



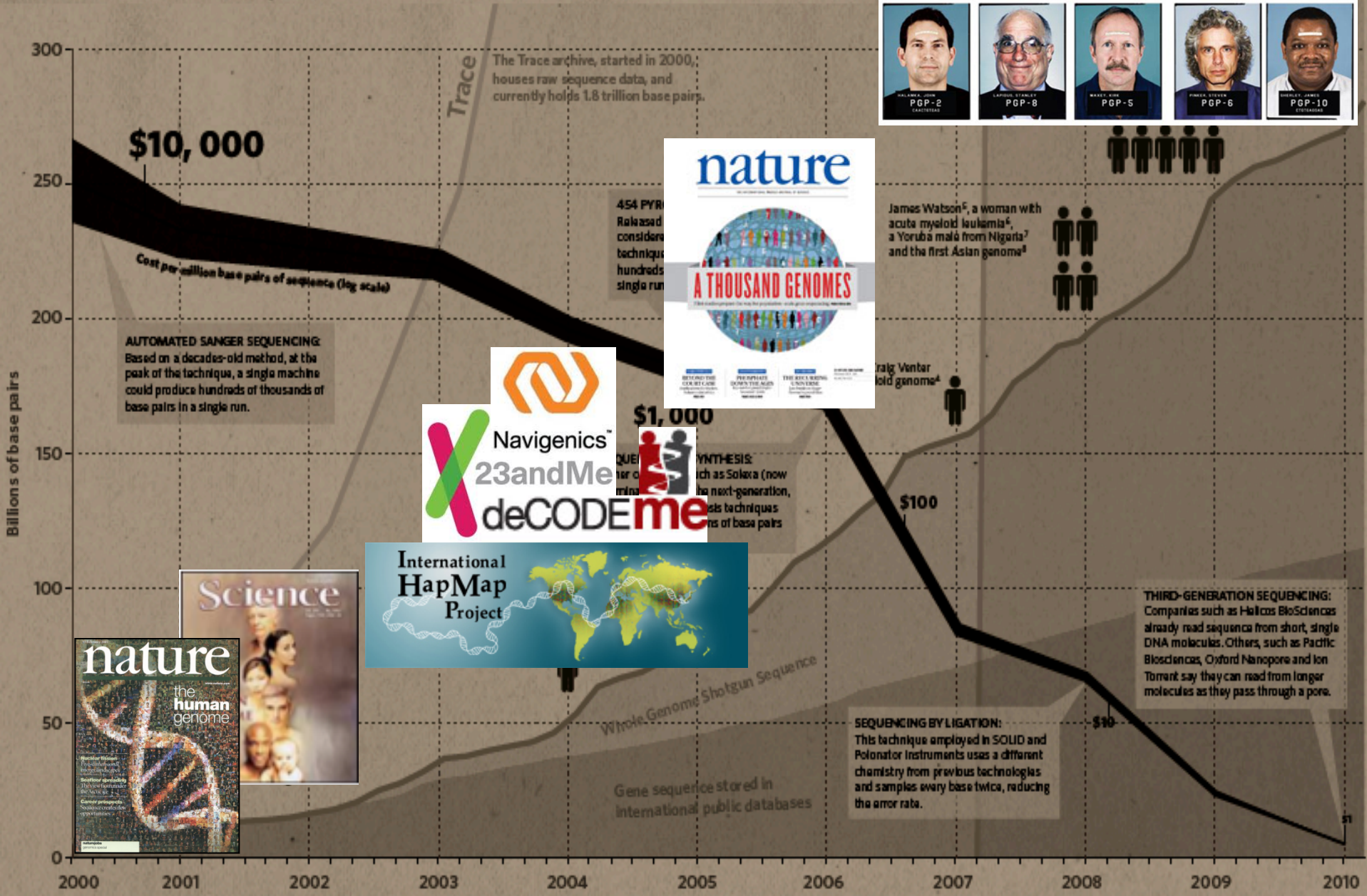
Analysis of Protein Networks:
**Using 3D-structure
into interpret
networks & deep-
sequencing data**

Slides freely downloadable from
Lectures.GersteinLab.org
& “tweetable” (via @markgerstein)

See Last Slide for References & More Info.

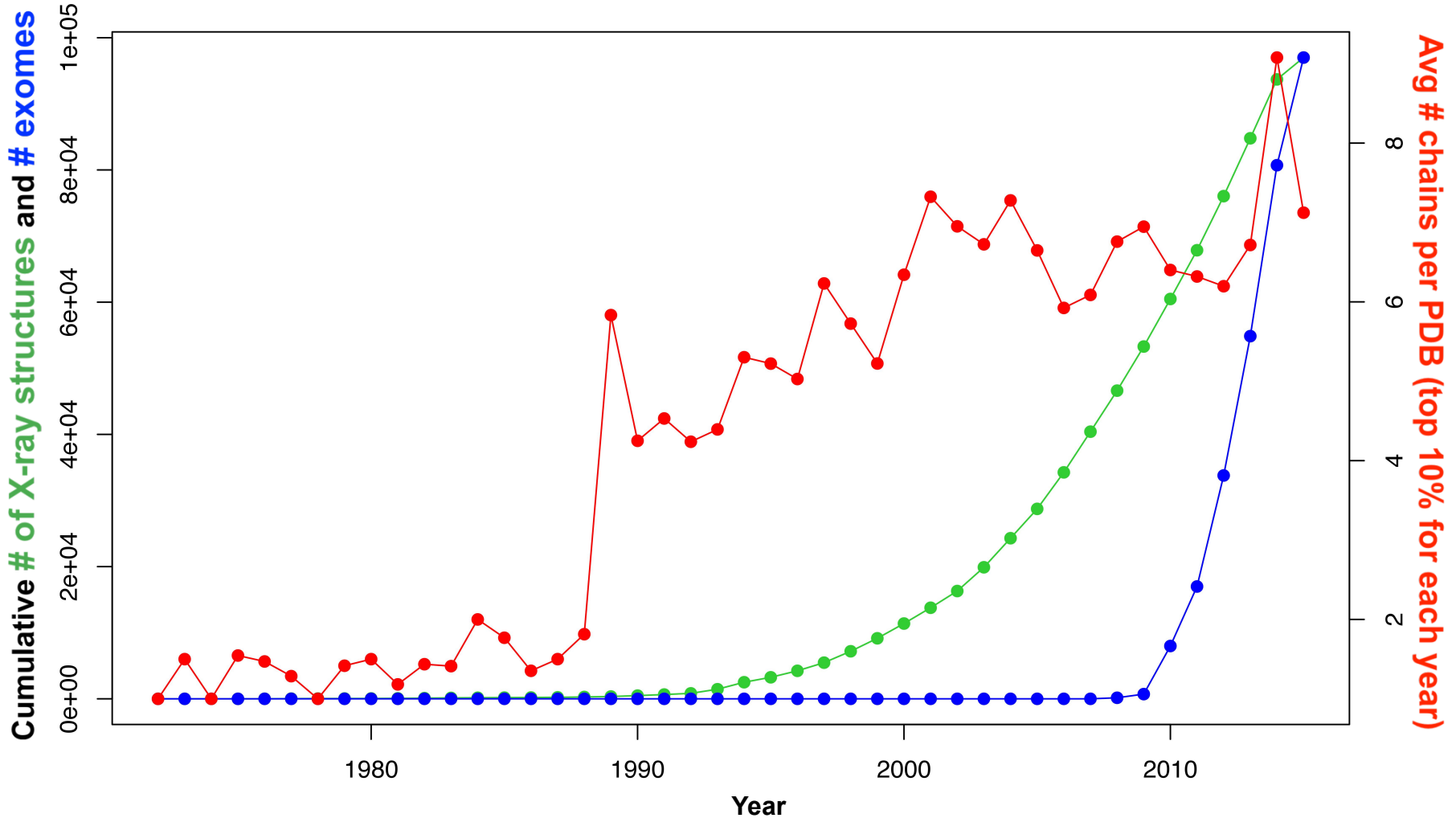
Mark B Gerstein
Yale

THE SEQUENCE EXPLOSION



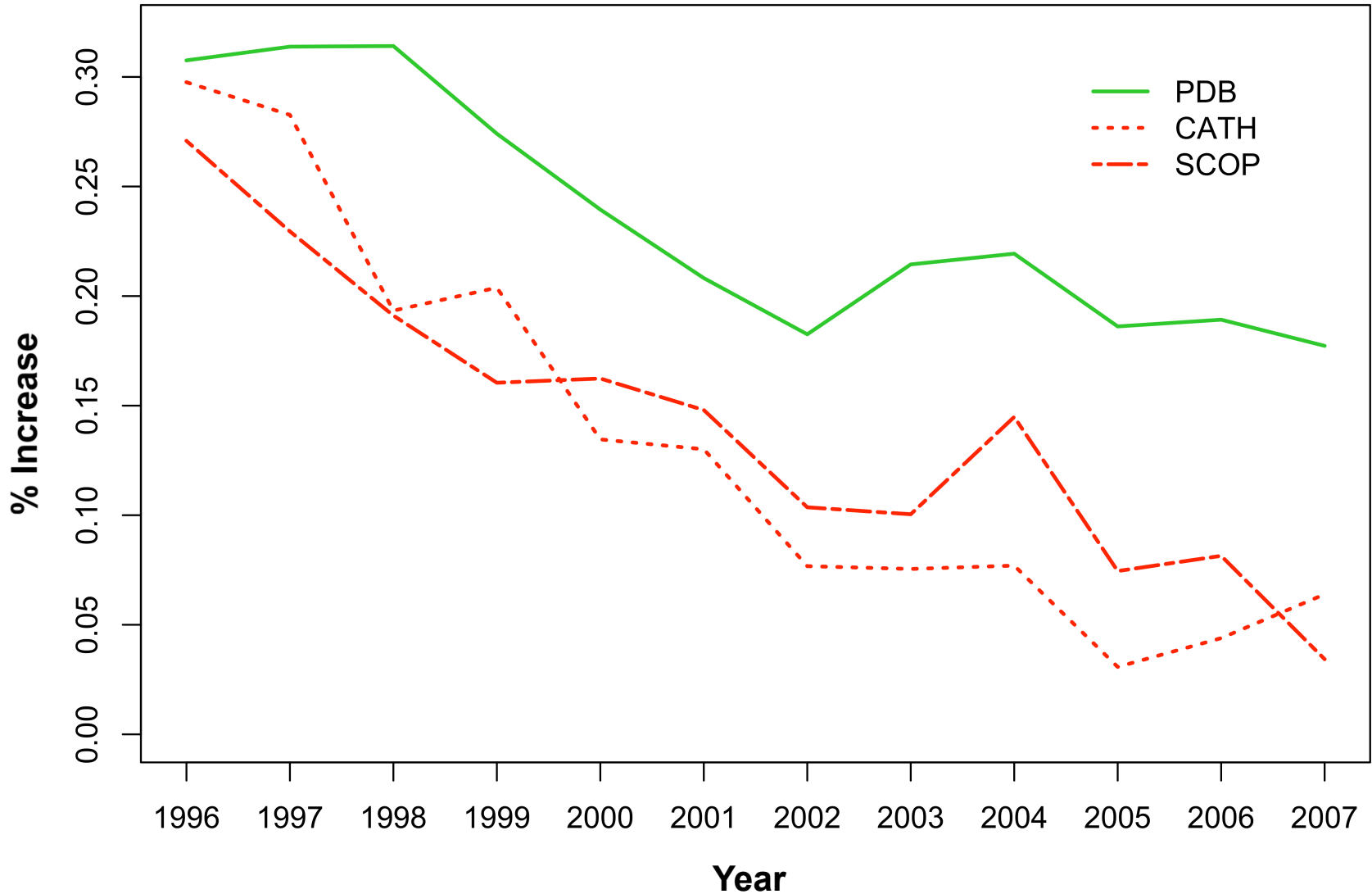
Trends in data generation point to growing opportunities for leveraging sequence variants to study structure (and vice versa)

The volume of sequenced exomes is outpacing that of structures, while solved structures have become more complex in nature.



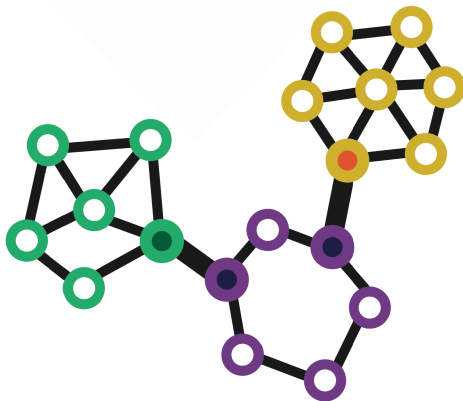
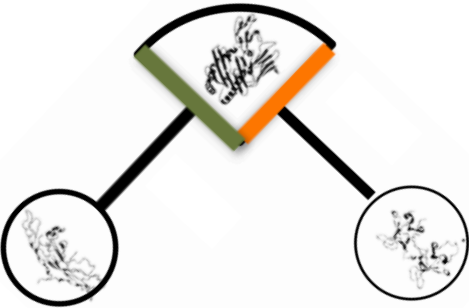
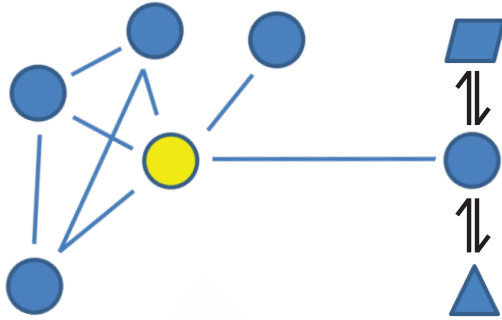
Exome data hosted on NCBI Sequence Read Archive (SRA)

Growing sequence redundancy in the PDB (as evidenced by a reduced pace of novel fold discovery) offers a more comprehensive view of how such sequences occupy conformational landscapes



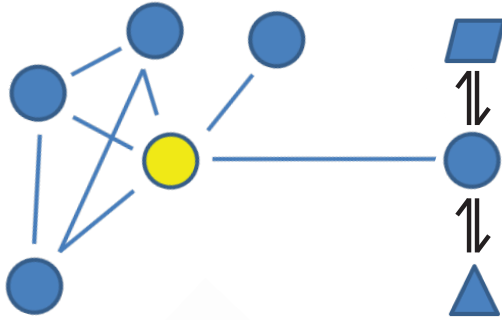
PDB: Berman HM, et al. NAR. (2000)
CATH: Sillitoe I, et al. NAR. (2015)
SCOP: Fox NK et al. NAR. (2014)

Using 3D-structure
into interpret
networks & deep-
sequencing data



- Structural Interaction Network & Protein Motions (DynaSIN)
 - Multi-interface permanent hubs have more motion than single-interface transient ones
 - Also have more conflicting motions
- LOF variants & Categories of Essential & Disease-sensitive Genes
- Variation at Protein Interfaces in the context of Network Connectivity & its use for Disease-gene Predictions
 - Highly connected parts of PPI under stronger selection but signal weak
 - Stronger signal in SIN & even stronger in multiNet (integration of many networks)
 - Signal strong enough to build predictor
- Rationalizing Deleterious Variants in terms of Potential Allosteric Sites
 - Identifying potential allosteric residues on surface & inside
 - These are under stronger selection & may explain some HGMD SNPs

Using 3D-structure
into interpret
networks & deep-
sequencing data



- Structural Interaction Network & Protein Motions (DynaSIN)

- Multi-interface permanent hubs have more motion than single-interface transient ones
- Also have more conflicting motions

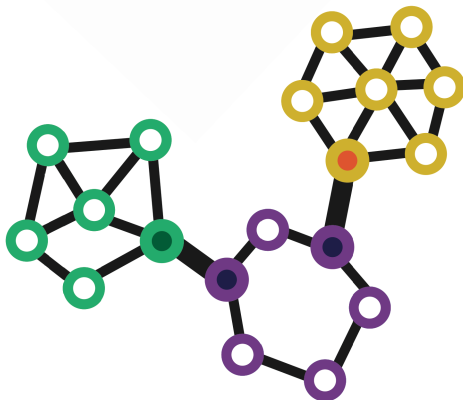
- LOF variants & Categories of Essential & Disease-sensitive Genes

- Variation at Protein Interfaces in the context of Network Connectivity & its use for Disease-gene Predictions

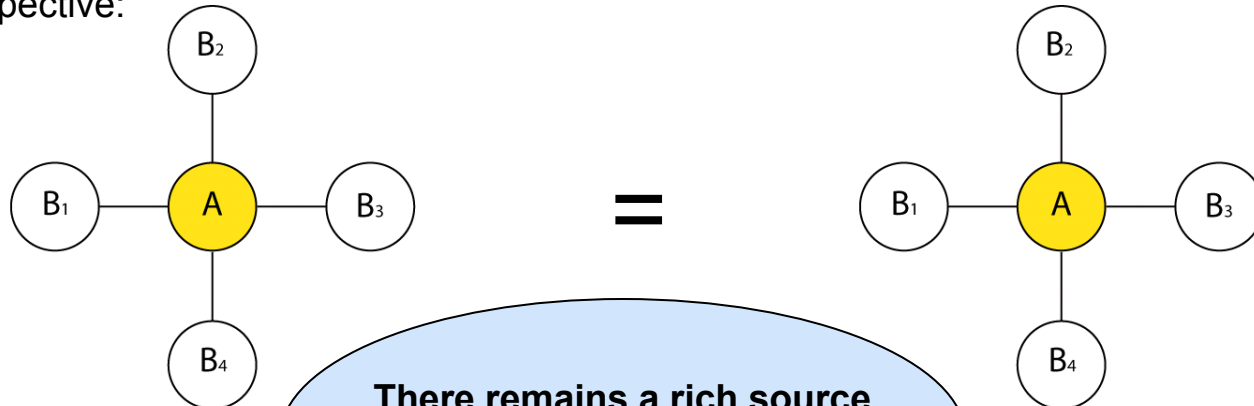
- Highly connected parts of PPI under stronger selection but signal weak
- Stronger signal in SIN & even stronger in multiNet (integration of many networks)
- Signal strong enough to build predictor

- Rationalizing Deleterious Variants in terms of Potential Allosteric Sites

- Identifying potential allosteric residues on surface & inside
- These are under stronger selection & may explain some HGMD SNPs

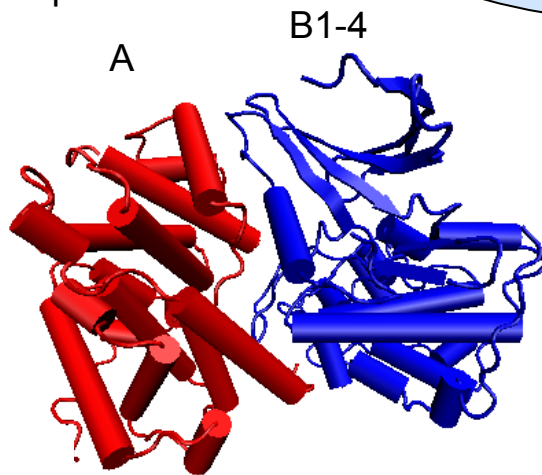


Network perspective:



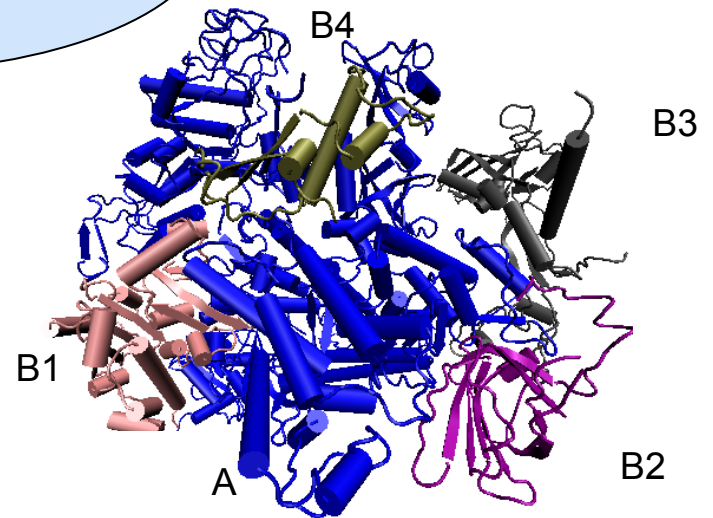
There remains a rich source of knowledge unmined by network theorists!

