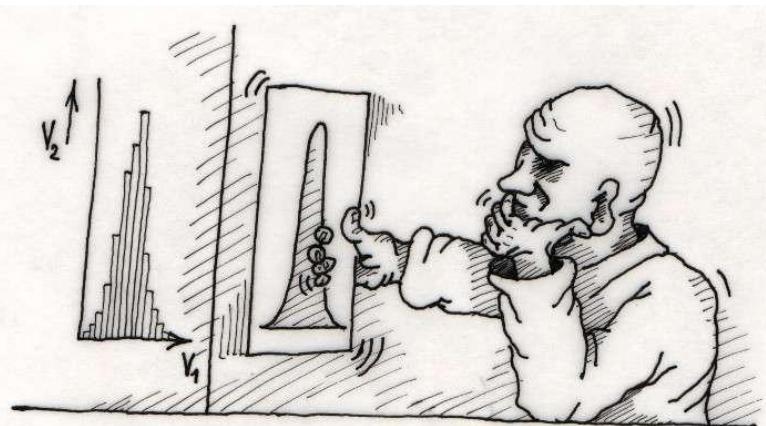


# 生物信息学： 组学时代的生物信息数据挖掘和理解

2020年秋



# 有关信息

- 授课教师: 宁康, 张礼斌, 陈鹏
  - Email: ningkang@hust.edu.cn
  - Office: 华中科技大学东十一楼504室
  - Phone: 87793041, 18627968927
- 课程网页
  - <http://www.microbioinformatics.org/Bioinformatics.html>
  - QQ群:



# 课程安排

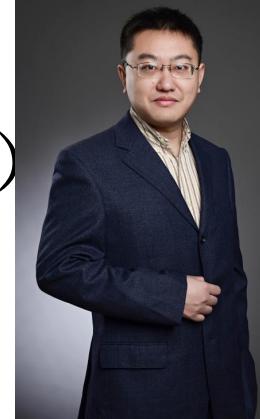
## (生物信息中的算法设计与概率统计模型)

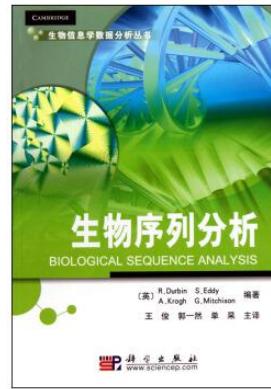
- 生物背景和课程简介
- 生物信息学和生物数据挖掘
  - 生物数据的格式及其意义
    - 序列数据
    - 树状数据
    - 网络数据
    - 表达数据等
  - 生物数据库及其用法
  - 生物信息基本算法
    - 双序列联配
    - 多序列联配
    - 基因组组装算法
    - 基因预测和功能注释
    - 系统发育树构建
    - 蛋白质结构预测
    - 生物调控网络解析
  - 组学数据分析方法
    - 基因组变异分析
    - 基因表达和比较分析
    - 非编码RNA分析
    - 蛋白组分析
    - 宏基因组分析
  - 系统生物学与交叉科学
- 面向生物大数据挖掘的深度学习

研究对象：  
生物序列，  
进化树，  
生物网络，  
基因表达

...

方法：  
生物计算与生物信息

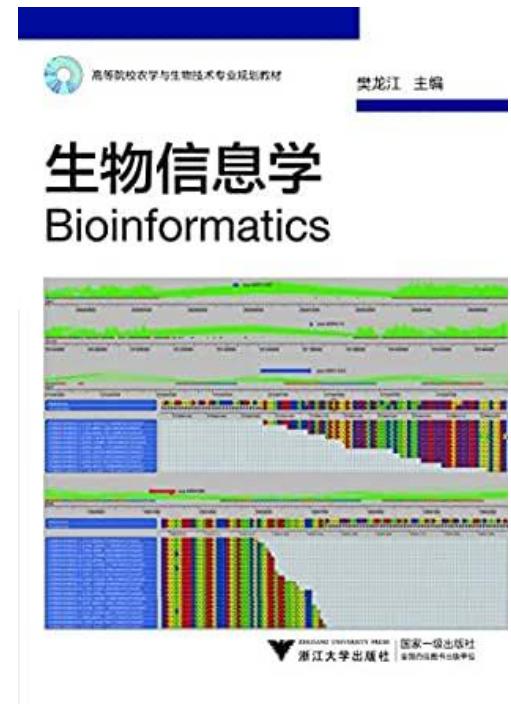
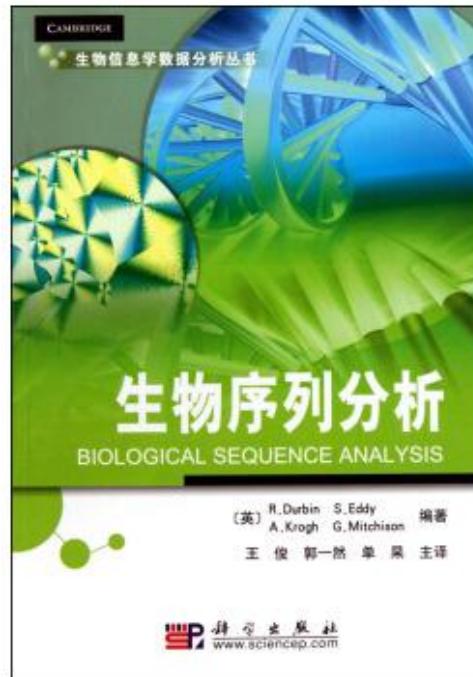
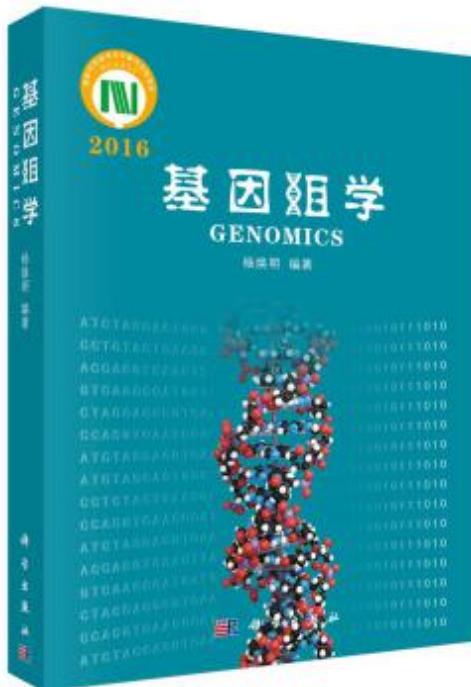




# 教材及参考书目

- **教学参考书:**
- 《生物序列分析》（第1版）.科学出版社. 2010年8月出版. R. Durbin等编著，王俊等主译.
- **课外文献阅读:**
- 《生物信息学》（第1版）.浙江大学出版社. 2017年3月出版. 樊龙江主编.
- 《基因组学》（第1版）.科学出版社. 2016年10月出版. 杨焕明主编.

# References



# Slides credits

- 生物信息学研究方法概述：北京大学生物信息中心
- 生物统计学：卜东波@中国科学院计算技术研究所，邓明华@北京大学
- 神经网络与深度学习：邱锡鹏@复旦大学
- Introduction to Computational Biology and Bioinformatics: Xiaole Shirley Liu Lab@Harvard University
- Combinatorial Methods in Computation Biology: Ken Sung Lab@NUS
- Deep Learning in the Life Sciences: MIT

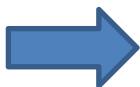
# 基因组学 (genomics)

- 基因组学主要是研究生物基因组和如何利用基因的一门综合学问,是涉及基因作图、测序和整个基因组功能分析的遗传学分支。

## 组学 (-Omics)

- 近二十多年以来, 基于高通量分析的系统生物学 (**system biology**) 研究飞速发展, 从最初的基因组学 (**genomics**) 已经发展到当前的多组学时代, 围绕核酸、蛋白、代谢物 (如糖、脂)、矿物元素 (离子)、表型等, 迄今已形成了基因组学、转录组学 (**transcriptomics**)、蛋白质组学 (**proteomics**)、代谢组学 (**metabonomics**)、离子组学 (**ionomics**)、生理组学以及表型组学, 它们已成为系统生物学研究的重要方向。

genome



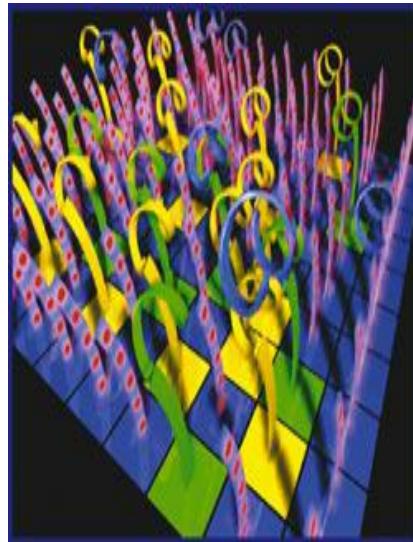
transcriptome



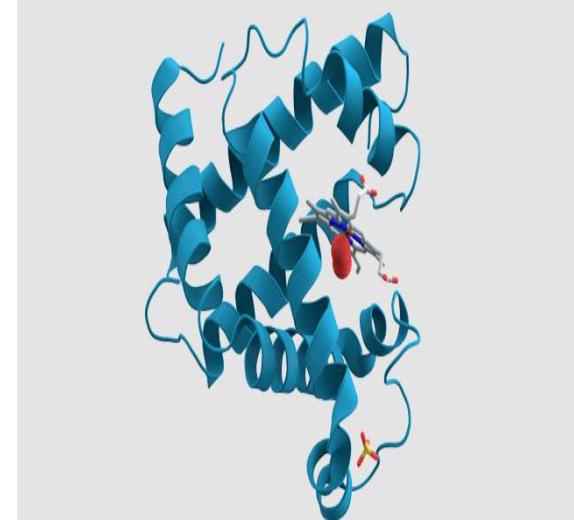
proteome



测序技术 (sequencing)

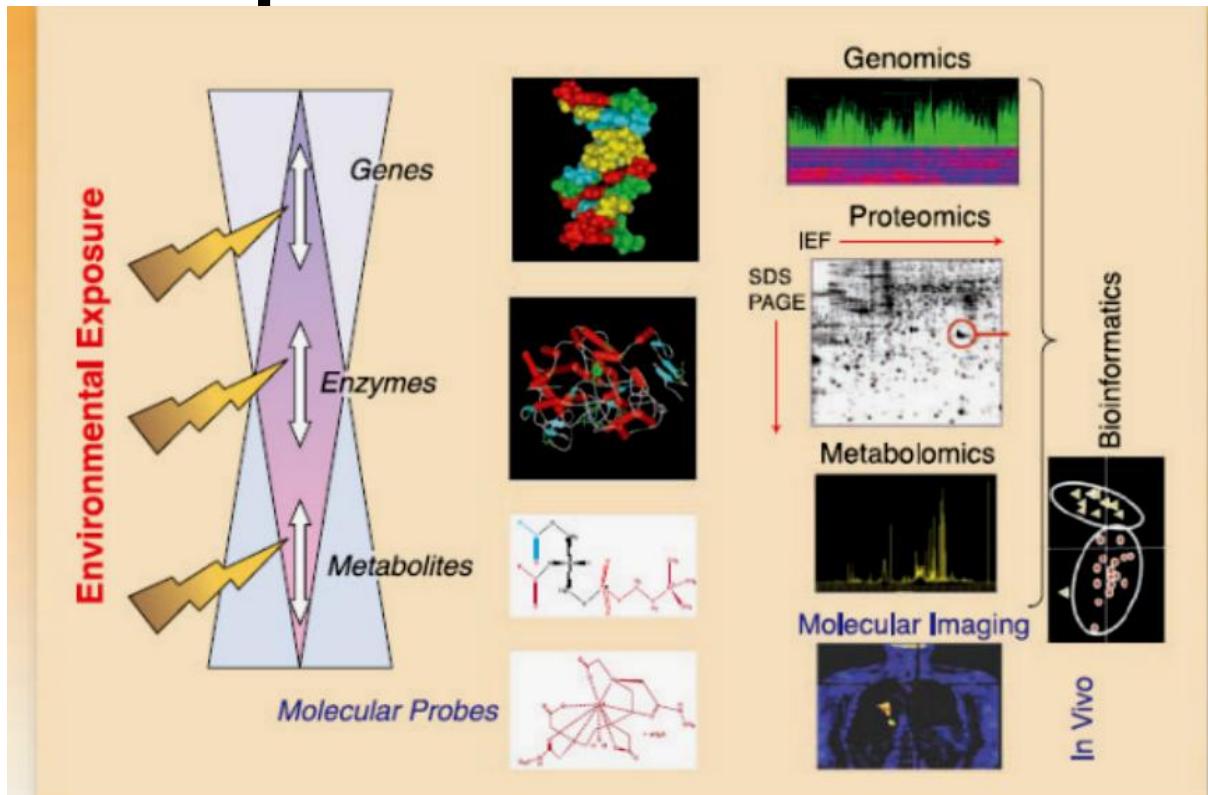


基因芯片 (DNA microarray)



二维凝胶电泳 (2-DE) 和质谱 (MS)

