A closer look at pathways and protein networks

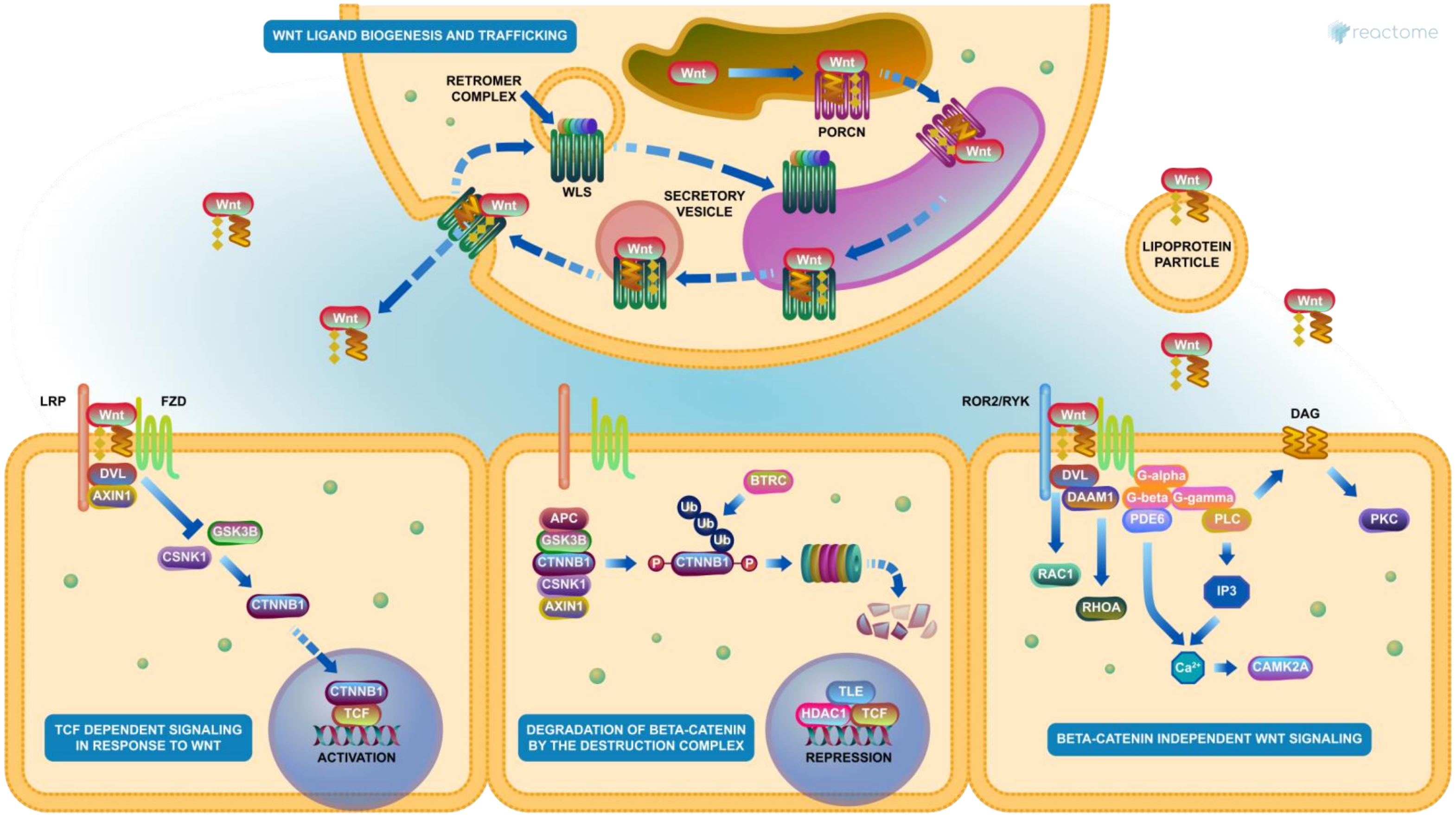
Harald Barsnes
Department of Biomedicine
& Department of Informatics
University of Bergen

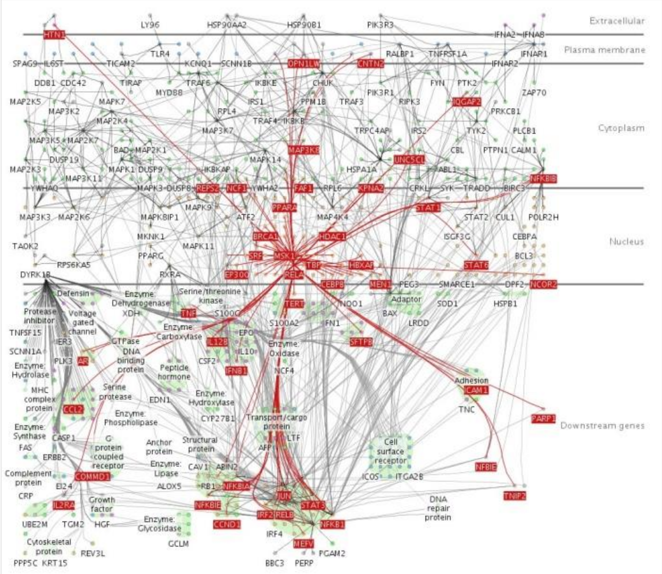
Faculty Lunch
March 6th 2019

Proteomics Unit PROBE University of Bergen

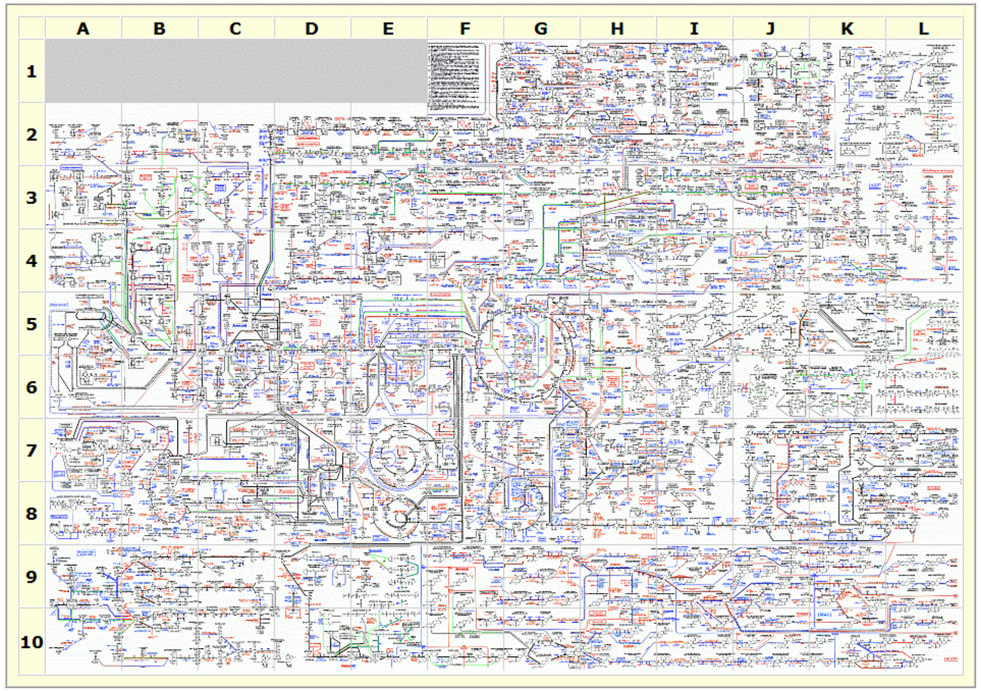
CBU COMPUTATIONAL BIOLOGY UNIT

What are pathways?

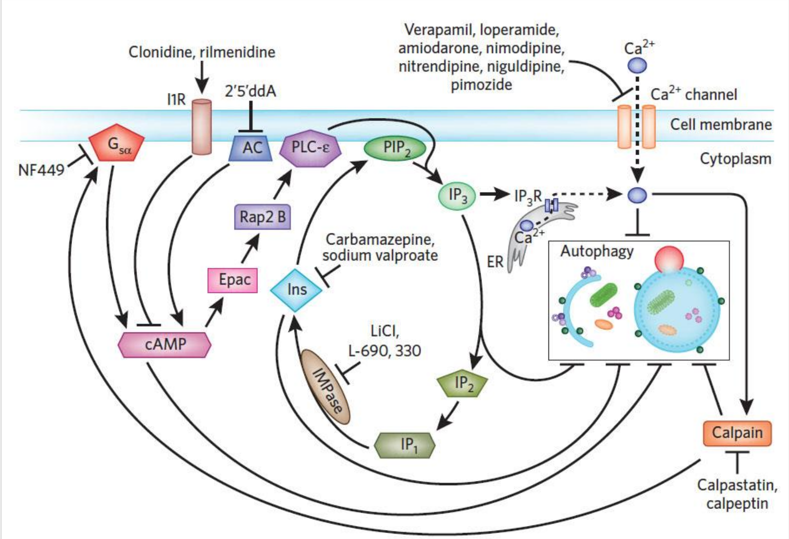




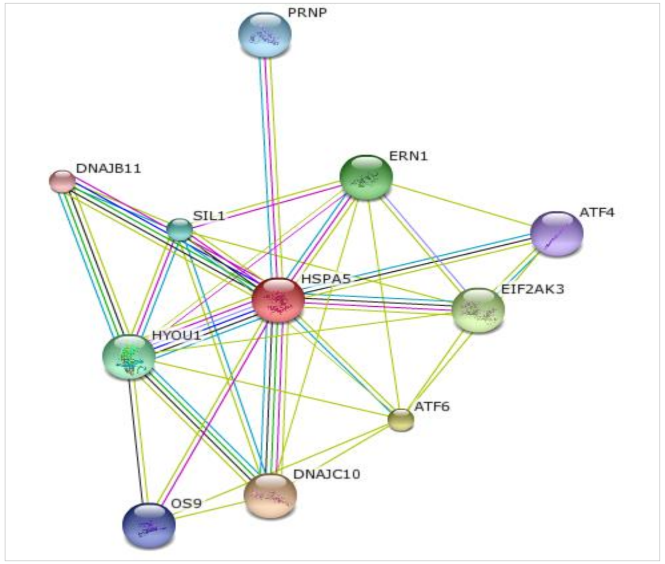
Extracellular
Plasma membrane
Cytoplasm
Nucleus
Downstream genes



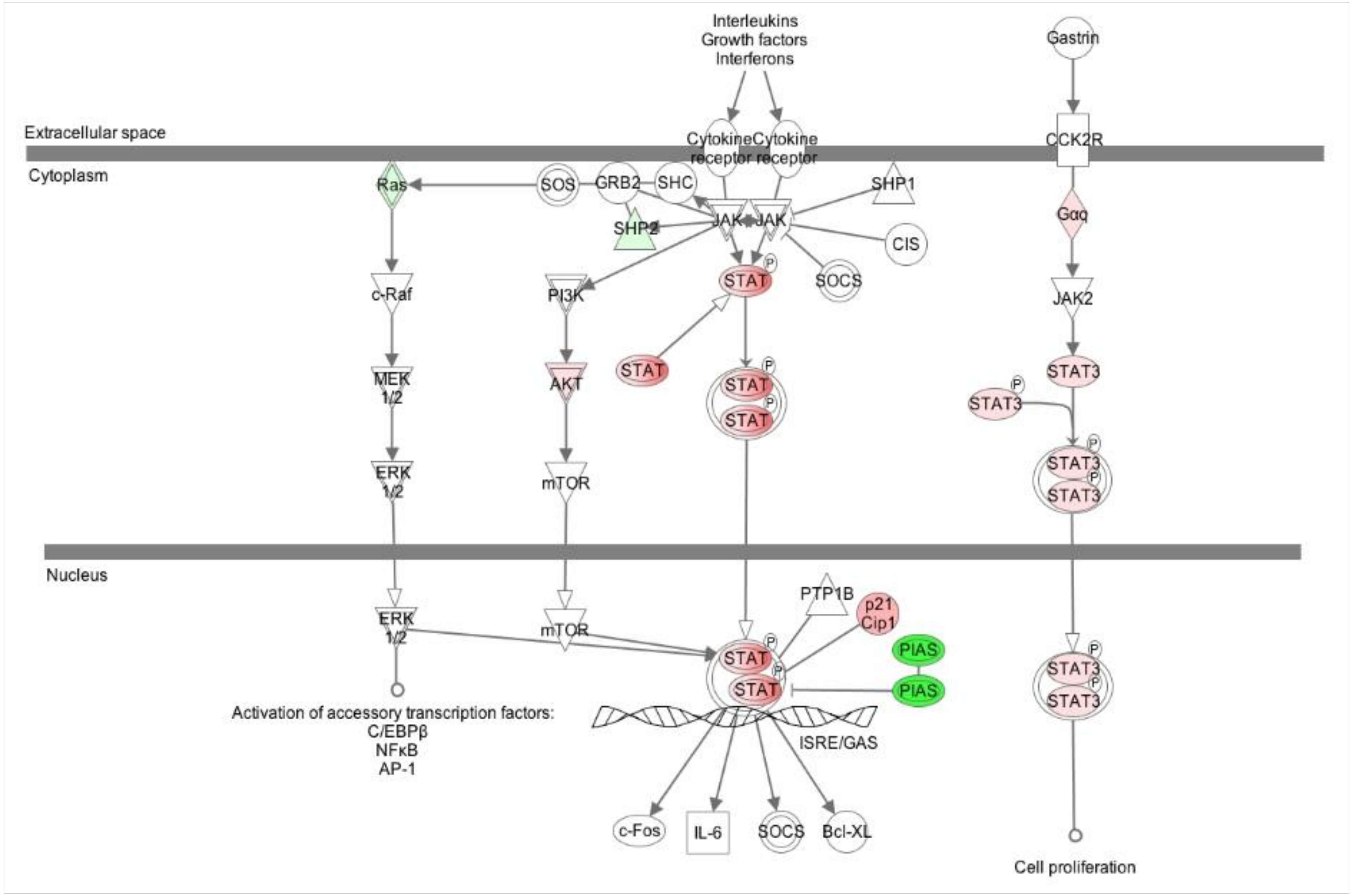
A
B
C
D
E
F
G
H
I
J
K
L
1
2
3
4
5
6
7
8
9
10



Clonidine, rilmenidine
I1R
2'5'ddA
Verapamil, loperamide, amiodarone, nimodipine, nitrendipine, niguldipine, pimozide
Ca2+
Ca2+ channel
Cell membrane
Cytoplasm
Gsα
NF449
AC
PLC-ε
PIP2
IP3
IP3R
Ca2+
ER
Rap2 B
Epac
cAMP
Ins
Carbamazepine, sodium valproate
LiCl, L-690, 330
IMPase
IP2
IP1
Autophagy
Calpain
Calpastatin, calpeptin

