基因测序技术的发展及应用

徐俊美

华大集团 猛犸教育

01 测序技术发展和测序原理

什么是基因测序?

什么是基因测序?

基因全序列分析

基因测序是一种通过分析血液或唾液样本，测定个体基因全序列的技术，能够揭示基因中蕴含的遗传信息，为疾病预测和个性化医疗提供科学依据。

病变基因锁定

通过基因测序，可以精准锁定个体的病变基因，提前发现潜在的健康风险，从而采取预防措施或早期治疗，降低疾病发生的可能性。

行为特征预测

基因测序不仅能预测疾病风险，还能分析个体的行为特征，如性格倾向、运动能力等，为个性化教育和职业规划提供参考。

发展历程 History

跟踪-跟随 Following
参与-接轨 Participating
同步-超越 Exceeding
跨越-引领 Leading

1990年 1998年 1999年 2003年 2006年 2007年 2014年 2018年

1999年9月9日 北京华大基因研究中心成立
BGI was founded on September 9, 1999 in Beijing

杭州 华大 Hangzhou BGI

2007年6月深圳华大基因研究院成立
BGI Shenzhen was founded on June 19, 2007

华大全球总部（建设中）
BGI new headquarter (under construction)

人类基因组计划在美国启动
Human Genome Project (HGP) was launched in the US

宋健题词 Inscription from Song Jian

电泳法 Electrophoresis Sequencing

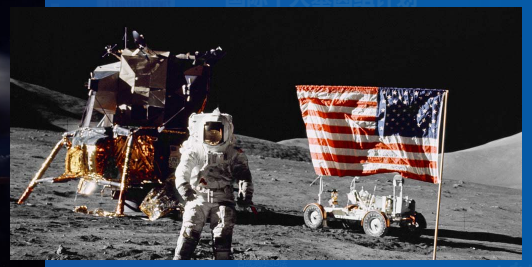
Cost $3B/Genome
Data output 102-103 bp/run

20世纪人类三大大科学工程

曼哈顿原子弹工程
Manhattan Project
1945USA

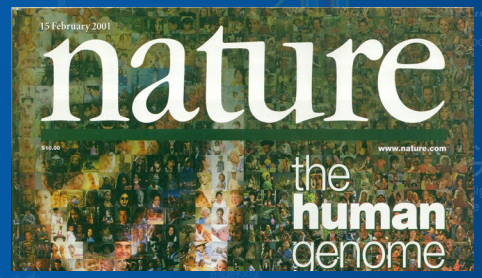
中国两弹一星
1960年-1970年

阿波罗登月计划
The Apollo Program
1965 USA



中国神舟五号载人飞船发射成功
2003年

人类基因组计划
1990-2003年 由美、英、法、德、日、中等6国共同参与
The Human Genome Project
1990-2003 US, UK, France, Germany, Japan, China

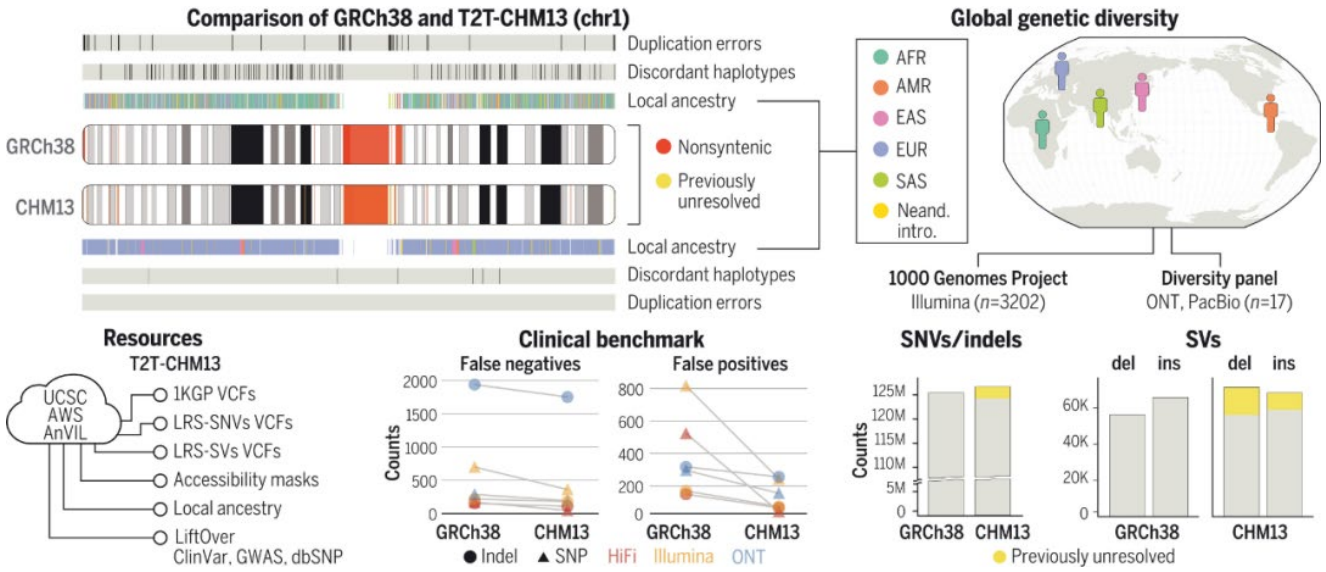


中国科学家承担人类基因组计划的1%测序任务
The Chinese scientists contributed to 1% of the Human Genome Project

人类基因组计划（Human Genome Progect）

- ✓人类科学史三大工程之一
- ✓1990年，国家人类基因组研究中心，准备投入30亿美元用15年完成对人类基因组测序
- ✓1999年，中国加入，承担1%（3000万碱基对，3号染色体）测序任务
- ✓2001年2月，人类基因组计划”（HGP）首次对人体90%以上的DNA碱基对完成了测序
- ✓2022年4月，Science连发6篇文章，宣布人类完整基因组测序计划正式完成

2001.2



从GRCH38到T2T-CHM13

2022年4月

6 国家

16 研究中心

1,100 科学家

3,000,000,000 美元经费

23Gb 数据

伦理、法律、社会影响研究
Ethical, Legal and Social Implications (ELSI)

人类基因组计划 (HGP): 1990-2003

基因测序技术的发展历程

01

第一代测序技术

1977年，Sanger测序法的诞生标志着基因测序技术的开端，其基于链终止原理，虽然精度高但通量低，主要用于小规模基因研究。

02

高通量测序技术

21世纪初，第二代测序技术的出现大幅提高了测序通量和效率，如Illumina平台，使得全基因组测序成为可能，推动了基因组学研究的快速发展。

03

单分子测序技术

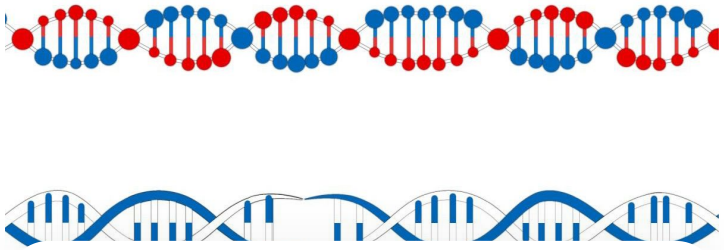
第三代测序技术如PacBio和Oxford Nanopore，实现了单分子实时测序，无需PCR扩增，能够直接读取长片段DNA，提高了测序的准确性和应用范围。

04

临床应用的普及

随着技术的成熟和成本的降低，基因测序从实验室研究逐步走向临床应用，成为疾病诊断、个性化治疗和健康管理的重要工具。

基因测序的基本原理



DNA片段化与扩增

基因测序的第一步是将DNA片段化，然后通过聚合酶链式反应（PCR）技术扩增特定基因片段，以便后续测序分析。

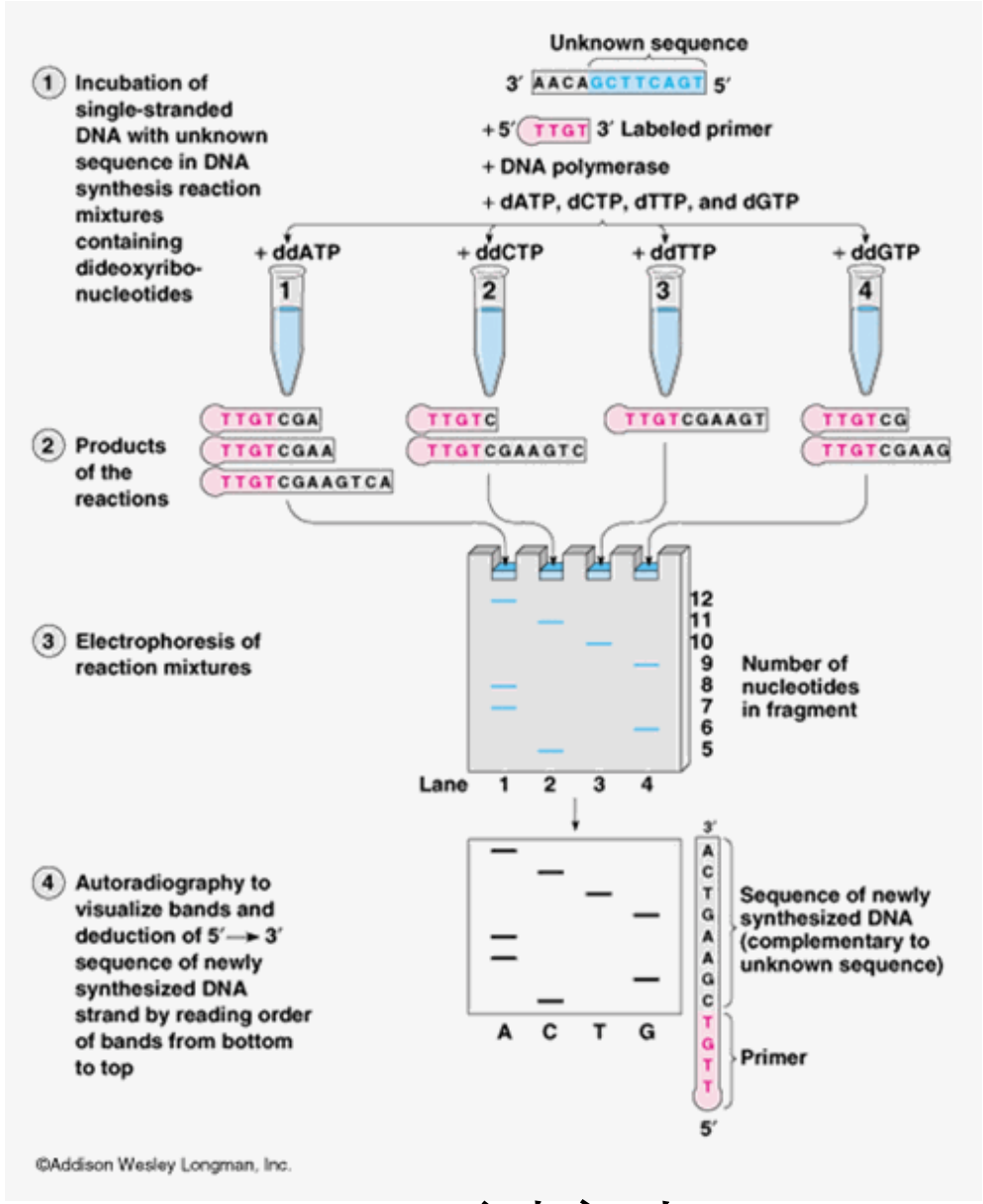
测序平台与检测

利用高通量测序平台，对扩增后的DNA片段进行测序，通过荧光标记或电化学信号检测碱基序列，生成原始的基因数据。

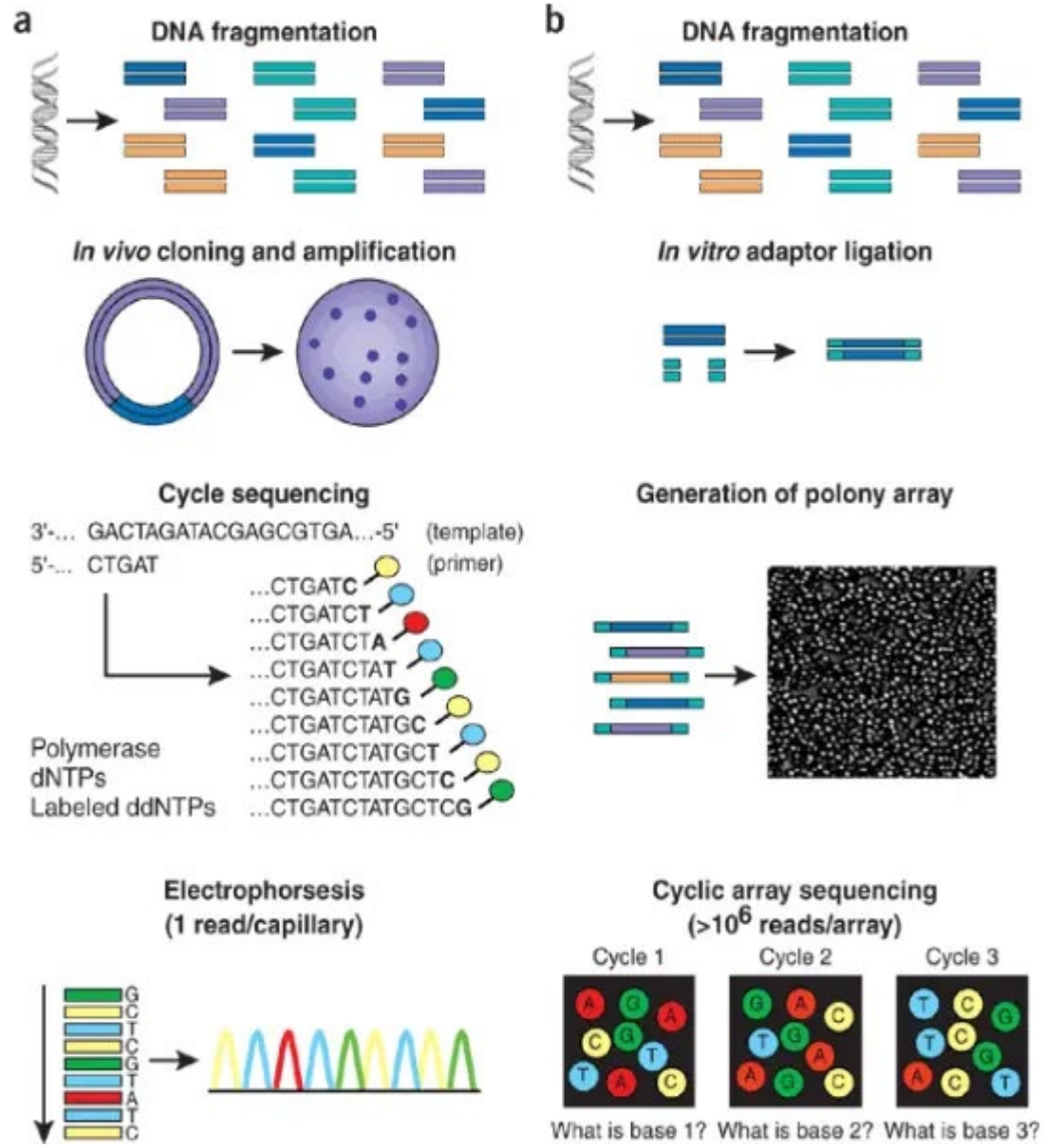
数据分析与解读

通过生物信息学工具对测序数据进行比对、组装和注释，解读基因序列中的突变、多态性等信息，最终生成个体基因报告。

一代&二代测序技术发展

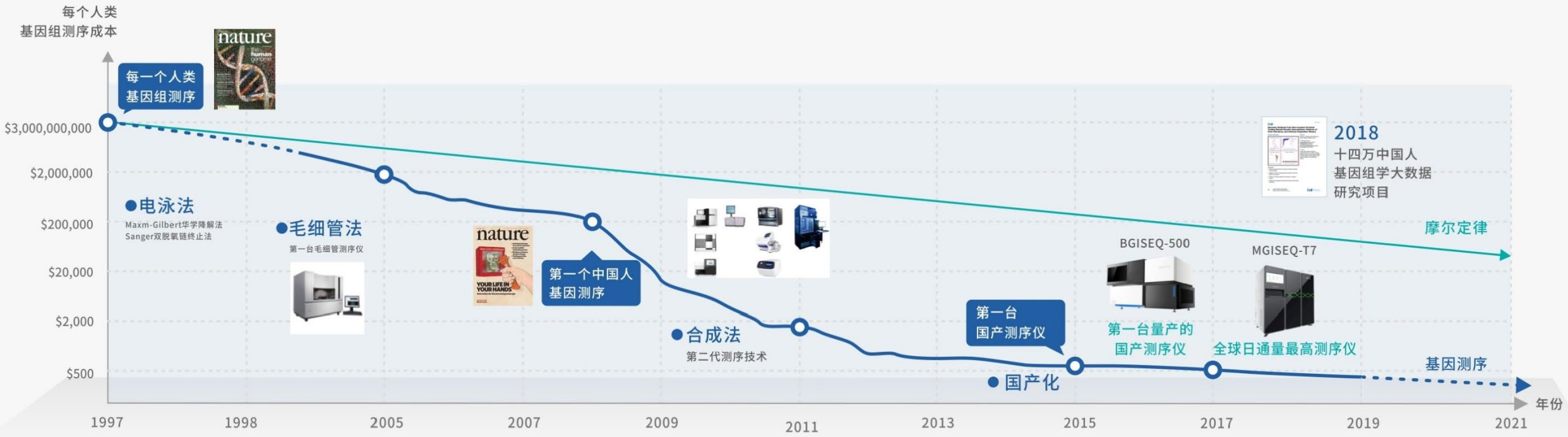


Sanger 测序法



二代测序

工具升级推动成本从数十亿下降至千元级



● 电泳法

Sanger测序平板法

数据通量:10^2-10^3碱基/运行周期

$30亿/基因组

● 毛细管法

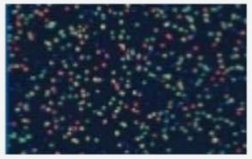
Sanger测序毛细管法

数据通量:10^3-10^6碱基/运行周期

$30亿/基因组

BT+IT规模化

● 合成法



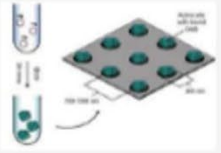
合成测序3D法

数据通量:10^9-10^{12}碱基/运行周期

$百万~万/基因组

BT+IT规模化大科学应用

● 国产化



