Mutations and Variations in Health and Disease: Protein Interaction Networks and 3D Structure Information

Franca Fraternali
Randall Centre for Cellular and Molecular Biophysics

Fraternalilab

http://fraternalilab.kcl.ac.uk/wordpress/

Randall Centre for Cellular and Molecular Biophysics

KING'S College LONDON

THE FRANCIS CRICK INSTITUTE

We are interested in the mechanisms underlying molecular interactions and regulation

Create curated datasets from public databases
(i.e. genomic variants and Protein Interactions)

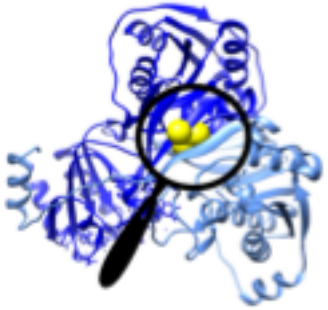
Generate dynamical trajectories of selected proteins and protein complexes

Develop methods to analyse large-scale data

Design comprehensive web tools, which enable access to all data (raw and processed) and our software

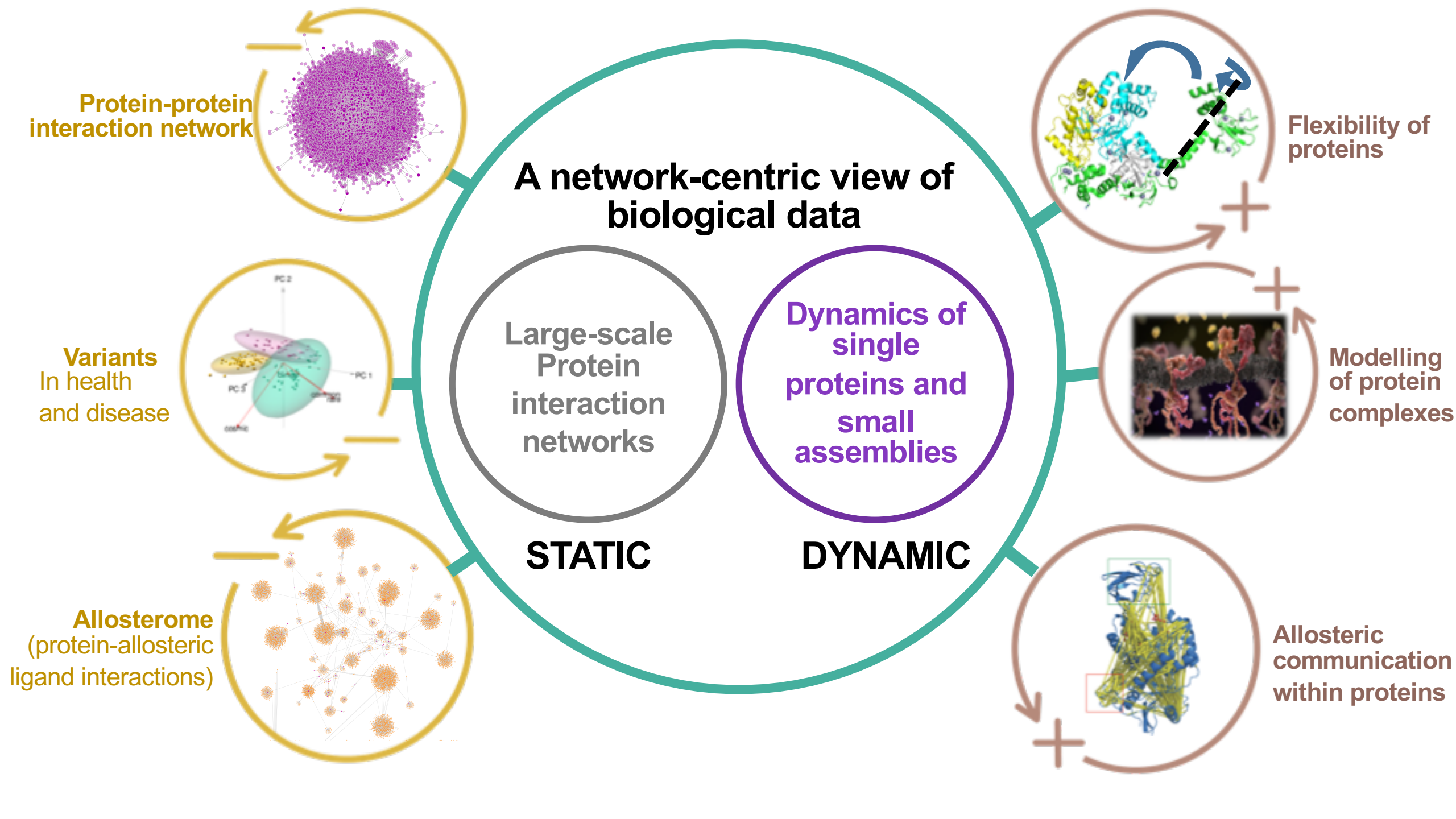
PopSCOMP

ZoomVar

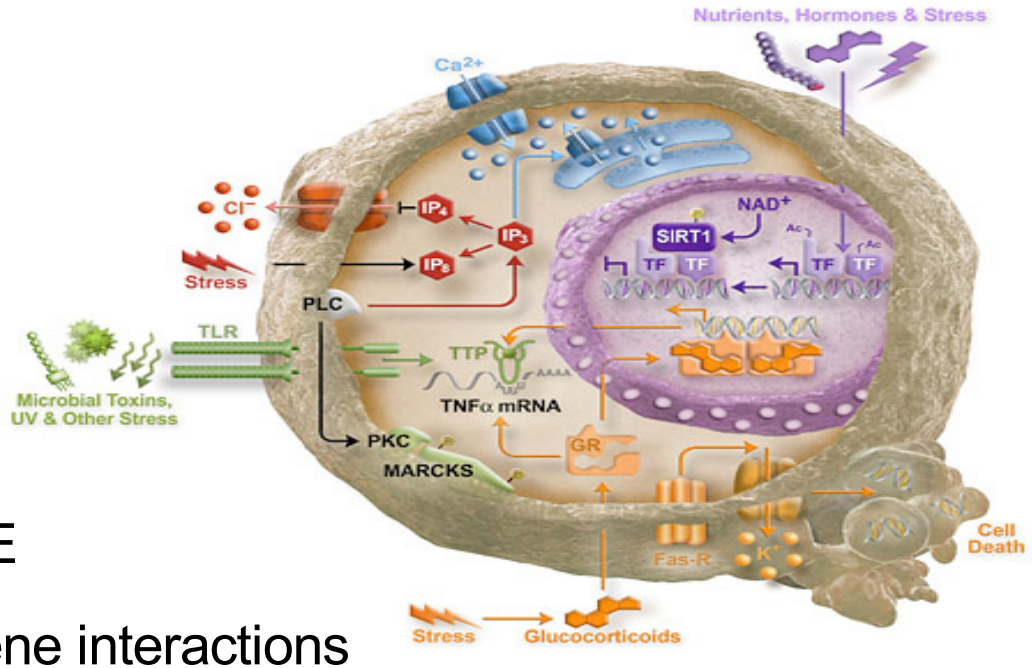


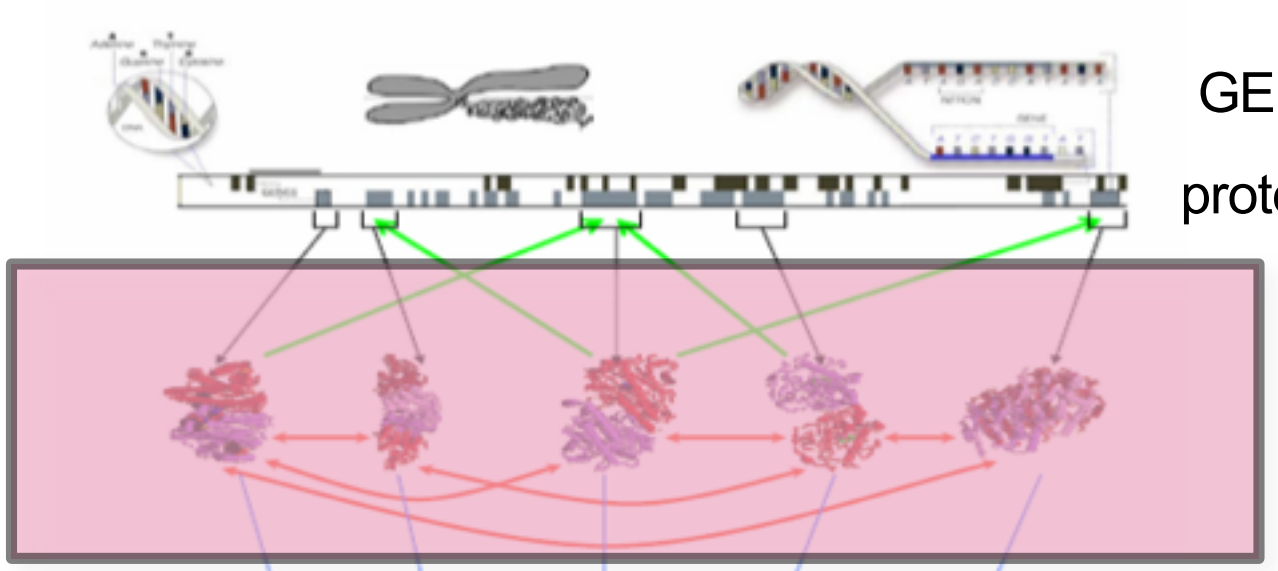
TITINdb

BRepertoire


A network-centric view of biological data
Large-scale Protein interaction networks
Dynamics of single proteins and small assemblies
STATIC
DYNAMIC
Protein-protein interaction network
Variants
In health and disease
Allosterome
(protein-allosteric ligand interactions)
Flexibility of proteins
Modelling of protein complexes
Allosteric communication within proteins

Interactions inside the cell...



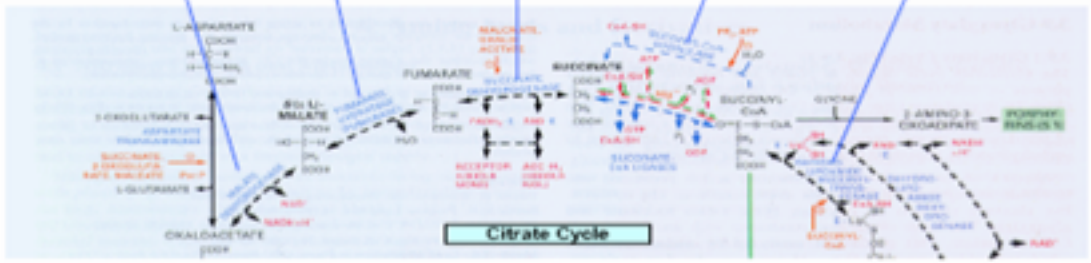


GENOME

protein-gene interactions

PROTEOME

protein-protein interactions



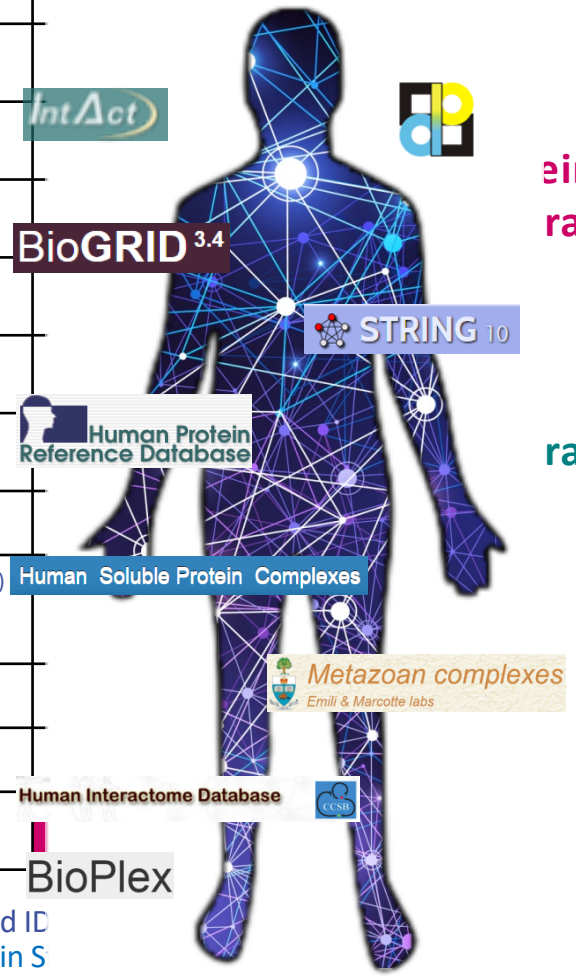
METABOLISM

bio-chemical reactions

Integration protein interaction networks

Integration of human protein-protein interaction networks (PPINs)

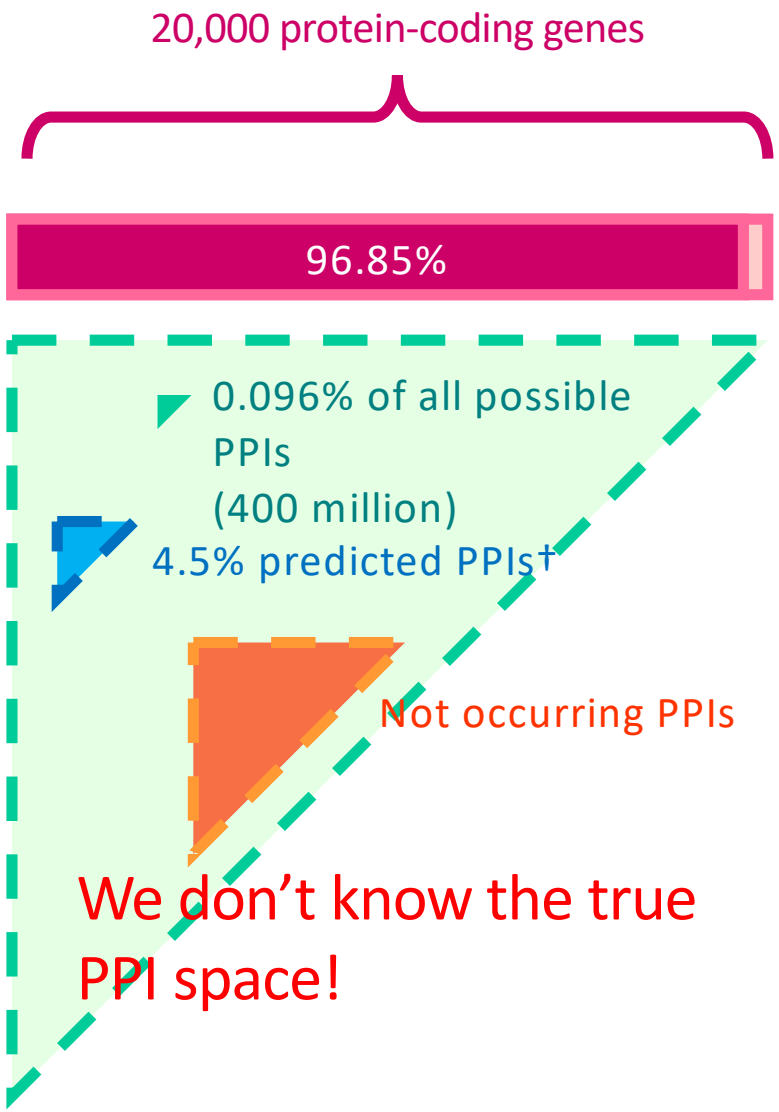
	# Proteins*	# PPIs
IntAct		
BioGRID		
String		
DIP		
HPRD		
BP-MS[1] 2012		
Y2H[2] 2014		
(Epitope-tag) AP-MS[3] 2015		
Co-fraction[4] 2015		
(GFP) AP-MS[5] 2015		
Total (unique)		



ein
rage

rage

19,370 proteins with 385,879 interactions



* Based on UniProt reviewed ID

† 8,548,003 PPIs predicted in S

1) Cell. 2012. A census of human soluble protein complexes.

