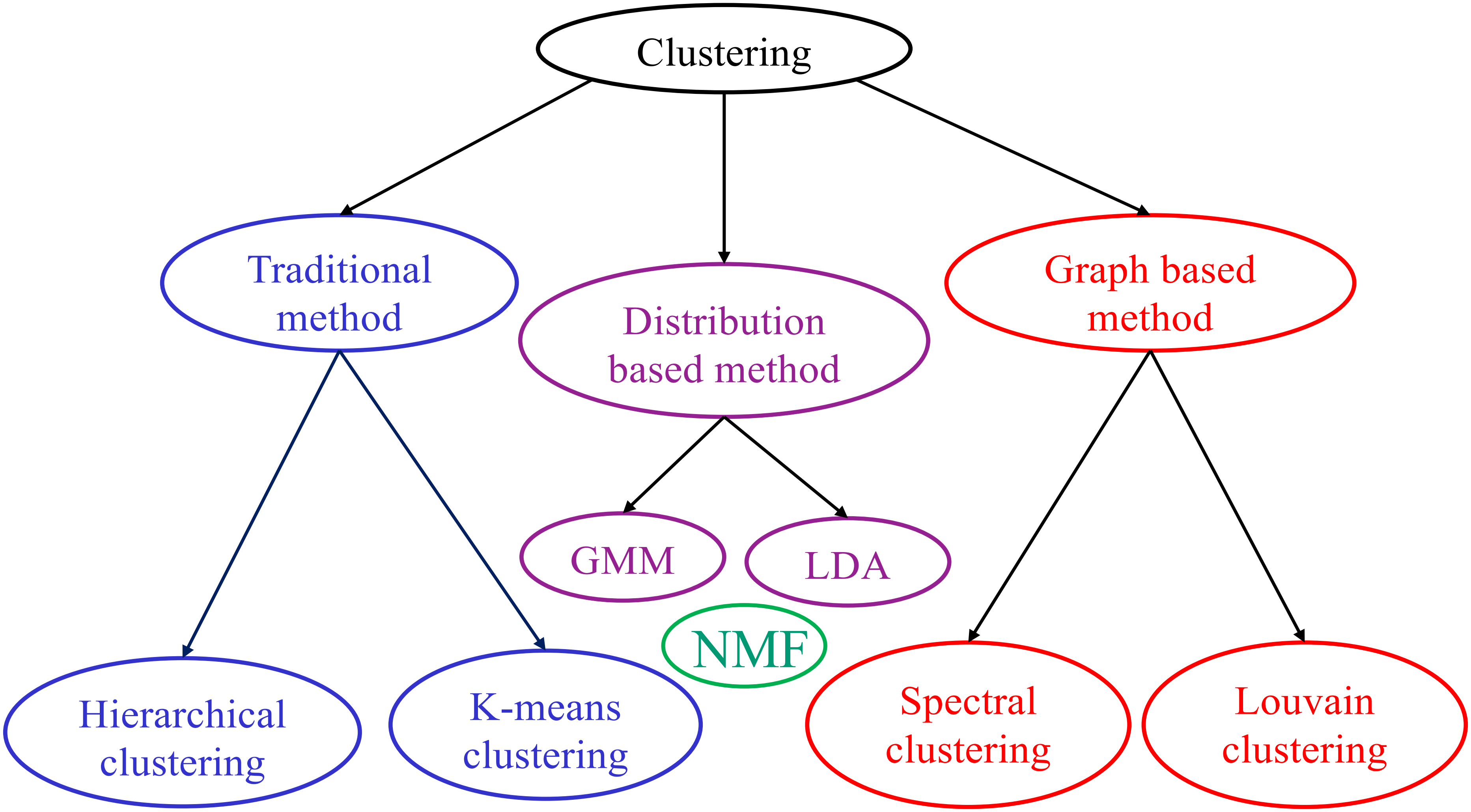
Clustering Methods:
From k-means to Gaussian Mixture Model and Louvain Algorithm

Maxwell Lee

High-dimension Data Analysis Group
Laboratory of Cancer Biology and Genetics
Center for Cancer Research
National Cancer Institute

November 2, 2020

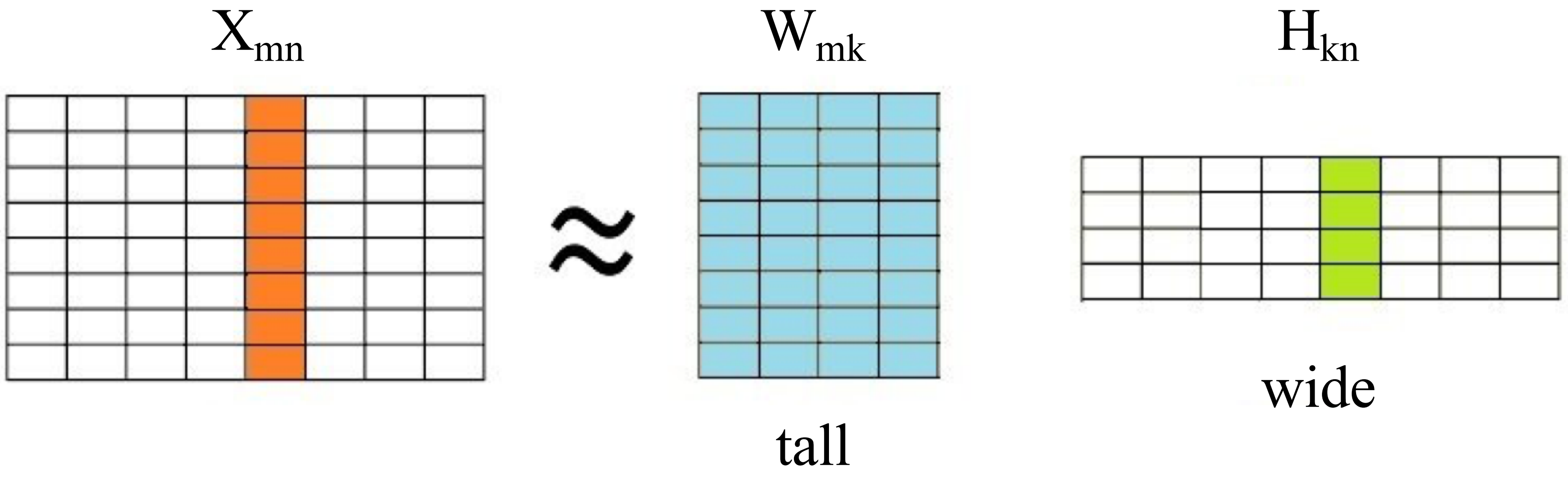
Outline of Clustering Methods



GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation
NMF: Non-negative matrix factorization

Contributed by Emily Tai

Mathematic Model of Non-Negative Matrix Factorization



X_{mn}: m features; n samples
W_{mk}: m features; k latent variables
H_{km}: k latent variables; n encodings

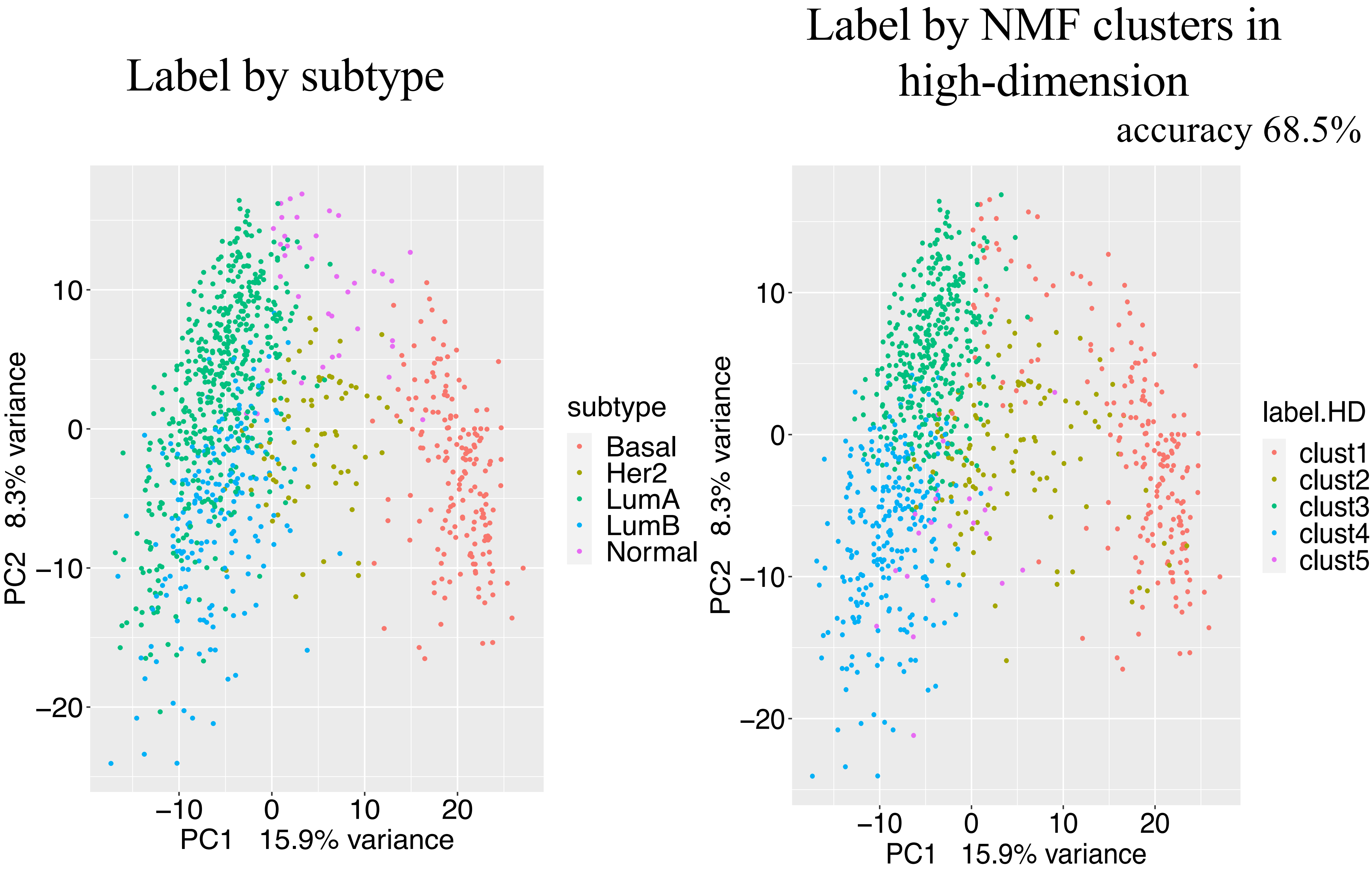
Each element of matrix is non-negative
$X >= 0; W >= 0; H >= 0$

latent variables:
basis images
topics
centroids
signatures

$k << \min(m,n) \longrightarrow$ Dimension reduction

Lee and Seung, Nature 1999; 401:788–791

PCA: Label by Subtype vs. by NMF Cluster



Label by subtype

Label by NMF clusters in high-dimension

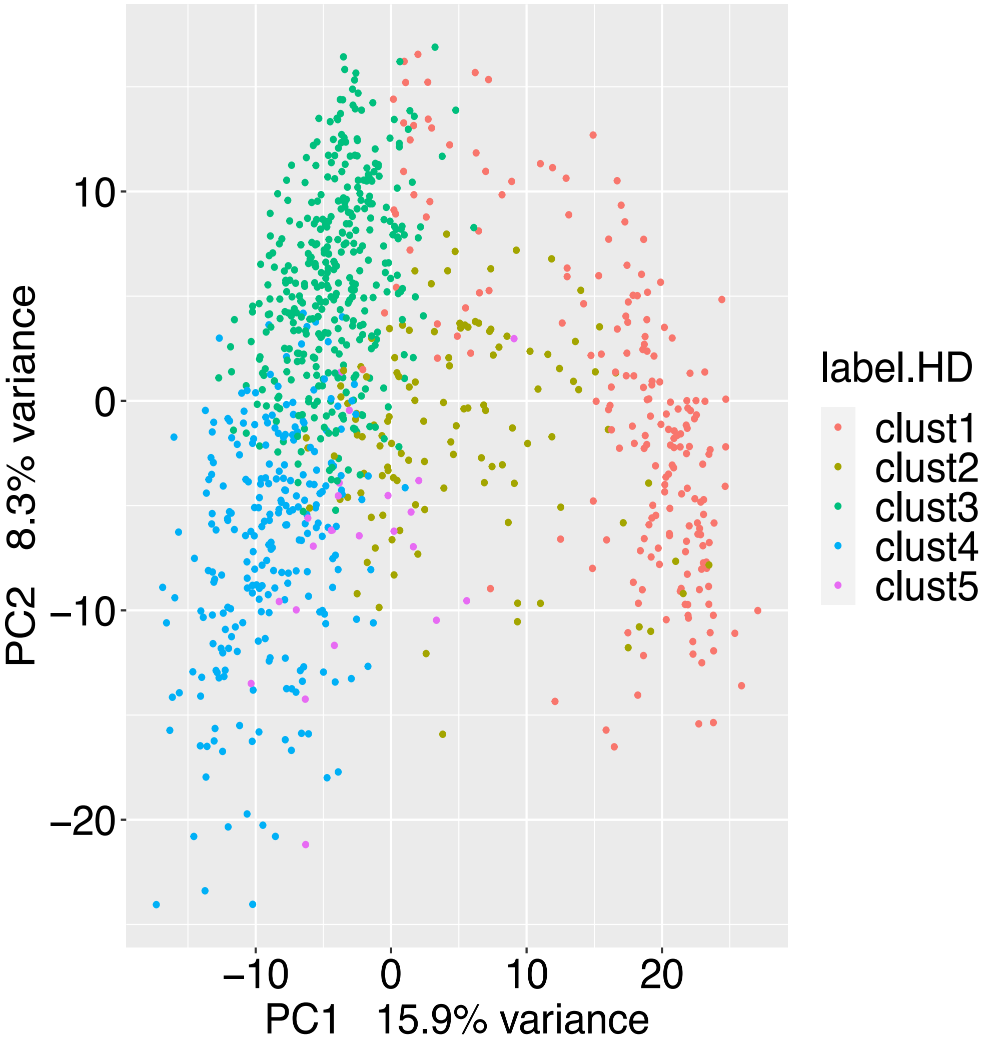
accuracy 68.5%

HD: high dimension, 5000 genes

PCA: Label by NMF Cluster vs. by k-means Cluster

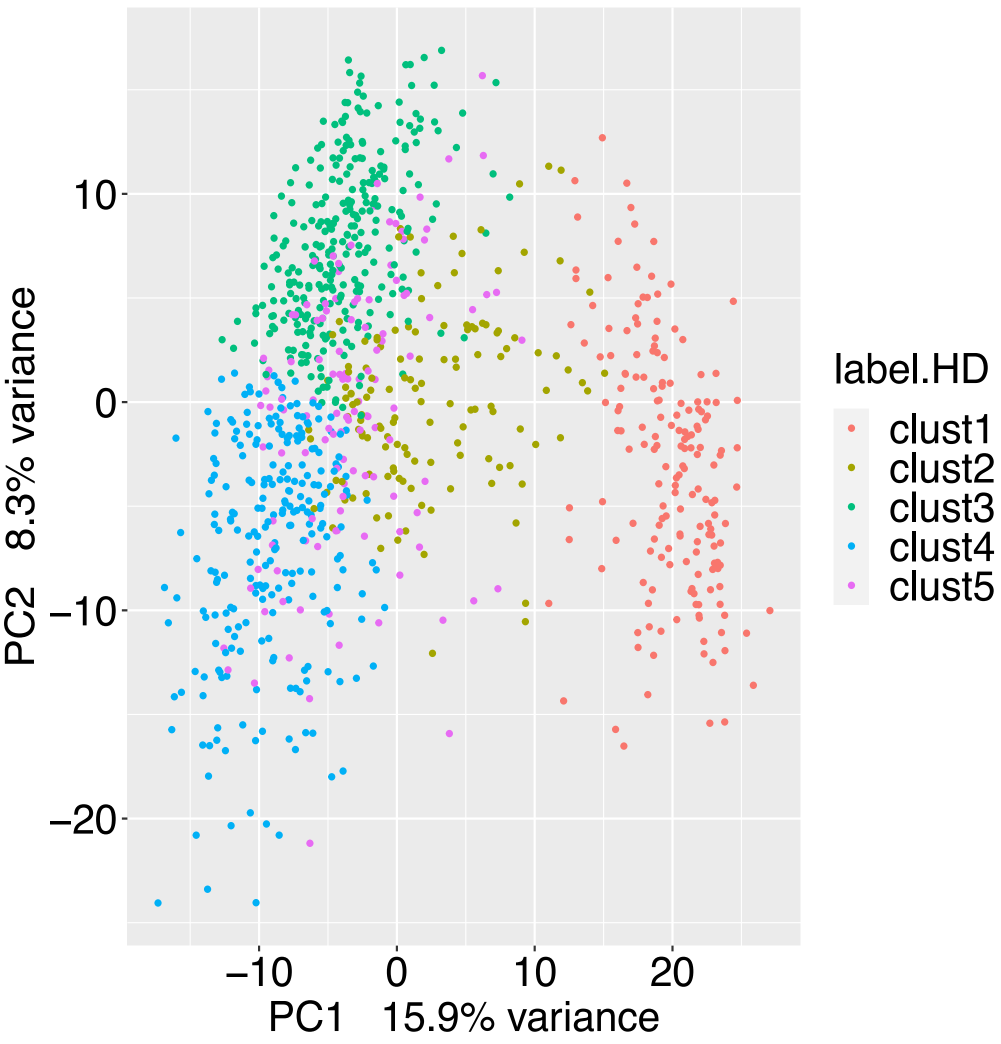
Label by NMF clusters in high-dimension

accuracy 68.5%



Label by k-means clusters in high-dimension

accuracy 65%



HD: high dimension, 5000 genes

Comparison Between Subtype and NMF vs. k-means Cluster (HD)

	Basal	Her2	LumA	LumB	Normal
clust1	155	2	11	1	**30**
clust2	18	62	7	32	2
clust3	0	4	**338**	38	6
clust4	0	2	132	114	0
clust5	0	3	12	8	0

NMF

Accuracy = (155 + 62 + 338 + 114) / 977 = 68.5%

Match

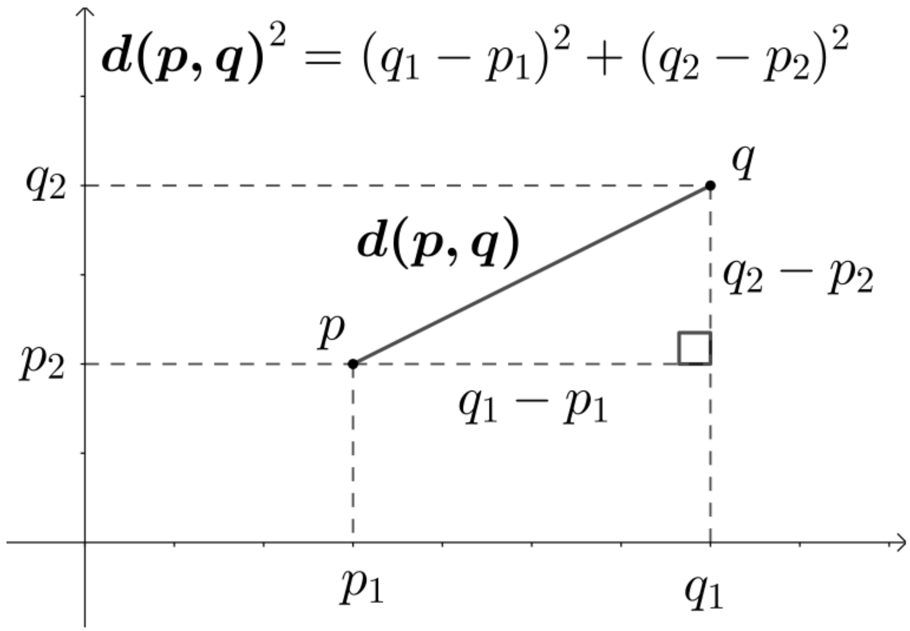
Mismatch

	Basal	Her2	LumA	LumB	Normal
clust1	169	0	0	0	6
clust2	4	69	17	40	5
clust3	0	0	268	11	21
clust4	0	0	125	119	0
clust5	0	4	**90**	23	6

K-means

Accuracy = (169 + 69 + 268 + 119 + 6) / 977 = 64.6%

Euclidean Distance

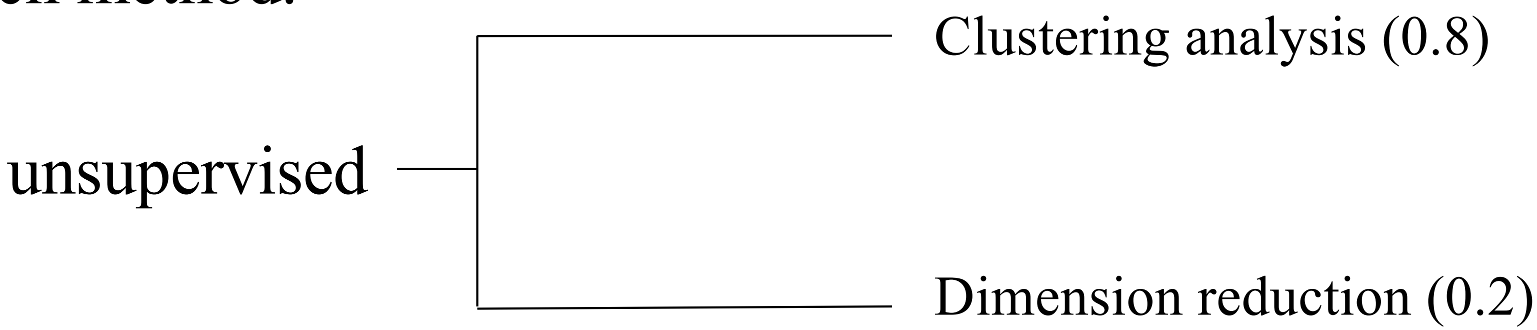


$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \cdots + (q_n - p_n)^2}$$

- Euclidean distance is not affected by the shift of coordinate system
- Euclidean distance is not affected by the rotation of coordinate system
- Euclidean distance is not affected by flipping axis

Understanding NMF from Topic Modeling (Mixture Model)

I will talk about spectral clustering, which is a graph-based method and consists of dimension reduction with Laplacian Eigenmap and k-means clustering in the reduced dimension space. I will also talk about Louvain algorithm, which is used in Seurat package to cluster single cell RNAseq data. Louvain algorithm is a network community approach. It is very fast and has capacity to do clustering analysis for million nodes in a network. I will provide practical examples to illustrate how each method works and how to interpret the results of clustering analysis and explain the pros and cons of each method.

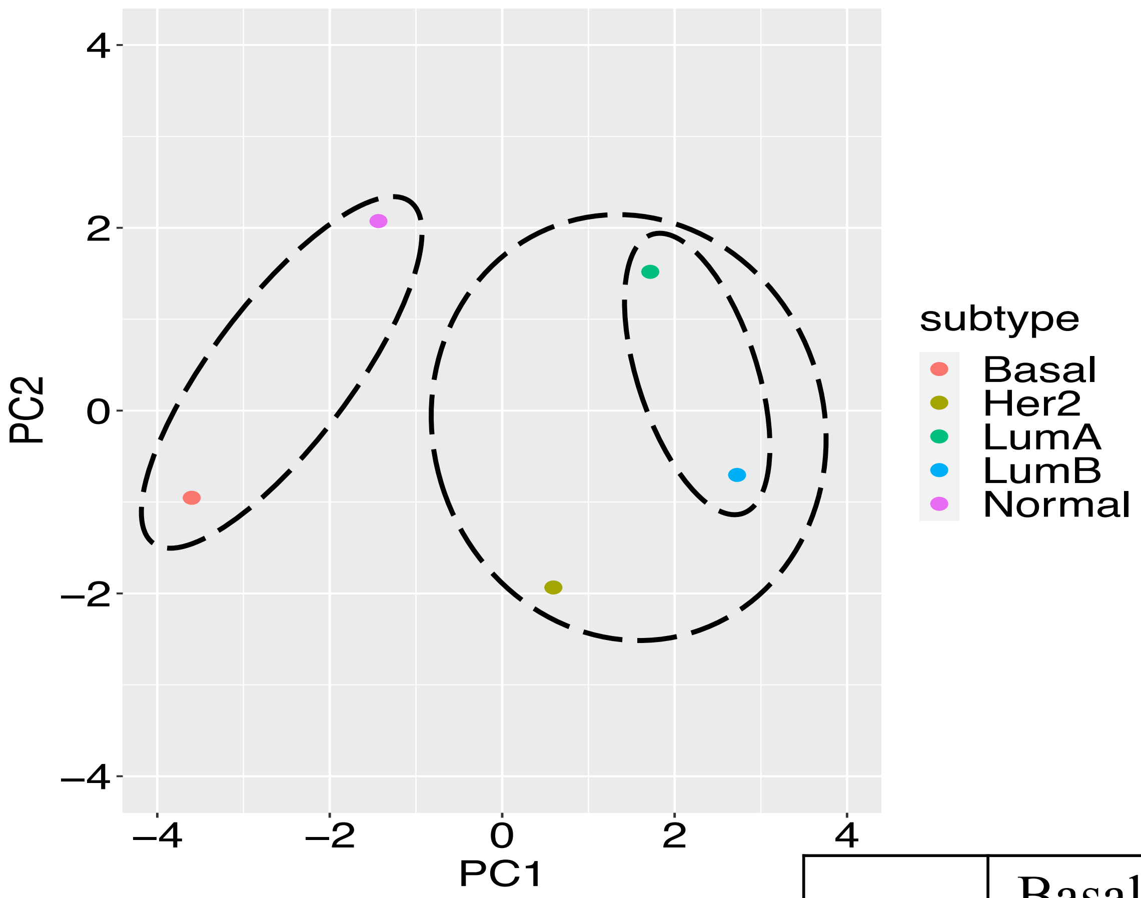


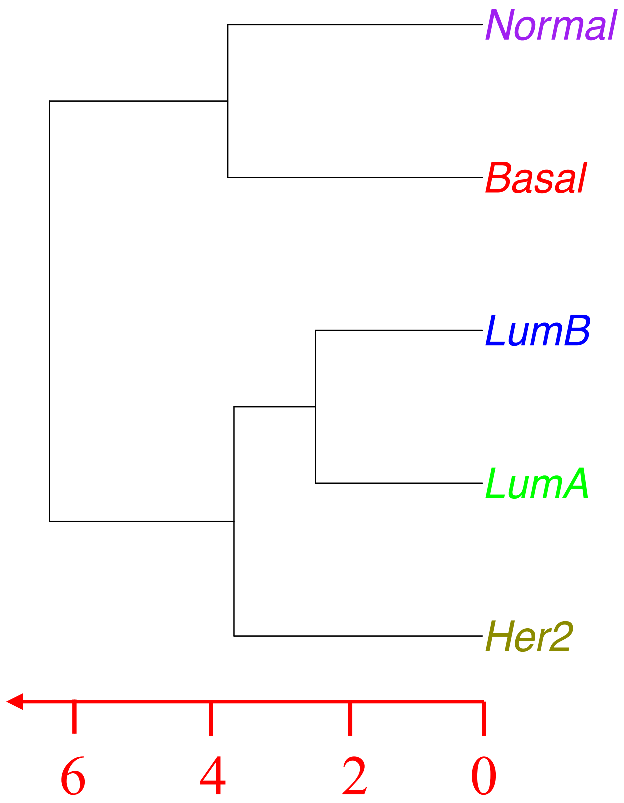
unsupervised

Clustering analysis (0.8)

