Supplementary Material

B: APPLICATION DATA DESCRIPTION

Note: citations in **boldface** refer to the bibliography in the main manuscript.

B.1 Samples and Data Description

This study uses deidentified data that were previously published. Originally, peripheral blood (PB) samples from 247 treatment-naïve CLL patients were obtained at the University of Texas M.D. Anderson Cancer (MDACC) and processed as described.**[43-45]** Informed consent was obtained and the studies were conducted according to the principles expressed in the Declaration of Helsinki and approved by the Institutional Review Boards.

Study entry occurred upon presentation for care at a tertiary cancer center (MDACC). Clinical and routine laboratory data were assessed at time of study entry, and later obtained retrospectively by review of the medical records for research purposes. These data included 22 clinical measurements drawn from physical exam, demographic factors, and routine laboratory testing conducted as part of the standard of care. A detailed definition and description of data type of each feature is included in **Table B.1.**

The somatic mutation status of immunoglobulin heavy chain variable region (*IGHV*) genes and ZAP70 expression were assessed on blood or bone marrow samples and measured by either flow cytometry or immunohistochemistry, according to established protocols.**[46-48]** Common CLL-associated cytogenetic abnormalities were assessed by array-based SNP genotyping.**[43, 48]** In these data, cytogenetic abnormalities are available both as binary indicators of presence or absence of each abnormality on the chromosome or as the categorical Döhner classification. NOTCH mutation status was removed from analysis , due to unavailability in 144 of 247 patients.

Table B.1. Data type and description for 22 mixed, clinical features collected on 247 patients with CLL.

Type	**Description**	**Values**
Binary	*IGHV* mutation status	Mutated or unmutated
	NOTCH mutation status	Mutated or unmutated
	Zap70 expression	Positive or negative
	Chromosome 13 status	Presence or absence of cytogenetic abnormality del(13)(q14.3)
	Chromosome 11 status	Presence or absence of cytogenetic abnormality del(11)(q22.3)
	Chromosome 12 status	Presence or absence of cytogenetic abnormality trisomy 12
	Chromosome 17 status	Presence or absence of cytogenetic abnormality del(17)(p13.1)
	Sex	Male or female
	Rai stage category	Low or high tumor staging at presentation
	Massive splenomegaly	Presence or absence on physical exam
	CD38	Low or high
	Beta-2 microglobulin	Low or high
	White blood cell count	Low or high
	Hypogammaglobulinemia	Presence or absence
	Matutes immunophenotype	Typical or atypical immunophenotype
	Light chain subtype	Kappa or lambda immunoglobulin light chain
Ordinal	Döhner classification	5 category hierarchy for prognostically significant cytogenetic abnormalities
Nominal	Race	White, Black, Asian, or Hispanic/Latino
Continuous	Age at Diagnosis	years
	Hemoglobin	g/dL
	Platelet count	number per mm^3
	Prolymphocyte count	number per mm^3

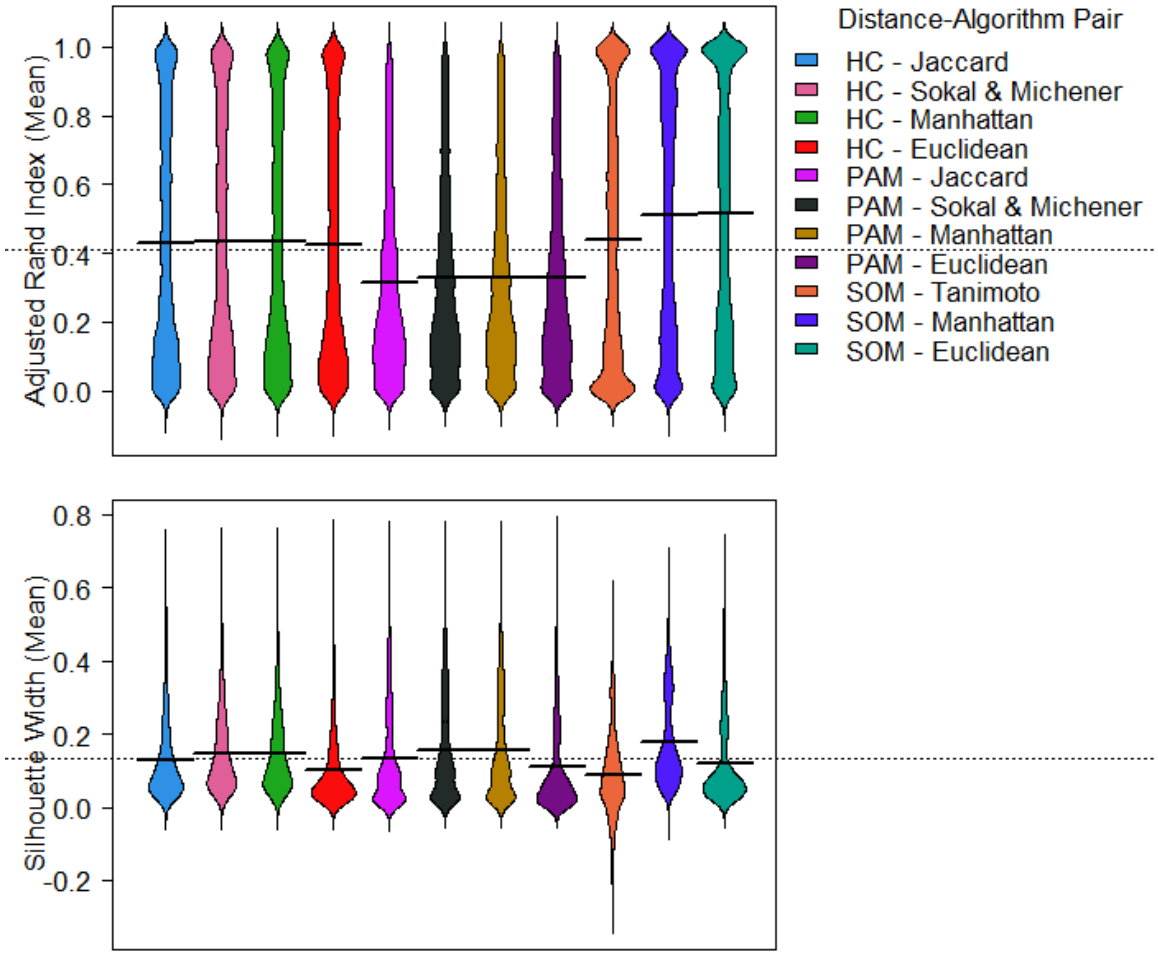
Supplementary Table B.1. Results of single-distance methods for simulations of single data types. 13 algorithm-single distance paired methods were tested on 5 single data types. Mean ± sd are presented as averages across all simulation parameters (number of patients, features, and clusters). On noisy simulations across each data type and distance metric, HC had higher ARI than PAM. HC had higher silhouette widths than PAM on continuous, ordinal, and mixed categorical data. The two algorithms had similar silhouette width performance for nominal and binary data. SOM had highest ARI and SW for all data types and distance metrics, except nominal data. Tested omitted in this table could not be performed because the distance metric in question was incompatible with the data type of interest.