2) Cell. 2014. A proteome-scale map of the human interactome network

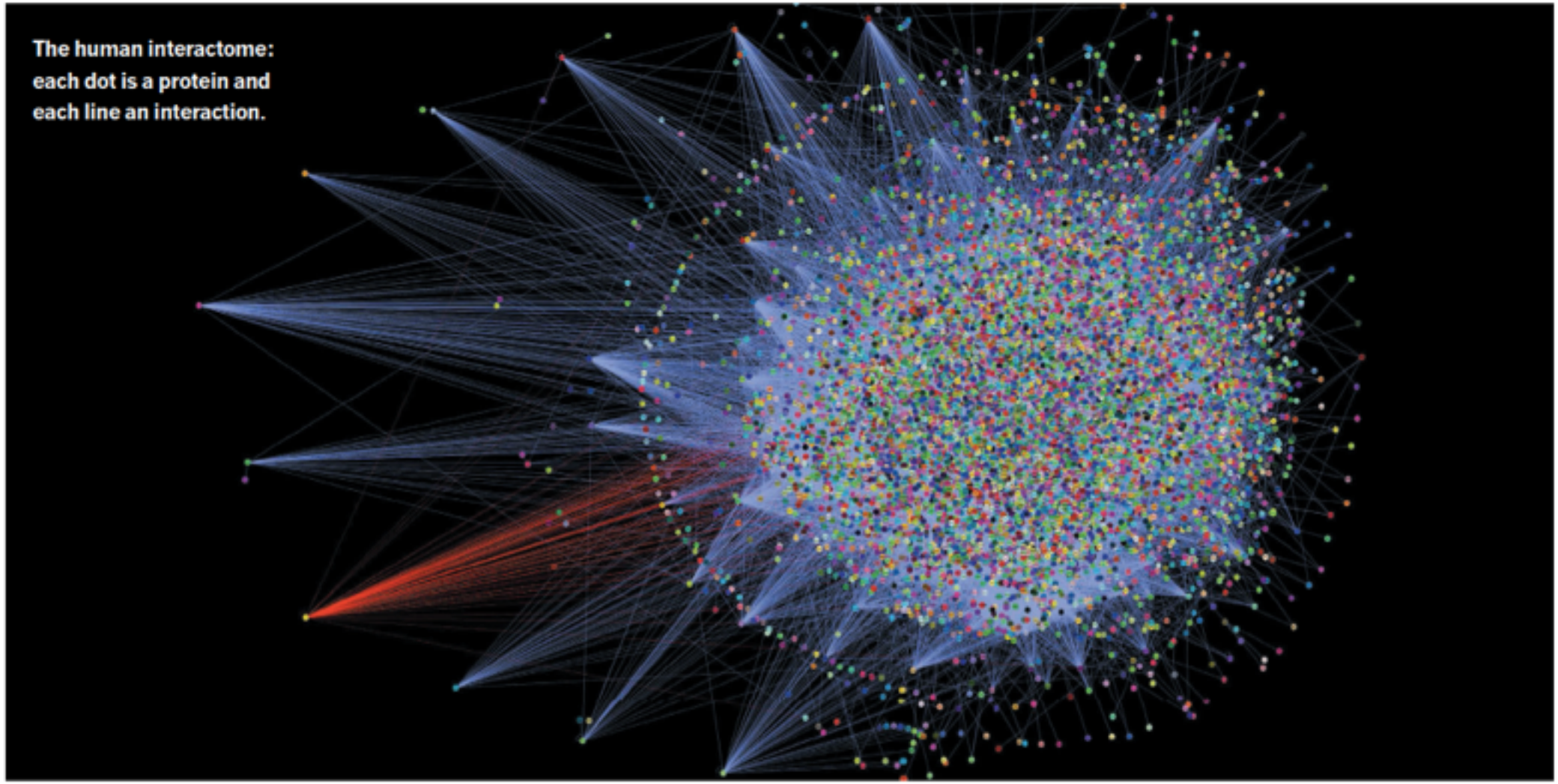
3) Cell. 2015. The BioPlex Network: A ome

4) Nature. 2015. Panorama of ancient

5) Cell. 2015. A Human Interactome in

Stoichiometries and Abundances

ANDREW GARROW



The human interactome: each dot is a protein and each line an interaction.

PROTEIN MAPS CHART THE CAUSES OF DISEASE

Marisa Fessenden

What about predictions?...recently there have been progresses

eLIFE
elifesciences.org

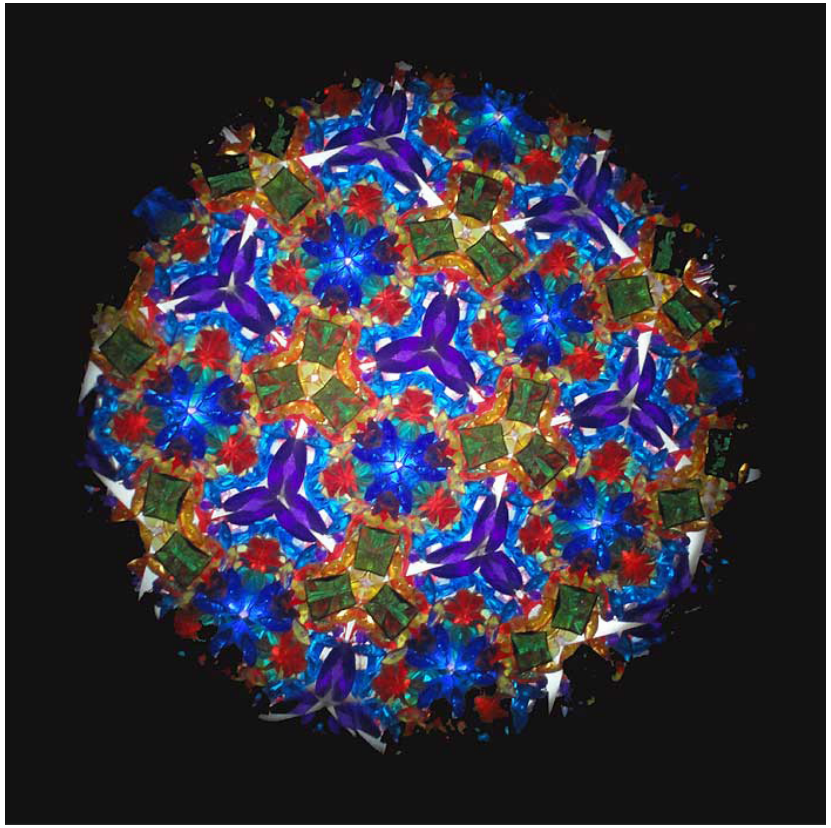
TOOLS AND RESOURCES

A computational interactome and functional annotation for the human proteome

José Ignacio Garzón[1], Lei Deng[1,2], Diana Murray[1], Sagi Shapira[1,3], Donald Petrey[1,4], Barry Honig[1,4,5,6,7*]

PrePPI makes about 127,000 reliable predictions based only on evidence that indicates a direct interaction (structural modeling – SM; protein peptide – PrP, Protein redundancy –PR)

predicted interactions for about 85% of the human proteome



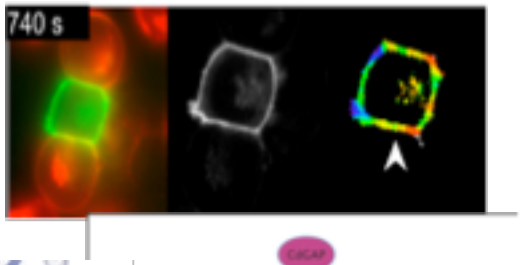
A Kaleidoscopic view of Protein Interactions to extract testable hypotheses for experiments

Development of tools for the analysis of Protein Protein Interactions

TOPOLOGICAL

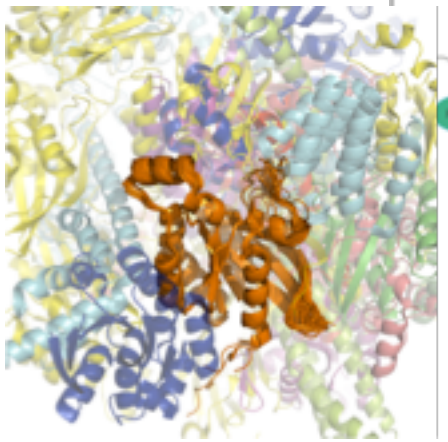
Carlin LM et al Sci Signal. 2011 Nov 29;4(201):ra81.

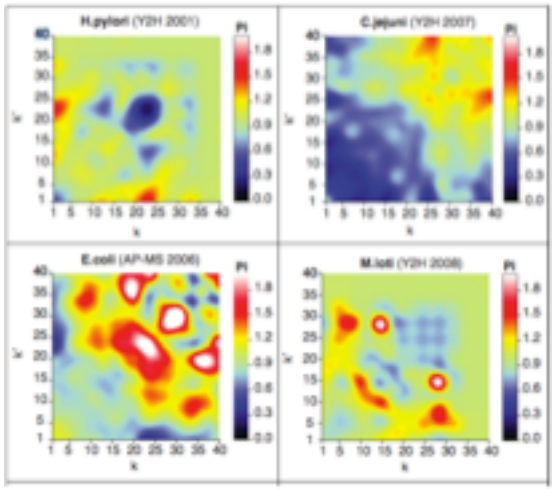
CELLULAR



MOLECULAR

ATOMISTIC



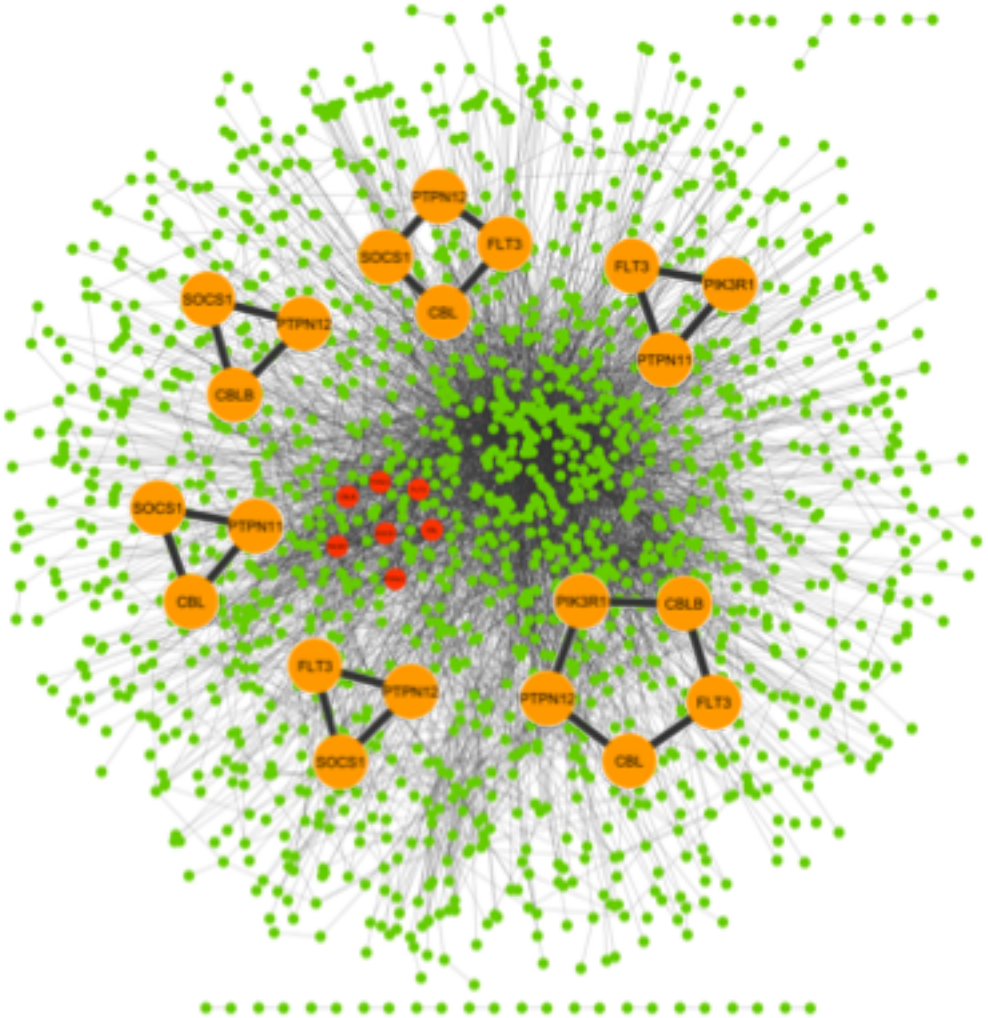


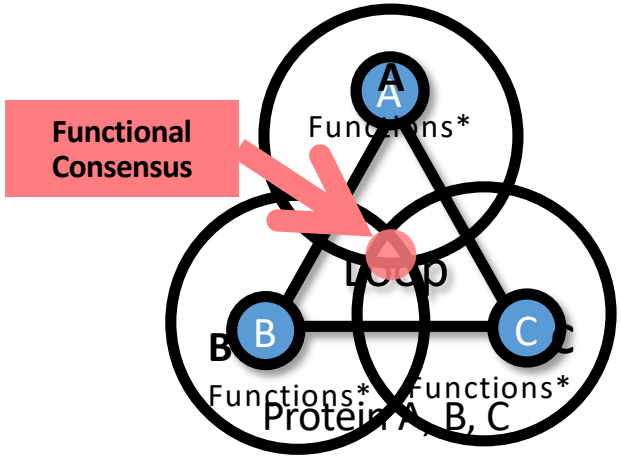
Fernandes LP, Annibale A, Kleinjung J, Coolen AC,
Fraternali F. PLoS One. 2010 Aug 18;5(8):e12083.

Lu, Fornili & Fraternali *Expert Rev Proteomics* 10, 511–520 (2013).

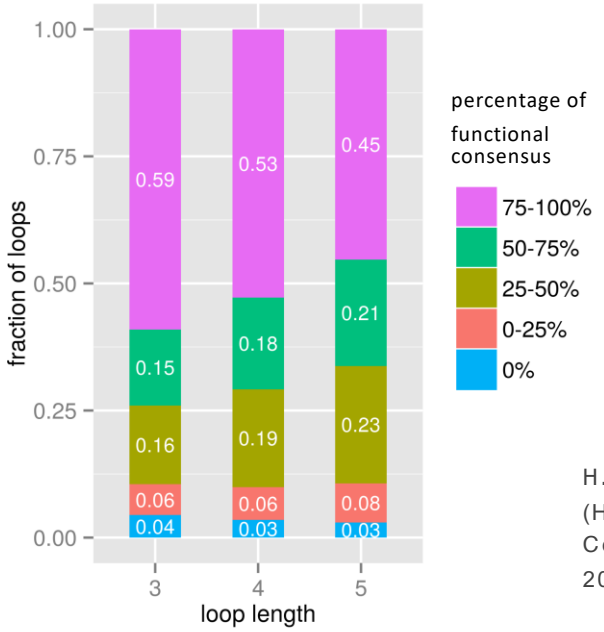
Vaz F et al. Mutation of the RAD51C gene in a Fanconi anemia-like disorder. Nat Genet. 2010 May;42(5):406-9.

Short loop network motif profiling





*Gene Ontology (GO)
Biological process terms (Lv.2 and above)



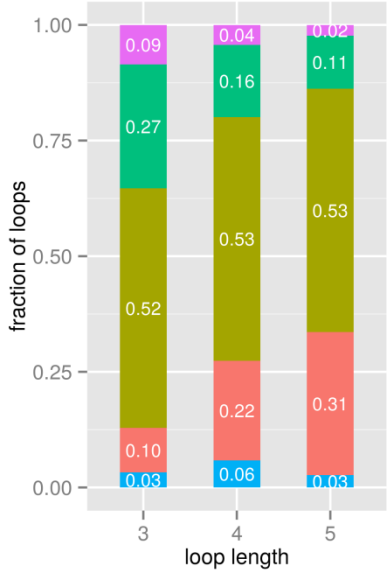
H. sapiens V (BP-MS) (Havugimana et al., Cell 150, 1068–1081 2012)

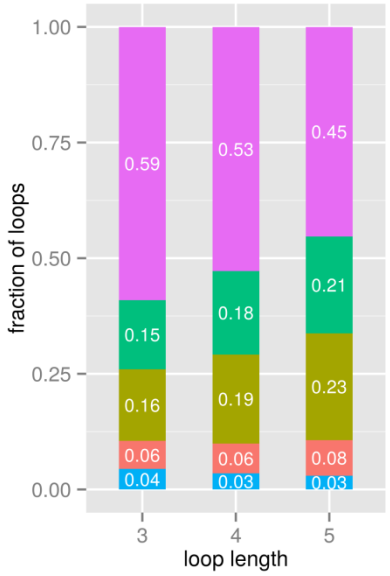
"Throughout large-scale rigorous analyses on PPINs, we found short loops are intrinsic features of protein-protein interaction networks (PPINs); the number of short loops can be a topological barometer of PPINs and their functional annotations can imply not only local enrichments but also wide-ranging associations of short loops."

