基因组拼接与注释
实验操作

贾磊

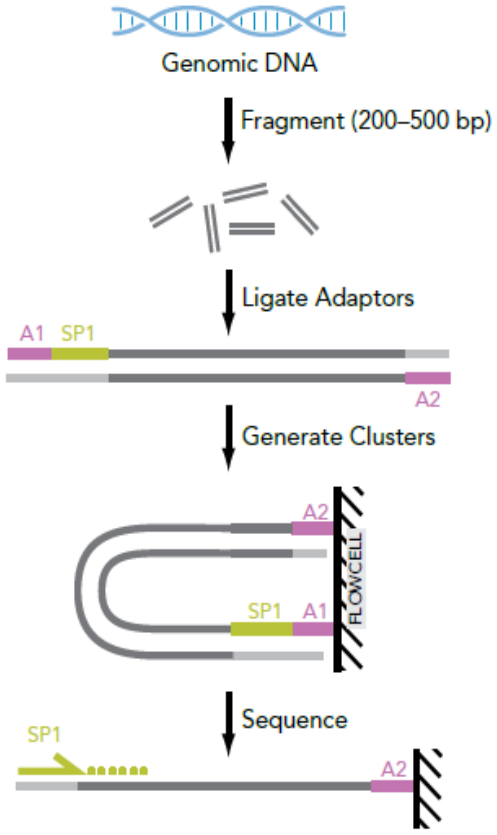
2020.9.24

浙江大学
ZHEJIANG UNIVERSITY

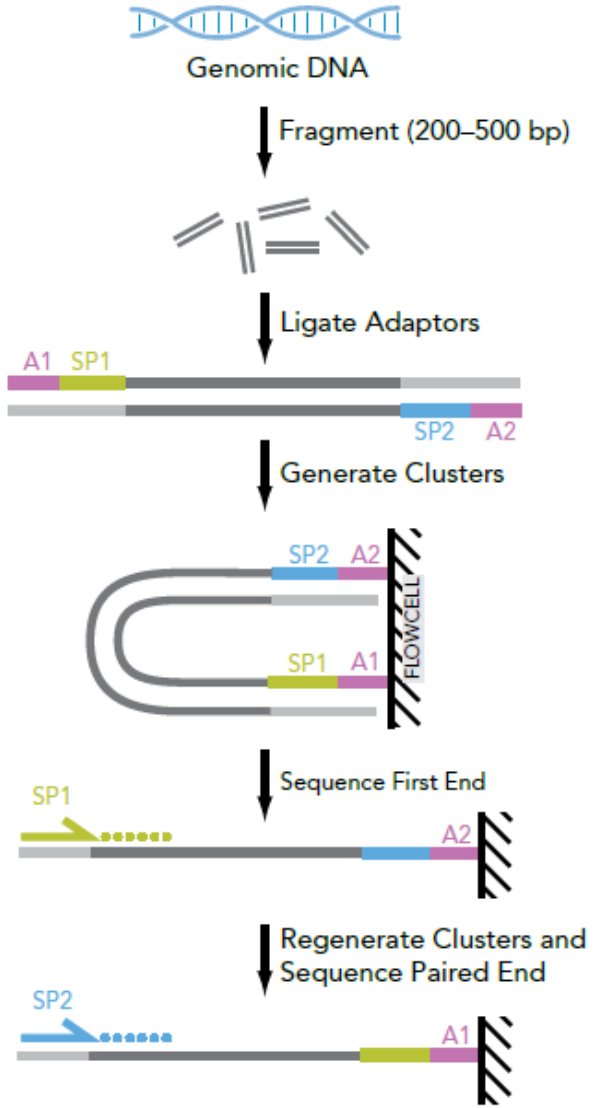
目录

基因组拼接

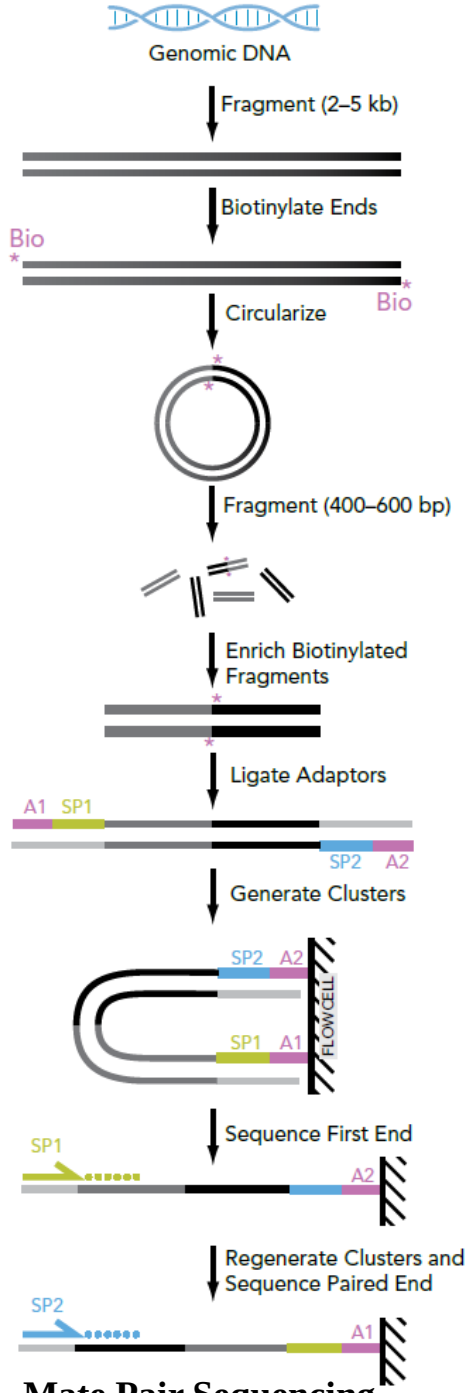
背景介绍 Single-Read vs Paired-End vs Mate Pair



