Protein Similarity Networks (PSNs)

Basics
Using PSNs
Challenges for Interpretation

Pennsylvania State Bioinorganic Workshop

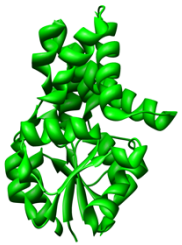
Patsy Babbitt & Shoshana Brown

babbitt@cgl.ucsf.edu

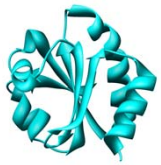
June 2016

Most protein superfamilies are now too large to manage & explore using multiple sequence alignments & trees

Enolase ~50,000 seqs

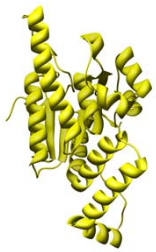


Haloalkanoic acid dehalogenase >80,000



Glutathione transferase >50,000

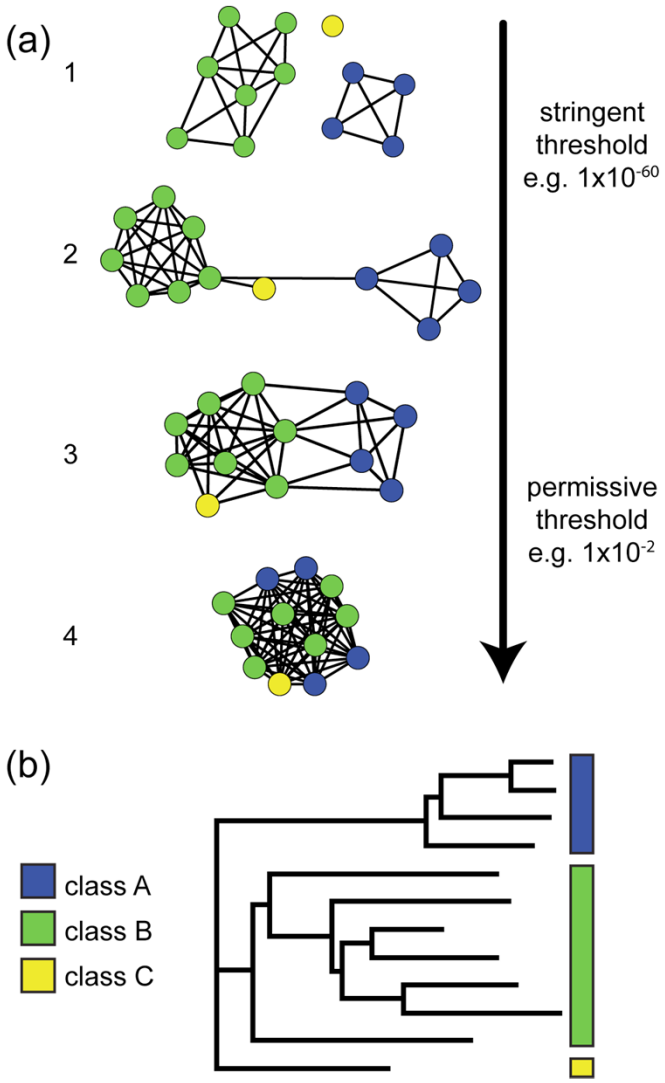
Isoprene synthase I: >15,000



Enoyl CoA hydratase >100,000

- Practical limits of multiple sequence alignment for visualization <100 sequences
- High quality phylogenetic trees limited by many factors, including alignment quality & sequence diversity

“Identification of functional trends from the context of sequence similarity”



- > Handles thousands of sequences, hundreds of structures
- > Fast to generate from pairwise comparisons
- > Interactive visualization (Cytoscape) allows mapping of nodes to many types of functional & other features
- > Thresholded networks enable exploration of similarity relationships across divergence levels
- > Many types: sequence, structure, ligand, reaction, active site signatures, HMMs...
- > Built on earlier work from Pajek, Ouzounis (Tribe-MCL), Marcotte, others

Node = sequence (or structure)

Edge = connections between sequences w scores as good as the E-value cutoff threhold

Simplest form: sequence similarity networks generated from all-by-all BLAST (using *E*-values as scores); structure: all-by-all FAST (or TM-Align) scores

Atkinson HJ & Babbitt PC. *PLoS ONE* (2009) 4: e4345; built on earlier work by Pajek, Ouzounis, Marcotte, others

Validation

Comparison of distance metrics for generating networks shows that BLAST correlates well with several other metrics

Uronatel / **MLE**	BLAST	SW	MA	PT
BLAST		0.999	0.971	0.953
Smith-Waterman (SW)	0.998		0.970	0.953
Multiple Alignment (MA)	0.800	0.798		0.974
Phylogenetic Tree (PT)	0.731	0.731	0.777	

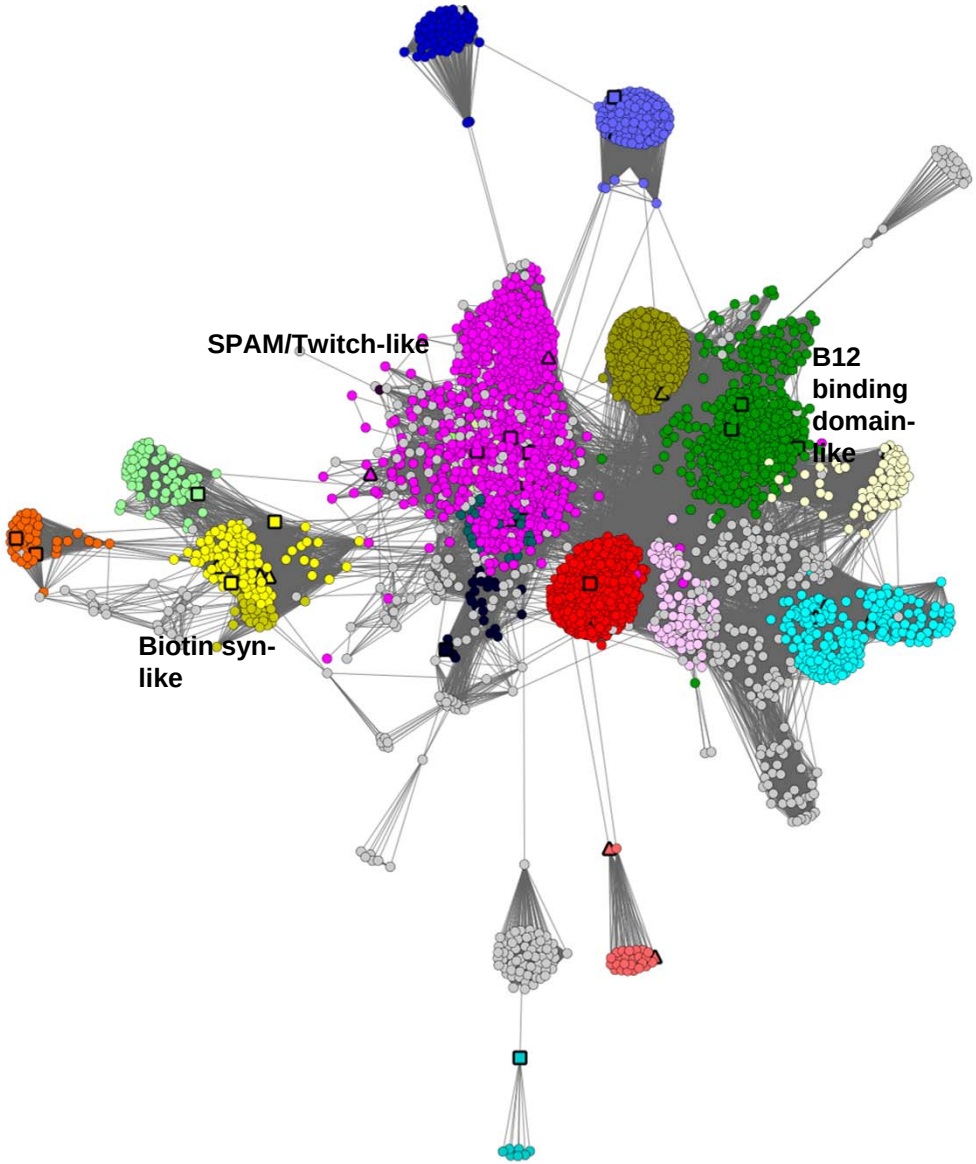
NagA	BLAST	SW	MA	PT
BLAST		0.997	0.841	0.748
Smith-Waterman (SW)			0.846	0.753
Multiple Alignment (MA)				0.719
Phylogenetic Tree (PT)				

R^2 values for linear regressions of distances generated from various metrics for scoring similarity among sequences

Other validation analyses show

- Network topologies are generally robust to missing data
- Two-dimensional distances in visualized networks correlate well with the underlying distances in high-dimensional space
  - See Atkinson et al, PLoS ONE, 4: e4345 (2009) for more statistical validation

Colors/shapes depict many types of features



Context for leveraging information about “knowns” to tell us about “unknowns” that are too numerous to characterize experimentally

Structure-function relationships across entire superfamilies can be summarized or dissected in detail starting from a single network

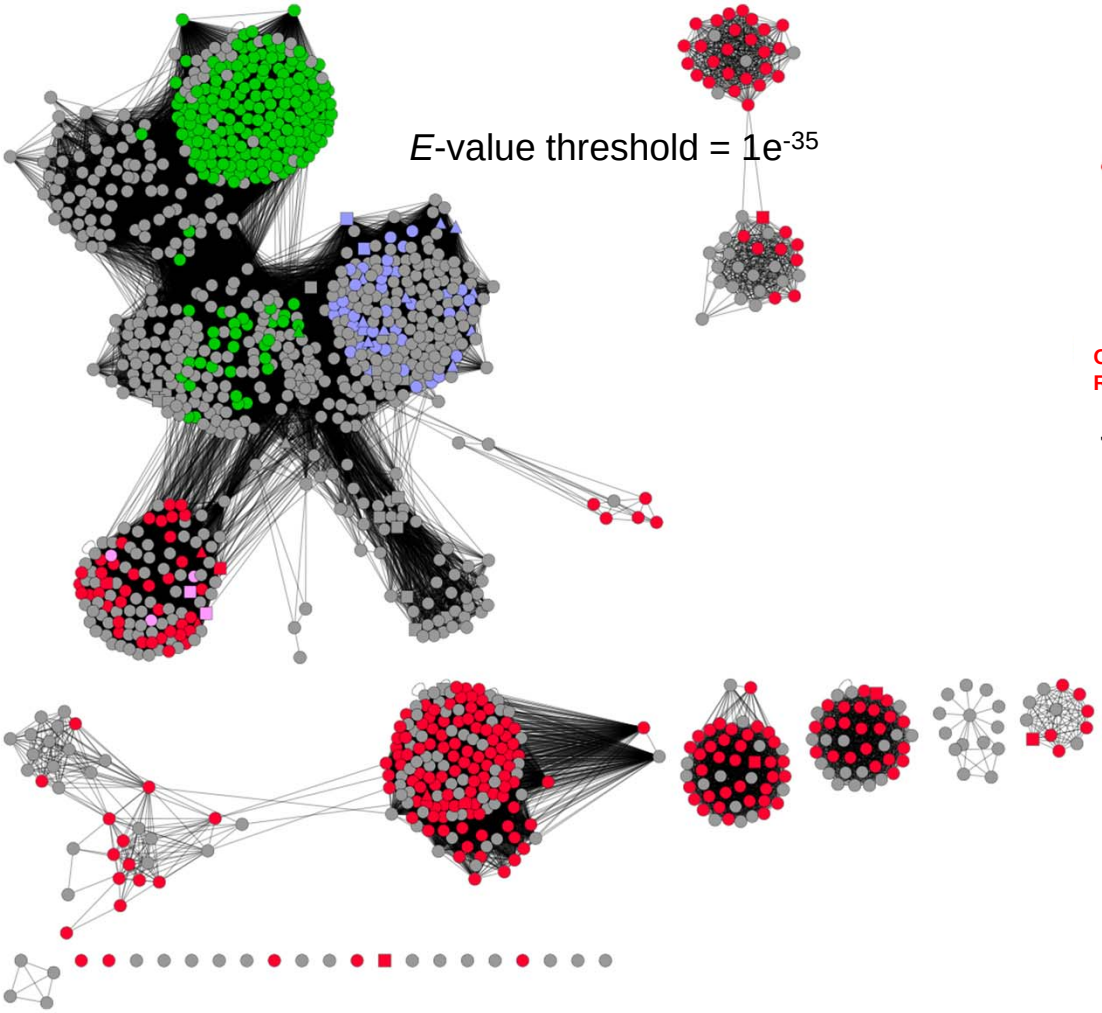
Intuitively accessible survey of experimental & structural coverage of a superfamily

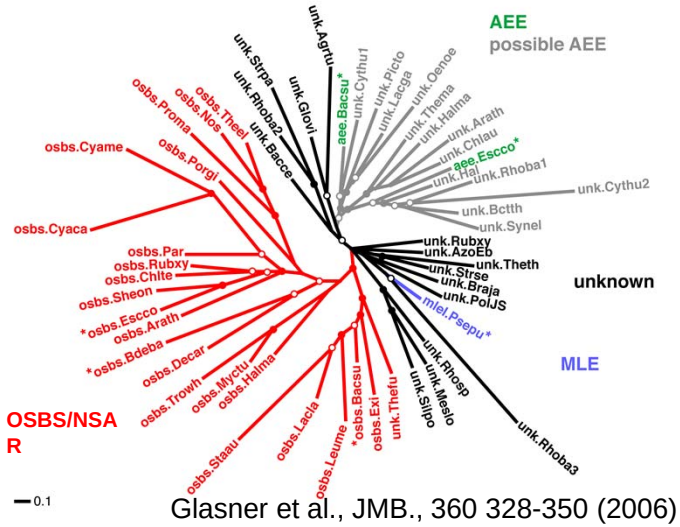
Radical SAM PSN : Gemma Holliday, with Squire Booker

Generally similar topologies in networks & trees

Each provides different types of information

But networks can't substitute for trees!





Glasner et al., JMB., 360 328-350 (2006)

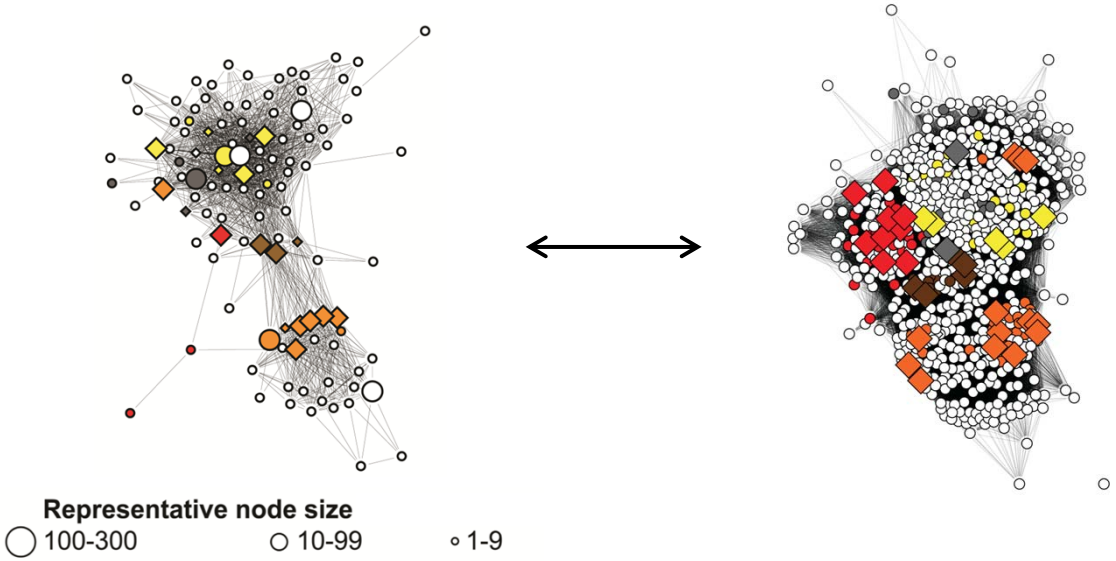
Dipeptide epimerase
MLE
OSBS
Unknown

Shoshana Brown

The big data problem: Working with SSNs in Cytoscape limited by the number of edges (~500,000)

Representative networks address size limitations of PSNs:PYTHOSCAPE*

Barber & Babbitt, Bioinf, 28:2845 (2012)



RepNets summarize sequence relationships for very large networks

- Each representative node contains 1- >1000 sequences binned at a user-chosen pairwise % identity
- Edges between nodes capture similarities computed from pairwise BLAST scores between repnodes, computed as the median of all the similarity scores in eachrepnode (or the 2 most similar, or least similar, etc)
  - A good resource for generating your own networks: http://efi.igb.illinois.edu