Chung, S.S. *et al*. Bridging topological and functional information in protein interaction networks by short loops profiling. *Nature Scientific Reports* (2015)

Functional Consensus

H. Sapiens V (BP-MS)

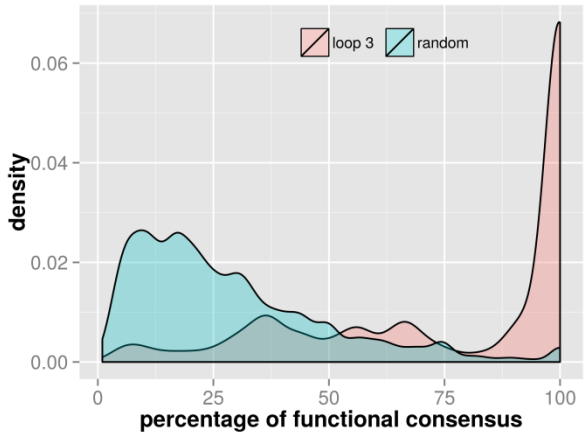


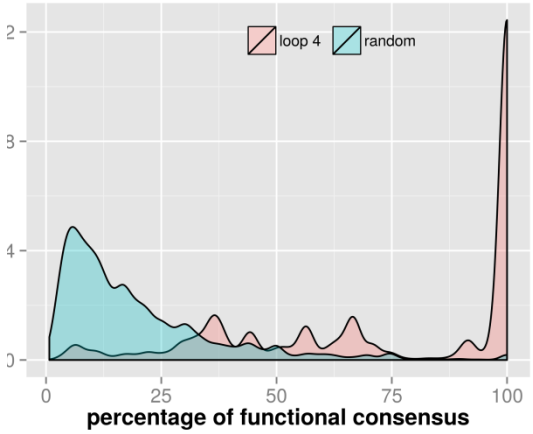


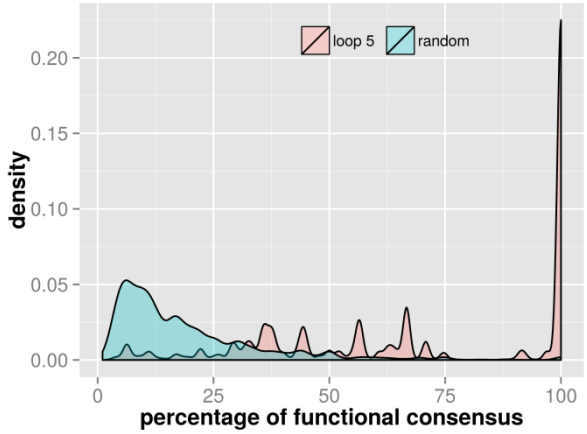
percentage of functional consensus

- 75-100%
- 50-75%
- 25-50%
- 0-25%
- 0%

How do these results compare to a random annotation?

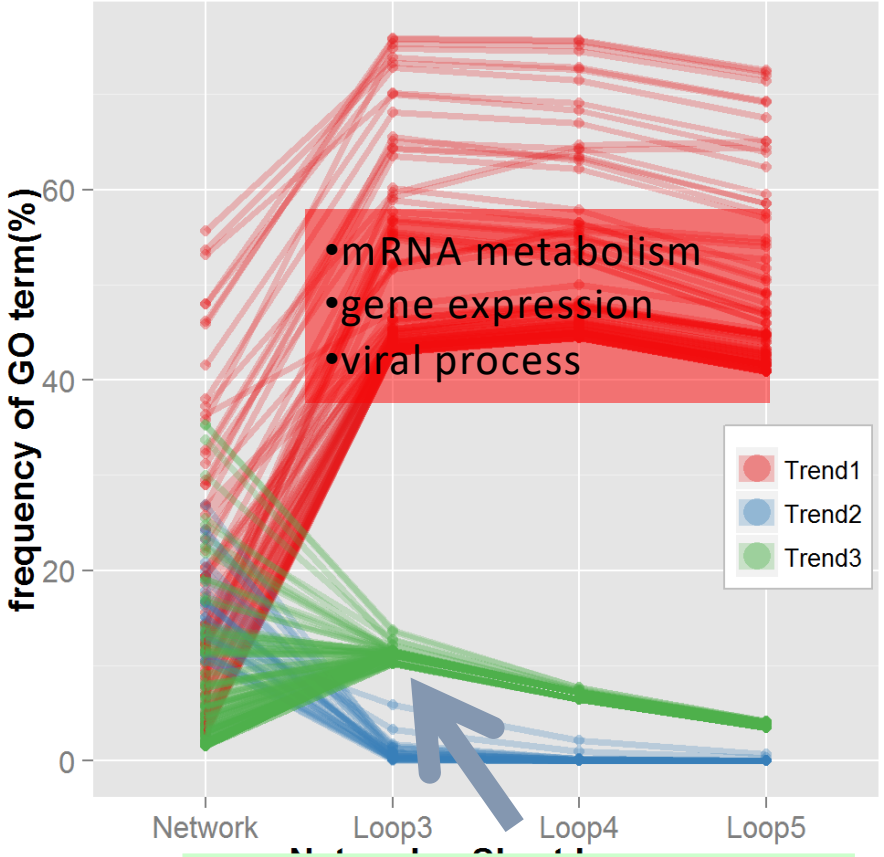






Functional enrichment in short loops

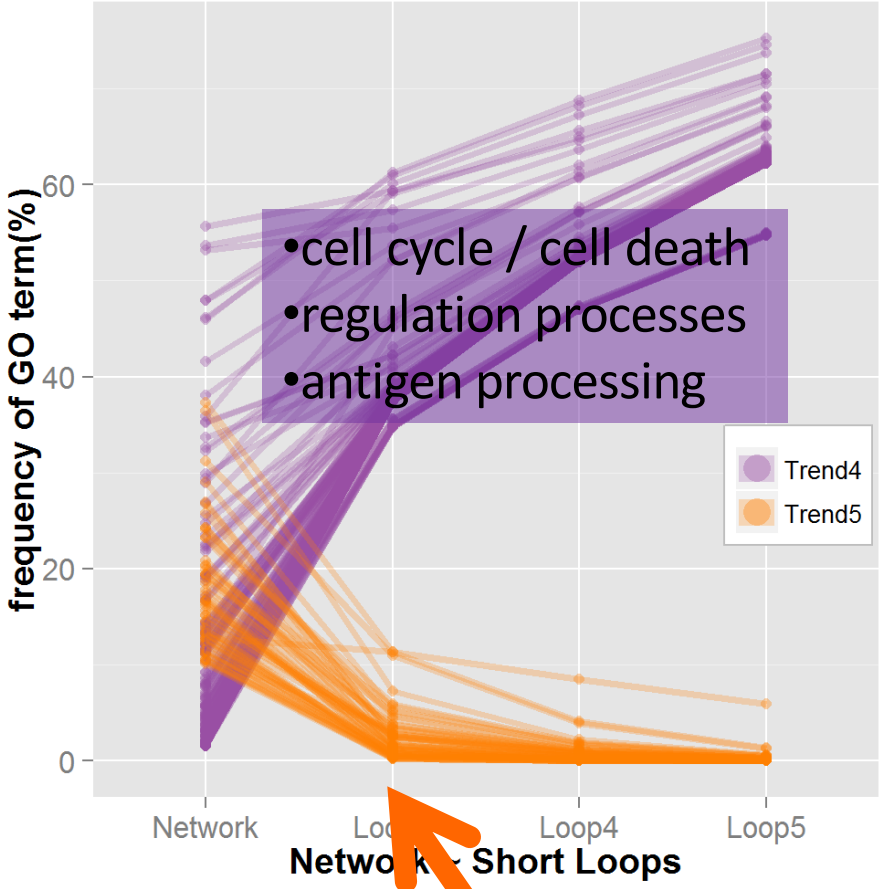
H. Sapiens V (BP-MS)



•mRNA metabolism
•gene expression
•viral process

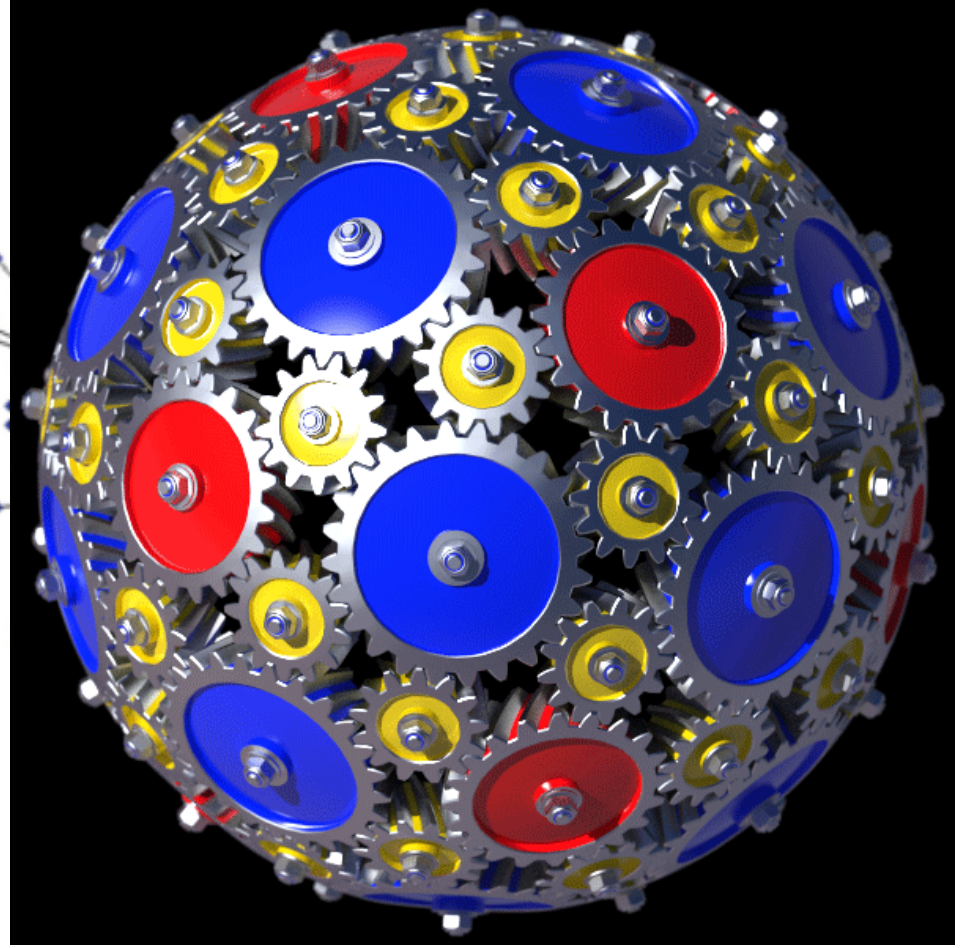
•organismal & developmental processes
•DNA-templated transcription

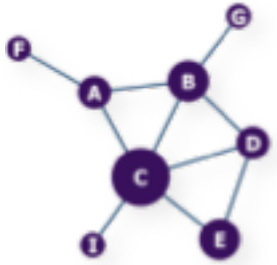
H. Sapiens V (BP-MS) - Ribosomal proteins



•cell cycle / cell death
•regulation processes
•antigen processing

•biosynthesis
•protein complex subunit
•localization (transport)





LONG-RANGE COMMUNICATION THROUGHOUT THE NETWORK

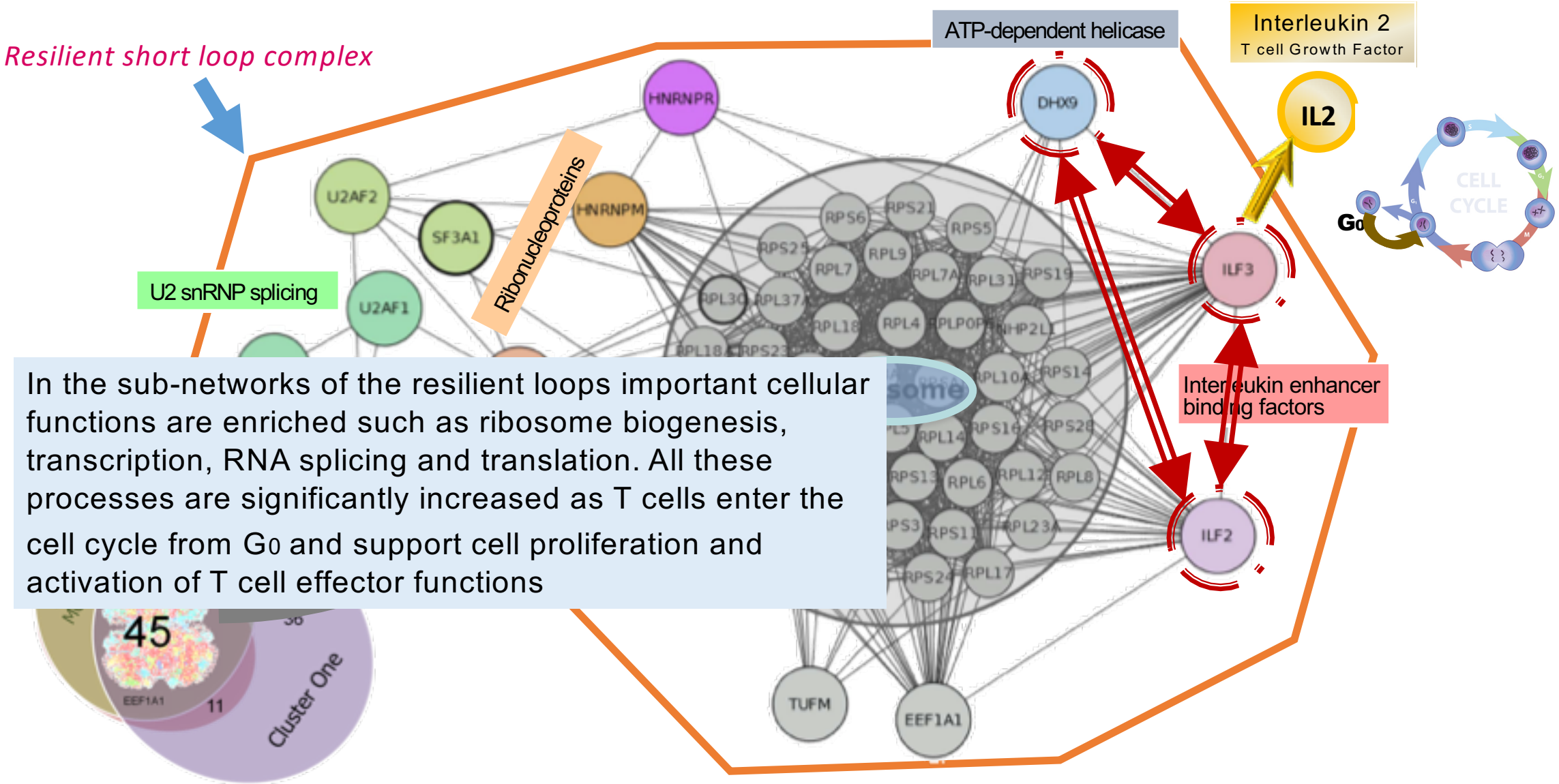
- organismal & developmental processes
- DNA-templated transcription

- biosynthesis
- protein complex subunit
- localization (transport)

- cell cycle / cell death
- regulation processes
- antigen processing

Chung SS, Pandini A, Annibale A, Coolen AC, Thomas NS, Fraternali F. Sci Rep. 2015 Feb 23;5:8540. doi: 10.1038/srep08540

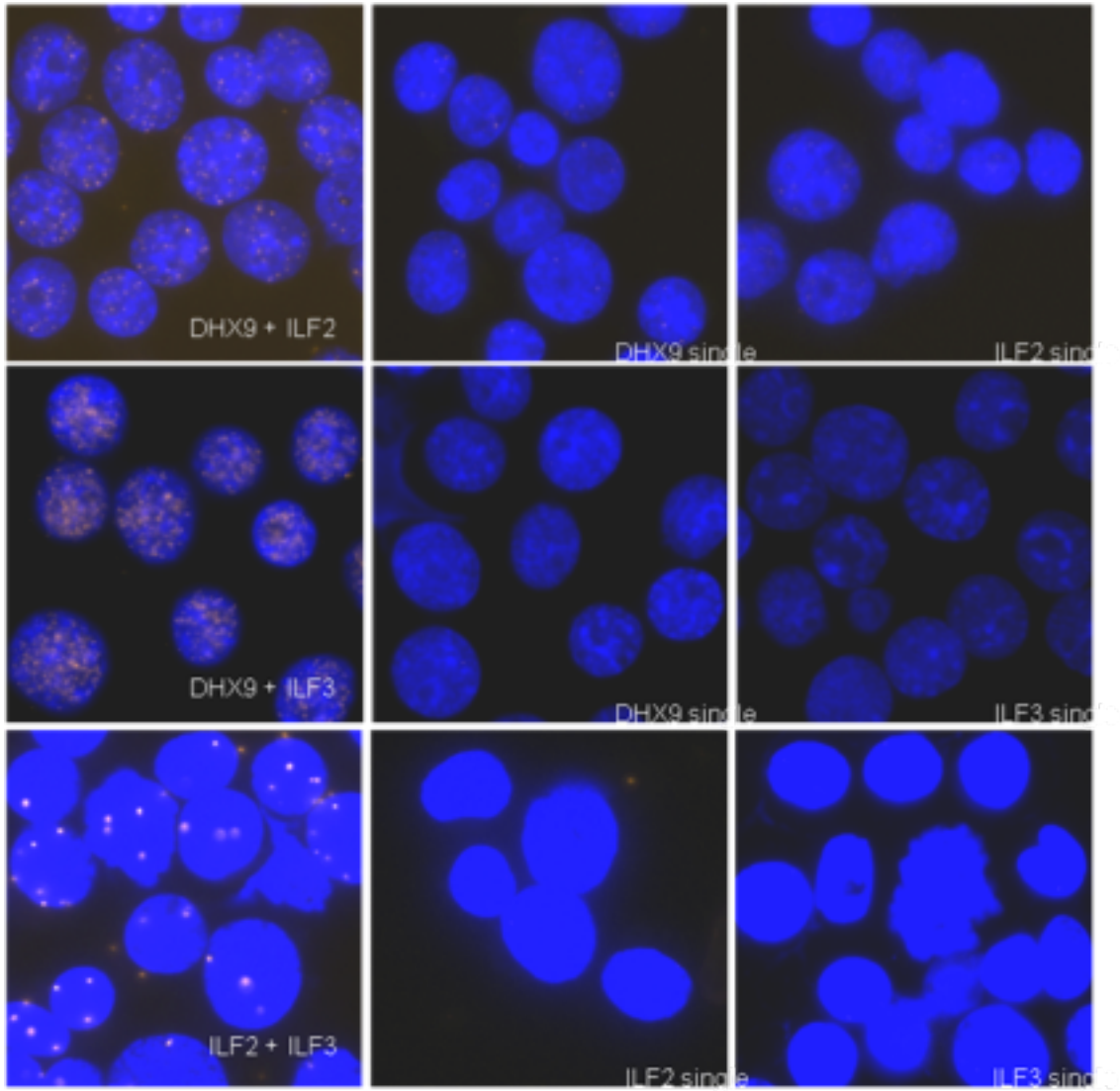
Loop resilience identifies core clusters of protein interactions



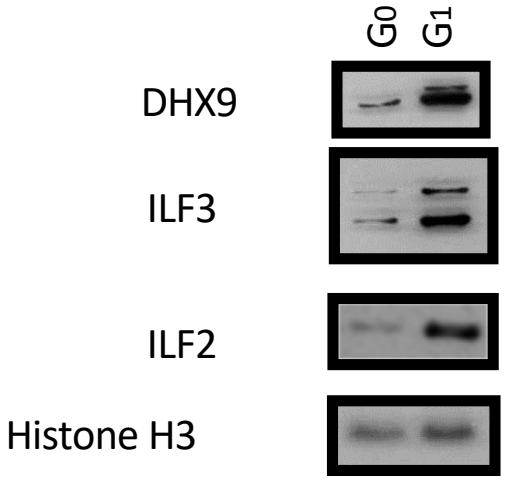
In the sub-networks of the resilient loops important cellular functions are enriched such as ribosome biogenesis, transcription, RNA splicing and translation. All these processes are significantly increased as T cells enter the cell cycle from G_0 and support cell proliferation and activation of T cell effector functions

Probing protein interactions of ILF2, ILF3 and DHX9

using Proximity Ligation Assay



Primary human T-cells
G_0 : Quiescent status
G_1 : 72hours after CD3 + CD28 stimulation

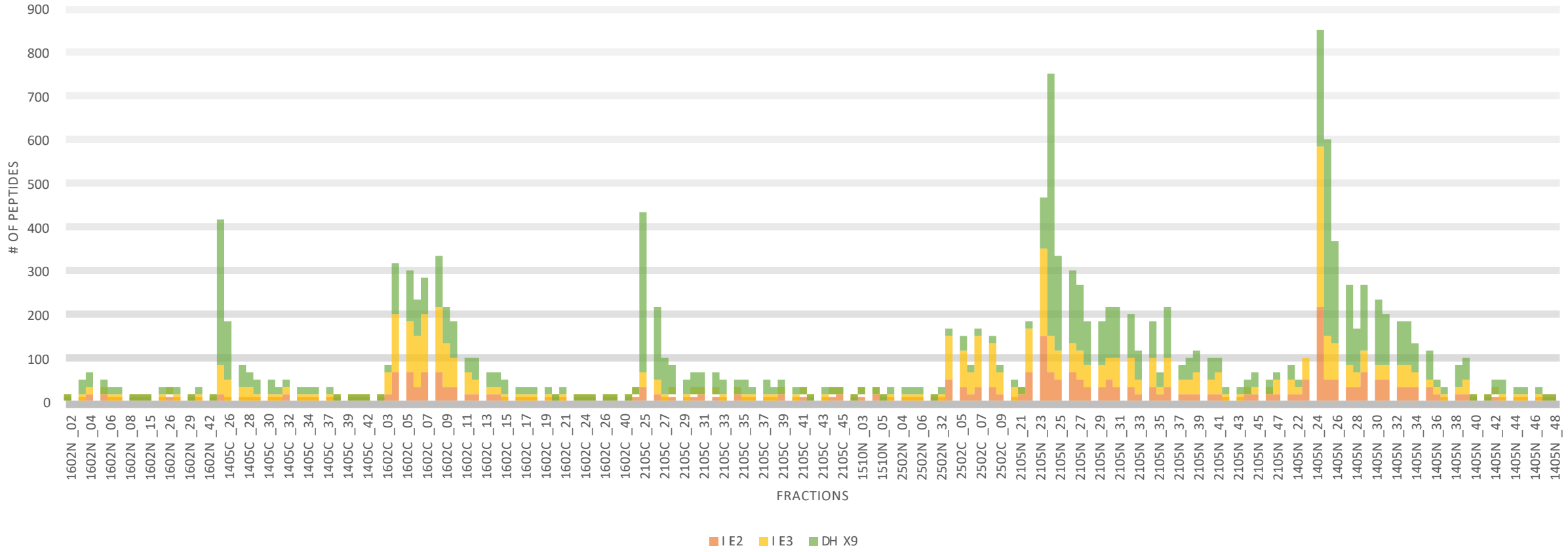


DHX9 : RNA helicase A
ILF3 : IL2 (T-cell growth factor) transcription regulator
ILF2 : Reshuttling of ILF3

In collaboration with Ed Marcotte's laboratory, Austin Texas

Probing protein interactions of ILF2, ILF3 and DHX9 in T-cells

using LC-MS/MS co-fraction assay



Among 8 out of 9 replicas, ILF2, ILF3 and DHX9 are co-eluted in 48 different co-fractionated cell lysates of T-cells. (manuscript in preparation)

Summary

- We have shown by a large-scale analysis of publicly available datasets that the present protein network data are strongly biased by their experimental methods, while still exhibiting species-specific similarity and reproducibility.

- We have introduced a new strategy to identify regulators of a signalling pathway by analysis of short loop motifs in a reliable dataset of human soluble protein interactions

We demonstrate that short loops are an intrinsic property of PPINs AND that contain significant information on functional mechanisms underlying the biology of the cell.

We believe that these communities can be used in drug targeting screens to expand the protein-drug space, and or suggest novel drug-disease associations that offer unprecedented opportunities for drug repurposing and the detection of adverse effects.

Which variants play a causative role in disease?

Problems

Approaches

- Pathogenic titin variants may be present in only a single/few individual(s).
- A number of known cases are due to the combined impact of two distinct mutations.
- We need to look at variant impact on the molecular level!
- *In-vitro* & *in-vivo* methods time-consuming and expensive.
- Prioritise variants using *in-silico* techniques.

Motivation

Explosion in the growth of variant data – gnomAD database.

This has challenged previous conceptions regarding disease-associated variants.

We aimed to perform a more comprehensive comparison of the spatial distribution and regional enrichment of variants in health and disease.

Goal – to uncover features which separate the datasets.

A Pan-Cancer Catalogue of Cancer Driver Protein Interaction Interfaces

Eduard Porta-Pardo, Luz Garcia-Alonso, Thomas Hrabe, Joaquin Dopazo, Adam Godzik

Published: October 20, 2015 • https://doi.org/10.1371/journal.pcbi.1004518

Spatial distribution of disease-associated variants in three-dimensional structures of protein complexes

A Gress, V Ramensky & O V Kalinina

Common sequence variants affect molecular function more than rare variants?

Yannick Mahlich, Jonas Reeb, Maximilian Hecht, Maria Schelling, Tjaart Andries Petrus De Beer, Yana Bromberg & Burkhard Rost

Challenges for variants mapping:
Titin - the largest protein in the Human Body

35991 amino acids (inferred complete (IC)isoform), weighs over 4000 kDa and spans half a sarcomere

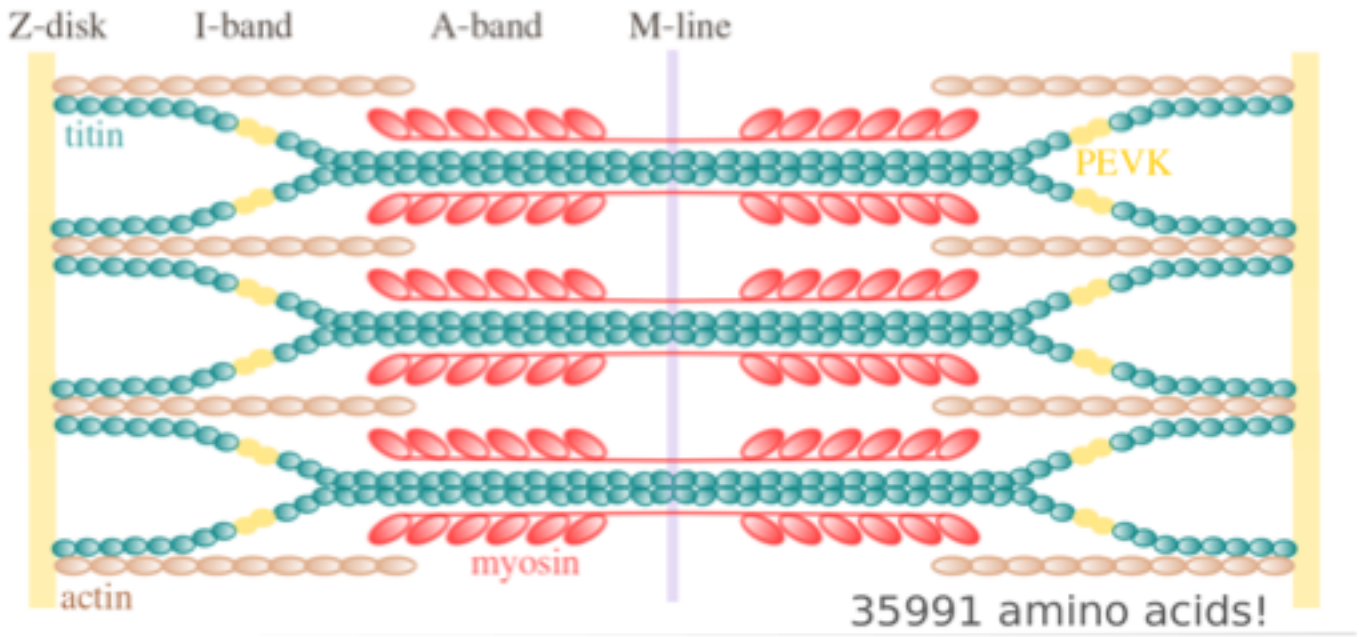
Circulation. 2013 February 26; 127(8): 938–944. doi:10.1161/CIRCULATIONAHA.112.139717.

Titin is a major human disease gene

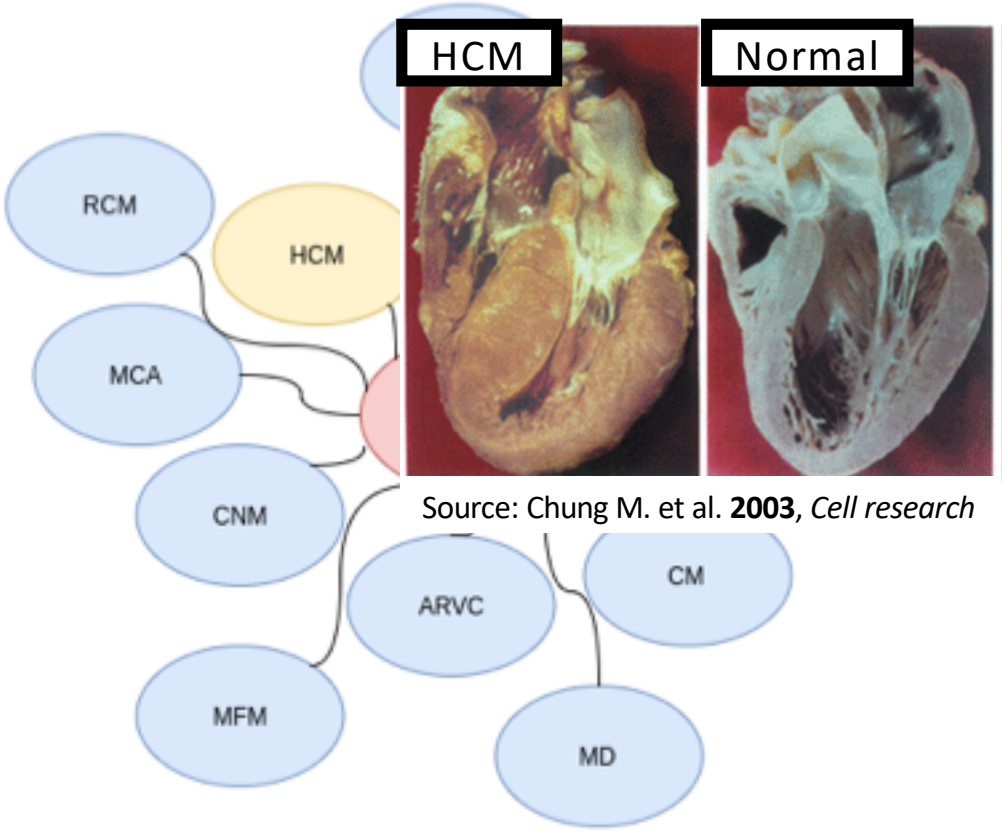
Martin M. LeWinter, M.D.[1] and **Henk L. Granzier, Ph.D.**[2]

[1]Cardiology Unit, Fletcher Allen Health Care, Burlington, VT

[2]Sarver Molecular Cardiovascular Research Program and Department of Physiology, University of Arizona, Tucson, AZ



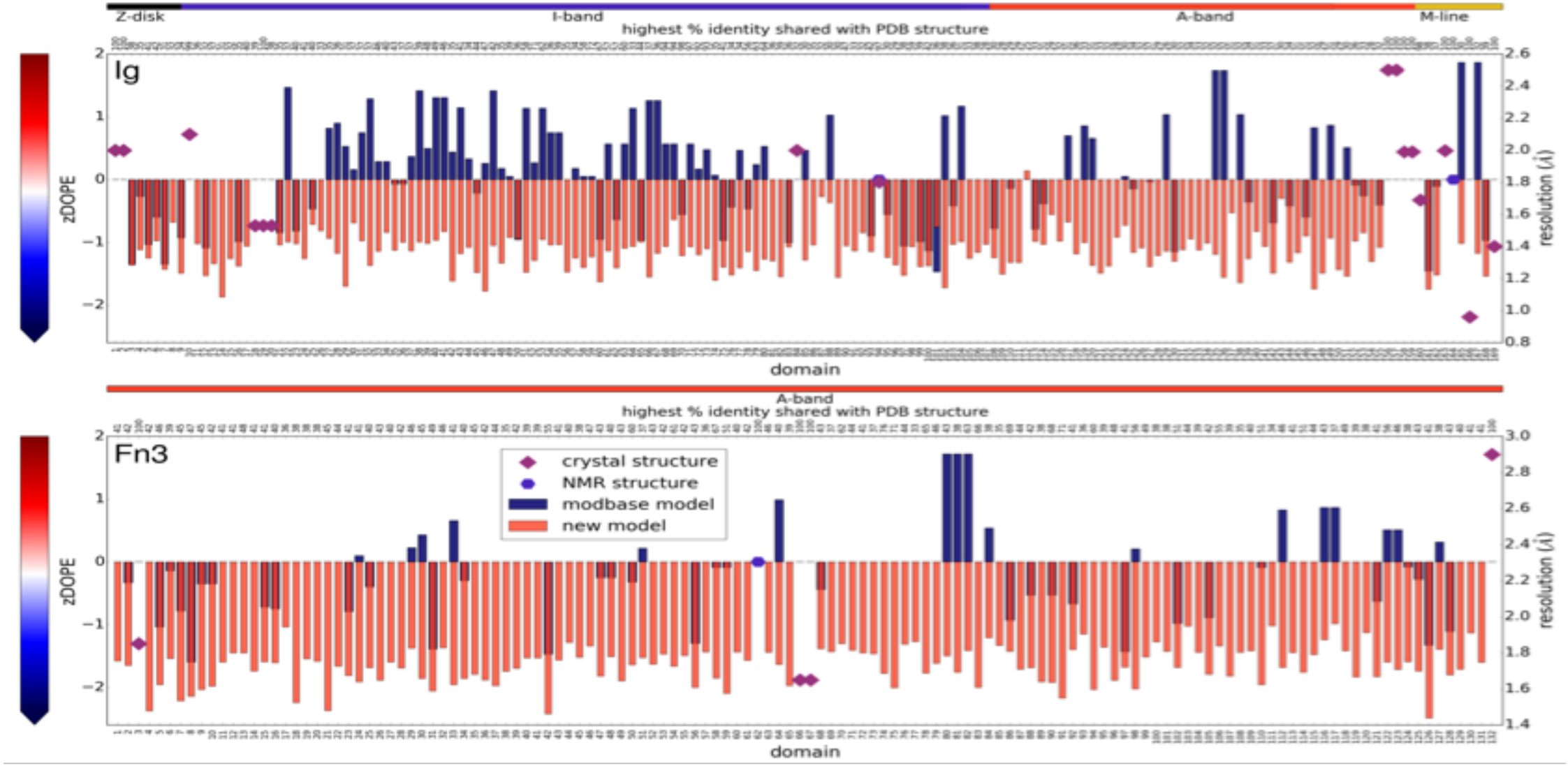
- Titin, the largest protein, spans half a cardiac sarcomere.
- Roles: scaffold, spring, signalling.

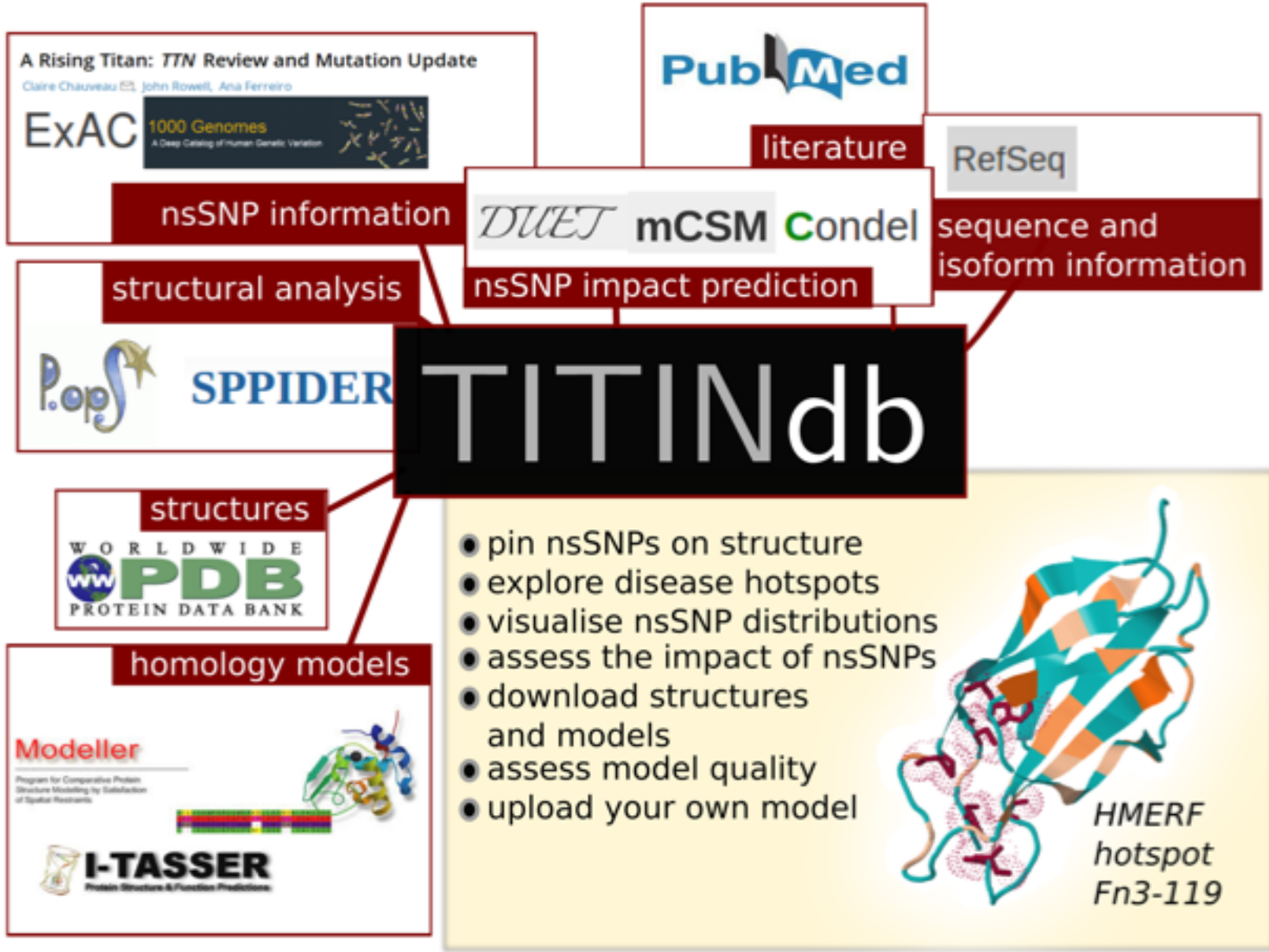


Source: Chung M. et al. **2003**, *Cell research*

- Titin missense variants associated with myopathies.
- Due to titin's large size, the majority of healthy individuals possess rare titin missense variants.
- **This results in the paradox that rare titin variants are commonly found!**

Increase in structural coverage and model quality





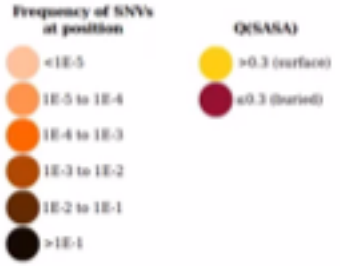
http://fraternalilab.kcl.ac.uk/TITINdb/ **TITINdb—a computational tool to assess titin's role as a disease gene**
Anna Laddach Mathias Gautel Franca Fraternali *Bioinformatics*, btx424. https://doi.org/10.1093/bioinformatics/btx424

TITINdb Home Search Tutorials Documentation Downloads

Fn3-1 | Search by domain

gnomAD SNVs | 1000 genome SNVs | disease SNVs | reset

show van der Waals | show spacefill | save PNG | save PDB



Export SNV table to CSV

show saturation mutagenesis table, this may take a few moments to load...

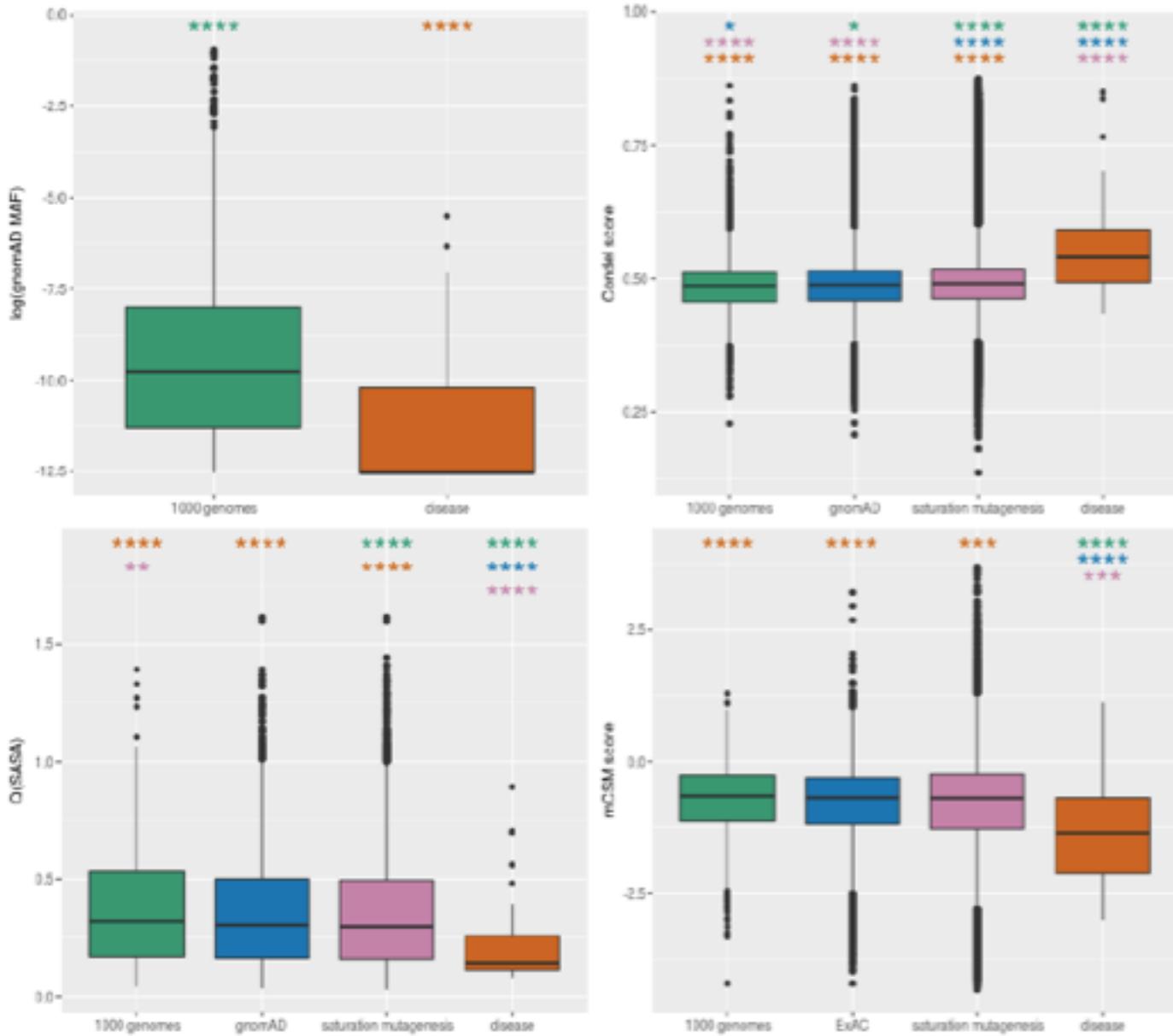
JSmol

SPGPCGKLTVSRVTQEKCTLAWSLPQEDGGAEITHYIVERRETSRLNWVIVEGECPTLSYVVTRLIKNNEYIFRVRAVNKYGPGVPVESEPI

SNV table

Pin SNV	Position	Domain position	SNV	RS	Disease	DUET	PolyPhen-2	Condel	Q(SASA)	Source	MAF
	31709	2	P/R	None	HMERF	-0.93	0.146	D	0.1489	Palmio (2013)	None
	31710	3	G/S	None	None	-1.127	1.0	N	0.1842	gnomAD	4.20748E-06
	31712	5	C/R	None	MFM/HMERF,HMERF	-1.28	0.9	N	0.1205	Pfeffer (2012) Ohlsson (2012) Toro (2013) Palmio (2013) Pfeffer (2014) Uruha (2015) Yue (2015)	None
	31712	5	C/R	None	None	-1.28	0.9	N	0.1205	gnomAD	4.16354E-06
	31712	5	C/Y	None	HMERF	-1.682	0.93	N	0.1205	Uruha (2015)	None
	31717	10	V/I	rs150930737	None	-0.378	0.004	N	0.3028	1000 genomes	1.99681E-04

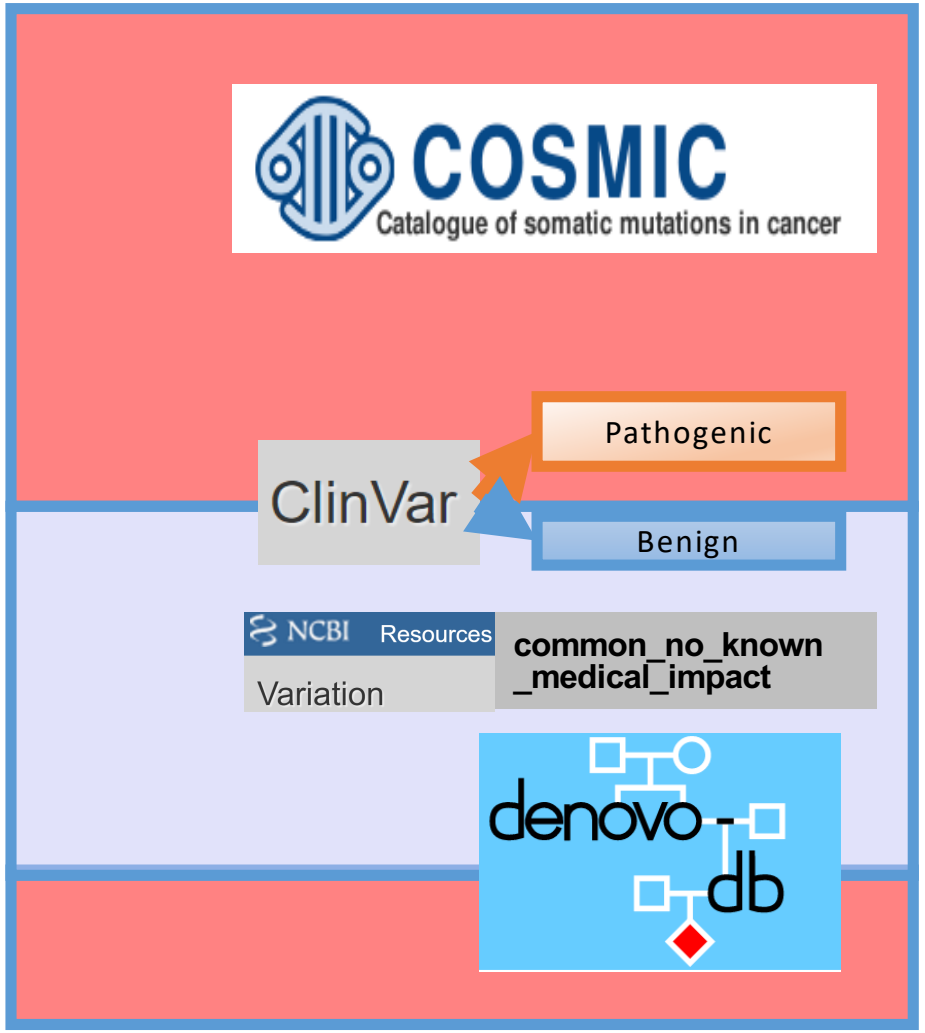
Comparison of properties across datasets



Whilst working on this…..

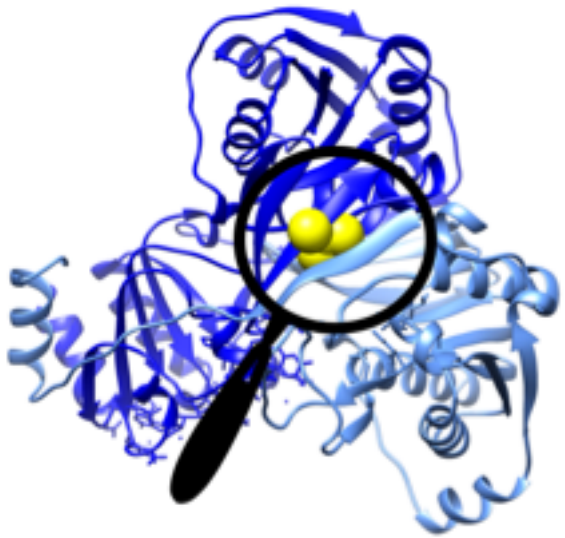
- A huge amount of common variant data (also nominally healthy individuals) became available.
- This has challenged preconceptions about variant associations with disease.
- Can this improve our understanding of variants in health and disease?
- Can we use this information to predict which variants are deleterious?

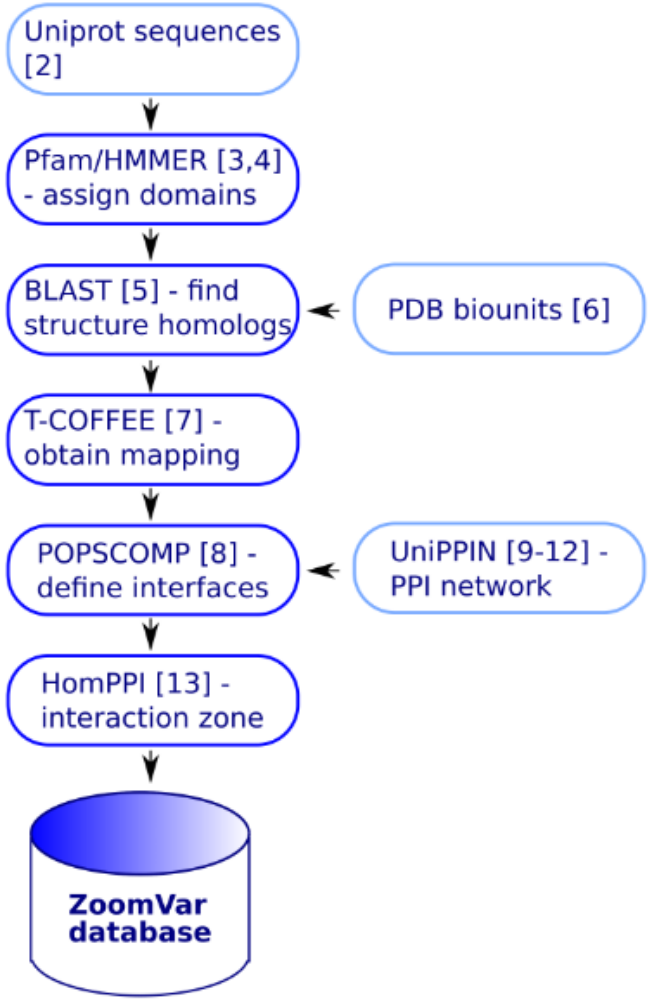
Mapping Genetic Variation to Proteins: common vs disease



COSMIC, the Catalogue Of Somatic Mutations In Cancer, is the world's largest and most comprehensive resource for exploring the impact of somatic mutations in human cancer
ClinVar-Pathogenic * ClinVar aggregates information about genomic variation and its relationship to human health.
ClinVar-Benign
An up-to-date report of common nsSNPs not known to cause clinical phenotypes.
A collection of Germline *de novo* variants. Variants which are present in children but not their parents. Some of these variants are known to be pathogenic.

ZoomVar database - http://fraternalilab.kcl.ac.uk/ZoomVar/

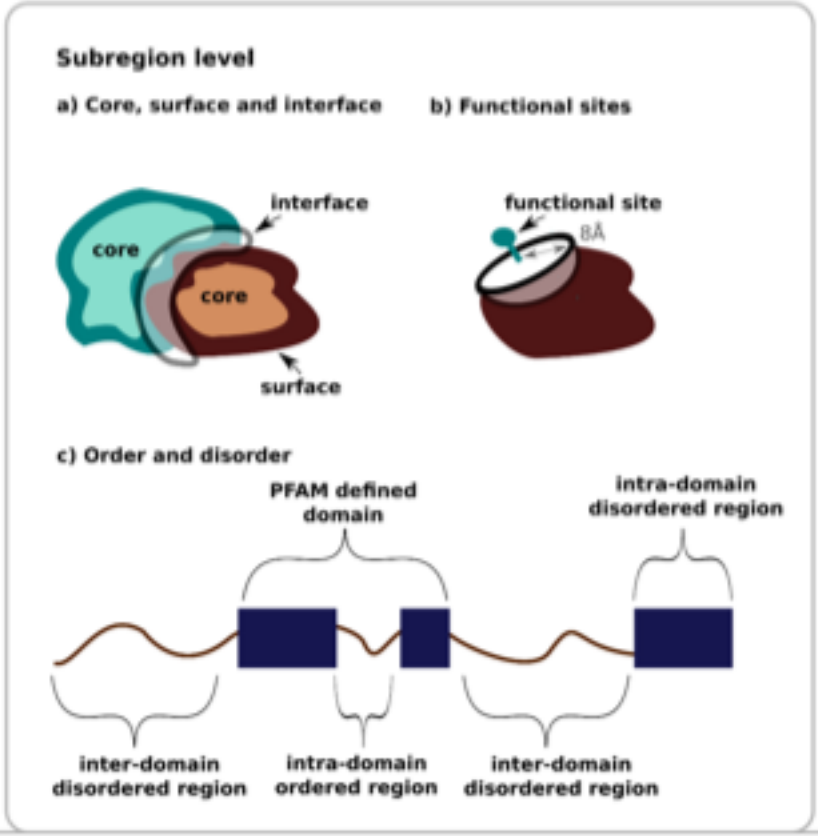


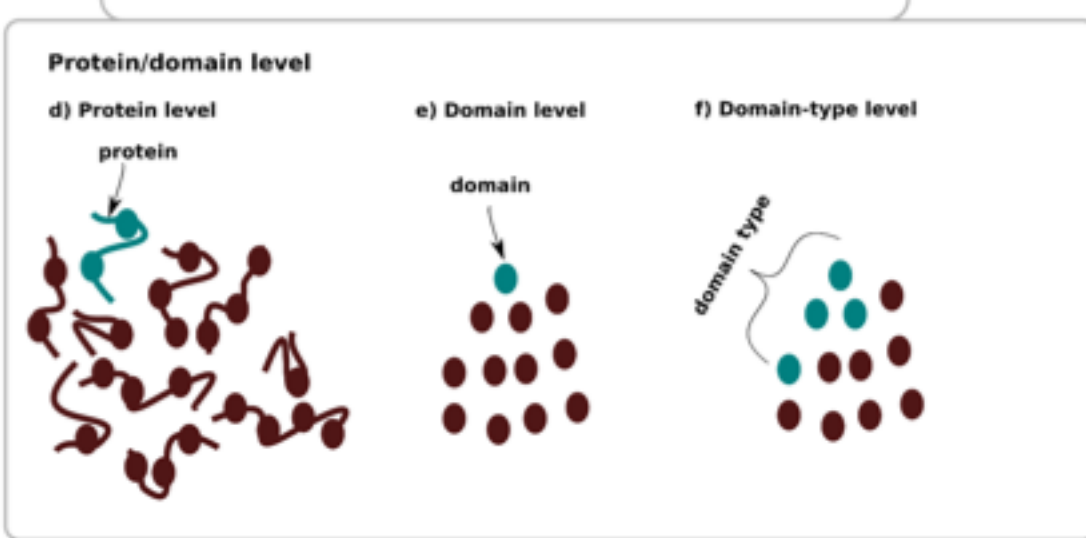


Data summary

	gnomAD common MAF > 0.05*	gnomAD rare MAF < 0.005	COSMIC	ClinVar
SAVs	21358	3860943	1731030	21272
Proteins	17048	17048	16679	5594
Proteins with SAVs and core coverage	3050	10487	10433	1661
SAVs core	1002	303311	152356	5194
Proteins with SAVs and interact coverage	818	3531	3561	677
SAVs interact	157	38315	22205	768
Proteins with SAVs and surface coverage	3092	10797	10717	1703
SAVs surface	4723	990915	491179	8558

* In at least one gnomAD population





Investigating the enrichment of individual proteins/domains:

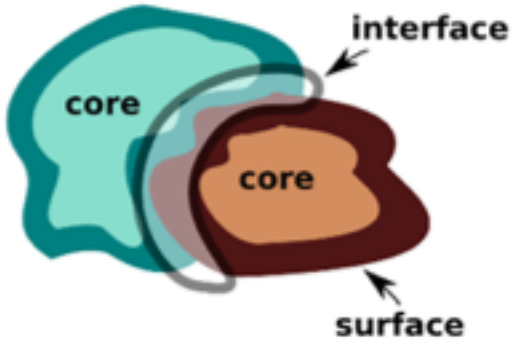
$$P(N(SNVs_{region}) = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

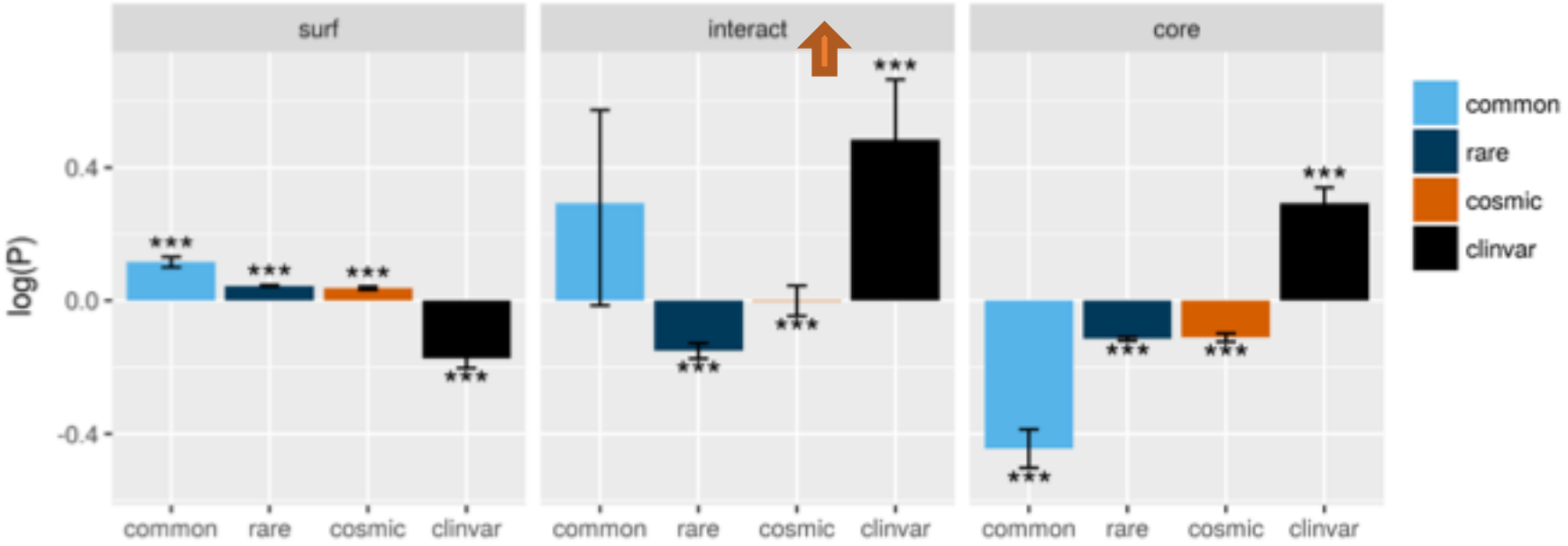
Investigating general trends:

$$P(SNV_{region}) = \frac{(N(SNVs)_{region}/size_{region})}{(N(SNVs)_{protein}/size_{protein})}$$

Bootstrap to obtain confidence interval

Core, Surface and Interface





Functional analysis

FGSEA package – binomial CDF as the enrichment statistic – performed at the whole protein and sub-protein levels.

↓

Normalised Enrichment Statistic and significance of enrichment obtained.

↓

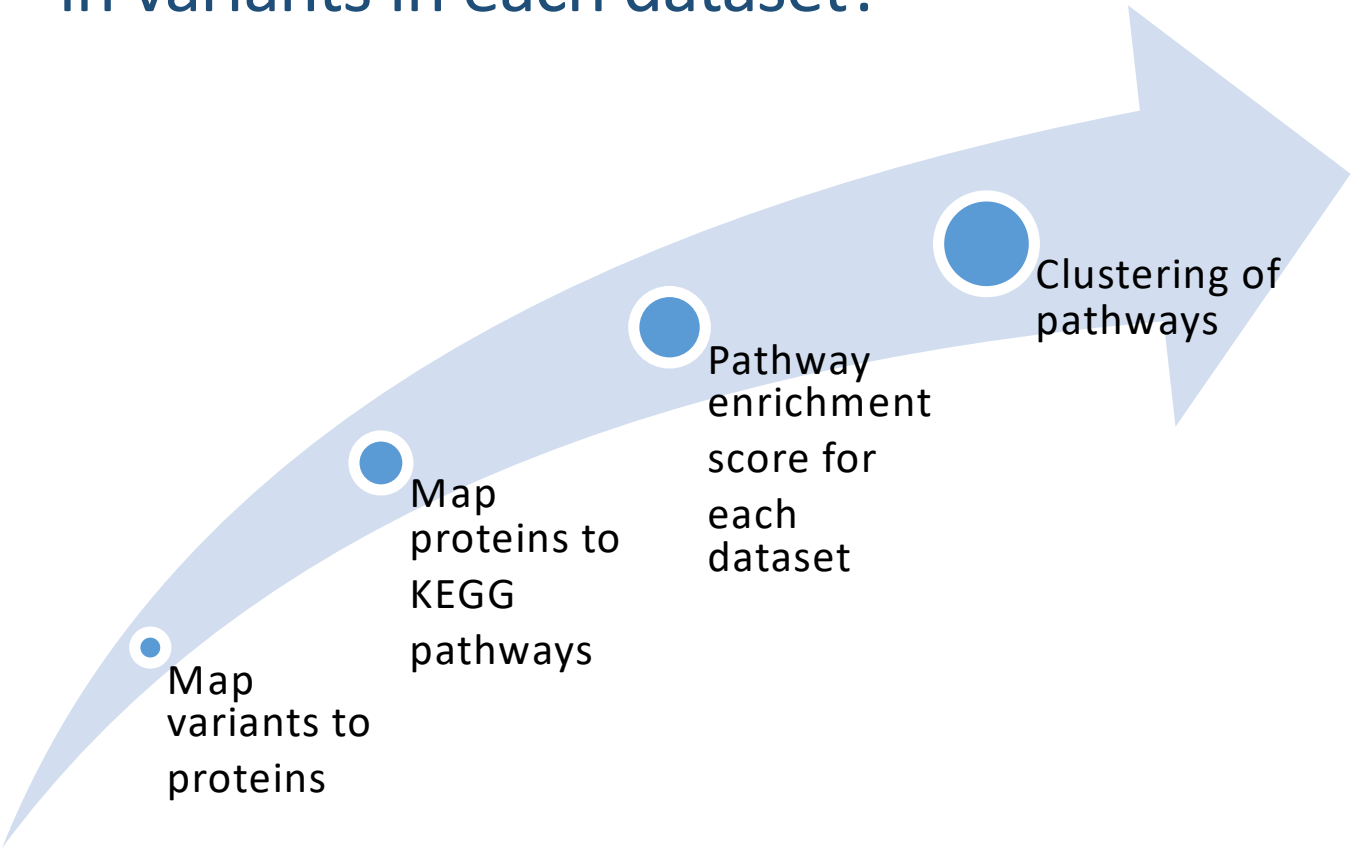
K-means clustering of Normalised Enrichment Statistic at the whole protein level.

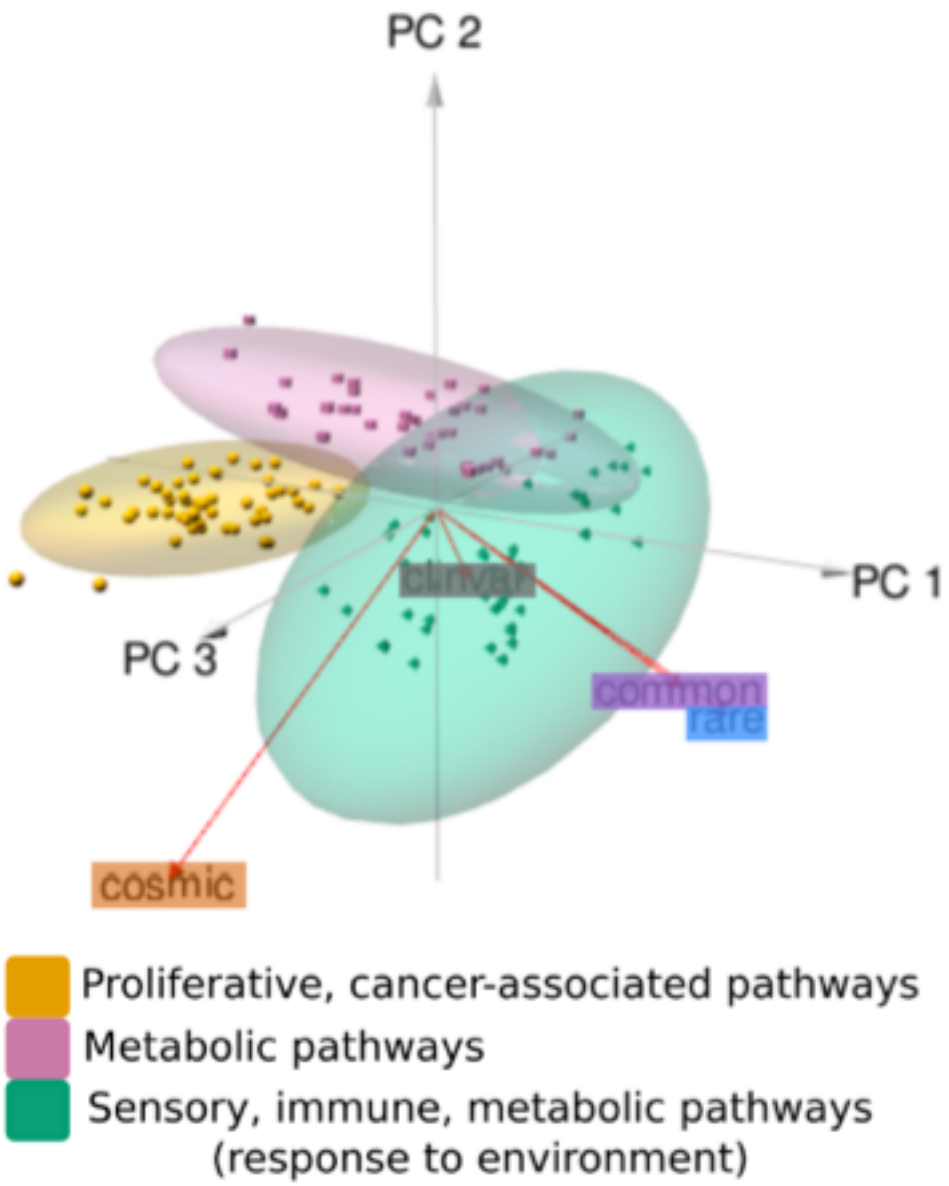
↓

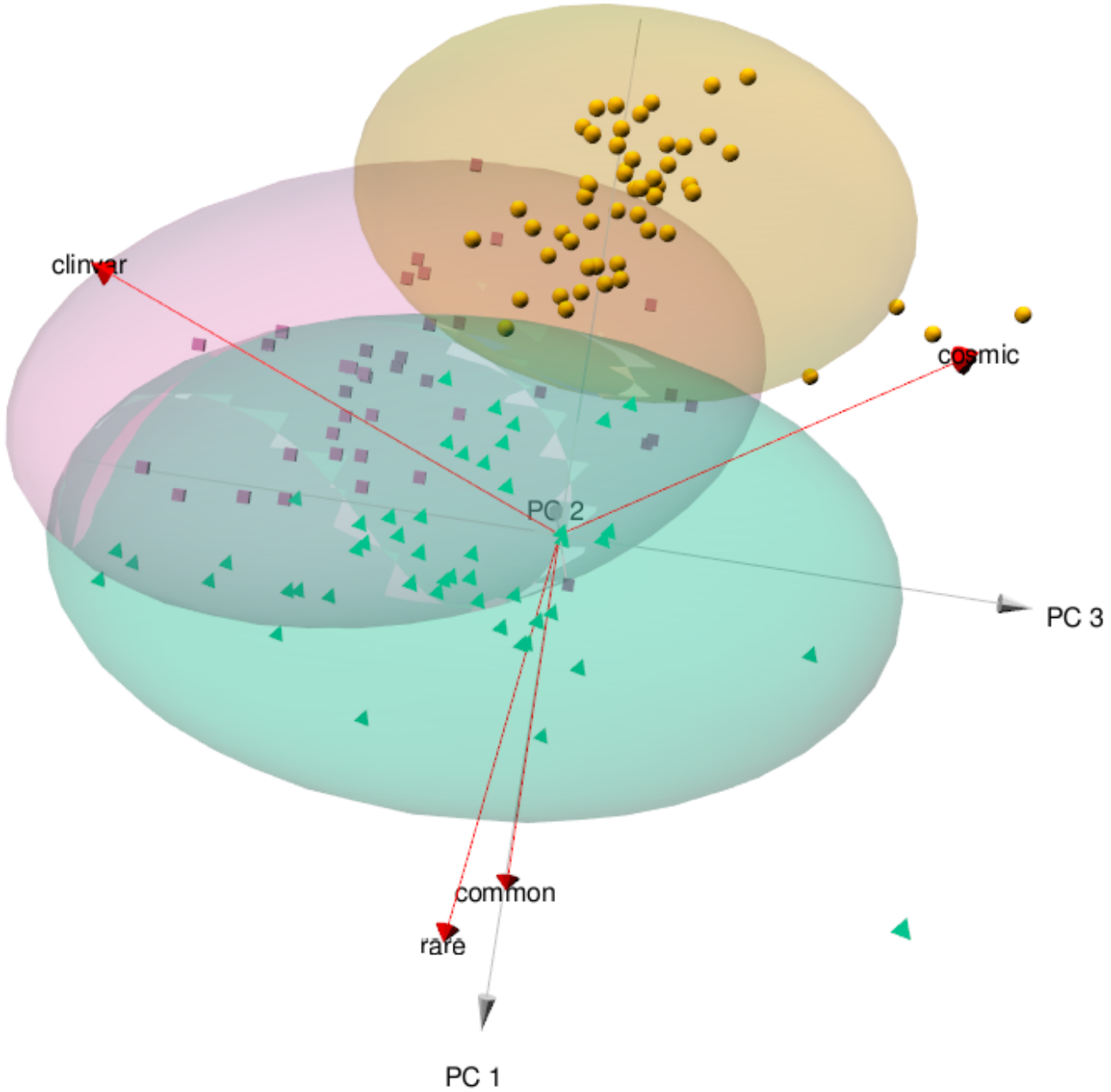
Map analysis at the sub-region level to K-means clusters at the whole protein level.

Missense variants target distinct functional pathways in health and disease

Which functional pathways are enriched in variants in each dataset?



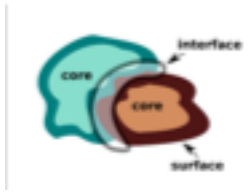


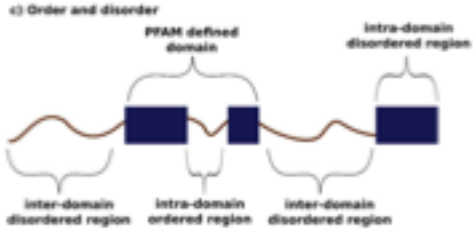


ted pathways

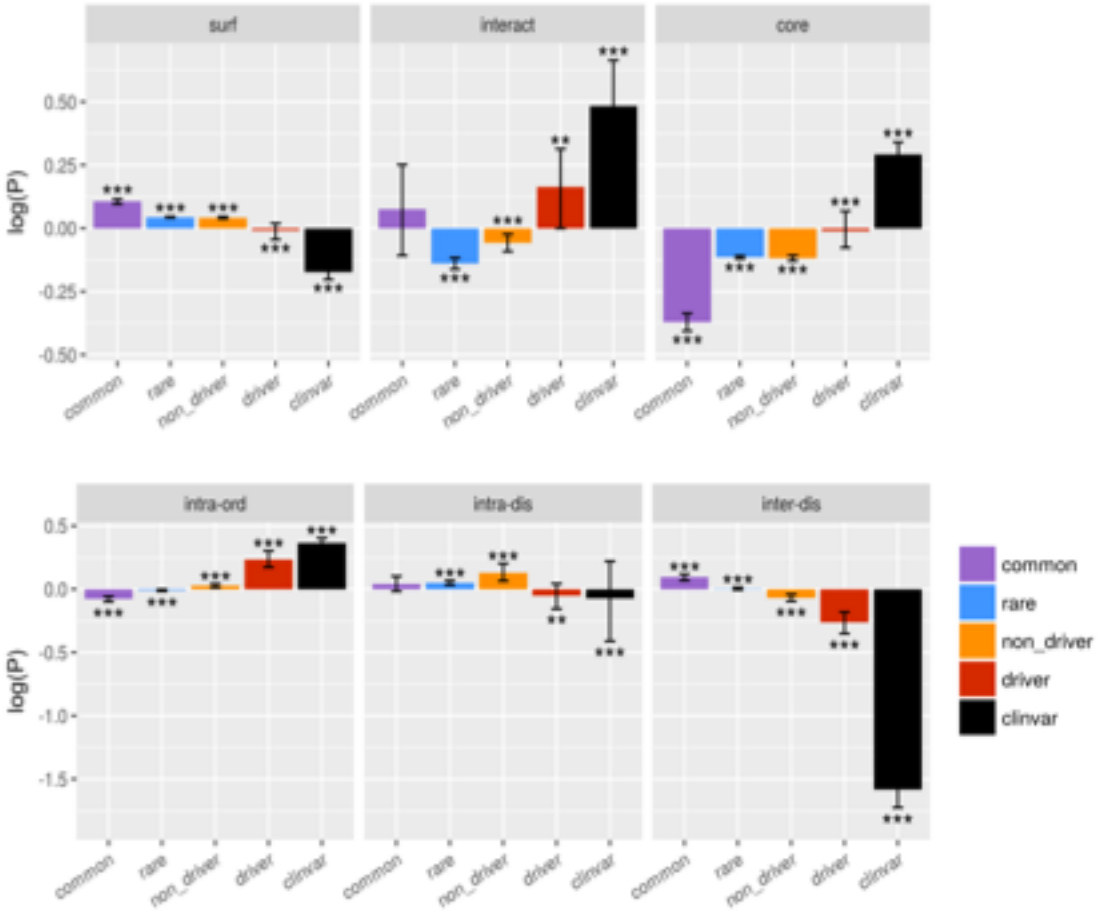
c pathways
nent)

PC2 (29.0% explained var.)

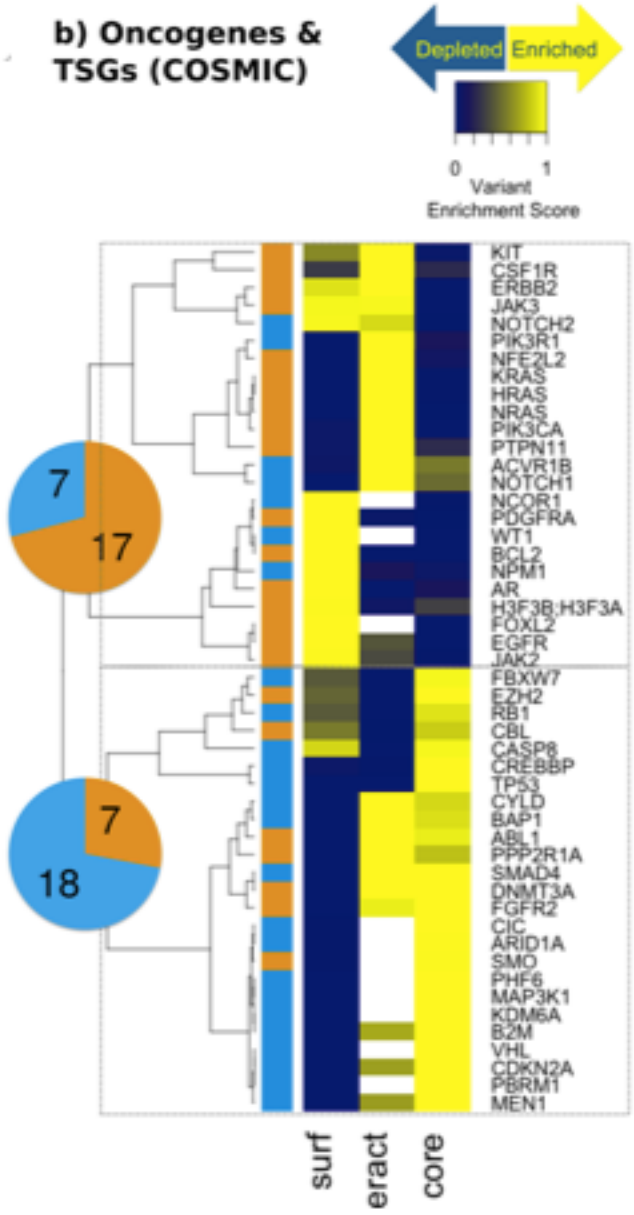




a) Region enrichment



b) Oncogenes & TSGs (COSMIC)



TITINrf: A Titin variant impact predictor

Class	Description
Healthy	49 common nsSNVs from the 1000 genomes project (MAF > 0.02)
Disease	45 SNVs in total. 41 titin disease nsSNVs from the literature, 4 unpublished SNVs known to be disease causing

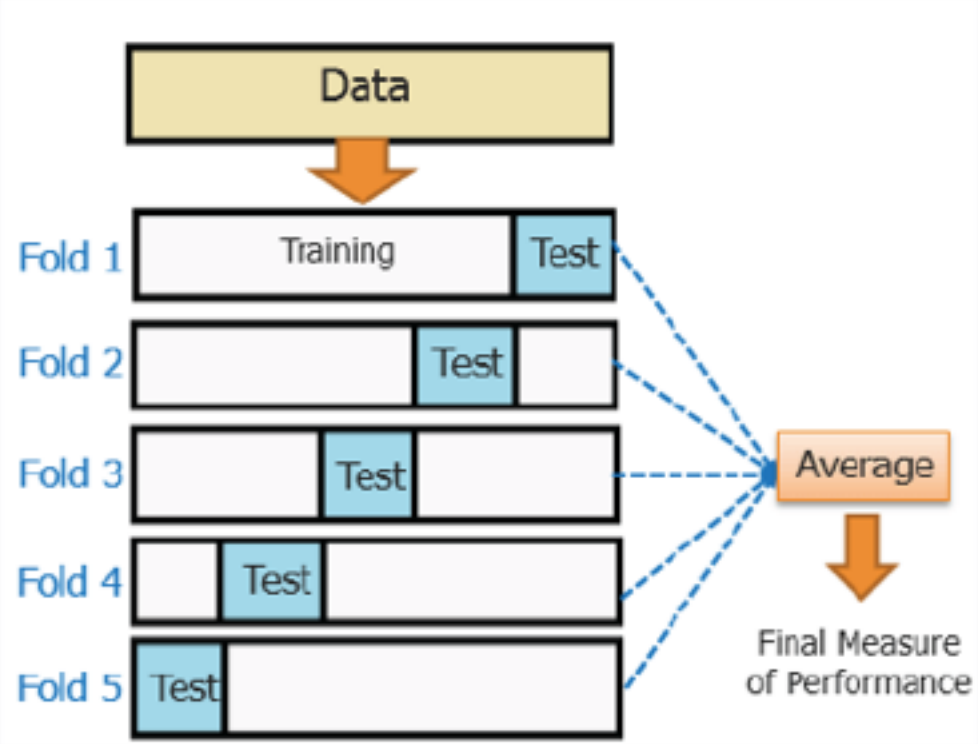


Figure 3: source:edureka.co

Network

- degree centrality
- node centrality
- betweeness centrality
- load centrality
- neighbour centrality