Single-Read Sequencing



Paired-End Sequencing



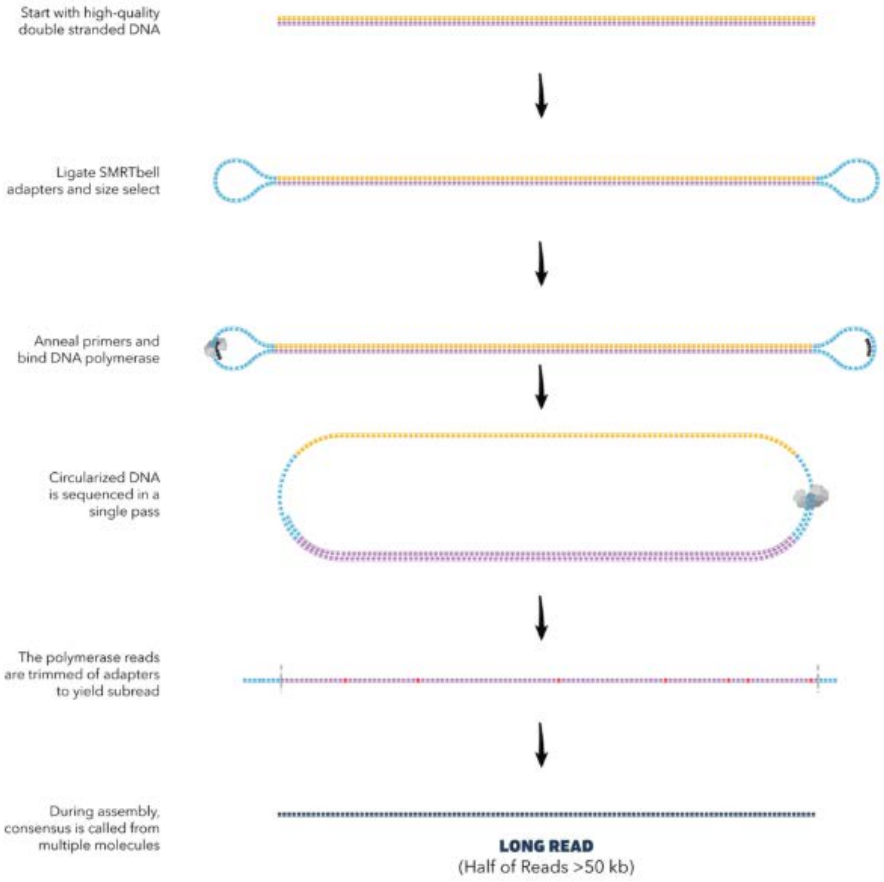
Mate Pair Sequencing

浙江大学 ZHEJIANG UNIVERSITY

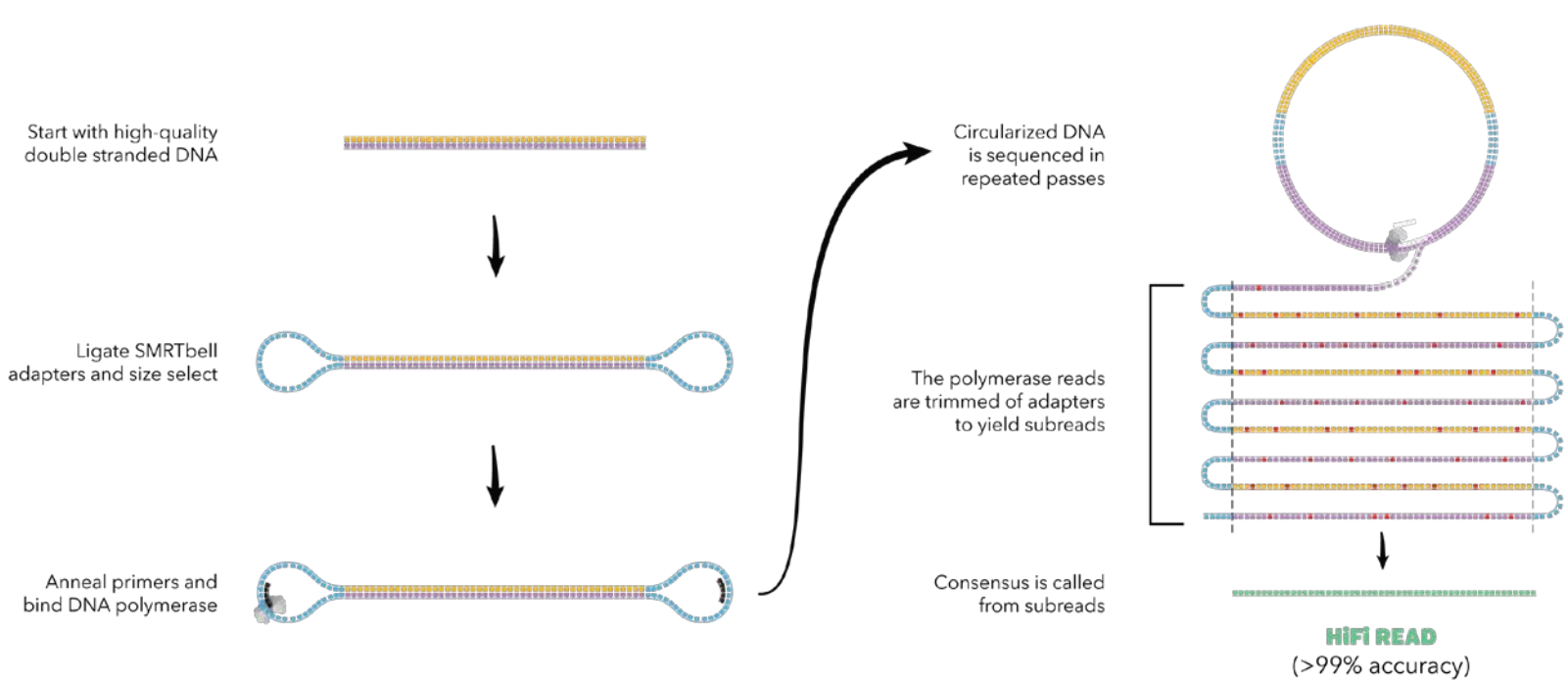
illumina

基因组拼接

背景介绍 PacBio SMRT sequencing technology



PacBio CLR (Continuous Long Read) 测序技术

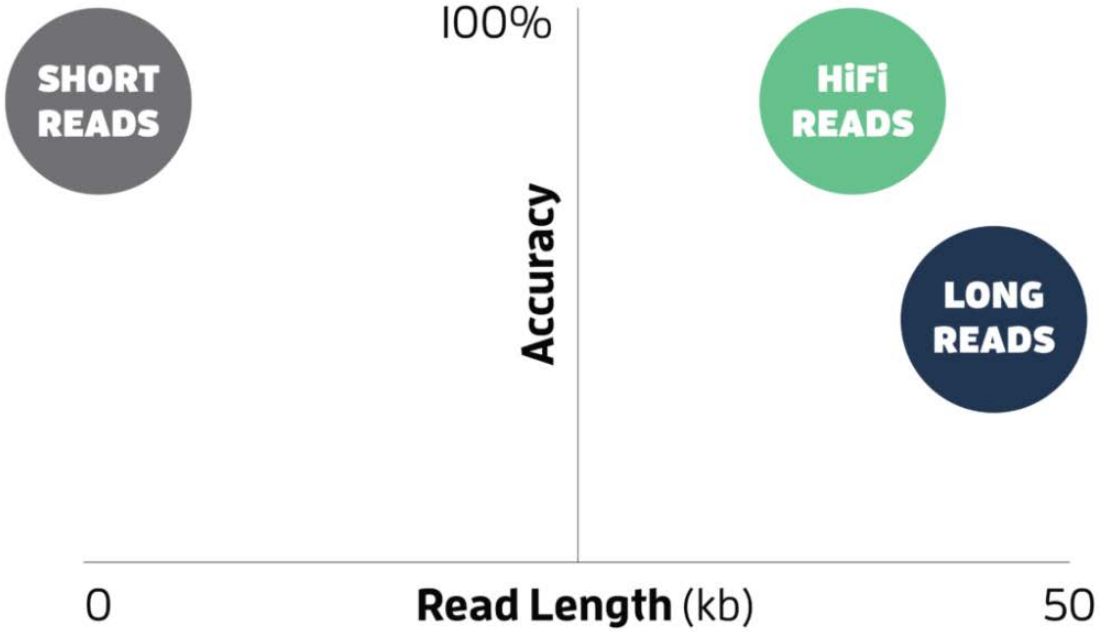


PacBio CCS (Circular Consensus Sequencing) 测序技术

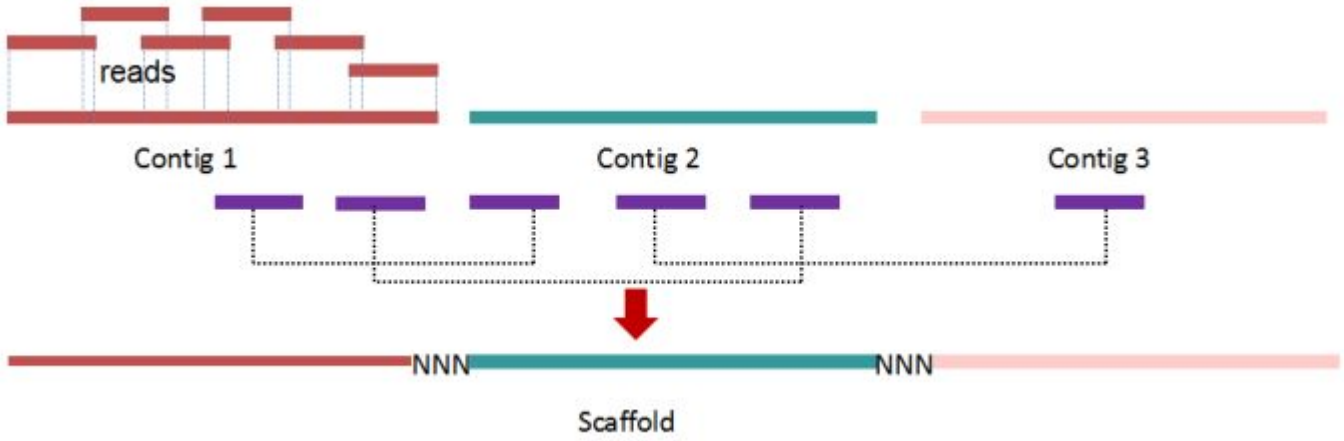
浙江大学 ZHEJIANG UNIVERSITY

PACBIO®

背景介绍 NGS vs TGS



背景介绍 Scaffold vs Contig



背景介绍 Linux基础

操作类型	操作命令	命令释义
目录操作	cd	切换目录（change directory）
	ls	列出目录中所有文件（list directory）
文件操作	less/more/cat/head/tail	查看文件
	mv	移动文件/重命名文件（move）
	rm	删除文件（remove）
其他操作	echo	显示字符串、变量内容
	clear	清屏
	date	显示系统时间

环境变量PATH 存放所有可执行程序（命令）的路径。

```
[root@localhost ~]# echo $PATH
/home/software/Reapr_1.0.18:/home/software/R-3.6.1/bin:/home/software/proftpd-1.3.6:/home
/software/nim-0.18.0/bin:/home/software/ncbi-blast-2.5.0+/bin:/home/software/mummer-4.0.0
beta/bin:/home/software/hmmer-3.1b2-linux-intel-x86_64/bin:/home/software/hisat2-2.0.5:/h
ome/software/DAGCHAINER:/home/software/clustalw-2.1/bin:/home/software/busco-master-v3/sc
```

命令补齐 Tab键，实现对文件名、命令名等的自动补齐。

背景介绍

确定进行一个基因组拼接项目前的工作

·基因组调查

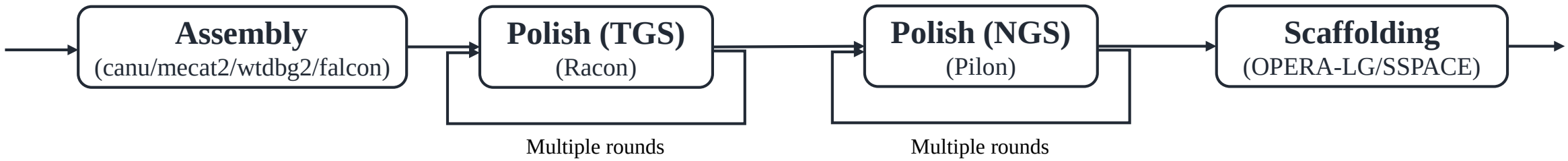
* 基因组大小（测序量，NGS>60×）
* 重复序列比例（测序手段）
* 杂合程度（高杂合度拼接难度大）
* 倍性（测序量、测序手段）
* GC含量（Illumina测序技术的GC偏好性）

·测序策略选择

基于二代测序数据的拼接

基于二代测序数据的拼接流程

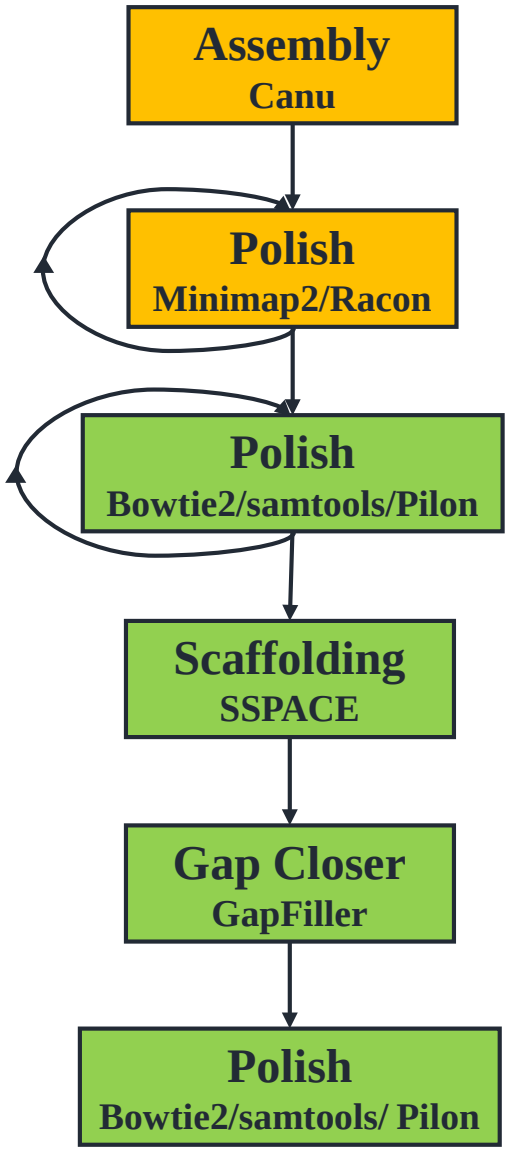
基于三代测序数据的拼接



基于三代测序数据的拼接流程

基因组拼接

基于三代测序数据的拼接



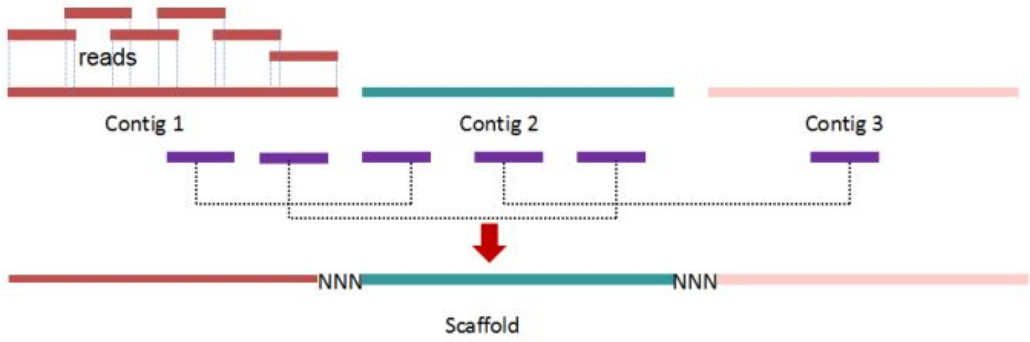
Canu用于PacBio/Nanopore等长序列组装。
Canu组装策略：1）检测测序序列的重叠；2）获取校正后一致性序列；3）对校正后序列进行修剪；4）组装；

