In [1]:
```
from scgenome import tantalus
import pandas as pd
from IPython.display import display
from scgenome import utils, cncluster, simulation, cnplot
import scipy.cluster.hierarchy as sch
import matplotlib.pyplot as plt
import numpy as np
import seaborn as sns
import time
from sklearn import metrics
```

In [2]:
```
all_cn_data_fp = "/Users/massoudmaher/data/sc_1935_1936_1937_cn_data_qc.csv"
all_cn_data = pd.read_csv(all_cn_data_fp)
all_cn_data = all_cn_data.iloc[:,1:]
###
all_cn_data = all_cn_data[all_cn_data["chr"]=="X"]
```

In [3]:
```
all_cn_data.head()
```

Out[3]:

	chr	start	end	reads	gc	copy	state	cell_id	sample_id	library_id
5776	X	1	500000	343	-1.000000	NaN	2	SA922-A90554B-R28-C07	SA922	A90554B
5777	X	500001	1000000	571	0.458294	1.682699	2	SA922-A90554B-R28-C07	SA922	A90554B
5778	X	1000001	1500000	280	-1.000000	NaN	2	SA922-A90554B-R28-C07	SA922	A90554B
5779	X	1500001	2000000	631	0.481712	1.862518	2	SA922-A90554B-R28-C07	SA922	A90554B
5780	X	2000001	2500000	616	-1.000000	NaN	2	SA922-A90554B-R28-C07	SA922	A90554B

In [4]:
```
hmmcopy_tickets = ['SC-1935', 'SC-1936', 'SC-1937']
sample_ids = [["SA922"], ['SA921'], ['SA1090']]

# spike in params
total_ncells = 100
proportions = [0.3, 0.3, 0.4]

# bhc params
n_states = 12
alpha = 0.3
prob_cn_change = 0.8
bhc_incon = 2 # inconsistent score used for making clusters from bhc
bhc_depth = 2

# naive clusering params
naive_method = "complete"
naive_metric = "cityblock"
naive_incon = 1.1
naive_depth = 2

# Params for testing threshold values
params = simulation.expand_grid({"transform":["log","none"], "criterion"
: ["inconsistent"], "threshold": np.arange(0.025, 2, step=0.05)})
params = pd.concat([params, simulation.expand_grid({"transform":["log",
"none"], "criterion": ["distance"], "threshold": np.arange(3, 20, step=1
)})])
```

In [5]:
```
subsample = utils.get_cn_data_submixture(all_cn_data, total_ncells, hmmc
opy_tickets, sample_ids, proportions=proportions)

mixed_cn_data = subsample["mixed_cn_data"]
mixed_cn_data["origin_id_int"] = mixed_cn_data["origin_id"].factorize()[
0]
cell_counts = subsample["cell_counts"]
```

```
/Users/massoudmaher/Documents/Code/scgenome/scgenome/utils.py:169: Sett
ingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: http://pandas.pydata.org/pandas-d
ocs/stable/indexing.html#indexing-view-versus-copy
  jira_cn_data[origin_field_name] = hmmcopy_tickets[i]
```

In [6]:
```
start = time.time()
bhc_linkage, bhc_root, bhc_cell_ids, matrix_data, measurement, variances
= (
    cncluster.bayesian_cluster(mixed_cn_data, n_states=n_states, alpha=a
lpha, prob_cn_change=prob_cn_change)
)
print(f"{time.time()-start}s for BHC on {total_ncells} cells")
```

```
15.014842987060547s for BHC on 100 cells
```

In [7]:

```
bhc_linkage, bhc_plot_data = simulation.get_plot_data(bhc_linkage)
lbhc_plot_data = bhc_plot_data.copy()
lbhc_plot_data[:,2] = np.log(lbhc_plot_data[:,2]) # Log because the high
est link is way higher

naive_linkage = sch.linkage(measurement, method=naive_method, metric=nai
ve_metric)
##
naive_linkage[:,2] = naive_linkage[:,2] + 1
lnaive_linkage = naive_linkage.copy()
lnaive_linkage[:,2] = np.log(lnaive_linkage[:,2])
```