How to use PSNs

Depends on your question

Interpretation of structure-function relationships from a large-scale context

- Selecting targets for biochemical/structural characterization
- Functional clues about knowns & unknowns
- Discovery of specificity determinants distinguishing different reaction famiies of a functionally diverse enzyme superfamily
- Comparative proteomics

Linkers

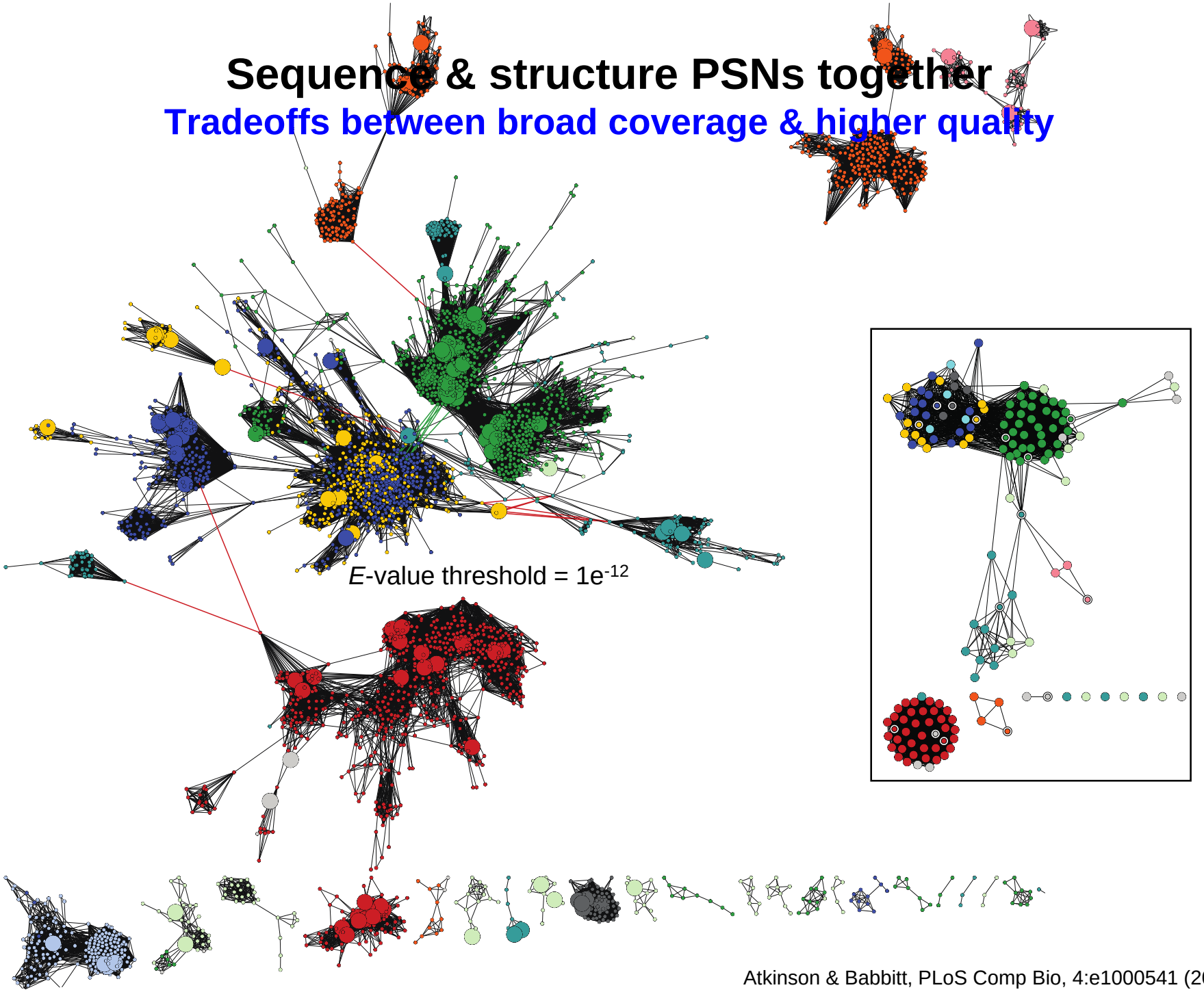
- Using the large-scale context to hypothesize functional “transitions” between functional families

Applications to metagenomics data

Genome context networks

Sequence & structure PSNs together

Tradeoffs between broad coverage & higher quality

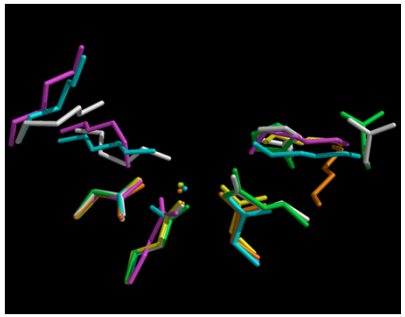


Atkinson & Babbitt, PLoS Comp Bio, 4:e1000541 (2009)

Uses of PSNs at different levels of granularity

Superfamily

All homologs with some common functional attributes

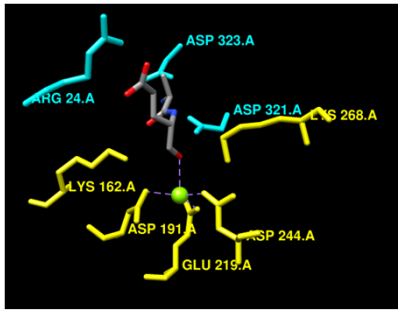


Enolase Superfamily

Enolase Subgroup | MR Subgroup | MLE Subgroup

Subgroup

Enolase Family | MR Family | GalD Family | RhamD Family | MLE1 Family | ChMLE Family | NSAR Family | OSBS Family | Dipeptide Epimerase Family

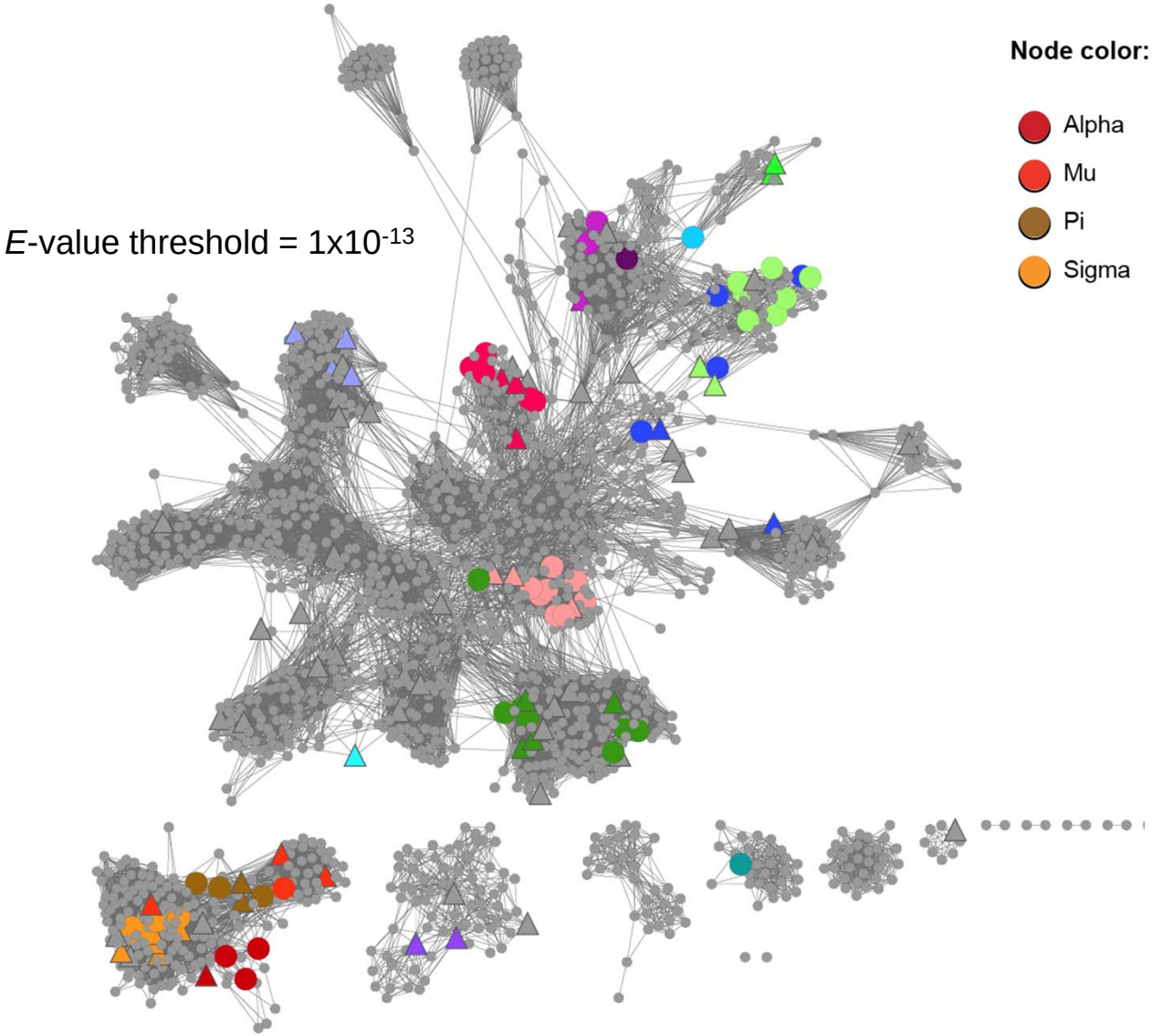


Family

Variation in active site/other features distinguish reaction specificity for each family

Large-scale structure-function mapping:GST superfamily

(Exercise #1: Target selection)



Node color: Swiss-Prot family annotation

- Alpha
- Mu
- Pi
- Sigma
- Beta
- DHAR
- FAX
- HSP26
- Lambda
- chloride channel CLIC
- Omega
- Phi
- Tau
- Theta
- Zeta

"New" classes

- Nu
- Xi

1,568 repnodes represent 13,000 sequences (50% ID filtered)

Median edge *E*-value = 4x10^{-25}

Median alignment length = 210 residues

Large nodes colored by SwissProt classification if >50% of sequences in each node belong to that class

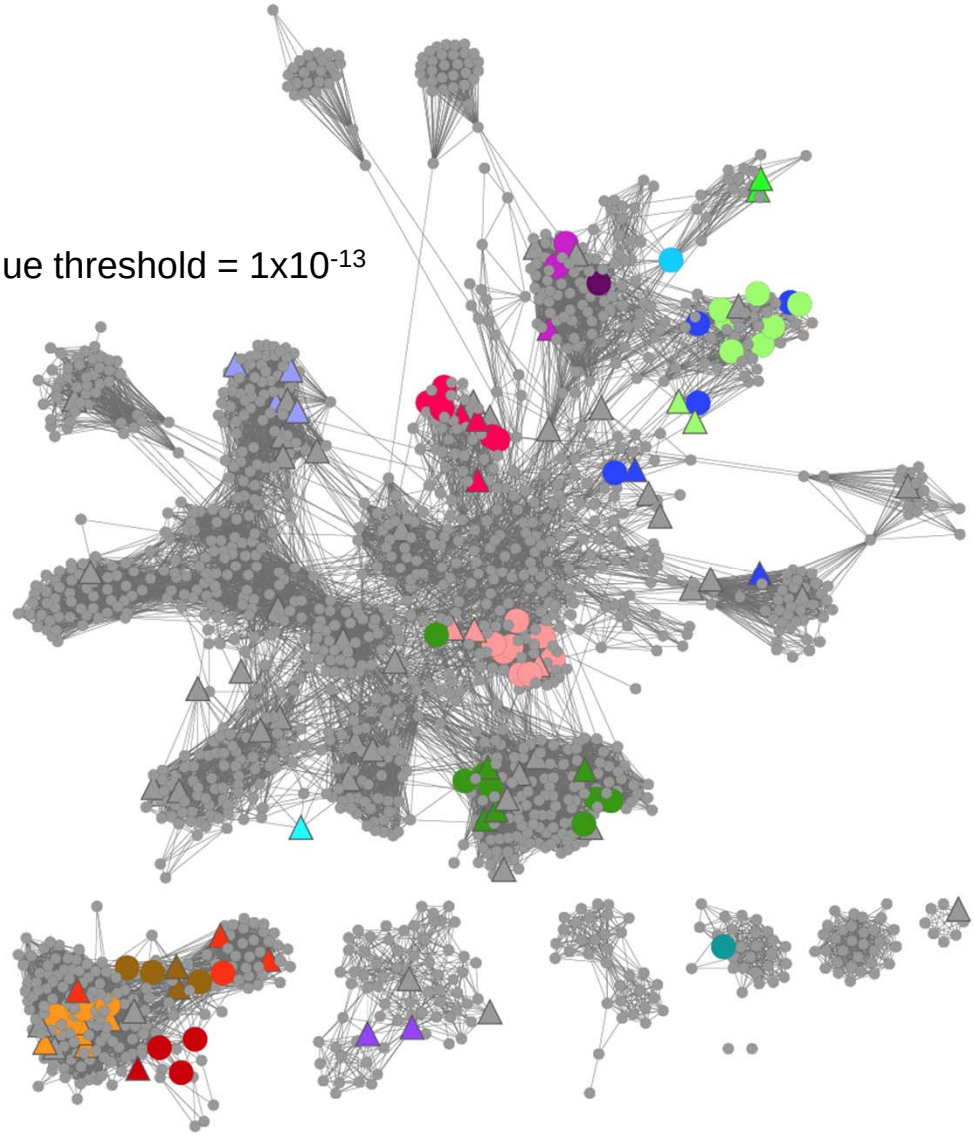
Gray: Not classified

Triangle: PDB structure

Mashiyama et al, PLoS Biol, 12: e1001843 (2014)

How little we know

E-value threshold = $1x10^{-13}$



Node color: Swiss-Prot family annotation

- Alpha
- Beta
- Omega
- Mu
- DHAR
- Phi
- Pi
- FAX
- Tau
- Sigma
- HSP26
- Theta
- Lambda
- Zeta
- chloride channel CLIC

"New" classes

- Nu
- Xi

1,568 nodes represent 13,000 sequences (50% ID filtered)

Median edge *E*-value = $4x10^{-25}$

Median alignment length = 210 residues

Large nodes colored by SwissProt classification if >50% of sequences in each node belong to that class