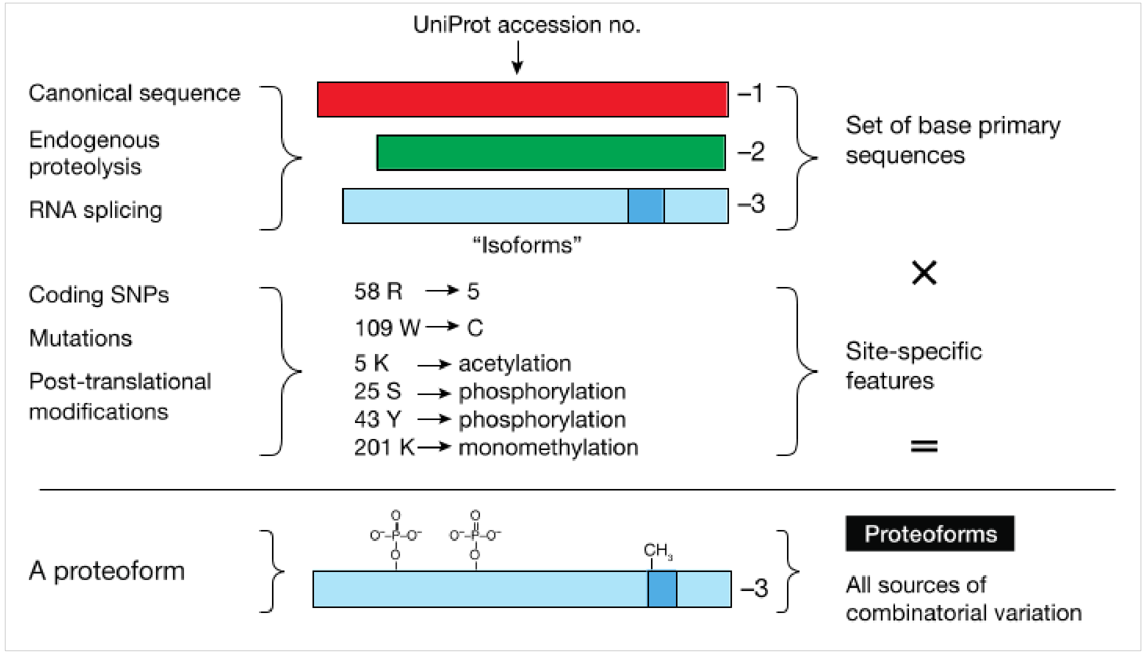


PRNP
DNAJB11
ERN1
SIL1
ATF4
HSPA5
EIF2AK3
HYOU1
ATF6
DNAJC10
OS9


Interleukins
Growth factors
Interferons
Gastrin
Extracellular space
Cytokine receptor
Cytokine receptor
CCK2R
Cytoplasm
Ras
SOS
GRB2
SHC
SHP1
SHP2
JAK
JAK
CIS
Gαq
STAT
P
SOCS
c-Raf
PI3K
JAK2
STAT
STAT
P
STAT
P
STAT3
STAT3
P
MEK 1/2
AKT
STAT3
P
STAT3
P
ERK 1/2
mTOR
Nucleus
PTP1B
p21 Cip1
ERK 1/2
mTOR
STAT
P
STAT
P
PIAS
PIAS
STAT3
P
STAT3
P
Activation of accessory transcription factors:
C/EBPβ
NFκB
AP-1
ISRE/GAS
c-Fos
IL-6
SOCS
Bcl-XL
Cell proliferation

Challenge I : Proteoforms

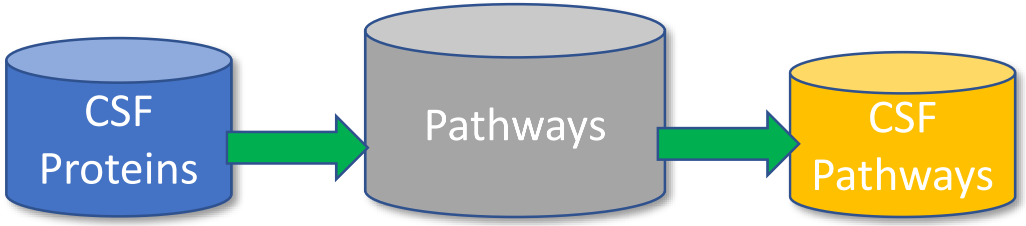
- Proteins are <u>not</u> genes
  - Proteoforms:
    - Modifications
    - Signal peptides
    - Isoforms
    - Mutations
    - SNPs
    - Location
    - Degradation

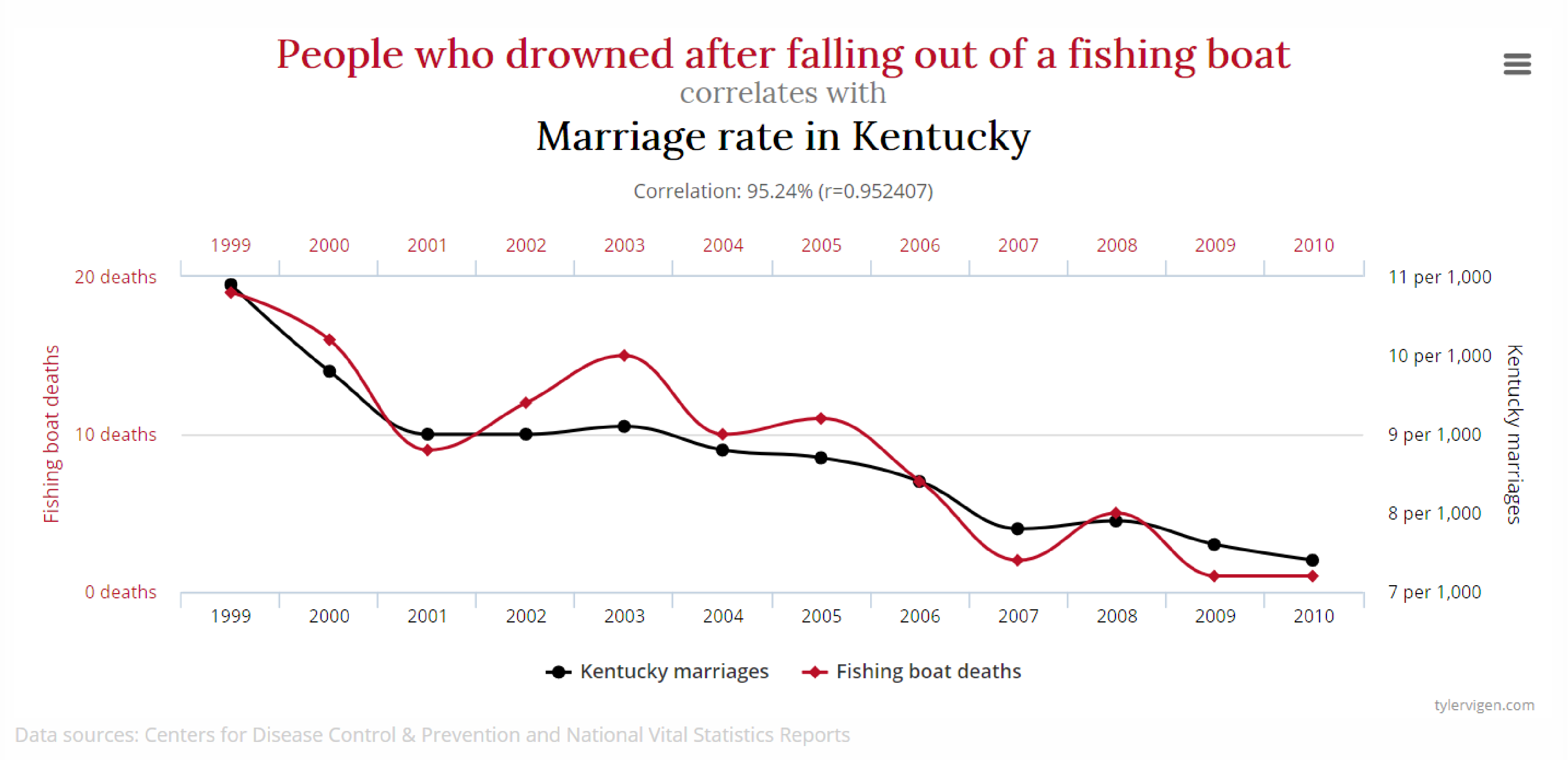


Challenge II : Biases

- Database bias
- Researcher bias
- Selection bias

"I already wrote the paper. That's why it's so hard to get the right data."





People who drowned after falling out of a fishing boat

correlates with

Marriage rate in Kentucky

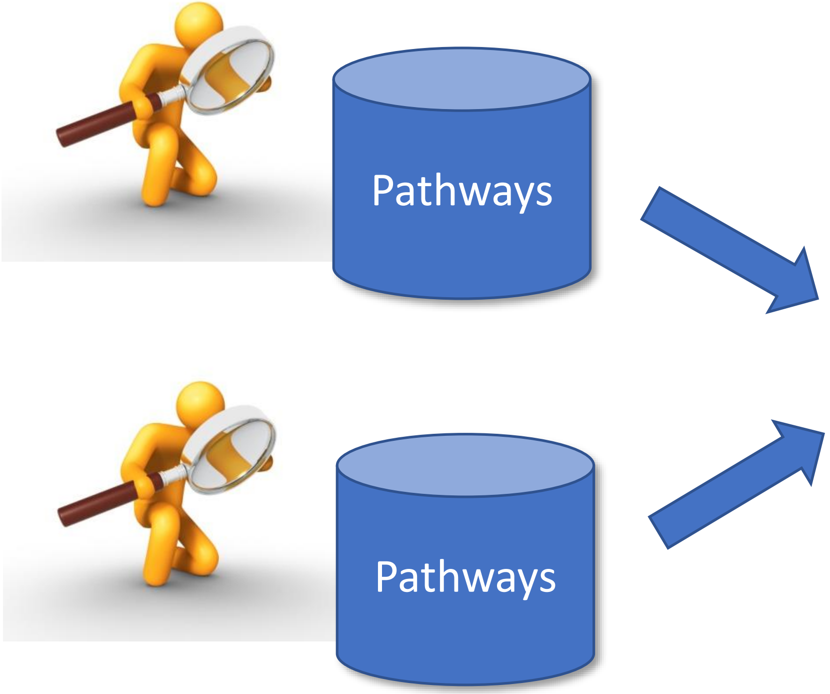
Correlation: 95.24% (r=0.952407)

tylervigen.com

Data sources: Centers for Disease Control & Prevention and National Vital Statistics Reports

Challenge III : Changing Databases

- Databases are not fixed



March 2019

Pathways

March 2020

Pathways

Same result..?

Impact of outdated gene annotations on pathway enrichment analysis

To the Editor: Pathway enrichment analysis is a common technique for interpreting gene lists derived from high-throughput experiments[1]. Its success depends on the quality of gene annotations. We analyzed the evolution of pathway knowledge and annotations over the past seven years and found that the use of outdated resources has strongly affected practical genomic analysis and recent literature: 67% of ~3,900 publications we surveyed in 2015 referenced outdated software that captured only 26% of biological processes and pathways identified using current resources.

Pathway analysis assesses the statistical enrichment of biological processes and pathways in a given gene list on the basis of information in Gene Ontology[2] (GO) and pathway databases such as Reactome[3] and PathwayCommons. GO is updated daily and Reactome versions are released quarterly, but many software tools interpret gene lists using functional information that has not been updated for years.

We surveyed the update times of 25 web-based pathway enrichment tools and citations of these tools in 3,879 publications (**Fig. 1a** and **Supplementary Tables 1** and **2**). Although nine tools (for example, g:Profiler[4] and PANTHER[5]) provided gene annotations that had been revised within six months (September 2015 through February 2016), most tools were outdated by several years. Ten (42%) were

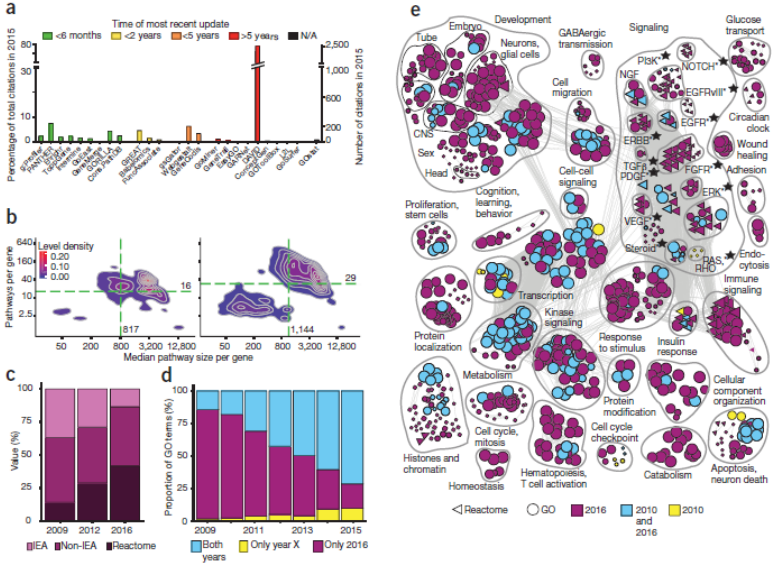
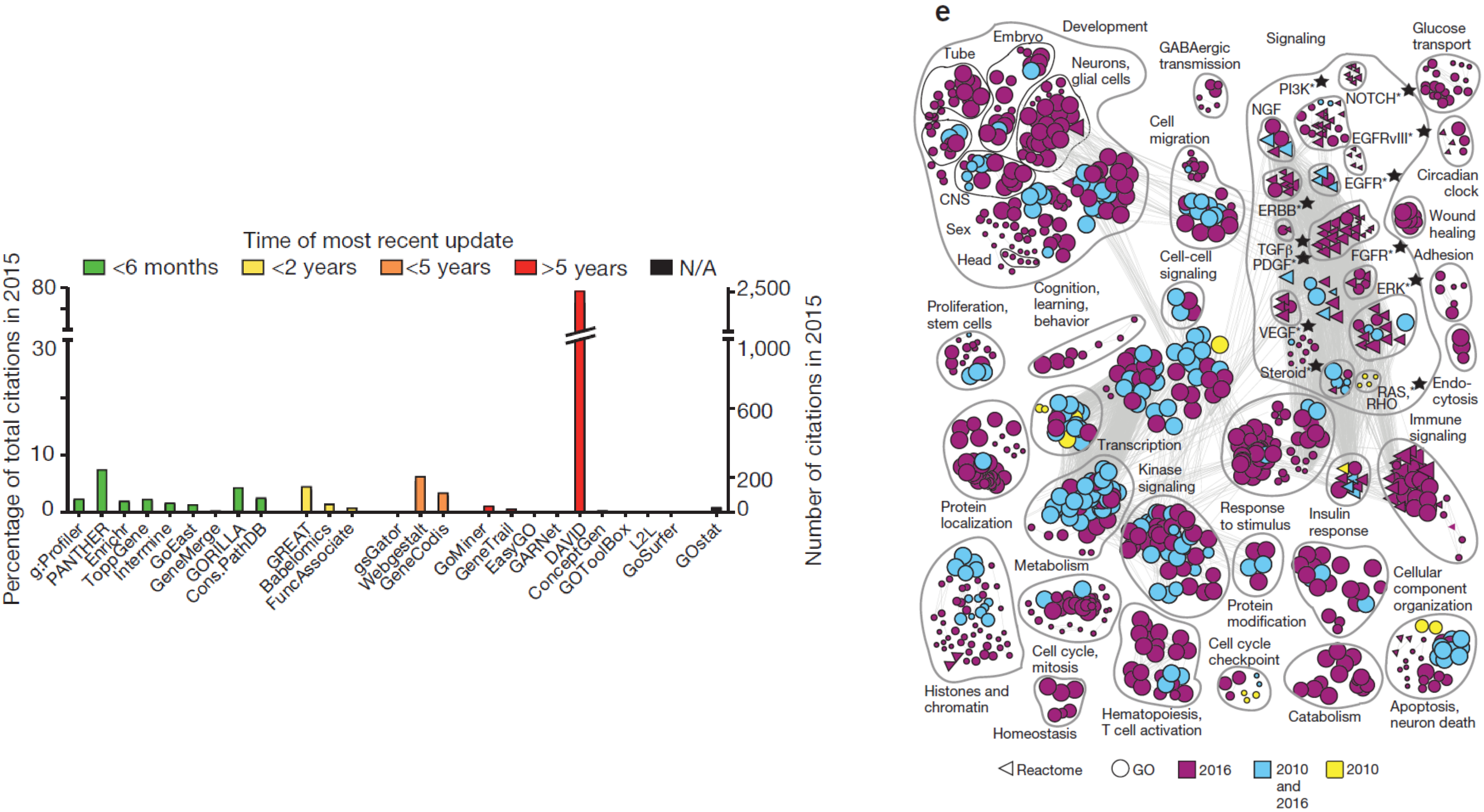


