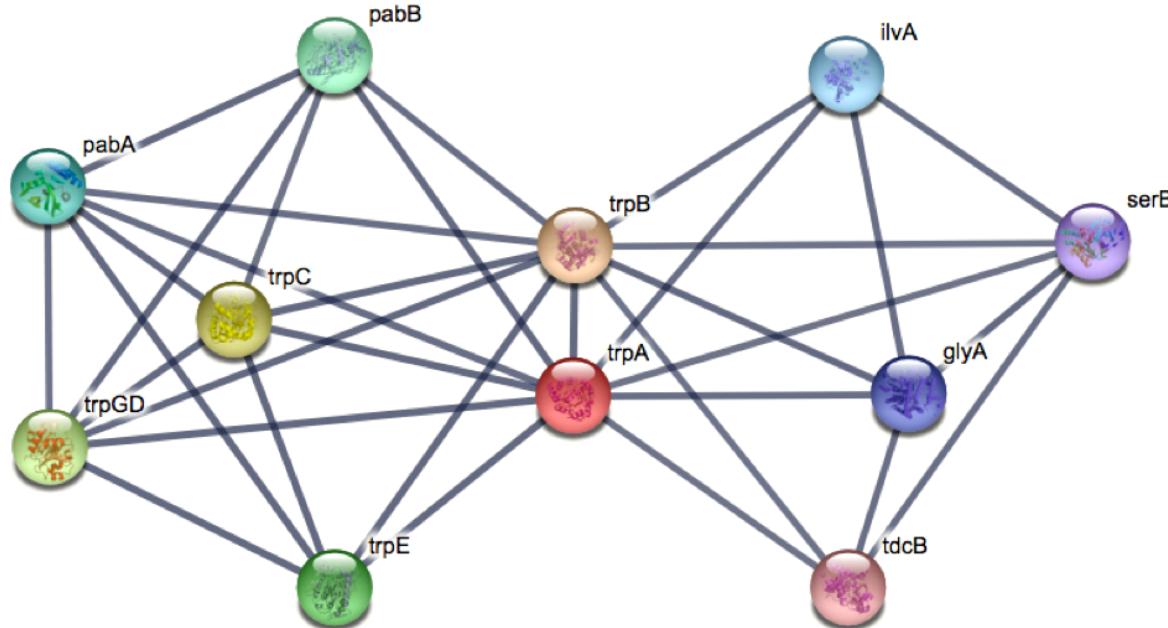


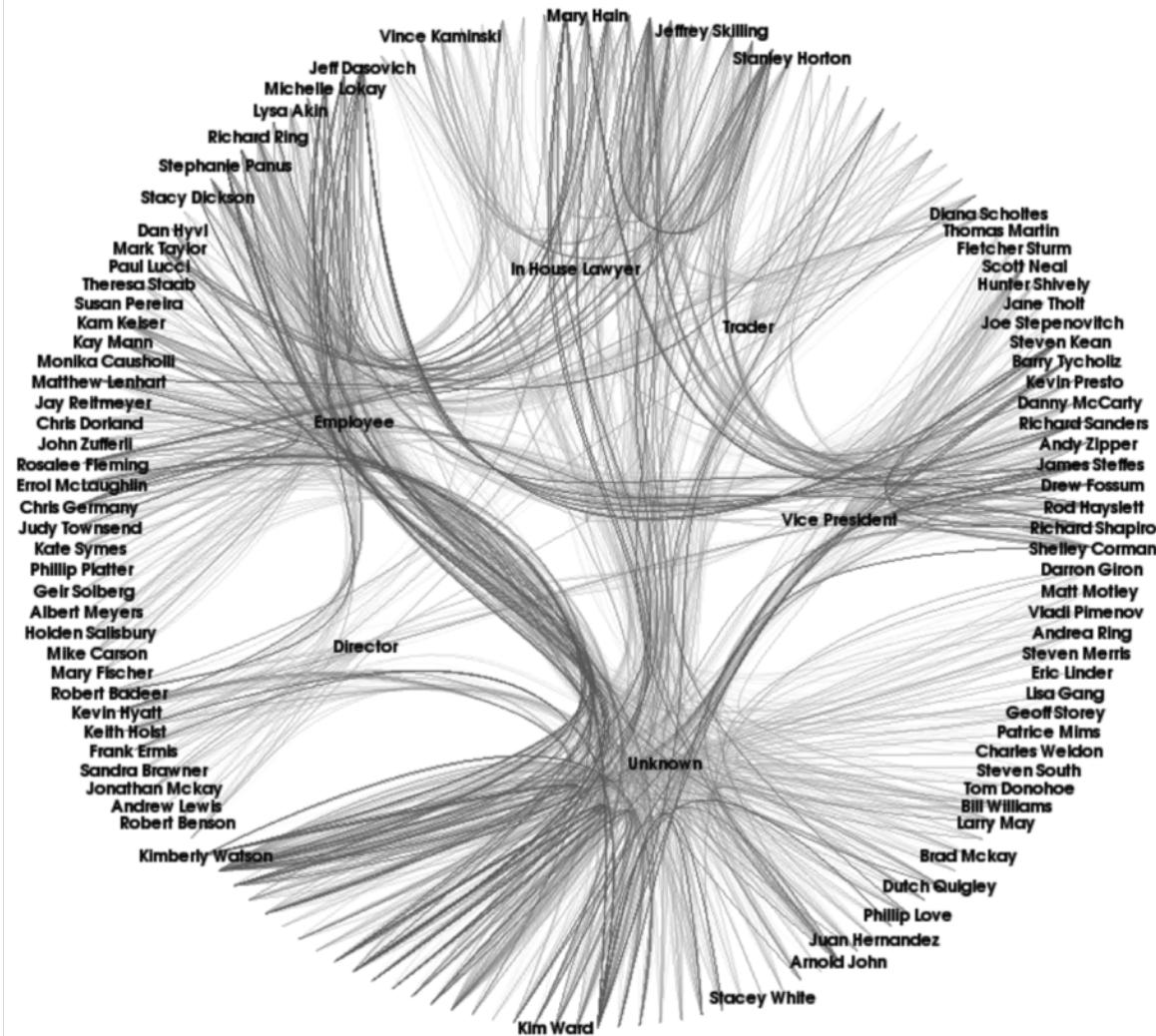
# The STRING database



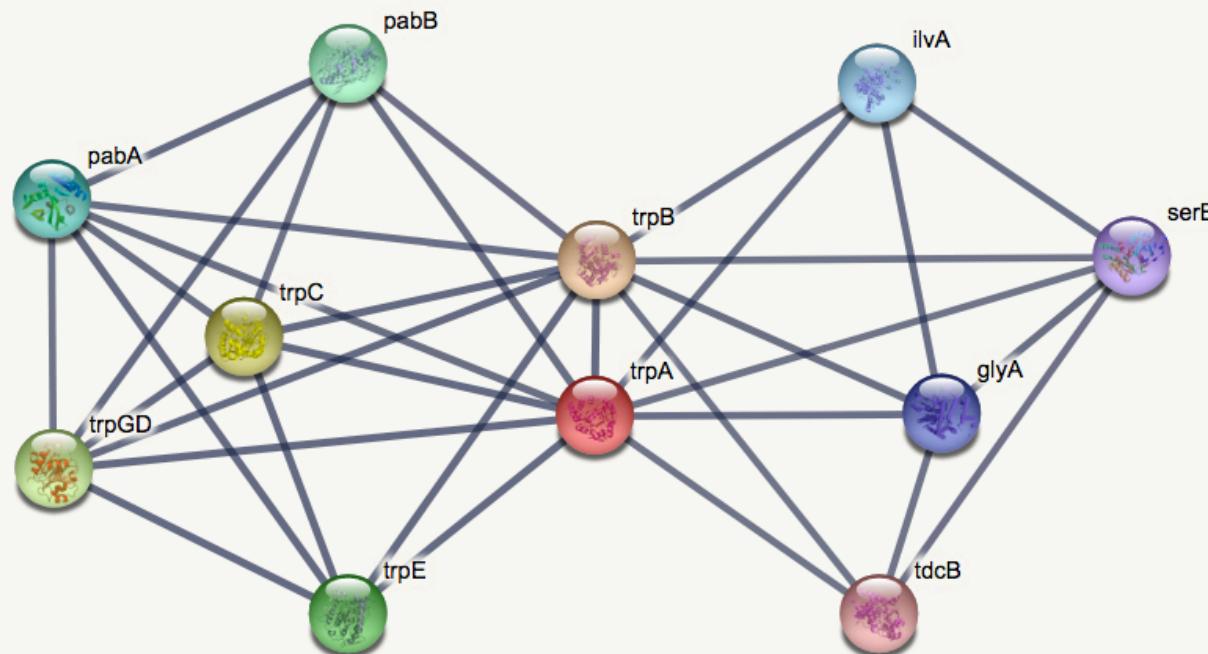
5090 genomes

24.6 million proteins

guilt by association



web interface



# Cytoscape stringApp



bulk download files

heavily used

100,000 users monthly

● Users

20,000

10,000

2011

2012