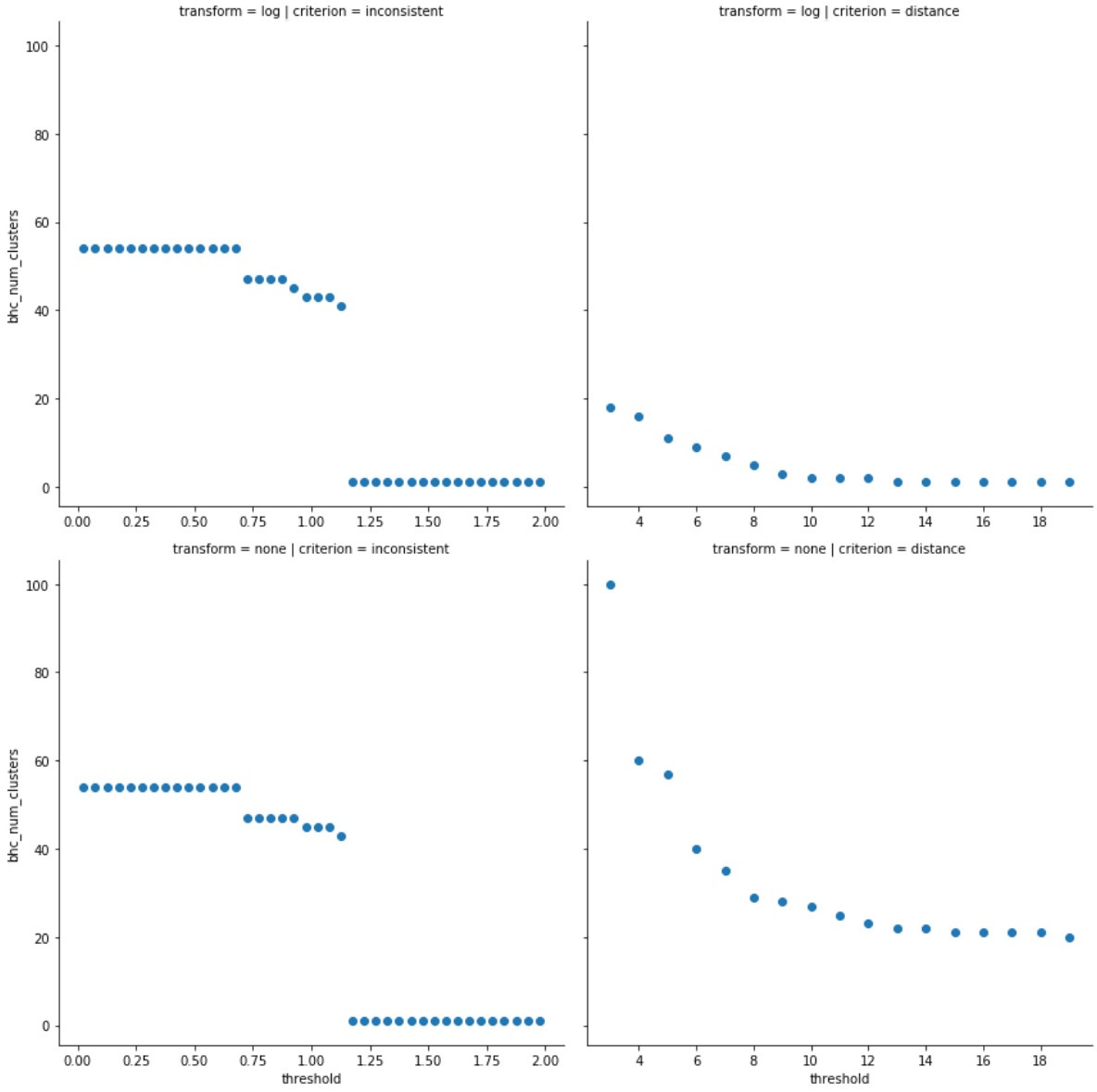
In [8]:
```
def apply_fn(row):
    if row["transform"] == "log":
        df = lbhc_plot_data
    else:
        df = bhc_plot_data
    return sch.fcluster(df, row["threshold"], criterion=row["criterion"
])
params["bhc_fcluster"] = params.apply(apply_fn, axis=1)
params["bhc_num_clusters"] = params["bhc_fcluster"].apply(lambda x: len(
set(x)))

def apply_fn(row):
    if row["transform"] == "log":
        df = lnaive_linkage
    else:
        df = naive_linkage
    return sch.fcluster(df, row["threshold"], criterion=row["criterion"
])
params["naive_fcluster"] = params.apply(apply_fn, axis=1)
params["naive_num_clusters"] = params["naive_fcluster"].apply(lambda x:
len(set(x)))

params.head()
```

