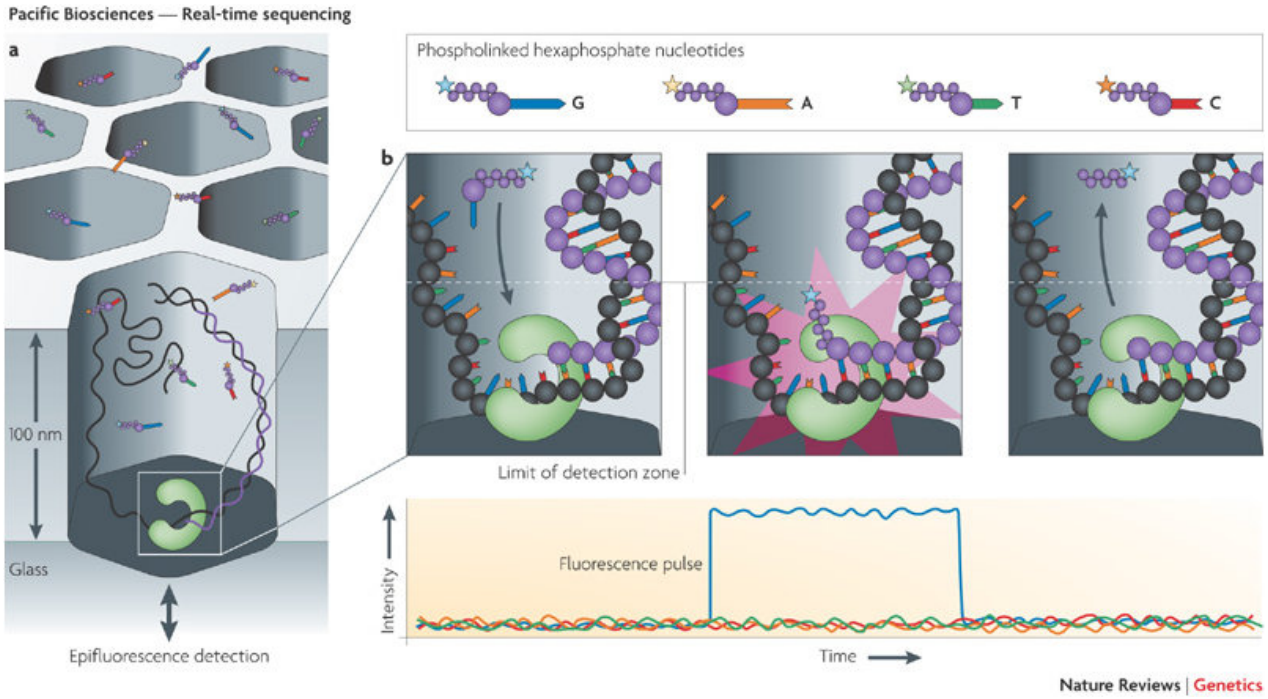
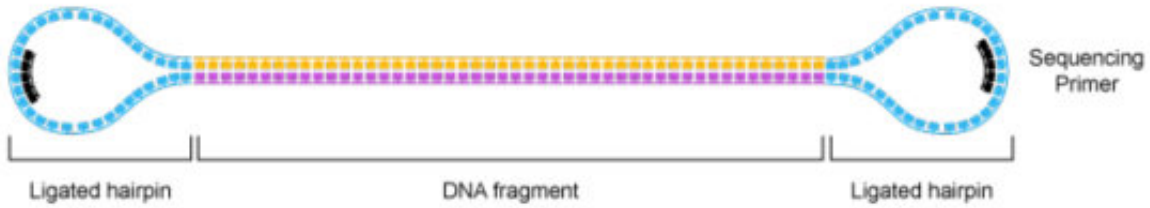
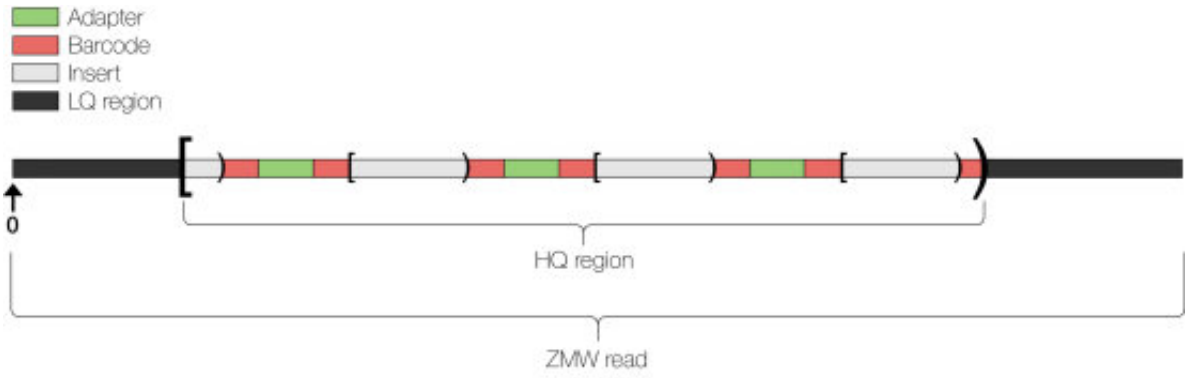
[PacBio][https://www.jianshu.com/p/15dd0baca47c]



PacBio数据的文库模型是两端加街头的哑铃型结构, 测序时会环绕着文库进行持续的进行, 由此得到的测序片段称为polymerase reads, 即一条含接头的测序序列, 其长度由反应酶的活性和上机时间决定.



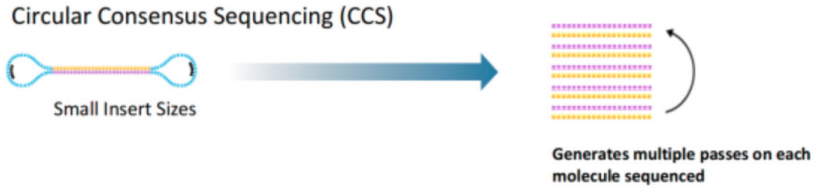
polymerase reads是需要进行一定的处理才能获得用于后续分析的数据的. 该过程首先去除低质量序列和接头序列:



处理后得到的序列称为subreads, 根据不同文库的插入片段长度, subreads的类型也有所不同.

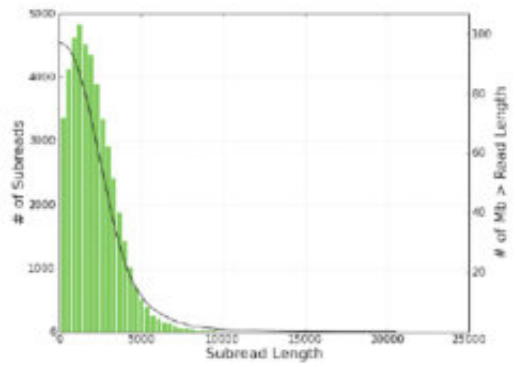
在用于基因组denovo组装时, 通常会构建10kb/20kb的文库, 对长长插入片段文库的测序基本时少于2 passes的(pass即环绕测序的次数, 两个adapters之间就是一个pass). 这时得到的reads也称为 Continuous Long Reads(CLR), 这样的reads测序错误率等同于原始的测序错误率.

而对于全长转录本或全长16s测序, 构建的文库插入片段较短, 测序会产生多个passes, 这时会对多个reads进行一致性校正, 得到一个唯一的read, 也称为Circular Consensus Sequencing(CCS) Reads, 这样的reads测序准确率会有显著的提升.

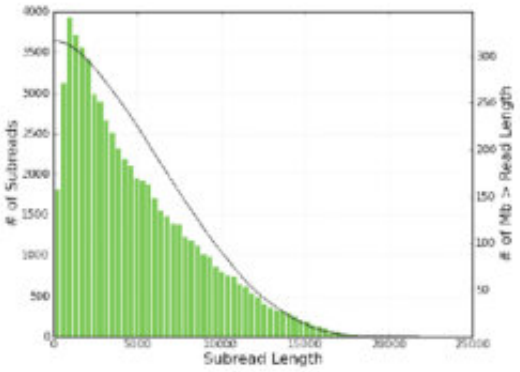


不同于二代测序的碱基质量标准Q20/Q30, 三代测序由于其随机分布的碱基错误率, 其单碱基的准确性不能直接用于衡量数据质量.

最直接的方法就是看长度. 长度短的测序数据不一定差(与文库大小有关), 但差的数据长度一定短. 上游实验环节, 最关键影响因素时文库的构建, 高质量的文库产出的数据长度长, 质量好; 而低质量的文库产出的数据长度短, 质量差. 另外其测序错误时随机的, 不像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纠错.