Structural biology perspective:



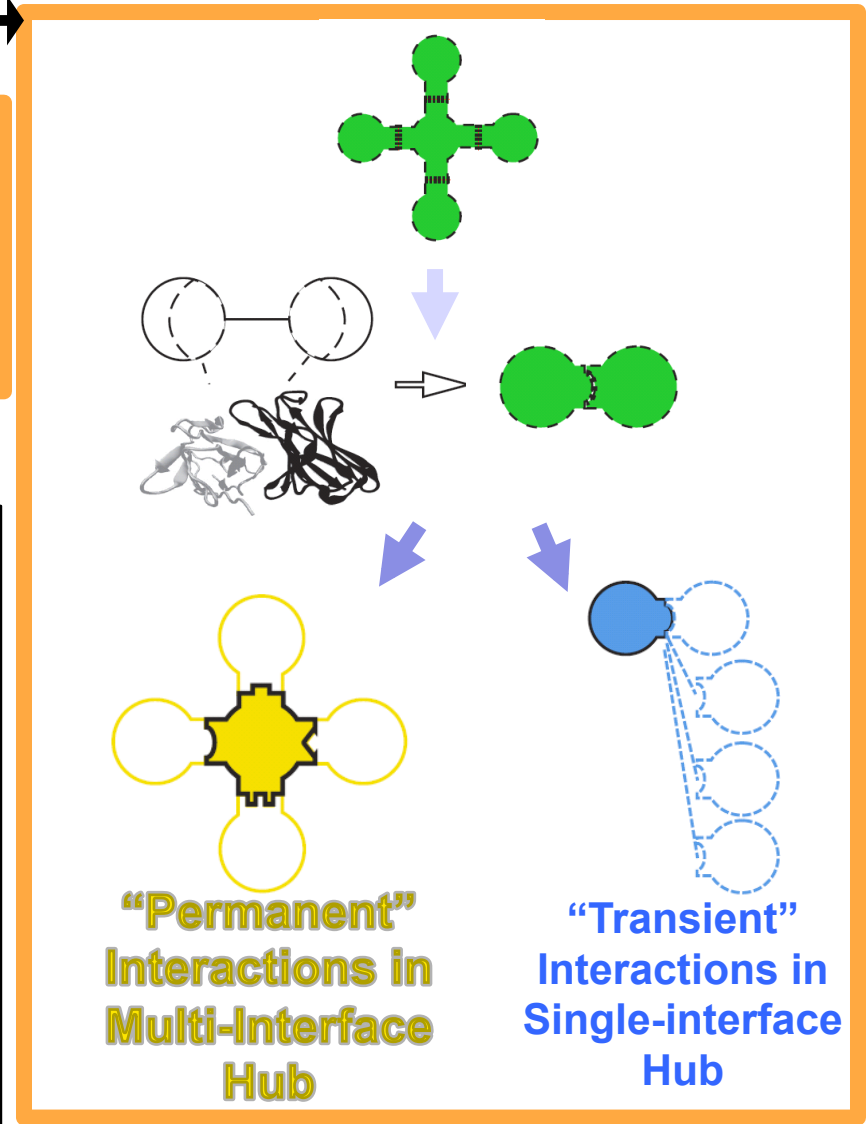
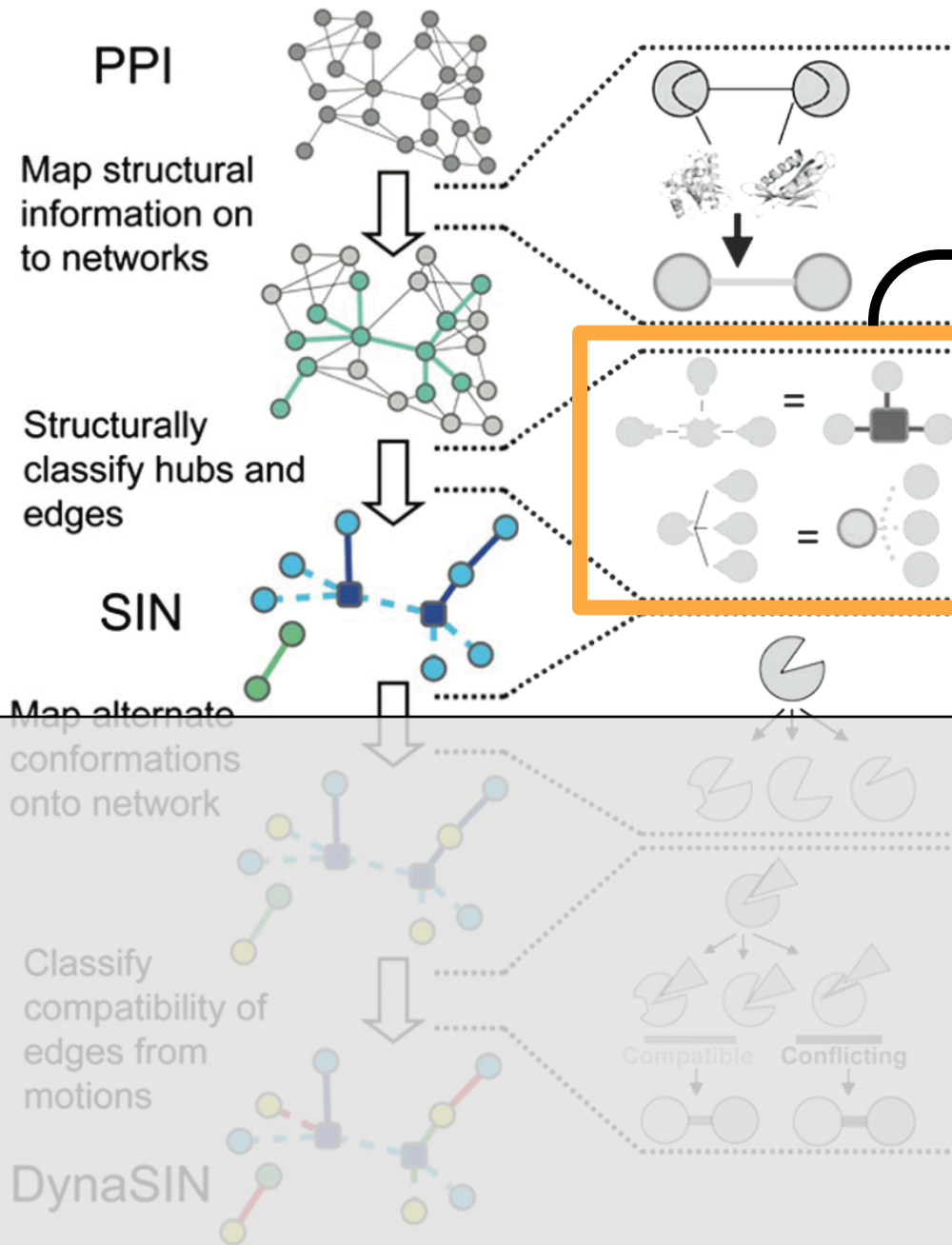
Cdk/cyclin complex

≠

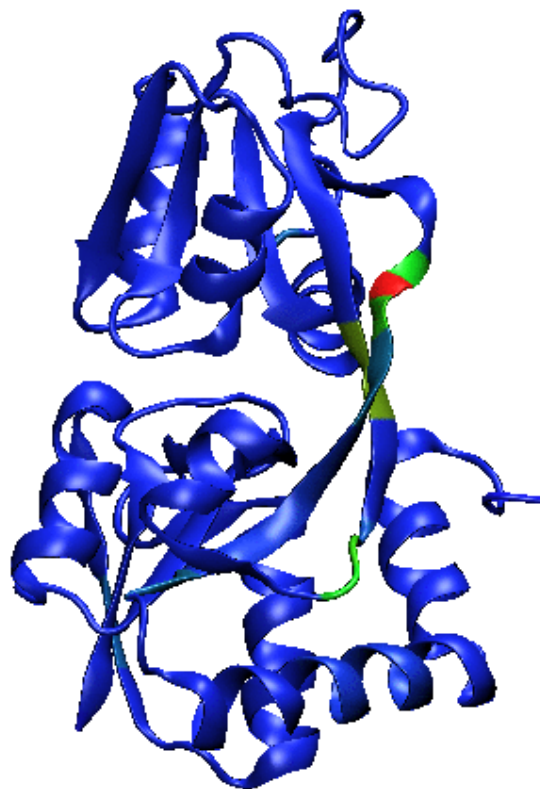


Part of the RNA-pol complex

Overview of DynaSIN Construction



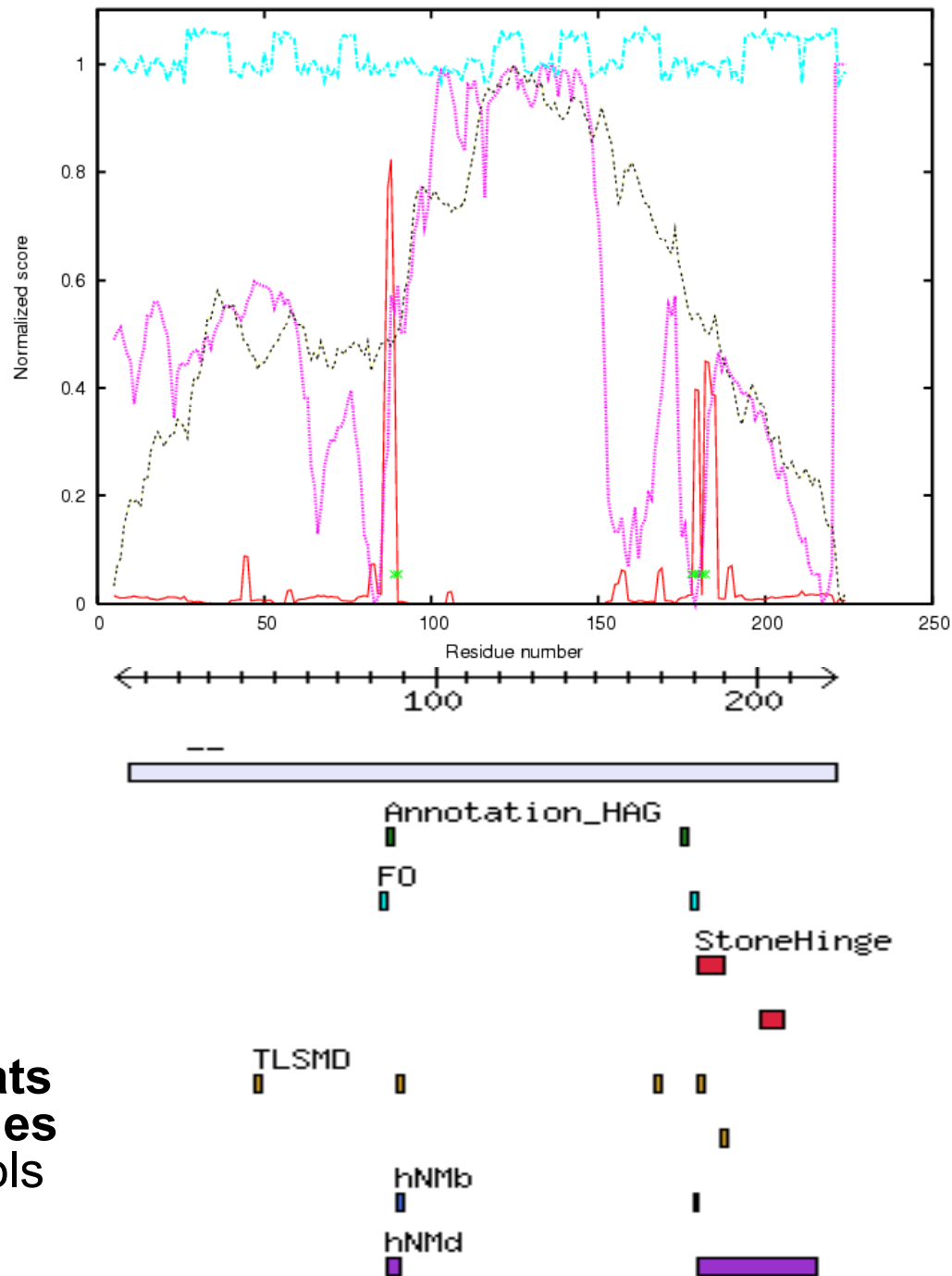
[Bhardwaj et al. ('11) Prot Sci]



EX: Glutamine Binding Protein







MolMovDB.org :

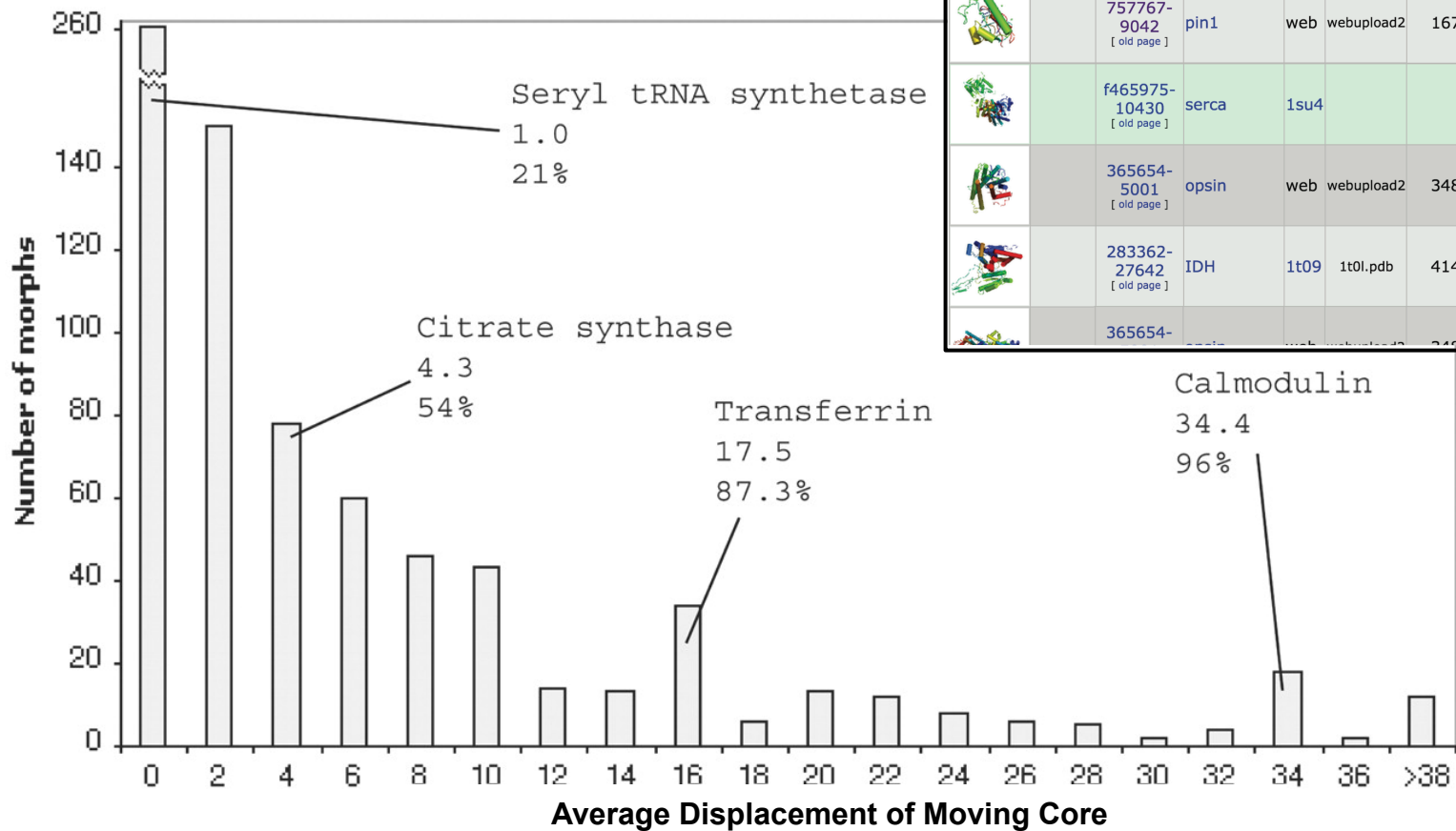
Analysis of a single motion to characterize with it **standard stats** (eg rot. Angle) & find **key residues** (eg hinges) using a variety of tools



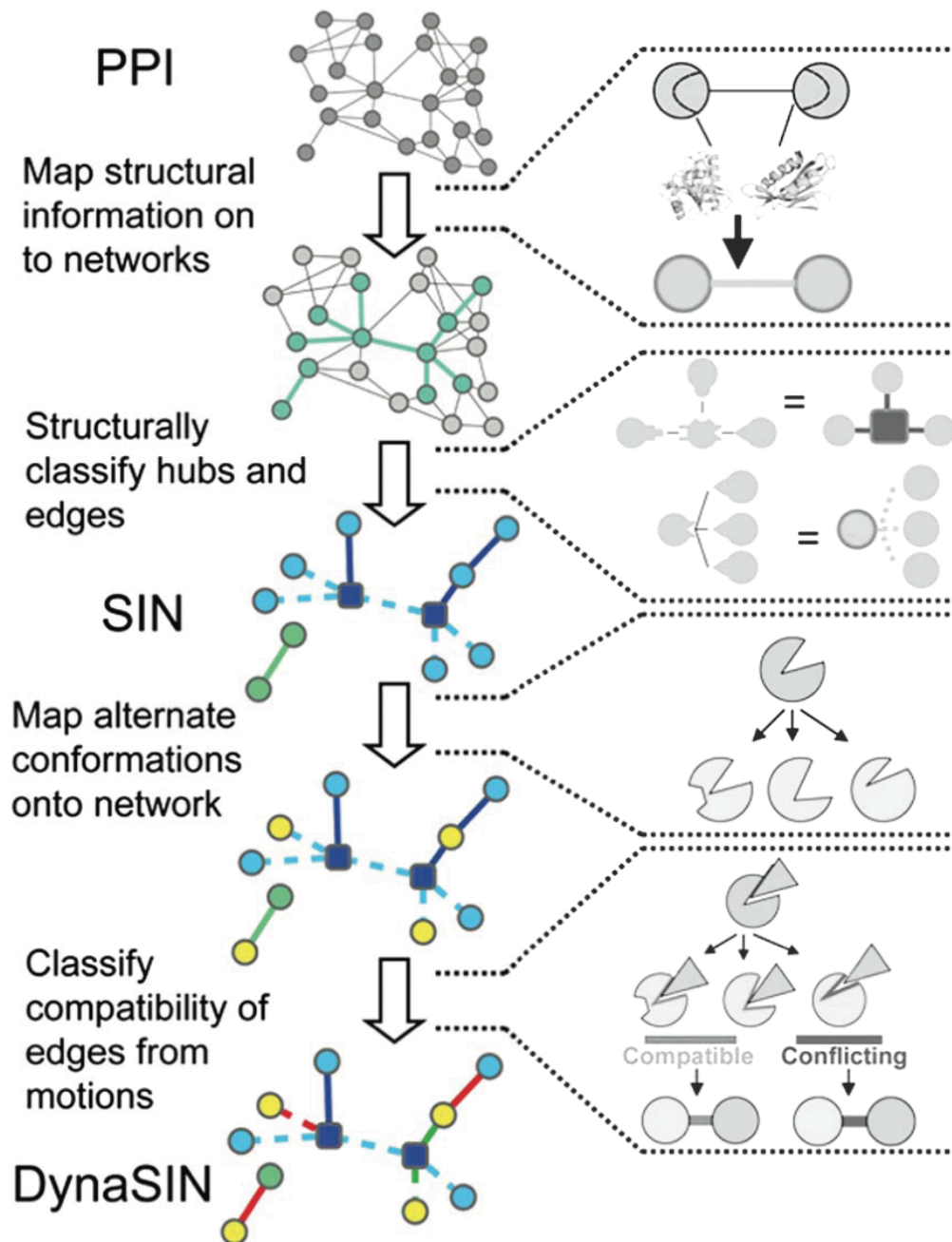
MolMovDB: determining population statistics on protein motions

Distribution across user-submitted morphs

Morph	Motion			PDB ID		# of residues	maximum CA deviation	# of frames
	motion ID	morph ID	name (in DB, as submitted)	#1	#2			
		882194-12707 [old page]	D4	web	webupload2	218	0.285705	10
		757767-9042 [old page]	pin1	web	webupload2	167	0.19603	10
		f465975-10430 [old page]	serca	1su4				
		365654-5001 [old page]	opsin	web	webupload2	348	0.593538	13
		283362-27642 [old page]	IDH	1t09	1t0l.pdb	414	4.79516	32
		365654-						

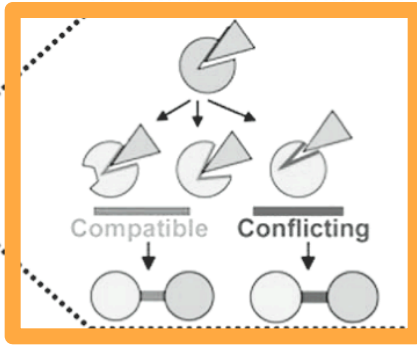
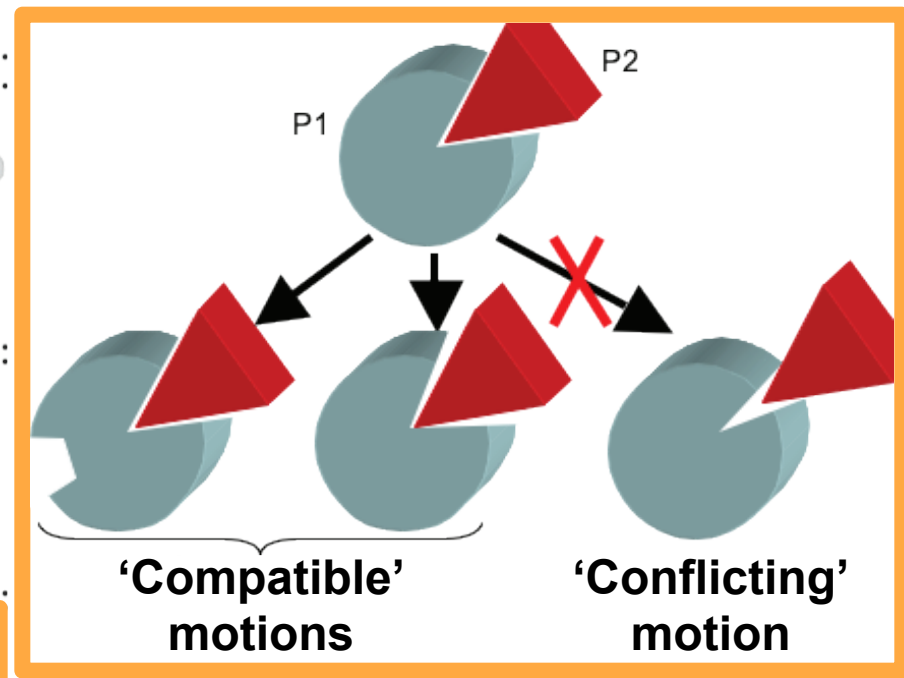
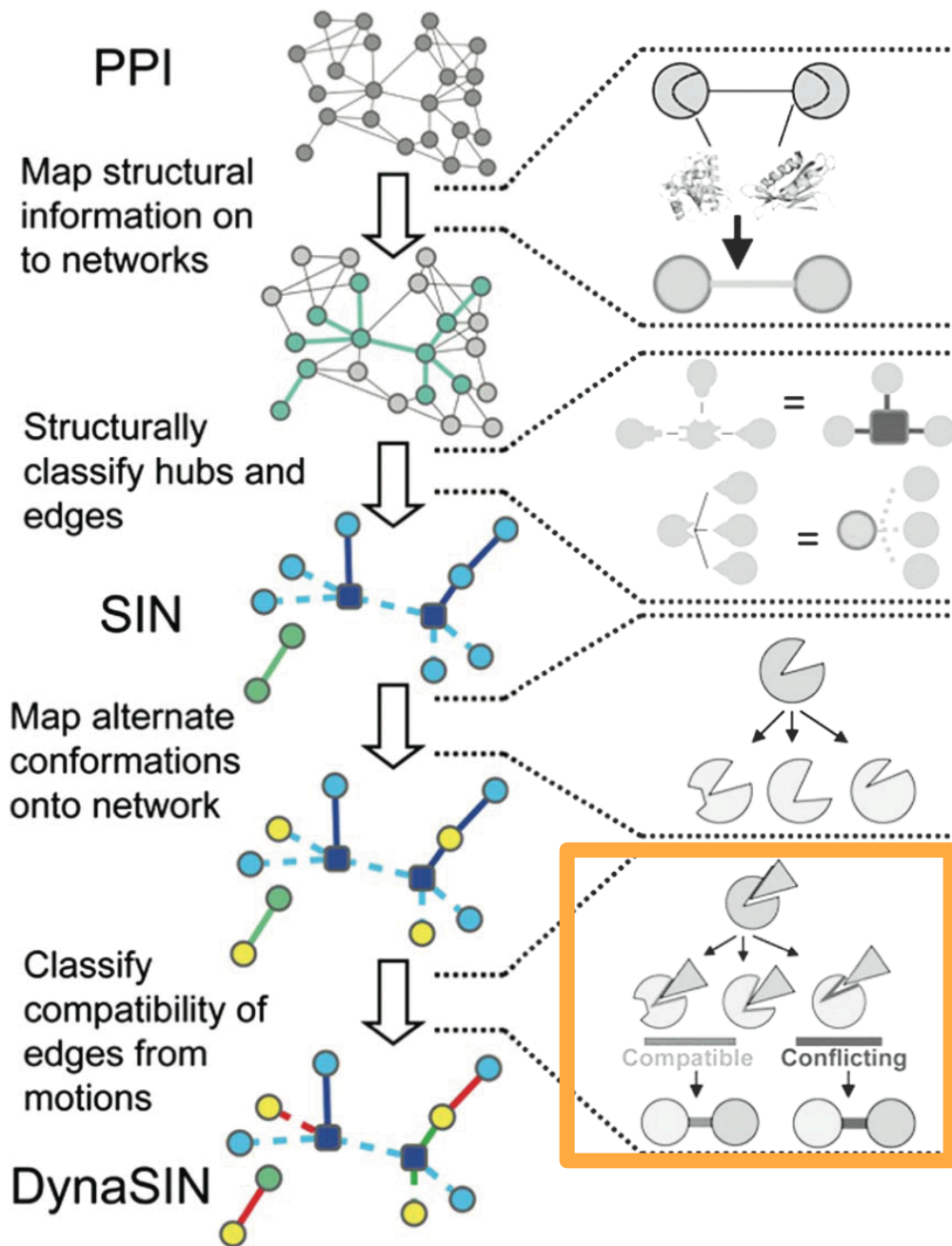


Overview of DynaSIN Construction



DynaSIN.MolMovDB.org
[Bhardwaj et al. ('11) Prot Sci]

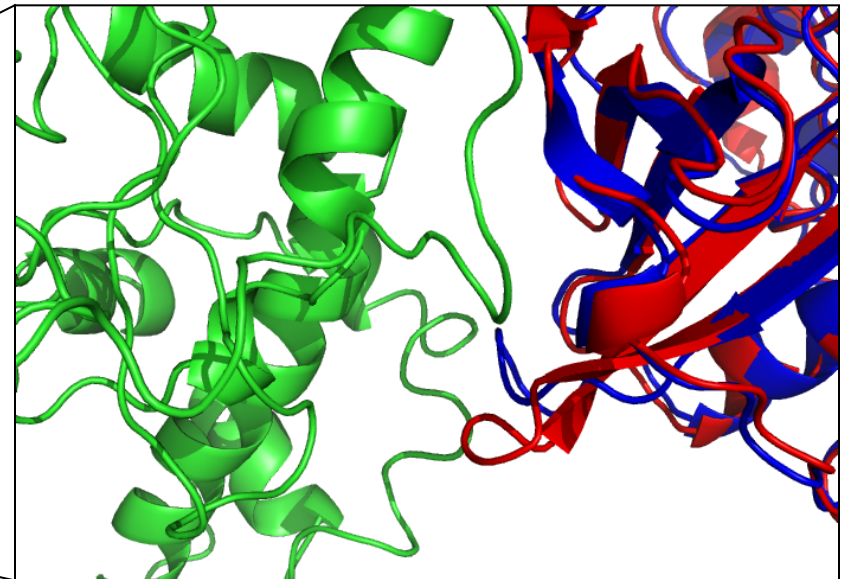
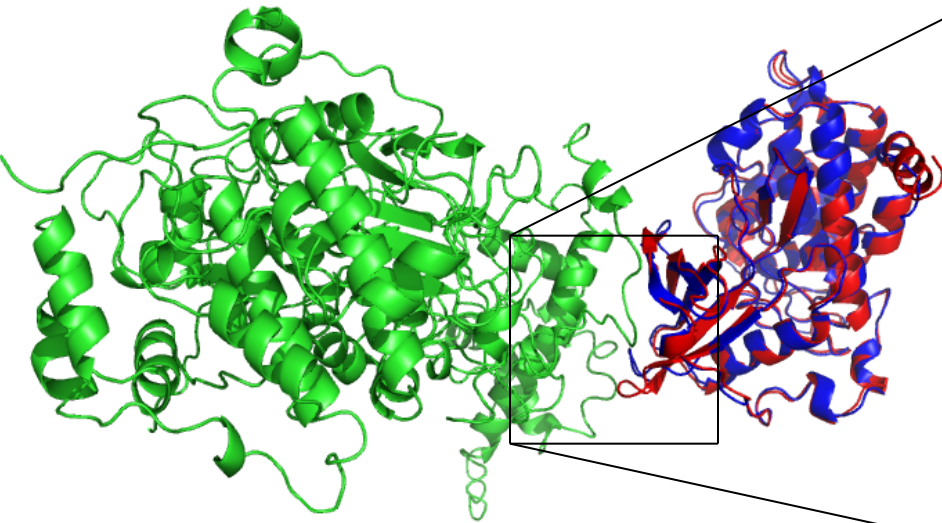
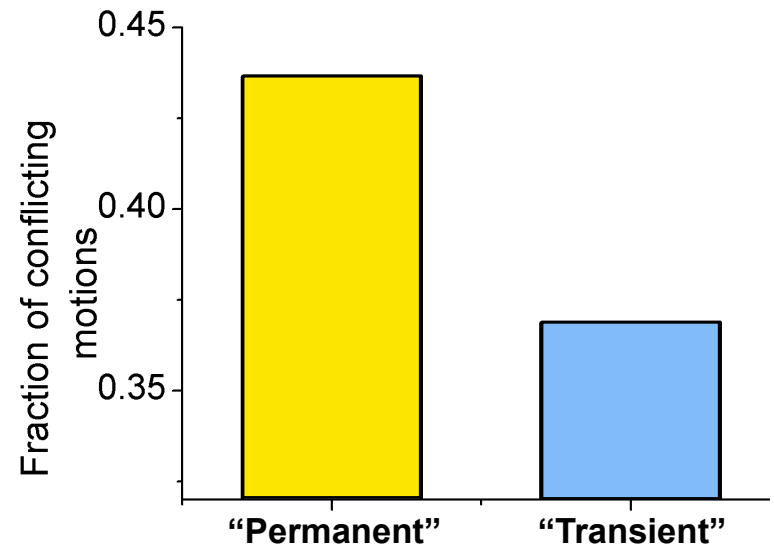
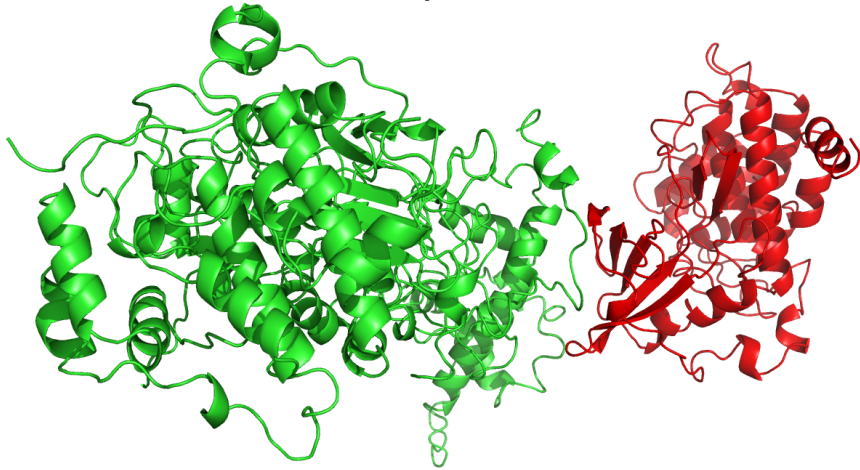
Overview of DynaSIN Construction



