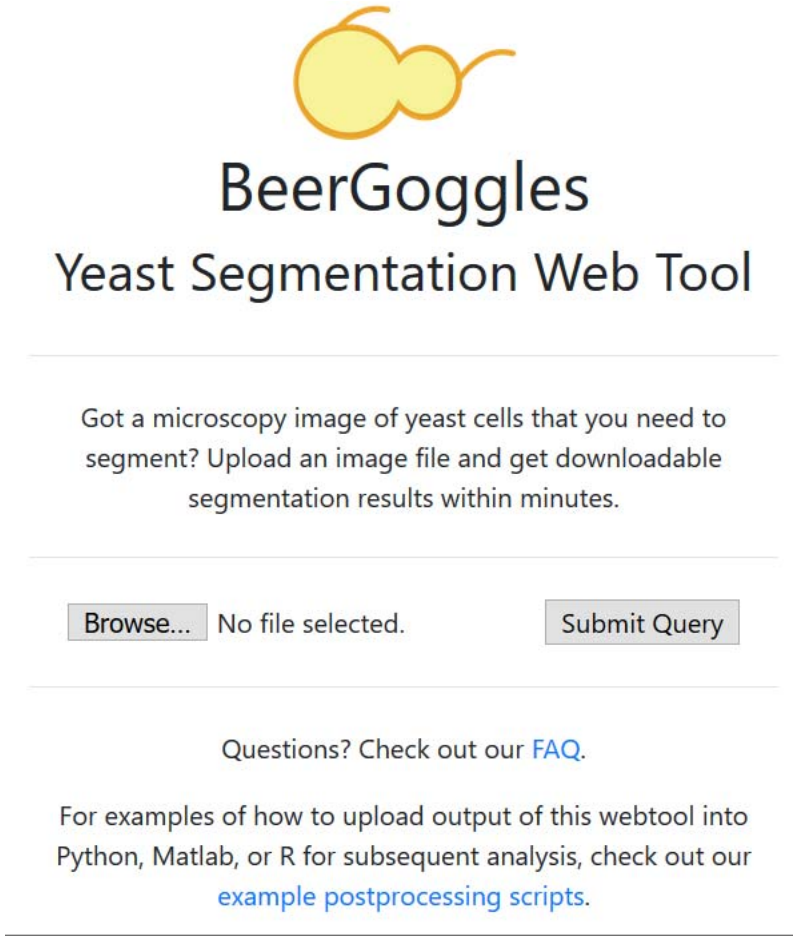
Human Protein Atlas

12,068 Proteins
81,312 Images
638,640
Single Cells

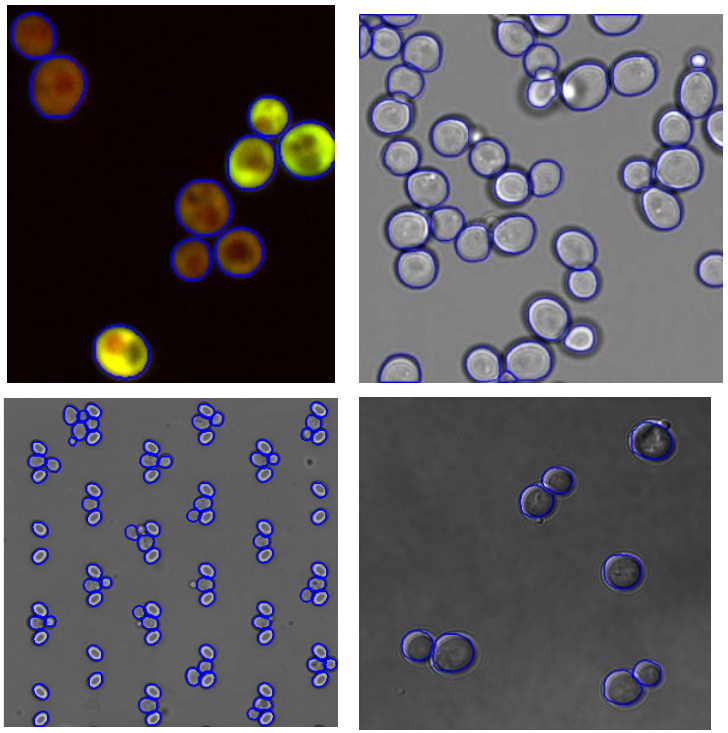
Thul et al. *Science* 2017

And many other datasets that have been analyzed by looking at images one by one

General cell segmentation tools



http://beergoggles.csb.utoronto.ca/



Mask R-CNN trained on human nuclei easily segments yeast across imaging modalities

Lu & Moses *submitted* 2019

How to get morphology and microscopy independent features?

- Our “designed” features relied on prior knowledge of cell morphology and RFP
- CNNs are known to produce good features
- CNNs can work very well for microscopy image classification Kraus et al. *MSB 2017*
  - Supervised classification, so features are likely to be most sensitive to what we trained on
  - Impractical to label training sets for each cellular perturbation

- Self-supervised learning: train the CNN on a “proxy” task
  - Teach the model indirectly by exploiting the structure of the data

Context Encoders: Feature Learning by Inpainting

Deepak Pathak Philipp Krähenbühl Jeff Donahue Trevor Darrell Alexei A. Efros
University of California, Berkeley
{pathak,philkr,jdonahue,trevor,efros}@cs.berkeley.edu

Abstract

We present an unsupervised visual feature learning algorithm driven by context-based pixel prediction. By analogy with auto-encoders, we propose Context Encoders – a convolutional neural network trained to generate the contents of an arbitrary image region conditioned on its surroundings. In order to succeed at this task, context encoders need to both understand the content of the entire image, as well as produce a plausible hypothesis for the missing part(s). When training context encoders, we have experimented with both a standard pixel-wise reconstruction loss, as well as a reconstruction plus an adversarial loss. The latter produces much sharper results because it can better handle multiple modes in the output. We found that a context encoder learns a representation that captures not just appearance but also the semantics of visual structures. We quantitatively demonstrate the effectiveness of our learned features for CNN pre-training on classification, detection, and segmentation tasks. Furthermore, context encoders can be used for semantic inpainting tasks, either stand-alone or as initialization for non-parametric methods.

(a) Input context (b) Human artist

(c) Context Encoder ($L2$ loss) (d) Context Encoder ($L2$ + Adversarial loss)

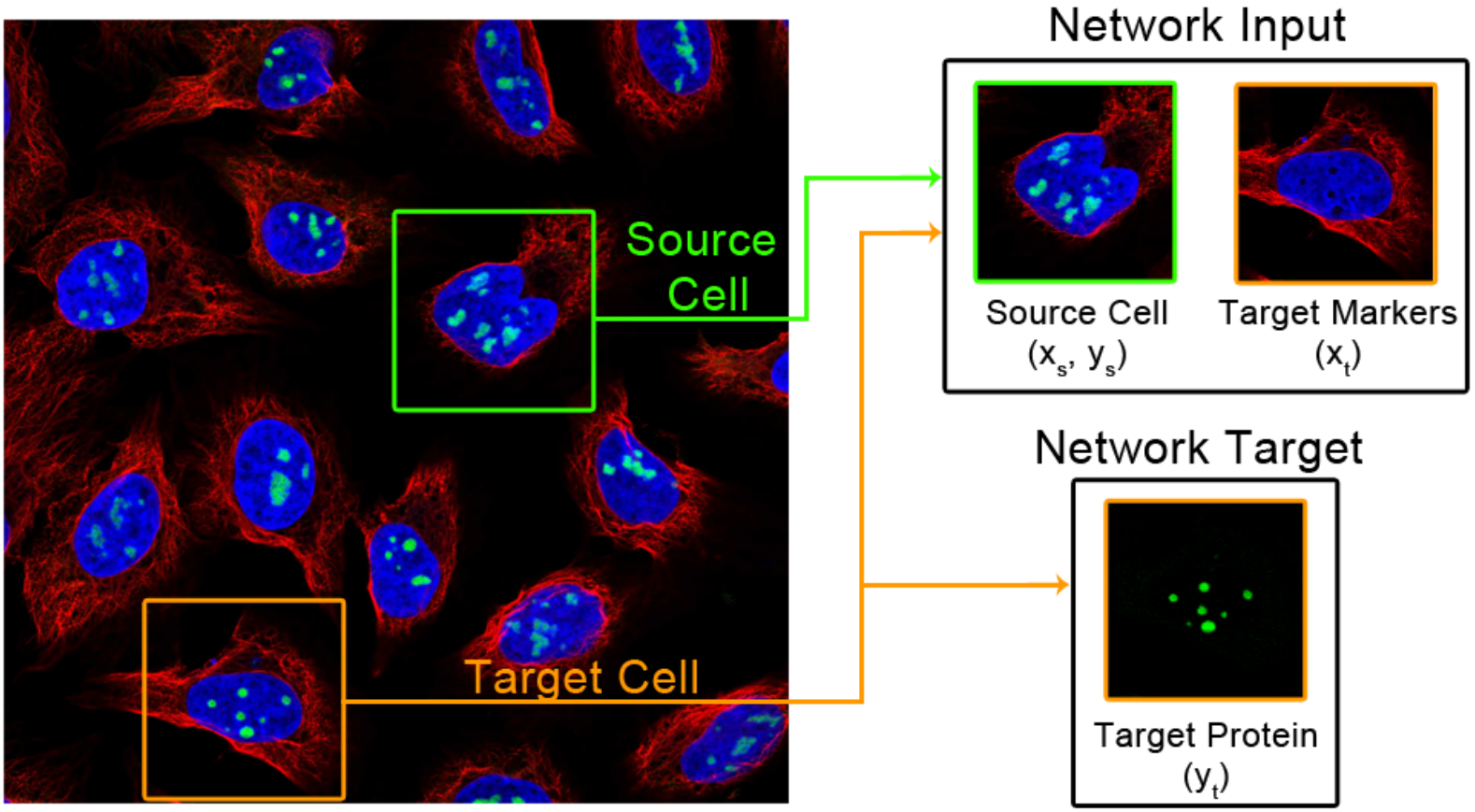
Paired-cell inpainting

Cold Spring Harbor Laboratory
bioRχiv
THE PREPRINT SERVER FOR BIOLOGY
HOME | CHA
Search

New Results
Comment on this paper

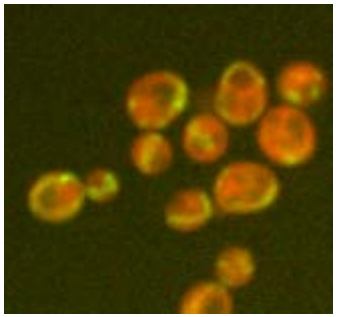
Learning unsupervised feature representations for single cell microscopy images with paired cell inpainting

Alex Lu, Oren Z Kraus, Sam Cooper, Alan M Moses
doi: https://doi.org/10.1101/395954
This article is a preprint and has not been peer-reviewed [what does this mean?].

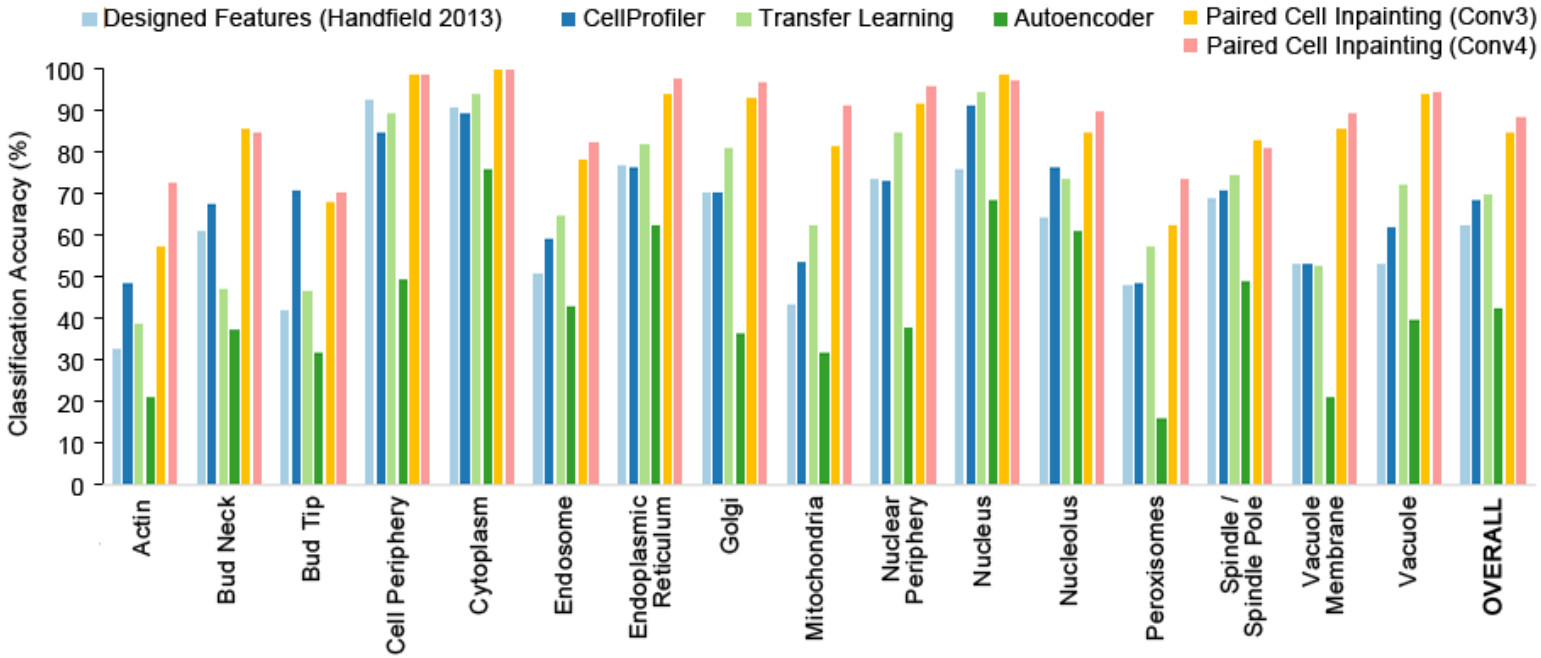


Lu et al. *submitted* 2018

Outperforms other feature sets

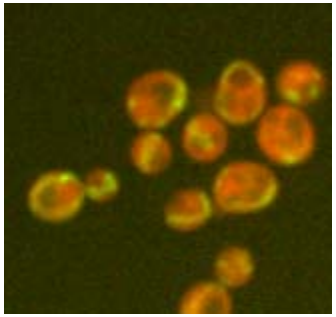


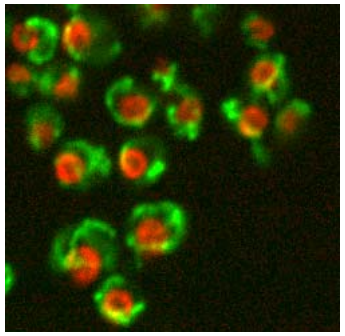
Single cell classification benchmark
(Kraus et al. *MSB* 2017)

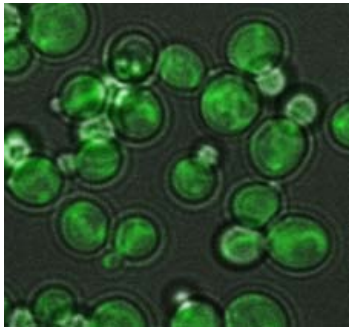


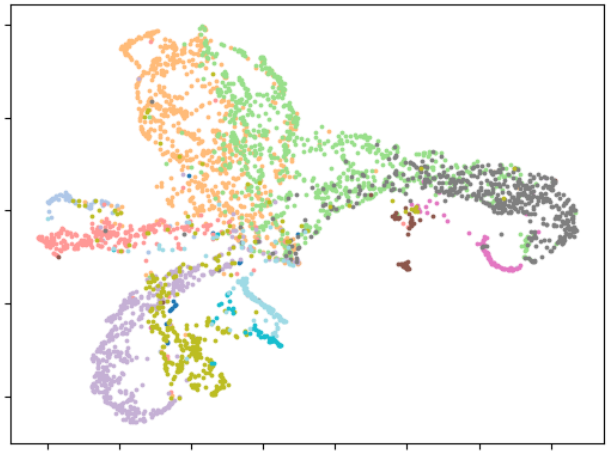
Overall accuracy approaches supervised CNN

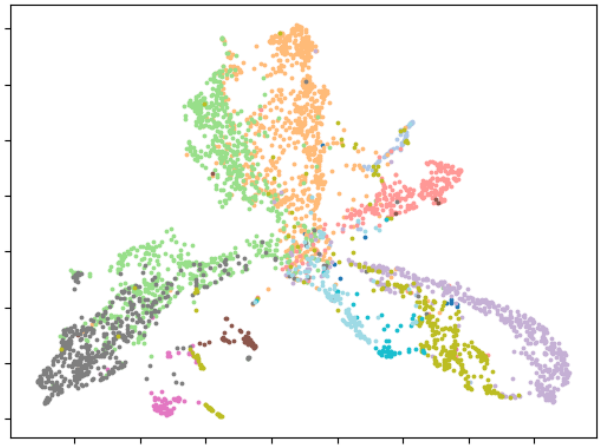
Lu et al. *submitted* 2018

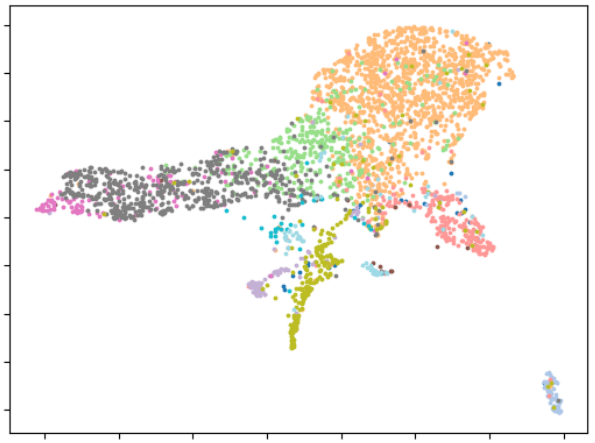




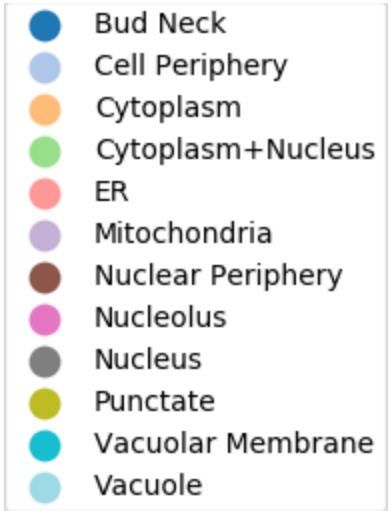








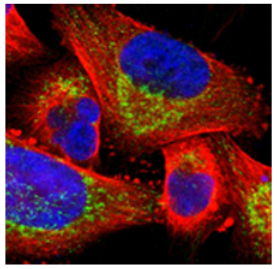
Now straightforward to learn a feature space for each new dataset
No training data is required!
(First objective assessment of consistency between these datasets)



