Exploring Unidentified Peptide Sequence Data from a Circadian Rhythm Study in *Kalanchoe fedtschenkoi*

*Armin G Geiger[1,2]; *Paul E Abraham[1]; Xiaohan Yang, Rongbin Hu, Richard J Gianonne[1] Daniel A Jacobson[1];
[1]Oak Ridge National Lab, Oak Ridge, TN; [2]Bredesen Center for Interdisciplinary Research and Graduate Education, Knoxville, TN;

Abstract

Liquid chromatography coupled with to two rounds of mass spectrometry (LC-MS/MS) applied in a technique known as `shotgun proteomics', has proven effective as a means to capture a considerable portion of the protein complement of a biological sample. The technique often produces millions of mass spectra per experiment, however, approximately 50-75% of MS2 spectra remain unidentified, even as a good portion of these spectra are of high quality and likely peptide-derived. There are many possible reasons why these spectra may go unassigned including not having good enough database matches for spectra arising from biological phenomena such as unknown post-translational modifications and single nucleotide polymorphisms(SNPs). In addition, problems such as spectral chimerism - a phenomena where the isolation window for a peptide contains more than one distinct peak - is also known to negatively impact spectral library searching. Clustering these high-quality unassigned (HQU) spectra together with their assigned counterparts, combined with a spectral purity analysis, may yield insights into the origins of these HQU spectra. This work combines clustering of mass spectra (using 3 distinct algorithms) with a spectral purity analysis. Moreover, the experimental design is leveraged by using the peptide intensities to identify unidentified spectra of possible biological relevance. The method is applied to an LC-MS/MS dataset obtained from a circadian rhythm experiment in the plant species, *K. fedtschenkoi*. This plant is an important model species for the study of Crassulacean Acid Metabolism - a special adaptation of plants that inhabit areas with low water availability. Mining of this untapped proteome resource may yield valuable insights into the proteomic changes that occur during the circadian rhythm of this plant.

Introduction

Clustering of mass spectra in proteomics refers to the grouping of mass spectra that have similar fragmentation patterns together into clusters. How these clusters are formed is an important aspect of the strategy to investigate unassigned spectra. Three clustering algorithms used in the proteomics field are discussed here, all three having the following basic steps: Firstly, a similarity metric is used to compare all spectra to one another in a pairwise fashion. Secondly, these calculated pairwise distance relationships are used to cluster the spectra.

Table 1. Summay of differences between the three clustering algorithms used.

	MS-clustering	MaRaCluster	Spectra-cluster
Scoring scheme	normalized dot product	p-value based	probabilistic
clustering type	bottom-up, greedy, incremental hierarchical	bottom-up hierarchical	bottom-up, greedy, incremental hierarchical

MScluster[1] uses a normalized dot product. MaRaCluster makes use of a distance calculation relying on the rarity of experimental fragment peaks following the intuition that peaks shared by only a few spectra offer more evidence than peaks shared by a large number of spectra. Spectra-cluster uses a hypergeometric distribution to model the probability that the number of matched peaks occurred at random. The probability that the rank distribution of matched peaks occurred by chance is assessed using Kendall's Tau correlation.

Methods and Materials

LC and MS setup:

All samples were analyzed on a Q Exactive Plus mass spectrometer (Thermo Fischer Scientific) coupled with a with a Proxeon EASY-nLC 1200 liquid chromatography (LC) pump (Thermo Fisher Scientific). Peptides were separated on a 75 μm inner diameter microcapillary column packed with 25 cm of Kinetex C18 resin (1.7 μm, 100 Å, Phenomenex). For each sample, a 2 μg aliquot was loaded in buffer A (0.1% formic acid, 2% acetonitrile) and eluted with a linear 150 min gradient of 2 – 20% of buffer B (0.1% formic acid, 80% acetonitrile), followed by an increase in buffer B to 30% for 10 min, another increase to 50% buffer for 10 min and concluding with a 10 min wash at 98% buffer A. The flow rate was kept at 200 nl/min. MS data was acquired with the Thermo Xcalibur software version 4.27.19, a topN method where N could be up to 15. Target values for the full scan MS spectra were 1 x 10^6 charges in the 300 – 1,500 m/z range with a maximum injection time of 25 ms. Transient times corresponding to a resolution of 70,000 at m/z 200 were chosen. A 1.6 m/z isolation window and fragmentation of precursor ions was performed by higher-energy C-trap dissociation (HCD) with a normalized collision energy of 30 eV. MS/MS sans were performed at a resolution of 17,500 at m/z 200 with an ion target value of 1 x 10^6 and a maximum injection time of 50 ms. Dynamic exclusion was set to 45 s to avoid repeated sequencing of peptides.

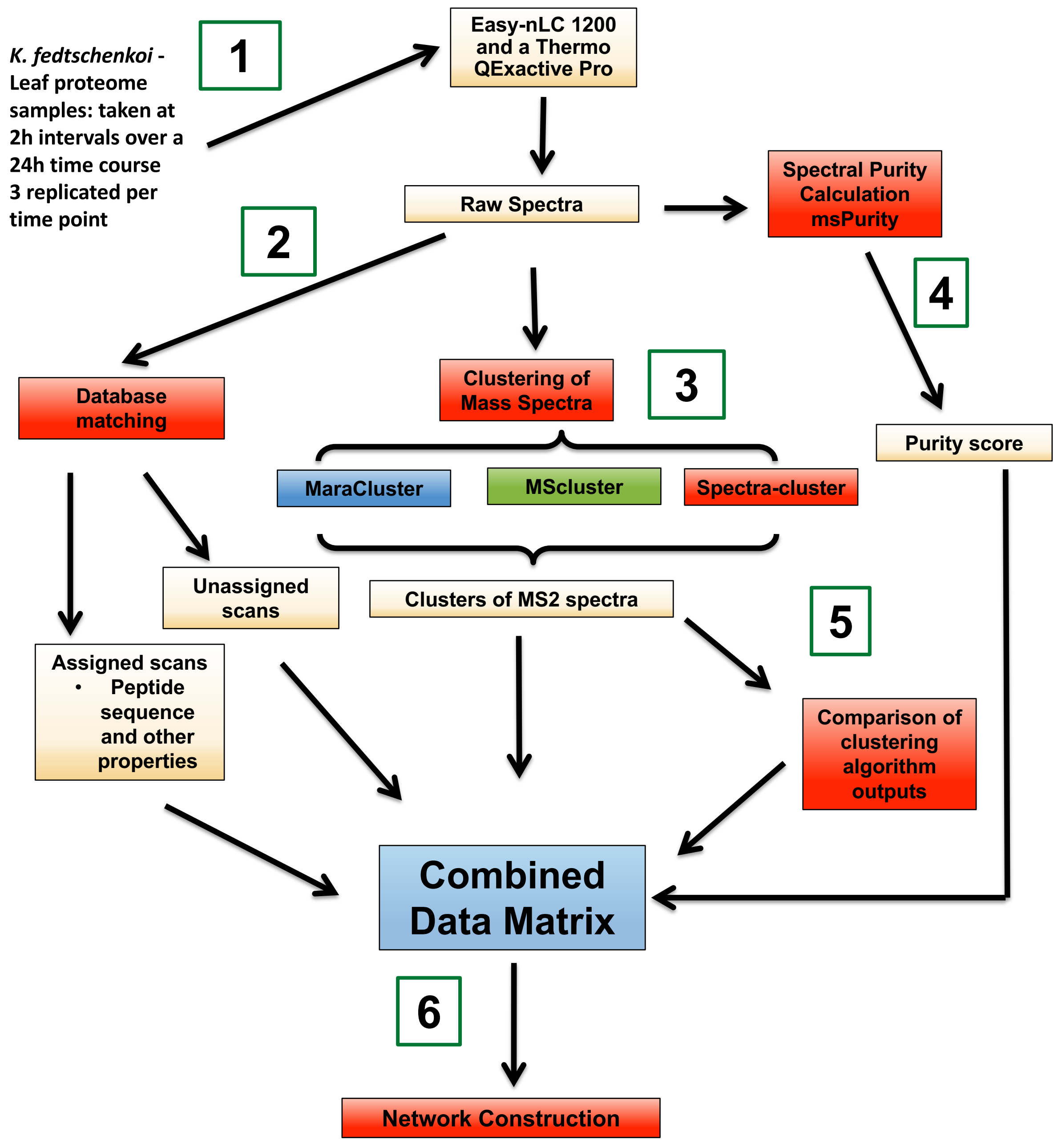
K. fedtschenkoi - Leaf proteome samples: taken at 2h intervals over a 24h time course 3 replicated per time point
1
Easy-nLC 1200 and a Thermo QExactive Pro
Raw Spectra
Spectral Purity Calculation msPurity
2
Database matching
Clustering of Mass Spectra
3
4
Purity score
MaraCluster
MScluster
Spectra-cluster
Unassigned scans
Clusters of MS2 spectra
5
Assigned scans
• Peptide sequence and other properties
Comparison of clustering algorithm outputs
Combined Data Matrix
6
Network Construction

Figure 1. Flow chart of the methodology used: 1. Sample generation and LC-MS/MS analysis; 2. Peptide spectral matching; 3. Spectral clustering; 4. Spectral purity calculation; 5. Cluster similarity calculation. 6. Cluster to Scan network construction.

Spectral matching:

MS raw data files were searched against the Kalanchoe fedtschenkoi v1.1 proteome FASTA database appended with the predicted chloroplast and mitochondrial proteins as well as common contaminates. A decoy database, consisting of the reversed sequences of the target database, was appended in order to discern the false-discovery rate (FDR) at the spectral level. For standard database searching, the peptide fragmentation spectra (MS/MS) were analyzed by the Crux pipeline v3.0. The MS/MS spectra were searched using the Tide algorithm and was configured to derive fully-tryptic peptides using default settings except for the following parameters: allowed clip nterm-methionine, a precursor mass tolerance of 10 parts per million (ppm), a static modification on cysteines (iodoacetamide; +57.0214 Da), and dynamic modifications on methionine (oxidation; 15.9949). The results were processed by Percolator to estimate q values. Peptide spectrum matches (PSMs) and peptides were considered identified at a q value <0.01. Across the entire experimental dataset, proteins were required to have at least 2 distinct peptide sequences and 2 minimum spectra per protein.

For label-free quantification, MS1-level precursor intensities were derived from MOFF using the following parameters: 10 ppm mass tolerance, retention time window for extracted ion chromatogram was 3 min, time window to get the apex for MS/MS precursors was 30 s. Protein intensity-based values, which were calculated by summing together quantified peptides, were normalized by dividing by protein length and total ion intensities and then LOESS and median central tendency procedures were performed on log2-transformed values. Using the freely available software Perseus[4] , missing values were replaced by random numbers drawn from a normal distribution (width = 0.3 and downshift = 2.5).

Spectral purity calculation:

Spectral purity describes the contribution of the selected precursor peak in an isolation window used for fragmentation. It involves dividing the intensity of the selected precursor peak by the total intensity of the isolation window. Spectral purity was calculated using the R package msPurity v1.5.4.

Spectral clustering:

MaraCluster version 0.03.1 on Windows was used with a --precursorTolerance 0.005 Da and a p-val clustering threshold of 0.00001. MSCluster v2.00 (Release 20101018) was used with a --fragment-tolerance 0.02 and --window 0.01. Spectra-cluster-cli-1.0.3 was used with the following parameters: -precursor tolerance of 2.0 Da; fragment_tolerance 0.01; x_min_comparisons=0.

Cluster similarity:

The Jaccard index is a set overlap similarity metric and was used here as a measure of scan overlap between clusters. It is calculated by dividing the size of the intersection by the size of the union of two sets. A set in this instance refers to the scans that have been grouped into clusters. The Jaccard index was calculated for every pair of mass spec clusters resulting in a matrix from which networks can be constructed. The Jaccard index varies between 0 and 1, with a value of 1 indicating complete set overlap, whilst zero indicates no set overlap.

Network construction: All networks were visualized using Cytoscape v.3.5.1 [6].

Results

Total number of scans: 2,323,718 produced with 1,845,503 passing noise filtering. 496,045 spectral assignments were made. Thus, 73% of spectra remained unassigned. Of the assigned spectra 36,053 were unique and were used to infer 4,915 different proteins.

Table 2. Summary of clustered scans produced by different algorithms

	Maracluster	%	MS-cluster	%	Spectra-cluster	%
Number of input scans (pass noise filter)	1,845,503	79	1,844,836	79	1,845,503	79
Total number of clusters produced	369,049	na	217,384	na	139,442	na
Member count of the largest cluster	896	na	2,029	na	6,041	na
Number of clusters of size==1	173,700	47	119,057	55	81,780	58
Number of clusters of size==2	48,608	13	17,283	8	6,872	5
Number of clusters of size==3	29,003	8	9,838	5	3,725	3
Number of clusters of size 3< size <100	117,505	32	69,237	32	44,607	32
Number of clusters of size >100	233	0	1,969	1	2,458	2

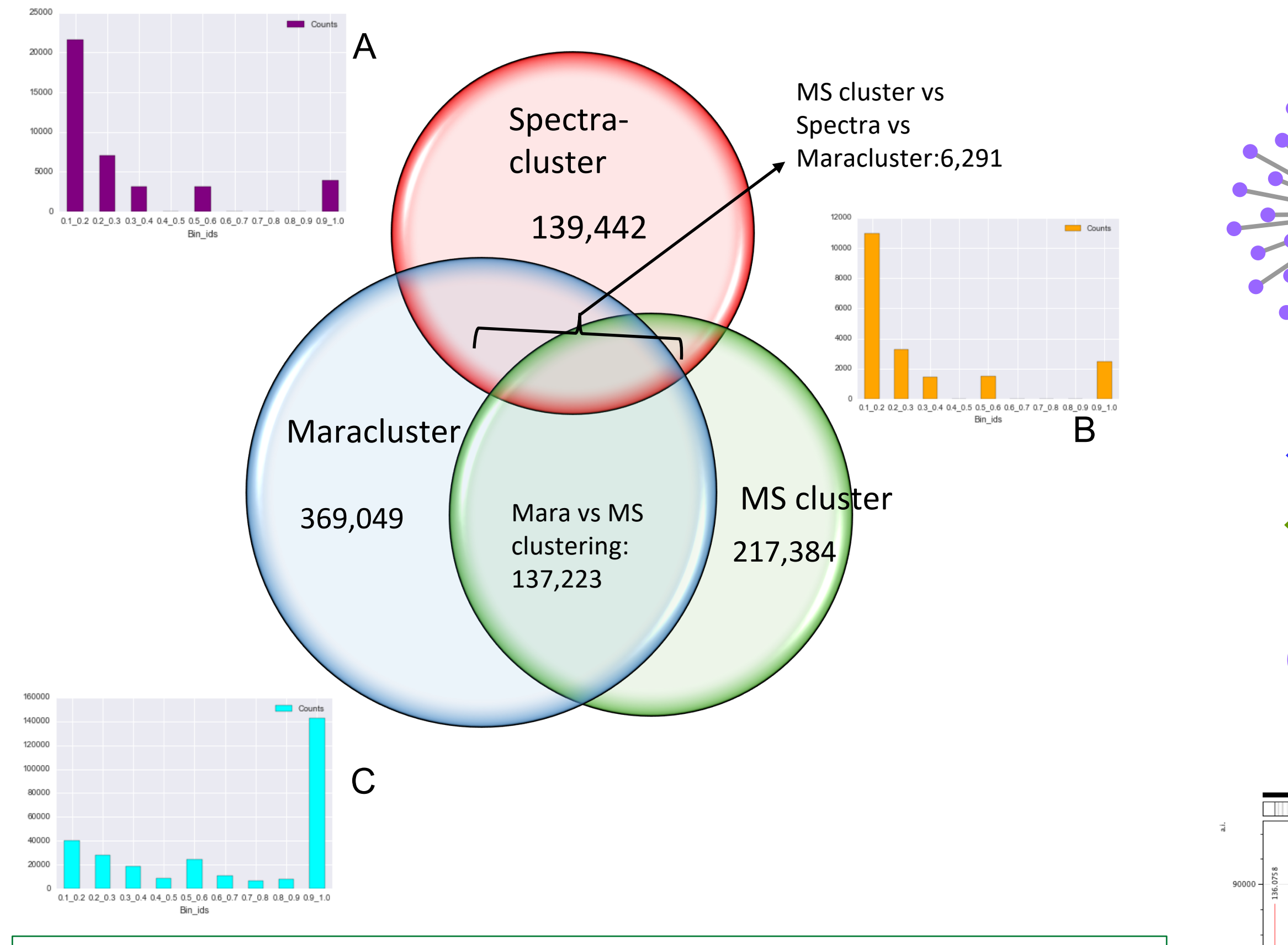

Figure 2. Similarity of the outputs produced by the three algorithms. The Venn diagram shows cluster similarity at a Jaccard coefficient of 1, indicating the number of clusters that have complete overlap. Inserts A,B and C are the Jaccard score frequency distributions for the respective algorithm comparisons.

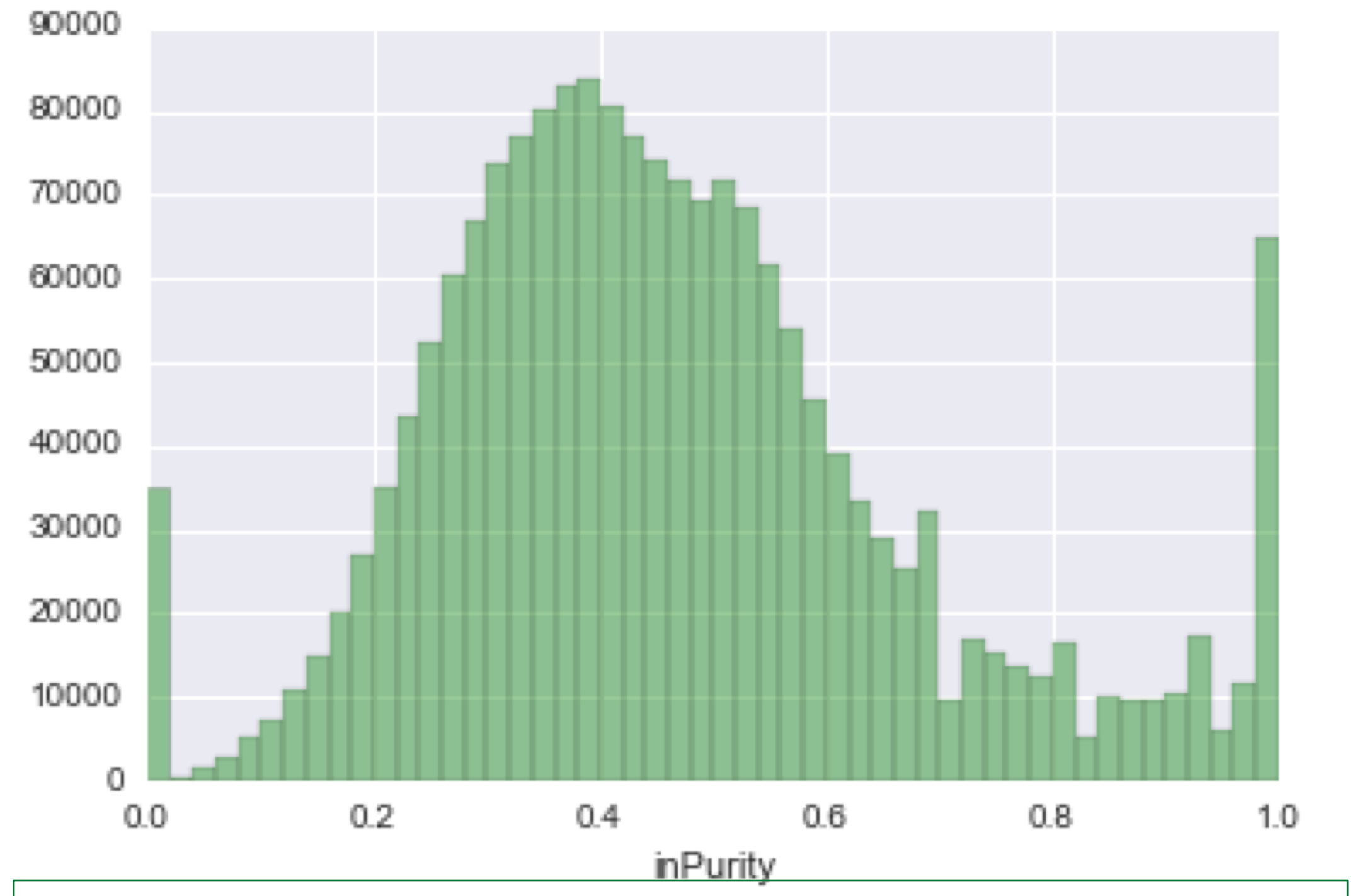

Figure 3. Distribution of msPurity scores. Score could only be reliably estimated for 509,139 of the scans.

Discussion

Maracluster grouped the scans into the largest number of clusters followed by MS-cluster and then Spectra-cluster. All three algorithms were unable to group nearly half of the scans (shown as clusters of size one). The bulk of the scans that could be grouped were formed part of clusters with less than 100 members. The cluster similarity analysis showed that the clusters produced by Spectra-cluster were very distinct from the clusters produced by the other algorithms. Maracluster and MS-cluster showed a large amount of concordance in the composition of the clusters formed. Maracluster grouped the scans into the largest number of clusters followed by MS-cluster and then Spectra-cluster. Figure 4 shows a representative example of the structure for much of the data. It is centered around a cluster formed by Spectra-cluster, it has six member spectra which are all grouped into different Maraclusters and MS-clusters respectively. Only “Sample_031_Scan_39257” had a calculated purity score. Two plotted examples of the scans are presented.

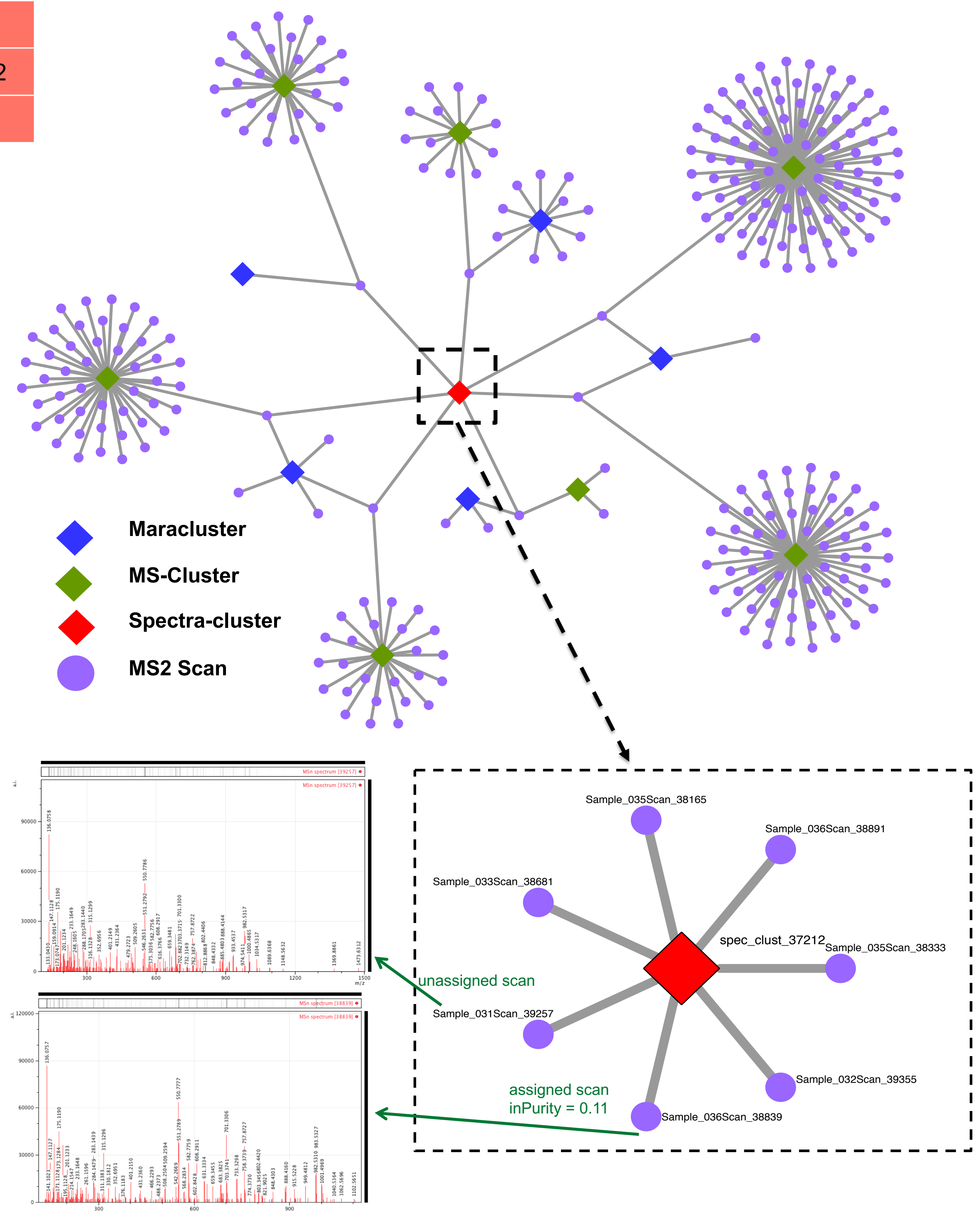

Figure 4. Connected component from Cluster-Scan network. “spec_clust_37212” is selected as an example of the data visualized as a network. The edges show which scans belong to clusters.

Conclusion

The choice of clustering algorithm has a major impact on the grouping of scans for this particular dataset. Clusters that do share scans are very different in terms of there size and the purity scores of the scans. No clear pattern between clustering behavior and spectral purity could be determined at this time since purity scores could only be determined for a third of the scans.

Contact

Armin Geiger (geigerag@ornl.gov)
Paul Abraham (abrahampe@ornl.gov)

References

[1] Frank, Ari M., et al. "Clustering millions of tandem mass spectra." *Journal of Proteome Research* 7.01 (2007): 113-122
[2] The, M. Käll L., et al. "MaRaCluster: A Fragment Rarity Metric for Clustering Fragment Spectra in Shotgun Proteomics." *Journal of Proteome Research* 15.3 (2016): 713-720.
[3] Griss, J., et al. "Recognizing millions of consistently unidentified spectra across hundreds of shotgun proteomics datasets." *Nature methods* 13.8 (2016):651-656
[4] http://www.perseus-framework.org
[5] Lawson, Thomas N, et al. "MsPurity: Automated Evaluation of Precursor Ion Purity for Mass Spectrometry-Based Fragmentation in Metabolomics." Analytical Chemistry. 89.4 (2017): 2432-439.
[6] Shannon, Paul, et al. "Cytoscape: a software environment for integrated models of biomolecular interaction networks." *Genome research* 13.11 (2003): 2498-2504.

Acknowledgements

This research used resources of the Oak Ridge Leadership Computing Facility (OLCF) at the Oak Ridge National Laboratory, which is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC05-00OR22725 as well as the Compute and Data Environment for Science (CADES) at ORNL.