组学前沿与应用研究

方向东　研究员
中国科学院基因组科学与信息重点实验室
中国科学院北京基因组研究所

Xiangdong FANG, M.D., Ph.D.

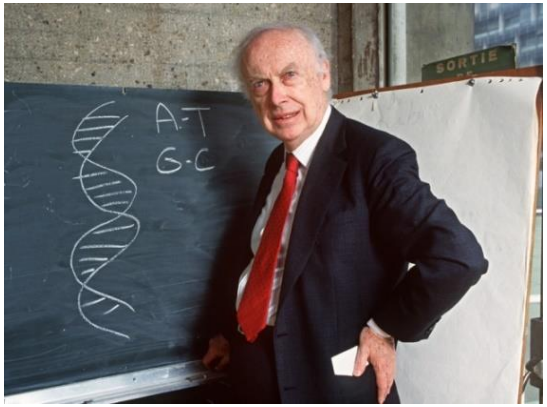
Professor, Beijing Institute of Genomics, Chinese Academy of Sciences

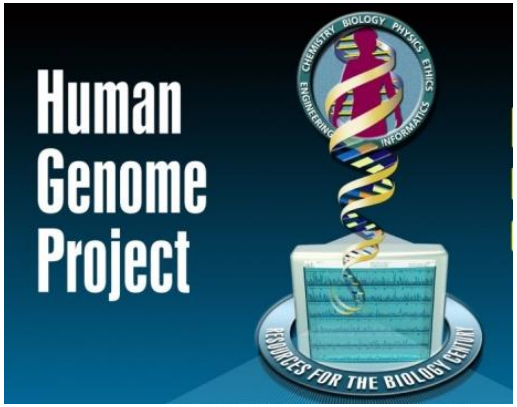
岭南科学论坛·双周创新论坛之精准医学，广州·2019.9.28

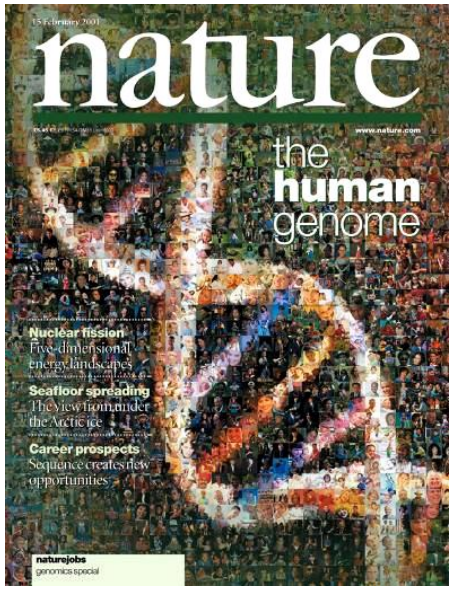
人类基因组计划 (HGP), 1990-2000

1984年，在美国犹他州的Alta，在美国能源部（DOE）DNA重组技术会议上，科学家第一次讨论了人类基因组测序的价值

1988年，DNA双螺旋结构发现者詹姆斯·沃森领导美国国家卫生研究院中新成立的基因组研究中心加入这个计划。沃森评价："不尽快将它（人类基因组计划）完成是非常不道德的"







Craig Venter (*Celera Genomics*), President Clinton and Francis Collins (*NIH*) making the historic announcement on June 26, **2000**.

1990年，投资**三十亿美元**的人类基因组计划由美国能源部和国家卫生研究院正式启动，预期15年完成

2000年6月26日，中美英日德法6国宣布HGP草图绘制完成

Collins FS, *et al.* ***Science*** 300: 286-290 **(2003)**

中国科学院北京基因组研究所 (BIG)

BIG

中国科学院北京基因组研究所
BEIJING INSTITUTE OF GENOMICS CHINESE ACADEMY OF SCIENCES

2003.11.28

华大基因 BGI

1999.9.9

中科院遗传所
人类基因组研究中心

1998.8

中科院重点实验室：基因组科学与信息、精准基因组医学

高通量组学技术解决重大科学问题

科研项目组	后期数据整合与综合研究	28 PI 组
大数据中心	数据处理与分析计算	70人
核酸测序部	全基因组水平的测序	17人


PacBio RS II


Illumina HiSeq


Ion Proton


3730*xl*


iScan


Sequenom


BeadXpress


Confocal

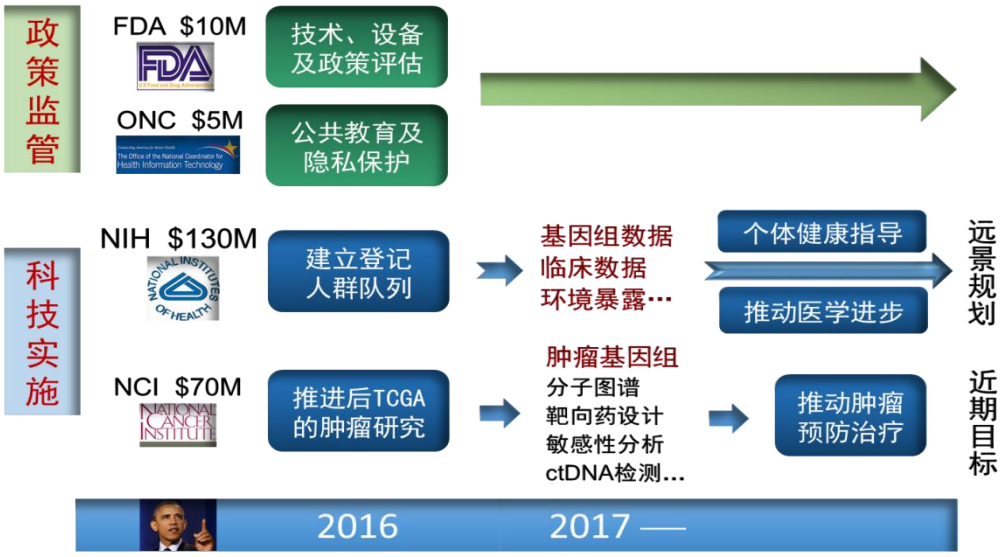

Server Cluster

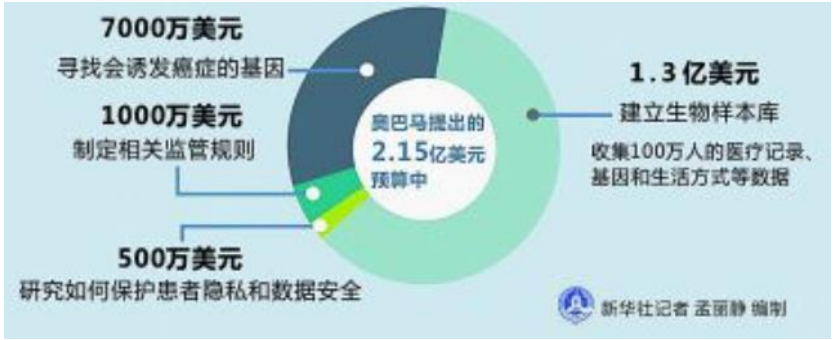
可产出数据800G/天。10P 磁盘+3P 磁带库存储
8000 CPU，理论浮点运算能力超过120万亿次每秒

精准医学，Precision Medicine

美国医学界2011年首次提出“精准医学”概念

奥巴马2015年1月国情咨文正式提出“精准医学计划” **Precision Medicine initiative**



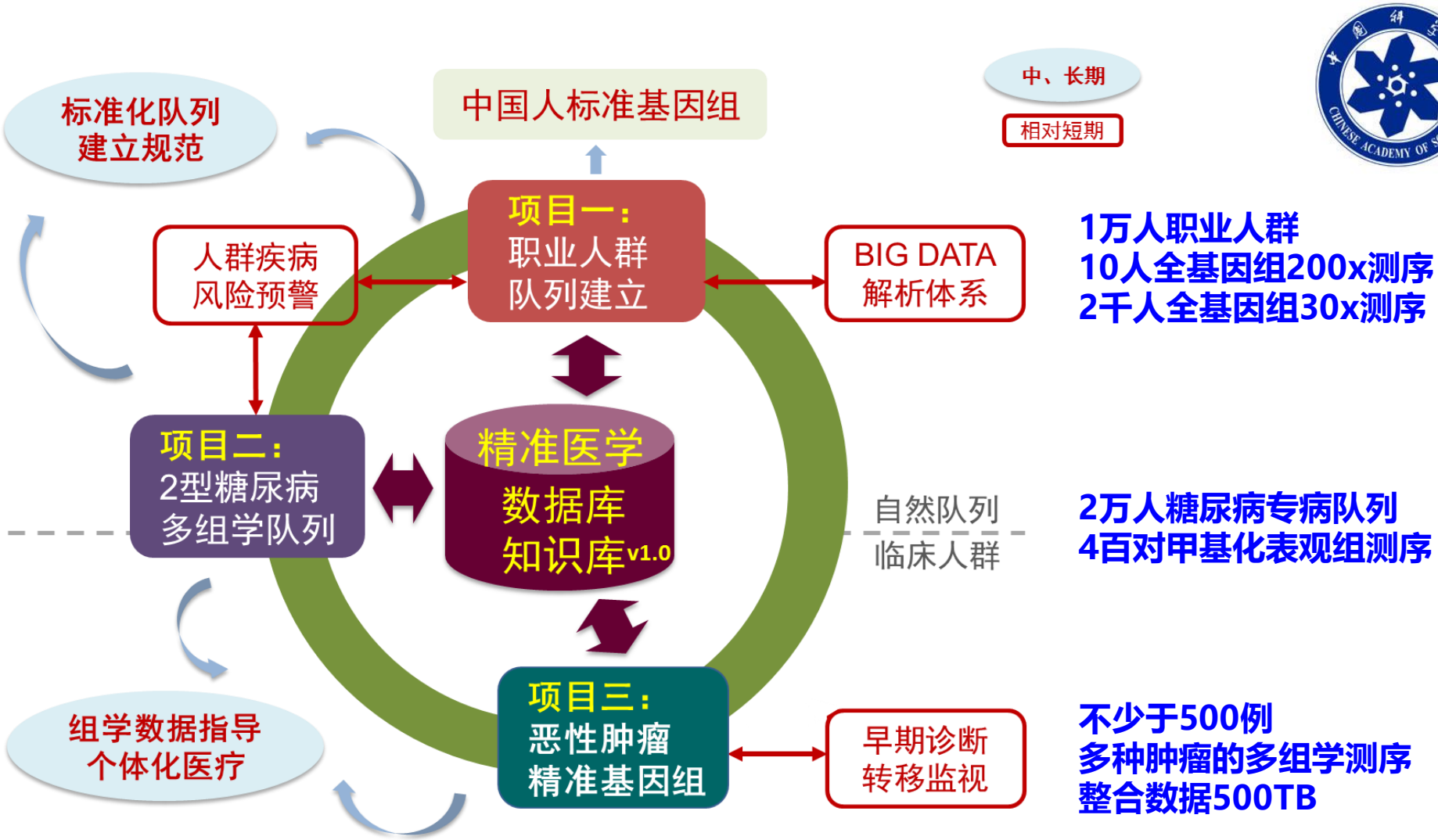


英国：世界上第一个将提供基因组医学作为日常护理一部分的主流健康服务体系”，以罕见病和临床肿瘤分析为主

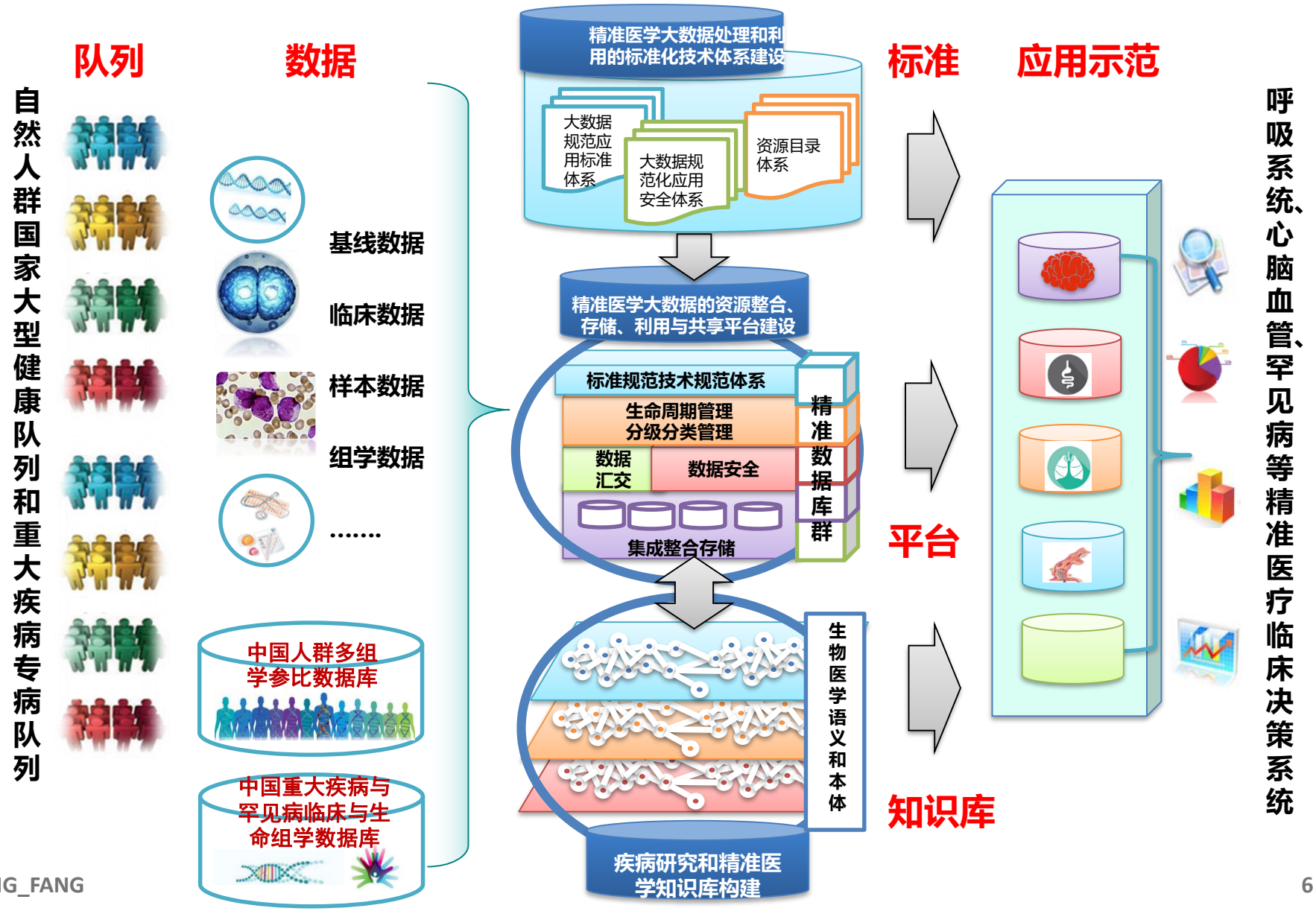
欧洲：以体系内病人测序（罕见病和肿瘤）为主

美加：兼顾，以Longitudinal健康计划（监测分析）为主

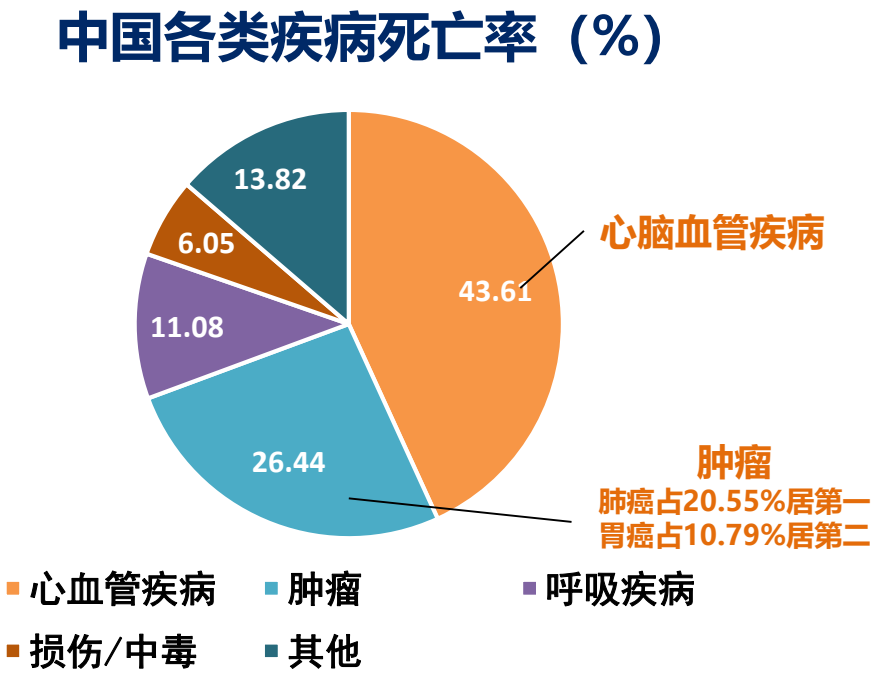
中科院人群精准医学研究计划（2015-2017）



“精准医学研究”重点研发计划 (2016-2020)



加强重大疾病防治是“健康中国” 重要战略



疾病防、诊、治全周期

亟需实时、精准的检测

- 全国每天1.2万人确诊癌症
- 我国患心血管疾病超2.9亿

基因组技术是“健康中国”战略的重要内容

我国健康服务产业发展增长潜力巨大。国务院总理李克强2013年8月28日主持召开国务院常务会议，提出“把健康产业作为国家支柱型战略产业，要求进一步加大改革力度，充分调动社会力量，加快发展内容丰富、层次多样的健康服务业”

