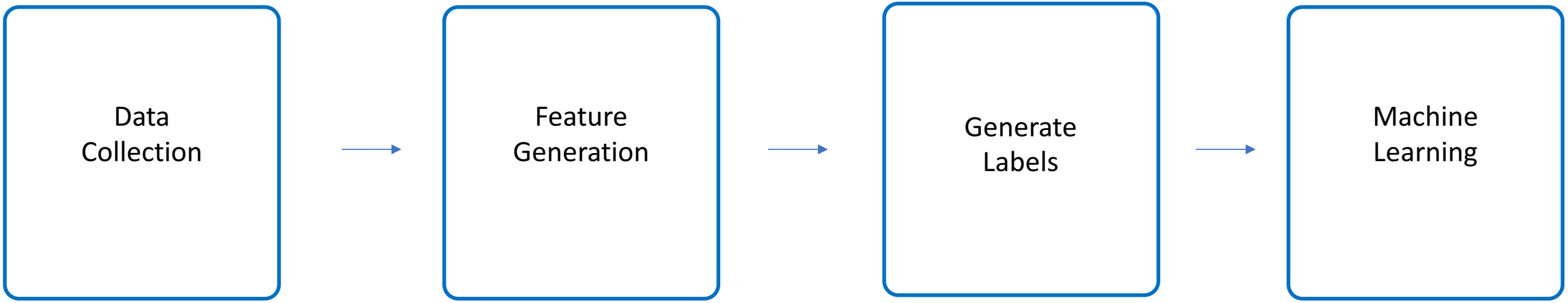
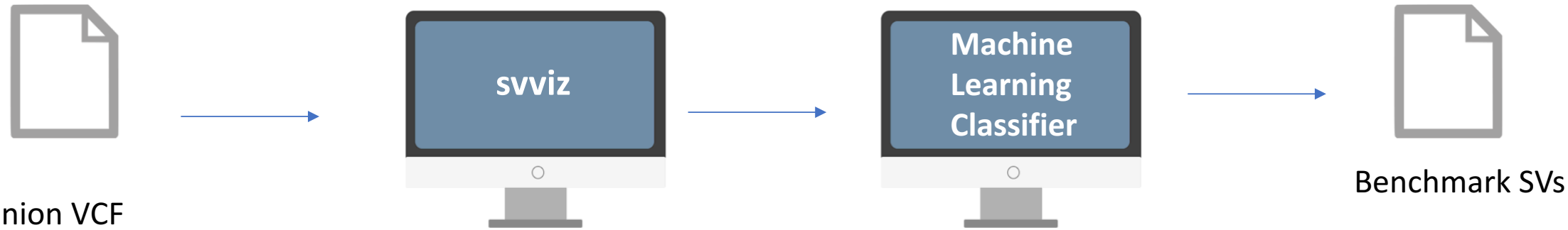
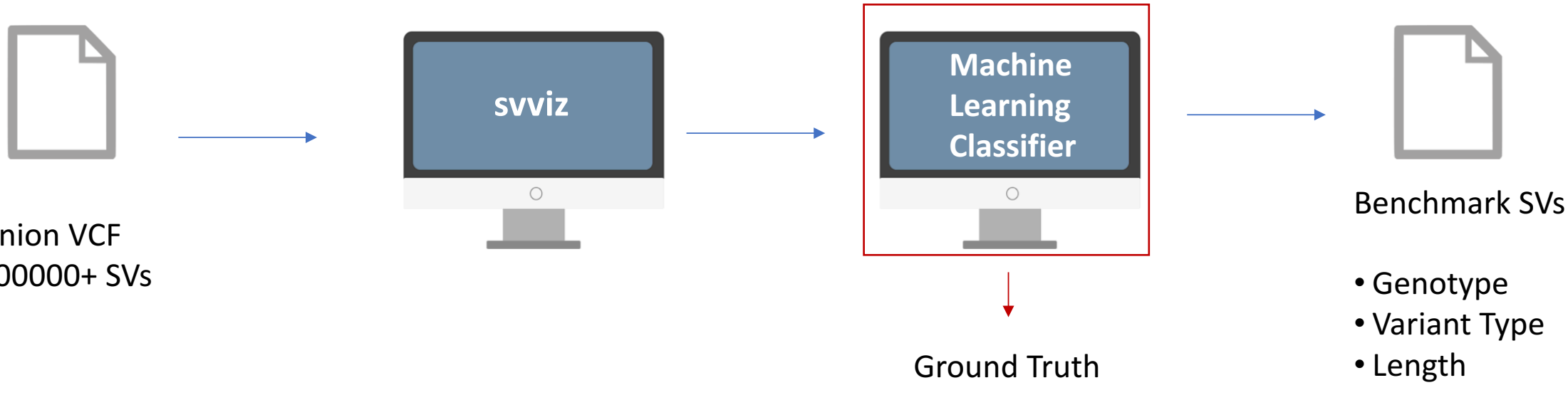

Data Collection
Feature Generation
Generate Labels
Machine Learning

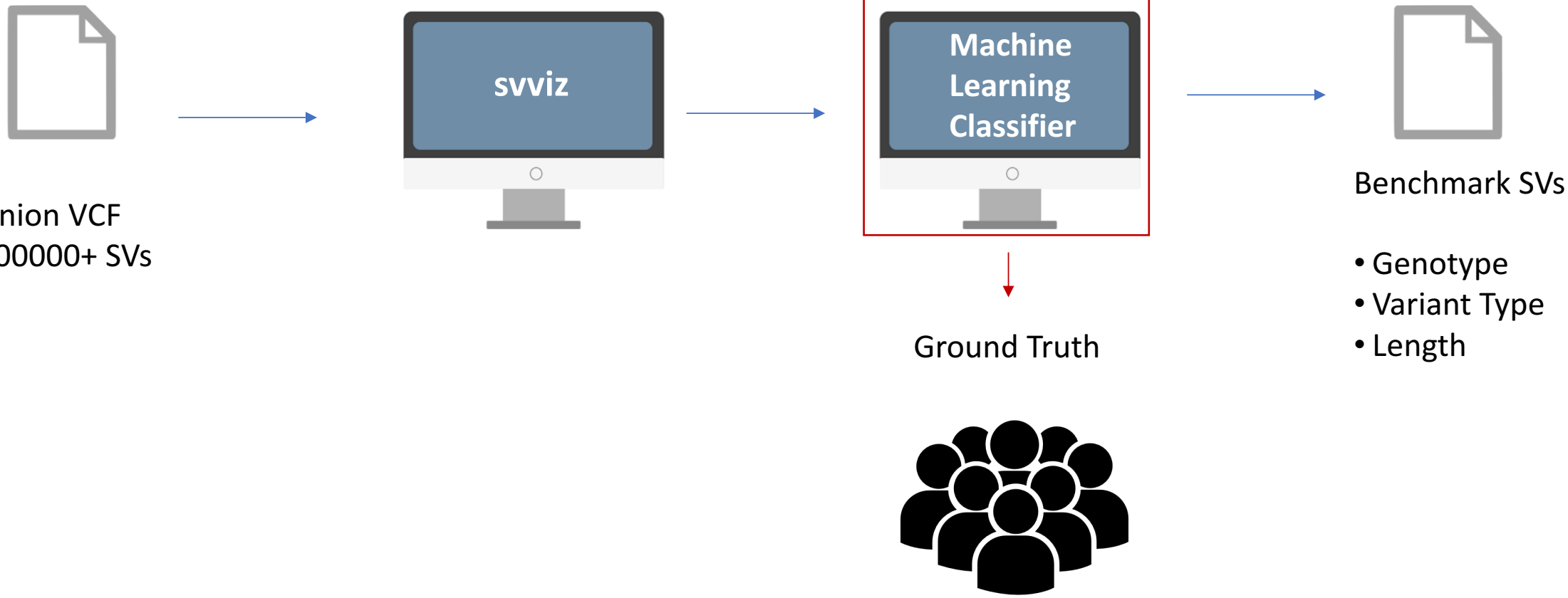


Union VCF
500000+ SVs

Benchmark SVs

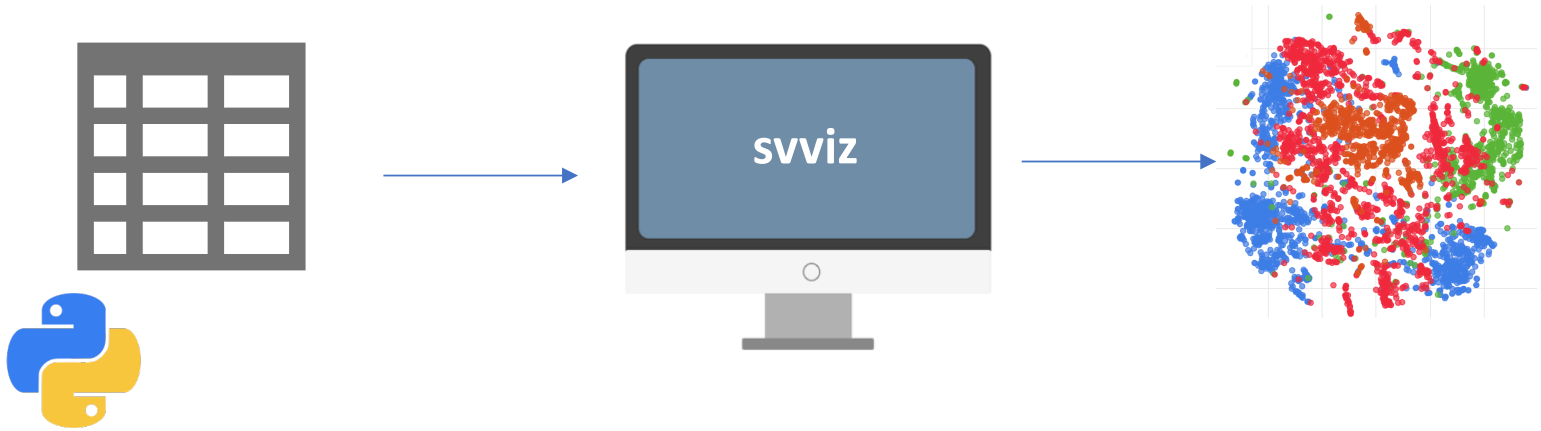
- Genotype
- Variant Type
- Length


svviz
Machine Learning Classifier
Union VCF
500000+ SVs
Ground Truth
Benchmark SVs
• Genotype
• Variant Type
• Length


svviz
Machine Learning Classifier
Union VCF
500000+ SVs
Ground Truth
Benchmark SVs
• Genotype
• Variant Type
• Length

Generate labeled data

Overview

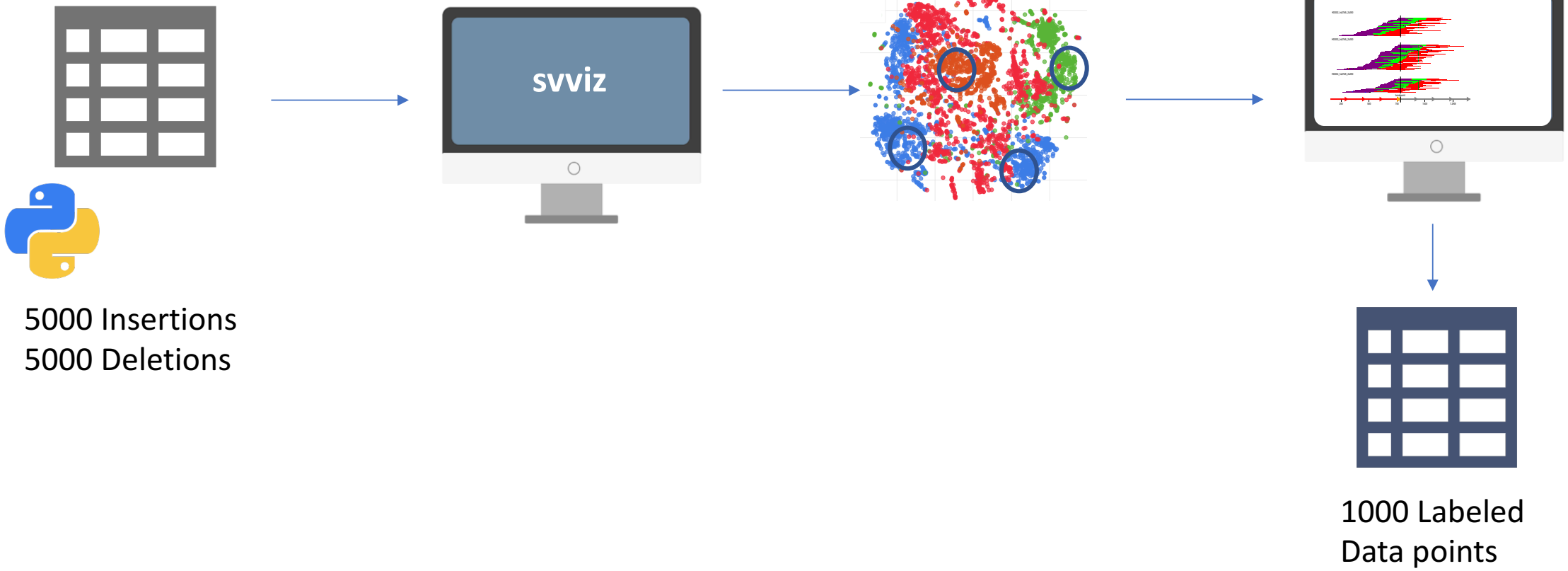


5000 Insertions
5000 Deletions

Overview

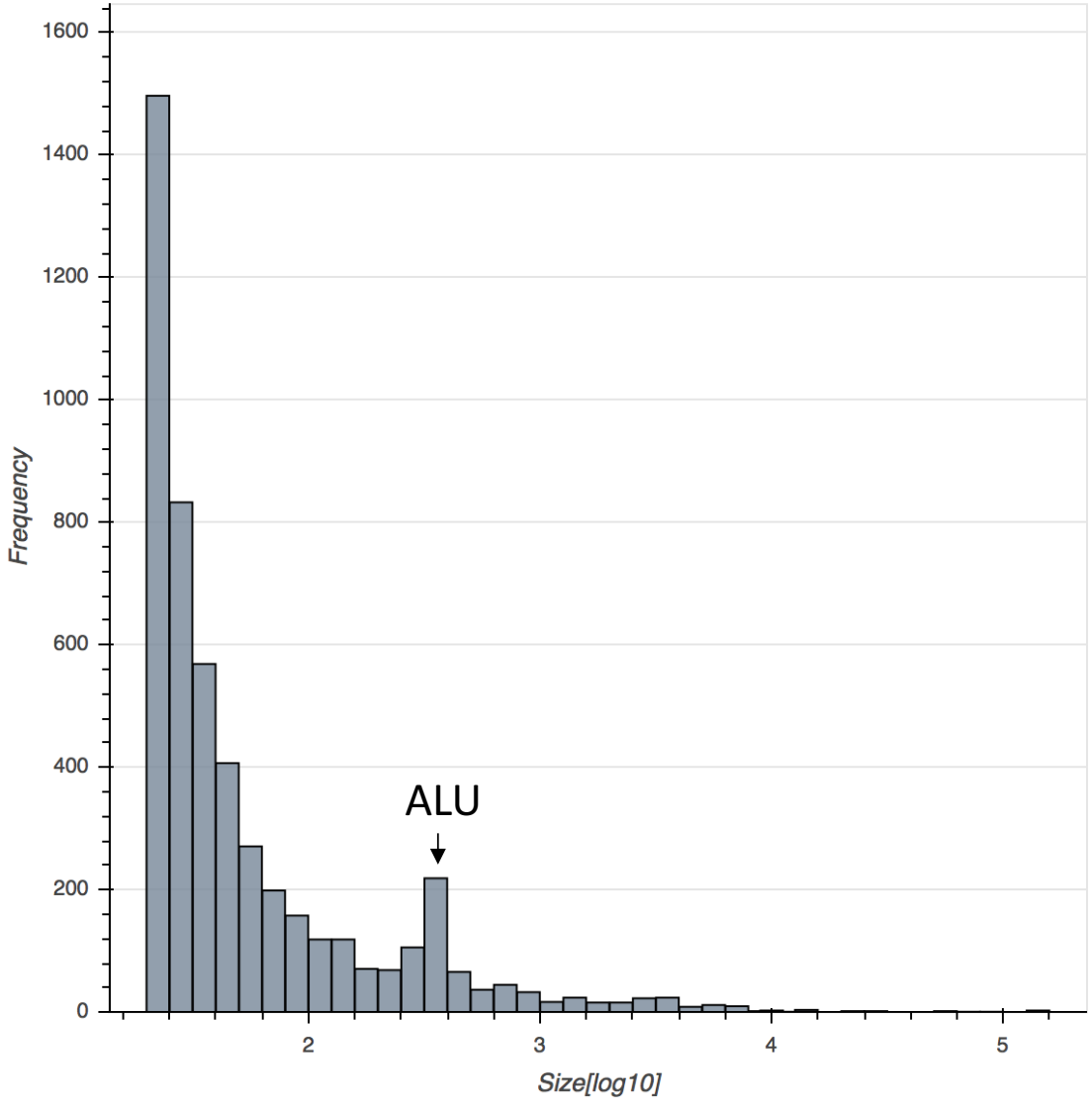
5000 Insertions
5000 Deletions

Overview

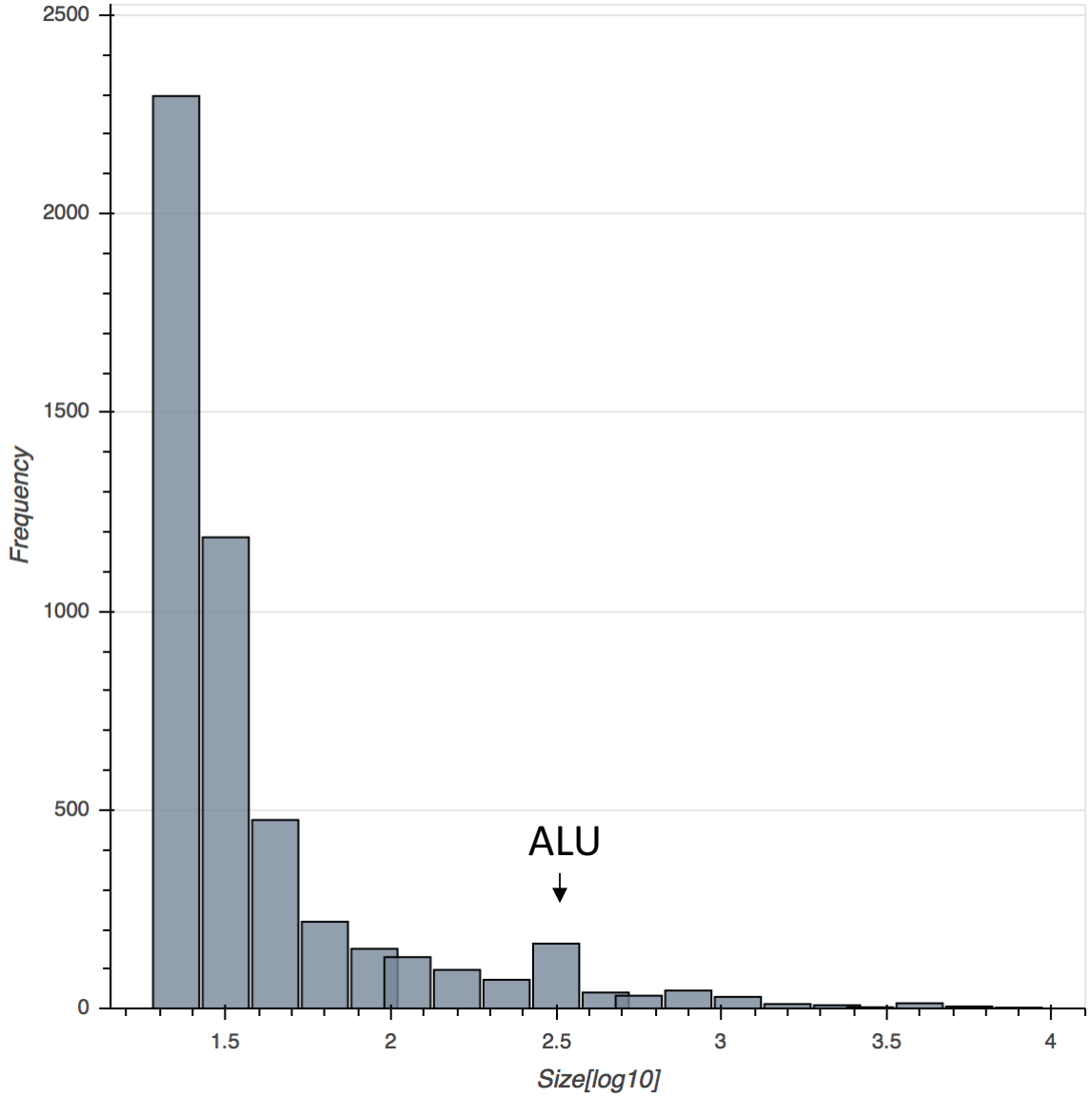

svviz
5000 Insertions
5000 Deletions
1000 Labeled
Data points

Results

Size

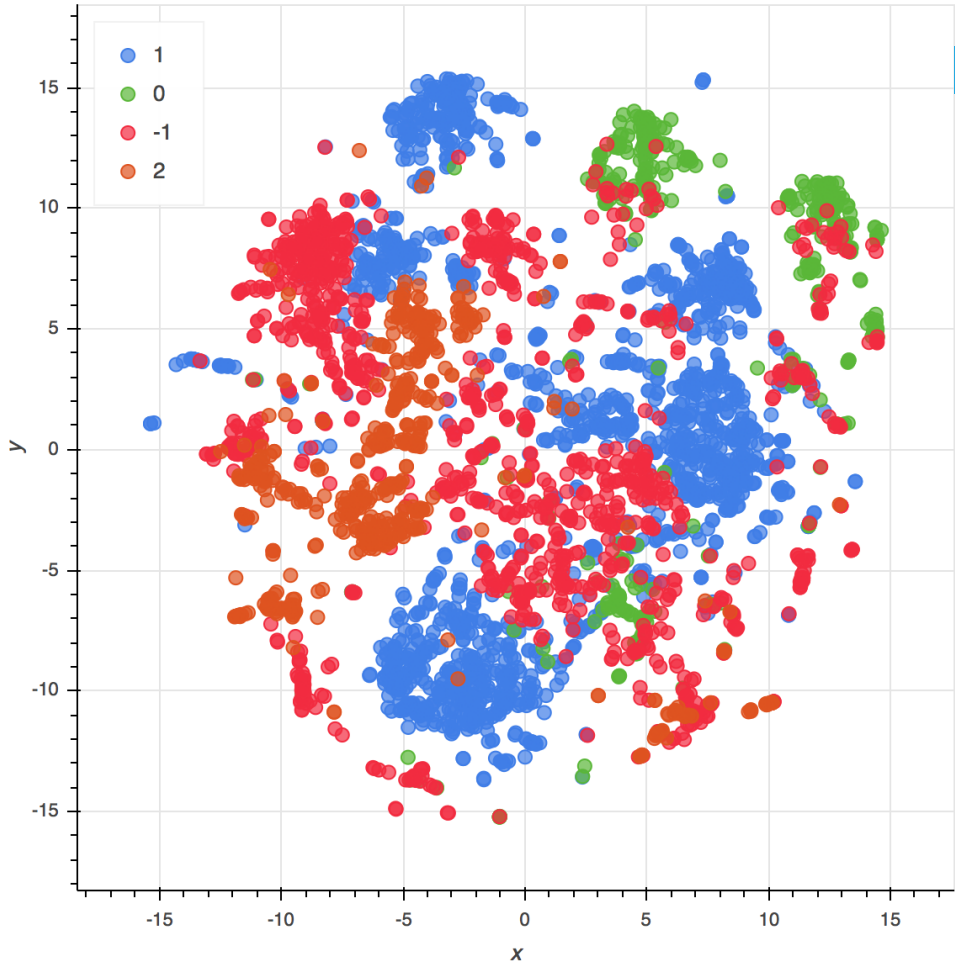


Deletions

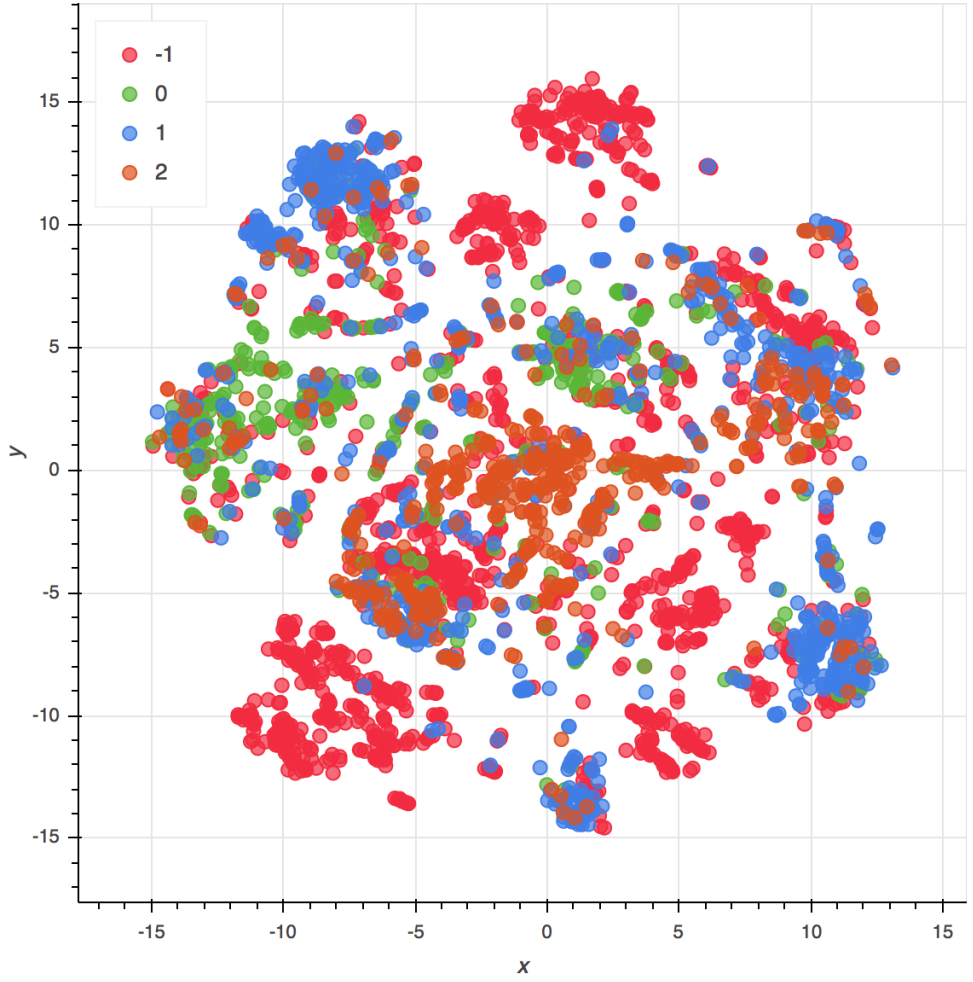


Insertions

Estimated Consensus GT



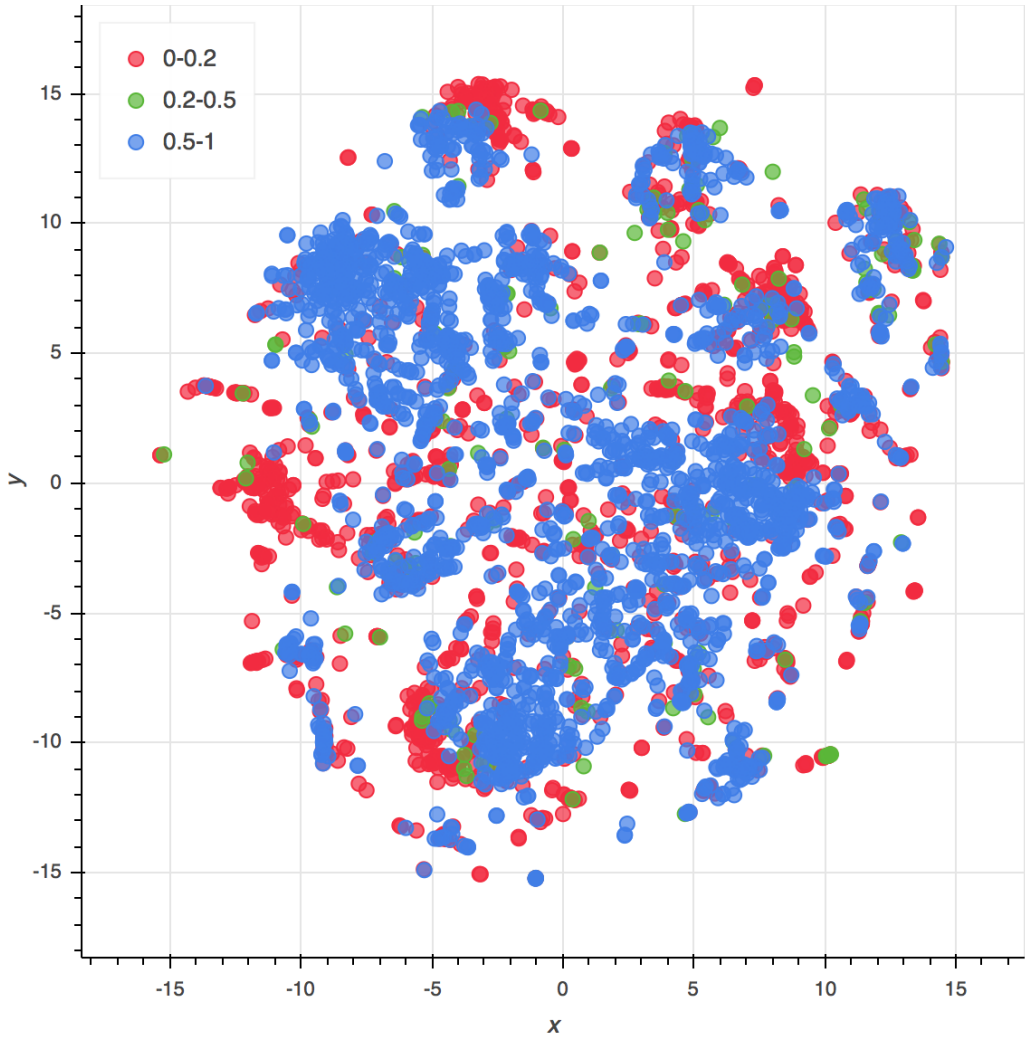
Deletions



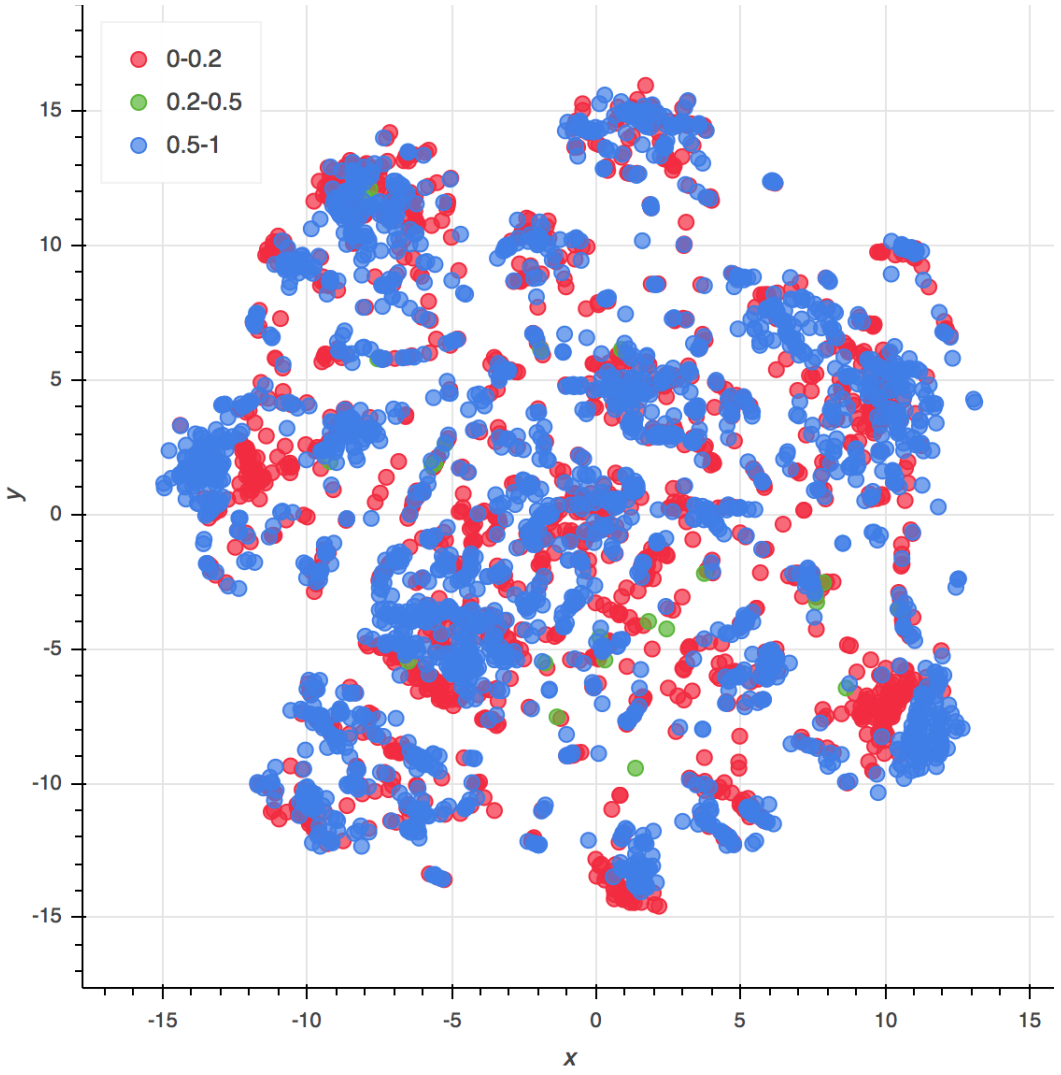
Insertions

Key	
0	Homozygous Reference
1	Heterozygous Variant
2	Homozygous Variant
-1	Undetermined

Tandem Repeats

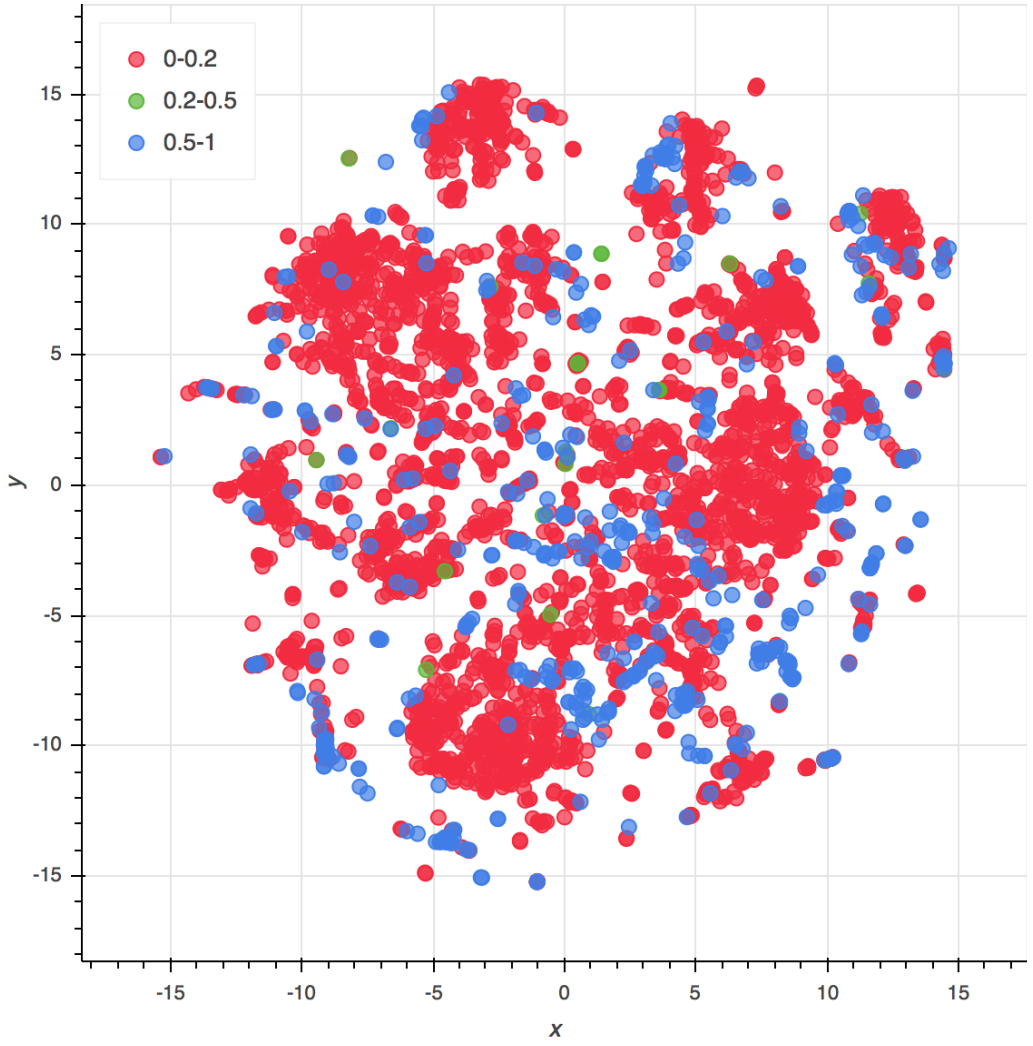


Deletions

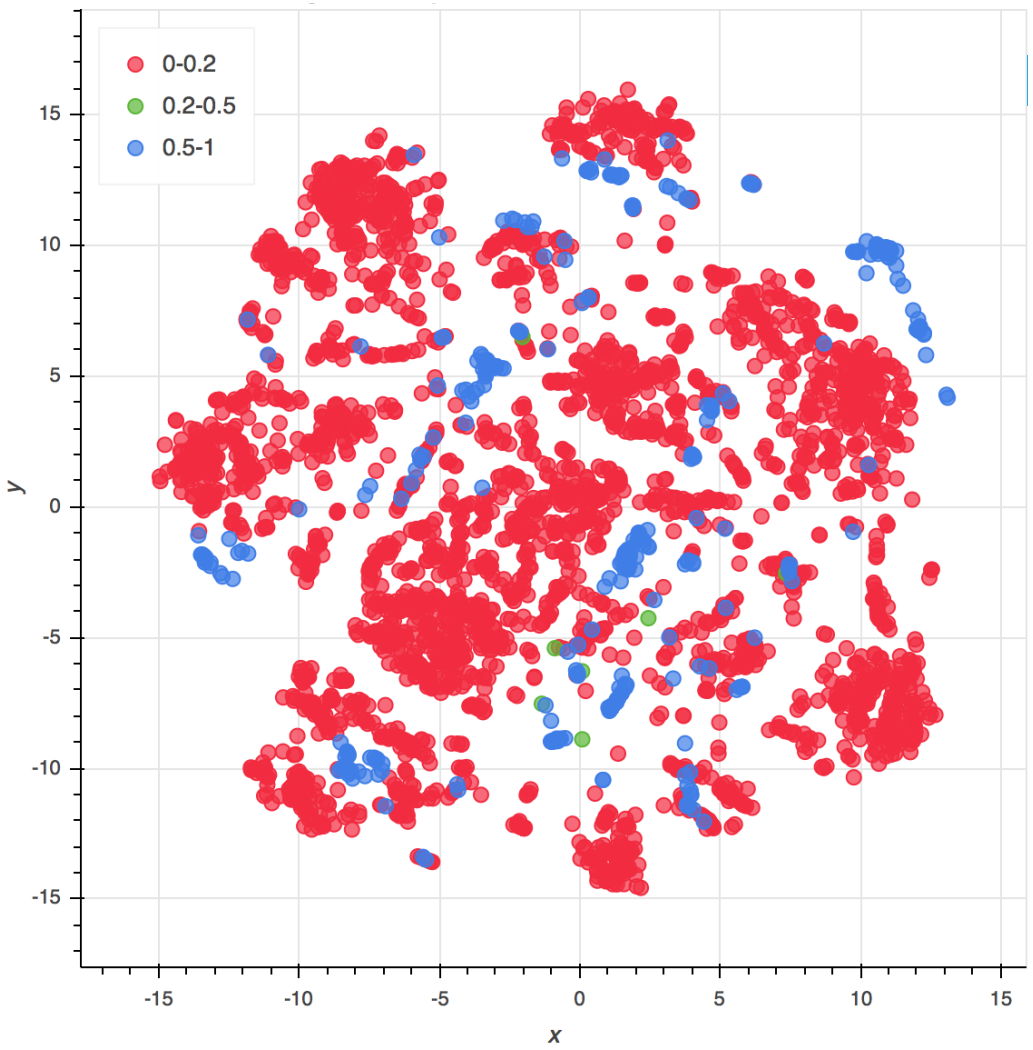


Insertions

Segmental Duplications

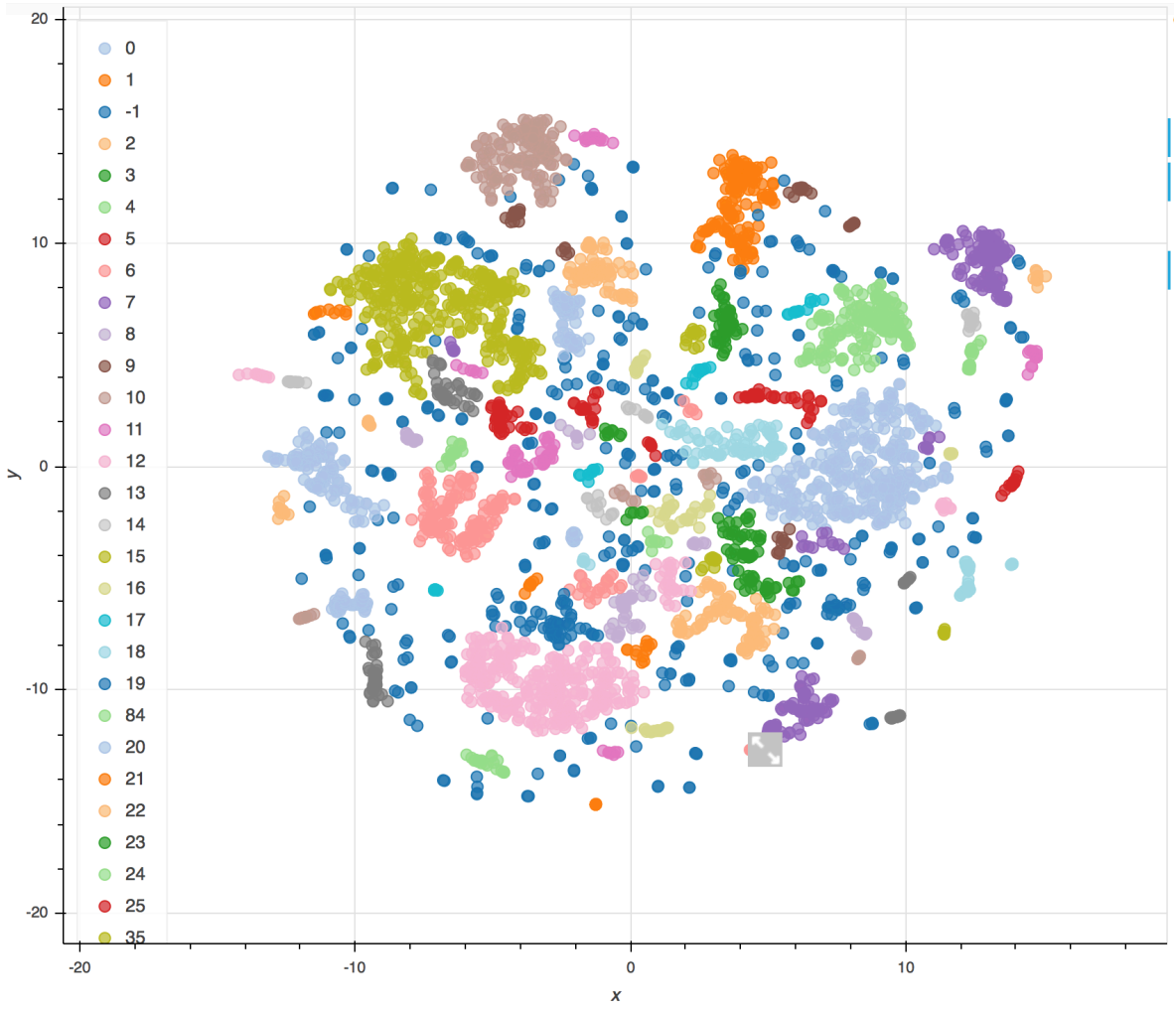


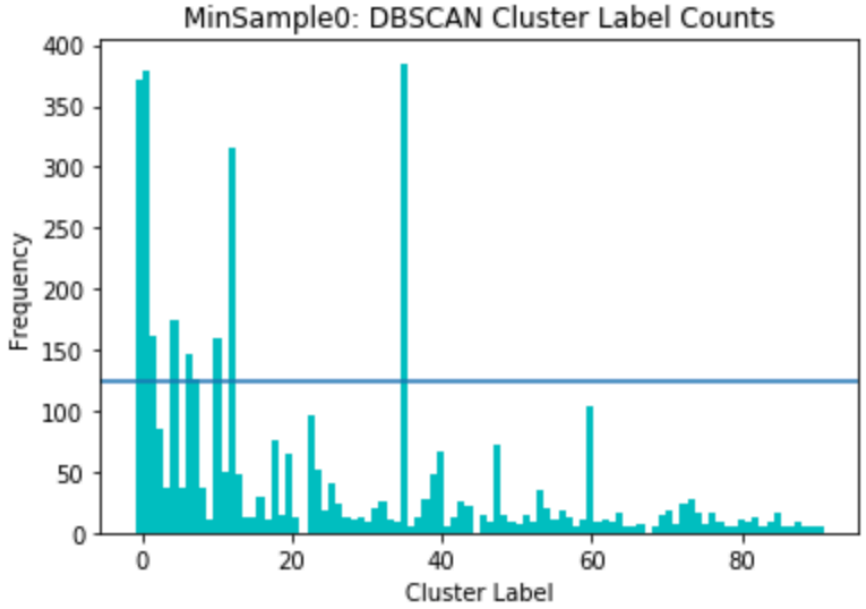
Deletions



Insertions

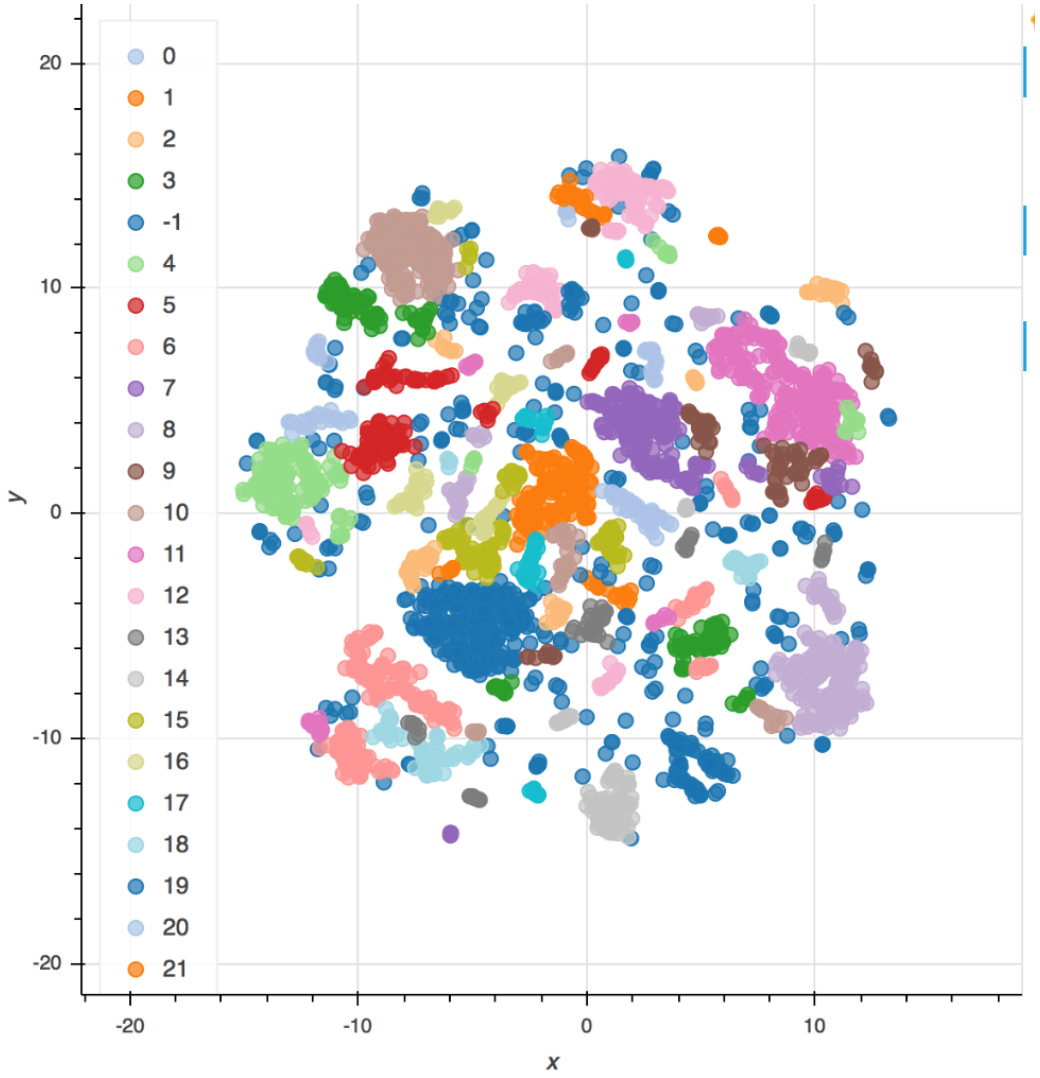
DBSCAN



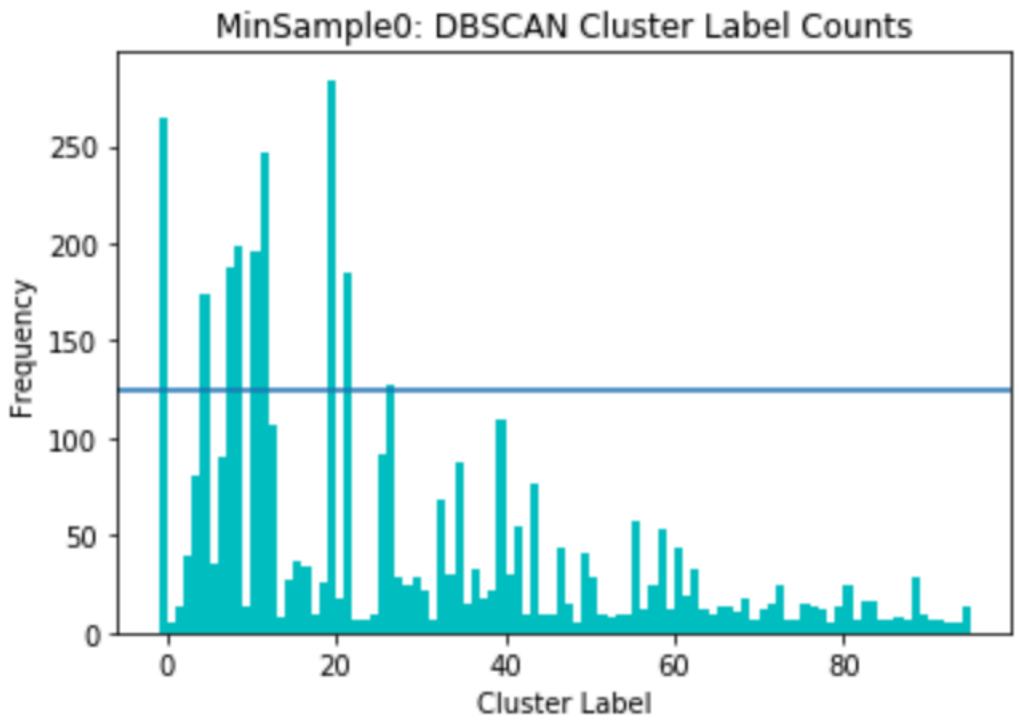


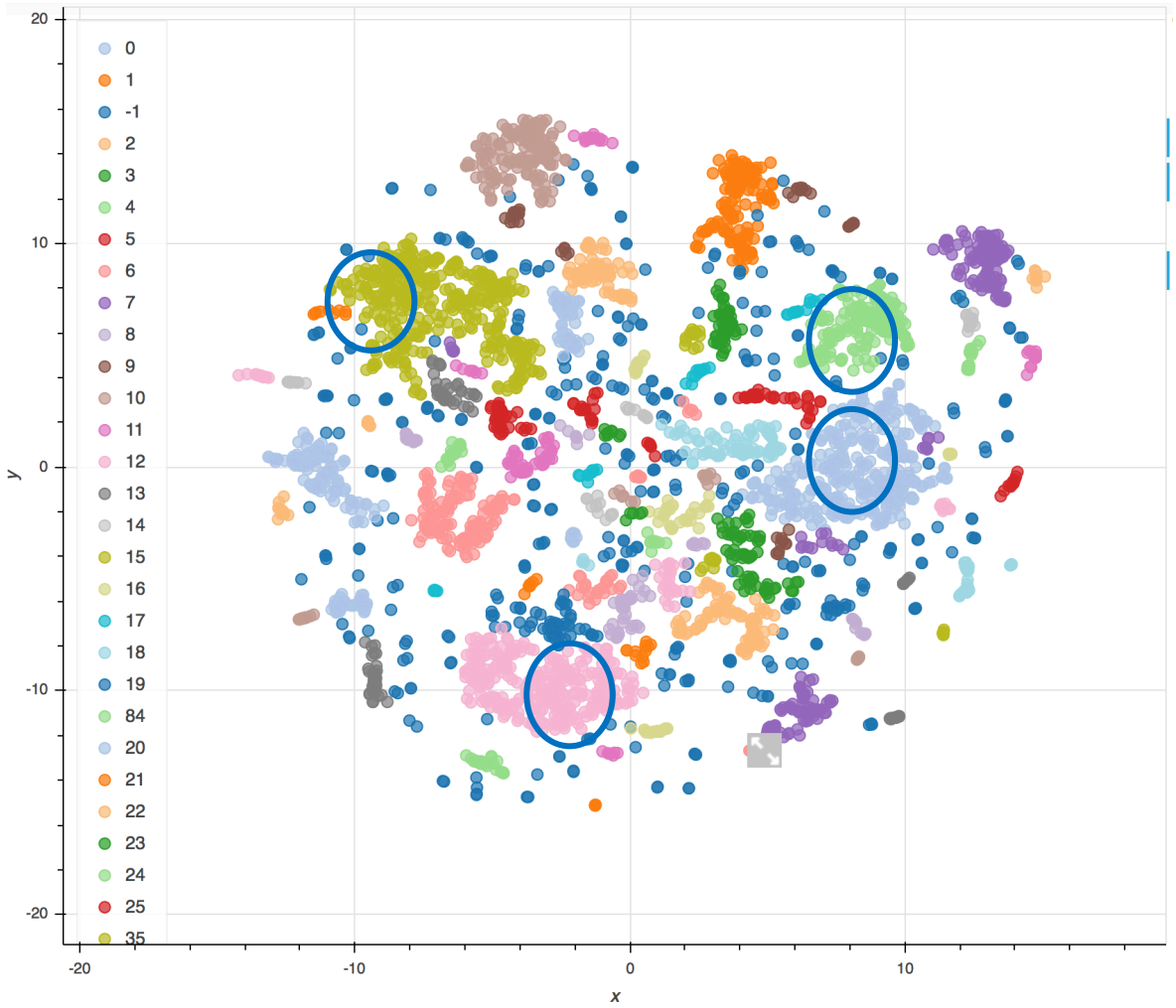
Deletions

DBSCAN

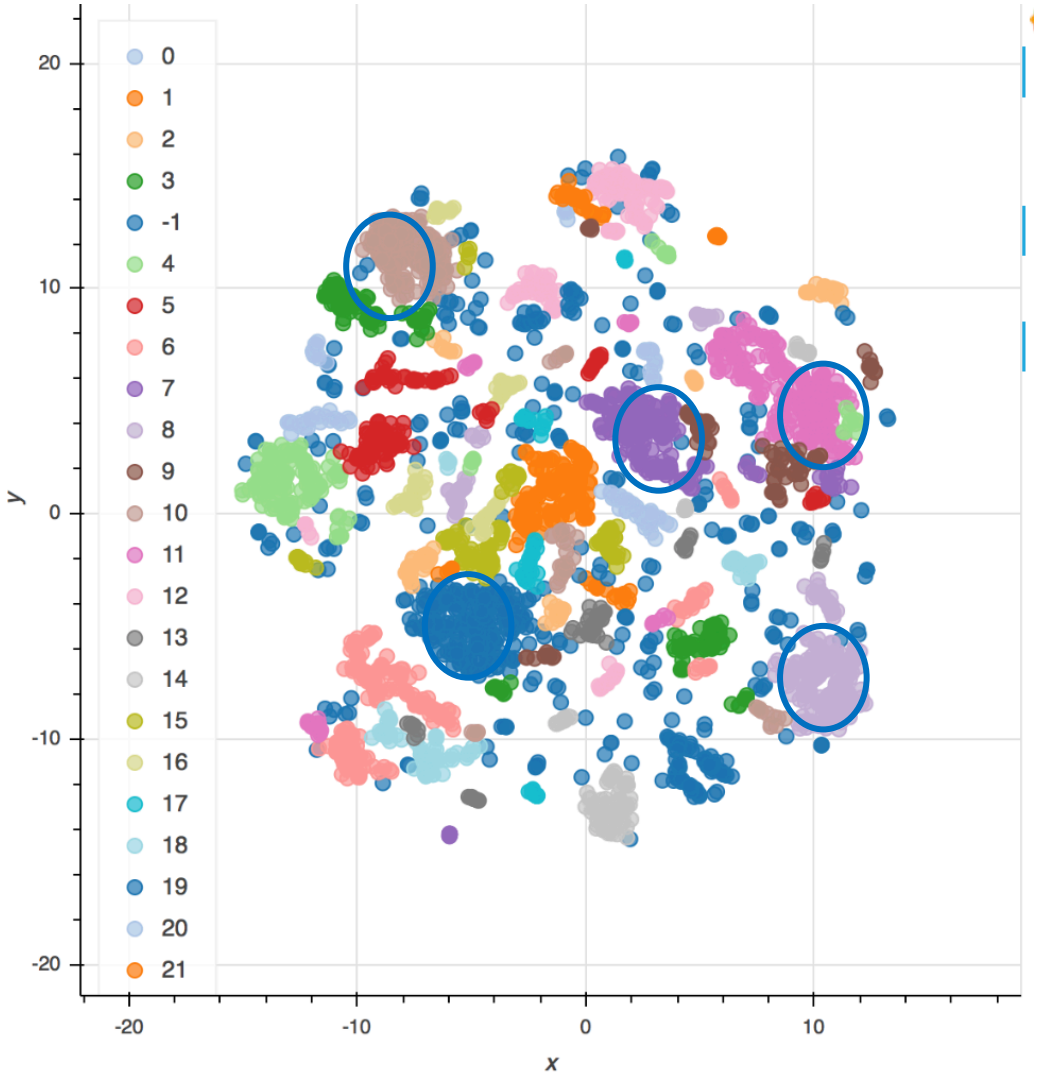


Insertions





Deletions



Insertions

App Development

Preliminary Machine Learning

CrowdSourced Labels

Cold Spring Harbor Laboratory

bioRxiv beta

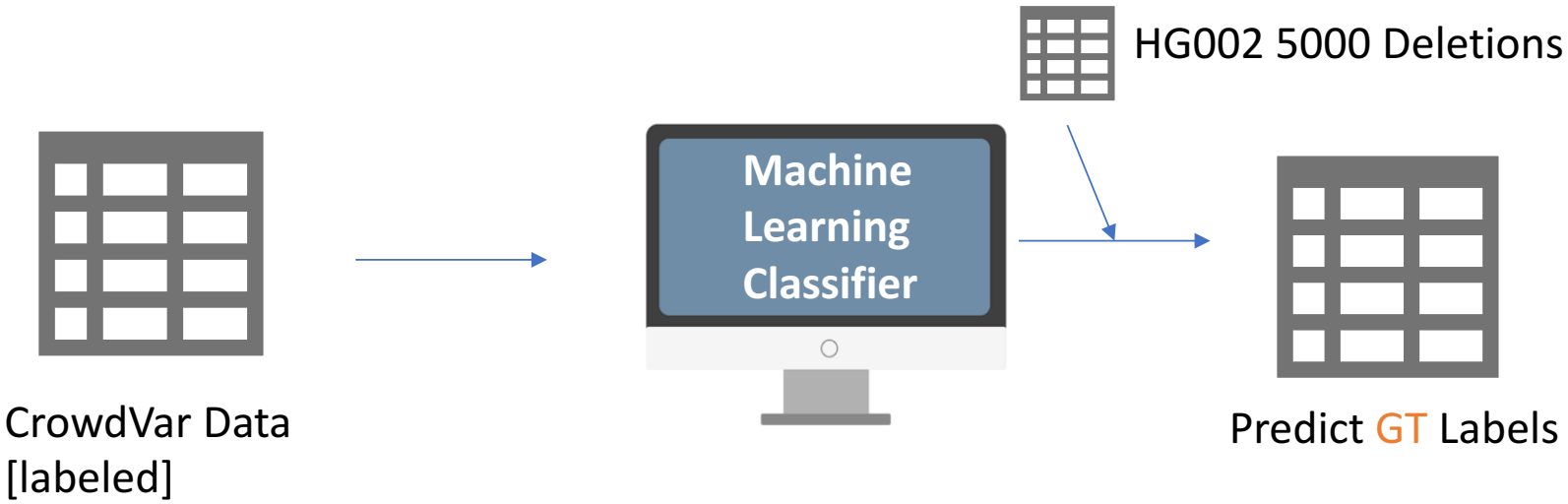
THE PREPRINT SERVER FOR BIOLOGY

- 1700 + DEL HG002
- 8 Hom Ref
- 1097 Het Var
- 647 Hom Var

New Results

CrowdVariant: a crowdsourcing approach to classify copy number variants

Peyton Greenside, Justin Zook, Marc Salit, Ryan Poplin,
Madeleine Cule, Mark DePristo


HG002 5000 Deletions
Machine Learning Classifier
CrowdVar Data [labeled]
Predict GT Labels

Train Random Forest Classifier

- Train-test split [70% Train 30% Test]

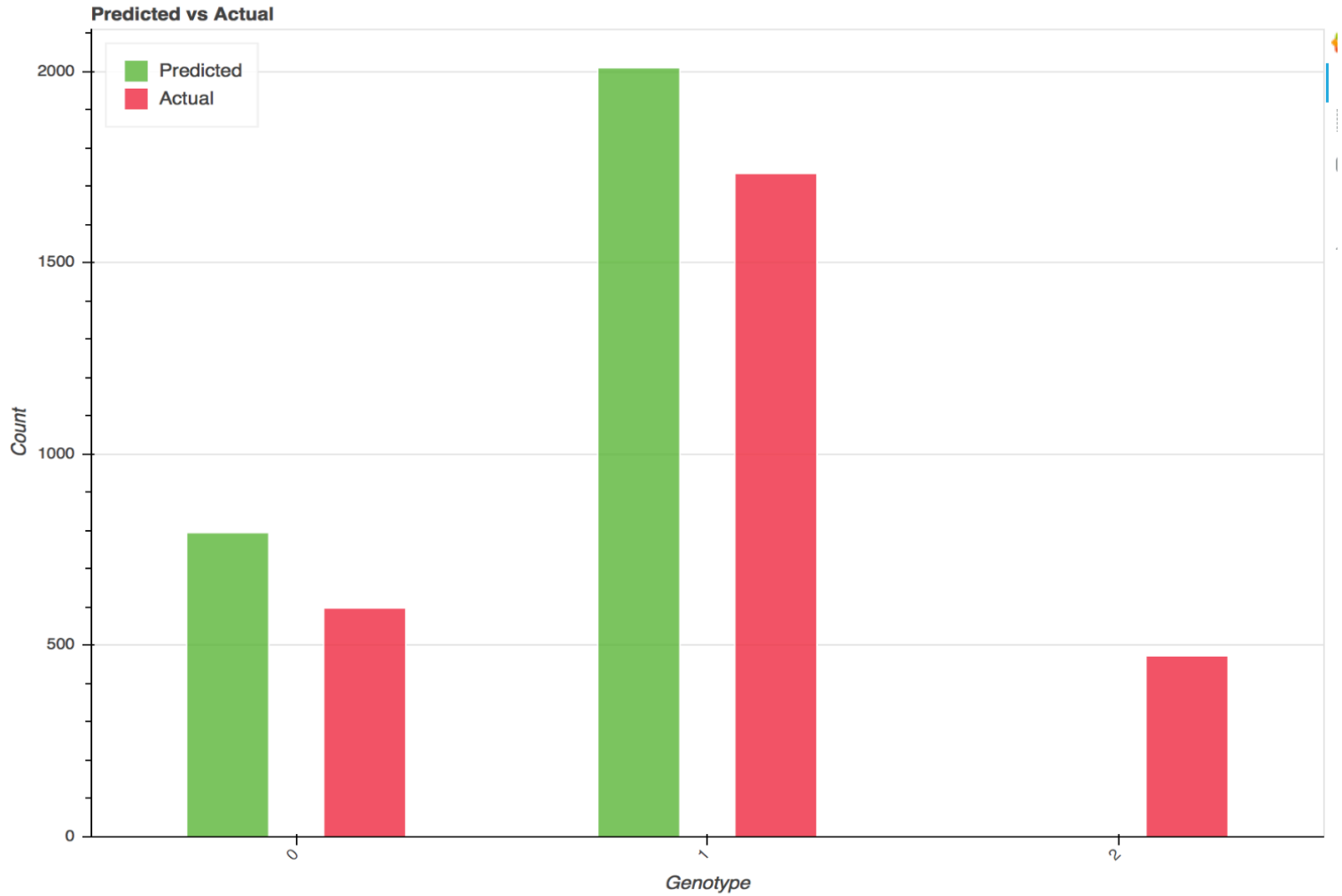
Precision Score
0.981619887

Predict on 5000 Randomly Selected DEL

```
Out[109]: 0.7618538324420677 2
```

Predict on 5000 Randomly Selected DEL

```
Out[109]: 0.76185383244206772
```



- Trained RF classifier
- Train-Test Split
- (overall prediction accuracy)
- Predict on HG002 given svviz features
- Prediction accuracy – 70%, why? Only trained on a few references
- Show examples of good and bad matches