Report 01 - Exploratory analysis of distance and its implication in duplication and speciation events

02/09/2022

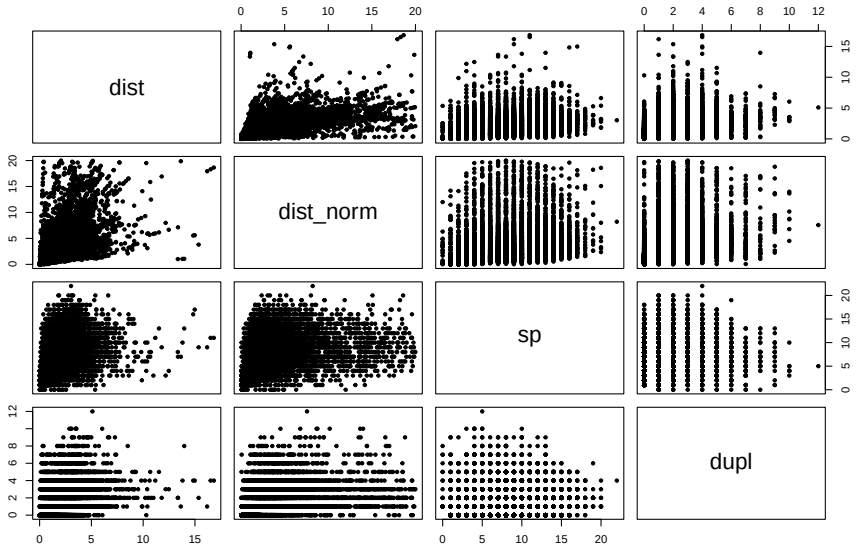
Yeast

Retrieved data

```
##   id             tree    prot             from from_sp
## 1  5 Phy000CX7L_YEAST YFL059W Phy000EXNX_YARLI   YARLI
## 2  5 Phy000CX7L_YEAST YFL059W Phy000EXNX_YARLI   YARLI
## 3  5 Phy000CX7L_YEAST YFL059W Phy000EXNX_YARLI   YARLI
## 4  5 Phy000CX7L_YEAST YFL059W Phy000EXNX_YARLI   YARLI
## 5  5 Phy000CZPL_YEAST YNR030W Phy000EYAK_YARLI   YARLI
## 6  5 Phy000CX7L_YEAST YFL059W Phy000EXNX_YARLI   YARLI
```

```
##                 to to_sp     dist dist_norm sp
## 1 Phy0005LO6_DEBHA DEBHA 0.701397  3.898599  3
## 2 Phy000HOB6_PICGU PICGU 0.715596  3.977522  3
## 3 Phy000JQJD_KLUWA KLUWA 0.789342  4.387427  4
## 4 Phy000NVAC_SACKL SACKL 0.769062  4.274704  5
## 5 Phy000JQB2_KLUWA KLUWA 2.110419  1.343531  3
## 6 Phy000NQDH_SACCA SACCA 0.801917  4.457323  5
```

Relationships between raw numerical variables

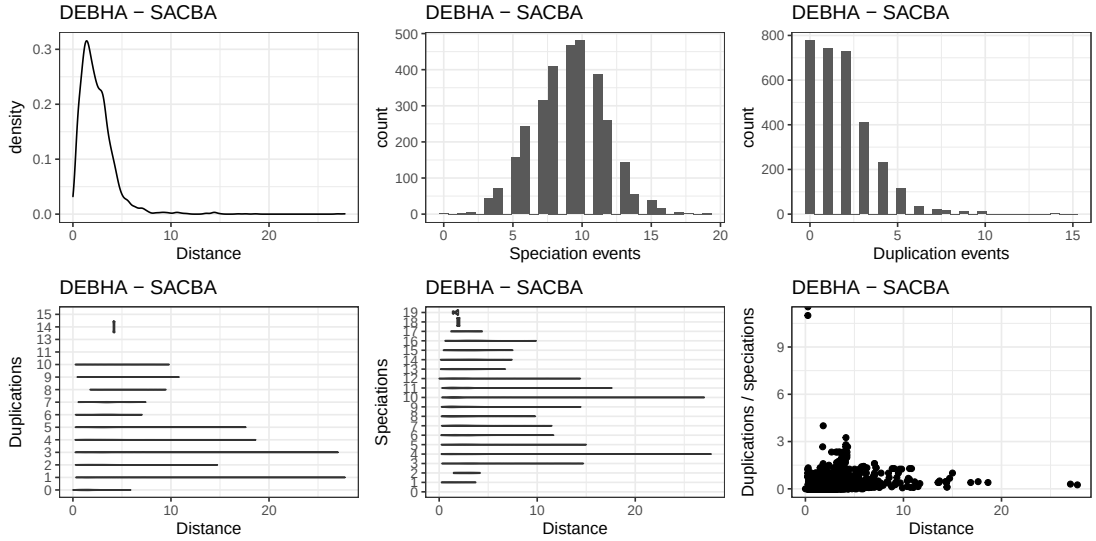


Summarizing data

We retrieve basic descriptive statistics below for each pair of species sets:

`Min. 1st Qu. Median Mean 3rd Qu. Max. skew kurt sum`

Also we get the basic representations of each sp to sp comparisons:



What we have for each phylome species pair

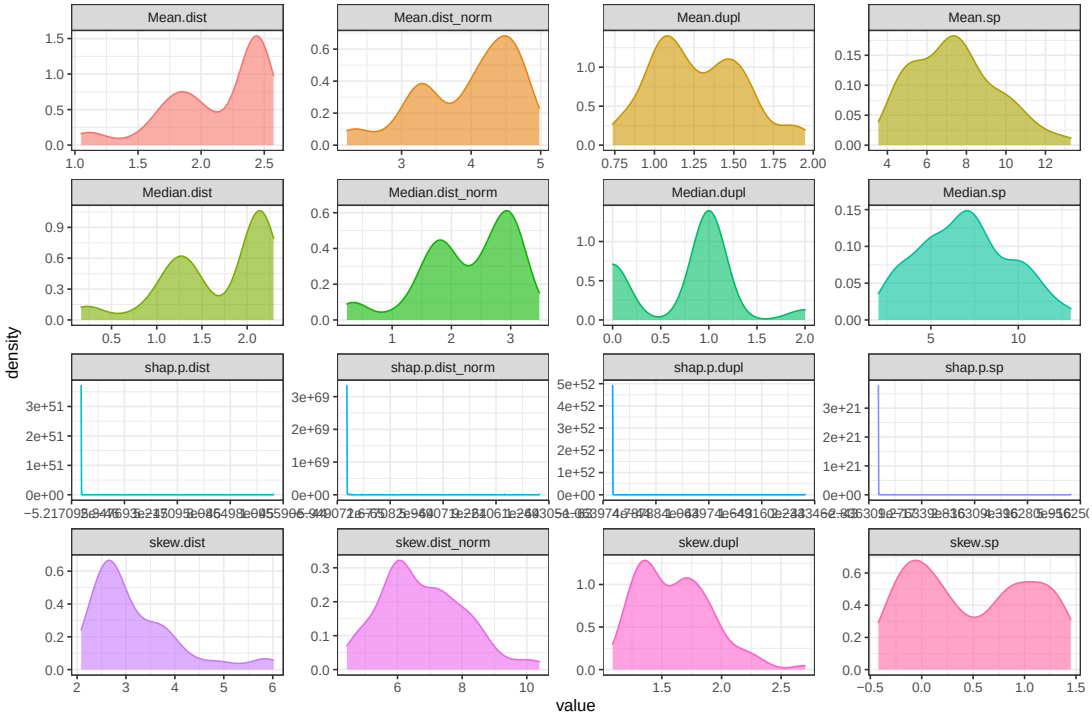
```
##   Median.dist Median.dist_norm Median.sp Median.dupl
## 1    2.250837         3.010814         5           1
## 2    2.292345         3.212353         5           1
## 3    1.970266         2.800025         5           1
## 4    2.190021         3.483268         8           2
## 5    2.027600         2.813689         5           1
## 6    2.218559         3.396224         8           2
```

```
##   Mean.dist Mean.dist_norm  Mean.sp Mean.dupl
## 1  2.535179       4.573514 4.690312  1.548642
## 2  2.573890       4.790191 5.384278  1.735832
## 3  2.356550       4.332298 4.655675  1.562215
## 4  2.489323       4.980213 7.731469  1.925493
## 5  2.380385       4.386917 5.072977  1.627173
## 6  2.519712       4.853984 8.013307  1.910317
```