Minimap2将三代数据比对到基因组上，Racon基于比对结果获取consensus序列，从而实现对拼接的校正。

Pilon使用二代数据进行矫正。

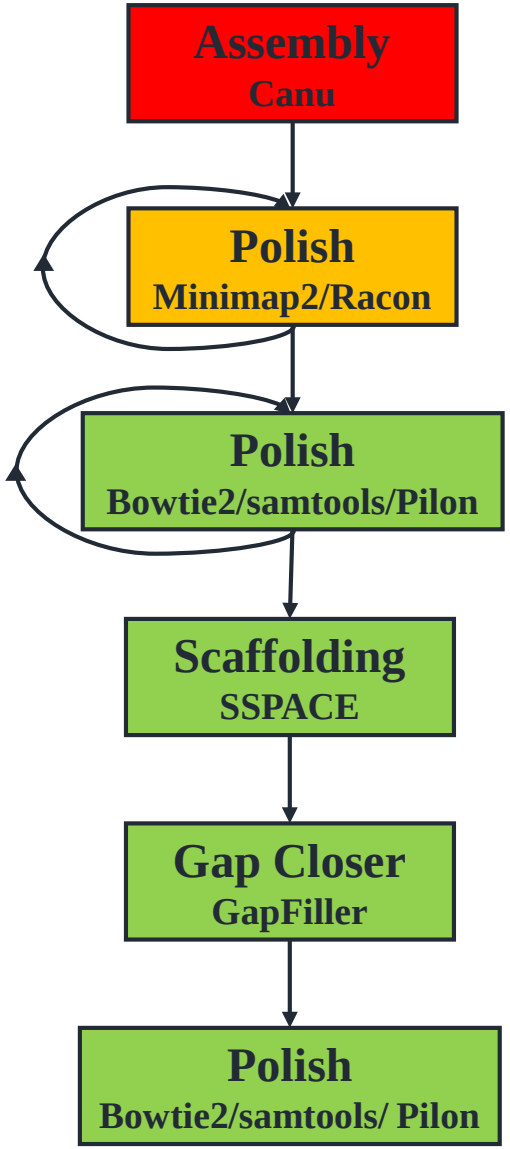
SSPACE使用二代长库进行scaffolding。

GapFiller对拼接进行补洞。



基因组拼接

基于三代测序数据的拼接



usage: canu [-correct | -trim | -assemble | -trim-assemble] \
-p <assembly-prefix> #输出前缀
-d <assembly-directory> #输出目录
genomeSize=<number>[g|m|k] #预估基因组大小
[-pacbio-raw |
-pacbio-corrected |
-nanopore-raw |
-nanopore-corrected] #序列文件类型
file1 file2 ... #序列文件

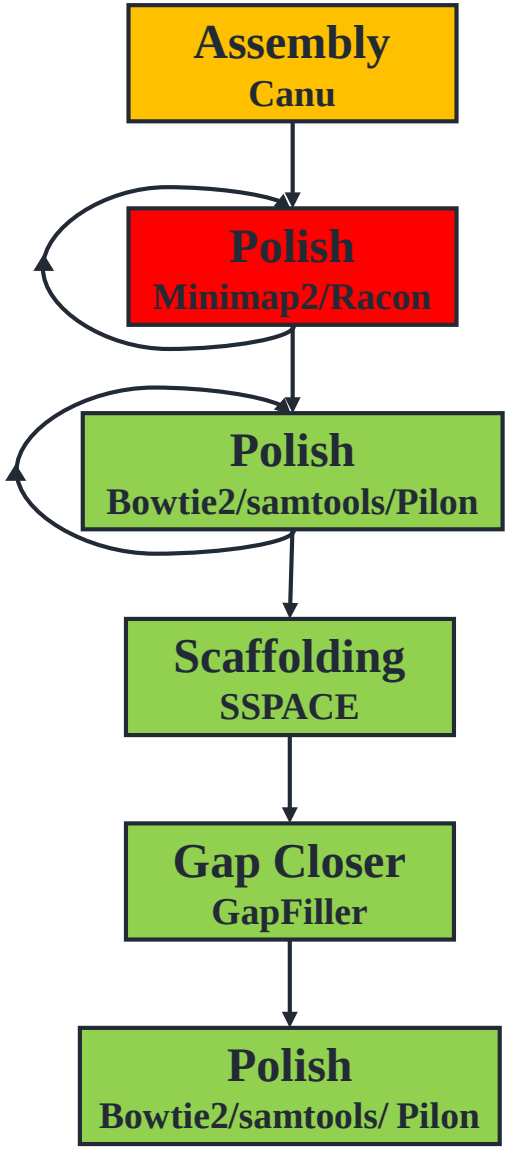
example: canu -d run1 -p godzilla genomeSize=1g -nanopore-raw reads/*.fasta.gz

canu -p ecoli -d ecoli-pacbio genomeSize=4.8m -pacbio-raw pacbio.fastq

Canu manual:
https://canu.readthedocs.io/en/latest/quick-start.html

基因组拼接

基于三代测序数据的拼接



Usage: minimap2 [options] <target.fa>|<target.idx> [query.fa] [...]
Options:
-a output in the SAM format (PAF by default) #输出格式
-o FILE output alignments to FILE [stdout] #指定输出文件
-t INT number of threads [3] #线程数
-x STR - map-pb/map-ont: PacBio/Nanopore vs reference mapping #比对方式
- ava-pb/ava-ont: PacBio/Nanopore read overlap
- asm5/asm10/asm20: asm-to-ref mapping, for ~0.1/1/5% sequence divergence
- splice: long-read spliced alignment
- sr: genomic short-read mapping

minimap2 -a -x map-pb -t 80 ref_1.fa test_pacbio.fastq -o test_pacbio.sam

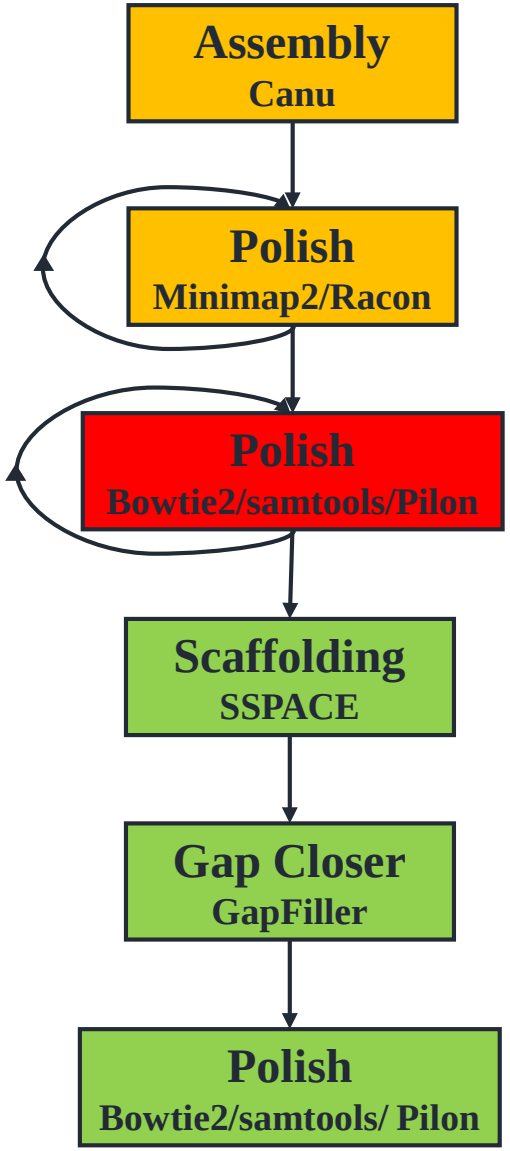
usage: racon [options ...] <sequences> <overlaps> <target sequences>
<sequences>
input file in FASTA/FASTQ format (can be compressed with gzip), containing sequences used for correction
<overlaps>
input file in MHAP/PAF/SAM format (can be compressed with gzip), containing overlaps between sequences and target sequences
<target sequences>
input file in FASTA/FASTQ format (can be compressed with gzip), containing sequences which will be corrected

options:
-t, --threads <int> default: 1 number of threads

racon -t 90 test_pacbio.fastq test_pacbio.sam ref_1.fa >ref_2.fa

基因组拼接

基于三代测序数据的拼接



Usage: bowtie2-build [options]* <reference_in> <bt2_index_base>
reference_in comma-separated list of files with ref sequences
bt2_index_base write bt2 data to files with this dir/basename

bowtie2-build ref_2.fa ref_2.fa

Usage:
bowtie2 [options]* -x <bt2-idx> {-1 <m1> -2 <m2> | -U <r> | --interleaved <i> | -b <bam>} [-S <sam>]

<bt2-idx> Index filename prefix (minus trailing .X.bt2).
<m1> Files with #1 mates, paired with files in <m2>.
<m2> Files with #2 mates, paired with files in <m1>.
<sam> File for SAM output (default: stdout)
Options (defaults in parentheses):
-p/--threads <int> number of alignment threads to launch (1)

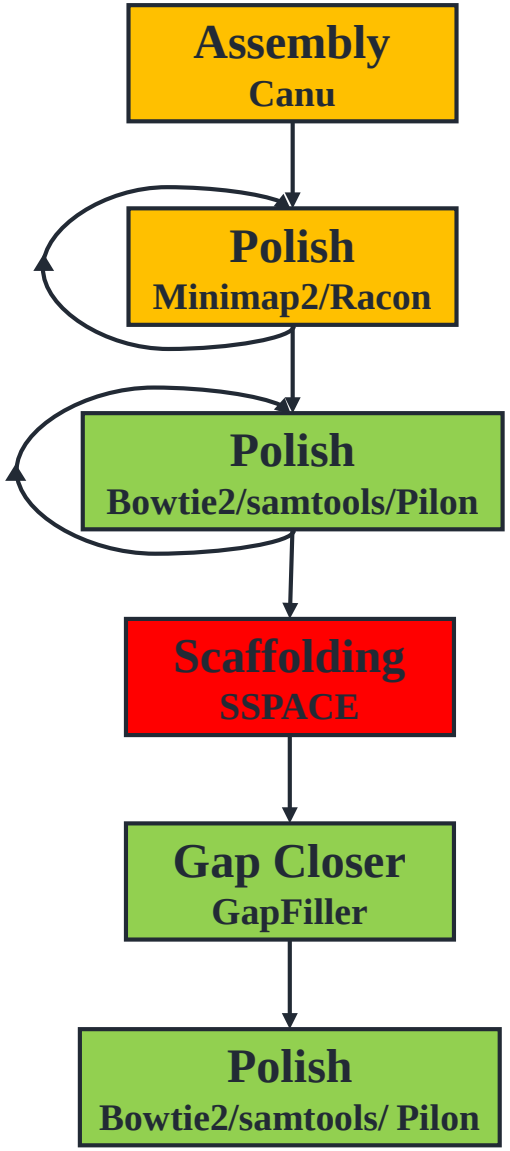
bowtie2 -p 20 -x ref_2.fa -1 test_NGS_1.fq test_NGS_2.fq -S test_NGS.sam

Usage: pilon --genome genome.fasta [--frags frags.bam] [--jumps jumps.bam] [--unpaired unpaired.bam] [other options]
--genome genome.fasta The input genome we are trying to improve.
--frags frags.bam A bam file consisting of fragment paired-end alignments.
--jumps jumps.bam A bam file consisting of jump (mate pair) paired-end alignments.
--output prefix Prefix for output files
--outdir directory Use this directory for all output files.
--changes If specified, a file listing changes in the <output>.fasta will be generated.
--threads Degree of parallelism to use for certain processing (default 1).

java -jar /public/software/apps/pilon-1.23/pilon-1.23.jar --genome ref_2.fa --frags test_NGS_1.bam --frags test_NGS_2.bam --output ref_3 --outdir ./ --changes --threads 60

基因组拼接

基于三代测序数据的拼接



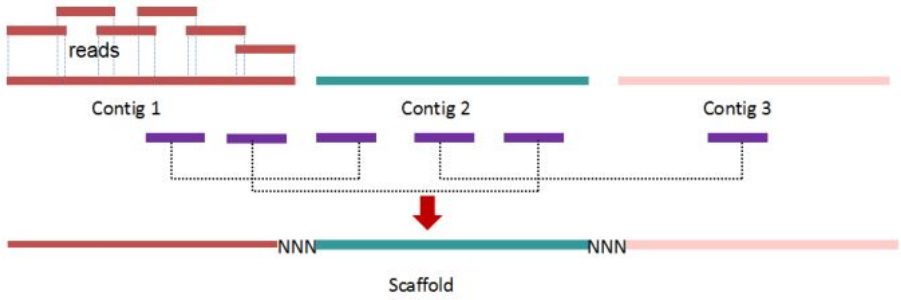
Usage: /public/software/apps/SSPACE-STANDARD-3.0_linux-x86_64/SSPACE_Standard_v3.0.pl [SSPACE_Standard_v3.0_linux]