到2030年，我国健康服务业的总规模可望超过12万亿元

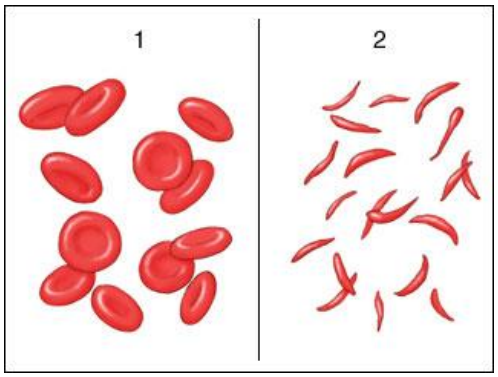
根据国家发展改革委《关于实施新兴产业重大工程包的通知》（发改高技[2015]1303号）的要求，2015年颁布的新型健康技术惠民工程将重点支持”基因检测技术应用示范中心”建设

2016年3月5日，国家发改委在“两会”公布“十三五”规划纲要(草案)。“加速推动基因组学等生物技术大规模应用”位列计划实施的100个重大工程及项目之中

镰刀状红细胞贫血，人类发现的第一个分子病

镰刀状红细胞贫血（SCD）是遗传性贫血症，属隐性遗传。患者β珠蛋白第六个氨基酸残基由正常的谷氨酸(Glu)突变为缬氨酸(Val)，导致血红蛋白结构异常，丧失输氧功能

患者的红细胞由于缺氧变成镰刀形，并因此破裂，造成严重贫血，甚至引起死亡

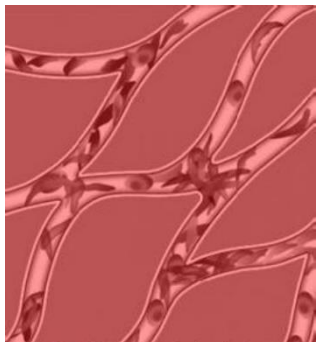


November 25, 1949, Vol. 110 SCIENCE 543

Sickle Cell Anemia, a Molecular Disease[1]

Linus Pauling, Harvey A. Itano,[2] S. J. Singer,[2] and Ibert C. Wells[3]

Gates and Crellin Laboratories of Chemistry,
California Institute of Technology, Pasadena, California[4]



1954, Nobel Prize in Chemistry
1962, Nobel Peace Prize

莱纳斯·卡尔·鲍林 Linus Carl Pauling
1901年2月28日 - 1994年8月19日
美国著名化学家，量子化学和结构生物学的先驱者

地中海贫血（Thalassemia）

临床常见的血液系统疾病，主要分α和β地贫两种。发病原因是由于血液红细胞中组成血红蛋白分子的珠蛋白肽链结构异常或合成速率异常，造成肽链不平衡，产生以溶血性贫血为主的症状群

我国重型和中间型地贫患者30万人，携带地贫基因超过3000万人，涉及近1亿人口，主要集中在长江以南广西、广东、福建、海南、湖南、江西、重庆、云南、贵州、四川等地，尤以两广地区最为严重，广东地贫基因携带者超过10%，广西地贫基因携带者超过20%

《中国地中海贫血蓝皮书》(2015)

Protein Cell 2015, 6(5):363–372
DOI 10.1007/s13238-015-0153-5

CrossMark Protein & Cell

RESEARCH ARTICLE

CRISPR/Cas9-mediated gene editing in human tripronuclear zygotes

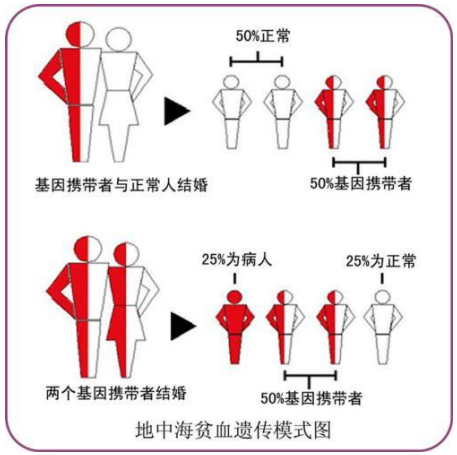
Puping Liang, Yanwen Xu, Xiya Zhang, Chenhui Ding, Rui Huang, Zhen Zhang, Jie Lv, Xiaowei Xie, Yuxi Chen, Yujing Li, Ying Sun, Yaofu Bai, Zhou Songyang, Wenbin Ma, Canquan Zhou✉, Junjiu Huang✉

Guangdong Province Key Laboratory of Reproductive Medicine, the First Affiliated Hospital, and Key Laboratory of Gene Engineering of the Ministry of Education, School of Life Sciences, Sun Yat-sen University, Guangzhou 510275, China
✉ Correspondence: hjunjiu@mail.sysu.edu.cn (J. Huang), zhoucanquan@hotmail.com (C. Zhou)

Received March 30, 2015 Accepted April 1, 2015

首例基因编辑胚胎

研究治疗地中海贫血



地中海贫血遗传模式图

在我国本土克隆疾病基因零的突破

夏家辉 院士，中南大学

1974年创建"遗传咨询门诊"，开展了染色体病的诊断和产前诊断

1975年发现与鼻咽癌相关标记染色体t(1;3)(q14;p11)

1981年将人类睾丸决定基因（TDF）定位于Yp11.32带

1985年开展了遗传资源的收集、保藏与利用

1998年克隆了人类耳聋疾病基因（GJB3）

letter

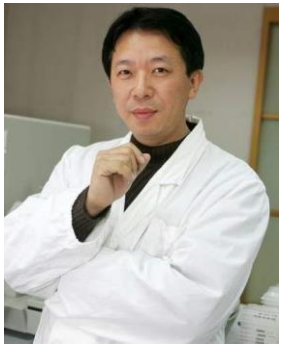
Mutations in the gene encoding gap junction protein β-3 associated with autosomal dominant hearing impairment

Jia-hui Xia[1], Chun-yu Liu[1], Bei-sha Tang[2], Qian Pan[1], Lei Huang[1], He-ping Dai[1], Bao-ro e[1], Dong-xu Hu[5], Duo Zheng[1], Xiao-liu Shi[1], De-an Wang[1], Kun Xia[1], Kuan-ping Yu[1], Xiao Yong Feng[3], Yi-feng Yang[5], Jian-yun Xiao[3], Ding-hua Xie[4] & Jian-zheng Huang[6]

Xia JH, *et al*. ***Nat Genet***. 20(4):370-373 **(1998)**

第一例正式以中国人姓氏命名的遗传病

通过对患病家系的遗传连锁分析，定位了第一例以中国人姓氏命名的罕见恒齿缺失的孟德尔**常染色体显性遗传病** **“贺-赵缺陷症”** 的致病基因



贺林 院士
交通大学

遗传性恒齿缺失一家系

赵双民　赵万里

先证者　男，61岁，汉族。现上下颌仅存8个磨牙，门牙（中切牙、侧切牙和尖牙）全无，嘴唇较正常人稍厚，说话语调低颤。身高、体重和智力发育无异常。据其母回忆，先证者乳牙萌出和发育均正常，6～8岁更换恒齿时门齿全部脱落未再萌出，其它恒牙也再生很少。

家系调查：已调查该家系6代269人，其中发现患者36人，其表现均与先证者相似，严重的满口无一牙齿，无牙的牙基部被牙龈样组织填充覆盖，嘴唇较正常稍厚，语调低颤，其他发育正常。家系成员中无近亲婚配。第4、第5代中有部分成员尚未到发病年龄，需继续追踪观察。还有部分家系成员有待继续调查。

讨论　该恒齿缺失的大家系是罕见的，其遗传方式符合常染色体显性遗传。因为：(1)连续6代均出现患者；(2)男女患者比例近似1∶1；(3)亲代不患病，子代中亦无患者。该病既不同已发现的“齿质形成不全”，也不同于“Christ-Semens 综合征”。齿质形成不全的特征是有牙而牙体发育不全，且牙齿具有一种特殊的乳褐色光泽[医学遗传学原理. 第1版. 哈尔滨：黑龙江科学技术出版社，1984：53～54]；而 Christ-Semens 综合征所指的牙齿缺陷是先天性的，乳牙萌出时就缺失，其遗传方式为 X 连锁隐性遗传[中华医学遗传学杂志，1992(4)：233]。我们所报告的病例其病变是后天发育过程中才出现的，所以也不同于先天性无齿畸形，其病因和相关的一些问题有待进一步研究。

本文初稿承复旦大学谈家桢、刘祖洞、薛京伦教授，江绍慧副教授；陕西师范大学谦振民教授、安书成副教授；南京师范大学陈俊才教授，张益清主任；华东师范大学周本湘、李难、范培昌教授以及中国科技大学李振刚教授审阅，特致感谢

（收稿：1992-10-09　修回：1993-04-16）

（本文编辑　王昌淑）

本文作者单位：711300　陕西省旬邑县中学（赵双民）；旬邑县医院（赵万里）

“贺-赵缺陷症” 是指人乳牙脱落后，机体丧失了替换恒牙能力的一种遗传疾病。1985年，我国陕西省旬邑县中学生物学教师赵双民和其堂叔-县医院内科医师赵万里，在陕西省发现一个**6代269人家系**中36人乳牙脱落后长不出恒牙

中华医学遗传学杂志 10(6): 373 **(1993);** ***J Dent Res*** 80(8): 1716-1720 **(2001)**

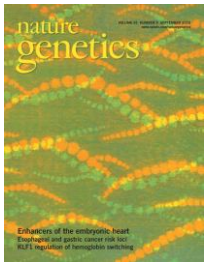
中国食管癌人群研究

林东昕 院士

詹启敏 院士

王立东 教授

中国医学科学院肿瘤医院/北京大学医学部/郑州大学/新乡医学院



Nature Genet
2010; 42: 759

Nature Genet
2011; 43: 679

Nature Genet
2012; 44: 1090

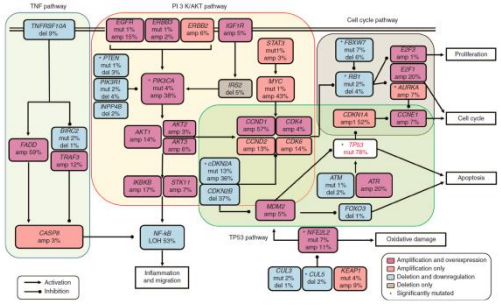
Nature Genet
2013; 45: 632

Nature Genet
2014; 46: 1001

Nature
2014; 509: 91

Nature Commu
2017; 8: 15290

- 60余万例食管癌的健康/医疗标准数据
- 10,000例以上食管癌GWAS数据
- 400例以上完整基因组、转录组、蛋白组数据
- 多篇论文发表在*Nature*，*Nature Genetics*



