Computational Biomarker Discovery in the Big Data Era:
from Translational Biomedical Informatics to Systems Medicine

Bairong Shen
Center for Systems Biology
Soochow University
Bairong.Shen@suda.edu.cn

Suzhou & Soochow University

- Near Shanghai, 2,500 years of history, the Venice of the East, the canals that pass through the city.

- Cold Spring Harbor Conference-Asia

- One of the 10 most beautiful campuses,one of the top 5% universities in China

Structure of the presentation

1. Why we need biomarkers?

2. What are biomarkers?

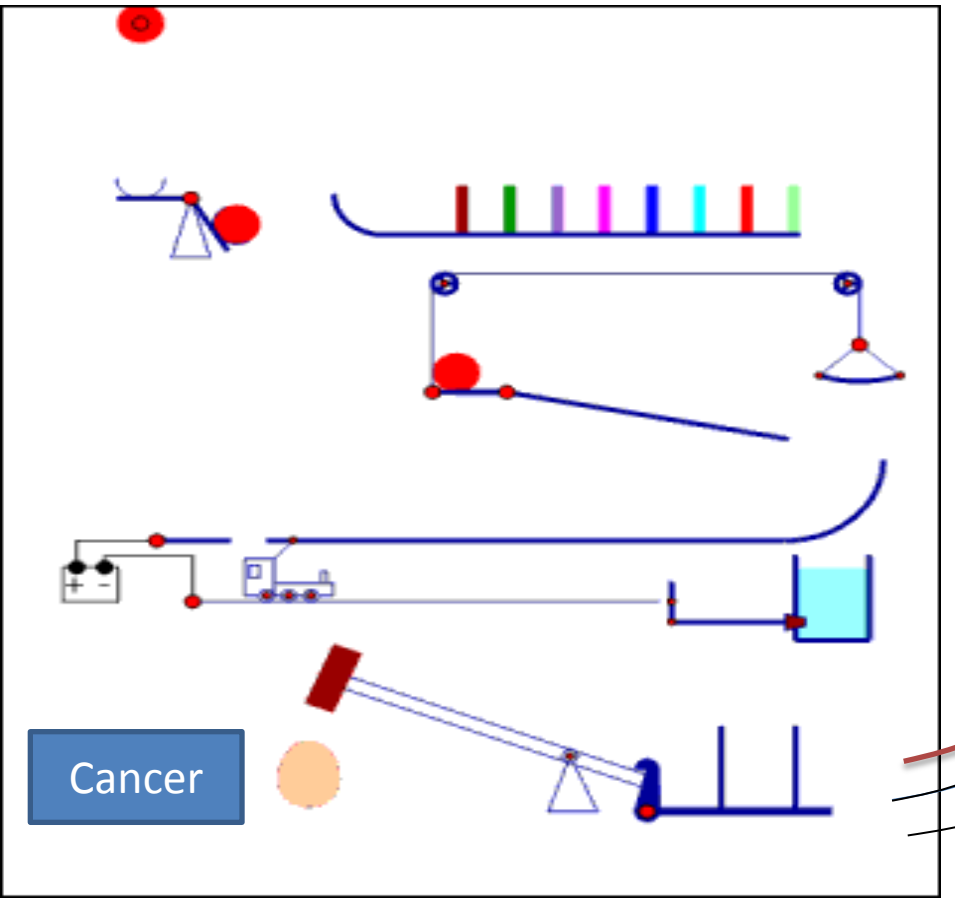
3. How will informatics be helpful?

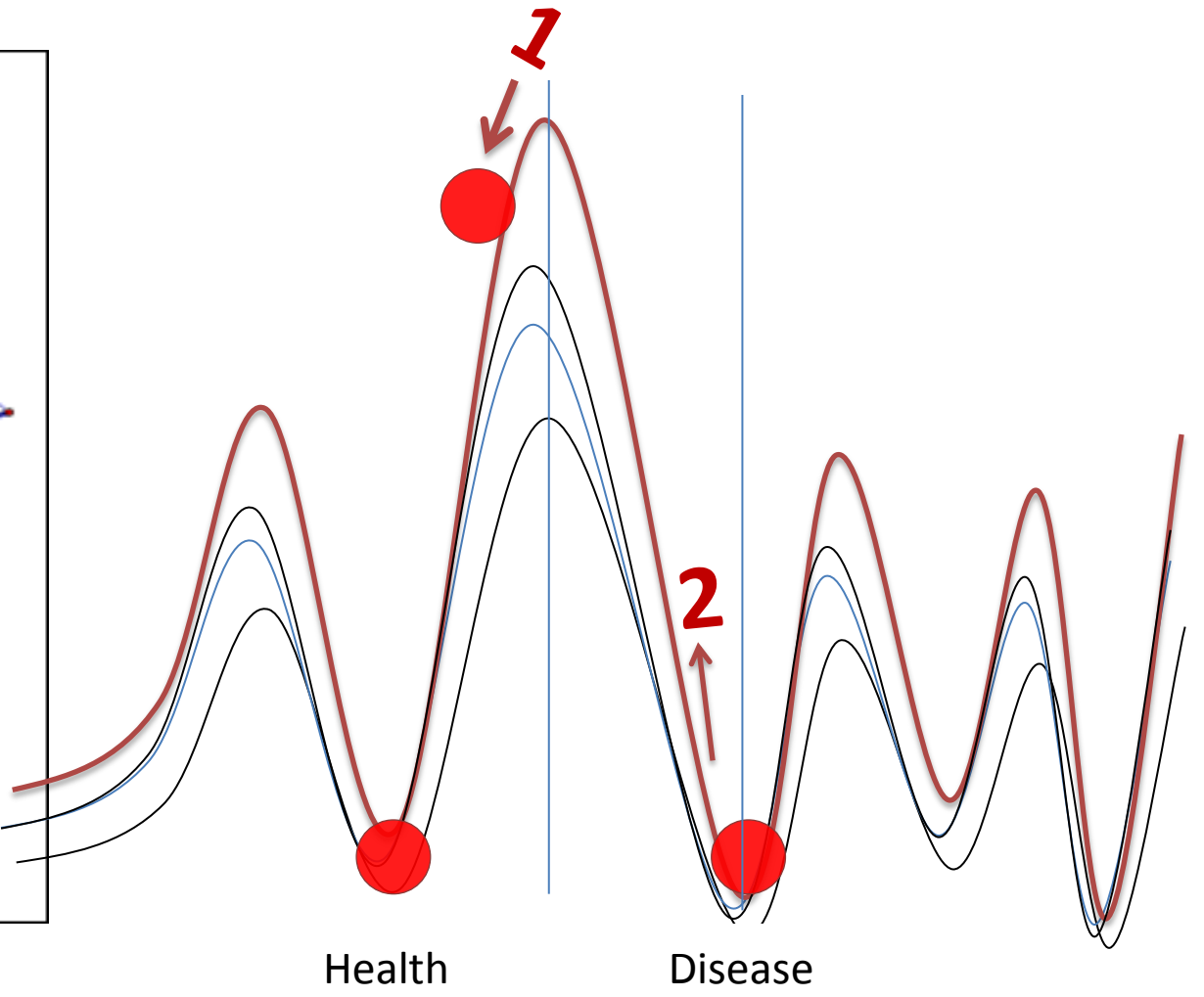
1.Why we need biomarkers?

- Over (mis)-diagnosis and over-treatment
  - Most widely understood in prostate cancer.
  - **Overdiagnosis in mammographic screening for breast cancer.**
  - Overdiagnosis in chest x-ray screening for lung cancer.
    - **20-40% of lung cancers detected by conventional x-ray screening represent overdianosis.**

安吉丽娜·朱莉(Angelina Jolie):发现自己带有乳腺癌易感基因好莱坞影星安吉丽娜·朱莉勇切双乳

A disease – health transfer model

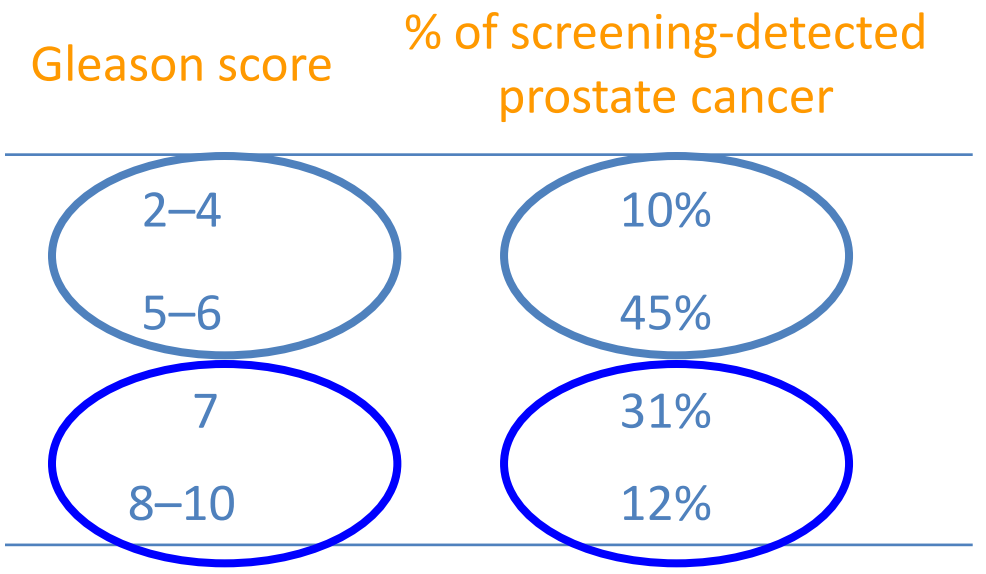




沈百荣，系统生物学研究中心，苏州大学，
Bairong.Shen@suda.edu.cn

Early Detection/Screening
Urgent Needs for Prostate Cancer

- **Most screening-detected prostate cancers are less aggressive**

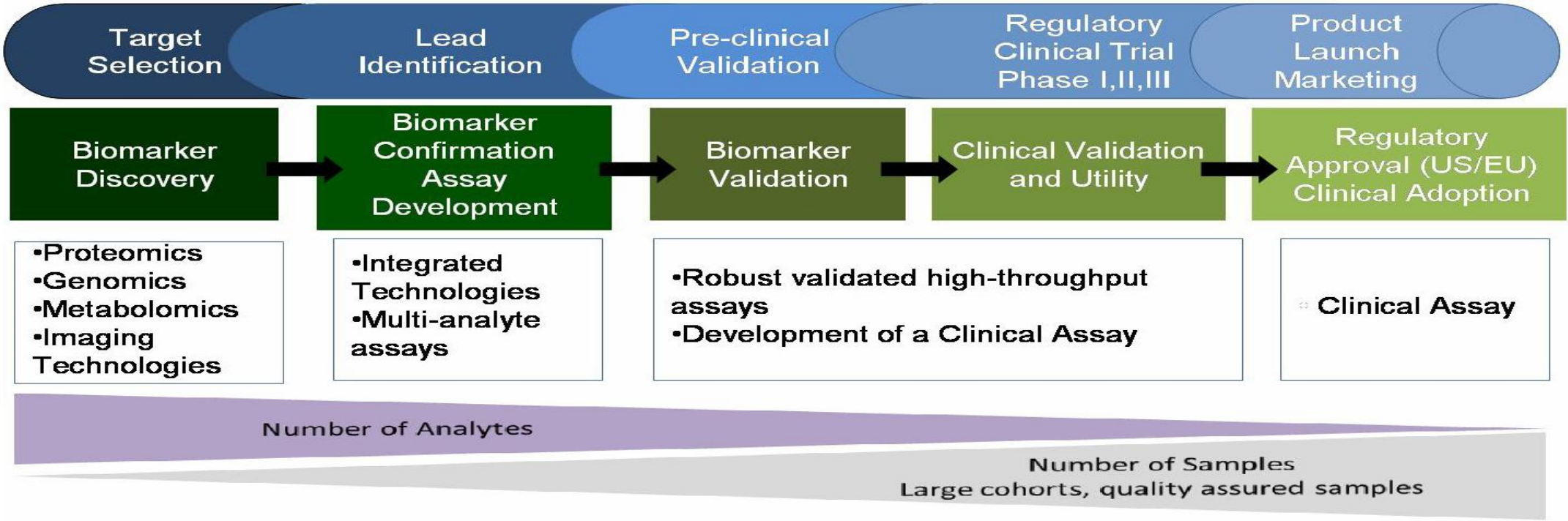


- Biomarkers to help distinguish aggressive from indolent disease
- Improve over current clinical and pathology data

- *Reproducible from laboratory to laboratory (Validation)*
- *Impact clinical care*
  - *Diagnosis*
  - *Prognosis*
  - *Treatment*

Andriole GL. J Natl Cancer Inst. 2005;97:433-8.

From CADD to CABD
Computer-Aided Drug (Biomarker) Discovery



Improved Translation of Biomarkers to the Clinic

http://www.molecularmedicineireland.ie/page/g/t/44

Towards to predictive medicine : the new paradigm

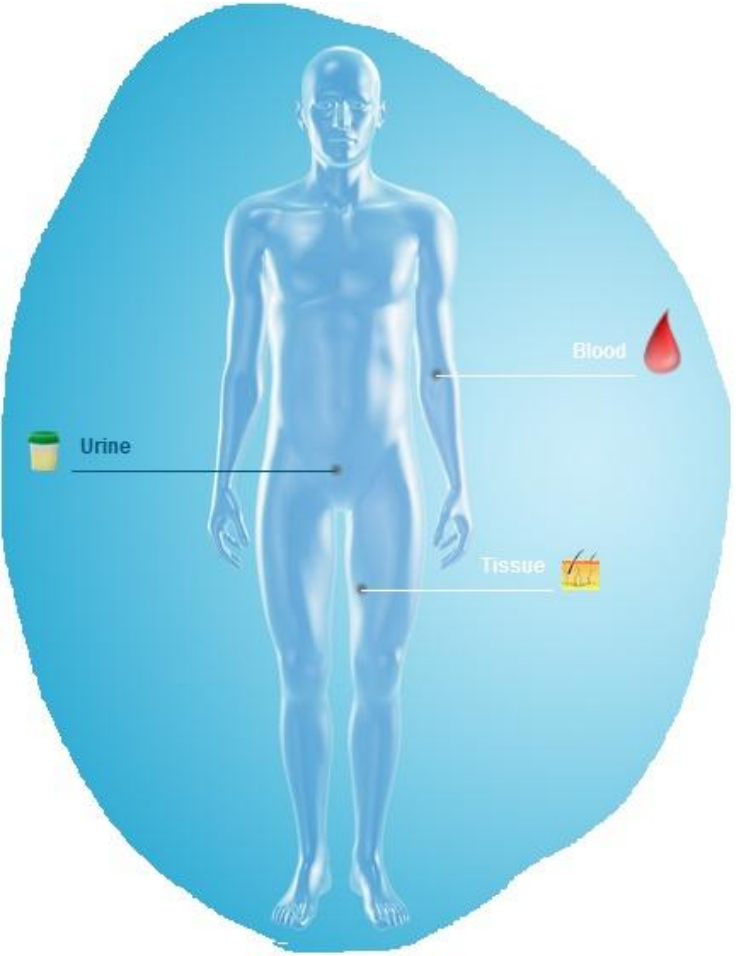
Comparing to CADD,
we have no general theories
for computer-aided biomaker discovery
at present!

DEGs are only potential biomarkers.

沈百荣，系统生物学研究中心，苏州大学，
Bairong.Shen@suda.edu.cn

2. What are biomarkers?

Biomarker Discovery： Types



- Prognostic:
  - Is it likely to develop this cancer ?
- Diagnostic:
  - What types/Subtypes of cancer is it ?
- Predictive:
  - Is this the optimal drug for the cancer ?
- Pharmaco-dynamics:
  - What's the optional dose for the patient ?
- Recurrence:
  - Will the cancer return ?

