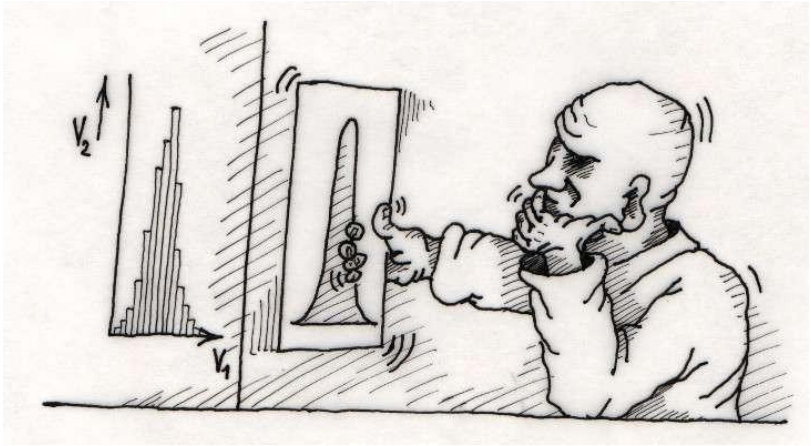
生物信息学：
组学时代的生物信息数据挖掘和理解

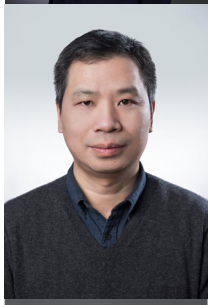
2021年春



有关信息

- 授课教师：宁康，刘欣，陈卫华，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：“医学八年 生物信息 2021”


考评

课程成绩

=

课堂考勤（10%）

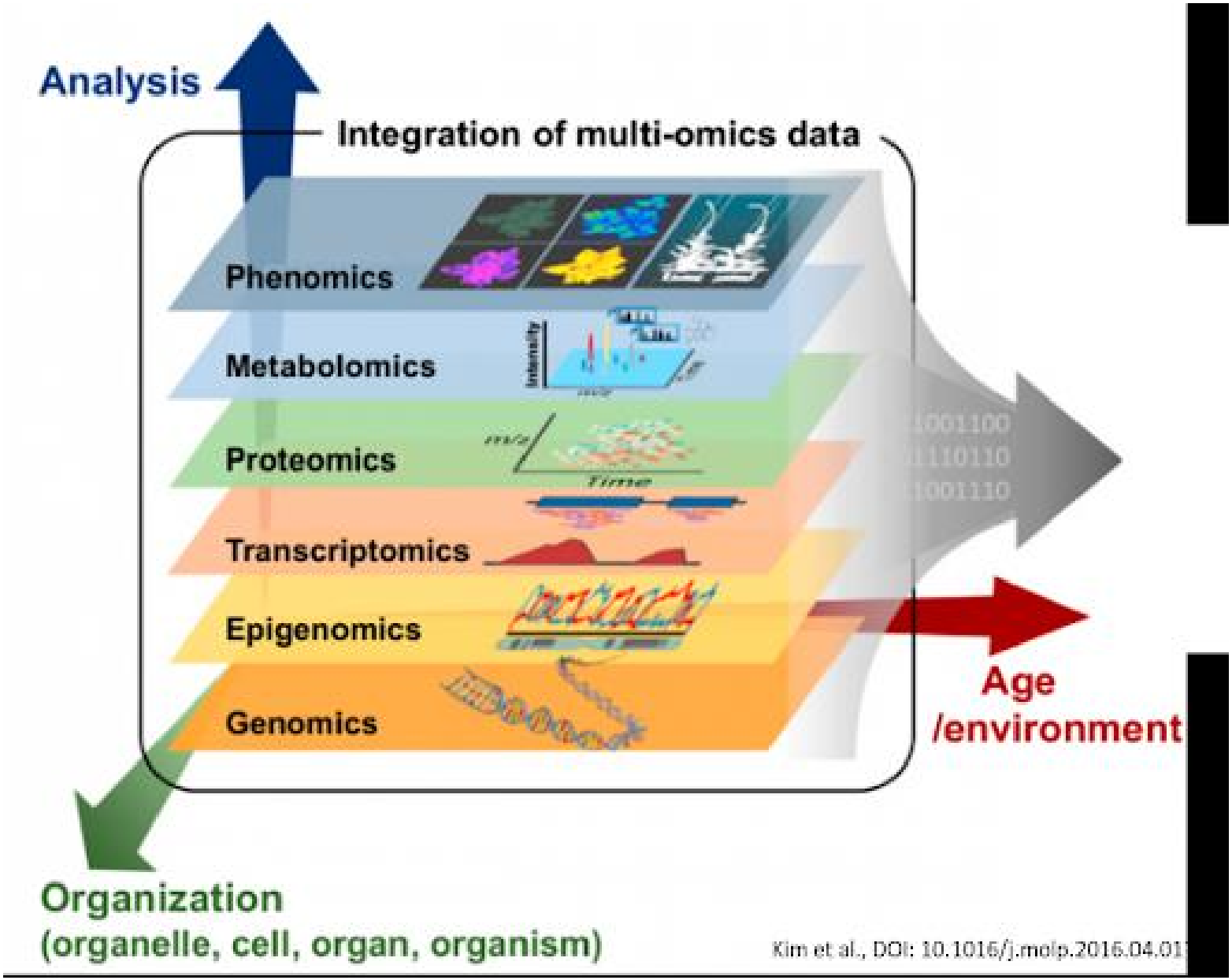
+课堂讨论&随堂测验（20%）

+终结性考试（70%）

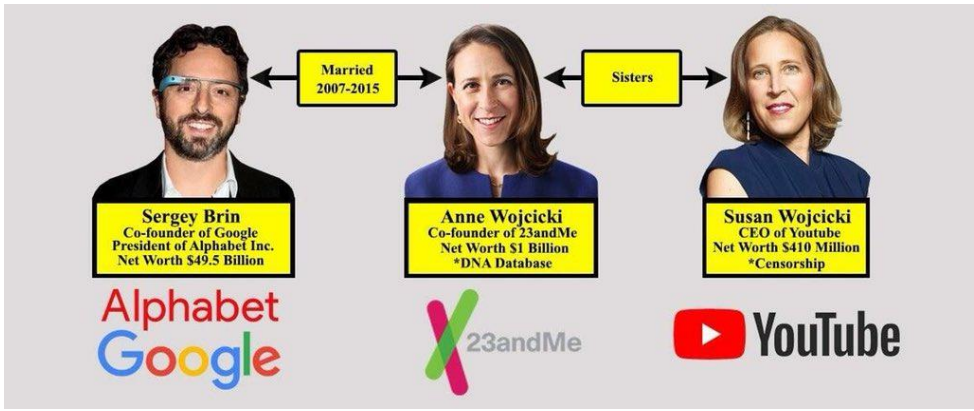
Bio-Big-Data Research

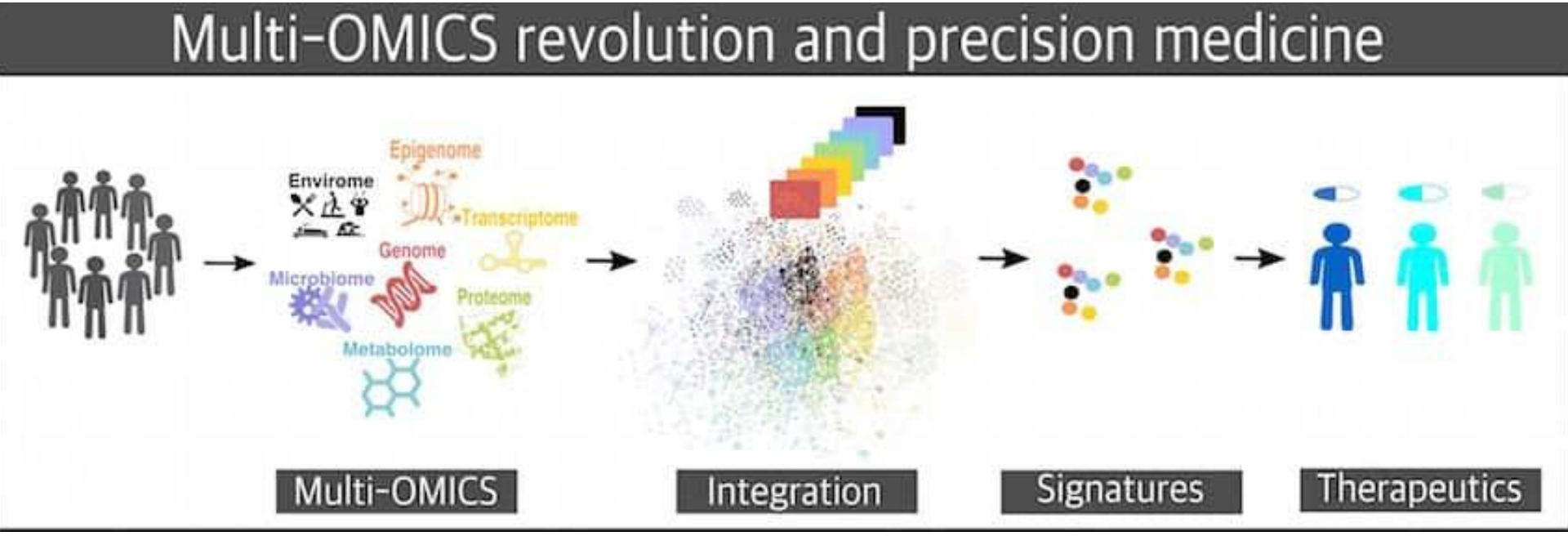
生物大数据：什么是组学

我们曾经认为组学（omics）是一小部分人的事情。。。

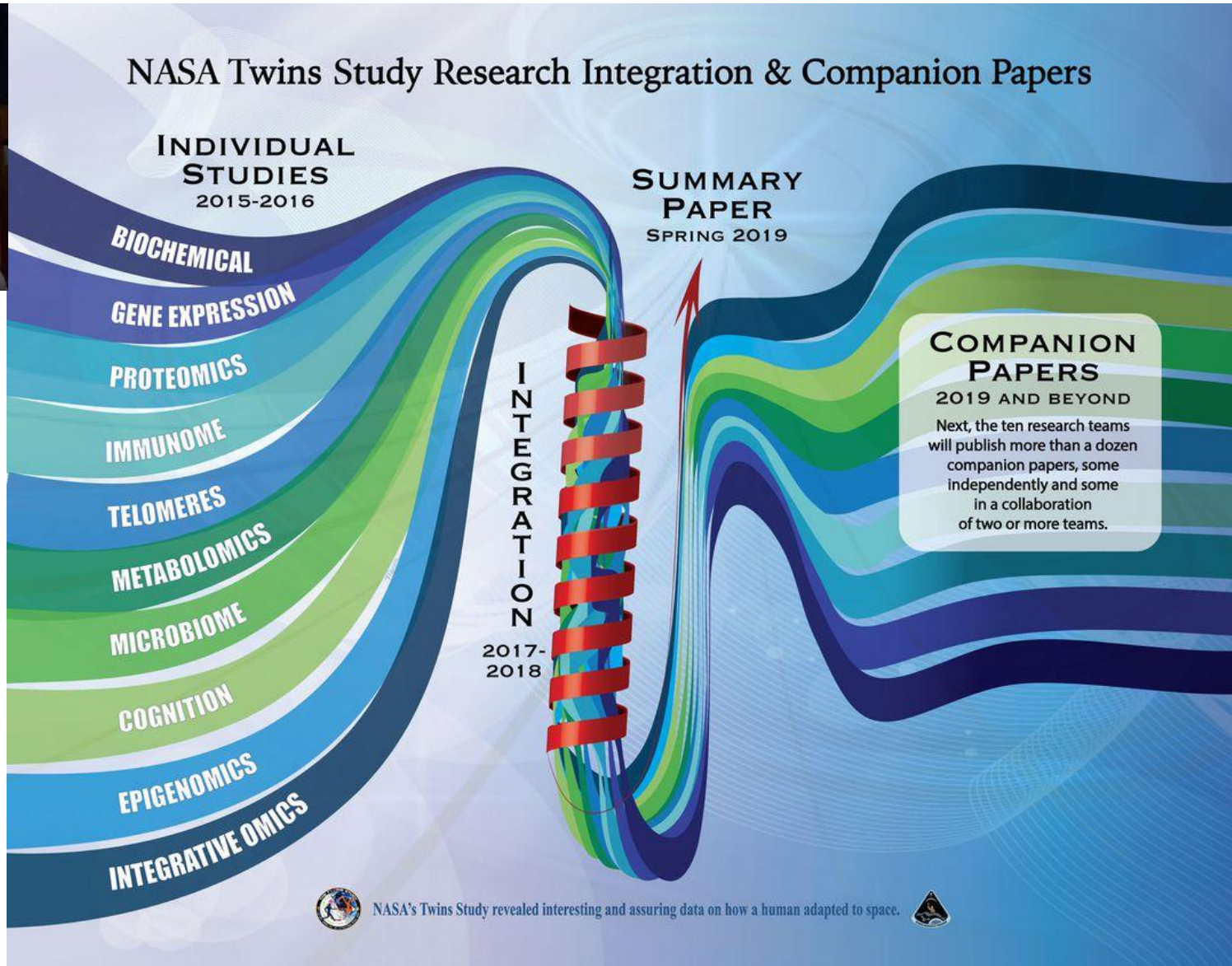


他们告诉我们：组学是每个人的事情





他们告诉我们：组学是每个人的事情



他们告诉我们：组学是每个人的事情

SARS-CoV-2 全基因组测序

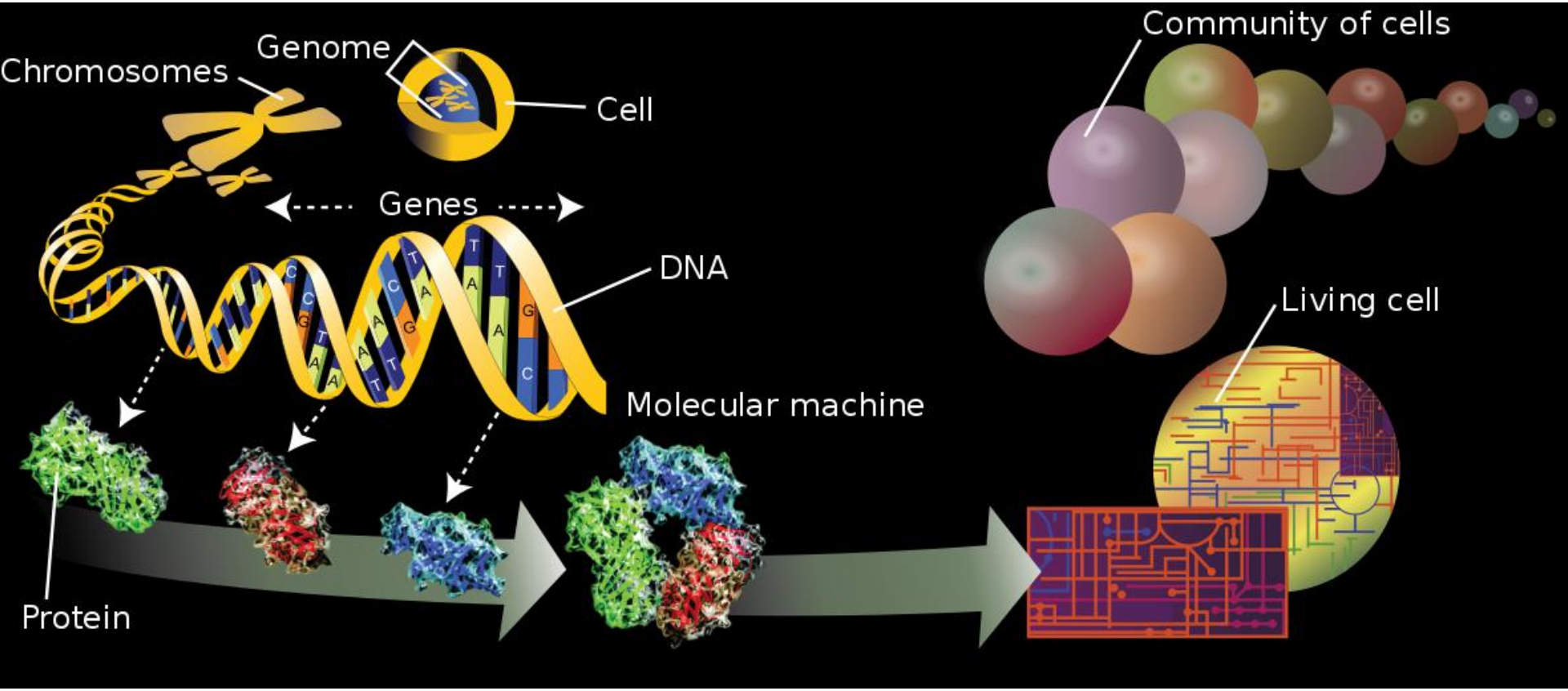
逆转录步骤	~ 1 小时
PCR	~ 2 小时
添加 barcodes	~ 1.5 小时
添加测序接头	~ 30 min
测序	~ 1 小时
分析	~ 1 小时