Proteome-scale pattern of localization is recovered from all these datasets

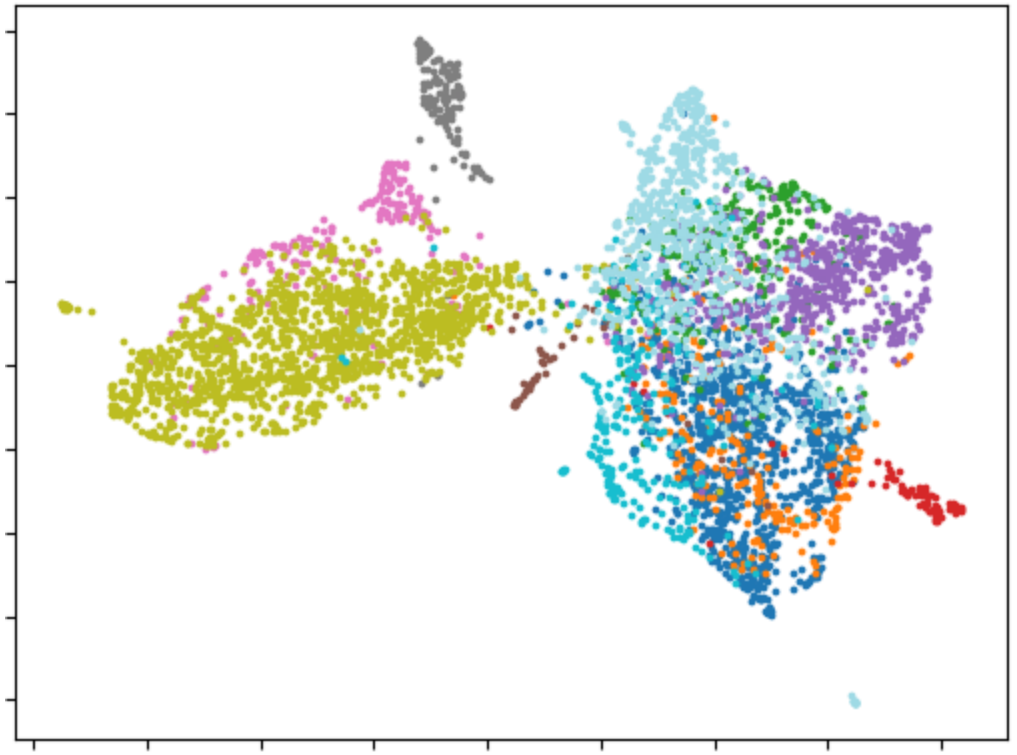
Alex Lu *unpublished*

Rediscovers most known human patterns



Human Protein Atlas

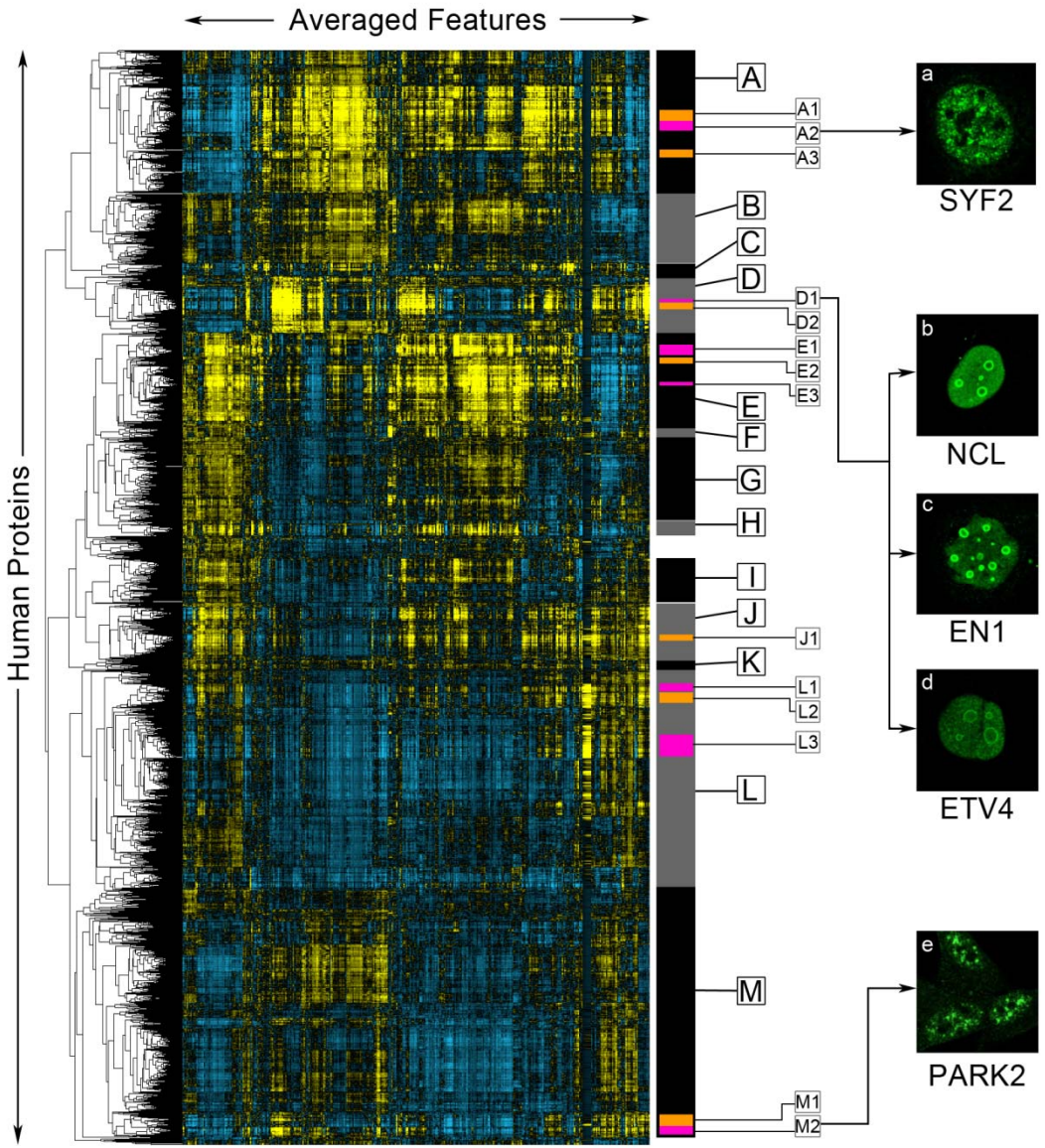
12,068 Proteins
81,312 Images
638,640
Single Cells



- Cytosol
- Endoplasmic reticulum
- Golgi apparatus
- Microtubules
- Mitochondria
- Nuclear membrane
- Nuclear speckles
- Nucleoli
- Nucleoplasm
- Plasma membrane
- Vesicles

Alex Lu *unpublished*

Discover rare, difficult patterns



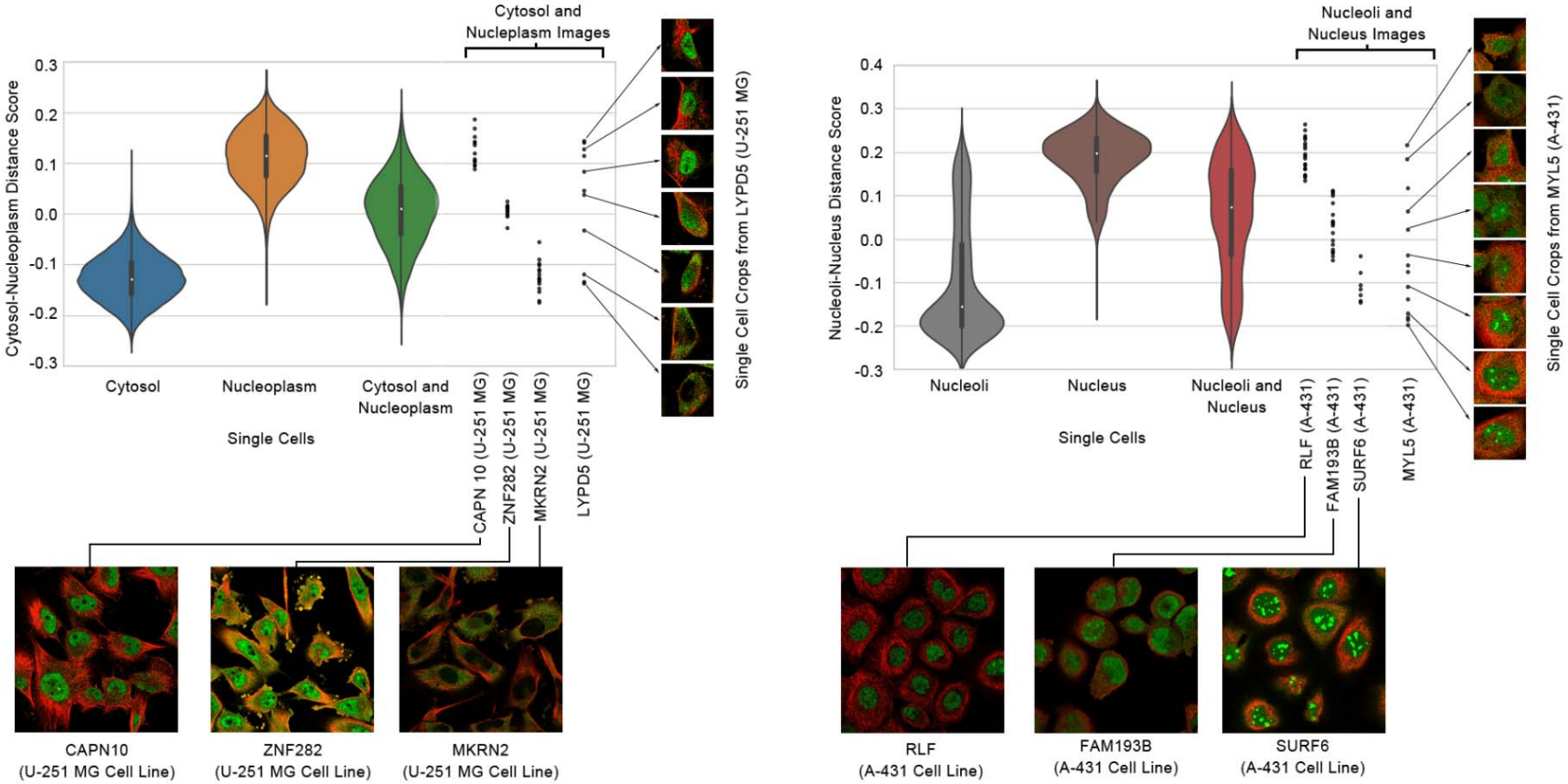
Splicing
"Speckles"

"Nucleolar rim": very rare pattern with no training data

Other
"Speckles"

Lu et al. *submitted* 2018

Features seem highly robust to morphological variation

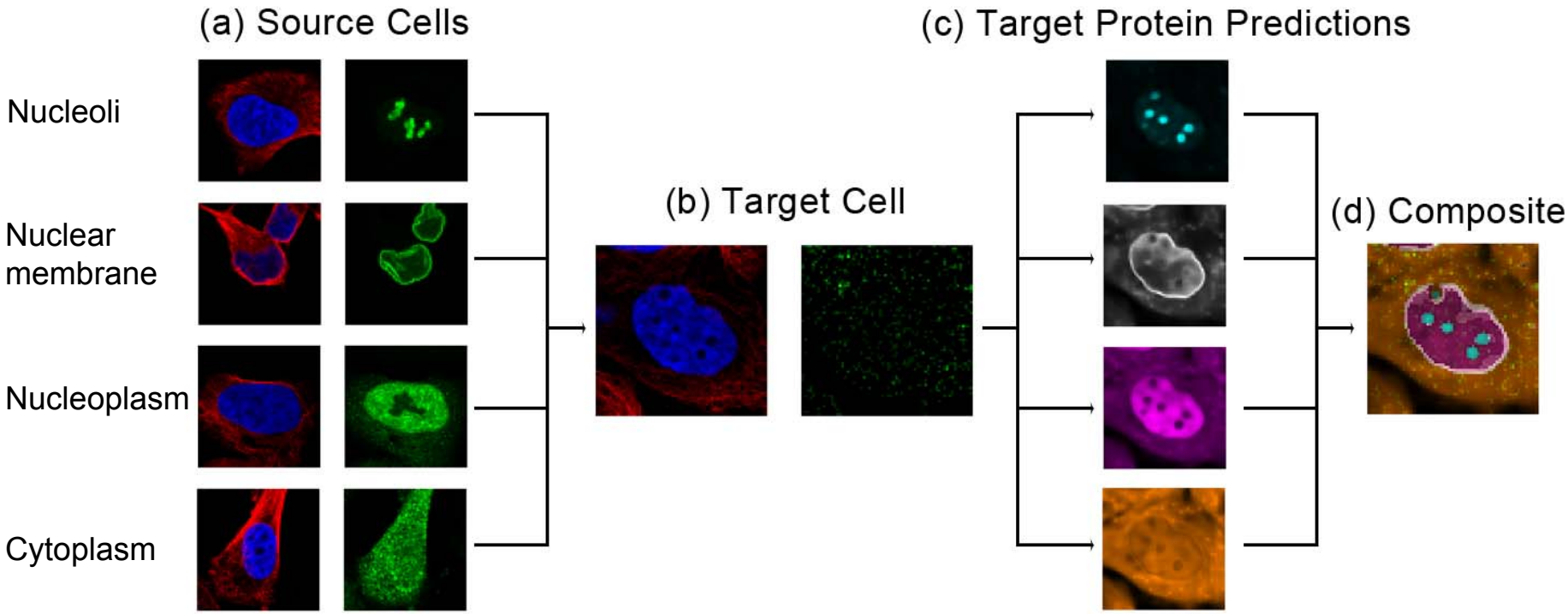


Lu et al. *submitted* 2018

Paired-Cell Inpainting

- For any multi-channel cell microscopy dataset, we can now generate a single cell feature representation ***without training data***
- Change detection and other applications:
  - Quantify cell to cell variation
  - Identify cell-type-specific localization patterns
  - Integrate data across imaging modalities
- What about the generative capacity of the model?

Applications approach science fiction



"Inpaint" more markers to highlight cell components?

Outline


@alexijielu

- Introduction: regulation of proteins
- Automatic identification of protein localization changes in microscopy images
  - Local statistics & data integration
    Lu & Moses *PLoS One 2016*
  - Patterns of localization change to cell biology
    Lu et al. *eLife 2018*
  - Self-supervised learning of image features
    Lu et al. *submitted* 2018
- Unsupervised classification of intrinsically disordered protein regions

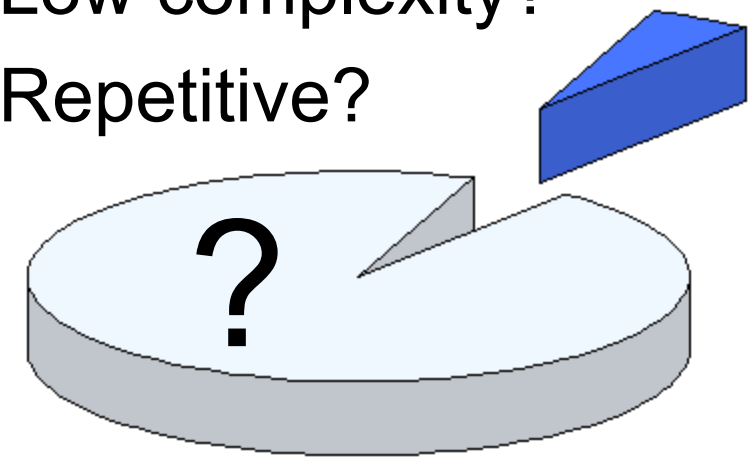
Outline

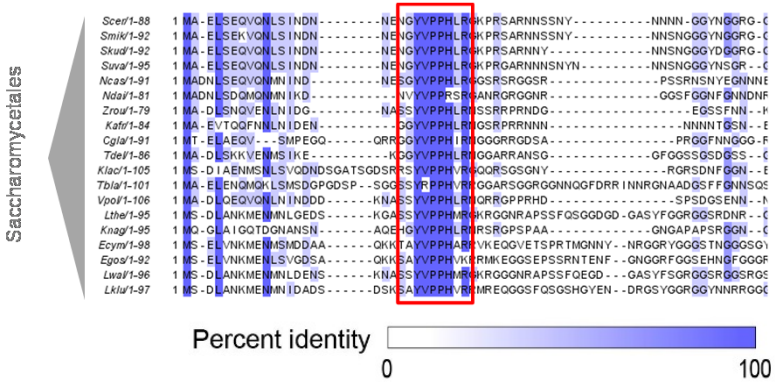
- Introduction: regulation of proteins
- Automatic identification of protein localization changes in microscopy images
- Unsupervised classification of intrinsically disordered protein regions

@taraneh_z

Use evolutionary comparisons to understand regulation of proteins

- Intrinsically disordered regions (IDRs) are mostly highly diverged
- ~1/3 known motifs are conserved, only 5% of total IDR amino acids are conserved in sequence alignments
- Low complexity?
- Repetitive?





