IMPROVER
systemsbiologyverification

IMPROVER and its Application to PMI R&D

Manuel C. Peitsch, VP, Biological Systems Research, PMI Research & Development

Gustavo Stolovitzky, Manager, Functional Genomics & Systems Biology, IBM Computational Biology Center

2nd October 2012

IMPROVER: Industrial Methodology for Process Verification in Research

Collaboration of IBM and Philip Morris International on a project funded by Philip Morris International

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

Verifies individual methods using double blind performance assessment

computational BIOLOGY

COMMENTARY

Verification of systems biology research in the age of collaborative competition

Pablo Meyer[1], Leonidas G Alexopoulos[2], Thomas Bonk[3], Andrea Califano[4], Carolyn R Cho[5], Alberto de la Fuente[6], David de Graaf[7], Alexander J Hartemink[8], Julia Hoeng[3], Nikolai V Ivanov[3], Heinz Koeppl[9], Rune Linding[10], Daniel Marbach[11], Raquel Norel[1], Manuel C Peitsch[3], J Jeremy Rice[1], Ajay Royyuru[1], Frank Schacherer[12], Joerg Sprengel[13], Katrin Stolle[3], Dennis Vitkup[4] & Gustavo Stolovitzky[1]

Collaborative competitions in which communities of researchers compete to solve challenges may facilitate more rigorous scrutiny of scientific results.

Nature Biotechnology 2011 Sep 8;29(9):811-5

What do we mean by Verification?

- ***Verification*** **is the process of determining if the theory, experiments and their associated data accurately represent** the hypotheses underlying a scientific research undertaking and the correctness of the scientific results and interpretations.

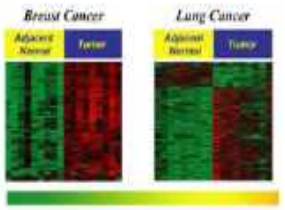
- ***Validation*** **is to confirm that the processes, methods or components meet documented requirements and specifications** levied on the design. It is about “Doing the science right”.

Why do we need verification in Systems Biology?

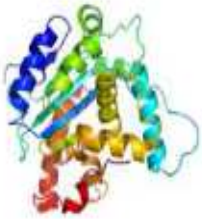
We are experiencing a data deluge…

Genomic

Literature



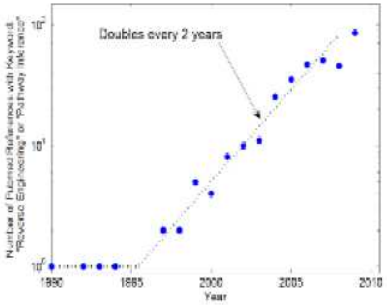
Molecular Profiles



Structures

Explosion of data

But we lack the corresponding verification tools to know that we are doing the right science…



