Figure 4

Jordan A. Lee | leejor@ohsu.edu
Ellrott Lab
Oregon Health & Science University
Portland, Oregon
12/11/20

Incorporating feedback from previous presentation

Figure 4. Feature set analysis

Purpose: **Compare feature sets between teams -** of the best performing model. Per cancer basis

1. Exact features (variant) - **N:GEXP::ERG:2078:**
2. Grouped features (variant clusters)

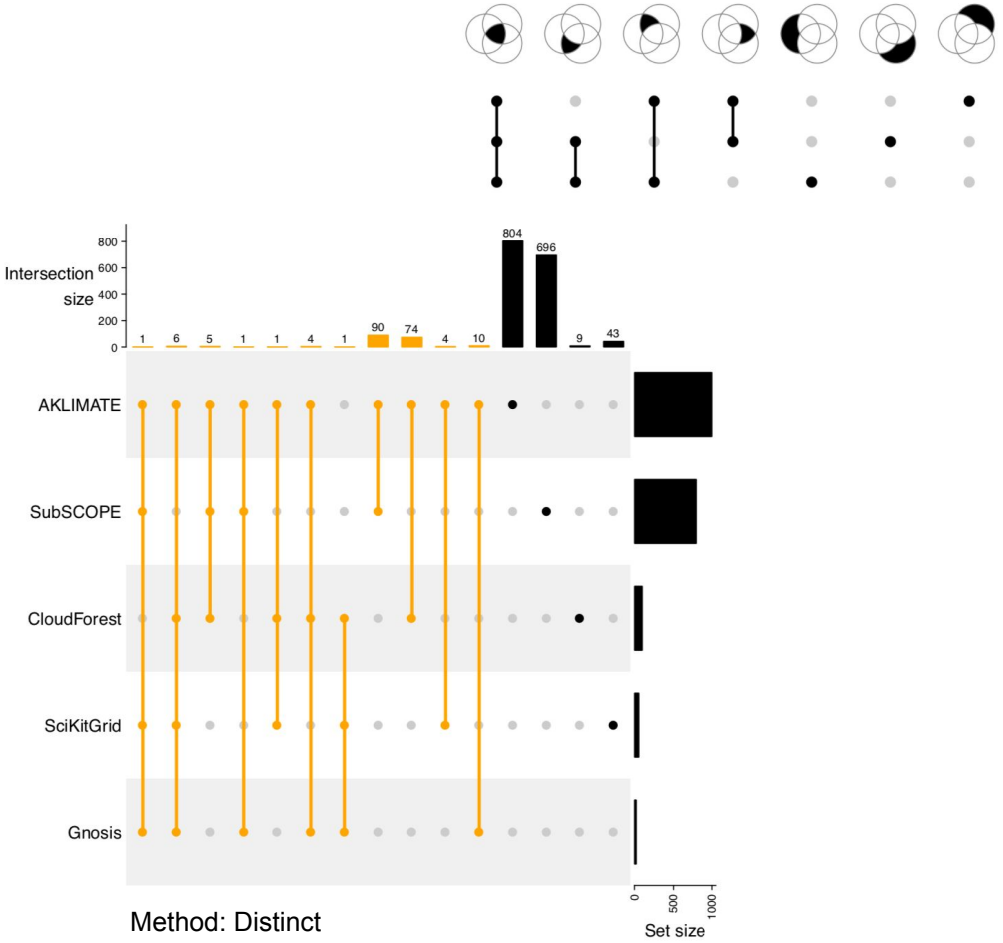
This analysis will be updated with new results once groups upload files post miRNA fix

Result slides for BRCA. If there is interest, can run for other cancer types.

Recall: Exact feature overlap

Small overlap

- Some overlap between 90 AKLIMATE:SubSCOPE and 74 AKLIMATE:CloudForest
- These are the 3 teams with the largest feature set size

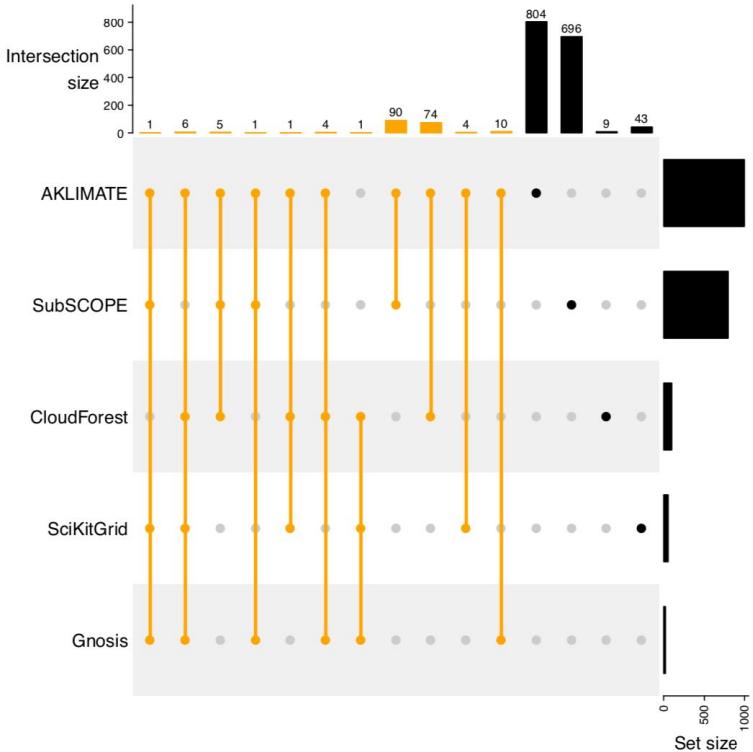


Method: Distinct

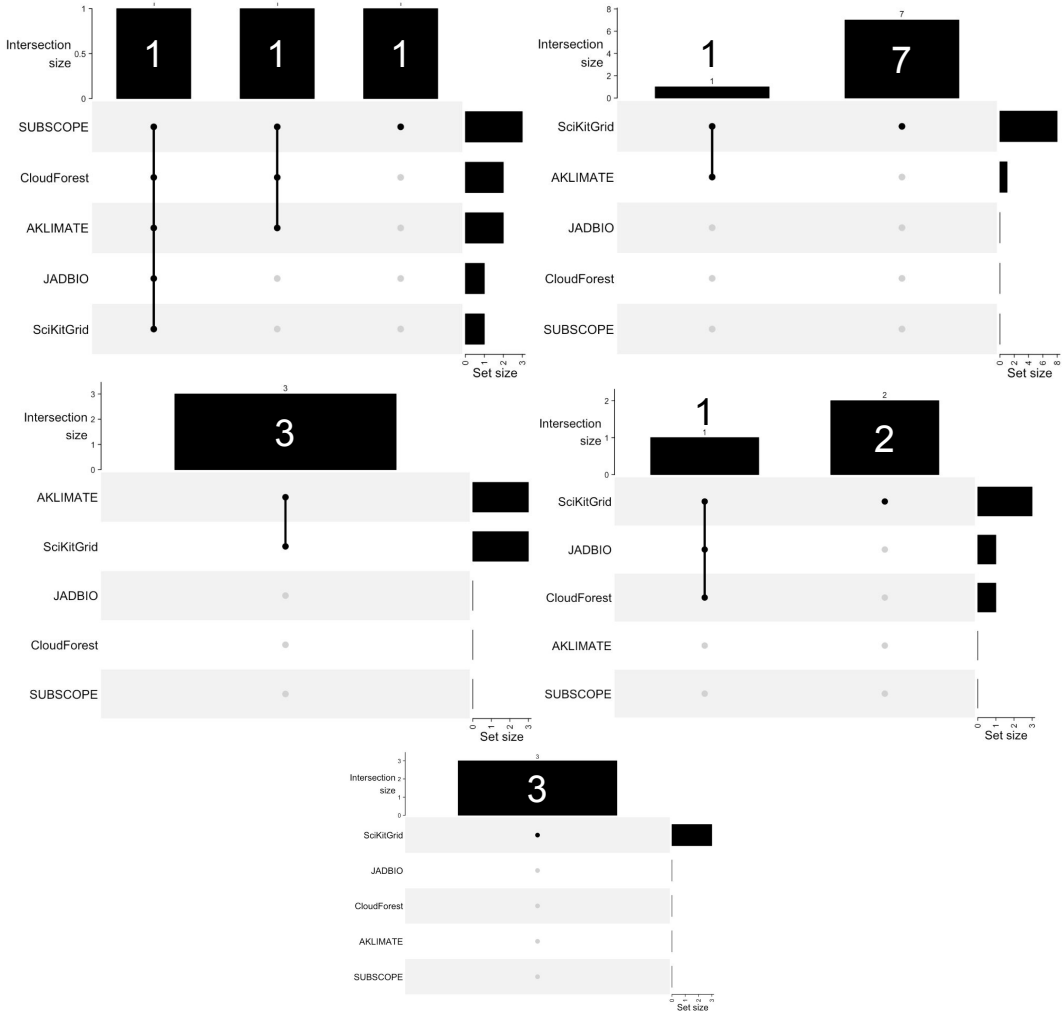
...Motivates cluster overlap

Figure 4

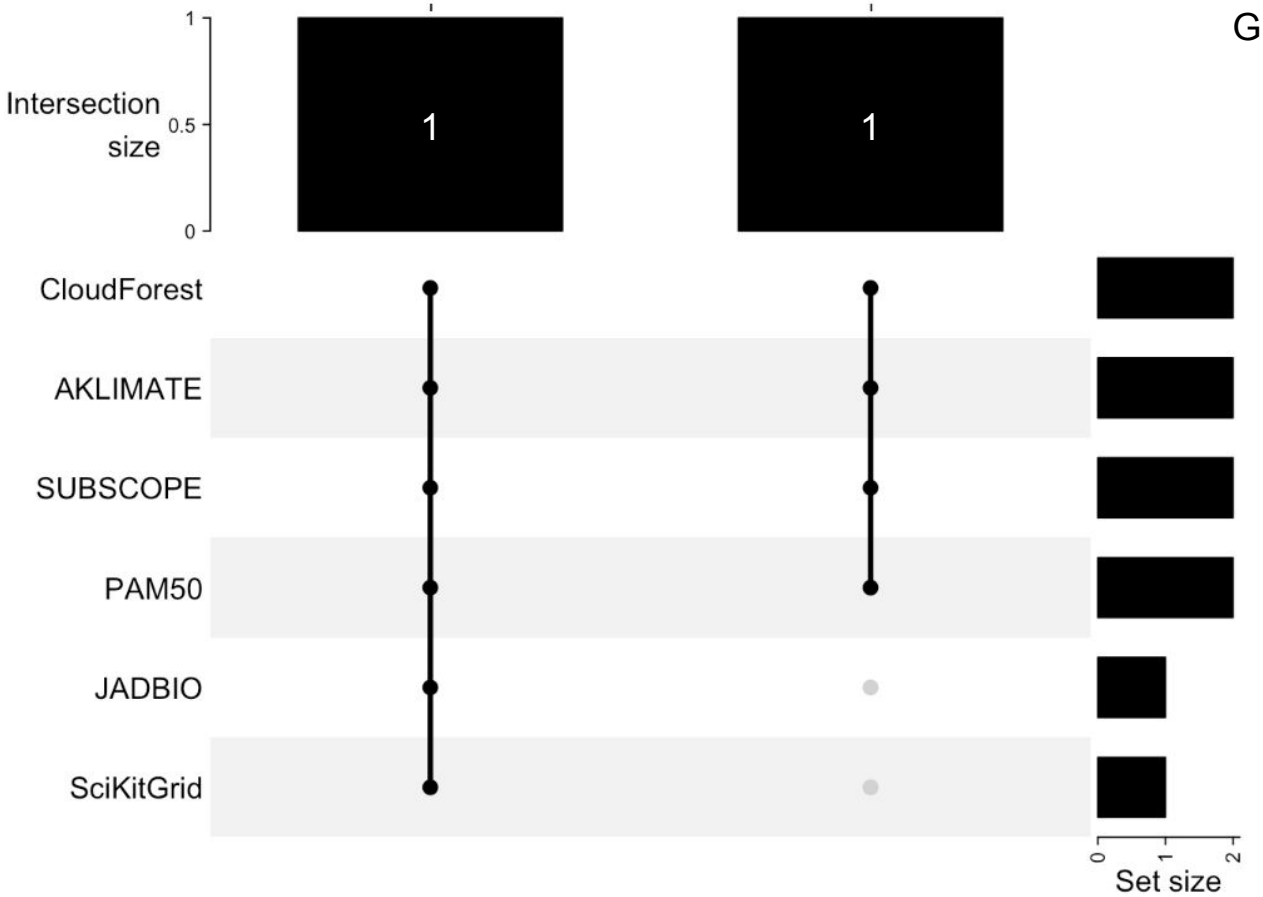
A. Exact ft. All datatypes pooled



B. Exact ft. Cluster. Datatypes GEXP (n=1682), CNVR (n=18), METH (n=30), MIR (n=14), MUTA (n=7)



Or with PAM50
GEXP at optimal number of clusters



How did we cluster?

Data used?

TMP_v9_20201029.tar.gz*

- feature_list_with_performance_with_subtype_names_20200828.tsv.gz (syn22337110)
- collected_features_matrix_20200722.tsv.gz (syn22271992)

Best model per team - mean overall weighted F1 score

- gnosis_1_BRCA
- CF|All_Top 100_BRCA
- AKLIMATE_BRCA_reduced_model_1000_feature_set_BRCA
- nn_jg_2020-03-20_top1kfreq:BRCA_BRCA
- fbedeBIC_BRCA

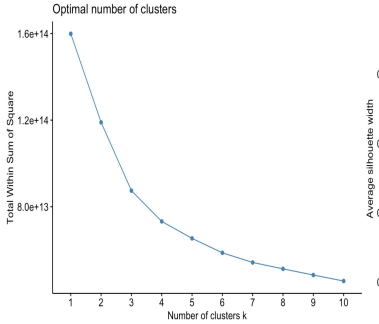
*TMP_v8_20200203.tar.gz if classifier miRNA not used as input

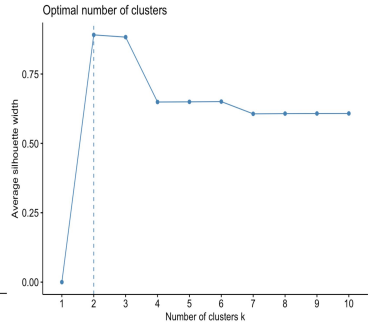
1. Best model per team 2. Pool feature across all teams 3. Cluster features based on raw tarball matrix values

A. Optimize clustering structure

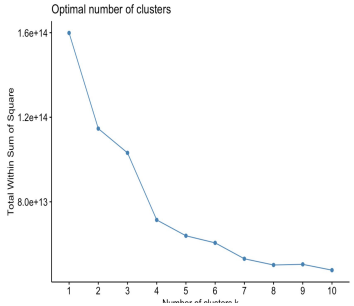
```
     ward    single  complete   average
0.9893030 0.9860369 0.9852940 0.9845092
```

B. Optimize number of clusters





hcut - WSS and silhouette



K-means - WSS

C. Cluster - AGNES and explore cluster membership

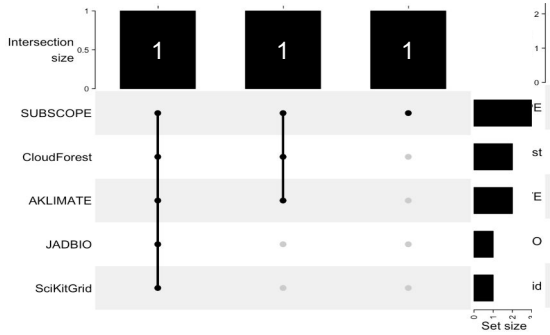
clusterID	n_members
1	1655
2	27

```
clusterID                    fts
        2     N:GEXP::CPB1:1360:
        2      N:GEXP::FTL:2512:
        2     N:GEXP::RPL8:6132:
        2     N:GEXP::XBP1:7494:
        2 N:GEXP::SLC39A6:25800:
        2  N:GEXP::SCGB2A2:4250:
```

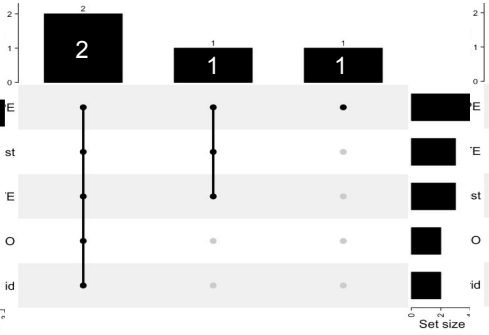
....but perhaps we don't want minimal n_clusters

Explore a range of n_clusters outside optimal

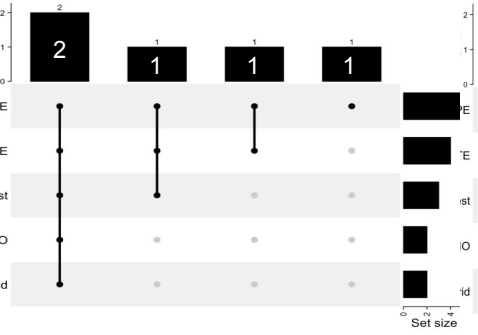
3 clusters



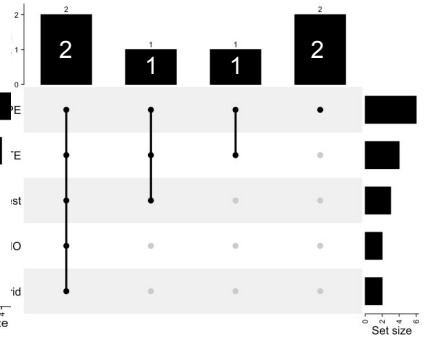
4 clusters



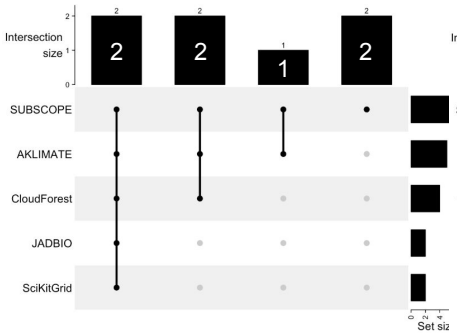
5 clusters



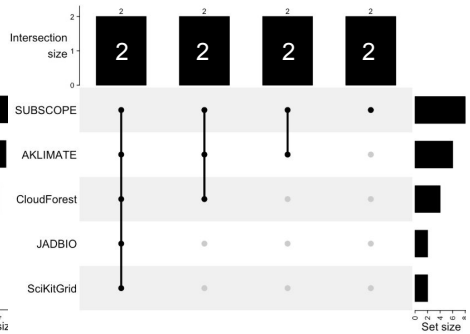
6 clusters



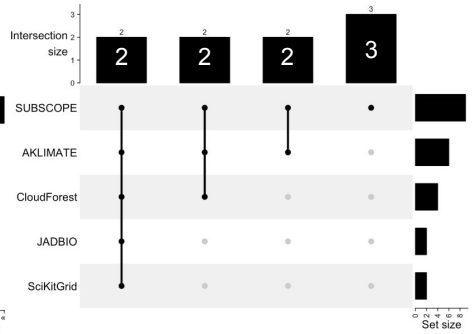
7 clusters



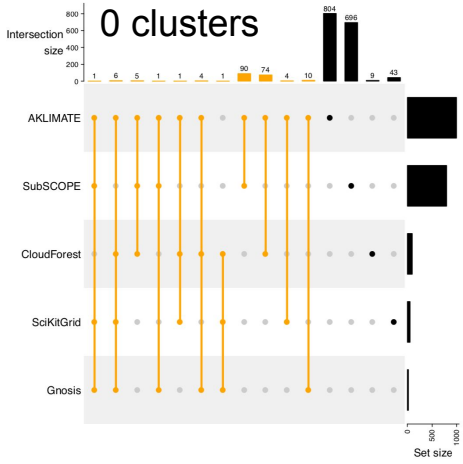
8 clusters



9 clusters

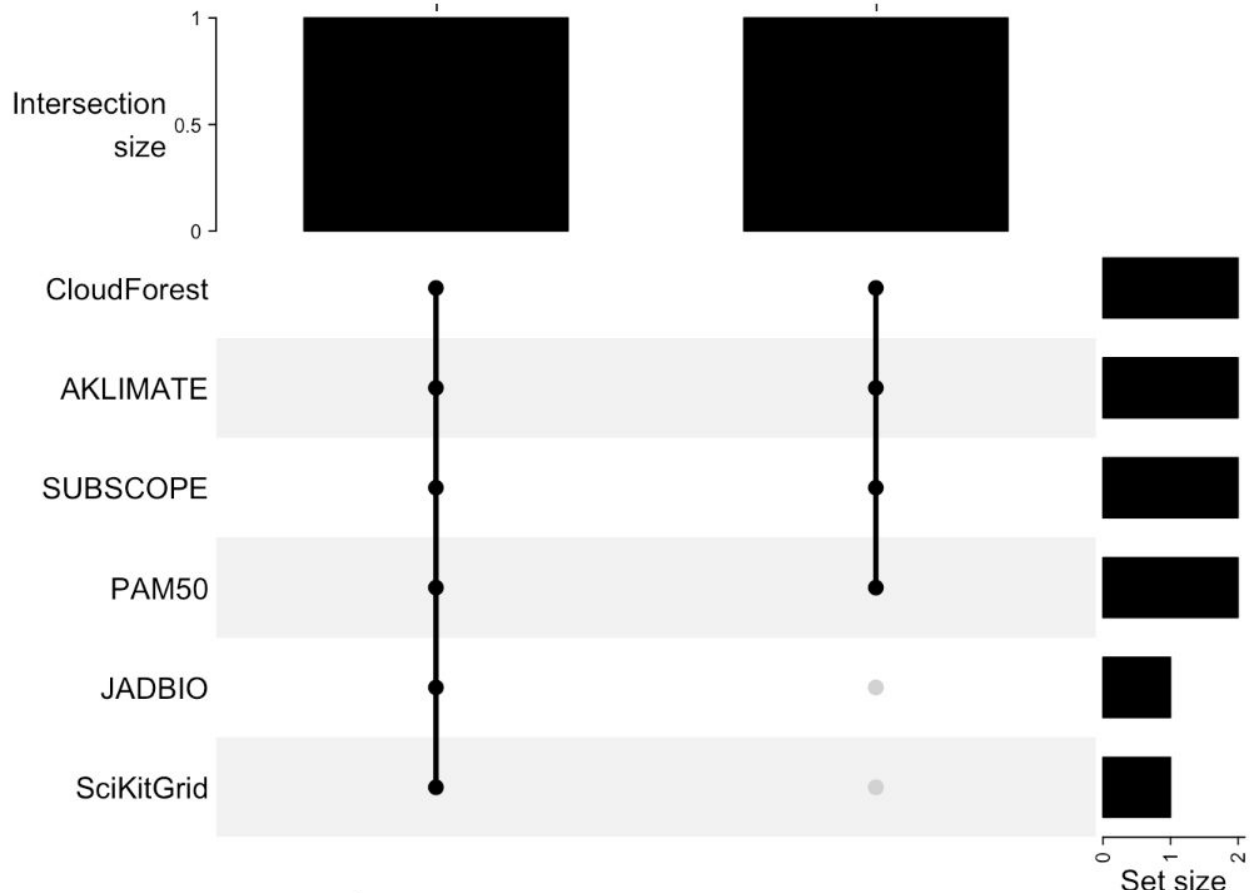


0 clusters



PAM50

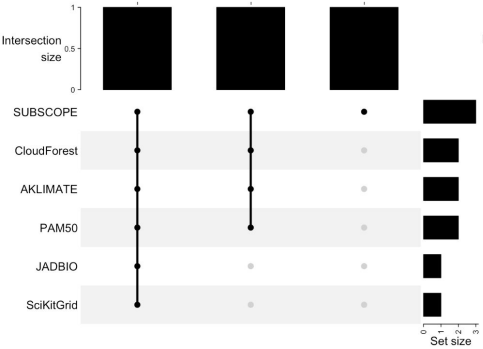
GEXP at optimal number of clusters

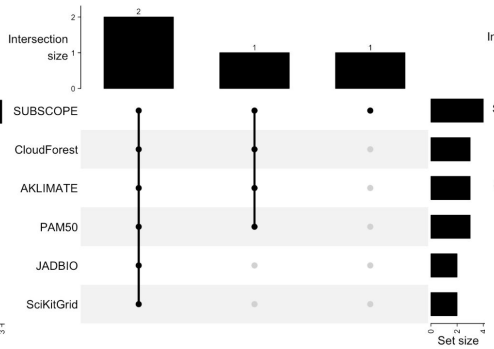


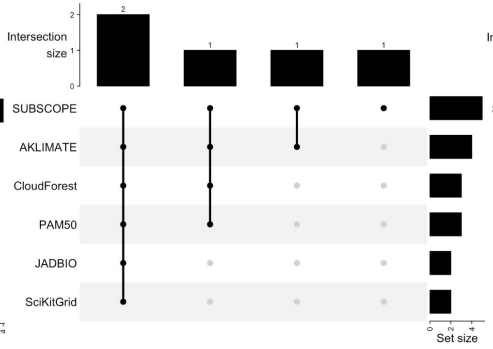
	gnosis_1_BRCA	CF|All_Top 100_BRCA	AKLIMATE_BRCA_reduced_model_1000_feature_set_BRCA	nn_jg_2020-03-20_top1kfreq:BRCA_BRCA	fbedeBIC_BRCA	PAM50
clust1	1	1	1	1	1	1
clust2	0	1	1	1	0	1

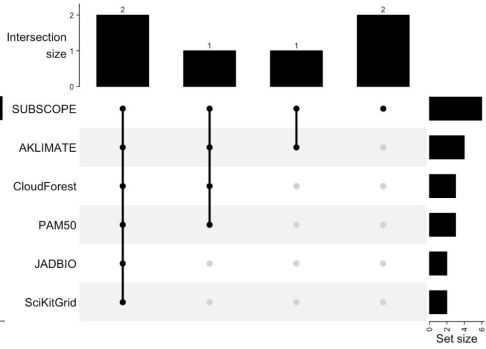
Explore a range of n_clusters outside optimal

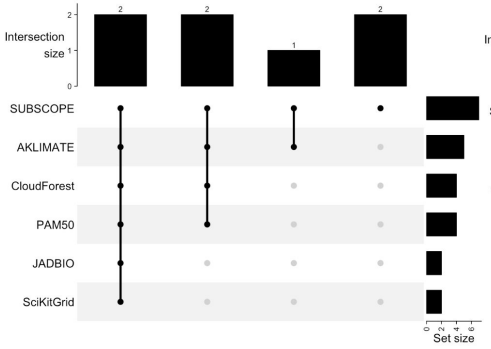
....but perhaps we don’t want minimal n_clusters

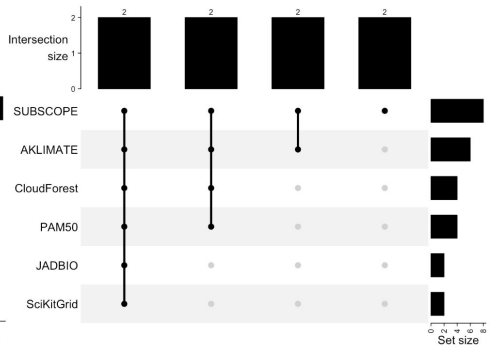


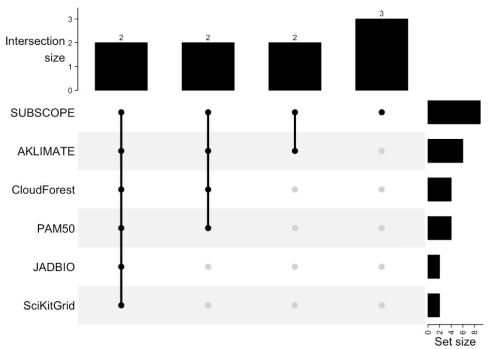






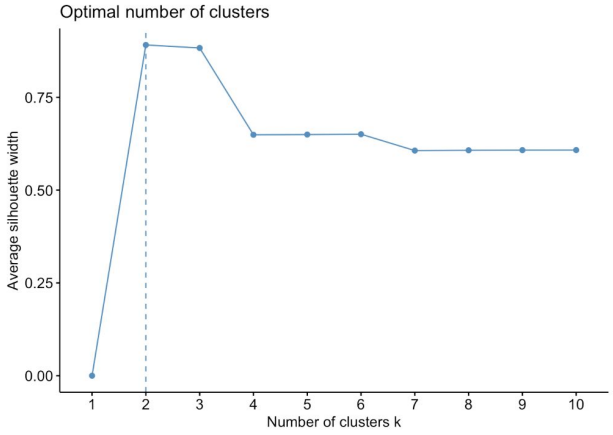


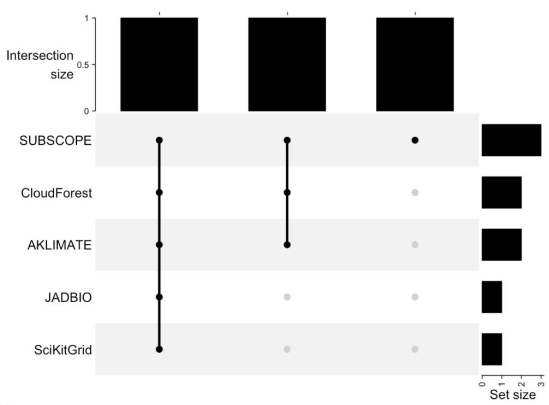




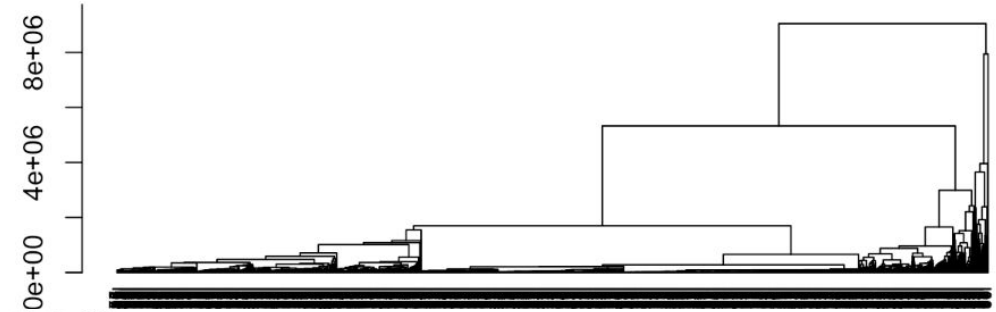
Additional Slides

GEXP n_cluster = optimal +1



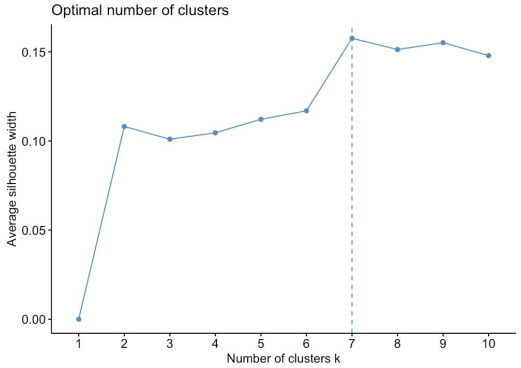


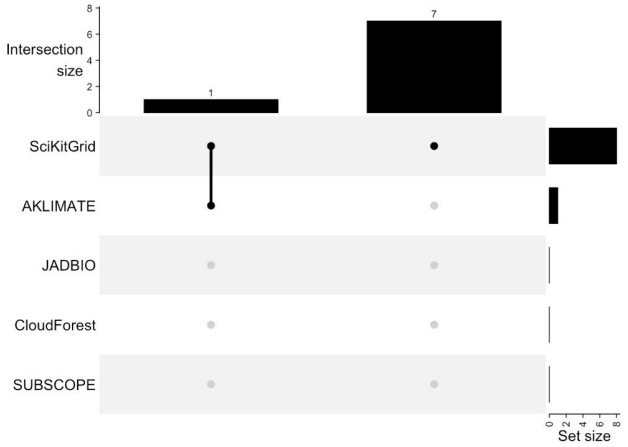
	gnosis_1_BRCA	CF|All_Top 100_BRCA	AKLIMATE_BRCA_reduced_model_1000_feature_set_BRCA	nn_jg_2020-03-20_top1kfreq:BRCA_BRCA	fbedeBIC_BRCA
clust1	1	1	1	1	1
clust2	0	1	1	1	0
clust3	0	0	0	1	0

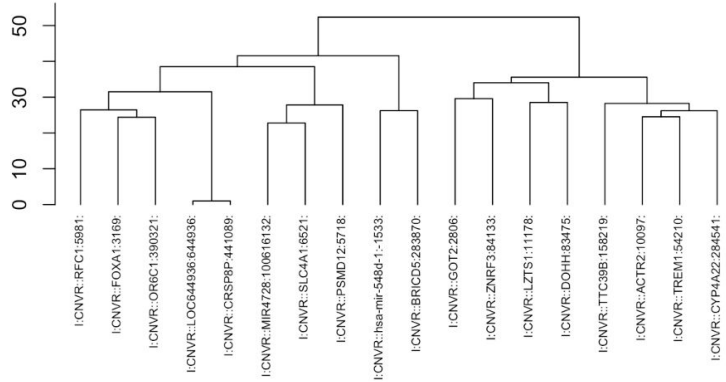


clusterID	n_members
1	1655
2	26
3	1

CNVR n_cluster = optimal +1



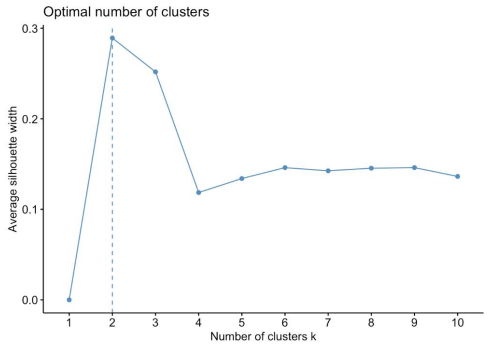


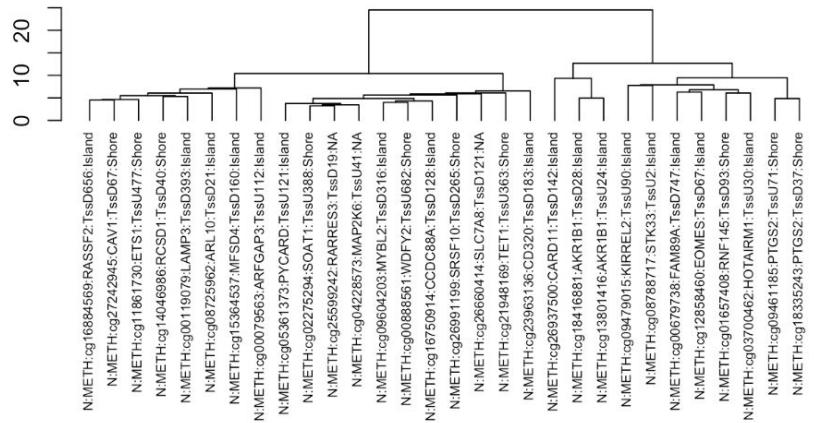


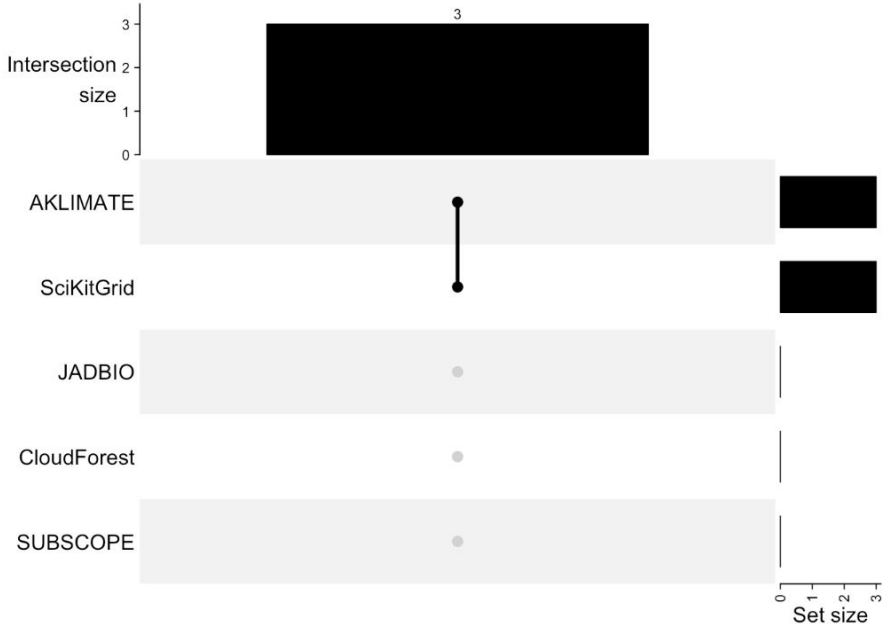
	gnosis_1_BRCA	CF|All_Top 100_BRCA	AKLIMATE_BRCA_reduced_model_1000_feature_set_BRCA	nn_jg_2020-03-20_top1kfreq:BRCA_BRCA	fbedeBIC_BRCA
clust1	0	0	0	0	1
clust2	0	0	0	0	1
clust3	0	0	1	0	1
clust4	0	0	0	0	1
clust5	0	0	0	0	1
clust6	0	0	0	0	1
clust7	0	0	0	0	1
clust8	0	0	0	0	1