Conflicting and compatible interactions

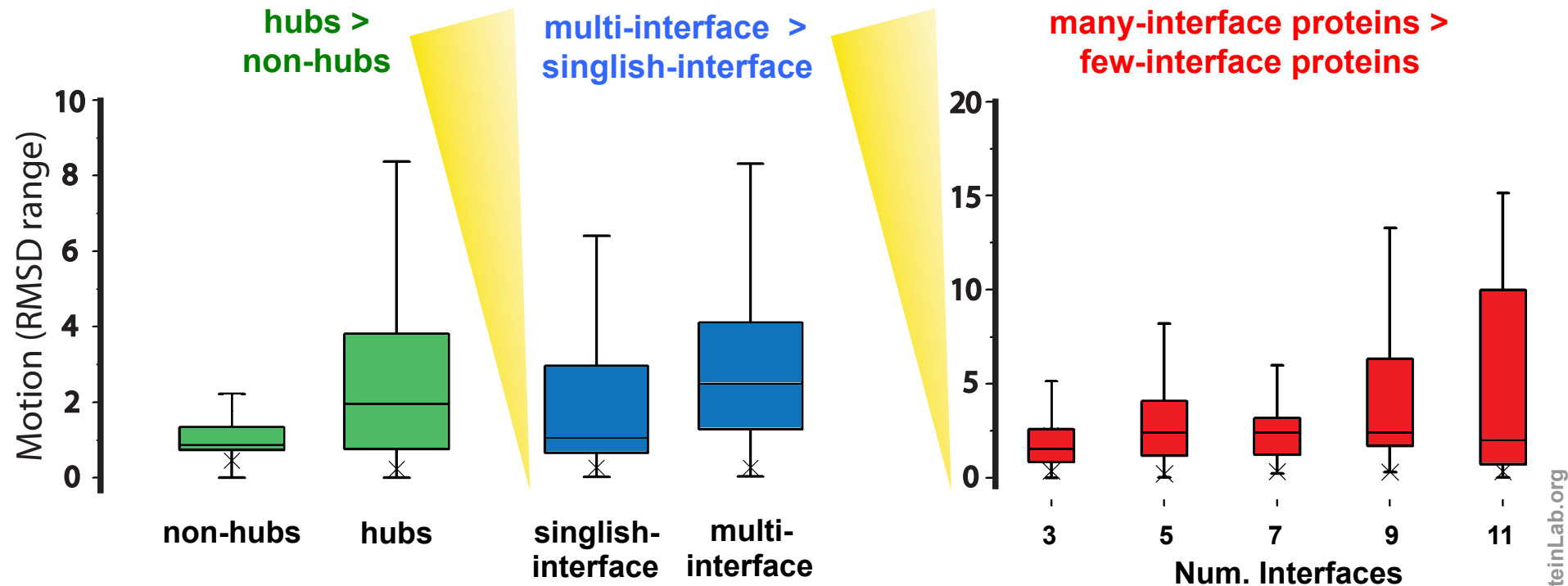
Protein Kinase CK2 β

Protein Kinase CK2 α



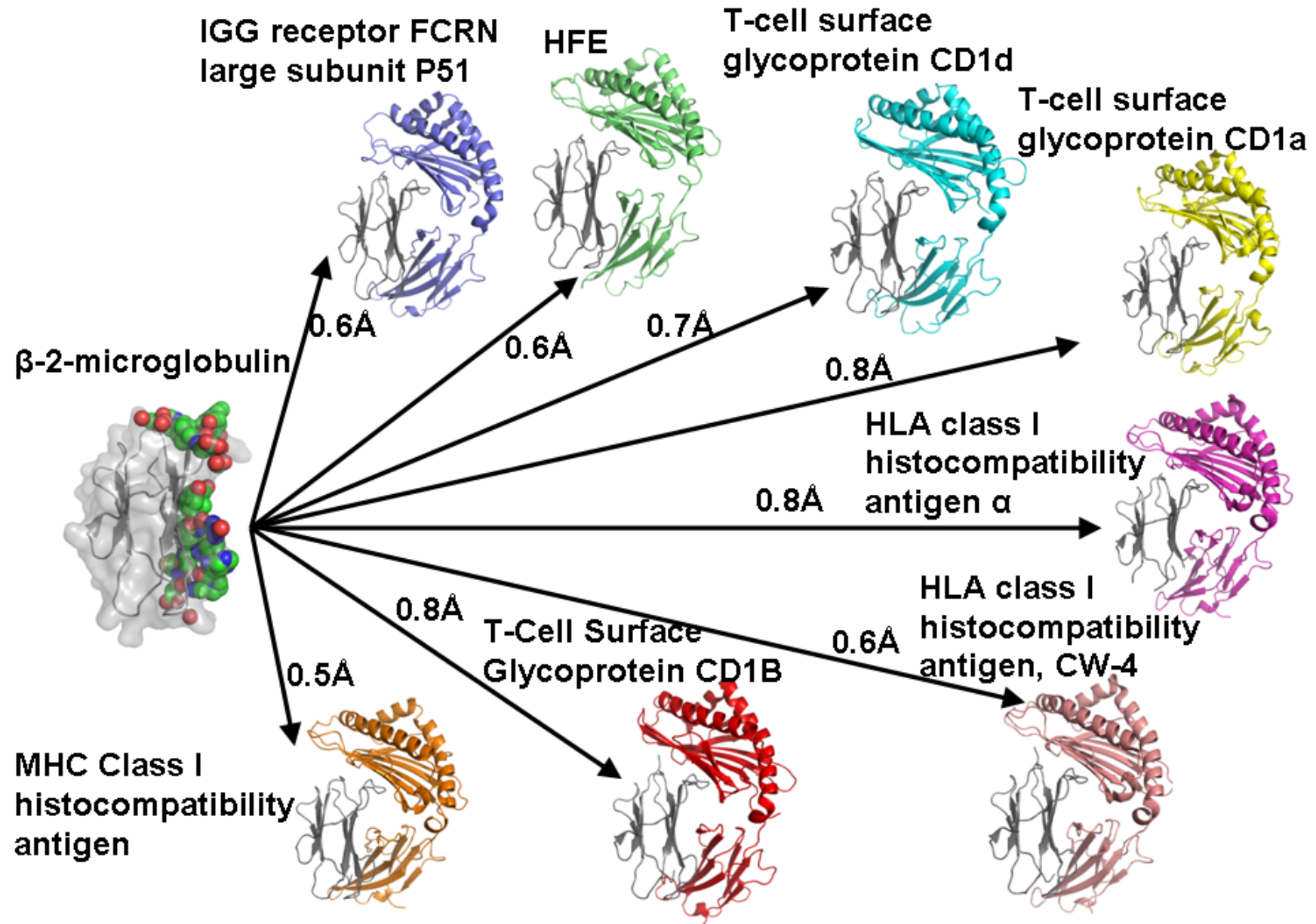
Blue: alternate conformation

The degree of conformational change correlates with hub properties



* Note: All p-values < 4E-3

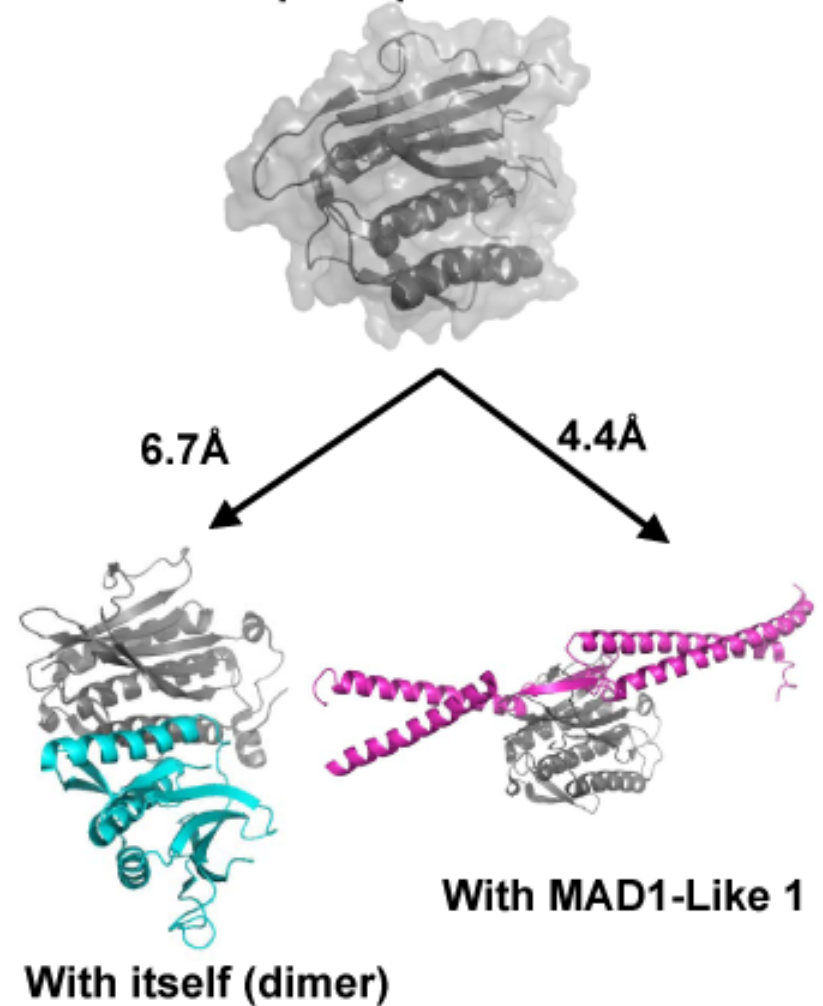
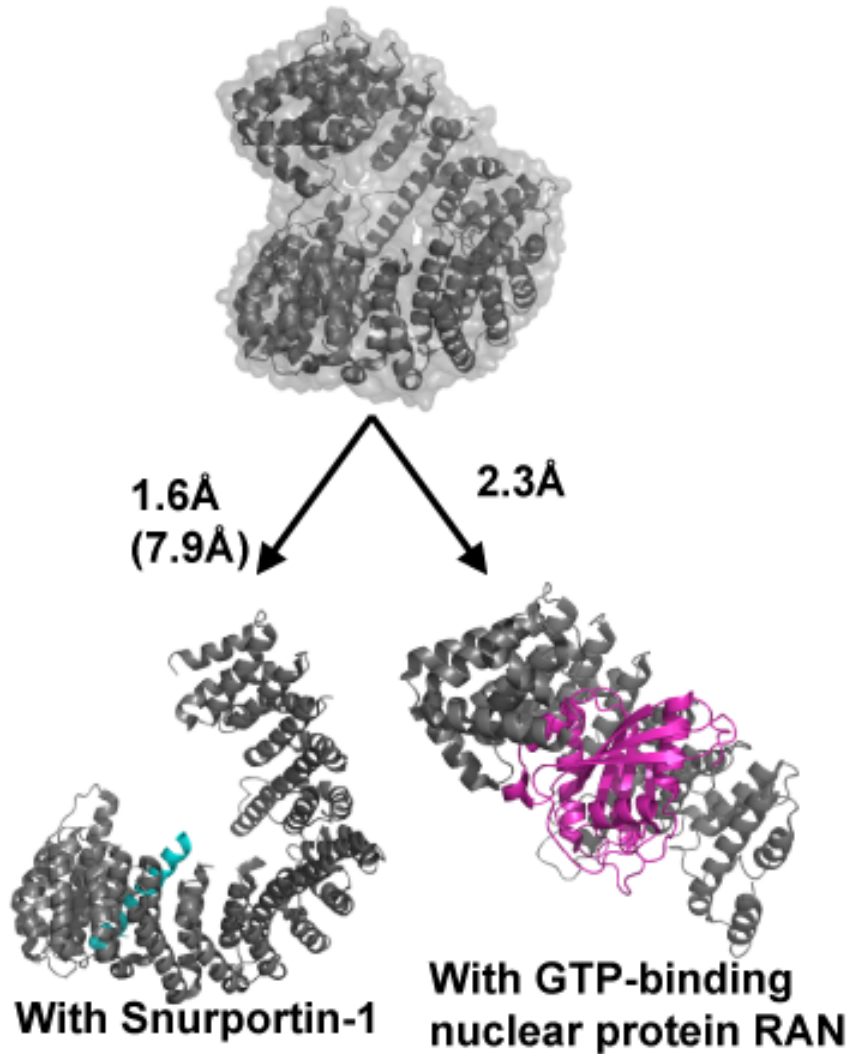
Examples: Single Interface Hubs



Examples: Multi-interface hubs

Importin subunit beta-1

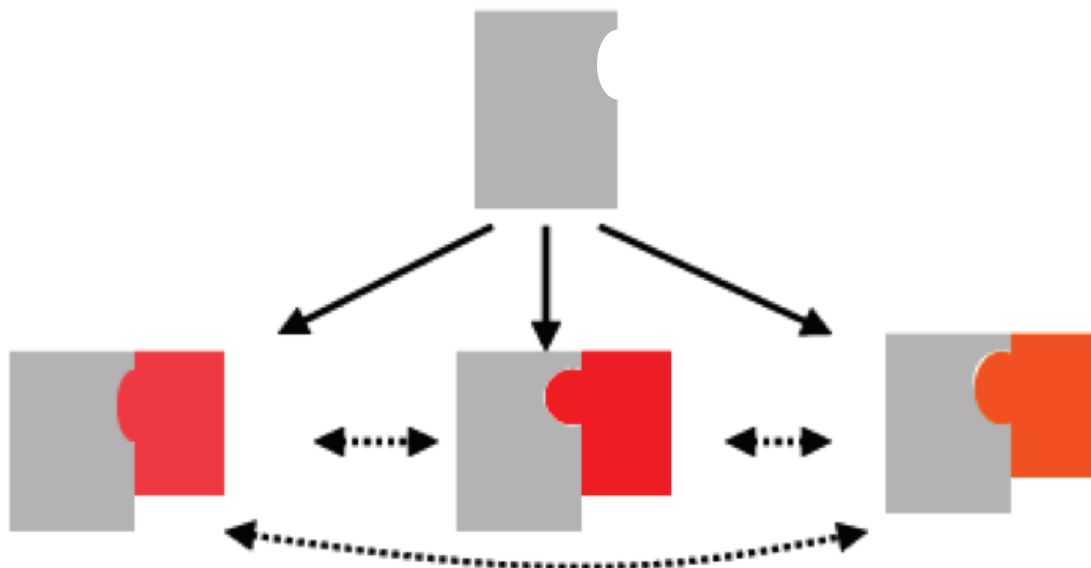
Mitotic spindle assembly checkpoint protein mad2a



Rationalization: “Permanent” vs “Transient” Interactions

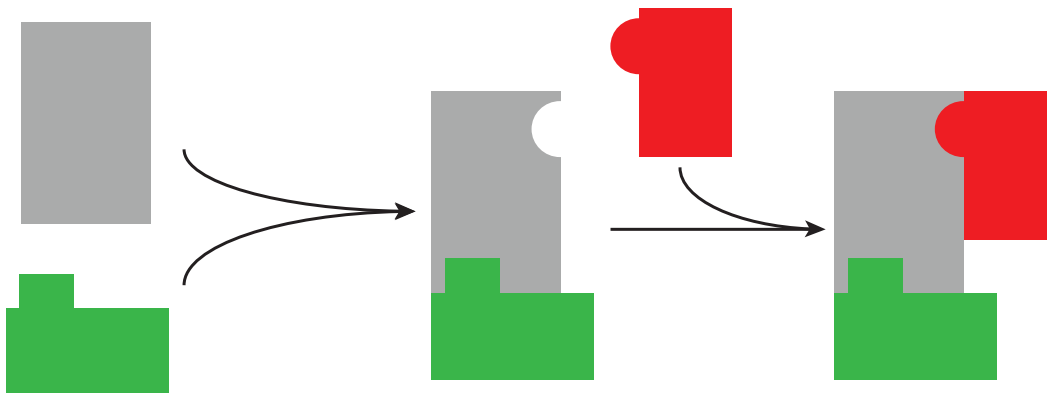
Single-interface hub interactions

→ Less interface modification (lower energy barrier) for frequently-changing interactions

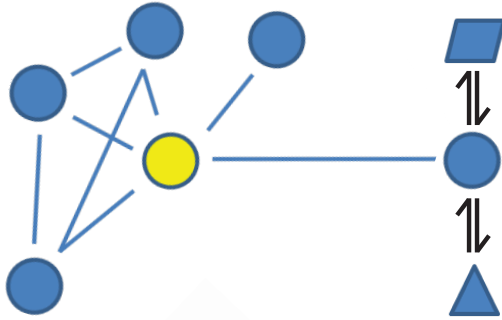


Multi-interface hub interactions

→ Only need to pay high energetic costs for larger changes infrequently



Using 3D-structure
into interpret
networks & deep-
sequencing data



- Structural Interaction Network & Protein Motions (DynaSIN)

- Multi-interface permanent hubs have more motion than single-interface transient ones
- Also have more conflicting motions

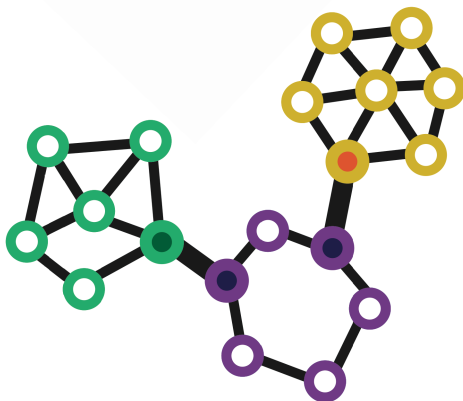
- LOF variants & Categories of Essential & Disease-sensitive Genes

- Variation at Protein Interfaces in the context of Network Connectivity & its use for Disease-gene Predictions

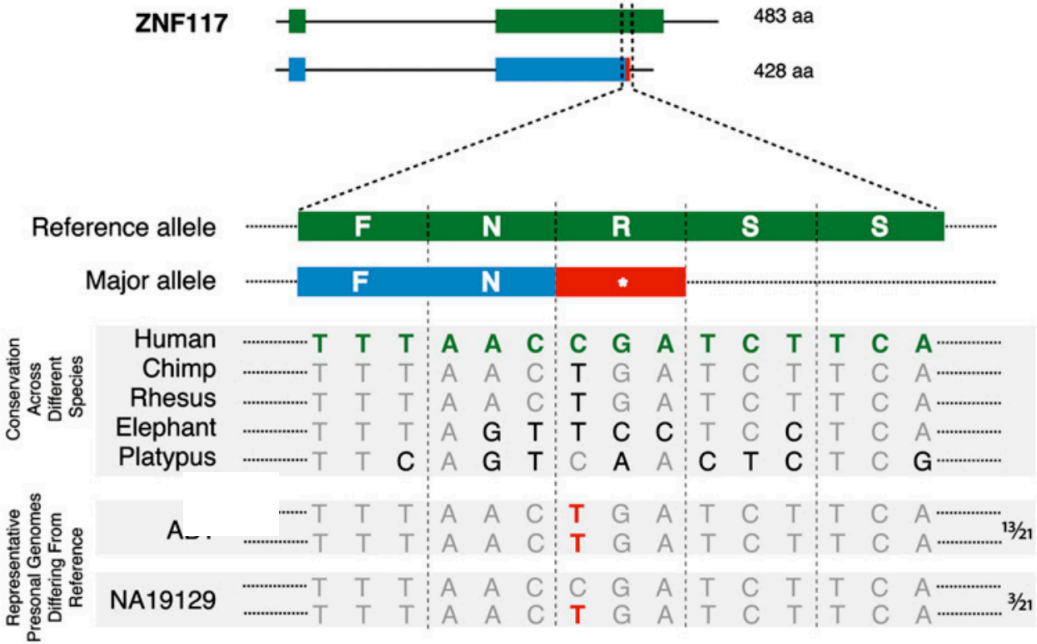
- Highly connected parts of PPI under stronger selection but signal weak
- Stronger signal in SIN & even stronger in multiNet (integration of many networks)
- Signal strong enough to build predictor

- Rationalizing Deleterious Variants in terms of Potential Allosteric Sites

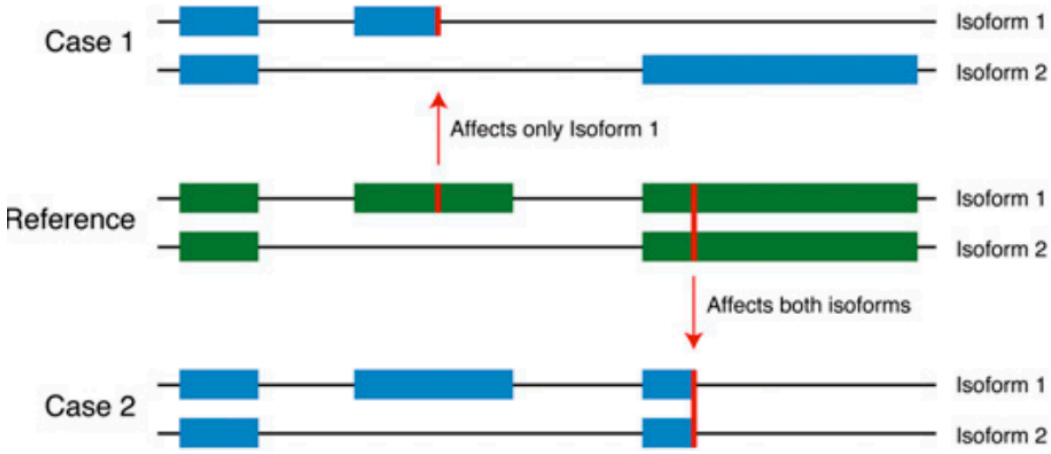
- Identifying potential allosteric residues on surface & inside
- These are under stronger selection & may explain some HGMD SNPs



LOF variants



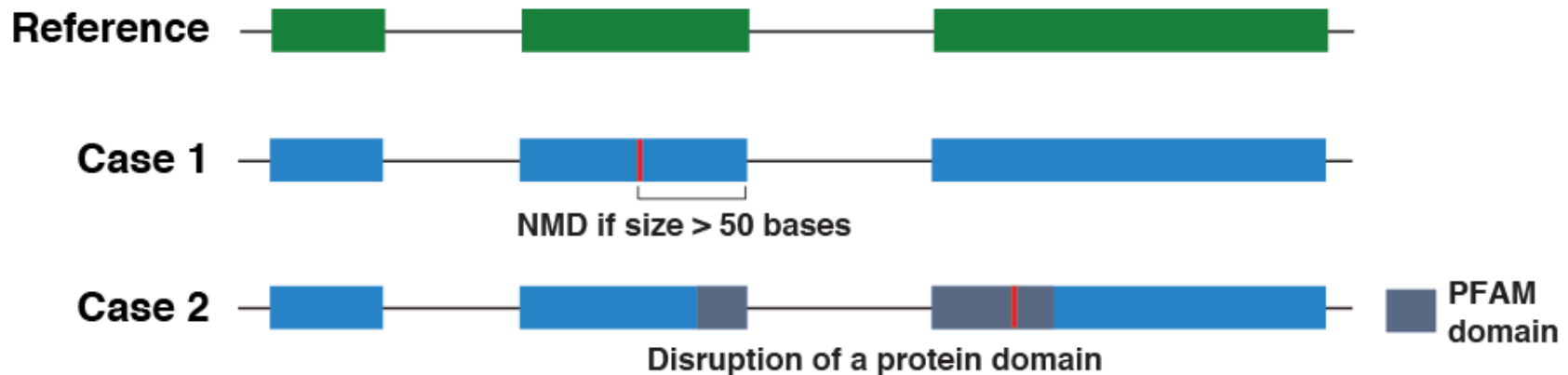
Impact of a SNP on alternate splice forms



[Balasubramanian et al., *Genes Dev.*, '11]

Making Sense Of Nonsense

- Tricky to find true LOFs
- Use VAT tool
- Some further complexities beyond alt. splicing



- Future: use more structural & protein knowledge

Number Of LoF Variants In An Individual

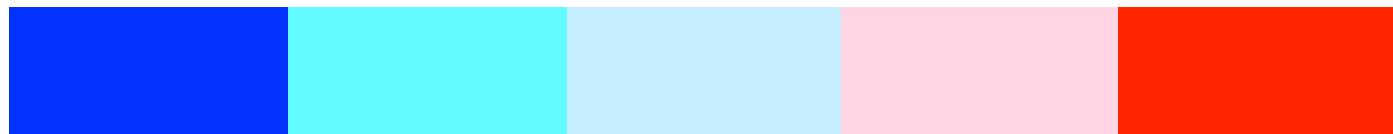
- Initial estimate of 250-300 LoF variants.
- Enriched for sequencing, annotation artifacts.
- Each of us probably carry ~ 100 genuine LoF variants.
- Majority of the LoF variants are rare.
- Each individual has ~ 20 homozygous LoF variants

variant type	Filtered LoF events			
	1000G low-coverage average per individual			NA12878 high coverage European
	CEU	CHB+JPT	YRI	
stop	26.2	27.4	37.2	23
splice	11.2	13.2	13.7	12
frameshift indel	38.2	36.2	44.0	38
large deletion	28.3	26.7	26.6	24
total	103.9	103.5	121.5	97

Phase 1 Update:
~150 LOF/individual
but only 10-20 of these rare

Gene Categories with known phenotypic effects

Decreasing tolerance to mutation



LoF-tol

Neutral

GWAS

HGMD

Essential

(common
disease-assoc.
variants)

(rare
disease-causing
variants)

- Homozygous inactivation in at least one healthy 1000 Genomes individual
- Weak selection constraints

- Homozygous inactivation leads to clinical features of death before puberty or infertility
- Very strong selection constraints

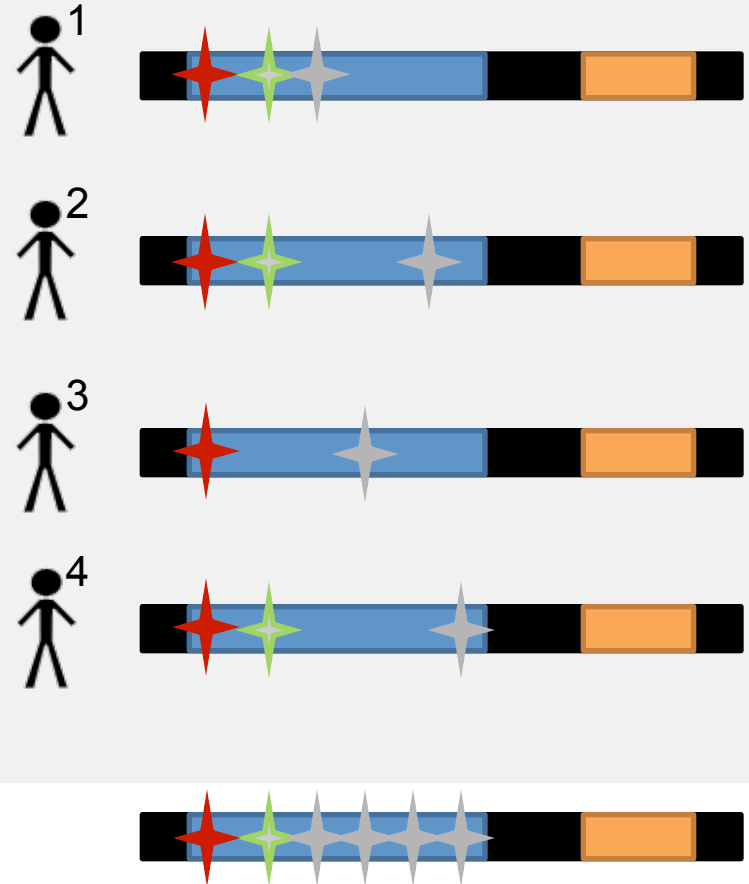
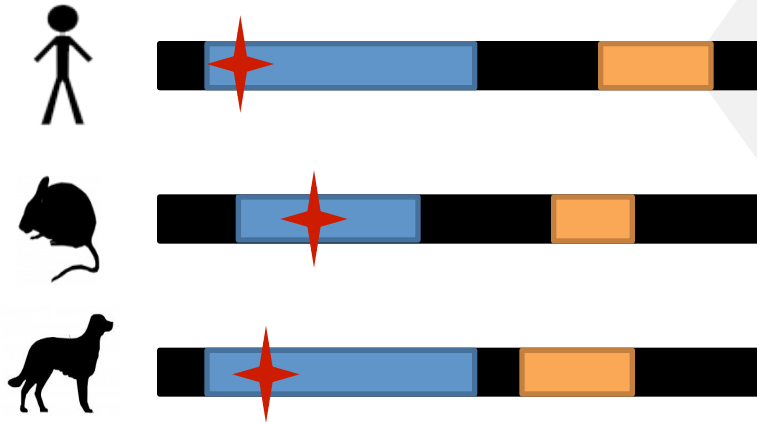
From MacArthur et al, Science, 2012

From Liao et al, PNAS, 2008

Quantifying Selection inter- and intra-species approaches

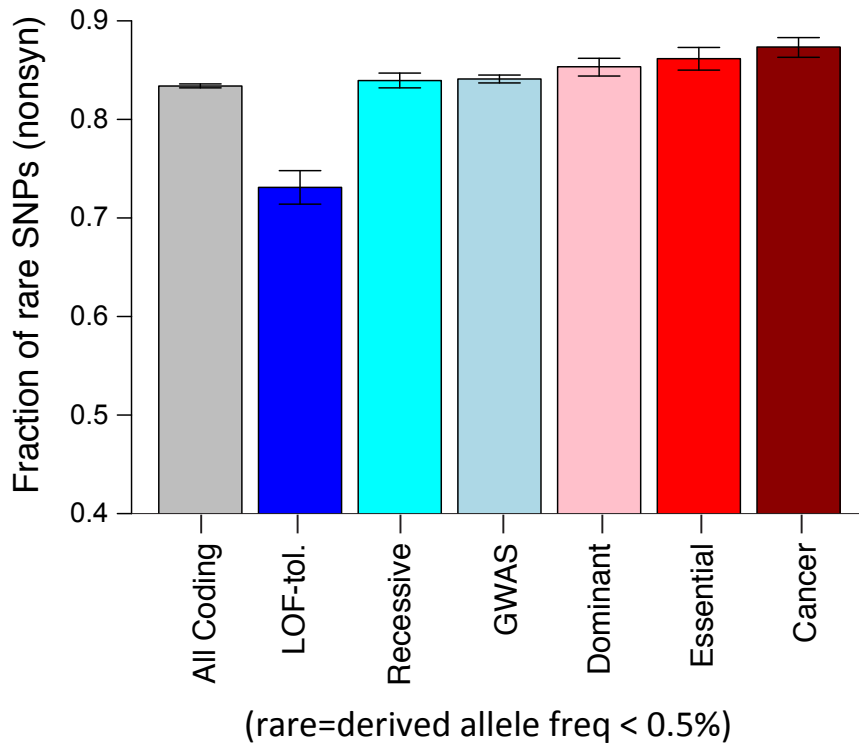
'Conservation'

- Typically defined by comparison across species
- dN/dS in coding regions
- GERP noncoding



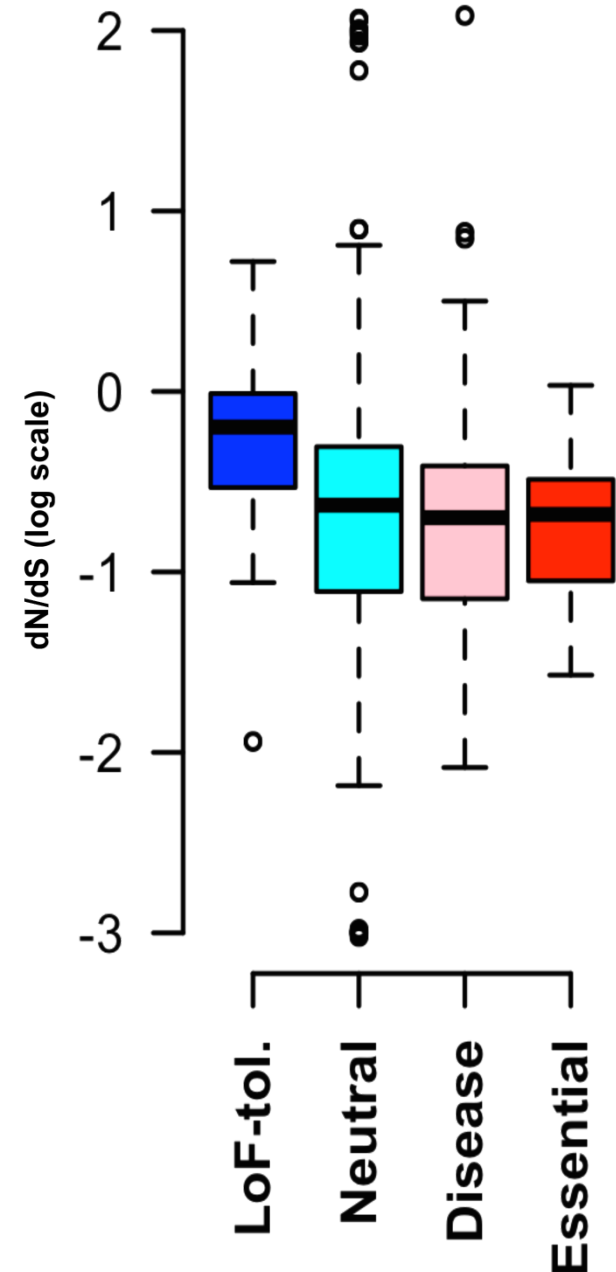
- Metrics for selection within population
 - SNP density (confounded by mutation rate)
- Depletion of common polymorphisms for regions under selection (also an enrichment of rare variants)

Selection vs Gene Categories: Signal there but weak

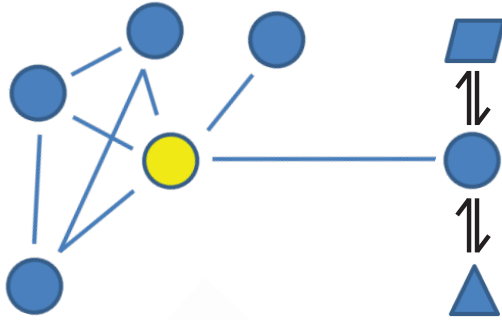


LOF-tol (Loss-of-function tolerant): least negative selection

Cancer: most selection



Using 3D-structure
into interpret
networks & deep-
sequencing data



- Structural Interaction Network & Protein Motions (DynaSIN)

- Multi-interface permanent hubs have more motion than single-interface transient ones
- Also have more conflicting motions

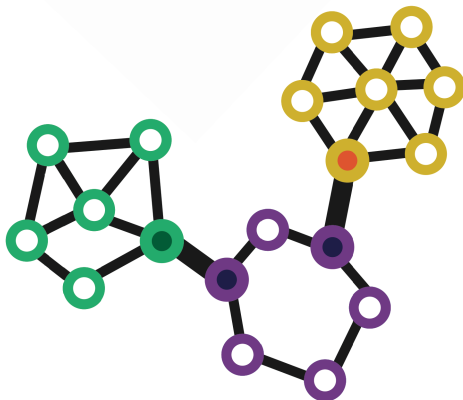
- LOF variants & Categories of Essential & Disease-sensitive Genes

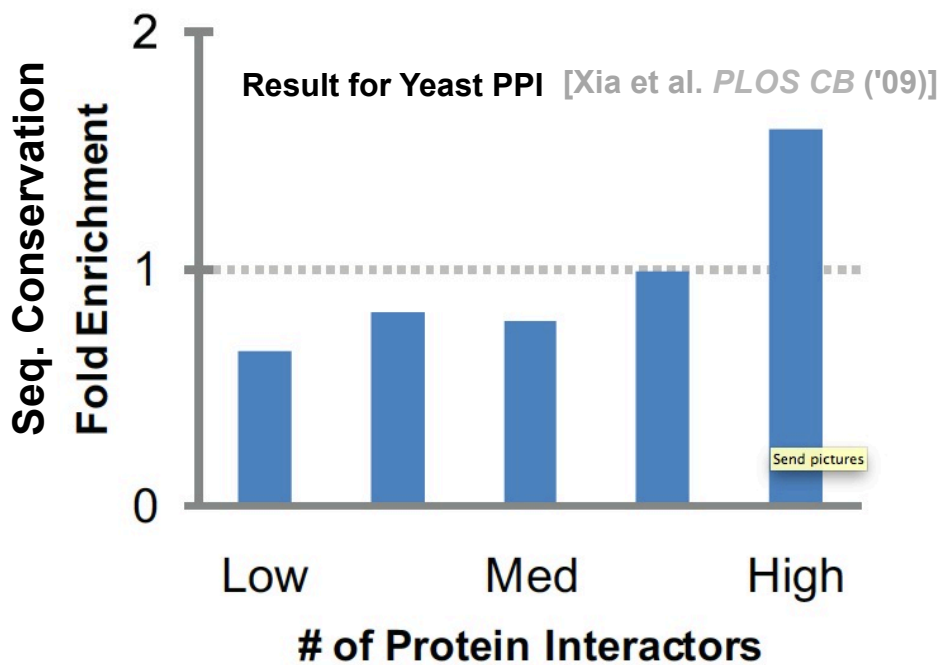
- Variation at Protein Interfaces in the context of Network Connectivity & its use for Disease-gene Predictions

- Highly connected parts of PPI under stronger selection but signal weak
- Stronger signal in SIN & even stronger in multiNet (integration of many networks)
- Signal strong enough to build predictor

- Rationalizing Deleterious Variants in terms of Potential Allosteric Sites

- Identifying potential allosteric residues on surface & inside
- These are under stronger selection & may explain some HGMD SNPs

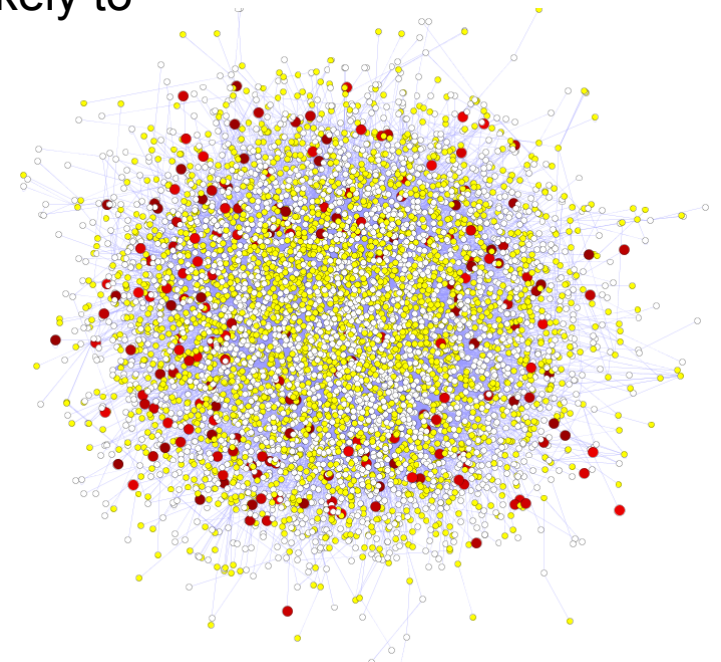




More Connectivity, More Constraint : A theme borne out in many studies

- High likelihood of positive selection
- Lower likelihood of positive selection
- Not under positive selection
- No data about positive selection

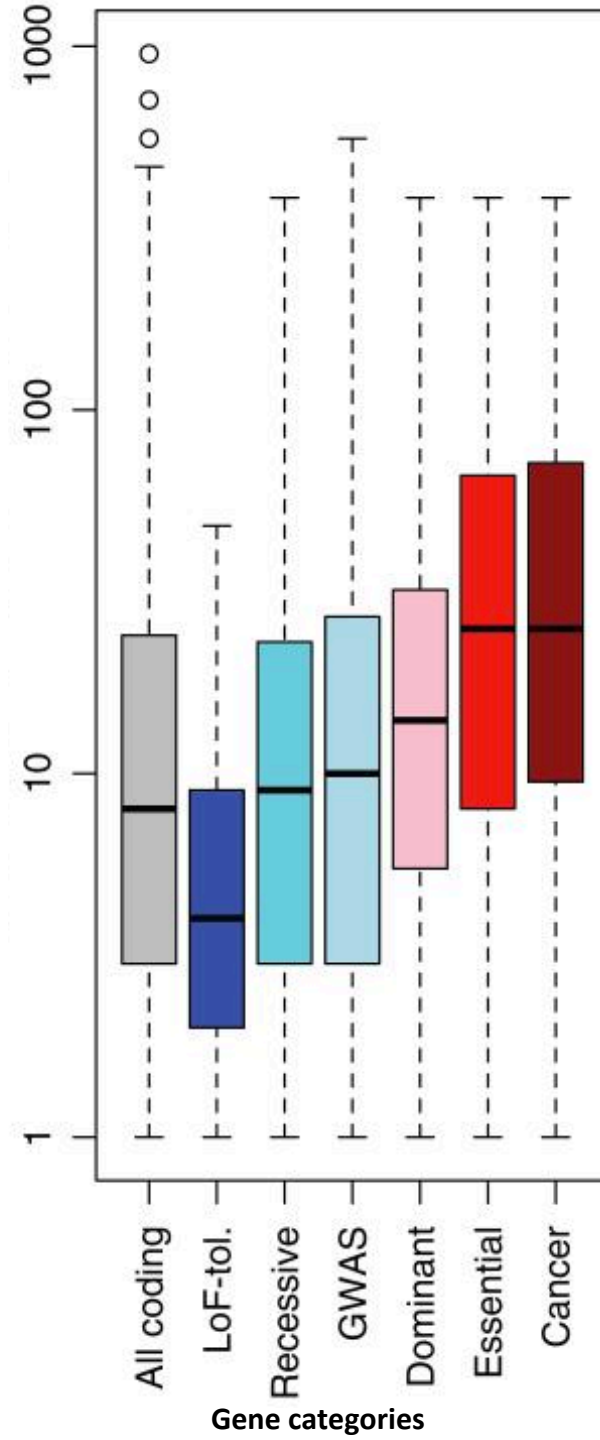
[Nielsen et al. *PLoS Biol.* (2005), HPRD, Kim et al. *PNAS* (2007)]



- Genes & proteins that have a more central position in the network tend to evolve more slowly and are more likely to be essential.
- This phenomenon is observed in **many organisms & different kinds of networks**
 - **yeast PPI** - Fraser et al ('02) *Science*, ('03) *BMC Evo. Bio.*
 - **Ecoli PPI** - Butland et al ('04) *Nature*
 - **Worm/fly PPI** - Hahn et al ('05) *MBE*
 - **Human RegNet** - Gerstein et al. ('12) *Nature*
 - **miRNA net** - Cheng et al ('09) *BMC Genomics*

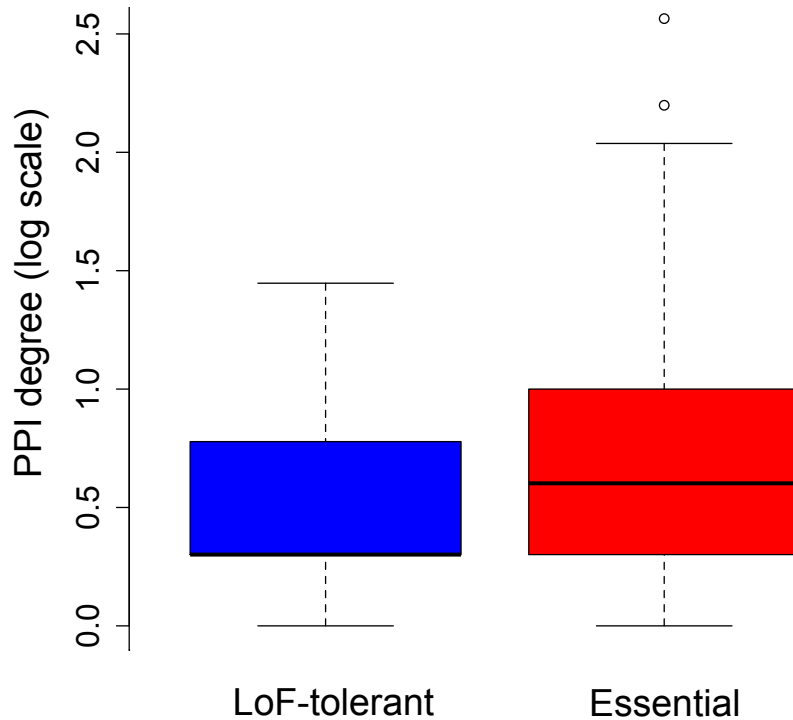
Categories vs Centrality in Human Genome, using 1000G Phase 1 Data: Signal there but weak

Degree centrality in protein interaction network

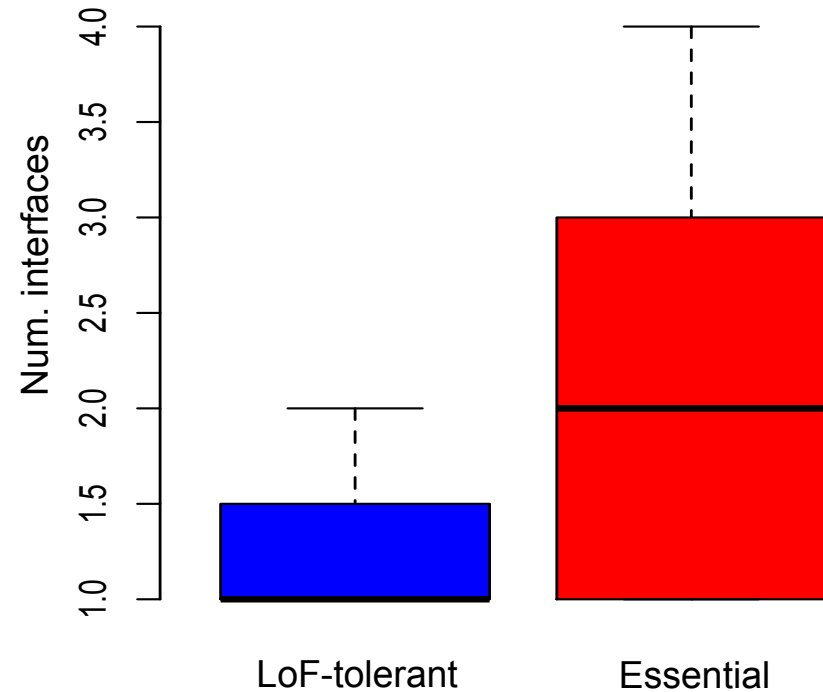


Stronger signal in the SIN than the Human PPI

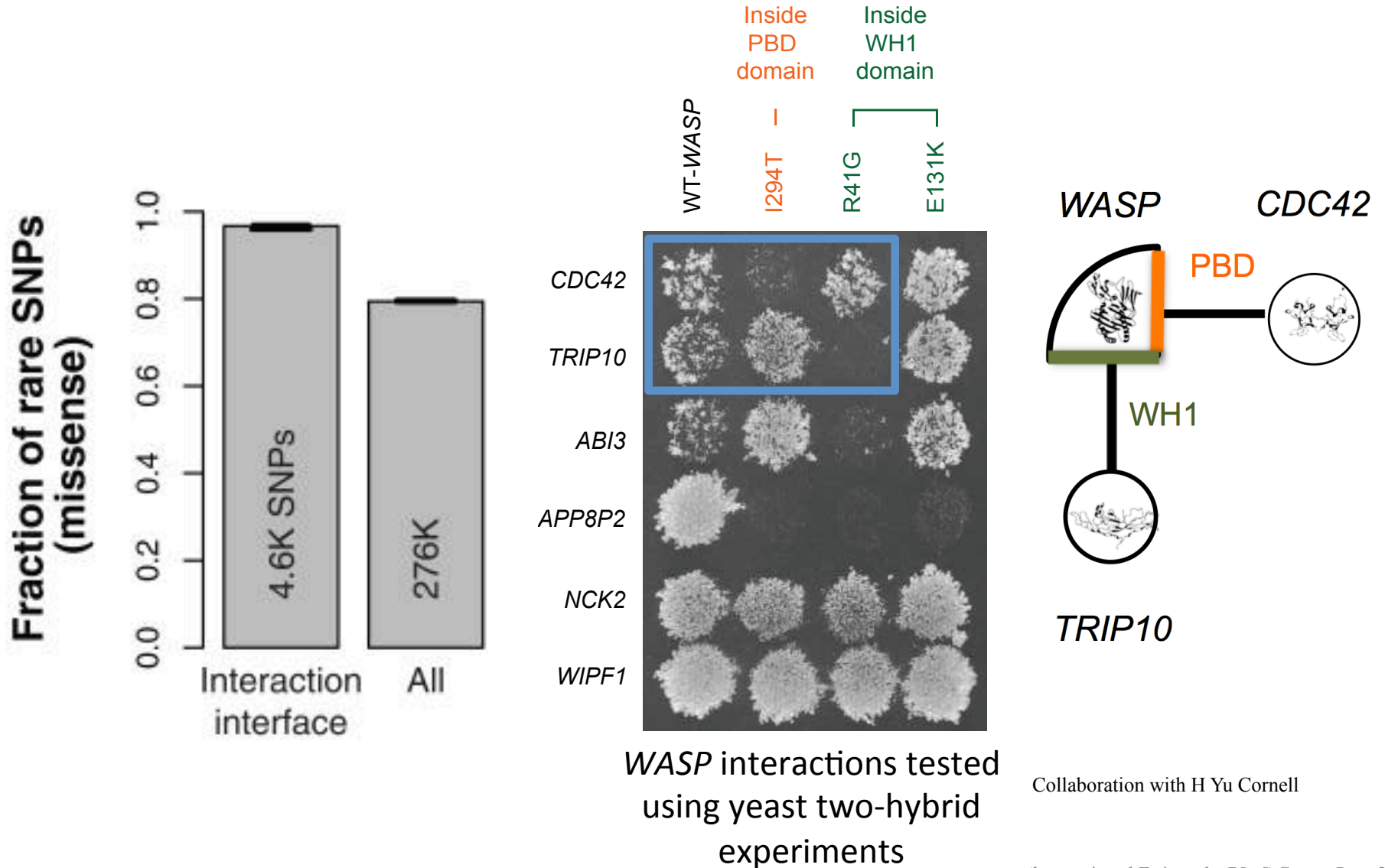
Higher
Centrality



More
interaction
interfaces



Recasting selection in SIN in terms of rare SNPs & direct experimental testing of such putative deleterious variants

