Parsnp

基于基因组的比对是追踪基因组进化，准确推导重组，识别基因岛，分析移动遗传单元，对同源序列进行综合分类，构建ancestral基因组和phylogenomic分析的基础。全基因组比对的目的是通过构建每个基因组序列之间的分类关系(orthology, paralog, xenolog)进而反应其进化历史。

core-genome比对为全基因组比对的子集，主要用于识别保守存在于所有比对基因组中的orthologous序列。相对于复杂的多重比对，core-genome排出了subset relationships，比对更可行。此外，core-genome包含了垂直遗传的必要基因。构建phylogenies的最可信的变异信息就是SNPs，因此，core-genome SNP分型是当前用于亲缘微生物构建大的phylogeny的标准方式。

parsnp采用suffix graph data structure(mummer)识别MUMs(maximal unique matches)，然后在MUMs的基础上招募类似的基因组，卯定多重比对。parsnp比对输入目录下的MultiFASTA文件，输出core-genome alignment，variant calls和SNP tree(可使用Gingr查看)。

parsnp多重比对(core-genome)输出包含所有SNP, Indel和structural variation。所有多重比对中的多态性列都被标记用与识别：

- repetitive sequence
- **small LCB size (locally collinear blocks, these LCBs form the basic of the core-genome alignment)**
- Poor alignment quality
- poor base quality
- possible recombination

最终得到的一组core-genom SNPs使用FastTree2重建全基因组范围的phylogeny。

Usage

```
parsnp —p <threads> —d <directory of genomes> —r <ref genome>
```

使用参考序列和genbank文件

```
parsnp -g <reference_replicon1,reference_replicon2,..> -d <genome_dir> -p <threads>
```

Autorecruit reference to a draft assembly:

```
parsnp -q <draft_assembly> -d <genome_db> -p <threads>
```

genbank文件需要含有GI numbers用于indexing。这意味着custom genbank (not download from NCBI)文件，尽管可以用于比对，但是在Gingr中不会出现注释信息。

genbank文件可以仅指定reference genome

-g/-r选项不能同时使用，-r指定fasta格式文件，-g指定genbank文件

所有输入文件必须包含在-d指定的目录中，若强制性包含该目录下所有基因组，使用参数-c

-c 不考虑MUMi值，使用目录中所有基因组，默认为NO

-d 包含genomes/contigs/scaffolds的目录

-r 参考基因组，设置为! 用于从genome目录中随机挑选一个作为reference genome

-g genbank文件，使用逗号分隔一些列genbank文件

-q (optional) specify (assembled) query genome to use, in addition to genomes found in genome dir (default = NONE)

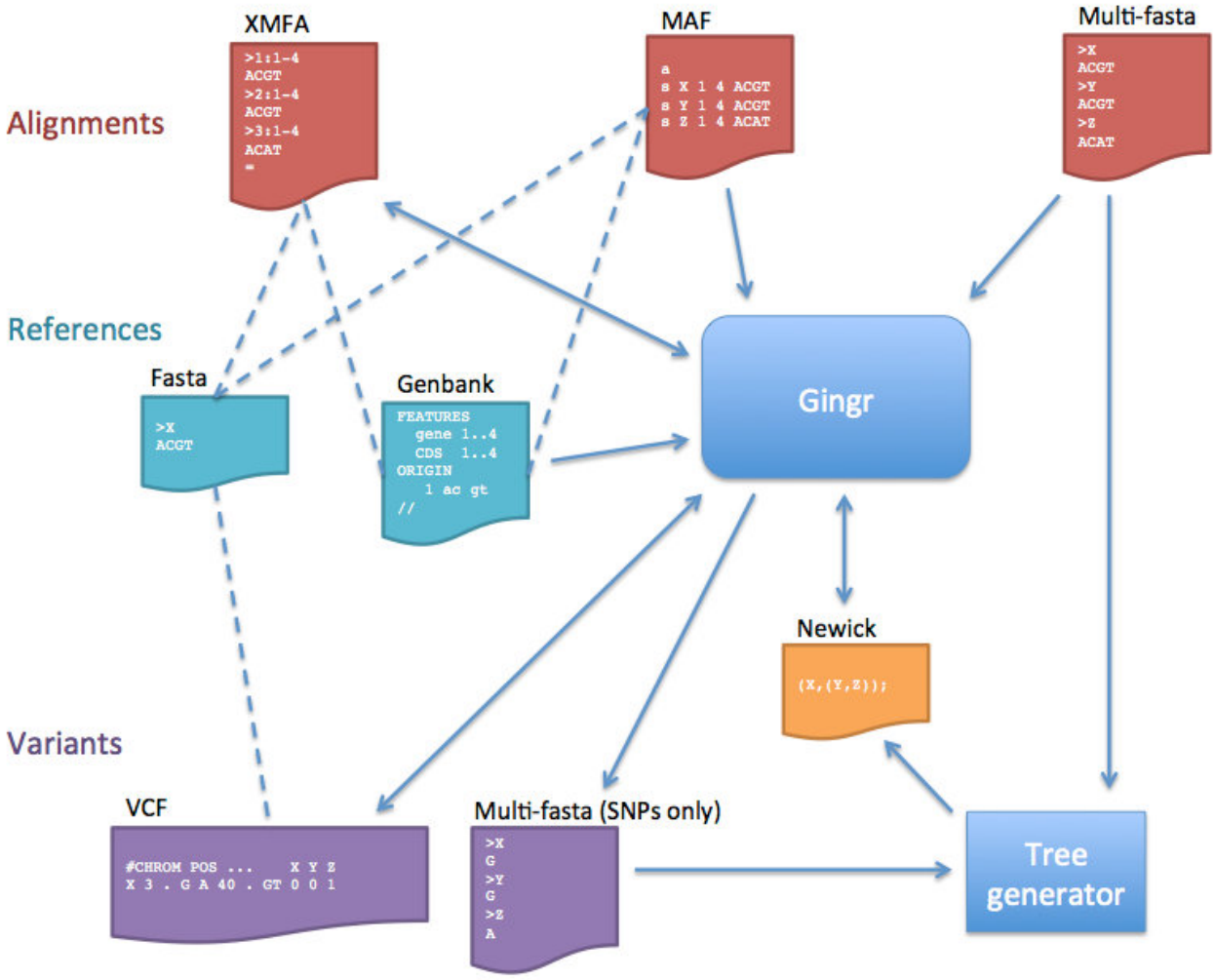
-x 使用PhiPack识别重组区域过滤掉该区域SNPs，默认NO

Output

1. Newick formatted core genome SNP tree: $outputdir/parsnp.tree
2. SNPs used to infer phylogeny: $outputdir/parsnp.vcf
3. Gingr formatted binary archive: $outputdir/parsnp.ggr
4. XMFA formatted multiple alignment: $outputdir/parsnp.xmfa

```
parsnp -c -d tmp_fasta_link/ -g GCF_000240185.1_ASM24018v2_genomic.gbff -o parsnp_analysis
```

Gingr



HarvestTools

基本用法

`harvesttools –x <input xmfa> -f <input reference fasta> -g <reference genbank formatted annotations> -n <newick formatted tree>`1

使用reference & genbank文件作为输入

`harvesttools -g <reference_genbank_file1> -r <reference fasta file> -x <XMFA file> -o hvt.ggr`

使用ggr文件输入，输出XMFA，同Gingr导出alignment(XMFA)

`harvesttools -i input.ggr -X output.xmfa`

使用ggr文件输入，输出fasta格式SNP文件，同Gingr到处variants(MFA)

`harvesttools -i input.ggr -S output.snps`

使用ggr文件输入，输出multi-fasta文件用于beast2/scotti输入，用于outbreak track分析(使用前过滤掉不用于比对genome序列)

`harvesttools -i input.ggr -M multi-align.fasta`