========== General Parameters ==========
-l Library file containing two mate pate files with insert size, error and either mate pair or paired end indication.
-s Fasta file containing contig sequences used for extension. Inserted pairs are mapped to extended and non-extended contigs (REQUIRED)
========== Additional Parameters ==========
-T Specify the number of threads to run SSPACE, used both for reading the input readfiles and mapping the reads against the contigs.
-b Base name for your output files (optional)

perl /public/software/apps/SSPACE-STANDARD-3.0_linux-x86_64/SSPACE_Standard_v3.0.pl -l lib_list -s ref_3.fa -T 80 -b ref_4

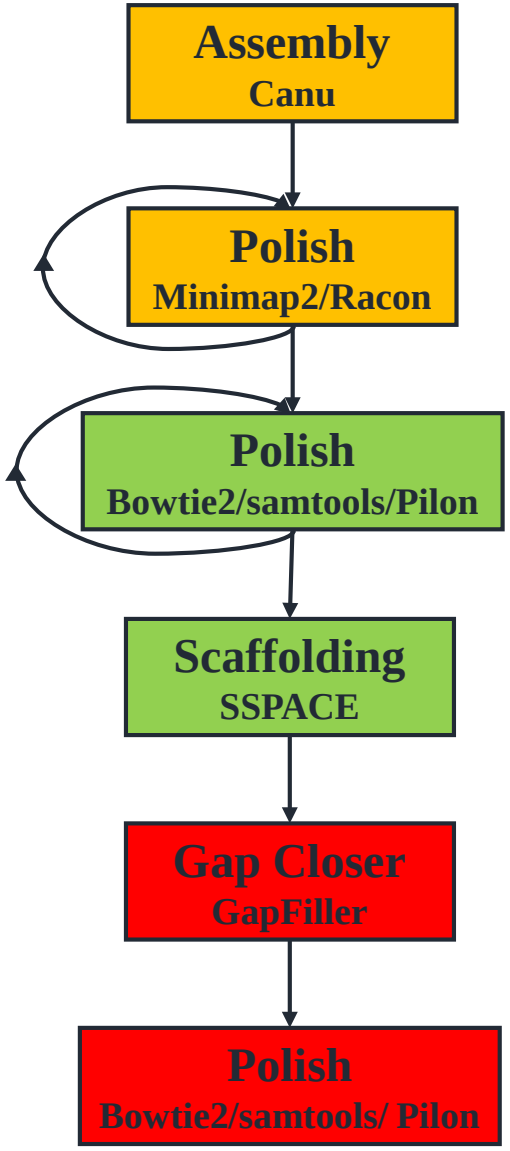
lib_list

Lib3	bowtie	500_1_left.fq	500_1_right.fq	500	0.25	FR
Lib4	bowtie	500_2_left.fq	500_2_right.fq	500	0.25	FR
Lib5	bowtie	2K_1_left.fq	2K_1_right.fq	2000	0.5	RF
Lib6	bowtie	2K_2_left.fq	2K_2_right.fq	2000	0.5	RF



基因组拼接

基于三代测序数据的拼接

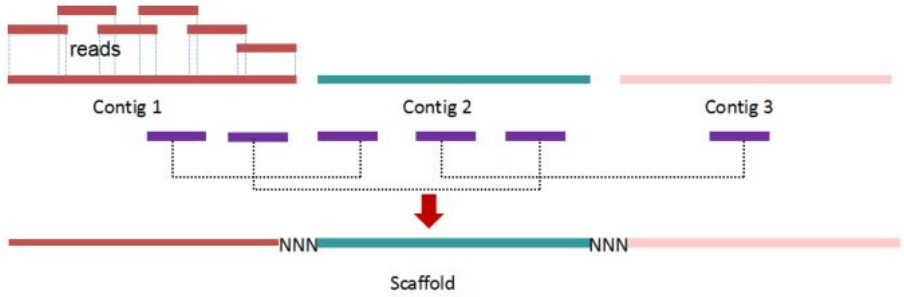


Usage: /public/software/apps/GapFiller_v1-10_linux-x86_64/GapFiller.pl [GapFiller_v1-10]

========== General Parameters ==========
-l Library file containing two paired-read files with insert size, error and orientation indication.
-s Fasta file containing scaffold sequences used for extension.
========== Additional Parameters ==========
-T Number of threads to run (default -T 1)
-b Base name for your output files (optional)

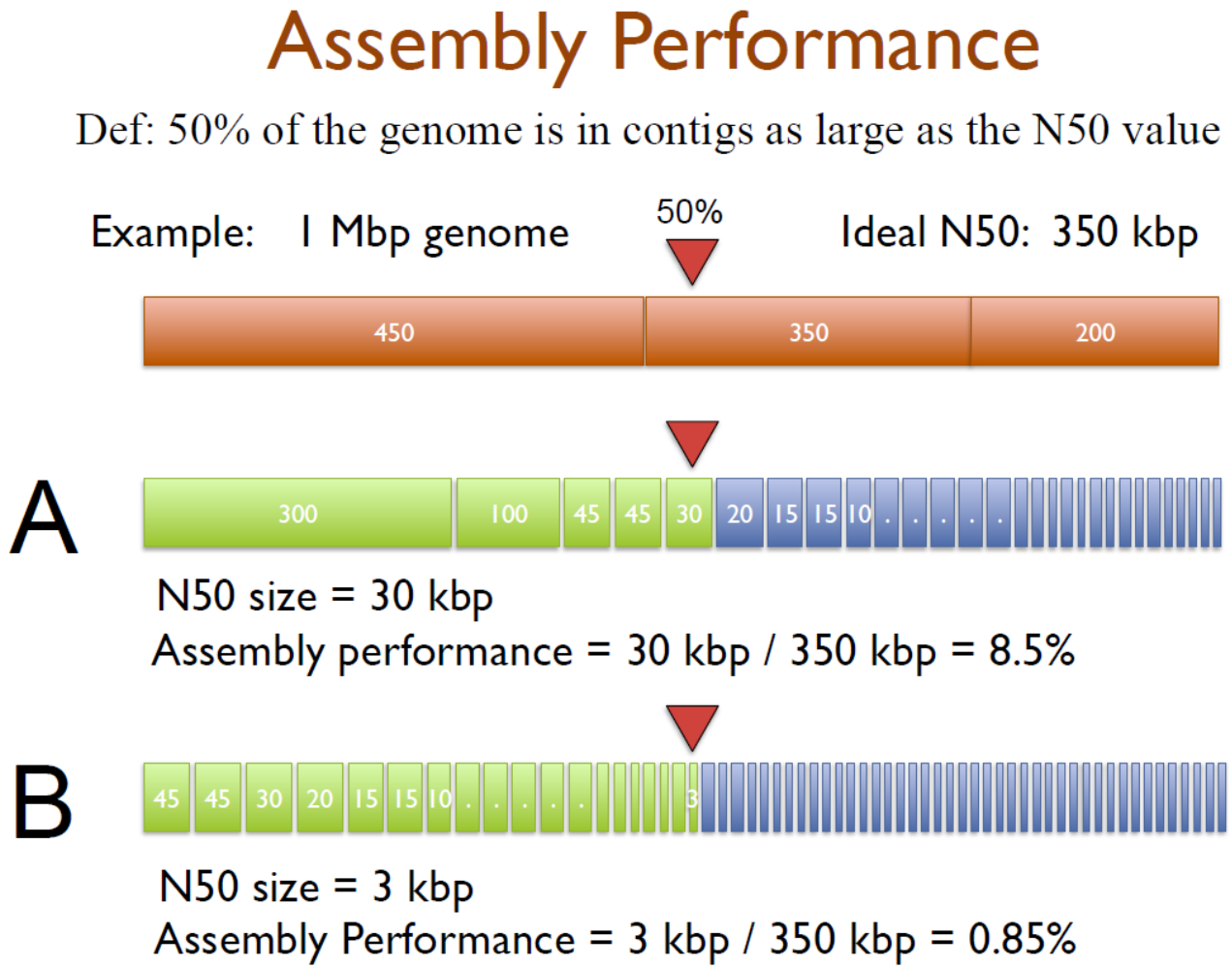
perl /public/software/apps/GapFiller_v1-10_linux-x86_64/GapFiller.pl
-l lib_list -s ref_4.fa -T 80 -b ref_5

Polish同Step3



拼接质量评估

N50


Assembly Performance
Def: 50% of the genome is in contigs as large as the N50 value
Example: 1 Mbp genome
50%
Ideal N50: 350 kbp
450
350
200
A
300
100
45
45
30
20
15
15
10
N50 size = 30 kbp
Assembly performance = 30 kbp / 350 kbp = 8.5%
B
45
45
30
20
15
15
10
3
N50 size = 3 kbp
Assembly Performance = 3 kbp / 350 kbp = 0.85%

基因组拼接

拼接质量评估

完整性

BUSCO/CEGMA进行基因组拼接完整性评估。

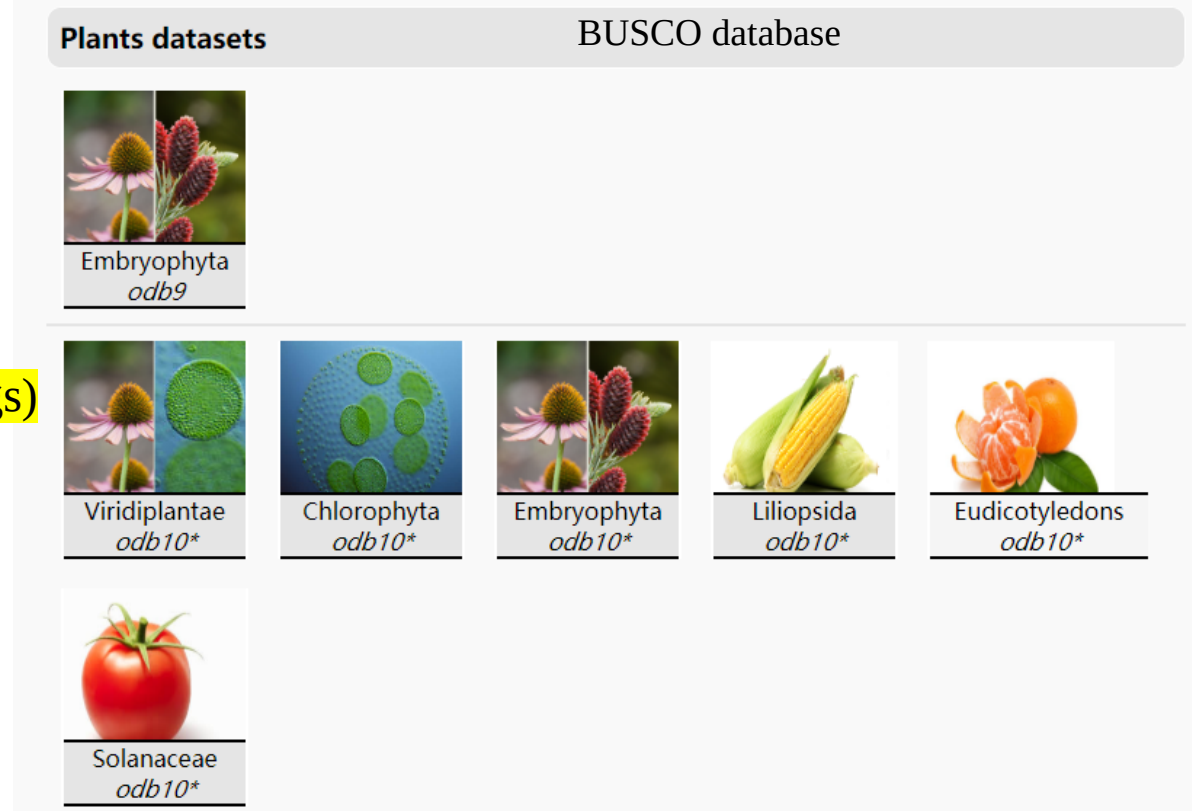
BUSCO(**B**enchmarking **U**niversal **S**ingle-**C**opy **O**rthologs)

CEGMA(**C**ore **E**ukaryotic **G**enes **M**apping **A**pproach)

准确性

·其它数据集比对验证

·近缘物种基因组比较



基因组注释

·结构注释

·功能注释

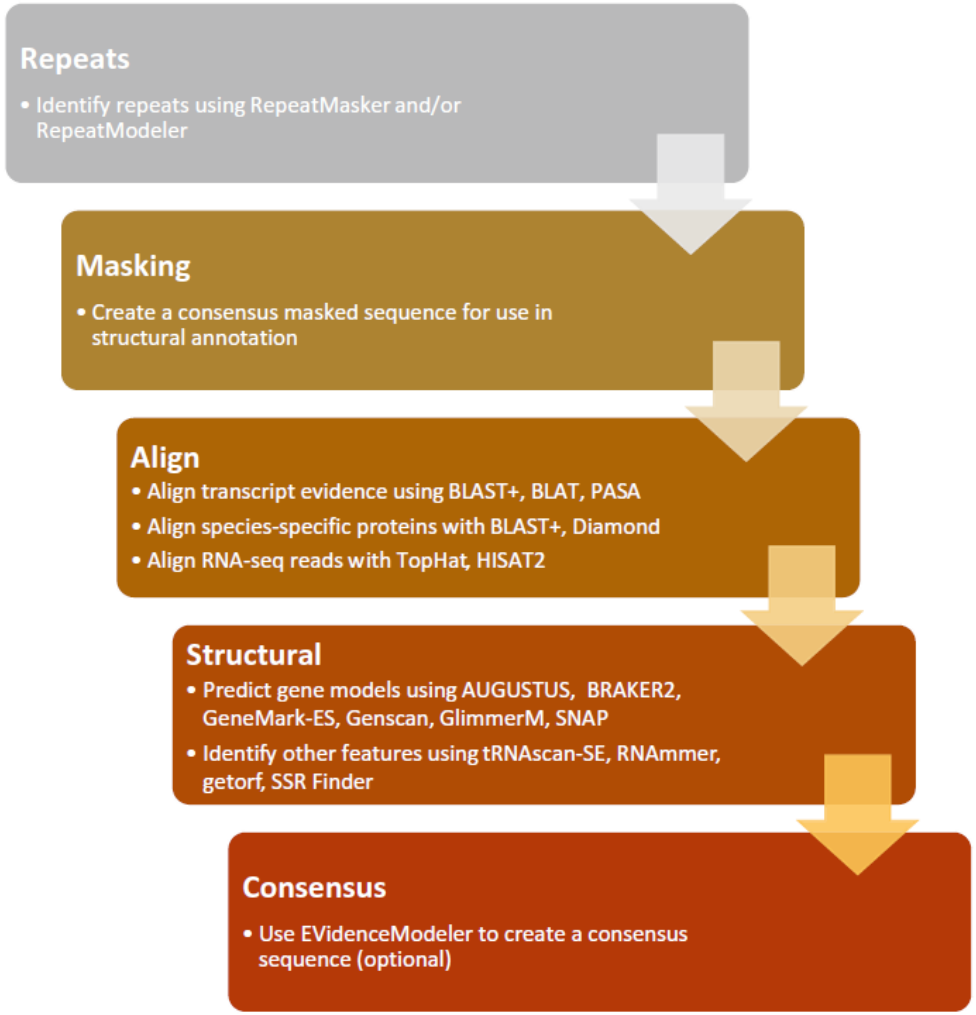
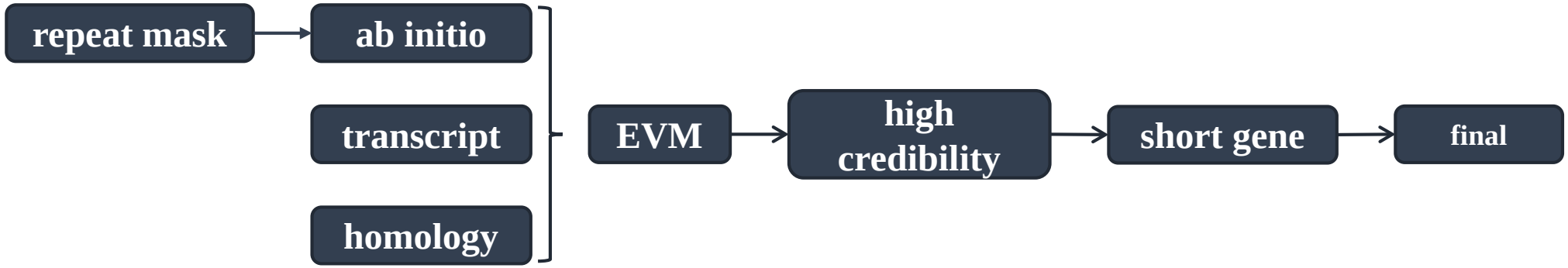


Fig. 7 An overview of the structural annotation steps of GenSAS

Jodi L. Humann *et al.*

基因组注释



基因组注释一般流程

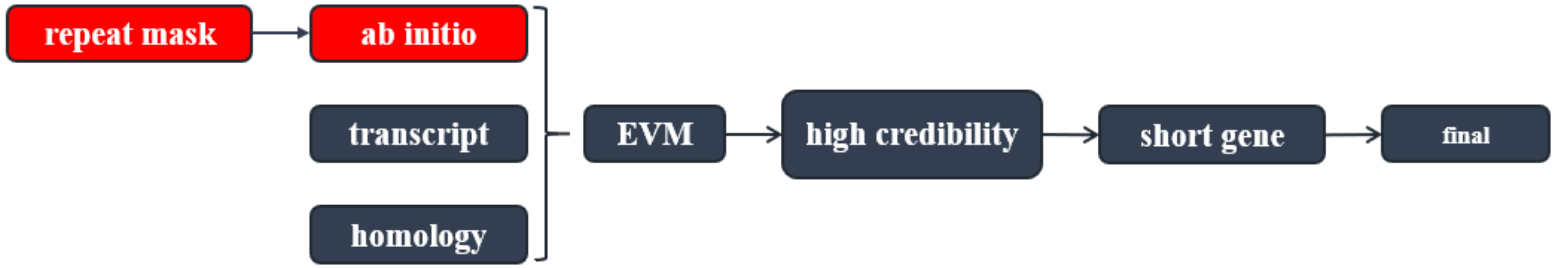
ab initio从头预测

·通过repeatmasker对基因组中重复序列进行masked

·使用Augustus/genemark/fgenesh对masked后的基因组进行基因从头预测

FGENESH:
http://www.softberry.com/berry.phtml?topic=fgenesh&group=programs&subgroup=gfind#
Augustus:
http://bioinf.uni-greifswald.de/augustus/submission.php

