

系统生物学

天津医科大学
生物医学工程与技术学院

2016-2017 学年上学期 (秋)
2013 级生信班

第二章 基因组学

伊现富 (Yi Xianfu)

天津医科大学 (TJMU)
生物医学工程与技术学院

2016 年 9 月



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库

● 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题

教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

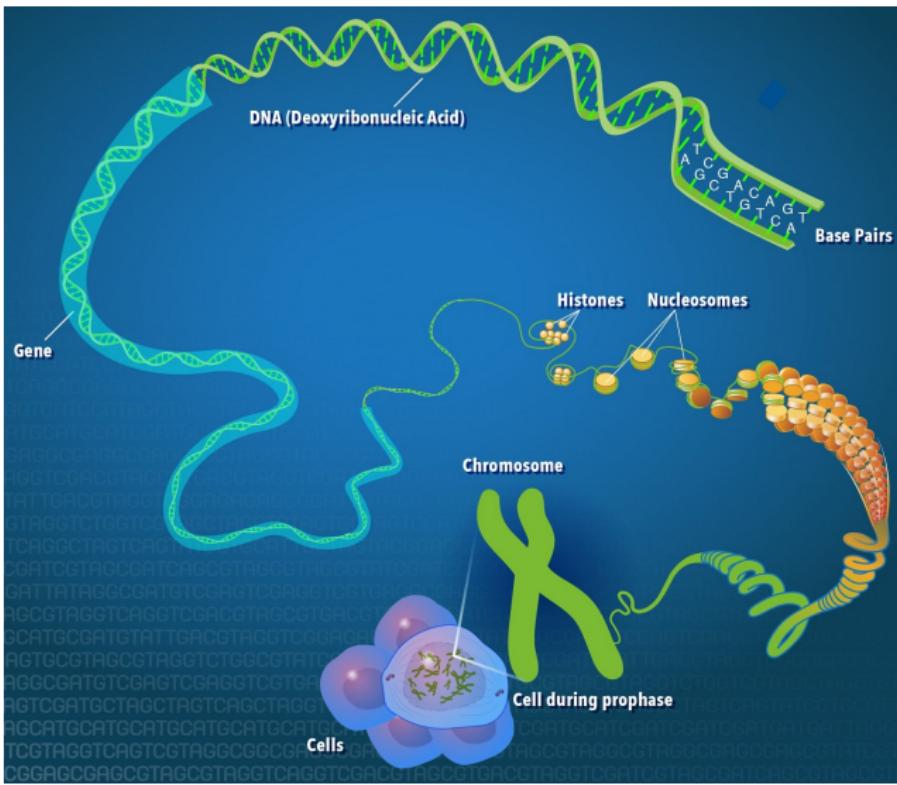
- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



基因组学 | 概述



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。



基因一词来自希腊语，意思为“生”。是指携带有遗传信息的DNA序列，是控制性状的基本遗传单位，亦即一段具有功能性的DNA序列。基因通过指导蛋白质的合成来表现所携带的遗传信息，从而控制生物个体的性状（差异）表现。人类约有两万至两万五千个基因。

染色体在体细胞中是成对存在的，每条染色体上都带有一定数量的基因。一个基因在细胞有丝分裂时有两个对应的位点，称为等位基因，分别来自父与母。依所携带性状的表现，又可分为显性基因和隐性基因。

一般来说，同一生物体中的每个细胞都含有相同的基因，但并不是每个细胞中的所有基因携带的遗传信息都会被表现出来。职司不同功能的细胞中，活化而表现的基因也不同。



性状 (《漫画玩转遗传学》)

- 棕色眼睛，蓝色眼睛
- 色觉，色盲
- 头发浓密，秃顶
- 卷起舌头，无法卷起舌头
- 多指，五指



在生物学中，一个生物体的基因组是指包含在该生物的 DNA（部分病毒是 RNA）中的全部遗传信息，又称基因体（genome）。基因组包括基因和非编码 DNA。

1920 年，德国汉堡大学植物学教授汉斯·温克勒（Hans Winkler）首次使用基因组这一名词。

更精确地讲，一个生物体的基因组是指一套染色体中的完整的 DNA 序列。例如，生物个体体细胞中的二倍体由两套染色体组成，其中一套 DNA 序列就是一个基因组。



基因组一词可以特指整套核 DNA (例如, 核基因组), 也可以用于包含自己 DNA 序列的细胞器基因组, 如粒线体基因组或叶绿体基因组。

当人们说一个有性生殖物种的基因组正在测序时, 通常是指测定一套常染色体和两种性染色体的序列, 这样来代表可能的两种性别。即使在只有一种性别的物种中, “一套基因组序列” 可能也综合了来自不同个体的染色体。

通常使用中, “遗传组成” 一词有时在交流中即指某特定个体或物种的基因组。

对相关物种全部基因组性质的研究通常被称为基因组学, 该学科与遗传学不同, 后者一般研究单个或一组基因的性质。



对于像人类这样的脊椎动物，基因组通常指的只是染色体 DNA。因此，尽管人类线粒体里包含了基因，但这些基因并不作为基因组的一部分。事实上，有时候称线粒体拥有自己的基因组，通常叫做**线粒体基因组**。而在叶绿体中的被称为**叶绿体基因组**。



- 1976 年, 瓦尔特·菲尔斯 (比利时根特大学), RNA 病毒噬菌体 MS2 【第一个完整测序的基因组】
- 1977 年, 弗雷德里克·桑格, Φ -X174 噬菌体 【第一个完成测序的 DNA 基因组】
- 1995 年, The Institute for Genomic Research 团队, 流感嗜血杆菌 (*Haemophilus influenzae*) 【第一个被测序的细菌基因组】
- 1996 年, 酿酒酵母 (*Saccharomyces cerevisiae*) 【第一个被测序的真核生物基因组】
- 1996 年, The Institute for Genomic Research 团队, 詹氏甲烷球菌 (*Methanococcus jannaschii*) 【第一个被测序的古菌基因组】
- 1998 年, 秀丽隐杆线虫 (*Caenorhabditis elegans*) 【第一个被测序的多细胞生物基因组】
- 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



- 1976 年, 瓦尔特·菲尔斯 (比利时根特大学), RNA 病毒噬菌体 MS2 【第一个完整测序的基因组】
- 1977 年, 弗雷德里克·桑格, Φ -X174 噬菌体 【第一个完成测序的 DNA 基因组】
- 1995 年, The Institute for Genomic Research 团队, 流感嗜血杆菌 (*Haemophilus influenzae*) 【第一个被测序的细菌基因组】
- 1996 年, 酿酒酵母 (*Saccharomyces cerevisiae*) 【第一个被测序的真核生物基因组】
- 1996 年, The Institute for Genomic Research 团队, 詹氏甲烷球菌 (*Methanococcus jannaschii*) 【第一个被测序的古菌基因组】
- 1998 年, 秀丽隐杆线虫 (*Caenorhabditis elegans*) 【第一个被测序的多细胞生物基因组】
- 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



- 1976 年, 瓦尔特·菲尔斯 (比利时根特大学), RNA 病毒噬菌体 MS2 【第一个完整测序的基因组】
- 1977 年, 弗雷德里克·桑格, Φ -X174 噬菌体 【第一个完成测序的 DNA 基因组】
- 1995 年, The Institute for Genomic Research 团队, 流感嗜血杆菌 (*Haemophilus influenzae*) 【第一个被测序的细菌基因组】
- 1996 年, 酿酒酵母 (*Saccharomyces cerevisiae*) 【第一个被测序的真核生物基因组】
- 1996 年, The Institute for Genomic Research 团队, 詹氏甲烷球菌 (*Methanococcus jannaschii*) 【第一个被测序的古菌基因组】
- 1998 年, 秀丽隐杆线虫 (*Caenorhabditis elegans*) 【第一个被测序的多细胞生物基因组】
- 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



- 1976 年, 瓦尔特·菲尔斯 (比利时根特大学), RNA 病毒噬菌体 MS2 【第一个完整测序的基因组】
- 1977 年, 弗雷德里克·桑格, Φ -X174 噬菌体 【第一个完成测序的 DNA 基因组】
- 1995 年, The Institute for Genomic Research 团队, 流感嗜血杆菌 (*Haemophilus influenzae*) 【第一个被测序的细菌基因组】
- 1996 年, 酿酒酵母 (*Saccharomyces cerevisiae*) 【第一个被测序的真核生物基因组】
- 1996 年, The Institute for Genomic Research 团队, 詹氏甲烷球菌 (*Methanococcus jannaschii*) 【第一个被测序的古菌基因组】
- 1998 年, 秀丽隐杆线虫 (*Caenorhabditis elegans*) 【第一个被测序的多细胞生物基因组】
- 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



- 1976 年, 瓦尔特·菲尔斯 (比利时根特大学), RNA 病毒噬菌体 MS2 【第一个完整测序的基因组】
- 1977 年, 弗雷德里克·桑格, Φ -X174 噬菌体 【第一个完成测序的 DNA 基因组】
- 1995 年, The Institute for Genomic Research 团队, 流感嗜血杆菌 (*Haemophilus influenzae*) 【第一个被测序的细菌基因组】
- 1996 年, 酿酒酵母 (*Saccharomyces cerevisiae*) 【第一个被测序的真核生物基因组】
- 1996 年, The Institute for Genomic Research 团队, 詹氏甲烷球菌 (*Methanococcus jannaschii*) 【第一个被测序的古菌基因组】
- 1998 年, 秀丽隐杆线虫 (*Caenorhabditis elegans*) 【第一个被测序的多细胞生物基因组】
- 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



- 1976 年, 瓦尔特·菲尔斯 (比利时根特大学), RNA 病毒噬菌体 MS2 【第一个完整测序的基因组】
- 1977 年, 弗雷德里克·桑格, Φ -X174 噬菌体 【第一个完成测序的 DNA 基因组】
- 1995 年, The Institute for Genomic Research 团队, 流感嗜血杆菌 (*Haemophilus influenzae*) 【第一个被测序的细菌基因组】
- 1996 年, 酿酒酵母 (*Saccharomyces cerevisiae*) 【第一个被测序的真核生物基因组】
- 1996 年, The Institute for Genomic Research 团队, 詹氏甲烷球菌 (*Methanococcus jannaschii*) 【第一个被测序的古菌基因组】
- 1998 年, 秀丽隐杆线虫 (*Caenorhabditis elegans*) 【第一个被测序的多细胞生物基因组】
- 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



- 1976 年, 瓦尔特·菲尔斯 (比利时根特大学), RNA 病毒噬菌体 MS2 【第一个完整测序的基因组】
- 1977 年, 弗雷德里克·桑格, Φ -X174 噬菌体 【第一个完成测序的 DNA 基因组】
- 1995 年, The Institute for Genomic Research 团队, 流感嗜血杆菌 (*Haemophilus influenzae*) 【第一个被测序的细菌基因组】
- 1996 年, 酿酒酵母 (*Saccharomyces cerevisiae*) 【第一个被测序的真核生物基因组】
- 1996 年, The Institute for Genomic Research 团队, 詹氏甲烷球菌 (*Methanococcus jannaschii*) 【第一个被测序的古菌基因组】
- 1998 年, 秀丽隐杆线虫 (*Caenorhabditis elegans*) 【第一个被测序的多细胞生物基因组】
- 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



- 1976 年, 瓦尔特·菲尔斯 (比利时根特大学), RNA 病毒噬菌体 MS2 【第一个完整测序的基因组】
- 1977 年, 弗雷德里克·桑格, Φ -X174 噬菌体 【第一个完成测序的 DNA 基因组】
- 1995 年, The Institute for Genomic Research 团队, 流感嗜血杆菌 (*Haemophilus influenzae*) 【第一个被测序的细菌基因组】
- 1996 年, 酿酒酵母 (*Saccharomyces cerevisiae*) 【第一个被测序的真核生物基因组】
- 1996 年, The Institute for Genomic Research 团队, 詹氏甲烷球菌 (*Methanococcus jannaschii*) 【第一个被测序的古菌基因组】
- 1998 年, 秀丽隐杆线虫 (*Caenorhabditis elegans*) 【第一个被测序的多细胞生物基因组】
- 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



基因组构成

基因组构成 (genome composition) 用于描述一个单倍体基因组的组成, 包括基因组大小、非重复 DNA 和重复 DNA 所占的比重等。

当讨论基因组的构成时, 首先要区别的是原核基因组还是真核基因组, 两者在基因组组成上有很大的不同。

非重复 DNA 比重

非重复 DNA 的总长除以基因组大小即为非重复 DNA 比重。蛋白质编码基因和非编码 RNA 基因一般都是非重复的 DNA。

不同生物中的非重复 DNA 的比重会有很大不同。更大的基因组并不意味着更多的基因, 随着高等真核生物的基因组大小的增加, 非重复 DNA 的比重相应减少。

基因组构成

基因组构成 (genome composition) 用于描述一个单倍体基因组的组成，包括基因组大小、非重复 DNA 和重复 DNA 所占的比重等。

当讨论基因组的构成时，首先要区别的是原核基因组还是真核基因组，两者在基因组组成上有很大的不同。

非重复 DNA 比重

非重复 DNA 的总长除以基因组大小即为非重复 DNA 比重。蛋白质编码基因和非编码 RNA 基因一般都是非重复的 DNA。

不同生物中的非重复 DNA 的比重会有很大不同。更大的基因组并不意味着更多的基因，随着高等真核生物的基因组大小的增加，非重复 DNA 的比重相应减少。

基因组大小

基因组大小是指一种生物单倍体基因组的全部 DNA 碱基对数。

在原核生物和低等真核生物中，基因组大小与生物形态的复杂性基本呈正相关关系；但是在软体动物以及其他更高等的真核生物中，这种相关性就不存在了。这一现象可能是由基因组中的重复 DNA 引起的。

C 值

C 值 (C-value) 是指真核生物细胞中，单倍细胞核（受精卵或二倍体体细胞中的一半量）里所拥有的 DNA 含量。有时候 C 值和基因组大小两个用词可替换使用。

一个物种单倍体基因组的 DNA 含量是相对恒定的，它通常称为该物种 DNA 的 C 值。

基因组大小

基因组大小是指一种生物单倍体基因组的全部 DNA 碱基对数。

在原核生物和低等真核生物中，基因组大小与生物形态的复杂性基本呈正相关关系；但是在软体动物以及其他更高等的真核生物中，这种相关性就不存在了。这一现象可能是由基因组中的重复 DNA 引起的。

C 值

C 值 (C-value) 是指真核生物细胞中，单倍细胞核（受精卵或二倍体体细胞中的一半量）里所拥有的 DNA 含量。有时候 C 值和基因组大小两个用词可替换使用。

一个物种单倍体基因组的 DNA 含量是相对恒定的，它通常称为该物种 DNA 的 C 值。

C 值悖论 (C-value paradox)

指一个关于真核生物各物种的基因组大小差异的难题，也就是生物的 C 值（或基因组大小）并不与生物复杂程度相关的现象。

C 值谜 (C-value enigma)

In general terms, the C-value enigma relates to the issue of variation in the amount of non-coding DNA found within the genomes of different eukaryotes.



C 值悖论 (C-value paradox)

指一个关于真核生物各物种的基因组大小差异的难题，也就是生物的 C 值（或基因组大小）并不与生物复杂程度相关的现象。

C 值谜 (C-value enigma)

In general terms, the C-value enigma relates to the issue of variation in the amount of non-coding DNA found within the genomes of different eukaryotes.



C 值谜 (C-value enigma)

The C-value enigma, unlike the older C-value paradox, is explicitly defined as a series of independent but equally important component questions, including:

- What types of non-coding DNA are found in different eukaryotic genomes, and in what proportions?
- From where does this non-coding DNA come, and how is it spread and/or lost from genomes over time?
- What effects, or perhaps even functions, does this non-coding DNA have for chromosomes, nuclei, cells, and organisms?
- Why do some species exhibit remarkably streamlined chromosomes, while others possess massive amounts of non-coding DNA?

基因组学 | 概述 | 基因组 | 补遗

类型	生物	学名	基因组大小(碱基对)	注
病毒	猪圆环病毒I型		1,759	已知最小的基因组 [11]
病毒	猿猴病毒SV40		5,224	[12]
病毒	噬菌体Φ-X174		5,386	最早完成测序的DNA基因组 [13]
病毒	人类免疫缺陷病毒HIV		9,749	[14]
病毒	噬菌体λ		48,502	常作为重组DNA的克隆载体。 [15] [16] [17]
细菌	大肠杆菌	<i>Escherichia coli</i>	4.6Mb	[18]
变形虫	无恒变形虫	<i>Amoeba dubia</i>	670Gb	已知的最大基因组 [19] (但有争议) [20]
植物	贝母属一种	<i>Fritillary assyriaca</i>	130Gb	
真菌	酿酒酵母	<i>Saccharomyces cerevisiae</i>	12.1Mb	第一个测序的真核生物基因组, 完成于1996年 [21]
线虫	咖啡短体线虫	<i>Pratylenchus coffeae</i>	20Mb	已知最小的动物基因组 [22]
线虫	秀丽隐杆线虫	<i>Caenorhabditis elegans</i>	100Mb	第一个测序的多细胞生物基因组, 完成于1998年12月 [23]
昆虫	黑腹果蝇	<i>Drosophila melanogaster</i>	130Mb	[24]
哺乳动物	小家鼠	<i>Mus musculus</i>	2.7Gb	[25]
哺乳动物	人	<i>Homo sapiens</i>	3.2Gb	[26] [27]
鱼类	金娃娃 (一种河豚)	<i>Tetraodon nigroviridis</i>	385Mb	已知最小的脊椎动物基因组约为340Mb [28] [29] -385Mb [30]
鱼类	石花肺鱼	<i>Protopterus aethiopicus</i>	130Gb	已知最大的脊椎动物基因组



基因组演化

基因组不仅仅是生物基因的集合，对其研究和比较能获得生物演化信息的更多细节。

一些基因组性质如“染色体数”（核型）、基因组大小、基因顺序、密码子偏好性与 GC 含量能反映出现存生物的许多基因组演化信息。



基因组学

基因组学 (genomics)，或基因体学，是研究生物基因组和如何利用基因的一门学问。

基因组学的主要工具和方法包括：生物信息学，遗传分析，基因表达测量和基因功能鉴定。

特点

基因组学的特点是强调进行细胞中全部基因及非编码区的整体性考查和系统性研究，从而全面揭示基因与基因间的相互关系、基因与非编码序列的关系、基因与基因组的相互关系。



基因组学

基因组学 (genomics)，或基因体学，是研究生物基因组和如何利用基因的一门学问。

基因组学的主要工具和方法包括：生物信息学，遗传分析，基因表达测量和基因功能鉴定。

特点

基因组学的特点是强调进行细胞中全部基因及非编码区的整体性考查和系统性研究，从而全面揭示基因与基因间的相互关系、基因与非编码序列的关系、基因与基因组的相互关系。



组学

“组”在基因组一词中，意指一个物种的“全部”遗传组成。由于诸如基因组测序这样的大规模定量生物项目的成功，“组”的这个意义的使用已经扩展到其他相关领域。例如，蛋白质组指的是一个物种组织或细胞内的全部蛋白质。

基因组分析

基因组项目涉及三个部分：

- DNA 测序
- 该序列的组件生成原有染色体的表示法
- 该表示法的注释和分析



组学

“组”在基因组一词中，意指一个物种的“全部”遗传组成。由于诸如基因组测序这样的大规模定量生物项目的成功，“组”的这个意义的使用已经扩展到其他相关领域。例如，蛋白质组指的是一个物种组织或细胞内的全部蛋白质。

基因组分析

基因组项目涉及三个部分：

- DNA 测序
- 该序列的组件生成原有染色体的表示法
- 该表示法的注释和分析



- 1977 年，噬菌体 Φ -X174 (5,368 碱基对) 完全测序，成为第一个测定的基因组
- 1995 年，嗜血流感菌 (Haemophilus influenzae, 1.8Mb) 测序完成，是第一个测定的自由生活物种
- 2001 年，人类基因组计划公布了人类基因组草图，为基因组学研究揭开新的一页
- 2012 年，千人基因组计划



- 1977 年，噬菌体 Φ -X174 (5,368 碱基对) 完全测序，成为第一个测定的基因组
- 1995 年，嗜血流感菌 (Haemophilus influenzae, 1.8Mb) 测序完成，是第一个测定的自由生活物种
- 2001 年，人类基因组计划公布了人类基因组草图，为基因组学研究揭开新的一页
- 2012 年，千人基因组计划



- 1977 年，噬菌体 Φ -X174 (5,368 碱基对) 完全测序，成为第一个测定的基因组
- 1995 年，嗜血流感菌 (Haemophilus influenzae, 1.8Mb) 测序完成，是第一个测定的自由生活物种
- 2001 年，人类基因组计划公布了人类基因组草图，为基因组学研究揭开新的一页
- 2012 年，千人基因组计划



- 1977 年，噬菌体 Φ -X174 (5,368 碱基对) 完全测序，成为第一个测定的基因组
- 1995 年，嗜血流感菌 (Haemophilus influenzae, 1.8Mb) 测序完成，是第一个测定的自由生活物种
- 2001 年，人类基因组计划公布了人类基因组草图，为基因组学研究揭开新的一页
- 2012 年，千人基因组计划



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



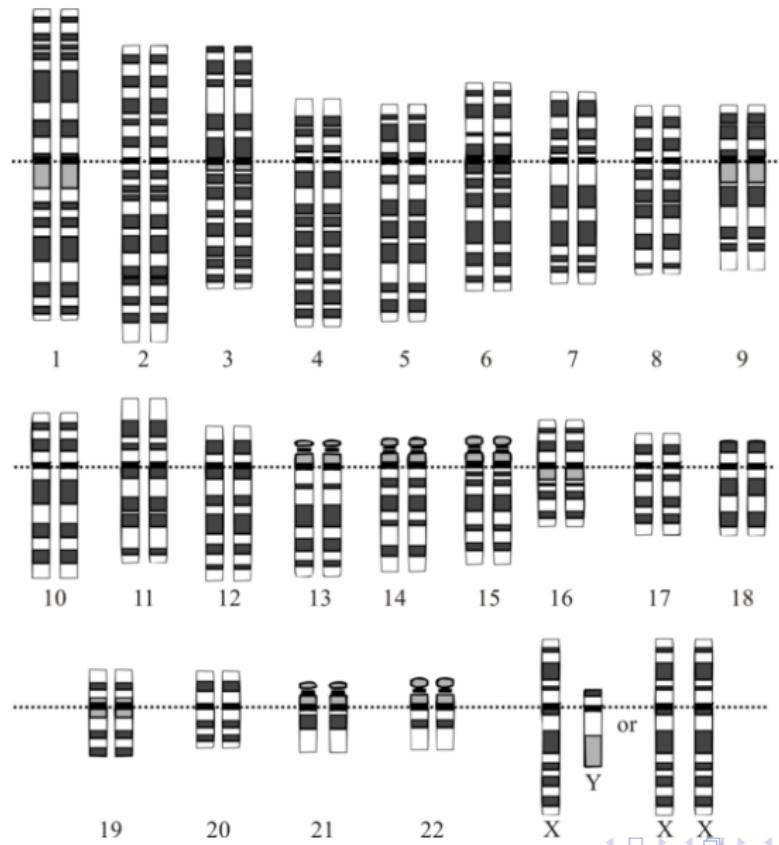
人类基因组

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 | 概述 | 人类基因组



染色体

人类拥有 23 对不同的染色体，其中 22 对属于常染色体，另外还有 1 对能够决定性别的性染色体，分别是 X 染色体与 Y 染色体。1 号到 22 号染色体的编号顺序，大致符合他们由大到小的尺寸排列。最大的染色体约含有 2 亿 5 千万个碱基对，最小的则约有 3800 万个碱基对。

在人类个体的体细胞中，通常含有来自亲代的 1 到 22 对体染色体，再加上来自母亲的 X 染色体，以及来自父亲的 X 或 Y 染色体，总共是 46 个（23 对）染色体。科学家将这些染色体分为 7 组：1 号到 3 号是 A 组；4 号与 5 号是 B 组；X 染色体以及 6 号到 12 号是 C 组；13 号到 15 号是 D 组；16 号到 18 号是 E 组；19 号与 20 号是 F 组；21 号、22 号与 Y 染色体是 G 组。对于一般人类来说，每个细胞核内只有两套染色体。



基因

人体内估计约有 20000 到 25000 个蛋白质编码基因。虽然人类的基因数量比起某些较为原始的生物更少，但是在人类细胞中使用了大量的选择性剪接 (alternative splicing)，这使得一个基因能够制造出多种不同的蛋白质，人类的蛋白质组规模也更加庞大。

大多数人类基因拥有许多的外显子，且人类的内含子比位于其两端的外显子更长。这些基因参差不齐地分布在染色体中，每一个染色体皆含有一些基因较多的区段与基因较少的区段。这些区段的差异，则与染色体带 (chromosome bands) 及 GC 含量相关。基因密度所显现的非随机模式之涵义与重要性尚未明了。



人类与其他物种的基因组比较(大约) [5][4]		
物种	碱基对数量	基因数量
<i>Mycoplasma genitalium</i> 霉浆菌(生殖器支原体)	580,000	500
<i>Streptococcus pneumoniae</i> 肺炎双球菌	2,200,000	2,300
<i>Haemophilus influenzae</i> 流感嗜血杆菌	1,830,140	1,700
<i>Escherichia coli</i> 大肠杆菌	4,600,000	4,400
<i>Saccharomyces cerevisiae</i> 酿酒酵母	12,000,000	5,538
<i>Caenorhabditis elegans</i> 秀丽隐杆线虫	97,000,000	18,250
<i>Arabidopsis thaliana</i> 阿拉伯芥(拟南芥)	125,000,000	25,500
<i>Drosophila melanogaster</i> 黑腹果蝇	180,000,000	13,350
<i>Oryza sativa</i> 亚洲稻	466,000,000	45,000-55,000
<i>Mus musculus</i> 家鼠	2,500,000,000	29,000
<i>Homo sapiens</i> 人类	2,900,000,000	27,000



功能未知区域

蛋白质编码序列（也就是外显子）在人类基因组中少于 1.5%。在基因与调控序列之外，仍然有许多功能未知的广大区域。科学家估计这些区域在人类基因组中约占有 97%，其中许多是属于重复序列（Repeated sequence）、转位子（transposon）与伪基因（pseudogene）。除此之外，还有大量序列不属于上述的已知分类。

人类基因组内大量功能未知的序列，是目前科学的研究的重点之一。



变异

大多数对于人类遗传变异的研究集中在单核苷酸多态性 (single nucleotide polymorphisms, SNP)，也就是 DNA 中的个别碱基变换。在人类的真染色质 (富含基因的染色质) 中，平均每 100 到 1000 个碱基会出现 1 个 SNP，不过密度并不均匀。由于 SNP 的存在，如“所有人类的基因有 99% 都是相同的”这样的说法并不精确。国际人类基因组单体型图计划 (International HapMap Project)，便是为了要将人类基因组中的 SNP 变异作编录，而组成的一个大规模合作计划。

研究人员发现在人类与其他哺乳类 DNA 序列中的拷贝数变异 (copy number variation, CNV)，可能非常重要。拷贝数变异又称为拷贝数多型性 (copy number polymorphisms, CNP)，是缺失 (deletion)、插入 (insertion)、重复 (duplication)，以及复杂多位置变异 (complex multi-site variants) 的合称，在所有人类以及其他已测序的哺乳动物中皆可发现。

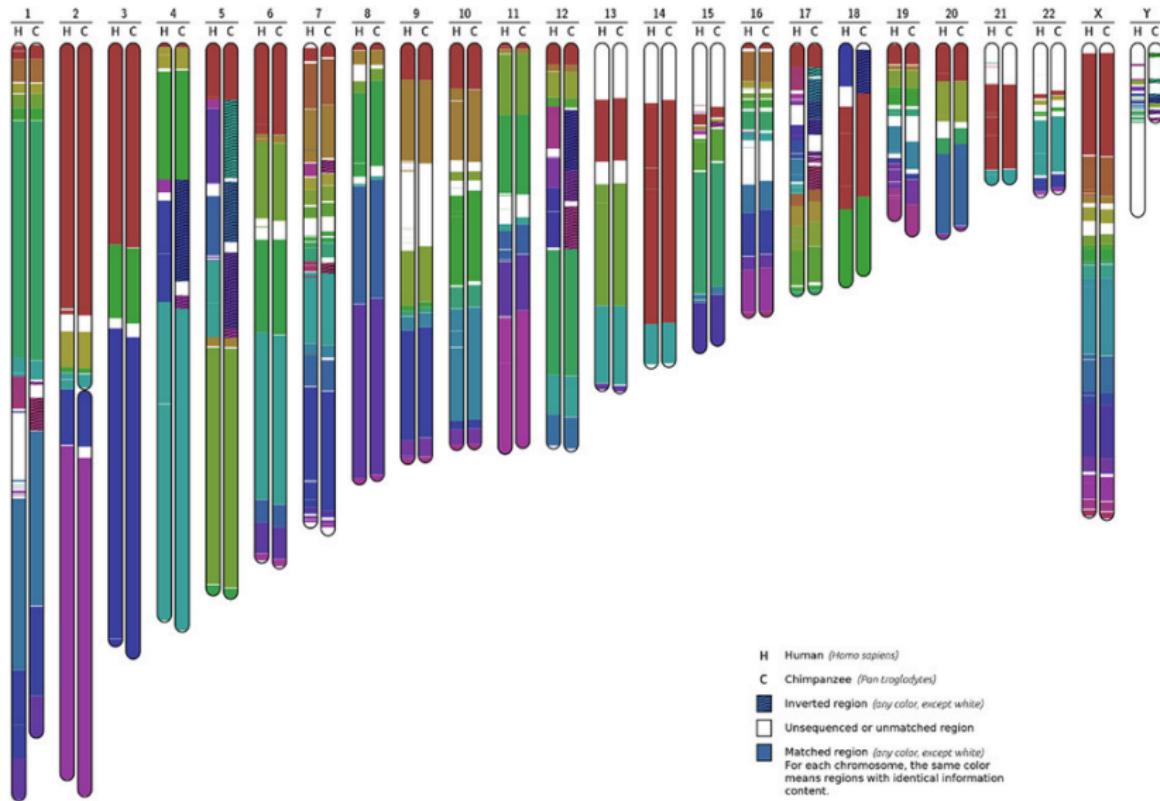
演化

比较基因组学（comparative genomics）对于哺乳类基因组的研究显示，人类与大约两亿年前就已经分化的各物种相比，有大约 5% 的比例在人类基因组中保留了下来，其中包含许多的基因与调控序列。而且人类与大多数已知的脊椎动物间，也享有了一些相同的基因。

黑猩猩的基因组与人类的基因组之间，有 98.77% 是相似的。而平均每一个属于人类的标准蛋白质编码基因，只与属于黑猩猩的同源基因相差两个氨基酸；并且有将近三分之一的人类基因与黑猩猩的同源基因，能够翻译出相同的蛋白质。人类的 2 号染色体，是人类与黑猩猩基因组之间的主要差异，它是由黑猩猩的染色体 2A 与 2B 融合而成的。

人类在最近的演化过程中失去了嗅觉受体基因，这解释了为何人类比起其他的哺乳动物来说，拥有较差的嗅觉。演化上的证据显示，人类与某些灵长类所拥有的彩色视觉，降低了这些物种对于嗅觉能力的需求。

基因组学 | 概述 | 人类基因组 | 补遗



线粒体基因组

大多数的基因存在于细胞核中，但是细胞中一个称为线粒体的细胞器，也拥有自己的基因组。线粒体基因组在线粒体疾病（mitochondrial disease）中具有一定的重要性；而且这些基因也可以用来研究人类的演化。线粒体位于细胞质中，当人类的精子与卵子结合时，源自母亲（女性）的卵子提供了绝大多数的细胞质，因此人类细胞中的线粒体基因皆是来自母亲。

由于线粒体缺乏用来检查复制错误的能力，因此**线粒体 DNA (mtDNA)** 的变异速率比细胞核 DNA（一般所指的 DNA）更快。线粒体的突变速率快了 20 倍，这使 mtDNA 能够用来较为精确地追溯出母系祖先。研究族群中的 mtDNA，也能使人们得知此族群过去的迁移路径。



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)



人类基因组计划 (human genome project, HGP) 是一项规模宏大，跨国跨学科的科学探索工程。其宗旨在于测定组成人类染色体（指单倍体）的 30 亿个碱基对形成的核苷酸序列，从而绘制人类基因组图谱，并且辨识其载有的基因，达到破译人类遗传信息的最终目的。

该计划起始于 1990 年，已基本测序了人类的所有基因。截止到 2005 年，人类基因组计划的测序工作已经基本完成（92%）。但关于其功能的细节，则仍有许多研究在进行中。

其中，2001 年人类基因组工作草图的发表被认为是人类基因组计划成功 **【?】** 的里程碑。



- 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，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



- 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，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



- 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，詹姆士·华生，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



- 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，詹姆士·华生，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



- 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，詹姆士·华生，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



- 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，詹姆士·华生，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



- 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，詹姆士·华生，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，且许多片断的顺序方位并没有得到确定）



- 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，且许多片断的顺序方位并没有得到确定）



- 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，且许多片断的顺序方位并没有得到确定）

- 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，且许多片断的顺序方位并没有得到确定）

- 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



- 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



- 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



- 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



- 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

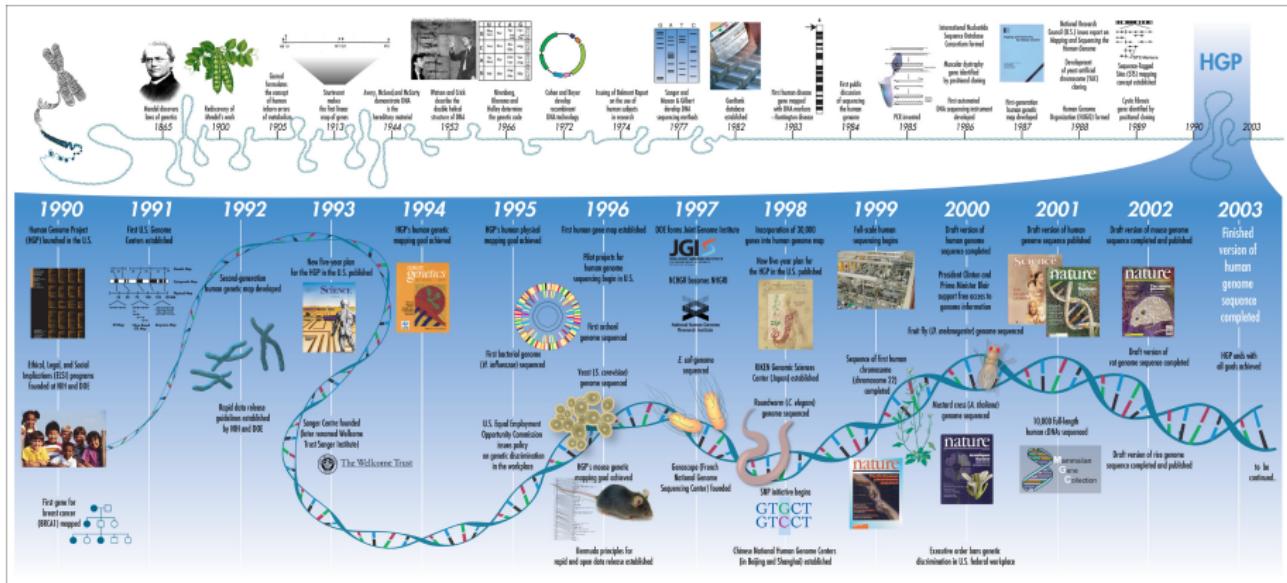


- 1994 年，中国的人类基因组计划启动
- 1998 年，中国南方基因组中心成立，中国科学院遗传研究所人类基因组中心成立
- 1999 年，北京华大基因研究中心（华大基因）成立，北方基因组中心成立
- 1998 年 3 月，中美港科学家合作，成功地将与华人和鼻咽癌有关的肿瘤抑制基因定位于人类第 3 号染色体的短臂 3p21.3 位点
- 1999 年 6 月 26 日，中国科学院遗传研究所人类基因组中心向美国国立卫生研究院（NIH）的国际人类基因组计划（HGP）递交加入申请。HGP 在网上公布中国注册加入国际测序组织，中国成为继美、英、日、德、法后第六个加入该组织的国家
- 1999 年 11 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



- 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

基因组学 | 概述 | 人类基因组计划 | 事件



遗传图谱的绘制 遗传图谱主要是用遗传标签来确定基因在染色体上的排列。

物理图谱的绘制 物理图谱是通过序列标签位点对构成基因组的 DNA 分子进行测定，从而对某基因所相对之遗传信息及其在染色体上的相对位置做一线性排列。

序列测定 通过测序得到基因组的序列，是一般意义上的人类基因组计划。

辨别序列中的个体差异 人类基因组计划只是为未来鉴定不同个体间基因组差异做一些基础的框架性工作。当前主要工作在于鉴定不同个体间包含的单核苷酸多态性。

基因鉴定 以获得全长的人类 cDNA 文库为目标。

基因的功能性分析 对海量的数据进行标注，分析注释序列。另一个目标是研发出更快更有效的方法来进行 DNA 测序和序列分析，并把这一技术加以产业化。



遗传图谱的绘制 遗传图谱主要是用遗传标签来确定基因在染色体上的排列。

物理图谱的绘制 物理图谱是通过序列标签位点对构成基因组的 DNA 分子进行测定，从而对某基因所相对之遗传信息及其在染色体上的相对位置做一线性排列。

序列测定 通过测序得到基因组的序列，是一般意义上的人类基因组计划。

辨别序列中的个体差异 人类基因组计划只是为未来鉴定不同个体间基因组差异做一些基础的框架性工作。当前主要工作在于鉴定不同个体间包含的单核苷酸多态性。

基因鉴定 以获得全长的人类 cDNA 文库为目标。

基因的功能性分析 对海量的数据进行标注，分析注释序列。另一个目标是研发出更快更有效的方法来进行 DNA 测序和序列分析，并把这一技术加以产业化。



遗传图谱的绘制 遗传图谱主要是用遗传标签来确定基因在染色体上的排列。

物理图谱的绘制 物理图谱是通过序列标签位点对构成基因组的 DNA 分子进行测定，从而对某基因所相对之遗传信息及其在染色体上的相对位置做一线性排列。

序列测定 通过测序得到基因组的序列，是一般意义上的人类基因组计划。

辨别序列中的个体差异 人类基因组计划只是为未来鉴定不同个体间基因组差异做一些基础的框架性工作。当前主要工作在于鉴定不同个体间包含的单核苷酸多态性。

基因鉴定 以获得全长的人类 cDNA 文库为目标。

基因的功能性分析 对海量的数据进行标注，分析注释序列。另一个目标是研发出更快更有效的方法来进行 DNA 测序和序列分析，并把这一技术加以产业化。



遗传图谱的绘制 遗传图谱主要是用遗传标签来确定基因在染色体上的排列。

物理图谱的绘制 物理图谱是通过序列标签位点对构成基因组的 DNA 分子进行测定，从而对某基因所相对之遗传信息及其在染色体上的相对位置做一线性排列。

序列测定 通过测序得到基因组的序列，是一般意义上的人类基因组计划。

辨别序列中的个体差异 人类基因组计划只是为未来鉴定不同个体间基因组差异做一些基础的框架性工作。当前主要工作在于鉴定不同个体间包含的单核苷酸多态性。

基因鉴定 以获得全长的人类 cDNA 文库为目标。

基因的功能性分析 对海量的数据进行标注，分析注释序列。另一个目标是研发出更快更有效的方法来进行 DNA 测序和序列分析，并把这一技术加以产业化。



遗传图谱的绘制 遗传图谱主要是用遗传标签来确定基因在染色体上的排列。

物理图谱的绘制 物理图谱是通过序列标签位点对构成基因组的 DNA 分子进行测定，从而对某基因所相对之遗传信息及其在染色体上的相对位置做一线性排列。

序列测定 通过测序得到基因组的序列，是一般意义上的人类基因组计划。

辨别序列中的个体差异 人类基因组计划只是为未来鉴定不同个体间基因组差异做一些基础的框架性工作。当前主要工作在于鉴定不同个体间包含的单核苷酸多态性。

基因鉴定 以获得全长的人类 cDNA 文库为目标。

基因的功能性分析 对海量的数据进行标注，分析注释序列。另一个目标是研发出更快更有效的方法来进行 DNA 测序和序列分析，并把这一技术加以产业化。



遗传图谱的绘制 遗传图谱主要是用遗传标签来确定基因在染色体上的排列。

物理图谱的绘制 物理图谱是通过序列标签位点对构成基因组的 DNA 分子进行测定，从而对某基因所相对之遗传信息及其在染色体上的相对位置做一线性排列。

序列测定 通过测序得到基因组的序列，是一般意义上的人类基因组计划。

辨别序列中的个体差异 人类基因组计划只是为未来鉴定不同个体间基因组差异做一些基础的框架性工作。当前主要工作在于鉴定不同个体间包含的单核苷酸多态性。

基因鉴定 以获得全长的人类 cDNA 文库为目标。

基因的功能性分析 对海量的数据进行标注，分析注释序列。另一个目标是研发出更快更有效的方法来进行 DNA 测序和序列分析，并把这一技术加以产业化。



遗传图谱的绘制 1994 年 9 月，完成了包含 3000 个（原计划为 600 – 1500）标签分辨率为 1-cM（即 1% 重组率）的遗传图谱的绘制。

物理图谱的绘制 1998 年 10 月，完成了包含 52,000 个（原计划为 30,000）序列标签位点的物理图谱的绘制。

序列测定 2003 年 4 月，包含基因序列中的 98%（原预计为 95%）获得了测定，精确度为 99.99%。

辨别序列中的个体差异 至 2003 年 2 月，已有约 3,700,000 个单核苷酸多态性位点得到测定。

基因鉴定 至 2003 年 3 月，已获得 15,000 个全长的人类 cDNA 文库。

基因的功能性分析 已获得开发的技术包括高通量寡聚核苷酸的合成（1994 年）、DNA 微阵列（1996 年）、标准化和消减化 cDNA 文库（1996 年）、真核（酵母）全基因组敲除技术（1999 年）、大型化双杂交定位（2002 年）。



遗传图谱的绘制 1994 年 9 月，完成了包含 3000 个（原计划为 600 – 1500）标签分辨率为 1-cM（即 1% 重组率）的遗传图谱的绘制。

物理图谱的绘制 1998 年 10 月，完成了包含 52,000 个（原计划为 30,000）序列标签位点的物理图谱的绘制。

序列测定 2003 年 4 月，包含基因序列中的 98%（原预计为 95%）获得了测定，精确度为 99.99%。

辨别序列中的个体差异 至 2003 年 2 月，已有约 3,700,000 个单核苷酸多态性位点得到测定。

基因鉴定 至 2003 年 3 月，已获得 15,000 个全长的人类 cDNA 文库。

基因的功能性分析 已获得开发的技术包括高通量寡聚核苷酸的合成（1994 年）、DNA 微阵列（1996 年）、标准化和消减化 cDNA 文库（1996 年）、真核（酵母）全基因组敲除技术（1999 年）、大型化双杂交定位（2002 年）。



遗传图谱的绘制 1994 年 9 月，完成了包含 3000 个（原计划为 600 – 1500）标签分辨率为 1-cM（即 1% 重组率）的遗传图谱的绘制。

物理图谱的绘制 1998 年 10 月，完成了包含 52,000 个（原计划为 30,000）序列标签位点的物理图谱的绘制。

序列测定 2003 年 4 月，包含基因序列中的 98%（原预计为 95%）获得了测定，精确度为 99.99%。

辨别序列中的个体差异 至 2003 年 2 月，已有约 3,700,000 个单核苷酸多态性位点得到测定。

基因鉴定 至 2003 年 3 月，已获得 15,000 个全长的人类 cDNA 文库。

基因的功能性分析 已获得开发的技术包括高通量寡聚核苷酸的合成（1994 年）、DNA 微阵列（1996 年）、标准化和消减化 cDNA 文库（1996 年）、真核（酵母）全基因组敲除技术（1999 年）、大型化双杂交定位（2002 年）。



遗传图谱的绘制 1994 年 9 月，完成了包含 3000 个（原计划为 600 – 1500）标签分辨率为 1-cM（即 1% 重组率）的遗传图谱的绘制。

物理图谱的绘制 1998 年 10 月，完成了包含 52,000 个（原计划为 30,000）序列标签位点的物理图谱的绘制。

序列测定 2003 年 4 月，包含基因序列中的 98%（原预计为 95%）获得了测定，精确度为 99.99%。

辨别序列中的个体差异 至 2003 年 2 月，已有约 3,700,000 个单核苷酸多态性位点得到测定。

基因鉴定 至 2003 年 3 月，已获得 15,000 个全长的人类 cDNA 文库。

基因的功能性分析 已获得开发的技术包括高通量寡聚核苷酸的合成（1994 年）、DNA 微阵列（1996 年）、标准化和消减化 cDNA 文库（1996 年）、真核（酵母）全基因组敲除技术（1999 年）、大型化双杂交定位（2002 年）。



遗传图谱的绘制 1994 年 9 月，完成了包含 3000 个（原计划为 600 – 1500）标签分辨率为 1-cM（即 1% 重组率）的遗传图谱的绘制。

物理图谱的绘制 1998 年 10 月，完成了包含 52,000 个（原计划为 30,000）序列标签位点的物理图谱的绘制。

序列测定 2003 年 4 月，包含基因序列中的 98%（原预计为 95%）获得了测定，精确度为 99.99%。

辨别序列中的个体差异 至 2003 年 2 月，已有约 3,700,000 个单核苷酸多态性位点得到测定。

基因鉴定 至 2003 年 3 月，已获得 15,000 个全长的人类 cDNA 文库。

基因的功能性分析 已获得开发的技术包括高通量寡聚核苷酸的合成（1994 年）、DNA 微阵列（1996 年）、标准化和消减化 cDNA 文库（1996 年）、真核（酵母）全基因组敲除技术（1999 年）、大型化双杂交定位（2002 年）。



遗传图谱的绘制 1994 年 9 月，完成了包含 3000 个（原计划为 600 – 1500）标签分辨率为 1-cM（即 1% 重组率）的遗传图谱的绘制。

物理图谱的绘制 1998 年 10 月，完成了包含 52,000 个（原计划为 30,000）序列标签位点的物理图谱的绘制。

序列测定 2003 年 4 月，包含基因序列中的 98%（原预计为 95%）获得了测定，精确度为 99.99%。

辨别序列中的个体差异 至 2003 年 2 月，已有约 3,700,000 个单核苷酸多态性位点得到测定。

基因鉴定 至 2003 年 3 月，已获得 15,000 个全长的人类 cDNA 文库。

基因的功能性分析 已获得开发的技术包括高通量寡聚核苷酸的合成（1994 年）、DNA 微阵列（1996 年）、标准化和消减化 cDNA 文库（1996 年）、真核（酵母）全基因组敲除技术（1999 年）、大型化双杂交定位（2002 年）。



遗传图谱

遗传图谱 (genetic map) 是利用基因的重组率来做分析，单位是分莫甘 (centimorgan)。这种图谱表现出来的是基因或特定 DNA 片段之间的相对位置，而不是它们各自的绝对位置。

物理图谱

物理图谱 (physical map) 则是 DNA 两点的实际距离，是实际将 DNA 片段排序而得，单位是碱基的数目 (如 Kb, kilobase)。



遗传图谱

遗传图谱 (genetic map) 是利用基因的重组率来做分析，单位是分莫甘 (centimorgan)。这种图谱表现出来的是基因或特定 DNA 片段之间的相对位置，而不是它们各自的绝对位置。

物理图谱

物理图谱 (physical map) 则是 DNA 两点的实际距离，是实际将 DNA 片段排序而得，单位是碱基的数目 (如 Kb, kilobase)。



完成

关于如何界定人类基因组测序完成，有多种定义。根据不同的定义，人类基因组的测序是否完成有不同的看法。曾有多个大众媒体报道人类基因组计划“完成”，而且由国际人类基因组计划所采用的定义，基因组的测序已经完成。

仍有许多区域未获得测序

- 着丝粒含有数百万（可能接近千万）的碱基对，其中的大多数完全没有得到测序
- 染色体末端区域（称为端粒）大都不完整，无法精确地知道在端粒前还有多少序列
- 每个人的基因组中都含有多个包含多基因家族成员的位点
- 还有一些间隙散布于基因组中，部分间隙较大

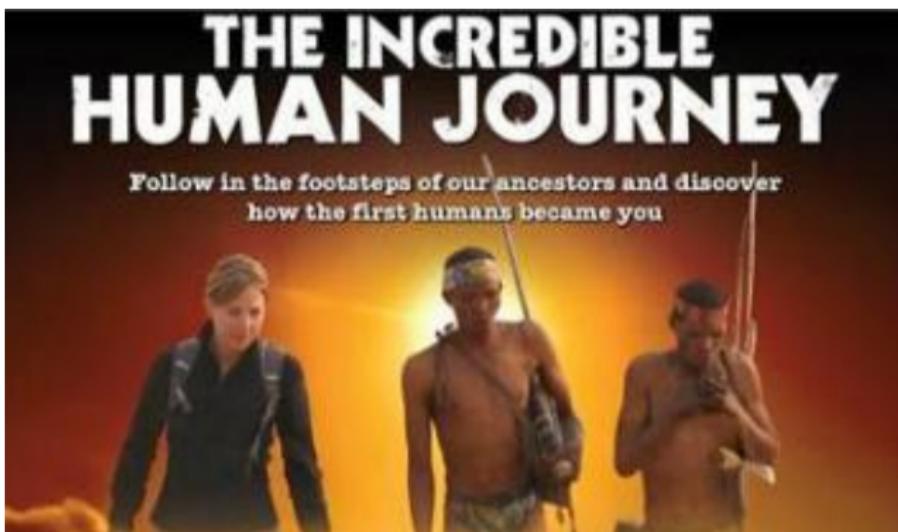
完成

关于如何界定人类基因组测序完成，有多种定义。根据不同的定义，人类基因组的测序是否完成有不同的看法。曾有多个大众媒体报道人类基因组计划“完成”，而且由国际人类基因组计划所采用的定义，基因组的测序已经完成。

仍有许多区域未获得测序

- 着丝粒含有数百万（可能接近千万）的碱基对，其中的大多数完全没有得到测序
- 染色体末端区域（称为端粒）大都不完整，无法精确地知道在端粒前还有多少序列
- 每个人的基因组中都含有多个包含多基因家族成员的位点
- 还有一些间隙散布于基因组中，部分间隙较大

- 基础科研领域：癌症、老年痴呆症等疾病的病因研究，推动新的疗法和新药的开发研究
- 医疗健康/商业价值：基因检测，个性化医疗/精准医疗
- 生物进化研究：揭示了许多重要的生物进化史上的里程碑事件（核糖体的出现，器官的产生，胚胎的发育，脊柱和免疫系统等）
- 人类遗传信息的应用：考古学（走出非洲），犯罪学以及社会执法



- 2004 年，国际人类基因组测序联盟的研究者宣布，人类基因组中所含基因的预计数目从先前的 30,000 至 40,000（在计划初期的预计数目则高达 2,000,000）调整为 20,000 至 25,000。预期还需要多年的时间来确定人类基因组中所含基因的精确数目。
- 目前基因组信息的注释工作仍然处于初级阶段。



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 基因组学概述

- 概述
- 人类基因组计划

● 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



结构基因组学

结构基因组学 (structural genomics) 是基因组学的一个重要组成部分和研究领域，它是一门通过基因作图、核苷酸序列分析确定基因组成、基因定位的学科。

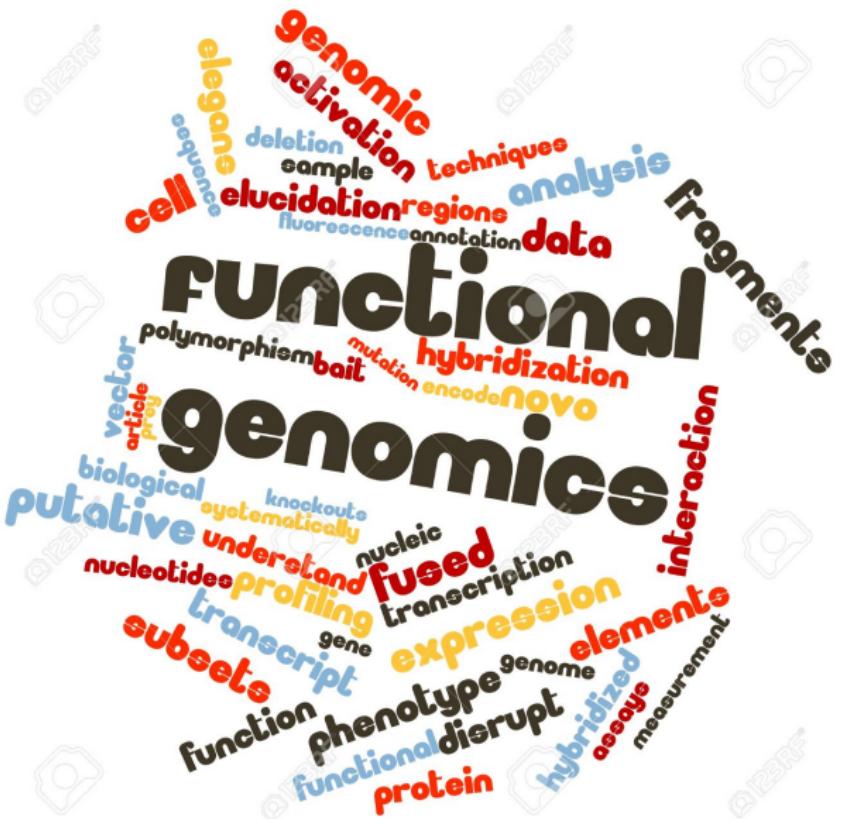


功能基因组学

功能基因组学 (functional genomics) 的研究又往往被称为后基因组学 (postgenomics) 研究，它是利用结构基因组学提供的信息和产物，在基因组或系统水平上全面分析基因的功能和相互作用。

功能基因组学是利用结构基因组学所获得的各种信息，建立与发展各种技术和实验模型来测定基因及基因非编码序列的生物学功能。





比较基因组学

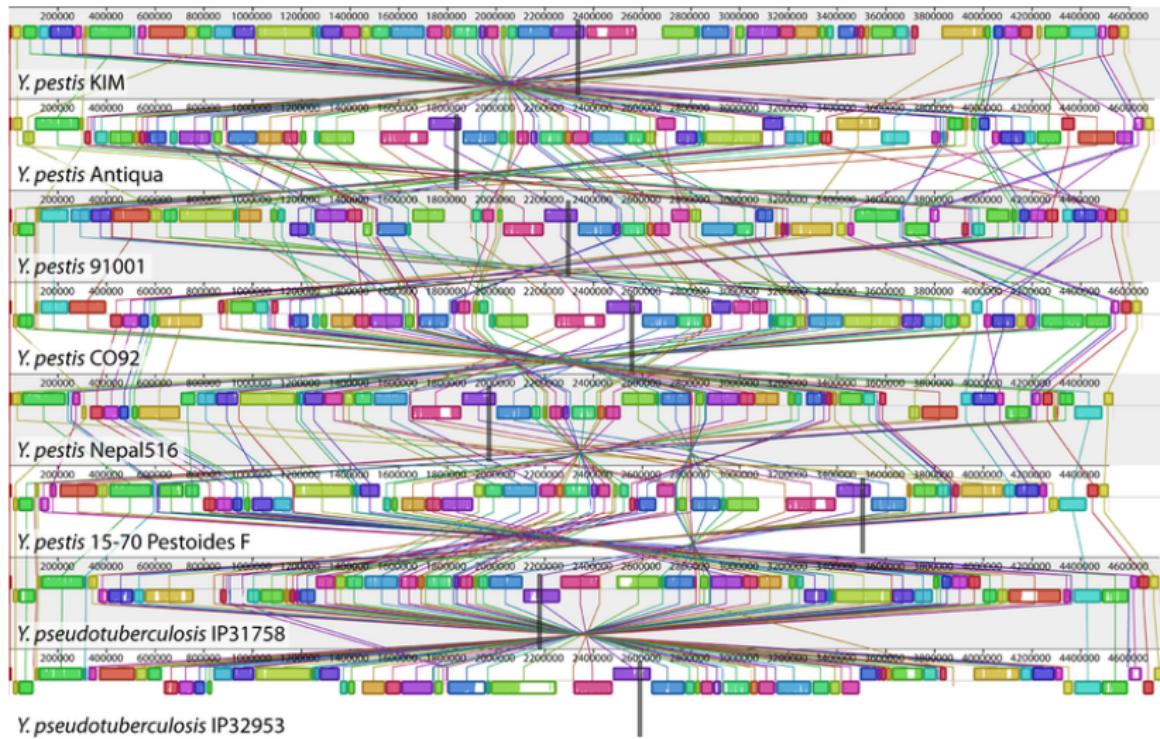
比较基因组学（comparative genomics）是在基因组图谱和测序技术的基础上，对已知的基因特征和基因组结构进行比较以了解基因的功能、表达机制和不同物种亲缘关系的生物学研究。

比较基因组学的基础是相关生物基因组的相似性。全基因组比对是比较基因组学的经典方法。比较基因组学的研究成果催生了水平基因转移理论，支持细胞器起源的内共生学说。

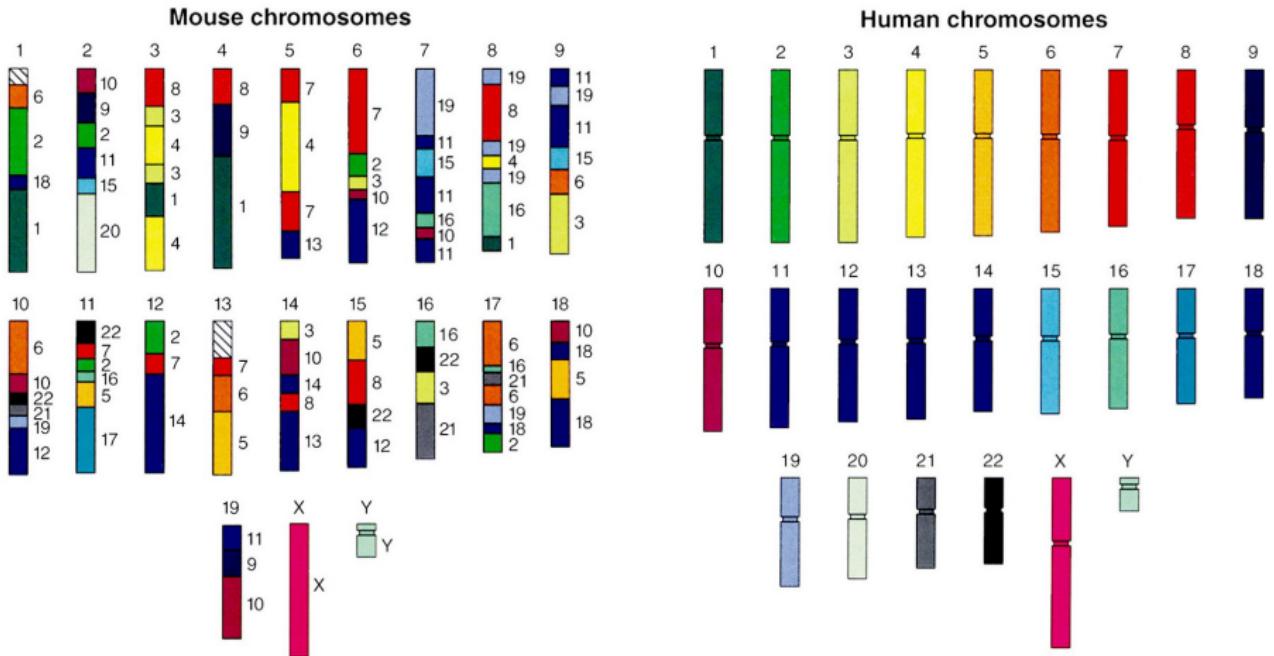
比较基因组学是研究比较不同物种基因组的异同，目的在于寻找物种间共有的、也就是在进化上保守的基因或 DNA 序列，这些基因往往具有重要的生物学功能。也可以从这些模式生物中寻找人类可能具有的新基因，以及为预测新的基因功能提供依据。



基因组学 | 概述 | 分支学科 | 比较基因组学



Mouse and Human Genetic Similarities



种间比较基因组学研究

通过对不同亲缘关系物种的基因组序列进行比较，能够鉴定出编码序列、非编码调控序列及给定物种独有的序列。

基因组范围之内的序列比对，可以了解不同物种在核苷酸组成、同线性关系和基因顺序方面的异同，进而得到基因分析预测与定位、生物系统发生进化关系等方面的信息。



Lift Genome Annotations

This tool converts genome coordinates and genome annotation files between assemblies. The input data can be pasted into the text box, or uploaded from a file. If a pair of assemblies cannot be selected from the pull-down menus, a direct lift between them is unavailable. However, a sequential lift may be possible. Example: lift from Mouse, May 2004, to Mouse, Feb. 2006, and then from Mouse, Feb. 2006 to Mouse, July 2007 to achieve a lift from mm5 to mm9.

Original Genome:

Human

Original Assembly:

Mar. 2006 (NCBI36/hg18)

New Genome:

Mouse

New Assembly:

July 2007 (NCBI37/mm9)

- [Human/Gorilla \(gorGor4\)](#)
- [Human/Mouse \(mm10\)](#)
- [Human/Rat \(rn5\)](#)
- [Human/Cow \(bosTau8\)](#)
- [Human/Cat \(felCat8\)](#)
- [Human/Dog \(canFam3\)](#)
- [Human/Opossum \(monDom5\)](#)
- [Human/Chicken \(galGal5\)](#)
- [Human/X. tropicalis \(xenTro7\)](#)
- [Human/Zebrafish \(danRer10\)](#)

• Multiple Alignments

- [Multiple alignments of 99 vertebrate genomes with Human](#)
- [Conservation scores for alignments of 99 vertebrate genomes with Human](#)
- [Basewise conservation scores \(phyloP\) of 99 vertebrate genomes with Human](#)
- [FASTA alignments of 99 vertebrate genomes with Human for CDS regions](#)
- [Multiple alignments of 19 mammalian \(16 primate\) genomes with Human](#)
- [Conservation scores for alignments of 19 mammalian \(16 primate\) genomes with Human](#)





80后励志网
www.201980.com



种内比较基因组学研究

同种群体内基因组存在大量的变异和多态性，正是这种基因组序列的差异构成了不同个体与群体对疾病的易感性和对药物与环境因子不同反应的遗传学基础。

- 单核苷酸多态性 (single-nucleotide polymorphism, SNP) 是指在基因组水平上由于单个核苷酸位置上存在转换或颠换等变异所引起的 DNA 序列多态性。
- 拷贝数多态性 (copy number polymorphism, CNP)：平均 2 个个体间存在 11 个 CNP 的差异，CNP 的平均长度为 465kb，其中半数以上的 CNP 在多个个体中重复出现，并经常定位于其他类型的染色体重排附近。



系统发育谱法

系统发育谱法 (phylogenetic profile method) 是在基因组全序列已完成测序的一系列基因组中分析某一蛋白质存在与否的模式。如果两个蛋白质在所研究的若干基因组中有相同的系统发育谱，便推断这两个蛋白质具有功能联系。

基因邻居法

基因邻居法 (gene neighbour method) 的原理是原核生物中如果两个基因在一个共同的操纵子内，且在不同的其他基因组也出现相邻现象，就可以推断它们编码的蛋白质之间具有功能联系。



系统发育谱法

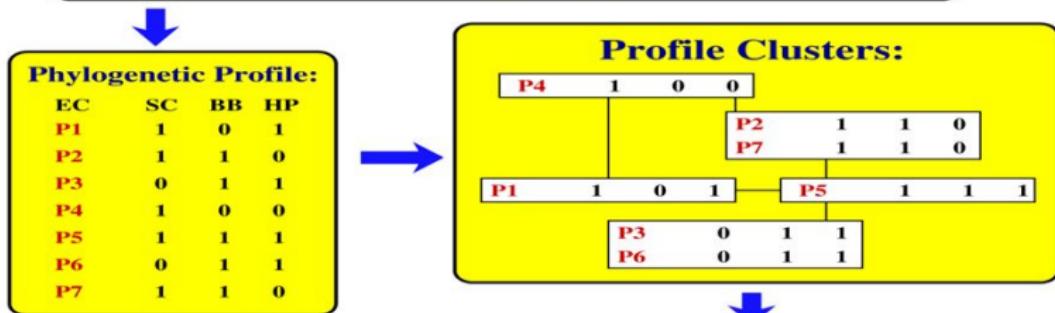
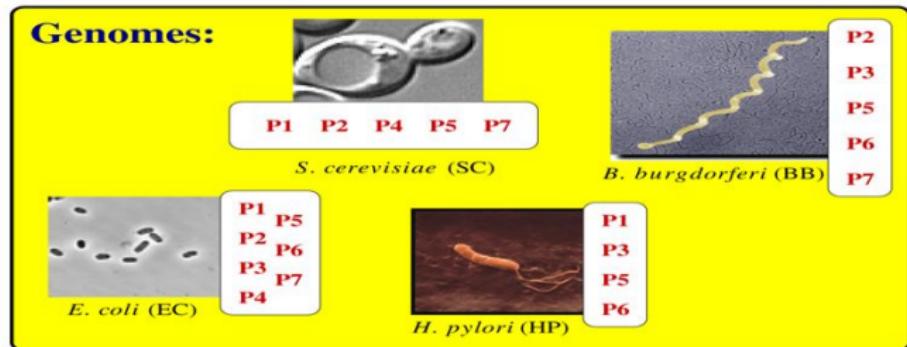
系统发育谱法 (phylogenetic profile method) 是在基因组全序列已完成测序的一系列基因组中分析某一蛋白质存在与否的模式。如果两个蛋白质在所研究的若干基因组中有相同的系统发育谱，便推断这两个蛋白质具有功能联系。

基因邻居法

基因邻居法 (gene neighbour method) 的原理是原核生物中如果两个基因在一个共同的操纵子内，且在不同的其他基因组也出现相邻现象，就可以推断它们编码的蛋白质之间具有功能联系。



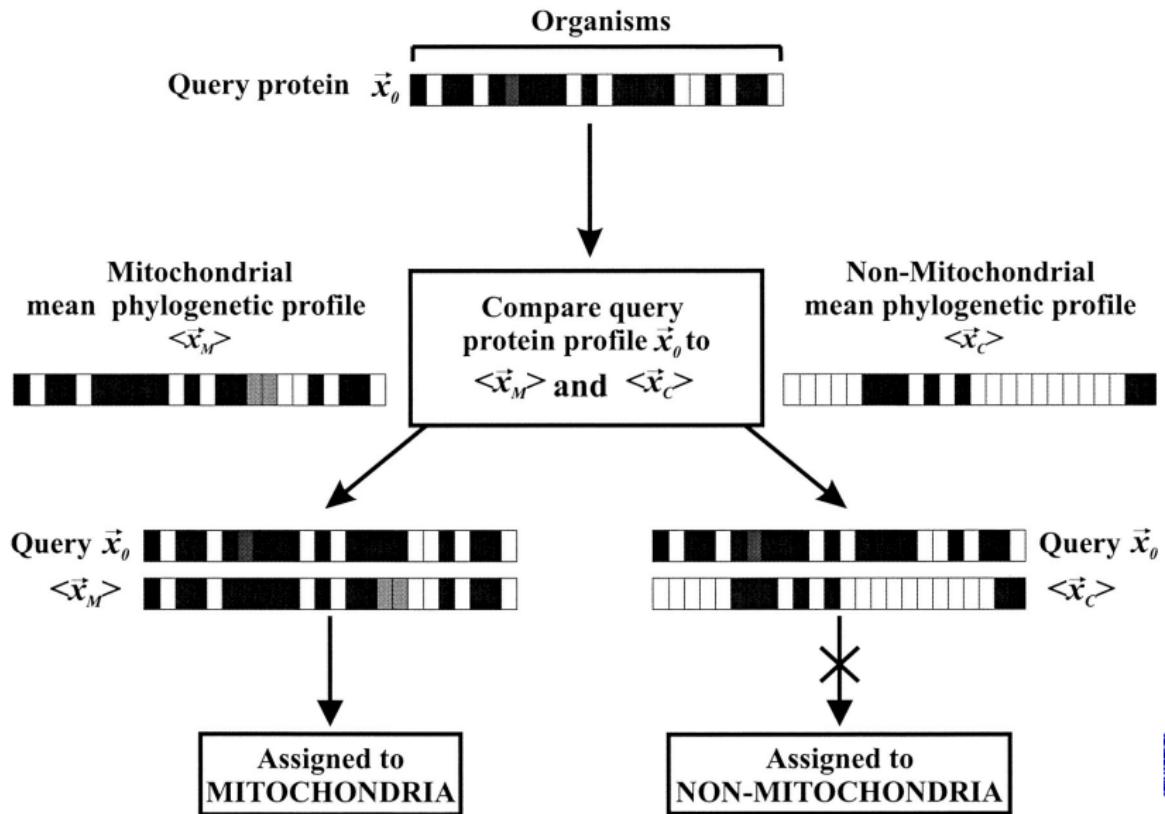
PHYLOGENETIC PROFILE METHOD



Conclusion: P2 and P7 are linked
P3 and P6 are linked

Pellegrini et al (1999) PNAS 96, 4285





药物基因组学

药物基因组学旨在理解个体对药物不同反应的遗传背景，即为什么某种药物对一部分人群有效，而对另一部分人群效果不佳或完全失效。

制药业将充分应用药物基因组学的理论知识和技术手段来设计临床实验并模拟和分析理论与实验数据，也为新药设计和筛选提供依据。



个体对药物不同反应的遗传背景研究

人类基因组 DNA 在个体间一般呈现 0.01% 差异，几乎每个基因都有一个核苷酸变异谱。这是人类个体对药物敏感性不同的遗传基础，也是现代医学提出药物治疗必须个体化的依据。

为药物设计和筛选提供依据

现代药物设计应考虑与遗传有关的若干因素：

- 与致病有关的等位基因常会影响到对药物作用的反应
- 药物代谢途径与疗效密切相关



个体对药物不同反应的遗传背景研究

人类基因组 DNA 在个体间一般呈现 0.01% 差异，几乎每个基因都有一个核苷酸变异谱。这是人类个体对药物敏感性不同的遗传基础，也是现代医学提出药物治疗必须个体化的依据。

为药物设计和筛选提供依据

现代药物设计应考虑与遗传有关的若干因素：

- 与致病有关的等位基因常会影响到对药物作用的反应
- 药物代谢途径与疗效密切相关



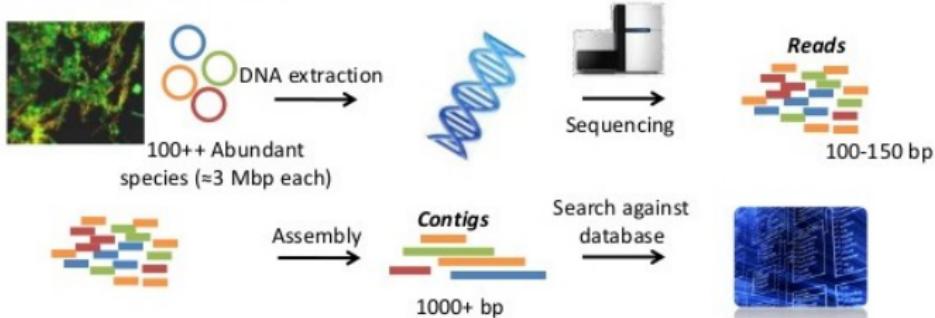
元基因组学

元基因组或总体基因组学 (metagenomics) 是一门直接取得环境中所有遗传物质的研究，意指直接研究环境中微生物群落基因组学的应用，而非于实验室中进行单一个体纯化与培养的实验方式。研究领域广泛，也可称为环境基因组学、生态基因组学或群落基因组学。

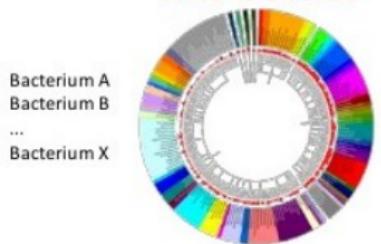
在早期研究微生物基因组必须将环境基因 DNA 或 RNA 克隆进入大肠杆菌体内，利用复制扩增方式，分析在自然环境中复制扩增特定基因 (通常为 16S rRNA) 的多样性。但是，以复制扩增的方式得不到精准的微生物多样性。总体基因组学是认识复杂微生物群落的主要途径，提供一个更客观的方式发现微生物的世界，有可能改变之前我们所认知的微生物世界。



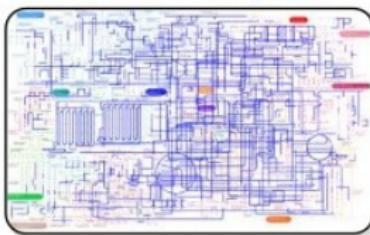
Metagenomics



Phylogenetic classification Who is there?

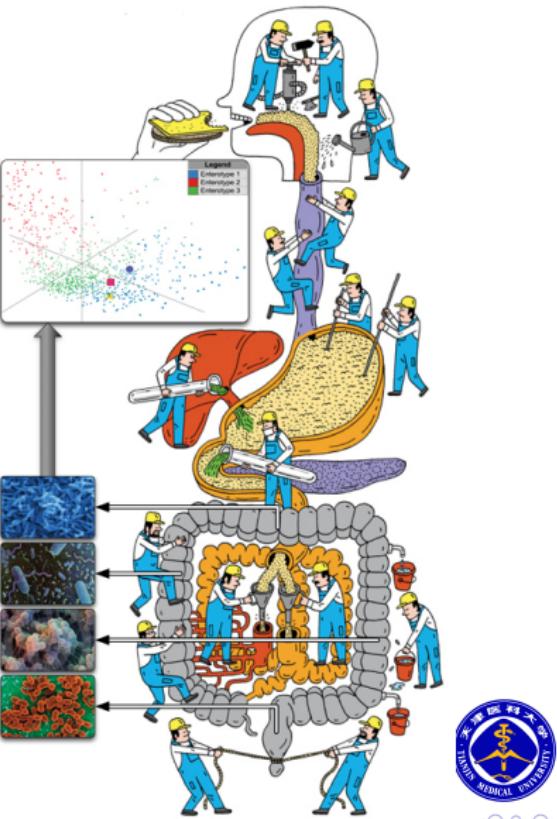
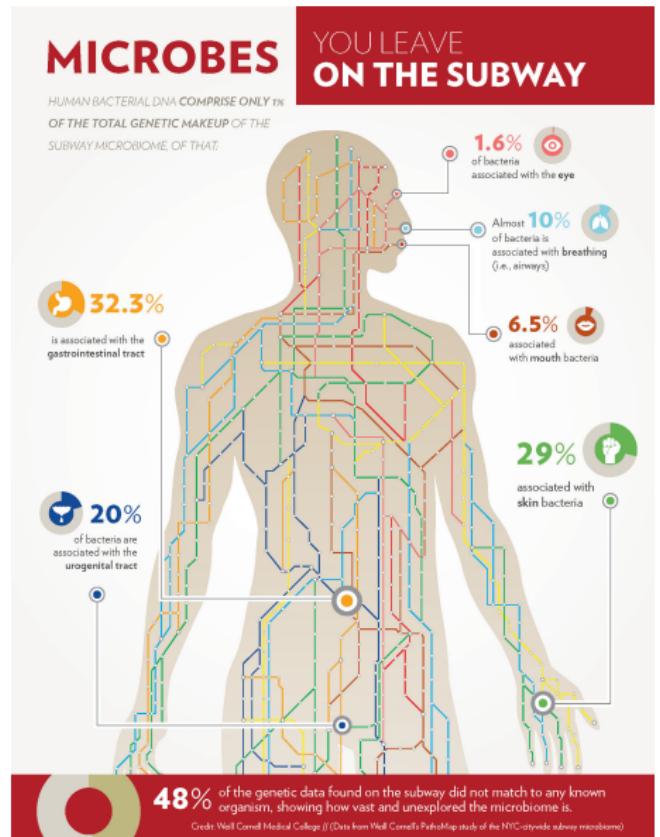


Functional classification What can they do?



CENTER FOR MICROBIAL COMMUNITIES | AALBORG UNIVERSITY





人类元基因组是指与人类共生的全部微生物的基因总和。又被称为“微生物组”或“人类第二基因组”。

人类体内的微生物多达 1000 多种，特别是胃肠道内的微生物最为丰富；因此我们所说的元基因组在狭义上指的是肠道元基因组。

人体内微生物的编码基因的总量大约是人类编码基因数目的 50-100 倍，这相当于在人类体内存在着另一个基因组通过表达调控人体的生命健康，即第二基因组。

目前关于元基因组的研究还处于一个比较浅的阶段，在现有的研究中普遍认为糖尿病和肥胖症与人体元基因组有关。

美国国立卫生研究院在 2007 年启动人体微生物计划，该计划一开始最主要的目的就是调查是否有人体微生物的存在、了解人体微生物的变化与人类健康的关系，并开发新的技术和生物信息的工具以支持这些目标。



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



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

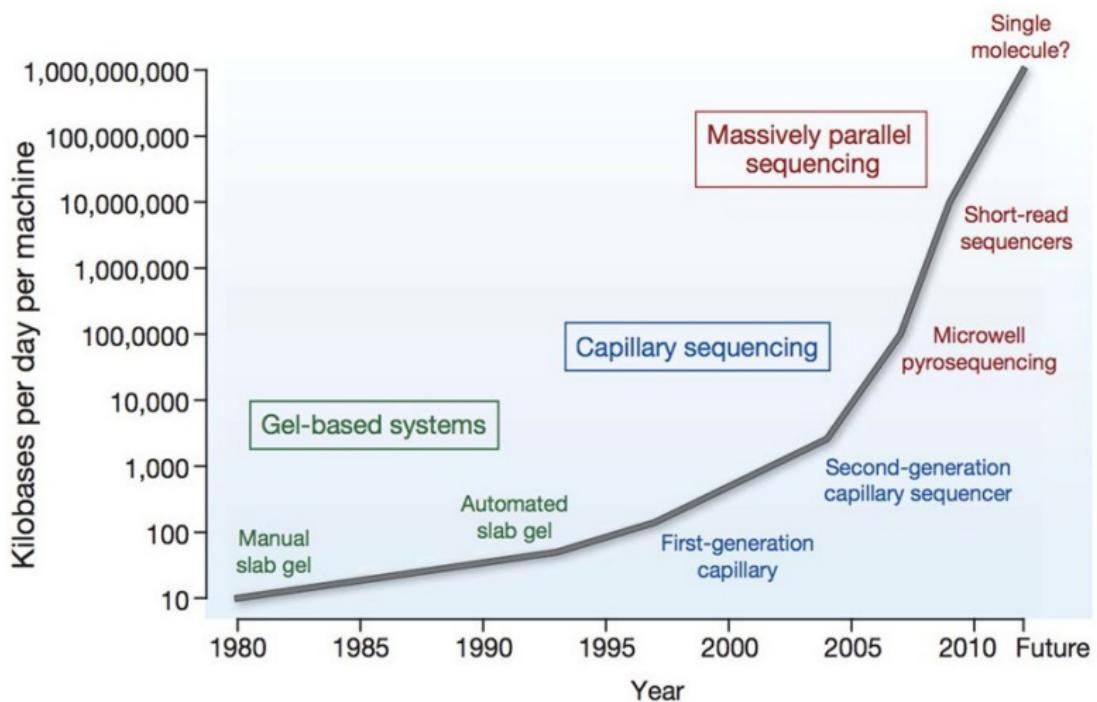
DNA 测序 (DNA sequencing) 是指分析特定 DNA 片段的碱基序列，也就是腺嘌呤 (A)、胸腺嘧啶 (T)、胞嘧啶 (C) 与鸟嘌呤 (G) 的排列方式。

RNA 测序

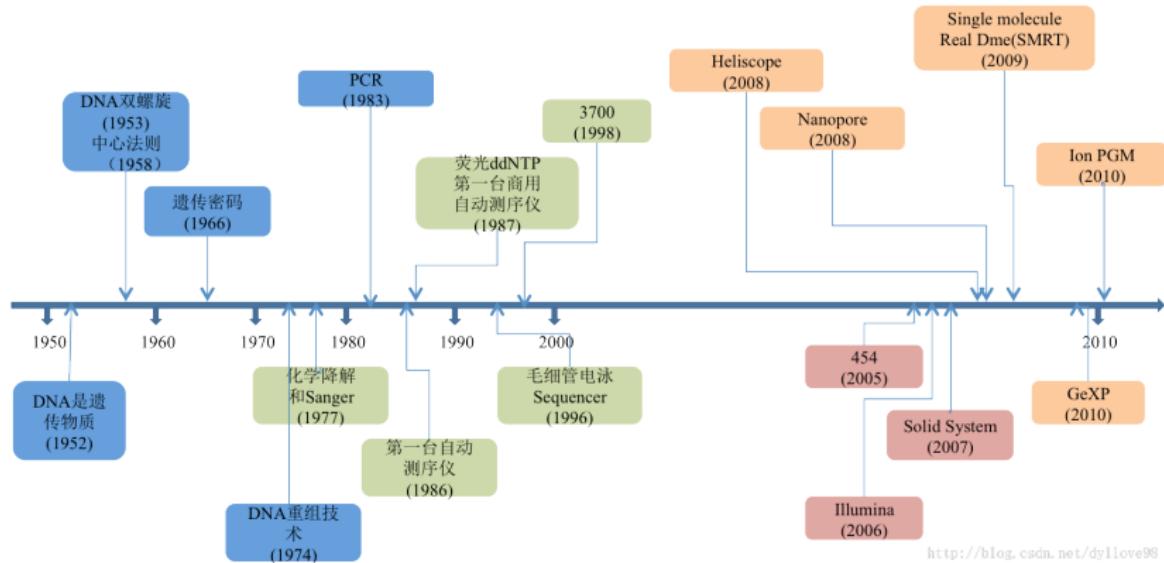
RNA 测序则通常将 RNA 提取后，反转录为 DNA 后使用 DNA 测序的方法进行测序。



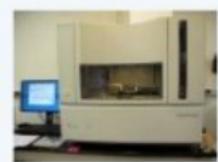
The History of DNA Sequencing Technology



基因组学 | 测序 | 历史



Advances in Sequencing Technology



AB3700

1998



SOLiD

2005



454

2008



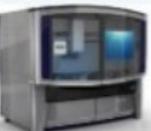
IonTorrent

2009



HiSeq

2010



SOLiD 5500xl

2011



MiSeq

2012

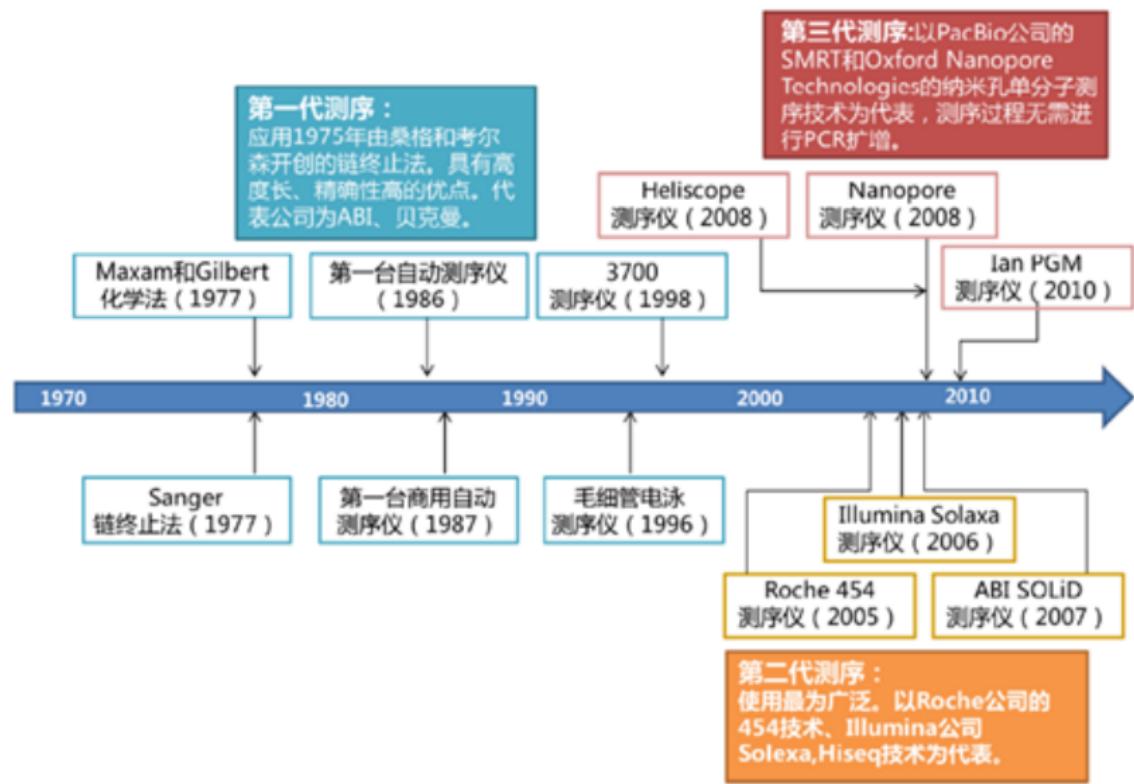


IonProton

Complete Genomics

PacBio RS





教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



- 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）



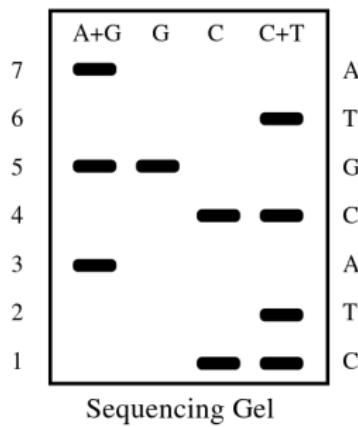
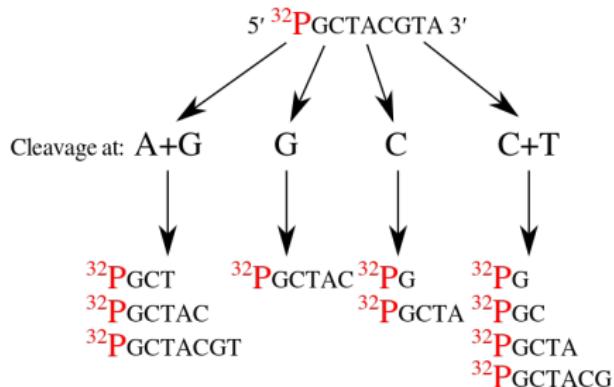
化学测序法

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。



基因组学 | 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

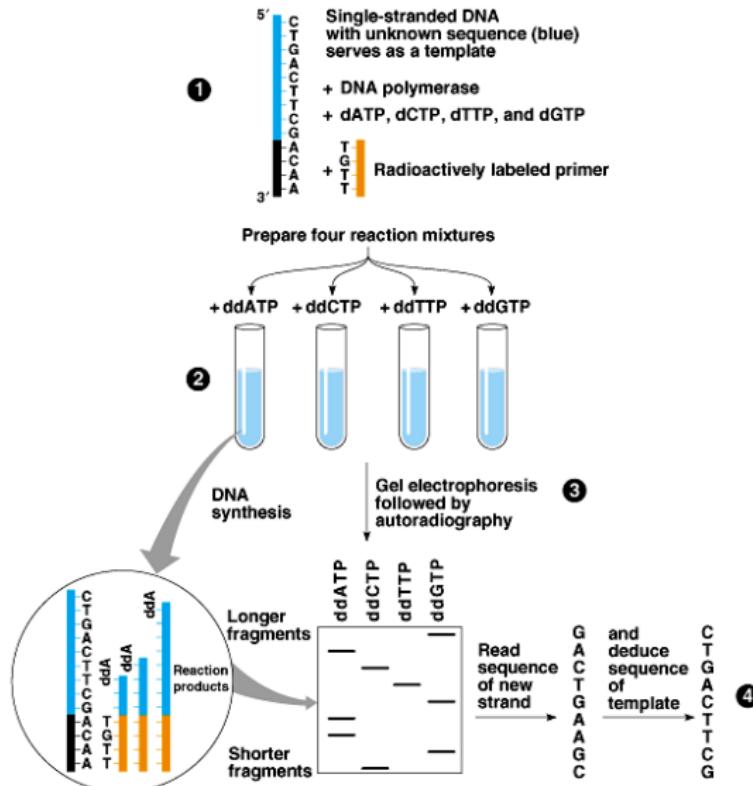
- 最长可测定 600-1000bp 的 DNA 片断
- 对重复序列和多聚序列的处理较好
- 序列准确性高，高达 99.999%
- 测序的“黄金标准”

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

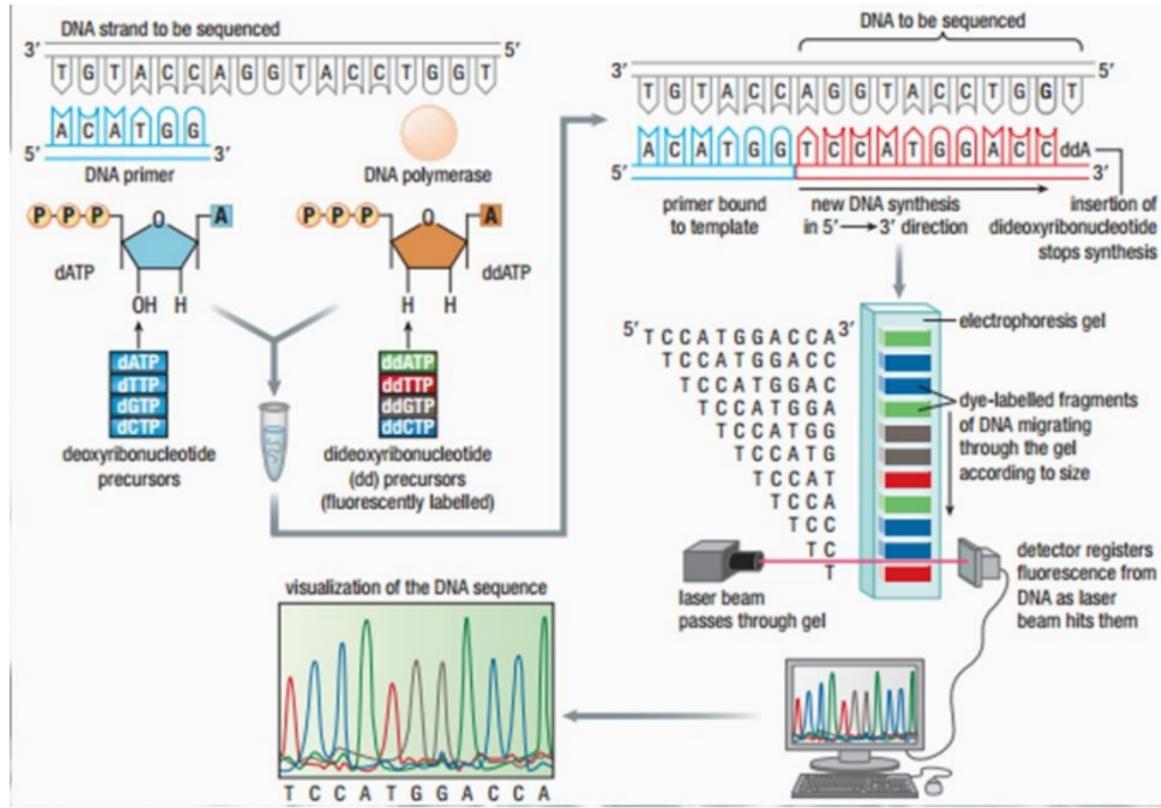


基因组学 | 测序 | 第一代 | 桑格测序法



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

基因组学 | 测序 | 第一代 | 桑格测序法



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



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,

三句半

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
I love you
not for who
am with you
who I am

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

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

love you not
but for who
you not for
for who I
with you.
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》



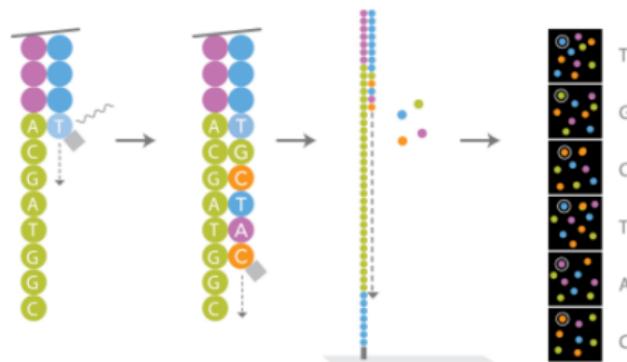
GCTGATGTGCCGCCTCACTTCGGTGGTGAGGTG	Reference sequence
CTGATGTGCCGCCTCACTTCGGTGGT	Short read 1
TGATGTGCCGCCTCACT ACGGTGGTG	Short read 2
GATGTGCCGCCTCACTTCGGTGGTGA	Short read 3
GCTGATGTGCCGCCTCACT ACGGTG	Short read 4
GCTGATGTGCCGCCTCACT ACGGTG	Short read 5



