Run GENESPACE vignettes - Cotton genomes treated as tetraploids

JTLovell

22-Feb 2022

1. Global Parameters

```
baseDir <- "/Users/jlovell/Desktop/GENESPACE_data/results/"
rawAnnotationDir <- "/Users/jlovell/Desktop/GENESPACE_data"
mcscanDir <- "/Users/jlovell/Documents/comparative_genomics/programs/MCScanX"
nThreads <- 6
path2of <- "orthofinder" # since orthofinder is in the path via conda
```

2. Set parameters

NOTE: Wheat, Maize and Switchgrass WGD occurred after MRCA of all genomes below. Therefore, treat them as polyploid.

Since some ploidy > 1, run orthofinder again in syntenic blocks.

```
cotton <- list(
  wd = file.path(baseDir, "cotton4x_outgroup"),
  speciesIDs = "cotton",
  genomes = data.table(do.call("rbind", list(
    c(genome = "Gbarbadense", version = "Gbarbadense", ploidy = 2),
    c("Gdarwinii", "Gdarwinii", 2),
    c("Gtomentosum", "Gtomentosum", 2),
    c("Tcacao", "Tcacao_v2.1", 1)))))

cottonParams <- list(
  pepString = "fa",
  orthofinderInBlk = T,
  blkSize = 10,
  nGaps = 10)
```

3 Initialize the rho run with the above specified parameters

```
gparCotton4x <- with(cotton, init_genespace(
  genomeIDs = genomes$genome,
  versionIDs = genomes$version,
  ploidy = genomes$ploidy,
```

```
  outgroup = "Tcacao",
  speciesIDs = rep(speciesIDs, length(genomes$genome)),
  orthofinderInBlk = TRUE,
  pepString = "fa",
  wd = wd,
  nCores = nThreads,
  path2orthofinder = path2of,
  path2mcscanx = mcscanDir,
  rawGenomeDir = rawAnnotationDir))
```

```
## set working directory to /Users/jlovell/Desktop/GENESPACE_data/results/cotton4x_outgroup
##
## found raw gff files:
##   /Users/jlovell/Desktop/GENESPACE_data/cotton/Gbarbadense/annotation/Gbarbadense.gff3.gz
##  /Users/jlovell/Desktop/GENESPACE_data/cotton/Gdarwinii/annotation/Gdarwinii.gff3.gz
##  /Users/jlovell/Desktop/GENESPACE_data/cotton/Gtomentosum/annotation/Gtomentosum.gff3.gz
##  /Users/jlovell/Desktop/GENESPACE_data/cotton/Tcacao_v2.1/annotation/Tcacao_523_v2.1.gene.gff3.gz
##
## found raw peptide files:
##   /Users/jlovell/Desktop/GENESPACE_data/cotton/Gbarbadense/annotation/Gbarbadense.fa.gz
##  /Users/jlovell/Desktop/GENESPACE_data/cotton/Gdarwinii/annotation/Gdarwinii.fa.gz
##  /Users/jlovell/Desktop/GENESPACE_data/cotton/Gtomentosum/annotation/Gtomentosum.fa.gz
##  /Users/jlovell/Desktop/GENESPACE_data/cotton/Tcacao_v2.1/annotation/Tcacao_523_v2.1.protein_primary
##
## GENESPACE run initialized:
##  Initial orthofinder database generation method: default inside R
##  Orthology graph method: inBlock
```

4 Parse the raw annotations.

```
parse_phytozome(gsParam = gparCotton4x)
```

```
## parsed annotations for Gbarbadense exist and !overwrite, skipping
## parsed annotations for Gdarwinii exist and !overwrite, skipping
## parsed annotations for Gtomentosum exist and !overwrite, skipping
## parsed annotations for Tcacao exist and !overwrite, skipping
```

