基因组序列组装

1. 序列组装的基本理论

1.1.流程：

多个基因组测序的副本 -》碎片化（fragments） -》长度过小的过滤 -》BAC/YAC 双末端测序序列（带有 BAC 末端的序列，reads） -》通过 overlap 寻找重叠群区域（contigs） -》锚定在染色体上的重叠群（scaffold） -》草图序列（可覆盖测序克隆片段 3-4 倍的 DAN 序列，含间隙/没有间隙，排列方向和位置未定） -》完成序列（错误碱基数<0.01%的 DNA 序列，排列方向确定，内部不含间隙，测序覆盖率在 8-10 个单倍体基因组）

Ps：根据确定 BAC 的排序方向以及重叠群（contigs）在支架 scaffold 中的排列方向; 酵母人工染色体（*YAC*）、细菌人工染色体（*BAC*）

1.2.问题：

测序错误

重复序列

多态性变异-》contigs

倒位 inversion

覆盖率

1.3.组装类型：

- 从头组装（De novo） vs 基于参考的组装（ reference-based(mapping)）：

1.3.1. de novo 组装：

即使可以获得参考基因组，也应该进行从头组装，因为它可以从基因组组装中丢失的基因组片段中恢复转录的转录本。

1.3.2. mapping 组装：

对现有的主干序列进行读取，构建一个类似的序列，但不一定与主干序列相同。转录组数据主要通过对参考基因组的 mapping 进行分析

1.3.3. mapping 缺点：

无法解释 mRNA 转录本结构改变的原因,如可变剪接。由于基因组包含可能存在于转录本的所有内含子和外显子，因此在基因组中不连续排列的剪接变体可能被折现为实际的蛋白质亚型。

- 基因组组装 vs 转录组组装：

1.3.4. 基因组：基因组序列覆盖水平可以根据 non-codingDNA 内含子区域的重复序列随便改变；这些重复序列也会造成基因组组装重叠群 contigs 组成的错误；

1.3.5. 转录组：转录组的覆盖水平可以表示为基因表达水平；转录组装中的重叠群 contigs 区域可能是剪接的异常或者基因家族成员之间的差异

1.3.6. 基因组组装软件不能被用于转录组组装

一个基因组的基因组测序深度通常相同，但转录的深度不同；

两条链在基因组测序都是按顺序排列的，但 RNA-seq 可以是特异的；

来自相同基因的转录变异体(transcript variants)可以共享外显子，难以处理

2. 序列组装的软件：

TIGR 组装

Minimus 组装：

Minimus2 组装：

Minimo
Newbler
BioPerl -》 Module:Bio::Assembly::IO
AllPathsLG -》 DISCOVAE de novo
Velvet https://www.ebi.ac.uk/~zerbino/velvet/
SOAPdenovo http://soap.genomics.org.cn/soapdenovo.html
SOAPdenovo ：SOAPdenovo2（基因组）、SOAPdenovo-Trans（转录组）
Trinity（转录组）
CAP3
应用于转录组的序列组装：
SeqMan NGen
SOAPdenovo-Trans
Velvet（基因组很小的 reads）-》Oases（应用于转录组）
Trans-ABySS
Trinity

3. CAP3 小规模序列组装 http://doua.prabi.fr/software/cap3
使用 genebank 的 unigene 数据库中搜索的某个基因的 EST 序列进行组装
4. 获得某个基因的 EST 序列：
NCBI-UniGene -》 某基因 -》下面有 mRNA 序列和 EST 序列（fasta）
5. 序列组装算法：
5.1.从头组装（De novo）
贪婪图算法：OLC、DBG
给定一组 reads ，从中挑选一个 read 作为“种子”【规则】，用与它两端中 的一段有足够数量的碱基序列相同的 read 来扩展；迭代进行，直到不可继续 扩展。再选择其他未参与拼接的 reads 序列拼 接，重复上述过程，直到所有 read 被 拼接完成。
计算多有 fragments 的成对对齐程度
-》选择最大重叠的两个 fragments
-》合并选择的 fragments
【不断迭代，中间的重叠区就是 contig】
5.1.1.OLC 方法【Sanger-data 组装】Hamilton path：
1.把每个 DNA 片段 (reads) 看成一个节点 ；
2.如果两个 DNA 片段之间存在重叠，就在相应的节点之间建立 一条边 ；
3.所有 DNA 片段通过这种重叠关联，构造出一个有向图 ；
4.通过寻找图中经过每个节点一次且仅一次的一条路径 （ Hamilton 路径） ；
5.即可获得目标 DNA 序列。
1.对参与拼接的 reads 进行比对，分析它们之间的重叠信息（ overlap ） ；
2.把存在重叠的 reads 进行组合，形成拼接结果 contigs （ layout ）；
3.对 contigs 形成的图上的 reads 进行排列，通过在图中寻找 Hamilton 路径来确定最终序列（ consensus ）。
应用的软件 Celera Assembler, Arachne, CAP，PCAP，TIGR， PHRAP……
5.1.2.DBG 方法 Eularian path：

1. 对给定的 reads ，按照长度 k 进行连续划分（步长 =1 ），得到若干等长度段序列（ k-mer)。一个长度为 1 的 read ，将被分成 lk+1 个 k-mer

2. 对于任意两个 k-mers ： k1 和 k2 ，如果 k1 的后 k-1 个碱基序列与 k2 的前 k-1 个碱基序列相同，则建立一条从 k1 指向 k2 的有向边。

通过以上两步即可构建出一个 de Bruijn 图，拼接结果序列可以通 过在图中寻找 Eulerian path 获得。

3. 第一个 k-mer 的序列全部读出，后面的每个 k-mer 只读取 最后一个碱基

5.1.3.DBG 方法的问题：

Q1.测序错误：tip 结构和 bubble 结构

错误的 reads-》错误的 k-mer 节点-》deBruijin 图大且复杂-》降低组装效率

Q2.测序 gap：

基因组覆盖不全-》k-mer 信息不全-》deBruijin 图连通性降低-》产生 dead-end 路径-》k-mer 越长问题越严重

Q3.分支问题：

数据错误/重复序列-》deBruijin 图出现分支-》无法处理

=》Solution：设定过滤标准（k-mer 出现次数）-》过滤掉可能出错的 reads；

直接删除 dead-end 路径-》可能导致 gap 问题；

5.1.4.对比 Hamilton path 和 Eularian path：

当短的 reads 数量很多时，Hamilton 图基于 reads，巨大复杂-》时间复杂度高-》空间复杂度低；Eularian 图基于 k-mer，不受影响-》时间复杂低-》空间复杂度高

