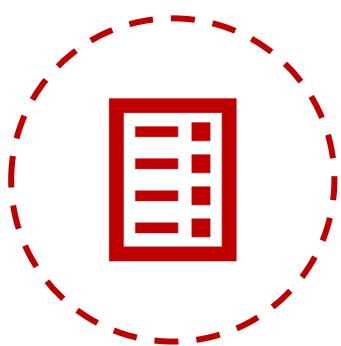


基于三代结构变异的某物种群体遗传学研究

王 鹏

科技服务技术支持部



01

研究背景

02

研究方案

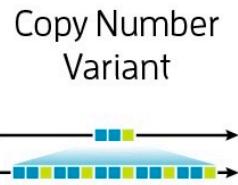
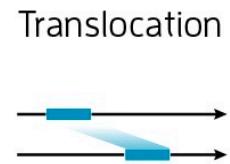
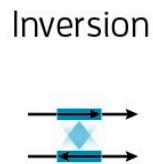
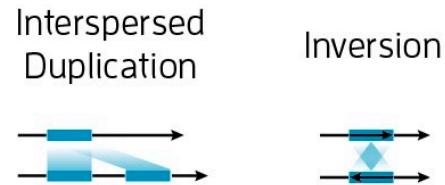
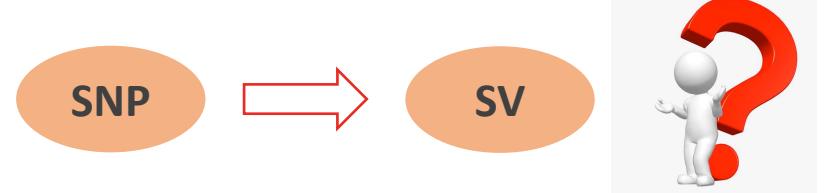
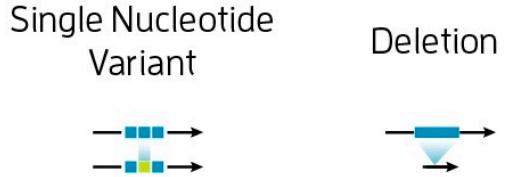
03

PB和ONT比较

04

贝瑞介绍

研究背景--变异的类型



SV对个体间基因组差异影响更大（是SNP的3-10倍）

SV更能影响基因表达、表型/性状和基因组进化

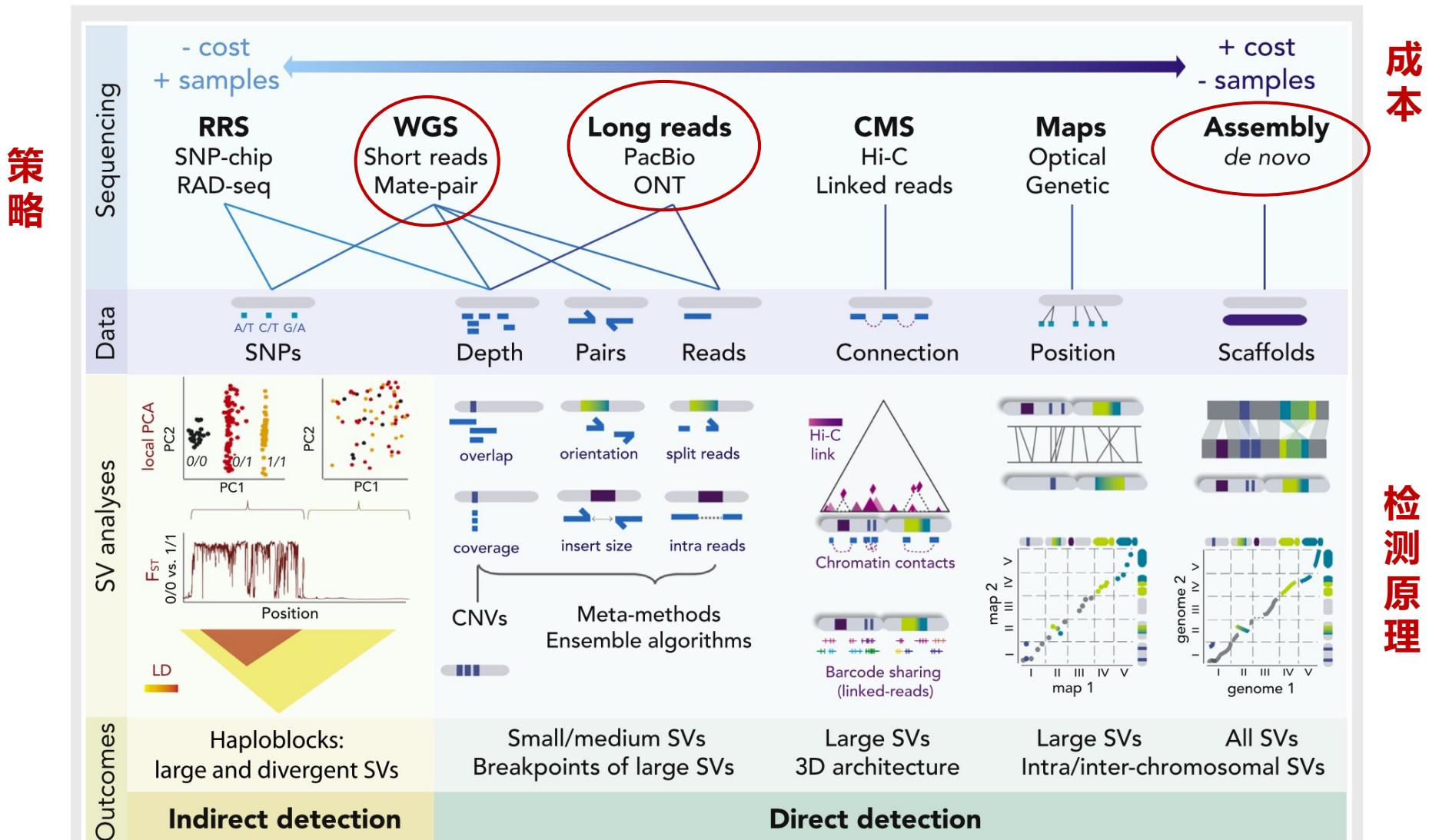
Types of Variants

突变类型	突变描述	突变差异
单核苷酸多态性 (简称SNP)	单个核苷酸的变异	1个碱基的差异
插入缺失 (简称Indel)	小片段 (<50bp) 的插入或缺失	1-50个碱基的差异
结构变异 (简称SV)	大片段 (> 50bp) 的插入、缺失、倒位、易位等	50个以上碱基的差异
拷贝数变异 (简称CNV)	基因组片段的拷贝数增加或者减少	大于1000bp的差异

研究背景--变异检测的技术

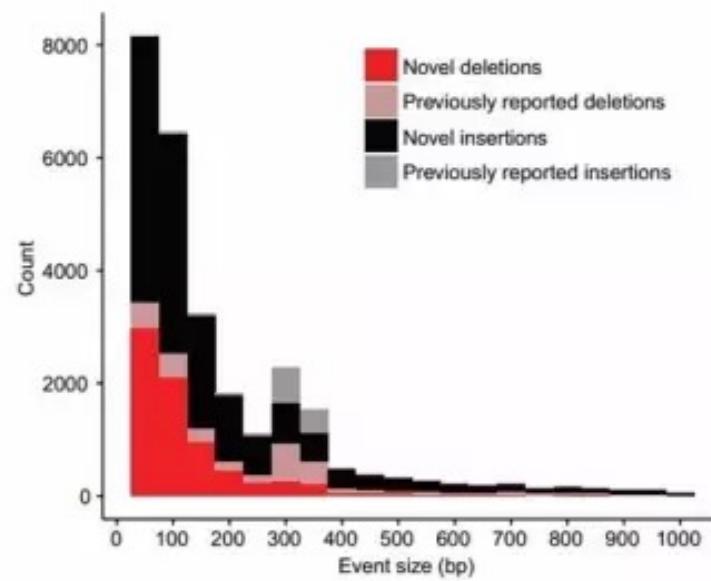
技术策略	介绍	缺点
Illumina WGS	利用高深度全基因组测序中的paired reads和splice alignment reads精确定位断裂点	读长短， SVs检测灵敏度及准确性较低
RNA-seq	利用paired reads和splice alignment reads直接寻找融合基因	有 组织特异性 和基因之间转录丰度差异，很多低丰度和不表达基因无法测到。
Strand-seq	仅对双链中的模板链进行测序，可用于检测倒位，大片段的缺失和重复等	不适用于小片段的SVs检测
10x Genomics	利用barcode将处于一个大片段的短序列reads利用信息学手段拼成长reads，利用长读长splice alignment reads精确定位断裂点	覆盖不均一 ，存在很大的间隙，插入检测难度大（最广泛使用的算法之一Long Ranger目前不能检测插入）
Bionano	利用酶切位点光谱比对参考基因组酶切谱精确定位断裂点，用于检测大型SVs，特别是易位	依赖于线性基因组中酶切位点的存在， 分辨率低 ，很难达到碱基级别
Hi-C	Hi-C的reads长度可以达到Mbp级别，从而使得其适合用于检测大片段的SVs，尤其是易位	不适合检测小片段的SVs ，检测出的SV片段大小一般大于2Mb
PacBio	利用长读长splice alignment reads精确定位断裂点。	单碱基准确性低 ，但SV检测准确性高

研究背景--变异检测的技术



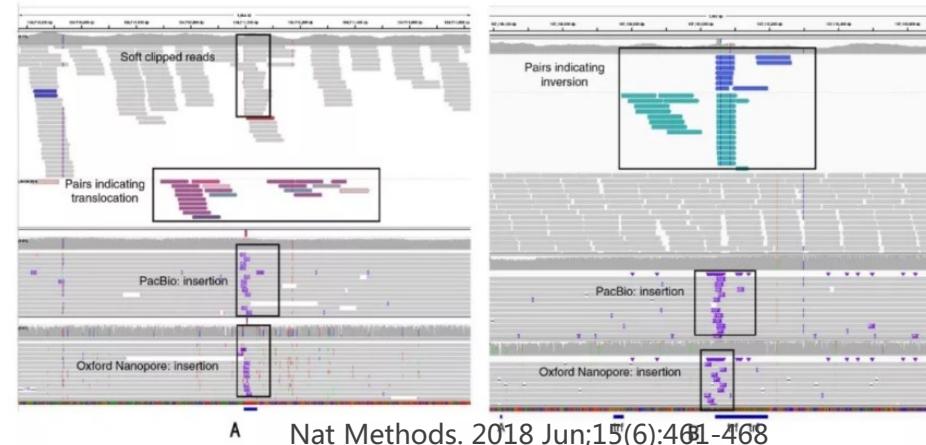
研究背景--Illumina检测SV

- 敏感度较低、错误率高
- 难以检测插入、串联重复及复杂SV
- 存在检测偏好性及系统性错误



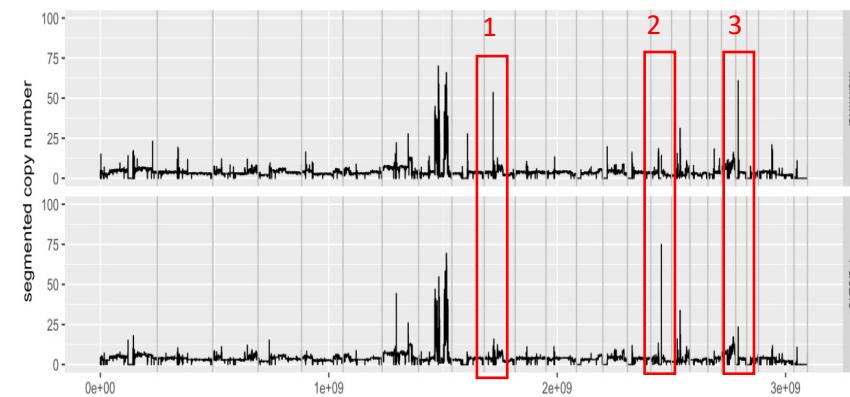
Genome Res. 2017 May;27(5):677-685.

PacBio 检测SV有83%是基于二代未发现的



A Nat Methods. 2018 Jun;15(6):461-468

Illumina将插入 (48.87%) 错误识别为倒位



基于Illumina检测有重复区域和GC偏好性

1、3号峰由于重复假阳性；2号峰由于高GC区域导致假阴性

研究背景--PacBio检测SV

PacBio的两种测序模式：

1、Continuous Long Read Sequencing (CLR)

特点：插入片段长（20/30/40/60kb），测序时间短（默认15h）

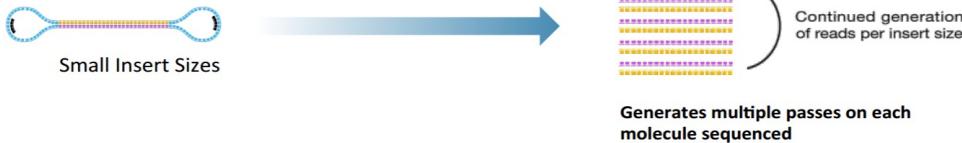
Standard Sequencing for Continuous Long Reads (CLR)



2、Circular Consensus Sequencing (CCS)

特点：插入片段短（15kb），测序时间长（默认30h）

Circular Consensus Sequencing (CCS)



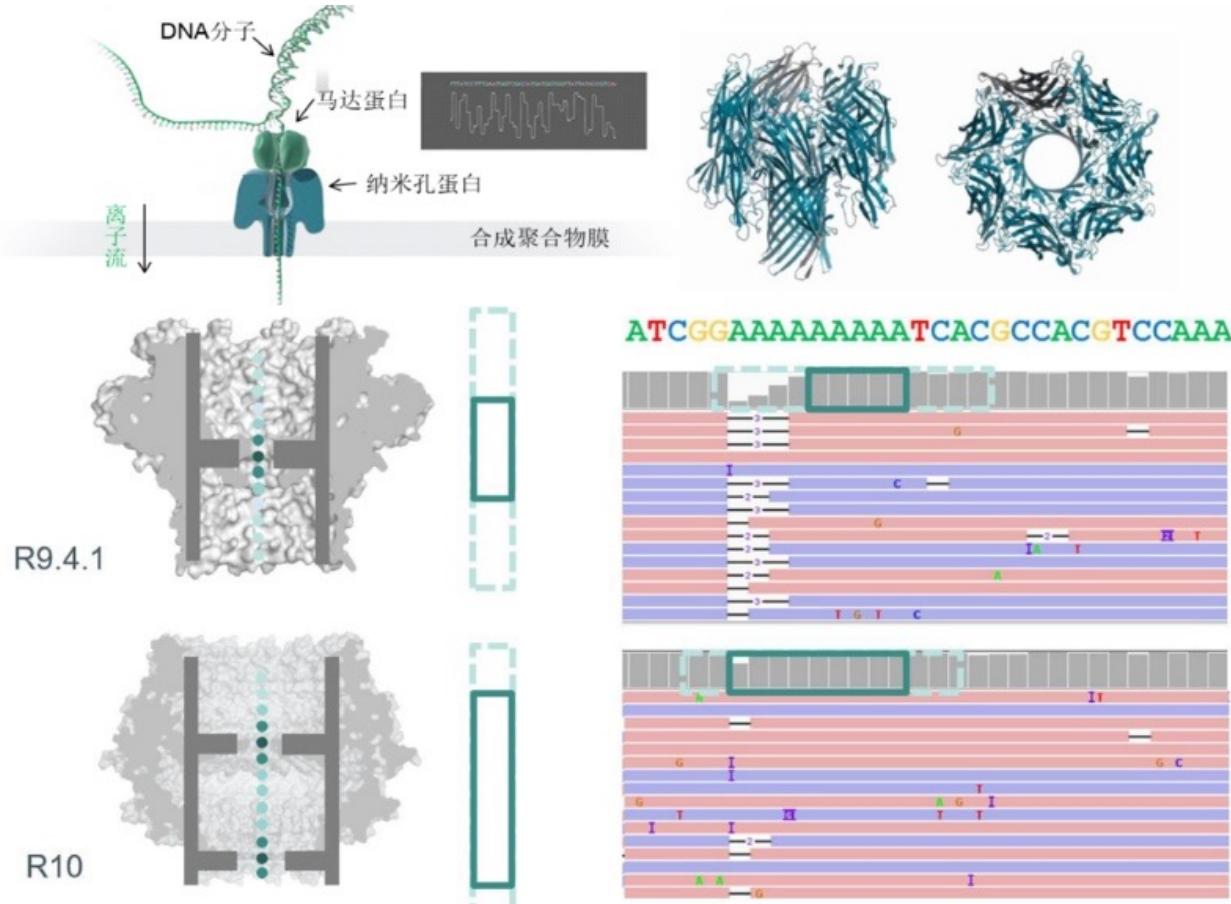
- 超长读长、无PCR扩增、无GC偏好，轻松跨过高重复和高复杂度区域
- 更高的变异检出率
- 更准确的变异检测信息



Platform	Caller	All variants			Deletions			Insertions		
		Prec.	Recall	F1 ^	Prec.	Recall	F1	Prec.	Recall	F1
PacBio (CCS)	Integrated	96.13%	95.99%	96.06%	97.66%	96.88%	97.27%	94.97%	95.30%	95.13%
PacBio (CCS)	pbsv	96.26%	94.93%	95.59%	96.71%	94.98%	95.84%	95.95%	94.89%	95.42%
PacBio (CLR)	pbsv	94.64%	94.48%	94.56%	96.70%	95.57%	96.13%	93.11%	93.64%	93.37%
PacBio (CCS)	Sniffles	94.28%	91.76%	93.01%	96.56%	92.19%	94.32%	92.59%	91.44%	92.01%
PacBio (CCS)	paftools/Canu ‡‡	93.16%	92.32%	92.74%	95.84%	92.76%	94.28%	91.48%	91.99%	91.73%
PacBio (CCS)	paftools/FALCON †	93.25%	89.14%	91.15%	95.99%	89.00%	92.36%	91.64%	89.25%	90.43%
PacBio (CLR)	Sniffles	95.66%	79.33%	86.73%	98.19%	80.07%	88.21%	93.80%	78.76%	85.62%
Illumina	Manta	85.34%	55.88%	67.53%	85.95%	76.90%	81.17%	92.12%	39.65%	55.44%
10X	paftools/Supernova	64.52%	52.74%	58.04%	55.37%	73.71%	63.24%	82.74%	36.57%	50.72%
10X	LongRanger	83.79%	39.83%	53.99%	94.66%	70.18%	80.60%	59.39%	16.41%	25.71%
Illumina	Delly	65.92%	19.90%	30.58%	65.92%	45.70%	53.98%	0.00%	0.00%	0.00%