Nguyen Ba et al. *Sci. Signaling* 2012

Are IDRs mostly “junk”?

- Yes (Norman Davey, personal communication)
  - IDRs are just inert “linkers” to hold motifs
  - Many can be greatly shortened

OPEN ACCESS Freely available online

PLOS | BIOLOGY

The Robustness of a Signaling Complex to Domain Rearrangements Facilitates Network Evolution

Paloma M. Sato, Kogulan Yoganathan, Jae H. Jung, Sergio G. Peisajovich*

- No way! (Julie Forman-Kay)
  - “bulk properties” (such as low-complexity, repeats, charge) are the key to phase separation

THE EMBO JOURNAL

Phosphorylation of the FUS low-complexity domain disrupts phase separation, aggregation, and toxicity

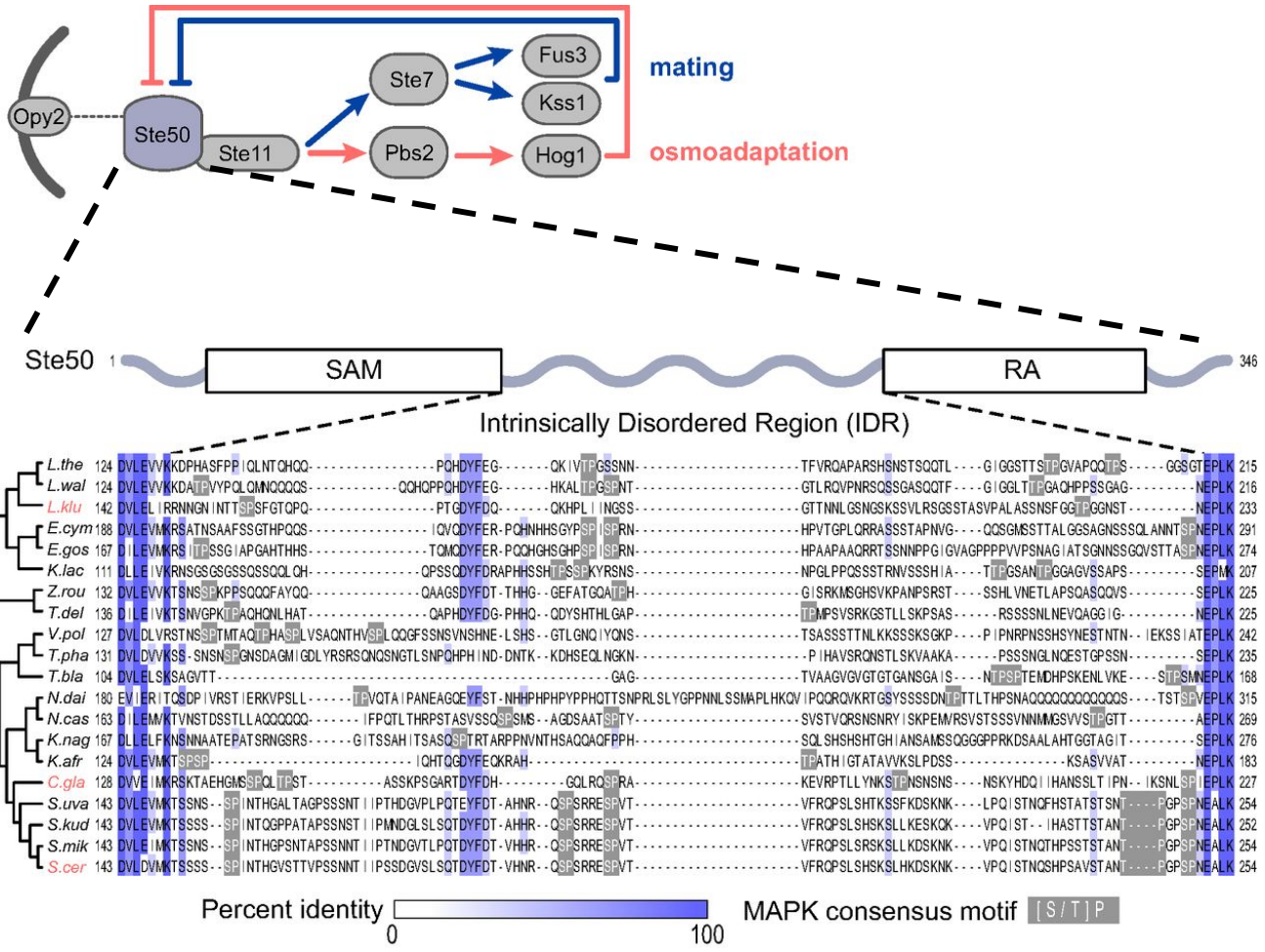
Zachary Monahan[1,†], Veronica H Ryan[2,†], Abigail M Janke[3], Kathleen A Burke[3], Shannon N Rhoads[1], Gül H Zerze[4], Robert O'Meally[5], Gregory L Dignon[4], Alexander E Conicella[6], Wenwei Zheng[7], Robert B Best[7], Robert N Cole[5], Jeetain Mittal[4], Frank Shewmaker[2,*] & Nicolas L Fawzi[2,3,6,**]

- Let’s find out (Taraneh Zarin, PhD Student)

If bulk sequence properties are important they should be preserved by evolution…

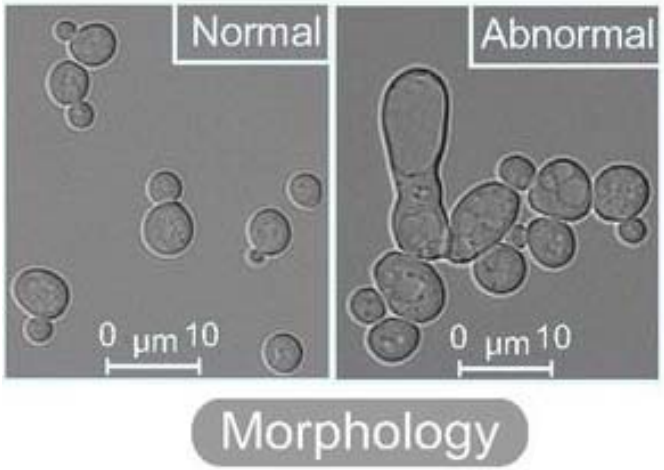
Ste50 as a typical IDR

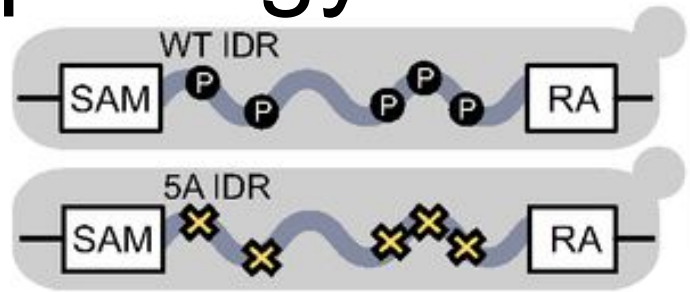
MAPK signaling network

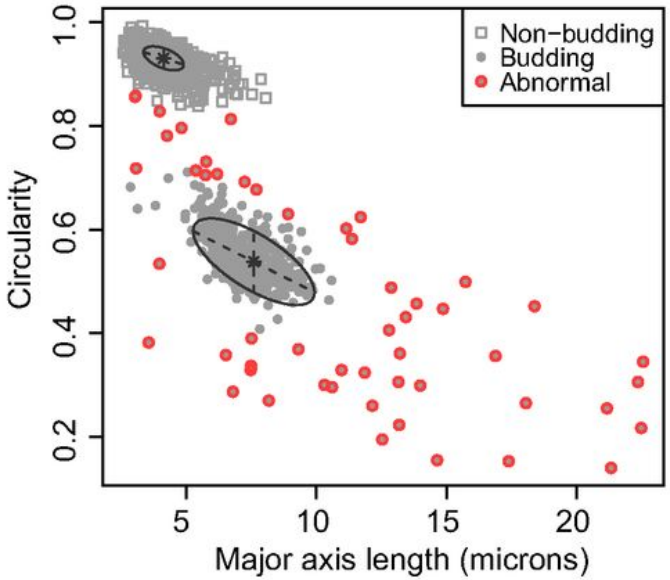


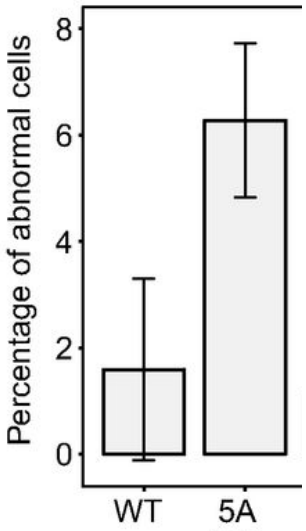
Zarin *et al. PNAS 2017*

Phosphorylation sites are needed for normal morphology



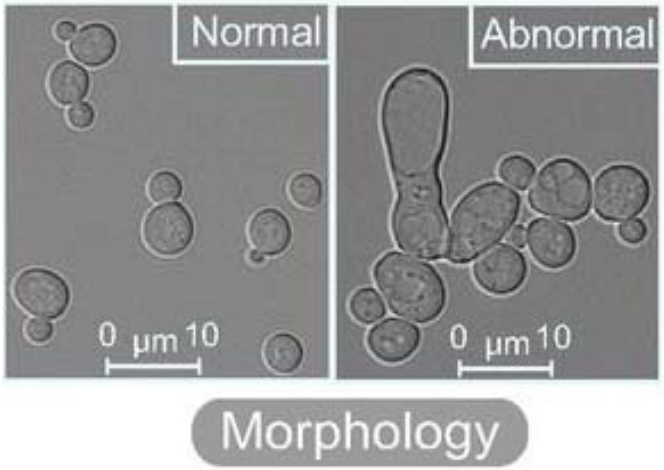


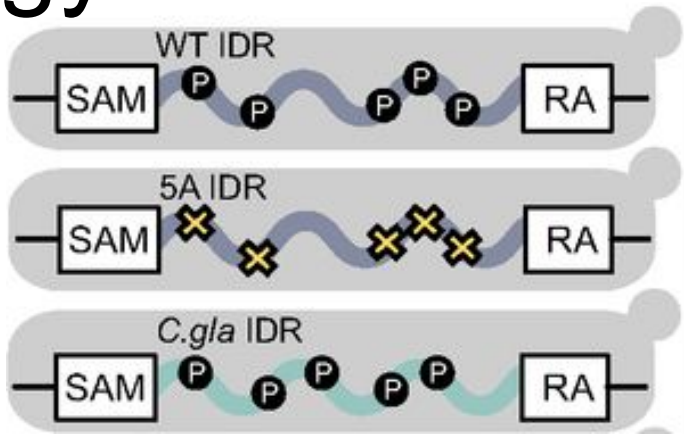


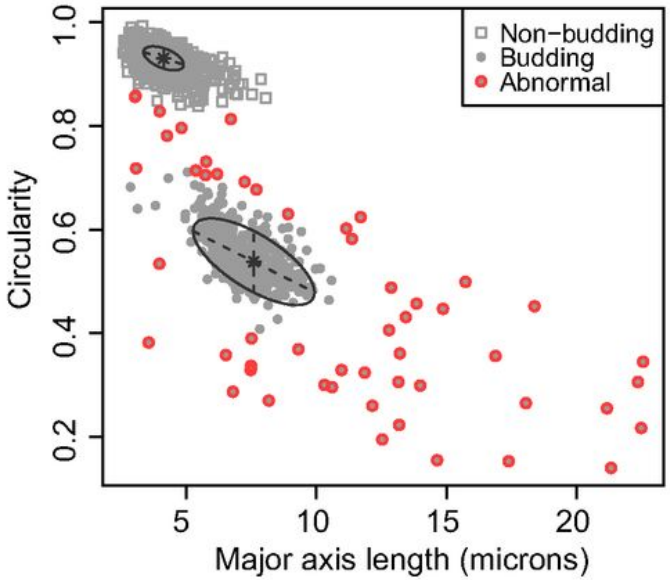


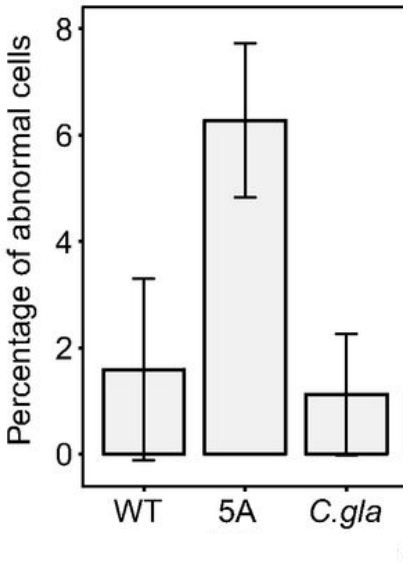
Quantify abnormal morphology using a two-component mixture model

Other species' IDRs rescue morphology



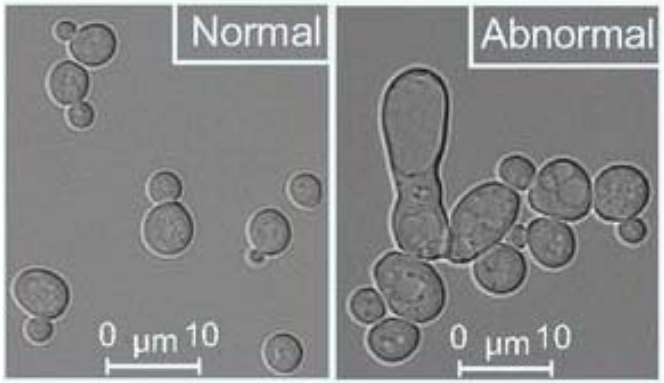


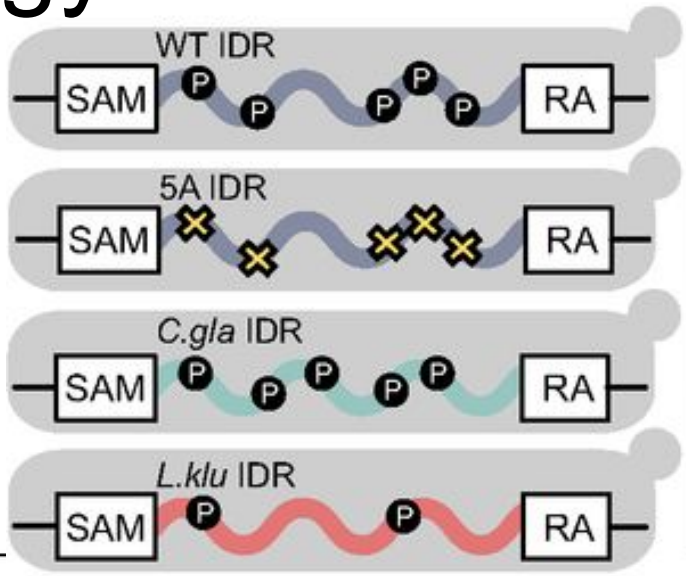


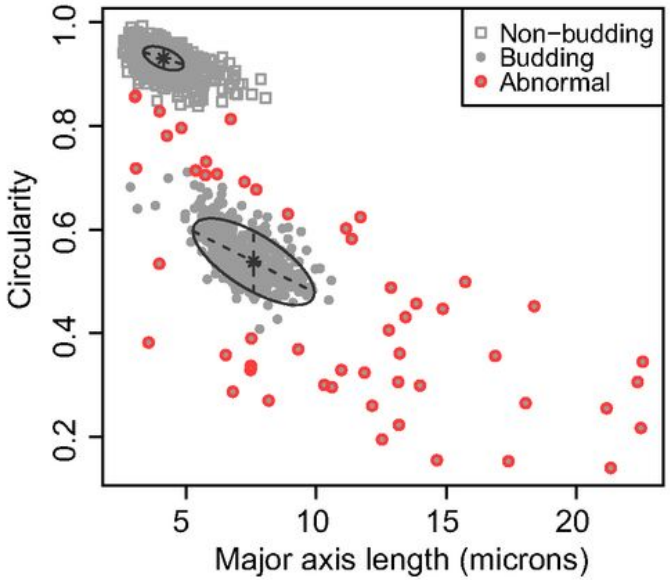


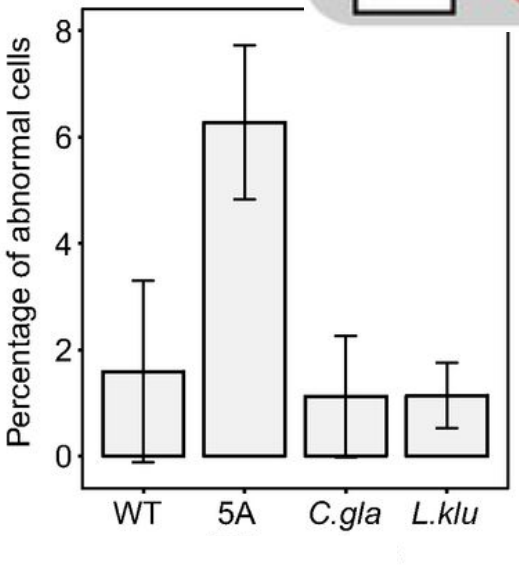
Zarin *et al. PNAS 2017*

Other species' IDRs rescue morphology







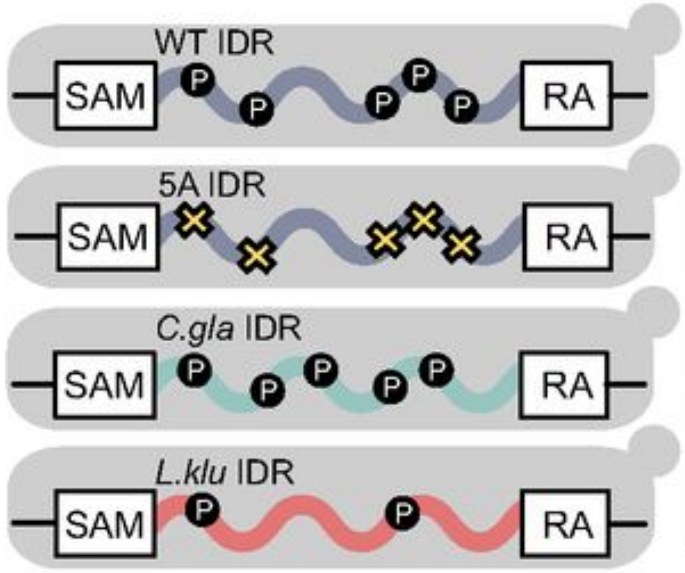


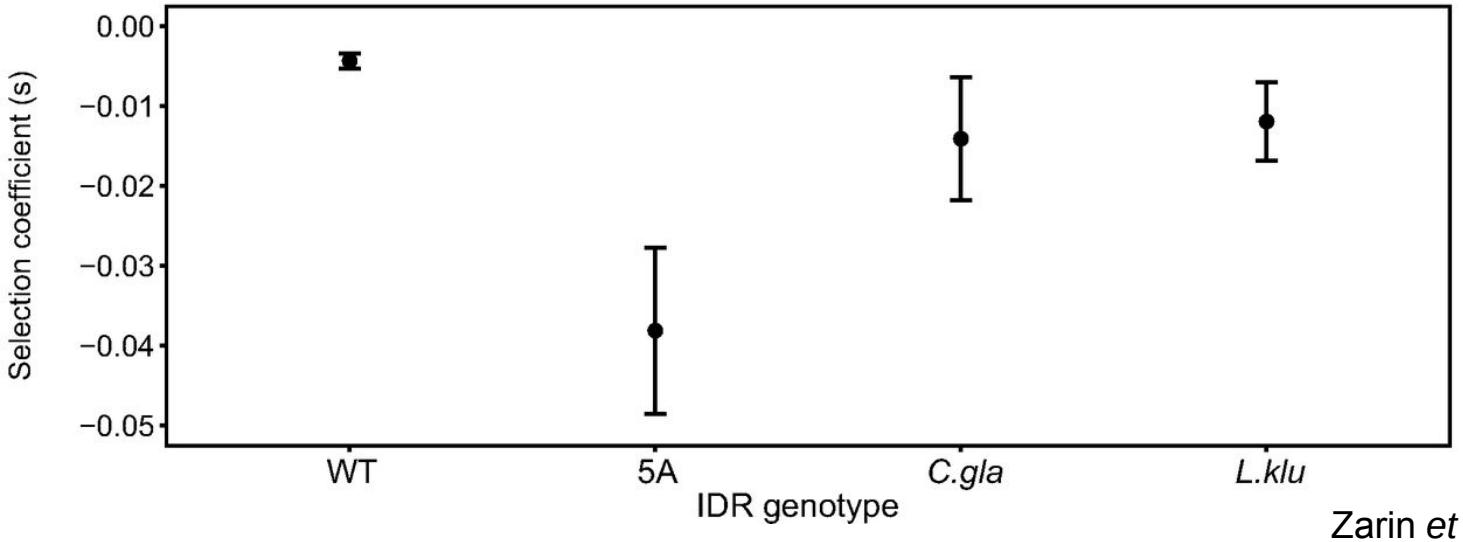
Phosphorylation site number is not correlated with morphology

Zarin *et al. PNAS 2017*

Other species' IDRs support normal growth

Other species IDRs can mostly rescue fitness defect

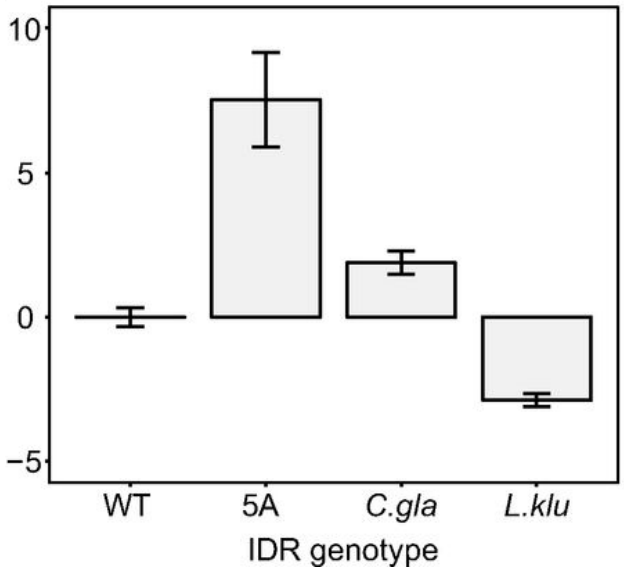


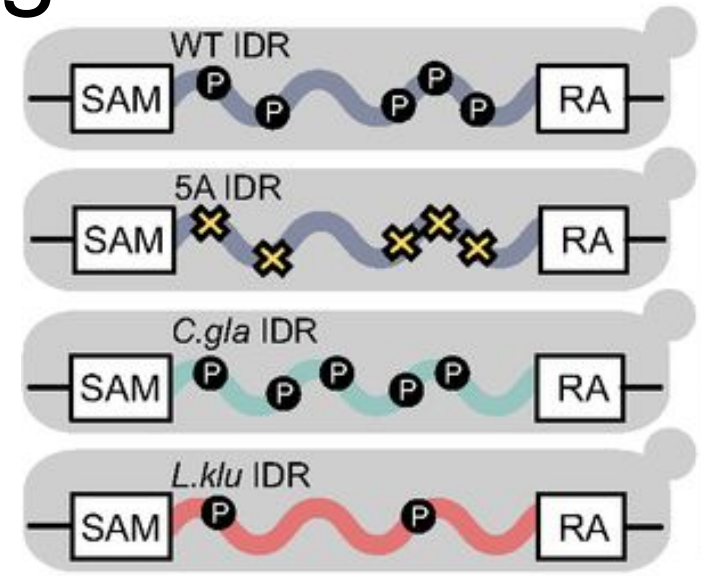


Zarin *et al. PNAS 2017*

Other species' IDRs support normal signaling

Signaling reporter
(flow cytometry!)

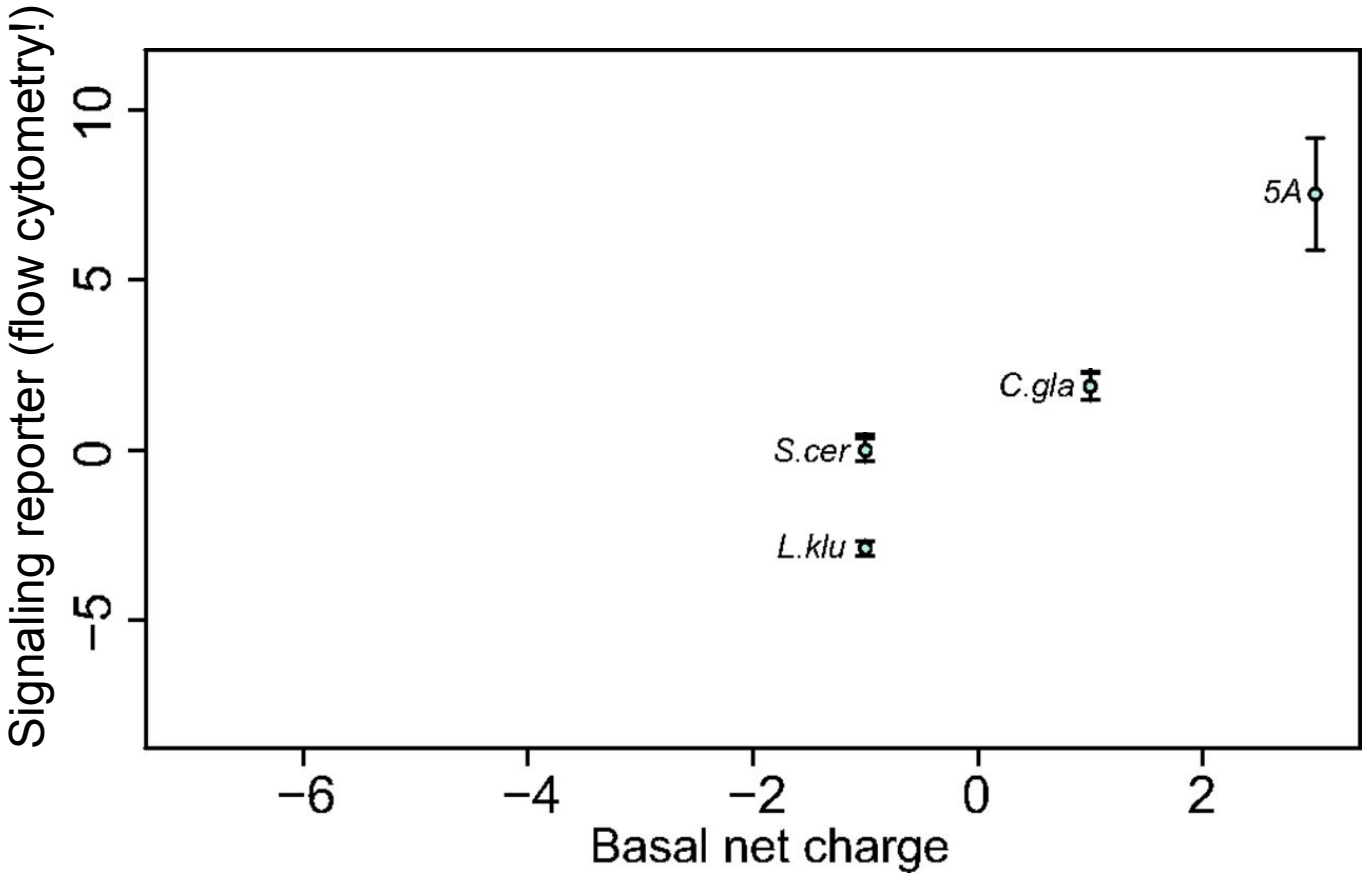




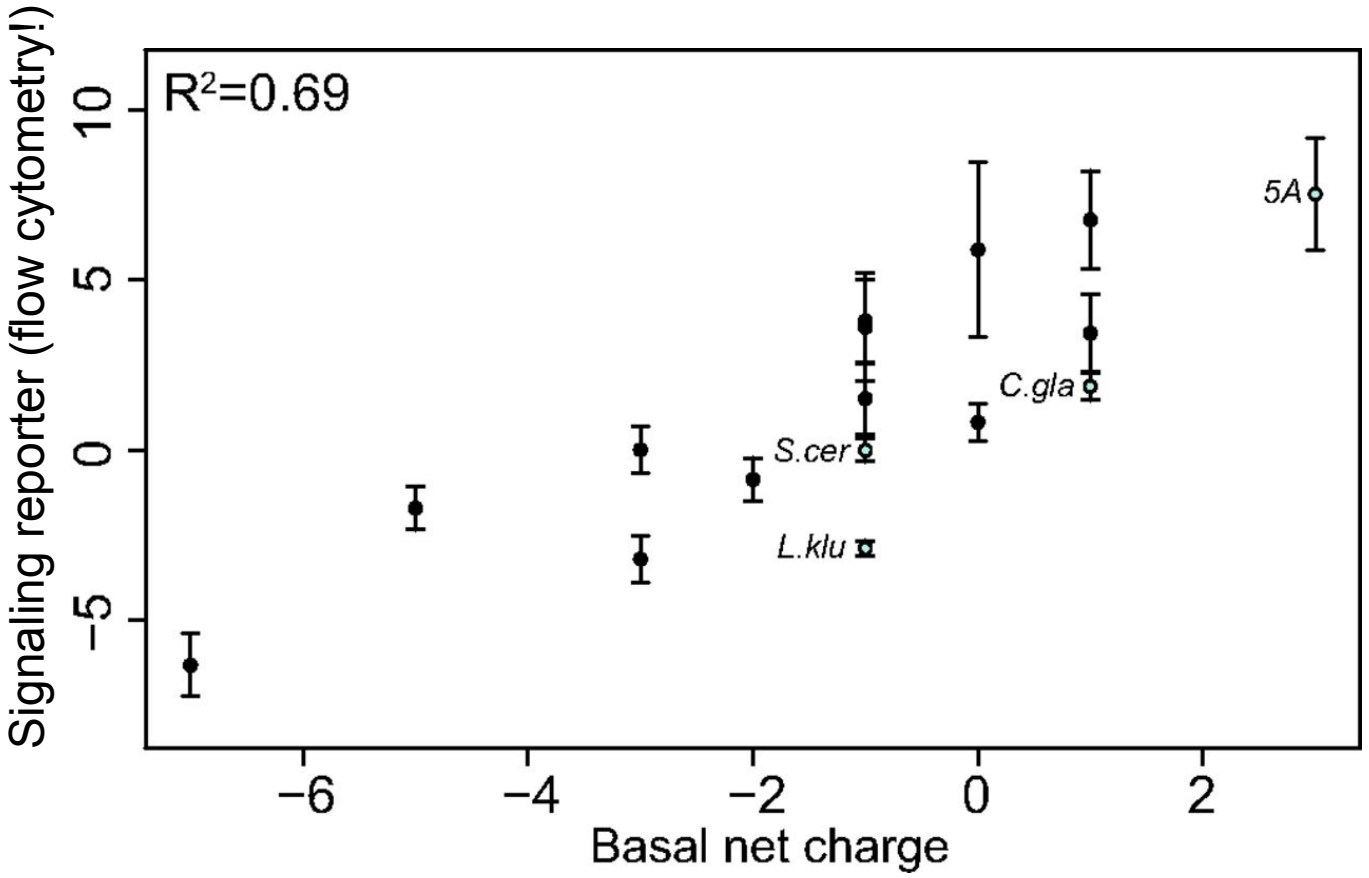
Mutation of phosphorylation sites appears to be a gain of function

Zarin *et al. PNAS 2017*

Correlation of signaling with charge?



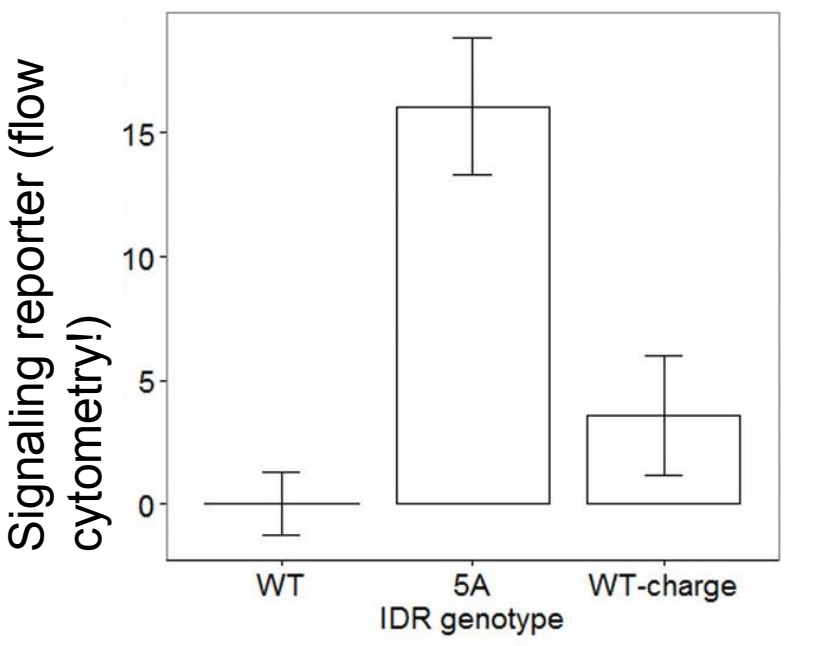
Make lots of mutants and measure reporter expression

