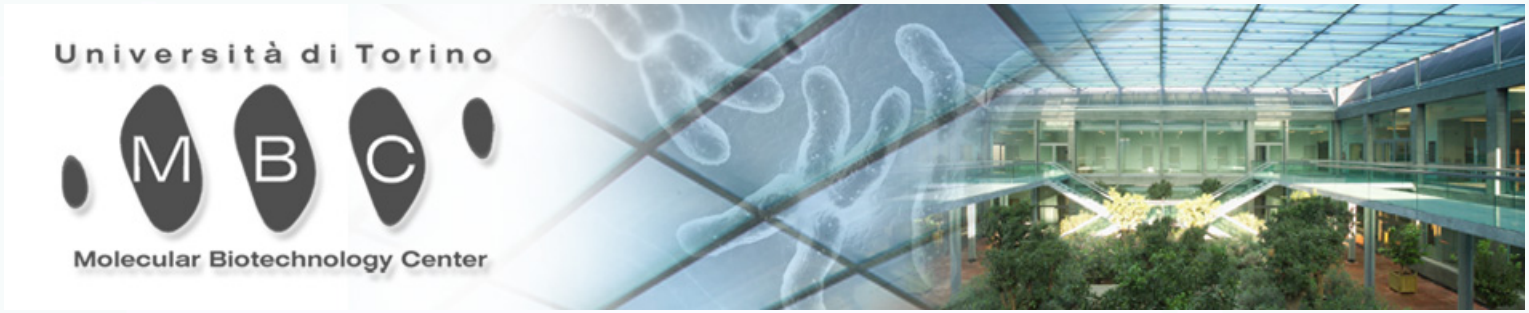


Dreaming about models: a biologist's perspective

raffaele.calogero@unito.it

Bioinformatics & Genomics unit

Agenda

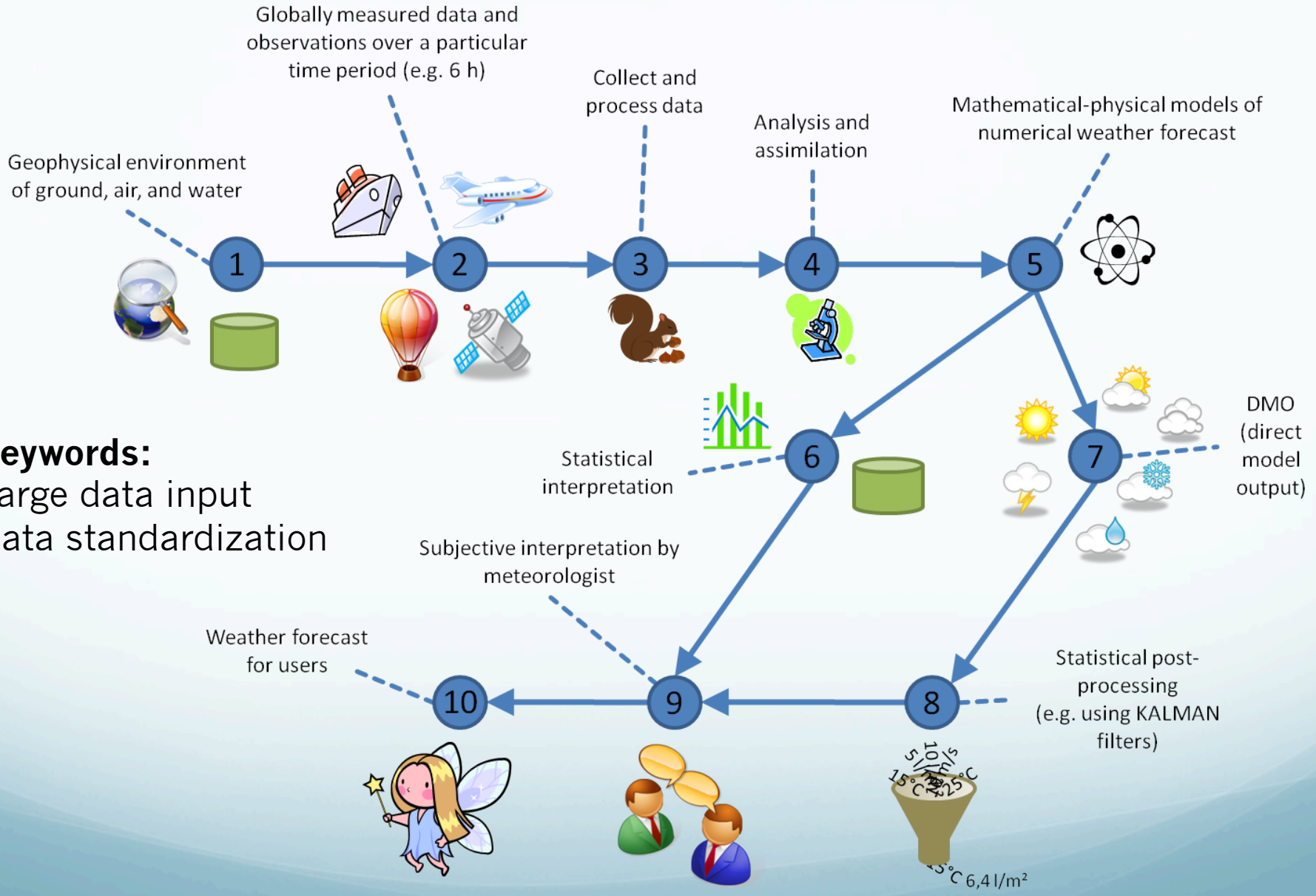
- About the biology domain
- How a biologist designs an experiment
- What a biologist would like to get from a model
- What a biologist could really provide for model design and development

Agenda

- **About the biology domain**
- How a biologist designs an experiment
- What a biologist would like to get from a model
- What a biologist could really provide for model design and development

Models in daily life

Weather prediction



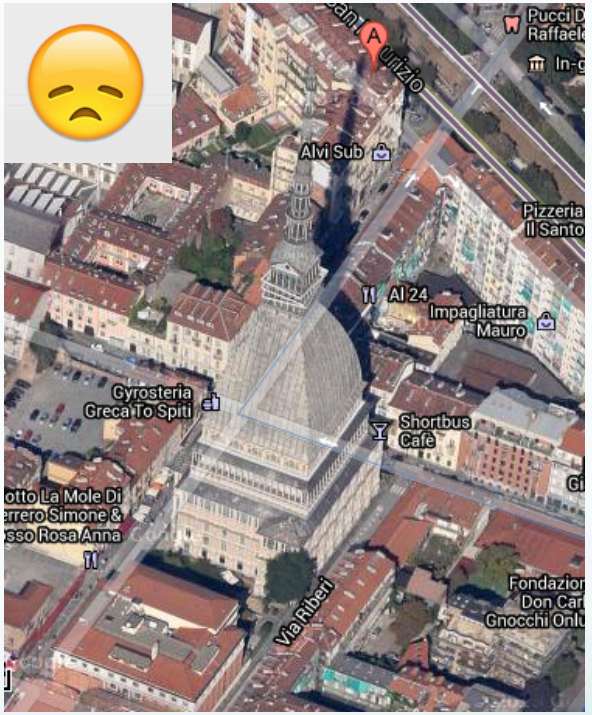
Keywords:
Large data input
Data standardization

Weather prediction

- Why is prediction successful?
  - We have good numerical models predicting the weather changes.
  - We have a significant set of measurable standardized parameters to feed the models.
- Type of prediction:
  - Weather development

Weather prediction

- Does weather forecast prediction work at any space scale?

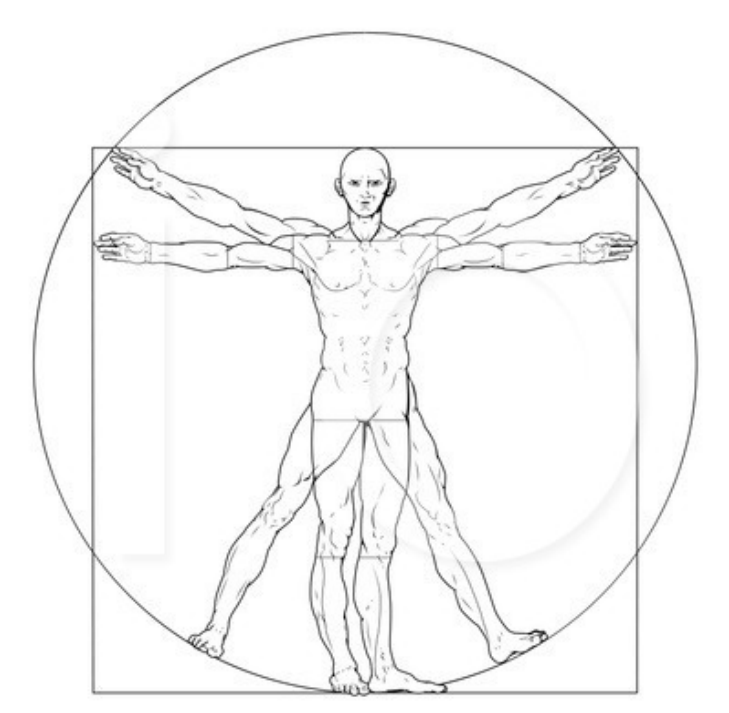


- Space scale is important!

Can biological processes be modeled?

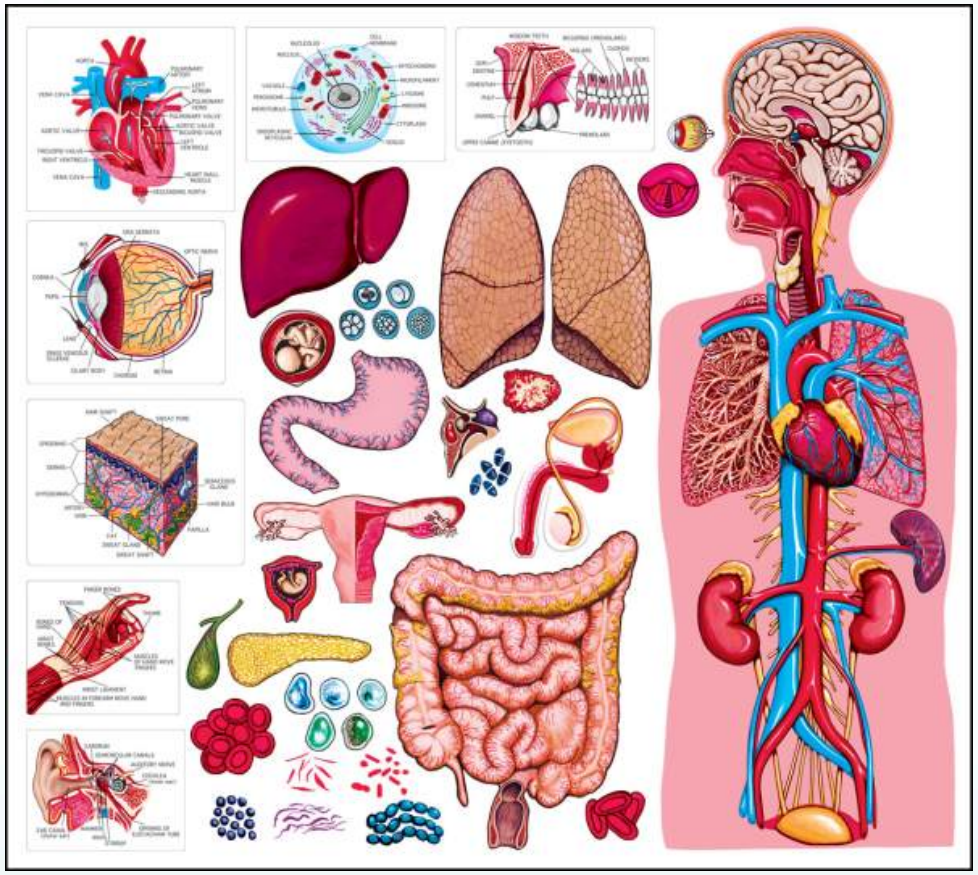
- Over the past 10-20 years, biology has become increasingly quantitative, and mathematical sciences have in turn been increasingly influenced by biology.
- However to be able to do really quantitative biology:
  - Questions need to be addressed in the right biology space
  - Data need to be provided in a sufficient amount
  - Analytical methodologies need to be standardized