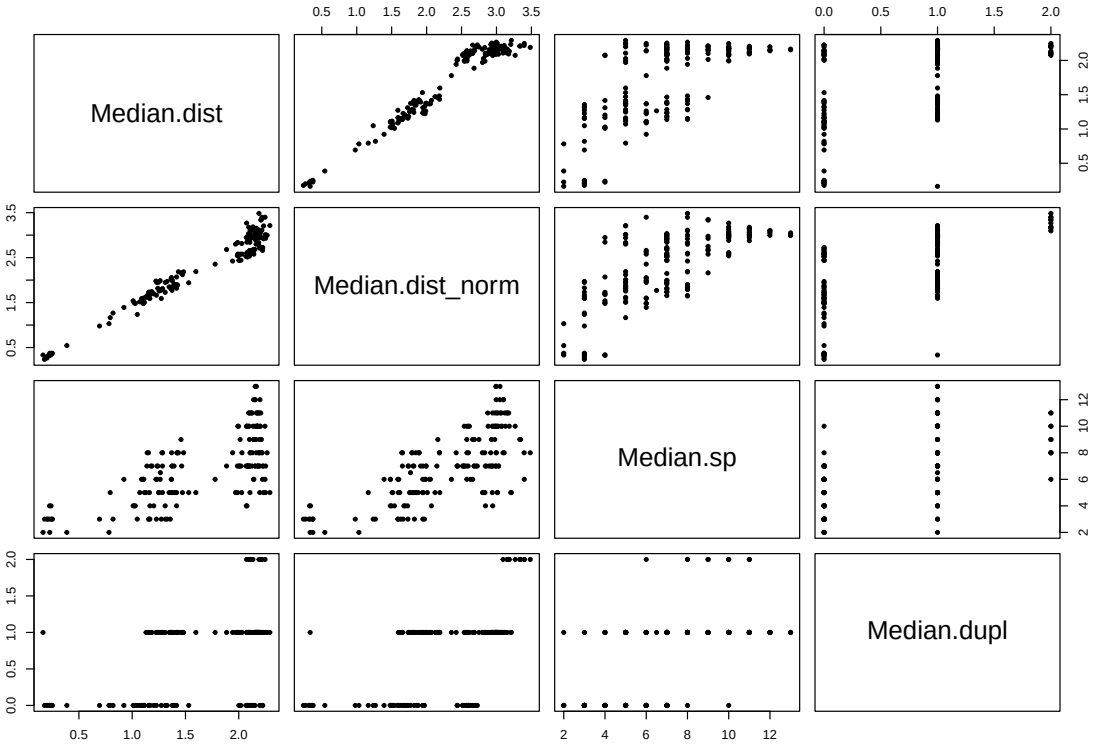
```
##   skew.dist skew.dist_norm     skew.sp skew.dupl
## 1  3.127490       8.533871  0.55403312  1.323812
## 2  3.193884       8.166580  0.02443264  1.244945
## 3  3.728122       6.441972  0.34248112  1.268274
## 4  3.126163       7.913055 -0.30378627  1.197467
## 5  3.997622       8.833735  0.31330892  1.429918
## 6  3.706171       7.658348 -0.24477196  1.087064
```

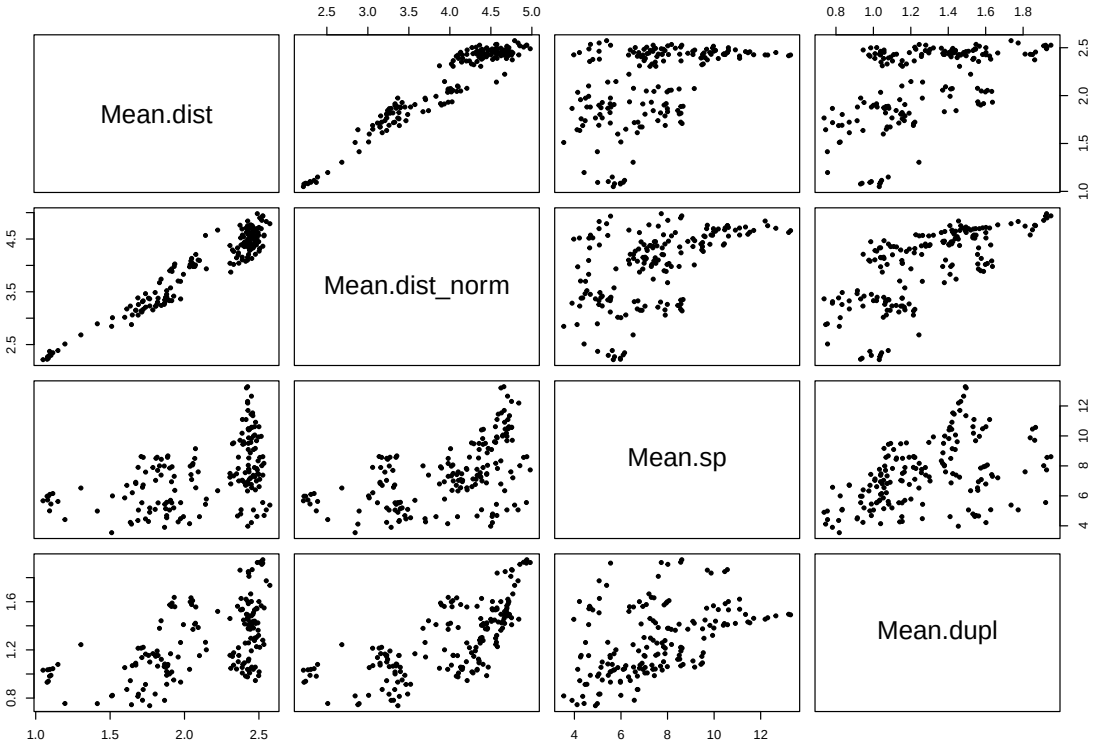
```
##    shap.p.dist shap.p.dist_norm    shap.p.sp  shap.p.dupl
## 1 3.072818e-60     7.581544e-81 1.286326e-37 4.583110e-59
## 2 4.791015e-61     1.518197e-81 1.468878e-31 6.329370e-56
## 3 2.426649e-64     4.275922e-80 1.114222e-42 3.548152e-58
## 4 6.013959e-59     1.716784e-80 2.814881e-30 1.873694e-53
## 5 2.695103e-65     1.280683e-79 4.901101e-40 2.077144e-57
## 6 6.231422e-62     7.797918e-79 1.609301e-29 2.279310e-52
```