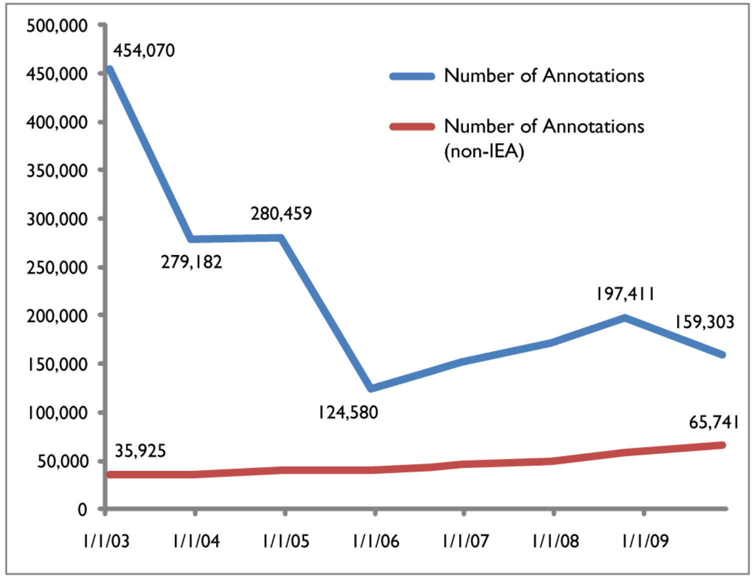
Figure 1 | Outdated pathway analysis resources strongly affect practical genomic analysis and literature. (**a**) The majority of public software tools for pathway enrichment analysis use outdated gene annotations, and the majority of surveyed papers published in 2015 used annotations that were more than five years old. (**b**) Density plots showing the evolution of pathway knowledge (GO + Reactome) between 2009 (left) and 2016 (right). The values for the median gene are indicated by green dashed lines. The bottom left group in the 2016 plot corresponds to Reactome pathways. (**c**) Gene annotation quality is improving rapidly as manually curated Reactome annotations are becoming more frequent and fewer genes in GO are IEA. (**d**) Pathway enrichment analysis of frequently mutated GBM genes showing the proportion of results missed in outdated GO annotations. Each bar compares annotations from a given year to 2016 annotations. (**e**) Enrichment map of frequently mutated GBM pathways and processes according to gene annotations from 2010 and 2016. Three-quarters of current findings are missed in out-of-date analyses (purple). Nodes represent processes and pathways, and edges connect nodes with many shared genes. Stars indicate clinically actionable pathways.



Time of most recent update
<6 months
<2 years
<5 years
>5 years
N/A
Percentage of total citations in 2015
80
30
10
0
Number of citations in 2015
2,500
1,000
600
200
0
g:Profiler
PANTHER
Enrichr
ToppGene
Intermine
GoEast
GeneMerge
GORILLA
Cons.PathDB
GREAT
Babelomics
FuncAssociate
gsGator
Webgestalt
GeneCodis
GoMiner
GeneTrail
EasyGO
GARNet
DAVID
ConceptGen
GOToolBox
L2L
GoSurfer
GOstat
e
Development
Embryo
Tube
Neurons, glial cells
CNS
Sex
Head
GABAergic transmission
Cell migration
Cell-cell signaling
Signaling
Glucose transport
PI3K*
NOTCH*
NGF
EGFRvIII*
EGFR*
Circadian clock
ERBB*
Wound healing
TGFβ
PDGF*
FGFR*
Adhesion
ERK*
VEGF*
Steroid*
RAS, RHO*
Endo- cytosis
Immune signaling
Proliferation, stem cells
Cognition, learning, behavior
Transcription
Kinase signaling
Protein localization
Response to stimulus
Insulin response
Metabolism
Protein modification
Cellular component organization
Cell cycle, mitosis
Cell cycle checkpoint
Histones and chromatin
Homeostasis
Hematopoiesis, T cell activation
Catabolism
Apoptosis, neuron death
◁ Reactome
○ GO
■ 2016
■ 2010 and 2016
■ 2010

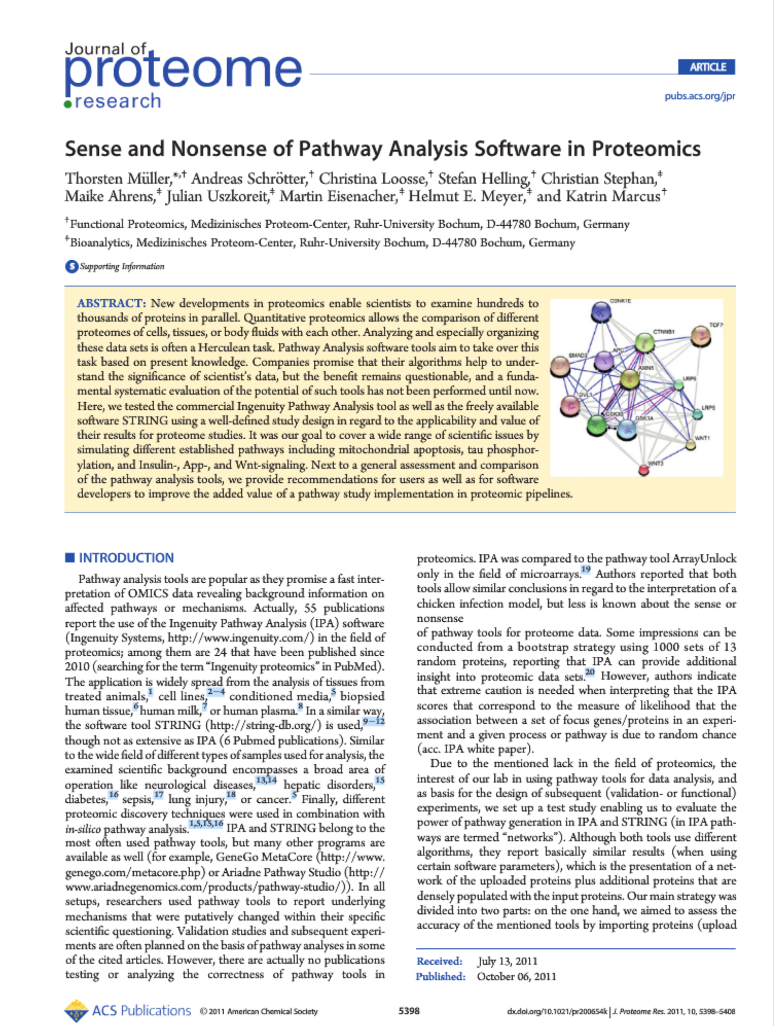
Challenge IV : Data Quality

- Pathway quality is variable
  - Nature paper vs. Student project?
    - How many peptides? How many spectra?
    - Which database?
    - What statistics?
    - Which experimental protocol?
    - …
- Electronically inferred..?



Khatri et al.: Ten years of pathway analysis: current approaches and outstanding challenges. PLoS Comput Biol. 2012;8(2):e1002375.

Commercial = better..?



Journal of proteome research

ARTICLE

pubs.acs.org/jpr

Sense and Nonsense of Pathway Analysis Software in Proteomics

Thorsten Müller,*,† Andreas Schrötter,† Christina Loosse,† Stefan Helling,† Christian Stephan,‡ Maike Ahrens,‡ Julian Uszkoreit,‡ Martin Eisenacher,‡ Helmut E. Meyer,‡ and Katrin Marcus†

†Functional Proteomics, Medizinisches Proteom-Center, Ruhr-University Bochum, D-44780 Bochum, Germany

‡Bioanalytics, Medizinisches Proteom-Center, Ruhr-University Bochum, D-44780 Bochum, Germany

S *Supporting Information*

ABSTRACT: New developments in proteomics enable scientists to examine hundreds to thousands of proteins in parallel. Quantitative proteomics allows the comparison of different proteomes of cells, tissues, or body fluids with each other. Analyzing and especially organizing these data sets is often a Herculean task. Pathway Analysis software tools aim to take over this task based on present knowledge. Companies promise that their algorithms help to understand the significance of scientist's data, but the benefit remains questionable, and a fundamental systematic evaluation of the potential of such tools has not been performed until now. Here, we tested the commercial Ingenuity Pathway Analysis tool as well as the freely available software STRING using a well-defined study design in regard to the applicability and value of their results for proteome studies. It was our goal to cover a wide range of scientific issues by simulating different established pathways including mitochondrial apoptosis, tau phosphorylation, and Insulin-, App-, and Wnt-signaling. Next to a general assessment and comparison of the pathway analysis tools, we provide recommendations for users as well as for software developers to improve the added value of a pathway study implementation in proteomic pipelines.

INTRODUCTION

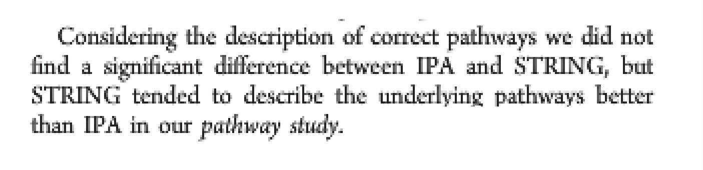
Pathway analysis tools are popular as they promise a fast interpretation of OMICS data revealing background information on affected pathways or mechanisms. Actually, 55 publications report the use of the Ingenuity Pathway Analysis (IPA) software (Ingenuity Systems, http://www.ingenuity.com/) in the field of proteomics; among them are 24 that have been published since 2010 (searching for the term "Ingenuity proteomics" in PubMed). The application is widely spread from the analysis of tissues from treated animals,[1] cell lines,[2−4] conditioned media,[5] biopsied human tissue,[6] human milk,[7] or human plasma.[8] In a similar way, the software tool STRING (http://string-db.org/) is used,[9−12] though not as extensive as IPA (6 Pubmed publications). Similar to the wide field of different types of samples used for analysis, the examined scientific background encompasses a broad area of operation like neurological diseases,[13,14] hepatic disorders,[15] diabetes,[16] sepsis,[17] lung injury,[18] or cancer.[5] Finally, different proteomic discovery techniques were used in combination with *in-silico* pathway analysis.[1,5,15,16] IPA and STRING belong to the most often used pathway tools, but many other programs are available as well (for example, GeneGo MetaCore (http://www.genego.com/metacore.php) or Ariadne Pathway Studio (http://www.ariadnegenomics.com/products/pathway-studio/)). In all setups, researchers used pathway tools to report underlying mechanisms that were putatively changed within their specific scientific questioning. Validation studies and subsequent experiments are often planned on the basis of pathway analyses in some of the cited articles. However, there are actually no publications testing or analyzing the correctness of pathway tools in proteomics. IPA was compared to the pathway tool ArrayUnlock only in the field of microarrays.[19] Authors reported that both tools allow similar conclusions in regard to the interpretation of a chicken infection model, but less is known about the sense or nonsense

of pathway tools for proteome data. Some impressions can be conducted from a bootstrap strategy using 1000 sets of 13 random proteins, reporting that IPA can provide additional insight into proteomic data sets.[20] However, authors indicate that extreme caution is needed when interpreting that the IPA scores that correspond to the measure of likelihood that the association between a set of focus genes/proteins in an experiment and a given process or pathway is due to random chance (acc. IPA white paper).

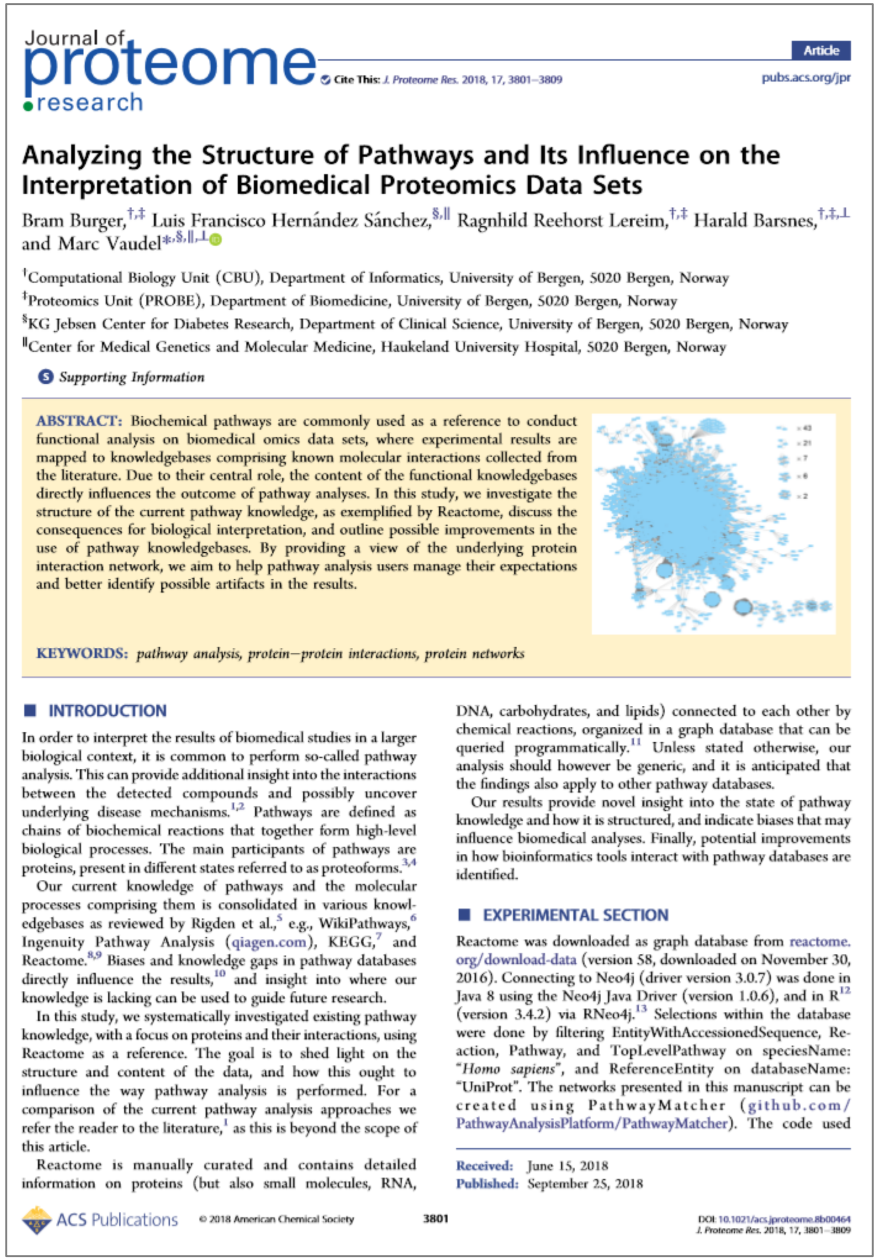
Due to the mentioned lack in the field of proteomics, the interest of our lab in using pathway tools for data analysis, and as basis for the design of subsequent (validation- or functional) experiments, we set up a test study enabling us to evaluate the power of pathway generation in IPA and STRING (in IPA pathways are termed "networks"). Although both tools use different algorithms, they report basically similar results (when using certain software parameters), which is the presentation of a network of the uploaded proteins plus additional proteins that are densely populated with the input proteins. Our main strategy was divided into two parts: on the one hand, we aimed to assess the accuracy of the mentioned tools by importing proteins (upload

Received: July 13, 2011
Published: October 06, 2011

dx.doi.org/10.1021/pr200654k | *J. Proteome Res.* 2011, 10, 5398–5408



Considering the description of correct pathways we did not find a significant difference between IPA and STRING, but STRING tended to describe the underlying pathways better than IPA in our *pathway study*.



Journal of proteome research

Article

Cite This: J. Proteome Res. 2018, 17, 3801−3809

pubs.acs.org/jpr

Analyzing the Structure of Pathways and Its Influence on the Interpretation of Biomedical Proteomics Data Sets

Bram Burger,[†,‡] Luis Francisco Hernández Sánchez,[§,∥] Ragnhild Reehorst Lereim,[†,‡] Harald Barsnes,[†,‡,⊥] and Marc Vaudel[*,§,∥,⊥]

[†] Computational Biology Unit (CBU), Department of Informatics, University of Bergen, 5020 Bergen, Norway
[‡] Proteomics Unit (PROBE), Department of Biomedicine, University of Bergen, 5020 Bergen, Norway
[§] KG Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, 5020 Bergen, Norway
[∥] Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, 5020 Bergen, Norway

Supporting Information

ABSTRACT: Biochemical pathways are commonly used as a reference to conduct functional analysis on biomedical omics data sets, where experimental results are mapped to knowledgebases comprising known molecular interactions collected from the literature. Due to their central role, the content of the functional knowledgebases directly influences the outcome of pathway analyses. In this study, we investigate the structure of the current pathway knowledge, as exemplified by Reactome, discuss the consequences for biological interpretation, and outline possible improvements in the use of pathway knowledgebases. By providing a view of the underlying protein interaction network, we aim to help pathway analysis users manage their expectations and better identify possible artifacts in the results.

KEYWORDS: *pathway analysis, protein−protein interactions, protein networks*

INTRODUCTION

In order to interpret the results of biomedical studies in a larger biological context, it is common to perform so-called pathway analysis. This can provide additional insight into the interactions between the detected compounds and possibly uncover underlying disease mechanisms.[1,2] Pathways are defined as chains of biochemical reactions that together form high-level biological processes. The main participants of pathways are proteins, present in different states referred to as proteoforms.[3,4]

Our current knowledge of pathways and the molecular processes comprising them is consolidated in various knowledgebases as reviewed by Rigden et al.,[5] e.g., WikiPathways,[6] Ingenuity Pathway Analysis (qiagen.com), KEGG,[7] and Reactome.[8,9] Biases and knowledge gaps in pathway databases directly influence the results,[10] and insight into where our knowledge is lacking can be used to guide future research.

In this study, we systematically investigated existing pathway knowledge, with a focus on proteins and their interactions, using Reactome as a reference. The goal is to shed light on the structure and content of the data, and how this ought to influence the way pathway analysis is performed. For a comparison of the current pathway analysis approaches we refer the reader to the literature,[1] as this is beyond the scope of this article.

Reactome is manually curated and contains detailed information on proteins (but also small molecules, RNA, DNA, carbohydrates, and lipids) connected to each other by chemical reactions, organized in a graph database that can be queried programmatically.[11] Unless stated otherwise, our analysis should however be generic, and it is anticipated that the findings also apply to other pathway databases.

Our results provide novel insight into the state of pathway knowledge and how it is structured, and indicate biases that may influence biomedical analyses. Finally, potential improvements in how bioinformatics tools interact with pathway databases are identified.

EXPERIMENTAL SECTION

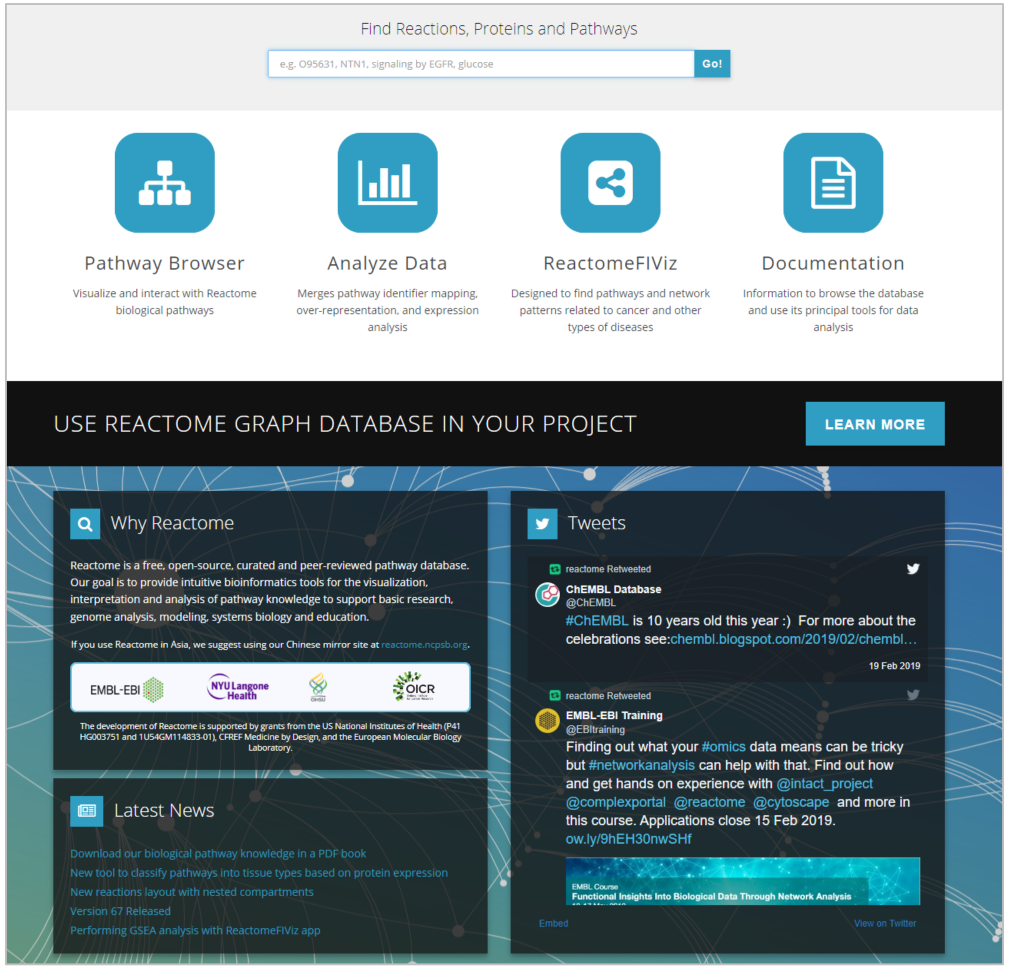
Reactome was downloaded as graph database from reactome.org/download-data (version 58, downloaded on November 30, 2016). Connecting to Neo4j (driver version 3.0.7) was done in Java 8 using the Neo4j Java Driver (version 1.0.6), and in R[12] (version 3.4.2) via RNeo4j.[13] Selections within the database were done by filtering EntityWithAccessionedSequence, Reaction, Pathway, and TopLevelPathway on speciesName: "*Homo sapiens*", and ReferenceEntity on databaseName: "UniProt". The networks presented in this manuscript can be created using PathwayMatcher (github.com/PathwayAnalysisPlatform/PathwayMatcher). The code used

Received: June 15, 2018
Published: September 25, 2018

DOI: 10.1021/acs.jproteome.8b00464
J. Proteome Res. 2018, 17, 3801−3809

Burger et al.: J Proteome Res. 2018 Nov 2;17(11):3801-3809

Reactome - curated and peer-reviewed pathway database



Find Reactions, Proteins and Pathways

e.g. O95631, NTN1, signaling by EGFR, glucose | Go!

Visualize and interact with Reactome biological pathways

Merges pathway identifier mapping, over-representation, and expression analysis

Designed to find pathways and network patterns related to cancer and other types of diseases

Information to browse the database and use its principal tools for data analysis

USE REACTOME GRAPH DATABASE IN YOUR PROJECT

Why Reactome

Reactome is a free, open-source, curated and peer-reviewed pathway database. Our goal is to provide intuitive bioinformatics tools for the visualization, interpretation and analysis of pathway knowledge to support basic research, genome analysis, modeling, systems biology and education.

If you use Reactome in Asia, we suggest using our Chinese mirror site at reactome.ncpsb.org.

EMBL-EBI NYU Langone Health OHSU OICR

The development of Reactome is supported by grants from the US National Institutes of Health (P41 HG003751 and 1U54GM114833-01), CFREF Medicine by Design, and the European Molecular Biology Laboratory.

Latest News

- Download our biological pathway knowledge in a PDF book
- New tool to classify pathways into tissue types based on protein expression
- New reactions layout with nested compartments
- Version 67 Released
- Performing GSEA analysis with ReactomeFIViz app

Tweets

reactome Retweeted

ChEMBL Database
@ChEMBL

#ChEMBL is 10 years old this year :) For more about the celebrations see:chembl.blogspot.com/2019/02/chembl…

19 Feb 2019

reactome Retweeted

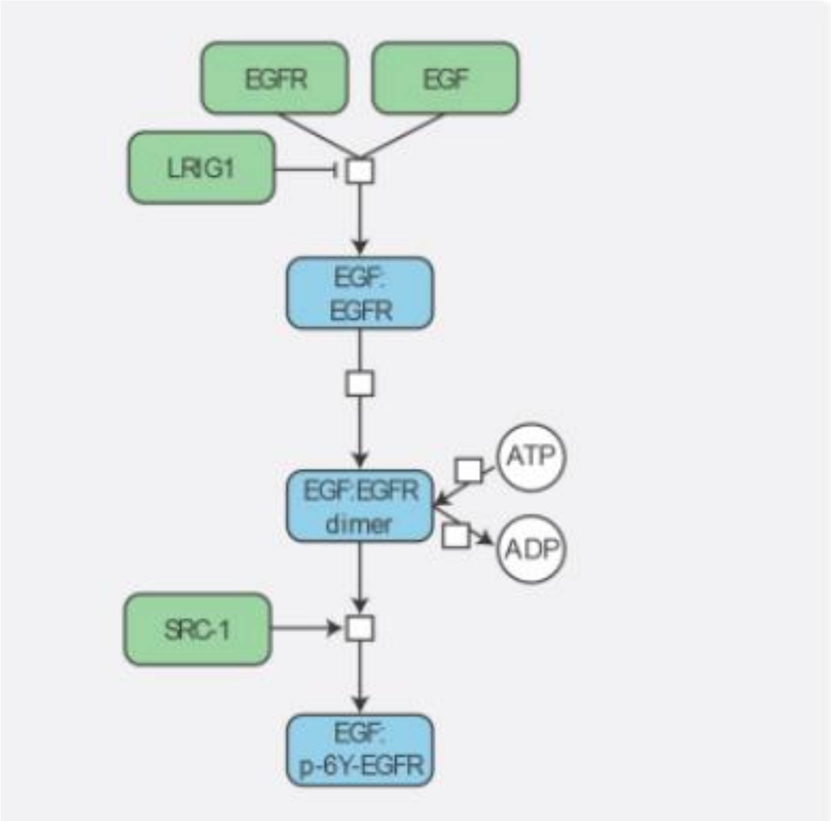
EMBL-EBI Training
@EBItraining

Finding out what your #omics data means can be tricky but #networkanalysis can help with that. Find out how and get hands on experience with @intact_project @complexportal @reactome @cytoscape and more in this course. Applications close 15 Feb 2019. ow.ly/9hEH30nwSHf

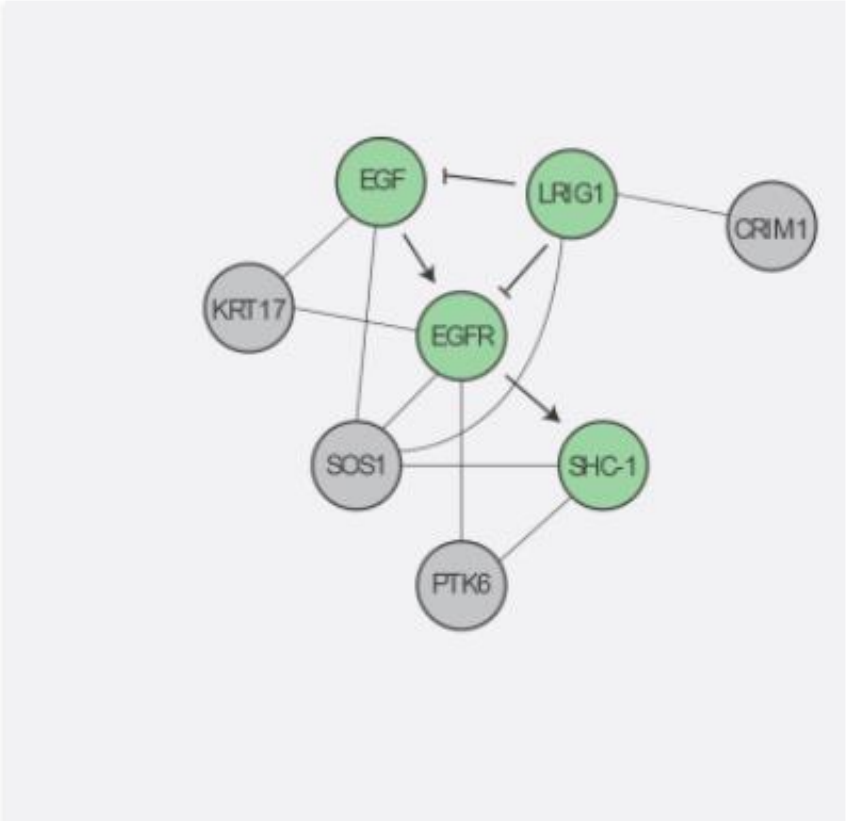
EMBL Course
Functional Insights Into Biological Data Through Network Analysis

What are interaction networks?