7小时

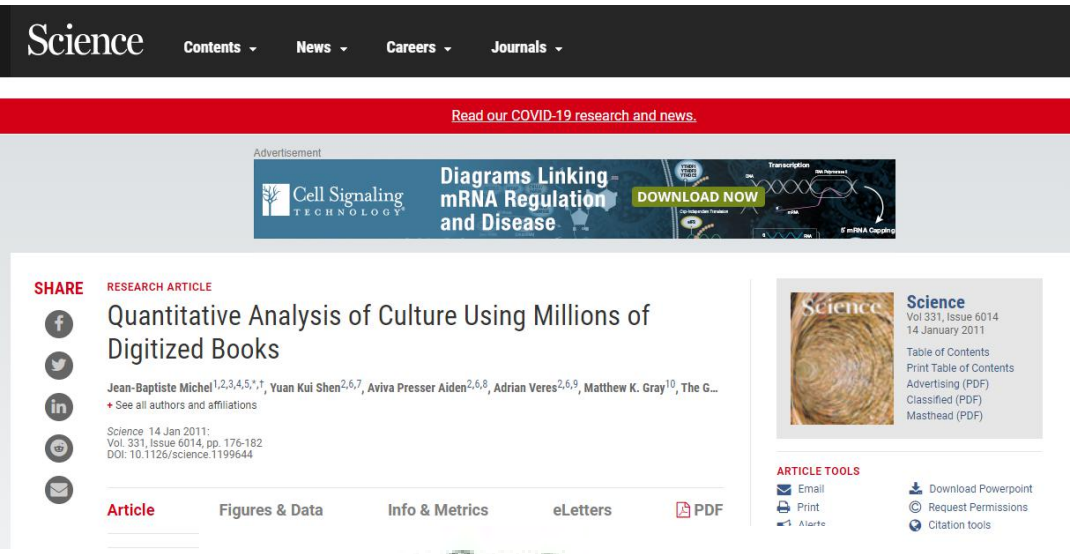
RNA 到
获得结果

其中约 **1 小时**
测序时间

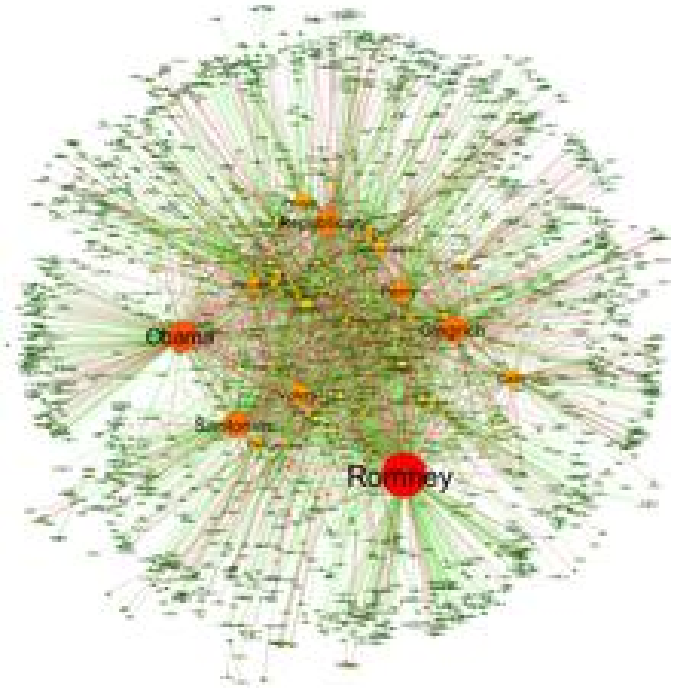
我们曾经认为组学（omics）是生物学的事情。。。



他们告诉我们：组学是更多学科的事情

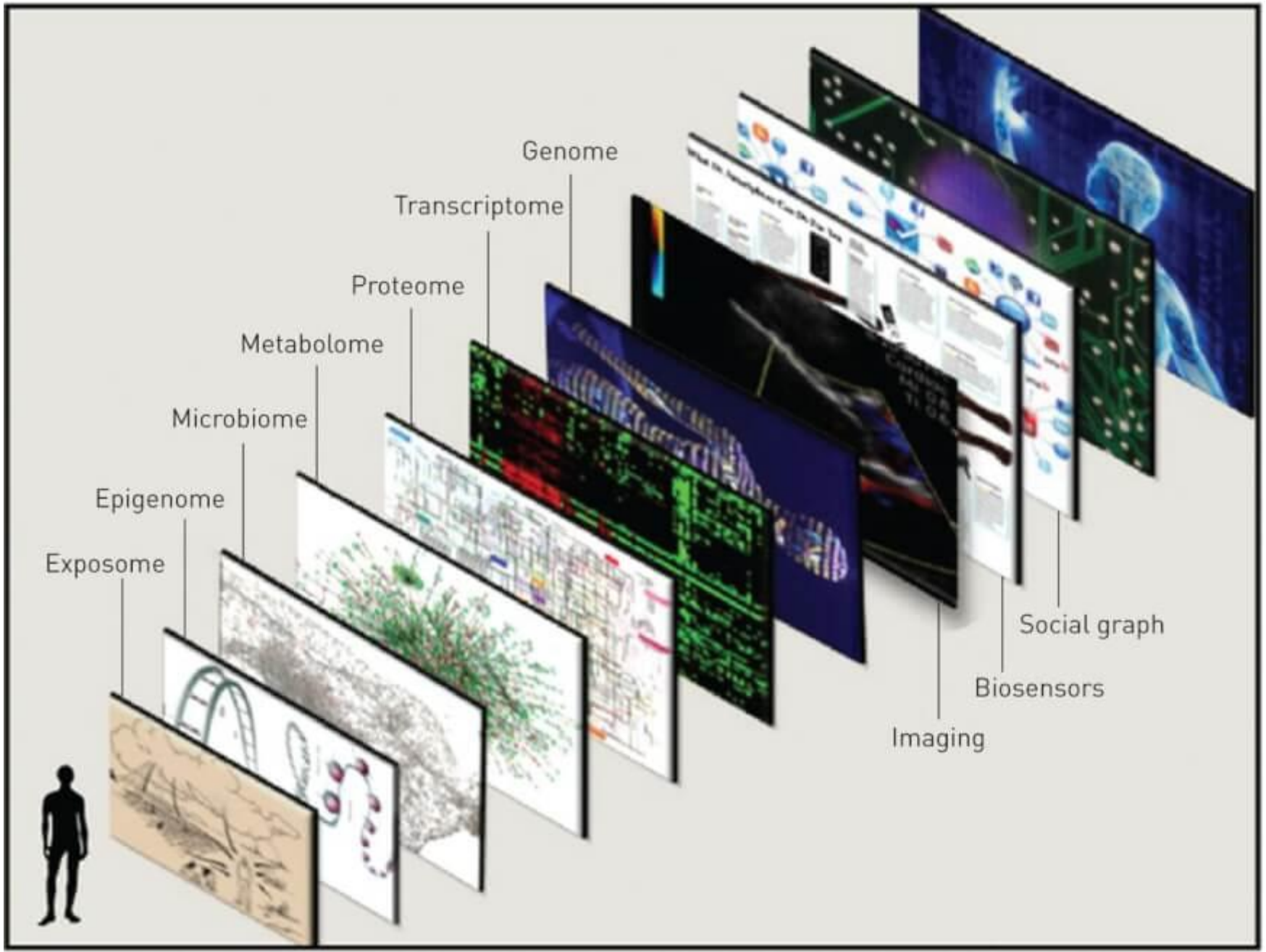


culturomics

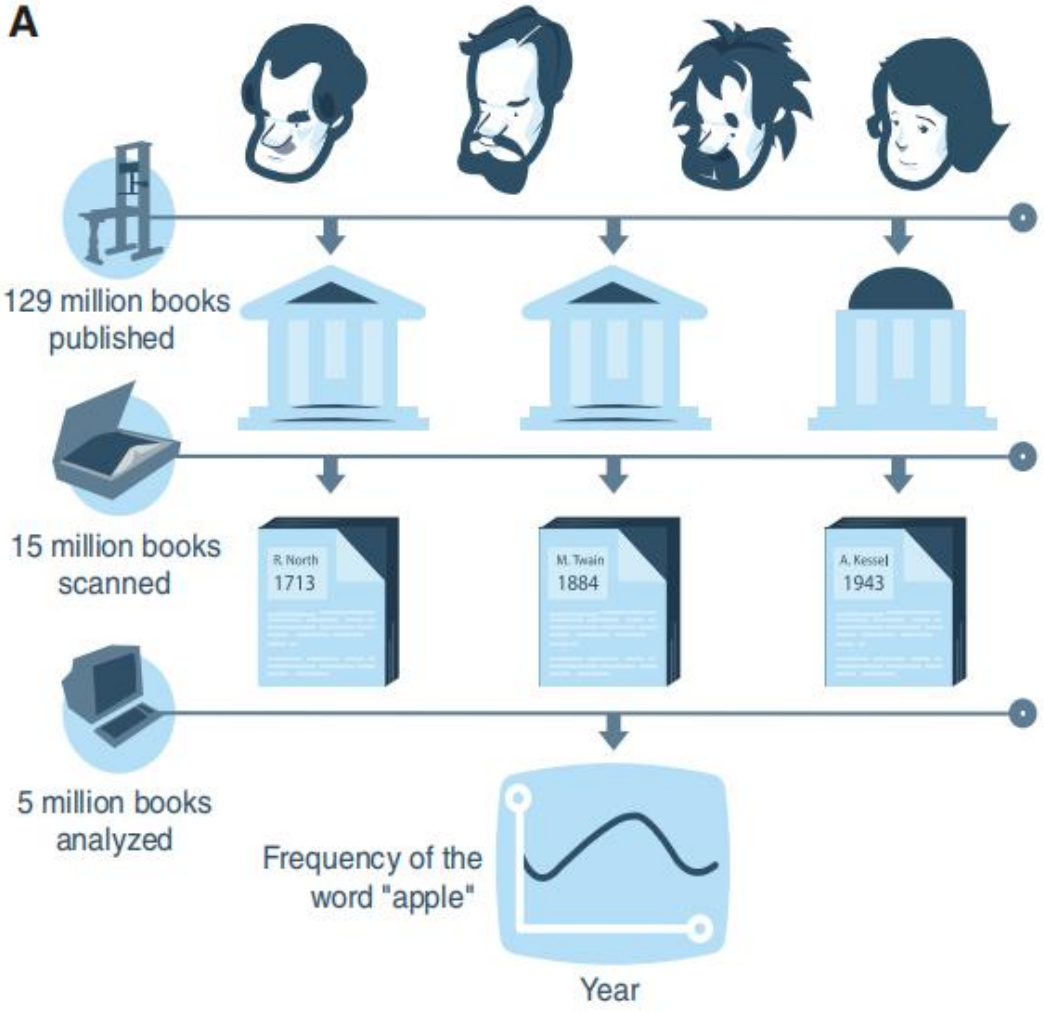


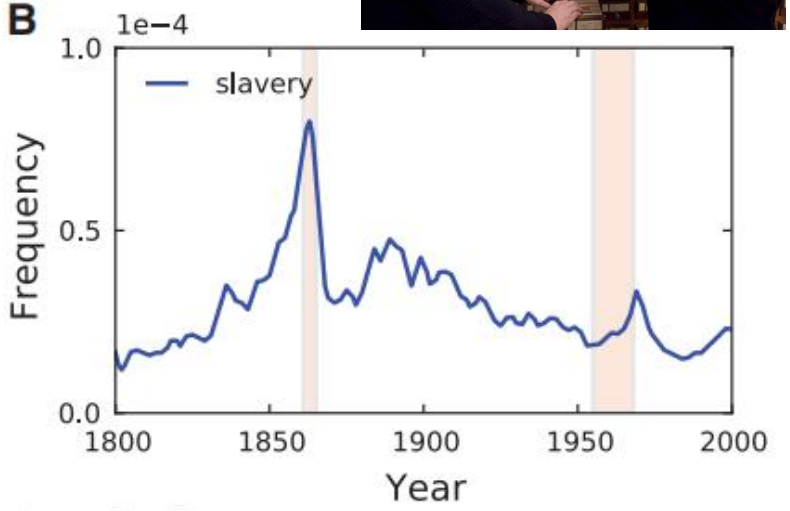
social omics

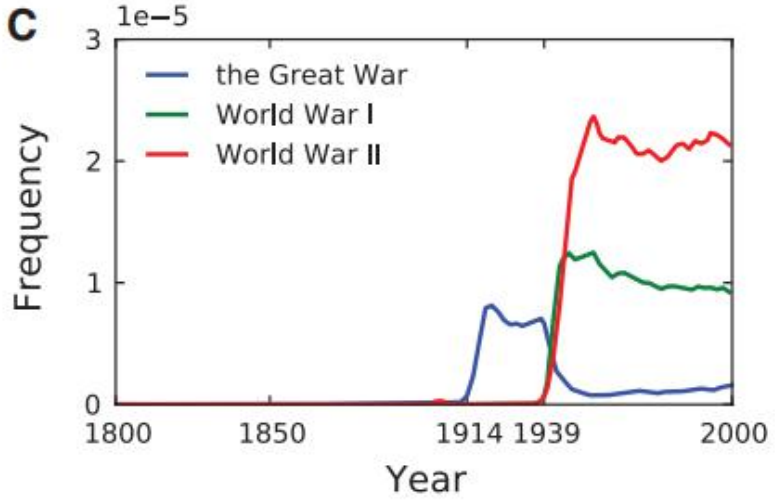
他们告诉我们：组学是更多学科的事情



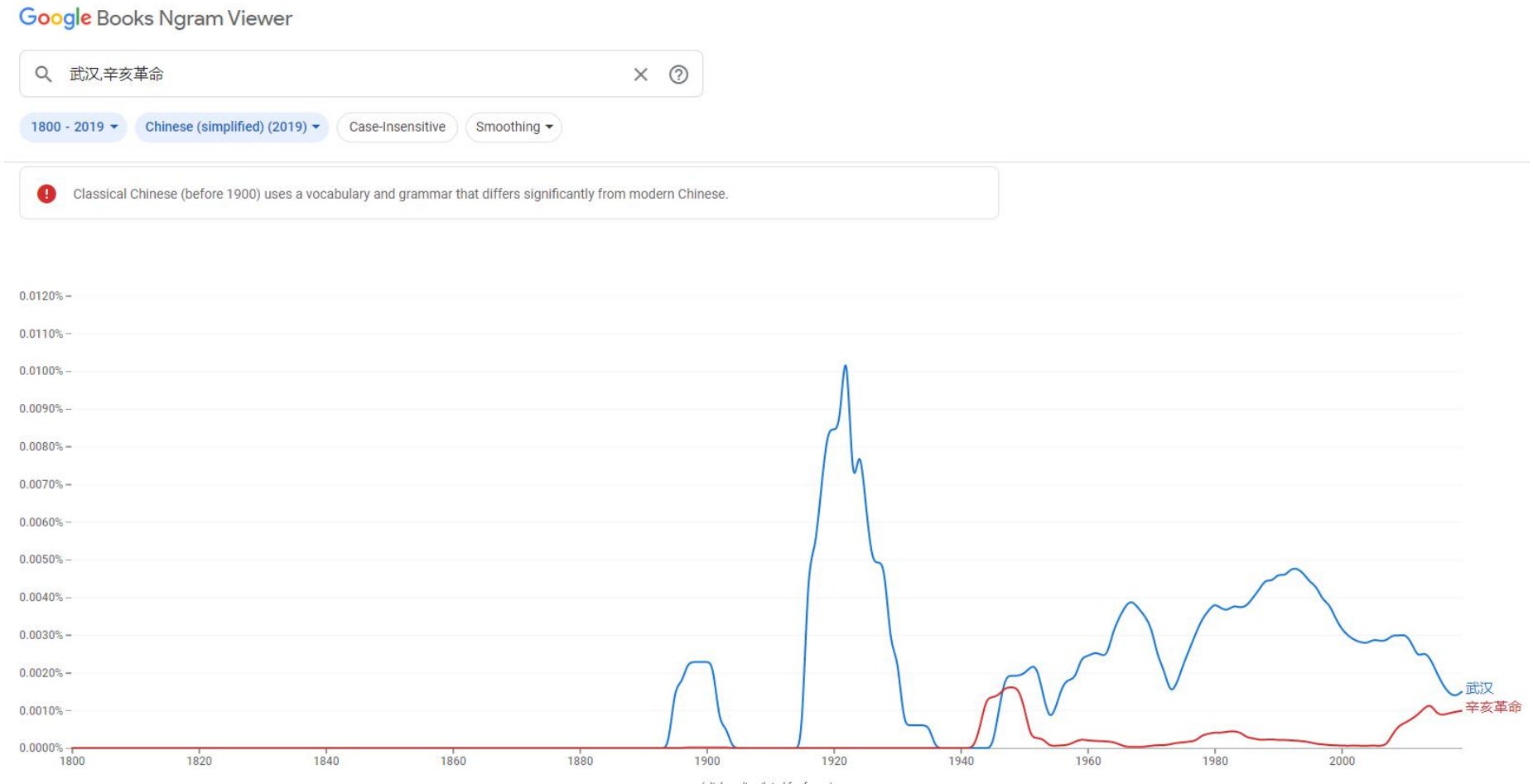
他们告诉我们：组学是更多学科的事情



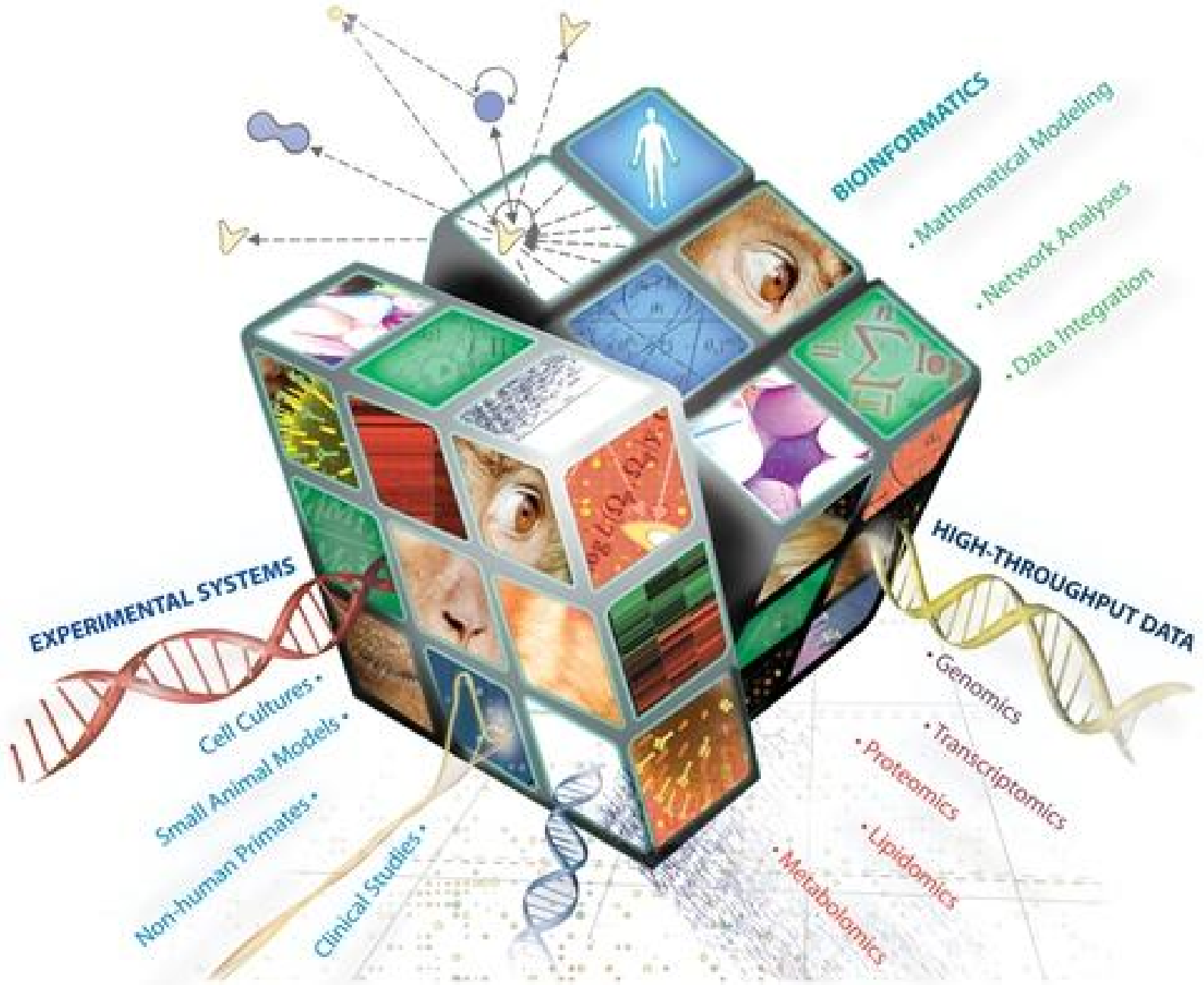




他们告诉我们：组学是所有学科的事情



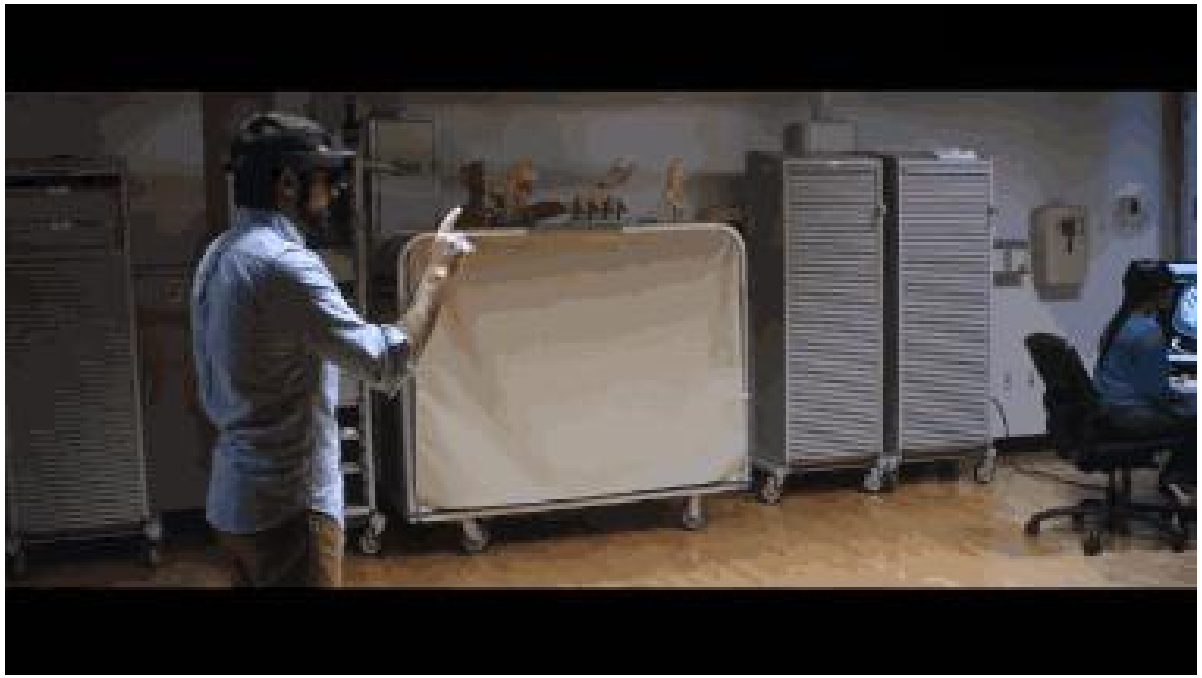
组学就是非结构化的数据，组学数据挖掘就是规律的探寻



Biomedical big-data…

"Have you ever figured how information-rich your stool is?," Larry asks me with a wide smile, his gray-green eyes intent behind rimless glasses. "There are about 100 billion bacteria per gram. Each bacterium has DNA whose length is typically one to 10 megabases—call it 1 million bytes of information. This means human stool has a data capacity of 100,000 terabytes of information stored per gram. That's many orders of magnitude more information density than, say, in a chip in your smartphone or your personal computer. So your stool is far more interesting than a computer."

-- Larry Smarr



Microbiome and big-data…

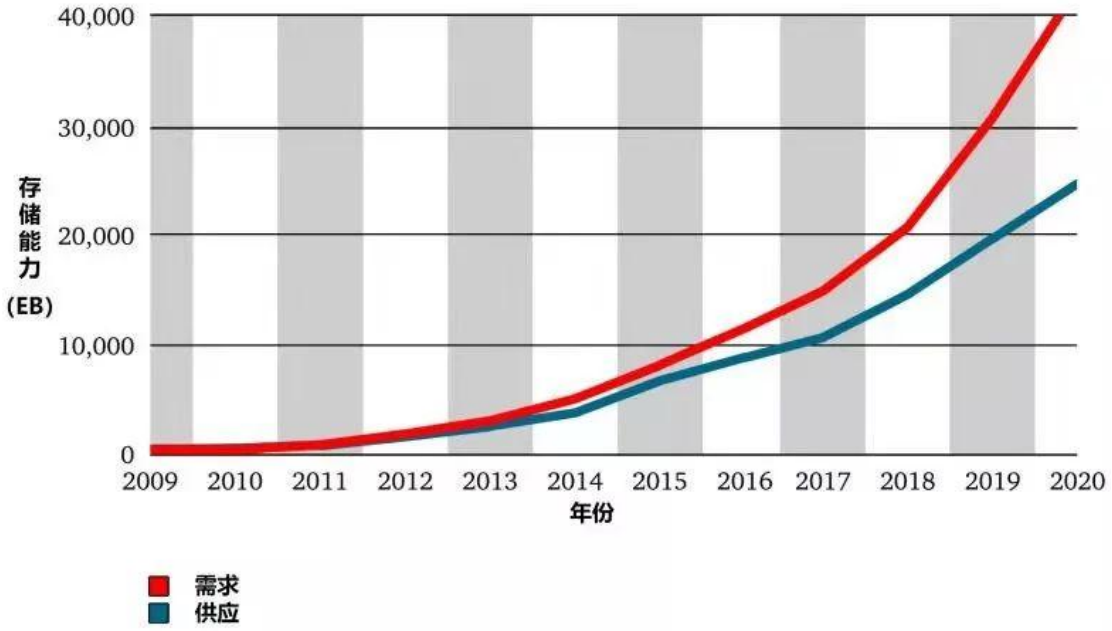
Microbiome in 4D

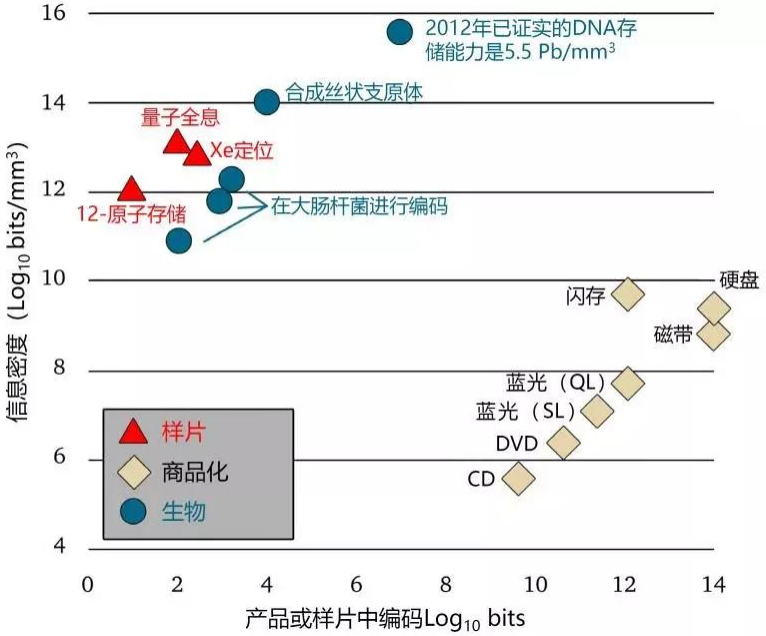
The development of antibiotic-resisitance in time-series

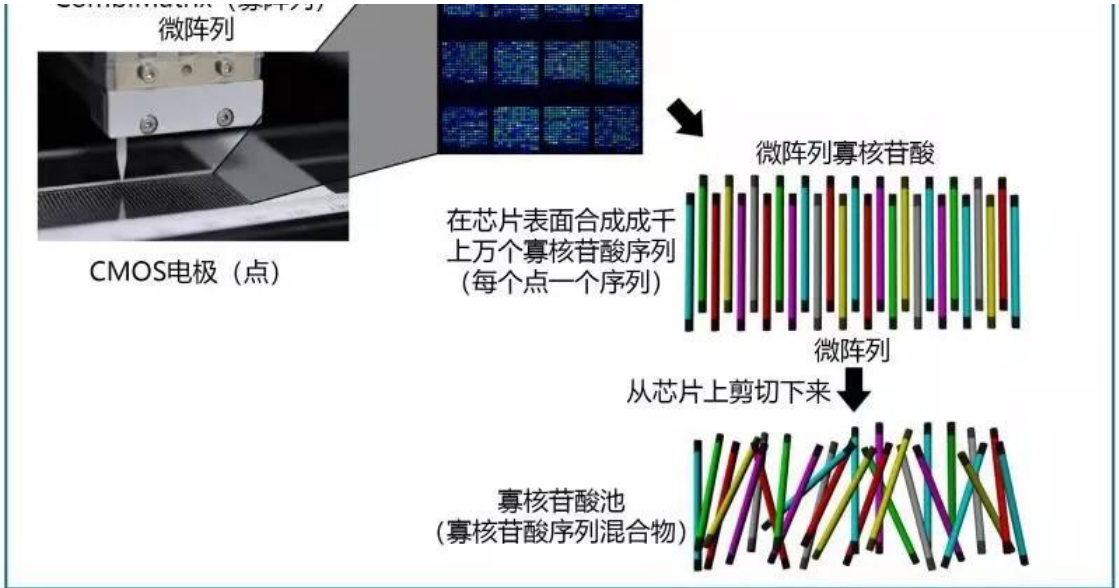
Science, 2016

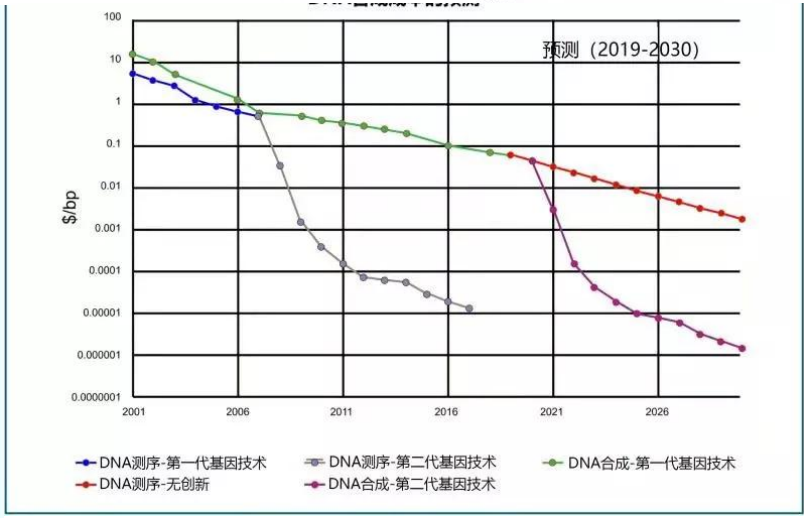
Science, 2019

Understand it, create it!



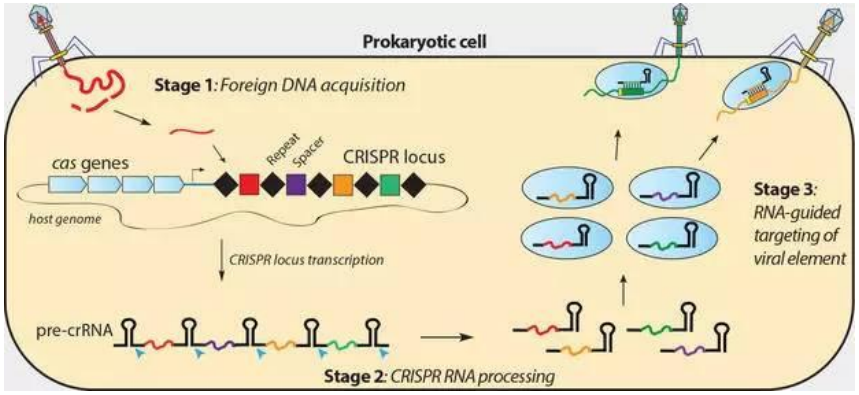


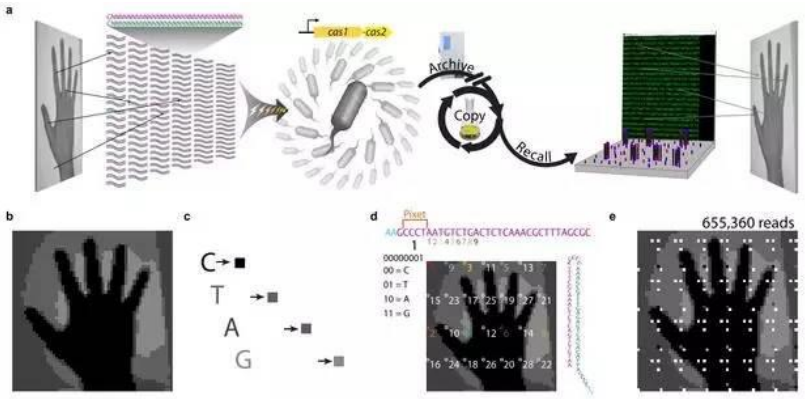


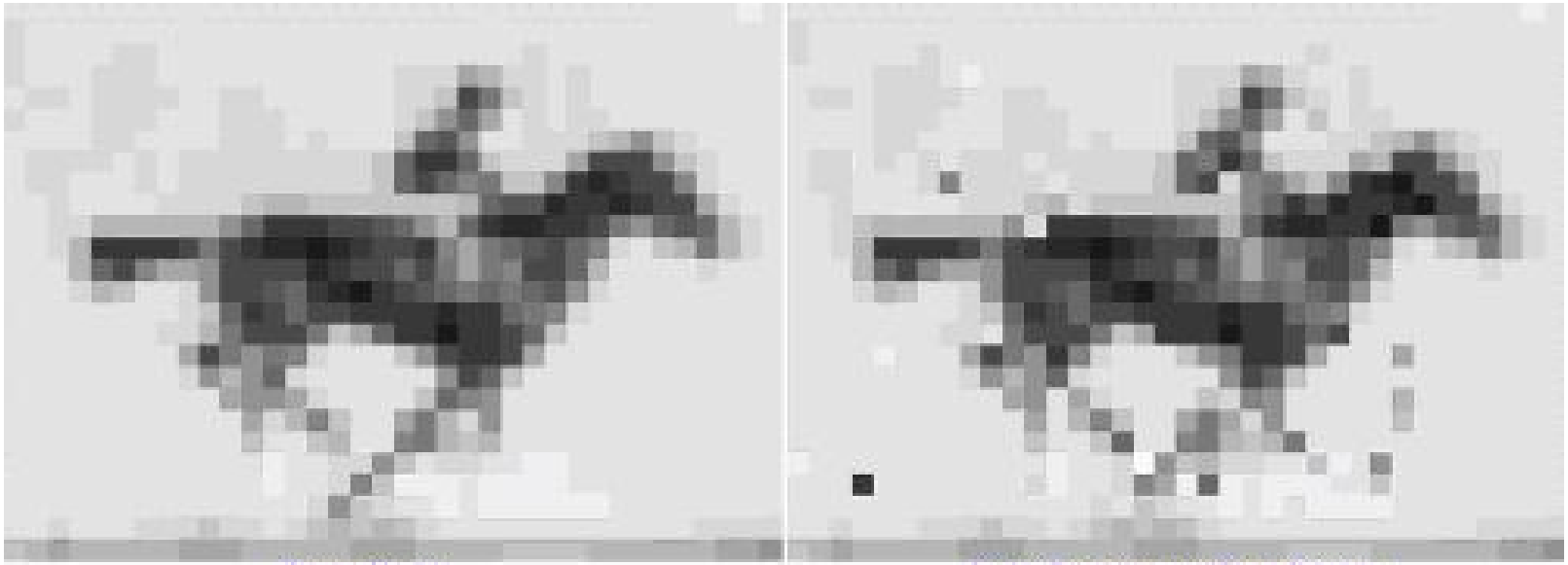


DNA数据存储的现在和未来

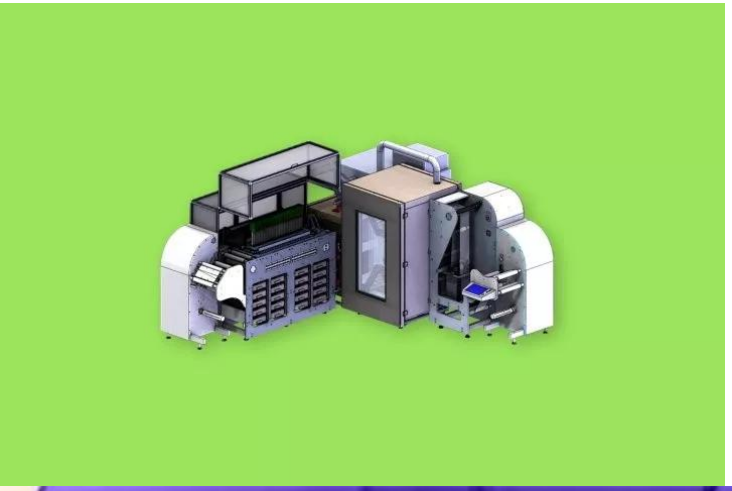
Understand it, create it!







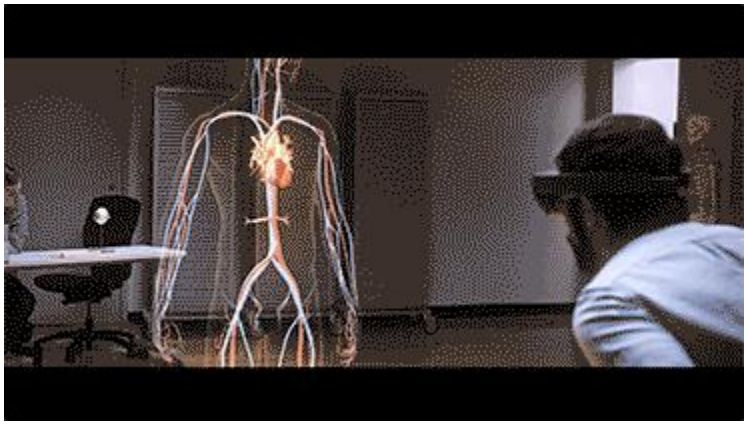
Understand it, create it!



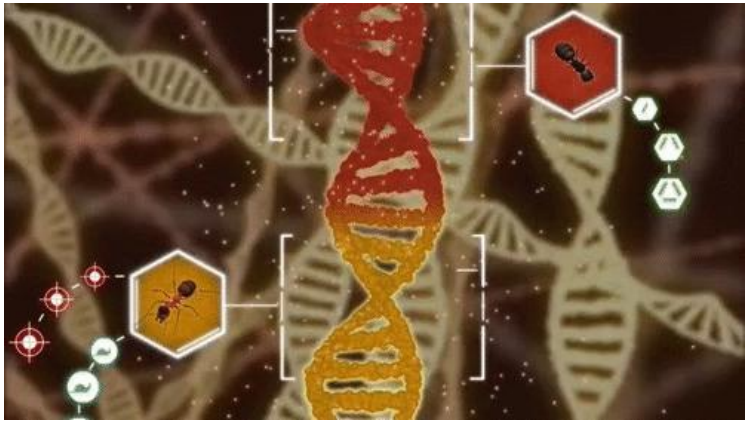
用DNA存储数据
CATALOG DNA Data Writer

我们的数字世界正在飞速产生新的数据，而存储这些数据的成本很高，而且会占用物理空间。

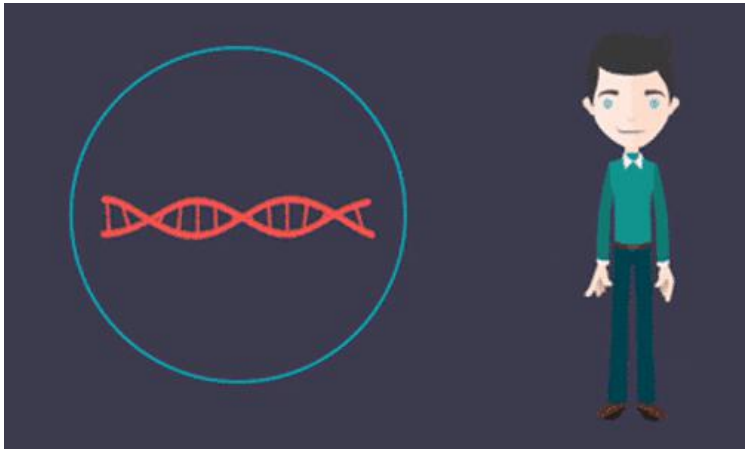
该公司最近宣布，他们已经将全部维基百科英文版数据存储进了DNA链上，耗时约12小时——大约是之前速度的1000倍。



See it!