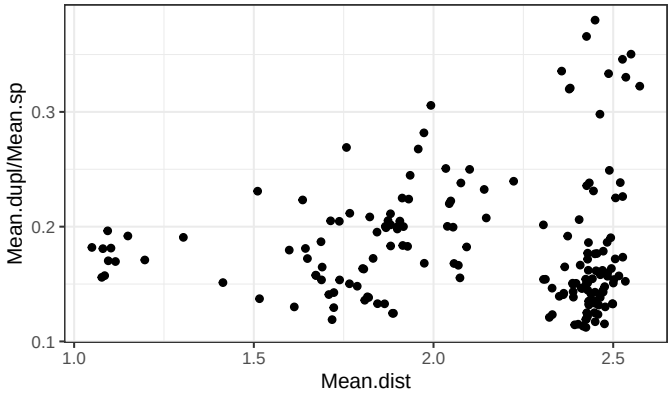
Graphical representations for summarised data

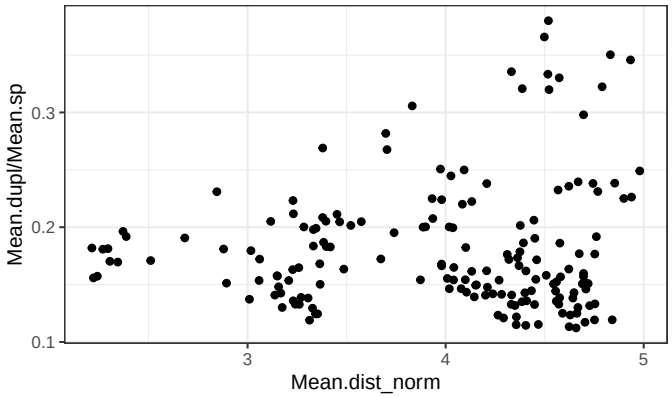


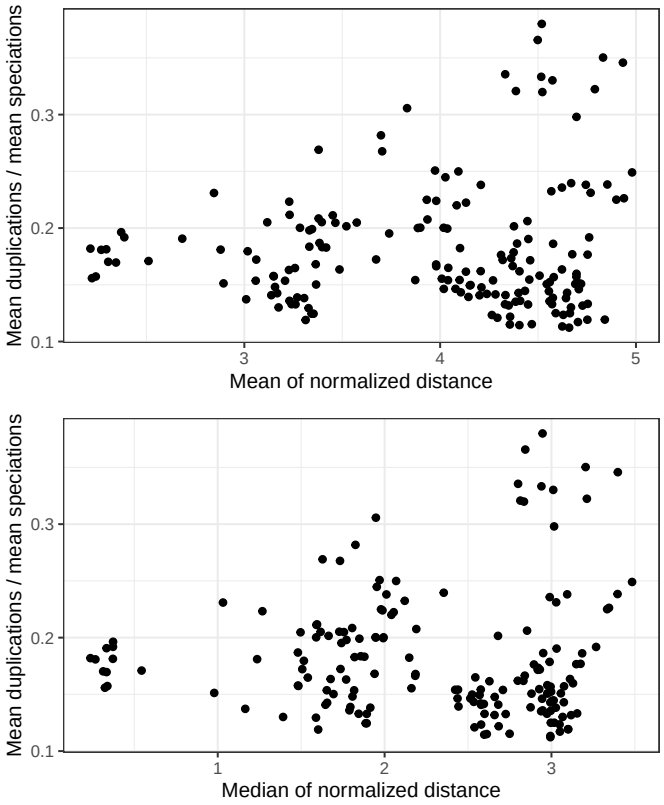
