Prognostic Expression and Methylation Signatures Partition Luminal-A Breast Tumors Into Clinically Distinct Subgroups

Dvir Netanely[a], Ayelet Avraham[b], Adit Ben-Baruch[c], Ella Evron[b], Ron Shamir[a]

[a]Blavatnik School of Computer Science, Tel Aviv University, Tel Aviv, Israel

[b]Oncology Department, Assaf Harofeh Medical Center, Tsrifin, Israel

[c]Department of Cell Research and Immunology, George S. Wise Faculty of Life Sciences, Tel Aviv University, Tel Aviv, Israel

Breast Cancer

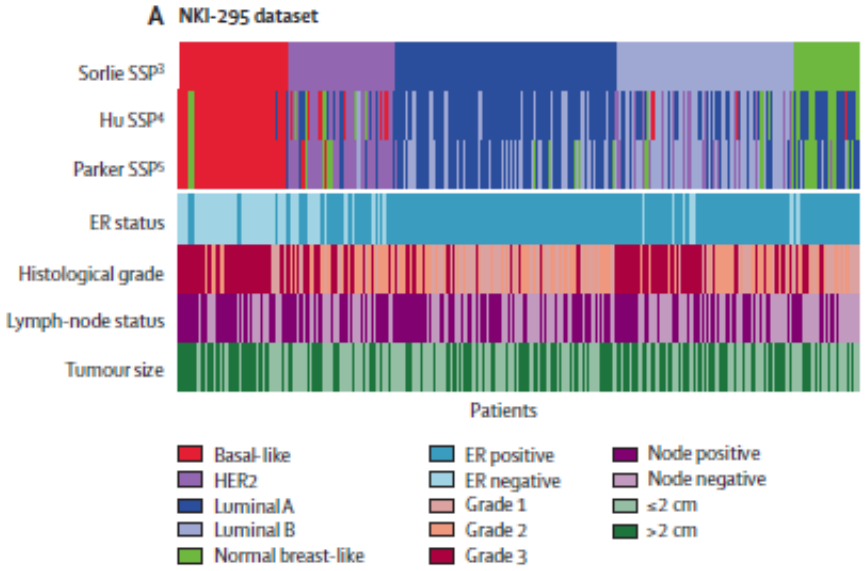
- Most common cancer among women, ranks among the leading causes of cancer-related deaths.
- A heterogeneous disease with different subtypes showing distinct biological and clinical features.
- Prognosis of breast cancer patients has been improving over time with the development of subtype specific treatments
  - Tamoxifen for patients with hormone receptor-positive tumors
  - Trastuzumab (Herceptin) for patients displaying overexpression and amplification of the HER-2 oncogene

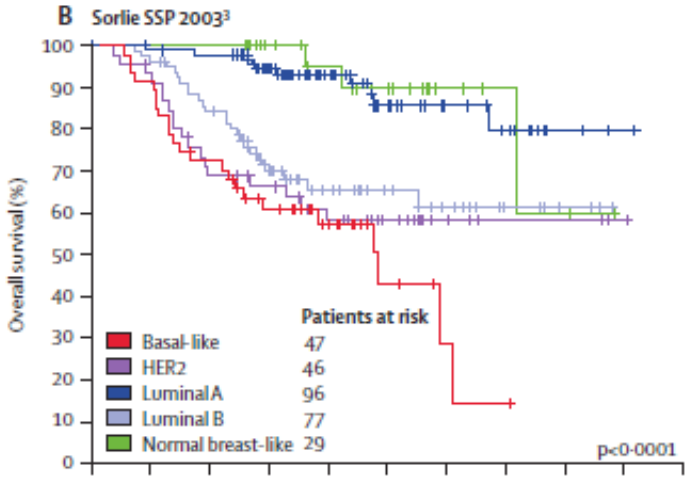
The importance of accurate subtype identification

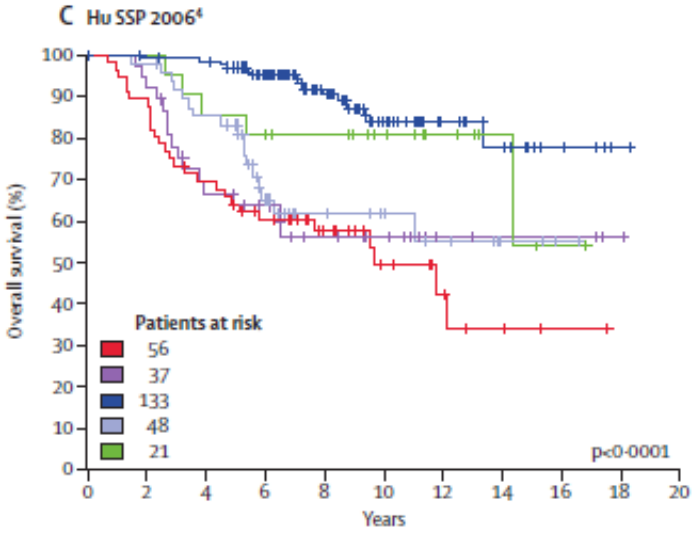
- An important problem in breast cancer treatment is the definition of patient subsets that will require aggressive treatment options and close follow-up after treatment
- A major milestone on the way to this goal was the definition of five biologically and clinically meaningful breast cancer subtypes based on genome-wide expression analyses:
  - Luminal-A
  - Luminal-B
  - HER-2
  - Basal-like (Triple Negative: ER-, PR-, Her2-)
  - Normal-like

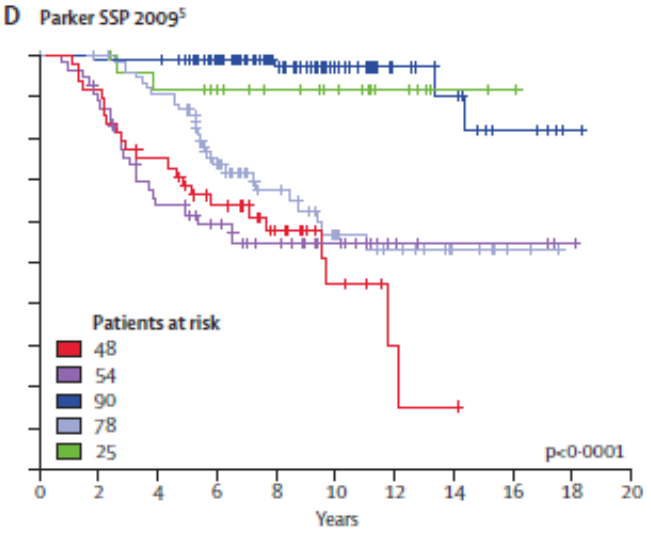
PAM50 and other SSPs

Weigelt, B., Mackay, A., A'hern, R., Natrajan, R., Tan, D. S. P., Dowsett, M., … Reis-Filho, J. S. (2010). Breast cancer molecular profiling with single sample predictors: a retrospective analysis. *The Lancet. Oncology*, *11*(4), 339–49. doi:10.1016/S1470-2045(10)70008-5









Epigenetics and breast cancer subtypes

- Molecular profiling of breast cancer subtypes have so far focused mainly on the expression level.
- Less is known about the contribution of epigenetic changes to the development of biologically distinct breast cancer subtypes

CpG sites and CpG islands

- **CpG sites** are regions of DNA where a cytosine nucleotide occurs next to a guanine nucleotide
- Cytosines in CpG dinucleotides can be methylated to form 5-methylcytosine.
- In mammals, methylating the cytosine within a gene can turn the gene off, a mechanism that is part of a larger field of science studying gene regulation that is called **epigenetics**.