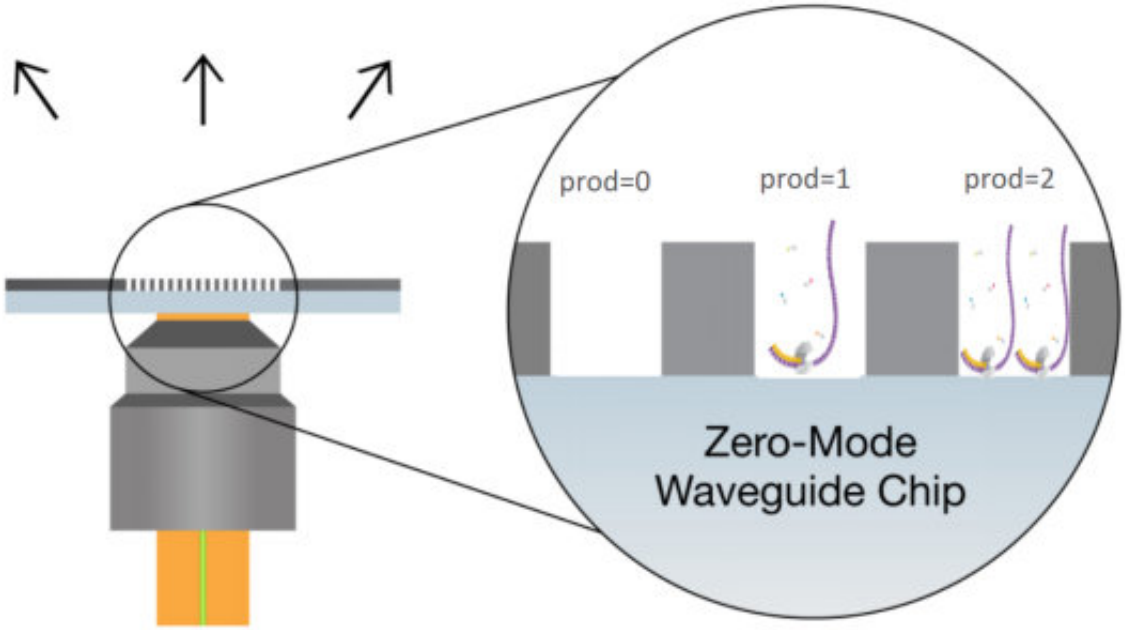


Degraded gDNA

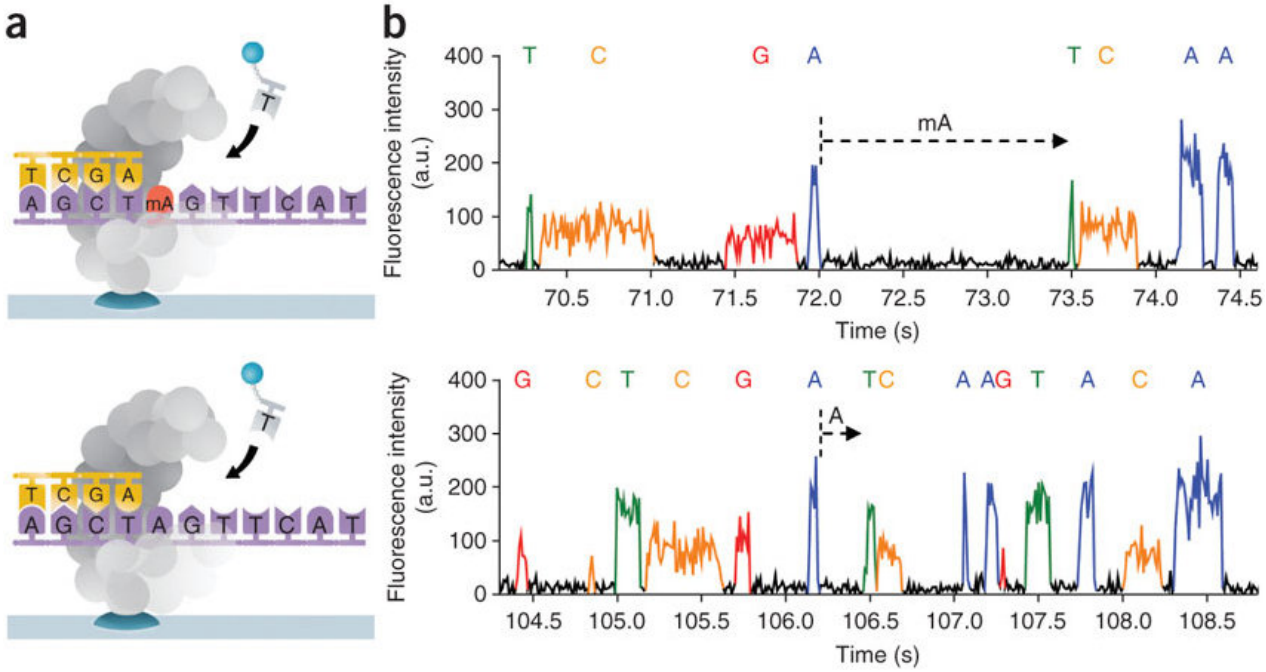


High Quality gDNA

其次, 看比率. 需关注两个比例, 一个是subreads与polymerase reads数据量的比例, 比例过低反应测序过程中的低质量的序列较多; 一个是zmw孔载入的比率, 根据孔中载入的DNA片段数分为P0, P1, P2. P1合理的比例在40%-60%之间. 上样浓度异常会导致P0或P2比例过高, 有效数据量减少. 需要注意的是P2比例过低时, 可能存在P2转P1的情况, 测序结果包含较多的嵌合型reads.



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.



组装

PacBio-only de novo assembly: 只使用PacBio产生的long reads进行拼接, 在拼接之前要进行预处理, 然后采用Overlap-Layout-Consensus算法进行拼接

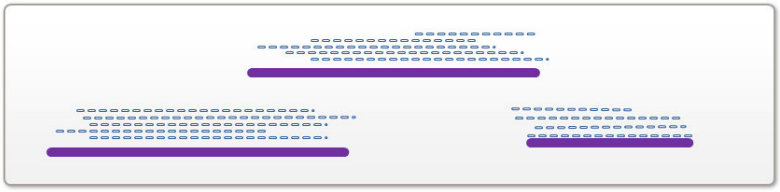
Hybrid de novo assembly: 结合PacBio的长reads和二代的短reads

Gap filling: 使用二代的短reads拼接得到的scaffold, 然后用PacBio的长reads进行补洞

Scaffolding: 用二代的短reads拼接得到的contigs/scaffold, 用PacBio的长reads确定contigs/scaffold之间的位置关系

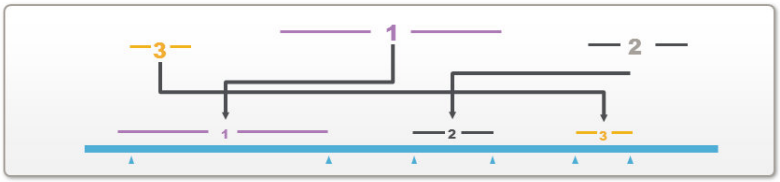
De novo Assembly

Complete genomes using only PacBio reads or combine technologies



Scaffold

Establish framework for genome and resolve ambiguities



Span Gaps

Polish genomic regions with up to 10x improvement

不同组装策略可选用工具:

PacBio-only

HGCA: 先进行reads的预组装(preassembly), 然后用Celera Assembler进行进一步组装, 最后用Quiver进行校正

Canu: 以Celera Assambler为基础, 为三代单分子测序而开发的分支工具

Celera Assembler: Celera Assembler 8.1已经可以直接用于subreads的组装

Hybrid

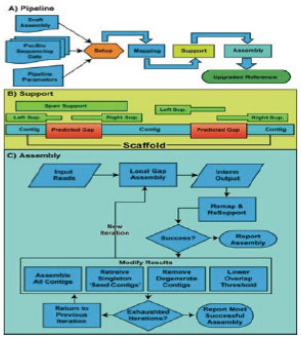
SPAdes: 3.0版本后增加了对PacBio的混合组装支持

Gap Filling

PBJelly2: 对已经组装后的基因组, 用PacBio的long reads进行补洞

PacBio Assembly Algorithms

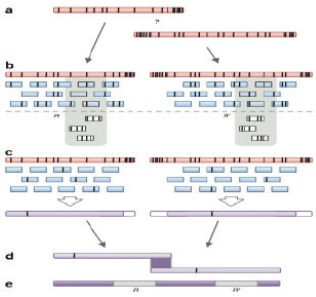
PBJelly



Gap Filling and Assembly Upgrade

English *et al* (2012)
PLOS One. 7(11): e47768

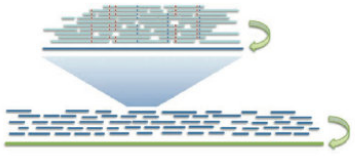
PacBioToCA & ECTools



Hybrid/PB-only Error Correction

Koren, Schatz, *et al* (2012)
Nature Biotechnology. 30:693–700

HGAP & Quiver



$$\Pr(\mathbf{R} \mid T)$$

$$\Pr(\mathbf{R} \mid T) = \prod_k \Pr(R_k \mid T)$$

Quiver Performance Results
Comparison to Reference Genome
M. ruber; 3.1 MB; SMRT® Cells)

	Initial Assembly	Quiver Consensus
QV	43.4	54.5
Accuracy	99.99540%	99.99964%
Differences	141	11