# The flow of the “omics” sciences: genomics, proteomics, and



# 转录组学（ transcriptomics ）

- 转录组学是对细胞（生物体）在某种条件下所有转录产物（即转录组， transcriptome ）进行系统研究，即在 RNA 水平研究基因表达的变化。
- 转录组学的研究需要大规模的基因表达分析技术，应用于转录组学的研究技术有：差异显示（ DD ）、扩增片段长度多态性 cDNA-AFLP ）、抑制消减杂交（ SSH ）、基因表达序列分析（ SAGE ）、基因芯片（ DNA microarray ）， RNA 测序（ RNA-seq ）等技术。
- 其中 RNA-seq 优点突出，能对转录本进行大规模测序及表达水平的精确定量分析

## 应用于转录组分析的三种技术的比较

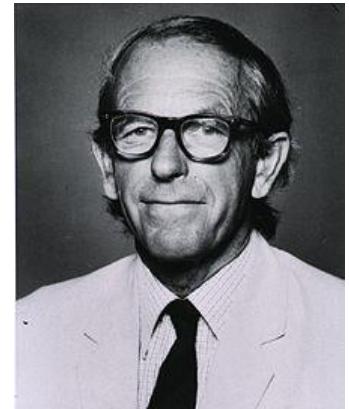
类别	RNA_seq	基因芯片	cDNA/EST 测序
原理	高通量测序	杂交	Sanger 测序
分辨率	单碱基	数 bp 至 100bp	单碱基
通量	高	高	低
对基因组序列的依赖性	低	高	无
背景噪音	低	高	低
基因表达检测水平	>8000 倍	≤数百倍	不能检测
能否分辨同分异构体	能	有限	能
能分辨等位基因表达	能	有限	能
需要的 RNA 样本量	低	高	高
费用	相对较低	高	高

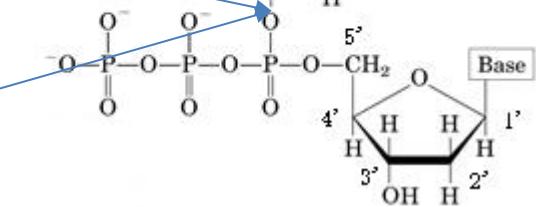
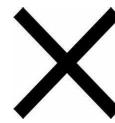
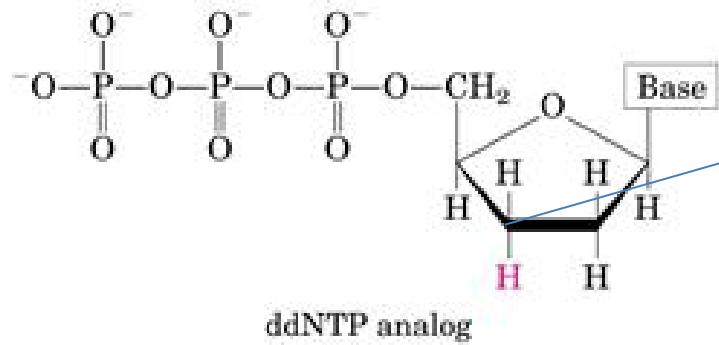
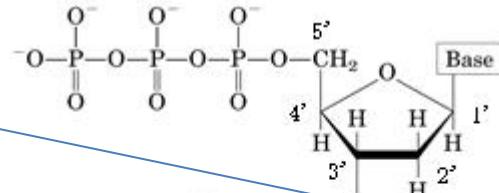
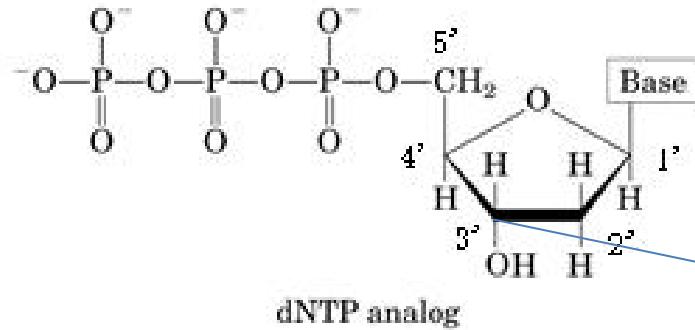
# 第一代测序法：sanger测序

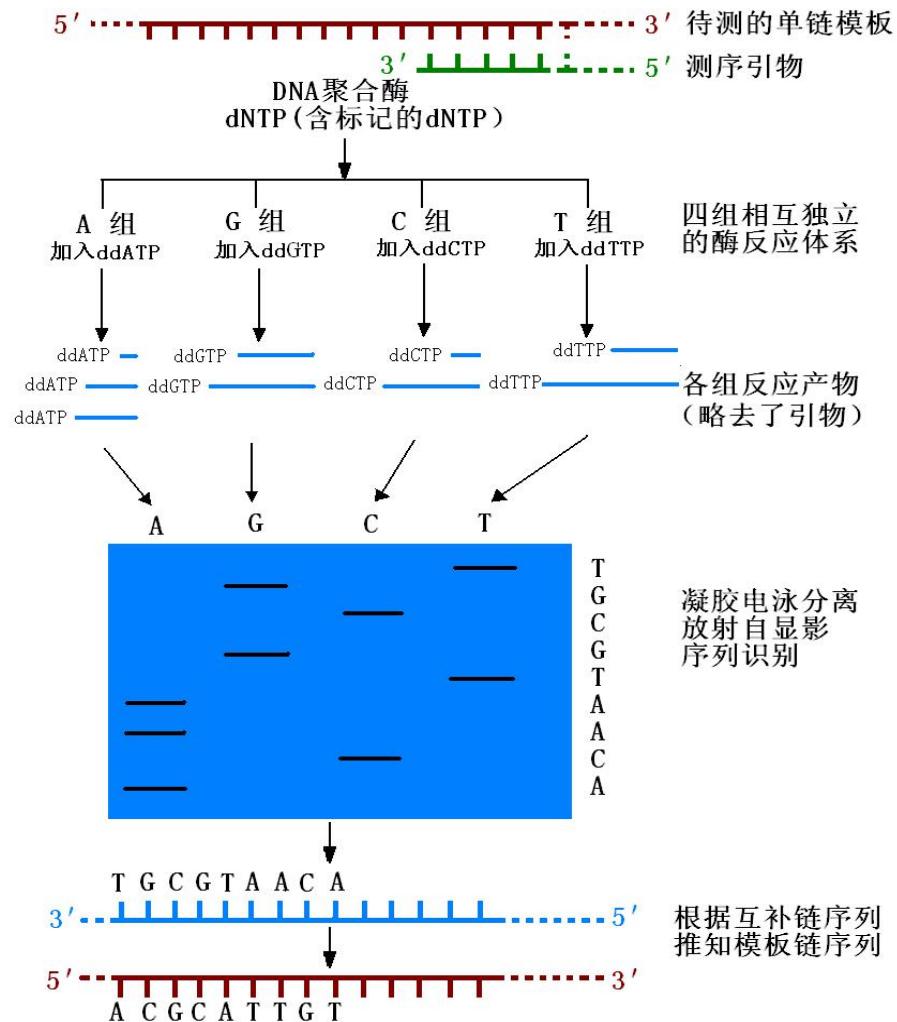


英国人Frederick Sanger 创建了双脱氧链末端合成终止法（chain termination method），简称Sanger法。

他发现如果在DNA复制过程中掺入ddNTP，就会产生一系列末端终止的DNA链，并能通过电泳按长度分辨。不同末端终止DNA链的长度是由掺入到新合成链上随机位置的ddNTP决定的。







# RNA-seq (RNA转录组测序)

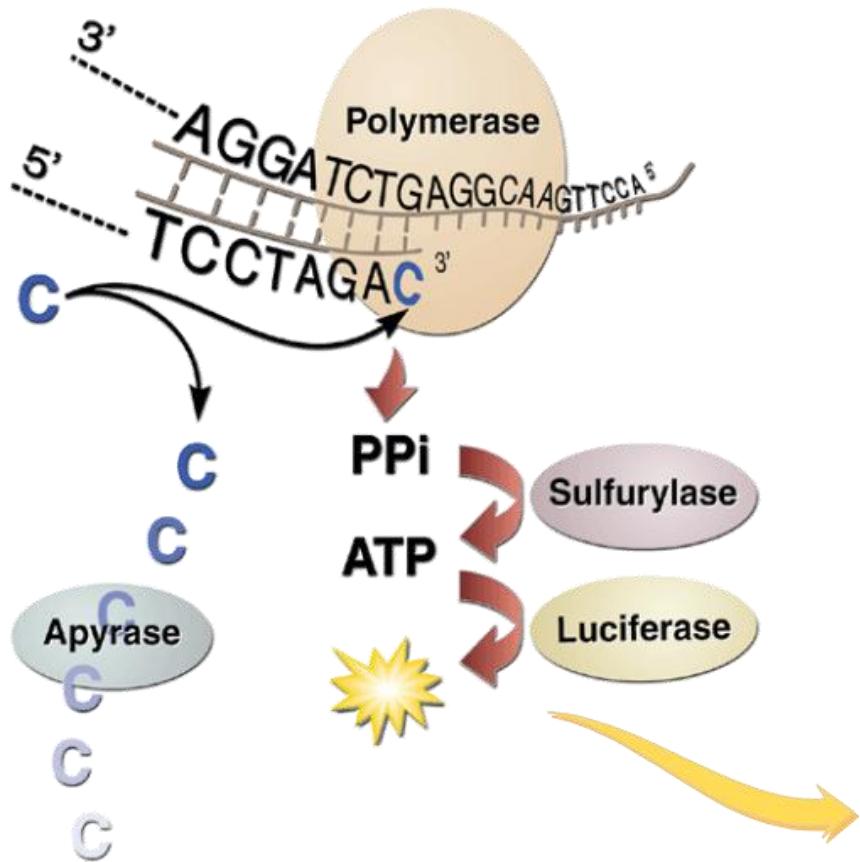
RNA测序又称转录组测序，就是把mRNA, small RNA和non-coding RNA (ncRNA) 全部或者其中一些用高通量测序技术进行测序分析的技术。

首先，我们获得细胞总RNA，然后对RNA进行片段化，然后反转录形成cDNA，获得cDNA文库，然后在cDNA片段的两端接上接头，最后用新一代高通量测序依进行测序。

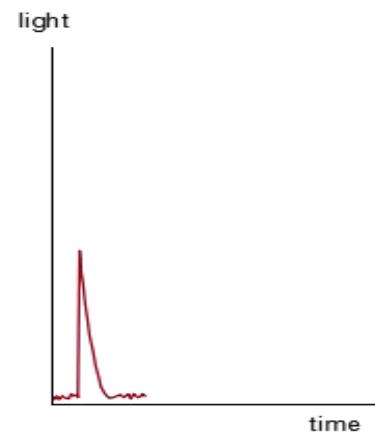
## 第二代测序技术（高通量测

- 高通量测序技术是对传统测序一次革命性的改变，一次对几十万到几百万条DNA分子进行序列测定，因此在有些文献中称其为下一代测序技术(next generation sequencing)。
- 自2005年以来，以 Roche公司的454技术、 Illumina公司的Solexa技术和ABI公司的SOLiD技术为标志的新一代测序技术相继诞生。之后 Helicos Biosciences公司又推出单分子测序 (Single molecule sequencing, SMS)技术。

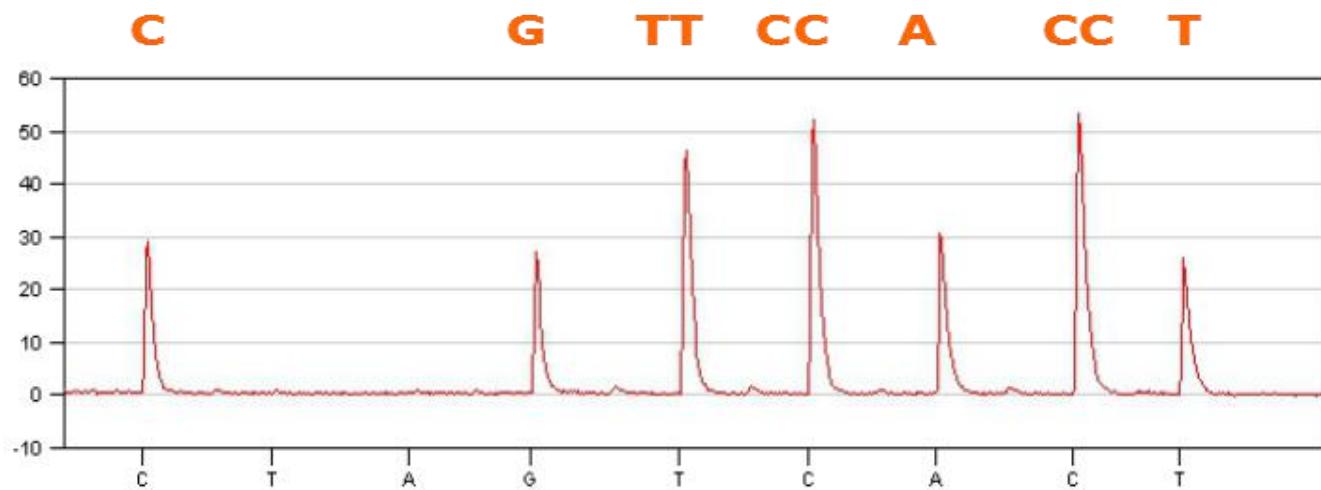
## Principle of Pyrosequencing



基于酶级联反应的分析技术



- 每次向反应体系中加入一种dNTP，相应位置的峰代表该种dNTP的掺入情况，峰高与掺入的核苷酸数量成正比，多余的dNTP和ATP在加入下一种dNTP前就被降解。