Biomarker Discovery： A cross-disciplinary challenge

- Static/ dynamic
- Molecule, imaging, clinical,
- Single or complex or multiple-scales

  - Medicine
  - Biology
  - Statistics: Personalized medicine: Subgroups
  - System Control Engineering: Controllability of complex networks
    - identifying the set of driver nodes with time-dependent control that can guide the system's entire dynamics
    - the driver nodes tend to avoid the high-degree nodes.

3. How will informatics be helpful?

Case Studies: Biomarker identificatin based on microRNA-mRNA network

BIOINFORMATICS

Vol. 21 Suppl. 2 2005, pages ii93–ii100
doi:10.1093/bioinformatics/bti1116

Genes and Genomes

Prediction of regulatory modules comprising microRNAs and target genes

Sungroh Yoon[1,*] and Giovanni De Micheli[2]

[1]Computer Systems Laboratory, Stanford University, Stanford, CA 94305, USA and [2]Integrated System Center, EPF Lausanne, Switzerland

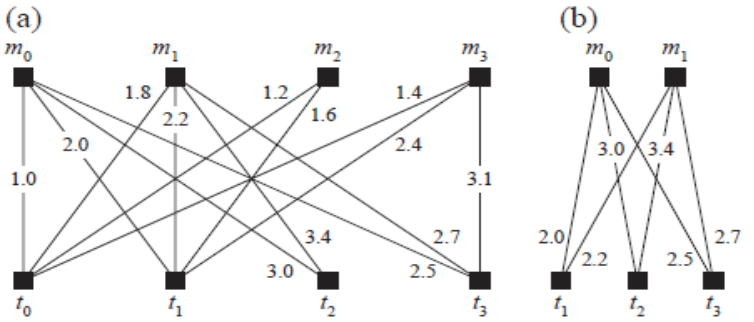
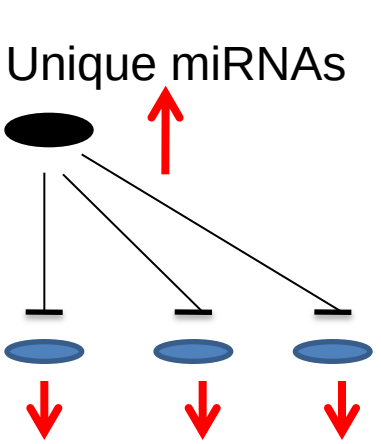


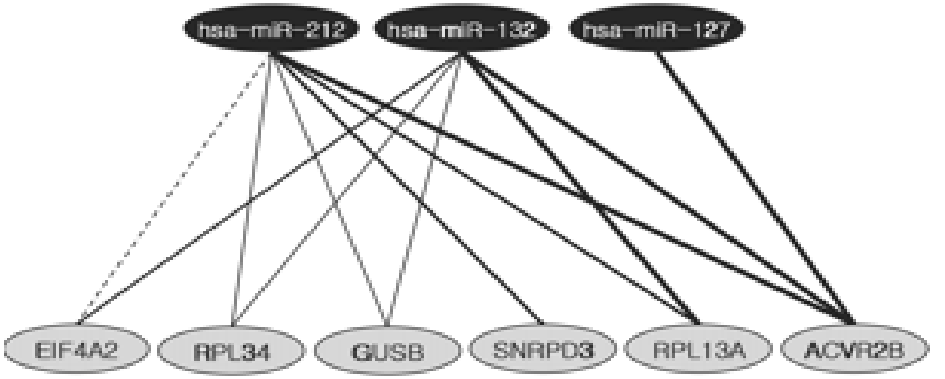
Fig. 2. (a) Example relation graph $G = (M \cup T, E, w)$, where $M = \{m_0, m_1, m_2, m_3\}$, and $T = \{t_0, t_1, t_2, t_3\}$ with some hypothetical weights. (b) An MRM found in G with the parameter $\delta = 0.5$.

microRNA regulatory modules (MRMs):

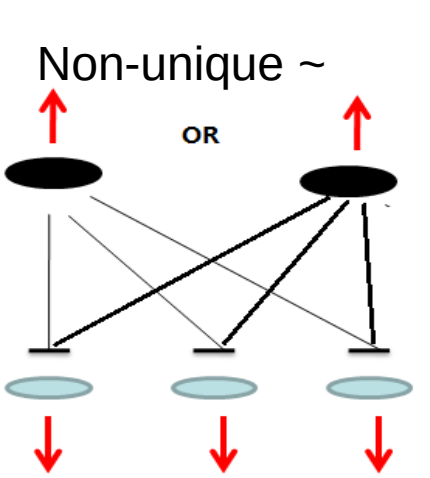
defined as groups of miRNAs and their target genes that are believed to have similar functions or to be involved in similar biological processes.

The microRNA-mRNA network



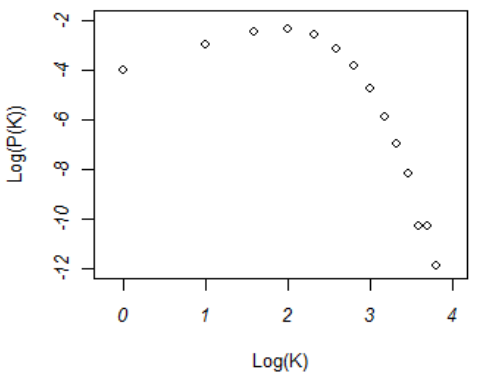


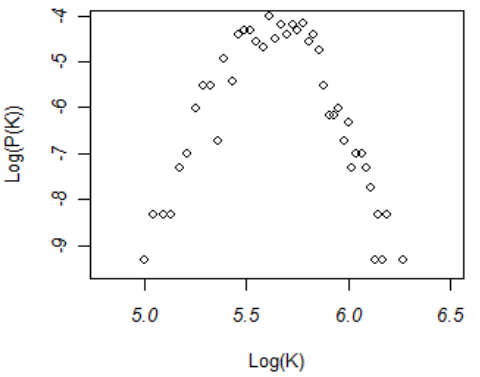
miRNA–mRNA modules



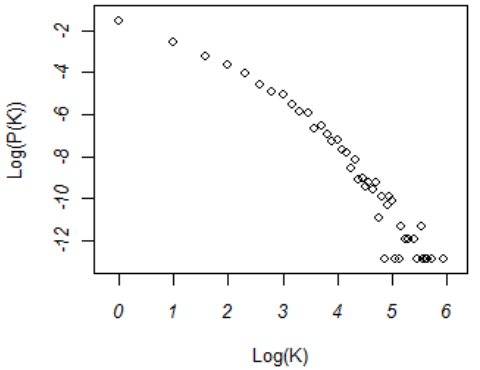
As it was observed in our study that most genes are regulated by unique miRNAs, miRNAs activity can be reflected by the proportion of their unique target genes from conditional miRNA-mRNA interaction network.

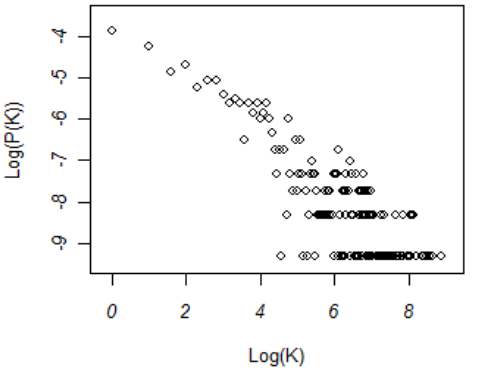
Degree Distribution of Random Simulation and Real miRNA-mRNA interaction network