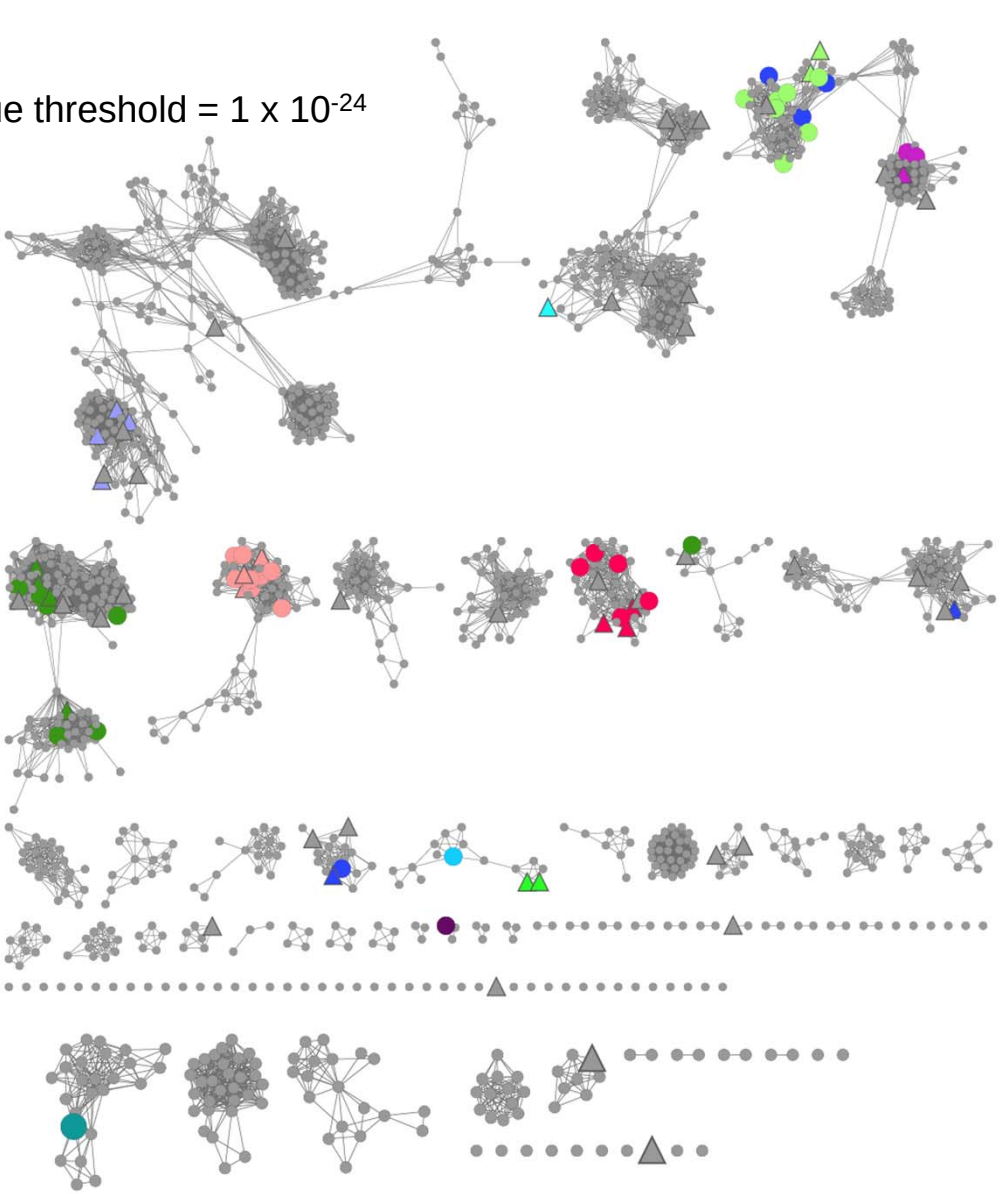
Gray: Not classified

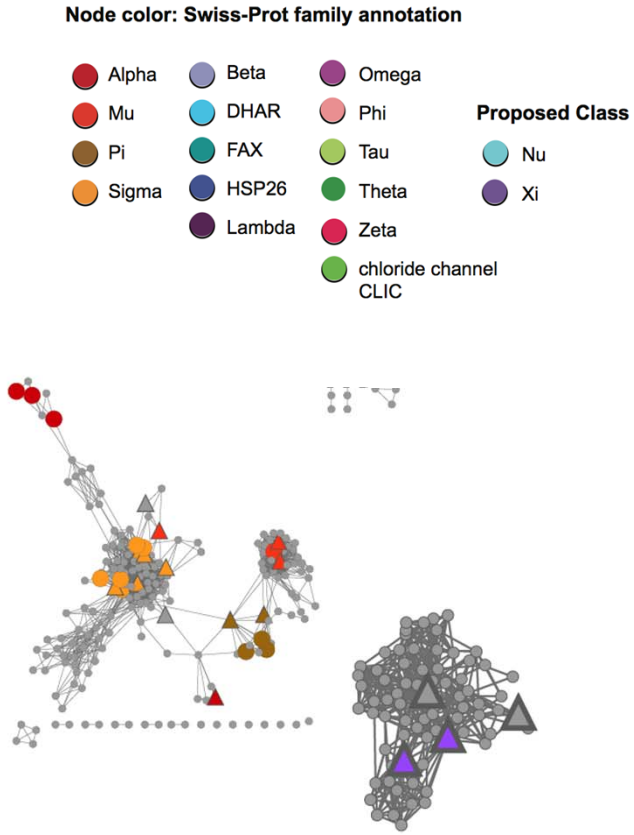
Triangle: PDB structure

Mashiyama et al, PLoS Biol, 12: e1001843 (2014)

Changing thresholds for drawing edges enables facile exploration at varying levels of detail

E-value threshold = 1 x 10^{-24}

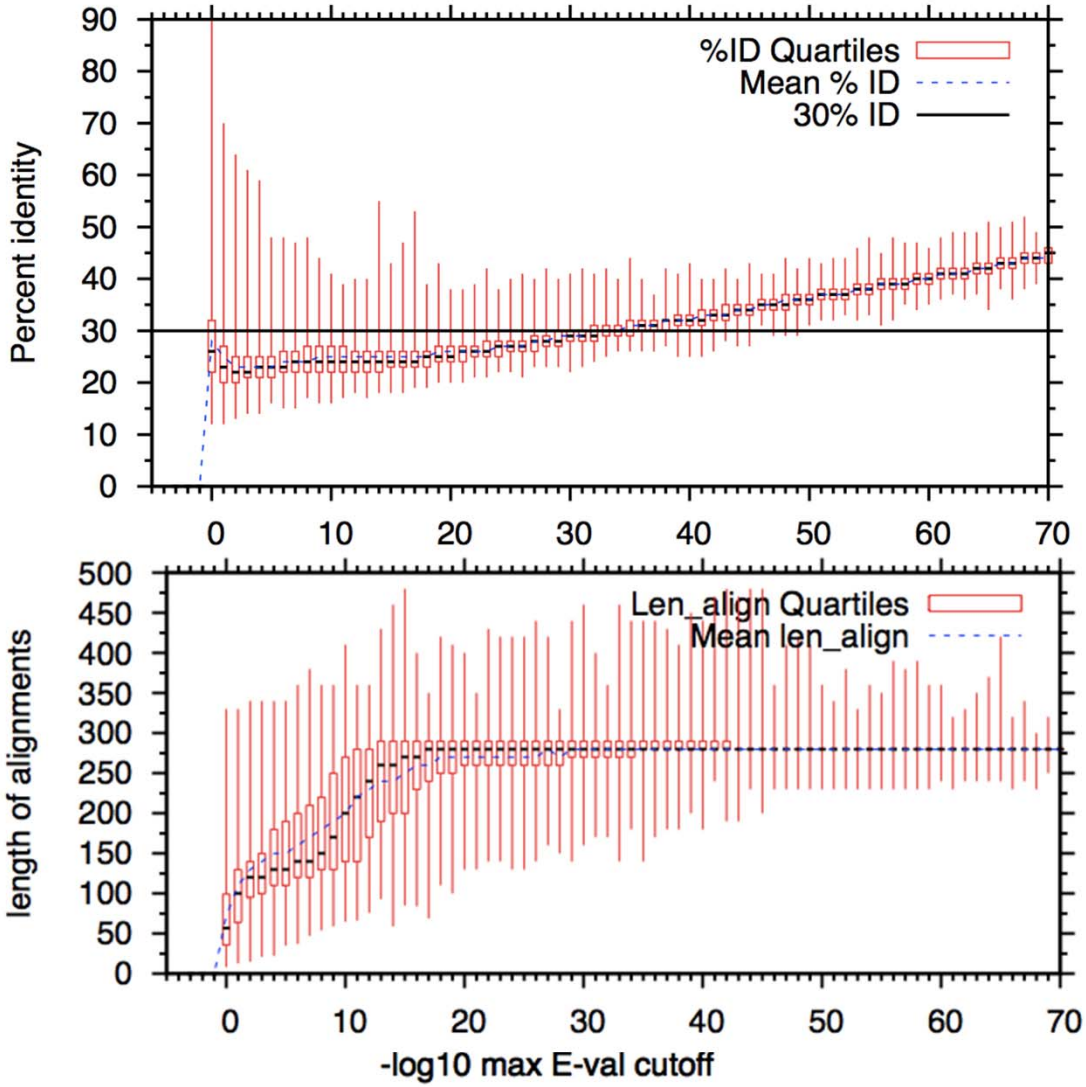




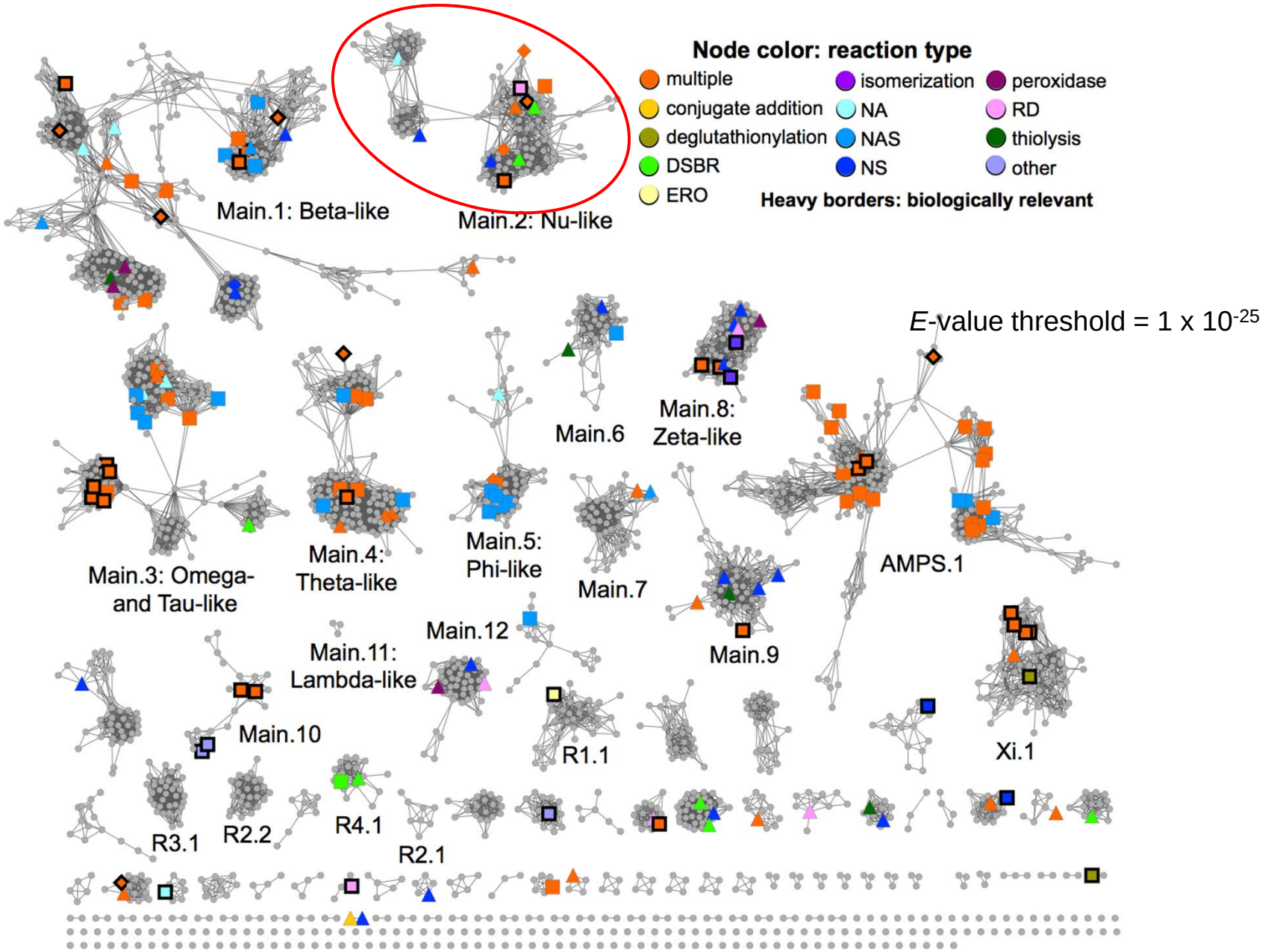
Median *E*-value = 6 x 10^{-35}

Median alignment length = 212

Choosing initial thresholds for visualization

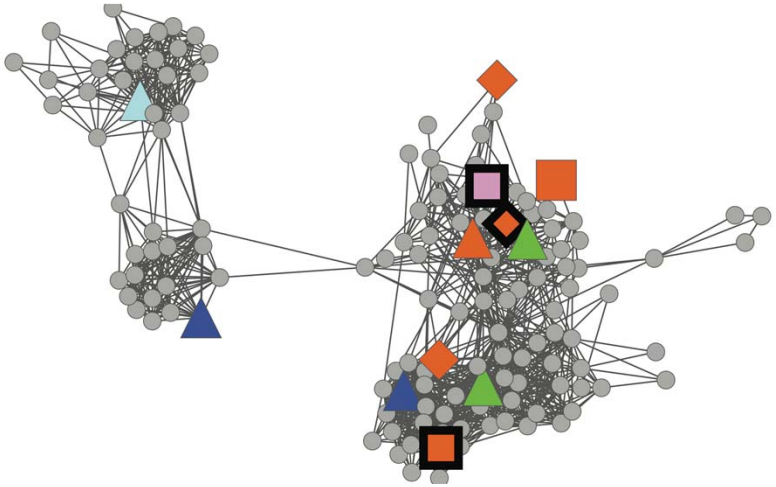


GSTs are highly promiscuous

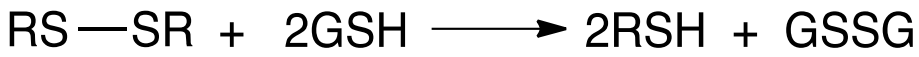


E-value threshold = 1 x 10^{-25}

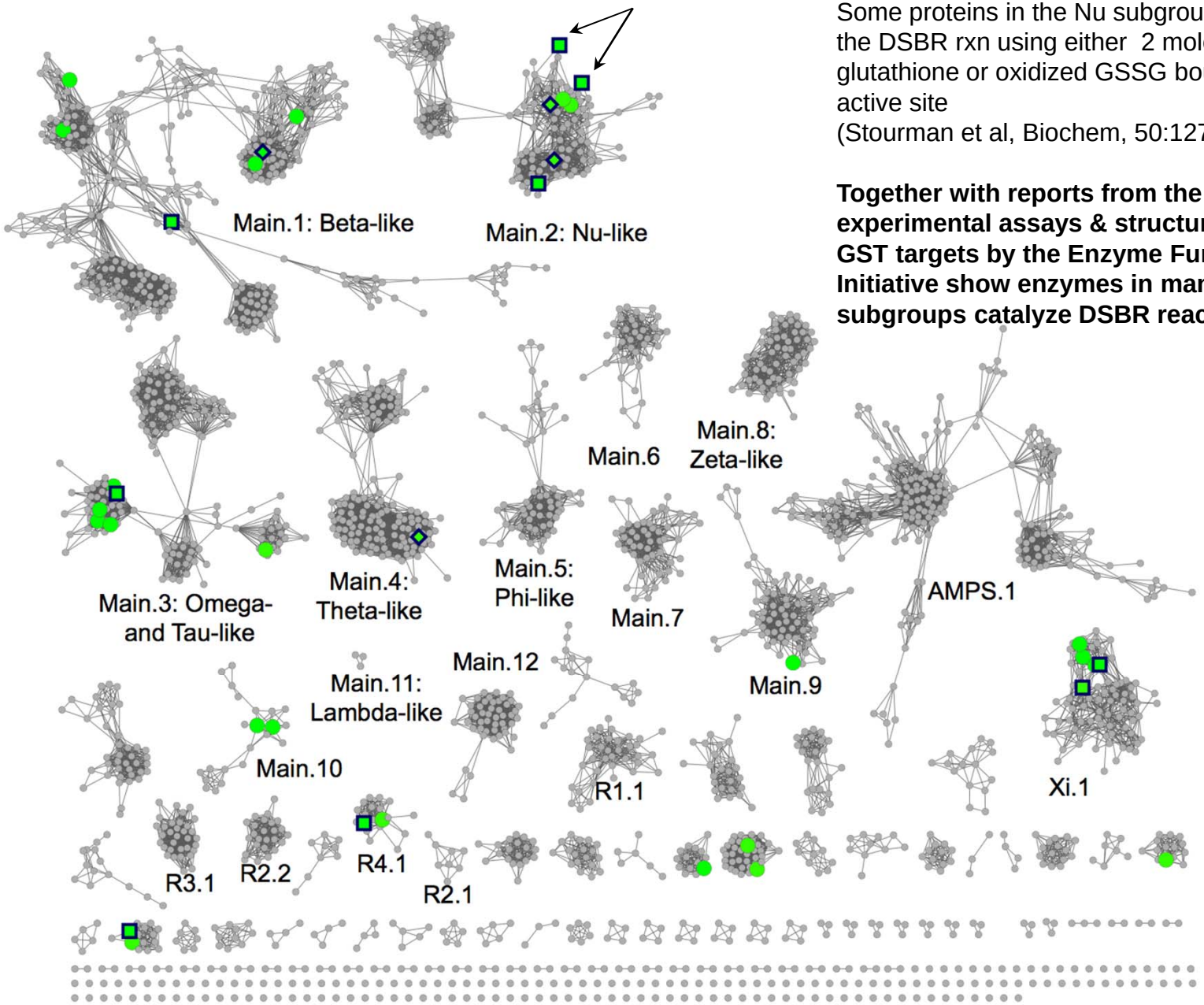
Disulfide bond reductase (DSBR) activity



Some proteins in the Nu subgroup catalyze the DSBR rxn using either 2 molecules of glutathione or oxidized GSSG bound in the active site (Stourman et al, Biochem, 50:1274 (2011))