Pathway

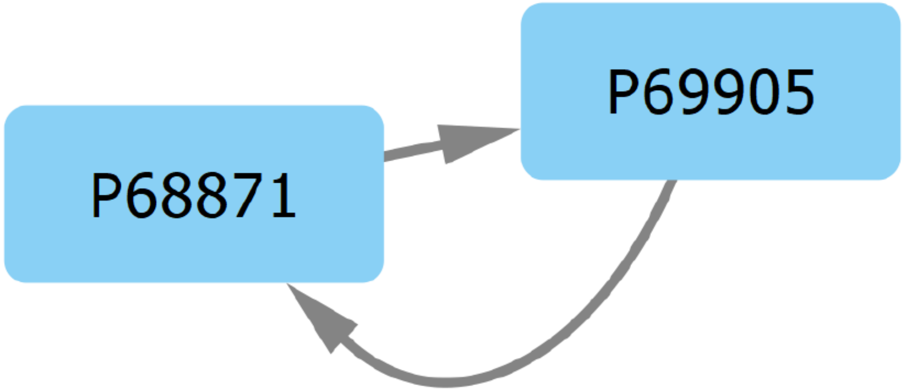


Network

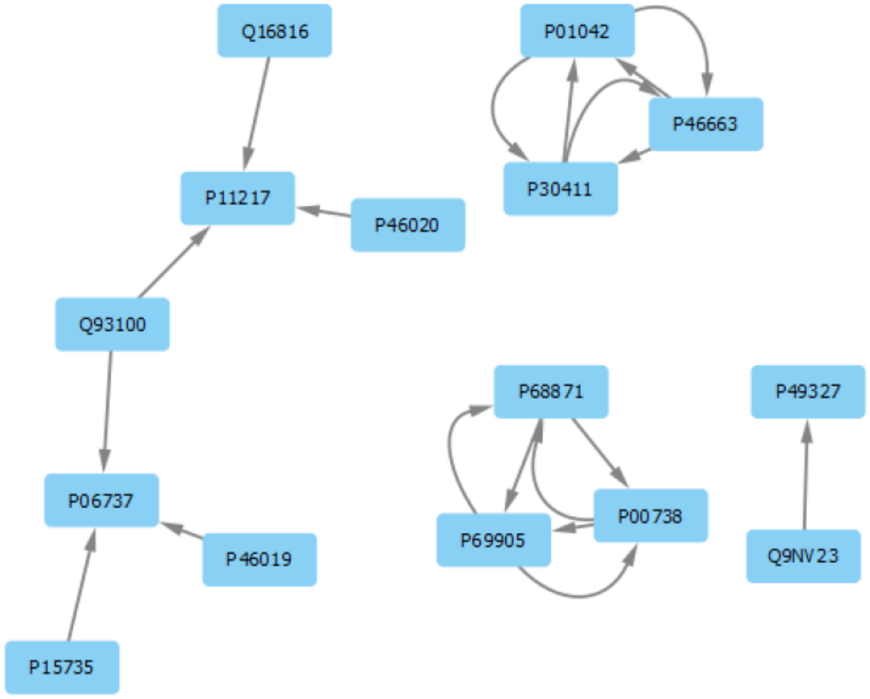


reactome.org

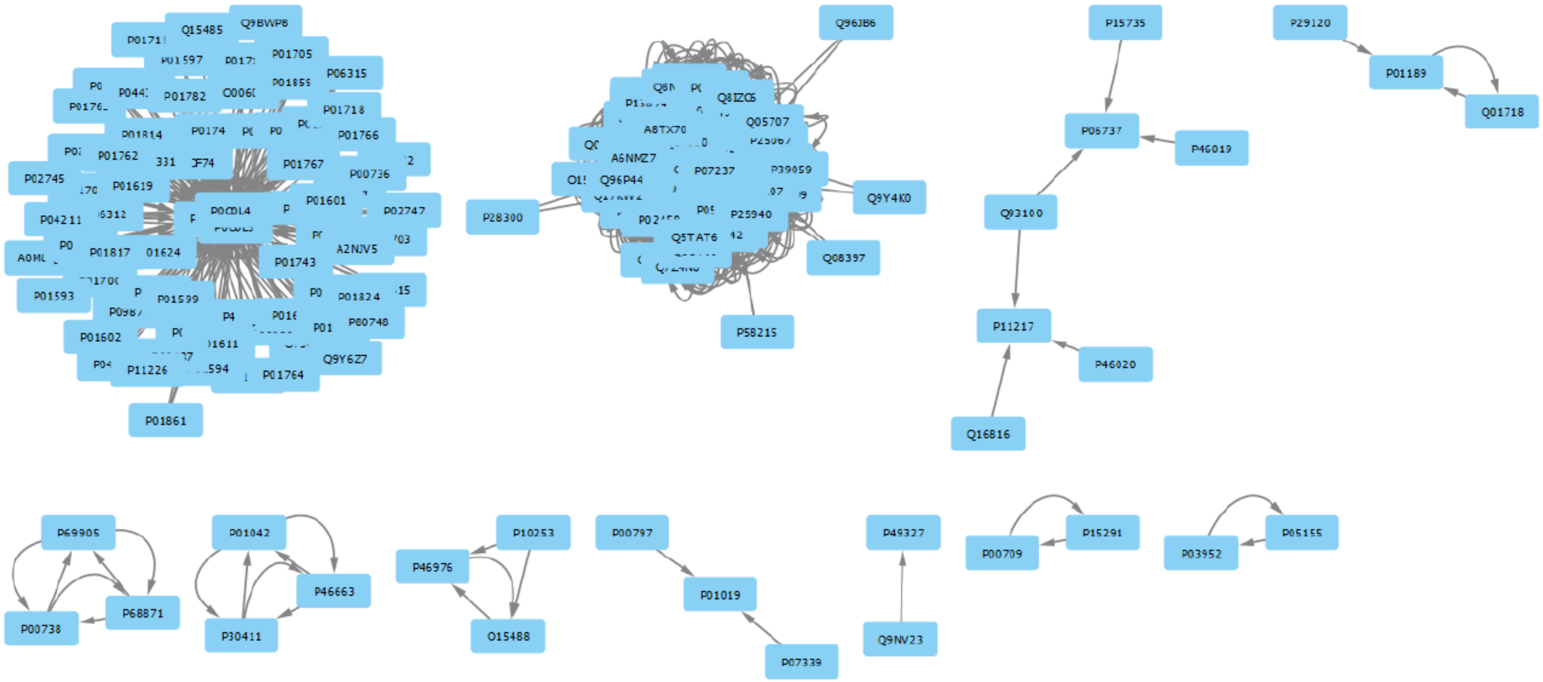
Pathways in 1934



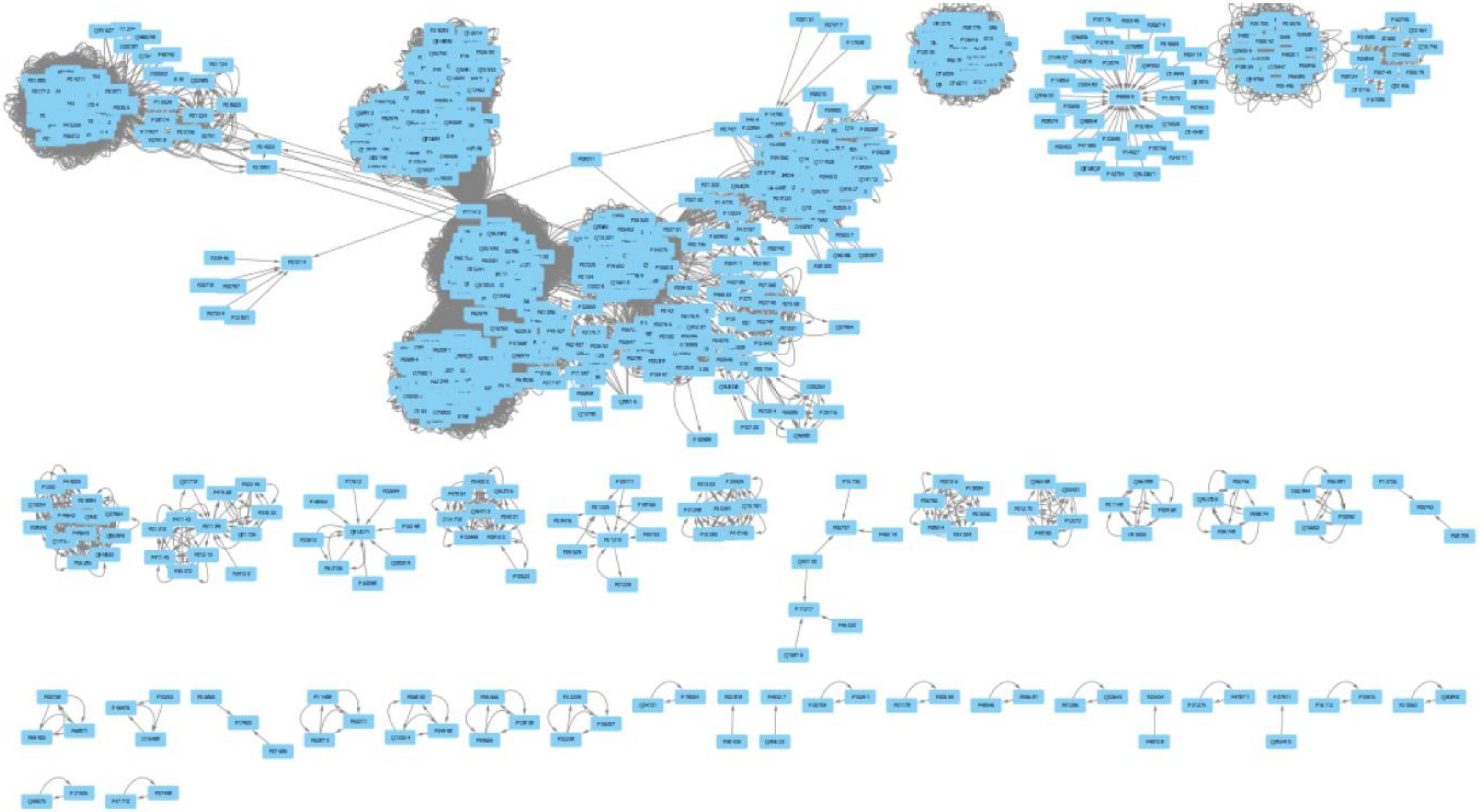
Pathways in 1960



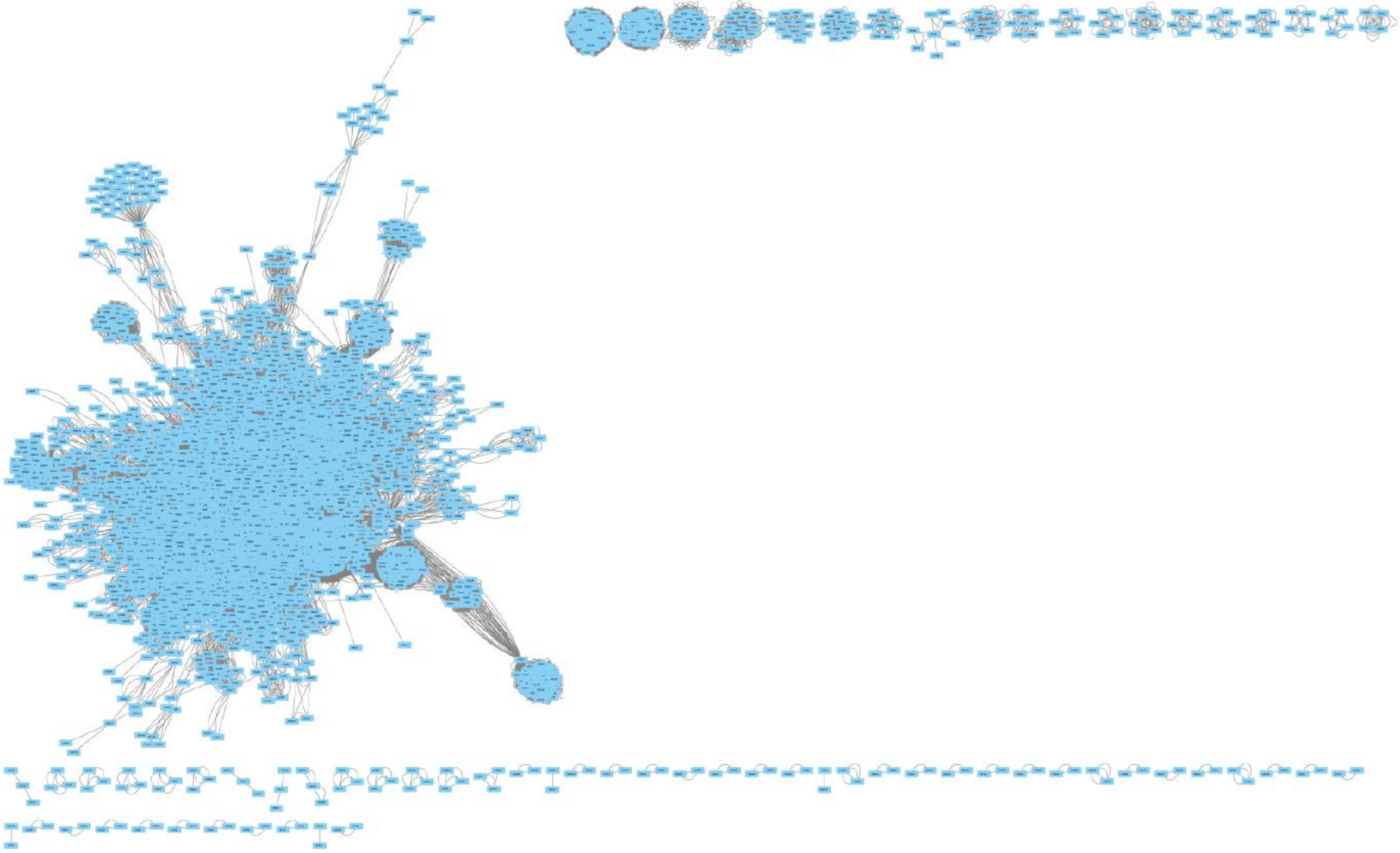
Pathways in 1970



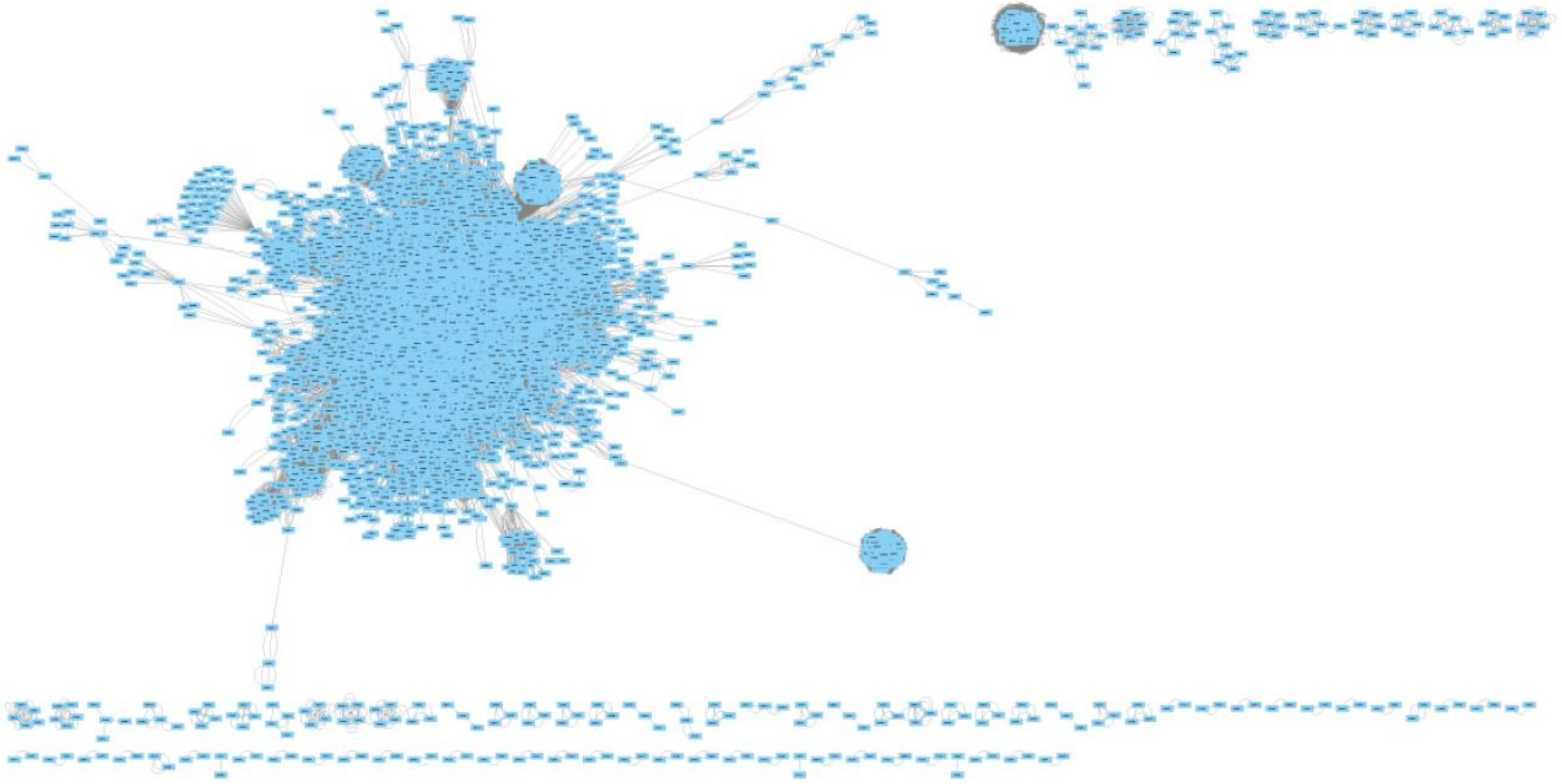
Pathways in 1985



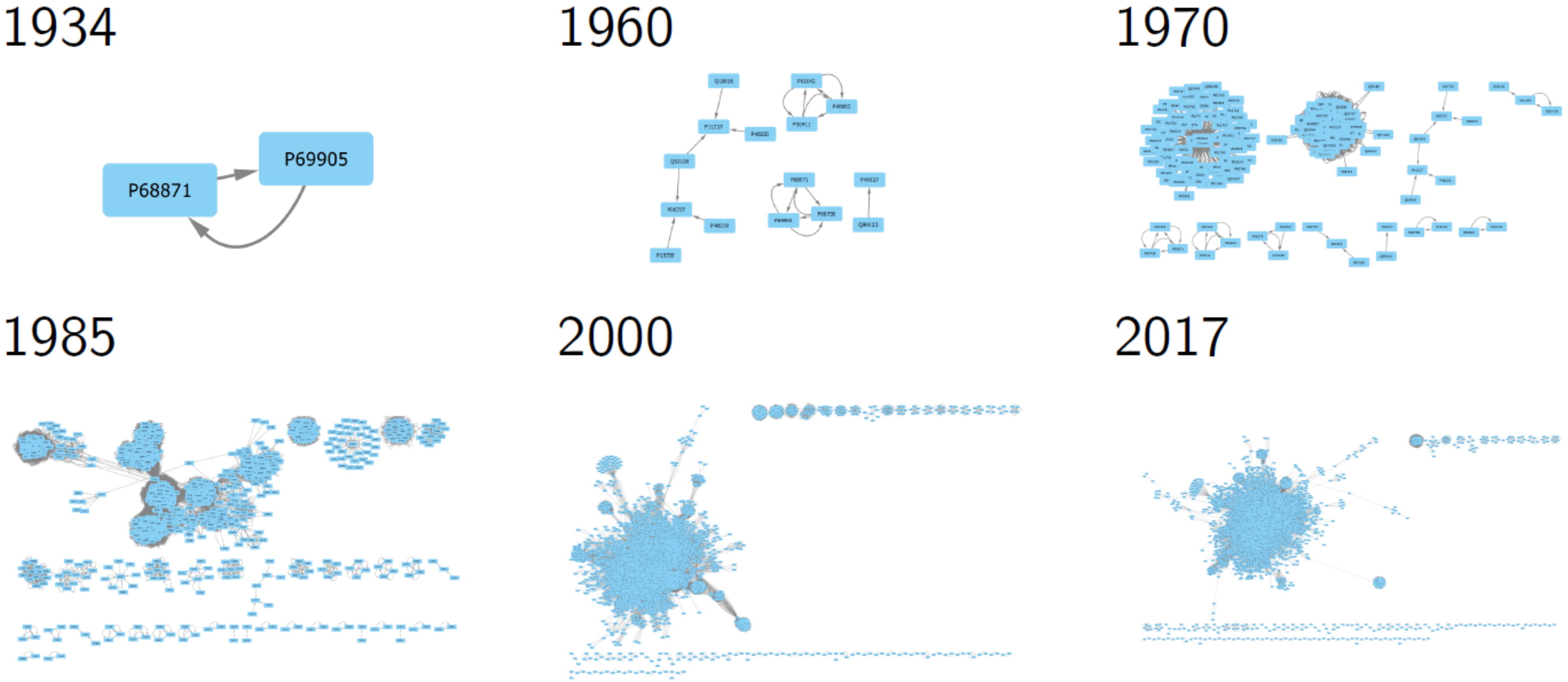
Pathways in 2000



Pathways in 2017



Pathways 1934 - 2017



1934

1960

1970

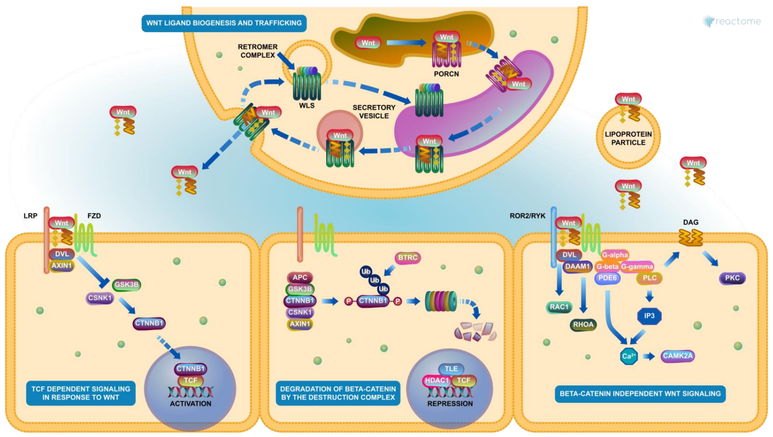
1985

2000

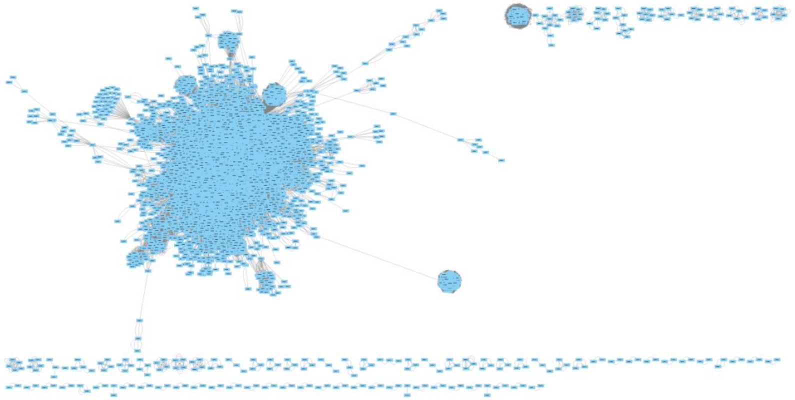
2017

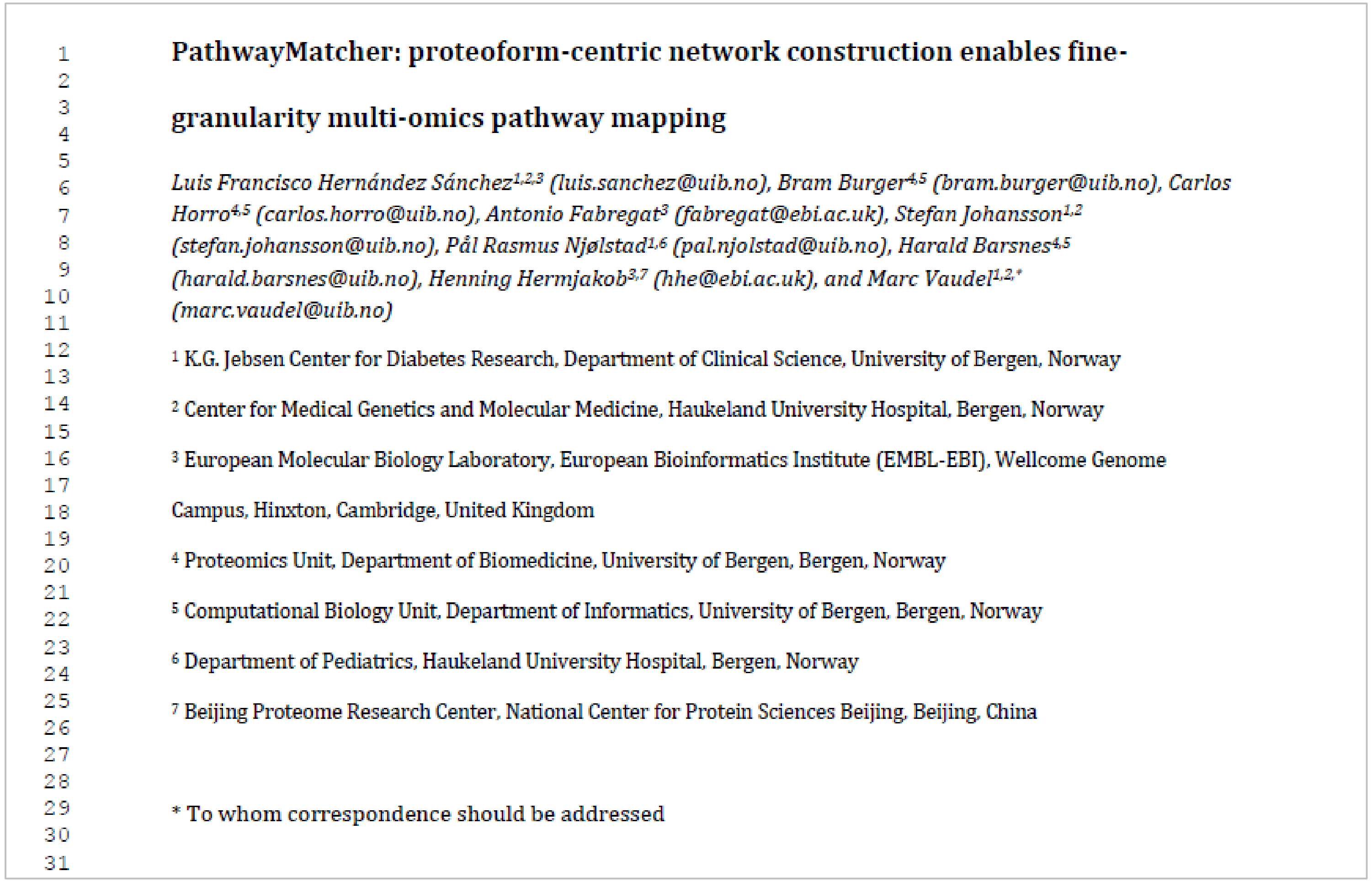
Burger et al.: J Proteome Res. 2018 Nov 2;17(11):3801-3809

What are pathways?



vs





PathwayMatcher: proteoform-centric network construction enables fine-granularity multi-omics pathway mapping

Luis Francisco Hernández Sánchez[1,2,3] (luis.sanchez@uib.no), Bram Burger[4,5] (bram.burger@uib.no), Carlos Horro[4,5] (carlos.horro@uib.no), Antonio Fabregat[3] (fabregat@ebi.ac.uk), Stefan Johansson[1,2] (stefan.johansson@uib.no), Pål Rasmus Njølstad[1,6] (pal.njolstad@uib.no), Harald Barsnes[4,5] (harald.barsnes@uib.no), Henning Hermjakob[3,7] (hhe@ebi.ac.uk), and Marc Vaudel[1,2,] (marc.vaudel@uib.no)*

[1] K.G. Jebsen Center for Diabetes Research, Department of Clinical Science, University of Bergen, Norway

[2] Center for Medical Genetics and Molecular Medicine, Haukeland University Hospital, Bergen, Norway

[3] European Molecular Biology Laboratory, European Bioinformatics Institute (EMBL-EBI), Wellcome Genome Campus, Hinxton, Cambridge, United Kingdom