# • 蛋白质组学 (proteomics)

蛋白质组学是研究细胞或生物体内的所有蛋白质（即蛋白质组，proteome）的存在及其活动方式的学科，是在蛋白质水平上的后基因组学研究。

- 基因组计划的局限：

无法解决“基因精细调控”问题——大规模基因表达检测技术如：微阵列法、DNA 芯片无法反映蛋白质的质与量。

- 基因是遗传信息的源头，功能性蛋白是基因功能的执行体，以往人类对于蛋白质的研究只是针对生命活动中某一种或某几种蛋白质，难以形成一种整体观，难以系统透彻地阐释生命活动的基本机制。所以大规模、全方位白质研究势在必行

# 蛋白质组学的研究方法

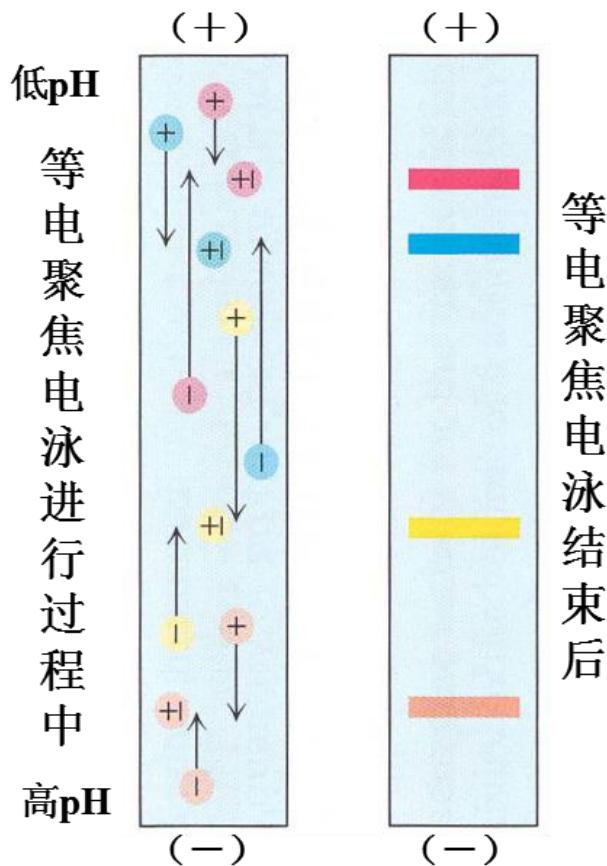
- 二维凝胶电泳（2-DE）和质谱（MS）技术是蛋白质组研究的核心技术，分别针对样品的分离和鉴定
- 双向凝胶电泳（two-dimensional electrophoresis, 2-DE）利用蛋白质等电点和分子量差异，结合凝胶化学特性，分离各种蛋白质的方法。

工作原理：

根据蛋白质等电点的不同进行第一向等电聚焦电泳分离  
转移到二向SDS-聚丙烯酰胺凝胶上，再根据相对分子量大小不同进行分离

## 第一向电泳：等电聚焦电泳 ( isoelectrophoresis focusing, IEF )

聚丙烯酰胺凝胶内的缓冲液在电场作用下沿电场方向在凝胶内制造一个pH梯度。每种蛋白质都将迁移至与它的 $\text{pI}$ 相一致的pH处。



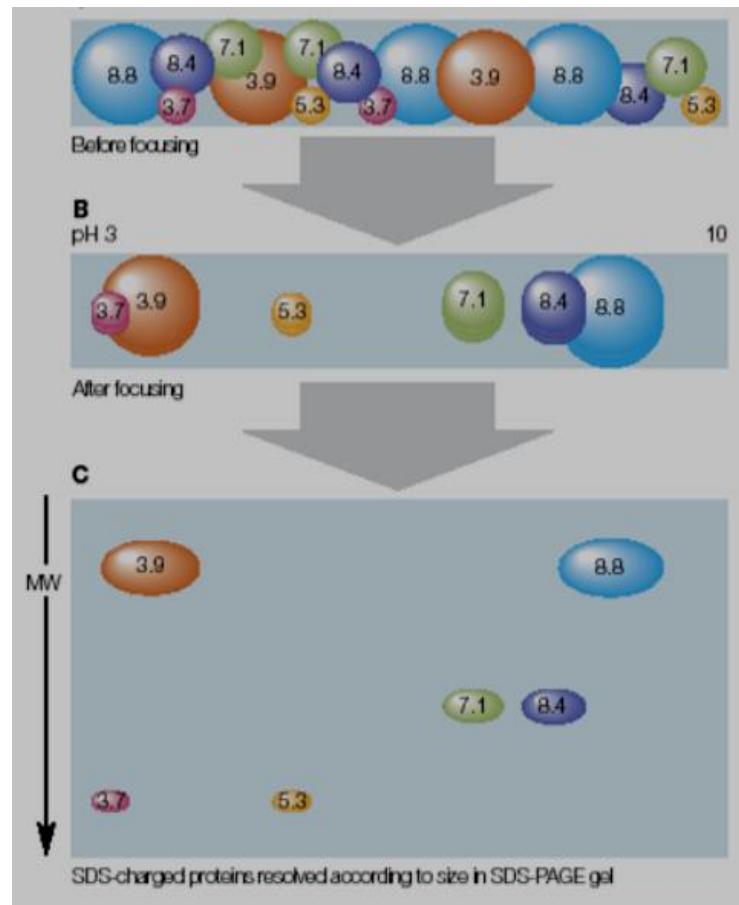
## 第二向电泳：聚丙烯酰胺凝胶电泳（SDS-PAGE）

根据蛋白质分子质量将样品分离

双向电泳后的凝胶经染色后蛋白呈现二维分布图：

水平方向反映出蛋白在 $pI$ 上的差异，

垂直方向反映出它们在分子量上的差别。



## 质谱（MS）法：

基本原理：样品分子离子化后，根据离子间质荷比（ $m/z$ ）的差异来分离并确定样品的分子量。

灵敏度高、快速、能同时提供样品的精确分子量和结构信息、既可定性又可定量、并能有效地与各种色谱联用来分析复杂体系。

# 代谢组学 (metabonomics)

代谢组学：“是通过考察生物体系（细胞、组织或生物体）受到刺激或扰动后（如将某个特定的基因变异或者环境变化后），其代谢产物的变化或其随时间的变化，来研究生物体系的一门科学” 许国旺 2008

- **Analytical platforms:**

- (1) Nuclear magnetic resonance (NMR); 核磁共振仪
- (2) Gas Chromatography–Mass Spectrometry ( GC-MS ); 气相色谱-质谱联用仪
- (3) Liquid Chromatography-Mass Spectrometry ( LC-MS ); etc. 液相色谱-质谱联用仪

代谢组指一个细胞、组织或器官中所有代谢组分的集合，尤其指分子质量为1,000以下的小分子物质。



- 质谱(MS)或与色谱联用技术具有普适性、高灵敏度和特异性等特点，已经在代谢组学研究中成为首选技术。
- 核磁共振技术 (NMR)，能够对样品实现非破坏性、非选择性分析。它是唯一既能定性，又能在微摩尔范围定量有机化合物的技术。缺陷是灵敏度相对较低，不适合分析低浓度代谢物。
- 代谢组学在肿瘤研究中的应用

近来的研究表明，不同肿瘤制备样品(比如培养细胞、组织切片、体内肿瘤块等)的代谢图谱与肿瘤类型、增殖、代谢活性以及细胞死亡有较强的相关性。对肿瘤代谢表型图谱的研究有助于人们了解肿瘤发生、发展以及致死的机制；在临床条件下，这些代谢图谱可以作为肿瘤诊断、预后以及治疗的评判标准。

# Omics data

NCBI Resources How To Sign in to NCBI

All Databases Search

COVID-19 is an emerging, rapidly evolving situation.  
Get the latest public health information from CDC: <https://www.coronavirus.gov>.  
Get the latest research from NIH: <https://www.nih.gov/coronavirus>.  
Find NCBI SARS-CoV-2 literature, sequence, and clinical content: <https://www.ncbi.nlm.nih.gov/sars-cov-2/>.

NCBI Home

Resource List (A-Z)

All Resources

Chemicals & Bioassays

Data & Software

DNA & RNA

Domains & Structures

Genes & Expression

Genetics & Medicine

Genomes & Maps

Homology

Literature

Proteins

Sequence Analysis

Taxonomy

Training & Tutorials

Variation

## Welcome to NCBI

The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

[About the NCBI](#) | [Mission](#) | [Organization](#) | [NCBI News & Blog](#)

**Submit**  
Deposit data or manuscripts into NCBI databases  


**Download**  
Transfer NCBI data to your computer  


**Learn**  
Find help documents, attend a class or watch a tutorial  


**Develop**  
Use NCBI APIs and code libraries to build applications  


**Analyze**  
Identify an NCBI tool for your data analysis task  


**Research**  
Explore NCBI research and collaborative projects  


**Popular Resources**

PubMed

Bookshelf

PubMed Central

BLAST

Nucleotide

Genome

SNP

Gene

Protein

PubChem

**NCBI News & Blog**

Easily download large amounts of genomic data with NCBI Datasets  
10 Sep 2020

Do you need to download a lot of genomic data? Maybe you need all  
04 Sep 2020

Hiding sequences in an alignment now available in the MSA Viewer!  
04 Sep 2020

Do you ever wish there was a quick way to hide partial or poor quality sequences?  
04 Sep 2020

GenBank release 239 is available  
04 Sep 2020

GenBank release 239.0 (8/18/2020) is now available on the NCBI FTP site. This release has 9.89 trillion bases and 2.12

<https://www.ncbi.nlm.nih.gov/>

# Omics data

The European Bioinformatics Institute (EMBL-EBI) maintains the world's most comprehensive range of freely available and up-to-date molecular data resources.

Developed in collaboration with our colleagues worldwide, our services let you share data, perform complex queries and analyse the results in different ways. You can work locally by downloading our data and software, or use our web services to access our resources programmatically.

— You can read more about our services in the journal [Nucleic Acids Research](#)

## Tools & Data Resources

### Tools

#### Clustal Omega



Multiple sequence alignment of DNA or protein sequences. Clustal Omega replaces the older ClustalW alignment tools.

[Multiple sequence alignment](#)

#### InterProScan



InterProScan searches sequences against InterPro's predictive protein signatures.

[Protein feature detection](#) [Sequence motif recognition](#)

#### BLAST [protein]



Fast local similarity search tool for protein

### Data resources

#### Ensembl



Genome browser, API and database, providing access to reference genome annotation

#### UniProt



A comprehensive resource for protein sequence and functional annotation.

#### PDBe



The European resource for the collection, organisation and dissemination of 3D structural data (from PDB and EMDB) on biological macromolecules and their complexes.

#### Europe PMC



<https://www.ebi.ac.uk/services>

### Browse by type

DNA & RNA	Gene Expression	Proteins
Structures	Systems	Chemical biology
Ontologies	Literature	Cross domain

### Programmatic access

EMBL-EBI web services allow you to query our large biological data resources programmatically, so that you can develop data analysis pipelines or integrate public data with your own applications. The Web Services technology we use are built on open standards to ensure client and server software from various sources

# Omics data

The screenshot shows the homepage of the China National Center for Bioinformation (CNCB-NGDC). The top navigation bar includes links for Databases, Tools, Standards, Publications, and About, along with language options for English and Simplified Chinese. The main header features the logo of the National Science & Technology Infrastructure (NSTI) and the text "国家生物信息中心" (China National Center for Bioinformation). A large green button on the right says "数据提交" (Data Submission). Below the header, there is a search bar with placeholder text "Find a bioproject, biosample, gene, protein, tool, database..." and a "Search" button. A message at the top states: "面向我国人口健康和社会可持续发展的重大战略需求, 国家基因组科学数据中心(NGDC), 作为国家生物信息中心(CNCB)的重要组成, 建立生命与健康大数据汇交存储、安全管理、开放共享与整合挖掘研究体系, 研发大数据前沿交叉与转化应用的新方法和新技术, 建成支撑我国生命科学发展、国际领先的基因组科学数据中心。" The main content area is divided into several sections: "热门资源" (Hot Resources) featuring BioProject, Bio Sample, GSA, GSA-Human, GWH, GVM, MethBank, GEN, eGPS Cloud, and DatabaseCommons; "最新资源" (Latest Resources) featuring a link to the "2019新型冠状病毒资源库"; "特色资源" (Special Resources) featuring Human Genomics (EDK, EWAS Data Hub, PGG.SNV, PGG.Han) and Biodiversity Databases (2019nCoVR, iDog, iSheep, IC4R, eLMSG, LSD, PED); and "Non-coding RNAs" and "Knowledge Associations".

