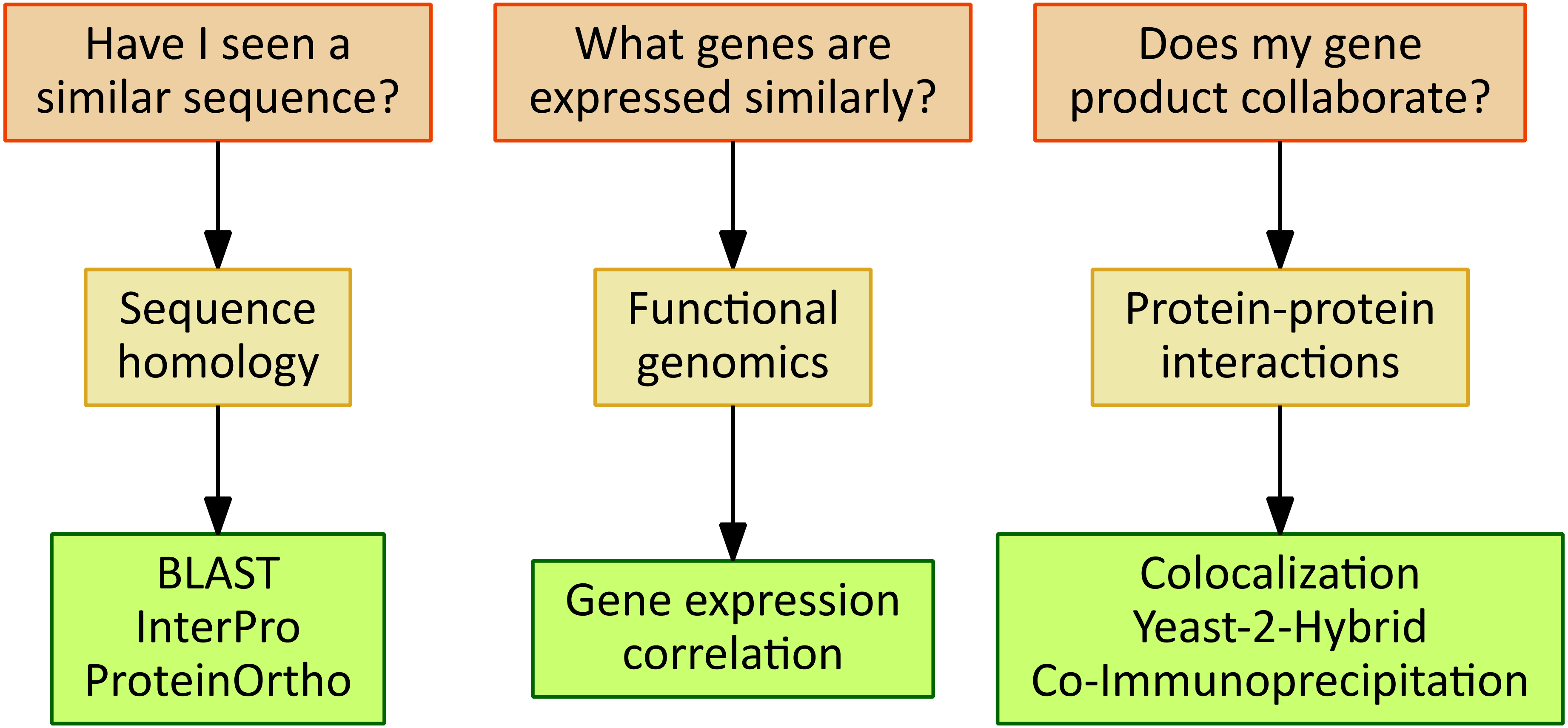
What does this gene do? Bioinformatics for function

DAVID L. TABB, PH.D.
MOLECULAR BIOLOGY AND HUMAN GENETICS
STELLENBOSCH UNIVERSITY

Outline



Have I seen a similar sequence?

Sequence homology

BLAST
InterPro
ProteinOrtho

What genes are expressed similarly?

Functional genomics

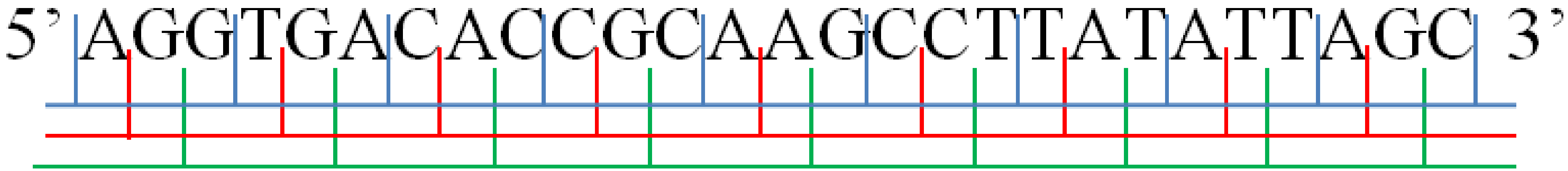
Gene expression correlation

Does my gene product collaborate?

Protein-protein interactions

Colocalization
Yeast-2-Hybrid
Co-Immunoprecipitation

If you have transcripts, generate polypeptides, too.



- EMBOSS `transeq` can 6-frame translate polypeptide sequence from assembled mRNAs.
- EMBOSS `checktrans` can eliminate any ORF below 100 amino acids.
- Downstream tools are faster if you translate.

From putative transcript to a translation

```
CTGGTCGGCTATTAGAAAAGAAAGATCGGCTAAGTCCTTCGGACCTGATCAACTTAATCC
GGGAGCTACTGATTTCAACTACTTCGACCTAACTCTCTGAAAAGATGAATTACACAAGTT
TTATCTTCGCCTTTCAGCTTTGCATAATTTTGTGTTCGTCTGGCTATTACTGTCAGTCCA
TAATTTTTAGGGAAATAGAAAACCTAAGGGACTATTTTAACGCAAGTAATCCAGATGTAG
CAGATGGTGGGTCGCTTTTCATAGATATTTTGAAGAATTGGAGAGAGGAGAGTGATAAAA
CAGTAATTCAAAGCCAAATTGTCTNNNNNNACTTGAAAATGTTTGAAAACTTGAAAGATA
ACCAGCTCATTCAAAGGAGCATCGATACCATCAAGGAAGACATGCTTGATAAGCTGTTAA
ATAGCAGCTCTGATAAACGGAATGACTTCCTCAAGCTGACTCAAATTCCTGTAAATGATC
TGCAGGTCCAGCGCAAAGCGATAAATGAACTCTTCAAAGTGATGAATGATCTCTCACCAA
GATCTAACCTCAGAAAGAGGAAAAGGAGCCAGAATCTGTTTCAAGGCCGGAGAGCATCGA
AATAATGGTCATCCTGCCTGCAATATTTGAATTTTTTATATAAATCTATTTATTAATATT
TAATATTTTACATTATTTATATGAAGAATATATTTTTAGACTCATCAATCAAAGTATTTAT
```

```
GRLLEKKDRLSPSDLINLIR
ELLISTTST*LSEKMNYTSF
IFAFQLCIILCSSGYYCQSI
IFREIENLRDYFNASNPDVA
DGGSLFIDILKNWREESDKT
VIQSQIVXXXLKMFENLKDN
QLIQRSMDTIKEDMLDKLLN
SSSDKRNDFLKLTQIPVNDL
QVQRKAINELFKVMNDLSPR
SNLRKRKRSQNLFQGRRASK
*WSSCLQYLNFLYKSIY*YL
IFYIIYMKNIFLDSSIKVF
```

N: indeterminate base call

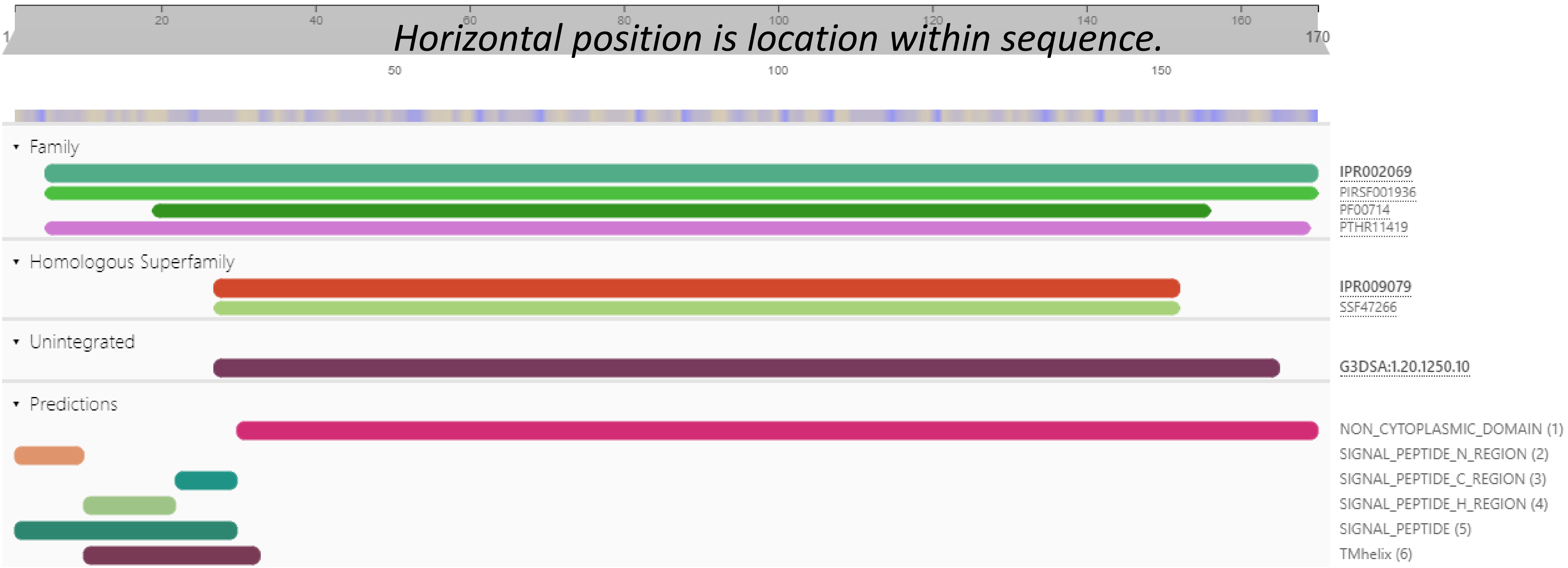
X: indeterminate amino acid

*: stop codon

This reading frame produces longest ORF.

>Hyena-TRINITY_DN38289_c0

InterPro can be run on each sequence or in batch (Linux).



Horizontal position is location within sequence.

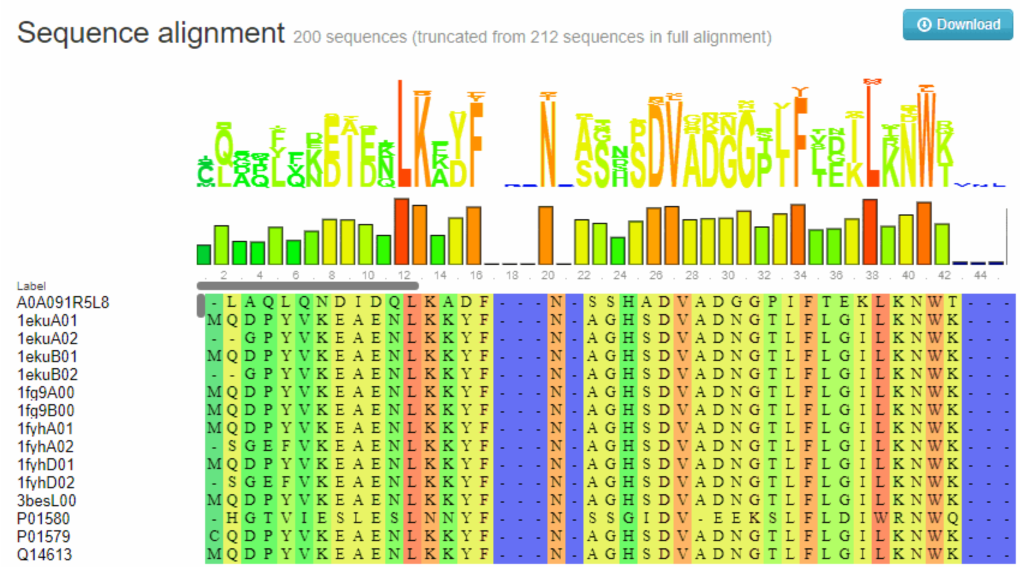
Family matches reflect hits to consensus for orthologous IFN-Gammas.

Superfamilies reflect 4-helical cytokines.

Unintegrated Gene3D reflects structure availability!

https://www.ebi.ac.uk/interpro/

Gene3D and CATH



Sequence alignment 200 sequences (truncated from 212 sequences in full alignment)