Pair plot for summarised data




Mean.dist
Mean.dist_norm
Mean.sp
Mean.dupl

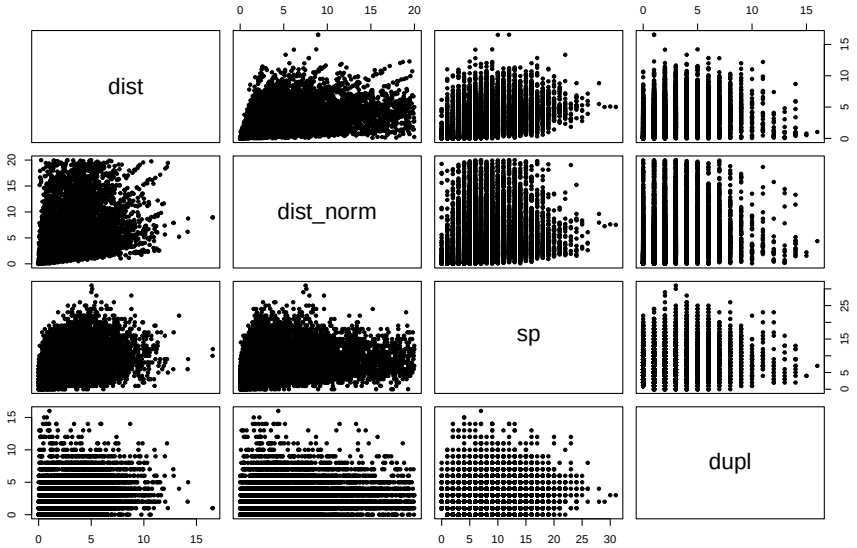

Mean.dupl/Mean.sp
Mean.dist