Explosion of algorithms for Reverse Engineering

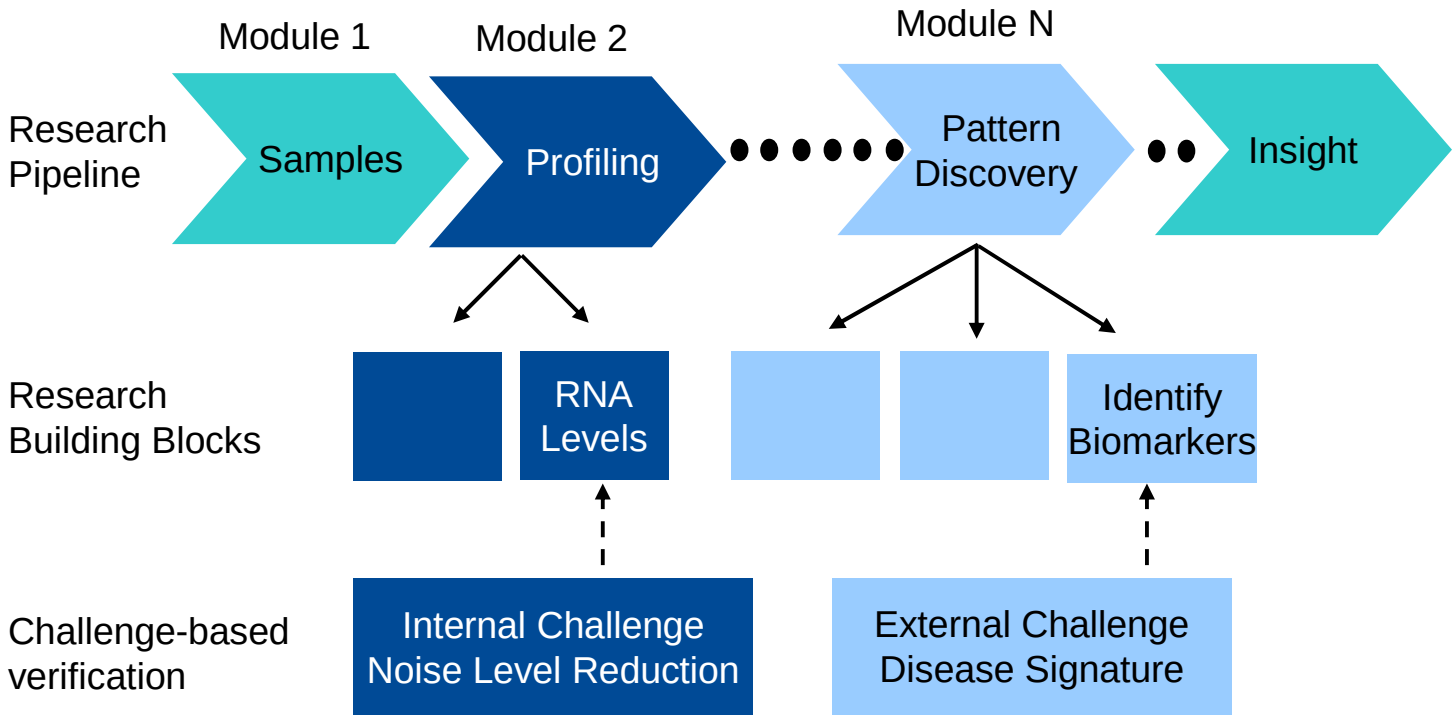
ϕ

Lack of methods for high throughput verification

Lack of independent assessment of computational methods

IMPROVER aims to develop a robust methodology that verifies systems biology-based approaches in an industrial context

IMPROVER starts by identifying a research pipeline and dividing it into Verifiable Building Blocks



Building blocks support each other towards a final goal

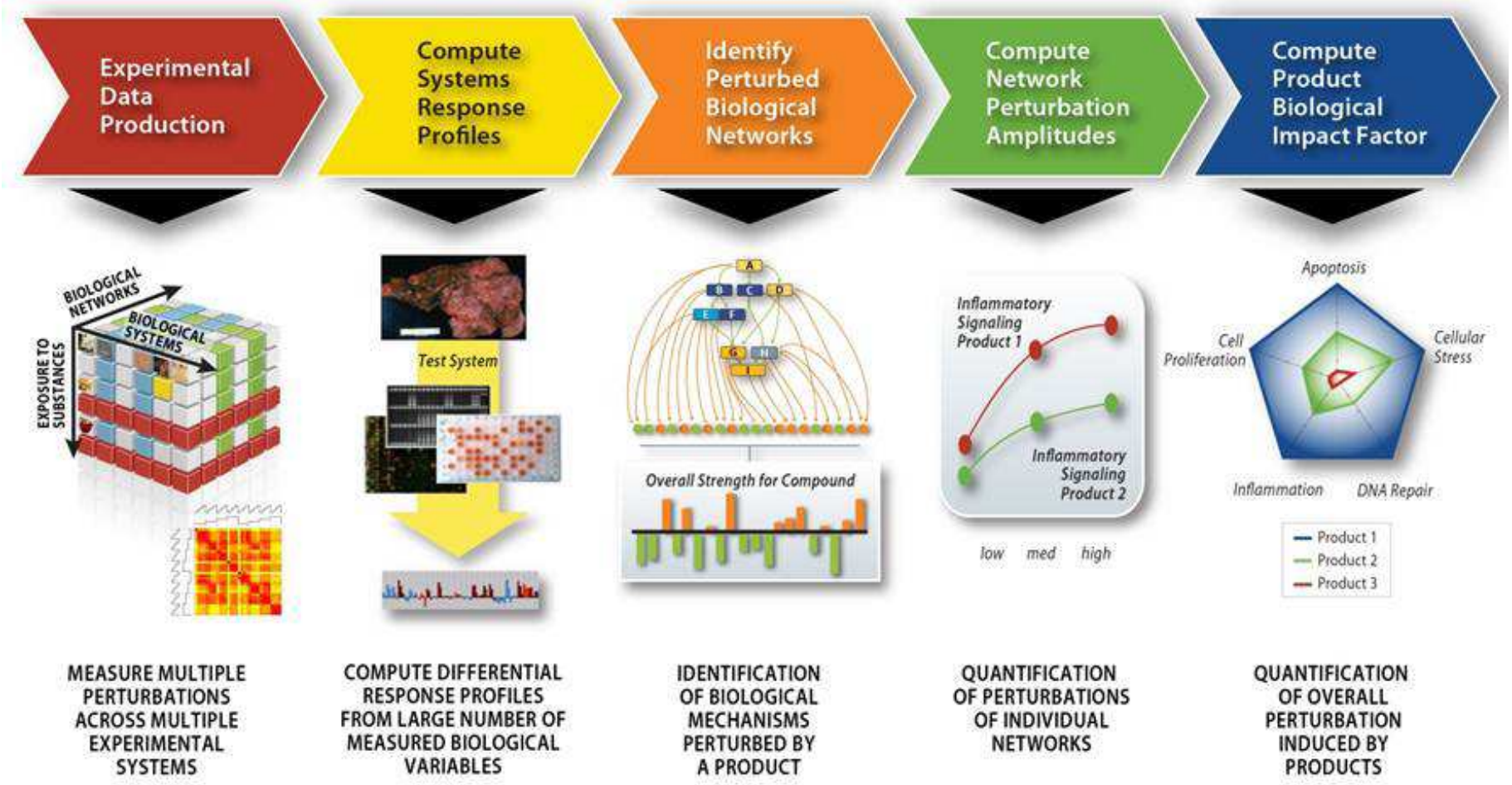
Each building block is verifiable by a challenge