5 Get orthofinder results

```
gparCotton4x <- set_syntenyParams(gsParam = gparCotton4x)
gparCotton4x <- run_orthofinder(gsParam = gparCotton4x)
```

```
##  Running 'defualt' genespace orthofinder method
##  ############################################################
##  Cleaning out orthofinder directory and prepping run
##  Calculating blast results and running OrthoFinder
##  ####################################################
##  ####################################################
##
##
```

```
##  OrthoFinder version 2.5.4 Copyright (C) 2014 David Emms
##
##  2022-02-22 20:18:53 : Starting OrthoFinder 2.5.4
##  6 thread(s) for highly parallel tasks (BLAST searches etc.)
##  1 thread(s) for OrthoFinder algorithm
##
##  Checking required programs are installed
##  ----------------------------------------
##  Test can run "mcl -h" - ok
##  Test can run "fastme -i /Users/jlovell/Desktop/GENESPACE_data/results//cotton4x_outgroup/orthofinde
##
##  Dividing up work for BLAST for parallel processing
##  --------------------------------------------------
##  2022-02-22 20:18:55 : Creating diamond database 1 of 4
##  2022-02-22 20:18:56 : Creating diamond database 2 of 4
##  2022-02-22 20:18:57 : Creating diamond database 3 of 4
##  2022-02-22 20:18:58 : Creating diamond database 4 of 4
##
##  Running diamond all-versus-all
##  ------------------------------
##  Using 6 thread(s)
##  2022-02-22 20:18:58 : This may take some time....
##  2022-02-22 20:18:59 : Done 0 of 16
##  2022-02-22 21:08:24 : Done 10 of 16
##  2022-02-22 21:24:36 : Done all-versus-all sequence search
##
##  Running OrthoFinder algorithm
##  -----------------------------
##  2022-02-22 21:24:37 : Initial processing of each species
##  2022-02-22 21:25:45 : Initial processing of species 0 complete
##  2022-02-22 21:26:57 : Initial processing of species 1 complete
##  2022-02-22 21:28:07 : Initial processing of species 2 complete
##  2022-02-22 21:28:29 : Initial processing of species 3 complete
##  2022-02-22 21:28:48 : Connected putative homologues
##  2022-02-22 21:28:56 : Written final scores for species 0 to graph file
##  2022-02-22 21:29:04 : Written final scores for species 1 to graph file
##  2022-02-22 21:29:12 : Written final scores for species 2 to graph file
##  2022-02-22 21:29:15 : Written final scores for species 3 to graph file
##
##  WARNING: program called by OrthoFinder produced output to stderr
##
##  Command: mcl /Users/jlovell/Desktop/GENESPACE_data/results//cotton4x_outgroup/orthofinder/Results_F
##
##  stdout
##  ------
##  b''
##  stderr
##  ------
##  b'[mcl] cut <2> instances of overlap\n'
##  2022-02-22 21:29:58 : Ran MCL
##
##  Writing orthogroups to file
##  ---------------------------
##  OrthoFinder assigned 246344 genes (95.3% of total) to 44944 orthogroups. Fifty percent of all genes
```

```
##
##  2022-02-22 21:30:27 : Done orthogroups
##
##  Analysing Orthogroups
##  =====================
##
##  Calculating gene distances
##  --------------------------
##  2022-02-22 21:34:00 : Done
##  2022-02-22 21:34:03 : Done 0 of 26200
##  2022-02-22 21:34:06 : Done 1000 of 26200
##  2022-02-22 21:34:10 : Done 2000 of 26200
##  2022-02-22 21:34:13 : Done 3000 of 26200
##  2022-02-22 21:34:16 : Done 4000 of 26200
##  2022-02-22 21:34:20 : Done 5000 of 26200
##  2022-02-22 21:34:23 : Done 6000 of 26200
##  2022-02-22 21:34:27 : Done 7000 of 26200
##  2022-02-22 21:34:30 : Done 8000 of 26200
##  2022-02-22 21:34:34 : Done 9000 of 26200
##  2022-02-22 21:34:37 : Done 10000 of 26200
##  2022-02-22 21:34:41 : Done 11000 of 26200
##  2022-02-22 21:34:44 : Done 12000 of 26200
##  2022-02-22 21:34:48 : Done 13000 of 26200
##  2022-02-22 21:34:51 : Done 14000 of 26200
##  2022-02-22 21:34:55 : Done 15000 of 26200
##  2022-02-22 21:34:58 : Done 16000 of 26200
##  2022-02-22 21:35:02 : Done 17000 of 26200
##  2022-02-22 21:35:06 : Done 18000 of 26200
##  2022-02-22 21:35:09 : Done 19000 of 26200
##  2022-02-22 21:35:13 : Done 20000 of 26200
##  2022-02-22 21:35:16 : Done 21000 of 26200
##  2022-02-22 21:35:20 : Done 22000 of 26200
##  2022-02-22 21:35:23 : Done 23000 of 26200
##  2022-02-22 21:35:27 : Done 24000 of 26200
##  2022-02-22 21:35:30 : Done 25000 of 26200
##  2022-02-22 21:35:33 : Done 26000 of 26200
##
##  Inferring gene and species trees
##  --------------------------------
##
##  17402 trees had all species present and will be used by STAG to infer the species tree
##
##  Best outgroup(s) for species tree
##  ---------------------------------
##  2022-02-22 21:41:23 : Starting STRIDE
##  2022-02-22 21:41:30 : Done STRIDE
##  Observed 4081 well-supported, non-terminal duplications. 4081 support the best root and 0 contradic
##  Best outgroup for species tree:
##    Tcacao
##
##  Reconciling gene trees and species tree
##  ---------------------------------------
##  Outgroup: Tcacao
##  2022-02-22 21:41:30 : Starting Recon and orthologues
```

```
##  2022-02-22 21:41:30 : Starting OF Orthologues
##  2022-02-22 21:41:31 : Done 0 of 26200
##  2022-02-22 21:41:41 : Done 1000 of 26200
##  2022-02-22 21:41:46 : Done 2000 of 26200
##  2022-02-22 21:41:51 : Done 3000 of 26200
##  2022-02-22 21:41:56 : Done 4000 of 26200
##  2022-02-22 21:42:00 : Done 5000 of 26200
##  2022-02-22 21:42:03 : Done 6000 of 26200
##  2022-02-22 21:42:07 : Done 7000 of 26200
##  2022-02-22 21:42:10 : Done 8000 of 26200
##  2022-02-22 21:42:13 : Done 9000 of 26200
##  2022-02-22 21:42:17 : Done 10000 of 26200
##  2022-02-22 21:42:20 : Done 11000 of 26200
##  2022-02-22 21:42:24 : Done 12000 of 26200
##  2022-02-22 21:42:27 : Done 13000 of 26200
##  2022-02-22 21:42:31 : Done 14000 of 26200
##  2022-02-22 21:42:34 : Done 15000 of 26200
##  2022-02-22 21:42:38 : Done 16000 of 26200
##  2022-02-22 21:42:41 : Done 17000 of 26200
##  2022-02-22 21:42:44 : Done 18000 of 26200
##  2022-02-22 21:42:47 : Done 19000 of 26200
##  2022-02-22 21:42:50 : Done 20000 of 26200
##  2022-02-22 21:42:53 : Done 21000 of 26200
##  2022-02-22 21:42:55 : Done 22000 of 26200
##  2022-02-22 21:42:58 : Done 23000 of 26200
##  2022-02-22 21:43:00 : Done 24000 of 26200
##  2022-02-22 21:43:02 : Done 25000 of 26200
##  2022-02-22 21:43:05 : Done 26000 of 26200
##  2022-02-22 21:43:05 : Done OF Orthologues
##
##  Writing results files
##  =====================
##  2022-02-22 21:43:18 : Done orthologues
##
##  Results:
##      /Users/jlovell/Desktop/GENESPACE_data/results/cotton4x_outgroup/orthofinder/Results_Feb22/
##
##  CITATION:
##   When publishing work that uses OrthoFinder please cite:
##   Emms D.M. & Kelly S. (2019), Genome Biology 20:238
##
##   If you use the species tree in your work then please also cite:
##   Emms D.M. & Kelly S. (2017), MBE 34(12): 3267-3278
##   Emms D.M. & Kelly S. (2018), bioRxiv https://doi.org/10.1101/267914
```

Also set the synteny parameters as default.

```
gparCotton4x <- set_syntenyParams(
  gsParam = gparCotton4x,
  blkSize = 10,
  nGaps = 10)
```

6 Build synteny data

```
gparCotton4x <- find_orthofinderResults(gsParam = gparCotton4x)
gparCotton4x <- synteny(gsParam = gparCotton4x, overwrite = T)
```

```
## Parsing the gff files ...
##  Reading the gffs and adding orthofinder IDs ... Done!
##  Found 53624 global OGs for 231202 genes
##  QC-ing genome to ensure chromosomes/scaffolds are big enough...
##              Genome: n. chrs PASS/FAIL, n. genes PASS/FAIL, n. OGs PASS/FAIL
##      Gbarbadense: 98/779, 72940/1621, 69354/1439
##      Gdarwinii: 39/131, 77953/350, 73571/258
##      Gtomentosum: 37/90, 78145/193, 73638/142
##  All look good!
##  Defining collinear orthogroup arrays ...
##  Found the following counts of arrays / genome:
##      Gbarbadense: 4944 genes in 1929 collinear arrays
##      Gdarwinii: 5694 genes in 2193 collinear arrays
##      Gtomentosum: 5848 genes in 2265 collinear arrays
## Pulling synteny for 6 unique pairwise combinations of genomes
##  Running 1 chunks of up to 6 combinations each:
##  Chunk 1 / 1 (21:43:40) ... Done!
##  Gtomento-Gtomento: 1604647 (tot), 257779/177 (reg), 139907/412 (blk)
##  Gtomento-Gdarwini: 1603797 (tot), 250634/97 (reg), 131739/352 (blk)
##  Gdarwini-Gdarwini: 1616753 (tot), 258974/222 (reg), 140479/430 (blk)
##  Gtomento-Gbarbade: 1566227 (tot), 234802/100 (reg), 126515/374 (blk)
##  Gdarwini-Gbarbade: 1572604 (tot), 236868/98 (reg), 128921/315 (blk)
##  Gbarbade-Gbarbade: 1546417 (tot), 234179/933 (reg), 133183/1144 (blk)
## Defining synteny-constrained orthogroups ...
##  Found 67457 synteny-split OGs for 231202 genes
## Running orthofinder by region ...
##  genome combinat. : n. non-self genes, nOGs global/syntenic/inblk
##  Gbarbade-Gbarbade: 59635 genes, 34336 / 37563 / 31370
##  Gdarwini-Gbarbade: 136359 genes, 44561 / 49580 / 40895
##  Gdarwini-Gdarwini: 61874 genes, 36015 / 39326 / 32624
##  Gtomento-Gbarbade: 133899 genes, 44620 / 49494 / 40094
##  Gtomento-Gdarwini: 139191 genes, 46443 / 51918 / 42130
##  Gtomento-Gtomento: 62072 genes, 35503 / 38874 / 32682
## Combining synteny-constrained and inblock orthogroups ...
##  syn OGs: 67457, inblk OGs: 68923, combined OGs: 55131
##  Found the following counts of arrays / genome:
##      Gbarbadense: 6155 genes in 2394 collinear arrays
##      Gdarwinii: 7207 genes in 2769 collinear arrays
##      Gtomentosum: 7345 genes in 2824 collinear arrays
## Pulling synteny for 6 unique pairwise combinations of genomes
##  Running 1 chunks of up to 6 combinations each:
##  Chunk 1 / 1 (22:03:34) ... Done!
##  Gtomento-Gtomento: 1604647 (tot), 258437/177 (reg), 137640/416 (blk)
##  Gtomento-Gdarwini: 1603797 (tot), 250708/98 (reg), 126695/352 (blk)
##  Gdarwini-Gdarwini: 1616753 (tot), 259800/222 (reg), 138071/428 (blk)
##  Gtomento-Gbarbade: 1566227 (tot), 234837/100 (reg), 122107/374 (blk)
##  Gdarwini-Gbarbade: 1572604 (tot), 237229/98 (reg), 124422/311 (blk)
##  Gbarbade-Gbarbade: 1546417 (tot), 234441/933 (reg), 131433/1145 (blk)
```

```
## Found 55131 OGs across 231202 genes. gff3-like text file written to:
##  /Users/jlovell/Desktop/GENESPACE_data/results//cotton4x_outgroup/results/gffWithOgs.txt.gz
## Calculating syntenic block breakpoints ...
##  Found 4878 blocks. Text file written to:
##  /Users/jlovell/Desktop/GENESPACE_data/results//cotton4x_outgroup/results/syntenicBlocks.txt.gz:
```