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
TGCTACTGGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIG9IC
```

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

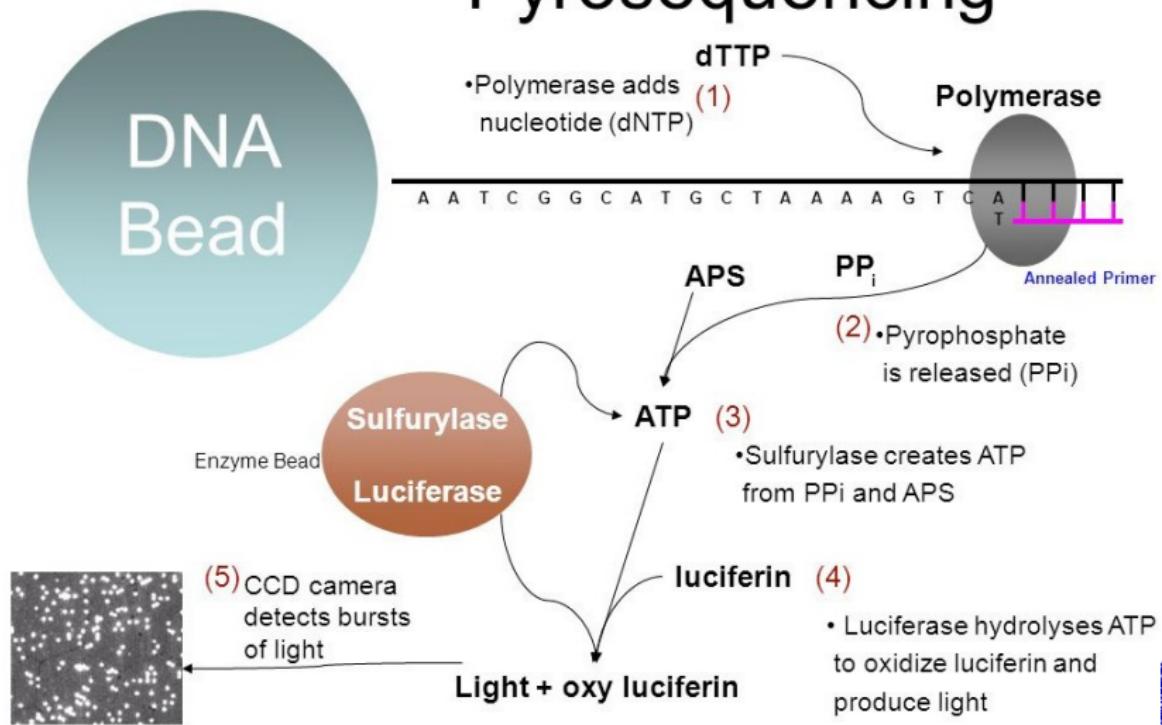
焦磷酸测序

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 PyrogramTM 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

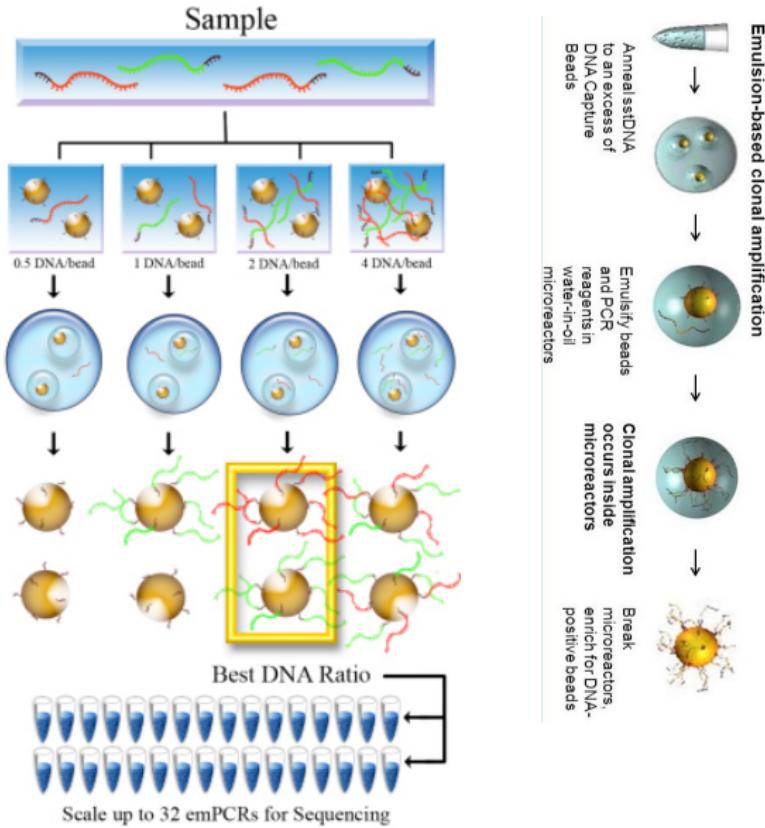


emPCR

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。

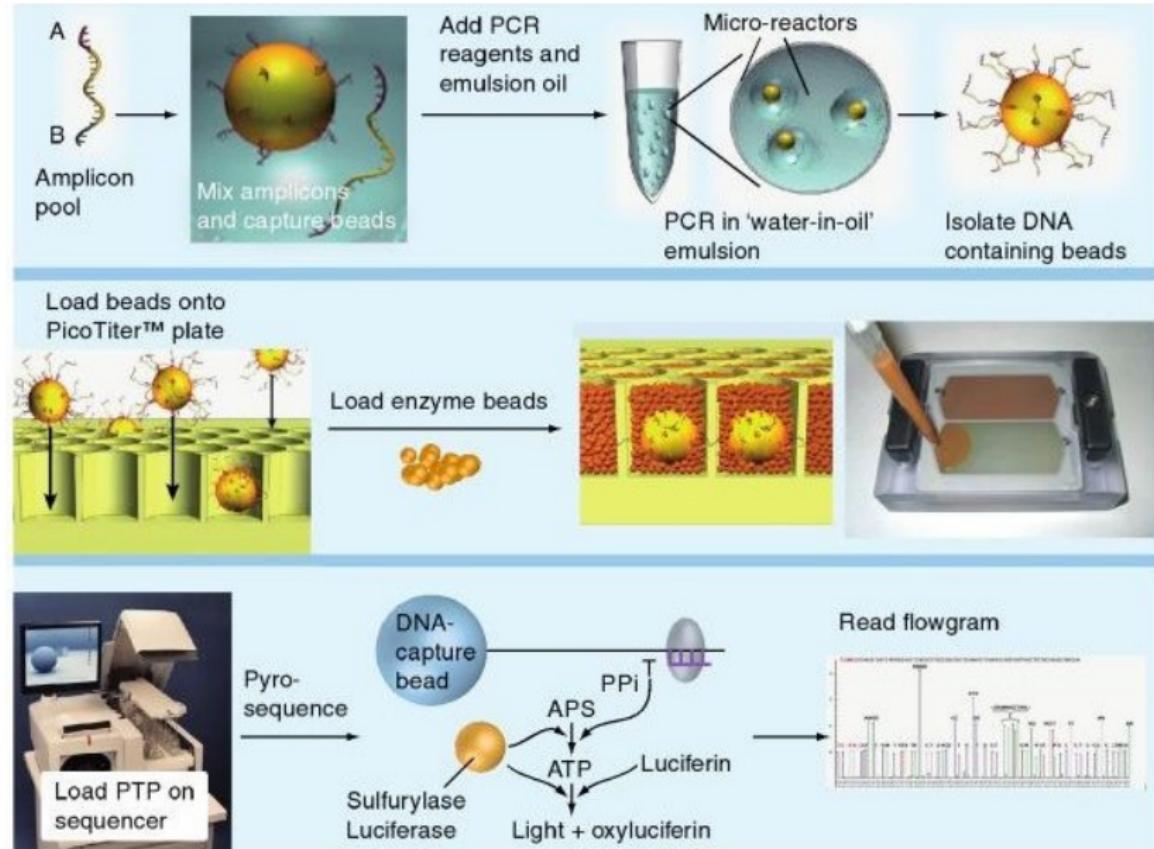




概述

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。





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

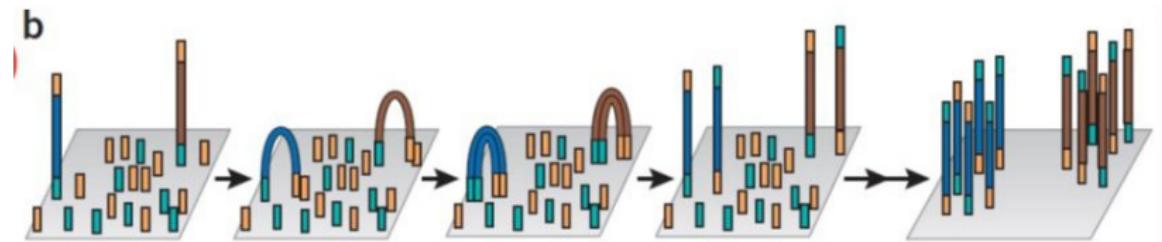


桥式扩增

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。

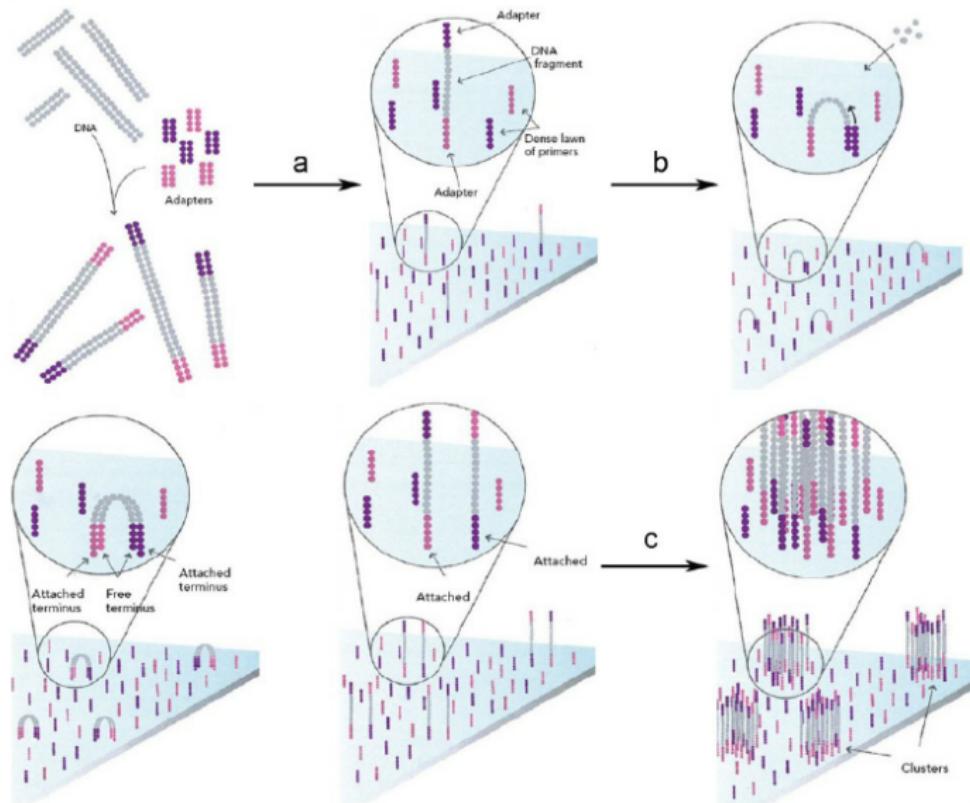


Bridge PCR



- 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.

基因组学 | 测序 | 第二代 | Illumina/Solexa

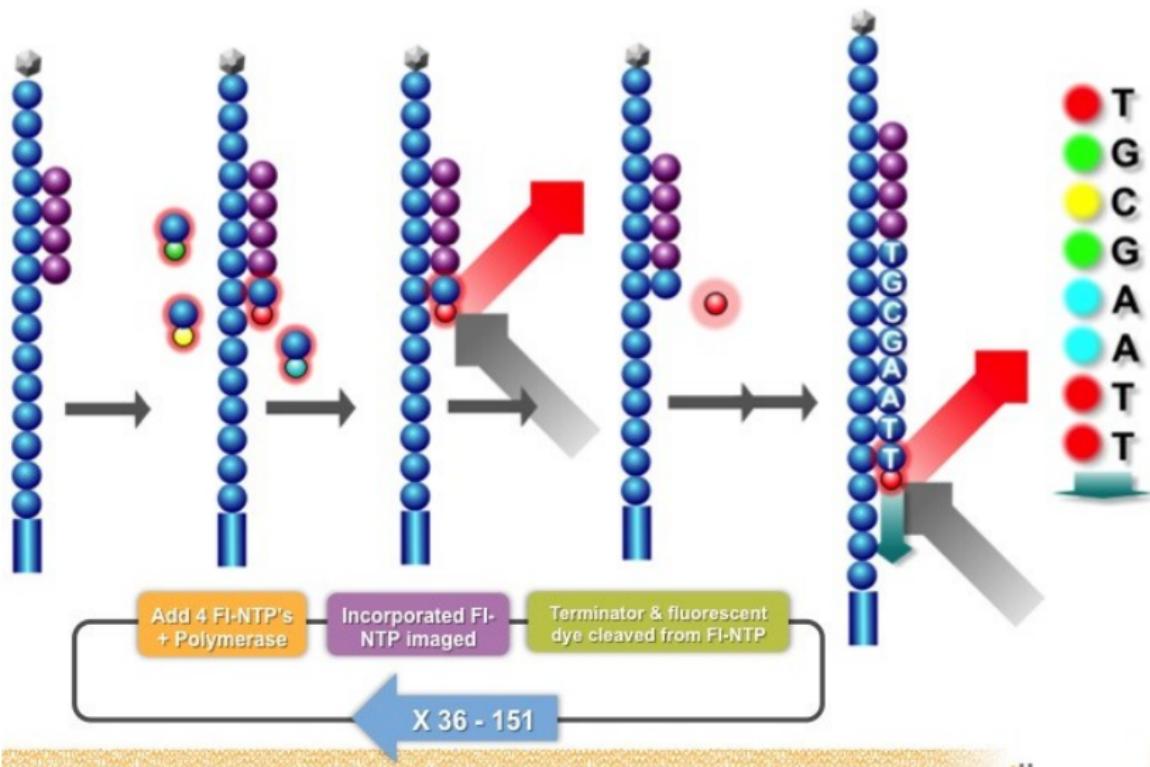


边合成边测序

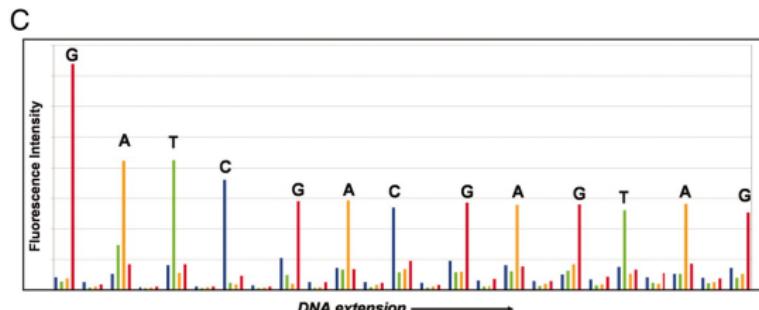
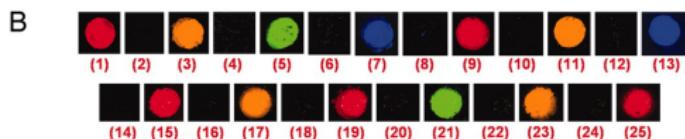
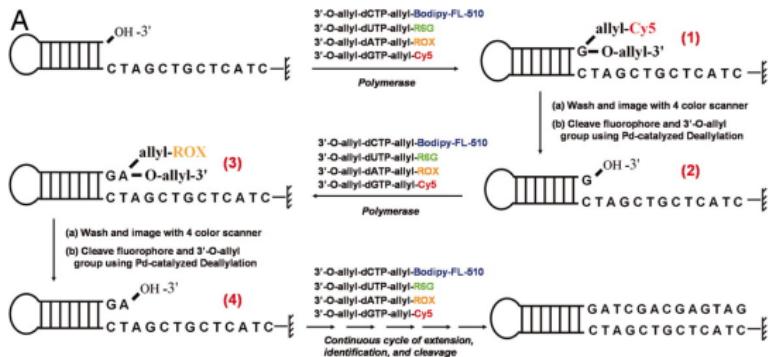
边合成边测序 (sequencing by synthesis, SBS) : 以 DNA 单链为模板, 在合成互补链的时候, 利用带荧光标记的 dNTP 发出不同的荧光来确定碱基类型。



Sequencing By Synthesis



基因组学 | 测序 | 第二代 | Illumina/Solexa

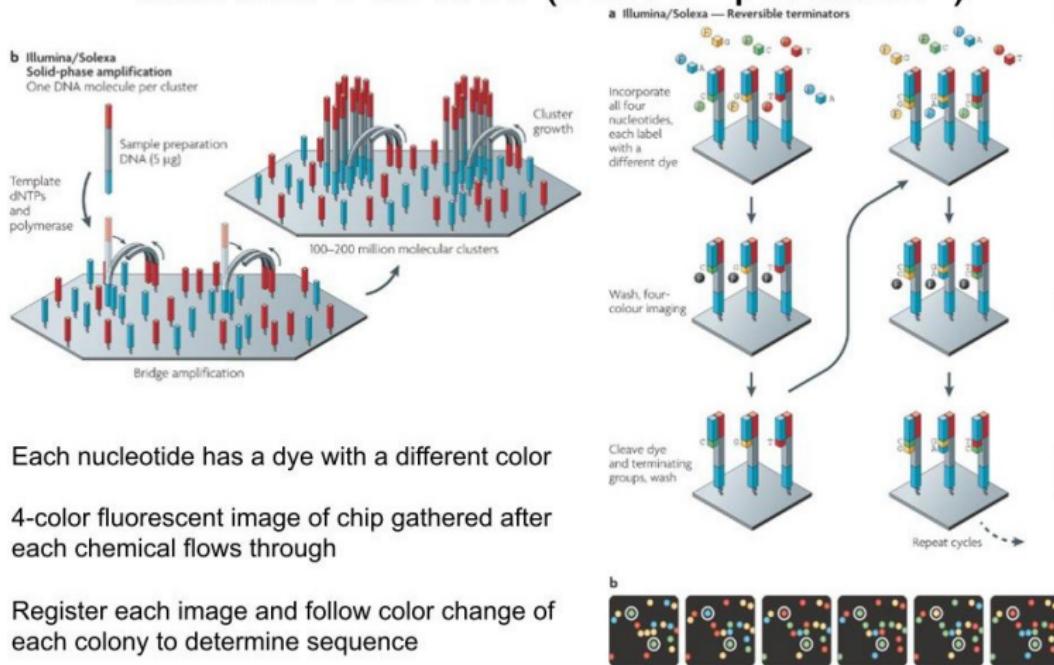


概述

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。

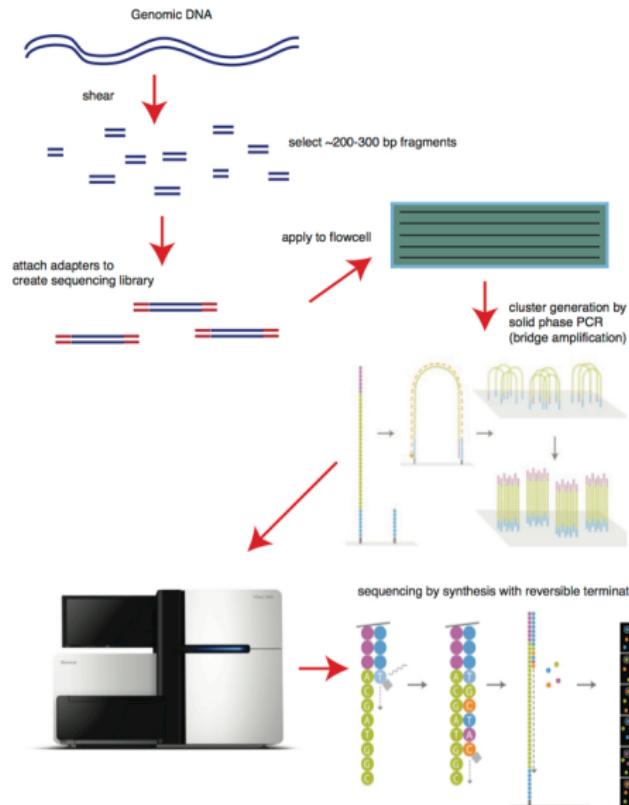


Illumina Colonies (called “polonies”)



ECE/BioE 416
Lecture 24

基因组学 | 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

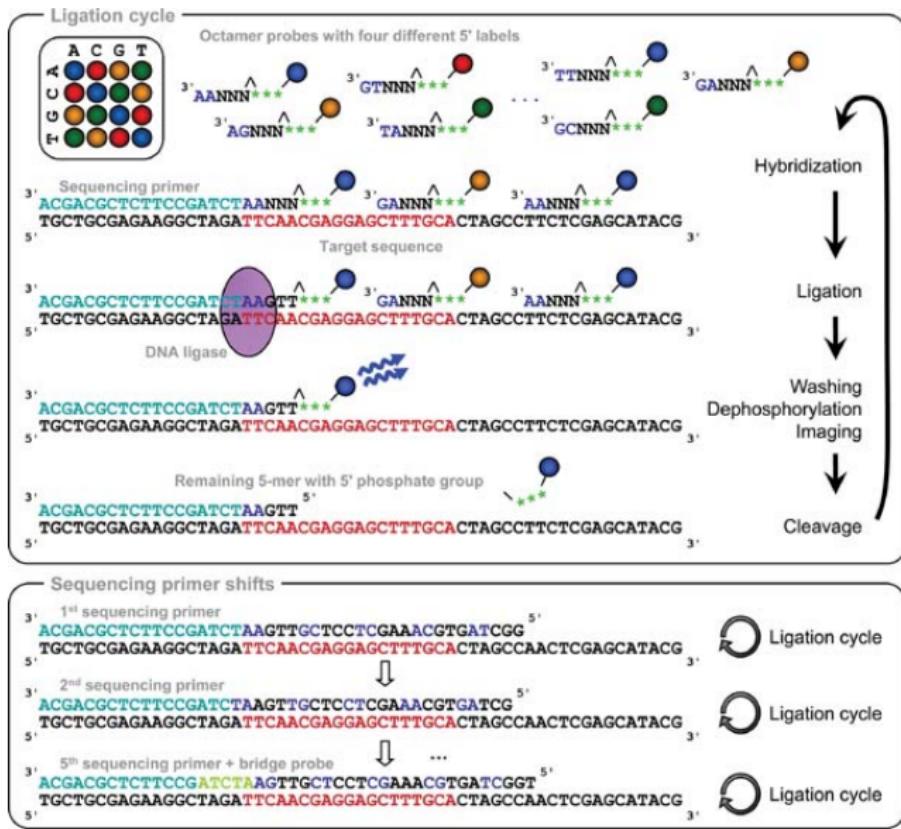


边连接边测序

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。



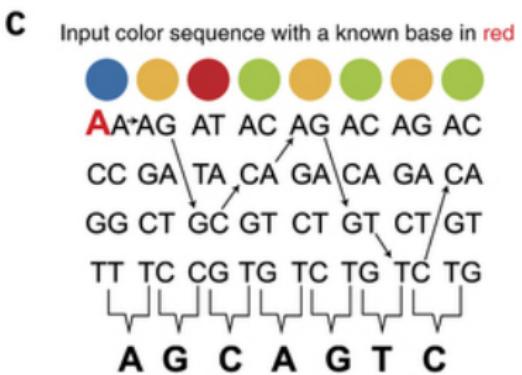
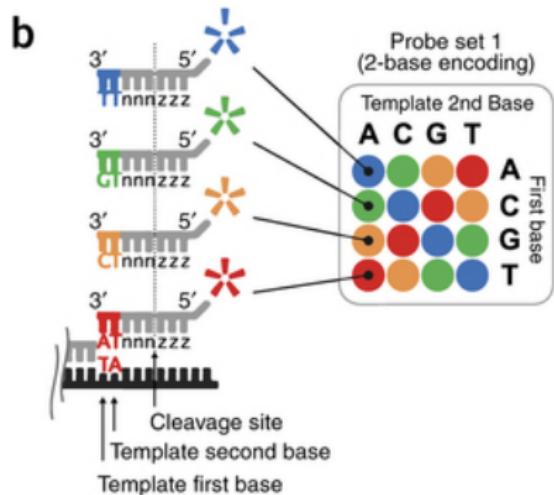
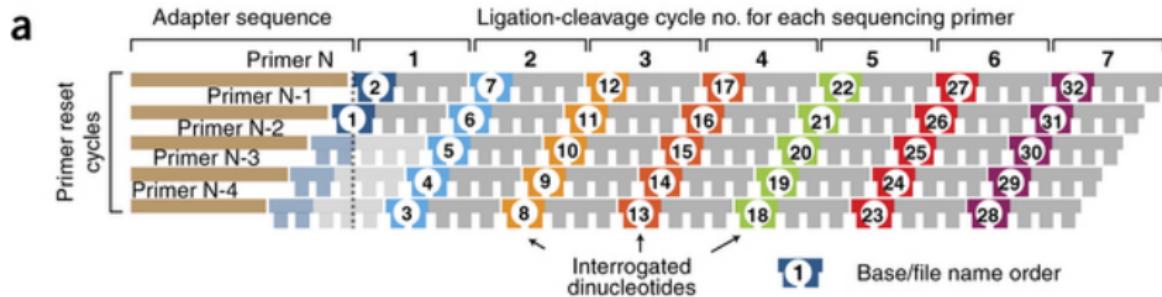
基因组学 | 测序 | 第二代 | ABI/SOLiD



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。

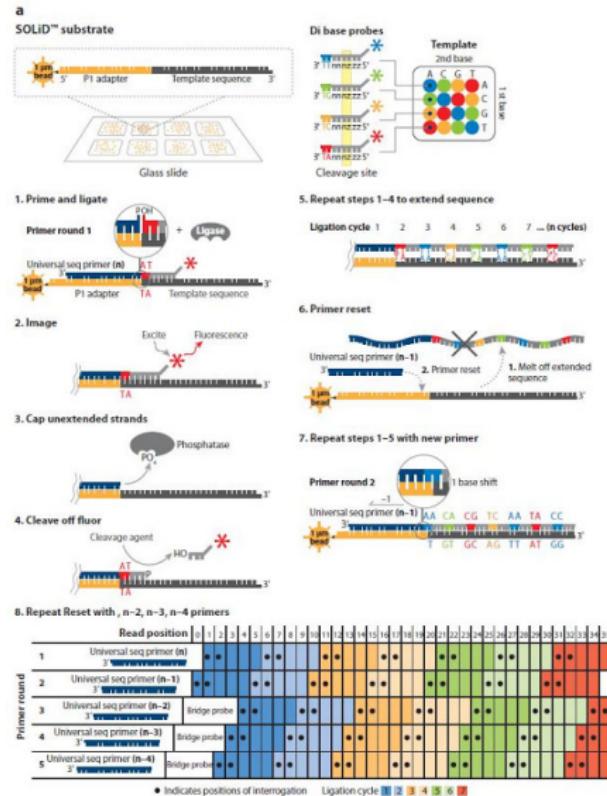
基因组学 | 测序 | 第二代 | ABI/SOLiD



AAGCAGTCA

Output sequence





SNP site indicated by 2 adjacent color changes



Single color change is typically a measurement error



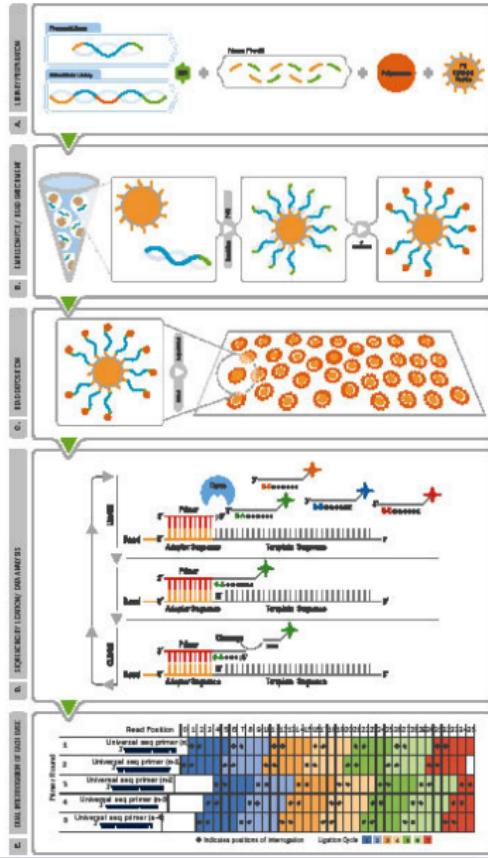
1 Base Deletion



Insertion



基因组学 | 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



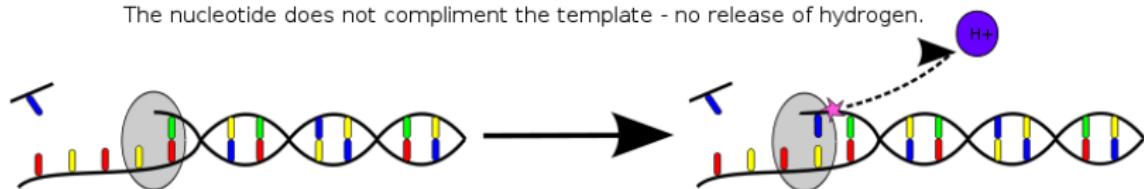
概述

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⁺ 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⁺ 离子信号，H⁺ 离子信号再直接转化为数字信号，从而读出 DNA 序列。

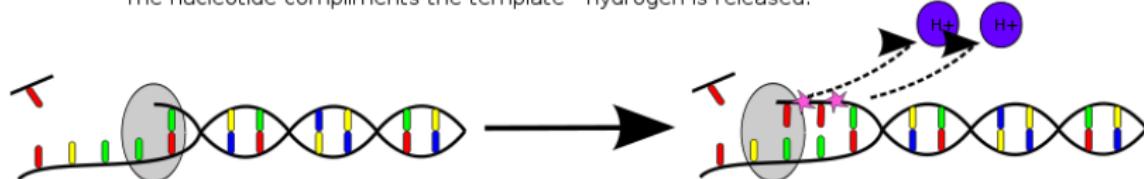




The nucleotide does not compliment the template - no release of hydrogen.

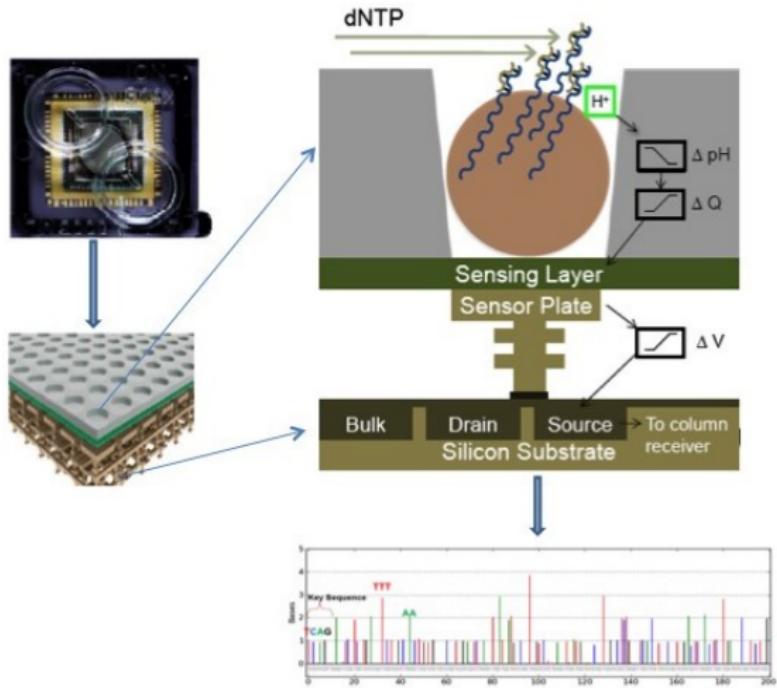


The nucleotide compliments the template - hydrogen is released.



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





dNTP流经反应池并发插入时，释放的 H^+ 引起pH值的变化(ΔpH)

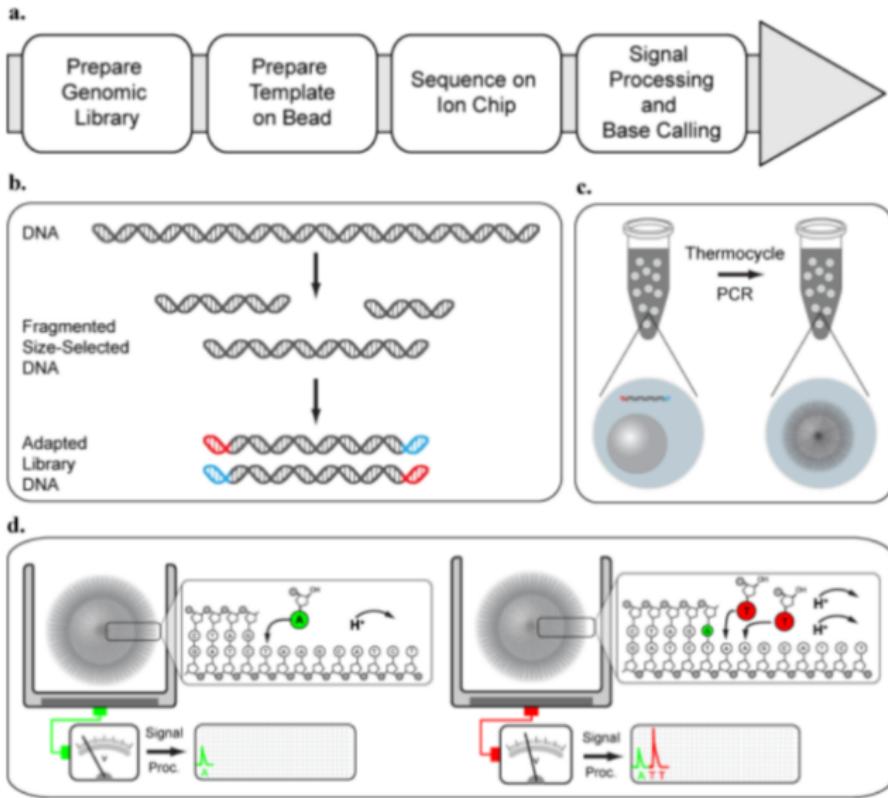
反应池底部金属-氧化物-传感层表面电势变化

电势的变化引起底部场效应晶体管终端的电压(ΔV)的变化

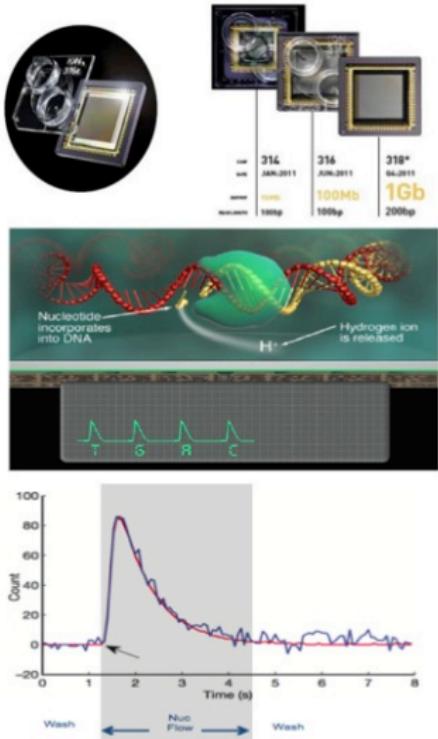
根据电压(ΔV)变化读取流经反应池的碱基，从而读取测序序列



基因组学 | 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。

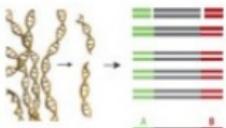
Ion Torrent 的化学测序原理自然简单，无修饰的核苷酸、无激光器或光学检测设备，因而可达到极小的测序偏差和出色的测序覆盖均衡度。



Next-generation DNA sequencing

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

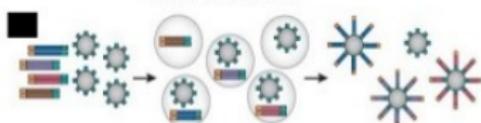
1



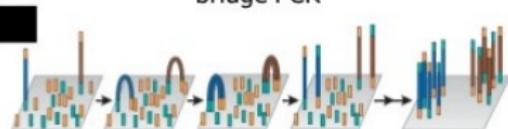
DNA
fragmentation
and in vitro
adaptor ligation

emulsion PCR

2



bridge PCR

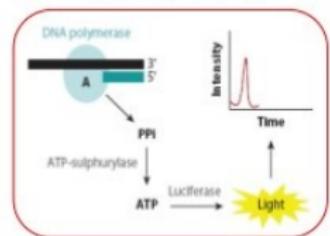


3

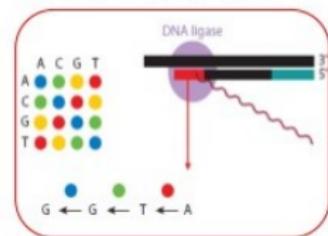
Pyrosequencing

Sequencing-by-ligation

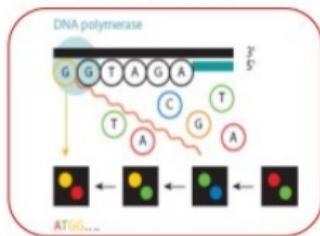
Sequencing-by-synthesis



454 sequencing



SOLiD platform



Solexa technology



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 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- **第三代测序技术**
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



单分子测序

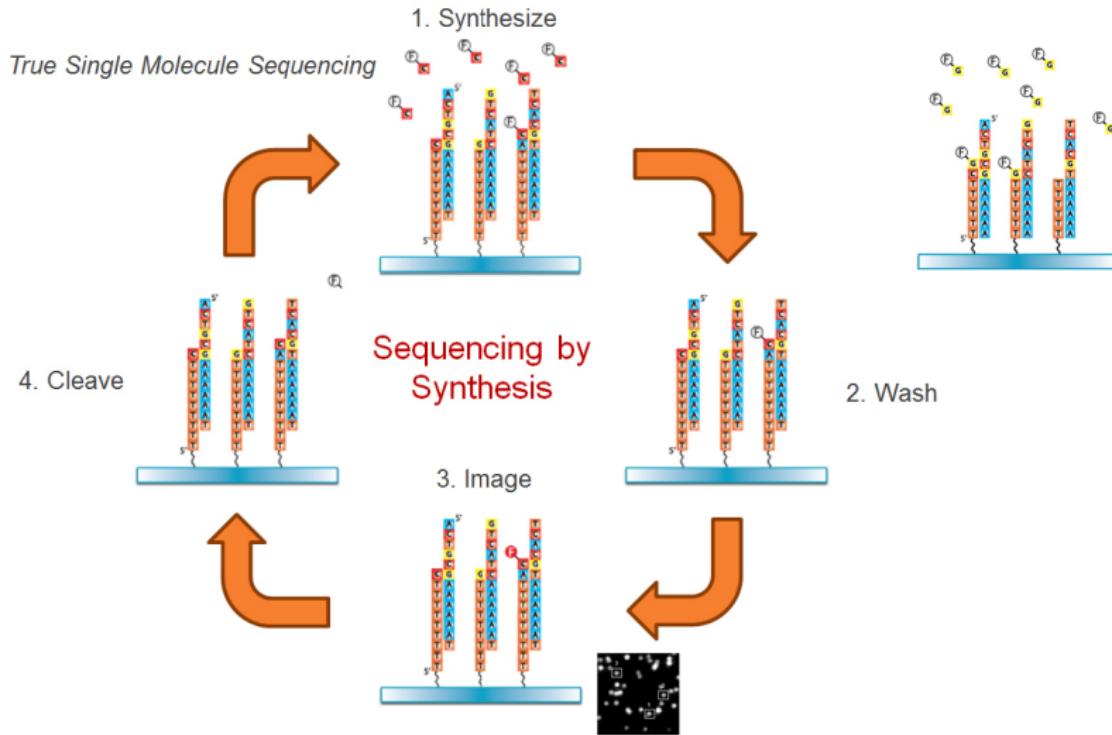
测序过程无需进行 PCR 扩增。



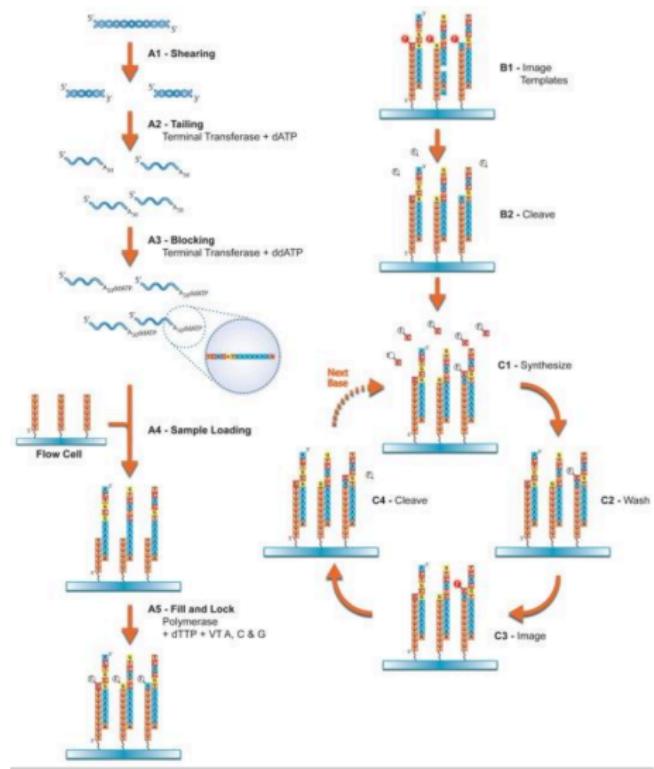
概述

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。





基因组学 | 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。



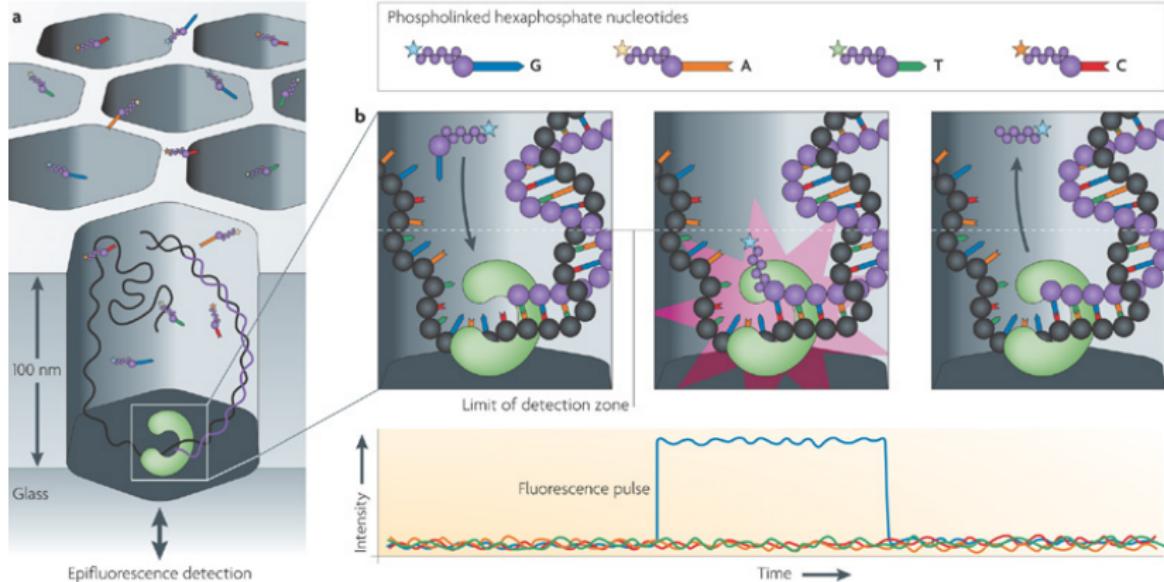
概述

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。



基因组学 | 测序 | 第三代 | SMRT

Pacific Biosciences — Real-time sequencing



Nature Reviews | Genetics



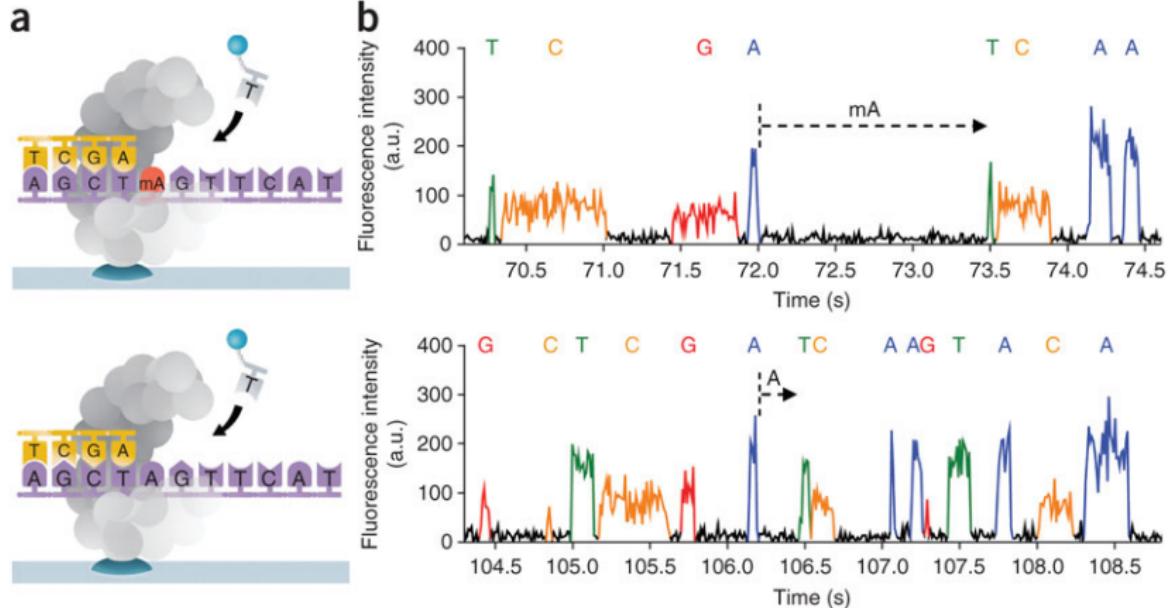
优缺点

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。





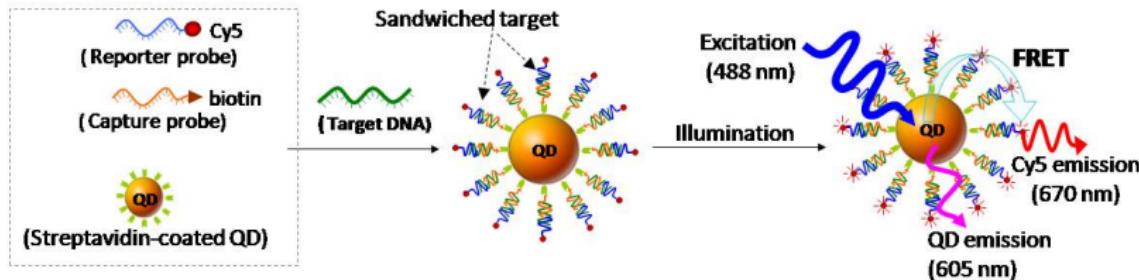
概述

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。



基因组学 | 测序 | 第三代 | FRET

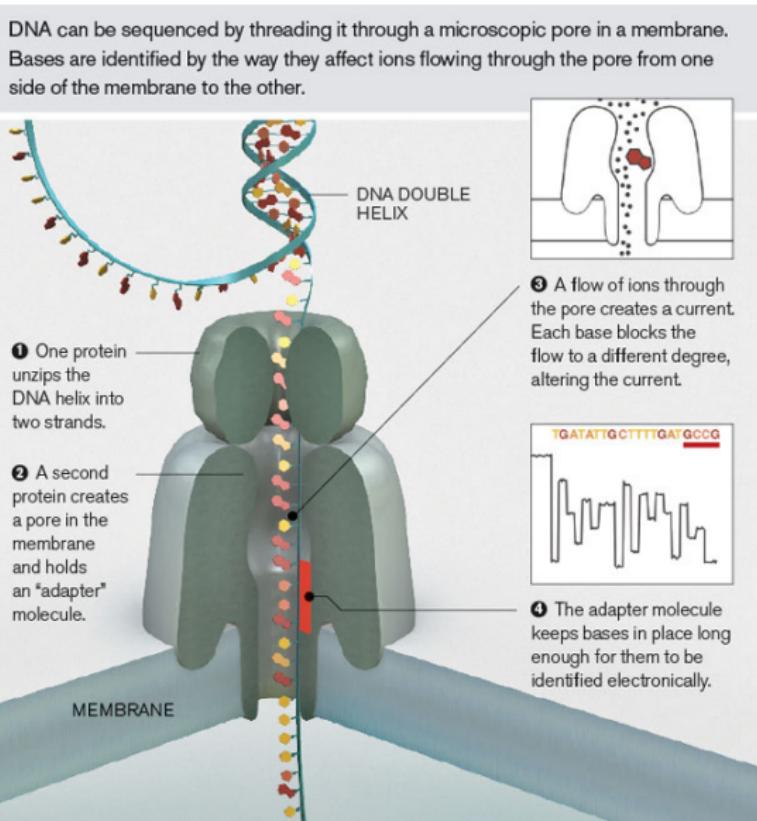


概述

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。



基因组学 | 测序 | 第三代 | 纳米孔测序



优缺点

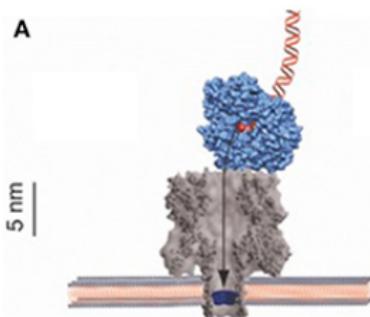
纳米孔测序的主要特点是：读长很长，大约在几十 kb，甚至 100kb；错误率目前介于 1% 至 4%，且是随机错误，而不是聚集在读段的两端；数据可实时读取；通量很高（30x 人类基因组有望在一天内完成）；起始 DNA 在测序过程中不被破坏；样品制备简单又便宜。理论上，它也能直接测序 RNA。

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。

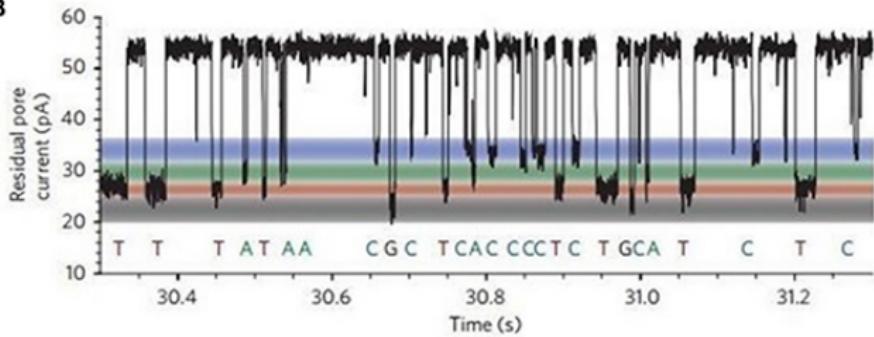


基因组学 | 测序 | 第三代 | 纳米孔测序

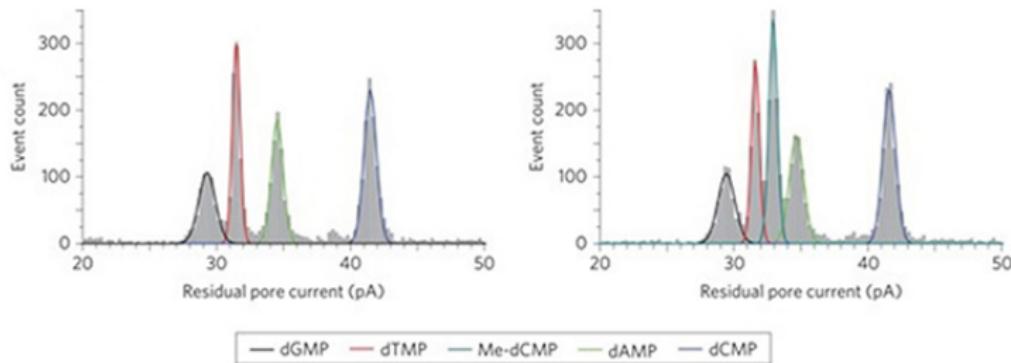
A



B



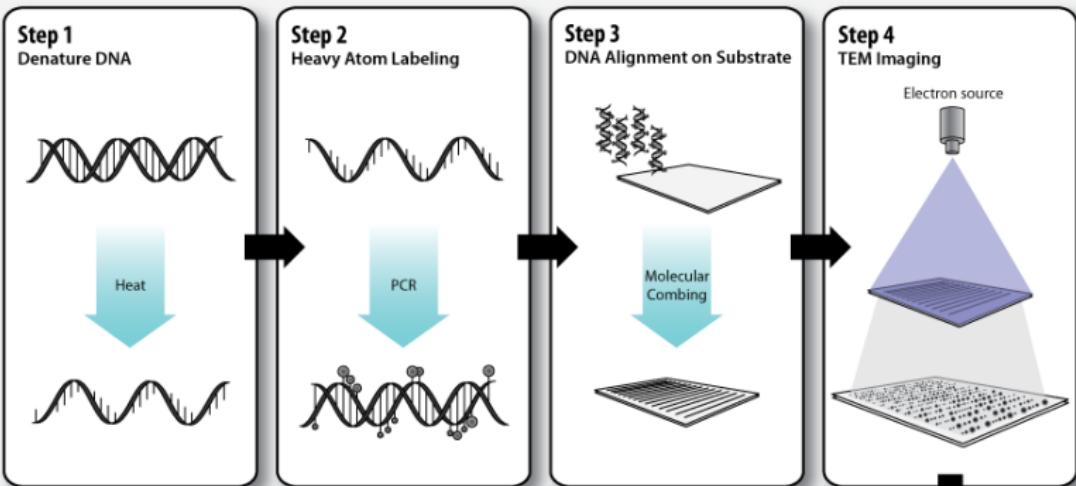
C



概述

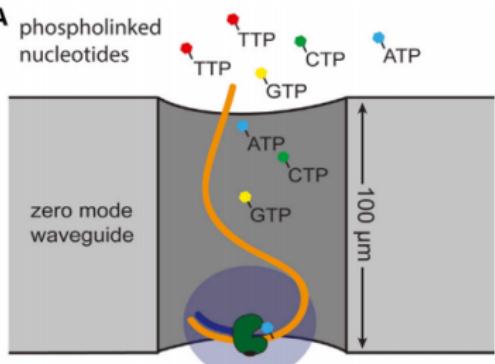
以单链线性 DNA 为模板，以三种重元素标记、一种不标记的脱氧核苷酸为原料，合成其互补链，经透射电镜（TEM, Transmission electron microscopy）检测，则可见重元素标记，其互补链则可由点的大小和强度被分辨出来。





基因组学 | 测序 | 第三代 | SMRT, Nanopore

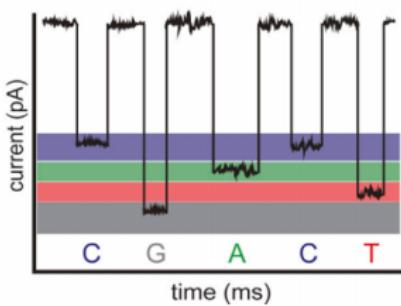
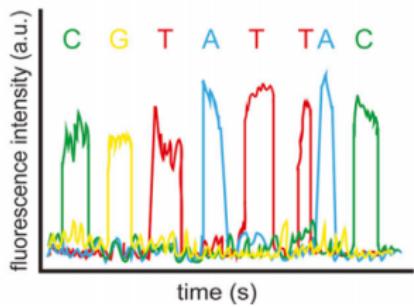
A phospholinked nucleotides



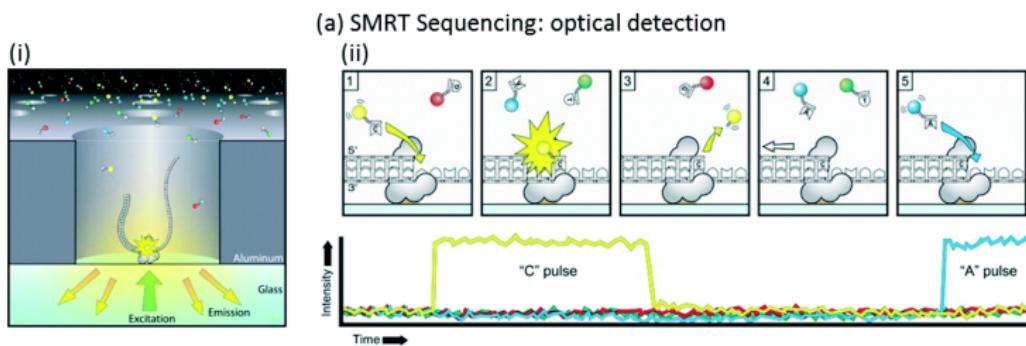
B  HP motor

template
read

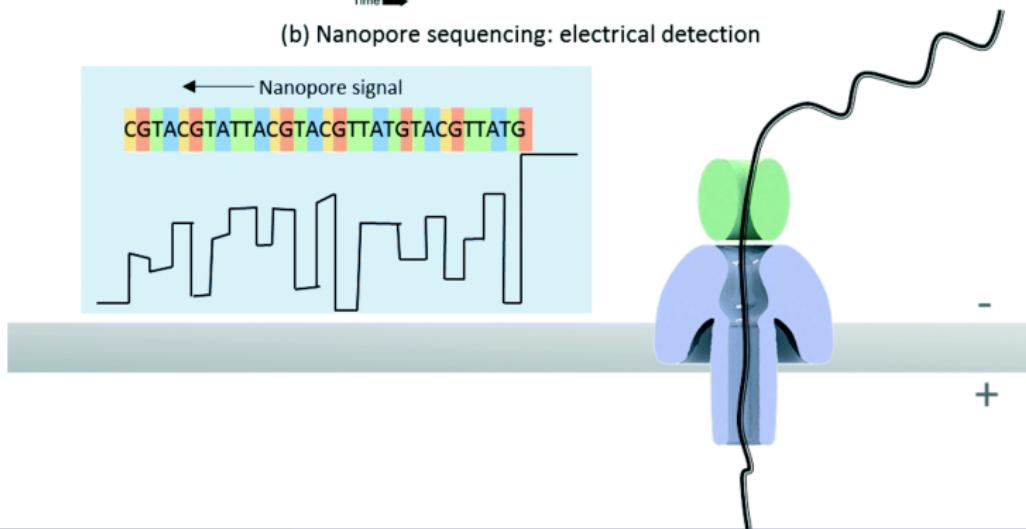
2-direction
read



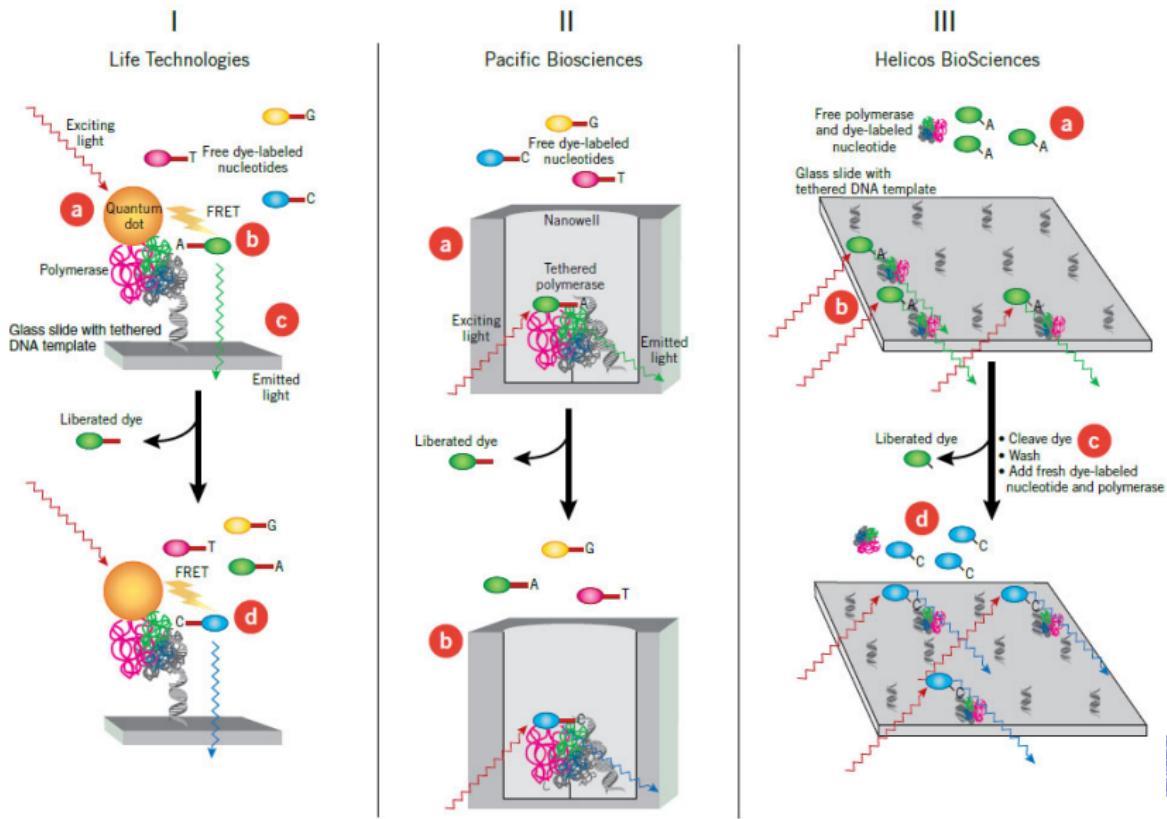
基因组学 | 测序 | 第三代 | SMRT, Nanopore



(b) Nanopore sequencing: electrical detection

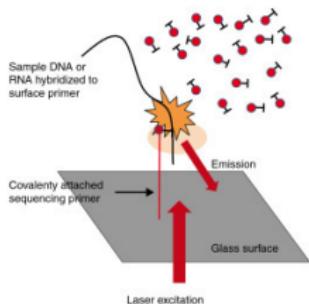


基因组学 | 测序 | 第三代 | FRET, SMRT, tSMS

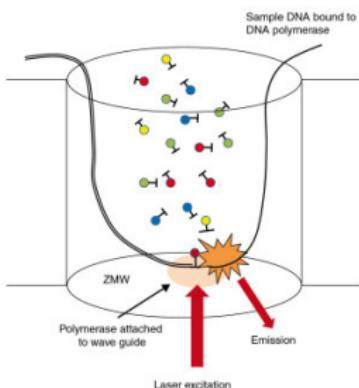


基因组学 | 测序 | 第三代 | tSMS, SMRT, FRET

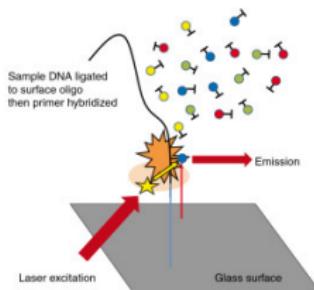
(a) Helicos BioSciences



(b) Pacific Biosciences

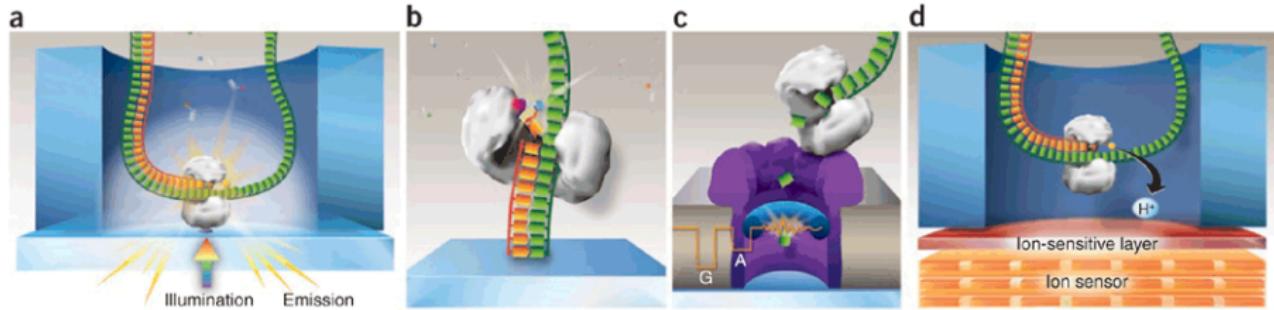


(c) Life Technologies



Key:

- Fluorescently labeled nucleotide
- DNA/RNA Polymerase
- Sequencing primer
- Sample DNA/RNA
- Quantum dot
- Surface-attached primer complement
- FRET transfer



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题

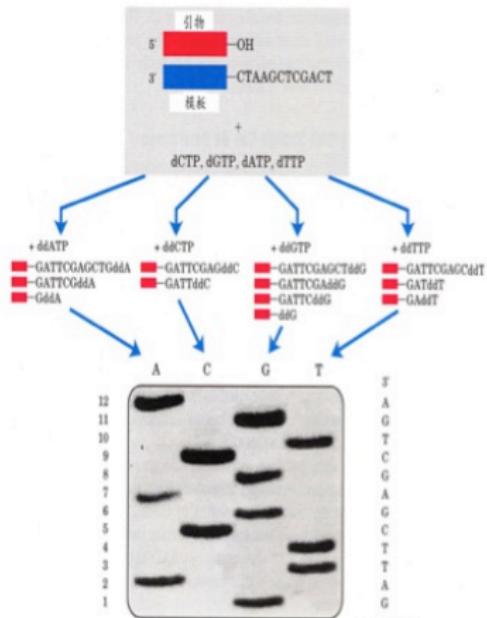
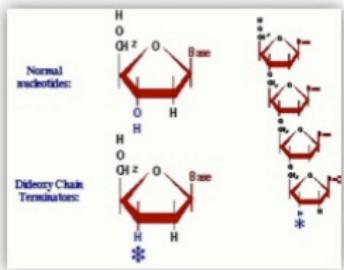




第一代测序技术—Sanger 测序法



Dr. Fred Sanger



优点：

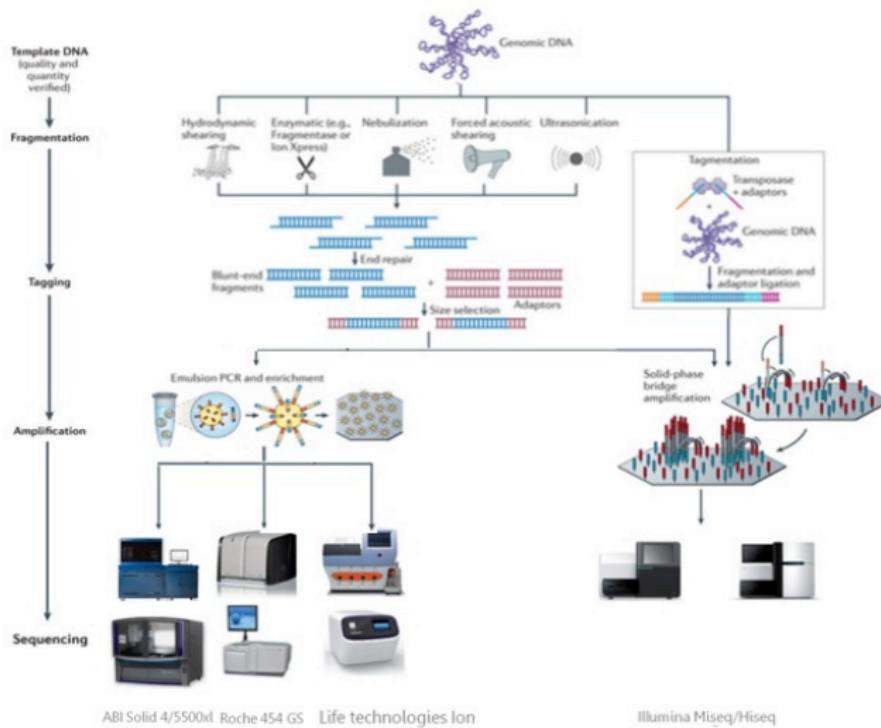
- ✓ 准确
- ✓ 读长长

缺点：

- ✓ 通量低
- ✓ 速度慢
- ✓ 成本高



第二代测序技术



优点：

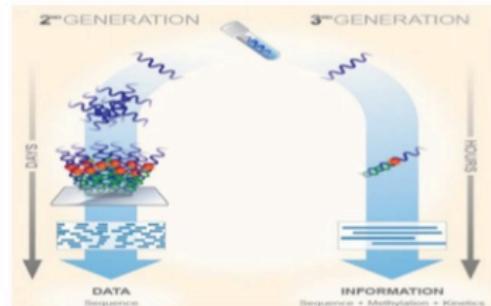
- ✓ 通量高
- ✓ 成本低
- ✓ 时间短

缺点：

- ✓ 读长短
- ✓ 效率不一致



第三代测序技术



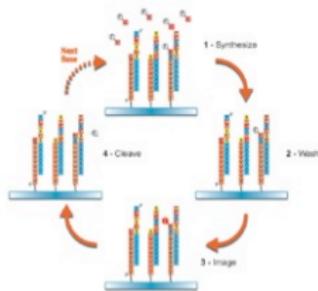
Difference between 2 and 3 generation

优点:

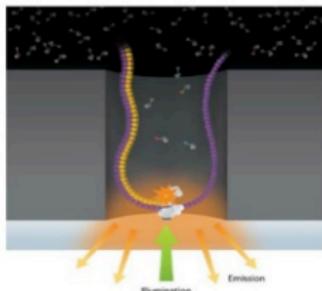
- ✓ 无扩增
- ✓ 直接观察
- ✓ 速度快
- ✓ 长读长

缺点:

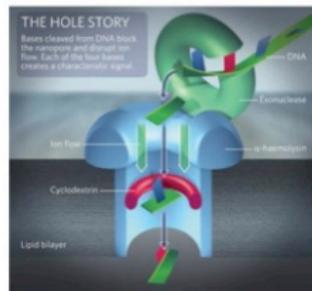
- ✓ 错误率高
- ✓ 可靠性差



Helicos Sequencing



Pacific Biosciences SMRT



Nanopore Sequencing

各代测序技术特点总结



	First generation	Second generation ^a	Third generation ^a
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



基因组学 | 测序 | 比较

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

基因组学 | 测序 | 比较

Method	Read length	Accuracy (single read not consensus)	Reads per run	Time per run	Cost per 1 million bases (in US\$)	Advantages	Disadvantages
Single-molecule real-time sequencing (Pacific Biosciences)	10,000 bp to 15,000 bp avg (14,000 bp N50); maximum read length >40,000 bases ^{[61][62][63]}	87% single-read accuracy ^[64]	50,000 per SMRT cell, or 500-1000 megabases ^{[65][66]}	30 minutes to 4 hours ^[67]	\$0.13-\$0.60	Longest read length. Fast. Detects 4mC, 5mC, 6mA. ^[68]	Moderate throughput. Equipment can be very expensive.
Ion semiconductor (Ion Torrent sequencing)	up to 400 bp	98%	up to 80 million	2 hours	\$1	Less expensive equipment. Fast.	Homopolymer errors.
Pyrosequencing (454)	700 bp	99.9%	1 million	24 hours	\$10	Long read size. Fast.	Runs are expensive. Homopolymer errors.
Sequencing by synthesis (Illumina)	MiniSeq, NextSeq: 75-300 bp; MiSeq: 50-600 bp; HiSeq 2500: 50-500 bp; HiSeq 3/4000: 50-300 bp; HiSeq X: 300 bp	99.9% (Phred30)	MiniSeq/MiSeq: 1-25 Million; NextSeq: 130-00 Million, HiSeq 2500: 300 million - 2 billion, HiSeq 3/4000 2.5 billion, HiSeq X: 3 billion	1 to 11 days, depending upon sequencer and specified read length ^[69]	\$0.05 to \$0.15	Potential for high sequence yield, depending upon sequencer model and desired application.	Equipment can be very expensive. Requires high concentrations of DNA.
Sequencing by ligation (SOLID sequencing)	50+35 or 50+50 bp	99.9%	1.2 to 1.4 billion	1 to 2 weeks	\$0.13	Low cost per base.	Slower than other methods. Has issues sequencing palindromic sequences. ^[70]
Chain termination (Sanger sequencing)	400 to 900 bp	99.9%	N/A	20 minutes to 3 hours	\$2400	Long individual reads. Useful for many applications.	More expensive and impractical for larger sequencing projects. This method also requires the time consuming step of plasmid cloning or PCR.



基因组学 | 测序 | 比较

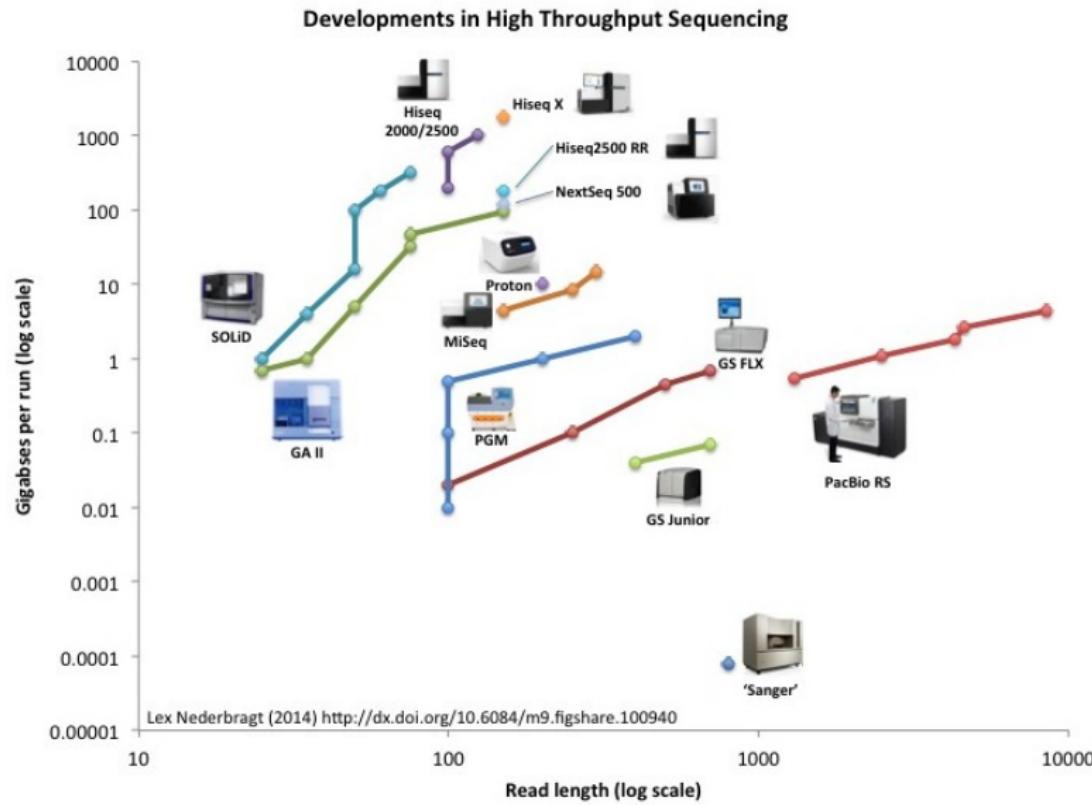
第X代	公司	平台名称	测序方法	检测方法	大约读长(碱基数)	优点	相对局限性
第一代	ABI/生命技术公司	3130xL-3730 xL	桑格-毛细管电泳测序法	荧光/光学	600-1000	高读长，准确度一次 性达标率高，能很好 处理重复序列和多聚 序列	通量低；样品制备成本高 ，使之难以做大量的平行 测序
第一代	贝克曼	GeXP遗传分析系统	桑格-毛细管电泳测序法	荧光/光学	600-1000	高读长，准确度一次 性达标率高，能很好 处理重复序列和多聚 序列；易小型化	通量低；单个样品的制备 成本相对较高
第二代	Roche/454	基因组测序仪 FLX系统	焦磷酸测序法	光学	230-400	在第二代中最高读长 ；比第一代的测序通量大	样品制备较难；难于处理 重复和同种碱基多聚区域 ；试剂冲洗带来错误累积 ；仪器昂贵
第二代	Illumina	HiSeq2000,H iSeq2500/Mi Seq	可逆链终止物和合成测序法	荧光/光学	2x150	很高测序通量	仪器昂贵；用于数据调节 和分析的费用很高
第二代	ABI/Solid	5500xL/Solid 系统	连接测序法	荧光/光学	25-35	很高测序通量；在广 为接受的几种第二代 平台上，所要拼接出 人类基因组的试剂成本最低	测序运行时间长；读长短 ，造成成本高，数据分析 困难和基因组拼接困难； 仪器昂贵
第二代	赫利克斯	HeliScope	单分子合成 测序法	荧光/光学	25-30	高通量；在第二代中 属于单分子性质的测 序技术	读长短，提高了测序成本 ，降低了基因组拼接的质 量；仪器非常昂贵



基因组学 | 测序 | 比较

第三代	太平洋生物科学公司	PacBio RS	实时单分子DNA测序	荧光/光学	~1000	高平均读长。比第一代的测序时间降低；不需要扩增；最长单个读长接近3000碱基 并不能高效地将DNA聚合酶加到测序阵列中；准确性一次性达标的机会低（81-83%）；DNA聚合酶在阵列中降解；总体上每个碱基测序成本高（仪器昂贵）；
第三代	全基因组学公司	GeXP遗传分析系统	复合探针锚杂交和连接技术	荧光/光学	10	在第三代中通量最高；在所有测序技术中，用于拼接一个人基因组的试剂成本最低；每个测序步骤独立，使错误的累积变得最低 低读长；模板制备妨碍长重复序列区域测序；样品制备费事；尚无商业化供应的仪器
第三代	Ion Torrent/生命技术公司	个人基因组测序仪（PGM）	合成测序法	以离子敏感场效应晶体管检测pH值变化	100-200	对核酸碱基的掺入可直接测定；在自然条件下进行DNA合成（不需要使用修饰过的碱基） 一步步的洗脱过程可导致错误累积；阅读高重复和同种多聚序列时有潜在困难；
第三代	牛津纳米孔公司	gridION	纳米孔外切酶测序	电流	尚未定量	有潜力达到高读长；可以低成本生产纳米孔；无需荧光标记或光学手段 切断的核苷酸可能被读错方向；难于生产出带多重平行孔的装置





基因组学 | 测序 | 比较 | 测序仪

制造商	测序平台		总输出	运转时间	输出/天	读长	# of Single Reads	仪器价格	运行成本
Roche	454	GS FLX+	700Mb	23hrs	700Mb	Up To 1kb	1M	~\$500k	~\$6k
		GS Jr.	35Mb	10hrs	35Mb	~700b	0.1M	\$125K	~\$1k
illumina	HiSeq X Ten		1.8Tb	3 Days	600Gb	2x150	6B	\$1M	~\$12k
	HiSeq 2500	HT v4	1Tb	6 Days	167Gb	2x125	4B	\$740k	~\$29k
		HTv3	600Gb	11 Days	55Gb	2x100	3B	\$740k	~\$26k
		Rapid	180Gb	40hrs	~110Gb	2x150	600M	\$740k	~\$8k
	NextSeq 500	High	129Gb	29hrs	~100Gb	2x150	400M	\$250k	\$4k
		Mid	39Gb	26hrs	~36Gb	2x150	130M	\$250k	
	MiSeq	15Gb	~65hrs	15Gb	2x300	25M	\$125k	~\$1.4k	
Life Tech	Solid 5500xl	95Gb	6Days	16Gb	2x60	800M	\$595k	~\$10k	
	Solid 5500xl Wildfire	240Gb	10Days	24Gb	2x50	2.4B	\$70k Upgrade	~\$5k	
	Solid 5500	48Gb	6Days	8Gb	2x60	400M	\$349kk	~\$5k	
	Solid 5500 Wildfire	120Gb	10Days	12Gb	2x50	1.2B	\$70k Upgrade	~\$2.5k	
	IonTorrent	PGM 314	Up to 100Mb	2-4hrs	Up to 200MB	Up to 400b	Up to 0.6M	\$50k	
		PGM 316	Up to 1Gb	3-5hrs	Up to 2Gb	Up to 400b	Up to 3M	\$50k	
		PGM 318	Up to 2Gb	4-7hrs	Up to 4Gb	Up to 400b	Up to 5.5M	\$50k	
		PI	~10G	2-4hrs	~20Gb	Up to 200b	Up to 82M	\$149k	
		PII	~32Gb(at launch)	2-4hrs	~64Gb	100b	Up to 330M	\$149k	
Pacific Biosciences	PacBio RS II;P6-C4		Up to 240min	~500Mb-1Gb	~2Gb	10-15kb		~\$700k	~\$400



基因组学 | 测序 | 比较 | 测序仪

Platform	Sequencing Chemistry	Number of Reads	Peak Data Output	Suitable Applications
NextSeq 500	2 × 150 bp	400 M paired	100 - 120 Gb	<ul style="list-style-type: none"> Eukaryotic Whole Genome de novo* RNA-seq de novo* Whole Genome metagenome RNA-seq Reseq
	2 × 75 bp	400 M paired	50 - 60 Gb	<ul style="list-style-type: none"> Eukaryotic Whole Genome reseq RNA-seq Reseq* ChIP-seq (histone modifications)
	1 × 75 bp	400 M	25 - 30 Gb	<ul style="list-style-type: none"> Small RNA Profiling* ChIP-seq *
MiSeq	2 × 150 bp	12 - 15 M paired	4.5 - 5 Gb	<ul style="list-style-type: none"> Targeted sequencing
	2 × 300 bp	22 - 25 M paired V3	13 - 15 Gb	<ul style="list-style-type: none"> 16S rRNA metagenome sequencing* Bacterial Whole genome De novo* De novo transcriptome HLA typing* Genome Gap closures
Ion Proton	1 × 120 - 160 bp	100 - 150 M	10 - 12 Gb	<ul style="list-style-type: none"> Whole Exome* RNA-seq reseq Bacterial Genome reseq
Ion Torrent PGM	1 × 300 - 400 bp	0.1 - 10 M	10 Mb - 1 Gb	<ul style="list-style-type: none"> Targeted sequencing Viral genomes and plasmids*

* Indicates the preferred platform for the application

教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题

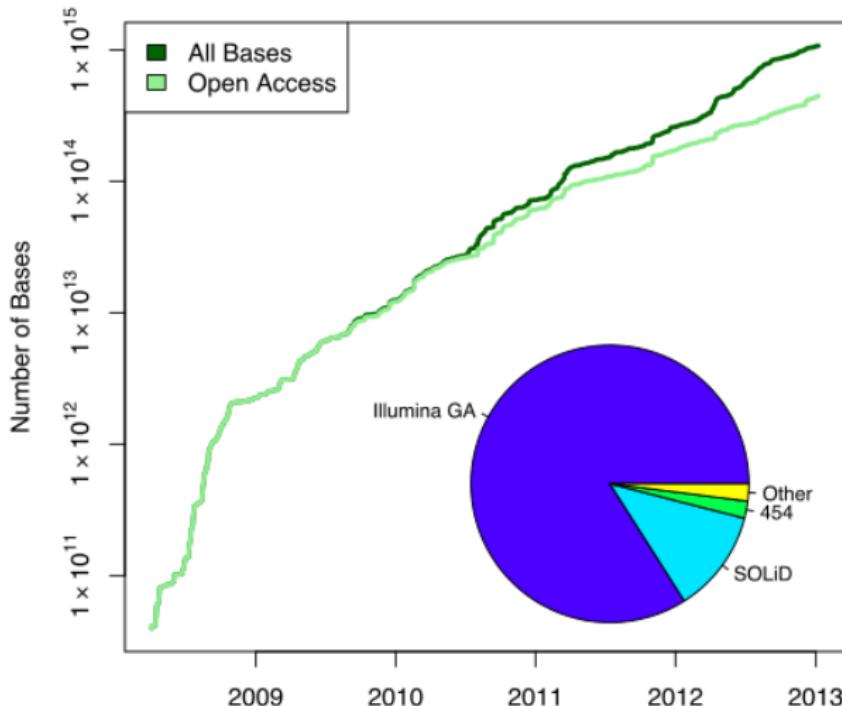


SRA

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®.



NCBI Resources ▾ How To ▾ Sign in to NCBI

SRA SRA Advanced Search Help



SRA

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®.

Getting Started

[How to Submit](#)[Login to SRA](#)[Login to Submission Portal](#)[SRA Handbook](#)[Download Guide](#)[SRA Fact Sheet \(.pdf\)](#)

Tools and Software

[Download SRA Toolkit](#)[SRA Toolkit Documentation](#)[SRA-BLAST](#)[SRA Run Browser](#)[SRA Run Selector](#)

Related Resources

[Submission Portal](#)[Trace Archive](#)[dbGaP Home](#)[BioProject](#)[BioSample](#)

[SRX214992](#): DGE sequencing of Human tumor 3

1 ILLUMINA (Illumina HiSeq 2000) run: 4.4M spots, 92.9M bases, 59.8Mb downloads

Submitted by: SYSU

Study: Homo sapiens Transcriptome or Gene expression

[PRJNA185379](#) • [SRP017786](#) • [All experiments](#) • [All runs](#)

[show Abstract](#)

Sample: DGE sequencing of Human tumor 3

[SAMN01883035](#) • [SRS383504](#) • [All experiments](#) • [All runs](#)

Organism: [Homo sapiens](#)

Library:

Instrument: Illumina HiSeq 2000

Strategy: RNA-Seq

Source: TRANSCRIPTOMIC

Selection: RANDOM

Layout: SINGLE

Spot descriptor:

1 forward

Runs: 1 run, 4.4M spots, 92.9M bases, [59.8Mb](#)

Run	# of Spots	# of Bases	Size	Published
SRR645375	4,424,492	92.9M	59.8Mb	2015-07-22

ID: 294252



 **Sequence Read Archive**

Main Browse Search Download Submit Documentation Software Trace Archive Trace Assembly Trace Home Trace BLAST
Studies Samples Analyses Run Browser Run Selector Provisional SRA

DGE sequencing of Human tumor 3 (SRR645375)

Metadata Reads Download

Run	Spots	Bases	Size	GC content	Published	Access Type
SRR645375	4.4M	92.9Mbp	62.7M	43.5%	2015-07-22	public

Quality graph (bigger)

This run has 1 read per spot:

L=21, 100%

Legend

Experiment Library

[SRX214992](#) Name Platform Strategy Source Selection Layout
[to BLAST](#) Illumina RNA-Seq TRANSCRIPTOMIC RANDOM SINGLE

Biosample Sample Description Organism
[SAMN01883035 \(SRS383504\)](#) Homo sapiens

Bioproject SRA Study Title
[PRJNA185379](#) [SRP017786](#) Homo sapiens Transcriptome or Gene expression
[Show abstract](#)



 **Sequence Read Archive**

Main Browse Search Download Submit Documentation Software Trace Archive Trace Assembly Trace Home Trace BLAST
Studies Samples Analyses Run Browser Run Selector Provisional SRA

DGE sequencing of Human tumor 3 (SRR645375)

Metadata Reads Download

Filter: Find Filtered Download [What does it do?](#)

[What can the filter be applied to?](#)

< 1 1 442450 >

View: biological reads technical reads quality scores [advanced](#)

Read

1. [SRR645375.1 SRS383504](#)
name: FCD1AAWACXX:3:1101:2007:2196.
member: default
x: 2007, y: 2196

2. [SRR645375.2 SRS383504](#)
name: FCD1AAWACXX:3:1101:2044:2209.
member: default
x: 2044, y: 2209

3. [SRR645375.3 SRS383504](#)
name: FCD1AAWACXX:3:1101:2330:2199.
member: default
x: 2330, y: 2199

4. [SRR645375.4 SRS383504](#)
name: FCD1AAWACXX:3:1101:2765:2226

>gnl|SRA|SRR645375.1 FCD1AAWACXX:3:1101:2007:2196
CATGTACTTTAGCTAGTTT

One channel quality score

C:34 A:34 T:34 G:34 T:31 A:30 C:30 T:32 T:35 T:35 T:35 A:35 G:30 C:33 T:34 A:37
G:29 T:38 T:37 T:38 T:35

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.



基因组学 | 测序 | 数据库 | GEO

NCBI Resources How To

[Sign in to NCBI](#)

[GEO Home](#) [Documentation](#) [Query & Browse](#) [Email GEO](#)

Gene Expression Omnibus

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.



Getting Started

[Overview](#)
[FAQ](#)
[About GEO DataSets](#)
[About GEO Profiles](#)
[About GEO2R Analysis](#)
[How to Construct a Query](#)
[How to Download Data](#)

Tools

[Search for Studies at GEO DataSets](#)
[Search for Gene Expression at GEO Profiles](#)
[Search GEO Documentation](#)
[Analyze a Study with GEO2R](#)
[GEO BLAST](#)
[Programmatic Access](#)
[FTP Site](#)

Browse Content

[Repository Browser](#)
DataSets: 3848
Series: 71107
Platforms: 16059
Samples: 1863122

Information for Submitters

[Login to Submit](#)

[Submission Guidelines](#)
[Update Guidelines](#)

[MIAME Standards](#)
[Citing and Linking to GEO](#)
[Guidelines for Reviewers](#)
[GEO Publications](#)



基因组学 | 测序 | 数据库 | GEO

Platforms (1)	GPL17021 Illumina HiSeq 2500 (Mus musculus)
Samples (15) + More...	GSM2176510 LSCs in gonadal adipose tissue (replicate 1) GSM2176511 LSCs in gonadal adipose tissue (replicate 2) GSM2176512 LSCs in gonadal adipose tissue (replicate 3)

Relations

BioProject	PRJNA322680
SRA	SRP075661

Download family	Format
SOFT formatted family file(s)	SOFT
MINIML formatted family file(s)	MINIML
Series Matrix File(s)	TXT

Supplementary file	Size	Download	File type/resource
GSE81842_AT_vs_NBm.txt.gz	904.8 Kb	(ftp)(http)	TXT
GSE81842_BL_vs_AT.txt.gz	901.7 Kb	(ftp)(http)	TXT
GSE81842_BL_vs_BM.txt.gz	906.4 Kb	(ftp)(http)	TXT
GSE81842_BL_vs_NBm.txt.gz	906.6 Kb	(ftp)(http)	TXT
GSE81842_BL_vs_Spl.txt.gz	884.1 Kb	(ftp)(http)	TXT
GSE81842_BM_vs_AT.txt.gz	904.1 Kb	(ftp)(http)	TXT
GSE81842_BM_vs_NBm.txt.gz	897.2 Kb	(ftp)(http)	TXT
GSE81842_Spl_vs_AT.txt.gz	881.7 Kb	(ftp)(http)	TXT
GSE81842_Spl_vs_BM.txt.gz	867.7 Kb	(ftp)(http)	TXT
GSE81842_Spl_vs_NBm.txt.gz	872.7 Kb	(ftp)(http)	TXT
GSE81842_Summary.txt.gz	271 b	(ftp)(http)	TXT
GSE81842_frm_gene_exp.diff.txt.gz	8.7 Mb	(ftp)(http)	TXT
GSE81842_frm_genes.count_tracking.txt.gz	763.1 Kb	(ftp)(http)	TXT
GSE81842_frm_genes.fpkm_tracking.txt.gz	1.4 Mb	(ftp)(http)	TXT
SRP/SRP075/SRP075661		(ftp)	SRA Study

Raw data provided as supplementary file

Processed data is available on Series record



基因组学 | 测序 | 数据库 | GEO

Library strategy RNA-Seq
Library source transcriptomic
Library selection cDNA
Instrument model Illumina HiSeq 2500

Description AT1
Data processing (Illumina -> FastQ) & DEMULTIPLEXING - bcltofastq-1.8.4
Quality filter raw read data: Trimmomatic-0.32
Read alignment: SHRIMP_2_2_3
Data normalization and differential expression analysis: cufflinks-
2.0.2.Linux_x86_64 (cuffdiff)
Genome build: mm10

Submission date May 24, 2016
Last update date Jul 07, 2016
Contact name Haobin Ye
E-mail haobin.ye@ucdenver.edu
Organization name University of Colorado
Department Hematology
Lab Craig Jordan Lab
Street address 12700 East 19th Ave, Room 9122
City Aurora
State/province Colorado
ZIP/Postal code 80045
Country USA

Platform ID [GPL17021](#)
Series (1) [GSE81842](#) Genome-wide comparison of gene expression level between leukemia stem cells in different tissues

Relations

BioSample [SAMN05172382](#)
SRA [SRX1798522](#)

Supplementary file	Size	Download	File type/resource
SRX/SRX179/SRX1798522	(ftp)		SRA Experiment

Raw data provided as supplementary file



千人基因组计划

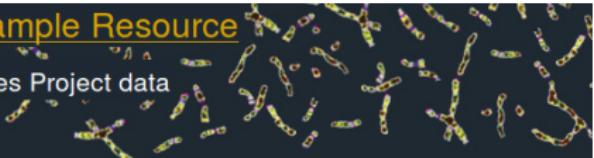
千人基因组计划（1000 Genomes Project），旨在绘制迄今（截至 2011 年）最详尽、最有医学应用价值的人类基因多态性图谱，该图谱由中美英等国科研机构发起的“千人基因组计划”共同协作完成，标志着人类基因研究取得重大突破。

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。



IGSR: The International Genome Sample Resource

Providing ongoing support for the 1000 Genomes Project data

[Home](#)[About](#)[Data](#)[Portal](#)[Analysis](#)[Contact](#)[Browser](#)[FAQ](#)

IGSR and the 1000 Genomes Project



Populations: ● - African; ● - American; ● - East Asian; ● - European; ● - South Asian;

The International Genome Sample Resource (IGSR) was established to ensure the ongoing usability of data generated by the 1000 Genomes Project and to extend the data set. More information is available [about the IGSR](#).

Links

Announcements

[IGSR Sample Collection Principles](#)

[1000 Genomes Project Publications](#)

File formats

Software tools

Download data

Twitter



Using data from IGSR

IGSR provides open data to support the community's research efforts. You can see our terms of use in our [data disclaimer](#).

Data portal *beta*

We are developing a new data portal to make it easier to find and browse data in IGSR. You can use the development version to [explore the data set](#). Let us know what you think at info@1000genomes.org.

Sample	Sex	Population	Exome	Low cov WL	High cov WL	HD genotype	Complete
HG00513	Female	CHS	●	●	●	●	
HG01112	Male	CLM	●	●	●	●	
HG00759	Female	CDX	●	●	●	●	
HG01500	Male	IBS	●	●	●	●	
HG03006	Male	BEB	●	●	●		
NA18525	Female	CHB	●	●	●	●	
NA19648	Female	MXL	●	●	●	●	



Data collections for HG00119

[1000 Genomes on GRCh38](#)[1000 Genomes phase 3 release](#)[1000 Genomes phase 1 release](#)

 Data reuse policy for 1000 Genomes on GRCh38

22 matching data files

[Download the list](#)

Data types

- Sequence
- Alignment

[« Previous](#)[Next »](#)

File

Analysis group

 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR043/SRR043348/SRR043348_1.fastq.gz

Low coverage WGS

 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR043/SRR043354/SRR043354_1.fastq.gz

Low coverage WGS

 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR043/SRR043372/SRR043372_1.fastq.gz

Low coverage WGS

 ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR043/SRR043378/SRR043378_2.fastq.gz

Low coverage WGS

 ftp://ftp.sra.ebi.ac.uk/vol1/fasta/SRR099/SRR099967/SRR099967_1.fasta

Exome

Analysis groups

- Exome
- Low coverage WGS



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。





THE CANCER GENOME ATLAS

National Cancer Institute

National Human Genome Research Institute

[Launch Data Portal](#) | [Contact Us](#) | [For the Media](#)

Search



Search

Home

About Cancer Genomics

Cancers Selected for Study

Research Highlights

Publications

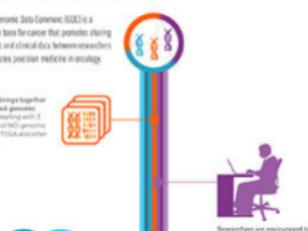
News and Events

About TCGA

NATIONAL CANCER INSTITUTE GENOMIC DATA COMMONS

The NCI Genomic Data Commons (GDC) is a knowledge base for cancer that promotes sharing of genomic and clinical data between researchers and facilitates precision medicine in oncology.

The GDC brings together harmonized genomic datasets, starting with cancer genome data from TCGA and other institutes.



The NCI Genomic Data Commons Launches

The Genomic Data Commons (GDC) is a data sharing platform that promotes precision medicine in oncology, and it will host all of the TCGA data.

[Learn More ▶](#)

Genomic Data
Commons
Launches

Analysis of
Adrenocortical
Carcinoma

Cancers
Selected
for
Study

About TCGA

Research Briefs



July 2016

Longitudinal Study Charts Brain Tumor Evolution

News and Announcements



June 06, 2016

Newly launched Genomic Data Commons to facilitate

Launch Data Portal

The Genomic Data Commons (GDC) Data Portal is an interactive data system for researchers to search, download, upload, and analyze harmonized cancer genomic data sets, including TCGA.

Questions About Cancer

Visit www.cancer.gov

Call 1-800-4-CANCER

Use [LiveHelp Online Chat](#)

Multimedia Library

Images

Videos and Animations

Podcasts



基因组学 | 测序 | 数据库 | TCGA

NATIONAL CANCER INSTITUTE
GDC Data Portal

Home Projects Data Analysis

Harmonized Cancer Datasets
Genomic Data Commons Data Portal

Get Started by Exploring:

Projects Data

Perform Advanced Search Queries, such as:

Cases of kidney cancer diagnosed at the age of 20 and below 182 Cases 1,501 Files

CNV data of female brain cancer cases 459 Cases 1,788 Files

Gene expression quantification data in TCGA-GBM project 166 Cases 522 Files

DATA PORTAL SUMMARY
Data Release 1.0 - June 6, 2016

PROJECTS 39

PRIMARY SITE 29

CASES 14,531

FILES 250,498

Infrastructure
Data is continuously being processed and harmonized by the GDC.
View GDC system statistics:

Compute Infrastructure	12,800 Cores	87.96 TB RAM
Storage Infrastructure	4.98 PB Used	5.42 PB Total

[View Data Download Statistics Report >](#)

Documentation
Learn how to use the GDC Data Portal to its full potential with common topics such as:

Browse Data using Facet Search
Search Data with Advanced Search Technology
Project Based Data Availability
Controlled Access Data
[Visit the Documentation Website >](#)

Cases by Primary Site

Primary Site	Cases
Kidney	~1,300
Brain	~1,100
Bladder	~1,100
Breast	~950
Colon	~850
Uterus	~650
Head and Neck	~600
Pancreas	~550
Stomach	~500
Lung	~450
Liver	~400
Esophagus	~350
Prostate	~300
Bladder	~300
Leukemia	~250
Adenocarcinoma	~200
Bone	~150
Esophageal	~100
Thyroid	~100
Lymphoma	~50
Bladder	~50
Bladder	~50



ICGC  Data Portal Get Cancer Data Data Access Compliance Office Apply for Access to Controlled Data  Contact Us Log In | Create an Account



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.

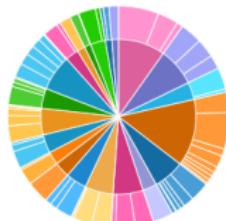


PCAWG
PanCancer Analysis
OF WHOLE GENOMES

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...](#)



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

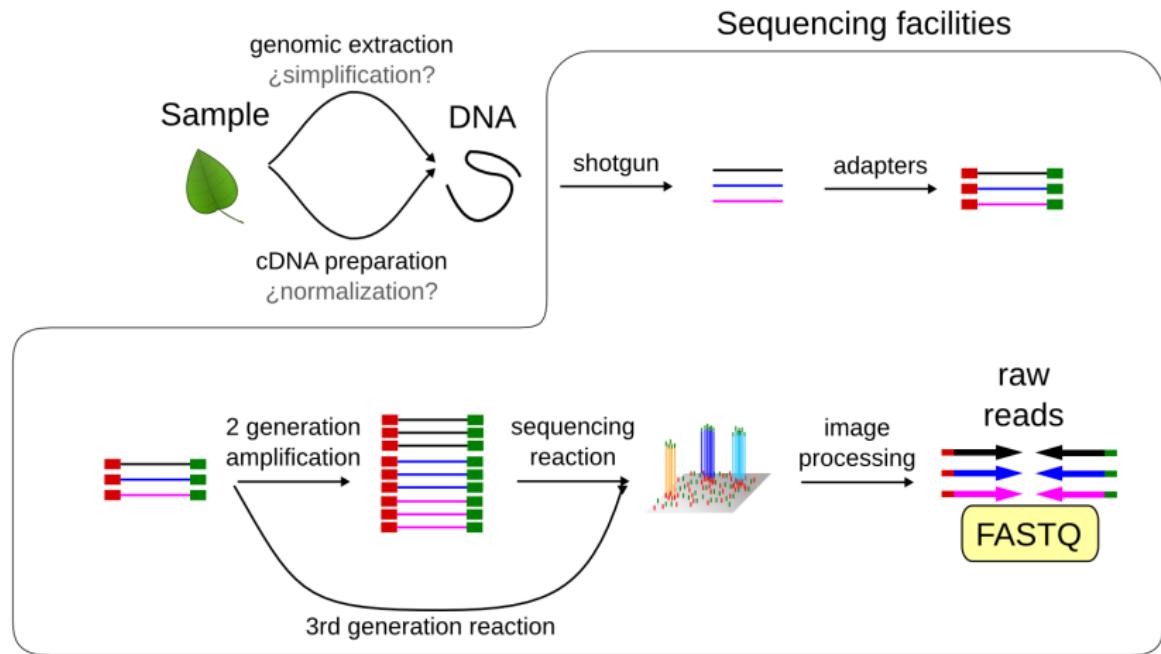
5 外显子组测序

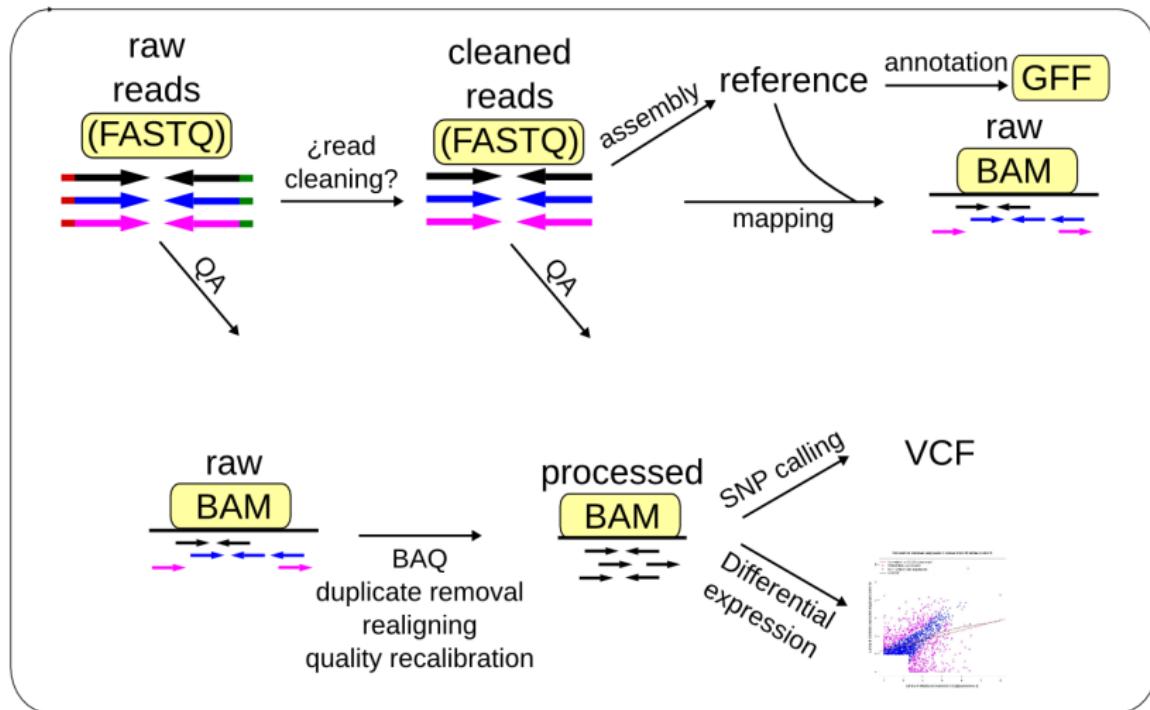
- 简介
- 操作流程
- 应用实例

6 回顾与总结

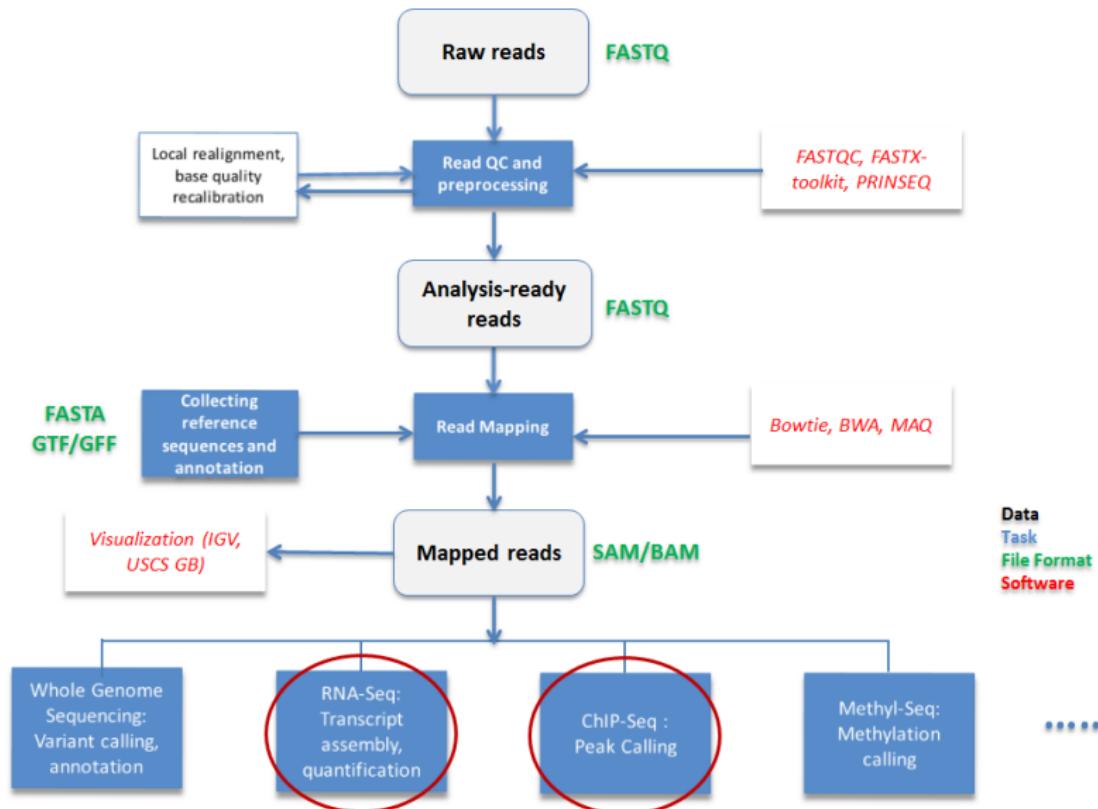
- 总结
- 思考题

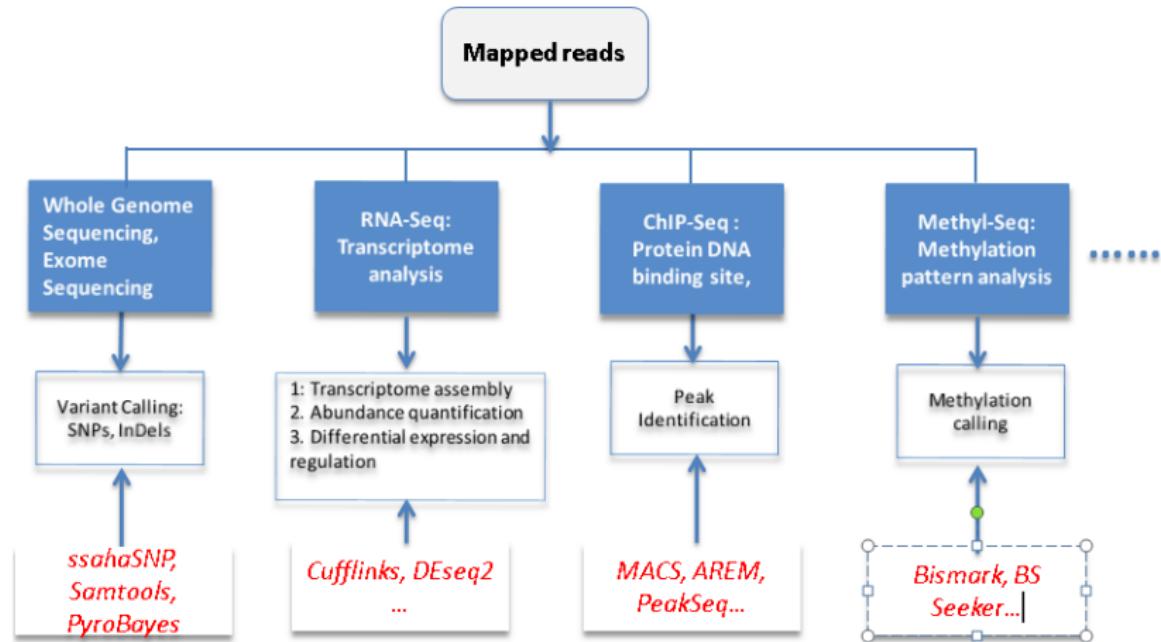




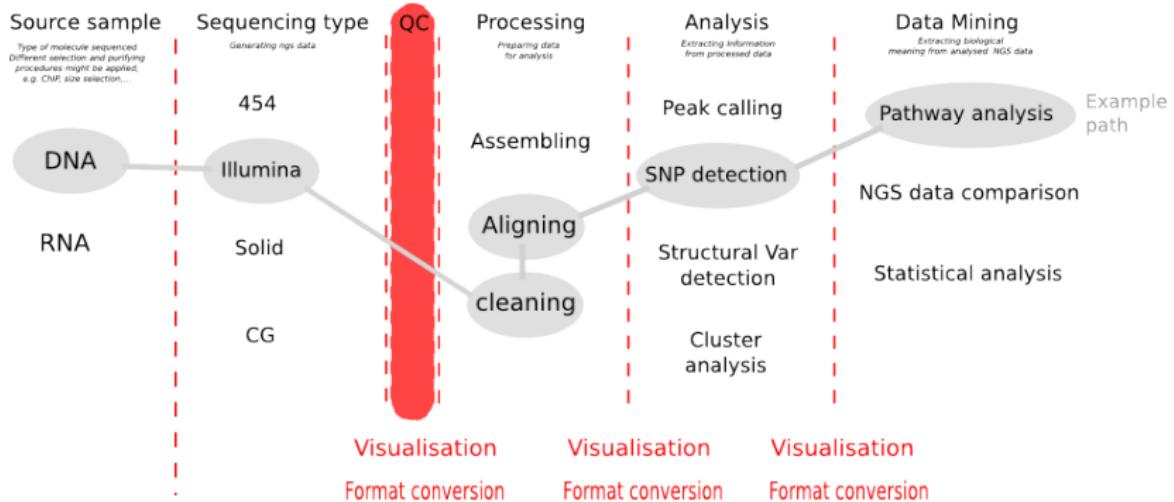


基因组学 | 测序 | 数据格式 | 概览





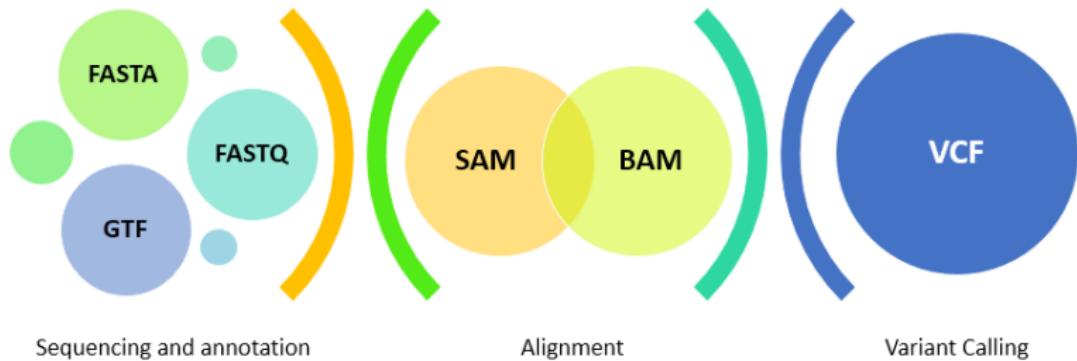
基因组学 | 测序 | 数据格式 | 概览



Common NGS Data Formats

File extension	Description	Reference	Publication
.fasta	Classic DNA sequence file format	http://www.ncbi.nlm.nih.gov/blast/fasta.shtml	n/a
.ace	File format for whole-genome assemblies	Annotated in the documentation for CONSED, currently: http://www.phrap.org/consed/distributions/README.19.0.txt	Gordon, Abajian, and Green, 1998
.wig	A reference-genome indexed data series for "dense" and continuous data (such as %GC)	http://genome.ucsc.edu/goldenPath/help/wiggle.html	Haussler, 2002
.bed	A reference-genome indexed data series for "sparse" data (such as transcriptome data)	http://genome.ucsc.edu/goldenPath/help/bedgraph.html	Haussler, 2002
.tab	Tab-delimited text	N/A	n/a
.pdf	Portable document format	Either ISO-32000-1 or http://www.adobe.com/devnet/pdf/pdf_reference.html	n/a
.sam	"Sequence Alignment/Map" format	http://samtools.sourceforge.net/SAM1.pdf	Li, 2009
.bam	Binary format of .sam	http://samtools.sourceforge.net/SAM1.pdf	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.	http://maq.sourceforge.net/fastq.shtml	Li, 2008, and Cock, 2009
.csfasta	Life Technologies SOLID colorspace fasta file - containing color calls (0, 1, 2, 3) rather than base calls	See: http://solidsoftwaretools.com/	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.	http://www.sanger.ac.uk/resources/software/gff/	n/a
.srf	"Short Read Format" - a new format proposed for short-read DNA sequence	http://srf.sourceforge.net	n/a
.sff	Standard Flowgram Format (specific for Roche/454)	http://www.ncbi.nlm.nih.gov/Traces/trace.cgi?cmd=show&f=format&m=doc&s=formats#header-global	n/a
.gtf	Gene transfer format, an alternate format to GFF for specifying gene features	http://mblab.wustl.edu/GTF22.html	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)



★ FASTQ Files

FASTQ format is a text-based format for storing a biological sequence and its corresponding quality scores. See http://en.wikipedia.org/wiki/FASTQ_format.

★ Bed Files

BED format provides a way for defining genomic regions. We will use BED format to define target regions e.g., exons being targeted for sequence capture. The first three required fields specify: name of chromosome, start position and end position. For more information on BED format see <http://genome.ucsc.edu/FAQ/FAQformat.html#format1>.

★ VCF Files

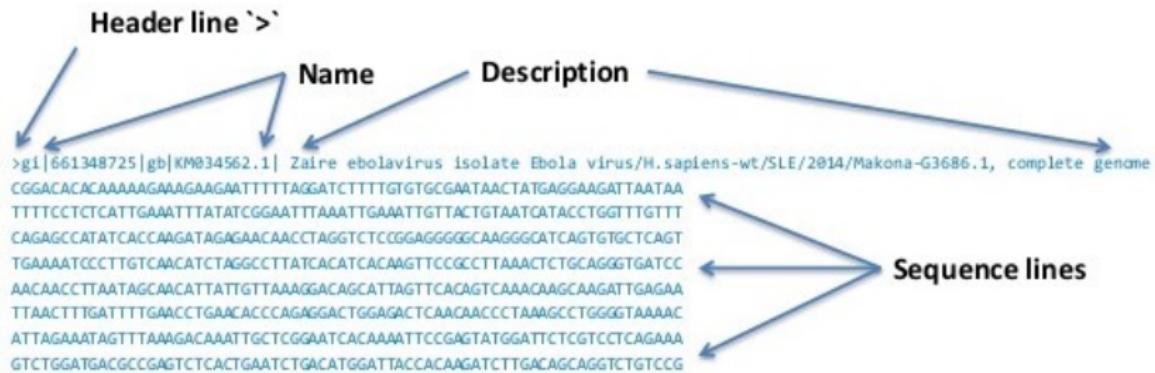
Variant Call Format (VCF), not to be confused with the standard file format for storing contact information, is a specification for storing sequence variations. For more information on VCF format see http://en.wikipedia.org/wiki/Variant_Call_Format.



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  
GGGAAGGAAGCAAGCGGACAGTGAGCTGAGGCCGGTGCCGGCAGGCCGCCACCTGGCCC
```