IMPROVER employs internal challenges and external challenges

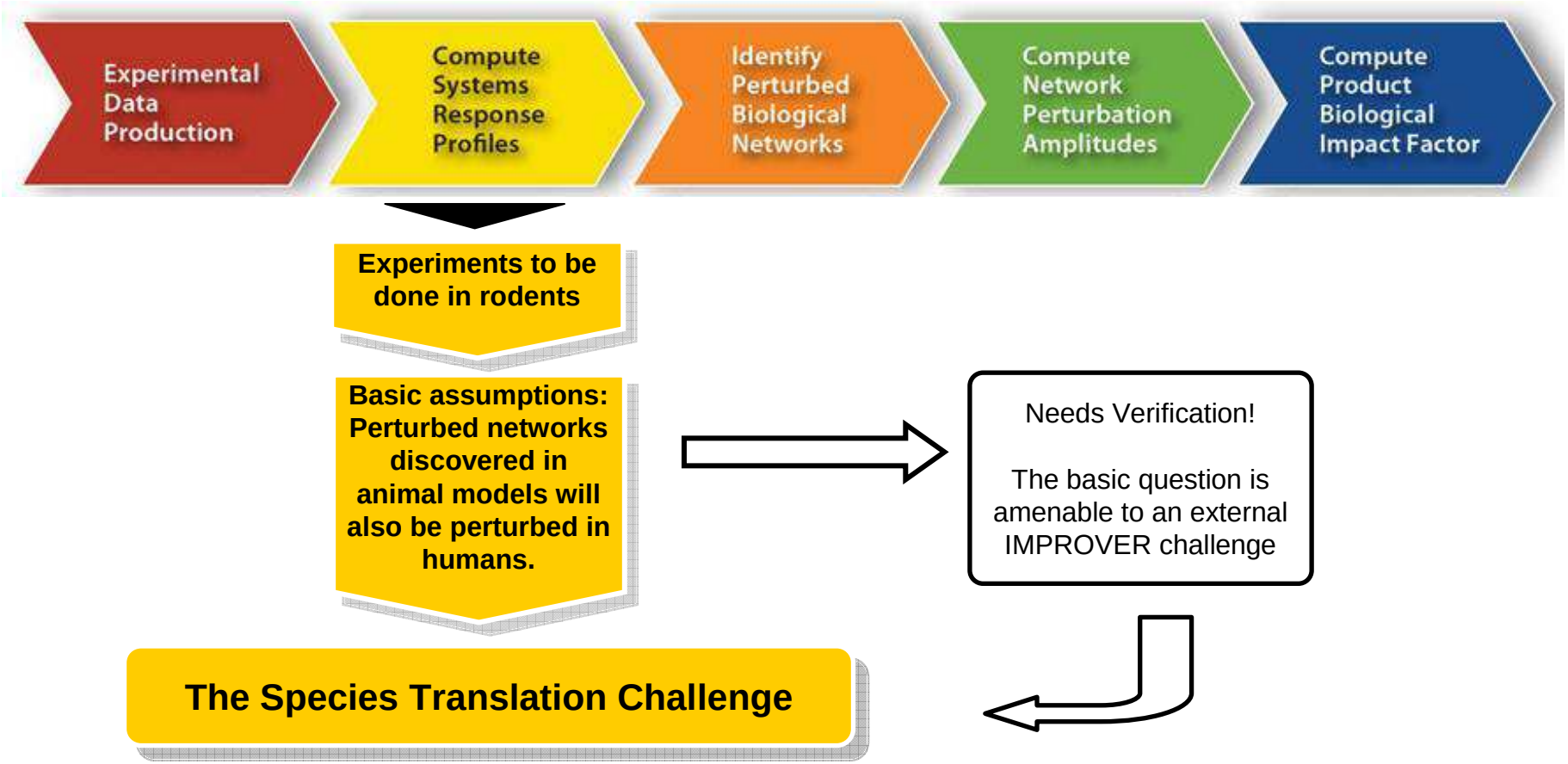
- **Verification by Internal (to an industry) Challenges**
  - Not crowd-sourced because of proprietary concerns
  - Scope does not require and external community
  - Lack of sufficient interest to entice the research community at large

- **Verification by External (to an industry) Challenges**
  - It invites new approaches to a problem not considered by internal researchers
  - Can leverage the wisdom of crowds
  - Elicits public discussion, building consensus on best approaches
  - Can flag a building block as “unsolvable” (at least at a given moment in time)

The PMI Systems Biology pipeline



The PMI Systems Biology pipeline: Preview of an upcoming challenge



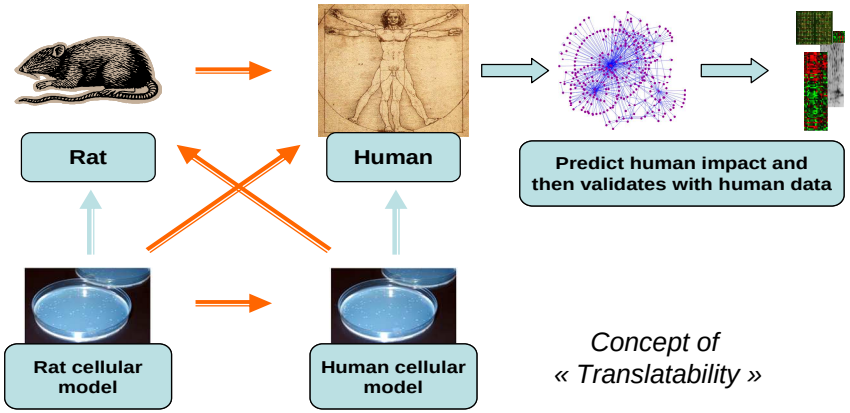
What is the IMPROVER Species Translation Challenge?

- Goal: Verify that a mapping exists and allows the translation of biological effects of stimulus-induced perturbations in one species given information of the same perturbations in another species.