PacBio检测SV精确性及召回率双高

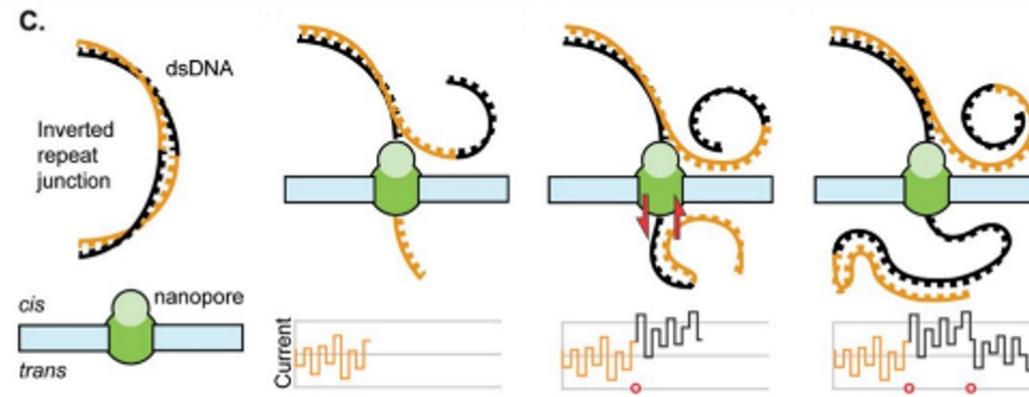
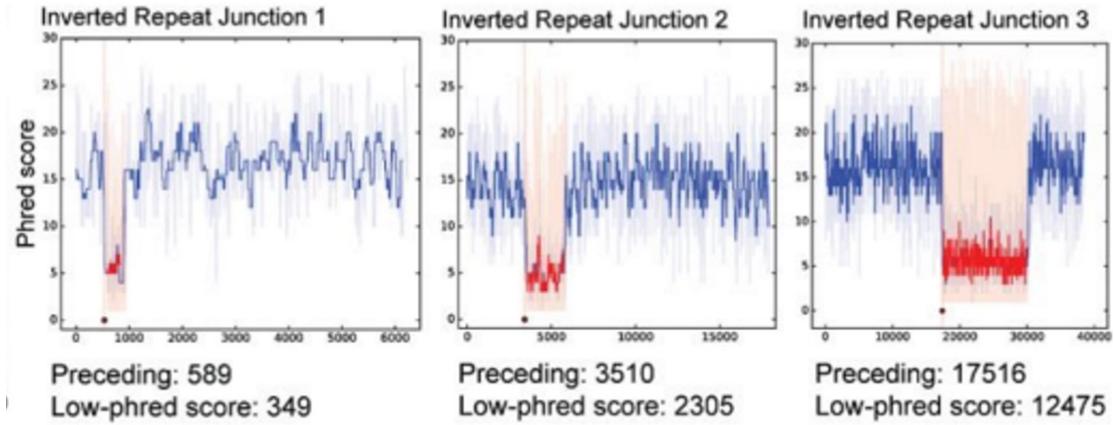
研究背景--Nanopore检测SV



Nanopore测序的寡核苷酸**系统错误**

Nanopore平台纳米孔蛋白读取头具有约 5 个碱基接触位点（R9 芯片），R10 芯片升级为双读取头，具有约 10 个碱基接触位点，在高速读取电信号的过程中，对均聚物、复杂度的重复序列的测序容易产生误判。

研究背景--Nanopore检测SV



Nanopore测序在酵母中结构变异研究应用 (bioRxiv, 2019)

反向重复序列可能形成的DNA二级结构导致了测序数据的不准确性。当反向重复结构的DNA双链解开，穿过Nanopore的纳米小孔后，容易形成单链的互补配对。由此，互补的二级机构将对后续检测的DNA分子形成拉力，从而使得碱基读取出现了不可弥补的系统性错误。

- 错误率高
- 存在检测偏好性
- 系统性错误

技术策略	PB	ONT
SV检出个数	15499	26657
被其他数据集验证比率	95%	57%

FJ Sedlazeck., Nature Methods, 2018.

研究背景--结构变异技术比较

● Illumina、PacBio、Nanopore

	Deletions			Insertions		
	Counts	FDR	Sensitivity	Counts	FDR	Sensitivity
PacBio (30-fold) ²⁸	8,737	3%	95%	12,378	3%	93%
PacBio (10-fold) ²⁸	6,798	3-10%	83%	11,252	3-10%	83%
ONT (30-fold) ²⁹	28,791	65%	93%	3,900	65%	11%
10X Genomics (30-fold)	3,166	Not reported	39%	Not reported	N/A	0%
Illumina ³⁰	1,910	2-4%	24%	1,090	1-4%	9%
BioNano ^{31,32}	522	3%	6%	769	2%	6%

PacBio has the highest sensitivity

Oxford Nanopore has the highest false discovery rate

Other technologies struggle with poor sensitivity

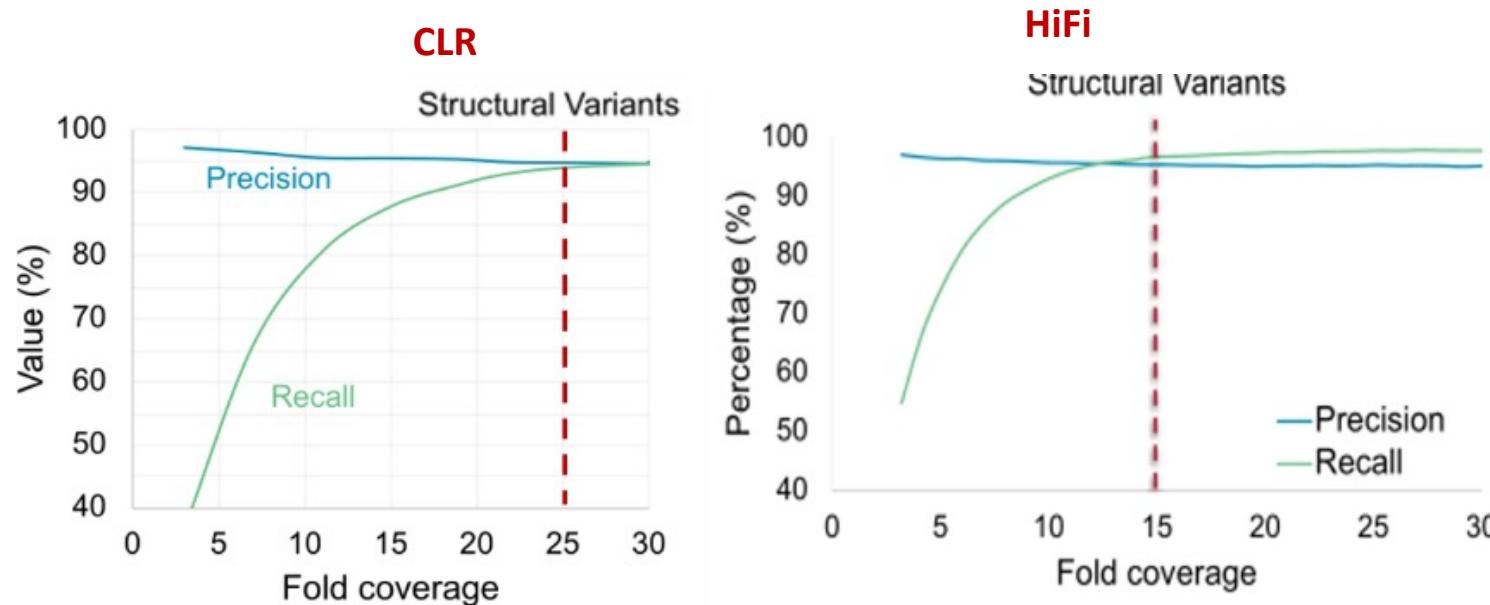
技术策略	Illumina	PacBio	Nanopore
检出个数	低	高	高
敏感度	低	高	高
准确度	低	高	低
串联重复检出	低	高	高
复杂构变异检出	难	易	易

- 在检测SV上，PacBio具有最高的敏感度
- ONT具有最高的错误率
- 其它检测（Illumina和Bionano）的敏感度也较低

研究背景--PacBio SV检测模式和深度

两种模式可灵活选择：CLR模式和CCS模式（左下图）。

- ✓ 基于CLR模式：构建20Kb/30K/40Kb/60Kb文库，测序深度推荐15X~30X，最佳测序深度 $\geq 25X$ ；
- ✓ 基于CCS模式：可构建15K的文库，获得HiFi reads，测序深度推荐 $\geq 10X$ ，最佳测序深度 $\geq 15X$ 。