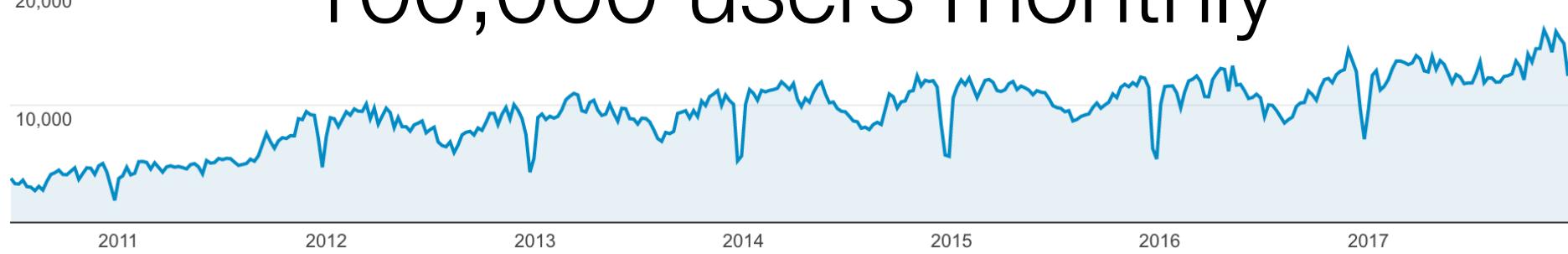
2013

2014

2015

2016

2017



well cited

>10,000 citations

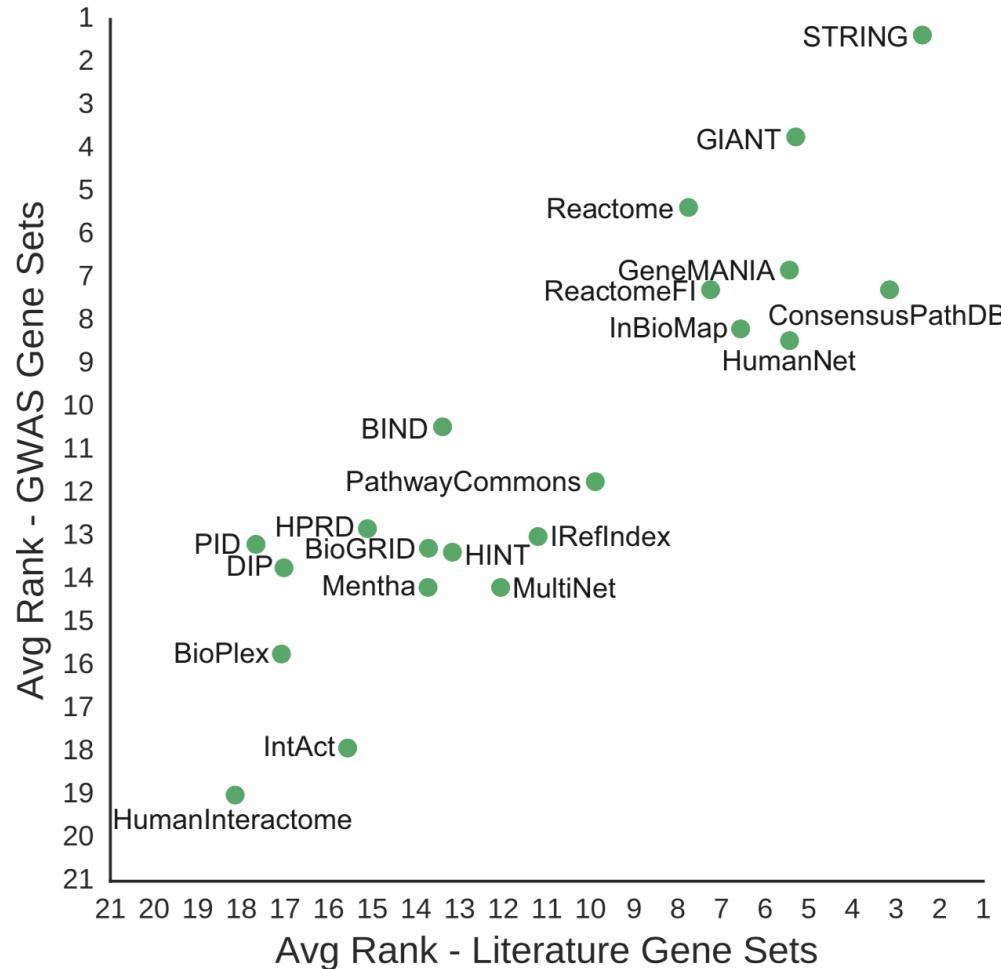
# ELIXIR Core Data Resource

works well

all tools work well

according to their authors

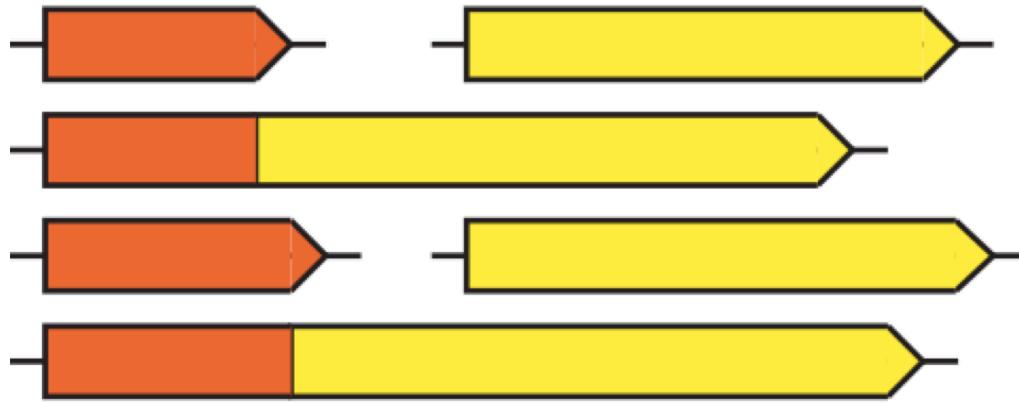
independent benchmarks



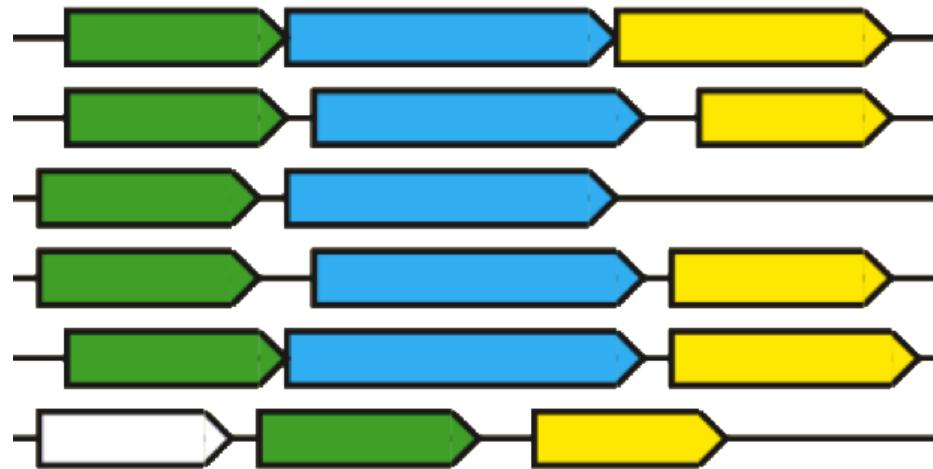
# data integration

genomic context

