sbv IMPROVER™

SYSTEMS BIOLOGY VERIFICATION

www.sbvimprover.com

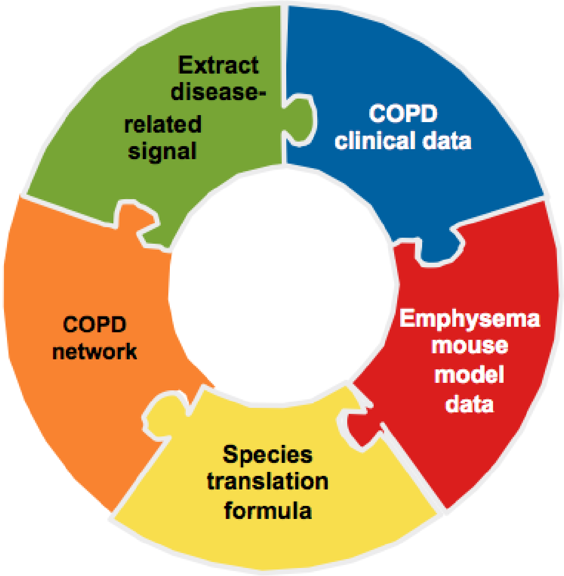
Verification of Systems Biology Research in the Age of Collaborative-Competition

Erhan Bilal, PhD
IBM Research, NY USA

Precision Medicine
October 11th 2013
New York, USA

Outline

- IMPROVER at a glance
- Crowdsourcing
- Diagnostic Signature Challenge
- Species Translation Challenge
- Network Verification Challenge
- Grand Challenge

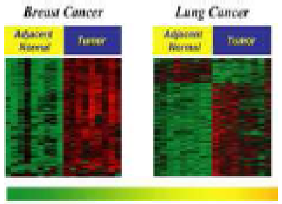


Why do we need IMPROVER?

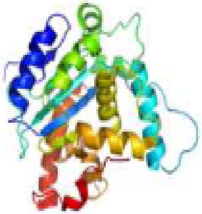
We are experiencing a data deluge…

Genomic

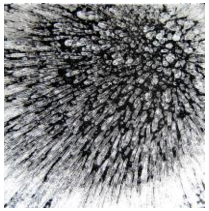
Literature



Molecular Profiles

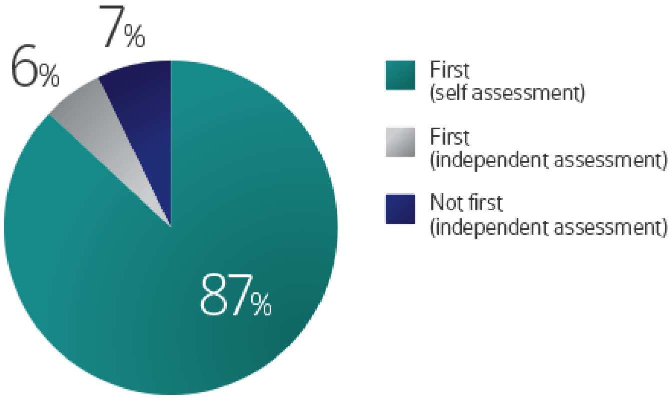


Structures



Explosion of data

But we lack the corresponding validation tools…

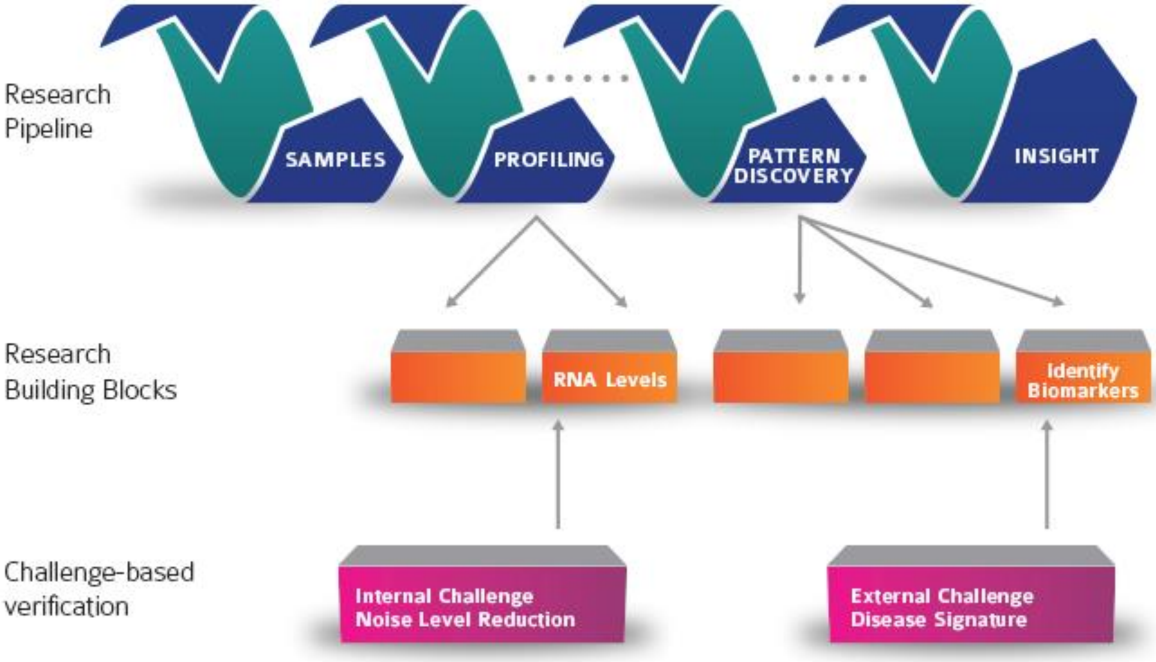


The self-assessment trap: can we all be better than average?

Molecular Systems Biology **7**: 537; published online 11 October 2011; doi:10.1038/msb.2011.70

Develop a robust methodology that verifies systems biology-based approaches

IMPROVER: Industrial Methodology for Process Verification in Research

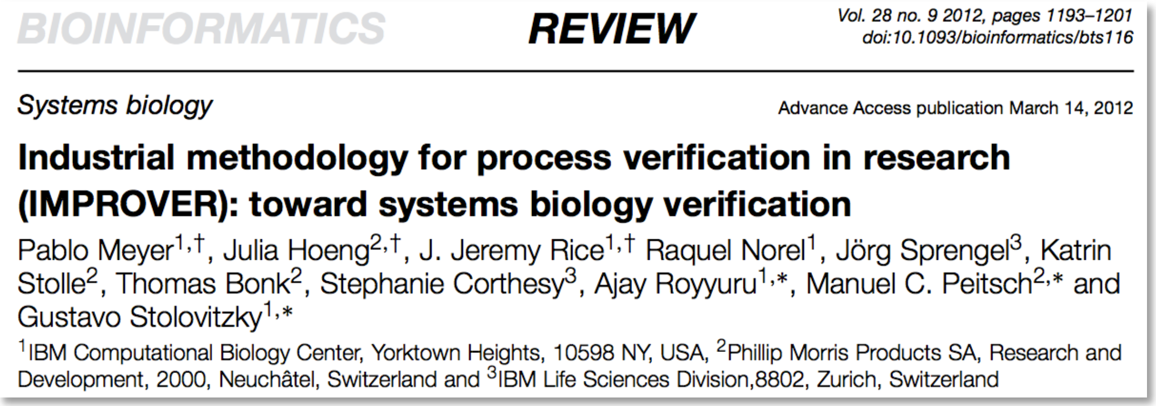


Aims to provide a measure of quality control in R&D by identifying the building blocks that need verification in a complex industrial research pipeline

Crowdsourcing challenges with double blind performance assessment of building blocks

IBM collaborating on a project funded by Philip Morris International

sbv IMPROVER is a structured process for deconstructing and evaluating research components



BIOINFORMATICS **REVIEW** *Vol. 28 no. 9 2012, pages 1193–1201*
doi:10.1093/bioinformatics/bts116

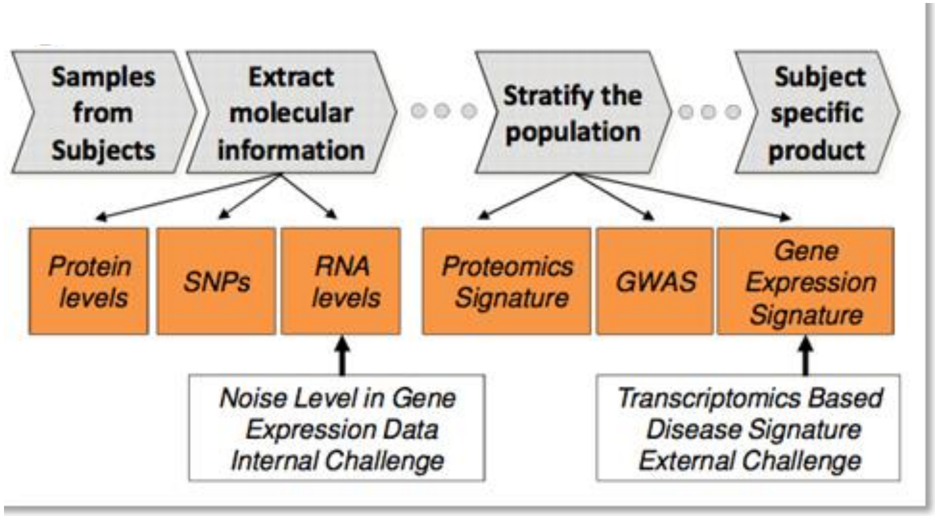
Systems biology Advance Access publication March 14, 2012

Industrial methodology for process verification in research (IMPROVER): toward systems biology verification

Pablo Meyer[1,†], Julia Hoeng[2,†], J. Jeremy Rice[1,†] Raquel Norel[1], Jörg Sprengel[3], Katrin Stolle[2], Thomas Bonk[2], Stephanie Corthesy[3], Ajay Royyuru[1,*], Manuel C. Peitsch[2,*] and Gustavo Stolovitzky[1,*]

[1]IBM Computational Biology Center, Yorktown Heights, 10598 NY, USA, [2]Phillip Morris Products SA, Research and Development, 2000, Neuchâtel, Switzerland and [3]IBM Life Sciences Division,8802, Zurich, Switzerland

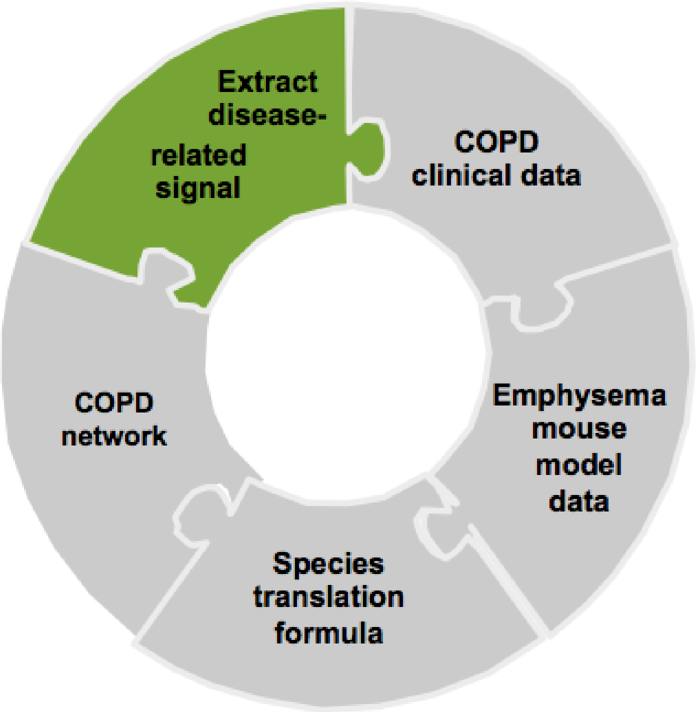
Bioinformatics 2012 28(9):1193-1201



Crowdsourcing advantages

- Many contributors with independent methods / knowledge
- Different solutions tackle various aspects of a complex problem
- The combination of solutions often outperforms the best performing submissions and is extremely robust → "Wisdom of Crowds"
- Nucleates a community around a given scientific problem
- Allows for unbiased benchmarking
- Establishes state-of-the-art technology and knowledge in a field
- Complements the classical peer-review process

Diagnostic Signature Challenge (completed)



www.sbvimprover.com

Diagnostic signature challenge

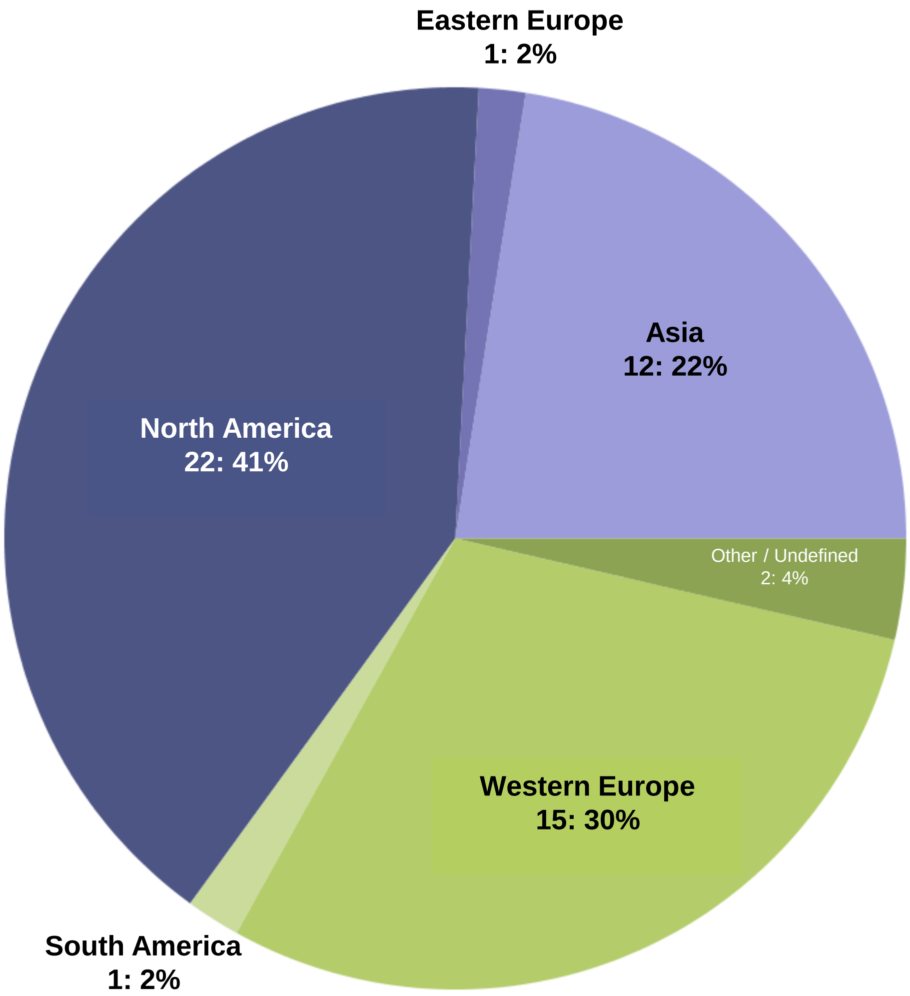
Aim to assess and verify computational approaches that classify clinical samples based on transcriptomics data.