PacBio对于SV检出的饱和度展示（来源于PacBio官方）

研究背景--为何做SV

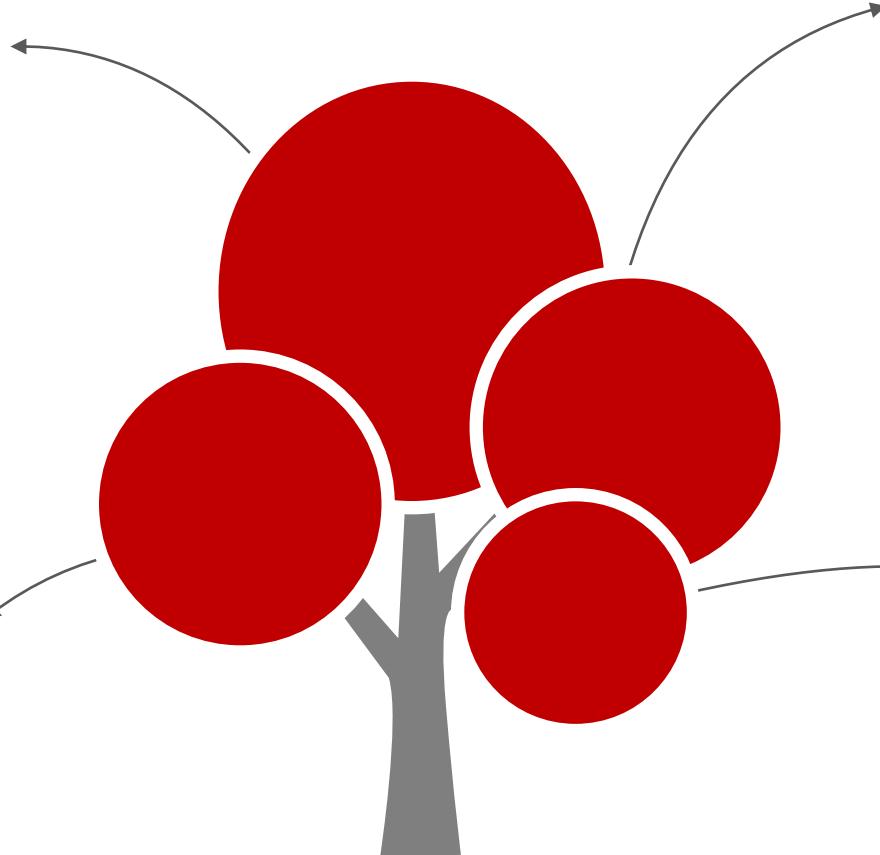
重要性逐渐被发掘

植物基因组中，大概1/3（三分之一）已报道的作物表型是由于结构变异引起的。

期刊	物种	SV类型	表型
Nature	水稻	重复	抗涝
Nature Genetics	水稻	缺失	谷粒大小
Molecular genetics and Genomics	谷子	缺失	粘性
Science	葡萄	插入	果实颜色
plant cell	血橙	重复	果皮颜色
Nature Plant	番茄	重复	花序
Nature Plant	葡萄	倒位	浆果颜色
Plant Biotechnol J	油菜	缺失	抗病

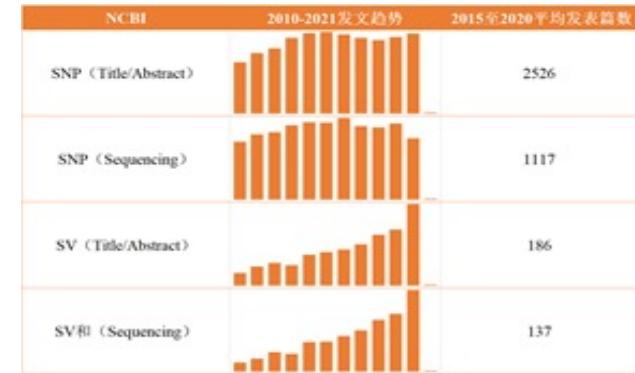
技术升级，成本下降

测序技术进步及成本的下降，使识别基因组中SV变得越来越可行，特别是三代测序技术的发展，使全基因组范围内产生准确的长读长数据变得更容易。



文章发表趋势增多

基于SNP的重测序研究瓶颈，研究方向逐渐转向基于大尺度SV的关联分析，文章频发。



材料减少，文章提升

以GWAS为例，SNP-GWAS的文章样本个数趋势增加，分数降低。SV-GWAS方向新颖，样本数较少可发不错的文章。

发表时间	期刊	IF	研究物种	样本个数
2021	Genome Biology	10.806	桃	149个
2020	Genome Biology	10.806	桃	336个
2020	Nature Communications	12.121	绵羊	248个
2020	Cell	38.637	大豆	2898个
2020	Nature Plants	13.256	油菜	2141个
2019	Nature Genetics	27.603	玉米	521个

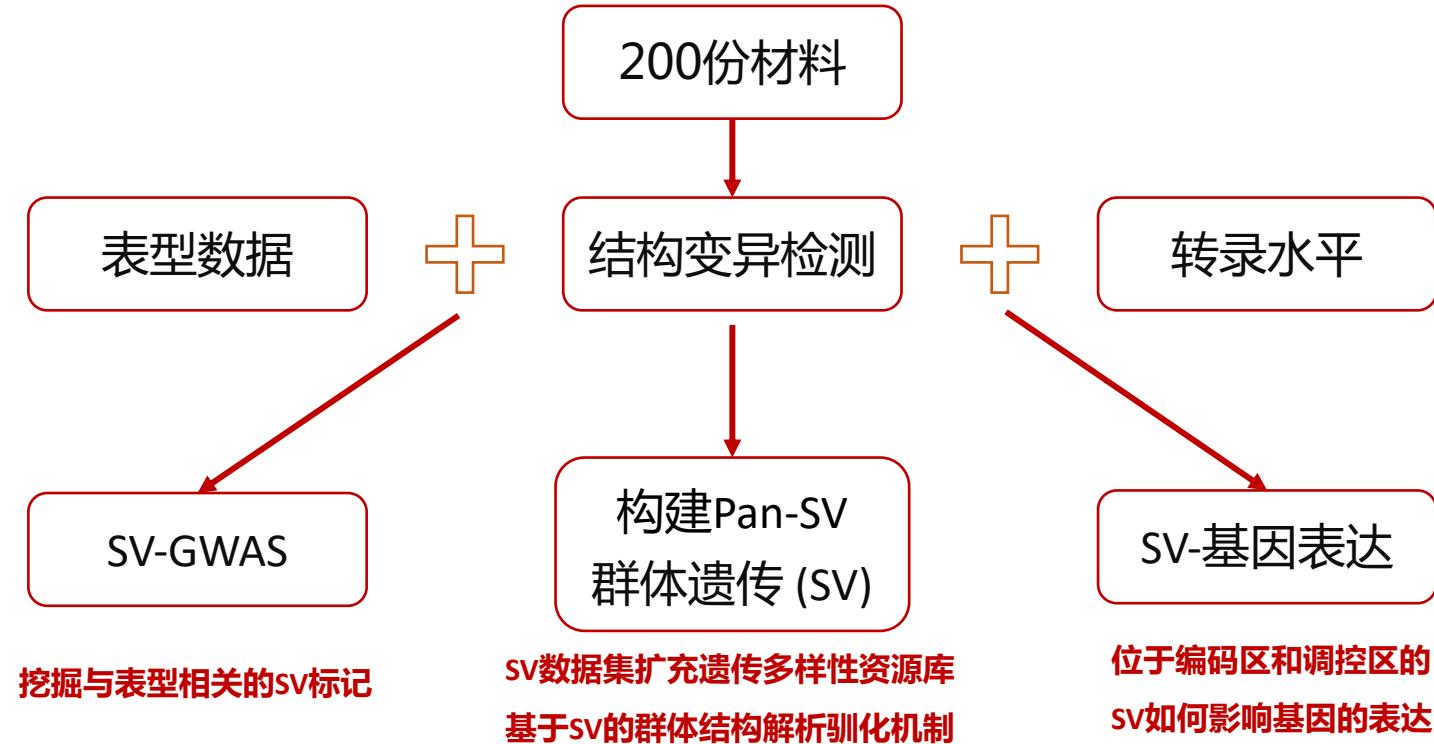
研究背景--三代SV的应用案例

物种	发表时间	期刊	影响因子	总结
鸣禽	2020.7	Nat Commun	12.121	鉴定方式：组装比较和reads mapping（二代+三代）；鉴定到了220,452个SV，其中一个2.25Kb的LTR逆转座子插入降低了NDP基因的表达，产生了生殖前隔离
水稻	2020.7	Mol Biol Evol	11.062	鉴定方式：组装序列和长reads比对；栽培稻比野生稻基因组中多了25%的SV和转移元件插入（MEIs），表明SV对水稻驯化有很大的贡献
油菜	2020.7	Plant Biotechnol J	8.154	鉴定方式：长度长比对；10%的基因都受到了小（30-10000bp）和中等（10000-30000）SV事件的影响，50%的SV在100-1000bp之间，这是二代无法检测到的
番茄	2020.6	Cell	38.637	鉴定方式：长度长比对；共得到238490个SVs，并构建了pan-SVs，SV对基因的调控、表达和性状都有较大影响，细胞色素P450基因串联重复会影响番茄果重
油菜	2020.1	bioRxiv	-	鉴定方式：组装，短reads和长reads比对；共鉴定到~120K的高置信SV，被SV影响到的基因多余应激反应、分生组织和花的发育有关
葡萄	2019.9	Nature Plant	13.256	鉴定方式：组装，短reads和长reads比对；纯化选择对SV有很强的作用，尤其是倒位和易位