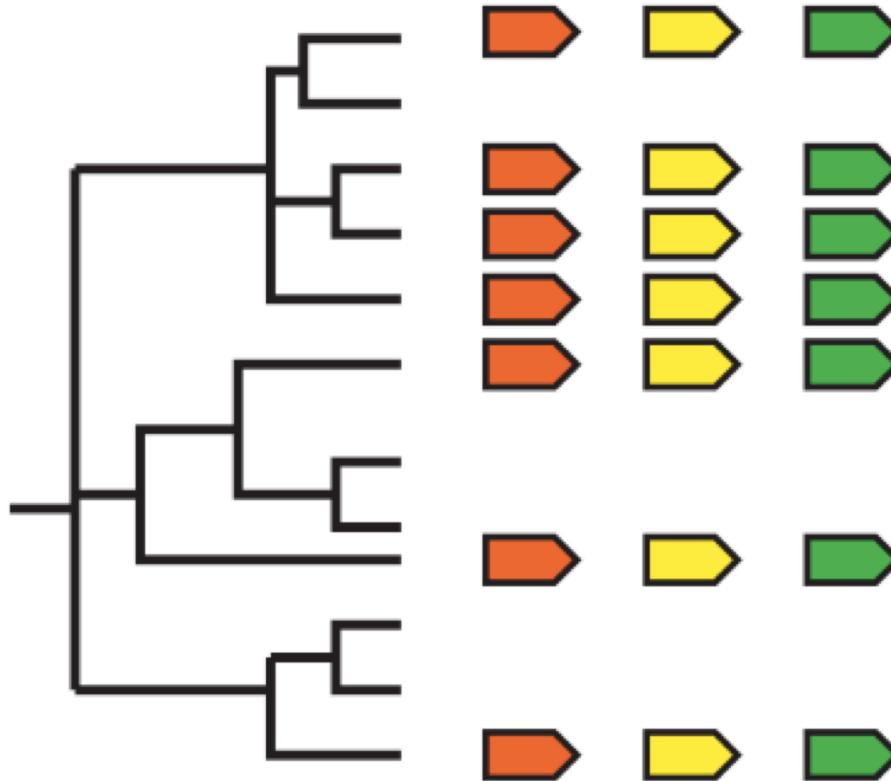
gene fusion



gene neighborhood



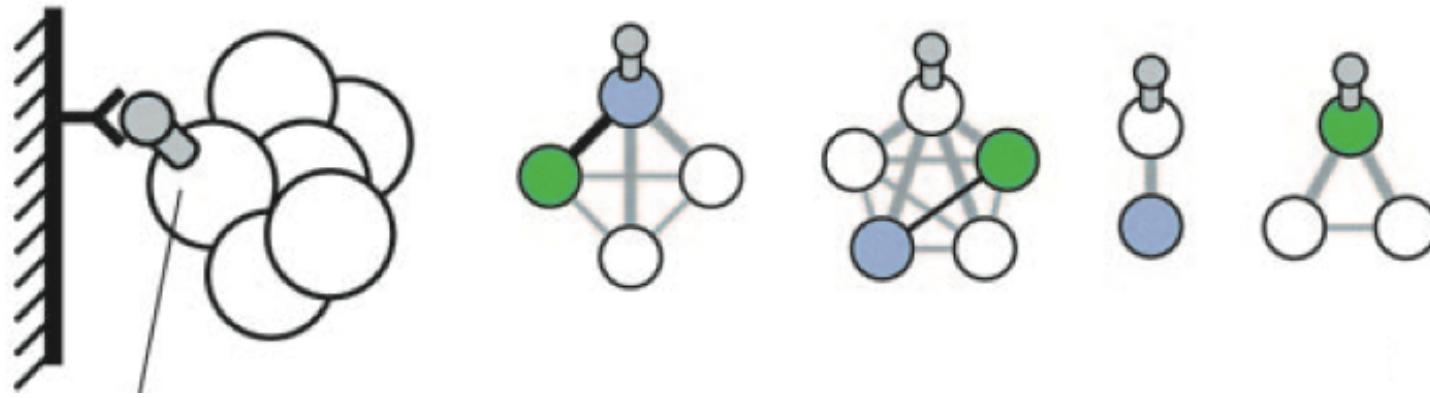
phylogenetic profiles



experimental data

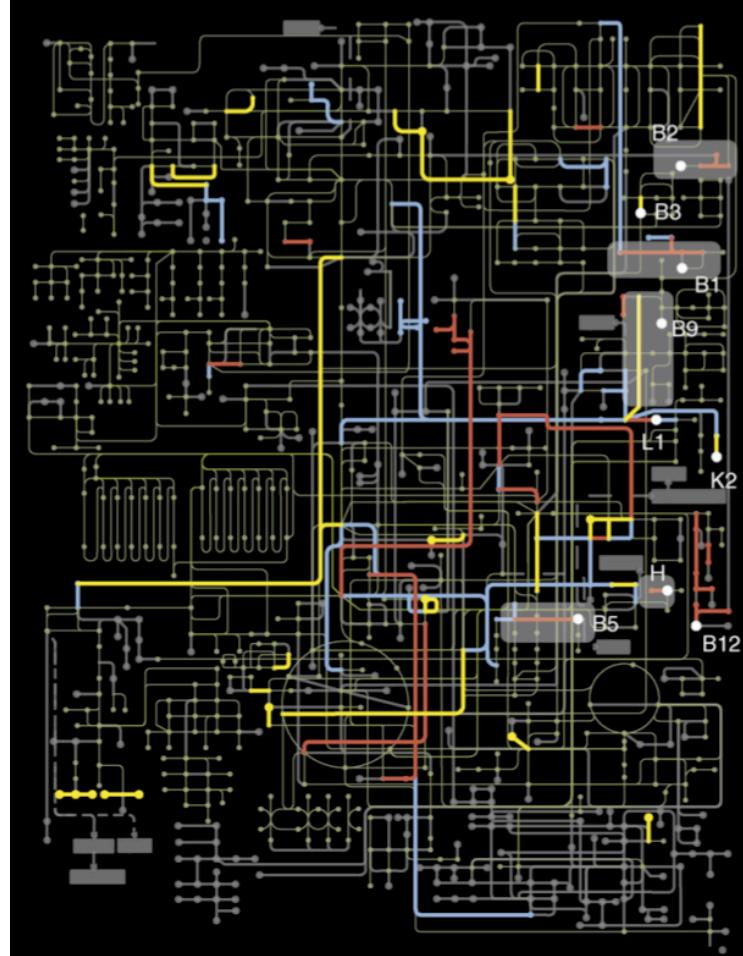
gene coexpression

physical interactions



curated knowledge

pathways



many databases

different formats

different identifiers

varying quality