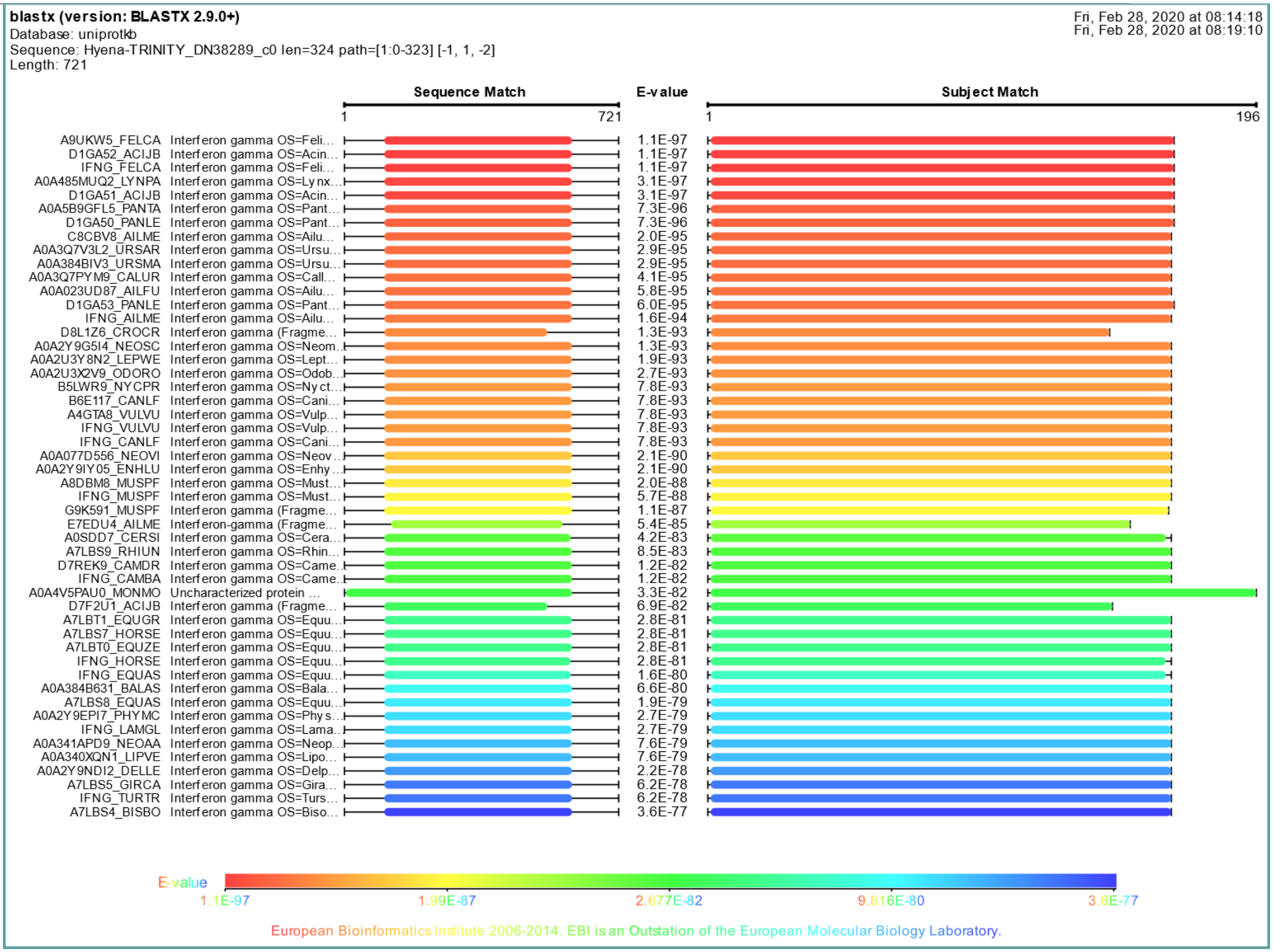
```
Label
A0A091R5L8  - L A Q L Q N D I D Q L K A D F - - - N - S S H A D V A D G G P I F T E K L K N W T - - -
1ekuA01     M Q D P Y V K E A E N L K K Y F - - - N - A G H S D V A D N G T L F L G I L K N W K - - -
1ekuA02     - - G P Y V K E A E N L K K Y F - - - N - A G H S D V A D N G T L F L G I L K N W K - - -
1ekuB01     M Q D P Y V K E A E N L K K Y F - - - N - A G H S D V A D N G T L F L G I L K N W K - - -
1ekuB02     - - G P Y V K E A E N L K K Y F - - - N - A G H S D V A D N G T L F L G I L K N W K - - -
1fg9A00     M Q D P Y V K E A E N L K K Y F - - - N - A G H S D V A D N G T L F L G I L K N W K - - -
1fg9B00     M Q D P Y V K E A E N L K K Y F - - - N - A G H S D V A D N G T L F L G I L K N W K - - -
1fyhA01     M Q D P Y V K E A E N L K K Y F - - - N - A G H S D V A D N G T L F L G I L K N W K - - -
1fyhA02     - S G E F V K E A E N L K K Y F - - - N - A G H S D V A D N G T L F L G I L K N W K - - -
1fyhD01     M Q D P Y V K E A E N L K K Y F - - - N - A G H S D V A D N G T L F L G I L K N W K - - -
1fyhD02     - S G E F V K E A E N L K K Y F - - - N - A G H S D V A D N G T L F L G I L K N W K - - -
3besL00     M Q D P Y V K E A E N L K K Y F - - - N - A G H S D V A D N G T L F L G I L K N W K - - -
P01580      - H G T V I E S L E S L N N Y F - - - N - S S G I D V - E E K S L F L D I W R N W Q - - -
P01579      C Q D P Y V K E A E N L K K Y F - - - N - A G H S D V A D N G T L F L G I L K N W K - - -
Q14613      M Q D P Y V K E A E N L K K Y F - - - N - A G H S D V A D N G T L F L G I L K N W K - - -
```

- This subsequence matches sequences with known structure.
- We can overlay those structures, and we can discern sequence motif commonalities.

http://www.cathdb.info/

BLAST can be run on each sequence or in batch.

- blastx matches mRNA to protein.
- E-values show strength of match.
- Taxonomy shown in accession suffix:
  FELCA: housecat
  ACIJB: cheetah
  LYNPA: lynx
  PANTA: tiger



blastx (version: BLASTX 2.9.0+)
Database: uniprotkb
Sequence: Hyena-TRINITY_DN38289_c0 len=324 path=[1:0-323] [-1, 1, -2]
Length: 721

Fri, Feb 28, 2020 at 08:14:18
Fri, Feb 28, 2020 at 08:19:10

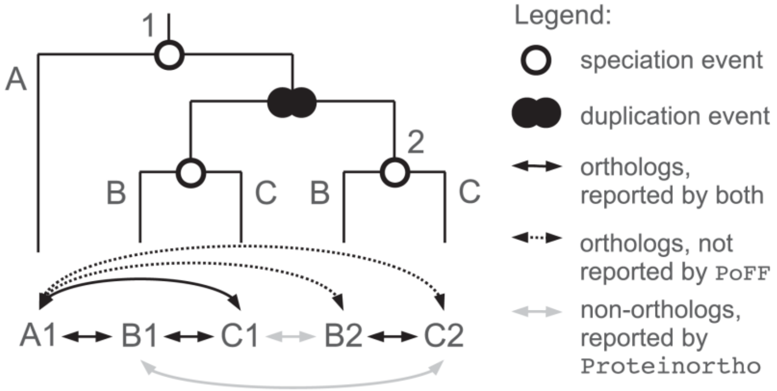
```
                                                   Sequence Match   E-value   Subject Match
                                                   1          721             1          196
   A9UKW5_FELCA  Interferon gamma OS=Feli...                        1.1E-97
   D1GA52_ACIJB  Interferon gamma OS=Acin...                        1.1E-97
     IFNG_FELCA  Interferon gamma OS=Feli...                        1.1E-97
A0A485MUQ2_LYNPA Interferon gamma OS=Lynx...                        3.1E-97
   D1GA51_ACIJB  Interferon gamma OS=Acin...                        3.1E-97
A0A5B9GFL5_PANTA Interferon gamma OS=Pant...                        7.3E-96
   D1GA50_PANLE  Interferon gamma OS=Pant...                        7.3E-96
   C8CBV8_AILME  Interferon gamma OS=Ailu...                        2.0E-95
A0A3Q7V3L2_URSAR Interferon gamma OS=Ursu...                        2.9E-95
A0A384BIV3_URSMA Interferon gamma OS=Ursu...                        2.9E-95
A0A3Q7PYM9_CALUR Interferon gamma OS=Call...                        4.1E-95
A0A023UD87_AILFU Interferon gamma OS=Ailu...                        5.8E-95
   D1GA53_PANLE  Interferon gamma OS=Pant...                        6.0E-95
     IFNG_AILME  Interferon gamma OS=Ailu...                        1.6E-94
   D8L1Z6_CROCR  Interferon gamma (Fragme...                        1.3E-93
A0A2Y9G5I4_NEOSC Interferon gamma OS=Neom...                        1.3E-93
A0A2U3Y8N2_LEPWE Interferon gamma OS=Lept...                        1.9E-93
A0A2U3X2V9_ODORO Interferon gamma OS=Odob...                        2.7E-93
   B5LWR9_NYCPR  Interferon gamma OS=Nyct...                        7.8E-93
   B6E117_CANLF  Interferon gamma OS=Cani...                        7.8E-93
   A4GTA8_VULVU  Interferon gamma OS=Vulp...                        7.8E-93
     IFNG_VULVU  Interferon gamma OS=Vulp...                        7.8E-93
     IFNG_CANLF  Interferon gamma OS=Cani...                        7.8E-93
A0A077D556_NEOVI Interferon gamma OS=Neov...                        2.1E-90
A0A2Y9IY05_ENHLU Interferon gamma OS=Enhy...                        2.1E-90
   A8DBM8_MUSPF  Interferon gamma OS=Must...                        2.0E-88
     IFNG_MUSPF  Interferon gamma OS=Must...                        5.7E-88
   G9K591_MUSPF  Interferon gamma (Fragme...                        1.1E-87
   E7EDU4_AILME  Interferon-gamma (Fragme...                        5.4E-85
   A0SDD7_CERSI  Interferon gamma OS=Cera...                        4.2E-83
   A7LBS9_RHIUN  Interferon gamma OS=Rhin...                        8.5E-83
   D7REK9_CAMDR  Interferon gamma OS=Came...                        1.2E-82
     IFNG_CAMBA  Interferon gamma OS=Came...                        1.2E-82
A0A4V5PAU0_MONMO Uncharacterized protein ...                        3.3E-82
   D7F2U1_ACIJB  Interferon gamma (Fragme...                        6.9E-82
   A7LBT1_EQUGR  Interferon gamma OS=Equu...                        2.8E-81
   A7LBS7_HORSE  Interferon gamma OS=Equu...                        2.8E-81
   A7LBT0_EQUZE  Interferon gamma OS=Equu...                        2.8E-81
     IFNG_HORSE  Interferon gamma OS=Equu...                        2.8E-81
     IFNG_EQUAS  Interferon gamma OS=Equu...                        1.6E-80
A0A384B631_BALAS Interferon gamma OS=Bala...                        6.6E-80
   A7LBS8_EQUAS  Interferon gamma OS=Equu...                        1.9E-79
A0A2Y9EPI7_PHYMC Interferon gamma OS=Phys...                        2.7E-79
     IFNG_LAMGL  Interferon gamma OS=Lama...                        2.7E-79
A0A341APD9_NEOAA Interferon gamma OS=Neop...                        7.6E-79
A0A340XQN1_LIPVE Interferon gamma OS=Lipo...                        7.6E-79
A0A2Y9NDI2_DELLE Interferon gamma OS=Delp...                        2.2E-78
   A7LBS5_GIRCA  Interferon gamma OS=Gira...                        6.2E-78
     IFNG_TURTR  Interferon gamma OS=Turs...                        6.2E-78
   A7LBS4_BISBO  Interferon gamma OS=Biso...                        3.6E-77
```