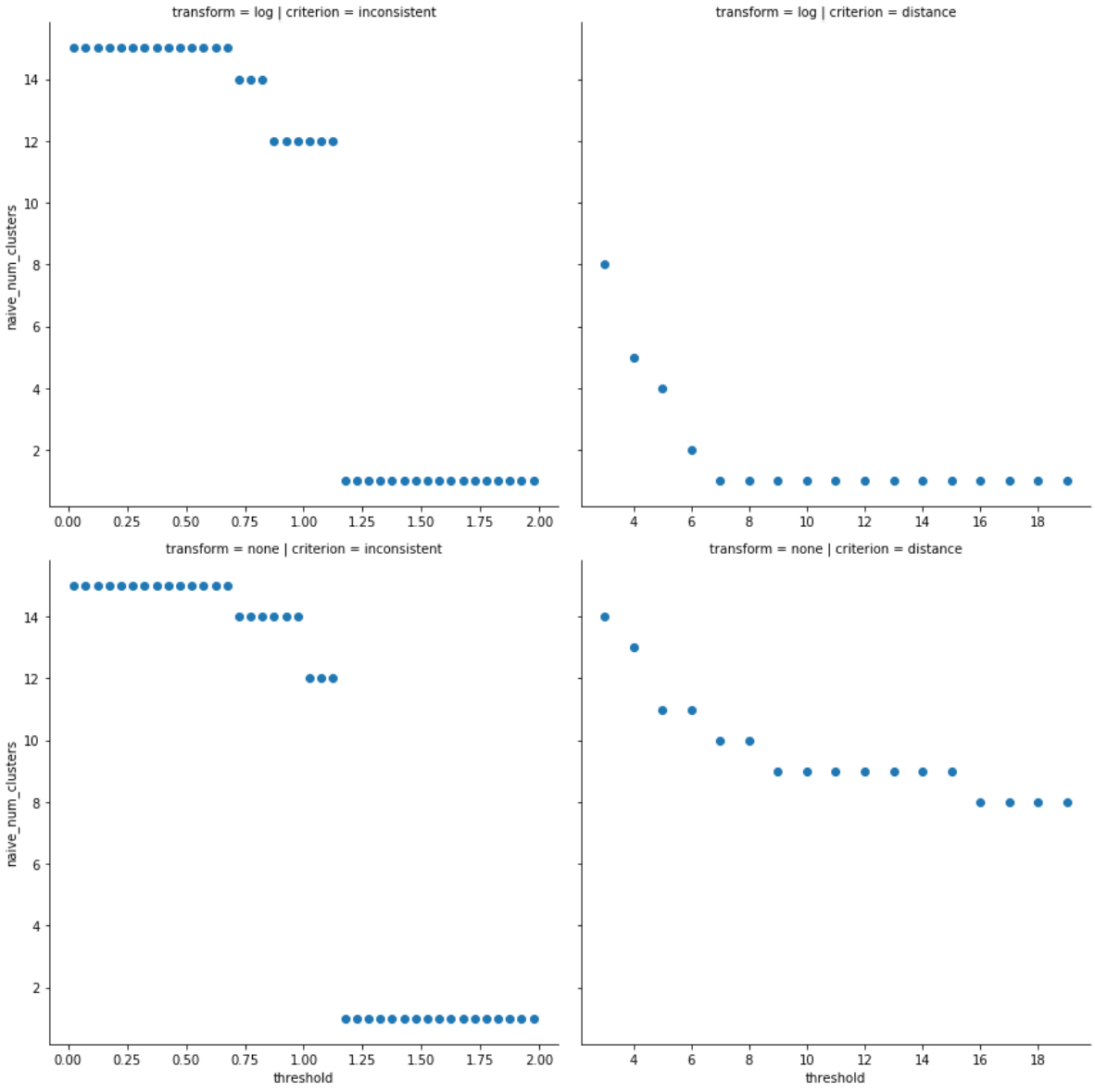
Out[8]:

	transform	criterion	threshold	bhc_fcluster	bhc_num_clusters	naive_fcluster	naive_num_cl
0	log	inconsistent	0.025	[2, 6, 1, 3, 18, 17, 6, 17, 54, 7, 53, 9, 16, ...	54	[5, 6, 5, 5, 10, 8, 6, 7, 2, 6, 1, 3, 4, 9, 3,...	
1	log	inconsistent	0.075	[2, 6, 1, 3, 18, 17, 6, 17, 54, 7, 53, 9, 16, ...	54	[5, 6, 5, 5, 10, 8, 6, 7, 2, 6, 1, 3, 4, 9, 3,...	
2	log	inconsistent	0.125	[2, 6, 1, 3, 18, 17, 6, 17, 54, 7, 53, 9, 16, ...	54	[5, 6, 5, 5, 10, 8, 6, 7, 2, 6, 1, 3, 4, 9, 3,...	
3	log	inconsistent	0.175	[2, 6, 1, 3, 18, 17, 6, 17, 54, 7, 53, 9, 16, ...	54	[5, 6, 5, 5, 10, 8, 6, 7, 2, 6, 1, 3, 4, 9, 3,...	
4	log	inconsistent	0.225	[2, 6, 1, 3, 18, 17, 6, 17, 54, 7, 53, 9, 16, ...	54	[5, 6, 5, 5, 10, 8, 6, 7, 2, 6, 1, 3, 4, 9, 3,...	

