Cytogenetic Analysis of B-Cell Malignancies Using Mercator, RCytoGPS, and CytoGPS

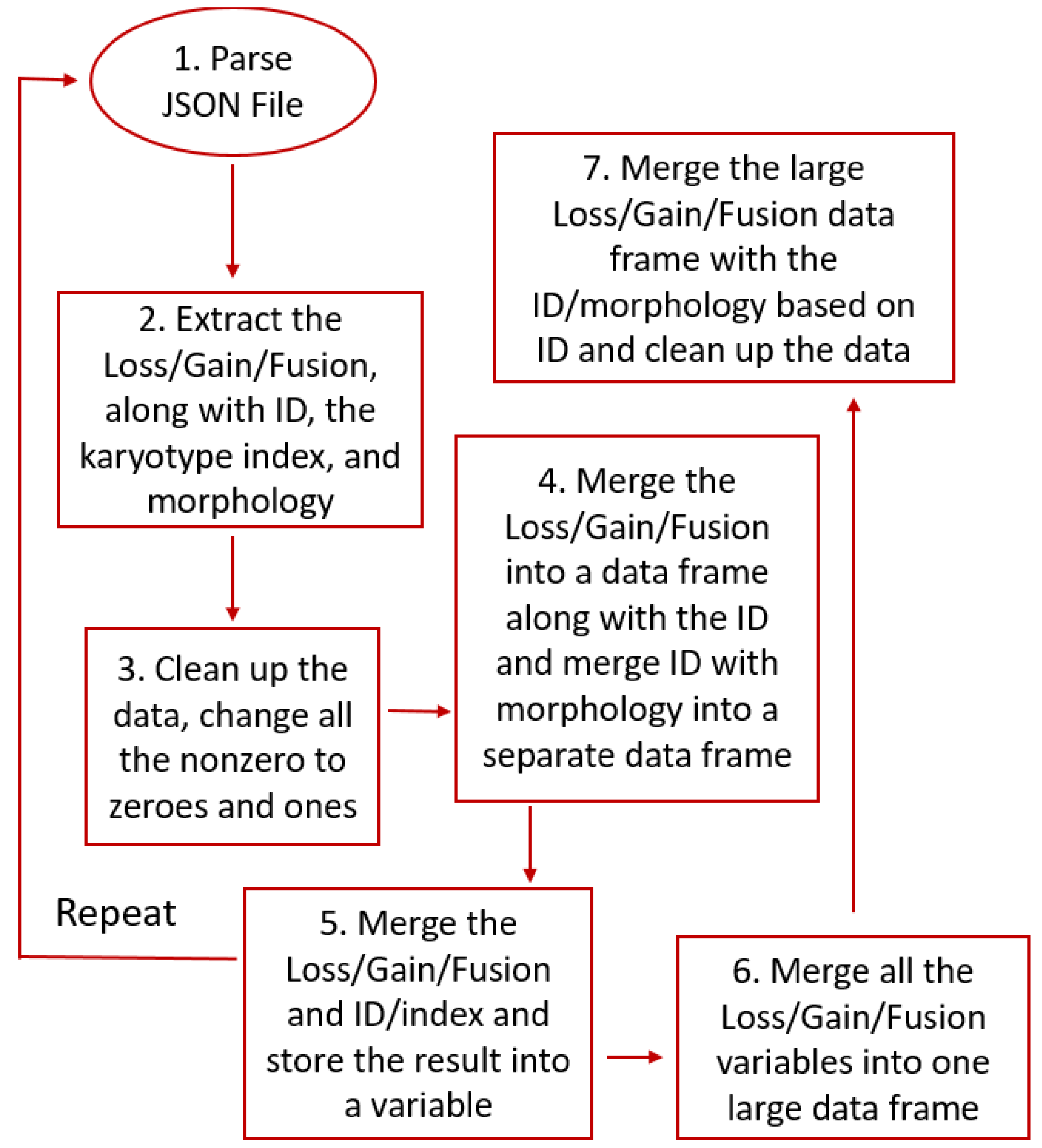
Dwayne G. Tally[1], Zachary B. Abrams[2], Caitlin E. Coombes[2], Suli Li[2], and Kevin R. Coombes[2]

[1]The Center for Genomic Advocacy at Indiana State University, [2]Department of Biomedical Informatics at Ohio State University

Background

Lymphoid cells are one of the two most common cell types from which leukemias and lymphomas are derived. The current World Health Organization (WHO) classification of lymphoid malignancies incorporates a variety of factors, including cytogenetics, cell-of-origin, location, clinical findings, immunophenotype, histological patterns and mutations or rearrangements of specific genes. We wanted to test how well a classifier using only cytogenetic data would perform. Our dataset came from the largest public database for cytogenetic data is The Mitelman Database of Chromosomal Aberrations and Gene Fusions in Cancer. We previously developed CytoGPS, a tool that converts text karyotypes into binary vectors using a Loss-Gain-Fusion model. To test our hypothesis, we applied CytoGPS to the lymphoid malignancies in the Mitelman Database. Here, we present an unsupervised clustering analysis using the R packages Thresher, Mercator, and RCytoGPS.

Methods



1. Parse JSON File

2. Extract the Loss/Gain/Fusion, along with ID, the karyotype index, and morphology

3. Clean up the data, change all the nonzero to zeroes and ones

4. Merge the Loss/Gain/Fusion into a data frame along with the ID and merge ID with morphology into a separate data frame

5. Merge the Loss/Gain/Fusion and ID/index and store the result into a variable

6. Merge all the Loss/Gain/Fusion variables into one large data frame

7. Merge the large Loss/Gain/Fusion data frame with the ID/morphology based on ID and clean up the data

Figure 1: Workflow for analyzing the Mitelman database obtained from the Cancer Genome Anatomy Project web site. The first step was to extract the data from Mitelman for all 69,174 patients. Then we ran CytoGPS to convert the data to the Loss/Gain/Fusion (LGF) model, stored in 14 JSON files. A total of 22,741 samples were associated with lymphoid malignancies. Afterwards, using Thresher and Mercator, we found that there were 134 clusters; we assigned samples to clusters using Partitioning Around Medoids (PAM). We visualized the results using t-distributed Stochastic Neighbor Embedding (t-SNE). We calculated high-frequency events and displayed them in a heatmap. Then we developed a pipeline R package called RCytoGPS which converts the data stored in the JSON file into a s4 object. The s4 object is then used to generate idiogram images to see the frequency of karyotype events along the chromosomes.

Results: t-SNE Plot and Cluster Identities

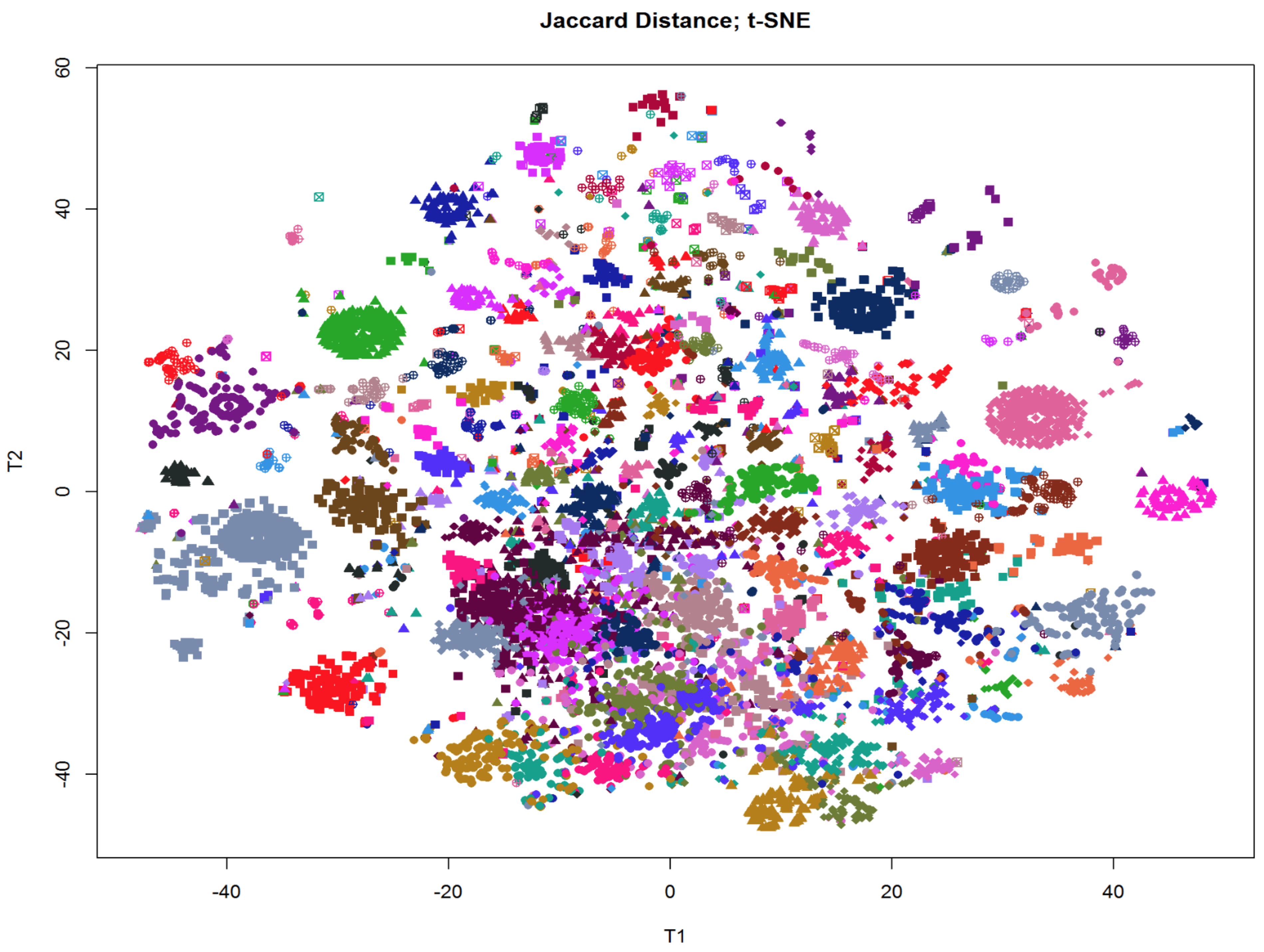


Figure 2: t-SNE map of the 134 clusters and related cytogenetic events. The clusters represent cytogenetic event. There are numerous clusters that form a sort of gradient that could also separate each cluster. As pointed out in the figure we defined several clusters as a gain or loss of the chromosome including sex chromosome.

Cluster	Symbol	Karyotype	Frequency
ST132	☒	+Y	100
ST131	☒	+22	100
ST114	⊕	+20	100
ST101	⊕	t(19q)	100
ST055	▲	+10,+4,+21,+6,+X,+18,-14,+17	100,100,96,87,86,79,74,63
ST053	▲	t(12p;21q)	100
ST047	■	t(8q;14q)	100
ST087	◆	+4,+21,+6,+X,-14	99,92,80,80,63
ST035	■	add(7q),+8,t(7p;18q)	99,96,84
ST060	▲	-Y	99
ST046	■	+8	99
ST081	◆	+X,+21	99,98
ST016	●	+X	99
ST075	◆	+X,-14,+21	99,98,75
ST097	⊕	+16	99
ST095	◆	+15	99
ST041	■	add(7q),t(7p;18q)	99,88
ST008	●	+21	99

Table 1: This table displays the top 18 clusters based on frequency of karyotypic event. The frequency represents the percent of cases classified to a cluster that contain the reported karyotype. The findings revealed 49 clusters that include at least 90% of the samples to have the same abnormalities, 70 clusters have a cytogenetic event with at least 80% similarity, and 84 clusters that have 70% similarity. This demonstrate that Mercator generates high fidelity clusters based on cytogenetic patterns.

Results: Heatmap – Abnormalities vs Clusters

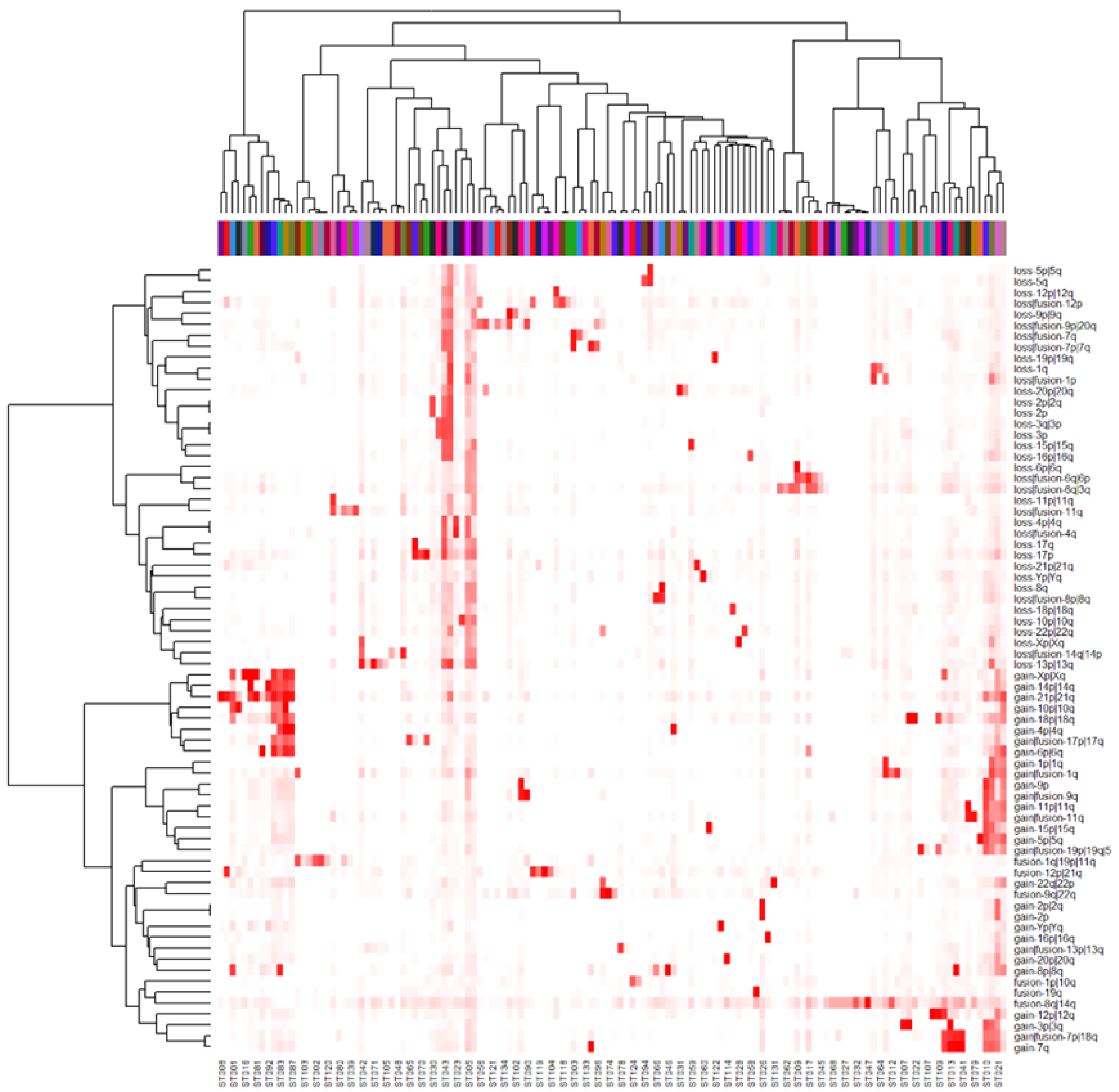


Figure 3: This heatmap displays frequent cytogenetic events by clusters. The dendrogram for frequent cytogenetic events clearly shows that the largest separation is based on gains versus losses. The highest level of distinction amongst clusters separates a group of clusters with multiple trisomies, further demonstrating the distinction between cases with monosomies compared to trisomies.

Conclusion

- Comprehensive analysis of karyotype data
  - Enables novel discovery
  - Produce visual models that are easier to process
- Mercator allows comprehensive analysis based on visualizations
- Our method recovers clusters though high fidelity
- RCytoGPS
  - Generates new visuals of the cytogenetic events in the form of Idiograms
  - Allows closer analysis on cytogenetic events

Acknowledgement

- NIH R25-MD011712-01: Big Data for Indiana State University (BD4ISU)
- The OSU Biomedical Informatics Summer 2019 Internship Program
- NCI Grant # R03CA235101
- Pelotonia Intramural Research Funds from the OSU James Cancer Center

Indiana State University