Understand it!

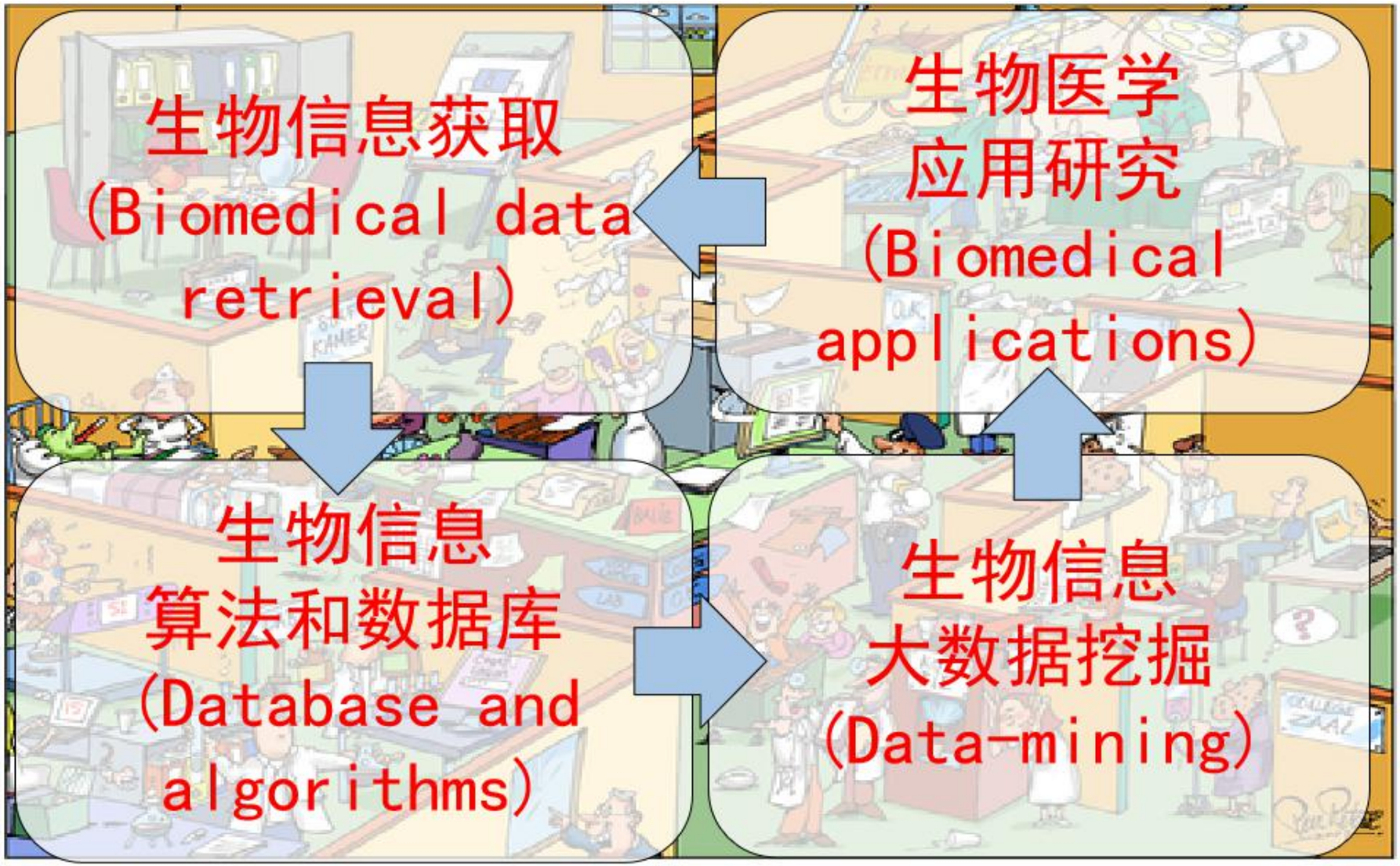


Create it!

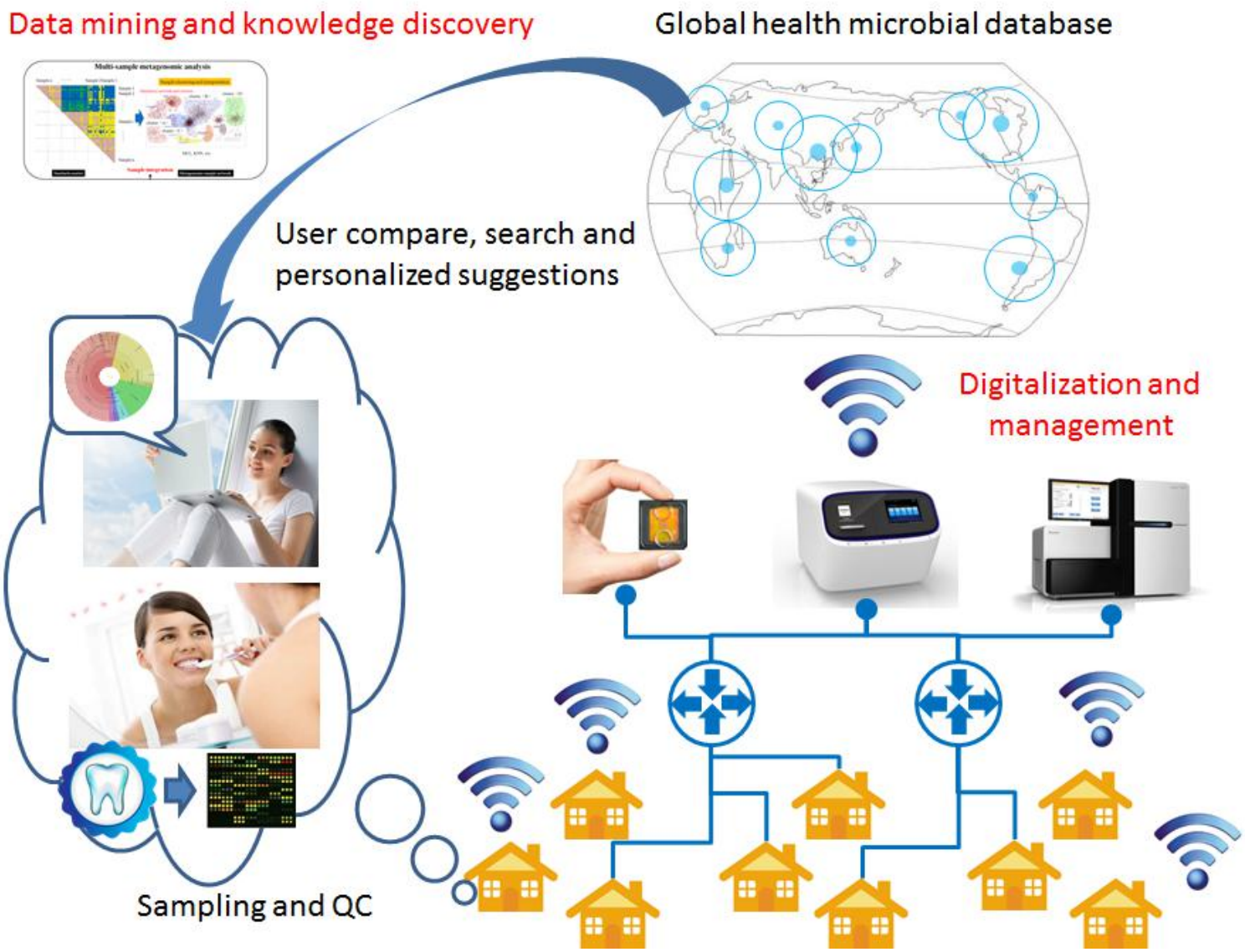
Biomedical big-data…

- 这不是科幻小说. . .
- 这就是当下正在发生的真事！
- 这就是你可以学到的东西！

生物信息学方向


生物信息获取
(Biomedical data retrieval)
生物医学
应用研究
(Biomedical applications)
生物信息
算法和数据库
(Database and algorithms)
生物信息
大数据挖掘
(Data-mining)

生物信息学方向


Data mining and knowledge discovery
Global health microbial database
User compare, search and personalized suggestions
Digitalization and management
Sampling and QC

Bioinformatics & Systems Biology @HUST

生命学院刘笔锋团队发表单细胞蛋白质组学研究成果

来源：华中大新闻网 浏览次数： 3743 发布时间：2014-05-27 编辑：党委宣传部

新闻网讯 5月21日，国际化学顶级杂志《德国应用化学》（Angewandte Chemie International Edition）在线刊发了生命学院刘笔锋教授团队的重要研究成果，论文题为“基于活性探针的单细胞化学蛋白质组学：鉴定原代神经元低拷贝膜蛋白”（Single Cell Chemical Proteomics with an Activity-based Probe: Identify Low Copy Membrane Proteins on Primary Neurons）。

据介绍，随着人类基因组计划的完成，从系统生物学角度认识生命现象的本质规律已成为新的科学范式之一，各种组学如基因组学、蛋白质组学和代谢组学等的发展方兴未艾。由于细胞间的个体差异性，在单个细胞水平探讨生命过程的物质基础是近年来的热点问题，例如当下如火如荼的单细胞基因组学和单细胞测序计划。刘笔锋教授团队在2013年提出了基于微流控芯片的超高通量单细胞基因组学新方法研究环境助迫下的基因组损伤，

《核酸研究》同期发表生命学院“健康大数据”团队5篇论文

来源：生命学院 浏览次数： 2524 发布时间：2018-01-05 编辑：

新闻网讯 1月4日，核酸及生物信息学等领域顶级期刊《核酸研究》（Nucleic Acids Research，影响因子：10.162）的“annual Database Issue”同期刊发生命学院“健康大数据”团队的5篇论文，其中3篇论文我校为第一作者单位、团队成员为通讯作者，标题分别为“lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs”、“iUUCD 2.0: an update with rich annotations for ubiquitin and ubiquitin-like conjugations”和“MVP: a microbe–phage interaction database”，同时还参与发表了“Database Resources of the BIG Data Center in 2018”和“dbCoRC: a database of core transcriptional regulatory circuitries modeled by H3K27ac ChIP-seq signals”两篇论文。

生命学院薛宇教授团队发现神经细胞自噬的重要调控因子

来源：生命学院 浏览次数： 2107 发布时间：2017-12-29 编辑：

新闻网讯（通讯员 陈蕾蕾）12月27日，生命学院薛宇教授与香港浸会大学李敏教授团队以封面论文的形式在细胞自噬领域的国际权威期刊《自噬》（Autophagy）上发表题为“Phosphoproteome-based kinase activity profiling reveals the critical role of MAP2K2 and PLK1 in neuronal autophagy”的论文。论文共同第一作者为陈蕾蕾博士、王勇博博士和宋聚贤博士，其中王勇博博士为我校11级硕博连读研究生，共同通讯作者为薛宇教授和李敏教授。

生命学院郭安源教授团队发布动物转录因子数据库第三版

来源：生命学院 浏览次数： 28535 发布时间：2018-09-13 编辑：

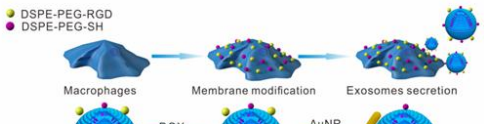
新闻网讯（通讯员 苗亚茹）9月12日，牛津大学出版社(Oxford University Press)出版的《核酸研究》(Nucleic Acids Research)（2017年影响因子11.56）在线发表了生命学院郭安源教授团队的动物转录因子数据库（AnimalTFDB）第三版（http://bioinfo.life.hust.edu.cn/AnimalTFDB/）。

郭安源教授为通讯作者，生命学院博士生胡慧和硕士生苗亚茹为共同第一作者。郭安源教授团队于2011年完成了动物转录因子数据库第一版，至今持续维护和更新了7年，三个版本的文章分别都发表在《核酸研究》。持续的更新、准确完善的数据和方便易用的在线平台使得该数据库成为了国际上转录因子相关研究领域最权威的资源。与其他相关数据库的比较显示AnimalTFDB是最准确可靠的动物转录因子数据库。AnimalTFDB自2012年初发表以来，用户遍及60多个国家，访问次数多达100万次，文章总引用近300次，引文包括多篇Cell和Nature等。

刘笔锋研究团队首次设计功能化外泌体将光热治疗与化疗结合高效靶向治疗癌症

来源：生命学院 浏览次数： 1657 发布时间：2018-03-30 编辑：

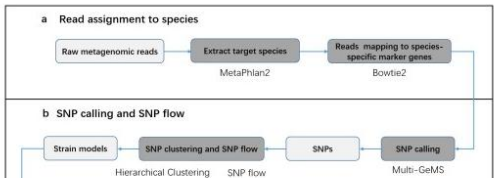
新闻网讯（通讯员 任一杰）3月12日，国际权威学术期刊《Advanced Functional Materials》在线刊发了刘笔锋教授团队的重要研究成果，题为“Designer Exosomes for Active Targeted Chemo Photothermal Synergistic Tumor Therapy”。（设计功能化外泌体将光热治疗与化疗结合高效靶向治疗癌症）



生命学院宁康教授团队微生物亚种识别方法研究取得新进展

来源：生命学院 浏览次数： 383 发布时间：2018-11-07 编辑：

新闻网讯（通讯员 谭重阳）10月5日，牛津大学出版社(Oxford University Press)出版的《Bioinformatics》（2017年影响因子5.481）在线发表了生命学院宁康教授团队的微生物亚种识别方法Strain-GeMS（https://github.com/HUST-NingKang-Lab/straingems）。