5.2.基于参考的 mapping 组装

6. 大规模序列组装软件：

6.1.AllPaths-LG:短 reads

给定一个参考基因组，pipeline 能在基因组组装的不同阶段对组装过程 和结果进行评估

BASIC:基础评估，不需要参考基因组；

frag_size:小片段文库插入片段长度的均值；

frag_stddev:小片段文库的插入片段长度估算的标准偏差；

insert_size:大片段文库插入片段长度的均值；

insert_stddev: 大片段文库插入片段长度估算的标准偏差；

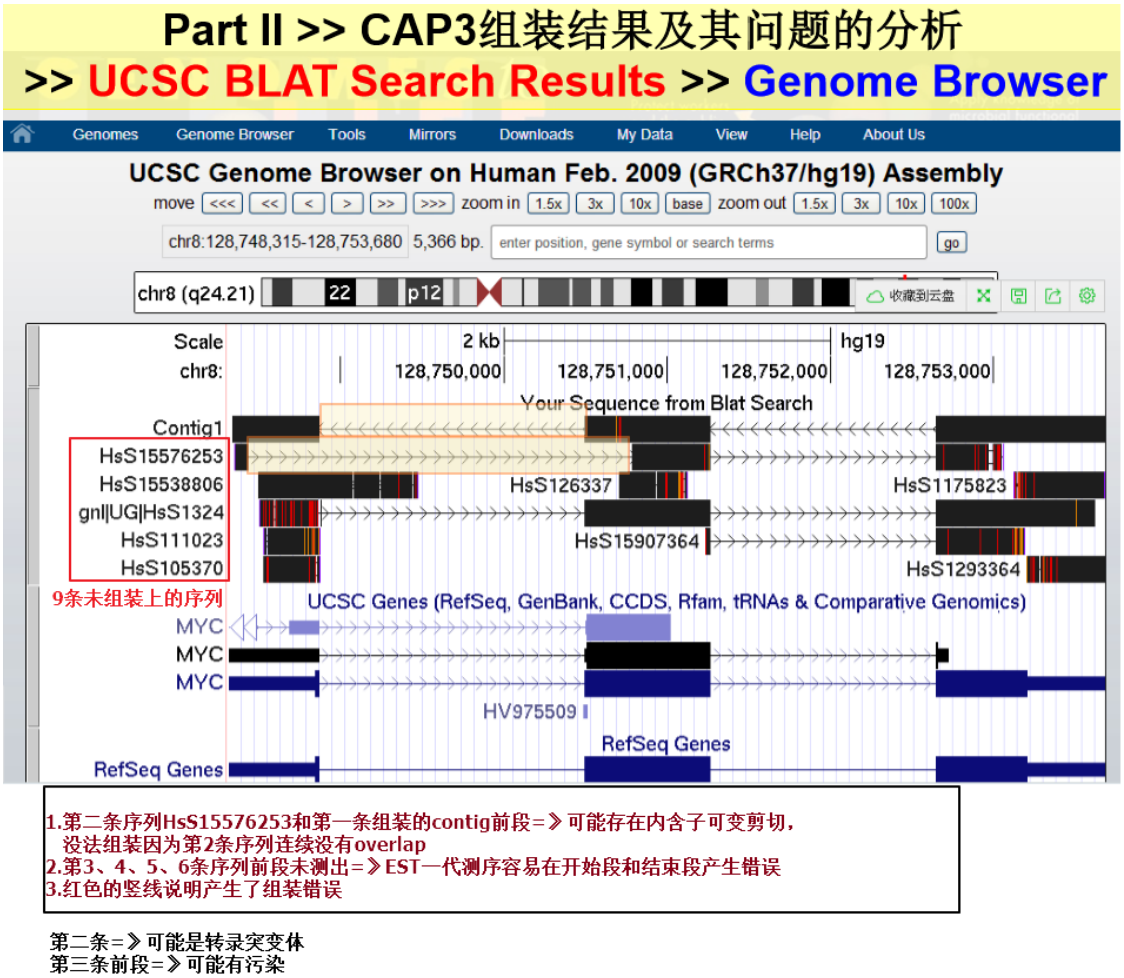
read_orientation:reads 的方向，小片段文库为 inward，大片段文库为 outward；

genomic_start:reads 从该位置开始，读入数据，如果不为 0，之前的碱基都被剪掉； genomic_end:reads 从该位置开始，停止读入数据，如果不为 0，之后的碱基都被剪掉。

6.2.Velvet + SOAPdenovo

7. CAP3 序列组装结果分析：

UCSC -》 Blat -》提交 cap3 生成的 contig 结果 -》browser



8. 序列比对基本原则：

相似性、同源性

8.1. 相似性：是指序列比对过程中， 用来描述检测序列和目标序列之间，相同 DNA 碱基或氨基酸 残基顺序所占比例的高低。 《= 统计

8.2. 同源性：是指从某一共同祖 先经趋异进化而形成的不同序列。 《=进化

8.3. 当相似程度高于 50%时，比较容易推测检测序列和目标 序列可能是同源序列；

当相似性程度低于 20%时，就难以确定或者根本无法 确定其是否具有同源性；

无论相似程度有多高或多低，都不能 100%确保两个序列 之间一定同源，它只能作为推测的依据之一。

8.4.

8.5. 在蛋白质家族或超家族中经常存在，不同的蛋白质序列之间只有局部区域存在高度相似性（保守性 Motif），而这些局部区域在整个序列中所占的比例有时很低！

序列比对过程中需要在检测序列或目标序列中引入空位， 以表示插入(Insertions)或删除(Deletions) [Indel]

序列比对的数学模型大体可以分为两类： 一类从全长序列出发，考虑序列的整体相似性，即整体比对/全局比对； 第二类考虑序列部分区域的相似性，即局部比对。

局部相似性比对的生物学基础，是蛋白质功能位点 往往是由较短的序 列片段组成的，这些部位的序列具有相当大的保守性，尽管在序列的 其它部位可能有插入、删除或突变。此时，局部相似性比对往往比整 体比对具有更高的灵敏度，其结果更具 生物

学意义

8.6.序列比对时的打分模型：

突变数据矩阵 MD、模块替换矩阵 BLOSUM

8.6.1.MD:

建立在已知的同源蛋白质/蛋白质家庭 的多序列比对的基础之上的;

统计某位点出现各种氨基酸的比例(%) [Pj, j=1 To 20，对应 20 种 Aa，伪随机概率;

该位点一个氨基酸发生改变，改变成另一个 Aa 的概率 P1,2 = Pj1/Pj2;

可接受点突变(Point Accepted Mutation,PAM) 1 个 PAM 的进化距离表示 100 个残基中发生一个残基突变的概率;

PAM 是在蛋白质高度相似的基础上选择的。蛋白质对齐要求显示至少是 85%

8.6.2.BLOSUM:

以序列片段为基础;基于蛋白质模块数据库 BLOCKS;从蛋白质模块数据库 BLOCKS 中找出一组替换矩阵，用于解决序列的远距离相关。

8.6.3.PAM 和 BLOSUM 换算：

PAM 后的数字——越大 越适用于序列相似性低的序列之间的比对

PAM 后的数字——越小 越适用于序列相似性高的序列之间的比对

Blosum 正好相反

PAM	BLOSUM
PAM100	BLOSUM90
PAM120	BLOSUM80
PAM160	BLOSUM60
PAM200	BLOSUM52
PAM250	BLOSUM45

PAM	BLOSUM
To compare closely related sequences, PAM matrices with lower numbers are created.	To compare closely related sequences, BLOSUM matrices with higher numbers are created.
To compare distantly related proteins, PAM matrices with high numbers are created.	To compare distantly related proteins, BLOSUM matrices with low numbers are created.

PAM	BLOSUM
Based on global alignments of closely related proteins.	Based on local alignments.
PAM1 is the matrix calculated from comparisons of sequences with no more than 15% divergence but corresponds to 99% sequence identity.	BLOSUM 62 is a matrix calculated from comparisons of sequences with a pairwise identity of no more than 62%.
Other PAM matrices are extrapolated from PAM1.	Based on observed alignments; they are not extrapolated from comparisons of closely related proteins.
Higher numbers in matrices naming scheme denote larger evolutionary distance.	Larger numbers in matrices naming scheme denote higher sequence similarity and therefore smaller evolutionary distance.

空位开放罚分

空位延伸罚分

v(g) = - d - (g-1)e

【g 代表连续空位数 d 代表空位开放罚分 e 代表空位延伸罚分】

8.7.序列比对原则：动态规划算法(最长共同子序列)、启发式搜索算法（最优方案、完整性、准确度精密度、执行时间）

启发式算法应用：神经网络、自学习。

启发式算法应用在序列比对 ：等长的高分片段对-》通过延伸或连接优化结果-》运用动态规划算法引入空位

8.8.回溯：从最高的 分数开始 ，进行 traceback 直到碰到 0 为止。

8.9.序列比对软件：EBI ClustalW2【动态规划算法 >> 全局联配工具】

8.10.BLAST 和指定的物种进行序列比对:【启发式搜索算法（局部打分策略）:】

8.11. EBI >> FASTA【启发式搜索算法（局部打分策略）:】

https://www.ebi.ac.uk/Tools/sss/fasta/

8.12.二代测序 NGS 比对（4）

BWA、Bowtie、Bowtie2(=》illumina,solid)、samtools、