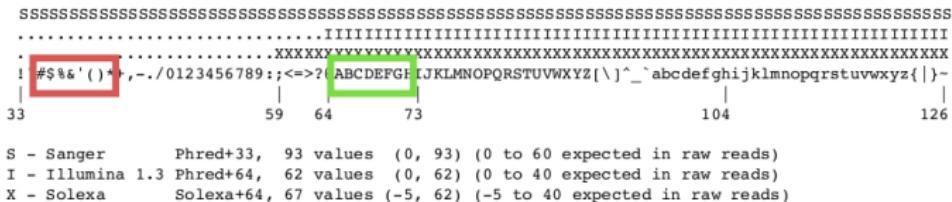


What is FASTQ?

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

```
@SEQ_ID  
GATTGGGGTTCAAAAGCAGTATCGATCAAATAGTAAATCCATTGTTCAACTCACAGTT  
+  
''''(((((****+)%%%)++)(%%)%)-1***-+*'') )**55CCE>>>>CCCCCCCC65
```

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



http://en.wikipedia.org/wiki/FASTQ_format



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



FASTQ

FASTQ format is a text-based format for storing both a biological sequence (usually nucleotide sequence) and its corresponding quality scores. Both the sequence letter and quality score are each encoded with a single ASCII character for brevity.

It was originally developed at the Wellcome Trust Sanger Institute to bundle a FASTA sequence and its quality data, but has recently become the *de facto* standard for storing the output of high-throughput sequencing instruments such as the Illumina Genome Analyzer.

There is no standard file extension for a FASTQ file, but **.fq** and **.fastq**, are commonly used.



- **FastA** format (everybody knows about it)
 - Header line starts with “>” followed by a sequence ID
 - Sequence (string of nt).
- **FastQ** format
 - First is the sequence (like Fasta but starting with “@”)
 - Then “+” and sequence ID (optional) and in the following line are QVs encoded as single byte ASCII codes
 - Different quality encode variants



基因组学 | 测序 | 数据格式 | FASTQ

FASTQ Format

The FASTQ format stores sequences and **Phred qualities scores** in a single file. FASTQ file uses four lines per sequence:

- Line 1 - begins with a '@' character and is followed by a sequence identifier and an optional description (like a sequence description),
- Line 2 - is the raw sequence letters,
- Line 3 - begins with a '+' character,
- Line 4 - encodes the quality values for the sequence in Line 2.

```
@WGG97JN1:192:C200YACXX:7:1101:1307:1960 1:N:0:TTAGGC  
CGAGGAGCTGAGTCACAGAGCAGAAGGGTTTCAGAGATTGGCTGTCCA  
+  
@FFFFFHCFHHIEGIIGIJIGHGGHJIJIIJJGGGHFI7@  
@WGG97JN1:192:C200YACXX:7:1101:1602:1991 1:N:0:TTAGGC  
CTGCGGTTCCCTCGTACTGAGCAGGATTACTAGCGCAACAACATCATC  
+  
=?DD@=<AF?DFFF;EBDHCCFFG:E<D<?DFC>GGHD@BG.=@C;FGEE
```

Sequence identifier contains: **WGG97JN1** the unique instrument name; **192** the run id; **C200YACXX** the flowcell id; **7** flowcell lane; **1101** tile number within the flowcell lane; **1307** 'x'-coordinate of the cluster within the tile; **1960** 'y'-coordinate of the cluster within the tile; **1** the member of a pair, 1 or 2 (paired-end or mate-pair reads only); **Y** if the read fails filter (read is bad), **N** otherwise; **0** when none of the control bits are on, otherwise it is an even number; **TTAGGC** index sequence.



Fastq files:

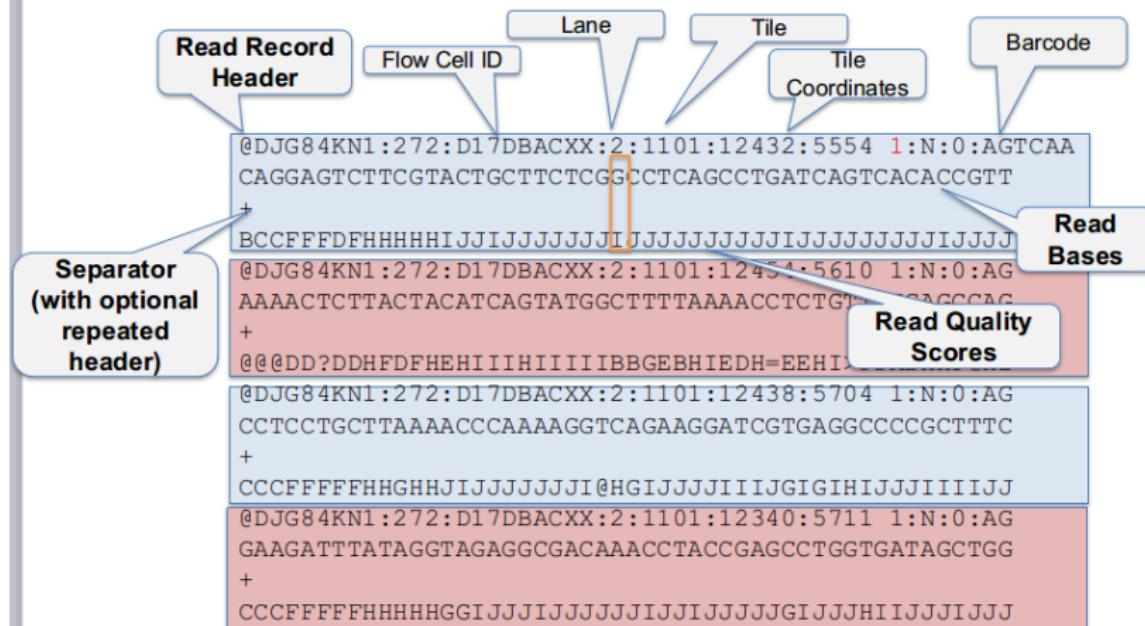
FASTQ format is a **text-based format** for storing both a biological **sequence** (usually nucleotide sequence) and its corresponding **quality scores**.

-Wikipedia

1. Single line ID with at symbol ("@") in the first column.
 2. There should be not space between "@" symbol and the first letter of the identifier.
 3. Sequences are in multiple lines after the ID line
 4. Single line with plus symbol ("+") in the first column to represent the quality line.
 5. Quality ID line can have or have not ID
 6. Quality values are in multiple lines after the + line



FASTQ Format (Illumina Example)

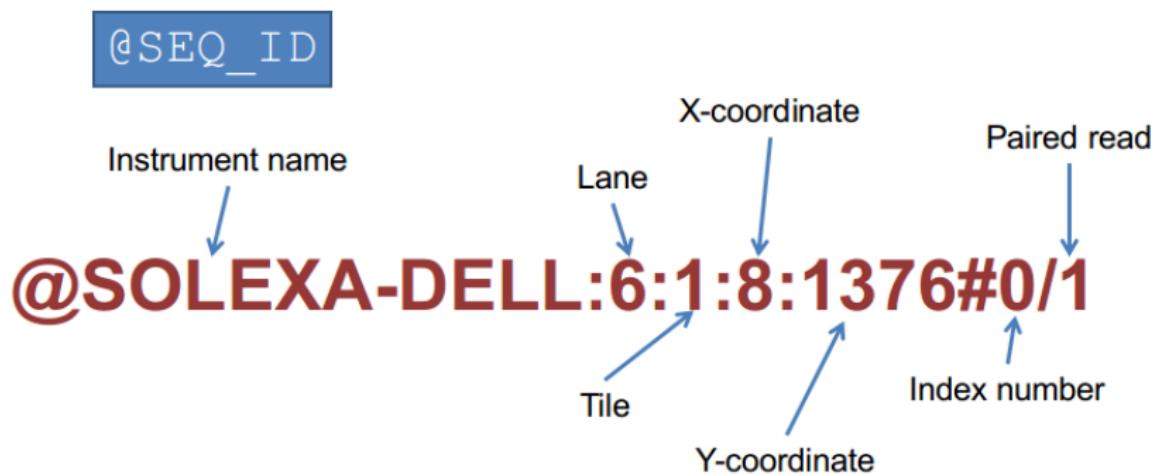


NOTE: for paired-end runs, there is a second file with one-to-one corresponding headers and reads.

(Passarelli, 2012)



Illumina sequence identifiers



Sequences from the Illumina software use a systematic identifier:

@HWUSI-EAS100R:6:73:941:1973#0/1

HWUSI-EAS100R	the unique instrument name
6	flowcell lane
73	tile number within the flowcell lane
941	'x'-coordinate of the cluster within the tile
1973	'y'-coordinate of the cluster within the tile
#0	index number for a multiplexed sample (0 for no indexing)
/1	the member of a pair, /1 or /2 (<i>paired-end or mate-pair reads only</i>)

Versions of the Illumina pipeline since 1.4 appear to use **#NNNNNN** instead of **#0** for the multiplex ID, where **NNNNNN** is the sequence of the multiplex tag.



基因组学 | 测序 | 数据格式 | FASTQ | ID

With Casava 1.8 the format of the '@' line has changed:

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:Y:18:ATCACG
```

EAS139	the unique instrument name
136	the run id
FC706VJ	the flowcell id
2	flowcell lane
2104	tile number within the flowcell lane
15343	'x'-coordinate of the cluster within the tile
197393	'y'-coordinate of the cluster within the tile
1	the member of a pair, 1 or 2 (<i>paired-end or mate-pair reads only</i>)
Y	Y if the read is filtered, N otherwise
18	0 when none of the control bits are on, otherwise it is an even number
ATCACG	index sequence

Note that more recent versions of Illumina software output a sample number (as taken from the sample sheet) in place of an index sequence. For example, the following header might appear in the first sample of a batch:

```
@EAS139:136:FC706VJ:2:2104:15343:197393 1:N:18:1
```



Illumina data format

- Fastq format:

```
@ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
GTAGAACTGGTACGGACAAGGGGAATCTGACTGTAG
+ILLUMINA-F6C19_0048_FC:5:1:12440:1460#0/1
hhhhhhhhhhhhhhhhhehhhedhhhhfhhhhh
```

/1 or /2 paired-end

@seq identifier
seq
+any description
seq quality values



```
@SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
GGGTGATGCCGCTGCCGATGGCGTCAAATCCCACC
+SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
```

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

What is a PHRED score?

- Started way back with Sanger sequencing
- Gives the confidence in the base call 1:10 = 10, 1:100 = 20, 1:1000 = 30, 1:10000 = 40
- Kinda clunky, two characters
 - hard to parse
 - big files

```
>SRR014849.1 EIXKN4201CFU84 length=93
GGGGGGGGGGGGGGGGCTTTTTGTTGAAACCGAAAGG
GTTTGAATTCAAACCCCTTCGGTTCCAACCTTCCAA
AGCAATGCCAATA
```

and as a QUAL entry holding the PHRED scores:

```
>SRR014849.1 EIXKN4201CFU84 length=93
18 10 5 3 2 1 1 1 1 1 1 1 1 1 1 22 37
31 22 16 11 6 1 26 34 30 11 33 26 30 21
33 26 25 36 32 16 36 32 16 36 32 20 6
24 33 25 30 25 2 24 36 32 15 35 31 17
36 32 20 6 25 29 20 30 25 4 32 26 32 23
32 26 30 24 33 26 35 31 14 28 27 30 22
28 24 27 17 32 23 28 28
```

$$Q_{\text{PHRED}} = -10 \times \log_{10}(P_e)$$



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%



Sequence Encoding (FASTQ)

- The base calling (A, T, G or C) is performed based on Phred Scores.

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

- Phred scores provide \log_{10} -transformed error probability values
 → If p is probability that the base call is wrong the Phred score is

$$Q = -10 \log_{10} (P) \Leftrightarrow P = 10^{-Q/10}$$

基因组学 | 测序 | 数据格式 | FASTQ | Quality | Encoding



Phred quality score, a measure of base call quality

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

Phred quality scores are logarithmically linked to error probabilities

Phred Quality Score	Probability of incorrect call	Base call accuracy
10	1 in 10	90%
20	1 in 100	99%
30	1 in 1000	99.9%
40	1 in 10000	99.99%
50	1 in 100000	99.999%

The quality score is ASCII encoded in the FASTQ format

FASTQ is a FASTA with score

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



Anatomy of SAM file

```
Read1 113 1 497 37 37M      15 38662 0 CGGGTCTGACC  0;====9;>>> NM:i:0
Read2 213 1 497 37 37M      15 37662 0 CGGGTCTGACC  0;====45;>>> NM:i:1 XM:i:3
Read3 337 1 497 37 37M      15    38662 0 CGGGTCTGACC ;==9;>>><>; NM:i:0
Read4 615 1 497 37 36MD1 15    447 0 CGGGTCTGACC  0;==5"=69;>> NM:i:0
Read5 844 1 497 37 37M      15   1445 0 CGGGTCTGACC =====9;>>> NM:i:0
```

One line per original read sequence

- Big!
- Where it aligned (if at all)
- How much of it aligned (soft/hard clipping)
- Mapping quality, likelihood correctly aligned
- Any differences to the reference (CIGAR string)
- Lots of other stuff (aligner dependent)
- Does not contain the reference sequence



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

GCCCCCTGGGGATGTTTGCACCAAGCCACTGTCTCCAGCTGG sequence

BBC@GIIHGCFcieHEAIEIFFGEONDNJFINIONHNGJNNNNKNJN 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)

基因组学 | 测序 | 数据格式 | SAM

Each row describes a single alignment of a raw read against the reference genome. Each alignment has 11 mandatory fields, followed by any number of optional fields.

SAM Format Specification

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

基因组学 | 测序 | 数据格式 | SAM

Col	Field	Type	Regexp/Range	Brief description
1	QNAME	String	[!-?A-~]{1,254}	Query template NAME
2	FLAG	Int	[0,2 ¹⁶ -1]	bitwise FLAG
3	RNAME	String	* [!-()+-<>-~] [!-~]*	Reference sequence NAME
4	POS	Int	[0,2 ³¹ -1]	1-based leftmost mapping POSition
5	MAPQ	Int	[0,2 ⁸ -1]	MAPping Quality
6	CIGAR	String	* ([0-9]+ [MIDNSHPX=]) +	CIGAR string
7	RNEXT	String	* = [!-()+-<>-~] [!-~]*	Ref. name of the mate/next read
8	PNEXT	Int	[0,2 ³¹ -1]	Position of the mate/next read
9	TLEN	Int	[-2 ³¹ +1,2 ³¹ -1]	observed Template LENgth
10	SEQ	String	* [A-Za-z.=.] +	segment SEQuence
11	QUAL	String	[!-~] +	ASCII of Phred-scaled base QUALity+33

```
@HD VN:1.0
@SQ SN:chr20 LN:62435964
@RG ID:L1 PU:SC_1_10 LB:SC_1 SM:NA12891
@RG ID:L2 PU:SC_2_12 LB:SC_2 SM:NA12891
read_28833_29006_6945 99 chr20 28833 20 10M1D25M = 28993 195 \
    AGCTTAGCTAGCTACCTATATCTTGGTCTTGGCCG <<<<<<<<<<<<<:<9 /,&,22;;<<< \
    NM:i:1 RG:Z:L1
read_28701_28881_323b 147 chr20 28834 30 35M      = 28701 -168 \
    ACCTATATCTTGGCCTTGGCCGATGCGGCCTTGCA <<<<;<<<7;:<<<6;<<<<<<<<<7<<<< \
    MF:i:18 RG:Z:L2
```



query name = sample name:bead coordinates



Bit	Description
1	0x1 template having multiple segments in sequencing
2	0x2 each segment properly aligned according to the aligner
4	0x4 segment unmapped
8	0x8 next segment in the template unmapped
16	0x10 SEQ being reverse complemented
32	0x20 SEQ of the next segment in the template being reverse complemented
64	0x40 the first segment in the template
128	0x80 the last segment in the template
256	0x100 secondary alignment
512	0x200 not passing filters, such as platform/vendor quality controls
1024	0x400 PCR or optical duplicate
2048	0x800 supplementary alignment

Example

$$99 = 64 + 32 + 2 + 1$$

Decoding SAM flags

<http://broadinstitute.github.io/picard/explain-flags.html>

FLAG meaning in English	FLAG
read paired	1
read mapped in proper pair	2
read unmapped	4
mate unmapped	8
read reverse strand	16
mate reverse strand	32
first in pair	64
second in pair	128
not primary alignment	256
read fails platform/vendor quality checks	512
read is PCR or optical duplicate	1024

Most common flags:

0 (mapped, not paired, forward strand), 4 and 16.

Op	BAM	Description
M	0	alignment match (can be a sequence match or mismatch)
I	1	insertion to the reference
D	2	deletion from the reference
N	3	skipped region from the reference
S	4	soft clipping (clipped sequences present in SEQ)
H	5	hard clipping (clipped sequences NOT present in SEQ)
P	6	padding (silent deletion from padded reference)
=	7	sequence match
X	8	sequence mismatch



CIGAR string

For example:

RefPos:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Reference:	C	C	A	T	A	C	T	G	A	A	C	T	G	A	C	T	A	A	C
Read:	ACTAGAAATGGCT																		

Aligning these two:

RefPos:	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19
Reference:	C	C	A	T	A	C	T	G	A	A	C	T	G	A	C	T	A	A	C
Read:				A	C	T	A	G	A	A		T	G	G	C	T			

With the alignment above, you get:

POS: 5	CIGAR: 3M1I3M1D5M
--------	-------------------



Last, but very important, SAM field is the TAG field

Each TAG has a meaning and summarizes some aspect of the alignment.

Some tags (e.g. NM) have a predefined meaning in the format, NM is the number of mismatches between the read and the template

Other tags (e.g XT) are program specific – XT:A:U/R in BWA tells whether there is one or many “best alignments” for the read.

There are numerous predefined, or program specific tags that convey much useful information about each alignment, and alternative mappings for the reads. These tags are used when you filter alignments based on number of mismatches, or unique versus repeat, etc.



基因组学 | 测序 | 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



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+



BED

- Developed primarily for the UCSC genome browser
- Used to store annotations on genomic coordinates
 - Annotate gene/mRNA/exon/... position
 - Annotate Transcription Factor binding sites
 - Annotate SNP genotypes
 - Annotate Gene Expression
 - Annotate ...



Need for gene/transcript annotation

Two main annotation files are found:

GFF Generic Feature Format Version 3 (GFF3)

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

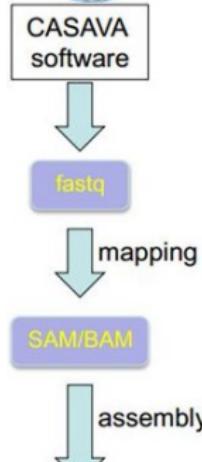
GTF Gene Transfer Format (GTF2)

- <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";



基因组学 | 测序 | 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"



基因组学 | 测序 | 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



基因组学 | 测序 | 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 ctgl23 1 1497228
ctgl23 Prokka gene    1000    9000   .    +    .    ID=gene00001;Name=EDEN
ctgl23 Prokka TF_binding_site 1000    1012   .    +    .    ID=tfbs00001;Parent=gene00001
ctgl23 Prokka mRNA     1050    9000   .    +    .    ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctgl23 Prokka mRNA     1050    9000   .    +    .    ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctgl23 Prokka mRNA     1300    9000   .    +    .    ID=mRNA00003;Parent=gene00001;Name=EDEN.3
ctgl23 Prokka exon    1300    1500   .    +    .    ID=exon00001;Parent=mRNA00003
ctgl23 Prokka exon    1050    1500   .    +    .    ID=exon00002;Parent=mRNA00001,mRNA00002
ctgl23 Prokka exon    3000    3902   .    +    .    ID=exon00003;Parent=mRNA00001,mRNA00003
ctgl23 Prokka exon    5000    5500   .    +    .    ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctgl23 Prokka exon    7000    9000   .    +    .    ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003
ctgl23 Prokka CDS    1201    1500   .    +    0    ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctgl23 Prokka CDS    3000    3902   .    +    0    ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctgl23 Prokka CDS    5000    5500   .    +    0    ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctgl23 Prokka CDS    7000    7600   .    +    0    ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctgl23 Prokka CDS    1201    1500   .    +    0    ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctgl23 Prokka CDS    5000    5500   .    +    0    ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctgl23 Prokka CDS    7000    7600   .    +    0    ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctgl23 Prokka CDS    3301    3902   .    +    0    ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctgl23 Prokka CDS    5000    5500   .    +    1    ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctgl23 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

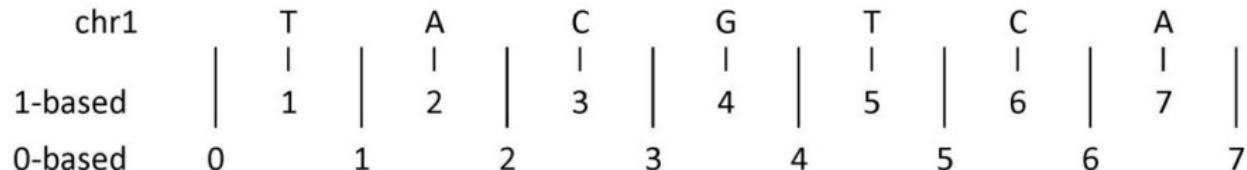
GTF/GFF: one based, inclusive

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



基因组学 | 测序 | 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



VCF format

- The Variant Call Format (VCF) is the emerging standard for storing variant data.
 - Originally designed for SNPs and short INDELS, it also works for structural variations.
-
- VCF consists of a header section and a data section.
 - The **header** must contain a line starting with one '#', showing the name of each field, and then the sample names starting at the 10th column.
 - The **data** section is TAB delimited with each line consisting of at least 8 mandatory fields
The FORMAT field and sample information are allowed to be absent.



基因组学 | 测序 | 数据格式 | VCF

Col	Field	Description
1	CHROM	Chromosome name
2	POS	1-based position. For an indel, this is the position preceding the indel.
3	ID	Variant identifier. Usually the dbSNP rsID.
4	REF	Reference sequence at POS involved in the variant. For a SNP, it is a single base.
5	ALT	Comma delimited list of alternative sequence(s).
6	QUAL	Phred-scaled probability of all samples being homozygous reference.
7	FILTER	Semicolon delimited list of filters that the variant fails to pass.
8	INFO	Semicolon delimited list of variant information.
9	FORMAT	Colon delimited list of the format of individual genotypes in the following fields.
10+	Sample(s)	Individual genotype information defined by FORMAT.



Example .VCF file

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myProgramV3
##reference=file:///seq/NCBI36.fasta
```

Header lines
(marked by ##):
Metadata of analysis

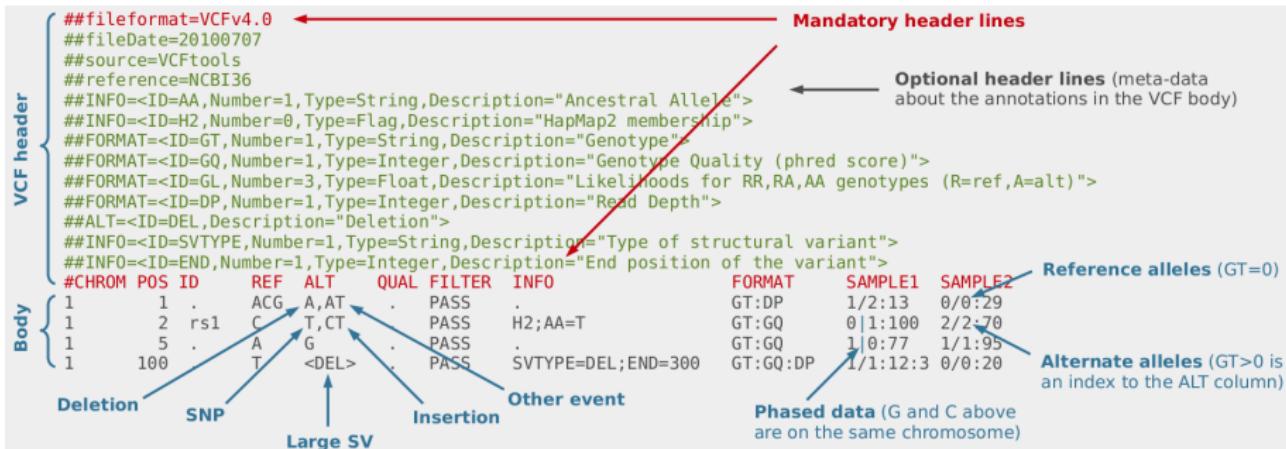
```
...
#CHROM POS ID REF ALT QUAL FILTER
20 14370 rs6054257 G A 29 PASS
20 17330 . T A 3 q10
```

INFO	FORMAT	SAMPLE1	...
NS=2;DP=14;AF=0.5;DB;H2	GT:GQ:DP	1 0:48:8	
NS=2;DP=11;AF=0.017	GT:GQ:DP	0 0:49:3	

Data lines:
Individual variant calls



基因组学 | 测序 | 数据格式 | VCF



基因组学 | 测序 | 数据格式 | VCF

```
##fileformat=VCFv4.1
##fileDate=20090805
##source=myImputationProgramV3.1
##reference=file:///seq/references/1000GenomesPilot-NCBI36.fasta
##contig=<ID=20,length=62435964,assembly=B36,md5=f126cdf8a6e0c7f379d618ff66beb2da,species="Homo sapiens",taxonomy=x>
##phasing=partial
##INFO=<ID=NS,Number=1,Type=Integer,Description="Number of Samples With Data">
##INFO=<ID=DP,Number=1,Type=Integer,Description="Total Depth">
##INFO=<ID=AF,Number=A,Type=Float,Description="Allele Frequency">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=DB,Number=0,Type=Flag,Description="dbSNP membership, build 129">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FILTER=<ID=q10,Description="Quality below 10">
##FILTER=<ID=p50,Description="Less than 50% of samples have data">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##FORMAT=<ID=HQ,Number=2,Type=Integer,Description="Haplotype Quality">
#CHROM POS ID REF ALT QUAL FILTER INFO FORMAT NA00001 NA00002 NA00003
20 14370 rs6054257 G A 29 PASS NS=3;DP=14;AF=0.5;DB;H2 GT:GQ:DP:HQ 0|0:48:1:51,51 1|0:48:8:51,51 1/1:43:5:...
20 17330 . T A 3 q10 NS=3;DP=11;AF=0.017 GT:GQ:DP:HQ 0|0:49:3:58,50 0|1:3:5:65,3 0/0:41:3
20 1110696 rs6040355 A G,T 67 PASS NS=2;DP=10;AF=0.333,0.667;AA=T;DB GT:GQ:DP:HQ 1|2:21:6:23,27 2|1:2:0:18,2 2/2:35:4
20 1230237 . T . 47 PASS NS=3;DP=13;AA=T GT:GQ:DP:HQ 0|0:54:7:56,60 0|0:48:4:51,51 0/0:61:2
20 1234567 microsat1 GTC G,GTCT 50 PASS NS=3;DP=9;AA=G GT:GQ:DP 0|1:35:4 0|2:17:2 1/1:40:3
```



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

● 常见术语

- 分析流程
- 补遗
- 预处理
- 比对后
- 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



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 1 个 read 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 1 个 read 覆盖。

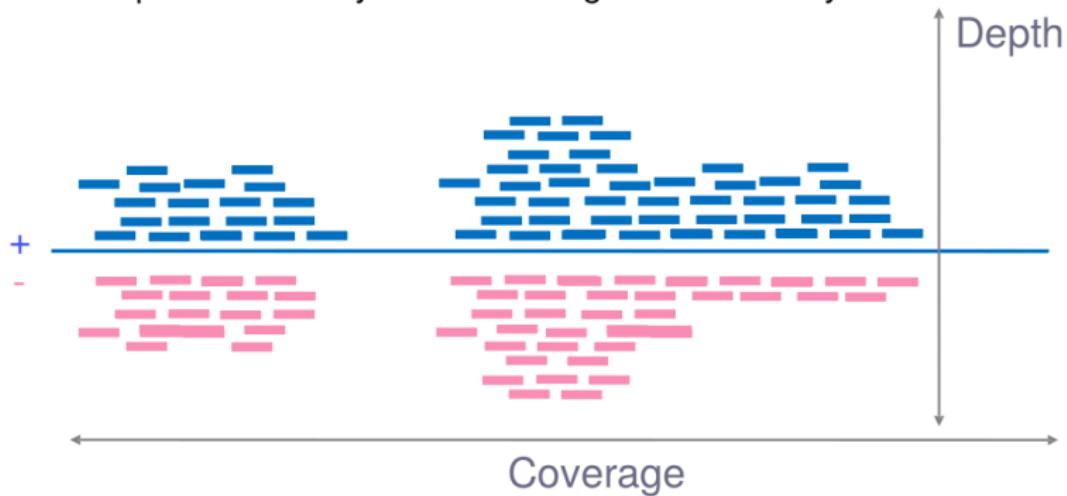
深度与覆盖度

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



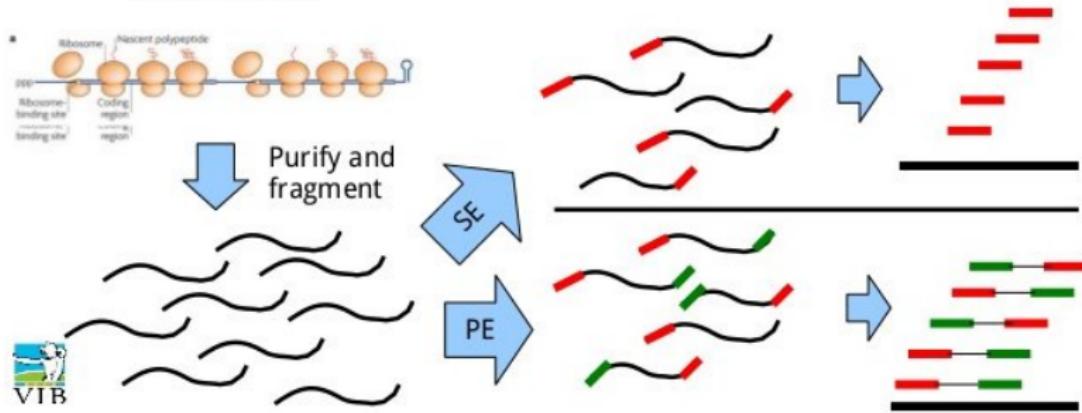
Statistics used as Quality Control

- **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



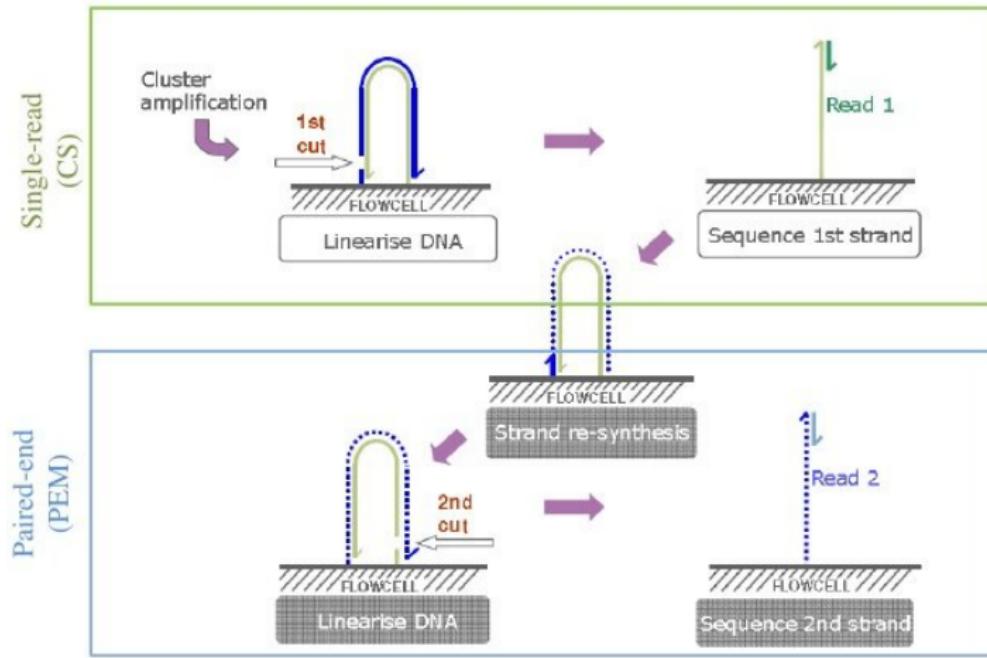
PE versus SE Illumina

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



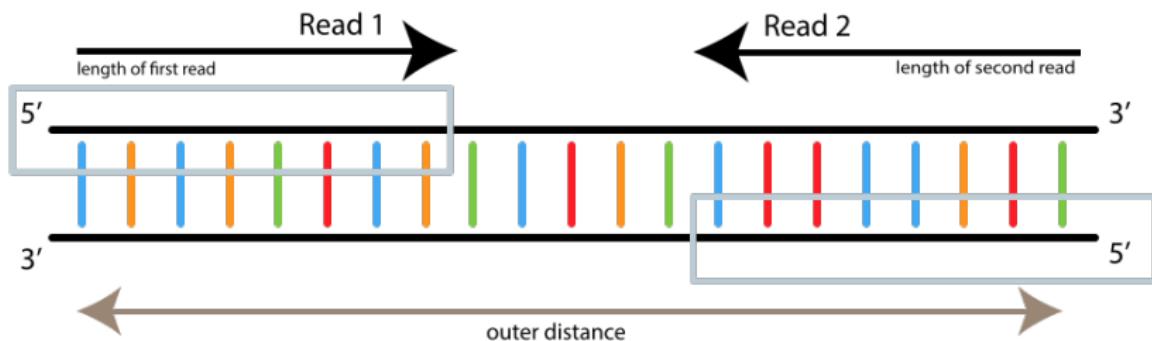
Paired-end strategy

Paired-end sequencing works into GA and uses chemicals from PE module to perform cluster amplification of the reverse strand



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



Paired-End DNA Sequencing

- provide superior alignment across DNA regions containing repetitive sequences
- produce longer contigs for *de novo* sequencing by filling gaps in the consensus sequence
- detect rearrangements such as insertions, deletions, and inversions

Paired-End RNA Sequencing

- enable discovery applications such as detecting gene fusions in cancer and characterizing novel splice isoforms.