[4] Proteomics Unit, Department of Biomedicine, University of Bergen, Bergen, Norway

[5] Computational Biology Unit, Department of Informatics, University of Bergen, Bergen, Norway

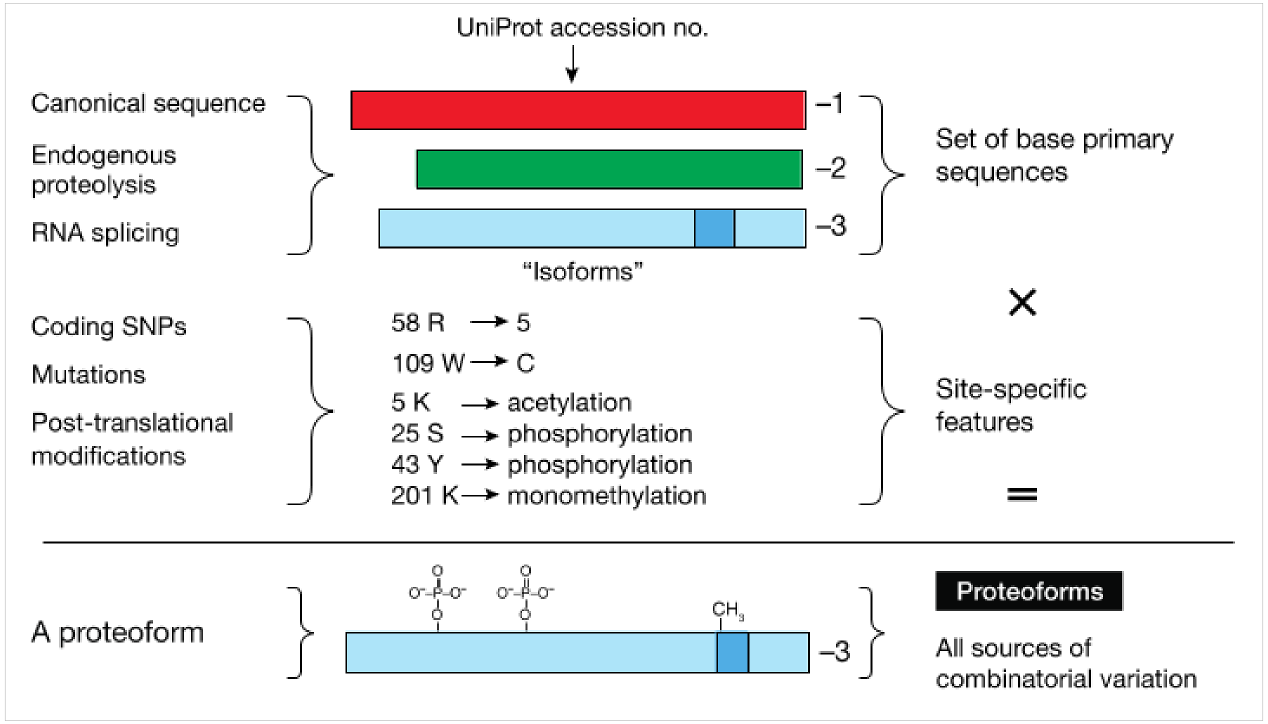
[6] Department of Pediatrics, Haukeland University Hospital, Bergen, Norway

[7] Beijing Proteome Research Center, National Center for Protein Sciences Beijing, Beijing, China

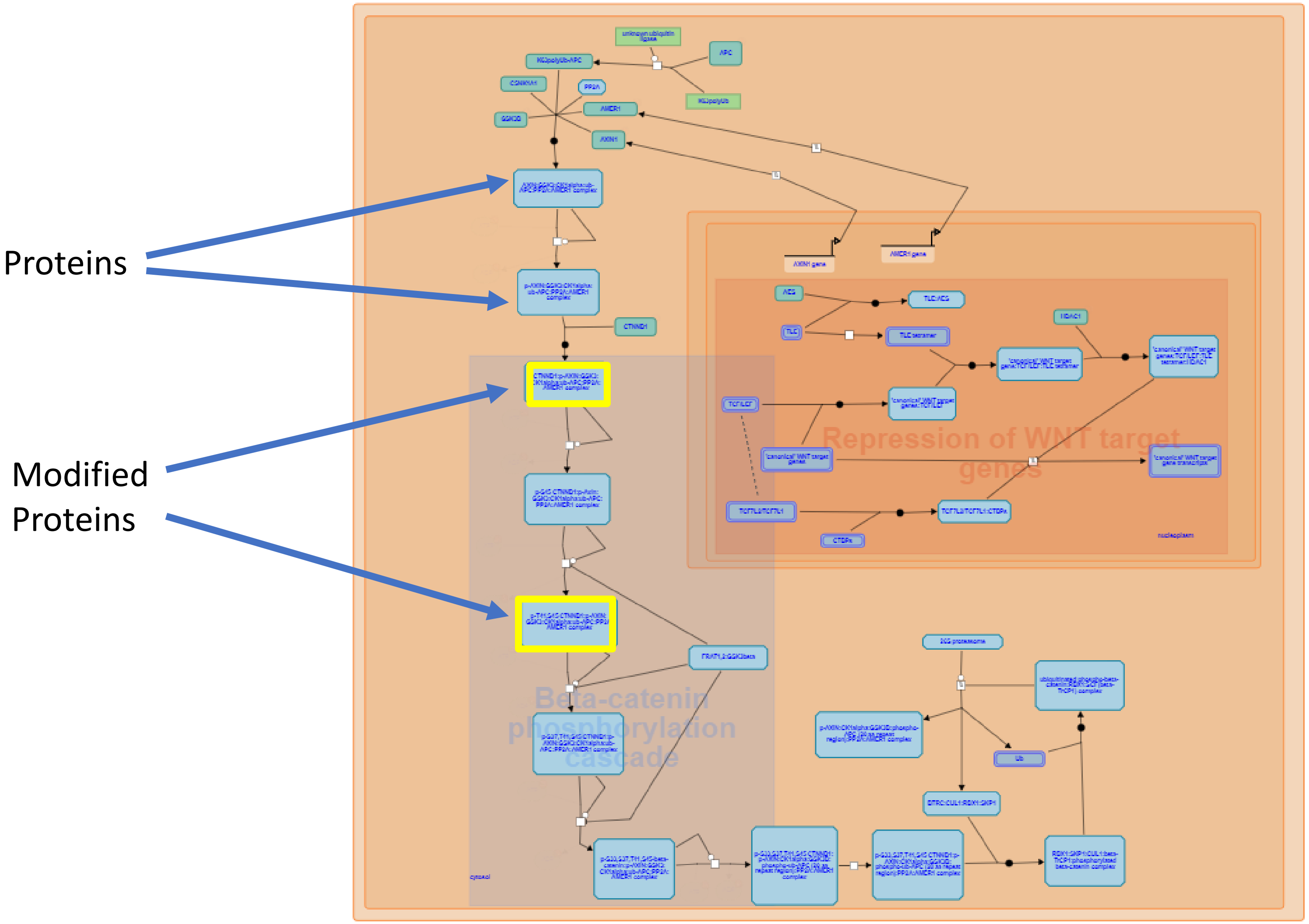
* To whom correspondence should be addressed

Hernandez Sanchez, Burger et al. (in review)

The proteoform-centric paradigm



Pathways have post-translational modifications

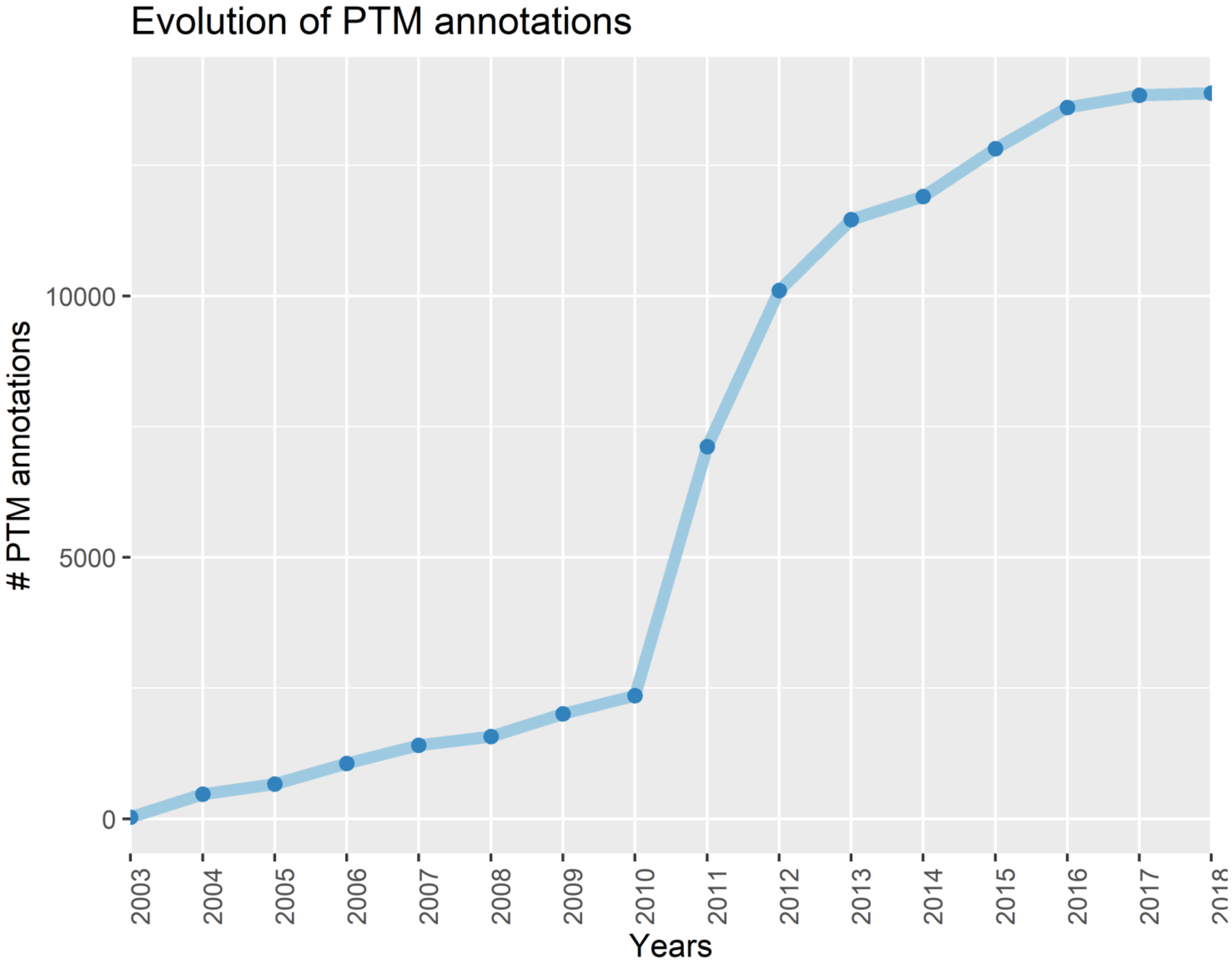


Proteins

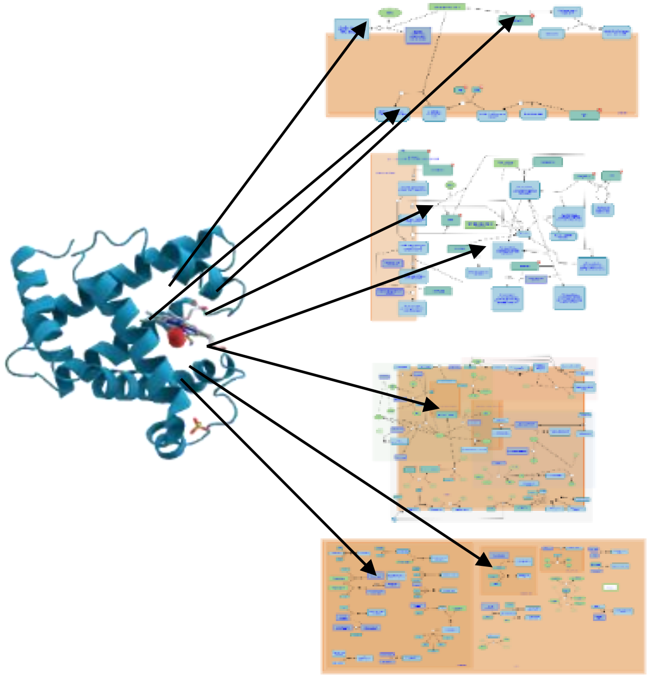
Modified Proteins

reactome.org

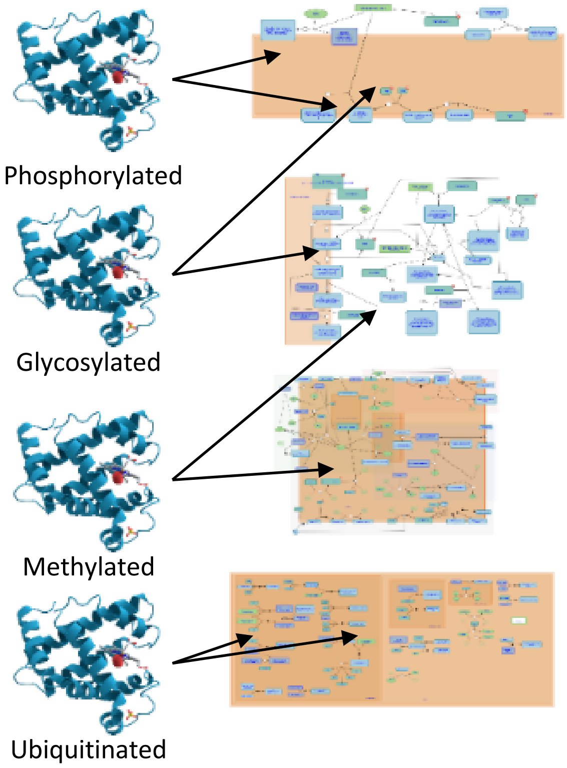
How many post-translational modifications?



