Comparative genomics: Two closely related species genome alignment, structural variant detection & TE annotation pipeline

March 2, 2022

## **Genome alignment and variant detection**

* Genome alignment is useful because it tells you how the two species are related at genomic level.
* Relevant for phylogenetics, evolutionary analysis, population genetics etc.
* Genomic variants include large-scale structural variations (inversions, translocations, etc.) that are not captured by genome-wide genetic polymorphism scans like SNPs and InDels identifications. Whole genome alignment has advantages to capture synonymous changes which are more common than nonsynonymous, interesting genomic variants that affect the genome evolution in a non-traditional central dogma (Transposable elements or selfish elements that don’t code for particular genes but affects the organism’s phenotype, etc.)
* Each step in this pipeline can be performed individually as long as you have the proper input data. But as a whole pipeline, it helps you to produce a readily useful statistics that indicates interesting information about your species of interest at genomic level.
* Widely applicable to any organisms you’re interested in. No upper or lower limit in terms of genome complexity (theoretically), though the more complicated the genome the more computational resource it requires to run.

## Limitation of this pipeline

* relies on the quality of genome you have. Only chromosome-resolved genomes can be used as the input.
* The two species of interest must contain sufficient genomic region that is syntenic
* requires appropriate amount of computational resource as shown below:

|  |  |  |  |
| --- | --- | --- | --- |
| Input genome size | Step | required time (hrs) | CPU memory (GB) |
| 1 Prunus (202Mb) | minimap2 |  |  |
|  | SyRI |  | 10 |
|  | EDTA | 36\* | 5.5 |
| 2 Arachis (1.01Gb) | minimap2/SyRI | 8.7 | 72 |
|  | EDTA | 142\* | 18.9 |

\*Wall-clock time using 16 CPU, run by chromosome-by-chromosome.

General pipeline structure:

Case study: Genome comparison between almond (*Prunus dulcis*) and apricot (*Prunus armeniaca*)

 

As it was mentioned in the introduction, this pipeline is not limited to one study system. For example, you could apply this pipeline to your animal genomes, microorganisms, insects, or plants as long as high-quality genome is available for both species you’re comparing, and they are reasonably close in their phylogeny. Here, we will walk through the pipeline using *Prunus* species in plant kingdom to illustrate what each command/code does and outputs, and let you adjust to your needs.

To get started, there are a few things to know about coding language used. The pipeline combines R (.R; mostly used in visualization procedures) and shell script (.sh; for most software programs). All codes are accessible in my github page: