水母基因

生物医学大数据与精准风险评估

赵 南，博士

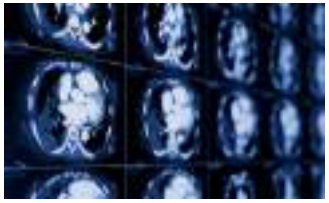
CSO 联合创始人

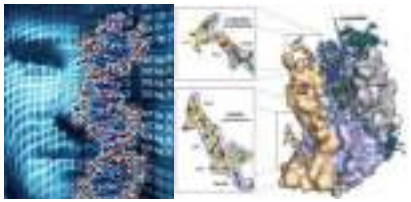
北京水姆科技有限公司

2015年奥巴马在国情咨文宣布将精准医疗列入重点工程。2016年2月25日，美国启动精准医疗计划。

2015年3月，中国首次精准医学战略专家会议，规划在2030年前精准医疗投入600亿。

健康大数据与数据科学





临床健康数据

生物大分子数据

个人健康大数据

数据科学分析建模

数据挖掘

机器学习

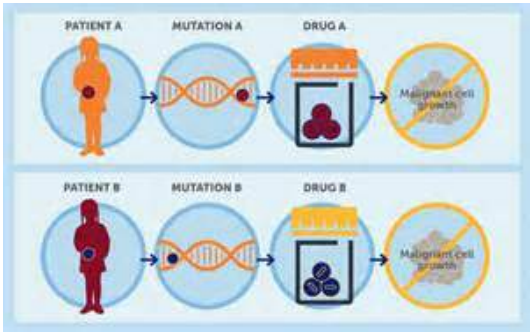
信息检索

精准医疗的三大应用领域

精准诊断
(ctDNA, cfDNA)

精准预防
(个性化疾病风险评估与管理)

精准治疗及用药
(肿瘤靶向药)



上医治未病，中医治欲病，下医治已病。

-- 《黄帝内经》

重大疾病预防的现状

健康管理需求

老龄化

慢病、大病高发

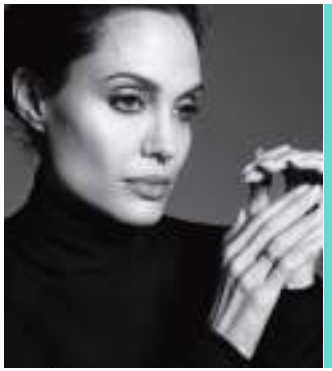
大众对基因检测的认知
利润率很高

健康管理公司不断增加
但缺乏有效的服务手段

疾病预防需求

目前的健康管理手段：常规体检

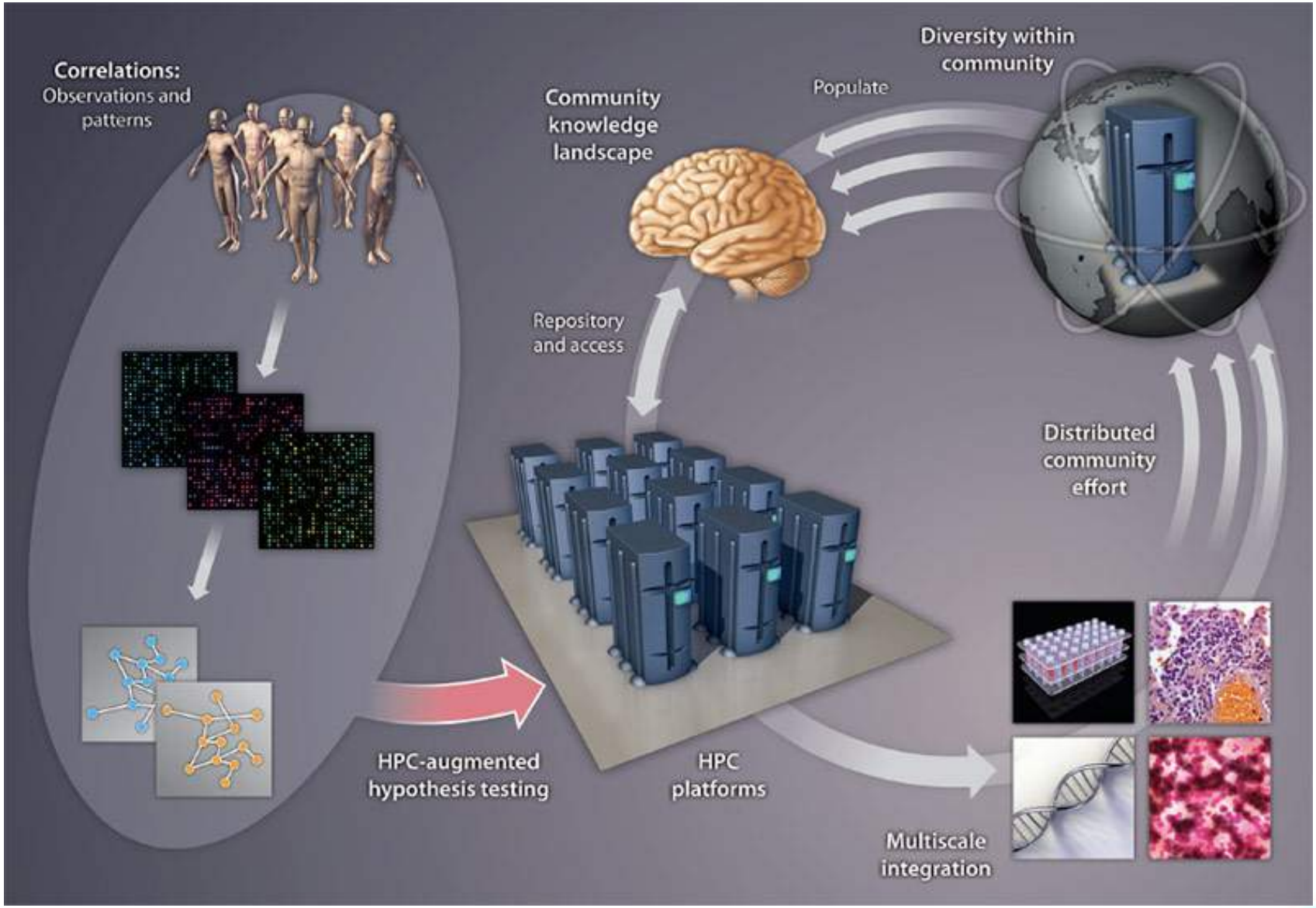
分子生物学检测在精准医疗中的应用



好莱坞红星安吉丽娜·朱莉（Angelina Jolie）通过基因检测发现自身带有家族性的BRCA1基因突变，这也意味着她拥有87%和50%的几率罹患乳腺癌和卵巢癌。朱莉拥有癌症家族史，家族中一共有三位女性亲人都死于女性相关癌症。朱莉对乳腺、卵巢和输卵管进行了预防性切除，使其患乳腺癌的几率从87%下降到5%以下。

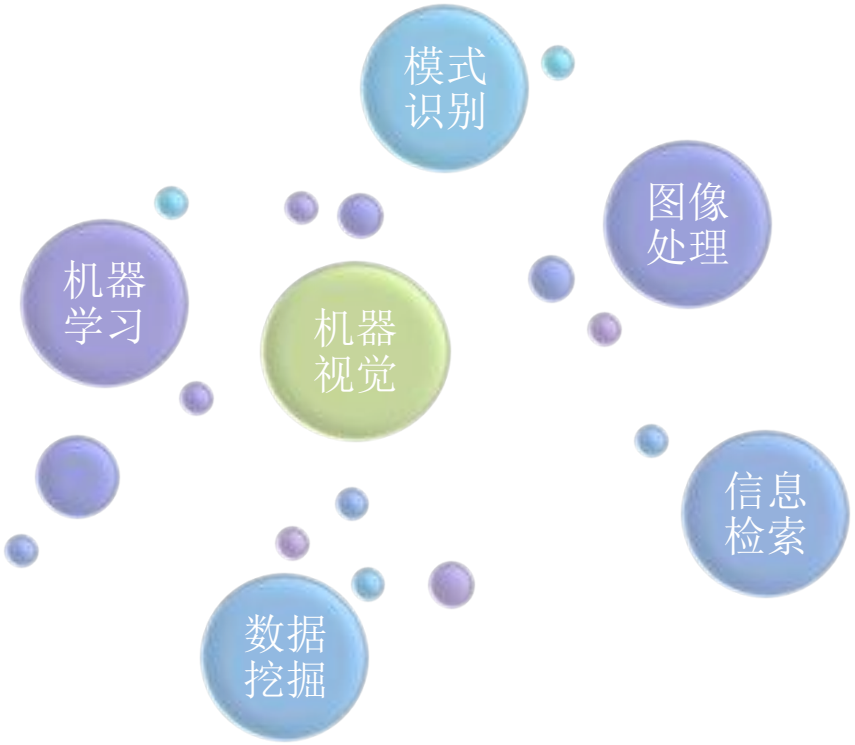
斯坦福大学分子生物学教授 Michael Snyder 最初通过基因组测序，得知他患糖尿病的风险很高。而呼吸道合胞病毒的感染，触发他患上了2型糖尿病。之后他检测了自己20个不同时间所采集的血样，得到基因组学、代谢组学以及蛋白质组学的生化数据，描绘出自身免疫系统、代谢和基因活动的状态。经过6个月的饮食调整和积极锻炼，使得血糖恢复到正常水平。

生物医学大数据

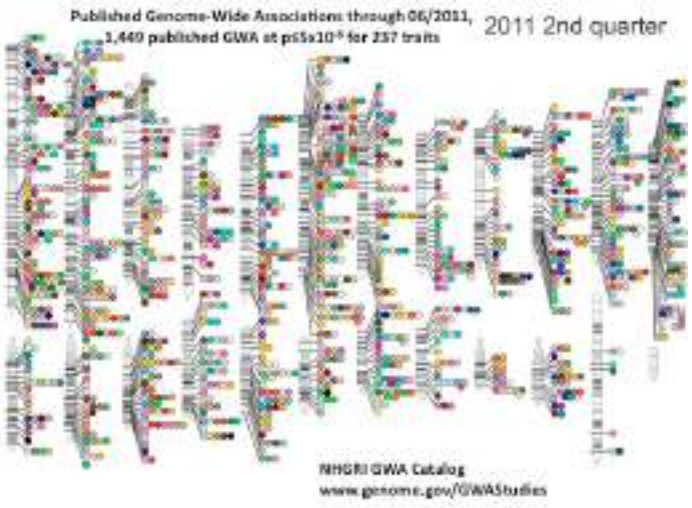

Correlations:
Observations and patterns
Community knowledge landscape
Populate
Diversity within community
Repository and access
Distributed community effort
HPC-augmented hypothesis testing
HPC platforms
Multiscale integration

大数据时代的机遇与挑战

- *Subjective*
- *Implicit*
- *Indirect*
- *Inconsistent*
- *Laborious*
- *Tedious*



生物信息分析及大数据挖掘技术

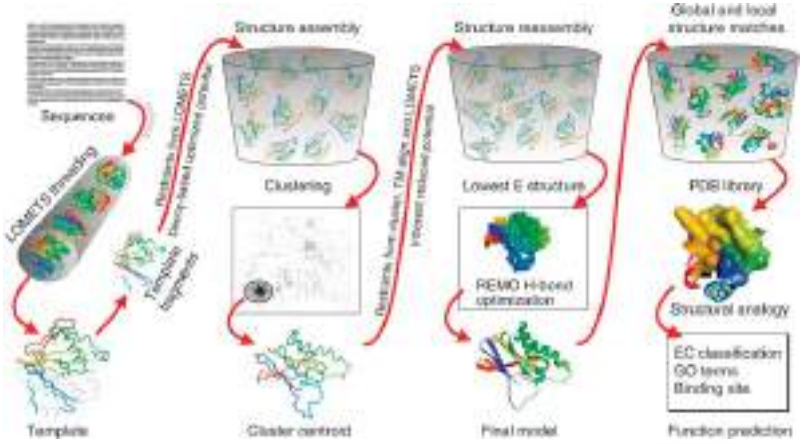


全基因组关联分析

应用全基因组中数以百万计的SNP位点为分子遗传标记，进行全基因组水平上的对照分析及关联分析，发现并鉴定与复杂性状相关联的遗传变异。

功能预测

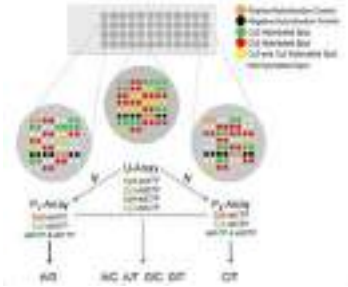
利用生物信息学的方法，通过计算机模拟和计算来预测DNA、RNA以及蛋白质序列的结构和功能信息，了解各个基因所要表达的生物学意义，真的的揭开遗传的奥秘。



易感基因疾病遗传风险评估

基因体检项目	
肿瘤	肺癌、乳腺癌、胃癌、食管癌、结直肠癌、膀胱癌、胰腺癌、甲状腺癌、肾癌、鼻咽癌、肝癌、口腔癌、喉癌、慢性淋巴细胞白血病、黑色素瘤、基底细胞癌、脑胶质瘤、脑膜瘤、神经母细胞瘤、鳞状细胞癌、滤泡性淋巴瘤、骨髓增生性肿瘤、腹主动脉瘤、脑动脉瘤、霍奇金淋巴瘤
内分泌系统疾病	1型糖尿病、2型糖尿病、高血压、高血脂、肥胖症、甲状腺机能减退症
泌尿系统疾病	慢性肾病、继发性肾病综合征、特发性膜性肾病、肾结石
消化系统疾病	溃疡性结肠炎、乳糜泻、胆结石、克罗恩病、酒精性脂肪肝、非酒精性脂肪性肝病
神经系统疾病	老年痴呆、帕金森、抑郁症、不宁腿综合征、双相型障碍、进行性核上麻痹、精神分裂症、偏头痛、丛集性头痛、肌萎缩侧索硬化症、酒精依赖、注意力缺陷多动障碍、强迫性神经障碍、抽动秽语综合征、尼古丁依赖、特发性震颤、发作性睡病
呼吸系统疾病	哮喘、慢性阻塞性肺疾病、肺纤维化
骨科疾病	骨质疏松、强直性脊柱炎、脊柱侧凸、变形性骨炎、骨关节炎、腰椎间盘突出症
免疫系统疾病	系统性红斑狼疮、类风湿关节炎、痛风、白塞氏病、风湿性关节炎、选择性IgA缺陷、干燥综合征、原发性胆汁性肝硬化、多发性硬化症、结节病
皮肤病	牛皮癣、遗传过敏性皮炎、白癜风、瘢痕瘤、掌跖腹掌缩症、皮肤型硬皮病
五官科疾病	剥脱性青光眼、耳硬化症、过敏性鼻炎、老年性黄斑病变、斑秃
心血管疾病	心房颤动、冠心病、脑中风、静脉血栓栓塞、外周动脉病变、心脏性猝死
男性疾病	前列腺癌、男性不育症、睾丸癌
女性疾病	宫颈癌、卵巢癌、子宫内膜癌、子宫肌瘤、多囊卵巢综合征、子宫内膜异位、妊娠肝内胆汁淤积症、妊娠糖尿病
营养代谢	碳水化合物、蛋白质、脂肪、维生素A、维生素B12、维生素D、维生素E、叶酸、钙、铁

DNA检测：基因芯片 + 二代高通量测序



- 根据评价发病率（IR，incident rate）和人均寿命（T）估算评价终身患病风险（LR，lifetime risk）
- 根据疾病相关位点比值比（ OR，odd ratio）计算个人终身患病风险
- 使用个人终身患病风险与平均终身患病风险计算相对患病风险

$$LR_{avg}(D) = 1 - e^{-IR(D)T}$$

$$odd(D) = \frac{LR_{avg}(D)}{1 - LR_{avg}(D)}$$

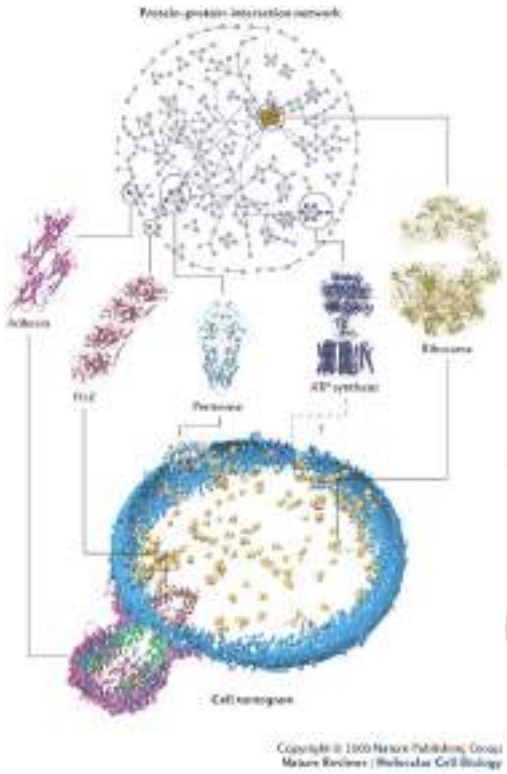
$$odd(D|\bigcap_{k=1}^{K} G_{m_k,k}) = odd(D) * \prod_{k=1}^{K} OR_{G_{m_k,k}}(D)$$

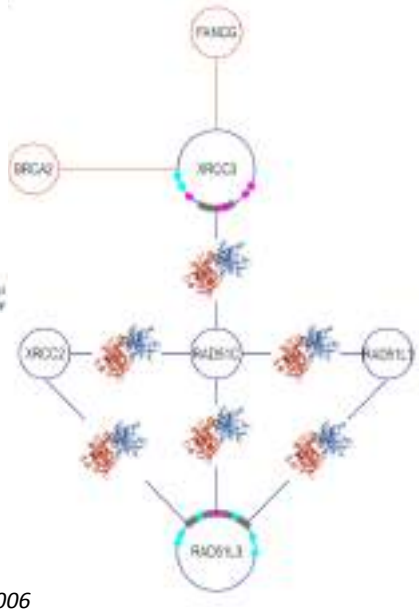
$$LR(D|\bigcap_{k=1}^{K} G_{m_k,k}) = \frac{odd(D|\bigcap_{k=1}^{K} G_{m_k,k})}{1 + odd(D|\bigcap_{k=1}^{K} G_{m_k,k})}$$

$$RR(D|\bigcap_{k=1}^{K} G_{m_k,k}) = \frac{LR(D|\bigcap_{k=1}^{K} G_{m_k,k})}{LR(D)}$$

多种生物分子数据整合、大尺度建模分析

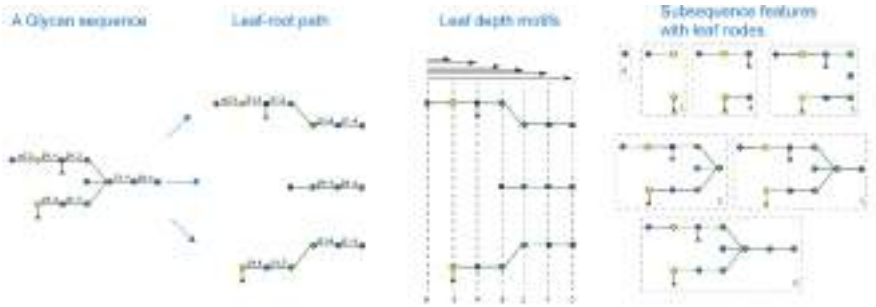
蛋白质组、结构、网络通路数据



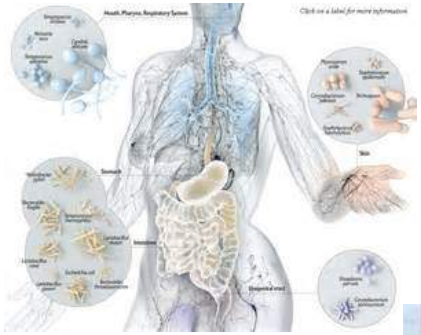


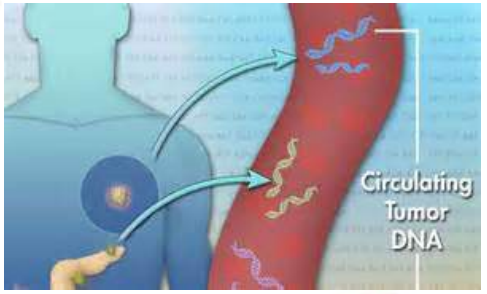
P. Aloy, and R. Russell, Nat Rev Mol Cell Biol. 7(3) 2006