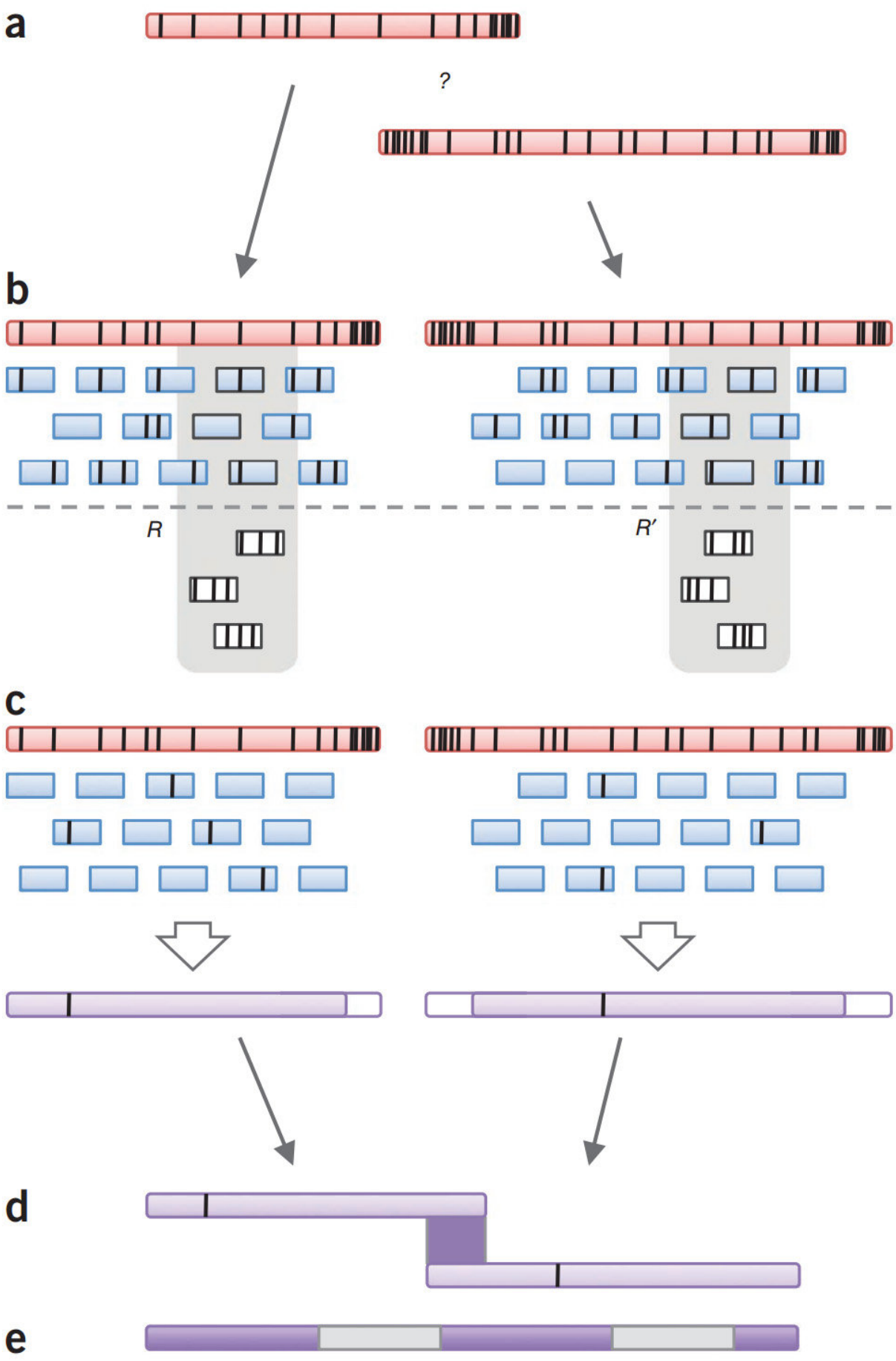
PB-only Correction & Polishing

Chin *et al* (2013)
Nature Methods. 10:563–569

< 5x PacBio Coverage > 50x

三代单分子测序会产生较高的随机错误, 平均正确率在82.1%-84.6%, 这么高的错误率显然不能直接用于后续的分析, 需要进行错误校正:

- 多测几个pass: 由于测序序列时发夹结构, 可以进行多轮的滚环测序, 靠覆盖度来自我交错, 如果通量不是限制因素, 那么PacBio时目前最准确的测序方式: 错误率可以无限接近罕见突变的发生率(即无法分辨时测序错误还是罕见突变), 不过这会极大缩短有效测序的插入序列长度
- 用二代的短reads校正: 2012年冷泉港实验室的Michael Schatz开发了一种纠错算法, 用二代测序的短读长高精确数据对三代读长数据进行接错, 这种称为"混合纠错拼接"(PBcR (PacBio corrected Reads)algorithm)
- Map short reads to long reads
- Trim long reads at coverage gaps
- Computer consensus for each long read



[甲基化][https://www.pacb.com/applications/epigenetics/microbial/]

DNA的甲基化是微生物对抗外源DNA的重要途径, 但是限制性修饰系统(Restriction-Modification R-M)也会通过增加双链断裂和C-T突变和改变影响毒力和致病性的基因的转录驱动细菌的进化

SMRT测序直接检出甲基化修饰(epigenetic modifications)

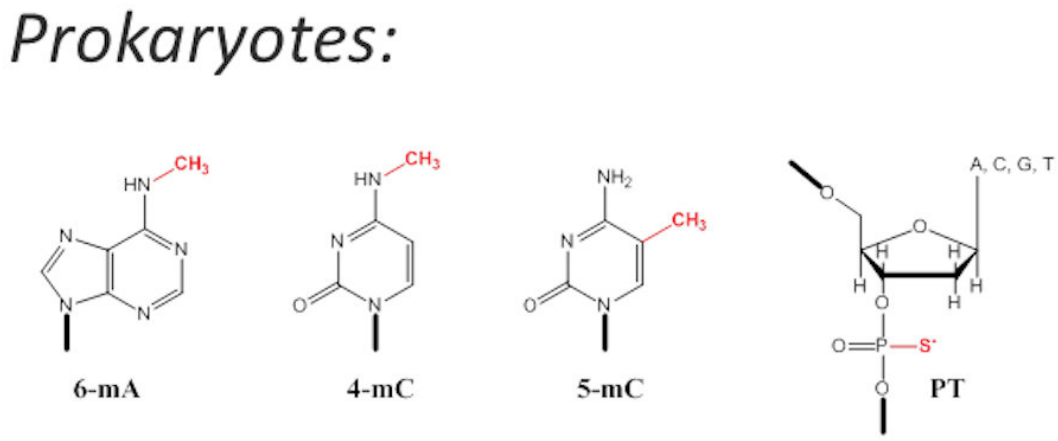
Single Molecule, Real-Time(SMRT) Sequencing 通过测量DNA碱基在合成时的聚合酶动力变化检出DNA修饰:

- 在组装覆盖度水平检测基因组范围内的m6A和m4A R-M系统motifs
- 在全基因组位置的marginally higher coverage水平, 检测m6A和m4C甲基化状态
- 获得甲基化修饰注视的全基因组
- 揭露涉及到致病性, 宿主适应性和抗生素耐药在不同阶段性R-M基因的变化

在遗传学上, 一个序列的基序(motif)为一个核酸或氨基酸序列模式, 广泛分布, 同时拥有或推测拥有显著的生物学意义. 在植物和其他生物中, DNA甲基化存在三种不同的序列组成: CG(or CpG), CHG or CHH(H指代A,T或C). 而哺乳动物, DNA甲基化几乎唯一出现在CpG二核苷酸中, 这里双链上的C(胞嘧啶, cytosines)都被甲基化. CpG岛常定义为: 长度大于200bp, G+C含量大于50%, 其中CpG比率大于0.6.

在细菌基因组中, N6-甲基腺嘌呤(m6A)和N4-甲基嘧啶(m4C)也很常见, 作为限制性修饰(RM)系统的一部分而行使功能.

目前PacBio测序仪可以分析的甲基化类型有: 5mC\4mC和6mA:



[Base Modification: From Sequencing Data to a High Confidence Motif List][https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/Base-Modification-:--From-Sequencing-Data-to-a-High-Confidence-Motif-List]

SMRT测序利用每一碱基添加过程中的动力学信息识别各个位置的碱基. 通过比较 该信息也可用于识别修饰后的和原始碱基(by compare results of SMRT Sequencing to an in silico kinetic reference for dynamics without modifications). 不同类型的修饰具有独特的印记(footprints with varying intensities), 常围绕被修饰碱基跨越多个位置.

Experiemental Design for Bacterial Modification

- For de novo assembly, you will need 60-100X coverage for HGAP
- Complete your assembly first, the upload the reference into SMRT Portal
- SMRT Portal v2.1 or v2.2 using RS_Modification_and_Motif_Analysis and the in silico reference.

完成以上分析将会得到初始的motif列表作为起始点. 如果motif列表中的motif具有较低的修饰比例或不常见的序列, 可通过以下来改善motif结果:

- Rerunning the job with the a more appropriate QV setting
- Using Motif Maker with a more appropriate min QV setting
- Using R to refine the list of 'hits' used for motif finding and using Motif Maker

Apply what you know about base modification signals from SMRT sequencing and our motif finding algorithm

m6A 在修饰的A碱基上游5个碱基处会出现显著的信号

5mC 在修饰的C碱基上游2个碱基出会出现最强的信号

通过所知的细菌restriction/modification系统, 在细菌系统中motifs的修饰一般为100%; motifs常为回文结构, 同时在两条链时以反向互补形式存在.

Make use of public databases on the methylation systems

[REBASE][http://tools.neb.com/genomes/]

[Methylome Analysis Technical Note] [https://github.com/PacificBiosciences/Bioinformatics-Training/wiki/Methylome-Analysis-Technical-Note]

Detecting DNA Base Modifications: SMRT Analysis of Microbial Methylomes

Background

微生物基因组包含多种碱基修饰, 大部分发生在腺嘌呤或胞嘧啶位置的甲基化. 这些甲基化典型地源自RM(restriction-modification)系统, 该系统作为微生物的防御机制, 保护细胞不受噬菌体或其他外源DNA的侵入. 一般而言, RM系统包含一个限制性酶(endonuclease, 核酸内切酶)和一个甲基化酶, 指向共同的序列motif. 核酸内切酶识别并裂解motif序列从而降解外源DNA. 细菌自身的DNA motif通过甲基化而被免除降解(防止被限制性核酸内切酶裂解). DNA修饰也用来控制细菌的其他生物过程, 例如细胞循环, DNA复制, 错配修复, 基因表达和致病性.

不同种类细菌的被修饰的motif变化非常大, 且大多数物种包含不知一个RM系统. SMRT在特殊的细菌中, 可以在基因组范围内检测出7个修饰位点. 因此也可用于判断特殊甲基转移酶的修饰位点.

The method used to calculate IPD ratios

IPD: interpulse duration ratios, 对应在前一个碱基完成结合后新碱基结合到测序聚合酶活性位置所需要点时间. 计算样本中的感兴趣IPD和参考样本中的IPD的比率为IPD ratio.

