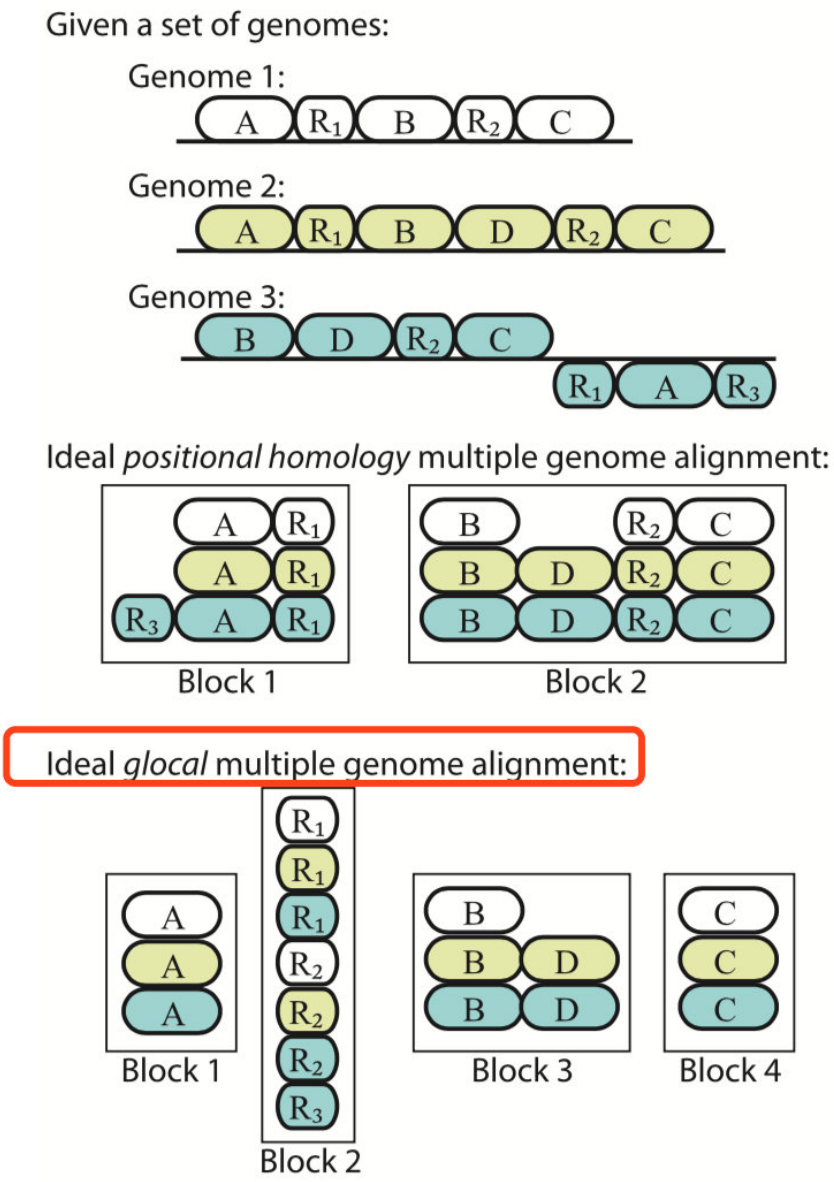
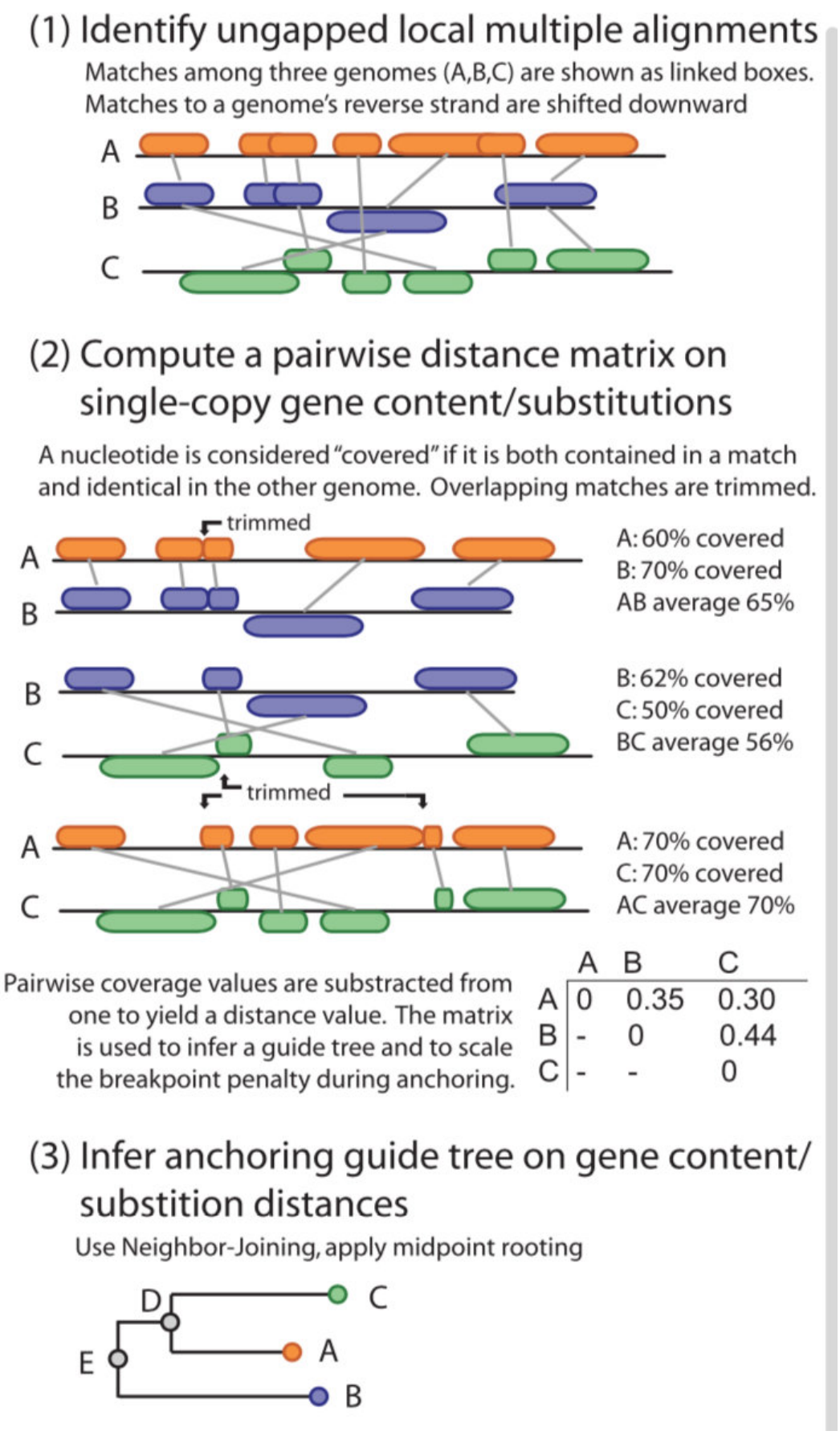
[Instruction][http://darlinglab.org/mauve/user-guide/introduction.html]