Paired-End DNA Sequencing

- provide superior alignment across DNA regions containing repetitive sequences
- produce longer contigs for *de novo* sequencing by filling gaps in the consensus sequence
- detect rearrangements such as insertions, deletions, and inversions

Paired-End RNA Sequencing

- enable discovery applications such as detecting gene fusions in cancer and characterizing novel splice isoforms.



PE versus SE Illumina

Single end (SE):

- **Gene level** differential expression

Paired end (PE):

- Novel splice junction detection
- *De novo* assembly of transcriptome
- Helps with correctly positioning reads on the reference genome sequence.

Note: PE not the same as mate pairs.



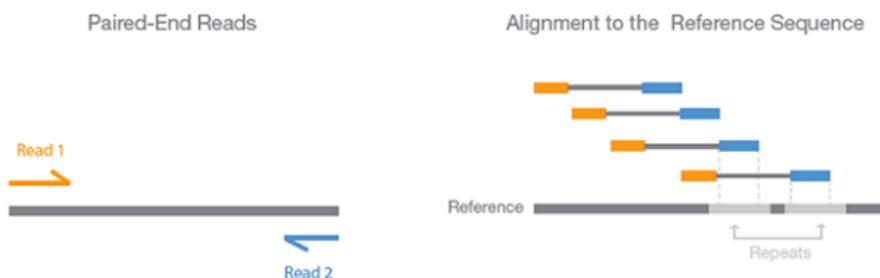
Huge impact on read mapping

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

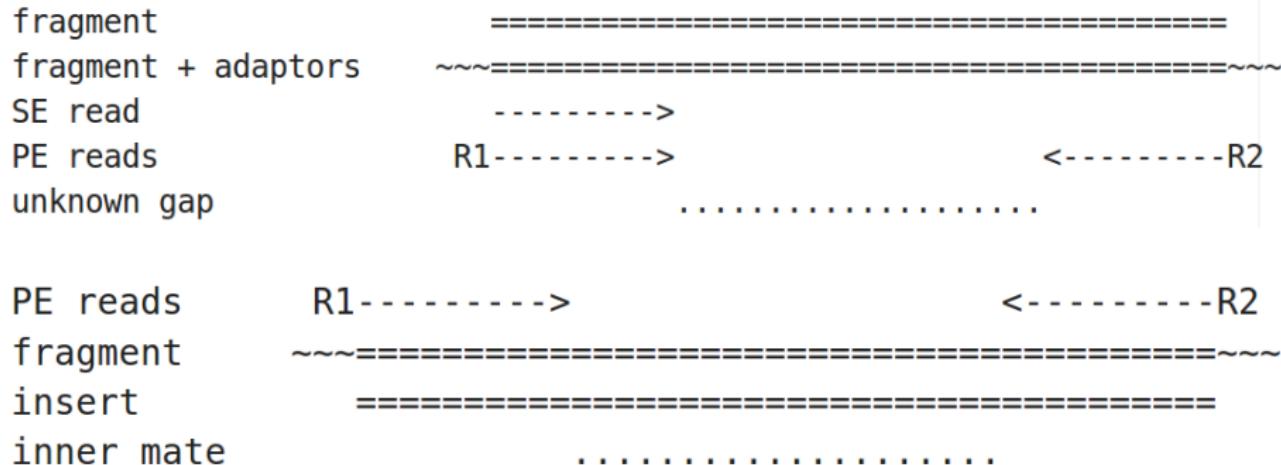


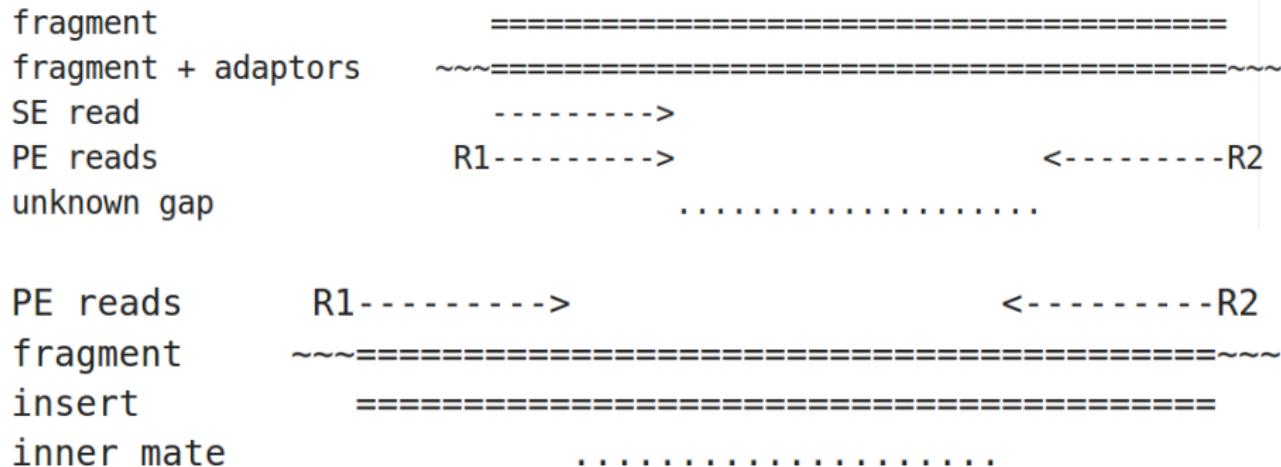
Paired-end sequencing enables both ends of the DNA fragment to be sequenced. Because the distance between each paired read is known, alignment algorithms can use this information to map the reads over repetitive regions more precisely. This results in much better alignment of the reads, especially across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

The diagram illustrates five types of sequencing reads:

- fragment**: Represented by a solid horizontal line.
- fragment + adaptors**: Represented by a dashed horizontal line with a tilde (~) at the start.
- SE read**: Represented by a dashed line ending in a right-pointing arrow.
- PE reads**: Represented by two dashed lines with arrows pointing in opposite directions, labeled R1 and R2.
- unknown gap**: Represented by a dotted horizontal line.



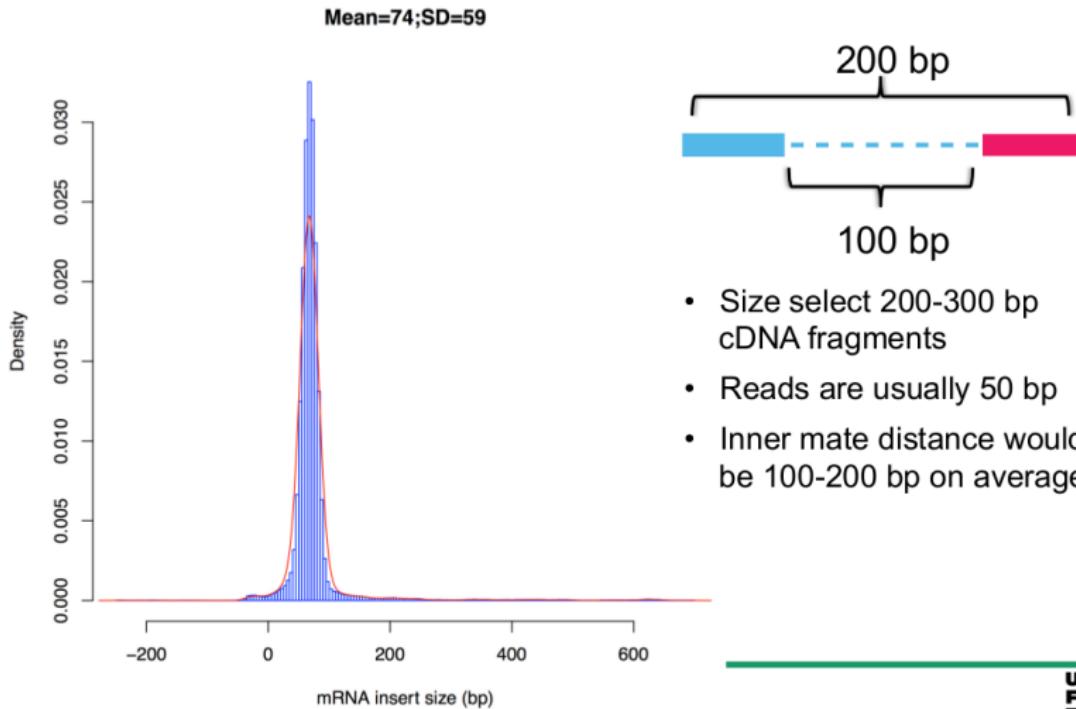




Conclusion

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 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- **分析流程**
- 补遗
 - 预处理
 - 比对后
 - 实验设计

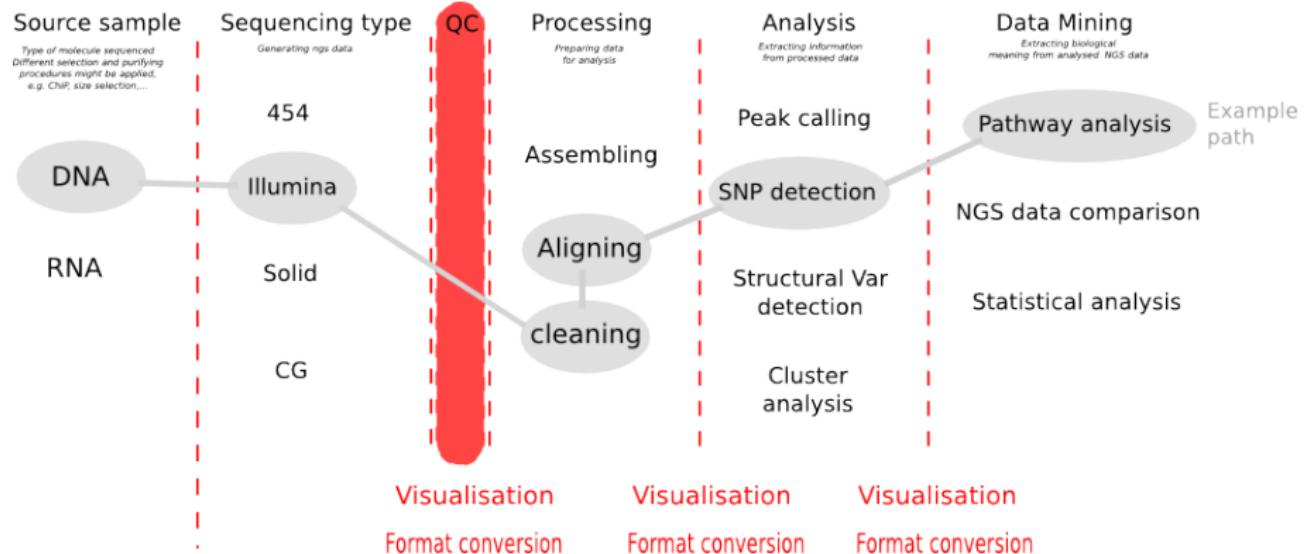
5 外显子组测序

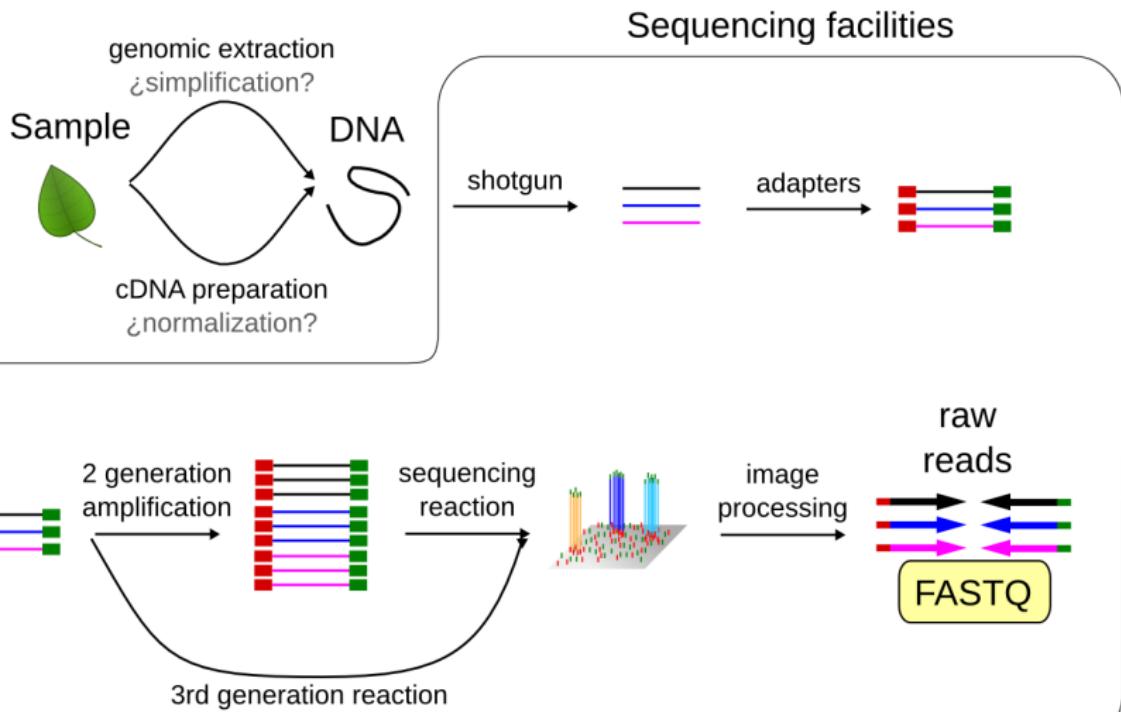
- 简介
- 操作流程
- 应用实例

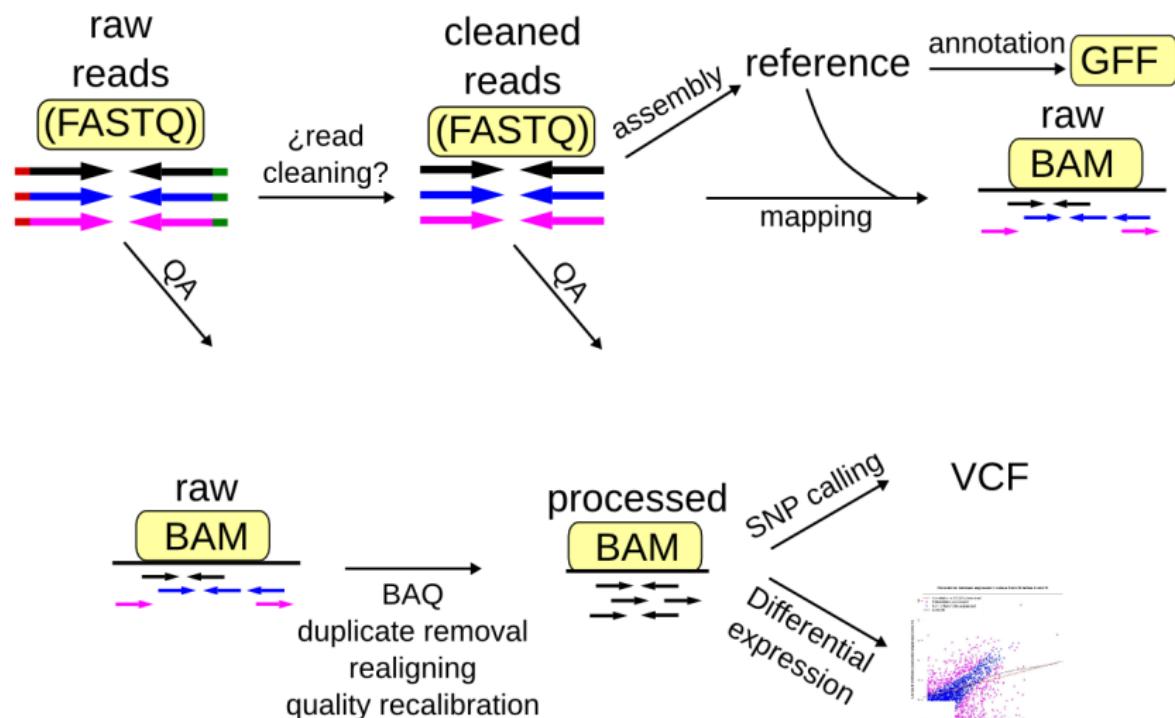
6 回顾与总结

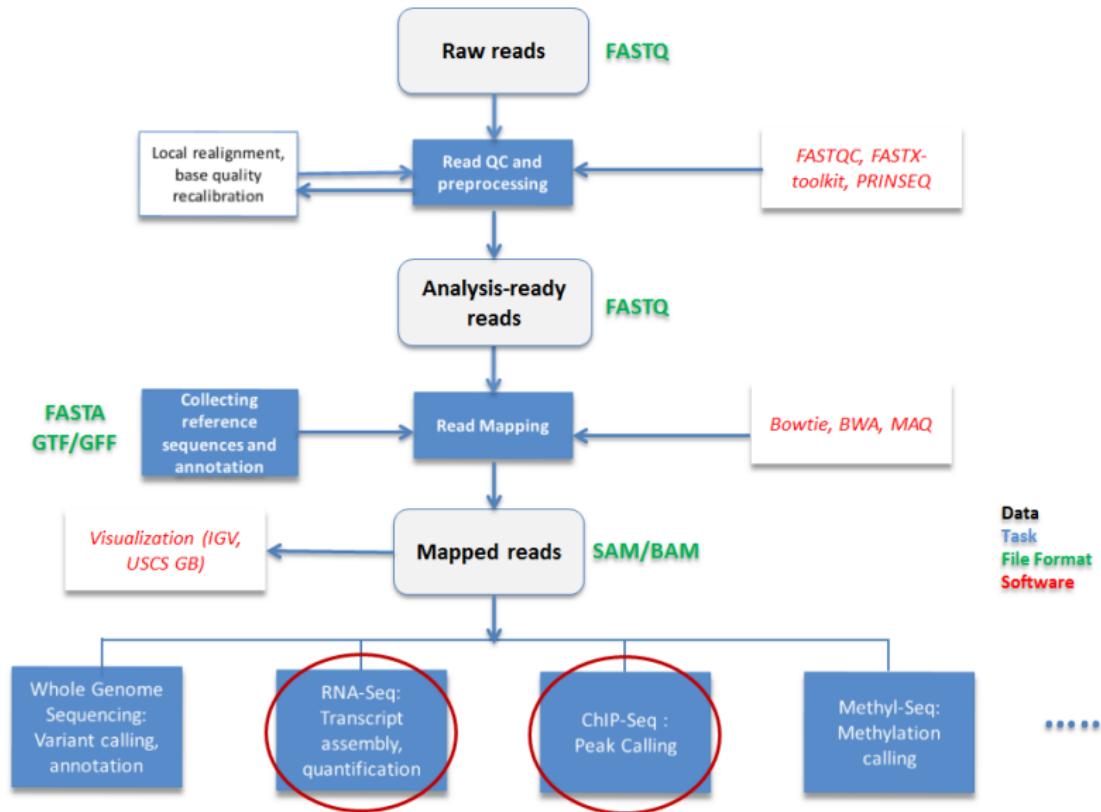
- 总结
- 思考题

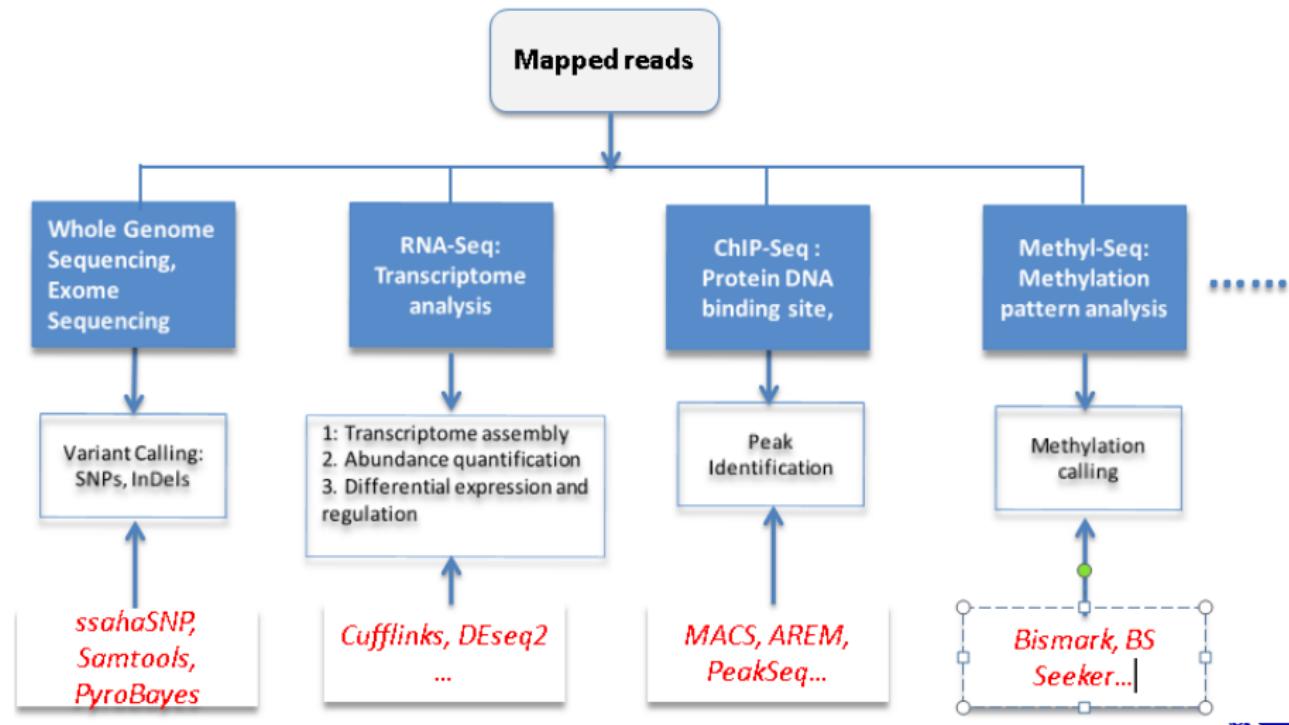


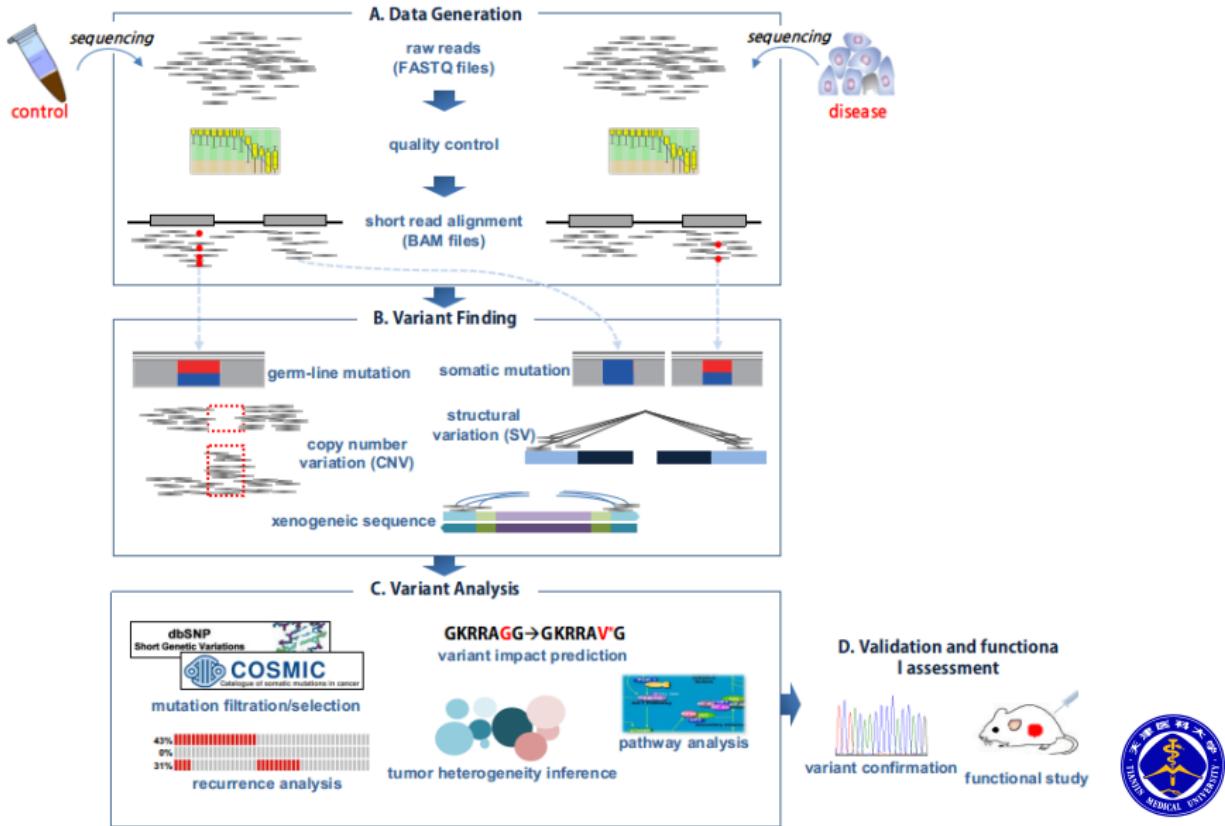


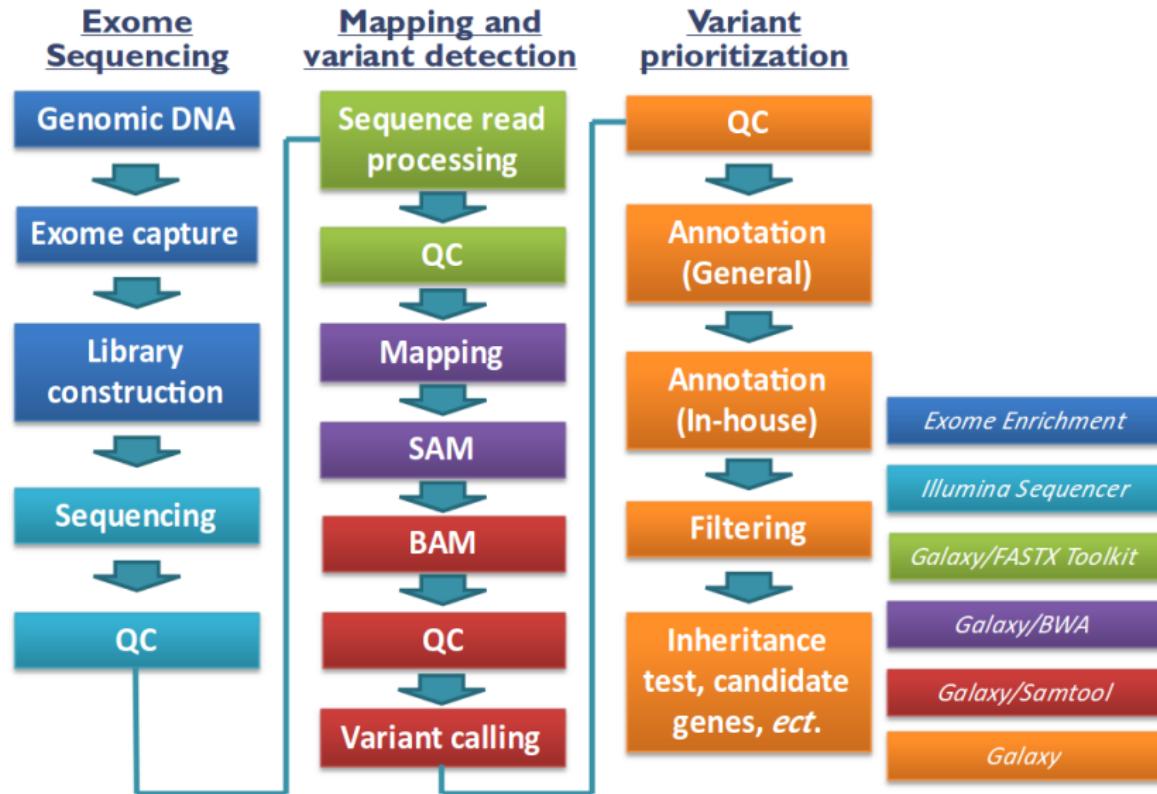




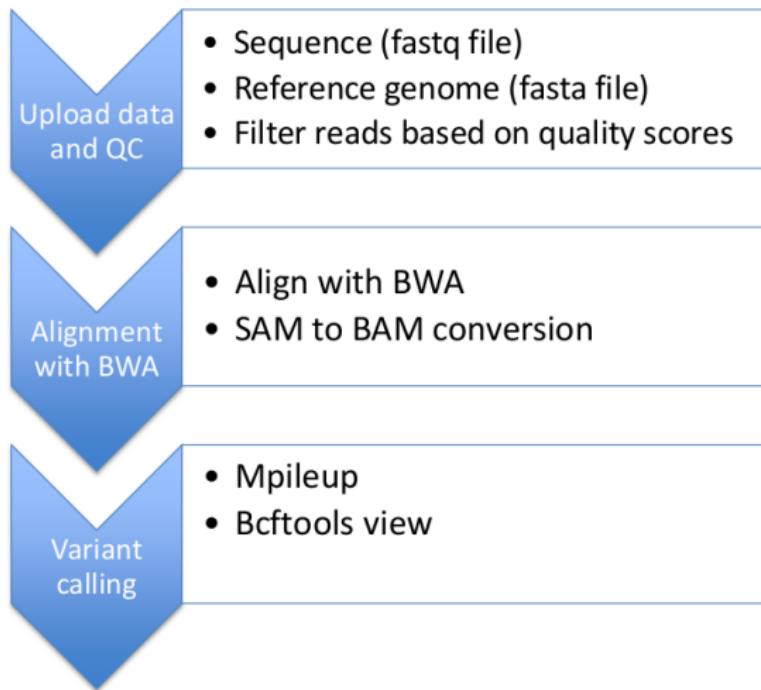






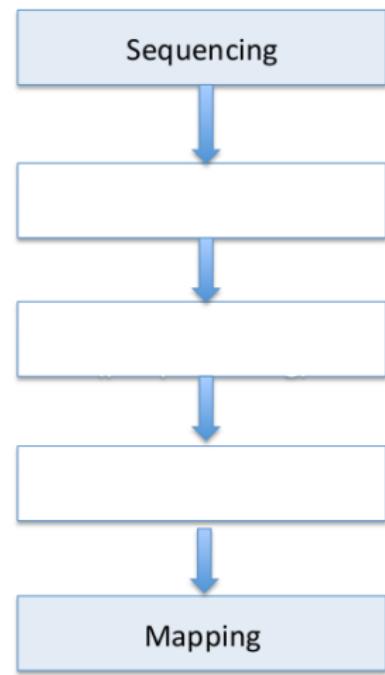


SNP-Seq pipeline



Quality control

- It is important to check the quality of your sequenced reads!
- FASTQC: free program that reports quality profile of reads
- Pre-processing
 - Trim reads
 - exclude low quality reads
 - contaminations



Data Quality Control

- Data Quality Assessment
 - Identify poor/bad sample
 - Identify contaminates
 - Trimming: remove bad bases from read
 - Filtering: remove bad reads from library

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



Feature\Tools	NGS QC Toolkit v2.2	FastQC v0.10.0	PRINSEQ-lite v0.17 ¹	TagDust	FASTX-Toolkit v0.0.13	SolexaQA v1.10	TagCleaner v0.12 ¹	CANGS v1.1
Supported NGS platforms	Illumina, 454	FASTQ ²	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 ³	No	Yes	Yes	No	Yes ⁴	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 ⁵	Yes	No	Yes	No	No	Yes
Low complexity filtering	No	No	Yes	No	Yes	No	No	No
N/X content filtering	No	No ⁶	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 ⁷	No	Yes	Yes	No ⁷	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.

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.

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.

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

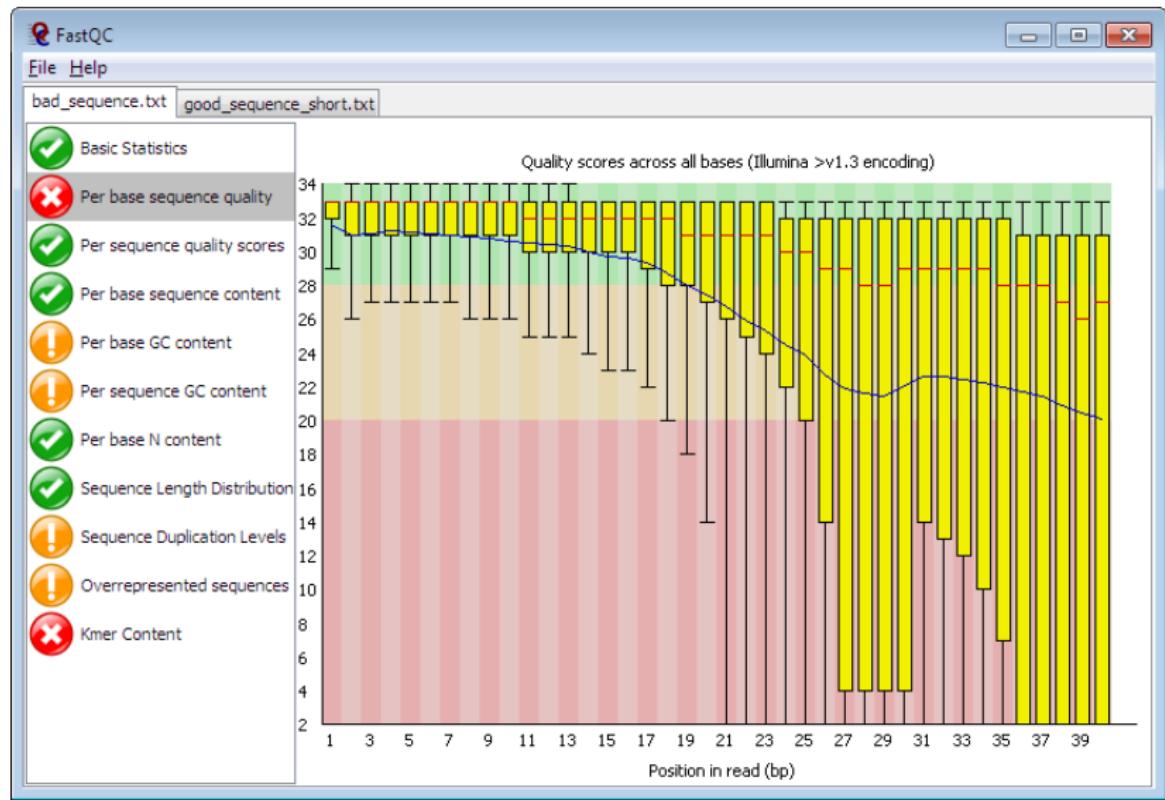
Measure	Value
Filename	sample.fastq
File type	Conventional base calls
Encoding	Illumina 1.5
Total Sequences	9053
Sequences flagged as poor quality	0
Sequence length	36
%GC	50



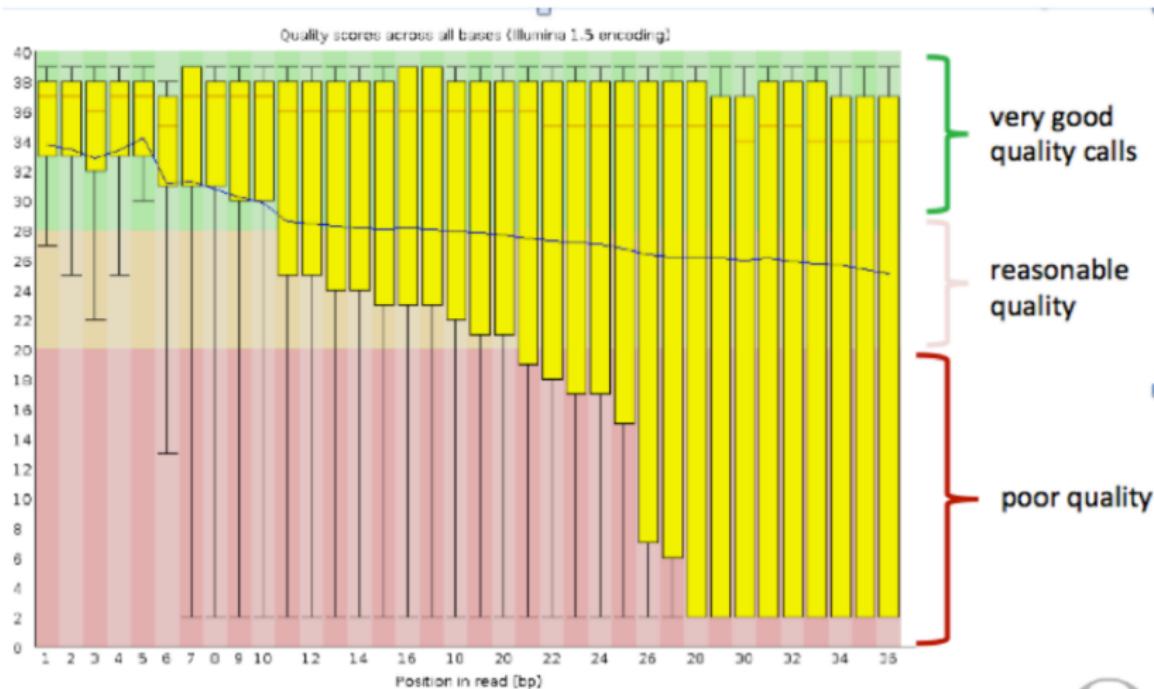
FASTQC: Quality control on raw data

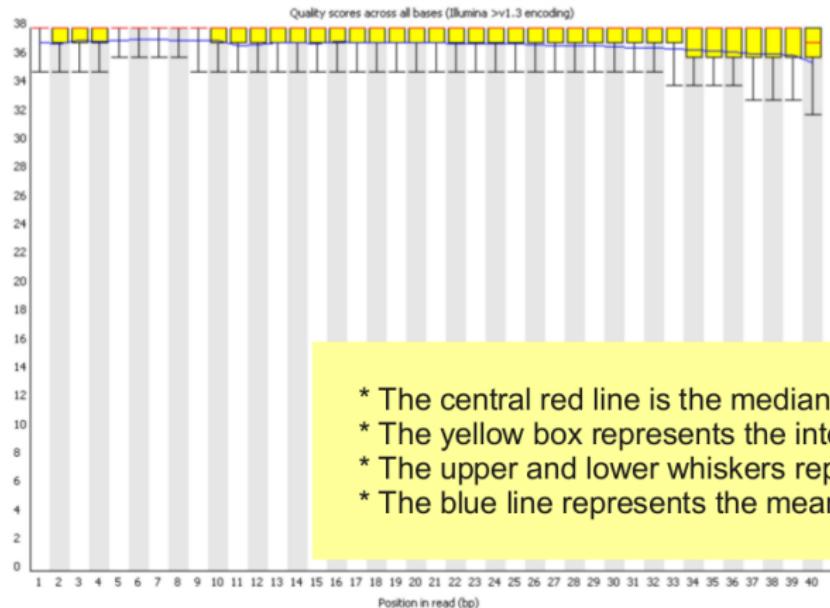
- **Sequence Length Distribution**: sequencers produce same (Illumina) or different (PGM) read length. This metric helps identify abnormal read length
- **Sequence content**: % A/C/G/T ; %GC; %N at each position in the read
 - Proportion biased for targeted sequencing
- **Quality score**:
 1. **Per base**: identify base calls with low quality (commonly towards the end of a read)
 2. **Per sequence**: to see if a subset of your sequences have universally low quality values
- **K-mers content**: a k-mer is a motif of length k observed more than once in a sequence (repeats : ACACAC ; spaced occurrences : tccGAGGaaggGAGGaa)
- **Over-represented sequences**: highly duplicated sequence in your library (primer, adapter..)





(2) FASTQC: Per base sequence quality

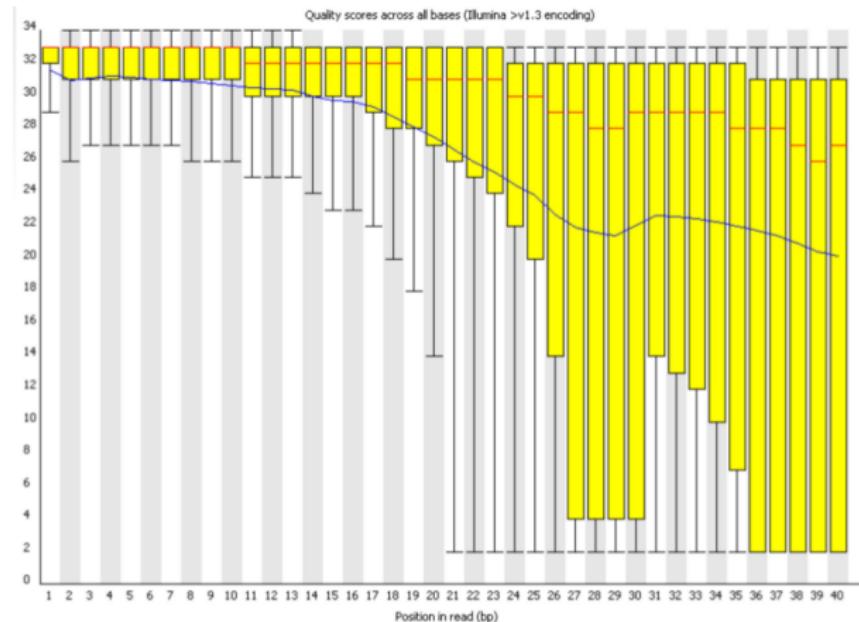




Good data

- Consistent
- High quality along the read



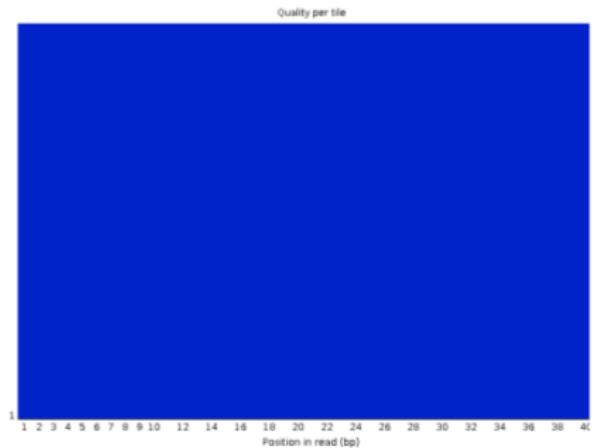


Bad data

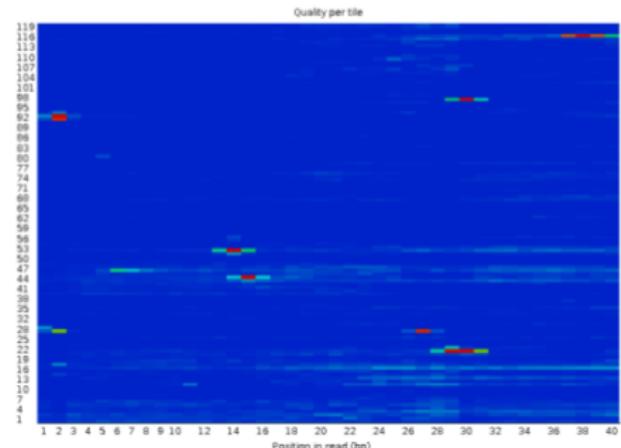
- High variance
- Quality decrease with length

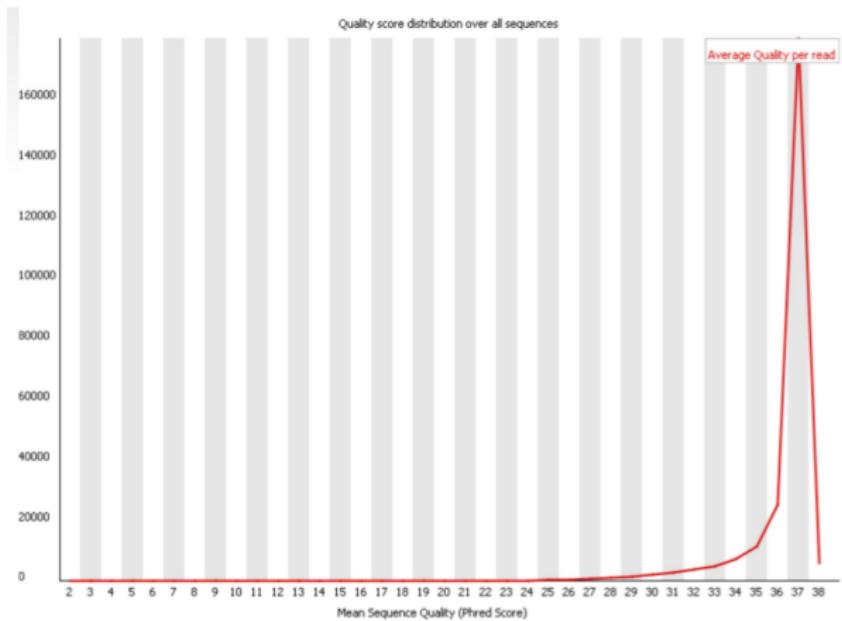


Per tile sequence quality



Per tile sequence quality

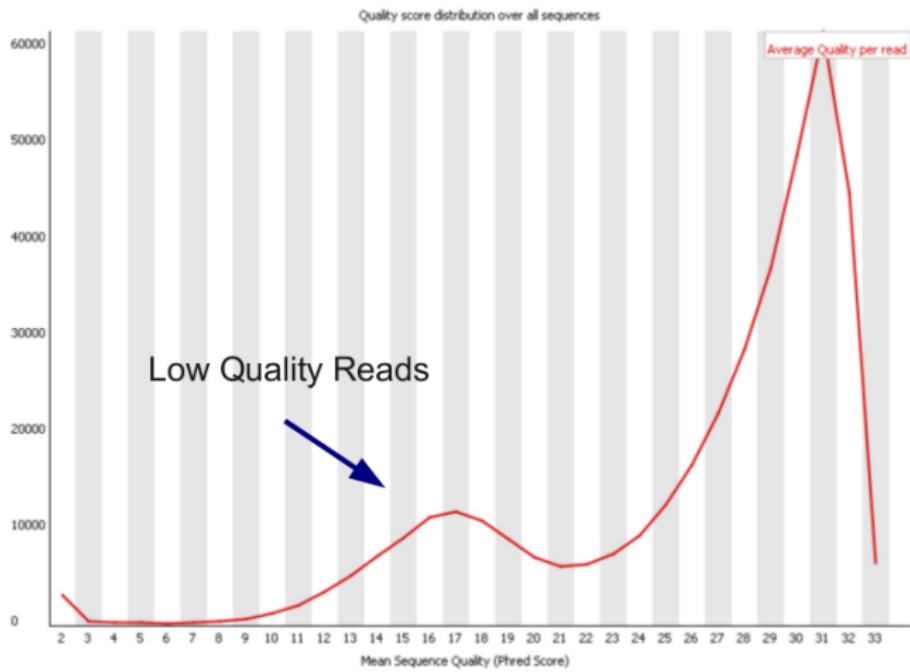




Good data

- Most are high-quality sequences

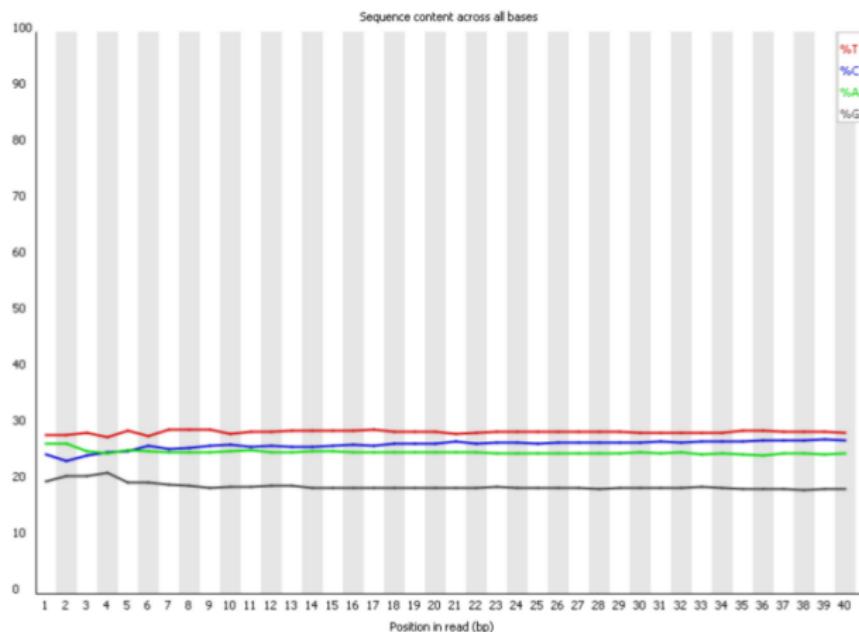




Bad data

- Not uniform distribution

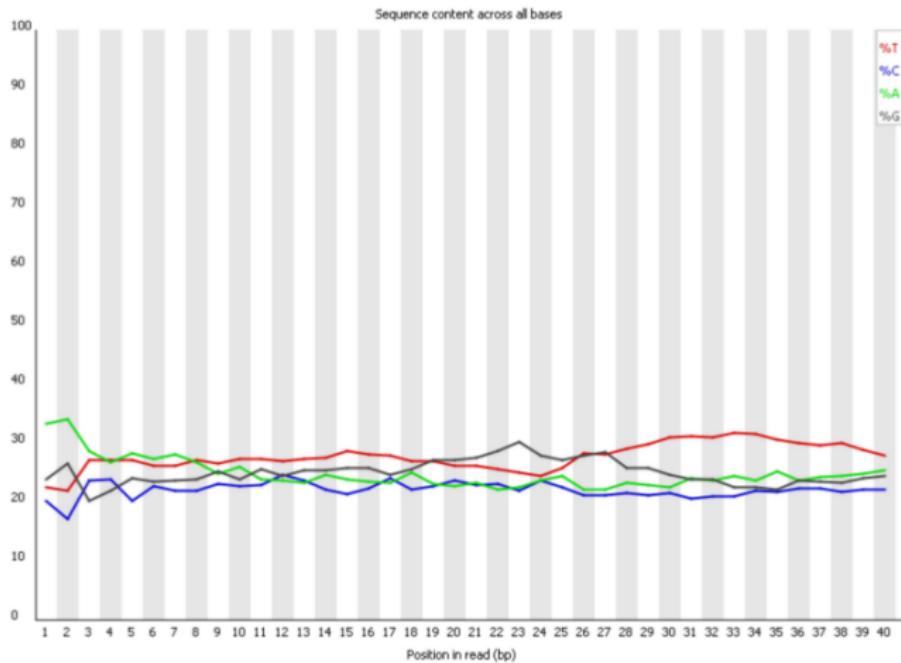




Good data

- Smooth over length
- Organism dependent (GC)



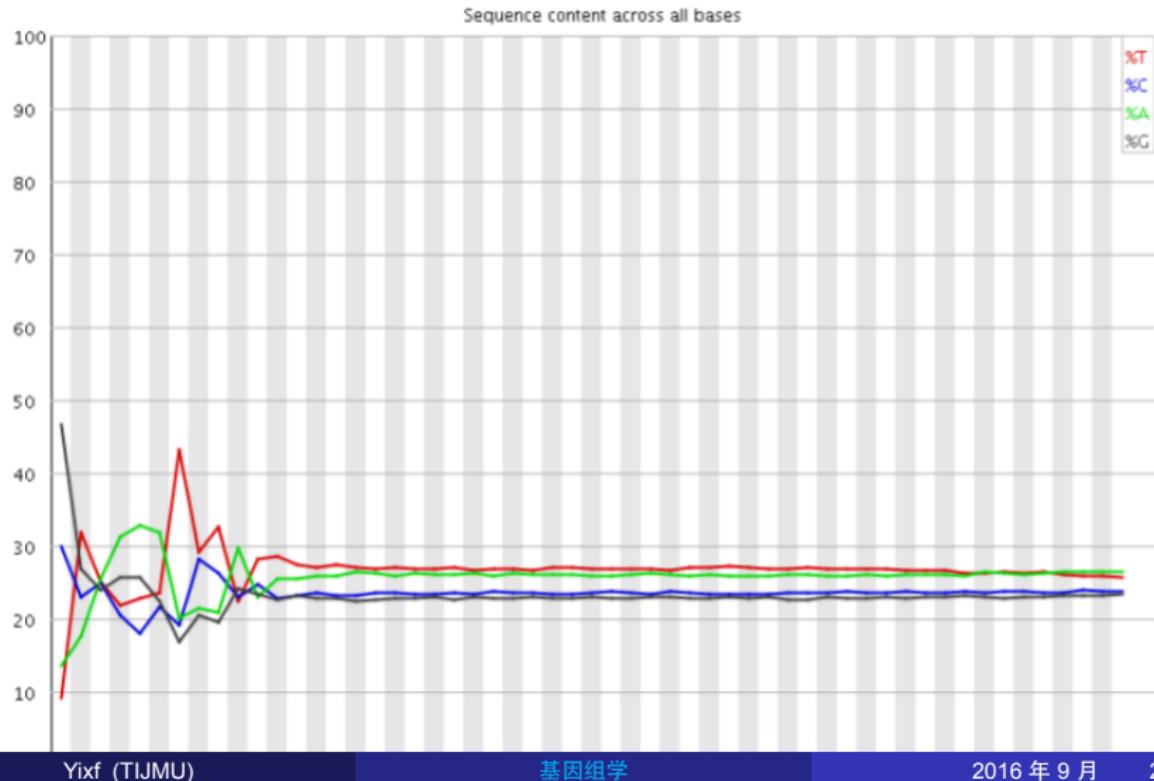


Bad data

- Sequence position bias

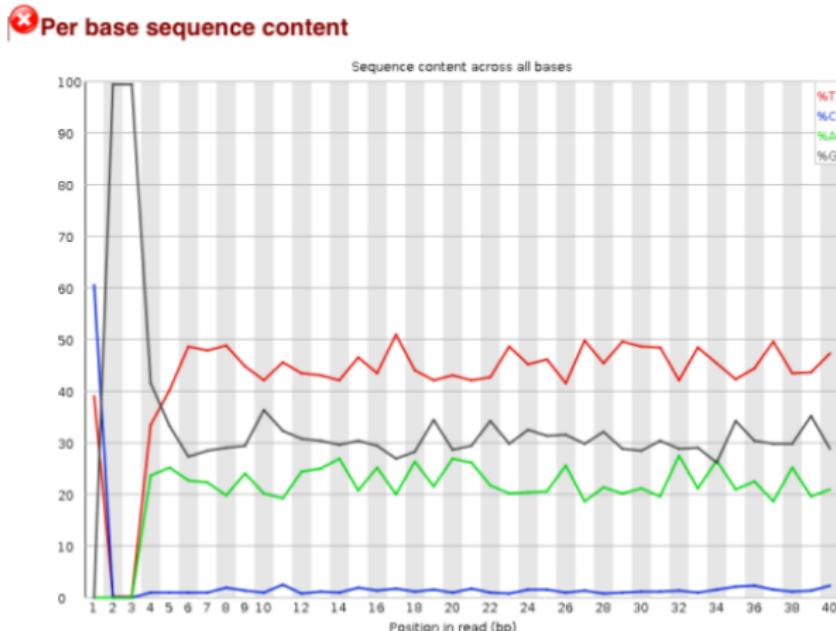


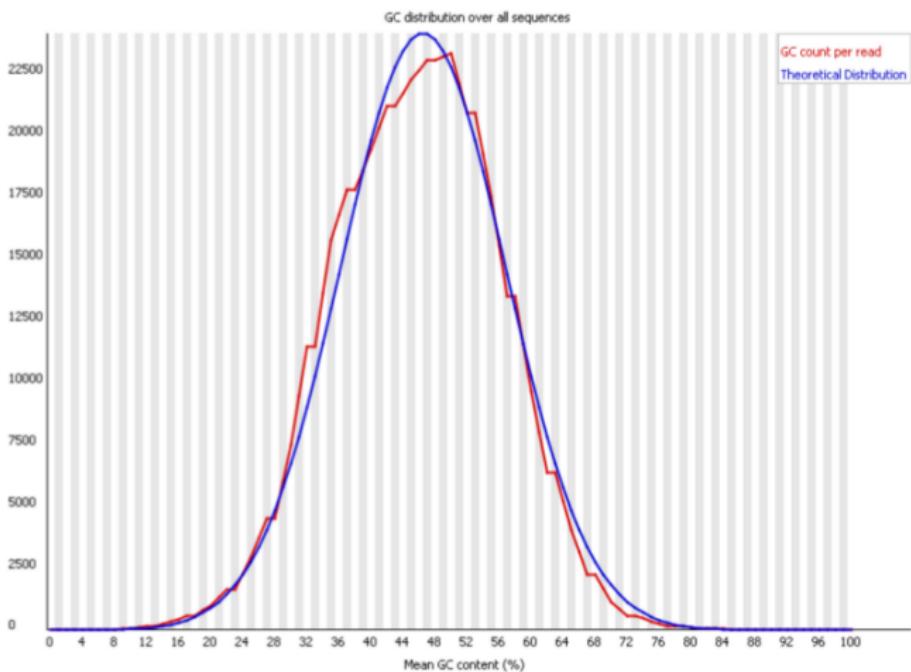
Unavoidable – RNA-Seq



Unavoidable – RRBS

Devoided of cytosines because the library was treated with sodium bisulphite
(which will have converted most of the C to T)

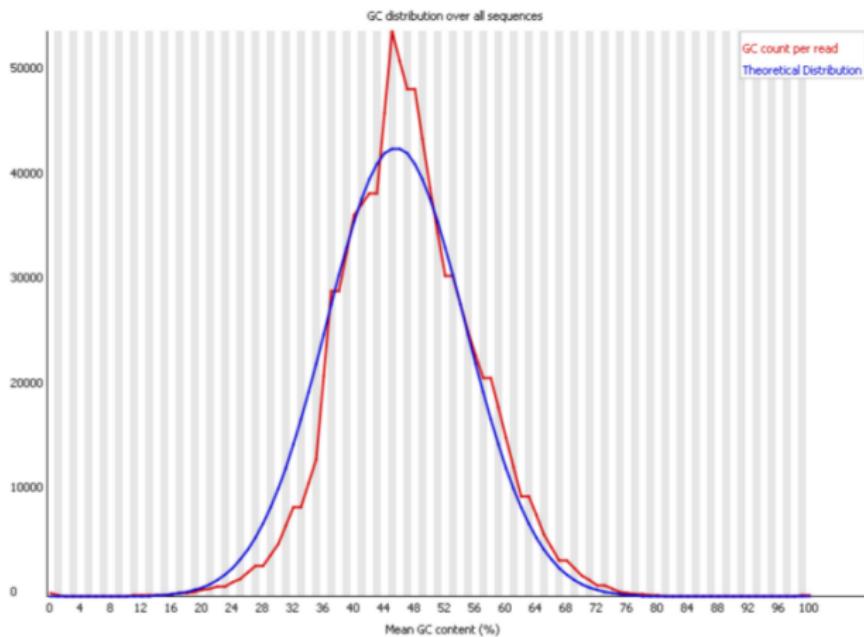




Good data

- Fits with the expected
- Organism dependent

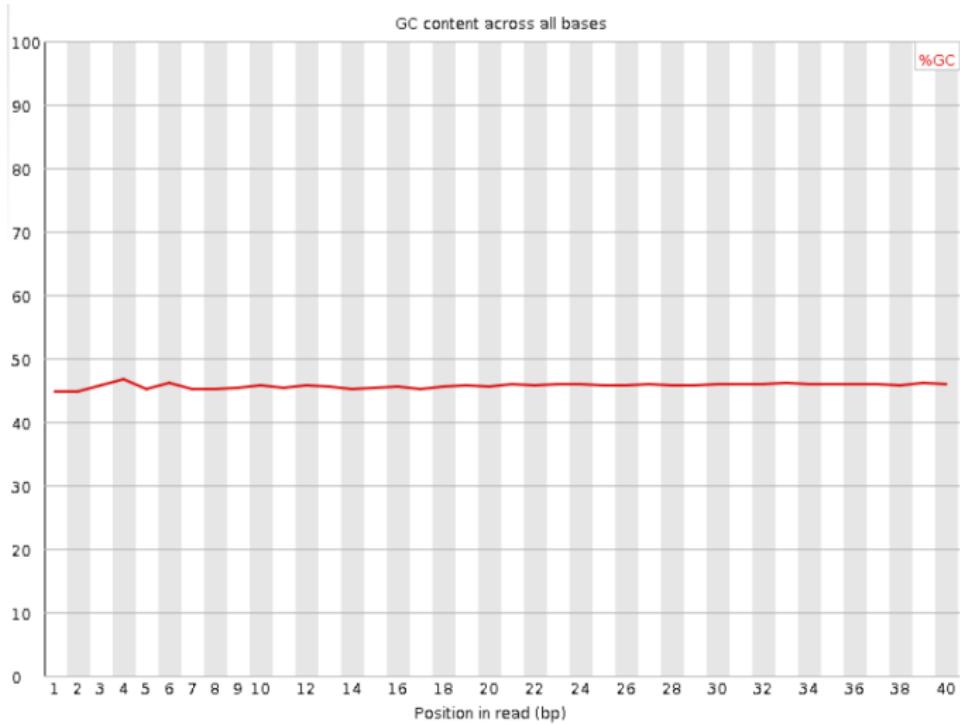




Bad data

- It does not fit with expected
 - Organism dependent
- Library contamination?