European Bioinformatics Institute 2006-2014. EBI is an Outstation of the European Molecular Biology Laboratory.

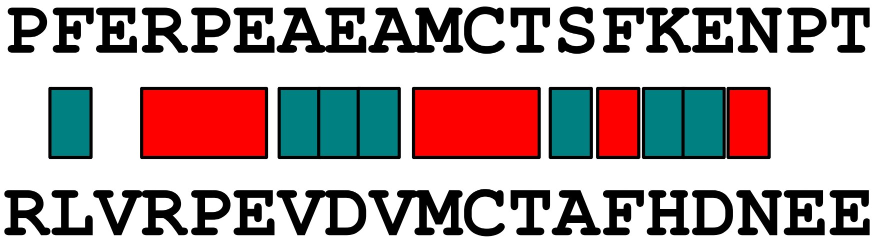
https://www.ebi.ac.uk/Tools/sss/ncbiblast

Orthologs are the “same gene across species boundaries”

- ProteinOrtho uses speciation and gene duplication models to relate sequences.


1
A
2
B
C
B
C
A1
B1
C1
B2
C2
Legend:
speciation event
duplication event
orthologs, reported by both
orthologs, not reported by PoFF
non-orthologs, reported by Proteinortho

- Diamond indexing accelerates all protein vs. all protein homology detection.



