

# 系统生物学

天津医科大学  
生物医学工程与技术学院

2017-2018 学年上学期（秋）  
研究生

# 自我介绍

姓 名 伊现富 (Yi Xianfu)

本 科 山东大学

硕 博 中国科学院

工作邮箱 [yixfbio@gmail.com](mailto:yixfbio@gmail.com)

生活邮箱 [yixf1986@gmail.com](mailto:yixf1986@gmail.com)

手 机 [156 2061 0763](tel:15620610763)

个人博客 <http://yixf.name>

@GitHub <https://github.com/Yixf-Education>

网络昵称 yixf, Yixf



# 授课资料



[https://github.com/Yixf-Education/course\\_System\\_Biology](https://github.com/Yixf-Education/course_System_Biology)



# 课程安排

日期	授课内容	学时	授课方式	授课教师
11.23	高通量测序技术及数据分析	3	理论授课	伊现富
11.30	蛋白质组学	3	理论授课	乔海晅
12.07	系统生物学的建模和仿真	3	理论授课	王举
12.14	生物大分子的药物肿瘤治疗	3	理论授课	高秀军
12.21	蛋白质网络分析、蛋白质分子模拟	3	理论授课	张涛
12.28	系统生物学相关应用	3	课堂讨论	张涛



# 教学提纲

1

## 系统生物学

2

## 基因组学

3

## 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

## 数据库与数据格式

5

## 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

## 外显子组测序

## ● 简介

## ● 操作流程

## ● 应用实例

## 7 转录组学

## 8 RNA-Seq

## ● 概述

## ● 数据分析

## ● 应用实例

## 9 顺反组

## ● 概述

## ● ChIP-Seq

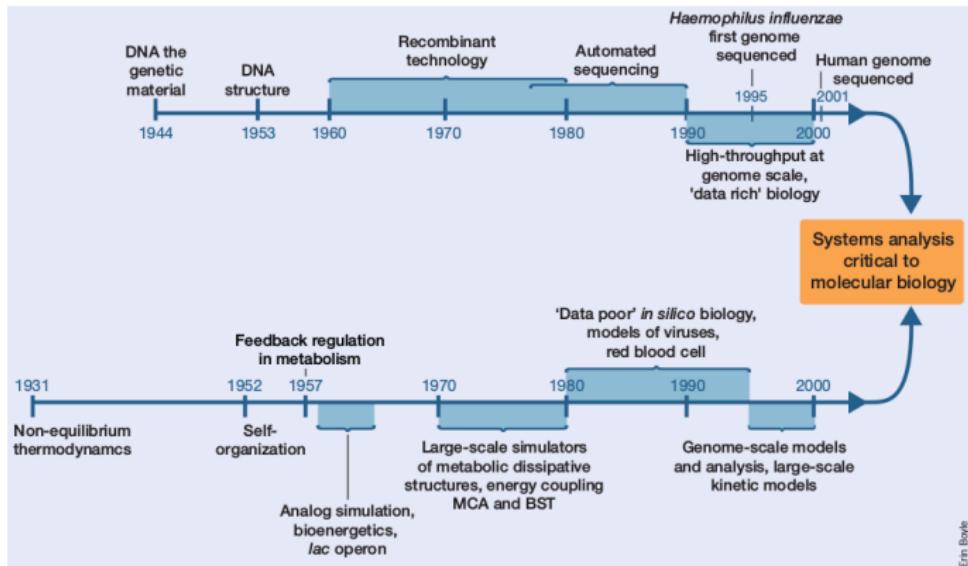
## 10 表观遗传学

## ● 概述

## ● Methyl-Seq



# 系统生物学 | 历史



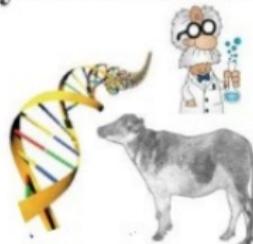
- 2000 年, 第一届国际系统生物学会 (1st International Conference on Systems Biology ; ICSB 2000)
- 2000 年, 莱诺伊·胡德、阿兰·阿德雷姆及鲁迪·艾伯索尔德, 系统生物学研究所 (Institute for Systems Biology)

## Molecular Biology vs. Systems Biology

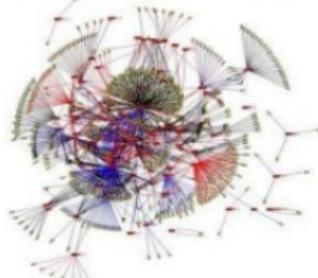
■ Molecular biology -  
biomolecule structure and  
function is studied at the  
molecular level

■ Systems biology -  
specific interactions of  
components in the  
biological system are  
studied -  
cells, tissues, organs, and  
ecological webs

- Integrative approach in which scientists study pathways and networks, will touch all areas of biology, including drug discovery



Era of Molecular Biology (1953 – 2001)

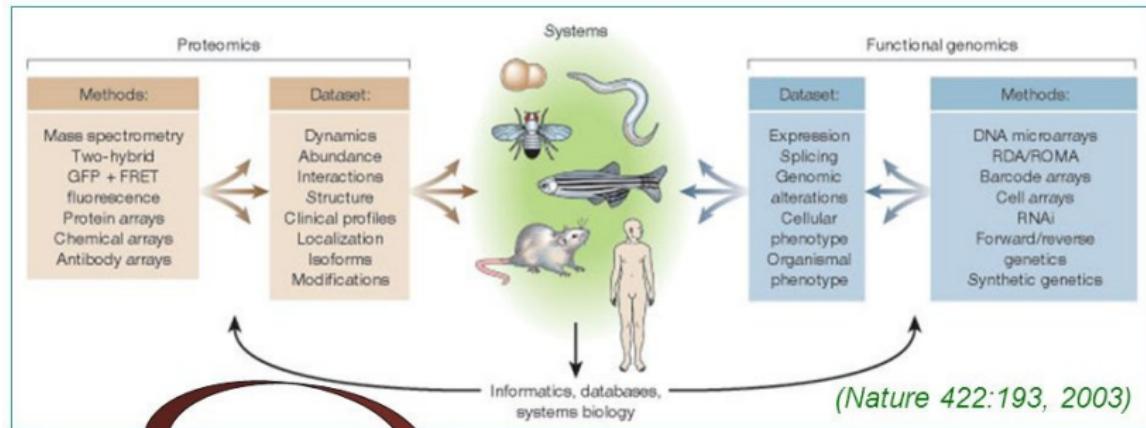


Era of Systems Biology (2001 – ??)

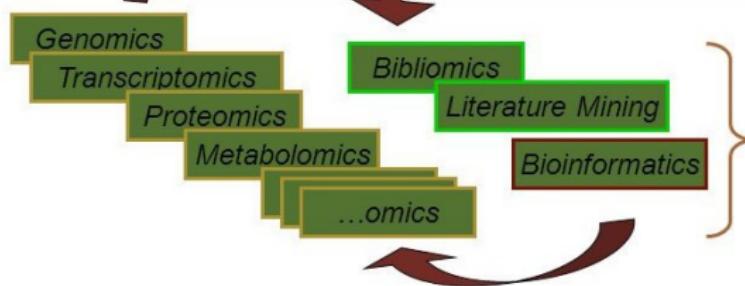
Molecular ⇒ System

Systems biology is a natural extension of molecular biology, and can be defined as “biology after the identification of key genes”.

系统生物学 | 历史 | 组学 → 系统生物学



Integrated knowledge  
and tools are needed  
for Systems Biology's  
research



# 系统生物学 | 定义

莱诺伊·胡德, 2001

系统生物学是将 DNA、RNA、蛋白质以及三者彼此之间的交互作用等信息加以整合，并运用这些资料去建立出数学计量模型，以期能掌握所有生物基因与组织间的关系及运作。

Hood, 2004

系统生物学是研究一个生物系统中所有组成成分（基因、mRNA、蛋白质等）的构成，以及在特定条件下这些组分间的相互关系，并通过计算生物学建立一个数学模型来定量描述和预测生物功能、表型和行为的学科。

杨胜利, 2004

系统生物学是在细胞、组织、器官和生物体水平上研究结构和功能各异的生物分子及其相互作用，并通过计算生物学定量阐明和预测生物功能、表型和行为。系统生物学将在基因组测序基础上完成 DNA 序列到生命的过程，这是逐步整合、优化的过程，系统生物学的发展预计需要一个世纪或更长的时期，因此常把系统生物学称为 21 世纪的生物学。

# 系统生物学 | 定义

莱诺伊·胡德, 2001

系统生物学是将 DNA、RNA、蛋白质以及三者彼此之间的交互作用等信息加以整合，并运用这些资料去建立出数学计量模型，以期能掌握所有生物基因与组织间的关系及运作。

Hood, 2004

系统生物学是研究一个生物系统中所有组成成分（基因、mRNA、蛋白质等）的构成，以及在特定条件下这些组分间的相互关系，并通过计算生物学建立一个数学模型来定量描述和预测生物功能、表型和行为的学科。

杨胜利, 2004

系统生物学是在细胞、组织、器官和生物体水平上研究结构和功能各异的生物分子及其相互作用，并通过计算生物学定量阐明和预测生物功能、表型和行为。系统生物学将在基因组测序基础上完成 DNA 序列到生命的过程，这是逐步整合、优化的过程，系统生物学的发展预计需要一个世纪或更长的时期，因此常把系统生物学称为 21 世纪的生物学。

# 系统生物学 | 定义

莱诺伊·胡德, 2001

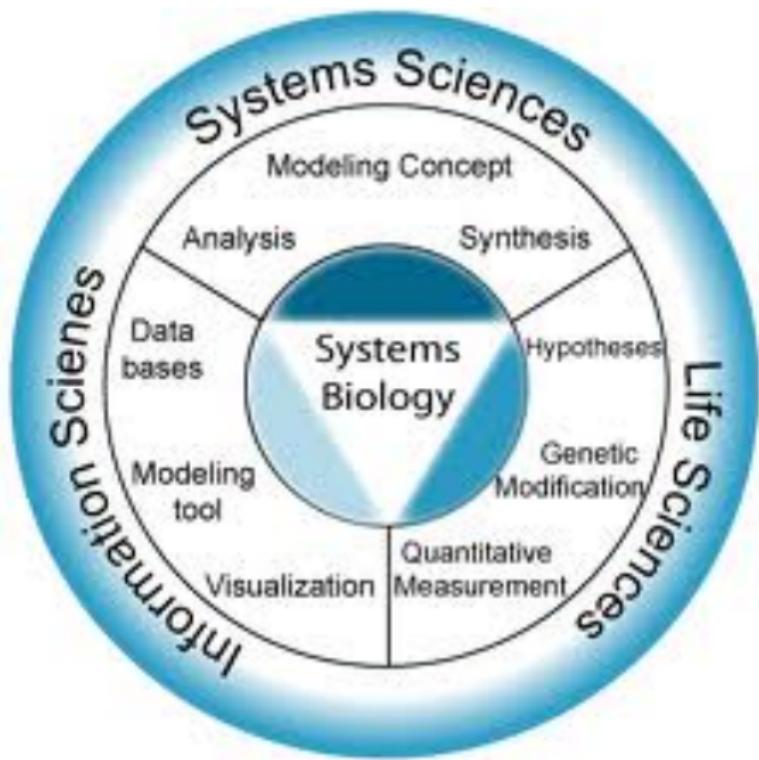
系统生物学是将 DNA、RNA、蛋白质以及三者彼此之间的交互作用等信息加以整合，并运用这些资料去建立出数学计量模型，以期能掌握所有生物基因与组织间的关系及运作。

Hood, 2004

系统生物学是研究一个生物系统中所有组成成分（基因、mRNA、蛋白质等）的构成，以及在特定条件下这些组分间的相互关系，并通过计算生物学建立一个数学模型来定量描述和预测生物功能、表型和行为的学科。

杨胜利, 2004

系统生物学是在细胞、组织、器官和生物体水平上研究结构和功能各异的生物分子及其相互作用，并通过计算生物学定量阐明和预测生物功能、表型和行为。系统生物学将在基因组测序基础上完成 DNA 序列到生命的过程，这是逐步整合、优化的过程，系统生物学的发展预计需要一个世纪或更长的时期，因此常把系统生物学称为 21 世纪的生物学。



## 湿实验

采用高通量实验技术，通过众多组学，在整体和动态研究水平上积累数据并在挖掘数据时发现新规律、新知识，提出新概念。

## 干实验

通过计算生物学建立生物模型，根据被研究的真实系统的模型，利用计算机进行实验研究。



## 湿实验

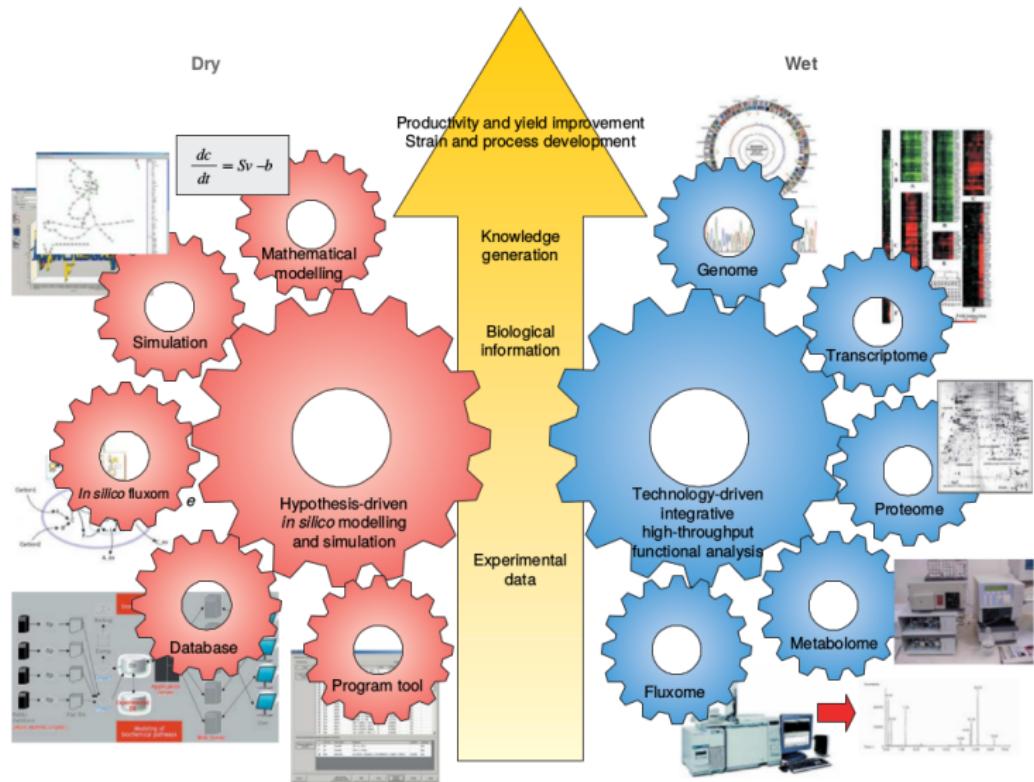
采用高通量实验技术，通过众多组学，在整体和动态研究水平上积累数据并在挖掘数据时发现新规律、新知识，提出新概念。

## 干实验

通过计算生物学建立生物模型，根据被研究的真实系统的模型，利用计算机进行实验研究。



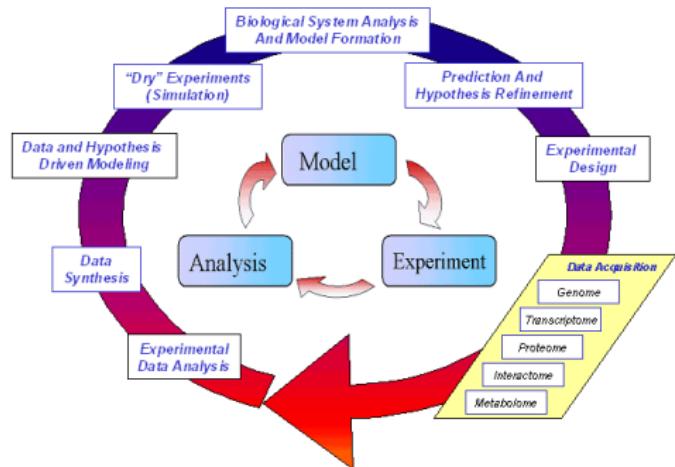
# 系统生物学 | 内容



TRENDS in Biotechnology

## 流程

- ① 研究组分，构建模型
- ② 改变条件，观测变化
- ③ 比较结果，修订模型
- ④ 重新实验，继续修订



## 整合 (incorporation)

把系统内不同性质的构成要素 (DNA、RNA、蛋白质和生物小分子等) 或不同层次的构成要素整合在一起进行研究。

## 干涉 (perturbation)

人为地设定某种或某些条件去作用于被实验的对象，从而研究特定的生命系统在不同时间和空间条件下具有的动力学特征。



## 整合 (incorporation)

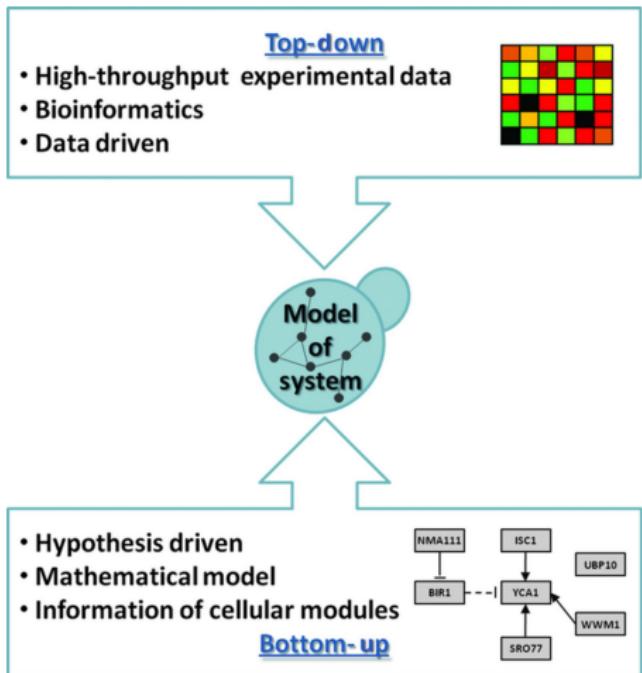
把系统内不同性质的构成要素 (DNA、RNA、蛋白质和生物小分子等) 或不同层次的构成要素整合在一起进行研究。

## 干涉 (perturbation)

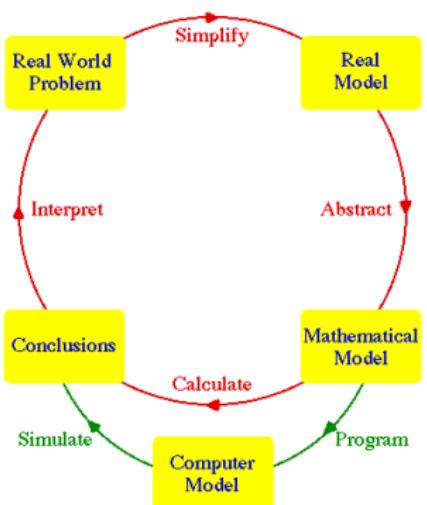
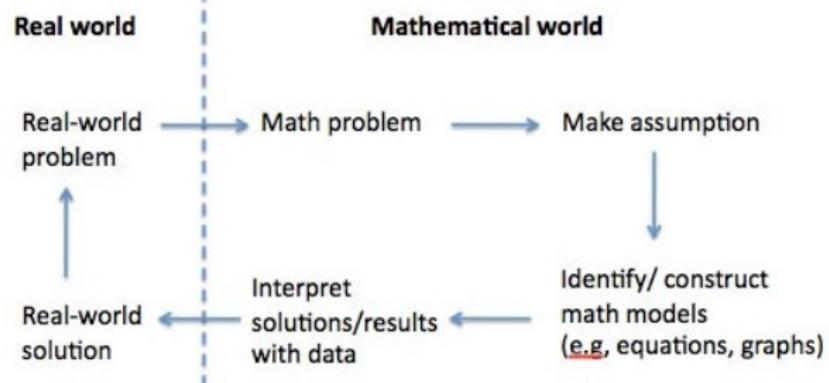
人为地设定某种或某些条件去作用于被实验的对象，从而研究特定的生命系统在不同时间和空间条件下具有的动力学特征。



- 自下而上 (hypothesis based) : 使用独立的实验数据, 适用于大多数基因和它们的调控关系相对比较清楚的情况
- 自上而下 (data driven) : 利用高通量的 DNA 芯片和其他新的测试技术获得数据来研究
- 混合使用 : 自下而上 + 自上而下



# 系统生物学 | 方法 | 建模和模拟



## System Identification

"Divide" into components

Genome-wide Profiling      Novel Circadian Identification of  
Bioinformatics      Clock Circuit      Season-sensor

Nature, 2002; Nature Genetics, 2005;  
Genes Dev., 2007; Nature 2008

## System Analysis

Accurately "measure"

Quantitative Measurement      Heart of Clock = Feedback  
Behavior Prediction      Repression

Nature Genetics, 2006

### Our Main Achievements

## System Control

"Operate" intentionally

Melanopsin (+), CT~17

Time (h)

Perturbation of Clock State by Light      Singularity = De-synchronization

Nature Cell Biology, 2007

## System Design

"Synthesize" from scratch

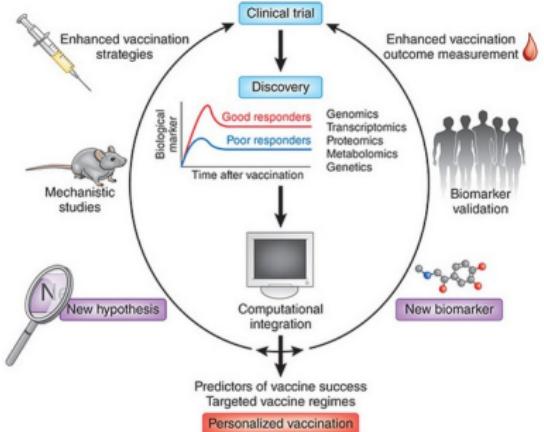
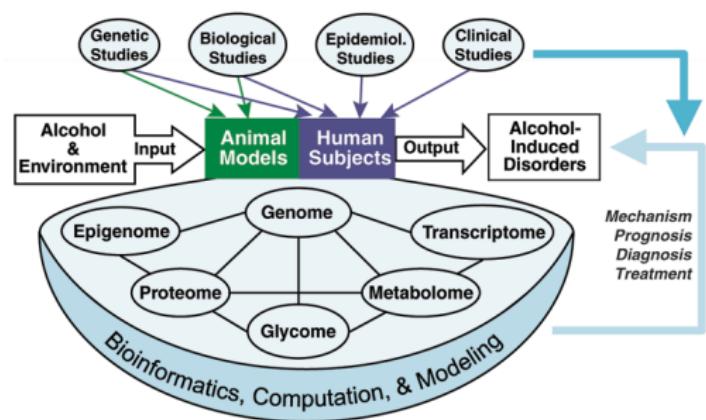
Design Artificial Circuits of Circadian clock in Cell

PNAS, Nature Cell Biology, 2008

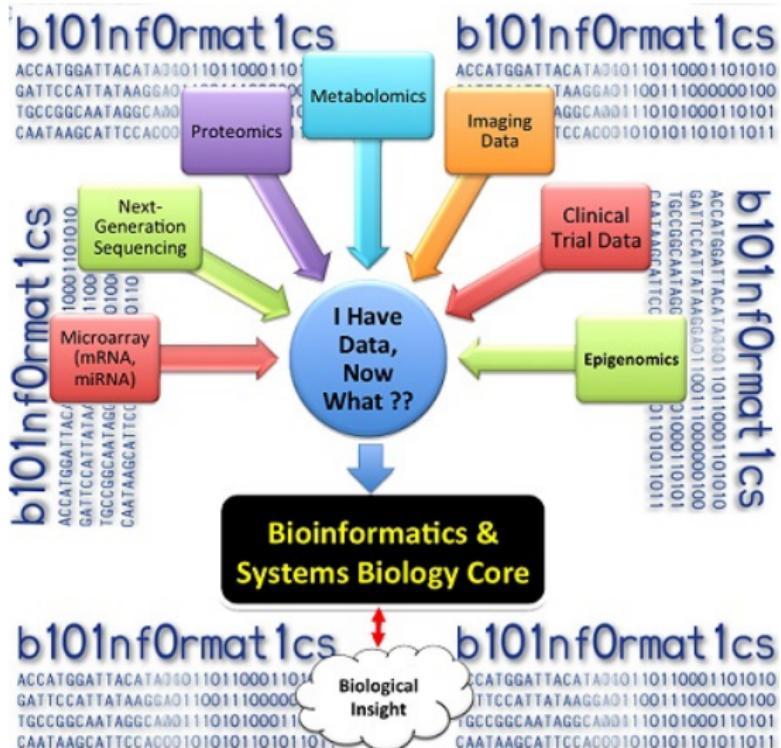
Combinatorial (Re)Generation of Circadian Phase



# 系统生物学 | 应用



# 系统生物学 | 前景



# 教学提纲

## 1 系统生物学

## 2 基因组学

## 3 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

## 4 数据库与数据格式

## 5 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

## 6 外显子组测序

- 简介
- 操作流程
- 应用实例

## 7 转录组学

## 8 RNA-Seq

- 概述
- 数据分析
- 应用实例

## 9 顺反组

- 概述
- ChIP-Seq

## 10 表观遗传学

- 概述
- Methyl-Seq



## 基因

基因 (gene) 是编码某种特定多肽链、tRNA、rRNA 和 ncRNA 的 DNA 区段，是 DNA 上的功能单位。

## 基因组

基因组 (genome) 是一种生物体或个体细胞所具有的一套完整的基因及其调控序列。

## 基因组学

基因组学 (genomics) 是研究基因组的结构组成、时序表达模式和功能，并提供有关生物物种及其细胞功能的进化信息。



## 基因

基因 (gene) 是编码某种特定多肽链、tRNA、rRNA 和 ncRNA 的 DNA 区段，是 DNA 上的功能单位。

## 基因组

基因组 (genome) 是一种生物体或个体细胞所具有的一套完整的基因及其调控序列。

## 基因组学

基因组学 (genomics) 是研究基因组的结构组成、时序表达模式和功能，并提供有关生物物种及其细胞功能的进化信息。



## 基因

基因 (gene) 是编码某种特定多肽链、tRNA、rRNA 和 ncRNA 的 DNA 区段，是 DNA 上的功能单位。

## 基因组

基因组 (genome) 是一种生物体或个体细胞所具有的一套完整的基因及其调控序列。

## 基因组学

基因组学 (genomics) 是研究基因组的结构组成、时序表达模式和功能，并提供有关生物物种及其细胞功能的进化信息。



- 1976 年, 瓦尔特·菲尔斯 (比利时根特大学), RNA 病毒噬菌体 MS2 【基因组】
- 1977 年, 弗雷德里克·桑格,  $\Phi$ -X174 噬菌体 【DNA 基因组】
- 1995 年, The Institute for Genomic Research 团队, 流感嗜血杆菌 (*Haemophilus influenzae*) 【细菌基因组】
- 1996 年, 酿酒酵母 (*Saccharomyces cerevisiae*) 【真核生物基因组】
- 1996 年, The Institute for Genomic Research 团队, 詹氏甲烷球菌 (*Methanococcus jannaschii*) 【古菌基因组】
- 1998 年, 秀丽隐杆线虫 (*Caenorhabditis elegans*) 【多细胞生物基因组】
- 1990 年, 人类基因组计划启动
- 2007 年, 完成了詹姆斯·杜威·沃森个人基因组的测序
- 2008-2012 年, 千人基因组计划



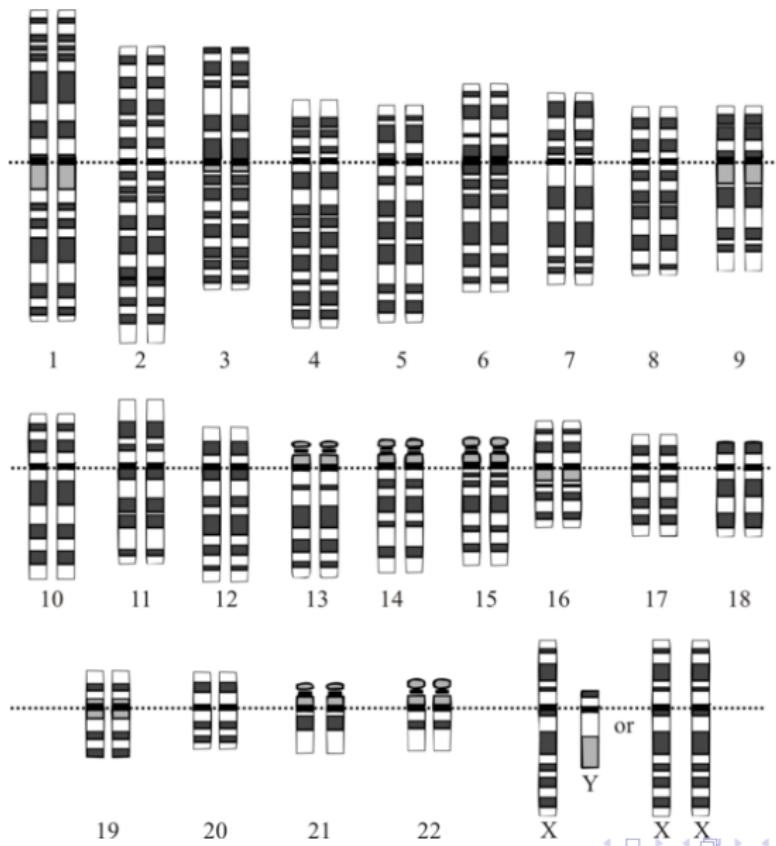
## 人类基因组

人类基因组，又称人类基因体，是智人 (*Homo sapiens*) 的基因组，由 23 对染色体组成，其中包括 22 对体染色体、1 条 X 染色体和 1 条 Y 染色体。

人类基因组含有约 30 亿个 DNA 碱基对，碱基对是以氢键相结合的两个含氮碱基，以胸腺嘧啶 (T)、腺嘌呤 (A)、胞嘧啶 (C) 和鸟嘌呤 (G) 四种碱基排列成碱基序列，其中 A 与 T 之间由两个氢键连接，G 与 C 之间由三个氢键连接，碱基对的排列在 DNA 中也只能是 A 对 T，G 对 C。其中一部分的碱基对组成了大约 20000 到 25000 个基因。



基因组学 | 人类基因组

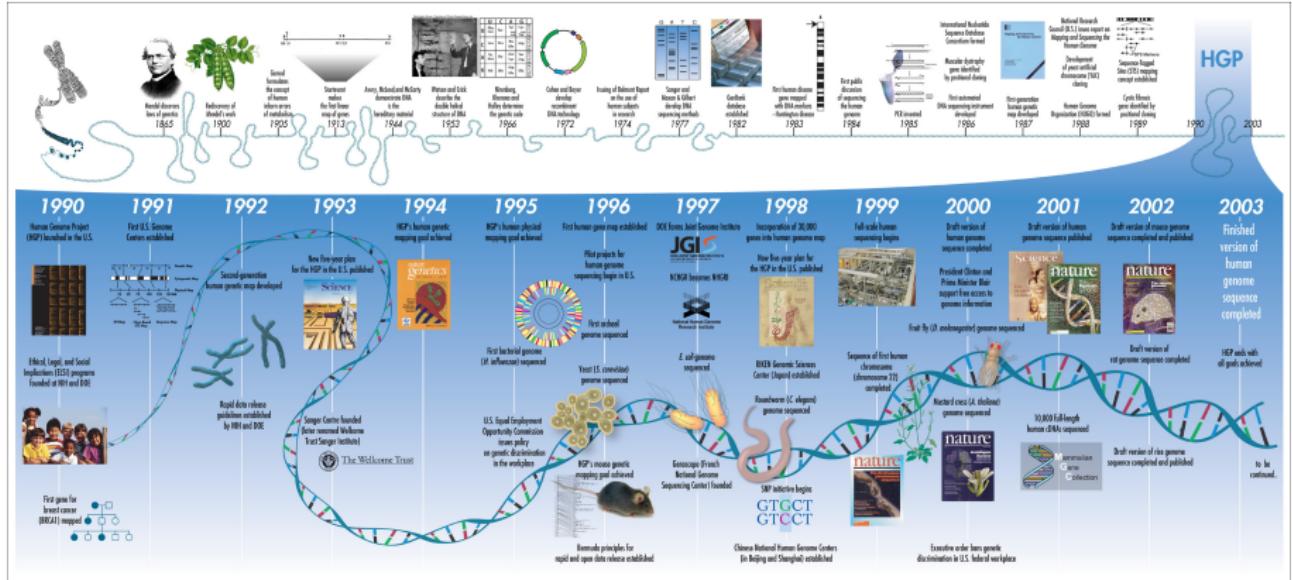


## Greatest Breakthroughs in Human History

- Manhattan Project ~1940/1944  
(Nuclear Energy)
- Apollo Project ~1960/1972  
(Moon Landing)
- Human Genome Project ~1990/2003  
(Decoding the Book of Life)



基因组学 | 人类基因组计划 | 事件



- 1984 年，第一次讨论人类基因组测序的价值
- 1985 年，首次对于人类基因组测序的可行性进行认真的探讨
- 1986 年，罗纳德·杜尔贝科 (Renato Dulbecco)，建议开展人类基因组研究计划
- 1986 年，美国能源部 (DOE) 加入人类基因组计划
- 1987 年，美国国家卫生研究院 (NIH) 加入人类基因组计划
- 1998 年，詹姆斯·华生 (沃森)，NIH 的基因组部门主管 (1988-1992)
- 1988 年，国际人类基因组组织 (HUGO) 成立



- 1990 年，投资 30 亿美元的人类基因组计划由美国能源部和国家卫生研究院正式启动，预期在 15 年内完成，随后扩展为国际合作的人类基因组计划
- 1996 年，百慕大会议，以 2005 年完成测序为目标，分配了各国负责的工作，并且宣布研究结果将会即时公布，且完全免费
- 1998 年，克莱格·凡特的塞雷拉基因组公司成立，希望能以更快的速度和更少的投资（3 亿美元）来完成此项工程；开发出全世界第一台全自动测序仪，宣布将在 2001 年完成测序工作
- 2000 年 6 月 26 日，塞雷拉公司的代表凡特，以及国际合作团队的代表弗朗西斯·柯林斯（Francis Collins），在美国总统克林顿的陪同下发表演说，宣布人类基因组的概要已经完成；所有人类基因组数据为人类共同财产，不允许专利保护，且必须对所有研究者公开
- 2001 年 2 月，国际人类基因组测序联盟与塞雷拉公司，分别将研究成果发表于《自然》与《科学》；覆盖基因组序列的 83%，包括常染色质区域的 90%（带有 150,000 个空缺，且许多片断的顺序和方位并没有得到确定）

- 1994 年，中国的人类基因组计划启动
- 1998 年，中国南方基因组中心成立，中国科学院遗传研究所人类基因组中心成立
- 1999 年，北京华大基因研究中心（华大基因）成立，北方基因组中心成立
- 1998 年 3 月，中美港科学家合作，成功地将与华人和鼻咽癌有关的肿瘤抑制基因定位于人类第 3 号染色体的短臂 3p21.3 位点
- 1999 年 6 月 26 日，中国科学院遗传研究所人类基因组中心向美国国立卫生研究院（NIH）的国际人类基因组计划（HGP）递交加入申请。HGP 在网上公布中国注册加入国际测序组织，中国成为继美、英、日、德、法后第六个加入该组织的国家
- 1999 年 11 月 10 日，1% 计划被列入中国国家项目，并确定由北京华大基因研究中心（华大基因）牵头，国家基因组南方中心、北方中心共同参与，承担全部工程 1% 的测序工作
- 2000 年 4 月，中国完成了人第 3 号染色体上 3000 万个碱基对的工作草图



## 延伸计划

**模式生物的基因组计划** 小鼠、果蝇、线虫、斑马鱼、酵母等。

**人类元基因组计划** 对人体内所有共生菌群的基因组进行序列测定，并研究与人体发育和健康相关基因的功能。

**国际人类基因组单体型图计划（HapMap 计划）** 目标是构建人类 DNA 序列中多态位点的常见模式，为研究人员确定对健康和疾病以及对药物和环境反应有影响的相关基因提供关键信息。

**人类基因组多样性研究计划** 对不同人种、民族、人群的基因组进行研究和比较，这一计划将为疾病监测、人类的进化研究和人类学研究提供重要信息。

**千人基因组计划（1000 Genomes Project）** 目标是建立最详尽的人类遗传变异目录。启动于 2008 年 1 月，计划在随后三年内，测定来自不同族群的至少一千名的匿名参与者的基因组序列。2010 年完成试点阶段，2012 年 10 月公布 1092 个基因组的测序。

# 教学提纲

1

系统生物学

2

基因组学

3

测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

数据库与数据格式

5

二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

外显子组测序

- 简介
- 操作流程
- 应用实例

7

转录组学

8

RNA-Seq

- 概述
- 数据分析
- 应用实例

9

顺反组

- 概述
- ChIP-Seq

10

表观遗传学

- 概述
- Methyl-Seq



# 教学提纲

1

系统生物学

2

基因组学

3

测序技术

- 测序技术的发展

- 第一代测序技术

- 第二代测序技术

- 第三代测序技术

- 测序技术的比较

4

数据库与数据格式

5

二代测序数据分析

- 常见术语

- 分析流程

- 补遗

6

外显子组测序

● 简介

● 操作流程

● 应用实例

7 转录组学

8 RNA-Seq

- 概述

- 数据分析

- 应用实例

9 顺反组

- 概述

- ChIP-Seq

10 表观遗传学

- 概述

- Methyl-Seq



## DNA 测序

DNA 测序 (DNA sequencing) 是指分析特定 DNA 片段的碱基序列，也就是腺嘌呤 (A)、胸腺嘧啶 (T)、胞嘧啶 (C) 与鸟嘌呤 (G) 的排列方式。

## RNA 测序

RNA 测序则通常将 RNA 提取后，反转录为 DNA 后使用 DNA 测序的方法进行测序。



## DNA 测序

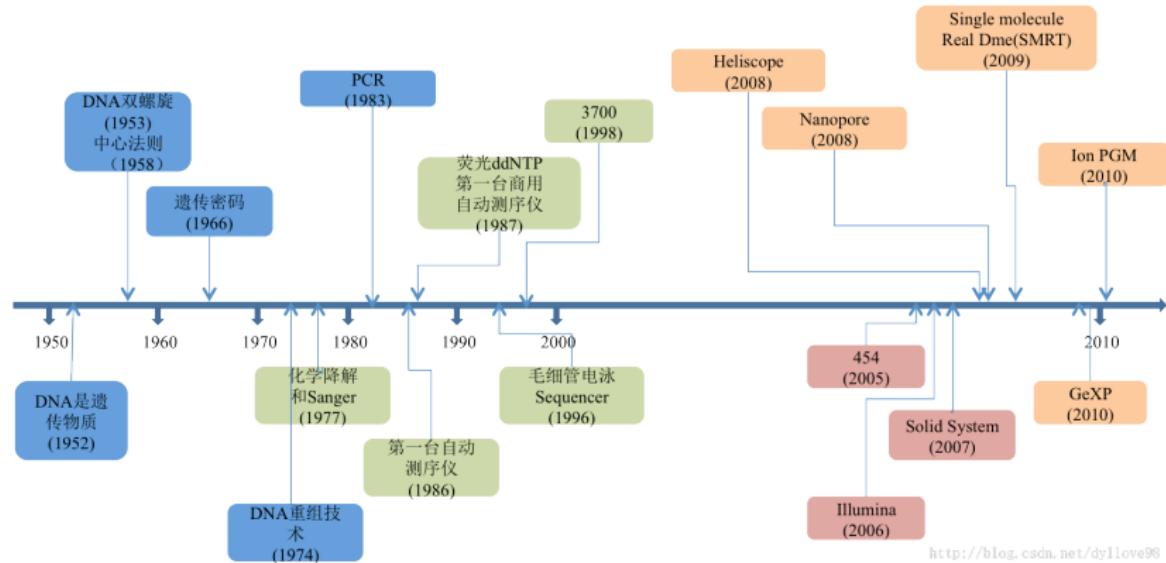
DNA 测序 (DNA sequencing) 是指分析特定 DNA 片段的碱基序列，也就是腺嘌呤 (A)、胸腺嘧啶 (T)、胞嘧啶 (C) 与鸟嘌呤 (G) 的排列方式。

## RNA 测序

RNA 测序则通常将 RNA 提取后，反转录为 DNA 后使用 DNA 测序的方法进行测序。



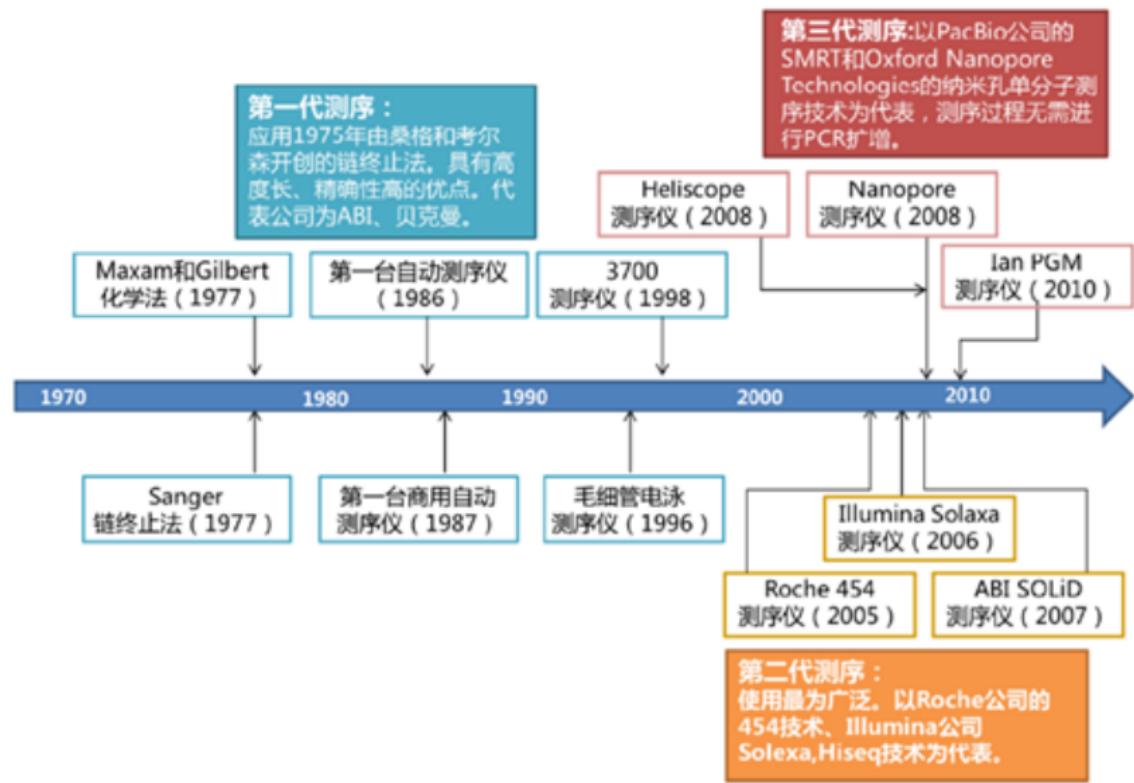
# 测序 | 历史



<http://blog.csdn.net/dyllove98>



# 测序 | 历史



# 教学提纲

1

系统生物学

2

基因组学

3

测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

数据库与数据格式

5

二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

外显子组测序

- 简介
- 操作流程
- 应用实例

7

转录组学

8

RNA-Seq

- 概述
- 数据分析
- 应用实例

9

顺反组

- 概述
- ChIP-Seq

10

表观遗传学

- 概述
- Methyl-Seq



- 1975 年，弗雷德里克·桑格（Frederick Sanger）和艾伦·库尔森（Alan Coulson），“加减测序法技术”；改进后为链终止法（chain termination method），即桑格测序法
- 1977 年，哈佛大学的沃尔特·吉尔伯特（Walter Gilbert）和艾伦·马克萨姆（Allan Maxam），链降解，马克萨姆-吉尔伯特测序（Maxam-Gilbert 法，又称化学测序法）



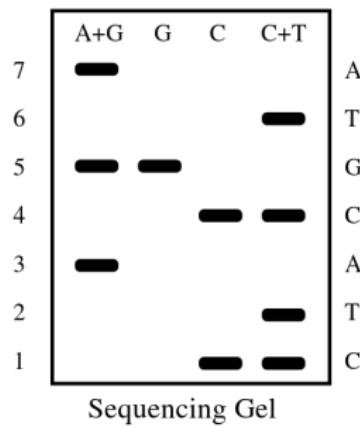
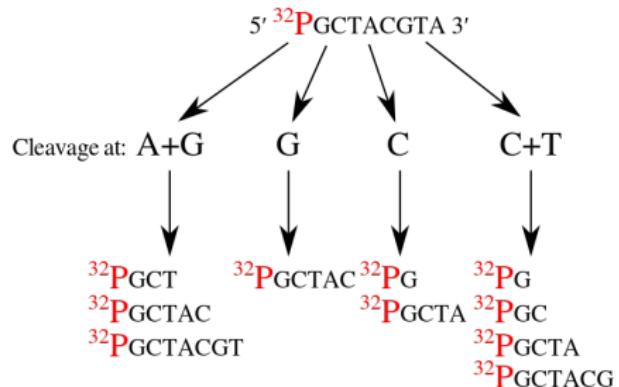
## 化学测序法

马克萨姆-吉尔伯特测序（Maxam-Gilbert sequencing）是一项由阿伦·马克萨姆与沃尔特·吉尔伯特于 1976~1977 年间开发的 DNA 测序方法。此项方法基于：对核碱基特异性地进行局部化学变性，接下来在变性核苷酸毗邻的位点处 DNA 骨架发生断裂。

最初的桑格法须要在每次测序之前克隆得到单链 DNA 产物，因此马克萨姆-吉尔伯特测序法发表后迅速得到了推广，因为被纯化的 DNA 可被直接使用。然而，随着链终止法的改良，马克萨姆-吉尔伯特测序逐渐失宠，这是由于：技术复杂性阻碍其成为标准分子生物学套装使用、大量使用危险药品以及难于扩大规模。



## 测序 | 第一代 | 化学测序法



## 桑格测序法

Sanger (桑格) 双脱氧链终止法是弗雷德里克·桑格 (Frederick Sanger) 于 1975 年发明的。测序过程需要先做一个聚合酶连锁反应 (PCR)。PCR 过程中，双脱氧核糖核苷酸可能随机的被加入到正在合成中的 DNA 片段里。由于双脱氧核糖核苷酸少了一个氧原子，一旦它被加入到 DNA 链上，这个 DNA 链就不能继续增加长度。最终的结果是获得所有可能获得的、不同长度的 DNA 片段。

目前最普遍最先进的方法，是将双脱氧核糖核苷酸进行不同荧光标记。将 PCR 反应获得的总 DNA 通过毛细管电泳分离，跑到最末端的 DNA 就可以在激光的作用下发出荧光。由于 ddATP, ddGTP, ddCTP, ddTTP (4 种双脱氧核糖核苷酸) 荧光标记不同，计算机可以自动根据颜色判断该位置上碱基究竟是 A, T, G, C 中的哪一个。



## 原理

双脱氧链终止法采用 DNA 复制原理。Sanger 测序反应体系中包括目标 DNA 片断、脱氧三磷酸核苷酸 (dNTP)、双脱氧三磷酸核苷酸 (ddNTP)、测序引物及 DNA 聚合酶等。

测序反应的核心就是其使用的 ddNTP：由于缺少 3'-OH 基团，不具有与另一个 dNTP 连接形成磷酸二酯键的能力，这些 ddNTP 可用来中止 DNA 链的延伸。此外，这些 ddNTP 上连接有放射性同位素或荧光标记基团，因此可以被自动化的仪器或凝胶成像系统所检测到。



## 概述

每个反应含有所有四种脱氧三磷酸核苷酸（dNTP）使之扩增，并混入限量的一种不同的双脱氧三磷酸核苷酸（ddNTP）使之终止。由于 ddNTP 缺乏延伸所需要的 3'-OH 基团，使延长的寡聚核苷酸选择性地在 G、A、T 或 C 处终止，终止点由反应中相应的 ddNTP 而定。

每一种 dNTPs 和 ddNTPs 的相对浓度可以调整，使反应得到一组长几个至千以上个、相差一个碱基的一系列片断。它们具有共同的起始点，但终止在不同的核苷酸上，可通过高分辨率变性凝胶电泳分离大小不同的片段，凝胶处理后可用 X-光胶片放射自显影或非同位素标记进行检测。



## 优势

- 最长可测定 600-1000bp 的 DNA 片断
- 对重复序列和多聚序列的处理较好
- 序列准确性高，高达 99.999%
- 测序的“黄金标准”

## 缺点

- 通量较低（在 24h 内可测定的 DNA 分子数一般不超过 10,000 个）
- 每碱基测序成本较高
- 不适合大规模平行测序



## 优势

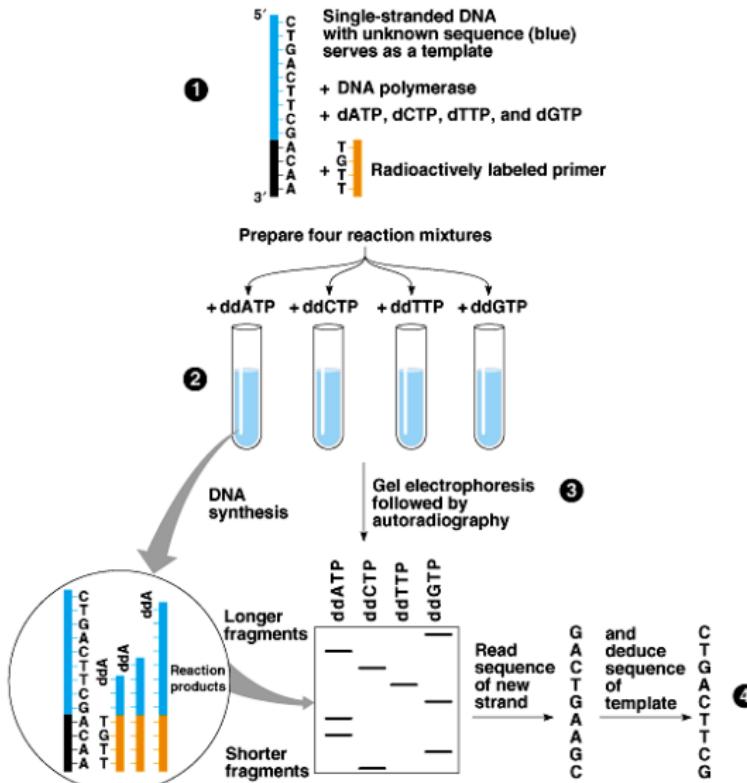
- 最长可测定 600-1000bp 的 DNA 片断
- 对重复序列和多聚序列的处理较好
- 序列准确性高，高达 99.999%
- 测序的“黄金标准”

## 缺点

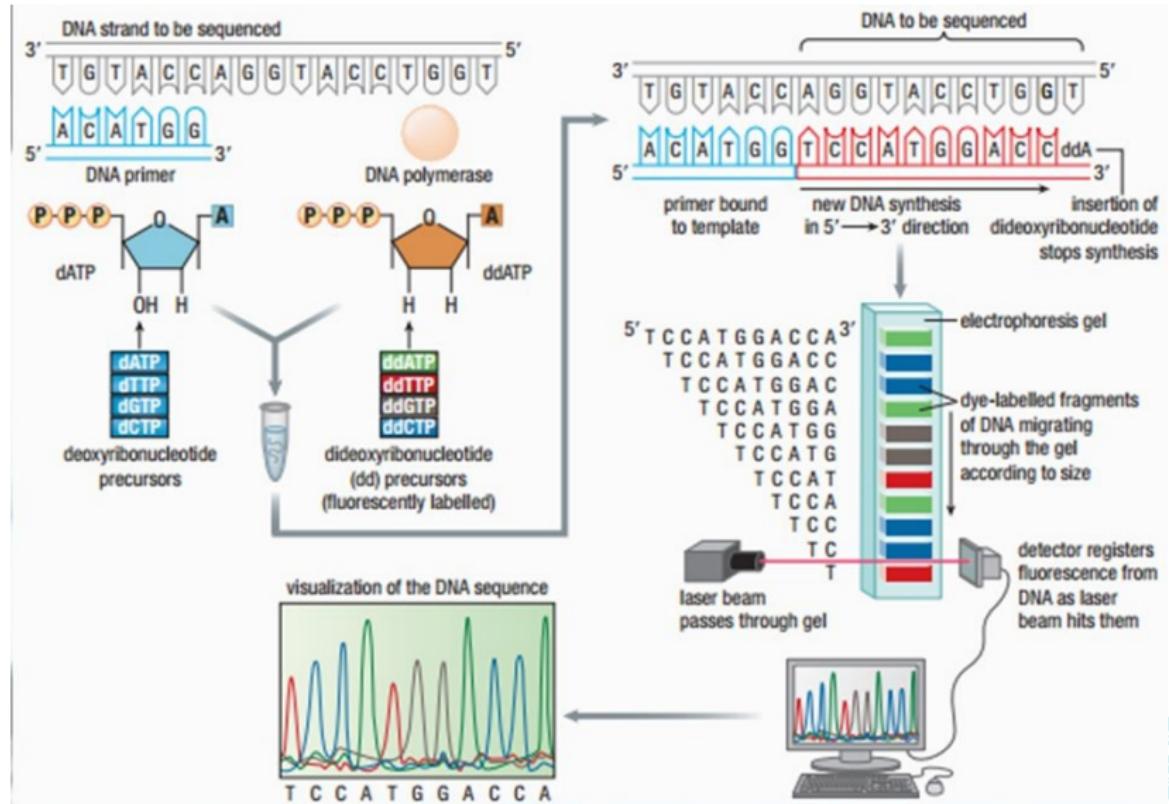
- 通量较低（在 24h 内可测定的 DNA 分子数一般不超过 10,000 个）
- 每碱基测序成本较高
- 不适合大规模平行测序



# 测序 | 第一代 | 桑格测序法



# 测序 | 第一代 | 桑格测序法



# 教学提纲

1

系统生物学

2

基因组学

3

测序技术

- 测序技术的发展
- 第一代测序技术
- **第二代测序技术**
- 第三代测序技术
- 测序技术的比较

4

数据库与数据格式

5

二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

外显子组测序

- 简介
- 操作流程
- 应用实例

7

转录组学

8

RNA-Seq

- 概述
- 数据分析
- 应用实例

9

顺反组

- 概述
- ChIP-Seq

10

表观遗传学

- 概述
- Methyl-Seq



## Reads

一个完美的人，不是寻找一眼光，欣赏那爱，不是寻找是学会用完美不完美的人。

是寻找一个完的人，而是学完美的眼光，，欣赏那个并完美的人，而那个并不完美

## 三思而行

爱，不是寻找一个完美的人，而是学会用完美的眼光，欣赏那个并不完美的人。——《哈尔的移动城堡》



## Reads

一个完美的人，不是寻找一眼光，欣赏那爱，不是寻找是学会用完美不完美的人。

是寻找一个完的人，而是学完美的眼光，，欣赏那个并完美的人，而那个并不完美

## 基因组

爱，不是寻找一个完美的人，而是学会用完美的眼光，欣赏那个并不完美的人。——《哈尔的移动城堡》



## Reads

一个完美的人，不是寻找一眼光，欣赏那爱，不是寻找是学会用完美不完美的人。

是寻找一个完的人，而是学完美的眼光，，欣赏那个并完美的人，而那个并不完美

## 基因组

爱，不是寻找一个完美的人，而是学会用完美的眼光，欣赏那个并不完美的人。——《哈尔的移动城堡》



## Reads

一个完美的人，不是寻找一眼光，欣赏那爱，不是寻找是学会用完美不完美的人。

是寻找一个完的人，而是学完美的眼光，，欣赏那个并完美的人，而那个并不完美

## 基因组

爱，不是寻找一个完美的人，而是学会用完美的眼光，欣赏那个并不完美的人。——《哈尔的移动城堡》



## Reads

一个完美的人  
，不是寻找一  
眼光，欣赏那  
爱，不是寻找  
是学会用完美  
不完美的人。

是寻找一个完  
的人，而是学  
完美的眼光，  
，欣赏那个并  
完美的人，而  
那个并不完美

## 基因组

爱，不是寻找一个完美的人，而是学会用完美的眼光，欣赏那个并不完美的人。——《哈尔的移动城堡》



## Reads

who you are are, but	love you not but for who
I love you not for who	you not for for who I
am with you	with you.
who I am	you are,

## 三思录

I love you not for who you are, but for who I am with you. ——《剪刀手爱德华》



## Reads

who you are are, but	love you not but for who
I love you not for who	you not for for who I
am with you	with you.
who I am	you are,

## 基因组

I love you not for who you are, but for who I am with you. —— 《剪刀手爱德华》



## Reads

who you are are, but	love you not but for who
I love you not for who	you not for for who I
am with you	with you.
who I am	you are,

## 基因组

I love you not for who you are, but for who I am with you. —— 《剪刀手爱德华》



## Reads

who you are are, but	love you not but for who
I love you not for who	you not for for who I
am with you	with you.
who I am	you are,

## 基因组

I love you not for who you are, but for who I am with you. —— 《剪刀手爱德华》



## Reads

who you are are, but	love you not but for who
I love you not for who	you not for for who I
am with you	with you.
who I am	you are,

## 基因组

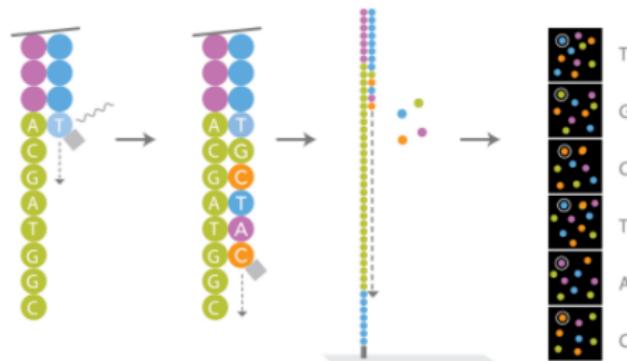
I love you not for who you are, but for who I am with you. —— 《剪刀手爱德华》



## What are reads?

The sequenced part of one DNA fragment.

