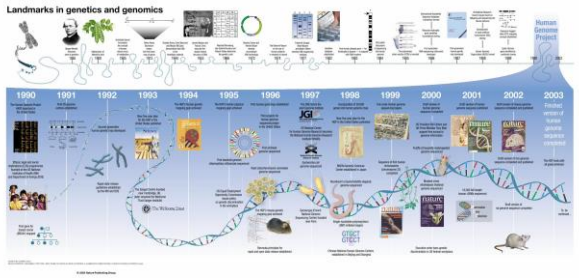
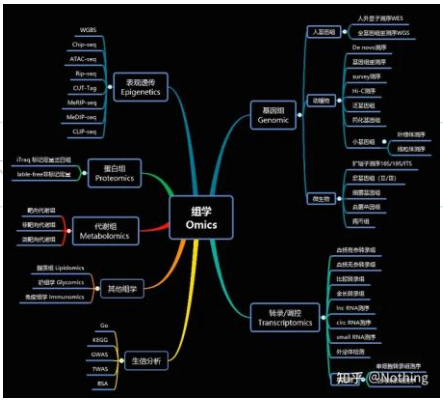
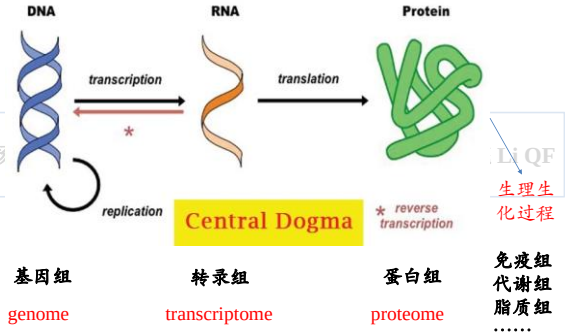
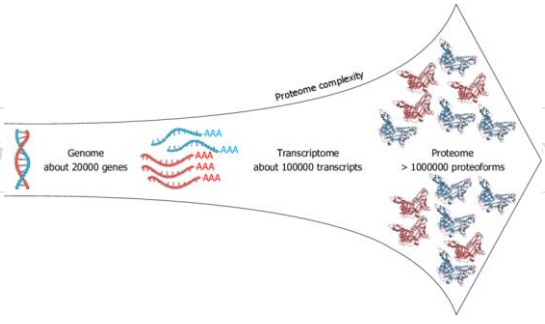
第九章 组学研究技术

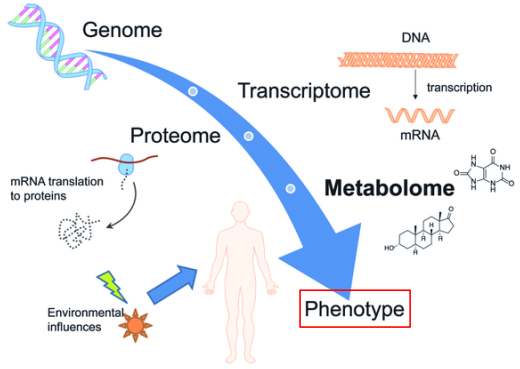


组学(Omics)，就是对某一类分子族群进行研究

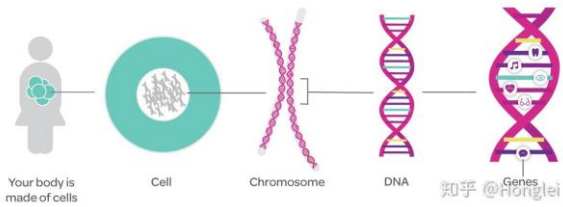






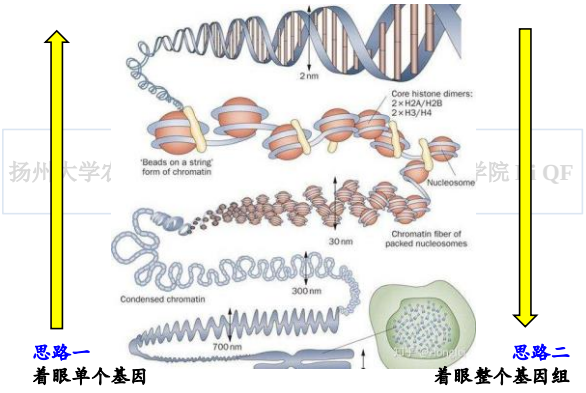


基因组



基因组就是指一个细胞中包含的所有DNA。

我们人类的DNA分布于23对(46条)染色体(Chromosome)上，其中一半来自父亲，一半来自母亲。

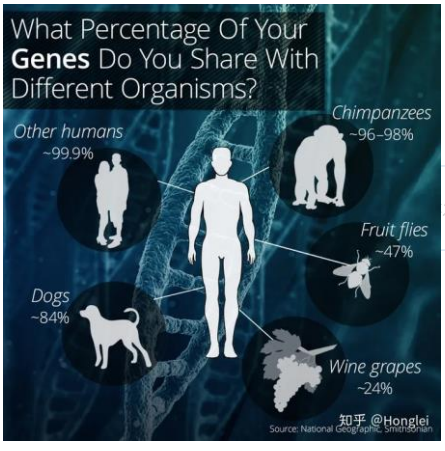


思路一
着眼单个基因

思路二
着眼整个基因组

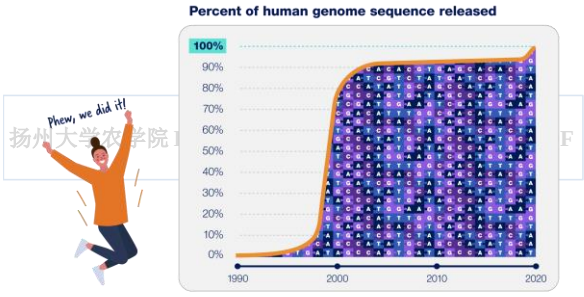
为何要研究基因组?

大约99.5%是和其他人一样的，大约有0.5%的DNA是你自己独有的。而正是这些不同的DNA部分，决定着你的性别，身高，眼睛和头发的颜色等等性状，让你成为独一无二的你。

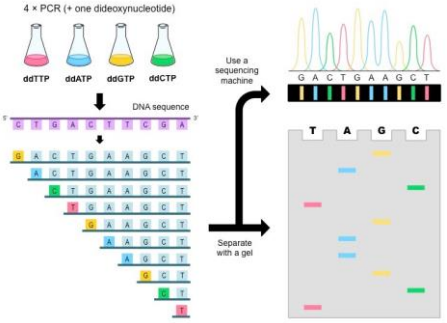


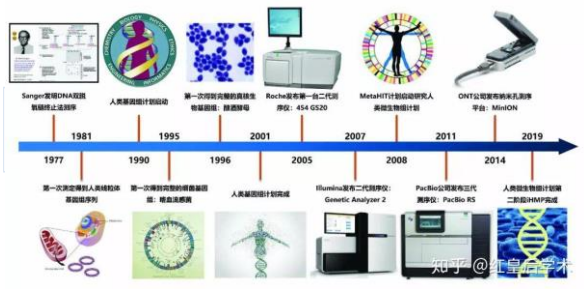
和其他动物的DNA序列有着很高的相似性，这既支持了达尔文进化论的观点，又使得人类可以通过动物实验，研究动物的基因组来开发相应的药物，治疗人类疾病。

如何研究基因组?



Sanger测序

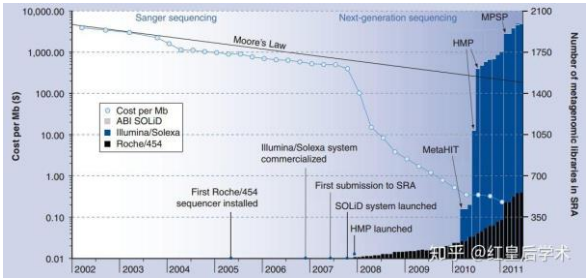




Sanger所发明的测序方法被称为**第一代测序技术**，该技术直到现在依然被广泛使用，但是其**一次只能获得一条长度在700～1000个碱基的序列**，无法满足现代科学发展对生物基因序列获取的迫切需求。

高通量测序 (High-Throughput Sequencing, HTS) 一次运行即可同时得到几十万到几百万条核酸分子的序列，因此也被称为新一代测序 (Next Generation Sequencing, NGS)或**第二代测序**。其获得**单条序列长度很短**，想要得到准确的基因序列信息依赖于较高的测序覆盖度和准确的序列拼接技术，因此最终得到的结果中会存在一定的错误信息。