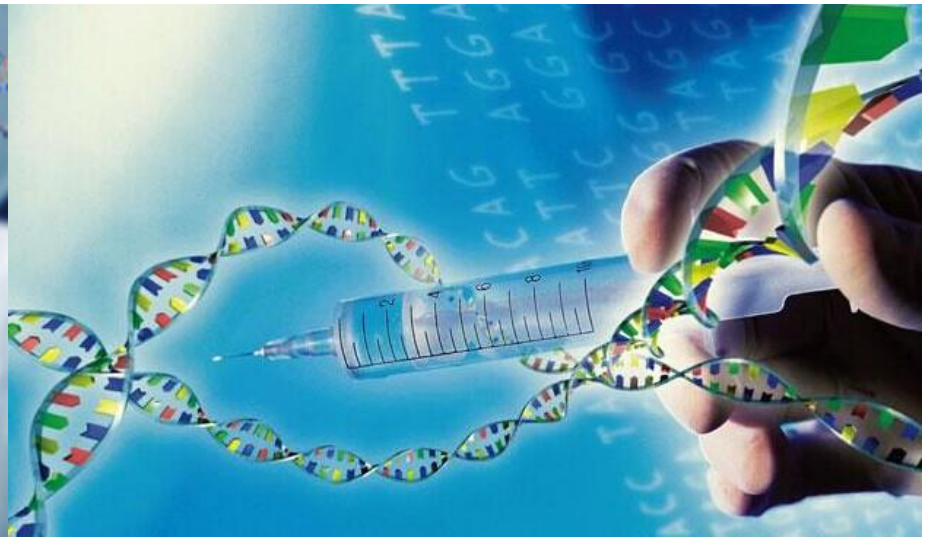
食管癌相关染色体区域、易感位点和信号通路

基因组技术应用于精准医学和健康管理

以基因组、转录组、蛋白组、表观组、代谢组等多组学信息整合分析为核心的“精准医学”，已成为现代医学发展的重要模式，将对医学各个领域产生颠覆性的影响

精准医学发展趋势和前景展望

- ✓ 基因检测是关键技术
- ✓ 数据解析是核心保障



基因检测的核心技术和发展阶段

基因检测的核心技术：

聚合酶链式反应，PCR

基因芯片技术，Gene Chip

DNA测序，DNA Sequencing

荧光原位杂交技术，FISH

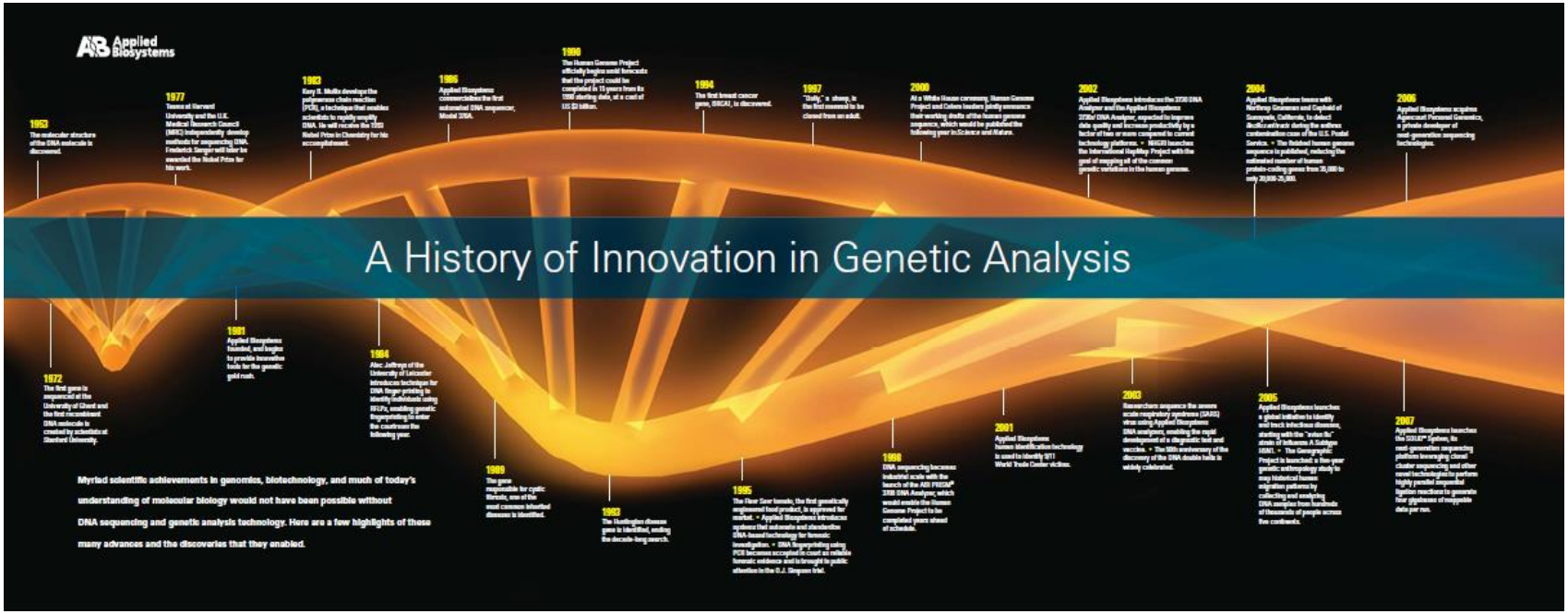
DNA印迹技术，DNA Blotting

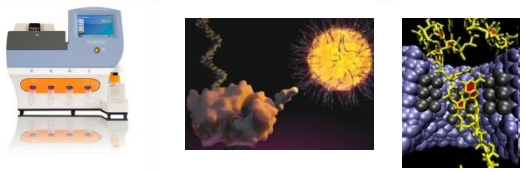
单核苷酸多态性，SNP

连接酶链反应，LCR

引自：火石创造《中国分子诊断产业图谱》行业研究报告2017

基因组学测序平台的发展历程





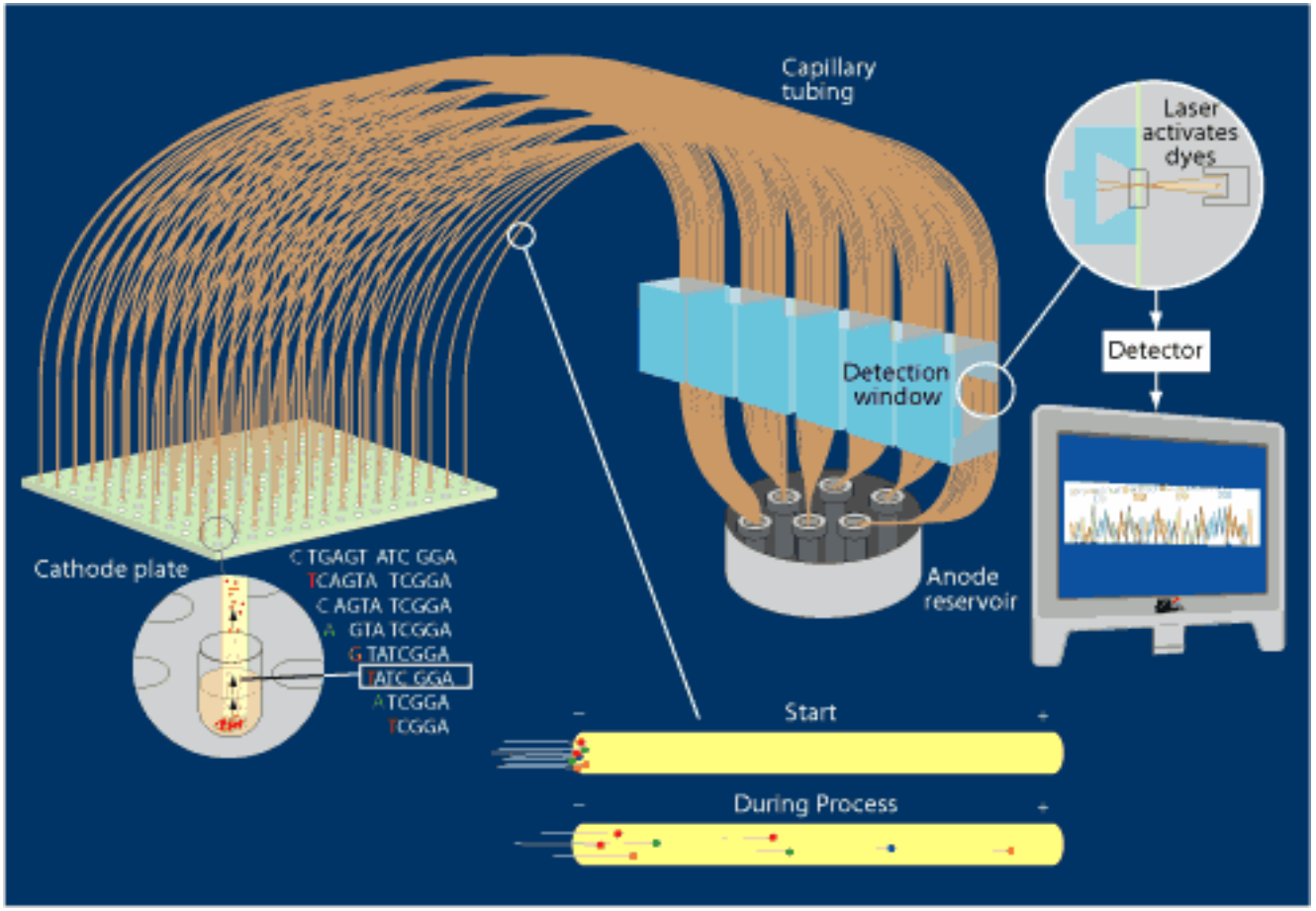
AND MORE… …

First Gen. Sequencing Platform

Next Gen. Sequencing Platform

3rd/4th Gen. Sequencing Platform

毛细管电泳DNA测序仪家族

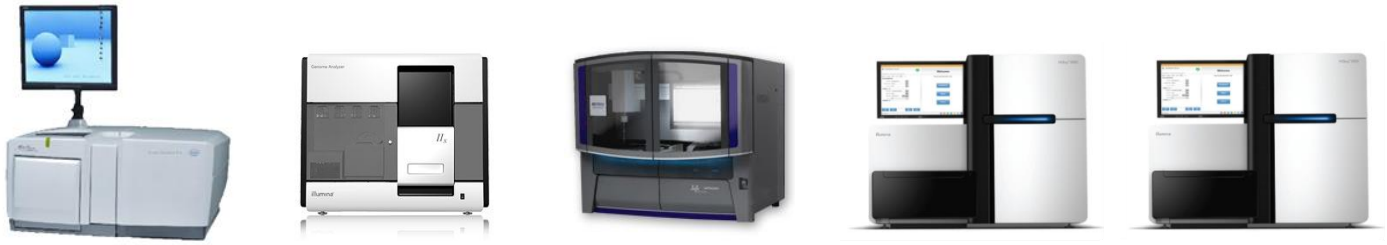


AB 3130*xL*

AB 3730*xL*

AB 3500*xL*

第二代高通量测序系统

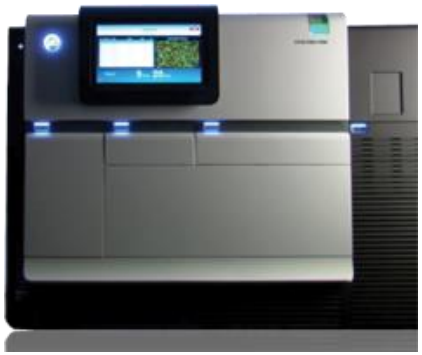


	Roche/454 GS FLX	Illumina GAIIx	ABI SOLiD 5500	Illumina HiSeq 2500	Illumina HiSeq X
片段的有效读长	700 bp	150-300 bp	75-120 bp	2x250 bp	2x150 bp
每次反应片段数	1 M	0.3 G	1.2 G	20 G	30 G
每反应检测通量	0.7 G	45-90 G	90-140 G	50-1000 G	1.6-1.8 T
每反应所需时间	1 Day	7-9 Day	9 Days	6 Days	<3 Days
检测反应准确性	97%	>99%	>99%	98%	98%
检测反应的成本 (/Gb)	$1500	$100	$300	<$50	<$10

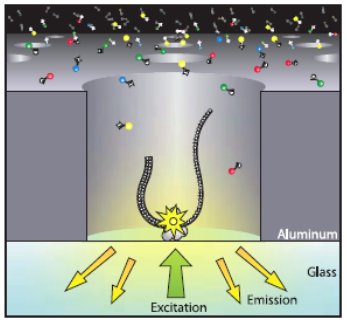
* 人类基因组大小是30亿碱基（3.0 Gb）

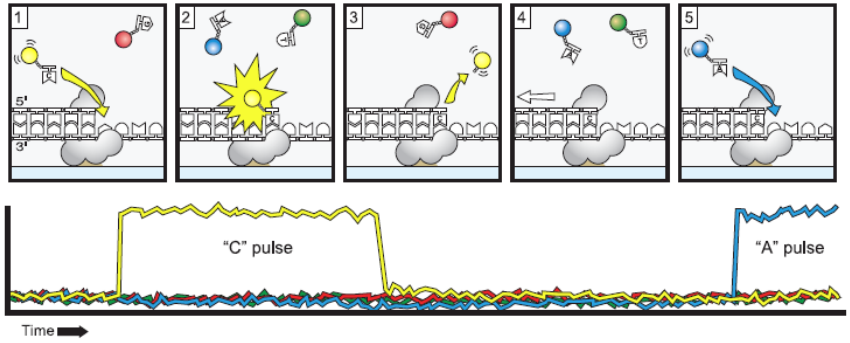
第三代高通量DNA测序仪：单分子、长片段

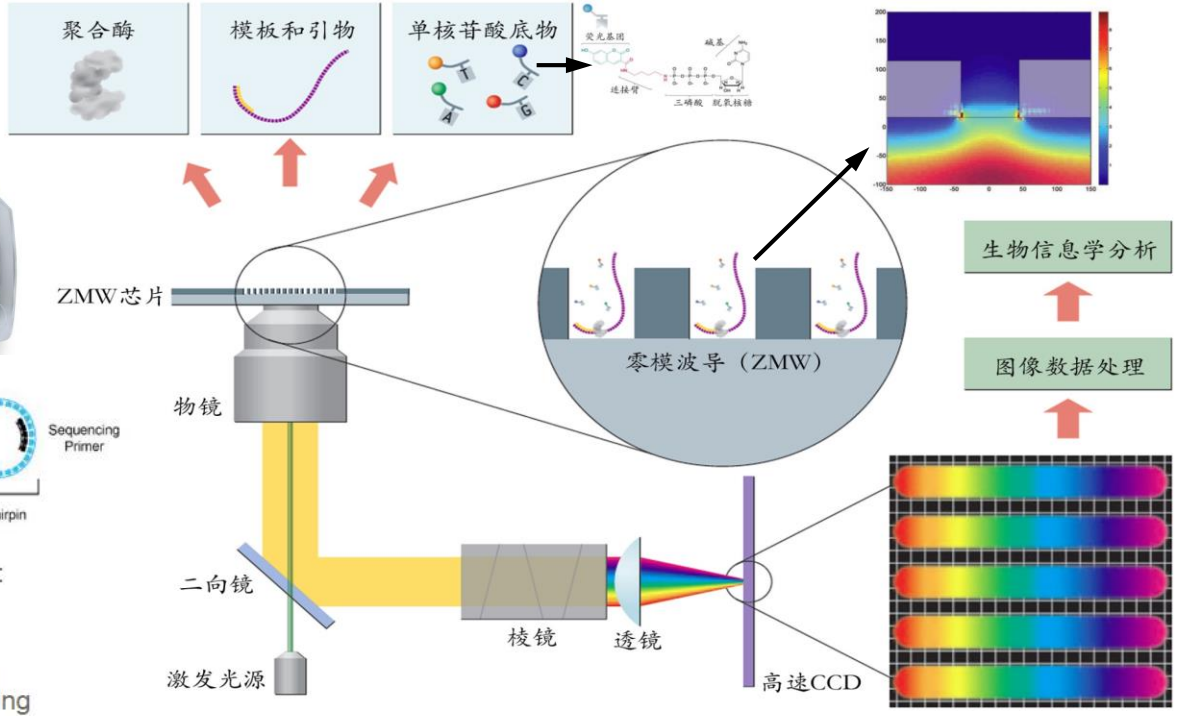
PACIFIC BIOSCIENCES

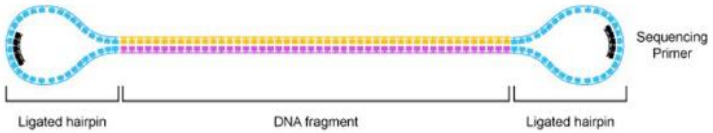


PacBio RS System







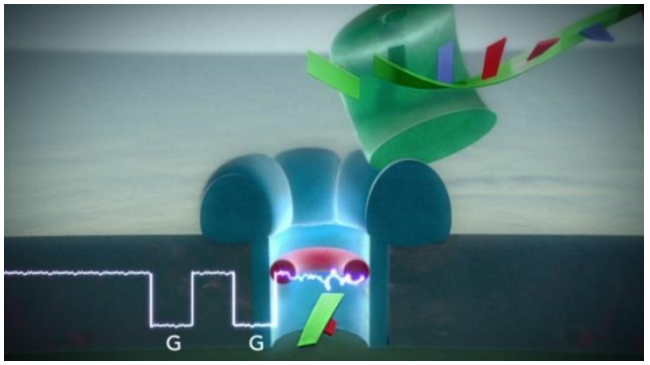


Key Advantages of SMRTbell™ template:

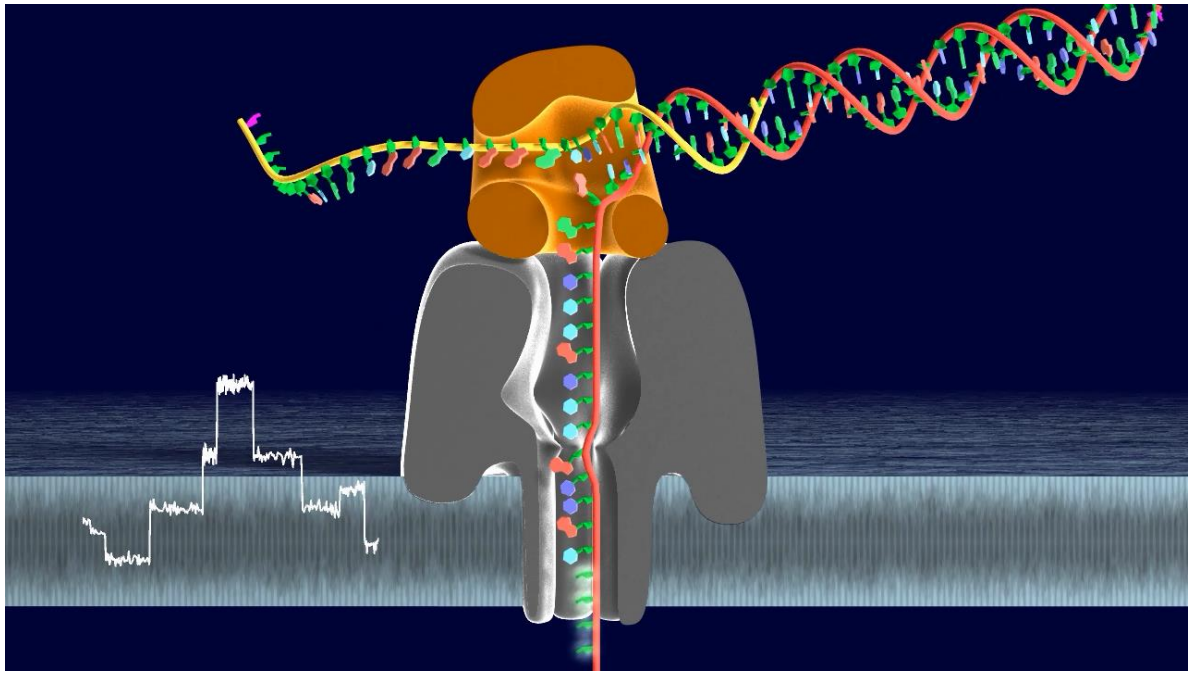
- Structurally linear
- Topologically circular
- Structural homogeneity of templates
- Forward and reverse strand sequencing

数小时产出数据的小型测序仪Oxford Nanopore

Oxford NANOPORE Technologies®



GridION & MinION



通过中国药监局认证的高通量测序平台

ThermoFisher SCIENTIFIC

ion torrent by life technologies

国家药品监督管理局 National Medical Products Administration

PGM

Ion Proton

华大基因 BGI

BGISEQ100

达安基因 DA AN GENE

DA8600

博奥生物 CapitalBio Corporation

BioelectronSeq 4000

illumina®

MiSeq™Dx

NextSeq500

BerryGenomics 贝瑞和康

NexSeq CN500

BGISEQ/MGISEQ系列基因测序仪

MGI 华大智造

国家药品监督管理局 National Medical Products Administration

国械注准20163402206

BGISEQ-500

最大通量
520 Gb
有效reads数
1300 Million
最大读长*
PE100

国械注准20173401605

BGISEQ-50

最大通量
8 Gb
有效reads数
160 Million
最大读长
SE50

MGISEQ-200

最大通量
60Gb
有效reads数
300 Million
最大读长
PE100

国械注准20183400258

国械注准20183400257

MGISEQ-2000

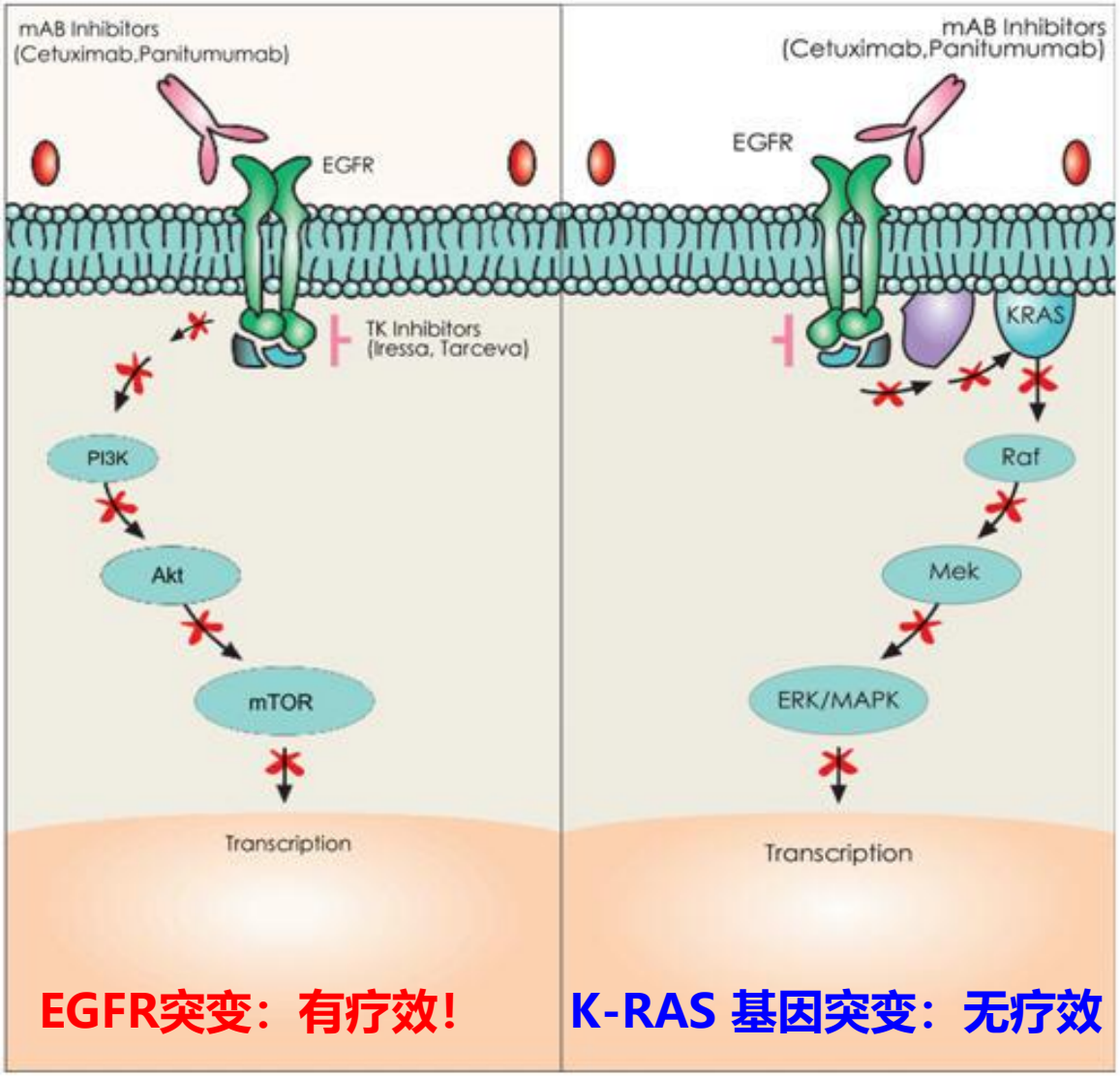
最大通量
1080Gb
有效reads数
1500 Million
最大读长*
PE150