not comparable

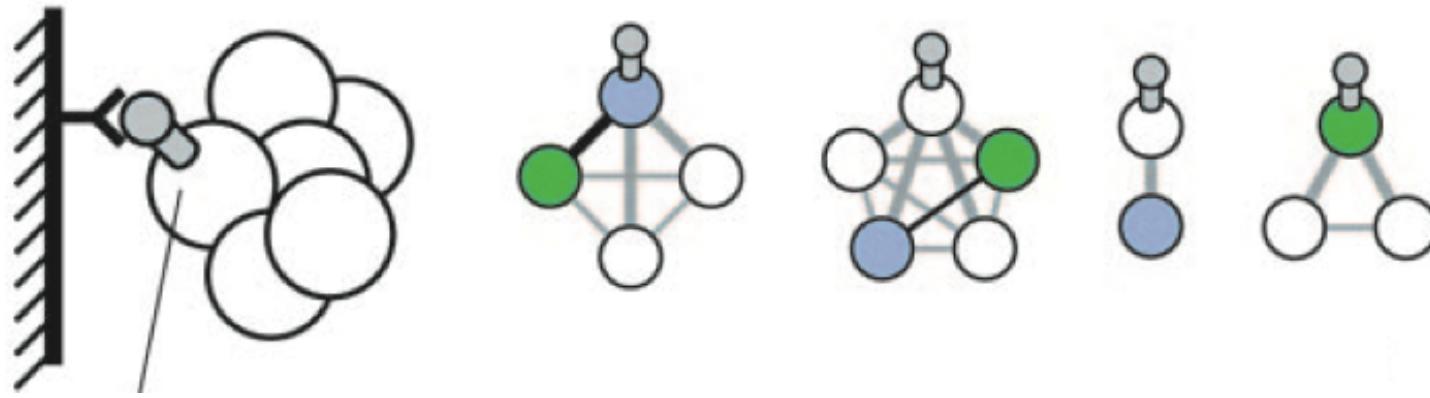
not same species

hard work

# parsers

mapping files

quality scores



benchmarking

gold standard pathways

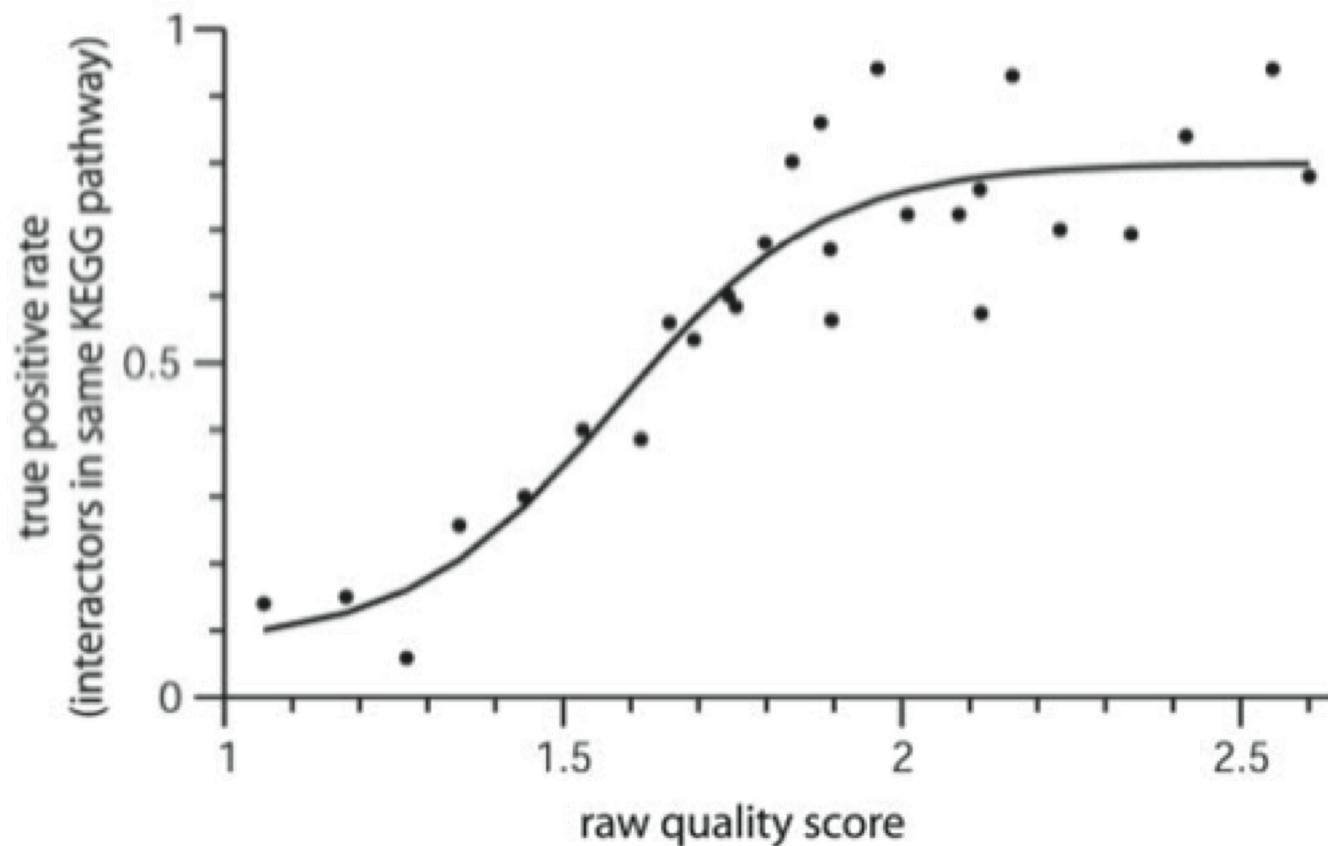
positive pairs

same pathway

negative pairs

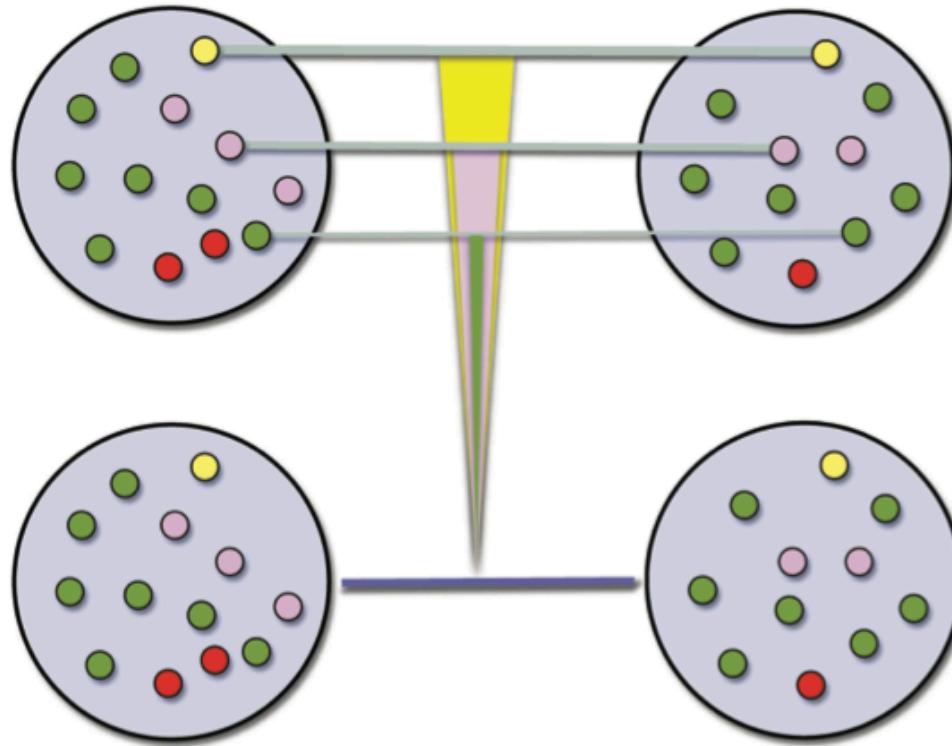
different pathways

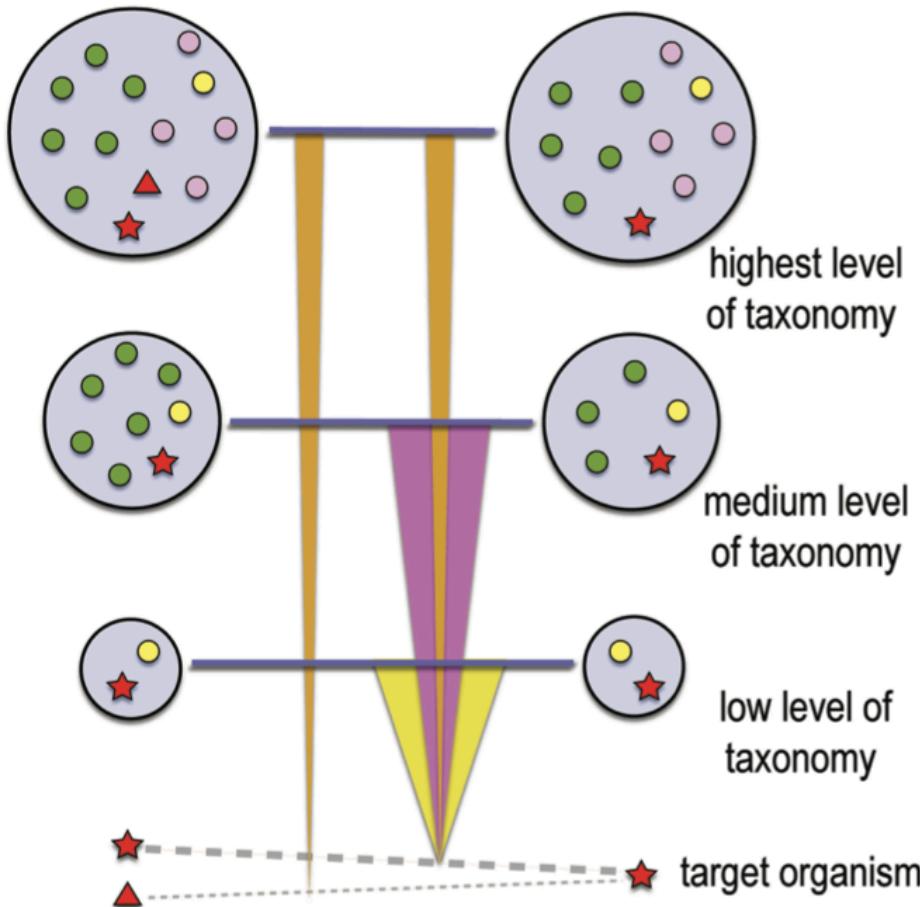
score calibration



common scale

orthology transfer





cross-species integration

missing most of the data

**text mining**

>10 km



too much to read

computer

as smart as a dog

teach it specific tricks

## What we say to dogs

Okay, Ginger! I've had it!  
You stay out of the garbage!  
Understand, Ginger? Stay out  
of the garbage, or else!