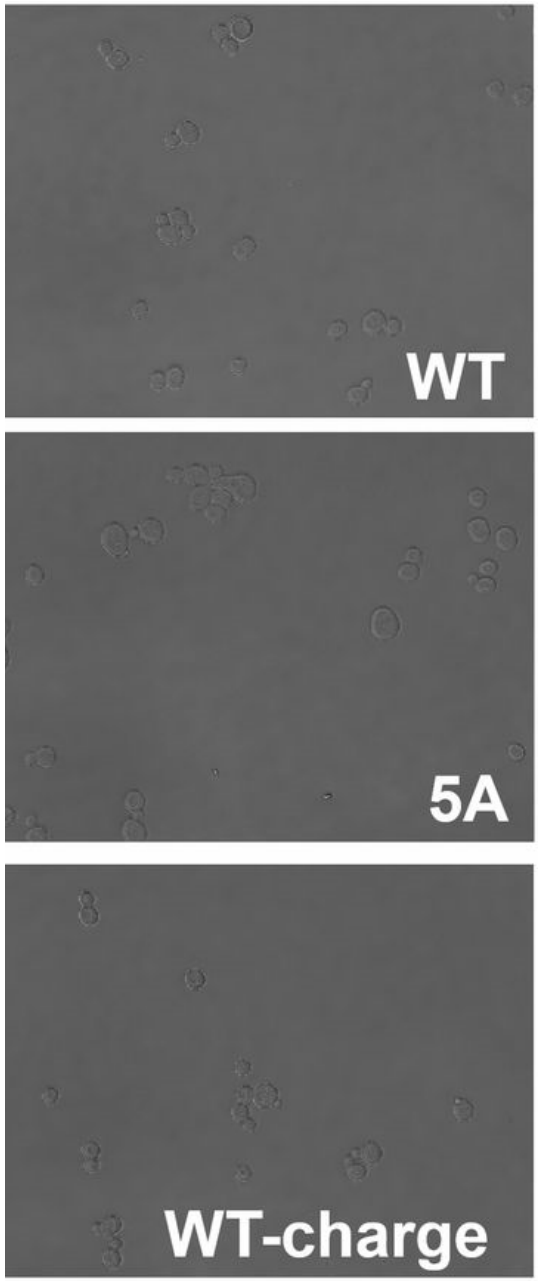


Turns out charge might be what's important…

Zarin *et al. PNAS 2017*

Non-phosphorylatable
WT-charge mutant
SSSAPINTHGVSTTVPSSNNTIIPSSDGVSLSQTD
YFDTVHNRQAPSRREAPVTVFRQPSLSHSKSLH
KDSKNKVPQISTNQSHPSAVSTANEEGPEEN

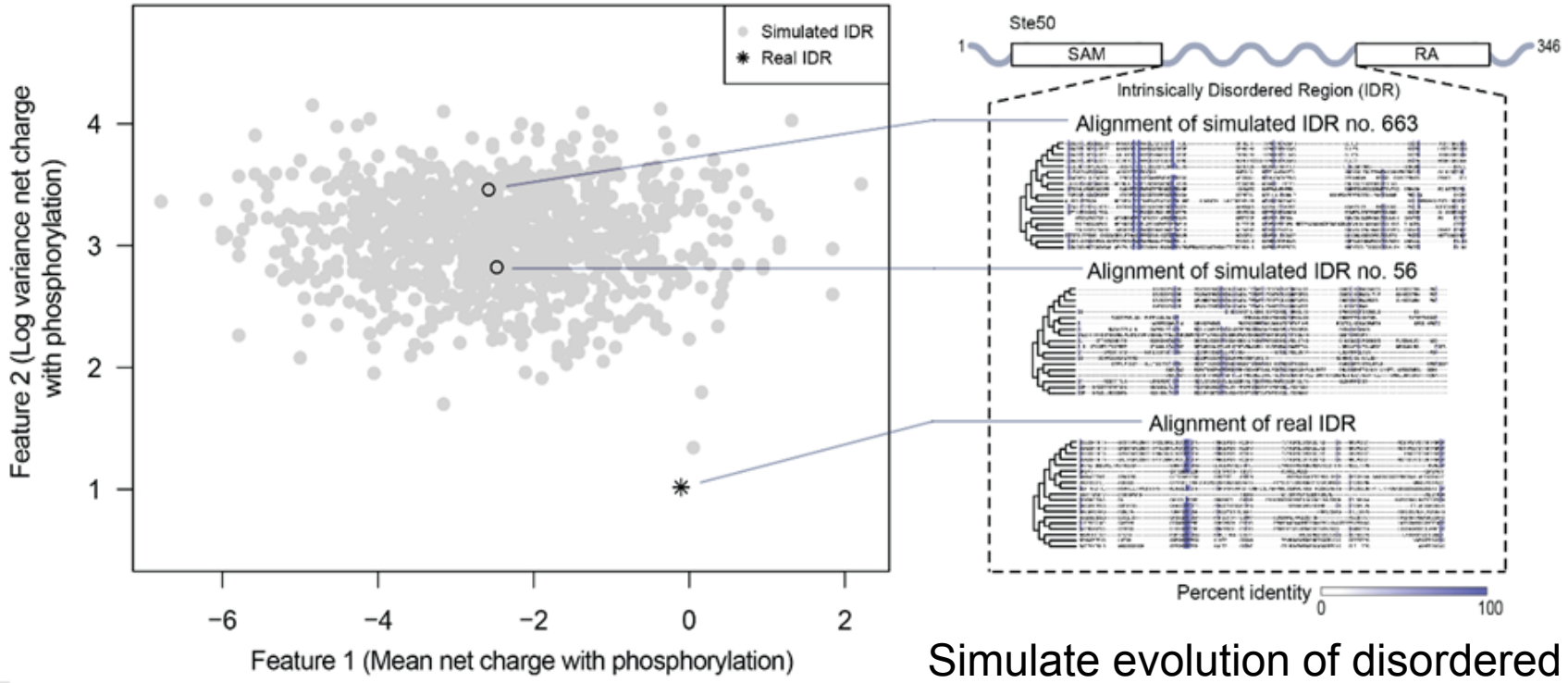




Turns out charge might be what's important…

Zarin *et al. PNAS 2017*

A test for unusual evolution in IDRs



Simulate evolution of disordered regions and compare the charge in the real orthologs to what we see in the simulations

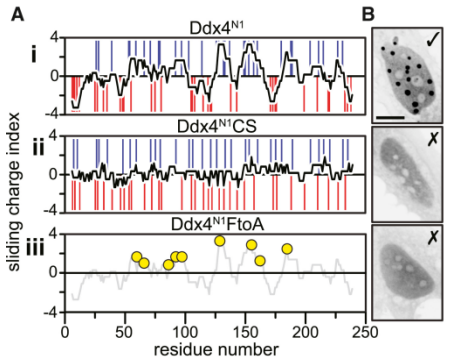
At least in the case of the Ste50 IDR, this points to constraint on charge

Nguyen Ba *et al. PLoS Comp. Bio 2014*

Zarin *et al. PNAS 2017*

Test for other conserved properties that are not visible in sequence alignments

- Recent studies report experimental evidence for functional “bulk” properties in IDRs



Nott *et al. Mol Cell 2015*

Sequence Determinants of Intracellular Phase Separation by Complex Coacervation of a Disordered Protein

Chi W. Pak,[1] Martyna Kosno,[1] Alex S. Holehouse,[2,3] Shae B. Padrick,[1] Anuradha Mittal,[3] Rustam Ali,[1] Ali A. Yunus,[1] David R. Liu,[4] Rohit V. Pappu,[3,*] and Michael K. Rosen[1,*]

JACS JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

Article

pubs.acs.org/JACS

Sequence Determinants of the Conformational Properties of an Intrinsically Disordered Protein Prior to and upon Multisite Phosphorylation

Erik W. Martin,[†,§] Alex S. Holehouse,[‡,§] Christy R. Grace,[†] Alex Hughes,[†] Rohit V. Pappu,[*,‡] and Tanja Mittag[*,†]

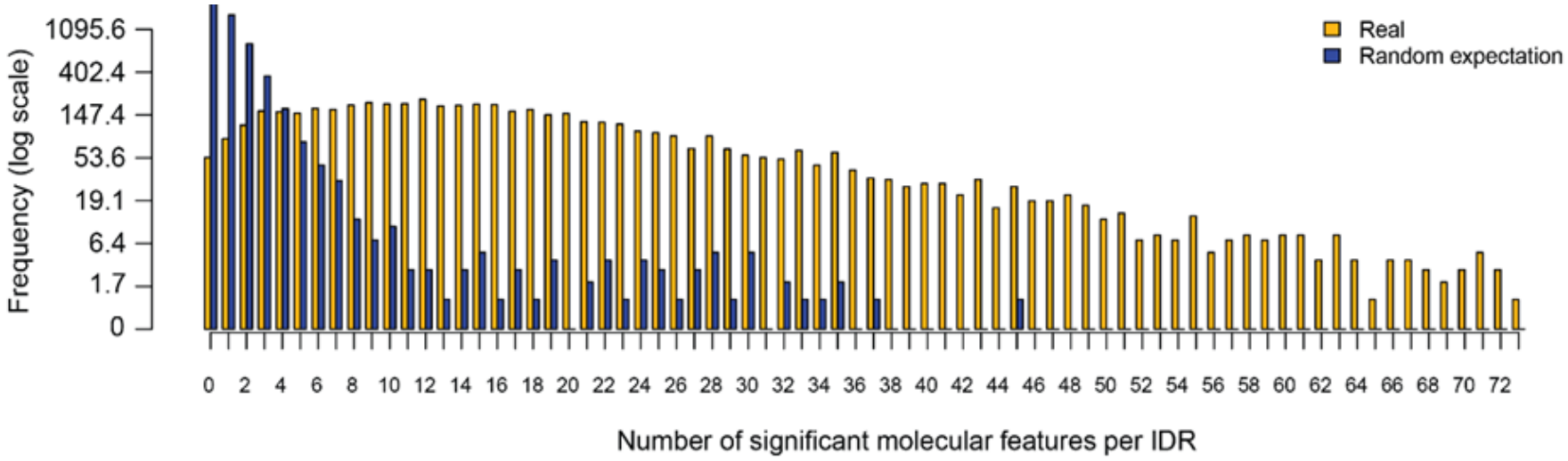
Cryptic sequence features within the disordered protein p27^{Kip1} regulate cell cycle signaling

Rahul K. Das[a,1], Yongqi Huang[b,1], Aaron H. Phillips[b,1], Richard W. Kriwacki[b,c,2], and Rohit V. Pappu[a,2]

[a]Department of Biomedical Engineering and Center for Biological Systems Engineering, Washington University in St. Louis, St. Louis, MO 63130; [b]Department of Structural Biology, St. Jude Children's Research Hospital, Memphis, TN 38105; and [c]Department of Microbiology, Immunology and Biochemistry, University of Tennessee Health Sciences Center, Memphis, TN 38163

Test for other conserved properties that are not visible in sequence alignments

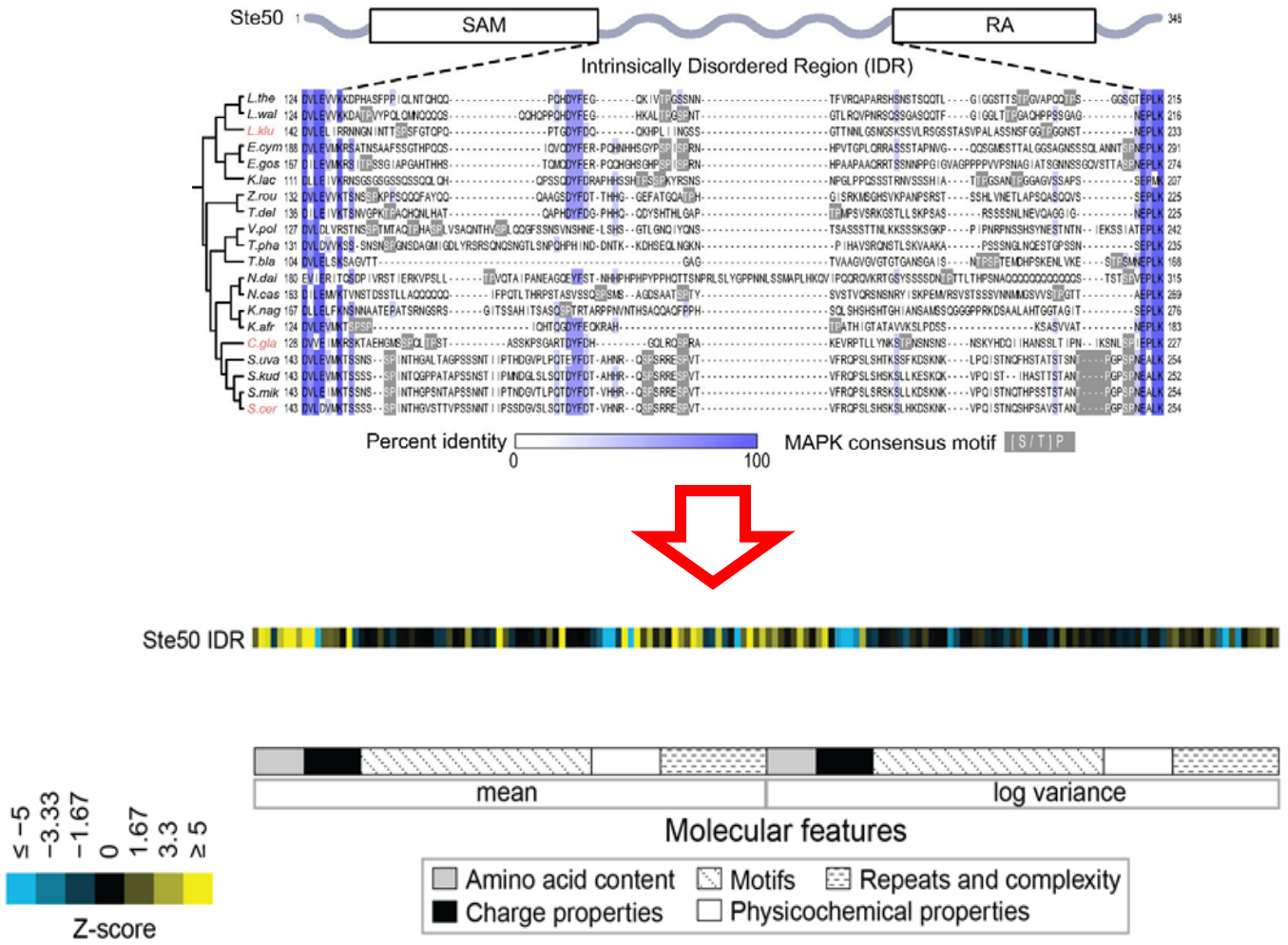
- 82 molecular features from literature that we can compute from protein sequences



Most IDRs in the yeast proteome have many molecular features that deviate from the expectation

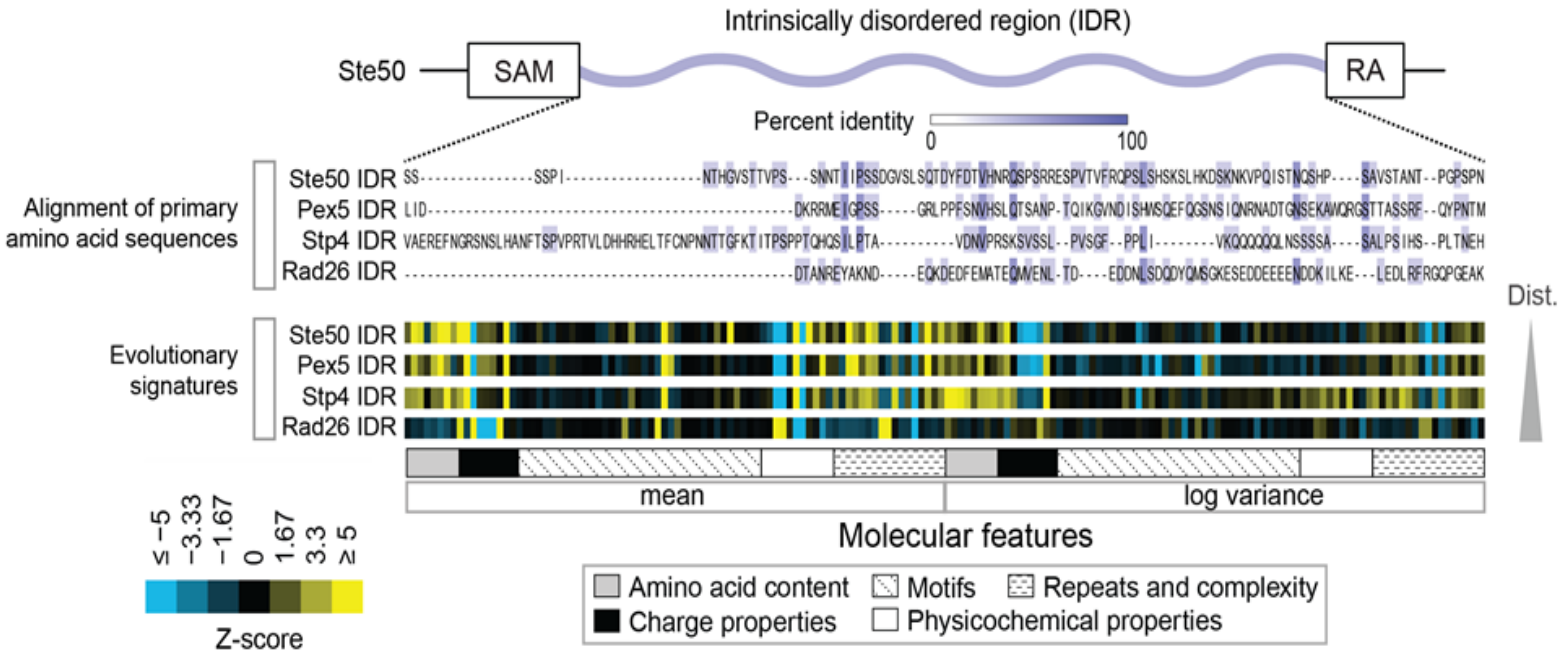
Taraneh Zarin *unpublished*

Evolution of molecular features is a “signature”