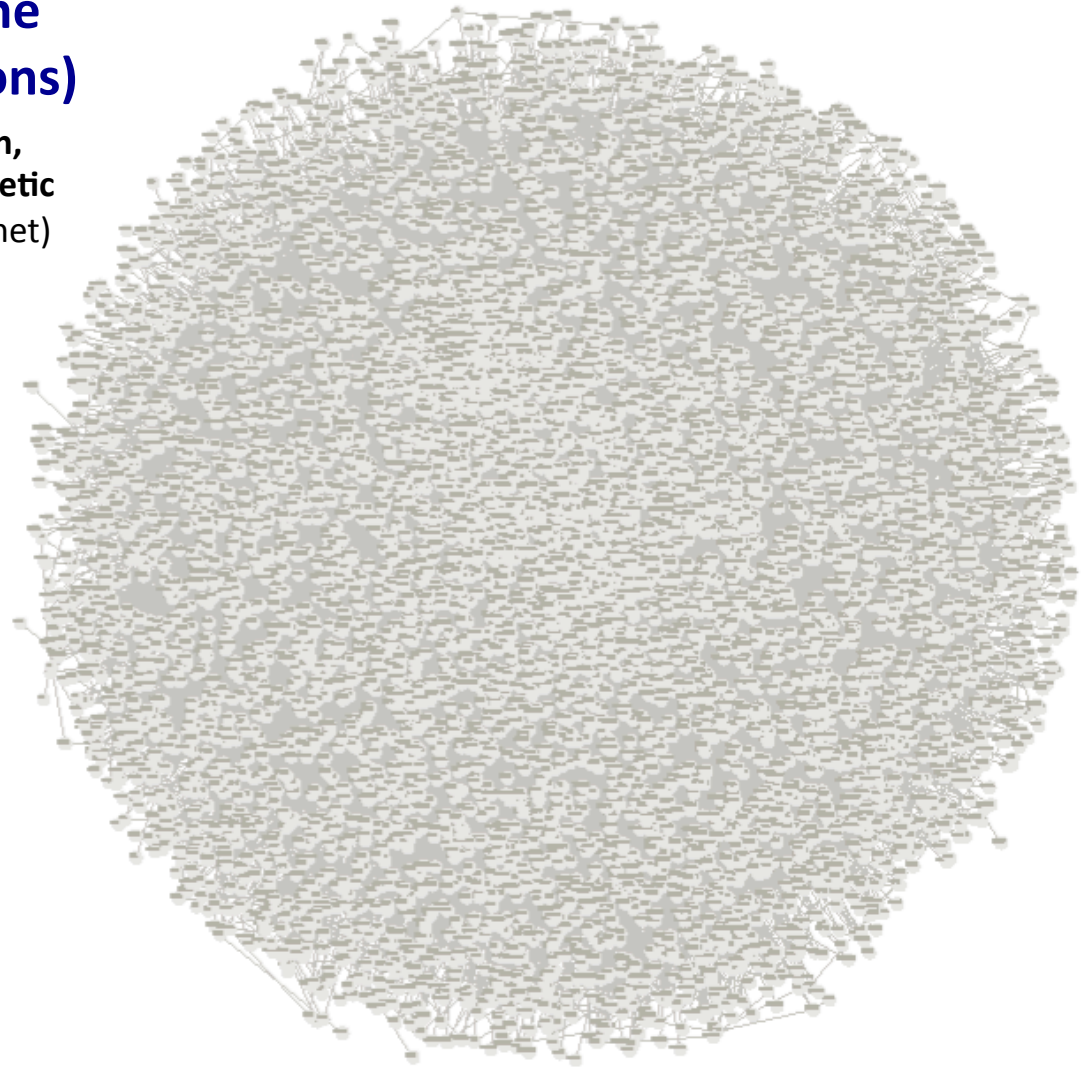


Collaboration with H Yu Cornell

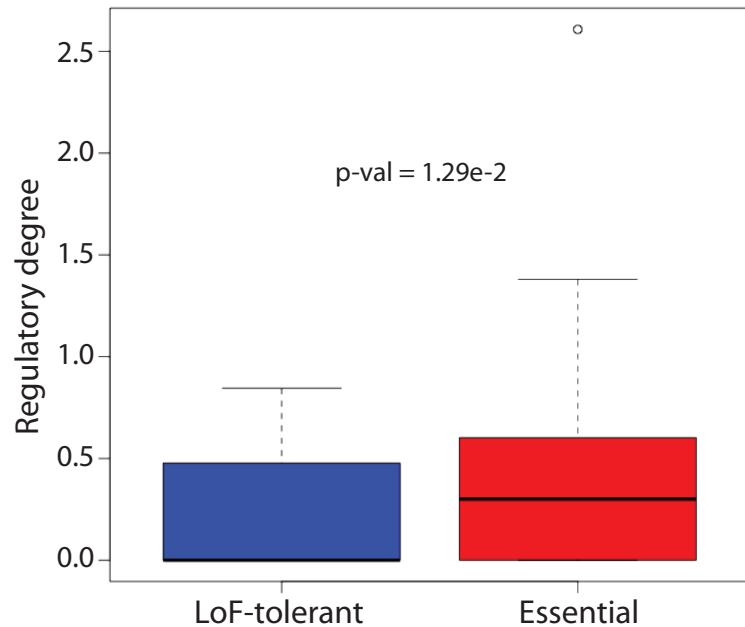
Genes participate in many networks (no single network captures the global picture of gene interactions)

Combine **regulatory, physical protein-protein, signaling, metabolic, phosphorylation and genetic** interactions to create a unified network (Multinet)

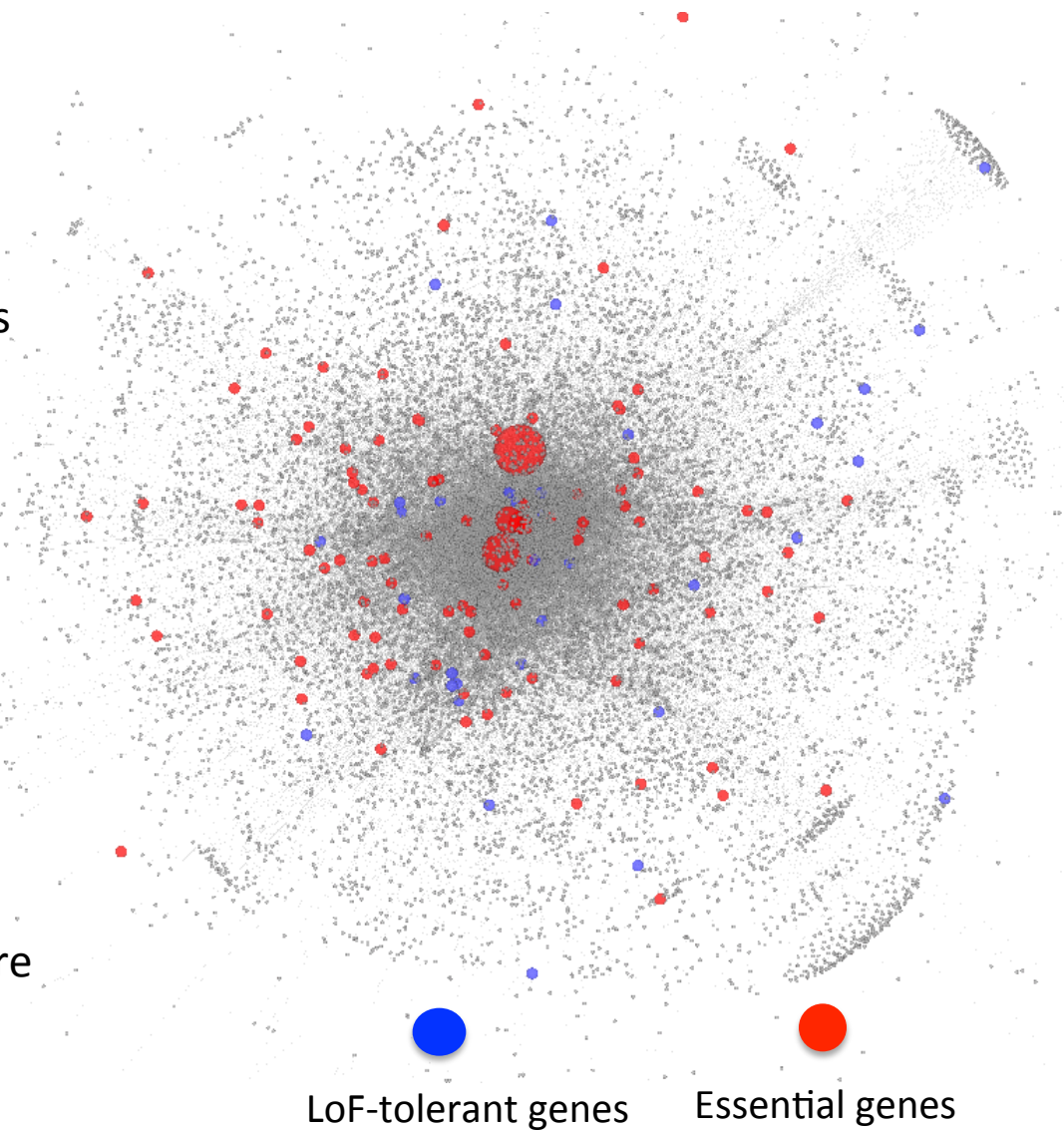
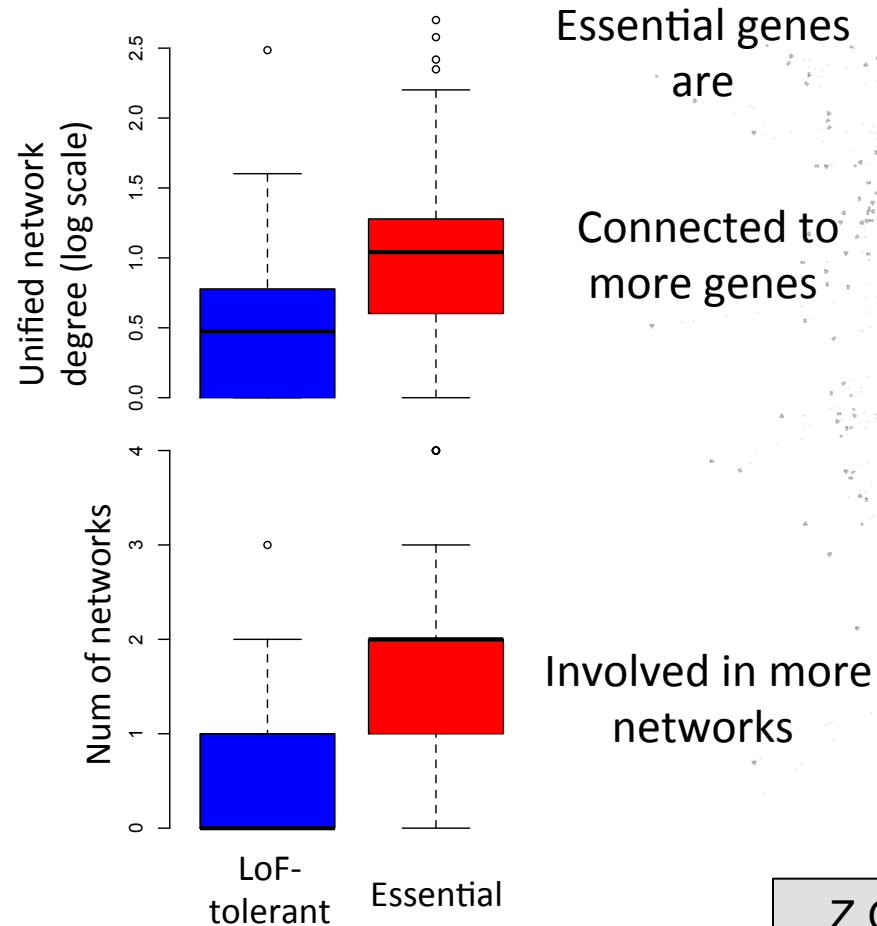
Multinet – the ultimate hairball!



Nodes: ~15,000 genes
Edges: ~110,000 interactions



Even Stronger Signal of Connectivity v Constraint in MultiNet



Z Gumus
iCAVE movie

Size of nodes scaled by
total degree

Integration of network & other properties to predict systems-level effects of deleterious mutations

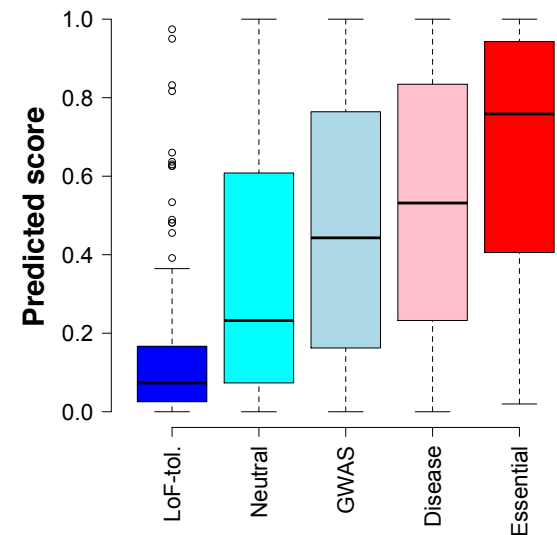
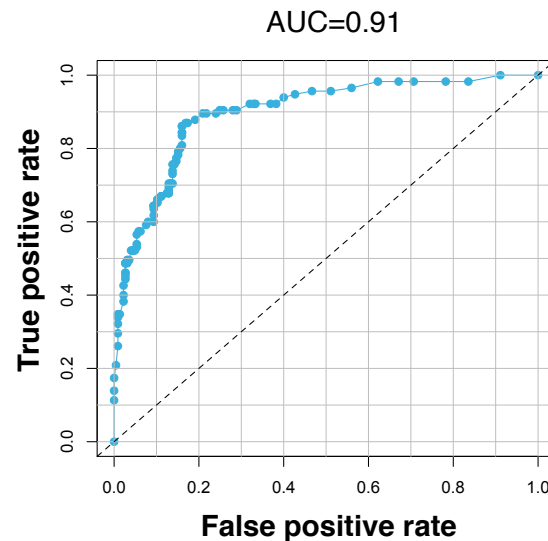
Train logistic regression model using network and evolutionary properties



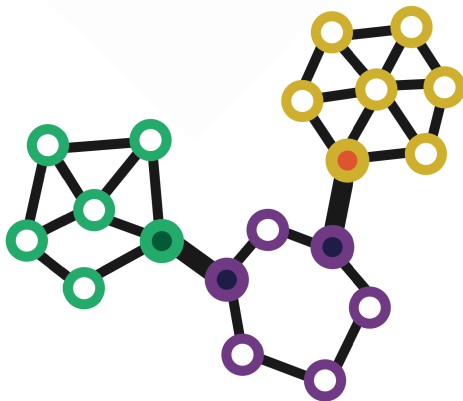
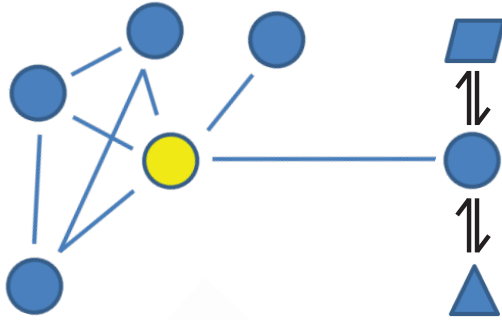
Can distinguish between LoF-tolerant and Essential genes with high accuracy



Application of the model on all genes



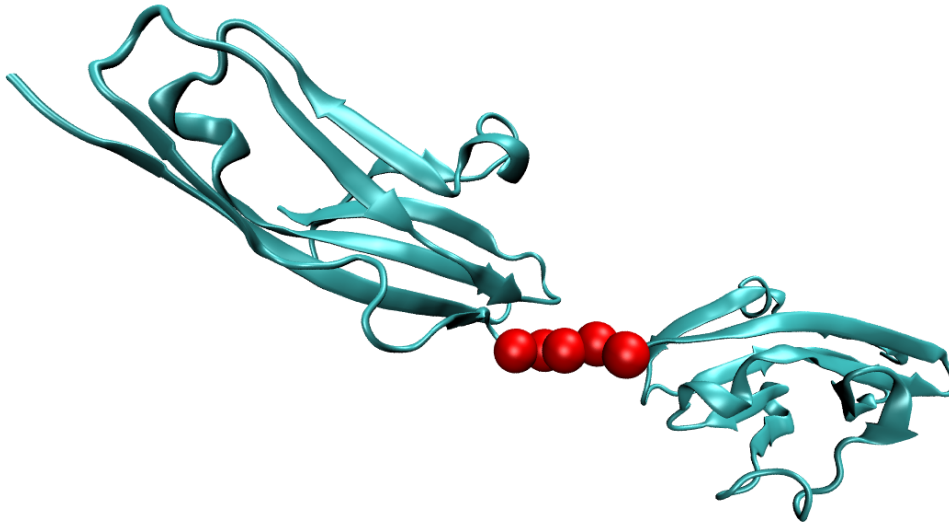
Using 3D-structure
into interpret
networks & deep-
sequencing data