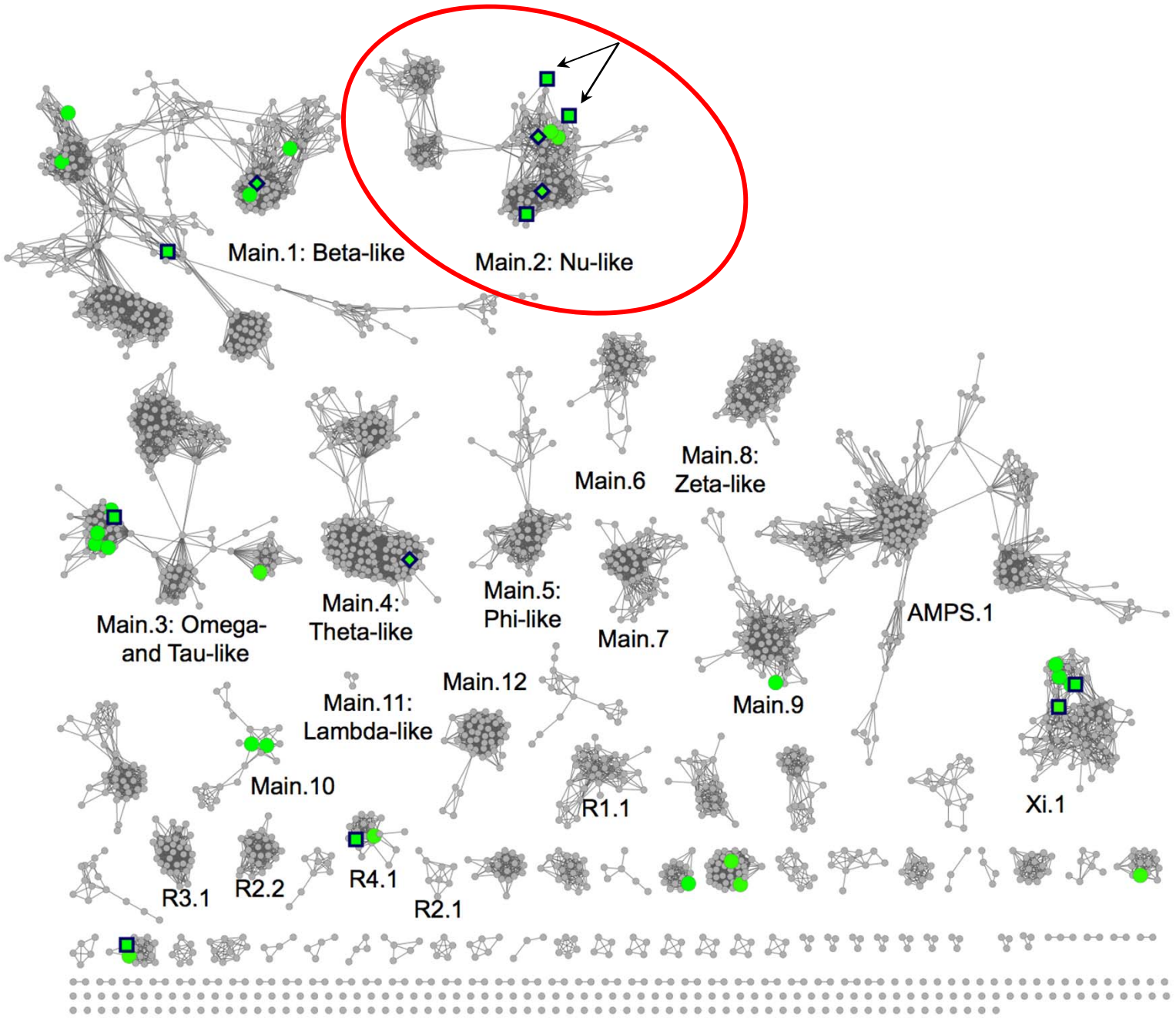
Many selected targets also catalyze DSBR rxn



Some proteins in the Nu subgroup catalyze the DSBR rxn using either 2 molecules of glutathione or oxidized GSSG bound in the active site
(Stourman et al, Biochem, 50:1274 (2011))

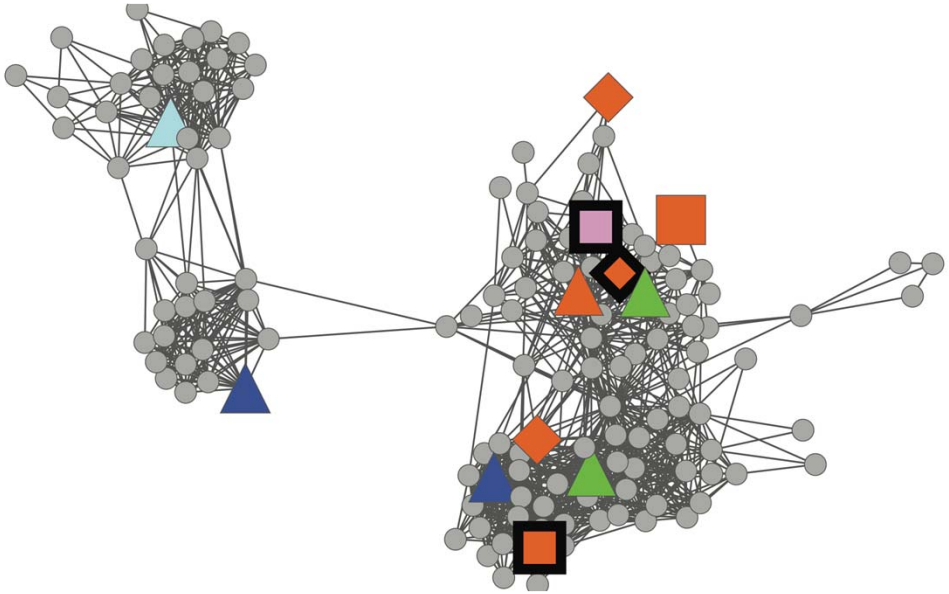
Together with reports from the literature, experimental assays & structures for new GST targets by the Enzyme Function Initiative show enzymes in many subgroups catalyze DSBR reactions

Target selection for investigating mechanism



Drilling down for more detail: RepNets vs FullNets

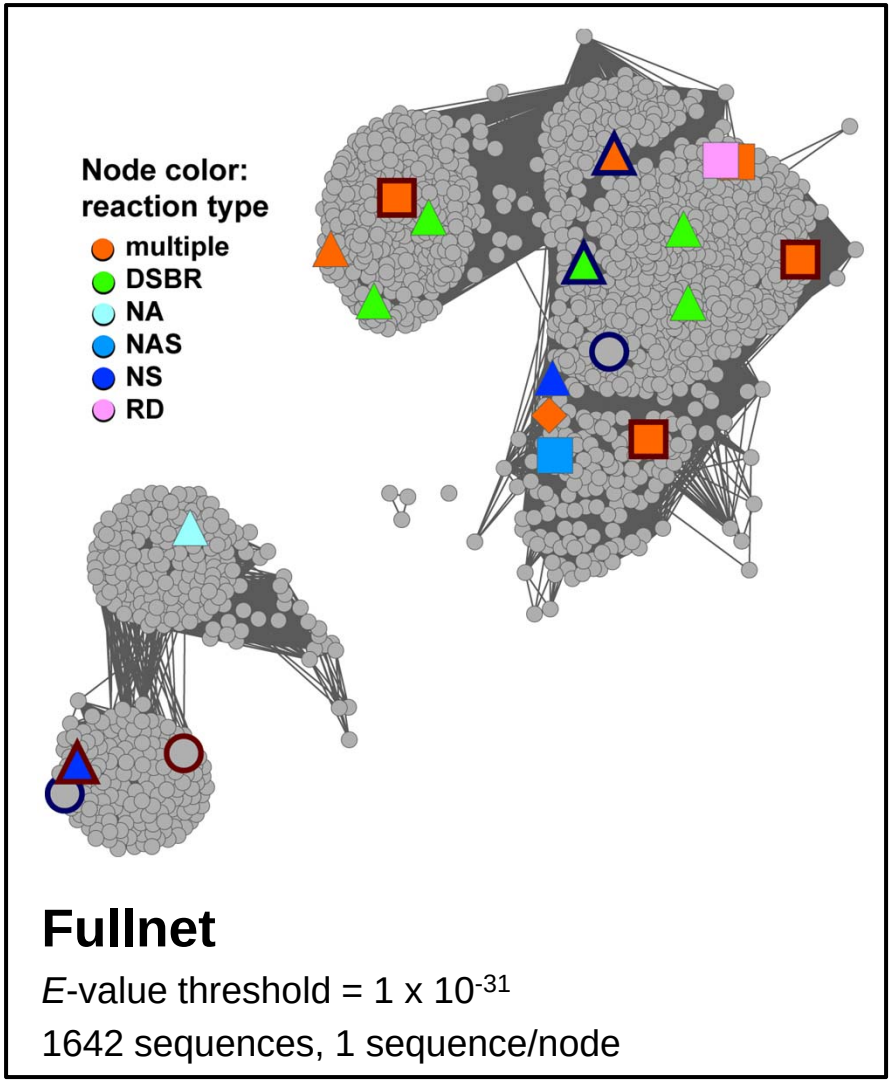
GST-Main2 (Nu-like)subgroup



Repnet

E-value threshold = 1 x 10^{-25}

139 representative nodes,
1-many sequences/node

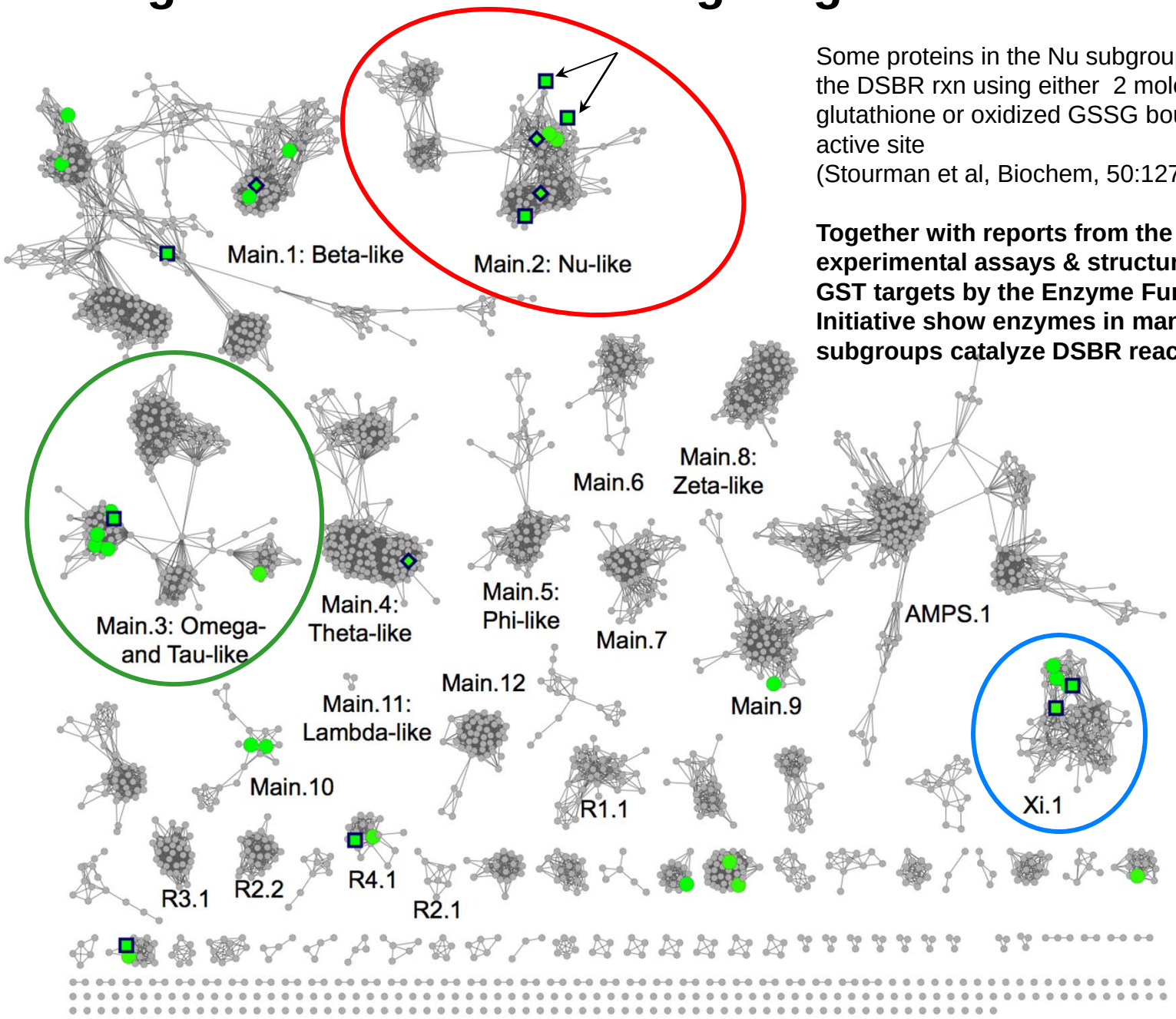


Fullnet

E-value threshold = 1 x 10^{-31}

1642 sequences, 1 sequence/node

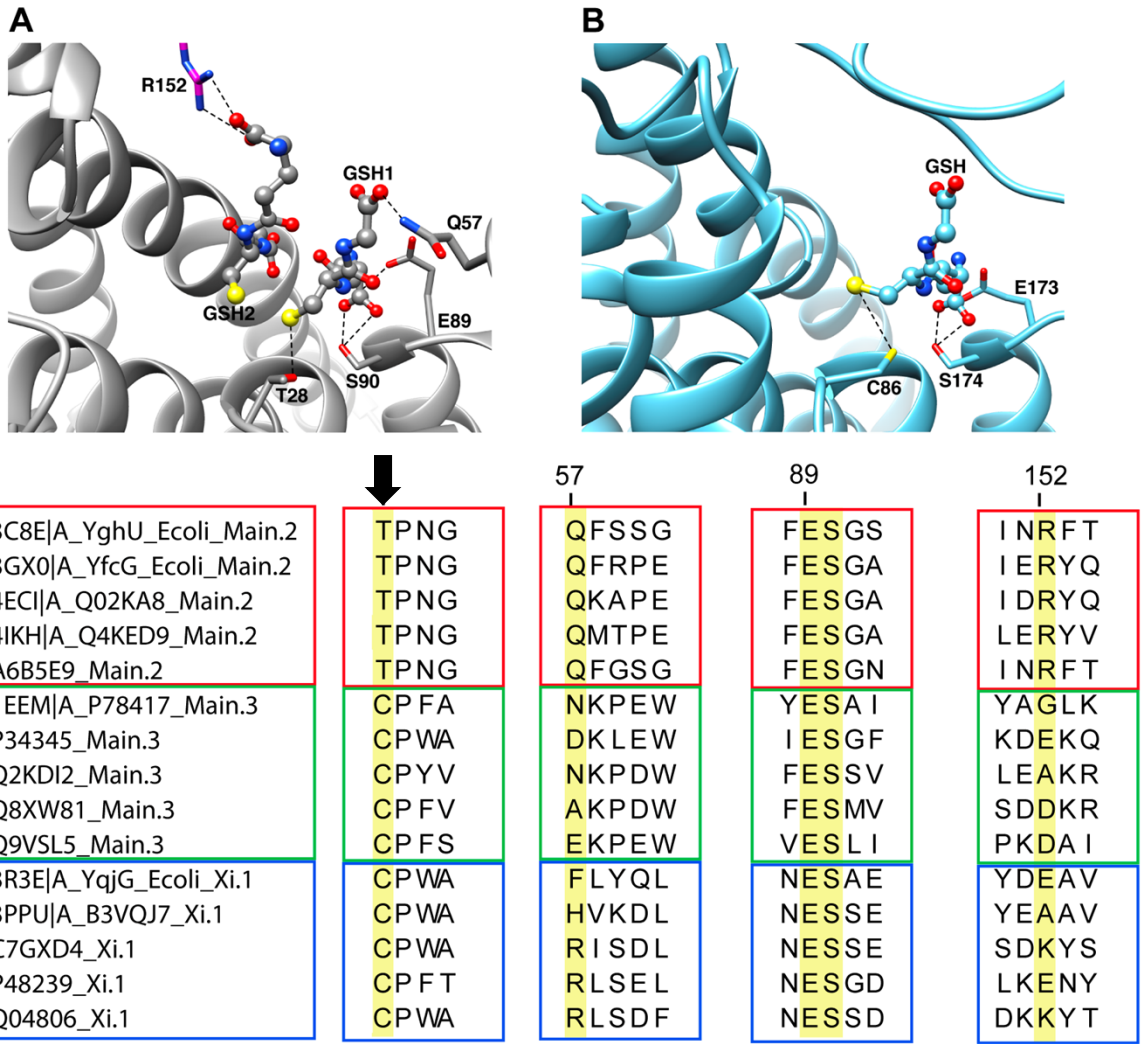
Target selection for investigating mechanism