REMINDER

DNA Methylation and gene expression

- Methylation of CpG sites in the promoter of a gene may inhibit gene expression
- Most of the methylation differences between tissues, or between normal and cancer samples, occur a short distance from the CpG islands (at "CpG island shores") rather than in the islands themselves

Infinium HumanMethylation450 BeadChip Kit

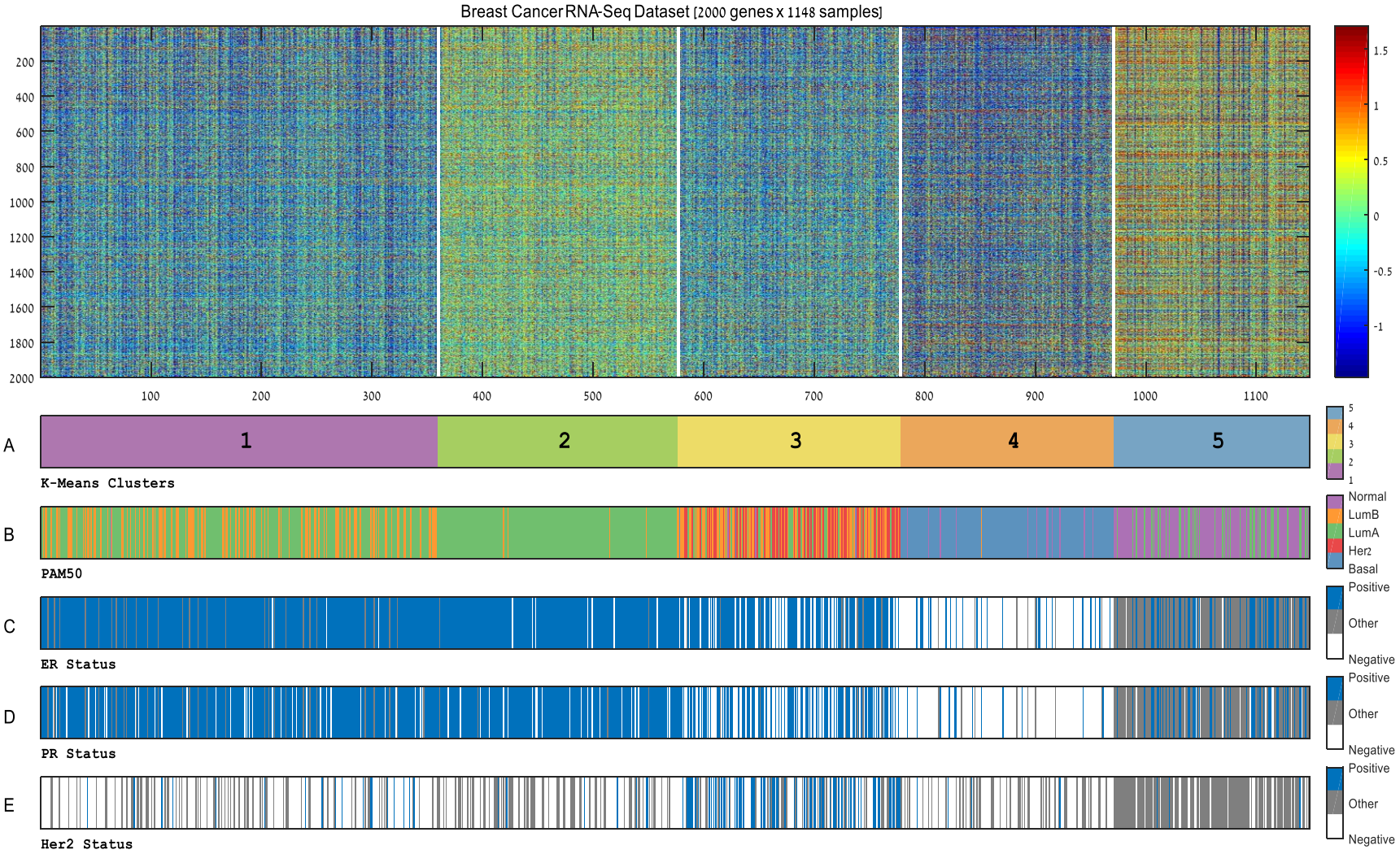
- Allows researchers to interrogate > 485,000 methylation sites per sample at single-nucleotide resolution
- Covers 99% of RefSeq genes, with an average of 17 CpG sites per gene region distributed across the promoter, 5'UTR, first exon, gene body, and 3'UTR.
- It covers 96% of CpG islands, with additional coverage in island shores and the regions flanking them.
- Methylation level of a CpG locus is estimated using beta values (β) which are the ratio of intensities between methylated and unmethylated alleles (rang: 0-1).

GOAL

- **Goal:** Revisit breast cancer classification based on large number of expression and methylation breast cancer profiles obtained from TCGA

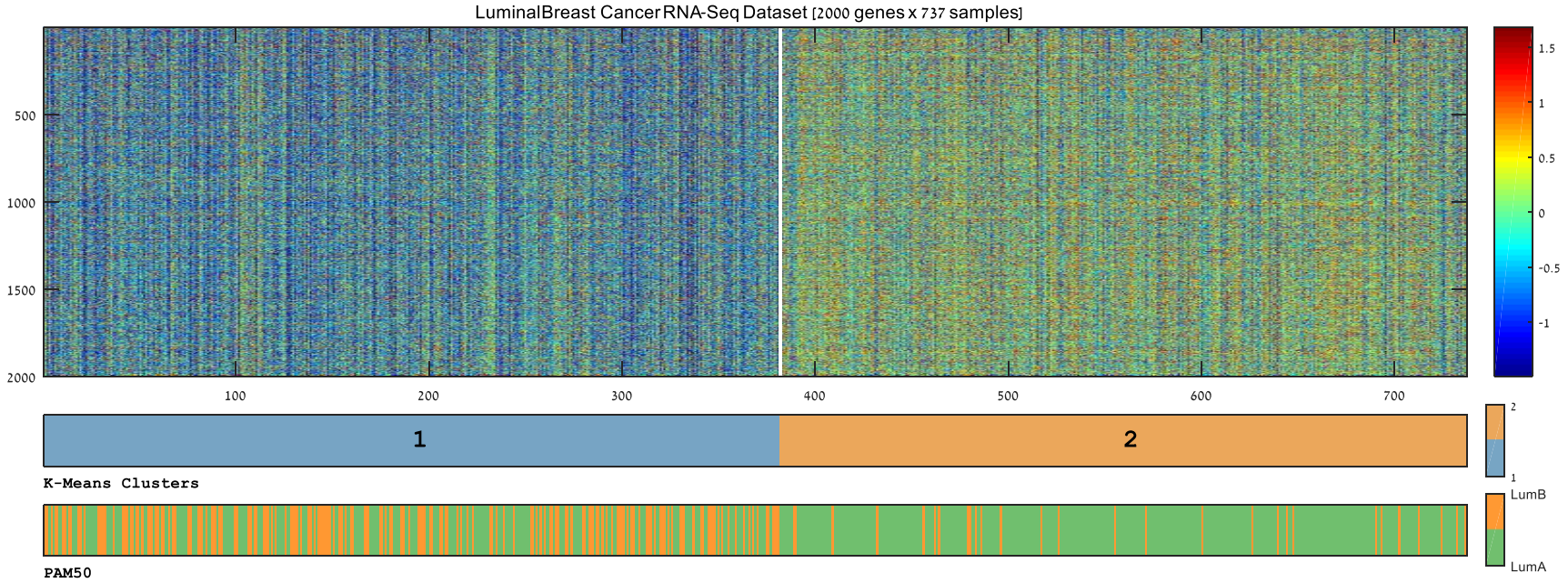
1148 Breast samples (Normal + Tumor)

Unsupervised analysis on RNA-Seq data

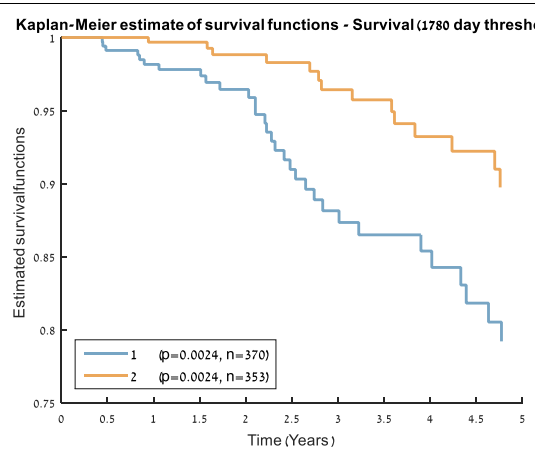
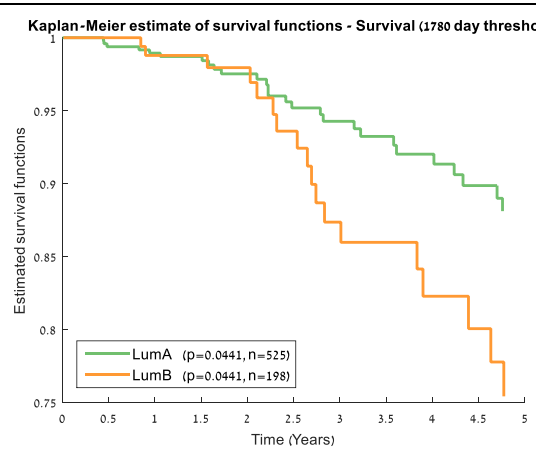
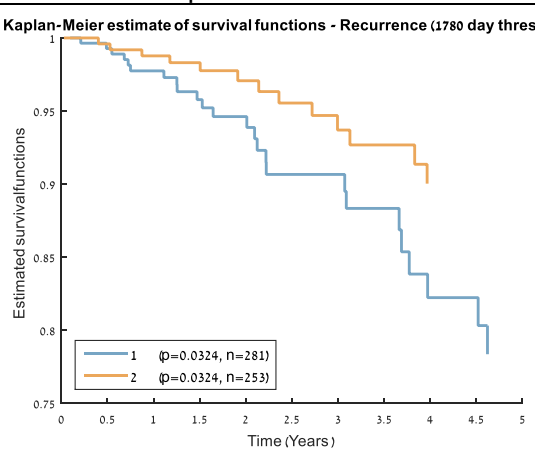
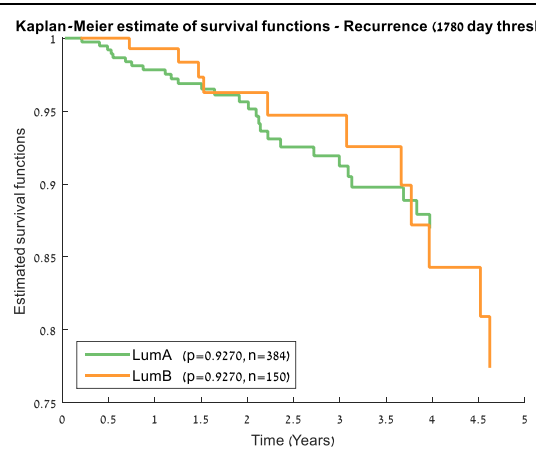


Luminal breast samples

Unsupervised analysis on RNA-Seq data

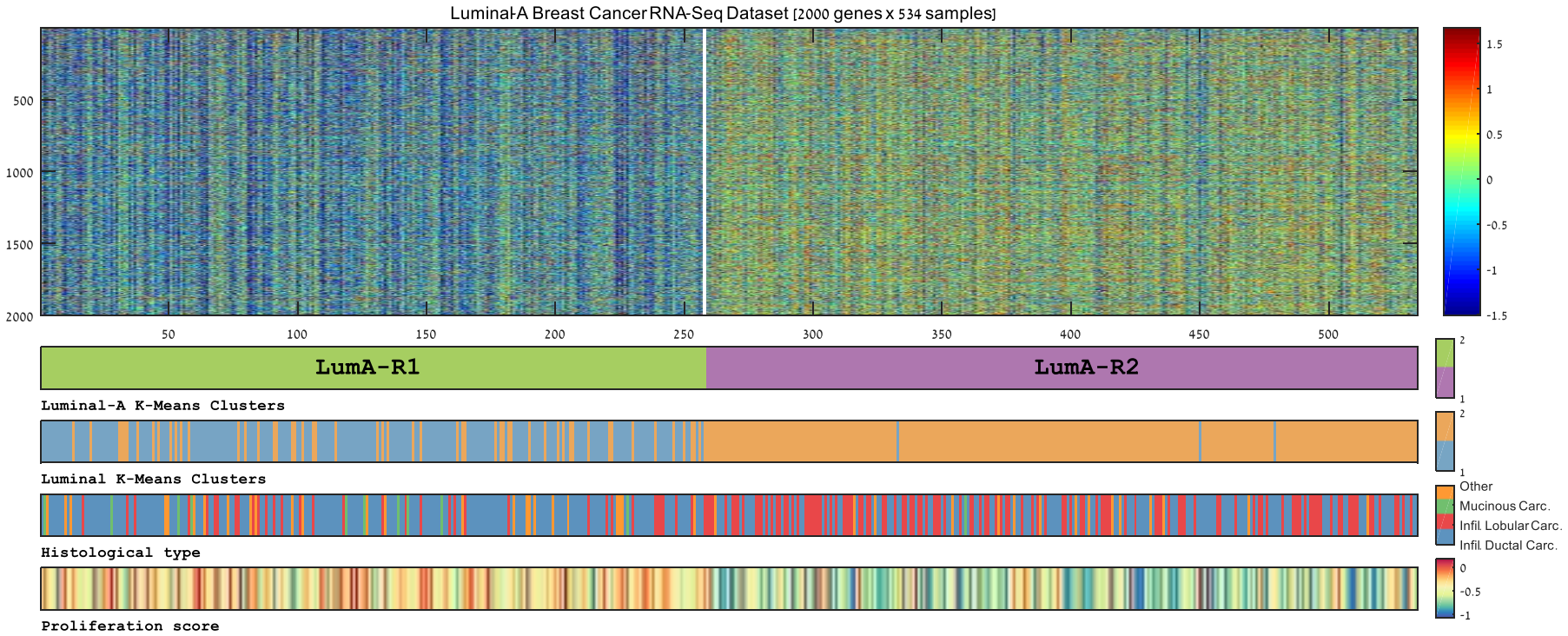


Five-year Kaplan-Meier plots for the two Luminal breast cancer partitions

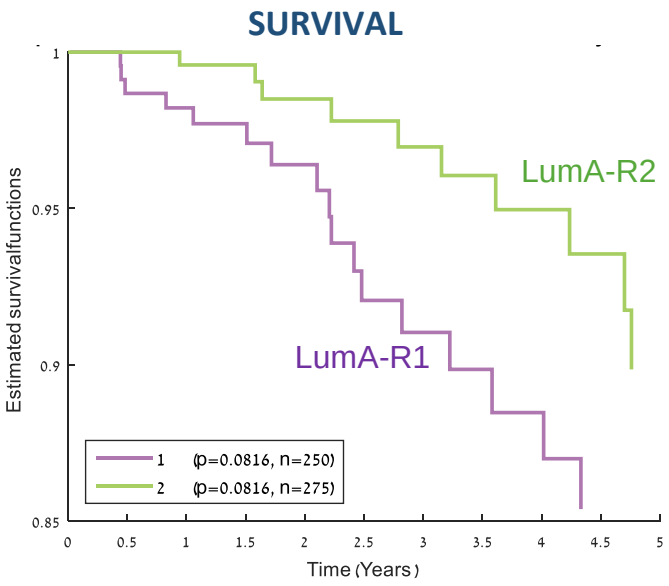
	RNA-SEQ clusters	PAM50
ear RVIVAL	 pValue=0.002	 pValue=0.044
ear CURRENCE	 pValue = 0.032	 pValue = 0.927

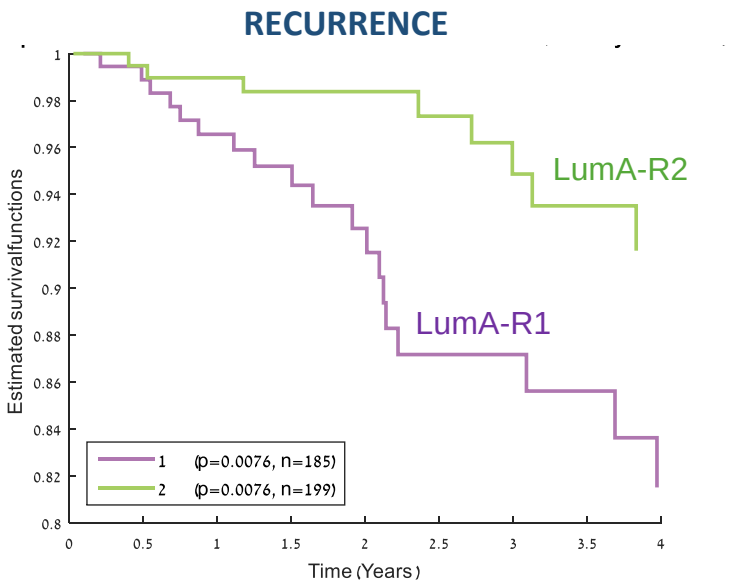
Luminal-A breast samples

Unsupervised analysis on RNA-Seq data



Five-year Kaplan-Meier plots for the two Luminal-A subgroups





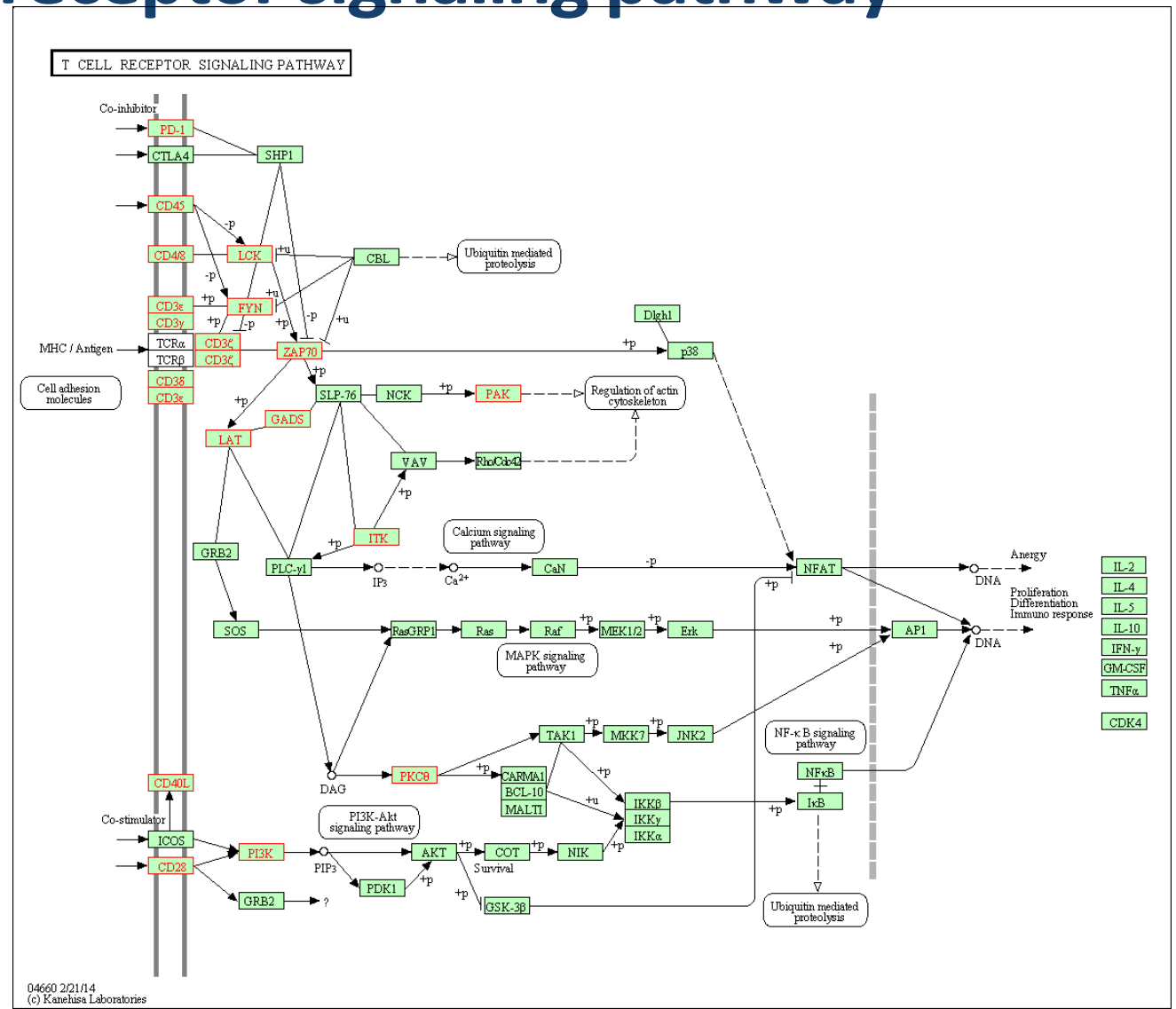
Main distinguishing characteristics between the Luminal-A subgroups

Group Characteristic	LumA-R1	LumA-R2	p-value
Recurrence free survival	Increased recurrence	Reduced recurrence	7.6e-3
Histology Enrichment p-values for each group	Ductal (p=2.1e-05)	Lobular (p=9.7e-12)	
Age average	61.5	57.4	2.6e-05
Proliferation score	-0.4	-0.6	8.9e-25
Tumor nuclei percent	80%	73%	2.6e-12
Normal cell percent	2.9%	6.1%	2.8e-08
Gene over-expression Out of 2000 genes used for clustering	194	1068	

The most enriched functional categories among the 1000 genes most differentially expressed between LumA-R1 and LumA-R2 samples

ENRICHMENT TYPE	TERM	#GENES	P-VALUE
GENE ONTOLOGY	regulation of immune system process	152	3.74E-50
	immune system process	201	3.65E-47
	regulation of leukocyte activation	71	2.37E-28
	regulation of multicellular organismal process	183	2.89E-28
	cell activation	91	4.59E-28
	regulation of response to external	73	8.18E-27
	regulation of biological quality	218	1.82E-26
	leukocyte activation	67	1.95E-26
	positive regulation of cell activation	56	5.13E-24
	T cell activation	45	4.93E-22
	regulation of cell proliferation	128	1.83E-21
KEGG PATHWAYS	Cytokine-cytokine receptor interaction	56	4.76E-22
	Hematopoietic cell lineage	29	1.50E-17
	Cell adhesion molecules (CAMs)	30	4.08E-13
	Primary immunodeficiency	16	8.70E-13
	Chemokine signaling pathway	31	1.14E-09
	Complement and coagulation cascades	17	1.36E-08
	T cell receptor signaling pathway	20	1.30E-07
	Allograft rejection	11	6.44E-07
	Natural killer cell mediated cytotoxicity	20	5.66E-06
	Pathways in cancer	34	1.49E-05
WIKI-PATHWAYS	TCR Signaling Pathway	10	1.55E-09
	B Cell Receptor Signaling Pathway	10	1.72E-06
	Focal Adhesion	11	5.88E-05
	Complement Activation, Classical Pathway	6	8.38E-05
CHROMOSOMAL LOCATION	11q23	18	1.84E-05
	Xq23	8	4.99E-05

LumA-R2 samples over-express genes in the T cell receptor signaling pathway

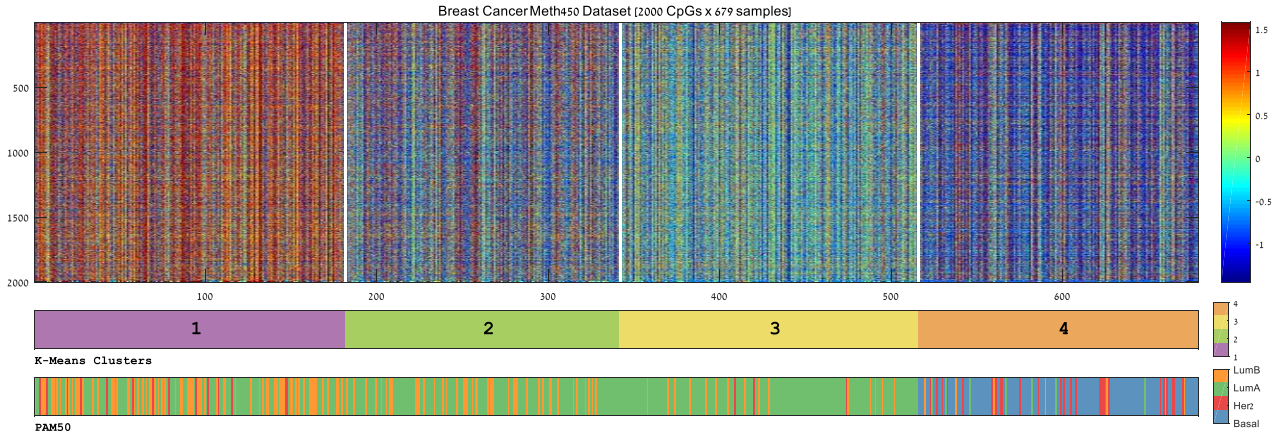


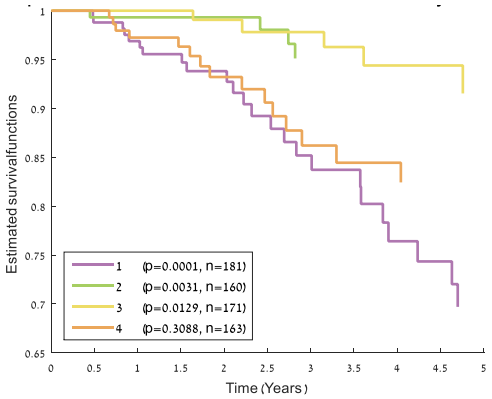
Intermediate summary:

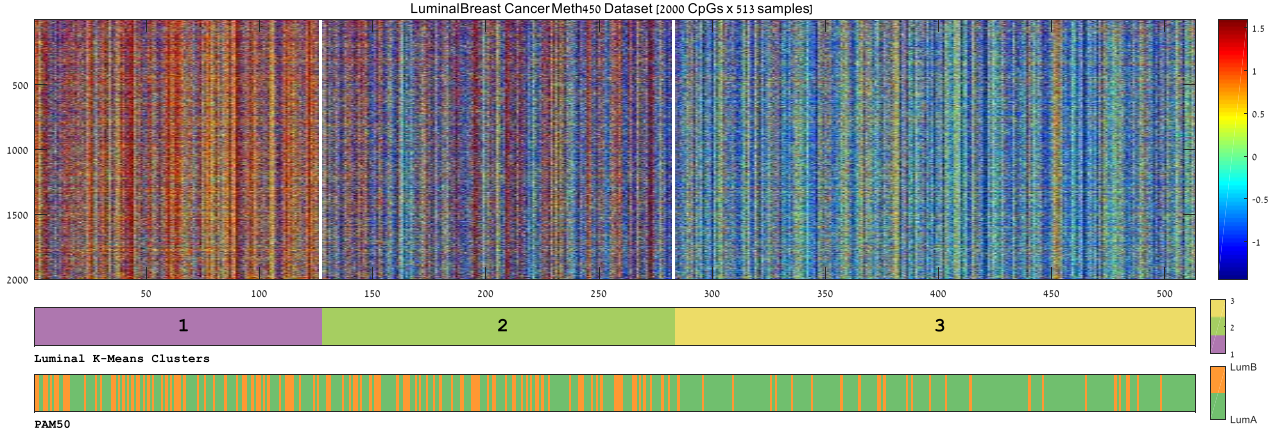
Unsupervised analysis of RNA-Seq expression data identifies a Luminal-A subgroup associated with reduced recurrence