基因组学研究与肿瘤精准医疗实践

- ✓ 已在多种肿瘤的临床实践中实现指导肿瘤诊断、治疗与预后评估，揭示耐药机制等
- ✓ 超过10种肿瘤类型的患者从中获益

Hyman, *et al.* ***Cell*** 168: 584-599 **(2017)**

利用基因检测技术指导肿瘤靶向用药



EGFR/K-RAS信号转导用药

非小细胞肺癌靶向药物**特罗凯**(Tarceva)和**易瑞沙**(Iressa)靶向阻断经由EGFR传导的癌基因信号通路。只有患者EGFR发生特定突变，且K-RAS没有突变时才能有很好的效果

携带K-RAS永久激活性突变的患者，建议使用**安卓健**(Antroquinonol)等靶向KRAS抑制剂药物治疗

泰瑞沙(Tagrisso, AZD-9291)对L858R/T790M、外显子19缺失型和野生型EGFR的IC50分别为11.44、12.92和493.8 nM；**血脑屏障通过率较好**

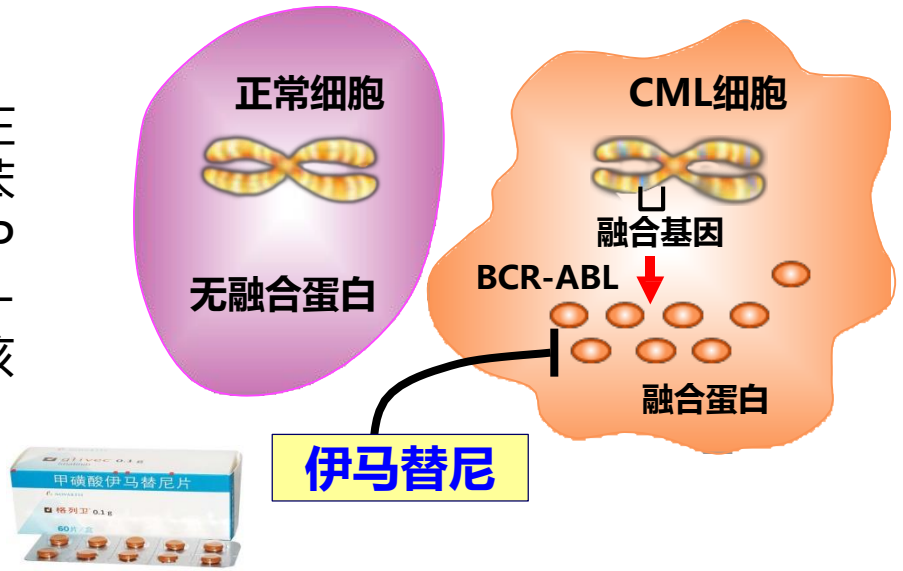
基因靶向药物有效治疗恶性肿瘤

特异性靶点

BCR-ABL激酶在正常细胞中不表达，但在CML时，会促CML成熟粒细胞增生。2-苯氨嘧啶衍生物**伊马替尼**可以选择性阻断ATP与ABL激酶结合位点，有效地抑制BCR-ABL激酶底物中酪氨酸残基的磷酸化，使该酶失活。**研发历时42年**

1960年，发现费城染色体(Bcr-Abl)阳性CML

2001年，美国FDA批准瑞士Novartis伊马替尼

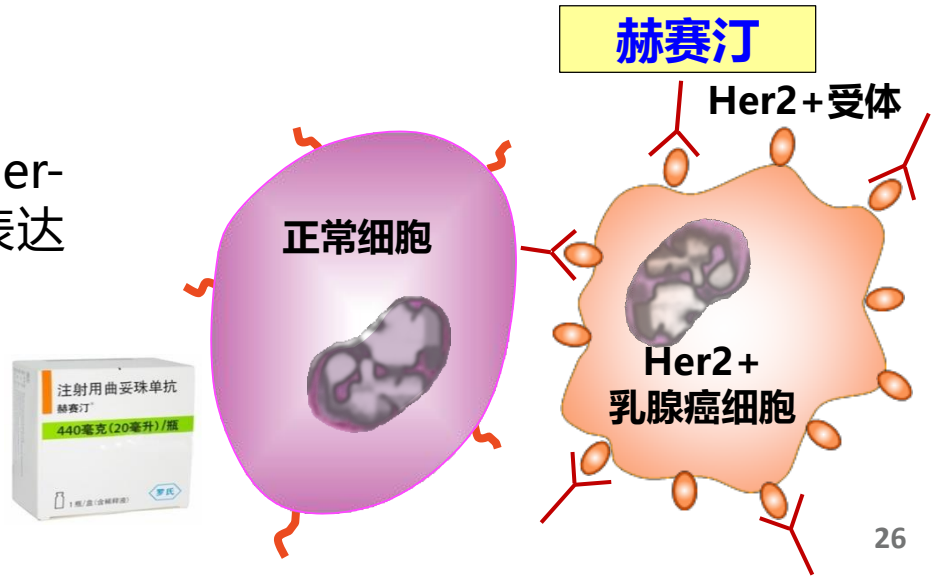


过表达靶点

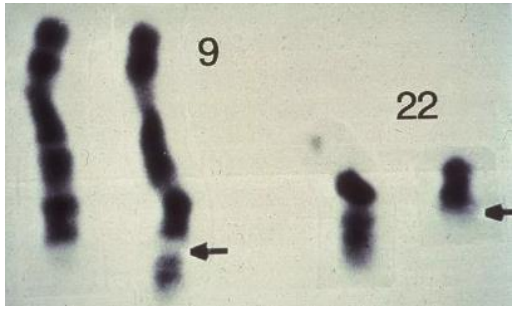
赫赛汀是人源化Her-2单克隆抗体，对Her-2有高亲合力，用于治疗Her-2受体过表达的乳腺癌。**研发历时17年**

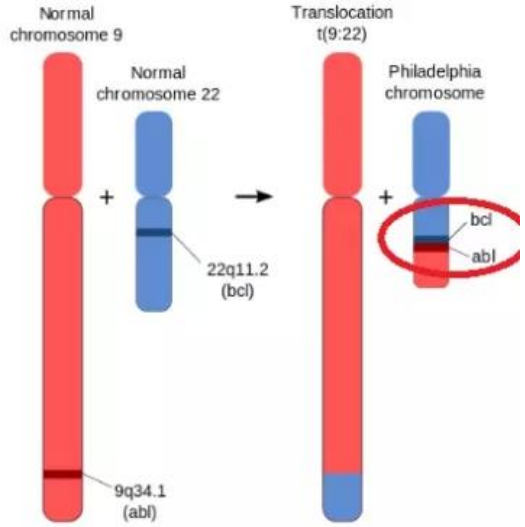
1982年，Robert A. Weinberg等发现HER-2

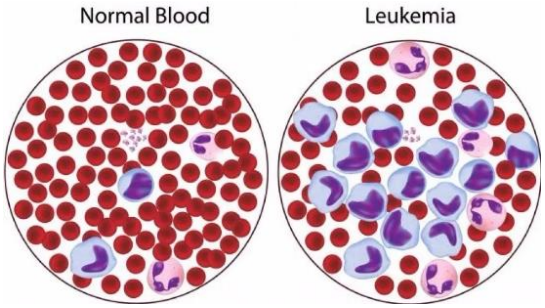
1998年，美国FDA批准赫赛汀上市

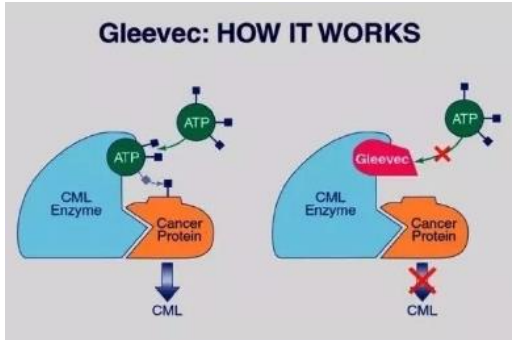


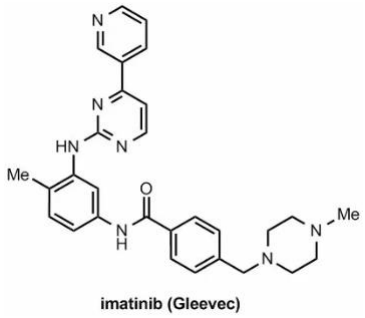
肿瘤靶向治疗药物的典范：Gleevec®







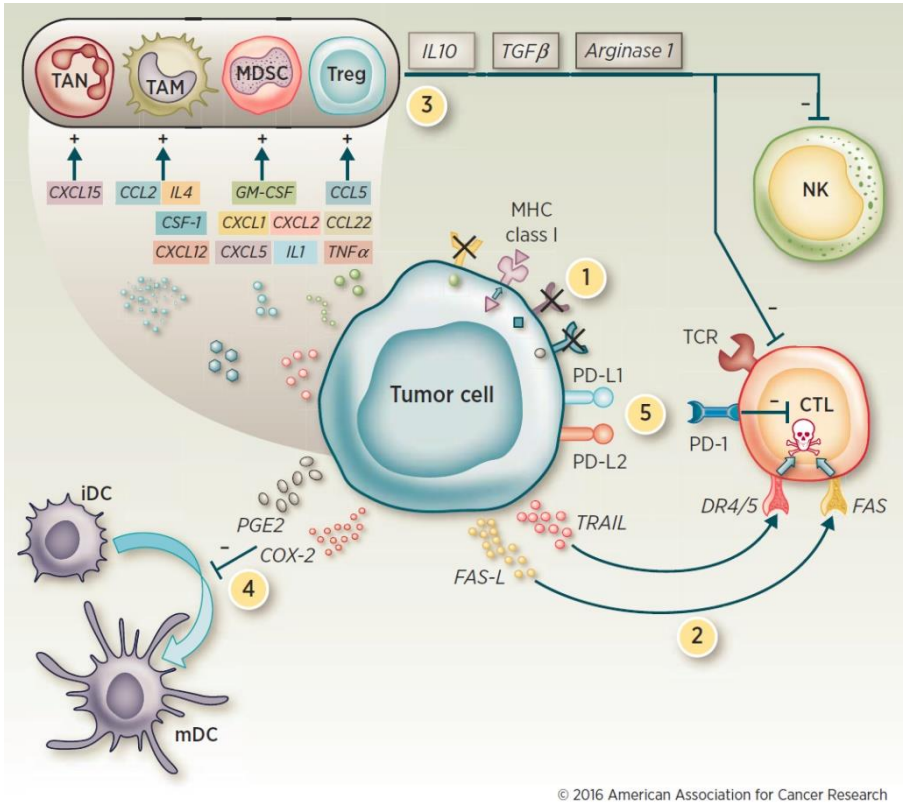




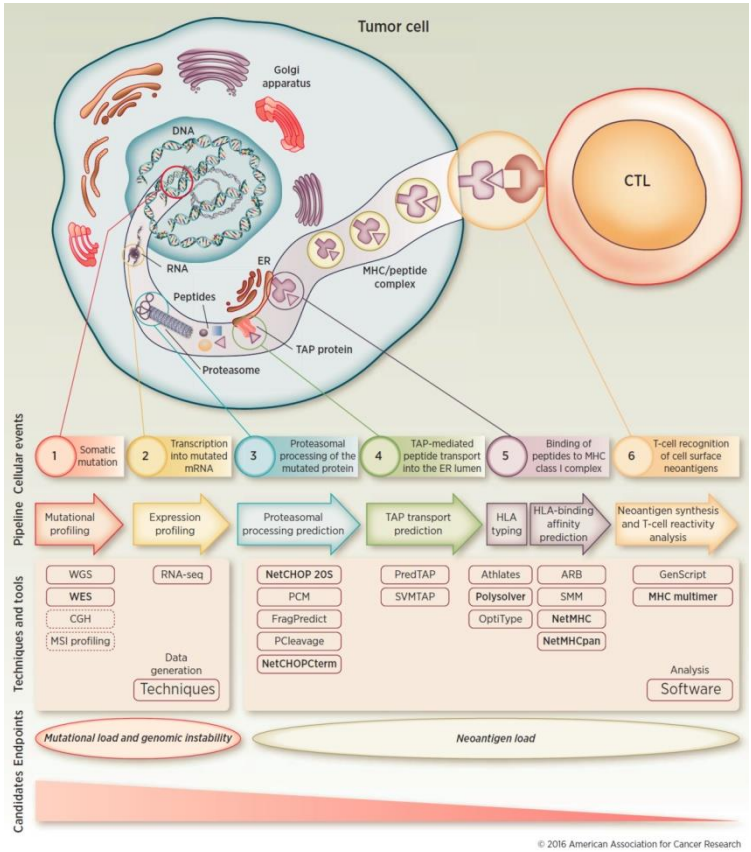
“正牌药”￥23500/月

“仿制药”￥ 400/月

免疫靶向治疗的疗效监测和新基因靶点发现



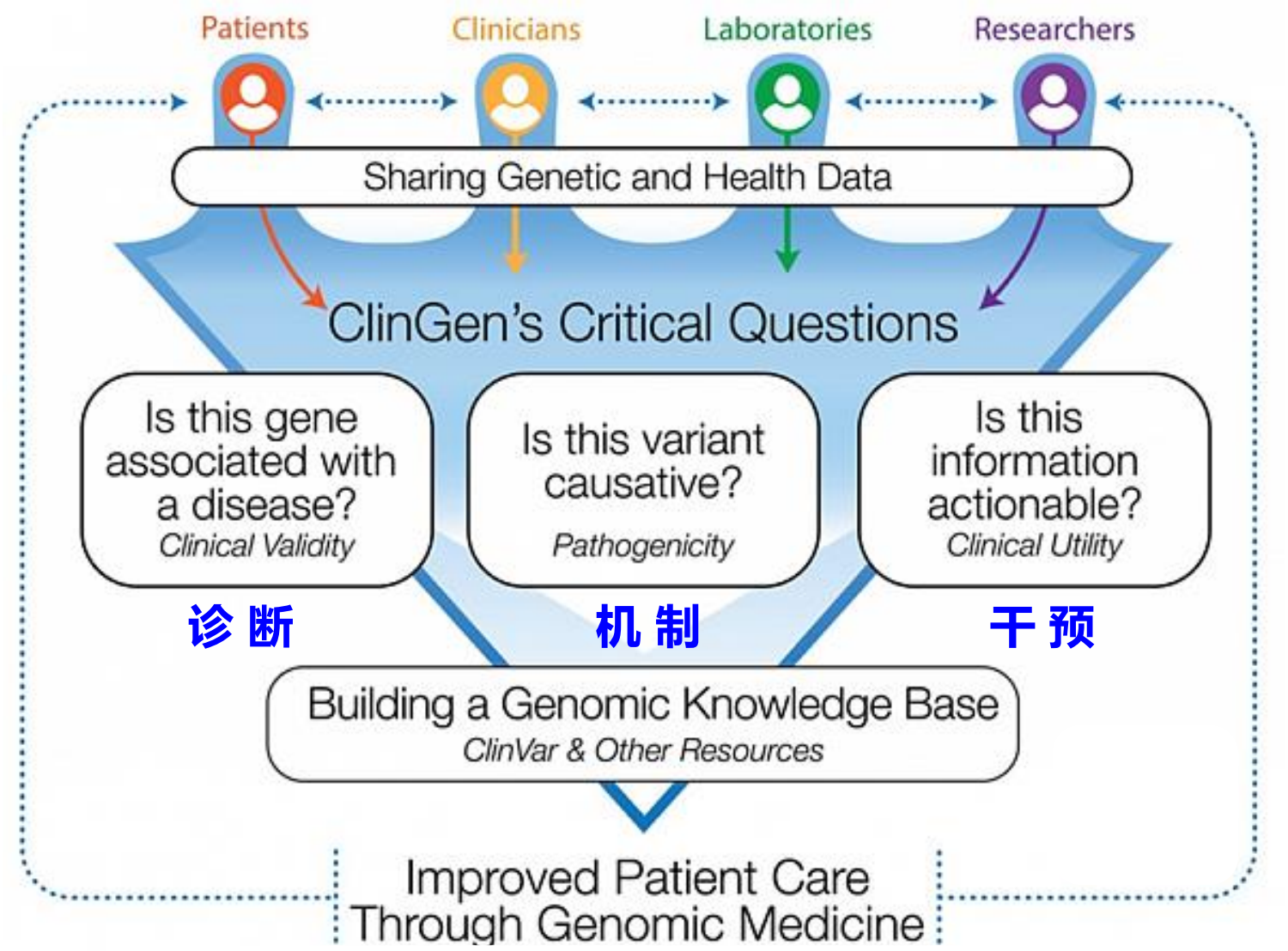
Mechanisms of Immune Escape in Tumor Micro-environment