Mean.dupl/Mean.sp
Mean.dist_norm

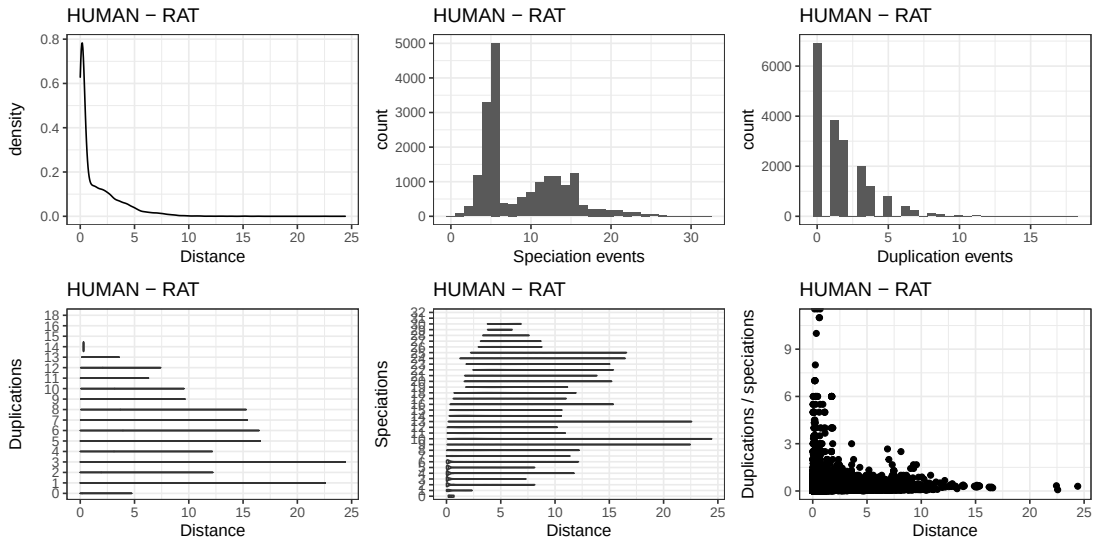


Human

Relationships between raw numerical variables

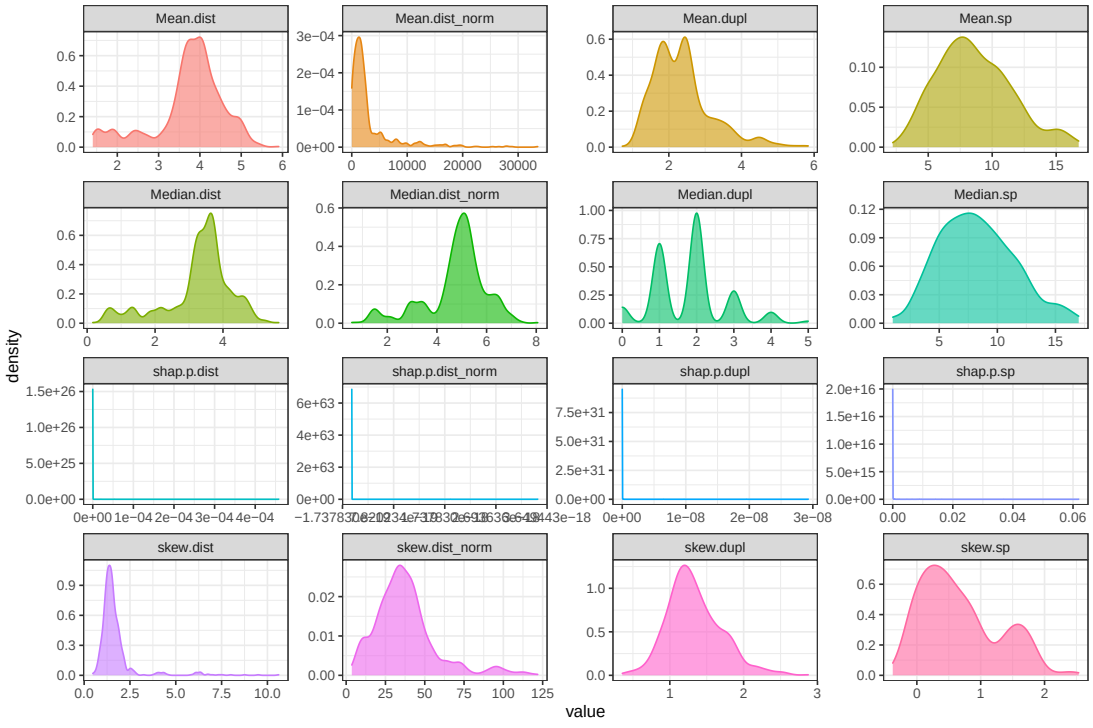


Summarizing data

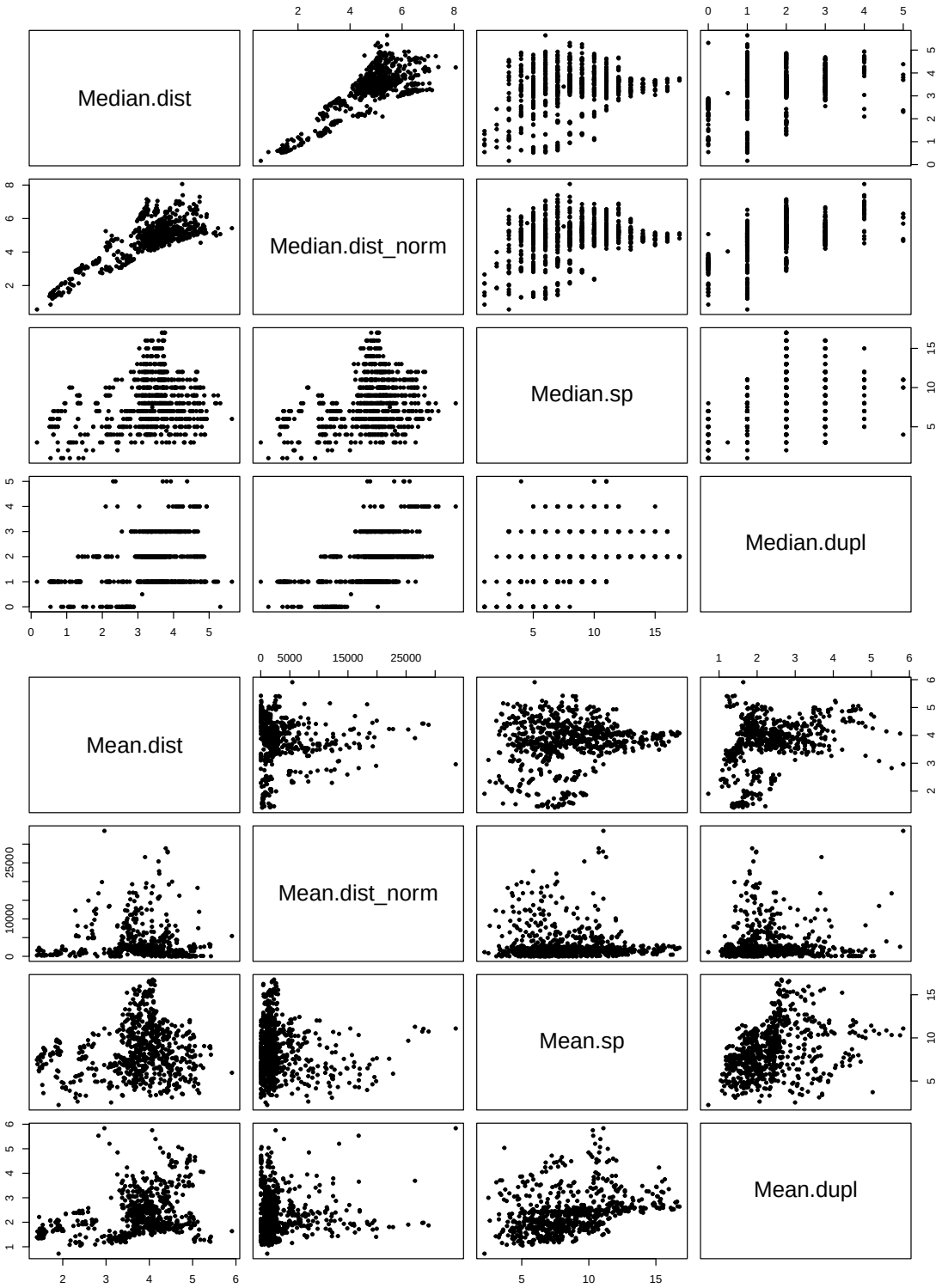


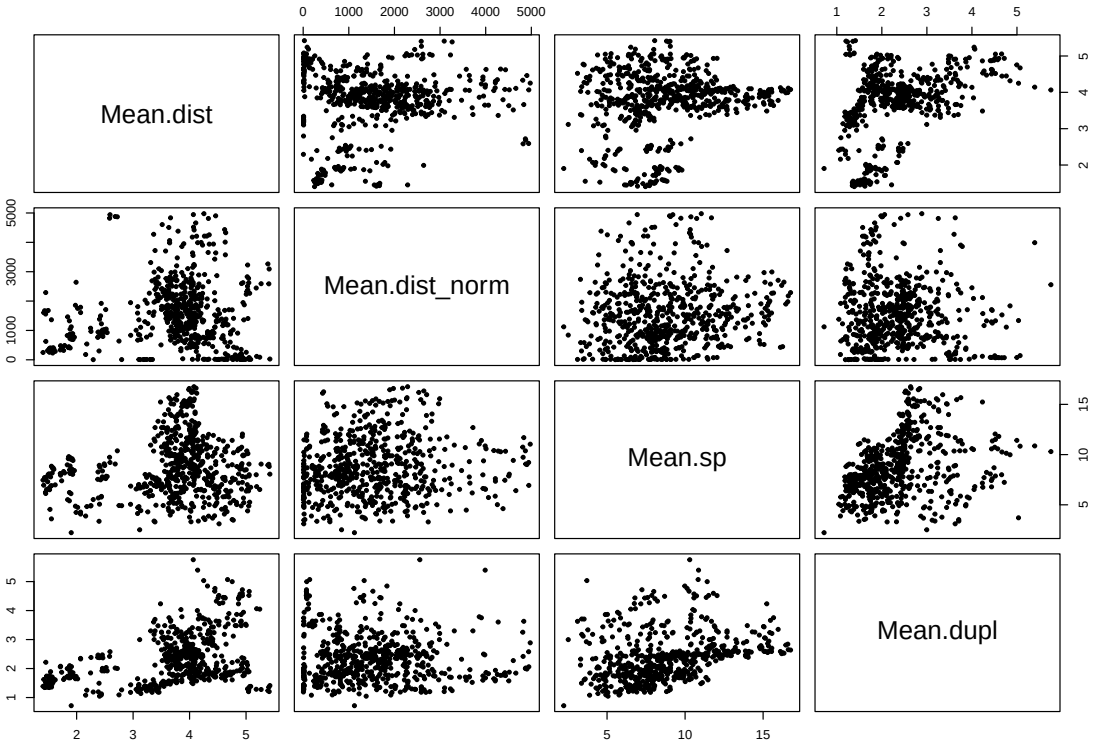
What we have for each phylome species pair

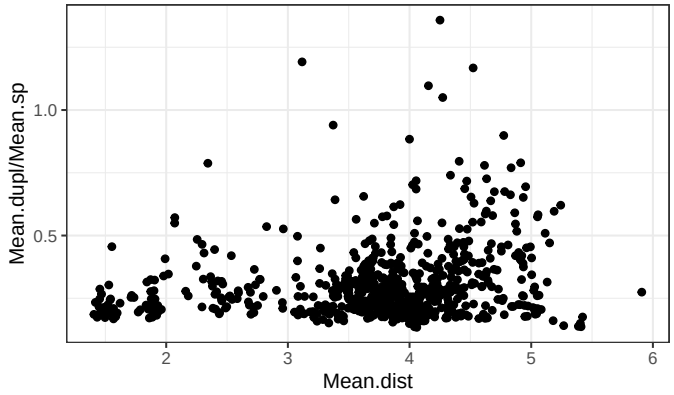
Graphical representations for summarised data

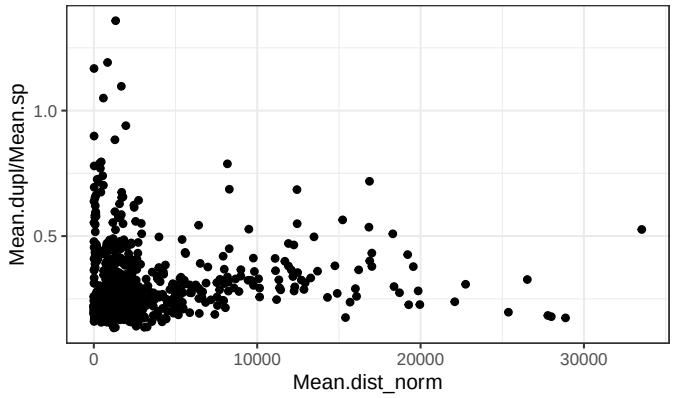


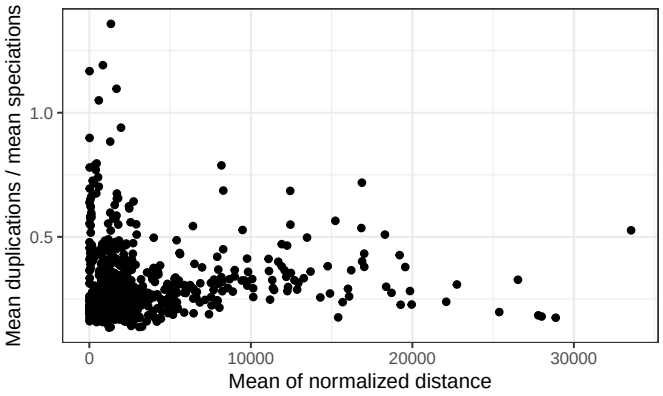
Pair plot for summarised data

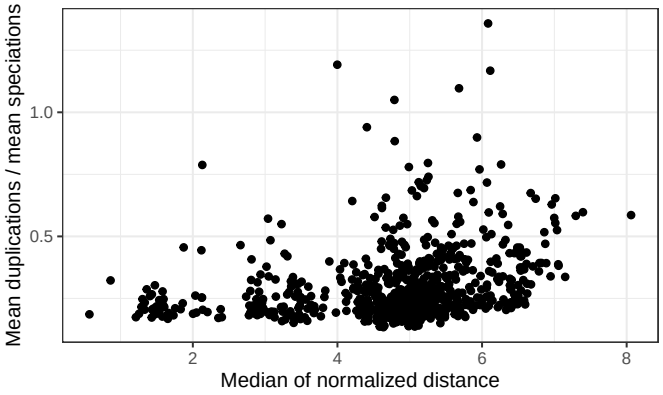



Mean.dist
Mean.dist_norm
Mean.sp
Mean.dupl


Mean.dupl/Mean.sp
Mean.dist


Mean.dupl/Mean.sp
Mean.dist_norm


Mean duplications / mean speciations
1.0
0.5
0
10000
20000
30000
Mean of normalized distance


Mean duplications / mean speciations
1.0
0.5
2
4
6
8
Median of normalized distance