Participants were asked to establish predictive signatures on unlabeled gene expression data sets in 4 disease areas

- Chronic Obstructive Pulmonary Disease
- Multiple Sclerosis
- Lung Cancer
- Psoriasis

Diagnostic Signature Challenge: overall participation

54 Teams from around the world participated



Diagnostic Signature Challenge participation

Submissions were spread evenly across all five sub-challenges:

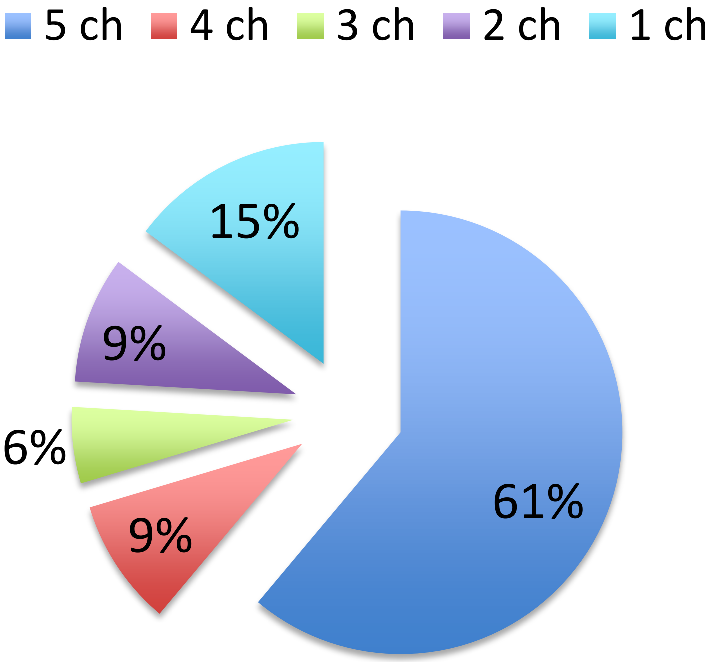
Psoriasis: 49 participants

COPD: 40 participants

Lung Cancer: 46 participants

MS Diagnosis: 40 participants

MSS Staging: 39 participants

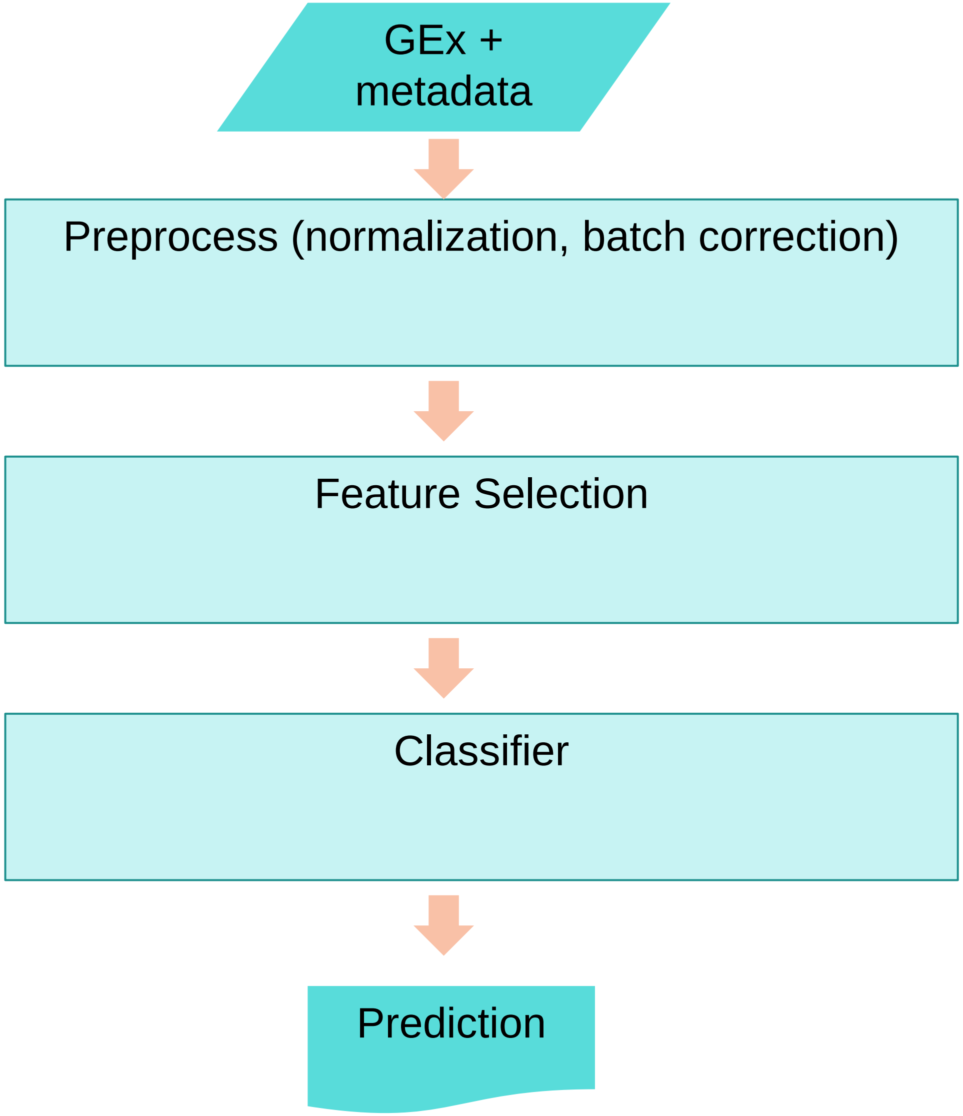


Most teams submitted predictions to all challenges (34/54)

Challenge structure

	sbv IMPROVER project
Project Type	Competition
Classification type	Confidence levels: 2-way + 4-way
Prediction type	Diagnosis, staging
Training Datasets	Public
Test Datasets	Created for or licensed to the sbv IMPROVER project Completely independent to the training datasets
Gain for community	• Available datasets can be used for benchmarking • Determine the existence of a robust signature for a particular disease/data set • Methods to be published in special issue of new journal Systems Biomedicine

Typical methodology



GEx + metadata

Preprocess (normalization, batch correction)

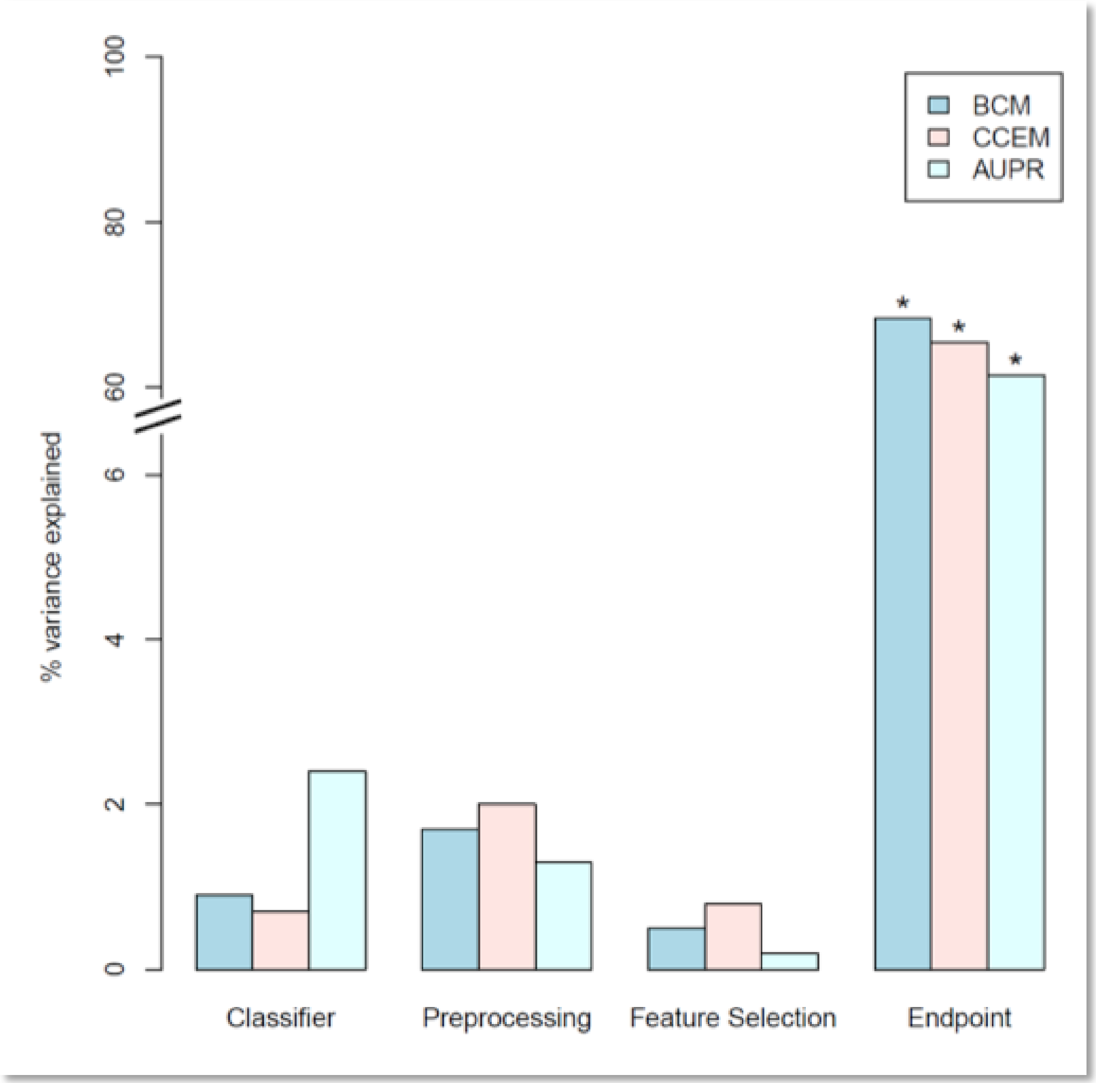
Feature Selection

Classifier

Prediction

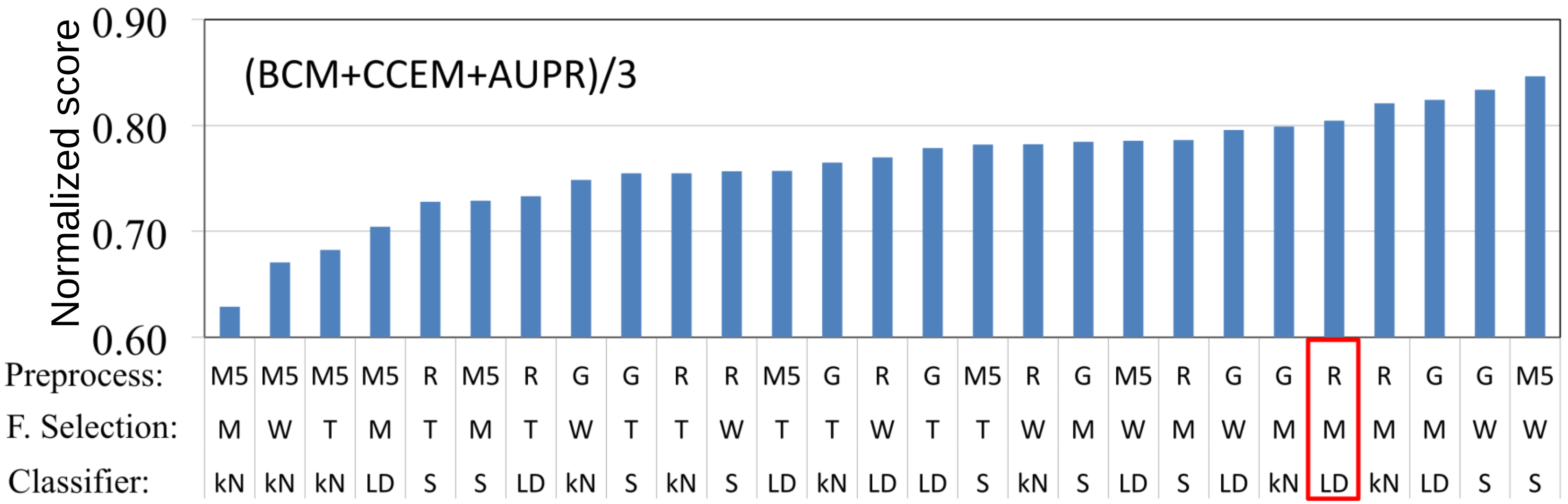
The disease endpoint was the biggest determinant of performance

sbv IMPROVER
SYSTEMS BIOLOGY VERIFICATION



Meta analysis of pipeline performance at conclusion of IMPROVER DSC

Virtual participants are created by combining different methods for data processing, feature selection and classifiers.