In [9]:
```
g = sns.FacetGrid(data=params, col="criterion", row="transform", size=6,
sharey=True, sharex=False).add_legend()
g = g.map(plt.scatter, "threshold", "bhc_num_clusters")

g = sns.FacetGrid(data=params, col="criterion", row="transform", size=6,
sharey=True, sharex=False).add_legend()
g = g.map(plt.scatter, "threshold", "naive_num_clusters")
```

```
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-packages/seaborn/axisgrid.py:230: UserWarning: The `size` paramter has been renamed to `height`; please update your code.
  warnings.warn(msg, UserWarning)
```

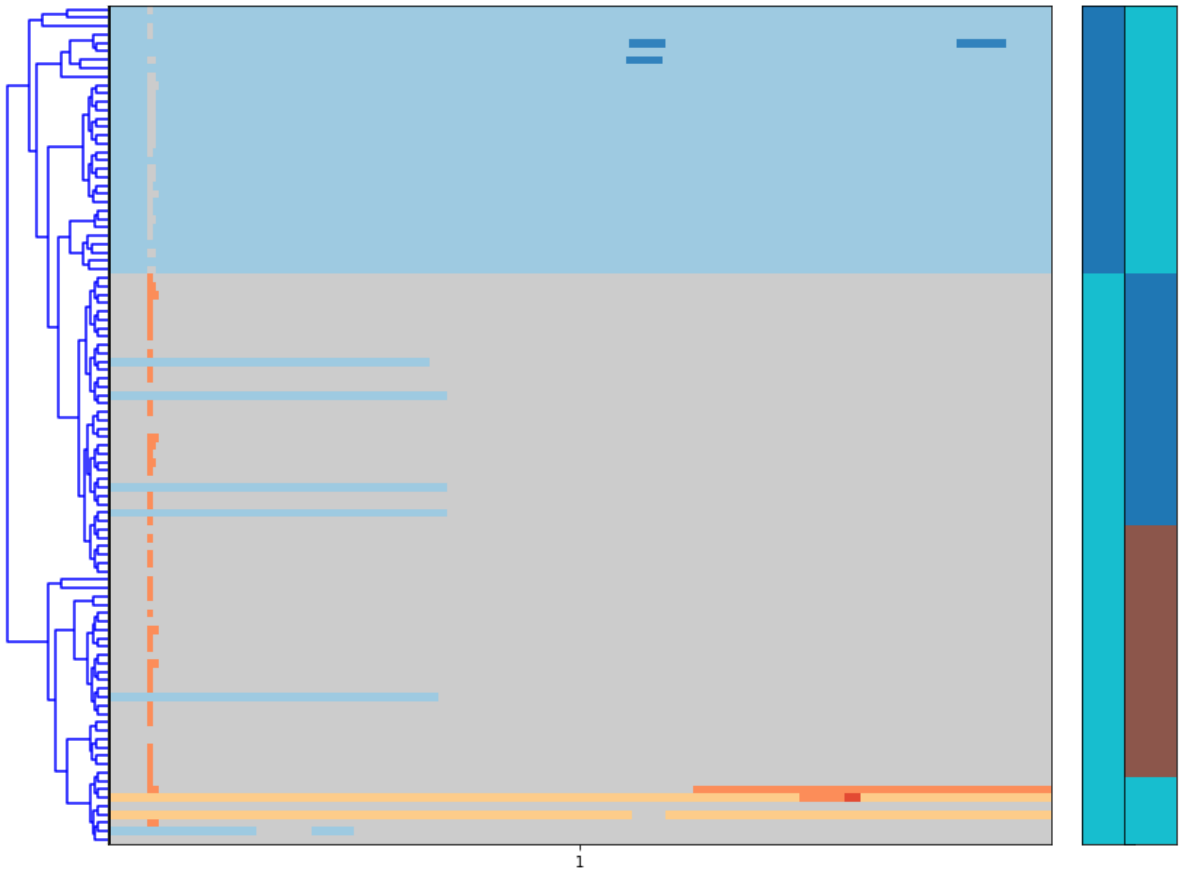