- Scientific Question:

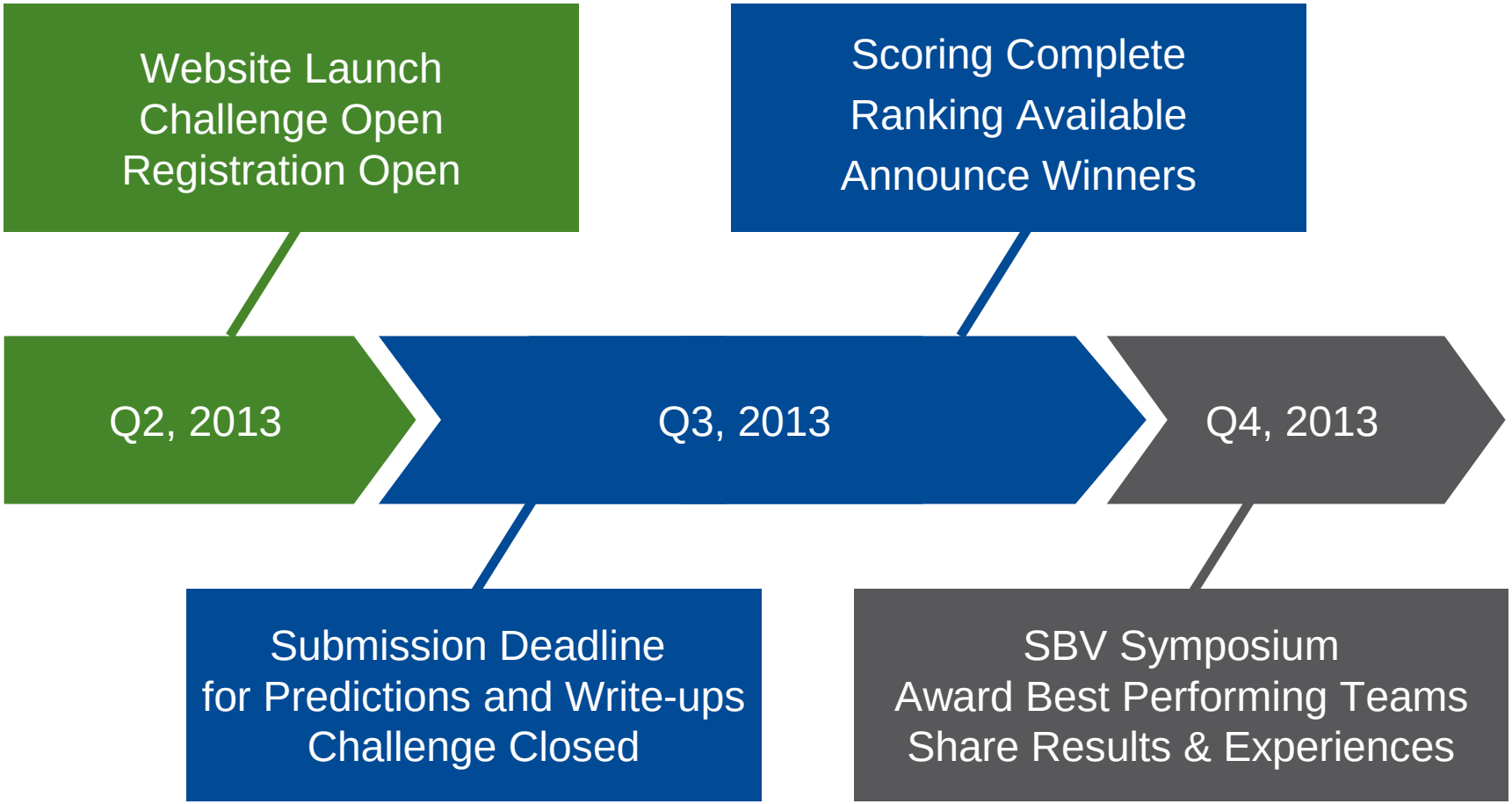
  How accurately can the observations from an in Vivo, in Vitro rodent models be translated to the human situation?

  1. Is it possible to predict the amplitude of perturbation (e.g. expression levels) in one species given the perturbation amplitude measured in another species ?
     - Methodology development of a translation function
  2. Which area of the biological system is translatable/predictable and which is not?
     - Range of applicability of the translatability concept

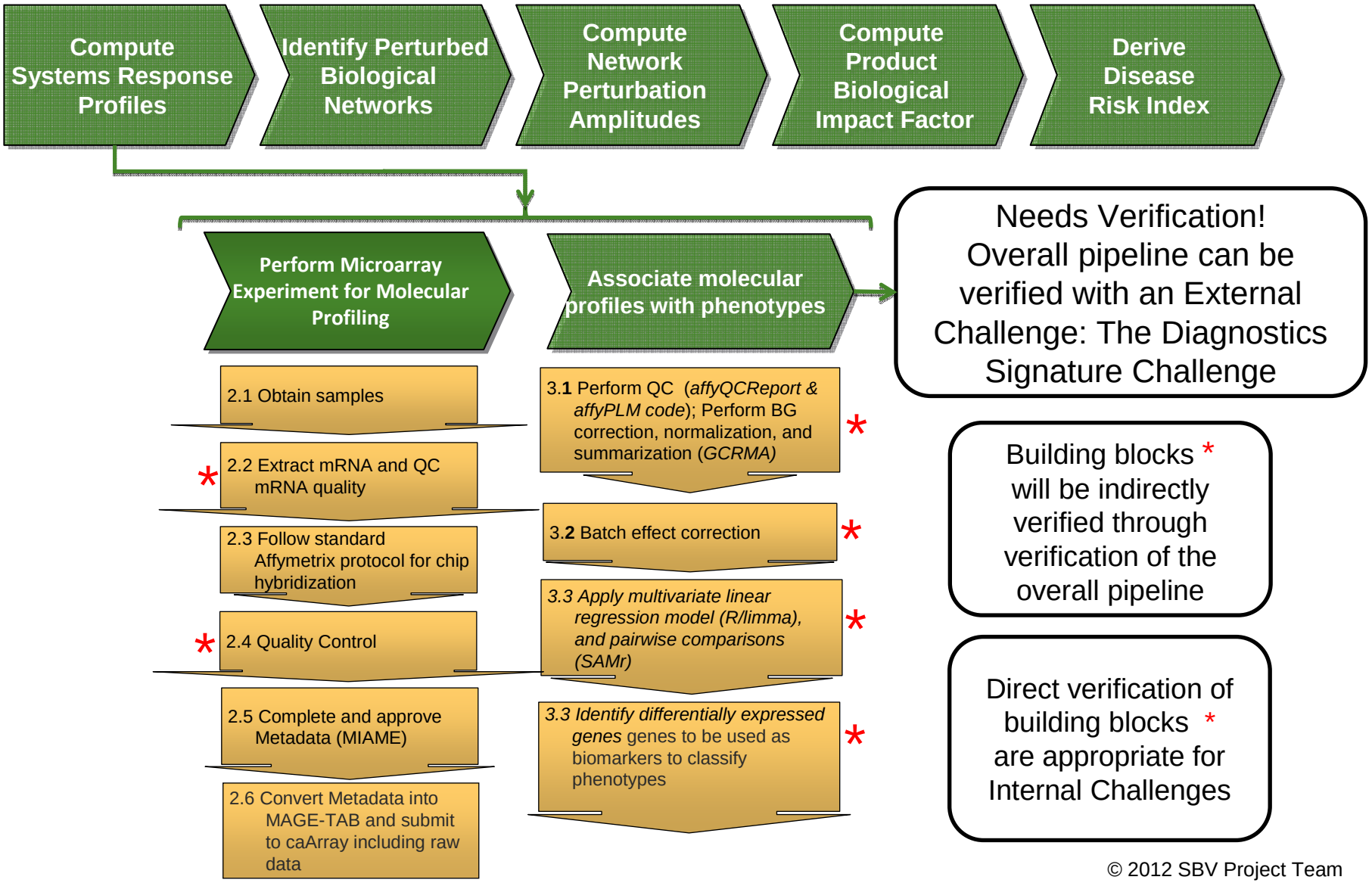


Concept of « Translatability »

Provisional Species Translation Timeline



The PMI Systems Biology pipeline: Rationale of the IMPROVER DSC



Why do we need to verify that it is possible to infer clinical phenotype from genomics data?

A few success stories of gene expression based biomarkers in clinical use

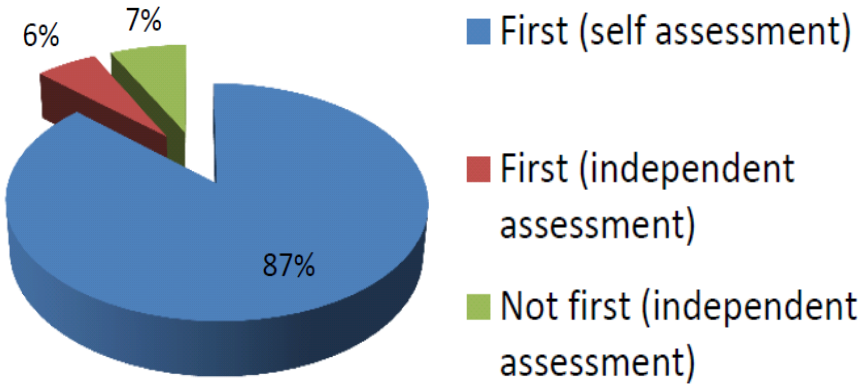
✓*MammaPrint* (breast cancer recurrence assay)
70-gene profile; requires fresh tissue
✓*Oncotype Dx* (breast cancer recurrence assay)
21-gene profile; works on both fresh and fixed tissue

Counter-balanced by a few failure stories of gene expression based biomarkers in clinical use

✓*Potti et al, Nat Med (2006)* claimed to identify genomic signatures for drug response. Three clinical trials begun in 2007, 2008 for lung and breast cancer. The research was later deemed statistically flawed and at least 10 high profiled publications were retracted and the clinical trials stopped.
✓Amgen scientists tried to confirm 53 landmark papers in pre-clinical oncology research: Only 6 (11%) were confirmed.
✓Bayer HealthCare reported that only about 25% of published preclinical studies could be validated.