		Data Type									
		Binary		**Nominal**		**Ordinal**		**Categorical**[1]		**Continuous**	
Algorithm	**Distance**	***ARI***[5]	***SW***[6]	***ARI***[5]	***SW***[6]	***ARI***[5]	***SW***[6]	***ARI***[5]	***SW***[6]	***ARI***[5]	***SW***[6]
HC[2]	Jaccard	0.430 ± 0.342	0.129 ± 0.108								
	Sokal & Michener	0.433 ± 0.341	0.147 ± 0.106								
	Gower			**0.283 ± 0.298**	0.024 ± 0.020	0.271 ± 0.293	0.023 ± 0.019	0.276 ± 0.295	0.023 ± 0.020		
	Manhattan	0.434 ± 0.341	0.148 ± 0.107	0.141 ± 0.219	0.035 ± 0.038	0.376 ± 0.336	0.064 ± 0.042	0.280 ± 0.305	0.049 ± 0.039	0.561 ± 0.341	0.06 ± 0.050
	Euclidean	0.426 ± 0.343	0.101 ± 0.106	0.047 ± 0.100	0.036 ± 0.044	0.335 ± 0.327	0.059 ± 0.042	0.211 ± 0.272	0.046 ± 0.043	0.602 ± 0.345	0.085 ± 0.053
PAM[3]	Jaccard	0.314 ± 0.269	0.133 ± 0.131								
	Sokal & Michener	0.331 ± 0.276	0.156 ± 0.138								
	Gower			0.103 ± 0.105	0.020 ± 0.017	0.096 ± 0.097	0.019 ± 0.017	0.099 ± 0.100	0.019 ± 0.017		
	Manhattan	0.331 ± 0.276	0.156 ± 0.138	0.044 ± 0.057	0.035 ± 0.038	0.207 ± 0.195	0.056 ± 0.040	0.121 ± 0.137	0.043 ± 0.039	0.373 ± 0.277	0.063 ± 0.051
	Euclidean	0.329 ± 0.277	0.109 ± 0.125	0.018 ± 0.025	0.038 ± 0.045	0.189 ± 0.192	0.055 ± 0.044	0.094 ± 0.123	0.044 ± 0.045	0.446 ± 0.302	0.076 ± 0.056
SOM[4]	Tanimoto	0.437 ± 0.361	0.086 ± 0.114								