Some proteins in the Nu subgroup catalyze the DSBR rxn using either 2 molecules of glutathione or oxidized GSSG bound in the active site
(Stourman et al, Biochem, 50:1274 (2011))

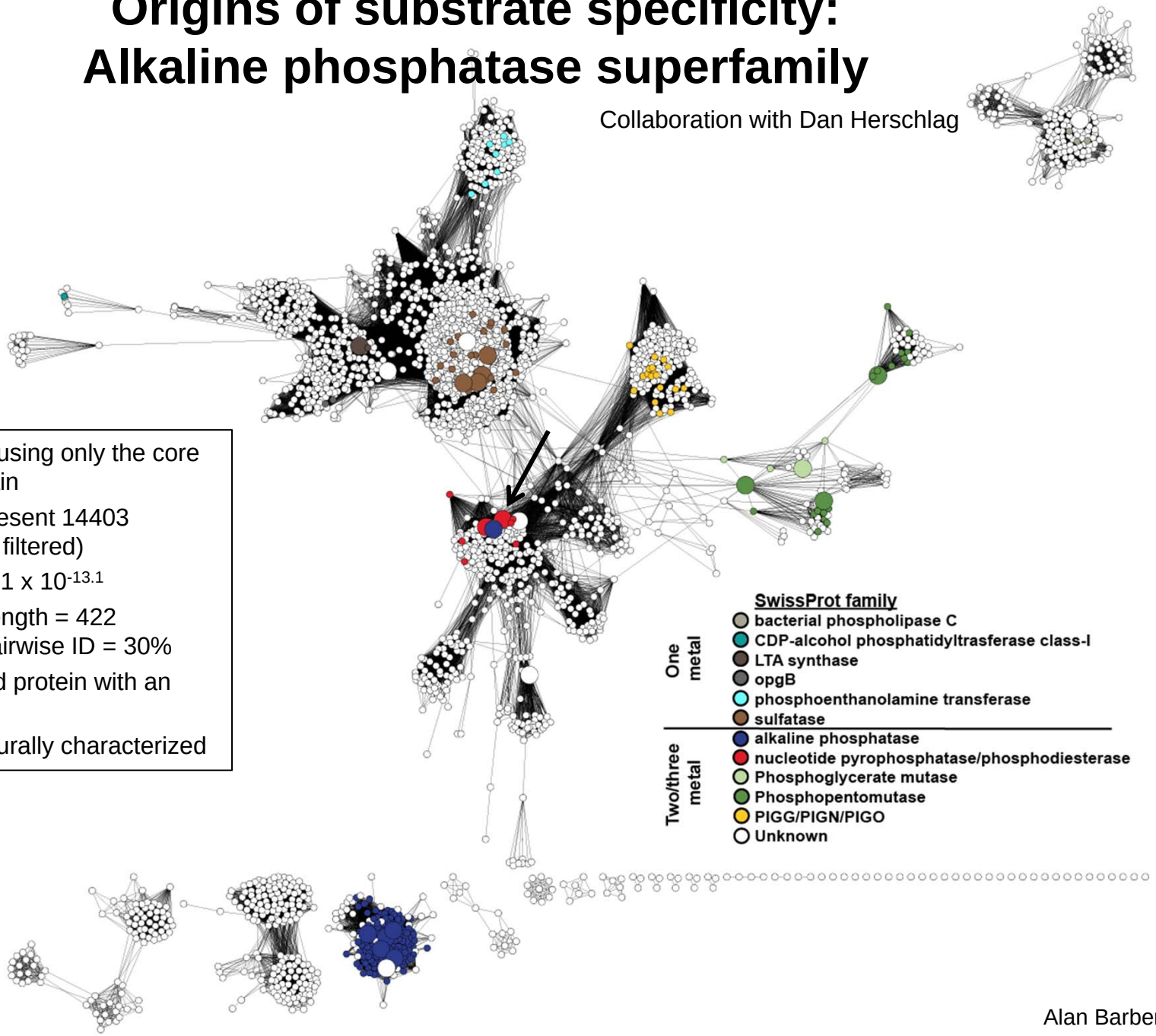
Together with reports from the literature, experimental assays & structures for new GST targets by the Enzyme Function Initiative show enzymes in many subgroups catalyze DSBR reactions

Enzymes of 3 different subgroups catalyze the DSBR reaction using different structural solutions



Origins of substrate specificity: Alkaline phosphatase superfamily

Collaboration with Dan Herschlag



Network generated using only the core Rossman fold domain

4590 repnodes represent 14403 sequences (40% ID filtered)

E-value threshold = 1 x $10^{-13.1}$

Median alignment length = 422 residues; median pairwise ID = 30%

Color: Characterized protein with an annotation

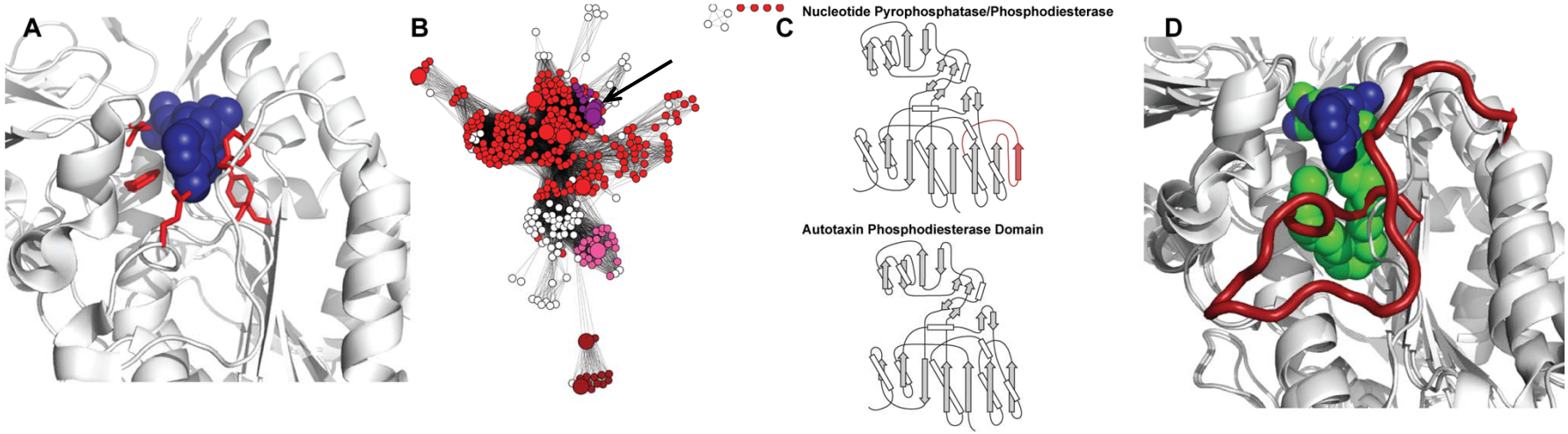
Large nodes: Structurally characterized

Alan Barber

A specialized NPP: Autotaxin

> In nucleotide pyrophosphatase/ phosphodiesterases (NPPs) recognize different ester substituents

> Autotaxin is a specialized NPP that generates lysophosphatidic acid for biological roles in cell proliferation, tumor cell motility, cytokine production



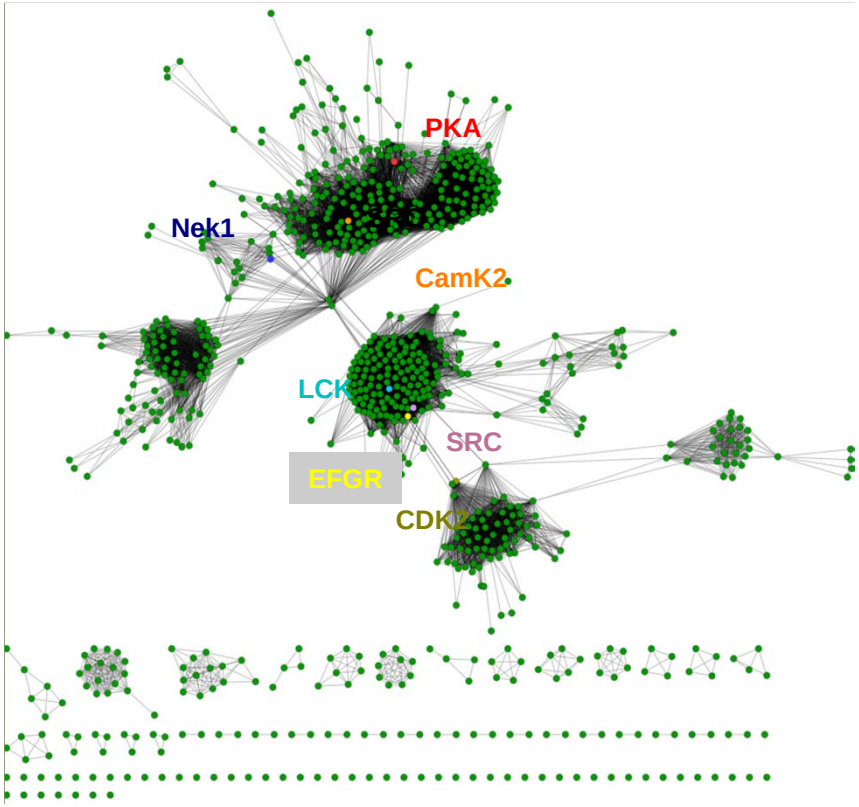
A. Side chains lining the nucleotide binding pocket (blue) of a representative NPP

B: NPP network colored by variations in sidechain residues lining the nucleotide binding pocket shown in A

C. In autotaxins, an 18-residue fragment is missing relative to other NPPs, contributing to binding phosphate diesters with lipid groups larger that those of many other NPPs

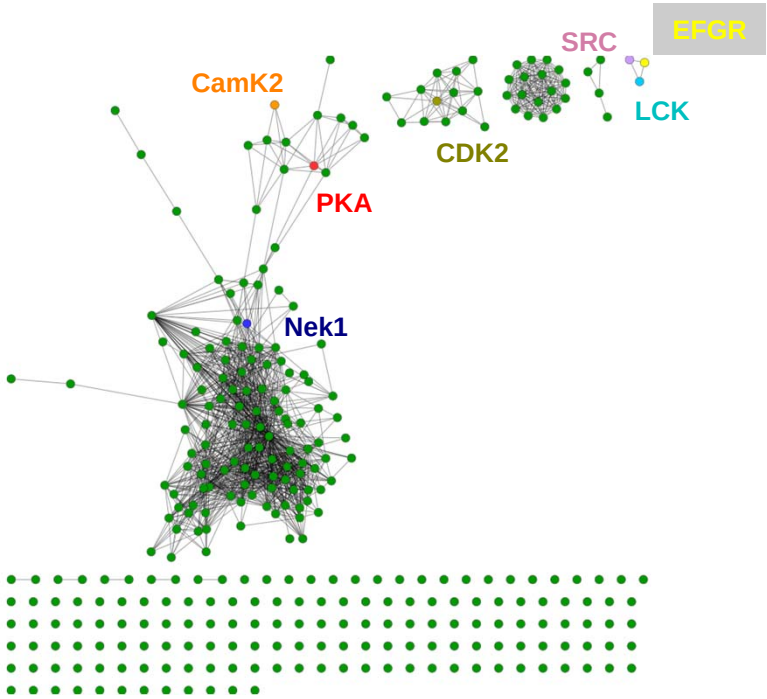
D. Superposition of autotaxin and the NPP shown in A. Red, NPP sidechain that is missing in autotaxin presumably to accommodate the large lipophosphatidic acid group (green) bound to autotaxin (3NKR)

Comparative proteomics: Human vs parasite kinomes



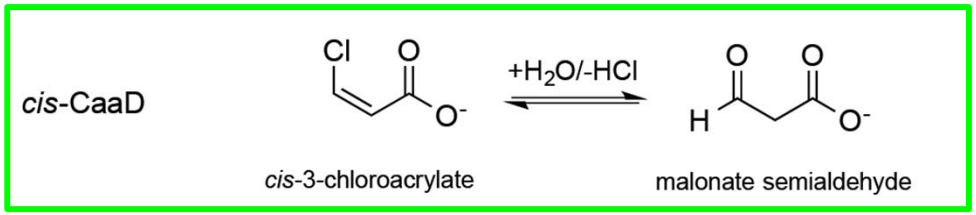
Nodes : 818 representatives of human sequences similar to 7 well-characterized human kinase domains (labeled)

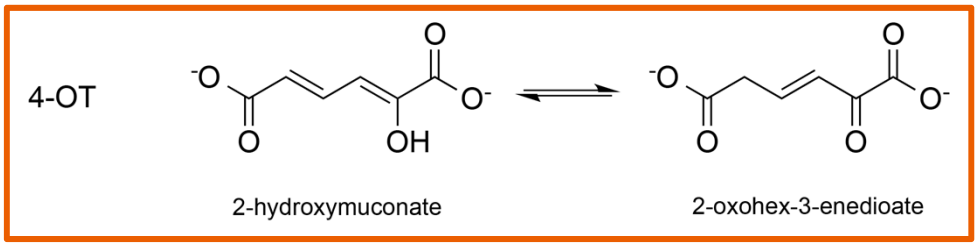
Nodes : 325 genus Giardia sequences similar to these 7 well-characterized human kinase domains

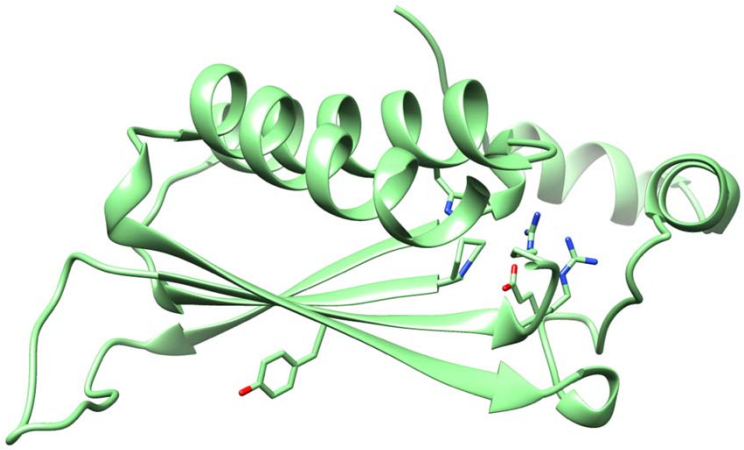


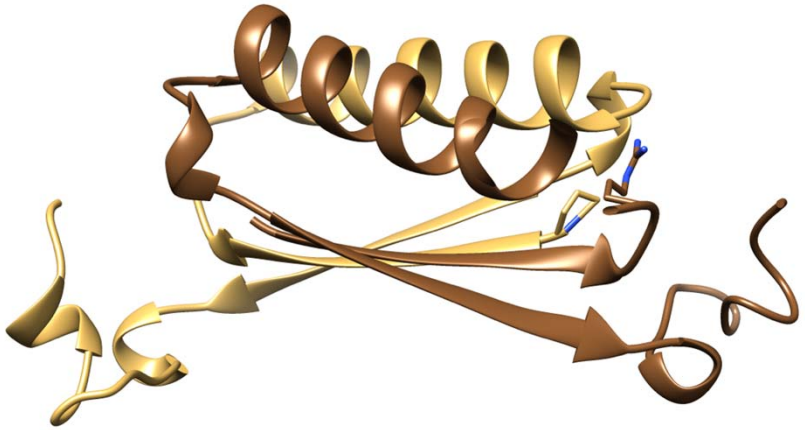
4-Oxalocrotonate tautomerase superfamily: What makes cisCaaDs different from 4-OTs?