科学问题



研究思路



建库测序策略

建库测序策略比较

文库类型	检测变异	文库个数	文库大小	测序模式	测序深度	测序仪	备注
PacBio-CLR文库	SV	1个/样	20Kb	CLR	15x (15G/样)	PacBio Sequel II	3样/cell
PacBio-HiFi文库	SV	1个/样	15Kb	CCS	10x (10G/样)	PacBio Sequel II	1-2样/cell
Illumina重测序	SNP+SV	1个/样	350bp	PE150	30x (30G/样)	Illumina Novaseq 6000	
Illumina转录组	-	1个/样	350bp	PE150	6G/样	Illumina Novaseq 6000	

本研究可选策略

策略	PacBio-CLR	PacBio-HiFi	Illumina重测序	Illumina转录组	推荐指数
策略1-纯三代	✓ (200)			✓ (待定)	★★★★
策略2-纯三代		✓ (200)		✓ (待定)	★★★
策略3 : 二+三	✓ (>20)		✓ (200)	✓ (待定)	★★★★
策略4 : 二+三		✓ (>20)	✓ (200)	✓ (待定)	★★★★★

三代SV分析内容



群体进化

GWAS

三代SV--群体进化和GWAS

农艺经济性状研究
优质功能基因挖掘
指导分子育种

GWAS

以连锁不平衡（LD）为基础，通过识别全基因组范围内由数百个甚至上千个个体组成的群体中高密度的分子标记，筛选出与复杂性状表现型变异相关联的分子标记，进而挖掘与**表型相关的基因**的方法。



pOr (pop. 70796)



dOr (pop. 70796)



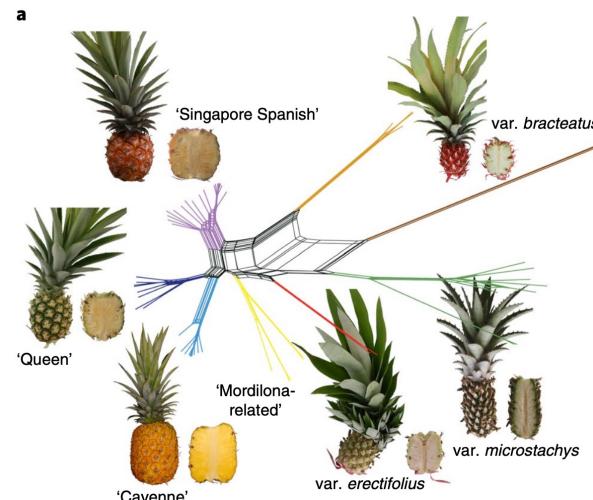
Y (pop. 97837)

Nat Genet. 2016 Jun;48(6):657-66

群体进化

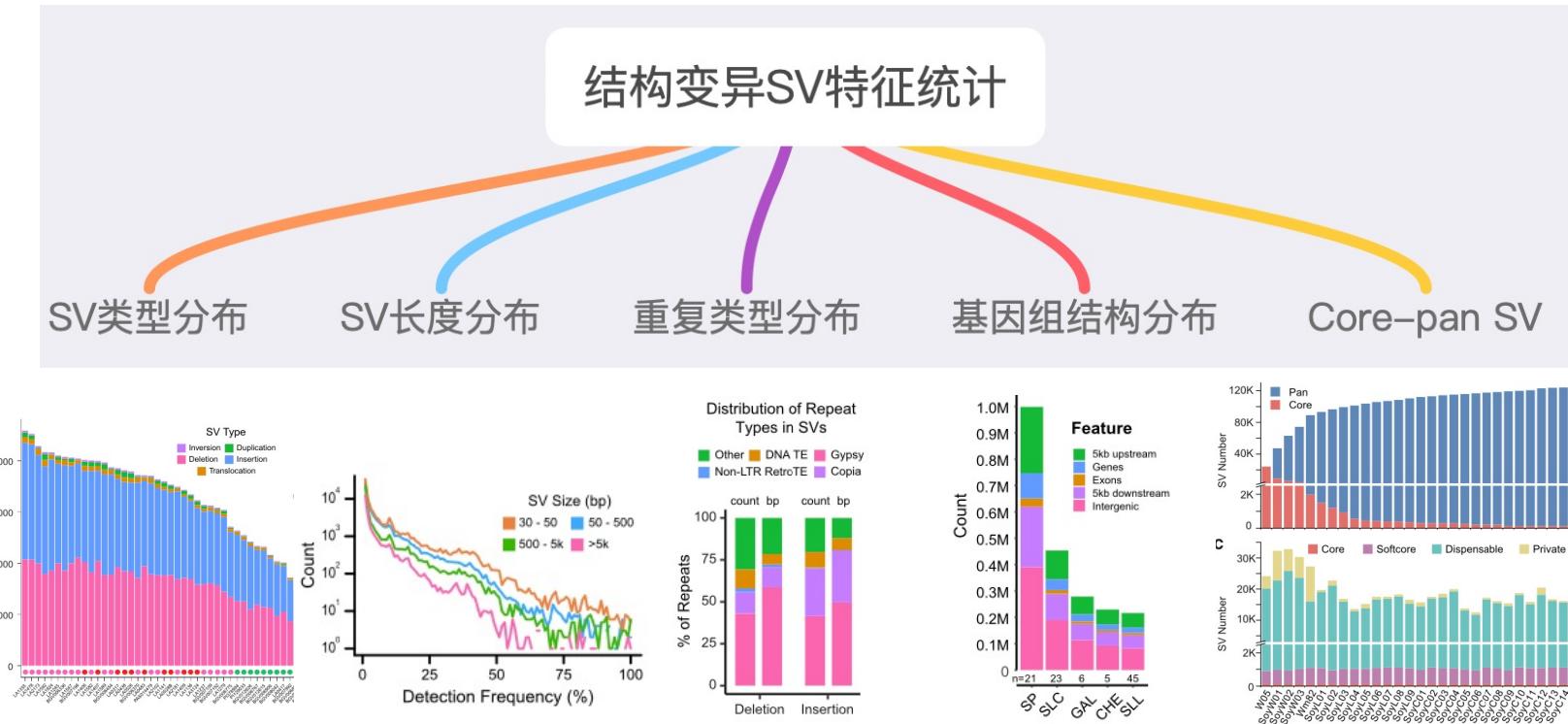
基于群体间的遗传变异信息研究群体**遗传结构、遗传多样性、物种的形成机制**等。从分子层面深入研究该物种的进化历程。

物种适应性进化
物种驯化和改良
种群历史研究



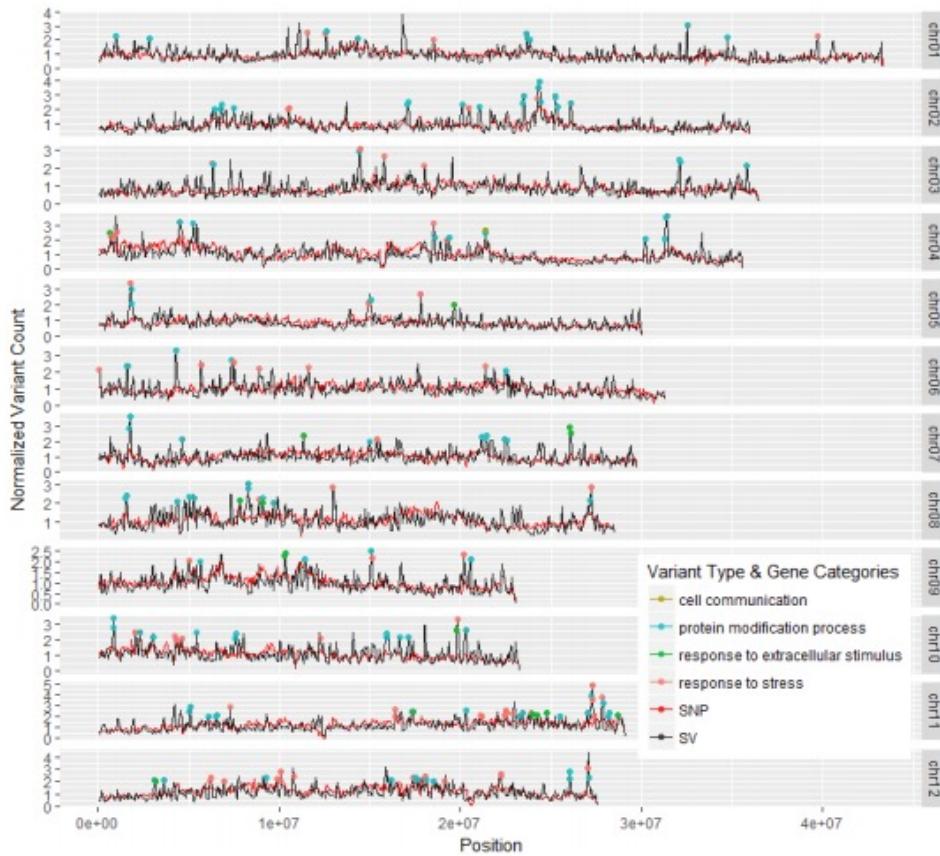
Nat Genet. 2019 Oct;51(10):1549-1558

研究内容--遗传多样性的扩充：panSV



- 通过构建panSV，可以按照在样本中的分布比例进行核心、次核心、非必需和私有SV
- 对SV进行注释和统计，以查看sv种类的分布及其在基因组结构中的分布类型

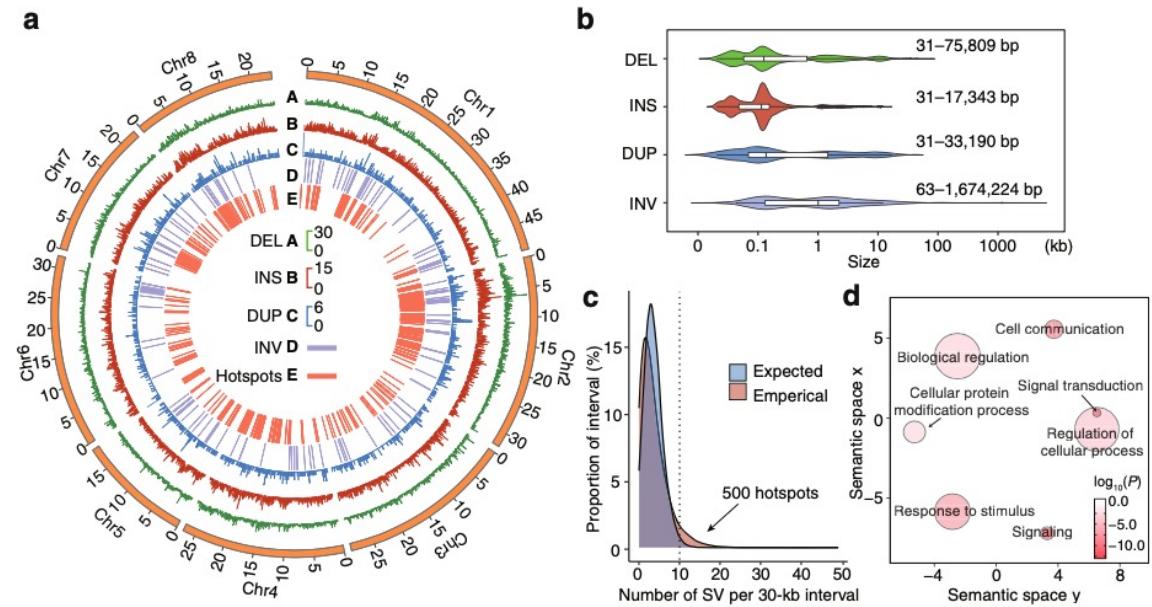
研究内容--遗传多样性的扩充-案例



重要结论：3K水稻SV图谱研究发现SVs与蛋白修饰，胁迫反应基因高度共定位，表明SVs与逆境胁迫相关的基因具有较高的关联性。

Genome Res. 2019 May;29(5):870-880

149份桃的SV研究



重要结论：构建高质量的SVs图谱，SVs热点往往处于节段重复区域且在信号转导，刺激响应，信号传递，细胞通讯等通路富集。研究表明SVs可促进桃子的抗性和环境适应力。

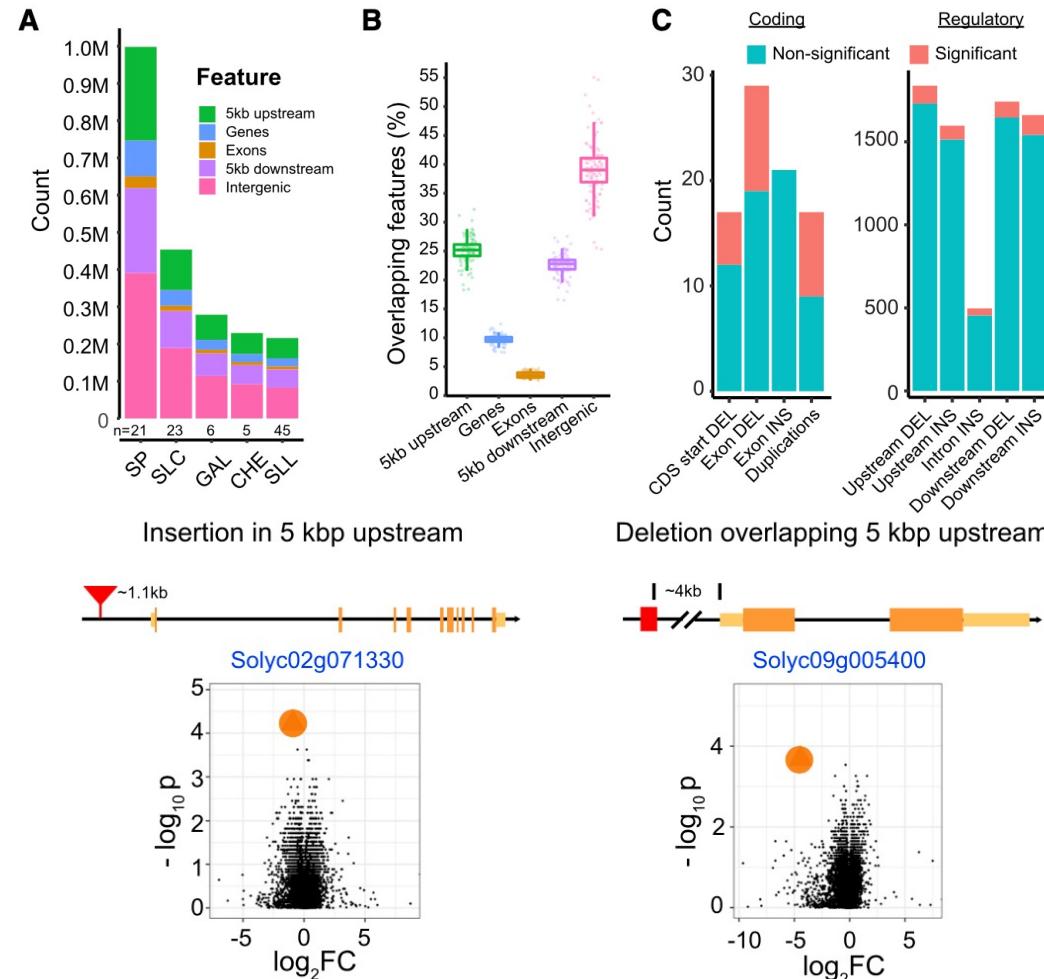
Genome Biol. 2021 Jan 5;22(1):13

研究内容--SV对基因表达的影响

SV分布
SV在基因组中的分布规律（编码区/调控区）

不同区的SV对基因表达是否有显著影响
基因表达

作物性状
受SV影响的基因是否显著改变作物的性状



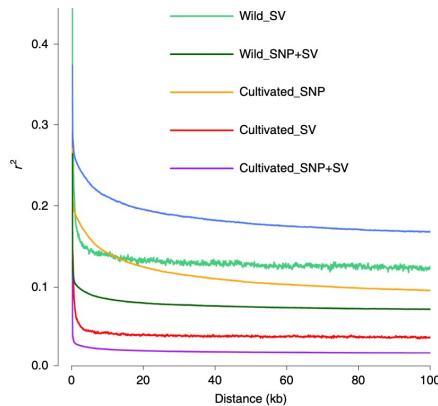
重要结论：~50%的SVs与基因区或其侧翼调控序列重叠，约95%的基因在其编码序列的5kb内至少有一个SV，大多数位于顺调控区。评估SV对基因表达的影响，发现数百个显著的基因表达改变。

研究内容--群体遗传结构：基于SV的群体分析

群体
结构

基于SV的群体结构分析：通过SV有无矩阵的聚类发现，100份
番茄品种可以分为野生、早期驯化、栽培种等不同类群

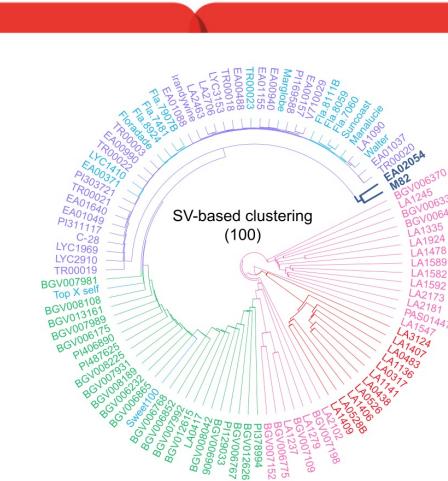
Cell. 2020 Jul 9;182(1):145-161.e23.



选择
消除

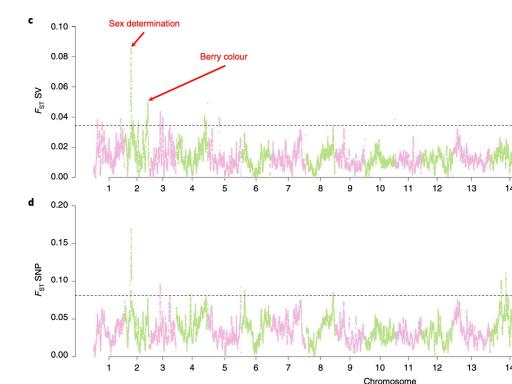
选择性分析： F_{ST} 研究发现2号染色体两个峰值，分别与性别决定
和浆果颜色相关。

Nat Plants. 2019 Sep;5(9):965-979

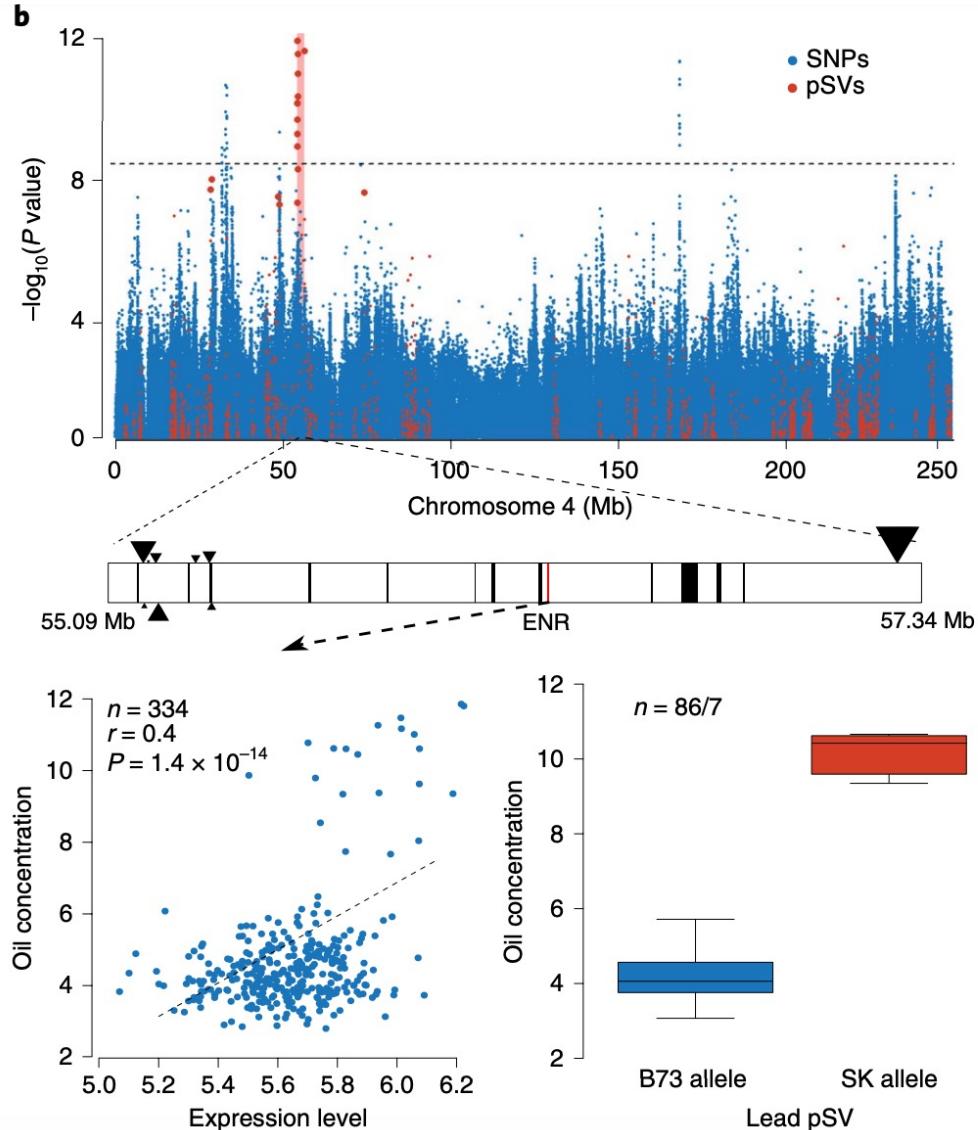


LD
衰减

SV的LD衰减：葡萄种群中，SVs通常比SNP的衰减更快，SVs在
驯化过程中比SNP经历了更强烈的纯化选择



研究内容--揭示表型关联的SV：SV-GWAS

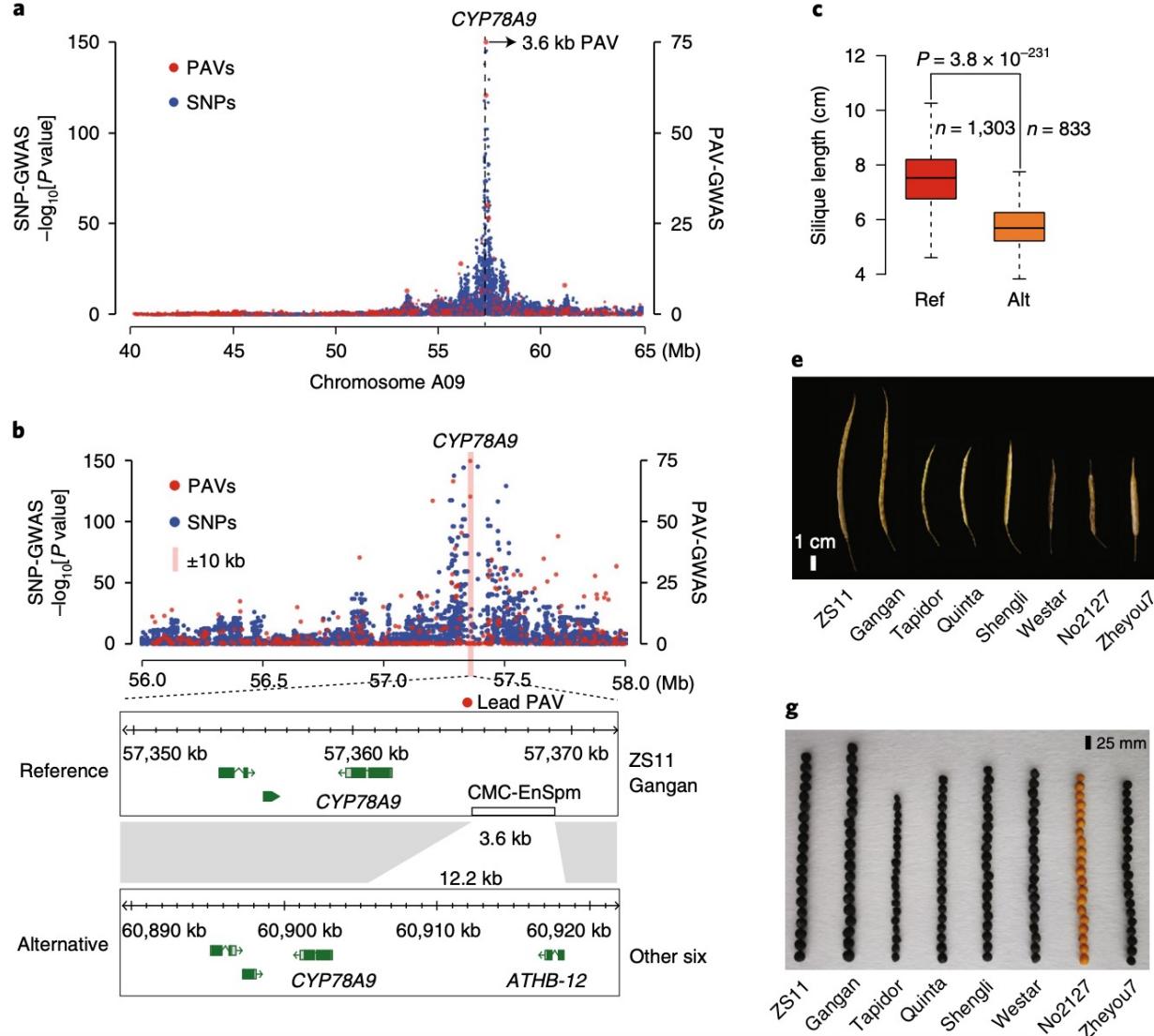


- SV-GWAS发现基于SNP-GWAS未发掘的玉米含油量相关的SV

重要结论：基于SV-GWAS结果发现了一个新的位于4号染色体上的显著相关的区域，验证结果表明是SVs造成玉米含油量的显著差异。

Nat Genet. 2019 Jun;51(6):1052-1059.

研究内容--揭示表型关联的SV：SV-GWAS

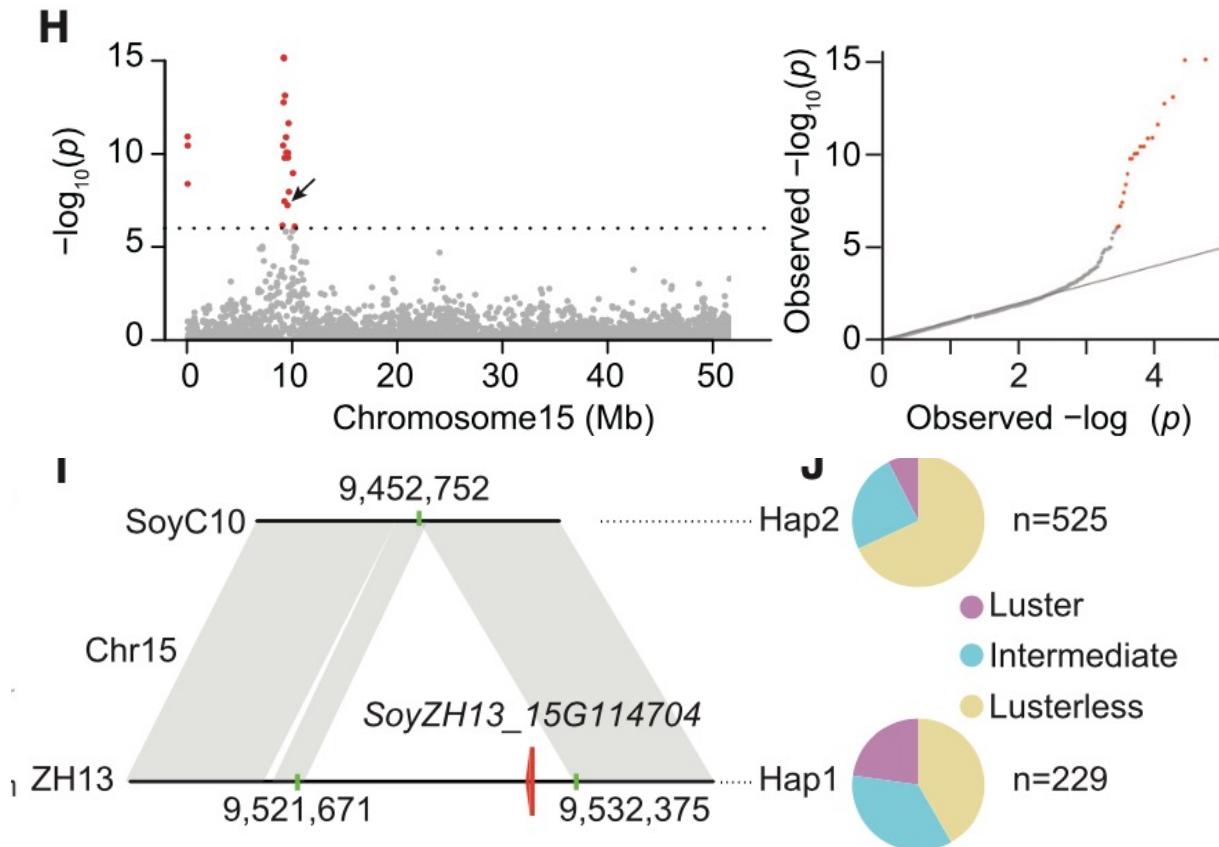


- PAV-GWAS挖掘到SNP-GWAS未发现的油菜果实长度、种子重量和开花时间相关的SV

重要结论：PAV-GWAS定位到了与角果长度、种子重量和开花时间相关的结构变异，而SNP-GWAS没有检测到这些信号，表明PAV-GWAS在确定与性状的关联方面与SNP-GWAS互补。

Nat Plants. 2020 Jan;6(1):34-45

研究内容--揭示表型关联的SV : SV-GWAS



- PAV-GWAS揭示结构变异对大豆种子光泽的影响

重要结论：PAV-GWAS确定了15号染色体上的一个重要信号，其中一个10kb的PAV导致了一个HPS编码基因的存在和缺失是控制大豆种子光泽变化的因果遗传变异之一。

Cell. 2020 Jul 9;182(1):162-176.e13

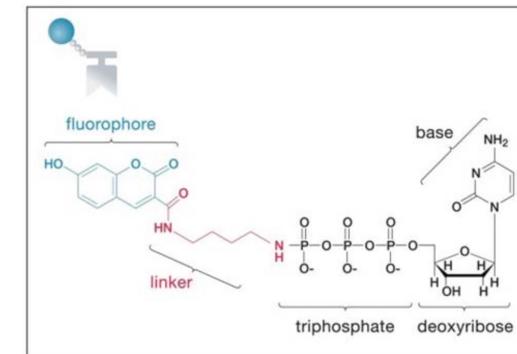
PB vs ONT-- PacBio测序

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。

◆两点关键创新：分别是零模波导孔（zero-mode waveguides, ZMWs）和荧光标记在核苷酸焦磷酸链上（Phospholinked nucleotides）。



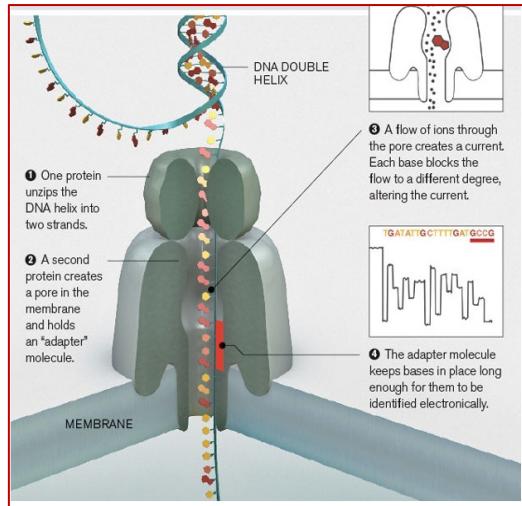
SMRT Cell含有纳米级的零模波导孔，每个ZMW都能够包含一个DNA聚合酶及一条DNA样品链进行单分子测序，并实时检测插入碱基的荧光信号。ZMW是一个直径只有10~50 nm的孔，当激光打在ZMW底部时，只能照亮很小的区域，DNA聚合酶就被固定在这个区域。



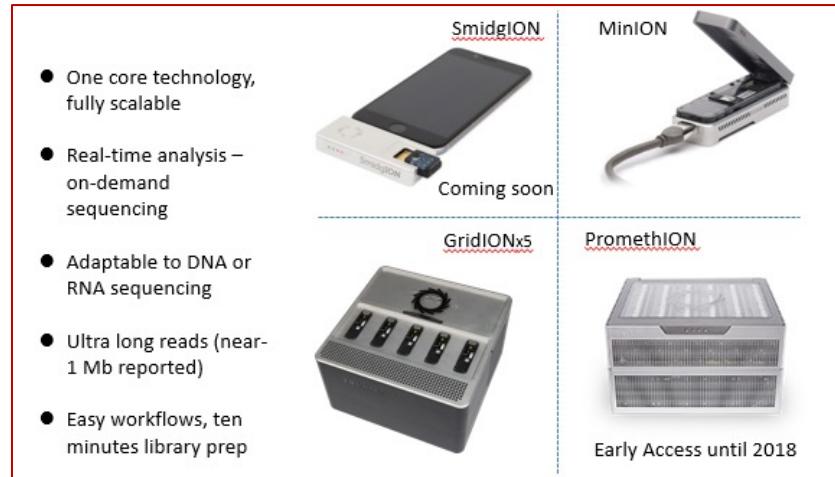
将荧光染料标记在核苷酸的磷酸链而不是碱基上，当核苷酸掺入到新生的链中，标记基团就会自动脱落，减少了DNA合成的空间位阻，维持DNA链连续合成，延长了测序读长。

PB vs ONT-- ONT测序

PacBio和Nanopore都属于三代测序技术，一个是采用单分子荧光测序，一个是采用纳米孔测序。两项技术都是不经过PCR扩增构成，并且获得长片段的序列信息。



Nanopore测序原理



Nanopore设备类型

Nanopore即纳米孔单分子测序法，采用电泳技术，借助电泳驱动单个分子逐一通过纳米孔来实现测序的。由于纳米孔的直径非常细小，仅允许单个核酸聚合物通过，而ATCG单个碱基的带电性质不一样，通过电信号的差异就能检测出通过的碱基类别，从而实现测序。

PB vs ONT-- 优劣势总结

PacBio优点：

- 1、PacBio模式选择更灵活，既可以选择追求长读长CLR模式，又可以选择高保真数据的HiFi模式。
- 2、PacBio测序错误是随机的，数据准确性可通过提高加大深度得到提高。
- 3、应用PacBio测序发表的文章数量远多于ONT。
- 4、PacBio HiFi模式是目前基因组组装的最佳选择，同时具有高准确性，高完整性，高连续性的特点，并且无需二代纠错，组装效率高，能够解决超大、高复杂基因组组装难题。
- 5、PacBio 在结构变异检测上的表现优于ONT

PacBio缺点：

- 1、深度不够时，数据准确率低。

Nanopore优点：

- 1、读长更长，连续性更好，可用较低的深度完成测序组装，适合非高杂高重的基因组。
- 3、可以直接检测DNA/RNA修饰。

Nanopore缺点：

- 1、Nanopore样品要求较高，普通模式要求DNA主带40K以上，超长模式需要100k以上。
- 2、Nanopore普通模式读长与Pacbio持平，超长模式测序时读长较长，数据产出不稳定。
- 3、Nanopore测序存在固定错误，无法通过高深度测序纠正。
- 4、Nanopore测序具有偏好性，在同聚物和串联重复区域有较高的缺失错误，在高GC区域也存在大量的缺失和错配。
- 5、Nanopore原始组装版本错误率高，需要加入Pacbio/Illumina数据，结合软件进行多轮纠错。纠错时间偏长，计算资源消耗高。

贝瑞基因--仪器平台

作为国内最早引入PacBio Sequel II的公司，贝瑞基因目前自主拥有**19台**最新的PacBio Sequel II测序仪，相比于其他公司，在仪器数量上有明显的优势，贝瑞是目前国内规模最大的PacBio三代测序服务商。能够为科研人员稳定、高效的提供读长更长，通量更高、数据质量更优的三代测序服务。



国内最大的三代测序服务商



贝瑞基因--公司介绍



经营理念 Business Philosophy

以创新为基石，以品质为价值

贝瑞基因科技服务基于全面丰富的高通量测序平台，目前已建立了**35种**科研服务类型，可满足多种领域的独特实验要求。同时，开发特有的数据库资源，结合高效的云计算平台，为国内科学的研究提供全面、精准的分析服务。

北京贝瑞和康生物技术有限公司（简称贝瑞基因）成立于2010年5月，是致力于应用高通量基因测序技术，为生命科学研究提供整体解决方案的研发型高新技术企业。公司先后于2011年9月获得中关村高新技术企业认证，2012年12月获得国家高新技术企业认证，**2017年8月成功登陆A股主板上市。**





Thank You!



官方网站



官方微信

www.berrygenomics.com

TCGATCGA GATCGATCGATCGATCGATCG
GATCGATCGATCGATCGATCGATCGATCG
CGATCGATCGATCGATCGATCGATCGATCGATCG