Taraneh Zarin *unpublished*

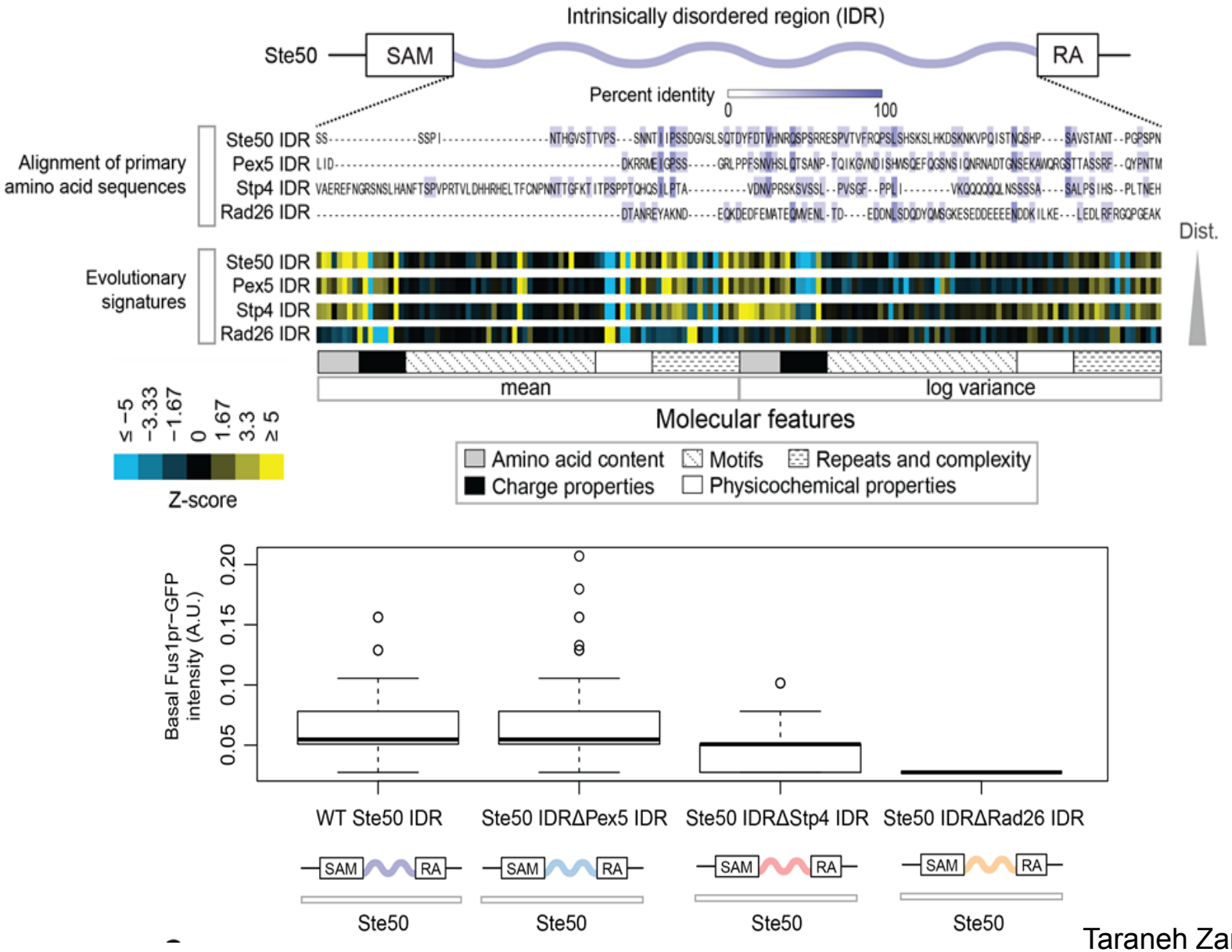
We can quantitatively compare IDRs



Evolutionary signatures can measure similarity of IDR sequences, even when there is no detectable similarity in alignments

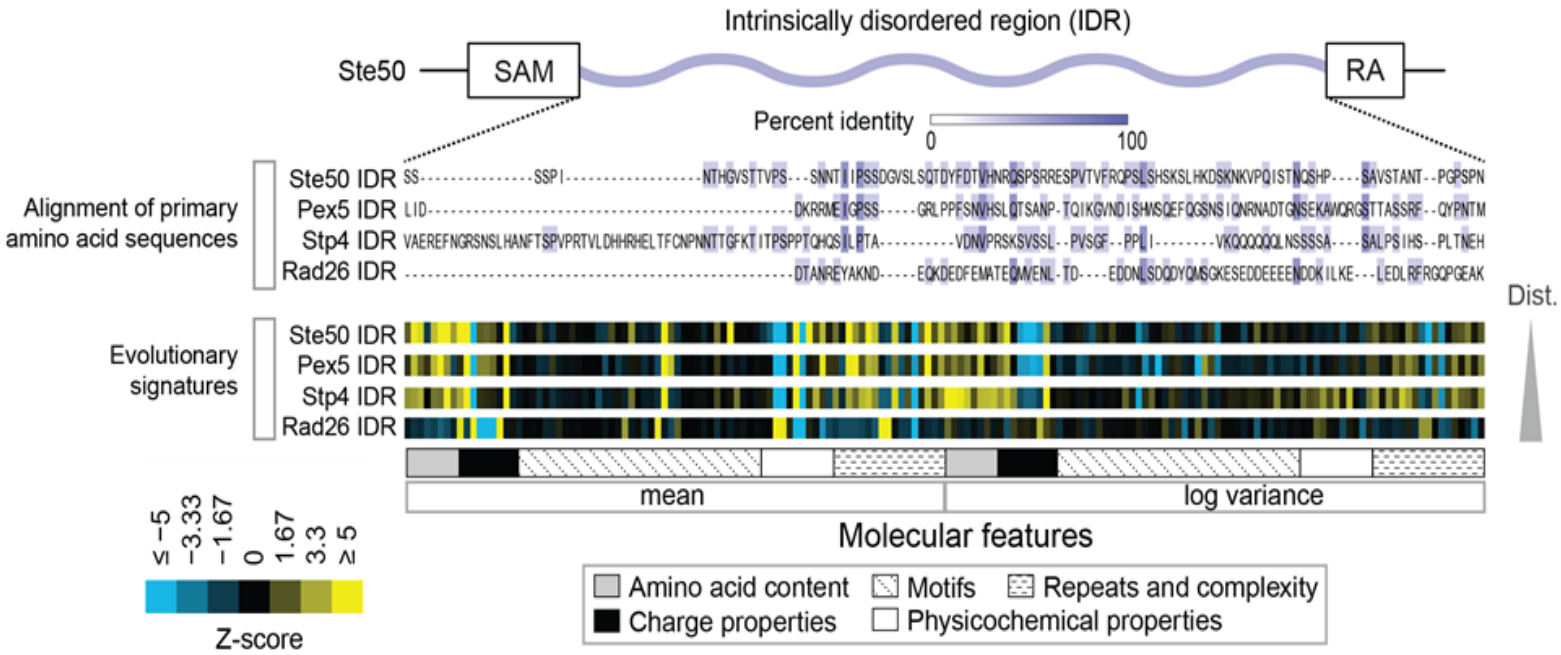
Taraneh Zarin *unpublished*

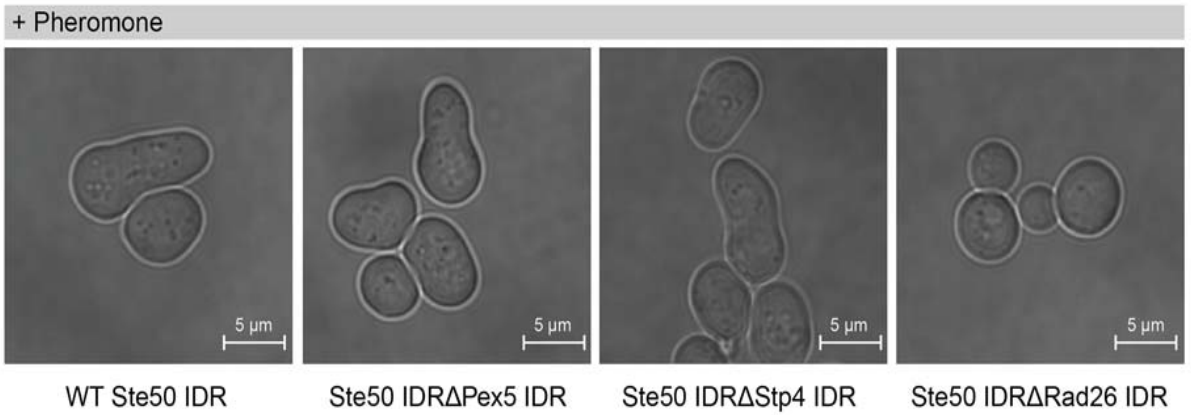
Other proteins' IDRs can rescue signaling!?



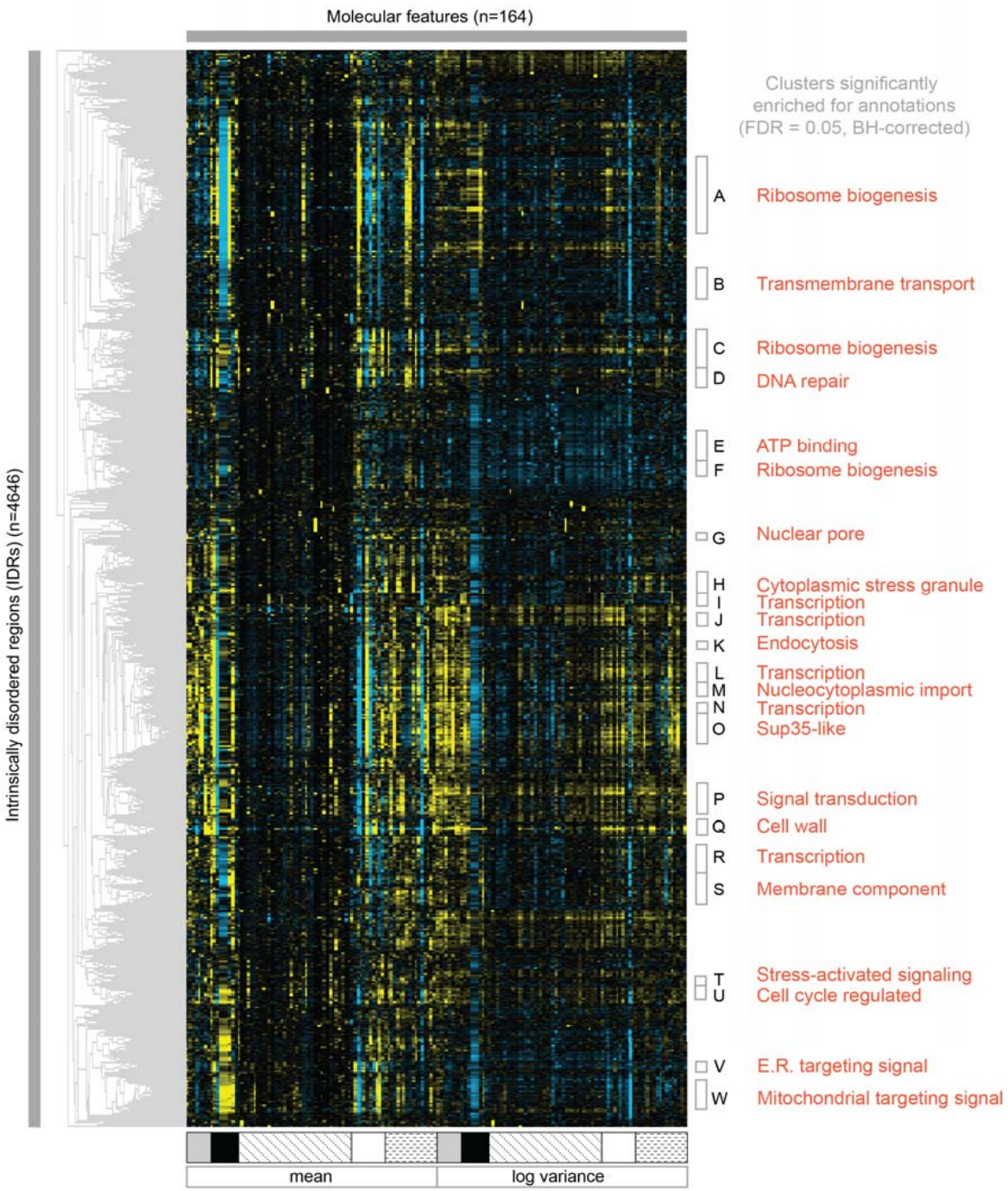
Taraneh Zarin *unpublished*

Other proteins' IDRs can rescue signaling!?

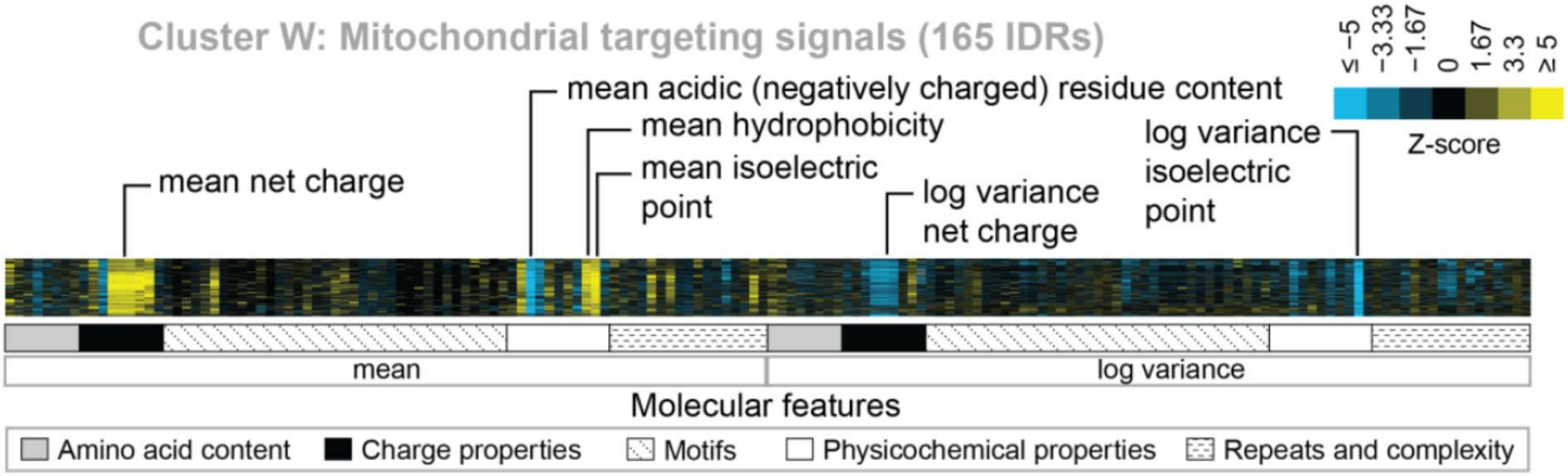


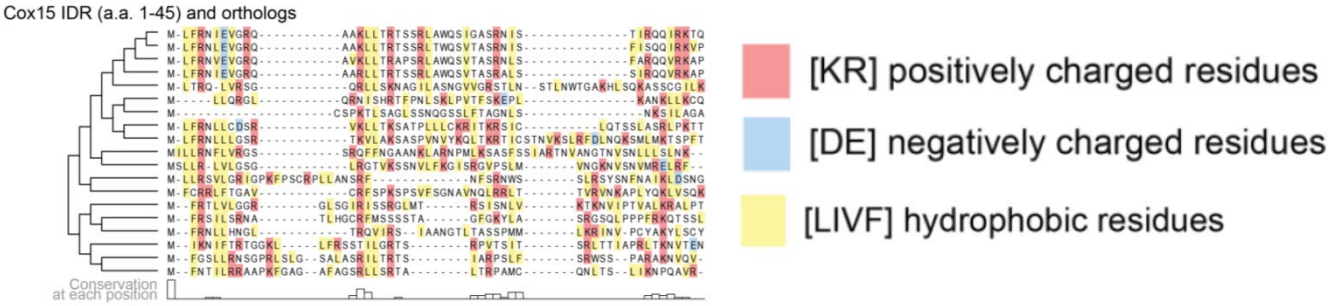


Taraneh Zarin *unpublished*

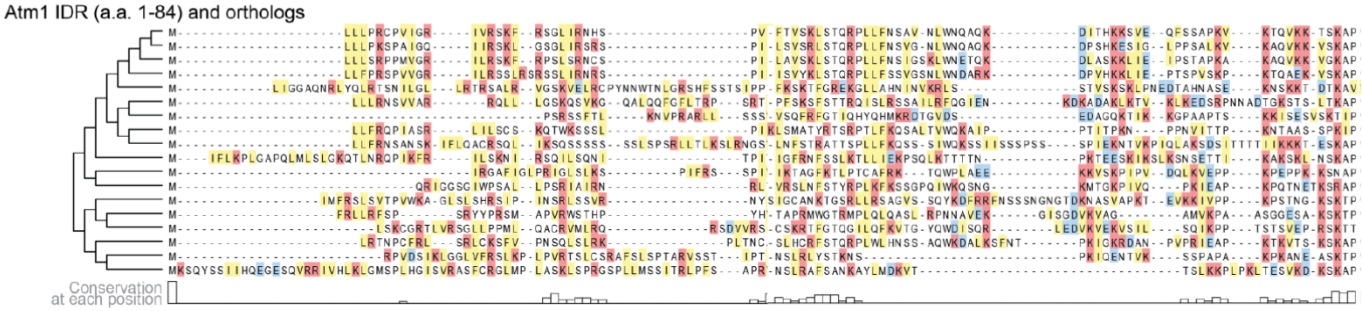


Evolutionary patterns of molecular features are associated with specific biological functions





144/165 IDRs in this cluster are in mitochondrial proteins

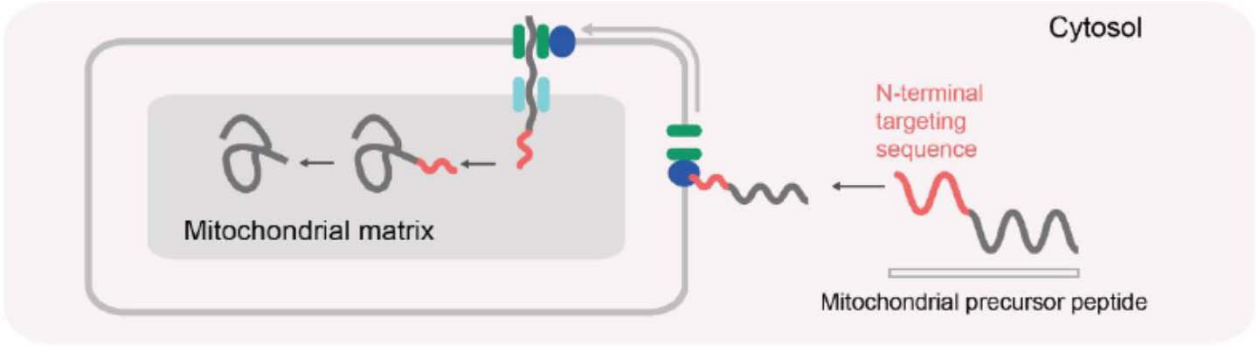


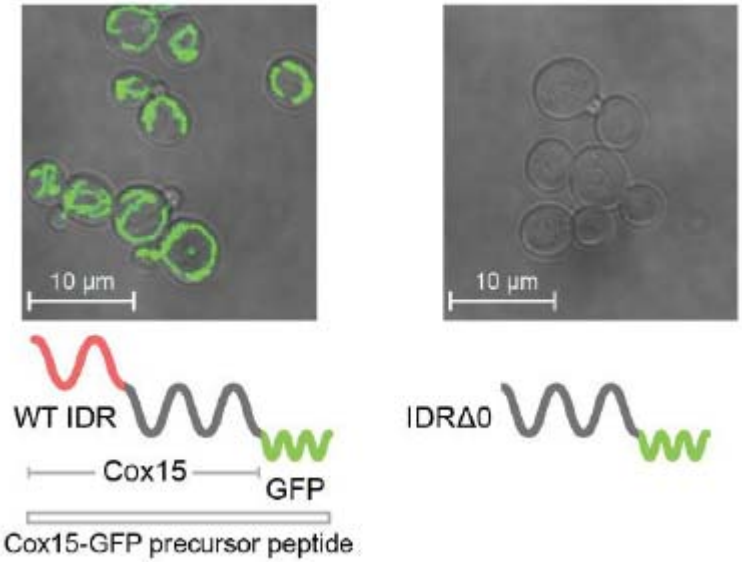
No clear motifs or detectable, sequence similarity, or evolutionary conservation

Evolutionary patterns of molecular features are associated with specific biological functions

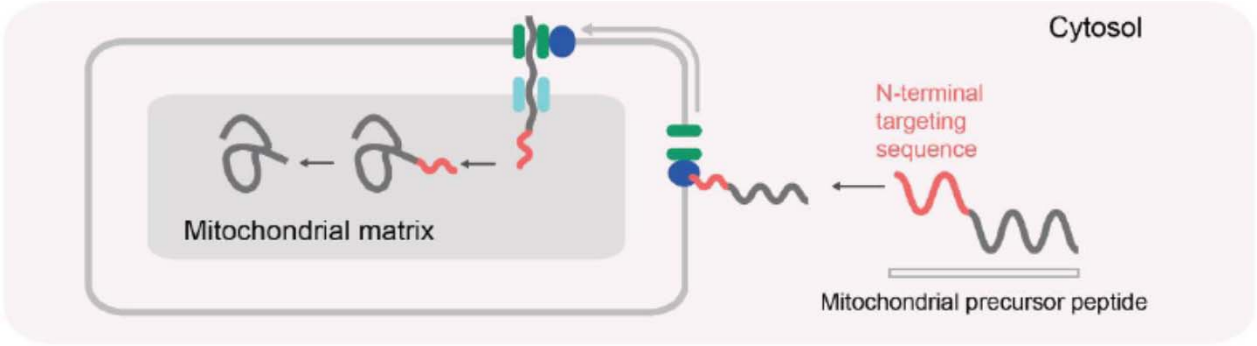
Taraneh Zarin *unpublished*

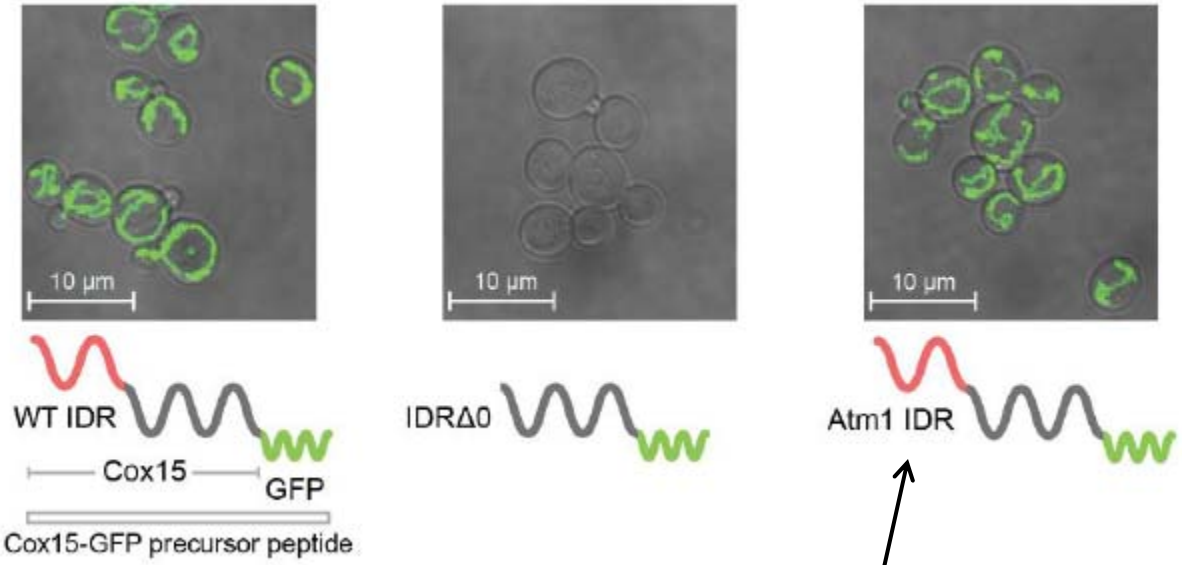
These IDRs are clearly targeting signals



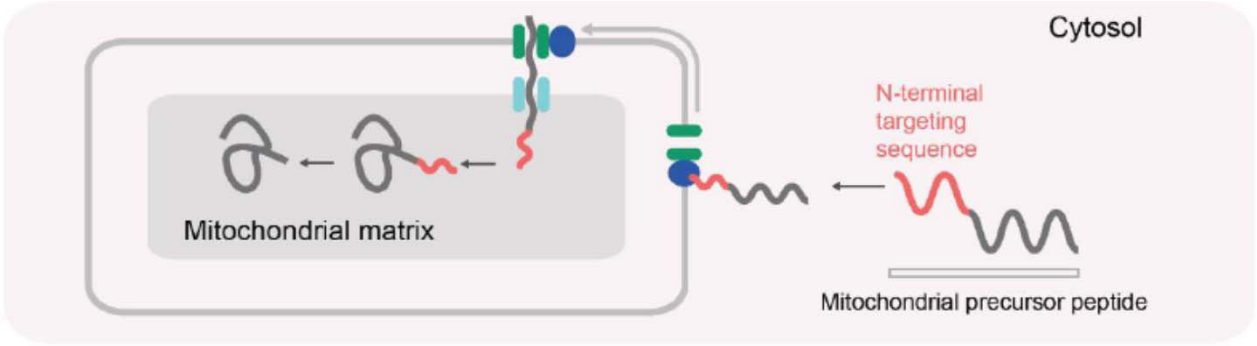


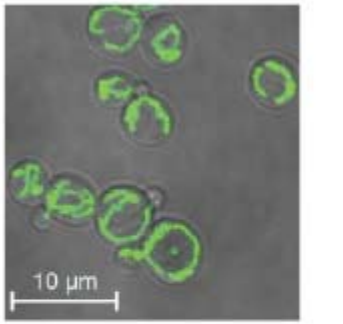
Other proteins' IDRs can rescue targeting

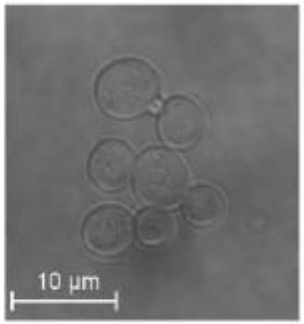


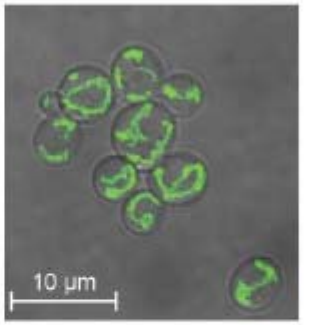


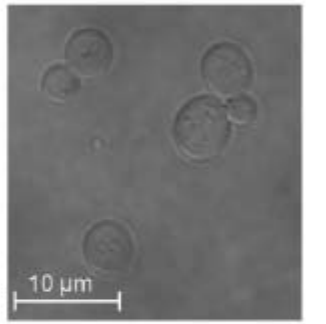
Other proteins' IDRs can rescue targeting