>tig00000001 LN:i:15464111 RC:i:16134 XC:f:0.999806
TAGAAACAGTTGGAAGAGTTGAAAGAAAAGGGGTATATACGTCCGAGTACTTCCCATGGGGTGCTCCTGTGATTTTTGTAGAGAAGAAAGACAAGACGAAGAGGATGTGCGTCGACTACAGAGCTCTCAATGAAGTCACCATCAAAAACAAGTACCCTCTTCCCCGGATCGATGATCTGTTCGATCAGCTTAAAGGTGCCACTGTATTCTCCAAGATTGATCTTCGTTCCGGATATCACCAATTACGTATTCGTGAAGAAGATATTCC
GAAGACCGCATTCACCACGCGATACGGGCTGTATGAATTCACGGTGATGTCGTTCGGTTTGACCAATGCACCTGCCTTCTTCATGAACTTGATGAACAAAGTGTTCATGGAATACTTGATAAGTTTGTTGTGGTTTTCATTGATGACATTCGATATACTCACAATCAGAGGAAGACCATCAGCACCATCTCCGTCTGGTGTTGGGAAAGTTGCGGGAACATCAATTGTATGCCAAGCTTAGCAAGTGTGAGTTCTGGTTGTCCGAAGT
CAAATTCTTAGGGCACGTGATATCTGCTAAAGGAGTTGCTGTGGATCCCGAAACAGTGACCGCCGTCACCGATTGGAAGCAACCCAAGACAGTCACTCAAGTTCGCAGTTTCTTAGGTTTGGCAGGATACTACCGGAGATTCATCGAGAACTTCTCCAGAATCGCTCGACCCATGACACAGTTGTTGAAGAAAGAAGAGAAATTTGTGTGGAGCCCGCAATGCGAGAAGGCGTTTCAAACTCTCAAGGAAAAGCTGGTTTCCTCGCC
AGTGTTAATCTTGCCGGATACTCGCAAGGACTTCATGGTGTATTGTGATGCTTCGCGCCAAGGATTGGGGTGCGTGCTCATGCAGGACGGTCATGTGGTAGCCTATGCCTCACGTCAGCTATGACCTCATGAAGTGTCGCGACCGGAAAAATTGCCTTTTCCCGAACGCTAGTGTATTAATCCCCGTCCCAGGAAAAGCCGGGGTACACCACACAAACATGATATAAAAGACACATCTTTATTATATCGATAATATTTACATTATTACA
AAGGCACTTAAGGCCTGACAATCAAAAATAAACGGCAGCGGACTAGCGGGCCTACGGCTCCATCTTCACAGGGCGCTCAACTGGGGTATAAGCCAGGACTCCACCTAAGACATCTTCTCCAAAGCTTCTCTTACTGAAGGGGGGAAAGATTGAGCAAGAATGAGTACAACCACCGTACTCAGCAAGCCACACCGGCGGATGCATGATTAATGCAAGGGGTGTACAAGGGGTAATAATATAGGGGTTAAGTTTTGCAGTAAACAGCAT
TTAAAAGTCACTTAGTTGCCCAAAGCCATTTTGTAAAGACGATCCTAGAGCTACACAGTATTATTAATCGAGGCCGTGAACCTTCACGAACCTGCCTTAACCCAAGGCCTACGATGATTTGGACCGAACTGGCAACCCGACCTGGGTCCCAGCTCGTCCCAAGCCAACCCAGGCCAACCATTCCAAATTTTAGTTGTTGAGCAAATTTTAAGAATTAAAACACTGACTTGGGTACATTGCTAGGCTTGCCCCATATCCGAGGGCTCGGC
TATTCGAATAGATTTACTCTGATCAGAGGTCTACATCTTTACCCACAAGACACATCTTCCTCACGTGTTGCCACGTGCCATACCACCACGGTATACGGACAGAAGCGTGACATAGTTTCCAACCCATCCTAGCCTTACGACAAGAGTACCGACCCAATCCTACCTACGGCCGGGATACCGGCACAGGCAGGCAGAACTCAGCCCCTCAGCATTGGGCACCAACCCTCCGCCGTATGACATCTCGACTACGCGGGCCGCAGCTCGTGTA
GCCTTCATTTGCCCTGGAGAATGTCCATCGACGCCCGACTTCATCCATCTCCAATCCGTGTACTTTTGTTTATGACTAGGCTGAGCCACAAACTAAGCCTTACCCATTAGACATGTGGAAGTACGGTAGTGCTTTGCAACAGAGGCCCAAGCTTAATCCTTATAGTACCCGAGGTACGACTAACAAAACCCAACCACGCACCTCGAGCCCAGCCTAAACCATTTTGGGGGTTTTGAAAAGAGGGGGAGGTCTGTGTGTCCAATTCCAACAT
AAGCCAACCATTCCATAGGGTCCAAATGATATGAGAATTACCAAAGTCTAAAGTTATAAAACCACCTAATGTTGCCTAATTAATCAGCGAAGCATCTACCTAAATTTATACTAGTGGAACCATGAATATAATGTCCACTAGTTAGGGTTTTGTTTATTCTAGGGTGGACAAGGTAATAATAACAATAGCAATAATAATAATAAGGTCATAACAAAGGTAAATAGGTATGGCTAGATAAAACAGTGATAACGCGGGAATTTAAATAAAGCG
ATAATGCATTAATTTAAATCAAAATAATTTTATAAACTGGGATTCAATATGCTCAAGGATGATGTGACTTGCCTTGCTCGCTTTCCCAAACGCCGGCTTCGACTTCCACGAAGAACGGATCTTCCGAAGCTGCAGCGTCTACACGACCAACGGAAAAGAAAAGGCTTTTACACTAATAAACTCCATACAACAACAGAAACAAAGCACAAAAATGGGTTCTTTACTTCTTAAGGAAAAATTAGAGACTTGAACGGTCCAATTCCGAGTTT
CAAATGGCCAAGATATGGCCATTTGAAGTTTATATGCTCTTTAAATGGATATTTGCGAATTTCTTCATTTAATTTTCAATTAAAAATCATGTTTATTGCGTCAGCCGAGGGGCGGGGCGGCGCGACCGGGTCCACGGGAAGTGGGGCCCACGCGTCAGCCGCTCGGTCCACGGTGGACCGGGCGCGTCGGCGGCGGTCGGGCGCCGTGGGTCCCGCGCGTCAGCCGCACCCGCGGGCGCGGGCGGCTGACGGCCGGGCCCACTTG
GCAGCCGCGAGGGCGCGCCCGAGGGCGGCCTCGCCGGCACGGCCGACGGGAGCGGCGCGCCCGCGCCCCGATGGTCACCGCCGGCGACCACCGGCGGCGGCGGAGCGGCGCCCGAGAGAGAGGAGGGAAGGGGAAAGCGAACGGCGGTCTTCGGCTCACCCCGAGACAACGGCGACGACGAGAACGGCGGACGGAGAGGAAGAAGGCAGCGGCGCGGTTCGGGATGACGTTGGCGACGGCGCTCCGACGGTCGACGGGCGAAG
CGGAGCGGCGGACGAGGTCGACGGCAACACGGCGAAGCGGACGGAGGCGACGCCGAGTCGGGAGGAAGACGGGAGCGGCGGCGGCGGCTAACCGAAGCACGGCGGCGACGGCGGGGGAGAAGGTGCTACGGCGCTAGCACGATTCCGACGGCGAGAGCGAGCGGTGAGCGGCGGAAACGGAAGGGCGGCGAGCTCGGGCAACGTTTTTATAGCGCCGGGAGGGAGAGAGGGCGGCCGGGGAGAGAGGAAACCGCACGCGGAGAT
CTCGGCCGCCATTGATTGCGCCGGCGACGATTACGGGCTTGATTCGAACGAATCCGAGGGAGAGAGAGGGGAAAGAGGGGAAACGGGAGAGGGGATTGCGGGGATTAATTCCCCCCTCTTGATTGCACGCGGGGCGGACGGGAGCGGCGGGATTCGCGGCGGCGGCGGCGCTAGGGCACGGACGCGGCGGCGGCGGCGGGAGGAGGGAGATGACAGGTGGGACCCACCCGTCAGCGAGAGTGGCGGGCGGGCCCGCCTG
TCAGCGGCGCGCGCGCGGGGAGCCGATGGGCCGCGGGAGAGGAGAGAGAGGGAGTGGGCCGAGCCGGCCCAAGAGGGAAGGGGGGGGGAAAGAGGACTTTTTAGGATTTTTCTTTTTATAAACCTTTTTCCACTTTGTTTATTTCTTTTCAATTTATTATTTGTGCTCTGAAAATTCCACTAAAATTTGAGGGCTCCTTTTAGACCAAGGAGAATTTAACAAAAATTCTCCGGGCCACATATGAATTTTTCTTGTA
CGTATTTTAGGGTTTGCTAATTTCTTTTCGAATTTTAATTAATTCTATTATTCCTTTTAGAAGATGATTTTTATTTCGGGATGAATTTATCAGGACGTGACATGAAGGCAACTACCCTACACATGACTTGGAGTTAGCCGCAGTGGTTCATGCTCTAAAAATCTGGCGGCACTATCTCATTGGGAATCGCTGTGAAATATATACCGACCATAAGAGTTTGAAATACATCTTCACCCAGTCGGATCTGAATCTTCGGCAAAGGAGATGGTT
AGAACTCATCAAGGACTATGATGTGGGAATTCACTATCATCCTGGTAAGGCCAACGTGGTTGCCGACGCTCTAAGTCGAAAGAGCCATTGCAACACCCTCTACGTCCGTGGCATTCCACCCGAACTTAATCAACAGATGGAAGCCCTGAACCTAAGCATAGTTGGTCCGCGGATTCTTGGCTGCGTTGGAAGCCAAGCCTACCTTGCTCGATCAAATCCGCGAGGCTCAGAAGAATGATCCGGACATGCATGGACTCCTCAAGAATATG
AAACAGGGTAAAGCCGCTCGTTTCACAGAGGATGAACACGGAACCCTATGGAACGGAAAGCGGGTATGCCGTTCCAGATAACAGAGAACTTAAACAGCTGATACTTCAGGACGCTCACGAAAGTCCTTATTCCATTCACCCGGCCAGTACAAAGATGTACTTAGACTTGAAGGAGAAATACTGGTGGTTAGCATGAAGCGGAAATTGCAGAGTTTGTGCCCTATGTGATGTGTCCAGCGTGTCAAGGCAGAGCAGCAACGACC
GGCCGACTTCTCCAACCTCTCCAAGTACCAGAATGGAAATGGGATGAAATTGGGATGGATTTCATCACCGGACTTCCAAAACCCAAGGTGGATATGATTCTATATGGGTAGTTGTGGATCGATTAACCAAGGTAGCTCGATTCATTCCTGTGAAGACCACTATGGAGGGAACAAACTAGCGGGAGCTCTACTTCGCTAGGATCGTGAGTCTCCATGGCGCTGCCCAAGAAAATCGTATCTGATAGGGAAACCAATTCACCTCTCAC
TTTTGGAAGAAGCTCCAAGAAGAGCTGGGTACTCGATTGAATTTCAGCACCGCCTATCACCCGCAGACAGATGGACAGACCGAGCGTCTGAACCAGATCCTAGAAGACATGCTTCGTGCATGTGTACTAGATTTCGGGAAGACCTGGGATAAGAGTCTCCCCTATGCCGAGTTCTCCTACAATAACAGTTATCAAGCCAGCATACAAATGGCACCTTATGAAGCGTTGTACGGGGCGCAAATGCCGGACACCACTATTGTGGGACCAA
GTTGGAGAAAGCCAGGTGTTCGGGACTGATATTCTGAGAGAAGCTGAAGCTAAAGTCAGAATCATTCGGGATAATCTAAAGGTGGCACAGTCCCGGCAGAAAAGTTATGCAGATAACCGTCGCCGTGACTTGGAATTCGCAGTGGATGACTTTGTGTATCTTTGGGTCACGCCATTGCGAGGTGTACACAGGTTTCAGACGAAAGGGAAGCTCGCACCTCGGTATGTGGGTCCTTTCCGTATCGTTGCTCGACGTGGAGAAGTCGCT
TACCAGTTGGAGCTGCCCGCCTCCTTGGGCAACGTGCATGATGTGTTTCATGTCTCGCAACTCAAGAAGTGTCTCCGAGTGCCATCCGAGCAGGCCGACTCAGAGCAAATTGAAGTTCGGCGAAGACTTGACCTATGTGGAAAGGCCAGTGAAAATCCTGGACACCATGGAGCGCGAGGACCAGGAACCGGGTAATCCGTTTTTGCAAGGTCCAGTGGAGCAACCACGCCGAAGAAGAAGCTACCTGGGAGCGTGAAGACGAGCTGAA
GGCAGCCCATCCCGGTCTCTTCGCCAGTTCTTCCGAATCTCGGGGTCGAGATTCCGTTTAAGGGGGGTAGGTTTGTCGCGTCCCAAATCGACTCTAAAATACATCCTCGAAATTGTGCATAGAATAATTAAAATTCCTGTGAAAGCCTCCAACAATTAAAATGTGCGAAGATAAAAGTCAAATAAGGCTTGGAGGATTTTTGTATATTTCCTAAAAGCCTAAAATGCGTAAATAAATTTTAGTGGAATTTTCAGAGCACAAATAATAA
TTGAGAAGAAATAAACAAAGTTAAAAAGGTTATAAAAAGAAAACCCTAAAAGAAGTCCTTTTCCCCCCCTTCTCTCTCTTGGGCCGGCTCGGCCCACTTCTCTTCCCTCCCTCTCTCTCCTCCCCTCCGGCCCATCGGCTCCTCCCCGCGCGCGCGCGCGCTGACAGGCGGGGCCCGCCCGCCACCCTCGCTGACGGGTGGCACCCACCTGTCATCGCCTTCCTCGCGCCCGCGTCCGCCGCTCGCCCGTGCCCTAGCGCCGCCGCCGCCGCGG
GATTCCGCCGCATCCCGCCGTCCCCGCGCGCCATAAAGAGGGGGGAAATATTCCCCACGATTCCCTCTCTCTTTCCCCCCGTTTTCCCTCTCTCTCTCCCTCGGATTCGTCCGATTTTGTGCCCGCAAAGTTCGCCGGCGCAATCAATGGCGGCCGAGATCTCCGCGCCAGTTTCCATCTCCTCCGGCCGCTCTCTCCTCCTCTCGGCCCTATTTAATTGTTCCCCGTGCCTCTCTCCTCCGTTTCCGCCGTTCGCCGCTCGCTCTCACCGTC
GCATTCGTGCTCGCGCCGTCGCACCTTCTCCGCCGTCGCCGCCGAGCTCCGGTTCGCCGCCGCCGTCGCTCCGGACTTCCTCCCGACTCGGCGTCACCTCCATCGGCTTCGCCGGCGTCGTCATCGACCCCGTCCGCTCCTCCGTTTCGCTCGCCGACCGTCGGAGCGGCCGTCGTCCTCGTCGTCCCGAACCGCGCCGCCGCCGCTTCCTCGTCCGTTCGCCGCTCGACCGGGGTGAGCCGAGGACCGCCCTTTCGC
TTCCCTCTCCTCTCCGCTCTCTCGGGCGCCGCTCCGTCGCGCCCGGAGGTCGCCGCCGGTGACCATCGGGGCGCGGGTGCGTGCCGCCTCCCGTCGGCCGTGCCGGCGAGGCCGCCTTCGGGCGCGCCCTCGCGGCTGCCAAGTGGGCCCGGCCGTCAGCCGCCCGCCCCGGCGGGTGCGGCTGACGCGTGGGACCCACGGCGCCGGCCGCCGCCAGCCGCGTAAACCCGGTCCACCGTGGACCGAGAGGCTGACGCGTGGGCCCCACT
CCCGTGGACCCGGTCCGCGCGCCCTCTCCCTCGGCTGACGCAATAAATGAGATTTTTAATTAAAATTAAATGAAGAAATTCGCAAATATCCATTAAAGAGCATATAAACTTCAAATGGCCATATCTTGGCCATTTGAACTCGGAATTGGACCGTTCAAGTCTCTAATTTTTCCTTAAGAAGTCAAGAACCCATTTTTGTGCTTTGTTTCTGCTATTATGTGGAGTTTATTAGAGTGAAAGCCTTTTC



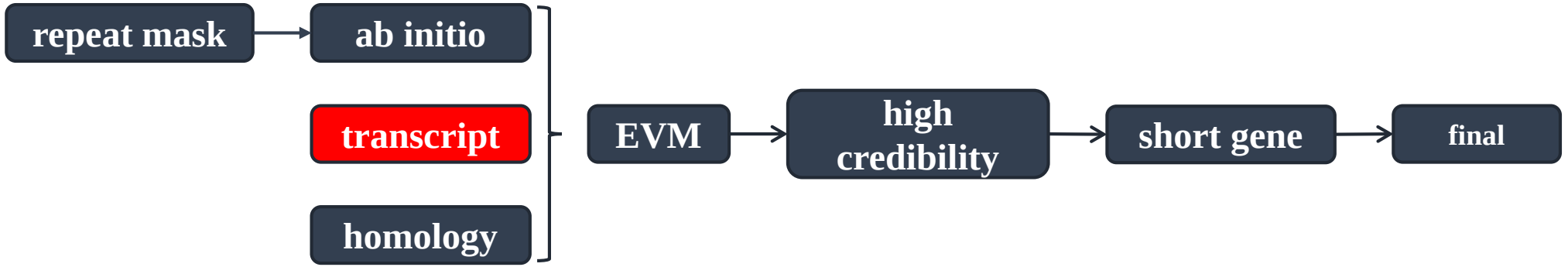
transcript转录组预测

·tophat比对，cufflinks组装转录本形成基因模型

·trinity组装转录本

·pasa整合上述结果进行预测

示例讲解（203）

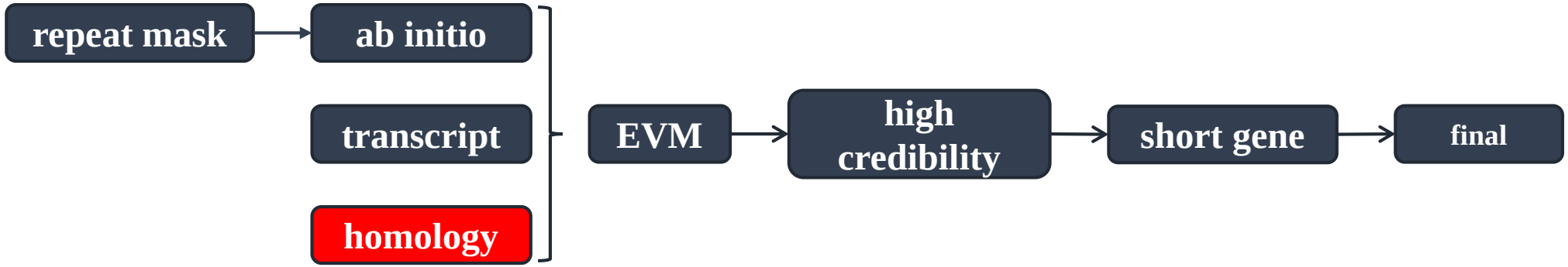


homology同源蛋白预测

·获取近缘物种蛋白序列

·使用gmap进行同源蛋白的比对，获得同源蛋白预测

示例讲解（203）



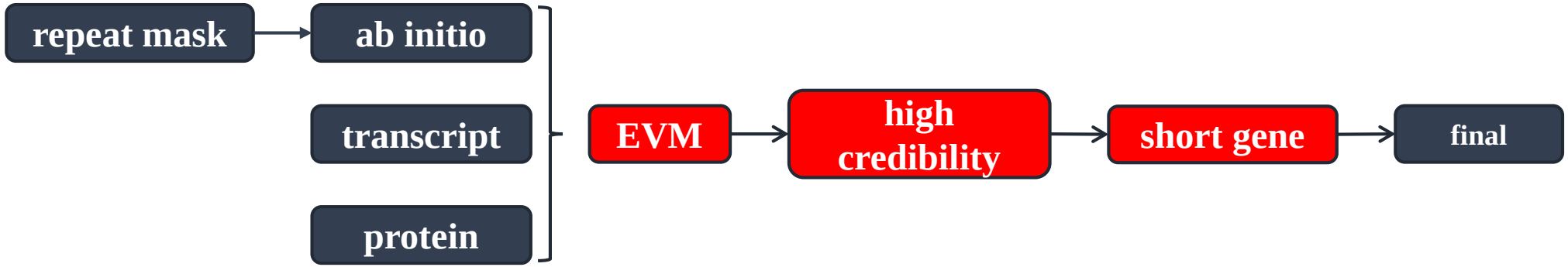
预测结果整合及过滤

·使用EVM（**EVidenceModeler**）将上述从头预测/转录组预测/同源蛋白预测结果整合成非冗余注释结果

·判断预测基因各种证据支持情况，确定可信基因集

·去除过短基因

EVM
https://evidencemodeler.github.io/



基因组注释

·结构注释

·功能注释

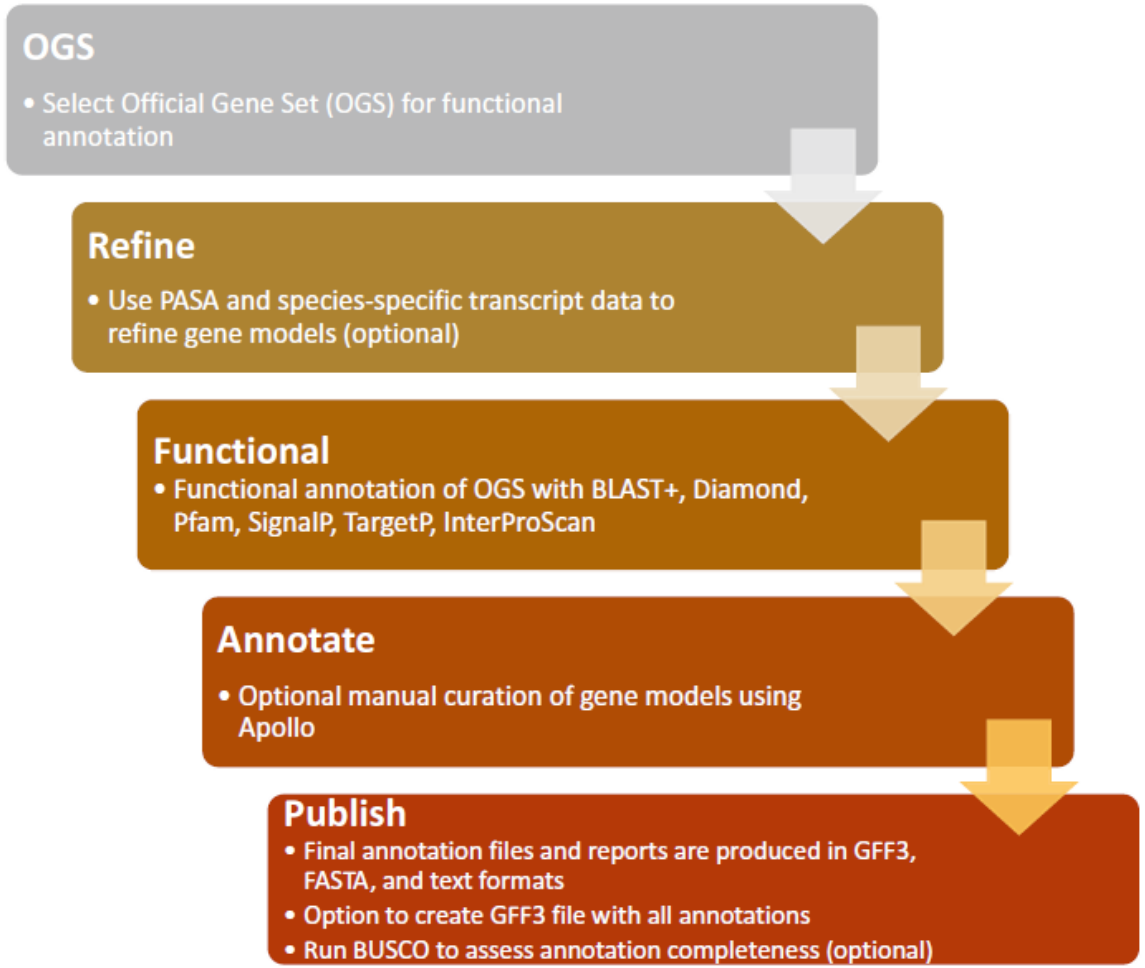
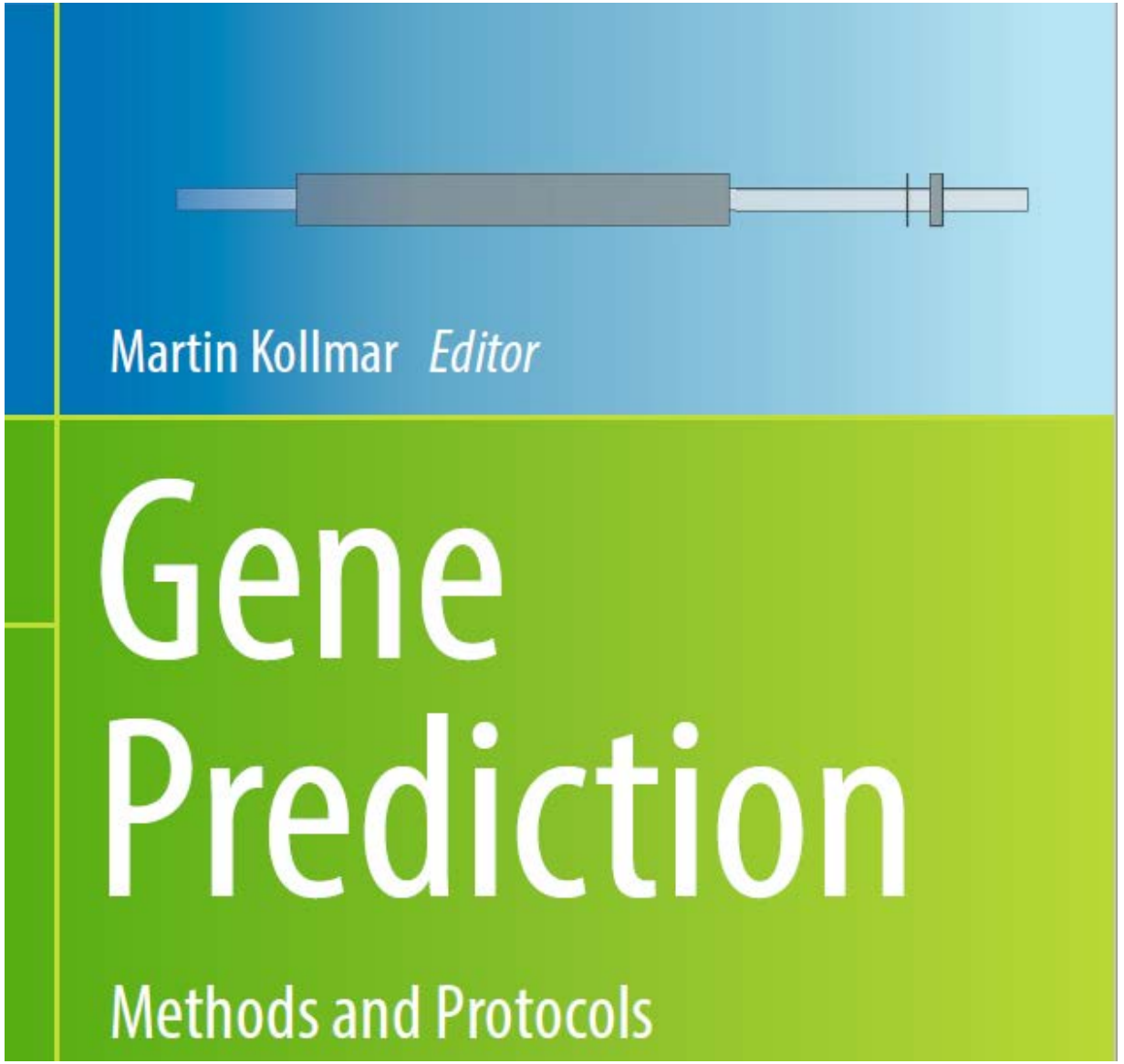