Bioinformatics & Systems Biology @HUST

生物信息与系统生物学

Data retrieval

生物医学信息获取技术团队

Data analysis

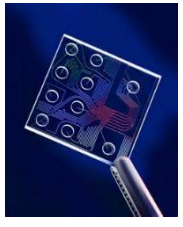
非编码RNA团队

Applications

系统生物学团队

蛋白质翻译后修饰组学团队

微生物信息学团队

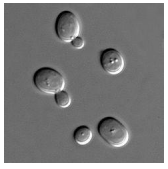


Lab on Chip

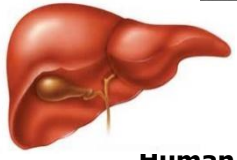
Linux Cluster & high-capacity storage

GPU computing system

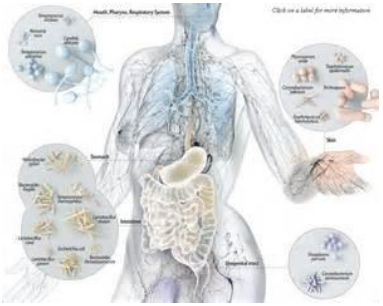
Yeast metabolism



Liver cancer



Human microbiome



Solexa GA-IIx

Thermo LTQ

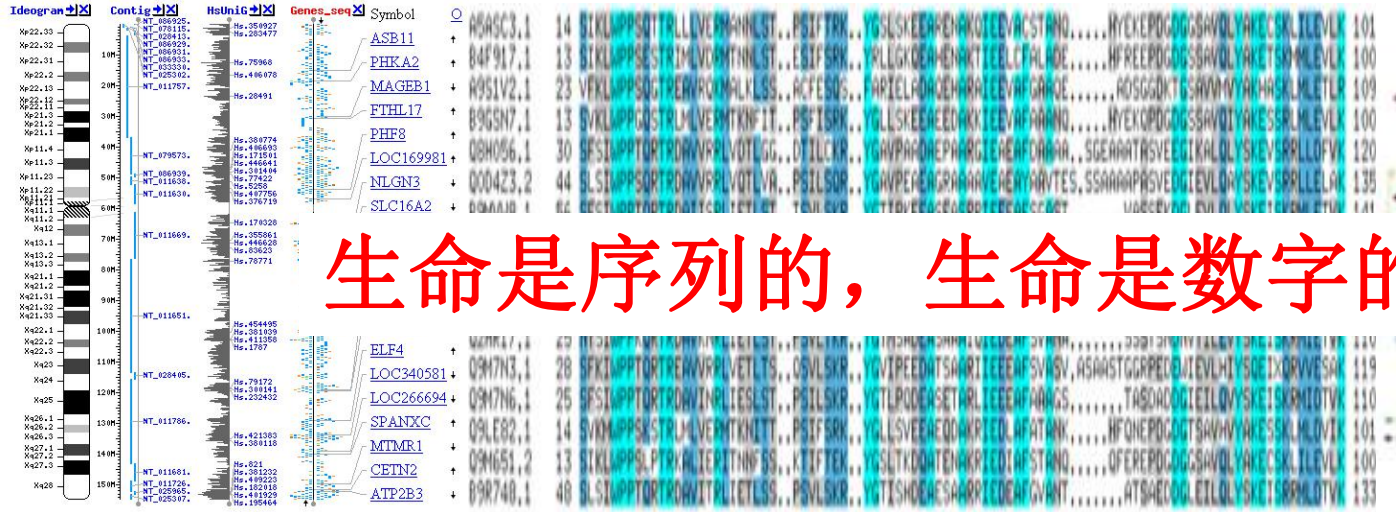
NMR

Chemostat

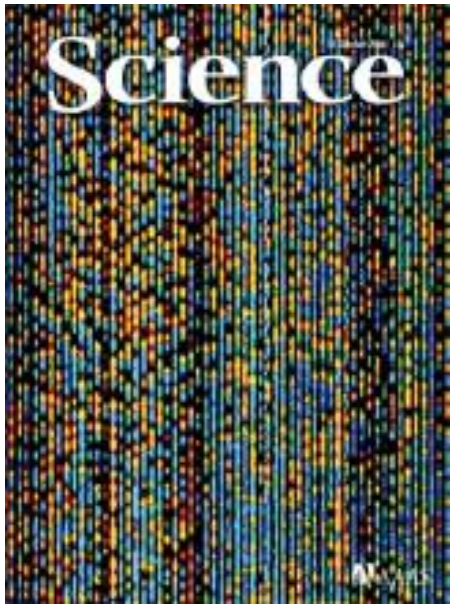
LC/GC

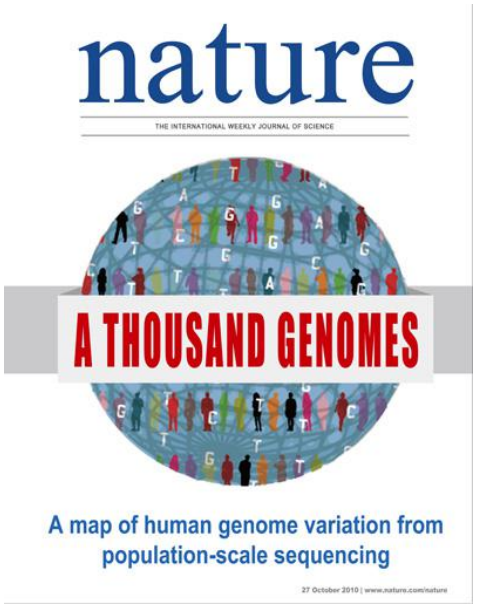
生物信息学：生物学视角

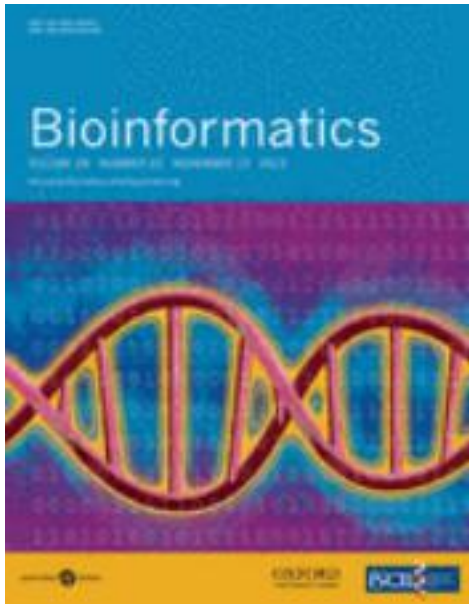
生物信息学@HUST



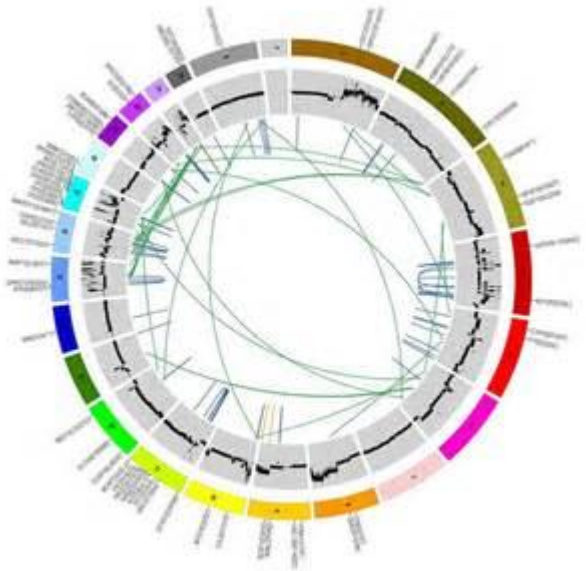
生命是序列的，生命是数字的！

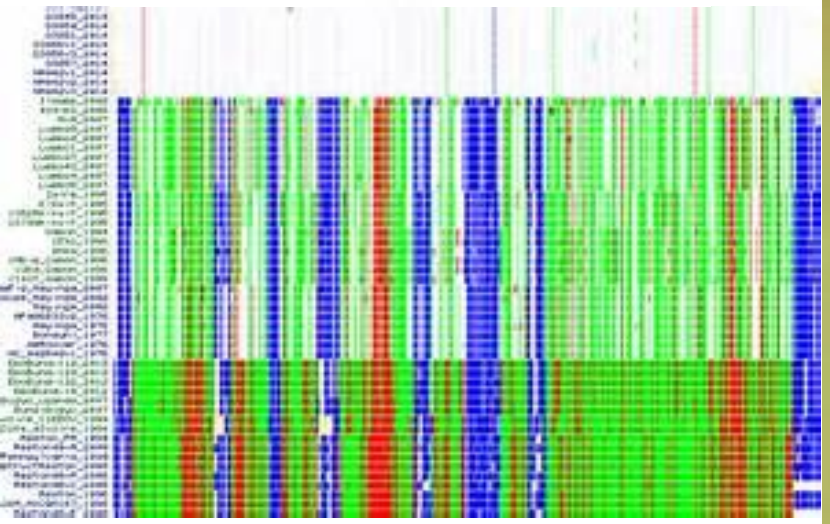


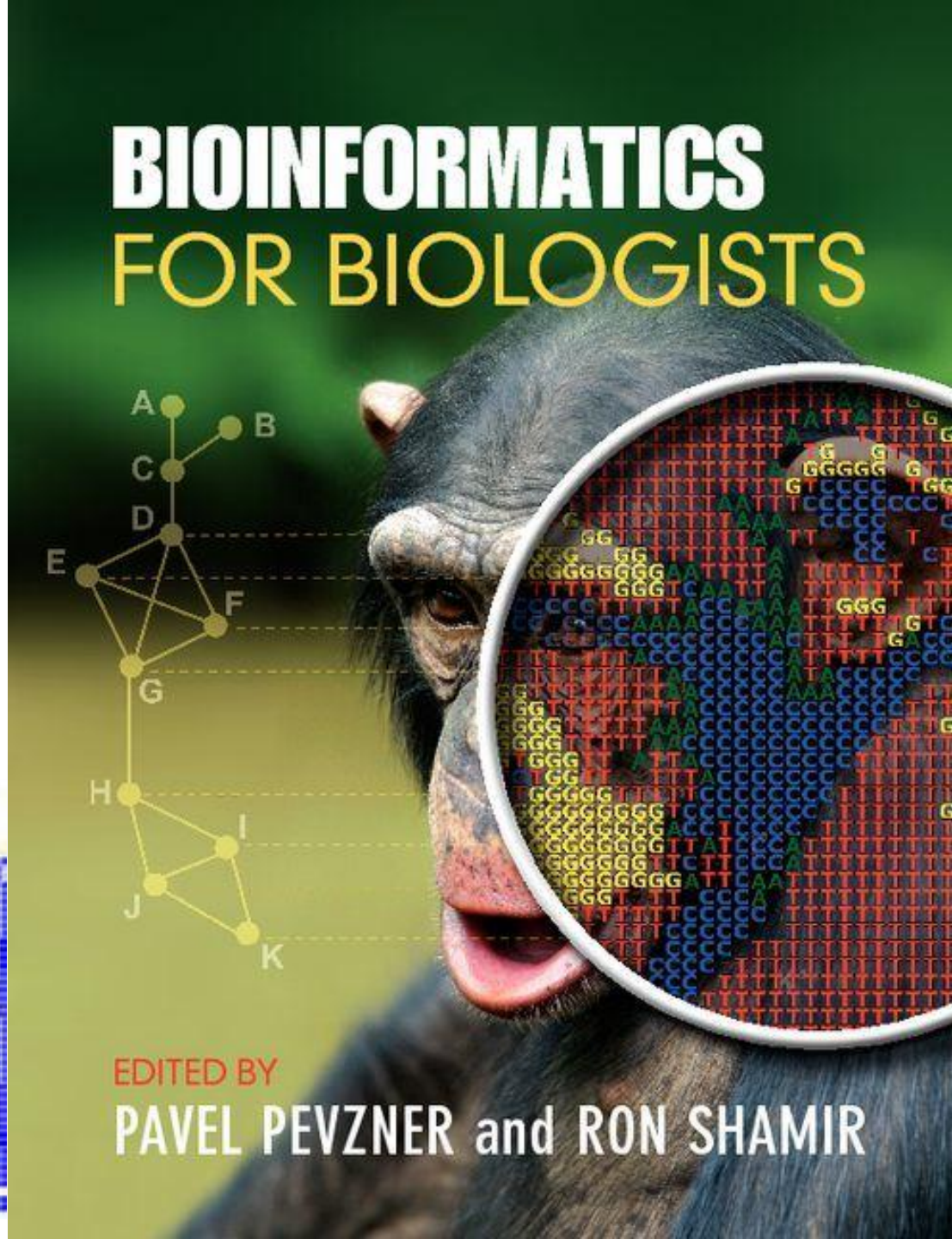




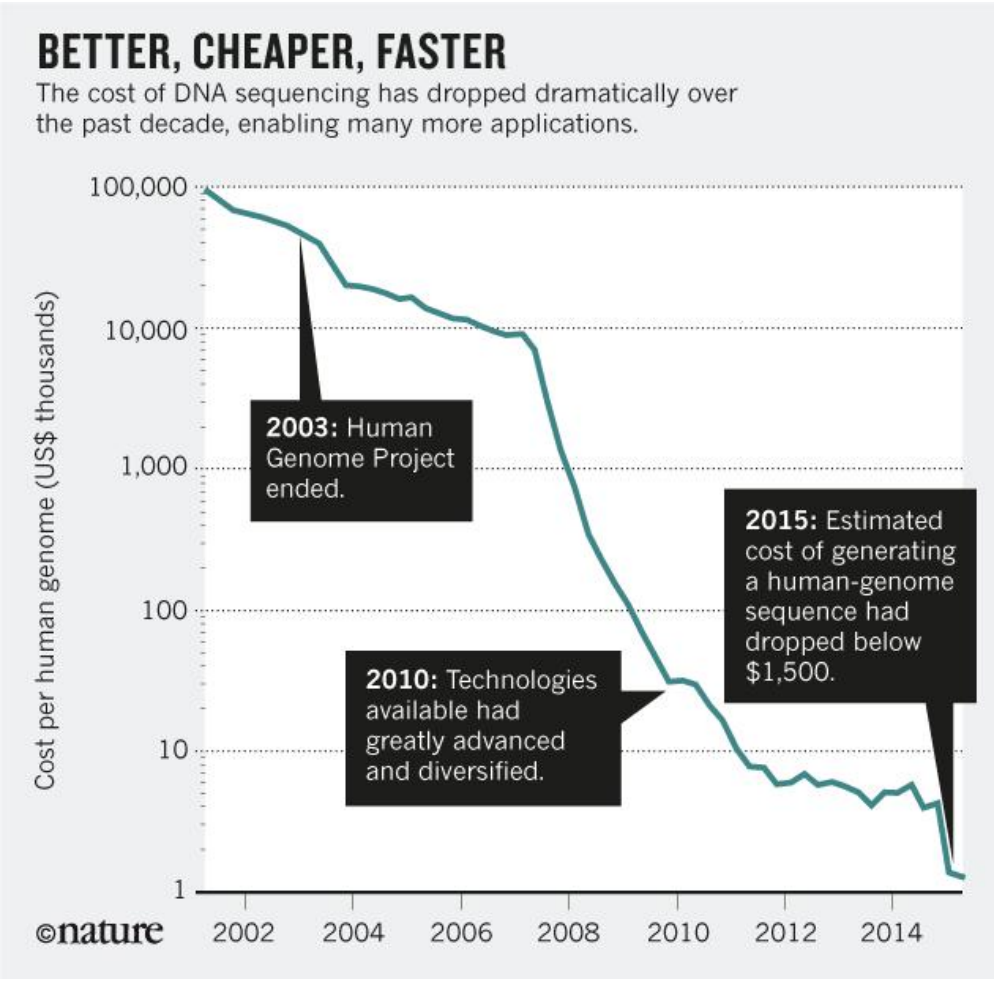
生命是序列的，
生命是数字的！

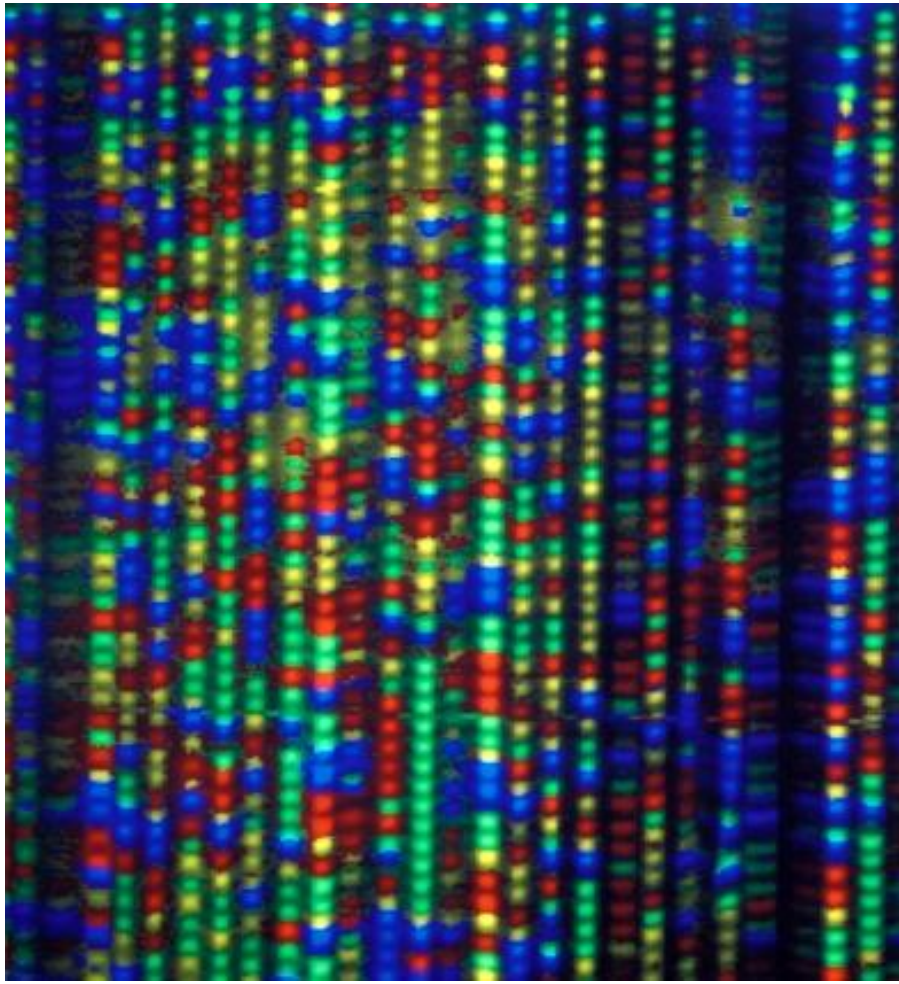




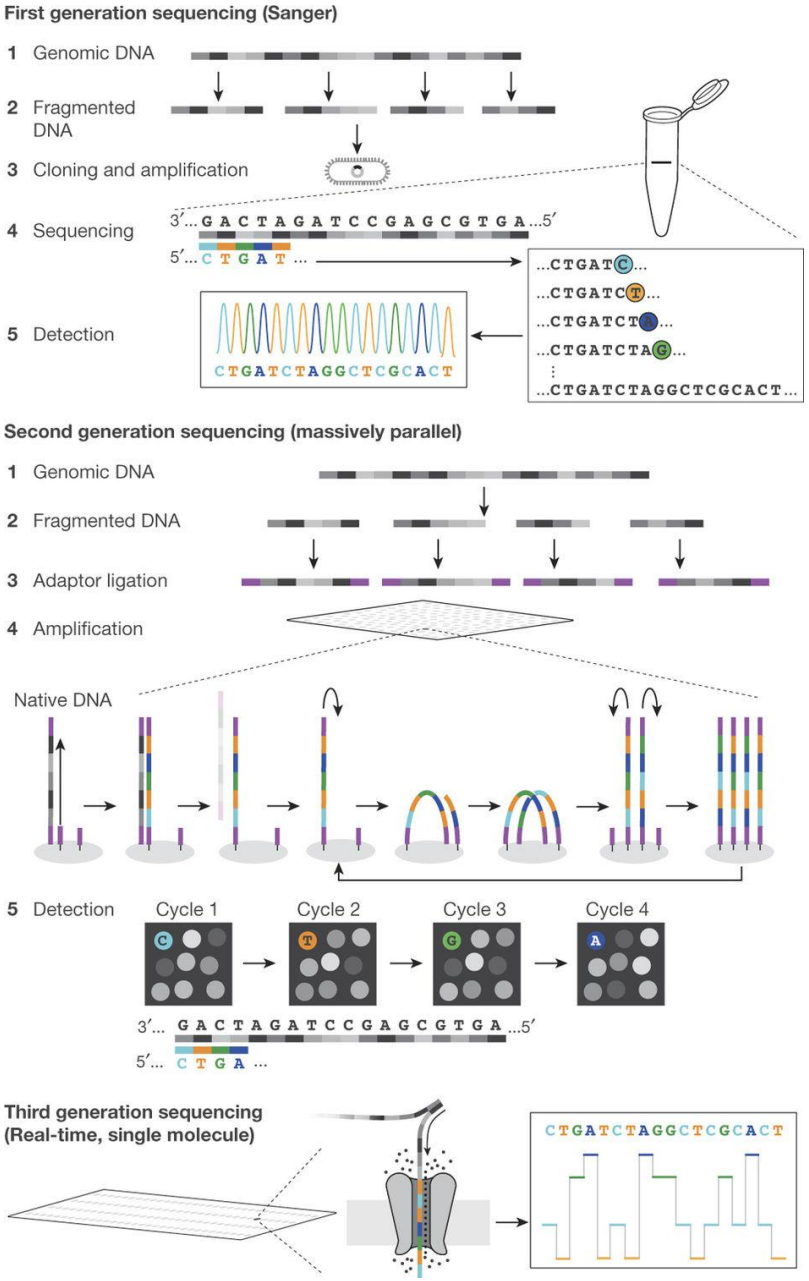


DNA sequencing and bioinforamtics

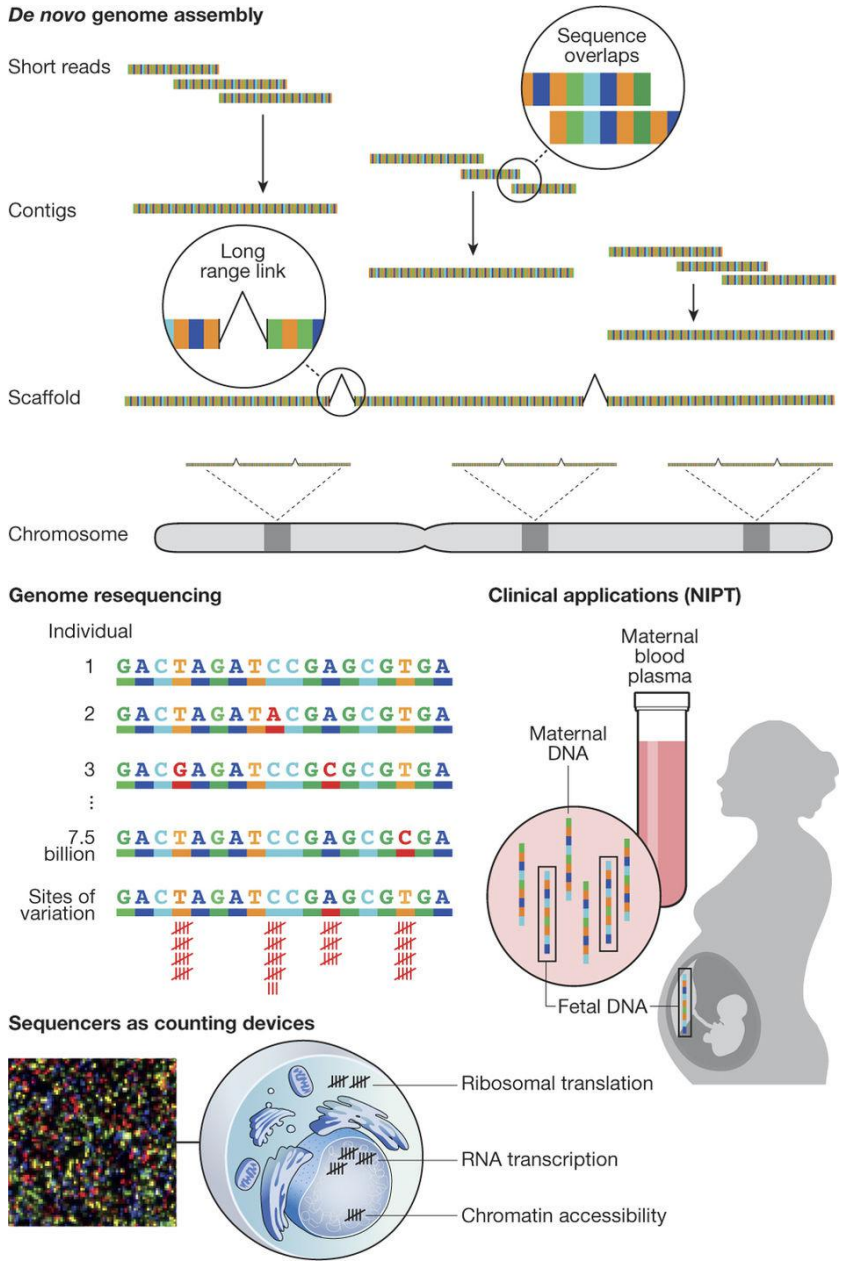




DNA Sequencing



Sequencing applications



在今天，DNA测序技术已经在诸多方面达到了临床应用的具体要求。

科学家估计，全世界每年大约有四百万到六百万孕妇正在通过外周血游离DNA对胎儿的21三体综合征进行诊断，而十年之内，这个数字将超过1500万。

在高收入国家，基因组测序已经广泛用于多种疾病的产前诊断，可以揭示大约30％的出生缺陷，同时，这一数字也正随着数据解读能力的成熟而逐渐上升。

在肿瘤学领域，液体活检在最近几年已经成为了肿瘤相关学术，产业以及投资界的新宠。基于DNA测序的液体活检被认为正逐步发展为癌症诊断与预后评估的标准方法，能够在可知的时间内逐步补充甚至取代传统的创伤性癌症诊断技术。

同样，手持DNA测序仪等设备的开发也使得流行病学家甚至能够在最为偏远的地区高效完成对人类样本，动物以及昆虫病原载体甚至是空气，水，食物的基因检测。

流行病学家和公共卫生专家也开始讨论如何通过对城市垃圾中微生物的DNA测序辅助传染性疾病的预防与控制。

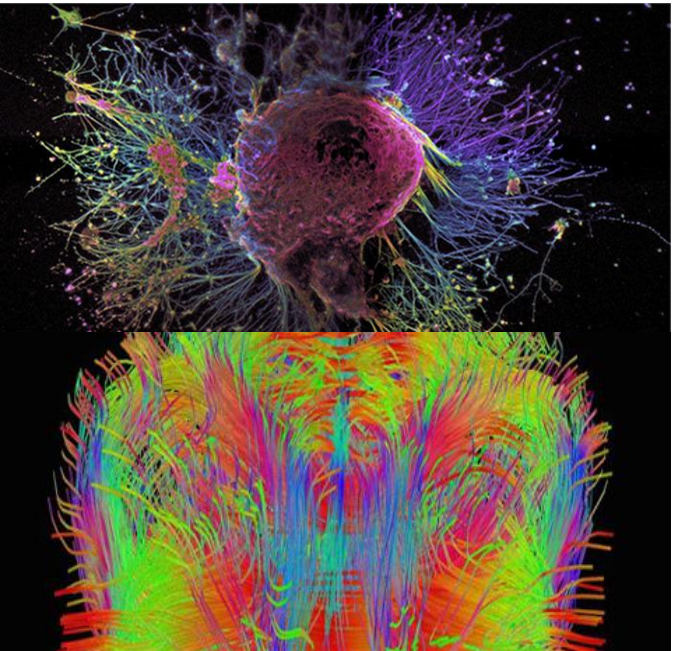
海洋生物学家也正在通过宏基因组学技术来对海洋的生态健康进行监测与研究。

在法医领域，便携式DNA测序仪可以将DNA分析带出法医实验室，使其成为一线警务工作人员的随身工具。帮助警方即使通过DNA监测确定嫌疑人，发展成为诸如酒精探测器一类的便捷工具。

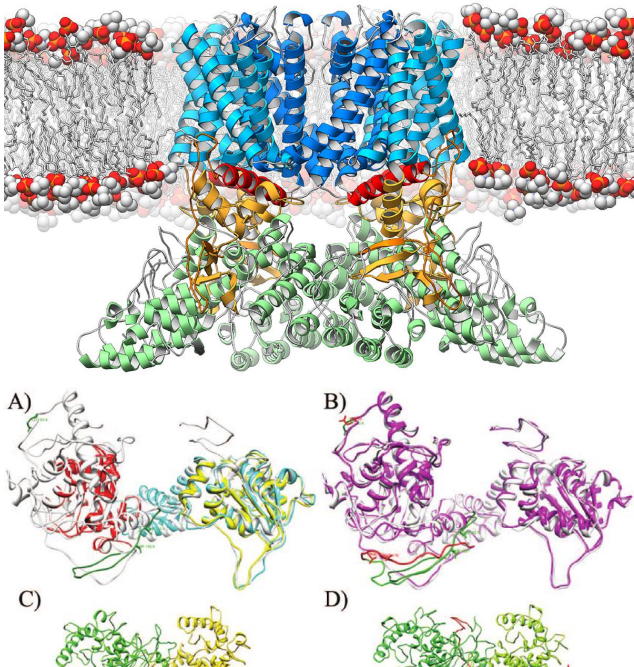
在人们的家里，DNA测序设备或许也可以成为下一个“智能”或“连接”设备，一些评论者甚至认定厕所是通过实时DNA测序监测家庭成员健康的理想场所。

生物信息学@HUST

生命不只是序列的，但是生命始终是数字的！

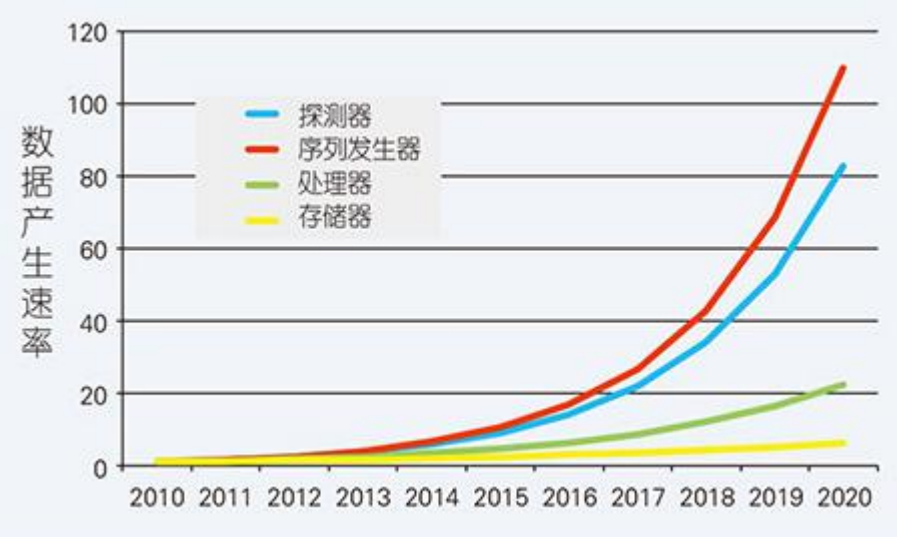


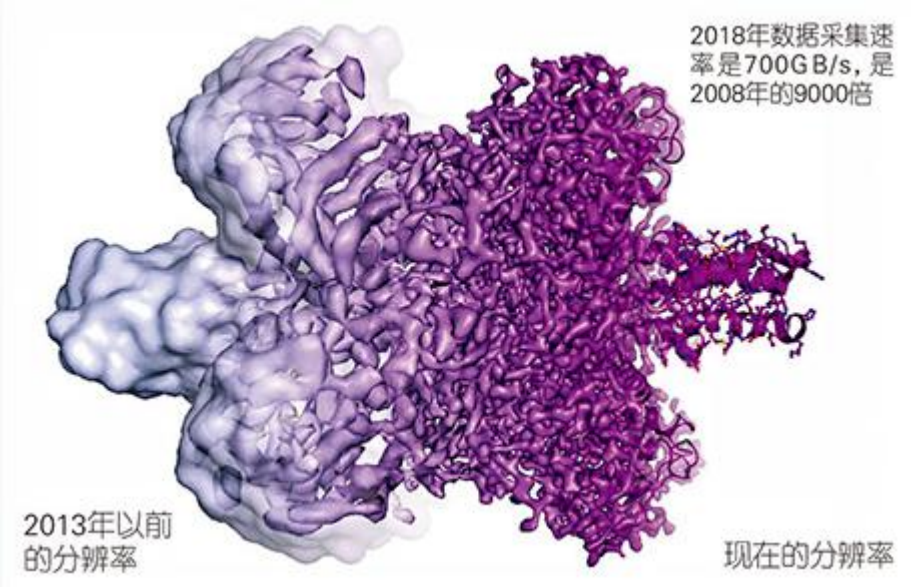
- 结构生物学（Structure biology）
- 生物图像（Bio-imaging）
- 。 。 。



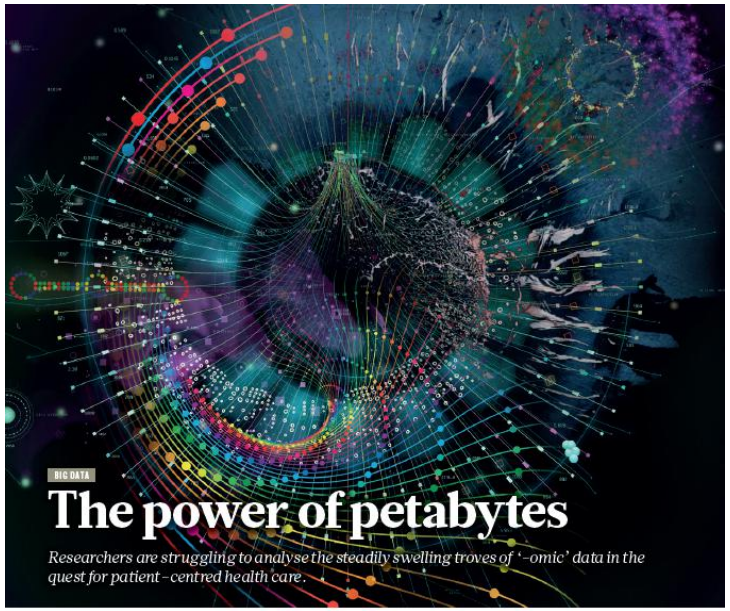
生物信息学@HUST

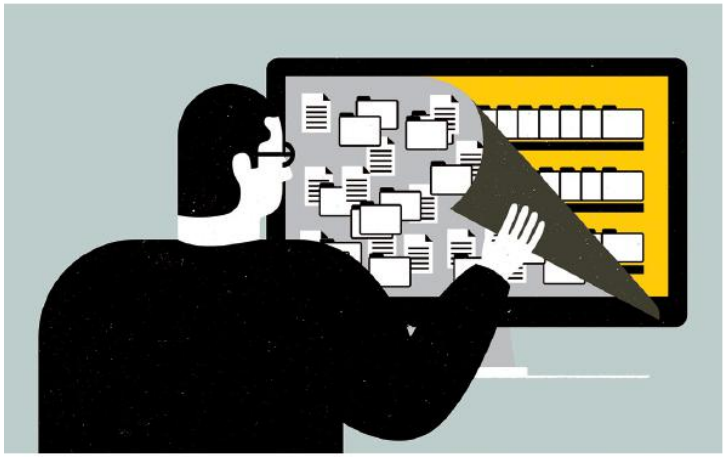
生命不只是序列的，但是生命始终是数字的！

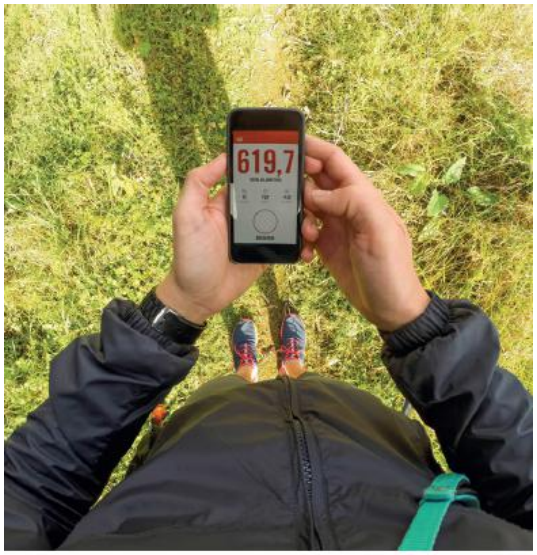




Big-data become popular…







Smartphone fitness apps enable researchers to gather health data from large numbers of people.

MOBILE DATA
Made to measure

Nature, 2015/11/05 collection on "Big-data in biomedicine"

Microbiome become popular as well…


MICROBIOME

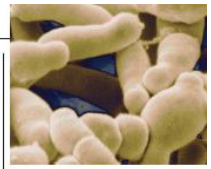
A unified initiative to harness Earth's microbiomes

Transition from description to causality and engineering


COMMENT

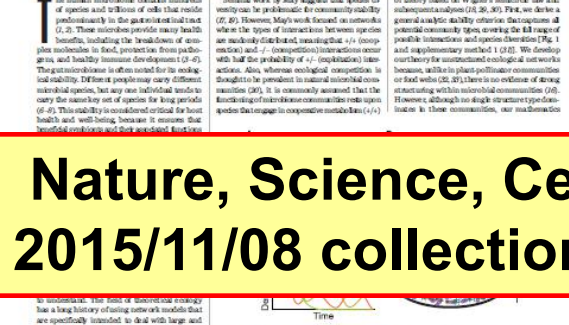
Create a global microbiome effort

Understanding how microbes affect health and the biosphere requires an international initiative, argue Nicole Dubilier, Margaret McFall-Ngai and Liping Zhao.


MICROBIOME

Microbes aid cancer drugs

Gut bacteria boost immunotherapies



MICROBIOME

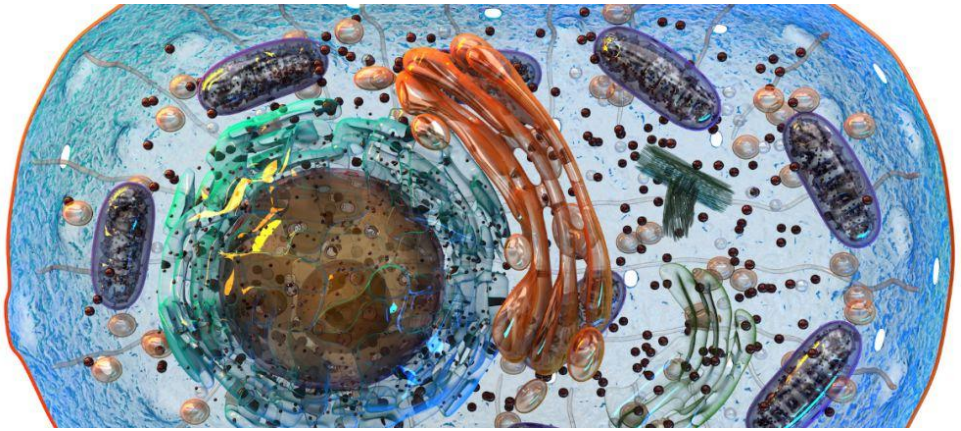
The ecology of the microbiome: Networks, competition, and stability

Nature, Science, Cell, etc.

2015/11/02- 2015/11/08 collection on "Microbiome"

Microbiome and big-data…

- A cell is already very complex
- 一个细胞已经非常复杂了

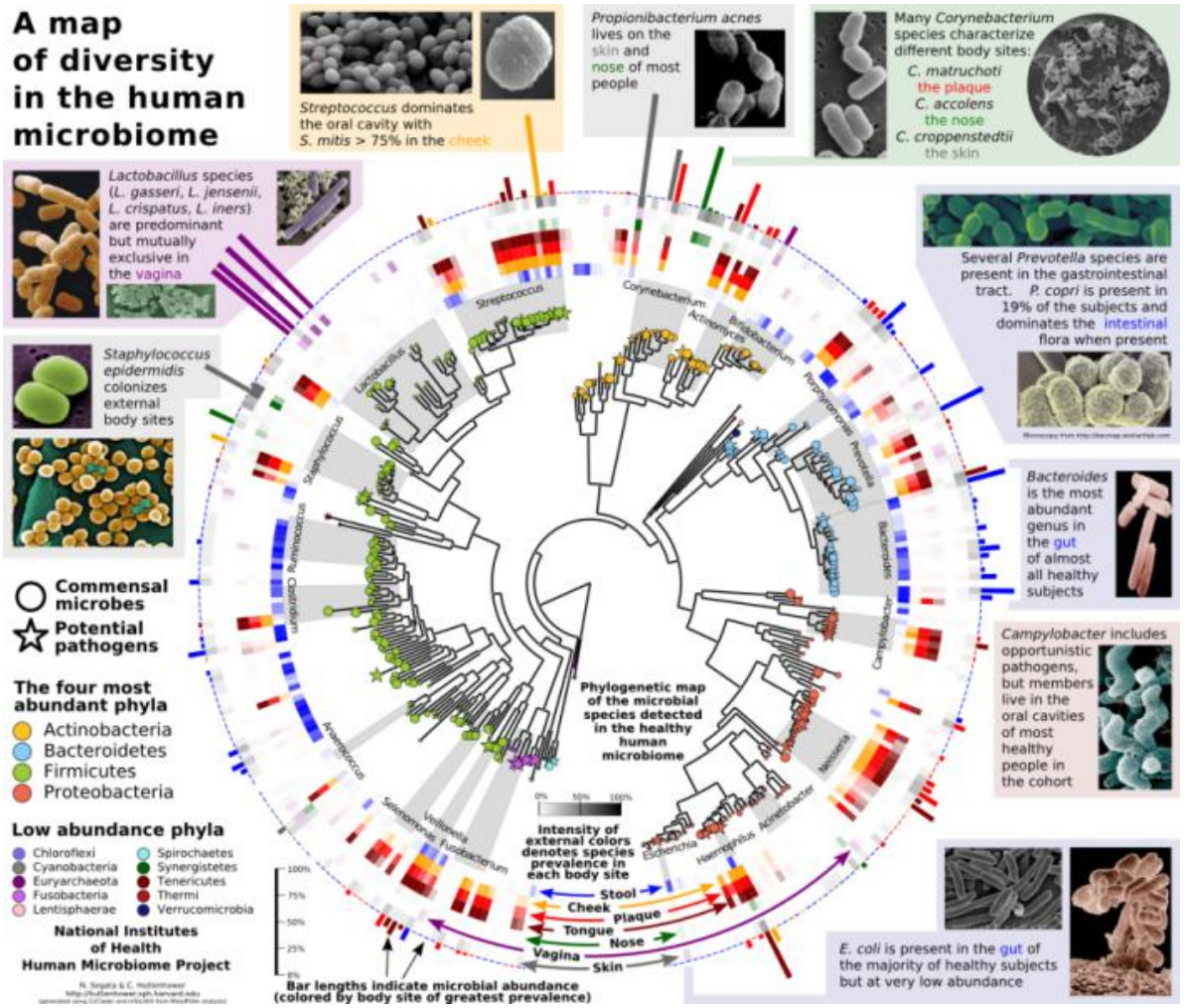


- A microbial community is much more complex than a cell
- 一个微生物群落就更为复杂了

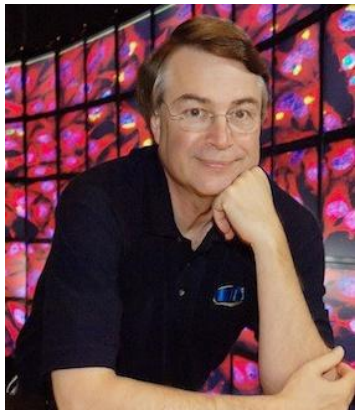
- But much more big-data
- 但是也代表了更多的数据

Microbiome and big-data…

在生物信息眼里，这全是大数据。。。



Microbiome and big-data…



Larry Smarr

Founding Director of the California Institute for Telecommunications and Information Technology (Calit2)

PUBLICATIONS

LARRY'S LATEST PAPERS

Large Memory High Performance Computing Enables Comparison Across Human Gut Microbiome Of Patients With Autoimmune Diseases And Healthy Subjects
Published in the XSEDE 2013 Proceedings of the Conference on Extreme Science and Engineering Discovery Environment: Gateway to Discovery. Article No. 25 (http://dl.acm.org/citation.cfm?doid=2484762.2484828)

Quantifying Your Body: A How-to Guide From A Systems Biology Perspective
Larry Smarr, Biotechnol. J. 2012, 7, 980-991

Supporting Information For Quantifying Your Body: A How-to Guide From A Systems Biology Perspective
Supporting Information for DOI 10.1002/biot.201100495

Essay: An Evolution Toward A Programmable Universe
Larry Smarr, Dec 5, 2011, The New York Times

Quantified Health: A 10-year Detective Story Of Digitally Enabled Genomic Medicine
Larry Smarr, with commentary by Mark Anderson, published as a Special Letter in the Strategic News Service Newsletter, September 30, 2011.

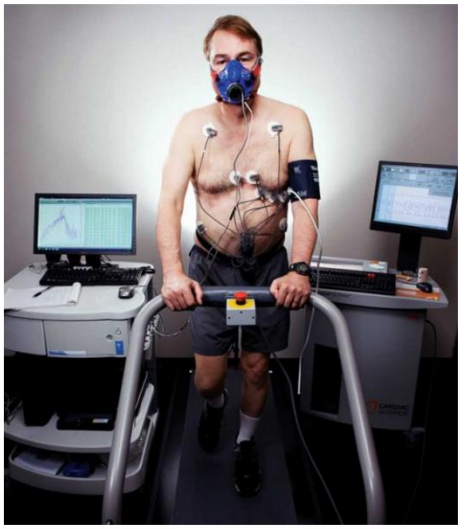
How I Improved My Health By Changing My Eating, Exercise, And Stress Management Habits: An Annotated Reading List
Larry Smarr, Requested by Mark Anderson, CEO Strategic News Service For Distribution to the Future in Review 2011 Attendees

Biomedicine

The Patient of the Future

Internet pioneer Larry Smarr's quest to quantify everything about his health led him to a startling discovery, an unusual partnership with his doctor, and more control over his life.

by Jon Cohen February 21, 2012

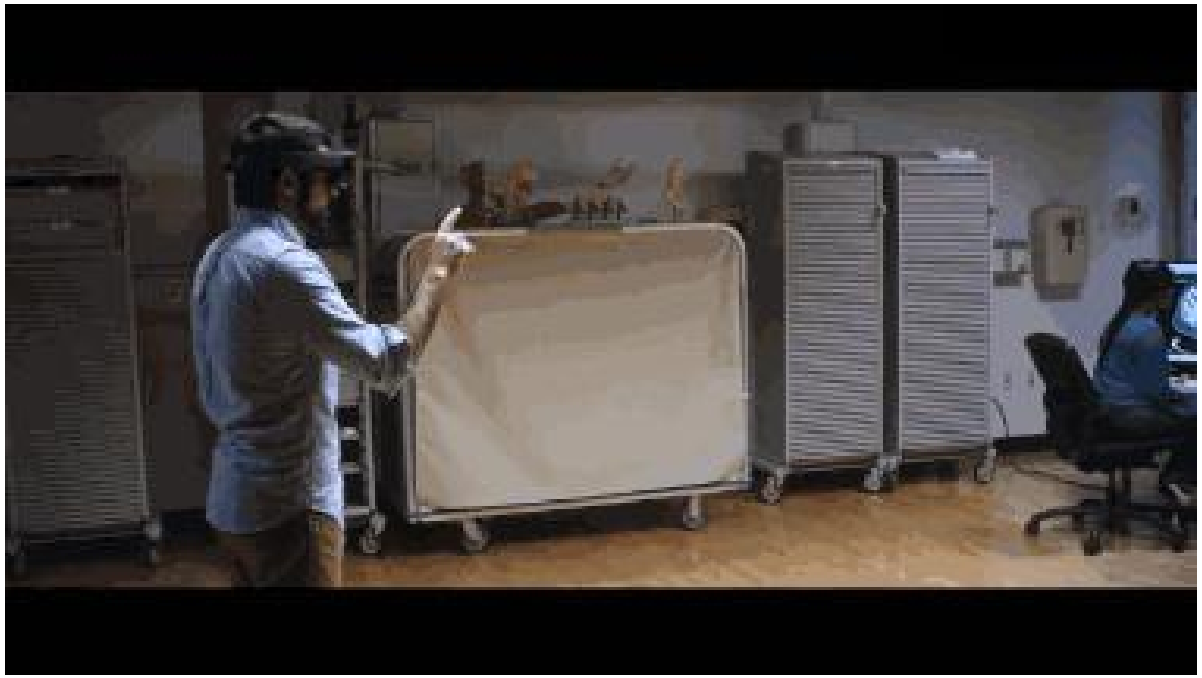


TEDMED Attend Speakers TEDMED Live Talks The Hive Partnerships About Blog

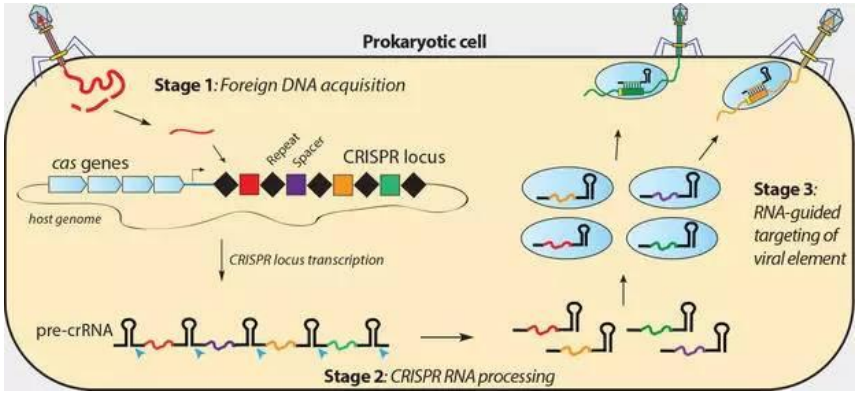
Biomedical big-data…

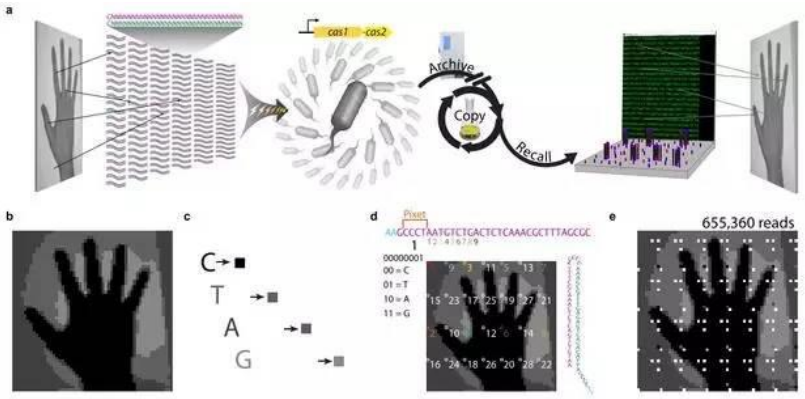
"Have you ever figured how information-rich your stool is?," Larry asks me with a wide smile, his gray-green eyes intent behind rimless glasses. "There are about 100 billion bacteria per gram. Each bacterium has DNA whose length is typically one to 10 megabases—call it 1 million bytes of information. This means human stool has a data capacity of 100,000 terabytes of information stored per gram. That's many orders of magnitude more information density than, say, in a chip in your smartphone or your personal computer. So your stool is far more interesting than a computer."

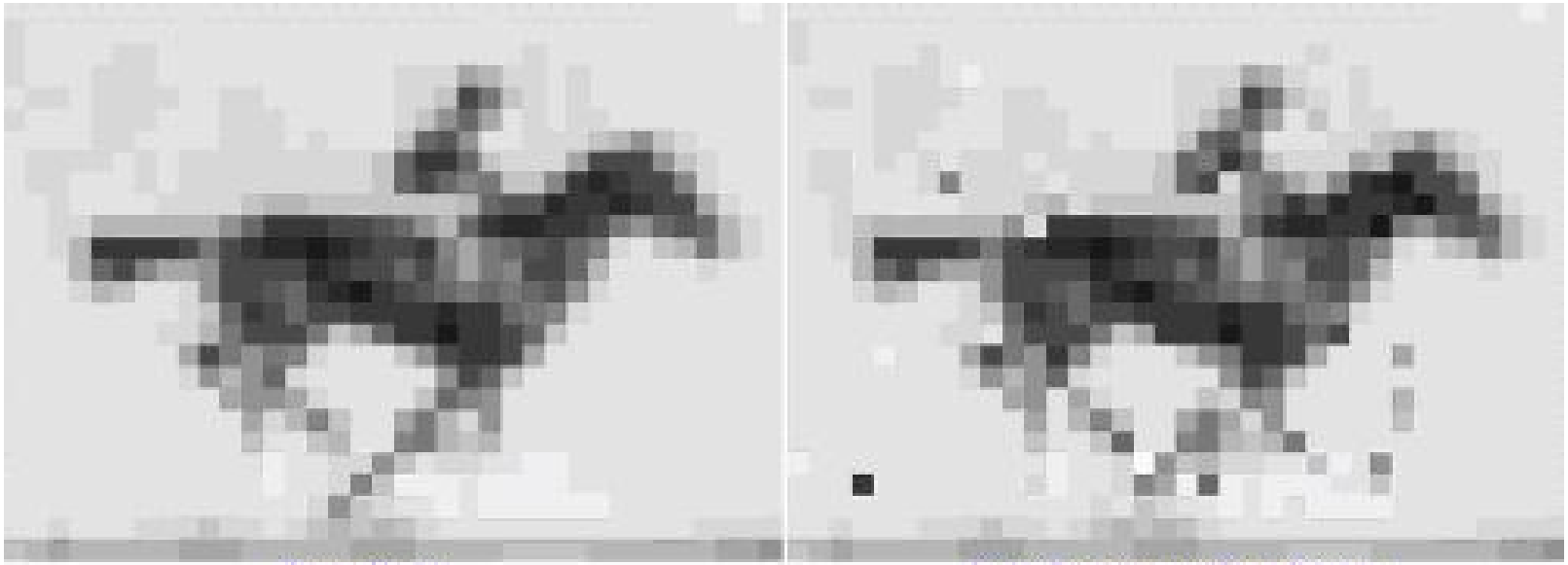
-- Larry Smarr

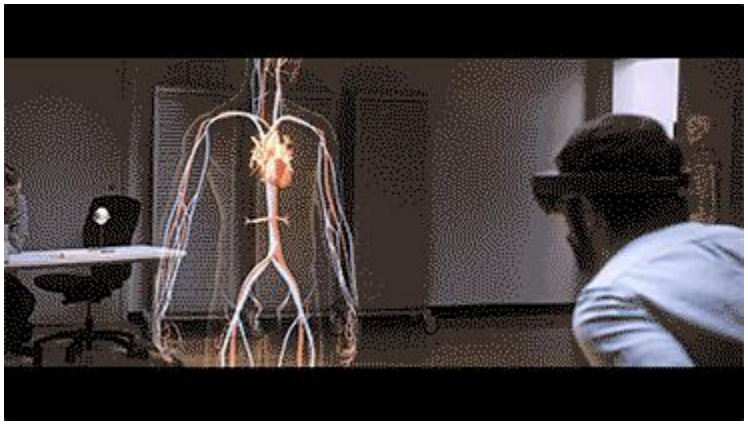


Understand it, create it!

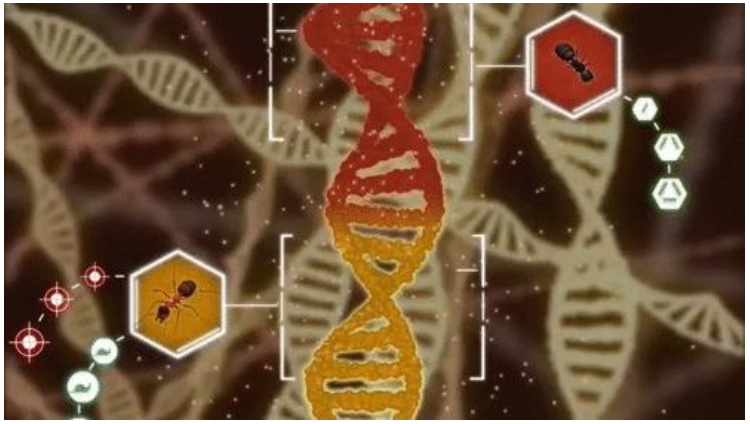




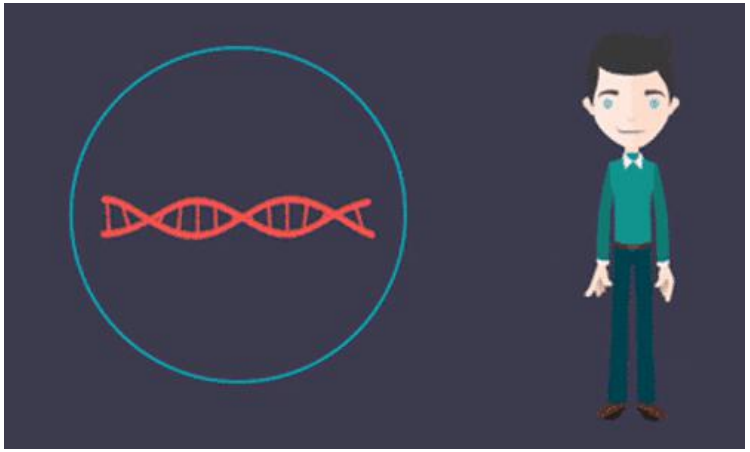




See it!



Understand it!



Create it!

生物信息学：计算科学视角

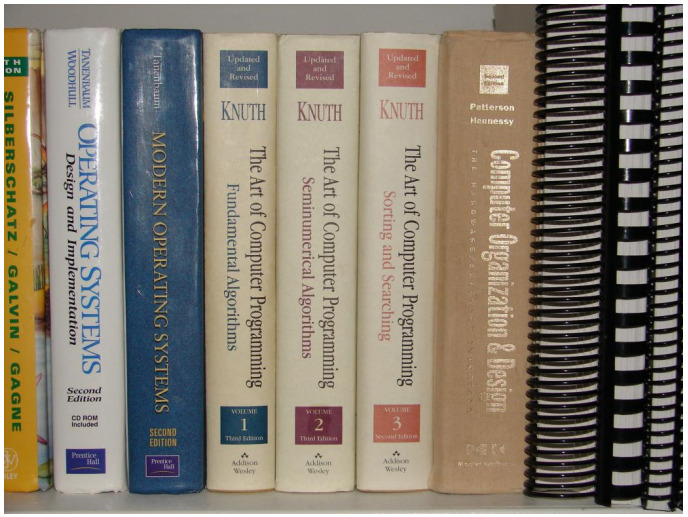
Donald Knuth（高德纳）

Donald Knuth, the "father of the analysis of algorithms.”

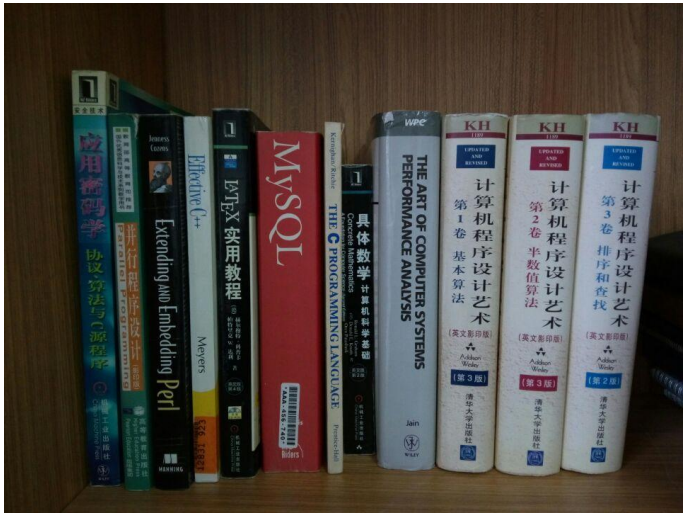
The Art of Computer Programming （计算机程序设计艺术）

Markup	Renders as
`The quadratic formula is $-b \pm \sqrt{b^2 - 4ac} \over 2a$ \bye`	The quadratic formula is $\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

“生物信息学为算法研究提供了500年的问题”– Don Knuth

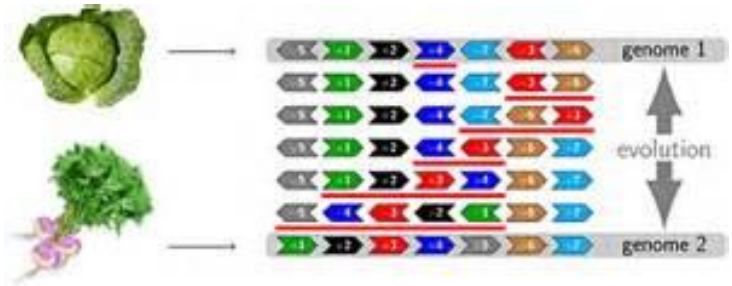


"definitely send me a résumé if you finish this fiendishly difficult book" – Bill Gates



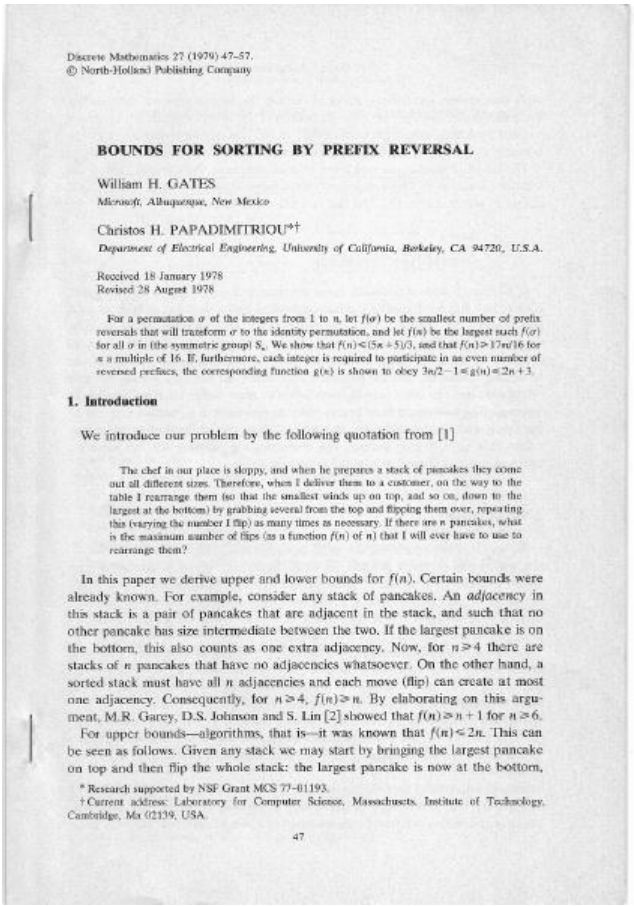
"definitely come to talk about algorithm if you read half of this book" – Kang Ning

Bill Gates（比尔盖茨）



比尔盖茨:下个世界首富出自基因检测领域

Sorting by reversal problem



Discrete Mathematics 27 (1979) 47–57.
© North-Holland Publishing Company

BOUNDS FOR SORTING BY PREFIX REVERSAL

William H. GATES
Microsoft, Albuquerque, New Mexico

Christos H. PAPADIMITRIOU*†
Department of Electrical Engineering, University of California, Berkeley, CA 94720, U.S.A.

Received 18 January 1978
Revised 28 August 1978

For a permutation σ of the integers from 1 to n, let $f(\sigma)$ be the smallest number of prefix reversals that will transform σ to the identity permutation, and let $f(n)$ be the largest such $f(\sigma)$ for all σ in (the symmetric group) S_n. We show that $f(n) \leqslant (5n+5)/3$, and that $f(n) \geqslant 17n/16$ for n a multiple of 16. If, furthermore, each integer is required to participate in an even number of reversed prefixes, the corresponding function $g(n)$ is shown to obey $3n/2 - 1 \leqslant g(n) \leqslant 2n+3$.

1. Introduction

We introduce our problem by the following quotation from [1]

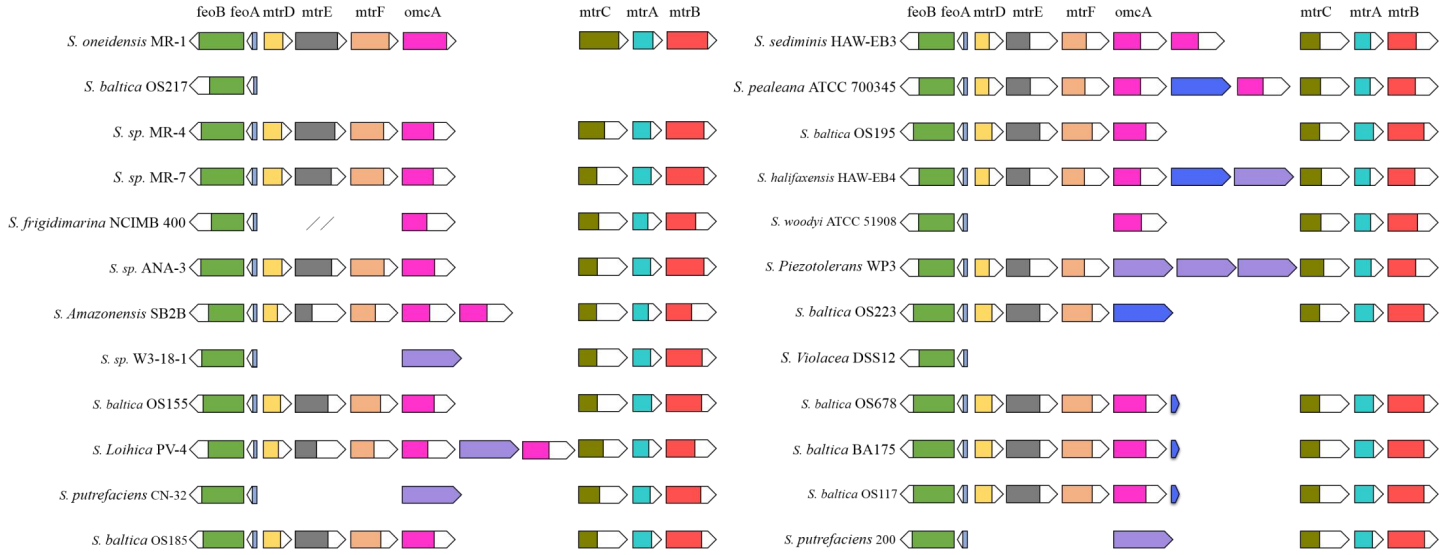
> The chef in our place is sloppy, and when he prepares a stack of pancakes they come out all different sizes. Therefore, when I deliver them to a customer, on the way to the table I rearrange them (so that the smallest winds up on top, and so on, down to the largest at the bottom) by grabbing several from the top and flipping them over, repeating this (varying the number I flip) as many times as necessary. If there are n pancakes, what is the maximum number of flips (as a function $f(n)$ of n) that I will ever have to use to rearrange them?

In this paper we derive upper and lower bounds for $f(n)$. Certain bounds were already known. For example, consider any stack of pancakes. An *adjacency* in this stack is a pair of pancakes that are adjacent in the stack, and such that no other pancake has size intermediate between the two. If the largest pancake is on the bottom, this also counts as one extra adjacency. Now, for $n \geqslant 4$ there are stacks of n pancakes that have no adjacencies whatsoever. On the other hand, a sorted stack must have all n adjacencies and each move (flip) can create at most one adjacency. Consequently, for $n \geqslant 4$, $f(n) \geqslant n$. By elaborating on this argument, M.R. Garey, D.S. Johnson and S. Lin [2] showed that $f(n) \geqslant n+1$ for $n \geqslant 6$.

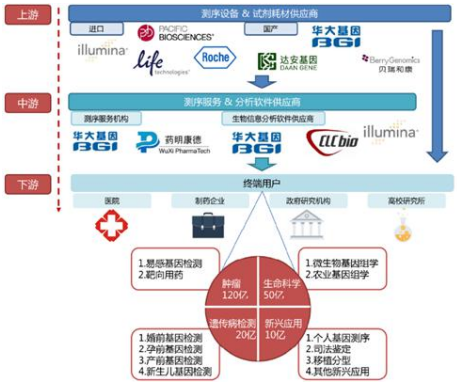
For upper bounds—algorithms, that is—it was known that $f(n) \leqslant 2n$. This can be seen as follows. Given any stack we may start by bringing the largest pancake on top and then flip the whole stack: the largest pancake is now at the bottom,

* Research supported by NSF Grant MCS 77-01193.
† Current address: Laboratory for Computer Science, Massachusetts Institute of Technology, Cambridge, Ma 02139, USA.

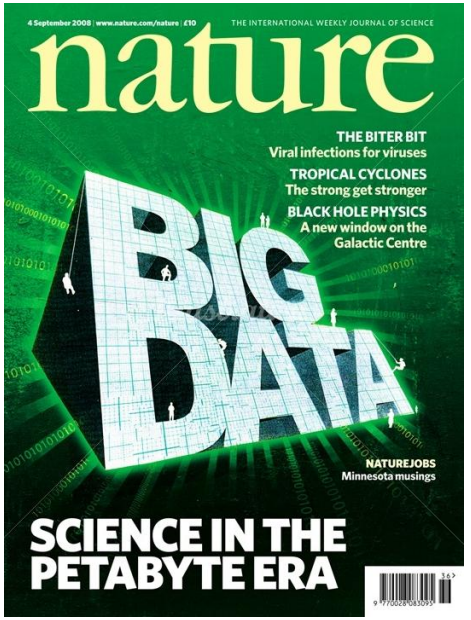
How many reversal steps for this REAL case?