Mauve用于多重基因组比对，查看进化过程带来的重排和插入。

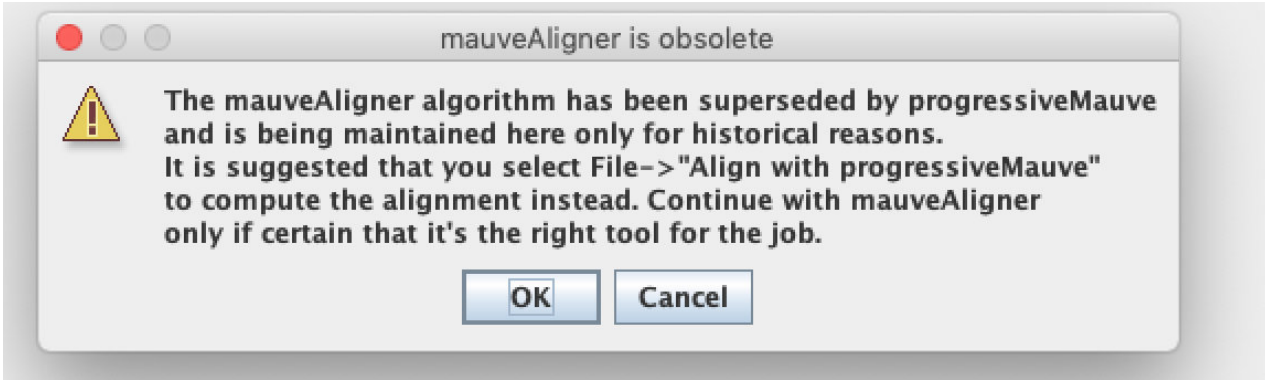


构建全基因组范围内的Neighbor-Joining phylogenetic tree

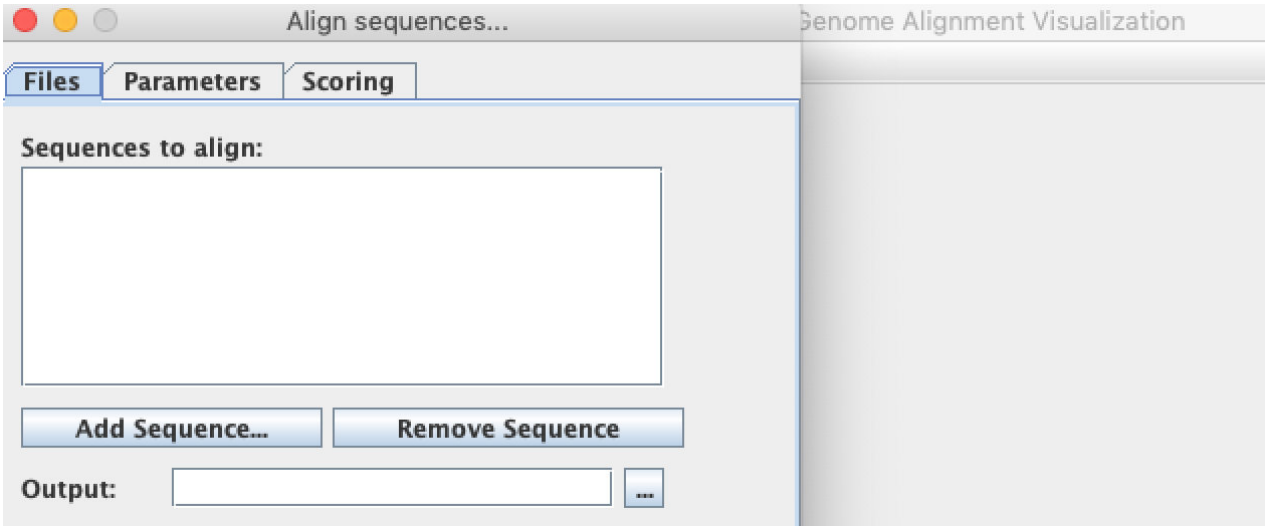


Usage

推荐使用"Align with progressiveMauve"



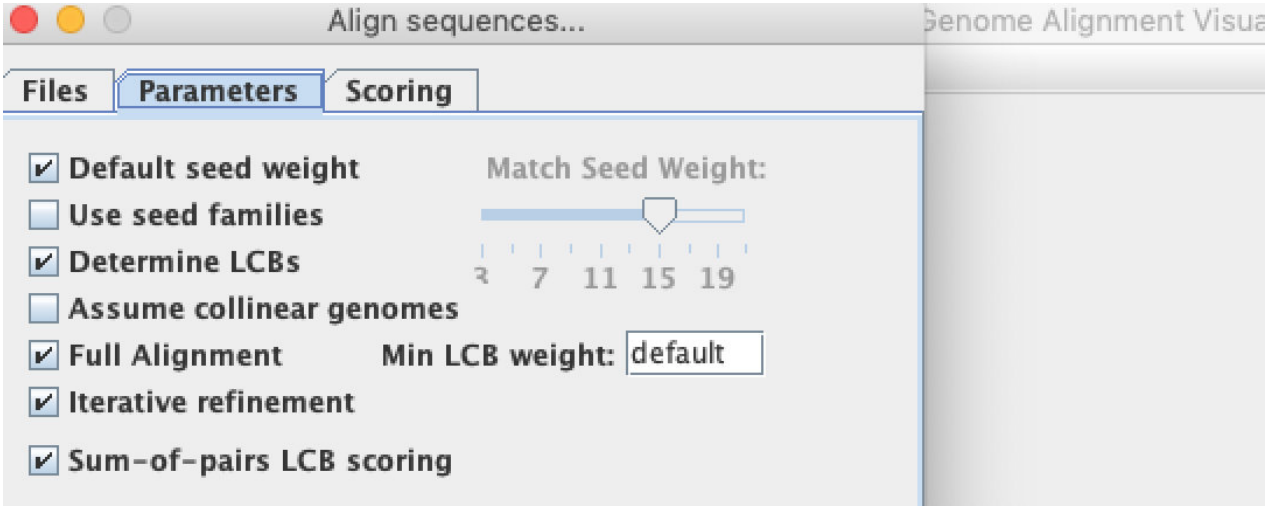
- Align with progressiveMauve



输入基因组文件可以是任意的FastA，Multi-FastA，GenBank 文件，软件基于文件名称后缀判断输入文件个数，未能判断文件默认为(Multi-)FastA格式。Mauve假设同一个输入的Multi-FastA为一个基因组。