Dynamics

- mean squared fluctuations
- sensor/ef-fector
- mechanical stiffness

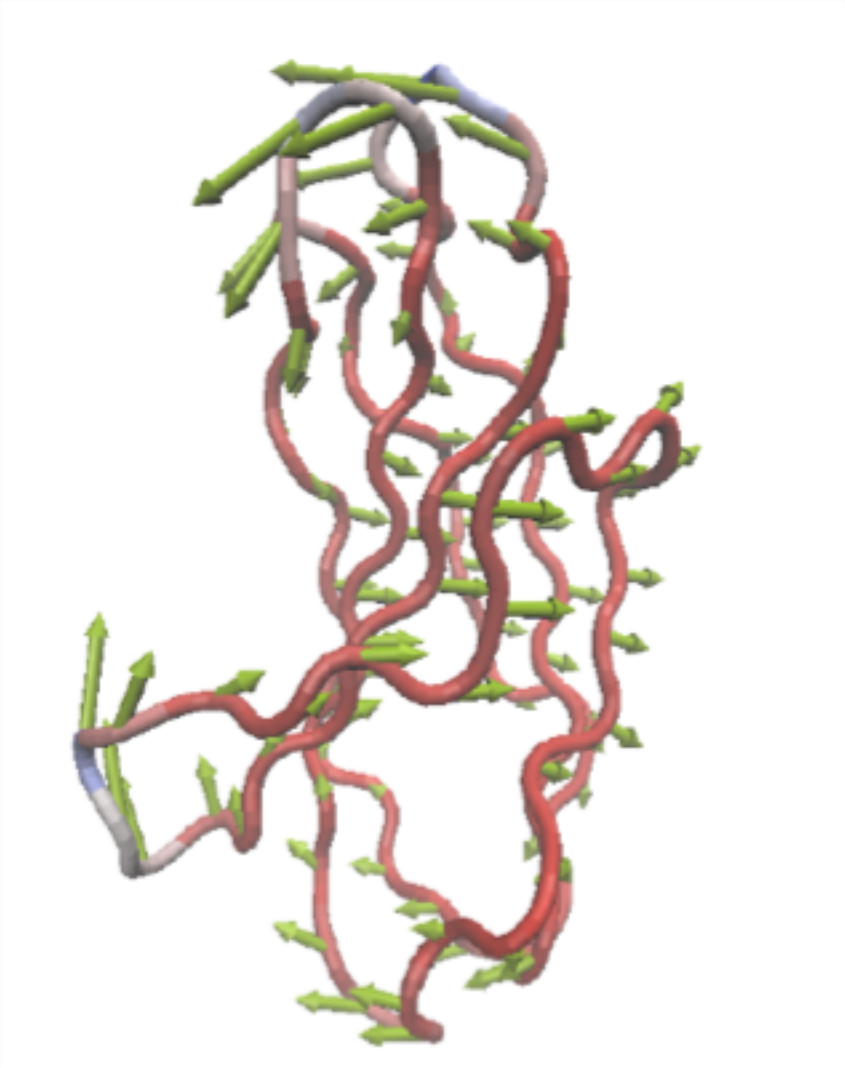
Structure

- SASA
- residue density

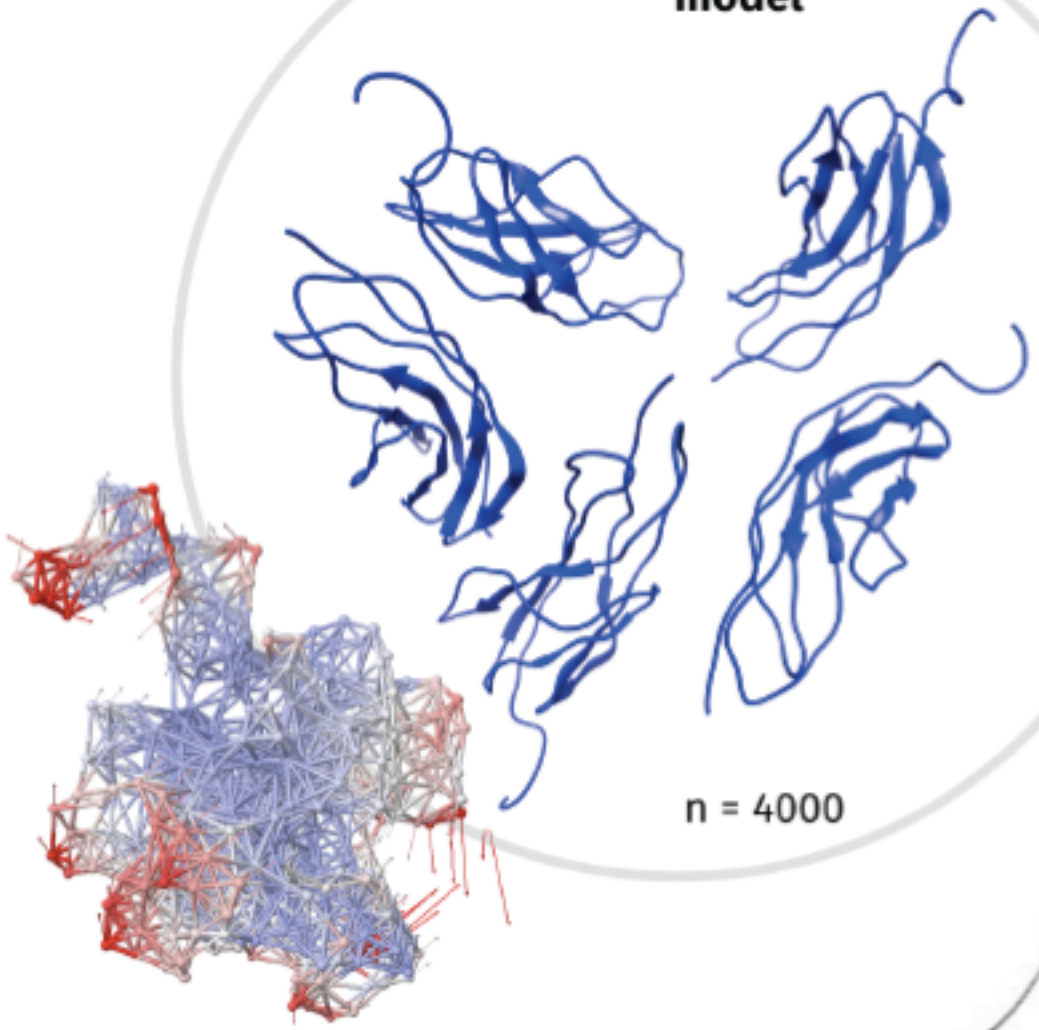
Sequence

- PSIC score
- Kidera factors

- C alpha atoms represented by nodes connected by elastic springs
- course-grained, computationally inexpensive
- can calculate on a large scale

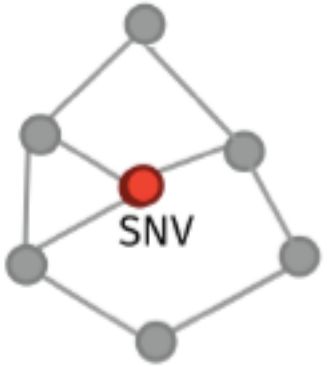


Conformational ensemble generated from elastic network model



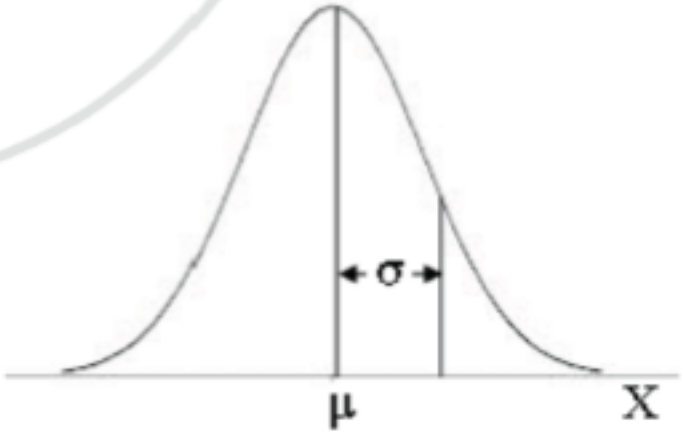
Network properties calculated

- Degree centrality
- Load centrality
- Betweeness centrality



Distribution for each feature

- Mean
- Standard deviation
- Maximum
- Minimum



predictor	accuracy	precision	recall	F1	MMC
TITINrf	0.80	0.84	0.80	0.81	0.61
Condel	0.73	0.75	0.65	0.69	0.46

oob score of final model =0.82

Precision
Out of all SAVs predicted to be disease-associated how many are actually disease-associated?

Recall
Out of all disease-associated SAVs how many are correctly predicted?

F1 score

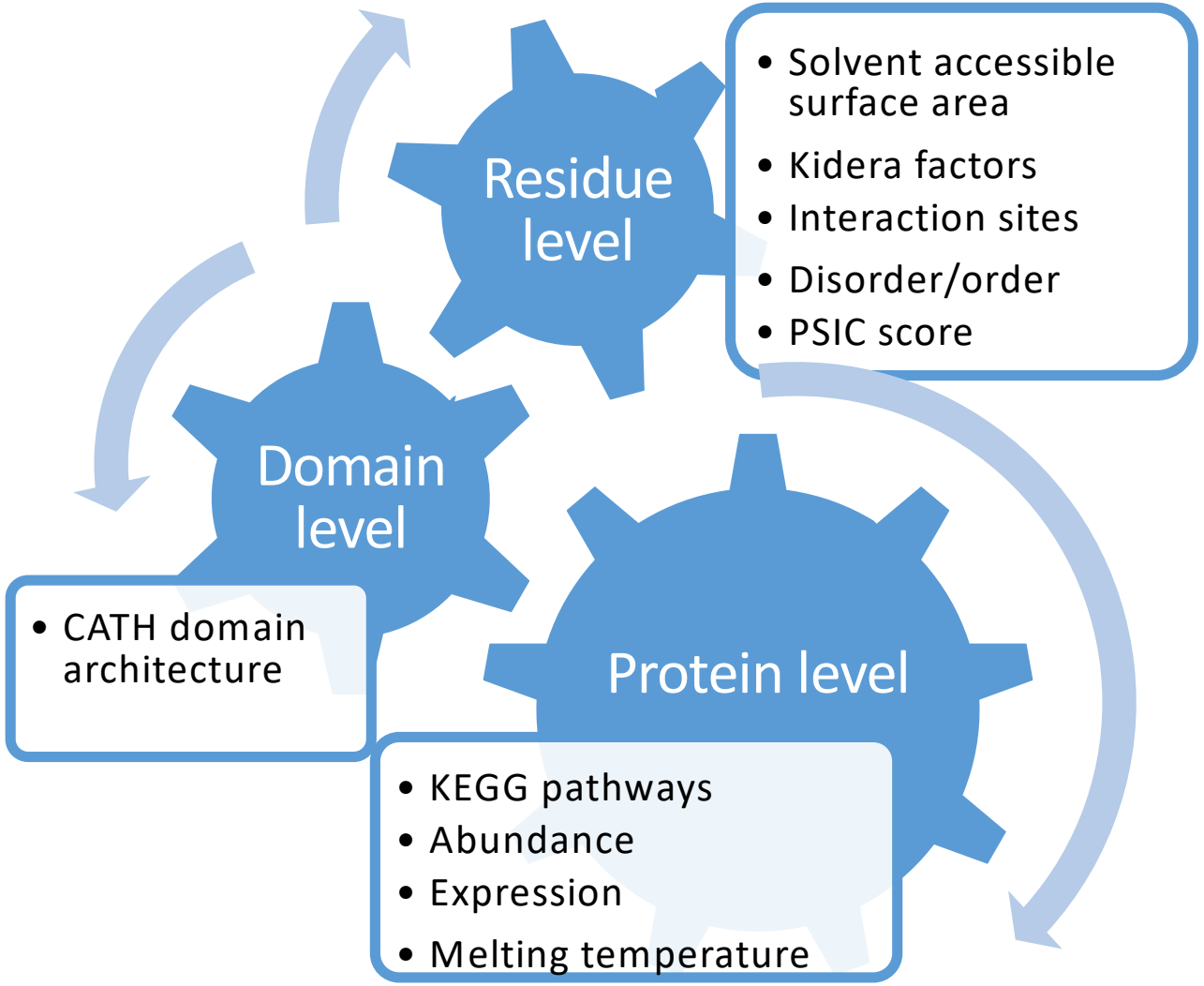
$$2 \cdot \frac{precision \cdot recall}{precision + recall}$$

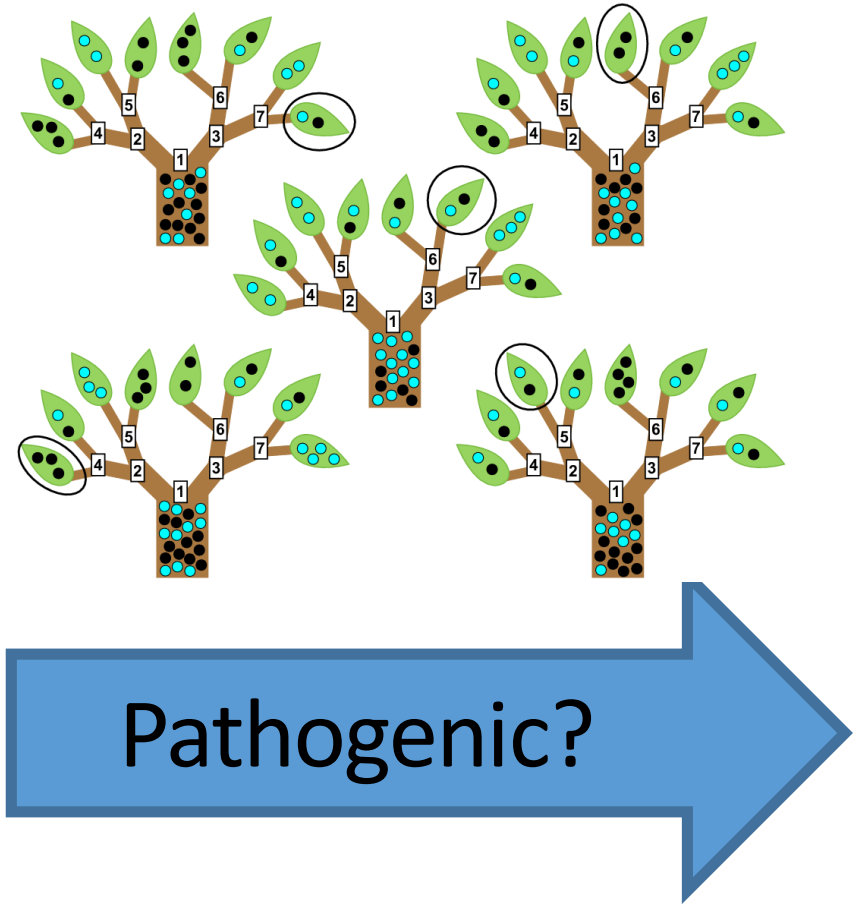
MCC

$$\frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP+FP)(TP+FN)(TN+FP)(TN+FN)}}$$

		Predicted	
		Deleterious	Neutral
Actual	Deleterious	TP	FN
	Neutral	FP	TN

Currently using features from this analysis to create ZoomVar predictor

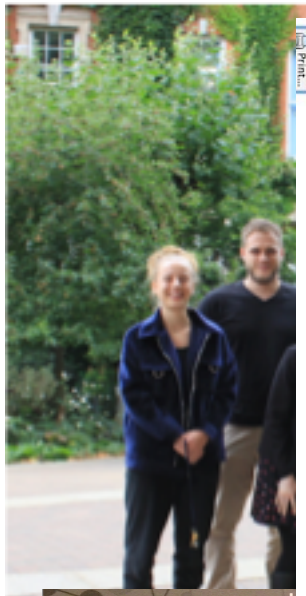


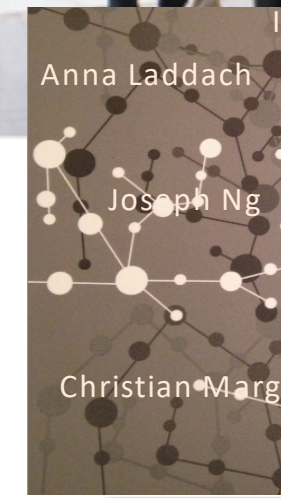


Random forest cartoon source: http://inspirehep.net

Conclusion

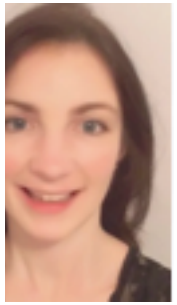
- The created ZoomVar database has allowed us to investigate features of variants the general population (gnomAD) and disease (ClinVar).
- We find features which clearly segregate population and disease-associated variants.
- This will enable us to create a random-forest based predictor of variant impact which uses these novel features.





...

ook Chung



na Laddach

Luis Fernandes KCL

Prof. Tony Ng Cancer Studies KCL
Prof. Thomas Shaun Haematological Medicine KCL
Prof.Mathias Gautel Randall KCL

Jens Kleinjung
Crick Institute London
Heptares Pharmaceuticals

MRC

EPS

BEATING BLOOD CANCER

Bloodwise

LEUKAEMIA & LYMPHOMA RESEARCH

THE FRANCIS CRICK INSTITUTE

KING'S College LONDON

bbsrc
biotechnology
research coun