Collaboration with Chris Whitman





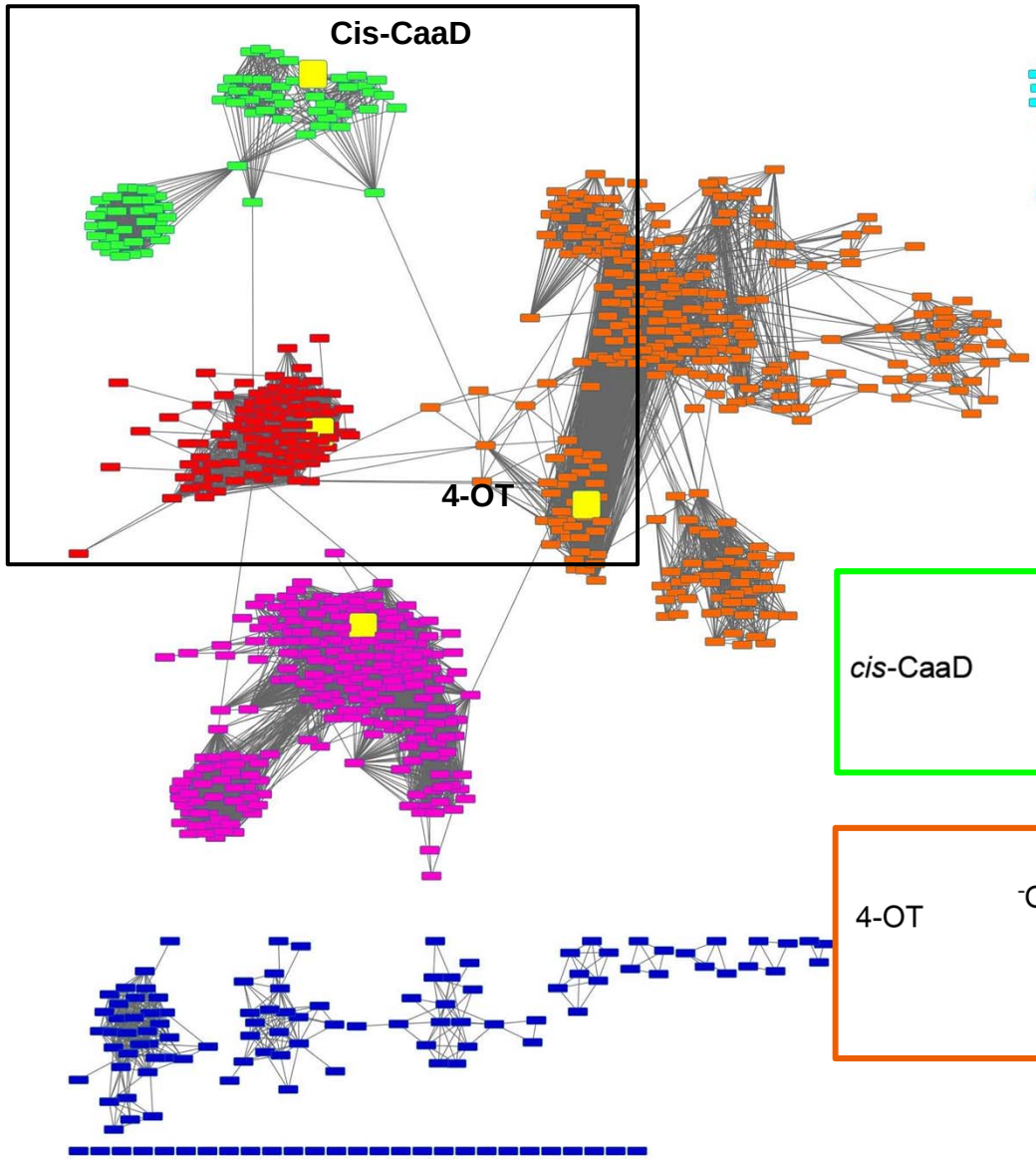




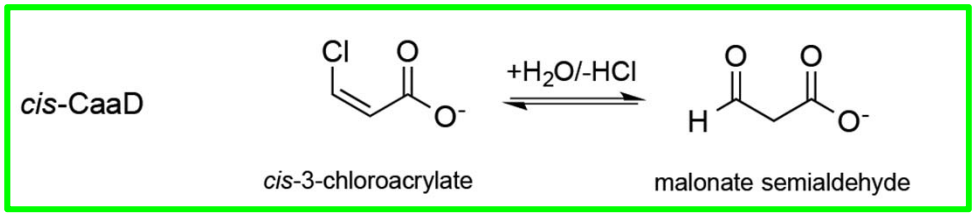
Rebecca Davidson & Bert-Jan Baas

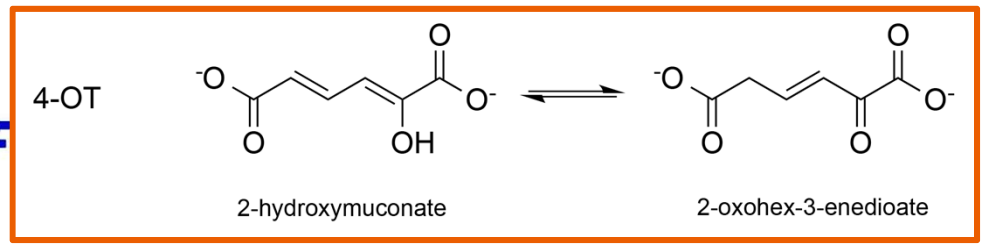
Major subgroups show differences in reaction type

Collaboration with Chris Whitman

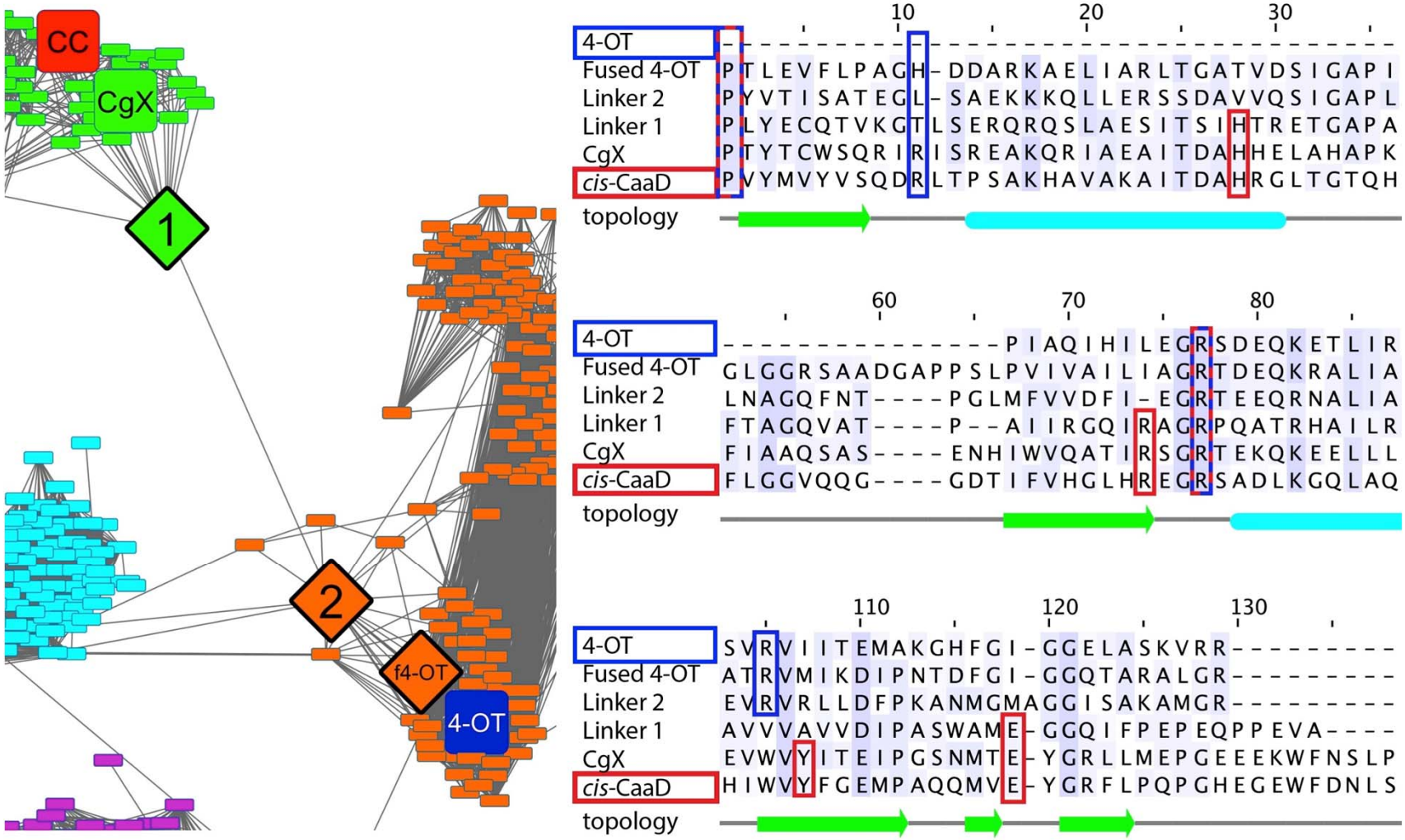


E-value threshold = 1e-10
Red: has PDB
Black border: has Swiss-Prot annotation
968 repnodes represent 7,187 sequences
Yellow: namesake enzyme





“Linkers” provide clues about structural features leading to functional variation

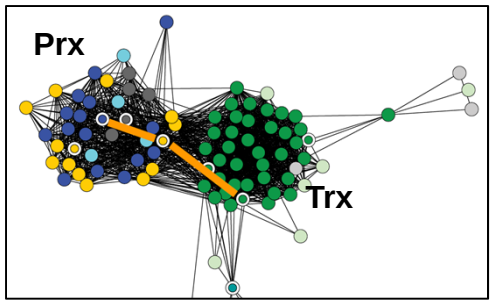


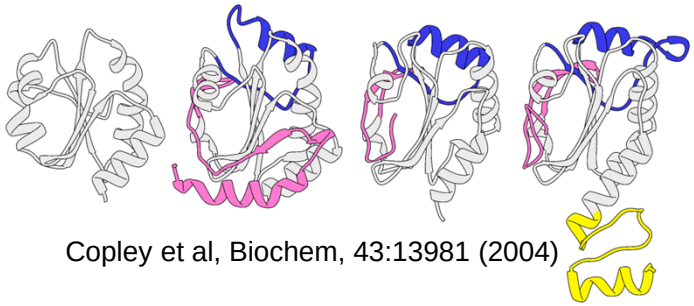
Preliminary kinetic analyses in Whitman lab show loss of 4-OT activity along the linker path

4-OT > fused 4-OT > Linker 2 > Linker 1, no activity with *cis*-CaaD

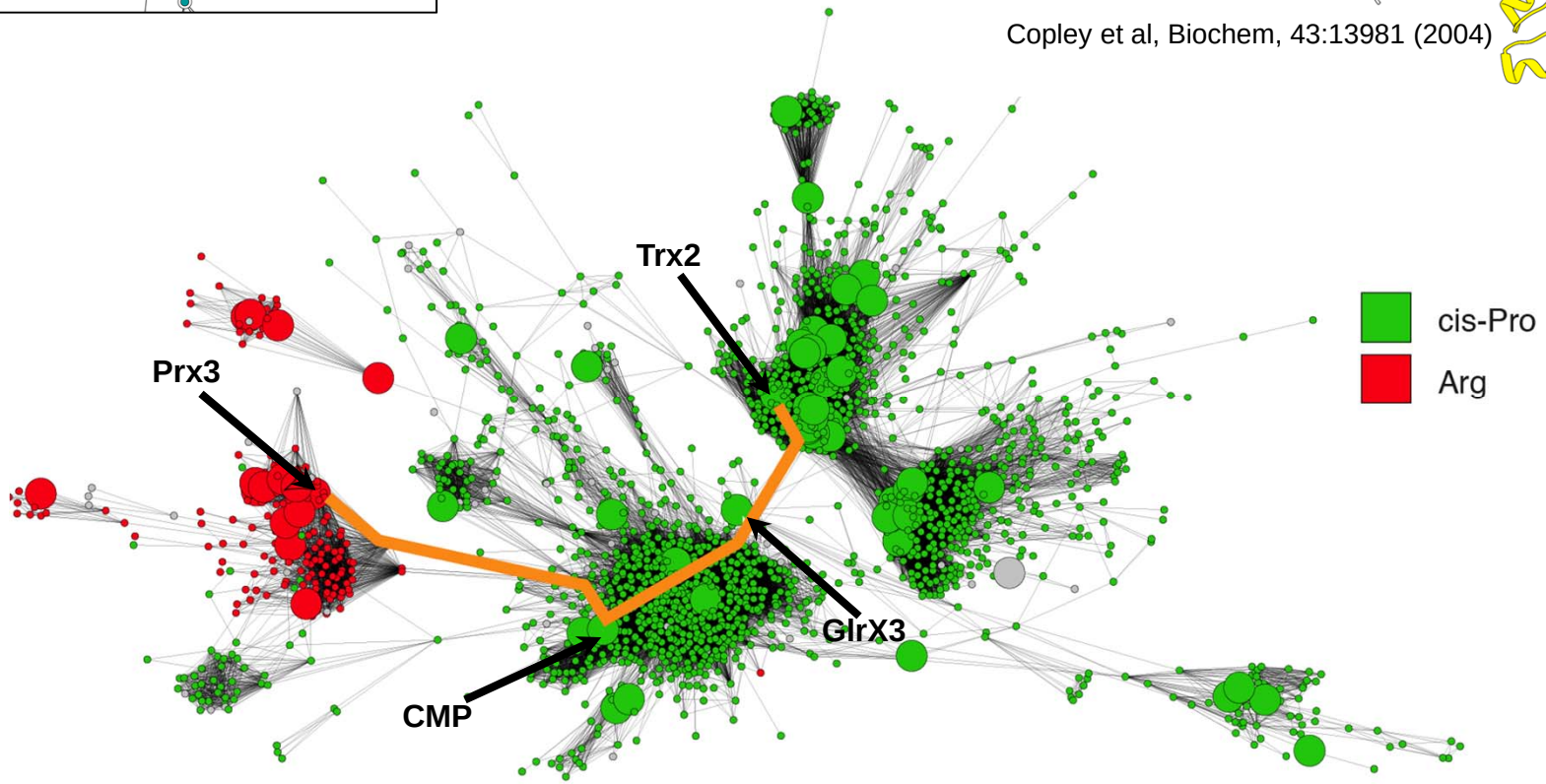
Linkers: Evolution of peroxiredoxins from a thioredoxin-like ancestor?

(Exercise #3: Linkers)





Copley et al, Biochem, 43:13981 (2004)

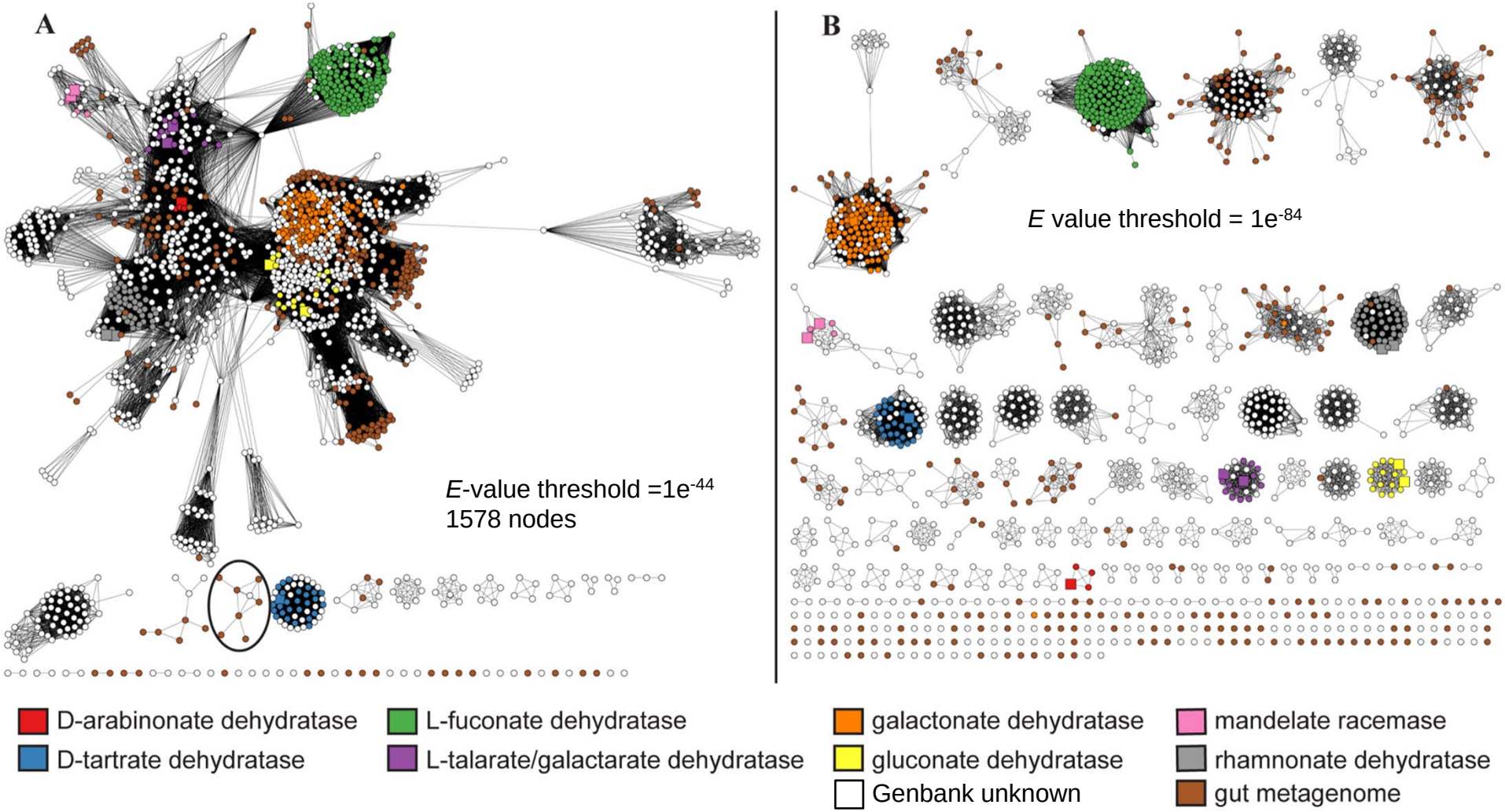


Atkinson & Babbitt, PLoS Comp Biol, 5: e1000541 (2009)

