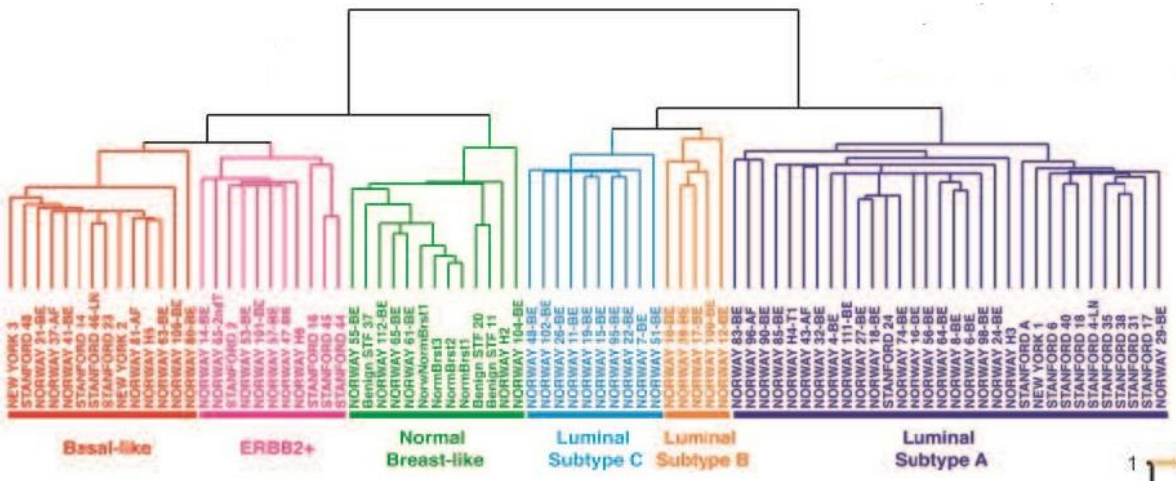
Memorial Sloan Kettering Cancer Center™

COLUMBIA UNIVERSITY | MAILMAN SCHOOL of PUBLIC HEALTH
BIOSTATISTICS

Application of a method for identifying disease subtypes that are etiologically heterogeneous

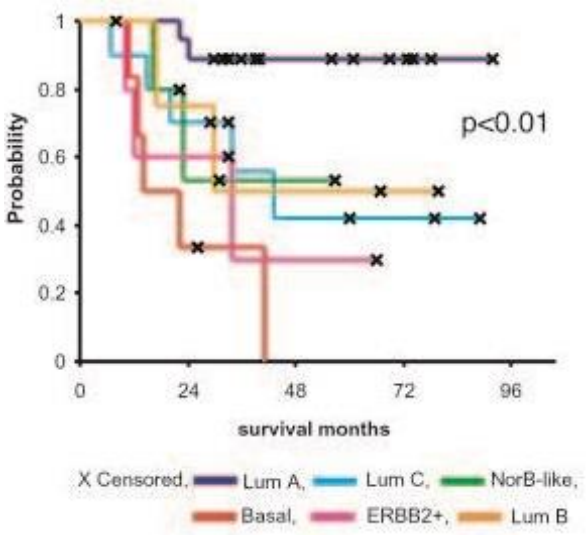
Emily C. Zabor, Shuang Wang, Colin B. Begg

Breast cancer is biologically diverse and subtypes of disease risk and prognosis are recognized



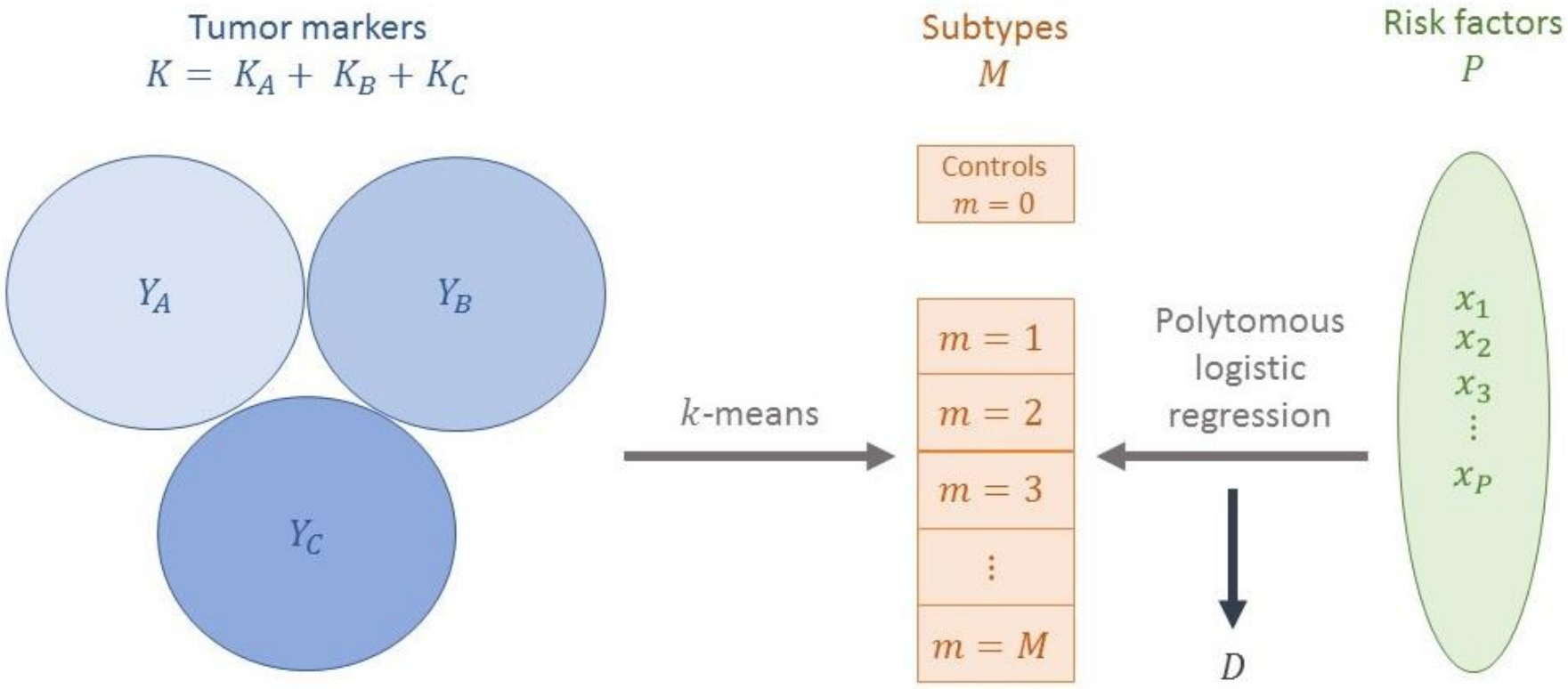
Primary molecular subtypes of breast cancer can be approximated with 4 classes based on 3 IHC markers:

- Estrogen receptor (ER)
- Progesterone receptor (PR)
- Human epidermal growth receptor (Her2)



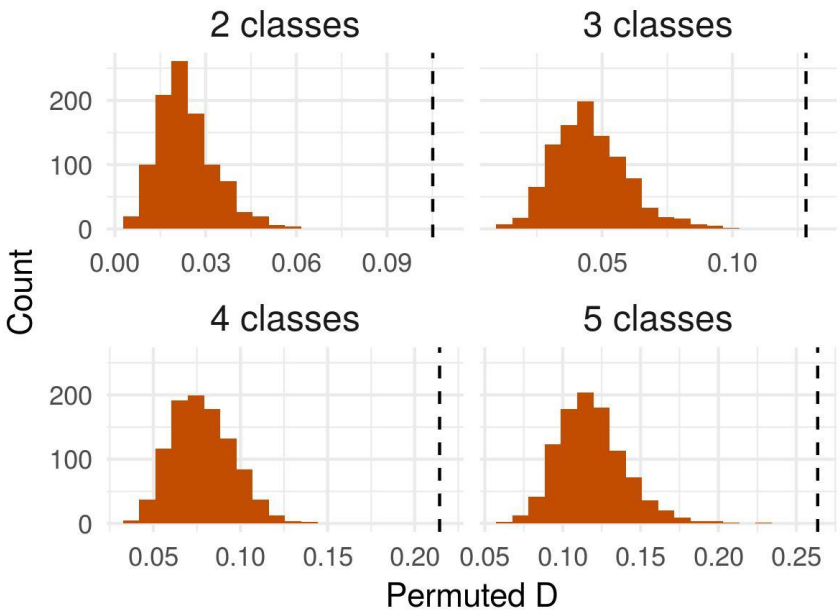
Sorlie et al (2001). Gene Expression Patterns of Breast Carcinomas Distinguish Tumor Subclasses with Clinical Implications. PNAS (19): 10869-74.

Seek to cluster tumor marker data and optimize a scalar measure of etiologic heterogeneity, D



For three subtypes A, B, and C: $D = (\pi_A V_A^2 + \pi_B V_B^2 + \pi_C V_C^2) - V^2$,
where π_j is the relative frequency and V_j^2 is the coefficient of variation for subtype j.

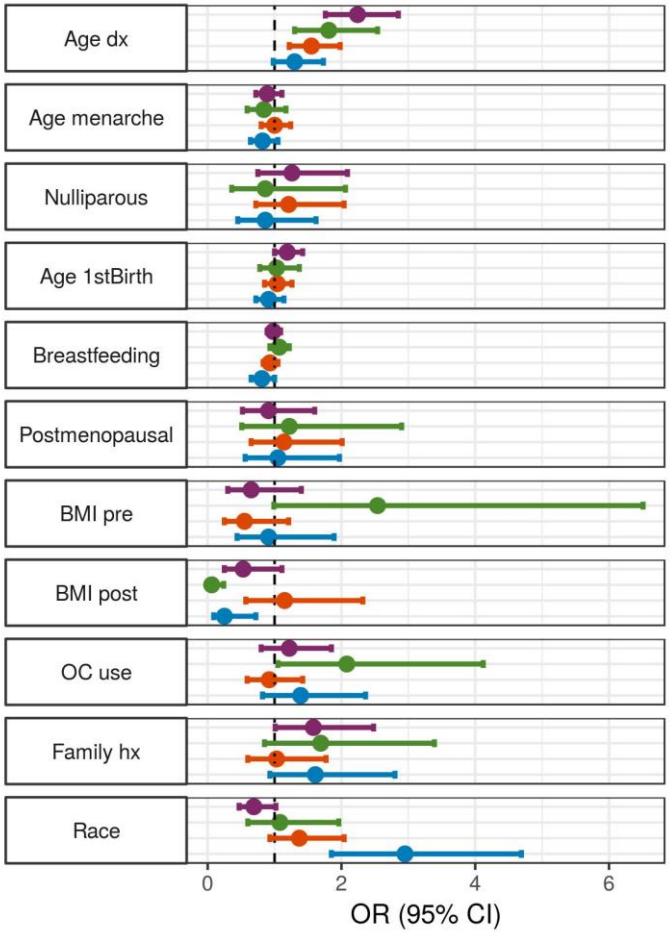
Discovery analysis selects 112 genes and identifies 4 classes



Class size	D difference	P-value
3 vs 2	0.023	0.454
4 vs 2	0.109	0.002
5 vs 2	0.159	0.012
4 vs 3	0.086	0.007
5 vs 3	0.135	0.015
5 vs 4	0.049	0.358

- 532 breast cancer cases from Carolina Breast Cancer Study used for discovery
- Top-ranked genes selected for inclusion in clustering based on individual gene D

Age at diagnosis, menopausal status, and race are driving the heterogeneity across subtypes



Class size	D
Discovery	0.214
Validation	0.245
Traditional IHC 4-class	0.148

- Validation produces reasonably similar gene rankings
- Optimal 4-class D exceeds that in traditional IHC 4-class
- ER is a key gene in distinguishing subtypes