These DNA sequences are given by the sequencing machine with a Phred quality score in so called **FASTQ** format



```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36  
TGCTTACGGCCGCTGCCGATGGCGTCAAATCCCACC  
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36  
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

<b>GCTGATGTGCCGCCTCACTTCGGTGGTGAGGTG</b>	Reference sequence
CTGATGTGCCGCCTCACTTCGGTGGT	Short read 1
TGATGTGCCGCCTCACT <b>ACGGTGGTG</b>	Short read 2
GATGTGCCGCCTCACTTCGGTGGTGA	Short read 3
GCTGATGTGCCGCCTCACT <b>ACGGTG</b>	Short read 4
GCTGATGTGCCGCCTCACT <b>ACGGTG</b>	Short read 5



## Roche 公司的 454 技术

- 1996 年，波尔·尼伦和穆斯塔法·罗纳吉，焦磷酸测序 (pyrosequencing)
- 2004/2005 年，商业化测序仪
- 2009 年，百万条、200-400bp

## Illumina 公司的 Solexa 技术

- 2006 年，商业化测序仪
- 2009 年，上亿条、50-100bp

## ABI 公司的 SOLiD 技术

- 2006/2007 年，商业化测序仪

## Roche 公司的 454 技术

- 1996 年，波尔·尼伦和穆斯塔法·罗纳吉，焦磷酸测序 (pyrosequencing)
- 2004/2005 年，商业化测序仪
- 2009 年，百万条、200-400bp

## Illumina 公司的 Solexa 技术

- 2006 年，商业化测序仪
- 2009 年，上亿条、50-100bp

## ABI 公司的 SOLiD 技术

- 2006/2007 年，商业化测序仪

## Roche 公司的 454 技术

- 1996 年，波尔·尼伦和穆斯塔法·罗纳吉，焦磷酸测序 (pyrosequencing)
- 2004/2005 年，商业化测序仪
- 2009 年，百万条、200-400bp

## Illumina 公司的 Solexa 技术

- 2006 年，商业化测序仪
- 2009 年，上亿条、50-100bp

## ABI 公司的 SOLiD 技术

- 2006/2007 年，商业化测序仪

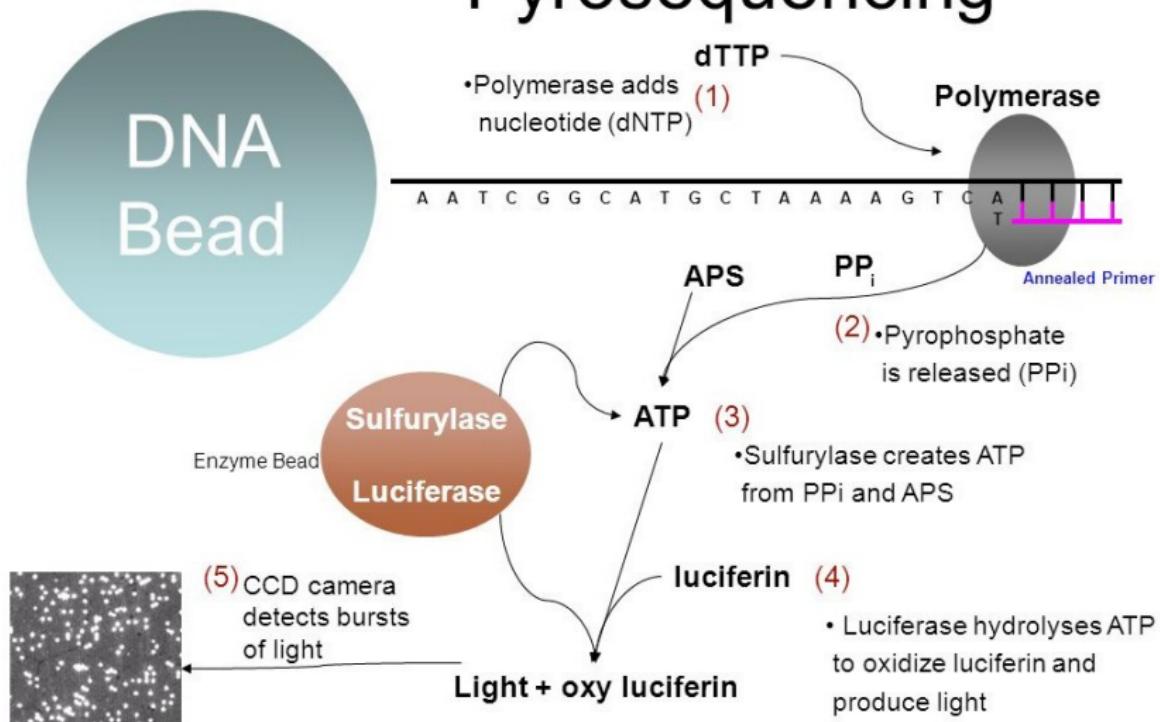
## 焦磷酸测序

焦磷酸测序 (pyrosequencing) 是一种基于聚合原理的 DNA 测序方法，它依赖于核苷酸掺入中焦磷酸盐的释放，而非双脱氧三磷酸核苷酸参与的链终止反应。

Pyrosequencing 技术是由 4 种酶催化的同一反应体系中的酶级联化学发光反应。在每一轮测序中，只加入一种 dNTP，若该 dNTP 与模板配对，聚合酶就可以将其掺入到引物链中并释放出等摩尔数的焦磷酸基团 (PPi)。PPi 可最终转化为可见光信号，并由 Pyrogram™ 转化为一个峰值。每个峰值的高度与反应中掺入的核苷酸数目成正比。然后加入下一种 dNTP，继续 DNA 链的合成。



# Pyrosequencing

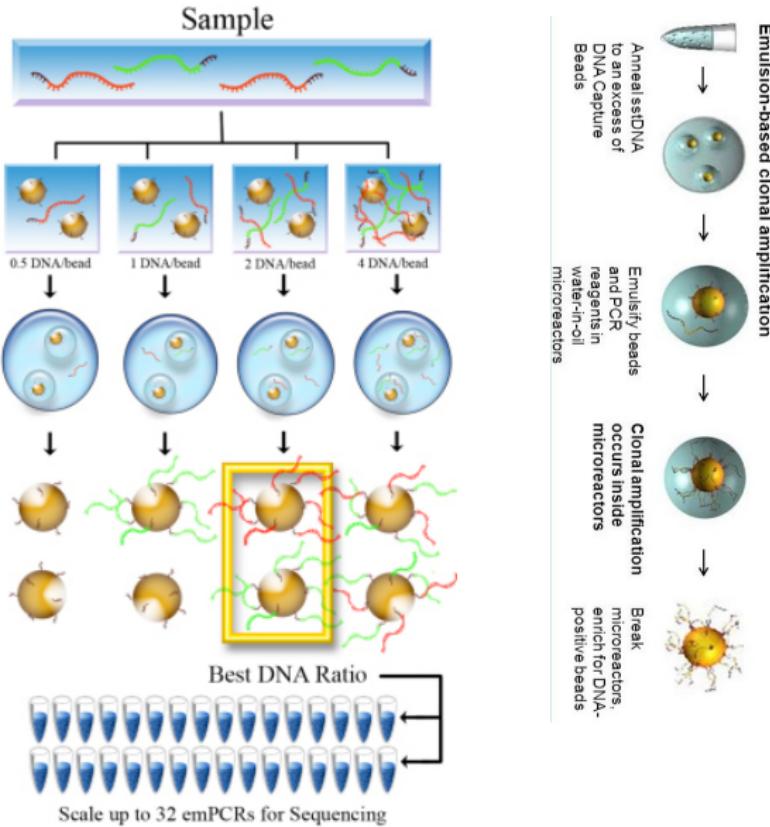


## emPCR

emPCR（乳液 PCR）主要通过将水相 PCR 溶液（包含引物、聚合酶、核苷酸和待扩增 DNA）与油混合，创建一种微小的悬浮水滴乳液。每个液滴都作为其自身 PCR 的“反应器”，从而创造了平行反应中的多个独立反应。

emPCR (emulsion PCR) 技术利用油包水 (water-in-oil) 结构作为 PCR 反应的微反应器，进行 PCR 扩增。emPCR 最大的特点是可以形成数目庞大的独立反应空间以进行 PCR 扩增。其关键技术是“注水到油”，基本过程是在 PCR 反应前，将包含 PCR 所需反应成分的水溶液注入到高速旋转的油相表面，水溶液瞬间形成数以万计个被油相包裹的小液滴。这些小液滴就形成了 PCR 反应空间。



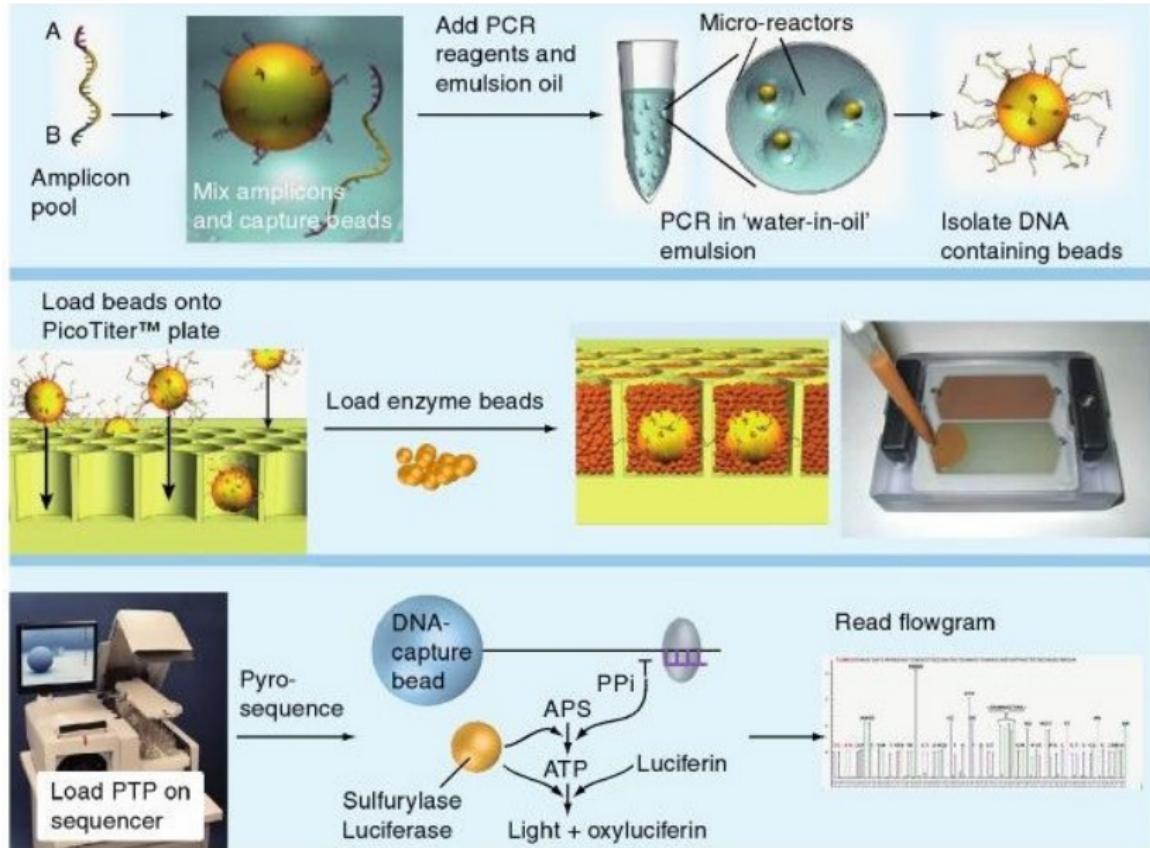


## 概述

在测序时，使用了一种叫做“Pico TiterPlate”（PTP）的平板，它含有160多万个由光纤组成的孔，孔中载有化学发光反应所需的各种酶和底物。测序开始时，放置在四个单独的试剂瓶里的四种碱基，依照T、A、C、G的顺序依次循环进入PTP板，每次只进入一个碱基。如果发生碱基配对，就会释放一个焦磷酸。这个焦磷酸在各种酶的作用下，经过一个合成反应和一个化学发光反应，最终将荧光素氧化成氧化荧光素，同时释放出光信号。此反应释放出的光信号实时被仪器配置的高灵敏度CCD捕获到。有一个碱基和测序模板进行配对，就会捕获到一分子的光信号；由此一一对应，就可以准确、快速地确定待测模板的碱基序列。



# 测序 | 第二代 | Roche/454



## 优点

- 读长长，使得后继的序列拼接工作更加高效、准确
- 速度快，一个测序反应耗时 10 个小时，获得 4-6 亿个碱基对
- 特别适合从头拼接和宏基因组学应用，多用于新的细菌基因组

## 缺点

- 无法准确测量同聚物的长度，所以检测插入缺失突变的误差率高
- 通量小且费用高
- 对重测序来说太贵，不适合



## 优点

- 读长长，使得后继的序列拼接工作更加高效、准确
- 速度快，一个测序反应耗时 10 个小时，获得 4-6 亿个碱基对
- 特别适合从头拼接和宏基因组学应用，多用于新的细菌基因组

## 缺点

- 无法准确测量同聚物的长度，所以检测插入缺失突变的误差率高
- 通量小且费用高
- 对重测序来说太贵，不适合

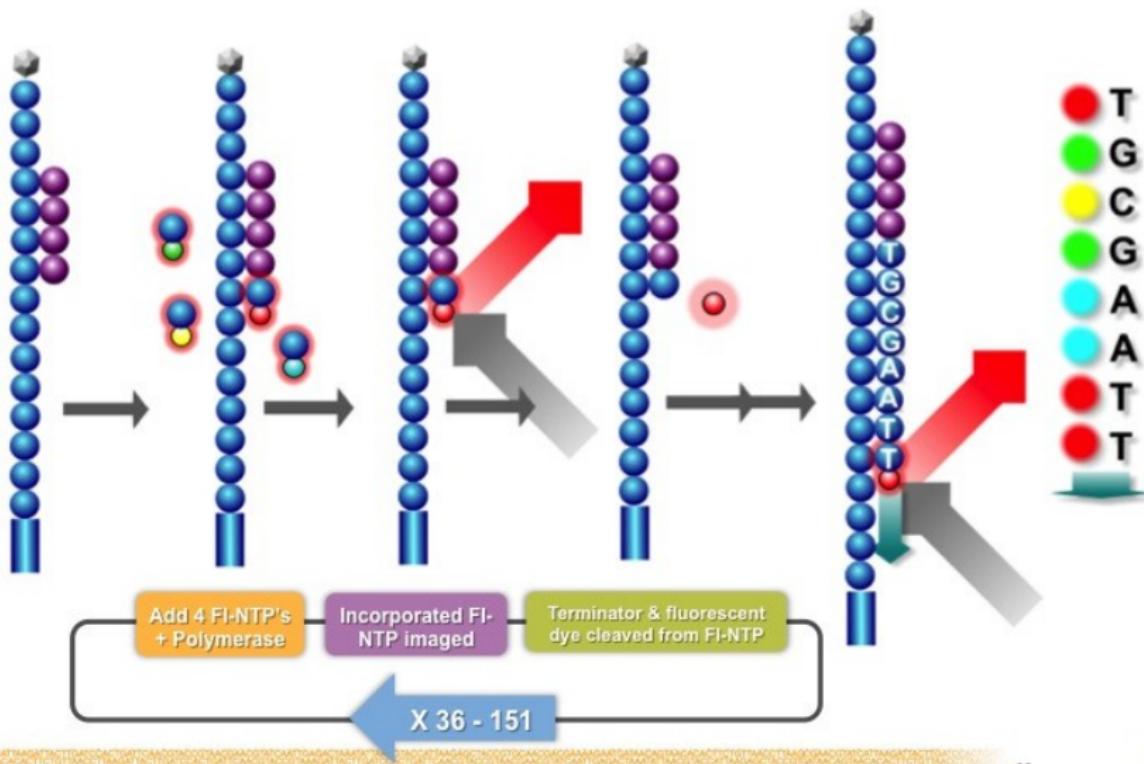


## 边合成边测序

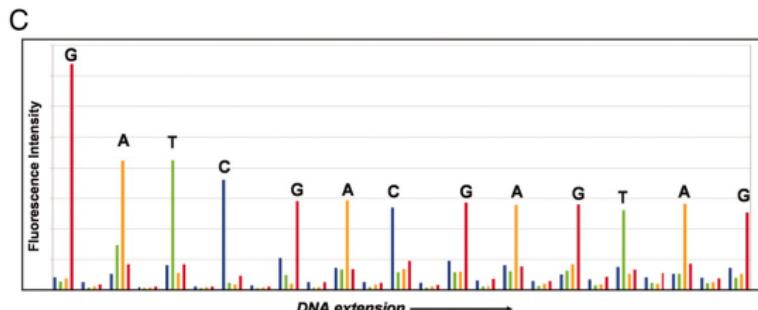
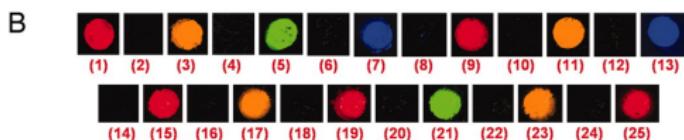
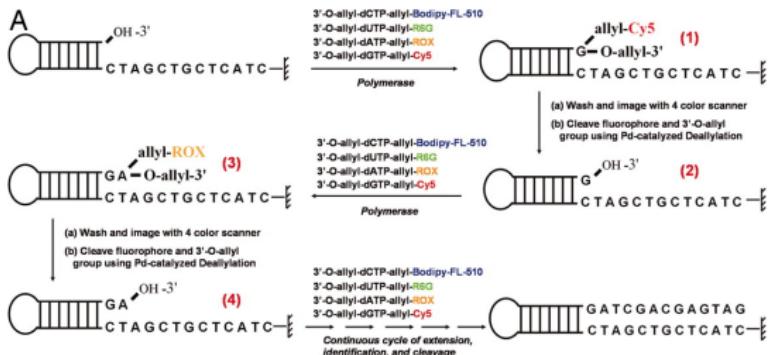
边合成边测序 (sequencing by synthesis, SBS) : 以 DNA 单链为模板, 在合成互补链的时候, 利用带荧光标记的 dNTP 发出不同的荧光来确定碱基类型。



## Sequencing By Synthesis



# 测序 | 第二代 | Illumina/Solexa

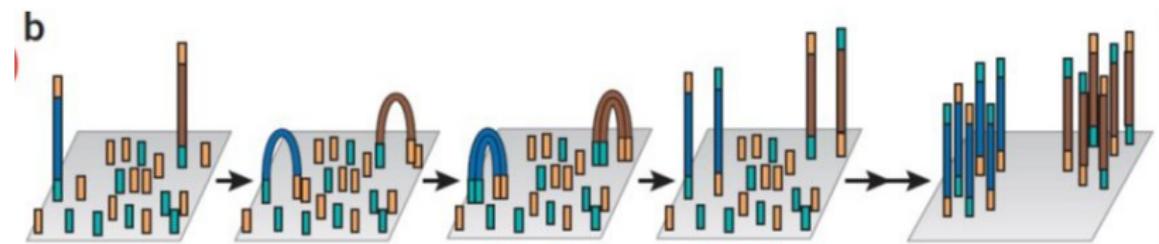


## 桥式扩增

桥式扩增 (bridge amplification)：随机打断的单链 DNA 片段通过两端接头与寡核苷酸的互补固定在芯片表面，形成桥形结构，之后以寡核苷酸为引物进行 PCR 扩增，得到单克隆的 DNA 簇群。



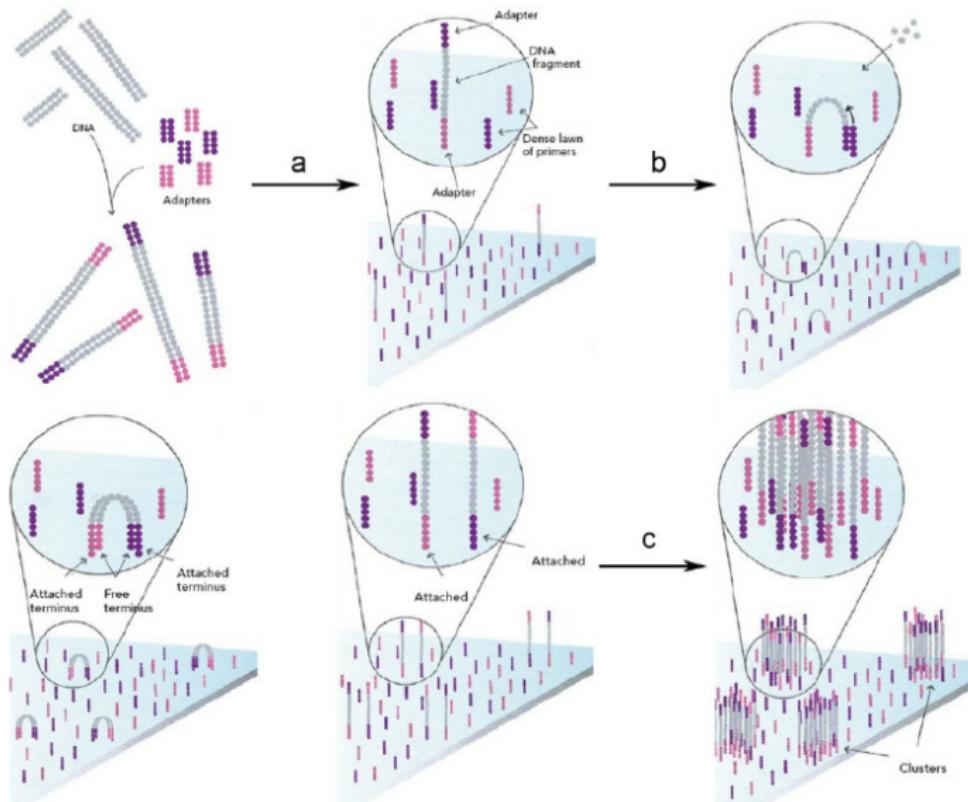
## Bridge PCR



- DNA fragments are flanked with adaptors.
- A flat surface coated with two types of primers, corresponding to the adaptors.
- Amplification proceeds in cycles, with one end of each bridge tethered to the surface.
- Used by Solexa.



# 测序 | 第二代 | Illumina/Solexa



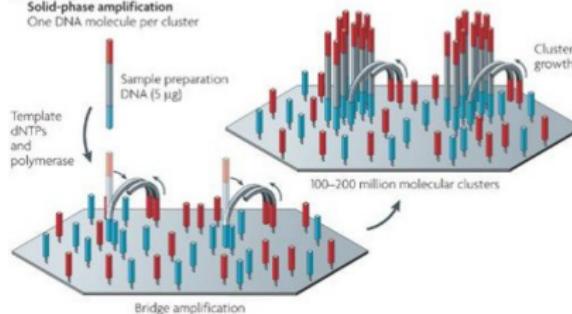
## 概述

这种测序技术通过将基因组 DNA 的随机片断附着到光学透明的表面，这些 DNA 片断通过延长和桥梁扩增，形成了具有数以亿计 cluster 的 Flowcell，每个 cluster 具有约 1000 拷贝的相同 DNA 模板，然后用 4 种末端被封闭的不同荧光标记的碱基进行边合成边测序。这种新方法确保了高精确度和真实的一个碱基接一个碱基的测序，排除了序列方面的特殊错误，能够测序同聚物和重复序列。

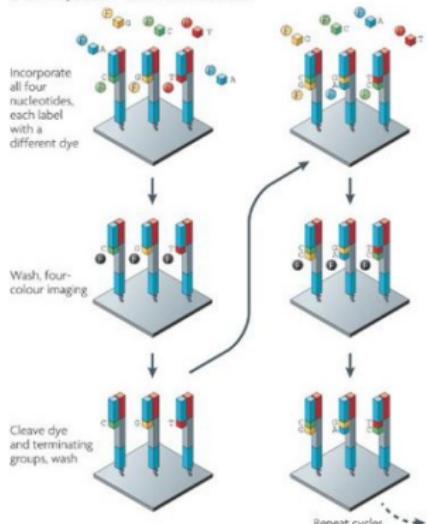


## Illumina Colonies (called “polonies”)

b Illumina/Solexa  
Solid-phase amplification  
One DNA molecule per cluster



a Illumina/Solexa — Reversible terminators



Each nucleotide has a dye with a different color

4-color fluorescent image of chip gathered after each chemical flows through

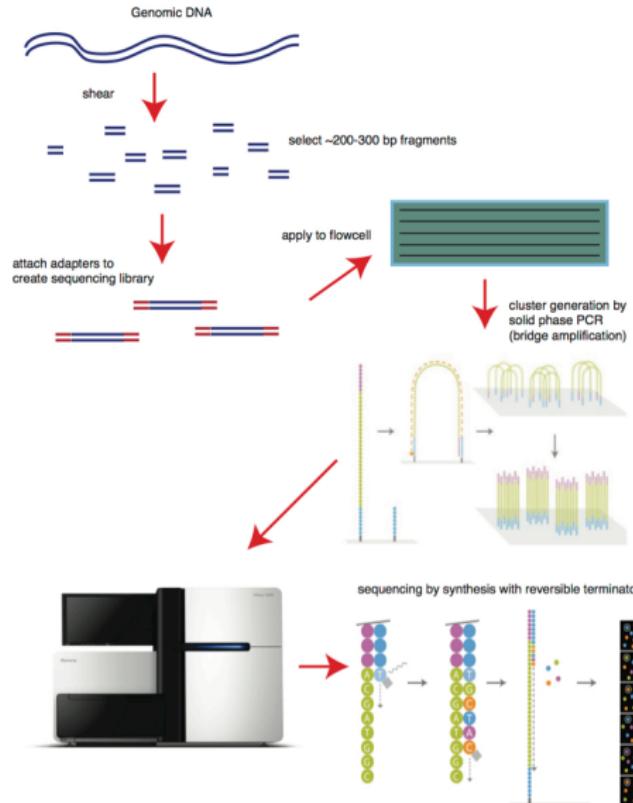
Register each image and follow color change of each colony to determine sequence



Top: CATCAT  
Bottom: CCCCCC



# 测序 | 第二代 | Illumina/Solexa



## 优点

- 通量大
- 测序方式灵活
- 分析软件多样化

## 缺点

- 样本制备过程复杂
- 样本要求相对较高



## 优点

- 通量大
- 测序方式灵活
- 分析软件多样化

## 缺点

- 样本制备过程复杂
- 样本要求相对较高

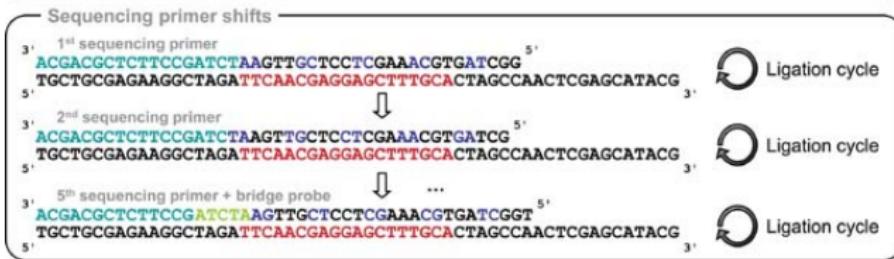
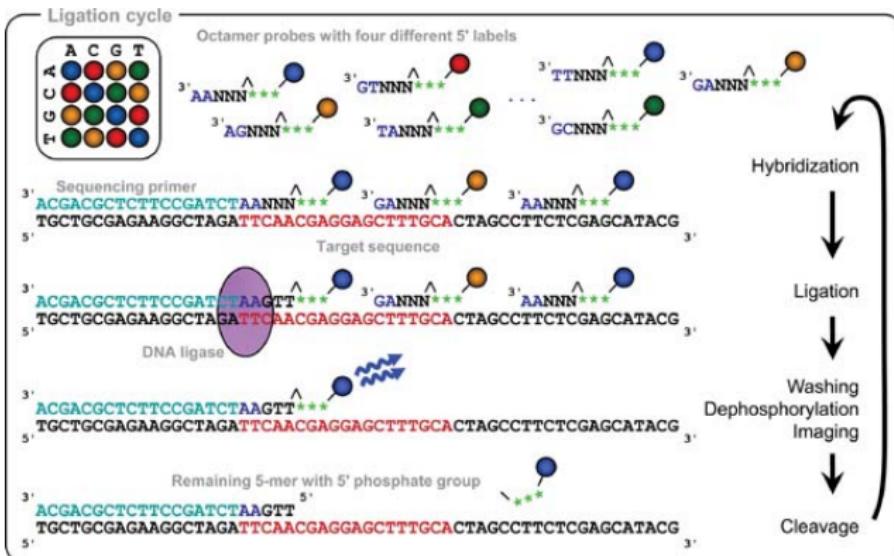


## 边连接边测序

边连接边测序 (sequencing by ligation)，基于连接酶法，即利用 DNA 连接酶在连接过程之中测序。



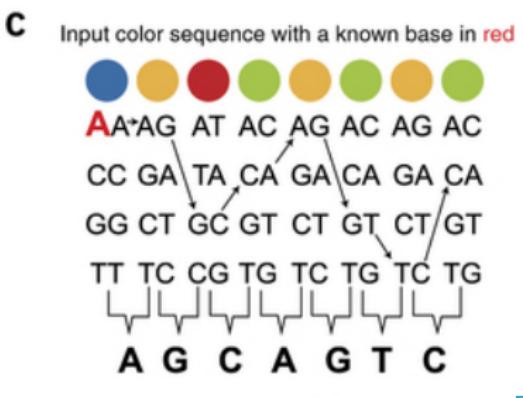
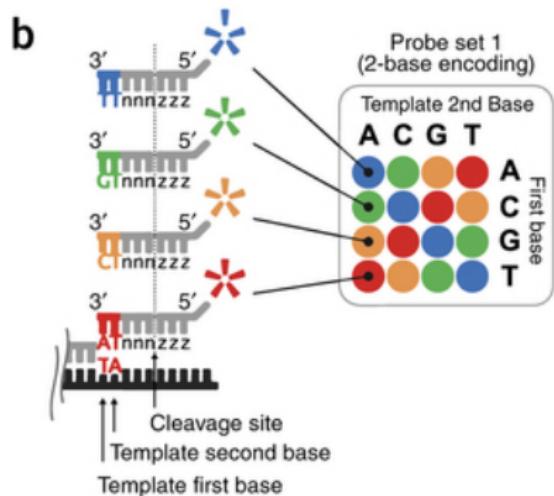
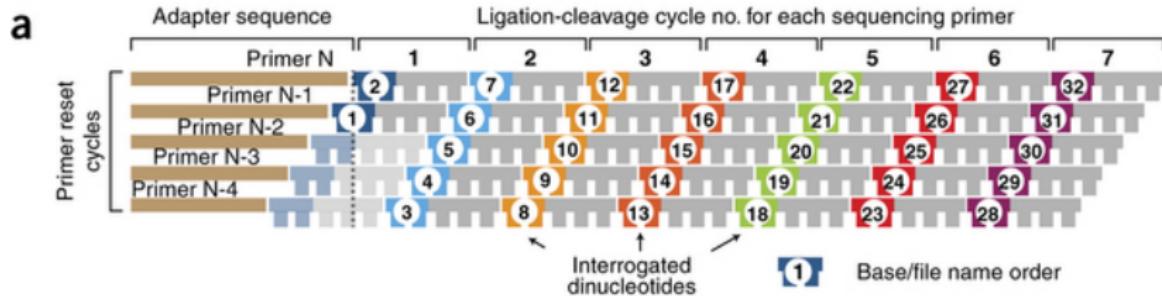
# 测序 | 第二代 | ABI/SOLiD



## 概述

SOLiD 连接反应的底物是 8 碱基单链荧光探针混合物 (3'-XXnnnzzz-5')，其中第 1 和第 2 位碱基 (XX) 上的碱基是确定的，并根据种类的不同在 6-8 位 (zzz) 上加上 CY5、Texas Red、CY3、6-FAM 四种不同的荧光标记。这是 SOLiD 的独特测序法，两个碱基确定一个荧光信号，相当于一次能决定两个碱基，因此也称为两碱基测序法。当荧光探针能够与 DNA 模板链配对而连接上时，就会发出代表第 1、2 位碱基的荧光信号。在记录下荧光信号后，通过化学方法在第 5 和第 6 位碱基之间进行切割，这样就能移除荧光信号，以便进行下一个位置的测序。这种测序方法每次测序的位置都相差 5 位：即第一次是第 1、2 位，第二次是第 6、7 位……在测到末尾后，要将新合成的链变性，洗脱。接着用引物 n-1 进行第二轮测序。引物 n-1 与引物 n 的区别是，二者在与接头配对的位置上相差一个碱基。也即是，通过引物 n-1 在引物 n 的基础上将测序位置往 3' 端移动一个碱基位置，因而就能测定第 0、1 位和第 5、6 位……第二轮测序完成，依此类推，直至第五轮测序，最终可以完成所有位置的碱基测序，并且每个位置的碱基均被检测了两次。

测序 | 第二代 | ABI/SOLiD

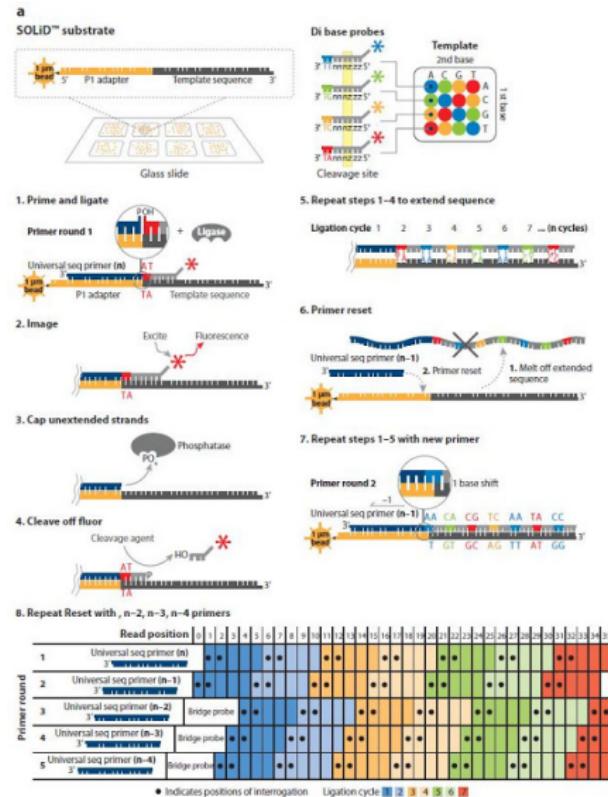


**AAGCAGTCA**

### Output sequence

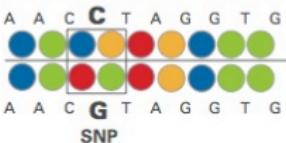


# 测序 | 第二代 | ABI/SOLiD



# 测序 | 第二代 | ABI/SOLiD

SNP site indicated by 2 adjacent color changes



Reference in base space

Reference in color space

Read in color space

Read in base space

Single color change is typically a measurement error



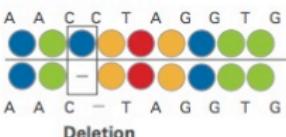
Reference in base space

Reference in color space

Read in color space

Read in base space

1 Base Deletion



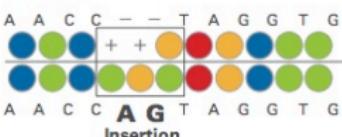
Reference in base space

Reference in color space

Read in color space

Read in base space

Insertion



Reference in base space

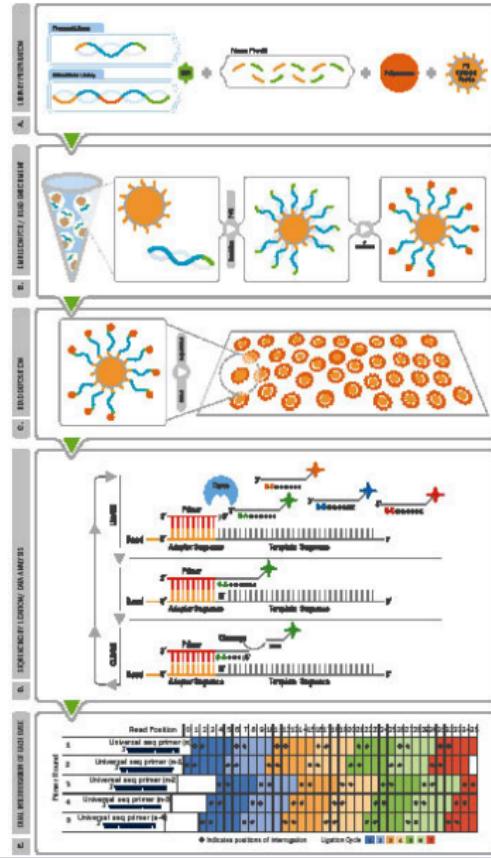
Reference in color space

Read in color space

Read in base space



# 测序 | 第二代 | ABI/SOLiD



## 优点

- 高准确性，每个 DNA 碱基检测 2 次，增加了序列读取的准确性

## 缺点

- 运行时间长，检测碱基替换突变的误差率高



## 优点

- 高准确性，每个 DNA 碱基检测 2 次，增加了序列读取的准确性

## 缺点

- 运行时间长，检测碱基替换突变的误差率高

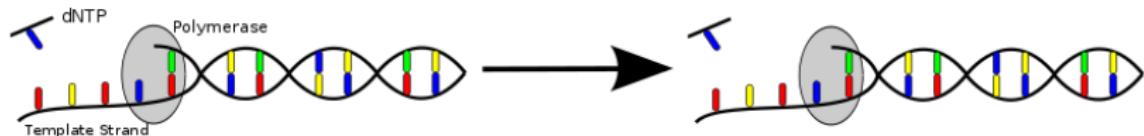


## 概述

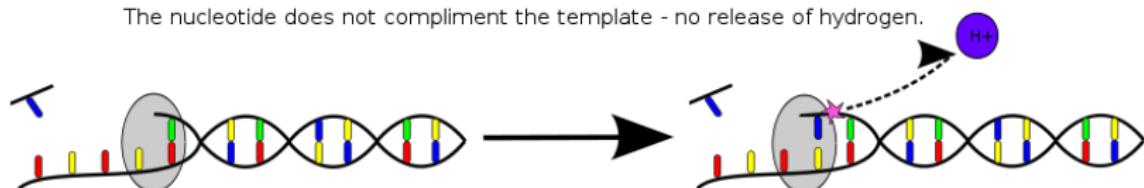
Ion Torrent (Ion semiconductor sequencing) 是一种基于半导体芯片的新一代革命性测序技术，通过检测 H<sup>+</sup> 信号的变化来获得序列碱基信息。该技术使用了一种布满小孔的高密度半导体芯片，一个小孔就是一个测序反应池，芯片置于一个离子敏感层和离子感受器之上。当 DNA 聚合酶把核苷酸聚合到延伸中的 DNA 链上时，会释放出一个氢离子，反应池中的 pH 发生改变，位于池下的离子感受器感受到 H<sup>+</sup> 离子信号，H<sup>+</sup> 离子信号再直接转化为数字信号，从而读出 DNA 序列。



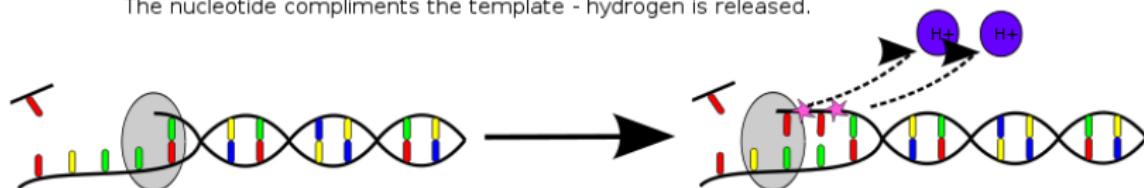
# 测序 | 第 2.5 代 | 离子半导体测序



The nucleotide does not compliment the template - no release of hydrogen.



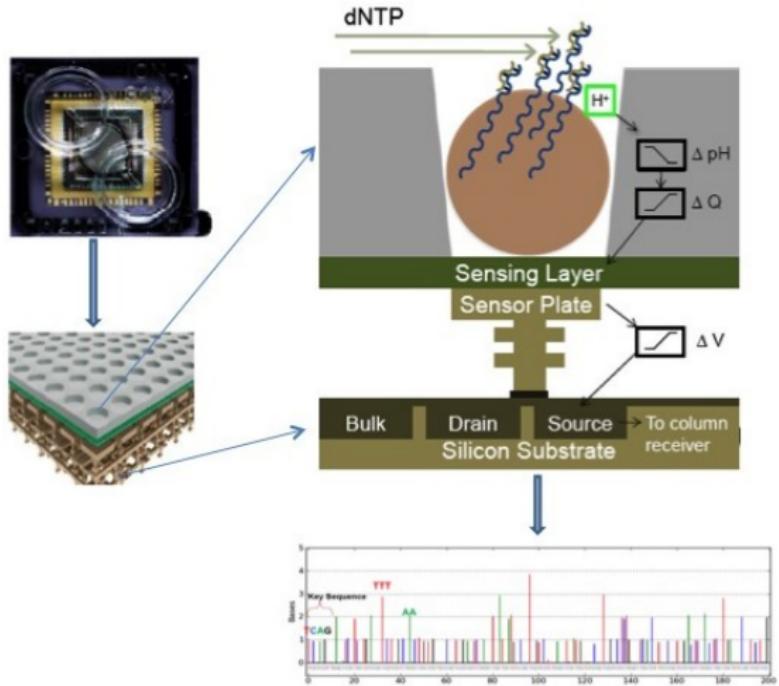
The nucleotide compliments the template - hydrogen is released.



The nucleotide compliments several bases in a row - multiple hydrogen ions are released.



# 测序 | 第 2.5 代 | 离子半导体测序



dNTP流经反应池并发生插入时，释放的 $H^+$ 引起pH值的变化( $\Delta pH$ )

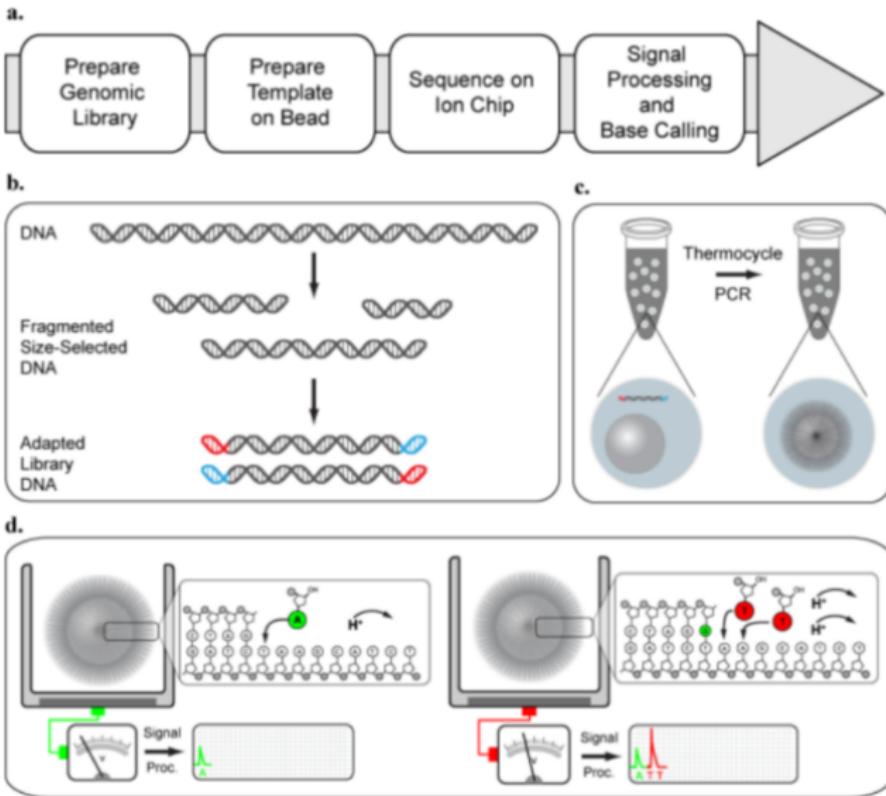
反应池底部金属-氧化物-传感层表面电势变化

电势的变化引起底部场效应晶体管终端的电压( $\Delta V$ )的变化

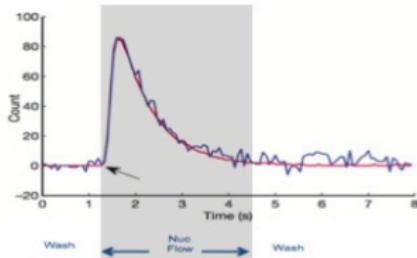
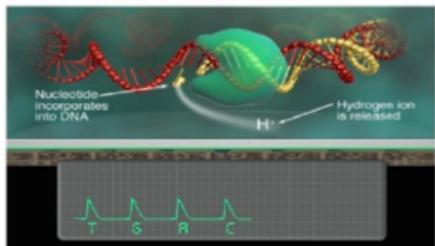
根据电压( $\Delta V$ )变化读取流经反应池的碱基，从而读取测序序列



# 测序 | 第 2.5 代 | 离子半导体测序



## PGM测序特点——优势



### 更易升级

- 上市第一年通量100X
- 半导体技术
- 升级遵循Moore定律

### 更简单

- 无标记核苷酸
- 无激光光源
- 无光学系统
- 无照相系统
- 无荧光
- 无酶促级联反应

### 更快速

- 碱基插入1-2/s
- 标准测序时间仅2-4.5h



## 优缺点

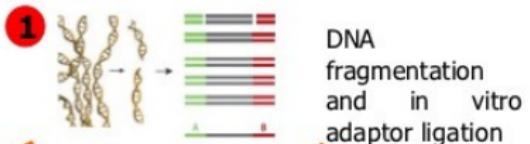
Ion Torrent 相比于其他测序技术来说，不需要昂贵的物理成像等设备，因此，成本相对来说会低，体积也会比较小，同时操作也要更为简单，速度也相当快速，除了 2 天文库制作时间，整个上机测序可在 2-3.5 小时内完成，不过整个芯片的通量并不高，目前是 10G 左右，但非常适合小基因组和外显子验证的测序。

Ion Torrent 的化学测序原理自然简单，无修饰的核苷酸、无激光器或光学检测设备，因而可达到极小的测序偏差和出色的测序覆盖均衡度。



## Next-generation DNA sequencing

- 1 Library preparation
  - 2 Clonal amplification
  - 3 Cyclic array sequencing

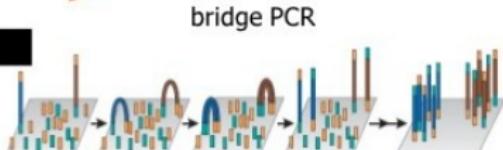


2



3

### Pyrosequencing



The diagram illustrates the sequencing by synthesis process. A polymerase enzyme (A) is shown with its active site at the bottom. A DNA template strand is positioned, with the 3' end labeled. The polymerase is using ATP as a substrate, catalyzed by PP<sub>i</sub>. The reaction product is a new DNA strand extending from the template. This new strand is labeled with a yellow starburst symbol representing fluorescence. To the right, a graph plots Intensity against Time, showing a single sharp peak corresponding to the fluorescence signal.



# 测序 | 第二代 | 比较

## Roche 454

- Long fragments
- Low throughput
- Expensive
- Poly nts errors
- De novo sequencing
- Amplicon sequencing
- Metagenomics
- RNASeq

## Illumina

- Short fragments
- High throughput
- Cheap
- GC bias
- Resequencing
- De novo sequencing
- ChipSeq
- RNASeq
- MethylSeq

## SOLID

- Short fragments
- High throughput
- Cheap
- Color-space
- Resequencing
- ChipSeq
- RNASeq
- MethylSeq



# 教学提纲

1

系统生物学

2

基因组学

3

测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- **第三代测序技术**
- 测序技术的比较

4

数据库与数据格式

5

二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

外显子组测序

- 简介
- 操作流程
- 应用实例

7

转录组学

8

RNA-Seq

- 概述
- 数据分析
- 应用实例

9

顺反组

- 概述
- ChIP-Seq

10

表观遗传学

- 概述
- Methyl-Seq



## 单分子测序

测序过程无需进行 PCR 扩增。

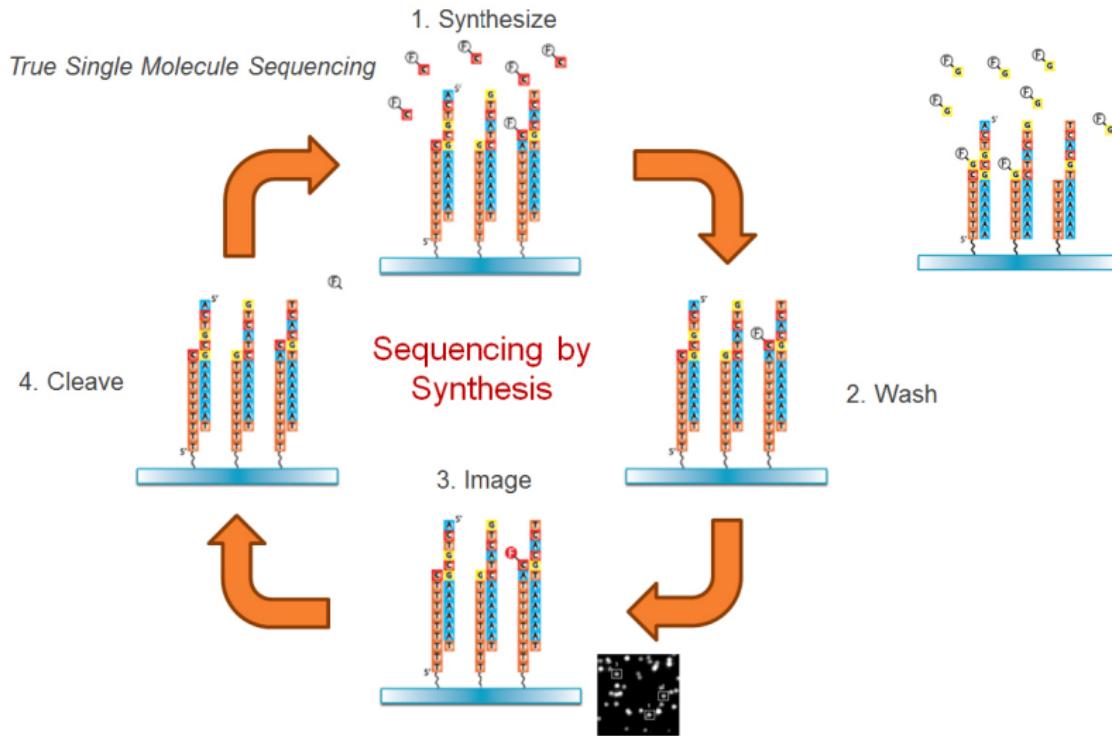


## 概述

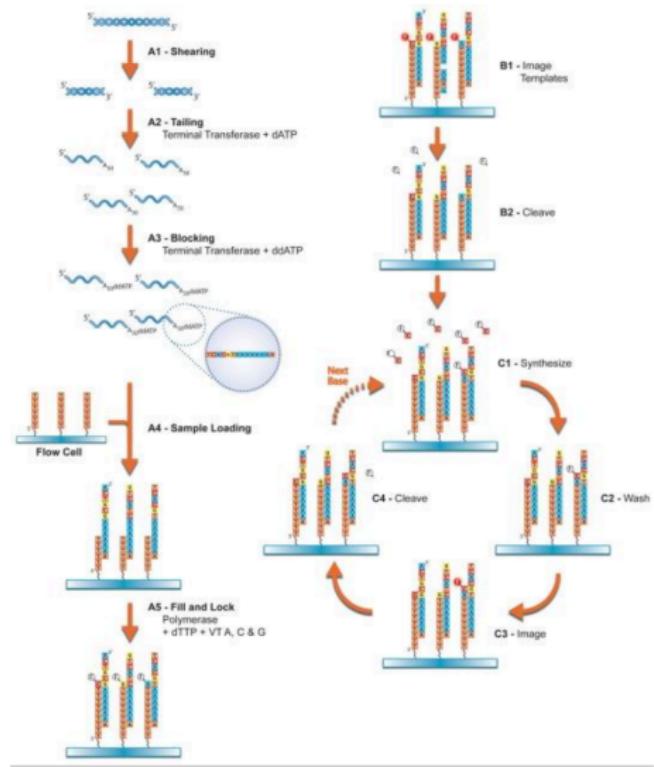
真正的单分子测序 (Helicos True Single Molecule Sequencing)。待测 DNA 被随机打断成小片段，在每个小片段（200bp）的末端加上 poly-dA，并于玻璃芯片上随机固定多个 poly-dT 引物，其末端皆带有荧光标记，以利于精确定位。首先，将小片段 DNA 模板与检测芯片上的 poly-dT 引物进行杂交并精确定位，然后逐一加入荧光标记的末端终止子。这个终止子与 Illumina 的终止子可不一样，不是四色的，是单色的，也就是说所有终止子都标有同一种染料。在掺入了单个荧光标记的核苷酸后，洗涤，单色成像，之后切开荧光染料和抑制基团，洗涤，加帽，允许下一个核苷酸的掺入。通过掺入、检测和切除的反复循环，即可实时读取大量序列。最后以软件系统辅助，可分析出完整的核酸序列。



# 测序 | 第三代 | tSMS



# 测序 | 第三代 | tSMS



## 优缺点

真正的单分子测序，无需前期扩增，不引入偏向性；特别适合 RNA-Seq 或 RNA 直接测序的应用，因为它能直接测序 RNA 模板，而无需将其转化成 cDNA。检测碱基替换突变的误差率非常低， $\sim 0.2\%$ 。

缺点：错误率高，Insertion 1.5%，Deletion 3.0%；Heliscope 在面对同聚物时也会遇到一些困难，但可以通过二次测序提高准确度；由于在合成中可能掺有未标记的碱基，因此其最主要错误来源是缺失。



## 概述

PacBio SMRT (single molecule real time sequencing) 技术也应用了边合成边测序的思想，并以 SMRT 芯片为测序载体。

基本原理是：DNA 聚合酶和模板结合，4 色荧光标记 4 种碱基（即是 dNTP），在碱基配对阶段，不同碱基的加入，会发出不同光，根据光的波长与峰值可判断进入的碱基类型。

DNA 聚合酶是实现超长读长的关键之一，读长主要跟酶的活性保持有关，它主要受激光对其造成的损伤所影响。



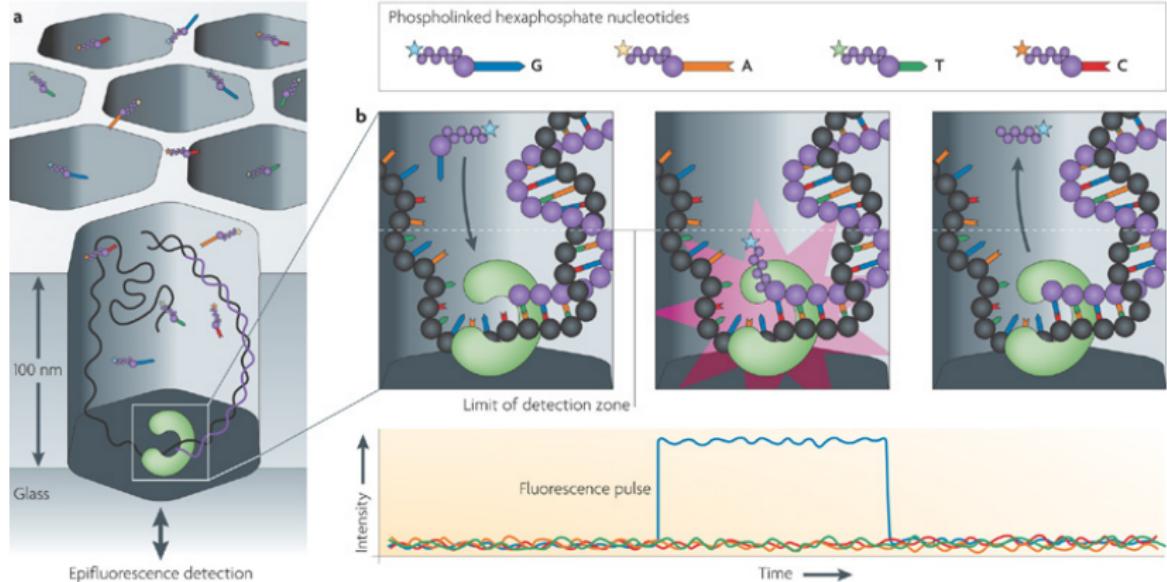
## 概述

PacBio SMRT 技术的一个关键是怎样将反应信号与周围游离碱基的强大荧光背景区别出来。它们利用的是 ZMW (Zero Mode Waveguide, 零模波导孔) 原理，如同微波炉壁上可看到的很多密集小孔。小孔直径有考究，如果直径大于微波波长，能量就会在衍射效应的作用下穿透面板而泄露出来，从而与周围小孔相互干扰。如果孔径小于波长，能量不会辐射到周围，而是保持直线状态（光衍射的原理），从而可起保护作用。同理，在一个反应管 (SMRT Cell, 单分子实时反应孔) 中有许多这样的圆形纳米小孔，即 ZMW (零模波导孔)，外径 100 多纳米，比检测激光波长小 (数百纳米)，激光从底部打上去后不能穿透小孔进入上方溶液区，能量被限制在一个小范围 (体积  $20 \times 10^{-21} L$ ) 里，正好足够覆盖需要检测的部分，使得信号仅来自这个小反应区域，孔外过多游离核苷酸单体依然留在黑暗中，从而实现将背景降到最低。



# 测序 | 第三代 | SMRT

Pacific Biosciences — Real-time sequencing



Nature Reviews | Genetics



## 优缺点

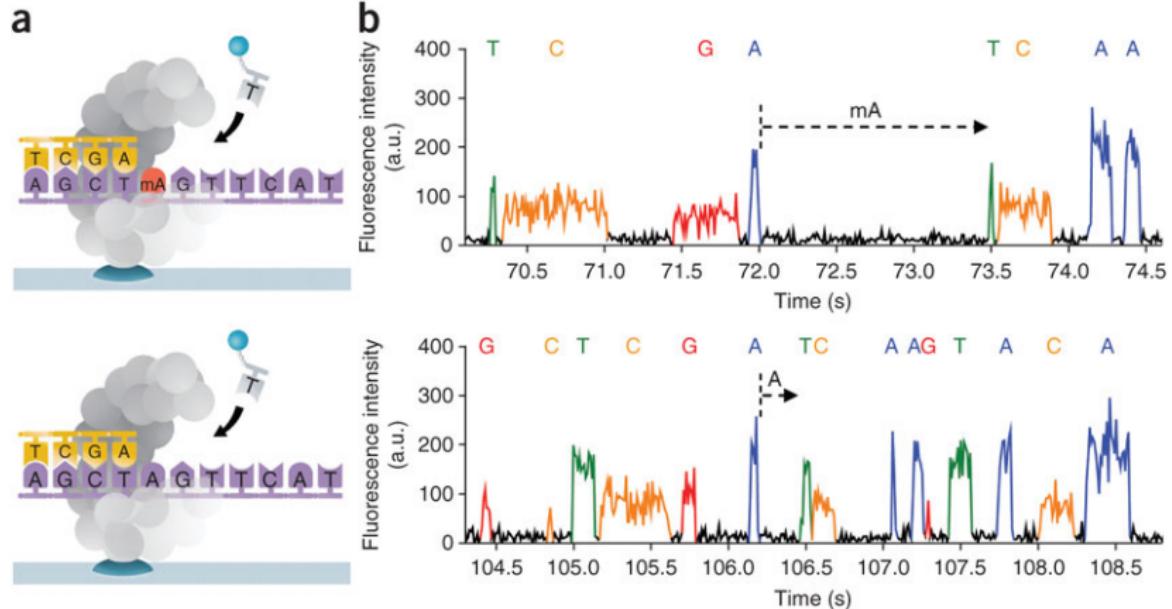
可以通过检测相邻两个碱基之间的测序时间，来检测一些碱基修饰情况，即如果碱基存在修饰，则通过聚合酶时的速度会减慢，相邻两峰之间的距离增大，可以通过这个来直接检测甲基化等信息。

SMRT 技术的测序速度很快，每秒约 10 个 dNTP。读长长。无需 PCR 扩增，也避免了由此带来的 bias。需要的样品量很少，样品制备时间花费少。通量灵活，时间快。可以远程快速获取数据和选择测序参数。

SMRT 技术的测序错误率比较高（这几乎是目前单分子测序技术的通病），达到 15%，但好在它的出错是随机的，并不会像第二代测序技术那样存在测序错误的偏向，因而可以通过多次测序来进行有效的纠错。



# 测序 | 第三代 | SMRT



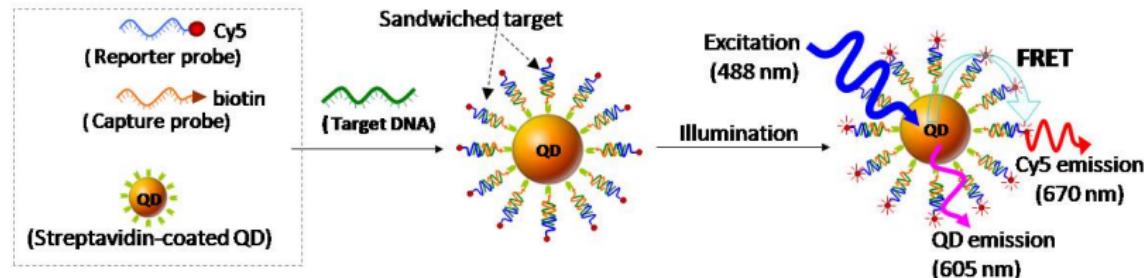
## 概述

VisiGen 基于荧光共振能量转移（FRET，Fluorescence Resonance Energy Transfer）的 DNA 测序技术。将标记了荧光供体基团的 DNA 聚合酶分子固定在载玻片上；再加含模板、引物、四种 dNTP（其磷酸上标记特异的荧光受体基团）的测序缓冲液。

测序延伸反应开始，带荧光受体基团的 dNTP 靠近含荧光供体基团的聚合酶，使后者释放能量，激发前者发出特异的荧光（即 FRET 信号），从而识别相应的碱基序列。当 dNTP 被加上后，荧光基团随磷酸离开，保证下一个 dNTP 能继续被加上。



# 测序 | 第三代 | FRET

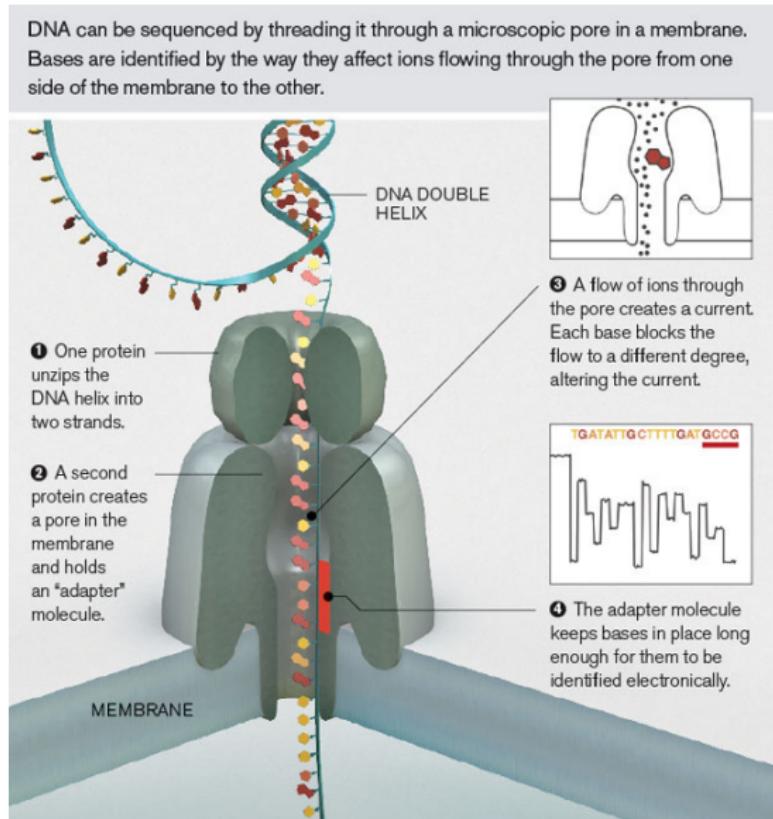


## 概述

Oxford Nanopore Technologies 公司所开发的纳米单分子测序技术 (Nanopore sequencing) 与以往的测序技术皆不同，是基于电信号而不是光信号的测序技术。该技术的关键之一是，它们设计了一种特殊的纳米孔，孔内共价结合有分子接头。当 DNA 碱基通过纳米孔时，它们使电荷发生变化，从而短暂地影响流过纳米孔的电流强度（每种碱基所影响的电流变化幅度是不同的），灵敏的电子设备检测到这些变化从而鉴定所通过的碱基。



# 测序 | 第三代 | 纳米孔测序



## 优缺点

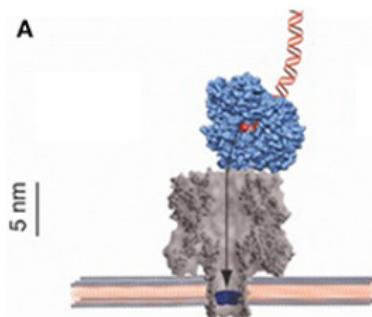
纳米孔测序的主要特点是：读长很长，大约在几十 kb，甚至 100kb；错误率目前介于 1% 至 4%，且是随机错误，而不是聚集在读段的两端；数据可实时读取；通量很高（30x 人类基因组有望在一天内完成）；起始 DNA 在测序过程中不被破坏；样品制备简单又便宜。理论上，它也能直接测序 RNA。

纳米孔单分子测序还有另外一大特点，它能够直接读取出甲基化的胞嘧啶，而不必像传统方法那样对基因组进行 bisulfite 处理。这对于在基因组水平直接研究表观遗传相关现象有极大的帮助。并且该方法的测序准确性可达 99.8%，而且一旦发现测序错误也能较容易地进行纠正。

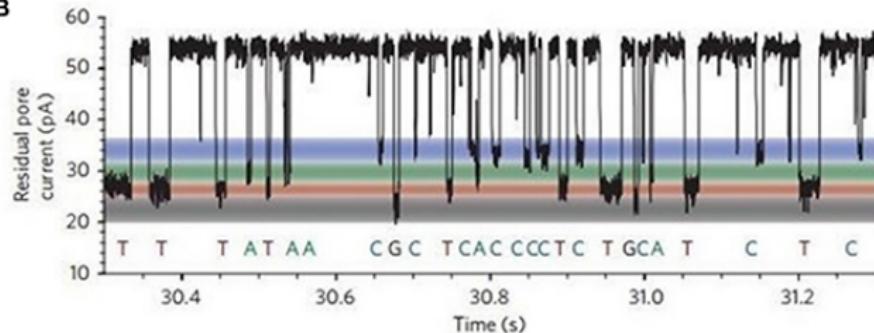


# 测序 | 第三代 | 纳米孔测序

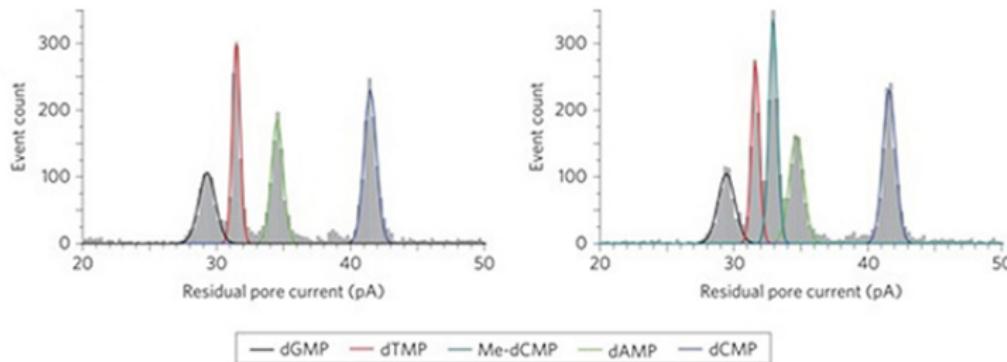
A



B



C

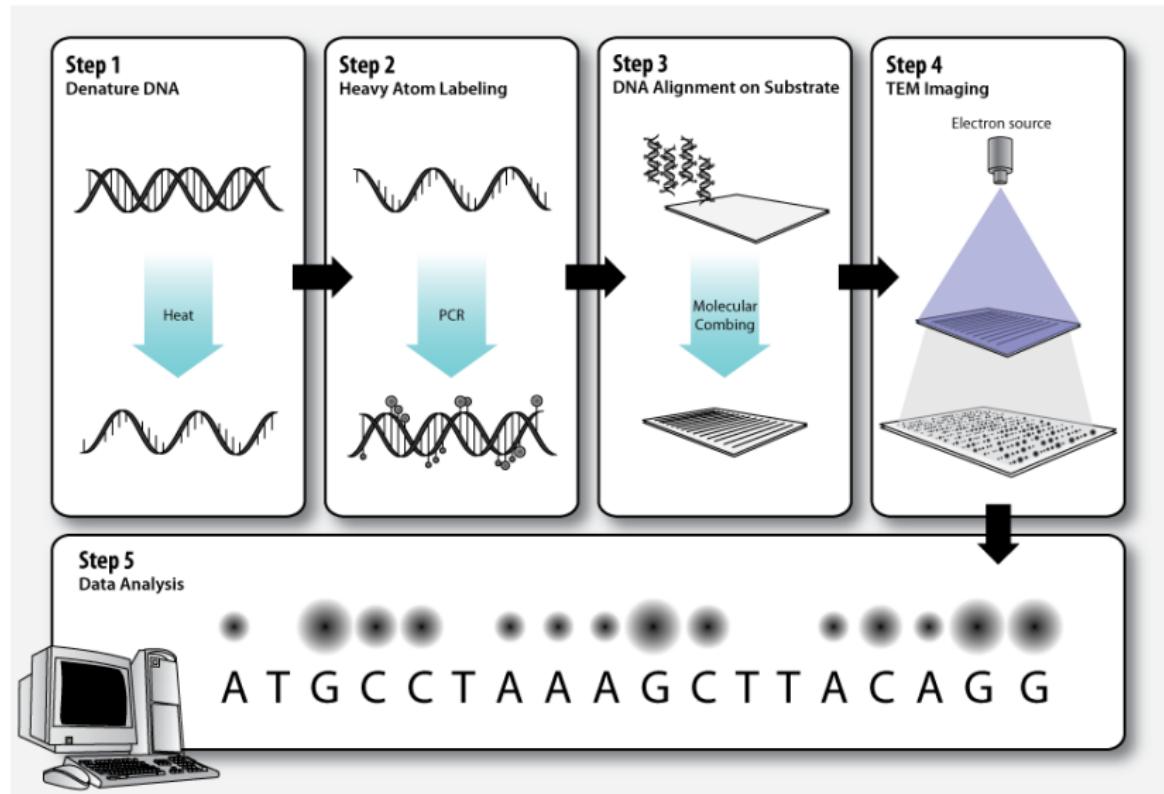


## 概述

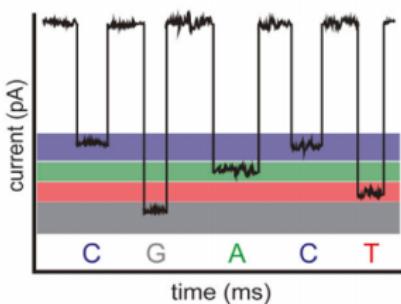
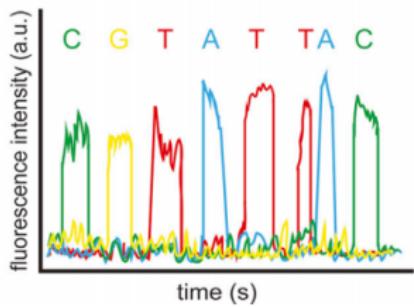
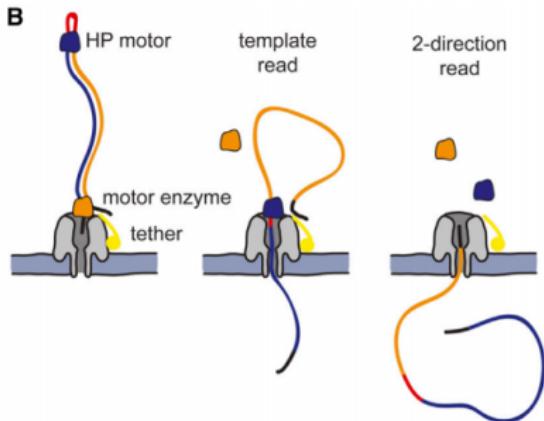
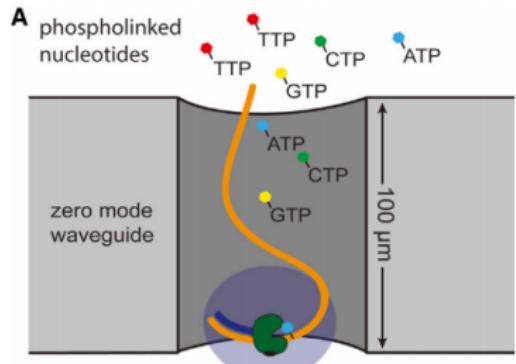
以单链线性 DNA 为模板，以三种重元素标记、一种不标记的脱氧核苷酸为原料，合成其互补链，经透射电镜（TEM, Transmission electron microscopy）检测，则可见重元素标记，其互补链则可由点的大小和强度被分辨出来。



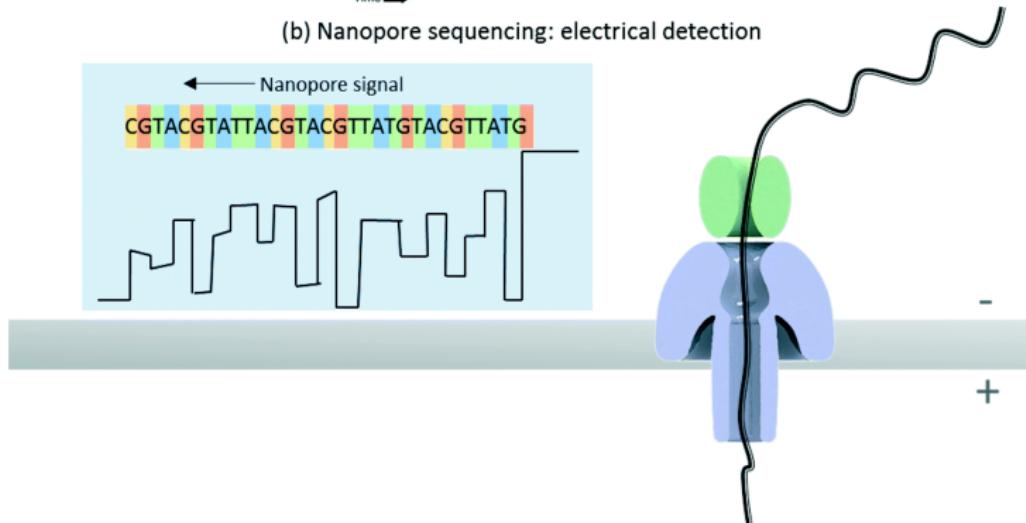
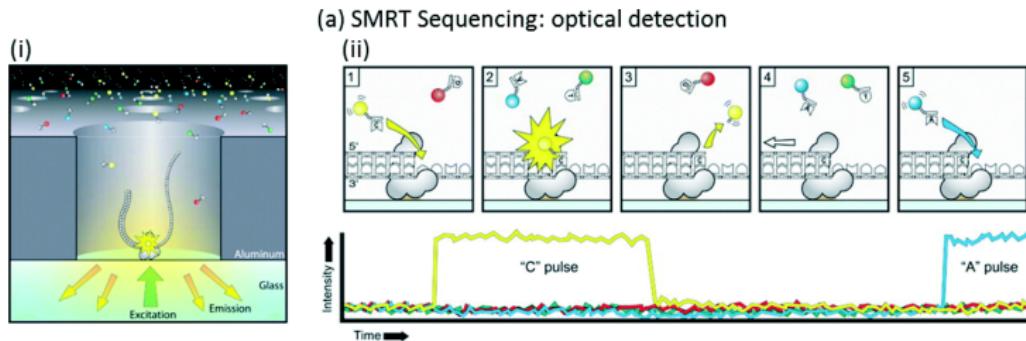
# 测序 | 第三代 | TEM



# 测序 | 第三代 | SMRT, Nanopore



# 测序 | 第三代 | SMRT, Nanopore



# 教学提纲

1

系统生物学

2

基因组学

3

测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

数据库与数据格式

5

二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

外显子组测序

● 简介

● 操作流程

● 应用实例

7 转录组学

8 RNA-Seq

● 概述

● 数据分析

● 应用实例

9 顺反组

● 概述

● ChIP-Seq

10 表观遗传学

● 概述

● Methyl-Seq

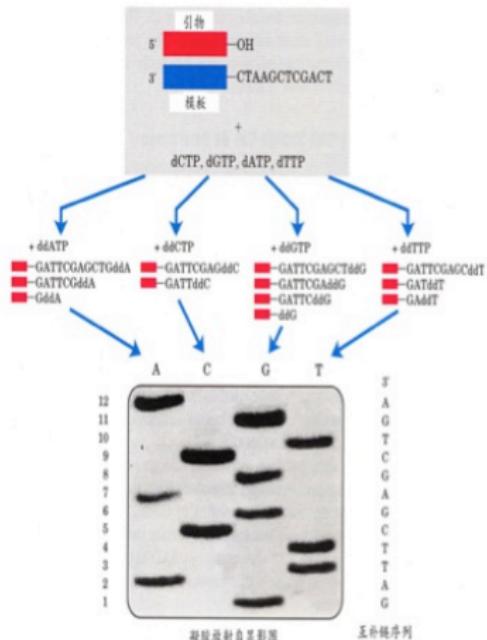
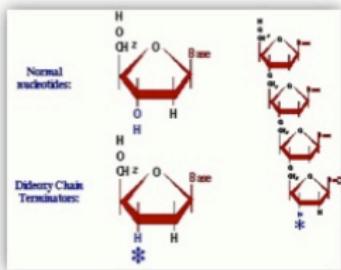




## 第一代测序技术—Sanger 测序法



Dr. Fred Sanger



优点：

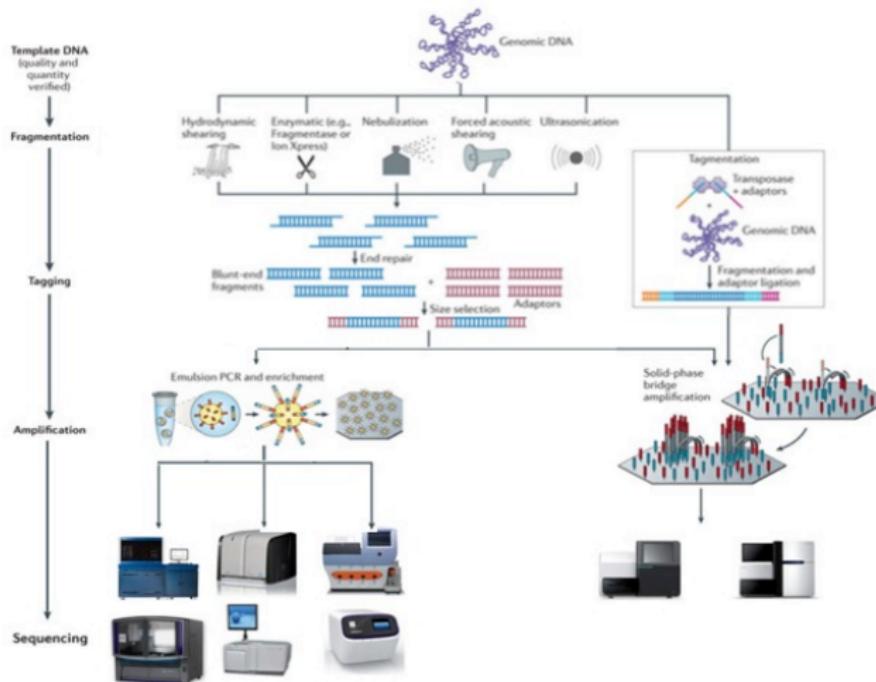
- ✓ 准确
- ✓ 读长长

缺点：

- ✓ 通量低
- ✓ 速度慢
- ✓ 成本高



## 第二代测序技术



优点：

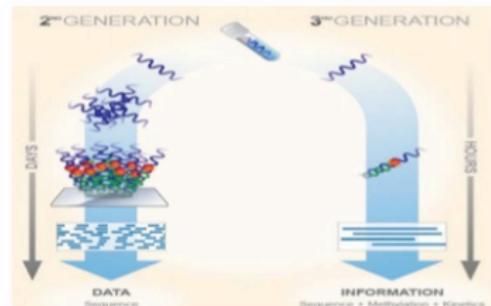
- ✓ 通量高
- ✓ 成本低
- ✓ 时间短

缺点：

- ✓ 读长短
- ✓ 效率不一致



## 第三代测序技术



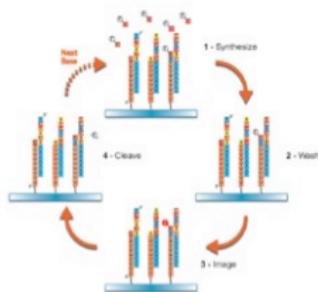
Difference between 2 and 3 generation

优点:

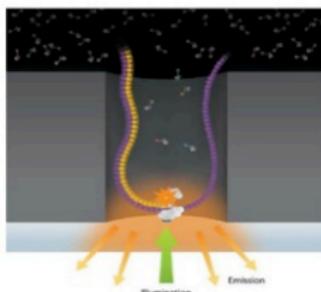
- ✓ 无扩增
- ✓ 直接观察
- ✓ 速度快
- ✓ 长读长

缺点:

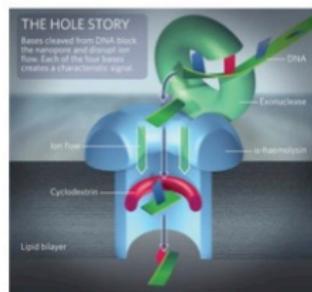
- ✓ 错误率高
- ✓ 可靠性差



Helicos Sequencing



Pacific Biosciences SMRT



Nanopore Sequencing



## 各代测序技术特点总结



	First generation	Second generation <sup>a</sup>	Third generation <sup>a</sup>
Fundamental technology	Size-separation of specifically end-labeled DNA fragments, produced by SBS or degradation	Wash-and-scan SBS	SBS, by degradation, or direct physical inspection of the DNA molecule
Resolution	Averaged across many copies of the DNA molecule being sequenced	Averaged across many copies of the DNA molecule being sequenced	Single-molecule resolution
Current raw read accuracy	High	High	Moderate
Current read length	Moderate (800–1000 bp)	Short, generally much shorter than Sanger sequencing	Long, 1000 bp and longer in commercial systems
Current throughput	Low	High	Moderate
Current cost	High cost per base	Low cost per base	Low-to-moderate cost per base
	Low cost per run	High cost per run	Low cost per run
RNA-sequencing method	cDNA sequencing	cDNA sequencing	Direct RNA sequencing and cDNA sequencing
Time from start of sequencing reaction to result	Hours	Days	Hours
Sample preparation	Moderately complex, PCR amplification not required	Complex, PCR amplification required	Ranges from complex to very simple depending on technology
Data analysis	Routine	Complex because of large data volumes and because short reads complicate assembly and alignment algorithms	Complex because of large data volumes and because technologies yield new types of information and new signal processing challenges
Primary results	Base calls with quality values	Base calls with quality values	Base calls with quality values, potentially other base information such as kinetics



# 测序 | 比较 | 三代

测序方法	代表仪器平台	测序原理	分析方法	定量属性				优势	劣势	应用场景
				通量	读长	测序时间	准确性			
一代测序	ABI/LIFE3730 ABI/LIFE3500	Sanger 双脱氧终止法	毛细管电泳，荧光检测	0.2Mb	400-900bp	1.6h	>99%	读长 准确度 仪器运转成本	通量 每个碱基的 测序成本	常规测序 各种确认性质测序 引物步查 配合二代测序检测 复杂基因组
二代测序	Illumina Hiseq Illumina Genome Analyer Life Solid Roche/454 GS 系列	边合成边测序，可逆终止法	文库制备，桥式 PCR	400Mb -1.8T	50-300bp	2h-3d	>99%	通量 每个碱基成本	仪器成本 仪器运转成本 读长 样本制备要求	二次测序 突变位点分析 变异分析 染色体免疫共沉淀 RNA 测序
三代测序	PACB PacBio RS Oxford Nanopore	单分子合成测序	无需 PCR, 直接转移到测序芯片测序	0.2-30 Gb	>1000bp	2h	<90%	读长 运行时间 样本制备要求 仪器运转成本	通量 仪器成本 准确度	微生物测序 复杂基因组



# 教学提纲

1

## 系统生物学

2

## 基因组学

3

## 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

## 数据库与数据格式

5

## 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

## 外显子组测序

## ● 简介

## ● 操作流程

## ● 应用实例

## 7 转录组学

## 8 RNA-Seq

## ● 概述

## ● 数据分析

## ● 应用实例

## 9 顺反组

## ● 概述

## ● ChIP-Seq

## 10 表观遗传学

## ● 概述

## ● Methyl-Seq



## SRA

NCBI 在 2007 年底推出了 SRA 数据库，专门用于存储、显示、提取和分析高通量测序数据。

SRA 数据库，最初命名为 Short Read Archive，现已改为 Sequence Read Archive。

Sequence Read Archive (SRA) makes biological sequence data available to the research community to enhance reproducibility and allow for new discoveries by comparing data sets. The SRA stores raw sequencing data and alignment information from high-throughput sequencing platforms, including Roche 454 GS System®, Illumina Genome Analyzer®, Applied Biosystems SOLiD System®, Helicos Heliscope®, Complete Genomics®, and Pacific Biosciences SMRT®.

## GEO

NCBI 的 GEO (Gene Expression Omnibus) 数据库是一个非常强大的高通量数据集合，它综合了大量的芯片数据和二代测序数据，供全球科研工作者免费使用。

NCBI 的 GEO 数据库用于存储高通量的芯片实验数据，在 SRA 未建立之前，GEO 数据库也用于存储高通量测序数据。

GEO is a public functional genomics data repository supporting MIAME-compliant data submissions. Array- and sequence-based data are accepted. Tools are provided to help users query and download experiments and curated gene expression profiles.



## 千人基因组计划

千人基因组计划（1000 Genomes Project），旨在绘制迄今（截至 2011 年）最详尽、最有医学应用价值的人类基因多态性图谱，该图谱由中美英等国科研机构发起的“千人基因组计划”共同协作完成，标志着人类基因研究取得重大突破。

这项计划于 2008 年启动，目前该项目拥有超过 1700 个样本，高达 200TB 数据量的 DNA 序列。2012 年开始全部数据免费对外开放。



## TCGA

Cancer Genome Atlas (TCGA) 和 International Cancer Consortium (ICGC) 是目前国际上最大的两个癌症基因信息检索数据库，共收集了 43 种癌症的超过 13 万个样本数据，此外还涉及到相关癌症基因的 mRNA/microRNA 表达谱、拷贝数变异、突变等大量的生物信息学数据。





International  
Cancer Genome  
Consortium

Enter keywords

Search

Home

Cancer Genome Projects

Committees and Working Groups

Policies and Guidelines

Media

## ICGC Cancer Genome Projects

Committed projects to date: [79](#)

Sort by: [Project](#) 

Biliary Tract Cancer

Japan 

Biliary Tract Cancer

Singapore 

Bladder Cancer

China 

Bladder Cancer

United States 

Blood Cancer

China 

Blood Cancer

Singapore 

Blood Cancer

South Korea 

Blood Cancer

United States 

Blood Cancer

United States 

Bone Cancer

France 

Bone Cancer

United Kingdom 

Brain Cancer

Canada 

**ICGC Goal:** To obtain a comprehensive description of genomic, transcriptomic and epigenomic changes in 50 different tumor types and/or subtypes which are of clinical and societal importance across the globe.

[Read more »](#)

[Launch Data Portal »](#)

[Apply for Access to Controlled Data »](#)

### Announcements

16/May/2016 - The ICGC Data Coordination Center (DCC) is pleased to announce ICGC data portal data release 21 (<http://dcc.icgc.org>).

ICGC data release 21 in total comprises data from more



## ICGC Data Portal

[Cancer Projects](#)[Advanced Search](#)[Data Analysis](#)[DCC Data Releases](#)[Data Repositories](#)

e.g. BRAF, KRAS G12D, DO35100, MU7870, FI998, apoptosis, Cancer Gene Census, imatinib, GO:0016049

### About Us

The ICGC Data Portal provides tools for visualizing, querying and downloading the data released quarterly by the consortium's member projects.

To access ICGC controlled tier data, please read these [instructions](#).

New features will be regularly added by the DCC development team. [Feedback is welcome.](#)



**PCAWG**  
PanCancer Analysis  
OF WHOLE GENOMES

The PanCancer Analysis of Whole Genomes (PCAWG) study is an international collaboration to identify common patterns of mutation in more than 2,800 cancer whole genomes from the International Cancer Genome Consortium.

### Data Release 21 May 16th, 2016

#### Donor Distribution by Primary Site



Cancer projects	68
Cancer primary sites	21
Donors with molecular data in DCC	15,613
Total Donors	18,677
Simple somatic mutations	42,584,179

### Tutorial

#### EXAMPLE QUERIES

1. BRAF missense mutations in colorectal cancer
2. Most frequently mutated genes by high impact mutations in stage III malignant lymphoma
3. Brain cancer donors with frameshift mutations and having methylation data available

**ICGC**  
International  
Cancer Genome  
Consortium



ICGC data is now available on commercial and academic compute cloud. [Read more...](#)



**GTEX Portal**

2017-09-18  
V7 Data Released.  
[Read More >>](#)

**Current Release**

Latest Version: V7  
Dataset Summary Statistics Report

How to cite or acknowledge the GTEX project?

Total samples in all eQTL tissues: 10294

Browse eQTL Tissues

View eQTL data of a gene...  
[Tau View Clean eQTL >>](#)

**Genetic Association**

Single-Tissue eQTLs

IGV eQTL Browser

Gene eQTL Visualizer

View eQTL data of a gene...  
[Tau View Clean eQTL >>](#)

**Transcriptome**

Search expression by gene ID...  
Top 100 Expressed Genes in a Tissue (e.g. Blood)

Gene Expression in Tissues

Exon and Isoform Expression

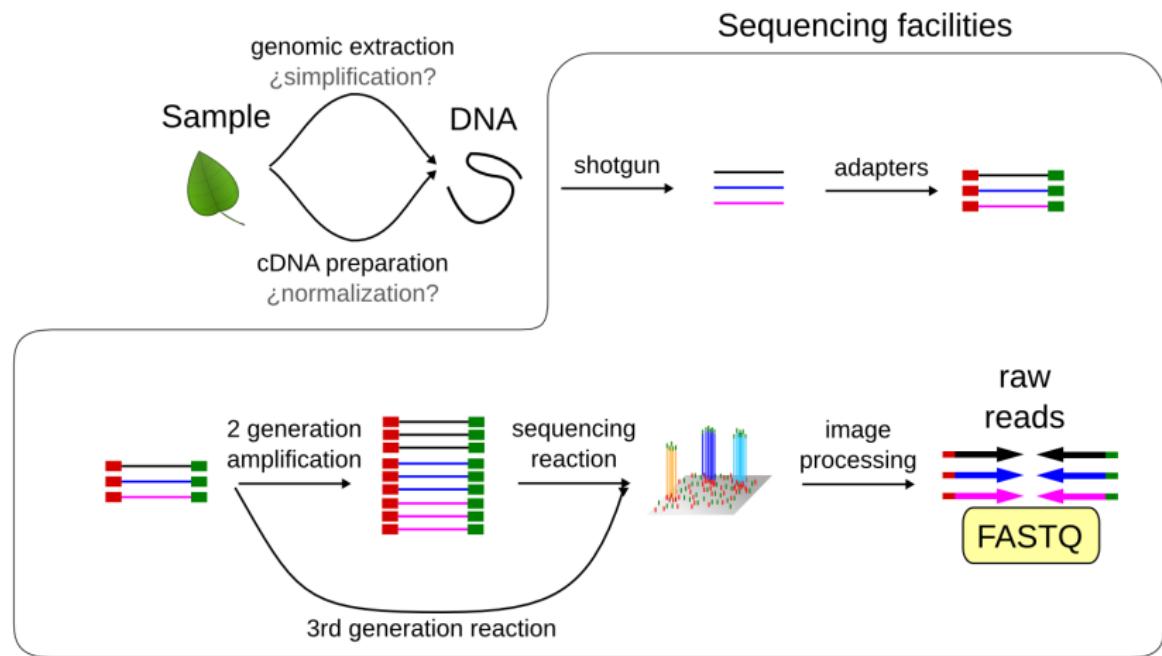
News & Events

GTEX Phase 2 Papers Published (2017-10-18)  
[Read More >>](#)

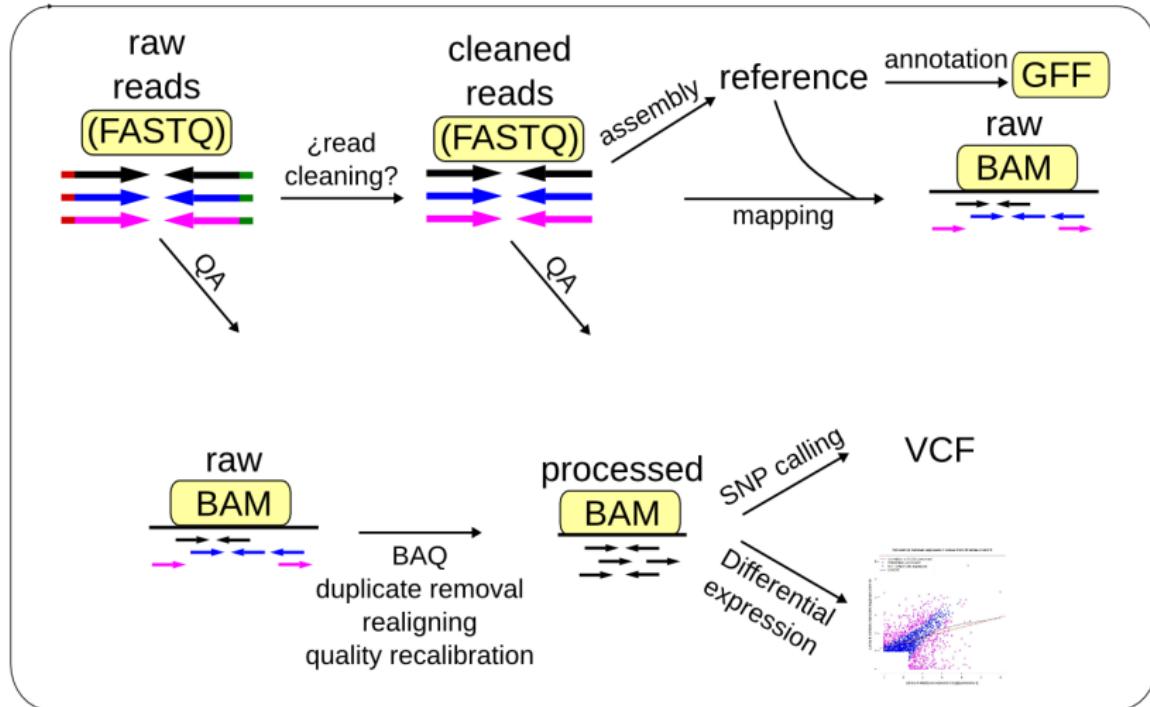
V7 Data Released (2017-09-18)  
[Read More >>](#)



# NGS | 数据格式 | 概览



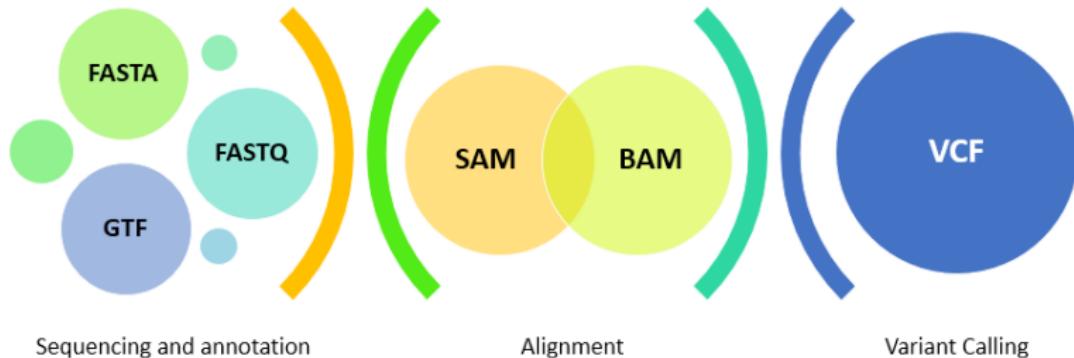
# NGS | 数据格式 | 概览



## Common NGS Data Formats

File extension	Description	Reference	Publication
.fasta	Classic DNA sequence file format	<a href="http://www.ncbi.nlm.nih.gov/blast/fasta.shtml">http://www.ncbi.nlm.nih.gov/blast/fasta.shtml</a>	n/a
.ace	File format for whole-genome assemblies	Annotated in the documentation for CONSED, currently: <a href="http://www.phrap.org/conseq/distributions/README.19.0.txt">http://www.phrap.org/conseq/distributions/README.19.0.txt</a>	Gordon, Abajian, and Green, 1998
.wig	A reference-genome indexed data series for "dense" and continuous data (such as %GC)	<a href="http://genome.ucsc.edu/goldenPath/help/wiggle.html">http://genome.ucsc.edu/goldenPath/help/wiggle.html</a>	Haussler, 2002
.bed	A reference-genome indexed data series for "sparse" data (such as transcriptome data)	<a href="http://genome.ucsc.edu/goldenPath/help/bedgraph.html">http://genome.ucsc.edu/goldenPath/help/bedgraph.html</a>	Haussler, 2002
.tab	Tab-delimited text	N/A	n/a
.pdf	Portable document format	Either ISO-32000-1 or <a href="http://www.adobe.com/devnet/pdf/pdf_reference.html">http://www.adobe.com/devnet/pdf/pdf_reference.html</a>	n/a
.sam	"Sequence Alignment/Map" format	<a href="http://samtools.sourceforge.net/SAM1.pdf">http://samtools.sourceforge.net/SAM1.pdf</a>	Li, 2009
.bam	Binary format of .sam	<a href="http://samtools.sourceforge.net/SAM1.pdf">http://samtools.sourceforge.net/SAM1.pdf</a>	Li, 2009
.fastq	Combination of sequences and quality scores in one file; mainly for data from Illumina sequencers in which case quality scores have been transformed.	<a href="http://maq.sourceforge.net/fastq.shtml">http://maq.sourceforge.net/fastq.shtml</a>	Li, 2008, and Cock, 2009
.csfasta	Life Technologies SOLID colorspace fasta file - containing color calls (0, 1, 2, 3) rather than base calls	See: <a href="http://solidsoftwaretools.com/">http://solidsoftwaretools.com/</a>	n/a
.qual	Per-base quality scores generated during basecalling. All but Illumina scores are scaled to estimate the probability of an incorrect base call, as is in common use for conventional sequencing as Phred quality scores.		Ewing & Green, 1998
.gff	A flexible format for annotating features (e.g. genes) on a sequence.	<a href="http://www.sanger.ac.uk/resources/software/gff/">http://www.sanger.ac.uk/resources/software/gff/</a>	n/a
.srf	"Short Read Format" - a new format proposed for short-read DNA sequence	<a href="http://srf.st.net">http://srf.st.net</a>	n/a
.sff	Standard Flowgram Format (specific for Roche/454)	<a href="http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&amp;f=format&amp;m=doc&amp;s=formats#header-global">http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&amp;f=format&amp;m=doc&amp;s=formats#header-global</a>	n/a
.gtf	Gene transfer format, an alternate format to GFF for specifying gene features	<a href="http://mblab.wustl.edu/GTF22.html">http://mblab.wustl.edu/GTF22.html</a>	n/a

For a full list, go to <http://genome.ucsc.edu/FAQ/FAQformat.html>



## Reference sequences

- FASTA
- 2bit

## Reads

- FASTQ (FASTA with quality scores)

## Alignments

- SAM (Sequence Alignment/Map format)
- BAM (Binary version of SAM)



## Reference sequences

- FASTA
- 2bit

## Reads

- FASTQ (FASTA with quality scores)

## Alignments

- SAM (Sequence Alignment/Map format)
- BAM (Binary version of SAM)



## Reference sequences

- FASTA
- 2bit

## Reads

- FASTQ (FASTA with quality scores)

## Alignments

- SAM (Sequence Alignment/Map format)
- BAM (Binary version of SAM)



## Features, annotation, coverage, scores

- GFF3/GTF (General Feature Format, Gene Transfer Format)
- BED/bigBed (Browser Extensible Data)
- WIG/bigWig (Wiggle format)
- bedGraph

## Variations

- VCF (Variant Call Format)
- BCF (Binary version of VCF)



## Features, annotation, coverage, scores

- GFF3/GTF (General Feature Format, Gene Transfer Format)
- BED/bigBed (Browser Extensible Data)
- WIG/bigWig (Wiggle format)
- bedGraph

## Variations

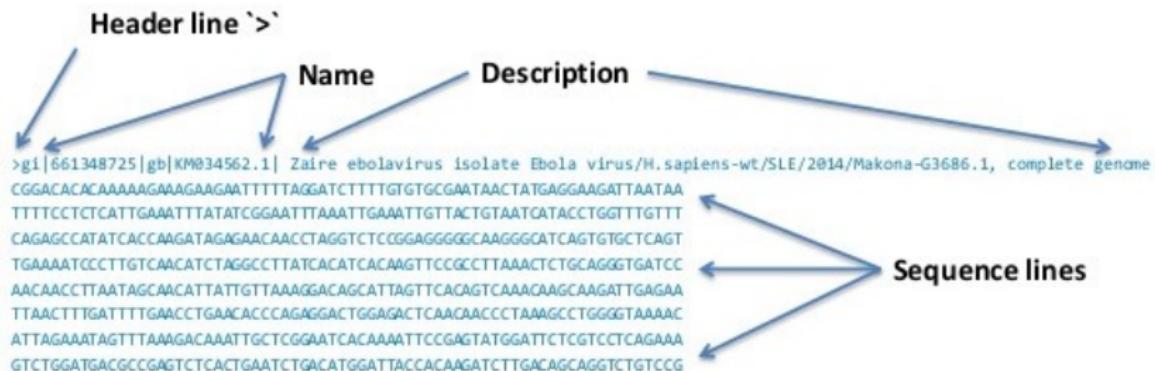
- VCF (Variant Call Format)
- BCF (Binary version of VCF)



## FASTA Format

The FASTA format is a standard for displaying (nucleotide or protein) sequences in a text file. An entry for a sequence takes up two lines in the file: the first line begins with a ">" symbol, followed by the sequence description, and the second line contains the sequence itself.

```
>gi|67328264|gb|AAFC02129962.1| Bos taurus breed Hereford Con136352, whole genome shotgun sequence  
CCCCCCCCCCCGGGCACGTACCTGCTGGATCAGCCCCACCTGGAGCTGGGTGAGGAACAGCTG  
GGGAAGGAAGCAAGCGGCAGTGAGCTGAGGCCGGTGCCGGCAGGCCGCCACCTGGCCC
```



## What is FASTQ?

- Specifies sequence (FASTA) and quality scores (PHRED)
  - Text format, 4 lines per entry

```
 @SEQ_ID  
 GATTTGGGTTCAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTT  
 +  
 +***(((((****+))%%%+)(%%%%, 1***-+*'))))**55CCE>>>>CCCCCCC65
```

- FASTQ is such a cool standard, there are 3 (or 5) of them!



[http://en.wikipedia.org/wiki/FASTQ\\_format](http://en.wikipedia.org/wiki/FASTQ_format)

**SITUATION:**  
THERE ARE  
4 COMPETING  
STANDARDS.

#### HOW STANDARDS PROLIFERATE:

THE AMERICAN JOURNAL OF HISTORICAL STUDIES IN LITERATURE

SOON:

**WE NEED TO REVAMP**

WE FAILED TO REACH  
THE UNIVERSAL STANDARD

NATION: ONE UNIVERSAL STANDARD  
FOR THE WORLD'S TRADE.

THAT COVERS EVERYONE'S

WE ARE USE CASES. THE

COMPETING WITHIN THE  
EUROPEAN UNION

STANDARDS | 100% SUSTAINABLE | STANDARDS

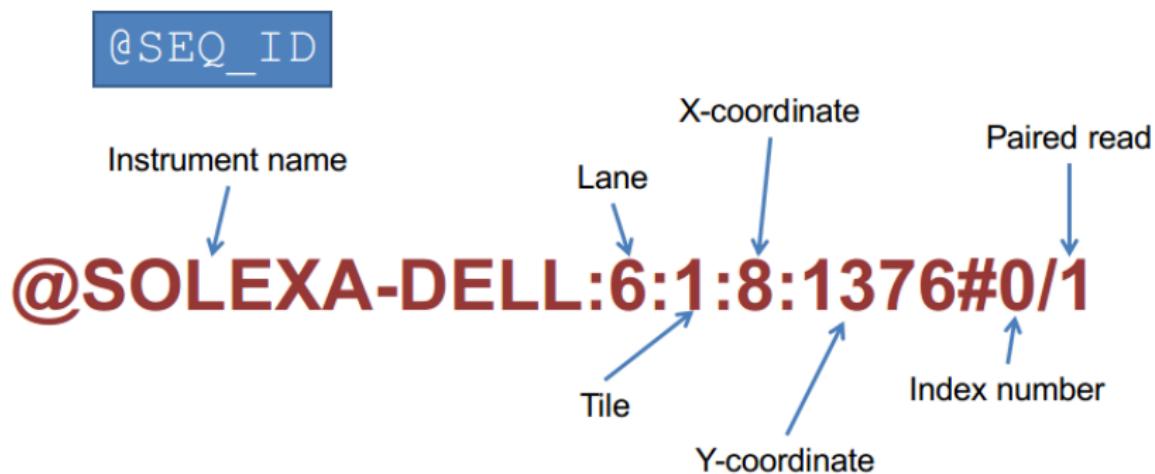
—  
—  
—



# Sequence Encoding (FASTQ)

- Extension from traditional FASTA format
- Each block has 4 elements (in 4 lines):
  - Sequence name (read name, group, etc...)
  - Sequence
  - + (optional: sequence name again)
  - Associated quality scores (phred-scaled) : different encoding possible
- Example record:
  - @FCD19MJACXX:2:1101:1735:1993#GTTCGACA/1
  - NGAGGCTGAGGCAGGGCAGAGGTCAAGGAGATCGAGACCATC
  - +
  - BP\cccccc]ceechheeZbe\_cZbd\_dbbdd\axab\_`b

## Illumina sequence identifiers



```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

## Note

- ID = NCBI-assigned identifier + the original identifier from Solexa/Illumina + the read length
- **fastq-dump**: lost the paired-end information, concatenate sequence of the forward and reverse reads together into a non-sense
- NCBI have converted this FASTQ data from the original Solexa/Illumina encoding to the **Sanger standard**

## Phred quality score

A quality value Q is an integer mapping of p (i.e., the probability that the corresponding base call is incorrect).

Phred quality score (the standard Sanger variant, assess reliability of a base call):

$$Q_{\text{sanger}} = -10 \log_{10} p$$

Phred Quality Score	Probability of Incorrect Base Call	Base Call Accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10,000	99.99%
50	1 in 100,000	99.999%
60	1 in 1,000,000	99.9999%



# NGS | 数据格式 | FASTQ | Quality | Encoding



- S - Sanger      Phred+33, raw reads typically (0, 40)
- X - Solexa      Solexa+64, raw reads typically (-5, 40)
- I - Illumina 1.3+ Phred+64, raw reads typically (0, 40)
- J - Illumina 1.5+ Phred+64, raw reads typically (3, 40)  
with 0=unused, 1=unused, 2=Read Segment Quality Control Indicator (**bold**)  
(Note: See discussion above).
- L - Illumina 1.8+ Phred+33, raw reads typically (0, 41)



## Quality control

- quality score distribution
- GC content
- k-mer enrichment
- ...

## Preprocessing

- adapter removal
- low-quality reads filtering
- ...

## Processing

- alignment
- further analysis

## Quality control

- quality score distribution
- GC content
- k-mer enrichment
- ...

## Preprocessing

- adapter removal
- low-quality reads filtering
- ...

## Processing

- alignment
- further analysis

## Quality control

- quality score distribution
- GC content
- k-mer enrichment
- ...

## Preprocessing

- adapter removal
- low-quality reads filtering
- ...

## Processing

- alignment
- further analysis

# Alignment output files

## SAM

- plain text file, tab separated columns
- "a huge spreadsheet"
- inefficient to read and store

## BAM

- a compressed version of SAM (~80% less storage)
- can be indexed (fast access to subsections)
- needs to be sorted to be useful however

## Standardized format

- readable by most software



## SAM/BAM aligned format

- SAM Format: aligned format, human readable

@SQ SN:chr12 LN:133851895

@RG ID:Sample\_ID LB:Sample\_Library PL:ILLUMINA SM:Sample\_Name PU:Platform\_Unit

Read name	Flag	Chr	5' pos	MAPQ	Cigar	paired	5' pos of the mate	Insert size
ERR166338.1	99	chr12	82670685	23	101M	=	82670850	266
GCCCTGGGATGTTGCACCAAGCCACTGTCTCCAGCTGG							sequence	
BBC@GIIHGFJCIEHEAIEIFFGEONDNJFINIONHNGJNNNNKJN							Base quality	
RG:Z:Sample_ID KT:A:U NM:i:0 X0:i:1 X1:i:1 XM:i:0 XO:i:0 XG:i:0 MD:Z:100 XA:Z							tags	

Group affiliation

- BAM Format: Binary SAM Format (not human readable but compressed = smaller)

### SAM Format Specification

<https://samtools.github.io/hts-specs/SAMv1.pdf>

# NGS | 数据格式 | SAM

reference name  
reference length  
header  
left most position  
• 5' for plus strands  
• 3' for minus strands  
mapping quality (phred scaled)  
cigar string  
strand: 0=plus, 16 =minus, 4=no match  
query sequence on same strand as reference  
query name = sample name:bead coordinates

SN:chr19 LN:61342430  
SN:chrX LN:166650250  
SN:chrY LN:15902555  
SN:chrM LN:16299  
SN:chr1\_random LN:400311  
SN:chr16\_random LN:3994  
SN:chr17\_random LN:628739  
SN:chr1\_random LN:1231697  
SN:chr3\_random LN:41899  
SN:chr4\_random LN:160594  
SN:chr5\_random LN:357350  
SN:chr7\_random LN:362490  
SN:chr8\_random LN:849593  
SN:chr9\_random LN:449403  
SN:chrUn\_random LN:5900358  
SN:chrX\_random LN:1785075  
SN:chrY\_random LN:58682461  
TD:hwa PN:hwa VN:0.5.9-r16  
No\_100\_4\_6\_25:446\_280\_705 0 chr1 3017770 27 33M \* 0 0 ATTTGTTTTTTTGTGTGTGTTCCGGGTGG  
!<%> No\_100\_4\_6\_25:119\_1089\_772 16 chr1 3137326 9 33M XG:i:0 CM:i:1 XM:i:8 X0:i:0 XT:A:O  
XG:i:1 X1:i:24 MD:Z:3T1A0T26 \* 0 0 TTAAGCAGCATGAACATTGTATGTACACTGAA  
XG:i:0 CM:i:3 XM:i:7 X0:i:0 XT:A:U

# NGS | 数据格式 | SAM

```
Coor      12345678901234 5678901234567890123456789012345  
ref       AGCATGTTAGATAA**GATAGCTGTGCTAGTAGGCAGTCAGGCCAT  
  
+r001/1      TTAGATAAAGGATA*CTG  
+r002      aaaAGATAA*GGATA  
+r003      gcctaAGCTAA  
+r004      ATAGCT.....TCAGC  
-r003      ttagctTAGGC  
-r001/2      CAGCGGCAT
```

@HD VN:1.5 SO:coordinate

@SQ SN:ref LN:45

```
r001 99 ref 7 30 8M2I4M1D3M = 37 39 TTAGATAAAGGATACTG *  
r002 0 ref 9 30 3S6M1P1I4M * 0 0 AAAAGATAAGGATA *  
r003 0 ref 9 30 5S6M * 0 0 GCCTAAGCTAA * SA:Z:ref,29,-,6H5M,17,0;  
r004 0 ref 16 30 6M14N5M * 0 0 ATAGCTTCAGC *  
r003 2064 ref 29 17 6H5M * 0 0 TAGGC * SA:Z:ref,9,+,5S6M,30,1;  
r001 147 ref 37 30 9M = 7 -39 CAGCGGCAT * NM:i:1
```



## BAM: the binary version of SAM

- SAM files are large: 1M short reads => 200MB; 100M short reads => 20GB.
- Makes sense for compression
- BAM: Binary sAM; compress using gzip library.
- Two parts: compressed data + index
- Index: random access (visualization, analysis, etc.)



# BED format

- Text-based, tab-delimited format for storing signals for intervals
  - 3 required fields: chrom, chromStart, chromEnd
  - 9 optional fields: name, score, strand, thickStart, thickEnd, itemRgb, blockCount, blockSizes, blockStarts (last ones for visualization)
  - Example:

```
chr22 1000 5000 cloneA 960 + 1000 5000 0 2 567,488, 0,3512  
chr22 2000 6000 cloneB 900 - 2000 6000 0 2 433,399, 0,3601
```

- There is also a binary format called BigBed with more efficient data access
- Many variations, such as the commonly-used bedGraph format with only 4 fields: chrom, chromStart, chromEnd, dataValue



# NGS | 数据格式 | BED

chr1	817371	819837	ENSG00000177757.2_FAM87B_lincRNA	0+
chr1	826206	827522	ENSG00000225880.5_LINC00115_lincRNA	0-
chr1	827608	859446	ENSG00000228794.5_LINC01128_processed_transcript	0+
chr1	868071	876903	ENSG00000230368.2_FAM41C_lincRNA	0-
chr1	873292	874349	ENSG00000234711.1_TUBB8P11_unprocessed_pseudogene	0+
chr1	904834	915976	ENSG00000272438.1_RP11-54O7.16_lincRNA	0+
chr1	911435	914948	ENSG00000230699.2_RP11-54O7.1_lincRNA	0+
chr1	914171	914971	ENSG00000241180.1_RP11-54O7.2_lincRNA	0+
chr1	916865	921016	ENSG00000223764.2_RP11-54O7.3_lincRNA	0-
chr1	924880	944581	ENSG00000187634.7_SAMD11_protein_coding	0+



## Need for gene/transcript annotation

Two main annotation files are found:

**GFF** Generic Feature Format Version 3 (GFF3)

- <http://www.sequenceontology.org/gff3.shtml>

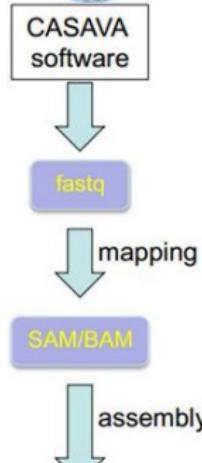
**GTF** Gene Transfer Format (GTF2)

- <http://mblab.wustl.edu/GTF22.html>

Both formats are nine-column, tab-delimited, plain text files

```
<seqname> <source> <feature> <start> <end> <score> <strand> <frame> [attributes] [comments]
```





# File formats – GTF

- GTF format
  - **Gene Transfer Format**
  - Widely used format for annotated genome and transcriptome
  - Downloadable from major browser sites, e.g. UCSC, Ensembl, NCBI
  - Illumina also provides a set of annotated genomes: igenomes
    - Available through Galaxy and command line

Seqname	Source	feature	start	end	score	strand	frame	attributes
chr1	unknown	exon	3204563	3207049	.	-	.	gene_id "Xkr4"; transcript_id "NM_001011874";



# NGS | 数据格式 | GTF

Seqname	Source	Feature	Start	End	Score	Strand	Frame	Attributes
chr4	protein_coding	CDS	24053	24477	.	+	0	exon_number "1"; gene_id "FBgn0040037"; gene_name "JYalpha"; p.
chr4	protein_coding	exon	24053	24477	.	+	.	exon_number "1"; gene_id "FBgn0040037"; gene_name "JYalpha"; p.
chr4	protein_coding	CDS	24979	25153	.	+	1	exon_number "2"; gene_id "FBgn0040037"; gene_name "JYalpha"; p.
chr4	protein_coding	exon	24979	25153	.	+	.	exon_number "2"; gene_id "FBgn0040037"; gene_name "JYalpha"; p.
chr4	protein_coding	CDS	25218	25450	.	+	0	exon_number "3"; gene_id "FBgn0040037"; gene_name "JYalpha"; p.
chr4	protein_coding	exon	25218	25450	.	+	.	exon_number "3"; gene_id "FBgn0040037"; gene_name "JYalpha"; p.
chr4	protein_coding	CDS	25501	25618	.	+	1	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p.
chr4	protein_coding	exon	25501	25621	.	+	.	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p.
chr4	protein_coding	stop_codon	25619	25621	.	+	0	exon_number "4"; gene_id "FBgn0040037"; gene_name "JYalpha"; p.
chr4	pseudogene	exon	26994	27101	.	-	.	exon_number "7"; gene_id "FBgn0052011"; gene_name "CR32011";
chr4	pseudogene	exon	27167	27349	.	-	.	exon_number "6"; gene_id "FBgn0052011"; gene_name "CR32011";
chr4	pseudogene	exon	28371	28609	.	-	.	exon_number "5"; gene_id "FBgn0052011"; gene_name "CR32011";



## GFF: a standard annotation format

- Stands for:
  - Gene Finding Format -or- General Feature Format
- Designed as a single line record for describing features on DNA sequence -- originally used for gene prediction output
- 9 tab-delimited fields common to all versions
  - seq source feature begin end score strand frame group
- The group field differs between versions, but in every case no tabs are allowed
  - GFF2: group is a unique description, usually the gene name.
    - NCOA1
  - GFF2.5 / GTF (Gene Transfer Format): tag-value pairs introduced, start\_codon and stop\_codon are required features for CDS
    - transcript\_id "NM\_056789" ; gene\_id "NCOA1"
  - GFF3: Capitalized tags follow Sequence Ontology (SO) relationships, FASTA seqs can be embedded
    - ID=NM\_056789\_exon1; Parent=NM\_056789; note="5' UTR exon"



# NGS | 数据格式 | GFF

```
ctg123 example gene          1050 9000 . + . ID=EDEN;Name=EDEN;Note=protein kinase
ctg123 example mRNA          1050 9000 . + . ID=EDEN.1;Parent=EDEN;Name=EDEN.1;Index=1
ctg123 example five_prime_UTR 1050 1200 . + . Parent=EDEN.1
ctg123 example CDS           1201 1500 . + 0 Parent=EDEN.1
ctg123 example CDS           3000 3902 . + 0 Parent=EDEN.1
ctg123 example CDS           5000 5500 . + 0 Parent=EDEN.1
ctg123 example CDS           7000 7608 . + 0 Parent=EDEN.1
ctg123 example three_prime_UTR 7609 9000 . + . Parent=EDEN.1

ctg123 example mRNA          1050 9000 . + . ID=EDEN.2;Parent=EDEN;Name=EDEN.2;Index=1
ctg123 example five_prime_UTR 1050 1200 . + . Parent=EDEN.2
ctg123 example CDS           1201 1500 . + 0 Parent=EDEN.2
ctg123 example CDS           5000 5500 . + 0 Parent=EDEN.2
ctg123 example CDS           7000 7608 . + 0 Parent=EDEN.2
ctg123 example three_prime_UTR 7609 9000 . + . Parent=EDEN.2

ctg123 example mRNA          1300 9000 . + . ID=EDEN.3;Parent=EDEN;Name=EDEN.3;Index=1
ctg123 example five_prime_UTR 1300 1500 . + . Parent=EDEN.3
ctg123 example five_prime_UTR 3000 3300 . + . Parent=EDEN.3
ctg123 example CDS           3301 3902 . + 0 Parent=EDEN.3
ctg123 example CDS           5000 5500 . + 1 Parent=EDEN.3
ctg123 example CDS           7000 7600 . + 1 Parent=EDEN.3
ctg123 example three_prime_UTR 7601 9000 . + . Parent=EDEN.3
```



## Feature formats: **GFF3 vs. GTF**

### ❖ **GFF3 – Gene feature format**

```
Chr1 amel_OGSv3.1 gene 204921 223005 . + . ID=GB42165
Chr1 amel_OGSv3.1 mRNA 204921 223005 . + . ID=GB42165-RA;Parent=GB42165
Chr1 amel_OGSv3.1 3'UTR 222859 223005 . + . Parent=GB42165-RA
Chr1 amel_OGSv3.1 exon 204921 205070 . + . Parent=GB42165-RA
Chr1 amel_OGSv3.1 exon 222772 223005 . + . Parent=GB42165-RA
```

### ❖ **GTF – Gene transfer format**

```
AB000381 Twinscan CDS 380 401 . + 0 gene_id "001"; transcript_id "001.1";
AB000381 Twinscan CDS 501 650 . + 2 gene_id "001"; transcript_id "001.1";
AB000381 Twinscan CDS 700 707 . + 2 gene_id "001"; transcript_id "001.1";
AB000381 Twinscan start_codon 380 382 . + 0 gene_id "001"; transcript_id "001.1";
AB000381 Twinscan stop_codon 708 710 . + 0 gene_id "001"; transcript_id "001.1";
```

*Always check which of the two formats is accepted by your application of choice, sometimes they cannot be swapped*



# NGS | 数据格式 | GTF vs. GFF

```
##GTF format
381 Twinscan exon    150    200    .    +    .    gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan exon    300    401    .    +    .    gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan CDS    380    401    .    +    0    gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan exon    501    650    .    +    .    gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan CDS    501    650    .    +    2    gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan exon    700    800    .    +    .    gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan CDS    700    707    .    +    2    gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan exon    900    1000   .    +    .    gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan start_codon 380    382    .    +    0    gene_id "381.000"; transcript_id "381.000.1";
381 Twinscan stop_codon 708    710    .    +    0    gene_id "381.000"; transcript_id "381.000.1";

##gff-version 3
##sequence-region ctg123 1 1497228
ctg123 Prokka gene      1000 9000  .  +  .  ID=gene00001;Name=EDEN
ctg123 Prokka TF_binding_site 1000 1012  .  +  .  ID=tfbs00001;Parent=gene00001
ctg123 Prokka mRNA      1050 9000  .  +  .  ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 Prokka mRNA      1050 9000  .  +  .  ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 Prokka mRNA      1300 9000  .  +  .  ID=mRNA00003;Parent=gene00001;Name=EDEN.3
ctg123 Prokka exon     1300 1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 Prokka exon     1050 1500  .  +  .  ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 Prokka exon     3000 3902  .  +  .  ID=exon00003;Parent=mRNA00001,mRNA00003
ctg123 Prokka exon     5000 5500  .  +  .  ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 Prokka exon     7000 9000  .  +  .  ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 Prokka CDS     1201 1500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 Prokka CDS     3000 3902  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 Prokka CDS     5000 5500  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 Prokka CDS     7000 7600  .  +  0  ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 Prokka CDS     1201 1500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 Prokka CDS     5000 5500  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 Prokka CDS     7000 7600  .  +  0  ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 Prokka CDS     3301 3902  .  +  0  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 Prokka CDS     5000 5500  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 Prokka CDS     7000 7600  .  +  1  ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
```



BED: zero based, start inclusive, stop exclusive

chr1	10491	10492	rs55998931	0	+
chr1	10582	10583	rs58108140	0	+

- ⇒ First base on the chromosome is 0
- ⇒ Length = stop - start

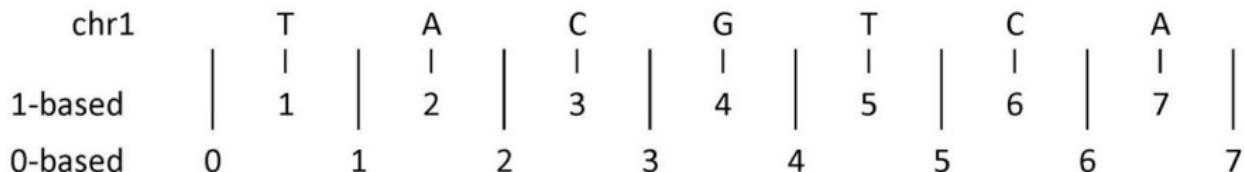
GTF/GFF: one based, inclusive

chr1	snp135Com	exon	10492	10492	0.000
chr1	snp135Com	exon	10583	10583	0.000

- ⇒ First base on the chromosome is 1
- ⇒ Length = stop – start+1



# NGS | 数据格式 | BED vs. GFF



	1-based	0-based
Indicate a single nucleotide	chr1:4-4 G	chr1:3-4 G
Indicate a range of nucleotides	chr1:2-4 ACG	chr1:1-4 ACG
Indicate a single nucleotide variant	chr1:5-5 T/A	chr1:4-5 T/A

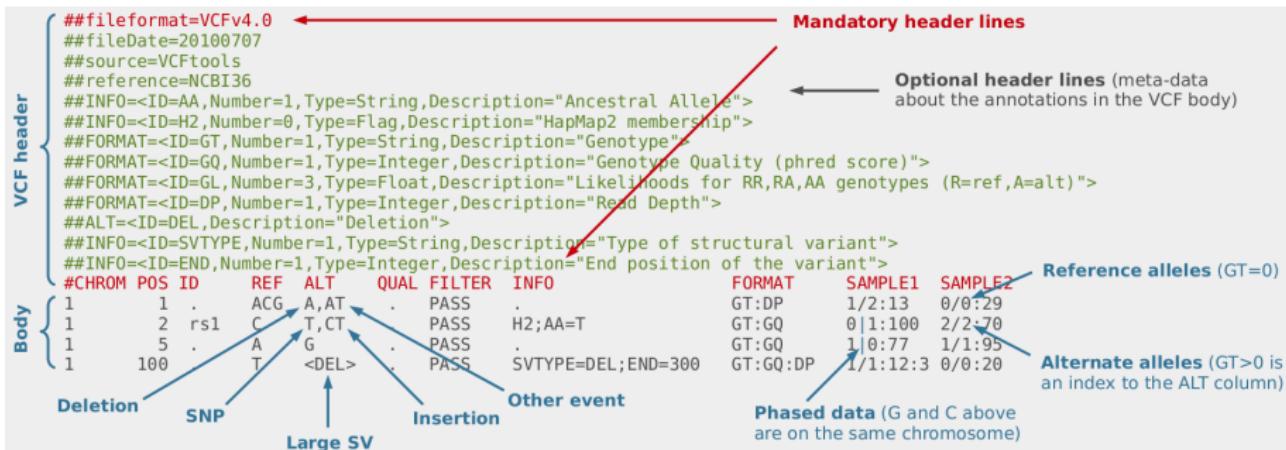


## VCF files

- There is a file format defined for genetic variants called VCF (Variant Call Format).
  - Specification available at  
<http://www.1000genomes.org/wiki/Analysis/Variant%20Call%20Format/vcf-variant-call-format-version-41>
  - Two main sections: header and content
  - Header provides basic information of the file, and defines content attributes and filters
  - Each line in the content section represents one variant in one or more samples



# NGS | 数据格式 | VCF



# 教学提纲

1

## 系统生物学

2

## 基因组学

3

## 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

## 数据库与数据格式

5

## 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

## 外显子组测序

## ● 简介

## ● 操作流程

## ● 应用实例

## 7 转录组学

## 8 RNA-Seq

## ● 概述

## ● 数据分析

## ● 应用实例

## 9 顺反组

## ● 概述

## ● ChIP-Seq

## 10 表观遗传学

## ● 概述

## ● Methyl-Seq



# 教学提纲

1

## 系统生物学

2

## 基因组学

3

## 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

## 数据库与数据格式

5

## 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

## 外显子组测序

## ● 简介

## ● 操作流程

## ● 应用实例

## 7 转录组学

## 8 RNA-Seq

## ● 概述

## ● 数据分析

## ● 应用实例

## 9 顺反组

## ● 概述

## ● ChIP-Seq

## 10 表观遗传学

## ● 概述

## ● Methyl-Seq



## 深度 (depth)

- 也叫乘数，衡量测序量的首要参数；测序得到的总碱基数与待测基因组大小的比值；每个碱基被测序的平均次数
- 假设一个基因大小为  $2M$ ，测序获得的总数据量为  $20M$ ，那么深度为  $10X$

## 覆盖度 (coverage)

- 测序获得的序列占整个捕获区域/基因组的比例
- 由于基因组中的高 GC、重复序列等复杂结构的存在，测序最终拼接组装获得的序列往往无法覆盖所有的区域，这部分没有获得的区域就称为 Gap。例如一个细菌基因组测序，覆盖度是 98%，那么还有 2% 的序列区域是没有通过测序获得的。

## 深度 (depth)

- 也叫乘数，衡量测序量的首要参数；测序得到的总碱基数与待测基因组大小的比值；每个碱基被测序的平均次数
- 假设一个基因大小为  $2M$ ，测序获得的总数据量为  $20M$ ，那么深度为  $10X$

## 覆盖度 (coverage)

- 测序获得的序列占整个捕获区域/基因组的比例
- 由于基因组中的高 GC、重复序列等复杂结构的存在，测序最终拼接组装获得的序列往往无法覆盖所有的区域，这部分没有获得的区域就称为 Gap。例如一个细菌基因组测序，覆盖度是 98%，那么还有 2% 的序列区域是没有通过测序获得的。

## 实验

对长 100bp 的目标区域进行捕获测序：采用单端测序，每个 read 长 5bp；总共得到了 200 个 reads；把所有的 reads 比对到目标区域后，100bp 的目标区域中有 98bp 的位置至少有 1 个 read 覆盖到，换言之，剩余的 2bp 没有任何 reads 覆盖。

## 深度与覆盖度

- 深度： $200 \times 5 / 100 = 10$
- 覆盖度： $98 / 100 \times 100\% = 98\%$



## 实验

对长 100bp 的目标区域进行捕获测序：采用单端测序，每个 read 长 5bp；总共得到了 200 个 reads；把所有的 reads 比对到目标区域后，100bp 的目标区域中有 98bp 的位置至少有 1 个 read 覆盖到，换言之，剩余的 2bp 没有任何 reads 覆盖。

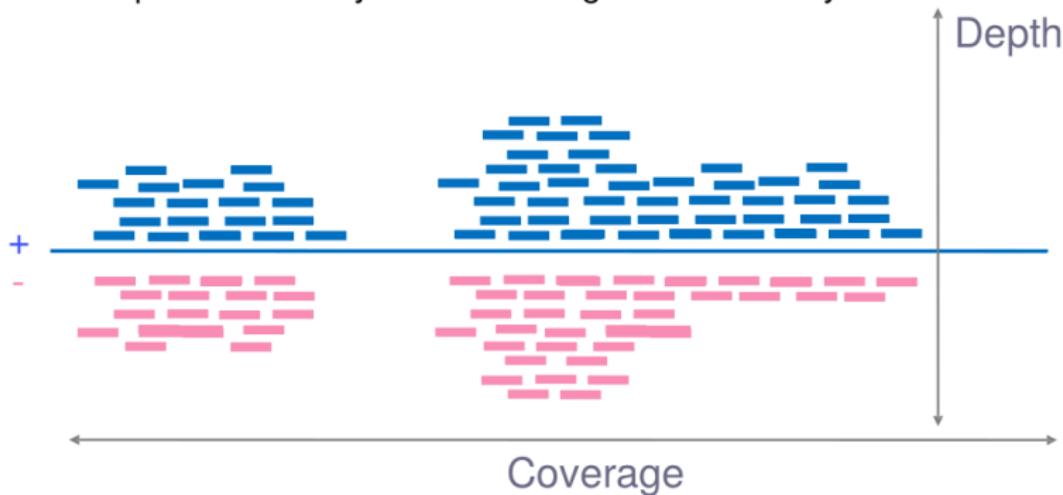
## 深度与覆盖度

- 深度： $200 \times 5 / 100 = 10$
- 覆盖度： $98 / 100 \times 100\% = 98\%$



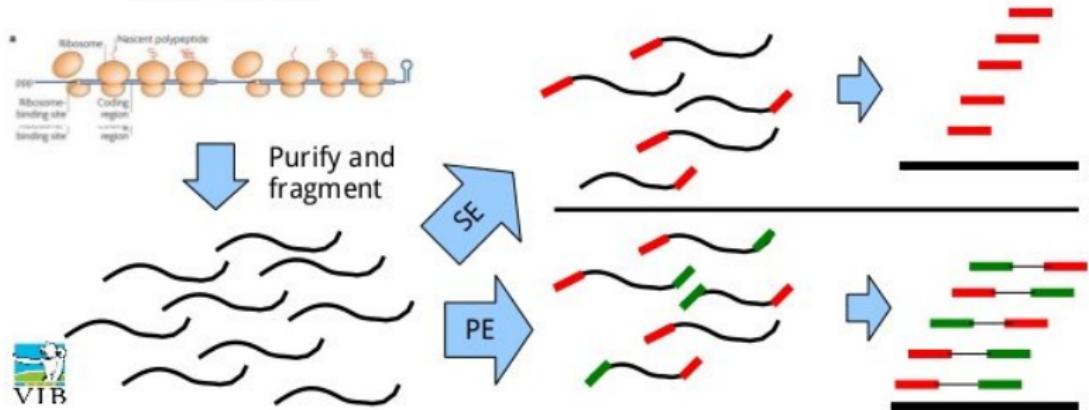
## Statistics used as Quality Control

- **Depth of coverage** = mean number of reads covering a base (X)  
Example: 30X for normal sample, 100X for tumor sample
- **Coverage** = part of the reference with at least one read  
Example:  $\geq 80\%$  of your exome target is covered by 20X



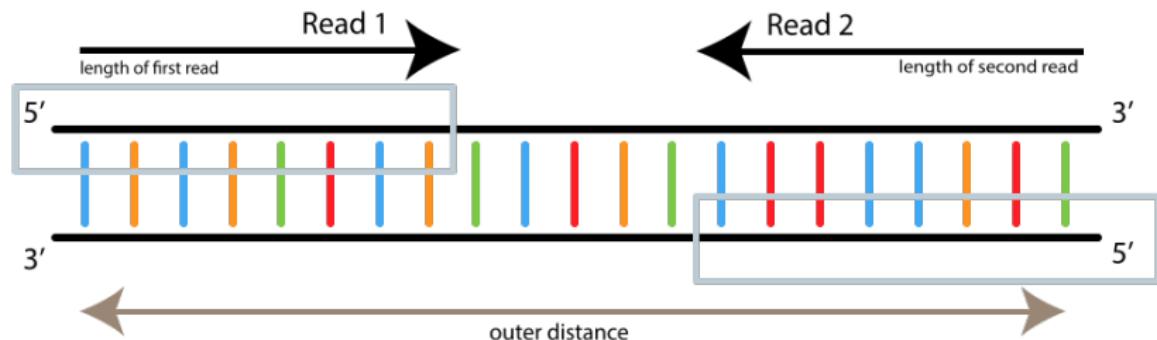
## PE versus SE Illumina

- **Single end (SE):** from each cDNA fragment only one end is read.
- **Paired end (PE):** the cDNA fragment is read from both ends.



## Paired-End Sequencing

- allows users to sequence both ends of a fragment and generate high-quality, alignable sequence data
- facilitates detection of genomic rearrangements and repetitive sequence elements, as well as gene fusions and novel transcripts



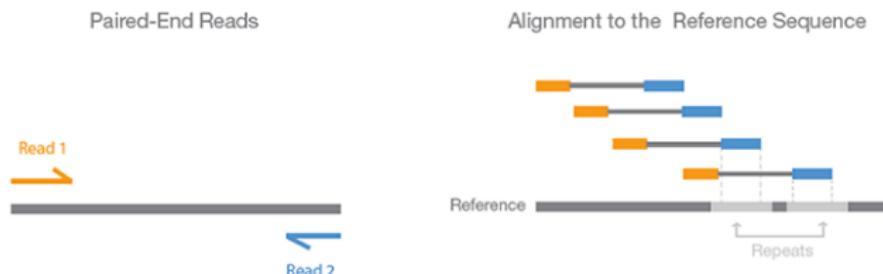
Huge impact on read mapping

**Pairs give two locations to determine whether read is unique**

Critical for estimating transcript-level abundance

**Increases number of splice junction spanning reads**

Figure 4. Paired-End Sequencing and Alignment



Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across difficult-to-sequence, repetitive regions of the genome.



fragment

---

fragment + adaptors

SE read

->

PE reads

R1 - - - - - >

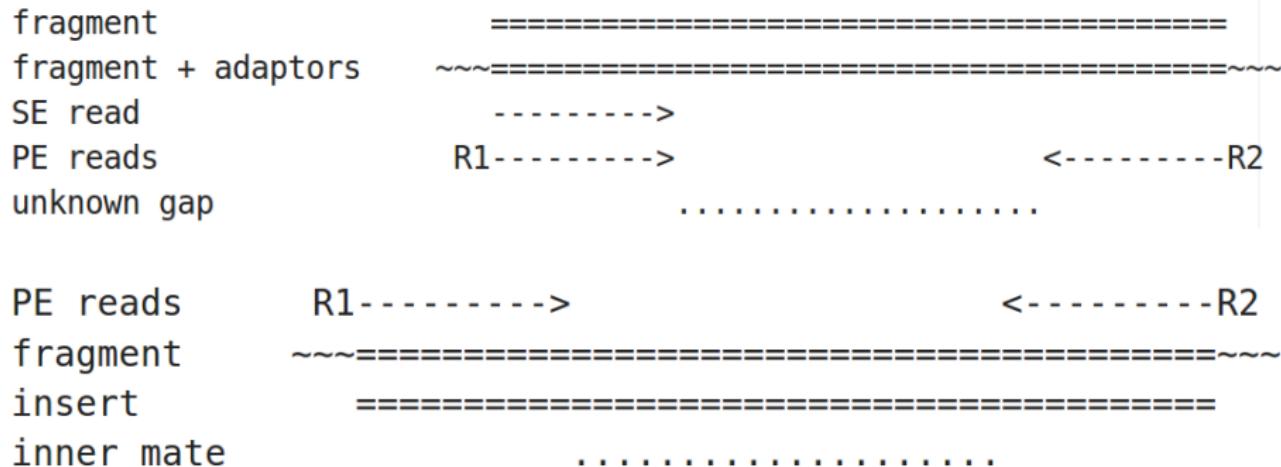
$\leftarrow \dots \text{R2}$

## unknown gap

.....

fragment	=====
fragment + adaptors	~~~~=====~~~~~
SE read	- - - - - >
PE reads	R1 - - - - - > < - - - - R2
unknown gap	.....
PE reads	R1 - - - - - > < - - - - R2
fragment	~~~~=====~~~~~
insert	=====
inner mate	.....

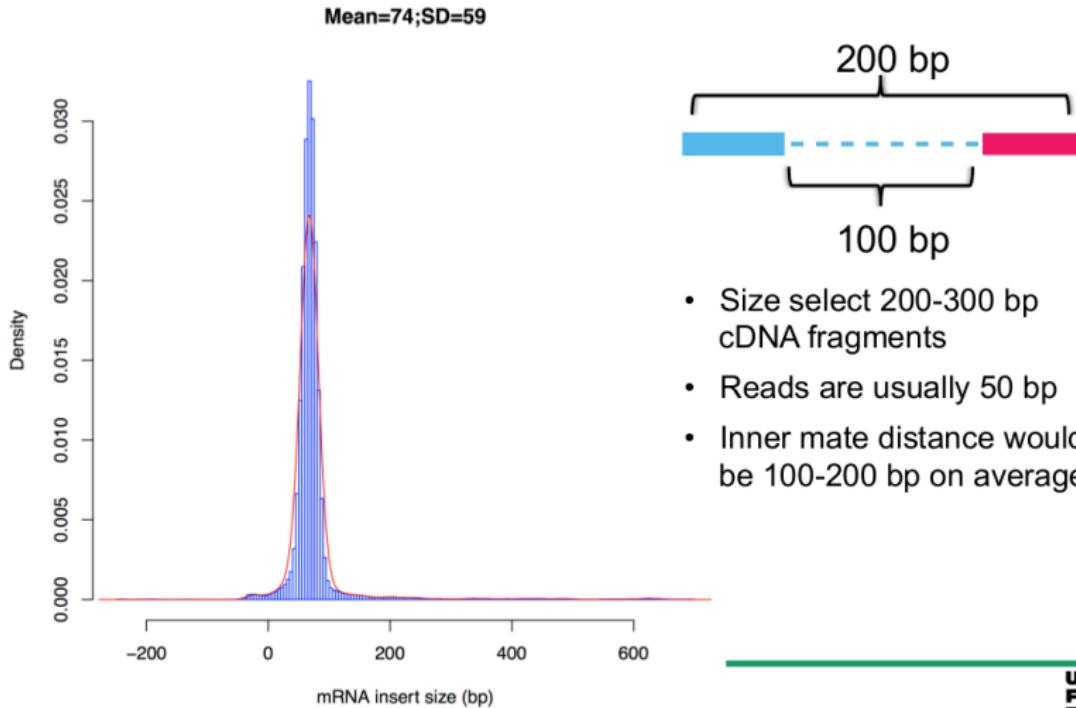




## Conclusion

Remember that “**insert**” refers to the DNA fragment between the adaptors, and not the gap between R1 and R2. Instead we refer to that as the “**inner mate distance**”.