# What they hear



named entity recognition

dictionary

# synonyms

cyclin dependent kinase 1

CDC2

orthographic variation

cyclin dependent kinase 1

cyclin-dependent kinase 1

CDC2

hCDC2

black list

SDS

text corpus

~2.5 million full-text articles

~25 million abstracts

co-mentioning

document

paragraph

sentence

weighted count

score calibration

orthology transfer

# host-pathogen PPIs

Viruses.STRING

239 viruses

319 hosts

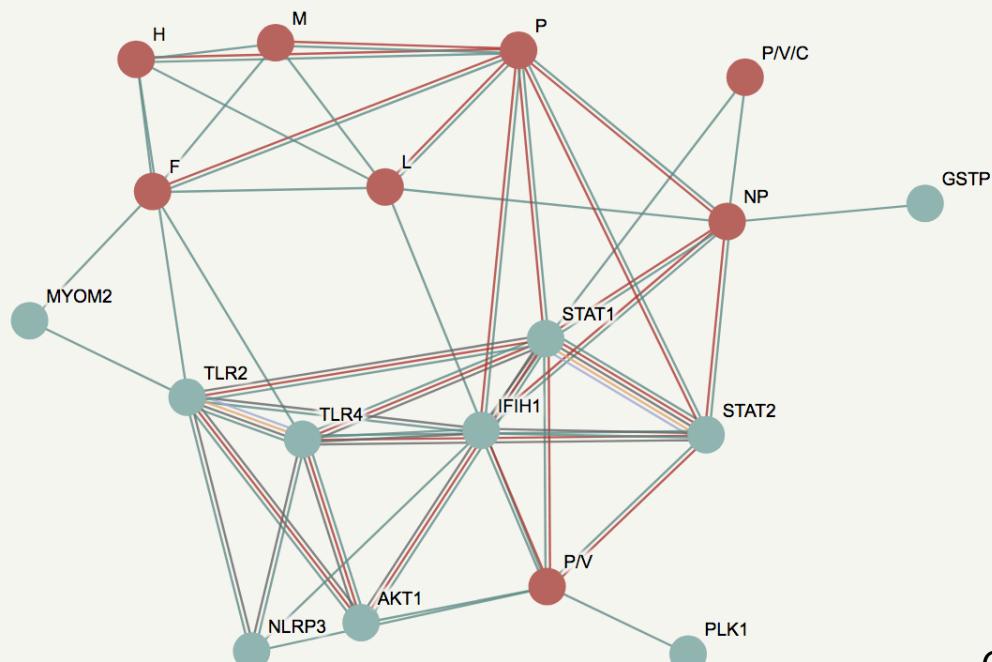
 Measles virus (strain Ichinose-B95a)

Measles virus (strain Ichinose-B95a)