当单个文件(individual file)包含多个序列输入(sequence entries)时，这些序列将会被合并到一起，然后使用整个连接到一起的序列比对其他序列或输入文件。类似地，包含多个contigs的不完整基因组也能被比对，但是不正确的contigs排序将会导致基因组的重排。

- Progressive Mauve alignment parameters



1. Match seed weight

表示用于生成局部多重比对(local multiple alignments, matches)的最小weight，在比对差异较大或者多个基因组时，该值越小，敏感度越高；同时要保证seed特异性，太小将降低敏感度

2. Default seed weight

默认条件下， 1MB基因组为11，5MB为15，一方面针对多个基因组比对时，默认值可能太大了，另一方面，较大的值将减少比对噪音，带来更好的比对结果

3. Use seed families

progressiveMauve会采用3个空间的seed模式而不是1个，这样将提供敏感度。

4. Determine LCBs (Locally Collinear Blocks)

不选择该选项时，Mauve将简单地在基因组间识别匹配(local multiple alignments)

5. Assume colinear genomes

假设比对的基因组之间不存在重排时，可选择该选项。

6. Full alignment and Iterative Refinement

选择"Full alignment"将使得progressive Mauve执行recursive anchor search，同时使用MUSCLE进行基因组序列full gapped 比对。Progressive Mauve将会识别alignment anchors，然后聚集为LCBs完成比对。

- The viewer

1. The display layout

每个颜色块代表比对到其他基因组的一个区域，并假设它们同源；当颜色块位于中线以上时表示相对和第一个基因组序列为正向比对，位于中线以下时表示为反向互补比对；颜色块以外的序列由于缺乏同源性而无法检测；颜色块相似性的高度对应比对序列之间的保守性；**完全为白色的表示未比对上，同时可能是该基因组特有的序列区域。**

2. The Backbone color scheme

使用Progressive Mauve比对时，同时输出Backbone文件，表示所有基因组共同存在的保守区域。

3. Zooming in on Annotated Features

针对GenBank输入文件，可对应放大查看注释信息，但是只有小于1Mb长度序列的注释信息才会显示。等查看小于1Mb的序列注释信息时，注释的CDS为白色盒，tRNAs为绿色盒，rRNAs为红色盒，misc_RNA为蓝色盒。

4. Mouse control

默认Mauve选择第一个基因组为参考基因组，同时其他基因组根据参考基因组调整方向。当在Mauve比对上移动鼠标时，每个基因组中的比对后的同源位置将会以黑色垂直框显示。

5. 略

- File formats

1. The .alignment文件

```
>seq_num:start1-end1 ± comments (sequence name, etc.)
AC-TG-NAC--TG
AC-TG-NACTGTG
...

> seq_num:startN-endN ± comments (sequence name, etc.)
AC-TG-NAC--TG
AC-TG-NACTGTG
...
= comments, and optional field-value pairs, i.e. score=12345
```

比对文件输出(.alignment)包含了Mauve生成的完整基因组比对，包含了相对于原始输入序列的方向和位置信息。XMFA(eXtended multi-FastA)文件格式。XMFA文件支持存储多个collinear sub-alignments文件，每个文件之间使用等号(=)区分，共同组成单个基因组比对。**同时对应的序列为经过方向调整后的序列，如：减号，意味着后续序列经过了反向互补后输出，同时序列输出包含了比对情况，因此含有的-表示gaps。**

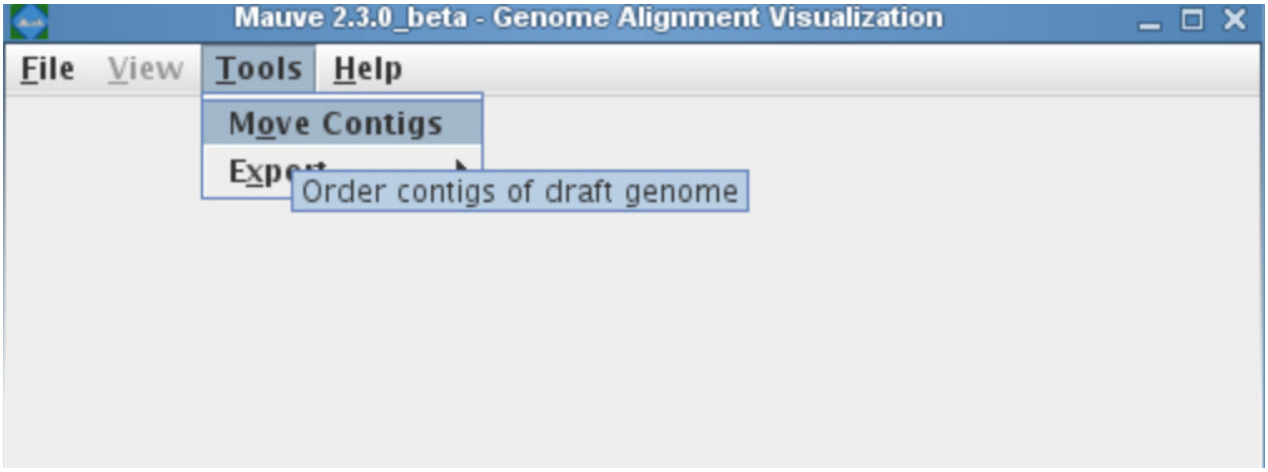
2. The Progressive Mauve backbone file

seq_0_leftend	seq_0_rightend	seq_1_leftend	seq_1_rightend	seq_2_leftend	seq_2_rightend
1	15378	1	15377	1	15377
16728	19795	15378	18446	15378	18445
0	0	18447	18668	18446	18667
19796	20566	18669	19439	18668	19438

第一行对应每列所包含的信息，分别对应每个基因组保守区域的位置值；同时第一个基因组的15379-16727位置为一个island(subset backbone)，同样第3行中，第一个基因组缺少可检测的同源片段(18447-18667)。

- Reordering contigs

Mauve Contig Mover(MCM)能用于根据相关参考基因组对draft基因组进行排序。



选定输出目录，MCM将会创建一系列比对输出，对应不同的子目录中，因此建议新建孔目录输出。**在reordering contigs时，只能输入两个序列，第一个序列必须为reference(仅含一个contig)，第二为draft genome；reference可以为任何可行的格式，draft必须为fasta或genbank文件。**

MCM将会输出迭代过程中的排序，每次迭代输出都包含标准的Mauve alignment files，输出文件name_of_genome_contigs.tab包含contigs的顺序和方向改变。

```
Ordered Contigs
type    label   contig  strand  left_end        right_end
contig  NODE_22_length_99526_cov_52.730128      chromosome      forward 1       99526
contig  NODE_23_length_86890_cov_52.109955      chromosome      forward 99527   186416
contig  NODE_57_length_4702_cov_48.184262       chromosome      complement      186417  191118
contig  NODE_35_length_34992_cov_51.576825      chromosome      complement      191119  226110
contig  NODE_15_length_129509_cov_54.602804     chromosome      forward 226111  355619
contig  NODE_12_length_192269_cov_55.791071     chromosome      forward 355620  547888
```

每次迭代输出的文件夹命名为alignment1-alignmentX；每个文件夹内包含以draft基因组名字命名的tab文件(name_of_genome_contigs.tab)，该文件包含排序和定向后的结果。该文件包含3个部分，每个部分包含一系列contigs。该数据包含了每个contig的名称(label)，位于参考基因组位置(position)，以及序列方向(orient, 同输入序列或反向互补)：

- Contigs to reverse： 该部分的contigs的排序与至之前迭代相反。同时这些contigs可能和原始输入文件的排序相同。
- **Ordered Contigs：draft文件内排序和定向后的contigs。由于该部分包含了原始draft文件中所有的contigs，那些不存在排序的(no aligned region)信息将会聚集在最后。不存在LCBs (Locally Collinear Blocks)的contigs将会出现在draft基因组最后。**
- Contigs with Conflicting Order information：包含一系列包含LCBs同时比对到多重位置的contigs。这些可用于核实contigs在参考基因组上的定位，或者查看潜在的错误组装或者重排。

假如输入文件为genebank格式，那么每个输出比对文件夹内会包含name_of_genome_fetures.tab文件，该文件包含么当前方向和位置的注释信息等。

因此，最大编号等文件夹内的fasta和1个或2个描述性文件即为最终的排序后draft genome(对于tab文件中的ccomplement，已经转换为反向互补序列)。

使用命令行对contigs重排序

```
java -jar -cp /Applications/Mauve.app/Contents/Java/Mauve.jar org.gel.mauve.contigs.ContigOrderer -output results_dir -ref reference.gbk -draft draft.fastas
```