## Distribution of cDNA fragment sizes



	定义	说明
测序质量	测序过程碱基识别过程中，对所识别的碱基给出的错误概率	比如质量值是 Q30，则错误识别的概率是 1/1000，碱基正确识别率是 99.9%
平均读长	测序时所有读段的平均长度	读长越长则单条读段覆盖的碱基数就比较多，也就越容易比对到基因组上
覆盖率	基因组上被测到的碱基数占总碱基的比例	覆盖率越高越好，这样可以保证测序结果判定完整性
基因组测序深度	测序得到的总碱基数与基因组大小的比值	测序深度越大，则对单个碱基判断的基底统计个数越多
测序通量	单次上机测序反应所产生的数据量	测序通量越高，产生的数据量越大
测序时间	单次上机测序反应所使用时间	测序时间越短则数据产生的速度越高，测序仪的使用效率也越高



# 教学提纲

1

## 系统生物学

2

## 基因组学

3

## 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

## 数据库与数据格式

5

## 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

## 外显子组测序

## ● 简介

## ● 操作流程

## ● 应用实例

## 7 转录组学

## 8 RNA-Seq

## ● 概述

## ● 数据分析

## ● 应用实例

## 9 顺反组

## ● 概述

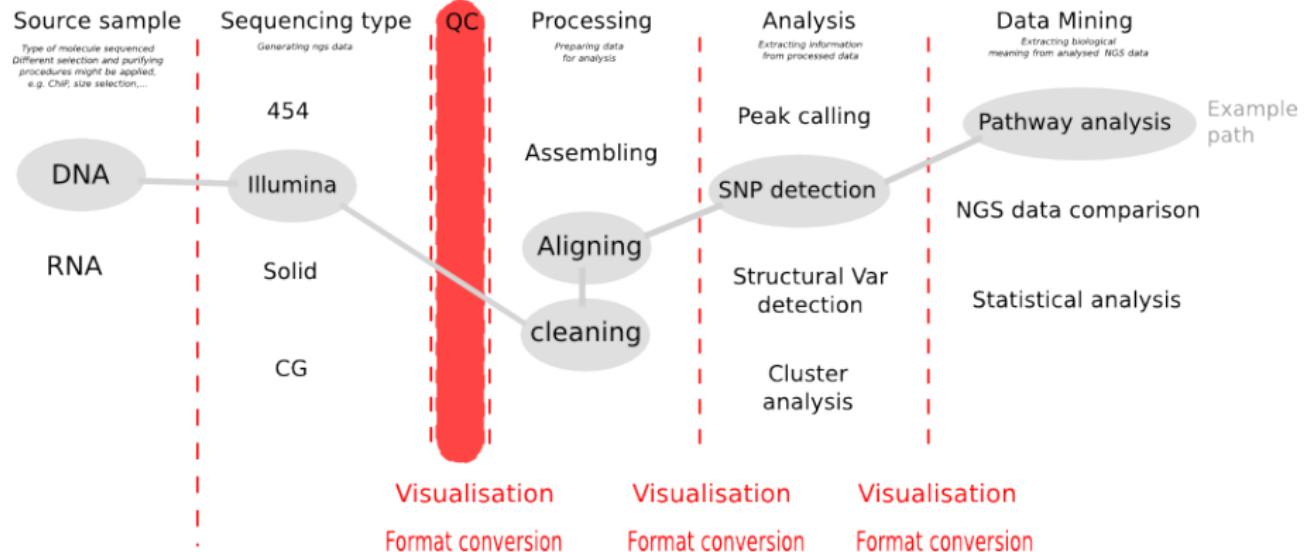
## ● ChIP-Seq

## 10 表观遗传学

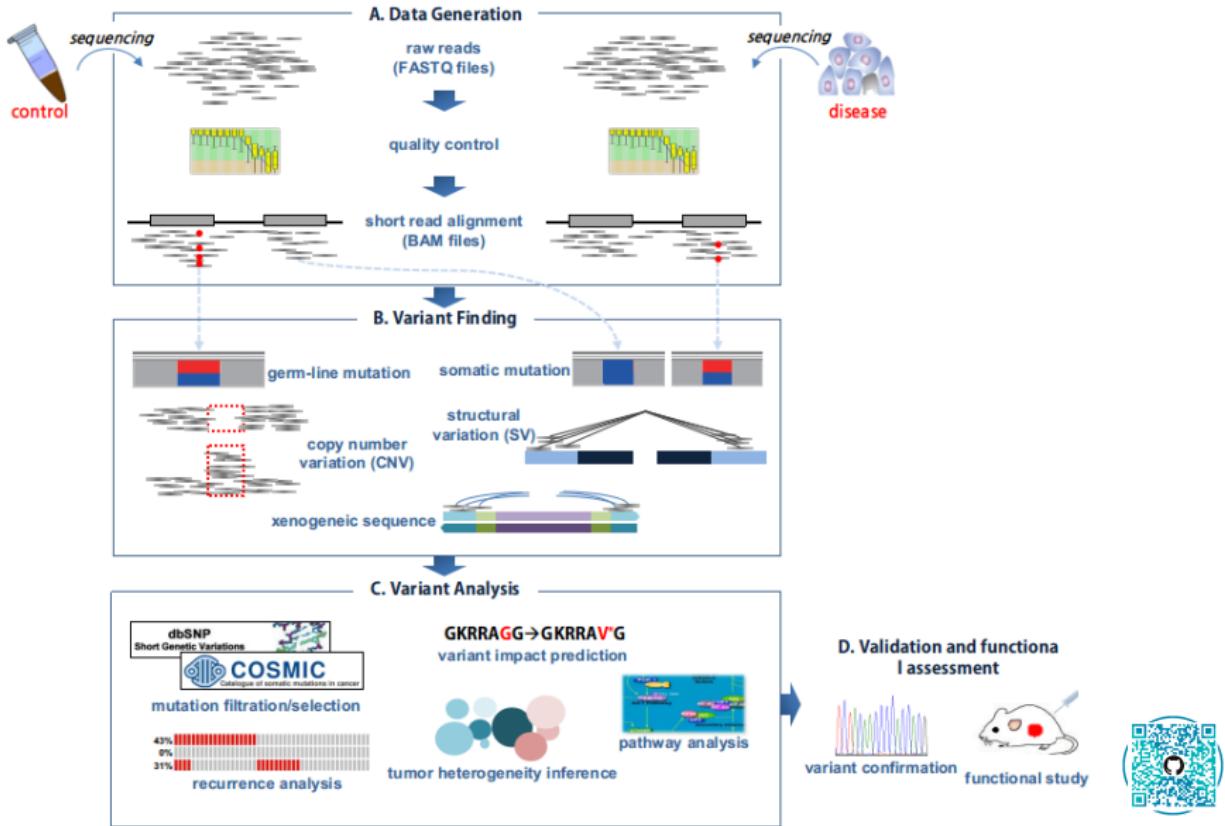
## ● 概述

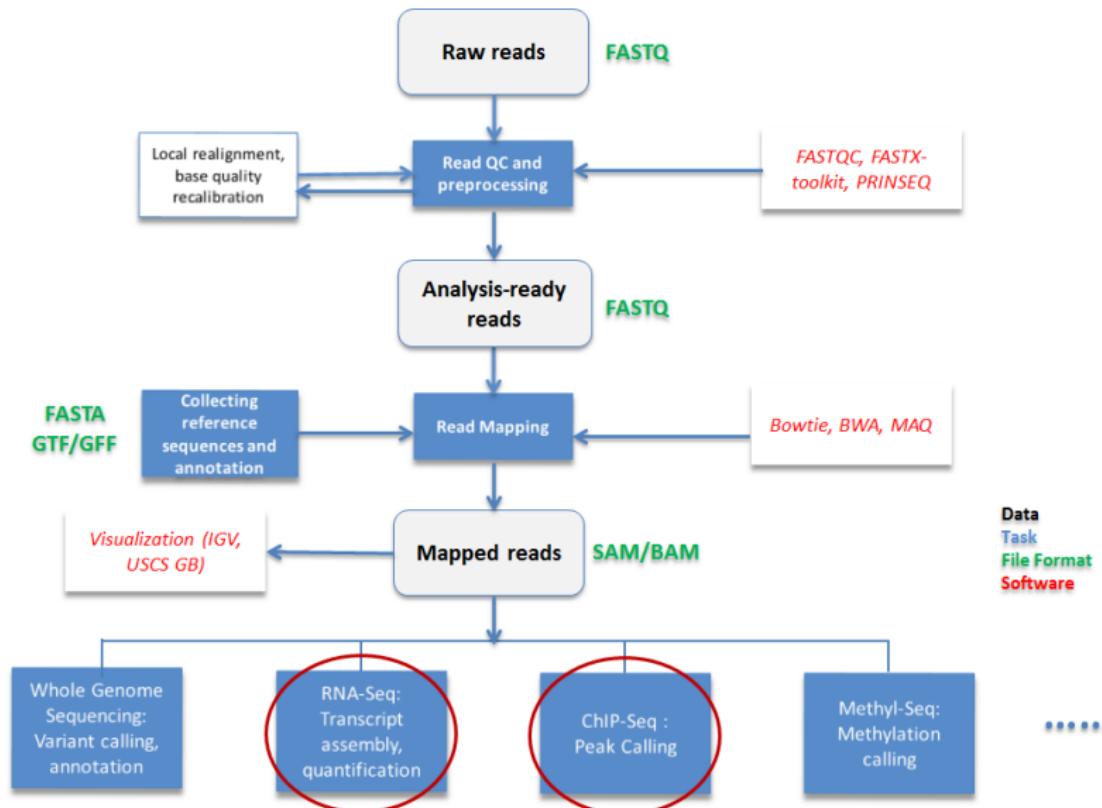
## ● Methyl-Seq

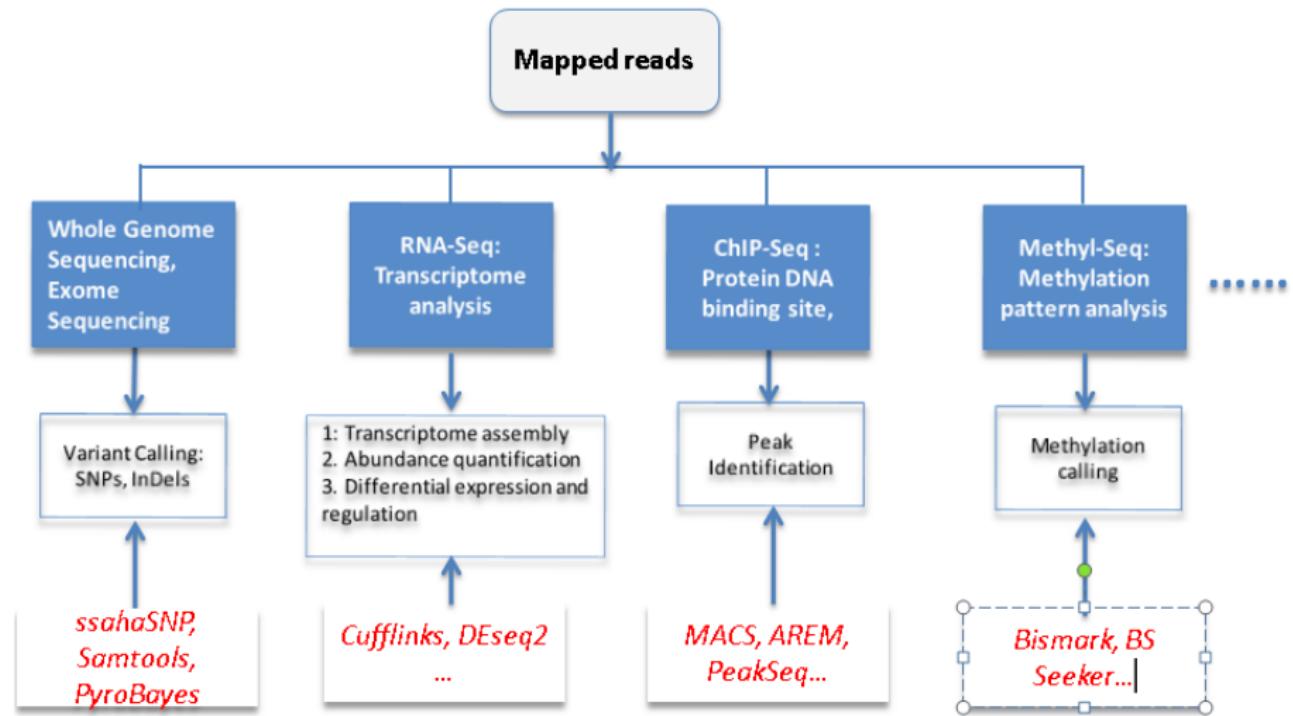




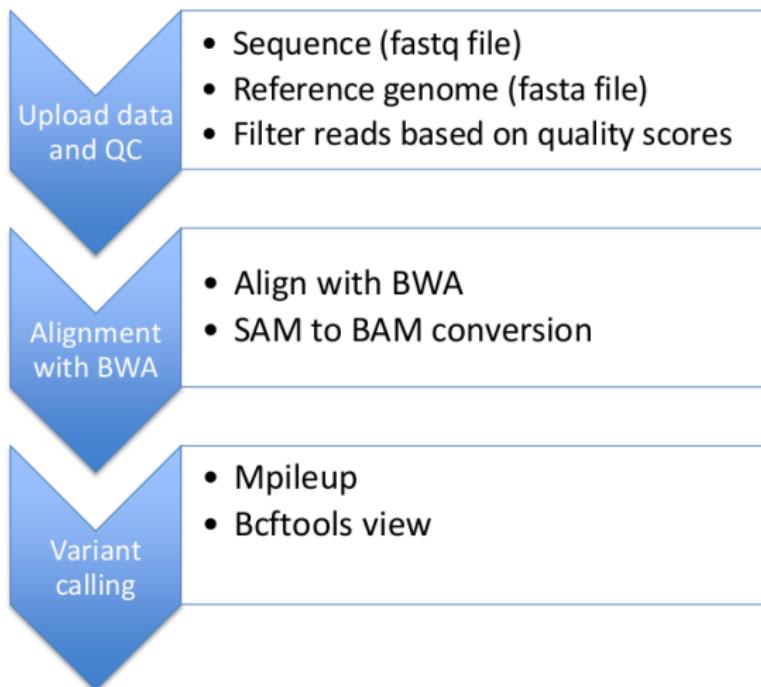
# NGS | 数据分析 | 流程 | 概述 | 总览

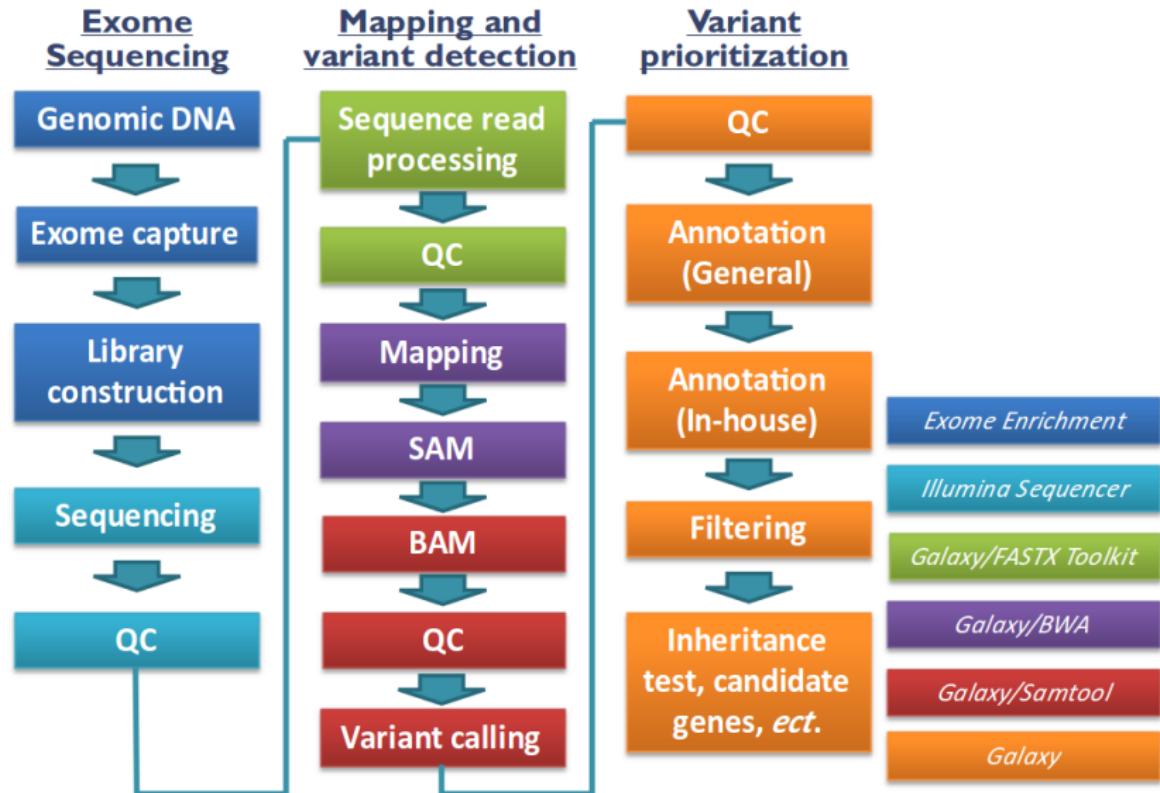






## SNP-Seq pipeline





## Quality Control of the data



First step after receiving the data

Sometimes partially done by the sequencing center (e.g., chastity)

Objective:

- Remove bad quality reads
- Remove contaminants
- Trim ends of reads
- Remove orphans (if possible or desirable)
- Correct errors

FastQC (<http://www.bioinformatics.bbsrc.ac.uk/projects/fastqc/>)



# QC Report

## ➤ Sequence Statistics

Total No. of Sequences	6970943
Avg. Sequence Length	54
Max Sequence Length	54
Min Sequence Length	54
Total Sequence Length	376430922
Total N bases	14254521
% N bases	3.78676
No of Sequences with Ns	278635
% Sequences with Ns	3.99709

## ➤ Quality Statistics

Total HQ bases	334195496
%HQ bases	88.78
Total HQ reads	6350256
%HQ reads	91.0961

# Alignment statistics

Total Reads	15849154
Reads aligned	7746088
% Reads Aligned	48.8738
Total Genome Size	64022747
Genome Covered	28234853
%Coverage	44.1013
Avg Read Depth	1.50491
% Coverage at 1X	44.1013
% Coverage at 5X	10.7884
% Coverage at 10X	1.76412
% Coverage at 15X	0.297722
% Coverage at 20X	0.122413
% Coverage at 30X	0.0557255
% Coverage at 40X	0.0372789



# NGS | 数据分析 | 流程 | 质控 | 工具

Feature\Tools	NGS QC Toolkit v2.2	FastQC v0.10.0	PRINSEQ-lite v0.17 <sup>1</sup>	TagDust	FASTX-Toolkit v0.0.13	SolexaQA v1.10	TagCleaner v0.12 <sup>1</sup>	CANGS v1.1
Supported NGS platforms	Illumina, 454	FASTQ <sup>2</sup>	Illumina, 454	Illumina, 454	Illumina	Illumina	Illumina, 454	454
Parallelization	Yes	Yes	No	No	No	No	No	No
Detection of FASTQ variants	Yes	Yes	Yes	No	No	Yes	No	No
Primer/Adaptor removal	Yes	No <sup>3</sup>	No	Yes	Yes	No	Yes <sup>4</sup>	Yes
Homopolymer trimming (Roche 454 data)	Yes	No	No	No	No	No	No	Yes
Paired-end data integrity	Yes	No	No	No	No	No	No	No
QC of 454 paired-end reads	Yes	No	No	No	No	No	No	No
Sequence duplication filtering	No	No <sup>5</sup>	Yes	No	Yes	No	No	Yes
Low complexity filtering	No	No	Yes	No	Yes	No	No	No
N/X content filtering	No	No <sup>6</sup>	Yes	No	Yes	No	No	Yes
Compatibility with compressed input data file	Yes	Yes	No	No	No	No	No	No
GC content calculation	Yes	Yes	Yes	No	No	No	No	No
File format conversion	Yes	No	No	No	No	No	No	No
Export HQ and/or filtered reads	Yes	No	Yes	Yes	Yes	No	Yes	Yes
Graphical output of QC statistics	Yes	Yes	No <sup>7</sup>	No	Yes	Yes	No <sup>7</sup>	No
Dependencies	Perl modules: Parallel::ForkManager, String::Approx, GD::Graph (optional)	-	-	-	Perl module: GD::Graph	R, matrix2png -		BLAST, NCBI nr database



## FastQC

A quality control tool for high throughput sequence data.

## NGS QC Toolkit

A toolkit for the quality control (QC) of next generation sequencing (NGS) data.

## AfterQC

Automatic Filtering, Trimming, Error Removing and Quality Control for fastq data.

## Others

- SolexaQA: calculates sequence quality statistics and creates visual representations of data quality for second-generation sequencing data.

...  
...

## FastQC

A quality control tool for high throughput sequence data.

## NGS QC Toolkit

A toolkit for the quality control (QC) of next generation sequencing (NGS) data.

## AfterQC

Automatic Filtering, Trimming, Error Removing and Quality Control for fastq data.

## Others

- SolexaQA: calculates sequence quality statistics and creates visual representations of data quality for second-generation sequencing data.

...  
...

## FastQC

A quality control tool for high throughput sequence data.

## NGS QC Toolkit

A toolkit for the quality control (QC) of next generation sequencing (NGS) data.

## AfterQC

Automatic Filtering, Trimming, Error Removing and Quality Control for fastq data.

## Others

- SolexaQA: calculates sequence quality statistics and creates visual representations of data quality for second-generation sequencing data.

...  
...

## FastQC

A quality control tool for high throughput sequence data.

## NGS QC Toolkit

A toolkit for the quality control (QC) of next generation sequencing (NGS) data.

## AfterQC

Automatic Filtering, Trimming, Error Removing and Quality Control for fastq data.

## Others

- SolexaQA: calculates sequence quality statistics and creates visual representations of data quality for second-generation sequencing data.
- ...

## FASTQC: Report

- 1) Basic statistics
- 2) Per base sequence quality
- 3) Per tile sequence quality
- 4) Per sequence quality scores
- 5) Per base sequence content
- 6) Per sequence GC content
- 7) Per base N content
- 8) Sequence Length Distribution
- 9) Sequence duplication levels
- 10) Over-represented sequences
- 11) Adapter/Kmer content

### Basic Statistics

Measure	Value
Filename	sample.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	9053
Sequences flagged as poor quality	0
Sequence length	36
%GC	50



## 目的

manipulating the sequences to produce better mapping results

## 内容

- Collapser: Collapsing identical sequences into a single sequence
- Clipper: Removing sequencing adapters/linkers
- Splitter: Splitting barcode containing multiple samples
- Filter: Filters sequences based on quality
- Trimmer: Trims (cuts) sequences based on quality
- Formatter: Rename identifiers, Reverse-complement, Mask nucleotides, Convert RNA ↔ DNA, ...
- ...

## 目的

manipulating the sequences to produce better mapping results

## 内容

- Collapser: Collapsing identical sequences into a single sequence
- Clipper: Removing sequencing adapters/linkers
- Splitter: Splitting barcode containing multiple samples
- Filter: Filters sequences based on quality
- Trimmer: Trims (cuts) sequences based on quality
- Formatter: Rename identifiers, Reverse-complement, Mask nucleotides, Convert RNA ↔ DNA, ...
- ...

## Read trimming or filtering

**Trimming** remove 5' and/or 3' ends of reads (bad quality or adapter)

**Filtering** remove full reads (e.g., contaminants)

Tools:

**FastX toolkit** ([http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/))

**PrinSeq** (<http://edwards.sdsu.edu/cgi-bin/prinseq/prinseq.cgi>)

**Sickle** (<https://github.com/najoshi/sickle>)

**ea-utils** (<https://code.google.com/p/ea-utils/>)

**cutadapt** (<https://cutadapt.readthedocs.org/>)

...



## FASTX-Toolkit

A collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.

## PRINSEQ

PReprocessing and INformation of SEQuence data. A publicly available tool that is able to filter, reformat and trim your sequences and to provide you summary statistics for your sequence data.

## cutadapt

Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads. It can also modify and filter reads in various ways.

## FASTX-Toolkit

A collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.

## PRINSEQ

PRereprocessing and INformation of SEQuence data. A publicly available tool that is able to filter, reformat and trim your sequences and to provide you summary statistics for your sequence data.

## cutadapt

Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads. It can also modify and filter reads in various ways.

## FASTX-Toolkit

A collection of command line tools for Short-Reads FASTA/FASTQ files preprocessing.

## PRINSEQ

PRereprocessing and INformation of SEQuence data. A publicly available tool that is able to filter, reformat and trim your sequences and to provide you summary statistics for your sequence data.

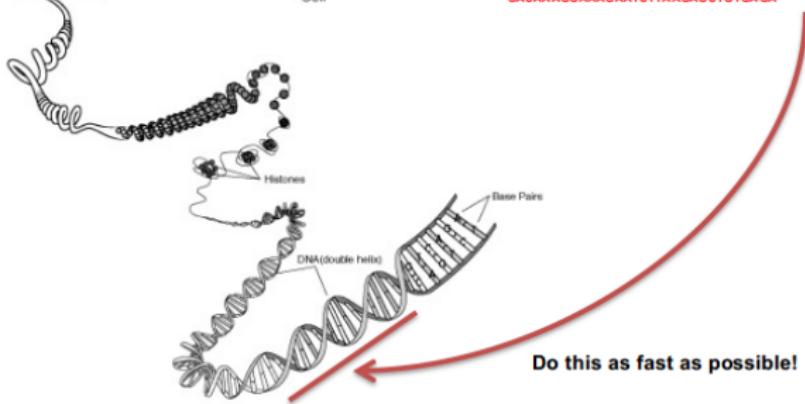
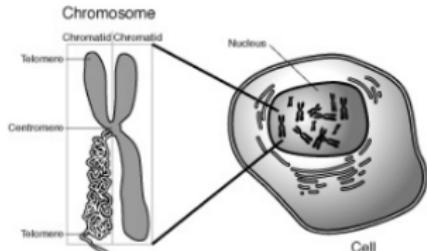
## cutadapt

Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads. It can also modify and filter reads in various ways.

- [Command Line Arguments](#)
  - [FASTQ-to-FASTA](#)
  - [FASTQ/A Quality Statistics](#)
  - [FASTQ Quality chart](#)
  - [FASTQ/A Nucleotide Distribution chart](#)
  - [FASTQ/A Clipper](#)
  - [FASTQ/A Renamer](#)
  - [FASTQ/A Trimmer](#)
  - [FASTQ/A Collapser](#)
  - [FASTQ/A Artifacts Filter](#)
  - [FASTQ Quality Filter](#)
  - [FASTQ/A Reverse Complement](#)
  - [FASTA Formatter](#)
  - [FASTA nucleotides changer](#)
  - [FASTA Clipping Histogram](#)
  - [FASTX Barcode Splitter](#)
- [Example: FASTQ Information](#)
- [Example: FASTQ/A manipulation](#)
- [Galaxy Usage](#)
- [FASTA/Q Information tools](#)
  - [Quality Statistics](#)
  - [Quality Boxplot](#)
  - [Nucleotide Distribution](#)
- [FASTA/Q Manipulation Tools](#)
  - [FASTA/Q Clipper](#)
  - [FASTA/Q Trimmer](#)
  - [FASTA/Q End Trimmer](#)
  - [FASTQ Quality Trimmer](#)
  - [FASTA/Q Renamer](#)
  - [FASTA/Q Collapser](#)
  - [FASTA UnCollapser](#)
  - [UnCollapse rows \(in a text file\)](#)
  - [Artifacts Filter](#)
  - [FASTQ Quality Filter](#)
  - [FASTQ/A Reverse Complement](#)
  - [FASTQ-to-FASTA converter](#)
  - [FASTA Formatter](#)
  - [FASTA nucleotide changer](#)
  - [FASTA Clipping Histogram](#)
  - [FASTQ/A barcode splitter](#)

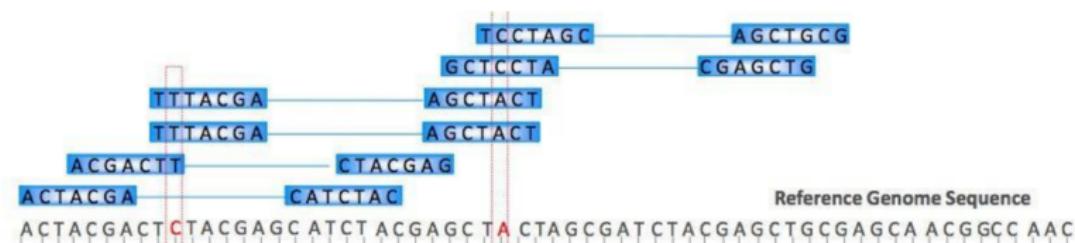


# Mapping back to genome

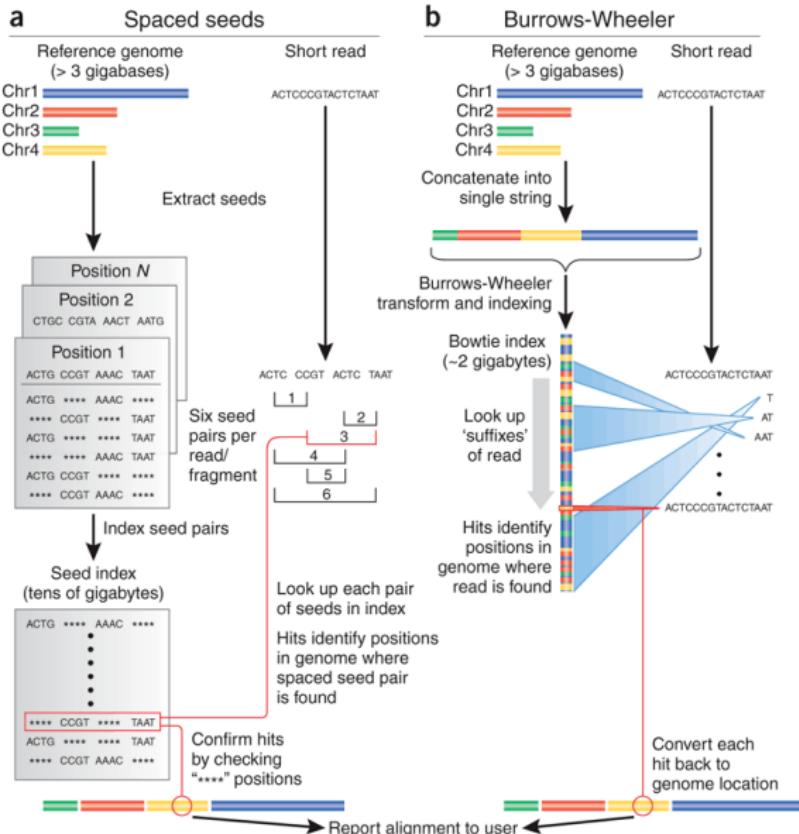


## Mapping on a reference Genome

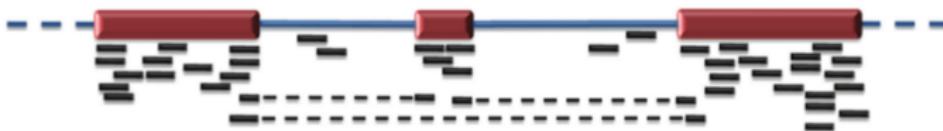
- Reads are aligned  $\geq 1$  times on the reference genome
- A **mapping quality** is associated to each alignment:
  - Quantify the probability that the alignment is correct
  - Decreases with the number of mismatches (wrong nucleotide) & gaps (small insertions/deletions) & the number of alignments



# NGS | 数据分析 | 流程 | 比对 | 算法



## Alignment Algorithms



**Bowtie2** (Langmead et al 2009) – BWT, multiseed heuristic alignment

**BWA** (Li and Durbin 2009) – BWT, Smith-Waterman alignment

**SOAP3** (Li et al. 2009) – BWT, proprietary alignment algorithm

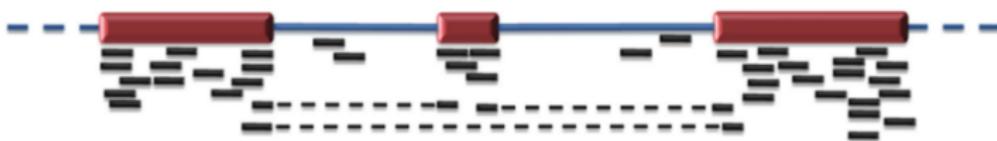
**BFAST** (Homer et al. 2009, based on BLAT) – multiple indexes, finds candidate alignment locations using seed and extend, followed by a gapped Smith-Waterman local alignment for each candidate

Many more!

[http://en.wikipedia.org/wiki/List\\_of\\_sequence\\_alignment\\_software](http://en.wikipedia.org/wiki/List_of_sequence_alignment_software)



## Alignment tools for splice junction mapping



Bowtie/Tophat/Cufflinks (Tuxedo suite)

MapSplice (good)

SpliceMap (less good)

GSNAP (good)

Star (very fast)

RUM (good)

HMMsplicer (slow)

HISAT (very fast)

...



## Mapping tools history

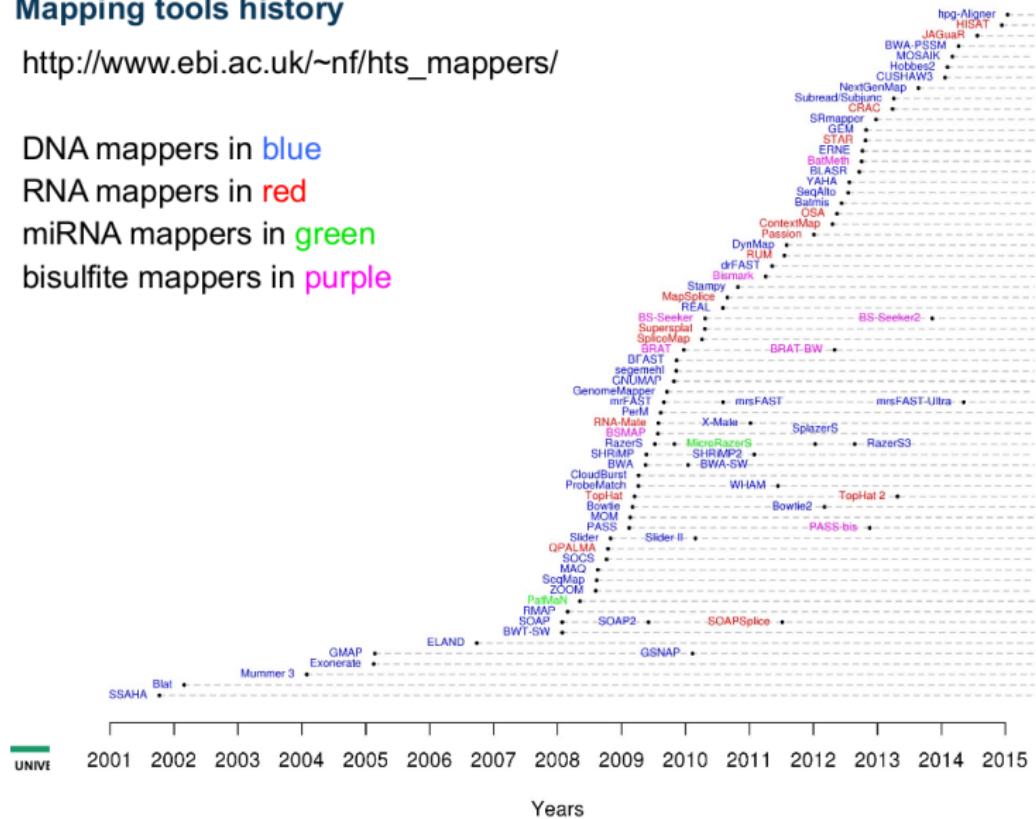
[http://www.ebi.ac.uk/~nf/hts\\_mappers/](http://www.ebi.ac.uk/~nf/hts_mappers/)

DNA mappers in blue

RNA mappers in red

miRNA mappers in green

bisulfite mappers in purple



## Alignment tools

- Multitude of alignment tools: BWA, Bowtie, Bowtie2, Bfast...
- How to choose the best tool ?
  - Is my sequencing technology supported ?
  - Do I have short or long reads ? Reads of different sizes ?
  - Do I want to allow gapped alignment ? Multiple alignments ?
  - Does it support single/paired-end reads ?
  - On which alignment algorithm is it based ?
  - Computational issues ? Is it used by the community ?
- A classical and performant tool for Illumina sequencing: BWA (Burrows-Wheeler Aligner)



## BWA

BWA(Burrows-Wheeler Aligner) is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome.

## Bowtie

- Bowtie is an ultrafast, memory-efficient short read aligner.
- Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.



## BWA

BWA(Burrows-Wheeler Aligner) is a software package for mapping low-divergent sequences against a large reference genome, such as the human genome.

## Bowtie

- Bowtie is an ultrafast, memory-efficient short read aligner.
- Bowtie 2 is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.



## SOAP

SOAP has been in evolution from a single alignment tool to a tool package that provides full solution to next generation sequencing data analysis.

- SOAPaligner/soap2: new alignment tool
- SOAPSnp: re-sequencing consensus sequence builder
- SOAPIndel: indel finder
- SOAPsv: structural variation scanner
- SOAPdenovo: *de novo* shot reads assembler
- SOAP3/GPU: GPU-accelerated alignment tool



# NGS | 数据分析 | 流程 | 提取变异

GMIAK1 :: ID : rs2909430, Reference C, Allele T, Position chr17:7519370

Reference	AACTGAAACAGATAAAGCACTTGGAAAGACGGCAGCAAAGAAAACCATG	TGAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA
GMI_1864715162(+)	TGGAAGACGGCAGCAAAGAAAACCATG	TGAAGC
GMI_813006088(+)	TGGAAGACGGCAGCAAAGAAAACCATG	TGAAGC
GMI_1079100245(-)	GAAGACGGCAGCAAAGAAAACCATG	TGAAGCAC
GMI_2159545344(+)	CAAAGAAAACCATG	GAAGCACCTCTGTACC
GMI_1776483420(-)	AAGAAACAAACATG	TGAAGCACCTCTGCACCCCA
GMI_2093120226(-)	AAGAAACAAACATG	TGAAGCACCTCTGCACCCCA
GMI_2204137276(-)	AAGAAACAAACATG	TGAAGCACCTCTGCACCCCA
GMI_594634658(-)	AAGAAACAAACATG	TGAAGCACCTCTGCACCCCA
GMI_878606194(-)	AAGAAACAAACATG	TGAAGCACCTCTGCACCCCA
GMI_1463224112(+)	AGAAACAAACATG	TGAAGCACCTCTGCACCCAC
GMI_1385601163(-)	ACATG	TGAAGCACCTCTGCACCCACTAGCGAGC
GMI_2247152549(+)	TG	TGAAGCACCTCTGCACCCACTAGCGAGCTAG
GMI_2552461258(+)	AAACAAACATG	TGAAGCACCTCTGCACCCACTAGCGAGCTAGGGAGAGTTGGCGTCA

GMIAK4 :: ID : rs2909430, Reference C, Allele C/T, Position chr17:7519370

Reference	AACTGAAACAGATAAAGCACTTGGAAAGACGGCAGCAAAGAAAACCATG	TGAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA
GMI_661419350(+)	AAGTGAACAGATAAAGCACTTGGAAAGACGGCAGCAAAGAAAACCATG	TGAAGCACCTCTGCACCCACCA
GMI_661419351(-)	AAGTGAACAGATAAAGCAACGGGAAGACGGCAGCAAAGAAAACCATG	TGAAGCACCTCTGCACCCACTAGCG
GMI_661419352(-)	AAGTGAACAGATAAAGCACTTGGAAAGACGGCAGCAAAGAAAACCATG	TGAAGC
GMI_661419354(-)	AAGTGAACAGAAAAGGCAACGGGAAGACGGCAGCAAAGAAAACCATG	TGAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCT
GMI_661419355(+)	AAGTGAACAGATAAAGCACTTGGAAAGACGGCAGCAAAGAAAACCATG	TGAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA
GMI_661419356(+)	AAGTGAACAGATAAAGCACTTGGAAAGACGGCAGCAAAGAAAACCATG	TGAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGGTGGCGCCA
GMI_661419357(-)	GAACAGATAAAGCACTTGGAAAGACGGCAGCAAAGAAAACCATG	TGAAGCACCTCTGCACCCACTAGCGAGC
GMI_661419358(+)	GATAAGCACTTGGAAAGACGGCAGCAAAGAAAACCATG	TGAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGGAGTGGCG
GMI_661419359(-)	AAGCACTTGGAAAGACGGCAGCAAAGAAAACCATG	TGAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGGAG
GMI_661419362(-)	CAGAAAGAAAACCATG	TGAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA
GMI_661419363(-)	CAGCAAAGAAAACCATG	TGAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA
GMI_661419364(+)	AGCAAAGAAAACCATG	TGAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA
GMI_661419365(+)	AGCAAAGAAAACCATG	TGAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA
GMI_661419366(+)	AAACAAACATG	TGAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA

## SAMtools

SAMtools is a suite of programs for interacting with high-throughput sequencing data.

## GATK

Genome Analysis Toolkit: Variant Discovery in High-Throughput Sequencing Data.

The GATK toolkit offers a wide variety of tools with a primary focus on variant discovery and genotyping.

## VarScan

VarScan is a platform-independent software tool developed at the Genome Institute at Washington University to detect variants in NGS data.

## SAMtools

SAMtools is a suite of programs for interacting with high-throughput sequencing data.

## GATK

Genome Analysis Toolkit: Variant Discovery in High-Throughput Sequencing Data.

The GATK toolkit offers a wide variety of tools with a primary focus on variant discovery and genotyping.

## VarScan

VarScan is a platform-independent software tool developed at the Genome Institute at Washington University to detect variants in NGS data.

## SAMtools

SAMtools is a suite of programs for interacting with high-throughput sequencing data.

## GATK

Genome Analysis Toolkit: Variant Discovery in High-Throughput Sequencing Data.

The GATK toolkit offers a wide variety of tools with a primary focus on variant discovery and genotyping.

## VarScan

VarScan is a platform-independent software tool developed at the Genome Institute at Washington University to detect variants in NGS data.

## SAMtools

SAMtools consists of three separate repositories:

**SAMtools** Reading/writing/editing/indexing/viewing  
SAM/BAM/CRAM format

**BCFtools** Reading/writing BCF2/VCF/gVCF files and  
calling/filtering/summarising SNP and short indel  
sequence variants

**HTSlb** A C library for reading/writing high-throughput sequencing  
data



# NGS | 数据分析 | 流程 | 提取变异 | GATK



GERMLINE		SOMATIC	
SNPs & INDELS	COPY NUMBER	SNVs & INDELS	COPY NUMBER
<b>EXOME/PANEL + WGS</b> BWA + HaplotypeCaller GVCF	<b>EXOME/PANEL</b> In development	<b>EXOME/PANEL + WGS</b> BWA + MuTect	<b>EXOME/PANEL</b> BWA + CallSegments
<b>RNASEQ</b> STAR + HaplotypeCaller	<b>WHOLE GENOME</b> In development	<b>EXOME/PANEL + WGS</b> BWA + MuTect2 BETA	<b>WHOLE GENOME</b> In development



## SnpEff

Genetic variant annotation and effect prediction toolbox. It annotates and predicts the effects of variants on genes (such as amino acid changes).

## ANNOVAR

ANNOVAR is an efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes (including human genome hg18, hg19, hg38, as well as mouse, worm, fly, yeast and many others).



## SnpEff

Genetic variant annotation and effect prediction toolbox. It annotates and predicts the effects of variants on genes (such as amino acid changes).

## ANNOVAR

ANNOVAR is an efficient software tool to utilize update-to-date information to functionally annotate genetic variants detected from diverse genomes (including human genome hg18, hg19, hg38, as well as mouse, worm, fly, yeast and many others).



## VEP

VEP (Variant Effect Predictor) determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions.

## SeattleSeq Annotation

The SeattleSeq Annotation server provides annotation of SNVs (single-nucleotide variations) and small indels, both known and novel.



## VEP

VEP (Variant Effect Predictor) determines the effect of your variants (SNPs, insertions, deletions, CNVs or structural variants) on genes, transcripts, and protein sequence, as well as regulatory regions.

## SeattleSeq Annotation

The SeattleSeq Annotation server provides annotation of SNVs (single-nucleotide variations) and small indels, both known and novel.



## SIFT

SIFT predicts whether an amino acid substitution affects protein function. SIFT prediction is based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences, collected through PSI-BLAST. SIFT can be applied to naturally occurring nonsynonymous polymorphisms or laboratory-induced missense mutations.

## PolyPhen-2

PolyPhen-2 (Polymorphism Phenotyping v2) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations.

## SIFT

SIFT predicts whether an amino acid substitution affects protein function. SIFT prediction is based on the degree of conservation of amino acid residues in sequence alignments derived from closely related sequences, collected through PSI-BLAST. SIFT can be applied to naturally occurring nonsynonymous polymorphisms or laboratory-induced missense mutations.

## PolyPhen-2

PolyPhen-2 (Polymorphism Phenotyping v2) is a tool which predicts possible impact of an amino acid substitution on the structure and function of a human protein using straightforward physical and comparative considerations.

## Genome Browser

interactively visualize genomic data

## IGV

The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets.

## Tablet

Tablet is a lightweight, high-performance graphical viewer for next generation sequence assemblies and alignments.

## Circos

Circos is a software package for visualizing data and information. It visualizes data in a circular layout — this makes Circos ideal for exploring relationships between objects or positions.

## Genome Browser

interactively visualize genomic data

## IGV

The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets.

## Tablet

Tablet is a lightweight, high-performance graphical viewer for next generation sequence assemblies and alignments.

## Circos

Circos is a software package for visualizing data and information. It visualizes data in a circular layout — this makes Circos ideal for exploring relationships between objects or positions.

## Genome Browser

interactively visualize genomic data

## IGV

The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets.

## Tablet

Tablet is a lightweight, high-performance graphical viewer for next generation sequence assemblies and alignments.

## Circos

Circos is a software package for visualizing data and information. It visualizes data in a circular layout — this makes Circos ideal for exploring relationships between objects or positions.

## Genome Browser

interactively visualize genomic data

## IGV

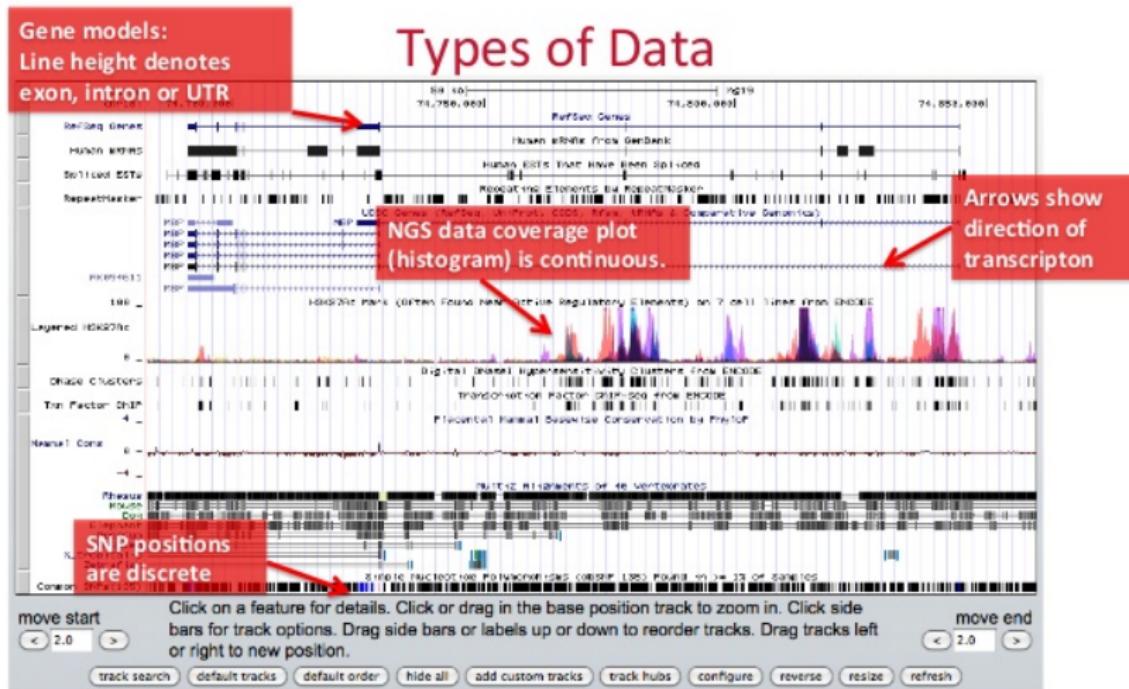
The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets.

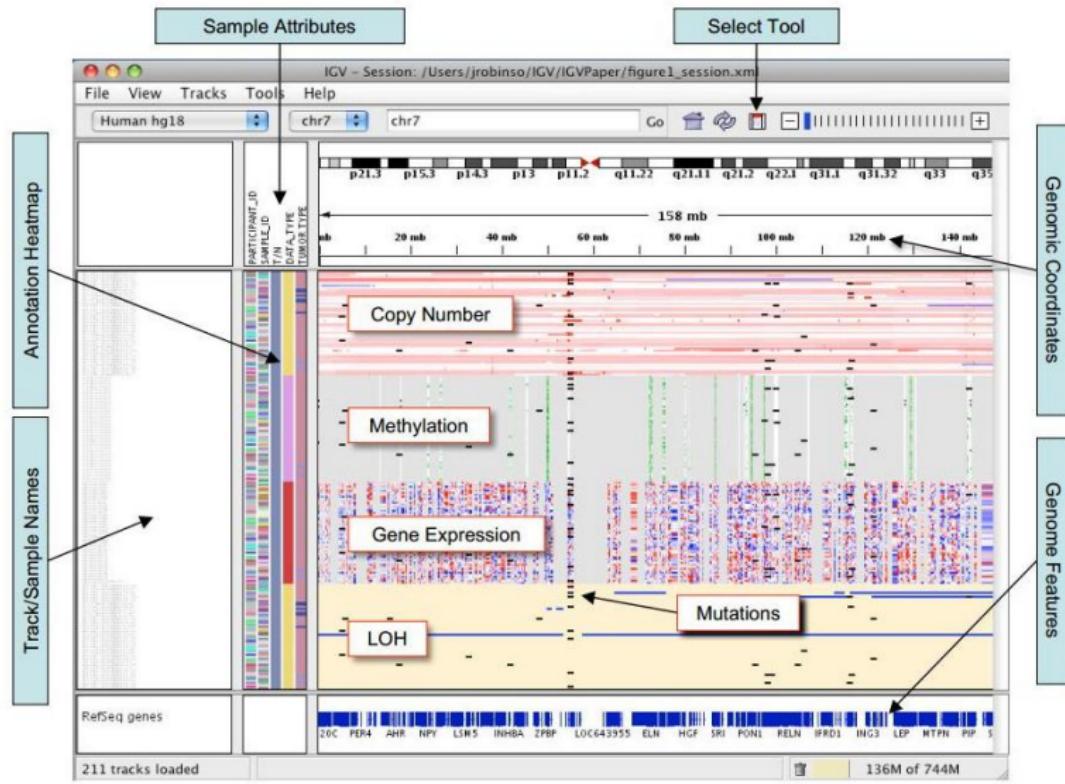
## Tablet

Tablet is a lightweight, high-performance graphical viewer for next generation sequence assemblies and alignments.

## Circos

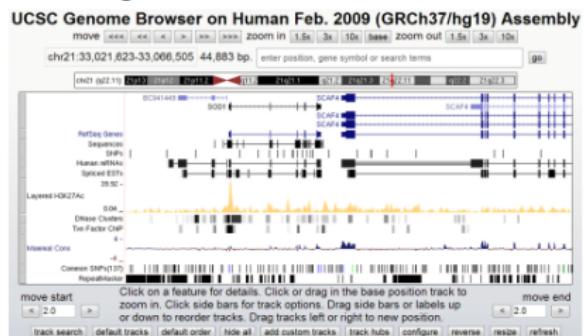
Circos is a software package for visualizing data and information. It visualizes data in a circular layout — this makes Circos ideal for exploring relationships between objects or positions.





## Genome browser

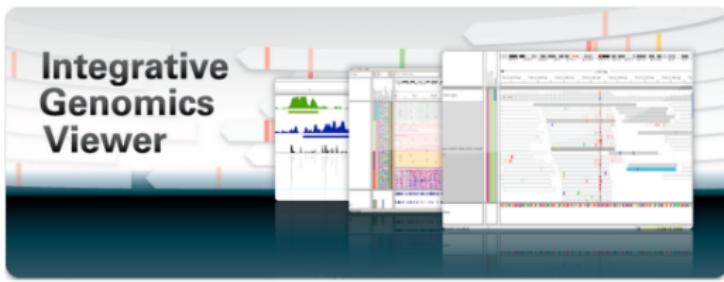
UCSC genome browser

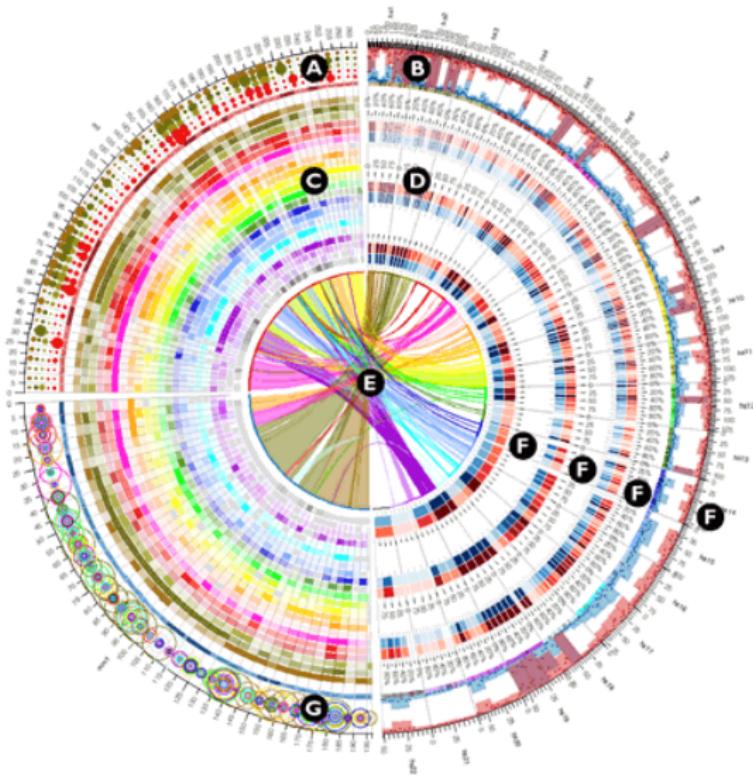


**Pros:** very comprehensive  
**Cons:** data have to be uploaded or transmitted via network dynamically

V.S.

**Pros:** locally installed  
**Cons:** less genome annotation





## bedtools

Collectively, the bedtools utilities are a swiss-army knife of tools for a wide-range of genomics analysis tasks. The most widely-used tools enable genome arithmetic: that is, set theory on the genome.

## BEDOPS

BEDOPS is an open-source command-line toolkit that performs highly efficient and scalable Boolean and other set operations, statistical calculations, archiving, conversion and other management of genomic data of arbitrary scale.



## bedtools

Collectively, the bedtools utilities are a swiss-army knife of tools for a wide-range of genomics analysis tasks. The most widely-used tools enable genome arithmetic: that is, set theory on the genome.

## BEDOPS

BEDOPS is an open-source command-line toolkit that performs highly efficient and scalable Boolean and other set operations, statistical calculations, archiving, conversion and other management of genomic data of arbitrary scale.



Utility	Description
<b>annotate</b>	Annotate coverage of features from multiple files.
<b>bamtobed</b>	Convert BAM alignments to BED (& other) formats.
<b>bamtofastq</b>	Convert BAM records to FASTQ records.
<b>bed12tobed6</b>	Breaks BED12 intervals into discrete BED6 intervals.
<b>bedpetobam</b>	Convert BEDPE intervals to BAM records.
<b>bedtobam</b>	Convert intervals to BAM records.
<b>closest</b>	Find the closest, potentially non-overlapping interval.
<b>cluster</b>	Cluster (but don't merge) overlapping/nearby intervals.
<b>complement</b>	Extract intervals _not_ represented by an interval file.
<b>coverage</b>	Compute the coverage over defined intervals.
<b>expand</b>	Replicate lines based on lists of values in columns.
<b>flank</b>	Create new intervals from the flanks of existing intervals.
<b>genomcov</b>	Compute the coverage over an entire genome.
<b>getfasta</b>	Use intervals to extract sequences from a FASTA file.
<b>groupby</b>	Group by common cols. & summarize oth. cols. (~ SQL "groupBy")
<b>igv</b>	Create an IGV snapshot batch script.
<b>intersect</b>	Find overlapping intervals in various ways.
<b>jaccard</b>	Calculate the Jaccard statistic b/w two sets of intervals.
<b>links</b>	Create a HTML page of links to UCSC locations.
<b>makewindows</b>	Make interval "windows" across a genome.
<b>map</b>	Apply a function to a column for each overlapping interval.
<b>maskfasta</b>	Use intervals to mask sequences from a FASTA file.
<b>merae</b>	Combine overlapping/nearby intervals into a single interval.



## Picard

Picard is a set of command line tools for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.

## csvkit

csvkit is a suite of command-line tools for converting to and working with CSV, the king of tabular file formats.



## Picard

Picard is a set of command line tools for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF.

## csvkit

csvkit is a suite of command-line tools for converting to and working with CSV, the king of tabular file formats.



## csvkit

`in2csv` the Excel killer

`csvlook` data periscope

`csvcut` data scalpel

`csvstat` statistics without code

`csvgrep` find the data you need

`csvsort` order matters

`csvjoin` merging related data

`csvstack` combining subsets

`csvsql & sql2csv` ultimate power

`csvjson` going online

`csvpy` going into code

`csvformat` for legacy systems

`csvclean` clean common syntax errors

# Combining tools in a pipeline

- Linux Command-line Tools
- Shell script, Makefile
- GUI Based pipeline
  - DNANexus
  - SevenBridges Genomics
- Galaxy
  - Open Source
  - Wrapper for command line utilities
  - Workflows
    - Save all steps you did in your analysis
    - Return the entire analysis on a new dataset
    - Share your workflow with other people



## Galaxy

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

## Taverna

Taverna is an open source and domain-independent Workflow Management System – a suite of tools used to design and execute scientific workflows and aid in silico experimentation.



## Galaxy

Galaxy is an open, web-based platform for data intensive biomedical research. Whether on the free public server or your own instance, you can perform, reproduce, and share complete analyses.

## Taverna

Taverna is an open source and domain-independent Workflow Management System – a suite of tools used to design and execute scientific workflows and aid in silico experimentation.



## WDL

Workflow Description Language (WDL) is a workflow language meant to be read and written by humans. The Workflow Description Language is a domain specific language for describing tasks and workflows.

## Bpipe

Bpipe provides a platform for running big bioinformatics jobs that consist of a series of processing stages - known as 'pipelines'.

## BioX::Workflow

A very opinionated template based workflow writer. This module was written with Bioinformatics workflows in mind, but should be extensible to any sort of workflow or pipeline.

## WDL

Workflow Description Language (WDL) is a workflow language meant to be read and written by humans. The Workflow Description Language is a domain specific language for describing tasks and workflows.

## Bpipe

Bpipe provides a platform for running big bioinformatics jobs that consist of a series of processing stages - known as 'pipelines'.

## BioX::Workflow

A very opinionated template based workflow writer. This module was written with Bioinformatics workflows in mind, but should be extensible to any sort of workflow or pipeline.

## WDL

Workflow Description Language (WDL) is a workflow language meant to be read and written by humans. The Workflow Description Language is a domain specific language for describing tasks and workflows.

## Bpipe

Bpipe provides a platform for running big bioinformatics jobs that consist of a series of processing stages - known as 'pipelines'.

## BioX::Workflow

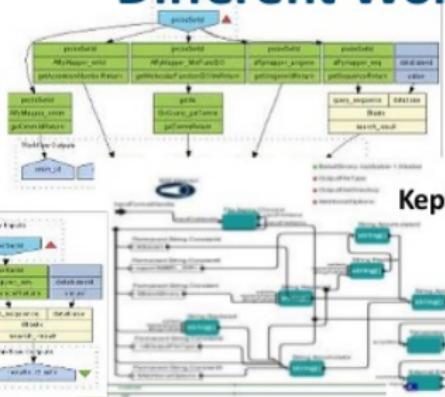
A very opinionated template based workflow writer. This module was written with Bioinformatics workflows in mind, but should be extensible to any sort of workflow or pipeline.

Tool	GUI	Command Line ( ** )	Built in			Online Data Source Integration	Need Programming Knowledge?	Easy Shell Script Portability
			Audit Trail	Cluster Support	Workflow Sharing			
Bpipe	No	Yes	Yes	Yes	No	No	No	Yes
Ruffus	No	Yes	Yes	No	No	No	Yes	No
Galaxy	Yes	No	Yes	Yes	Yes	Yes	No	No
Taverna	Yes	No	Yes	Yes	Yes	Yes	No	No
Pegasus	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No

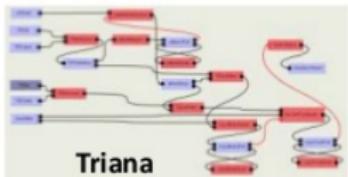




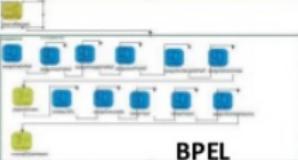
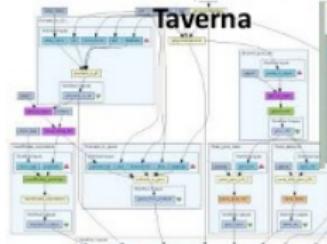
## Different Workflow Systems



Kepler



Triana



BPEL



Galaxy

myGrid



# Reproducibility in Genomics

18 *Nat. Genetics* experiments in microarray gene expression

<50% of reproducible

Problems

- missing data (38%)
- missing software, hardware details (50%)
- missing methods, processing details (66%)

14 re-sequencing experiments in *Nat. Genetics, Nature, Science*

0% reproducible?

Problems

- missing primary data (50%)
- tools unavailable (50%)
- missing parameter setting, tool versions (100%)

Ioannidis, J.P.A. et al. "Repeatability of published microarray gene expression analyses." *Nat Genet* 41, 149-155 (2009)

"Devil in the details," *Nature*, vol. 470, 305-306 (2011).



## Galaxy project: fundamental questions

When Biology (or any science) becomes dependent on computational methods:

- How can those methods best be made **accessible** to scientists?
- How best to ensure that analyses are **reproducible**?
- How best to facilitate **transparent** communication and reuse of analyses?



# What is Galaxy?

- A **data analysis and integration** tool
- A **(free for everyone) web service** integrating a wealth of tools, compute resources, terabytes of reference data and permanent storage
- **Open source software** that makes integrating your own tools and data and customizing for your own site simple



Learn the strengths and weaknesses of the tools you have ready access to. Are they a good match for the questions you are asking?

If not, then research alternatives, identify good options and then work with your bioinformatics/systems people to get access to those tools.



## bioconda

Bioconda is a channel for the conda package manager specializing in bioinformatics software. Bioconda consists of:

- a repository of recipes hosted on GitHub
- a build system that turns these recipes into conda packages
- a repository of >1500 bioinformatics packages ready to use with conda install
- Over 130 contributors that add, modify, update and maintain the recipes



## Using bioconda

bioconda supports only 64-bit Linux and Mac OSX.

- ① Install conda
- ② Set up channels (It is important to add them in this order)

```
1 conda config --add channels conda-forge
2 conda config --add channels defaults
3 conda config --add channels r
4 conda config --add channels bioconda
```

- ③ Install packages

```
1 # install into the current conda
  environment:
2 conda install bwa
3 # a new environment can be created
4 conda create -n aligners bwa bowtie
```

# 教学提纲

1

## 系统生物学

2

## 基因组学

3

## 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

## 数据库与数据格式

5

## 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

## 外显子组测序

## ● 简介

## ● 操作流程

## ● 应用实例

## 7 转录组学

## 8 RNA-Seq

## ● 概述

## ● 数据分析

## ● 应用实例

## 9 顺反组

## ● 概述

## ● ChIP-Seq

## 10 表观遗传学

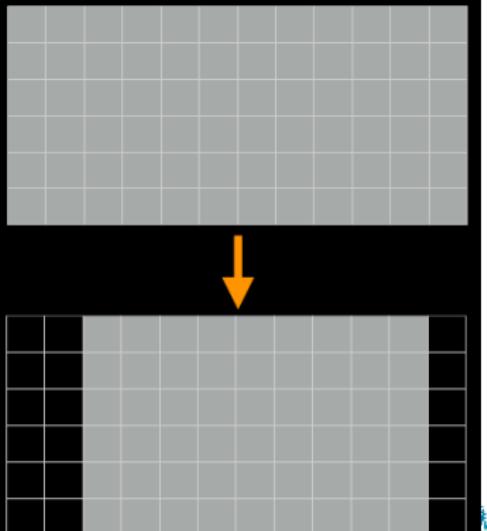
## ● 概述

## ● Methyl-Seq



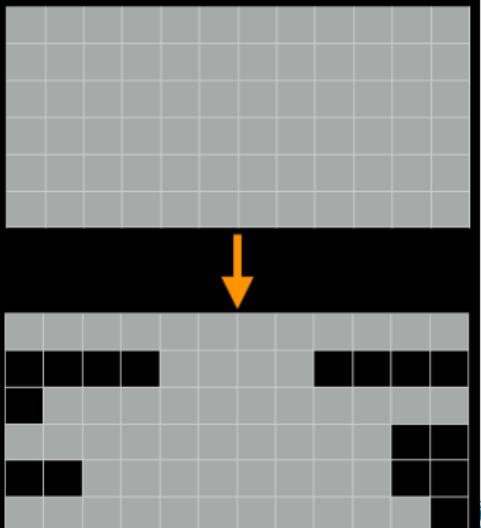
## NGS Data Quality: Trim as we see fit

- Trim as we see fit: Option 1
- NGS QC and Manipulation → **FASTQ Trimmer by column**
- Trim same number of columns from every record
- Can specify different trim for 5' and 3' ends



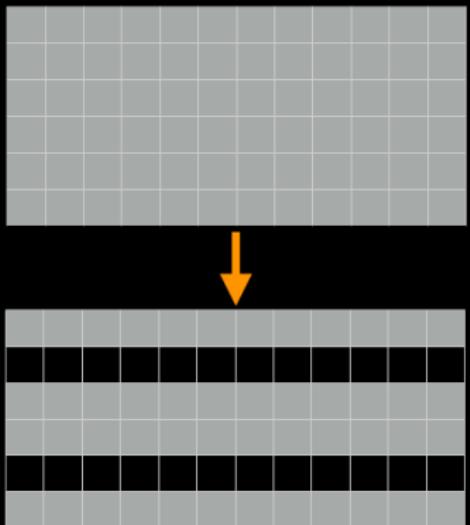
## NGS Data Quality: Base Quality Trimming

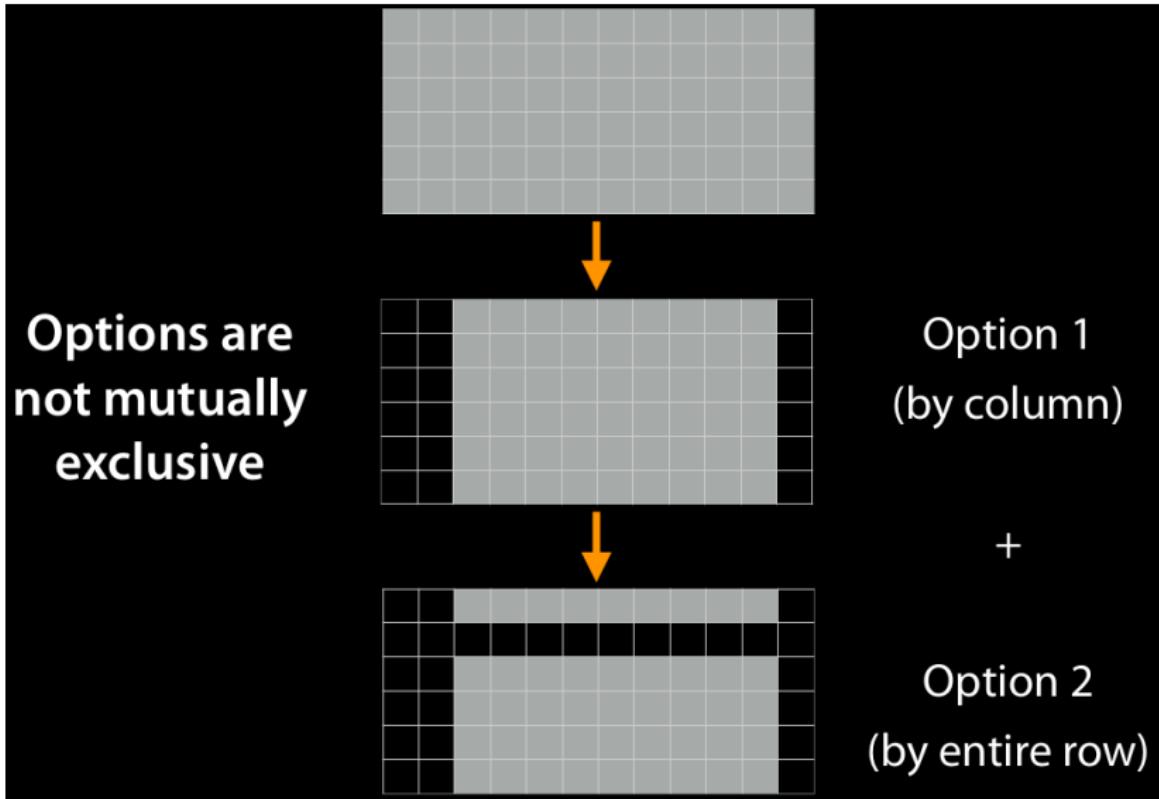
- Trim as we see fit: Option 3
  - NGS QC and Manipulation → **FASTQ Quality Trimmer by sliding window**
  - Trim from both ends, using sliding windows, until you hit a high-quality section.
  - Produces variable length reads



## NGS Data Quality: Base Quality Trimming

- Trim Filter as we see fit: Option 2
  - NGS QC and Manipulation →  
**Filter FASTQ reads by quality score and length**
  - Keep or discard whole reads
  - Can have different thresholds for different regions of the reads.
  - Keeps original read length.





