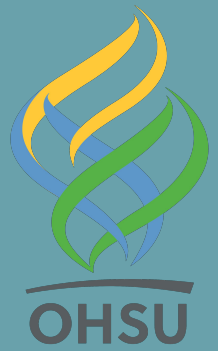
Figure 4. Feature Set Analysis

Jordan A. Lee
Oregon Health & Science University
Portland, Oregon
12/18/20



Developed in collaboration with Chris, Christina, Brian

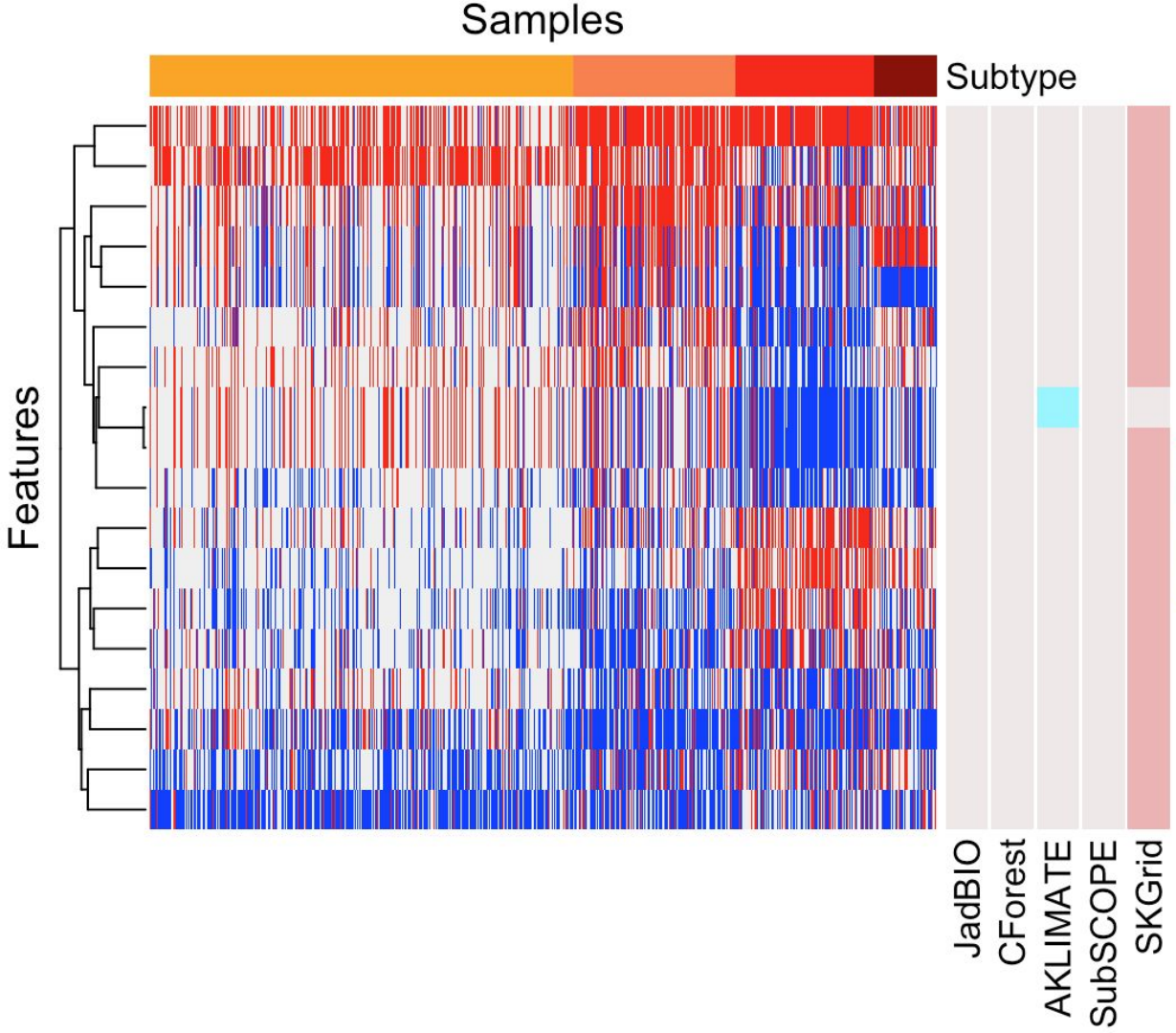
Data Analysis

Jordan, Chris, Christina, Brian

- Based on V8 tarball, these results are before the miRNA correction to certain cancers
- If interest, will rerun once all v9 results from classifier groups in

- Process:
  - Select a cancer cohort
  - Select best model per team (overall weighted F1)
  - Pull corresponding feature set for each model
  - Map back to molecular values (in v8 tarball)
  - Scale data if appropriate (z-scores for METH, GEXP, MIR)
  - Heatmaps
    - Plot scaled/raw data
    - Cluster feature rows
    - No clustering on sample/subtype columns

All heatmaps clustered using complete linkage method (max dist between clusters before merging) and euclidean distance

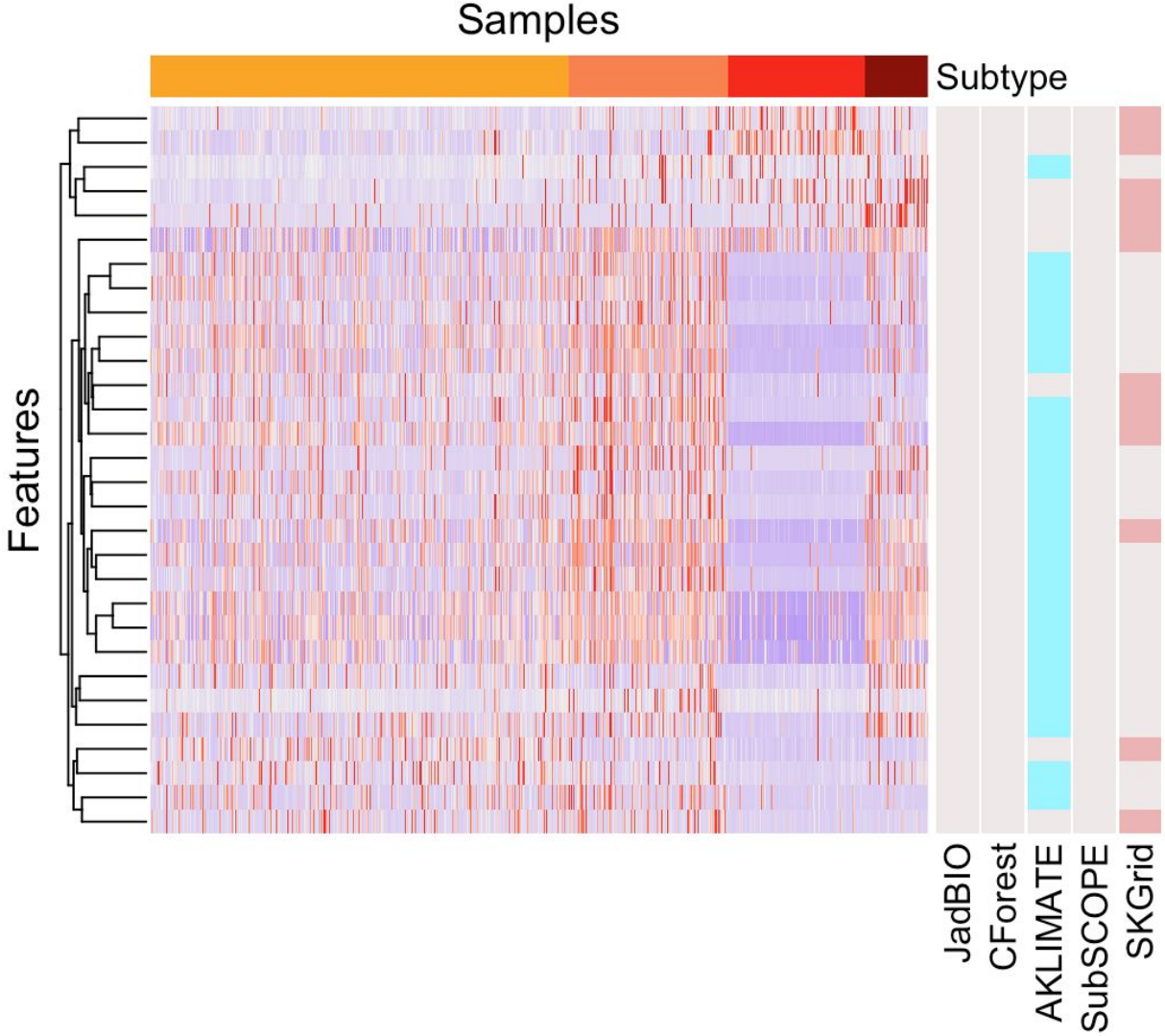


The one AKLIMATE feature is different than any picked from SKGrid

But appears to be very similar to a SKGrid feature

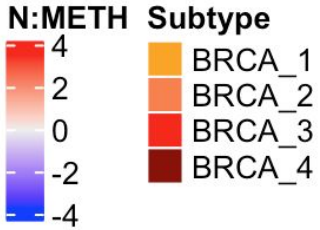
Ordered feature rows

```
[1] 18
 [1] "I:CNVR::hsa-mir-548d-1:-1533:" "I:CNVR::BRICD5:283870:"
 [3] "I:CNVR::PSMD12:5718:"          "I:CNVR::MIR4728:100616132:"
 [5] "I:CNVR::SLC4A1:6521:"          "I:CNVR::FOXA1:3169:"
 [7] "I:CNVR::OR6C1:390321:"         "I:CNVR::LOC644936:644936:"
 [9] "I:CNVR::CRSP8P:441089:"        "I:CNVR::RFC1:5981:"
[11] "I:CNVR::TREM1:54210:"          "I:CNVR::ACTR2:10097:"
[13] "I:CNVR::CYP4A22:284541:"       "I:CNVR::TTC39B:158219:"
[15] "I:CNVR::DOHH:83475:"           "I:CNVR::LZTS1:11178:"
[17] "I:CNVR::ZNRF3:84133:"          "I:CNVR::GOT2:2806:"
```



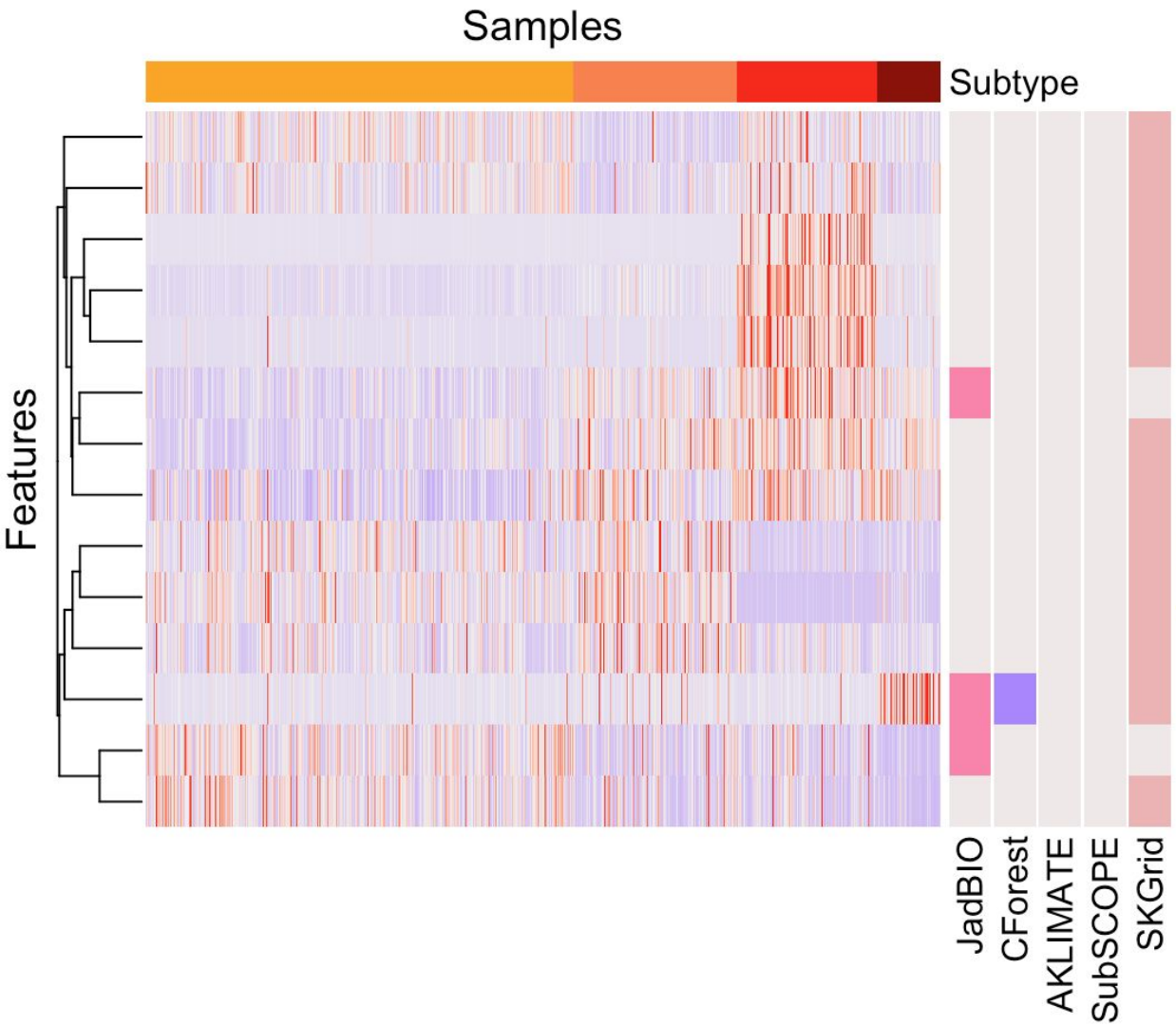
Teams don’t tend to group in separate clusters

The two groups often pick different yet similar features (cluster membership)



Ordered feature rows

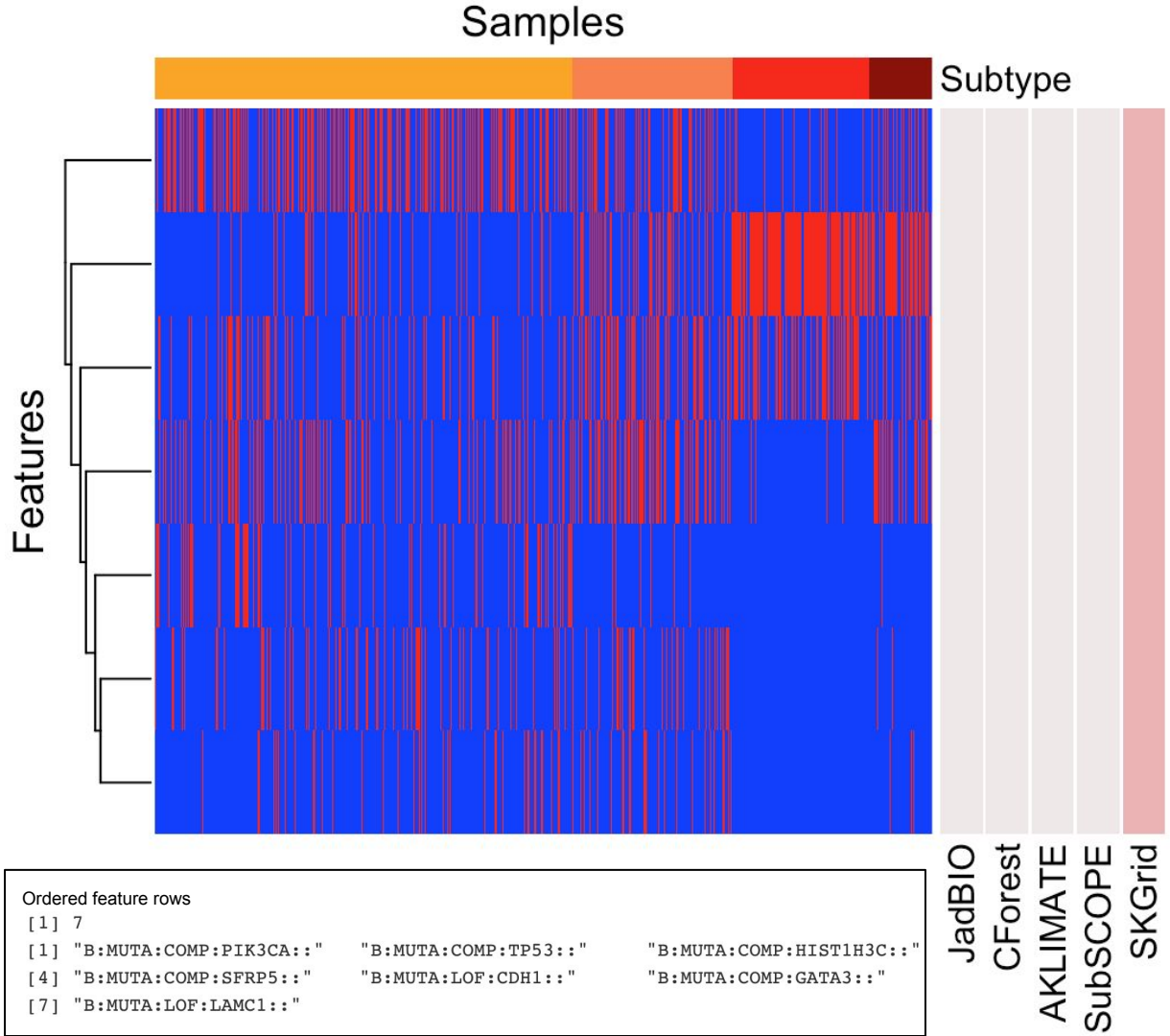
```
[1] 30
 [1] "N:METH:cg25599242:RARRES3:TssD19:NA"
 [2] "N:METH:cg02275294:SOAT1:TssU388:Shore"
 [3] "N:METH:cg05361373:PYCARD:TssU121:Island"
 [4] "N:METH:cg26660414:SLC7A8:TssD121:NA"
 [5] "N:METH:cg23963136:CD320:TssD183:Island"
 [6] "N:METH:cg26937500:CARD11:TssD142:Island"
 [7] "N:METH:cg18335243:PTGS2:TssD37:Shore"
 [8] "N:METH:cg09461185:PTGS2:TssU71:Shore"
                    ...
```