DSC best performer

Preprocess ->	M5: MAS5	R: RMA	G:GRMA
Feature Selection ->	M: Moderated t-test	T: regular t-test	W: Wilcoxon rank test
Classifier ->	kN: kNN	LD: LDA	S: SVM

Adapted from “Strengths and limitations of microarray-based phenotype prediction: Lessons learned from the IMPROVER Diagnostic Signature Challenge”, A. Tarca *et. al*., *submitted*

The Wisdom of Crowds for Diagnostics: aggregating predictions

Belief that subject has condition

	Team 1	Team 2	Team 3
Subject 1	0.7	0.8	0.6
Subject 2	0.5	0.7	0.8
Subject 3	0.3	0	0.1
Subject 4	0.9	0.4	0.7
….	….	….	….
Subject N-1	0.2	0.6	0.3
Subject N	1.0	0.9	0.7

Aggregate prediction by averaging beliefs

Belief aggregation

	Aggregate team
Subject 1	0.7
Subject 2	0.67
Subject 3	0.13
Subject 4	0.67
….	….
Subject N-1	0.37
Subject N	0.87

Transform into an ordered list

	Team 1	Team 2	Team 3
Subject 1	4	5	3
Subject 2	3	4	7
Subject 3	2	1	1
Subject 4	5	2	4
….	….	….	….
Subject N-1	1	3	2
Subject N	6	6	5

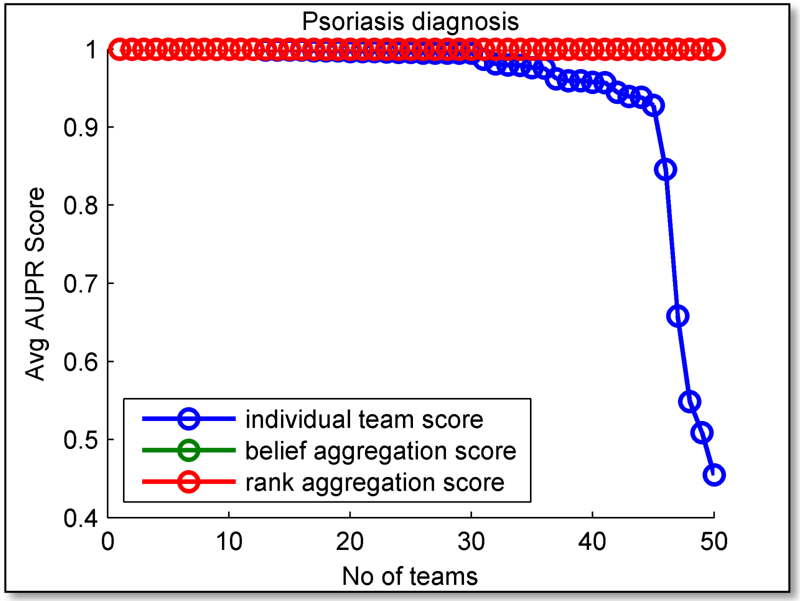
Aggregate prediction by averaging ranks

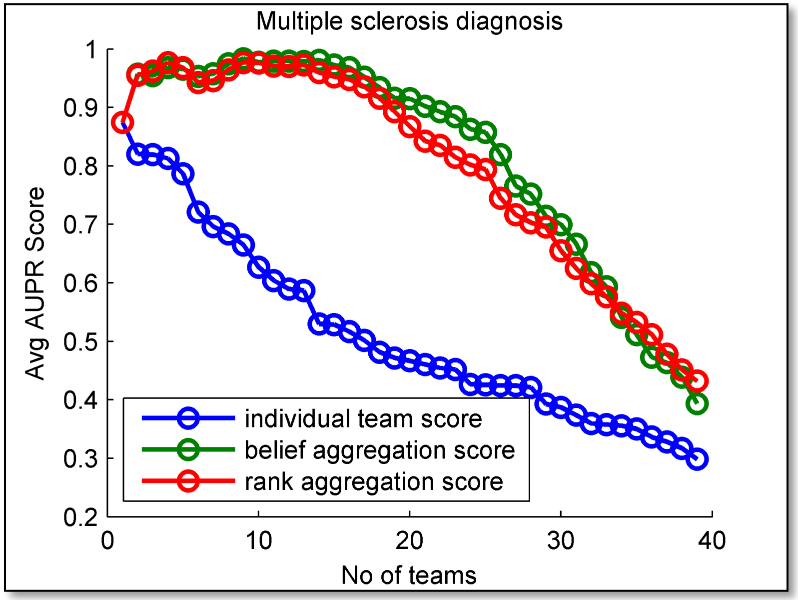
Rank aggregation

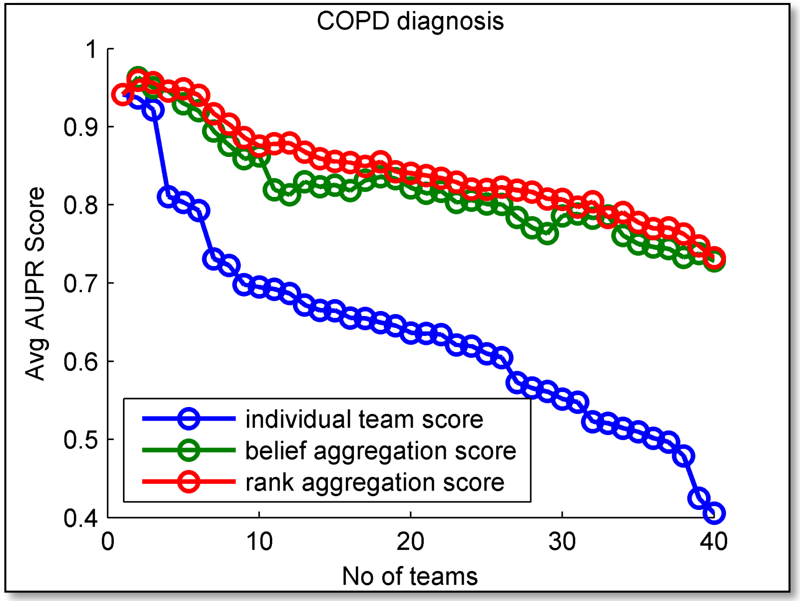
	Aggregate team
Subject 1	4
Subject 2	4.67
Subject 3	1.33
Subject 4	3.67
….	….
Subject N-1	2
Subject N	5.67

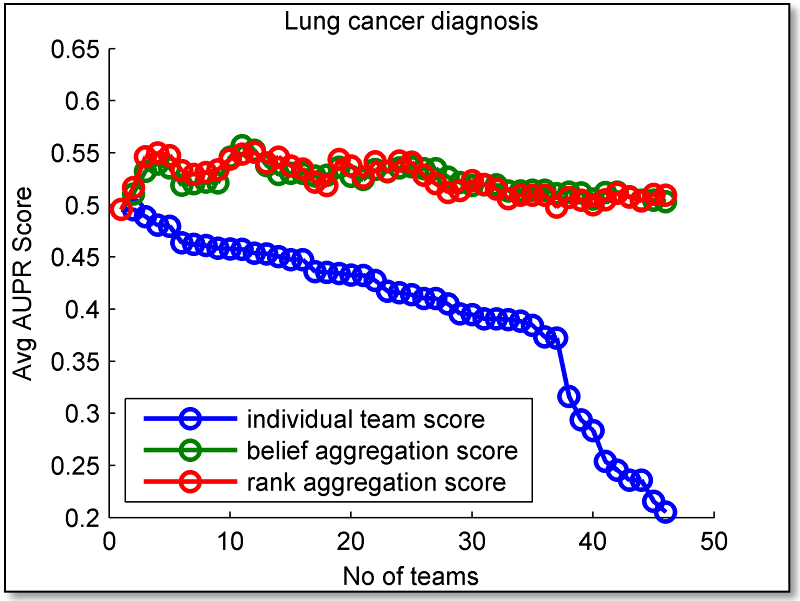
Rescore the aggregate predictions

Aggregation of Methods: Wisdom of Crowds at work in sbv IMPROVER


Psoriasis diagnosis


Multiple sclerosis diagnosis


COPD diagnosis


Lung cancer diagnosis

Lessons Learned from the 1st sbv IMPROVER Challenge

- The ability of computational methods to perform disease classification from transcriptomics data depends on endpoint of data (phenotype)
- Design of challenge data has to avoid confounding batch effects with phenotype effects.
- It may be wise not to provide all the data on the test set, as it can provide unintended information to the participants.
- Similar computational methods can have a wide range of performance within the same challenge: no single method was the clear winner

Diagnostic Signature Challenge

Symposium 2012 (2-3 October 2012 in Boston, MA, USA)

- Announced the best performing teams
- Discussed and shared experiences on SBV IMPROVER and the Diagnostic Signature Challenge
- Keynotes Speakers from Systems Biology Community

Systems Biology Verification: Diagnostic Signature Challenge completed

Congratulations to the best performing team from the sbv IMPROVER Diagnostic Signature Challenge: Adi L. Tarca and Roberto Romero

Manuel Peitsch, Philip Morris International (left), Adi L. Tarca, Wayne State University (center), Gustavo Stolovitzky, IBM (right)

Photo by Kate Flook Photography

In October 2012, the results of the first sbv IMPROVER challenge were shared with the scientific community at a symposium in Boston, USA. Adi L. Tarca, Department of Computer Science & Center for Molecular Medicine and Genetics, Director, Bioinformatics and Computational Biology Unit, Perinatology Research Branch, NICHD, Wayne State University, received the first sbv IMPROVER research award of USD 50,000. The team also included Roberto Romero, Chief, Perinatology Research Branch, Head, Program for Perinatal Research and Obstetrics Intramural Division, NICHD, NIH. The results are planned to be published in early 2013.

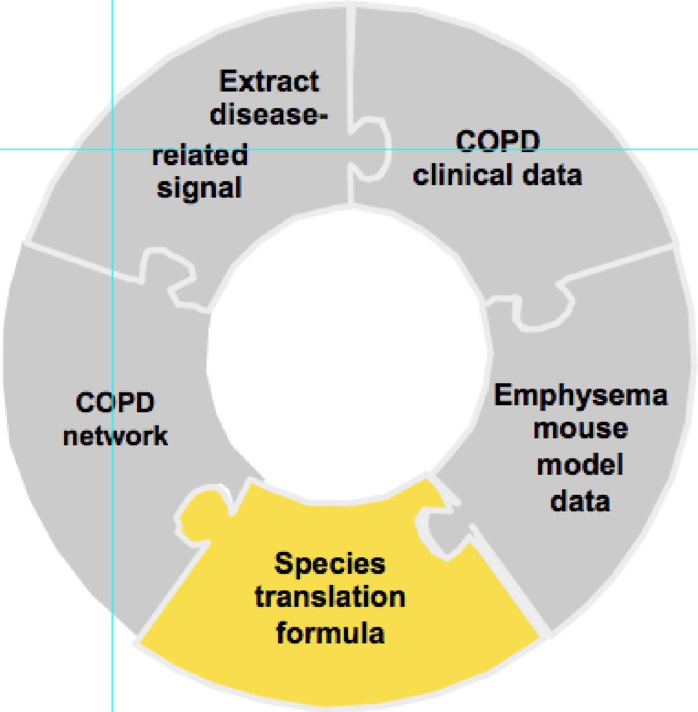
The next sbv IMPROVER challenge is planned to launch in April 2013 and will focus on Species Translation data. For more details see www.sbvimprover.com.

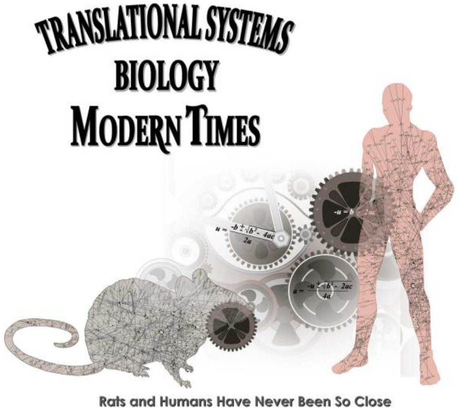
The sbv IMPROVER project and www.sbvimprover.com are part of a collaboration designed to enable scientists to learn about and contribute to the development of a new crowd sourcing method for verification of scientific data and results. The project team includes scientists from Philip Morris International's (PMI) Research and Development department and IBM's Thomas J. Watson Research Center. The project is funded by PMI.

As published in Nature, 24 Jan. 2013, page 565

Species Translation Challenge

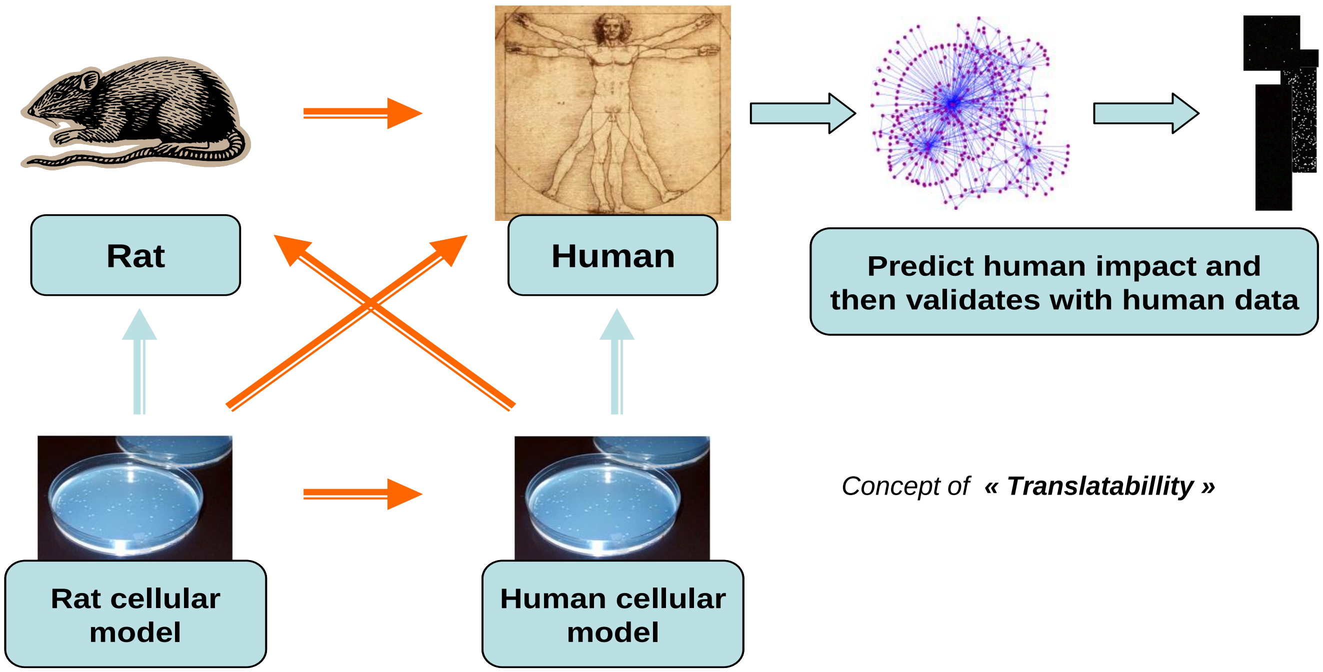
From Rat To Human: Understanding the Limits of Animal Models for Human Biology





www.sbvimprover.com

Species Translation Challenge: Background and Goal



Rat

Human

Predict human impact and then validates with human data

Rat cellular model

Human cellular model

Concept of ***« Translatabillity »***

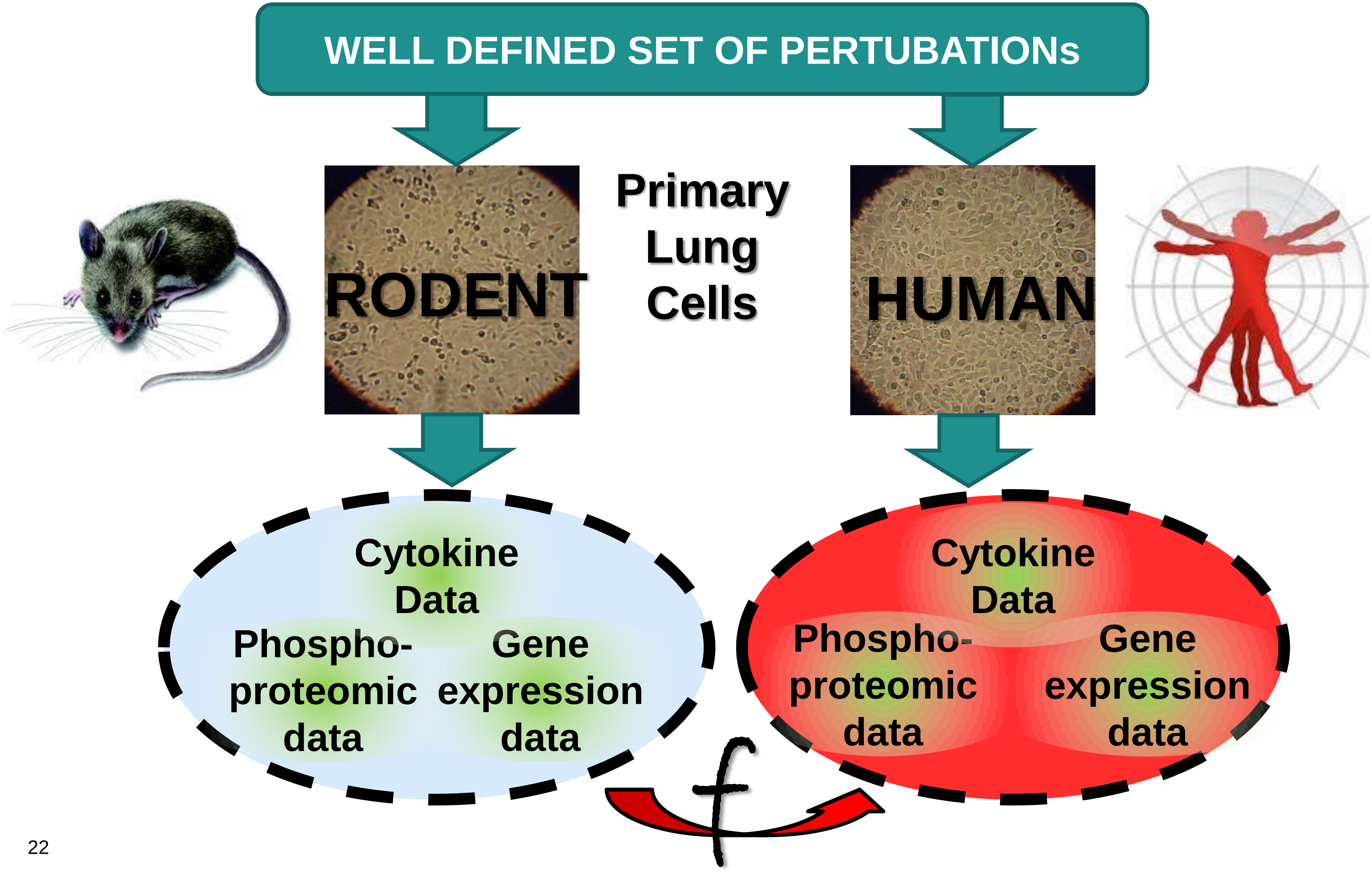
Goal: Verify the translatability of biological effects of perturbations in one species given information about the same perturbations in another species.

Scientific Questions

The **Species Translation Challenge** involves four sub-challenges that aim to shed light on important questions faced by the field:

- Can the perturbations of signaling pathways in one species predict the response to a given stimulus in another species?
- Which biological pathway functions and gene expression profiles are most robustly translated?
- Does translation depend on the nature of the stimulus or data type collected such as protein phosphorylation, gene expression and cytokine responses?
- Which computational methods are most effective for inferring gene, phosphorylation and pathway responses from one species to another?

Data production for Species Translation challenge



WELL DEFINED SET OF PERTUBATIONs

RODENT

Primary Lung Cells

HUMAN

Cytokine Data

Phospho-proteomic data

Gene expression data

Cytokine Data

Phospho-proteomic data

Gene expression data

f

Overall Experimental Workflow

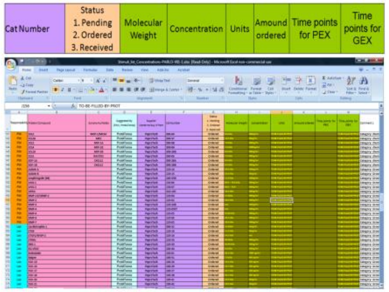
sbv IMPROVER
SYSTEMS BIOLOGY VERIFICATION

Step 1

Culture Cells
Human/Rat

Step 2

In-silico
Screening



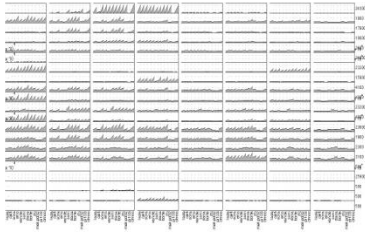
Cat Number	Status 1. Pending 2. Ordered 3. Received	Molecular Weight	Concentration	Units	Amound ordered	Time points for PEX	Time points for GEX

Step 3

Validation of
RNA & Protein
assays

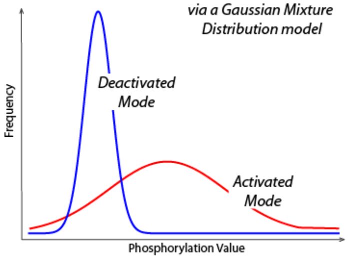
Step 4

Experimental
Screen of
Compounds



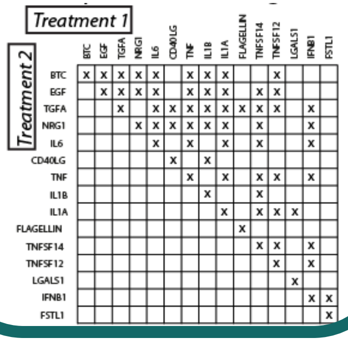
Step 5

Compound
Selection



Step 6

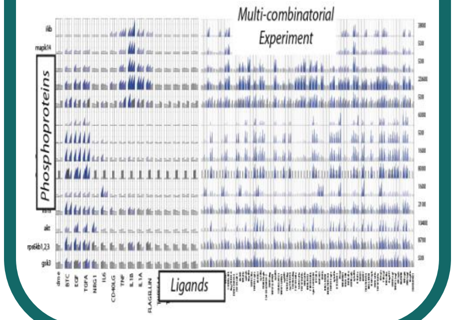
Experimental
Design



	BTC	EGF	TGFA	NRG1	IL6	CD40LG	TNF	IL1B	IL1A	FLAGELLIN	TNFSF14	TNFSF12	LGALS1	IFNB1	FSTL1
BTC	X	X	X	X	X		X	X	X			X			
EGF		X	X	X	X		X	X	X		X	X			
TGFA			X		X	X	X	X	X	X	X	X		X	
NRG1				X	X	X	X	X	X		X			X	
IL6					X		X		X		X			X	
CD40LG						X		X							
TNF							X		X		X	X		X	
IL1B								X			X				
IL1A									X		X	X	X		
FLAGELLIN										X					
TNFSF14											X	X		X	
TNFSF12												X		X	
LGALS1													X		
IFNB1														X	X
FSTL1															X

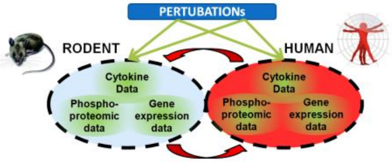
Step 7

Data
Acquisition



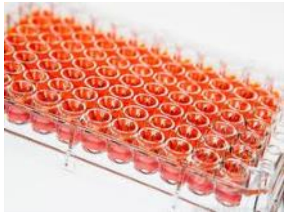
Step 8

Data Analysis
Species
Comparison

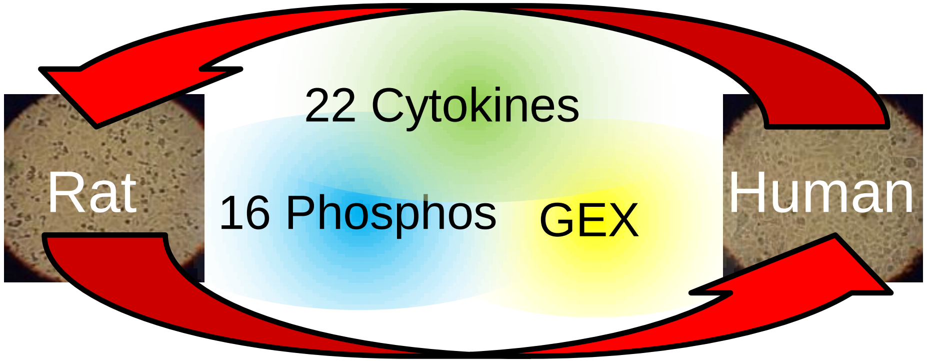


Species Translation Challenge

~100 flasks



~5000 human and
rat samples generated!!
(~50x96plates)



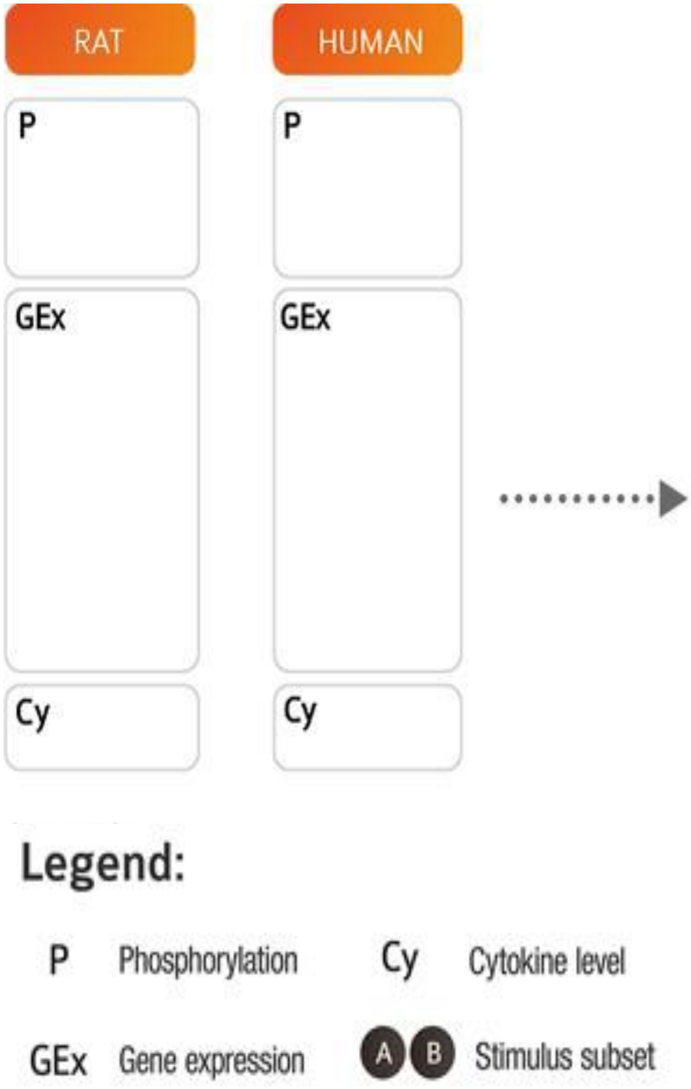
- → The <u>largest</u> multiplexed screen
- → RNA+Phospho+Cyto data
- → Novel hits

~1000 compounds
in-silico screen

Experimental
screen of
270 compounds

~ 50 selected compounds
with novel activity

Data Compendium



2 Species: human and rat

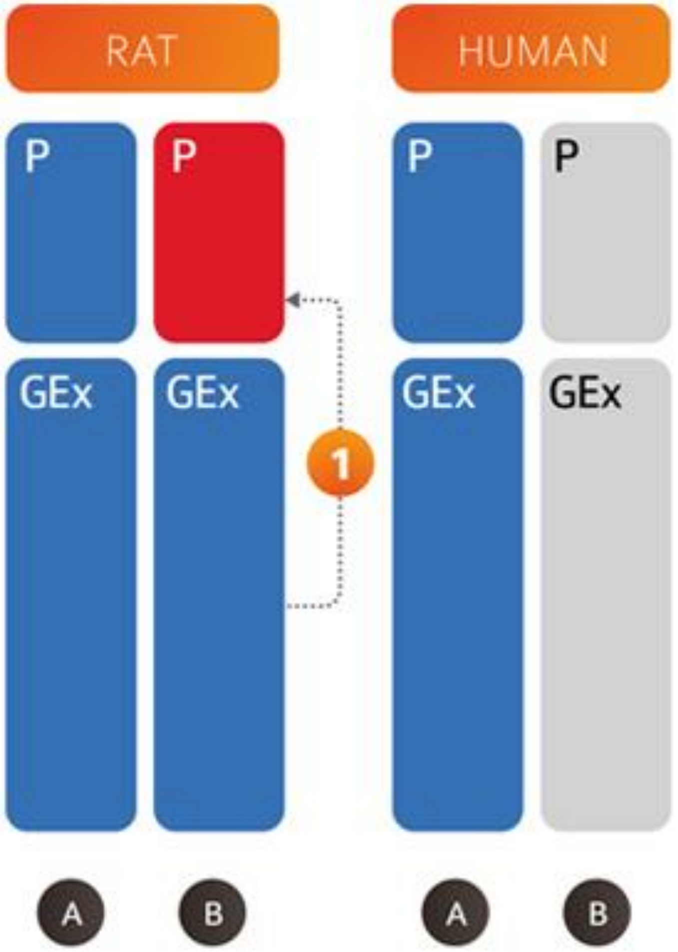
52 stimuli

Phospho-proteomics data (~10^4 data points)
16 phospho-proteins
2 time points: 5 and 25min
3 biological replicates

Gene Expression data (> 300 CEL files)
~20,000 (human) and ~19,000 (rat) genes
1 time point: 6h
3 biological replicates

Cytokine level data (~3,000 data points)
22 cytokines
1 time point: 24h
3 biological replicates

Sub-challenge 1
Intra-Species Protein Phosphorylation Prediction

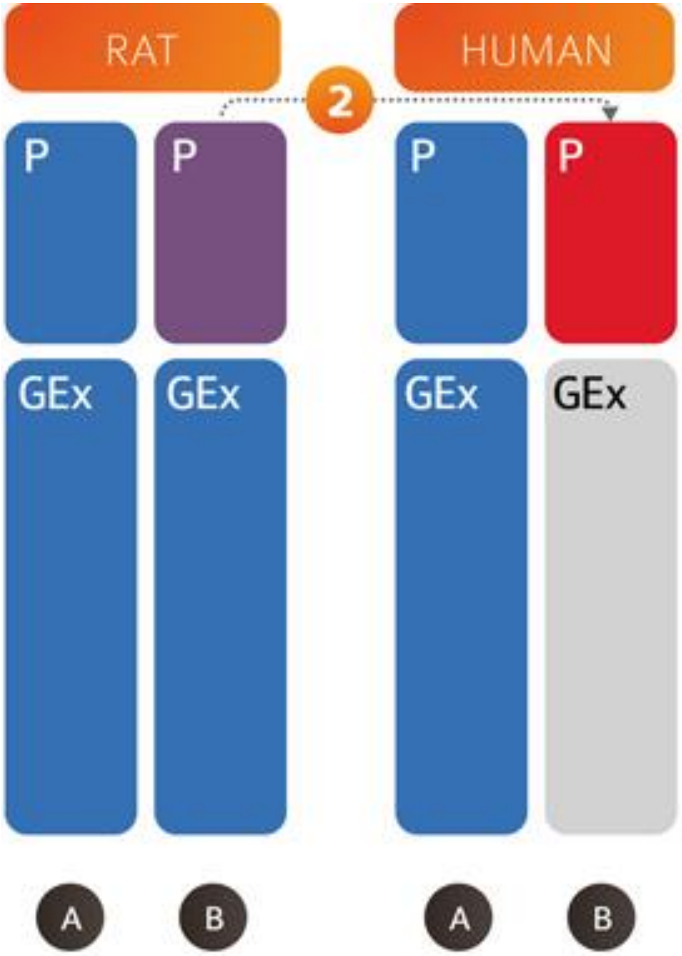


- Predict the protein phosphorylation status for each stimulus in Subset B of rat, from the corresponding gene expression information.
- Question:
  - Is gene expression data sufficiently informative to infer the phosphorylation status through a backward inference process?

Legend:

P	Phosphorylation	⬜	Not provided
GEx	Gene expression	🟦	Provided data
Ⓐ Ⓑ	Stimulus subset	🟥	Predicted data

Sub-challenge 2
Inter-Species Protein Phosphorylation Prediction



- Predict the protein phosphorylation status for each stimulus in subset B in human from the protein phosphorylation status for the same stimulus in subset B in rat.

- Question:
  - Are gene expression and phosphorylation data in one species sufficiently informative to infer the phosphorylation status in another species?

Legend:

P Phosphorylation

GEx Gene expression

Ⓐ Ⓑ Stimulus subset

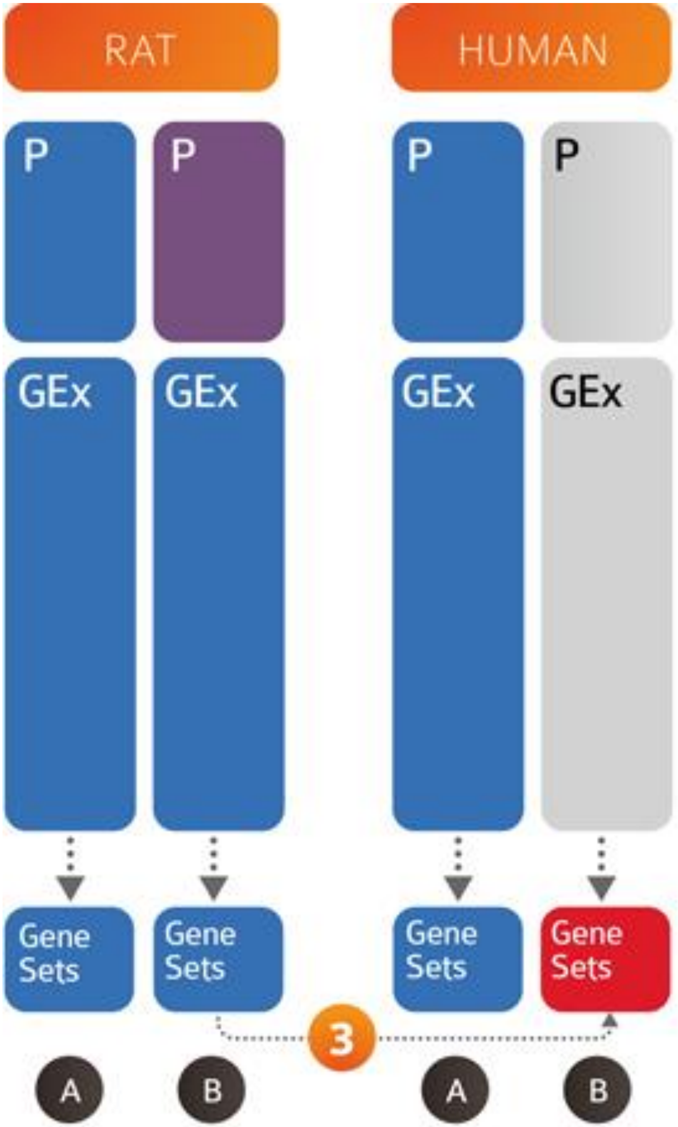
⬜ Not provided

🟦 Provided data

🟥 Predicted data

🟪 Provided after 1 July

Sub-challenge 3
Inter-Species Pathway Perturbation Prediction

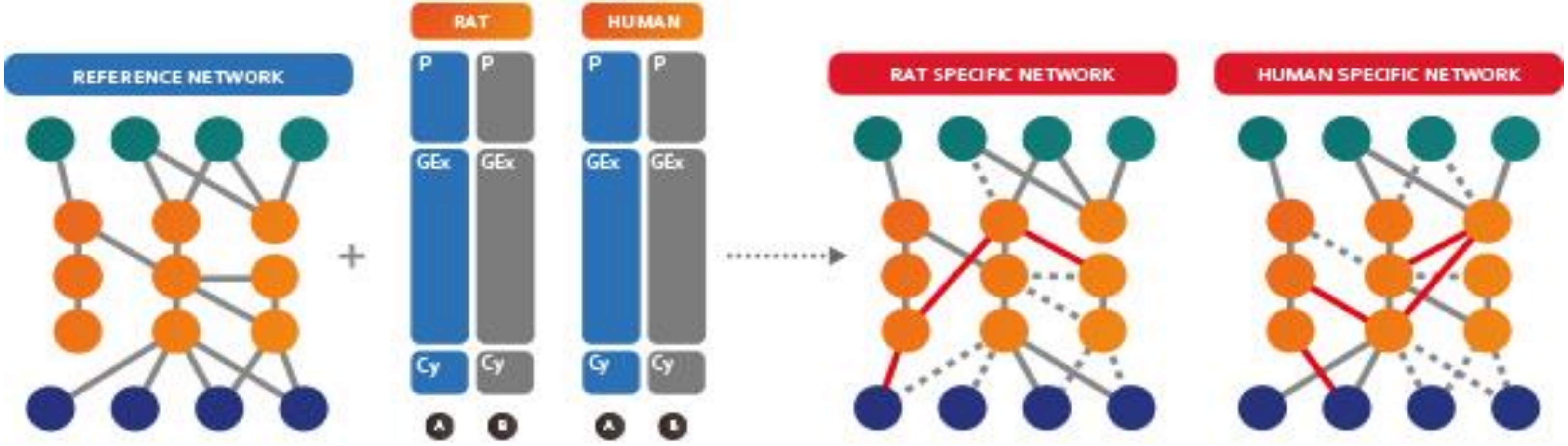


- Predict the gene sets representative of pathways/biological processes that are the most to least enriched among differentially expressed genes with respect to control for each stimulus in Subset B in human based on the corresponding data in rat.
- Question:
  - Can the perturbation of pathways be predicted in human from equivalent information in rat?

Legend:

P	Phosphorylation	⬜	Not provided
GEx	Gene expression	🟦	Provided data
Ⓐ Ⓑ	Stimulus subset	🟥	Predicted data
		🟪	Provided after 1 July

Sub-challenge 4
Species Specific Network Inference



Legend:

P	Phosphorylation	■	Not to be used	—	Edges
GEx	Gene expression	🟦	Provided data	- - -	Removed edges
Cy	Cytokine level	🟥	Inferred network	—	Added edges
		Ⓐ Ⓑ	Stimulus subset		

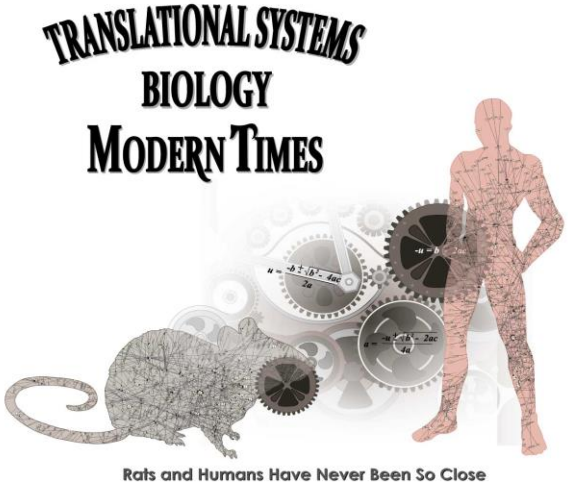
- The goal is to infer human and rat networks given phosphoprotein, gene expression and cytokine data and a reference map provided as prior knowledge. Participants will use network inference to add or remove edges from the reference map to produce specific rat and human networks.
- Question:
  - Can biological networks be built by leveraging diverse ‘omics’ data to assess the commonalities and differences between the species?