第三代测序技术也称为单分子测序技术，该技术在保证测序通量的基础上，对单条长序列进行从头测序，能够直接得到长度在**数万个碱基**的核酸序列信息。



基因测序技术成本变化

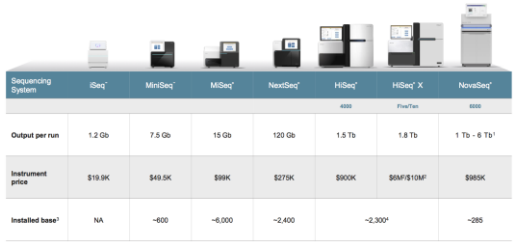
在2008年，全基因组测序的成本降至20万美元；到2010年，该费用已经可以控制在10000美元以内；目前，测定一个人类的全基因组只需要不到1000美元即可完成。

目前成熟的第二代测序技术共有3种

- Roche公司的454技术
- ABI公司的SOLiD技术
- Illumina公司的Solexa技术

Sequencing Power for Every Scale

The broadest portfolio offering available



Sequencing System	iSeq™	MiniSeq™	MiSeq™	NextSeq™	HiSeq™	HiSeq™ X	NovaSeq™
					4000	Five/Ten	6000
Output per run	1.2 Gb	7.5 Gb	15 Gb	120 Gb	1.5 Tb	1.8 Tb	1 Tb - 6 Tb[1]
Instrument price	$19.9K	$49.5K	$99K	$275K	$900K	$6M/$10M[2]	$985K
Installed base[3]	NA	~600	~6,000	~2,400	~2,300[4]		~285

1. Output per run for the S1, S2 and S4 flow cells equal 1 Tb, 2 Tb and 6 Tb, respectively assuming two flow cells per run
2. Based on purchase of 5 and 10 units for HiSeq X Five and HiSeq X Ten, respectively
3. Based on end of fiscal year 2017
4. Combined HiSeq family

illumina

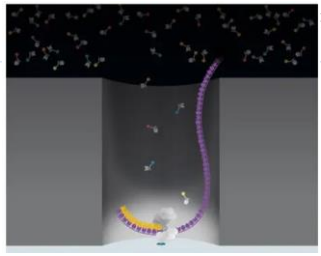
第三代测序技术

- PacBio公司的SMRT技术
- Oxford Nanopore Technologies公司的纳米孔单分子技术

PacBio

每个ZWM孔只允许一条DNA模板进入，然后DNA聚合酶与模板结合，加入4种不同颜色荧光标记4种dNTP，其通过布朗运动随机进入检测区域并与聚合酶结合从而延伸模板，与模板匹配的碱基生成化学键的时间远远长于其他碱基停留的时间，因此统计荧光信号存在时间的长短，可区分匹配的碱基与游离碱基。通过统计4种荧光信号与时间的关系，即可测定DNA模板序列。

SMRT芯片是一种带有很多ZMW孔的厚度为100nm的金属片，将DNA聚合酶、待测序列和不同荧光标记的dNTP放入ZMW孔的底部。



基因组从头测序

基因组从头测序即（*de novo* 测序），在无任何参考序列资料下对某个物种进行测序，从而获得该物种的基因组序列图谱。

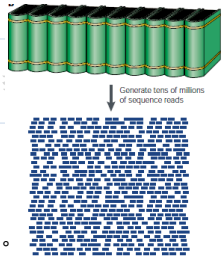
基因组从头测序适用范围：

- ✓ 物种的基因组序列完全未知；
- ✓ 微生物基因组，基因组的可塑性非常大，如细菌、真菌等；
- ✓ 基因组序列与已知基因组序列有较大的变异，如Cancer Cell；
- ✓ 物种基因组序列已知，拼接效果不理想。

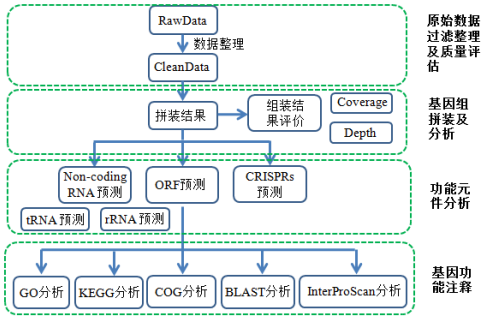
基因组从头测序-文库构建

文库：一定大小范围的片段的集合。

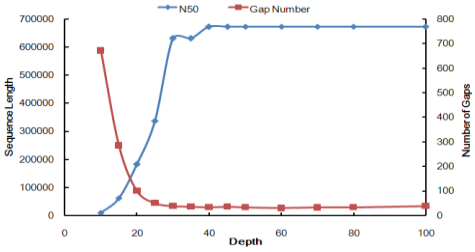
- ✓ Illumina Paired-End 文库；
- ✓ Illumina Mate-Paired 文库；
- ✓ 454 Shotgun 文库；
- ✓ 454 Shotgun/Paired-End 文库。