Current status（现今态势）



很难找到
与生物信息学和生物统计学
没有关系的
生物学与生物工程
研究和应用领域了。。。



Alphabet （谷歌）

Google 的基因组学梦想

The NEW ENGLAND JOURNAL of MEDICINE

HOME | ARTICLES & MULTIMEDIA | ISSUES | SPECIALTIES & TOPICS | FOR AUTHORS

CORRESPONDENCE

23andMe and the FDA

N Engl J Med 2014; 370:2248-2249 | June 5, 2014 | DOI: 10.1056/NEJMc1404692

Share:

Article | Citing Articles (3) | Metrics

To the Editor:

In their Perspective article (March 13 issue),[1] Annas and Elias state that the conflict between the genetic-testing company 23andMe and the Food and Drug Administration (FDA) concerns analytic and clinical validity, clinical utility, and ethical, legal, and social issues. However, their discussion is limited to a domestic U.S. perspective. After a person's raw genetic data have been determined from a DNA sample, the data are stored remotely and can be accessed easily anywhere in the world. For example, in Japan, maternal blood samples from Japanese mothers undergoing

nature biotechnology

Home | Current issue | News & comment | Research | Archive | Authors & referees | About the journal

home ▸ archive ▸ issue ▸ news ▸ full text

NATURE BIOTECHNOLOGY | NEWS

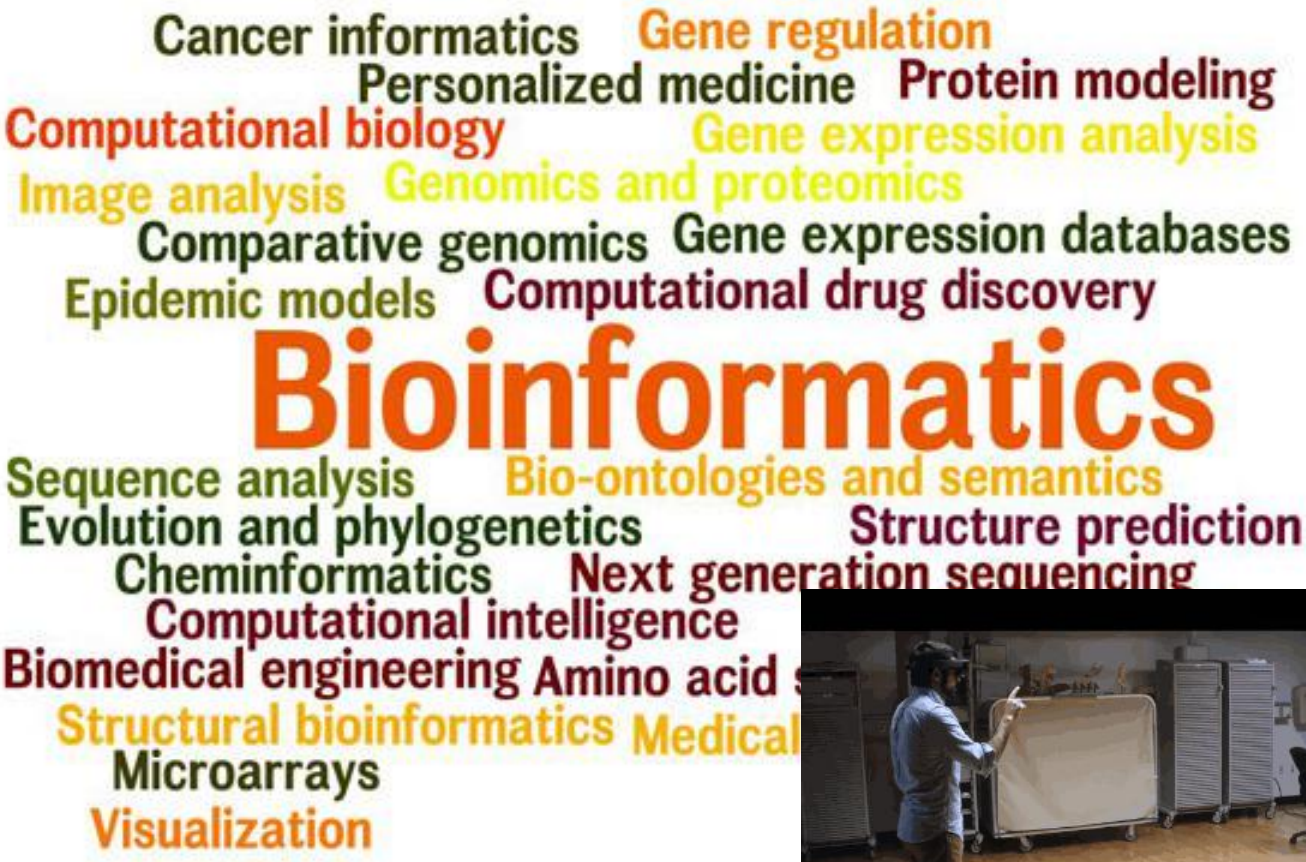
FDA approves 23andMe gene carrier test

Nature Biotechnology **33**, 435 (2015) | doi:10.1038/nbt0515-435a
Published online 12 May 2015

PDF | Citation | Reprints | Rights & permissions | Article metrics

23andMe, based in Mountain View, California, has received word from the US Food and Drug Administration (FDA) that their Bloom syndrome carrier screening test was approved as a class II device. The approval came in February, 15 months after the personal genomics company received a cease and desist letter from the regulator for its genetic tests because the company was dispensing health-related information to consumers without having obtained marketing clearance. The FDA website lists class II devices as moderate risk, requiring some regulatory controls, putting carrier screening tests in the same category as condoms. This turnaround follows the company's

Future（未来）


Cancer informatics
Gene regulation
Personalized medicine
Protein modeling
Computational biology
Gene expression analysis
Image analysis
Genomics and proteomics
Comparative genomics
Gene expression databases
Epidemic models
Computational drug discovery
Bioinformatics
Sequence analysis
Bio-ontologies and semantics
Evolution and phylogenetics
Structure prediction
Cheminformatics
Next generation sequencing
Computational intelligence
Biomedical engineering
Amino acid
Structural bioinformatics
Medical
Microarrays
Visualization

生物信息学

- 生物信息学(Bioinformatics)是研究生物信息的采集、处理、存储、传播，分析和解释等各方面的学科，也是随着生命科学和计算机科学的迅猛发展，生命科学和计算机科学相结合形成的一门新学科。
- 生物信息学通过综合利用生物学，计算机科学和信息技术而揭示大量而复杂的生物数据所赋有的生物学奥秘。

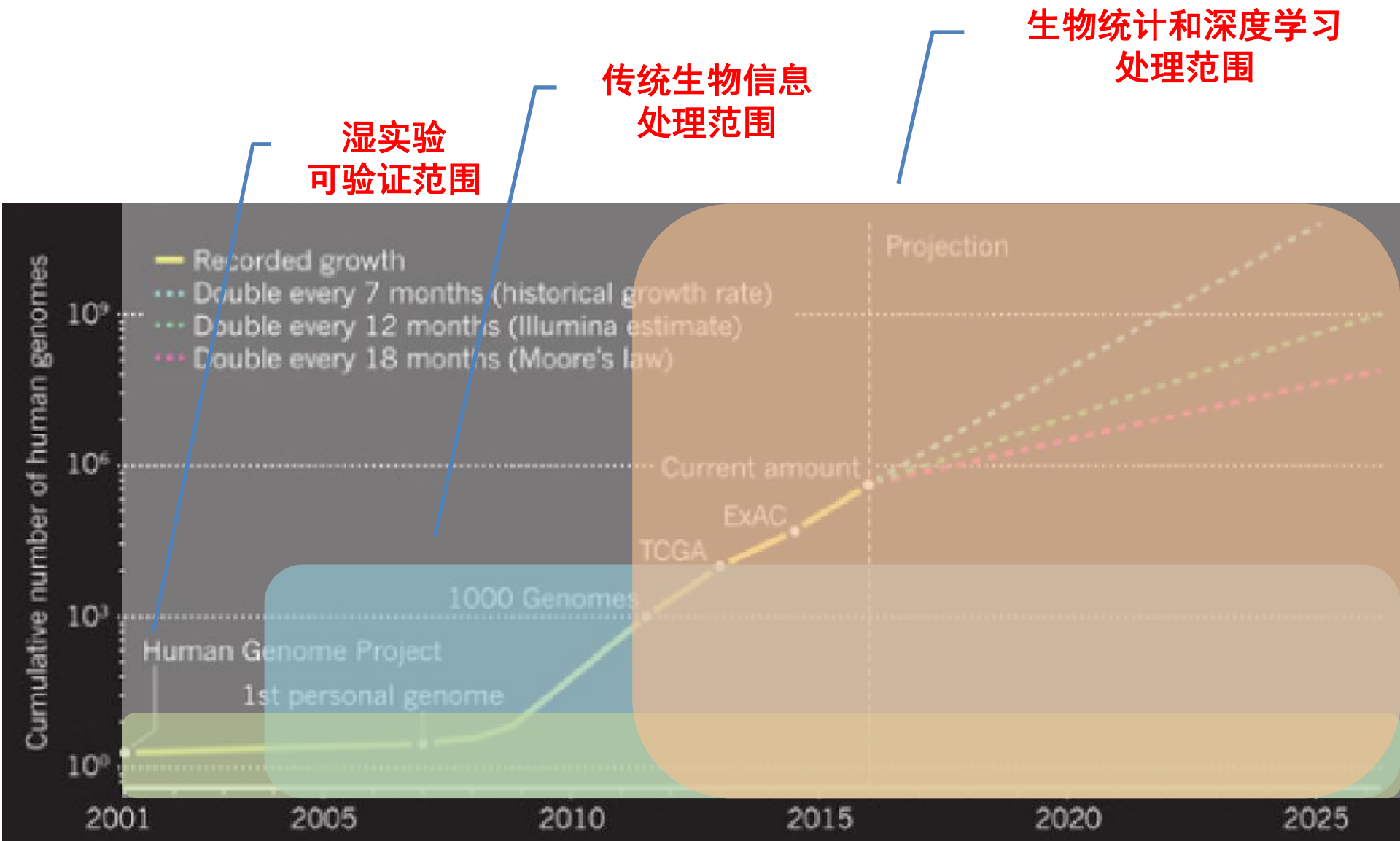
为什么要学习生物信息学

- 组学大数据的现状：4V， 4H
- 算法和数据库的需求：数据挖掘
- 生物信息学的思维：数据驱动的方法学研究
- 生物信息学的应用：几乎无限的需求

为什么要学习生物信息学

- 必须利用生物信息学才能回答的问题
  - 疾病已经进入哪个阶段了？
  - 哪些基因在疾病发生发展中起到关键作用？
  - 基因和环境是否有关？
  - 新药物是否更有效？
  - 遗传与环境哪个更重要？
  - ……

为什么要学习生物统计学



生物信息学研究的三个层面

初级层面
中级层面
高级层面

初级层面

基于现有的生物信息数据库和资源，利用成熟的生物信息学工具（专业网站、软件）解决生物信息学问题

——生物信息数据库（NCBI、EBI等）

——基因组序列分析、序列比对软件（GCG、BLAST、CLUSTAL等）

——系统发育树构造软件（PHYLIP、PALM、MEGA等）

——分子动力学模拟软件（GROMACS、NAMD等）

——搜集、整理有特色的生物信息学数据集

中级层面

利用数值计算方法、数理统计方法和相关的工具，研究生物信息学问题

——概率、数理统计基础

——科学计算基础

——现有的数理统计和科学计算工具（EXCEL、SPSS、SAS、MATLAB等）

——建立有特色的生物信息学数据库

高级层面

提出有重要意义的生物信息学问题；自主创新，发展新型方法，开发新型工具，引领生物信息学领域研究方向。

——面向生物学领域，解决生物学问题

——数学、物理、化学、计算科学等思想和方法

——建立模型，发展算法

——自行编程，开发软件，建立网页（Linux系统、C/C++、PERL、数据库技术）

从事生物信息学研究应具备多方面的科学基础：

（1）、一定的计算能力，包括相应的软、硬设备。要有各种数据库或者能与国际、国内的数据库系统进行有效的交流。要有发达、稳定的互联网络系统；
（2）、强有力的创新算法和软件。没有算法创新，生物信息学就无法获得持续的发展；
（3）、与实验科学，特别是与自动化的大规模高通量的生物学研究方法与平台技术建立广泛、紧密的联系。这些技术，既是产生生物信息数据的主要方法，又是验证生物信息学研究结果的关键手段。

从事生物信息学研究的人员必须具备多学科交叉的知识。

生物信息学的“降龙十八掌”

（1）

要掌握生物信息数据库及其查询搜索方法

(Database & searching)

——对分子生物信息数据库的种类以及某些具体数据库的掌握和了解

——从现有数据库中熟练获得需要的数据信息（尤其是二级数据库）

——能熟练地进行数据库查询和数据库搜索（数据库查询系统Entrez、SRS；搜索工具BLAST等）

——数据库技术、互联网技术

（2）

要学会生物信息学软件和工具的应用

(Software & application)

利用成熟的生物信息学工具（专业网站、软件）解决生物信息学问题

——基因组序列分析、序列比对软件（GCG、BLAST、CLUSTAL等）

——系统发育树构造软件（PHYLIP、PALM等……）

——基因芯片检测分析软件（商业软件ScanArray、Array-Pro等 ……）

——分子动力学模拟软件（GROMACS、NAMD等……）

（3）

概率论基础

(Probability theory)

——随机事件、概率

——随机变量、概率分布

——大数定律、中心极限定理

——几乎用于生物信息学的各个方面

“Most of the problems in computational sequence analysis are essentially statistical.”

——“Biological sequence analysis”

（4）

数理统计基础

(Statistical methods)

——样本和统计量（方差、均值……）

——参数估计、假设检验

——基本的统计分析（方差分析、协方差分析、回归分析）

——常用统计软件的运用（SPSS、SAS）

——几乎用于生物信息学的各个方面

（5）

基于频率的组分分析方法和权重矩阵方法

(Composition analysis & weight matrix method)

——符号（如碱基）频率反映具有生物学意义的序列特征，如内含子剪接位点的发现，KOZAK规则的发现等

——核酸组分、氨基酸组分、密码子使用频率

——主要用于具有特定生物学意义的序列特征的分析

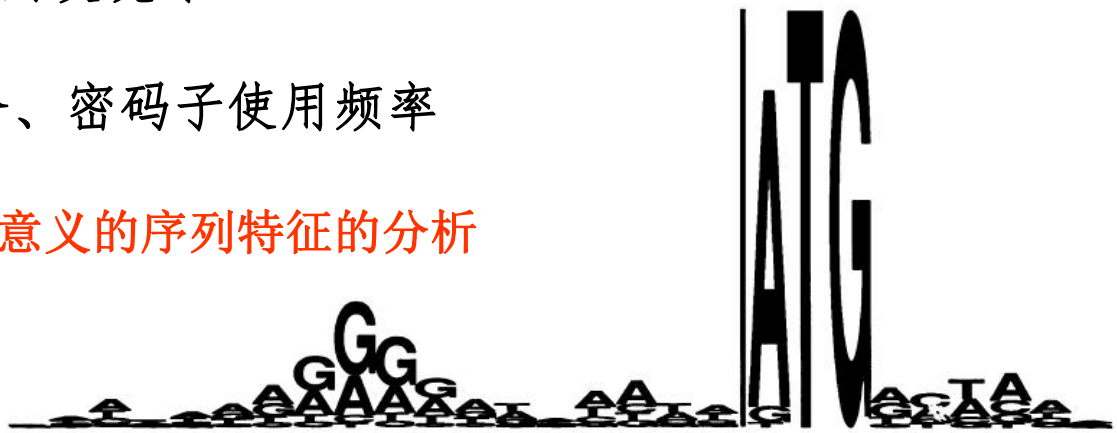


Figure 1. Logo for *E. coli* ribosome binding sites. Only −18 to +8 of the −20 to +13 site is shown. The first translated codon is just to the right of the 2 bit high vertical bar. 149 natural sites were used to create the logo (9).

权重矩阵分析方法举例

——针对序列信号（一段核酸、蛋白），计算每一位点所使用的词汇或叫符号（*碱基、氨基酸*）频率，频率的偏好性反映信号的序列特征（**sequence pattern**）。

例：人类基因内含子/外显子剪接位点的序列特征分析

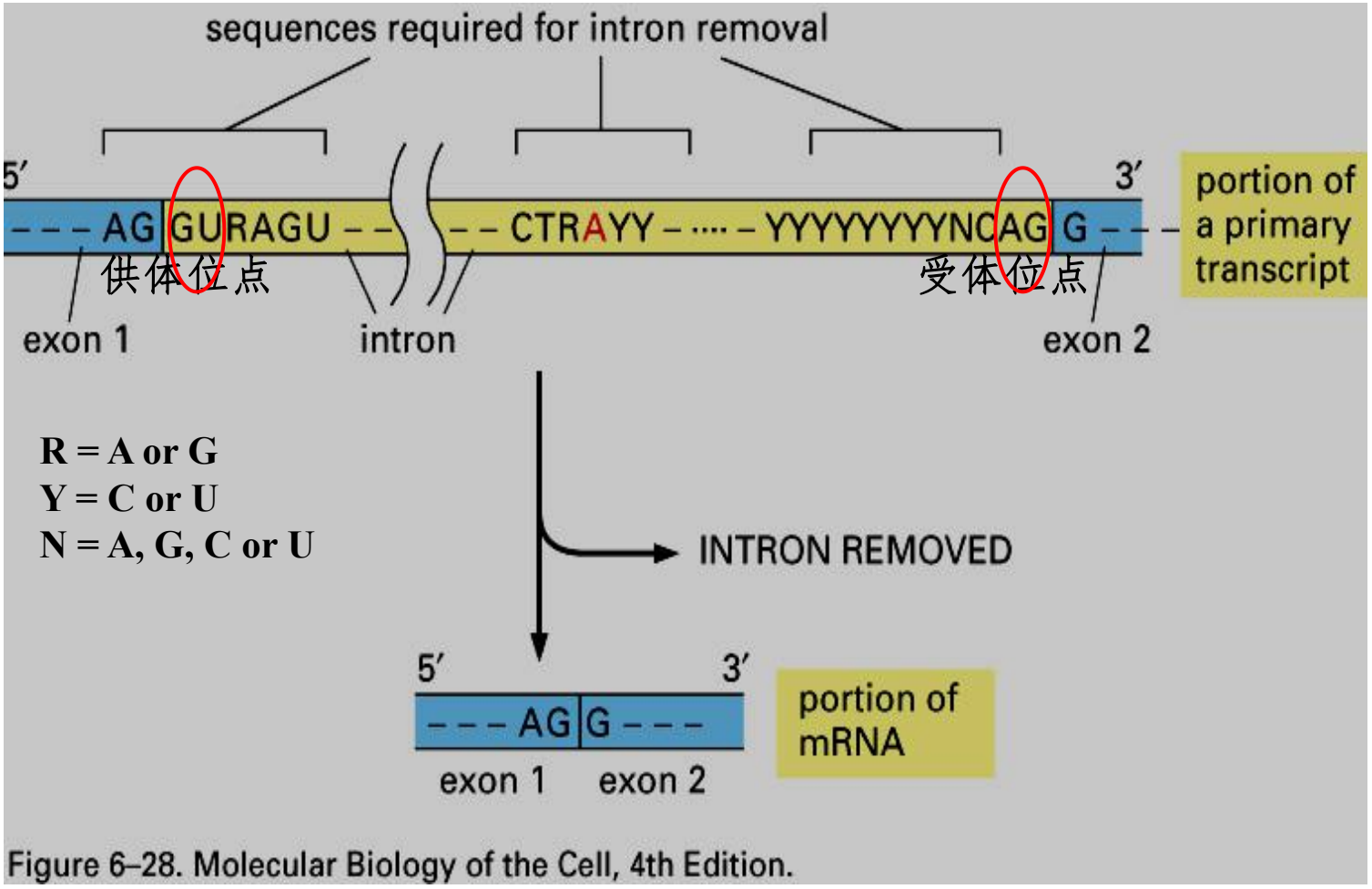


Figure 6–28. Molecular Biology of the Cell, 4th Edition.

Bayesian打分函数用于剪接位点预测的公式

The likelihood that a property value *v* (of a new structure) is drawn from the splicing site is:

$$P(site \mid \upsilon) = \frac{P(\upsilon \mid site)P(site)}{P(\upsilon \mid site)P(site) + P(\upsilon \mid nonsite\ )P(nonsite\ )'}$$

Score for the overall likelihood of the query sequence being a site is:

$$\sum_{\substack{\text{properties at} \\ \text{associated volumes}}} \log\left(\frac{P(site \mid \upsilon)}{P(site)}\right)$$

Say we have a sequence S= $S_1S_2\ldots S_n$. Then one need to calculate

$$\frac{P(S|\text{splice site})}{P(S|\text{background})}$$

So to look for a donor site in the sequence, we might calculate

（6）

信息论方法

(Information method)

——信息的度量：是信息符号出现何种状态的一种不确定性程度，信息的获得要对不确定性进行否定。

——生物信息的符号如ACGT四种符号，状态空间即其所有可能的排列

——用于结构预测

——信息熵 $$H = -\sum_{i} p_i \log p_i$$

——信息熵H刻画了由$\{p_i\}$表示的随机试验结果的先验不确定性，或观察到输出时所获得的信息量。

（7）

期望最大化（EM）方法

(Expectation Maximization)

——EM算法是一种在不完全数据情况下计算极大似然估计或者后验分布的迭代算法。

——适用于具有隐变量的模型和问题，

——用于结构的识别，如Motif识别的MEME方法、HMM中的Baum-Welch算法

（8）

动态规划方法

(Dynamic Programming)

——一种常用的多阶段决策的寻优算法
——动态规划用得最多的方面是DNA序列或者蛋白质序列比对

（9）

迭代方法

(Iteration)

——迭代的目的通常是在状态空间找到目标函数收敛的稳定解

——在运用模式识别方法时，对系统参数的学习通常要经过迭代来实现

——迭代必须能够不断逼近稳定解

——用于上述某些方法的方法

（10）

回归、拟合、相关性分析、关联分析

(Regression, fitting, correlation & association)

——经典的统计分析方法
——主要目的：描述和预测自变量与因变量间的关系
——用于上述某些方法的方法

（11）

判别分析方法

(Discriminant analysis)

——用于判别样品所属类型的统计分析方法

条件：已知研究对象总体的类别数目及其特征（如：分布规律，或各类的训练样本）

目的：判断未知类别的样本的归属类别

——用于基因识别、医学诊断、人类考古学

（12）

聚类分析方法

(Clustering method)

——聚类分析（群分析）是实用多元统计分析的一个新分支，正处于发展阶段。理论上尚未完善，但应用十分广泛。实质上是一种分类问题，目的是建立一种分类方法，将一批数据按照特征的亲疏、相似程度进行分类。

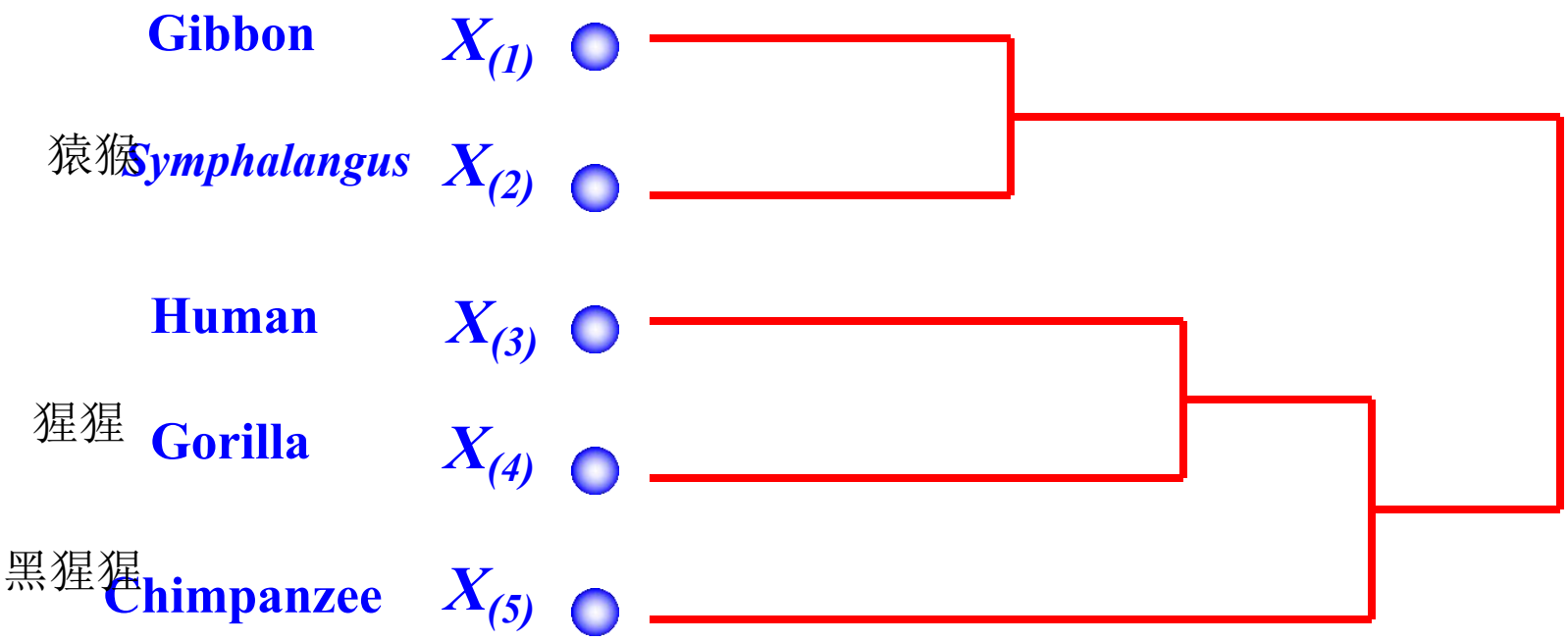
——条件：研究对象总体的类别数目未知，也不知总体样本的具体分类情况

——目的：通过分析，选定描述个体相似程度的统计量、确定总体分类数目、建立分类方法；对研究对象给出合理的分类。(“物以类聚”是聚类分析的基本出发点)

——定性、经验的分类的局限
分类较粗、数据量小、凭借经验

——谱系聚类法（系统聚类法）、动态聚类法、模糊聚类法

——生物信息学中的聚类分析问题：
根据DNA芯片获得的基因表达数据进行基因聚类（数据量庞大）
蛋白质相互作用网络的分类
根据不同物种的大分子序列进行相似性比较并构建系统发育树



（13）

Markov模型的应用

(Markov model)

——Markov过程：从一种状态转移到另一种状态时，过程仅取决于前面n种状态，是一种有序n模型。 n是影响下一个状态选择的状态数。

——最简单的Markov过程是一阶过程，状态的选择完全取决于前一状态，这种选择是依照概率来选择的。

——状态的选择是概率的，而非确定的。故Markov过程本质上是一种随机过程。

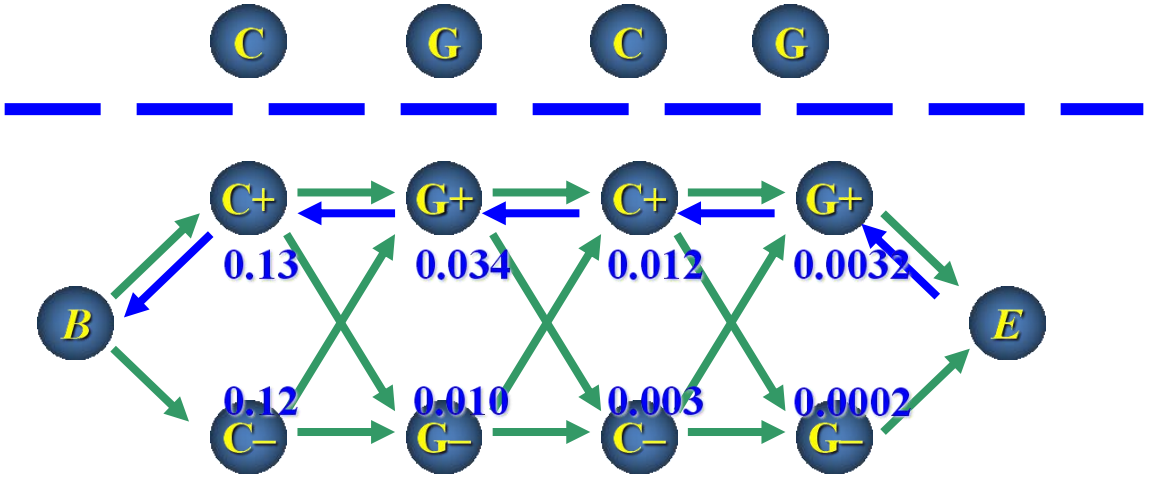
第十四式　损则有孚

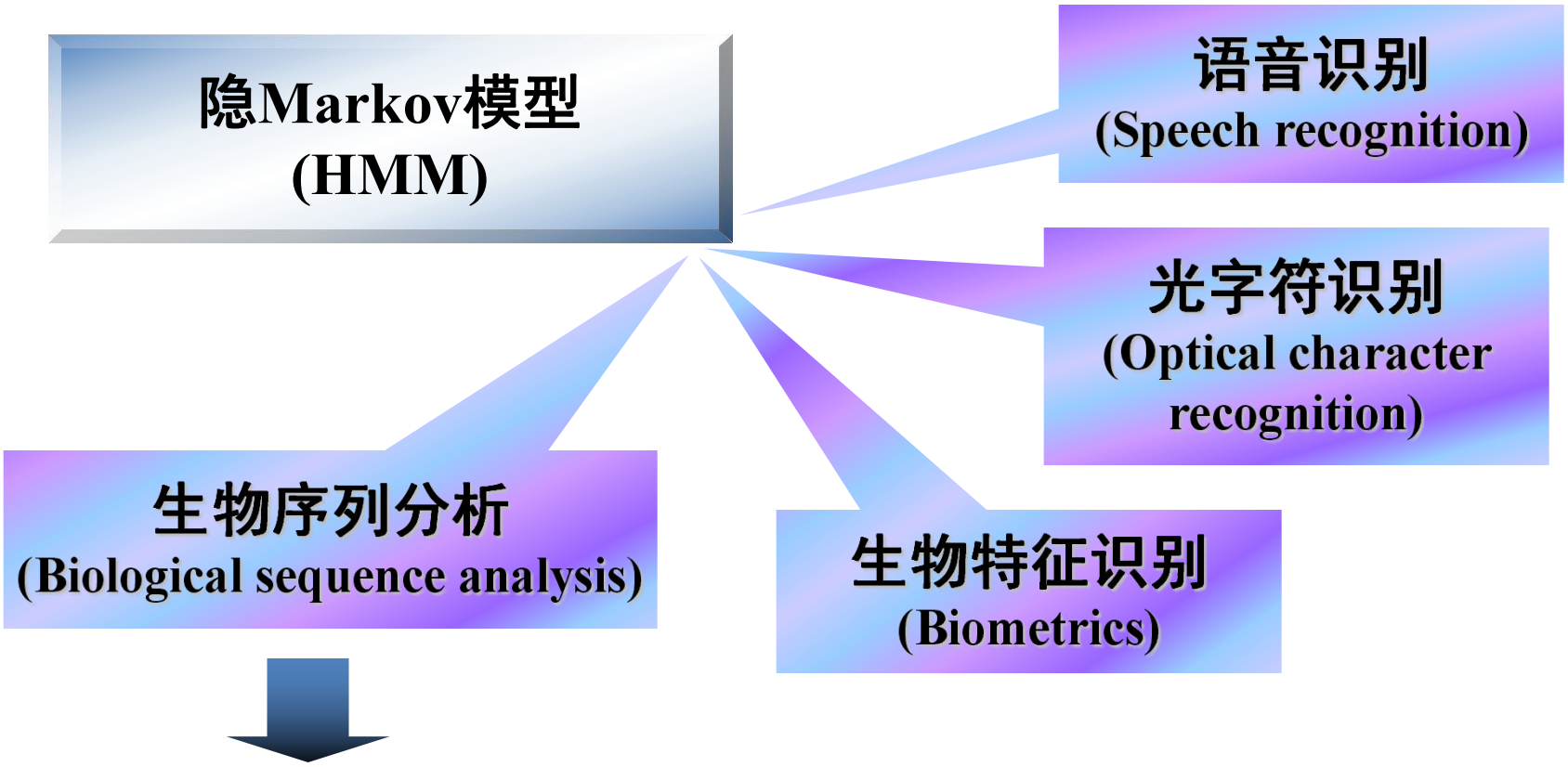
（14）

隐Markov模型方法

(HMM method)

——将核苷酸序列看成一个随机序列，DNA序列的编码部分与非编码部分在核苷酸的选用频率上对应着不同的Markov模型。由于这些Markov模型的统计规律是未知的，而HMM能够自动寻找出它们隐藏的统计规律。对于高等生物这样复杂的DNA序列，HMM必须学习不同的基因结构的信号。





（1）序列比较与搜寻（尤其是多序列比对）

（2）基因及信号的识别、预测（包括DNA编码与非编码区的识别、真核基因剪接位点信号识别、非编码区的转录调控信号识别、信号肽识别……）

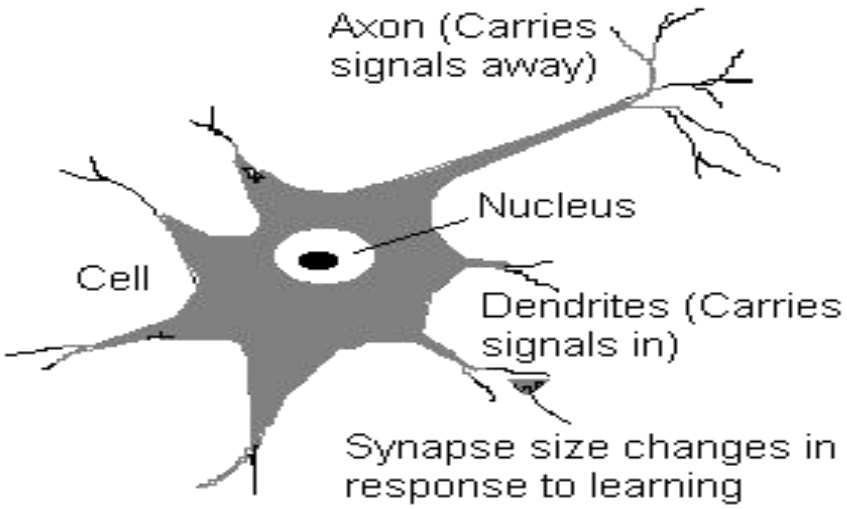
（3）蛋白质二级结构、家族、超家族预测、分类等……

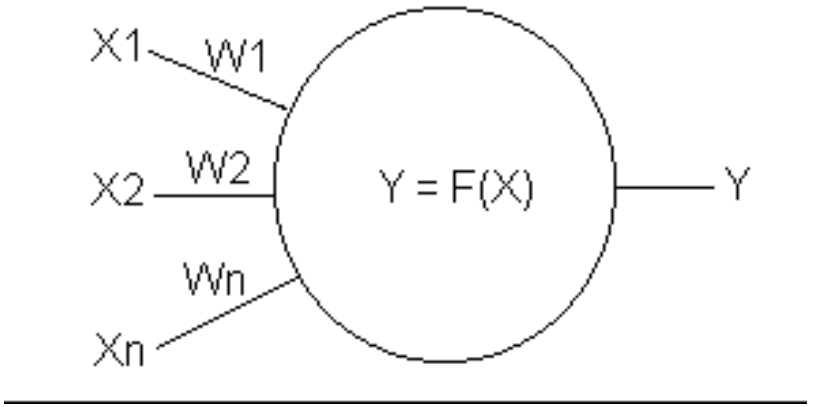
（15）

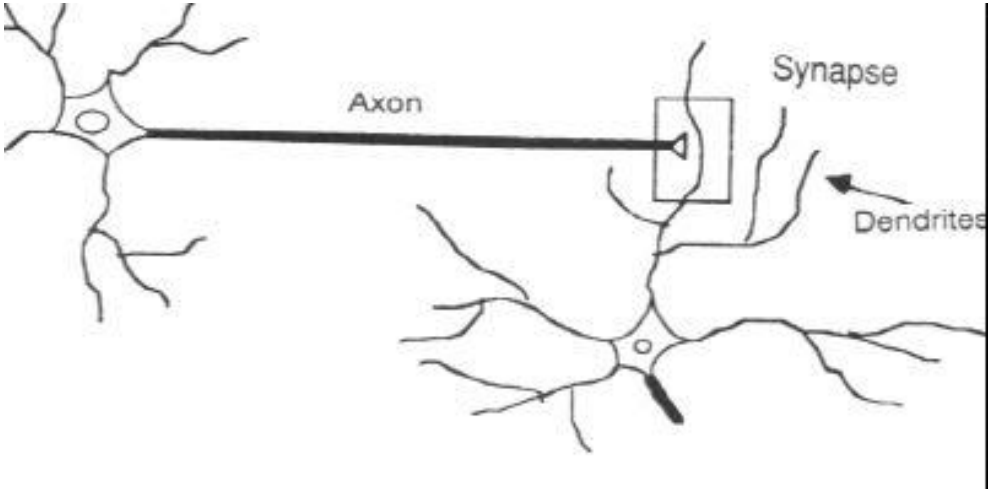
感知器与人工神经网络方法

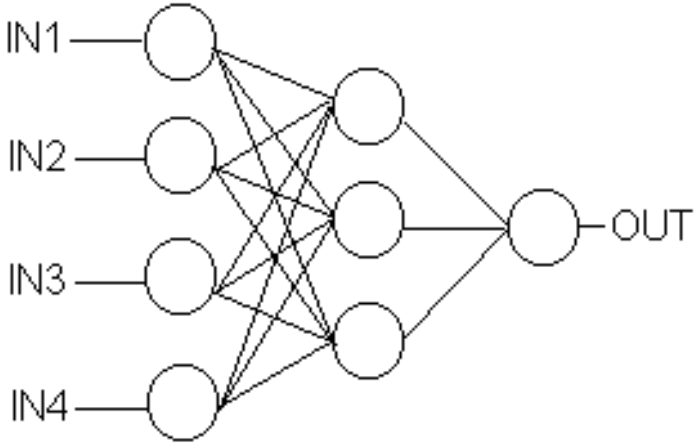
(Perceptron & ANN method)

——计算机人工神经网络是对大脑神经网络的模拟，在生物信息学研究中，无论是基因识别还是蛋白质结构预测，神经网络都取得了比其它方法更为准确的结果。


Axon (Carries signals away)
Nucleus
Cell
Dendrites (Carries signals in)
Synapse size changes in response to learning


X1
W1
X2
W2
Wn
Xn
Y = F(X)
Y


Axon
Synapse
Dendrites


IN1
IN2
IN3
IN4
OUT

（16）

决策树、支持向量机及其它模式识别方法

(Decision tree & SVM method)

——模式识别是在输入样本中寻找特征并识别对象的一种方法。

——模式识别主要有两种方法，一种是根据统计特征进行识别，另一种是根据对象的结构特征进行识别，而后者常用的方法为句法识别。

——在基因识别中，对于DNA序列上的功能位点和特征信号的识别都需要用到模式识别。

（17）

微分方程的数值方法

(Numerical methods)

——分子动力学模拟：研究生物大分子的构象，主要还是用基于半经验势函数的分子动力学方法，而量子力学则在确定势函数的参数和研究局部性质时起作用。对蛋白质进行动力学研究是利用计算机进行模拟实验的基础。

——分子动力学得到一组动力学微分方程，要求得到初值问题的解。

——微分方程的数值求解：有限差分法、有限元法

（18）

最终要诀：各类方法综合运用

All in one!

——综合运用不同的研究方法

——始终面向生物学问题

——知识和技能的学习方法

——文献的查阅和阅读方法

——中、英文论文的写作方法

生物信息学的“东邪西毒南帝北丐中顽童”

——东邪：de Bruijn图算法（基因组拼接）

——西毒：近似算法和似然估计（进化树分析）

——南帝：核函数（基因和物种分类问题）

——北丐：统计建模（序列分析）

——中顽童：深度学习和生成对抗神经网络GAN（判断问题）

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：中国科学院计算技术研所所

学习方法与要求

- 要弄懂算法的基本原理和基本公式；
- 要认真做好习题作业，加深对公式及算法步骤的理解，达到能熟练地应用算法；
- 注意培养科学的算法思维方法，理论联系实际，结合专业，了解算法方法的实际应用。

课程范围

- **生物信息学的范围**
  - 一切和生物相关数据的分析有关的算法和方法
- **面向生物信息和大数据挖掘的生物信息学特点**
  - 兼容并包、同时注重方法和应用
- **生物信息学的应用**
  - 精准医学的应用

课程结构

- 生物信息学基础；
- **生物信息中的算法设计与概率统计模型**；
- 生物大数据和深度学习。

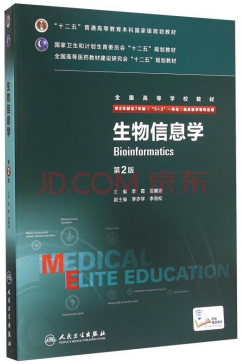
课程安排

周次	起止日期	教学内容	学时数	任课教师
9		生物信息学导论	8（讲课）	宁康
10		生物信息学基础问题和基本分析方法	8（讲课）	宁康
11		组学数据获取和蛋白组学分析（I）	8（讲课）	刘欣
12		组学数据获取和蛋白组学分析（II）	8（讲课）	刘欣
13		基因组和微生物组数据分析	8（讲课）	陈卫华
14		代谢组学数据获取和分析	8（讲课）	陈鹏
15		生物信息分析工具和数据库（I）	8（讲课）	张礼斌
16		生物信息分析工具和数据库（II）	8（讲课）	张礼斌

研究对象:
生物序列,
进化树,
生物网络,
基因表达
...

方法:
生物信息与大数据挖掘

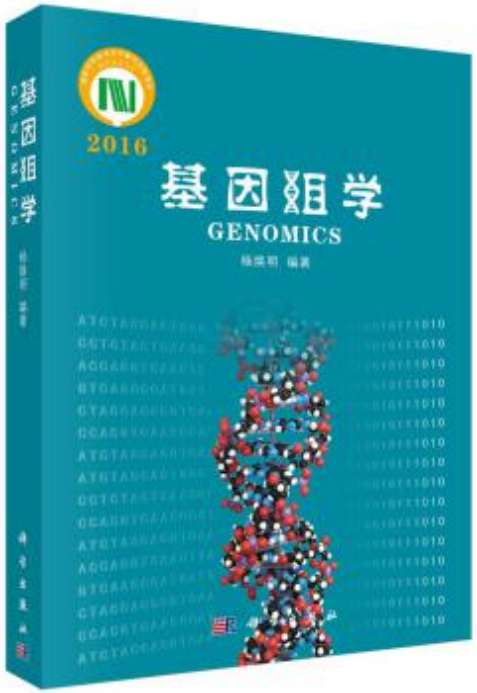
教材及参考书目

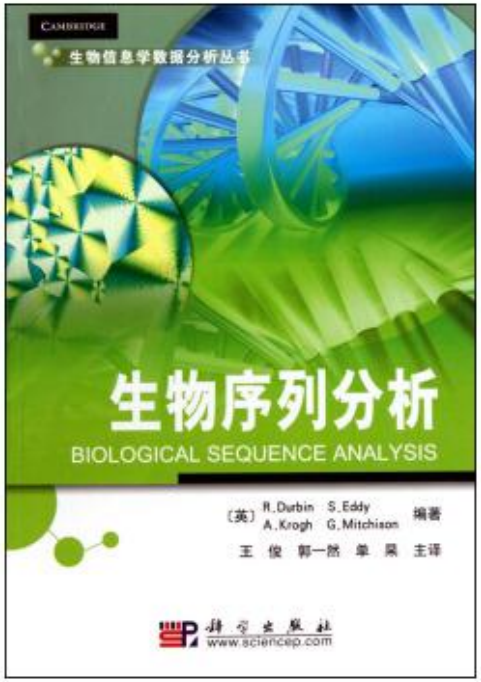


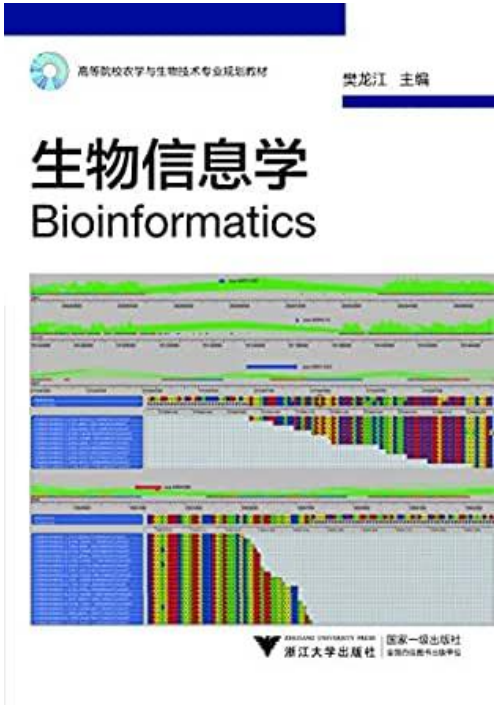
- **教学参考书：**
- 《生物信息学（八年制）》（第2版）. 人民卫生出版社. 2015年6月出版. 李霞等主编.

- **课外文献阅读：**
- 《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.
- 《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


2016
基因组学
GENOMICS


CAMBRIDGE
生物信息学数据分析丛书
生物序列分析
BIOLOGICAL SEQUENCE ANALYSIS
R.Durbin S.Eddy
A.Krogh G.Mitchison
编著
王俊 郭一然 单渠 主译
www.sciencep.com


樊龙江 主编
生物信息学
Bioinformatics
浙江大学出版社

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@Havard University

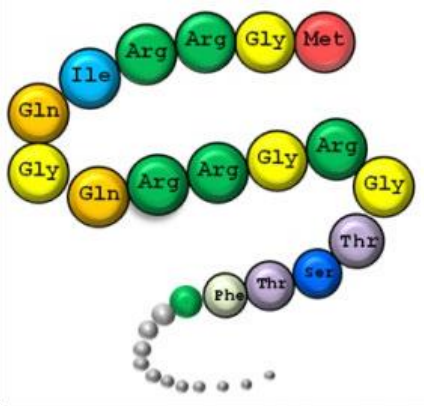
——Combinatorial Methods in Computation Biology: Ken Sung Lab@NUS

——Deep Learning in the Life Sciences: MIT

Cell

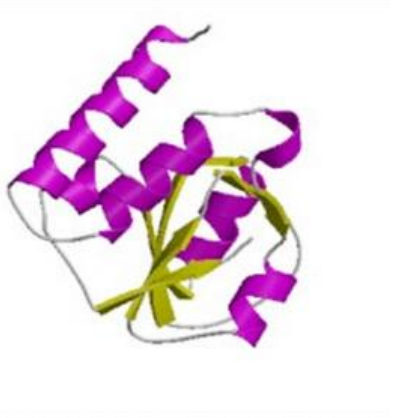
- Cell performs two type of functions:
  - Perform chemical reactions necessary to maintain our life
  - Pass the information for maintaining life to the next generation
- Actors:
  - Protein performs chemical reactions
  - DNA stores and passes information
  - RNA is the intermediate between DNA and proteins

Protein



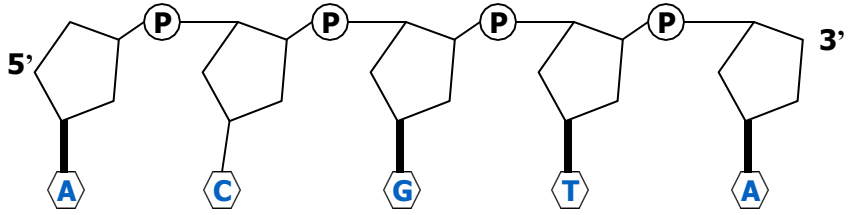
- Protein is a sequence composed of an alphabet of 20 amino acids.
  - The length is in the range of 20 to more than 5000 amino acids.
  - In average, protein contains around 350 amino acids.

- Protein folds into three-dimensional shape, which form the building blocks and perform most of the chemical reactions within a cell.
  - Structural: building blocks of cells
  - Signaling: Turn gene on or off, Pass signal between cells, Get signal from environment.
  - Catalyze reaction: Enzyme



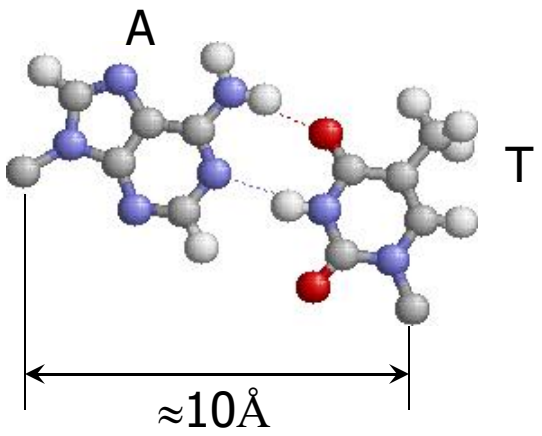
DNA

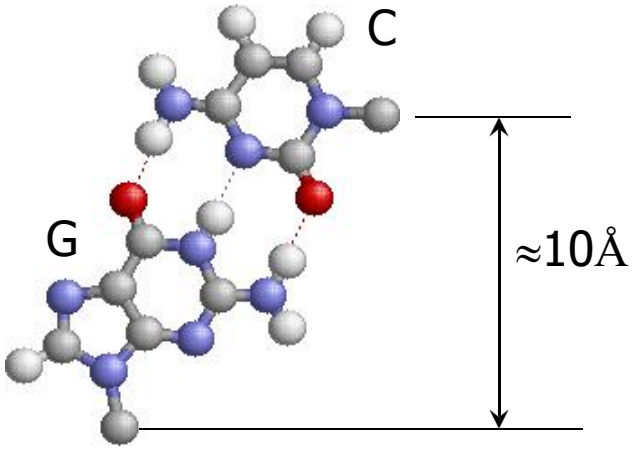
- DNA stores the instruction needed by the cell to perform daily life function.
- It consists of two strands which interwoven together and form a double helix.
- Each strand is a chain of some small molecules called nucleotides.
- There are 4 types of nucleotides: A, C, G, and T.



Watson-Crick rules

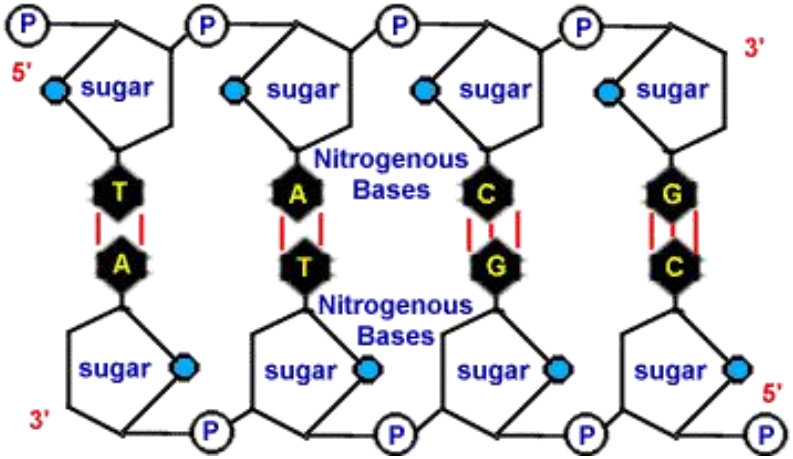
- Complementary bases:
  - A with T (two hydrogen-bonds)
  - C with G (three hydrogen-bonds)

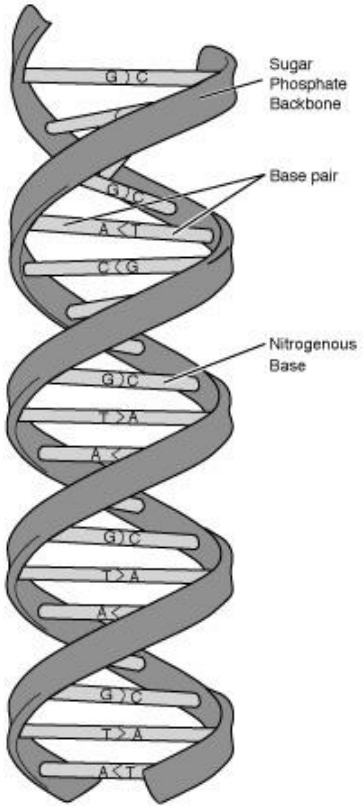




Double stranded DNA

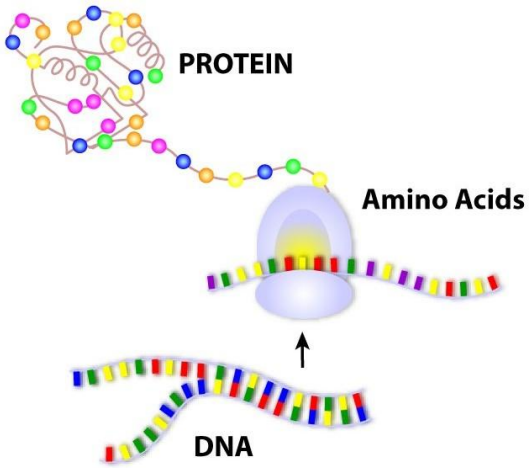
- Normally, DNA is double stranded within a cell. The two strands are antiparallel. One strand is the reverse complement of another one.
- The double strands are interwoven together and form a double helix.
- One reason for double stranded is that it eases DNA replicate.

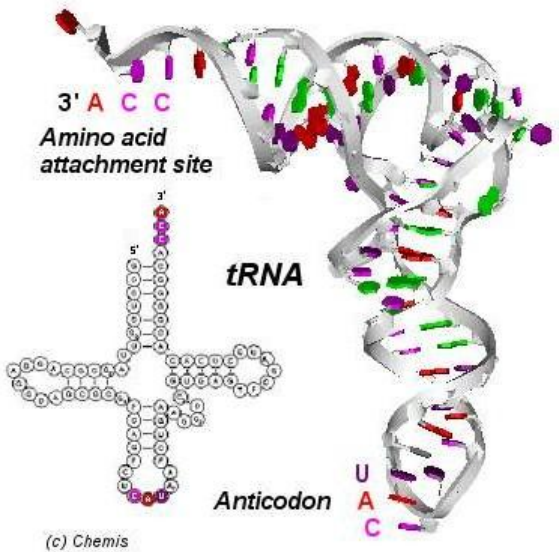




RNA

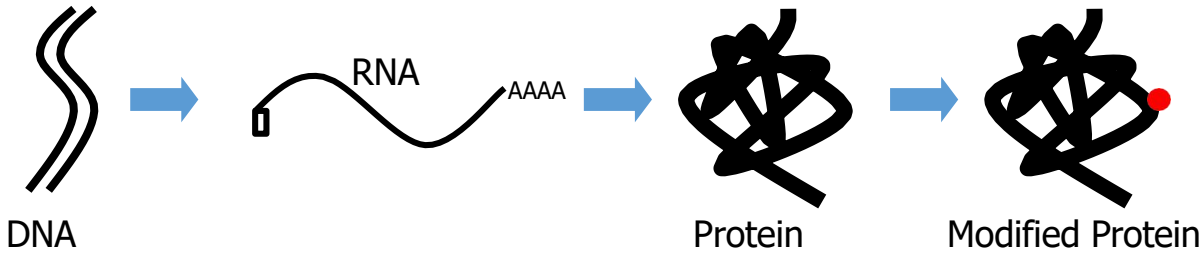
- RNA has two functions
  - As an intermediate between DNA and protein
  - Form complex 3-dimensional structure and perform some functions.





Central Dogma

- Central Dogma tells us how we get the protein from the gene. This process is called gene expression.
- The expression of gene consists of two steps
  - Transcription: DNA → mRNA
  - Translation: mRNA → Protein
  - Post-translation Modification: Protein → Modified protein



Replicate or Repair of DNA

- DNA is double stranded.
- When the cells divide,
  - DNA needs to be duplicated and passes to the two daughter cells.
  - With the help of DNA polymerase, the two strands of DNA serve as template for the synthesis of another complementary strands, generating two identical double stranded DNAs for the two daughter cells.
- When one strand is damaged,
  - it is repaired with the information of another strand.

What is bioinformatics? (from computer science point of view)

- [wiki] Bioinformatics is an interdisciplinary field that develops methods and software tools for understanding biological data.

- Bioinformatics combines
  - biology,
  - computer science,
  - information engineering,
  - mathematics and
  - statistics

  to analyze and interpret biological data.

The Promises of Bioinformatics

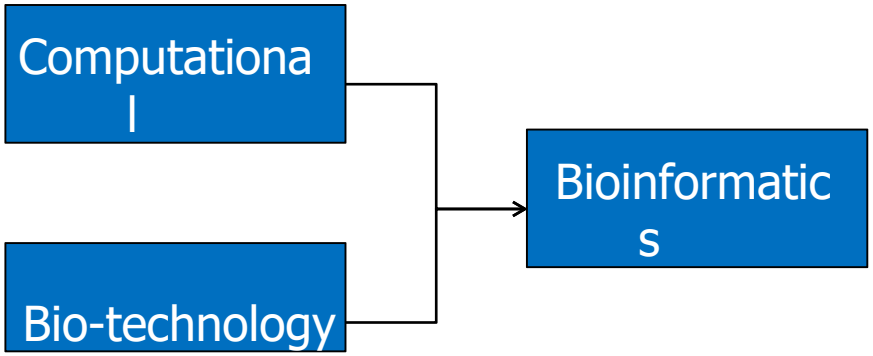
- To the patient:
  - Better drug, better treatment
- To the pharma:
  - Save time, save cost, make more $
- To the scientist:
  - Better science

Pervasiveness of Bioinformatics

- Bioinformatics is mandatory for large-scale biology
  - e.g., High-throughput, massively-parallel measurements, or “lab on a chip” miniaturization
- Computational data analysis is mandatory for indirect experimental methods
  - e.g., reconstruction haplotype from genotype data
- Limitless opportunities!