<https://bigd.big.ac.cn/>

# Omics data

PPT来源：

**BioMedical Big Data and Precision Medicine**

**Yixue Li**

[yxli@sibs.ac.cn](mailto:yxli@sibs.ac.cn)

**CAS-MPG Partner Institut of Computational Biology**

**Shanghai Center for Bioinformation**

**Technology**

# Outline

- Concept and Background
- Driven by genomics technology
- Insight into biomedical big data
- Challenge and opportunity
- Bioinformatics is a key

# 精准医疗研究已成为新一轮国家之间科技竞争热点和引领国际发展潮流的战略制高点



## Precision Medicine Initiative USA 美国精准医学计划

“我希望这个消灭小儿麻痹与绘制人类基因组图谱的国家，能领导医学新纪元，能够在正确的时间为患者提供正确的治疗。……今晚我要发起新‘**精准医学计划**’，让我们离治愈**癌症**、**糖尿病**与其他疾病更近一步，并让我们所有人能获得让自己与家人更健康所需要的**个性化信息**。”

“I want the country that eliminated polio and mapped the human genome to lead a new era of medicine – one that **delivers the right treatment at the right time..... Tonight, I'm launching a new Precision Medicine Initiative** to bring us closer to curing diseases like **cancer** and **diabetes** – and to give all of us access to the personalized information we need to keep ourselves and our families healthier.”

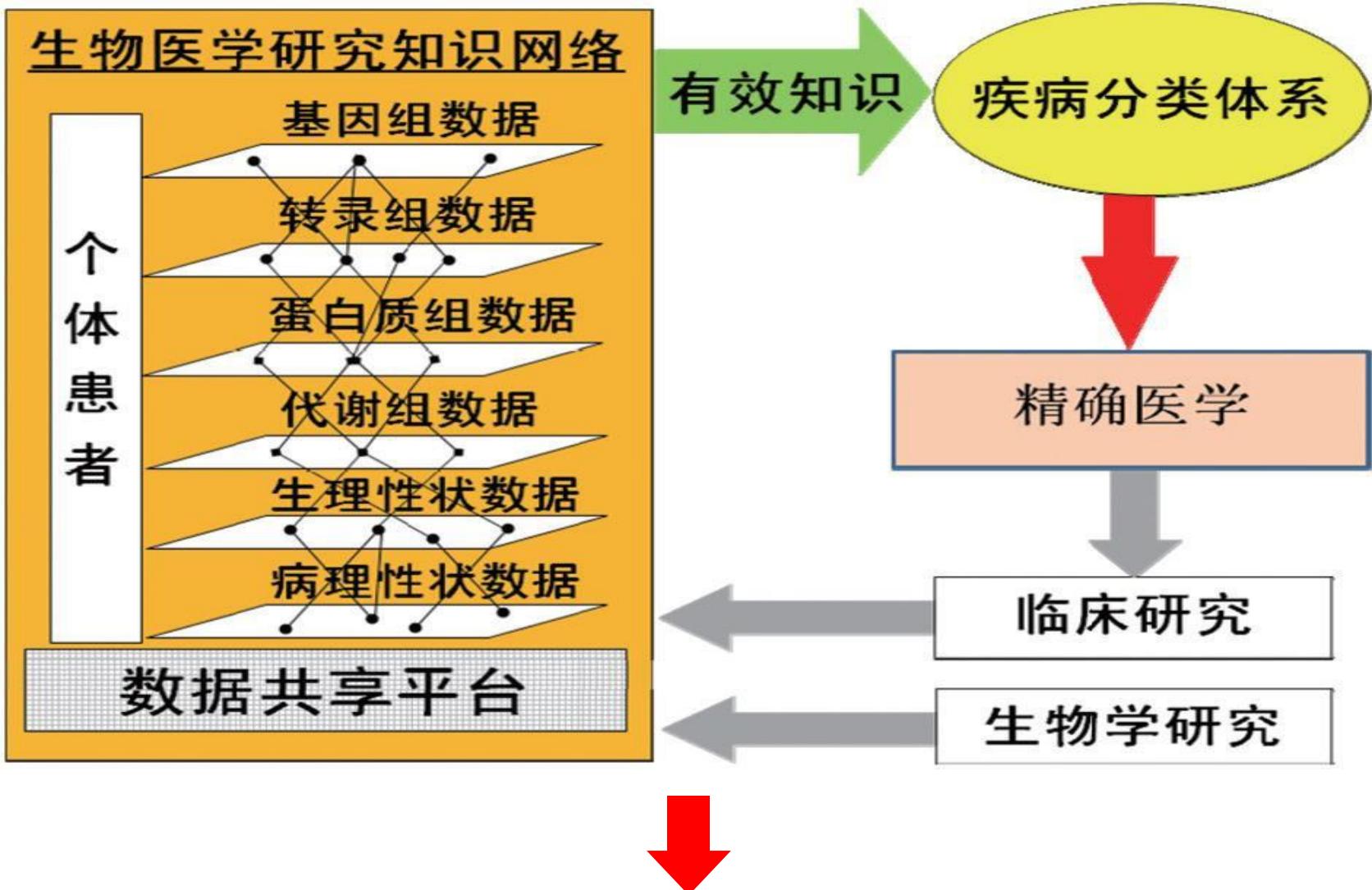
State of the Union Address (国情咨文2015) Tuesday, January 20, 2015

# 精准医疗概念

在大样本研究获得疾病分子机制的知识体系基础上，以生物医学特别是组学数据为依据，根据患者个体在基因型、表型、环境和生活方式等各方面的特异性，应用现代遗传学、分子影像学、生物信息学和临床医学等方法与手段，制定个性化精准预防、精准诊断和精准治疗方案。

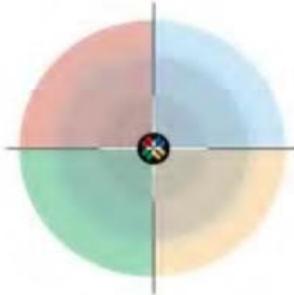
精准医疗背后是大数据！！！

# 精确医学与个体为中心的数据知识网络以及疾病分类关系



关键是生物医学大数据的整合、挖掘与知识发现与管理

# 基于大数据的精准的临床应 用 持续推动临床诊疗进步



药物基因组学检测  
能够帮助我们确定

The Right Drug

The Right Dose

The Right indication



诊断和治疗的遗传检测  
能够帮助我们确定

Familial screening for disease

Improve prognostic evaluation

Therapeutic decisions



肿瘤基因图谱可以  
指导靶向治疗

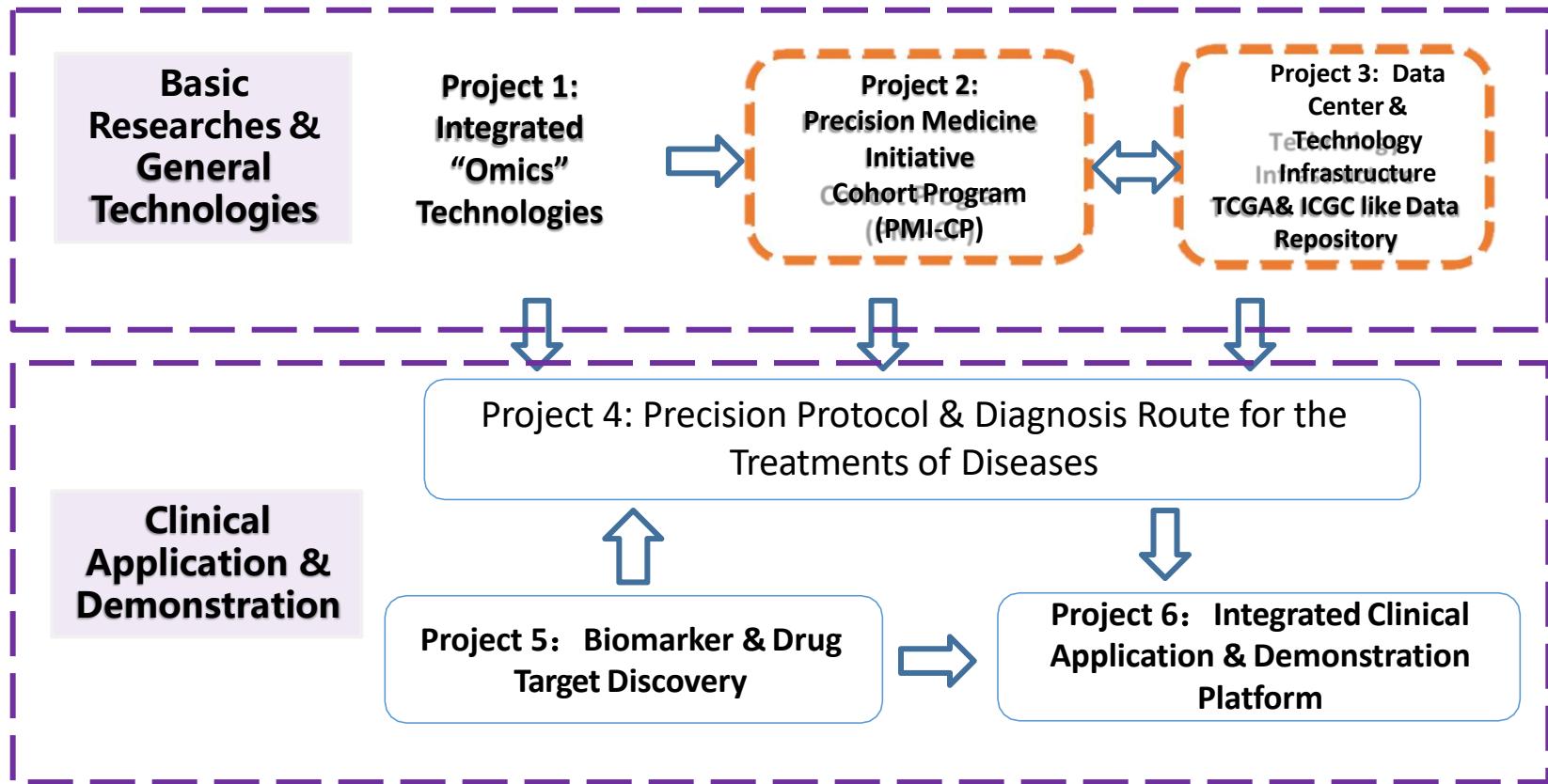
Clinically actionable results

Comprehensive coverage

Detects all types of cancer

causing mutations

# Chinese version of Precision Medicine



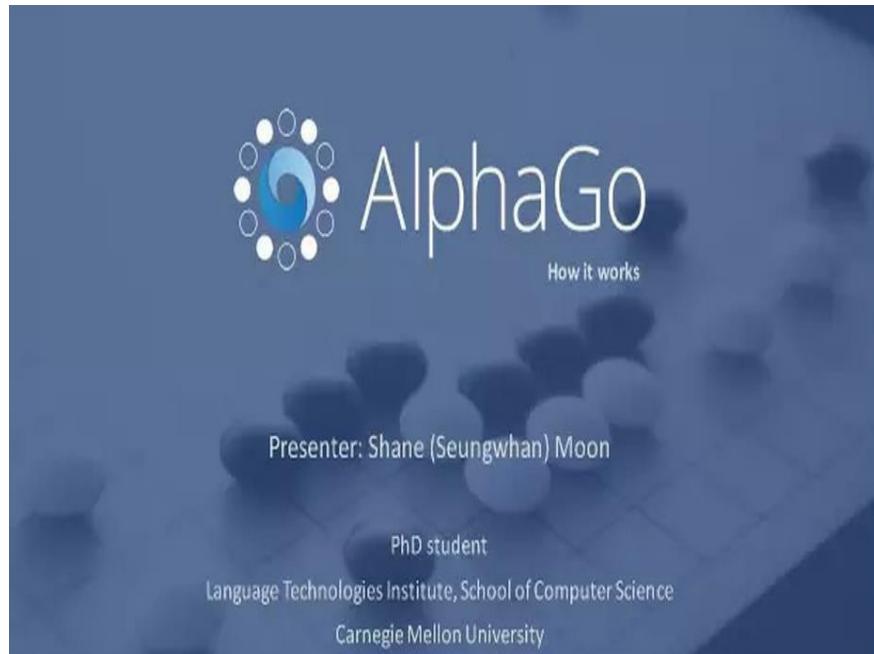
# 美国国家“癌症登月计划”与大数 据

2016年7月，在华盛顿举行的“癌症登月计划”（Cancer Moonshot）高峰会上，美国副总统拜登宣布了一系列推动提早终结癌症的新举措。其中**10**大重点方向：

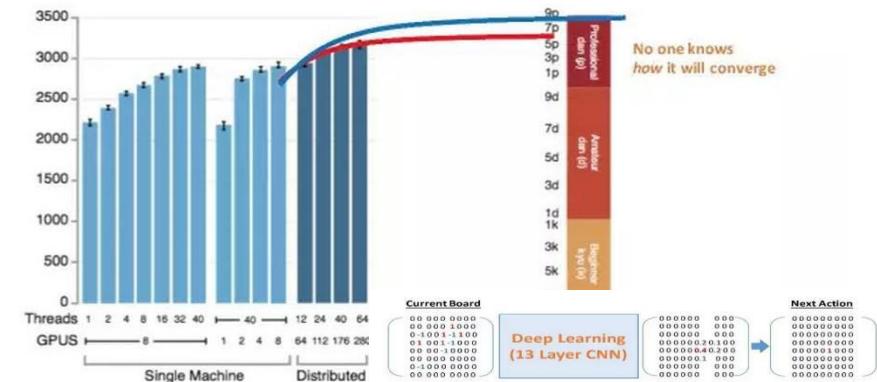
- 1、建立一个能让患者直接参与的癌症研究数据合作网络
- 2、建立一个专注于癌症免疫疗法临床试验合作网
- 3、研发克服肿瘤抗药性的方法
- 4、建立一个全国性的癌症数据生态系统
- 5、加强对于儿科癌症主要致病因素的研究
- 6、尽量减少癌症疗法使机体衰弱的副作用
- 7、扩大使用已被批准的癌症预防和早期诊断方法
- 8、通过挖掘过去的患者数据以预测未来患者的疗效
- 9、开发肿瘤三维图谱
- 10、开发新的癌症相关检测技术

大数据时代人的智  
力受到巨大挑战！





Taking CPU / GPU resources to virtually infinity?



AlphaGo learns millions of Go games every day

# Outline

- Concept and Background
- Driven by genomics technology
- Insight into biomedical big data
- Challenge and opportunity
- Bioinformatics is a key

# 新一代基因组技术的发展和应 用

# The Next Generation Sequencing Machine



Applied Biosystems  
ABI 3730XL  
1 Mb /day



Roche / 454  
Genome Sequencer  
FLX 1000 Mb/run



PacificBio™  
Single Molecule  
Sequencer  
3.5Gb /run



Illumina /  
Solexa/HiSeq2500  
Genetic Analyzer  
800 Gb/run

Applied Biosystems  
Ion Proton 12 Gb/run



HiSeq X Ten由10台HiSeq X测序仪组成，是定位为“测序工厂”模式的系统，适合运行于大型基因组测序中心，为各类生命科学和生物医学研究提供海量、高效率的测序服务。该测序仪每台每次运行仅需要3天时间，即可产出高达1.8Tb的数据，数据产出效率为现主流测序仪HiSeq 2000的12倍。整套系统每年可完成18,000人全基因组测序。

# Illumina NovaSeq6000 测序仪

## 100\$ 可实现一个人的全基因组测序

在2017年1月10号的摩根大通医疗健康年会上 Illumina 公布了2016年的市场表现和进入百美元基因组测序时代的最新测序仪 NovaSeq。其中 NovaSeq6000 型号的测序仪满负荷40小时内可以产出6TB的测序数据，即可实现200个人的样本的全基因组测序。每年可完成43800个人的全基因组测序，测序能力是 HiSeq X Ten 的单台机器的将近2.5倍。测序价格也许很快降到HiSeqX Ten的价格的十分之一左右。



NovaSeq 6000 系统

# 新一代测序技术可以干什么？

- 目标序列捕获测序技术 (Targeted Resequencing)
- 循环肿瘤DNA/细胞测序(ctDNA/ctcDNA)
- 免疫组库测序 (IR-SEQ)
- 单细胞转录因子结合位点测序(scATAC-seq)
- 大规模单细胞转录组测序(Drop-seq)
- .....

# IBM与Illumina合作将基因组数据解释智能化

- 2017年1月9日——IBM Watson健康和 Illumina公司宣布达成合作意向， 将整合基因组学智能计算平台Watson for Genomics与Illumina的BaseSpace Sequence Hub (Illumina cloud-based genomics computing environment for next-generation sequencing (NGS) data management and analysis ) 以及肿瘤测序流程来扩展对基因组数据的解释。希望能够促进简化基因组数据解释的流程并将其标准化。
- Watson for Genomics嵌入Illumina的新一代测序平台以后， 使用Illumina癌症基因组测序panel的研究人员将会迅速获得基因变异相关的功能和临床意义信息， 这有助于解释由TruSight Tumor 170 (170个基因的实体瘤基因检测panel) 生成的一系列变异数据。
- Watson for Genomics可在短短几分钟内读取TruSight Tumor 170生成的基因突变数据， 同时对大量专业指南、医学印刷品、临床试验手册以及其他知识来源进行梳理， 然后提供每个基因组变异的相关信息，并生成一份可供研究人员使用的报告——这一过程通常会花费一周以上的时间才能够完成。Watson for Genomics每个月可以解读大约10000篇左右科学论文以及100项新临床试验数据。

# Philips与Illumina合作将基因组数据解释临床化

- 此次合作旨在结合Illumina世界级的DNA测序技术、BaseSpace Sequence Hub和飞利浦基于云端的创新基因组学平台，获取、分析并解释癌症研究中的基因组学数据，并将基因组学连接到日常的医疗保健中，此过程将会整合患者数据，嵌入到临床路径中，并由实效证据和医疗补偿模式所支持。
- 整合Illumina用于大规模分析遗传变异与功能的测序系统和飞利浦的IntelliSpace Genomics临床信息学平台，飞利浦将与Illumina合作开发全新的解决方案，旨在从肿瘤学病例中获得、分析、注释与解释基因组学数据。这些数据将会从与仪器相连的Illumina BaseSpace® Sequence Hub中获得，然后通过飞利浦针对肿瘤学的IntelliSpace Genomics系统进行处理。
- 这样一个解决方案会从包括放射学、免疫组织化学、数字病理学、医疗记录以及实验室检测等多个渠道整合数据，然后在一个控制面板视图中显示出来。这个系统让研究人员更加高效地获取有价值信息，最终降低医疗成本，改善治疗效果。

# IntelliSpace Genomics & BaseSpace® Sequence Hub

- IntelliSpace Genomics是一个飞利浦数字健康平台驱动的生态系统，安全、基于云端技术，包括系统、临床应用以及数字化工具，具有大数据管理、预测分析、人工智能和物联网（IoT）功能，帮助临床医生在第一时间作出正确的决定。上述功能让飞利浦IntelliSpace Genomics能够随时随地为医生和专家提供关于治疗方案的可行信息。
- BaseSpace® Sequence Hub是一个基于云端的平台，延伸了Illumina仪器获取并分析基因组数据的功能，能够管理测序运行和Illumina测序平台并优化操作体验。一系列超值的Illumina与生态系统分析应用都可以在该平台上共享基因组数据。

序 技术， 投入30多亿美元， 15  
年间的 直接和间接产生经济效益高  
达投入 的**146** 倍， 达到**5000** 亿  
美元以上！

# Outline

- Concept and Background
- Driven by genomics technology
- **Insight into biomedical big data**
- Challenge and opportunity
- Bioinformatics is a key

# *Big Biological Data*

# Personalized genetic background

- 3 billions genomic DNA base pairs, 22k genes, 300,000 proteins,
- Personalized genomic difference: 6 millions bases

**“Theoretically,  $10^6$  scale may is a good population sample size for detecting certain phenotype related genotype.”**

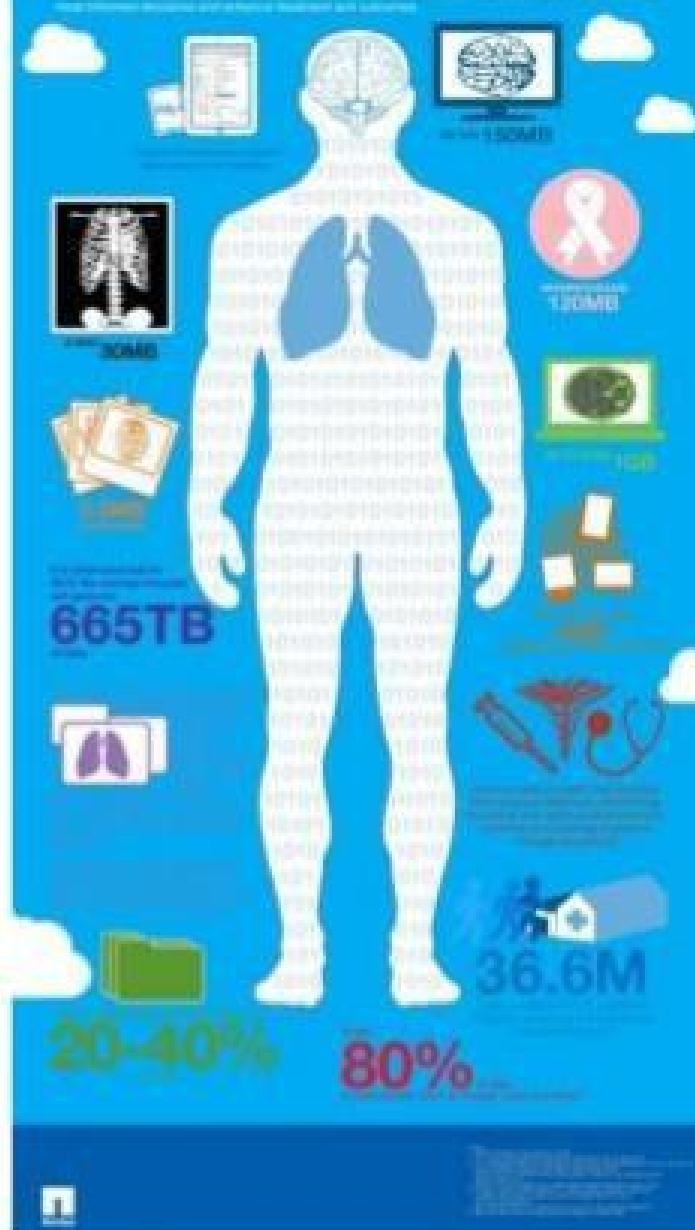
# Self omics? Big data in oncology



Oncology  
2005-2015  
140M patients  
100k hospitals  
1-10GB per patient  
140-1400PB  
80% unstructured

Source: Institute for Health Technology Transformation

## The Body as a Source of Big Data



# Outline

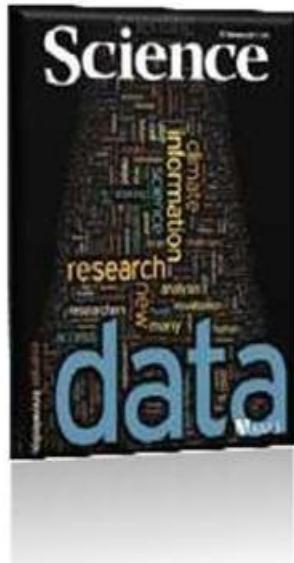
- Concept and Background
- Driven by genomics technology
- Insight into biomedical big data
- Challenge and opportunity
- Bioinformatics is a key

# 我们将面临何种挑战？

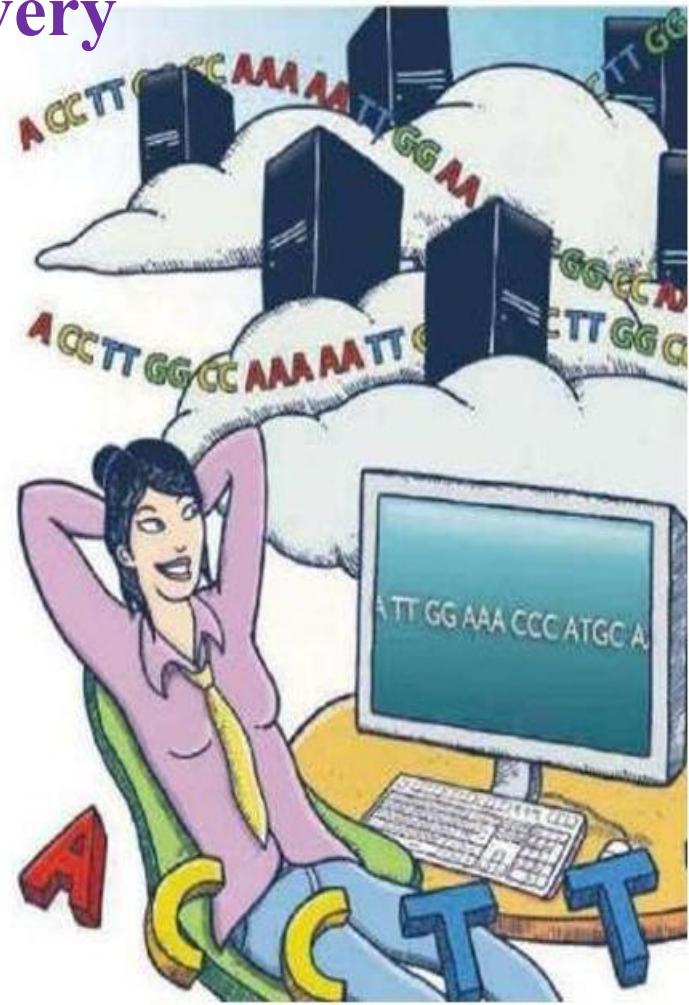
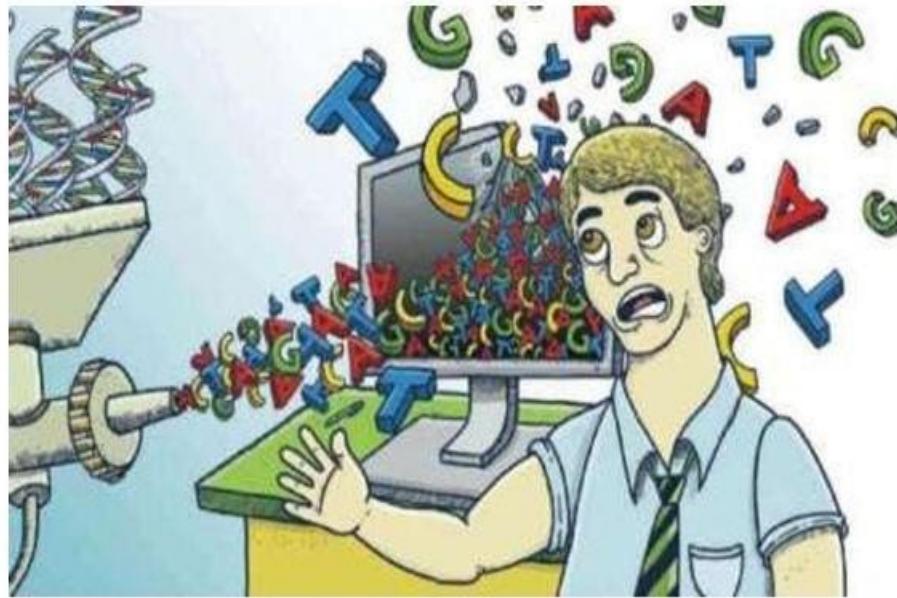
计算能力与急剧上升的生物数据之间的巨大差距



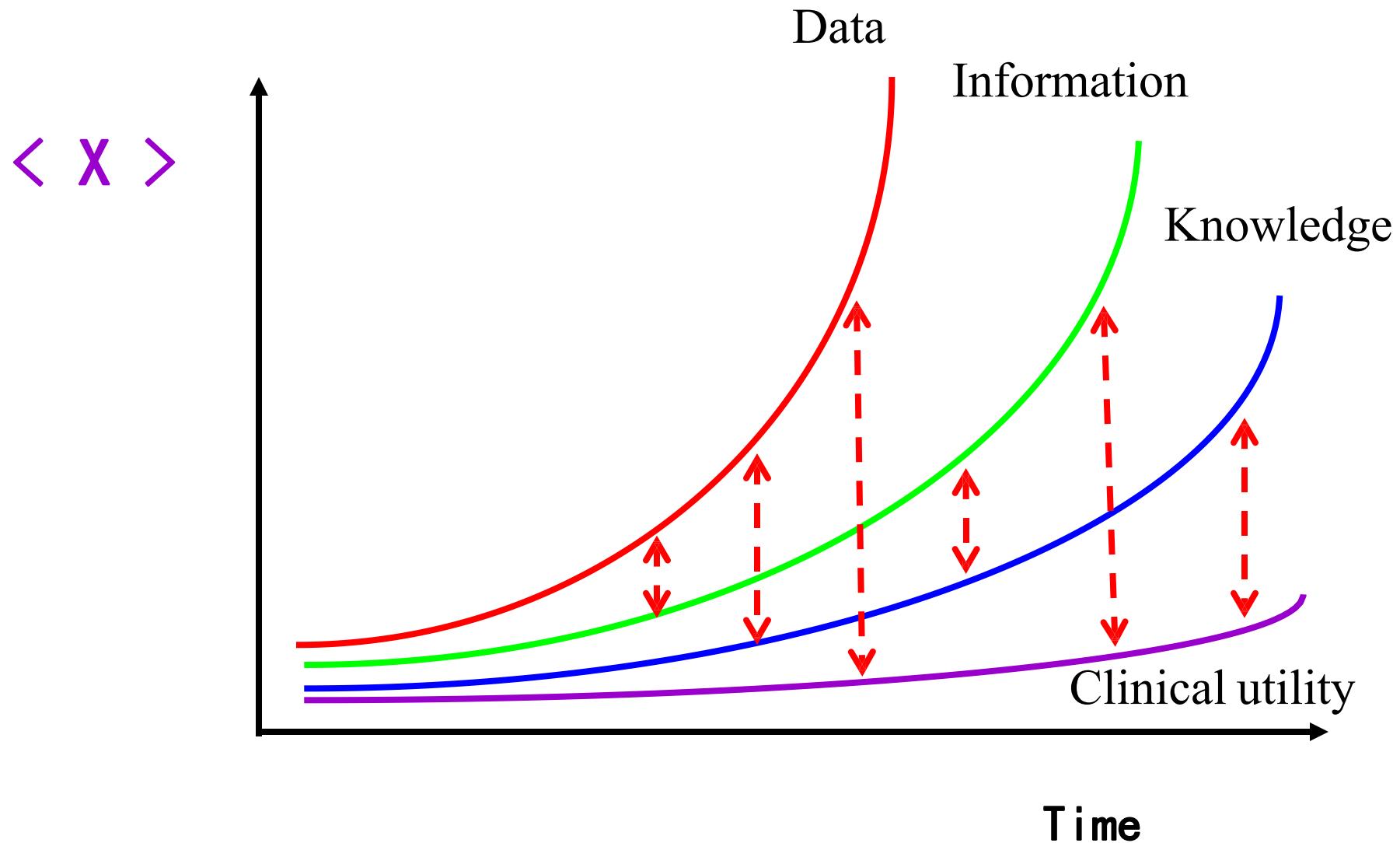
# Big gap between the dramatically increased biological data and our abilities of computation, data mining, and knowledge discovery



methods climate  
models including neuroscience  
brain cell analysis results 2010 technologies  
research information SPECIAL SECTION  
new work sharing many scientific access  
knowledge community example digital  
**Dealing with Data**



# Challenge resulted from “big data”



在大数据时代,生物医学发展的竞争

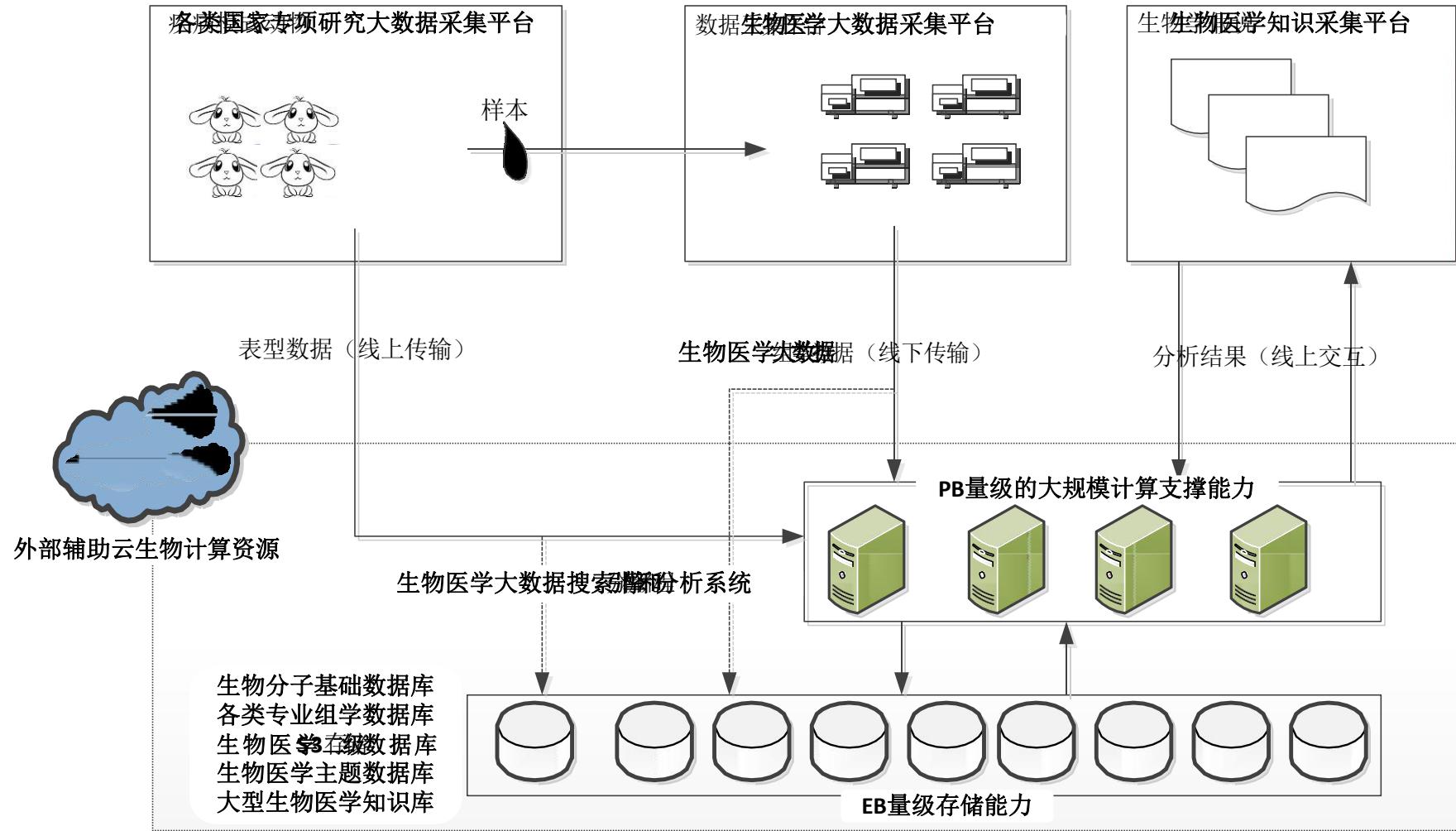
将 主要体现在如下三个方面:

数据资源和计算存储资

源 大规模数据分析处理

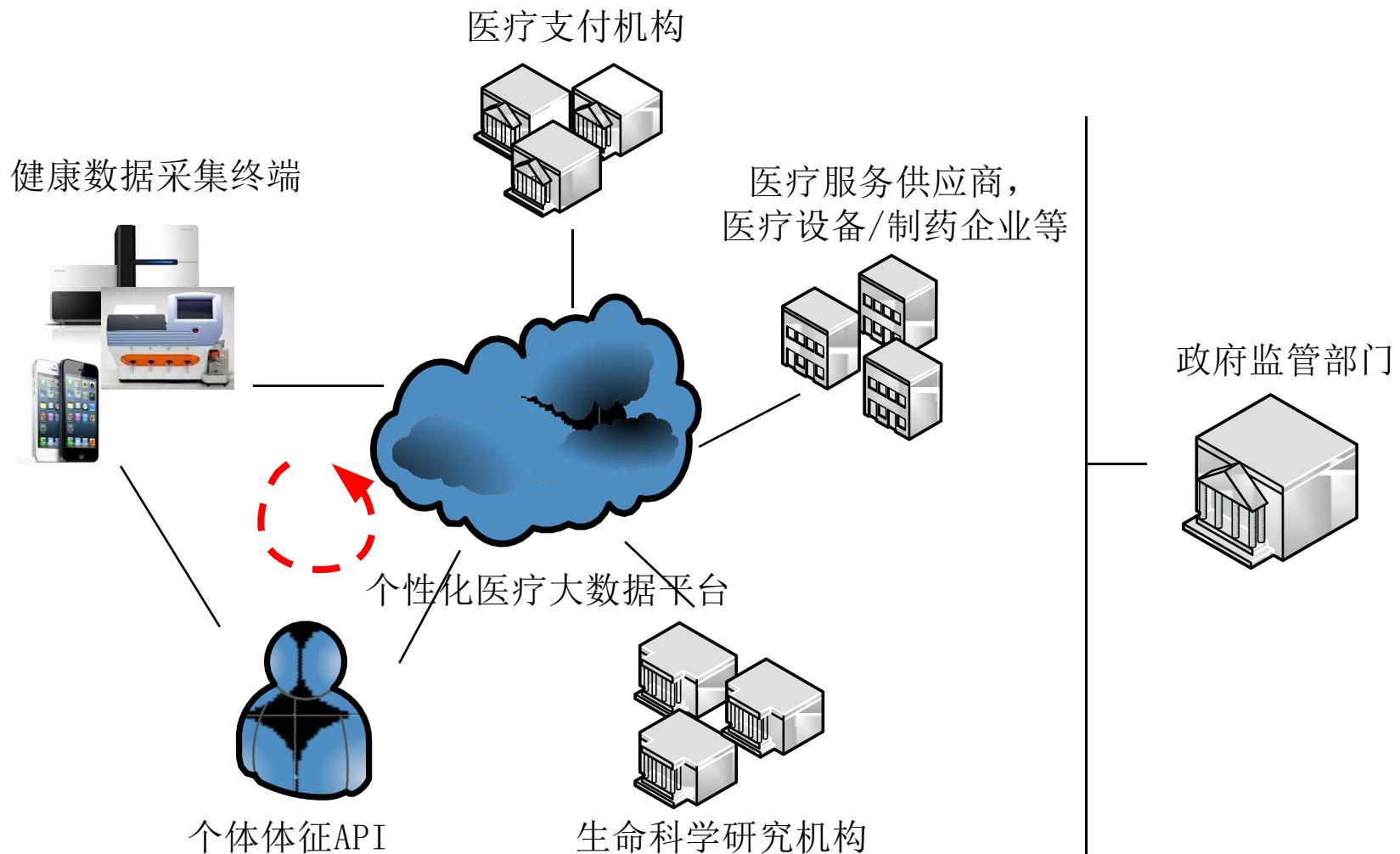
能力

# 各类生物医学大数据管理、汇交和分析整合平台



基于云平台的生物医学大数据多组学大数据管理平台

# 大数据基础设施为依托的健康服务 链



# 生物医学信息和医疗信息的整合 利用平台



Instrumented

人体状态随时随地感知与度量



Interconnected

信息互联互通，多方协同工作



Intelligent

健康评价、疾病预警，主动式干预

普适化

主动化

集成化

持续化

个性化

建立医疗信息与生物医学信息高度整合、知识快速形成与转化、决策支持全面渗透、临床多学科整体协同的疾病临床研究支撑平台

# 我们将面临何种机遇？

大样本 → 表型组+组学技术 → 大数据



## 临床应用

中科院、复旦大学、一批企业将启动百万健康人群表型组+GWAS大数据计划





23andMe

# Illumina HumanOmniExpress-24 format chip

**OmniExpress-24**芯片可容纳**12**个样本，每个样本可获得**70**多万个变异，单张芯片上总共有超过**800**万个数据点。每周能处理**1400**个样本，一台仪器*iSCAN*可以年处理**7**万个样本，获得**5600**亿个数据点。2012年的时候，**23andMe**就把个人基因检测价格降到了99美元，至今已经对超过一百二十万以上的人群进行了基因检测。可以说，目前来看，**23andMe**公司具有全世界最大的、质量最好的人类全基因组遗传变异信息数据库。

# 23andMe个人基因检测报告

- 始祖分析
  - 父母的祖先来源
- 健康分析
  - 饮酒能力
  - HIV/AIDS抗性
  - 药物反应
  - 遗传性疾病风险
  - 健康风险评估

- 2013年11月，美国食品药品管理局(FDA)要求23andMe暂停为新用户提供健康方面的基因检测服务。但FDA并未全面禁止其运营，仍允许23andMe为用户提供血统报告和原始基因数据。后来，FDA又批准了23andMe另外一款单一健康产品，用来预测布卢姆综合症(Bloom syndrome)。
- 通过合作，23andMe将允许合作伙伴访问其1000多种疾病相关数据，以便于他们寻找基因标记之间的新关联。这些合作伙伴将通过23andMe新建的一个研究网站来访问其数据。
- 在与辉瑞的合作中，23andMe允许辉瑞访问其研究平台，包括23andMe的服务和80多万人口的基因数据分析。在这庞大的数据库中，80%多的测试者(约65万人)同意参与研究。在合作初期，辉瑞将研究来23andMe的5000名狼疮患者的数据，以进一步了解狼疮基因。
- 与Genentech的合作中，将联合对3000名帕金森病患者的基因组测序数据进行分析，旨在找出治疗这种神经退行性疾病的新方案。在此次合作中，23andMe将负责收集帕金森病患者的数据，以及基因组测序工作，而Genentech将基于这些信息来制定潜在的治疗方案。

On April 6, 2017, the FDA identified 10 risk tests  
that 23andMe could include:

Parkinson's disease

Late-onset Alzheimer's disease

Celiac disease

Alpha-1 antitrypsin deficiency Early-onset

primary dystonia Factor XI deficiency

Gaucher disease type 1

Glucose-6-Phosphate Dehydrogenase deficiency

Hereditary hemochromatosis

Hereditary thrombophilia



基因检测迈出了一大步，23andMe获得了  
FDA的销售认可



基于液态活检的

肺癌精准用药案

例

# 肿瘤精准用药用于非小细胞肺癌治疗

- 患者，女，62岁。2015-8-3 因咳嗽、痰血2周，就诊入院，气管镜活检显示为非小细胞肺鳞癌。
- 诊断：右肺上叶鳞癌cT2bN1M1-IV期（对肺、右侧肾上腺）-IV期，活检组织标本基因检测显示：EGFR野生型；无可用的靶向药物

# 非靶向治疗阶 段

两药联合化疗4个周期（第一期治疗）：

- 化疗第一周期（2015-08-15）：吉西他滨联合顺铂；
- 化疗第二周期（2015-09-16）：吉西他滨联合奈达铂；

其后复查胸部CT，提示肺内病灶缩小，考评疗效：稳定（SD）。

- 化疗第三周期（2015-10-22）：吉西他滨联合奈达铂；
- 化疗第四周期（2015-11-24）：吉西他滨联合奈达铂；

# 非靶向治疗阶

## 段

化疗2个周期后复查胸部CT，提示肺内病灶缩小，考评疗效：稳定（SD）。

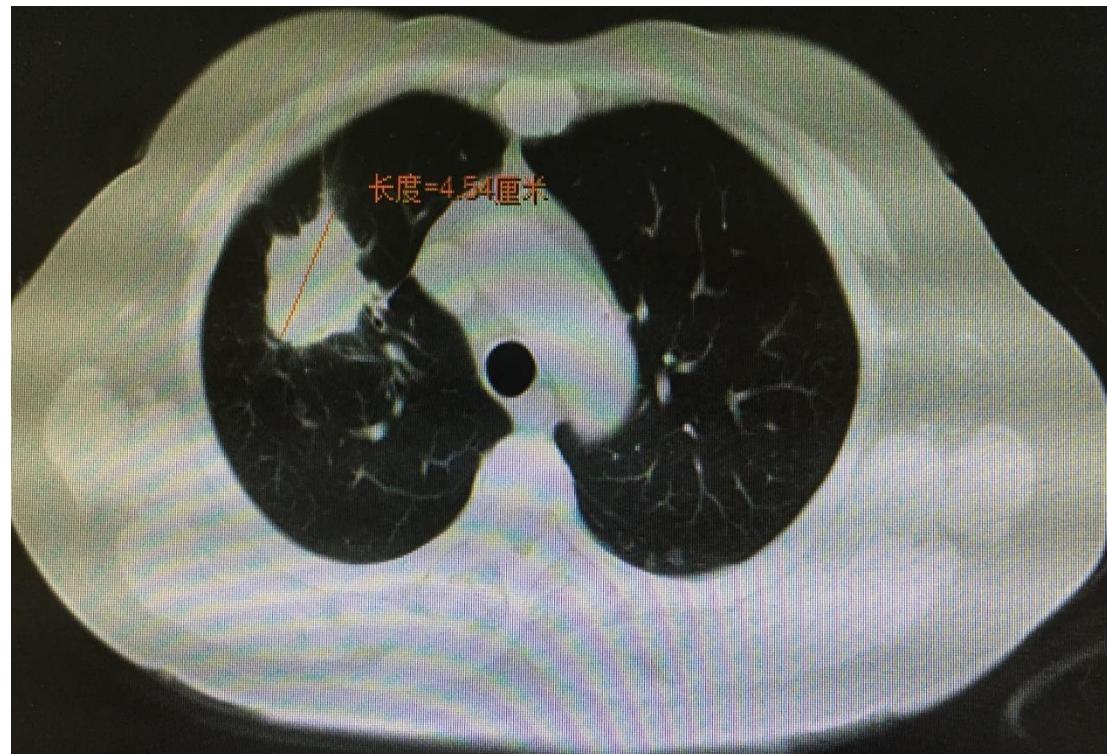


图1. 2015-10-14胸CT（联合化疗2段后，右肺肿块直径大小为4.54厘米）

# 非靶向治疗阶段

经过四个周期的化疗后，由于患者预后一般情况佳，肺内病灶在治疗过程中不断缩小，故其后给予吉西他滨单药维持治疗。

- 化疗第五至第十周期（2015-12-30—2016-05-30）：  
吉西他滨单药治疗五个月，共维持治疗5次。
- 一年后，2016-8-7复查胸CT，提示：右肺肿块及左肺部分小结节增大，两侧肾上腺转移，考评疗效：病情进展（**PD**），此时，患者预后情况较差，胸闷，乏力明显，食欲差，消瘦，化药治疗到这一阶段，已经产生耐药，且不能耐受进一步二线化疗。

# 非靶向治疗阶段

经过四个周期治疗后，由于患者预后一般情况佳，肺内病灶在治疗过程中不断缩小，故其后给予吉西他滨单药维持治疗（第二期治疗）：

- 化疗第五至第十周期（2015-12-30—2016-05-30）： 吉西他滨单药治疗五个月，共维持治疗5次。



图2. 2016-1-15胸CT（吉西他滨维持治疗1段后，右肺肿块直径大小由4.54厘米缩小至3.47厘米）



图3. 2016-8-7胸CT（吉西他滨维持治疗5段后，右肺肿块大小由3.47厘米增大至6.56厘米，病灶明显增大）

# 靶向治疗的可能性

评述：

在目前情况下，对于作为非小细胞鳞癌患者的一线化疗已经告一段落，治疗半年后，已经产生耐药，后半年保守治疗，预后效果不佳，病人身体状况无法继续二线化疗。此时，唯一的办法是寻找其它对症的治疗方案，以挽救患者生命，延长生存期。

近年来，液态活检技术成为指导肿瘤个性化用药的关键技术。液态活检是一种先进的癌症诊断技术，利用体液如血液、脑脊髓液（脑脊液）、血浆以及尿液进行癌症检测，也是癌症早期检测中具有前景的技术。作为一种非侵入性的诊断技术，液态活检以循环核酸、循环肿瘤细胞（CTC）及外泌体作为癌症诊断的生物标志物。研究人员通过临床试验发现，对于非小细胞肺癌患者，液体活检结果与组织活检结果匹配度近100%，更重要的是它还能实现组织活检不能实现的效果—检测癌症发展和指导临床治疗。

有研究表明：在对肺癌病人进行治疗时，有时组织活检时并未检测到EGFR、ALK或ROS1基因突变，但cfDNA检测却能够捕获到EML4-ALK易位和TP53突变，表明在一些情况下，cfDNA液态活检技术具有更为可靠的突变检出率。因此，在本病例情况下，对该患者进行液态活检，以期发现可能的治疗性驱动突变，指导临床决策，就成为临床大夫几乎是唯一的选择。

# cfDNA液态活检发现

一线化疗治疗一年后，在无法开展二线化疗的情况下，2016-08-10，利用产业技术研究院同肺科医院合作开发完成的技术平台，对肺科医院的这位非小细胞鳞癌患者进行了血液（cfDNA）检测，检查结果发现：病人的肿瘤细胞中出现了表皮生长因子受体基因EGFR上的敏感突变：p.L858R（突变频率：3.124%），表明患者体内目前处在活跃状态的是非小细胞腺癌肿瘤细胞亚克隆。

一年前，在对此病人进行组织活检样本检查时，并未发现EGFR基因的药物敏感性基因突变，也就是没有发现非小细胞腺癌的治疗性药物作用靶点。因此，我们在治疗时，对病人按照非小细胞鳞癌进行了一线化疗。现在看来，也许当时病人体内已经存在EGFR基因的药物敏感性基因突变，只是频率较低，组织活检并未检出，而液态活检是有可能检出的。

归纳一下，也许在治疗初期，组织活检没有检出药物敏感性基因突变，如果同时也做一下血液（cfDNA）检测，也许有可能检出药物敏感性基因突变，因此，可在一线治疗中对病人进行易瑞沙联合化疗治疗，也许可能是使此病人的治疗效果最大化的给药方式。

# 基于cfDNA的亚克隆分析

cfDNA液态活检结果的病理解释：

- 非小细胞肺鳞癌病人在初次液态活检时，没有发现腺癌亚克隆，也许并不等于肿瘤患者体内不存在腺癌肿瘤细胞亚克隆，实际上也可能存在，但是受到了鳞癌肿瘤细胞的竞争性抑制，数量稀少，难以通过组织活检检测发现。
- 然而在对鳞癌进行化疗后，如果疗效较好，反而鳞癌肿瘤细胞的增长受到了抑制，而原有的腺癌细胞得以不受抑制地增长。此时再进行液态活检，便可发现腺癌细胞及相应的亚克隆，此时治疗腺癌的靶向药物就可能发挥了作用。
- 在这样的情况下，是否在鳞癌治疗早期，鉴于cfDNA液态活检对于靶向药物敏感性基因突变的检出率要高于组织活检样本的检出率，有可能通过cfDNA液态活检技术，早期发现非小细胞腺癌的靶向药物敏感性基因突变，从而指导化疗药物和腺癌靶向药物联用，可能治疗效果和预后会更好。
- 综上所述，在非小细胞肺癌治疗的早期，无论病人确诊为鳞癌还是腺癌，均进行液态活检以便发现靶向药物敏感性基因突变可能是必要的。

# 基于cfDNA检测结果的精准用药

依据cfDNA液态活检结果的治疗：

- cfDNA液态活检显示病人的肿瘤细胞中出现了表皮生长因子受体基因（EGFR）的靶向药物敏感突变：p.L858R（突变频率：3.124%），表明患者体内目前处在活跃状态的是非小细胞腺癌肿瘤细胞亚克隆。
- 2016-8-27 更改治疗方案：进行易瑞沙靶向治疗；

# 基于ctDNA检测结果的精准用药

治疗效果：

- 易瑞沙靶向药治疗一周后，患者胸闷症状明显缓解，食欲改善，乏力减轻，皮疹1度；
- 易瑞沙靶向药治疗一月后，复查胸部CT示：肺内病灶明显缩小，结节吸收，疗效显著，疗效PR。
- 目前患者已经服用易瑞沙靶向药治疗3月，病情稳定，一般情况佳，无特殊不适。

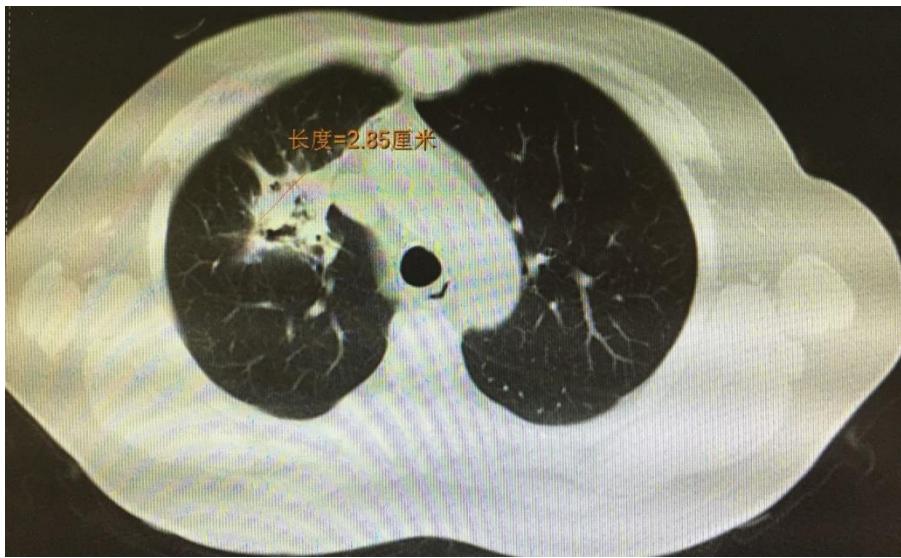


图4. 2016-9-10胸CT（易瑞沙靶向治疗1月后右肺肿块明显缩小，并出现坏死、空洞，直径由6.56厘米缩小至2.85厘米）

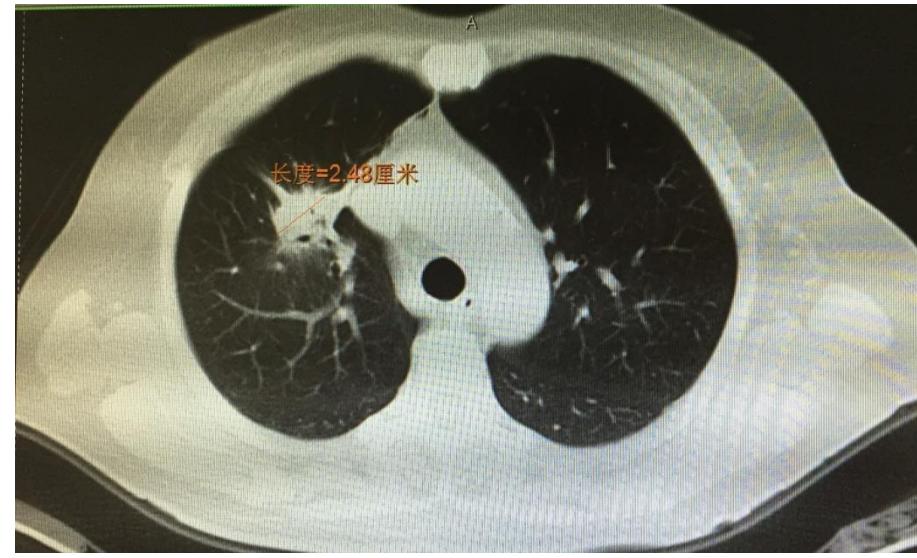
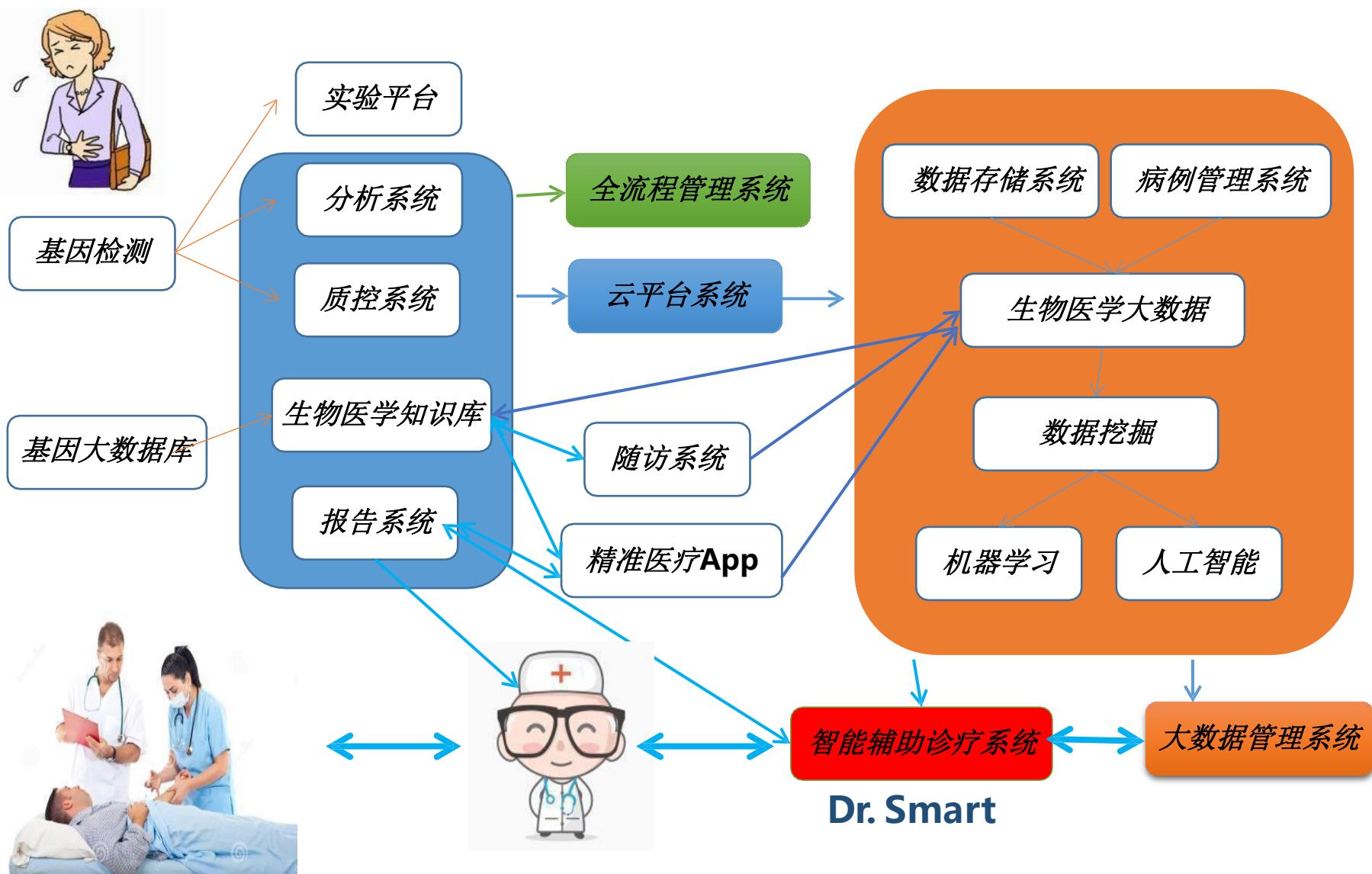


图5. 2016-11-12胸CT（易瑞沙靶向治疗3月后右肺肿块继续明缩小，直径由2.85厘米缩小至2.48厘米）

# 肿瘤精准医疗及生物医学大数据平台框架 图



# 常规化疗药物治疗阶段

非小细胞肺鳞癌  
EGFR基因突变野生型  
无可用的靶向药物

一线化疗（四个月）：吉西他滨+  
奈达铂  
治疗四个疗程，预后佳（SD）

一线化疗（五个月）：  
吉西他滨单药维持治疗，治疗五个  
疗程，肿瘤转移，预后差（PD）  
患者体质差，无法继续二线化疗

组织活检/病理检查



# 精准靶向药物治疗阶段

患者易瑞沙靶向治疗3月  
病情稳定，预后佳（PR）  
肺内病灶明显缩小  
结节吸收

6个“R”



一年后，血液cfDNA检查  
检出EGFR靶向药物敏感性突变  
患者呈现非小细胞腺癌基因型  
更改治疗方案，易瑞沙靶向治疗

# 精准医学/用药的6个“R”

- Right Targets
- Right Molecules
- Right Biomarkers
- Right Patients
- Right Combinations
- Right Dosing&Schedules

# 本病例治疗要点总结

- 1、在肿瘤晚期等组织标本不易获得的情况下，cfDNA液态活检是首选；血液样本更能够均匀的反映肿瘤的全貌，较好地避免了肿瘤异质性问题；
- 2、应该考虑在肿瘤治疗早期就开展cfDNA液态活检的可能性，因为cfDNA液态活检更有可能检出药物敏感性基因突变，依据cfDNA液态活检结果有可能为病人提出更为合理的治疗方案；
- 3、确诊为非小细胞鳞癌的病人也有可能存在未被检出的非小细胞腺癌肿瘤细胞克隆，可通过cfDNA液态活检跟踪非小细胞肺癌肿瘤细胞克隆演化的动态过程，以便早期发现靶向药物敏感性突变，对症开展药物的精准治疗。

# Outline

- Concept and Background
- Driven by genomics technology
- Insight into biomedical big data
- Challenge and opportunity
- Bioinofrmatics is a key

**Bioinformatics is a key !**

What is behind of the annotation  
of biological big data?

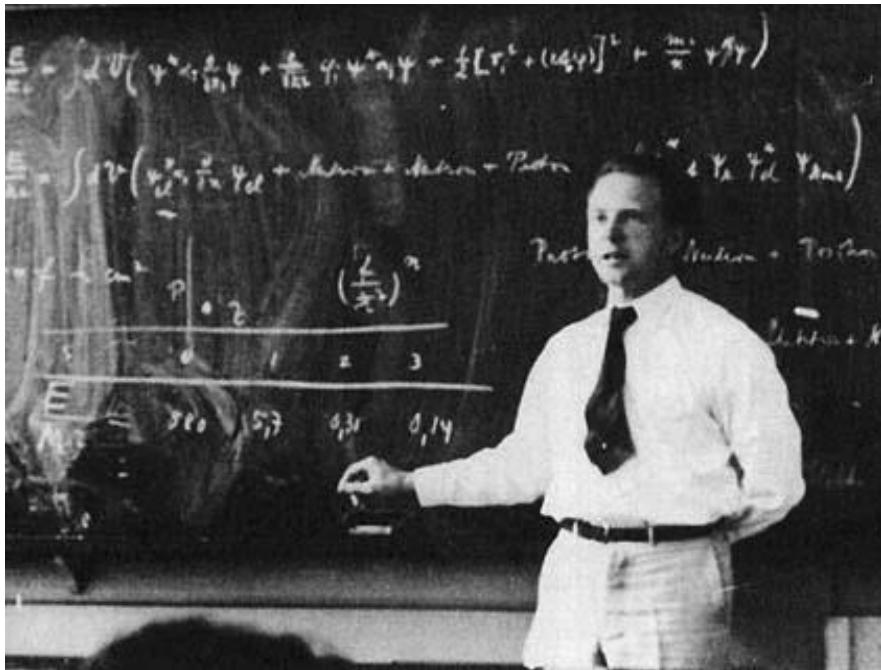
“Where is the wisdom we have lost in  
the knowledge? Where is the  
knowledge we have lost in  
**the information?** ”

**T.S. Eliot, a US poet**

Where is the information we  
have lost in the data?

# The Blind Men and the Elephant





***“What we observe  
is not nature itself,  
but nature exposed  
to our method of  
questioning.”***

Werner Karl Heisenberg, 1901-1976, 1932  
Nobel Prize in Physics

**Struggle for comprehensiveness — Systems Biology**



*“Good Design always helps, but that is usually restricted to a single study, and must be done **before** the data are generated.”*

*Terry Speed, TSINGHUA SANYA International Mathematics Forum on Big Data: Opportunities, Challenges and Innovations, Dec. 27-30, 2014*

# 大数据分析的九条规则

- **Rule 1:** Targeting
- **Rule 2:** Previous knowledge
- **Rule 3:** Data preparation
- **Rule 4:** Try-and-error(NFL, No Free Lunch)
- **Rule 5:** Intrinsic modules(DW rule, David Watkins)
- **Rule 6:** Vision to the fields
- **Rule 7:** Prediction and Assessment
- **Rule 8:** Values of discoveries
- **Rule 9:** Dynamic of the methodology