7 Print session info

```
gparCotton4x_outgroup <- gparCotton4x
save(gparCotton4x, file = file.path(baseDir, "gparCotton4x_outgroup.rda"))
sessionInfo()
```

```
## R version 4.1.2 (2021-11-01)
## Platform: x86_64-apple-darwin17.0 (64-bit)
## Running under: macOS Big Sur 10.16
##
## Matrix products: default
## BLAS:   /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRblas.0.dylib
## LAPACK: /Library/Frameworks/R.framework/Versions/4.1/Resources/lib/libRlapack.dylib
##
## locale:
## [1] en_US.UTF-8/en_US.UTF-8/en_US.UTF-8/C/en_US.UTF-8/en_US.UTF-8
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] GENESPACE_0.9.3   data.table_1.14.2
##
## loaded via a namespace (and not attached):
##  [1] Rcpp_1.0.8                  compiler_4.1.2
##  [3] restfulr_0.0.13             GenomeInfoDb_1.30.1
##  [5] XVector_0.34.0              MatrixGenerics_1.6.0
##  [7] R.methodsS3_1.8.1           bitops_1.0-7
##  [9] R.utils_2.11.0              tools_4.1.2
## [11] zlibbioc_1.40.0             digest_0.6.29
## [13] lattice_0.20-45             evaluate_0.14
## [15] pkgconfig_2.0.3             rlang_1.0.0
## [17] Matrix_1.4-0                igraph_1.2.11
## [19] DelayedArray_0.20.0         cli_3.1.1
## [21] rstudioapi_0.13             yaml_2.2.2
## [23] parallel_4.1.2              xfun_0.29
## [25] fastmap_1.1.0               GenomeInfoDbData_1.2.7
## [27] rtracklayer_1.54.0          stringr_1.4.0
## [29] knitr_1.37                  Biostrings_2.62.0
## [31] S4Vectors_0.32.3            IRanges_2.28.0
## [33] grid_4.1.2                  stats4_4.1.2
## [35] Biobase_2.54.0              BiocParallel_1.28.3
## [37] XML_3.99-0.8                rmarkdown_2.11
## [39] magrittr_2.0.2              matrixStats_0.61.0
## [41] GenomicAlignments_1.30.0    Rsamtools_2.10.0
## [43] GenomicRanges_1.46.1        htmltools_0.5.2
```

```
## [45] BiocGenerics_0.40.0         SummarizedExperiment_1.24.0
## [47] stringi_1.7.6               RCurl_1.98-1.5
## [49] rjson_0.2.21                crayon_1.4.2
## [51] dbscan_1.1-10               BiocIO_1.4.0
## [53] R.oo_1.24.0
```