Only one feature overlap (teams select the feature with strong subtype 4 banding)

Ordered feature rows

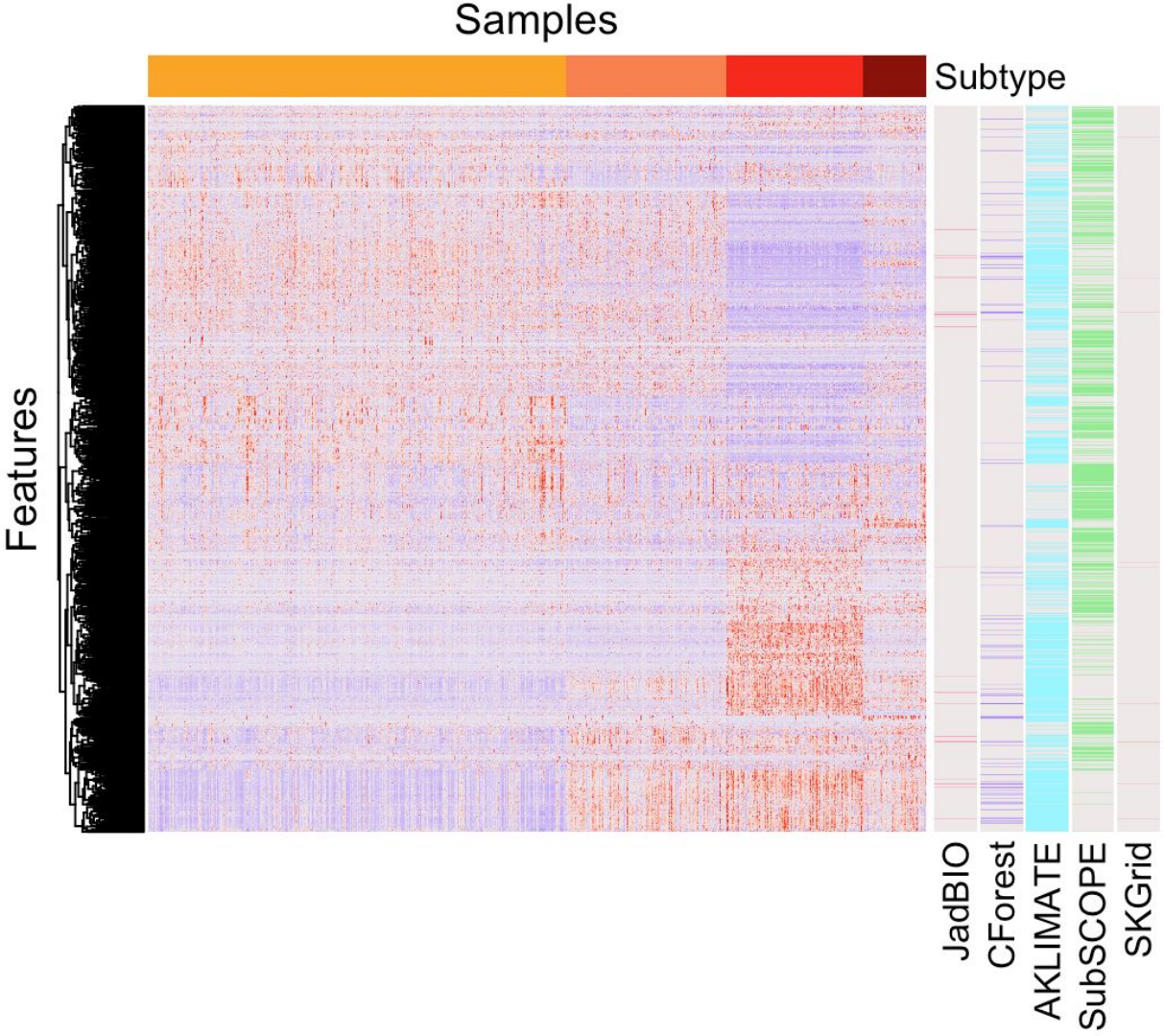
```
[1] 14
 [1] "N:MIR::hsa-let-7c-5p:Corrected:"     "N:MIR::hsa-miR-181c-5p:Corrected:"
 [3] "N:MIR::hsa-miR-934:Corrected:"       "N:MIR::hsa-miR-18a-5p:Corrected:"
 [5] "N:MIR::hsa-miR-135b-3p:Corrected:"   "N:MIR::hsa-miR-4677-3p:Corrected:"
 [7] "N:MIR::hsa-miR-1307-3p:Corrected:"   "N:MIR::hsa-miR-185-5p:Corrected:"
 [9] "N:MIR::hsa-miR-342-5p:Corrected:"    "N:MIR::hsa-miR-190b:Corrected:"
[11] "N:MIR::hsa-miR-4668-3p:Corrected:"   "N:MIR::hsa-miR-4728-3p:Corrected:"
[13] "N:MIR::hsa-miR-30c-2-3p:Corrected:"  "N:MIR::hsa-miR-30a-5p:Corrected:"
```



Not very interesting because only one group selected MUTA features

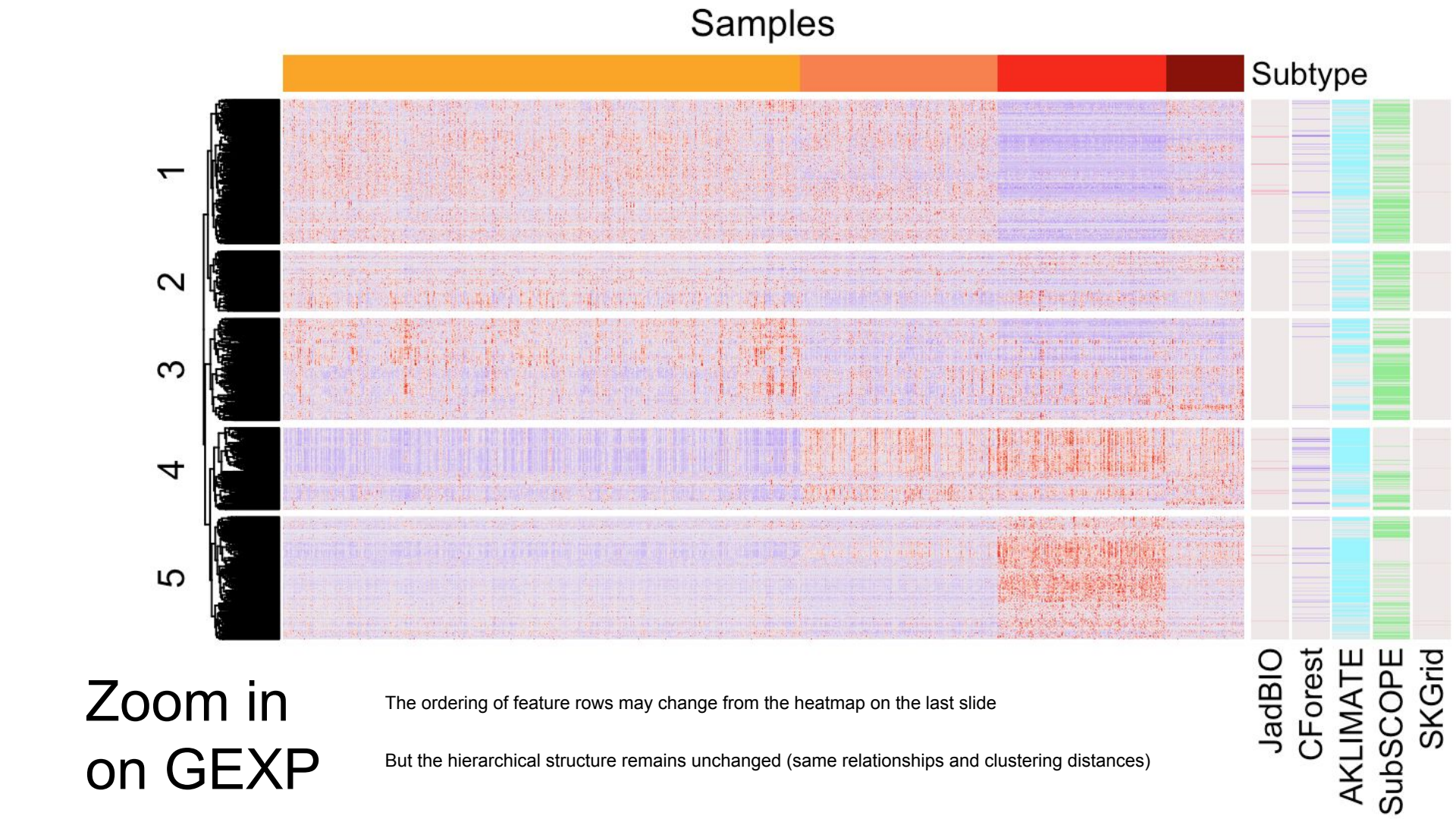
Ordered feature rows

```
[1] 7
[1] "B:MUTA:COMP:PIK3CA::"    "B:MUTA:COMP:TP53::"      "B:MUTA:COMP:HIST1H3C::"
[4] "B:MUTA:COMP:SFRP5::"     "B:MUTA:LOF:CDH1::"       "B:MUTA:COMP:GATA3::"
[7] "B:MUTA:LOF:LAMC1::"
```



CForest, AKLIMATE, and SubSCOPE tend to pick features across each cluster

JadBIO and SKGrid much more sparse across clusters. They do appear to pick features within the same clusters as each other



Zoom in on GEXP

The ordering of feature rows may change from the heatmap on the last slide

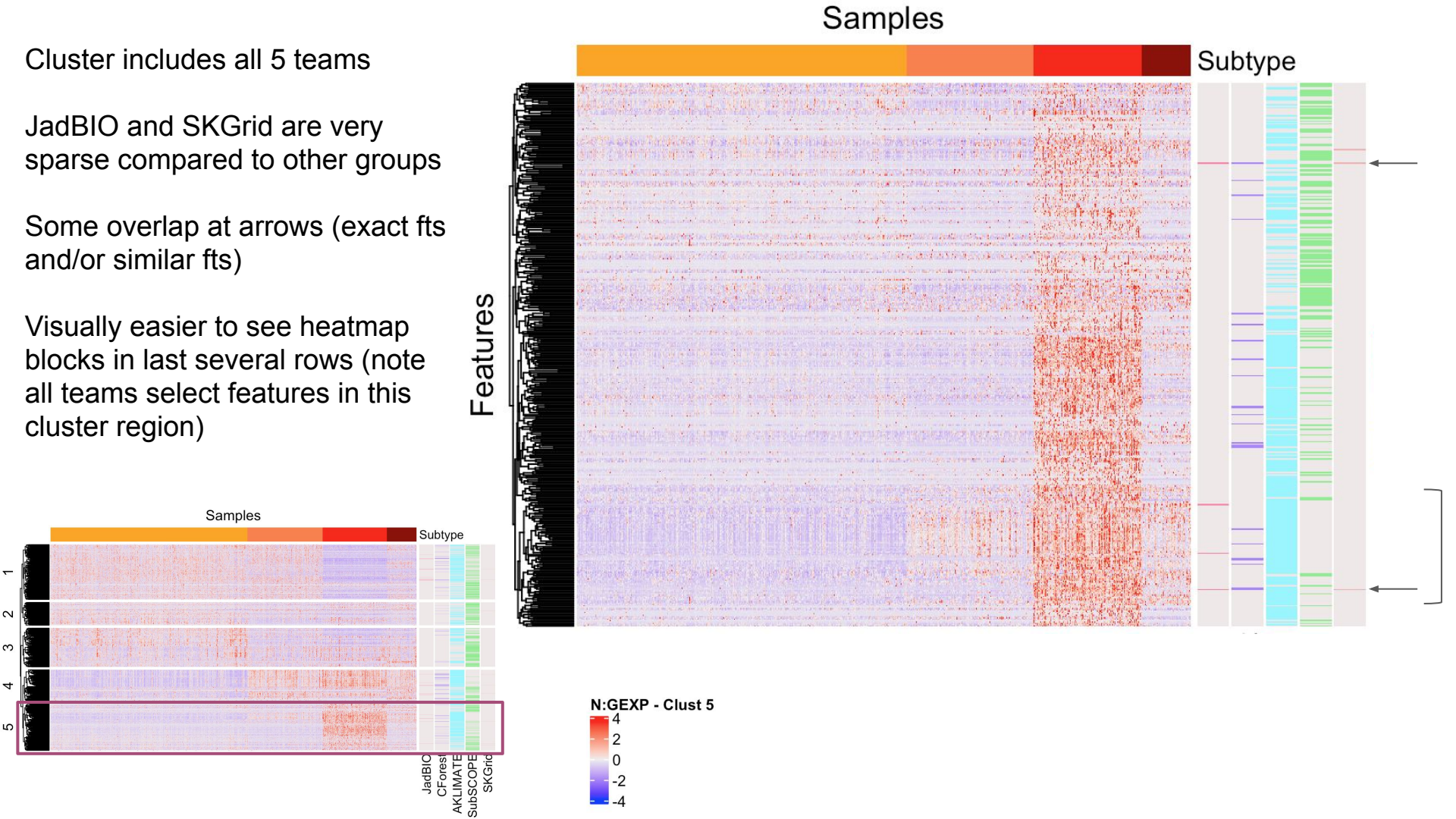
But the hierarchical structure remains unchanged (same relationships and clustering distances)

Cluster includes all 5 teams

JadBIO and SKGrid are very sparse compared to other groups

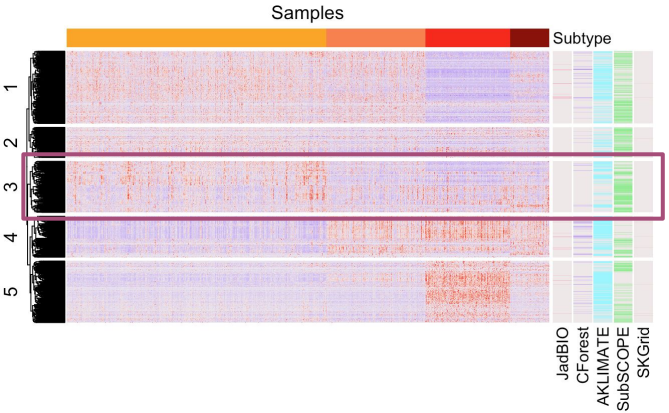
Some overlap at arrows (exact fts and/or similar fts)

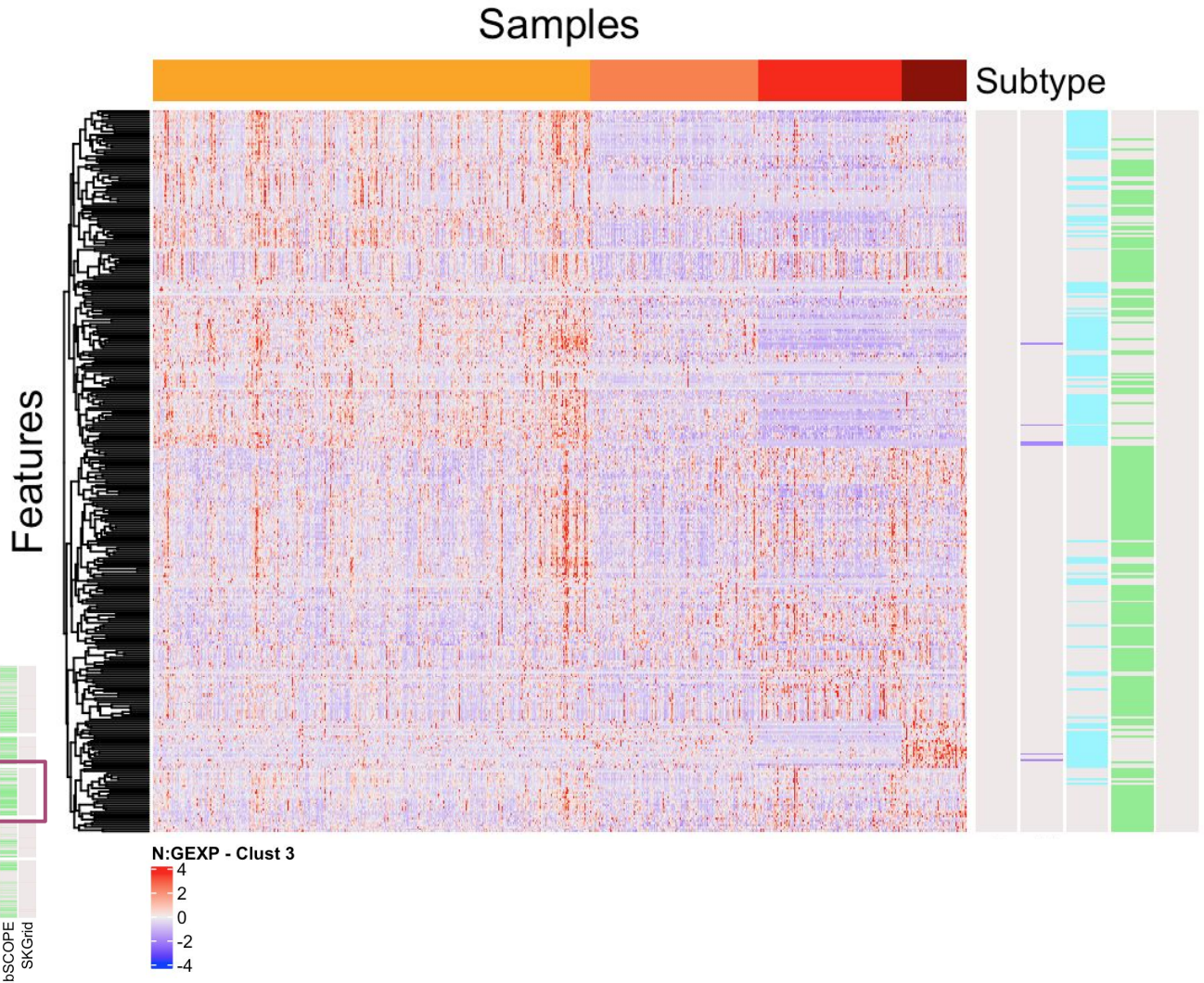
Visually easier to see heatmap blocks in last several rows (note all teams select features in this cluster region)



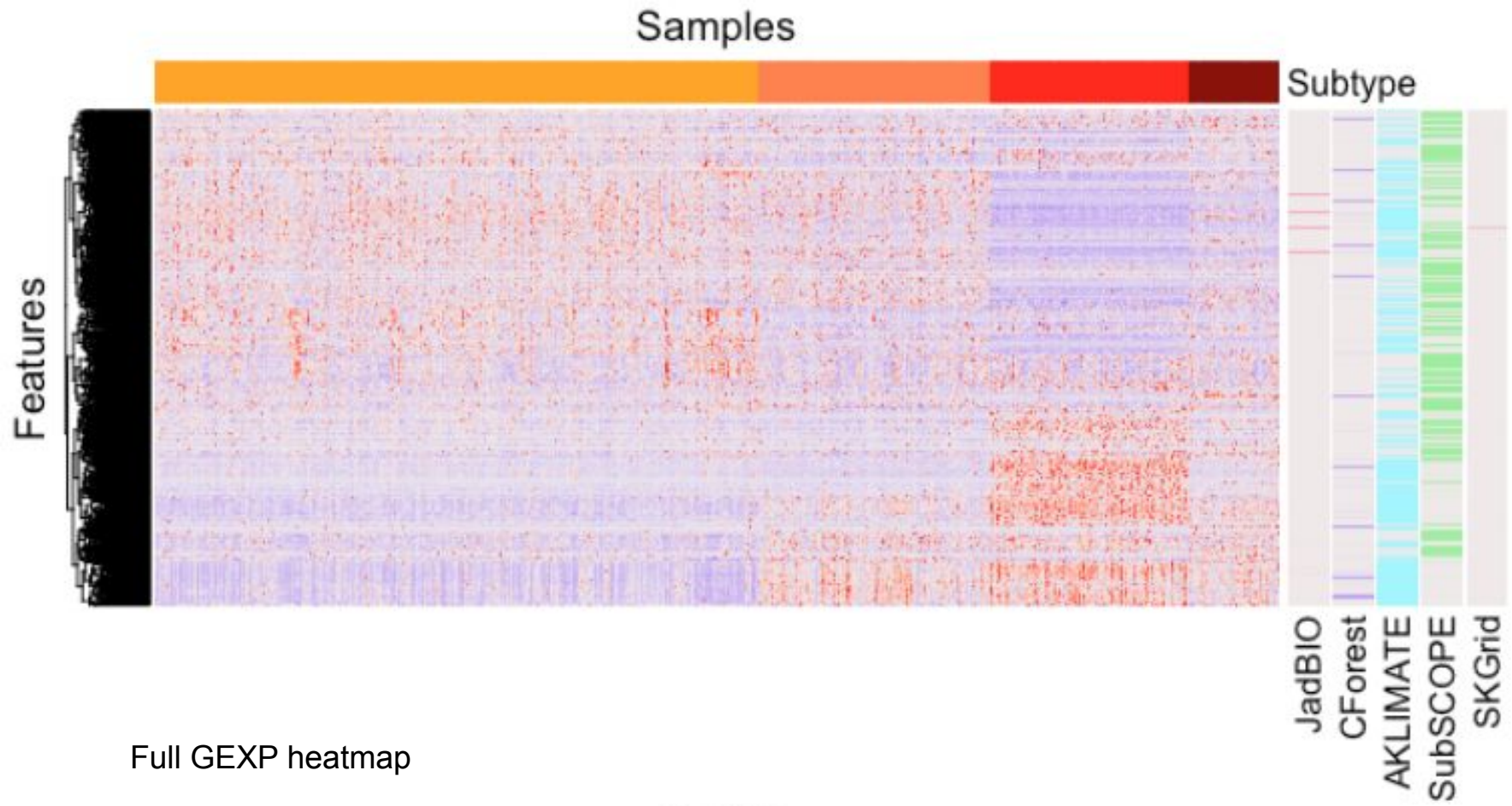
Cluster contains no features selected by JadBIO or SKGrid

AKLIMATE and SubSCOPE tend to select related features



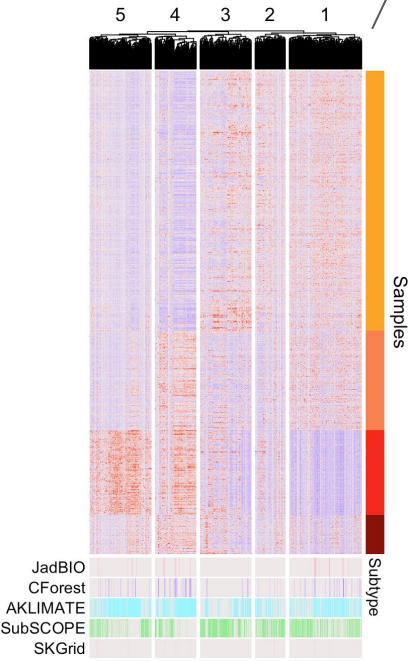


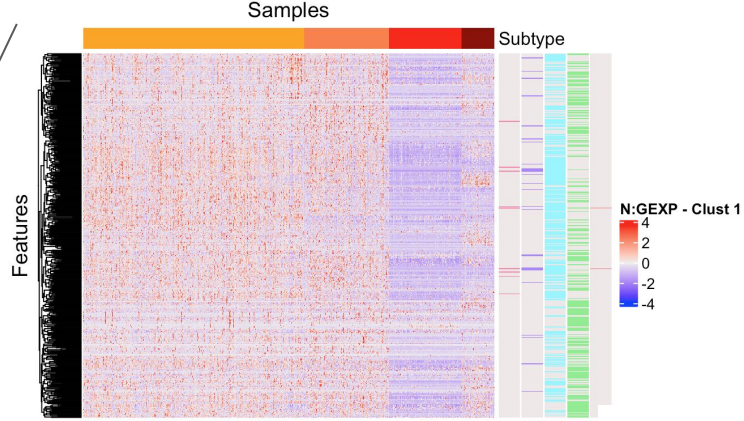
Additional Slides

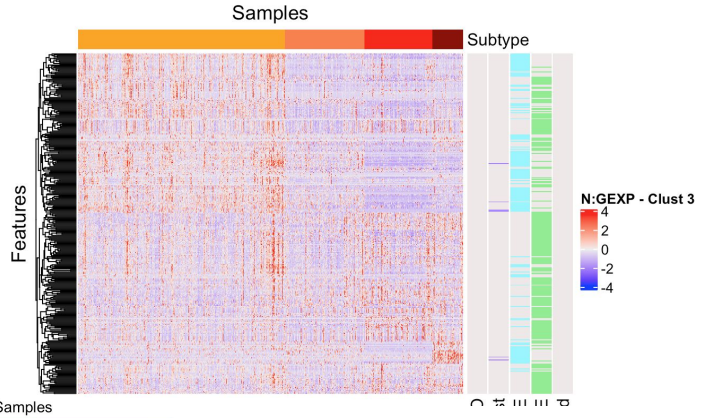


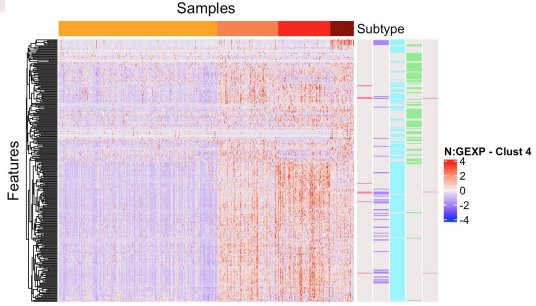
Full GEXP heatmap

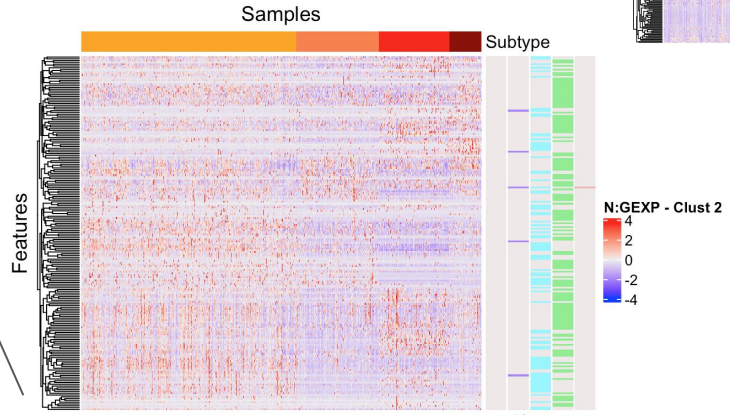
All clusters when have 5 total clusters

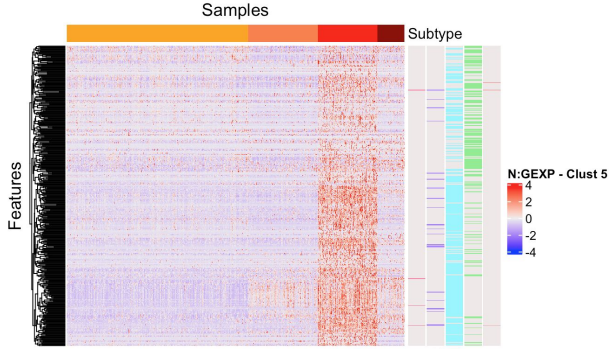




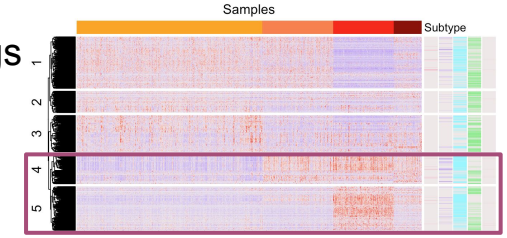


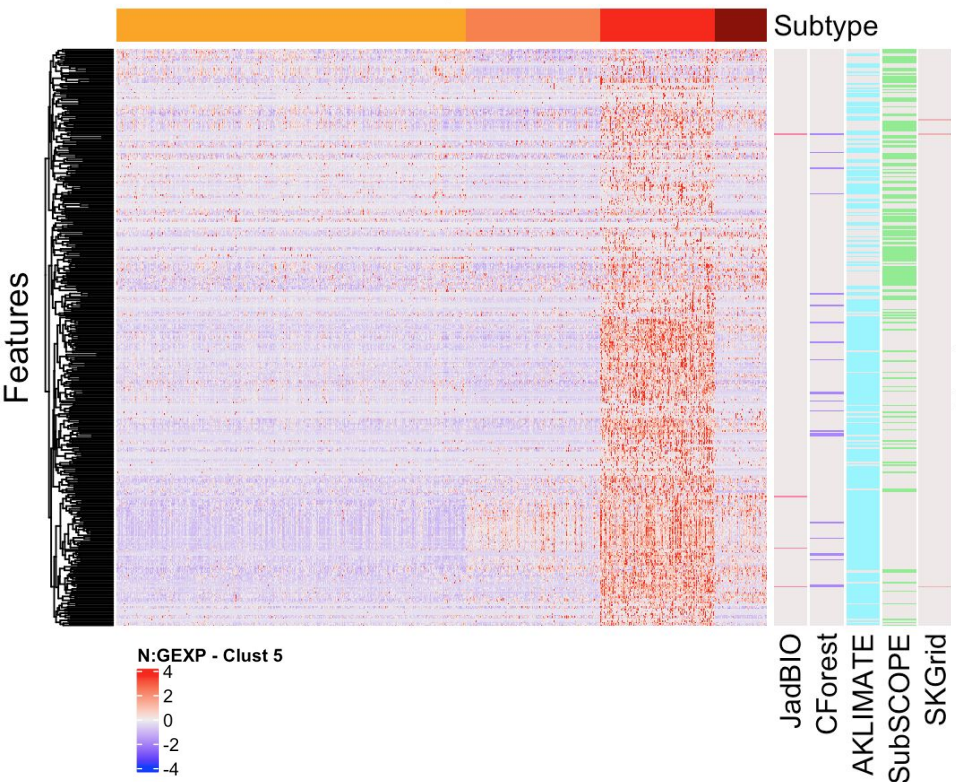


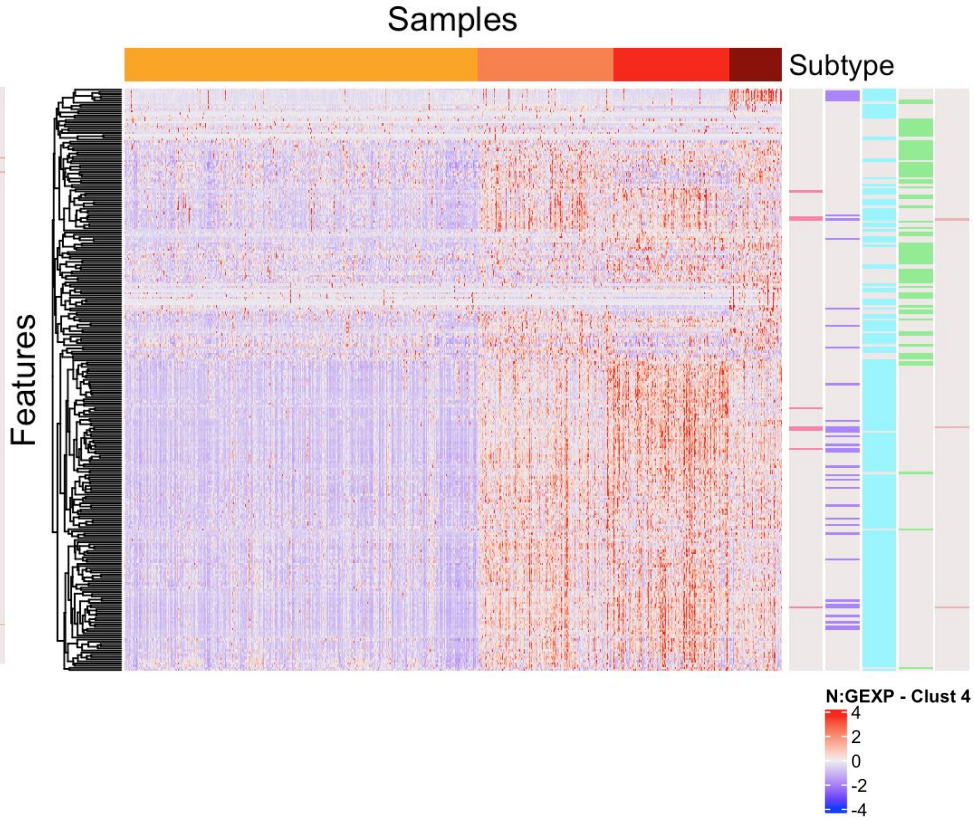




With 5 groupings





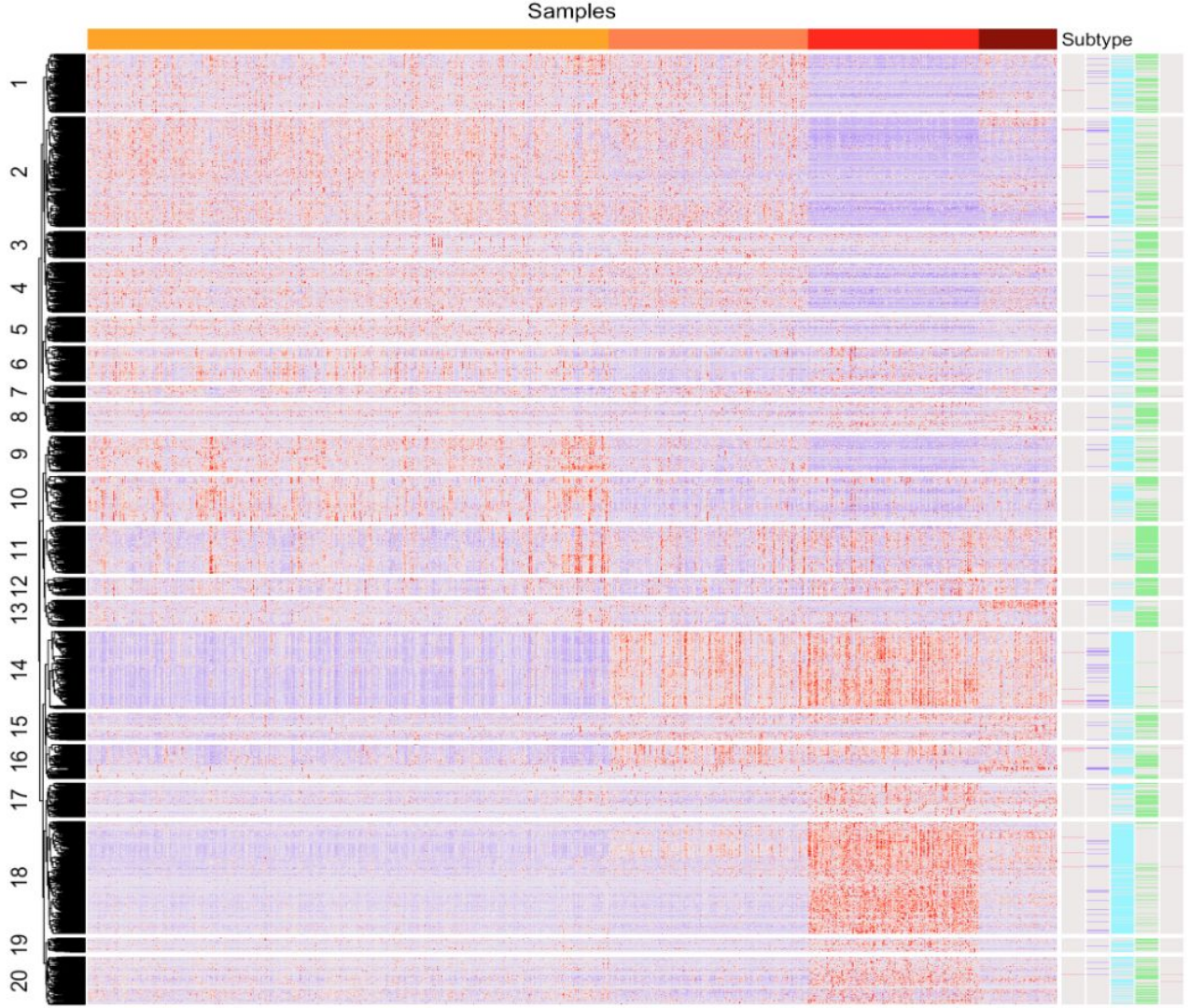


Zoom in on GEXP

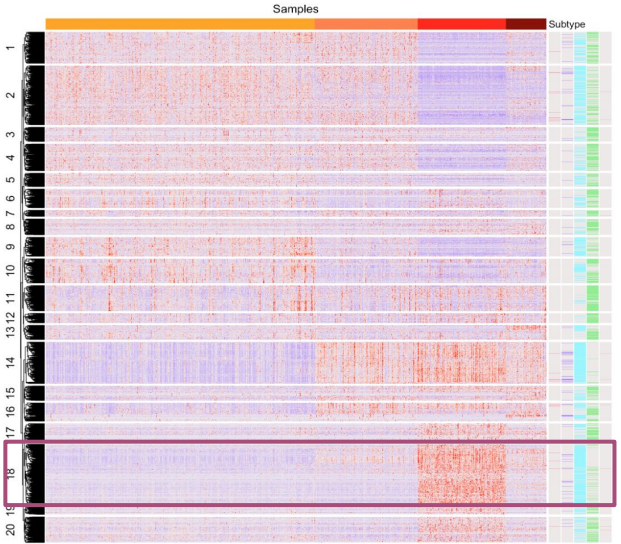
The ordering of feature rows may change from the heatmap on the last slide

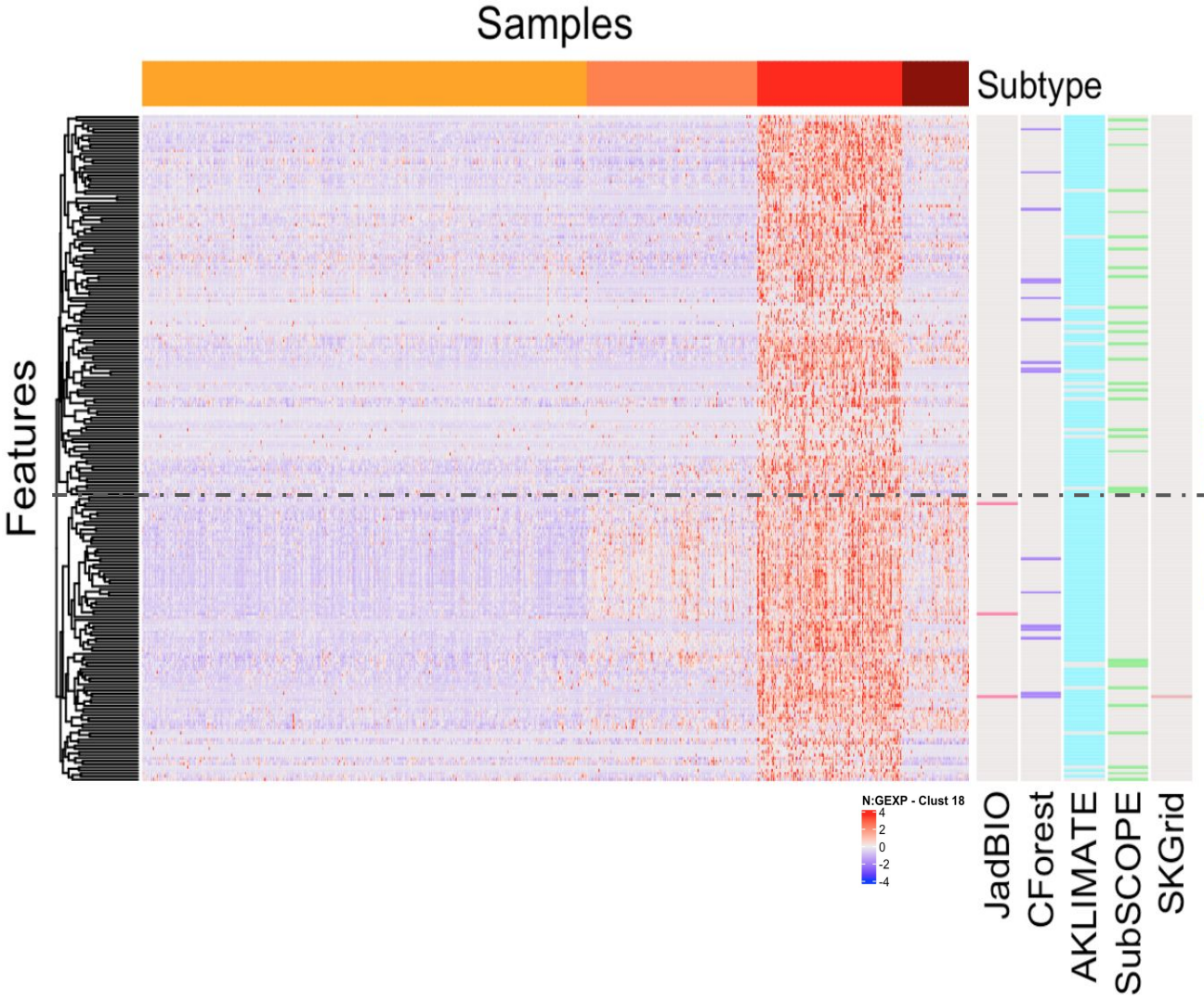
But the hierarchical structure remains unchanged (same relationships and clustering distances)