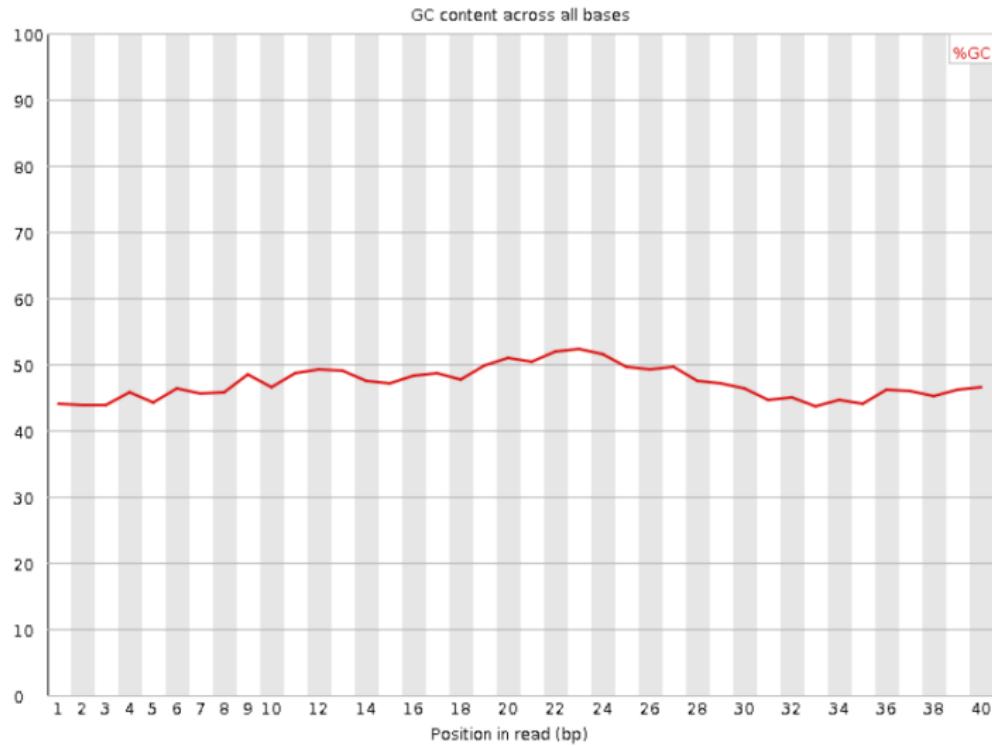




Good data

- No variation across read sequence

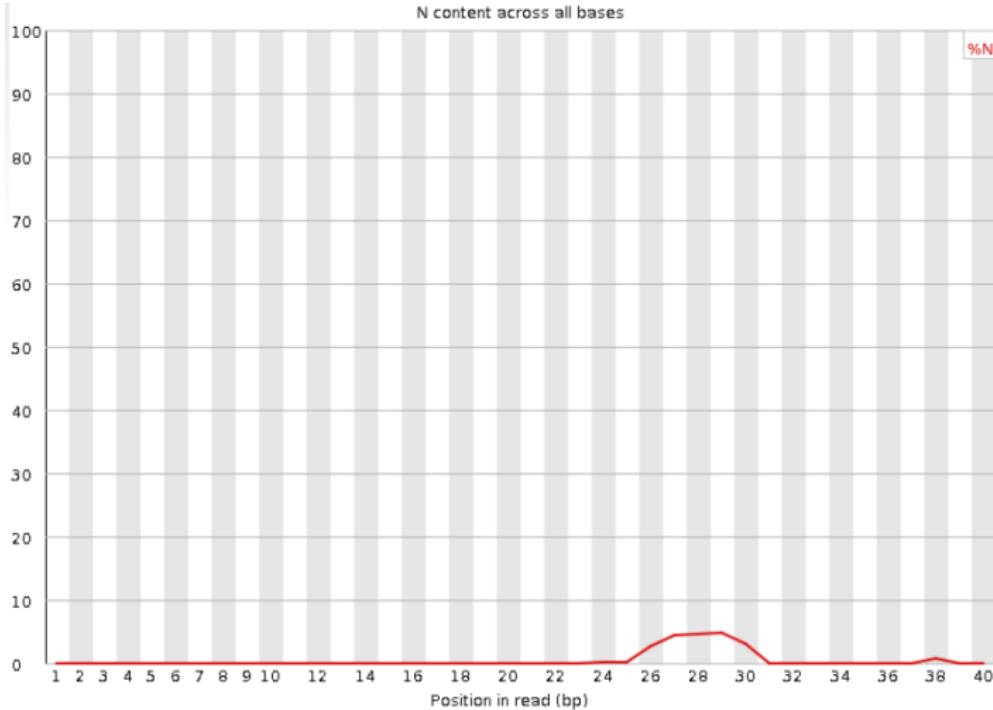




Bad data

- Variation across read sequence



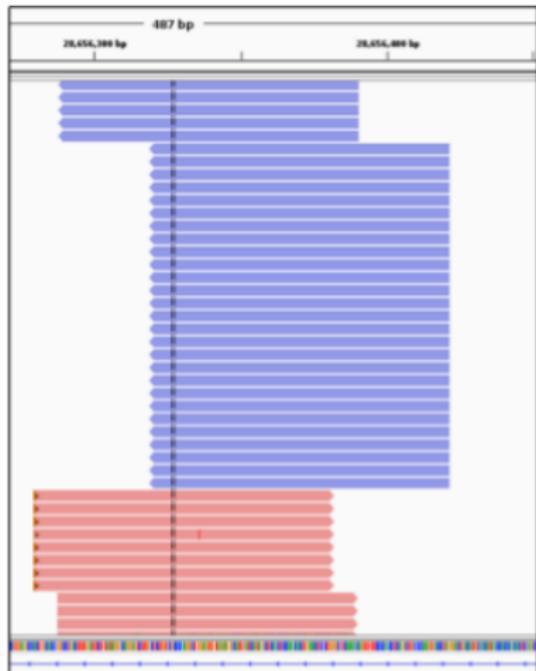


It's not good if there are N bias per base position



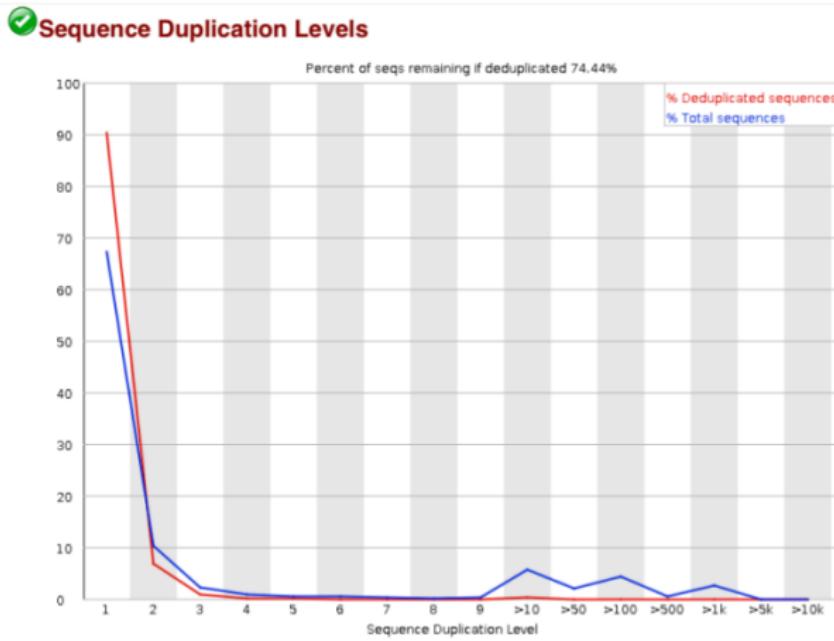
(9) FASTQC: Sequence duplication levels

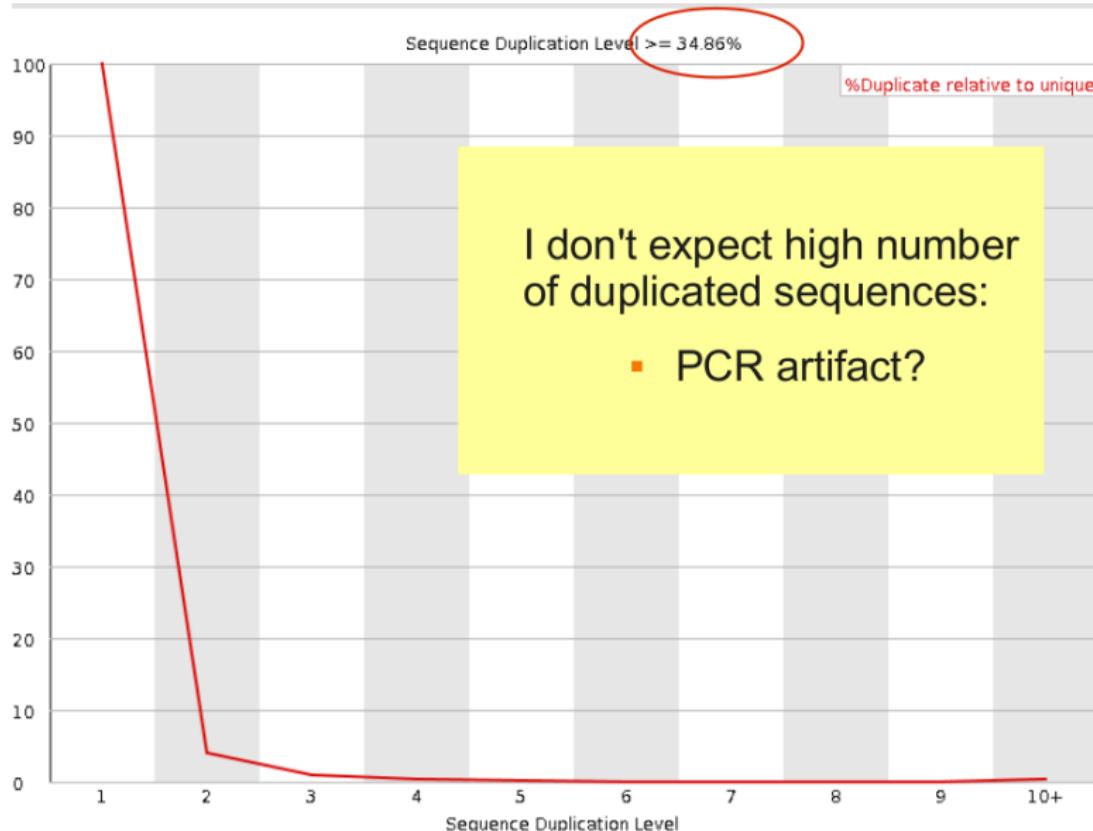
- PCR duplicates during sample preparation
- Optical duplicates: read the same cluster twice in the sequencer
- High duplication can lead to problems in downstream analysis (e.g. skew allele frequencies)



(9) FASTQC: Sequence duplication levels

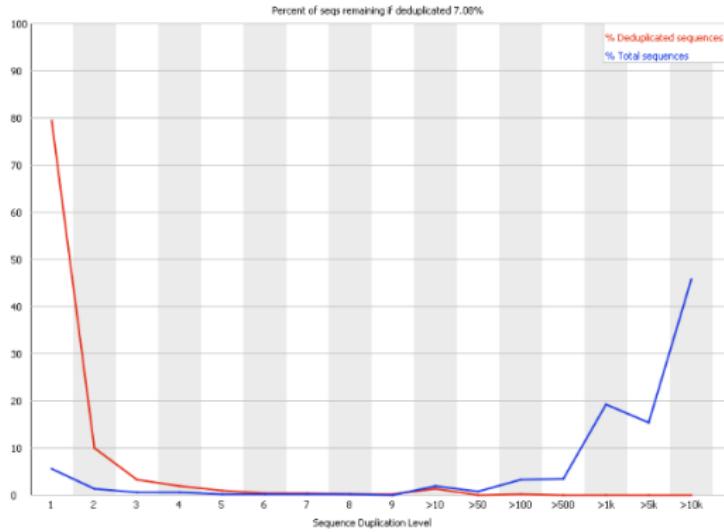
Very diverse library





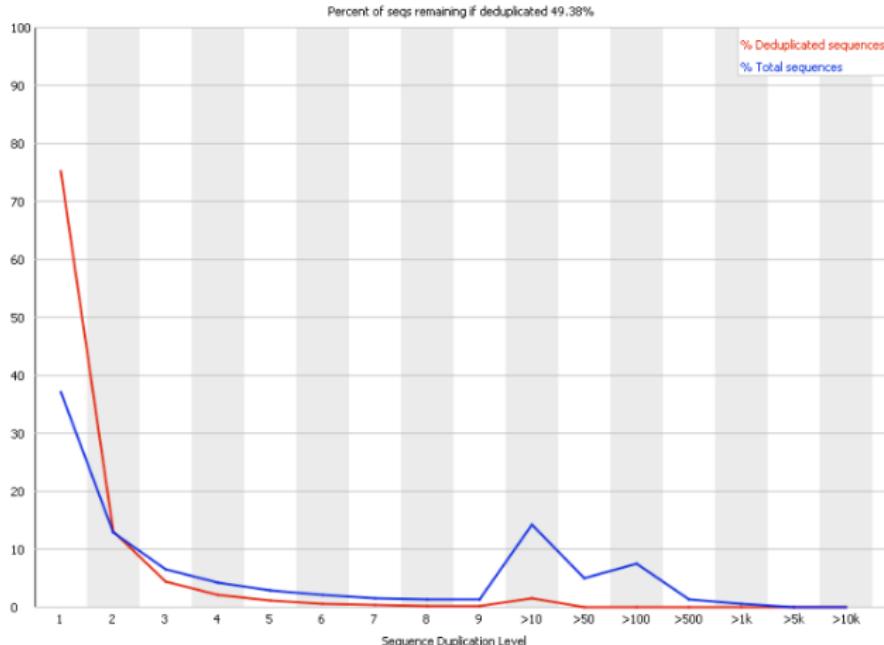
(9) FASTQC: Sequence duplication levels

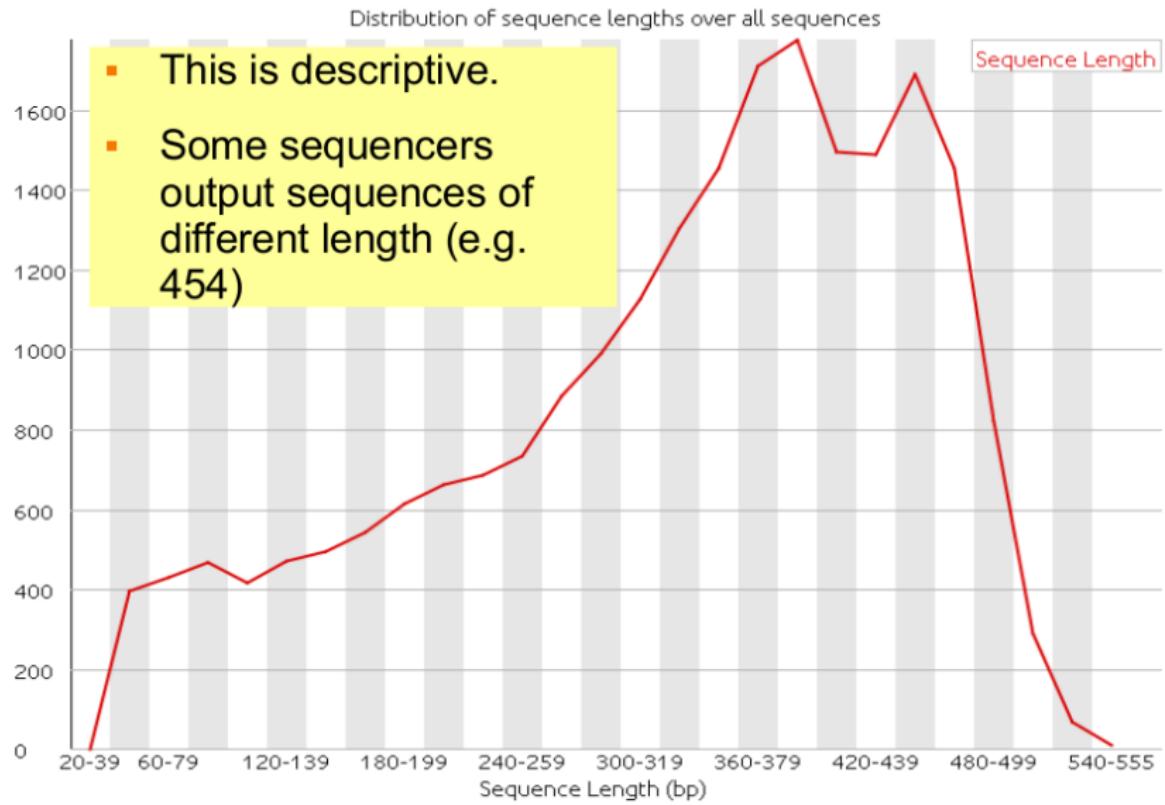
PCR duplication



(9) FASTQC: Sequence duplication levels

A good RNA-Seq library (although dup levels > 50%)





(10) FASTQC: Over-represented sequences

Good dataset



Overrepresented sequences

No overrepresented sequences

Bad datasets:



Overrepresented sequences

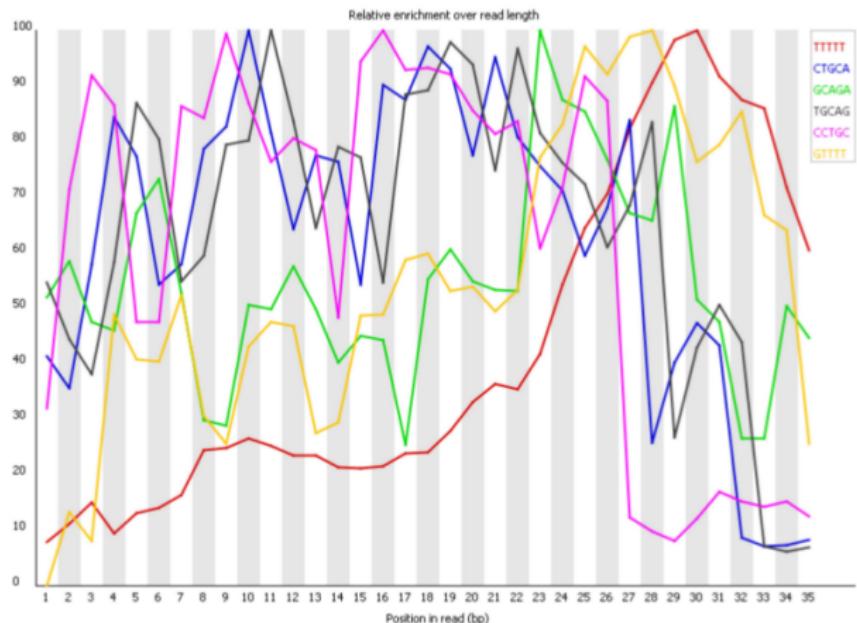
[Back to summary](#)



Overrepresented sequences

Sequence	Count	Percentage	Possible Source
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACA	28971	28.971000000000004	TruSeq Adapter, Index 5 (100% over 36bp)
GCTAACAAAATCCGACTAAATCAGTCAGTAAATA	392	0.392	No Hit
GTTAGCTATTACTTGACTGATTTAGTCGGGTATT	356	0.356	No Hit
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACC	108	0.108	TruSeq Adapter, Index 1 (97% over 36bp)
GATCGGAAGAGCACACGTCTGAACTCCAGTCACACG	107	0.107	TruSeq Adapter, Index 15 (97% over 36bp)

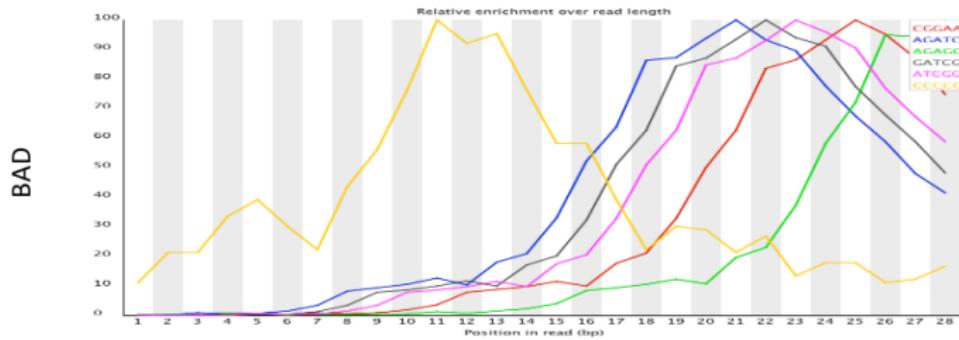
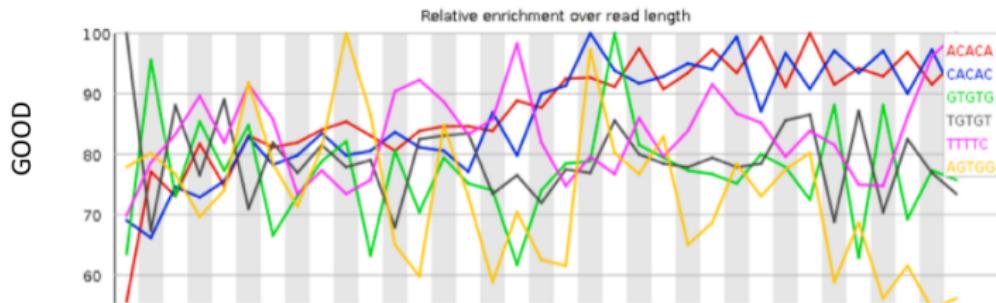




- Helps to detect problems
- Adapters?

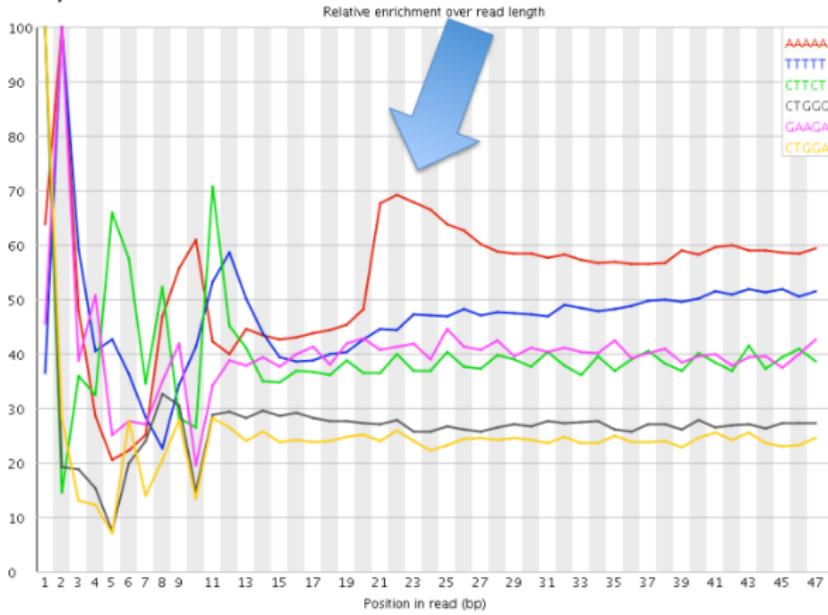


(11) FASTQC: Kmer content



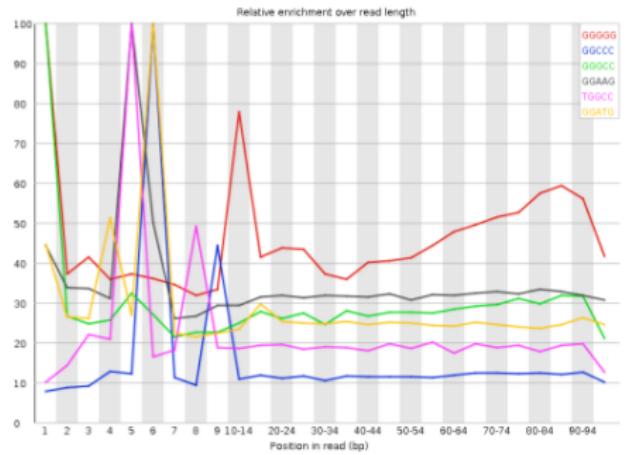
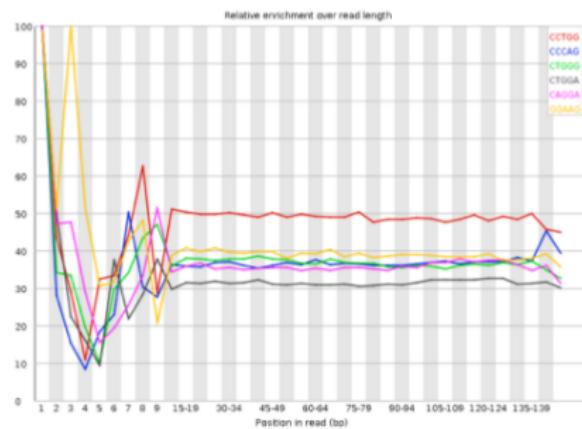
(11) FASTQC: Kmer content

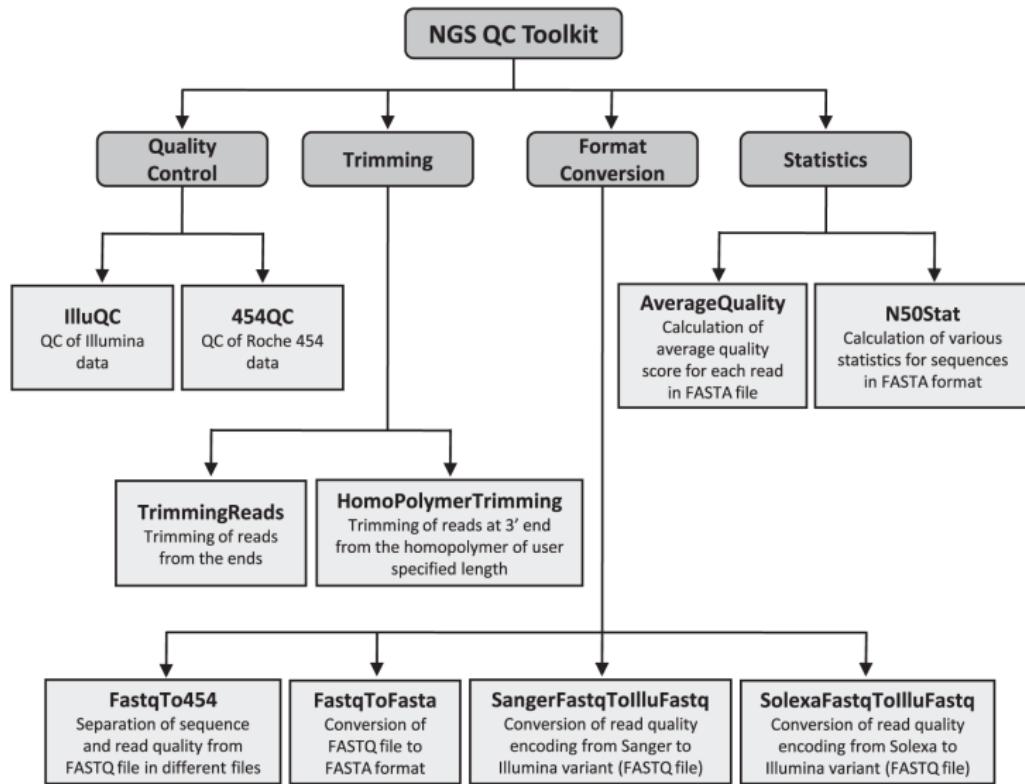
AAAAA k-mer that you're seeing at around 21 base pairs are arrested transcripts caused by cyclohexamide treatment.

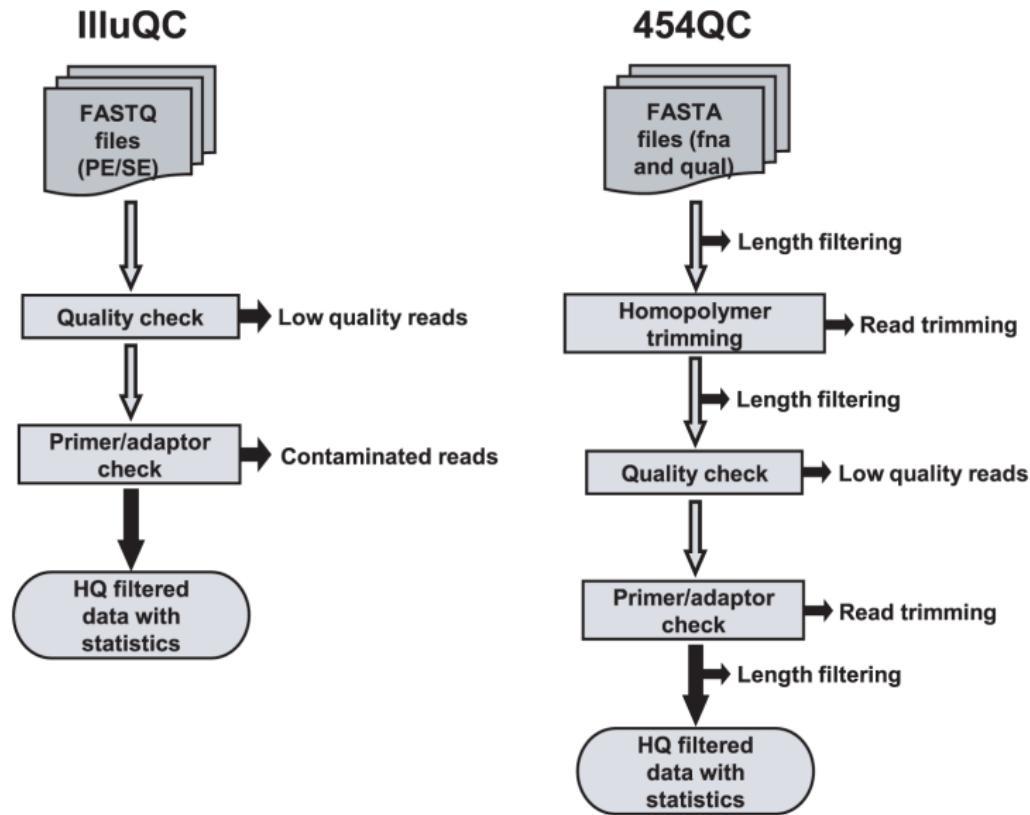


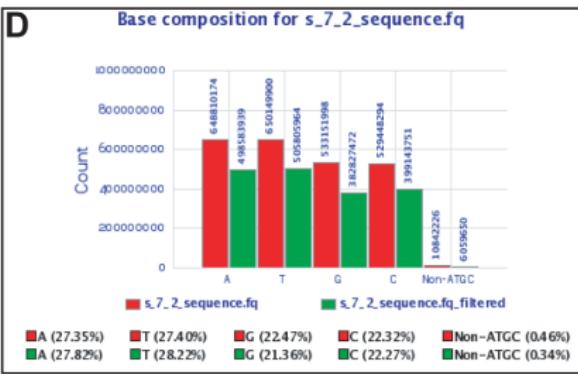
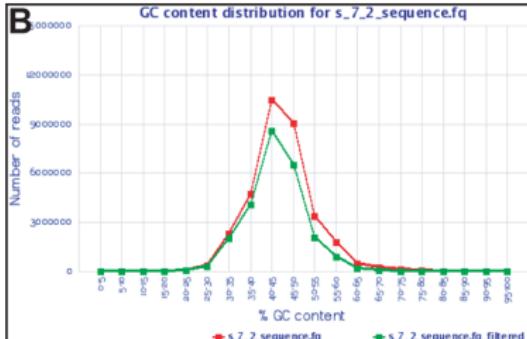
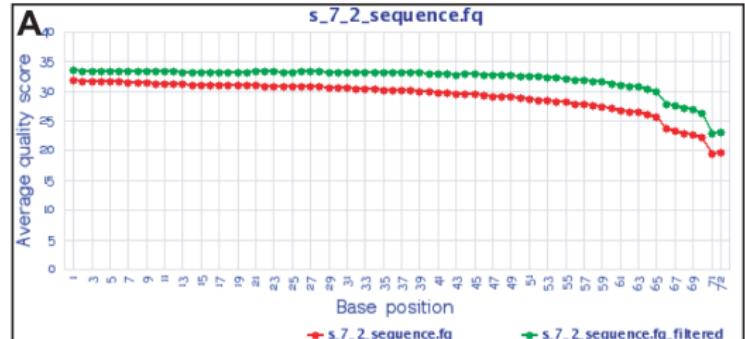
(11) FASTQC: Kmer content

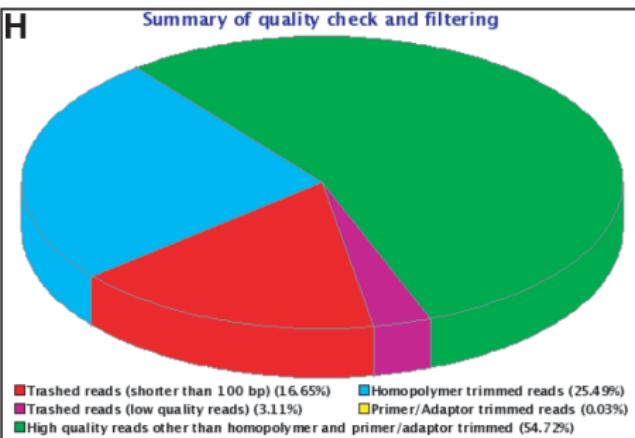
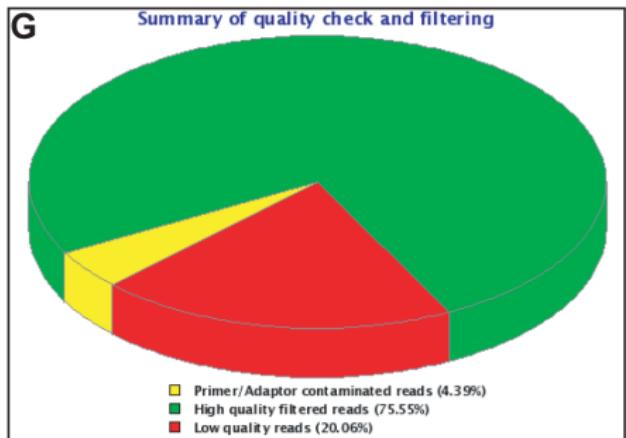
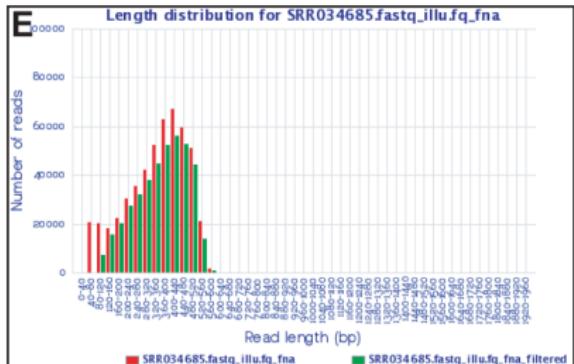
“Random” hexamer primer in RNA-seq libraries
(not that random after all)



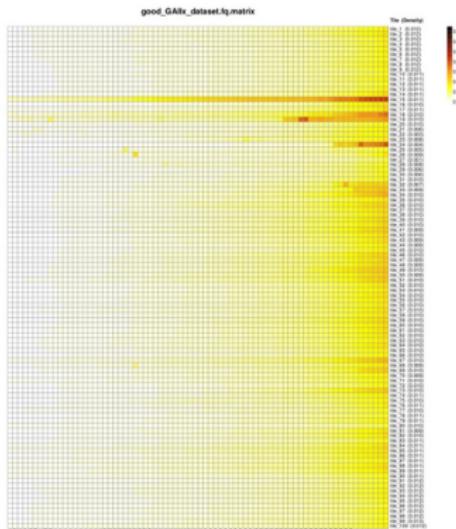




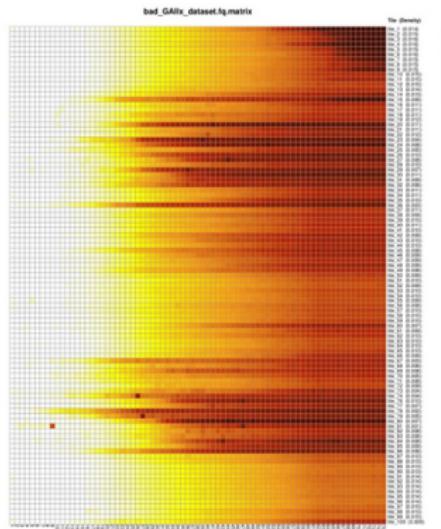


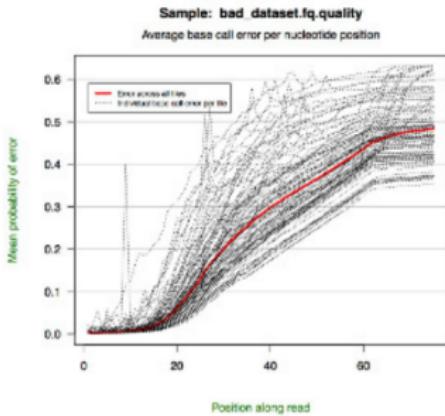
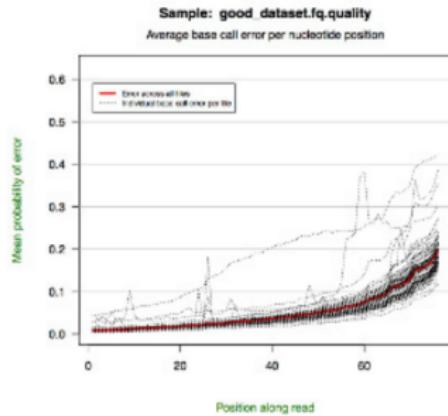


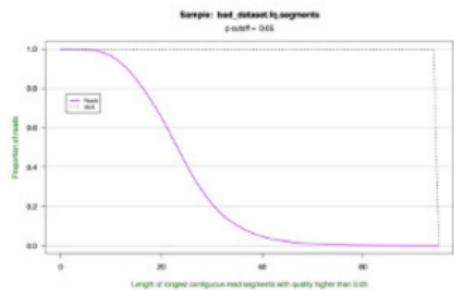
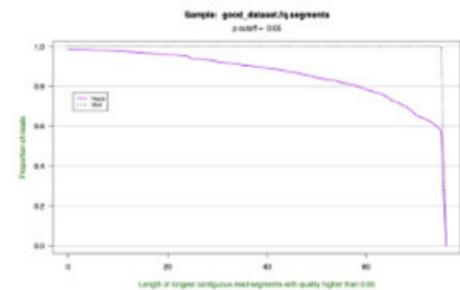
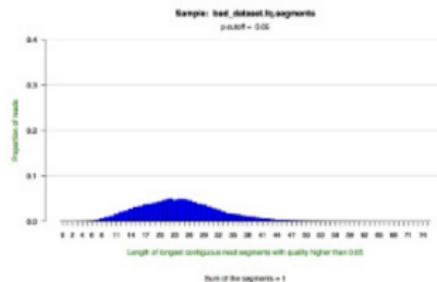
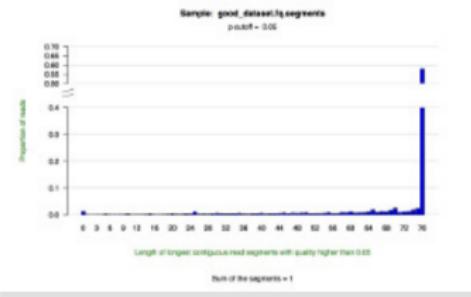
Good Dataset



Bad Dataset







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, ...
- ...

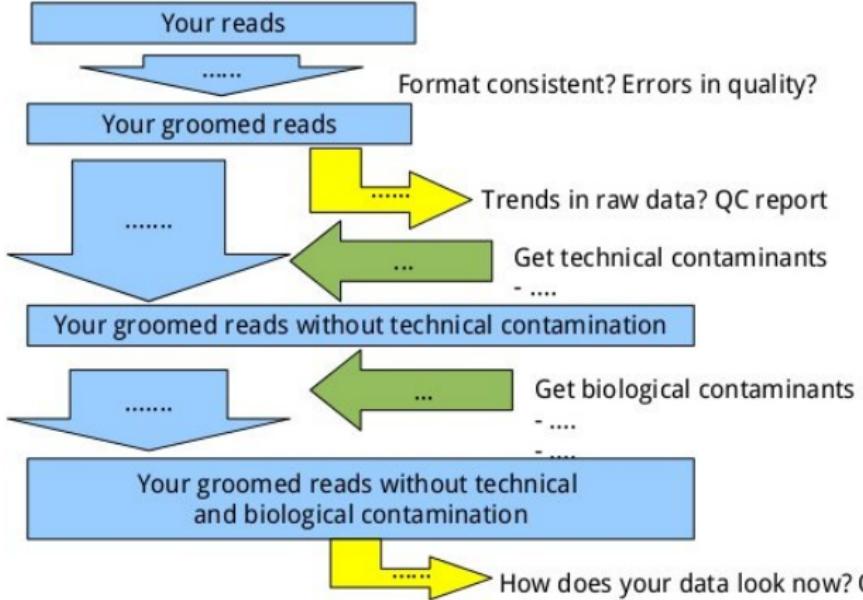
目的

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, ...
- ...

Summary preprocessing



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/)

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

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.



FASTX-Toolkit

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

PRINSEQ

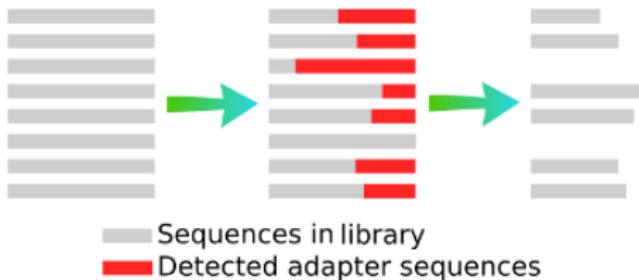
PRereprocessing and INformation of SEQuence data. A publicly available tool that is able to filter, reformat and trim your sequences and to provide you summary statistics for your sequence data.



- [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](#)



Clipping Illustration:



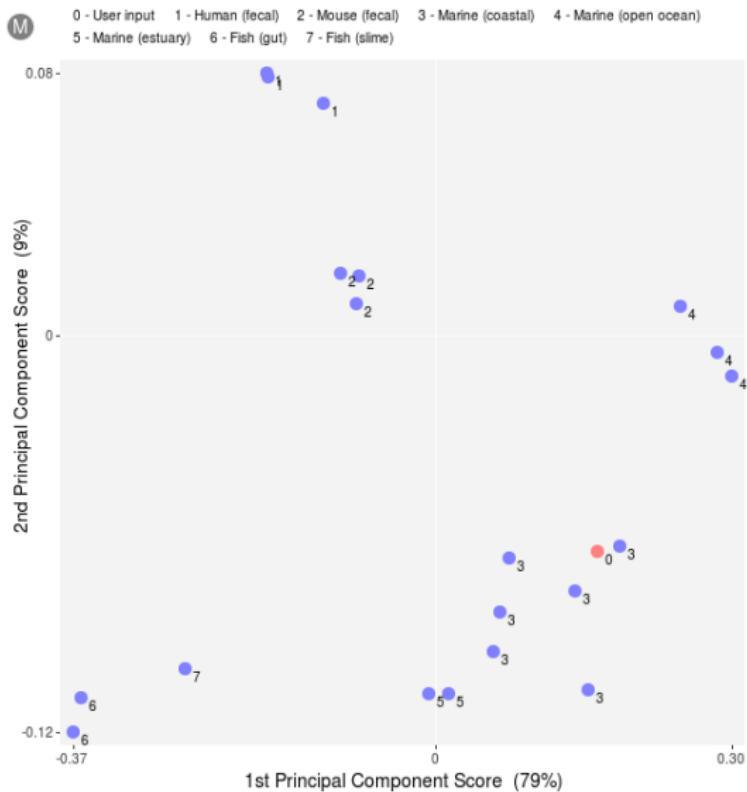
Clipping Example:

```
>1
ATGTAATGTTATATATATCGTAAATCCAACACAAT
>2
TATTTTGGATTCCACGACCCTGTAGGCACCATCAA
>3
ACGTTGTTGGTGCCTGACTGTAGGCACCATC
>4
TTTCTTCTTATCTCTCGAGTCTGTAGGCACCATCA
>5
TGGAACCTGCTGTAGGCACCATCATTATTTATAA
>6
TTTACCGGAAGCATAACTCTTCTGTAGGCACCATCA
>7
TGTATTAGCGGTGGGCCCGACTGTAGGCACCATCA
```

→

```
>2
TATTTTGGATTCCACGACC
>3
ACGTTGTTGGTGCCTGA
>4
TTTCTTCTTATCTCTCGAGT
>6
TTTACCGGAAGCATAACTCTT
>7
TGTATTAGCGGTGGGCCCGA
```





Data cleaning

We use Sickle for data cleaning.

For adapter clipping, we can use Cutadapt, Trimmomatic or the FastX toolkit (not in this practical).

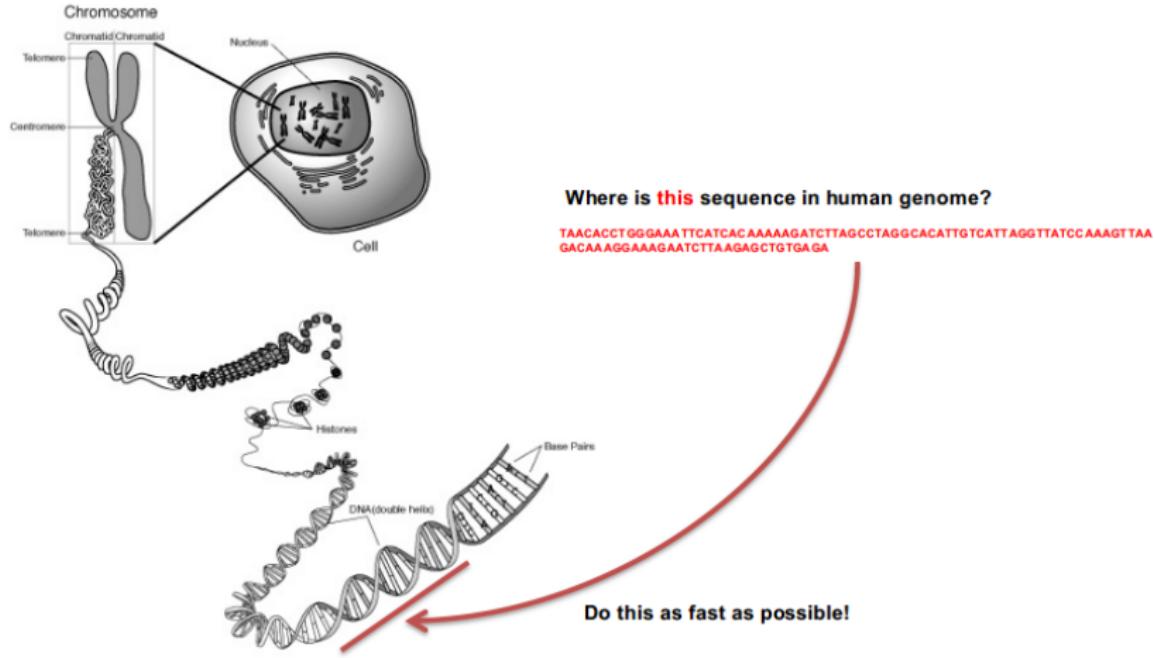
- Remove linker/adapter sequences.
- Trim low quality reads at the end of the read.
- Evaluate the part of the read that is left.

The FastQC tool kit is used for quality control (both before and after the data cleaning step).

- GC content.
- GC distribution.
- Quality scores distribution.

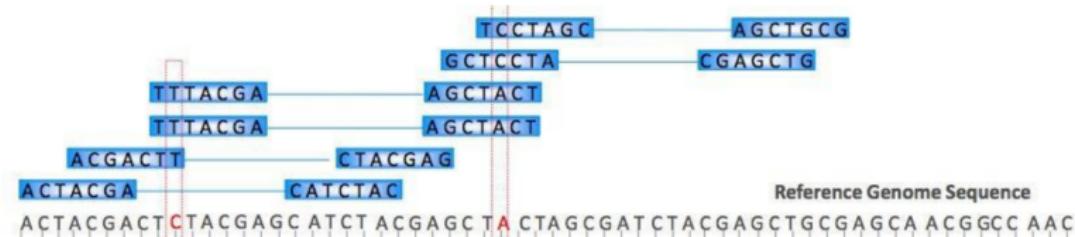


Mapping back to genome



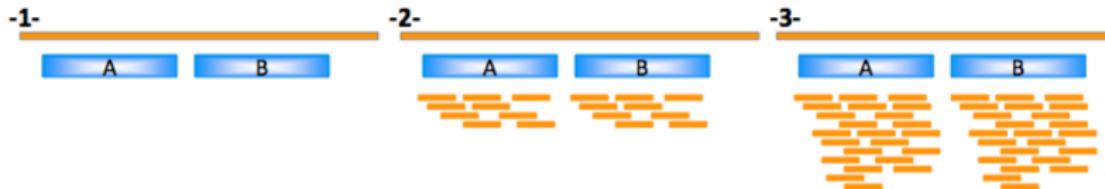
Mapping on a reference Genome

- Reads are aligned ≥ 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



Multiple Alignments

- A read can align **multiple times** on the genome (repeated elements...)
- How to deal with these multiple alignments reads ?
- Three strategies:
 - 1- Report only unique alignment
 - 2- Report best alignments & randomly assign reads across equally good loci
 - 3- Report all (best) alignments

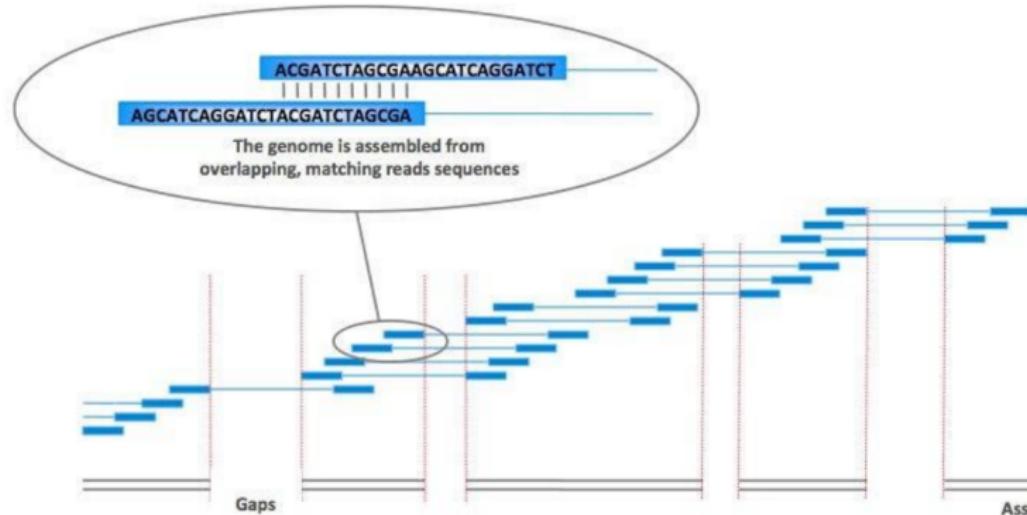


- **Mapping Quality:** quantify the probability that a read is misplaced.
→ Low if a read has multiple alignments



De Novo Reads Assembly

- **De Novo Reads Assembly:** used when there's no reference genome ; aims at reconstructing long scaffolds from single reads

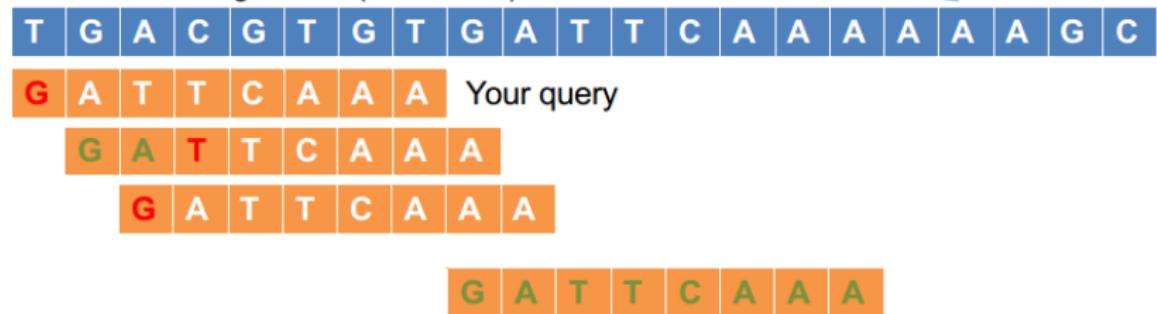


brute force way

Find “GATTCAAA” in human genome

This is very long (3 billion)

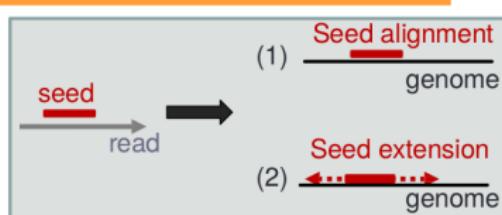
The reference genome (chr1, start)



Reads Alignment - Vocabulary

Mapping method: seed & extend

- 1.Align the seed (small part of the read)
- 2.Extend the seed to align the whole read



Mismatch: Incoherence between two nucleotides

Indels: Insertion/Deletion into the reference genome

Gap: Bridge within the read alignment (*i.e.* small indels)

Mappability: Uniqueness of a region

- repeated region = low mappability
- unique region = good mappability



HBS: A naive hash function

Let's assume: A = 1, C = 2, G = 4, T = 8, then: $HBS(S) = \sum_i HBS(S_i)$, e.g:

$$HBS(\text{AAAAAA}) = 1 + 1 + 1 + 1 + 1 + 1 = 6$$

$$HBS(\text{GTACG}) = 4 + 8 + 1 + 2 + 4 = 19$$

...