Immune-Relevant Neoantigens Identification

Chabanon RM, *et al*. ***Clin Cancer Res***. 22(17): 4309-4321 **(2016)**

基因组科学助力精准医疗实践



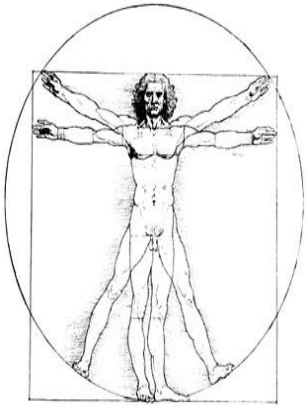
早期快速诊断、精准有效治疗、全生命周期健康管理

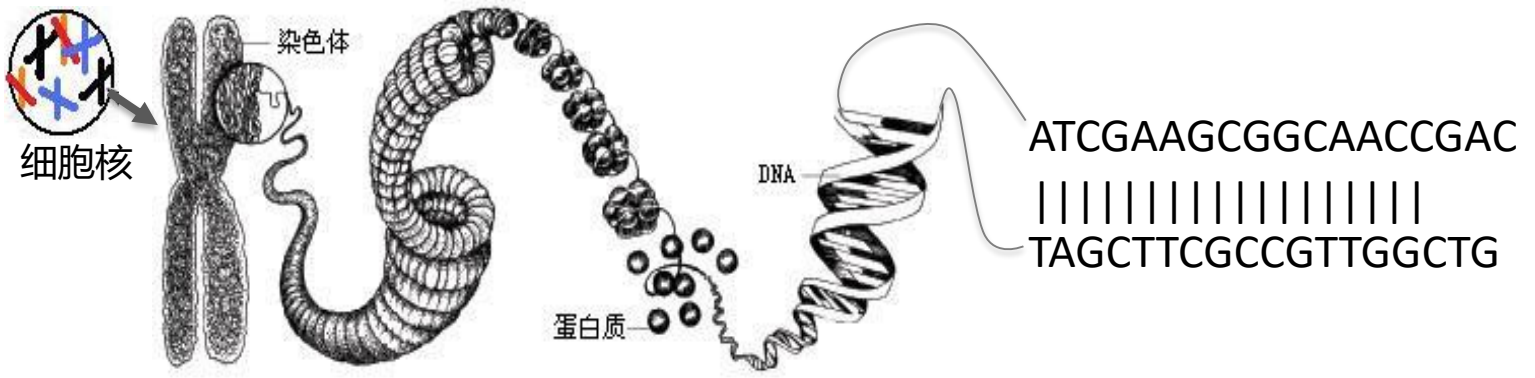
精准医学发展趋势和前景展望

- ✓ 基因检测是关键技术
- ✓ **数据解析是核心保障**

生物医学进入大数据时代（BIG Data）

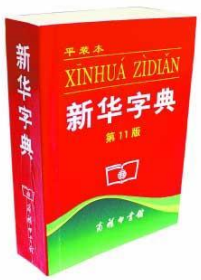
“人类基因组计划”于1990年启动，草图2000年完成，完成图2003年实现





一个人的基因组长度3GB：约30亿个字母（A/T/C/G），或$3x10^9$(字节)

这些序列编成新华字典，每本72万字，约**2144本**



大数据特点：

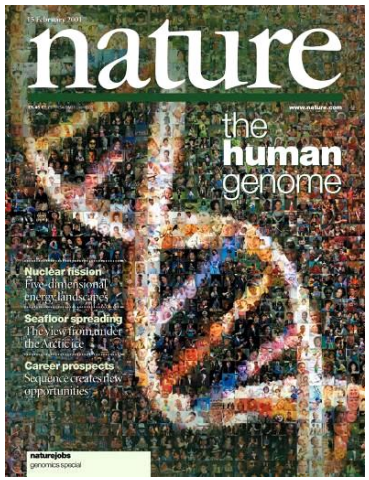
- 非结构性
- TB级以上
- 5V (Volume, Variety,Velocity, Veracity, Value)

人类基因组计划
10年, 30亿美元

现在2天30人
每人100美元

测序技术快速发展 → 通量提高、成本降低 → 生命与健康大数据

国际大型生命组学研究计划



人类基因组计划（HGP，1990-2000）

人类基因组单体图计划（HapMap，2002-2009）

DNA元件百科全书计划（ENCODE，2003-2012）

全基因组关联分析（GWAS，2005-）

癌症基因组计划（TCGA，2005-2015）

国际千人基因组计划（1000 Genomes，2008-2012）

人类表观基因组计划（HEP，2010-2015）

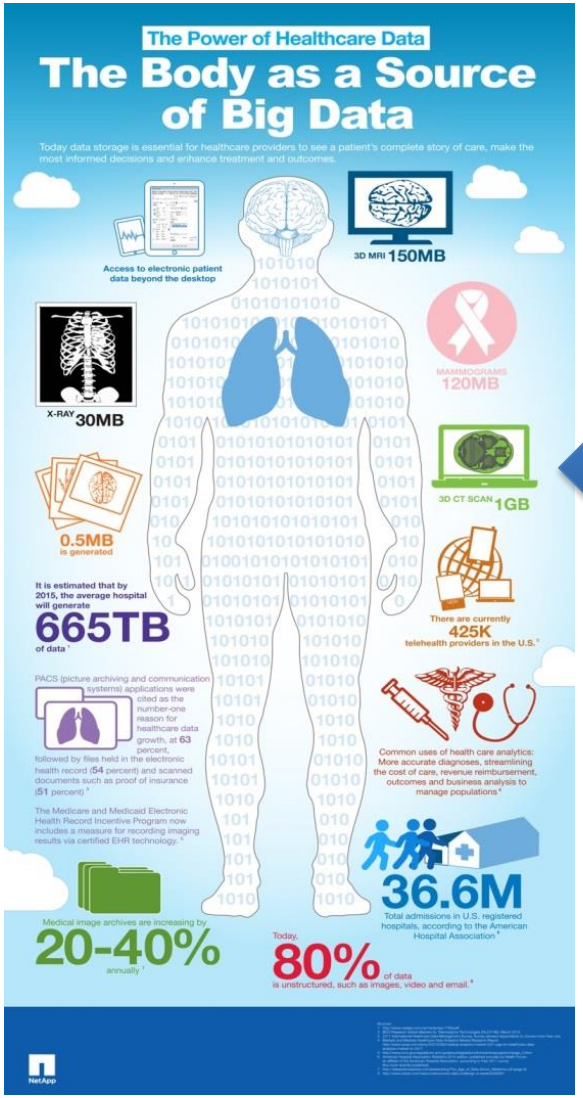
微生物组计划（National Microbiome Initiative，2016-）

人类细胞图谱计划（Human Cell Atlas，2016-）

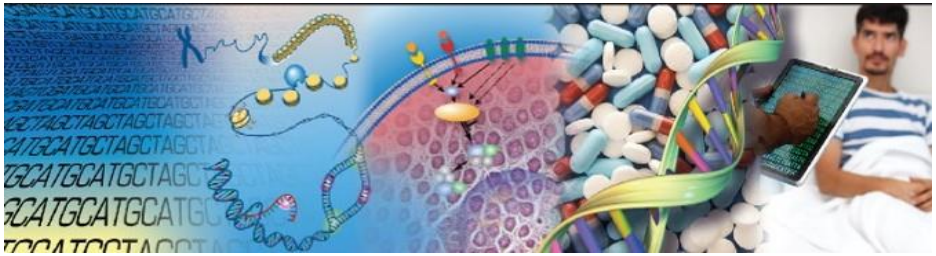
HUMAN CELL ATLAS

1990 2002 2003 2005 2008 2010 2016

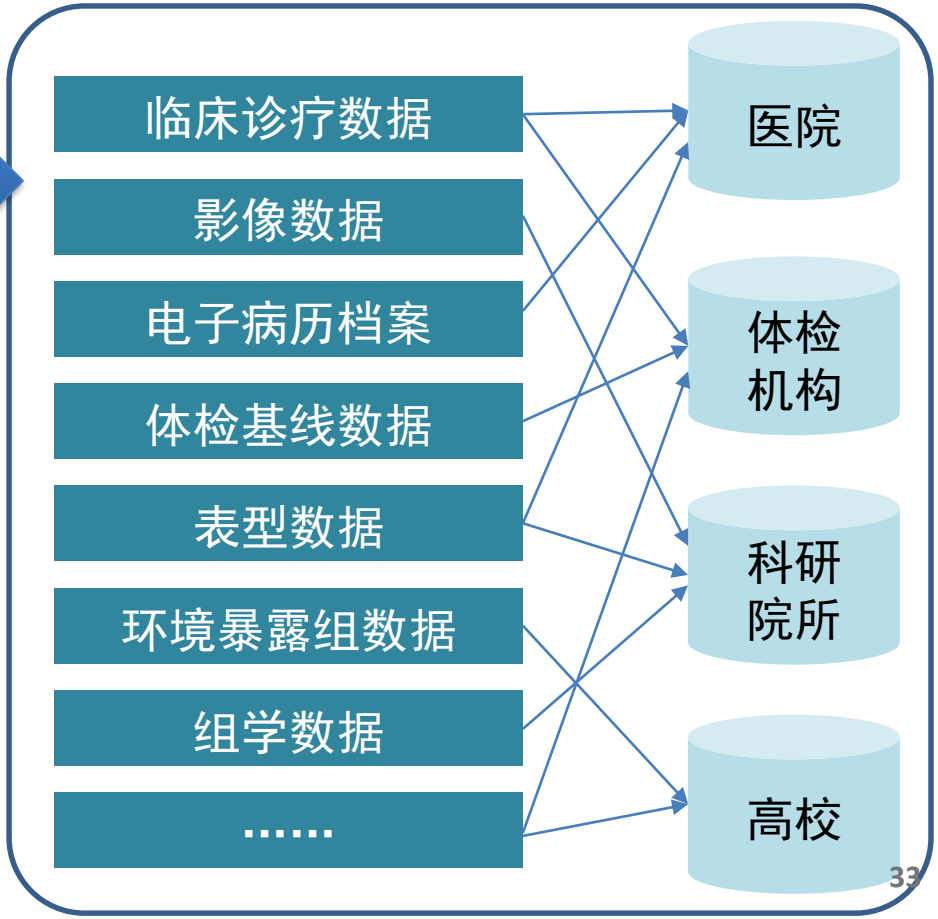
精准医学需要集成生命组学和医疗数据



来源：netapp.com

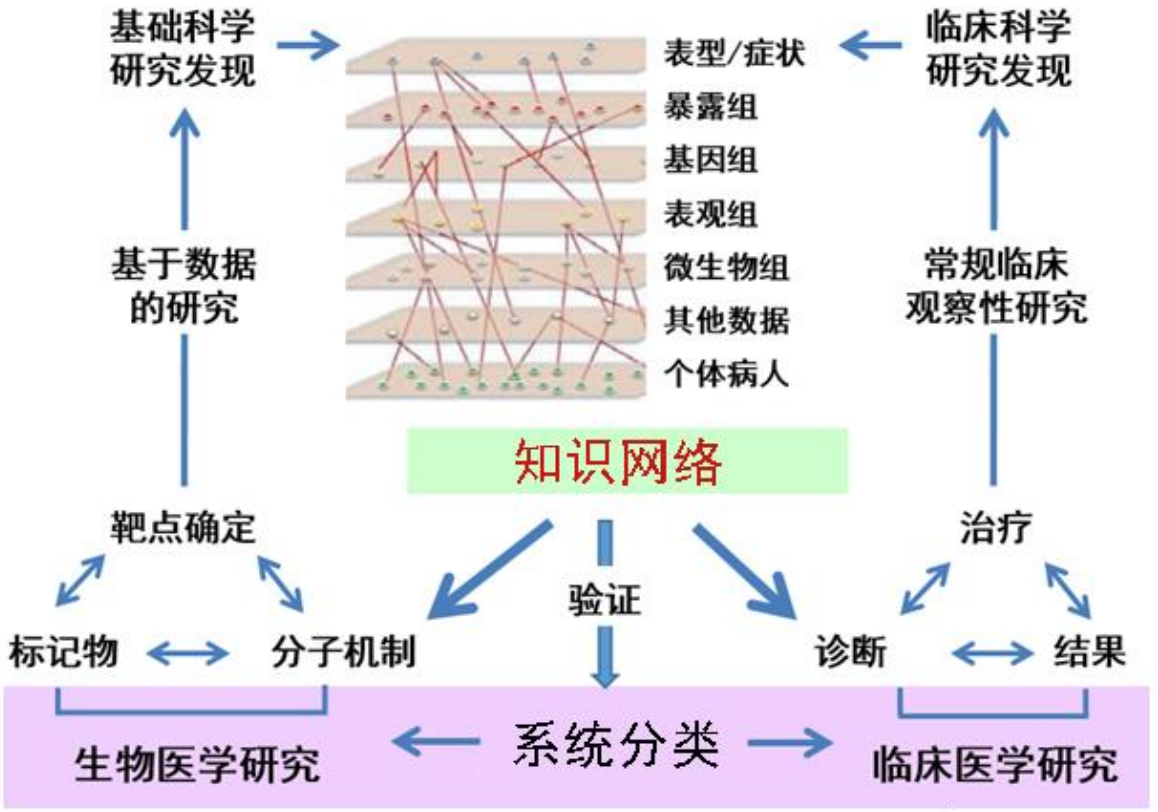


约10TB/人



大数据应用是精准医疗和健康管理的严峻挑战

- **数据共享开放程度较低**
  - ✓ 数据质量不高
  - ✓ **数据共享与开放不畅**
- **数据应用水平不高**
- **支撑保障体系尚不完善**
  - ✓ 法律法规不健全
  - ✓ 标准规范不完善
  - ✓ 安全与隐私保护薄弱
  - ✓ **复合型人才缺乏**



精准健康需要集成生命组学和医疗数据

精准医学与健康管理的关键大数据技术

生命组学数据规范化

GEO Gene Expression Omnibus　ArrayExpress　PDB PROTEIN DATA BANK

NCBI　EMBL-EBI　prosite

THE HUMAN PROTEIN ATLAS

MIAME　MIARE

MIMS　MIAPE

建立了大量的多组学数据集

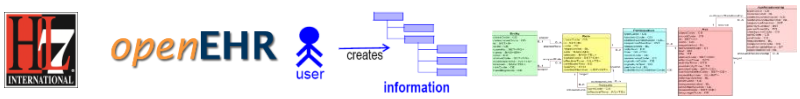
缺乏面向精准医学数据标准

临床医疗信息规范化

SIEMENS　mirth　TERADATA

Corepoint HEALTH　INTERSYSTEMS　Apache Camel

HL7 INTERNATIONAL　openEHR



医疗信息化发展积累了海量数据

数据集成度低，规范化程度不高

数据匹配与关联

索引与术语匹配

ORION HEALTH　OpenEMPI

ATLAS MEDICAL　HealthUnity MPI

MetaMap　UMLS

eleMap　LOINC

主题词提取和关联

cTAKES　MedLee　HITEx

MedEx　SecTag　NegEx

eTRIKS　iCOD　iDASH

全基因组关联分析

GeneCards HUMAN GENE DATABASE　GWAS CENTRAL

Genetic Association Database

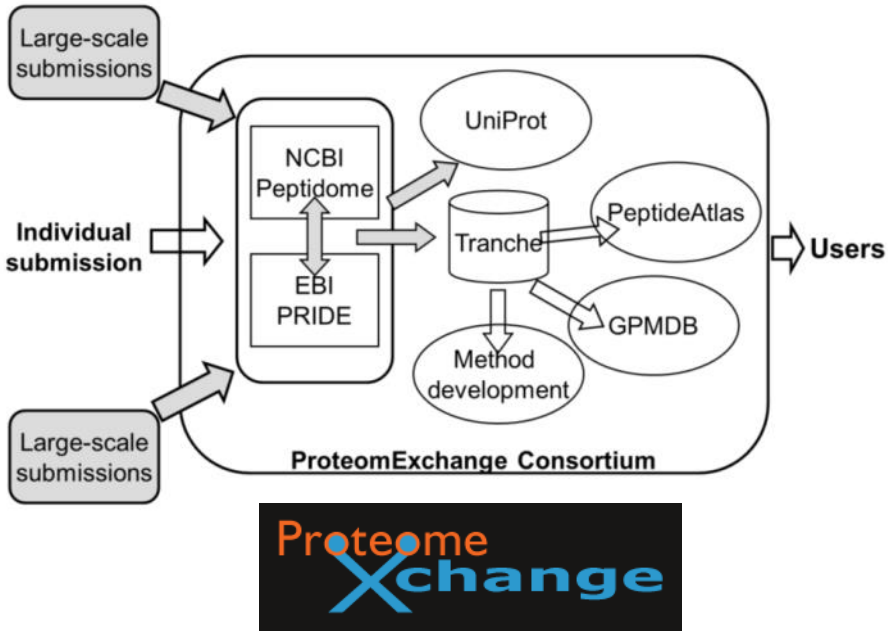
GWA db　DisGeNET

主索引与术语匹配、主题词提取与关联、GWAS等关联技术已经发展多年

缺乏完整的跨库关联系统及融合生命组学数据和临床医疗信息的技术体系

国际组学大数据中心的现状与趋势

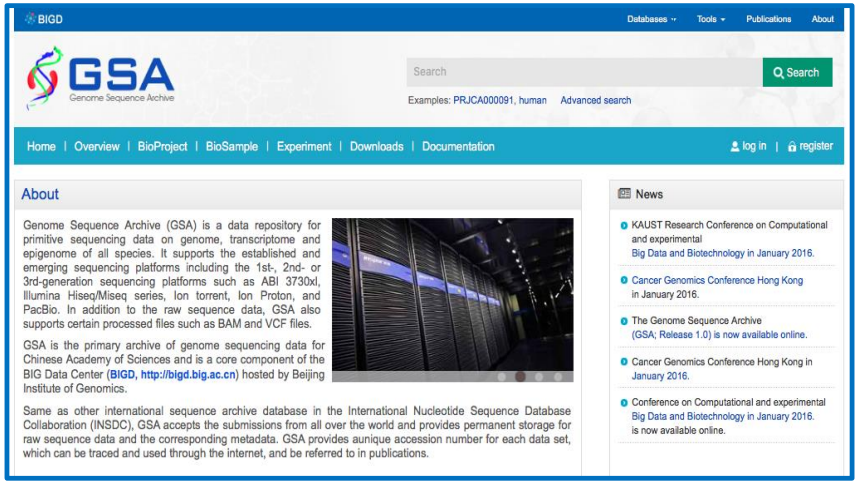
- 基因组
  - DDBJ/EMBL/GenBank
  - Ensembl/UCSC, TCGA
- 转录组
  - GEO/ArrayExpress
- 蛋白质组
  - PRIDE/PeptideAtlas
- EBI与NCBI
- 以物理集中管理模式为主
  - 避免网络瓶颈
  - 整合需要
  - 便于管理