Expanding the context: Metagenomic data

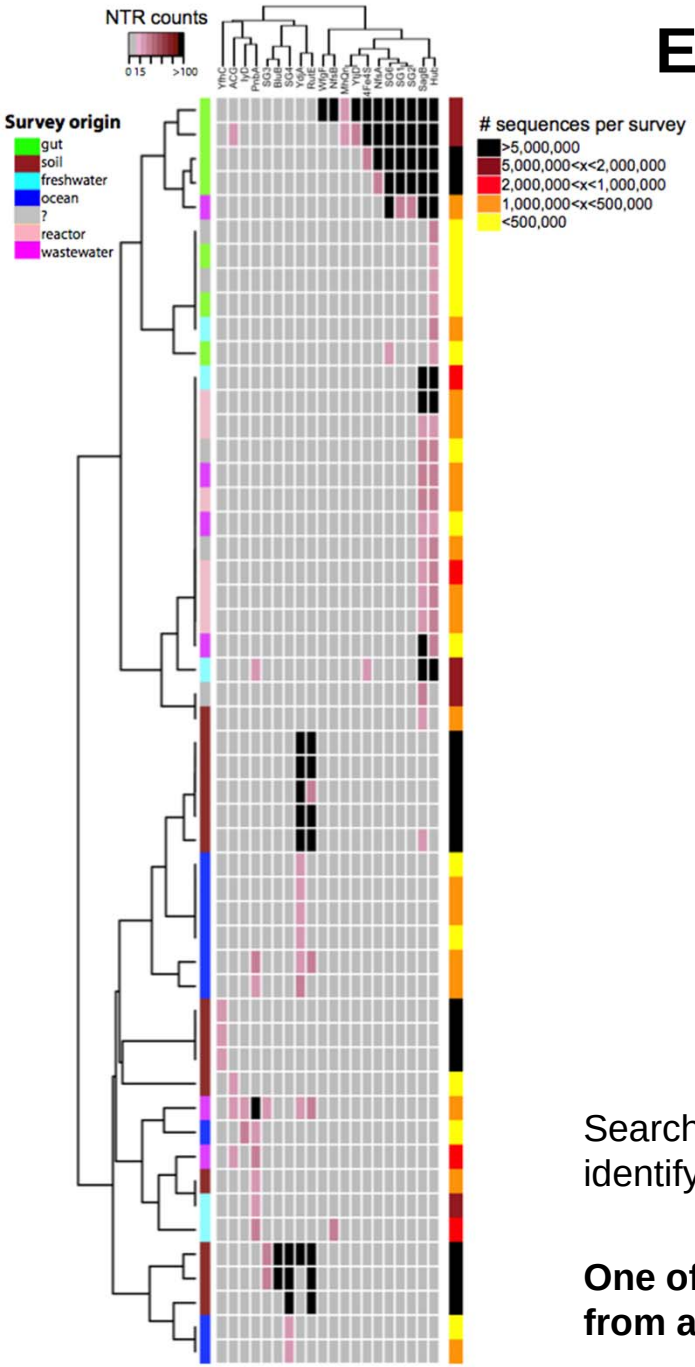
Identification of “new” carbon sources in the gut microbiome?

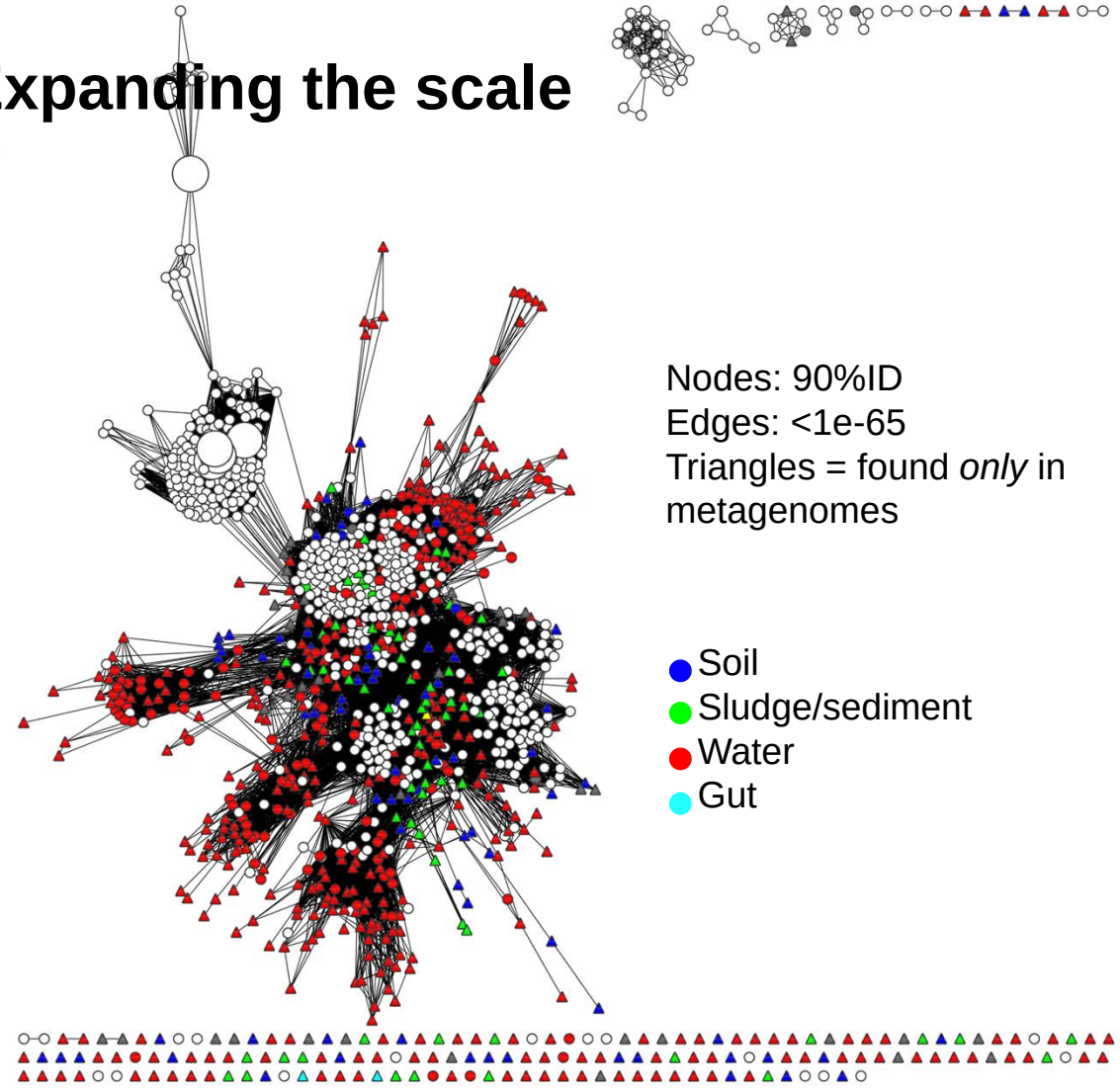
(Exercise #2: Metagenomic data)



Brown & Babbitt, JBC minireview, 287:35 (2011)

Expanding the scale





Nodes: 90%ID
Edges: <1e-65
Triangles = found *only* in metagenomes

- Soil
- Sludge/sediment
- Water
- Gut

Searched1806 metagenomics with subgroup HMMs using MG-RAST to identify surveys enriched in NTR SF subgroups

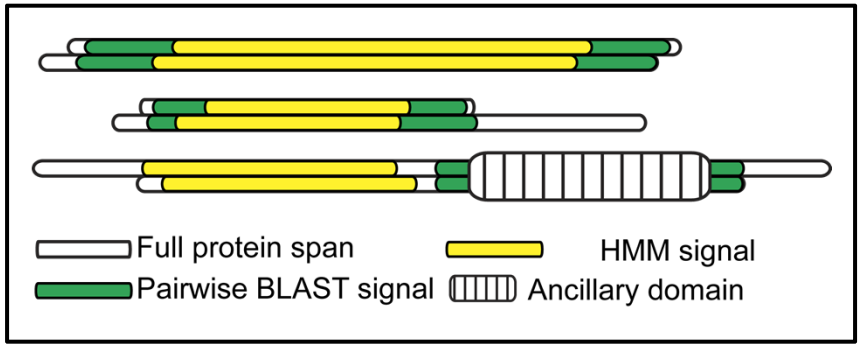
(Keegan et al, Meth Mol Biol,1399:207-33 (2016)

One of many other approaches: Addition of sequences predicted from ancestral reconstruction

Eyal Akiva

Challenges

- Technical: speed, size, visualization as sequence data increases
- Signals captured across N-1 pairwise comparisons may not map to the same regions across the sequence set



- How to choose/automate appropriate thresholds for visualization?
  - > Clustering algorithms (e.g. MCL) are limited for summarizing large-scale relationships, tricky to parameterize
  - > Changing thresholds changes clustering in complex ways — we are just starting to understand how to choose thresholds that are appropriate for a particular question
- Because biology is messy…
  - > Interpretation is sensitive to many factors: repnets vs fullnets, complex domain structures in many superfamiliessuperfamily-specific issues
  - > Predicting functional boundaries from sequence/structure similarity is often wrong, due to uneven rates of evolution in famlies in a superfamily, many other issues

Don’t let the “prettyness” of PSNs fool you!

Visualization of relationships can vary depending on layout, RepNet vs FullNet, thresholds, how different servers generate PSNs.

PSNs calculate pairwise relationships in N-1 dimensions, while visualization software displays these relationships only in (see Atkinson et al, PLoS ONE, 4: e4345 (2009)

Length & degree of similarity within & between subgroupings/families can vary enormously across a diverse set of superfamily sequences

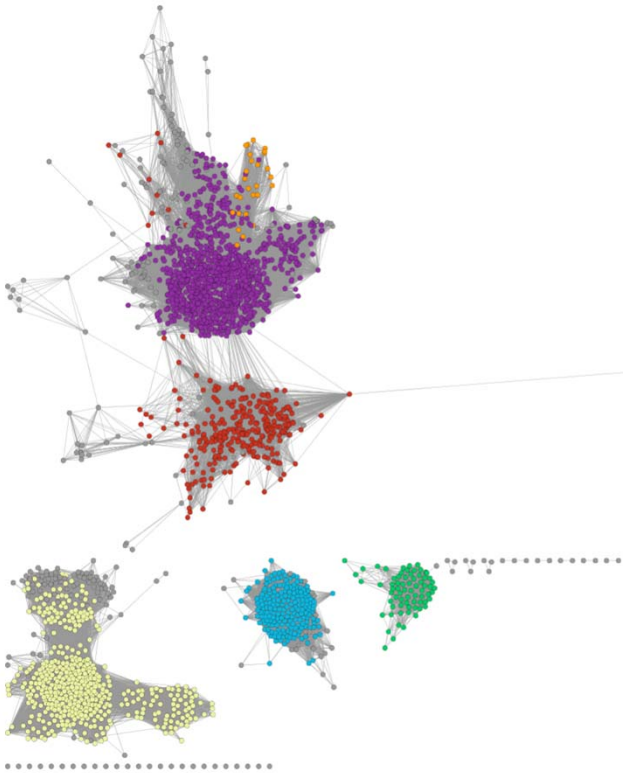
- > Can result in large differeneces in the strength of the similarity signal captured across a single network
- > These complex and heterogenous types of relationships across a visually attractive network are easily missed

Blast is fast & compares reasonably to more sophisticated metrics, but only provides rough guidance

- > Structure or HMM-based networks add robustness but only poorly sample sequence space, take much longer to generate

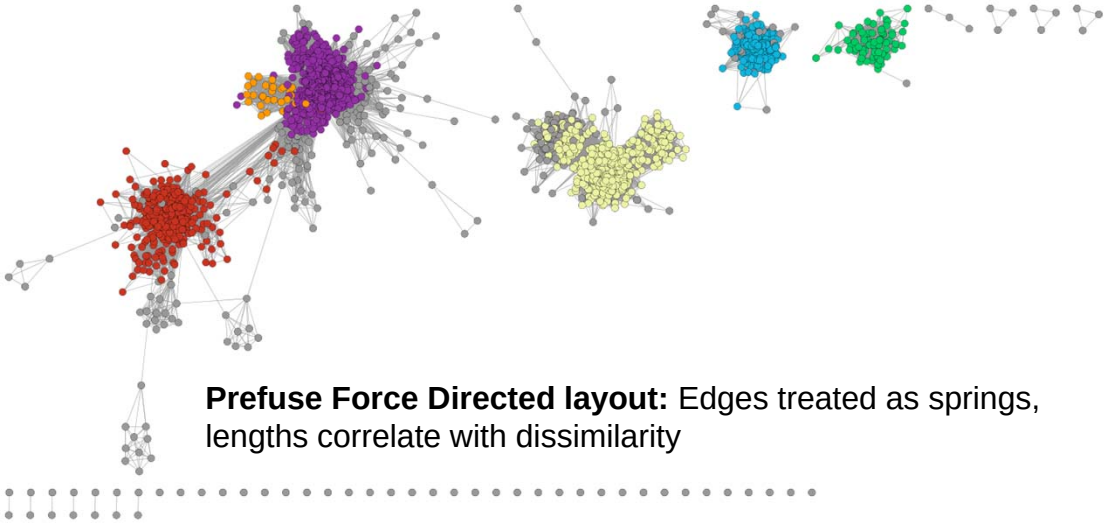
Choosing an appropriate sequence set to compare for network generation is key to answering the questions you want to address

Layouts



Organic: Edge lengths represent degree of connectivity, track with dissimilarity

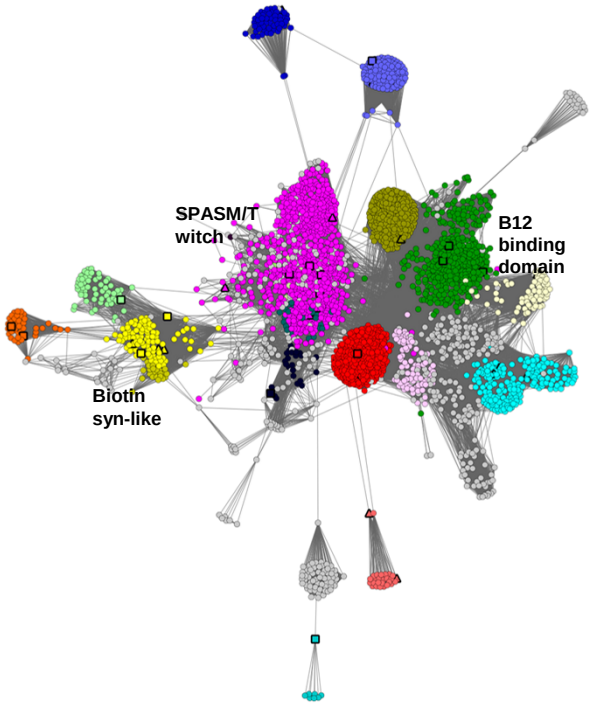
- Many other layouts available from Cytoscape
- Differences illustrate challenges for guiding experimental design from using visual reasoning alone