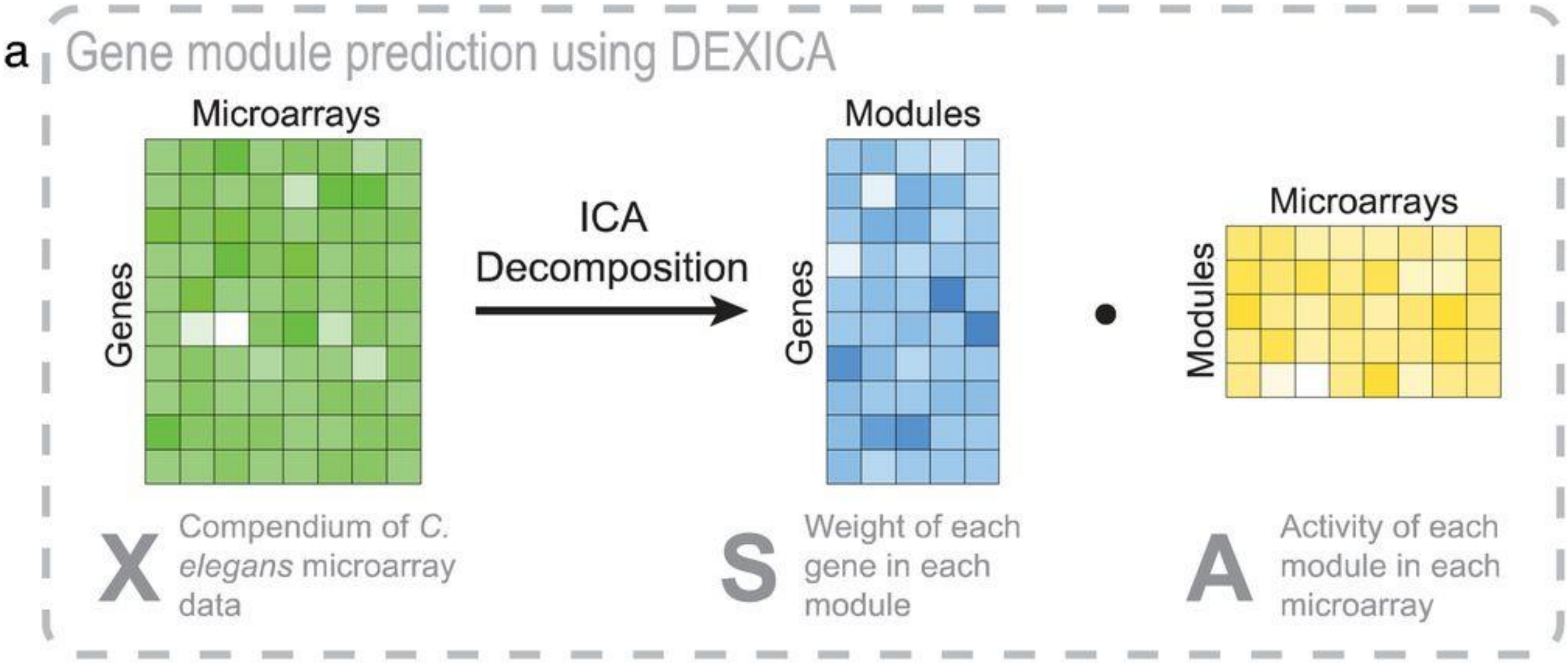
PFERPEAEAMCTSFKENPT

RLVRPEVDVMCTAFHDNEE

M. Lechner et al. *PLOS One* (2014) 9: e105015

B. Buchfink et al. *Nature Meth*. (2015) 12: 59-60

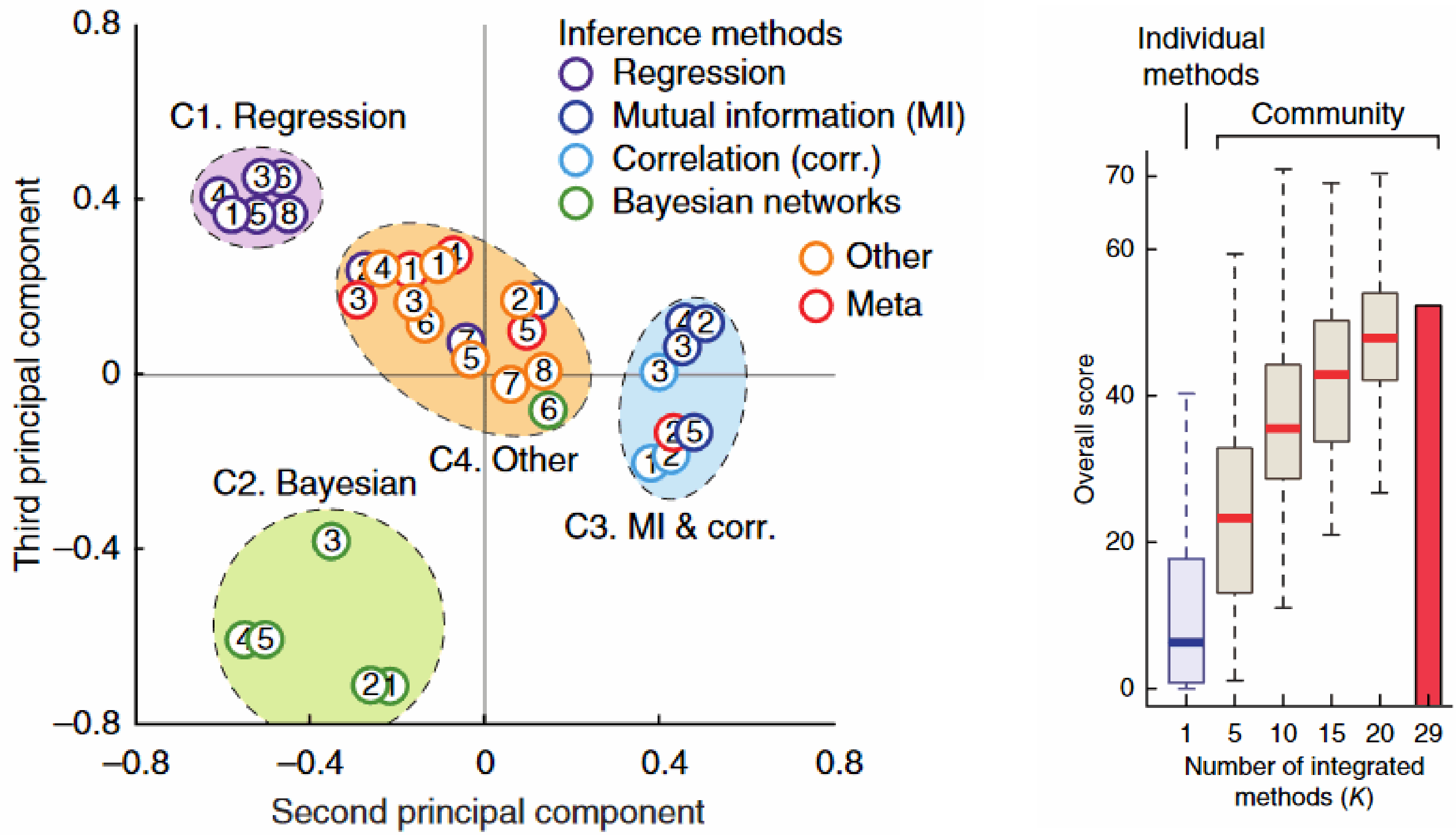
Infer function by coexpression


a
Gene module prediction using DEXICA
Microarrays
Genes
ICA
Decomposition
Modules
Genes
•
Microarrays
Modules
X
Compendium of *C. elegans* microarray data
S
Weight of each gene in each module
A
Activity of each module in each microarray

“Gene-expression data can be used to define groups of genes that show similar patterns of expression, or co-variation, across multiple conditions.”

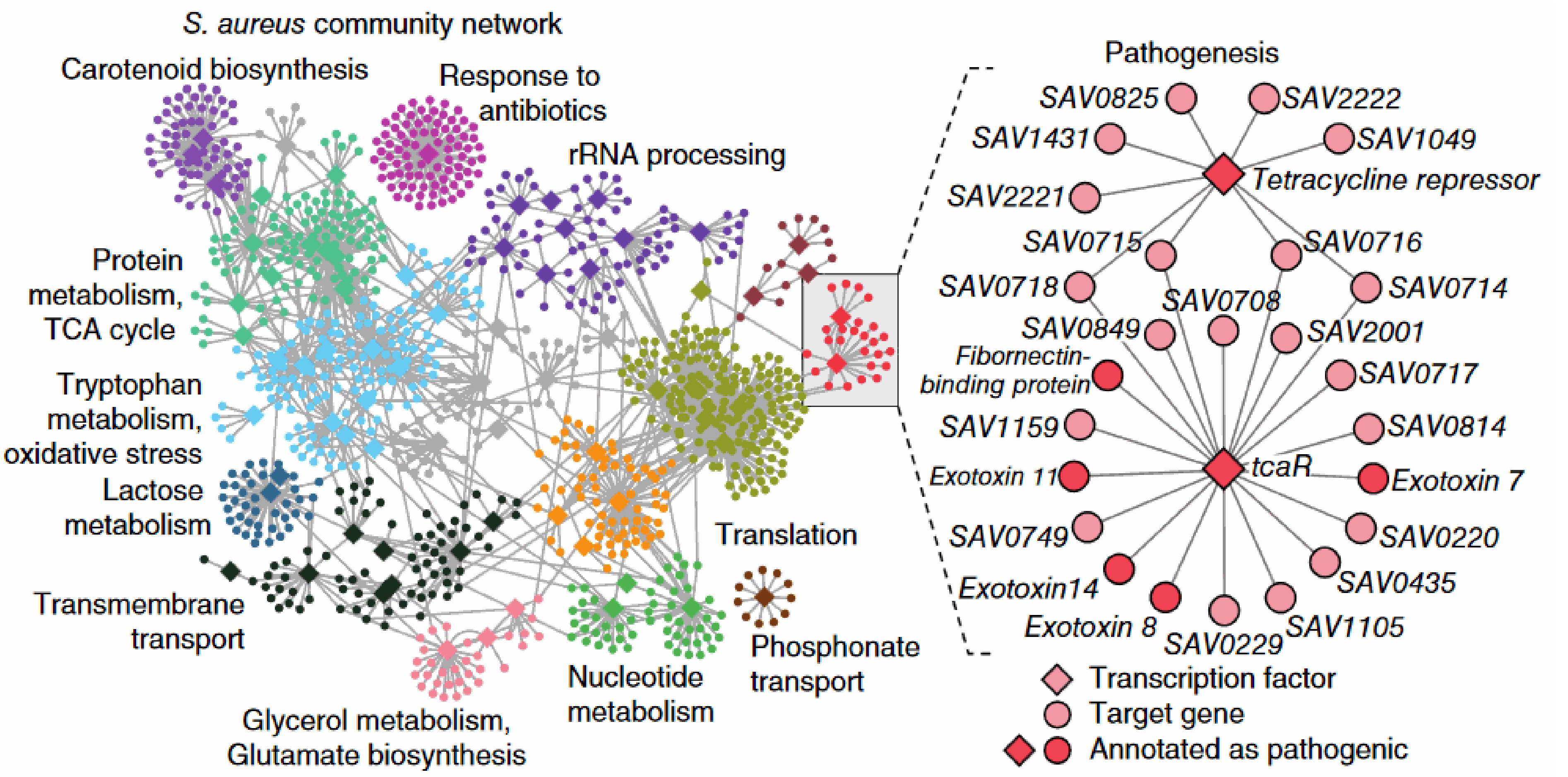
M Cary et al. BioRxiv. https://doi.org/10.1101/678482

Correlation of gene expression benefits from multiple methods



D Marbach et al. *Nature Methods* (2012) 9: 796-804.