基因组从头测序信息分析流程



基因组拼装的影响因素—测序深度



- ✓ 测序深度并非越深越好，投入和效果并非成正比；
- ✓ 测序深度达到一定值后，拼接效果呈现饱和现象。

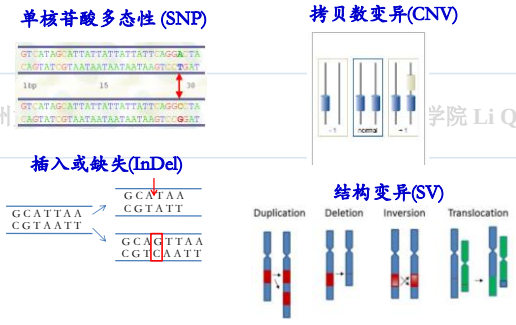
基因组重测序

基因组重测序 (Genome Resequencing) 是指对基因组序列已知的物种个体进行基因组测序，从而获得该物种个体或群体的差异的测序方法

基因组重测序的适用范围:

- ✓ 该物种的基因组序列已知
- ✓ 待测序个体与已知物种的基因组序列无显著的结构性差异

基因组变异类型



泛基因组(Pan-genome)

泛基因组(Pan-genome)：**某一物种全部基因的总称**，这里这个全部基因是有别于个体基因组的基因。

以一个基因组为模板的分析不能全面反应物种基因水平的全部遗传信息，尤其是同一物种中差异巨大的不同亚种或者变种，此类特有片段的差异往往比共有片段中的差异更重要。

核心基因（core genome）：在所有动植物品系或者菌株中都存在的基因；

非核心基因（dispensable/ variable/ accessory/genome）：在1个以及1个以上的动植物品系或者菌株中存在的基因。

分析核心基因和非核心基因的基本情况，并从特有基因序列的角度来研究物种内的差异。

宏基因组 (Metagenome)

特定环境下所有生物遗传物质的总和，包含了可培养的和未可培养的微生物的基因。

一般从环境样品中提取基因组DNA, 进行高通量测序，从而分析微生物多样性、种群结构、功能信息、与环境之间的关系等。

第二节、转录组学分析

转录组（transcriptome）
广义：指某一生理条件下，细胞内所有转录产物的集合，包括信使RNA、核糖体RNA、转运RNA及非编码RNA；
狭义：指所有mRNA的集合。

转录组学（transcriptomics）是指一门在整体水平上研究细胞中基因转录的情况及转录调控规律的学科。

无参考基因组的转录组分析　有参考基因组的转录组分析

无参考基因组的转录组分析

- **转录组拼接**
  将Reads拼接成为转录本。
- **转录本注释**
  通过blast近缘物种或Nr库注释转录本。
  Blast二级数据，进行注释
- **Unigene聚类**
  通过注释结果进行Unigene聚类
- **基因表达差异**
  表达量计算用RPKM。
  样本间每个基因的表达差异分析。
- **ORF分析**
  描述样本中可能存在的不同剪切形式。
- **cSNP分析**
  发现样本中存在的SNP。
- **SSR分析**

有参考基因组的转录组分析

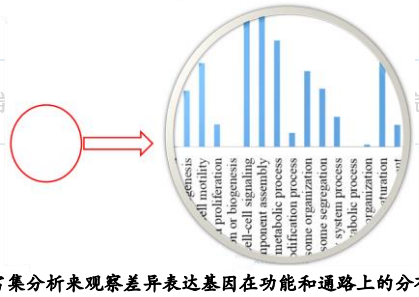
- **基因组整理**
  定位注释、分类注释、功能注释。
- **比对到基因组**
  把来自于DNA的序列还原到其应该在的位置
- **基因表达差异**
  表达量计算用RPKM或其他统计学方式表示。
  样本间每个基因的表达差异分析。
- **功能富集分析**
  通过富集分析，找出有差异的通路或者功能
- **功能聚类分析**
  通过聚类分析，归类表达模式的基因或样品
- **结构分析**
  UTR、Alternative Splicing、Fusion Gene、New Transcript 分析及可视化

- **cSNP分析**
  发现样本中存在的SNP

相关的数据库网站

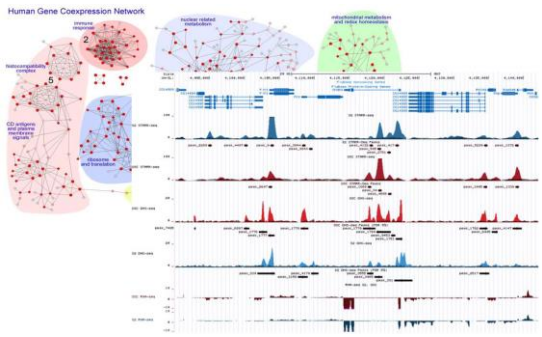
如何分析两组表达数据之间的差异?

富集分析

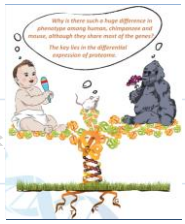


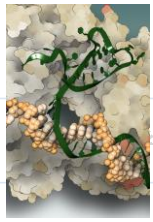
用富集分析来观察差异表达基因在功能和通路上的分布，看看集中在哪些功能和通路，从而得知生物体的整体状态!

共表达分析



第三节、蛋白组学分析





蛋白质组（proteome）指一种基因组所表达的全套蛋白质，即包括一种细胞乃至一种生物所表达的全部蛋白质；

蛋白质组学（proteomics）指系统研究某一基因组所表达的所有蛋白质，包括组成蛋白质一级结构的氨基酸序列，蛋白质的丰度，蛋白质的修饰以及蛋白质间的相互作用。

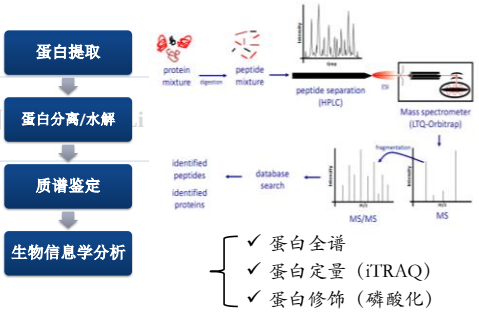
一、蛋白质组研究内容

- ✓ 全谱分析：定性，有什么蛋白？
- ✓ 差异分析：定量，哪些蛋白有或无，上调或下调？
- ✓ 修饰分析：磷酸化？糖基化？乙酰化？…

如何获取蛋白质信息

- 蛋白质信息？（太复杂）
- 提取---分离---酶解
- 肽段信息（简单，质谱技术）
- m/z ⟷ 肽段序列/修饰信息
- 蛋白质数据库
- 搜索匹配

二、蛋白质组学分析的工作流程



- ✓ 蛋白全谱
- ✓ 蛋白定量（iTRAQ）
- ✓ 蛋白修饰（磷酸化）

三、蛋白质组研究-蛋白全谱分析

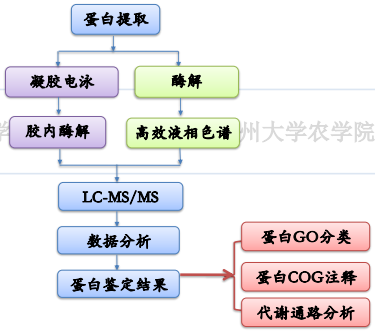
目的：

识别研究对象中尽可能多的肽和蛋白质混合物的组分。

应用领域：

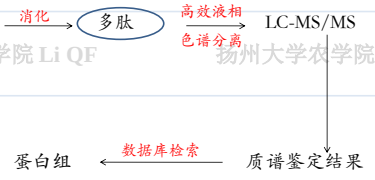
- ✓ 生长发育不同阶段蛋白的整体表达情况
- ✓ 不同组织中蛋白的整体表达情况
- ✓ 特殊生理状态或病害中蛋白的整体应答规律

3.1 蛋白全谱分析—工作流程



3.2 蛋白全谱分析—
基于凝胶电泳的分离

3.3 蛋白全谱分析—
基于高效液相色谱的分离



3.4 蛋白全谱分析—
SDS-PAGE *vs* HPLC

	SDS-PAGE	HPLC
分离	Protein level	Peptide level
消化	胶内	溶液
适用范围	分子量小于10KD或大于100KD的蛋白、低丰度蛋白、疏水蛋白（如膜蛋白），可能会检测不到。	较高丰度蛋白的存在，可能会影响低丰度蛋白的检测；柱外效应。