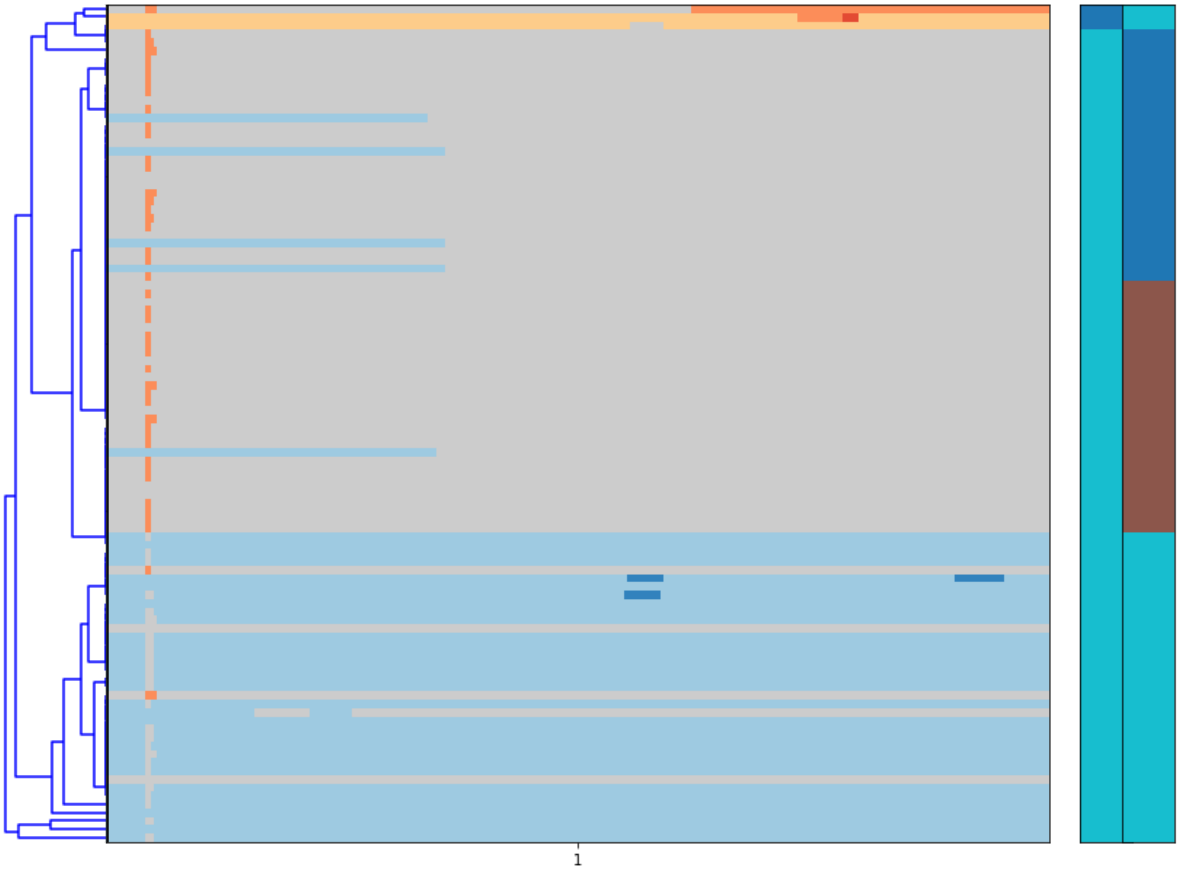
Rightmost bar represents where sample originally came frome

In [10]:
```
#cmixed_cn_data["bhc_cluster_id"].unique()
```

In [16]:
```
cmixed_cn_data = mixed_cn_data.copy()
clustering = sch.fcluster(lbhc_plot_data,12, criterion="distance")
cmixed_cn_data = cncluster.prune_cluster(clustering, bhc_cell_ids, mixed
_cn_data)

fig = plt.figure(figsize=(10, 8))
bimatrix_data = cnplot.plot_clustered_cell_cn_matrix_figure(
    fig, cmixed_cn_data, "state", cluster_field_name="bhc_cluster_id",
    linkage=lbhc_plot_data, origin_field_name="origin_id_int")
```



In [12]:
```
naive_clusters = sch.fcluster(lnaive_linkage, 6, criterion="distance")
assert len(set(naive_clusters)) > 1
nmixed_cn_data = cncluster.prune_cluster(naive_clusters, bhc_cell_ids, m
ixed_cn_data,
                                         cluster_field_name="naive_clust
er_id")
fig = plt.figure(figsize=(10, 8))
bimatrix_data = cnplot.plot_clustered_cell_cn_matrix_figure(
    fig, nmixed_cn_data, "state", cluster_field_name="naive_cluster_id",
    linkage=lnaive_linkage, origin_field_name="origin_id_int")
```



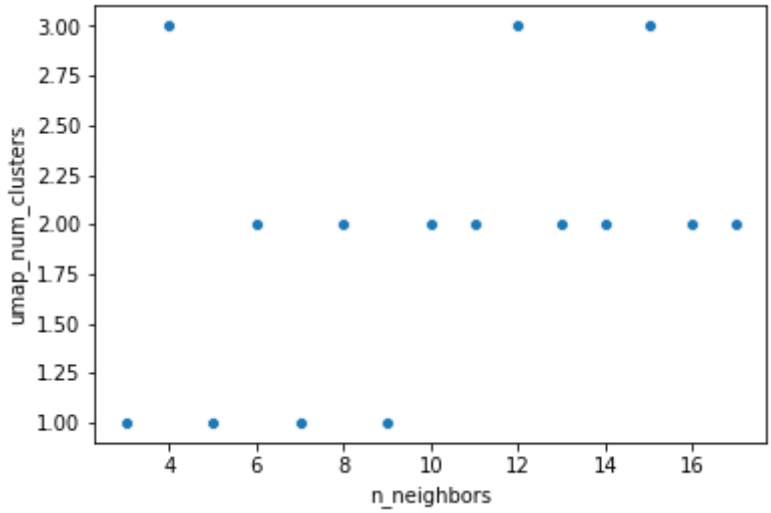
In [13]:

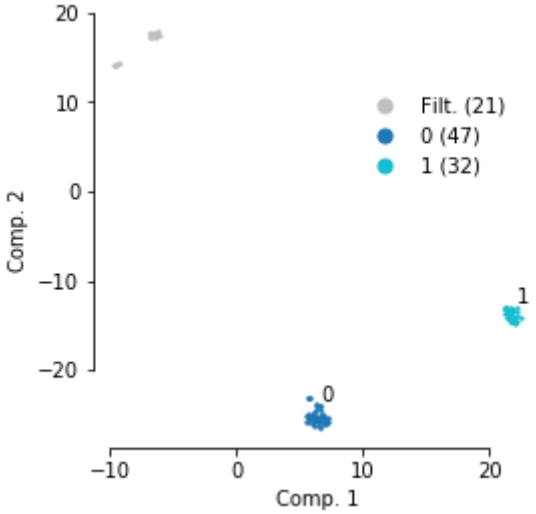
```
umap_params = utils.expand_grid({"n_neighbors": np.arange(3,18,step=1)})
def apply_fn(row):
    return cncluster.umap_hdbscan_cluster(matrix_data["state"], n_neighb
ors=row["n_neighbors"])

umap_params["umap_clusters"] = umap_params.apply(apply_fn, axis=1)
umap_params["umap_num_clusters"] = umap_params["umap_clusters"].apply(la
mbda x: len(set(x["cluster_id"])))

sns.scatterplot(data=umap_params, x="n_neighbors", y="umap_num_clusters"
)

umap_df = cncluster.umap_hdbscan_cluster(matrix_data["state"], n_neighbo
rs=15)
umixed_cn_data = mixed_cn_data.merge(umap_df, how="inner")

fig = plt.figure(figsize=(4, 4))
cncluster.plot_umap_clusters(plt.gca(), umap_df)

fig = plt.figure(figsize=(4, 4))
sns.barplot(x='cluster_id', y='count', data=umap_df.groupby('cluster_id'
).size().rename('count').reset_index())


fig = plt.figure(figsize=(10, 8))
bimatrix_data = cnplot.plot_clustered_cell_cn_matrix_figure(
    fig, umixed_cn_data, "state", cluster_field_name="cluster_id",
    linkage=None, origin_field_name="origin_id_int")
#def umap_hdbscan_cluster(
#        cn,
#        n_components=2,
#        n_neighbors=15,
#        min_dist=0.1,
#):
```