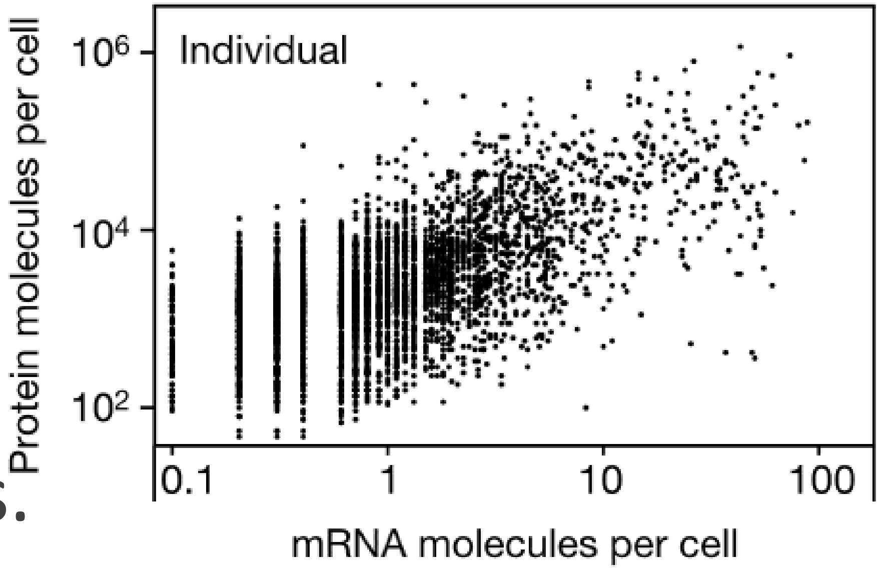
Networks of hierarchical, modular relationships



D Marbach et al. *Nature Methods* (2012) 9: 796-804.

mRNA and protein quant give different results

- Proteins in *S. cerevisiae* ranged from 50 to more than a million molecules per cell.
- Western blots quantified proteins, while microarrays quantified messenger RNA.
- Spearman rank corr. yields r=0.57. *Turnover* uses separate processes.

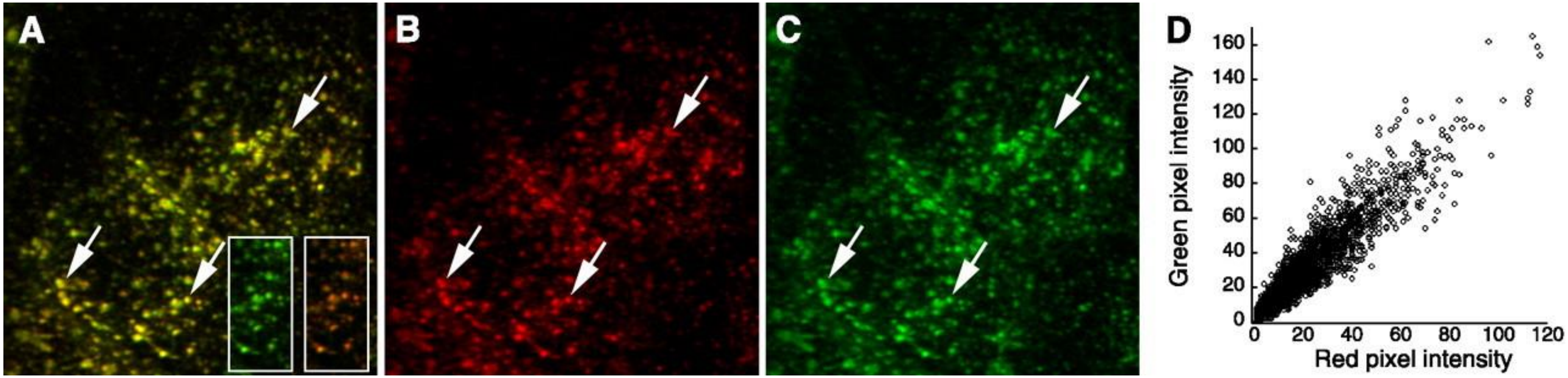


S. Ghaemmaghami et al. *Nature* (2003) 425: 737-741

Colocalization:
aspects of spatial association

- **Co-occur** overlaps in space
- **Correlate** scales similarly across structures
- **Interact** requires FRET or EM resolution

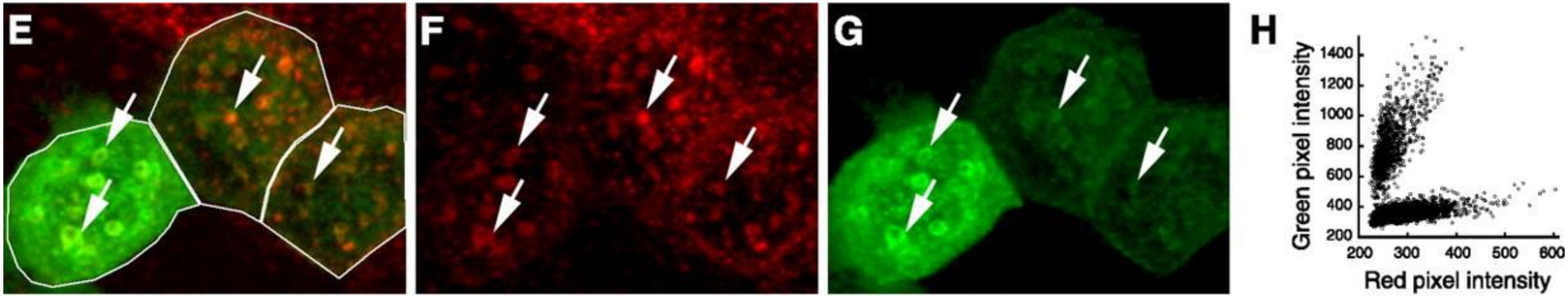
Endosomes in MDCK cells: Pearson CC= 0.944



KW Dunn et al. *Am. J. Physiol. Cell Physiol*. (2011) 300: C723-C742

Regions of Interest and single cell analysis

- Each cell may reveal different relationship.
- *Segmenting* cells from background boosts signal-to-noise.
- Specifying ROIs can separate cellular info.