Challenge results and award ceremony

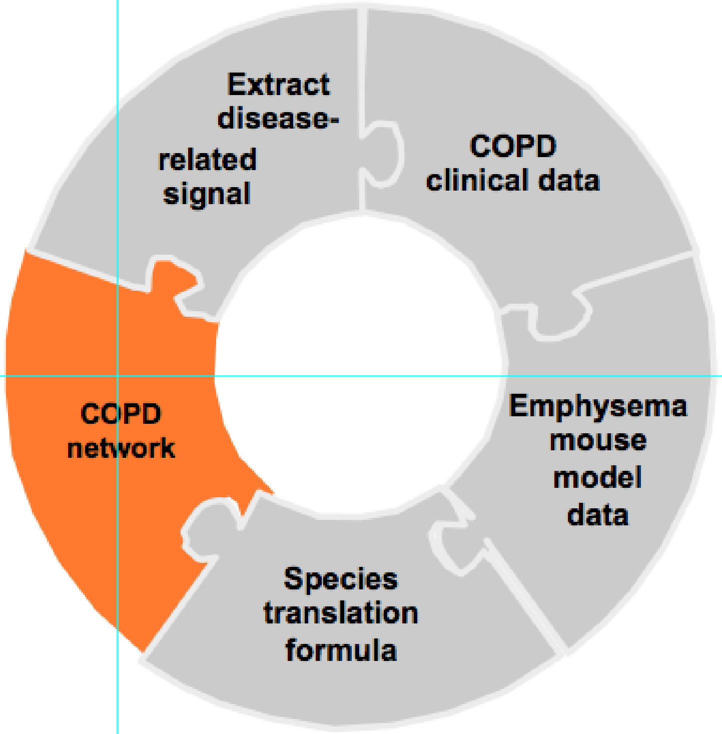
Symposium 2013 (29-31 October 20123 in Athens, Greece)

- Award ceremony
- Presentations by the best performing teams
- Keynote Speakers from Systems Biology Community

For more details on **sbv IMPROVER** and the Species Translation Challenge, visit **www.sbvimprover.com**



Network Verification Challenge

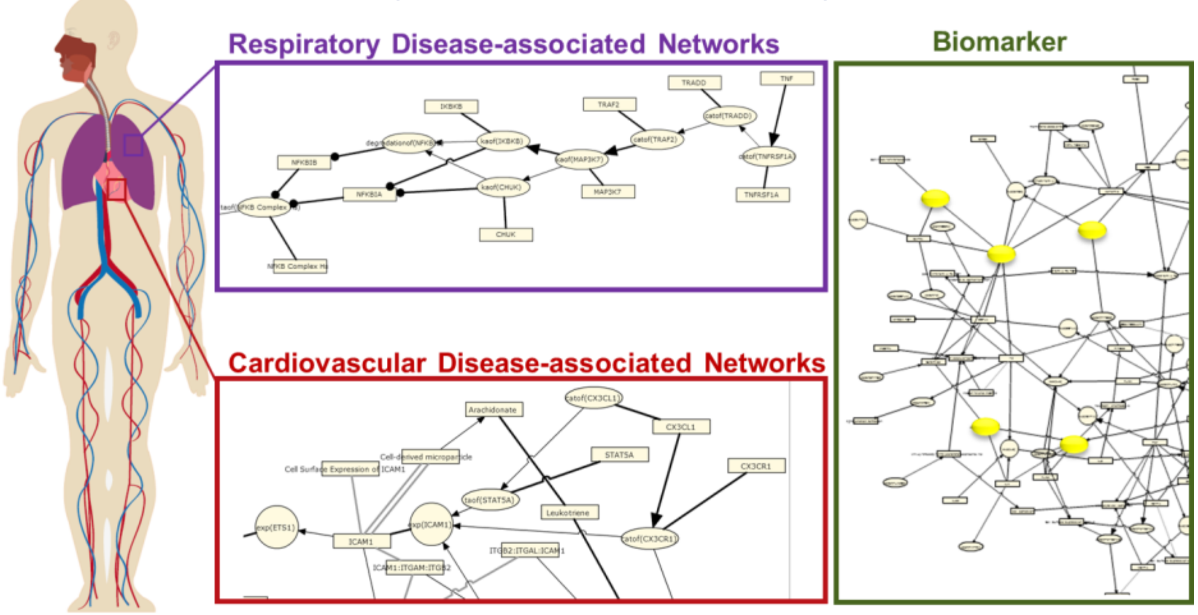


www.sbvimprover.com

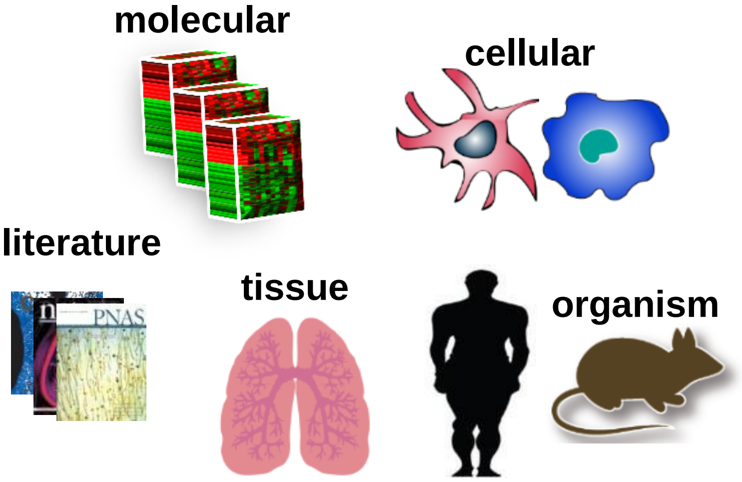
Overview of Network Verification Challenge

- The disparate information on molecular mechanisms of the respiratory system has been organized and captured within a coherent collection of network models.
- The purpose of the Network Verification Challenge is to engage the scientific community to review, challenge, and make corrections to the conventional wisdom
- The verified network will be used in the “COPD Grand Challenge”

Network Biology for Systems Toxicology and Biomarker Discovery



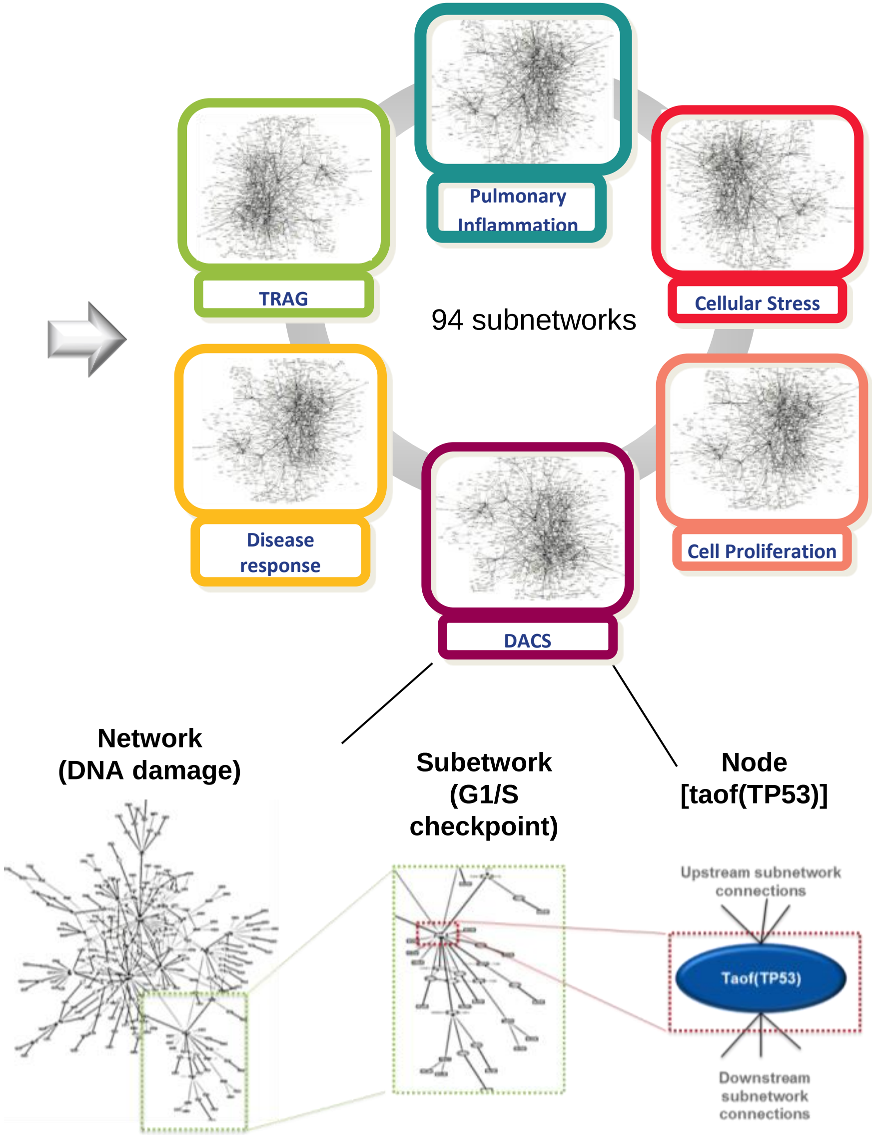
Networks Contain Relevant Biology Expressed in a Causal Framework



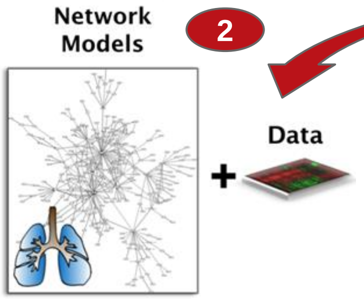
-Capture wide range of biology

-Represent specific species and/or tissue context

-Causal relationships in the network can be traced to measurable entities



Network Verification Challenge in a nutshell

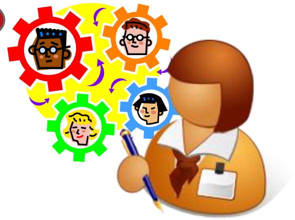


2

Import Network
Select a subset of the COPD biological network models for verification

3

Crowd Verification
Participants verify nodes and edges based on scientific findings



4

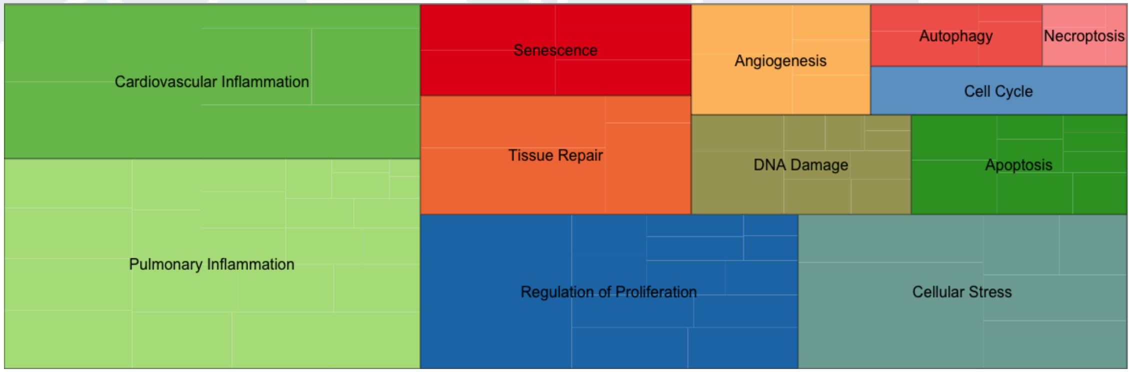
Interpretation of results
The project team will select “questionable” edges based on their consensus score

5

Jamboree
Selected participants will analyze scientific evidences for “questionable” edges

1

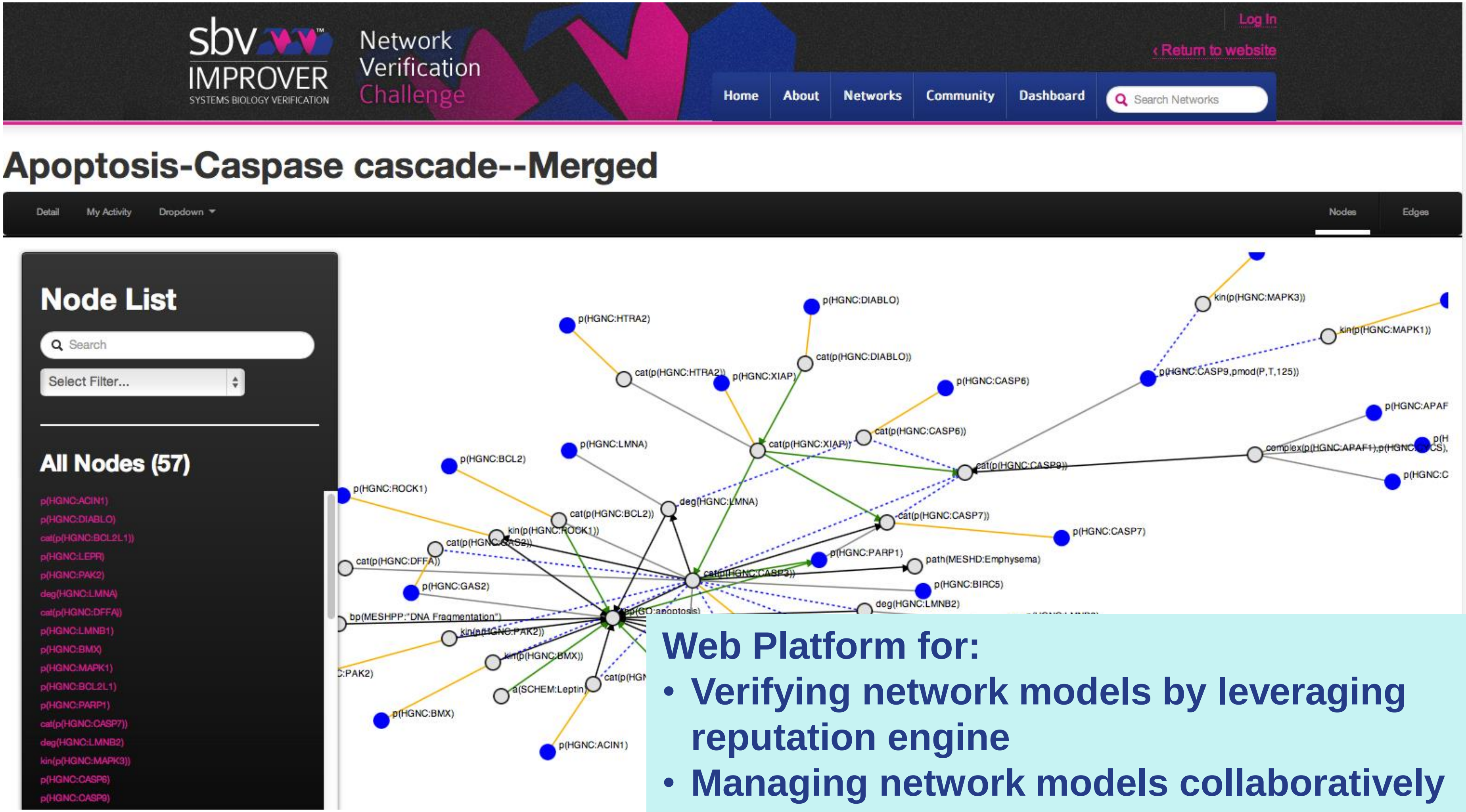
Web Infrastructure



6

Publication
Selected participants submit articles with results and conclusions to peer-reviewed journals

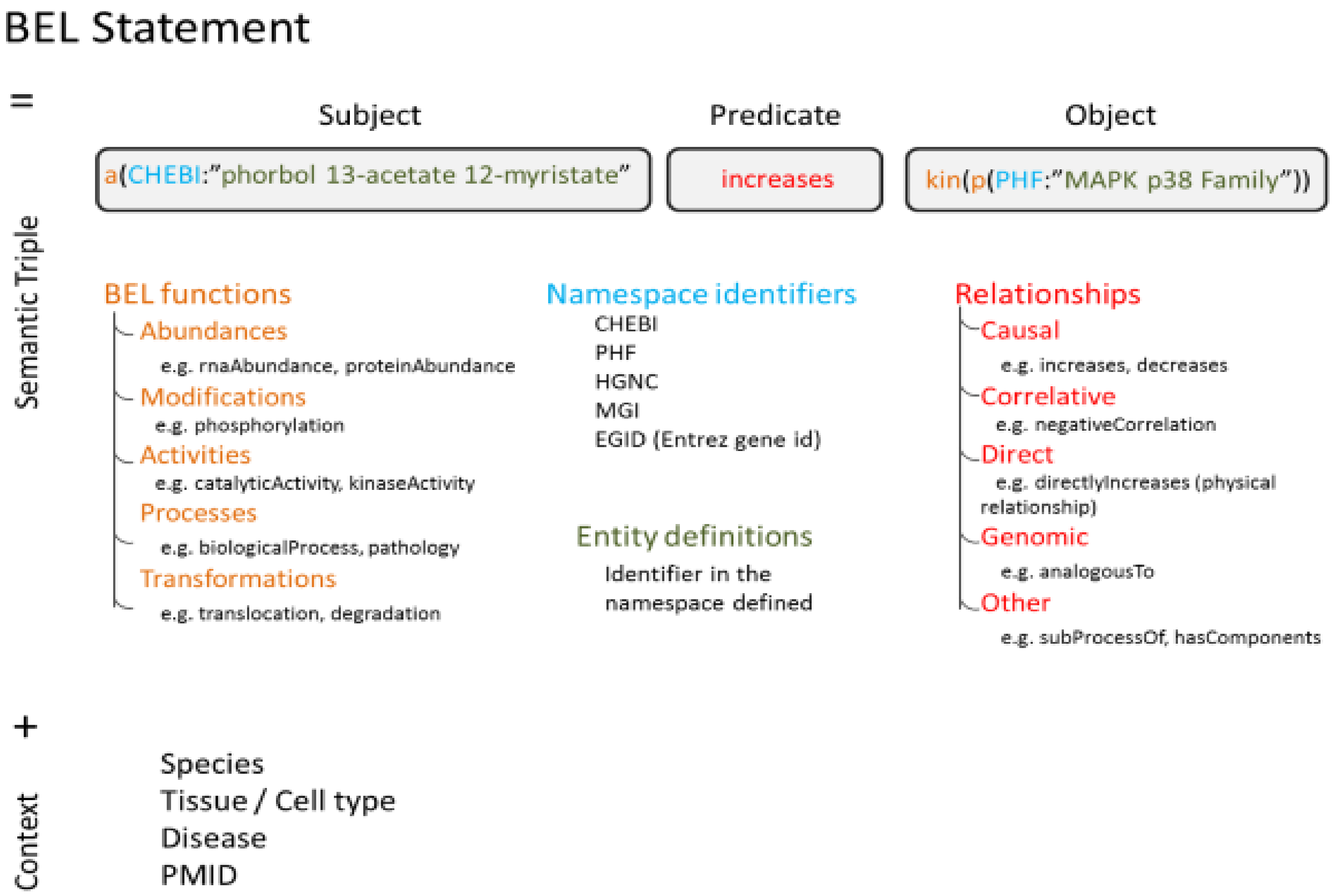
NVC Website



Web Platform for:

- Verifying network models by leveraging reputation engine
- Managing network models collaboratively
- Generating BEL knowledge
- Developing network models

BEL (Biological Expression Language) Statement



BEL Statement

=

Subject	Predicate	Object
a(CHEBI:"phorbol 13-acetate 12-myristate"	increases	kin(p(PHF:"MAPK p38 Family"))

Semantic Triple

BEL functions

- Abundances
  - e.g. rnaAbundance, proteinAbundance
- Modifications
  - e.g. phosphorylation
- Activities
  - e.g. catalyticActivity, kinaseActivity
- Processes
  - e.g. biologicalProcess, pathology
- Transformations
  - e.g. translocation, degradation

Namespace identifiers

CHEBI
PHF
HGNC
MGI
EGID (Entrez gene id)

Entity definitions

Identifier in the namespace defined

Relationships

- Causal
  - e.g. increases, decreases
- Correlative
  - e.g. negativeCorrelation
- Direct
  - e.g. directlyIncreases (physical relationship)
- Genomic
  - e.g. analogousTo
- Other
  - e.g. subProcessOf, hasComponents

+

Context

Species
Tissue / Cell type
Disease
PMID

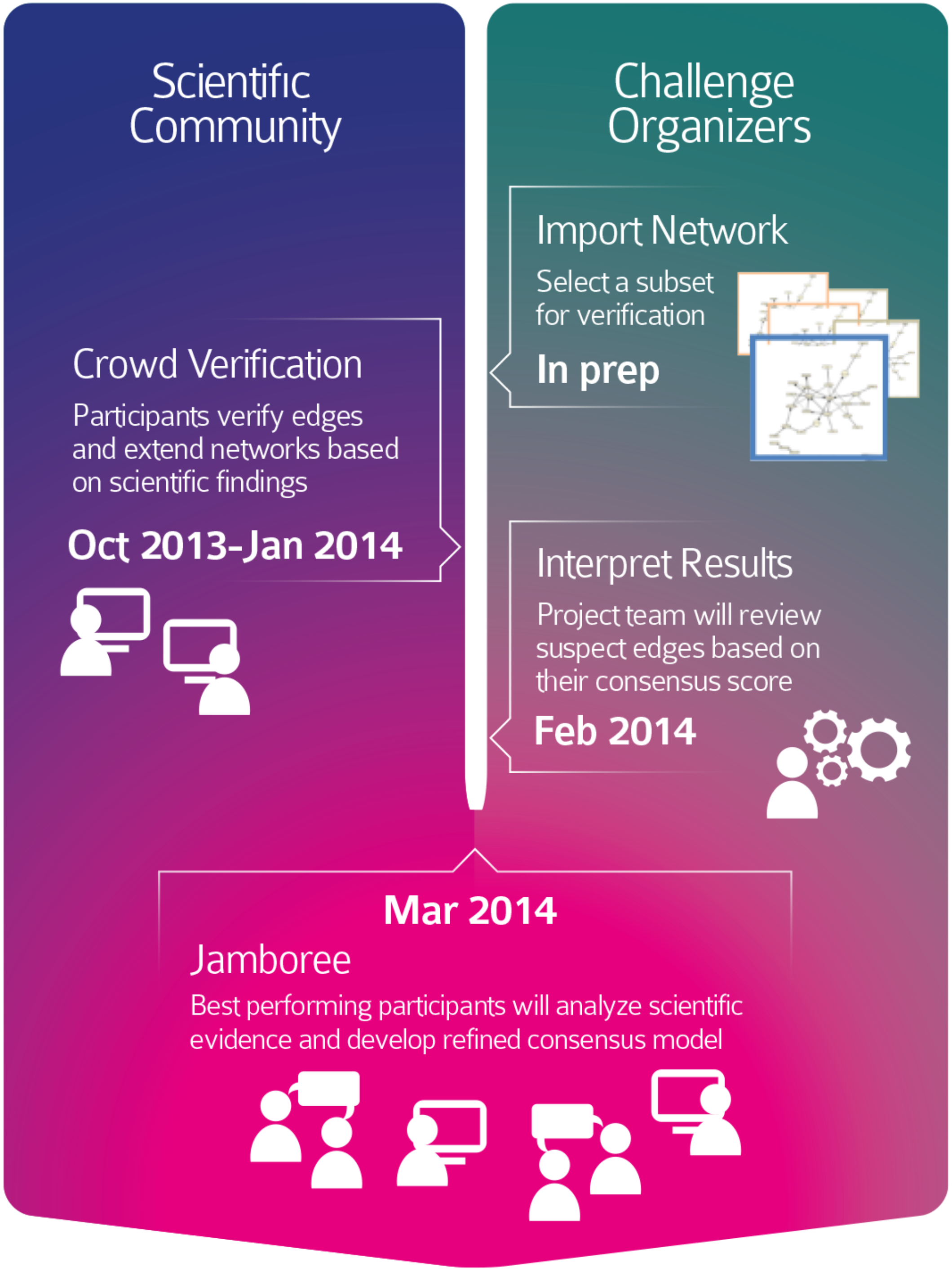
Who can participate?

- Any biology researcher / student
- Researchers / students working on inflammation / lung tissues / COPD
- Researchers / students interested in pathways verification
- Researchers / students interested in applying text mining to an applied biological case

Why should you participate?