形成了事实上的数据资源垄断
中国数据主权受到严重威胁！

建立并更新组学原始数据存储归档系统（GSA）

组学原始数据归档库（Genome Sequence Archive），采用国际兼容规范，是组学原始数据汇交、存储、管理与共享系统，是国内首个被国际期刊认可的组学数据发布平台



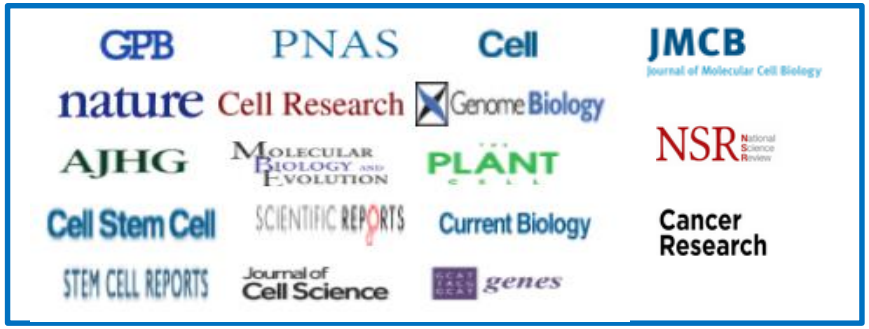
打破了组学数据国际垄断，保护了数据主权

GSA: Genome Sequence Archive

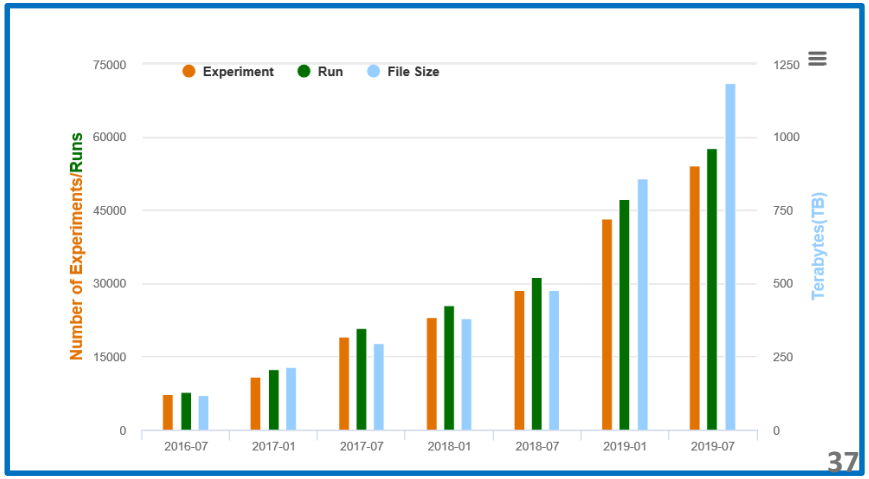
对应

美国NCBI中的SRA库
欧洲EBI的ENA/SRA库
日本DDBJ的DRA库

提交到GSA系统的数据，其发表的论文已经被**101**种国际知名期刊所收录，包括Cell、Nature、PNAS、AJHG、Cell Research等



已经接收了来 自 **149** 个研究机构 **511** 余位科研人员的共 超过 **1PB** 数据，国家重点研发计划项目实施以来，接受数据量持续上升



国家科学数据中心建设

中华人民共和国科学技术部
Ministry of Science and Technology of the People's Republic of China

请输入关键字 搜索

首页 组织机构 信息公开 科技政策 科技计划 政务服务 党建工作 公众参与 专题专栏

信息名称:	科技部 财政部关于发布国家科技资源共享服务平台优化调整名单的通知		
索 引 号:	306-07-2019-031	**信息类别:**	规范性文件2019
发布机构:	科技部;财政部	**发文日期:**	2019年06月05日
文　　号:	国科发基（2019）194号	**效　　力:**	

科技部 财政部关于发布国家科技资源共享服务平台优化调整名单的通知

国科发基〔2019〕194号

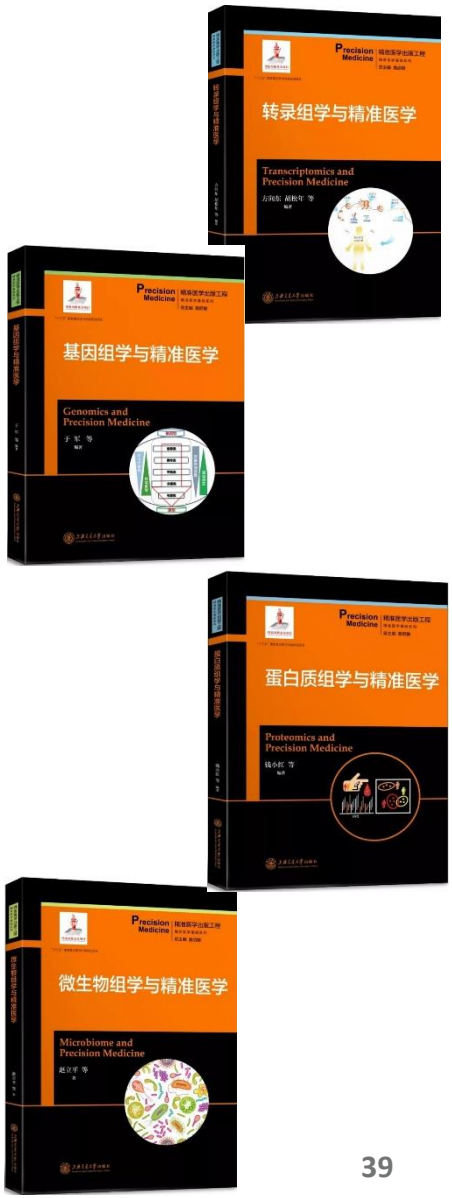
教育部、自然资源部、农业农村部、卫生健康委、市场监管总局、林草局、中科院、地震局、气象局、药监局科技、财务主管部门，广东省科技厅、财政厅：

为落实《科学数据管理办法》和《国家科技资源共享服务平台管理办法》的要求，规范管理国家科技资源共享服务平台（简称国家平台），完善科技资源共享服务体系，推动科技资源向社会开放共享，科技部、财政部对原有国家平台开展了优化调整工作，通过部门推荐和专家咨询，经研究共形成“国家高能物理科学数据中心”等20个国家科学数据中心、“国家重要野生植物种质资源库”等30个国家生物种质与实验材料资源库。

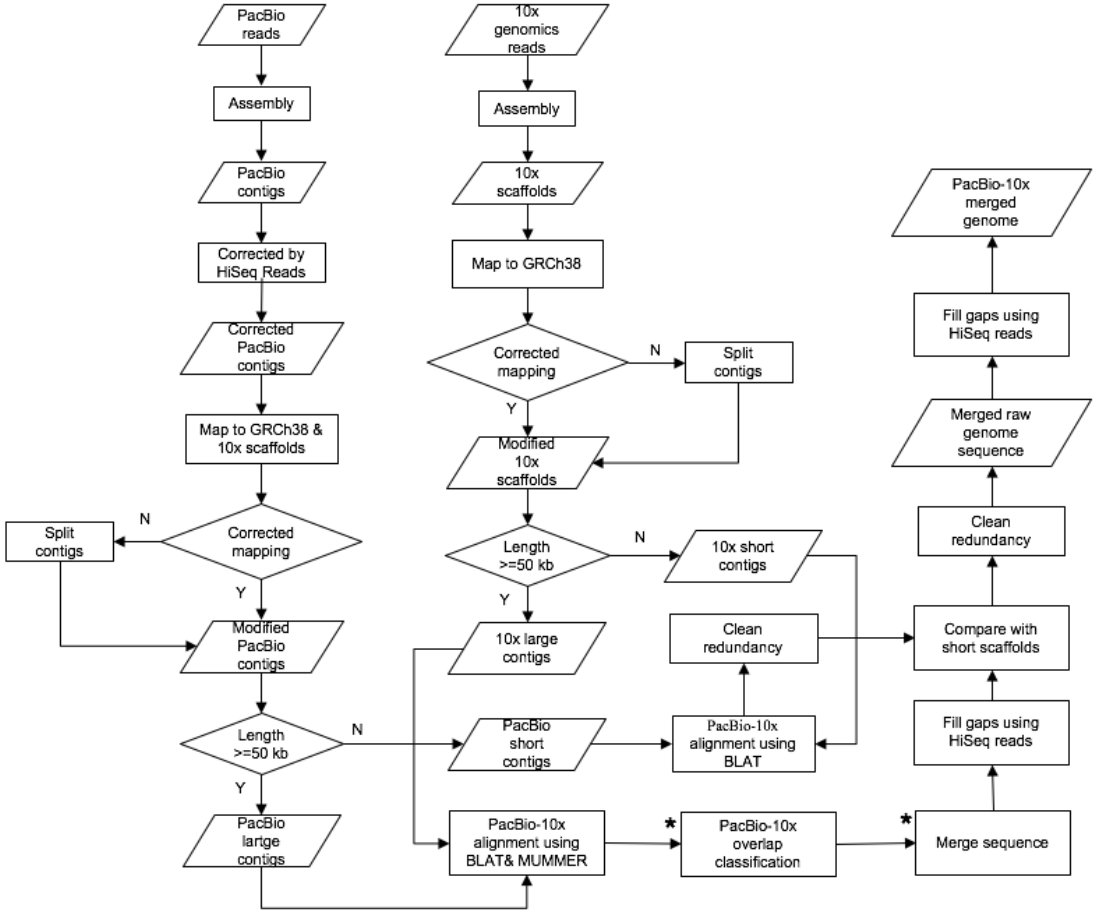
2	国家基因组科学数据中心	中国科学院北京基因组研究所	中科院
3	国家微生物科学数据中心	中国科学院微生物研究所	中科院

生命组学数据质量标准化体系建设

样本表型数据采集标准	• 实验设计与样本筛选标准，数据采集方法标准
组学测序质控标准	• 样本处理、DNA提取准备、文库构建等湿实验的测序质量控制标准
组学数据分析流程规范标准	• 数据预处理、遗传变异分析、基因型推断、遗传图谱构建等分析流程标准规范
数据库与知识库信息模型标准	• 多组学数据的查询、访问、检索、存储等标准
组学数据管理与共享策略与标准	• 组学数据汇交集成、异地备份、隐私保护、安全加密、访问控制等标准



建立中国人的参比基因组



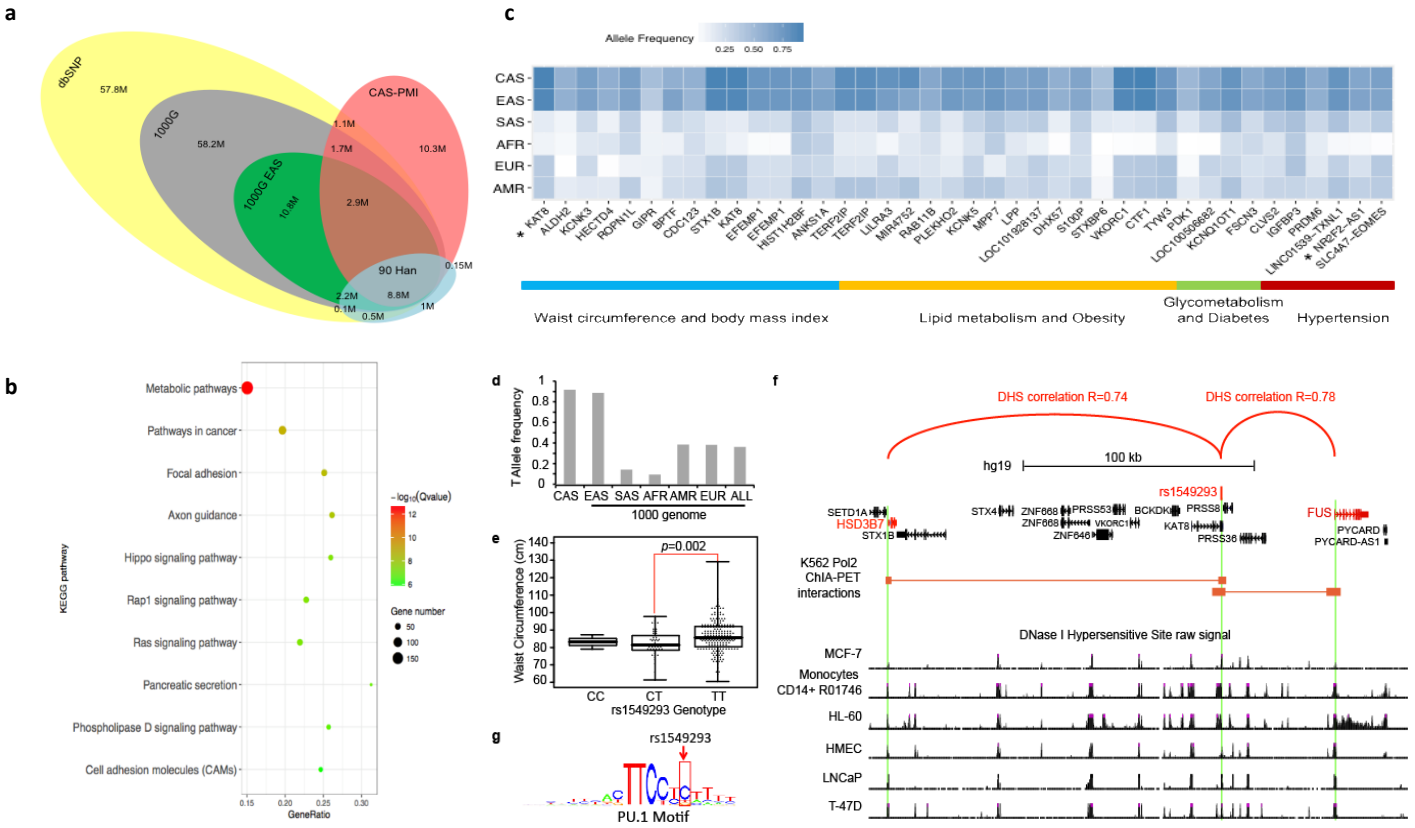
PacBio	**50X**
10X Genomics	**~60X**
Illumina Pair-end	**~60X**
Bionano	**~100X**

	Population	Method	Scaffold N50(MB)
YH2.0	Southern Chinese	Hiseq (fosmid)	20.52
HX1	Southern Chinese	PacBio + Bionano	21.98
NH1.0	**Northern Chinese**	**PacBio + 10X Genomics + Bionano**	**46.63**
GRCh38	European	Sanger (BAC + fosmid)	67.79

在NH1.0中填补了GRCh38上99个gaps，总长 609,822 bp

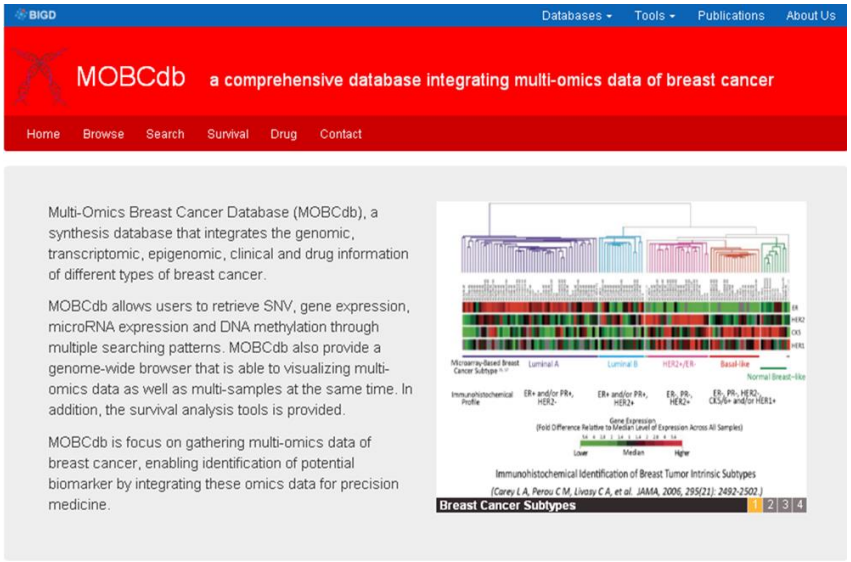
最长的gap为188,143bp

构建中国人群基因组变异图谱

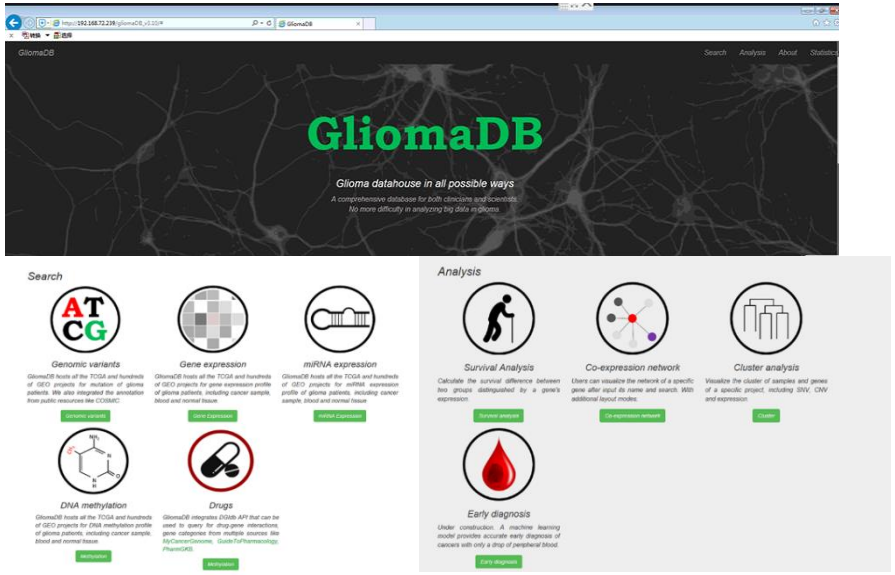


- 建立了600个中国人的基因变异参照集
- 对中国人群高频发生的变异进行功能注释，经算法预测突变作用的靶位点，与表型关联等分析，发现与腰围相关的变异位点所行使的潜在的调控功能