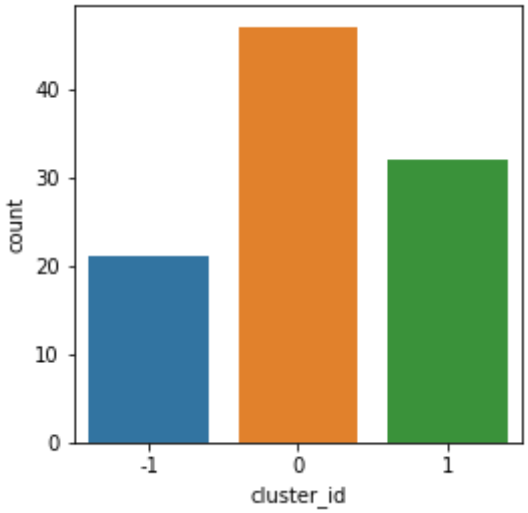
```
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/umap_.py:349: NumbaWarning:
Compilation is falling back to object mode WITH looplifting enabled bec
ause Function "fuzzy_simplicial_set" failed type inference due to: Unty
ped global name 'nearest_neighbors': cannot determine Numba type of <cl
ass 'function'>

File "../scg/lib/python3.7/site-packages/umap/umap_.py", line 467:
def fuzzy_simplicial_set(
    <source elided>
    if knn_indices is None or knn_dists is None:
        knn_indices, knn_dists, _ = nearest_neighbors(
        ^


  @numba.jit()
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/numba/compiler.py:742: NumbaWarning: Function "fuzzy_simplicial_se
t" was compiled in object mode without forceobj=True.

File "../scg/lib/python3.7/site-packages/umap/umap_.py", line 350:
@numba.jit()
def fuzzy_simplicial_set(
^


  self.func_ir.loc))
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/numba/compiler.py:751: NumbaDeprecationWarning:
Fall-back from the nopython compilation path to the object mode compila
tion path has been detected, this is deprecated behaviour.

For more information visit http://numba.pydata.org/numba-doc/latest/ref
erence/deprecation.html#deprecation-of-object-mode-fall-back-behaviour-
when-using-jit

File "../scg/lib/python3.7/site-packages/umap/umap_.py", line 350:
@numba.jit()
def fuzzy_simplicial_set(
^


  warnings.warn(errors.NumbaDeprecationWarning(msg, self.func_ir.loc))
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 7 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 7 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 4 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 4 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
```

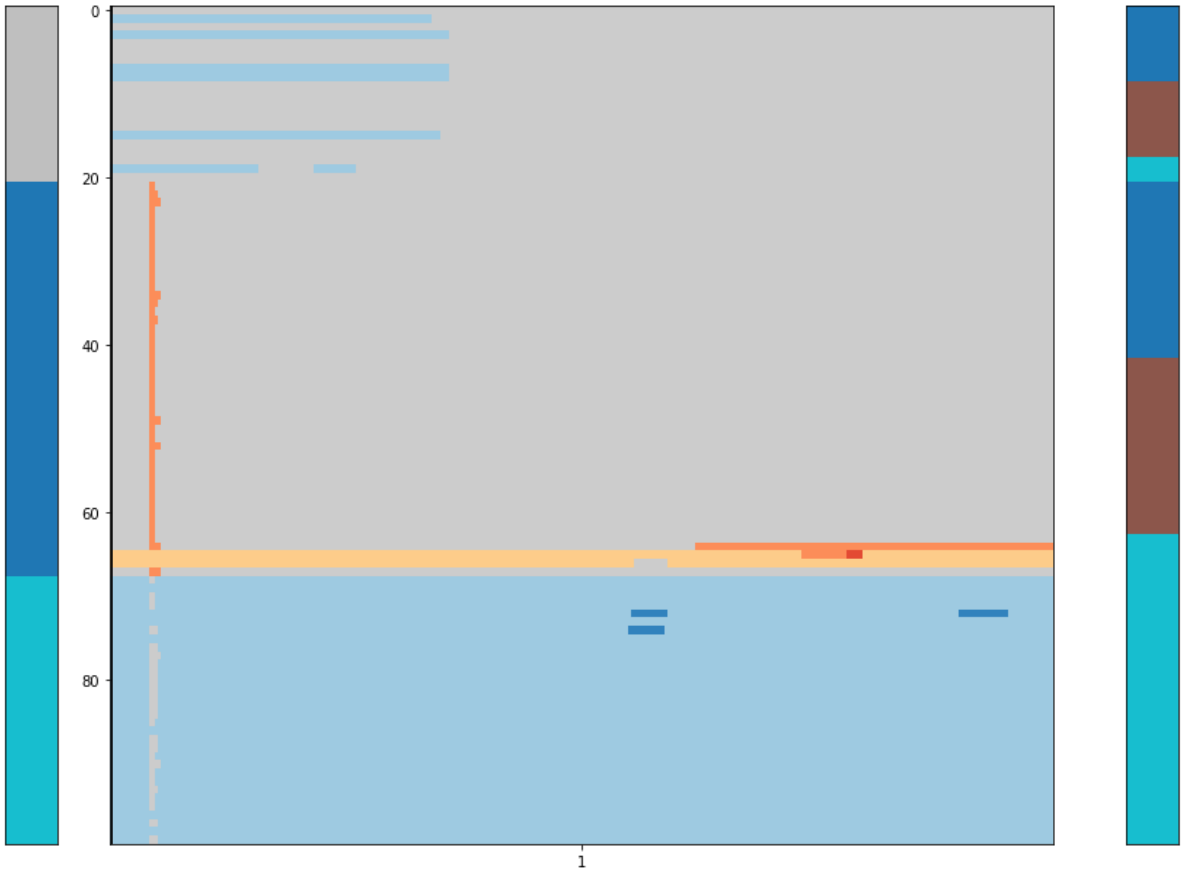
```
ages/umap/spectral.py:229: UserWarning: Embedding a total of 4 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 2 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 2 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 2 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 2 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 2 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 2 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 2 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 2 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 2 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 2 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 2 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/umap/spectral.py:229: UserWarning: Embedding a total of 2 separate
connected components using meta-embedding (experimental)
  n_components
/Users/massoudmaher/Documents/Code/scgenome/scg/lib/python3.7/site-pack
ages/numpy/core/fromnumeric.py:56: FutureWarning: Series.nonzero() is d
eprecated and will be removed in a future version.Use Series.to_numpy
().nonzero() instead
  return getattr(obj, method)(*args, **kwds)
/Users/massoudmaher/Documents/Code/scgenome/scgenome/cnplot.py:50: Clus
```

```
terWarning: scipy.cluster: The symmetric non-negative hollow observatio
n matrix looks suspiciously like an uncondensed distance matrix
  Y = sch.linkage(D, method='complete')
```








0
20
40
60
80
1

In [14]:
```
display(umixed_cn_data.head())
display(nmixed_cn_data.head())
display(cmixed_cn_data.head())
print(cmixed_cn_data.shape)
```

	chr	start	end	reads	gc	copy	state	cell_id	sample_id	library_id	ori
0	X	1	500000	76	-1.000000	NaN	2	SA922-A90554B-R27-C62	SA922	A90554B	SC
1	X	500001	1000000	107	0.458294	1.593015	2	SA922-A90554B-R27-C62	SA922	A90554B	SC
2	X	1000001	1500000	85	-1.000000	NaN	2	SA922-A90554B-R27-C62	SA922	A90554B	SC
3	X	1500001	2000000	145	0.481712	2.270180	2	SA922-A90554B-R27-C62	SA922	A90554B	SC
4	X	2000001	2500000	117	-1.000000	NaN	2	SA922-A90554B-R27-C62	SA922	A90554B	SC

	chr	start	end	reads	gc	copy	state	cell_id	sample_id	library_id
55424	X	1	500000	76	-1.000000	NaN	2	SA922-A90554B-R27-C62	SA922	A90554B
55425	X	500001	1000000	107	0.458294	1.593015	2	SA922-A90554B-R27-C62	SA922	A90554B
55426	X	1000001	1500000	85	-1.000000	NaN	2	SA922-A90554B-R27-C62	SA922	A90554B
55427	X	1500001	2000000	145	0.481712	2.270180	2	SA922-A90554B-R27-C62	SA922	A90554B
55428	X	2000001	2500000	117	-1.000000	NaN	2	SA922-A90554B-R27-C62	SA922	A90554B

	chr	start	end	reads	gc	copy	state	cell_id	sample_id	library_id
55424	X	1	500000	76	-1.000000	NaN	2	SA922-A90554B-R27-C62	SA922	A90554B
55425	X	500001	1000000	107	0.458294	1.593015	2	SA922-A90554B-R27-C62	SA922	A90554B
55426	X	1000001	1500000	85	-1.000000	NaN	2	SA922-A90554B-R27-C62	SA922	A90554B
55427	X	1500001	2000000	145	0.481712	2.270180	2	SA922-A90554B-R27-C62	SA922	A90554B
55428	X	2000001	2500000	117	-1.000000	NaN	2	SA922-A90554B-R27-C62	SA922	A90554B

```
(31100, 13)
```

In [ ]: