HPA_analyzing_note

YMX

4/26/2021

Index bam file

```
samtools index ./data/Y601.hap.bam
samtools index ./data/Y413.hap.bam
```

Simplify the bam file

```
tail -15 ~/genomics/genome/genome.info | awk ’{print $1}’ | awk ’{print ″bash src/split_bam.sh
Y601 Y413 ./data/params ″ $i ″ &″}’
```

The specific command lines are listed below:

```
## bash src/split_bam.sh Y601 Y413 ./data/params LG1 &
##  bash src/split_bam.sh Y601 Y413 ./data/params LG2 &
##  bash src/split_bam.sh Y601 Y413 ./data/params LG3 &
##  bash src/split_bam.sh Y601 Y413 ./data/params LG4 &
##  bash src/split_bam.sh Y601 Y413 ./data/params LG5 &
##  bash src/split_bam.sh Y601 Y413 ./data/params LG6 &
##  bash src/split_bam.sh Y601 Y413 ./data/params LG7 &
##  bash src/split_bam.sh Y601 Y413 ./data/params LG8 &
##  bash src/split_bam.sh Y601 Y413 ./data/params LG9 &
##  bash src/split_bam.sh Y601 Y413 ./data/params LG10 &
##  bash src/split_bam.sh Y601 Y413 ./data/params LG11 &
##  bash src/split_bam.sh Y601 Y413 ./data/params LG12 &
##  bash src/split_bam.sh Y601 Y413 ./data/params LG13 &
##  bash src/split_bam.sh Y601 Y413 ./data/params LG14 &
##  bash src/split_bam.sh Y601 Y413 ./data/params LG15 &
```

The syntenic haplotype blocks extraction

```
tail -15 ~/genomics/genome/genome.info | awk ’{print $1}’ | awk ’{print ″Rscript src/haplotype_
extraction.R Y601 Y413 ./data/params ″ $i ″ &″}’
```

The specific command lines are listed below:

```
## Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG1 &
##  Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG2 &
##  Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG3 &
##  Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG4 &
##  Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG5 &
##  Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG6 &
##  Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG7 &
##  Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG8 &
##  Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG9 &
##  Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG10 &
##  Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG11 &
##  Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG12 &
##  Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG13 &
##  Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG14 &
##  Rscript src/haplotype_extraction.R Y601 Y413 ./data/params LG15 &
```

Take the result of LG1 as an example:

```
## 1.1 Haplotype blocks extraction of reference
```

```
##  Read in sam file for reference
```

```
##                                 Hap Ref StartPos EndPos
## 1  LG1_ploidy_4_Bl_68_Hap_0_foInx_0 LG1     4303   4396
## 2  LG1_ploidy_4_Bl_68_Hap_1_foInx_0 LG1     4303   5472
## 3  LG1_ploidy_4_Bl_68_Hap_2_foInx_0 LG1     4316   5474
## 4  LG1_ploidy_4_Bl_68_Hap_3_foInx_0 LG1     4316   4398
## 5 LG1_ploidy_2_Bl_127_Hap_0_foInx_0 LG1     5490   6157
## 6 LG1_ploidy_2_Bl_127_Hap_1_foInx_0 LG1     5602   5711
##                         Block
## 1  LG1_ploidy_4_Bl_68_foInx_0
## 2  LG1_ploidy_4_Bl_68_foInx_0
## 3  LG1_ploidy_4_Bl_68_foInx_0
## 4  LG1_ploidy_4_Bl_68_foInx_0
## 5 LG1_ploidy_2_Bl_127_foInx_0
## 6 LG1_ploidy_2_Bl_127_foInx_0
```

```
## Haplotype blocks for reference
```

```
##   Ploidy                       Block Ref Start  End Length
## 1      4  LG1_ploidy_4_Bl_68_foInx_0 LG1  4316 4396     81
## 2      2 LG1_ploidy_2_Bl_127_foInx_0 LG1  5602 5711    110
## 3      2 LG1_ploidy_2_Bl_134_foInx_0 LG1  6160 6838    679
## 4      4  LG1_ploidy_4_Bl_44_foInx_0 LG1  6859 7432    574
## 5      3 LG1_ploidy_3_Bl_100_foInx_0 LG1  7173 7808    636
## 6      3 LG1_ploidy_3_Bl_105_foInx_0 LG1  9134 9565    432
```

```
## The number of haplotype blocks for reference is: 36058
```

```
## The result is saved as: Y601_LG1_ref_blocks.txt
## 1.2 Haplotype blocks extraction of wild relative
```

```
##  Read in sam file for wild relative
```

```
##                                  Hap Ref StartPos EndPos
## 1 LG1_ploidy_2_Bl_130_Hap_0_foInx_0 LG1     3620   4352
## 2 LG1_ploidy_2_Bl_130_Hap_1_foInx_0 LG1     3620   4352
## 3   LG1_ploidy_4_Bl_0_Hap_0_foInx_0 LG1     4371   6157
## 4   LG1_ploidy_4_Bl_0_Hap_1_foInx_0 LG1     5602   6157
## 5   LG1_ploidy_4_Bl_0_Hap_2_foInx_0 LG1     5602   5693
## 6   LG1_ploidy_4_Bl_0_Hap_3_foInx_0 LG1     5602   6157
##                         Block
## 1 LG1_ploidy_2_Bl_130_foInx_0
## 2 LG1_ploidy_2_Bl_130_foInx_0
## 3   LG1_ploidy_4_Bl_0_foInx_0
## 4   LG1_ploidy_4_Bl_0_foInx_0
## 5   LG1_ploidy_4_Bl_0_foInx_0
## 6   LG1_ploidy_4_Bl_0_foInx_0
```

```
## Haplotype blocks for wild relative
```

```
##   Ploidy                       Block Ref Start   End Length
## 1      2 LG1_ploidy_2_Bl_130_foInx_0 LG1  3620  4352    733
## 2      4   LG1_ploidy_4_Bl_0_foInx_0 LG1  5602  5693     92
## 3      3 LG1_ploidy_3_Bl_102_foInx_0 LG1  6859  7155    297
## 4      2 LG1_ploidy_2_Bl_134_foInx_0 LG1  7174  8993   1820
## 5      4   LG1_ploidy_4_Bl_1_foInx_0 LG1  9117  9572    456
## 6      3  LG1_ploidy_3_Bl_92_foInx_0 LG1  9574 10321    748
```

```
## The number of haplotype blocks for wild relative is: 29865
```

```
## The result is saved as: Y413_LG1_WR_blocks.txt
```

```
## 2. Homologous blocks between reference and wild relative
```

```
## 2.1 Extract the syntenic haplotype blocks between reference and wild relative
##        WRPloidy                      WRBlock Ref  WRStart    WREnd RefPloidy
## 128959        4 LG1_ploidy_4_Bl_36_foInx_228 LG1 22827576 22892408         6
## 130717        4  LG1_ploidy_4_Bl_0_foInx_249 LG1 24900504 24934984         6
## 16595         4 LG1_ploidy_4_Bl_27_foInx_157 LG1 15740075 15774842         6
## 112358        4  LG1_ploidy_4_Bl_42_foInx_42 LG1  4245061  4278308         6
## 122179        4 LG1_ploidy_4_Bl_55_foInx_139 LG1 13944379 13992111         6
## 112364        4  LG1_ploidy_4_Bl_42_foInx_42 LG1  4245061  4278308         6
##                            RefBlock RefStart   RefEnd    Start      End Length
## 128959 LG1_ploidy_6_Bl_22_foInx_228 22856242 22892438 22856242 22892408  36167
## 130717  LG1_ploidy_6_Bl_0_foInx_249 24900402 24934687 24900504 24934687  34184
## 16595   LG1_ploidy_6_Bl_9_foInx_157 15741247 15774876 15741247 15774842  33596
## 112358  LG1_ploidy_6_Bl_35_foInx_42  4244216  4276068  4245061  4276068  31008
## 122179 LG1_ploidy_6_Bl_52_foInx_139 13955461 13986336 13955461 13986336  30876
## 112364  LG1_ploidy_6_Bl_30_foInx_42  4247702  4286479  4247702  4278308  30607
```

```
## The number of syntenic haplotype blocks is: 53507
```

```
## 2.22 Remove blocks with shorter sequence than length threshhold
```

```
## The length threshhold is: 3
```

```
## The number of filtered syntenic haplotype blocks is: 53465
```

```
## The result is saved as: Y601_Y413_LG1_ref_WR_all_blocks.txt
```

```
## 2.3 Remove duplicated syntenic haplotype blocks
```

```
## The number of filtered syntenic haplotype blocks is: 35738
```

```
## The result is saved as: Y601_Y413_LG1_ref_WR_filtered_blocks.txt
```

```
## 3. Calculate cigar position
```

```
## 3.1 Calculate cigar position for reference species
```

```
## An example of cigar information extracted of reference species
```

```
##    Cigar IDpos Length CLength
## 1 235M1I   235      1       1
## 2  26M1D   261     -1       0
## 3   3M2I   264      2       2
```

```
## 3.2 Calculate cigar position for wild relative
```

```
## An example of cigar information extracted of wild relative
##    Cigar IDpos Length CLength
## 1 235M1I   235      1       1
## 2  26M1D   261     -1       0
## 3   3M2I   264      2       2
```

```
## 4. Extract sequence
```

```
## 4.1 Extract sequence of reference species
```

```
##                       WRBlock                   RefBlock
## 1 LG1_ploidy_2_Bl_130_foInx_0 LG1_ploidy_4_Bl_68_foInx_0
## 2 LG1_ploidy_2_Bl_130_foInx_0 LG1_ploidy_4_Bl_68_foInx_0
## 3 LG1_ploidy_2_Bl_130_foInx_0 LG1_ploidy_4_Bl_68_foInx_0
## 4 LG1_ploidy_2_Bl_130_foInx_0 LG1_ploidy_4_Bl_68_foInx_0
##                                      Seq
## 1 TCCTTTTTGCTCAAAANNNNNNNNNNNNNAAAATGTA
## 2 TCCTTTTTGCTCAAAANNNNNNNNNNNNNAAAATGTA
## 3 TCCTTTATGCTCAAAANNNNNNNNNNNNNAAAATGTA
## 4 TCCTTTATGCTCAAAANNNNNNNNNNNNNAAAATGTA
```

```
## 4.2 Extract sequence of  wild relative
```

```
##                       WRBlock                   RefBlock
## 1 LG1_ploidy_2_Bl_130_foInx_0 LG1_ploidy_4_Bl_68_foInx_0
## 2 LG1_ploidy_2_Bl_130_foInx_0 LG1_ploidy_4_Bl_68_foInx_0
##                                       Seq
## 1 TCCTTTATGCTCAAAAGCTTCAAAAACACTCAAATGTA
## 2 TCCTTTATGCTCAAAAGCTTAAAAAACACTCAAATGTA
```

```
## 4.3 Remove blocks with the same sequence of all haplotypes and save as fasta file
```

```
## The number of syntenic haplotype block: 35738
```

```
## The final number of syntenic haplotype block: 33377
```

```
## The result is saved as: Y601_Y413_LG1_ref_WR_final_blocks.txt
```

High-throughput phylogenetic reconstruction

```
bash src/tree_building.sh ../../src/infer_UPGMA_nucleotide.mao 128
```

Tree topology analysis

```
Rscript src/topology_analysis.R Y601 Y413 ./data/params
```

```
## 1.Read commandArgs and config
## Topology analysis is performed for the following chromosomes:
```

```
## [1] "LG1"
```

```
## 2.Topology analysis
```

```
## LG1
```

```
## Topological result of each syntenic haplotype block
```

```
##                                       SeqName Ploidy Monophy  BranchLen UMonophy
## 23  Y601_Y413_LG1_10023315_10023914_11935.fa    2_2    TRUE 0.02278428     TRUE
## 31  Y601_Y413_LG1_10033338_10033824_11944.fa    2_2   FALSE 0.00000000    FALSE
## 106 Y601_Y413_LG1_10107581_10108462_12008.fa    2_2   FALSE 0.00000000    FALSE
## 109 Y601_Y413_LG1_10109814_10110683_12011.fa    2_2   FALSE 0.00000000    FALSE
## 210 Y601_Y413_LG1_10196643_10196939_12113.fa    2_2   FALSE 0.00000000    FALSE
## 282 Y601_Y413_LG1_10241306_10241335_12183.fa    2_2   FALSE 0.00000000    FALSE
##     UBranchLen
## 23        11.5
## 31         0.0
## 106        0.0
## 109        0.0
## 210        0.0
## 282        0.0
```

```
## The result is saved as: Y601_Y413_tree_topology_info.txt
```

```
## Statistic summary of LG1
##      WR Chr Ploidy   Num    Percent MeanLength GenomeCoverage  IQMonoRatio
## 1  Y413 LG1    2_2   844 0.03168168   360.3661     0.01005809   0.39691943
## 2  Y413 LG1    2_3   835 0.03134384   375.6383     0.01037254   0.33173653
## 3  Y413 LG1    2_4   783 0.02939189   421.8493     0.01092315   0.28863346
## 4  Y413 LG1    3_2  1534 0.05758258   312.1714     0.01583607   0.19882660
## 5  Y413 LG1    3_3  2347 0.08810060   326.4648     0.02533836   0.10950149
## 6  Y413 LG1    3_4  1836 0.06891892   346.8965     0.02106209   0.08605664
## 7  Y413 LG1    4_2  1725 0.06475225    280.382     0.01599441   0.12695652
## 8  Y413 LG1    4_3  2683 0.10071321   274.4793     0.02435338   0.06597093
## 9  Y413 LG1    4_4  3013 0.11310060   312.0873     0.03109597   0.04281447
## 10 Y413 LG1    5_2  1122 0.04211712   270.7086     0.01004439   0.10516934
## 11 Y413 LG1    5_3  1716 0.06441441    253.817     0.01440346   0.05186480
## 12 Y413 LG1    5_4  2363 0.08870120    307.182     0.02400426   0.02073635
## 13 Y413 LG1    6_2  1200 0.04504505     458.42     0.01819172   0.14416667
## 14 Y413 LG1    6_3  1754 0.06584084   437.9487     0.02540281   0.08551881
## 15 Y413 LG1    6_4  2885 0.10829580   753.9948     0.07193547   0.05095321
## 16 Y413 LG1    All 26640 1.00000000   373.4686     0.32901619 936.33333333
##      IQBranchLen UPGMAMonoRatio UPGMABranchLen
## 1  0.0225909904     0.39691943      1.7218602
## 2  0.0056669706     0.33173653      1.0889222
## 3  0.0047117214     0.28863346      1.1401660
## 4  0.0051823195     0.19882660      0.7199587
## 5  0.0018280767     0.10950149      0.2599063
## 6  0.0015501617     0.08605664      0.2808869
## 7  0.0084902852     0.12695652      0.4300000
## 8  0.0013480183     0.06597093      0.1509038
## 9  0.0006172463     0.04281447      0.1200077
## 10 0.0024120394     0.10516934      0.4124777
## 11 0.0009483052     0.05186480      0.1444444
## 12 0.0004945897     0.02073635      0.1236846
## 13 0.0022843471     0.14416667      0.7131389
## 14 0.0059681211     0.08551881      0.1873416
## 15 0.0004940539     0.05095321      0.4957486
## 16 0.0031093643   936.33333333      0.3983151
```

```
## The result is saved as: Y601_Y413_topology_info_ploidy_based_summary.txt
```