糖组测序、结构数据



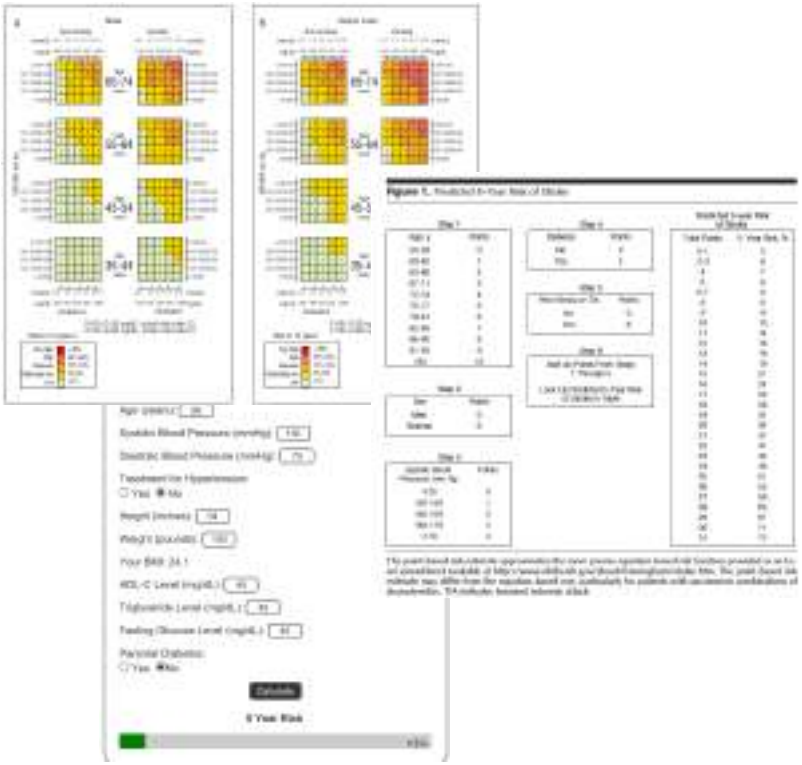
肿瘤、微生物基因检测





临床健康数据疾病风险评估

Framingham risk score: 对冠心病、高血压、心力衰竭、跛行、中风、心房颤动、糖尿病、心血管疾病8类项目进行风险预测

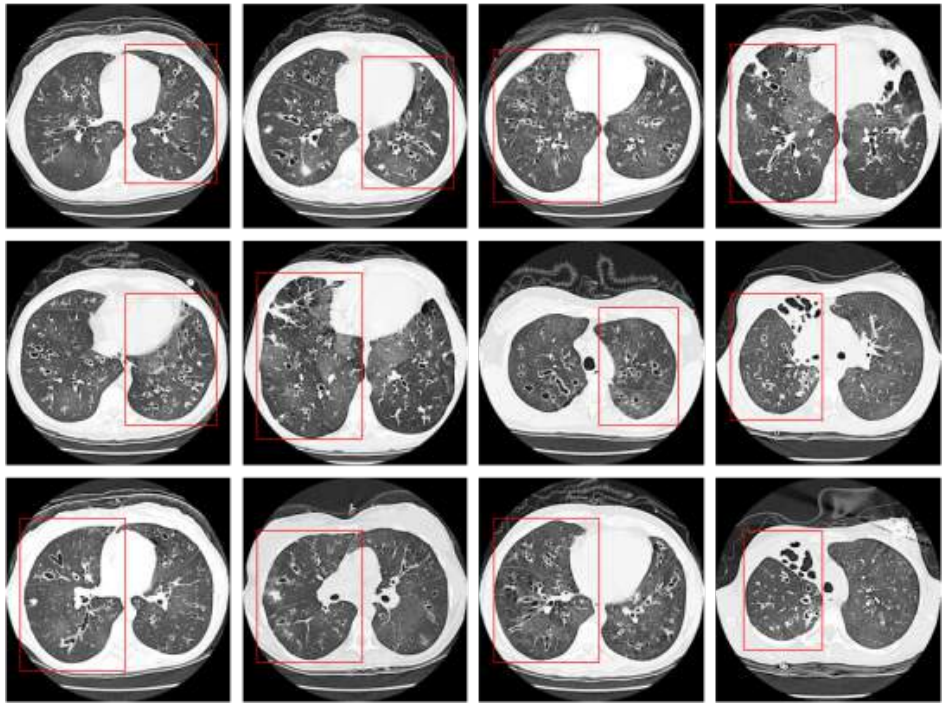


ClinRisk提供癌症、糖尿病、骨折、肾病、心血管疾病未来1-10内的风险预测。

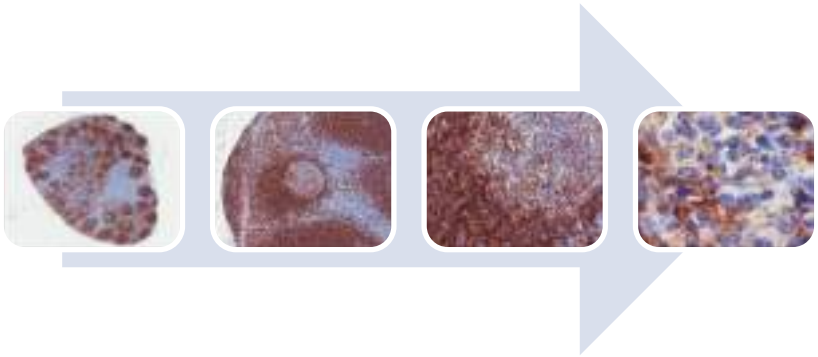
Cancer	Type	Risk
No cancer		4.71%
Any cancer		95.29%
	other	60.83%
	blood	21.87%
	colorectal	4.38%
	pancreatic	3.89%
	prostate	1.93%
	lung	1.33%
	gastro-oesophageal	0.99%
	testicular	0.04%

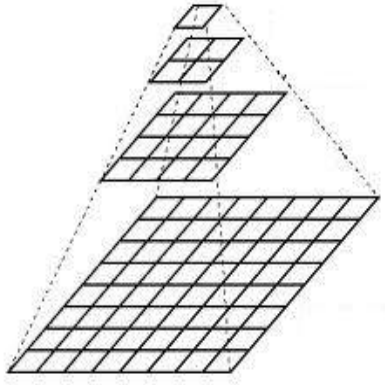
复杂临床影像数据

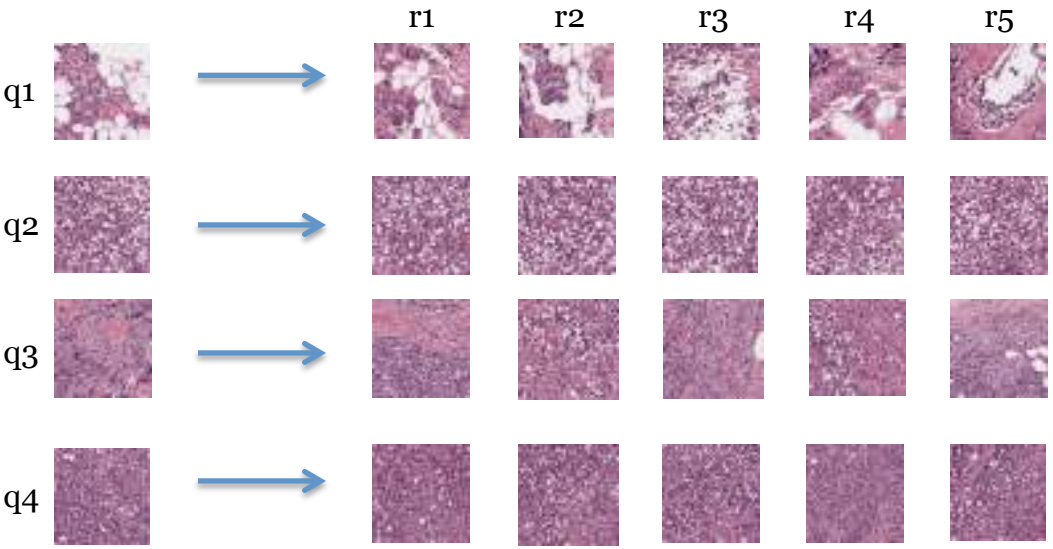
CBIR：基于内容的医学影像检索



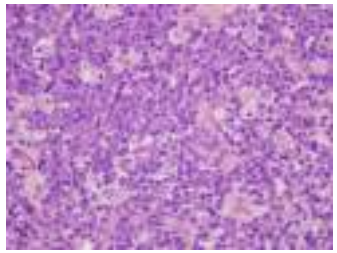
病理切片显微图像

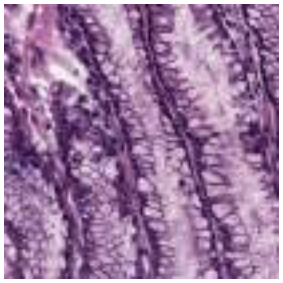


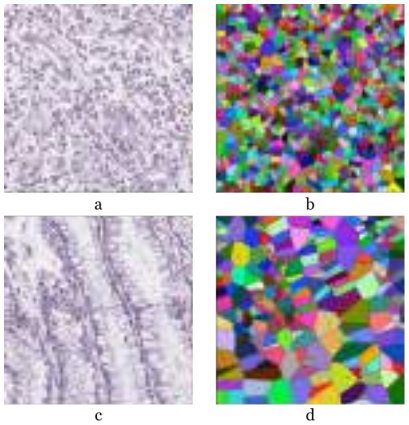




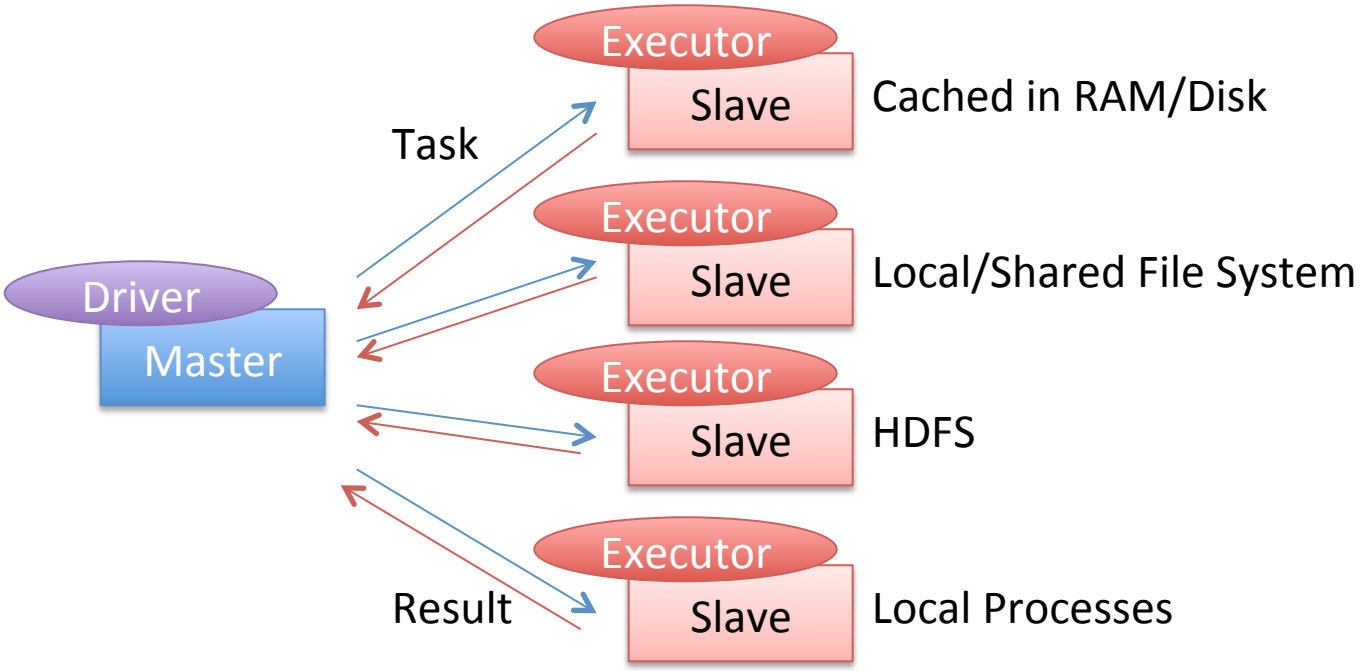
复杂医学影像数据存储与处理







- 单张组织切片图像超过10-20G
- 分割识别需要矩阵运算和复杂处理
- 海量病理图像内容检索



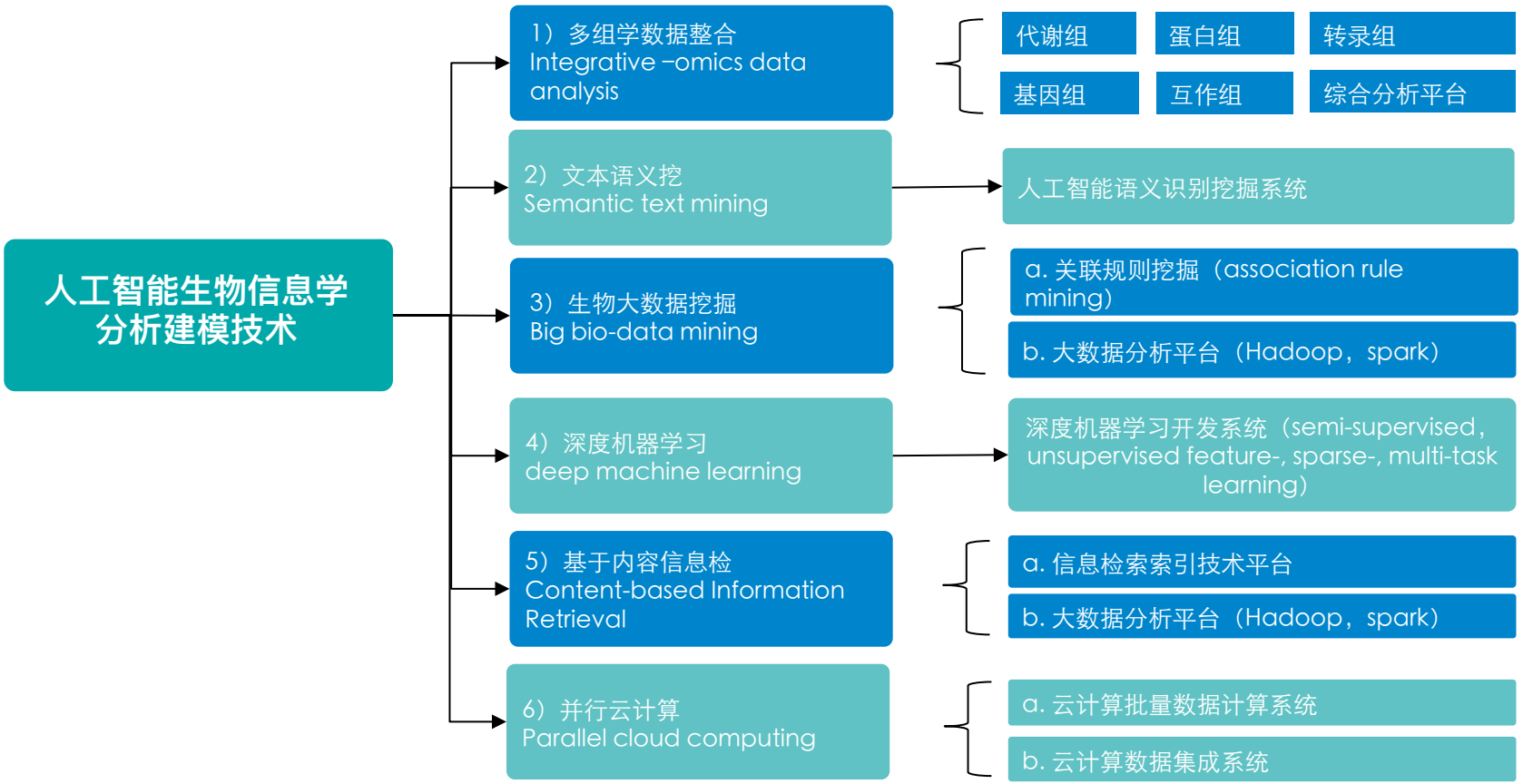
Spark

Scala

OpenCV

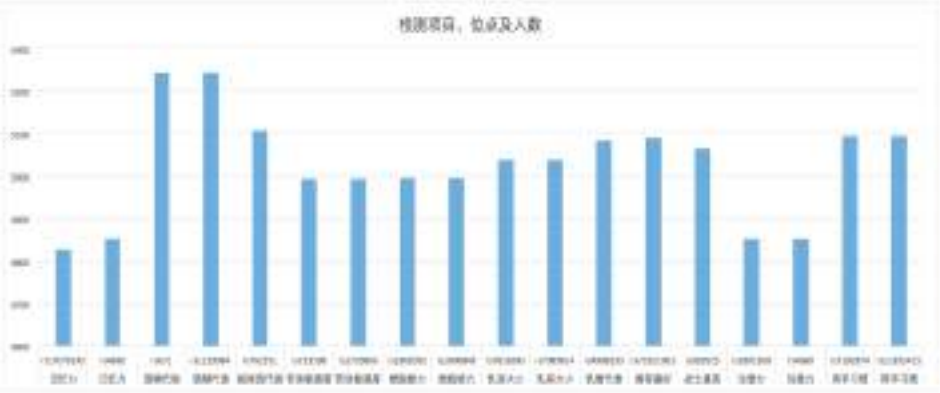
OpenSlide

水母基因在生物数据科学领域核心方法研究

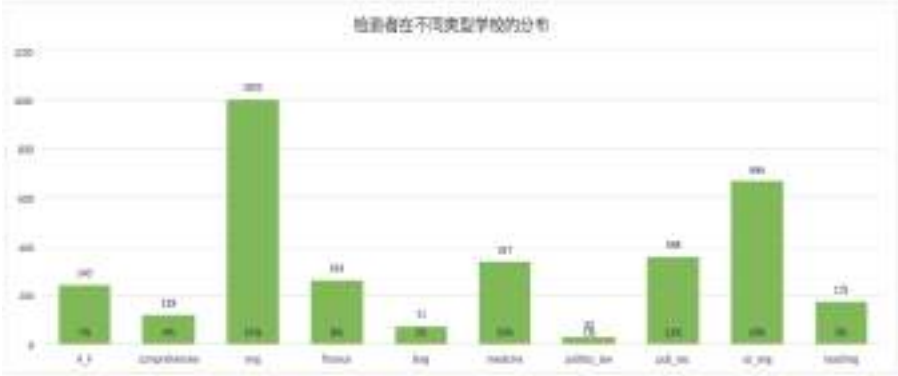


一组在校大学生人群的基因检测分布

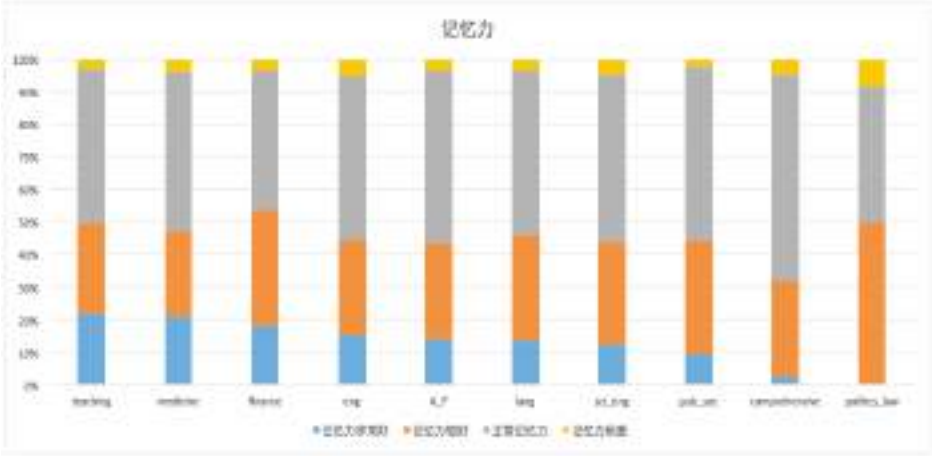
- 上万例样本，11项目17个位点SNP分型



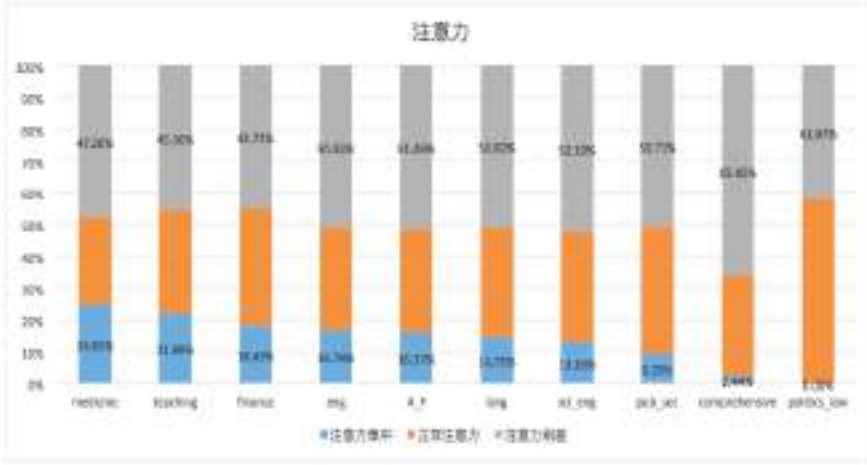
- 分析数据集包括学生检测者来自10种类型高校



- SNP rs4680 A/G分型检测
- 师范类、医药类、财经类学校记忆力非常好的学生占比较大

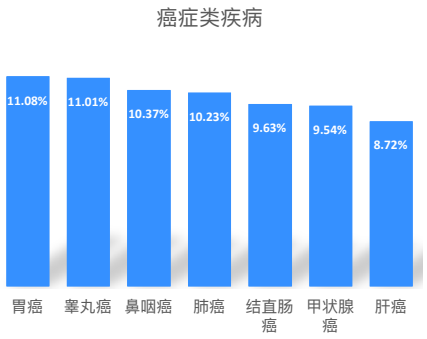


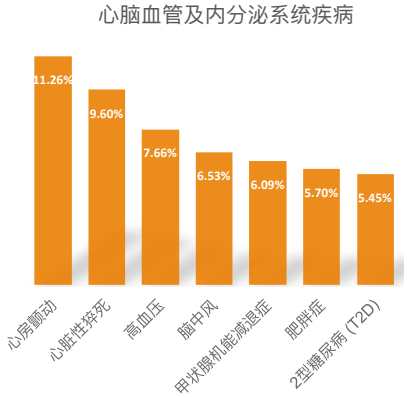
- COMT和CLOCK基因SNP分型检测
- 注意力集中学生占比最高的是医药类大学，其次师范类大学



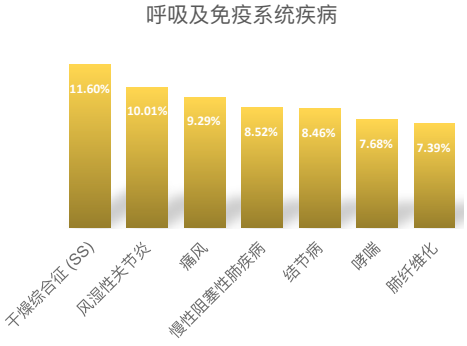
上万例样本精准疾病风险评估模型

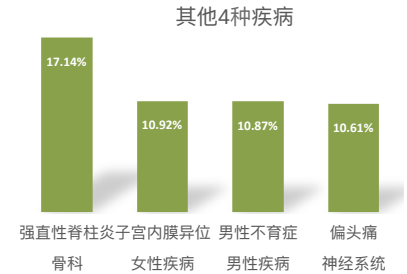
高危人群占比大的4类疾病

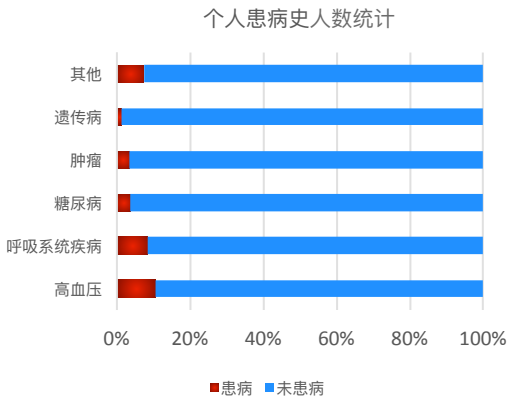


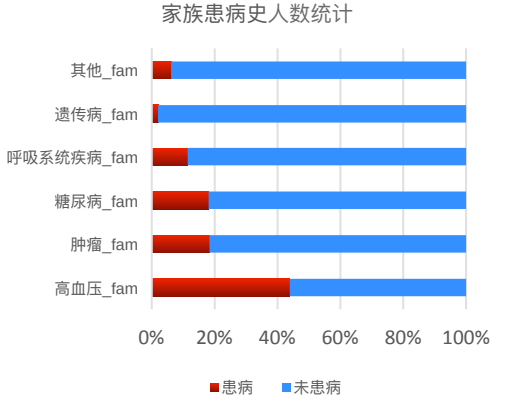


具有调查问卷的基本信息









基因+临床数据整合分析建模：患病风险评估

X 1546基因位点SNP + 30项临床特征向量

Relationship?

Y 7项疾病患病历史特征向量

n 基因编码+临床数据

m 样本

$$\begin{bmatrix} x_{11} & \cdots & x_{1j} & \cdots & x_{1n} \\ \vdots & \ddots & \vdots & & \vdots \\ x_{i1} & \cdots & x_{ij} & \cdots & x_{in} \\ \vdots & & \vdots & \ddots & \vdots \\ x_{m1} & \cdots & x_{mj} & \cdots & x_{mn} \end{bmatrix}$$

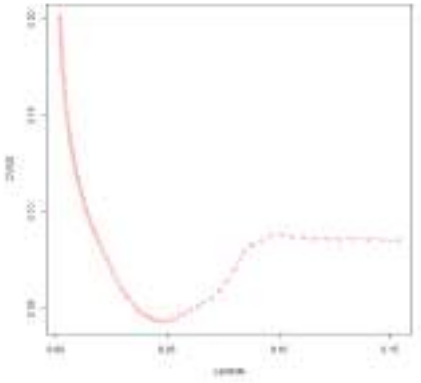
l 项疾病患病史

m 样本

$$\begin{bmatrix} y_{11} & \cdots & y_{1k} & \cdots & y_{1l} \\ \vdots & \ddots & \vdots & & \vdots \\ y_{i1} & \cdots & y_{ik} & \cdots & y_{il} \\ \vdots & & \vdots & \ddots & \vdots \\ y_{m1} & \cdots & y_{mk} & \cdots & y_{ml} \end{bmatrix}$$

group 1 ... Group 3

- 特征提取
  - 连续性变量的离散化、定性变量的重新编码
  - SNP位点数据则根据该位点等位基因频率进行编码
  - y共包括7个变量：高血压病史、糖尿病史、呼吸系统疾病史、肿瘤病史、遗传病史、其他以及无疾病。
- 模型评估结论
  - LASSO模型，优化Lambda得到最小cross validation standard error 0.043



Multi-task sparse learning (LASSO & Ridge)

X → Y

$$E(Y|X=x)=\beta_0+x^T\beta$$

Fit the model by solving:

$$\min \frac{1}{2N}\sum_{i=1}^{N}\left\|y_i-\beta_0-\beta^T x_i\right\|_F^2+\lambda\left[(1-\alpha)\|\beta\|_F^2/2+\alpha\sum_{j=1}^{p}\left\|\beta_j\right\|_2\right]$$

生物大数据整合分析技术在转化医学中的应用

自闭症转录组疾病机制研究与风险因素评估

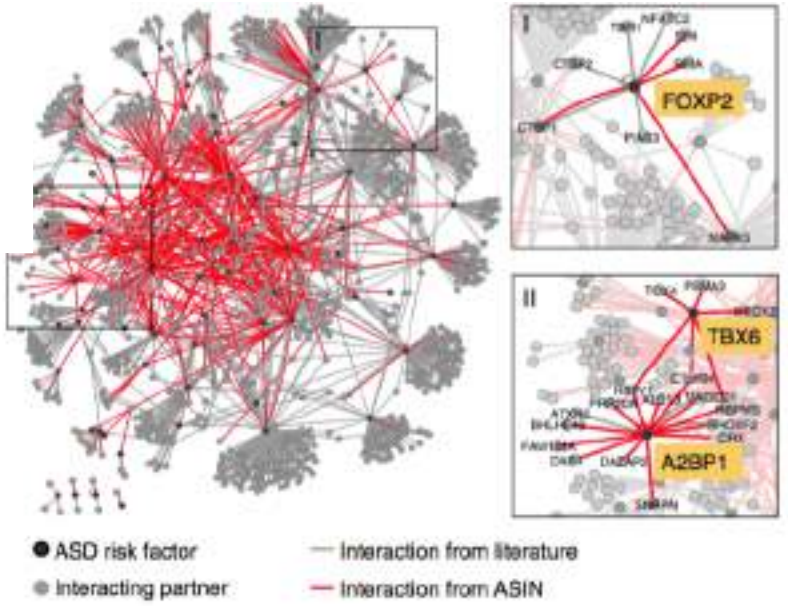
nature COMMUNICATIONS

ARTICLE

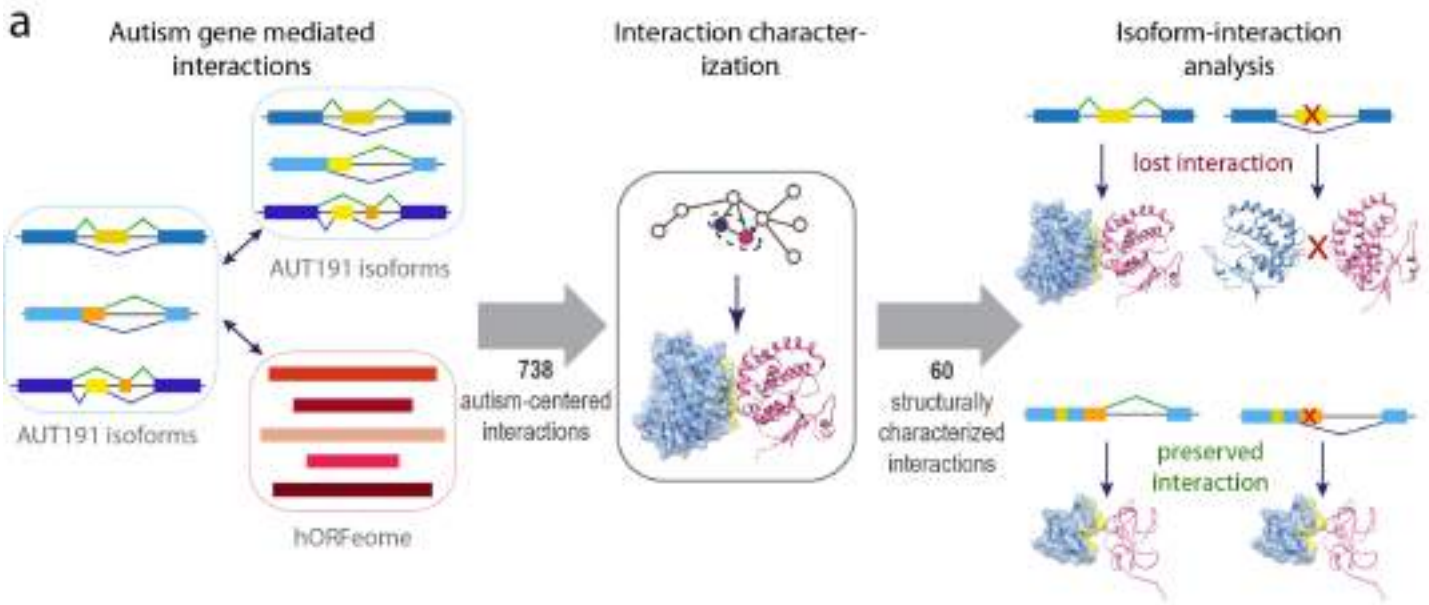
Received 24 Aug 2013 | Accepted 14 Mar 2014 | Published 11 Apr 2014 DOI: 10.1038/ncomms4930 OPEN

Protein interaction network of alternatively spliced isoforms from brain links genetic risk factors for autism

Roser Corominas[1,*], Xinping Yang[2,3,*], Guan Ning Lin[1,*], Shuli Kang[1,*], Yun Shen[2,3], Lila Ghamsari[2,3,†], Martin Broly[2,3], Maria Rodriguez[2,3], Stanley Tam[2,3], Shelly A. Trigg[2,3,†], Changyu Fan[2,3], Song Yi[2,3], Murat Tasan[4], Irma Lemmens[5], Xingyan Kuang[6], Nan Zhao[6], Dheeraj Malhotra[7], Jacob J. Michaelson[7,†], Vladimir Vacic[8], Michael A. Calderwood[2,3], Frederick P. Roth[2,3,4], Jan Tavernier[5], Steve Horvath[9], Kourosh Salehi-Ashtiani[2,3,†], Dmitry Korkin[6], Jonathan Sebat[7], David E. Hill[2,3], Tong Hao[2,3], Marc Vidal[2,3] & Lilia M. Iakoucheva[1]



晶体结构数据整合辅助自闭症转录组大数据建模



水母基因精准健康管理系统

易感基因检测

下一代高通量基因测序技术，精准检测
疾病易感基因

疾病风险预警

基于基因检测与客户所有健康相关数据，准
确预测各类疾病发生风险

精准预防方案

针对高风险疾病制定的精准健康管理方案，
精准预防疾病，避免疾病发生

为健康管理机构提供完整的终端客户健康管理方案

水母基因四大核心功能

精准管理

以基因检测结果为导向，为客户提供更精准的健康管理方案

整合数据

收集终端客户信息，为企业整合海量客户信息和相关数据，为相关服务提供第一手资料

健康档案

为客户建立个人健康档案，一次建档，终生服务

优化服务

通过精准的健康管理，帮助企业完善客户服务，提升服务品质

客户健康档案

- 每位终端客户独有的健康档案
- 明确的客户健康标签
- 可随时更新，收集最新数据
- 一次建档，服务终生

健康评估系统

- 结合基因数据、基础信息、往期评估等，系统自动进行精准评估
- 可定期评估，动态监测明确预警高风险疾病
- 评估报告通俗易懂

全面的基因体检报告

- 囊括100余种疾病
- 全面了解自身疾病的发生风险
- 为生活做全面的指导

详细的疾病信息

- 帮助客户预防疾病
- 优化客户体检方案

水母基因打造基于健康大数据的精准预防

将基因数据个性化的应用于各种生活场景
为各行业合作伙伴提供八大健康生活解决方案。

基因+ 行业解决方案

它不必成为基因专家，基因也能为您的行业提供差异化服务，为行业带来新的利润增长点。

健康管理
以人为本，让基因特质成为新的健康参数，个性化制定健康管理重点。

运动健身
量身定制，让基因分析成为新的服务亮点，量体裁衣地提供健身方案。

教育培训
因材施教，让基因检测成为新的参考依据，帮助客户更好地发现特长。

社交平台
连接你我，让基因标点成为新的交友元素，推出诚实有趣的性格标签。

美容产品
专属保养，让基因性状成为新的保养根据，真实有效地提高服务效果。

智能硬件
智能定义，让基因解读成为新的设计元素，结合不同体质的不同要求。

孕婴行业
护理优化，让检测结果成为新的参考标准，使孕婴产品更加贴心温馨。

营养保健
精准配比，让基因特性成为新的设计标准，解读客户身体的真正需求。

全面的高通量测序平台

领先的生物信息分析平台

自动化样本处理平台

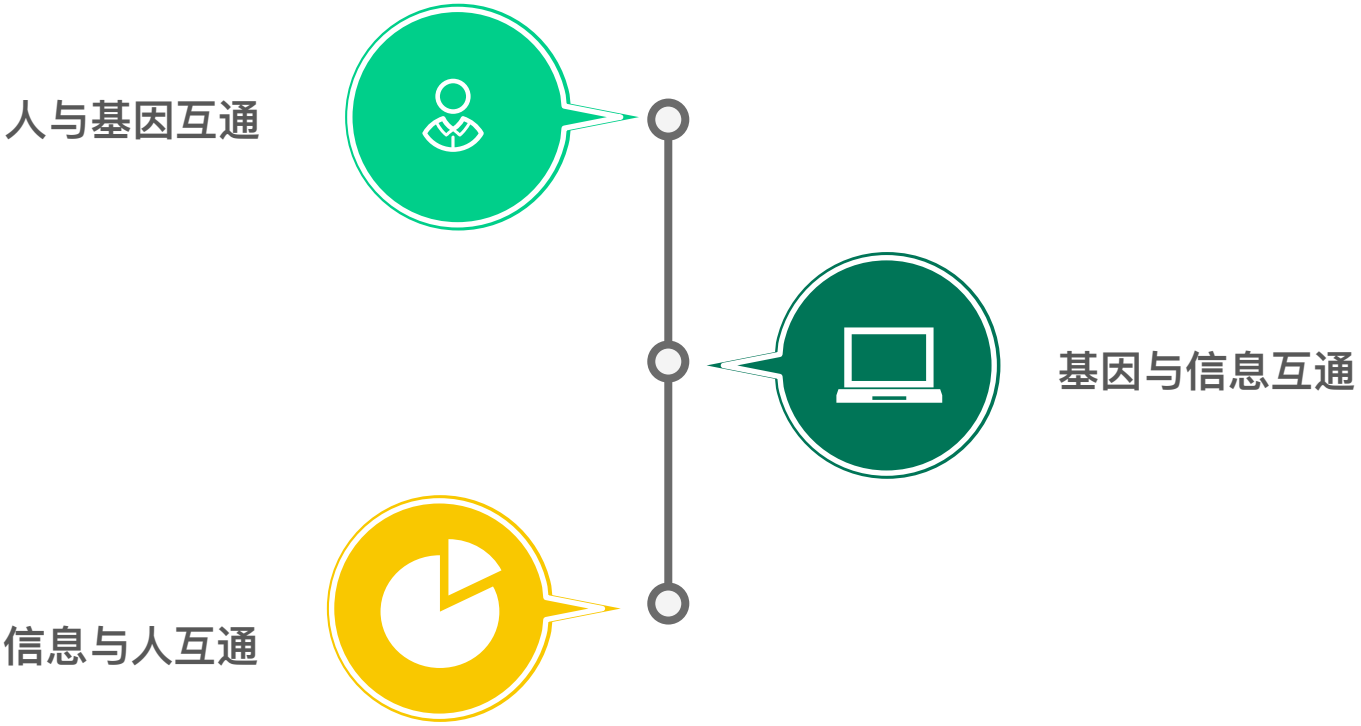
现代化生物样本库

水母基因致力于用基因科技真正改变人们的生活。我们将与各行业的合作伙伴一起，把最新的生物信息、人工智能、大数据等领域的专业知识，融入到各种生活化的应用场景中，为用户提供更精准化的服务。

我们是 水母基因，

邀您共同开启一个更大的基因世界。

我们正在做的事情
用基因科技真正改变人们的生活方式！



水母基因

Thanks