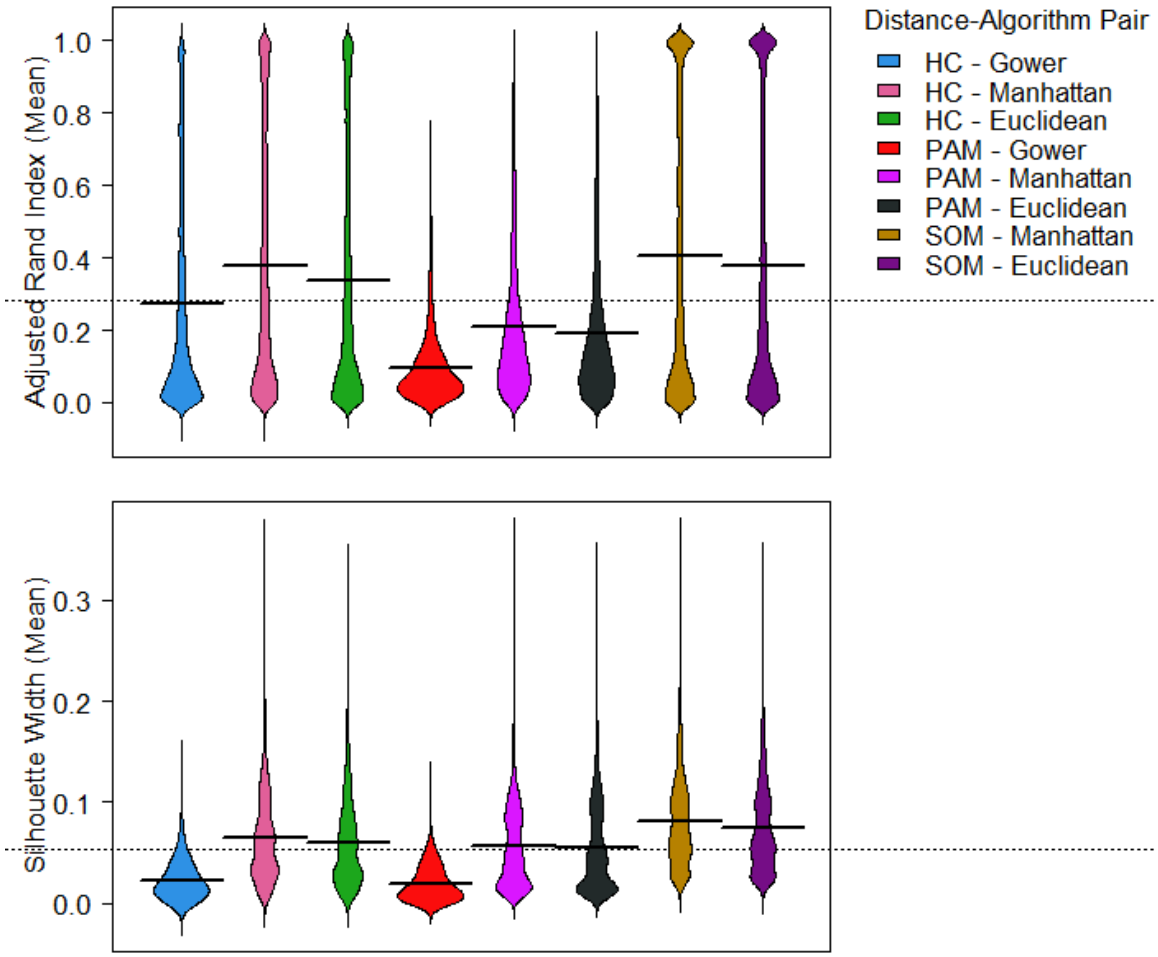
	Manhattan	0.513 ± 0.359	**0.177 ± 0.120**	0.131 ± 0.224	**0.052 ± 0.046**	**0.405 ± 0.368**	**0.081 ± 0.044**	**0.301 ± 0.342**	**0.066 ± 0.044**	0.568 ± 0.344	0.083 ± 0.049
	Euclidean	**0.516 ± 0.357**	0.117 ± 0.112	0.092 ± 0.188	0.052 ± 0.049	0.380 ± 0.368	0.075 ± 0.044	0.274 ± 0.337	0.062 ± 0.046	**0.611 ± 0.336**	**0.093 ± 0.051**

[1] A mixture of nominal and ordinal features.
[2] Agglomerative hierarchical clustering with Ward's criterion
[3] Partitioning Around Medoids
[4] Kohonen self-organizing maps
[5] Adjusted Rand Index; mean ± standard deviation
[6] Average silhouette width; mean ± standard deviation

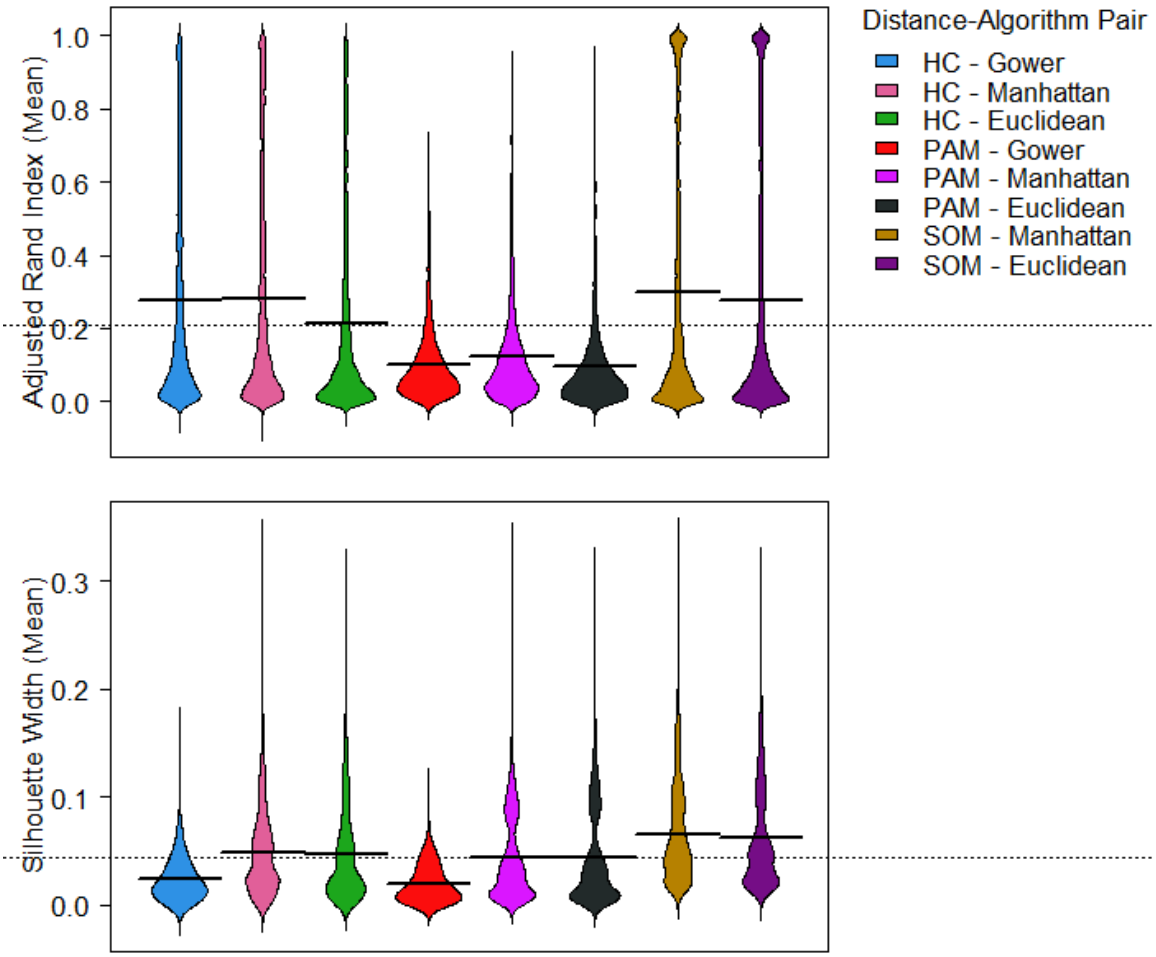
Supplemental Figure B.1. Violin plots of adjusted rand index (top) and silhouette width (below) for simulated binary data. All distance methods and algorithms produced solutions spanning a range of ARI from 0 to 1. SOM with Euclidean distance resulted in the highest mean ARI (0.516 ± 0.357) followed by SOM with Manhattan distance (0.513 ± 0.359). SOM with Manhattan distance also produced the highest mean SW (0.177 ± 0.120). Across HC or PAM, performance of the 4 distance metrics in question (Jaccard, Sokal & Michener, Manhattan, and Euclidean) produced similar results. PAM ARI's were heavily weighted towards inaccurate solutions (ARI between 0 and 0.4). HC and SOM produced bipolar results, with ARI clustered either near 1 or near 2. The bolus of solutions near 1 was larger for SOM than HC. The strongest bipolar distribution of ARI resulted from the Tanimoto distance. SOM with the Manhattan distance produced many solutions with lower silhouette widths, but resulted in a group of simulations with higher SW than other solutions, including PAM. The Tanimoto distance presented with the lowest range of SW, with a tail of many values less than 0.



Supplemental Figure B.2. Violin plots of adjusted rand index (top) and silhouette width (below) for simulated ordinal data. Clustering solutions of ordinal data produced intermediate ARI and SW. SOM with the Manhattan distance produced the solutions with highest mean ARI (0.405 ± 0.368) and SW (0.081 ± 0.044). The Gower distance had lower ARI and SW performance by quantitative measures and bean plot visualization than the Manhattan or Euclidean distance.(Figure 4.4) HC, PAM, and SOM all visualized with a range of ARI from 0 to 1. PAM solutions weighted towards 0. SOM solutions displayed a bipolar distribution, with solutions clustered either near 0 or a bolus of solutions near 1. All implementations of the Manhattan and Euclidean distance resulted in range of SW weighted between 0 and 0.2.



Supplemental Figure B.3. Violin plots of adjusted rand index (top) and silhouette width (below) for simulated mixed categorical data. Mixed categorical data resulted in low ARI and SW. SOM with the Manhattan distance produced the highest mean ARI (0.301 ± 0.342) and SW (0.066 ± 0.044). Visualization revealed a range of ARI with a heavy distribution near 0. SOM produced a small fraction of solutions near 1. The Euclidean and Manhattan distances with all 3 algorithms produced a range of SW between 0 and 1, with PAM producing many low solutions and a portion of solutions with elevated SW.

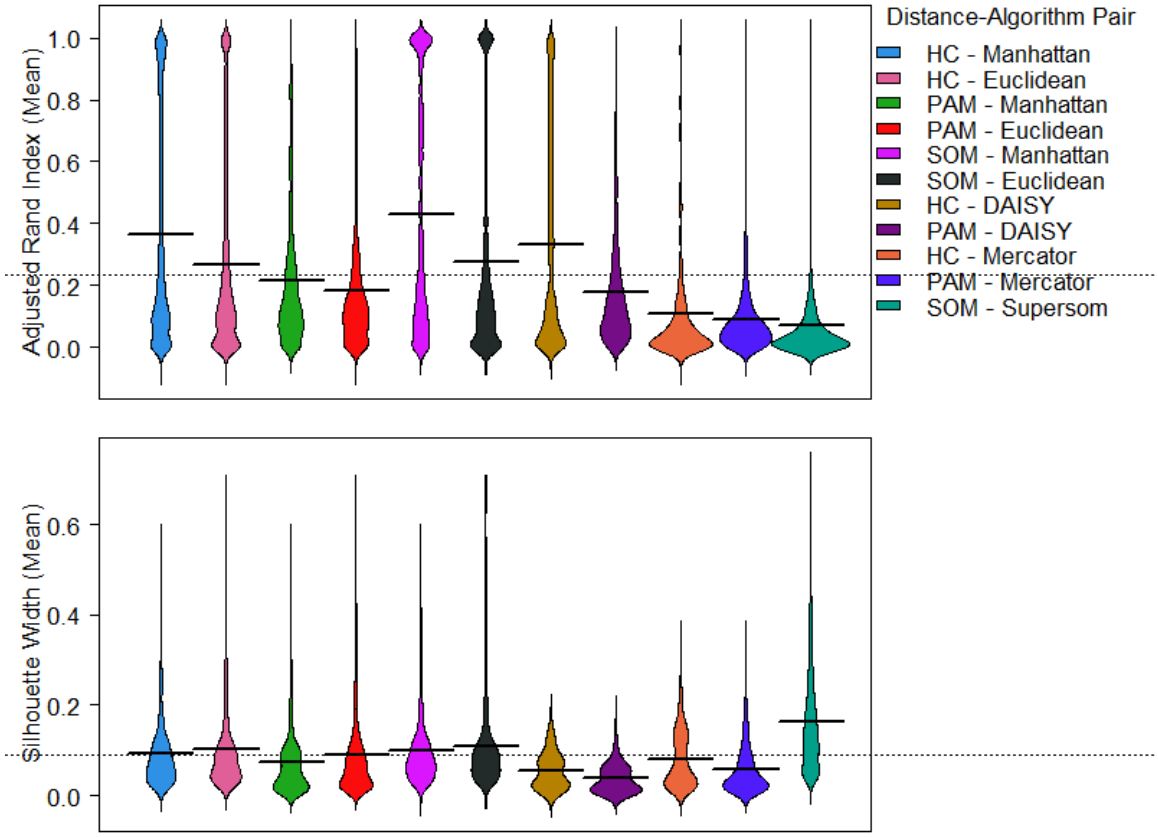


Supplementary Table B.2. Results of single- and mixed-distance methods for plausible, simulated mixed data types. 3 mixed-distance metrics of dissimilarity calculation and 2 single-distance controls were evaluated on 3 clustering algorithms. Mean ± sd are presented as averages across all simulation parameters (number of patients, features, and clusters). On noisy simulations across each data mixture and distance metric, HC had higher ARI than PAM. DAISY with HC had superior performance on all data mixtures except unbalanced continuous, which had highest mean ARI from SOM with Manhattan distance. In 3 of 4 data types, Supersom resulted in the highest SW. However, SOM with the Manhattan distance had higher ARI than Supersom across all data types.

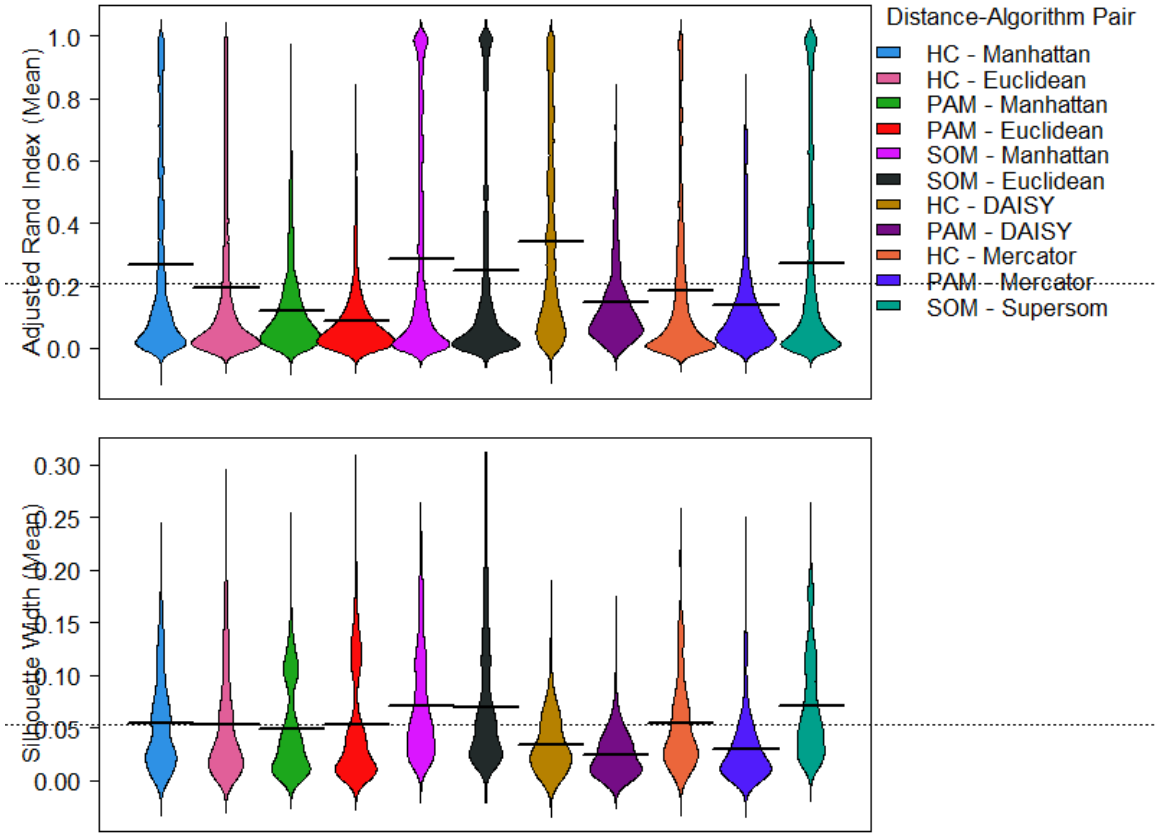
		Data Mixture Type							
		Balanced		**Binary Unbalanced**		**Categorical Unbalanced**		**Continuous Unbalanced**	
Distance	**Algorithm**	***ARI***[1]	***SW***[2]	***ARI***[1]	***SW***[2]	***ARI***[1]	***SW***[2]	***ARI***[1]	***SW***[2]
Manhattan	HC[3]	0.430 ± 0.357	0.081 ± 0.068	0.349 ± 0.366	0.142 ± 0.123	0.267 ± 0.303	0.055 ± 0.044	0.472 ± 0.385	0.105 ± 0.085
	PAM[4]	0.203 ± 0.210	0.068 ± 0.072	0.204 ± 0.228	0.118 ± 0.127	0.121 ± 0.135	0.049 ± 0.045	0.271 ± 0.258	0.080 ± 0.087
	SOM[5]	0.460 ± 0.278	0.098 ± 0.070	0.402 ± 0.417	0.153 ± 0.117	0.288 ± 0.338	0.071 ± 0.049	**0.564 ± 0.392**	0.110 ± 0.080
Euclidean	HC[3]	0.232 ± 0.299	0.079 ± 0.083	0.075 ± 0.175	0.156 ± 0.141	0.195 ± 0.263	0.053 ± 0.050	0.335 ± 0.359	0.119 ± 0.105
	PAM[4]	0.115 ± 0.157	0.077 ± 0.089	0.073 ± 0.134	0.140 ± 0.144	0.088 ± 0.114	0.052 ± 0.053	0.219 ± 0.240	0.101 ± 0.109
	SOM[5]	0.278 ± 0.353	0.097 ± 0.085	0.083 ± 0.204	0.160 ± 0.136	0.248 ± 0.325	0.069 ± 0.053	0.353 ± 0.385	0.123 ± 0.100
DAISY	HC[3]	**0.474 ± 0.352**	**0.099 ± 0.085**	**0.574 ± 0.324**	0.091 ± 0.053	**0.341 ± 0.311**	0.034 ± 0.026	0.393 ± 0.359	0.060 ± 0.043
	PAM[4]	0.279 ± 0.248	0.084 ± 0.093	0.387 ± 0.279	0.077 ± 0.060	0.146 ± 0.139	0.025 ± 0.020	0.205 ± 0.197	0.041 ± 0.034
Mercator	HC[3]	0.467 ± 0.366	0.093 ± 0.069	0.327 ± 0.165	0.089 ± 0.064	0.183 ± 0.253	0.054 ± 0.045	0.127 ± 0.219	0.085 ± 0.068
	PAM[4]	0.274 ± 0.248	0.074 ± 0.071	0.165 ± 0.187	0.069 ± 0.064	0.136 ± 0.163	0.030 ± 0.032	0.101 ± 0.135	0.065 ± 0.061
Supersom	SOM[5]	0.243 ± 0.312	0.098 ± 0.080	0.061 ± 0.158	**0.193 ± 0.137**	0.270 ± 0.326	**0.071 ± 0.049**	0.079 ± 0.190	**0.174 ± 0.123**

[1] Adjusted Rand Index; mean ± standard deviation
[2] Average silhouette width; mean ± standard deviation
[3] Agglomerative hierarchical clustering with Ward's criterion
[4] Partitioning Around Medoids
[5] Kohonen self-organizing maps

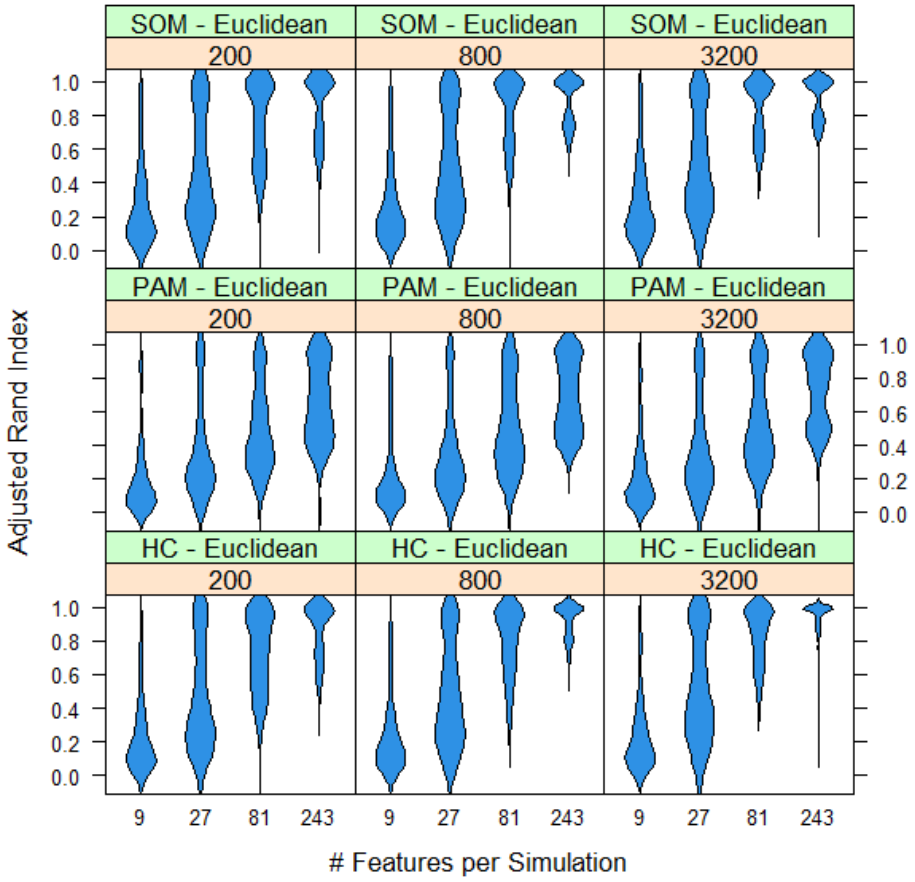
Supplemental Figure B.4 Violin plots of adjusted rand index (top) and silhouette width (below) for simulated unbalanced, continuous-dominant data mixtures. The highest mean ARI solutions were produced by SOM with the single Manhattan distance (0.564 ± 0.392) with the highest mean SW produced by Supersom (0.174 ± 0.123). By visualization, SOM, HC with single distances, and HC with DAISY produce bipolar distributions of ARI, with solutions with PAM, Mercator, and Supersom weighted towards 0.(Figure 4.9) DAISY and Mercator result in low SW, below the overall mean, compared to single distance metrics, SOM, or Supersom.



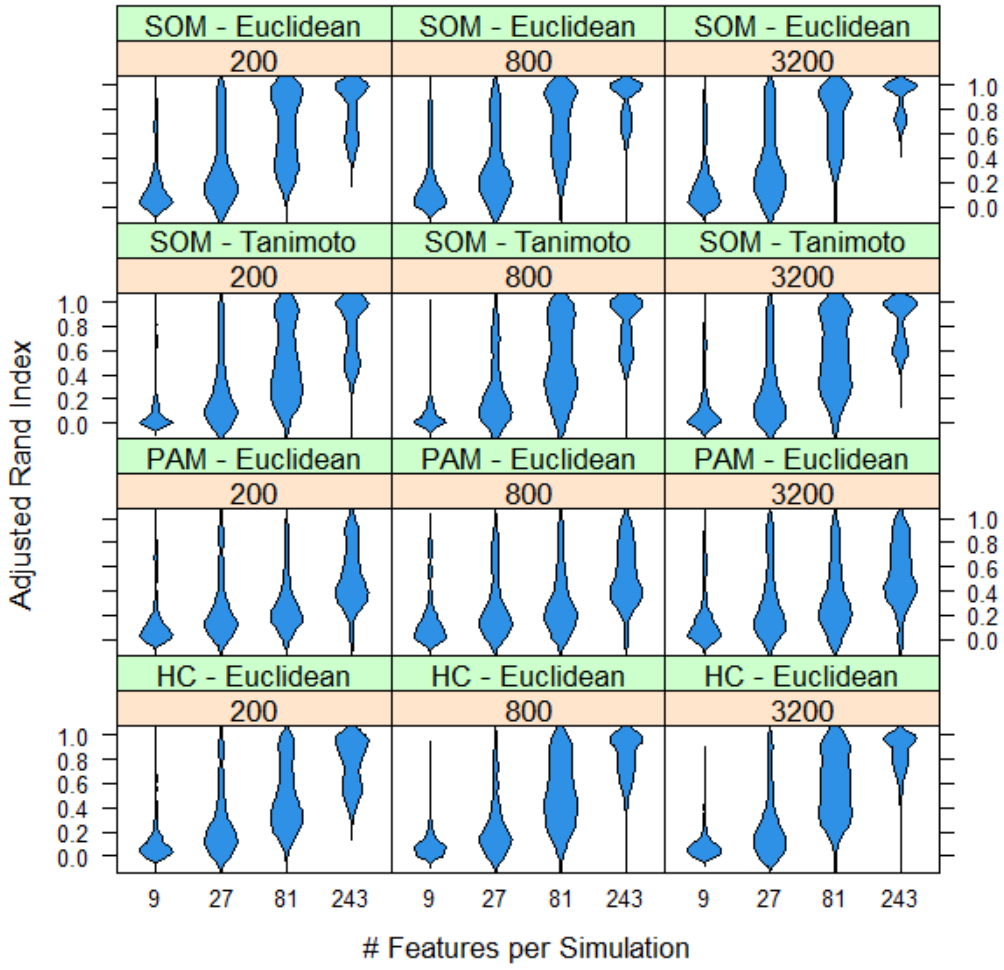
Supplemental Figure B.5. Violin plots of adjusted rand index (top) and silhouette width (below) for simulated unbalanced, categorical-dominant data mixtures. The highest mean ARI was produced by DAISY with HC (0.341 ± 0.311) with highest mean SW produced by Supersom (0.071 ± 0.049). When visualized HC with DAISY or the Manhattan distance produces solutions with a range of ARI between 0 and 1.(Figure 4.8) SOM and Supersom produce bipolar distributions of ARI. SW are low, with single distance metrics, SOM, and Supersom outperforming DAISY and Mercator.



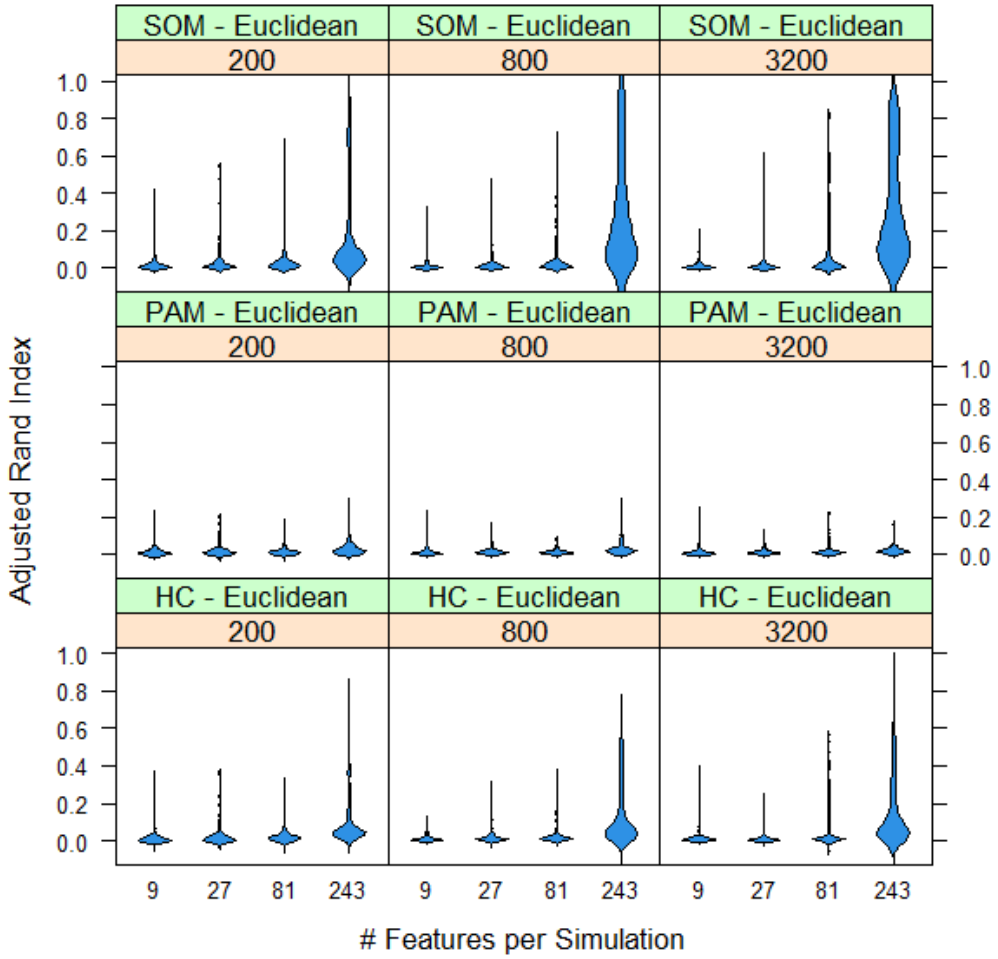
Supplemental Figure B.6. Lattice violin plot of Adjusted Rand Index (ARI) of continuous simulations by number of features and patients with 3 algorithms and the Euclidean distance. Continuous data were plotted with the Euclidean distance across 3 algorithsms (hierarchical clustering "HC", Partitioning Around Medoids "PAM", and self-organizing maps "SOM") Across algorithms, ARI varied strongly by number of features, but not by number of patients: lowest in simulations with 9 features and highest in simulations with 243 features. Intermediate feature spaces displayed higher degrees of variability, represented by broad spectra of ARI across many simulations.



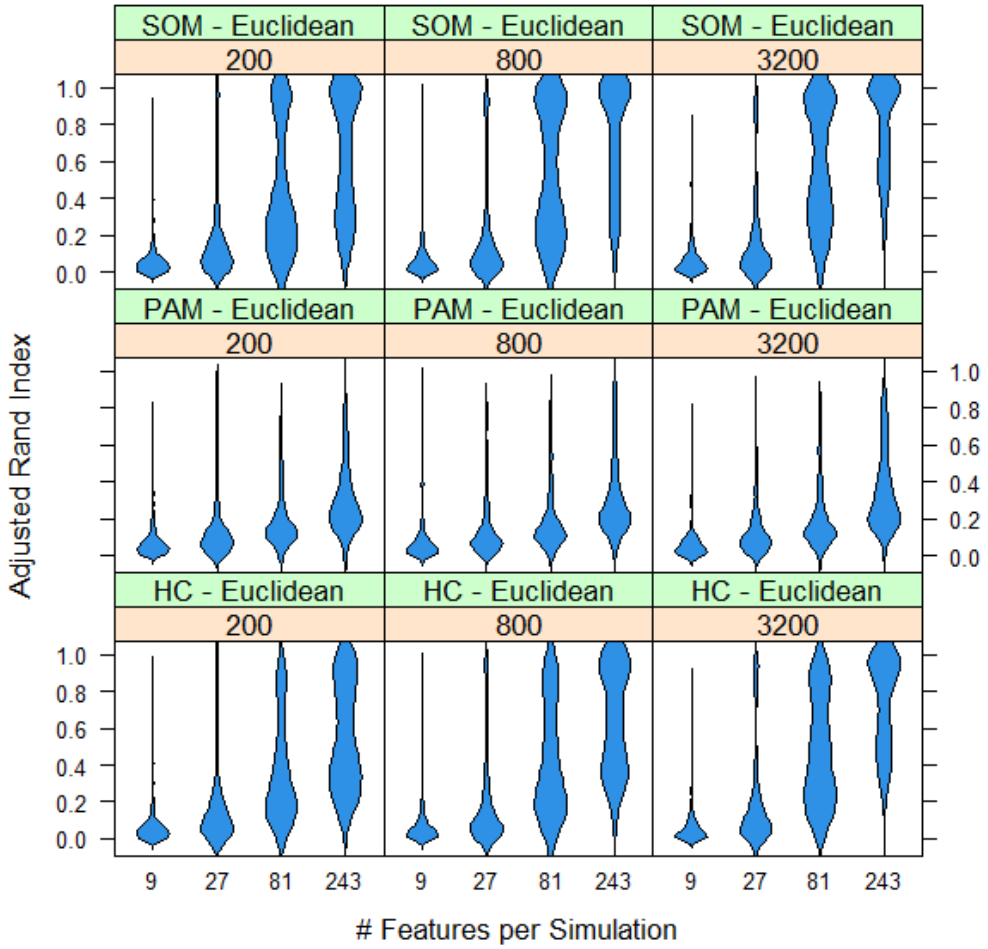
Supplemental Figure B.7. Lattice violin plot of ARI of binary simulations by number of features and patients with 3 algorithms with Euclidean and Tanimoto distance. ARI varies strongly by number of features, but not by number of patients. ARI was lowest among simulations with 9 features and highest among simulations with 243 features. Intermediate features spaces displayed higher degrees of variability, represented by broad spectra of ARI across many simulations.



Supplemental Figure B.8. Lattice violin plot of ARI of nominal simulations by number of features and patients with 3 algorithms and Euclidean distance. ARI varies strongly by number of features, but not by number of patients. ARI was lowest among simulations with 9 features and highest among simulations with 243 features. Intermediate features spaces displayed higher degrees of variability, represented by broad spectra of ARI across many simulations. Categorical simulations displayed poorer performance, even at larger feature spaces. Nominal data, characterized by poor performance at 81 or fewer features, presented with improved, though variable, performance at 243 features.



Supplemental Figure B.9. Lattice violin plot of ARI of ordinal simulations by number of features and patients with 3 algorithms and Euclidean distance. ARI varies strongly by number of features, but not by number of patients. ARI was lowest among simulations with 9 features and highest among simulations with 243 features. Intermediate features spaces displayed higher degrees of variability, represented by broad spectra of ARI across many simulations. Categorical simulations displayed poorer performance, even at larger feature spaces. Even at simulations with 243 features, ordinal simulations presented broad, variable spectra.



Supplemental Figure B.10. Lattice violin plot of ARI of 4 data mixtures with the Mercator distance algorithm and hierarchical clustering. ARI varies strongly by number of features, but not by number of patients. ARI was lowest among simulations with 9 features and highest among simulations with 243 features. Intermediate features spaces displayed higher degrees of variability, represented by broad spectra of ARI across many simulations.