Biology model



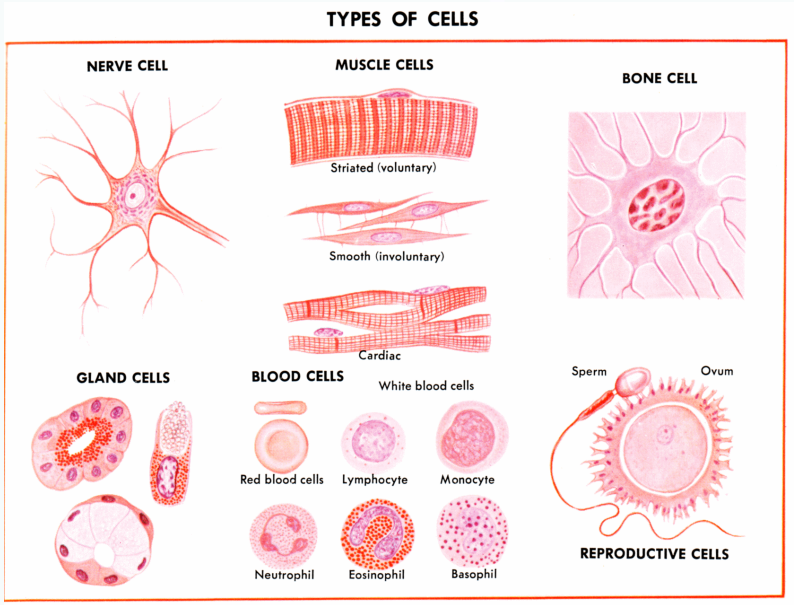
Body space

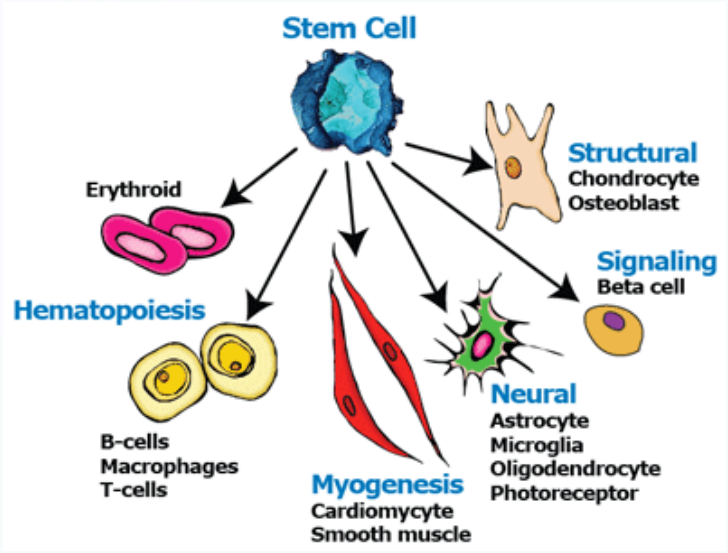
Biology model

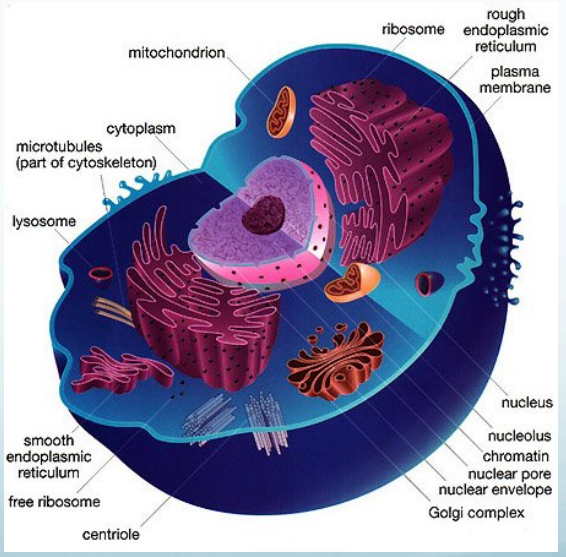


Organ space

Biology model

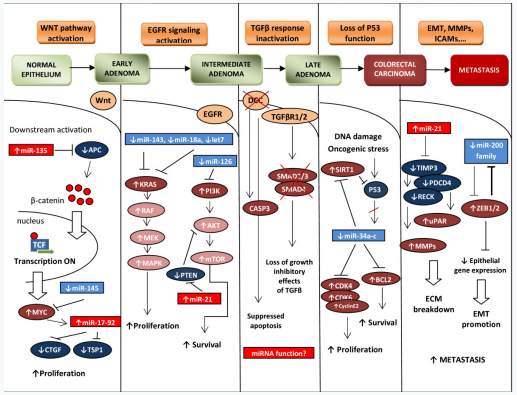




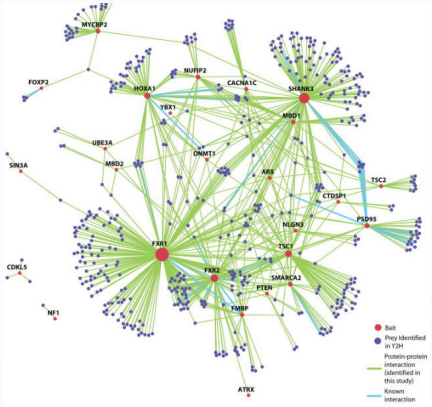


Cell & tissue spaces

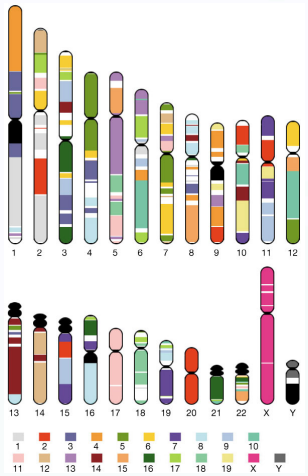
Biology model



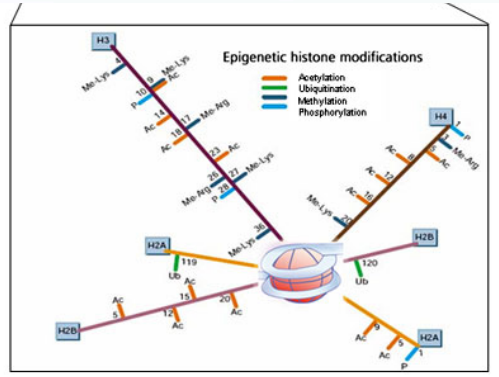
signal transduction



protein-protein interaction

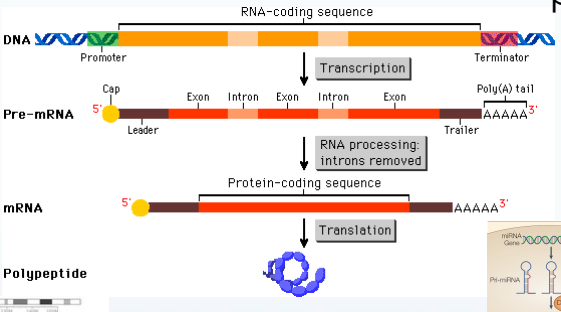


genome primary structure

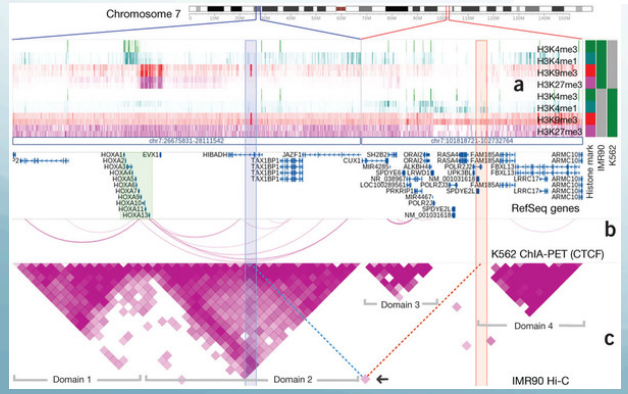


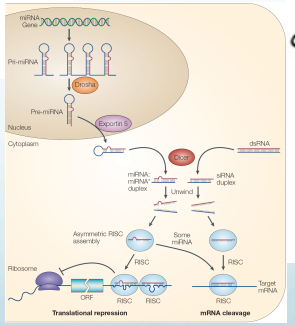
epigenetics

genome long-range interactions

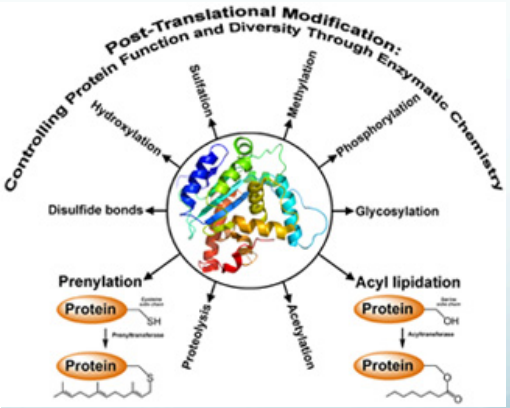


expression regulation





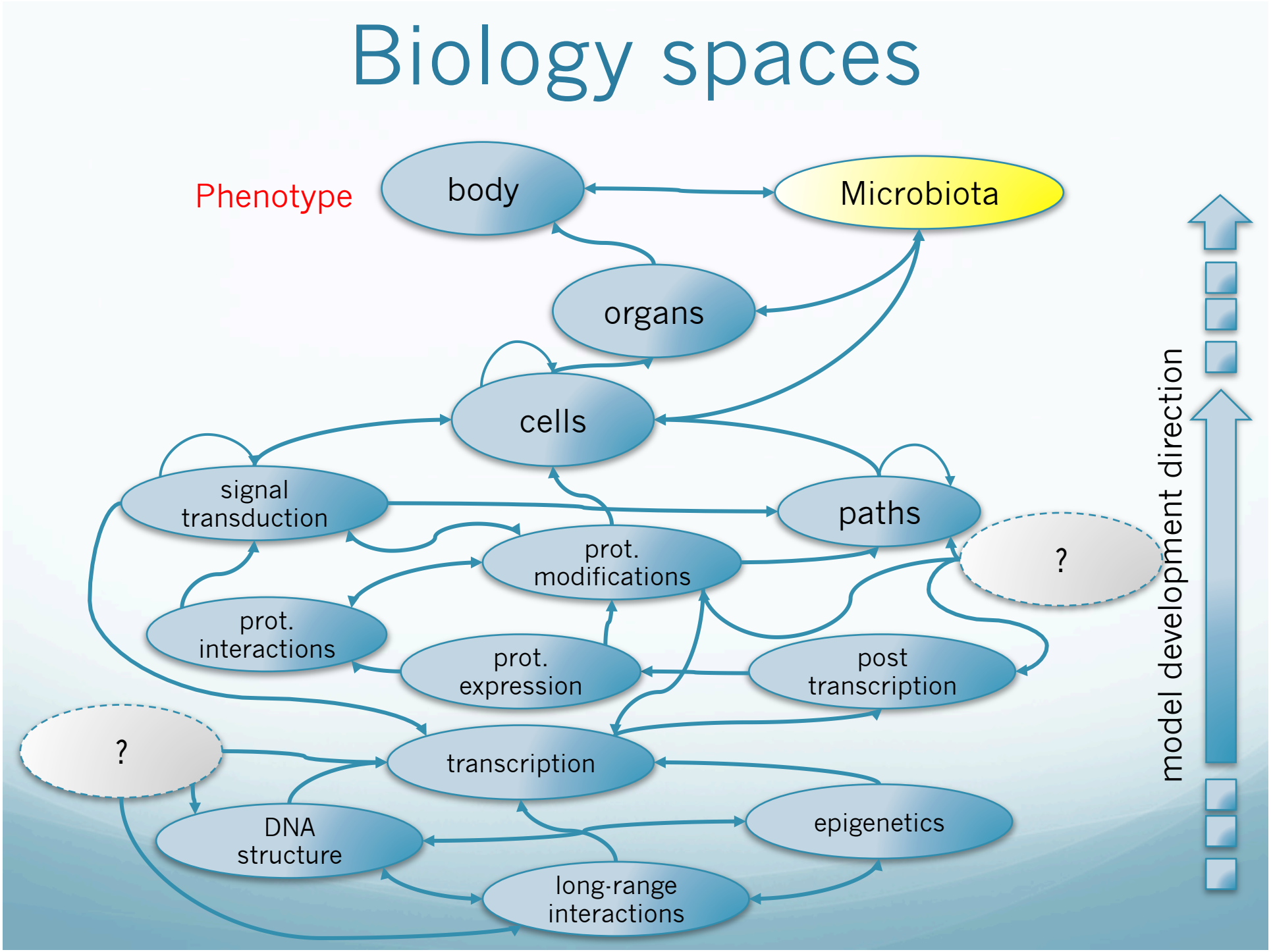
post-transcription regulation



post-traslation modification

Molecular spaces

Biology spaces



Agenda

- About the biology domain
- **How a biologist designs an experiment**
- What a biologist would like to get from a model
- What a biologist could really provide for model design and development

Biological research strategies

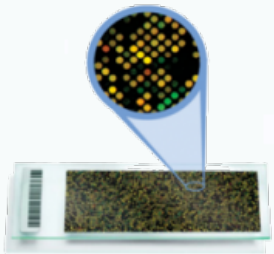
- How a biological problem is defined?
  - A hypothesis is proposed.
  - Experiments are designed to validate hypothesis:
    - Budget phase
    - Technology phase
    - Discovery phase
    - Validation phase

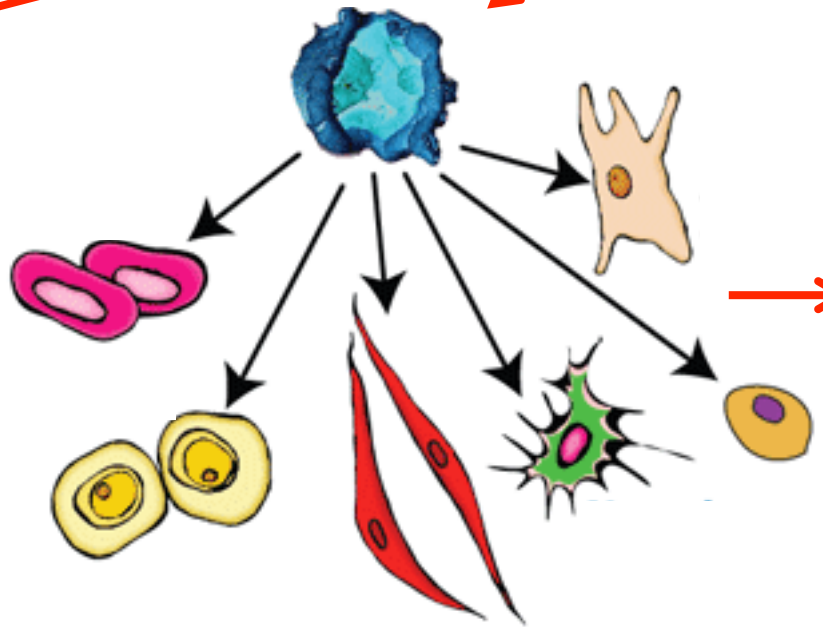
Budget phase

- This issue is part of every biological project in daily life.
- Biological experiments are getting more and more expensive.
- Technology offers constantly new opportunities to have more robust and effective experiments, but not always it is possible to use new methodologies because of their intrinsic cost.