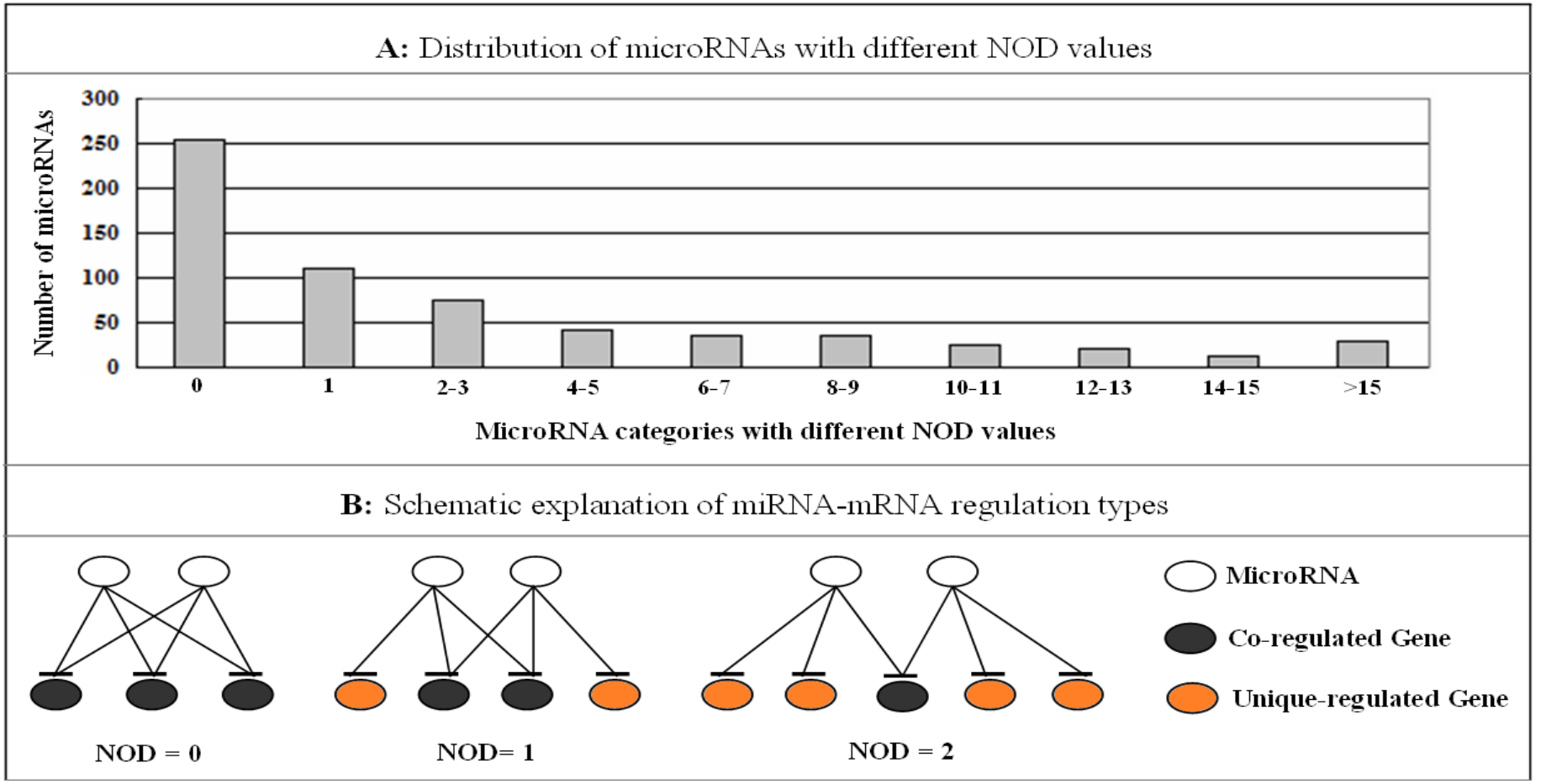
the approximate symmetrical degree distributions of random simulation network(A: in-degree; B: out-degree)





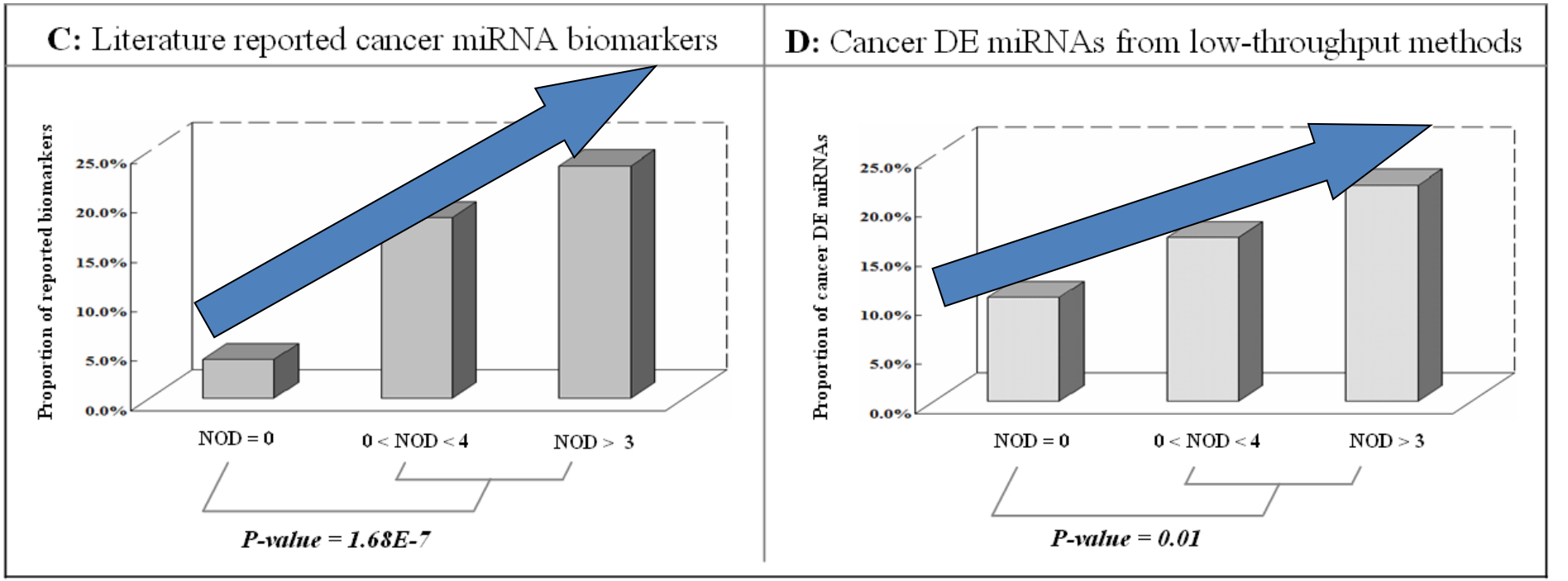
while the degree distributions for real miRNA-mRNA interaction network were proximate scale-free (C: in-degree; D: out-degree)

Novel out degree (NOD): measure the independent regulation power of an individual miRNA

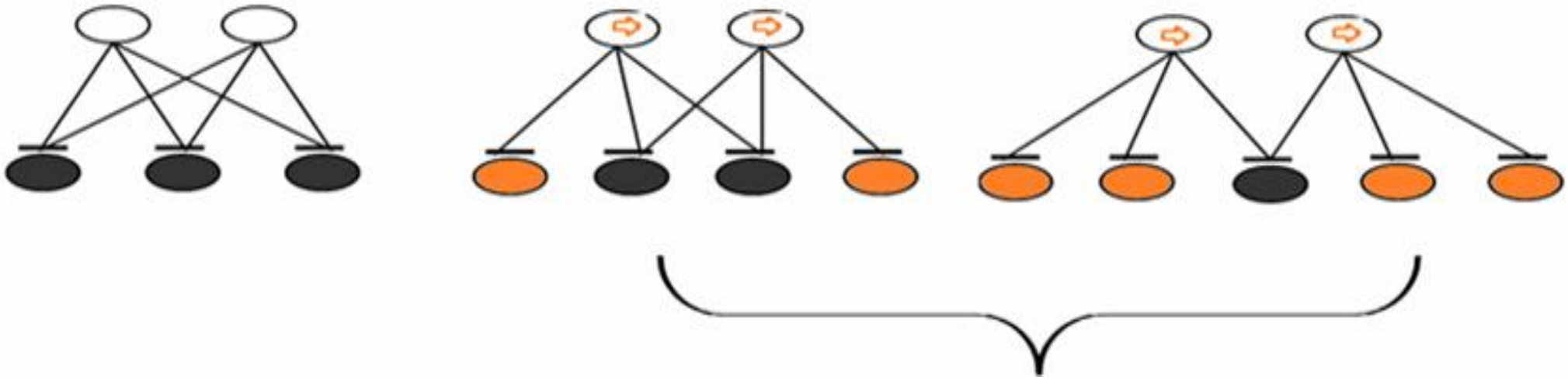


NOD: the uniquely regulated genes for one specific miRNA

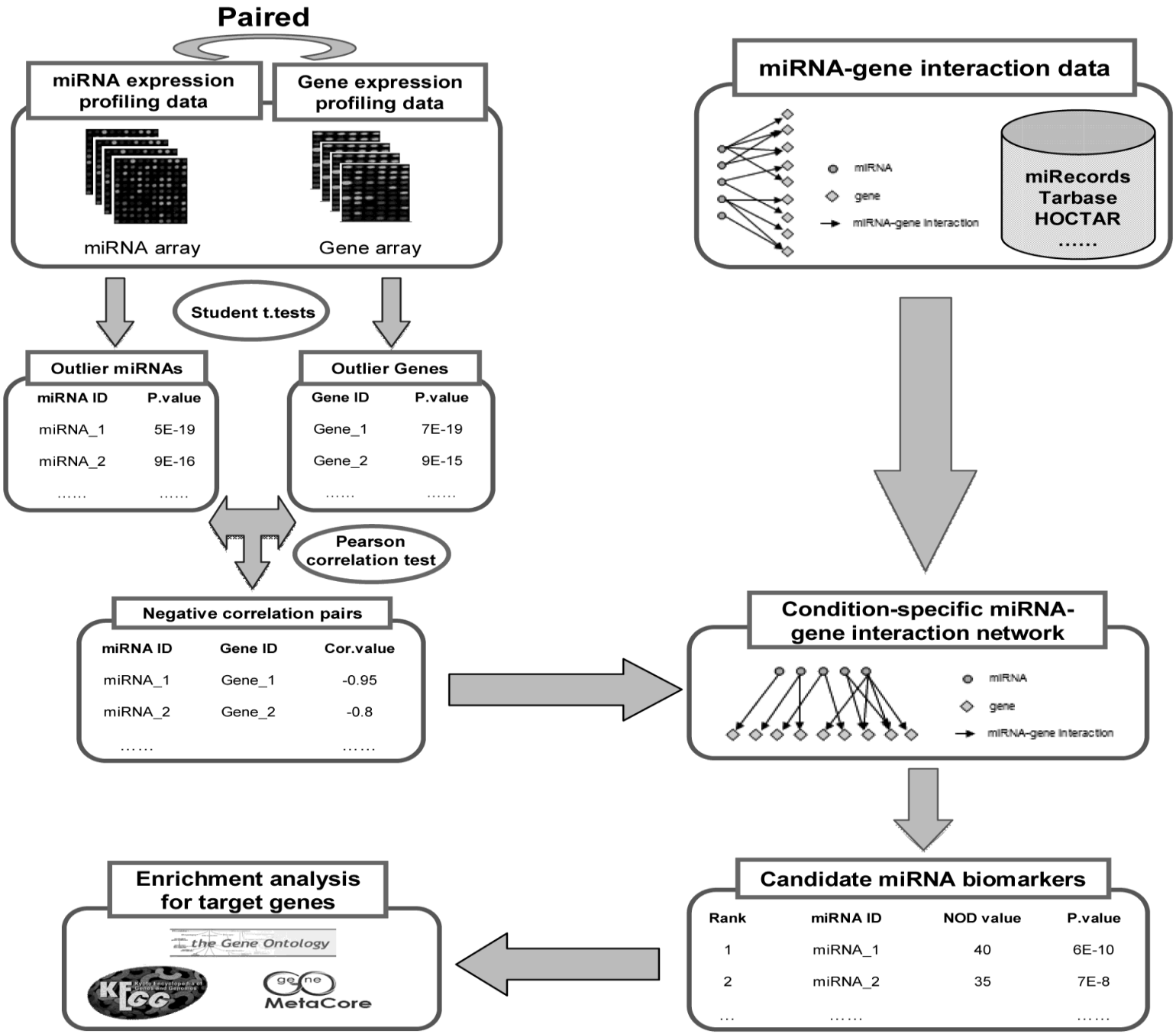
A statistically significant difference for their biomarker potential could be observed between miRNAs without independent regulation power and those with independent regulation power.



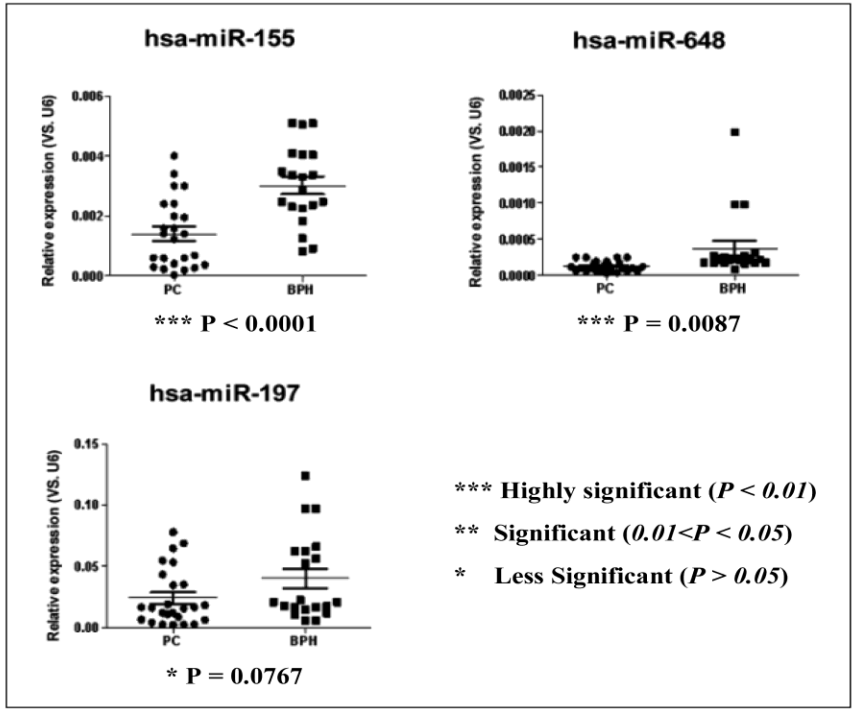
Spies or underground workers are always **single line contacted** by important persons



The bioinformatics model



In vitro validation of candidate prostate cancer miRNA biomarkers

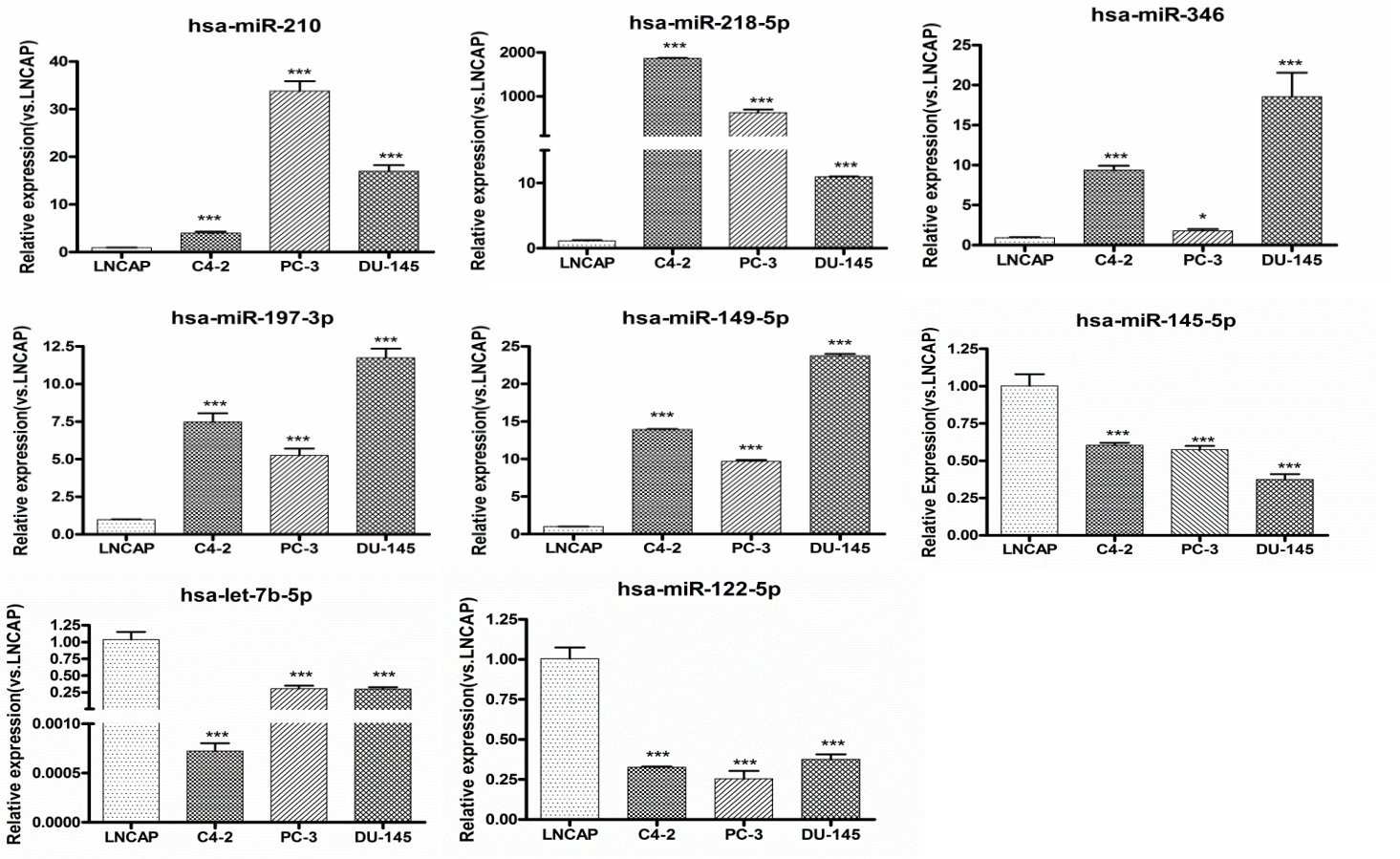


Among these two miRNAs, the down-regulating outlier pattern of *miR-155* in prostate cancer was in coincidence with previous report

while *miR-648* is the novel prostate cancer miRNA biomarker by our study.

Even though *miR-197* did not show significant outlier activity in prostate cancer, it was previously declared to be a potential miRNA biomarker for lung cancer in another study

Application: Identification of androgen dependent & independent specific miRNA biomarkers



- Mir-210
- miR-197-3p
- miR-149-5p
- miR-346
- miR-218-5p
- let-7b-5p
- miR-145-5p
- miR-122-5p

AD hPCa cell line: LNCAP; AI hPCa cell lines: C4-2, PC-3, DU-145

JOURNAL OF TRANSLATIONAL MEDICINE
IMPACT FACTOR 3.99

Medical bioinformatics Section edited by Samir K. Brahmachari

The explosion of genome sequencing data along with genotype to phenotype correlation studies has created data deluge in the area of biomedical sciences. The aim of the Medical bioinformatics section is to aid the development and maturation of the field by providing a platform for the translation of these datasets into useful clinical applications. The increase in computing capabilities and availability of different data from advanced technologies will allow researchers to build System Biology models of various diseases in order to efficiently develop new therapeutic interventions and reduce the current prohibitively large costs of drug discovery.

1. **Research** Open Access
   Drug-repurposing identified the combination of Trolox C and Cytisine for the treatment of type 2 diabetes
   Ling Jin, Jian Tu, Jianwei Jia, Wenbin An, Huanran Tan, Qinghua Cui, Zhixin Li
   Journal of Translational Medicine 2014, **12**:153 (31 May 2014)
   Abstract | Full text | PDF | PubMed

2. **Research** Open Access Highly accessed
   Identification of candidate miRNA biomarkers from miRNA regulatory network with application to prostate cancer
   Wenyu Zhang, Jin Zang, Xinhua Jing, Zhandong Sun, Wenying Yan, Dongrong Yang, Feng Guo, Bairong Shen
   Journal of Translational Medicine 2014, **12**:66 (11 March 2014)
   Abstract | Full text | PDF | ePUB | PubMed

3. **Meeting report** Open Access
   Dry computational approaches for wet medical problems
   Frank Emmert-Streib, Shu-Dong Zhang, Peter Hamilton
   Journal of Translational Medicine 2014, **12**:26 (25 January 2014)
   Abstract | Full text | PDF | ePUB | PubMed

4. **Research** Open Access
   VIRsiRNApred: a web server for predicting inhibition efficacy of siRNAs targeting human viruses
   Abid Qureshi, Nishant Thakur, Manoj Kumar
   Journal of Translational Medicine 2013, **11**:305 (11 December 2013)
   Abstract | Full text | PDF | ePUB | PubMed

5. **Research** Open Access Highly accessed
   A network model of genomic hormone interactions underlying dementia and its translational validation through serendipitous off-target effect
   Erfan Younesi, Martin Hofmann-Apitius
   Journal of Translational Medicine 2013, **11**:177 (26 July 2013)
   Abstract | Full text | PDF | ePUB | PubMed | Cited on BioMed Central

6. **Research** Open Access Highly accessed
   Clear cell renal cell carcinoma associated microRNA expression signatures identified by an integrated bioinformatics analysis
   Jiajia Chen, Daqing Zhang, Wenyu Zhang, Yifei Tang, Wenying Yan, Lingchuan Guo, Bairong Shen
   Journal of Translational Medicine 2013, **11**:169 (10 July 2013)

The future of CABD

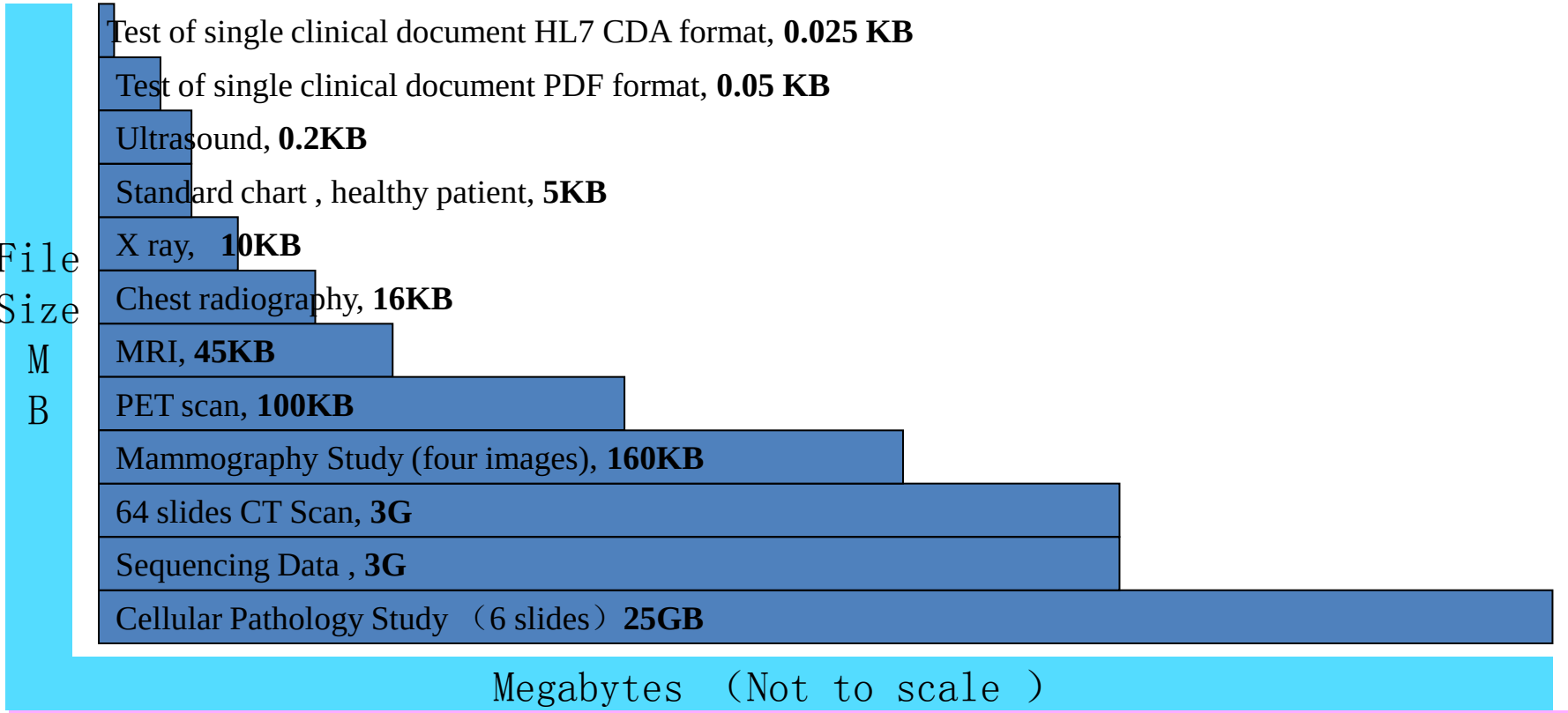
- alphanumeric strings,text,
- characters,Strings，
- booleans,floating-point numbers,
- imaging,voice, video

- Bioinformatics
- Imaging informatics
- Clinical informatics
- Public health informatics

- Storing
- Retrieving
- organizing
- Analyzing
  - Statistics
  - Mathematics: ODE/PDE
  - Physics and Chemical modeling: QSAR, classification
  - Biological: GO, pathway enrichment，evolutionary
  - Medical: personalized information
  - Systems level： network analysis

沈百荣，系统生物学研究中心，苏州大学，
Bairong.Shen@suda.edu.cn

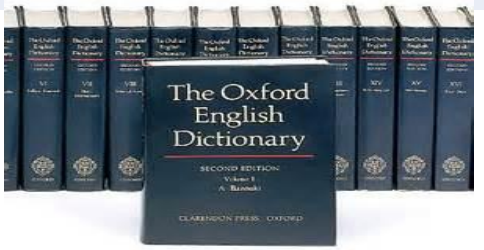
EHR: The size of various files and pieces of data that would be a part of any electronic and personal health record of the future.

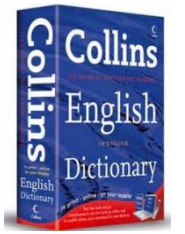


Modified from :
The Creative Destruction of Medicine: How the Digital Revolution Will Create Better Health Care

Big data and small data

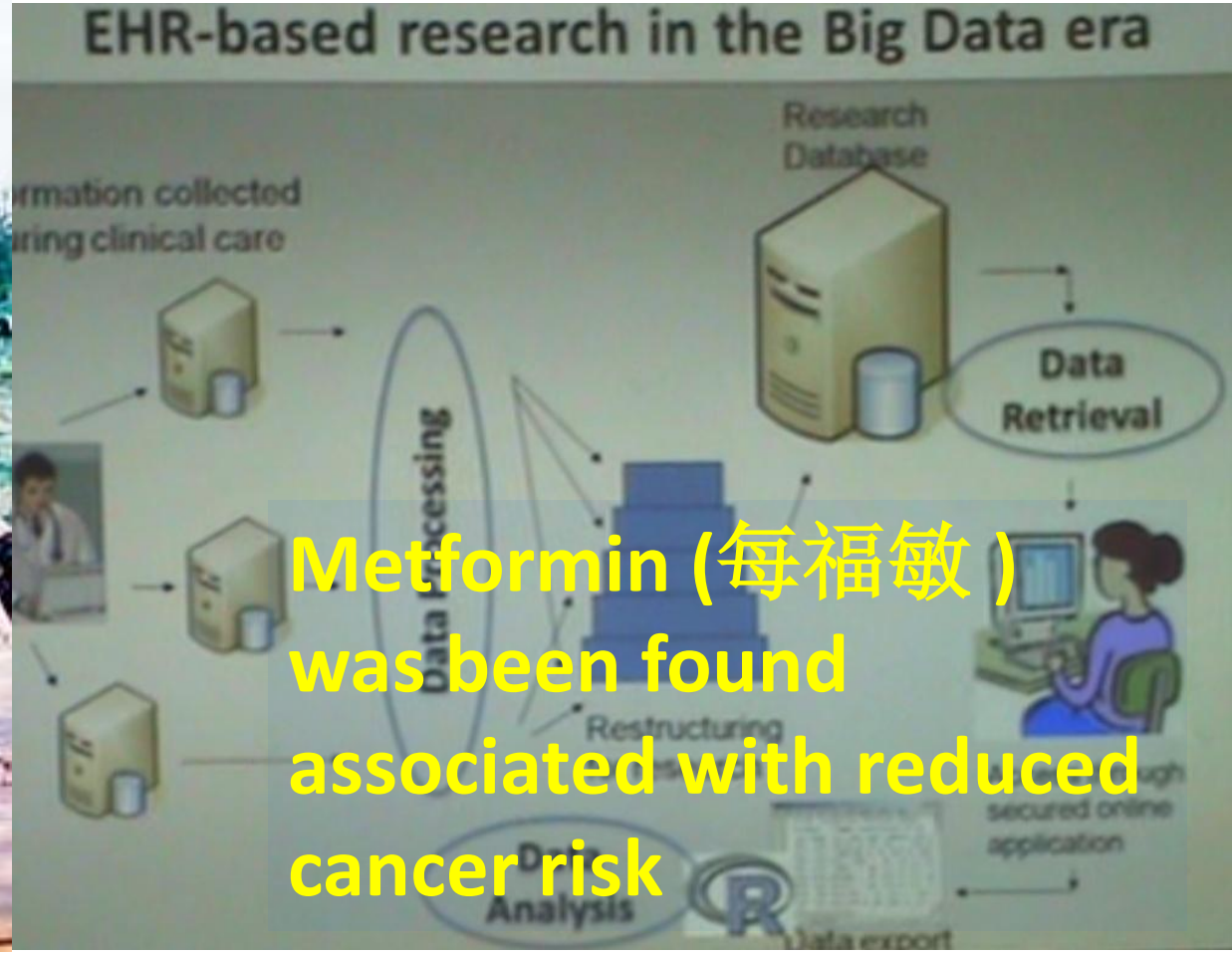
	Big Data	Small Data
Data Condition	Always unstructured, not ready for analysis, many relational database tables that need merged	Ready for analysis, flat file, no need for merging tables.
Location	Cloud, Offshore, SQL Server, etc.	Database, local PC
Data Size	Over 50K Variables, over 50K individuals, random samples, unstructured	File that is in a spreadsheet, that can be viewed on a few sheets of paper
Data Purpose	No intended purpose	Intended purpose for Data Collection