默认分析模式, 是使用聚合酶动力计算模型(polymerase kinetics computational model)来计算IPD ratios. 该计算模型使用silico control计算的. IPD针对指定碱基的结合是基于跨越约12个碱基的序列内容, 这匹配DNA聚合酶的'binding footprint'.

当使用in silico control时, 检测准确性可能通过activating 修饰的识别而增加(detecting accuracy may be increased by activating modification identification in SMRT Analysis). 该分析比较修饰信号和额外的计算模型(of the expected positive signature), 检测三种修饰类型:6-mA, 4-mC,和Tet-converted 5-mC. 当前, SMRT分析仅支持使用in silico control的修饰识别.

最后, IPD ratios也可通过比较两个不同的样本来计算. 这种情况下, 两DNA样本分别测序然后互相比对, 检测不同的修饰, 该修饰可来源不同的生长条件, 或细菌菌株之间的比较. 使用该分析方法, 菌株共有的修饰将不会被检出. 为实现识别的修饰和motifs的差异分析, 最简单是先使用in silico control执行分别分析, 然后再比对两分析结果.使用native DNA control将会有助于定位差异修饰的位置, 但是不兼容于当前SMRT分析过程.

The particular modification you are analyzing

由于kinetic signatures之间的不同, 针对不同的修饰类型所需的覆盖度不同. 例如, N6-methyladenine(6-mA)和N4-methylcytosine(4-mC)提供了强kinetic signals, 然而5-methylcytosine(5-mC)的kinetic signals较弱. 因此检测5-mC需要更高的覆盖度以达到可信的检出. 通过使用Tet酶将5-mC氧化为5-caryboxylcytosine(5-caC), 其修饰信号可增强到和6-mC, 4-mC一样的水平.

例如, 6-mC, 4-mC或Tet-converted 5-mC的可信检出需要约25X覆盖度(每条链), 然而由于更小和更分散的5-mC kinetic signature, 因此需要至少10倍的覆盖度(250X每链)才能用于检出. **Pacific Biosciences - Detecting and Identification of Base Modifications with Single Molecule Real Time Sequencing Data**

由于贯穿基因组的SMRT测序覆盖符合柏松分布, 因此推荐使用约100X的总覆盖度以确保最低每链25X的覆盖度.

The size of the genome

针对5Mb $E.coli$基因组, 需要使用PacBio RS II 约2 SMRT Cells run才可以实现6-mA, 4-mC或Tet-converted 5-mC可信检出. 该情况下, 每个SMRT Cell将估计产生约200Mb的比对序列(PacBio RS通量减半, Cell数需增倍).

Whether you are using Tet convertion to indentfy 5-mC

由于当前使用Tet-conversion 5-mC对4-mC具有脱靶影响, 因此, 若需要同时检测4-mC和5-mC， 需要分开测序.

With P6-C4 chemistry, high coverage(minimum 250X per-strand coverage is recommended) motifs will still be detected as 'modified' but may not be correctly classified(identified) as 5-mC.

Methylome Analysis - RSII and SMRT Analysis 2.3.0 or older

使用SMRT Portal采用RS_Modification_and_Motif_Analysis.1直接分析甲基化和motif情况:

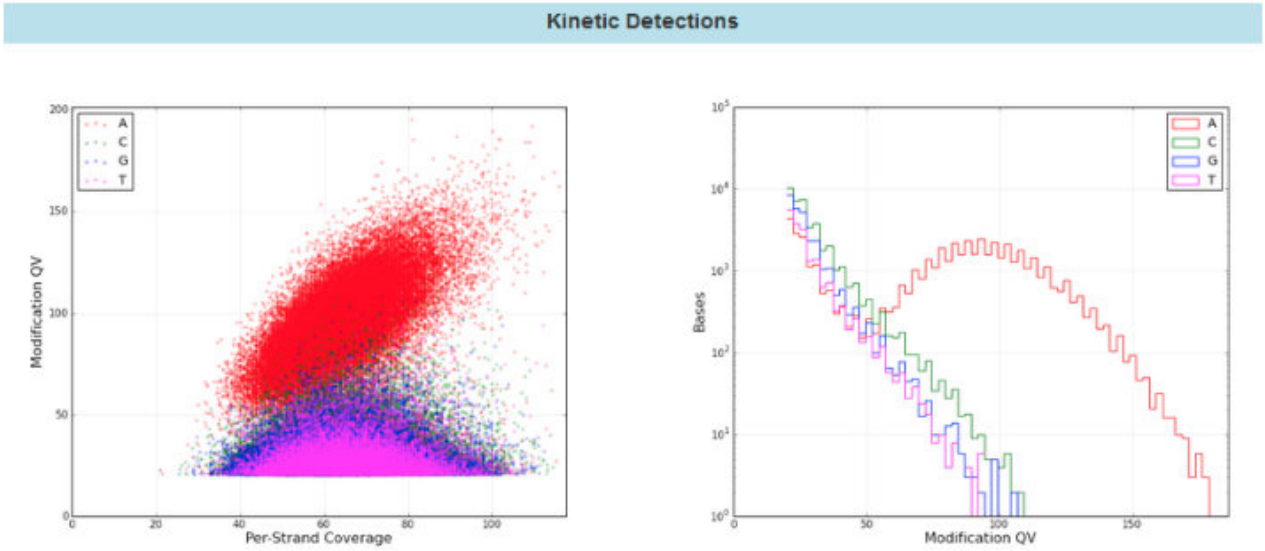
1. 比对SMRT测序subread到参考基因组, 生成cpm.h5比对文件
2. 检出变异, 生成GFF轨迹和VCF轨迹文件
3. 保存数据为相同格式, 用于计算IPD ratios
4. 生成modifications.csv文件和modifications.gff文件. 这些文件包含测序过程中聚合酶动力信息(statistics on the polymerase kinetics during sequencing at every position in the genomic sample). **文件中高的IPD ratio位置代表假定修饰位置(locations of putative modification). 相同的修饰会在15个碱基的footprint(-10 to +4)内的多个位置具有高的IPD ratios**
5. 在基因组范围内分析重现的修饰内容, 构建modified motifs报告(motif_summary.csv)
6. 使用来自第四步的信息识别6-mC, 4-mC和Tet-converted 5-mC位置, 合并这些信息到第五步的motif信息. RS_Modification_and_Motif_Analysis.1需要参考基因组. 因此在SMRT Portal运行前上传参考基因组. 如果用于比较两de novo组装, 其中作为参考的组装需先完成组装. 这里参考基因组的低质量的或变异序列区域将会导致明显的修饰检出(since the true sequence context of the base calls will not match the expected sequence context that the in silico control is using)

Setting Up the Job

Output Files

RS_Modification_and_Motif_Analysis.1 分析流程生成二个报告和四个数据文件. 两个报告为1)Modificaitons和2)Motifs. Modifications表明根据两graphics, 这些碱基具有modification

Figure 3. Modifications Report



四个文件为:

1. `modifications.csv`为参考中每个位置的统计分析. 该文件为进一步的分析准备. 当分析subreads时, 所有subreads中所有IPDs都根据当前subread的平均IPD标准化, 这解决了IPDs中read-to-read的变异.
2. `modifications.gff`为Feature Format(GFF)文件. 该GFF文件用于motif分析和modification的可视化(in SMRT View). GFF文件时用于序列可视化查看使用. 包含根据满足modQV 20或更高的假定修饰位点(p-value为0.01或更低), 这表明该位置的IPD ratio显著性区别于背景信息. 因此选择合适的modQV阈值很重要.
3. `motif_summary.csv`包含甲基转移酶基识别的motif
4. `motifs.gff`类似`modifications.gff`, 但是在motif分析开始时生成的. 该文件包含所有检测为modified, motif和modifications和motifs重叠的位点信息(all sites detected as modified, all locations of a discovered motif including those which are apparently unmodified, and also the overlap between the modifications and motifs).

`motifs.gff`文件格式

- Seqid 参考序列标签
- Source 分析软件'kinModCall'
- Type 修饰类型'modified_base'为未定义的碱基.
- Start 修饰起点
- End 修饰终点
- Score -10log(p-value), 默认20为最小阈值(p=0.01).
- Strand 样本链方向'+'表示和原始fasta方向相同, 否则相反

- Phase Not applicable
- Attributes IPDRatio为IPD Ratio, 为围绕修饰位点-20bp到+20bp的参考序列内容, 修饰位点位于21位置, 同时还有序列在该位置的覆盖度. 序列为模版链的5'-3'方向. 同时若modificaton type被判定了, identificationQv值用于描述该修饰类型的可信度(计算方式类似Score)

```
context=GCCGTGGCGAGAAAATGTCGATCGCCATTATGGCCGGCGTA;
fracLow=0.798;motif=GATC;coverage=28;IPDRatio=8.04;
id=GATC;fracUp=1.000;frac=1.000;identificationQv=59
```

`motif_summary.csv`文件格式

- motifString 该位置检出的motif序列
- centerPos motif中修饰位置(0-based)
- modificationType 修饰类型, 'modified_base'用于unidentified bases; 针对identified bases, 为m6A, m4C, m5C
- fraction 在基因组中该motif检出未modified的时间百分率
- nDetected 该motif被检出未modified的次数
- nGenome 在参考基因序列中该moitf发生的次数
- groupTag 用于识别完整双链motif的名称
- partnerMotifString 配对motif的motifString(motif with reverse-complementary motifString)
- meanScore 该motif被检出未modified的平均Modification QV
- meanIpdRatio 平均IPD ratio, 同上
- meanCoverage 平均覆盖度, 同上
- objectiveScore motif finder 算法中该motif的分值

Performing Motif Analysis

The other alternative is to download the [motifmaker] [https://github.com/PacificBiosciences/MotifMaker] command-line Java program from GitHub. It is quicker to run because it uses the outpu of the SMRT Portal modification analysis as an input.