KW Dunn et al. *Am. J. Physiol. Cell Physiol*. (2011) 300: C723-C742

Defining protein-protein interactions (PPIs)

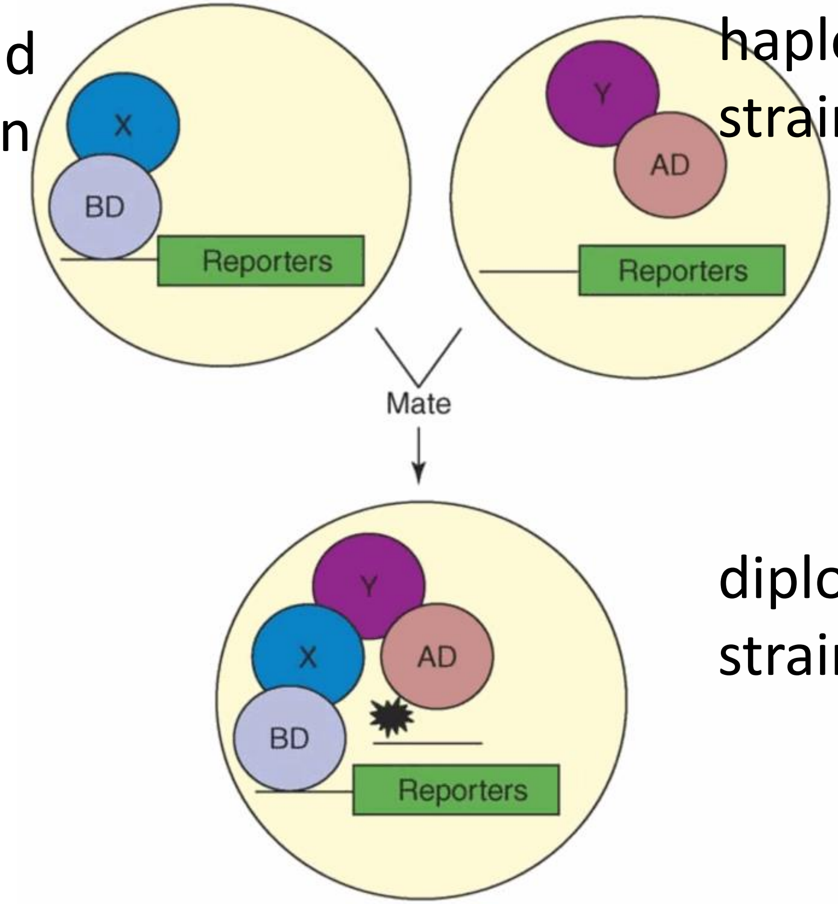
"The physical contact considered in PPIs should be specific, not just all proteins that bump into each other by chance. It also should exclude interactions that a protein experiences when it is being made, folded, quality checked, or degraded."

J de las Rivas and C Fontanillo. *PLOS Computat. Bio*. (2010) 6: e1000807

Yeast two-hybrid establishes binary relationship

- X: first gene
- Y: second gene
- BD: DNA-binding domain
- AD: Transcription activating domain

If X and Y interact, BD and AD cooperate to report.



hybrid fusion

haploid strains

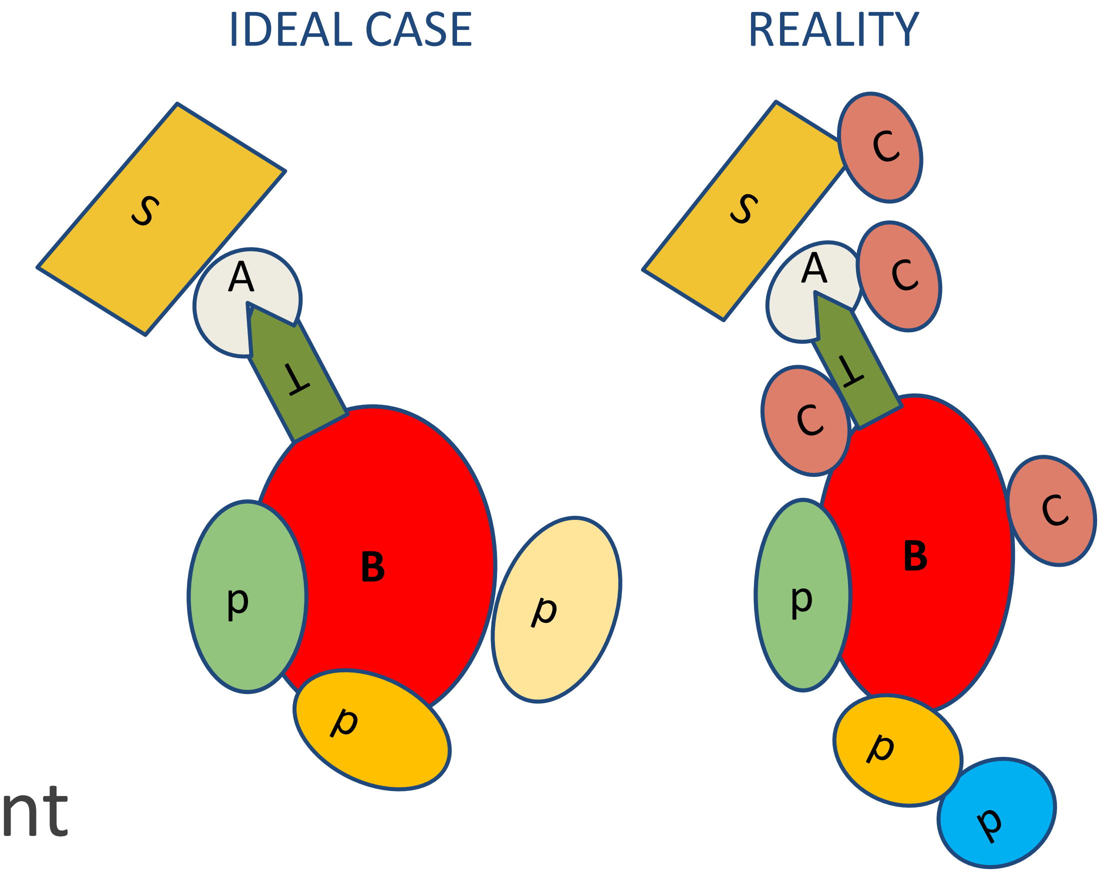
diploid strain

Current Opinion in Biotechnology

JR Parrish et al. *Curr. Opinion Biotech*. (2006) 17: 387-393.