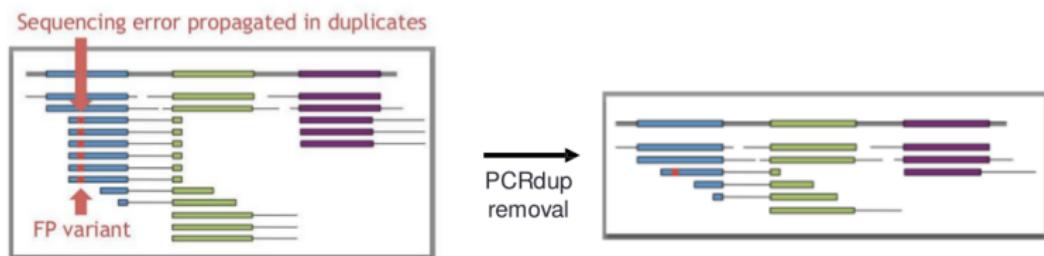
## Trim? As we see fit?

- Choice depends on downstream tools
- Find out assumptions & requirements for downstream tools and make appropriate choice(s) now.
- How to do that?
  - Read the tool documentation
  - <http://biostars.org/>
  - <http://seqanswers.com/>
  - <http://galaxyproject.org/search>



## Removing Duplicates

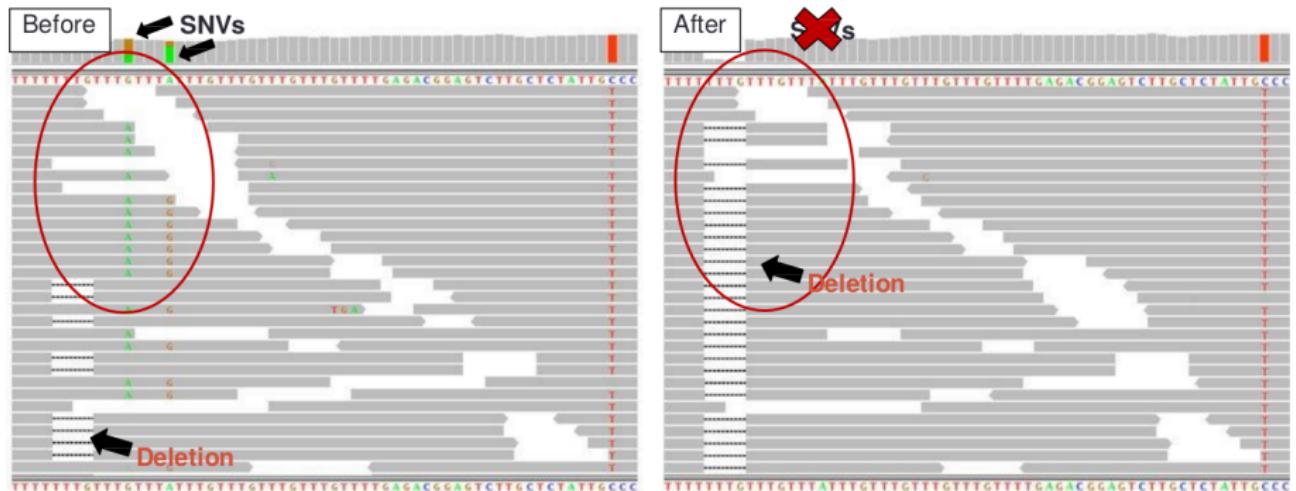
- **Duplicates reads:** different reads having the same sequence caused by PCR amplification during sequencing library preparation
- The removal of the duplicates depends on the application (not suitable for sequencing on small target)



- **Galaxy:** Use “Mark Duplicates reads” from “NGS:Picard” to **mark** duplicates (don’t remove them)  
→ If duplicates are marked, samtools and GATK tools will ignore them
- **Galaxy:** Run “Flagstat” on the output BAM to see the number of PCR duplicates

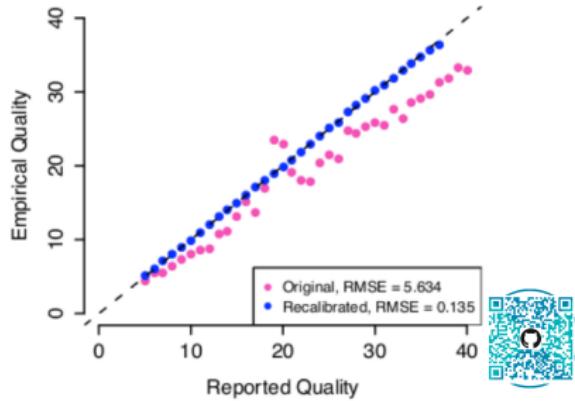


## Local realignment around indels



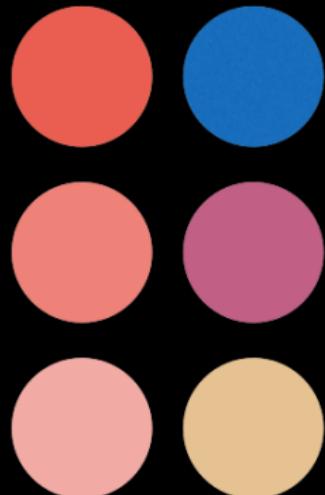
## GATK Preprocess: Base Quality Score Recalibration

- Analyze covariation among several features of a base, e.g:
  - Original quality score*
  - Position within the read (machine cycle)*
  - Preceding and current nucleotides (chemistry effect)*
  - Sequencing technology...*
- Adjust the quality score associated to each sequenced base to be more accurate  
→ Remove systematic biases



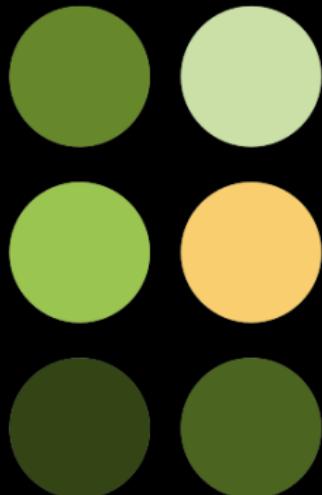
# Replicates, replicates, replicates

Mutant



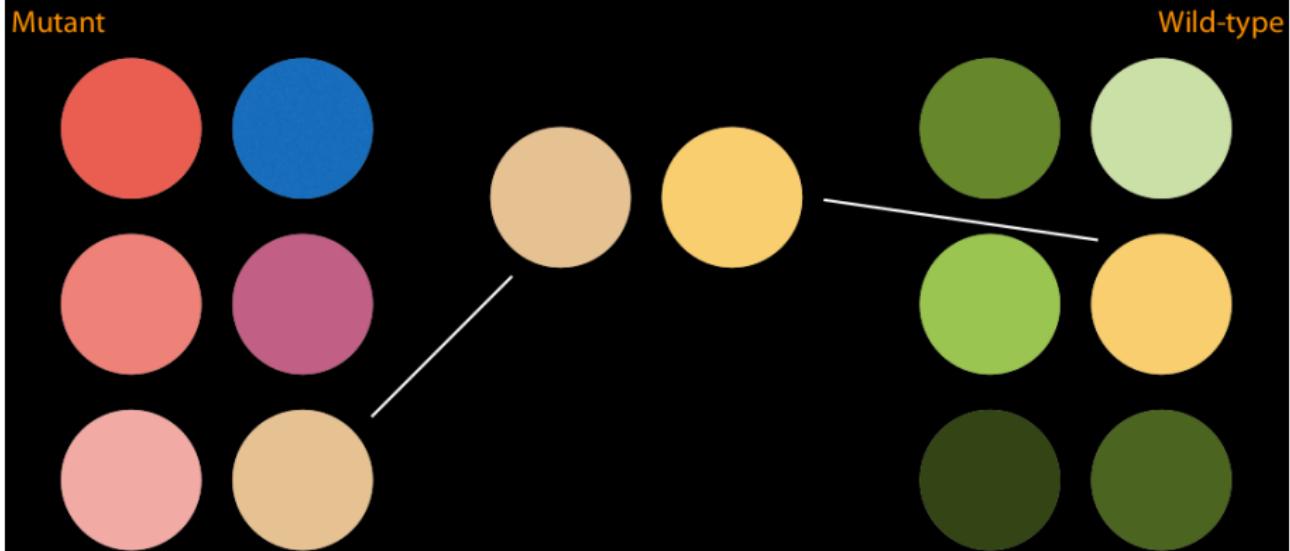
Why replicate?

Wild-type



Biological heterogeneity

# Unreplicated design



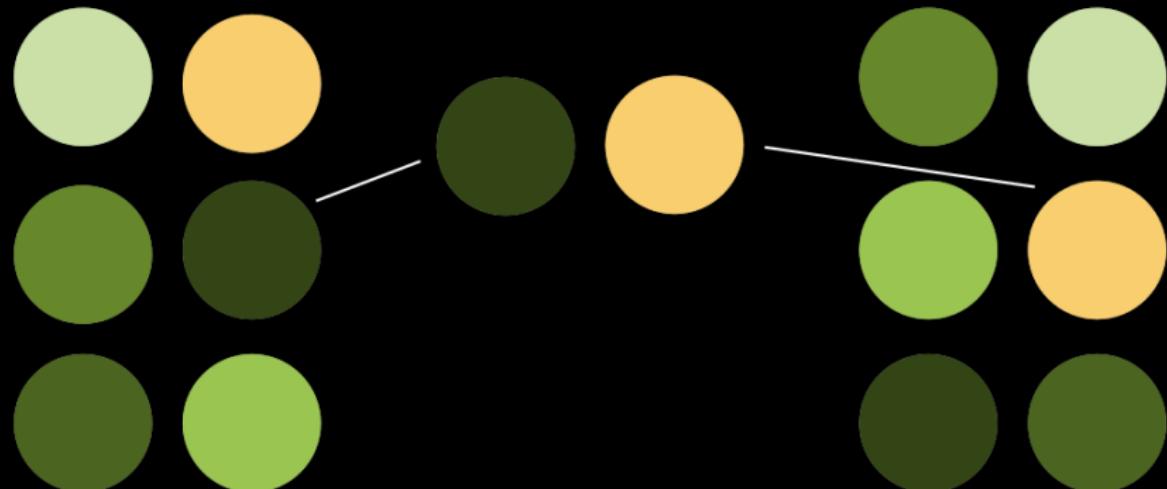
Here, groups differ, but single replicates from each group very similar



# Unreplicated design

Mutant

Wild-type



Here, groups are similar, but outlying observation from group on right makes it look like there's a big difference in unreplicated experiment



Experimental design

## Biological replicates

Reference genome?

Good gene annotation?

Read depth

Barcoding

Read length

Paired vs. single-end



High Accuracy  
Low Precision



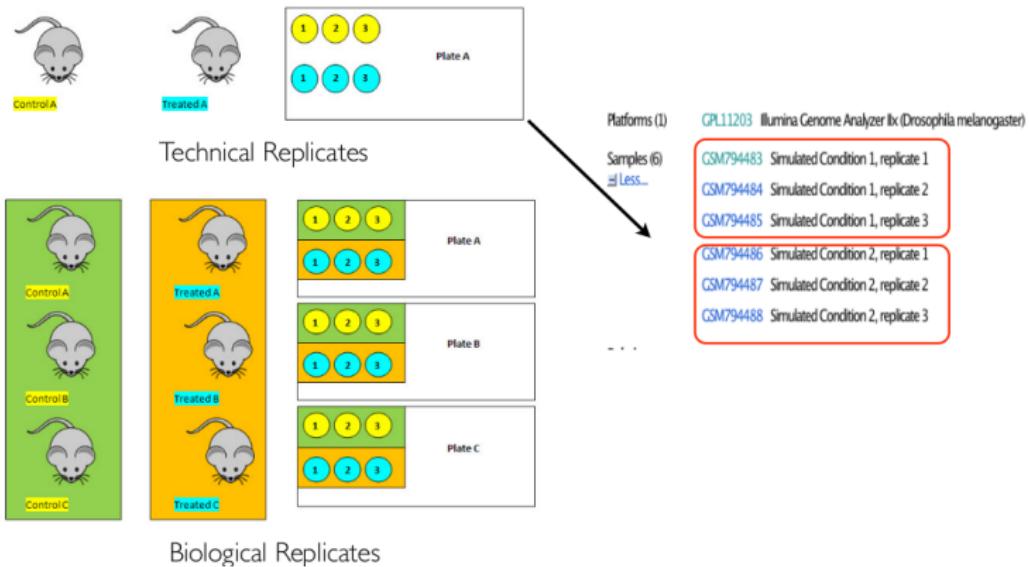
Low Accuracy  
High Precision

Biological variation

Technical variation



# Biological replicates vs. technical replicates



## How much data do we need?

~15-20K genes expressed in a tissue or cell line.  
Genes are on average 3Kbp

For 1x coverage using 100 bp reads, one would need 600K sequence reads

In reality, we need MUCH higher coverage to accurately detect all genes and estimate their expression levels.

**30-50 million reads for >90% genes detected**

Experimental design

Biological replicates

Reference genome?

Good gene annotation?

Read depth

Barcoding

Read length

Paired vs. single-end

$\text{Uniq seq} = 4^{\text{read length}}$

Read length	Unique seq
25	$1.1 \times 10^{15}$
50	$1.3 \times 10^{30}$
100	$1.6 \times 10^{60}$

~60 million ( $6 \times 10^7$ ) coding bases  
in vertebrate genome

usually 50bp is enough, but more  
helps for exon-exon junctions



Experimental design

Biological replicates

Reference genome?

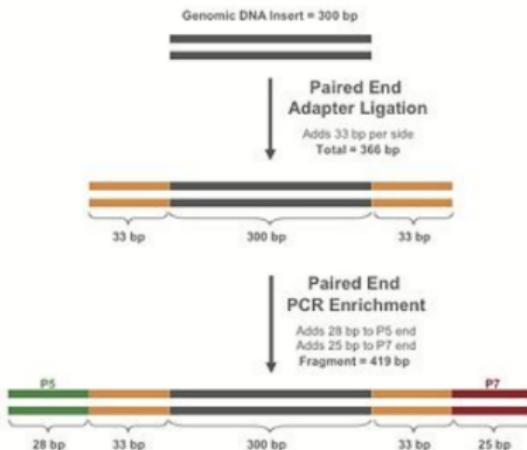
Good gene annotation?

Read depth

Barcode/Index

Read length

Paired vs. single-end



BROAD  
INSTITUTE illumina®

180-250 million reads / lane  
Run 4-8 samples / lane

Know your research question:

**comparing whether two genes are expressed differently from each other in same group?**

**comparing expression of the same gene between different groups?**

**time course experiments?**

Increase number of biological replicates rather than sequencing depth

Discuss your experimental design with a bioinformatician/biostatistician  
**BEFORE** running the experiment!



# 教学提纲

1

## 系统生物学

2

## 基因组学

3

## 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

## 数据库与数据格式

5

## 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

## 外显子组测序

## ● 简介

## ● 操作流程

## ● 应用实例

## 7 转录组学

## 8 RNA-Seq

## ● 概述

## ● 数据分析

## ● 应用实例

## 9 顺反组

## ● 概述

## ● ChIP-Seq

## 10 表观遗传学

## ● 概述

## ● Methyl-Seq



# 教学提纲

1

## 系统生物学

2

## 基因组学

3

## 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

## 数据库与数据格式

5

## 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

## 外显子组测序

## ● 简介

- 操作流程
- 应用实例

7

## 转录组学

8

## RNA-Seq

- 概述
- 数据分析
- 应用实例

9

## 顺反组

- 概述
- ChIP-Seq

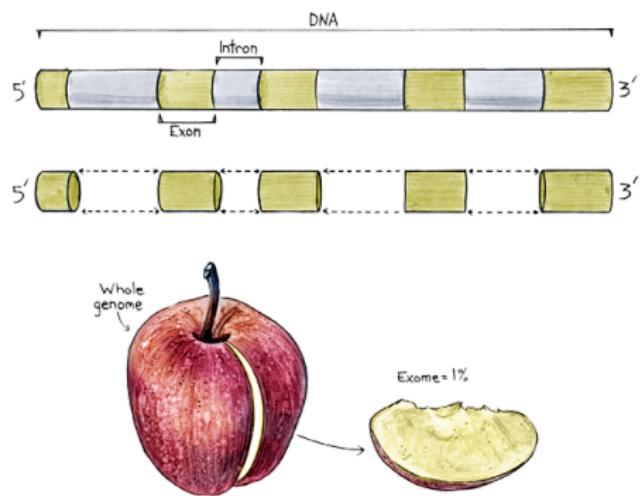
10

## 表观遗传学

- 概述
- Methyl-Seq

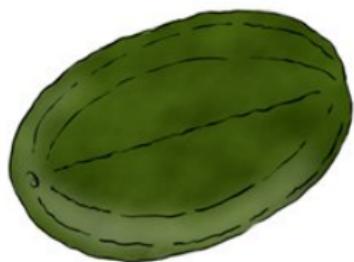


The exome is the part of the genome formed by exons, the sequences which when transcribed remain within the mature RNA after introns are removed by RNA splicing. It consists of all DNA that is transcribed into mature RNA in cells of any type as distinct from the transcriptome, which is the RNA that has been transcribed only in a specific cell population.



The exome of the human genome consists of roughly 180,000 exons constituting about 1% of the total genome, or about 30 megabases of DNA. Though comprising a very small fraction of the genome, mutations in the exome are thought to harbor 85% of mutations that have a large effect on disease.

WATERMELON = GENOME



SLICE = EXOME 1-2%



SEEDS = GENES



## WES

Exome sequencing, also known as whole exome sequencing (WES or WXS), is a technique for sequencing all the expressed genes in a genome (known as the exome).

## Projects

Examples of research projects using exome sequencing include:

- PGP (Personal Genome Project)
- RGI (Rare Genomics Institute)
- NIH-funded Exome Project
- NHGRI-funded Mendelian Exome Project
- NHLBI Grand Opportunity Exome Sequencing Project
- microarray-based Nimblegen SeqCap EZ Exome from Roche Applied Science

## WES

Exome sequencing, also known as whole exome sequencing (WES or WXS), is a technique for sequencing all the expressed genes in a genome (known as the exome).

## Projects

Examples of research projects using exome sequencing include:

- PGP (Personal Genome Project)
- RGI (Rare Genomics Institute)
- NIH-funded Exome Project
- NHGRI-funded Mendelian Exome Project
- NHLBI Grand Opportunity Exome Sequencing Project
- microarray-based Nimblegen SeqCap EZ Exome from Roche Applied Science

Exome sequencing consists of first selecting only the subset of DNA that encodes proteins (known as exons) and then sequencing that DNA using any high-throughput DNA sequencing technology.

Humans have about **180,000 exons**, constituting about **1% of the human genome**, or approximately **30 million base pairs**.

The goal of this approach is to identify genetic variation that is responsible for both **Mendelian and common diseases** such as Miller syndrome and Alzheimer's disease without the high costs associated with whole-genome sequencing.

Exome sequencing has proved to be an efficient strategy to determine the genetic basis of more than two dozen **Mendelian or single gene disorders**.



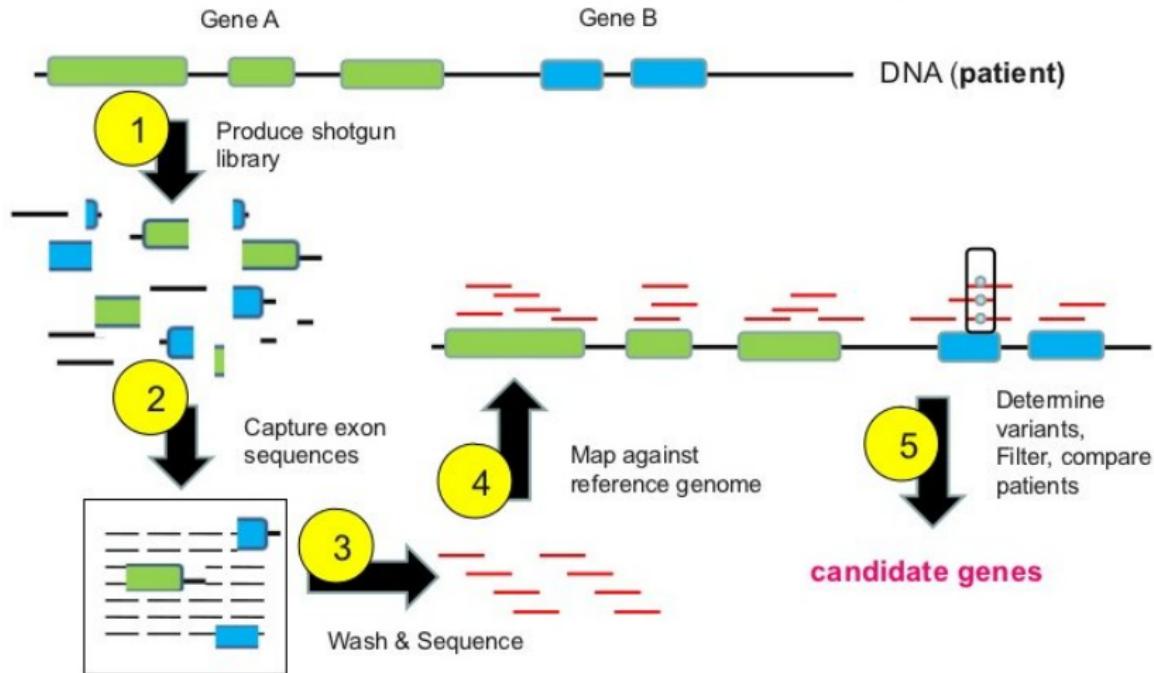
## Why exome sequencing?

- Whole-genome sequencing of individual humans is increasingly practical . But cost remains a key consideration and added value of intergenic mutations is not cost-effective.
- Alternative approach: targeted resequencing of all protein-coding subsequences (exome sequencing, ~1% of human genome)
- Linkage analysis/positional cloning studies that focused on protein coding sequences were highly successful at identification of variants underlying monogenic diseases (when adequately powered)
- Known allelic variants known to underlie Mendelian disorders disrupt protein-coding sequences
- Large fraction of rare non-synonymous variants in human genome are predicted to be deleterious
- Splice acceptor and donor sites are also enriched for highly functional variation and are therefore targeted as well

The exome represents a highly enriched subset of the genome in which to search for variants with large effect sizes

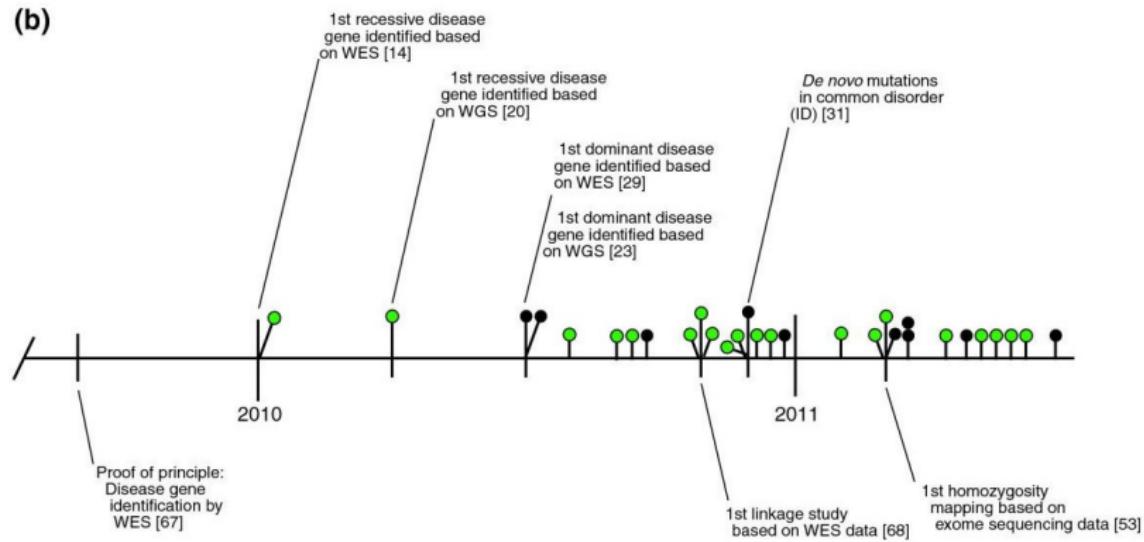


## How does exome sequencing work?



# WES | 简介 | Exome sequencing

(b)



Key:

| = Notable publication

● = Recessive disease gene identification by WES

● = Dominant disease gene identification by WES



## WGS

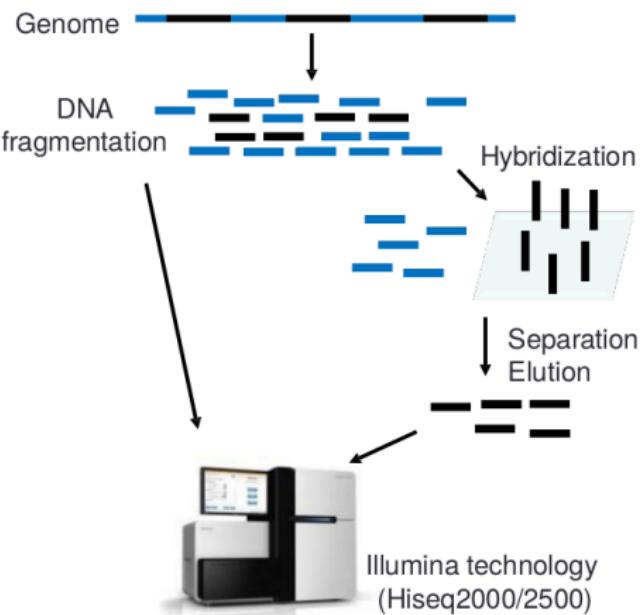
Whole genome sequencing (also known as WGS, full genome sequencing, complete genome sequencing, or entire genome sequencing) is a laboratory process that determines the complete DNA sequence of an organism's genome at a single time.

This entails sequencing all of an organism's chromosomal DNA as well as DNA contained in the mitochondria and, for plants, in the chloroplast.



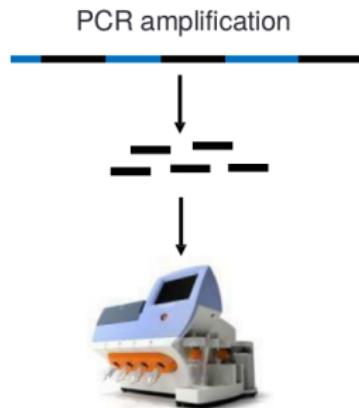
## Genome or Target DNA sequencing

### Whole Genome VS Exome Sequencing



### Amplicon Sequencing

→ Sequencing of a dedicated panel of genes/hotspots



Mostly Ion Torrent technology  
(PGM/Proton)



## Exome sequencing

- the most efficient way to identify the genetic variants in all of an individual's genes ⇒ especially effective in the study of **rare Mendelian diseases**
- severe disease causing variants are much more likely (but by no means exclusively) to be in the protein coding sequence ⇒ focusing on this **1% costs** far less than whole genome sequencing but still produces a high yield of relevant variants
- both to find mutations in genes already **known** to cause disease as well as to identify **novel genes** by comparing exomes from patients with similar features



## Exome

Exome sequencing is only able to identify those variants found in the coding region of genes which affect protein function. It is not able to identify the **structural and non-coding variants** associated with the disease, which can be found using other methods such as whole genome sequencing. There remains **99% of the human genome** that is not covered using exome sequencing.



By using exome sequencing, fixed-cost studies can sequence samples to much higher depth than could be achieved with whole genome sequencing. This additional depth makes exome sequencing well suited to several applications that need reliable variant calls.

- Rare variant mapping in complex disorders
- Discovery of Mendelian disorders
- Clinical diagnostics
- Direct-to-consumer exome sequencing



# 教学提纲

1

系统生物学

2

基因组学

3

测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

数据库与数据格式

5

二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

外显子组测序

● 简介

● 操作流程

● 应用实例

7 转录组学

8 RNA-Seq

● 概述

● 数据分析

● 应用实例

9 顺反组

● 概述

● ChIP-Seq

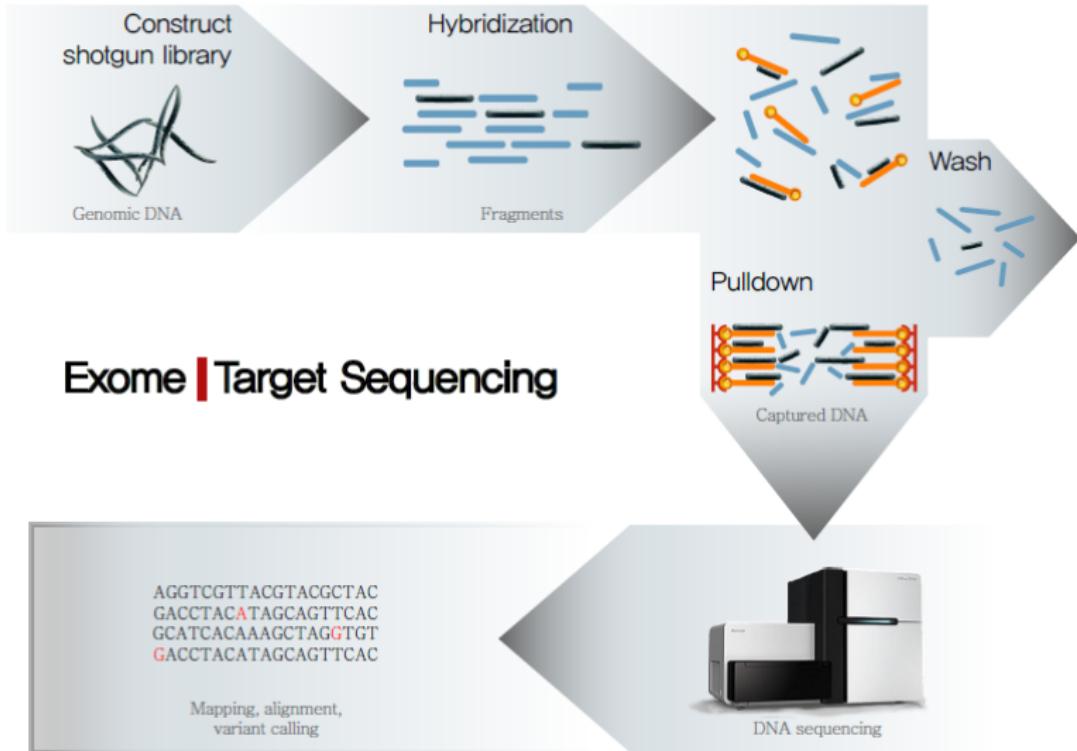
10 表观遗传学

● 概述

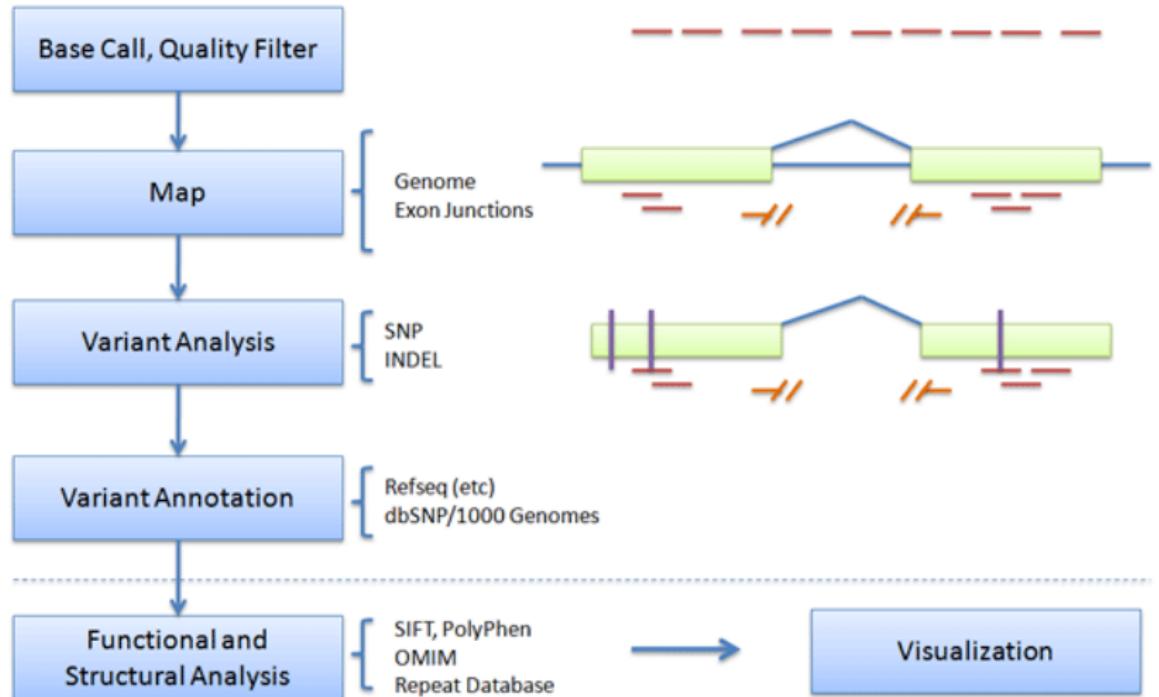
● Methyl-Seq



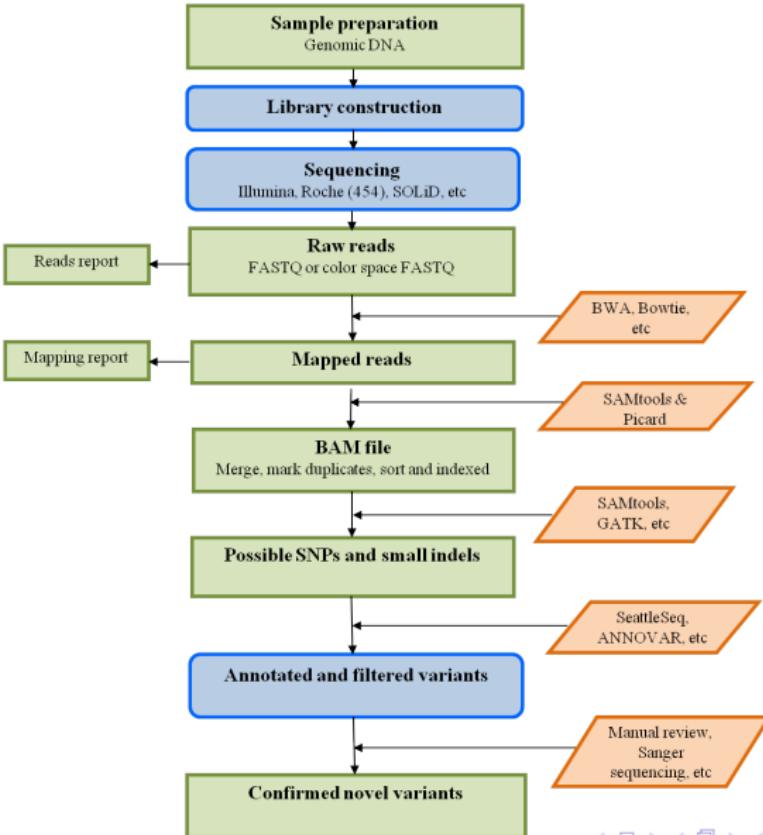
# WES | 流程 | 实验



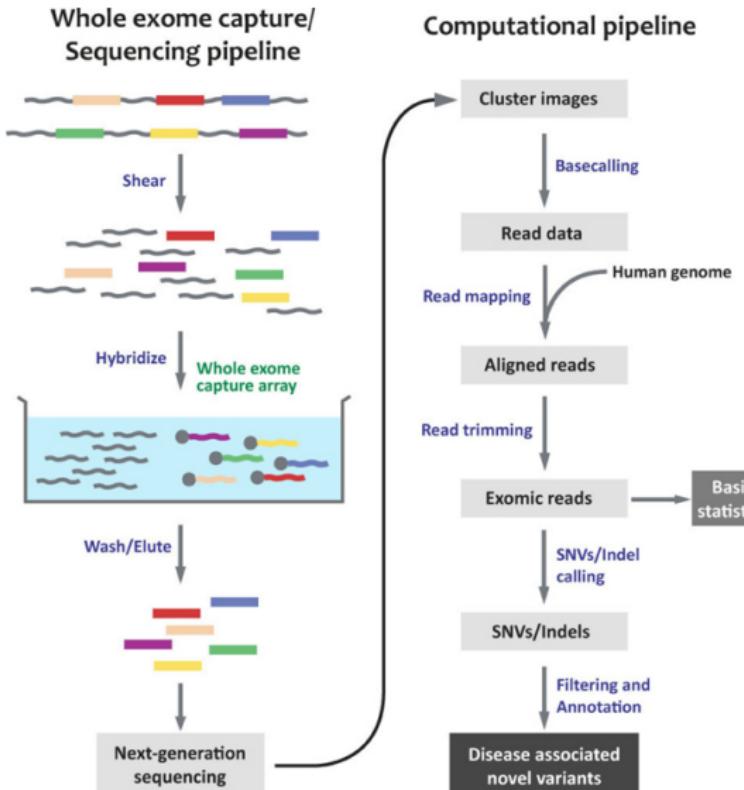
# Edge Exome Analysis Pipeline



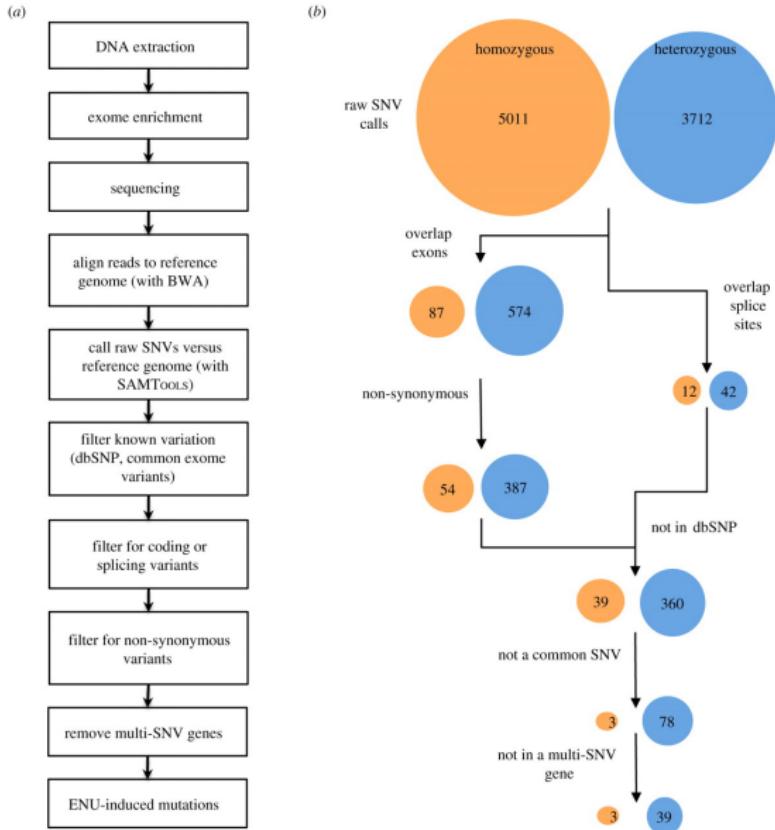
# WES | 流程 | 生信



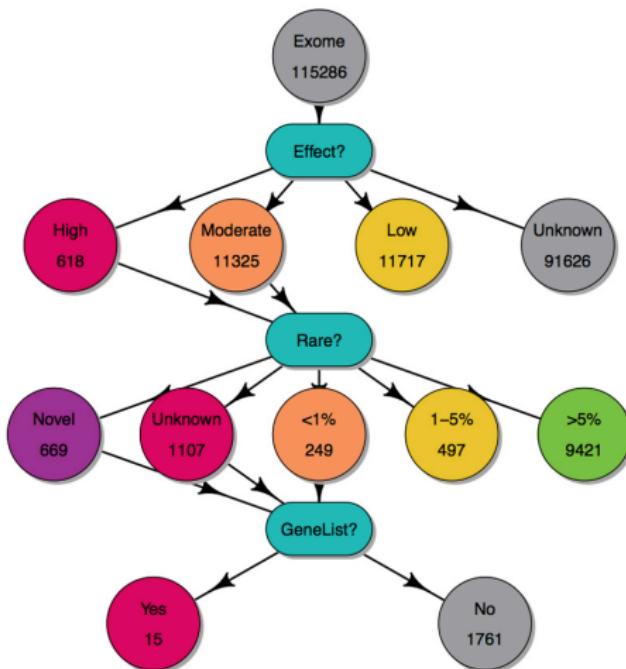
# WES | 流程 | 实验 & 生信



# WES | 流程 | 变异分析



## Filtering your variants



# 教学提纲

1

## 系统生物学

2

## 基因组学

3

## 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

## 数据库与数据格式

5

## 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

## 外显子组测序

## ● 简介

## ● 操作流程

## ● 应用实例

## 7 转录组学

## 8 RNA-Seq

### ● 概述

### ● 数据分析

### ● 应用实例

## 9 顺反组

### ● 概述

### ● ChIP-Seq

## 10 表观遗传学

### ● 概述

### ● Methyl-Seq



## 2009, Nature

Ng SB, Turner EH, Robertson PD, Flygare SD, Bigham AW, Lee C, Shaffer T, Wong M, Bhattacharjee A, Eichler EE, Bamshad M, Nickerson DA, Shendure J (10 September 2009). "Targeted capture and massively parallel sequencing of 12 human exomes". *Nature*. 461 (7261): 272–276.

A study published in September 2009 discussed a proof of concept experiment to determine if it was possible to identify causal genetic variants using exome sequencing. They sequenced four individuals with Freeman-Sheldon syndrome (FSS) (OMIM 193700), a rare autosomal dominant disorder known to be caused by a mutation in the gene MYH3. Eight HapMap individuals were also sequenced to remove common variants in order to identify the causal gene for FSS. After exclusion of common variants, the authors were able to identify MYH3, which **confirms that exome sequencing can be used to identify causal variants of rare disorders**. This was the first reported study that used exome sequencing as an approach to identify an unknown causal gene for a rare mendelian disorder.

# WES | 应用实例 | 2009-Nature

# genes in which each affected has at least one ...	FSS24895	FSS24895	FSS24895	FSS24895	ANY 3 OF 4
	FSS10208	FSS10208	FSS10066	FSS22194	FSS24895
	FSS10066	FSS22194	FSS10066	FSS22194	FSS10208
	FSS22194	FSS22194	FSS22194	FSS22194	FSS10066
	FSS22194	FSS22194	FSS22194	FSS22194	FSS22194
nonsynonymous cSNP, splice site variant or coding indel (NS/SS/I)	4,510	3,284	2,765	2,479	3,768
NS/SS/I not in dbSNP	513	128	71	53	119
NS/SS/I not in 8 HapMap exomes	799	168	53	21	160
NS/SS/I neither in dbSNP nor 8 HapMap exomes	360	38	8	1 (MYH3)	22
... AND predicted to be damaging	160	10	2	1 (MYH3)	3



## 2009, PNAS

Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P, Nayir A, Bakkaloğlu A, Ozen S, Sanjad S, Nelson-Williams C, Farhi A, Mane S, Lifton RP (10 November 2009). "Genetic diagnosis by whole exome capture and massively parallel DNA sequencing". Proc Natl Acad Sci U S A. 106 (45): 19096–19101.

Subsequently, another group reported successful clinical diagnosis of a suspected Bartter syndrome patient of Turkish origin. Bartter syndrome is a renal salt-wasting disease. Exome sequencing revealed an unexpected well-conserved recessive mutation in a gene called SLC26A3 which is associated with congenital chloride diarrhea (CLD). This molecular diagnosis of CLD was confirmed by the referring clinician. This example **provided proof of concept of the use of whole-exome sequencing as a clinical tool in evaluation of patients with undiagnosed genetic illnesses**. This report is regarded as the first application of next generation sequencing technology for molecular diagnosis of a patient.

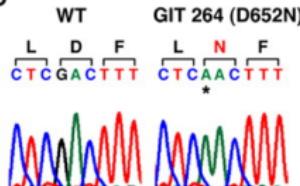


WES | 应用实例 | 2009-PNAS

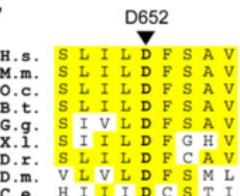
A

Reference GIT 264-1	P L N I E V P K I S L H S L I L D F S A V S F L D V S S V R G L K
Sense	P L N I E V P K I S L H S L I L D F S A V S F L D V S S V R G L K
Antisense	<p>5' - CCTCTCAACATTGAGGTCCCCAAATCAGGCTCCACAGCTCATTCCTCGACTTTTCACTGGAGCTGTCCTTCTTGATGTTCTCAGTAGGGGCCCTAAA-3'</p> <p>3' - GGAGAGTTGTAACTCCAGGGGTTTACTGGGAGGTGCGGAGTAAGAGCTGAAAGTCGTACAGGAAAGAAGTACAACAAAGAAGTCACTCCCGGAATT-5'</p> <p>3' - GGAGCGTGTAACTCAGGGGTTTACTGGGAGGTGCGGAGTAAGAGCTGAAAG-5'</p> <p>3' - GTGTAACTCAGGGGTTTACTGGGAGGTGCGGAGTAAGAGCTGAAAG-5'</p> <p>3' - AACTCCAGGGTTTCTCGCAGGGGTCGGAGTAAGAGCTGAAAGTCGT-5'</p> <p>5' - ctcagggttttagtcggaggtgcggagaagactgtaaaaagtctca-3'</p> <p>3' - CCAGGGGTTTACTGGGAGGTGCGGAGTAAGAGCTGAAAGTCGTACA-5'</p> <p>5' - ggggttttagtcggaggtgcggagaagactgtaaaaagtctcacaggaa-3'</p> <p>3' - TTTTGGTGGAGGTGCGGAGTAAGAGCTGAAAGTCGTACAGGAAAG-5'</p> <p>3' - TTTCAGGGGAGGTGCGGAGTAAGAGCTGAAAGTCGTACAGGAAAGA-5'</p> <p>3' - GTCCGAGGCCTGGAGTAAGAGCTGAAAGTCGTACAGGAAAGAACTAC-5'</p> <p>5' - cggagggtgcggagaagactgtaaaaagtctcacaggaaaactacaa-3'</p> <p>3' - GGGGGGTCGGAGTAAGAGCTGAAAGTCGTACAGGAAAGAACTACAA-5'</p> <p>5' - gaggttgcggagaagactgtaaaaagtctcacaggaaaactacaaag-3'</p> <p>3' - GGGTGGAGTAAGAGCTGAAAGTCGTACAGGAAAGAACTACAAAGAAG-5'</p> <p>5' - tcggagaagactgtaaaaagtctcacaggaaaactacaaagactca-3'</p> <p>3' - GAGTAAGAGCTAGAAAAGTCGTACAGGAAAGAACTACAAAGAAGTCACTC-5'</p> <p>5' - agatgtaaaaagtctcacaggaaaactacaaagactactcccccgg-3'</p> <p>3' - GTGAAAAGTCGTACAGGAAAGAACTACAAAGAAGTCACTCCCGGAATT-5'</p>

B



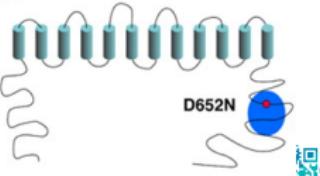
C



D



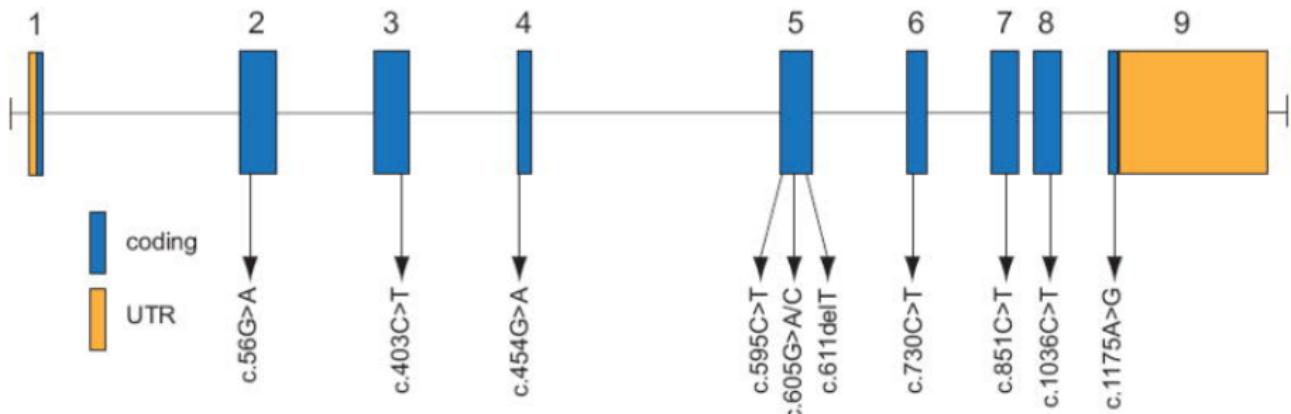
E



## 2010, NG

Sarah B Ng; Kati J Buckingham; Choli Lee; Abigail W Bigham; Holly K Tabor; Karin M Dent; Chad D Huff; Paul T Shannon; Ethylin Wang Jabs; Deborah A Nickerson; Jay Shendure; Michael J Bamshad (2010). "Exome sequencing identifies the cause of a mendelian disorder". *Nature Genetics*. 42 (1): 30–35.

A second report was conducted on exome sequencing of individuals with a mendelian disorder known as Miller syndrome (MIM#263750), a rare disorder of autosomal recessive inheritance. Two siblings and two unrelated individuals with Miller syndrome were studied. They looked at variants that have the potential to be pathogenic such as non-synonymous mutations, splice acceptor and donor sites and short coding insertions or deletions. Since Miller syndrome is a rare disorder, it is expected that the causal variant has not been previously identified. Previous exome sequencing studies of common single nucleotide polymorphisms (SNPs) in public SNP databases were used to further exclude candidate genes. After exclusion of these genes, the authors found mutations in DHODH that were shared among individuals with Miller syndrome. Each individual with Miller syndrome was a compound heterozygote for the DHODH mutations which were inherited as each parent of an affected individual was found to be a carrier.



# 教学提纲

1

系统生物学

2

基因组学

3

测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

数据库与数据格式

5

二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

外显子组测序

- 简介
- 操作流程
- 应用实例

7

转录组学

8

RNA-Seq

- 概述
- 数据分析
- 应用实例

9

顺反组

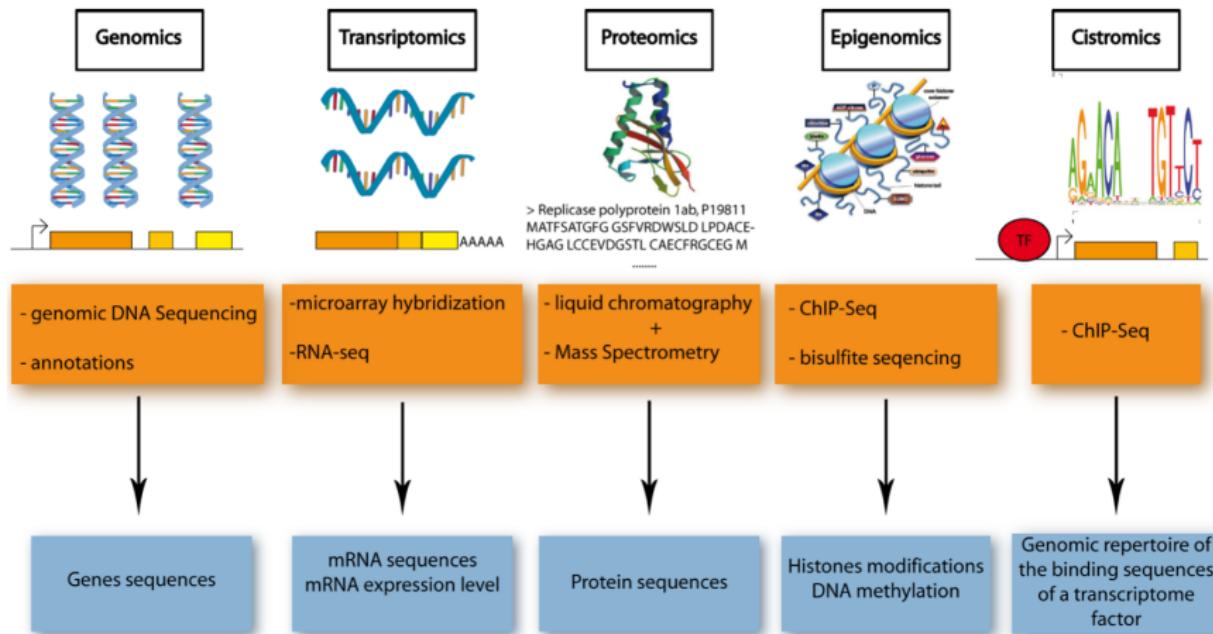
- 概述
- ChIP-Seq

10

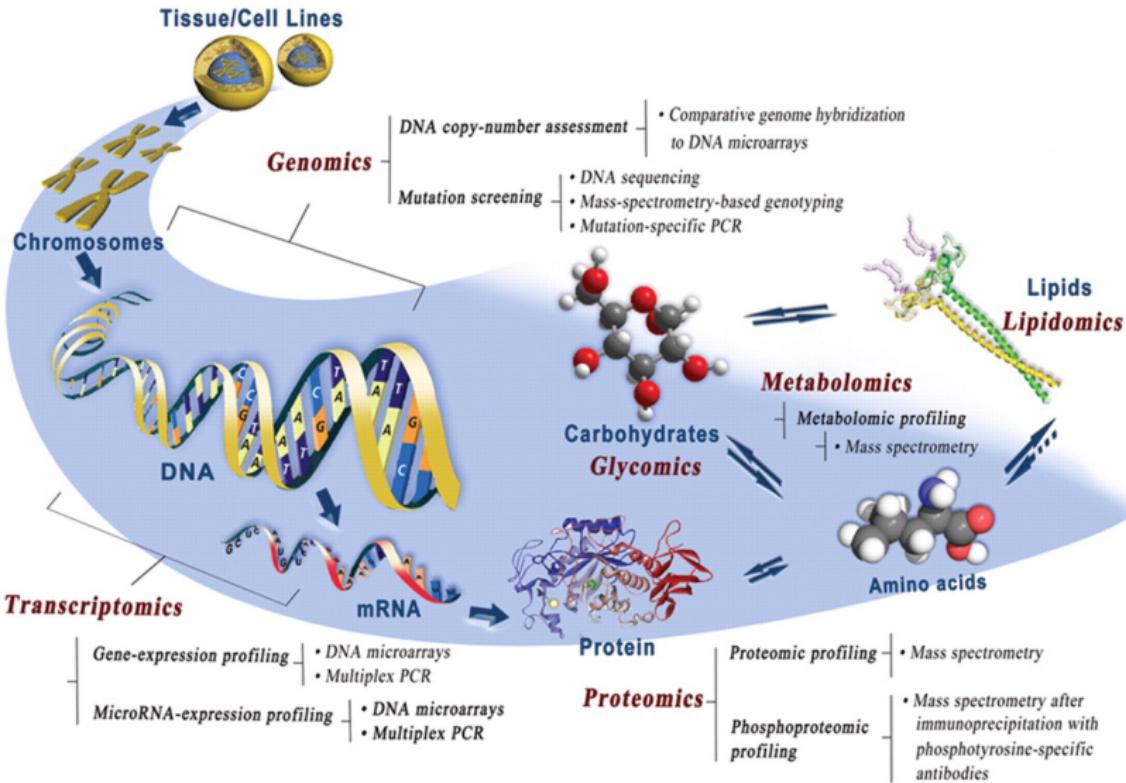
表观遗传学

- 概述
- Methyl-Seq



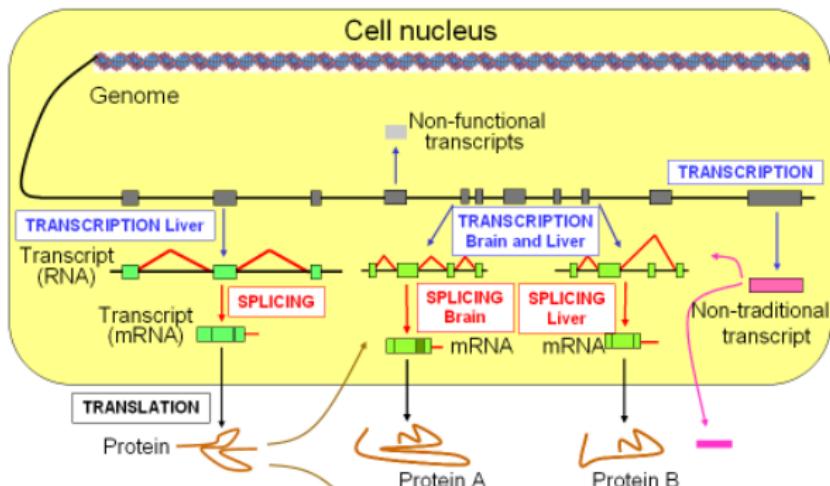


# 转录组学 | 组学



## 转录组

转录组 (transcriptome)，也称为“转录物组”，广义上指在相同环境 (或生理条件) 下的在一个细胞、或一群细胞中所能转录出的所有 RNA 的总和，包括信使 RNA (mRNA)、核糖体 RNA (rRNA)、转运 RNA (tRNA) 及非编码 RNA；狭义上则指细胞所能转录出的所有信使 RNA (mRNA)。



## 转录组学

转录组学（或“转录物组学”，transcriptomics）是分子生物学的分支，负责研究在单个细胞或一个细胞群的特定细胞类型内所生产的 mRNA 分子。

转录组学是对转录水平上发生的事件及其相互关系和意义进行整体研究的一门学科。

## 问题

- 一个细胞、组织或生物体的全部 RNA 集合体中包括多少种 RNA，各种 RNA 的数量有多少？
- 在不同发育时期和不同外界环境作用下 RNA 集合体会出现怎样的变化？
- 在细胞中转录是怎样被调节的？
- .....

## 转录组学

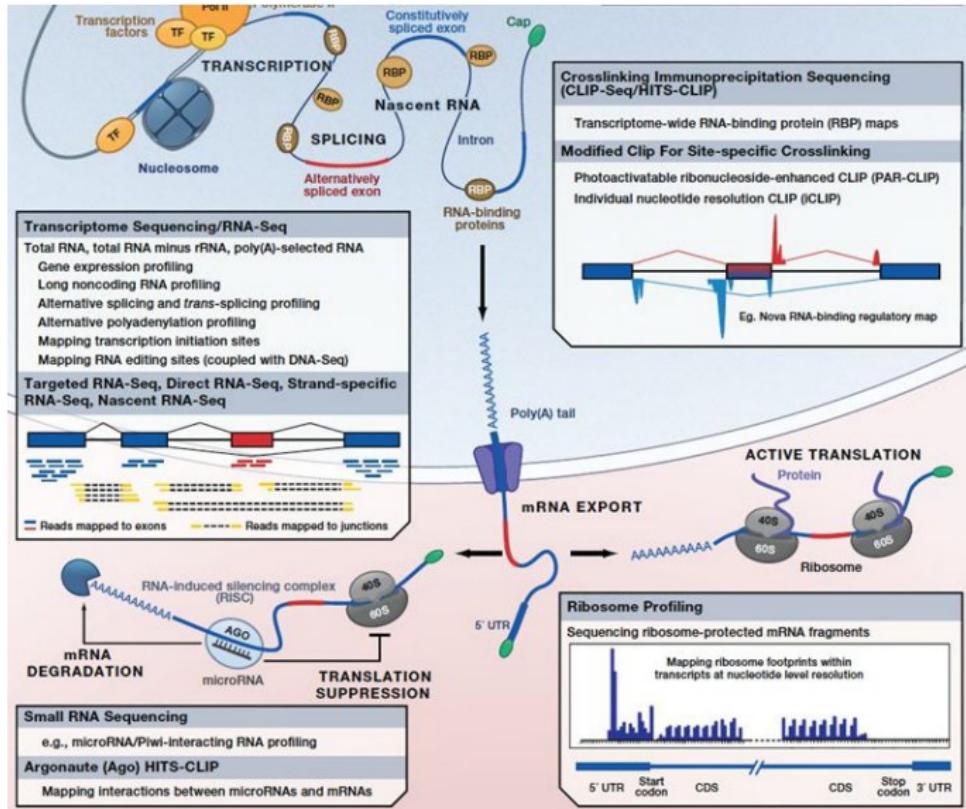
转录组学（或“转录物组学”，transcriptomics）是分子生物学的分支，负责研究在单个细胞或一个细胞群的特定细胞类型内所生产的 mRNA 分子。

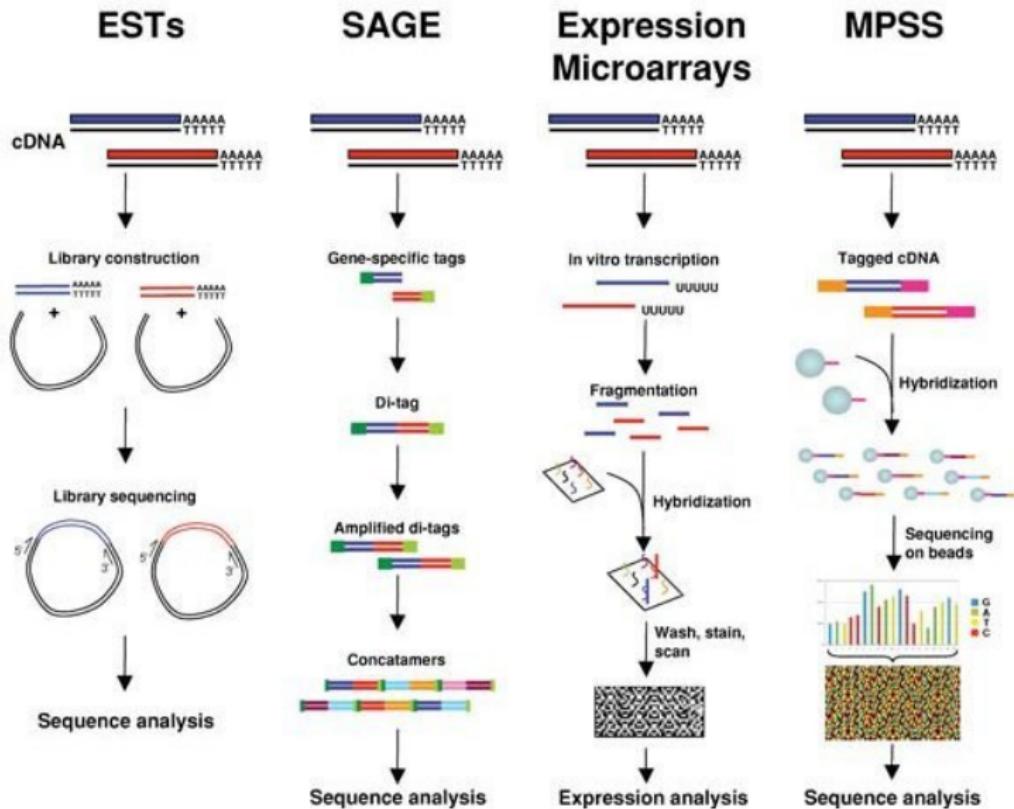
转录组学是对转录水平上发生的事件及其相互关系和意义进行整体研究的一门学科。

## 问题

- 一个细胞、组织或生物体的全部 RNA 集合体中包括多少种 RNA，各种 RNA 的数量有多少？
- 在不同发育时期和不同外界环境作用下 RNA 集合体会出现怎样的变化？
- 在细胞中转录是怎样被调节的？
- .....