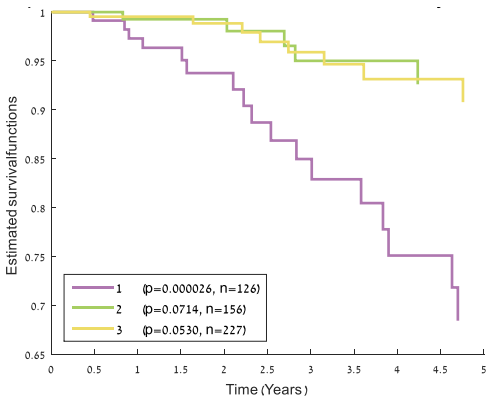
Next:

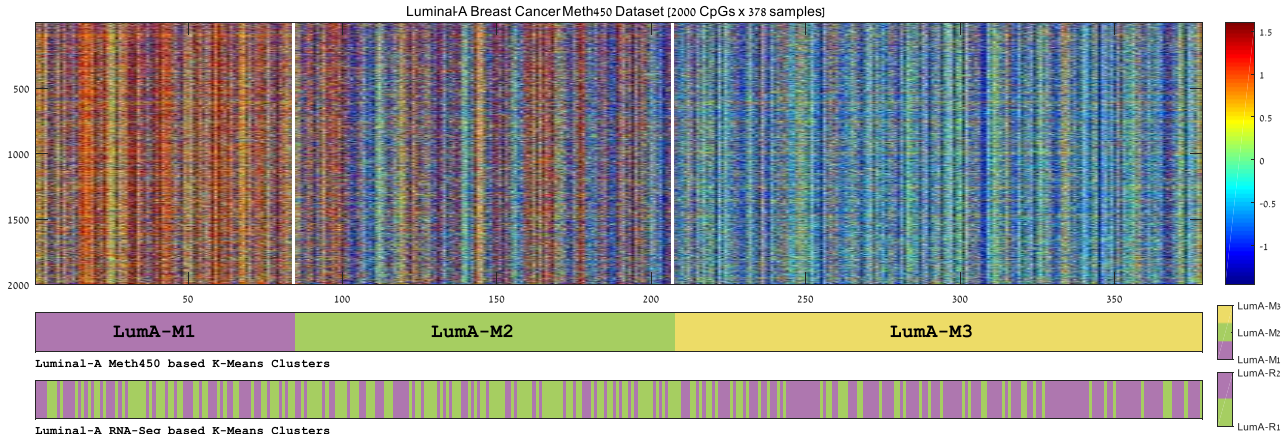
Unsupervised analysis of DNA-methylation data identifies a Luminal-A subgroup associated with bad survival

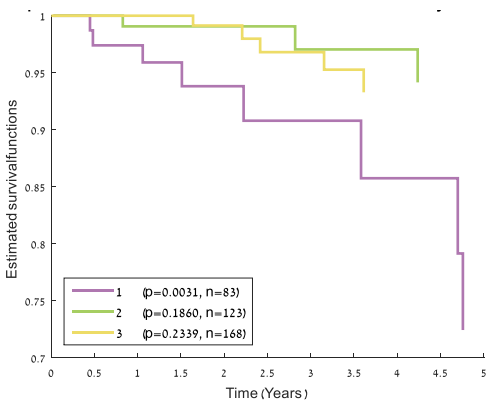

Breast Cancer Meth450 Dataset [2000 CpGs x 679 samples]
1
2
3
4
K-Means Clusters
LumB
LumA
Her2
Basal
PAM50


Estimated survivalfunctions
Time (Years)
1 (p=0.0001, n=181)
2 (p=0.0031, n=160)
3 (p=0.0129, n=171)
4 (p=0.3088, n=163)

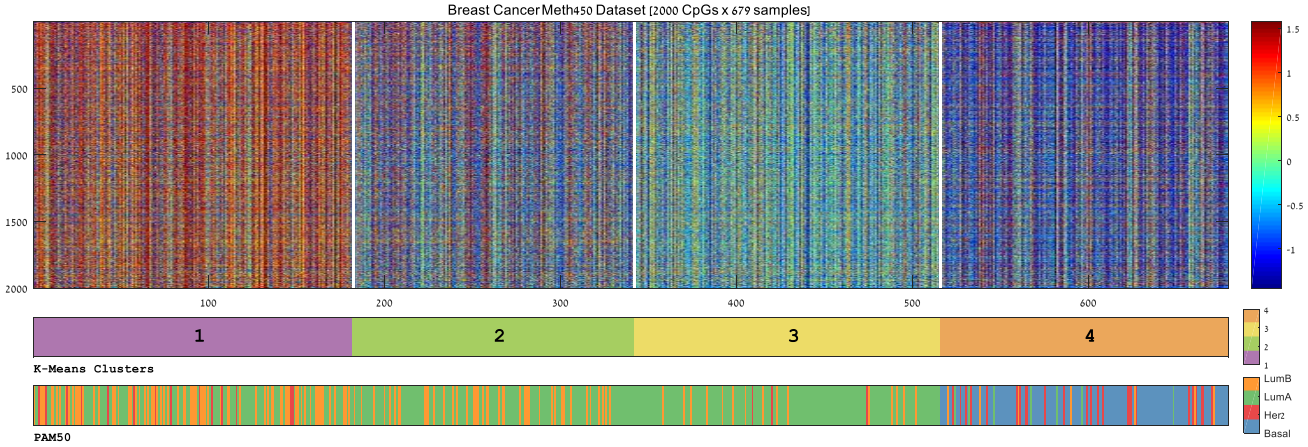

LuminalBreast Cancer Meth450 Dataset [2000 CpGs x 513 samples]
1
2
3
Luminal K-Means Clusters
LumB
LumA
PAM50

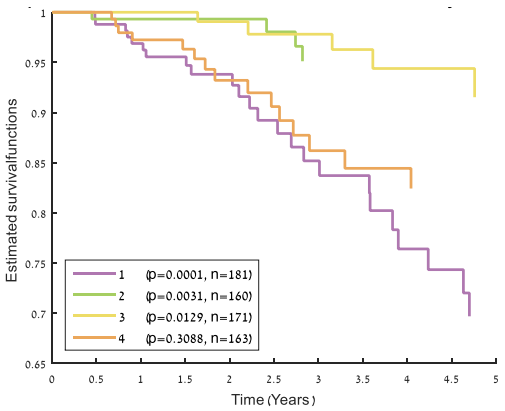

Estimated survivalfunctions
Time (Years)
1 (p=0.000026, n=126)
2 (p=0.0714, n=156)
3 (p=0.0530, n=227)


LuminalA Breast Cancer Meth450 Dataset [2000 CpGs x 378 samples]
LumA-M1
LumA-M2
LumA-M3
Luminal-A Meth450 based K-Means Clusters
LumA-M3
LumA-M2
LumA-M1
Luminal-A RNA-Seq based K-Means Clusters
LumA-R2
LumA-R1

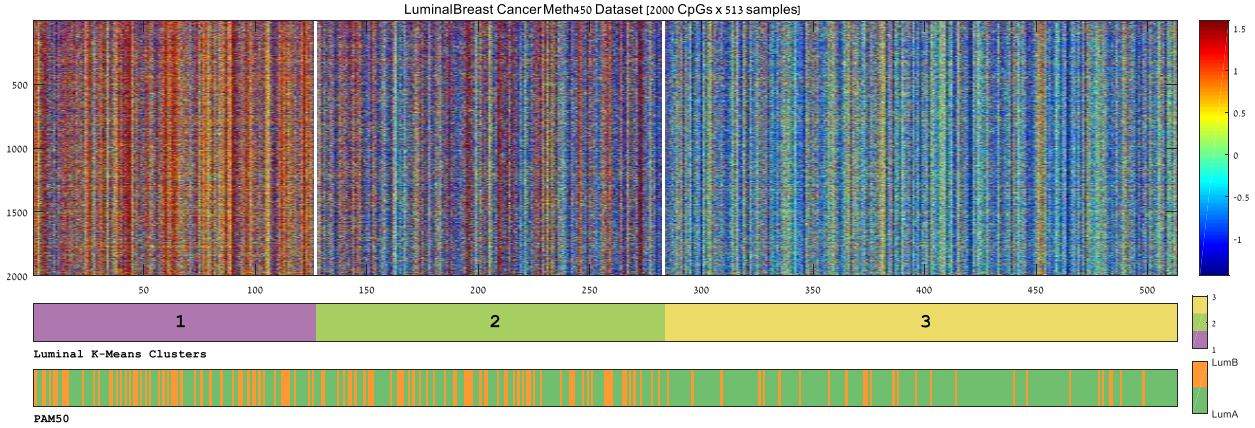

Estimated survivalfunctions
Time (Years)
1 (p=0.0031, n=83)
2 (p=0.1860, n=123)
3 (p=0.2339, n=168)

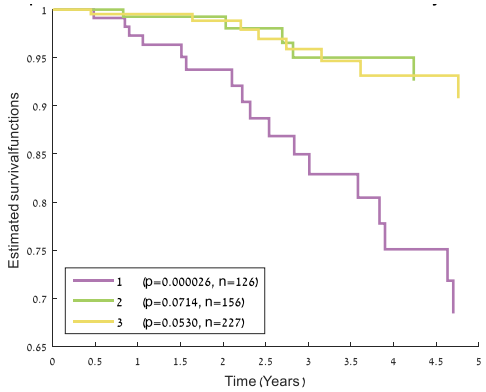
Basal, Her2, Luminal-A, Luminal-B



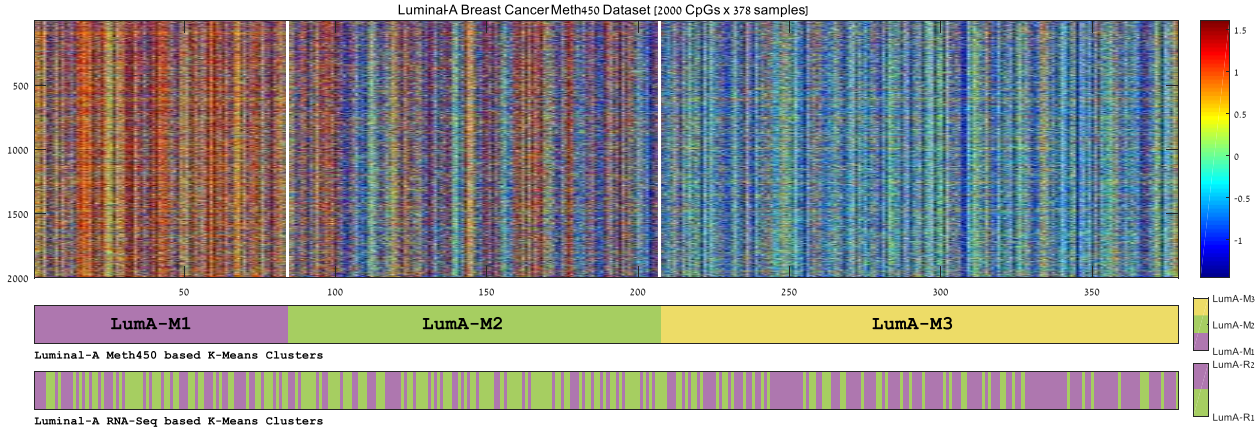


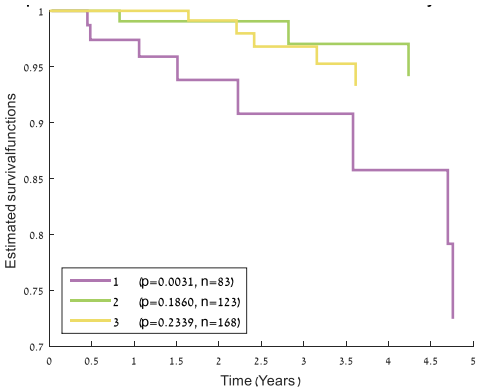
Luminal-A, Luminal-B





Luminal-A





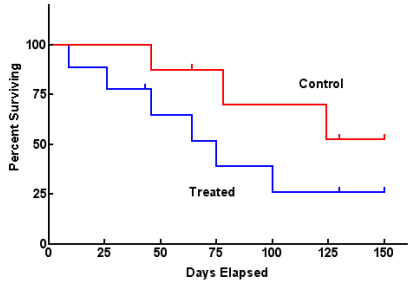
	(1) Hyper Meth. CpGs	
Gene ontology	anatomical structure development	6.1E-28
	developmental process	2.0E-25
	multicellular organismal process	9.6E-24
	single-multicellular organism process	1.6E-22
	single organism signaling	1.7E-21
	signaling	1.9E-21
	cell-cell signaling	1.7E-21
	neuron differentiation	1.2E-20
	single-organism developmental process	1.4E-19
	regulation of transcription from RNA polymerase II promoter	1.2E-16

		Hyper Meth. CpGs		Neg: R < -0.2		Pos: R > 0.2	
Label	**Term**	**Over-representation FDR corrected pValue**	**Under-representation FDR corrected pValue**	**Over-representation FDR corrected pValue**	**Under-representation FDR corrected pValue**	**Over-representation FDR corrected pValue**	**Under-representation FDR corrected pValue**
UCSC RefGene Group	1stExon	**1.E-04**	1.E+00	**1.E-07**	1.E+00	1.E+00	3.E-02
	3'UTR	1.E+00	2.E-03	1.E+00	**6.E-04**	2.E-02	1.E+00
	5'UTR	1.E+00	8.E-01	3.E-01	1.E+00	1.E+00	2.E-02
	Body	1.E+00	**7.E-05**	1.E+00	**1.E-16**	**9.E-20**	1.E+00
	TSS	2.E-02	1.E+00	**4.E-05**	1.E+00	1.E+00	**7.E-14**
Regulatory Feature Group	Gene Associated	1.E+00	2.E-01	1.E+00	5.E-01	1.E+00	1.E+00
	Gene Associated Cell type specific	1.E+00	5.E-02	1.E+00	2.E-01	2.E-01	1.E+00
	NonGene Associated	1.E+00	3.E-01	1.E+00	1.E-01	1.E+00	8.E-01
	NonGene Associated Cell type specific	3.E-03	1.E+00	5.E-01	1.E+00	2.E-01	1.E+00
	Promoter Associated	1.E+00	**2.E-146**	1.E+00	**3.E-31**	1.E+00	**4.E-34**
	Promoter Associated Cell type specific	1.E+00	5.E-02	**1.E-04**	1.E+00	1.E+00	7.E-02
	Unclassified	1.E+00	4.E-01	6.E-04	1.E+00	1.E+00	1.E+00
	Unclassified Cell type specific	9.E-35	1.E+00	4.E-06	1.E+00	1.E-10	1.E+00
	Unassigned	7.E-52	1.E+00	5.E-06	1.E+00	2.E-09	1.E+00
DMR Differentially Methylated Region	CDMR	**2.E-16**	1.E+00	4.E-03	1.E+00	1.E-13	1.E+00
	DMR	**9.E-183**	1.E+00	2.E-75	1.E+00	1.E-15	1.E+00
	RDMR	**2.E-04**	1.E+00	2.E-01	1.E+00	2.E-11	1.E+00
	Unassigned	1.E+00	**2.E-205**	1.E+00	2.E-75	1.E+00	5.E-40
Enhancer		**1.E-09**	1.E+00	8.E-06	1.E+00	2.E-04	1.E+00
DHS (DNAse hypersensitive site)		**1.E-07**	1.E+00	2.E-03	1.E+00	2.E-05	1.E+00

		Hyper Meth. CpGs		Neg: R < -0.2		Pos: R > 0.2	
el	**Term**	**Over-representation FDR corrected pValue**	**Under-representation FDR corrected pValue**	**Over-representation FDR corrected pValue**	**Under-representation FDR corrected pValue**	**Over-representation FDR corrected pValue**	**Under-representation FDR corrected pValue**
SC RefGene Group	1stExon	**1.E-04**	1.E+00	**1.E-07**	1.E+00	1.E+00	3.E-02
	3'UTR	1.E+00	2.E-03	1.E+00	**6.E-04**	2.E-02	1.E+00
	5'UTR	1.E+00	8.E-01	3.E-01	1.E+00	1.E+00	2.E-02
	Body	1.E+00	**7.E-05**	1.E+00	**1.E-16**	**9.E-20**	1.E+00
	TSS	2.E-02	1.E+00	**4.E-05**	1.E+00	1.E+00	**7.E-14**
Regulatory ature Group	Gene Associated	1.E+00	2.E-01	1.E+00	5.E-01	1.E+00	1.E+00
	Gene Associated Cell type specific	1.E+00	5.E-02	1.E+00	2.E-01	2.E-01	1.E+00
	NonGene Associated	1.E+00	3.E-01	1.E+00	1.E-01	1.E+00	8.E-01
	NonGene Associated Cell type specific	3.E-03	1.E+00	5.E-01	1.E+00	2.E-01	1.E+00
	Promoter Associated	1.E+00	**2.E-146**	1.E+00	**3.E-31**	1.E+00	**4.E-34**
	Promoter Associated Cell type specific	1.E+00	5.E-02	**1.E-04**	1.E+00	1.E+00	7.E-02
	Unclassified	1.E+00	4.E-01	6.E-04	1.E+00	1.E+00	1.E+00
	Unclassified Cell type specific	9.E-35	1.E+00	4.E-06	1.E+00	1.E-10	1.E+00
	Unassigned	7.E-52	1.E+00	5.E-06	1.E+00	2.E-09	1.E+00
R (Differentially hylated Region	CDMR	**2.E-16**	1.E+00	4.E-03	1.E+00	1.E-13	1.E+00
	DMR	**9.E-183**	1.E+00	2.E-75	1.E+00	1.E-15	1.E+00
	RDMR	**2.E-04**	1.E+00	2.E-01	1.E+00	2.E-11	1.E+00
	Unassigned	1.E+00	**2.E-205**	1.E+00	2.E-75	1.E+00	5.E-40
Enhancer		**1.E-09**	1.E+00	8.E-06	1.E+00	2.E-04	1.E+00
HS (DNAse ersensitive site)		**1.E-07**	1.E+00	2.E-03	1.E+00	2.E-05	1.E+00

Methods for survival analysis

- ✓Kaplan-Meier estimator
  - ✓ a non-parametric statistic used to estimate the survival function from lifetime data



- ✓Log rank test
  - ✓An hypothesis **test** to compare the survival distributions of two samples. It is also nonparametric and appropriate to use when the data are right skewed and censored (technically, the censoring must be non-informative).

- ✓COX univariate/multivariate regression model
  - ✓Cox regression (or proportional hazards regression) is method for investigating the effect of several variables upon the time a specified event takes to happen

Cox proportional hazards regression analysis

	Survival				Recurrence			
	Univariate		Multivariate		Univariate		Multivariate	
Variable	**HR**	**pValue**	**HR**	**pValue**	**HR**	**pValue**	**HR**	**pValue**
LumA-R (1 vs 2)	0.44	0.10939	0.62	0.43821	**0.20**	**0.00421**	**0.06**	**0.00735**
LumA-M (2,3 vs 1)	**4.53**	**0.00258**	**6.67**	**0.00494**	1.64	0.34338	3.07	0.07164
Age (<60 vs.>=60 years)	**5.79**	**0.00624**	**12.93**	**0.00296**	2.18	0.10301	1.02	0.97870
Pathologic stage (I,II vs. III,IV)	1.30	0.62799	4.02	0.05463	2.09	0.11941	1.85	0.38642
Pathologic T (I,II vs III,IV)	0.27	0.20444	0.20	0.16871	1.38	0.53411	1.11	0.91331
ER Status	1.72	0.60363	7.86	0.16239	0.00	0.99217	0.00	0.99573
PR Status	1.03	0.96671	0.46	0.48454	0.37	0.33789	0.29	0.28818
Her2 Status	0.79	0.82080	1.11	0.92180	0.99	0.98916	0.63	0.68516

How can we combine the two signatures?

- COX model showing that assignment to both groups independently contributes to risk prediction (DONE)
- Joint clustering (BETA)
- Signature projection (Experimental)

Summary

- Current classification of breast tumors can be improved utilizing large modern genomic databases.
- An expression based signature composed of immune system genes can partition the Luminal-A samples into groups showing different recurrence risks.
- A methylation based signatures composed of developmental genes can partition the Luminal-A samples into groups showing different survival risks.
- The availability of several different assay technologies per sample calls for the development of computational approaches that would partition sample groups and predict risk based on the integration of several assay types.