Why do we need to perform verification in a double blind way: the "self assessment" trap

- ✓ Researchers wishing to publish their methods are usually required to compare their methods against others
- ✓ Authors' method tends to be the best in an unreasonable majority of cases
  - ✓ Selective reporting of performance: inadvertent or disingenuous
  - ✓ Choice of only one, best metric



# of metrics	Total # of papers surveyed	Method best in all metrics & most datasets	Method best in most metrics & most datasets
1	25	19	6
2	15	13	2
3	7	4	3
4	4	1	3
5	4	1	3
6	2	1	1

57 Self assessed papers

Norel, J.J.Rice, G. Stolovitzky, Mol. Sys. Bio, Oct 11;7:537 (2011)

The DSC is a timely external IMPROVER challenge to lend credibility to the field of genomic profiling of disease

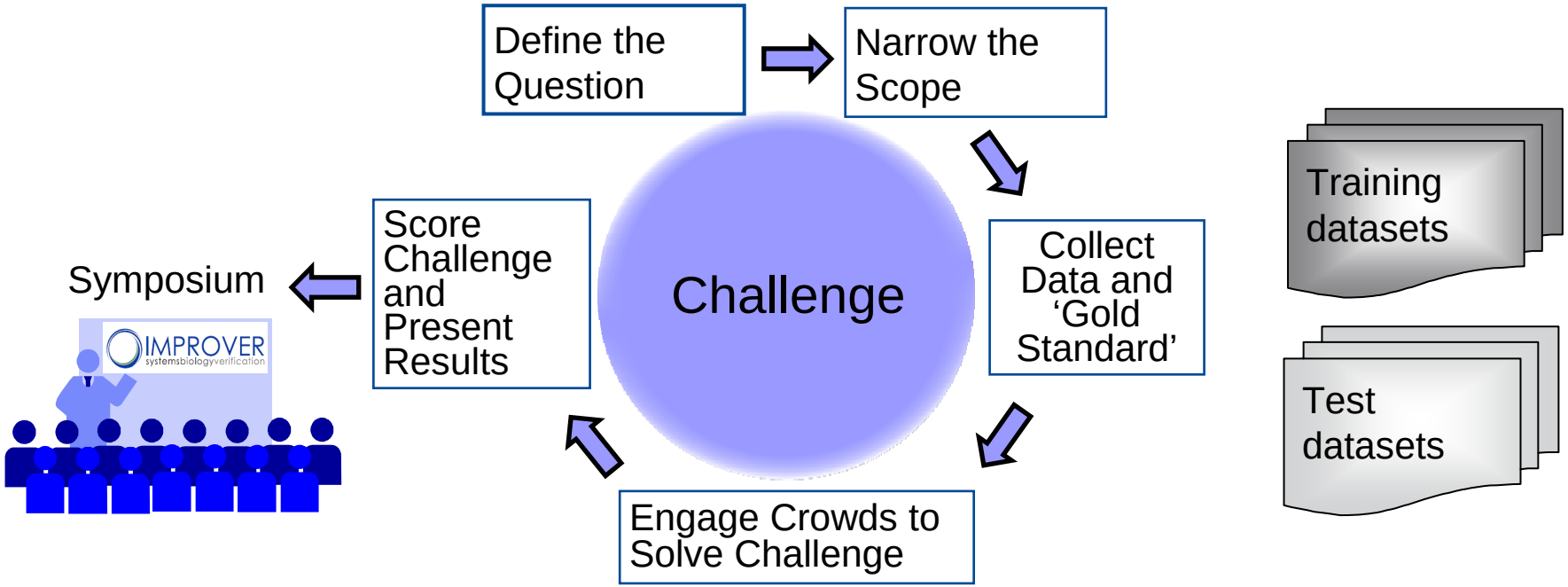
Ultimate goal

•Verify if transcriptomics data contains enough information for the diagnosis and/or prognosis of 5 human diseases

Useful by-products

•Identify best methods for particular data types
•Determine the dependence of performance on the methods of choice
•Study if the wisdom of crowds applies to diagnostics signatures
•Study the overlap of genes in the signatures (when applicable)

The Elements of a Challenge



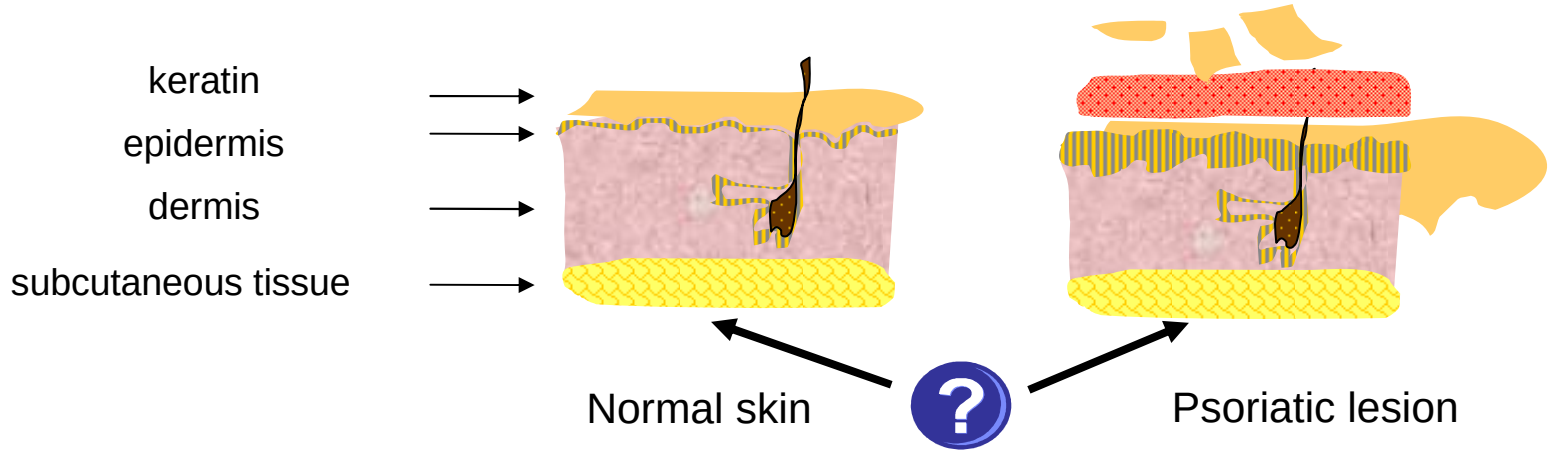
Crowd sourcing brings new ideas and leverages the wisdom of crowds