# 转录组学 | 概述 | 转录组学 | 研究内容





# 转录组学 | 研究方法 | 概述

## ADVANTAGES:

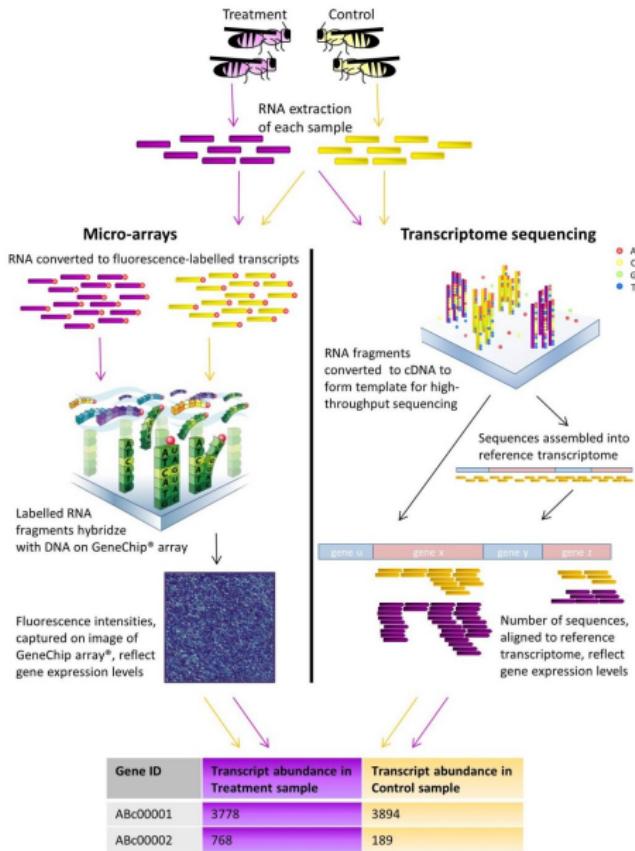
- |   |  |  |   |
|---|--|--|---|
| <ul style="list-style-type: none"><li>• Can detect novel genes and exons</li><li>• No hybridization required</li><li>• High specificity</li></ul> | <ul style="list-style-type: none"><li>• Can detect novel genes and exons</li><li>• No hybridization required</li></ul> | <ul style="list-style-type: none"><li>• Powerful method to find specific sequences</li><li>• Relatively fast and inexpensive</li></ul> | <ul style="list-style-type: none"><li>• Can detect novel genes and exons</li><li>• Tags are longer and more unique</li><li>• Identifies genes with lower expression levels</li><li>• Creates digital data that is easy to share and compare</li></ul> |
|---|--|--|---|

## DISADVANTAGES:

- |   |  |  |   |
|---|--|--|---|
| <ul style="list-style-type: none"><li>• Can not detect genes with low expression levels</li></ul> | <ul style="list-style-type: none"><li>• Costly and time-consuming</li><li>• Ambiguous tag assignment</li></ul> | <ul style="list-style-type: none"><li>• Can not detect novel genes</li><li>• Requires hybridization - false positives and negatives</li><li>• Difficult to compare data from different platforms</li></ul> | <ul style="list-style-type: none"><li>• Costly and time-consuming</li></ul> |
|---|--|--|---|



# 转录组学 | 研究方法 | 概述



## Choose the right technology

RNA-seq	Microarray
Identification of novel genes, transcripts & exons	Well validated QC and analysis methods
Greater dynamic range	Well characterized biases
Less bias due to genetic variation	Quick turnaround from established core facilities
Repeatable	Currently less expensive (for model organisms)
No species-specific primer/probe design	
More accurate relative to qPCR	
Many more applications	



# 教学提纲

1

## 系统生物学

2

## 基因组学

3

## 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

## 数据库与数据格式

5

## 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

## 外显子组测序

## ● 简介

## ● 操作流程

## ● 应用实例

## 7 转录组学

## 8 RNA-Seq

## ● 概述

## ● 数据分析

## ● 应用实例

## 9 顺反组

## ● 概述

## ● ChIP-Seq

## 10 表观遗传学

## ● 概述

## ● Methyl-Seq



# 教学提纲

1

## 系统生物学

2

## 基因组学

3

## 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

## 数据库与数据格式

5

## 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

## 外显子组测序

## ● 简介

## ● 操作流程

## ● 应用实例

## 7 转录组学

## 8 RNA-Seq

### ● 概述

## ● 数据分析

## ● 应用实例

## 9 顺反组

### ● 概述

## ● ChIP-Seq

## 10 表观遗传学

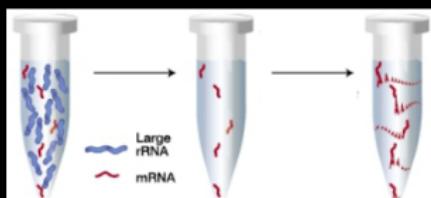
### ● 概述

## ● Methyl-Seq



# RNA

- RNA in cells consists of
  - 95% ribosomal rRNA and tRNA
  - other non-coding ncRNA
  - protein coding mRNA
- Sequence is transcribed from genome but
  - Introns spliced out
  - mRNA is polyadenylated ("A"s added to end)



## RNA-Seq

RNA 测序 (RNA sequencing, 简称 RNA-Seq, 也被称为全转录物组鸟枪法测序, Whole Transcriptome Shotgun Sequencing, 简称 WTSS) 是基于第二代测序技术的转录组学研究方法。RNA 测序是使用第二代测序的能力, 在给定时刻从一个基因组中, 揭示 RNA 的存在和数量的一个快照的技术。

RNA-seq (RNA sequencing), also called whole transcriptome shotgun sequencing (WTSS), uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment in time.



## poly(A)-poly(T)

Frequently, in mRNA analysis the 3' polyadenylated (poly(A)) tail is targeted in order to ensure that coding RNA is separated from noncoding RNA. This can be accomplished simply with poly (T) oligos covalently attached to a given substrate. Presently many studies utilize magnetic beads for this step.

## poly(T) & rRNA

Studies including portions of the transcriptome outside poly(A) RNAs have shown that when using poly(T) magnetic beads, the flow-through RNA (non-poly(A) RNA) can yield important noncoding RNA gene discovery which would have otherwise gone unnoticed.

Also, since ribosomal RNA represents over 90% of the RNA within a given cell, studies have shown that its removal via probe hybridization increases the capacity to retrieve data from the remaining portion of the transcriptome.

## poly(A)-poly(T)

Frequently, in mRNA analysis the 3' polyadenylated (poly(A)) tail is targeted in order to ensure that coding RNA is separated from noncoding RNA. This can be accomplished simply with poly (T) oligos covalently attached to a given substrate. Presently many studies utilize magnetic beads for this step.

## poly(T) & rRNA

Studies including portions of the transcriptome outside poly(A) RNAs have shown that when using poly(T) magnetic beads, the flow-through RNA (non-poly(A) RNA) can yield important noncoding RNA gene discovery which would have otherwise gone unnoticed.

Also, since ribosomal RNA represents over 90% of the RNA within a given cell, studies have shown that its removal via probe hybridization increases the capacity to retrieve data from the remaining portion of the transcriptome.

## Two methods

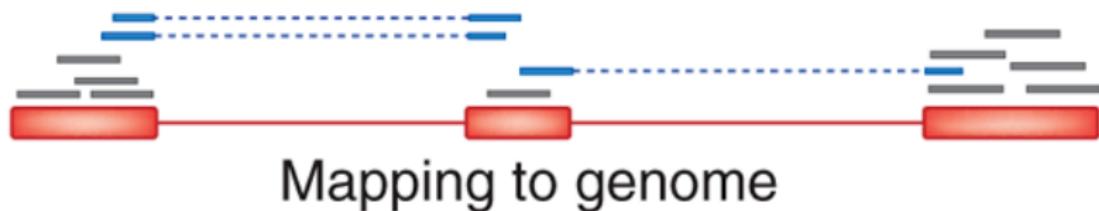
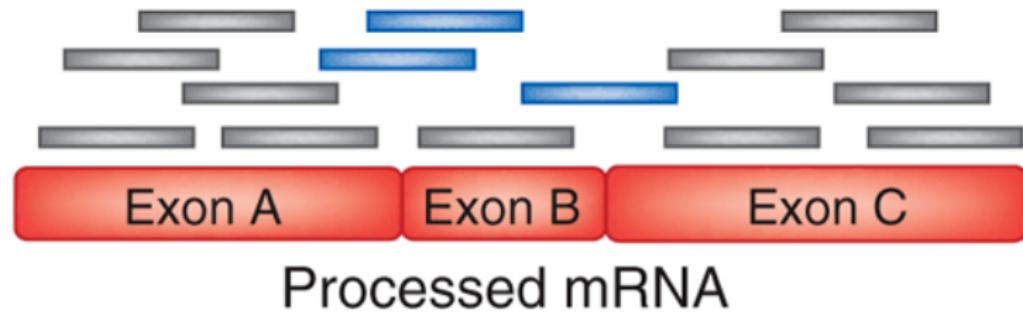
Two different assembly methods are used for producing a transcriptome from raw sequence reads: genome-guided and *de-novo*.



# Transcriptome Assembly

- RNA-Seq
  - Reference genome
  - Reference transcriptome
- RNA-Seq
  - Reference genome
  - No reference transcriptome
- RNA-Seq
  - No reference genome
  - No reference transcriptome





# 转录组学 | RNA-Seq | Method | Experimental considerations

## pros & cons

- Tissue specificity: Gene expression is not uniform throughout an organism's cells, it is strongly dependent on the tissue type being measured. RNA-Seq can provide a complete snapshot of all the transcripts being available at that precise moment in the cell.
- Time dependent: During a cell's lifetime and context, its gene expression levels change. Any single sequencing experiment will offer information regarding one point in time.
- Coverage: coverage/depth can affect the mutations seen.
- Subjectivity of the analysis: Numerous attempts have been taken to uniformly analyze the data. However, the results can vary due to the multitude of algorithms and pipelines available.
- Data management: The main issue with NGS data is the volume of data produced.
- Downstream interpretation of the data: Different layers of interpretations have to be considered when analyzing RNA-Seq data.

- RNA-Seq
  - Transcriptome assembly
    - Qualitative identification of expressed sequence
  - Differential expression analysis
    - Quantitative measurement of transcript expression
- RNA-Seq Applications
  - **Annotation:** Identify novel genes, transcripts, exons, splicing events, ncRNAs
  - Detecting RNA editing and SNPs
  - **Measurements:** RNA quantification and differential gene expression



## Why do an RNA-seq experiment?

Detect differential expression

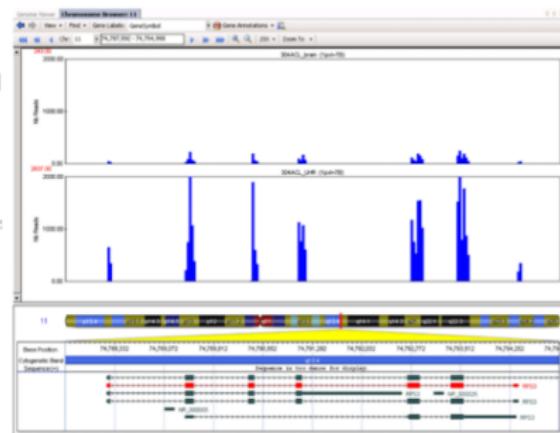
Assess allele-specific expression

Quantify alternative transcript usage

Discover novel genes/transcripts, gene fusions, circRNA

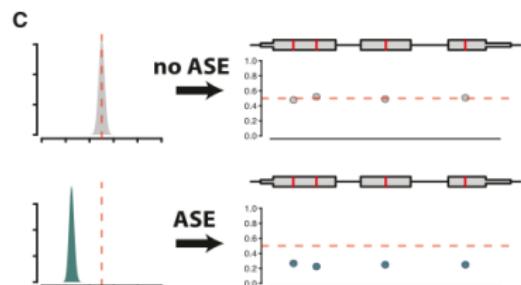
Profile transcriptome

Ribosome profiling to measure translation



## Why do an RNA-seq experiment?

- Detect differential expression
- Assess allele-specific expression
- Quantify alternative transcript usage
- Discover novel genes/transcripts, gene fusions, circRNA
- Profile transcriptome
- Ribosome profiling to measure translation



## Why do an RNA-seq experiment?

Detect differential expression

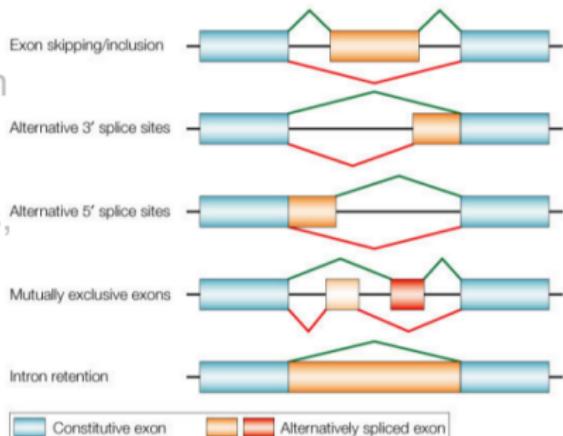
Assess allele-specific expression

**Quantify alternative transcript usage**

Discover novel genes/transcripts,  
gene fusions, circRNA

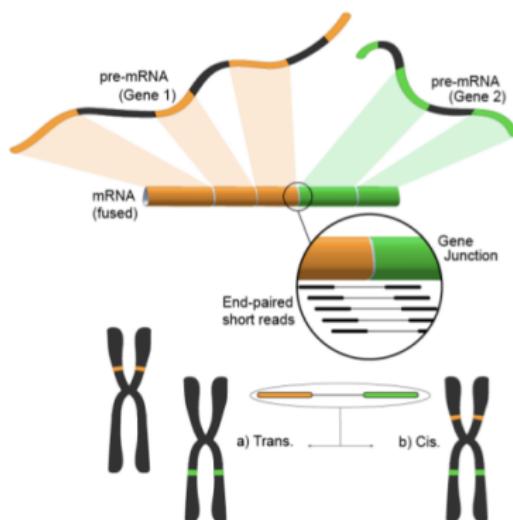
Profile transcriptome

Ribosome profiling to measure  
translation



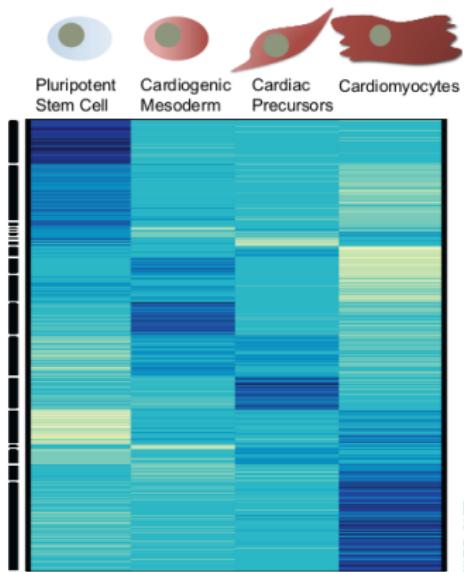
## Why do an RNA-seq experiment?

- Detect differential expression
- Assess allele-specific expression
- Quantify alternative transcript usage
- Discover novel genes/transcripts, gene fusions, circRNA
- Profile transcriptome
- Ribosome profiling to measure translation



## Why do an RNA-seq experiment?

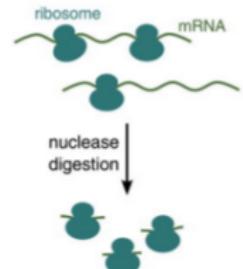
- Detect differential expression
- Assess allele-specific expression
- Quantify alternative transcript usage
- Discover novel genes/transcripts, gene fusions, circRNA
- Profile transcriptome**
- Ribosome profiling to measure translation



## Why do an RNA-seq experiment?

- Detect differential expression
- Assess allele-specific expression
- Quantify alternative transcript usage
- Discover novel genes/transcripts, gene fusions, circRNA
- Profile transcriptome
- Ribosome profiling to measure translation (Ribo-seq)

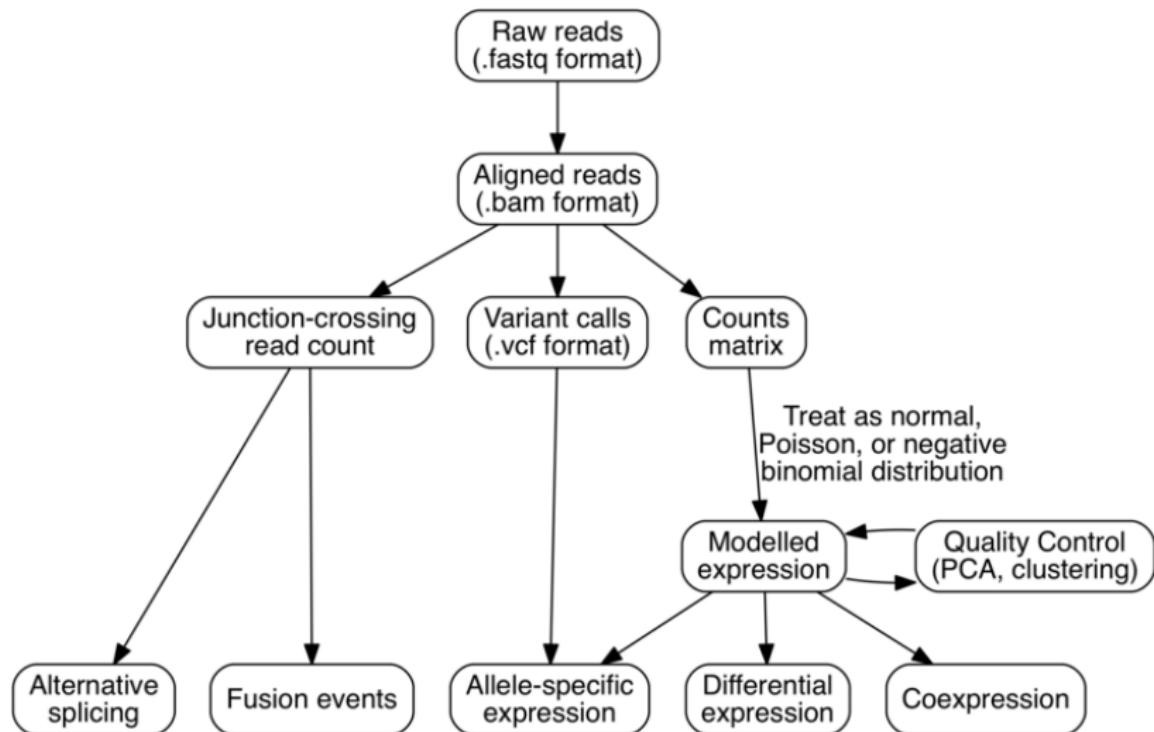
### A Ribosome footprinting



# Experimental design

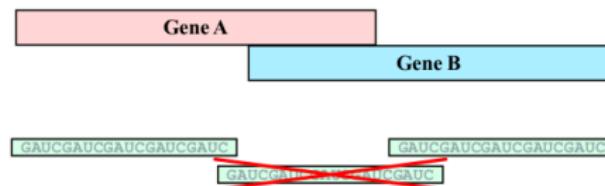
- What are my goals?
  - Transcriptome assembly?
  - Differential expression analysis?
  - Identify rare transcripts?
- What are the characteristics of my system?
  - Large, complex genome?
  - Introns and high degree of alternative splicing?
  - No reference genome or transcriptome?





## Measure expression levels in RNA-Seq data

- 1 Align read to reference genome
- 2 Measuring expression = counting aligned reads
  - ▶ Count in annotated exons
  - ▶ Positive integers (read counts of 3.1415 or  $-42$  are impossible)
  - ▶ Quantitative (read count has an absolute meaning)
- ▶ Observation (read count) must be statistically independent
  - ▶ No multi-map reads
  - ▶ Skip overlapping gene annotations

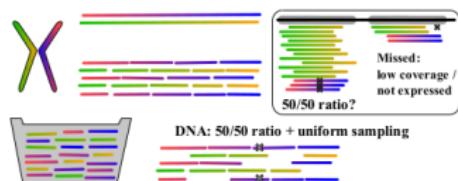


## Differential gene expression analysis tools

- ▶ Alignment
  - ▶ TopHat [15, 5]
  - ▶ STAR [3]
  - ▶ ... many many more
- ▶ Measuring expression (quantification)
  - ▶ HTSeq-count [2]
  - ▶ Cufflinks [16]
  - ▶ featureCounts [8]
- ▶ Group-wise comparison (hypothesis testing)
  - ▶ EdgeR [12]
  - ▶ DESeq2 [11]
  - ▶ Cuffdiff [14]



## Single Nucleotide Polymorphisms in RNA-Seq



- ▶ Major difference(s) between DNA-Seq:
  - ▶ Detected SNPs are expressed
    - ▶ Biological context
    - ▶ SNPs RNA-Seq only within exons and ncRNAs
    - ▶ Allele specific expression profiles
- ▶ Detection:
  - ▶ Expression affects coverage; in DNA-seq coverage should be uniform



# Single Nucleotide Polymorphisms in RNA-Seq

## Using: Samtools, VarScan

	A	C	T	G	A	
read1	a	c	c	g	c	
read2	a	c	t	g	a	
read3	a	c	c	g	a	
read4	a	c	c	g	a	
read5	a	c	t	a	a	
read6		c	c	g	a	
read7			c	g	a	
read quality (q)	0.99	0.99	0.85	0.8	0.99	
aligned	5	6	7	7	7	
q*aligned	4.95	5.94	5.95	5.6	6.93	
(1-q)*aligned	0.05	0.06	1.05	1.4	0.07	
exp match (abs)	5	6	6	6	7	
exp mismatch (abs)	0	0	1	1	0	
obs match	5	6	2	6	6	
obs mismatch	0	0	5	1	1	
P(obs exp) fisher exact	1.000	1.000	0.049	0.538	0.500	
P < 0.05	REF	REF	SNP	REF	REF	

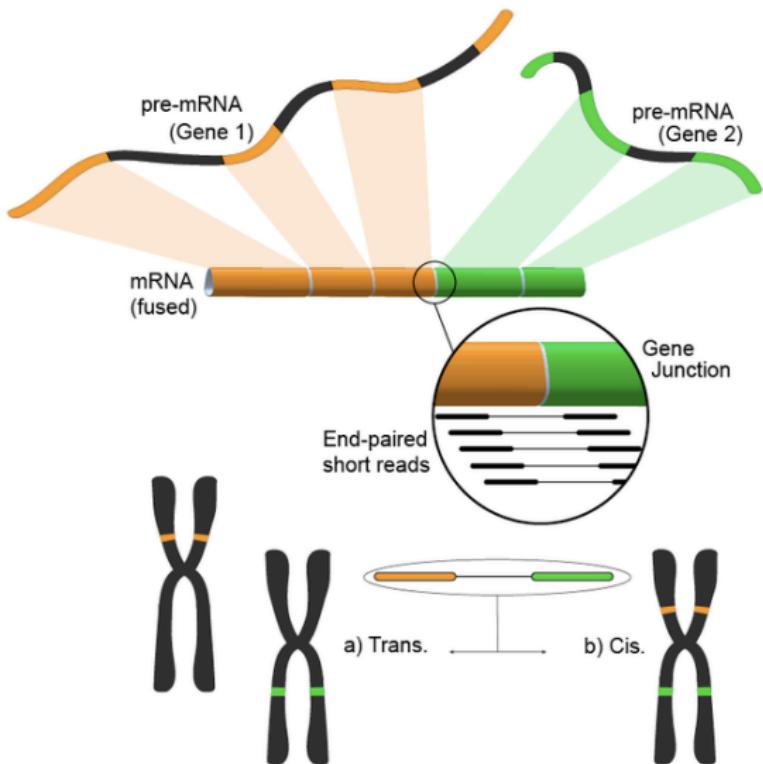


## Single Nucleotide Polymorphisms in RNA-Seq

### Detection tools

- ▶ Alignment
  - ▶ TopHat [15, 5]
  - ▶ STAR [3]
  - ▶ ... many many more
- ▶ SNV calling
  - ▶ VarScan2 [6]
  - ▶ samtools [7]
  - ▶ exactSNP *(part of subread [9] package)*
  - ▶ GATK [17]





## ENCODE

ENCODE aimed to identify genome-wide regulatory regions in different cohort of cell lines and transcriptomic data are paramount in order to understand the downstream effect of those epigenetic and genetic regulatory layers.

## RNA Seq Atlas

RNA-Seq Atlas is a web-based repository of RNA-Seq gene expression profiles and query tools. The website offers free and easy access to RNA-Seq gene expression profiles and tools to both compare tissues and find genes with specific expression patterns.

## TCGA

TCGA aimed to collect and analyze thousands of patient's samples from 30 different tumor types in order to understand the underlying mechanisms of malignant transformation and progression.

## ENCODE

ENCODE aimed to identify genome-wide regulatory regions in different cohort of cell lines and transcriptomic data are paramount in order to understand the downstream effect of those epigenetic and genetic regulatory layers.

## RNA Seq Atlas

RNA-Seq Atlas is a web-based repository of RNA-Seq gene expression profiles and query tools. The website offers free and easy access to RNA-Seq gene expression profiles and tools to both compare tissues and find genes with specific expression patterns.

## TCGA

TCGA aimed to collect and analyze thousands of patient's samples from 30 different tumor types in order to understand the underlying mechanisms of malignant transformation and progression.

## ENCODE

ENCODE aimed to identify genome-wide regulatory regions in different cohort of cell lines and transcriptomic data are paramount in order to understand the downstream effect of those epigenetic and genetic regulatory layers.

## RNA Seq Atlas

RNA-Seq Atlas is a web-based repository of RNA-Seq gene expression profiles and query tools. The website offers free and easy access to RNA-Seq gene expression profiles and tools to both compare tissues and find genes with specific expression patterns.

## TCGA

TCGA aimed to collect and analyze thousands of patient's samples from 30 different tumor types in order to understand the underlying mechanisms of malignant transformation and progression.

## ENCODE

DNA 元件百科全书 (Encyclopedia of DNA Elements, 简称为 ENCODE 项目) 是一个由美国国家人类基因组研究所 (NHGRI) 在 2003 年 9 月发起的一项公共联合研究项目，旨在找出人类基因组中所有功能组件。这是继完成人类基因组计划后国家人类基因组研究所开始的最重要的项目之一。

## 三个阶段

- 试验阶段：测试和比较现有方法以便严格分析一个所定义的人类基因组序列的一部分
- 技术发展阶段：分析整个基因组，并进行“额外中试规模研究”
- 生产阶段：2012 年 9 月 5 日，该项目的初步结果被整理为 30 篇论文并同时发表于多个刊物，包括 6 篇论文在《自然》、6 篇论文在《基因组生物学》及 18 篇论文在《基因组研究》上

## ENCODE

DNA 元件百科全书 (Encyclopedia of DNA Elements, 简称为 ENCODE 项目) 是一个由美国国家人类基因组研究所 (NHGRI) 在 2003 年 9 月发起的一项公共联合研究项目，旨在找出人类基因组中所有功能组件。这是继完成人类基因组计划后国家人类基因组研究所开始的最重要的项目之一。

## 三个阶段

- 试验阶段：测试和比较现有方法以便严格分析一个所定义的人类基因组序列的一部分
- 技术发展阶段：分析整个基因组，并进行“额外中试规模研究”
- 生产阶段：2012 年 9 月 5 日，该项目的初步结果被整理为 30 篇论文并同时发表于多个刊物，包括 6 篇论文在《自然》、6 篇论文在《基因组生物学》及 18 篇论文在《基因组研究》上

# 教学提纲

1

系统生物学

2

基因组学

3

测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

数据库与数据格式

5

二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

外显子组测序

● 简介

● 操作流程

● 应用实例

转录组学

● RNA-Seq

● 概述

● 数据分析

● 应用实例

顺反组

● 概述

● ChIP-Seq

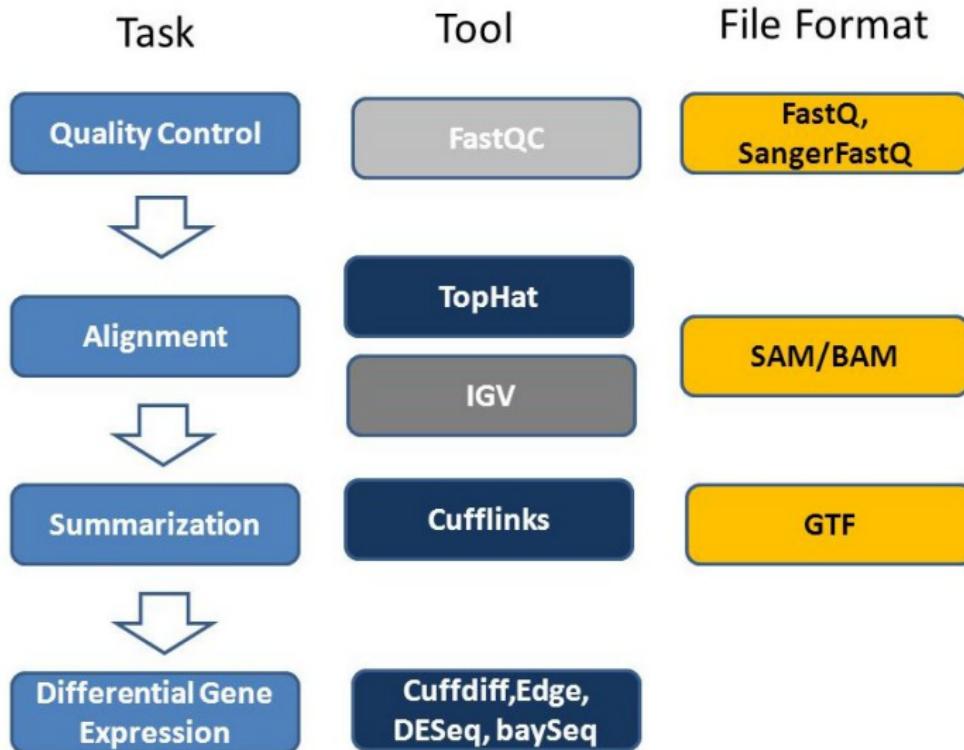
● 表观遗传学

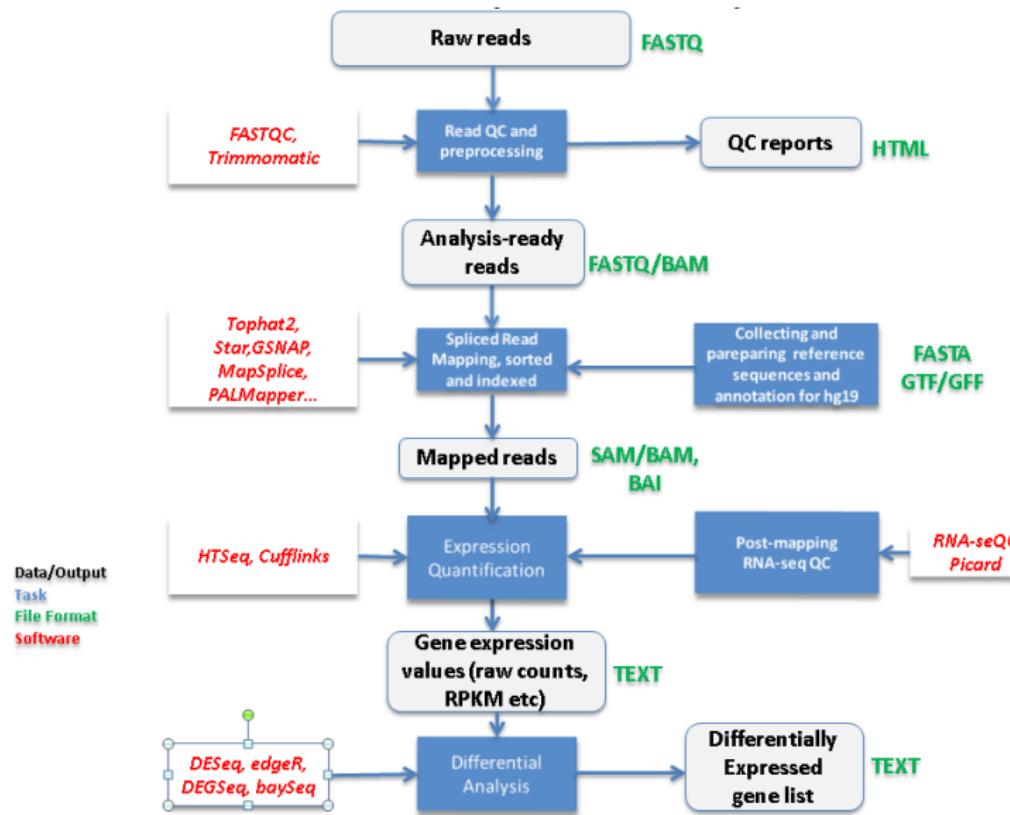
● 概述

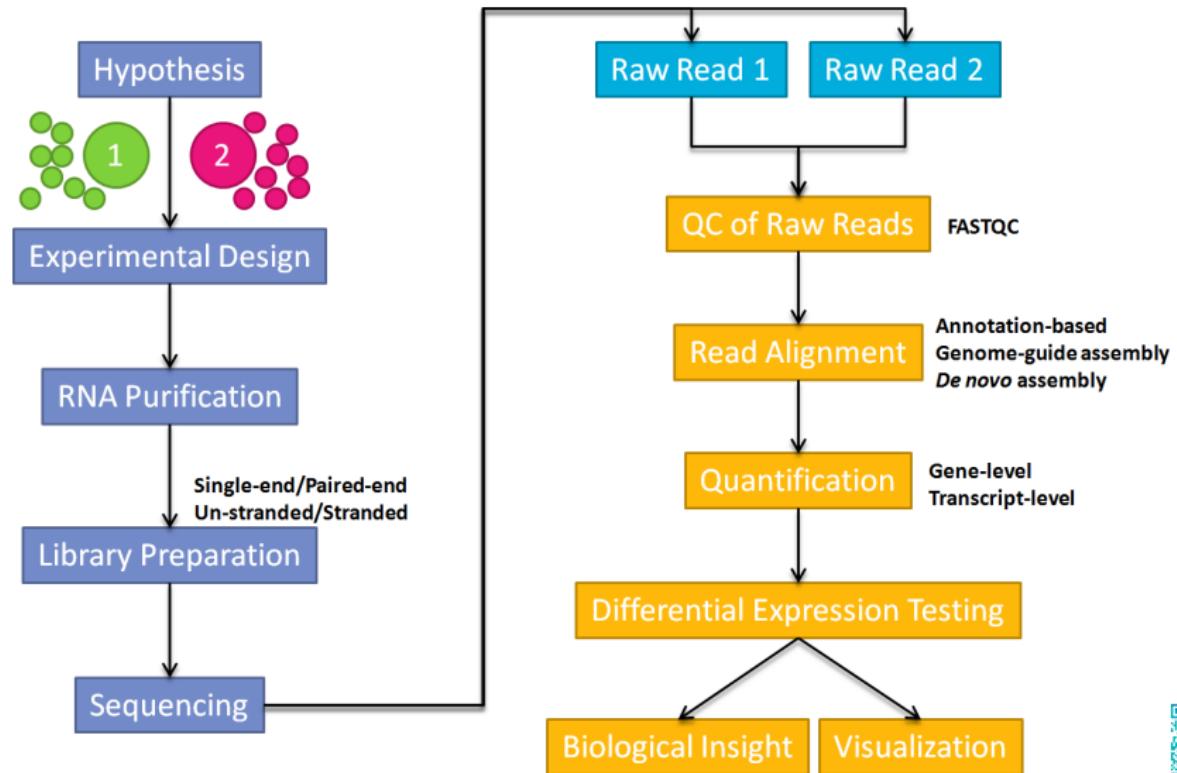
● Methyl-Seq



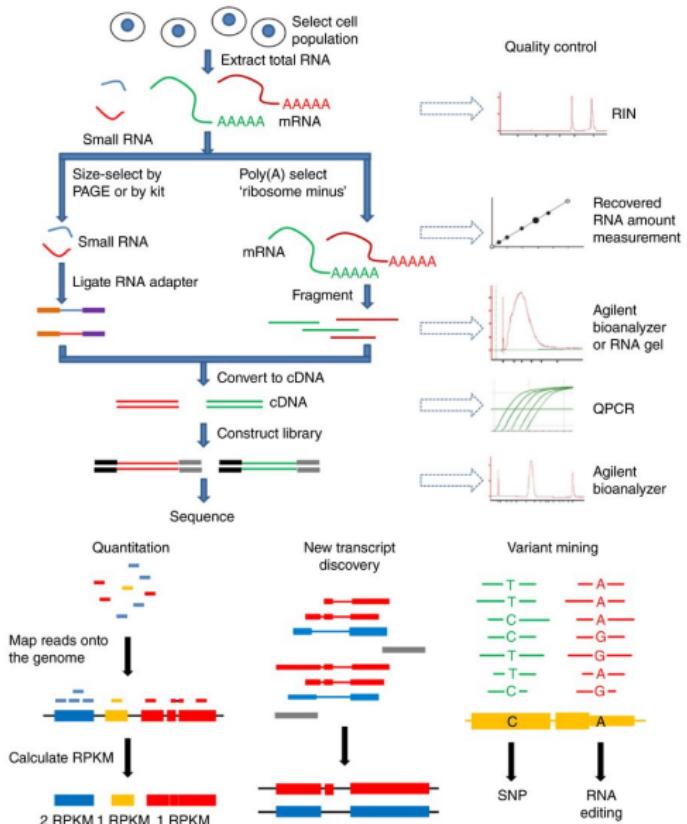
## RNASeq Tasks, Tools and File Formats



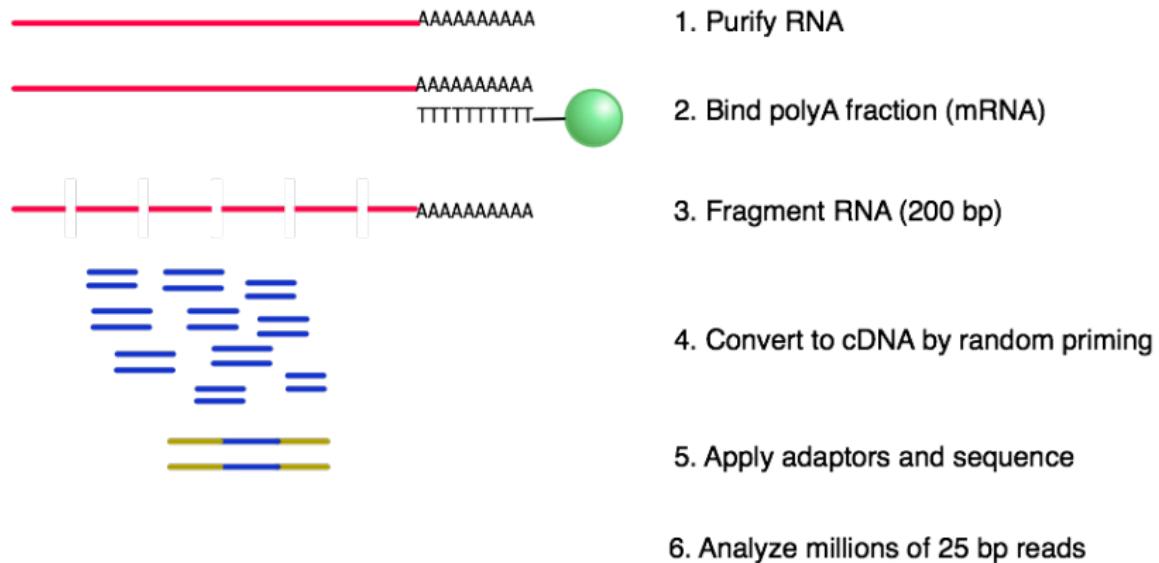




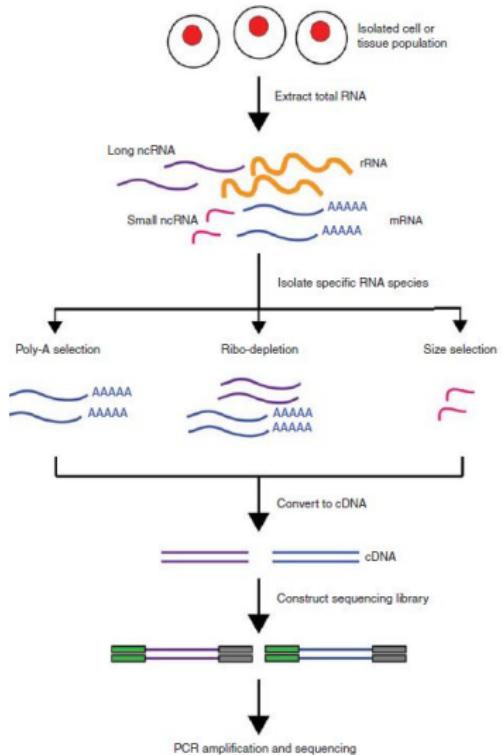
# 转录组学 | RNA-Seq | 分析 | 流程

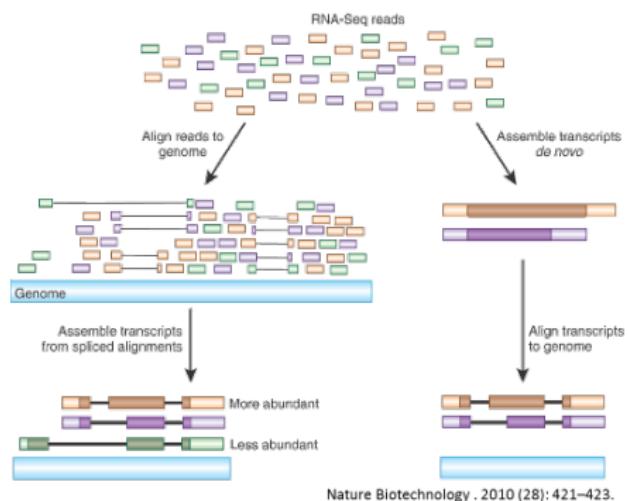


## Steps in Preparing an RNA-Seq Library

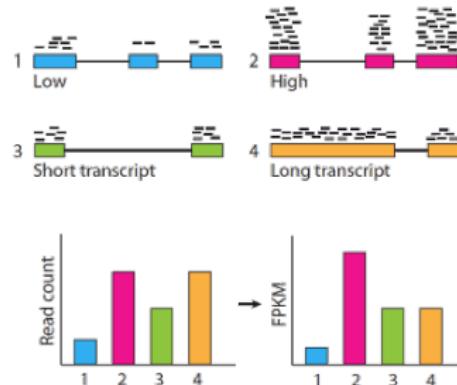


# 转录组学 | RNA-Seq | 分析 | 流程 | 实验



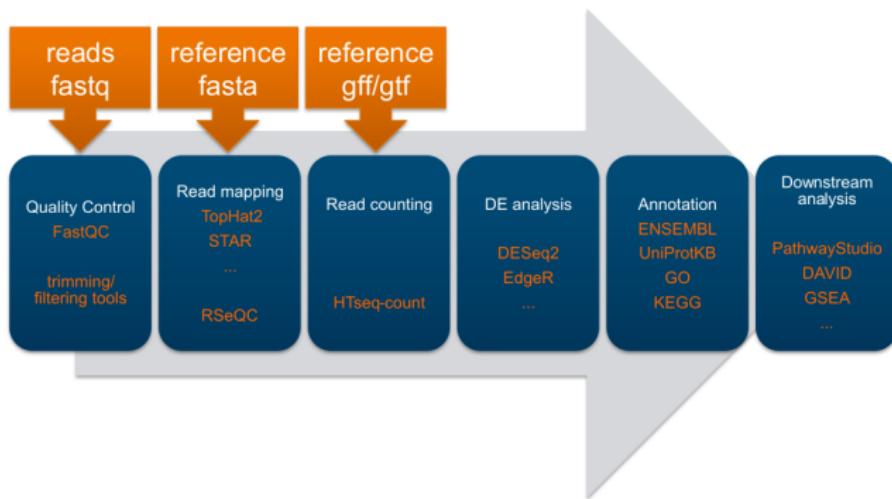
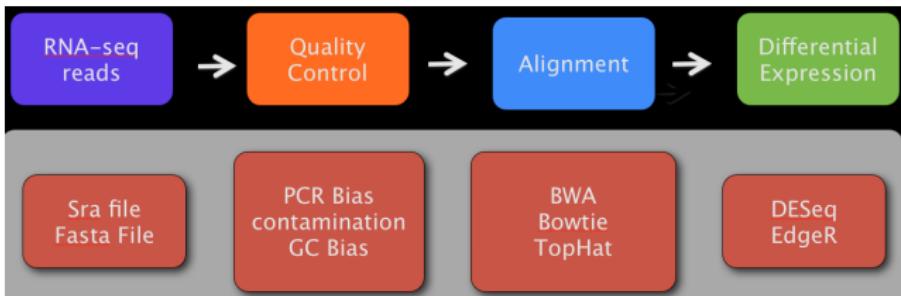


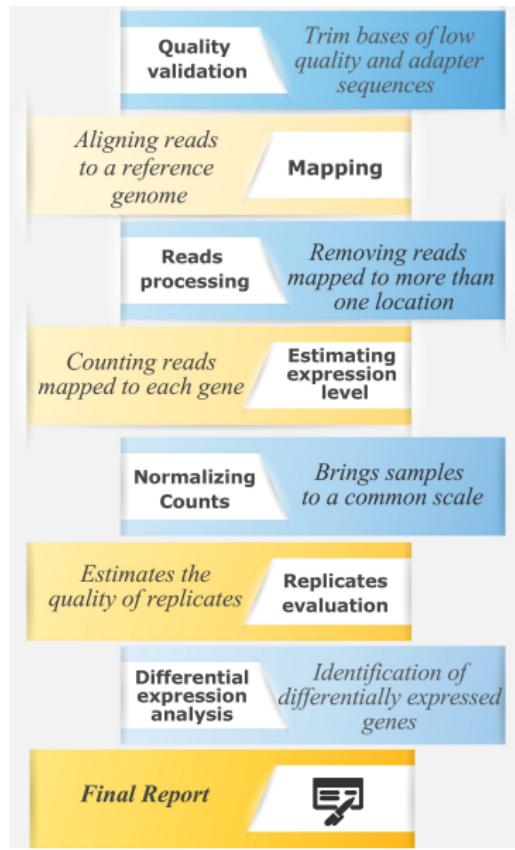
Nature Biotechnology . 2010 (28): 421–423.

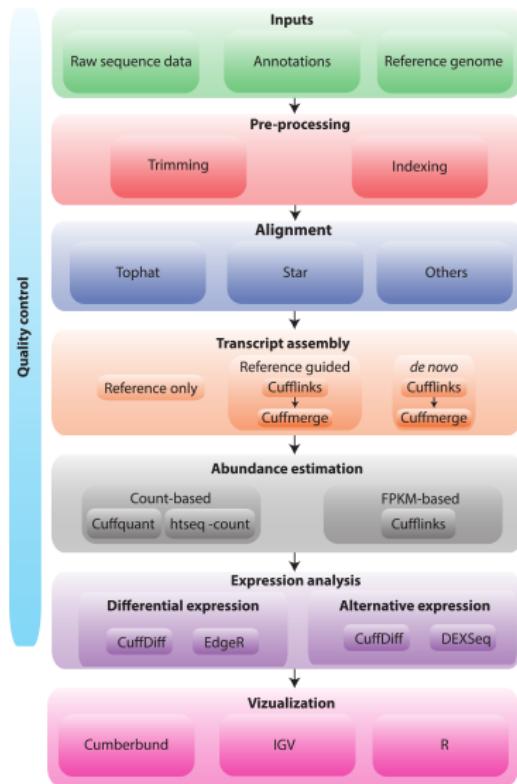


Nature Methods. 2011 (8):469–477.

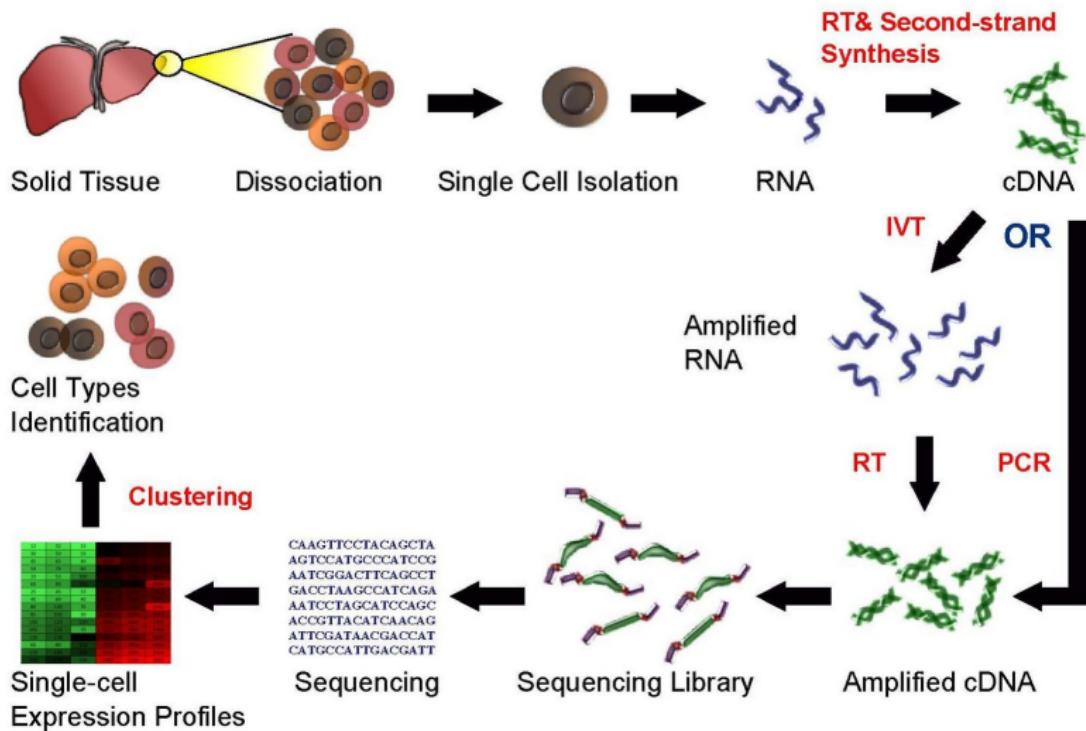




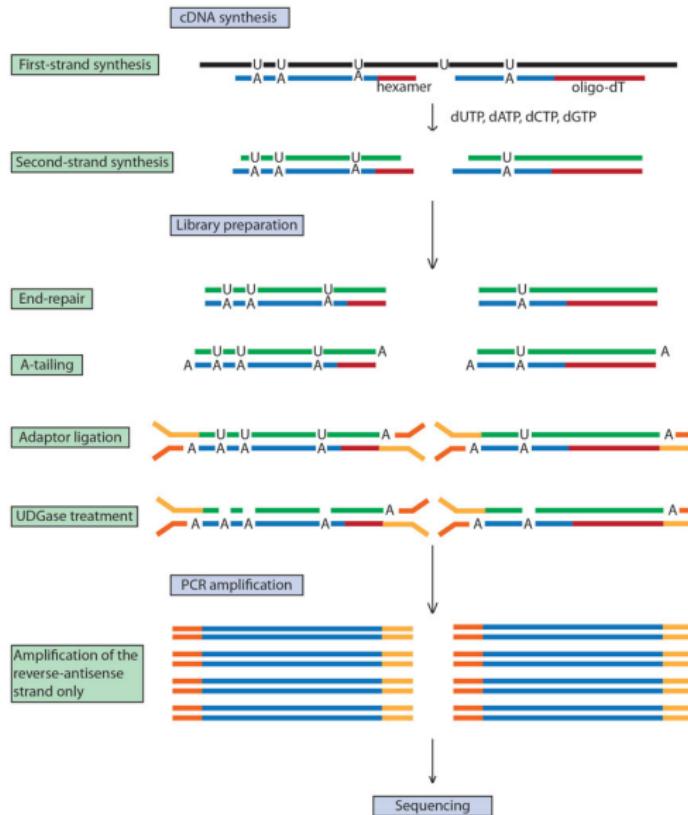


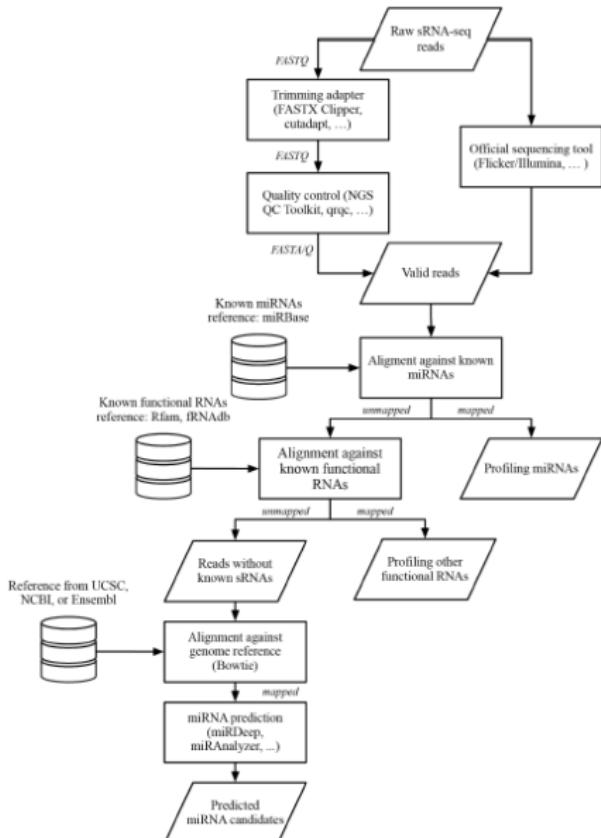


## Single Cell RNA Sequencing Workflow



# 转录组学 | RNA-Seq | 分析 | 流程 | 补遗 | Stranded





## Read counting options

Count per gene

**analysis with DESeq2, edgeR**



Count per transcript

**analysis with CummeRbund, DESeq2, edgeR**

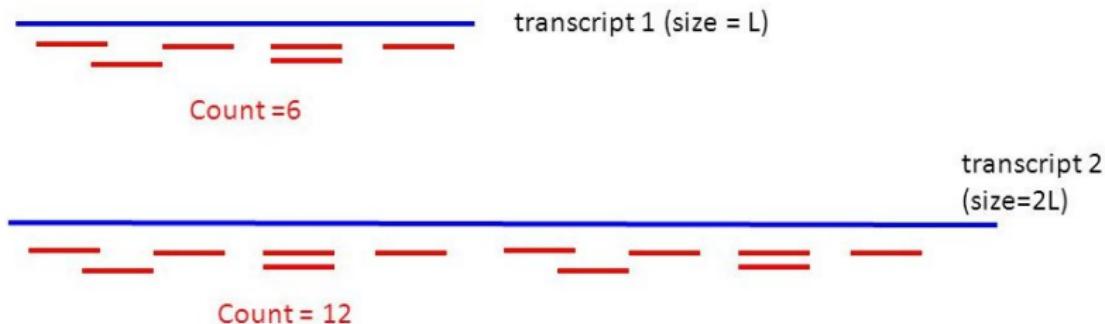
Count per exon

**analysis with DEXSeq**



## Normalization for gene length and library size: RPKM / FPKM

One sample, two transcripts



You can't conclude that **gene 2** has a higher expression than **gene 1**!



## Normalization: the problem

Number of reads (coverage) will not be exactly the same for each sample

**Problem:** Need to scale RNA counts per gene to total sample coverage

- **Solution** – divide counts per million reads

**Problem:** Longer genes have more reads, gives better chance to detect DE

- **Solution** – divide counts by gene length

Result = **RPKM** (Reads Per KB per Million)



## Normalization: classic (used by Cufflinks)

RPKM: Reads per kilobases per million mapped reads

**1 kbp transcript with 2000 alignments in a sample of 10 million reads  
(out of which 8 millions are mapped)**

$$\text{RPKM} = 2000 / (1 * 8) = 250$$

by extension:

FPKM: Fragment per kilobases per million mapped reads

**a fragment is a pair of reads (Paired-end)**



## Normalization: different goals

- **R/FPKM:** (Mortazavi et al. 2008)
  - **Correct for:** differences in sequencing depth and transcript length
  - **Aiming to:** compare a gene across samples and diff genes within sample
- **TMM:** (Robinson and Oshlack 2010)
  - **Correct for:** differences in transcript pool composition; extreme outliers
  - **Aiming to:** provide better across-sample comparability
- **TPM:** (Li et al 2010, Wagner et al 2012)
  - **Correct for:** transcript length distribution in RNA pool
  - **Aiming to:** provide better across-sample comparability
- **Limma voom (logCPM):** (Law et al 2013)
  - **Aiming to:** stabilize variance; remove dependence of variance on the mean



## Other normalization methods

Sequencing depth, gene length, and count distribution are the main biases that must be accounted for in the normalization and/or differential expression calculations.

**Biological Replicates** are essential to increase the robustness of statistics

**FPKM** (Trapnell et al., 2010) - Fragments per Kilobase of exon per Million mapped reads, analogous to RPKM.

**Upper-quartile** (Bullard et al., 2010) - Counts are divided by upper-quartile of counts for transcripts with at least one read.

**TMM (EdgeR)** (Robinson and Oshlack, 2010) - Trimmed mean of M values.

**RLE (DESeq)** (Anders & Huber, 2010) – Relative Log Expression (Median of ratios)

**Quantiles**, as in microarray normalization (Irizarry et al., 2003).

**Many methods which is the best?**



**Table 3:** Summary of comparison results for the seven normalization methods under consideration

Method	Distribution	Intra-Variance	Housekeeping	Clustering	False-positive rate
TC	-	+	+	-	-
UQ	++	++	+	++	-
Med	++	++	-	++	-
<b>DESeq</b>	++	++	++	++	++
<b>TMM</b>	++	++	++	++	++
Q	++	-	+	++	-
RPKM	-	+	+	-	-

A '-' indicates that the method provided unsatisfactory results for the given criterion, while a '+' and '++' indicate satisfactory and very satisfactory results for the given criterion.

Both methods are found in Bioconductor packages (DESeq2 and EdgeR)



# RPKM / FPKM / TPM

- RPKM (Reads per kilobase of transcript per million reads of library)
  - Corrects for total library coverage
  - Corrects for gene length
  - Comparable between different genes within the same dataset
- FPKM (Fragments per kilobase of transcript per million reads of library)
  - Only relevant for paired end libraries
  - Pairs are not independent observations
  - RPKM/2
- TPM (transcripts per million)
  - Normalises to transcript copies instead of reads
  - Corrects for cases where the average transcript length differs between samples



## RPKM:

### Reads Per Kilobase and Million mapped reads

Unit of measurement

$$RPKM = \frac{\# MappedReads * 1000bases * 10^6}{length\ of\ transcript * Total\ number\ of\ mapped\ reads}$$

- RPKM reflects the molar concentration of a transcript in the starting sample by normalizing for
  - RNA length
  - Total read number in the measurement
- This facilitates transparent comparison of transcript levels within and between samples

$$RPKM = \frac{\frac{number\ of\ reads\ of\ the\ region}{total\ reads}}{1,000,000} \times \frac{region\ length}{1,000}$$



## RPKM Example

Gene A 600 bases

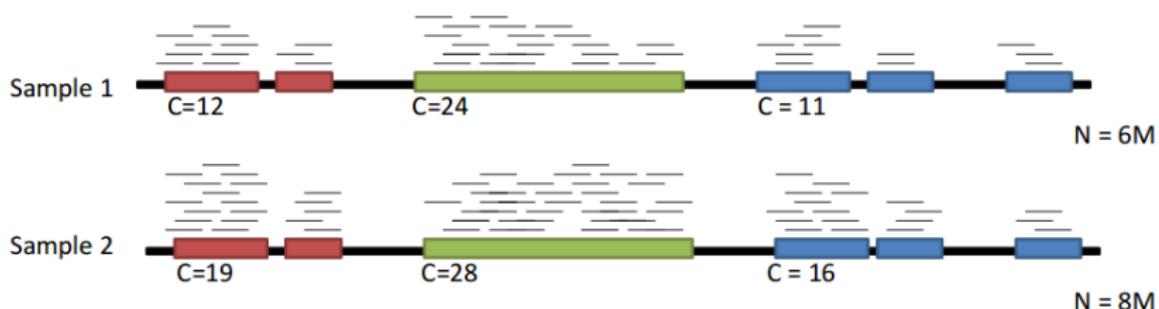
Gene B 1100 bases

Gene C 1400 bases

$$\text{RPKM} = 12/(0.6*6) = 3.33$$

$$\text{RPKM} = 24/(1.1*6) = 3.64$$

$$\text{RPKM} = 11/(1.4*6) = 1.31$$



$$\text{RPKM} = 19/(0.6*8) = 3.96$$

$$\text{RPKM} = 28/(1.1*8) = 1.94$$

$$\text{RPKM} = 16/(1.4*8) = 1.43$$



## Reporting quantitative expression: FPKM/RPKM

- In NGS RNA-seq experiments, quantitative gene expression data is normalized for total gene/transcript length and the number of sequencing reads, and reported as
  - RPKM: Reads Per Kilobase of exon per Million mapped reads. Used for reporting data based on single-end reads
  - FPKM: Fragments Per Kilobase of exon per Million fragments. Used for reporting data based on paired-end fragments



## FPKM: Fragments per K per M

What's the difference between FPKM and RPKM?

- Paired-end RNA-Seq experiments produce two reads per fragment, but that doesn't necessarily mean that both reads will be mappable. For example, the second read is of poor quality.
- If we were to count reads rather than fragments, we might double-count some fragments but not others, leading to a skewed expression value.
- Thus, FPKM is calculated by counting fragments, not reads.

$$\text{RPKM} = \frac{\text{total exon reads}}{\text{mapped reads (millions)} * \text{exon length (KB)}}$$

$$\text{FPKM} = \frac{\text{total fragments}}{\text{mapped reads (millions)} * \text{exon length (KB)}}$$



## TPM

当我们进行 RNA-Seq 时，会使用 RPKM 或者 FPKM 来代表某个 gene 或是 isoform 的表达量多寡。可是当我们想要比较不同次实验内的某个基因，其表达量相比于“整体基因表达”而言，是否维持在“固定比例”时，便无法使用这样的计算方式。因此 Wagner *et. al.* 在 2012 年的时候提出 TPM (Transcript Per Million) 的概念来弥补这个缺点。



## TPM – Transcripts Per Million

Theory Biosci.  
DOI 10.1007/s12064-012-0162-3

SHORT COMMUNICATION

**Measurement of mRNA abundance using RNA-seq data:  
RPKM measure is inconsistent among samples**

Günter P. Wagner · Koryn Kin · Vincent J. Lynch

A slightly modified RPKM measure that  
accounts for differences in gene length  
distribution in the transcript population

TPM

$$\text{TPM} = \frac{\frac{\text{total exon reads}}{\text{exon length (KB)}}}{\left( \frac{\text{GeneA mapped reads (millions)}}{\text{exon length (KB)}} + \frac{\text{GeneB mapped reads (millions)}}{\text{exon length (KB)}} + \frac{\text{GeneC mapped reads (millions)}}{\text{exon length (KB)}} + \dots \right)}$$

$$\text{TPM}_i = \left( \frac{\text{FPKM}_i}{\sum_j \text{FPKM}_j} \right) \cdot 10^6$$



## Metrics

- RPKM (Reads Per Kilobase Million)
- FPKM (Fragments Per Kilobase Million)
- TPM (Transcripts Per Kilobase Million)

These three metrics attempt to normalize for sequencing depth and gene length.



## RPKM

Here's how you do it for RPKM:

- ① Count up the total reads in a sample and divide that number by 1,000,000 – this is our “per million” scaling factor.
- ② Divide the read counts by the “per million” scaling factor. This normalizes for sequencing depth, giving you reads per million (RPM)
- ③ Divide the RPM values by the length of the gene, in kilobases. This gives you RPKM.



## FPKM

FPKM is very similar to RPKM. RPKM was made for single-end RNA-seq, where every read corresponded to a single fragment that was sequenced. FPKM was made for paired-end RNA-seq. With paired-end RNA-seq, two reads can correspond to a single fragment, or, if one read in the pair did not map, one read can correspond to a single fragment. The only difference between RPKM and FPKM is that FPKM takes into account that two reads can map to one fragment (and so it doesn't count this fragment twice).



## TPM

TPM is very similar to RPKM and FPKM. The only difference is the order of operations. Here's how you calculate TPM:

- ➊ Divide the read counts by the length of each gene in kilobases. This gives you reads per kilobase (RPK).
- ➋ Count up all the RPK values in a sample and divide this number by 1,000,000. This is your “per million” scaling factor.
- ➌ Divide the RPK values by the “per million” scaling factor. This gives you TPM.

So you see, when calculating TPM, the only difference is that you normalize for gene length first, and then normalize for sequencing depth second. However, the effects of this difference are quite profound.

## Compare

When you use TPM, the sum of all TPMs in each sample are the same. This makes it easier to compare the proportion of reads that mapped to a gene in each sample. In contrast, with RPKM and FPKM, the sum of the normalized reads in each sample may be different, and this makes it harder to compare samples directly.

Here's an example. If the TPM for gene A in Sample 1 is 3.33 and the TPM in sample B is 3.33, then I know that the exact same proportion of total reads mapped to gene A in both samples. This is because the sum of the TPMs in both samples always add up to the same number (so the denominator required to calculate the proportions is the same, regardless of what sample you are looking at.)

With RPKM or FPKM, the sum of normalized reads in each sample can be different. Thus, if the RPKM for gene A in Sample 1 is 3.33 and the RPKM in Sample 2 is 3.33, I would not know if the same proportion of reads in Sample 1 mapped to gene A as in Sample 2. This is because the denominator required to calculate the proportion could be different for the two samples.



# 转录组学 | RNA-Seq | 分析 | 工具 | 概览

Workflow	Category	Package	Reference
Preprocessing of raw data	Raw data QC	FastQC HTQC	[8] [9]
	Read trimming	FASTX-Toolkit FLEXBAR	[10] [11]
Read alignment	Unspliced aligner	MAQ BWA Bowtie	[13] [14] [15]
	Spliced aligner	TopHat MapSplice STAR GSNAP	[16] [17] [18] [19]
RNA-seq specific quality control		RNA-SeQC RSQC	[20] [21]
		Qualimap 2	[22]
Transcriptome reconstruction	Reference-guided	Cufflinks Scripture StringTie	[24] [25] [26]
	Reference-independent	Trinity Oases	[27] [28]
		transAbYSS	[29]
Expression quantification	Gene-level quantification	ALEXA-seq Enhanced read analysis of gene expression (ERANGE)	[32] [33]
		Normalization by expected uniquely mappable area (NEUMA)	[34]
	Isoform-level quantification	Cufflinks StringTie RSEM	[24] [26] [35]
		Sailfish	[36]
Differential expression	Gene-level	NOIseq edgeR DESeq	[23] [39] [40]
	Isoform-level	SAMseq Cuffdiff EBSeq Ballgown	[41] [24] [42] [45]



## Quality control

- FastQC: FastQC is a quality control tool for high-throughput sequence data (Babraham Institute) and is developed in Java. Import of data is possible from FastQ files, BAM or SAM format. This tool provides an overview to inform about problematic areas, summary graphs and tables to rapid assessment of data. Results are presented in HTML permanent reports. FastQC can be run as a stand-alone application or it can be integrated into a larger pipeline solution.
- NGSQC: cross-platform quality analysis pipeline for deep sequencing data.



## Quality control

- RNA-SeQC: RNA-SeQC is a tool with application in experiment design, process optimization and quality control before computational analysis. Essentially, provides three types of quality control: read counts (such as duplicate reads, mapped reads and mapped unique reads, rRNA reads, transcript-annotated reads, strand specificity), coverage (like mean coverage, mean coefficient of variation, 5'/3' coverage, gaps in coverage, GC bias) and expression correlation (the tool provides RPKM-based estimation of expression levels). RNA-SeQC is implemented in Java and is not required installation, however can be run using the GenePattern web interface. The input could be one or more BAM files. HTML reports are generated as output.
- RSeQC: RSeQC analyzes diverse aspects of RNA-Seq experiments: sequence quality, sequencing depth, strand specificity, GC bias, read distribution over the genome structure and coverage uniformity. The input can be SAM, BAM, FASTA, BED files or Chromosome size file (two-column, plain text file). Visualization can be performed by genome browsers like UCSC, IGB and IGV. However, R scripts can also be used to visualization.

## Trimming and adapters removal

- FASTX: FASTX Toolkit is a set of command line tools to manipulate reads in files FASTA or FASTQ format. These commands make possible preprocess the files before mapping with tools like Bowtie. Some of the tasks allowed are: conversion from FASTQ to FASTA format, information about statistics of quality, removing sequencing adapters, filtering and cutting sequences based on quality or conversion DNA/RNA.
- PRINSEQ: PRINSEQ generates statistics of your sequence data for sequence length, GC content, quality scores, n-plicates, complexity, tag sequences, poly-A/T tails, odds ratios. Filter the data, reformat and trim sequences.
- cutadapt: Cutadapt finds and removes adapter sequences, primers, poly-A tails and other types of unwanted sequence from your high-throughput sequencing reads.



# Read Mapping

- Alignment algorithm must be
  - fast
  - able to handle SNPs, indels, and sequencing errors
  - allow for introns for reference genome alignment
- Input
  - fastq read library
  - reference genome index
  - insert size mean and stddev(for paired-end libraries)
- Output
  - SAM (text) / BAM (binary) alignment files



### *De novo* Splice Aligners that also use annotation optionally

TopHat: TopHat is prepared to find *de novo* junctions. TopHat aligns reads in two steps. Firstly, unspliced reads are aligned with Bowtie. After, the aligned reads are assembled with Maq resulting islands of sequences. Secondly, the splice junctions are determined based on the initially unmapped reads and the possible canonical donor and acceptor sites within the island sequences.

graph-based alignment of next generation sequencing reads to a population of genomes

HISAT2: HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes (as well as to a single reference genome).

### *De novo* Splice Aligners that also use annotation optionally

TopHat: TopHat is prepared to find *de novo* junctions. TopHat aligns reads in two steps. Firstly, unspliced reads are aligned with Bowtie. After, the aligned reads are assembled with Maq resulting islands of sequences. Secondly, the splice junctions are determined based on the initially unmapped reads and the possible canonical donor and acceptor sites within the island sequences.