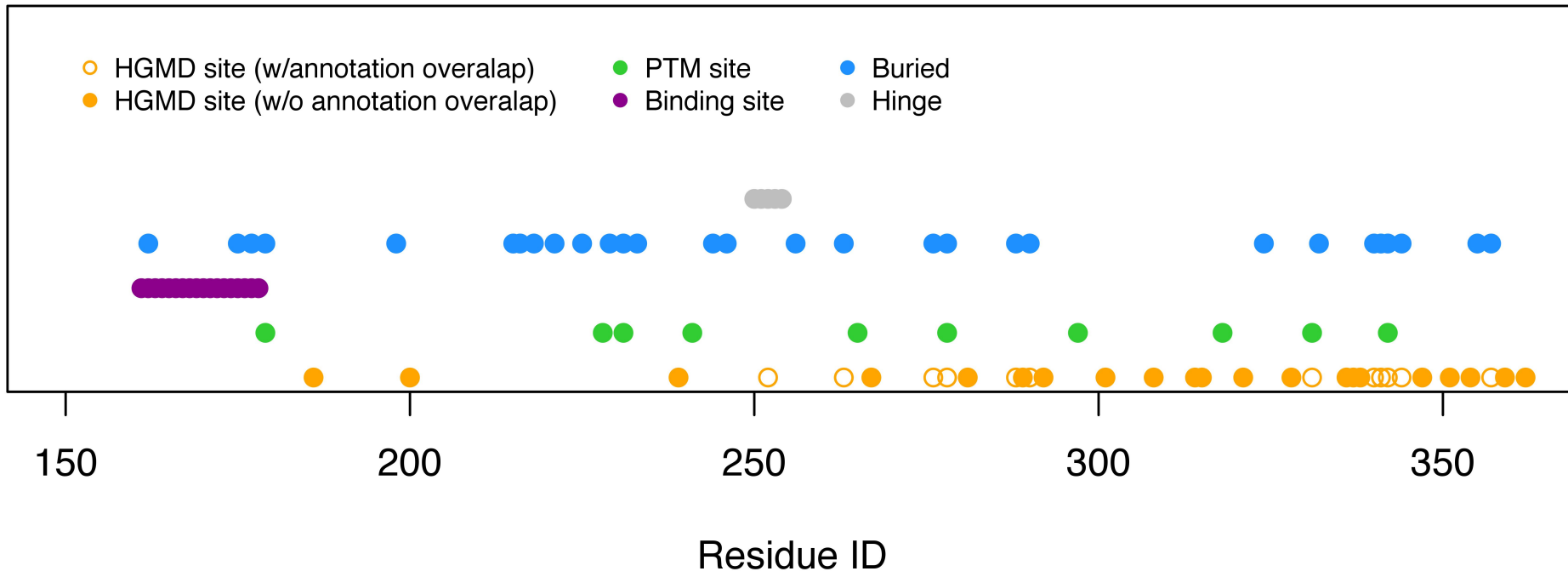
- Structural Interaction Network & Protein Motions (DynaSIN)
 - Multi-interface permanent hubs have more motion than single-interface transient ones
 - Also have more conflicting motions
- LOF variants & Categories of Essential & Disease-sensitive Genes
- Variation at Protein Interfaces in the context of Network Connectivity & its use for Disease-gene Predictions
 - Highly connected parts of PPI under stronger selection but signal weak
 - Stronger signal in SIN & even stronger in multiNet (integration of many networks)
 - Signal strong enough to build predictor
- Rationalizing Deleterious Variants in terms of Potential Allosteric Sites
 - Identifying potential allosteric residues on surface & inside
 - These are under stronger selection & may explain some HGMD SNPs

Interpreting Disease Variants in terms of aspects of Protein Structure

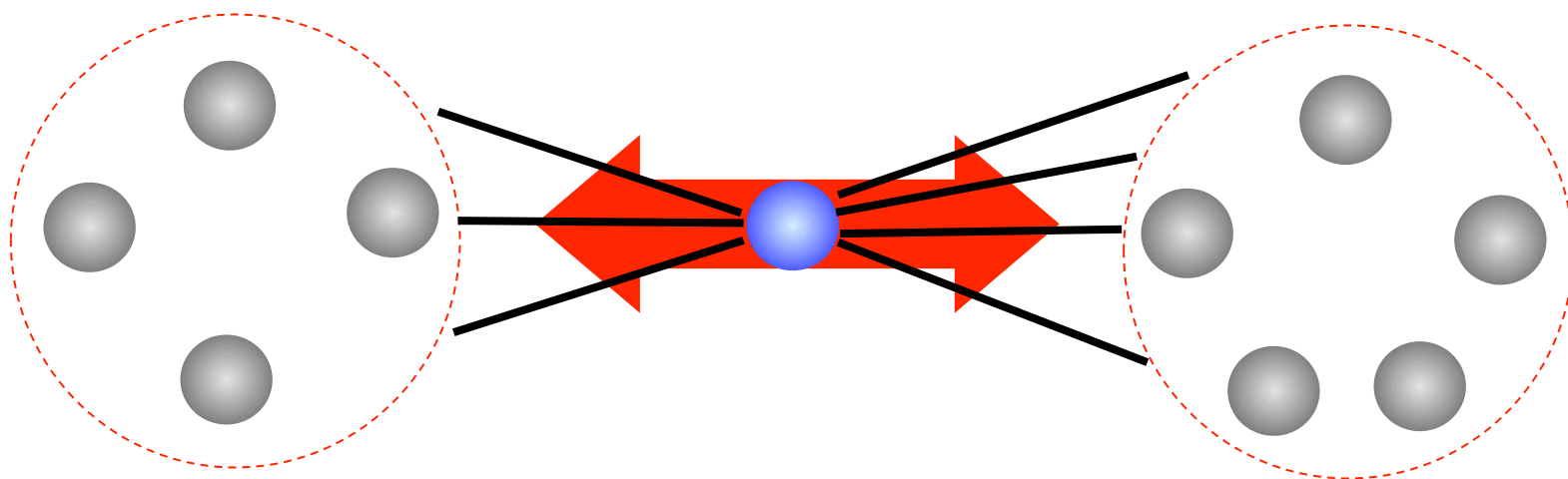
Fibroblast growth factor receptor (pdb 1IIL)



**“STRESS”
pipeline
to tackle
unaccounted for
variants, in terms
of potential
allosteric sites**



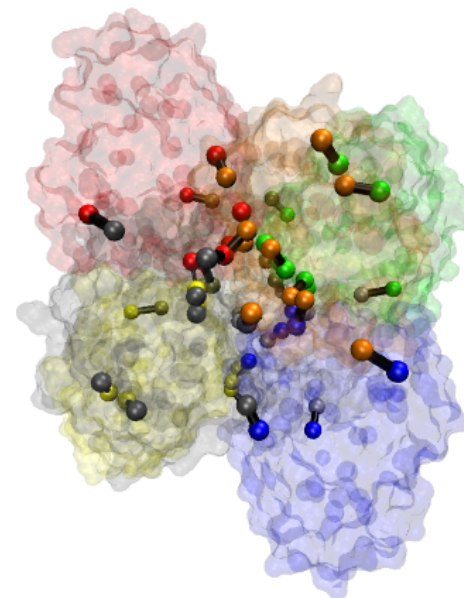
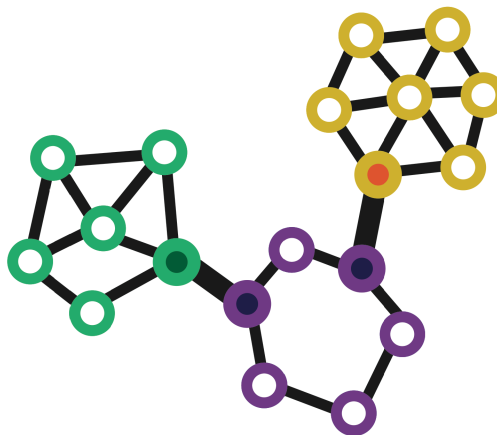
Identifying Potential Allosteric Residues in the Protein Interior



Edge 'distance' between residues i & j is:

$$W_{ij} = -\ln(|C_{ij}|)$$

C_{ij} is the correlation between the motions of residues i & j .
A *large* 'distance' (i.e., low correlated motion) *increases* the shortest path lengths between such residues.

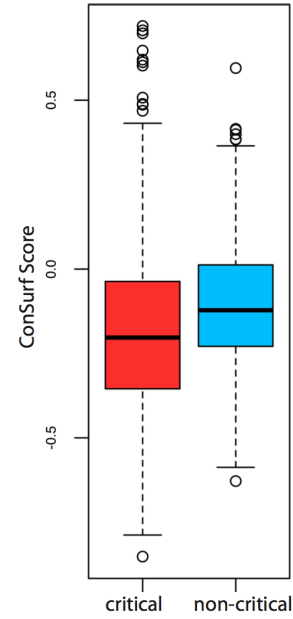
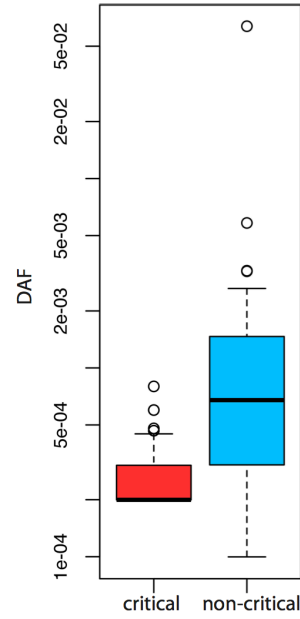
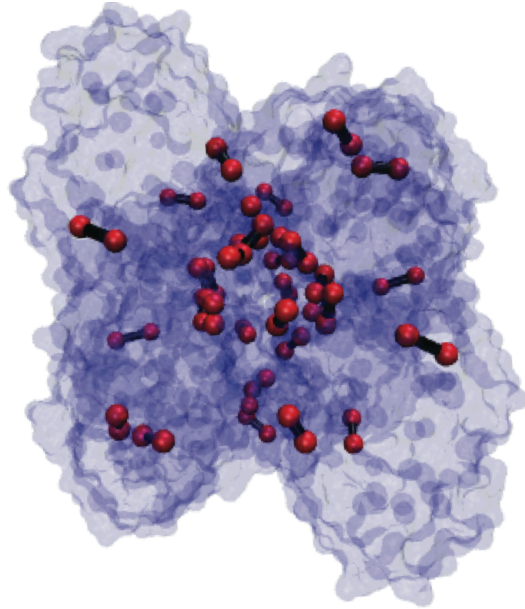
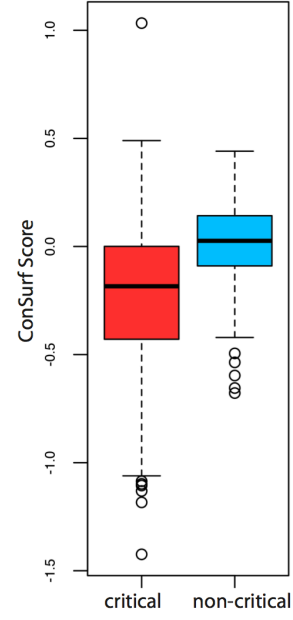
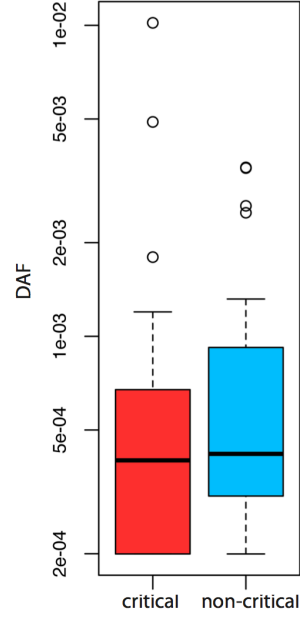
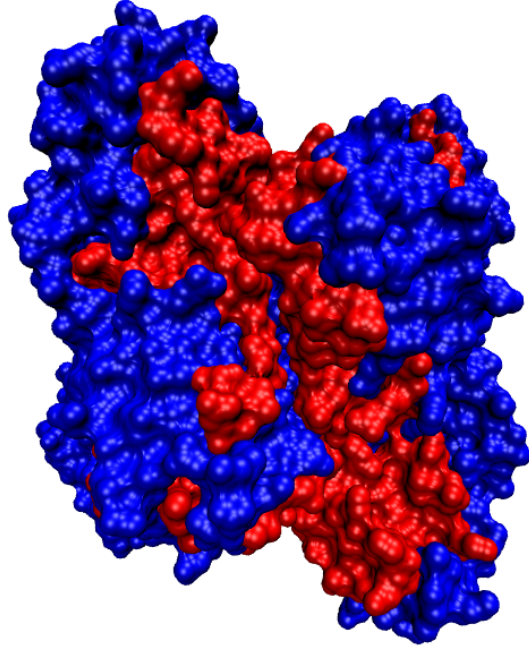


Freeman LC (1977) Set of measures of centrality based on betweenness.

Sociometry 40: 35–41.

Girvan & Newman (2002) PNAS 99: 7821.

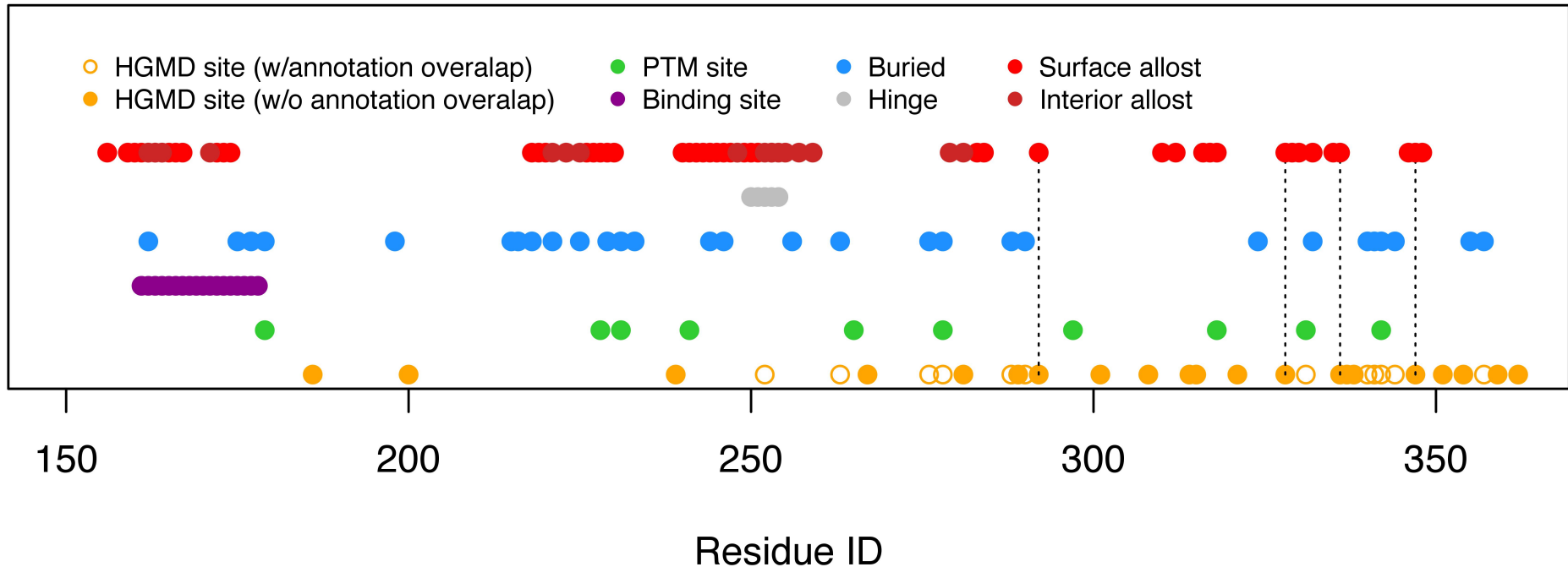
Conservation of network-identified residues implicated in allosteric signal transmission



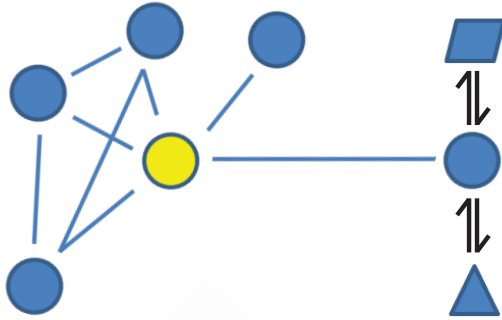
Conservation, allosteric hotspots, and disease variants in sequence space

Fibroblast growth factor receptor (pdb 1IIL)

Dotted lines designate HGMD sites without clear biophysical mechanisms of pathogenicity, but which are nevertheless captured by our pipeline.



Using 3D-structure
into interpret
networks & deep-
sequencing data



- Structural Interaction Network & Protein Motions (DynaSIN)

- Multi-interface permanent hubs have more motion than single-interface transient ones
- Also have more conflicting motions

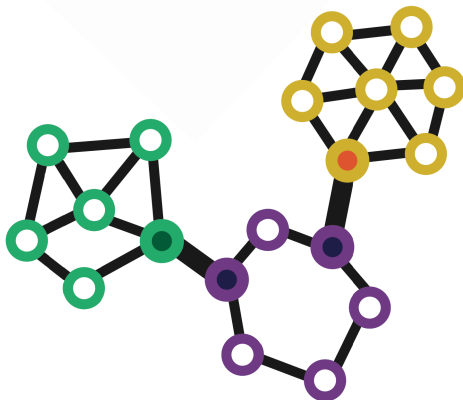
- LOF variants & Categories of Essential & Disease-sensitive Genes

- Variation at Protein Interfaces in the context of Network Connectivity & its use for Disease-gene Predictions

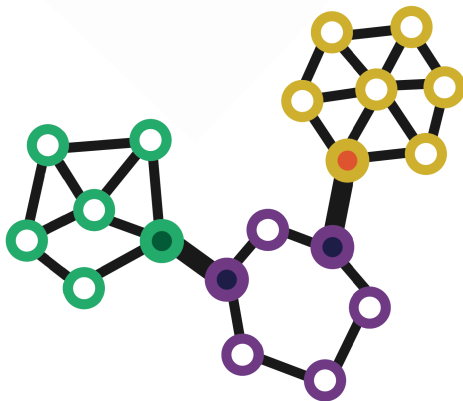
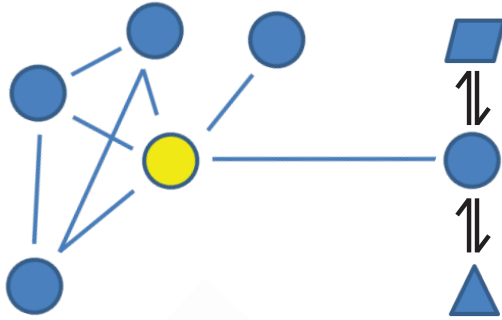
- Highly connected parts of PPI under stronger selection but signal weak
- Stronger signal in SIN & even stronger in multiNet (integration of many networks)
- Signal strong enough to build predictor

- Rationalizing Deleterious Variants in terms of Potential Allosteric Sites

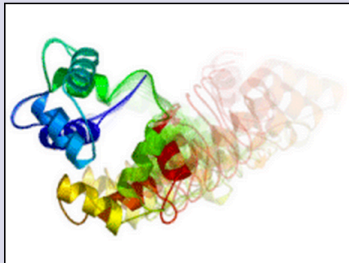
- Identifying potential allosteric residues on surface & inside
- These are under stronger selection & may explain some HGMD SNPs



Using 3D-structure
into interpret
networks & deep-
sequencing data



- Structural Interaction Network & Protein Motions (DynaSIN)
 - Multi-interface permanent hubs have more motion than single-interface transient ones
 - Also have more conflicting motions
- LOF variants & Categories of Essential & Disease-sensitive Genes
- Variation at Protein Interfaces in the context of Network Connectivity & its use for Disease-gene Predictions
 - Highly connected parts of PPI under stronger selection but signal weak
 - Stronger signal in SIN & even stronger in multiNet (integration of many networks)
 - Signal strong enough to build predictor
- Rationalizing Deleterious Variants in terms of Potential Allosteric Sites
 - Identifying potential allosteric residues on surface & inside
 - These are under stronger selection & may explain some HGMD SNPs



Database of Macromolecular Movements with Associated Tools for Flexibility and Geometric Analysis

This describes the motions that occur in proteins and other macromolecules, particularly using movies. Associated with it are a variety of free software tools and servers for structural analysis.

Movies of Conformational Changes

molmovdb.org

Server for morphing complexes

molmovdb.org

Morph Server for multiple subunits and nucleic acids

Database of Alternative Conformations

I. Motions of Fragments Smaller than Domains

A. Motion is predominantly shear

F-s-2. Proteins for which two or more conformations are known



Adenosylcobinamide Kinase [[motion](#)] [[morph](#)]



Small G-protein Arf6 [[motion](#)] [[morph](#)]



Bacteriorhodopsin (BR) [[motion](#)] [[morph](#)]

Quantifying Internal Packing of Residues

PACKING-EFFICIENCY CALCULATOR

The image shows a stylized logo with red lines and spheres, and three diagrams illustrating packing efficiency: a circular arrangement of spheres, a polyhedral arrangement, and a molecular structure within a polyhedral volume.

Identification of Protein Cavities

3V: Voss Volume Voxelator 3v website

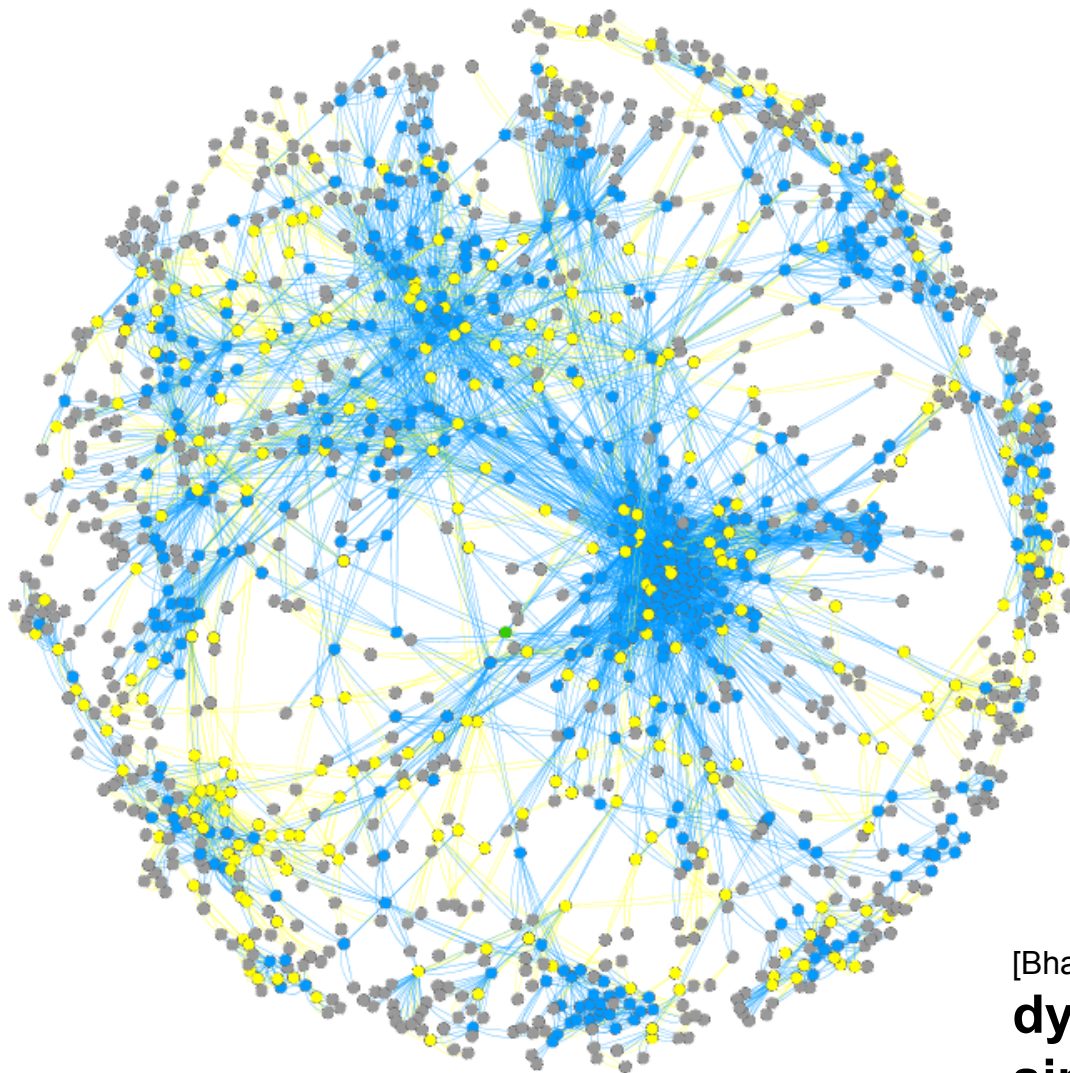
Volume Calculation and Extraction Procedures

3vee is collection of program for the assessment of volumes in protein files.

The image shows a diagram of a protein cavity with 'INSIDE' and 'OUTSIDE' labels, and two 3D surface models of protein structures.

Structural Interaction Network

(v2.0, '11, available for yeast, human, E coli)



○ Non-hub

● Multi-interface hub

● Single-interface hub

— Simultaneously possible interaction: "Permanent"

— Mutually exclusive interaction: "Transient"

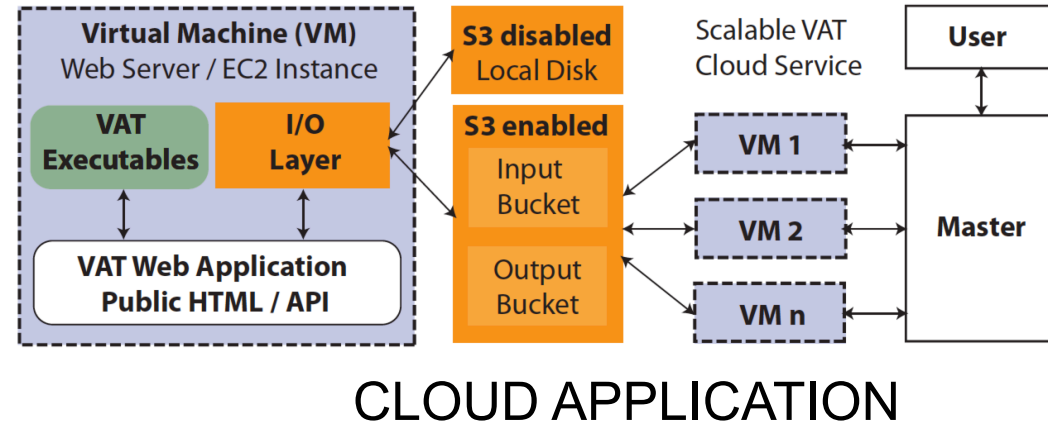
~8.5K Human &
~1.5K yeast
edges

[Bhardwaj et al. ('11) Prot Sci]

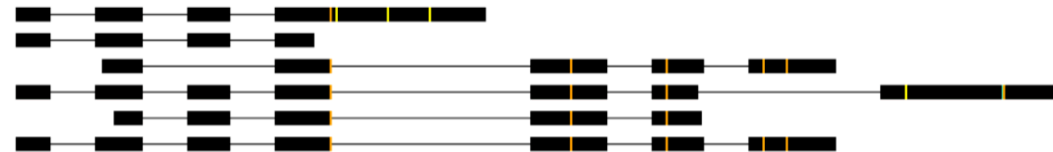
dynasin.molmovdb.org
sin.gersteinlab.org

Variant Annotation Tool (VAT)

- All our annotation from this pipeline
- Input
 - Uses GENCODE (with option of CCDS & other annotations)
 - Overlaps with 1000G SNPs, MNPs, indels & SVs (other input VCFs possible)
- Output
 - Annotated VCFs
 - Graphical representations of functional impact on transcripts
- Access
 - Source freely available
 - Webserver
 - AWS cloud instance



Graphical representation of genetic variants

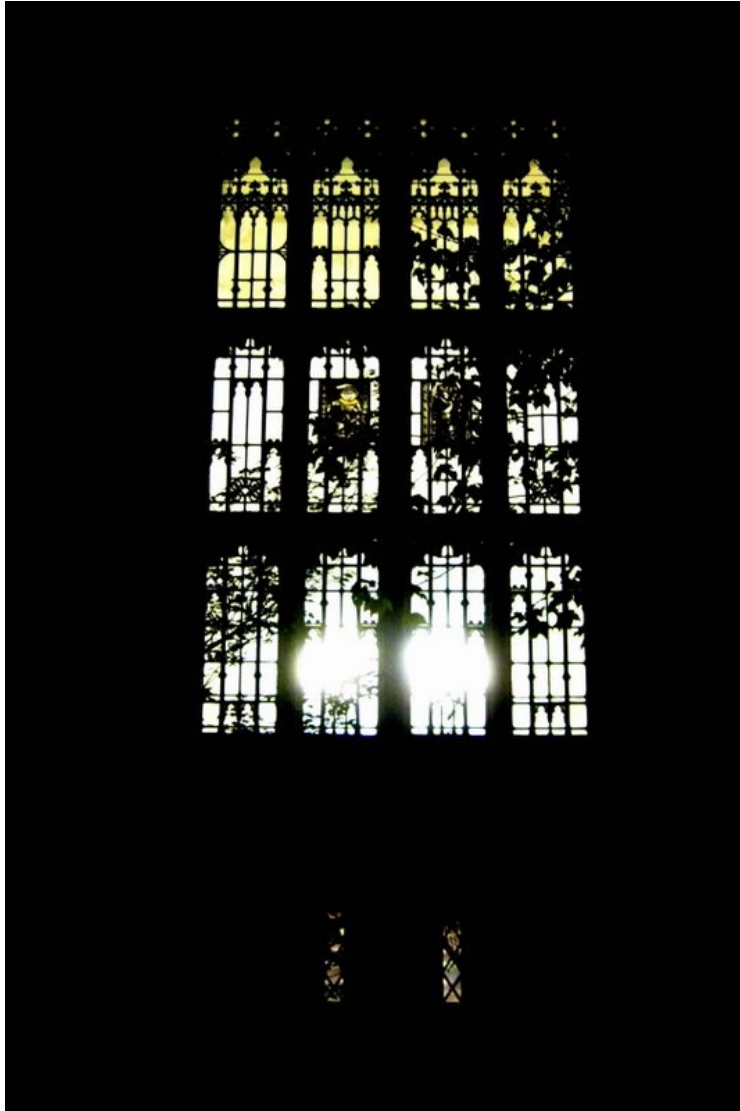


LEGEND FOR VARIATION TYPES:

spliceOverlap synonymous nonsynonymous prematureStop removedStop insertionFS insertionFS deletionFS deletionFS

VAT.gersteinlab.org

Acknowledgments



Hiring Postdocs. See gersteinlab.org/jobs

DynaSIN.molmovdb.org

N **Bhardwaj**, A Abyzov, D Clarke, C Shou

VAT.gersteinlab.org + LOF

L **Habegger**, S **Balasubramanian**, DZ Chen, E Khurana, A Sboner, A Harmanci, J Rozowsky, D Clarke, M Snyder

S **Balasubramanian**, L Habegger, A Frankish, DG

MacArthur, R Harte, C Tyler-Smith, J Harrow,

archive.gersteinlab.org/proj/**NetSNP**

FunSEQ.gersteinlab.org

E **Khurana**, Y **Fu**, V Colonna, XJ Mu, HM Kang, T Lappalainen,

A Sboner, L Lochovsky, J **Chen**, A Harmanci, J Das, A Abyzov, S Balasubramanian, K Beal, D Chakravarty, D Challis, Y Chen, D Clarke, L Clarke, F Cunningham, US Evani, P Flicek, R Fragoza, E Garrison, R Gibbs, ZH Gumus, J Herrero, N Kitabayashi, Y Kong, K Lage, V Liluashvili, SM Lipkin, DG MacArthur, G Marth, D Muzny, TH Pers, GR

Ritchie, JA Rosenfeld, C Sisu, X Wei, M Wilson, Y Xue, F Yu, **1000**

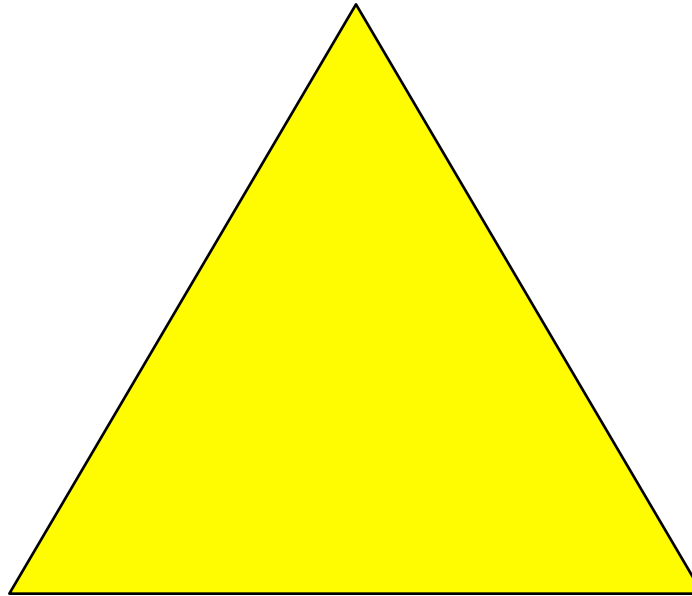
Genomes Project Consortium, ET Dermitzakis, H Yu, MA Rubin, C Tyler-Smith

STRESS

A **Sethi**, D **Clarke**, S Kumar, S Li, R Chang, KK Yan, J Chen

Default Theme

- Default Outline Level 1
 - Level 2



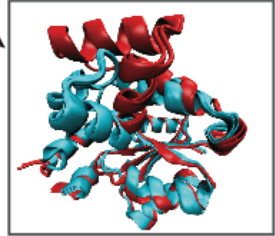
Pipeline (“STRESS”) for culling alternative conformations and predicting allosterically important residues

Sequence alignments

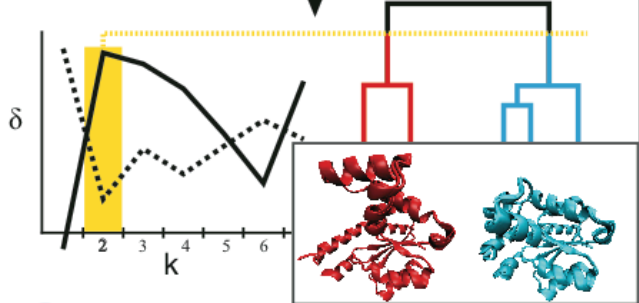
Filtered Structure Databases (PDB, SCOP)

Sequence Group 1:
 RLGTGFAARVMLVKKKESAGRDLFPVDRDOL
 RLGTGDFLRVMLVKKKESGRDLFWGTDHIWOL
 RVGTRDFARVMLVKKKESGRDLFPGTDBIDOL
 RLGTGDFARVMLLKKKESGRDLFACTDHIADOL
 ...
 Sequence Group 2:
 GSEQESVKEFLRKAKEDGLAGGVTYLWDEANR
 GSAQESVKEFLVKAKEDTCAGGVVYLDDEANR
 GSEQESVKEFLAKAFEDFAAGVVTYLDDEANR
 GSEQESVKEFLAKAKEDFDAGGVVYLDDEANR
 ...
 Sequence Group 3:
 VRDLKVENLLIDQQCYDQVWALGOLIVEMAAE
 VRDLVVENLLIDQQCEIQVWALGVLRLVWQAG
 YRALRQQLHIDDMGCIQVWVLCVLIYWMMAE
 VRDLAEPNDLIDQQCNAGVWALGVDENVMAAE
 ...

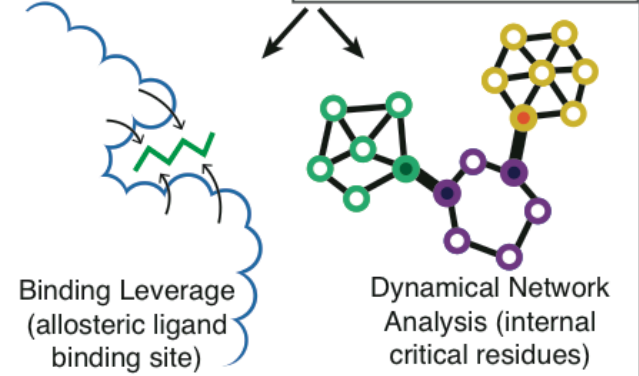
Structure alignment

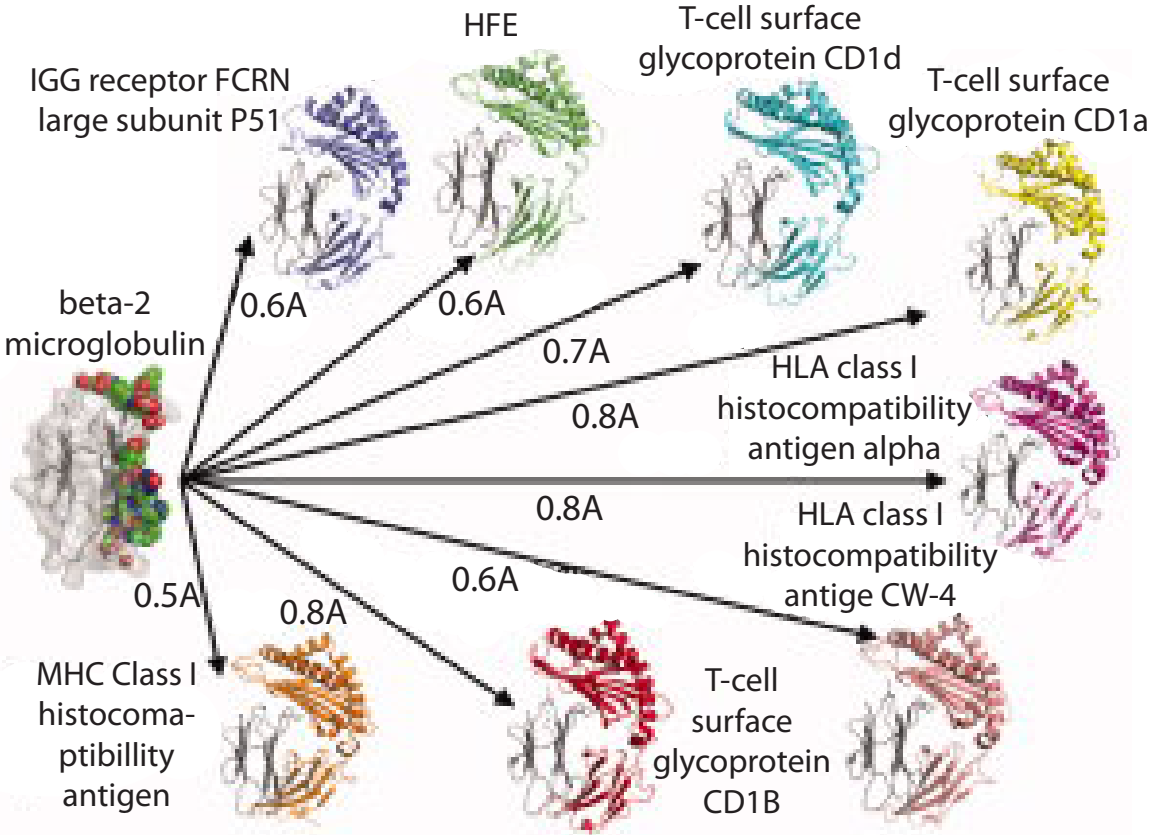


K-means Clustering with Gap Statistic



Allosteric hotspot prediction

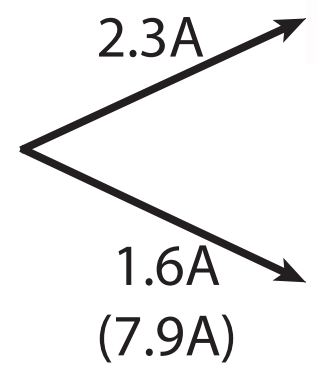




Conformational changes associated with "transient" interactions

Conformational changes associated with "permanent" interactions

Importin subunit beta-1



with GTP-binding nuclear protein RAN



with Snurportin-1

Objective: Using 3D-structure into interpret PPI networks & mutations from deep-sequencing data

- Growing numbers of complex structures with many interfaces allow structure to be related to networks
- Growing proportion of structures with same fold allow probing conformational plasticity & motions
- Vast increase exome data provides new ways to think about coding mutations – eg in terms of selection & allele freq. In turn, structure & networks provide interpreting these data
- Sites associated with allosteric motions provide a way interpreting disease mutations, not accounted for otherwise

More Information on this Talk

SUBJECT: Networks

DESCRIPTION:

NOTES:

This PPT should work on mac & PC. Paper references in the talk were mostly from Papers.GersteinLab.org.

PERMISSIONS: This Presentation is copyright Mark Gerstein, Yale University, 2010. Please read permissions statement at <http://www.gersteinlab.org/misc/permissions.html> . Feel free to use images in the talk with PROPER acknowledgement (via citation to relevant papers or link to gersteinlab.org).

PHOTOS & IMAGES. For thoughts on the source and permissions of many of the photos and clipped images in this presentation see <http://streams.gerstein.info> . In particular, many of the images have particular EXIF tags, such as **kwpotppt** , that can be easily queried from flickr, viz: <http://www.flickr.com/photos/mbgmbg/tags/kwpotppt> .