Clear challenge description and forum discussion in a user friendly website

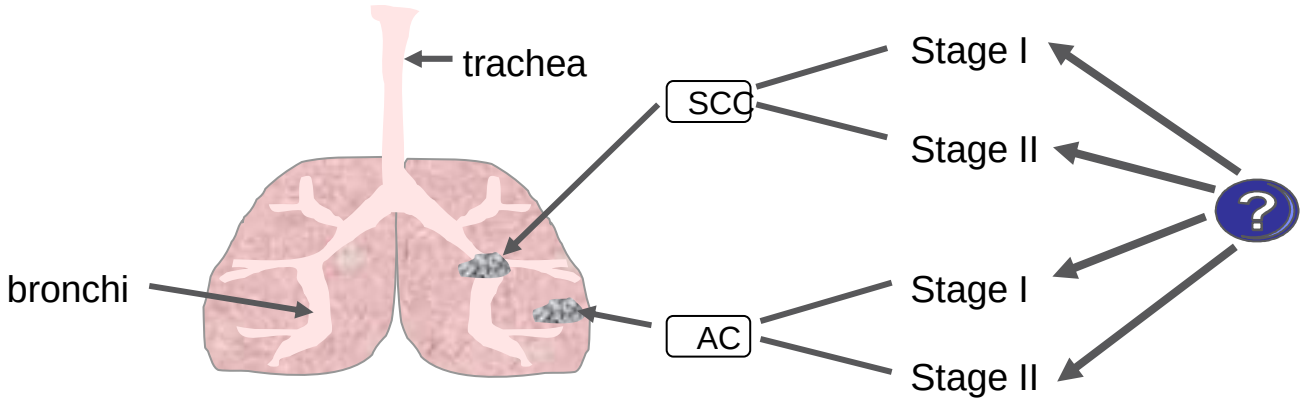
Four Diseases – Five Sub-challenges

- Four disease areas were selected:
  - Psoriasis (Diagnosis)
  - Multiple Sclerosis (Diagnosis or Stage)
  - Lung Cancer (Diagnosis and Stage)
  - Chronic Obstructive Pulmonary Disease (Diagnosis)
- Scoring was made against unpublished Gold Standards.

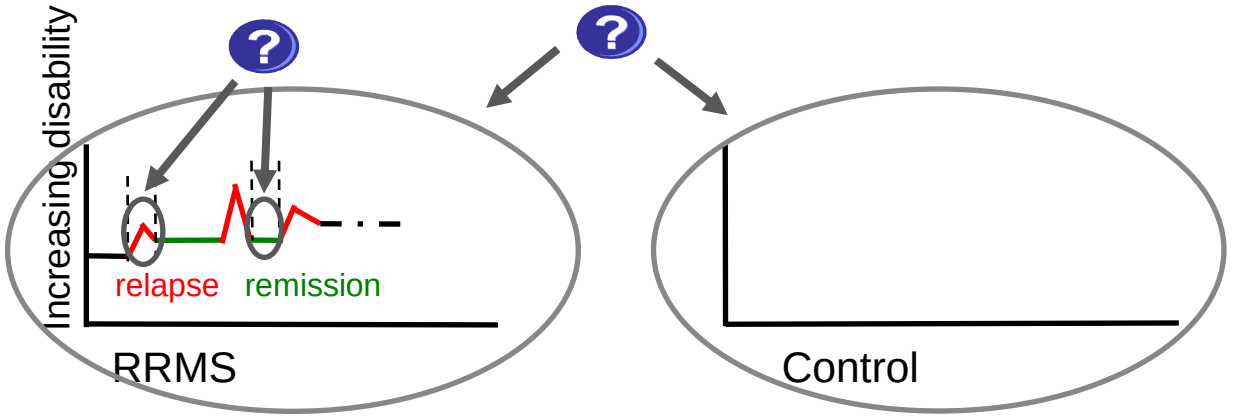
Psoriasis Challenge: Identify normal vs psoriatic skin based only on the transcriptome of a skin biopsy



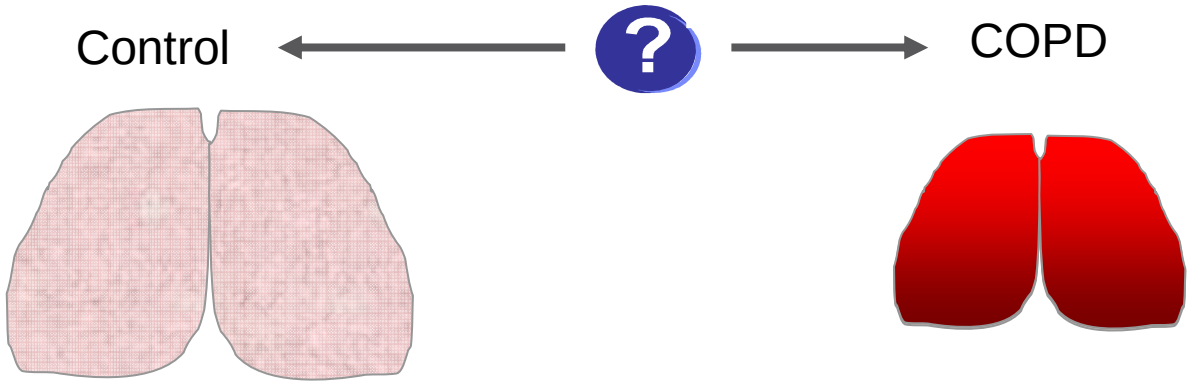
Lung Cancer Challenge: Identify SQCC and AdenoCarcinoma stages based the tumor transcriptome