clusterID	n_members
1	3
2	1
3	2
4	4
5	3
6	2
7	2
8	1

METH n_cluster = optimal +1



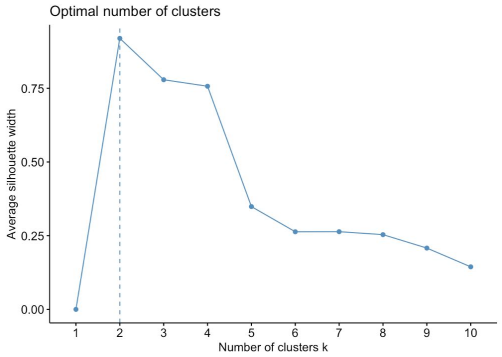


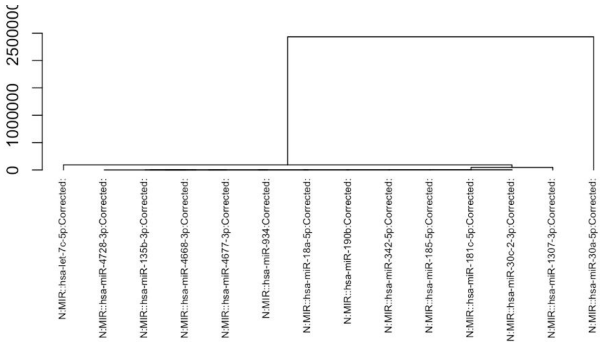


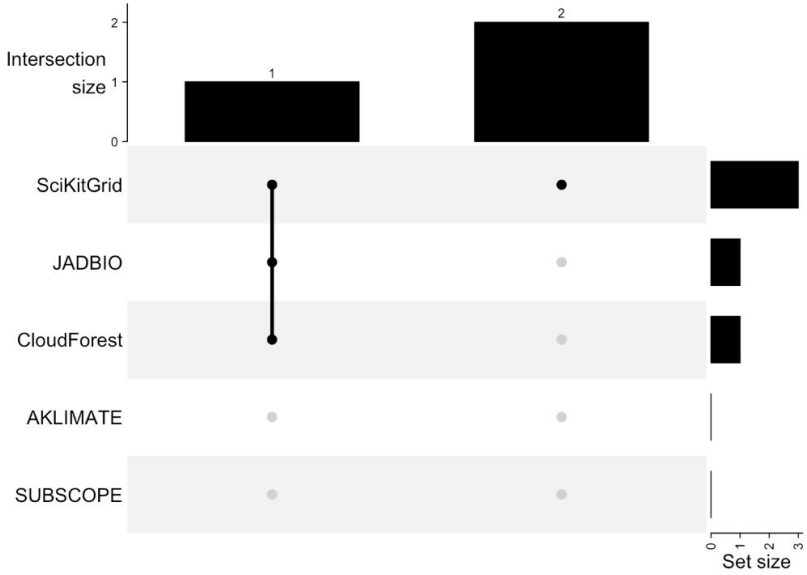
	gnosis_1_BRCA	CF|All_Top 100_BRCA	AKLIMATE_BRCA_reduced_model_1000_feature_set_BRCA	nn_jg_2020-03-20_top1kfreq:BRCA_BRCA	fbedeBIC_BRCA
clust1	0	0	1	0	1
clust2	0	0	1	0	1
clust3	0	0	1	0	1

clusterID	n_members
1	19
2	3
3	8

MIR n_cluster = optimal +1



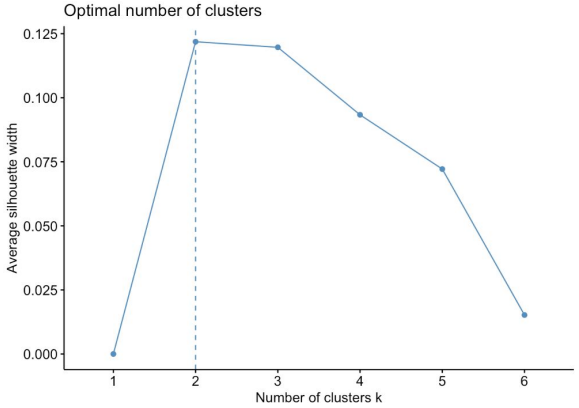


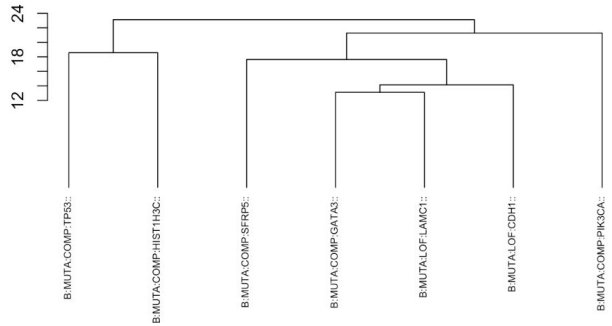


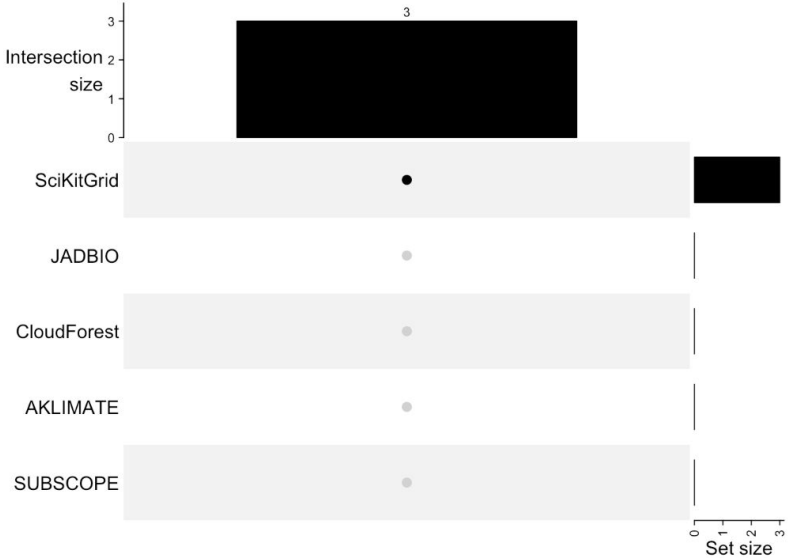
	gnosis_1_BRCA	CF|All_Top 100_BRCA	AKLIMATE_BRCA_reduced_model_1000_feature_set_BRCA	nn_jg_2020-03-20_top1kfreq:BRCA_BRCA	fbedeBIC_BRCA
clust1	0	0	0	0	1
clust2	1	1	0	0	1
clust3	0	0	0	0	1

clusterID	n_members
1	1
2	12
3	1

MUTA n_cluster = optimal +1



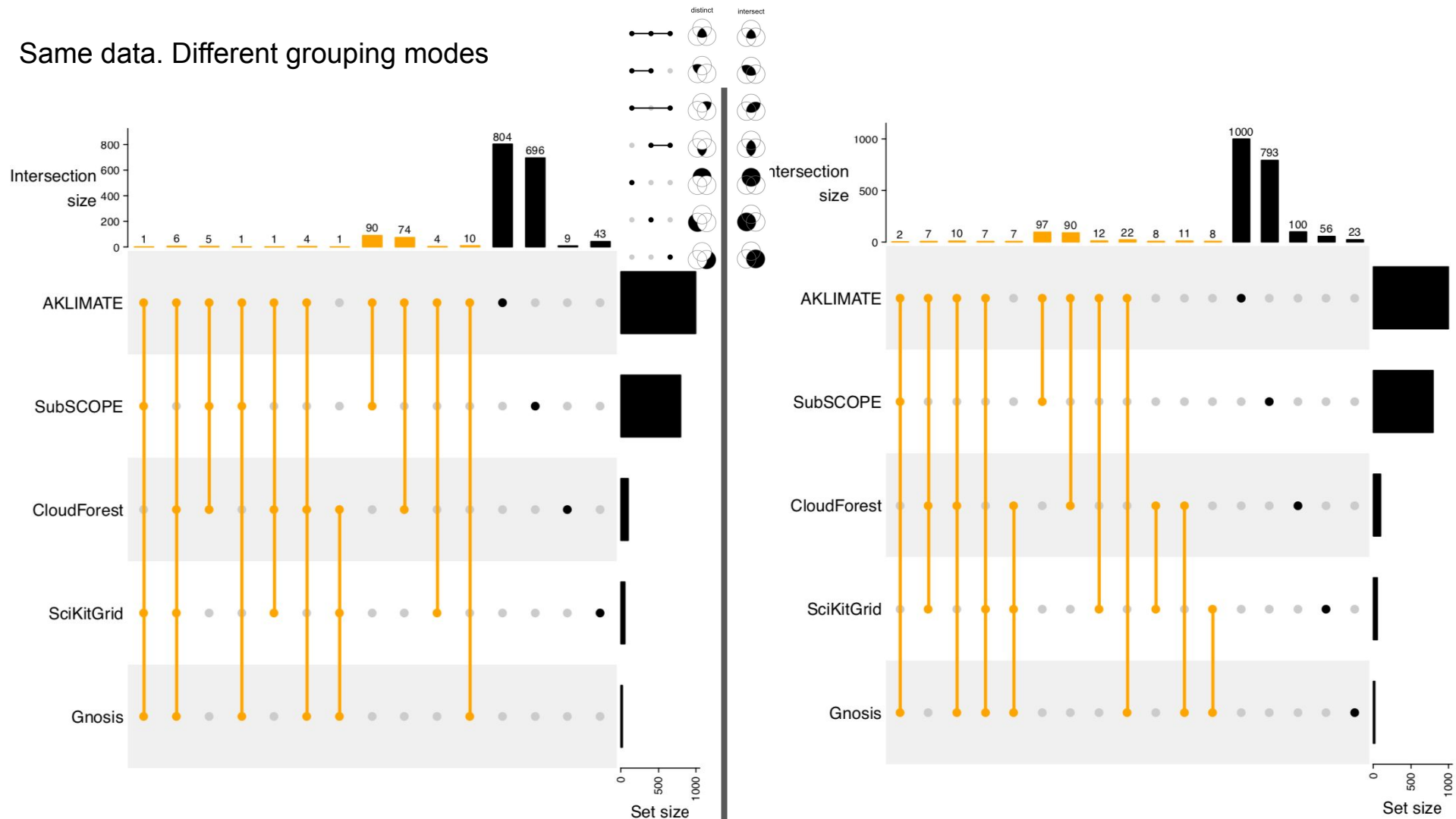


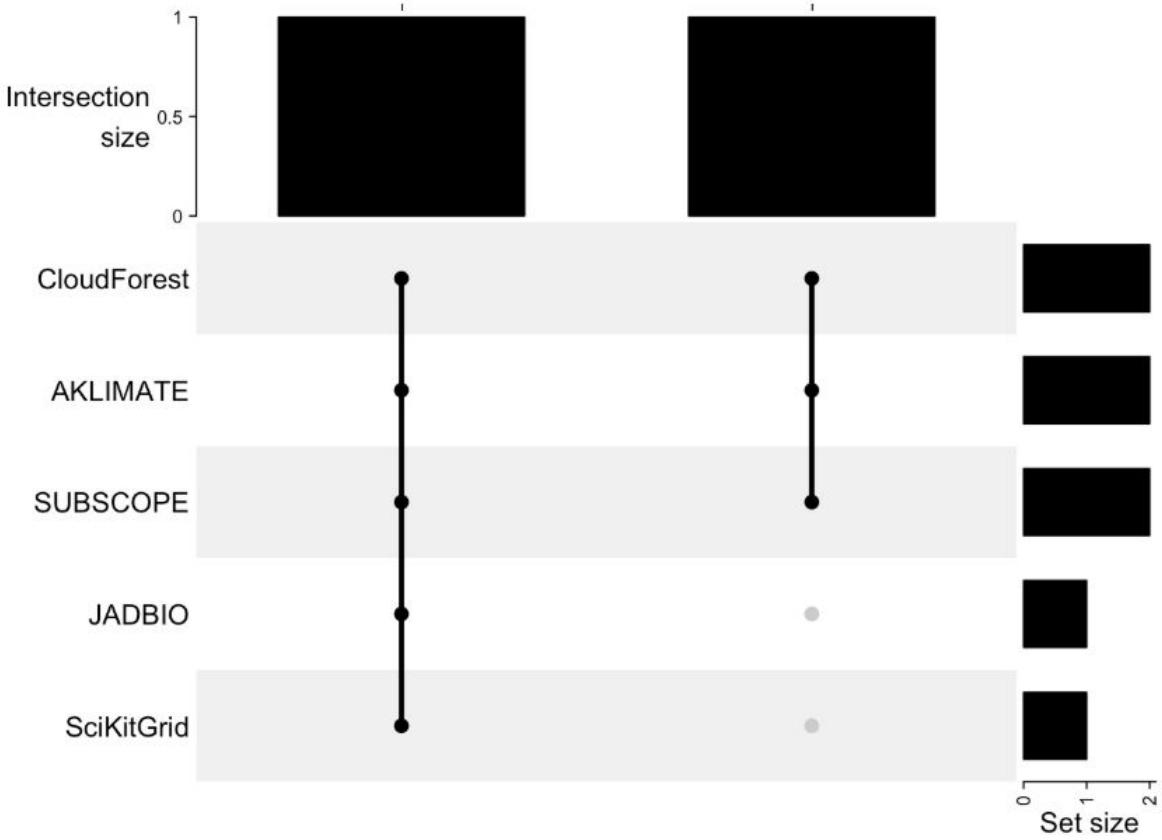


	gnosis_1_BRCA	CF|All_Top 100_BRCA	AKLIMATE_BRCA_reduced_model_1000_feature_set_BRCA	nn_jg_2020-03-20_top1kfreq:BRCA_BRCA	fbedeBIC_BRCA
clust1	0	0	0	0	1
clust2	0	0	0	0	1
clust3	0	0	0	0	1

clusterID	n_members
1	2
2	4
3	1

Same data. Different grouping modes





GEXP
Optimal number of clusters = 2

	gnosis_1_BRCA	CF|All_Top 100_BRCA	AKLIMATE_BRCA_reduced_model_1000_feature_set_BRCA	nn_jg_2020-03-20_top1kfreq:BRCA_BRCA	fbedeBIC_BRCA
clust1	1	1	1	1	1
clust2	0	1	1	1	0