四、蛋白质组研究-蛋白定量分析

目的：

比较两个或多个不同的蛋白质组，以确定一个已知的刺激或疾病不同阶段蛋白表达水平的差异。

应用领域：

- ✓ 寻找功能蛋白
- ✓ 探索发育过程及代谢调控机制
- ✓ 抗逆机理的研究
- ✓ 生物标志物或特殊蛋白的筛选

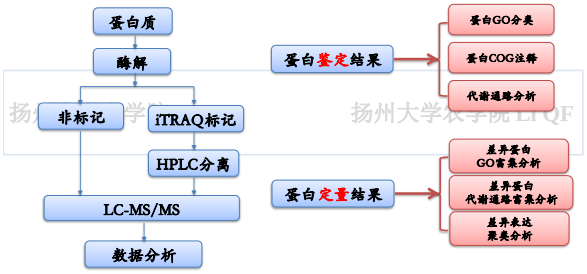
4.1 蛋白定量分析的策略

➢ Labeled 蛋白定量

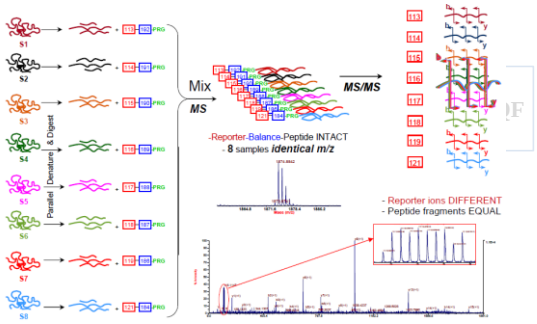
- MS based：ICAT, SILAC…
- MS/MS based：iTRAQ（isobaric tag for relative and absolute quantitation）

➢ Label free （不加标签）蛋白定量

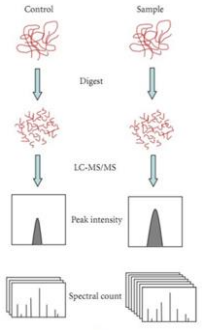
4.2 蛋白定量分析-工作流程



4.3 iTRAQ蛋白定量技术路线（多样本同时进入质谱仪）



4.4 Label-free蛋白定量技术路线



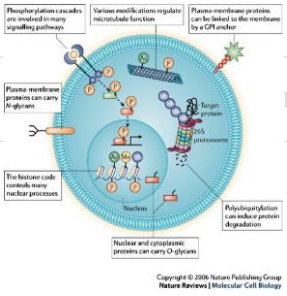
- ✓ 基于峰面积和峰数量（加内标）判断表达量的技术；
- ✓ 对质谱仪稳定性和分辨率、精度和灵敏度等要求高（由于不同的样本在不同批次进入质谱仪）；

4.5 蛋白定量分析—Label-free *vs* iTRAQ

	准确性	实验复杂度	可比较样品数	完成周期	价格
iTRAQ	好	复杂	一次8个	短	便宜
Label-free	一般	简单	很多	长	一般

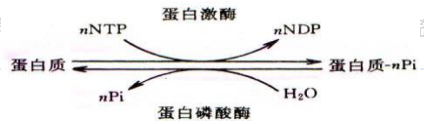
五、蛋白质组研究-蛋白修饰分析

蛋白翻译后修饰（post-translational modification，PTM）由于功能需要，蛋白质翻译后在特异氨基酸链上共价结合各种非肽类基团，如磷酸化、糖基化、泛素化等。



5.1 蛋白修饰分析-技术背景

蛋白质的磷酸化是指由蛋白质激酶催化的把ATP或GTP上γ位的磷酸基转移到底物蛋白质氨基酸残基上的过程，其逆转过程是由蛋白质磷酸酶催化的，称为蛋白质去磷酸化。



- ✓ 一种可逆的共价修饰方式。
- ✓ 是最常见修饰方式，哺乳动物中1/3的蛋白在细胞周期中会发生。
- ✓ 蛋白的磷酸化与去磷酸化参与到几乎所有的生命过程。

5.2 蛋白修饰分析的目的和应用

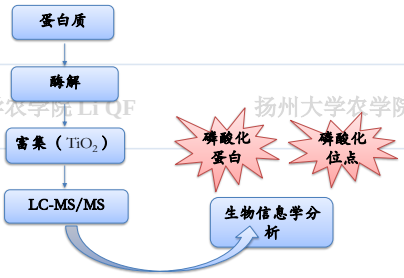
目的：

整体全面的找出相应组织中的蛋白磷酸化修饰谱，鉴定磷酸化蛋白及相应的磷酸化位点。

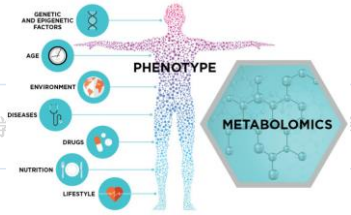
应用领域：

- ✓ 鉴定新的磷酸化蛋白
- ✓ 寻找新的磷酸化修饰位点
- ✓ 组蛋白修饰研究
- ✓ 信号转导途径分析研究

5.3 蛋白修饰分析-工作流程



第四节、代谢组学分析



代谢组（metabolome）指某一生物或细胞、组织在特定生理时期内的所有小分子代谢物（一般<1200 Da）集合称为代谢组；

代谢组学（metabonomics）则是对代谢组进行定性和定量分析，并研究该代谢组在干预或疾病生理条件下动态变化规律。

代谢组学检测的技术平台

GC-MS　GC-MS，GC×GC-TOF-MS，GC×GC—Quadropole-MS

平台： Agilent7890/5975®**气-质联用仪系统**

LC-MS　LC-MS，LC×LC-MS，LC-CE-MS，LC-NMR-MS，HPLC-TOF/MS，HPLC-ESI/MS/MS，UPLC-TOF/MS

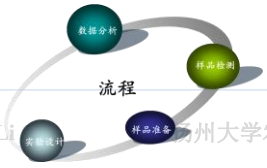
平台： Waters ACQUITY UPLC**超效液相色谱仪和**Micromass Q-TOF Premier **质谱联用仪**

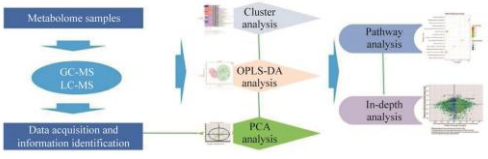
NMR　氢谱（1H NMR）、碳谱（13C NMR）及磷谱（31P NMR）

HR-MAS NMR（高分辨魔角旋转**核磁共振波谱法**），HPLC-DAD-MS-SPE-NMR

平台： Varian Inova600MHz®（配置超低温探头）NMR系统， Bruker 600MHz ® NMR系统

代谢组分析-工作流程





不同技术平台的应用

化合物分类及其最适的分析技术

不同技术平台优缺点

NMR

优点： 无损伤性，无辐射性，无偏向性，方法灵活，处理简单，成本低

缺点： 灵敏度较低，动态范围有限

GC-MS

优点： 较好的分离效率和检测灵敏度，易定性

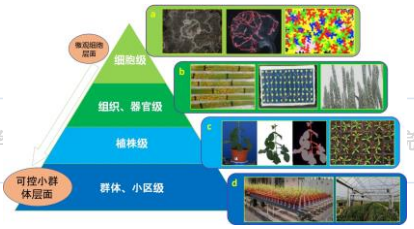
缺点： 衍生化限制了应用范围，具有偏向性

LC-MS

优点： 灵敏度较高，无需衍生化

缺点： 分离率不高，时间较长，具有偏向性，不易定性

第五节、表型组学分析

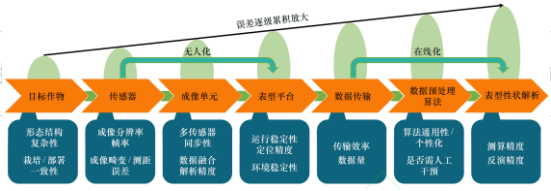


表型组（phenome）指某一生物的全部性状特征，即生物体从微观（即分子）组成到宏观、从胚胎发育到衰老死亡全过程中所有表型的集合。

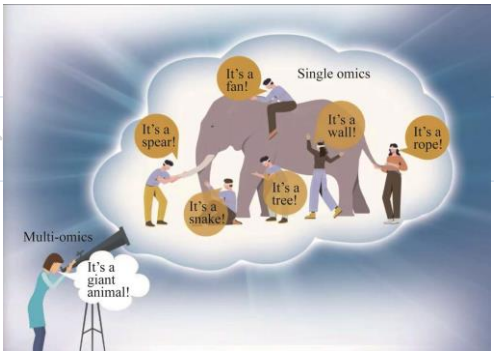
表型组学（phenomics）是一门在基因组水平上系统研究某一生物或细胞在各种不同环境条件下所有表型的学科。

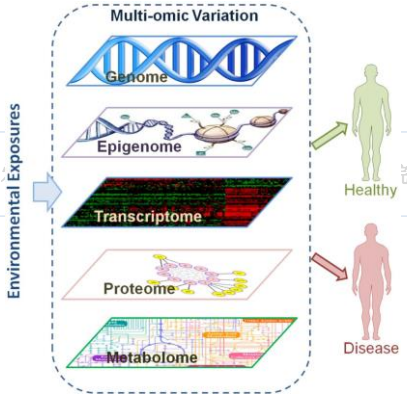
一、表型组大数据技术及装备从研发到应用路线图

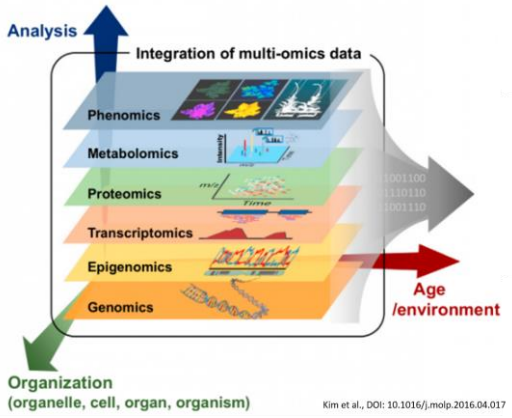
二、表型组大数据技术及装备实用性分析



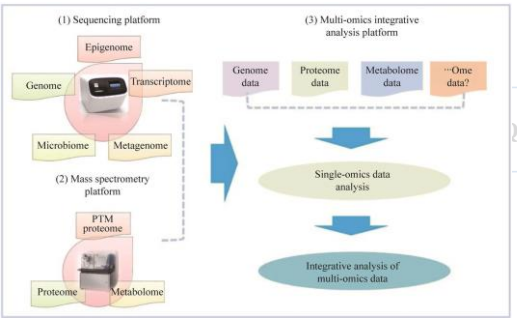
第六节、单组学和多组学的关系



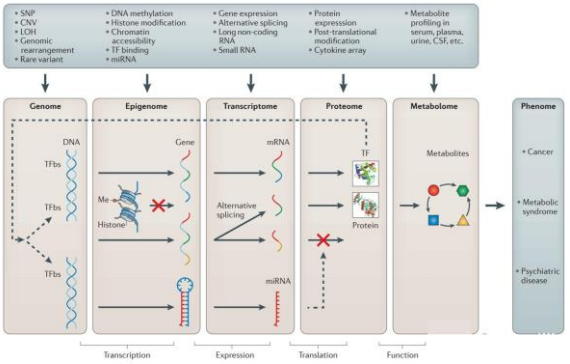




多组学技术平台的构建



多组学数据的关联分析



多组学数据库

Table 1 List of multi-omics repositories[50]

Data repository	Web link	Disease	Types of multi-omics data available
The cancer genome atlas (TCGA)	https://cancergenome.nih.gov/	Cancer	RNA-seq, DNA-seq, miRNA-seq, SNV, CNV, DNA methylation, and RPPA
Clinical proteomic tumor analysis	https://cptac-data-portal.georgetown.edu/cptacPublic/	Cancer	Proteomics data corresponding to TCGA cohorts
International cancer genomics consortium (ICGC)	https://icgc.org/	Cancer	Whole genome sequencing, genomic variations data (somatic and germline mutation)
Cancer cell line encyclopedia (CCLE)	https://portals.broadinstitute.org/ccle	Cancer cell line	Gene expression, copy number, and sequencing data; pharmacological profiles of 24 anticancer drugs
Molecular taxonomy of breast cancer international consortium (METABRIC)	http://molonc.bccrc.ca/aparicio-lab/research/ metabric/	Breast cancer	Clinical traits, gene expression, SNP, and CNV
TARGET	https://ocg.cancer.gov/programs/target	Pediatric cancers	Gene expression, miRNA expression, copy number, and sequencing data
Omics discovery index	https://www.omicsdi.org	Consolidated data sets from 11 repositories in a uniform framework	Genomics, transcriptomics, proteomics, and metabolomics

CNV: copy number variation; miRNA: microRNA; RPPA: reverse phase protein array; SNP: single-nucleotide polymorphism; SNV: single-nucleotide variant.