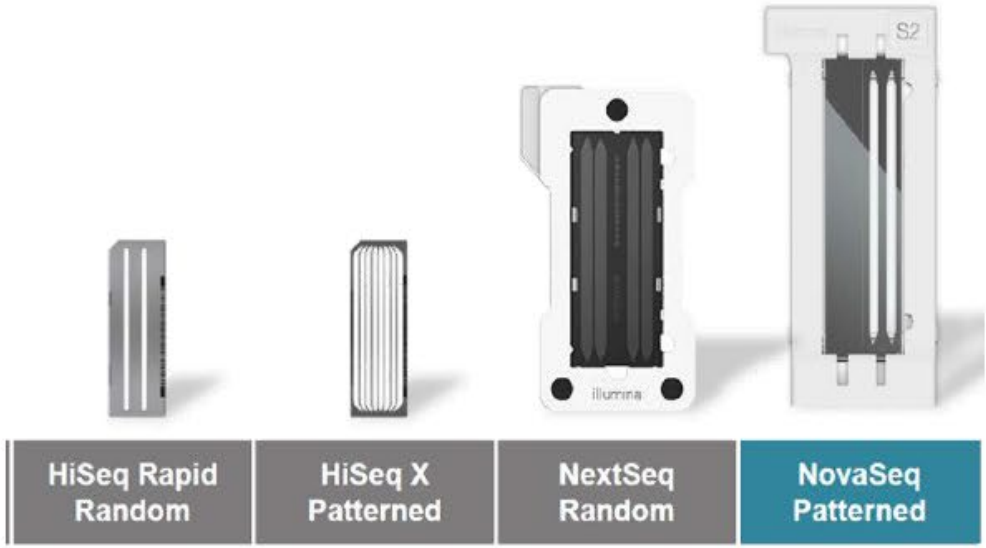
合成测序3D法

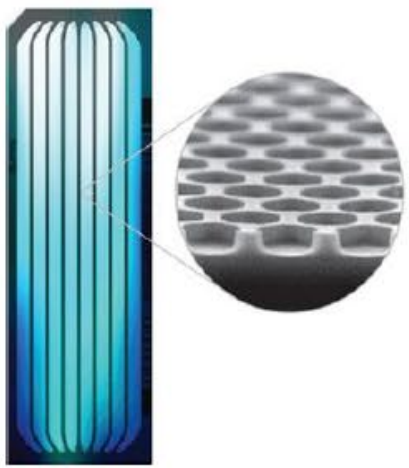
数据通量:10^{15}-10^{18}碱基/运行周期

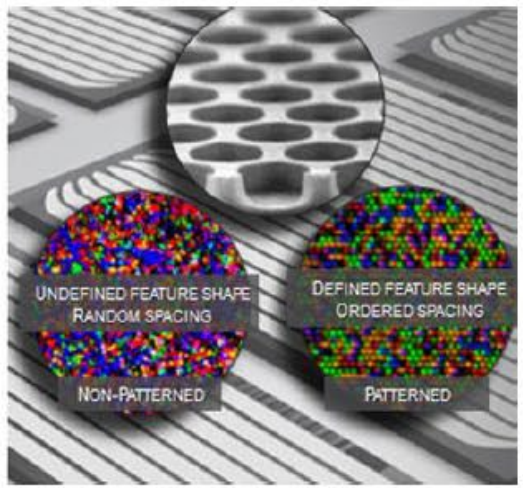
￥万~千/基因组

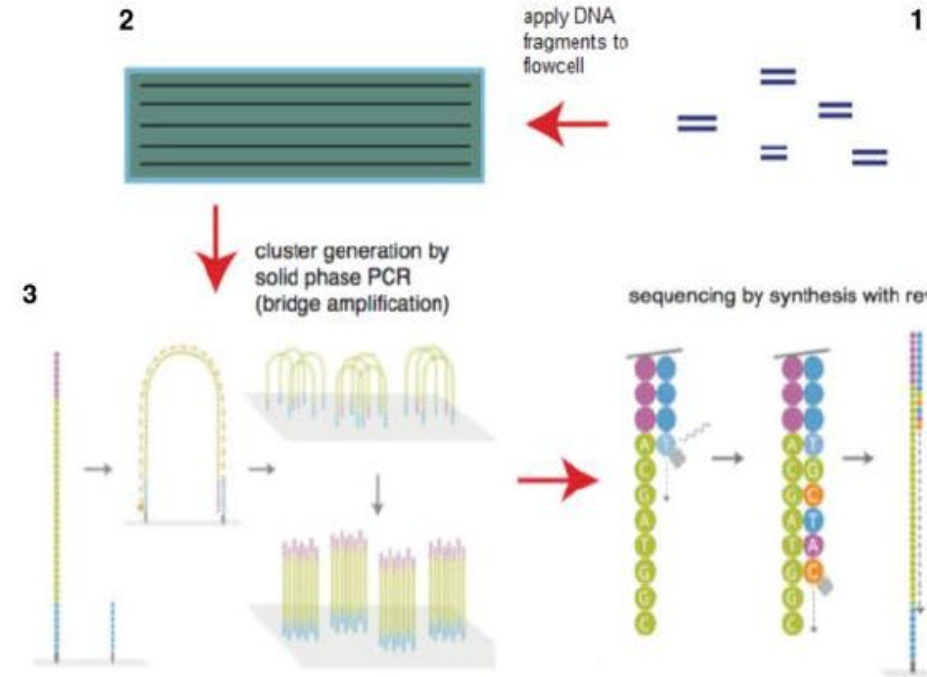
BT+IT规模化大科学、大产业应用

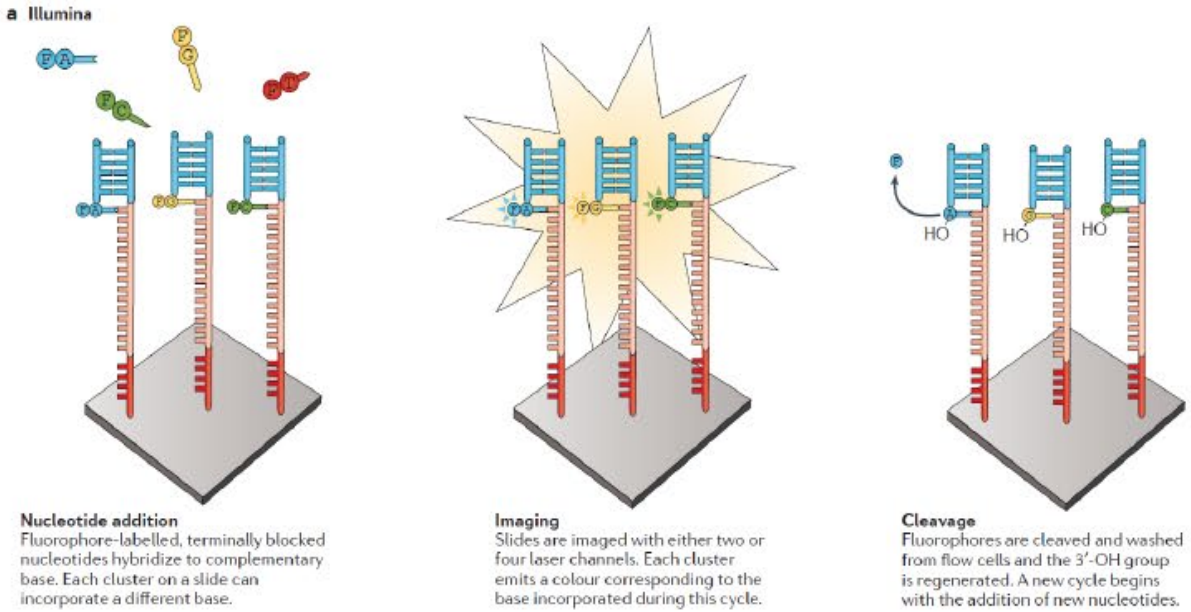
Illumina——桥式扩增SBS



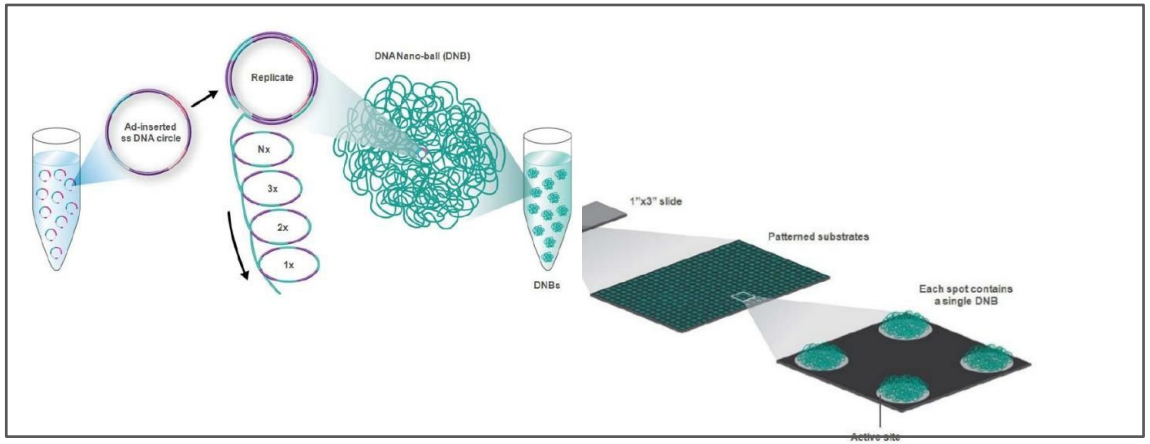


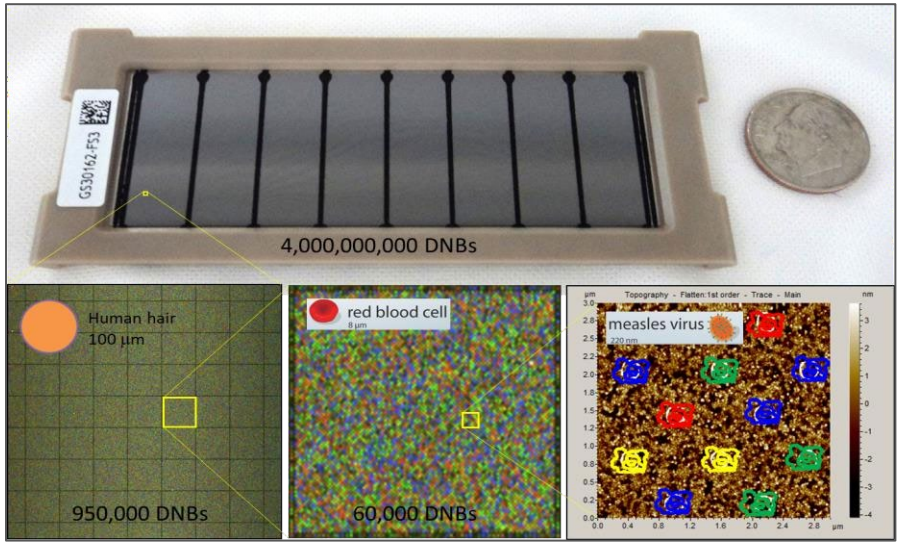


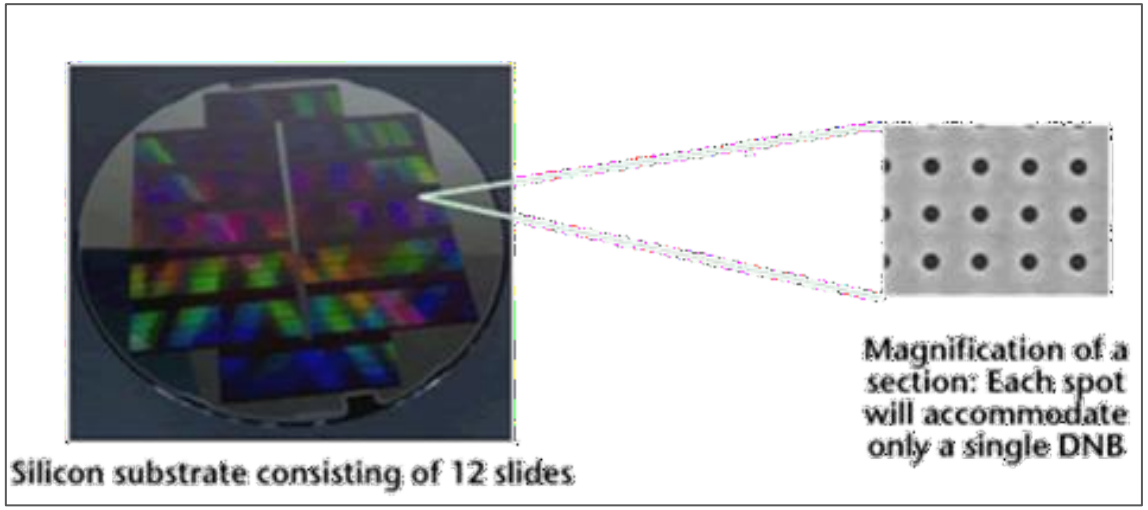


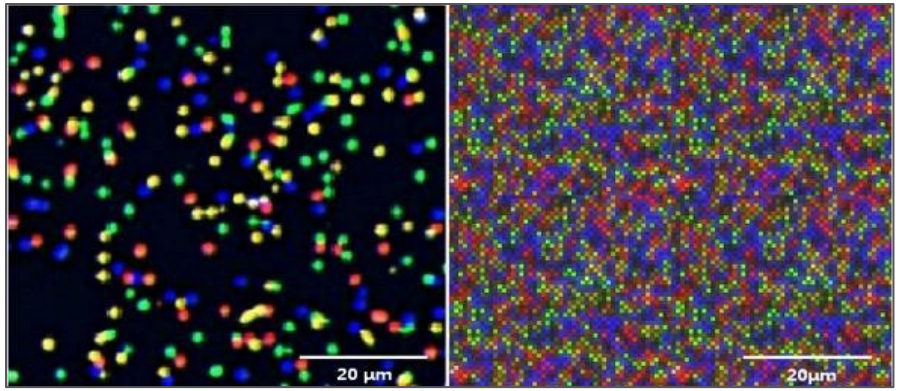


MGI—DNB测序技术


Ad-inserted ss DNA circle
Replicate
Nx
3x
2x
1x
DNA Nano-ball (DNB)
DNBs
1"x3" slide
Patterned substrates
Each spot contains a single DNB
Active site

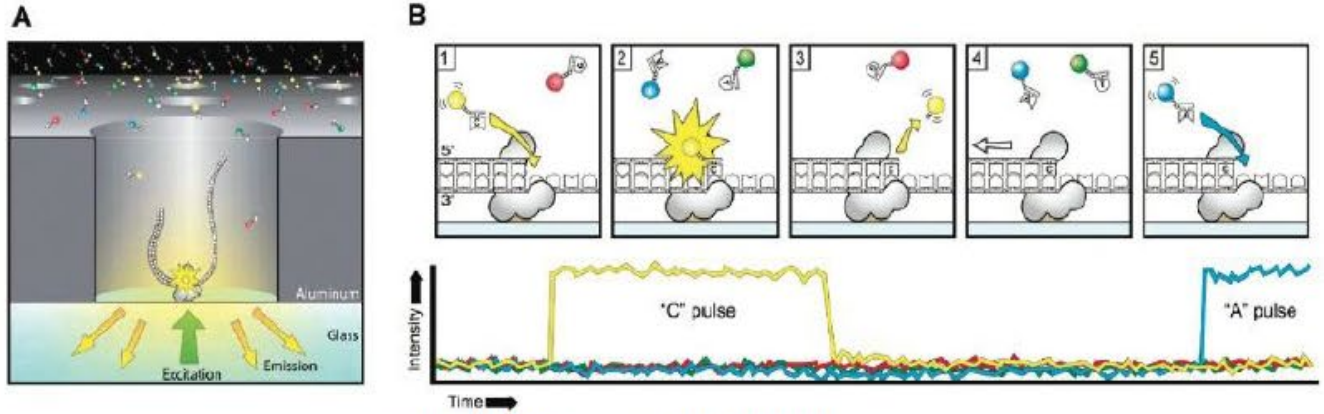

4,000,000,000 DNBs
Human hair
100 µm
950,000 DNBs
red blood cell
8 µm
60,000 DNBs
measles virus
220 nm