### graph-based alignment of next generation sequencing reads to a population of genomes

HISAT2: HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes (as well as to a single reference genome).

## Genome-Guided assemblers

- Cufflinks: Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.
- Scripture: Scripture is a method for transcriptome reconstruction that relies solely on RNA-Seq reads and an assembled genome to build a transcriptome *ab initio*. The statistical methods to estimate read coverage significance are also applicable to other sequencing data. Scripture also has modules for ChIP-Seq peak calling.

## Normalization, Quantitative analysis and Differential Expression

- Cufflinks/Cuffdiff: Cufflinks is appropriate to measure global *de novo* transcript isoform expression. It performs assembly of transcripts, estimation of abundances and determines differential expression (Cuffdiff) and regulation in RNA-Seq samples.
- DESeq: DESeq is a Bioconductor package to perform differential gene expression analysis based on negative binomial distribution.
- EdgeR: EdgeR is a R package for analysis of differential expression of data from DNA sequencing methods, like RNA-Seq, SAGE or ChIP-Seq data. edgeR employs statistical methods supported on negative binomial distribution as a model for count variability.
- DEGseq: DEGseq is an R package to identify differentially expressed genes from RNA-Seq data.
- baySeq: This package identifies differential expression in high-throughput ‘count’ data, such as that derived from next-generation sequencing machines, calculating estimated posterior likelihoods of differential expression (or more complex hypotheses) via empirical Bayesian methods.

## Analysis pipeline/Integrated solutions

- easyRNASeq: easyRNASeq calculates the coverage of high-throughput short-reads against a genome of reference and summarizes it per feature of interest (e.g. exon, gene, transcript). The data can be normalized as 'RPKM' or by the 'DESeq' or 'edgeR' package.
- Galaxy: Galaxy is a general purpose workbench platform for computational biology. There are several publicly accessible Galaxy servers that support RNA-Seq tools and workflows, including NBIC's Andromeda, the CBIIT-Giga server, the Galaxy Project's public server, the GeneNetwork Galaxy server, the University of Oslo's Genomic Hyperbrowser, URGI's server (which supports S-MART), and many others.
- GenePattern: GenePattern offers integrated solutions to RNA-Seq analysis.
- Taverna: Taverna is an open source and domain-independent Workflow Management System –a suite of tools used to design and execute scientific workflows and aid *in silico* experimentation.

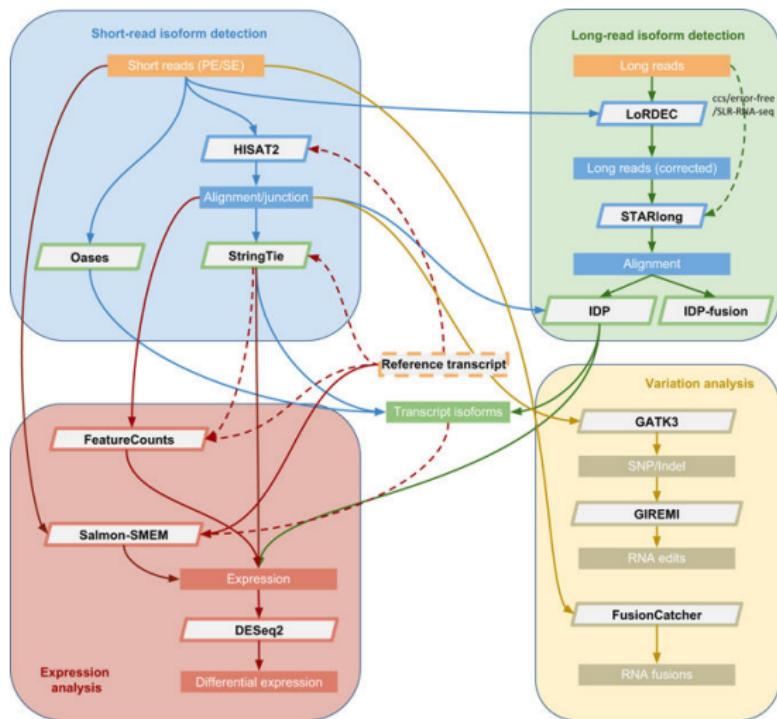
## RNA Cocktail

The RNA Cocktail pipeline is composed of a high-accuracy tools for different steps of RNA-Seq analysis. It performs a broad spectrum RNA-Seq analysis on both short- and long-read technologies to enable meaningful insights from transcriptomic data. It was developed after analyzing a variety of RNA-Seq samples (ranging from germline, cancer to stem cell datasets) and technologies using a multitude of tool combinations to determine a pipeline which is comprehensive, fast and accurate. RNA Cocktail supports:

short-read	long-read
alignment	error correction
transcriptome reconstruction	alignment
denovo transcriptome assembly	transcriptome reconstruction
alignment-free quantification	fusion prediction
differential expression analysis	—
fusion prediction	—
variant calling	—
RNA editing prediction	—



# 转录组学 | RNA-Seq | 分析 | 工具 | Framework | RNA Cocktail



## Visualization tools

- ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. It allows to easily visualize NGS samples at functional genomic regions.
- GBrowse: GBrowse is a combination of database and interactive web pages for manipulating and displaying annotations on genomes.
- IGB: The Integrated Genome Browser (IGB, pronounced ig-bee) is an application intended for visualization and exploration of genomes and corresponding annotations from multiple data sources.
- IGV: The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets.
- SeqMonk: SeqMonk is a program to enable the visualisation and analysis of mapped sequence data. It was written for use with mapped next generation sequence data but can in theory be used for any dataset which can be expressed as a series of genomic positions.
- Tablet: Tablet is a lightweight, high-performance graphical viewer for next generation sequence assemblies and alignments.

## RNA-Seq Databases

- ENCODE: Encyclopedia of DNA Elements
- GTEx: Genotype-Tissue Expression (GTEx) project
- RNA-Seq Atlas: A reference database for gene expression profiling in normal tissue by next-generation sequencing.
- SRA: The Sequence Read Archive (SRA) stores raw sequence data from “next-generation” sequencing technologies including 454, IonTorrent, Illumina, SOLiD, Helicos and Complete Genomics. In addition to raw sequence data, SRA now stores alignment information in the form of read placements on a reference sequence.



## Tuxedo

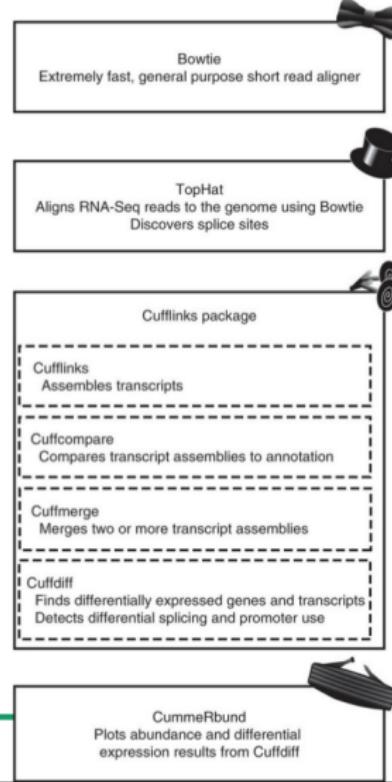
The RNA-seq pipeline “Tuxedo” consists of the **TopHat** spliced read mapper, that internally uses **Bowtie/Bowtie2** short read aligners, and several **Cufflinks** tools that allows one to assemble transcripts, estimate their abundances, and tests for differential expression and regulation in RNA-Seq samples.

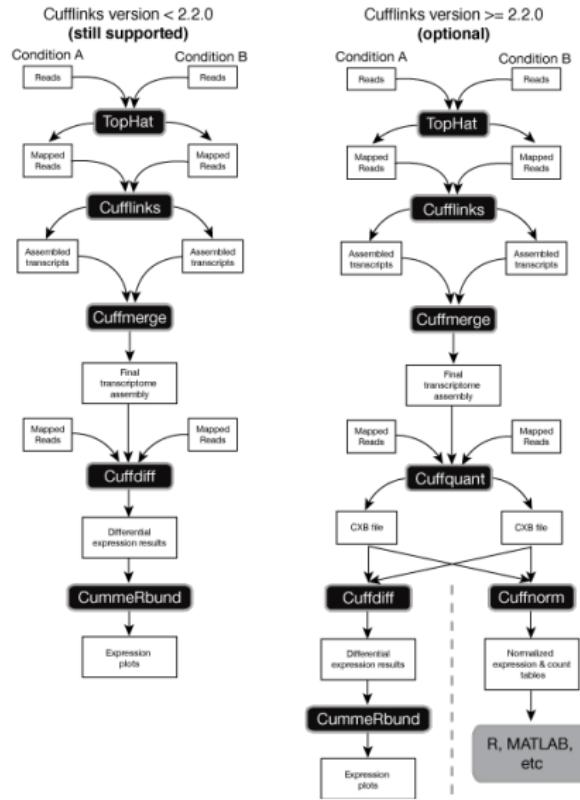
Tool	Tool description	
Bowtie	Ultrafast short read aligner	
Tophat	Aligns RNA-seq reads to the genome using Bowtie. Discovers splice sites	
Cufflinks package	Cufflinks	Assembles transcripts
	Cuffcompare	Compares transcript assemblies to annotation
	Cuffmerge	Merges two or more transcript assemblies
	Cuffdiff	Finds differentially expressed genes and transcripts. Detects differential splicing and promoter use

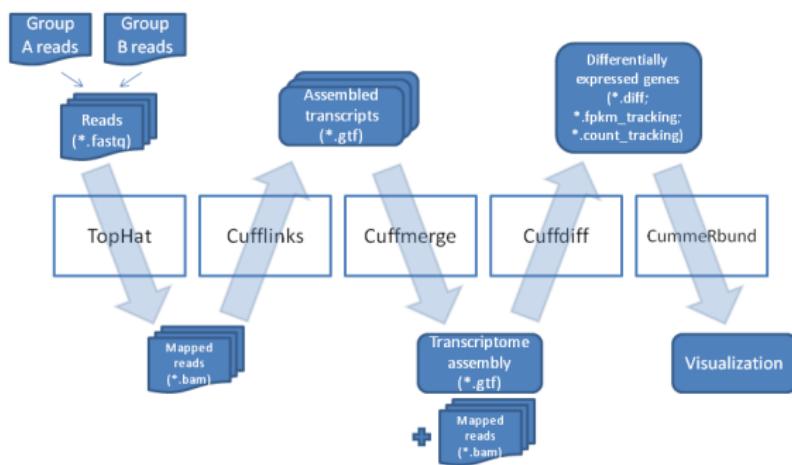
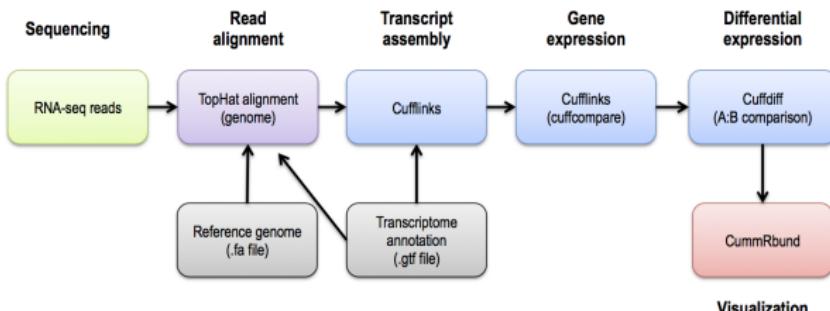


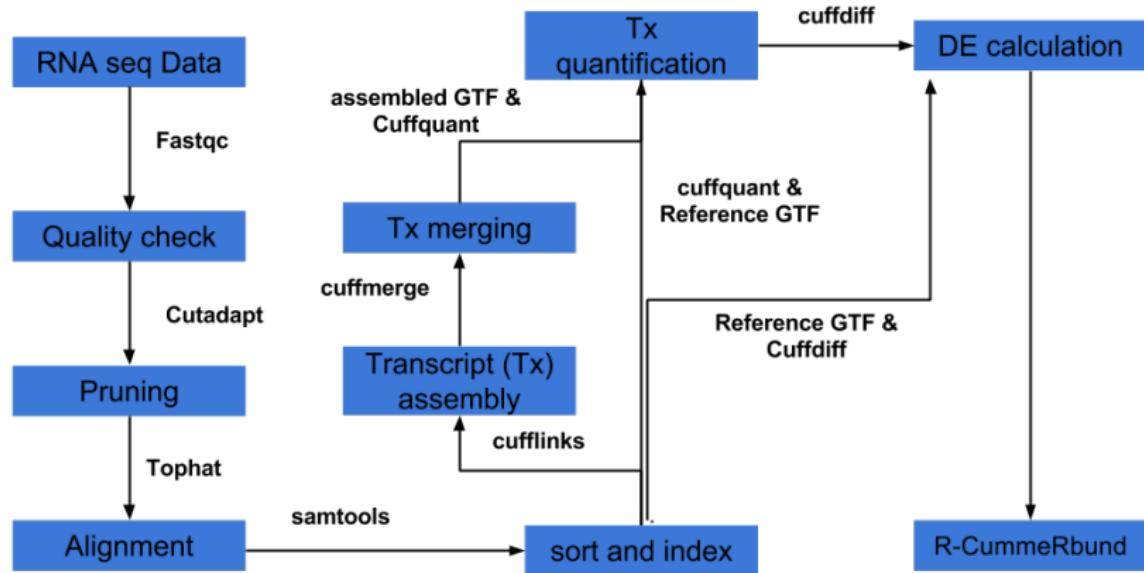
## The Tuxedo suite a complete solution (not the best anymore)

Bowtie  
TopHat  
Cufflinks  
**Cuffmerge**  
**Cuffdiff**  
CummRbund









## Bowtie: ultrafast short read alignment

Bowtie is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

## TopHat: alignment of short RNA-Seq reads

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

## Bowtie: ultrafast short read alignment

Bowtie is an ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences. It is particularly good at aligning reads of about 50 up to 100s or 1,000s of characters, and particularly good at aligning to relatively long (e.g. mammalian) genomes. Bowtie 2 indexes the genome with an FM Index to keep its memory footprint small: for the human genome, its memory footprint is typically around 3.2 GB. Bowtie 2 supports gapped, local, and paired-end alignment modes.

## TopHat: alignment of short RNA-Seq reads

TopHat is a fast splice junction mapper for RNA-Seq reads. It aligns RNA-Seq reads to mammalian-sized genomes using the ultra high-throughput short read aligner Bowtie, and then analyzes the mapping results to identify splice junctions between exons.

## Cufflinks: transcriptome assembly and differential expression analysis for RNA-Seq

Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.

## CummeRbund: visualization of RNA-Seq differential analysis

CummeRbund is an R package that is designed to aid and simplify the task of analyzing Cufflinks RNA-Seq output.



## Cufflinks: transcriptome assembly and differential expression analysis for RNA-Seq

Cufflinks assembles transcripts, estimates their abundances, and tests for differential expression and regulation in RNA-Seq samples. It accepts aligned RNA-Seq reads and assembles the alignments into a parsimonious set of transcripts. Cufflinks then estimates the relative abundances of these transcripts based on how many reads support each one, taking into account biases in library preparation protocols.

## CummeRbund: visualization of RNA-Seq differential analysis

CummeRbund is an R package that is designed to aid and simplify the task of analyzing Cufflinks RNA-Seq output.



## Cufflinks

Cufflinks is both the name of a suite of tools and a program within that suite. Cufflinks the program assembles transcriptomes from RNA-Seq data and quantifies their expression.

## Cuffcompare

Cuffcompare helps you compare the transcriptomes assembled from different RNA-Seq libraries and assess the quality of your assembly.

## Cuffmerge

Cuffmerge merges the assembled transcriptomes into a master transcriptome. This step is required for a differential expression analysis of the new transcripts you've assembled.

## Cufflinks

Cufflinks is both the name of a suite of tools and a program within that suite. Cufflinks the program assembles transcriptomes from RNA-Seq data and quantifies their expression.

## Cuffcompare

Cuffcompare helps you compare the transcriptomes assembled from different RNA-Seq libraries and assess the quality of your assembly.

## Cuffmerge

Cuffmerge merges the assembled transcriptomes into a master transcriptome. This step is required for a differential expression analysis of the new transcripts you've assembled.

## Cufflinks

Cufflinks is both the name of a suite of tools and a program within that suite. Cufflinks the program assembles transcriptomes from RNA-Seq data and quantifies their expression.

## Cuffcompare

Cuffcompare helps you compare the transcriptomes assembled from different RNA-Seq libraries and assess the quality of your assembly.

## Cuffmerge

Cuffmerge merges the assembled transcriptomes into a master transcriptome. This step is required for a differential expression analysis of the new transcripts you've assembled.

## Cuffquant

Cuffquant computes the gene and transcript expression profiles and save these profiles to files that you can analyze later with Cuffdiff or Cuffnorm. It is recommended for analyses involving more than a handful of libraries.

## Cuffdiff

Cuffdiff is a highly accurate tool for performing the comparisons of expression levels of genes and transcripts.

## Cuffnorm

Cuffnorm normalizes a set of samples to be on as similar scales as possible, which can improve the results you obtain with other downstream tools.

## Cuffquant

Cuffquant computes the gene and transcript expression profiles and save these profiles to files that you can analyze later with Cuffdiff or Cuffnorm. It is recommended for analyses involving more than a handful of libraries.

## Cuffdiff

Cuffdiff is a highly accurate tool for performing the comparisons of expression levels of genes and transcripts.

## Cuffnorm

Cuffnorm normalizes a set of samples to be on as similar scales as possible, which can improve the results you obtain with other downstream tools.

## Cuffquant

Cuffquant computes the gene and transcript expression profiles and save these profiles to files that you can analyze later with Cuffdiff or Cuffnorm. It is recommended for analyses involving more than a handful of libraries.

## Cuffdiff

Cuffdiff is a highly accurate tool for performing the comparisons of expression levels of genes and transcripts.

## Cuffnorm

Cuffnorm normalizes a set of samples to be on as similar scales as possible, which can improve the results you obtain with other downstream tools.

# Tophat Output

- unmapped.bam (BAM)
- accepted\_hits.bam (BAM): a list of read alignments in BAM/SAM format
- junctions.bed (BED): list BED track of junctions reported by Tophat where each junction consists of two connected BED blocks where each block is as long as the max overhang of any read spanning junction
- deletions.bed (BED): mentions the last genomic base before the deletion
- insertions.bed (BED): mentions the first genomic base of deletion

```

23: Tophat2 on data 5, data 1, and ❶ / ❷ ✎
data 4: unmapped bam

24: C1 R1 accepted hits ❶ / ❷ ✎

23: Tophat2 on data 5, data 1, and ❶ / ❷ ✎
data 4: splice junctions

22: Tophat2 on data 5, data 1, and ❶ / ❷ ✎
data 4: deletions

21: Tophat2 on data 5, data 1, and ❶ / ❷ ✎
data 4: insertions
130 regions, 1 comments
format: bed, database: dm3
Log: tool progress Log: tool progress [2013-12-17 16:07:04] Beginning TopHat run (v2.0.10) —
—— [2013-12-17 16:07:04] Checking for
Bowtie Bowtie version: 2.1.0.0 [2013-12-17
16:07:04] Checking for Samtools
  
display at UCSC main test
display in IGB Local Web
display at Ensembl Current

1.Chrom 2.Start 3.End 4.Name
track name=insertions description=Tophat ins
chr2L 2575279 2575279 G

```



## Cufflinks

Input:

- Aligned reads.
  - Gmap / Gsnap.
  - Tophat.

What it can do:

- Assemble transcripts.
- Estimate transcript abundance.



## Cufflinks

Modes of operation:

- Use predefined transcripts.
- Assemble transcripts assisted by known transcripts.
- Assemble transcripts with no prior knowledge.

When to use:

- Only interested in expression.
- Alternative splicing.

Discover new transcripts (Cuffcompare).



## Cuffdiff

Find significant changes in transcript expression, splicing, and promoter use.

- Support multiple samples and replicates
- Models to estimate the distribution
  - pooled, per-condition, blind
- Output a number of statistic tests
  - fold change, p values, q values



# Cuffdiff output

- Genes: gene differential FPKM
- Isoforms: Transcript differential FPKM
- CDS: Coding sequence differential FPKM

[62: Cuffdiff on data 39, data 1, and others](#) / ×  
others: transcript FPKM tracking

[61: Cuffdiff on data 39, data 1, and others](#) / ×  
others: transcript differential expression testing

[60: Cuffdiff on data 39, data 1, and others](#) / ×  
others: gene FPKM tracking  
14,200 lines  
format: tabular, database: dm3  
cuffdiff v2.1.1 (4046M) cuffdiff --no-update-check -q --library-norm-method geometric --dispersion-method pooled -p 4 -c 10 --FDR 0.050000 --labels "C1","C2"  
/BIO/galaxy/files/003/dataset\_3778.dat  
/BIO/galaxy/files/003/dataset\_3789.dat, /BIO/galaxy/files/



1	2	3	4
tracking_id	class_code	nearest_ref_id	gene
128up	-	-	128



## Cuffdiff: differentially expressed genes

Column	Contents
test_stat	value of the test statistic used to compute significance of the observed change in FPKM
p_value	Uncorrected P value for test statistic
q_value	FDR-adjusted p-value for the test statistic
status	Was there enough data to run the test?
significant	and, was the gene differentially expressed?

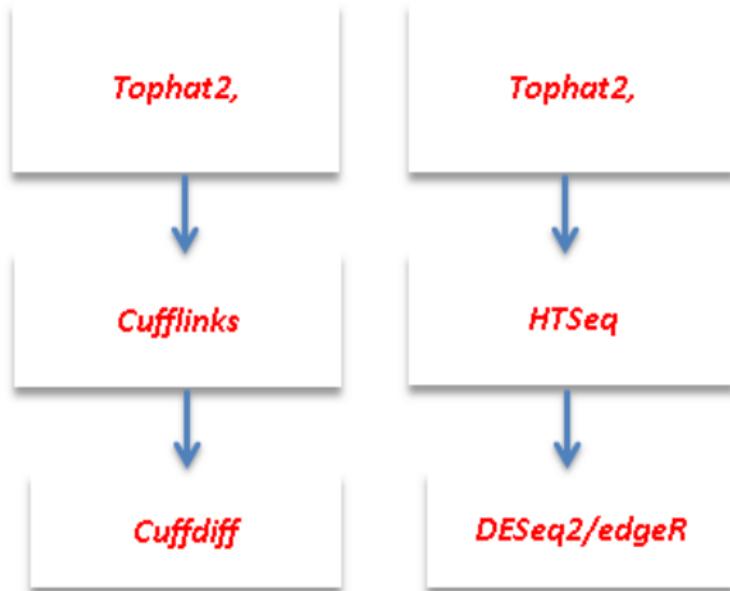


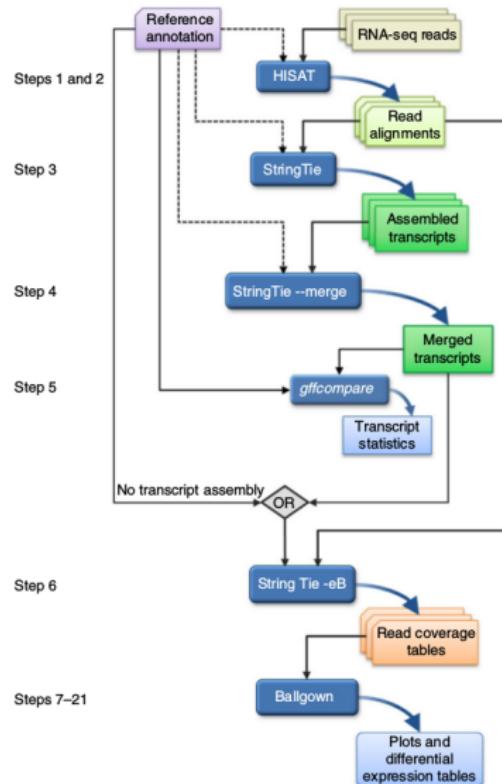
## Cuffdiff

- Column 7 ("status") can be FAIL, NOTEST, LOWDATA or OK
  - Filter and Sort → Filter
    - `c7 == 'OK'`
- Column 14 ("significant") can be yes or no
  - Filter and Sort → Filter
    - `c14 == 'yes'`

Returns the list of genes with  
1) enough data to make a call, and  
2) that are called as differentially expressed.







## HISAT2

HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes (as well as to a single reference genome).

## StringTie

StringTie is a fast and highly efficient assembler of RNA-Seq alignments into potential transcripts.

## Ballgown

Ballgown is a software package designed to facilitate flexible differential expression analysis of RNA-Seq data. It also provides functions to organize, visualize, and analyze the expression measurements for your transcriptome assembly.

## HISAT2

HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes (as well as to a single reference genome).

## StringTie

StringTie is a fast and highly efficient assembler of RNA-Seq alignments into potential transcripts.

## Ballgown

Ballgown is a software package designed to facilitate flexible differential expression analysis of RNA-Seq data. It also provides functions to organize, visualize, and analyze the expression measurements for your transcriptome assembly.

## HISAT2

HISAT2 is a fast and sensitive alignment program for mapping next-generation sequencing reads (both DNA and RNA) to a population of human genomes (as well as to a single reference genome).

## StringTie

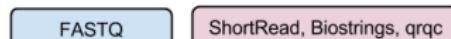
StringTie is a fast and highly efficient assembler of RNA-Seq alignments into potential transcripts.

## Ballgown

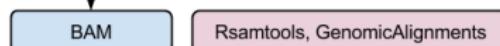
Ballgown is a software package designed to facilitate flexible differential expression analysis of RNA-Seq data. It also provides functions to organize, visualize, and analyze the expression measurements for your transcriptome assembly.

# 转录组学 | RNA-Seq | 分析 | 工具 | Pipeline | R/Bioconductor

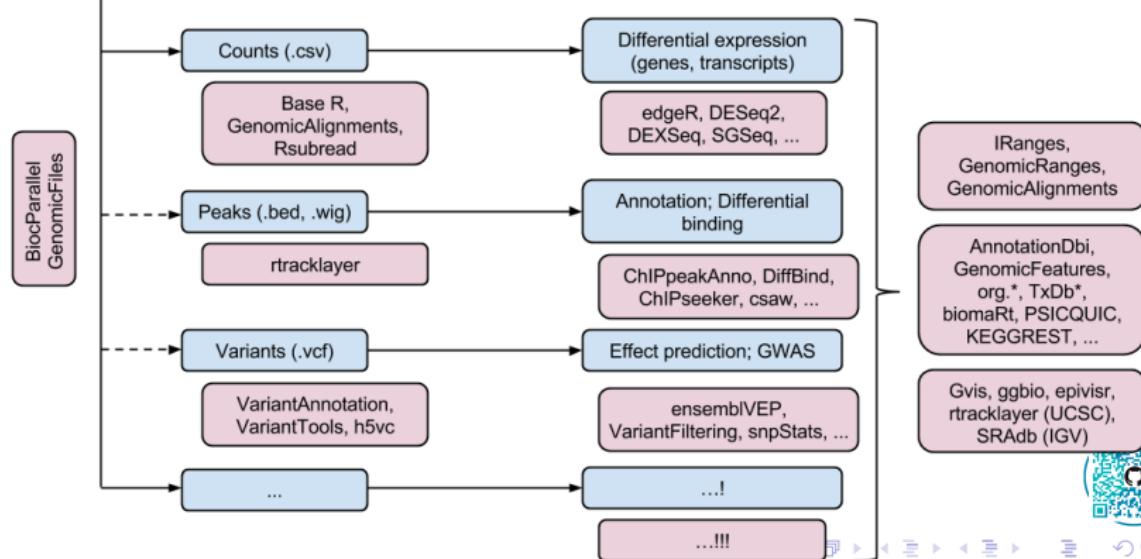
Sequencing



Alignment



Reduction



## Multiple testing problem

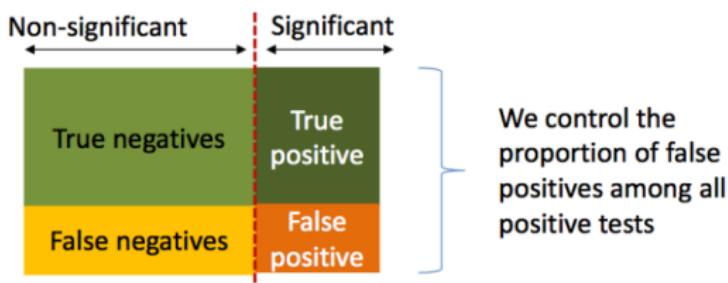
p-value: probability of observing a test statistic that is at least as extreme as the observed one if the null hypothesis were true.

In RNA-seq, we perform 1 test per gene

For example: ~20'000 human genes means ~1'000 significant tests (at P<5%) expected by chance **even if there is NO differential expression!**

#### False discovery rate (FDR) correction

For example Benjamini & Hochberg (1995) J. Roy. Stat. Soc. B



## Correcting for multiple test

By default DESeq2 uses a FDR of 10%

This means that max 10% of the significant genes could be false positives

	baseMean	log2FC	lfcSE	stat	pvalue	padj
WBGene00000001	366.24625	0.73464	0.27865	2.63646	0.00838	0.02992
WBGene00000002	289.48852	-0.58337	0.32379	-1.80172	0.07159	0.16243
WBGene00000003	93.21683	0.11464	0.27268	0.42044	0.67417	0.79280
WBGene00000004	165.72585	-0.69165	0.32537	-2.12575	0.03352	0.09026
WBGene00000005	439.78883	-0.74071	0.29184	-2.53804	0.01115	0.03771
WBGene00000006	244.67827	-1.16500	0.40588	-2.87031	0.00410	0.01658
WBGene00000007	367.81227	-0.00548	0.38734	-0.01416	0.98870	0.99445
WBGene00000008	19.25137	-0.09673	0.56984	-0.16975	0.86521	0.92061

Adjusted p-values using  
Benjamini-Hochberg  
procedure



# After DGE RNA-seq?

You get your “gene list”, finished?

- Validate
- Typically expect some false-positives
- Genes not in your list may be differentially expressed

Important to always remember

- Your list of genes is produced with an arbitrary significance threshold!

Next?

- Gene-set enrichment tests
- Novel transcripts, novel splice-variants, ...



## Shift focus from single genes to collections of genes (gene sets)

Why gene set analysis?

**Genes are believed to work together**

**Many small, but concordant, effects may be detectable together  
even if they are not detectable individually!**

**Fewer gene sets than individual genes - less severe multiple testing problem.**



## Gene sets and software tools

MSigDB (<http://www.broadinstitute.org/gsea/msigdb/index.jsp>)

GSEA (<http://www.broadinstitute.org/gsea/index.jsp>)

DAVID (<http://david.abcc.ncifcrf.gov>)

GO gene ontology (<http://www.geneontology.org>)

KEGG (<http://www.genome.jp/kegg/>)

Reactome (<http://www.reactome.org>)

Publications

Integrated in many software suits and commercial packages like  
GeneGo MetaCore, Pathway Studio, Ingenuity

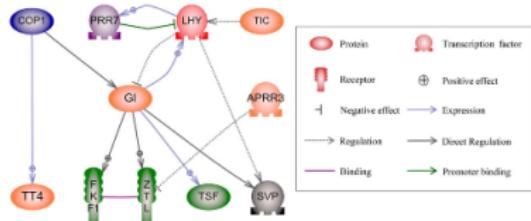
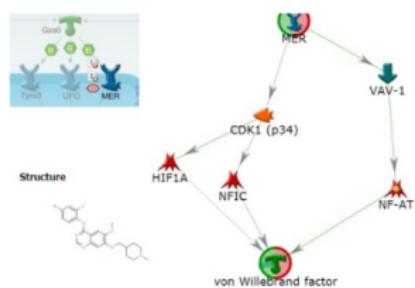
Where to find gene sets?

Some are available in R database packages (KEGG.db, GO.db,  
reactome.db)

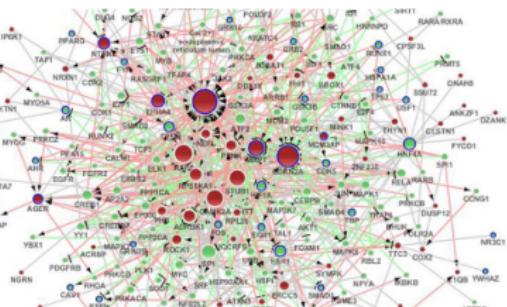


## Pathway analysis

Hoping for this...



You often get this...



# 教学提纲

1

系统生物学

2

基因组学

3

测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

数据库与数据格式

5

二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

外显子组测序

● 简介

● 操作流程

● 应用实例

转录组学

● RNA-Seq

● 概述

● 数据分析

● 应用实例

顺反组

● 概述

● ChIP-Seq

● 表观遗传学

● 概述

● Methyl-Seq

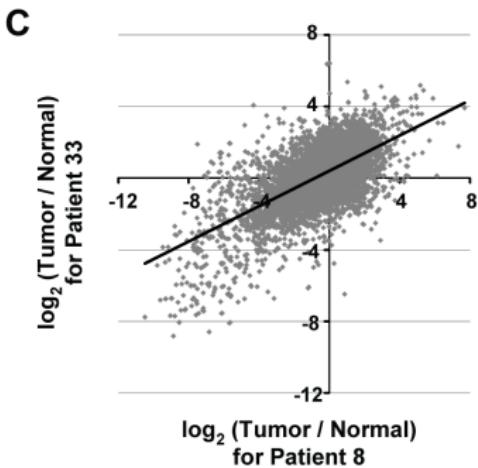
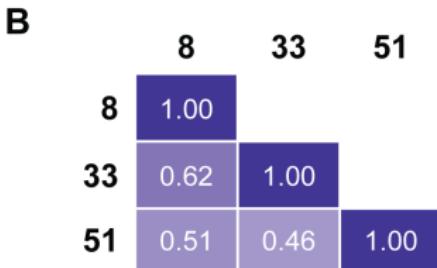
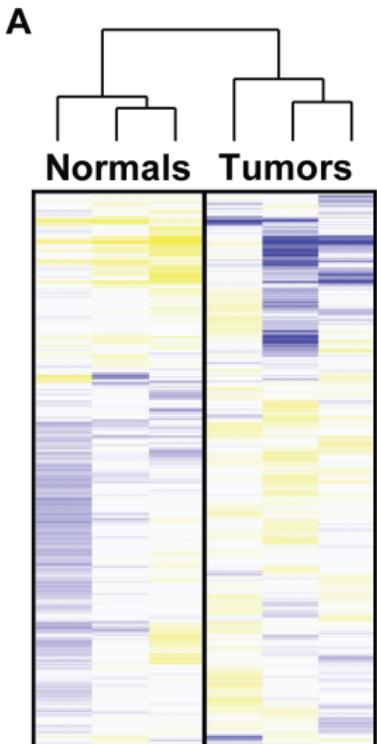


PLoS ONE, 2010

As an example of clinical applications, researchers at the Mayo Clinic used an RNA-Seq approach to identify differentially expressed transcripts between oral cancer and normal tissue samples. They also accurately evaluated the allelic imbalance (AI), ratio of the transcripts produced by the single alleles, within a subgroup of genes involved in cell differentiation, adhesion, cell motility and muscle contraction identifying a unique transcriptomic and genomic signature in oral cancer patients.

Tuch BB, Laborde RR, Xu X, et al. (2010). "Tumor transcriptome sequencing reveals allelic expression imbalances associated with copy number alterations". PLoS ONE. 5 (2): e9317.





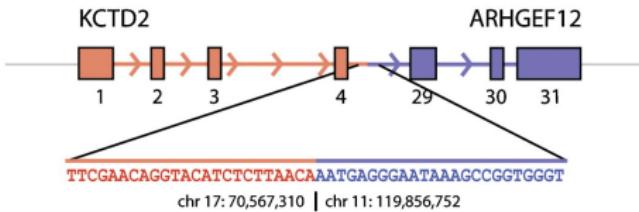
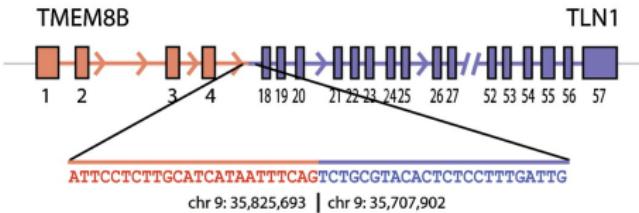
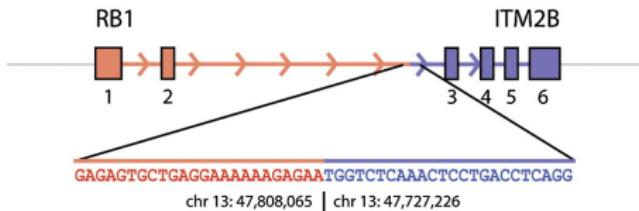
## Genome Research, 2010

Novel insight on skin cancer (melanoma) also come from RNA-Seq of melanoma patients. This approach led to the identification of eleven novel gene fusion transcripts originated from previously unknown chromosomal rearrangements. Twelve novel chimeric transcripts were also reported, including seven of those that confirmed previously identified data in multiple melanoma samples.

Berger MF, Levin JZ, Vijayendran K, et al. (April 2010). "Integrative analysis of the melanoma transcriptome". *Genome Res.* 20 (4): 413–27.



# 转录组学 | RNA-Seq | 实例 | fusion gene



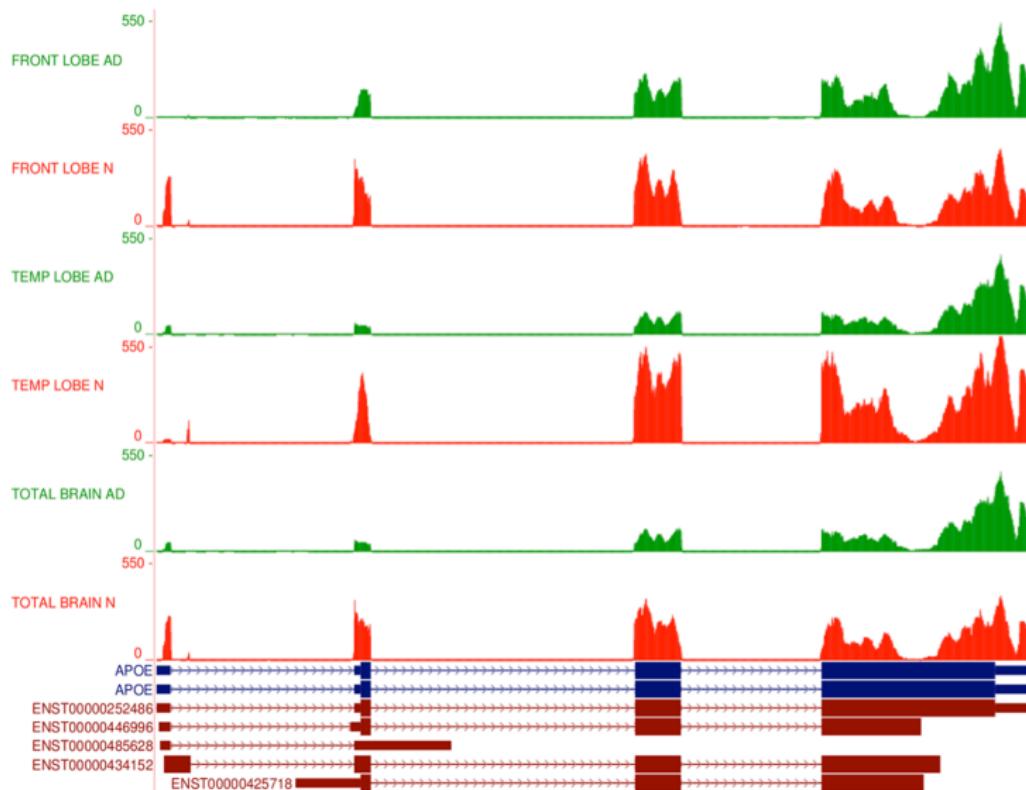
PLoS ONE, 2011

RNA-Seq has been used to study other important chronic diseases such as Alzheimer (AD) and diabetes. In the former case, Twine and colleagues compared the transcriptome of different lobes of deceased AD's patient's brain with the brain of healthy individuals identifying a lower number of splice variants in AD's patients and differential promoter usage of the APOE-001 and -002 isoforms in AD's brains.

Twine NA, Janitz K, Wilkins MR, Janitz M (2011). "Whole transcriptome sequencing reveals gene expression and splicing differences in brain regions affected by Alzheimer's disease". PLoS ONE. 6 (1): e16266.



# 转录组学 | RNA-Seq | 实例 | splice variant



## Mol. Endocrinol, Cell Metab, 2012

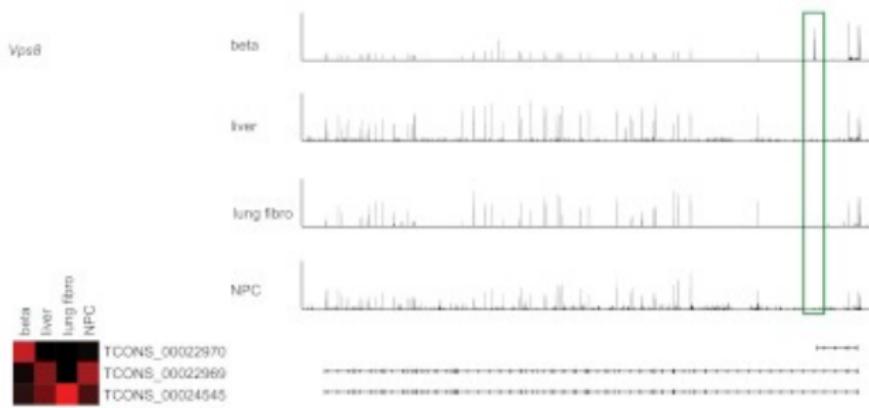
In the latter case, different groups showed the unicity of the beta-cells transcriptome in diabetic patients in terms of transcripts accumulation and differential promoter usage and long non coding RNAs (lncRNAs) signature.

Ku GM, Kim H, Vaughn IW, et al. (October 2012). "Research resource: RNA-Seq reveals unique features of the pancreatic  $\beta$ -cell transcriptome". Mol. Endocrinol. 26 (10): 1783–92.

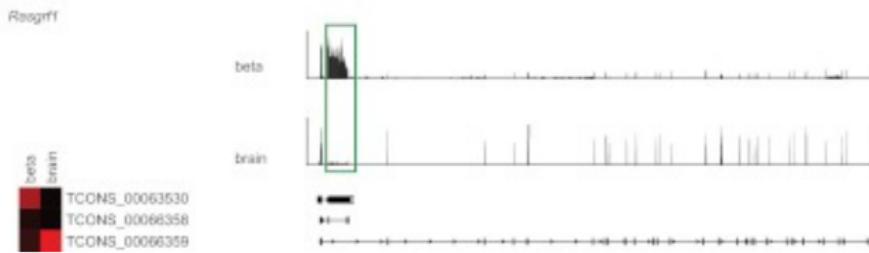
Morán I, Akerman I, van de Bunt M, et al. (October 2012). "Human  $\beta$  cell transcriptome analysis uncovers lncRNAs that are tissue-specific, dynamically regulated, and abnormally expressed in type 2 diabetes". Cell Metab. 16 (4): 435–48.

# 转录组学 | RNA-Seq | 实例 | splicing events and alternative promoter use

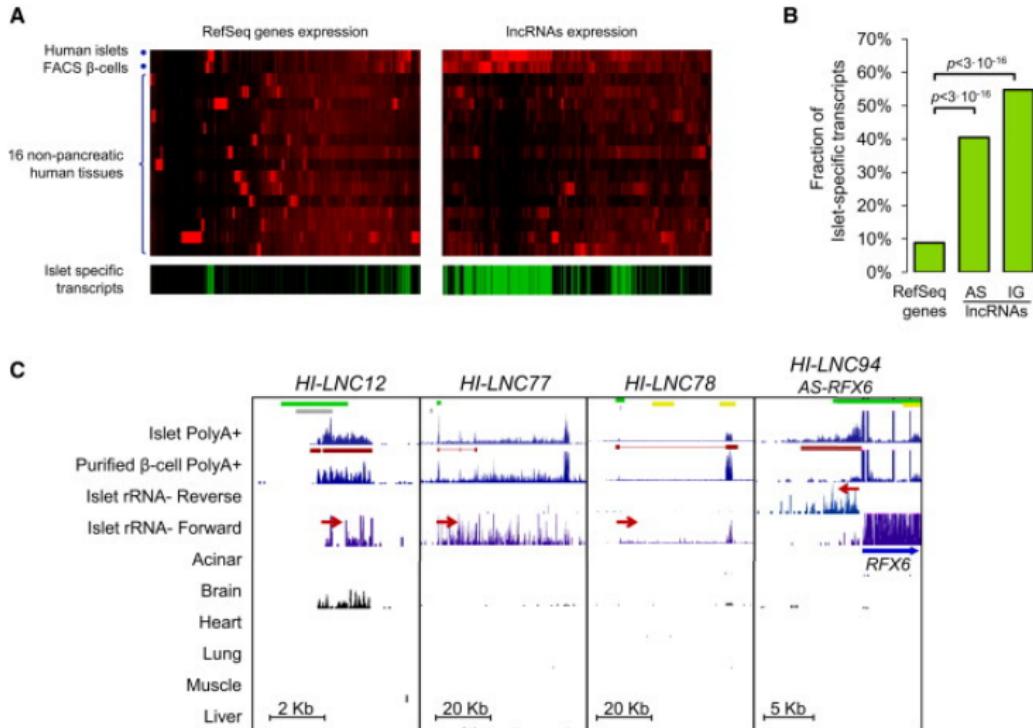
A



B



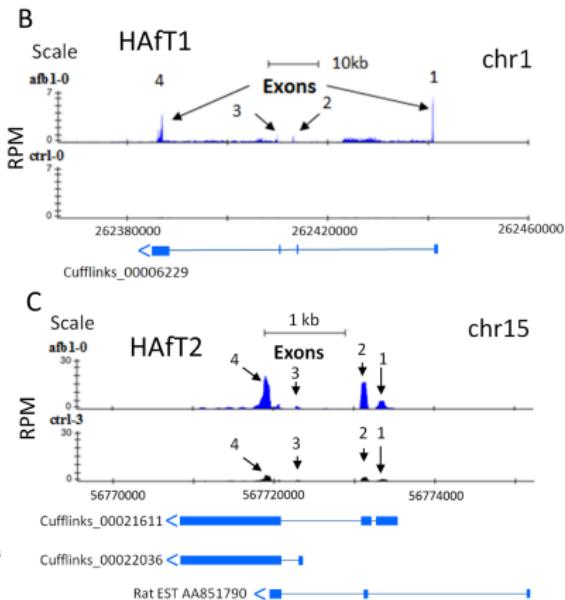
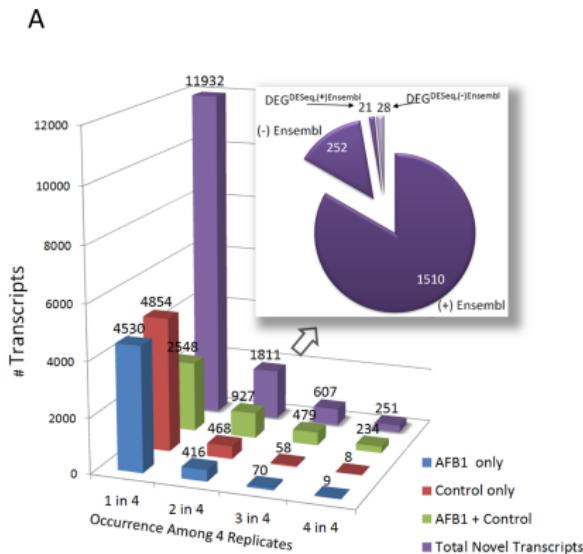
转录组学 | RNA-Seq | 实例 | lncRNA



PLoS ONE, 2013

NGS technology identified several previously undocumented differentially-expressed transcripts in rats treated with AFB1, a potent hepatocarcinogen. Nearly 50 new differentially-expressed transcriptions were identified between the controls and AFB1-treated rats. Additionally potential new exons were identified, including some that are responsive to AFB1. The next-generation sequencing pipeline identified more differential gene expressions compared with microarrays, particularly when DESeq software was utilized. Cufflinks identified two novel transcripts that were not previously annotated in the Ensembl database; these transcripts were confirmed using cloning PCR.

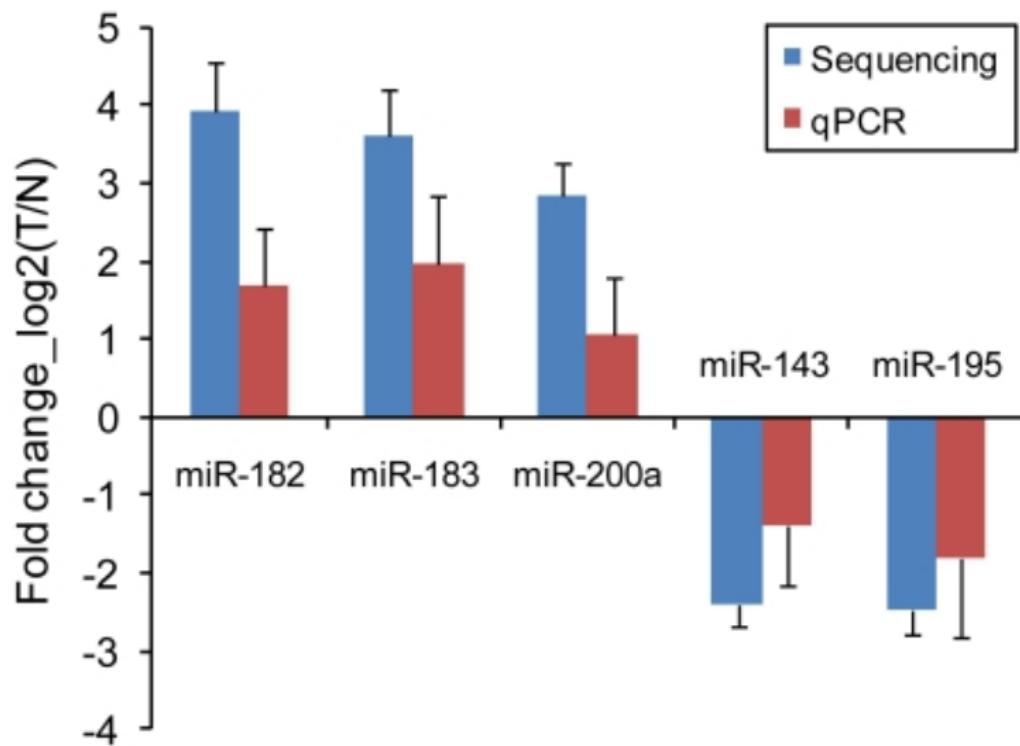
Merrick B. A.; Phadke D. P.; Auerbach S. S.; Mav D.; Stieglmeyer S. M.; Shah R. R.; Tice R. R. (2013). "RNA-seq profiling reveals novel hepatic gene expression pattern in Aflatoxin B1 treated rats". PLoS ONE. 8: e61768.



PLoS ONE, 2011

Han et al. (2011) examined microRNA expression differences in bladder cancer patients in order to understand how changes and dysregulation in microRNA can influence mRNA expression and function. Several microRNAs were differentially expressed in the bladder cancer patients. Upregulation in the aberrant microRNAs was more common than downregulation in the cancer patients. One of the upregulated microRNAs, has-miR-96, has been associated with carcinogenesis, and several of the overexpressed microRNAs have also been observed in other cancers, including ovarian and cervical. Some of the downregulated microRNAs in cancer samples were hypothesized to have inhibitory roles.

Han Y.; Chen J.; Zhao X.; Liang C.; Wang Y.; Sun L.; Jiang Z.; Zhang Z.; Yang R.; Chen J.; Li Z.; Tang A.; Li X.; Ye J.; Guan Z.; Gui Y.; Cai Z. (2011). "MicroRNA expression signatures of bladder cancer revealed by deep sequencing". PLoS ONE. 6: e18286.



# 教学提纲

1

## 系统生物学

2

## 基因组学

3

## 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

## 数据库与数据格式

5

## 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

## 外显子组测序

## ● 简介

## ● 操作流程

## ● 应用实例

## 7 转录组学

## 8 RNA-Seq

## ● 概述

## ● 数据分析

## ● 应用实例

## 9 顺反组

## ● 概述

## ● ChIP-Seq

## 10 表观遗传学

## ● 概述

## ● Methyl-Seq



# 教学提纲

1

系统生物学

2

基因组学

3

测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

数据库与数据格式

5

二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

外显子组测序

- 简介
- 操作流程
- 应用实例

7

转录组学

8

RNA-Seq

- 概述
- 数据分析
- 应用实例

9

顺反组

- 概述
- ChIP-Seq

10

表观遗传学

- 概述
- Methyl-Seq



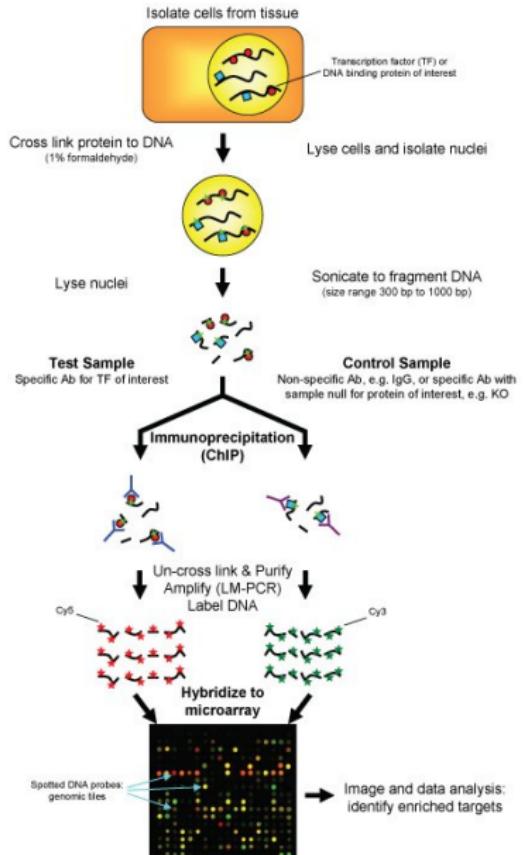
## 维基百科

顺反组（cistrome）指的是“全基因组尺度下反式作用因子的顺式作用靶点的集合，也可以说是在体情况下转录因子结合位点或组蛋白修饰在全基因组上的位置”。“顺反组”这一术语是 cistron（顺反子）和 genome（基因组）的混成词，最初由达纳-法伯癌症研究所和哈佛医学院的研究者命名。

染色质免疫沉淀等技术结合微阵列分析“ChIP-on-chip”或大规模并行 DNA 测序“ChIP-Seq”极大地方便了对转录因子及其它染色质相关蛋白的顺反组的定义。

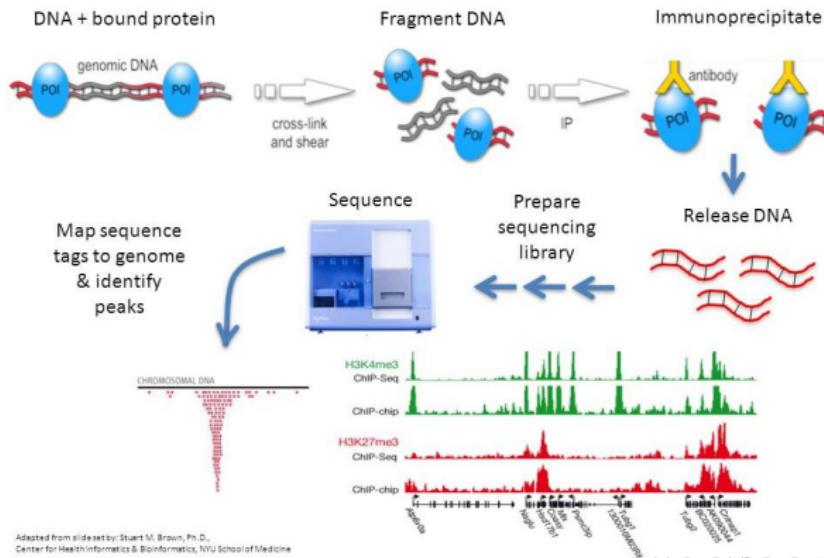


# 顺反组 | 研究方法 | ChIP-on-chip



## ChIP-Seq

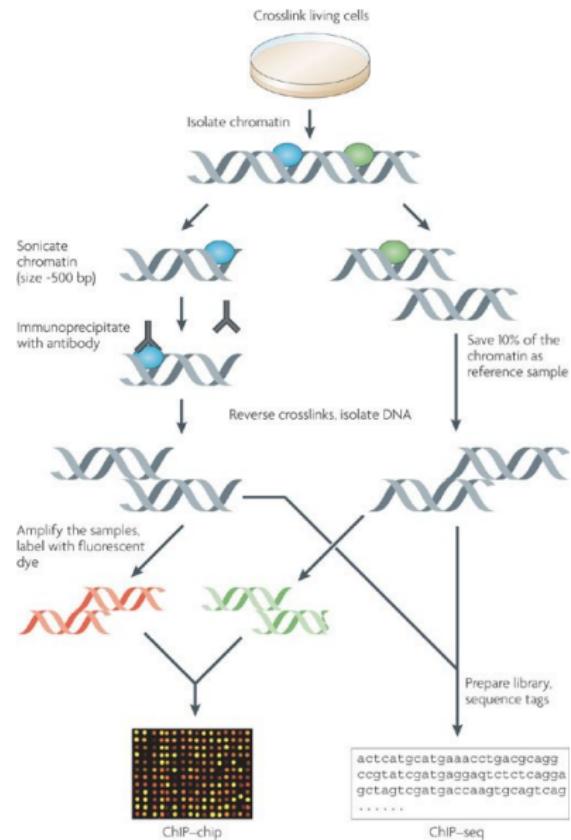
染色质免疫沉淀-测序（ChIP-sequencing，简称为 ChIP-seq）被用于分析**蛋白质与 DNA 的交互作用**。该技术将染色质免疫沉淀（ChIP）与大规模并行 DNA 测序结合起来以鉴定与 DNA 相关蛋白的结合部位。其可被用于精确绘制任意目的蛋白在全基因组上的结合位点。在此之前，ChIP-on-chip 是研究这些蛋白-DNA 联系的最常用的技术。



Adapted from slide set by Stuart M. Brown, Ph.D.,  
Center for Health Informatics & Bioinformatics, NYU School of Medicine



顺反组 | 研究方法 | ChIP-Seq vs. ChIP-chip



# 教学提纲

1

系统生物学

2

基因组学

3

测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

数据库与数据格式

5

二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

外显子组测序

- 简介
- 操作流程
- 应用实例

7

转录组学

8

RNA-Seq

- 概述
- 数据分析
- 应用实例

9

顺反组

- 概述
- ChIP-Seq

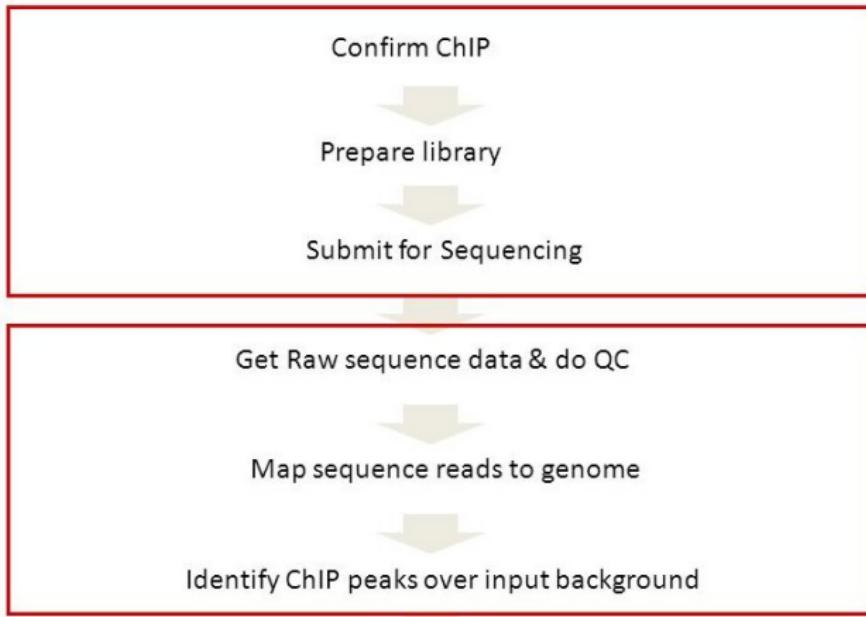
10

表观遗传学

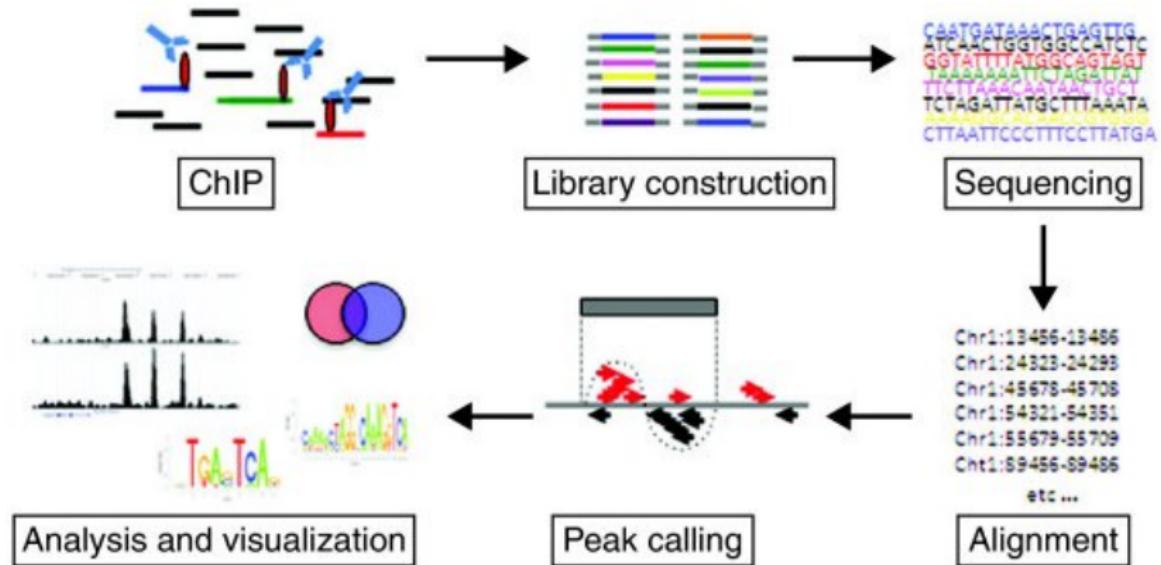
- 概述
- Methyl-Seq

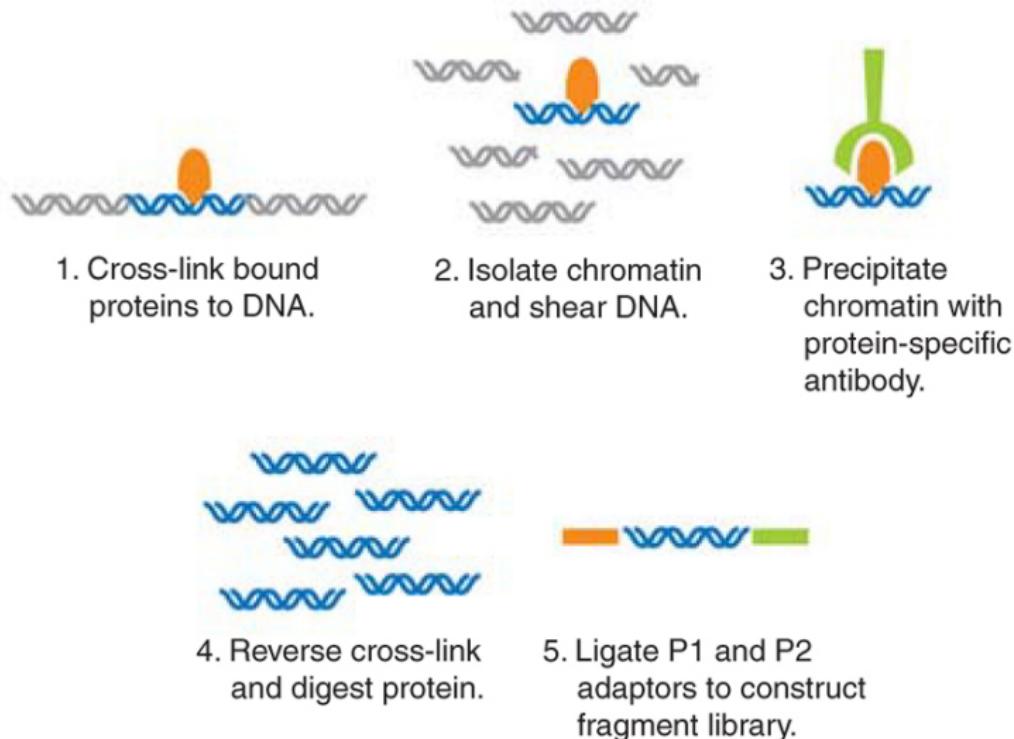


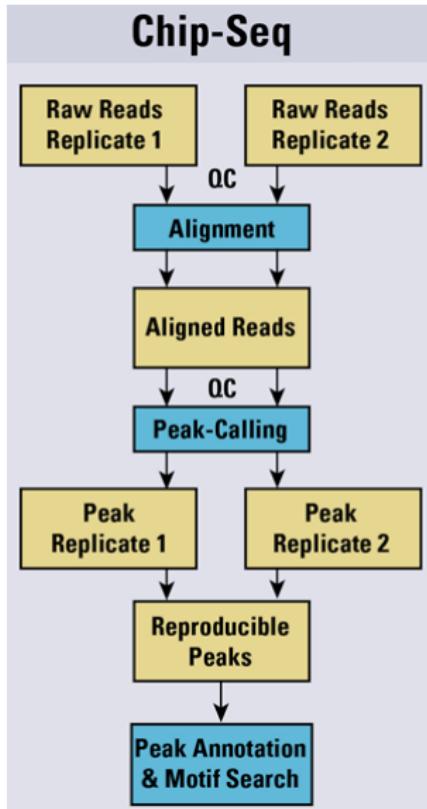
# ChIP-seq Workflow

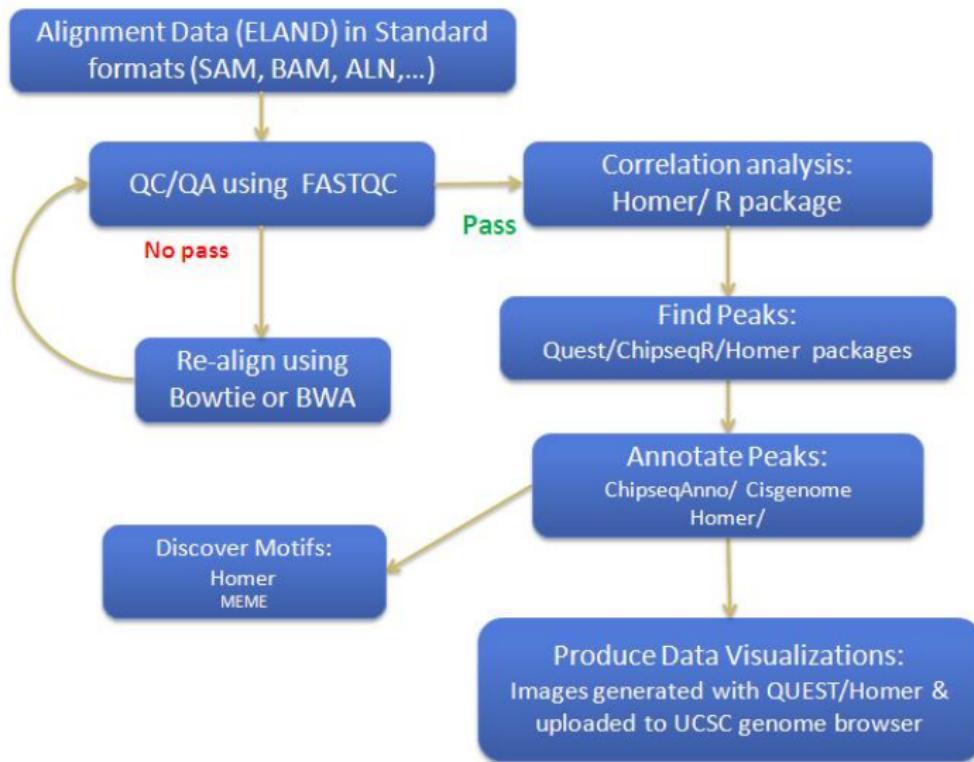


# 顺反组 | ChIP-Seq | 分析 | 流程

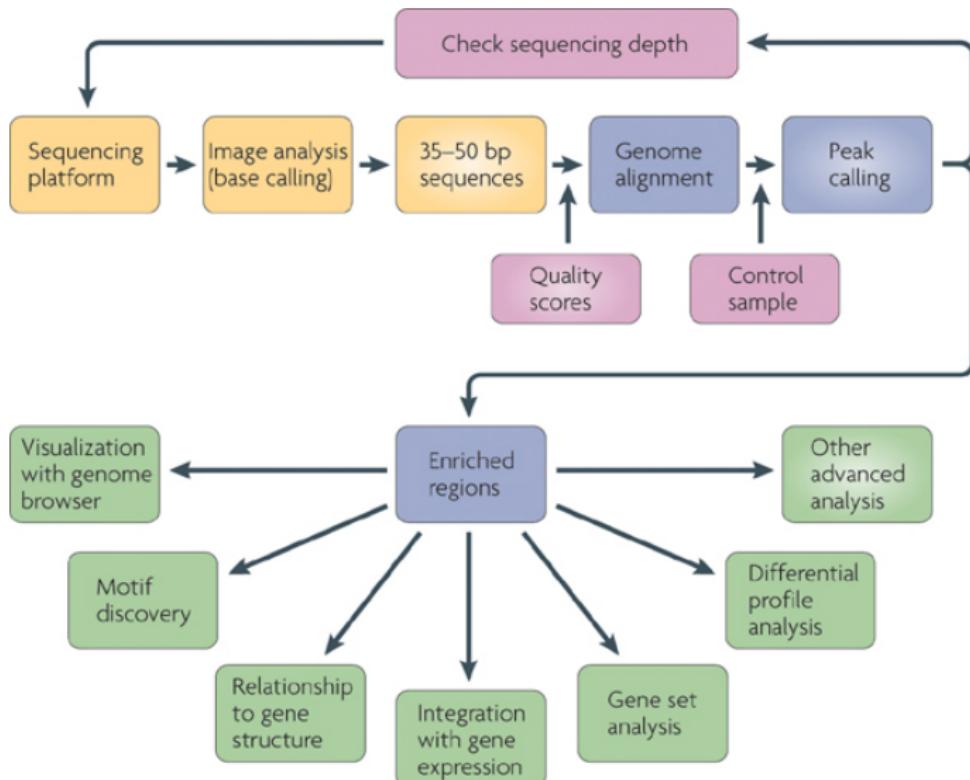


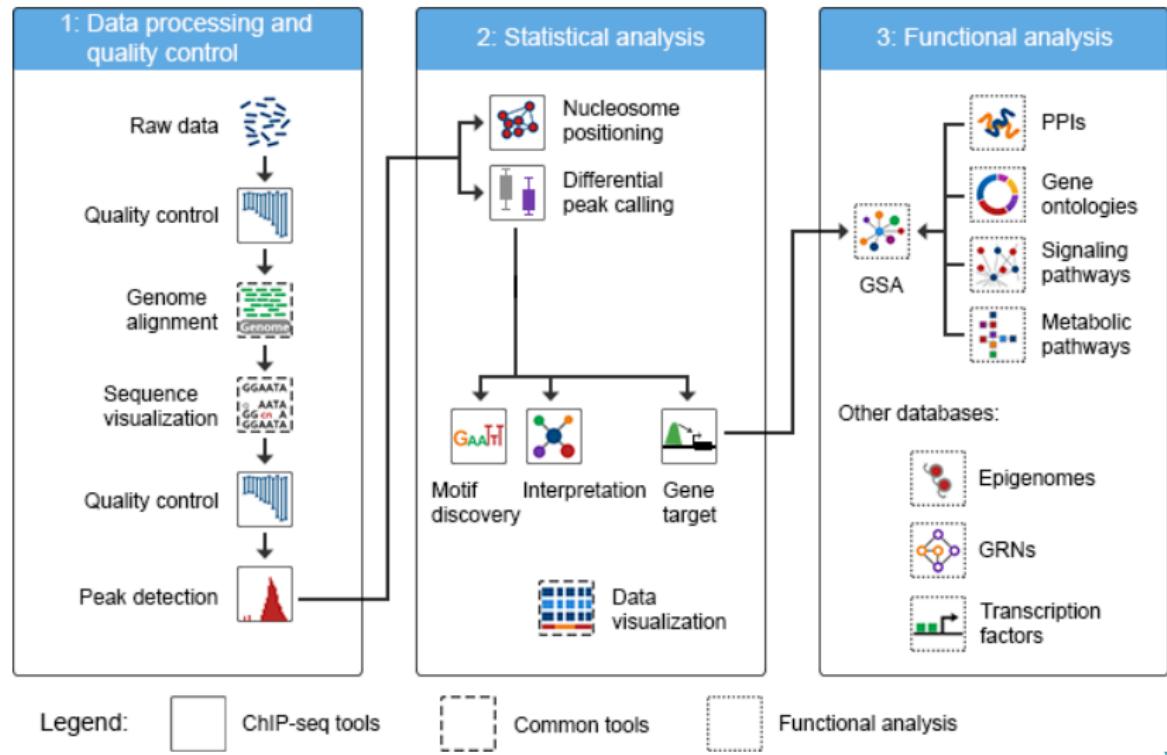






# 顺反组 | ChIP-Seq | 分析 | 流程 | 生信





## Peak calling

Peak calling 是一种用于鉴定经染色质免疫沉淀-测序或 MeDIP-测序实验后所得到的比对读段富集在基因组哪些区域中的一种计算方法。当免疫沉淀的蛋白质是一种转录因子时，那么 DNA 的富集区域就是转录因子结合位点 (TFBS)。主流的 Peak calling 软件有 MACS 等。

Peak calling 可应用于转录组/外显子组测序，亦可用于对 MeRIP-测序或 m6A-测序的 RNA 表观基因组测序数据进行分析；利用如 exomePeak 等的软件程序，可检测出转录后的 RNA 修饰位点。



## Differential peak calling

Differential peak calling is about identifying significant differences in two ChIP-seq signals. One can distinguish between one-stage and two-stage differential peak callers.