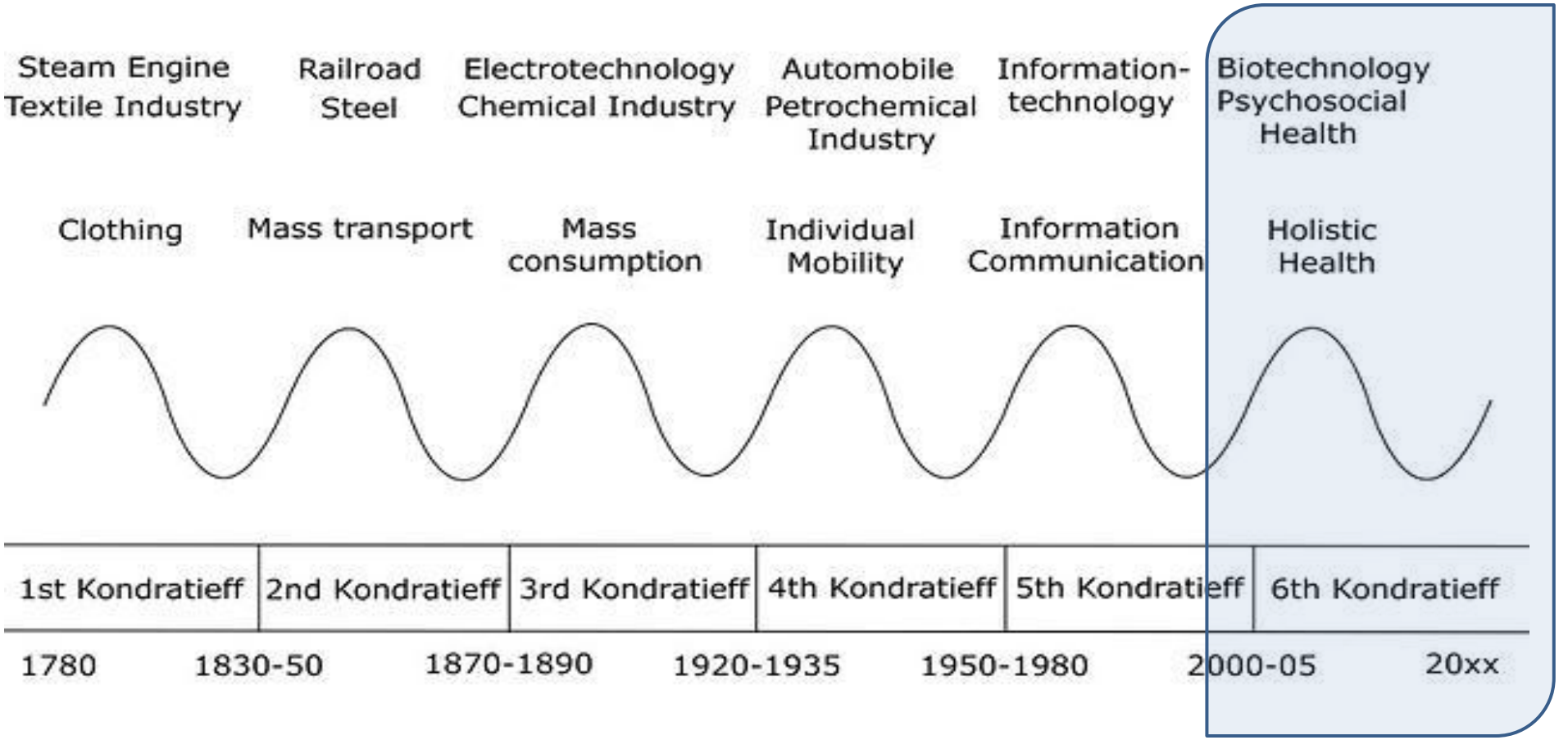


沈百荣，系统生物学研究中心，苏州大学，
Bairong.Shen@suda.edu.cn

Directly extraction of knowledge from EHRs

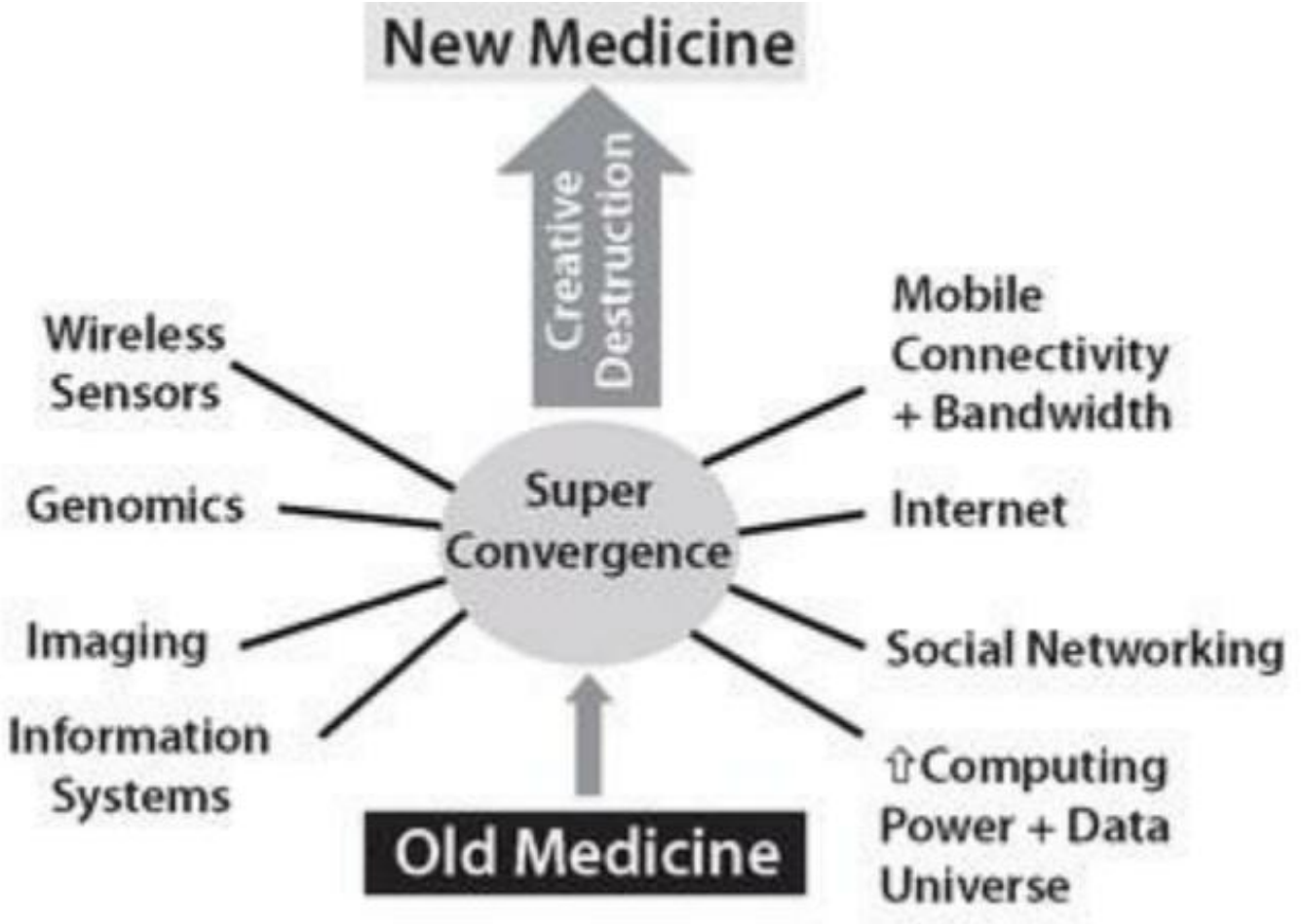


The sixth Kondratieff – long waves of prosperity



http://static.wixstatic.com/media/8fa589_d2b69a0ca7a546bf865c53b623381bd7.jpg_srz_638_319_75_22_0.50_1.20_0.00_jpg_srz

P4 (Personalized, Predictive, Preventive and Participatory) medicine and Biomarkers



沈百荣，系统生物学研究中心，苏州大学，
Bairong.Shen@suda.edu.cn

What do we need next steps ?

- All the identified biomarkers need to be validated for the future personalized diagnosis
  - Both Omics and EHR are required
    - PCA specific EHR
  - Both static and dynamic simulation are helpful
    - ODE based simulation

沈百荣，系统生物学研究中心，苏州大学，
Bairong.Shen@suda.edu.cn

With Big data available……

- Can we identify biomarkers from the big data directly:
  - For lifestyle, environmental and genetic factors.

NO !

  - Computer-aided deep mining is needed.

Summary of the presentation

1. Why we need biomarkers?
   - Over(mis)-diagnosis / and over-treatment

2. What are biomarkers?
   - Characterization of biomarkers

3. How will informatics be helpful?
   - Case Studies and the future of CABD

The superior doctor prevents illness: the new P4 medicine

Acknowledge: Our team in Suzhou