NCBI taxonomy Id: 11234

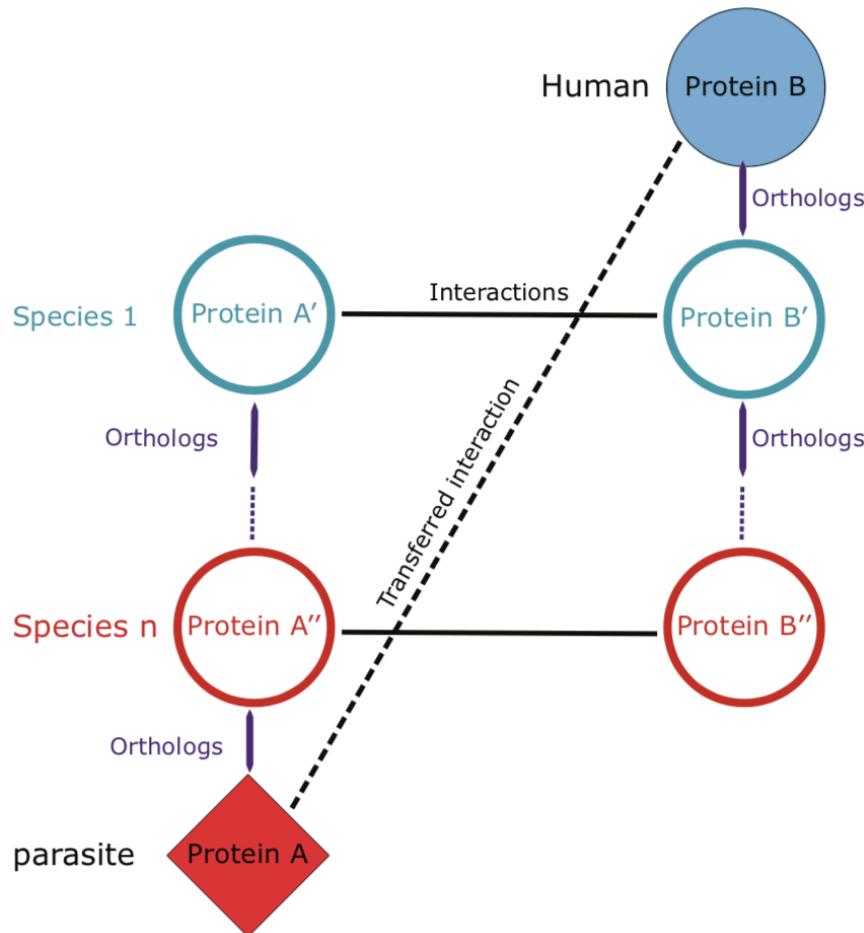
 Homo sapiens

NCBI taxonomy Id: 9606



OrthoHPI

orthology transfer

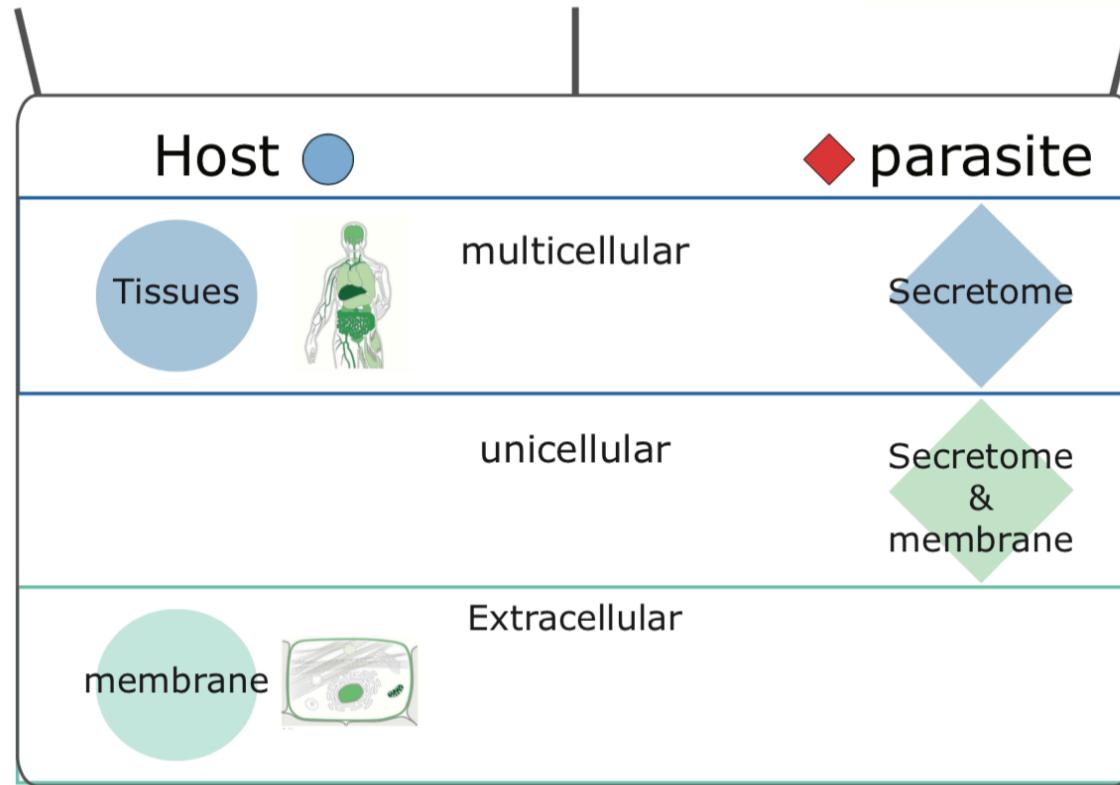


# localization filters

host proteome

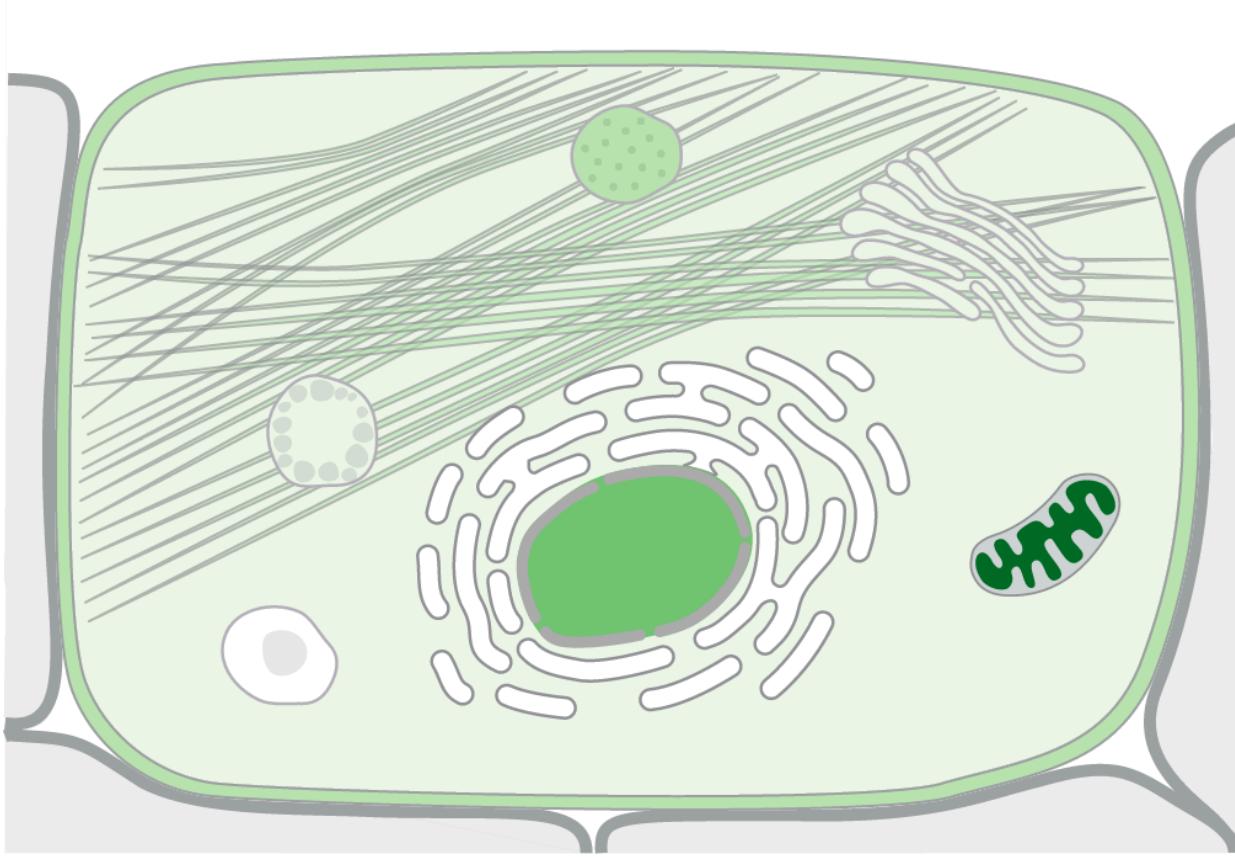
parasite attributes

parasite proteome

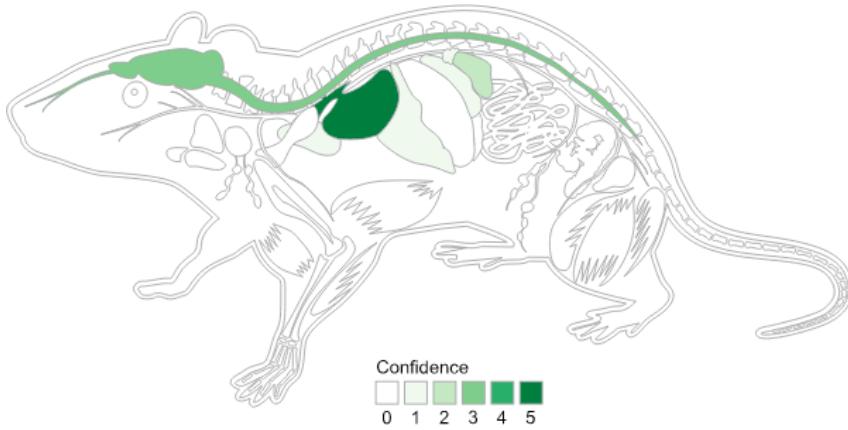
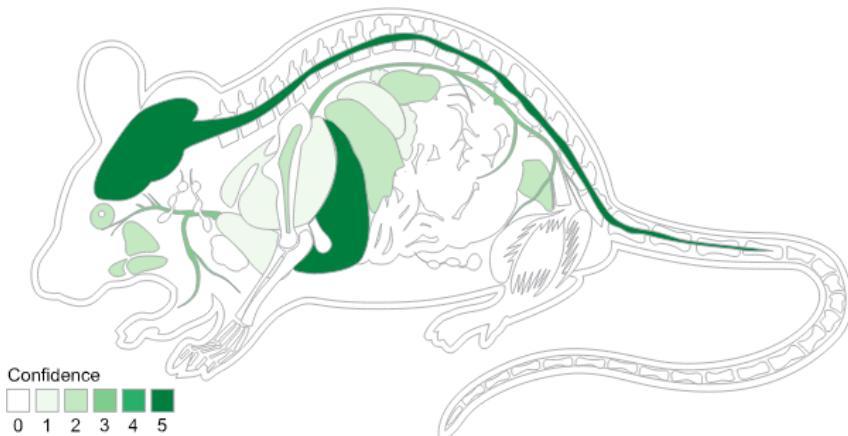
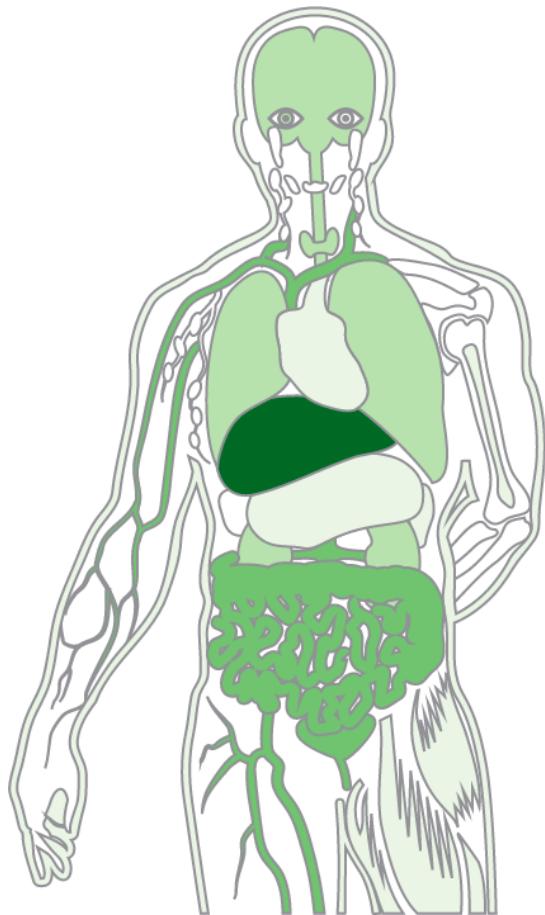