Technology phase

Serial approach



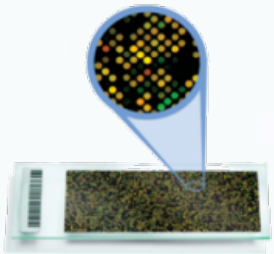


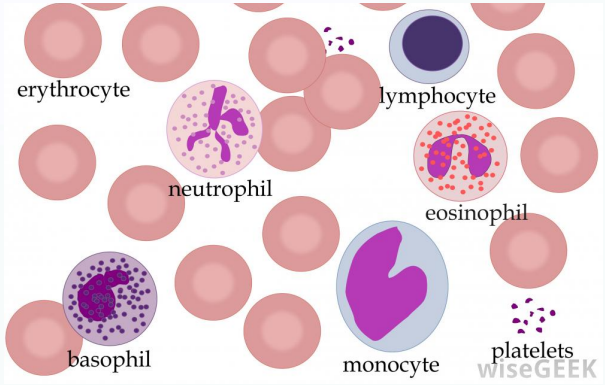
up

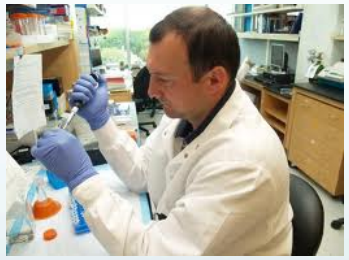
bottom

Technology phase

Parallel approach

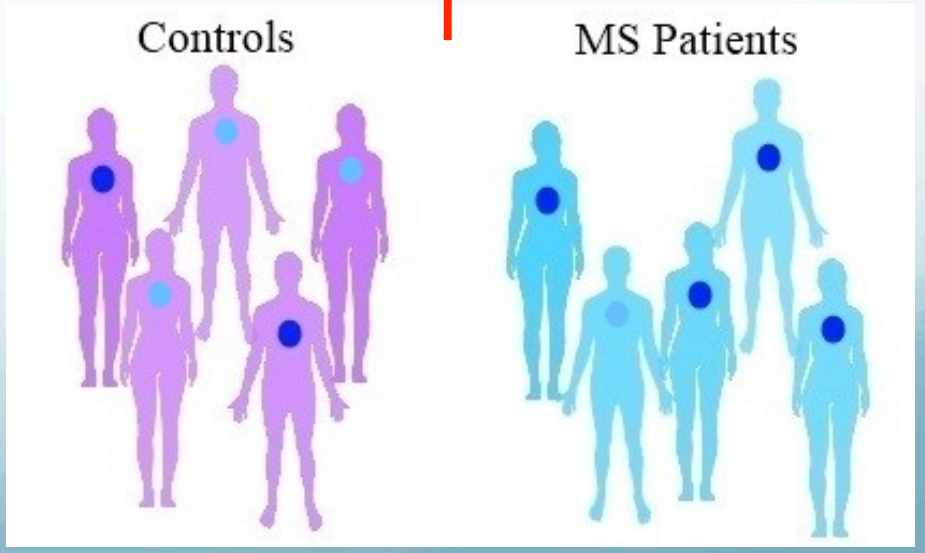






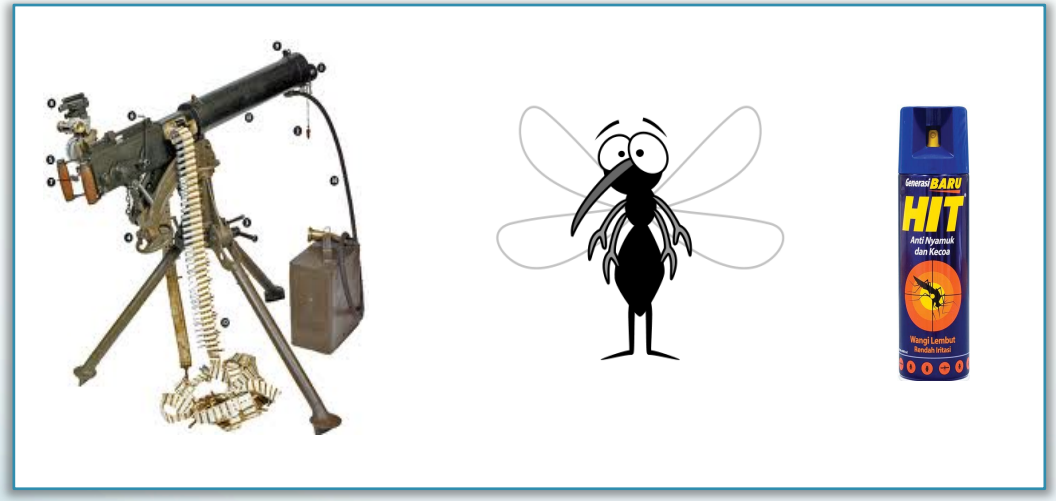
top

down

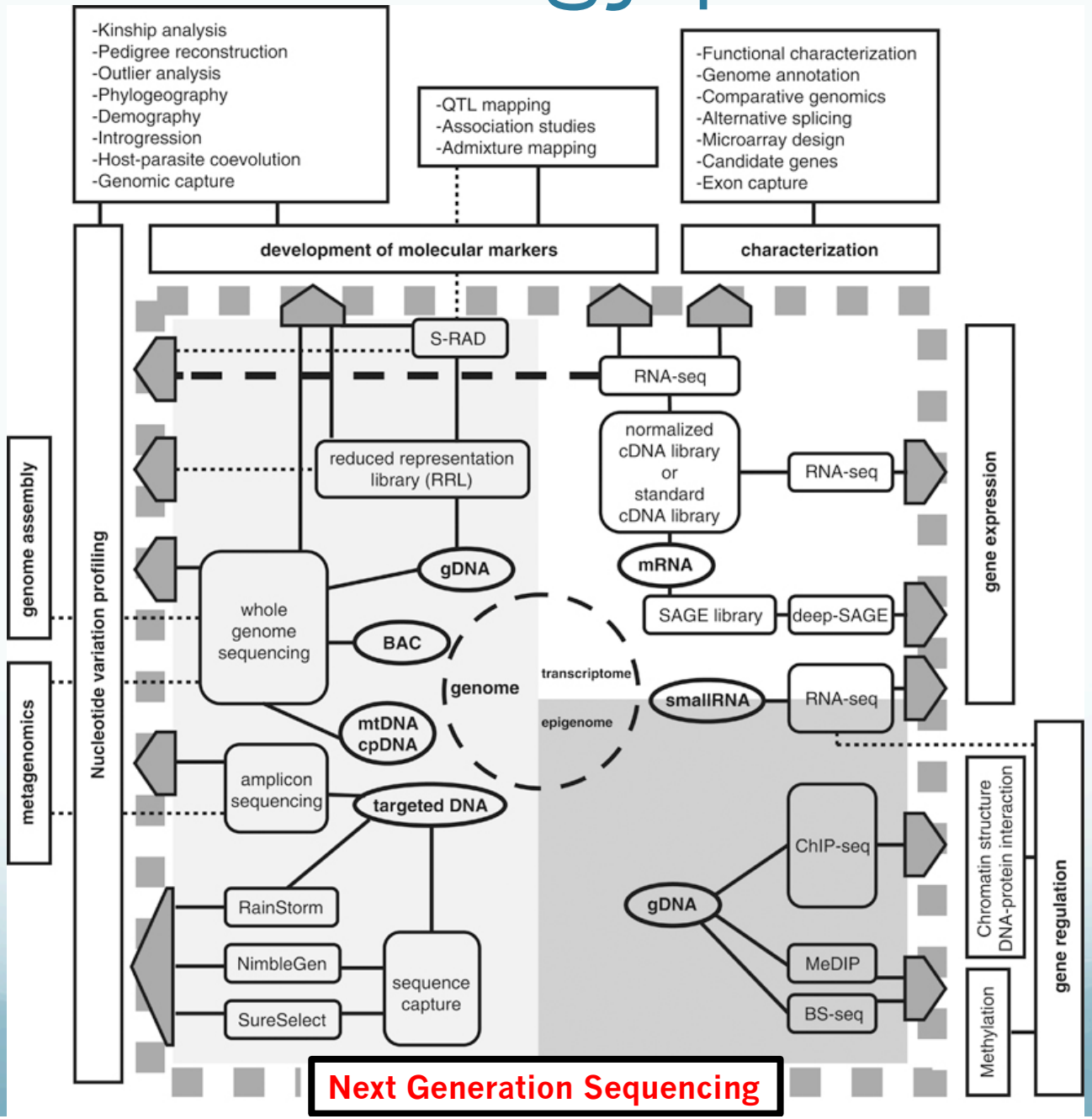


Technology phase

- Since the serial and parallel approach have similar time frame, why not using always the parallel approach?

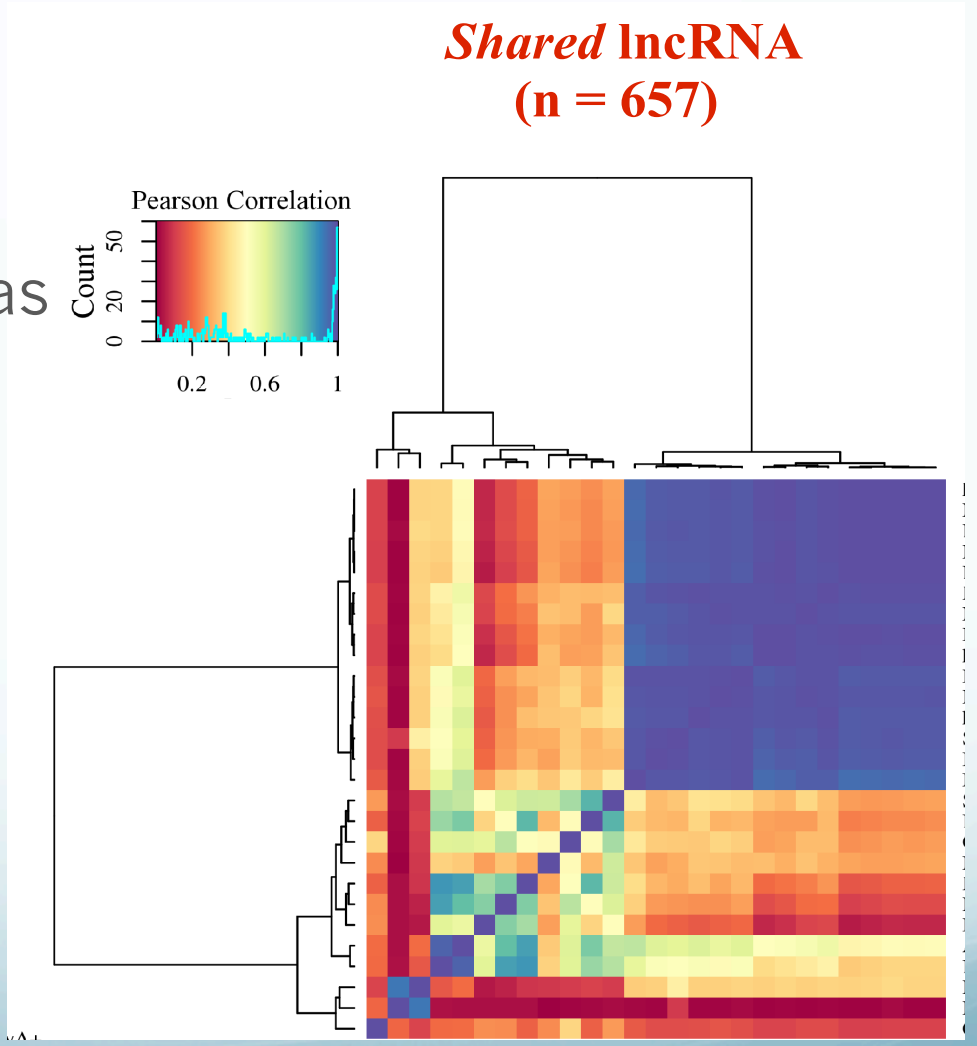


Technology phase



Technology phase

- Technologies have limitations!
- Methodological bias



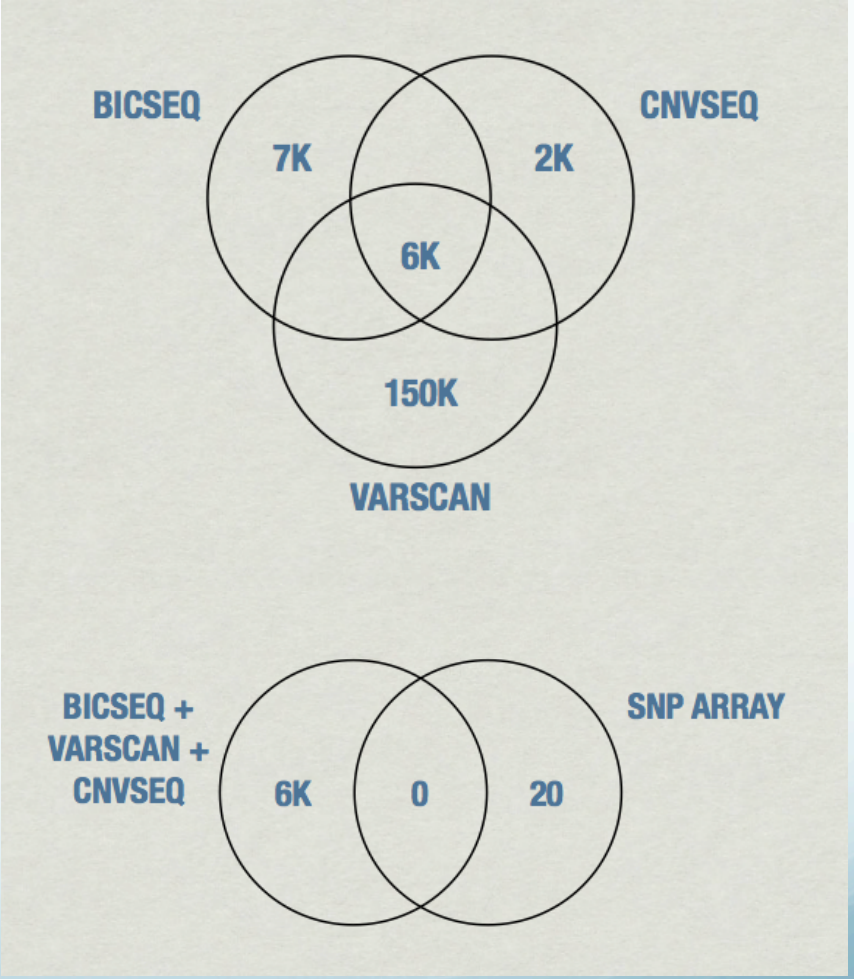
Technology phase

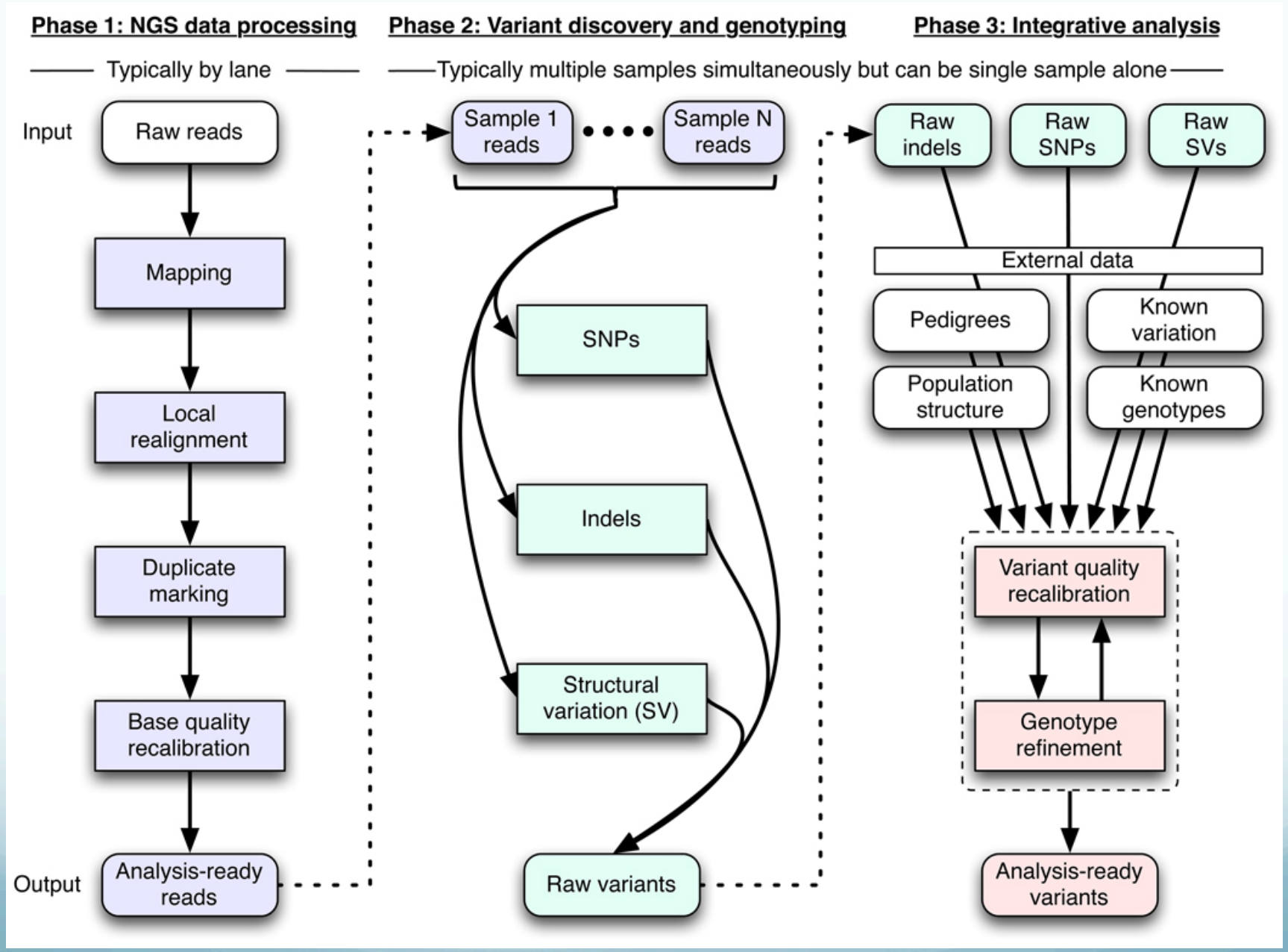
- Technologies have

  limitations!

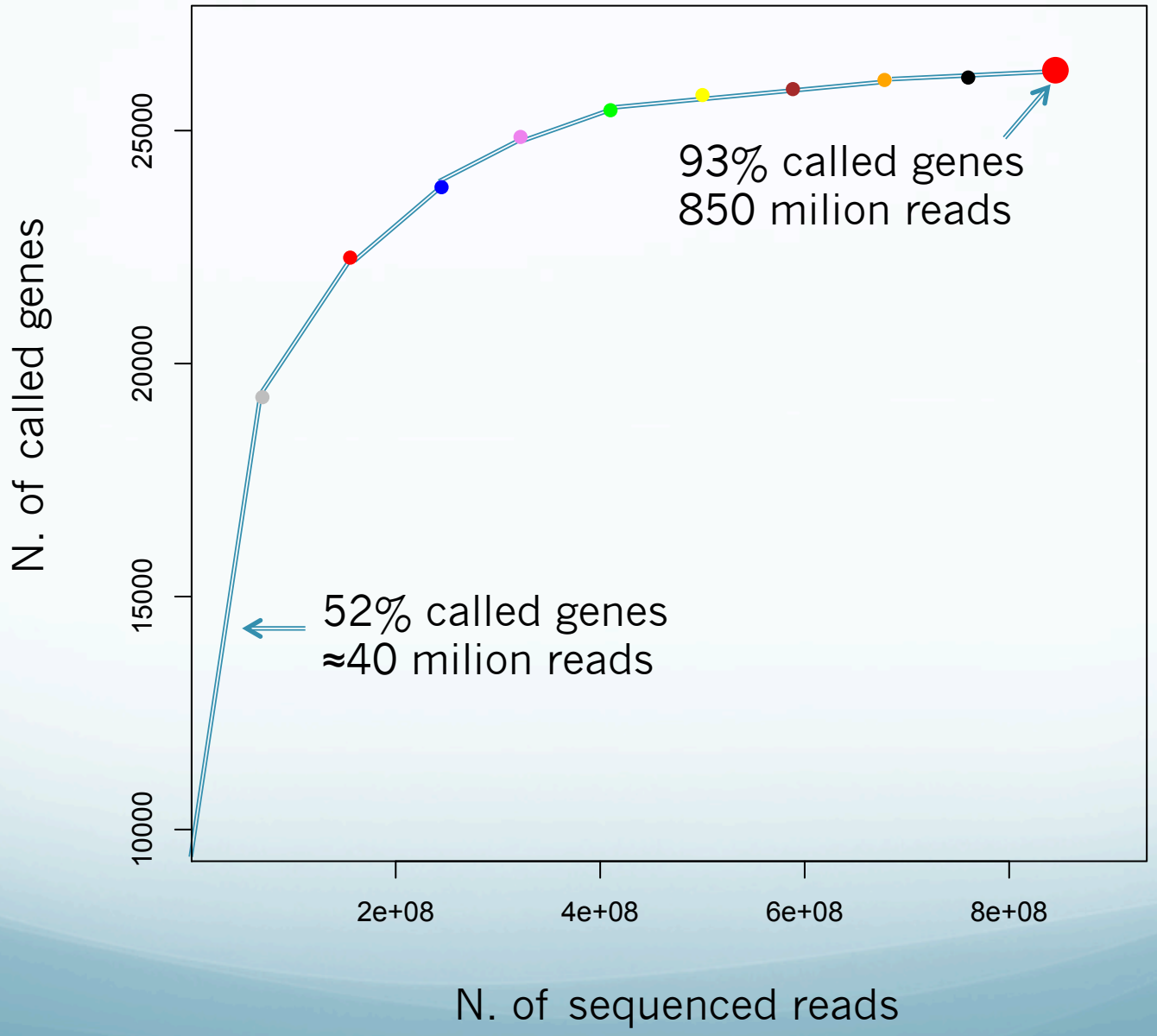
- Lack of standardized methods

- Inconsistency between technologies



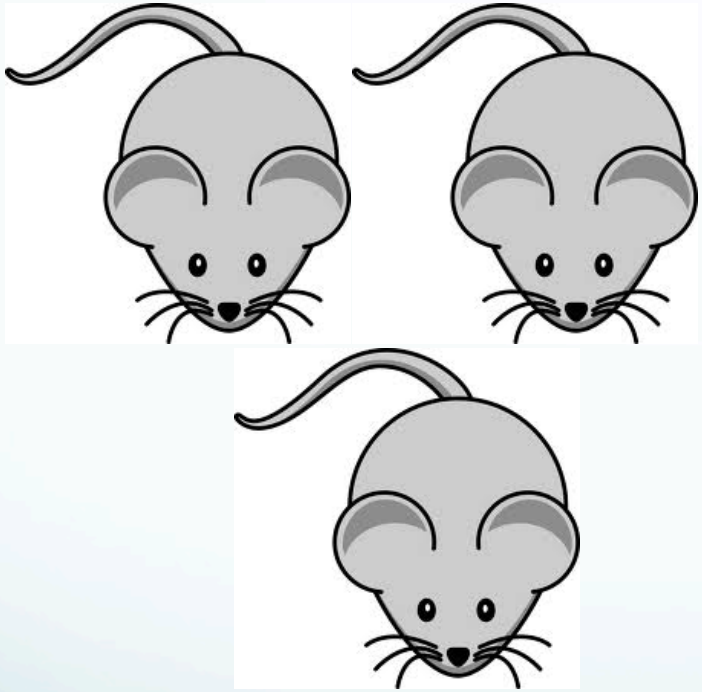

Phase 1: NGS data processing
Phase 2: Variant discovery and genotyping
Phase 3: Integrative analysis
Typically by lane
Typically multiple samples simultaneously but can be single sample alone
Input
Raw reads
Mapping
Local realignment
Duplicate marking
Base quality recalibration
Output
Analysis-ready reads
Sample 1 reads
Sample N reads
SNPs
Indels
Structural variation (SV)
Raw variants
Raw indels
Raw SNPs
Raw SVs
External data
Pedigrees
Known variation
Population structure
Known genotypes
Variant quality recalibration
Genotype refinement
Analysis-ready variants

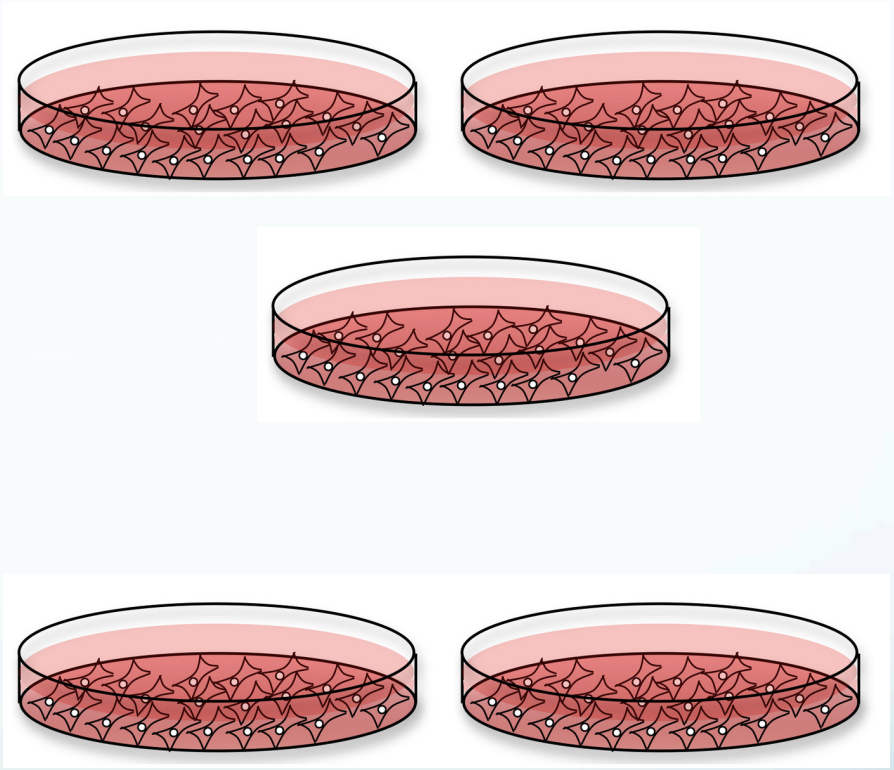
Limited data sampling



mm9

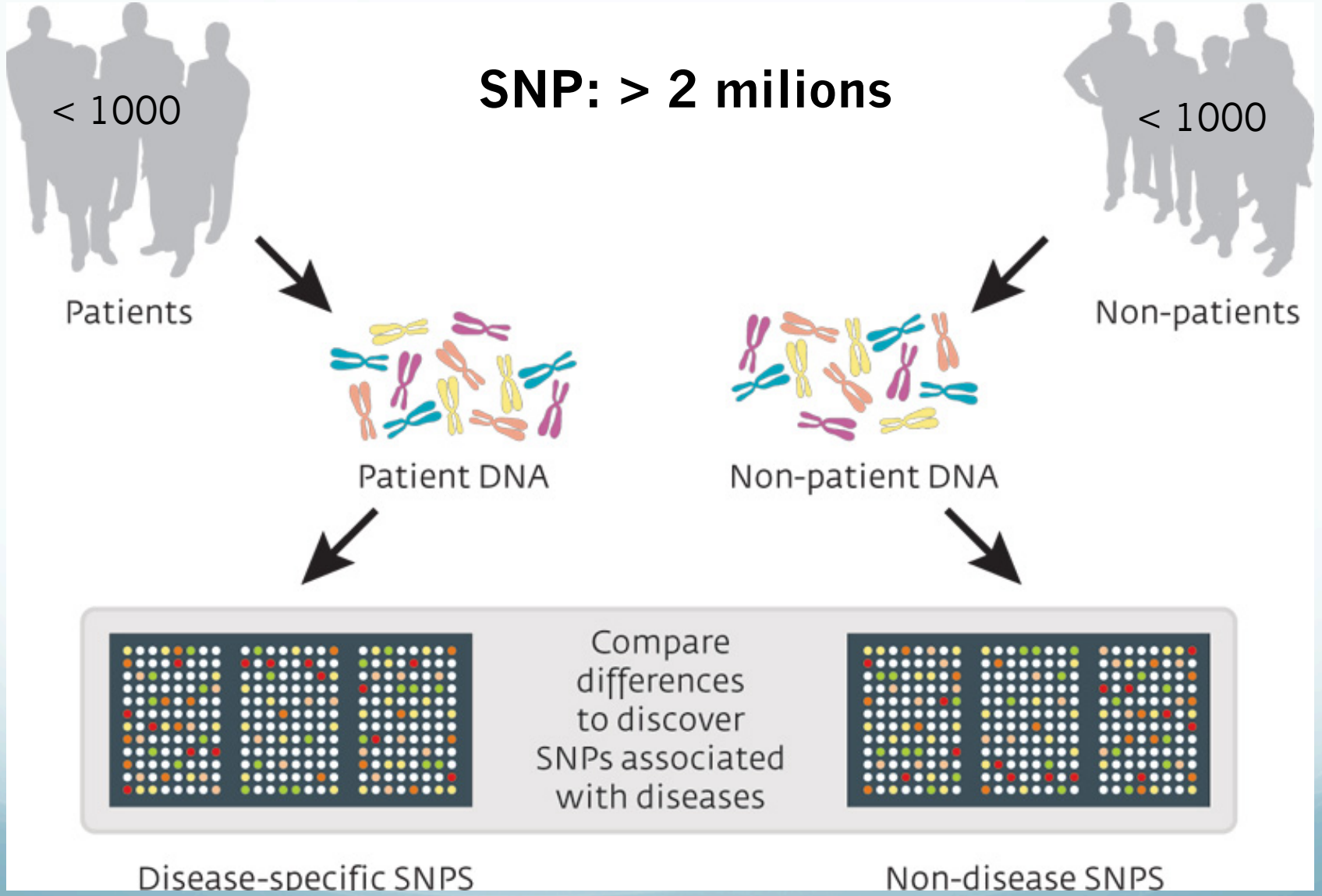
Limited sample sampling





Genes: > 25 K

Limited sample sampling



Discovery phase

- If top ➔ down approach is used, data can have different in depth depending on the lab producing them.
- **Small biological laboratories:**
  - Data are limited:
    - Time series: 4 points
    - Perturbations: 1 perturbation
    - Number of replicates: 2:3
    - In some cases difficult accessibility to raw data.
- **Public consortiums:**
  - High quality data
  - Large amount of data
  - In some cases difficult accessibility to raw data.

Validation phase

- After having defined some key elements for a specific biological problem various experiments are designed to confirm and motivate the involvement of the key elements in the problem under study.
  - Time frame: 2-3 yrs

Agenda

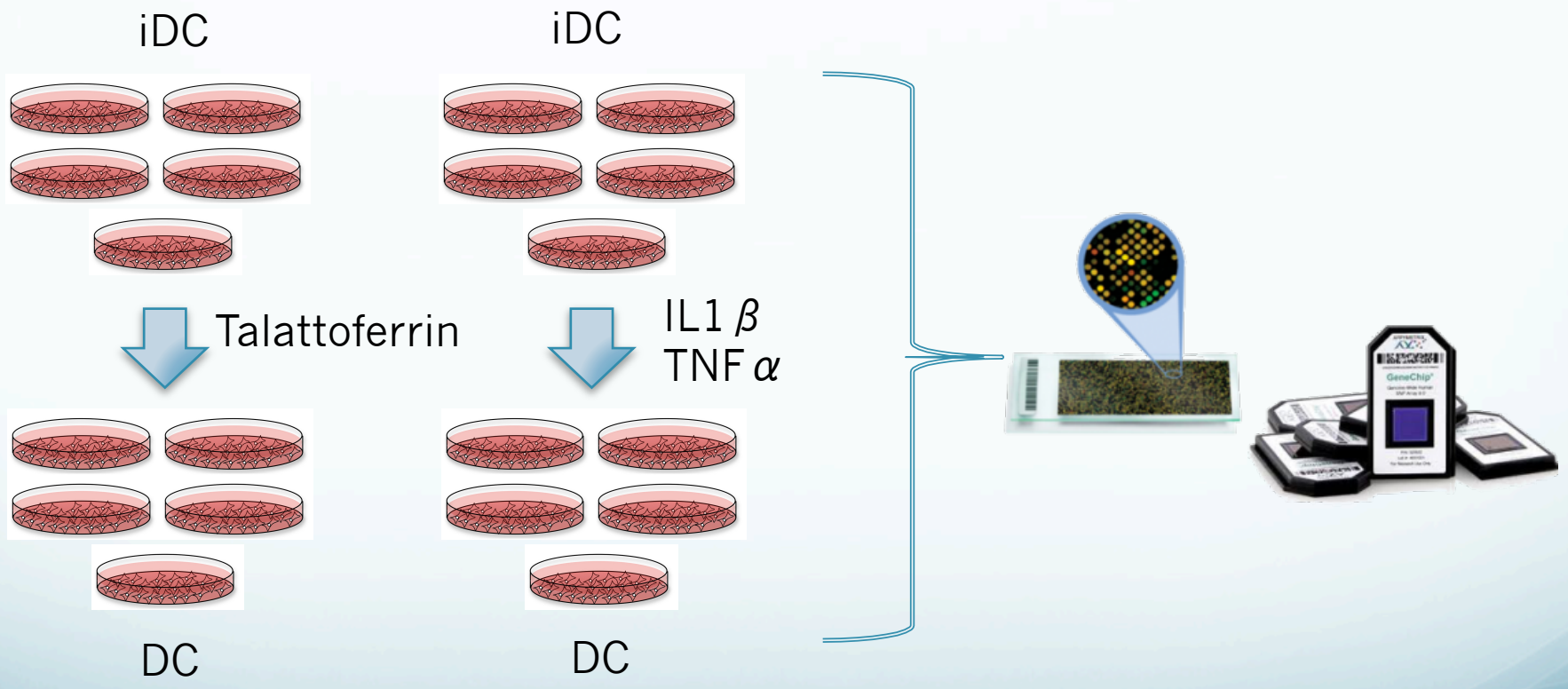
- About the biology domain
- How a biologist designs an experiment
- **What a biologist would like to get from a model**
- What a biologist could really provide for model design and development

Models in biology

- **What do we want to predict?**
  - Disease appearance
  - Disease development
  - Drug effect
- **What do we want to study?**
  - Effect of perturbations on molecular spaces affecting at least cell space
  - Understanding cell mechanisms:
    - Development
    - Differentiation
- **What do we want to detect?**
  - Key elements involved in disease development
  - Key elements involved in molecular pathways

Biologist's model idea

- **Key element involved in a molecular pathway**

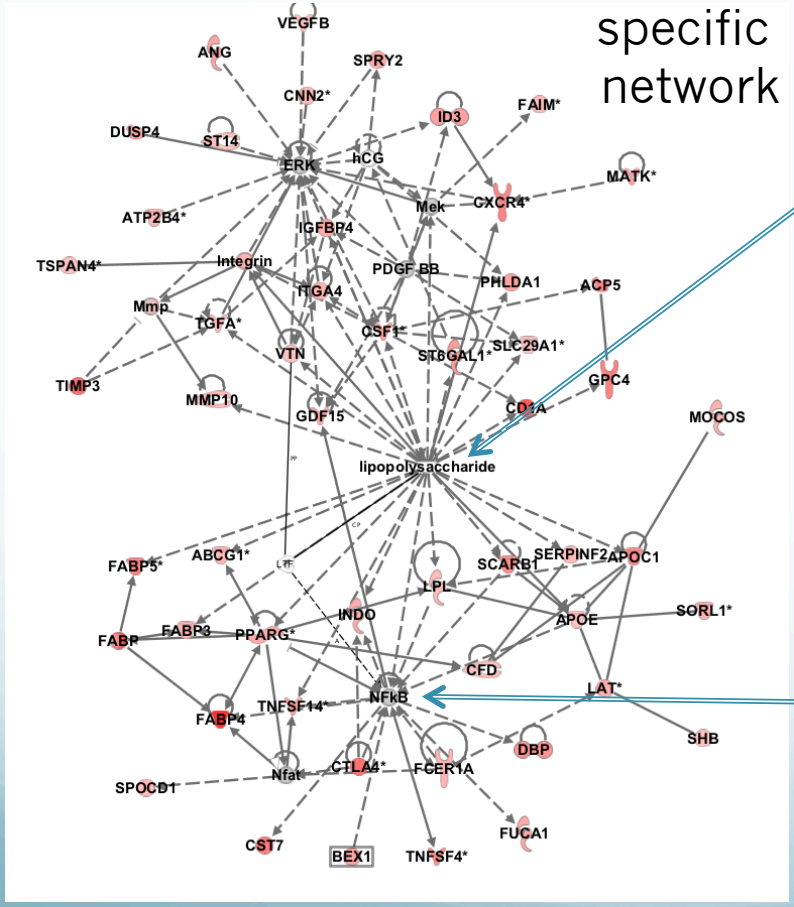


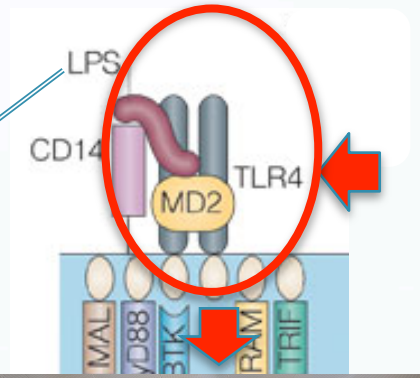
Spadaro et al. submitted

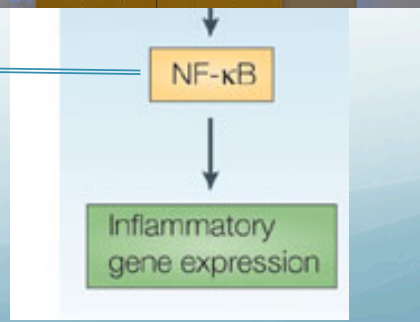
Biologist's model idea

- Models represent a way to define new working hypotheses.

Talattoferrin specific network



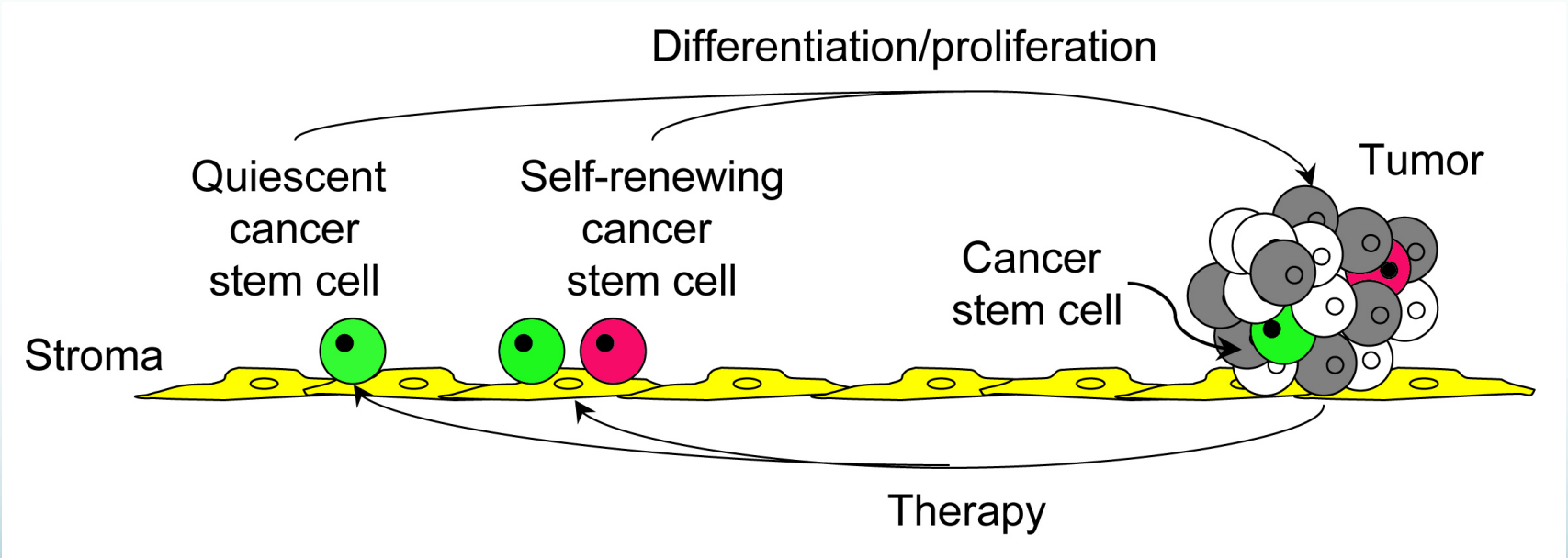




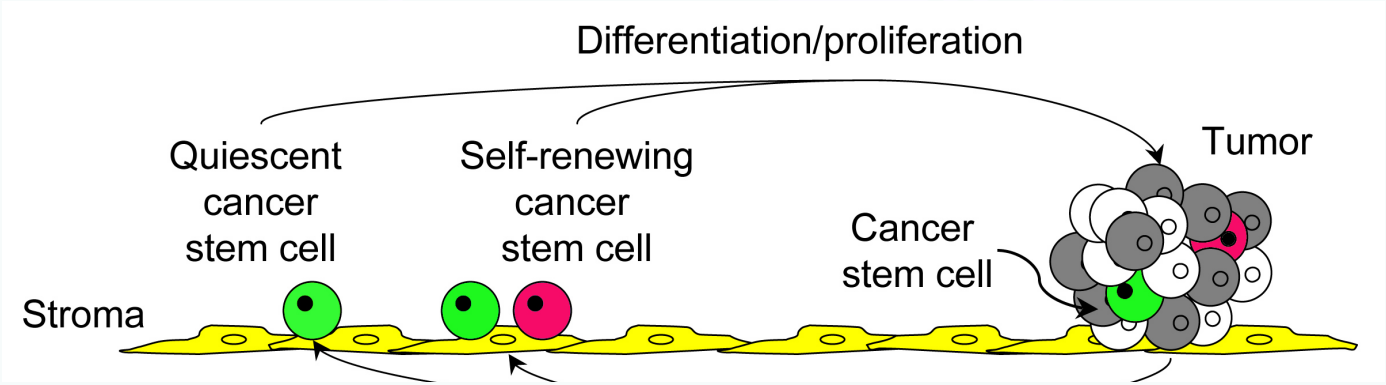
Spadaro et al. submitted

Models in biology

- Particularly intriguing are multi-level models:
  - Cell population
  - Molecular networks controlling cell population



Models in biology



$$\frac{dNcsc}{dt} = P_{sy}\omega_{CSC}N_{CSC} + \gamma_{PC}\sum_{j=1}^{3} N_{PC_j} - \eta_1 N_{CSC} - d_1 N_{CSC}$$

$$\frac{dN_{PC_1}}{dt} = P_{asy}\omega_{CSC}N_{CSC} - \omega_{PC}N_{PC_1} - \gamma_{PC}N_{PC_1} - \eta_1 N_{CSC} - \eta_2 N_{PC_1} - d_2 N_{PC_1}$$

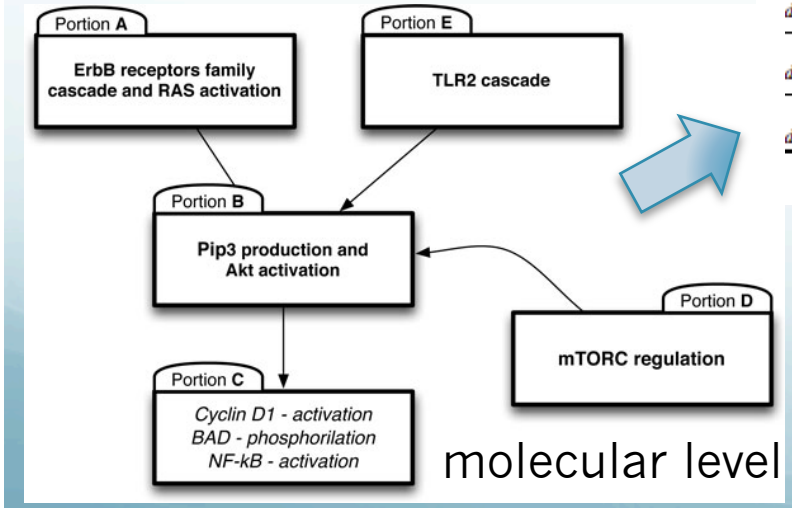
$$\frac{dN_{PC_j}}{dt} = 2\omega_{PC}N_{PC_{j-1}} - \omega_{PC}N_{PC_j} - \gamma_{PC}N_{PC_j} + \eta_2 N_{PC_{j-1}} - \eta_2 N_{PC_j} - d_2 N_{PC_j} \quad j = 2 \dots 3$$

$$\frac{dN_{PC_i}}{dt} = 2\omega_{PC}N_{PC_{i-1}} - \omega_{PC}N_{PC_i} + \eta_2 N_{PC_{i-1}} - \eta_2 N_{PC_i} d_2 N_{PC_i} \quad i = 4 \dots 6$$

$$\frac{dN_{PC_7}}{dt} = 2\omega_{PC}N_{PC_6} + \eta_2 N_{PC_6} - \eta_3 N_{PC_7} - d_2 N_{PC_7}$$

$$\frac{dN_{TC}}{dt} = \eta_3 N_{PC_7} - d_3 N_{TC}$$

population level

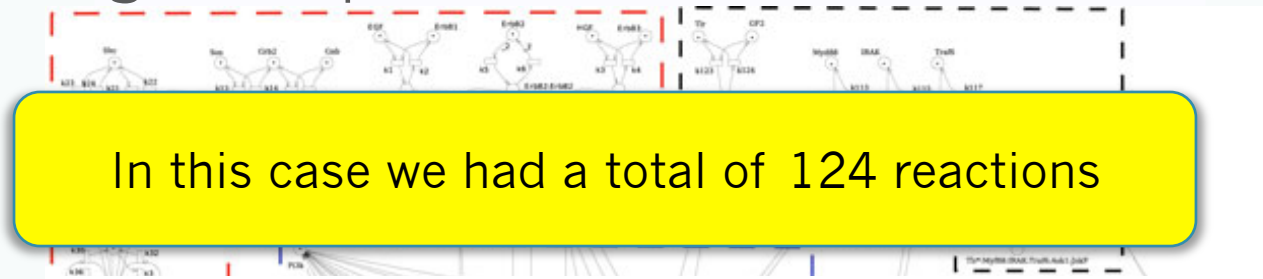


molecular level

Cordero et al. BMC Bioinformatics 2013, 14(Suppl 6):S11

Parameter definition

- The critical issue that require interaction with biologists is parameters definition



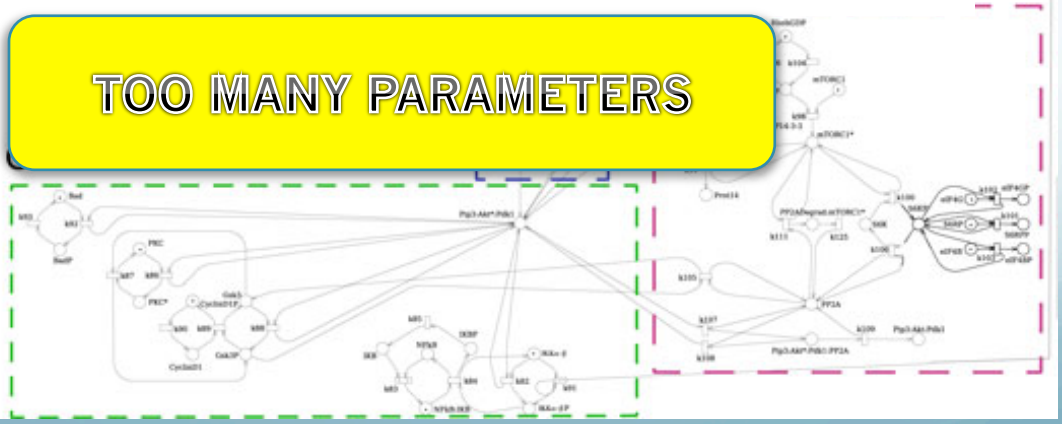
In this case we had a total of 124 reactions

Molecular Systems Biology **3** Article number: 144 doi:10.1038/msb4100188
Published online: 13 November 2007
Citation: *Molecular Systems Biology* 3:144

Ligand-dependent responses of the ErbB signaling network: experimental and modeling analyses

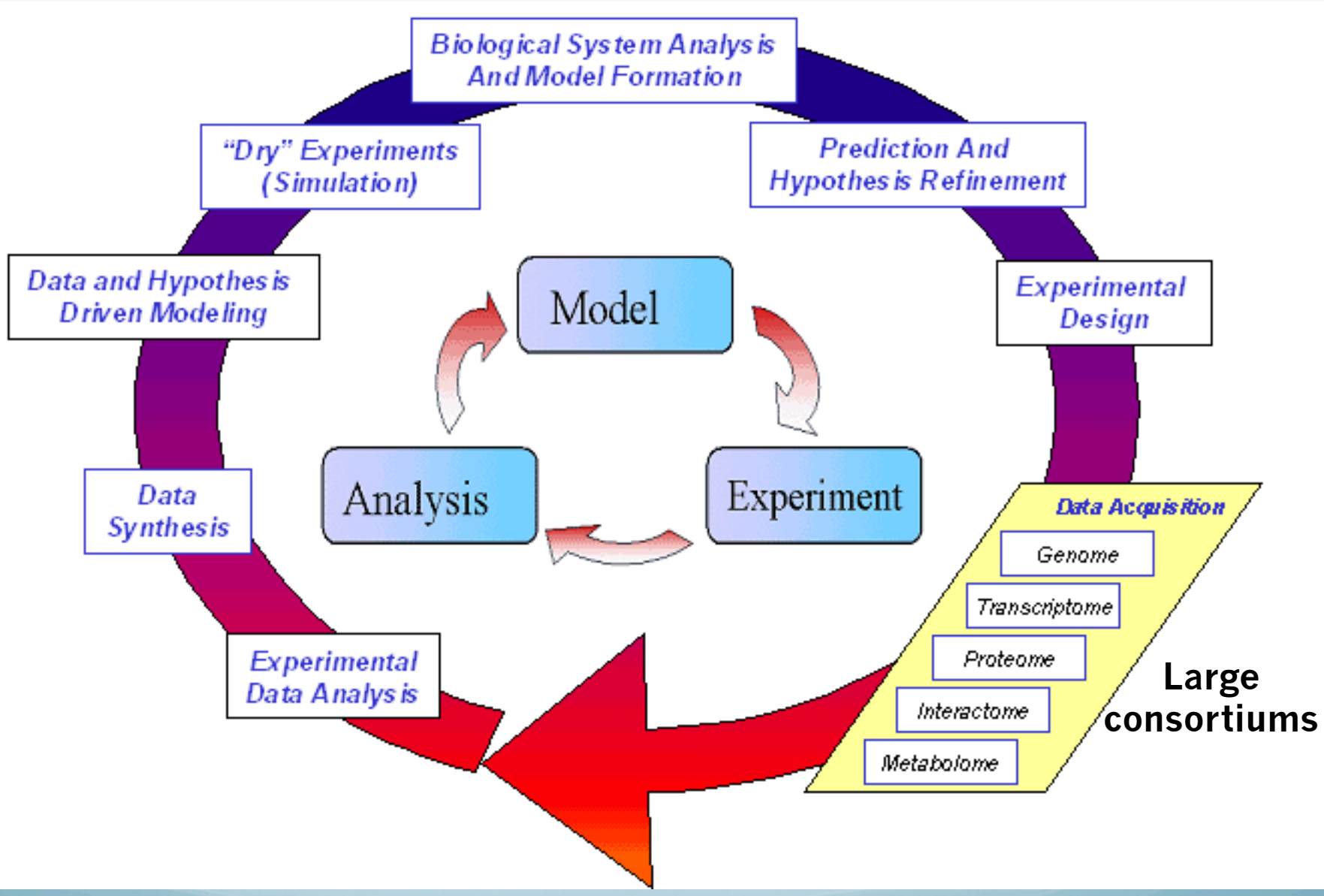
Marc R Birtwistle[1,2,a], Mariko Hatakeyama[3,a], Noriko Yumoto[3], Babatunde A Ogunnaike[1], Jan B Hoek[2] & Boris N Kholodenko[2]

TOO MANY PARAMETERS



Agenda

- About the biology domain
- How a biologist designs an experiment
- What a biologist would like to get from a model
- **What a biologist could really provide for model design and development**

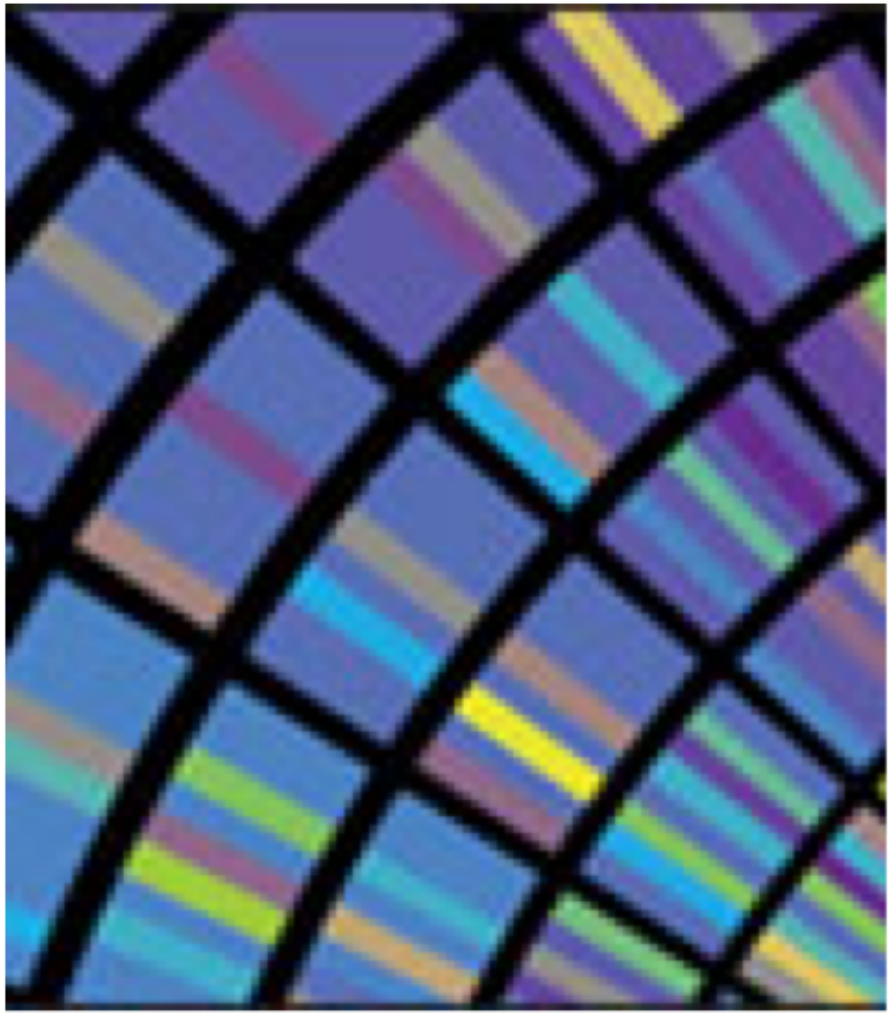


Large consortiums

Encode

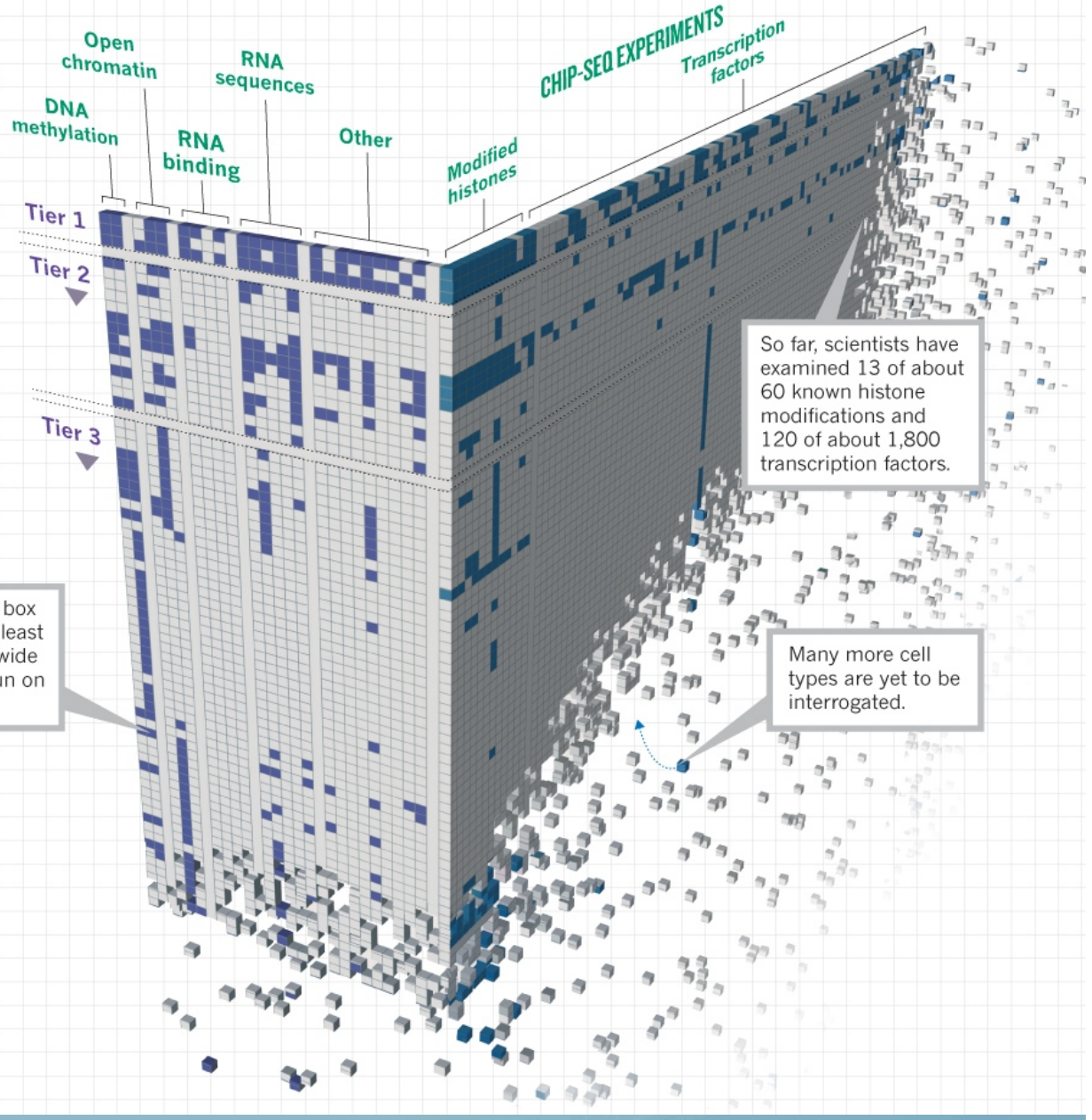
- 32 Institutes
- 442 Consortium Members
- 1649 Experiments
- 11,972 files
- 15TB of disk space

- 80% of genome participates in at least one biochemical and/or chromatin event in at least 1 cell type



MAKING A GENOME MANUAL

Scientists in the Encyclopedia of DNA Elements Consortium have applied 24 experiment types (across) to more than 150 cell lines (down) to assign functions to as many DNA regions as possible — but the project is still far from complete.



EXPERIMENTAL TARGETS

DNA methylation: regions layered with chemical methyl groups, which regulate gene expression.

Open chromatin: areas in which the DNA and proteins that make up chromatin are accessible to regulatory proteins.

RNA binding: positions where regulatory proteins attach to RNA.

RNA sequences: regions that are transcribed into RNA.

ChIP-seq: technique that reveals where proteins bind to DNA.

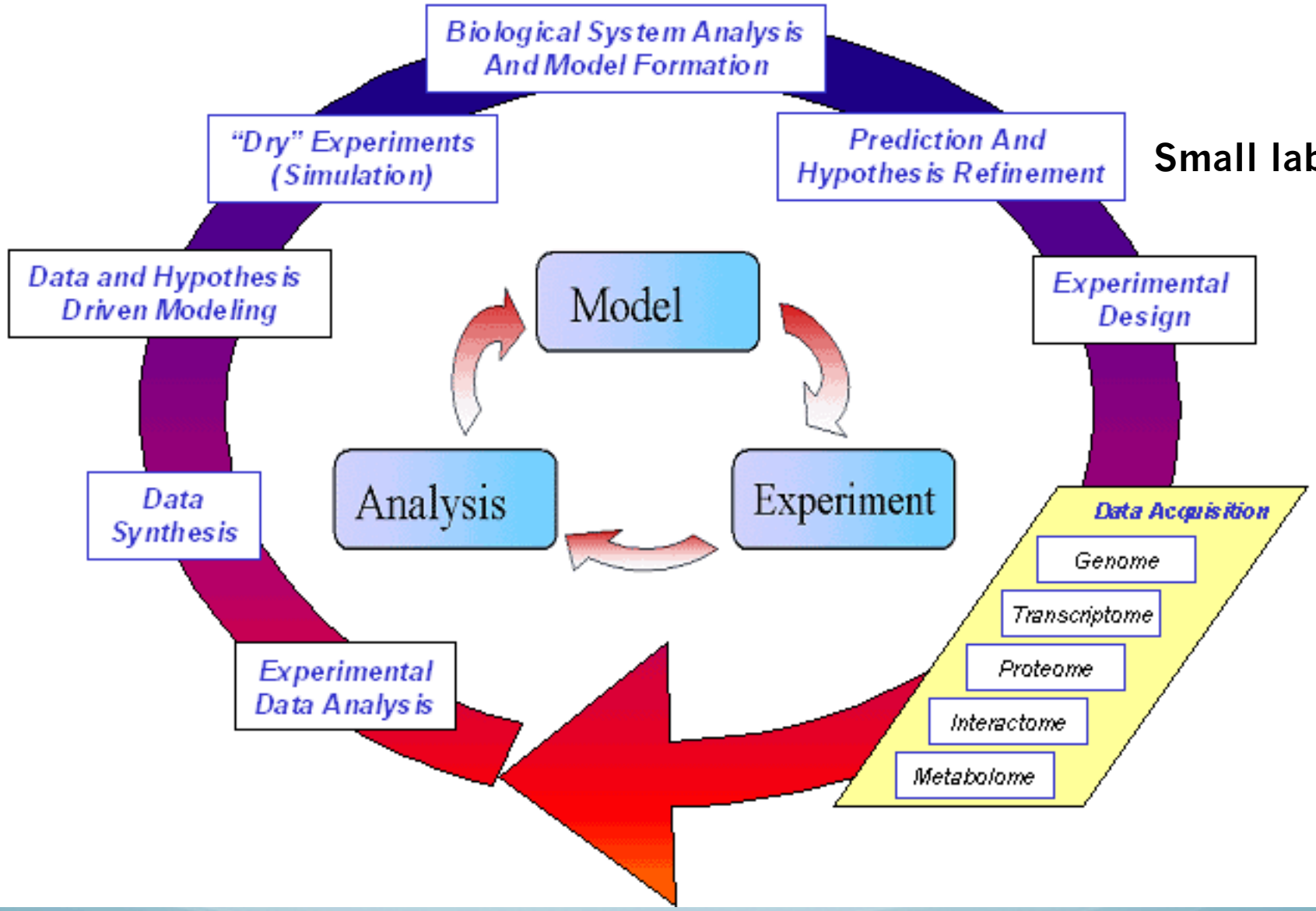
Modified histones: histone proteins, which package DNA into chromosomes, modified by chemical marks.