Dimension reduction (0.2)

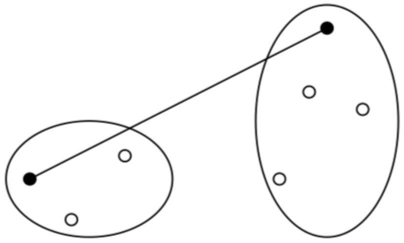
Hierarchical Agglomerative Clustering Analysis





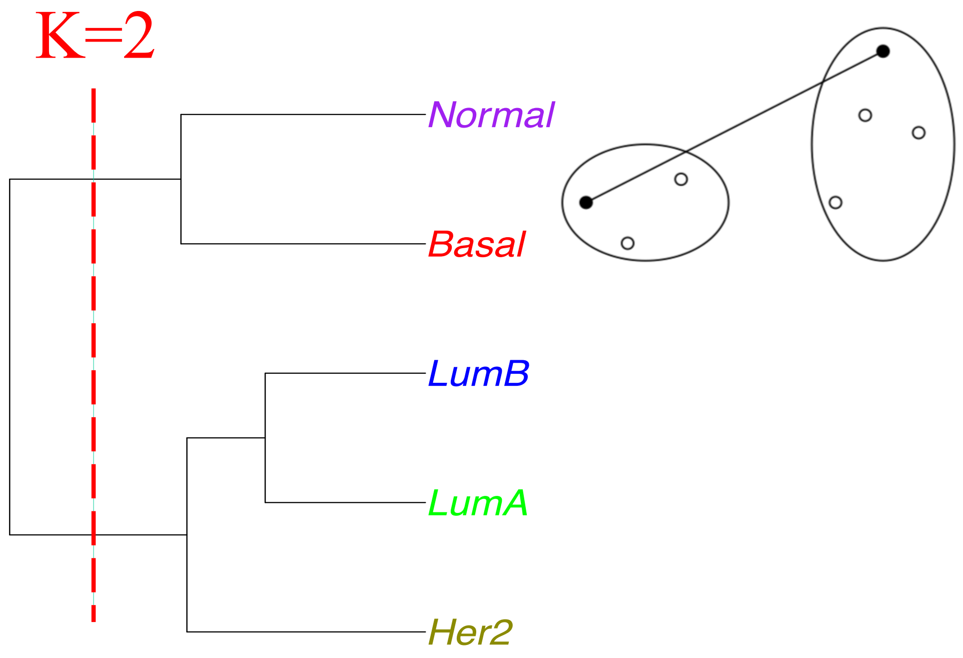
Single-linkage

Complete-linkage

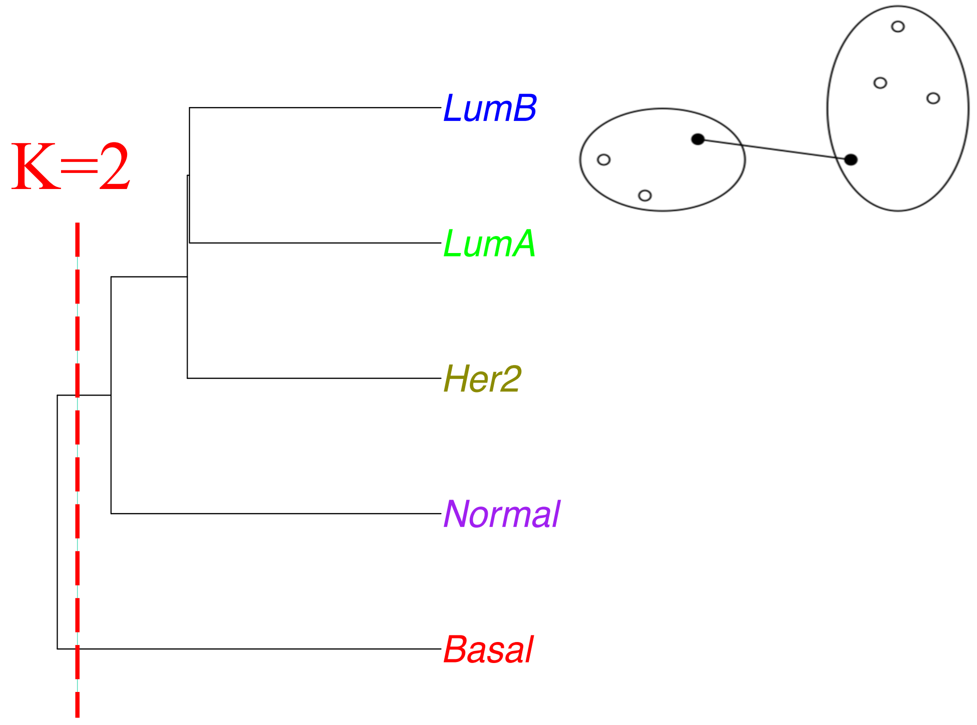


	Basal	Her2	LumA	LumB	Normal
Basal	0	4.31	5.86	**6.33**	**3.72**
Her2	4.31	0	**3.63**	2.46	4.49
LumA	5.86	3.63	0	**2.44**	3.2
LumB	6.33	2.46	2.44	0	5
Normal	3.72	4.49	3.2	5	0

Hierarchical Agglomerative Clustering Analysis



Complete-linkage

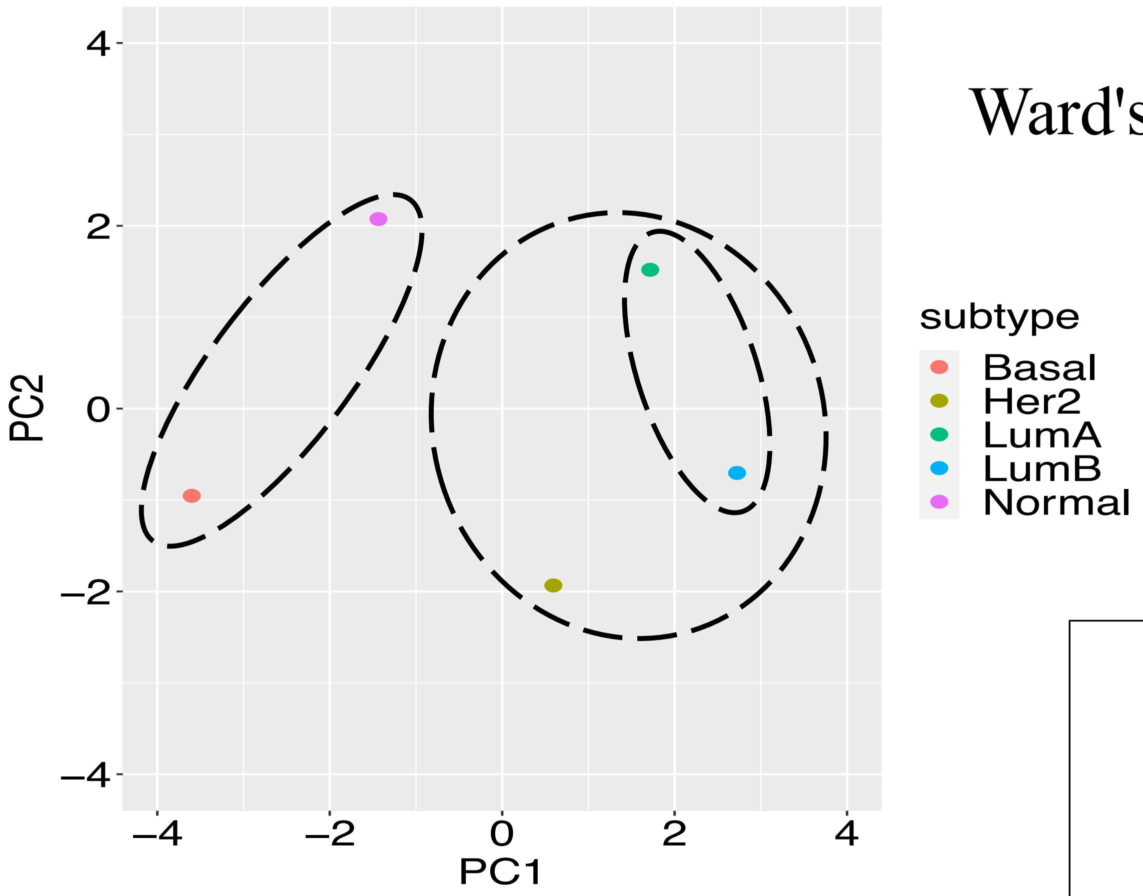


Single-linkage

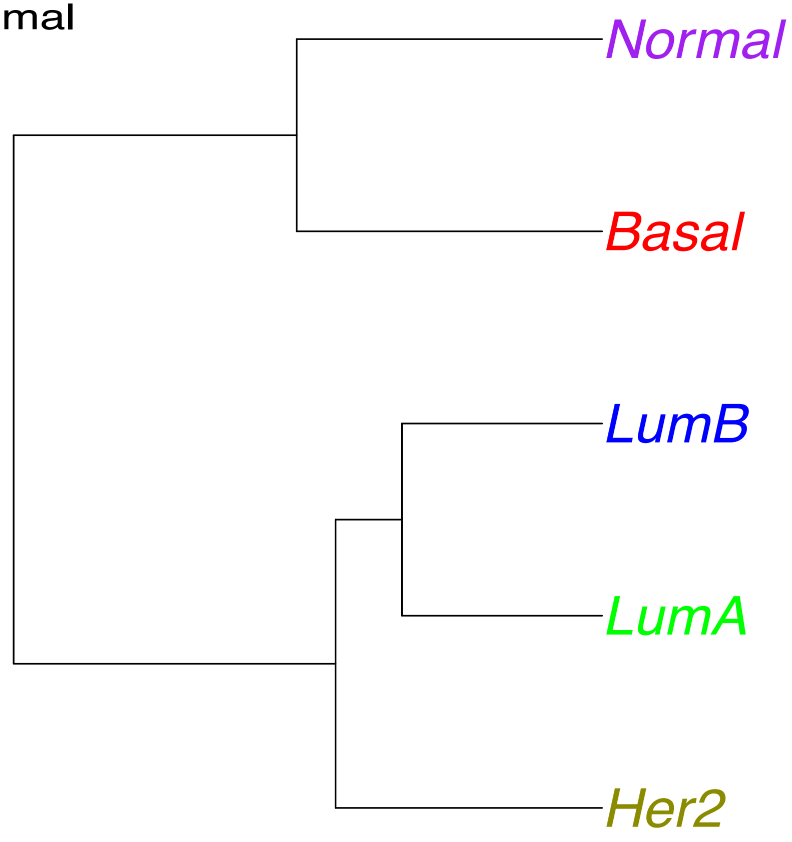
	Basal	Her2	LumA	LumB	Normal
Basal	0	4.31	5.86	**6.33**	**3.72**
Her2	4.31	0	**3.63**	2.46	4.49
LumA	5.86	3.63	0	**2.44**	3.2
LumB	6.33	2.46	2.44	0	5
Normal	3.72	4.49	3.2	5	0

	Basal	Her2	LumA	LumB	Normal
Basal	0	4.31	5.86	**6.33**	3.72
Her2	4.31	0	3.63	**2.46**	4.49
LumA	5.86	3.63	0	**2.44**	**3.2**
LumB	6.33	2.46	2.44	0	5
Normal	3.72	4.49	3.2	5	0

Hierarchical Agglomerative Clustering Analysis

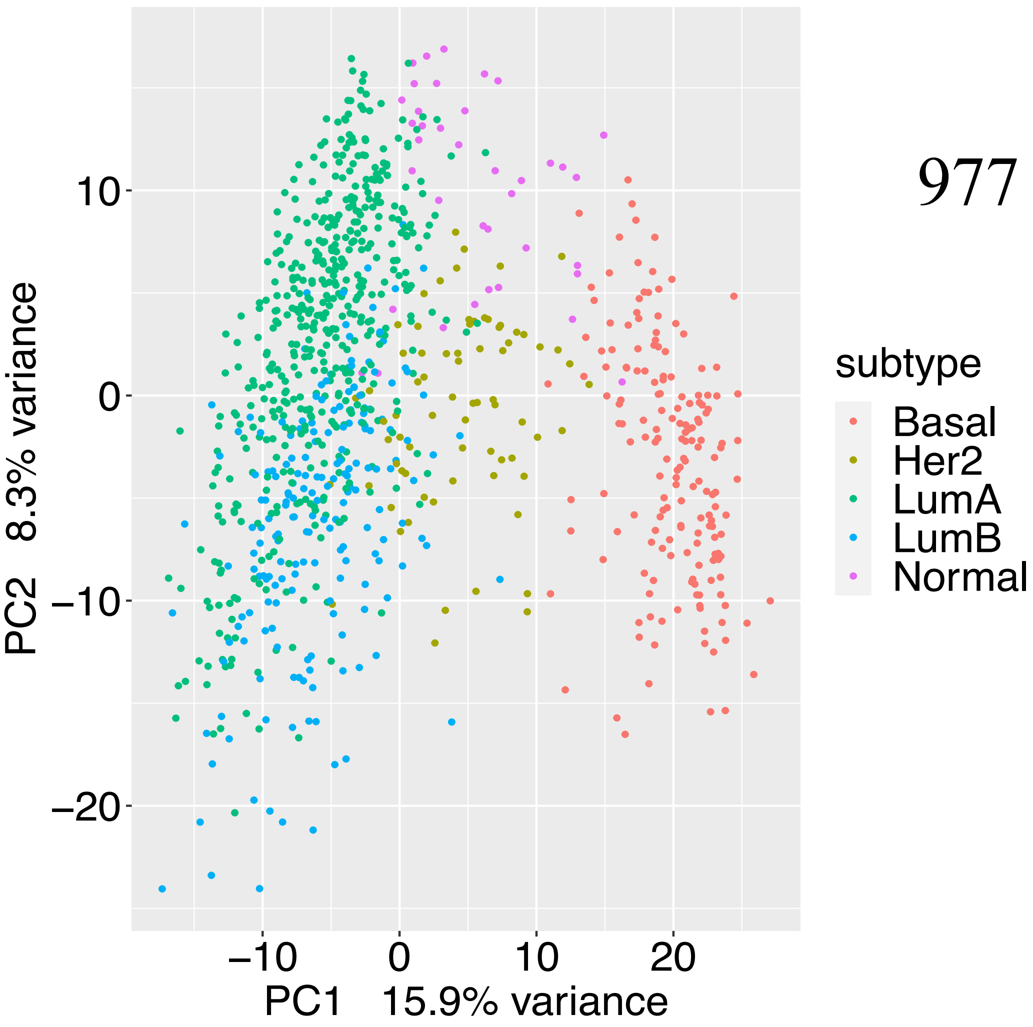


Ward's minimum variance method



PCA Label by Subtype vs. Hierarchical Clustering (Ward)

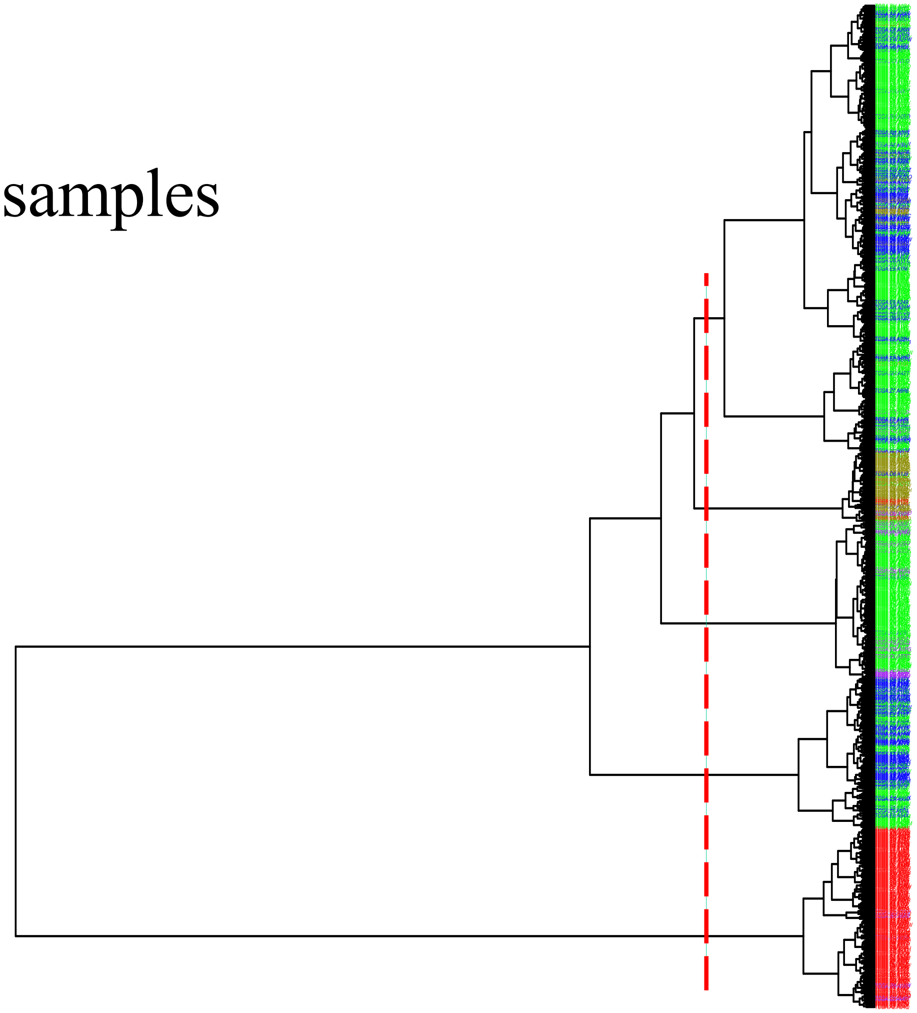
Label by subtype



HC clusters in high-dimension

accuracy 63.7%

977 samples



HD: high dimension, 5000 genes

Comparison Between Subtype and HC vs. k-means Cluster (HD)

5000 genes

HC

	Basal	Her2	LumA	LumB	Normal
clust1	168	0	0	0	7
clust2	5	55	4	2	2
clust3	0	18	**299**	**114**	4
clust4	0	0	71	75	0
clust5	0	0	**126**	2	**25**

Accuracy = (168 + 55 + 299 + 75 + 25) / 977 = 63.7%

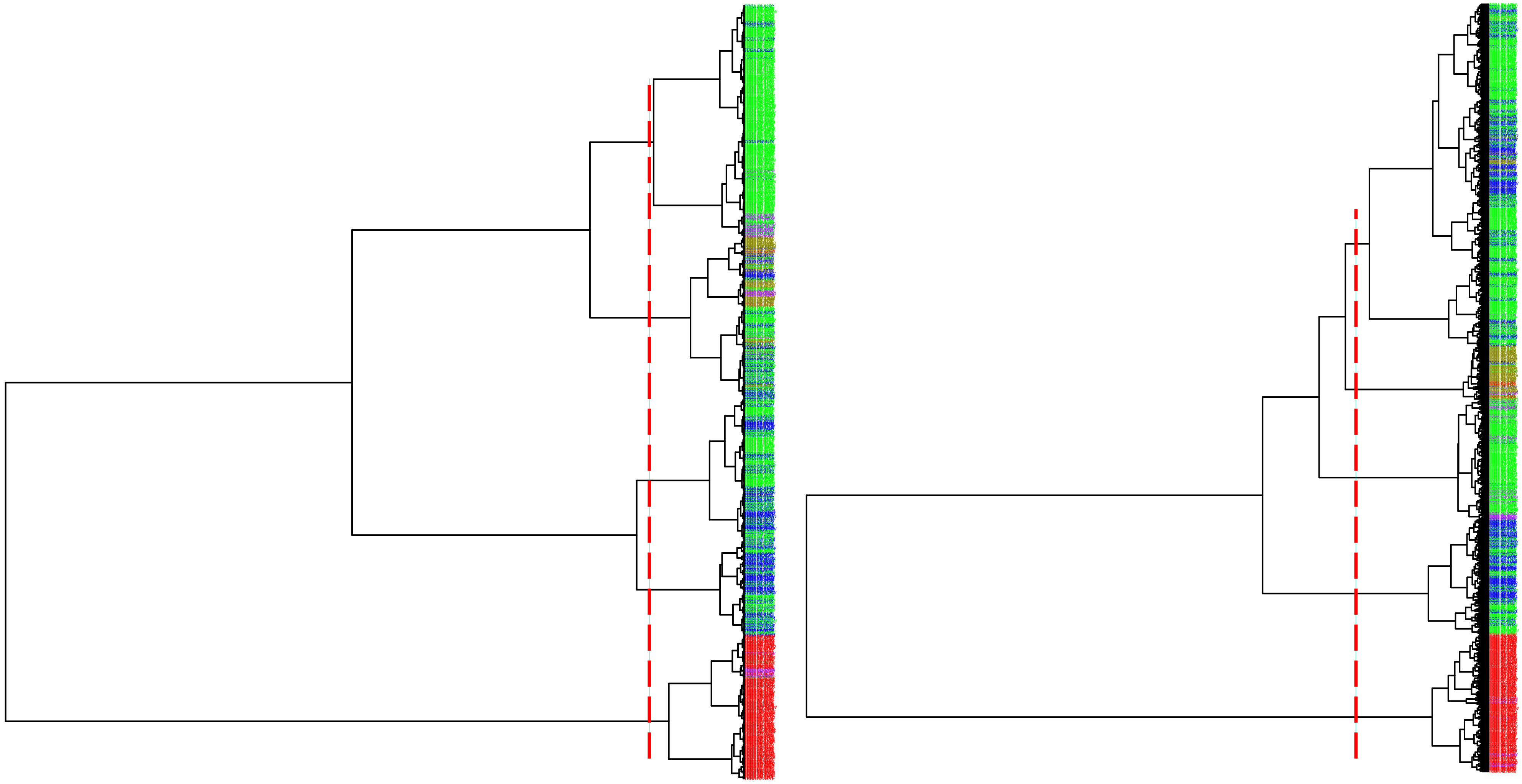
Match
Mismatch

K-means

	Basal	Her2	LumA	LumB	Normal
clust1	169	0	0	0	6
clust2	4	69	17	40	5
clust3	0	0	268	11	21
clust4	0	0	125	**119**	0
clust5	0	4	90	23	6

Accuracy = (169 + 69 + 268 + 119 + 6) / 977 = 64.6%

Hierarchical Clustering (Ward): Low vs. High Dimensions



HC clusters in low-dimension

accuracy 58.4%

HC clusters in high-dimension

accuracy 63.7%

HD: high dimension, 5000 genes

Comparison Between Subtype and HC Low vs. High Dimension

5000 genes

LD

	Basal	Her2	LumA	LumB	Normal
clust1	171	2	0	0	8
clust2	2	**68**	75	49	10
clust3	0	0	263	9	20
clust4	0	1	51	69	0
clust5	0	2	111	66	0

Accuracy = (171 + 68 + 263 + 69) / 977 = 58.4%

Match

Mismatch

HD

	Basal	Her2	LumA	LumB	Normal
clust1	168	0	0	0	7
clust2	5	55	4	2	2
clust3	0	**18**	**299**	**114**	4
clust4	0	0	71	75	0
clust5	0	0	126	2	**25**