### One stage differential peak callers

One stage differential peak callers work in two phases: first, call peaks on individual ChIP-seq signals and second, combine individual signals and apply statistical tests to estimate differential peaks. DBChIP and MAnorm are examples for one stage differential peak callers.

### Two stage differential peak callers

Two stage differential peak callers segment two ChIP-seq signals and identify differential peaks in one step. They take advantage of signal segmentation approaches such as Hidden Markov Models. Examples for two-stage differential peak callers are ChIPDiff, ODIN, and THOR. Differential peak calling can also be applied in the context of analyzing RNA-binding protein binding sites.

## Differential peak calling

Differential peak calling is about identifying significant differences in two ChIP-seq signals. One can distinguish between one-stage and two-stage differential peak callers.

### One stage differential peak callers

One stage differential peak callers work in two phases: first, call peaks on individual ChIP-seq signals and second, combine individual signals and apply statistical tests to estimate differential peaks. DBChIP and MAnorm are examples for one stage differential peak callers.

### Two stage differential peak callers

Two stage differential peak callers segment two ChIP-seq signals and identify differential peaks in one step. They take advantage of signal segmentation approaches such as Hidden Markov Models. Examples for two-stage differential peak callers are ChIPDiff, ODIN, and THOR. Differential peak calling can also be applied in the context of analyzing RNA-binding protein binding sites.

## Differential peak calling

Differential peak calling is about identifying significant differences in two ChIP-seq signals. One can distinguish between one-stage and two-stage differential peak callers.

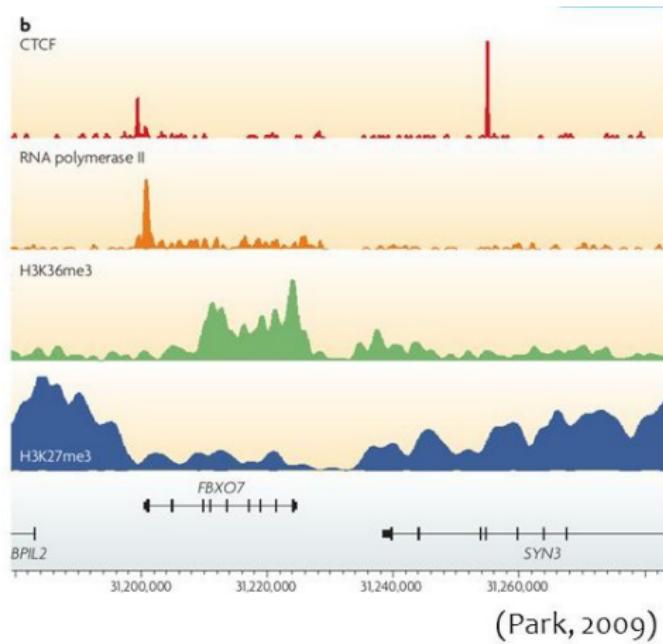
### One stage differential peak callers

One stage differential peak callers work in two phases: first, call peaks on individual ChIP-seq signals and second, combine individual signals and apply statistical tests to estimate differential peaks. DBChIP and MAnorm are examples for one stage differential peak callers.

### Two stage differential peak callers

Two stage differential peak callers segment two ChIP-seq signals and identify differential peaks in one step. They take advantage of signal segmentation approaches such as Hidden Markov Models. Examples for two-stage differential peak callers are ChIPDiff, ODIN, and THOR. Differential peak calling can also be applied in the context of analyzing RNA-binding protein binding sites.

# 顺反组 | ChIP-Seq | 分析 | Peak calling



Sharp (e.g. TF binding)

Mixture (e.g. polymerase binding)

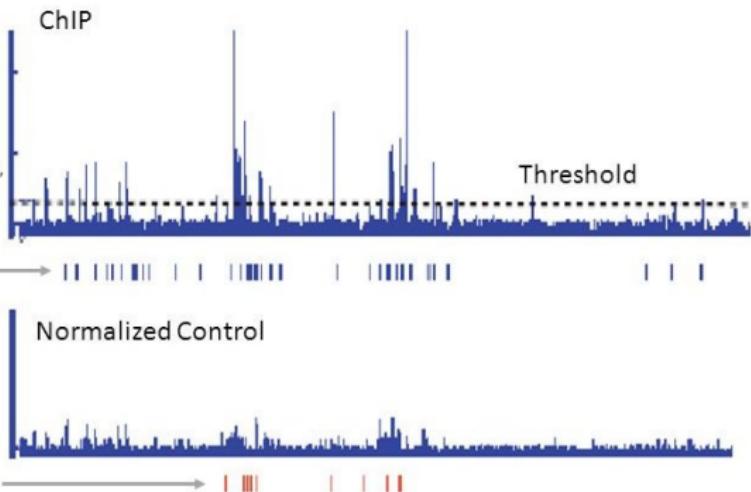
Broad (e.g. histone modification)



## Peak Calling

- Generate and threshold the signal profile and identify candidate target regions

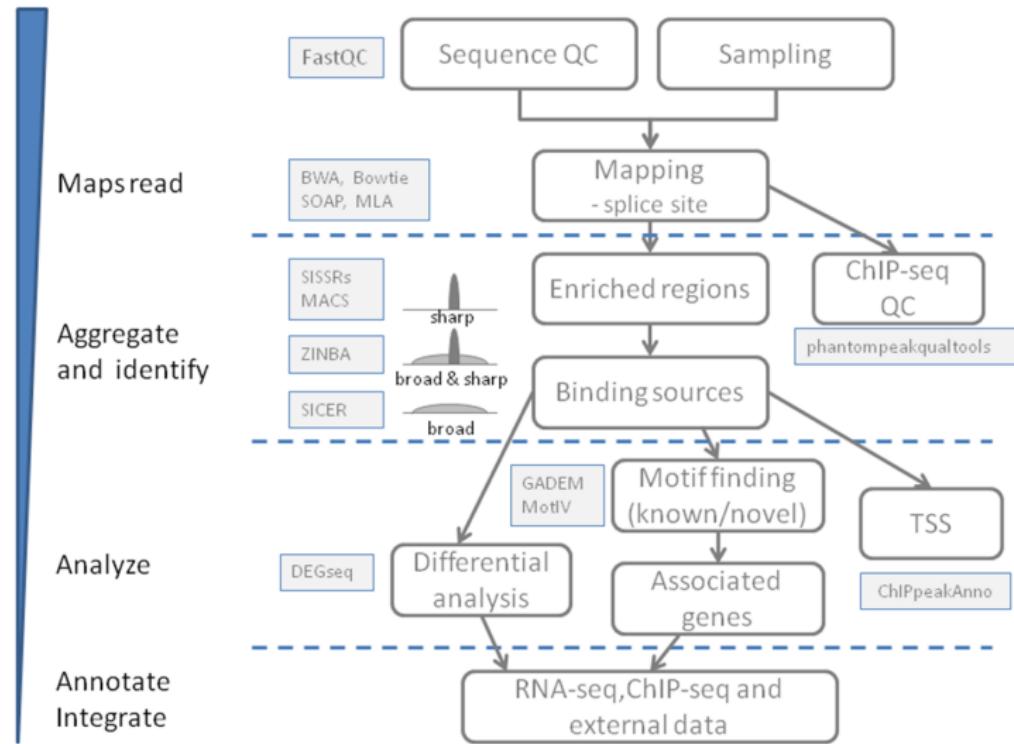
- Simulation (PeakSeq),
- Local window based Poisson (MACS),
- Fold change statistics (SPP)



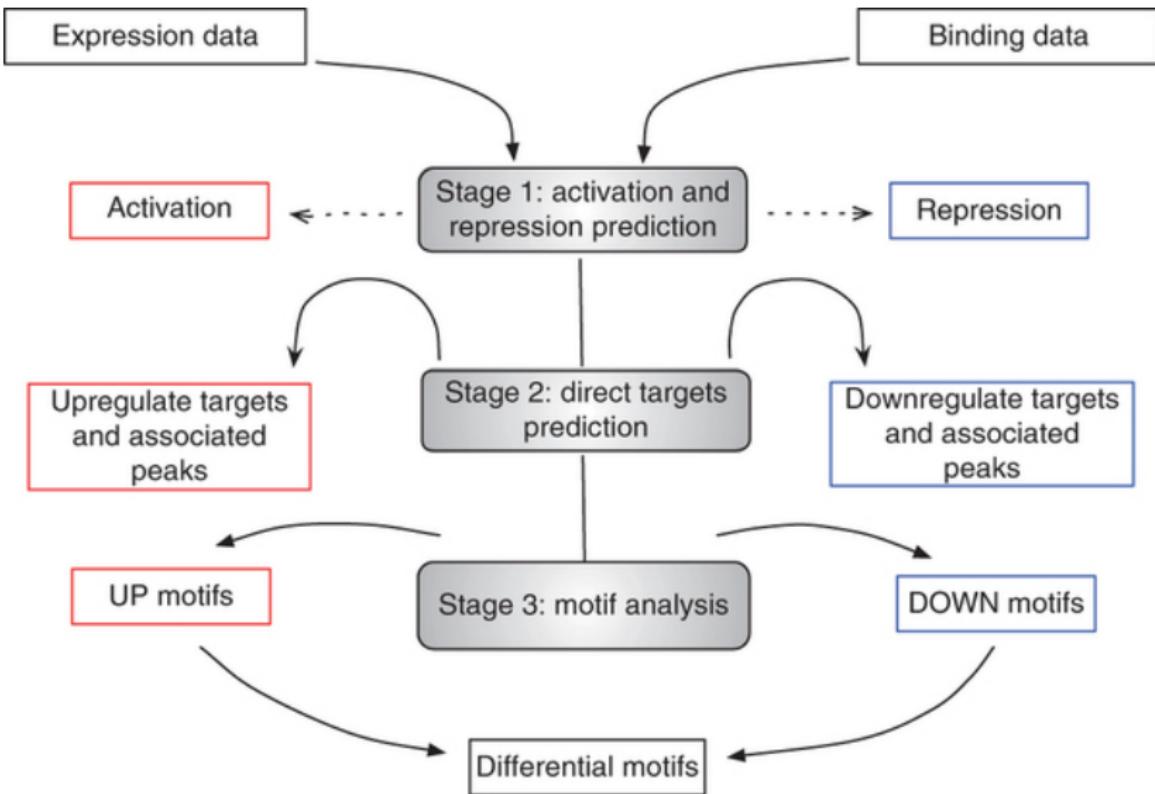
- Score against the control

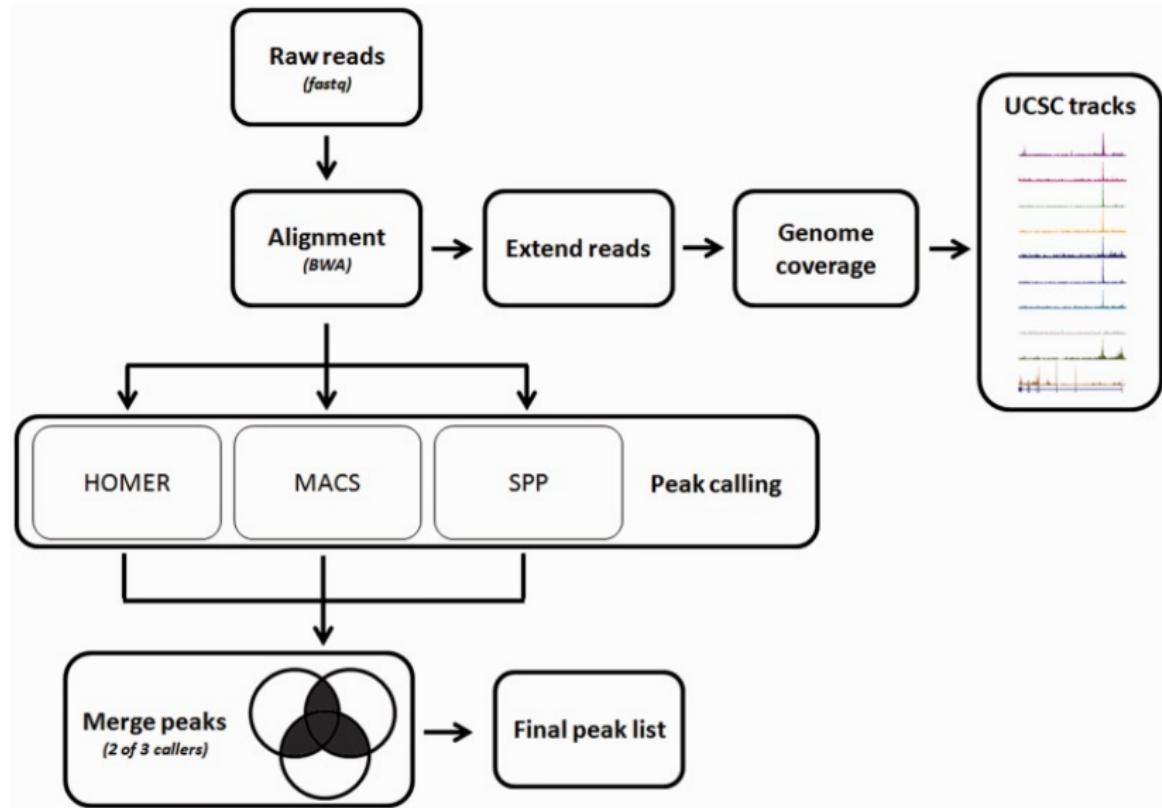


# 顺反组 | ChIP-Seq | 分析 | 应用



# 顺反组 | ChIP-Seq | 分析 | 应用





## MACS

Model-based Analysis of ChIP-Seq (MACS) empirically models the length of the sequenced ChIP fragments, which tends to be shorter than sonication or library construction size estimates, and uses it to improve the spatial resolution of predicted binding sites. MACS also uses a dynamic Poisson distribution to effectively capture local biases in the genome sequence, allowing for more sensitive and robust prediction. MACS compares favorably to existing ChIP-Seq peak-finding algorithms, is publicly available open source, and can be used for ChIP-Seq with or without control samples.

## SPP

An R package for analysis of ChIP-seq and other functional sequencing data.

## MACS

Model-based Analysis of ChIP-Seq (MACS) empirically models the length of the sequenced ChIP fragments, which tends to be shorter than sonication or library construction size estimates, and uses it to improve the spatial resolution of predicted binding sites. MACS also uses a dynamic Poisson distribution to effectively capture local biases in the genome sequence, allowing for more sensitive and robust prediction. MACS compares favorably to existing ChIP-Seq peak-finding algorithms, is publicly available open source, and can be used for ChIP-Seq with or without control samples.

## SPP

An R package for analysis of ChIP-seq and other functional sequencing data.

## PeakSeq

PeakSeq is a program for identifying and ranking peak regions in ChIP-Seq experiments. It takes as input, mapped reads from a ChIP-Seq experiment, mapped reads from a control experiment and outputs a file with peak regions ranked with increasing Q-values.

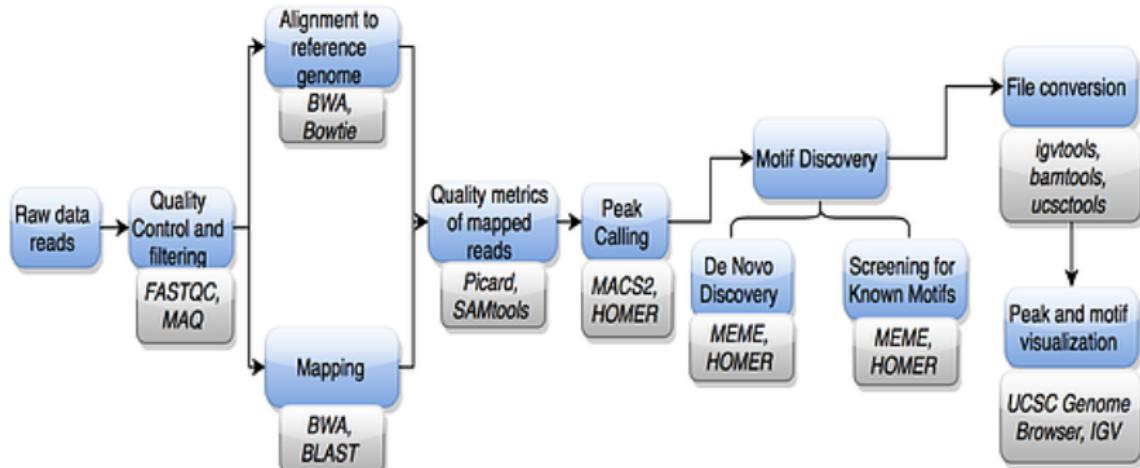


## HOMER

HOMER (Hypergeometric Optimization of Motif EnRichment) is a suite of tools for Motif Discovery and next-gen sequencing analysis. It is a collection of command line programs for unix-style operating systems written in Perl and C++. HOMER was primarily written as a *de novo* motif discovery algorithm and is well suited for finding 8-20 bp motifs in large scale genomics data. HOMER contains many useful tools for analyzing ChIP-Seq, GRO-Seq, RNA-Seq, DNase-Seq, Hi-C and numerous other types of functional genomics sequencing data sets.

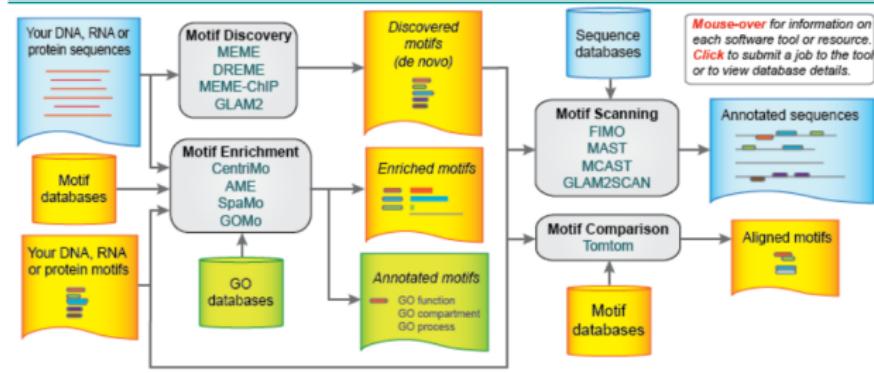


## Motif Discovery and Analysis Using ChIP-seq



# The MEME Suite

## Motif-based sequence analysis tools



Digitized by srujanika@gmail.com



Copyright © 2010 Pearson Education, Inc., publishing as Pearson Benjamin Cummings.



Digitized by srujanika@gmail.com



Copyright © 2010 Pearson Education, Inc., publishing as Pearson Benjamin Cummings.



ANSWER



ANSWER



— 1 —



[View Details](#) | [Edit](#) | [Delete](#)



[View Details](#) | [Edit](#) | [Delete](#)



— 1 —

## Nature Methods, 2007

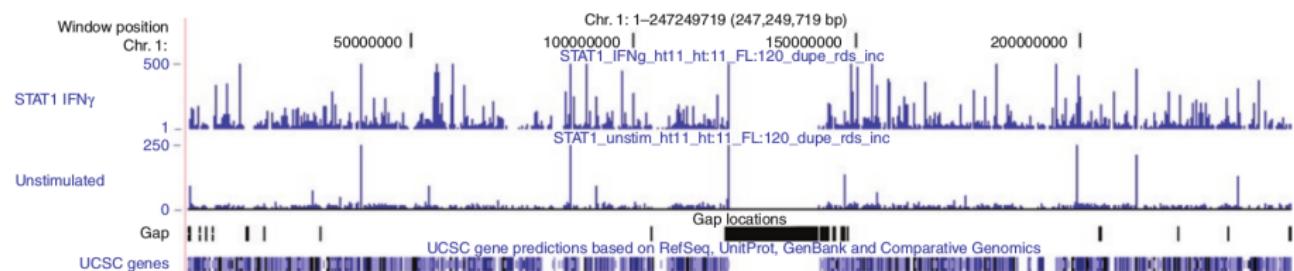
STAT1 DNA association: ChIP-seq was used to study STAT1 targets in HeLa S3 cells. The performance of ChIP-seq was then compared to the alternative protein-DNA interaction methods of ChIP-PCR and ChIP-chip.

ChIP-seq offers an alternative to ChIP-chip. STAT1 experimental ChIP-seq data have a high degree of similarity to results obtained by ChIP-chip for the same type of experiment, with >64% of peaks in shared genomic regions. Because the data are sequence reads, ChIP-seq offers a rapid analysis pipeline (as long as a high-quality genome sequence is available for read mapping, and the genome doesn't have repetitive content that confuses the mapping process) as well as the potential to detect mutations in binding-site sequences, which may directly support any observed changes in protein binding and gene regulation.

Robertson G et al.(2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature Methods 4: 651–657.



# 顺反组 | ChIP-Seq | 实例



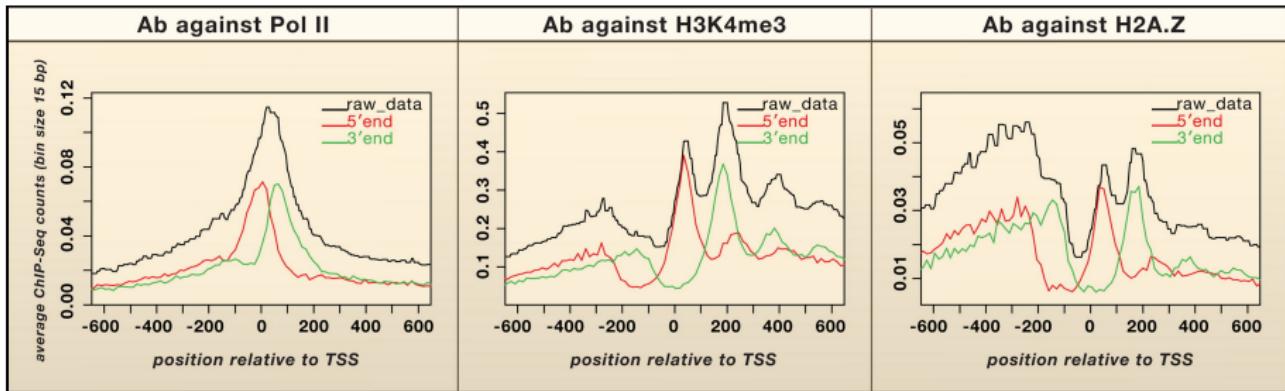
Cell, 2007

Nucleosome Architecture of Promoters: Using ChIP-seq, it was determined that Yeast genes seem to have a minimal nucleosome-free promoter region of 150bp in which RNA polymerase can initiate transcription.

Schmid et al. (2007) ChIP-Seq Data reveal nucleosome architecture of human promoters. Cell 131: 831–832



# 顺反组 | ChIP-Seq | 实例



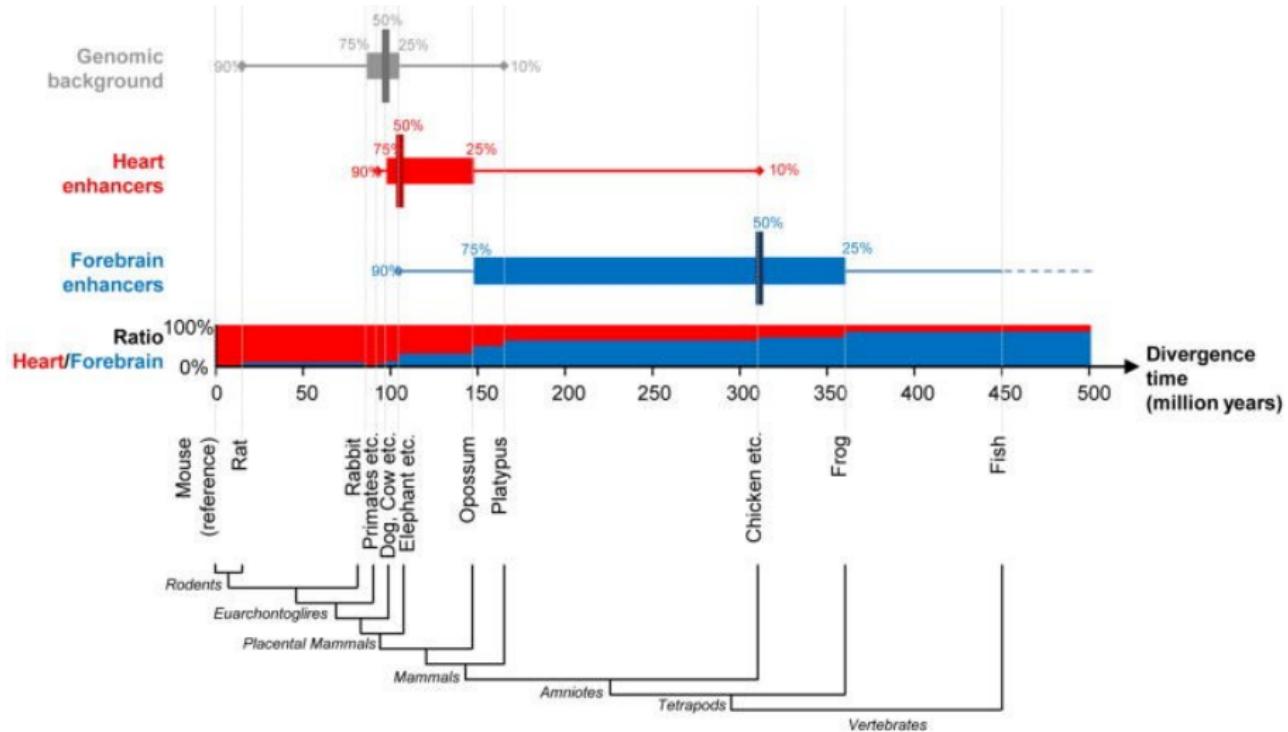
Nature Genetics, 2010

Transcription factor conservation: ChIP-seq was used to compare conservation of TFs in the forebrain and heart tissue in embryonic mice. The authors identified and validated the heart functionality of transcription enhancers, and determined that transcription enhancers for the heart are less conserved than those for the forebrain during the same developmental stage.

Blow, M. J., McCulley, D. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Bristow, J., Ren, B., Black, B. L., Rubin, E. M., Visel, A., & Pennacchio, L. A. (2010). ChIP-seq identification of weakly conserved heart enhancers. Nature Genetics, 42, 806-810.



# 顺反组 | ChIP-Seq | 实例

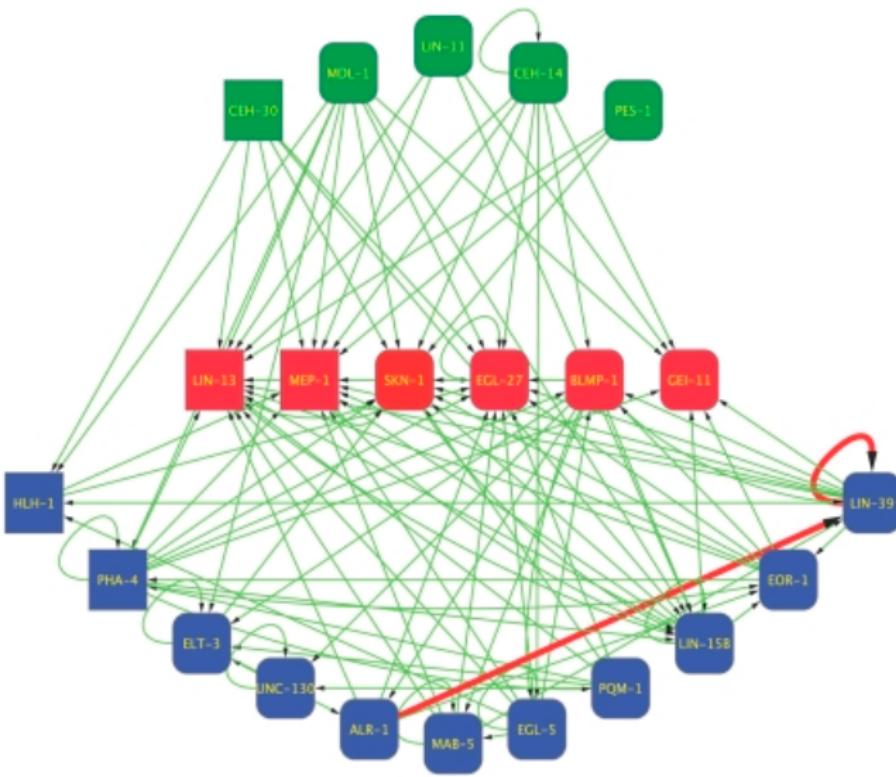


## Genome Research, 2011

Genome-wide ChIP-seq: ChIP-sequencing was completed on the worm *C. elegans* to explore genome-wide binding sites of 22 transcription factors. Up to 20% of the annotated candidate genes were assigned to transcription factors. Several transcription factors were assigned to non-coding RNA regions and may be subject to developmental or environmental variables. The functions of some of the transcription factors were also identified. Some of the transcription factors regulate genes that control other transcription factors. These genes are not regulated by other factors. Most transcription factors serve as both targets and regulators of other factors, demonstrating a network of regulation.

Niu, W., Lu, Z. J., Zhong, M., Sarov, M., Murray, J. I., Brdlik, C. M., Janette, J., Chen, C., Alves, P., Preston, E., Slightham, C., Jiang, L., Hyman, A. A., Kim, S. K., Waterston, R. H., Gerstein, M., Snyder, M., & Reinke, V. (2011). Diverse transcription factor binding features revealed by genome-wide ChIP-seq in *C. elegans*. *Genome Research*, 21, 245-254.

# 顺反组 | ChIP-Seq | 实例



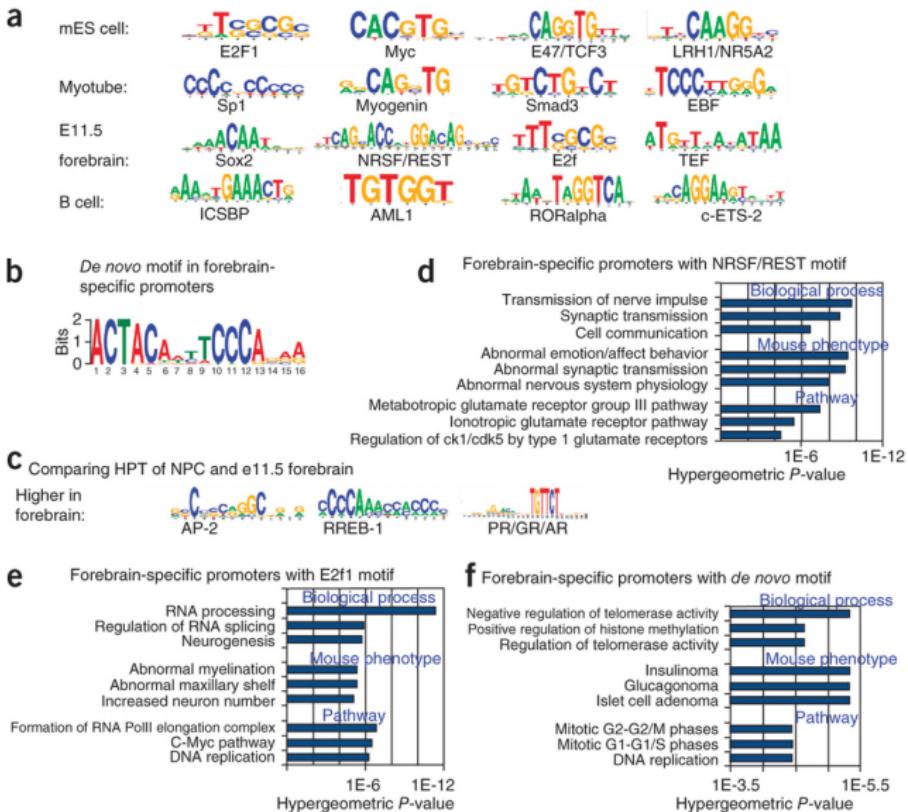
Nature Biotechnology, 2013

Inferring regulatory network: ChIP-seq signal of Histone modification were shown to be more correlated with transcription factor motifs at promoters in comparison to RNA level. Hence author proposed that using histone modification ChIP-seq would provide more reliable inference of gene-regulatory networks in comparison to other methods based on expression.

Vibhor Kumar, Masafumi Muratani, Nirmala Arul Rayan, Petra Kraus, Thomas Lufkin, Huck Hui Ng and Shyam Prabhakar. Uniform, optimal signal processing of mapped deep-sequencing data. Nature biotechnology, 2013.



# 顺反组 | ChIP-Seq | 实例



# 教学提纲

1

## 系统生物学

2

## 基因组学

3

## 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

## 数据库与数据格式

5

## 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

## 外显子组测序

## ● 简介

## ● 操作流程

## ● 应用实例

## 7 转录组学

## 8 RNA-Seq

## ● 概述

## ● 数据分析

## ● 应用实例

## 9 顺反组

## ● 概述

## ● ChIP-Seq

## 10 表观遗传学

## ● 概述

## ● Methyl-Seq



# 教学提纲

1

## 系统生物学

2

## 基因组学

3

## 测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

## 数据库与数据格式

5

## 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

## 外显子组测序

## ● 简介

## ● 操作流程

## ● 应用实例

## 7 转录组学

## 8 RNA-Seq

## ● 概述

## ● 数据分析

## ● 应用实例

## 9 顺反组

## ● 概述

## ● ChIP-Seq

## 10 表观遗传学

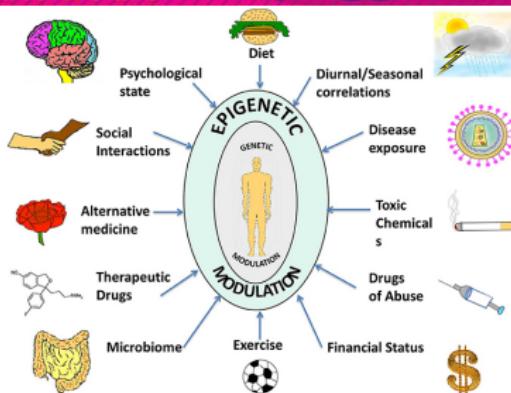
## ● 概述

## ● Methyl-Seq



# EPIGENETICS

How the experiences  
of previous  
generations can  
affect who we are



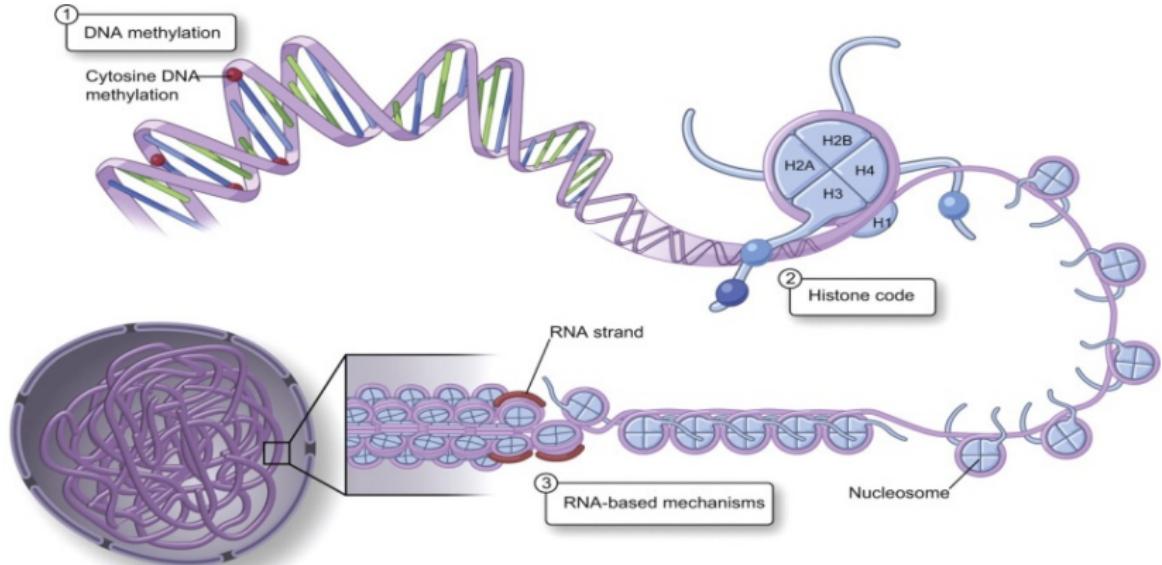
## 表观遗传学

表观遗传学 (epigenetics) 研究的是在不改变 DNA 序列的前提下，通过某些机制引起可遗传的基因表达或细胞表现型的变化。表观遗传学是 20 世纪 80 年代逐渐兴起的一门学科，是在研究与经典的孟德尔遗传学遗传法则不相符的许多生命现象过程中逐步发展起来的。

表观遗传现象包括 DNA 甲基化、组蛋白修饰、RNA 干扰等。与经典遗传学以研究基因序列影响生物学功能为核心相比，表观遗传学主要研究这些“表观遗传现象”建立和维持的机制。其研究内容主要包括两类，一类为基因选择性转录表达的调控，有 DNA 甲基化、基因印记、组蛋白共价修饰和染色质重塑；另一类为基因转录后的调控，包括基因组中非编码 RNA、微小 RNA、反义 RNA、内含子及核糖开关等。

表观遗传学指基因组相关功能改变而不涉及核苷酸序列变化，即“由染色体改变所引起的稳定的可遗传的表现型，而非 DNA 序列的改变”。

# 表观遗传学 | 简介



## DNA 甲基化

DNA 甲基化 (DNA methylation) 为 DNA 化学修饰的一种形式，能在不改变 DNA 序列的前提下，改变遗传表现。为外遗传编码 (epigenetic code) 的一部分，是一种外遗传机制。DNA 甲基化过程会使甲基添加到 DNA 分子上，例如在胞嘧啶环的 5' 碳上：这种 5' 方向的 DNA 甲基化方式可见于所有脊椎动物。特定胞嘧啶受甲基化的情形，可利用亚硫酸盐测序 (bisulfite sequencing) 方式测定。DNA 甲基化可能使基因沉默化，进而使其失去功能。

在人类细胞内，大约有 1% 的 DNA 碱基受到了甲基化。在成熟体细胞组织中，DNA 甲基化一般发生于 CpG 双核苷酸 (CpG dinucleotide) 部位；而非 CpG 甲基化则在胚胎干细胞中较为常见。

植物体内胞嘧啶的甲基化则可分为对称的 CpG (或 CpNpG)，或是不对称的 CpNpNp 形式 (C 与 G 是碱基；p 是磷酸根；N 指的是任意的核苷酸)。

## DNA 甲基化与肿瘤

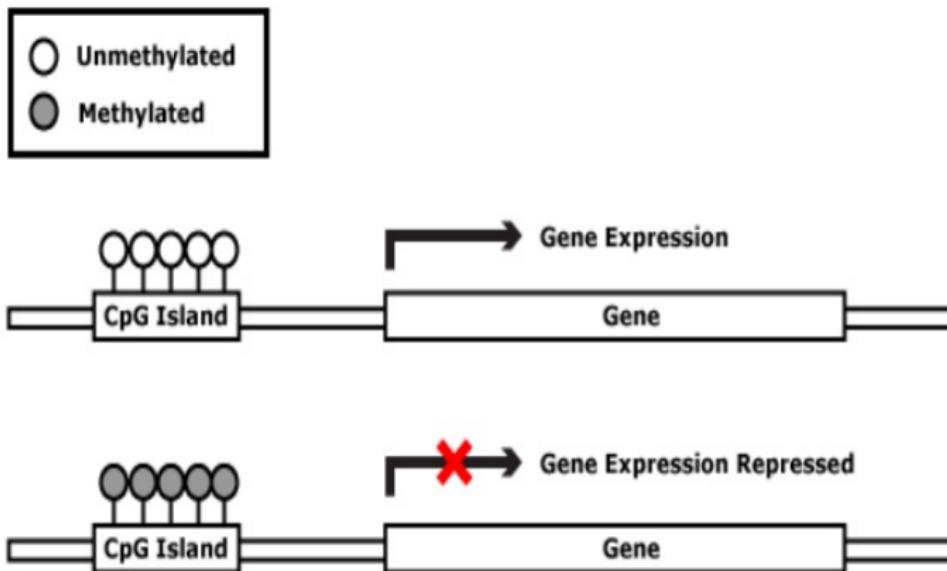
DNA 甲基化是一种基因转录的重要的调节器，许多证据已经证实，异常的 DNA 甲基化与不定期的基因沉默有关，若在启动子区域具有高水平的 5-甲基胞嘧啶，将发生基因沉默。

DNA 甲基化在胚胎发育期间是必需的，在体细胞中，DNA 甲基化的方式通常是高保真的传给子细胞。

异常的 DNA 甲基化模式与大量的人类恶性肿瘤有关，并发现其与正常组织相比存在两种不寻常的形式：超甲基化和低甲基化。超甲基化是主要的表观遗传修饰中的一种，其通过肿瘤抑制基因的启动子区抑制转录。超甲基化通常发生在启动子区的 CpG 岛，且与基因失活有关。整体的低甲基化也通过不同机制与癌症的发生和发展有关。



## DNA Cytosine Methylation



Mettivier, R. et al. Cyclical DNA methylation of a transcriptionally active promoter. *Nature* 452, 45–50 (2008).



# 教学提纲

1

系统生物学

2

基因组学

3

测序技术

- 测序技术的发展
- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术的比较

4

数据库与数据格式

5

二代测序数据分析

- 常见术语
- 分析流程
- 补遗

6

外显子组测序

- 简介
- 操作流程
- 应用实例

7

转录组学

8

RNA-Seq

- 概述
- 数据分析
- 应用实例

9

顺反组

- 概述
- ChIP-Seq

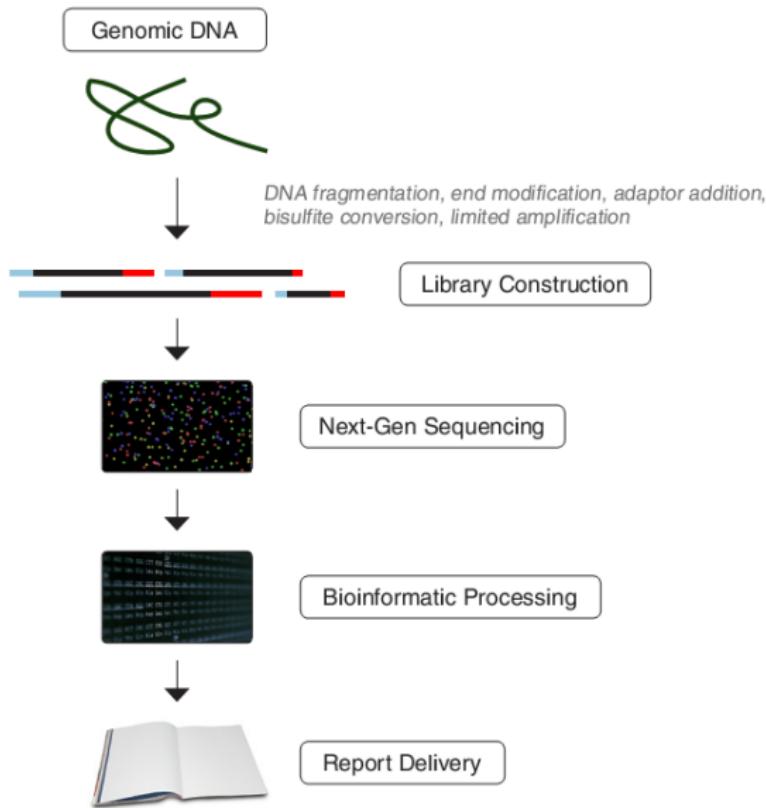
10

表观遗传学

- 概述
- Methyl-Seq

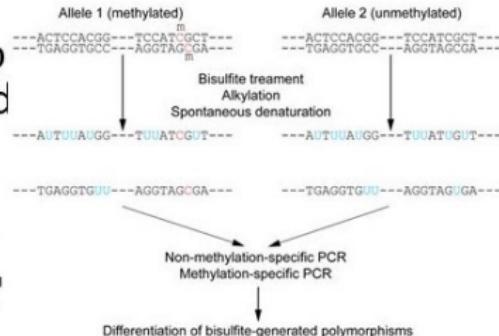
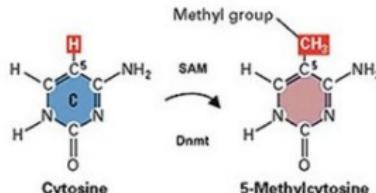


# 表观遗传学 | Methyl-Seq | 分析 | 流程



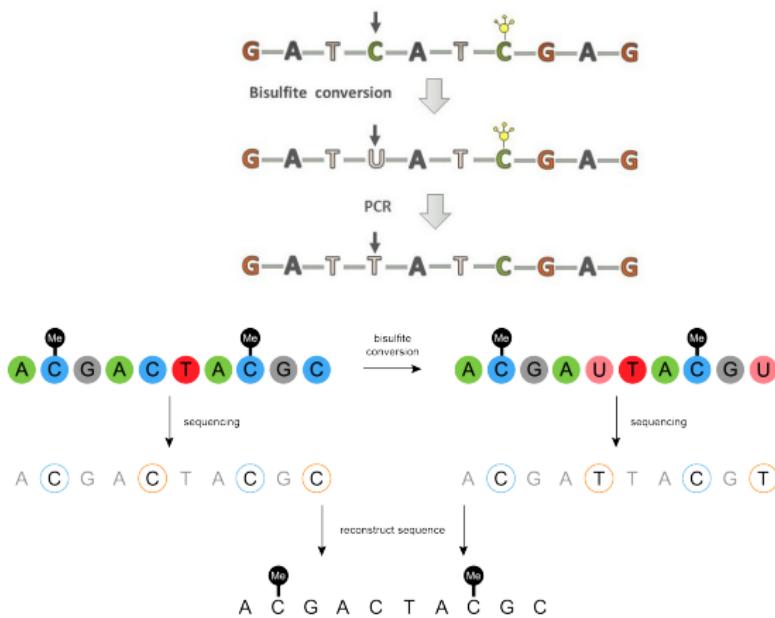
# Bisulfite sequencing

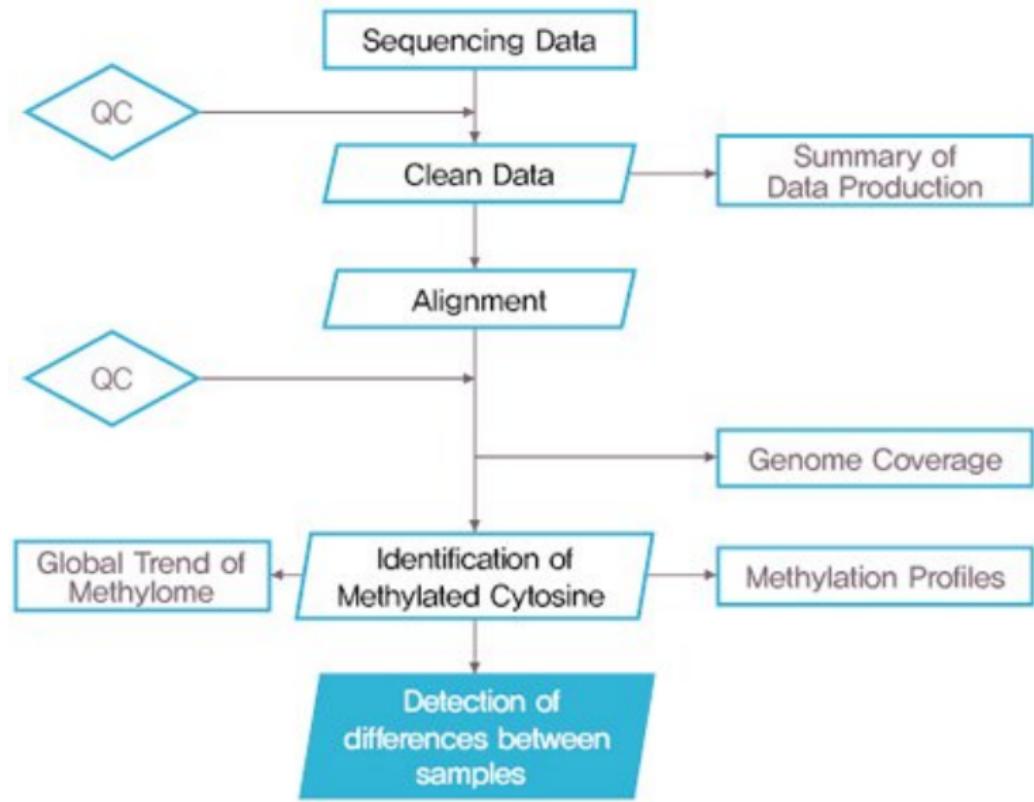
- To find out cytosines methylated at the carbon-5 position
    - Usually occurring at CpG, CpHpG and CpHpH nucleotide patterns

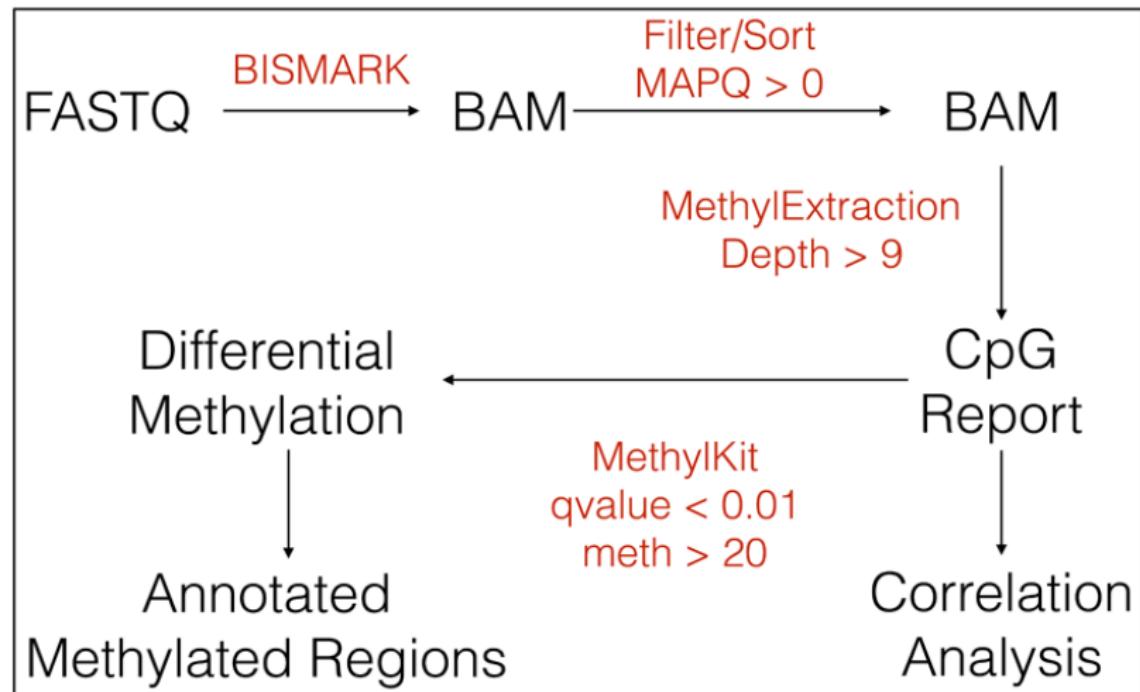


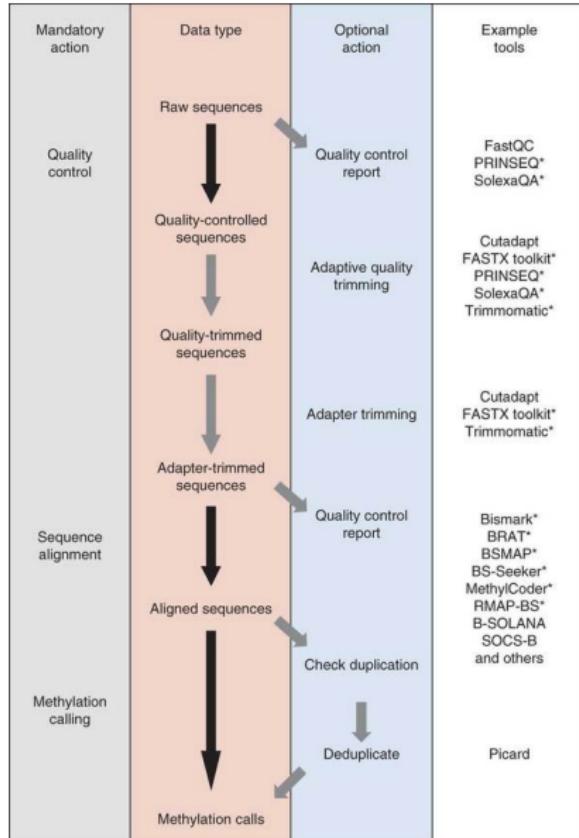
## METHYL-Seq

- Bisulfite Conversion

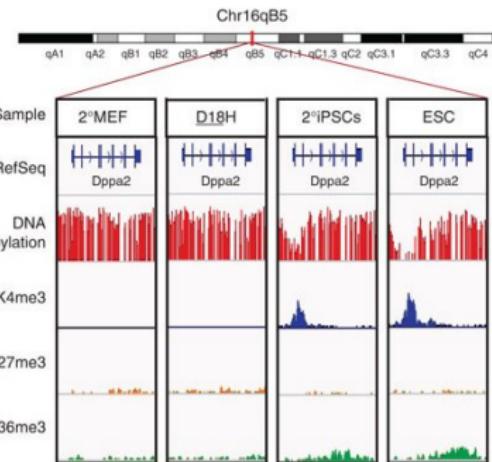
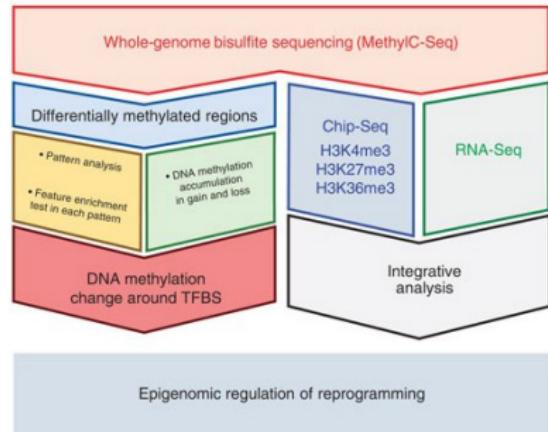








# 表观遗传学 | Methyl-Seq | 分析 | 应用



## Bismark

Bismark is a program to map bisulfite treated sequencing reads to a genome of interest and perform methylation calls in a single step. The output can be easily imported into a genome viewer, such as SeqMonk, and enables a researcher to analyse the methylation levels of their samples straight away.

## bwa-meth

Fast and accurate alignment of BS-Seq reads.

## methylKit

methylKit is an R package for DNA methylation analysis and annotation from high-throughput bisulfite sequencing. The package is designed to deal with sequencing data from RRBS and its variants, but also target-capture methods such as Agilent SureSelect methyl-seq. In addition, methylKit can deal with base-pair resolution data for 5hmC obtained from Tab-seq or oxBs-seq. It can also handle whole-genome bisulfite sequencing data if proper input format is provided.

## Bismark

Bismark is a program to map bisulfite treated sequencing reads to a genome of interest and perform methylation calls in a single step. The output can be easily imported into a genome viewer, such as SeqMonk, and enables a researcher to analyse the methylation levels of their samples straight away.

## bwa-meth

Fast and accurate alignment of BS-Seq reads.

## methylKit

methylKit is an R package for DNA methylation analysis and annotation from high-throughput bisulfite sequencing. The package is designed to deal with sequencing data from RRBS and its variants, but also target-capture methods such as Agilent SureSelect methyl-seq. In addition, methylKit can deal with base-pair resolution data for 5hmC obtained from Tab-seq or oxBs-seq. It can also handle whole-genome bisulfite sequencing data if proper input format is provided.

## Bismark

Bismark is a program to map bisulfite treated sequencing reads to a genome of interest and perform methylation calls in a single step. The output can be easily imported into a genome viewer, such as SeqMonk, and enables a researcher to analyse the methylation levels of their samples straight away.

## bwa-meth

Fast and accurate alignment of BS-Seq reads.

## methylKit

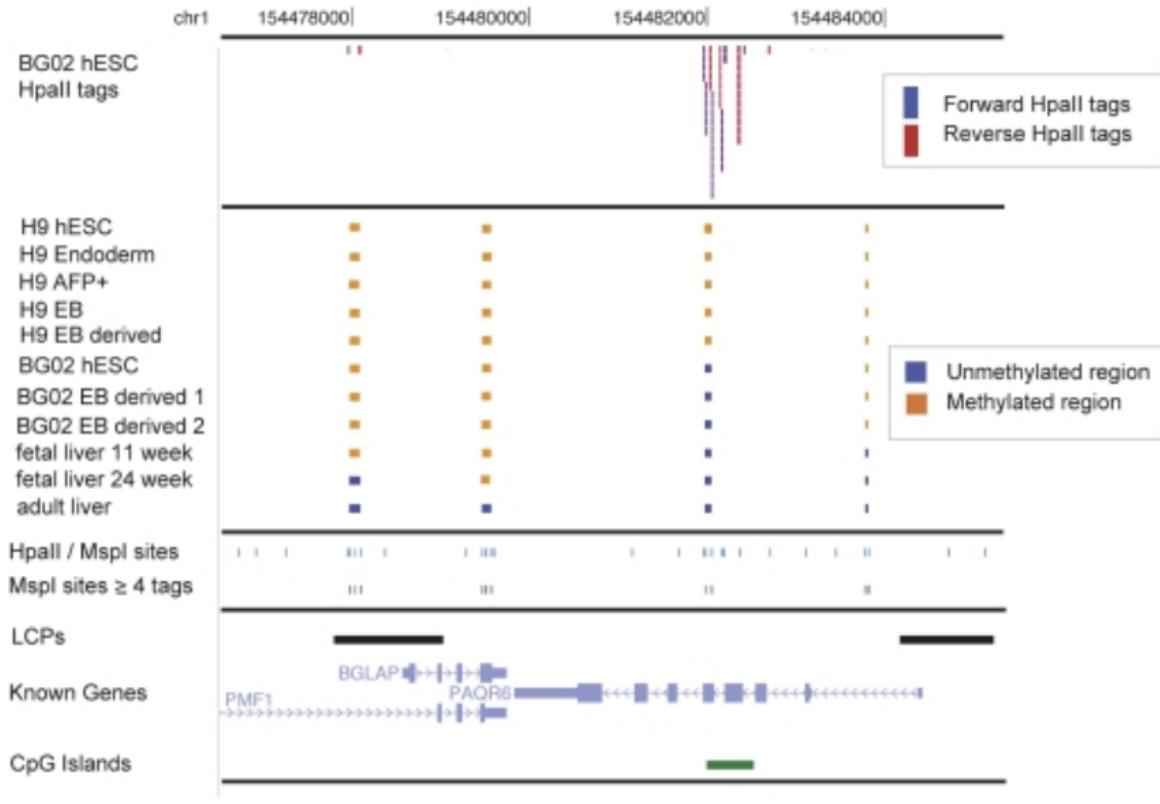
methylKit is an R package for DNA methylation analysis and annotation from high-throughput bisulfite sequencing. The package is designed to deal with sequencing data from RRBS and its variants, but also target-capture methods such as Agilent SureSelect methyl-seq. In addition, methylKit can deal with base-pair resolution data for 5hmC obtained from Tab-seq or oxBs-seq. It can also handle whole-genome bisulfite sequencing data if proper input format is provided.

## Genome Research, 2009

To investigate the role of DNA methylation during human development, we developed Methyl-seq, a method that assays DNA methylation at more than 90,000 regions throughout the genome. Performing Methyl-seq on human embryonic stem cells (hESCs), their derivatives, and human tissues allowed us to identify several trends during hESC and *in vivo* liver differentiation. Taken together, our results indicate that hESC differentiation has a unique DNA methylation signature that may not be indicative of *in vivo* differentiation.

Brunner AL, Johnson DS, Kim SW, Valouev A, Reddy TE, Neff NF, et al. Distinct DNA methylation patterns characterize differentiated human embryonic stem cells and developing human fetal liver. *Genome Res.* 2009;19: 1044–1056.

# 表观遗传学 | Methyl-Seq | 实例



# Powered by



T<sub>E</sub>X L<sup>A</sup>T<sub>E</sub>X X<sub>E</sub>T<sub>E</sub>X Beamer