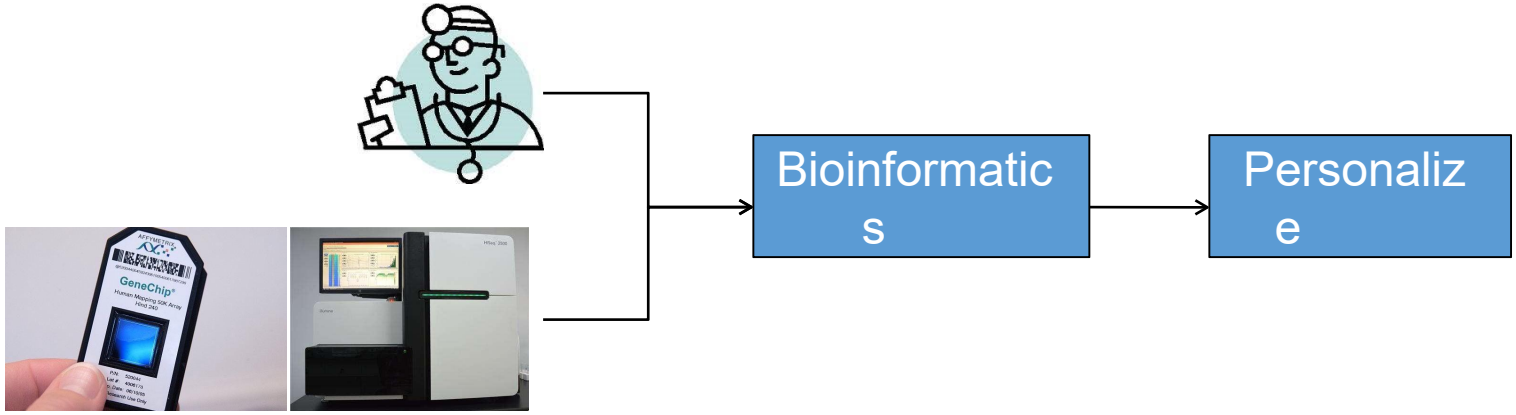
What do we study?

- We study the application of computer science and bio-technology to solve bioinformatics problems



Why these problems are important?

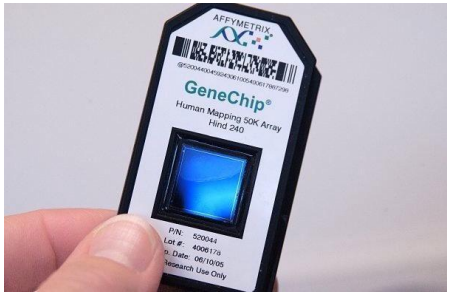
- Personalize sequencing is a big market.



- A number of big companies and start-up companies.
  - Bioinformatics is the main driving force.

Technologies

DNA array

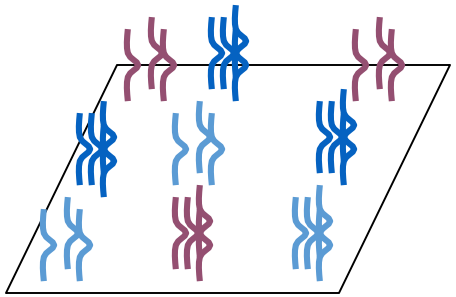


- The idea of hybridization leads to the DNA array technology.

- In the past, “one gene in one experiment”
- Hard to get the whole picture
- DNA array is a technology which allows researchers to do experiment on a set of genes or even the whole genome.

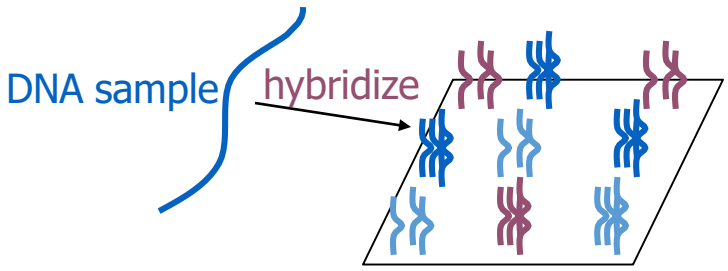
DNA array's idea (I)

- An orderly arrangement of thousands of spots.
- Each spot contains many copies of the same DNA fragment.



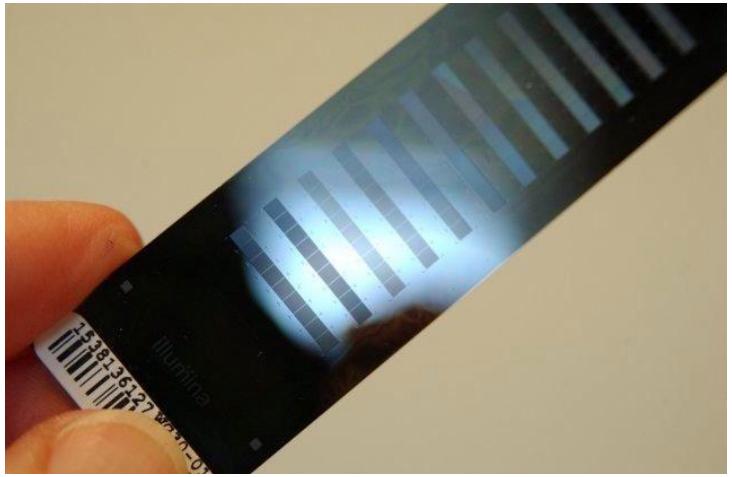
DNA array's idea (II)

- When the array is exposed to the target solution, DNA fragments in both array and target solution will match based on hybridization rule:
  - A=T, C≡G (hydrogen bond)
- Such idea allows us to do thousands of hybridization experiments at the same time.



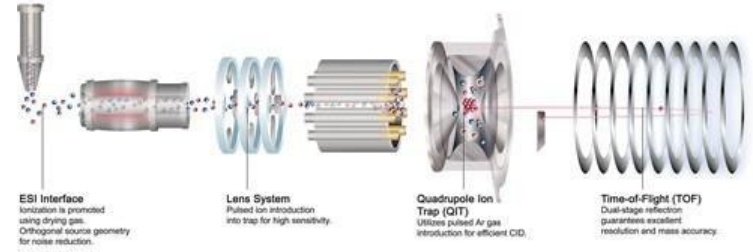
Genotyping chip

- Based on microarray technology.
- Allows us to know the genotype for millions of positions in our genome.

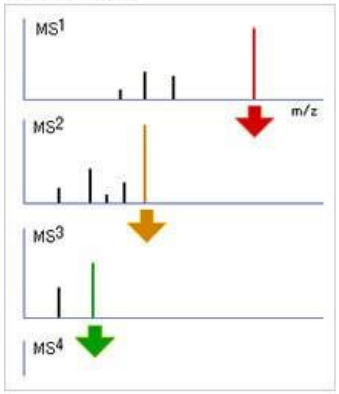


Mass Spec

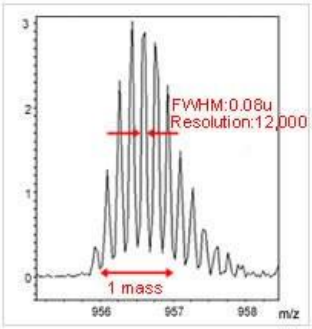
- Measure mass of different molecules accurately



MS^n measurement:
One peak acquired by MS^1 is performed MS^2, and one peak acquired by MS^2 is performed MS^3.
LCMS-IT-TOF can perform by MS^{10}.
This function supports structural analysis strongly.



High resolution and accuracy
This data shows a Mass spectra of Insulin Hexavalent Ion.
Resolution of >12,000 was achieved.
6 peaks are separated clearly in one mass difference.



Main unit: 1685mm , LC unit (by module): 260mm

Sequencing Technology

- **Next-generation sequencing (NGS)** can generate tens of billions of DNA bases efficiently.
- These machines can generate large amount of data per day.
- For example, Illumina sequencer can sequences 60G DNA bases per run

Illumina HiSeq

Pacific BioSciences

Oxford Nanopore

Short read machine

Long read machine

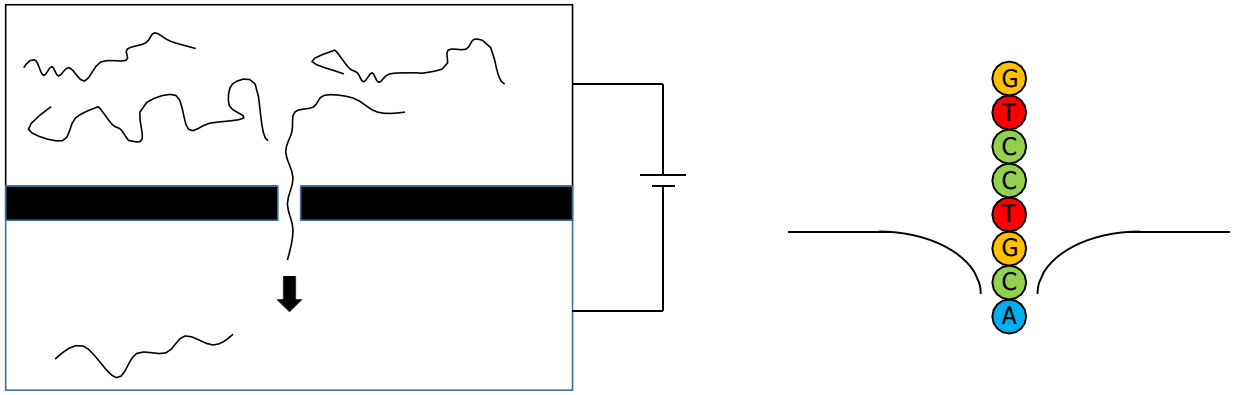
Illumina machine sequences short reads

Illumina HiSeq

gatggcccaggagaaccccaagatgcacaactcggagatcagcaagcgcctgggcgccga

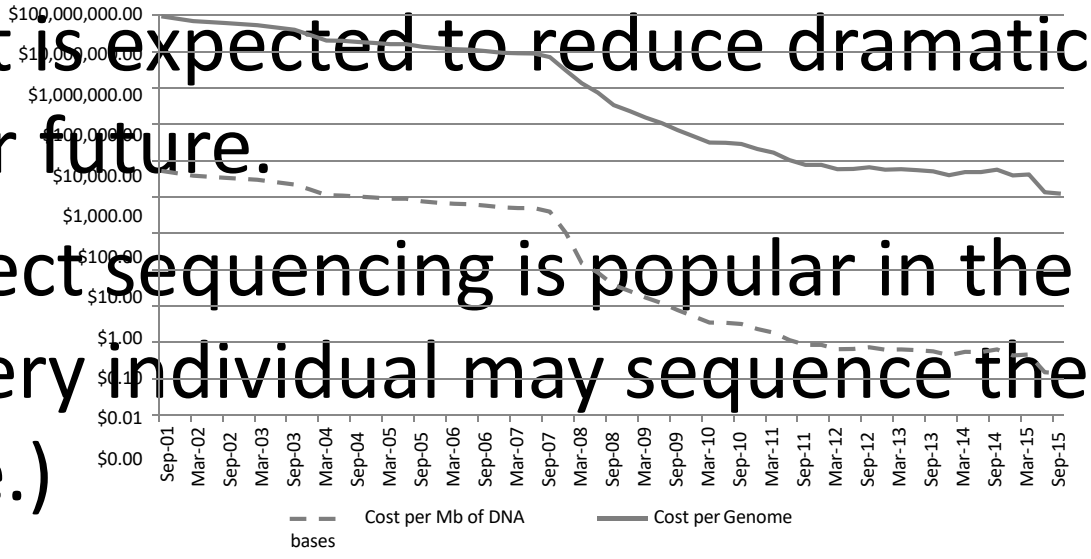
Nanopore sequences long reads

- This technology detect nucleotides by measuring the ionic current flowing through the pore.



The cost of high-throughput sequencing is continue to reduce

- Below figure shows the cost of sequencing.
- Now, to sequence an individual genome, the cost is about US$1000.
- The cost is expected to reduce dramatically in the near future.
- We expect sequencing is popular in the future. (E.g. every individual may sequence their genome.)



Computational techniques

Computational techniques

- Algorithm
  - Greedy algorithm
  - Dynamic Programming
  - EM algorithm
- Data-structure
  - Perfect hashing
  - Suffix tree
- Machine learning
  - SVM
  - k-mean
  - Neural network
- Statistics
  - Normal distribution, etc

Bioinformatics Problems

Example biology problems that can be solved by algorithm

- Learn the mutations in our genome
- Construct and comparing phylogenetic trees
- Whole genome alignment
- Genome rearrangement
- Population genetics
- RNA secondary structure prediction
- Peptide sequencing
- Virus sequencing using microarray

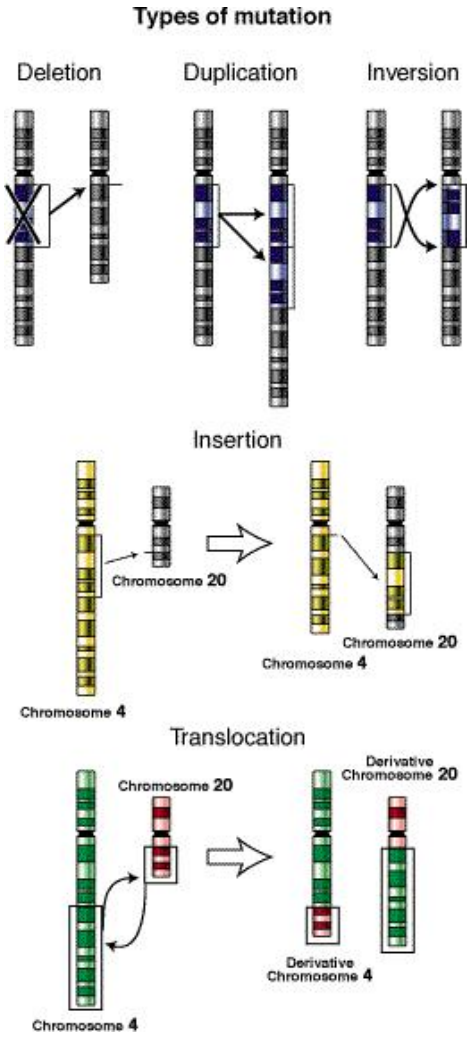
Learning mutation

Technology: Sequencing of geome

- Despite the near-perfect replication, infrequent unrepaired mistakes are still possible.
  - Those mistakes are called mutations.

- The most common type of mutation is point mutation.
- Other mutations are structural variations.

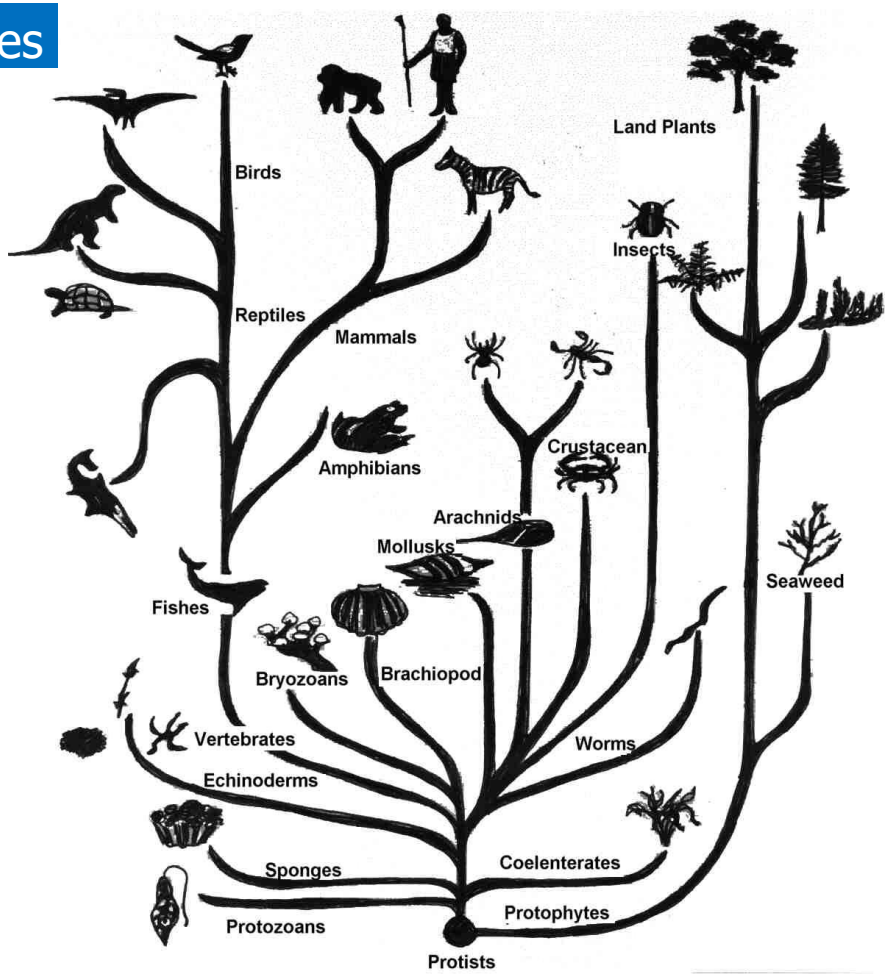
- Note: mutation can occur in DNA, RNA, and Protein



Technology: Sequencing of genes & genomes

Evolutionary tree

- Occasionally, mutations make the cells or organisms survive better in the environment.
  - The selection of the fittest individuals to survive is called natural selection.
- Mutation and natural selection have resulted in the evolution of a diversified organisms.
- Given the mutations, we can study the evolutionary tree of the individuals.
- Note that mutation is also the cause of diseases (like cancer, flu). We can study diseases by analyzing evolutionary tree.

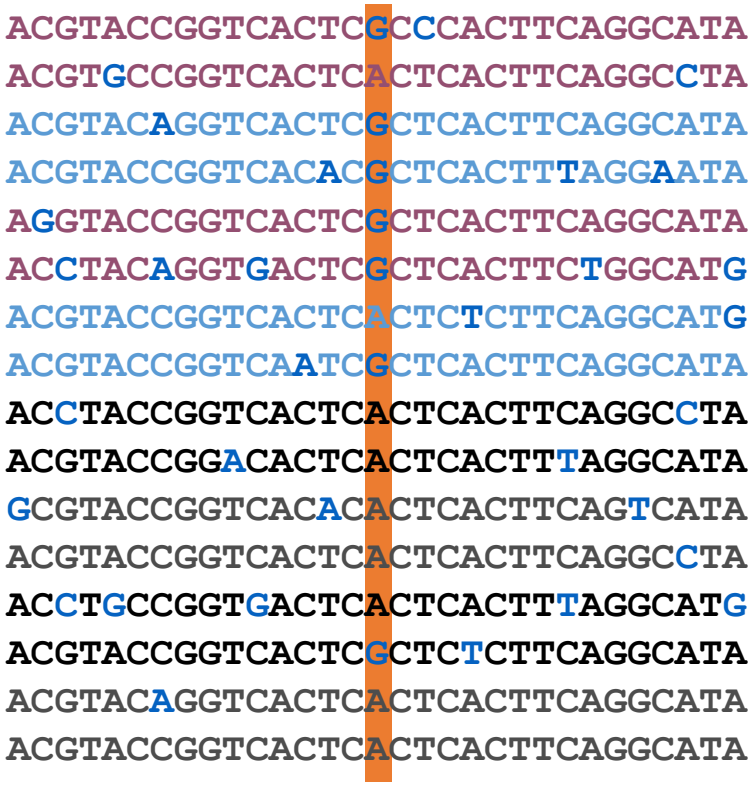


Technology: Genotyping

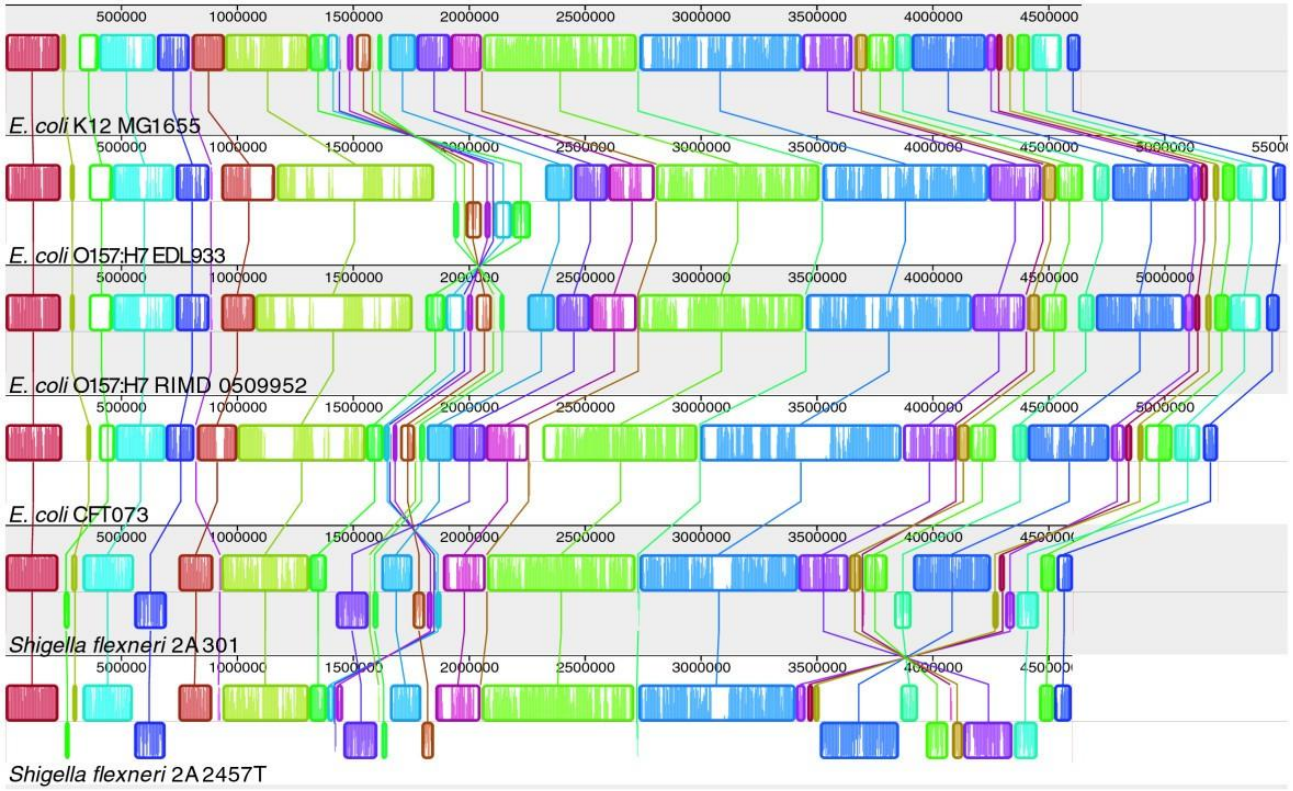
Population genetics: Finding causal variants

Case
(Disease sample)

Control
(Normal sample)



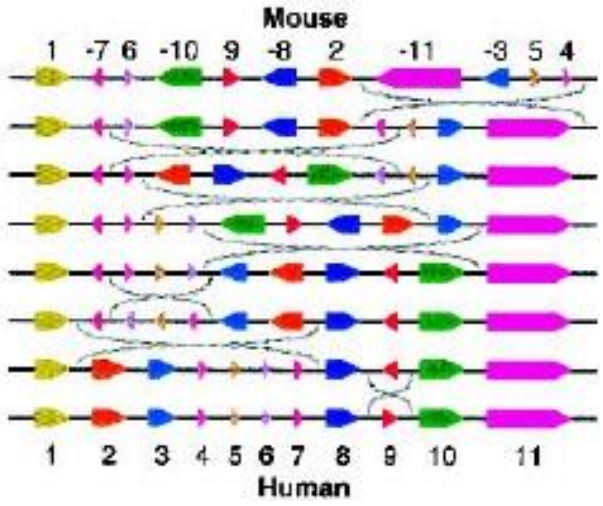
Whole genome alignment



Genome rearrangement

Technology: Sequencing of genome

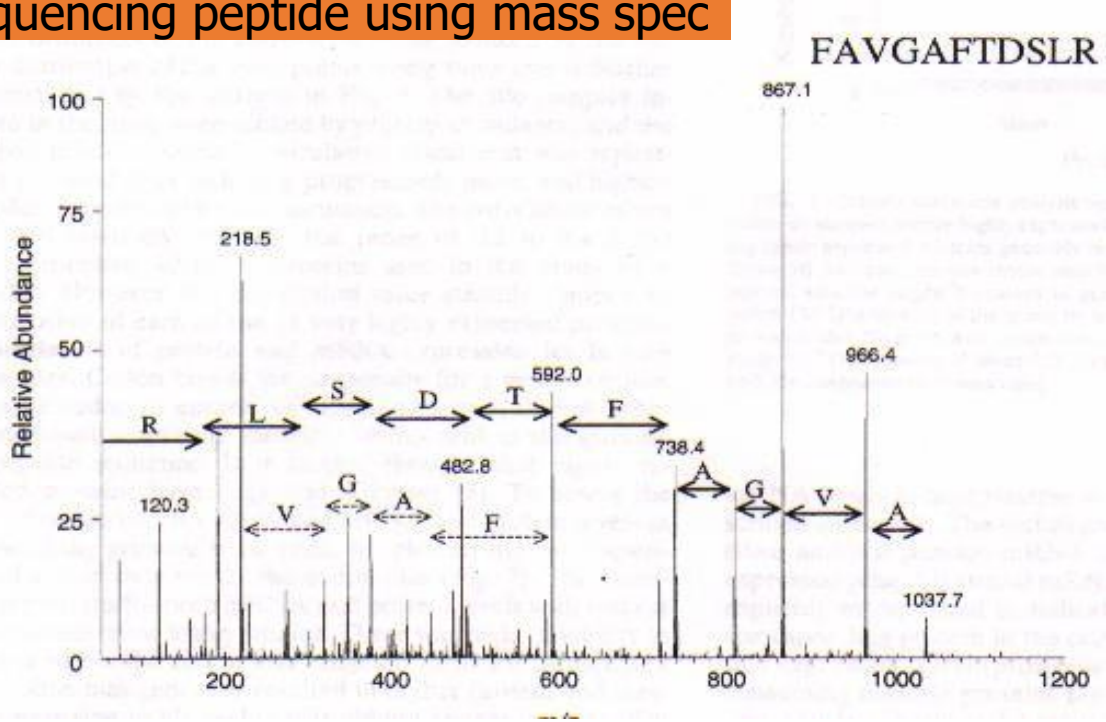
- chromosome X of human can be transformed to chromosome X of mouse using 7 reversals



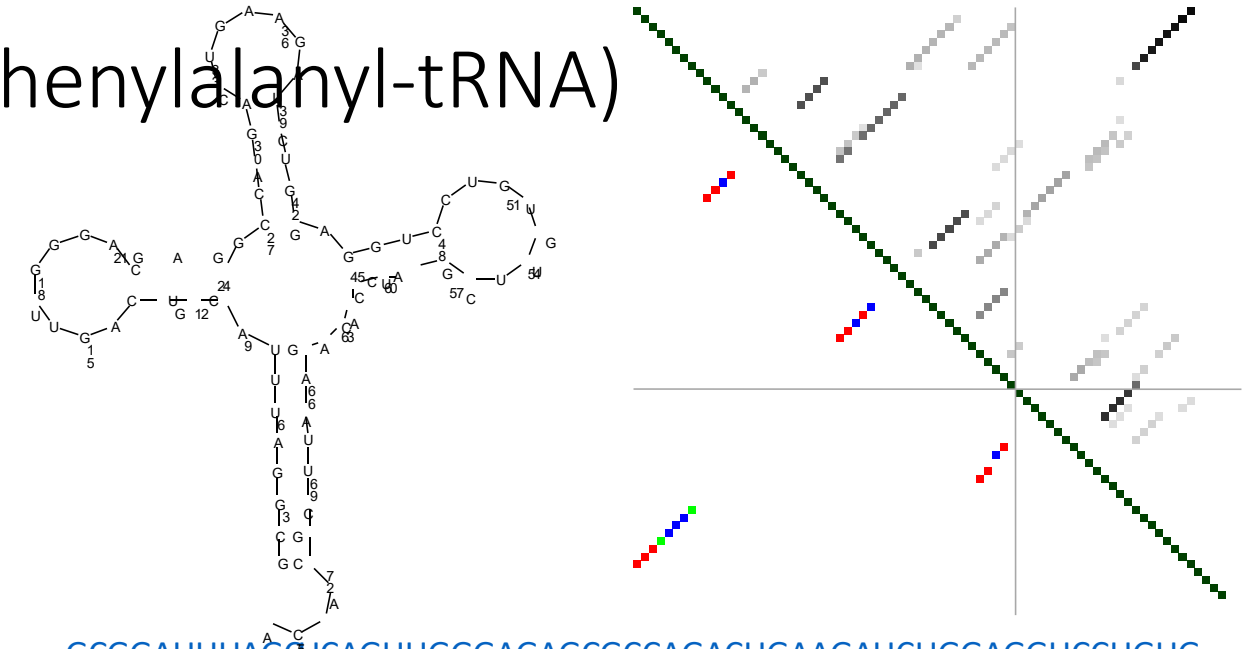
Technology: Mass spec

Peptide sequencing

Sequencing peptide using mass spec



Example (Secondary structure for phenylalanyl-tRNA)



GCGGAUUUAGCUCAGUUGGGAGAGCGCCAGACUGAAGAUCUGGAGGUCCUGUG
UUCGAUCCACAGAAUUCGCACCA