Cluster 14 and 18 on next slide

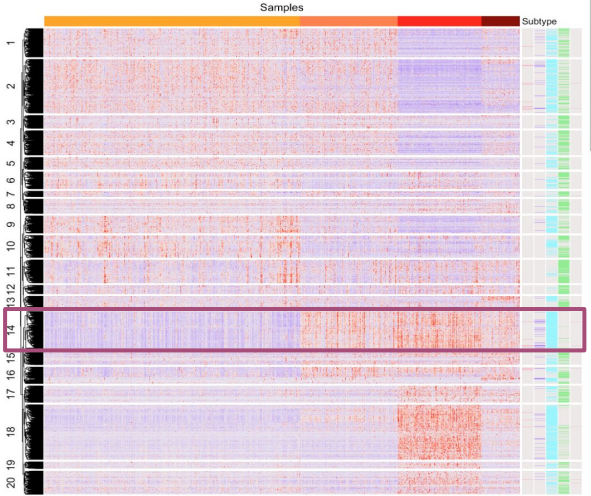


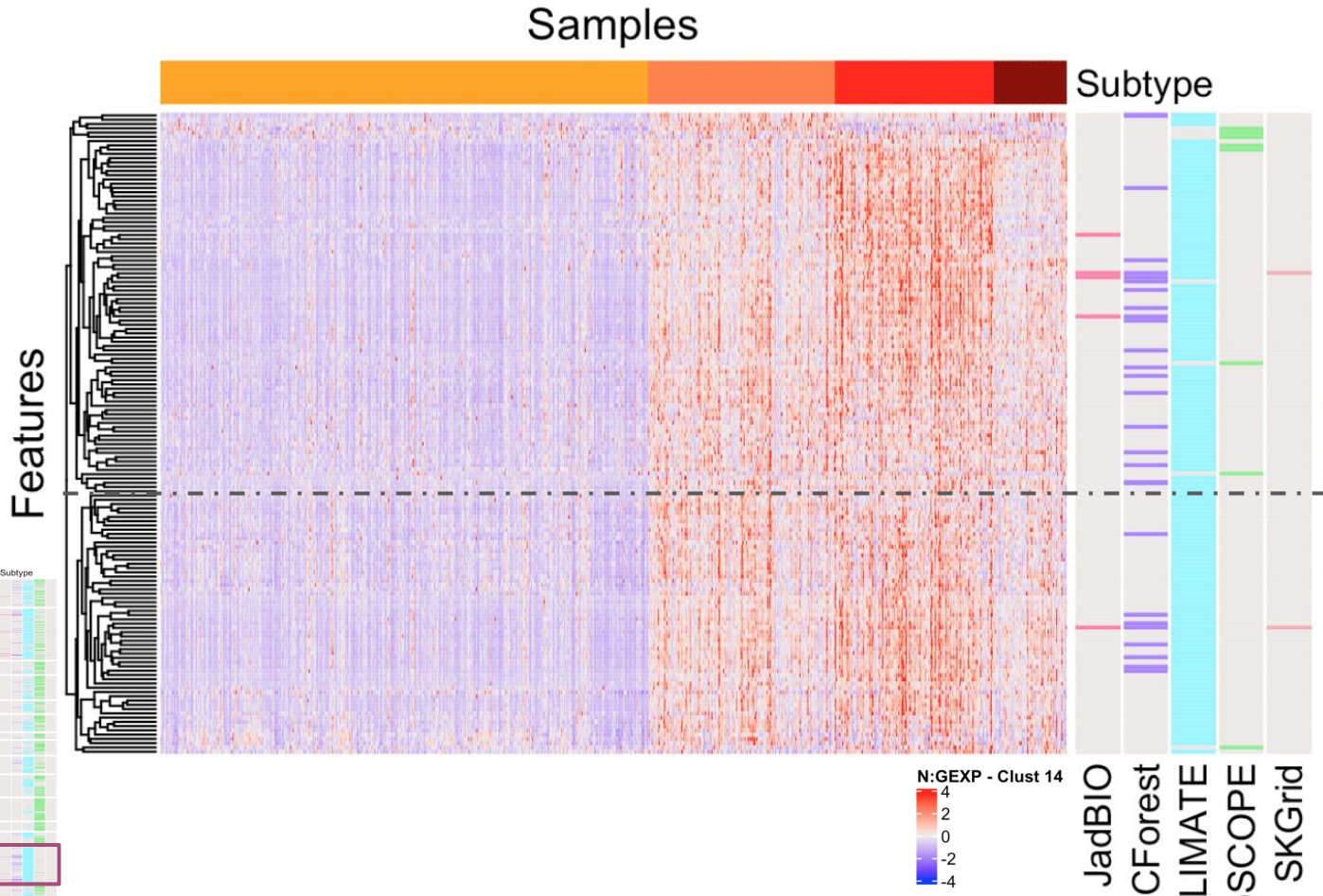
With 20 groupings

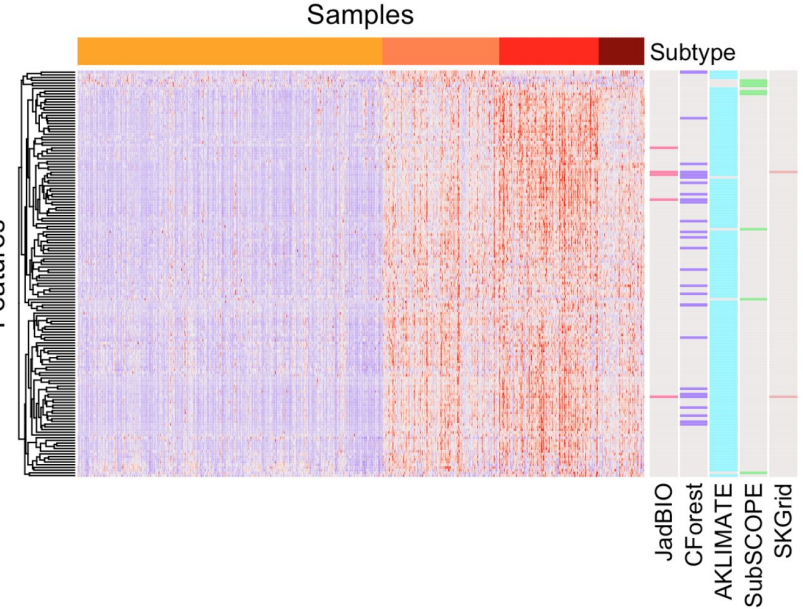


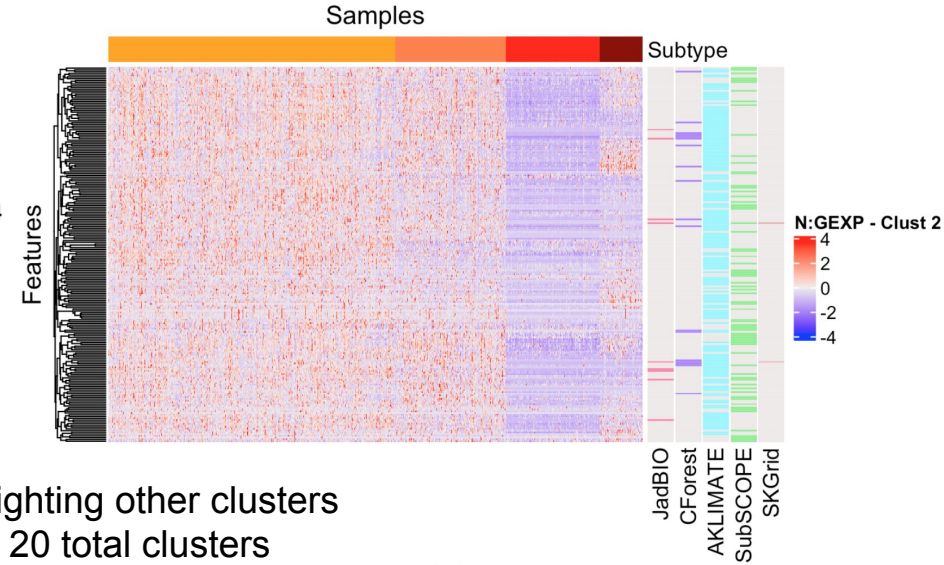


With 20 groupings









Highlighting other clusters
when 20 total clusters