精准医学知识库的构建



http://bigd.big.ac.cn/MOBCdb/



http://192.168.76.217:8080/gliomaDB/

乳腺癌综合在线数据库

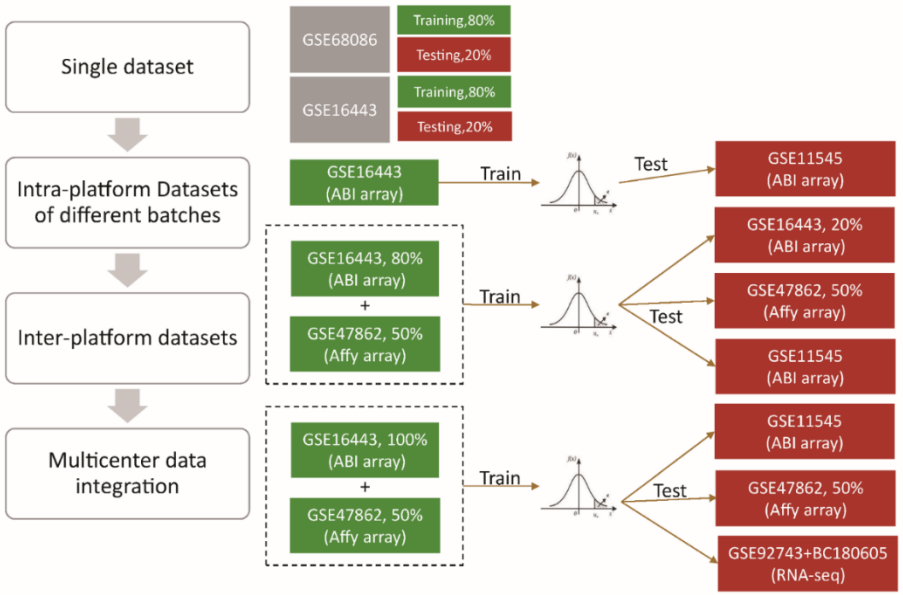
- ✓ 已收集乳腺癌组学和临床数据共 10TB
  - ✓ 临床信息1000例
  - ✓ 基因组、转录组、甲基化和 miRNA 数据
  - ✓ 肿瘤靶向用药信息
- ✓ 多组学数据的搜索和基因组浏览器内展示
  - ✓ 查询、浏览
  - ✓ 生存分析
  - ✓ 药物指导

胶质瘤综合在线数据库

- ✓ 已整合数据
  - ✓ 胶质瘤表达量信息 32411725、基因60191、病人样本1405、项目13、样本17570
- ✓ 实现功能
  - ✓ 查询、分析、药物指导
  - ✓ 分析功能包括生存分析、共表达网络、聚类分析、早期诊断

Fang Group. *Breast Cancer Res Treat* (2018)
Fang Group. *Genomics Proteomics Bioinformatics* (2018)

基于外周血表达谱的肿瘤检测 (rankDetect)



将绝对的基因表达量转换为两两基因表达量之间的大小关系，进一步通过特征选择、机器学习方法构建检测模型

$$G_{ij}=\begin{cases}1, & G_i > G_j \\ 0, & G_i \le G_j\end{cases} \quad (1)$$

数据转换

$$\underset{\beta}{\operatorname{argmin}} \frac{1}{2n_{samples}}\left\|X_\beta - y\right\|_2^2 + \lambda\alpha\left\|\beta\right\|_1 + \frac{\lambda(1-\alpha)}{2}\left\|\beta\right\|_2^2 \quad (3)$$

特征选择

$$\frac{1}{N}\sum_{i=1}^{N} f(w^T b_i X^i + v b_i) + \lambda\sum_{k=1}^{p}\left|\beta_k\right| \quad (4)$$

模型构建

随机森林.....

理论基础

肿瘤是系统性疾病

血液成分参与维持机体稳态

血液成分影响免疫和炎症反应

血液成分参与应激信号传导

血液基因表达反映机体状态

rankDetect算法创新

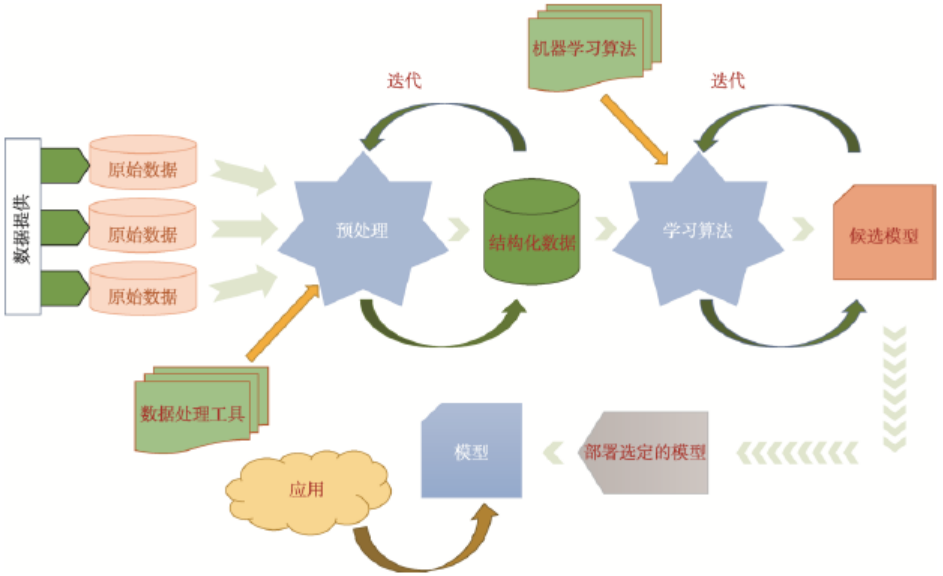
基于rank的标准化方式

去除了批次效应和平台差异

对种族差异不敏感

可同时应用于机器学习的训练和新数据的预测过程

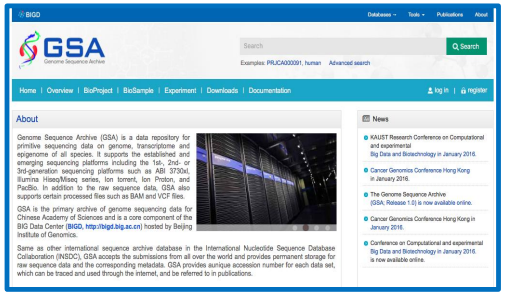
多组学数据辅助临床决策支持



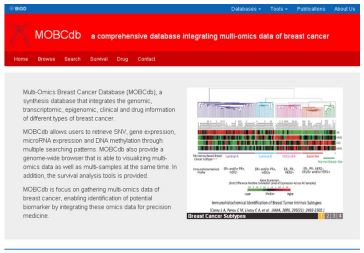
人工智能在临床决策领域迅速发展

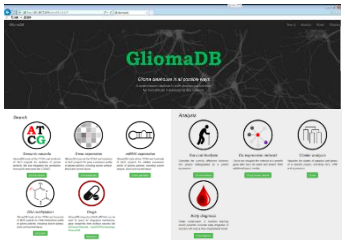
通过学习输入数据的数据结构和其内在模式，选择对应的学习方式和训练方法以构建最优的数学模型，开不断调整模型参数，通过数学方法求解模型最优化后的反馈结果，以提高泛化能力防止发生过拟合，以达到对疾病预测和分类、用药指导、疾病诊断等目的，为临床决策支持提供技术基础

人工智能与精准医学大数据应用



组学数据存储归档
精准医学知识库



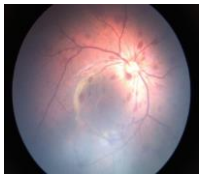


影像

组学

医疗

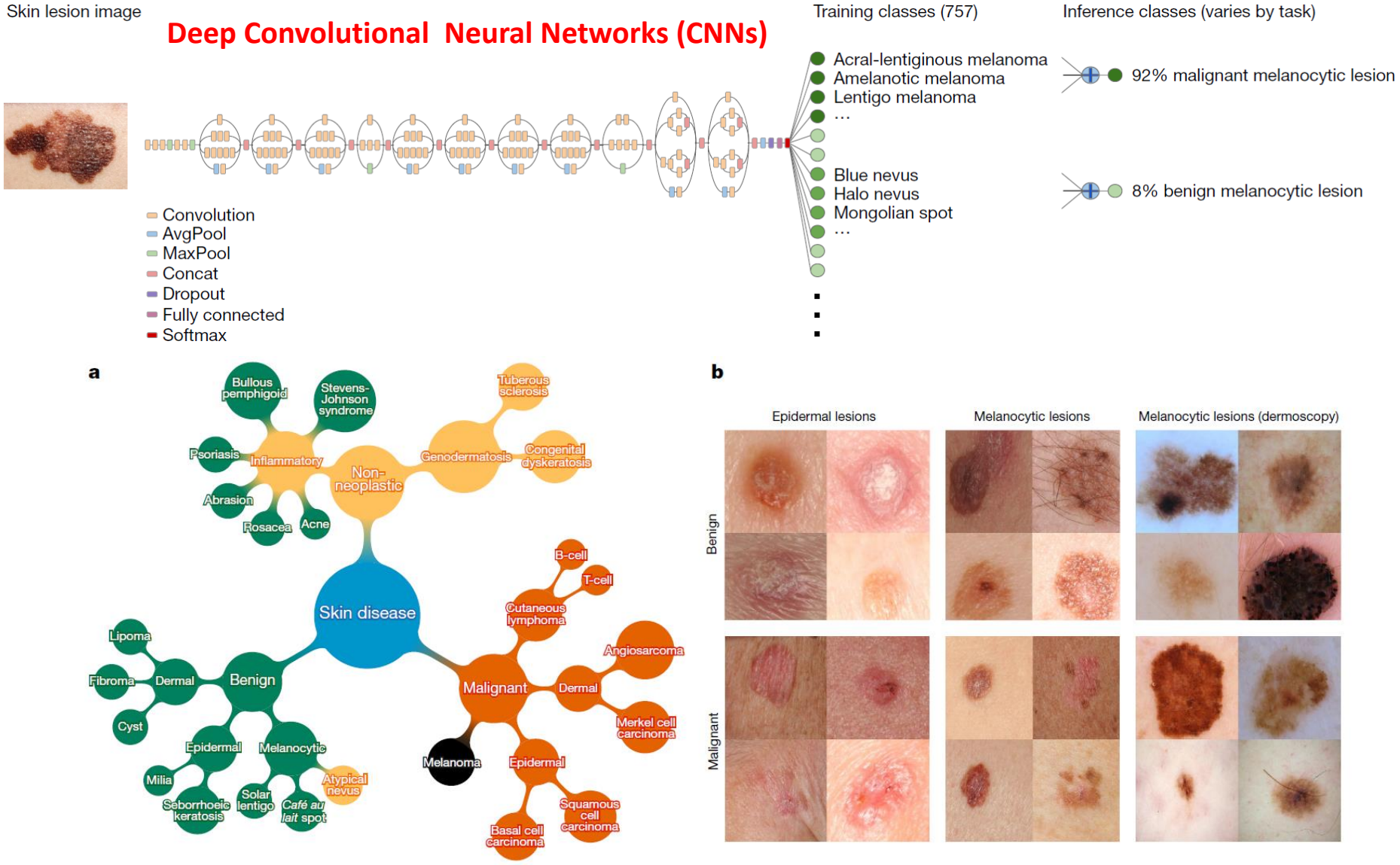
像素层面标注器官
检测分割病变组织



数据文本机构化
电子病历文本挖掘

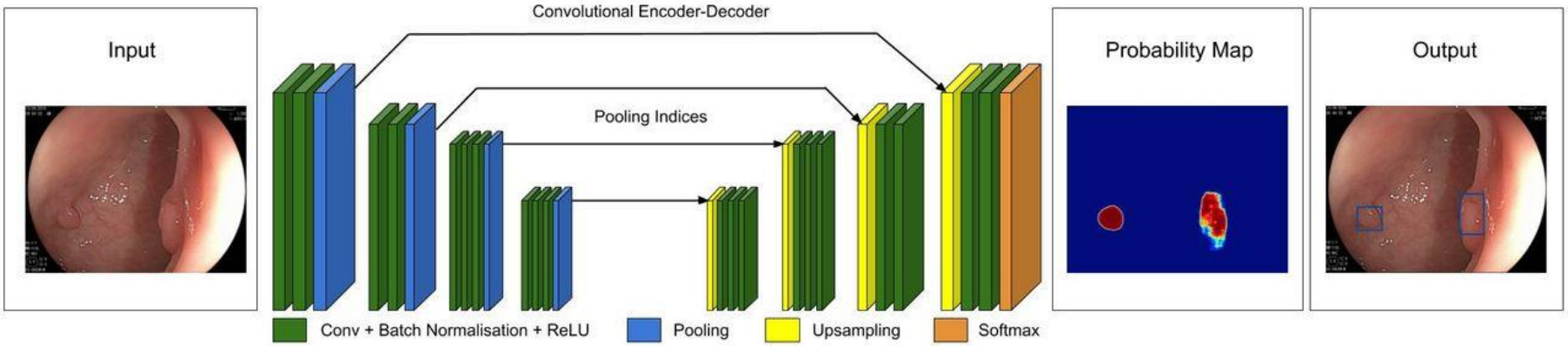
健康筛查	患病前	患病中	患病后
• 风险评估	• 早期诊断	• 风险分层-药物指导	• 生存预后分析

机器学习等AI技术应用于皮肤肿瘤的精准医疗



Esteva A, *et al*. ***Nature*** 542: 115-118 **(2017)**

人工智能应用于肿瘤早期筛查

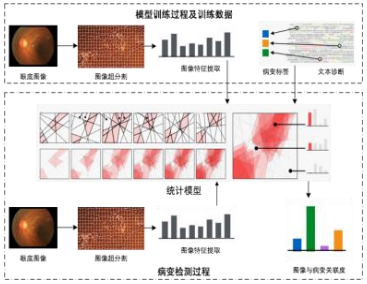


第一个研究AI辅助诊断是否能提高核心临床指标的前瞻性随机对照试验

通过卷积神经网络分析结直肠镜检查的图像，大大提高了息肉和腺瘤的检出率

Wang P, *et al*. ***Gut*** Feb 27. doi: 10.1136/gutjnl-2018-317500 [Epub ahead of print] (**2019**)

