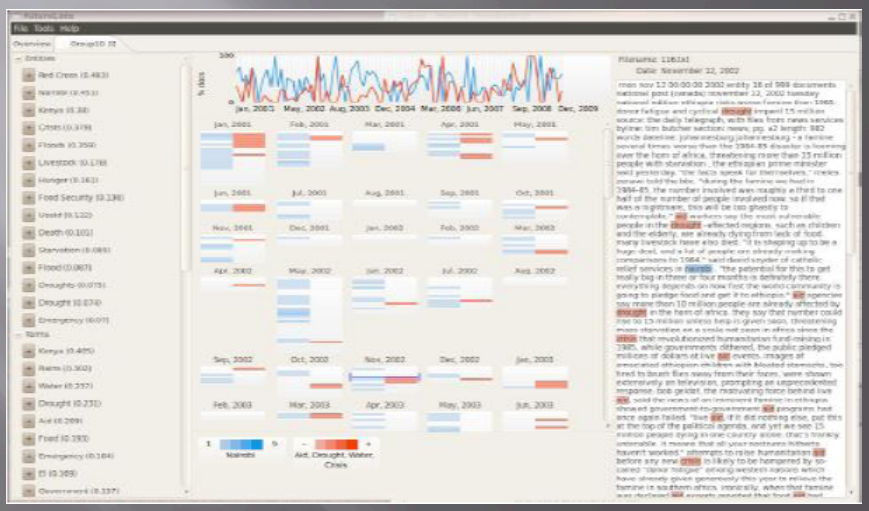
POLYLENS: SOFTWARE FOR MAP-BASED VISUALIZATION OF GENOME-SCALE POLYMORPHISM DATA

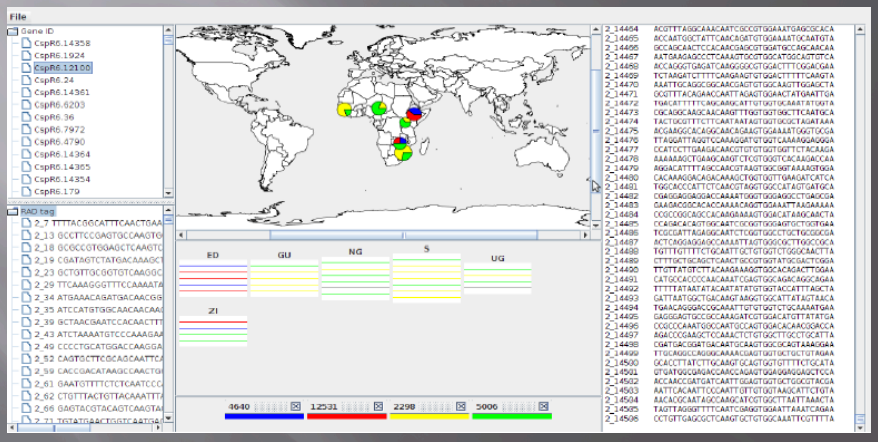
Overview

- Visualization tool for genome-scale population data
- Written in Java
- Swing GUI
- Uses JRI to generate maps using the rworldmap library in R

FutureLens



PhyloLens



The Data

- 4 Parts: Sample Genomes, Locations, Gene IDs, Stop List
- Test-Case Data: 31 samples of individuals of *Drosophila melanogaster* (the common fruit fly)

Sample Genomic Sequences

- Full DNA sequences
- Locus
  - Unique organism ID
  - Location within sequence
- RADTag
  - Subsequences of base pairs
  - Restriction site Associated DNA marker
  - Immediately flank each instance of a restriction enzyme
  - Long enough to be unique across multiple genomes (38 in *Drosophila* data set)

```
12_0  ATCACCGAAATACCAAAATATGTGGTGTAACAGTCAAA
12_1  GTAACTTCAAAGCCAAGGCATGTGGTACAATACCATAA
12_2  GTTCGGTGATTGCCAAGCAATGTGGTTTGCAGATGATA
12_3  ATTAGGCGTTAGACAAAGACGGTGGACAAAGCCATATA
12_4  TTACAAGCCTAGACAAGCCCAGTGGCCCCTCCGGCAGA
12_5  TGGCTCTCTTGTGCAATTACTGTGGCGCACTCCGTAGA
12_6  TTGTATTAACCAACAATGTCGGTGGCAGCAGGGCAACA
12_7  CAATGTCGGGAAGCAATCAATGTGGCTTCATGTTTCGA
12_8  TTTATCTTCATTACAATTTTTGTGGTAGGAAACTGATA
12_9  AGAGATATAAAACCAATCGGAGTGGGGCCTCATTAAGA
12_10 GTAGCGCATTCATCAATGTCCGTGGCGCTTCCGTTGGA
12_11 AGGGCGGTGGACACAATGAATGTGGACACCATCAGGGA
12_12 GATGAACGCGTGTCAAGGGGTGTGGGGTGTGGTGGGTA
12_13 TTCGAATTAATTGCAACTCCAGTGGAGAAATGTGTCTA
12_14 AAACGGATAAATACAACTTATGTGGGAATGGCAAAATA
12_15 TATAAGGATGCTCCAAAAGGGGTGGGGCGGGGCGTGGA
12_16 TACTGTACTAGCGCAAGTACAGTGGAGCCAACAATAGA
12_17 TTATGTGCTAACACAAAGTGTGTGGATGCCCGAATGGA
12_18 ACAAATTTCAGGGCAAAAAATGTGGCTATGGGAAAGCA
12_19 CCCGGTACATTTCCAACAGAGGTGGGGTTGCTGCCAAA
12_20 AGTTCGTGAATCTCAAGGGCGGTGGACTCCTGGTGGAA
12_21 AAGACGAGCATCACAACGGGCGTGGTGGATCGCTCGGA
12_22 TTTATTTGTATTGCAAAGGAAGTGGCTCCTTAACTGCA
```

Location

- Latitude/Longitude at which sample was collected
- In *Drosophila* data sets, samples were collected from Ethiopia, Guinea, Nigeria, South Africa, Uganda, Zambia, and France

```
  6.98   39.18    1
  6.98   39.18    2
  6.98   39.18    3
  6.98   39.18    4
  6.98   39.18    5
 10.70  -12.25    6
 10.70  -12.25    7
 10.70  -12.25    8
 10.70  -12.25    9
 10.70  -12.25   10
 11.85   13.16   11
 11.85   13.16   12
 11.85   13.16   13
 11.85   13.16   14
 11.85   13.16   15
 11.85   13.16   16
```

Gene IDs

- Sets of RADTags that differ by no more than 4 nucleotides
- Different expressions of a gene

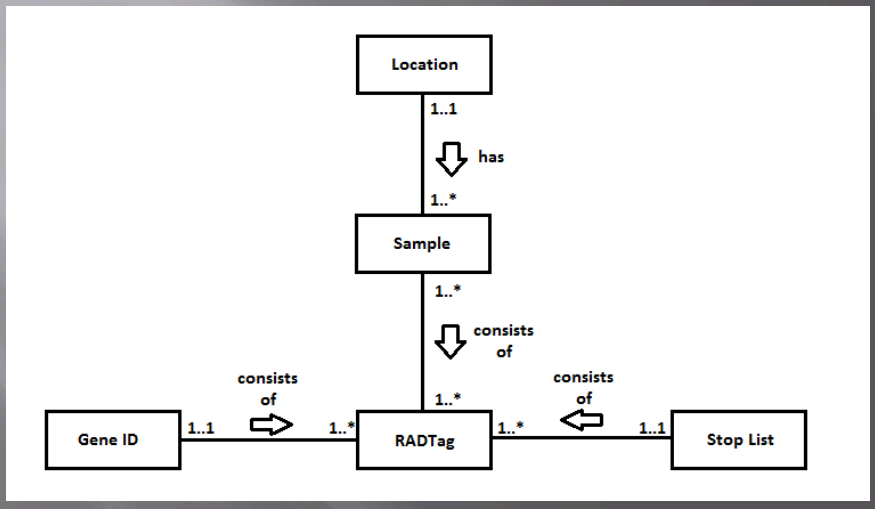
```
CspR6.20480 21_7004,22_6804,31_6896
CspR6.20481 21_7016,27_6645
CspR6.20485 21_7269,30_7109,31_7125
CspR6.20486 21_7313,22_7091,31_7181
CspR6.20487 21_7434,28_7221
CspR6.20489 21_7487,22_7262,23_7567,30_7334
CspR6.20490 21_7694,23_7763
CspR6.20491 21_8110,30_7950
CspR6.20492 21_8528,27_8013,29_8238
CspR6.20493 21_8993,31_8752
CspR6.20495 21_9395,29_9068,30_9152
CspR6.20500 21_10174,29_9821
CspR6.20505 21_10724,23_10786,28_10372
CspR6.20509 21_12244,30_11892
CspR6.20512 21_12595,29_12179
CspR6.20513 21_12733,31_12389
CspR6.20514 21_12913,26_11509,28_12471
CspR6.20515 21_12991,23_13074,31_12643
CspR6.20518 21_13114,28_12653
CspR6.20519 21_13124,23_13204
CspR6.20521 21_13443,28_12962,30_13048
CspR6.20526 21_14282,30_13828
```

Stop List

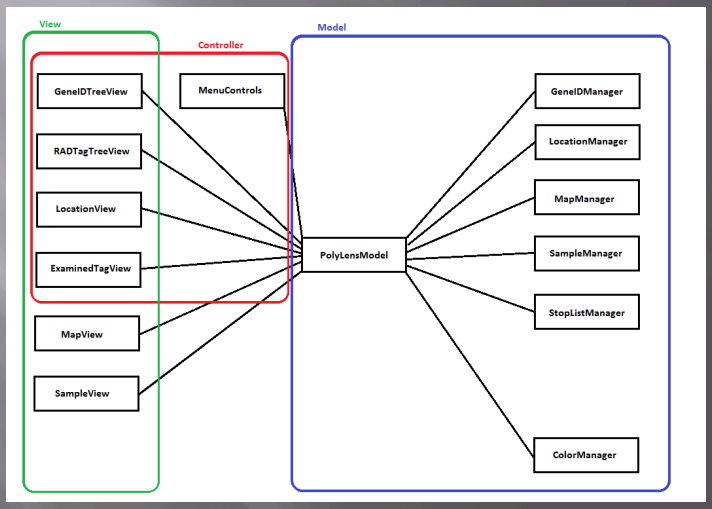
- Borrowed from text mining
- a, an, the
- High-frequency RADTags
- Incredibly rare RADTags


```
**AAAAATTCCCTTTCAAGTGTTGTGGCAAGGACATCCTA
1_8114, 2_8136, 3_8208, 4_8278, 5_8109, 6_8220, 7_7683, 8_8301, 9_8327, 10_8043, 11_8346, 12_8220, 13_8032, 14_8108, 15_8176, 16_8046, 17_8778, 18_8794, 19_8820, 20_8769, 21_8765, 22_8446, 23_8820, 24_8058, 25_8061, 26_7801, 27_8226, 28_8492, 29_8454, 30_8554, 31_8535
**AAAACATATGCAACAAGAGGTGTGGCCAAGACTCTCAA
1_353, 2_349, 3_349, 4_356, 5_341, 6_344, 7_329, 8_350, 9_356, 10_358, 11_368, 12_344, 13_345, 14_317, 15_345, 16_331, 17_374, 18_379, 19_381, 20_367, 21_369, 22_351, 23_359, 24_342, 25_340, 26_353, 27_346, 28_360, 29_346, 30_361, 31_368
**AAAACGCTGTGACCAATTCAAGTGGAGTTTTCCGGCGA
1_7453, 2_7435, 3_7500, 4_7581, 5_7436, 6_7549, 7_7064, 8_7618, 9_7626, 10_7379, 11_7656, 12_7549, 13_7356, 14_7448, 15_7486, 16_7386, 17_8041, 18_8061, 19_8083, 20_8024, 21_7990, 22_7739, 23_8077, 24_7384, 25_7376, 26_7129, 27_7528, 28_7771, 29_7742, 30_7829, 31_7831
**AAAAGCAATATTACAATAACTGTGGTAGAATTGGCTCA
1_7377, 2_7354, 3_7421, 4_7498, 5_7359, 6_7474, 7_6993, 8_7545, 9_7542, 10_7302, 11_7576, 12_7474, 13_7286, 14_7366, 15_7407, 16_7306, 17_7954, 18_7983, 19_7995, 20_7942, 21_7912, 22_7664, 23_7982, 24_7302, 25_7289, 26_7059, 27_7450, 28_7679, 29_7664, 30_7738, 31_7749
**AAAAGGCATTTGGCAAATGAGGTGGAAATGCTCTCAAA
1_10505, 2_10607, 3_10727, 4_10773, 5_10609, 6_10708, 7_10048, 8_10872, 9_10828, 10_10522, 11_10840, 12_10708, 13_10519, 14_10535, 15_10698, 16_10451, 17_11462, 18_11481, 19_11487, 20_11453, 21_11448, 22_11031, 23_11533, 24_10476, 25_10579, 26_10228, 27_10674, 28_11073, 29_11068, 30_11136, 31_11152
```


Data Relations



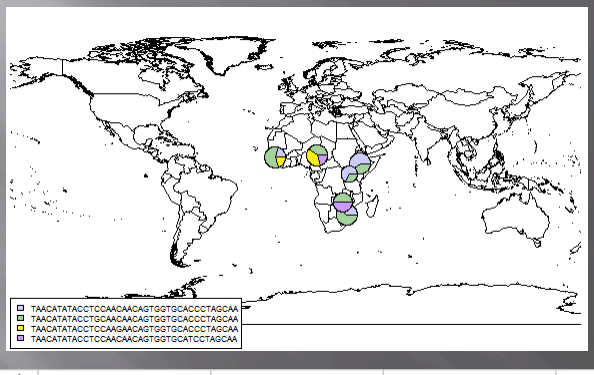
Implementation



Data Managers

- GeneIDManager, LocationManager, SampleManager, StopListManager
  - maintain data mappings, file associations, update status
- ColorManager
  - color coordination between components
- MapManager
  - Interfaces with R using JRI
  - Frequency table to pie charts on world map

MapManager



Latitude	Longitude	TAACATATACCTCCAACA	TAACATATACCTGCAACA	TAACATATACCTCCAAGA	TAACATATACCTCCAACA
6.98	39.18	3	2	0	0
10.7	-12.25	1	3	1	0
11.85	13.16	0	2	2	1
-23.94	31.14	2	3	0	0
0.53	32.6	2	1	0	0
-16.54	28.72	0	2	0	2

Uses

- Allele distribution patterns
- Phylogeography
- In *Drosophila*:
  - recent admixture event with French flies
  - some African flies more French than African
- Dataset 1: GeneID 359
- Dataset 2: GeneID 16520

Demo

Future Work

- Produce same output with clustering algorithm output
- RADTag split/merge
- Pan/Zoom map
- Better Pie Charts
- Possibly abandon R altogether
- Scalability