12345678901234567
TAACCCTAACCCCTAA

1234567890
AACCCTAACCC

Reference Genome

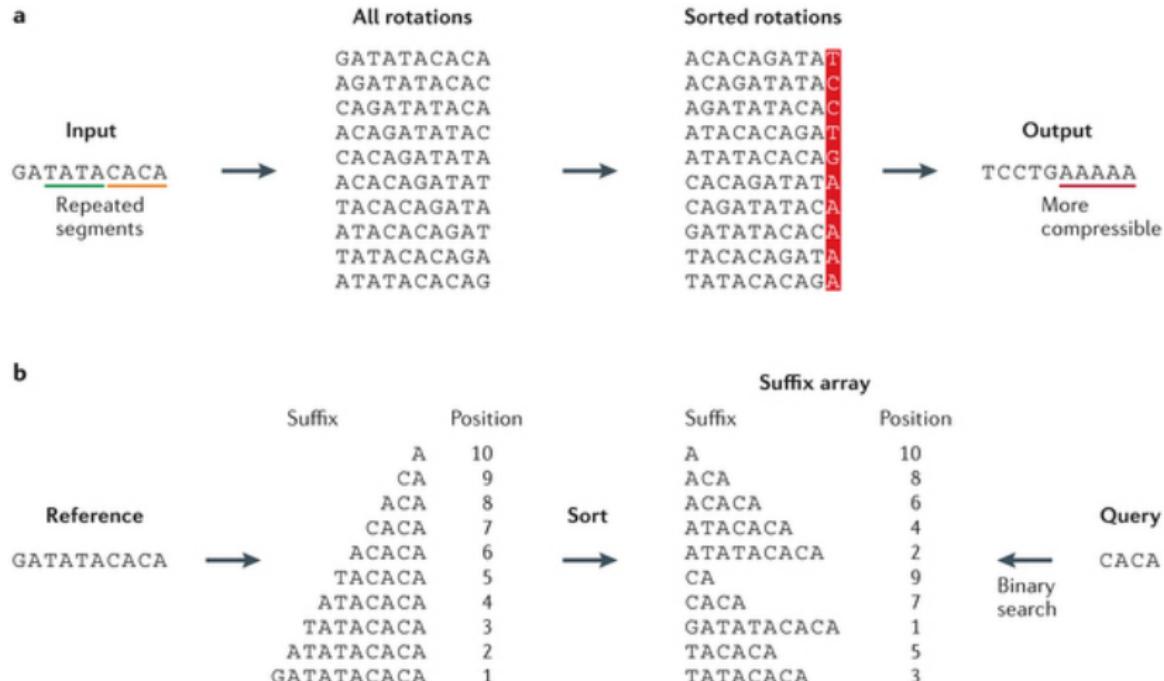
Index Table
...
20
...

HBS

$$2+2+8+4+4 = 20$$

Address Table
(CCTAA 11)
...







Hash Based Alignment Algorithms

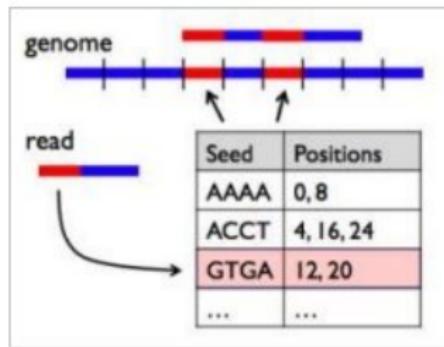
- Hash based

- Pick k-mer size, build lookup of every k-mer in the reference to its positions
- ~16GB of RAM required for hg19

- Seed-and-extend strategy

- Popular tools:

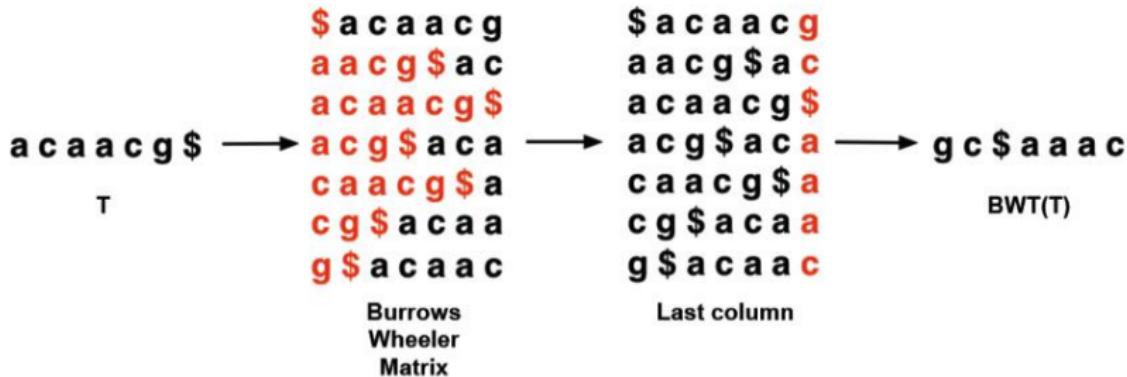
- BLAST: tunable for different uses
- MAQ (2008): Heng Li, et al
- NovaAlign: Slower, but very accurate
- Isaac (2013): High mem, but fast
- MOSAIK (2014): Hash clustering+SW

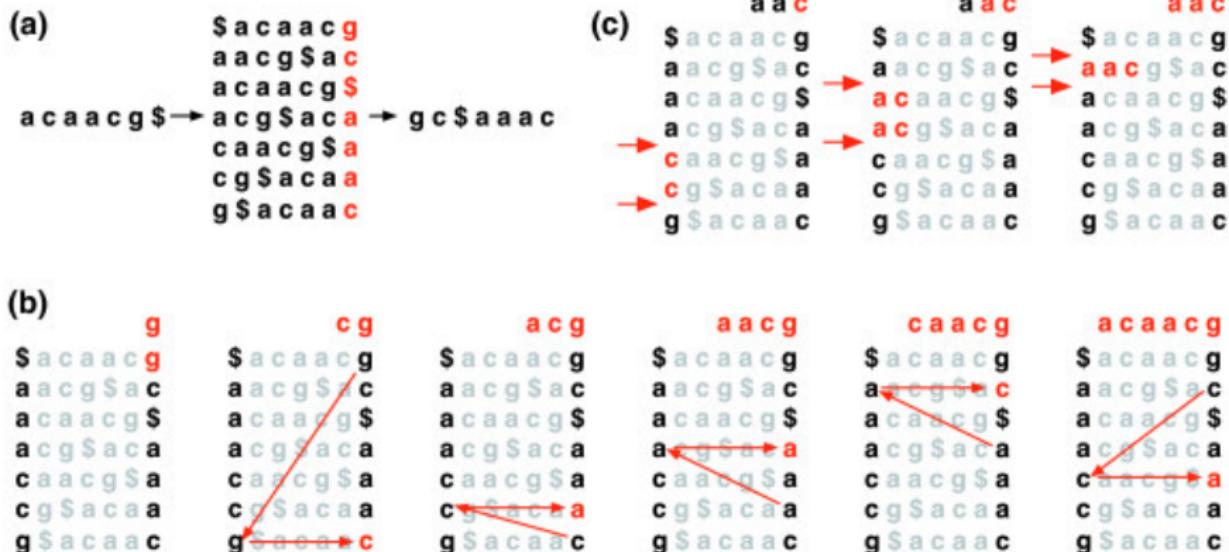


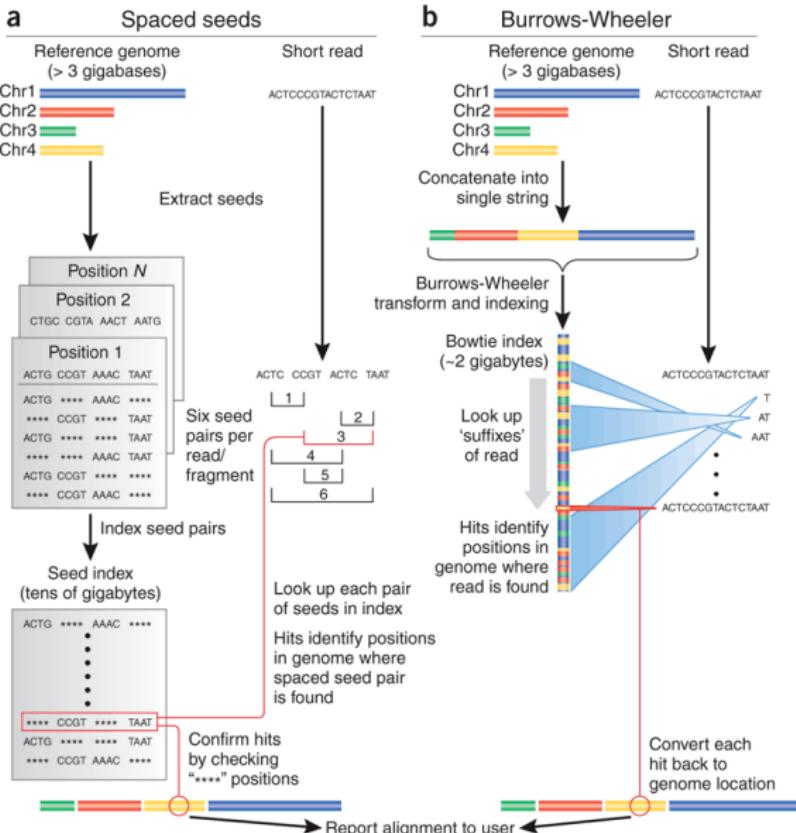
What is BWT?

- The Burrows and Wheeler transform (BWT) is a block sorting lossless and reversible data transform.
- The BWT can permute a text into a new sequence which is usually more “compressible”.
- Surfaced not long ago, 1994, by Michael Burrows and David Wheeler.
- The transformed text can be better compressed with fast locally-adaptive algorithms, such as run-length-encoding (or move-to-front coding) in combination with Huffman coding (or arithmetic coding).

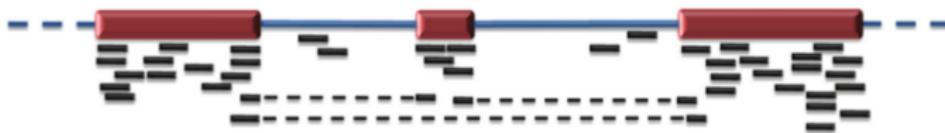








Alignment Algorithms



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

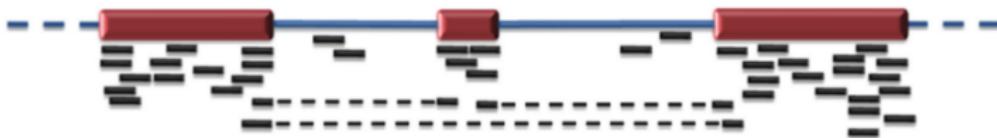
BFAST (Homer et al. 2009, based on BLAT) – multiple indexes, finds candidate alignment locations using seed and extend, followed by a gapped Smith-Waterman local alignment for each candidate

Many more!

http://en.wikipedia.org/wiki/List_of_sequence_alignment_software



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

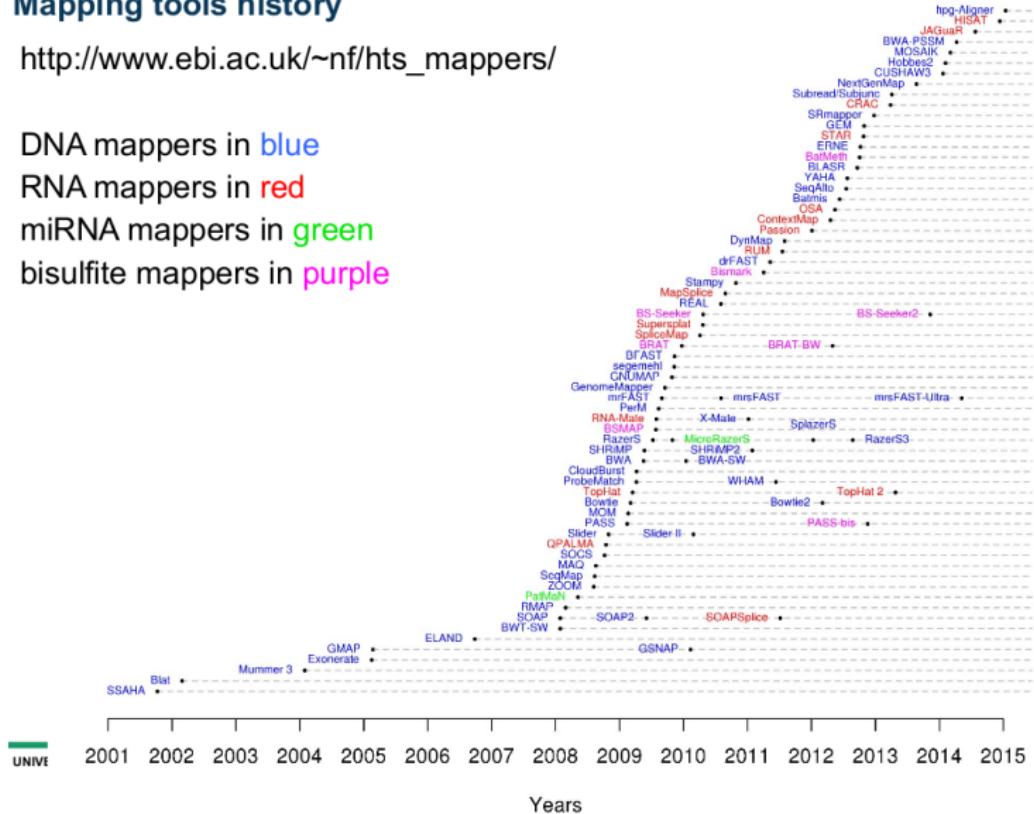
http://www.ebi.ac.uk/~nf/hts_mappers/

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



BWA

BWA consists of three algorithms: BWA-backtrack, BWA-SW and BWA-MEM.

- BWA-backtrack is designed for Illumina sequence reads up to 100bp, while BWA-SW and BWA-MEM for longer sequences ranged from 70bp to 1Mbp.
- BWA-MEM and BWA-SW share similar features such as long-read support and split alignment.
- BWA-MEM, which is the latest, is generally recommended for high-quality queries as it is faster and more accurate. BWA-MEM also has better performance than BWA-backtrack for 70-100bp Illumina reads.



Bowtie

Bowtie aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end).

Bowtie 2

Bowtie 2 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.

Bowtie

Bowtie aligns short DNA sequences (reads) to the human genome at a rate of over 25 million 35-bp reads per hour. Bowtie indexes the genome with a Burrows-Wheeler index to keep its memory footprint small: typically about 2.2 GB for the human genome (2.9 GB for paired-end).

Bowtie 2

Bowtie 2 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.

SOAP

SOAP is a program for efficient gapped and ungapped alignment of short oligonucleotides onto reference sequences.

SOAP2

SOAPAligner/soap2 is an updated version of SOAP software for short oligonucleotide alignment. The new program features in super fast and accurate alignment for huge amounts of short reads generated by Illumina/Solexa Genome Analyzer.

SOAP3

SOAP3 is a GPU-based software for aligning short reads with a reference sequence. It can find all alignments with k mismatches, where k is chosen from 0 to 3. When compared with its previous version SOAP2, SOAP3 can be up to tens of times faster.

SOAP

SOAP is a program for efficient gapped and ungapped alignment of short oligonucleotides onto reference sequences.

SOAP2

SOAPAligner/soap2 is an updated version of SOAP software for short oligonucleotide alignment. The new program features in super fast and accurate alignment for huge amounts of short reads generated by Illumina/Solexa Genome Analyzer.

SOAP3

SOAP3 is a GPU-based software for aligning short reads with a reference sequence. It can find all alignments with k mismatches, where k is chosen from 0 to 3. When compared with its previous version SOAP2, SOAP3 can be up to tens of times faster.

SOAP

SOAP is a program for efficient gapped and ungapped alignment of short oligonucleotides onto reference sequences.

SOAP2

SOAPAligner/soap2 is an updated version of SOAP software for short oligonucleotide alignment. The new program features in super fast and accurate alignment for huge amounts of short reads generated by Illumina/Solexa Genome Analyzer.

SOAP3

SOAP3 is a GPU-based software for aligning short reads with a reference sequence. It can find all alignments with k mismatches, where k is chosen from 0 to 3. When compared with its previous version SOAP2, SOAP3 can be up to tens of times faster.

基因组学 | 数据分析 | 流程 | 提取变异

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

Reference	AACTGAAACAGATAAAGCACTGGAAAGACGGCAGCAAAGAAAACCATG	T	GTAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA
GMI_1864715162(+)	TGGAAGACGGCAGCAAAGAAAACCATG	T	GTAAGC
GMI_813006088(+)	TGGAAGACGGCAGCAAAGAAAACCATG	T	GTAAGC
GMI_1079100245(-)	GAAGACGGCAGCAAAGAAAACCATG	T	GTAAGCAC
GMI_2159545344(+)	CAAAGAAAACCATG	T	GAAGACCTCTGCTACC
GMI_1776483420(-)	AAGAAACAAACATG	T	GTAAGCACCTCTGCACCCCA
GMI_2093120226(-)	AAGAAACAAACATG	T	GTAAGCACCTCTGCACCCCA
GMI_2204137276(-)	AAGAAACAAACATG	T	GTAAGCACCTCTGCACCCCA
GMI_594634658(-)	AAGAAACAAACATG	T	GTAAGCACCTCTGCACCCCA
GMI_878606194(-)	AAGAAACAAACATG	T	GTAAGCACCTCTGCACCCCA
GMI_1463224112(+)	AGAAACAAACATG	T	GTAAGCACCTCTGCACCCAC
GMI_1385601163(-)	ACATG	T	GTAAGCACCTCTGCACCCACTAGCGAGC
GMI_2247152549(+)	TG	T	GTAAGCACCTCTGCACCCACTAGCGAGCTAG
GMI_2552461258(+)	AAACAAACATG	T	GTAAGCACCTCTGCACCCACTAGCGAGCTAGGGAGAGTTGGCGTCA

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

Reference	AACTGAAACAGATAAAGCACTGGAAAGACGGCAGCAAAGAAAACCATG	C	GTAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA
GMI_661419350(+)	AACTGAAACAGATAAAGCACTGGAAAGACGGCAGCAAAGAAAACCATG	C	GTAAGCACCTCTGCACCCACCA
GMI_661419351(-)	AACTGAAACAGATAAAGCAACCGGAAGACGGCAGCAAAGAAAACCATG	C	GTAAGCACCTCTGCACCCACTAGCG
GMI_661419352(-)	AACTGAAACAGATAAAGCACTGGAAAGACGGCAGCAAAGAAAACCATG	C	GTAAGC
GMI_661419354(-)	AACTGAAACAGAAAAGGCAACCGGAAGACGGCAGCAAAGAAAACCATG	C	GTAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCT
GMI_661419355(+)	AACTGAAACAGATAAAGCACTGGAAAGACGGCAGCAAAGAAAACCATG	C	GTAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA
GMI_661419356(+)	AACTGAAACAGATAAAGCACTGGAAAGACGGCAGCAAAGAAAACCATG	C	GTAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGGTGGCGCCA
GMI_661419357(-)	AAACAGATAAAGCACTGGAAAGACGGCAGCAAAGAAAACCATG	C	GTAAGCACCTCTGCACCCACTAGCGAGC
GMI_661419358(+)	GATAAAAGCACTGGAAAGACGGCAGCAAAGAAAACCATG	C	GTAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGGTTGGCG
GMI_661419359(-)	AAGCAACTGGAAAGACGGCAGCAAAGAAAACCATG	C	GTAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGG
GMI_661419362(-)	CAGCAAAGAAAACCATG	C	GTAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA
GMI_661419363(-)	CAGCAAAGAAAACCATG	C	GTAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA
GMI_661419364(+)	AGCAAAGAAAACCATG	C	GTAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA
GMI_661419365(+)	AGCAAAGAAAACCATG	C	GTAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA
GMI_661419366(+)	AAACAAACATG	C	GTAAGCACCTCTGCACCCACTAGCGAGCTAGAGAGAGTTGGCGTCA



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





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



VarScan

VarScan is a platform-independent mutation caller for targeted, exome, and whole-genome resequencing data generated on Illumina, SOLiD, Life/PGM, Roche/454, and similar instruments. It can be used to detect different types of variation:

- Germline variants (SNPs and indels) in individual samples or pools of samples.
- Multi-sample variants (shared or private) in multi-sample datasets (with mpileup).
- Somatic mutations, LOH events, and germline variants in tumor-normal pairs.
- Somatic copy number alterations (CNAs) in tumor-normal exome data.

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).

SeattleSeq Annotation

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

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).

SeattleSeq Annotation

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

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).

SeattleSeq Annotation

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

Features

- Supports over 38,000 genomes
- Standard ANN annotation format
- Cancer variants analysis
- GATK compatible (-o gatk)
- HGVS notation
- Sequence Ontology standardized terms



ANNOVAR

Given a list of variants with chromosome, start position, end position, reference nucleotide and observed nucleotides, ANNOVAR can perform:

Gene-based annotation identify whether SNPs or CNVs cause protein coding changes and the amino acids that are affected.

Region-based annotation identify variants in specific genomic regions (conserved regions among 44 species, database of genomic variants, or many other annotations on genomic intervals).

Filter-based annotation identify variants that are documented in specific databases, (whether a variant is reported in dbSNP, calculate the SIFT/PolyPhen/... scores, or many other annotations on specific mutations).

Other functionalities Retrieve the nucleotide sequence in any user-specific genomic positions in batch, identify a candidate gene list for Mendelian diseases from exome data, and other utilities.

wANNOVAR

wANNOVAR is a web server that provides easy and intuitive web-based access to the most popular functionalities of the ANNOVAR software. Users can upload a VCF file and obtain annotated results as tab-delimited or comma-separated files; in addition, simple variants reduction can be performed to prioritize deleterious variants from the input files. Currently, wANNOVAR supports only human genome annotation.



SeattleSeq Annotation

This annotation includes:

- dbSNP rs IDs
- gene names and accession numbers
- variation functions (e.g. missense)
- protein positions and amino-acid changes
- conservation scores
- HapMap frequencies
- PolyPhen predictions
- clinical association



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.



基因组学 | 数据分析 | 流程 | 注释变异 | SIFT

Human Genome DB	Tool Description
SIFT/PROVEAN Human SNPs	Get SIFT and PROVEAN predictions for SNPs and indels (Ensembl 66) (Sample format)
SIFT Human SNPs	Get SIFT predictions for nonsynonymous SNPs (Ensembl 63) (Sample format)
	Other human genome tools: <ul style="list-style-type: none">• Restrict to Coding Variants (Sample format)• Classify Human indels (Sample format)
SIFT Human Protein DB	Tool Description (Ensembl 63)
SIFT Human Protein	Get SIFT predictions for nonsynonymous AA substitutions (Ensembl ENSP ID)
SIFT dbSNP DB	Tool Description (dbSNP Build 132)
SIFT dbSNP rs IDs	Get SIFT predictions for dbSNP SNPs including non-human species (NCBI rs ID)
SIFT dbSNP Protein	Get SIFT predictions for dbSNP proteins including non-human species (RefSeq ID or GI number)
SIFT Single Protein Tools	Tool Description
SIFT BLINK	Run SIFT analysis on single protein using precomputed BLAST from NCBI BLINK (RefSeq ID or GI number)
SIFT Sequence	Run SIFT analysis on single protein through a PSI-BLAST search (fasta)
SIFT Related Sequences	Run SIFT analysis on protein query and a group of related sequences (multi-fasta)
SIFT Aligned Sequences	Run SIFT analysis on protein query already in multi-sequence alignments (MSA)



Batch Query Data		Sample Batch
Batch query		<pre>Q92889 706 I T Q92889 875 E G XRCC1_HUMAN 399 R Q NP_005792 59 L P rs1799931 chr1:1267483 G/A chr1:1158631 A/C,G,T</pre>
Upload batch file	<input type="button" value="Browse..."/>	No file selected.
Query description	<input type="text"/>	
E-mail address	<input type="text"/>	
Protein Sequences (optional)		
Upload FASTA file	<input type="button" value="Browse..."/>	No file selected.
File description	<input type="text"/>	
Advanced Options		
Classifier model	<input type="button" value="HumDiv"/>	
Genome assembly	<input type="button" value="GRCh37/hg19"/>	
Transcripts	<input type="button" value="Canonical"/>	
Annotations	<input type="button" value="Missense"/>	
<input type="button" value="Submit Batch"/> <input type="button" value="Clear"/> <input type="button" value="Check Status"/>		



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

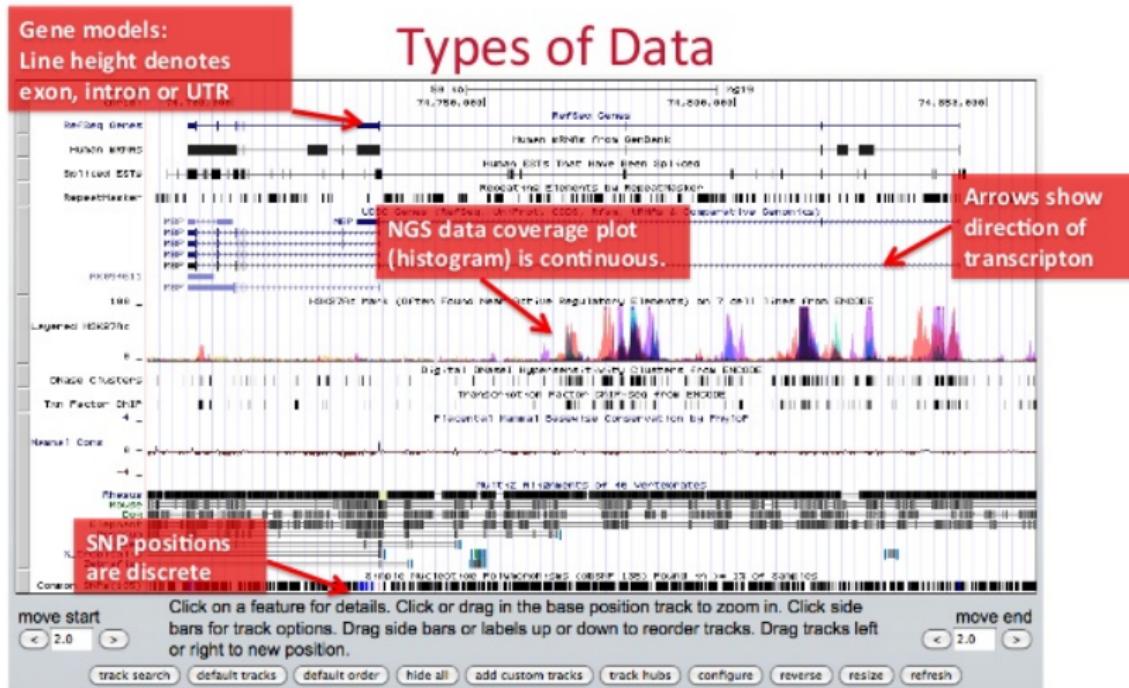
The Integrative Genomics Viewer (IGV) is a high-performance visualization tool for interactive exploration of large, integrated genomic datasets.

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.



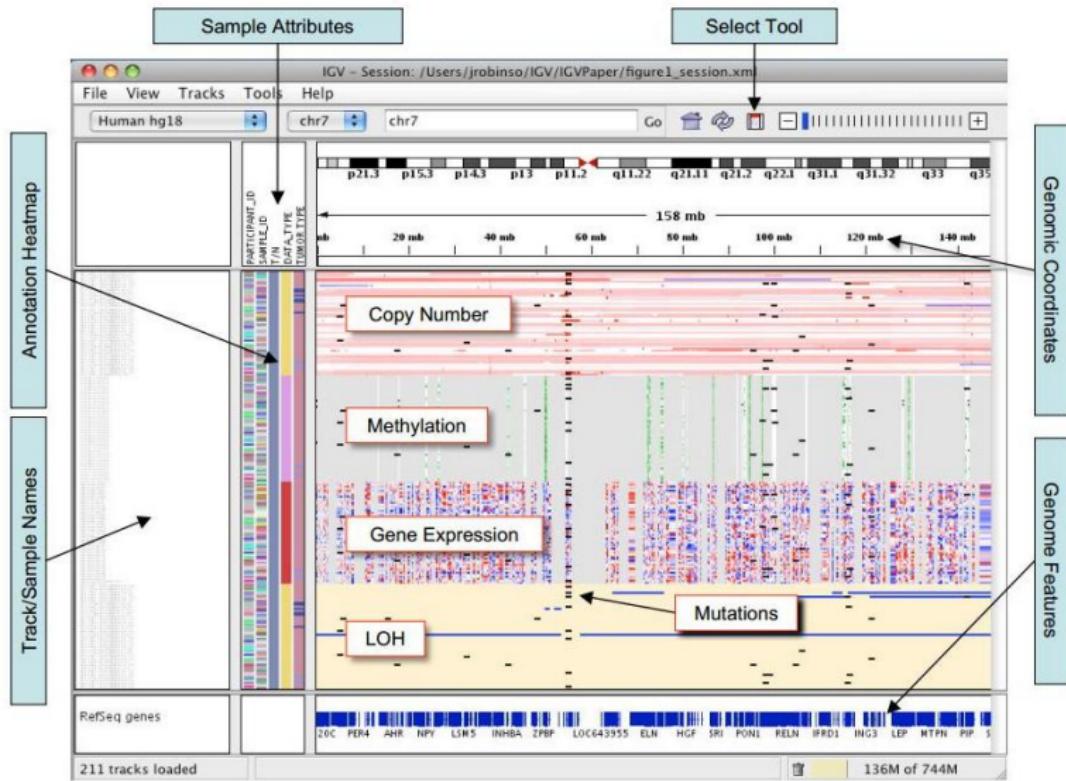
Why IGV

- IGV is an integrated visualization tool of large data types
- View large dataset easily
- Faster navigation on browsing
- Run it locally on your computer
- Easy to use interface



IGV interface

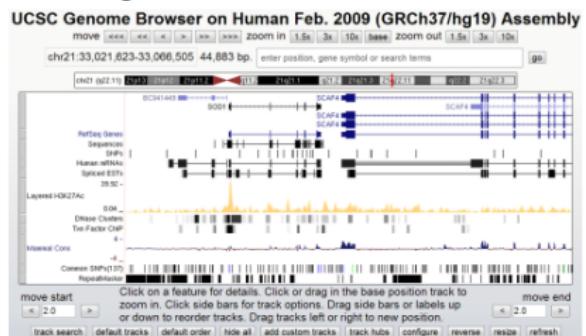






Genome browser

UCSC genome browser

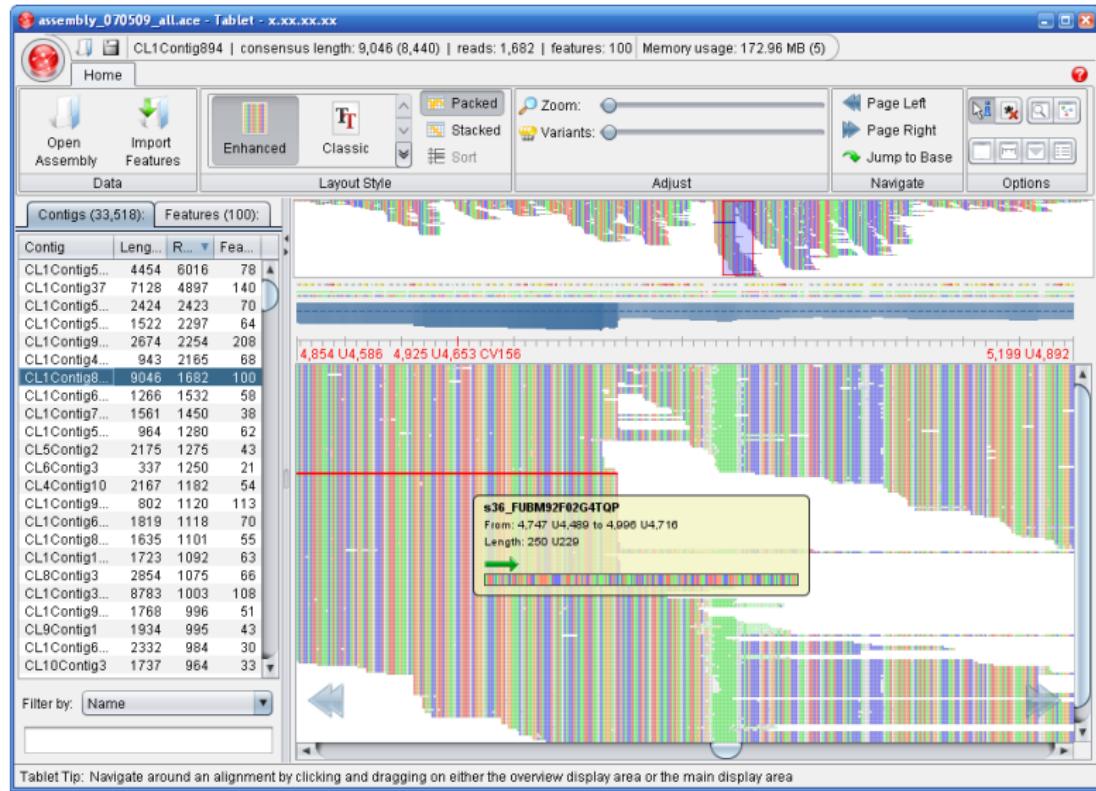


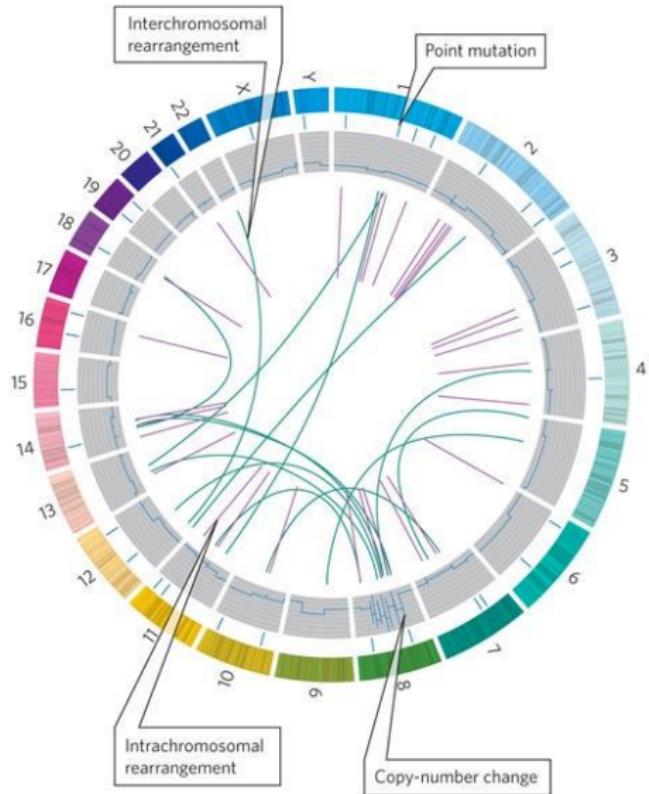
Pros: very comprehensive
Cons: data have to be uploaded or transmitted via network dynamically

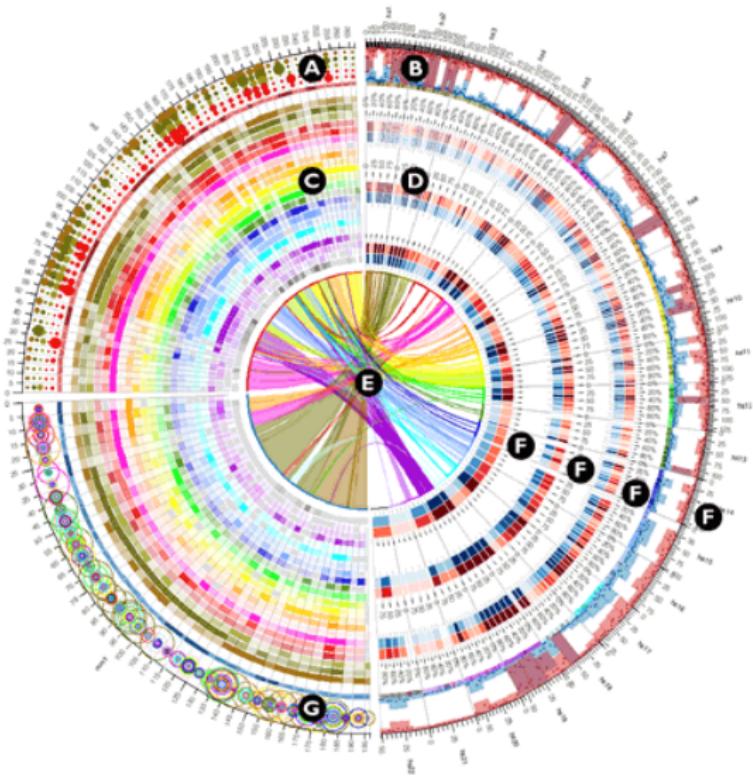
V.S.

Pros: locally installed
Cons: less genome annotation









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.

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.



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.

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.



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



Accessibility

Need an analysis? There's a tool for that.

- Get Data**
- Upload File from your computer
 - [UCSC Main table browser](#)
 - [UCSC Test table browser](#)
 - [UCSC Archaea table browser](#)
 - [BX main browser](#)
 - [EBI SRA ENA SRA](#)
 - [Get Microbial Data](#)
 - [BioMart Central server](#)
 - [BioMart Test server](#)
 - [CBI Rice Mart rice mart](#)
 - [GrameneMart Central server](#)
 - [modENCODE fly server](#)
 - [Flymine server](#)
 - [Flymine test server](#)
 - [modENCODE modMine server](#)
 - [Ratmine server](#)
 - [YeastMine server](#)
 - [metabolicMine server](#)
 - [modENCODE worm server](#)
 - [WormBase server](#)
 - [Wormbase test server](#)
 - [EuPathDB server](#)
 - [EncodeDB at NHGRI](#)
 - [EpiGRAPH server](#)
 - [EpiGRAPH test server](#)

- NGS: Mapping**
- Lastz map short reads against reference sequence
 - Lastz paired reads map short paired reads against reference sequence
 - [Map with Bowtie for Illumina](#)
 - [Map with Bowtie for SOLID](#)
 - [Map with BWA for Illumina](#)
 - [Map with BWA for SOLID](#)
 - [Map with BFAST](#)
 - Megablast compare short reads against htgs, nt, and wgs databases
 - [Parse blast XML output](#)
 - [Map with PerM for SOLID and Illumina](#)
 - [Re-align with SRMA](#)
 - [Map with Mosaik](#)

- NGS: RNA Analysis**
- RNA-SEQ
- Tophat for Illumina Find splice junctions using RNA-seq data
 - Tophat for SOLID Find splice junctions using RNA-seq data
 - Cufflinks transcript assembly and FPKM (RPKM) estimates for RNA-Seq data
 - Cuffcompare compare assembled transcripts to a reference annotation and track Cufflinks transcripts across multiple experiments
 - Cuffdiff find significant changes in transcript expression, splicing, and promoter use
- FILTERING
- Filter Combined Transcripts using tracking file

- NGS: GATK Tools (beta)**
- ALIGNMENT UTILITIES
- [Depth of Coverage](#) on BAM files
 - [Realigner Target Creator](#) for use in local realignment
 - [Indel Realigner](#) – perform local realignment
- BASE RECALIBRATION
- [Count Covariates](#) on BAM files
 - [Table Recalibration](#) on BAM files
 - Analyze Covariates – draw plots
- GENOTYPING
- [Unified Genotyper](#) SNP and indel caller
- ANNOTATION
- [Variant Annotator](#)
- FILTRATION
- [Variant Filtration](#) on VCF files
- VARIANT QUALITY SCORE RECALIBRATION
- [Variant Recalibrator](#)
 - [Apply Variant Recalibration](#)
- VARIANT UTILITIES
- [Validate Variants](#)
 - [Eval Variants](#)
 - [Combine Variants](#)



Visualization within Galaxy

Galaxy

- Tool integration framework
- Heavy focus on usability
- Sharing, publication framework

Genome Browser

- Physical depiction of data
- Visually identify correlations
- Find interesting regions, features

Trackster



HTS Analysis

- Set parameters
- Run tools / workflow
- Wait...
- Visualize output



Three ways to use Galaxy

Public website

Download and Run Locally

Run on the Cloud



Public Galaxy Servers

<https://wiki.galaxyproject.org/PublicGalaxyServers>

Interested in:

60+ Public Servers

- ✓ ChIP-chip and ChIP-seq?
- ✓ Cistrome
- Statistical Analysis?
- ✓ Genomic Hyperbrowser
- Sequence and tiling arrays?
- ✓ Oqtans
- Text Mining?
- ✓ DBCLS Galaxy
- Reasoning with ontologies?
- ✓ GO Galaxy
- Internally symmetric protein structures?
- ✓ SymD



How to use Galaxy?

GALAXY MAIN: User disk quotas 250GB for registered users, maximum concurrent jobs: 8

	NO WAIT TIMES	NO STORAGE QUOTAS	NO JOB SUBMISSION LIMITS	NO DATA TRANSFER BOTTLENECKS	NO IT EXPERIENCE REQUIRED	NO REQUIRED INFRASTRUCTURE	COST
GALAXY MAIN	✗	✗	✗	✗	✓	✓	Free
LOCAL GALAXY	?	?	?	✓	✗	✗	Free ?
CLOUD GALAXY (AMAZON)	✓	✓	✓	✗	✗	✓	동일사양 대비 약 2배 (KT의)
SLIPSTREAM GALAXY	✓	✓	✓	✓	✓	✓	\$19,995 (2천2백만원)
KT GenomeCloud GALAXY	✓	✓	✓	✓	✓	✓	시간당 740원 부터



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.

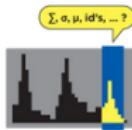


Utility	Description
annotate	Annotate coverage of features from multiple files.
bamtobed	Convert BAM alignments to BED (& other) formats.
bamtofastq	Convert BAM records to FASTQ records.
bed12tobed6	Breaks BED12 intervals into discrete BED6 intervals.
bedpetobam	Convert BEDPE intervals to BAM records.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



SET OPERATIONS

- **bedops** - apply set operations on any number of BED inputs
- **bedextract** - efficiently extract BED features
- **closest-features** - matches nearest features between BED files



STATISTICS

- **bedmap** - map overlapping BED elements onto target regions and optionally compute any number of common statistical operations



FILE MANAGEMENT

- **sort-bed** - apply lexicographical sort to BED data
- **starch** and **unstarch** - compress and extract BED data
- **starchcat** - merge compressed archives
- **Conversion tools** - convert common genomic formats to BED



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.



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

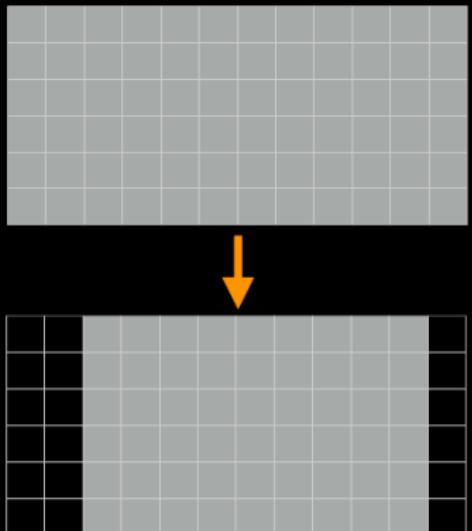
6 回顾与总结

- 总结
- 思考题



NGS Data Quality: Trim as we see fit

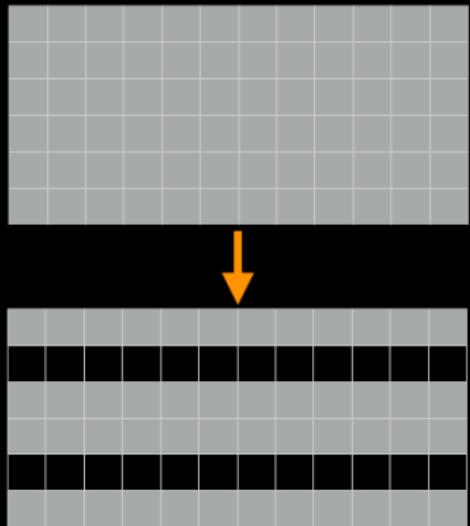
- Trim as we see fit: Option 1
- NGS QC and Manipulation →
FASTQ Trimmer by column
- Trim same number of columns
from every record
- Can specify different trim for 5'
and 3' ends



NGS Data Quality: Base Quality Trimming

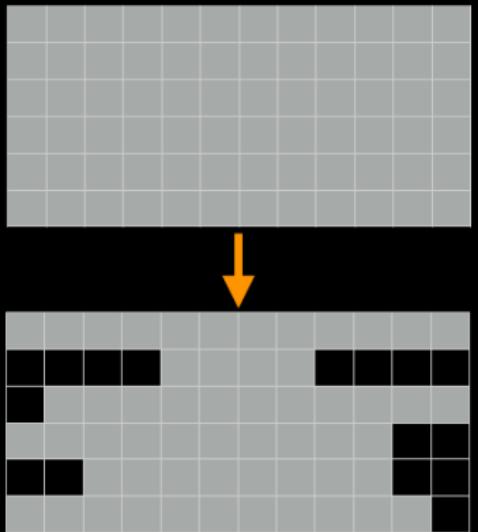
- Trim Filter as we see fit: Option 2

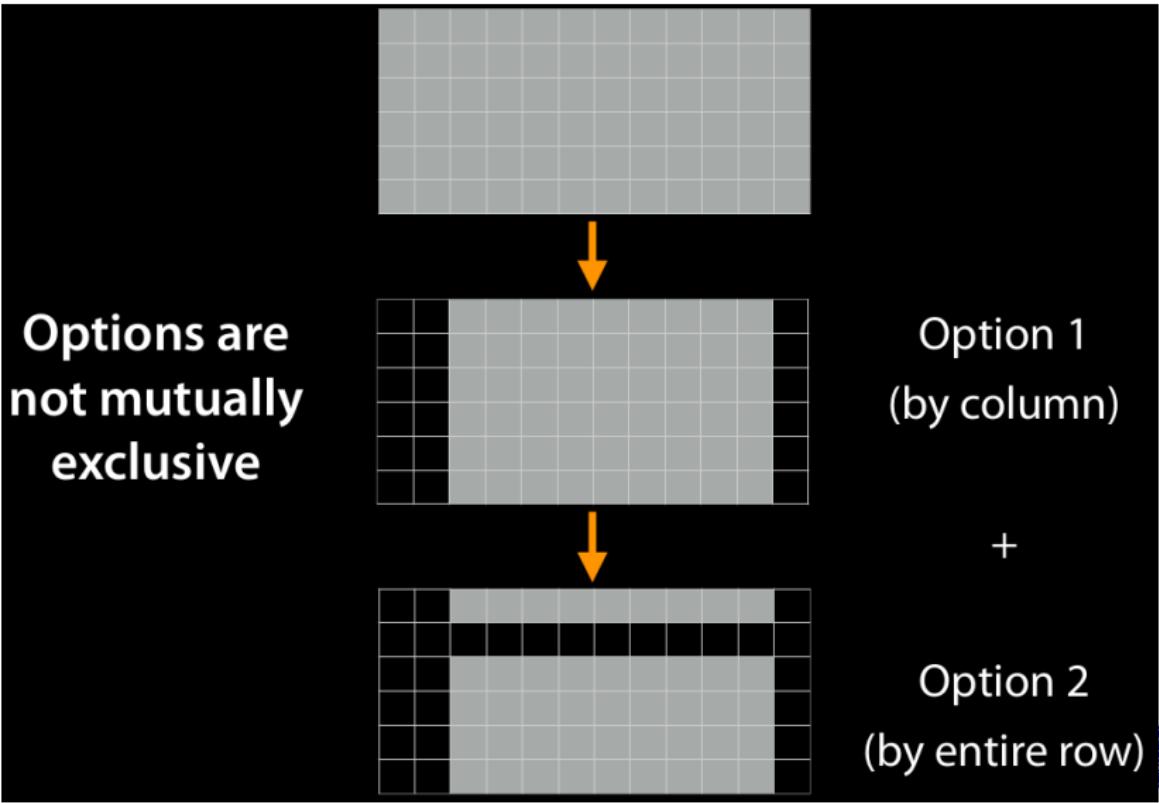
- NGS QC and Manipulation →
Filter FASTQ reads by quality score and length
- Keep or discard whole reads
- Can have different thresholds for different regions of the reads.
- Keeps original read length.



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





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

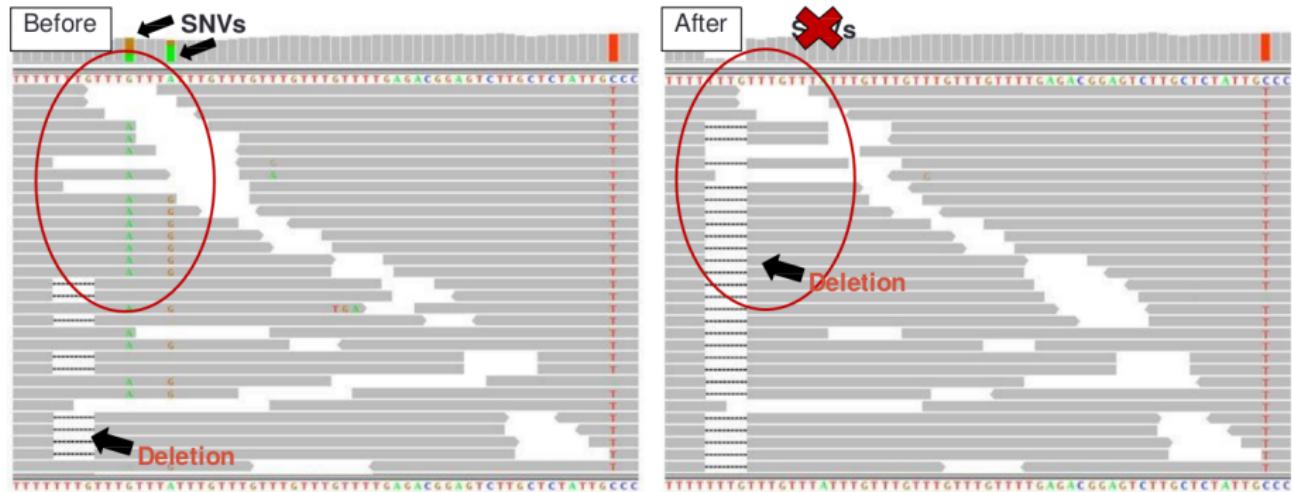


Why realign around indels?

- Small Insertion/deletion (Indels) in reads (especially near the ends) can trick the mappers into mis-aligning with mismatches
 - The mapper has to find a trade-off between adding several mismatches and opening one gap
 - Alignment scoring – cheaper to introduce multiple Single Nucleotide Variants (SNVs) than an indel : induce a lot of false positive SNVs
 - The mapper align reads (or pairs) independently
- Realignment around indels helps improve the accuracy of the alignment by having a global view of the alignment at one specific position



Local realignment around indels



Three types of realignment targets

- The idea is 1) to identify loci in need of local realignment then 2) apply realignment
- Types of realignment targets:
 - **Known sites:** Common polymorphisms (dbSNP, 1000Genomes)
→ On Galaxy: add « dbsnp[...].vcf » & « Mills[...].vcf » as Binding Reference Ordered Data in GATK
 - **Indels** seen in original alignments (in CIGAR, indicated by I for Insertion or D for Deletion)
 - Sites where evidences suggest a hidden indel (lot of SNVs in a small loci)



Indel realignment steps/tools

1. Identify what regions need to be realigned

→ **RealignerTargetCreator**

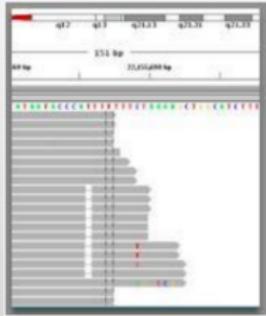
+ known sites



↓
Intervals

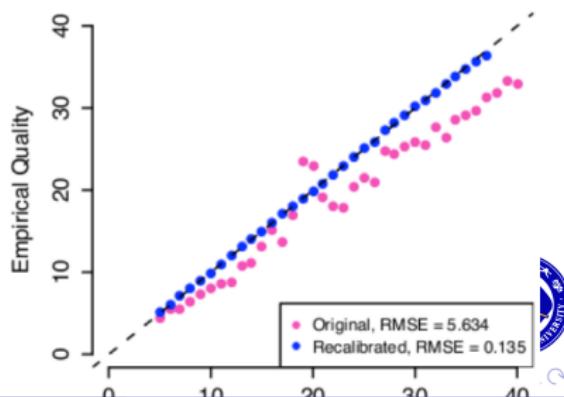
2. Perform the actual realignment (BAM output)

→ **IndelRealigner**

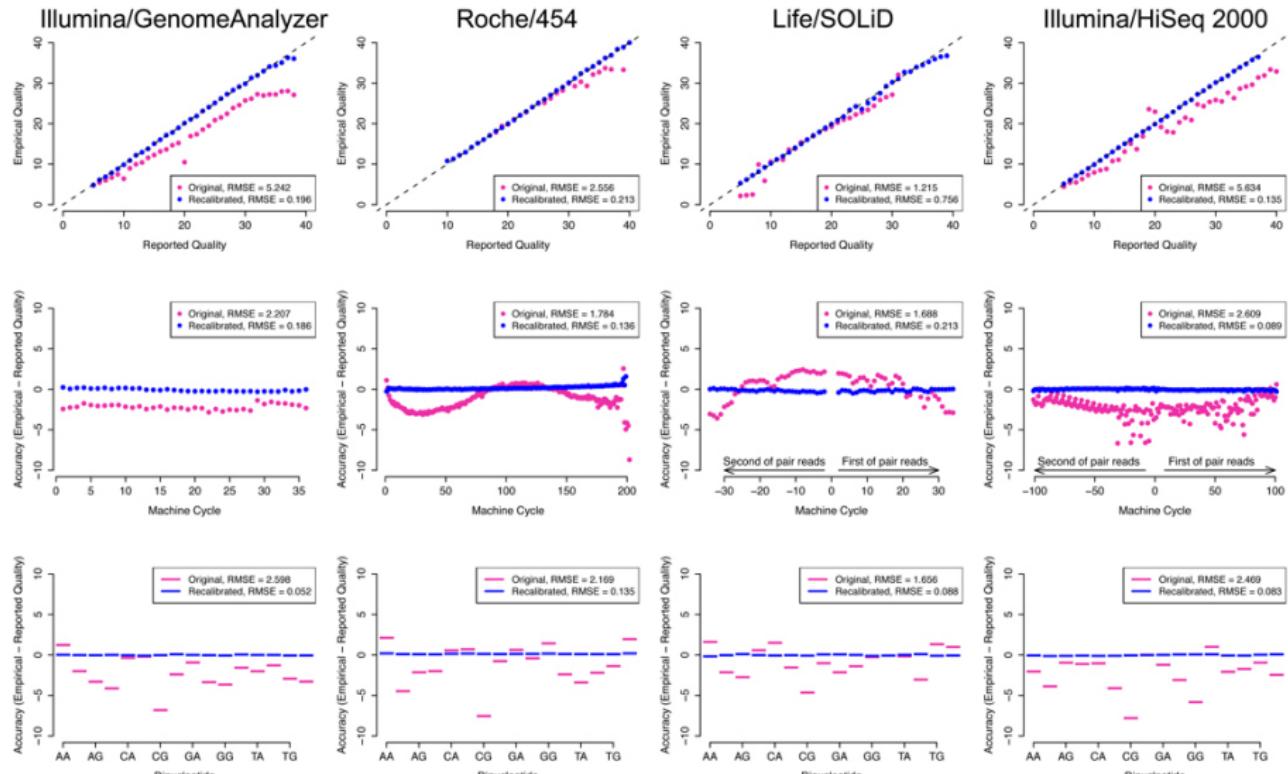


GATK Preprocess: Base Quality Score Recalibration

- 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



基因组学 | 数据分析 | 补遗 | 比对 | Base quality recalibration



BQSR tools

1. Calculate the covariates

→ **BaseRecalibrator**

+ known sites



↓ Covariates

1. Apply the covariates to the alignments

→ **PrintReads**



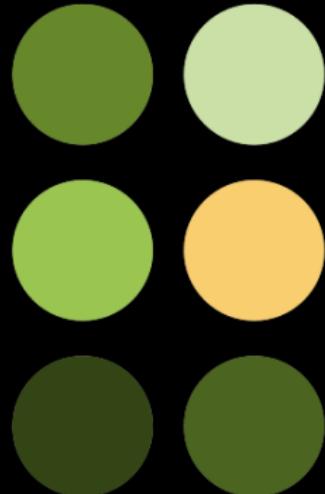
Replicates, replicates, replicates

Mutant



Why replicate?

Wild-type



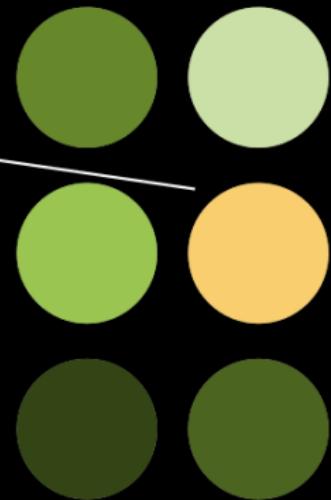
Biological heterogeneity

Unreplicated design

Mutant



Wild-type

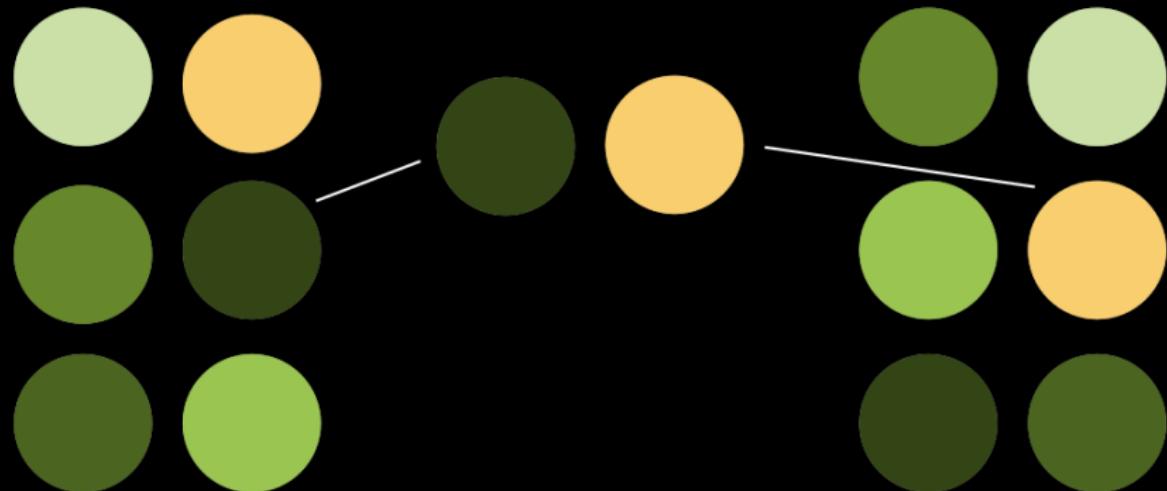


Here, groups differ, but single replicates from each group very similar

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

Importance of replication

Power increases substantially with number of replicates
 This is true even when total sequencing effort is constant!

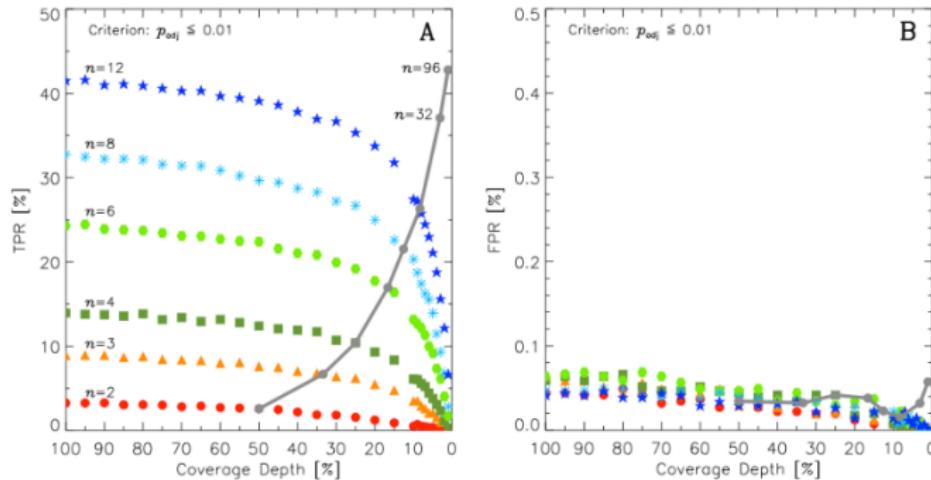


Figure 3 TPR and FPR detected by DESeq as a function of sequencing depth and replication. Different symbols represent the number n of control vs. treatment samples ($n = 2, 3, 4, 6, 8$, and 12) across sequence depths [100% → 1%]. **A:** TPR (Eq. 6 at $\alpha = 1\%$) $p_{adj} \leq 0.01$. **B:** FPR (Eq. 5 at $\alpha = 1\%$) $p_{adj} \leq 0.01$. The solid grey line ("multiplex line") connecting the TPR values of n biological replicates at $\frac{1}{n} \times 100\%$ sequencing depth shows the increase of TPR as more biological replicates n are used despite the loss power due to the sequencing depth reduction required by the multiplexing of lanes. This trend remains true even for the $n = 32$ and $n = 96$ cases.

Experimental design

Biological replicates

Reference genome?

Good gene annotation?

Read depth

Barcode

Read length

Paired vs. single-end



High Accuracy
Low Precision



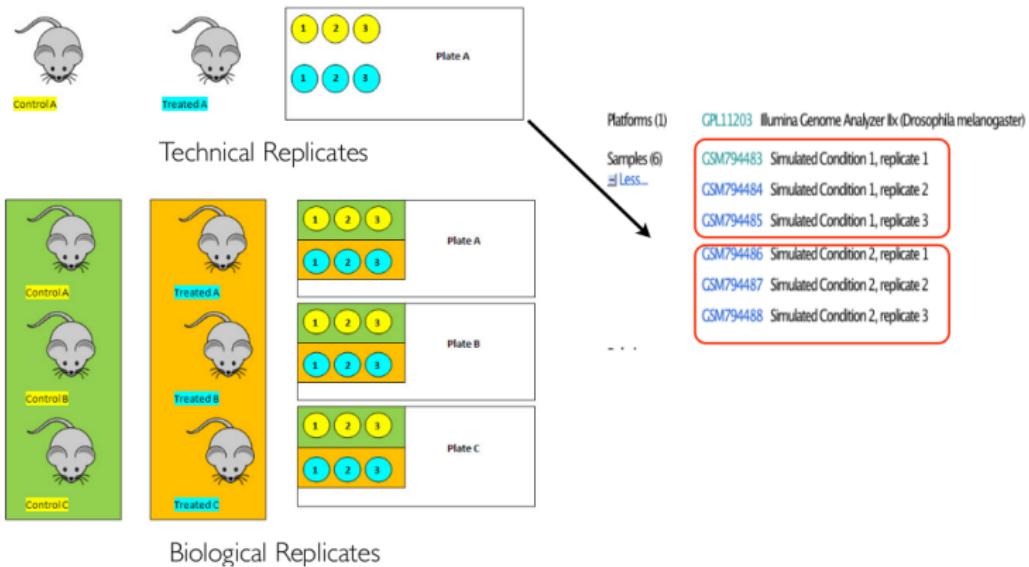
Low Accuracy
High Precision

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

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	1.1×10^{15}
50	1.3×10^{30}
100	1.6×10^{60}

~60 million (6×10^7) coding bases
in vertebrate genome

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



Experimental design

Biological replicates

Reference genome?

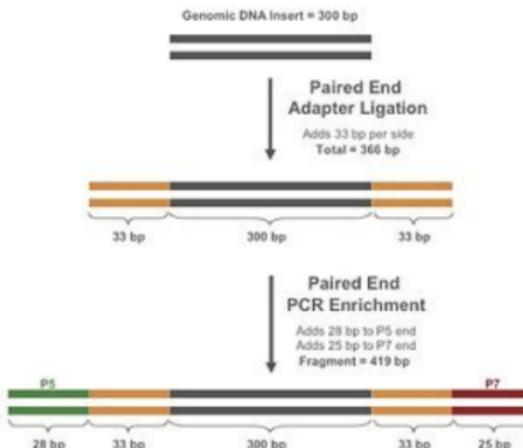
Good gene annotation?

Read depth

Barcode/Index

Read length

Paired vs. single-end



BROAD 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 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

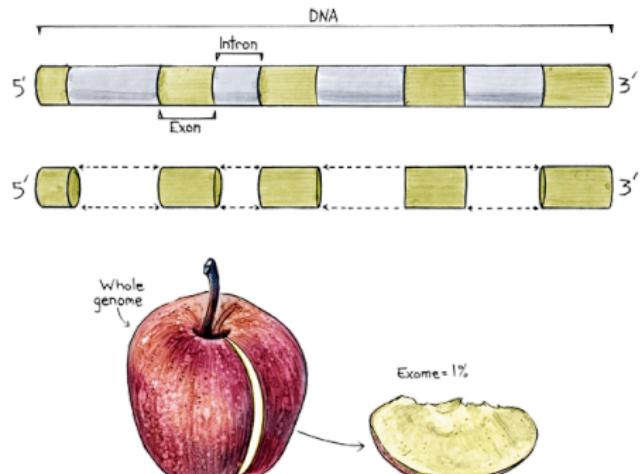
6 回顾与总结

- 总结
- 思考题



外显子组测序 | 简介 | Exome

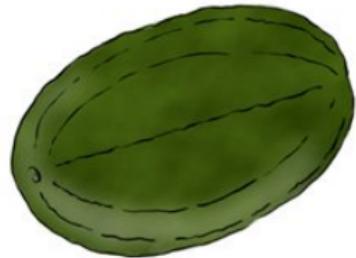
The exome is the part of the genome formed by exons, the sequences which when transcribed remain within the mature RNA after introns are removed by RNA splicing. It consists of all DNA that is transcribed into mature RNA in cells of any type as distinct from the transcriptome, which is the RNA that has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



外显子组测序 | 简介 | Exome

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.



What is WES?

- Whole Exome Sequencing
- Targets the protein coding regions of the human genome: ~180,000 exons in ~20,000 genes (30Mb)
- Coding regions of interest are targeted and “captured” for massively parallel sequencing (NextGen technology)
- Generates a huge amount of data which needs to first be filtered by bioinformatics and then analyzed by both bioinformatics and humans



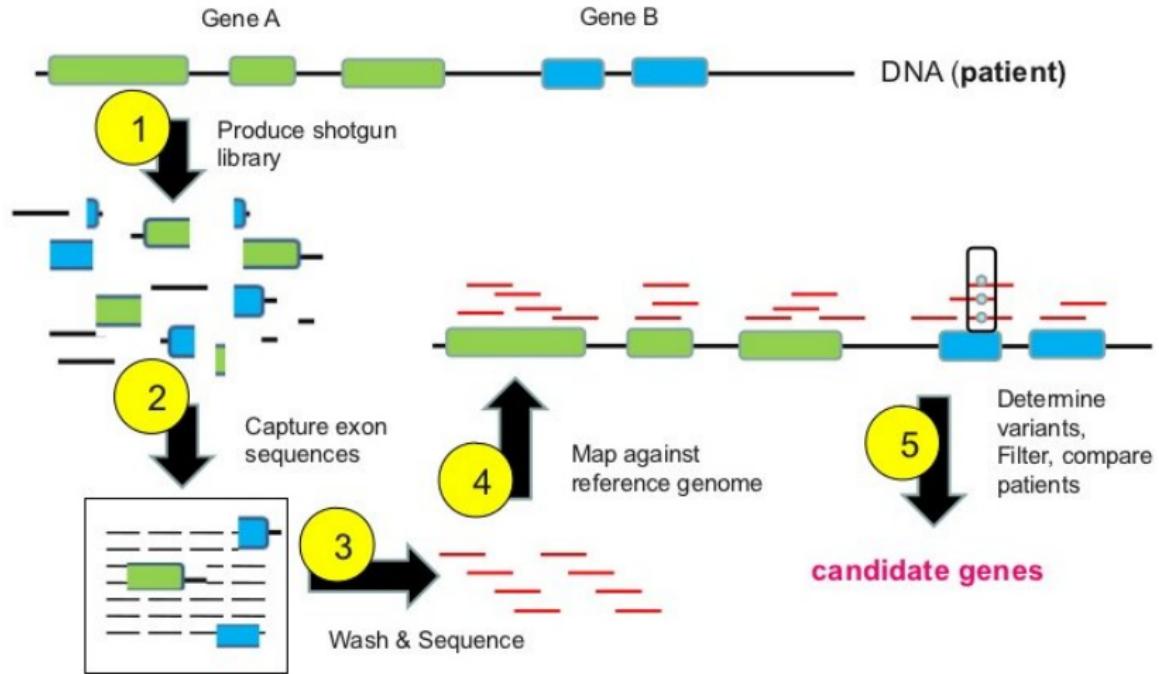
Why exome sequencing?

- 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

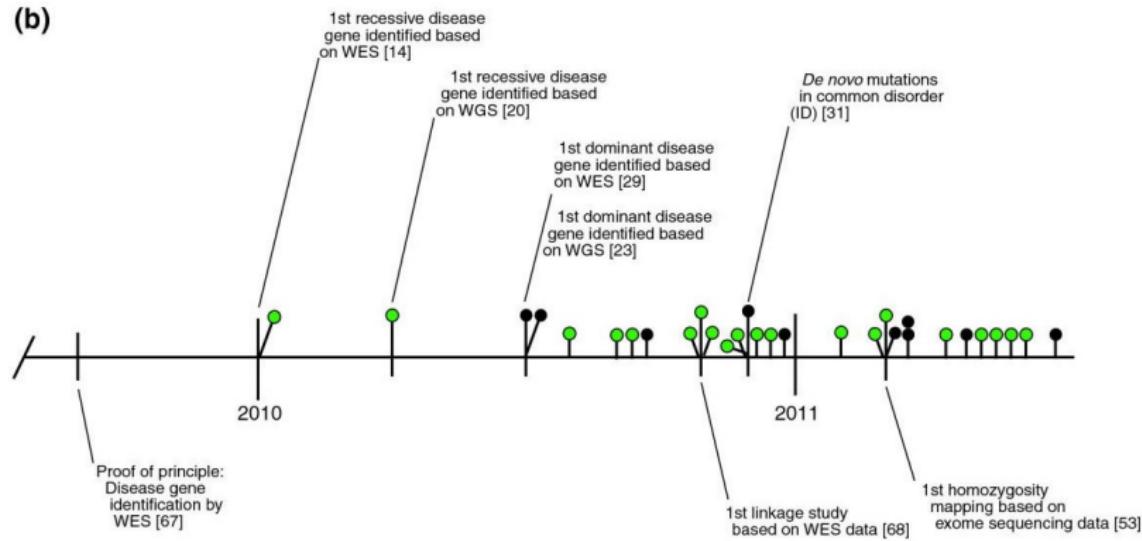


How does exome sequencing work?



外显子组测序 | 简介 | Exome sequencing

(b)



Key:

= Notable publication

= Recessive disease gene identification by WES

= Dominant disease gene identification by WES



WGS

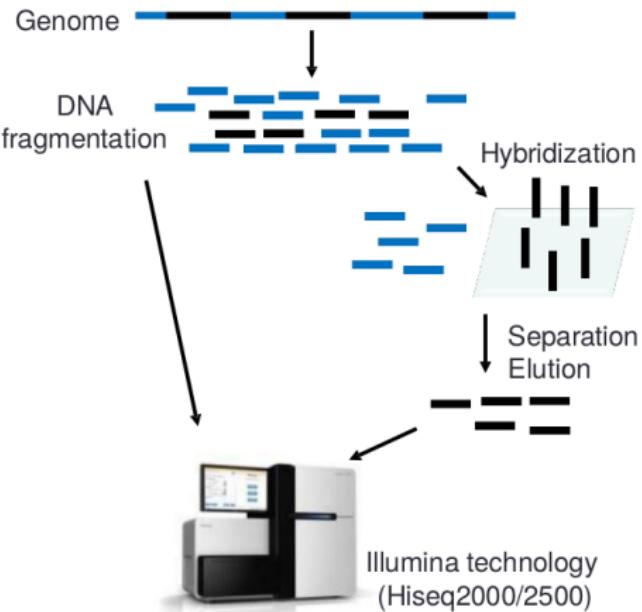
Whole genome sequencing (also known as WGS, full genome sequencing, complete genome sequencing, or entire genome sequencing) is a laboratory process that determines the complete DNA sequence of an organism's genome at a single time.

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



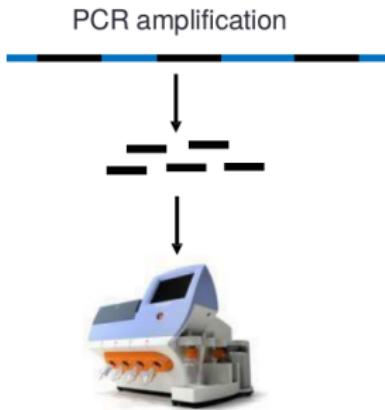
Genome or Target DNA sequencing

Whole Genome VS Exome Sequencing



Amplicon Sequencing

→ Sequencing of a dedicated panel of genes/hotspots



Mostly Ion Torrent technology
(PGM/Proton)



外显子组测序 | 简介 | Exome sequencing | Comparison

SNP array

- require hybridization probes of a known sequence
- can only detect shared genetic variants that are common to many individuals in the wider population

whole genome sequencing

high costs and time associated with sequencing large numbers of genomes



外显子组测序 | 简介 | Exome sequencing | Comparison

SNP array

- require hybridization probes of a known sequence
- can only detect shared genetic variants that are common to many individuals in the wider population

whole genome sequencing

high costs and time associated with sequencing large numbers of genomes



外显子组测序 | 简介 | Exome sequencing | Comparison

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



Techiniques

- Target-enrichment strategies
- PCR
- Molecular inversion probes (MIP)
- Hybrid capture
- In-solution capture
- Sequencing



Target-enrichment strategies

Target-enrichment methods allow one to selectively capture genomic regions of interest from a DNA sample prior to sequencing. Several target-enrichment strategies have been developed since the original description of the direct genomic selection (DGS) method by the Lovett group in 2005.



PCR

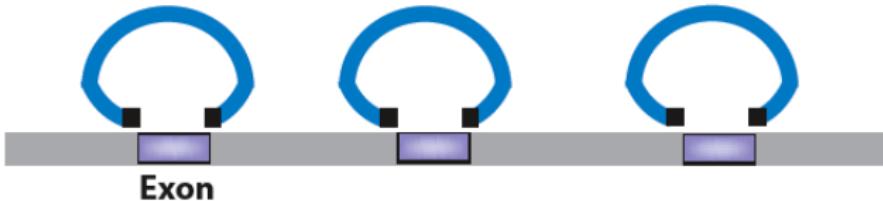
Polymerase chain reaction (PCR) is technology to amplify specific DNA sequences. It uses a single stranded piece of DNA as a start for DNA amplification. Uniplex PCR uses only one starting point (primer) for amplification and multiplex PCR uses multiple primers. This way multiple genes can be targeted simultaneously.



Molecular inversion probes (MIP)

Molecular inversion probe uses probes of single stranded DNA oligonucleotides flanked by target-specific ends. The gaps between the flanking sequences are filled and ligated to form a circular DNA fragment. Probes that did not undergo reaction remain linear and are removed using exonucleases. This is an enzymatic technique that targets the amplification of genomic regions by multiplexing based on target circularization.

Molecular Inversion Probes



Hybrid capture

Microarrays contain single-stranded oligonucleotides with sequences from the human genome to tile the region of interest fixed to the surface. Genomic DNA is sheared to form double-stranded fragments. The fragments undergo end-repair to produce blunt ends and adaptors with universal priming sequences are added. These fragments are hybridized to oligos on the microarray. Unhybridized fragments are washed away and the desired fragments are eluted. The fragments are then amplified using PCR.

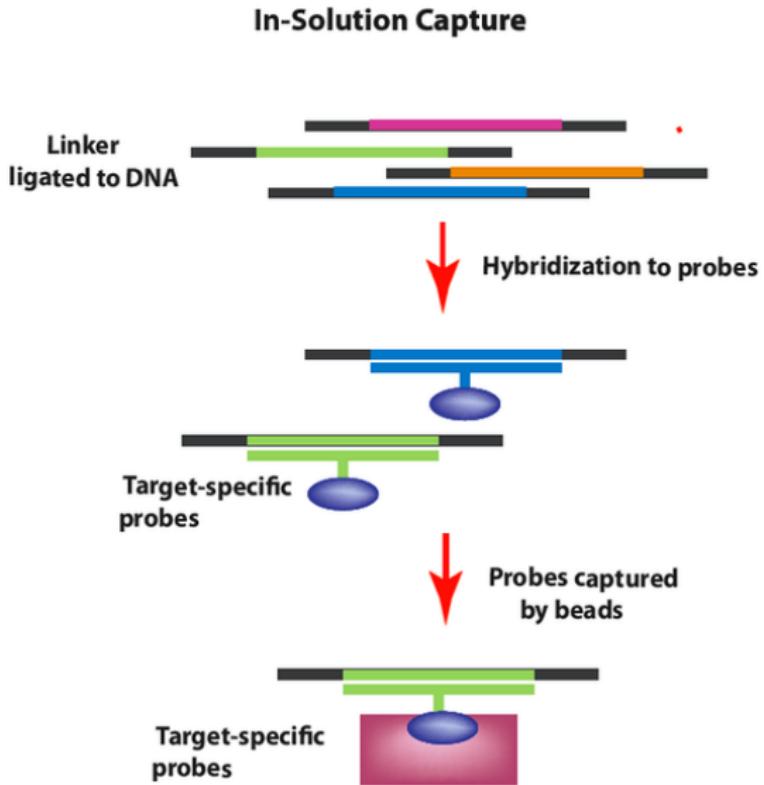


In-solution capture

To capture genomic regions of interest using in-solution capture, a pool of custom oligonucleotides (probes) is synthesized and hybridized in solution to a fragmented genomic DNA sample. The probes (labeled with beads) selectively hybridize to the genomic regions of interest after which the beads (now including the DNA fragments of interest) can be pulled down and washed to clear excess material. The beads are then removed and the genomic fragments can be sequenced allowing for selective DNA sequencing of genomic regions (e.g., exons) of interest.

This method was developed to improve on the hybridization capture target-enrichment method.





Sequencing

There are several sequencing platforms available including the classical Sanger sequencing. Other platforms include the Roche 454 sequencer, the Illumina Genome Analyzer II and the Life Technologies SOLiD & Ion Torrent all of which have been used for exome sequencing.



外显子组测序 | 简介 | Exome sequencing | Significance

Exome sequencing has the potential to locate causative genes in complex diseases, which previously has not been possible due to limitations in traditional methods.

Targeted capture and massively parallel sequencing represents a cost-effective, reproducible and robust strategy with high sensitivity and specificity to detect variants causing protein-coding changes in individual human genomes.



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.



Statistical analysis

The statistical analysis of the large quantity of data generated from sequencing approaches is a challenge. False positive and false negative findings are associated with genomic resequencing approaches and is a critical issue. A few strategies have been developed to improve the quality of exome data such as:

- Comparing the genetic variants identified between sequencing and array-based genotyping
- Comparing the coding SNPs to a whole genome sequenced individual with the disorder
- Comparing the coding SNPs with Sanger sequencing of HapMap individuals



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 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

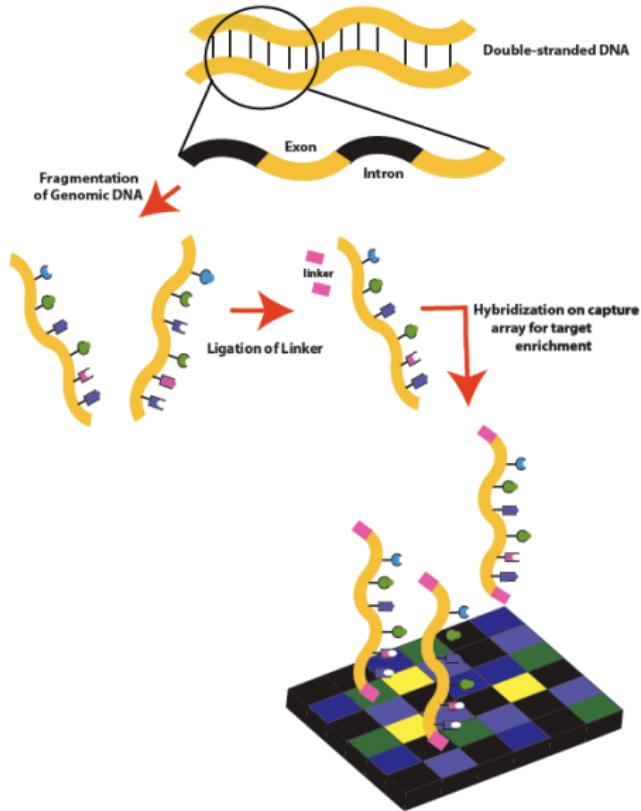
- 简介
- 操作流程
- 应用实例

6 回顾与总结

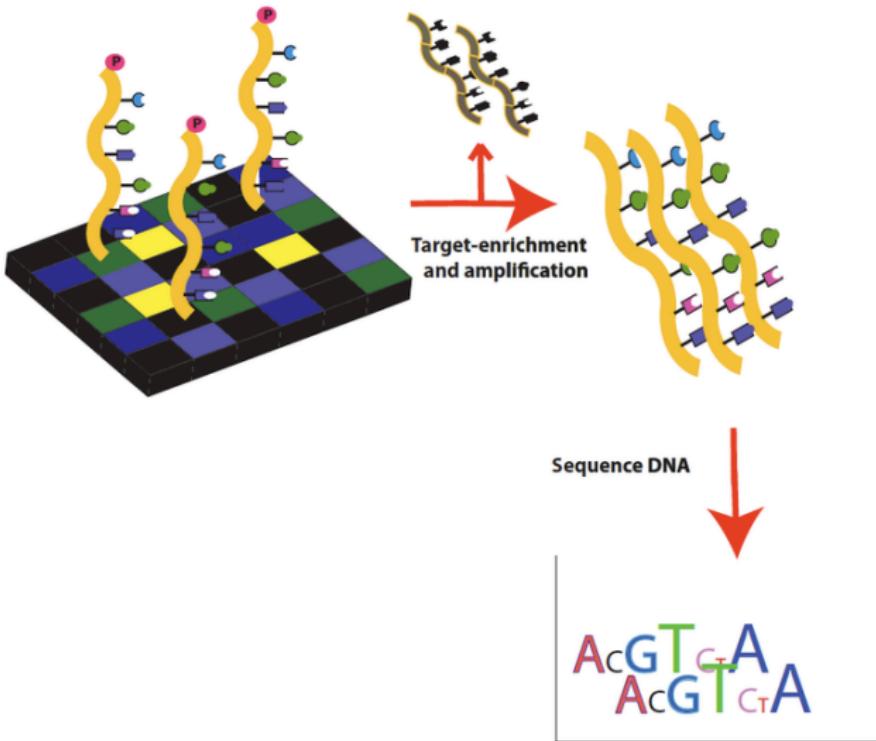
- 总结
- 思考题



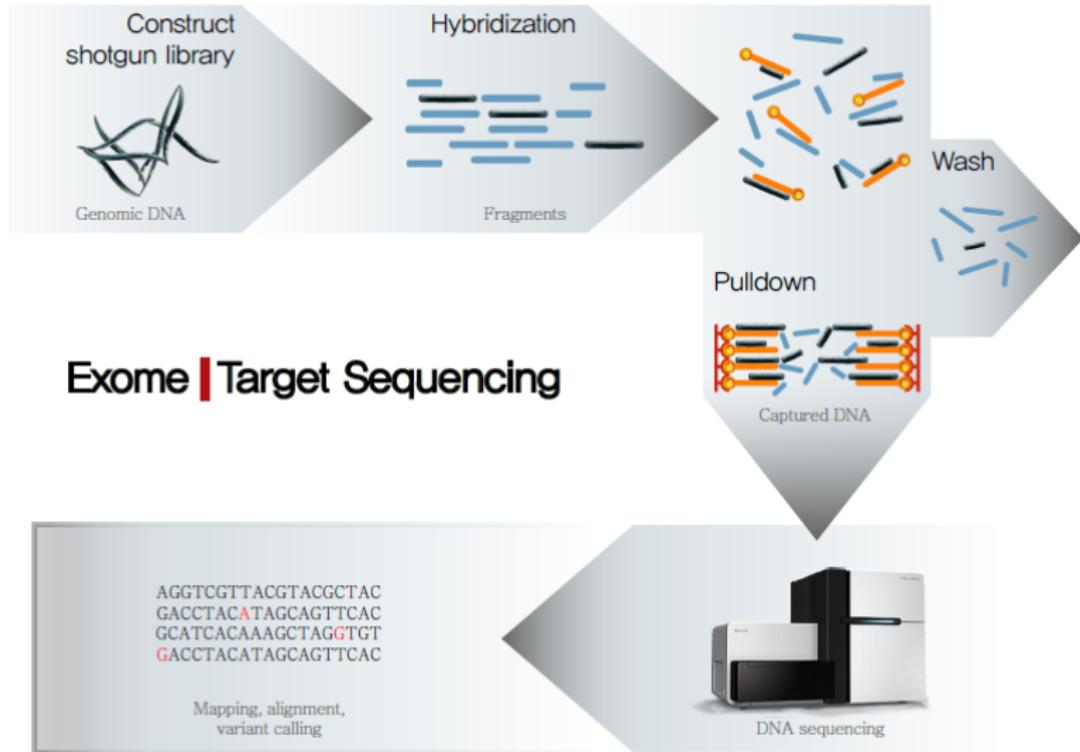
外显子组测序 | 流程 | 实验



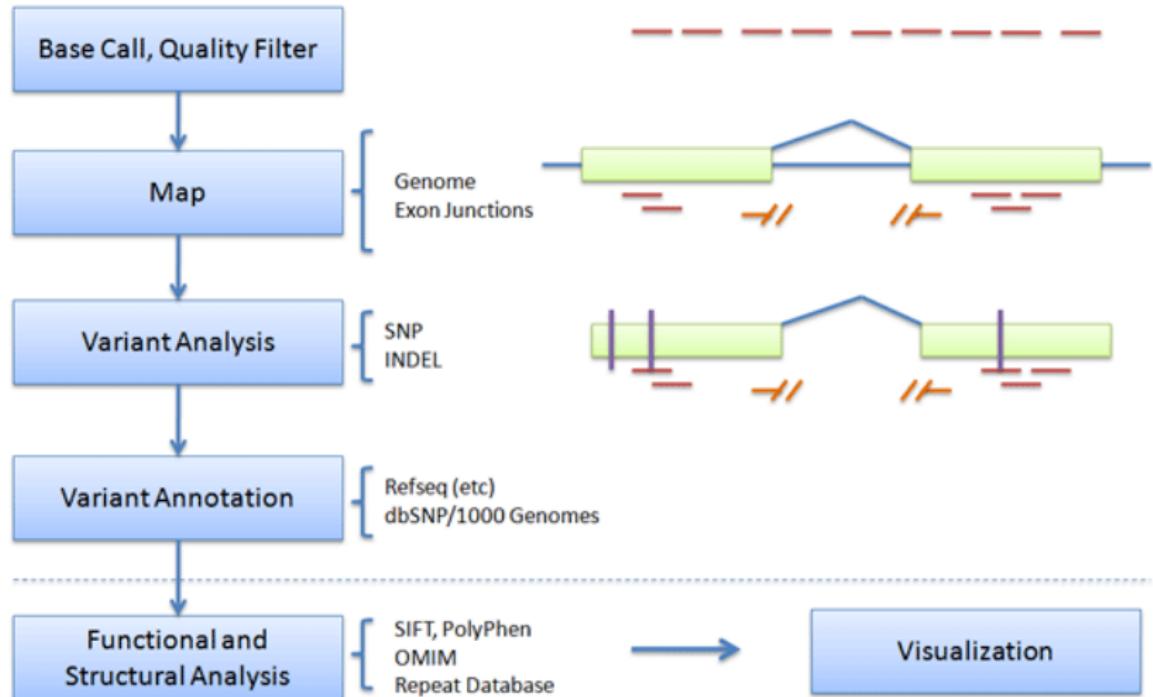
外显子组测序 | 流程 | 实验



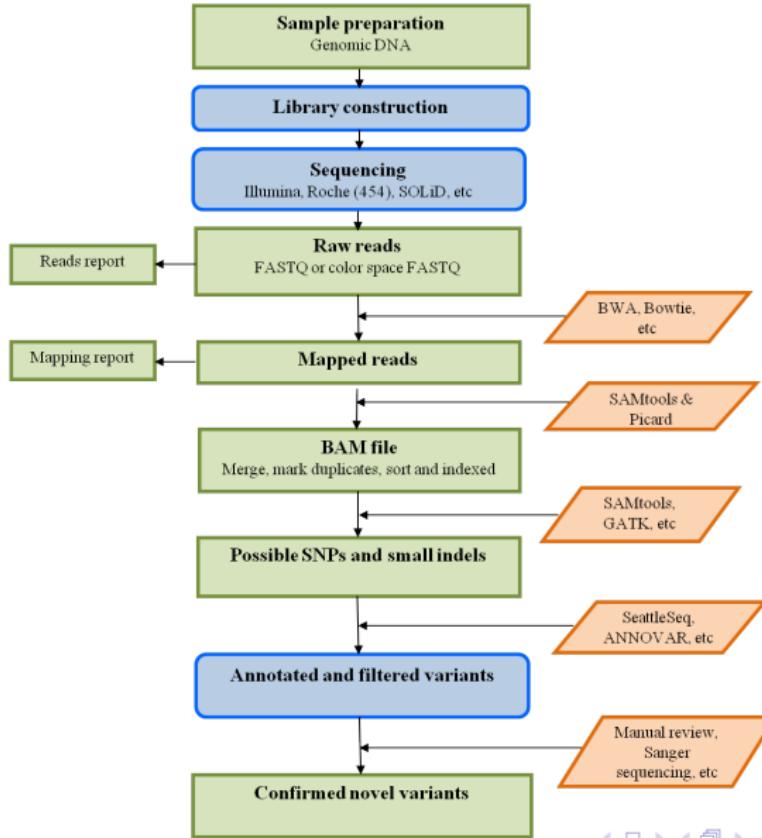
外显子组测序 | 流程 | 实验



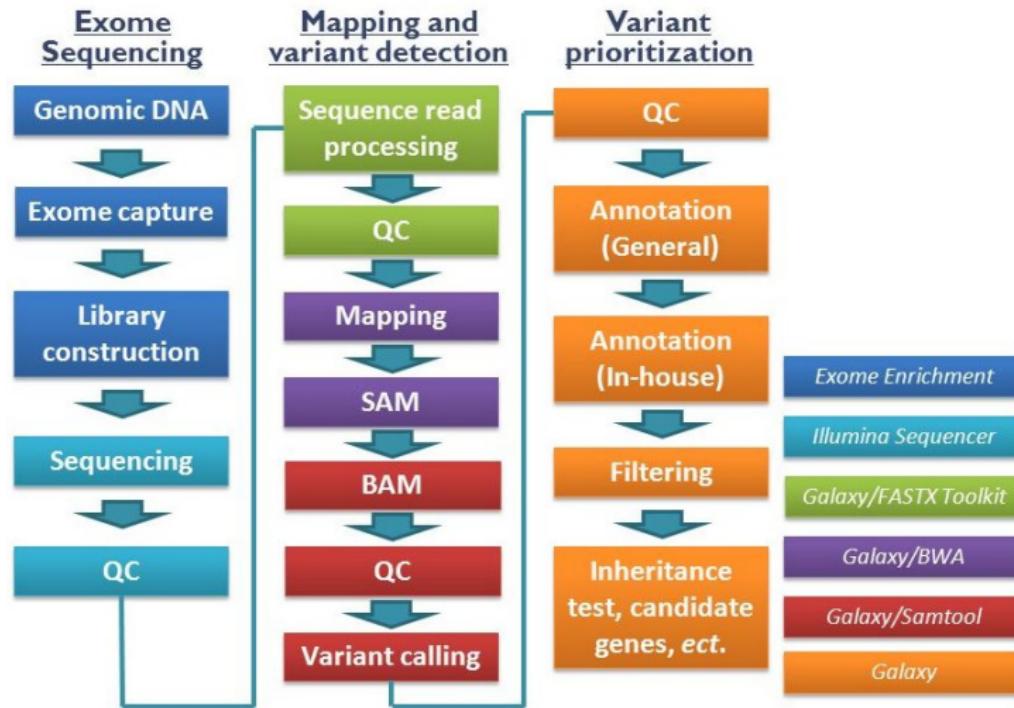
Edge Exome Analysis Pipeline



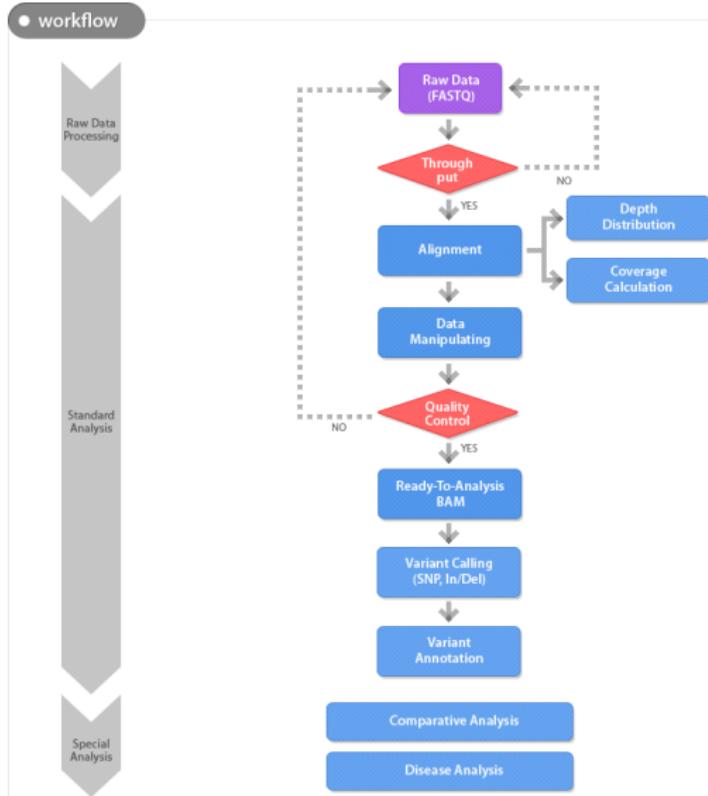
外显子组测序 | 流程 | 生信



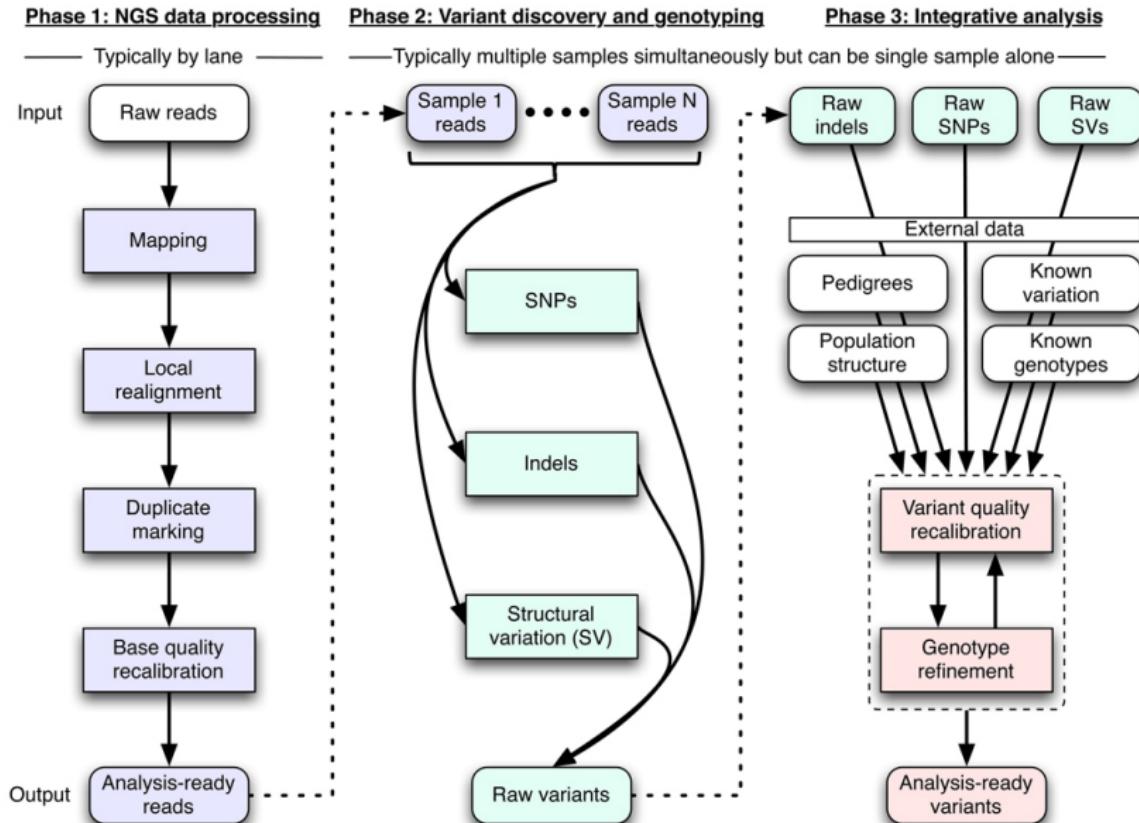
Exome NGS Workflow



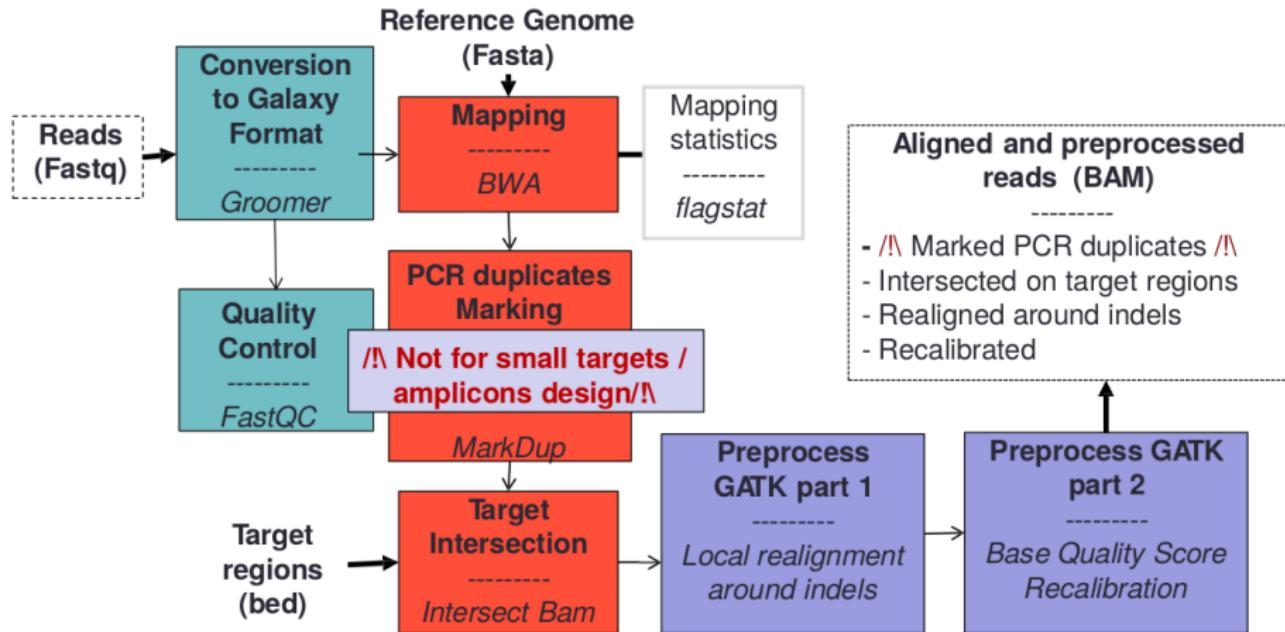
外显子组测序 | 流程 | 生信



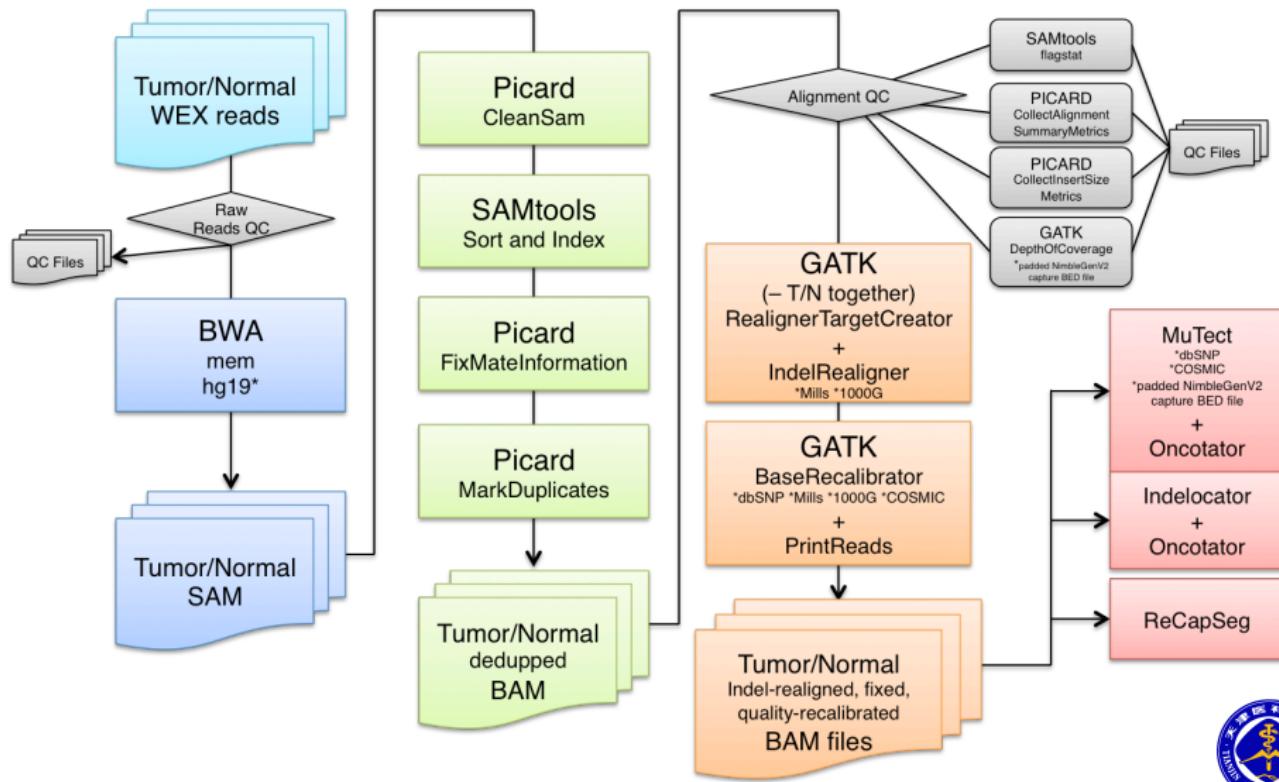
外显子组测序 | 流程 | 生信



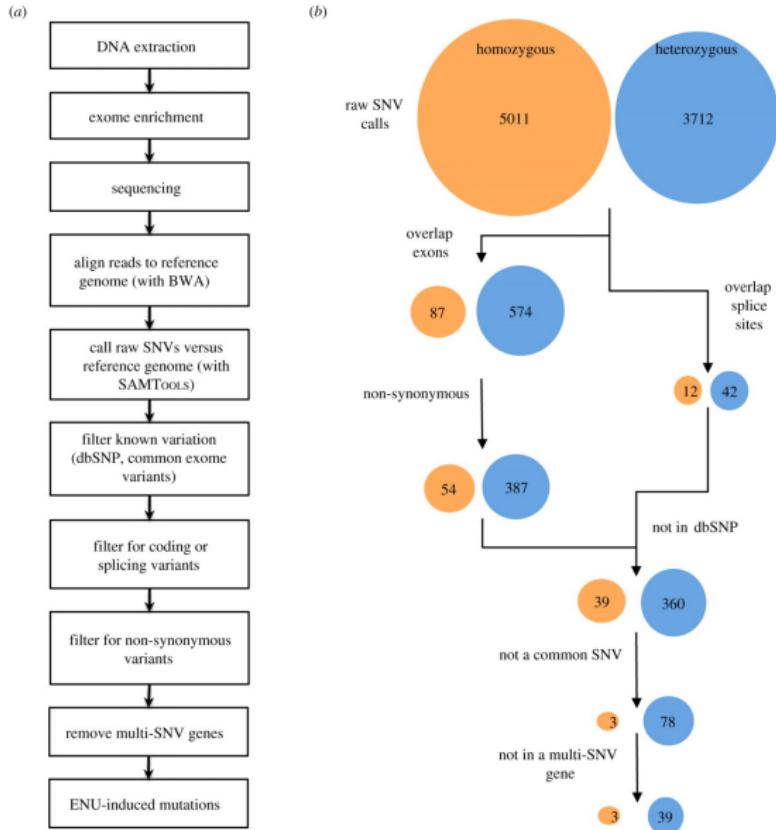
外显子组测序 | 流程 | 生信



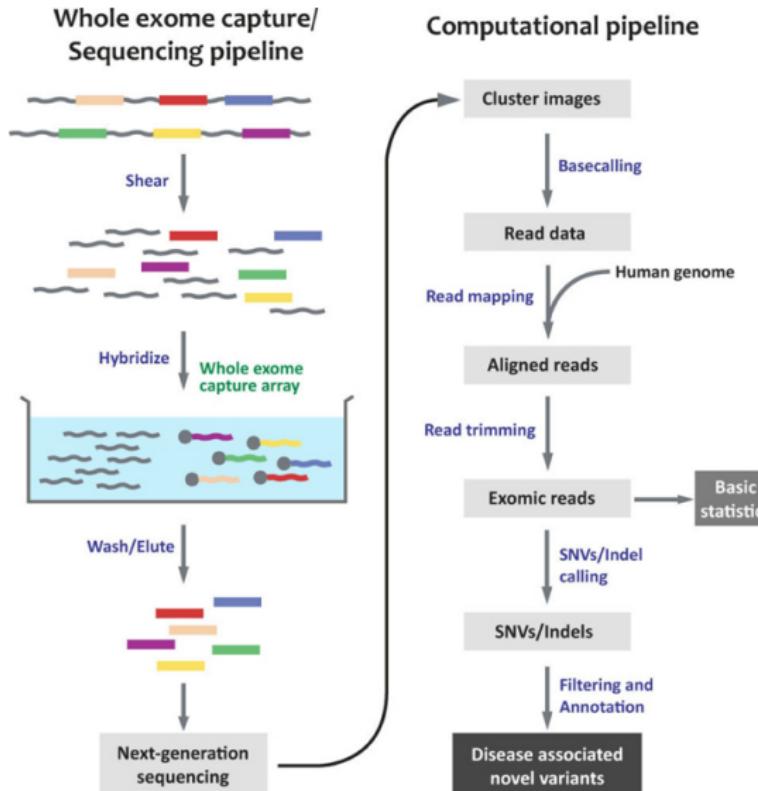
外显子组测序 | 流程 | 生信



外显子组测序 | 流程 | 生信



外显子组测序 | 流程 | 实验 & 生信



Your exome in numbers

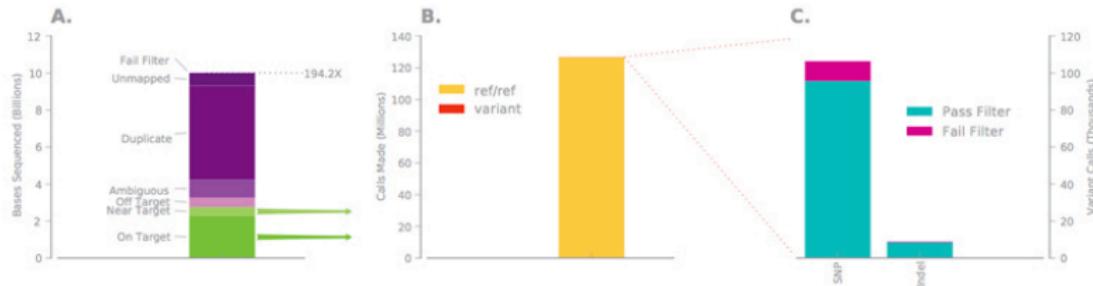


Figure 1: Getting from raw reads to called variants. A) The number of bases obtained by sequencing your exome. The top line indicates total coverage. B) Total number of called bases in your exome. The vast majority are the same as the reference genome. C) An expansion of the small sliver of variants depicted in B. These are the variants present in your VCF file.

Characterizing your variants

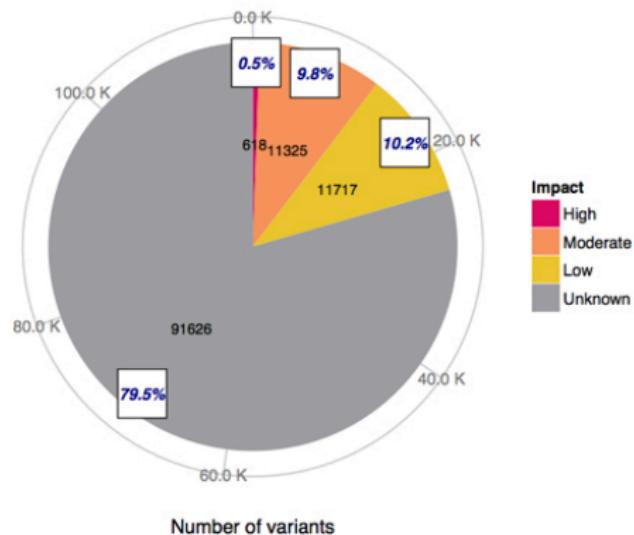
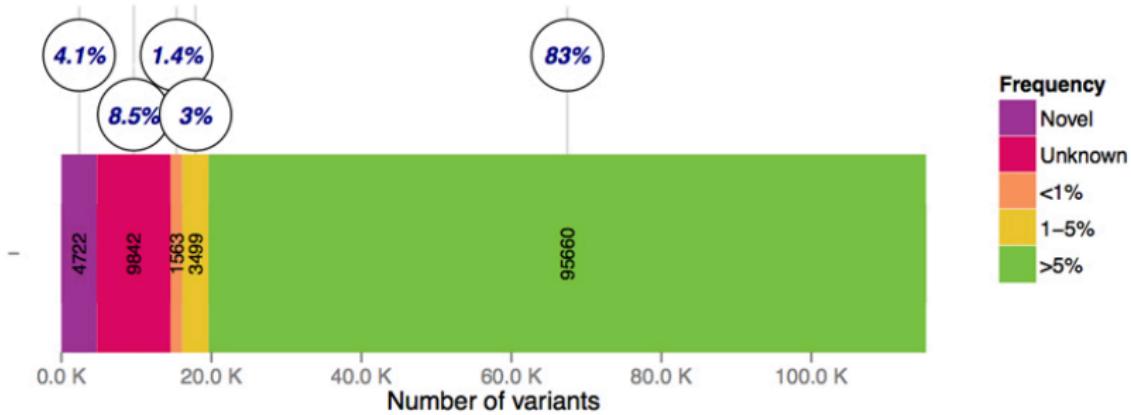


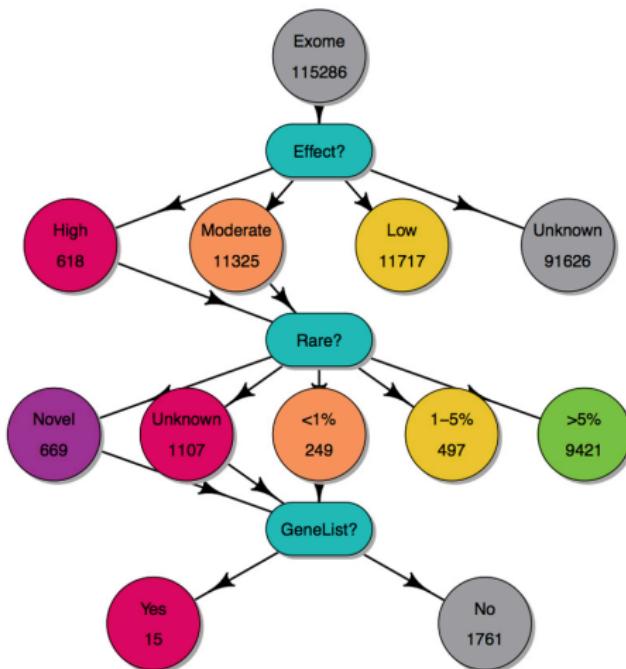
Figure 2: Predicting impact of variants on gene function. An overview of your variants and their predicted impact on gene function.



How rare are your variants?



Filtering your variants



外显子组测序 | 流程 | 实例

Symbol	Name	OMIM Link ¹	dbSNP ID ²	AA change (conservative?) ³	1K Genomes	Effect on Frequency ⁴ phenotype ⁵
BCKDHA	branched chain keto acid dehydrogenase E1, alpha polypeptide	608348	rs34442879	T122M (nonconservative)	0.00560	recessive
CHH23	cadherin-related 23	605516	rs41281338	E2588Q (conservative)	0.00670	recessive
EVC	Ellis van Creveld syndrome	604831	rs41269549	D184N (conservative)	9e-04	recessive



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



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.

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.



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.

hoi 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.



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.

Sarah B Ng; Kati J Buckingham; Choli Lee; Abigail W Bigham; Holly K



基因组学 | 外显子组测序 | 实例



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



知识点

- 基因组学：人类基因组计划，相关学科
- 测序技术：发展历史，三代测序技术的主要原理
- 外显子组测序：外显子组，实验步骤，主要应用
- 数据库：SRA、GEO
- 数据格式：FASTQ、SAM、BED、GFF、VCF
- 测序数据分析：常见术语，主要步骤，常用工具

技能

- 能够对测序数据进行质控和预处理
- 能够对测序数据进行完整分析
- 能够掌握常见测序数据分析软件的使用方法

教学提纲

1 基因组学概述

- 概述
- 人类基因组计划
- 分支学科

2 测序技术

- 第一代测序技术
- 第二代测序技术
- 第三代测序技术
- 测序技术比较

3 数据库与数据格式

- 数据库
- 数据格式

4 二代测序数据分析

- 常见术语
- 分析流程
- 补遗
 - 预处理
 - 比对后
 - 实验设计

5 外显子组测序

- 简介
- 操作流程
- 应用实例

6 回顾与总结

- 总结
- 思考题



- ① 根据自己的理解对人类基因组计划进行评价。
- ② 简述 Sanger 测序法的原理。
- ③ 列举第二代测序方法的主要技术。
- ④ 简述 Illumina/Solexa 测序的基本过程。
- ⑤ 列举第三测序方法的主要技术。
- ⑥ 简述外显子组测序的流程和应用。
- ⑦ 根据实例解释 FASTQ 格式。
- ⑧ 根据实例解释 BED、GFF 和 VCF 格式。
- ⑨ 解释测序深度和覆盖度。
- ⑩ 简述测序数据分析的主要步骤。
- ⑪ 列举测序数据分析的常用工具并进行简介。



- 回顾 DNA 测序的实验方法和数据分析步骤。
- 回顾表达芯片的实验过程和数据分析步骤。



Powered by



T_EX L^AT_EX X_ET_EX Beamer