Transcription factors: proteins that bind to DNA and regulate transcription.

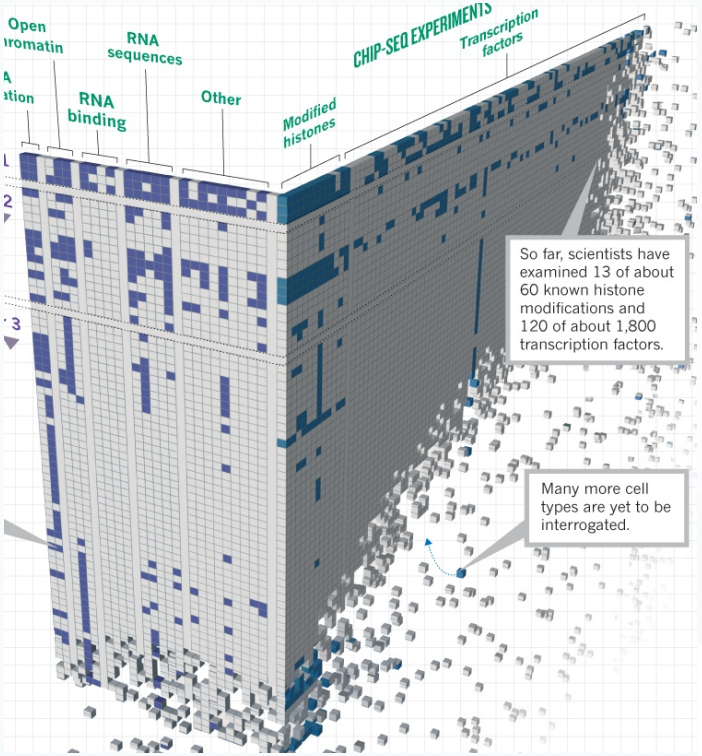
CELL LINES

Tiers 1 and 2: widely used cell lines that were given priority.

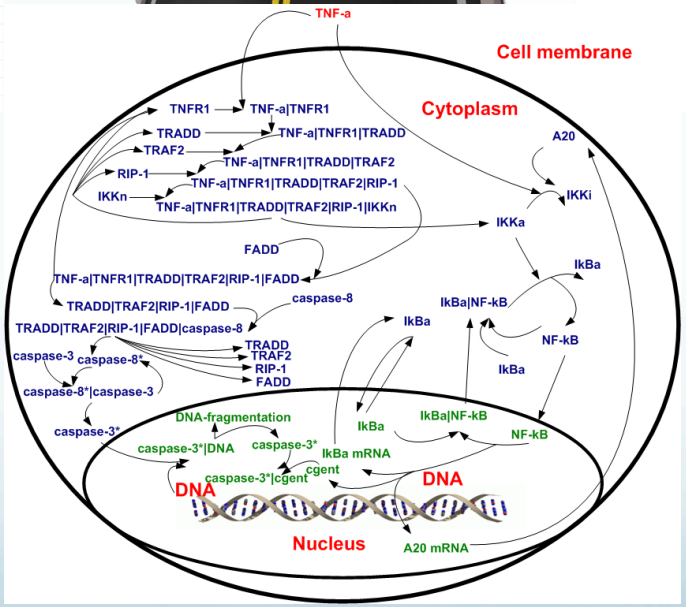
Tier 3: all other cell types.


Biological System Analysis And Model Formation
Prediction And Hypothesis Refinement
Experimental Design
Data Acquisition
Genome
Transcriptome
Proteome
Interactome
Metabolome
Experimental Data Analysis
Data Synthesis
Data and Hypothesis Driven Modeling
"Dry" Experiments (Simulation)
Model
Experiment
Analysis

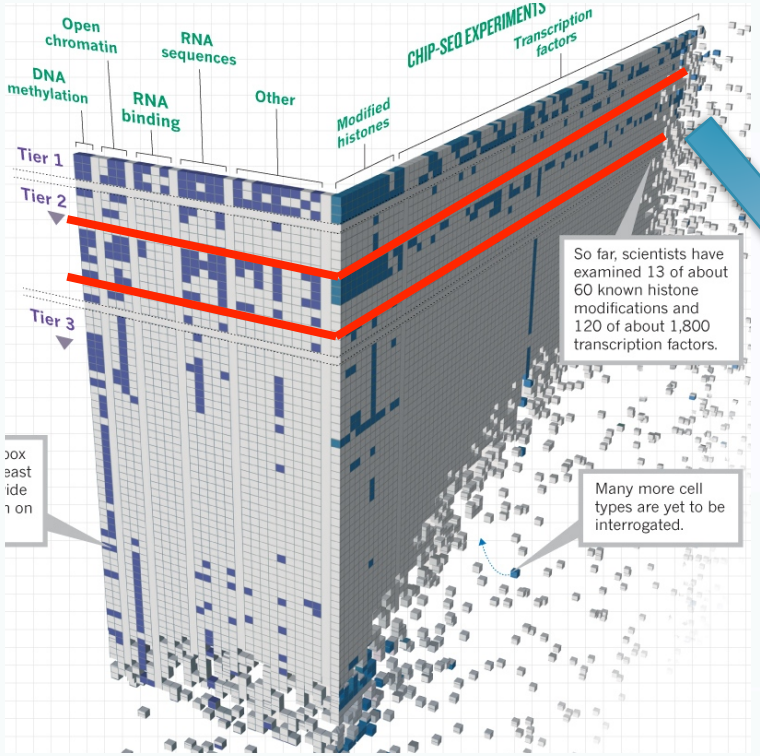
Small labs



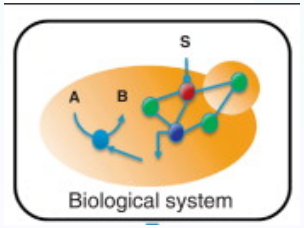
High-throughput data

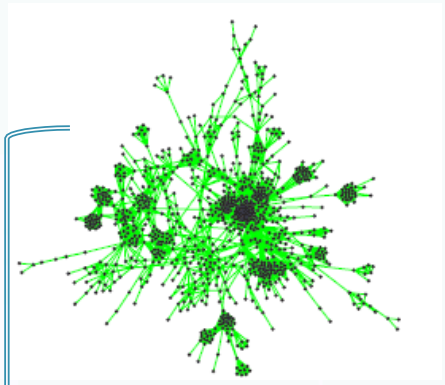


model

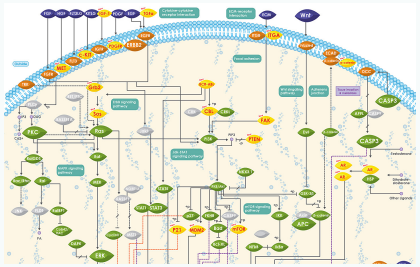


Genomic/transcriptomic data

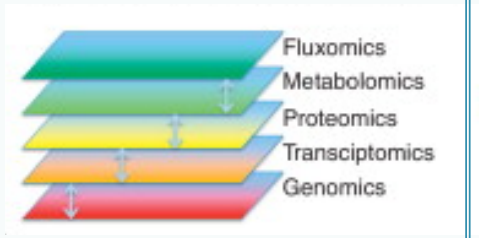




interactome



pathways



parameters

model

Model perturbation

New working hypothesis

Conclusions

- Mathematical models offer new perspective to biologists.
- With the present amount of data models can give new incites in providing new working hypotheses.
- Strong interaction between biologists and mathematician is needed to
  - correctly define data for parameters definition
  - highlight methodological limits

Università di Torino

MBC

Molecular Biotechnology Center

INFORMATICA

11001001

Gianfranco Balbo
Susanna Donatelli
Francesca Cordero
Marco Beccuti

Thank you!

raffaele.calogero@unito.it

B&Gu

Matteo Carrara
Chiara Fornari

Immunology Lab

Federica Cavallo
Elena Quaglino
Irene Merighi
Maddalena Arigoni
Stefania Lanzardo
Laura Conti

EPIGEN
Progetto Bandiera Epigenomica

Ministero Istruzione Università Ricerca

Bioinformatics wp

NEXT GENERATION SEQUENCING
for Targeted Personalized Therapy of Leukemia

FP7 grant 306242

2nd Annual
Next Generation Sequencing
Data Congress
June 2014, London UK

A Pathway-Based Approach using Latent Variable Structural Equation Modeling: An Application to 1000 Genomes Exon Sequencing Data

by

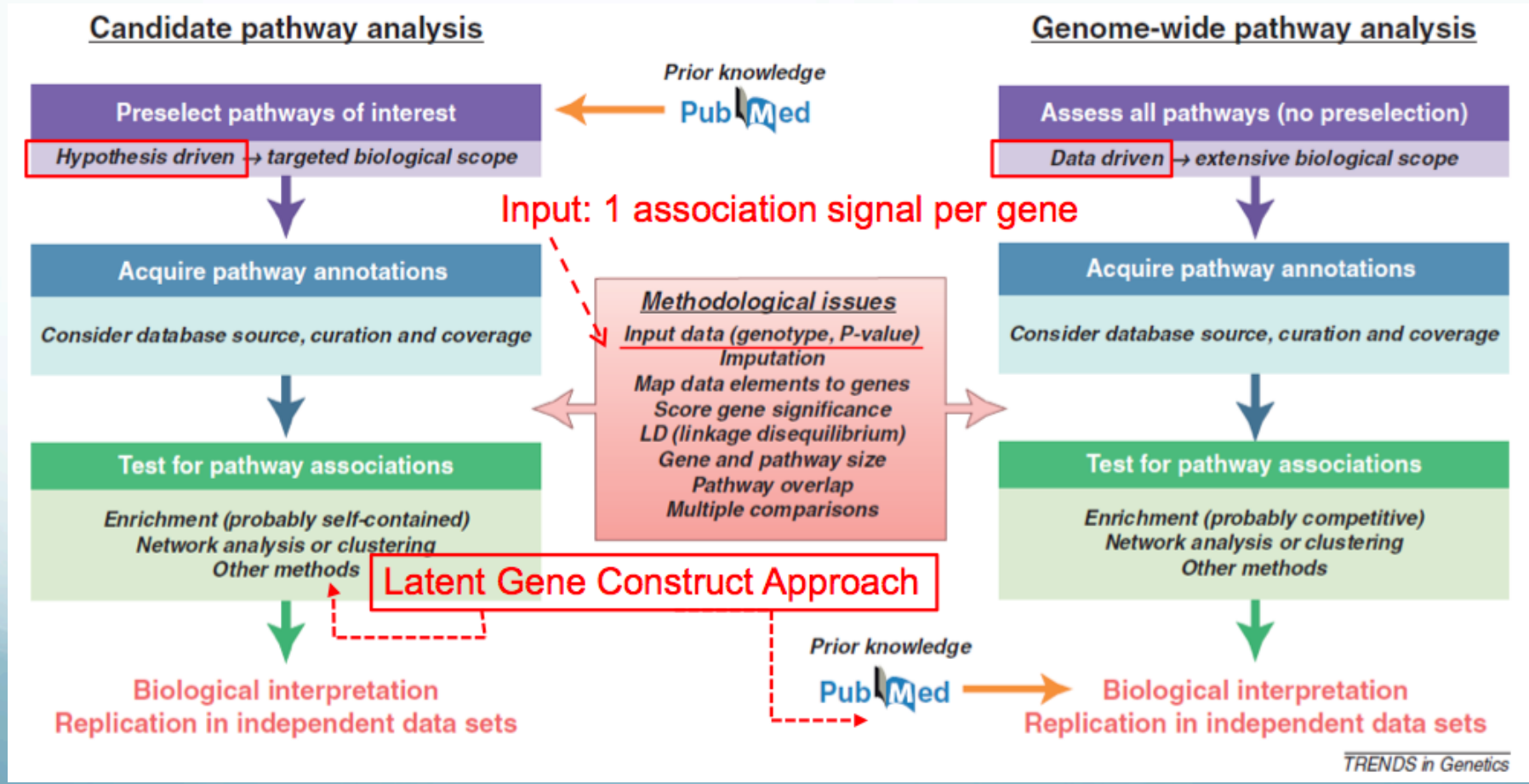
Nora L. Nock, Ph.D.

OXFORD GLOBAL

2nd Annual

Next Generation Sequencing Data Congress

June 2014, London UK



OXFORD GLOBAL

2nd Annual

Next Generation Sequencing Data Congress

June 2014, London UK