- Gain access to high quality and novel data
- Enhance your visibility and gain recognition
- Engage with peers to advance the field
- Top performers get invited to the Jamboree with travel expenses covered

Timelines



Scientific Community

Challenge Organizers

Import Network

Select a subset for verification

In prep

Crowd Verification

Participants verify edges and extend networks based on scientific findings

Oct 2013-Jan 2014

Interpret Results

Project team will review suspect edges based on their consensus score

Feb 2014

Mar 2014

Jamboree

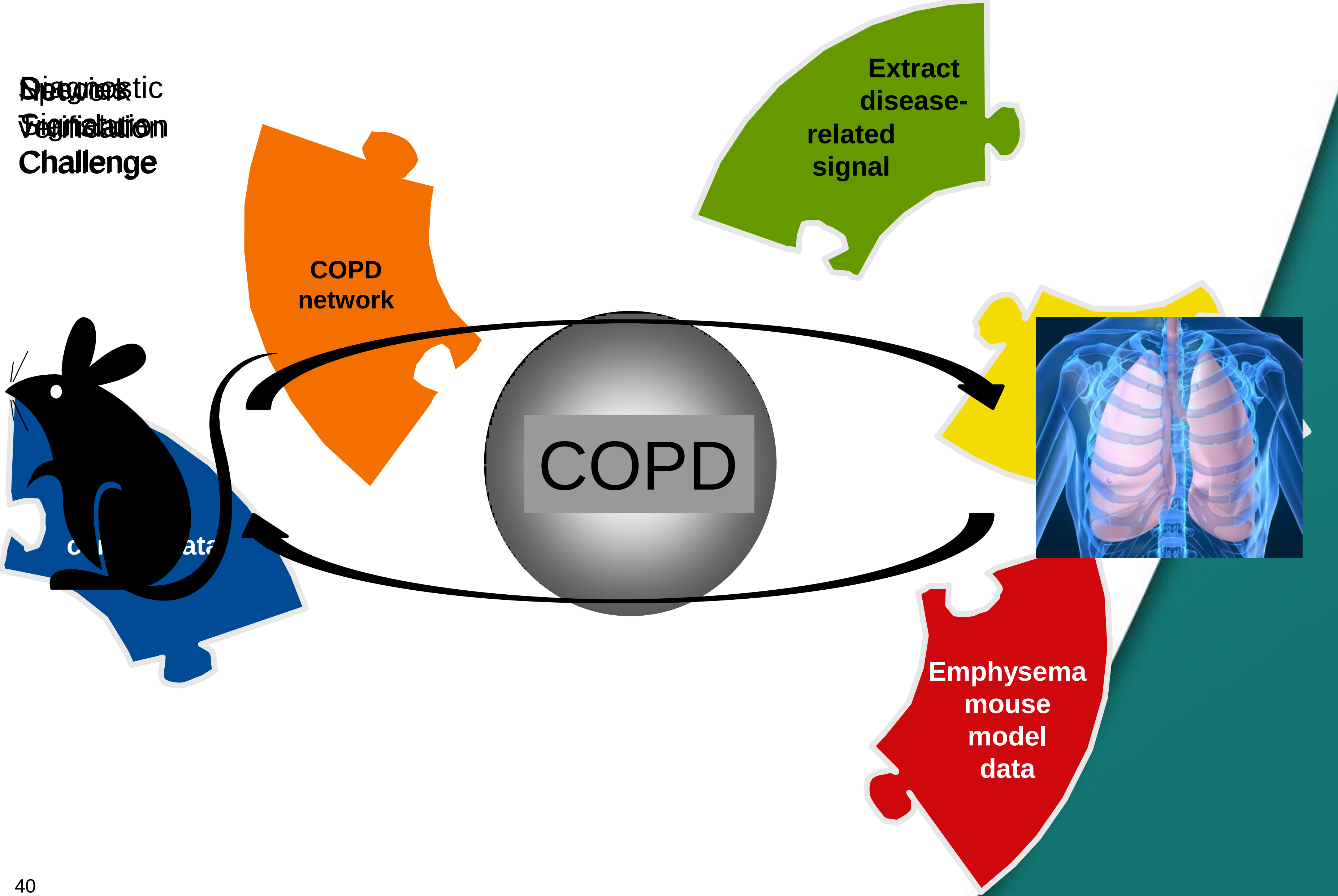
Best performing participants will analyze scientific evidence and develop refined consensus model

THE “GRAND CHALLENGE”

Diagnostic Signature Challenge

Species Translation Challenge

Network Verification Challenge



What do we want to address in the Grand Challenge?

- We want to:
  - identify indicators for onset of COPD
  - develop a comprehensive model of COPD onset

- We will have:
  - all the previously developed “puzzle” pieces
  - newly collected clinical data
  - newly collected rodent data

Clinical Endpoint Comparison to Emphysema Mouse Model

	Emphysema Mouse Model	COPD marker Identification Study
Genomics / Transcriptomics	White blood cells Nasal epithelium	White blood cells Nasal scrapes
Proteomics	Blood Bronchoalveolar lavage	Blood Sputum
Protein markers associated with inflammation	Bronchoalveolar lavage	Sputum
Cellular	Differential cell count in BALF	Differential cell count in sputum
Clinical / Symptomatic	Lung histopathology	High resolution computerized tomography, measurement of lung damage
	Full lung function	Full lung function
	Blood screening	Blood screening

Grand Challenge Summary

- Planned launch date in Q2 2014
- Leverage the “wisdom of crowds” to develop methodologies for predicting the prognostic impact of different stimuli on COPD.
- Network information verified by the Network Verification Challenge will be included as one of the inputs
- From this and the preceding challenges, we as a scientific community will better understand the biology that underlies COPD.

CURRENT SBV IMPROVER PROJECT TEAM

IBM

- Ajay Royyuru
- Elise Blase
- Erhan Bilal
- Gustavo Stolovitzky
- Jeremy Rice
- Kahn Rhrissorrakrai
- Pablo Meyer
- Raquel Norel

- Alf Scotland
- Claudia Frei
- Immanuel Luhn
- Joanna Taylor
- Marianne Charaf
- Peter Curle

PMI

- Bruce O'Neil
- Carine Poussin
- Carole Mathis
- Filipe Bonjour
- Florian Martin
- Hugh Browne
- Jean Binder
- Julia Höng
- Lionel Schilli
- Manuel Peitsch
- Marja Talikka
- Nikolai Ivanov
- Stephanie Boué
- Yang Xiang

External collaborators

- Protatonce
- Selventa

The sbv IMPROVER project, the website and the Symposia are part of a collaborative project designed to enable scientists to learn about and contribute to the development of a new crowd sourcing method for verification of scientific data and results.

The project team includes scientists from Philip Morris International's (PMI) Research and Development department and IBM's Thomas J. Watson Research Center. The project is funded by PMI.

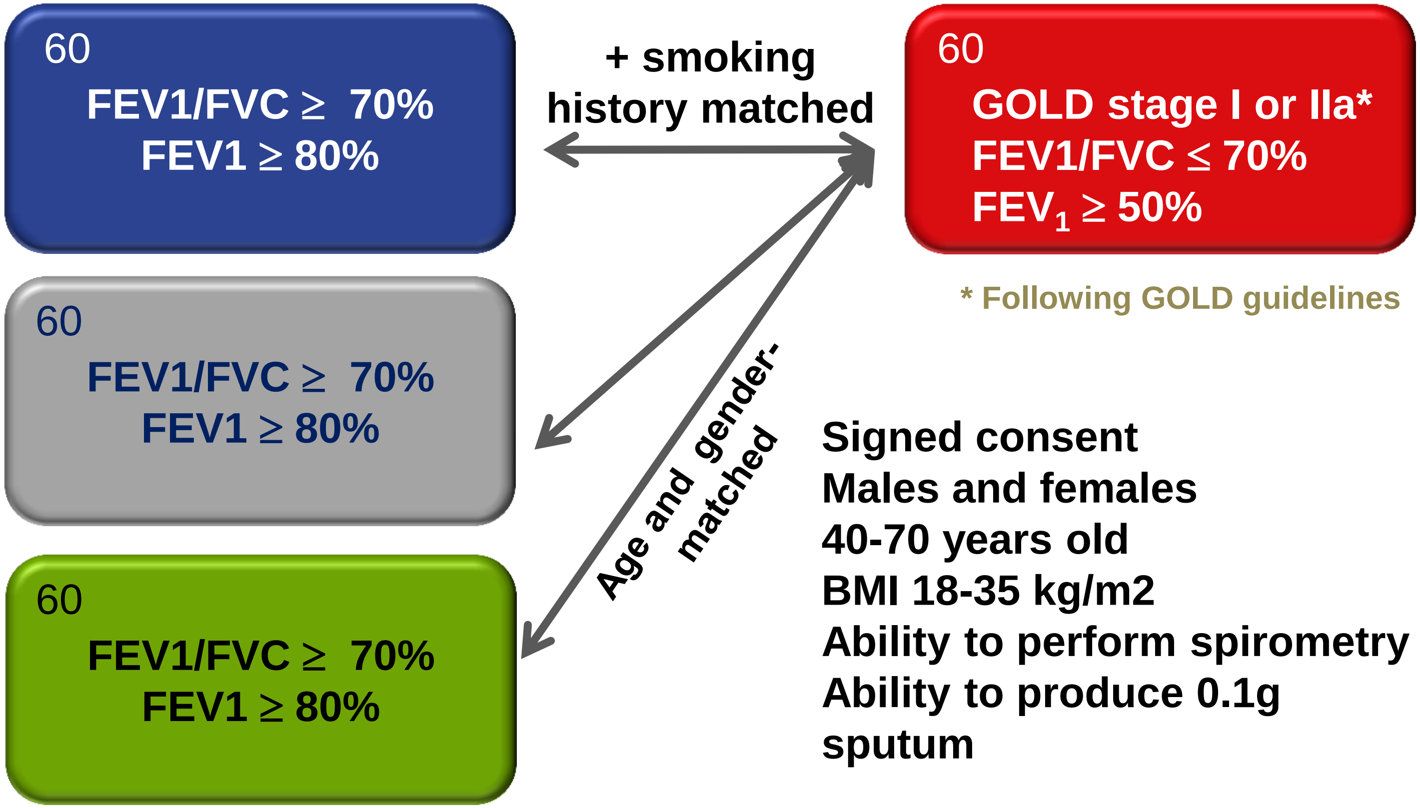
BACK UP SLIDES

Biological Samples

Induced Sputum	**Proteomics**	**Best representation of the inflammatory state of the small airways. Contains cells of interest and protein mediators**
Whole Blood (Lymphocytes)	**Transcriptomics Proteomics Lipidomics**	**Identify correlations with data identified from the sputum. Blood represents the ideal and most convenient matrix for large-scale biomarker measurements**
Nasal fluid	**Proteomics**	**A number of respiratory researchers have shown strong correlations between inflammation in the lung & nose. May represent a convenient matrix for large-scale studies**
Nasal scrapes	**Transcriptomics Proteomics**	**May represent a convenient matrix for large-scale studies if correlations are seen with the sputum**
Nasal lavage	**Transcriptomics Proteomics**	**May represent a convenient matrix for large-scale studies if correlations are seen with the sputum**

COPD Biomarker Identification Study - Design

Non-interventional, observational case-control design study conducted in the United Kingdom, and has been approved by the UK National Health Service (NHS) Ethics Committee



Controls

COPD

Current Smokers
(≥ 10 pack-year smoking history)

60
FEV1/FVC ≥ 70%
FEV1 ≥ 80%

\+ smoking history matched

60
GOLD stage I or IIa*
FEV1/FVC ≤ 70%
$FEV_1 \geq 50\%$

* Following GOLD guidelines

Former Smokers

60
FEV1/FVC ≥ 70%
FEV1 ≥ 80%

Age and gender-matched

Never Smokers

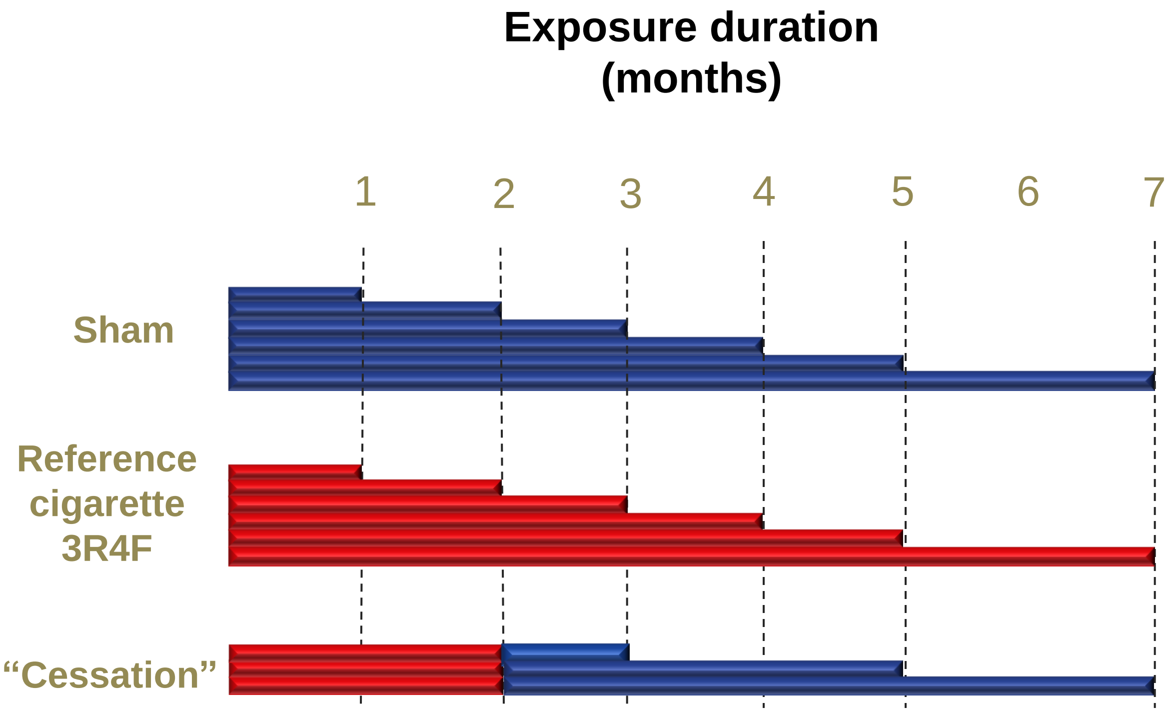
60
FEV1/FVC ≥ 70%
FEV1 ≥ 80%

Signed consent
Males and females
40-70 years old
BMI 18-35 kg/m2
Ability to perform spirometry
Ability to produce 0.1g sputum

Biological Samples

Induced Sputum	**Proteomics**
Whole Blood (Lymphocytes)	**Transcriptomics** **Proteomics** **Lipidomics**
Nasal fluid	**Proteomics**
Nasal scrapes	**Transcriptomics** **Proteomics**
Nasal lavage	**Transcriptomics** **Proteomics**

Study Design and Measured Endpoints in Emphysema Mouse Model



Inflammation:
- BALF** analysis
- Differential leucocyte count in whole blood

Pulmonary function
- Flow-volume loops
- FEV0.1 ***
- Resistance, Compliance
- Elastance

Lung histopathology and morphometry

Genomics and Transcriptomics (lung, nasal epithelium, aortic arch, liver, blood)

Lipidomics (lung, liver, aorta, blood)

** BALF: bronchoalveolar lavage fluid
*** FEV0.1 forced expiratory volume in 0.1s