Fig. 9 Overview of the functional annotation, manual curation, and final steps of a GenSAS project

Jodi L. Humann *et al.*


Martin Kollmar Editor
Gene Prediction
Methods and Protocols

资源推荐

F1000Research F1000Research 2018, 7(ELIXIR):148 Last updated: 15 OCT 2019

Check for updates

OPINION ARTICLE

Ten steps to get started in Genome Assembly and Annotation [version 1; peer review: 2 approved]

UCDAVIS Bioinformatics Core

Training Workshops Course Material

To learn more about the UC Davis Bioinformatics data analysis services and training program, visit the Bioinformatics Core website. Dates and prices for our upcoming workshops can be found on Genome Center Event Registration site.

2020 september isoseq

Three day workshop offered in partnership with PacBio. Learn all about the possibilities of full length isoform sequencing from PacBio experts and experienced Core personnel!

2020 August Advanced scRNAseq

Building on material from Introduction to Single Cell RNA-Seq, this workshop aims to cover advanced topics, including TCR and trajectory analysis. Instruction in introductory analysis will not be provided.

2020 Genome Assembly Workshop

Explore the world of genome assembly and gain an understanding of the power and pitfalls of sequencing technologies and assembly techniques.

2020 mRNA Seq Workshop

Focused multi-day workshop covering the material needed for a successful RNA-Seq experiment.

2020 Variant Analysis Workshop

From SNP and structural variant calling to GWAS. Workshop cancelled due to public health concerns.

2019 Winter Bioinformatics Command Line and R Prerequisites Workshop

An introduction to the basics of bioinformatics for biologists. No data analysis experience necessary.

https://ucdavis-bioinformatics-training.github.io/

资源推荐

浙江大学 ZHEJIANG UNIVERSITY

EVOLUTION AND GENOMICS
Intensive and immersive training opportunities

WORKSHOPS | LEARNING | PEOPLE | APPLY | INFORMATION

GENOMICS

Here you will find learning activities designed for the Workshop on Genomics. These activities were designed by faculty in collaboration with the Workshop Team and will receive major updates during each Workshop and iterative refinement throughout the year.

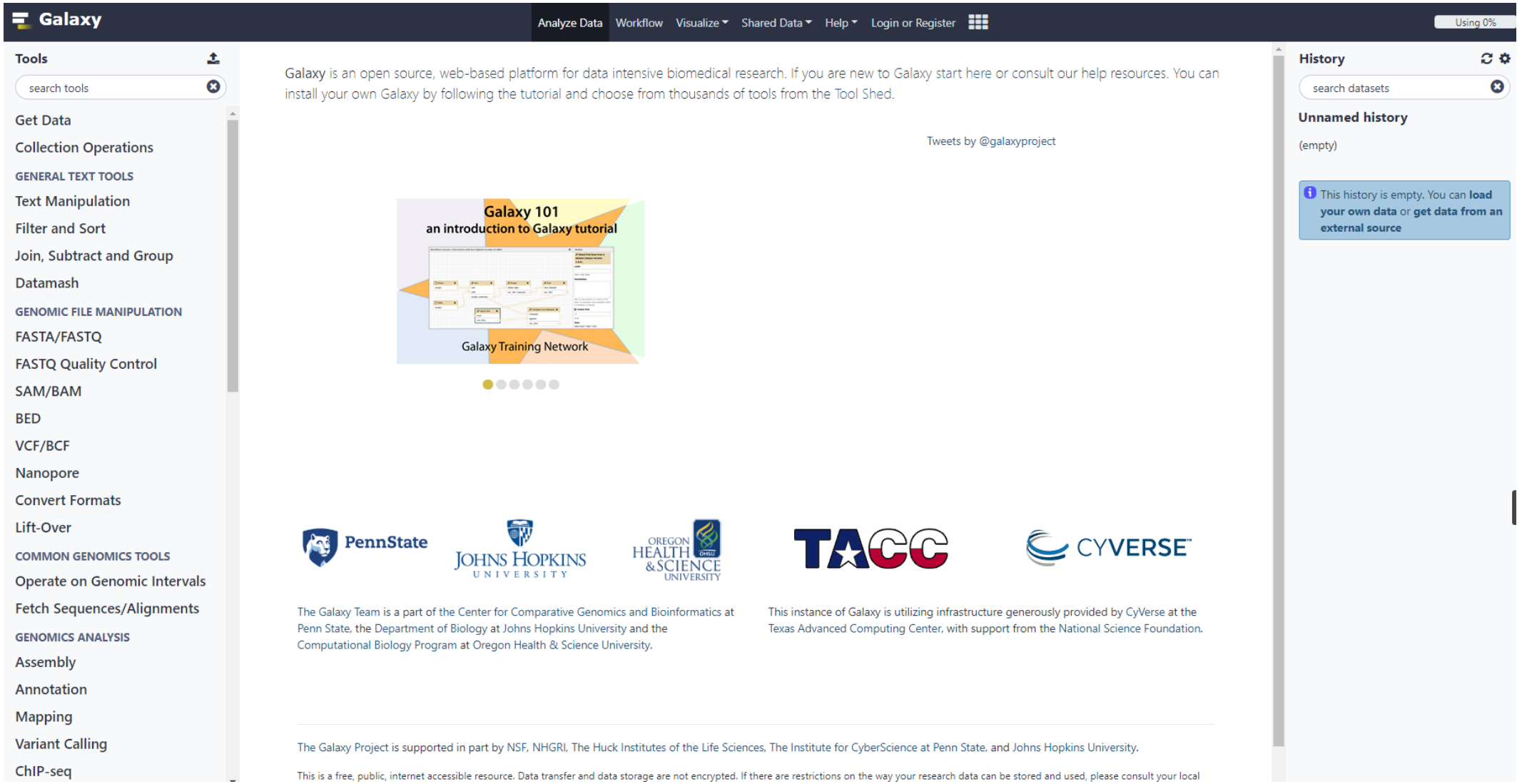
Please select the Learning Activity you are interested from the menu above. Alternatively, you can be taken to specific sections using the links below.

- Assembly
- Gene finding with AUGUSTUS
- Quality Assessment and Quality Control
- Read Mapping and Variant Calling
- Synteny Alignments using SatsumaSynteny
- Scripture
- Transcriptomics
- Stacks
- Metagenomics
  - Metagenome exploration with mg-RAST
  - PhyloPythiaS
  - Metagenomics using QIIME

http://evomics.org/learning/genomics/

资源推荐

浙江大学 ZHEJIANG UNIVERSITY



https://usegalaxy.org/

资源推荐

Linux命令大全(手册)

准确，丰富，稳定，在技术之路上为您护航!

请输入一个命令或相关功能 查询

1：文件管理

- ls命令 – 显示指定工作目录下的内容及属性信息
- pwd命令 – 显示当前路径
- cp命令 – 复制文件或目录
- mkdir命令 – 创建目录
- mv命令 – 移动或改名文件

2：文档编辑

- cat命令 – 在终端设备上显示文件内容
- echo命令 – 输出字符串或提取Shell变量的值
- rm命令 – 移除文件或目录
- tail命令 – 查看文件尾部内容
- rmdir命令 – 删除空目录

3：系统管理

- vmstat命令 – 显示虚拟内存状态
- startx命令 – 初始化X-windows
- uname命令 – 显示系统信息
- rpm命令 – RPM软件包管理器
- find命令 – 查找和搜索文件

4：磁盘管理

- df命令 – 显示磁盘空间使用情况
- fdisk命令 – 磁盘分区
- lsblk命令 – 查看系统的磁盘
- hdparm命令 – 显示与设定硬盘参数
- quota命令 – 显示磁盘已使用的空间与限制

每日学习

- lftpget命令 – 下载指定的文件
- builtin命令 – 执行bash内建命令
- apk命令 – 下载包管理工具
- apropos命令 – 在whatis数据库中查找字符串
- bmodinfo命令 – 显示给定模块的详细信息
- cancel命令 – 取消已存在的打印任务
- clockdiff命令 – 检测两台linux主机的时间差
- uucico命令 – 文件传输服务程序
- semanage命令 – 安全上下文查询与修改
- rpmverify命令 – 验证已安装的RPM软件包的正确性
- lsusb命令 – 显示USB设备列表
- setpci命令 – 配置PCI设备
- lvcreate命令 – 创建逻辑卷

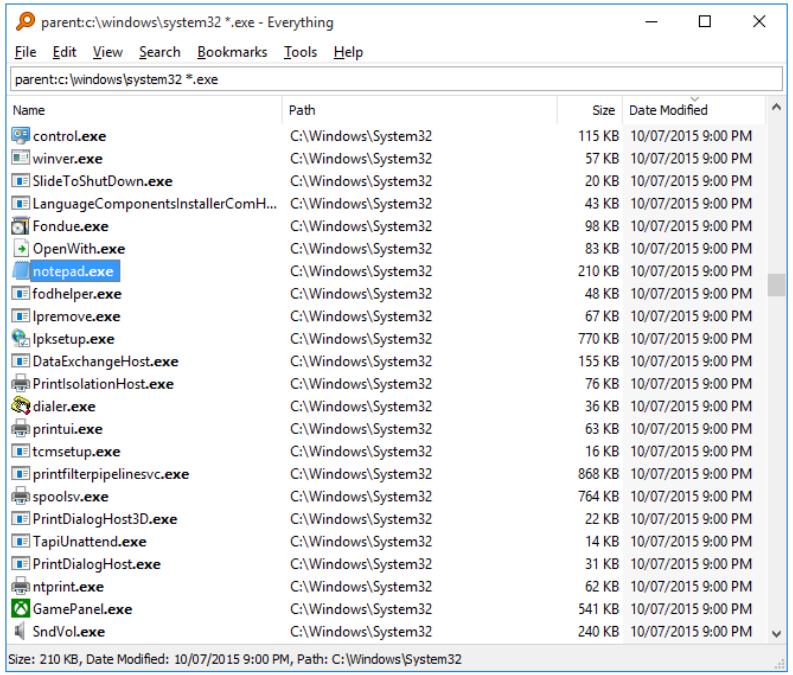
https://www.linuxcool.com/

资源推荐

voidtools

主页 下载 常见问题 支持 论坛 捐赠 联系方式

Everything



"*Everything*" 是 Windows 上文件名搜索引擎。

Everything 和其他搜索引擎有何不同

- 轻量安装文件。
- 干净简洁的用户界面。
- 快速文件索引。
- 快速搜索。

https://www.voidtools.com/zh-cn/support/everything/

资源推荐

浙江大学 ZHEJIANG UNIVERSITY

Thanks

贾磊

2020.9.24