localization databases

# COMPARTMENTS



# TISSUES



same principles

data integration

experimental data

curated knowledge

text mining

named entity recognition

co-mentioning

common identifiers

calibrated scores

# Acknowledgments

## STRING

Damian Szklarczyk  
Annika Gable  
David Lyon  
Alexander Junge  
Stefan Wyder  
Jaime Huerta-Cepas  
Milan Simonovic  
Peer Bork  
Christian von Mering

## Host-pathogen networks

Helen Cook  
Yesid Cuesta-Astroz  
Alberto Santos  
Damian Szklarczyk  
Guilherme Oliveira

## Cytoscape stringApp

Nadezhda Doncheva  
John “Scooter” Morris

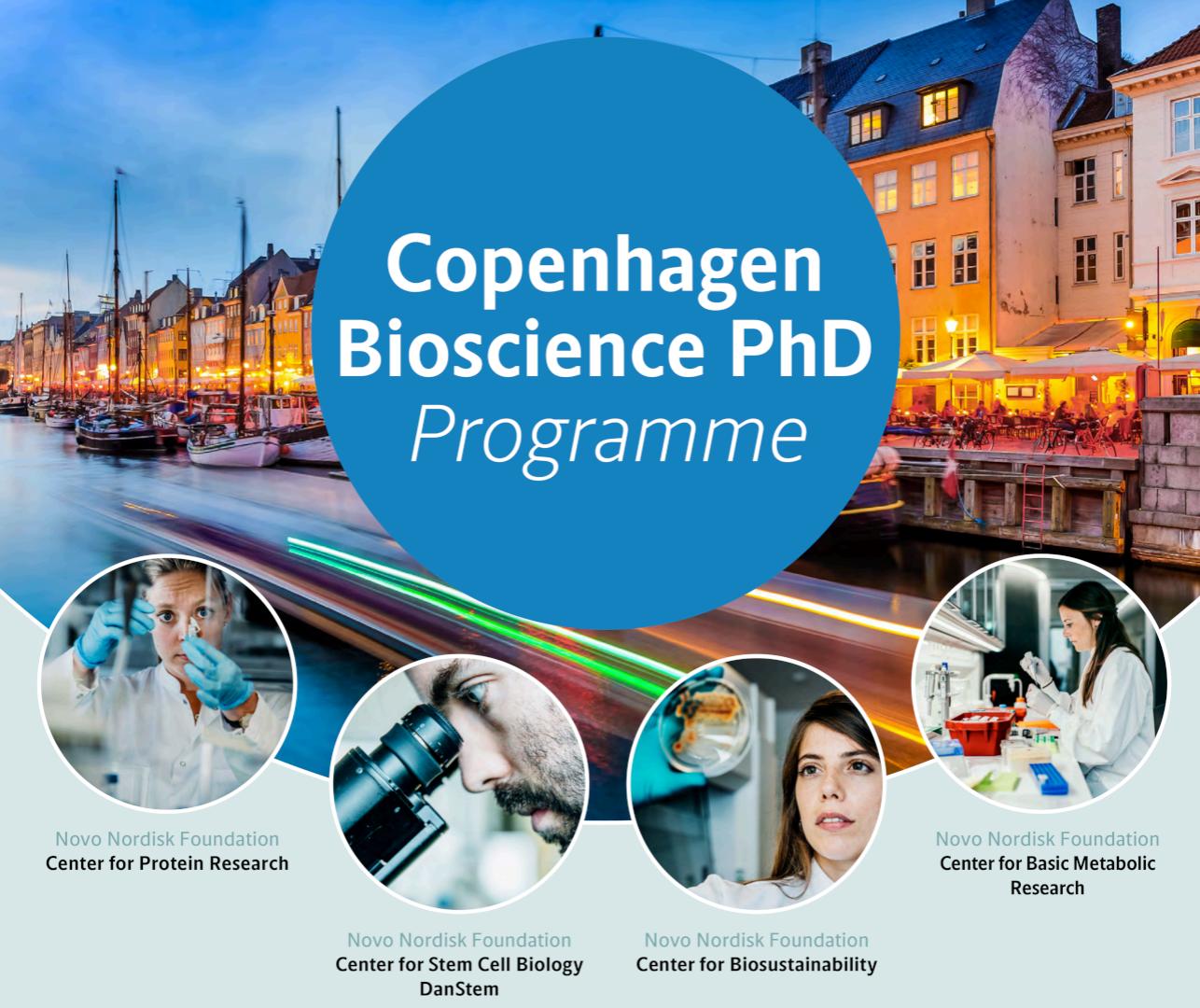


novo nordisk fonden



EMBL





**Opens:**  
2019-11-12

# Copenhagen Bioscience PhD Programme

**Deadline:**  
2020-01-09



Novo Nordisk Foundation  
Center for Protein Research



Novo Nordisk Foundation  
Center for Stem Cell Biology  
DanStem



Novo Nordisk Foundation  
Center for Biosustainability



Novo Nordisk Foundation  
Center for Basic Metabolic  
Research