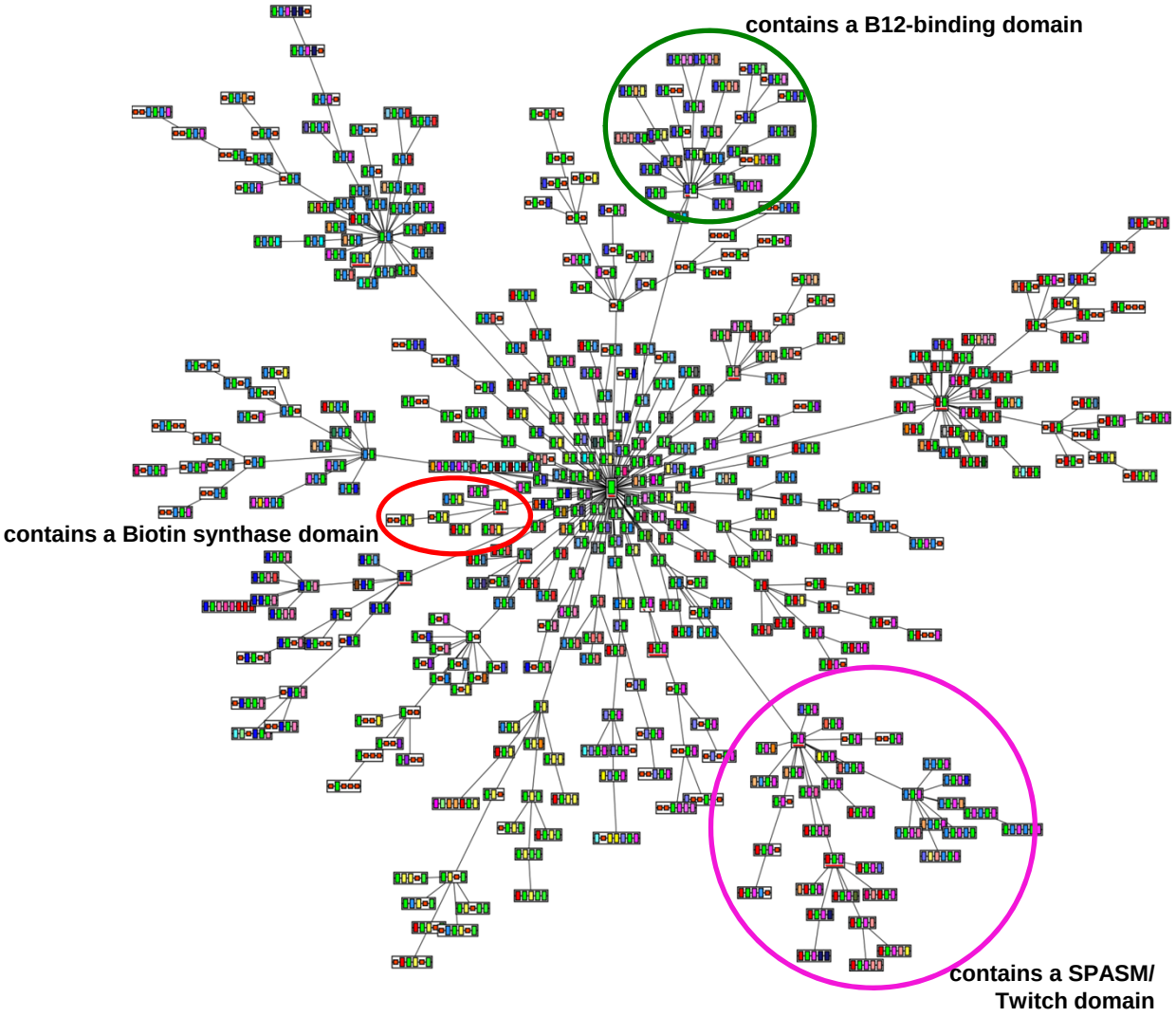


Prefuse Force Directed layout: Edges treated as springs, lengths correlate with dissimilarity

Radical SAM superfamily: Many & varied domains

>155,000 sequences





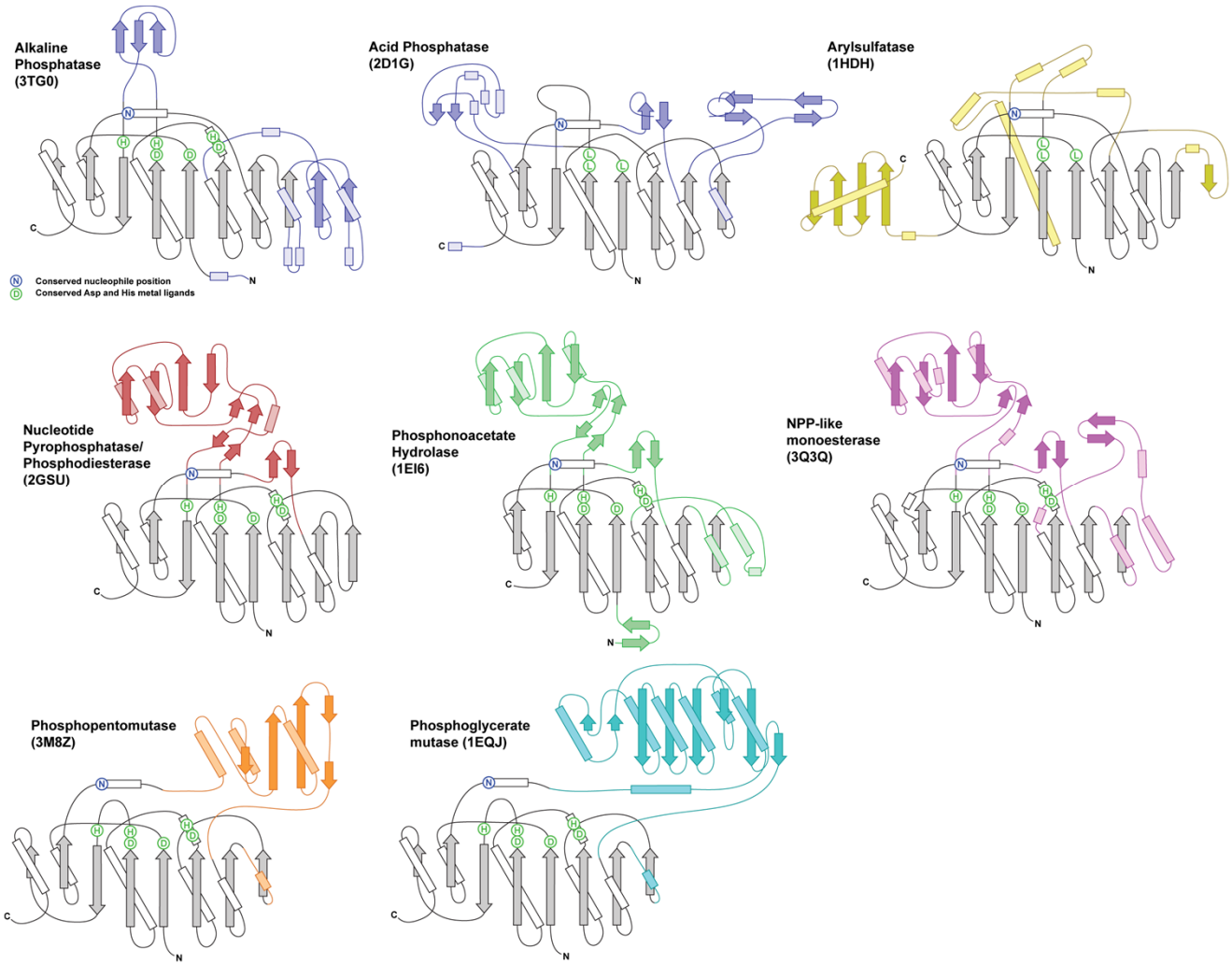
Very few of these enzymes are structurally or functionally characterized

Gemma Holliday PhD, in collaboration with Squire Booker, Andre Sali, Matt Jacobson, Steve Almo

Alkaline phosphatase superfamily: Inserts to the common core distinguish known reaction classes

>150,000 sequences (Pfam)

*Collaboration with Dan Herschlag



Many of these inserts are unrelated to each other

> Insert locations vary

> Multiple insertions in a single subgroup

> Our understanding is limited about how these structural insert patterns contribute to variations in reaction specificity or mechanism

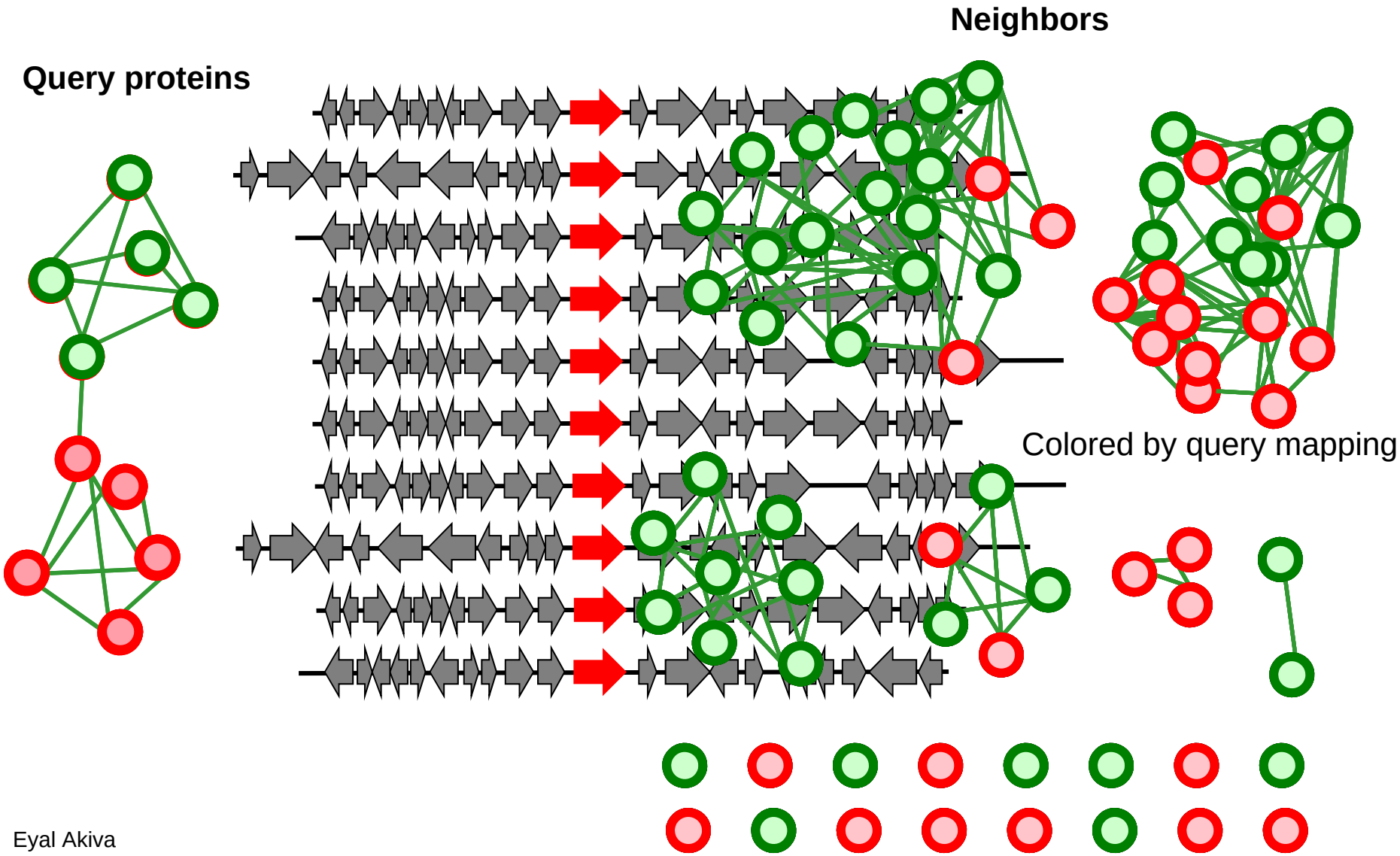
Jonathan Lassila

While providing valuable intuition, PSNs are best used as a starting point, not a final answer!

- Different superfamilies evolved in unique & complex ways to enable functional diversity, preventing simple generalization, e.g., use of general thresholds for predicting functional boundaries
- PSNs as a guide to experiment should be used with caution — use orthogonal & more rigorous approaches to evaluate PSN-derived hypotheses
- For many reasons, accurate clustering of homologous sequences & structures may not track with functional boundaries, complicating functional prediction
- Even for well-studied superfamilies, only a tiny proportion of proteins have been experimentally characterized making it hard to predict functions of unknowns by analogy to knowns

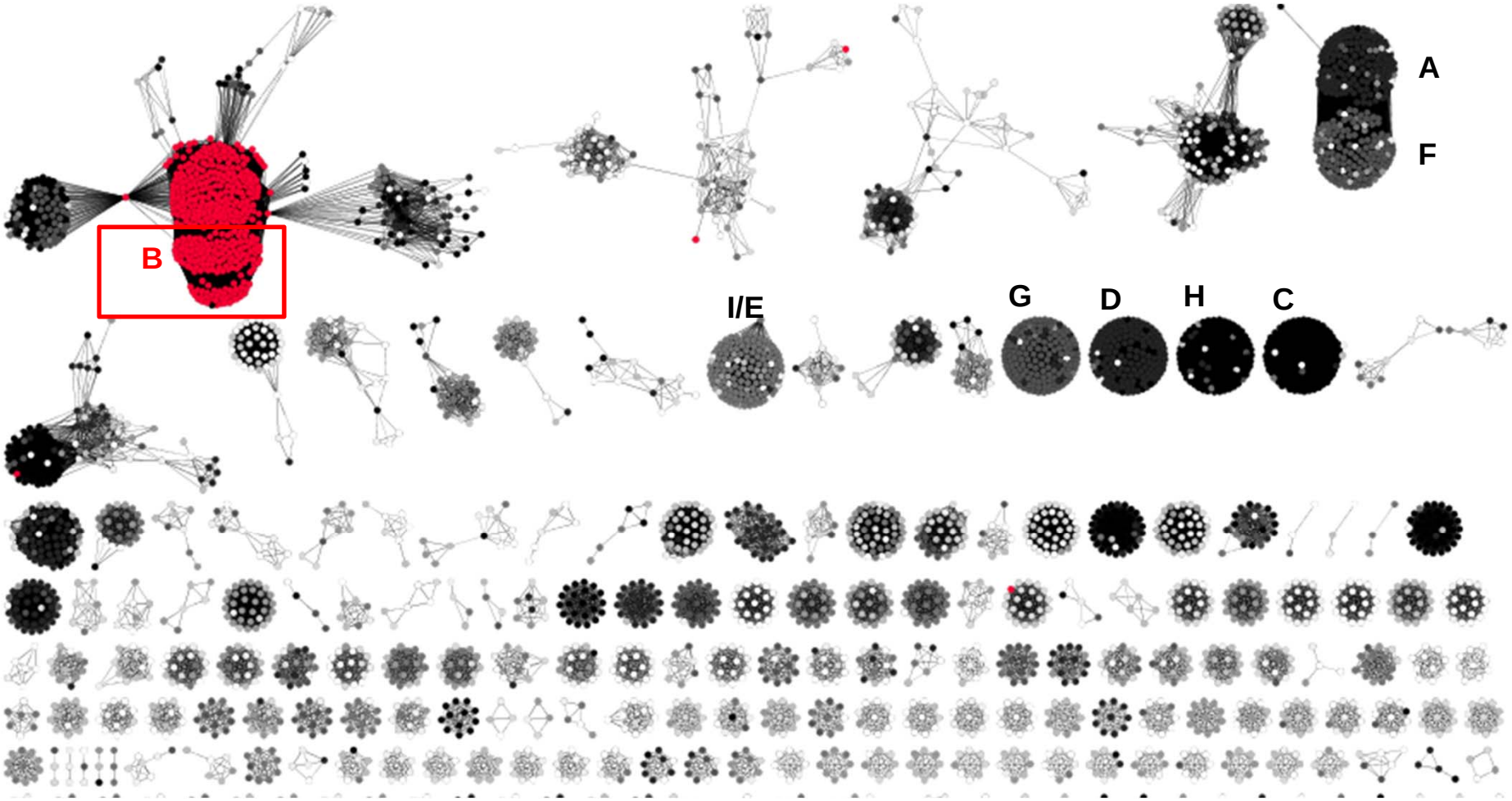
Genome context networks

(Exercise #4: Target selection)



Example GCN

E-value cutoff 1e-14
Organic Layout
Seed sequences colored red
Others colored by # hops from seed sequence (closer = darker)



Shoshana Brown

Acknowledgments

Eyal Akiva, PhD
Holly Atkinson, PhD*
Alan Barber, PhD*
Shoshana Brown, PhD
Gemma Holliday, PhD
Michael Hicks, PhD*
Florian Lauck*
Susan Mashiyama, PhD*
Elaine Meng, PhD
David Mischel
Rebecca Davidson
Alexandra Schnoes, PhD*
Jeff Yunes

Collaborators

USCF Resource for Biocomputing, Visualization & Informatics
Tom Ferrin
John Morris
..

Steve Almo
Richard Armstrong
Squire Booker
John Gerlt
Dan Herschlag
Matt Jacobson
Sarah O'Connor
Andrej Sali

*Former members

$$ NIH, NSF $$