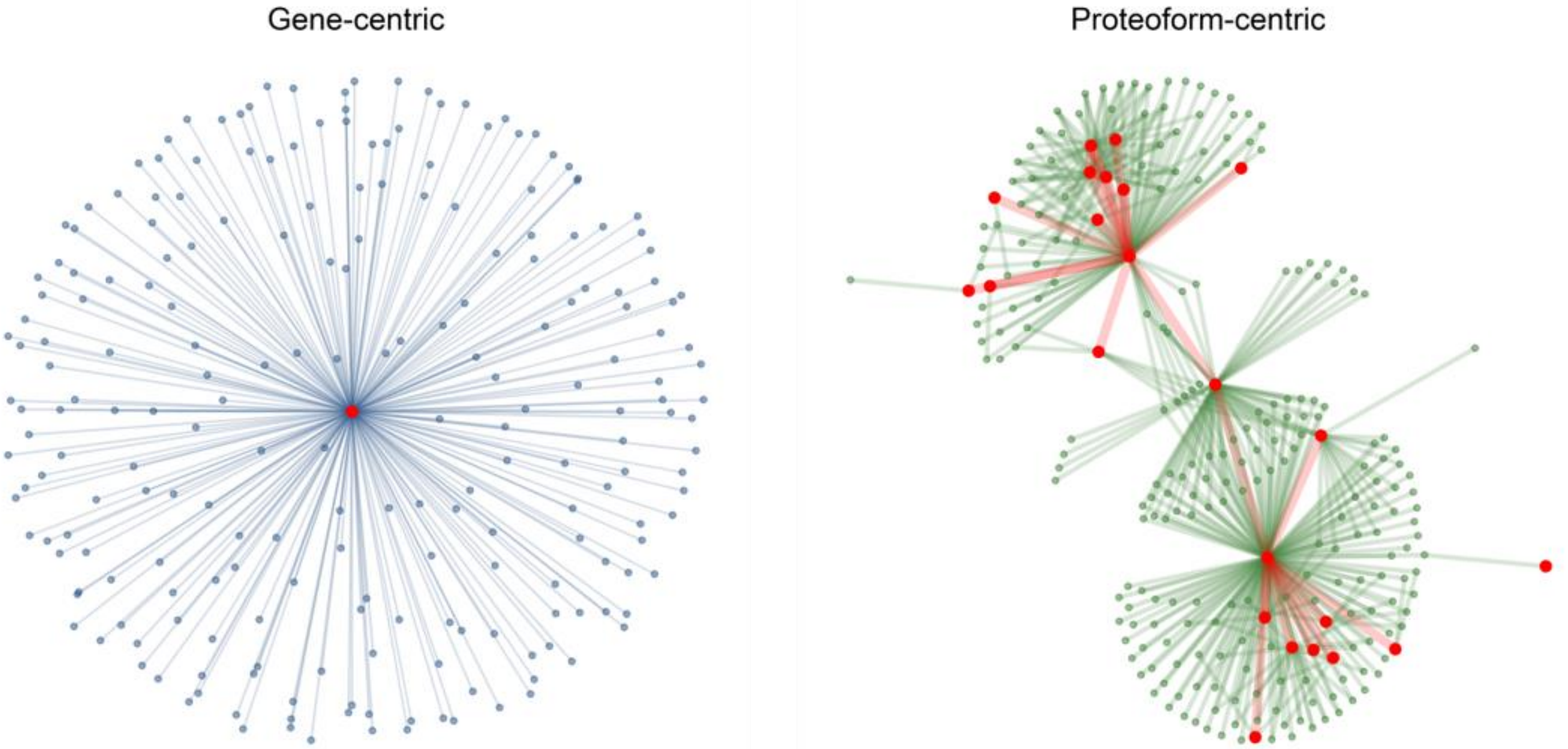
Proteoforms can differentiate the different functions of a protein



VS

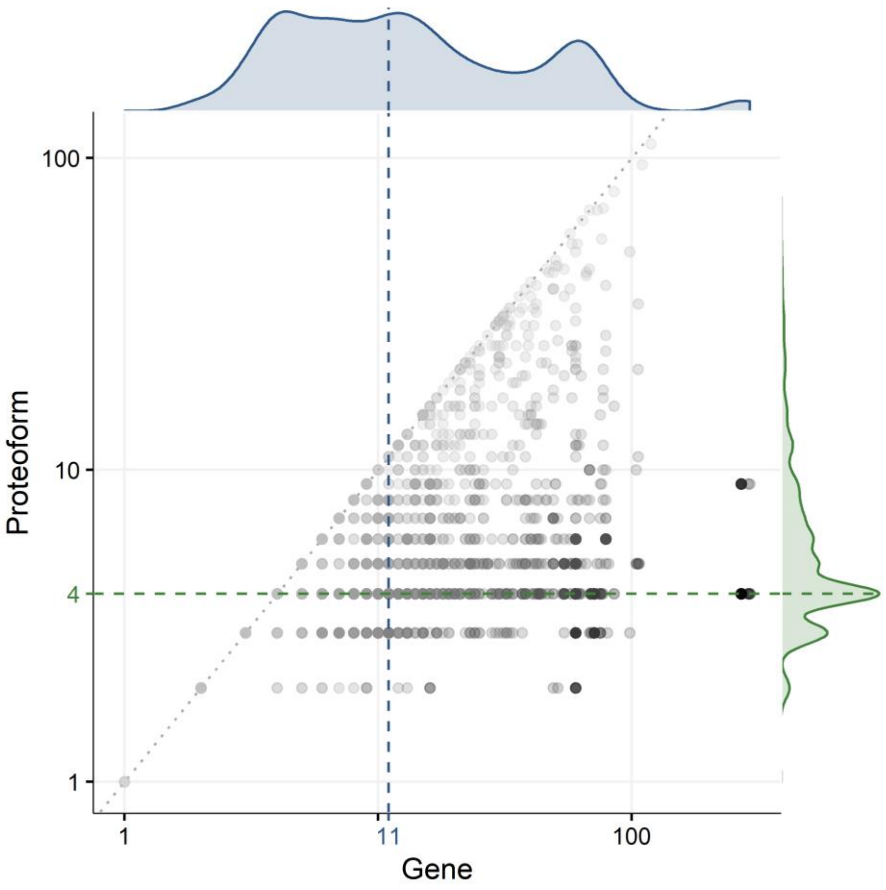


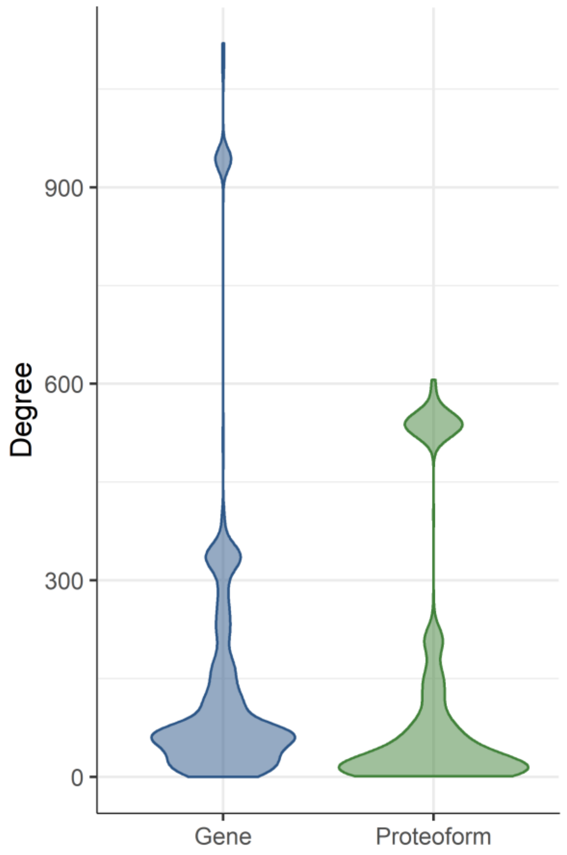
Proteoform-centric interaction networks show more accurate topology



Hernandez Sanchez, Burger et al. (in review)

Proteoforms provide more specific pathway results





Hernandez Sanchez, Burger et al. (in review)

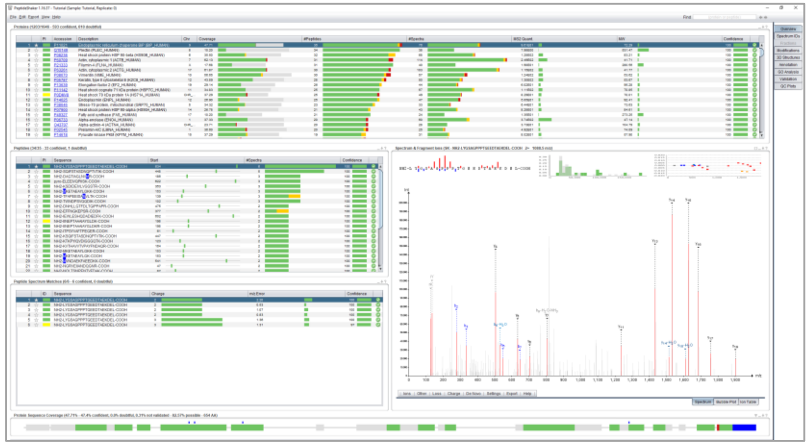
Try it yourself: PathwayMatcher!

- Proteoform-centric command line tool:
  - Pathway search and over-representation analysis
  - Export protein and proteoform interaction networks

GitHub	github.com/PathwayAnalysisPlatform/PathwayMatcher
docker	hub.docker.com/r/lfhs/pathwaymatcher
BIOCONDA	anaconda.org/bioconda/pathwaymatcher

Hernandez Sanchez, Burger et al. (in review)

Proteomics data + Proteoform networks..?



PeptideShaker - Tutorial (Sample: Tutorial, Replicate: 0)

File Edit Export View Help

Proteins

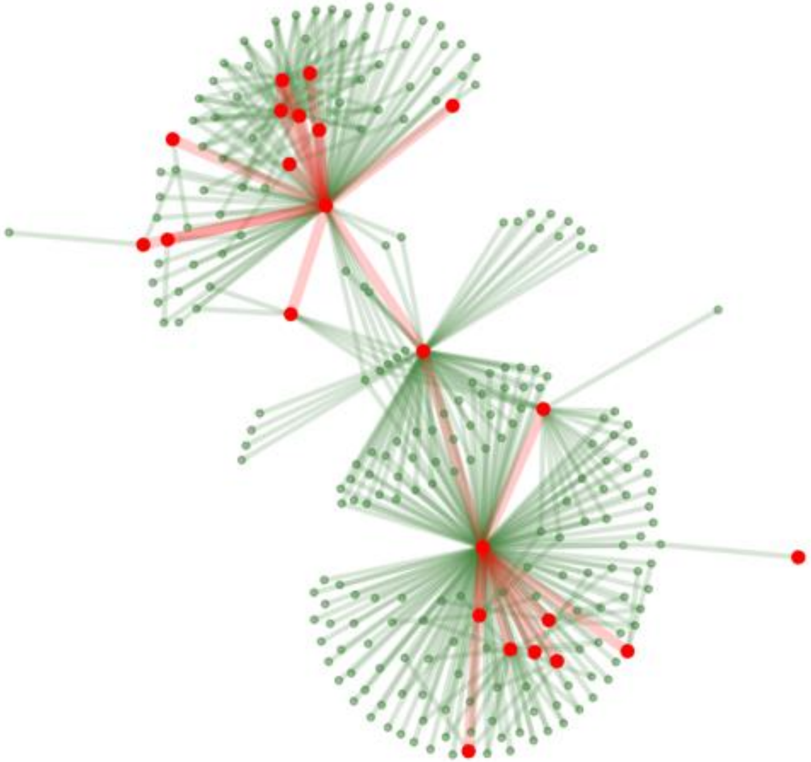
	Accession	Description
1	P11021	Endoplasmic reticulum chaperone BiP (BIP_HUMAN)
2	Q15149	Plectin (PLEC_HUMAN)
3	P08238	Heat shock protein HSP 90-beta (HS90B_HUMAN)
4	P60709	Actin, cytoplasmic 1 (ACTB_HUMAN)
5	P21333	Filamin-A (FLNA_HUMAN)
6	P63261	Actin, cytoplasmic 2 (ACTG_HUMAN)
7	P08670	Vimentin (VIME_HUMAN)
8	P05787	Keratin, type II cytoskeletal 8 (K2C8_HUMAN)
9	P13639	Elongation factor 2 (EF2_HUMAN)
10	P11142	Heat shock cognate 71 kDa protein (HSP7C_HUMAN)
11	P0DMV8	Heat shock 70 kDa protein 1A (HS71A_HUMAN)
12	P14625	Endoplasmin (ENPL_HUMAN)
13	P38646	Stress-70 protein, mitochondrial (GRP75_HUMAN)
14	P07900	Heat shock protein HSP 90-alpha (HS90A_HUMAN)
15	P49327	Fatty acid synthase (FAS_HUMAN)
16	P06733	Alpha-enolase (ENOA_HUMAN)
17	O43707	Alpha-actinin-4 (ACTN4_HUMAN)
18	P02545	Prelamin-A/C (LMNA_HUMAN)
19	P14618	Pyruvate kinase PKM (KPYM_HUMAN)

Peptides

Peptide Spectrum Matches

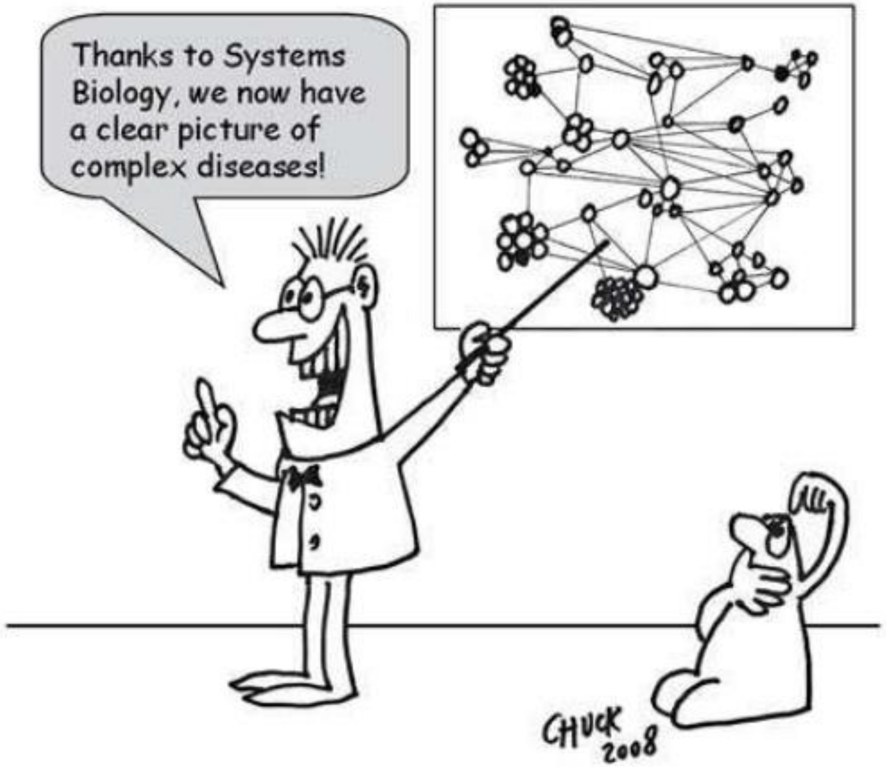
Spectrum & Fragment Ions

+



= ?

Pathways to heaven..?


Thanks to Systems Biology, we now have a clear picture of complex diseases!
CHUCK 2008

HOW
WHY
HOW
WHEN
WHERE
WHO
WHERE
WHERE
WHY
?
WHO
WHO
WHEN
WHAT
WHAT
WHY
WHAT
WHEN
HOW
WHAT
WHEN
WHO
WHERE
WHY
HOW