AI 辅助糖尿病视网膜病变早期筛查

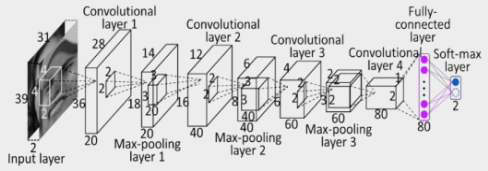


80万既往诊断报告学习训练

自动生成诊断报告

- ✓ DR分级诊断
- ✓ 病变位置检测
- ✓ 量化病变数据

云端深度学习自动检测系统



医生反馈诊断错误

在线学习
提高准确率

眼底拍照设备

与主治医生直接交流

通过面部图像诊断遗传病

数据

超过17,000张图像；包含上百种遗传病

原理

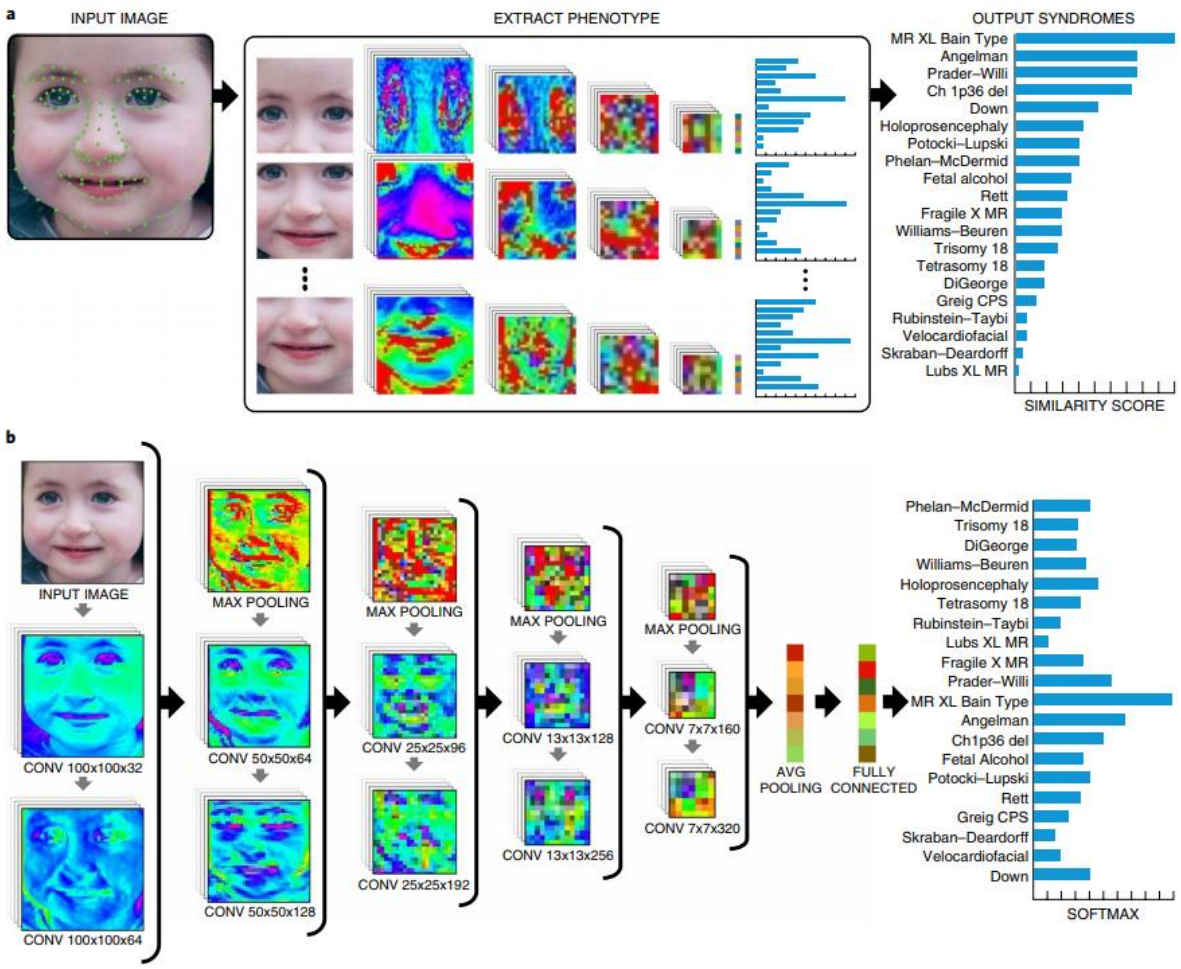
遗传病患者脸部具有明显基因特征

模型

DeepGestalt-基于级联深度卷积神经网络

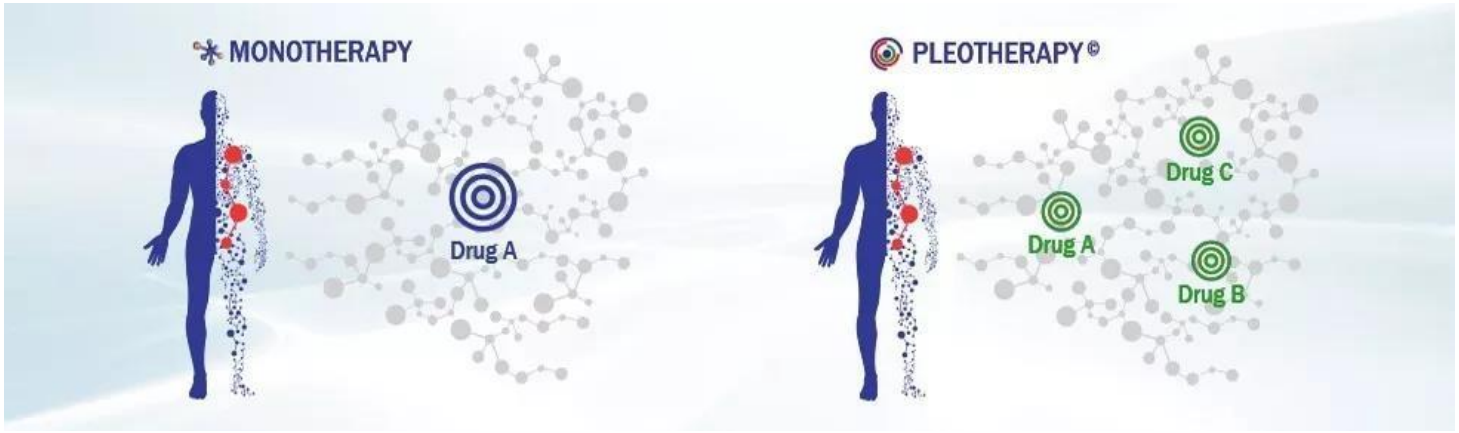
亮点

数据量大；精确度高；找到phenotype和genotype关联

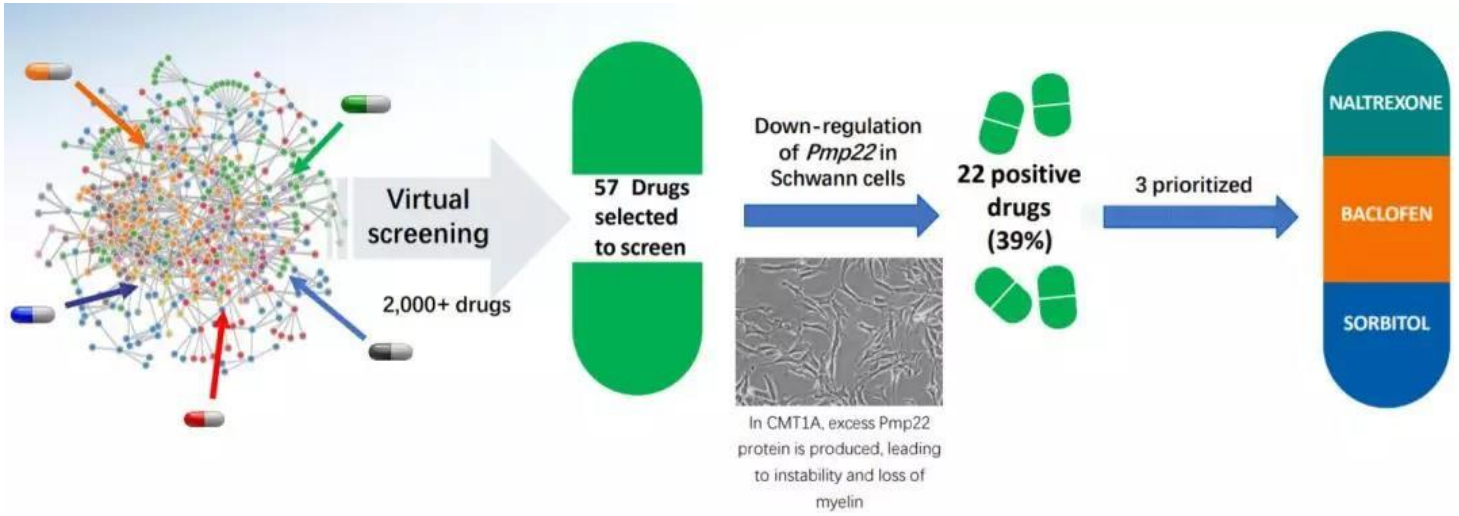


Gurovich Y, *et al*. ***Nat Med.*** 25(1): 60-64 **(2019)**

AI 辅助开发创新药物组合

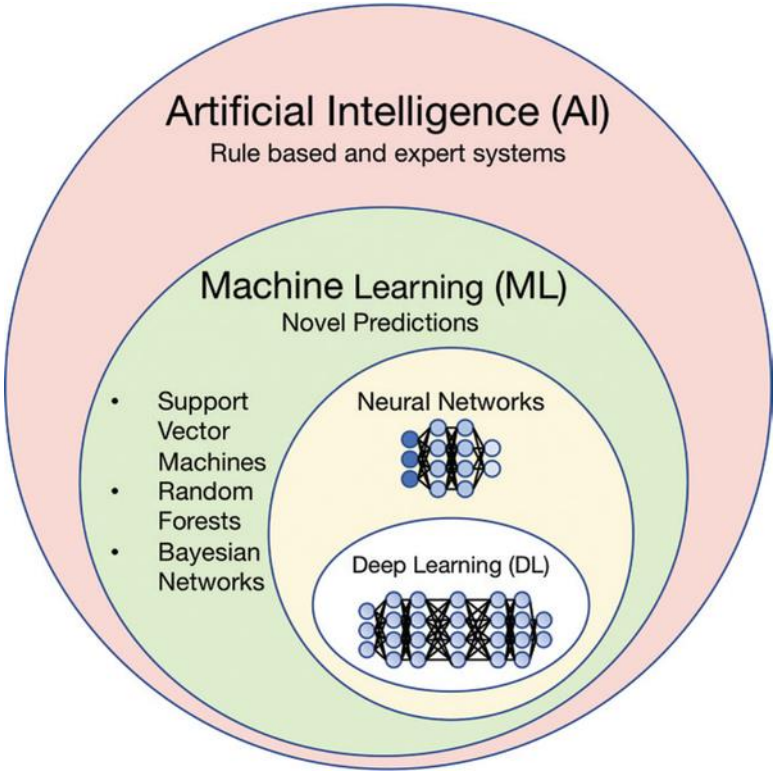


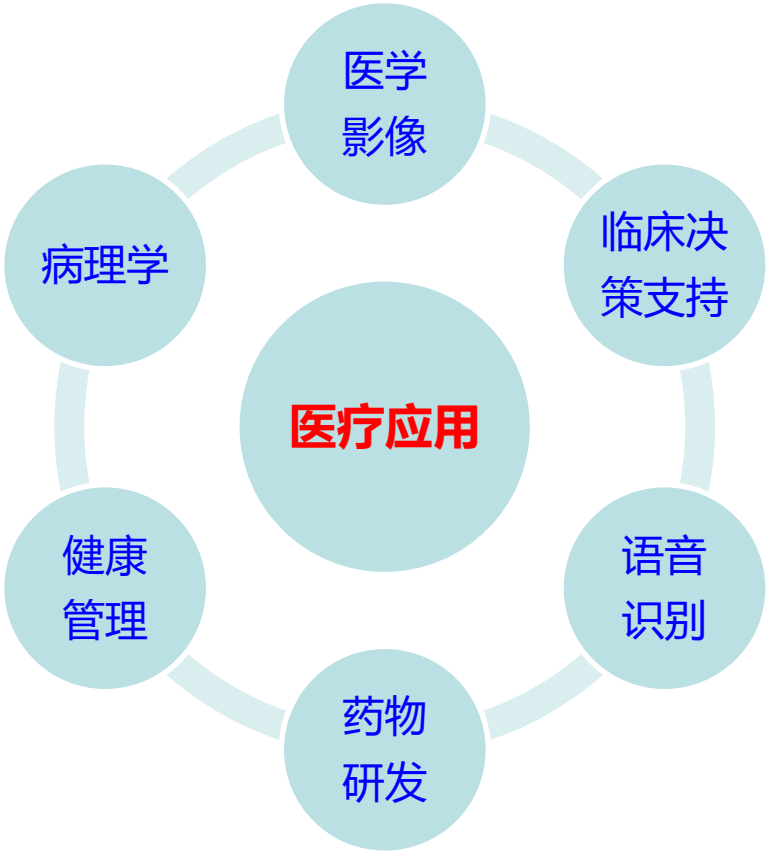
老药新用，AI助力开发创新组合疗法治疗罕见病



AI构建网络模型，3年研发PXT3003治疗腓骨肌萎缩症1A亚型（CMT1A），III期

人工智能在精准医学中的应用

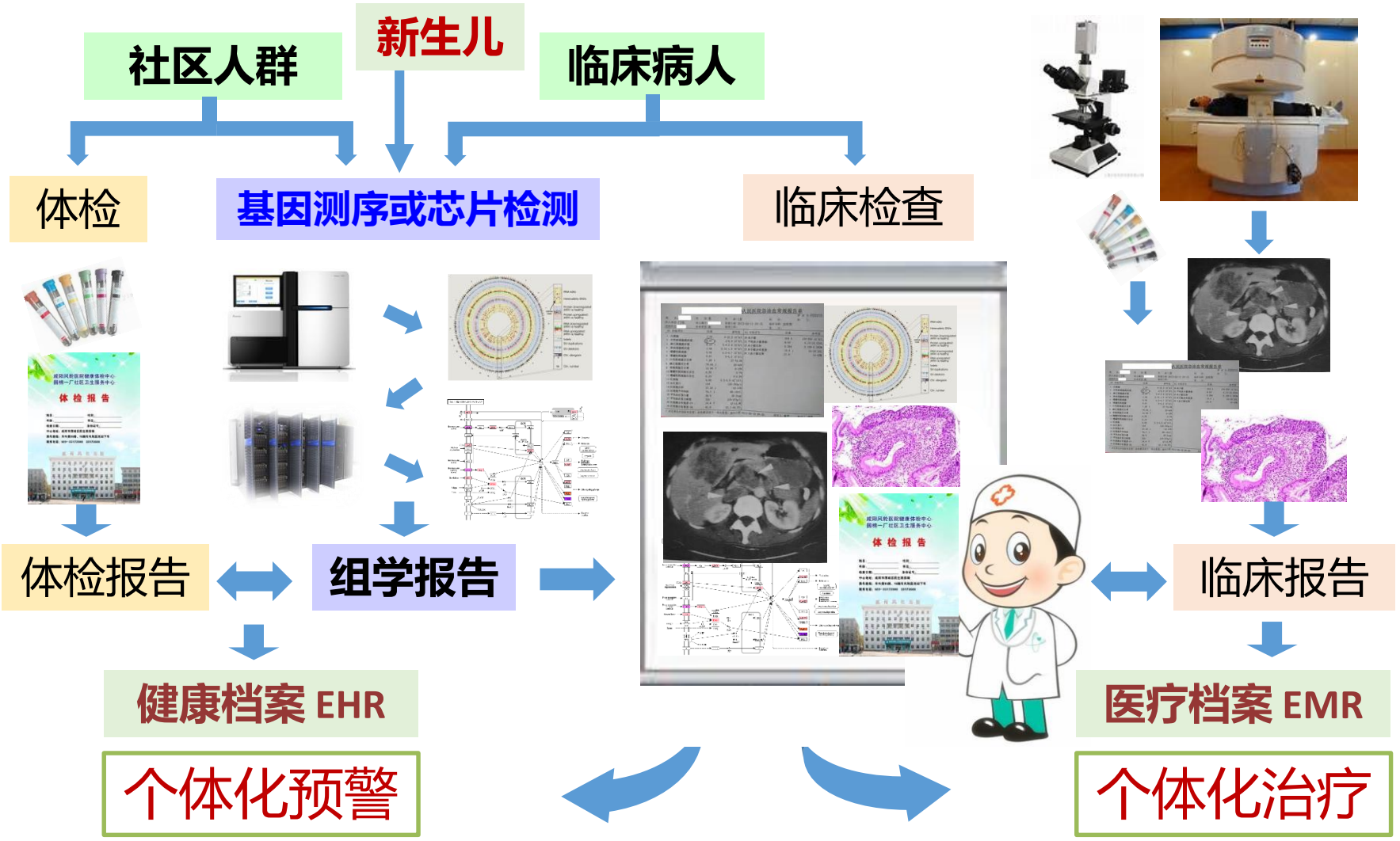




精准医学试图整合尽可能多的信息实现疾病预防和治疗，**“无AI，不精准”**

Rudie JD *et al.* ***Radiology*** 290(3): 607–618 (**2019**)
Mesko B. ***Expert Review of Precision Medicine and Drug Development*** 2(5): 239-241 (**2017**)

大数据时代的精准医学与健康管理



Genomics, Proteomics & Bioinformatics

HOSTED BY

ELSEVIER

Genomics Proteomics Bioinformatics

Impact Factor: 6.597

GPB's authors

Tomas Lindahl

Leroy Hood

Maynard Olson

Runsheng Chen

GPB's article types

Original research, letter, method, protocol;

Database, web server, resource;

Review, life-time achievement, resource review;

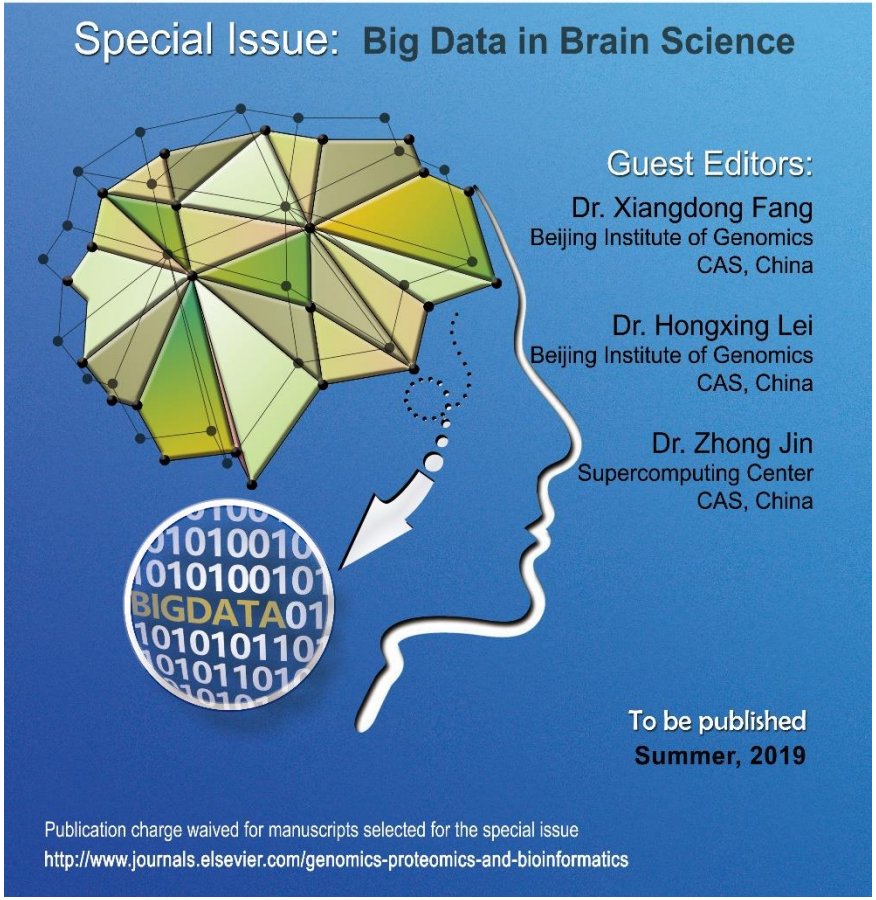
Historical note, news and view, perspective…

HOSTED BY

ELSEVIER

GPB

Genomics, Proteomics & Bioinformatics



衷心感谢！

会议主办单位

中华人民共和国科学技术部
Ministry of Science and Technology of the People´s Republic of China

中华人民共和国
国家卫生和计划
生育委员会

NSFC

国家科学基金

科研/合作团队

敬请批评指正

方向东
fangxd@big.ac.cn