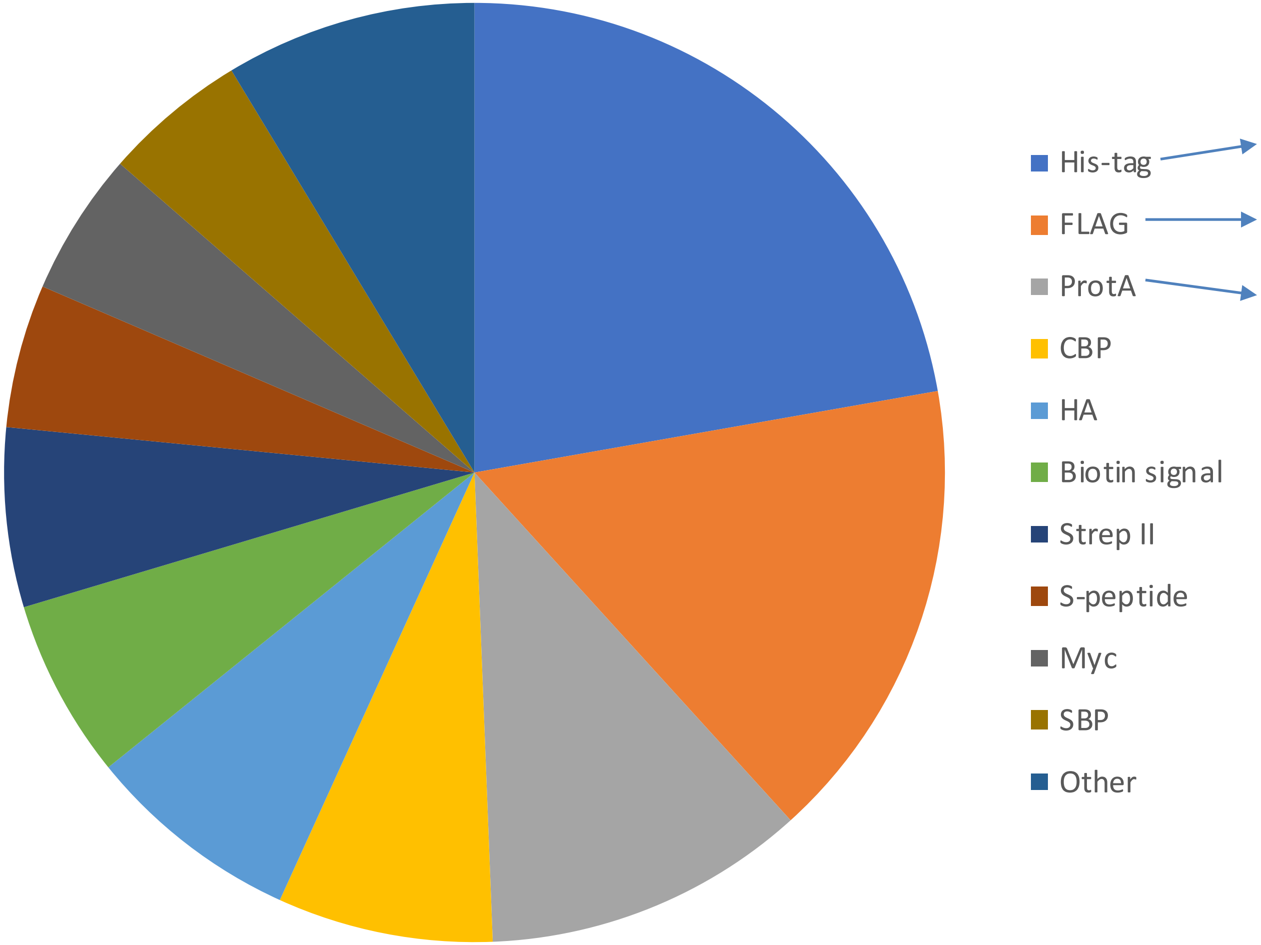
Co-immunoprecipitation probes multiple interactions

- Support
- Antibody
- Tag
- Bait
- Prey
- Contaminant



Diagrams courtesy of Alexey Nesvizhskii

Many possible combinations for tandem affinity purification



Used with:

IMAC column

Monoclonal antibody

Immunoglobulins

Two rounds of purification reduce non-specific binding, but they may lose low-affinity prey.

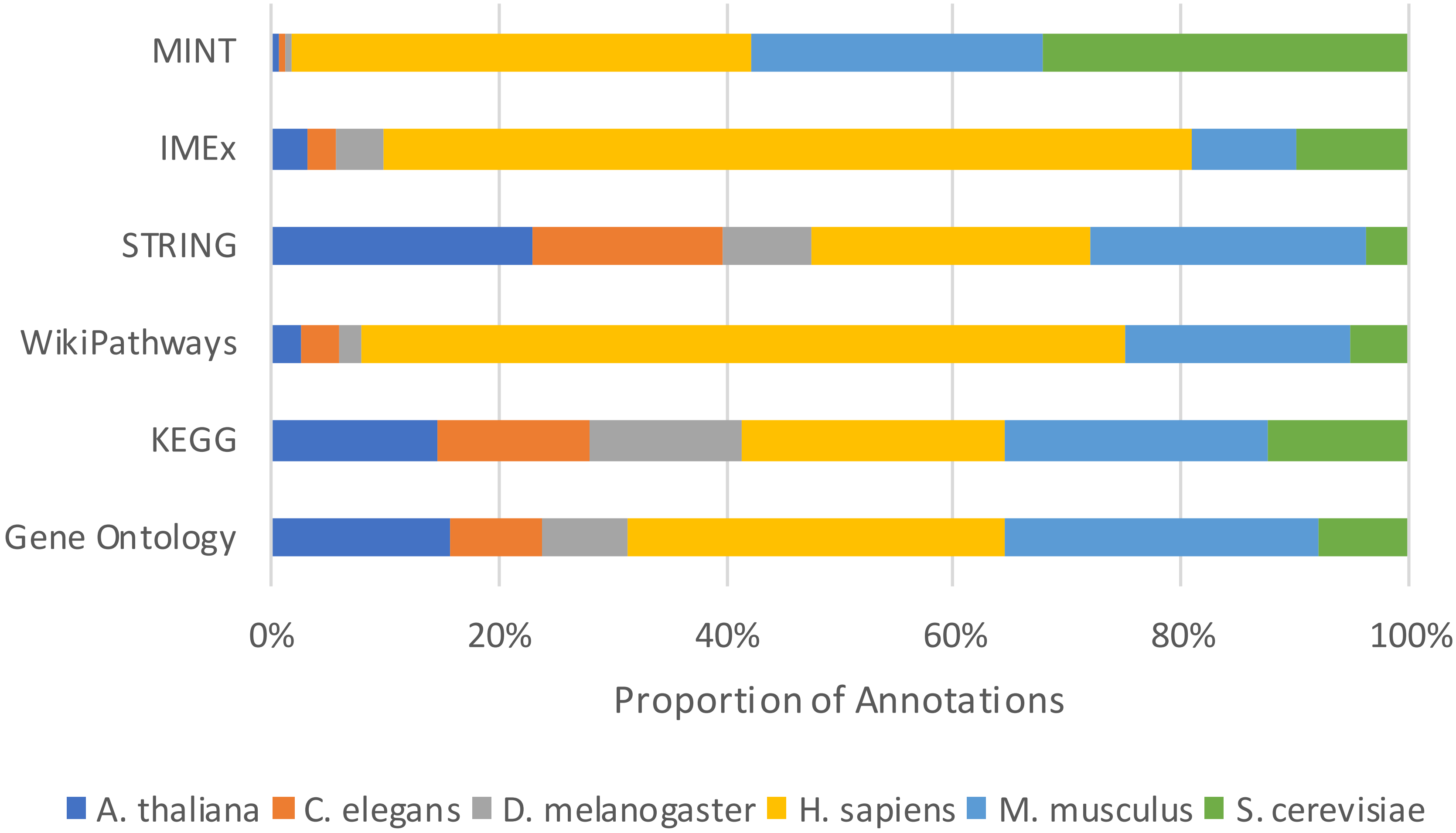
Yifeng Li. Biotechnol. *Appl. Biochem*. (2010) 55: 73-83.

When will structural data enable protein-protein inference?

- "Localized regions on protein surfaces [are] conserved among structural neighbors that participate in protein-protein interactions."
- "As long as structural information is available for a given pair of proteins… the set of 'template complexes' available in the current structural databases can be used to generate coarse-grained models of protein-protein interactions."

QC Zhang et al. *PNAS* (2010) 107:10896-10901

Pathway and network annotation vary considerably



Showing just 6 of 702 resources from http://pathguide.org/

Takeaway Messages

- Learning a transcript sequence allows you to leverage considerable sequence resources.
- Expression data may reveal co-expression partners besides responsiveness to stimuli.
- Colocalization and protein-protein interactions require experimentation.
- Pathway and network data skew by organism.