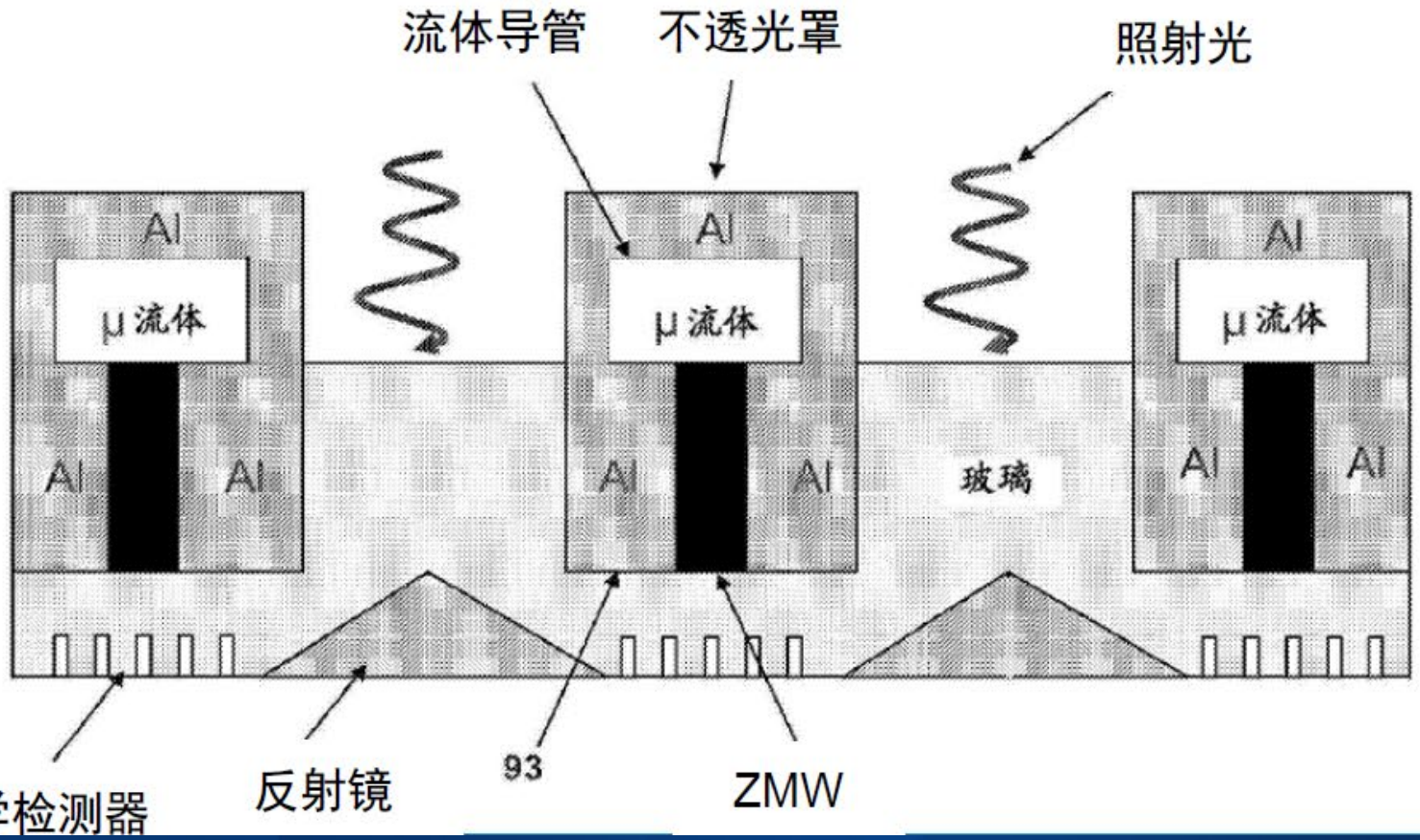

Silicon substrate consisting of 12 slides
Magnification of a section: Each spot will accommodate only a single DNB


20 µm
20µm

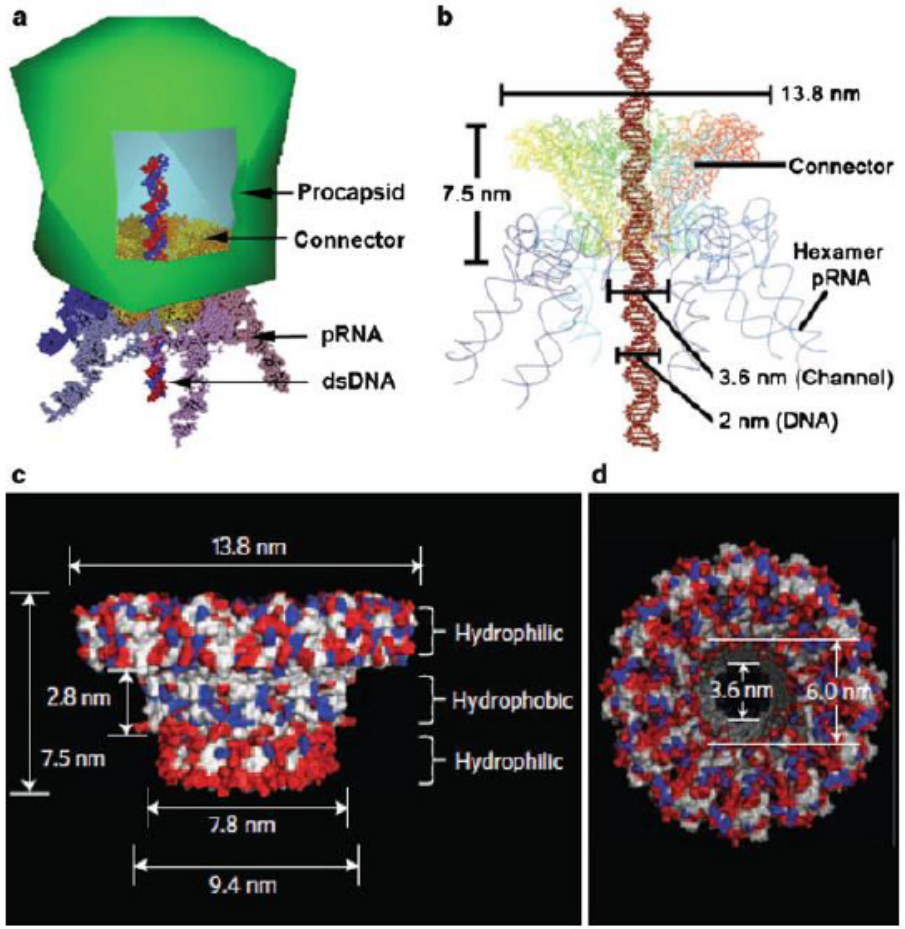
研究成果学术论文 Scientific Accomplishments

Pacbio——ZMW单分子测序

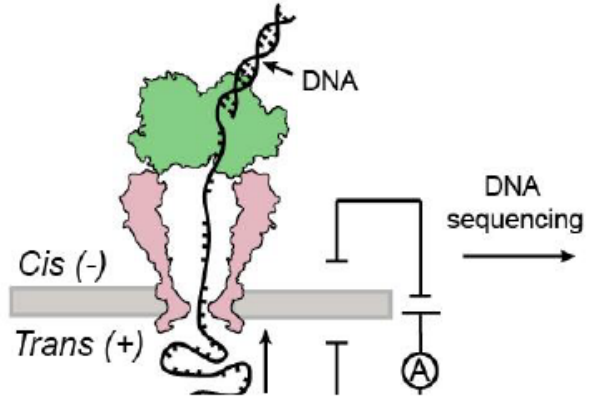


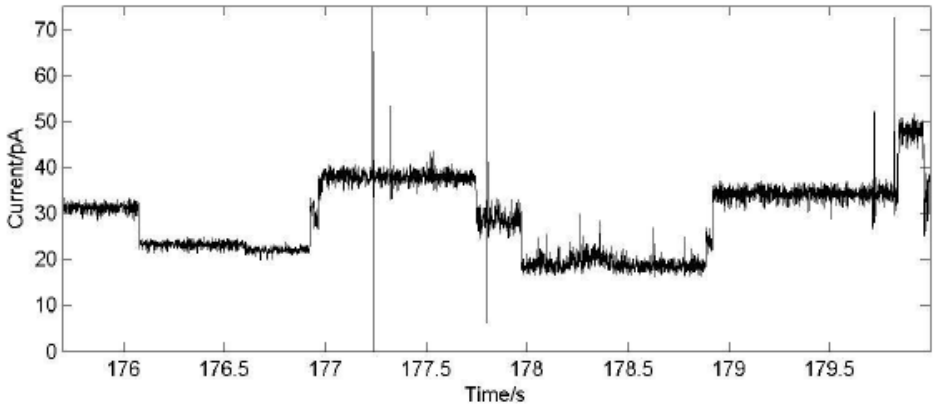


ONT——纳米孔测序



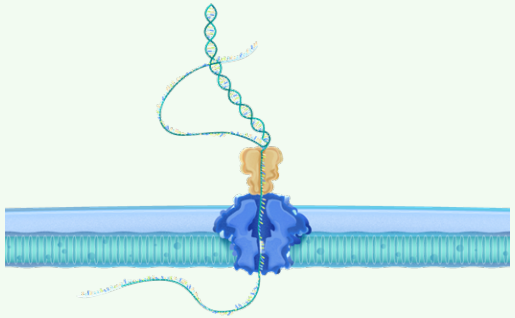
纳米孔技术起源于1996年，在一个典型的纳米孔测序实验中，纳米孔（粉色）是磷脂膜（灰色）两侧离子通过的唯一通道。测序酶（绿色）充当DNA的马达蛋白，拉动DNA链使其以单个核苷酸的步长依次通过纳米孔，每当一个核苷酸穿过纳米孔，相应的堵孔信号会被记录下来。通过分析这些序列相关的电流信号，我们可以反推出DNA的序列。





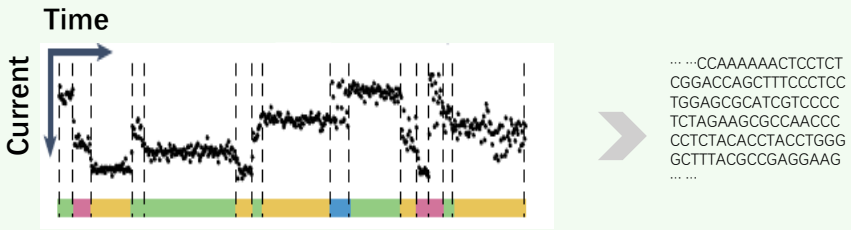
CycloneSEQ-WT02测序

流程



基于**纳米孔测序**原理，核酸分子通过电压引导穿过孔道蛋白，‘过孔’时引起不同的电流变化，因此识别碱基

测序信号



通过**电流信号**的变化经深度神经网络算法实时识别碱基排列信息

数据类型

以**单条read**为数据单元，测序时间不固定，run内的reads读长不完全一致，与**样品长度**直接相关

实时测序，灵活方便，读长全覆盖

— 小孔大世界 —

纳米孔基因测序仪

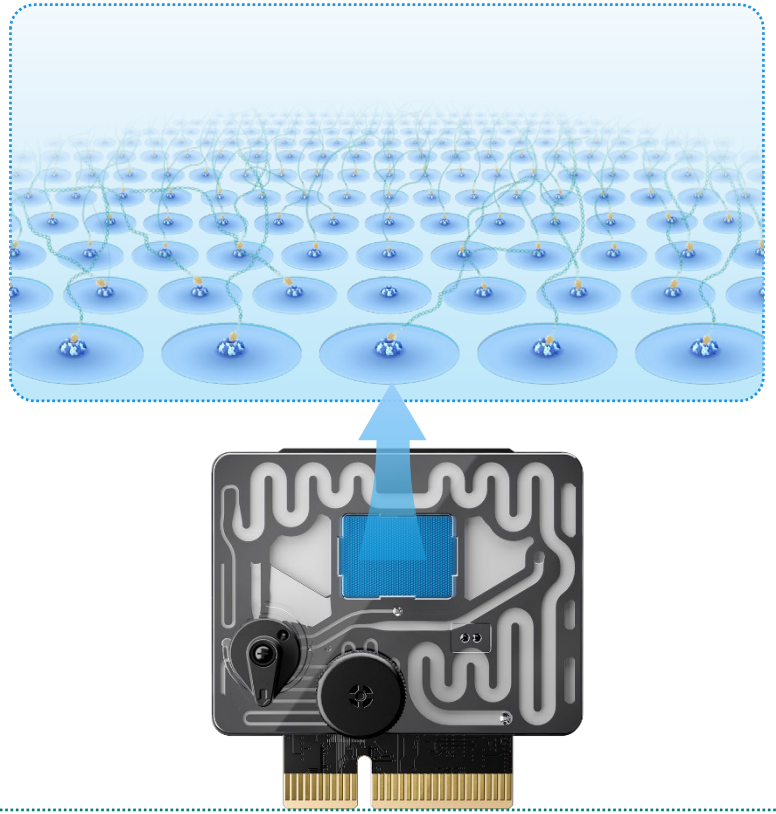
G400-ER

高通量·随心测·中国芯

G400-ER测序原理

G400-ER纳米孔测序系统

30,000+孔道蛋白数



01 DNA文库

02 马达蛋白

- 解旋待测DNA

03 孔道蛋白

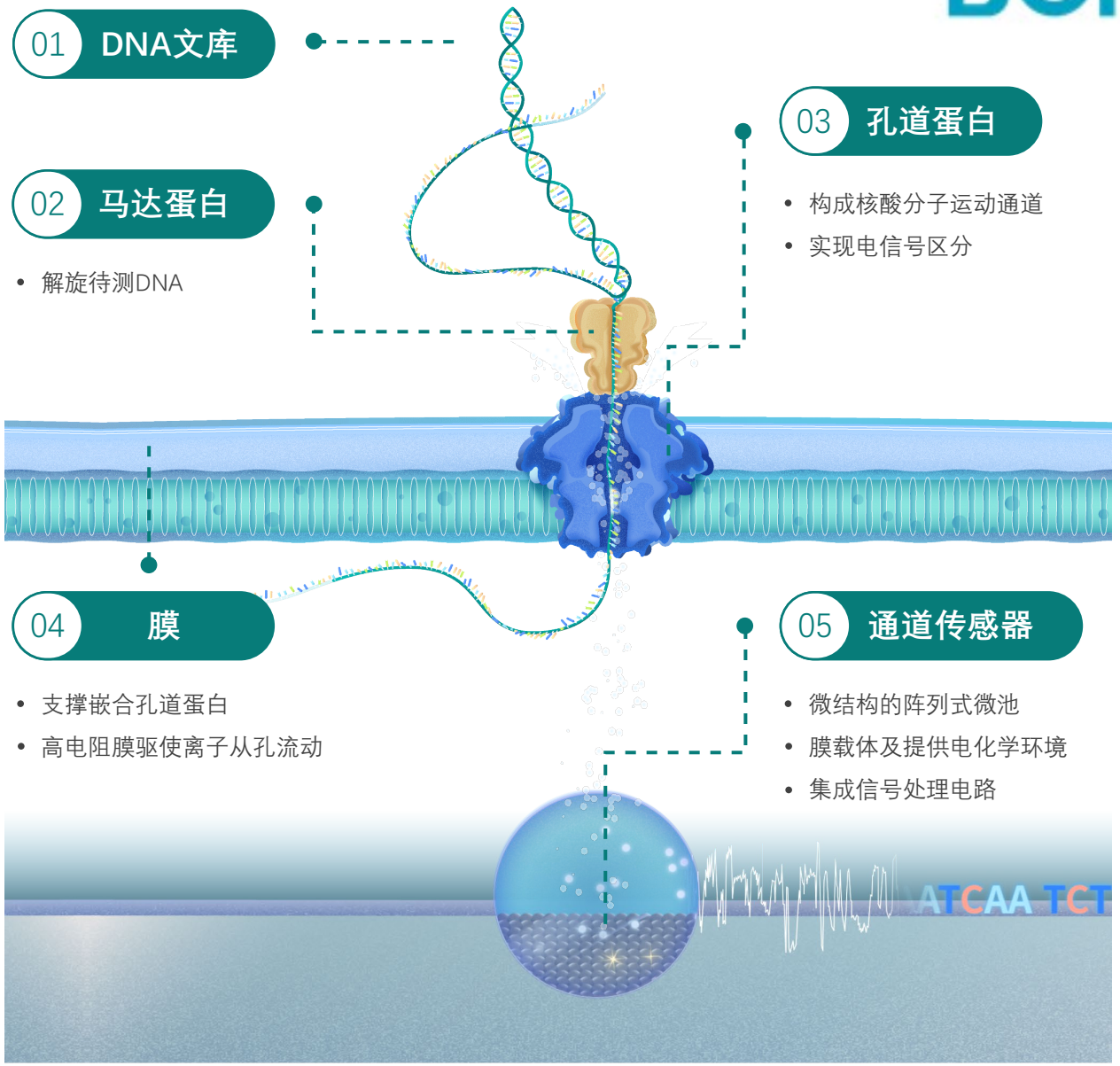
- 构成核酸分子运动通道
- 实现电信号区分

04 膜

- 支撑嵌合孔道蛋白
- 高电阻膜驱使离子从孔流动

05 通道传感器

- 微结构的阵列式微池
- 膜载体及提供电化学环境
- 集成信号处理电路



不同测序方法优劣势对比

	测序方法	准备方法	检测方法	主要优点	主要缺点	主要应用场景
一代	Sanger/毛细血管电泳	PCR+电泳	同位素标记/荧光标记	读长较长、准确度高，重复多聚序列测得好	通量低、成本很高	司法
二代	Illumina (Solexa)	芯片+桥式PCR	荧光标记	准确度较高，通量高、成本低	读长短	临床+科研
	BGI/MGI (Complete Genomics)	芯片+DNB	荧光标记	准确度较高，通量高、成本低	读长短	
	Thermo Fisher(Ion Torrent)	芯片+乳滴PCR	PH变化	时间短	读长短，成本较高、通量低	
三代	Pacific Biosciences	芯片+SMRT Bell	ZMW+荧光标记	读长长（30-50kb)，样本制备简单	成本高，文库制备复杂	科研
	Oxford Nanopore	芯片+leader-hairpin DNA	电流变化	读长长（100kb)，样本制备简单	错误率较高，成本高	

02 测序技术的应用

疾病预测与预防

风险评估

基因测序技术能够通过分析个体的基因序列，评估其患某些遗传性疾病的风险。例如，通过检测BRCA1和BRCA2基因突变，可以预测女性患乳腺癌和卵巢癌的概率，从而采取早期预防措施。

生活方式调整

基因测序结果可以为个体提供个性化的健康建议，如饮食、运动等生活方式的调整。例如，携带某些基因变异的个体可能对某些食物或环境因素更为敏感，通过调整生活方式可以降低患病风险。

早期干预

对于某些具有遗传倾向的疾病，基因测序可以帮助医生在疾病早期进行干预。例如，对于携带APOE4基因变异的个体，医生可以建议其采取预防阿尔茨海默病的措施，如认知训练和药物治疗。

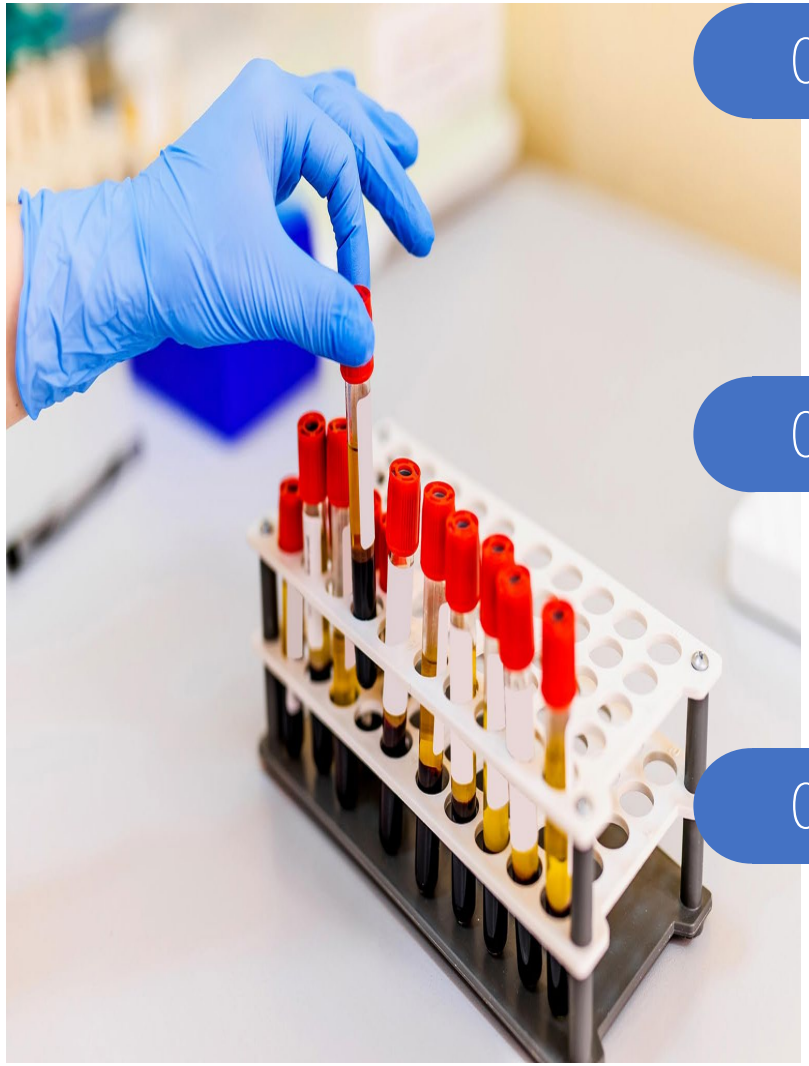

1:33
常见疾病风险
检测28项，2项需要关注
单基因风险
多基因预测
关注
关注
家族性高甘油三酯血症
年龄相关黄斑变性
晚发型阿尔茨海默病
遗传性心律失常
遗传性胰腺炎
鱼鳞病
早发型痴呆
帕金森病
MUTYH相关性息肉病
遗传性易栓症
BRCA1和BRCA2相关癌症


1:35
BGE
结果解读
科学细节
家族性高甘油三酯血症
家族性高甘油三酯血症是一种受遗传因素影响的、血液中甘油三酯水平升高的疾病。甘油三酯升高会使血液粘稠、血流减慢、加重动脉粥样硬化和血管内斑块形成，导致冠心病、脑梗死等心脑血管疾病风险增高，甘油三酯过高时还会诱发急性胰腺炎危及生命。
检测结果
1个杂合患病风险增加突变
突变基因：APOA5
向 小E 健康咨询
风险解读
患家族性高甘油三酯血症的风险增加
* 本次检测用于确定受检者在BGE数据库中APOA5基因的风险突变检出情况，但并不能包含家族性高甘油


1:38
BGE
结果解读
科学细节
高血压
高血压（Hypertension），也称血压升高，是血液在流动时对血管壁造成的压力值持续高于正常值的现象。这种持续的高压状态会对血管和各个器官造成…
展开+
「检测结果」
0
10
20
30
40
50
60
70
80
90
100
63.55%
您该疾病的相对风险位于人群 63.55%，意味着相对风险处于人群普通水平。但这并不等同于该疾病的绝对风险，因为绝对风险还需参照该病的总体发病率，同时生活方式和家族史等因素也对


1:37
点击开启智能体加持的 BGE AI 模式
小E健康助理
你的亮点
查看全部
爆发力运动潜能者
大象牌记忆力勋章
精选专刊
查看全部
心源性猝死要警惕
16项
你的组学报告
首页
溯源
行动
研究
应用

个性化医疗



01

药物选择

基因测序可以帮助医生根据患者的基因信息选择最合适的药物。例如，通过检测CYP2C19基因变异，可以预测患者对抗血小板药物氯吡格雷的反应，从而选择更有效的治疗方案。

02

剂量调整

基因测序还可以帮助医生调整药物剂量，以确保治疗效果和减少副作用。例如，对于携带VKORC1基因变异的患者，医生可以根据其基因信息调整华法林的剂量，以降低出血风险。

03

治疗监测

基因测序可以用于监测治疗效果和预测复发风险。例如，对于癌症患者，通过检测肿瘤基因突变，可以评估治疗效果和预测复发风险，从而及时调整治疗方案。


1:40
用药指南
检测123项，27项需要关注
按药物场景
按调整用药
四高用药
检测19项
日常用药
检测12项
精神类用药
检测14项
癌症化疗用药
检测12项
感染类用药
检测31项
抗抑郁药物
检测16项
心律失常用药
检测3项
手术用药
检测10项
消化系统用药
检测6项


1:40
BGE
使用指南
四高用药
共19项
高血压、糖尿病、高尿酸、高血脂等
四高疾病相关用药
筛选 全部
分类 全部
什么是药物遗传学检测?
什么是药物代谢型?
别嘌呤醇
建议换药
骨骼肌类药物
儿童禁用
孕妇禁用
哺乳禁用
过敏禁用
硝酸甘油
换药
心血管类药物
儿童禁用
过敏禁用
孕妇慎用
哺乳慎用
驾驶慎用
醋硝香豆素
降低剂量
心血管类药物
孕妇禁用
哺乳禁用
饮酒慎用
华法林
复杂情况，请遵医嘱

遗传病筛查

新生儿筛查

基因测序技术可以用于新生儿遗传病的筛查，如苯丙酮尿症、先天性甲状腺功能低下等。通过早期筛查和干预，可以有效预防这些疾病的发展。

孕期筛查

基因测序可以用于孕期筛查，如唐氏综合征、爱德华氏综合征等。通过检测胎儿基因，可以早期发现这些遗传病，为家长提供决策依据。

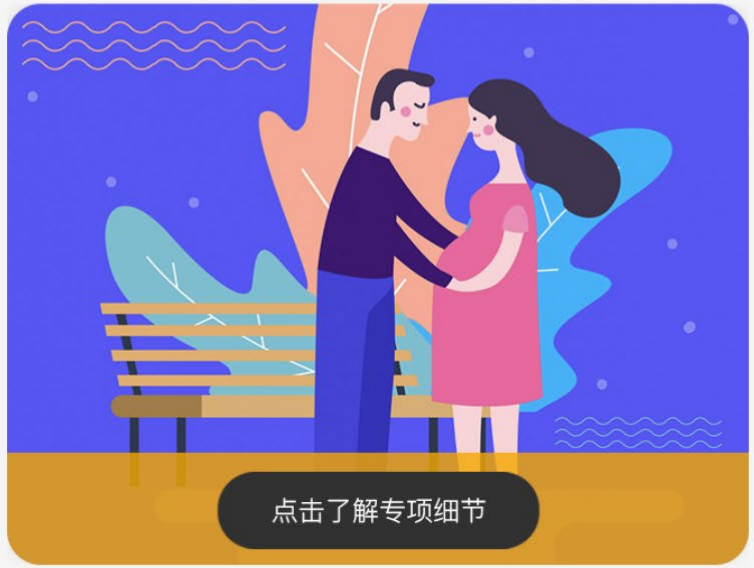
家族遗传病筛查

对于有家族遗传病史的个体，基因测序可以帮助其了解自身携带的遗传病风险。例如，通过检测亨廷顿舞蹈症相关基因，可以预测个体是否携带该病基因，从而采取相应的预防措施。

BGE

单基因遗传病携带者筛查 · 专项

安心孕育可爱宝贝，从了解自己开始



预防出生缺陷
避免单病遗传

由「BGE」与「股份携带者筛查专项」联合提供

影响您和后代 共0项

仅影响后代 共1项

BGE

使用指南

仅影响后代 共1项

本人可能并不表现出临床致病症状，但可能遗传给后代导致风险上升。

Joubert综合征3型 **1个杂合疑似致病突变**

基因组学研究

01.

基因组测序

通过对不同生物体的基因组进行全序列测定，科学家能够全面了解基因组的组成、结构和功能，揭示物种的遗传信息及其调控机制。

02.

基因功能注释

利用基因测序数据，结合生物信息学分析，科学家可以对基因组中的基因进行功能注释，明确基因在细胞代谢、发育和疾病中的具体作用。

03.

比较基因组学

通过比较不同物种的基因组序列，科学家可以识别保守基因和物种特异性基因，揭示物种间的进化关系和功能差异，推动生物学研究的深入发展。

物种进化分析

01

亲缘关系推断

通过基因测序技术，科学家可以比较不同物种或个体的DNA序列差异，推断它们之间的亲缘关系，构建物种的进化谱系，揭示生物进化的历史轨迹。

02

系统发育树构建

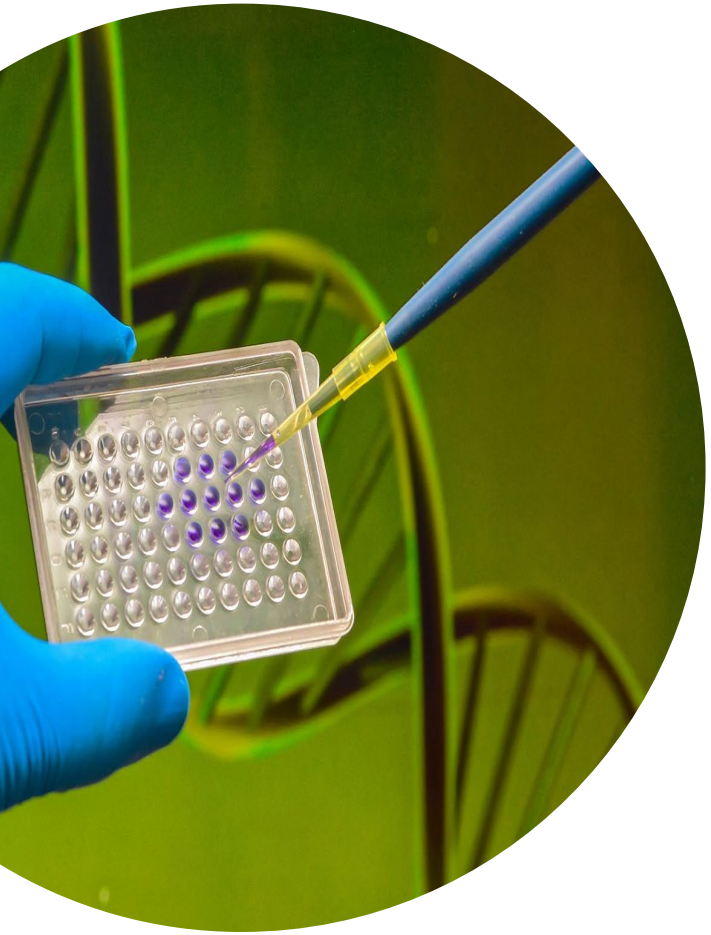
基于基因测序数据，科学家可以构建系统发育树，描述物种之间的演化关系，明确共同祖先和分支节点，为研究生物多样性和进化机制提供重要依据。

03

分子钟分析

利用基因序列的变异速率，科学家可以估算物种分化的时间，揭示进化事件的时间尺度，为研究地球历史和生物演化提供时间框架。

生物多样性研究



物种鉴定

基因测序技术可以用于物种的精确鉴定，尤其是在形态特征难以区分的情况下，通过DNA条形码技术快速识别物种，支持生物多样性调查和保护工作。

遗传多样性分析

通过对不同种群或个体的基因测序，科学家可以评估物种的遗传多样性，了解种群的遗传结构、基因流动和适应能力，为保护濒危物种提供科学依据。

生态系统研究

基因测序技术可以帮助科学家研究生态系统中物种的相互作用和功能角色，揭示生物群落的组成和动态变化，推动生态系统保护和恢复策略的制定。

03 体验一下?

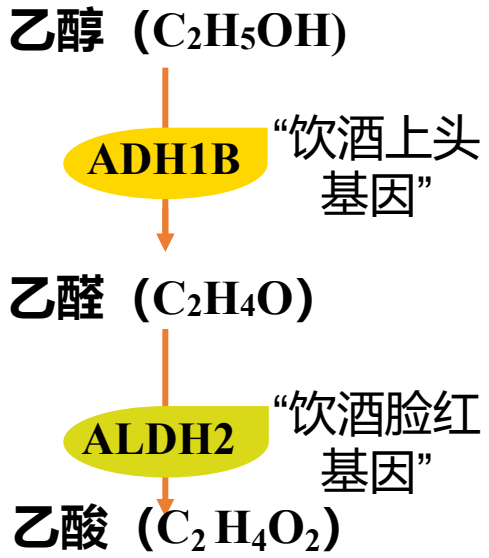

个体特征
检测90项
心理特质
检测8项
运动健身
检测10项
皮肤特性
检测11项
营养代谢
检测24项
饮酒能力
检测3项
易胖风险
检测7项
生活习惯
检测14项
外形特征
检测6项
抵抗力
检测7项

饮鉴 醉翁基因不在酒？在乎“能饮”基因也

饮酒基因速测星盒

（相当于一台mini的基因测序仪）

- 15-20分钟/一次
- 随时随地， 立等可取
- 更多代谢基因……



饮鉴 醉翁基因不在酒？在乎“能饮”基因也

酒量评级	设备亮灯	先天乙醇代谢能力 喝酒上头基因 ADH1B	先天乙醛代谢能力 喝酒脸红基因 ALDH2
五星酒神	5 盏灯亮起	快，TT 纯合子	快，GG 纯合子
四星酒仙	4 盏灯亮起	中，TC 杂合子	快，GG 纯合子
三星酒徒	3 盏灯亮起	快，TT 纯合子	慢，AA 纯合子
		慢，CC 纯合子	快，GG 纯合子
		快，TT 纯合子	中，AG 杂合子
二星酒怂	2 盏灯亮起	中，TC 杂合子	慢，AA 纯合子
		中，TC 杂合子	中，AG 杂合子
一星酒渣	1 盏灯亮起	慢，CC 纯合子	中，AG 杂合子
		慢，CC 纯合子	慢，AA 纯合子

通过基因型组合将代谢能力分为 **1-5 星**（酒渣/酒怂/酒徒/酒仙/酒神）

04 生物数据分析流程

高通量测序 大规模平行测序生信分析

```
@FS2000L1C002R004000052
GGACAGTTCACCCCTCCTTAGGCAACCCGGTGGTCCCCTGCTCCTGGCAG
+
IIIIIII=IIFIIIIIIIIIIIIIIIIIIIIIIIIIIGIIIIIIIII
@FS2000L1C002R004000106
CATTAAACCCAGCACCTACCCTCAGAAATCGCCTCCCAAGCGTTTACATC
+
IIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIIII
@FS2000L1C002R004000116
TTCAGCCCACACCTCTCCTCAGCCCATTACTGTGCAAAGTAGTTCCTAGA
+
IIIIIIIIIIII?IIIIIIIIBIII@EII7I@I?IIIIBIII%IIIIII
@FS2000L1C002R004000133
CGAAAACCTTTTTCCAAGGACAAATCAGAGAAAAAGTCTTTAACTCCACC
+
IIIIIIIIII;9IIIIIIIIIIBIIIIIIIIIII<IIIIIII=IAIII
```

fastq

Soapnuke 过滤

```
/Pipeline/FIS.Traits_v1.0/tools/SOAPnuke filter -
1 ./raw_data/test.fq.gz -l 10 -q 0.5 -n 0.01 -T 1 -
o ./01_clean/ -C test_clean.fq.gz
```

比　对

```
/Pipeline/FIS.Traits_v1.0/tools/bwa-mem2 mem -M -Y -t
1 ./00_ref/MGI358.SNP.fa ./01_clean/test_clean.fq.gz
> ./02_align/test.clean.sam
```

鉴定突变

```
/Pipeline/FIS.Traits_v1.0/tools/freebayes -m 30 -q 20 -
f ./00_ref/MGI358.SNP.fa -@  ./00_ref/alleles_all.vcf.gz -
t ./00_ref/target.358.SE50.subSNP.bed  --report-all-haplotype-
alleles ./02_align/test.clean.sort.uniq.bam
> ./03_SNPCalling/test.clean.SNP.vcf
```

什么是GWAS?

全基因组关联分析的定义

GWAS（Genome-Wide Association Study）是一种通过大规模基因组数据分析，揭示遗传变异与复杂疾病或性状关联的生物医学研究方法。

核心技术与方法

GWAS利用高通量基因分型技术，筛选与特定表型显著相关的单核苷酸多态性（SNP）等遗传标记，通过统计学方法分析各SNP位点与表型的关联强度。

研究步骤

GWAS通常包括样本招募、全基因组范围的SNP分型、统计学分析、独立样本验证和功能实验确认等步骤。

GWAS的应用场景

GWAS在多个领域具有广泛的应用价值，特别是在复杂疾病研究和遗传育种中，通过揭示遗传变异与表型之间的关联，为疾病预防和治疗、作物改良等提供了重要依据。

疾病研究

GWAS已在多种复杂疾病中取得突破，如肿瘤、心血管疾病等，发现了多个与疾病风险相关的遗传位点。

遗传育种

在作物遗传育种领域，GWAS通过探究基因与表型之间的关联，加速了作物育种的进程，提高了作物产量和抗逆性。

跨学科应用

GWAS结合多学科数据，如环境因素、生活方式等，进一步揭示遗传变异与表型之间的复杂关系，为精准医学和个性化健康管理提供了科学依据。

GWAS分析流程：数据准备与清洗

基因型数据格式转换

将原始基因型数据（如VCF或PLINK格式）转换为标准化的分析格式（如bed/bim/fam），确保数据的一致性和兼容性。使用PLINK工具进行格式转换和质量控制。

数据质量控制

通过过滤低质量SNP和样本，确保数据的可靠性。具体步骤包括剔除低基因型呼叫率（如<95%）的SNP和样本，排除不符合哈迪-温伯格平衡的SNP，以及去除高缺失率的位点。

缺失值填补

使用IMPUTE2等工具对基因型数据中的缺失值进行填补，以提高数据的完整性和分析结果的准确性。填补过程中需参考参考基因组（如1000 Genomes Project）进行比对和插补。

基因型与表型数据的关联分析

01

单位点关联分析

对每个SNP进行统计检验，通常采用线性回归（连续表型）或逻辑回归（二元表型）模型，计算SNP与表型之间的关联强度（P值）。分析时需考虑协变量（如年龄、性别、种群结构）的影响。

02

多重检验校正

由于GWAS涉及大量SNP的检验，需进行多重检验校正以控制假阳性率。常用方法包括Bonferroni校正、FDR（False Discovery Rate）控制和置换检验（Permutation Test）。

03

群体分层校正

使用主成分分析（PCA）或混合模型（如GCTA）校正群体分层效应，避免因种群结构差异导致的假阳性关联。

GWAS数据分析模型

- 特定表型与基因型之间的关联关系是如何确定的?

序号	特征性状（25 个）/总共 206 个	影响表型	文献备注
1	rs4988235	乳糖代谢能力	[5][6][7][8][9]
2	rs182549	乳糖代谢能力	[5][6][7][8][9]
3	rs16891982	头发颜色/瞳孔颜色	[10][11]
4	rs28777	头发颜色	[10][11]

- GWAS模型

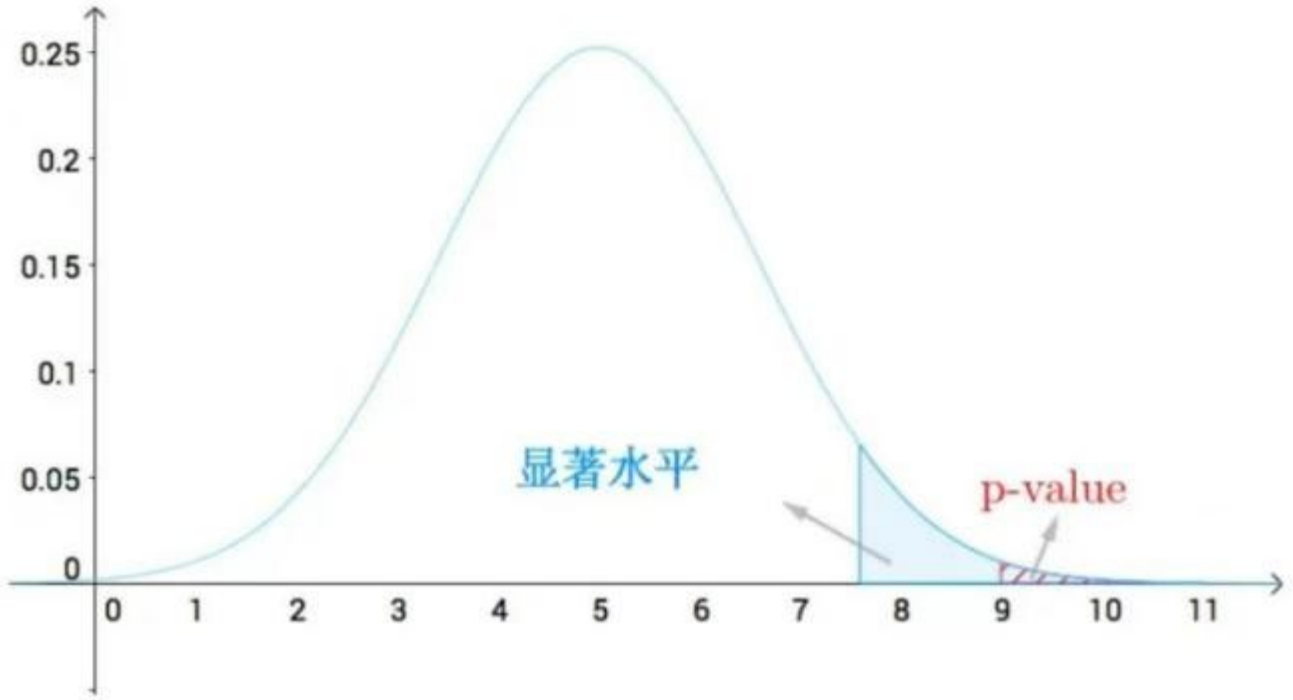
$$Y \sim W O + X , B + g + e$$

表型　环境效应　SNP位点效应　遗传背景随机效应　其他随机效应及误差

https://www.nature.com/articles/s43586-021-00056-9

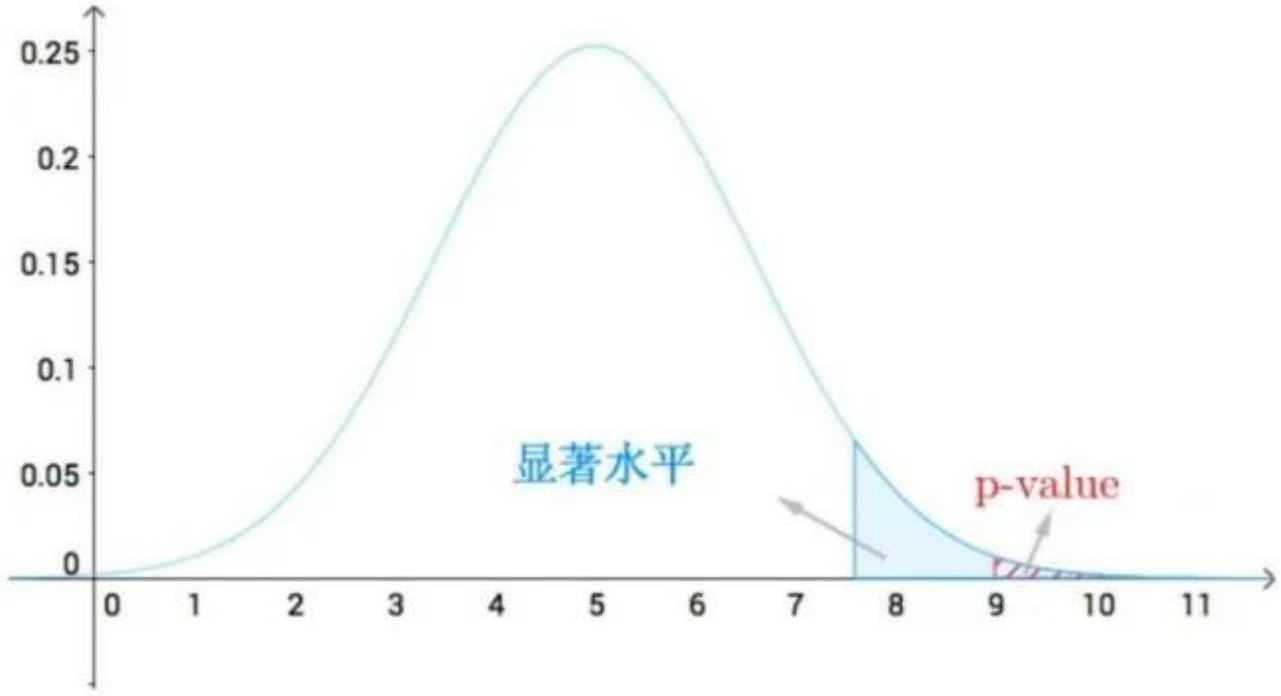
GWAS数据分析模型

- 0假设：特定基因型（Xsβs ）与表型（Y）无关联
- 选择统计模型，进行检验
- 确定拒绝域（p值< ？ ，则拒绝0假设）
- 求出统计检验的p值
- 查看样本结果是否位于拒绝域
- 判定0假设是否成立，做出决策



GWAS数据分析模型

- 0假设： rs1800414位点与本次课程涉及到的5种表型无关联
- 选择统计模型，进行检验
- 确定拒绝域（p值< 0.05，则拒绝0假设）
- 求出统计检验的p值
- 查看样本结果是否位于拒绝域（查看p值是不是小于0.05）
- 判定0假设是否成立，做出决策（判断rs1800414位点是否真的与所有5种表型无关？）



GWAS分析特点

GWAS分析的难点

- 样本量大
- 群体结构复杂
- 大多数表型是受多位点影响的数量性状，导致计算模型复杂度增加
- 表型量化难度大
- ……

本次GWAS实验特点

- 样本量小
- 样本组成简单，无家系样本，无需考虑亲缘关系
- 关注表型明确，分析所涉表型多受单一位点或数量有限的位点影响

GWAS实验分析思路与方法

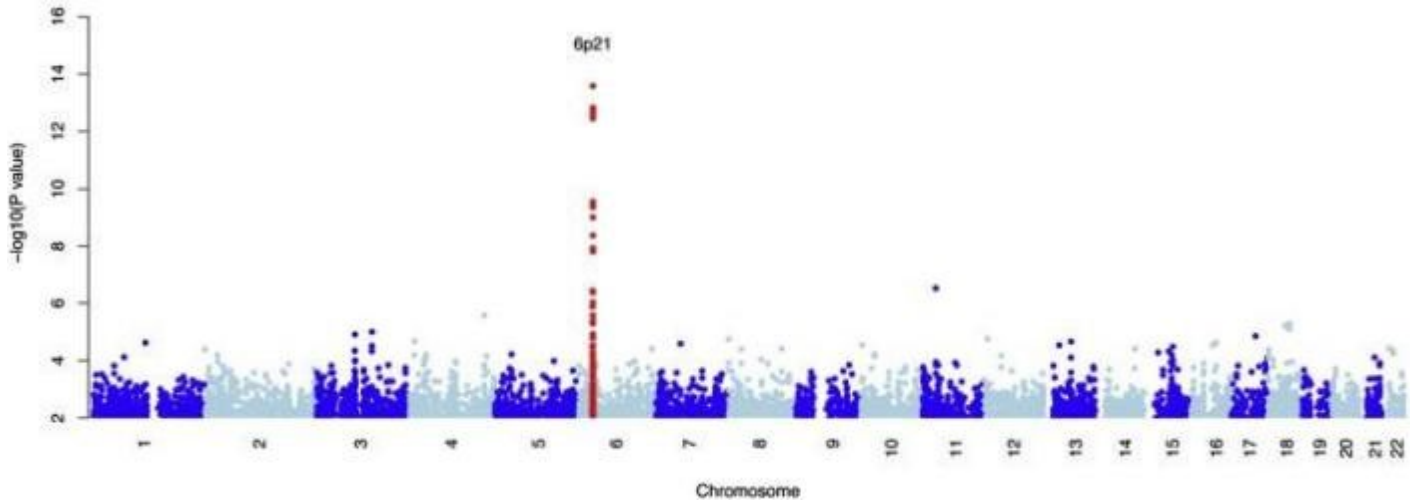
基因型数据准备　　表型数据准备

假设检验（统计模型选择）

p值计算（linux环境、 R语言与工具包）

曼哈顿图绘制

实验结果解读



服务器R语言环境

- R语言代码框架与功能模块解析

登录linux环境，进入R环境　　R

↓

安装R包　　镜像地址选择

↓

读取基因型、表型数据集　　备用数据　or 备用数据+现场整合数据

↓

参数赋值与数据格式整理

↓

统计检验与p值计算Manhattan图绘制

所需数据文件

基因型数据文件

- 个体基因型详表采用了长窄型表示： <样本编码，SNP名称，基因型>。

SampleCode	SNP_Marker	IP_Genotype
342_57	rs1490413	AA
342_57	rs5745448	TC
342_57	rs3737576	TC
342_57	rs1698647	CT
342_57	rs1343469	AG
342_57	rs11239930	GA
342_57	rs7554936	TT
342_57	rs3829868	CC
342_57	rs2814778	TT
342_57	rs560681	AA
342_57	rs10801520	CT
342_57	rs1106201	CT
342_57	rs2013162	AA
342_57	rs2292564	AA
342_57	rs1294331	CC
342_57	rs10495407	GG

所需数据文件

表型数据文件

●表型数据也类似：<样本编码，表型名称，表型值>。

245_58	耳垢类型	干
398_59	耳垢类型	干
367_60	耳垢类型	干
345_61	耳垢类型	湿
336_62	耳垢类型	干
235_63	耳垢类型	干
390_64	耳垢类型	干
205_65	耳垢类型	干
296_66	耳垢类型	干
271_67	耳垢类型	干
358_68	耳垢类型	干
342_57	肌肉性状	爆发
245_58	肌肉性状	爆发
398_59	肌肉性状	爆发

GWAS分析操作

所需数据文件


```
# 自己配置R语言环境的话可以安装下面的包:
# install.packages("openxlsx")
# install.packages("qqman")
# install.packages("tidyr")

# 设置工作目录Windows
# setwd("D:/workspace/R/gwas/")

# 设置Linux服务器上的工作目录，请把demo改成个人文件夹
setwd("/home/ecoli/SHSMU/demo/gwas/")
```

所需数据文件


```
# 导入读入xlsx的包openxlsx
library("openxlsx")

# 导入基因型文件
rawgenotype  <- read.xlsx("20211118.R2.SNP.xlsx")

# 导入表型文件
rawphenotype<- read.xlsx("20211118.phenotype.xlsx", colNames=FALSE)

# 通过数据框操作，提取文件的前3列有用信息，去掉后面的列
mg<- rawgenotype[,c(1,2,3)]
mp<- rawphenotype[,c(1,2,3)]
```

整合转换数据格式


```
# 第一列是样本代码，获取样本数目
nsample<-length(levels(factor(mg[,1])))

# 基因型数据第二列是SNP位点名称，获取数目
nloci <- length(levels(factor(mg[,2])))

# 表型数据第二列是性状特征名称，获取数目
ntrait<- length(levels(factor(mp[,2])))
```

整合转换数据格式


```
# 导入tidyr包进行数据格式的转换
library(tidyr)

# 第一列为样本名称，转换成行名称
# 第二列为基因型或表型名称，转换为列名称
# 第三列为基因型或表型取值，对应为矩阵项取值
md<-pivot_wider(mgp,names_from=names(mgp)[2],values_from=names(mgp)[3])

# 先单独生成一个公式试试，波浪号~前面为空，表示按出现项数来计数
xtabs(~Popu.rs1800414 + 肤色深浅, md)
```

对离散性状生成频数列联表和进行关联显著性检验

以下开始使用as.formula函数和paste函数来规律地生成公式表达式

第一列是Sample Code，第二列到nloci+1列是SNP，后面直到nloci+ntrait+1列是表型性状

```
dname<-names(md)
```

前面已经保存了nsample、nloci和ntrait三个值，方便我们来用下标进而可循环批量生成公式。

公式为:

```
xf <- paste(" ~ ",dname[2]," + ",dname[nloci+2])
```

as.formula在此很关键，#这样我们就得到了Chi-squared test或fisher's exact test需要的列联表mc

```
mc <- xtabs(as.formula(xf), md)
```

对离散性状生成频数列联表和进行关联显著性检验


```
# 开始进行Chi-squared test
chisq.test(mc)

# 这样会提示“Chi-squared近似算法有可能不准”。我们换成:
chisq.test(mc,simulate.p.value = TRUE, B = 10000)

# 参数simulate.p.value是指定用Monte Carlo模拟来计算p-value，参数B设置模拟次数。
# 因为样本数目较少，即列联表中频数较小，实际上应该直接用fisher's exact test。
fisher.test(mc)
```

计算所有位点-性状关联并画曼哈顿图


```
# 计算所有位点-性状关联并画曼哈顿图
# 接下来我们写一个循环来完成所有SNP与所有性状之间的列联表以及显著性检验。
 # 首先造一个矩阵来存储显著性q-values值， 并加上行列名称， 以便筛选后知道是哪个SNP关联上哪个性状。
assocTraitQ  <- data.frame(matrix(rep(0,ntrait*nloci),ncol =ntrait))
names(assocTraitQ)<-dname[(1+nloci+1):(1+nloci+ntrait)]
row.names(assocTraitQ)<-dname[(1+1):(1+nloci)]

#  使用pdf命令开一个pdf画布， 因为实时显示只能有一张图， 而我们有多个性状关联结果， 所以输出到pdf文件中
pdf(file="E5GWAS.pdf", family="GB1")
```

计算所有位点-性状关联并画曼哈顿图

每个性状画一张曼哈顿图。

```
for(k in 1:ntrait){
        pvalues=array(0)
        for(i in 1:nloci){
                xf<- paste(" ~ ",dname[1+i]," + ",dname[1+nloci+k])
                mc<-xtabs(as.formula(xf), md)
                if(dim(mc)[1]==1){
                        pvalues[i]<- 1
                }else{
                        pvalues[i]<-fisher.test(mc)$p.value
                }
                assocTraitQ[i,k]<-pvalues[i]
        }
```

计算所有位点-性状关联并画曼哈顿图


```
        nSNP<-length(pvalues)
        gwasMatrix<-data.frame(dname[(1+1):(1+nloci)],rep(1,times=nSNP), 1:nSNP, pvalues)
        names(gwasMatrix)<-c('SNP', 'CHR', 'BP', 'P')
        manhattan(gwasMatrix,
                annotatePval = -log10(0.05/nSNP), #标注p值最小的SNP位点
                suggestiveline = FALSE,
                genomewideline = -log10(0.05/nSNP), #画出邦费罗尼校正显著性的阈值线
                ylim = c(0,5),
                xlab = 'SNP index',
                main = dname[1+nloci+k]
                )
}
```

保存并关闭pdf文件

```
dev.off()
```

计算所有位点-性状关联并画曼哈顿图


```
        nSNP<-length(pvalues)
        gwasMatrix<-data.frame(dname[(1+1):(1+nloci)],rep(1,times=nSNP), 1:nSNP, pvalues)
        names(gwasMatrix)<-c('SNP', 'CHR', 'BP', 'P')
        manhattan(gwasMatrix,
                annotatePval = -log10(0.05/nSNP), #标注p值最小的SNP位点
                suggestiveline = FALSE,
                genomewideline = -log10(0.05/nSNP), #画出邦费罗尼校正显著性的阈值线
                ylim = c(0,5),
                xlab = 'SNP index',
                main = dname[1+nloci+k]
                )
}
```

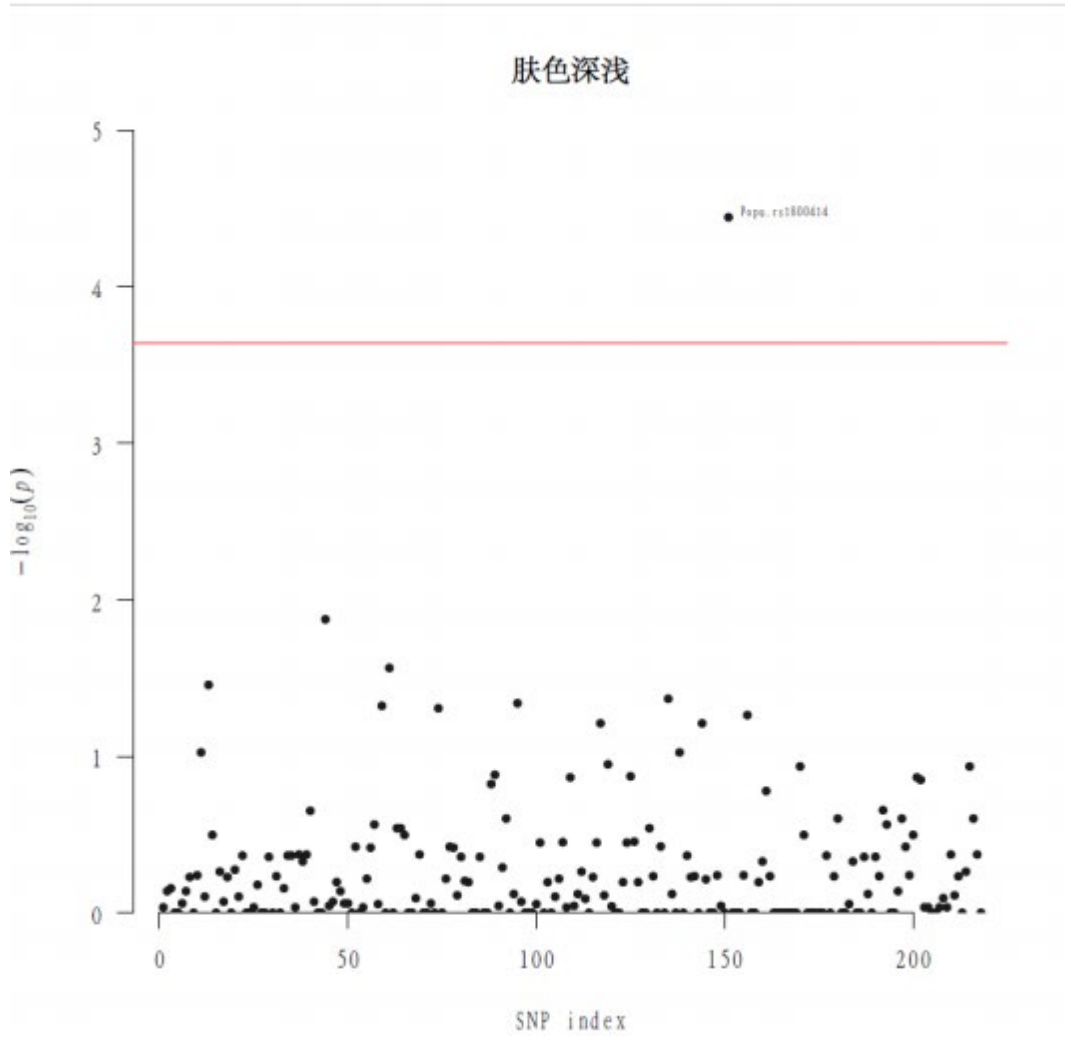
保存并关闭pdf文件

```
dev.off()
```

计算所有位点-性状关联并画曼哈顿图

我们筛选出q-value小于0.05的SNP

assocTraitQ[apply(assocTraitQ,1,min)< 0.05/nSNP,]



谢谢大家

徐俊美

华大集团 猛犸教育