Gene conversion analysis

```
Rscript src/gene_conversion_analysis.R Y601 Y413 ./data/params
```

```
## 1. Read commandArgs and config
```

```
## 2. Gene conversion analysis
```

```
## 2.1 Extract blocks of specific ploid
```

```
## The number of total tree: 370865
```

```
## The number of 6:4 tree: 38763
## 2.2 Extract blocks within gene region
```

```
## The number of 6:4 tree in gene region: 21519
```

```
## 2.3 Gene conversion identified by tree topology
```

```
##                                                    Tree Tmono Topo DeRefl
## 1 Y601_Y413_LG10_10006595_10008143_10269.fa-221145.nwk     1  1:9      T
## 2 Y601_Y413_LG10_10048207_10049241_10328.fa-236299.nwk     0  4:6   ATTT
## 3 Y601_Y413_LG10_10103281_10103887_10386.fa-250244.nwk     0  4:6   AAAT
## 4 Y601_Y413_LG10_10113124_10114127_10397.fa-253066.nwk     2  2:8     TT
## 5    Y601_Y413_LG10_1011351_1017478_1074.fa-253395.nwk     0  2:8     AT
## 6  Y601_Y413_LG10_10200255_10200745_10522.fa-21700.nwk     1  1:9      T
##       Type
## 1 di2tetra
## 2 tetra2di
## 3 di2tetra
## 4   Others
## 5 di2tetra
## 6 di2tetra
```

```
## The result is saved as: Y601_Y413_topology_info_ploidy_based_summary.txt
```

```
## The ratio from B2 to B1:
```

```
## 0.09716994
```

```
## The ratio from B1 to B2:
```

```
## 0.3934662
```

```
## How many trees are used for gene conversion analyse?
```

```
## 21519
```

```
sessionInfo()
```

```
## R version 4.0.2 (2020-06-22)
## Platform: x86_64-pc-linux-gnu (64-bit)
## Running under: Ubuntu 16.04.7 LTS
##
## Matrix products: default
## BLAS:   /usr/lib/libblas/libblas.so.3.6.0
## LAPACK: /usr/lib/lapack/liblapack.so.3.6.0
##
## locale:
##  [1] LC_CTYPE=en_US.UTF-8       LC_NUMERIC=C
##  [3] LC_TIME=en_US.UTF-8        LC_COLLATE=en_US.UTF-8
##  [5] LC_MONETARY=en_US.UTF-8    LC_MESSAGES=en_US.UTF-8
##  [7] LC_PAPER=en_US.UTF-8       LC_NAME=C
##  [9] LC_ADDRESS=C               LC_TELEPHONE=C
## [11] LC_MEASUREMENT=en_US.UTF-8 LC_IDENTIFICATION=C
##
## attached base packages:
## [1] stats     graphics  grDevices utils     datasets  methods   base
##
## other attached packages:
## [1] plyr_1.8.6     phangorn_2.5.5 ape_5.4-1      stringr_1.4.0
##
## loaded via a namespace (and not attached):
##  [1] igraph_1.2.6      Rcpp_1.0.6        knitr_1.31        magrittr_2.0.1
##  [5] lattice_0.20-41   R6_2.5.0          quadprog_1.5-8    rlang_0.4.10
##  [9] fastmatch_1.1-0   tools_4.0.2       parallel_4.0.2    grid_4.0.2
## [13] nlme_3.1-152      xfun_0.21         jquerylib_0.1.3   htmltools_0.5.1.1
## [17] yaml_2.2.1        digest_0.6.27     Matrix_1.3-2      sass_0.3.1
## [21] evaluate_0.14     rmarkdown_2.7     stringi_1.5.3     compiler_4.0.2
## [25] bslib_0.2.4       jsonlite_1.7.2    pkgconfig_2.0.3
```