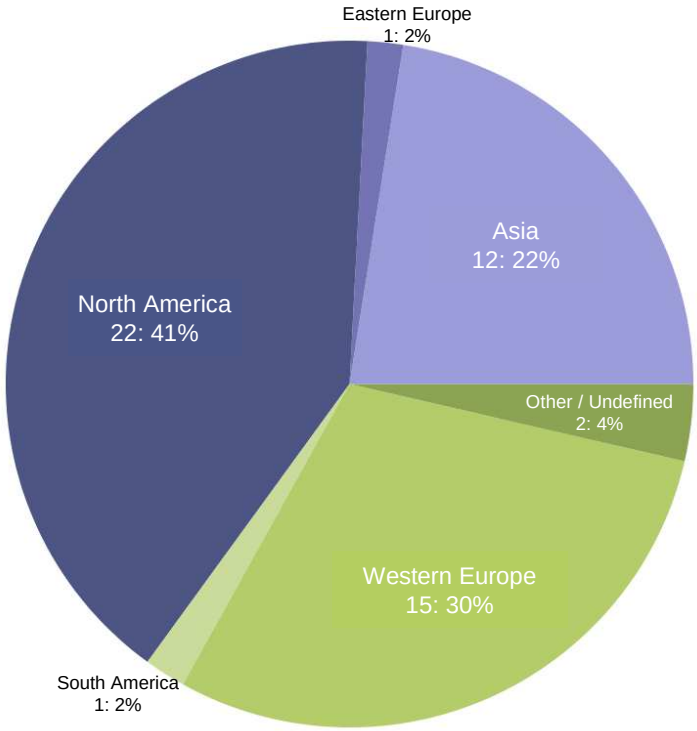
MS Challenge: Identify control vs. affected or remitting vs. relapsing patients, using Peripheral Blood Mononuclear cells.



Chronic Obstructive Pulmonary Disease (COPD) Challenge: Identify affected vs non-affected subjects based the tumor large airways transcriptome



Diagnostic Signature Challenge: overall participation



54 Teams from around the world participated in the Diagnostics Signature Challenge.

Diagnostic Signature Challenge participation

Submissions are spread evenly across all five sub-challenges:

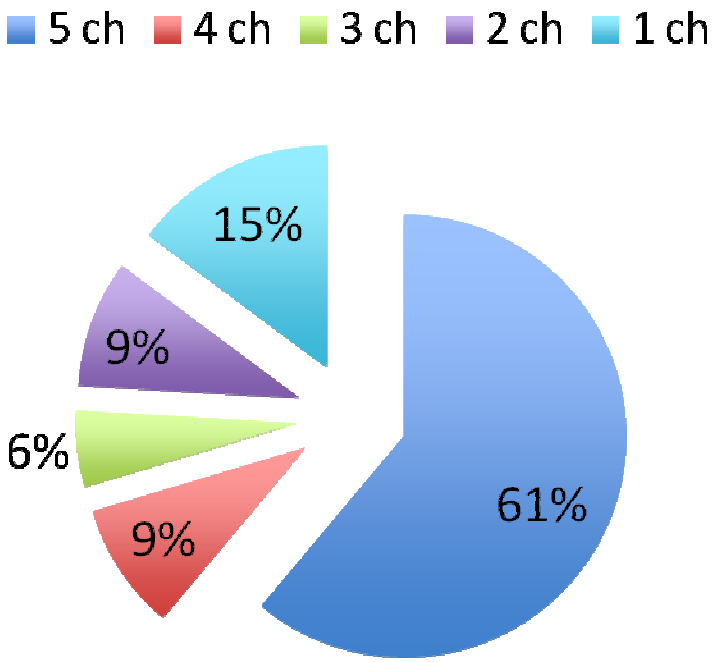
Psoriasis: 49 participants

COPD: 40 participants

Lung Cancer: 46 participants

MS Diagnosis: 40 participants

MSS Staging: 39 participants



Most teams submitted predictions to all challenges

Overview of the Lessons from the Diagnostics Signature Challenge

- There is no one-size-fits-all method for classifying disease:
  - No single normalization method conferred a performance advantage
  - No single classification method conferred a performance advantage
  - The specifics of the methodology used to classify disease seems to be decisive in extracting signal from the data
- If the signal is strong, most methods will get the classification right, as was the case with Psoriasis.
- We can determine that the signal is too weak or inexistent, by finding that statistical significance was not attained by any prediction, as was the case with MS Stages.
- When the signal is faint, the method used can be decisive. Crowd-sourcing is particularly relevant in these cases (COPD).
- The advantage of having many participants can be offset by the multiple testing problem that ensues
- The wisdom of crowds enhances the performance at least from the perspective of one of the performance metrics
- It is important to keep the test set data from the participants to better represent the situation at the clinic
- Many of these lessons learned are consistent with the conclusions reached in the MACQ-II study (2010) to be discussed in a forthcoming session

Can we define the optimal method on * after this challenge?