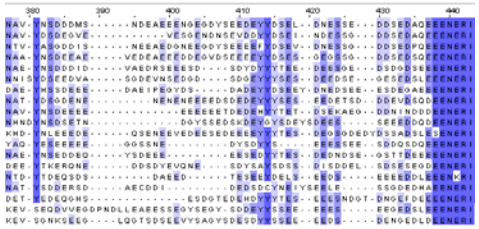


Other IDR from cluster W

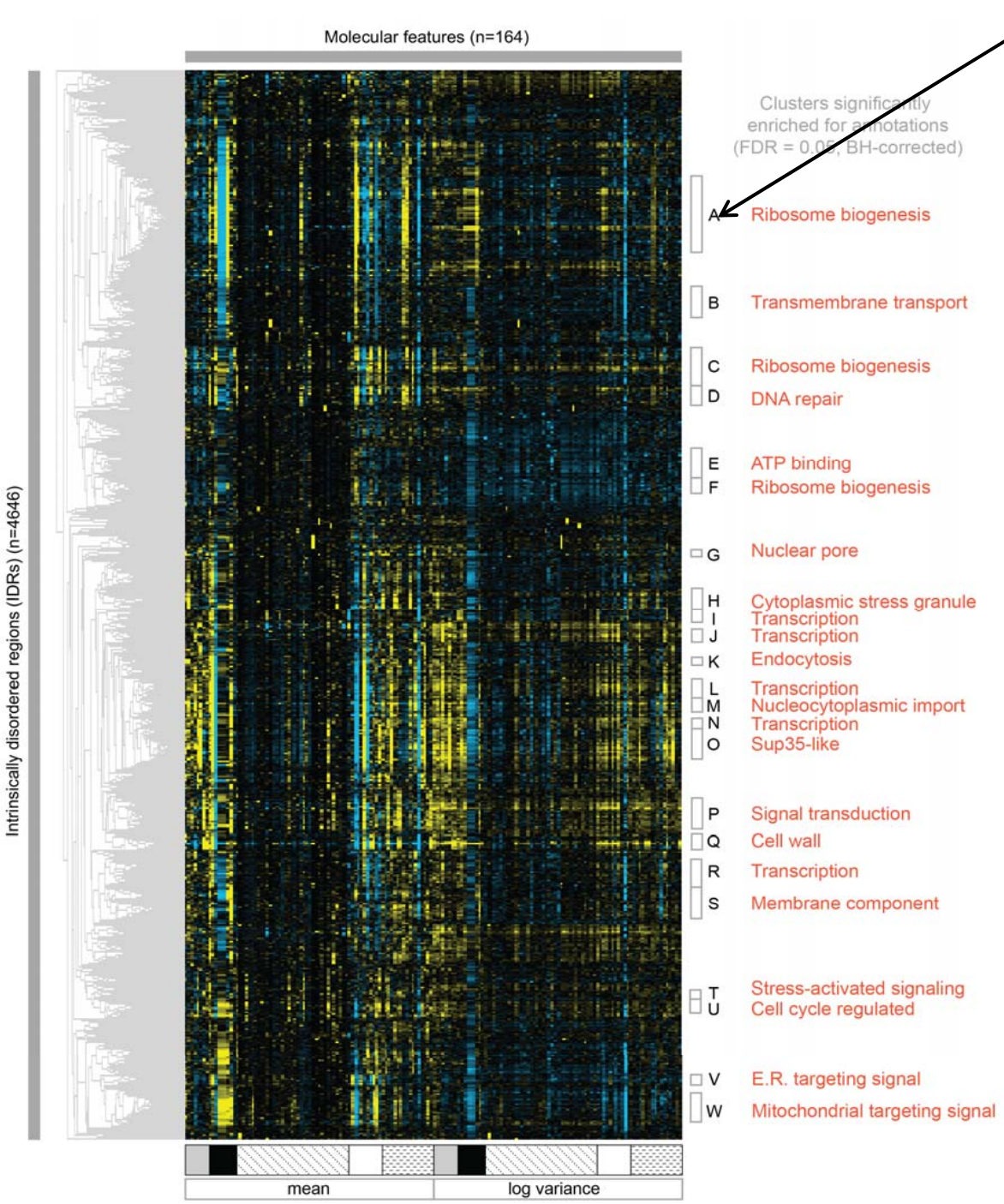
IDR from cluster V

Rio2

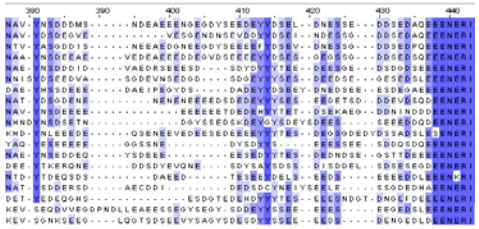
50/295 proteins are nucleolar



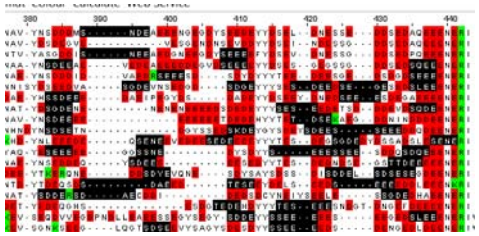
conserved



Rio2



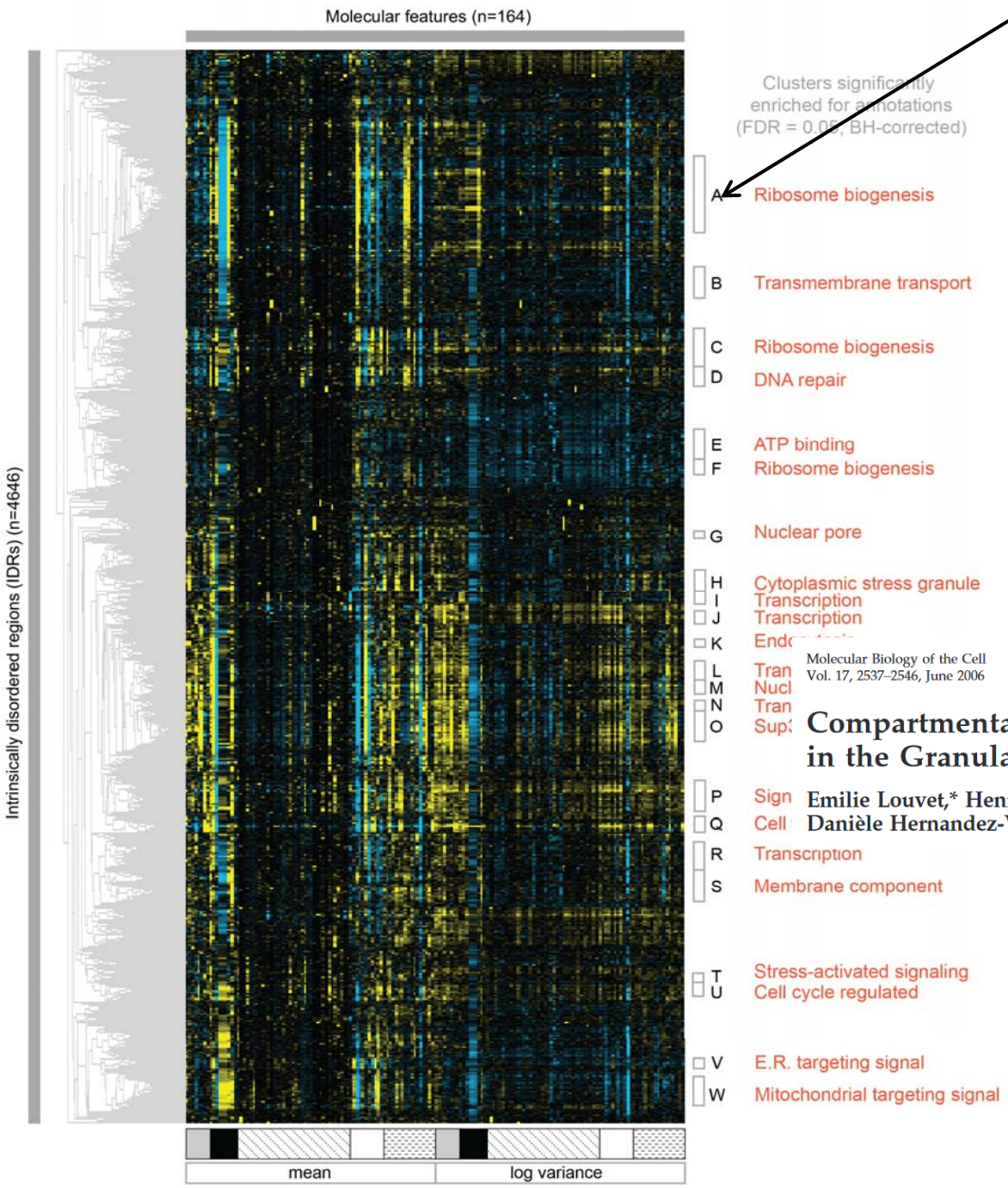
conserved



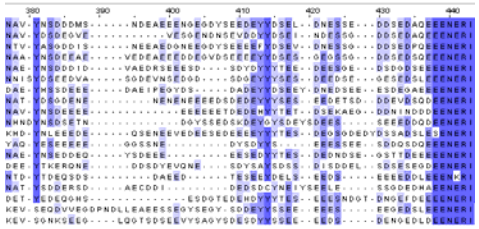
Negatively charged

Positively charged

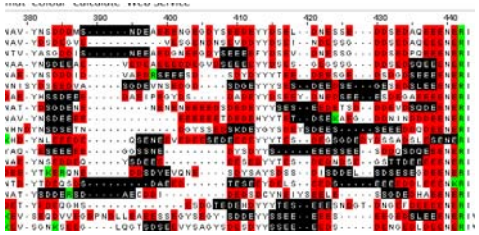
CKII consensus



Rio2



conserved



Negatively charged

Positively charged

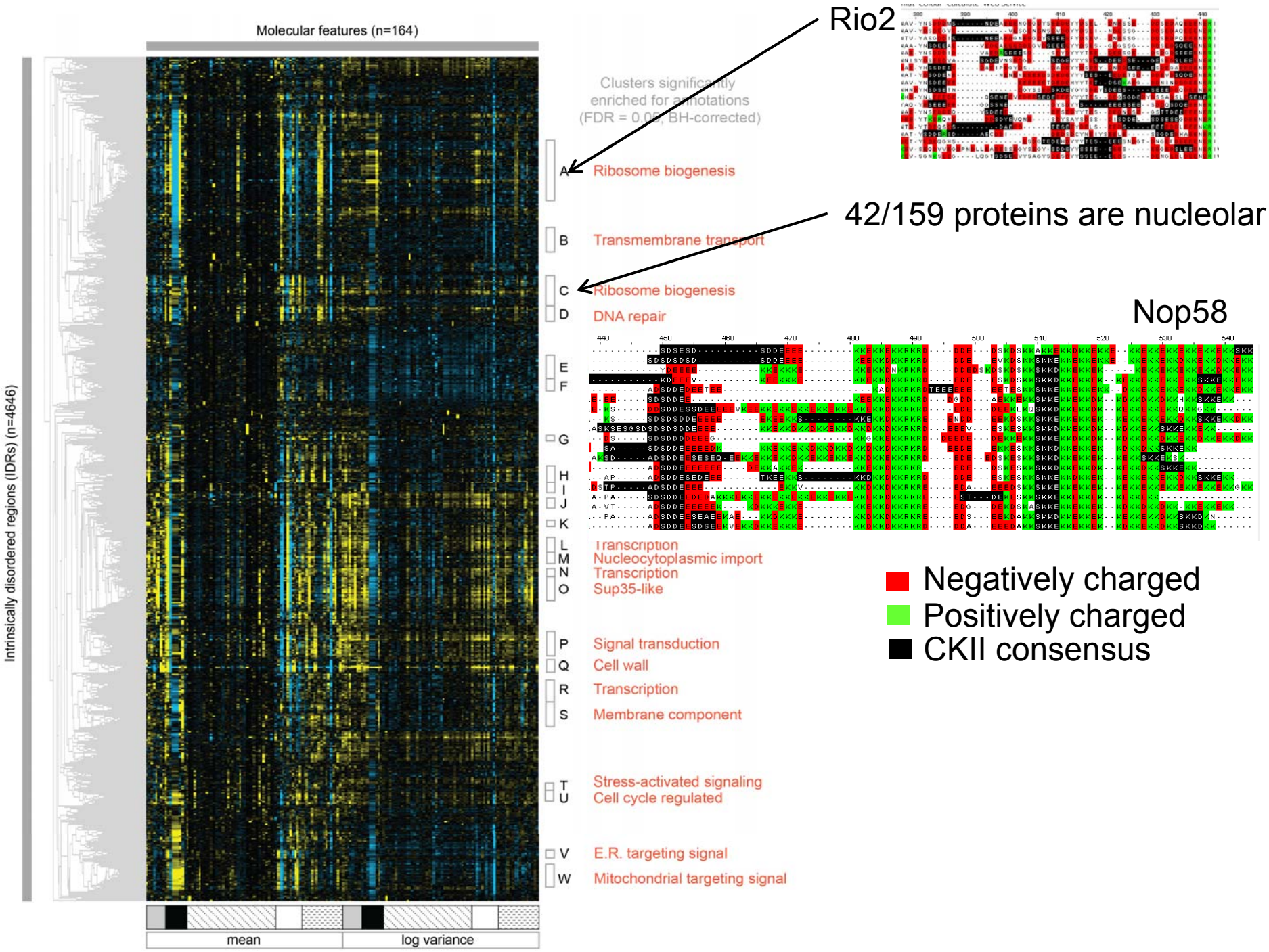
CKII consensus

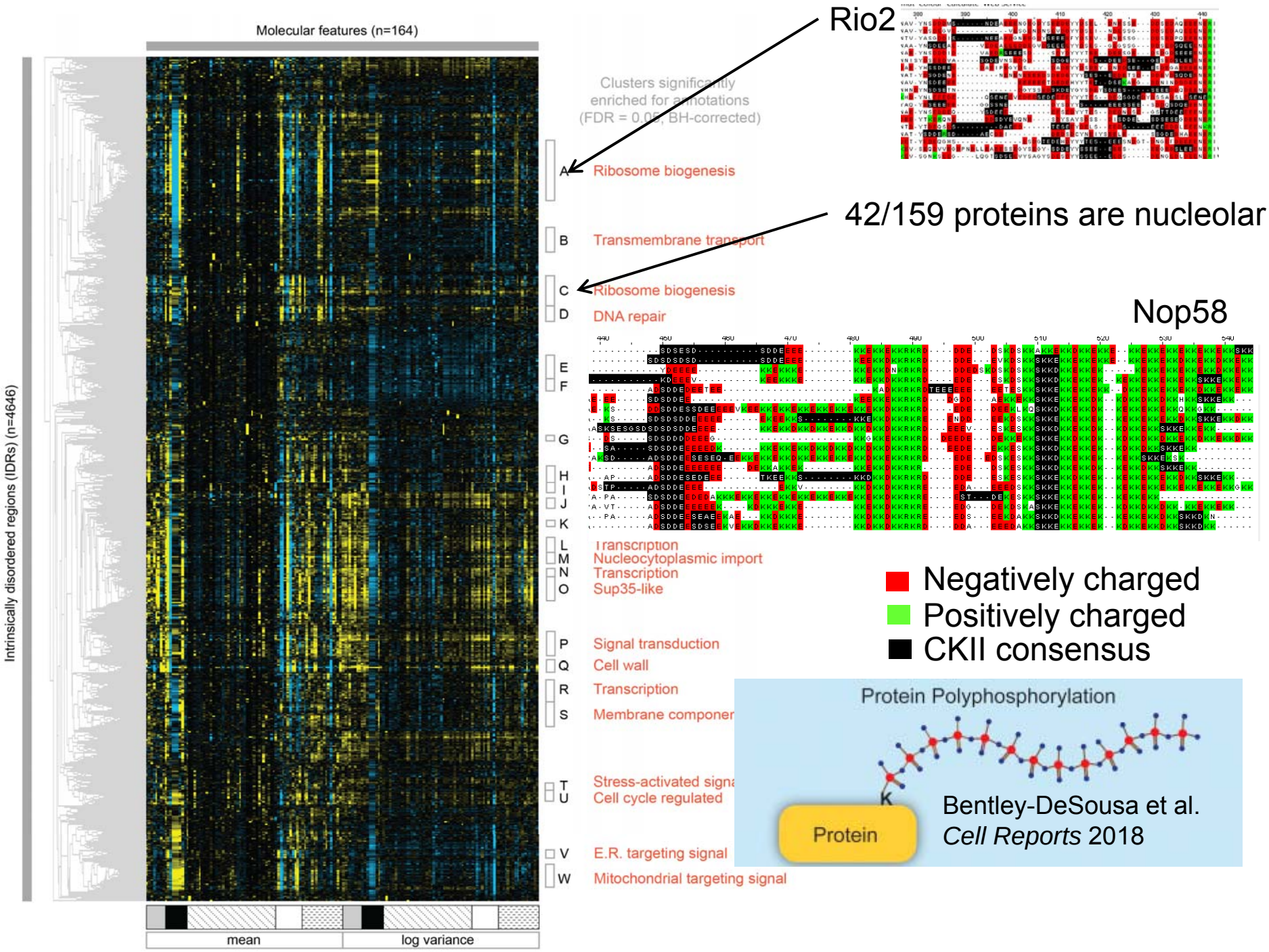
Molecular Biology of the Cell
Vol. 17, 2537–2546, June 2006

Compartmentation of the Nucleolar Processing Proteins in the Granular Component Is a CK2-driven Process

Emilie Louvet,* Henriette Roberte Junéra,* Isabelle Berthuy, and Danièle Hernandez-Verdun

Hundreds of proteins may be regulated by CKII in a similar manner


Molecular features (n=164)
Intrinsically disordered regions (IDRs) (n=4646)
Clusters significantly enriched for annotations (FDR = 0.05, BH-corrected)
A Ribosome biogenesis
B Transmembrane transport
C Ribosome biogenesis
D DNA repair
E
F
G
H
I
J
K
L Transcription
M Nucleocytoplasmic import
N Transcription
O Sup35-like
P Signal transduction
Q Cell wall
R Transcription
S Membrane component
T Stress-activated signaling
U Cell cycle regulated
V E.R. targeting signal
W Mitochondrial targeting signal
mean
log variance
Rio2
42/159 proteins are nucleolar
Nop58
Negatively charged
Positively charged
CKII consensus


Molecular features (n=164)
Intrinsically disordered regions (IDRs) (n=4646)
mean
log variance
Clusters significantly enriched for annotations (FDR = 0.05, BH-corrected)
A Ribosome biogenesis
B Transmembrane transport
C Ribosome biogenesis
D DNA repair
E
F
G
H
I
J
K
L
M
N
O
Transcription
Nucleocytoplasmic import
Transcription
Sup35-like
P Signal transduction
Q Cell wall
R Transcription
S Membrane componer
T Stress-activated signa
U Cell cycle regulated
V E.R. targeting signal
W Mitochondrial targeting signal
Rio2
42/159 proteins are nucleolar
Nop58
Negatively charged
Positively charged
CKII consensus
Protein Polyphosphorylation
K
Protein
Bentley-DeSousa et al.
Cell Reports 2018

Signatures of function in IDRs

- Rapidly evolving IDR sequences contain rich biological information
  - Mitochondrial targeting signals
  - Postranslational modifications associated with nucleolus
- Seems to rule out “mostly junk” hypothesis
- Shared molecular features must be due to convergent evolution
  - E.g., mitochondrial targeting peptides
- Should be possible to predict IDR function from sequence as is now done for folded protein domains
  - Won’t work using BLAST or HMMer

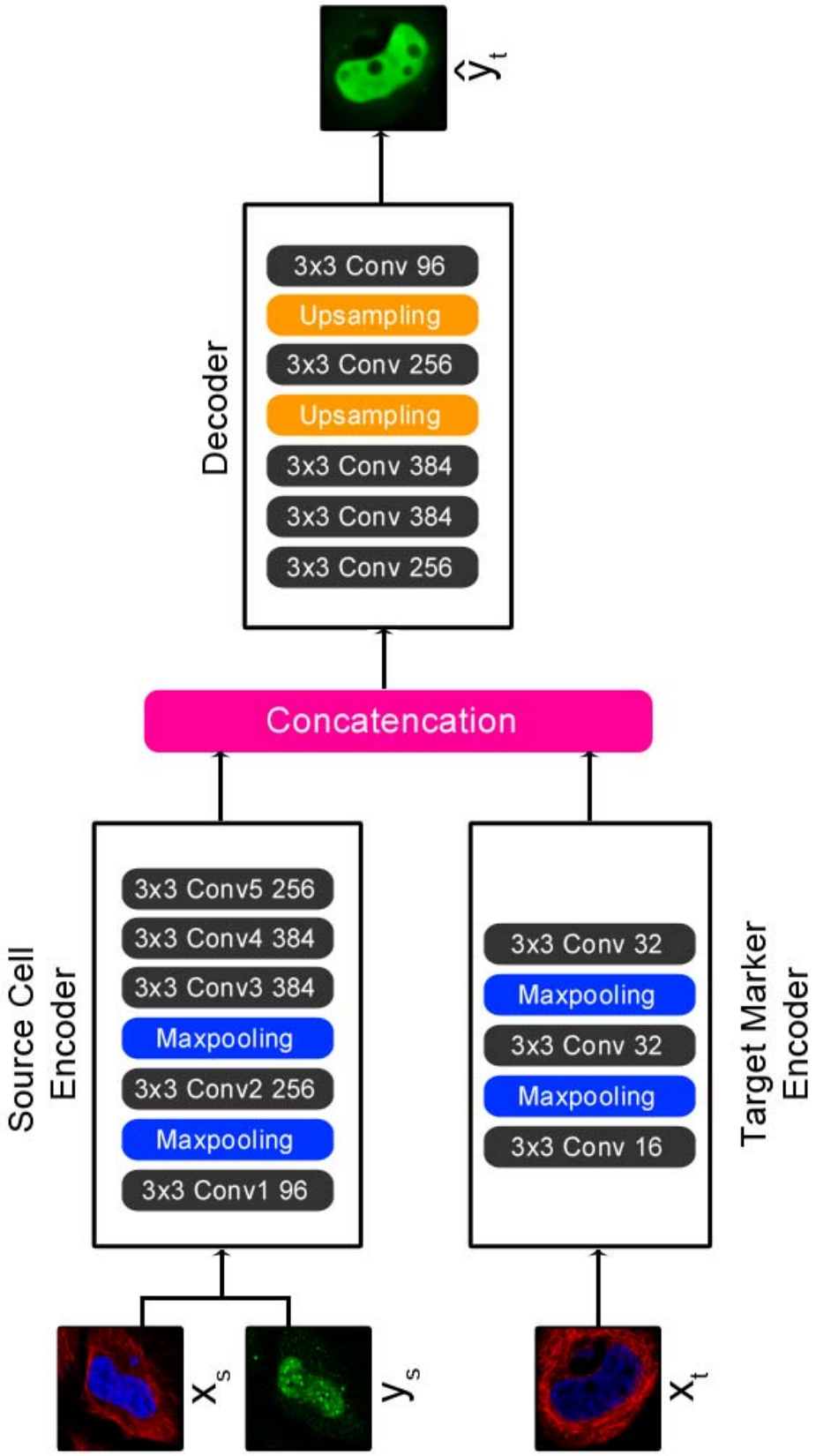
Outline

- Introduction: regulation of proteins
- Automatic identification of protein localization changes in microscopy images
- Unsupervised classification of intrinsically disordered protein regions
  - Evidence for conservation of bulk properties in highly diverged disordered regions

Zarin *et al. PNAS 2017*

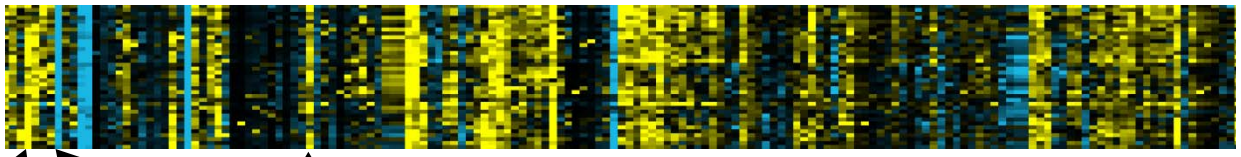
  - Evolutionary signatures of function


MOSES LAB
Iva Pritisanac
Nirvana Nursimulu
Shadi Zabad
Amy Lu
CSB
Cell & Systems Biology
Sergei Plotnikov
YGOB
wellcome sanger institute
Louis-Francois Handfield
FAS Center for Systems Biology
Alex Nguyen Ba
SickKids
Julie Forman-Kay
Donnelly Centre for Cellular + Biomolecular Research
UNIVERSITY OF TORONTO
Judice Koh
Yolanda Chong
Brenda Andrews
Helena Friesen
Caressa Tsai
Taraneh Zarin
Bob Strome
Alex Lu
Ian Hsu
#blessed
NSERC CRSNG
Political Message
Open access: "real scientists do it in public"
www.plos.org www.biomedcentral.com
Canada Research Chairs
Chaires de recherche du Canada
Canada Foundation for Innovation
Fondation canadienne pour l'innovation
Ontario
Canadian Institutes of Health Research
Instituts de recherche en santé du Canada


$\hat{y}_t$
Decoder
3x3 Conv 96
Upsampling
3x3 Conv 256
Upsampling
3x3 Conv 384
3x3 Conv 384
3x3 Conv 256
Concatencation
Source Cell Encoder
3x3 Conv5 256
3x3 Conv4 384
3x3 Conv3 384
Maxpooling
3x3 Conv2 256
Maxpooling
3x3 Conv1 96
Target Marker Encoder
3x3 Conv 32
Maxpooling
3x3 Conv 32
Maxpooling
3x3 Conv 16
x_s
y_s
x_t

What about Ste50?

- Ste50 IDR is in one of the clusters associated with transcription: 18/39 proteins are sequence-specific transcription factors
- Signature for this cluster is very complicated:



Enriched proline

Depleted in acidic residues

S/T-P sites

Demixed charged/ proline residues

High R/K ratio

Reduced variance of charge