Accuracy = (168 + 55 + 299 + 75 + 25) / 977 = 63.7%

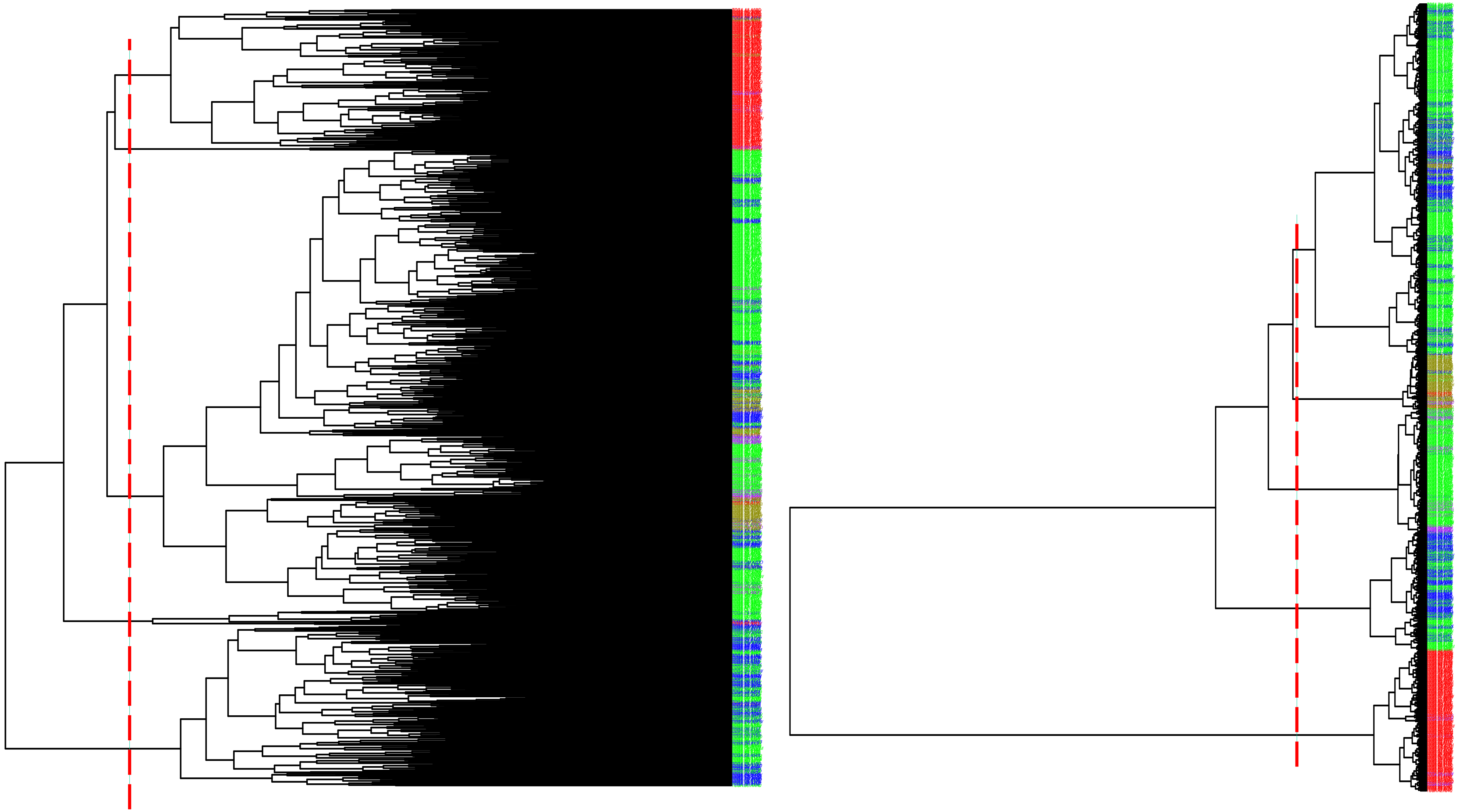
Hierarchical Clustering HD: Complete-Linkage vs. Ward

Complete-linkage

accuracy 66.3%

Ward

accuracy 63.7%



HD: high dimension, 5000 genes

Comparison Between Subtype and HC Ward vs. Complete-Linkage

HD: 5000 genes

	Basal	Her2	LumA	LumB	Normal
clust1	164	4	0	1	4
clust2	2	1	10	3	1
clust3	4	**67**	**385**	92	32
clust4	0	1	**105**	**97**	0
clust5	3	0	0	0	1

complete

Accuracy = (164 + 1 + 385 + 97 + 1) / 977 = 66.3%

Match

Mismatch

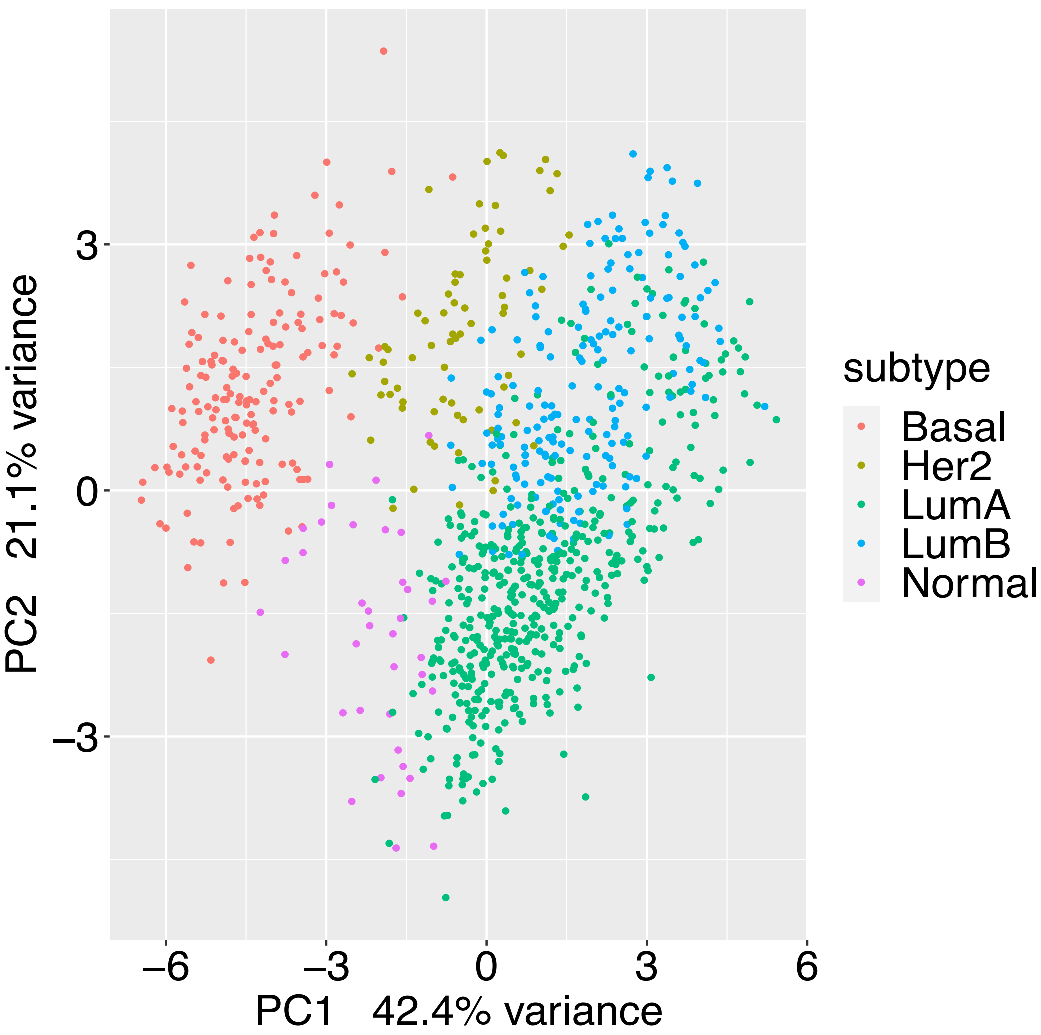
	Basal	Her2	LumA	LumB	Normal
clust1	168	0	0	0	7
clust2	5	**55**	4	2	2
clust3	0	18	299	**114**	4
clust4	0	0	71	75	0
clust5	0	0	126	2	**25**

Ward

Accuracy = (168 + 55 + 299 + 75 + 25) / 977 = 63.7%

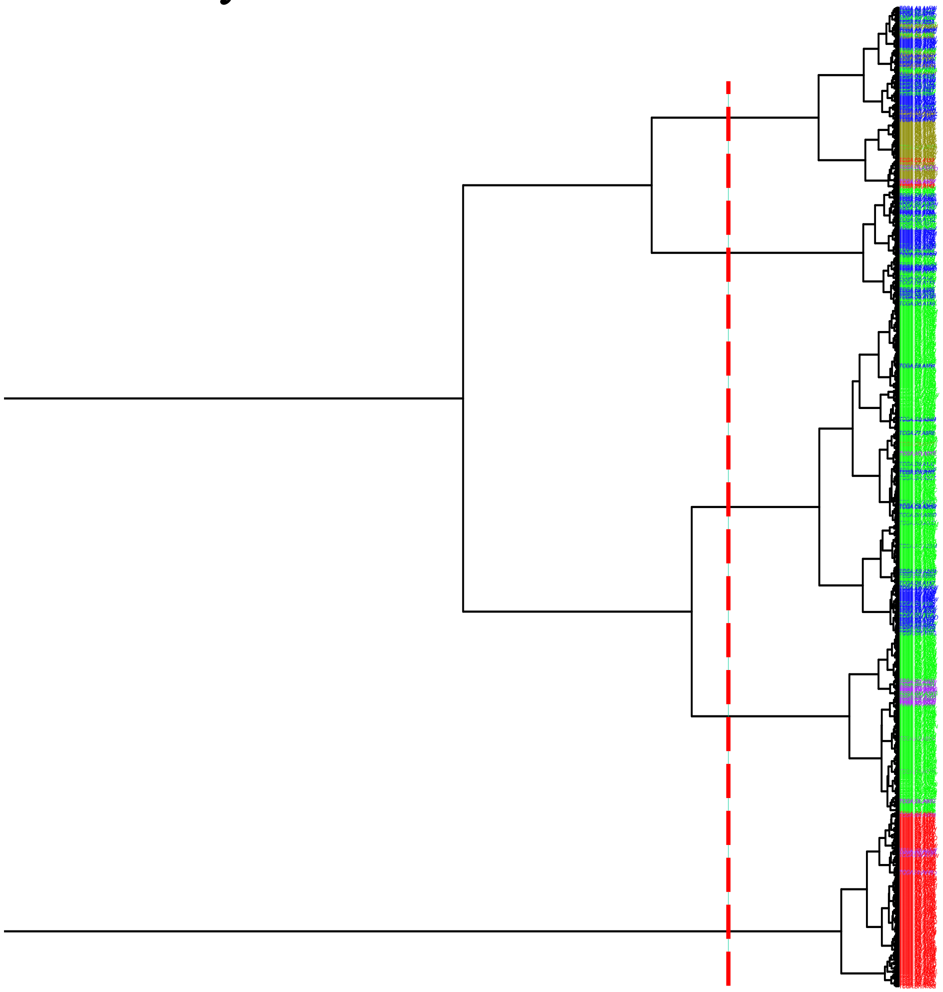
PCA with pam50: Label by Subtype vs. HC (Ward)

Label by subtype



HC clusters in high-dimension

accuracy 54.1%



HD: high dimension, 39 genes

Comparison Between Subtype and HC vs. k-means Cluster (HD)

	Basal	Her2	LumA	LumB	Normal
clust1	167	0	0	0	7
clust2	0	**0**	57	60	0
clust3	0	1	**265**	60	2
clust4	6	**72**	24	73	5
clust5	0	0	154	0	24

60%

HC

Accuracy = (167 + 265 + 73 + 24) / 977 = 54.1%

Match
Mismatch

	Basal	Her2	LumA	LumB	Normal
clust1	170	0	0	0	8
clust2	2	72	11	31	4
clust3	0	0	221	75	0
clust4	1	1	81	87	0
clust5	0	0	**187**	0	**26**

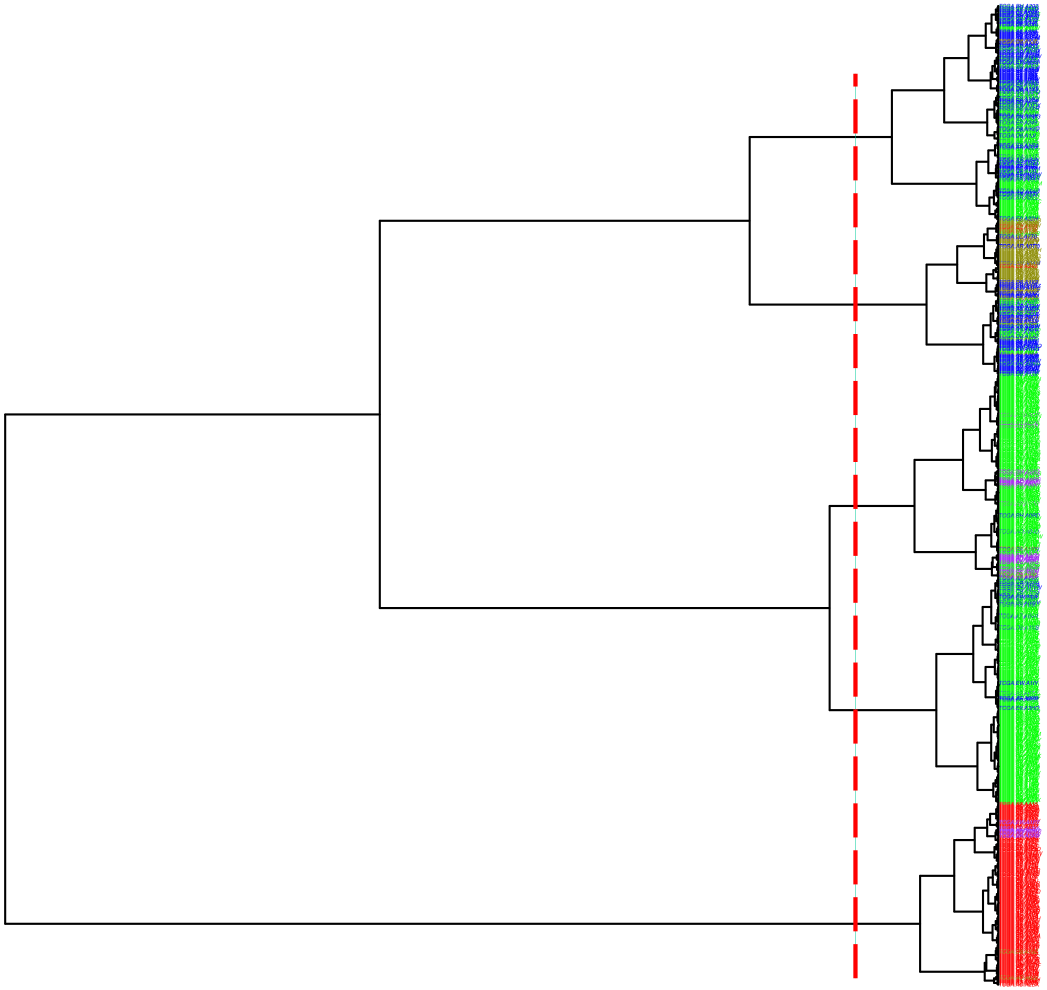
K-means

Accuracy = (170 + 72 + 221 + 87 + 26) / 977 = 59%

Pam50: HC (Ward) Low vs. High Dimension

HC clusters in low-dimension

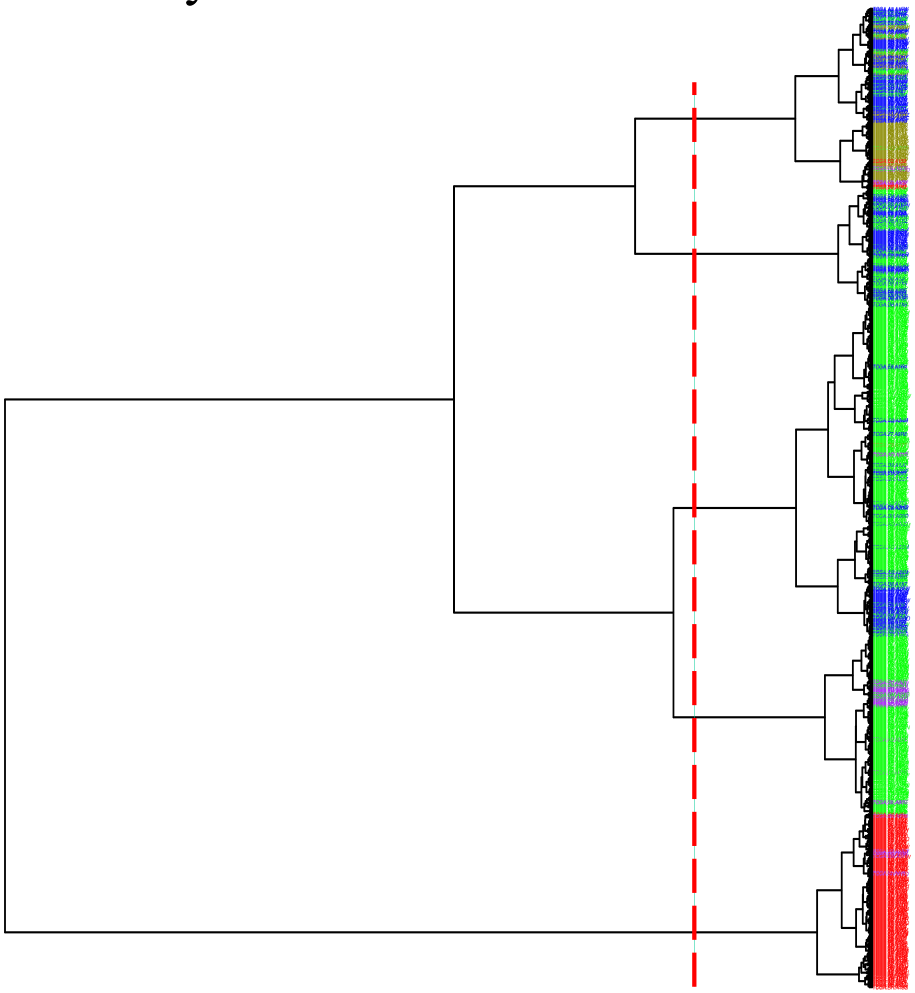
accuracy 58.8%



Low-dimension: 2 PCs

HC clusters in high-dimension

accuracy 54.1%



high dimension: 39 genes

Comparison Between Subtype and HC Low vs. High Dimension

	Basal	Her2	LumA	LumB	Normal
clust1	170	2	0	0	9
clust2	3	**66**	22	62	1
clust3	0	0	204	22	0
clust4	0	2	**106**	**106**	0
clust5	0	3	168	3	28

LD

Accuracy = (172 + 58 + 356+ 91) / 977 = 58.8%

Match

Mismatch

	Basal	Her2	LumA	LumB	Normal
clust1	167	0	0	0	7
clust2	0	**0**	57	60	0
clust3	0	1	**265**	**60**	2
clust4	6	**72**	24	73	5
clust5	0	0	154	0	24

HD

Accuracy = (167 + 265 + 73 + 24) / 977 = 54.1%

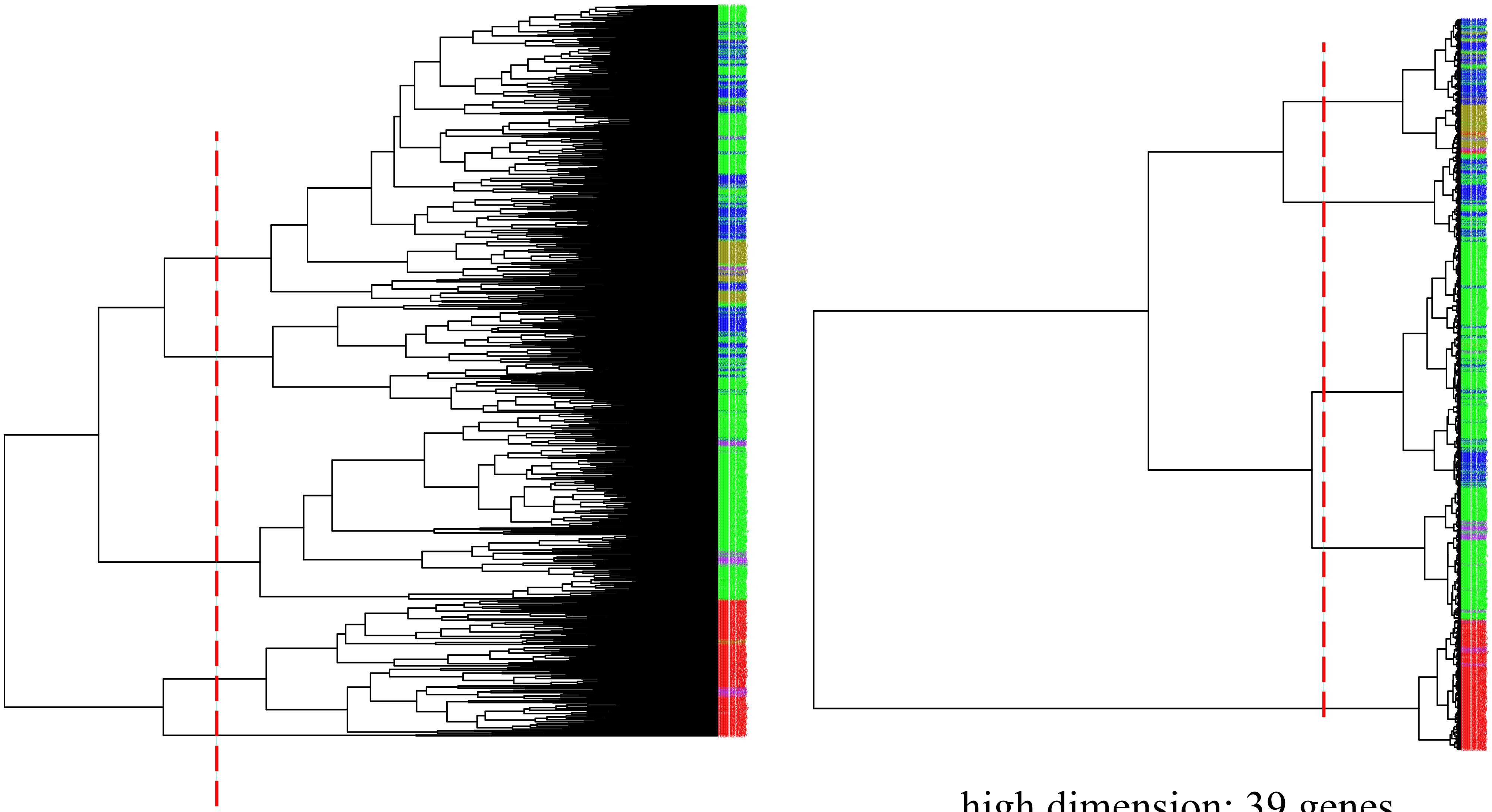
Pam50: HC High Dimension Complete-Linkage vs. Ward

HC complete-linkage

accuracy 54.2%

HC Ward

accuracy 54.1%



high dimension: 39 genes

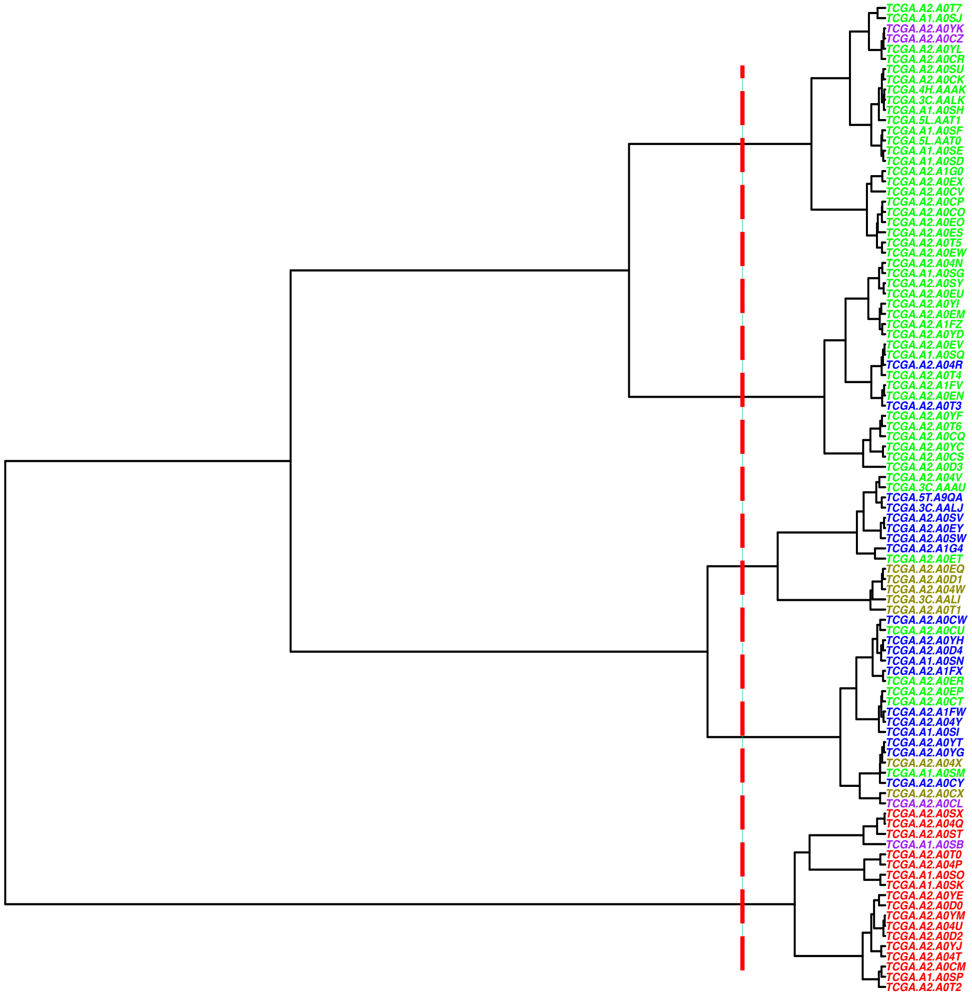
10% TCGA Samples (n=97) with Pam50

subtype	size
Basal	17
Her2	7
LumA	50
LumB	19
Normal	4

Pam50 (n=97): HC (Ward) Low vs. High Dimension

HC clusters in low-dimension

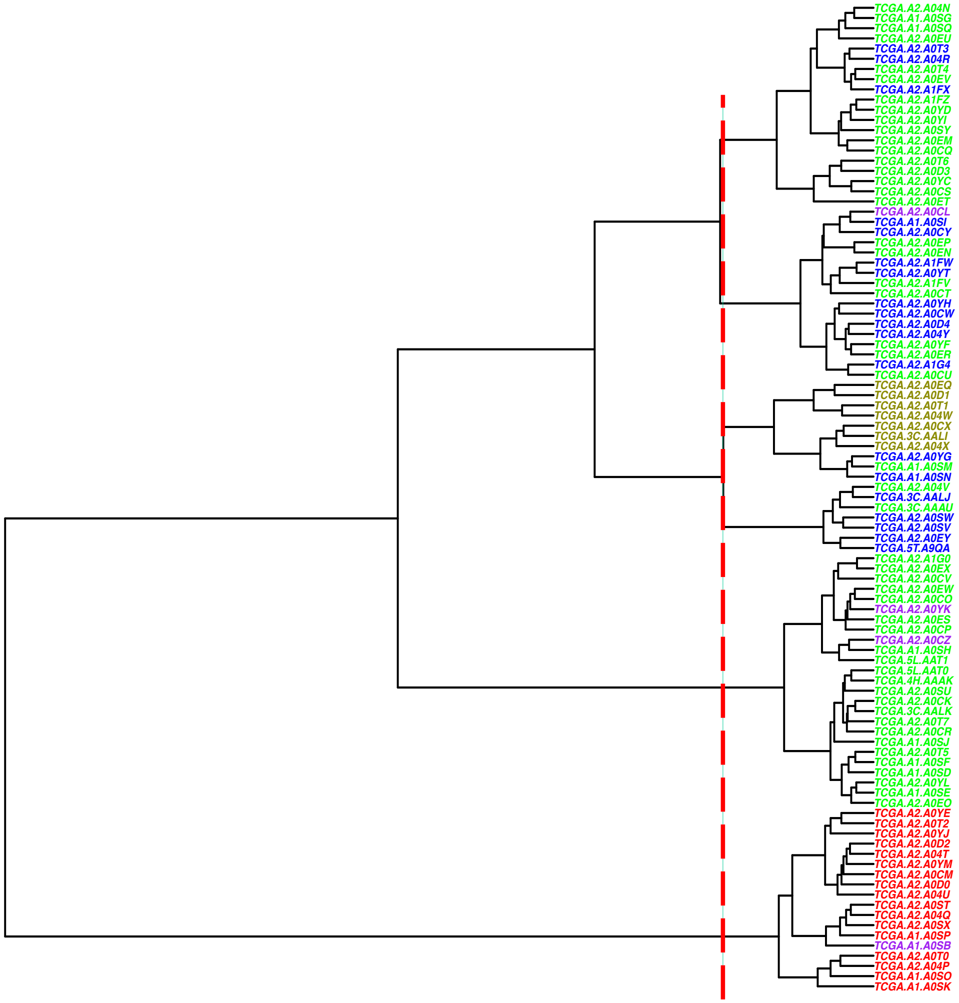
accuracy 57.7%



Low-dimension: 2 PCs

HC clusters in high-dimension

accuracy 57.7%



high dimension: 39 genes

Subtype and HC (Ward) Low vs. High Dimension

97 samples

LD

	Basal	Her2	LumA	LumB	Normal
clust1	17	0	0	0	1
clust2	0	5	3	6	0
clust3	0	0	23	0	2
clust4	0	**2**	5	**11**	1
clust5	0	0	19	2	0

Accuracy = (17 + 5 +23 + 11) / 97 = 57.7%

Match

Mismatch

HD: 39 genes

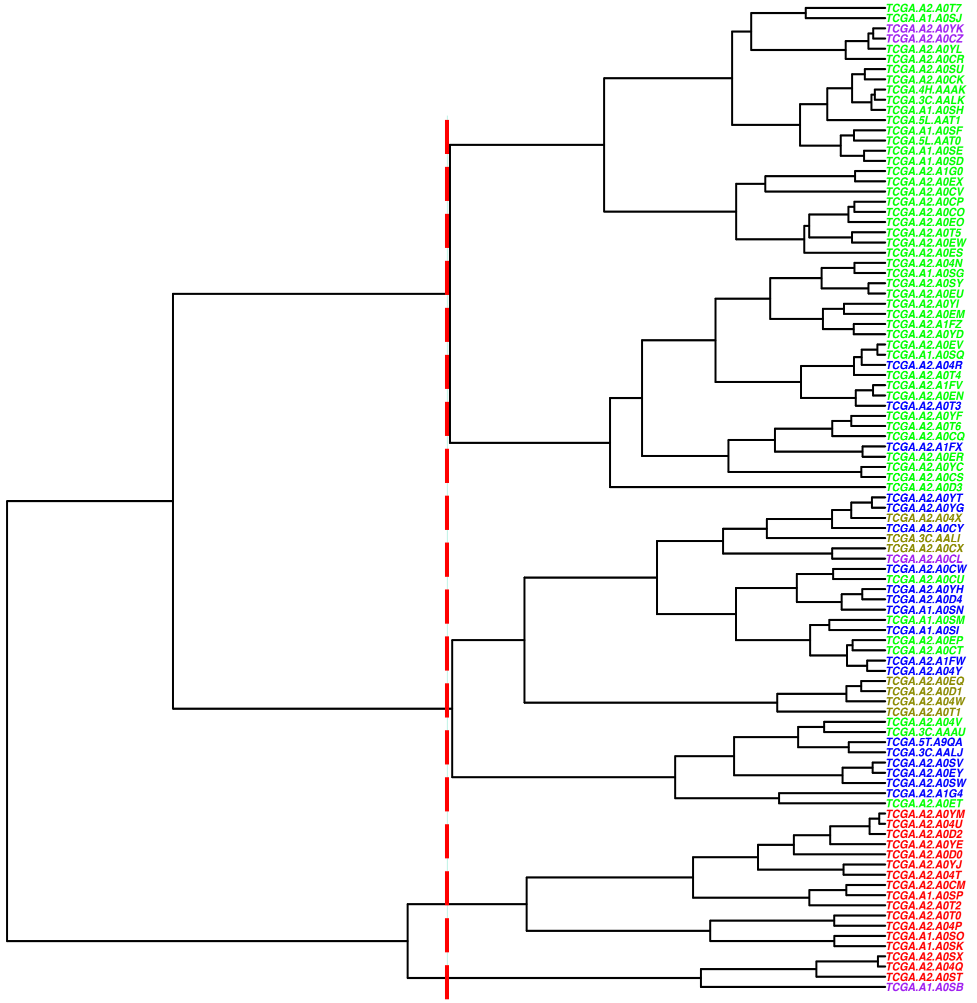
	Basal	Her2	LumA	LumB	Normal
clust1	17	0	0	0	1
clust2	0	**7**	3	7	0
clust3	0	0	23	0	2
clust4	0	0	**7**	9	1
clust5	0	0	17	3	0

Accuracy = (17 + 7 + 23 + 9) / 97 = 57.7%

Pam50 (n=97): HC (Complete-Linkage) Low vs. High Dimension

HC clusters in low-dimension

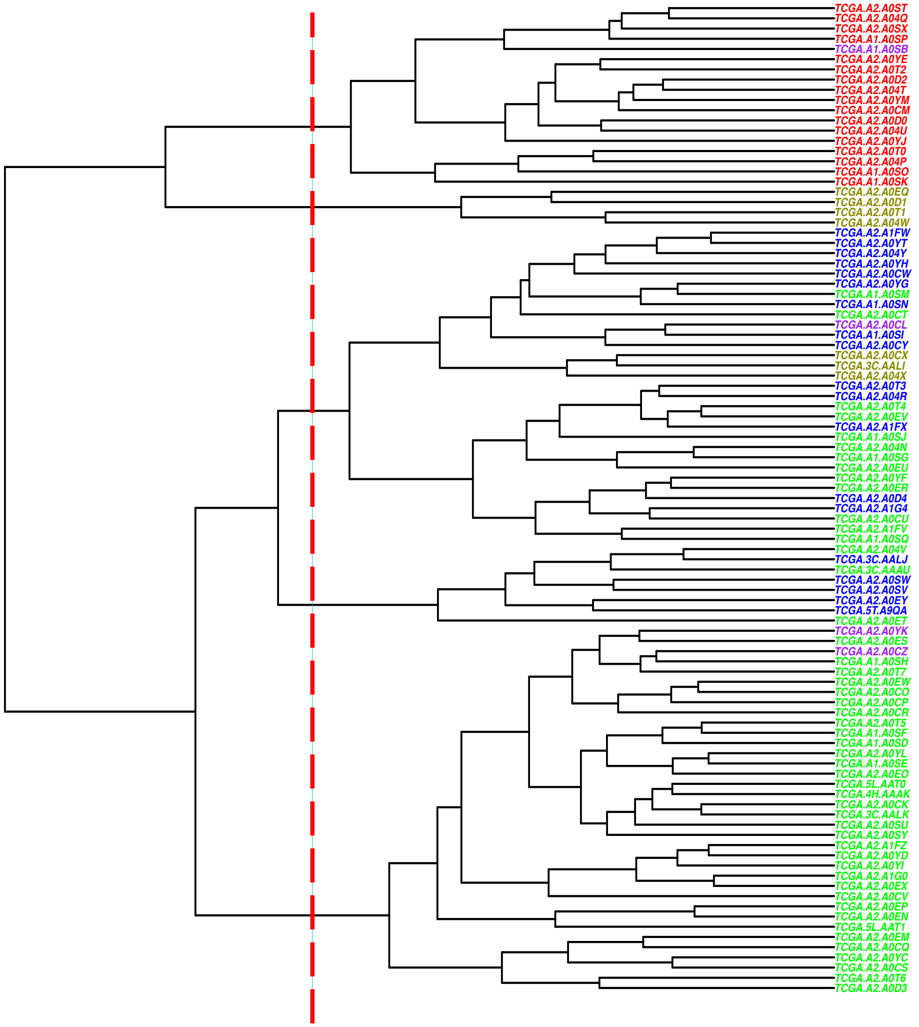
accuracy 55.7%



Low-dimension: 2 PCs

HC clusters in high-dimension

accuracy 71.1%



high dimension: 39 genes

Subtype and HC (Complete-Linkage) Low vs. High Dimension

97 samples

LD

	Basal	Her2	LumA	LumB	Normal
clust1	14	0	0	0	0
clust2	0	0	**20**	3	0
clust3	0	0	23	0	2
clust4	0	**7**	7	16	1
clust5	3	0	0	0	1

Accuracy = (14 + 23 + 16 + 1) / 97 = 55.7%

Match
Mismatch

	Basal	Her2	LumA	LumB	Normal
clust1	17	0	0	0	1
clust2	0	**4**	0	0	0
clust3	0	0	**34**	0	2
clust4	0	3	**13**	14	1
clust5	0	0	3	5	0

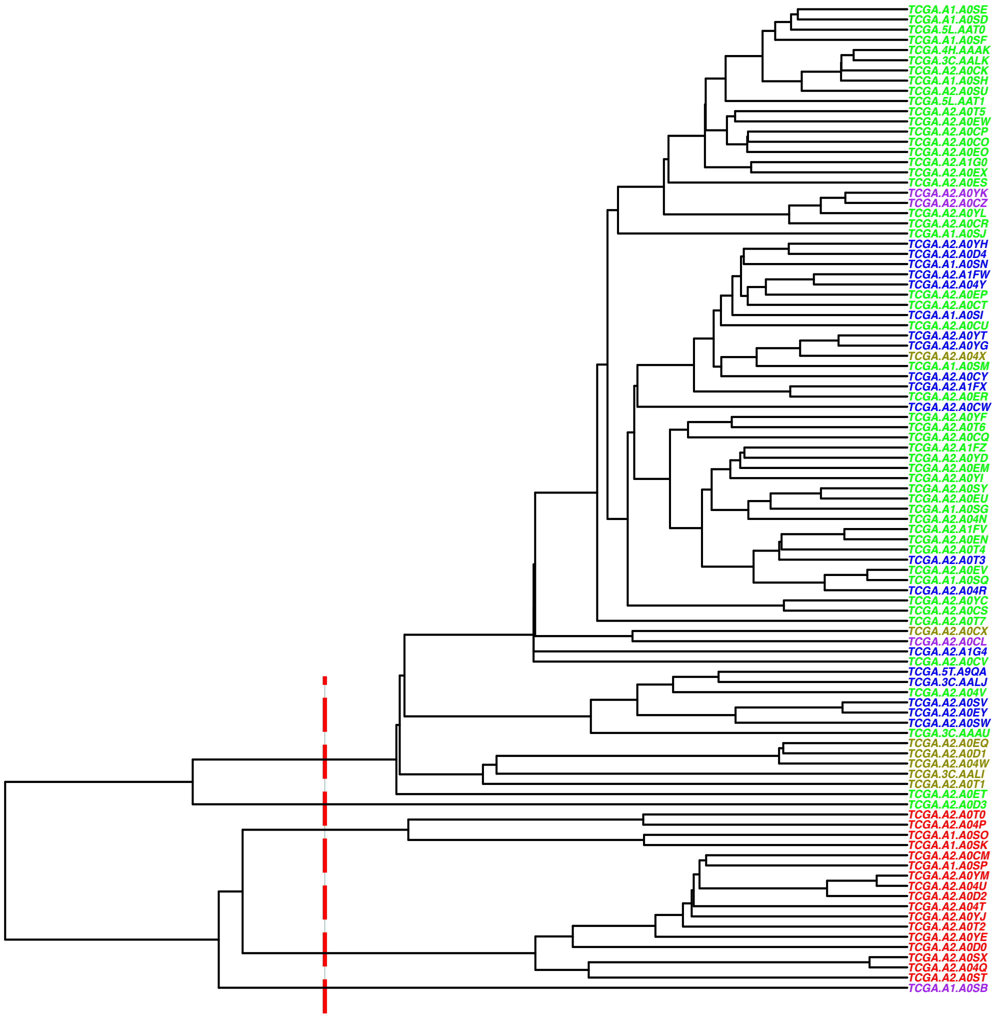
HD: 39 genes

Accuracy = (17 + 4 + 34 + 14) / 97 = 71.3%

Pam50 (n=97): HC (Single-Linkage) Low vs. High Dimension

HC clusters in low-dimension

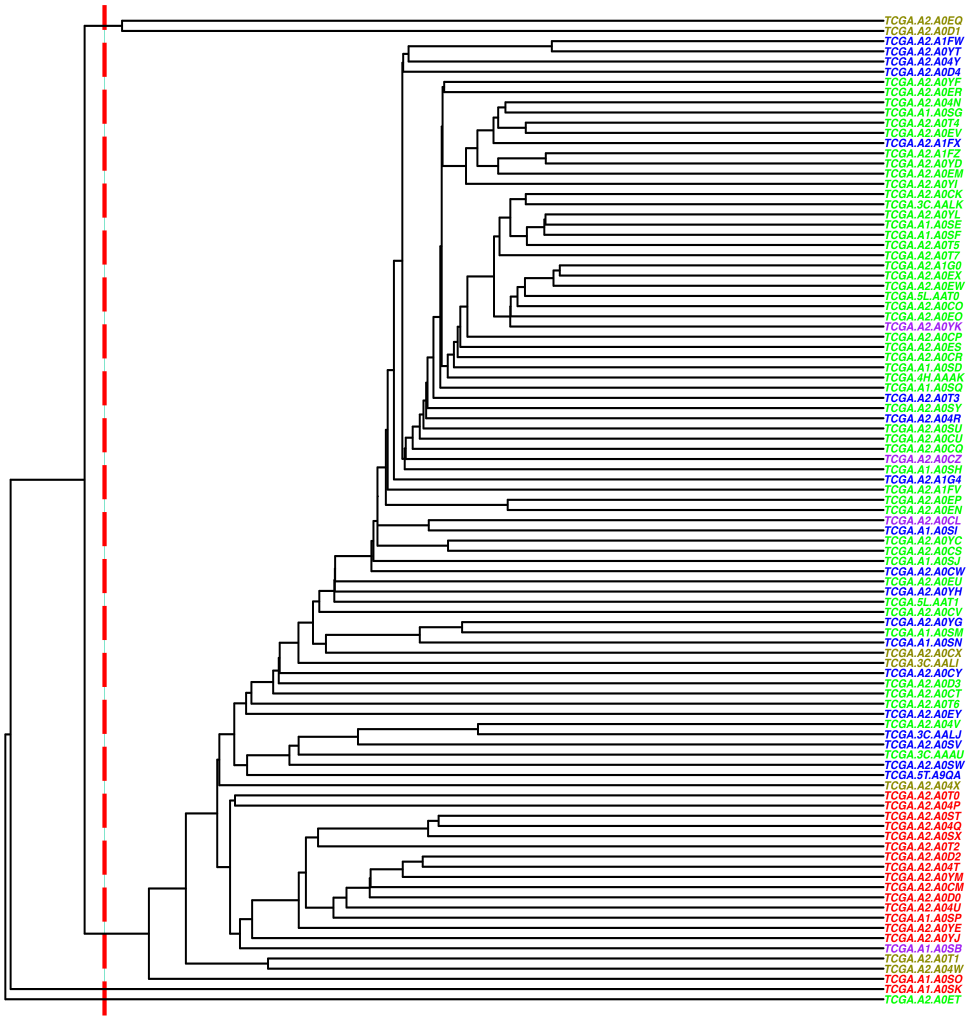
accuracy 64.9%



Low-dimension: 2 PCs

HC clusters in high-dimension

accuracy 52.6%



high dimension: 39 genes

Subtype and HC (Single-Linkage) Low vs. High Dimension

97 samples

LD

	Basal	Her2	LumA	LumB	Normal
clust1	**13**	0	0	0	0
clust2	4	0	0	0	0
clust3	0	7	**49**	19	3
clust4	0	0	1	0	0
clust5	0	0	0	0	1

Accuracy = (13 + 49 + 1) / 97 = 64.9%

Match

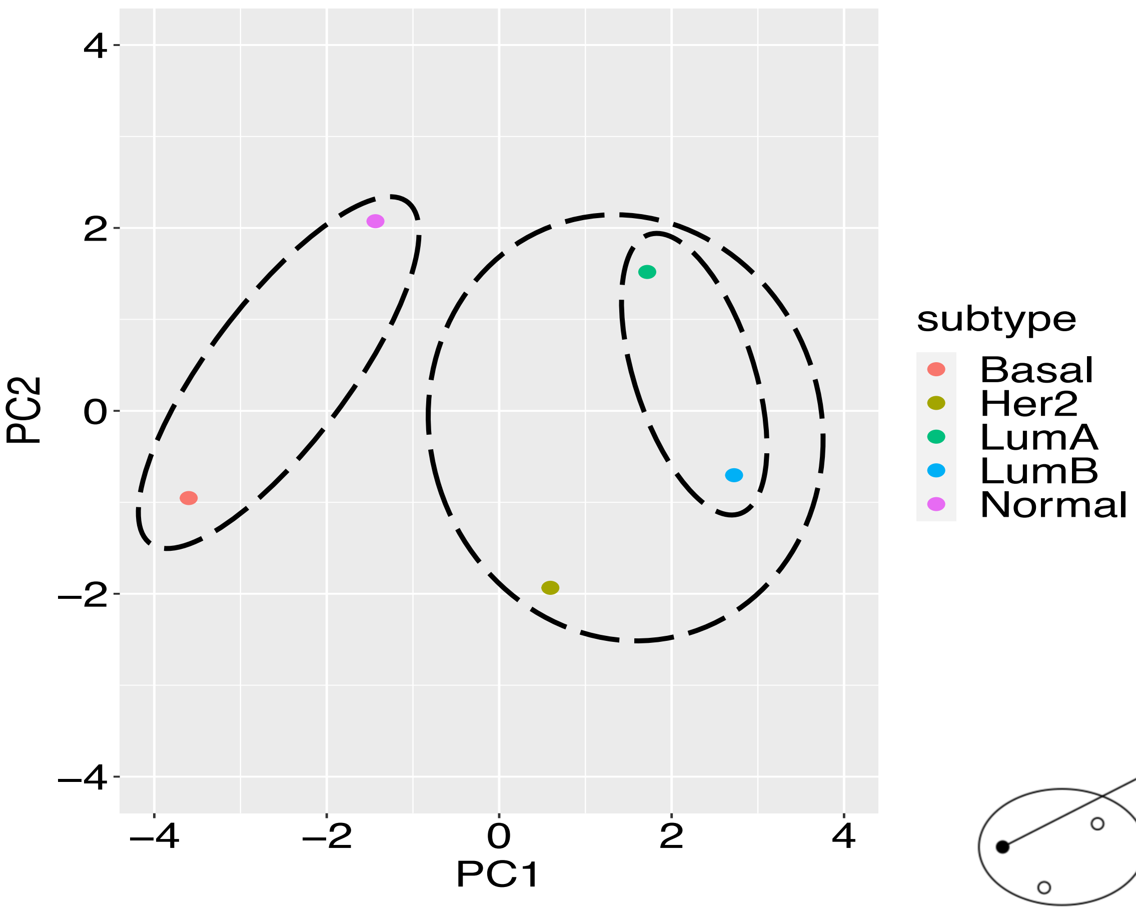
Mismatch

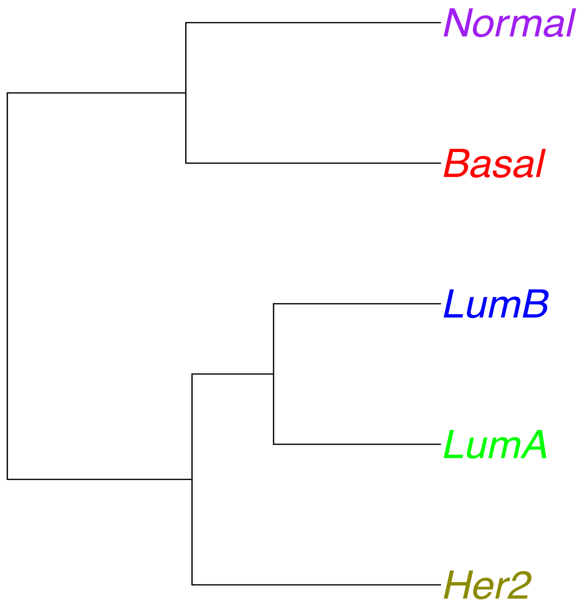
HD: 39 genes

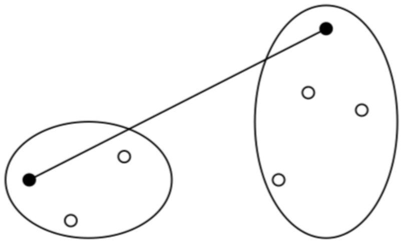
	Basal	Her2	LumA	LumB	Normal
clust1	1	0	0	0	0
clust2	0	1	0	0	0
clust3	**16**	5	49	19	4
clust4	0	1	0	0	0
clust5	0	0	1	0	0

Accuracy = (1 + 1 + 49) / 97 = 52.6%

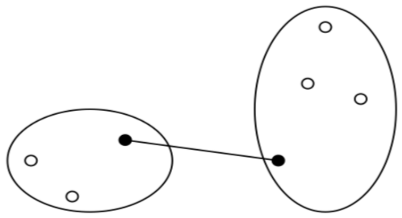
Hierarchical Agglomerative Clustering Algorithm







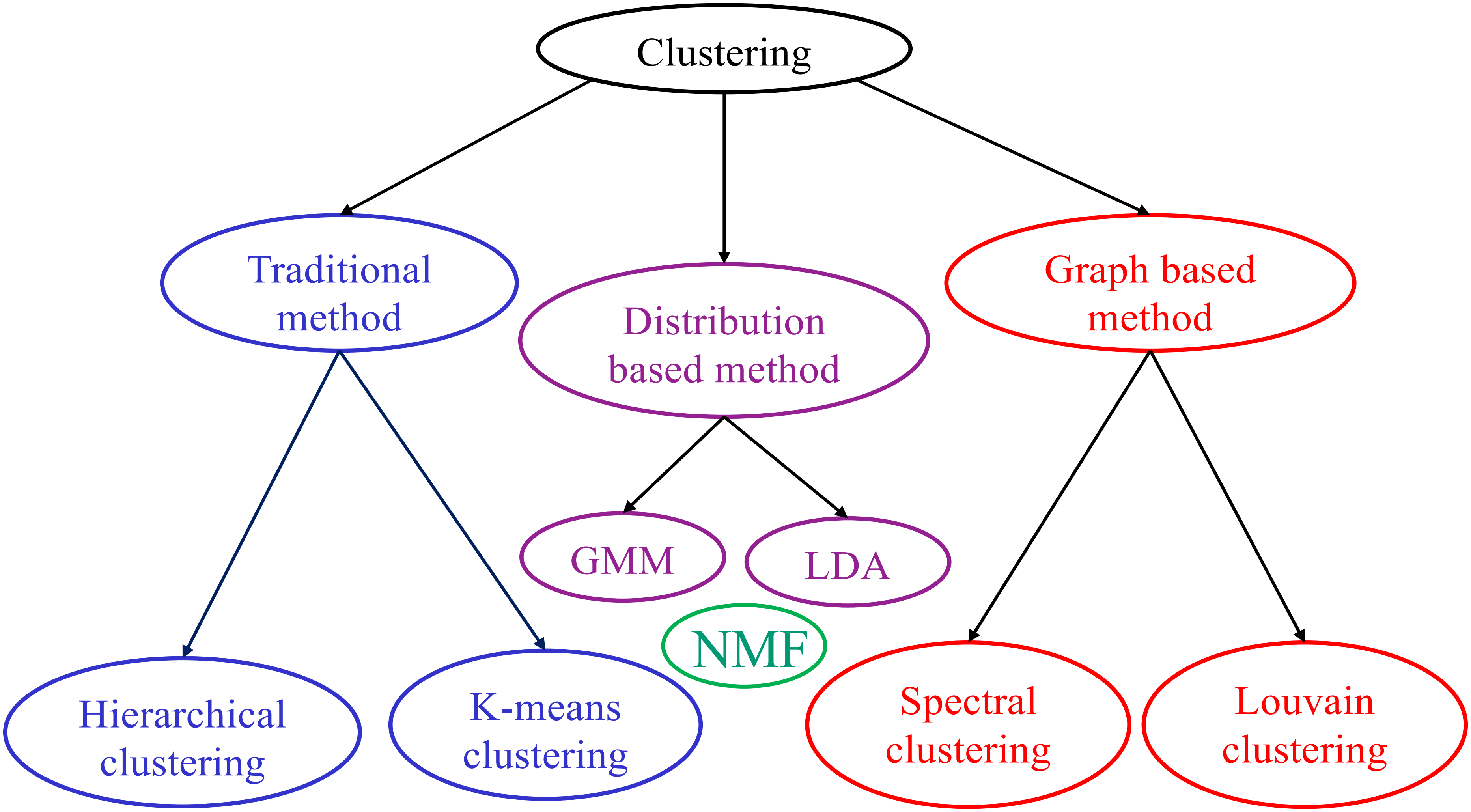
complete-linkage



single-linkage

- Start with n leaf nodes
- Sequentially merge a pair of nodes with the smallest distance or minimal variance
- End with a single cluster

Outline of Clustering Methods



GMM: Gaussian Mixture Model
LDA: Latent Dirichlet Allocation
NMF: Non-negative matrix factorization

Contributed by Emily Tai