Cancer Genomics and Class Discovery

2/20/17 – Chuck Perou (Department of Genetics) – Introduction to Genomics and Big Data, and Cancer Subtype Class Discovery using gene expression data (with some "homework")

- Katie Hoadley (Department of Genetics) – Introduction to TCGA Data Portal

2/27/17 – Katie Hoadley (Department of Genetics) – Multi-platform Data Analysis and Across Technology Data Integration

3/6/17 – Steve Marron (Department of Statistics and Operations Research) – Methods for Addressing Data Heterogeneity and Integration

3/13/17 – No Class

3/20/17 – Andrew Nobel (Department of Statistics and Operations Research) – Exploratory Analysis of Genomic Data

3/27/17– Joel Parker (Department of Genetics) – Methods and Challenges in the Analysis of NextGen Sequence Data for DNAseq and RNAseq

4/3/17 – No Class

4/10/17 – In Class Student Presentations (70%) and 2-3 page Written Report (30%) covering a unique analysis performed on TCGA Cancer Genomics Data

Cancer Genomics and Class Discovery

2/20/17 – Chuck Perou (Department of Genetics) – Introduction to Genomics and Big Data, and Cancer Subtype Class Discovery using gene expression data

- Katie Hoadley (Department of Genetics) – Introduction to TCGA Data Portal

Reading list

Eisen et al., PNAS 1998 (PMID:9843981)

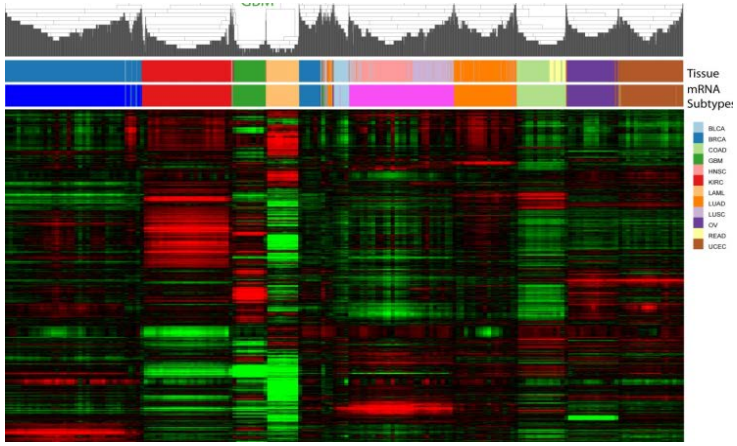
Perou et al., NATURE 2000 (PMID:10963602)

Parker et al., JCO 2009 (PMID:19204204)

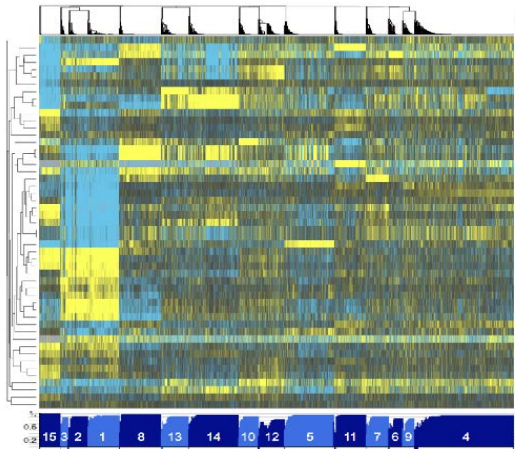
TCGA Breast Cancer Genomic Data Sites

https://tcga-data.nci.nih.gov/docs/publications/brca_2015/
https://lbg.unc.edu/~hoadley/BRCA.817.rsemg.uqnorm.counts.txt
https://lbg.unc.edu/~hoadley/BRCA.817.rsemg.uqnorm.counts.intrinsic.txt

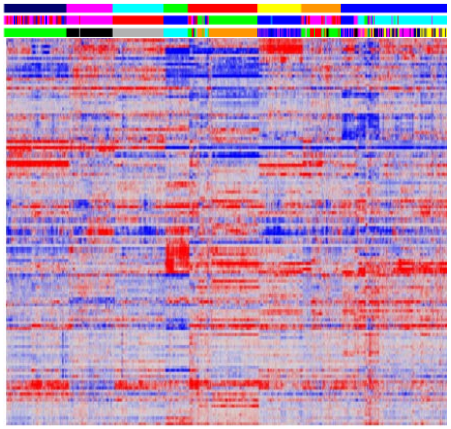
What is Big Data?



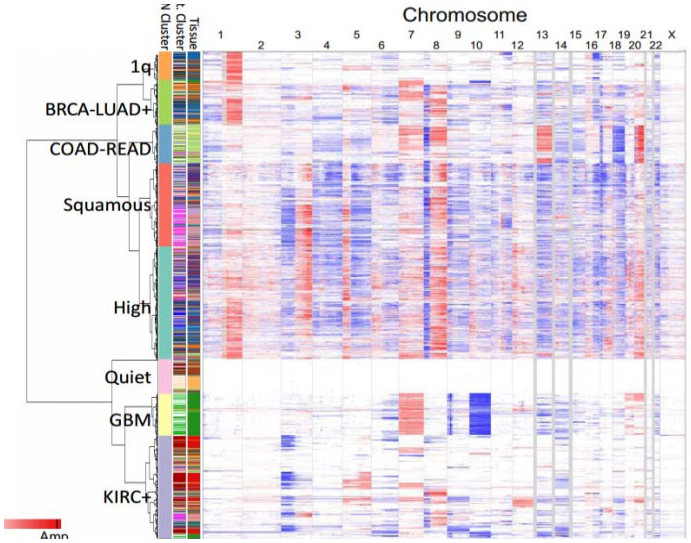
Gene Expression (mRNAs)
25,000-75,000/sample



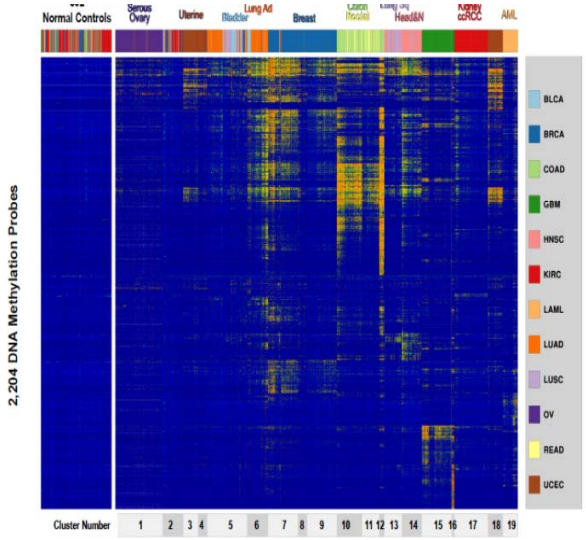
microRNAs
1500/sample



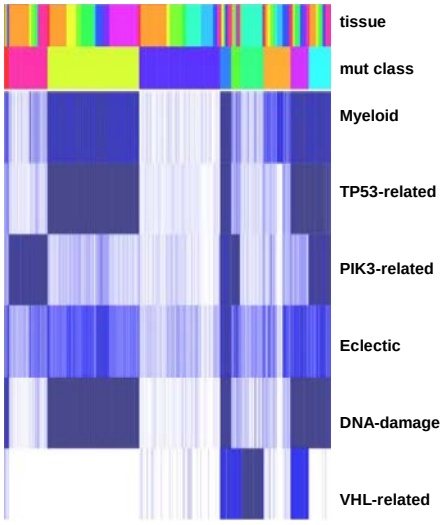
Proteins
200/sample



DNA Copy Number
1,000,000/sample

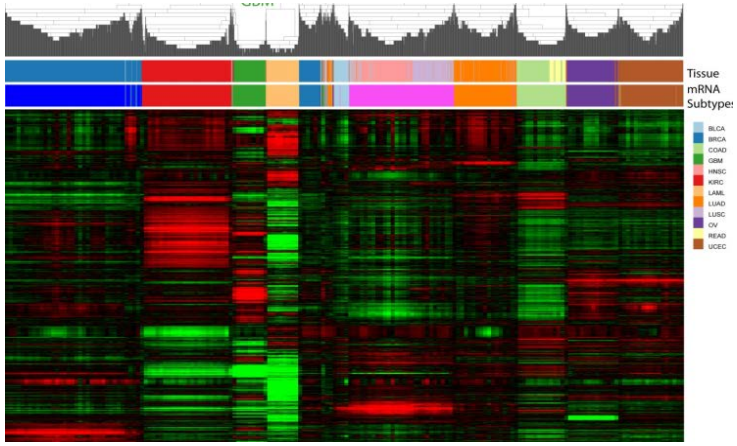


DNA Methylation
450,000/sample

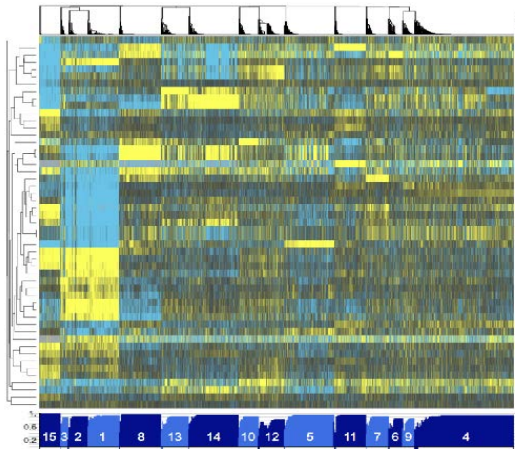


Somatic Mutations
25,000/sample

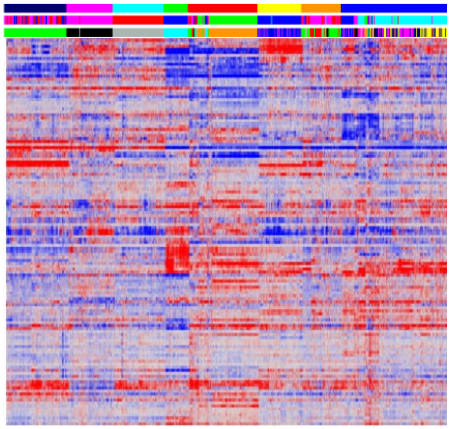
What is Big Data?



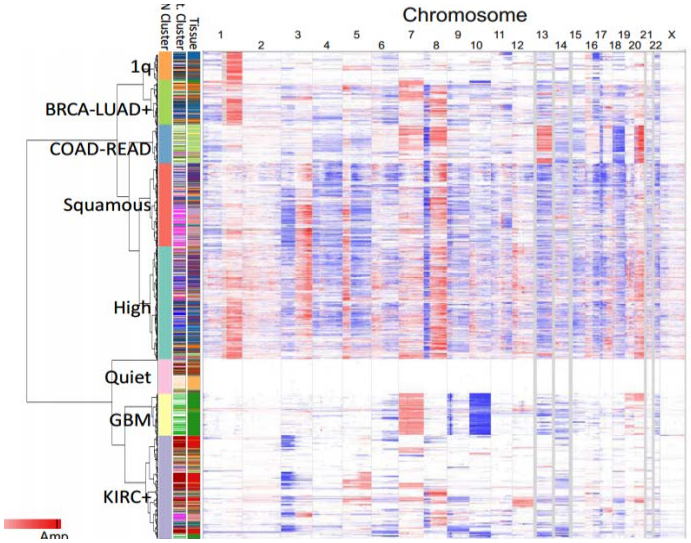
Gene Expression (mRNAs)



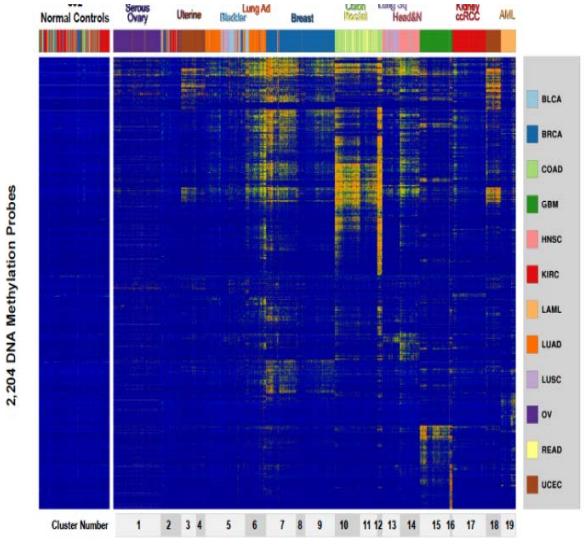
microRNAs



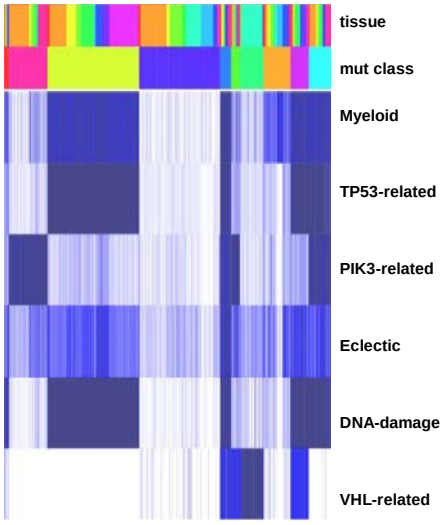
Proteins



DNA Copy Number



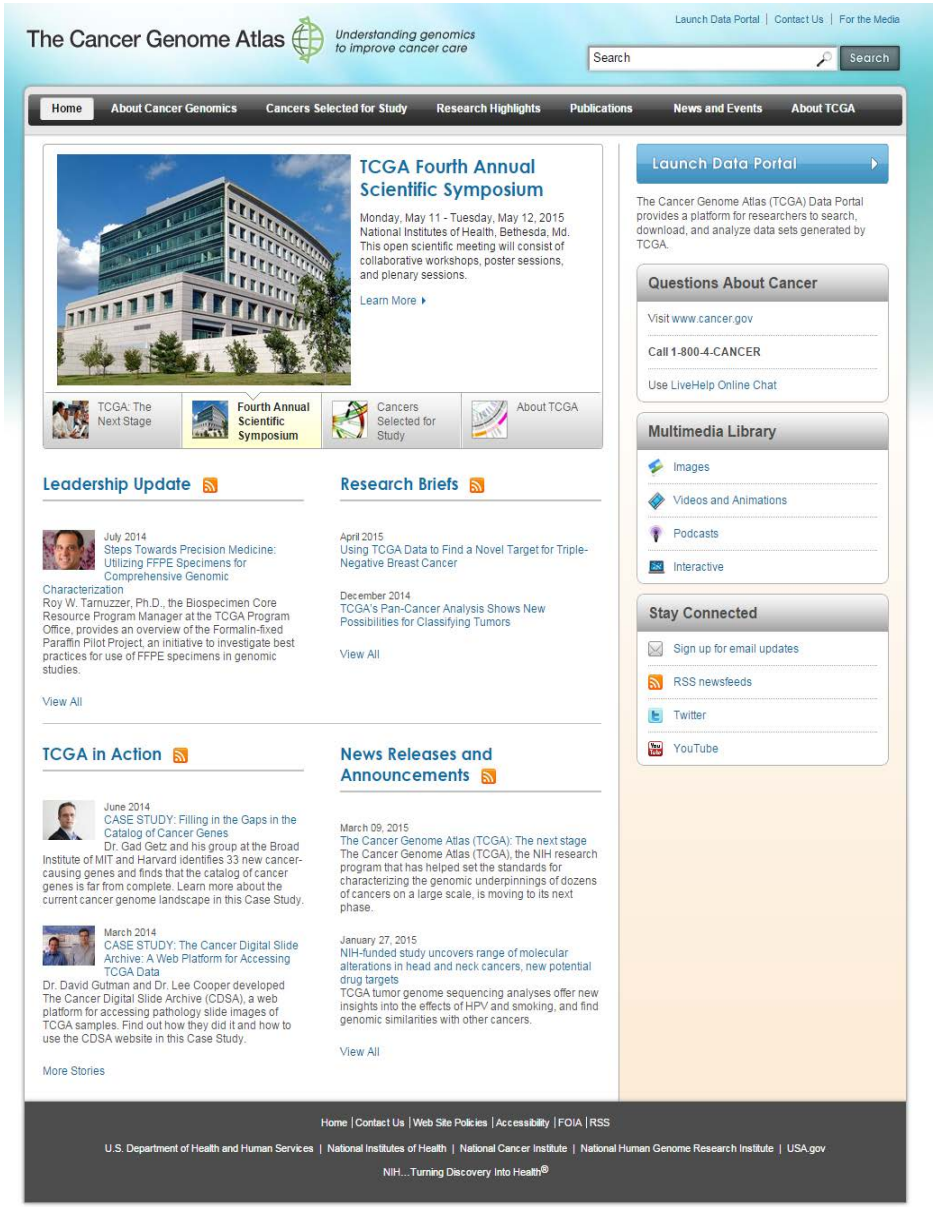
DNA Methylation



Somatic Mutations

~1.5 Million Data Points per sample X 10,000 samples = 15 Billion Data Points Total

http://cancergenome.nih.gov/



TCGA Data is

1. 10,000 individual tumors
2. 33 diverse tumor types
3. Clinical and Pathology data
4. Molecular assays performed:
    1. DNA exomes (mutations)
    2. mRNA-seq (gene expression)
    3. microRNA-seq (microRNAs)
    4. DNA methylation arrays
    5. AFFY SNP arrays (genotypes and DNA copy number)
    6. RPPA protein data on ~60%
    7. H&E images of each tumor

Rank	Cause of Death	Percent of Total Deaths
	All Causes	100
1	Heart Disease	23.5
2	Cancer	22.5
3	CLRD (Lung diseases)	5.7
4	Accidents	5
5	Cerebrovascular disease (stroke)	5
6	Alzheimer's disease	3.3
7	Diabetes	2.9
8	Influenza and pneumonia	2.2
9	Kidney disease	1.8
10	Suicide	1.6
	All other causes	26.4

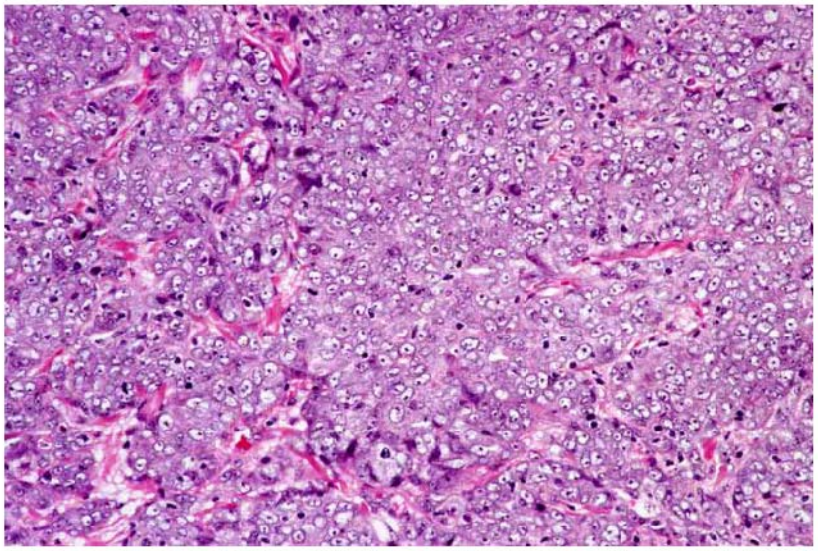
Estimated new cases*

Males			Females		
Prostate	238,590	28%	Breast	232,340	29%
Lung & bronchus	118,080	14%	Lung & bronchus	110,110	14%
Colorectum	73,680	9%	Colorectum	69,140	9%
Urinary bladder	54,610	6%	Uterine corpus	49,560	6%
Melanoma of the skin	45,060	5%	Thyroid	45,310	6%
Kidney & renal pelvis	40,430	5%	Non-Hodgkin lymphoma	32,140	4%
Non-Hodgkin lymphoma	37,600	4%	Melanoma of the skin	31,630	4%
Oral cavity & pharynx	29,620	3%	Kidney & renal pelvis	24,720	3%
Leukemia	27,880	3%	Pancreas	22,480	3%
Pancreas	22,740	3%	Ovary	22,240	3%
All Sites	**854,790**	**100%**	**All Sites**	**805,500**	**100%**

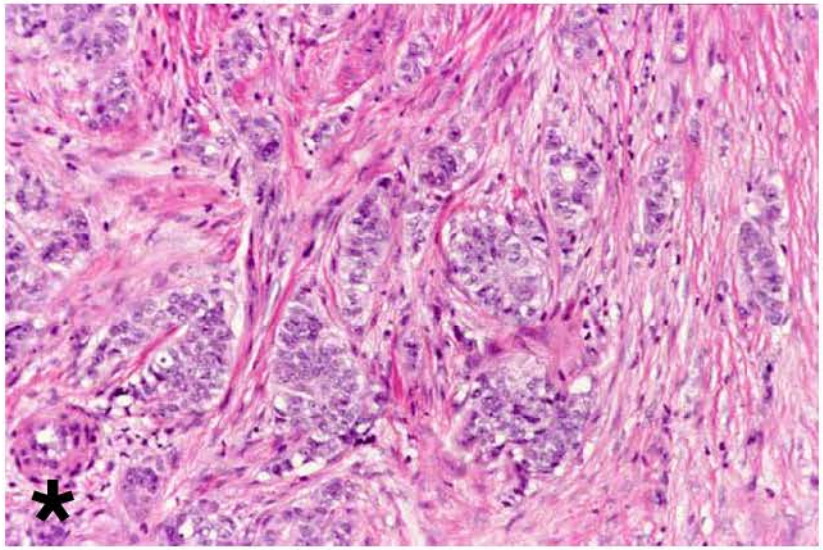
Estimated deaths

Males			Females		
Lung & bronchus	87,260	28%	Lung & bronchus	72,220	26%
Prostate	29,720	10%	Breast	39,620	14%
Colorectum	26,300	9%	Colorectum	24,530	9%
Pancreas	19,480	6%	Pancreas	18,980	7%
Liver & intrahepatic bile duct	14,890	5%	Ovary	14,030	5%
Leukemia	13,660	4%	Leukemia	10,060	4%
Esophagus	12,220	4%	Non-Hodgkin lymphoma	8,430	3%
Urinary bladder	10,820	4%	Uterine corpus	8,190	3%
Non-Hodgkin's lymphoma	10,590	3%	Liver & intrahepatic bile duct	6,780	2%
Kidney & renal pelvis	8,780	3%	Brain & other nervous system	6,150	2%
All Sites	**306,920**	**100%**	**All Sites**	**271,430**	**100%**

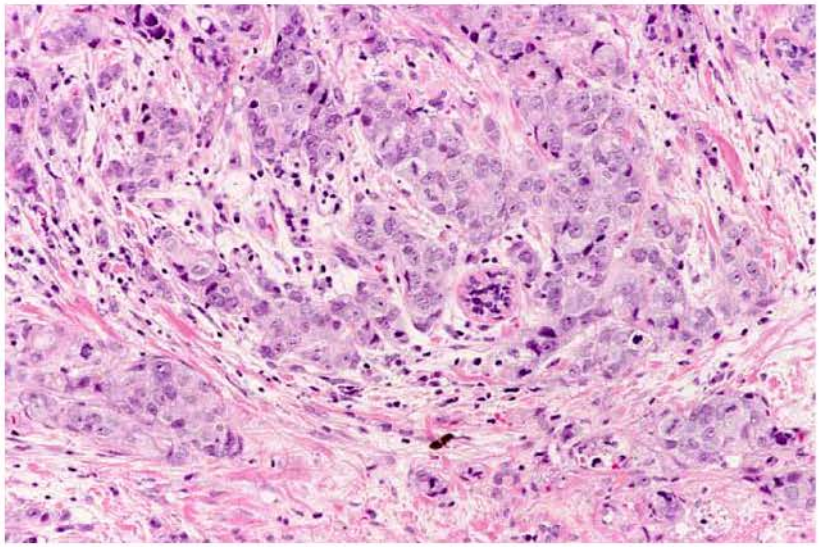
Breast Tumor, Grade III, Estrogen and Progesterone Receptor Positive, HER2-negative
T3 tumors (diameter >5cm), N1 (lymph node metastasis positive),
All received neoadjuvant chemotherapy and adjuvant tamoxifen



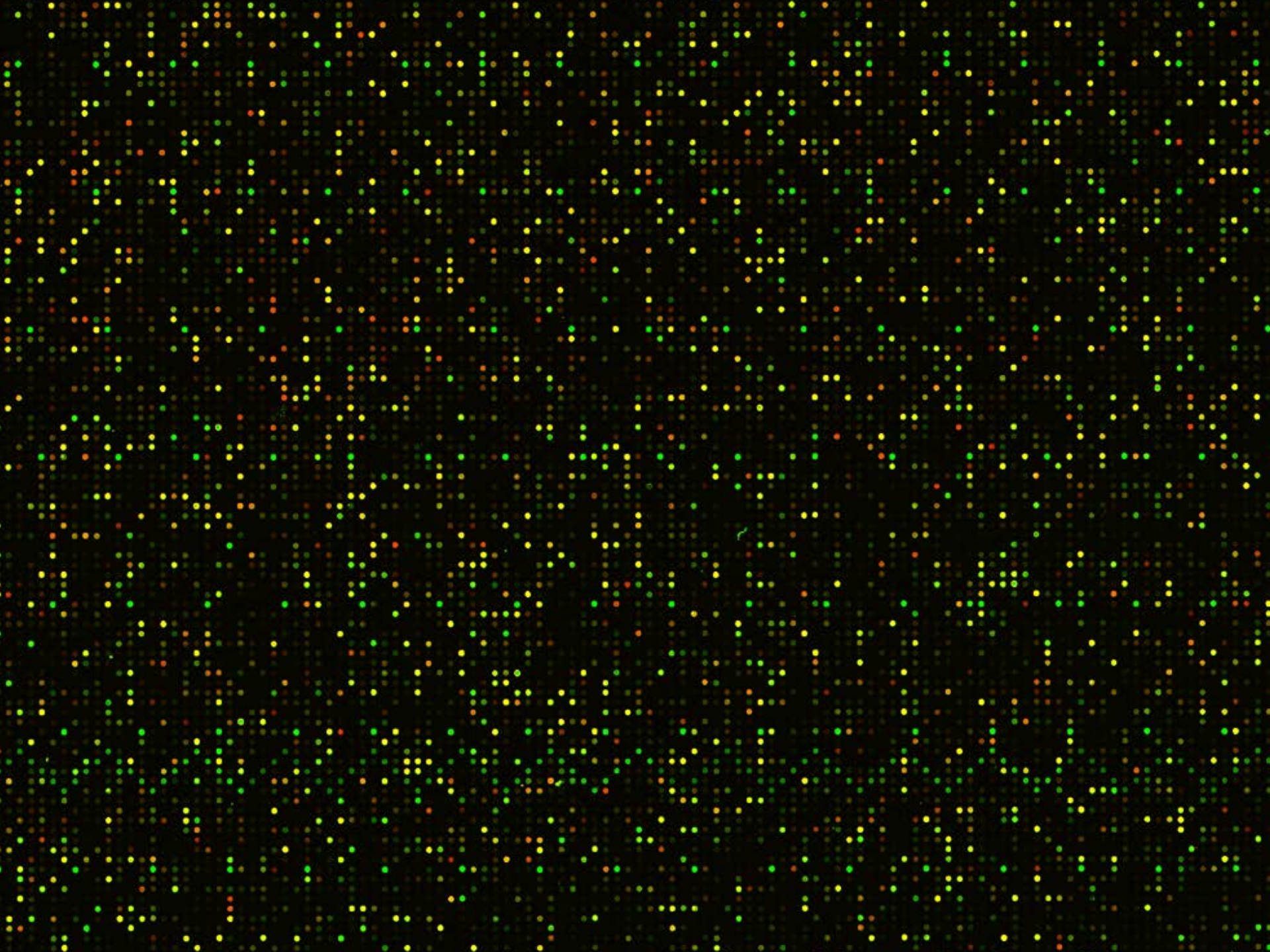
Died after 9 months

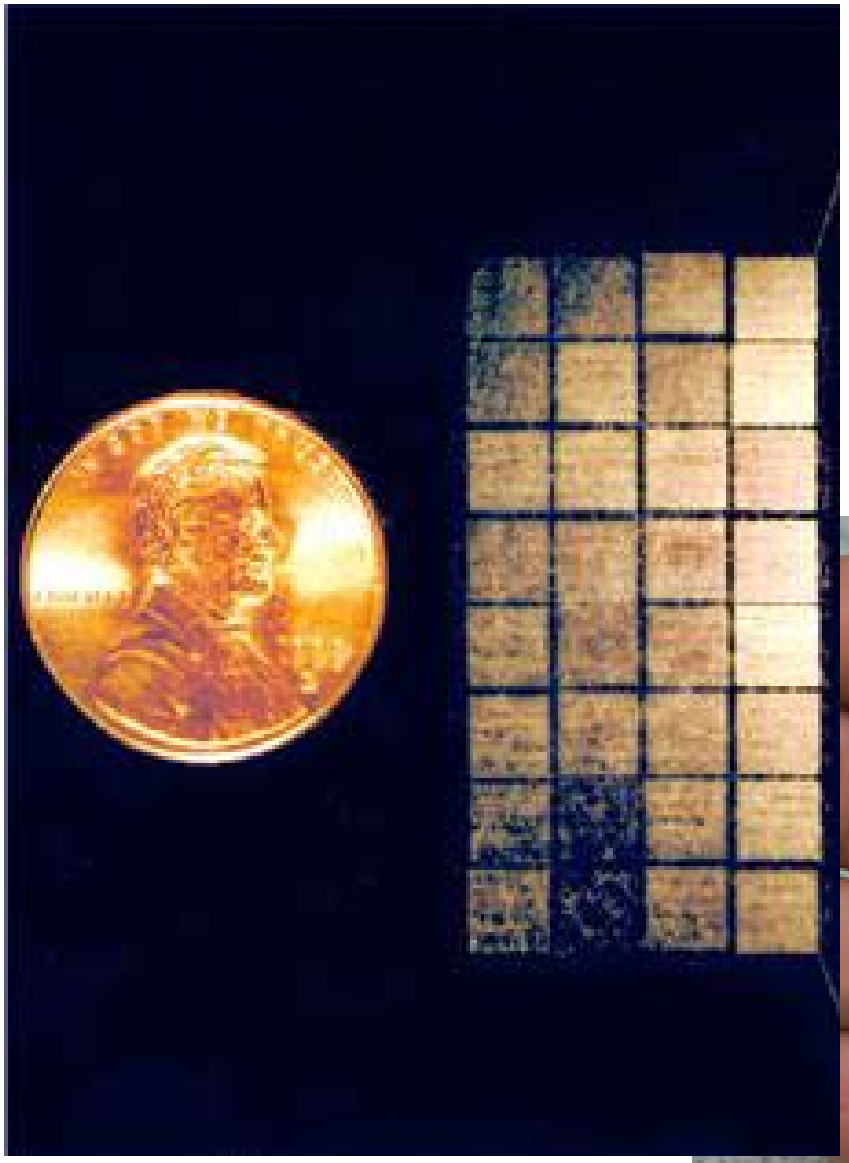


Died after 17 months



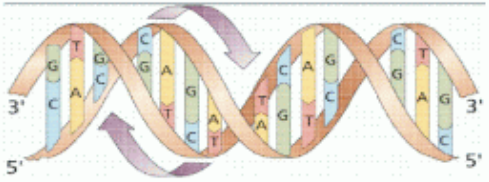
Alive for 96 months





AFFYMETRIX
GeneChip®
phylochip
P/N: 510706
Lot #: 4027020
Exp. Date: 04/29/07
For Research Use Only
UNITED STATES OF AMERICA
LIBERTY
QUARTER DOLLAR

Central Dogma: DNA -> RNA -> Protein



DNA

transcription

RNA

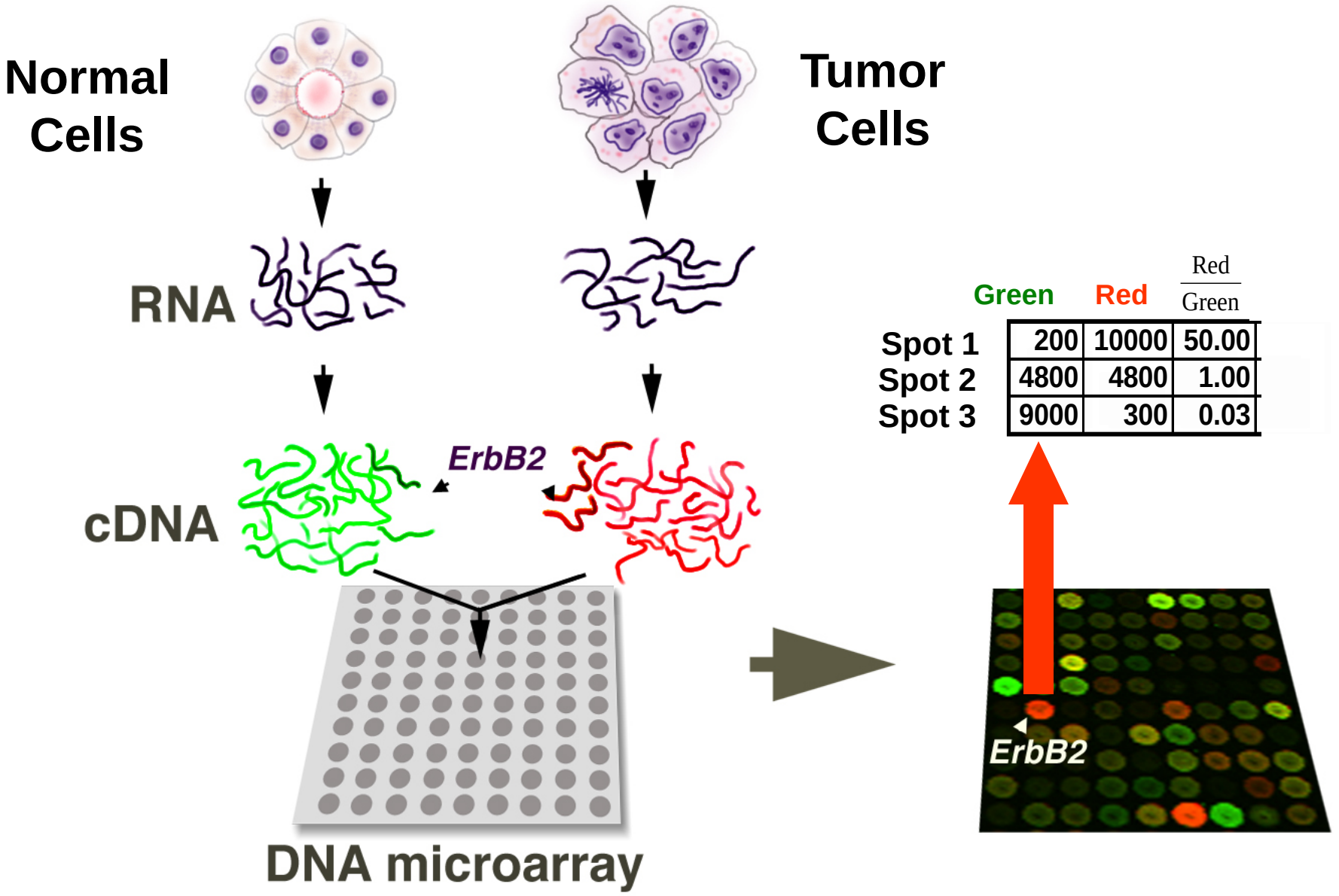
translation

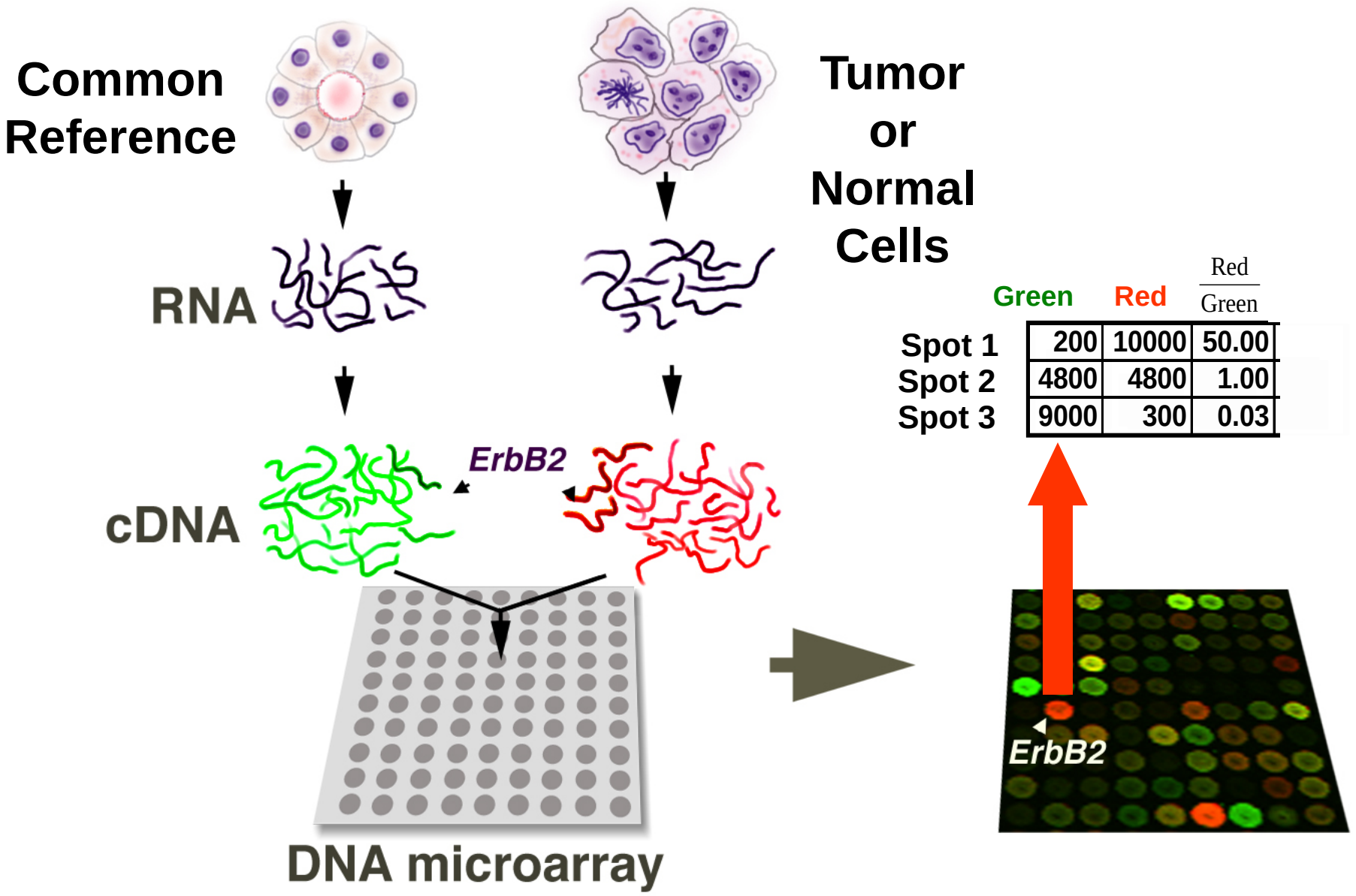
Protein

CCTGAGCCAACTATTGATGAA

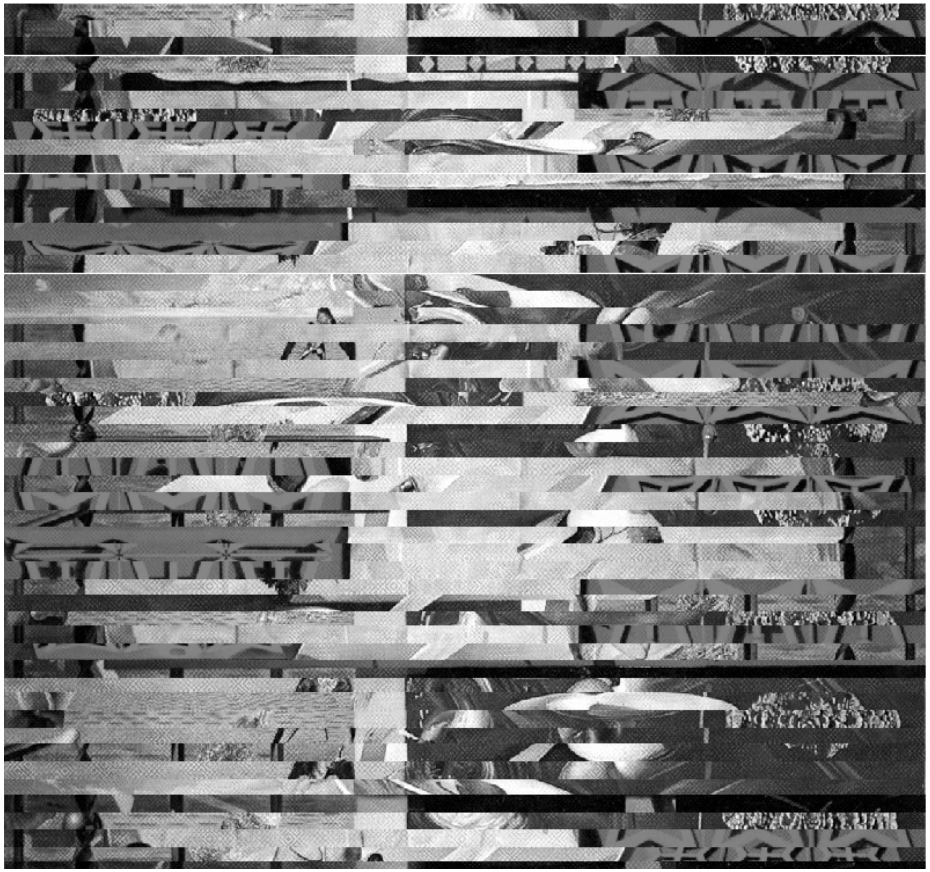
CCUGAGCCAACUAUUGAUGAA

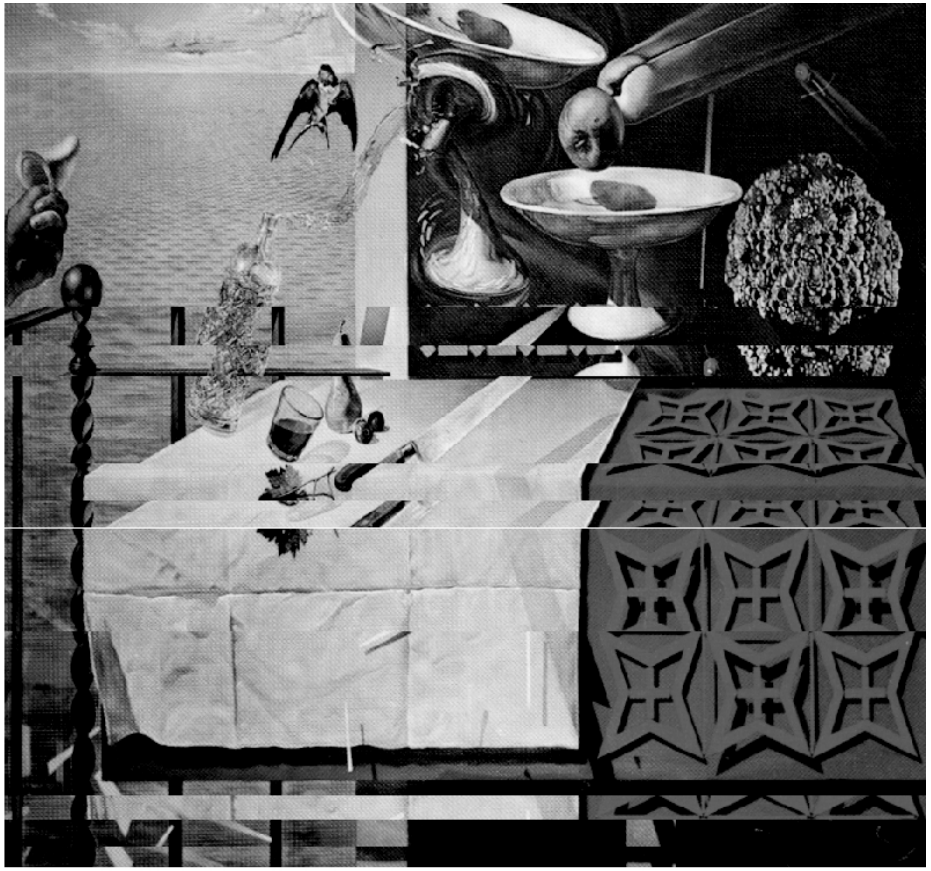
PEPTIDE

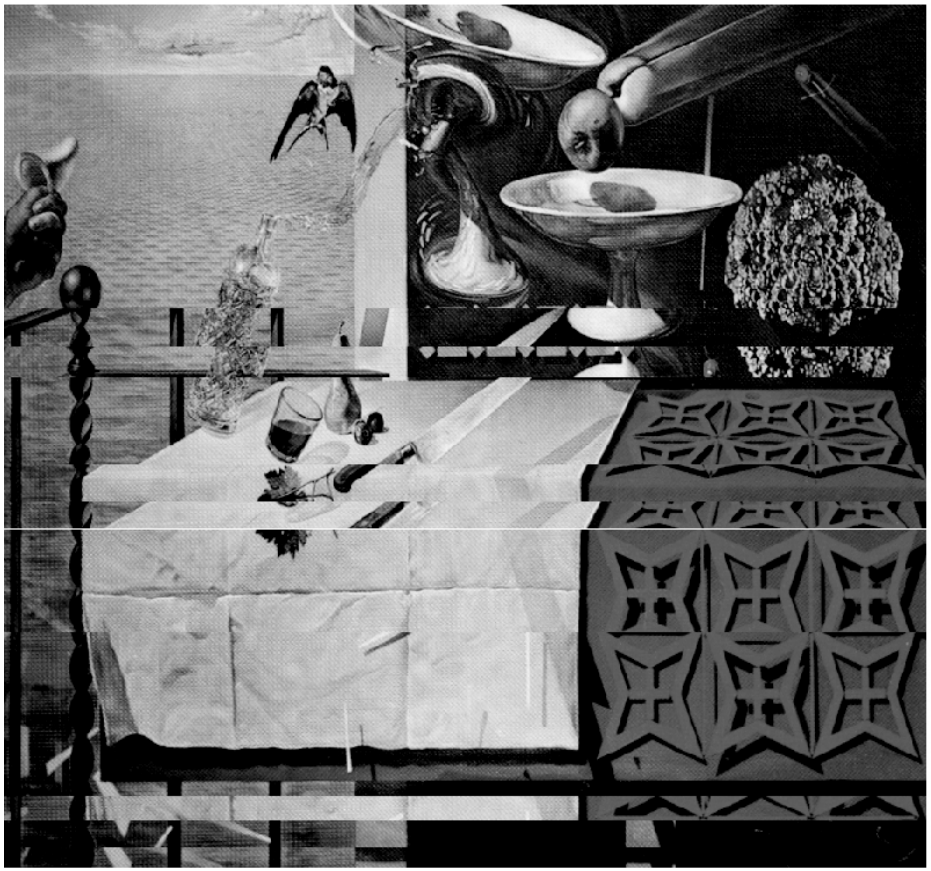

Normal Cells
Tumor Cells
RNA
cDNA
ErbB2
DNA microarray
Green
Red
Red / Green
Spot 1 200 10000 50.00
Spot 2 4800 4800 1.00
Spot 3 9000 300 0.03
ErbB2


Common Reference
Tumor or Normal Cells
RNA
cDNA
ErbB2
DNA microarray
Green
Red
Red / Green
Spot 1 200 10000 50.00
Spot 2 4800 4800 1.00
Spot 3 9000 300 0.03
ErbB2

NAME	BC/FUMI0	BC/FUMI4	BC/FUMI4	BC601B-A	BC601A-B	BC/FUMI1	BC/FUMI2	BC/FUMI2	BC/FUMI1	BC/FUMI1	BC102B-E	BC/FUMI2	BC/FUMI3	BC/FUMI3	BC/FUMI1	BC/FUMI1
adipose differentiation-related prote	0.242	1.21	-0.253	-0.841	-0.423	-0.363	-0.852	-1.383		-2.642	0.501	-0.25	-0.605	-0.636	0.229	-0.626
plasminogen activator, urokinase re	0.908	0.485	-0.397	-0.767	-0.886	-0.251	-0.683	0.057	-0.317	-1.2	0.125		-0.536	-0.248		-0.365
plasminogen activator, urokinase re	0.4635	0.3545	-0.8975	-1.23	-0.8335	0.0175	-1.002	0.1555	-0.4325	-1.008	-0.1785		-0.7445	-0.1485	0.0555	0.2055
coronin, actin binding protein, 1C A	0.551	0.151	-0.422	0.007	-0.638	0.087		-0.689	-0.91	-0.853	0.052	-0.492	-0.201	-0.152	-0.368	-0.741
**coatomer protein complex, subun	-1.061	-0.8655	-0.1235	-0.9895	0.3815	-0.4955	-0.2775	-0.1465	-1.109	-0.8635	0.2615	-0.0905	-0.3225	-0.6035	0.0195	-0.9345
coactosin-like protein R78490	-0.8835	-0.4545	0.2375	-1.177	0.2155	-0.2975	-0.9385	-0.2815	-1.494	-0.5985	0.4095	-0.3465	0.2185	-0.1345	-0.2895	-0.5525
folylpolyglutamate synthase R44864	0.686	1.583	1.313	0.048	-0.272	-0.143	-0.394	0.423	-0.445	-0.854	0.322	-0.03	-0.412	0.214	-1.098	-0.175
lysozyme (renal amyloidosis) N639	-0.18	1.155	1.575	-1.635	0.355	0.295	-0.805	0.135	-2.145	-0.955	0.575	0.735	-0.435	-0.855	-0.8	-1.705
chemokine (C-C motif) receptor 1 AA036881		0.524	1.233	-1.459	-0.095	-0.122	-0.196	0.101		-0.942	-0.2	-0.133	-0.549	-0.763		-0.059
interferon, gamma-inducible protei	-0.181	-0.062	0.37	0.064	0.418	-0.33	-0.098	-0.289	-1.042	-0.332	0.907	1.056	-0.8	-0.193	-0.789	-1.25
cystatin B (stefin B) H22919	-0.188	-0.489	-0.603	0.074	-0.212	-0.295	-0.54	-0.535	-0.453	-0.479	-0.021	0.291	-0.651	-0.536	-0.401	-0.511
cathepsin S AA236164	-0.791		0.334	-0.316	0.723	-0.46	0.39	-0.452		-0.413		1.063	-0.849	-1.088	-0.94	-1.291
small inducible cytokine A2 (monoc	0.2665	0.2955	0.5315	-0.1285	0.4255	-1.099	-0.7265	-0.6035	-1.052	-1.438	0.1355	0.0365	-0.4335	0.0875	-1.218	-0.7785
natural killer cell transcript 4 AA458	0.483	0.348	0.575	-0.685	0.971	-0.335	-0.222	-0.116	-1.644	-0.66	-0.322	0.885	-0.08	-0.02	-0.441	-0.51
superoxide dismutase 2, mitochonc	0.431	0.301	-0.836	0.519		-0.492	-0.834	-0.86		0.781	0.005	-1.163	-1.283	-0.969		-0.586
superoxide dismutase 2, mitochondrial AA487		0.3185	-0.6835	0.4865	0.6925	-0.7895	-0.6005	-0.5815	0.4995	0.0165	0.3755	-0.1225	-1.129	-1.137		-0.6935
transforming growth factor, beta-in	0.0235	0.6525	-0.3785	-0.5505	-0.3675	-0.4755	-0.1105	0.3435	0.0785	-0.4735	0.7925	1.532	-0.3355	-0.0885	0.2495	-0.1985
glycine dehydrogenase (decarboxy	-1.122	-1.412	-1.275	-1.764	-0.611	1.259	-1.25	-0.76	-2.159	-1.72	-1.017	-0.972	-0.715	-0.543	-0.658	-0.818
syndecan 2 (heparan sulfate protec	-1.828	-1.7	-1.409	-1.964	-0.975	1.516	-1.24	-1.75	-2.219	-2.477	-1.08	0.29	-1.641	-2.045	-0.315	-1.356
glutathione S-transferase pi R3364	-1.726	-1.892	-1.568			1.528	-1.346	-2.157	-3.114	-3.146	-0.943	0.236	-1.349	-1.674	-0.416	-1.557
chitinase 3-like 2 AA668821				-0.771	-1.436	-1.454	-0.813	-1.578	0.312	-0.167	0	-0.469	0.129	-0.566		-0.489
nuclear factor I/B W87528	0.464	-1.314	-0.187	-1.429	-0.189	0.551	-1.94	-1.372		-2.152	-1.825		-0.441	-0.928	0.316	-1.188
ras homolog gene family, member	-1.382	-0.471	-0.421	0.304	-0.448	-0.805	-0.945	-0.737	-1.222	-0.915	-0.713	-0.167	0.09	1.074		-0.393
ras homolog gene family, member	-1.311	-0.763	-0.61	0.198		-0.764	-0.391	-0.867	-1.469	-1.106	-0.486	-0.778	-0.579	0.812	0.348	-0.222
**zinc finger, DHHC domain containing 5 AA4		-0.965	-0.571	-0.304	-0.328	-0.417	-0.518	-0.473	-0.973	-0.94	-0.926	-1.153	-0.462	-0.683	0.828	0.347
keratin 5 (epidermolysis bullosa sin	-0.309	-0.485	-0.748	-0.909	-0.403	-0.127	-0.371	-0.778	-1.596	-1.787	-0.782	0.242	-0.559	-0.804	0.79	0.374
keratin 5 (epidermolysis bullosa simplex, Dowl		-0.655	-2.421	0.301	0.689	-0.38	-0.131	-1.647		-1.396	0.248	-1.118	-0.389	-1.423	1.963	-0.068
keratin 17 AA026100		-0.593	-2.294	0.181	-0.45	0.457	-1.132	-0.754		-2.708	-0.641	-0.148	-0.201	0.161	2.264	1.758
tripartite motif-containing 29 AA055	-0.523	-0.763		-0.726	-0.155	-0.401		-1.8	-1.591	-1.789	-1.076	-0.929	-1.132	-1.051		-0.24
pleiomorphic adenoma gene-like 1 AA463204	-0.7035	-0.5595	-0.7765	-0.2835			-0.1885		-1.466	-2.035	-0.1475	-0.7075	-0.4025	-1.054	0.3535	-0.5835
secreted frizzled-related protein 1 AA002080	-1.951	-2.022	-1.982	0.069	-0.117	-1.543	-2.996			-2.657	-0.275	-1.187	-0.262	-0.688	3.135	0.295
Homo sapiens cDNA FLJ11796 fis, clone HEN	-1.425		-0.74	-0.798	0.243	-0.225	-0.061			-0.957	-0.001		-0.491	-0.28	0.595	-0.721
ESTs AA074677		-0.411	-0.412	-0.879	-0.78	-0.401	-0.135	-0.508		-2.237	0.077		-0.72	-1.057		-1.301
pellino homolog 1 (Drosophila) W8	-0.3805	-1.159	-0.6945	-0.3935	-0.1785	-0.3665	-0.3835	-0.2825	0.1245	0.3185	0.2735	-1.329	-0.9455	-1.313		-0.4235
matrix metall																-0.04
moesin R22															0.24	-0.314
prion protein															2825	-0.0605
chitinase 3-l																-1.262
annexin A8															.619	0.519
hypothetical protein FLJ20481 N32	-0.078	-0.939	-1.002	0.058	-0.058	-0.158	-1.65	-0.794		-1.612	0.17	1.318	0.404	-0.312		-0.039
ADP-ribosylation factor-like 7 N353	-0.9415	-0.0585	-0.3685	-0.9365	-0.2155	0.0715	-0.2825	-0.5505		-1.107	-0.5855	0.2285	-0.2475	0.1635		-0.1405
cystatin A (stefin A) W72207		-0.532	-0.941	0.909	1.783	0.164	-0.106	-0.577		-1.496	0.588	3.351	-0.73			-0.855
inhibitor of DNA binding 3, domina	-0.46	-0.587	-0.421	-0.358	0.326		0.638		-0.642		-0.224	-0.143	-0.445	-0.58	0.377	
complement component 1, r subco	0.116		0.475	-1.506	0.089	-0.624	0.876	-1.115		-1.773	-0.505	-0.276	-0.204	-1.308	0.584	-0.431
nicotinamide N-methyltransferase T	0.675	-0.083	0.035	-0.244	0.053	-0.021	-0.365	-1.174	-1.235	-1.789	-0.688	0.972	-0.261	-0.532	0.606	0
myosin IE AA029956	-0.6075		-0.5465	-0.8195	-0.3755	-0.3535	-0.5545	-0.6505		-1.089	0.0005	-0.0205	0.1535	-0.1775		-0.0005
major histocompatibility complex, c	-0.494	-0.582	-1.091	-0.32	0.305	-0.098	-0.085	0.262	-1.668	-1.457	-0.039	-0.362	-0.218	-0.838	-0.197	-0.537
fatty acid binding protein 7, brain W72051			-1.595	-2.086	-1.717	-0.387	-2.433	-0.184			-1.441		-0.603	0.446		0.728
kynureninase (L-kynurenine hydrolase) H874		-0.342	-0.591	1.233	0.358	-0.954	-1.687	-1.194	-1.515	-2.291	-0.198	0.075	-0.657	-1.675	-0.58	-1.138
cytochrome P450, subfamily I (dio	1.065	-0.579	0	-0.767	0.392	-0.386	-0.479	-0.752	-0.401	-0.549	0.165	0.11	-0.605	-0.779	0.499	-0.131
cytochrome P450, subfamily I (dio	2.202	-0.047	-0.231	-0.604	-0.234	-0.713	-0.836	-1.99	-1.558	-1.474	0.425	0.622	-0.872	-1.706		-0.579
S100 calcium binding protein A8 (c	-1.641	0.014	-1.05	4.29	-0.162	-0.899	-1.625	-1.818		-2.268	-1.165	-1.2	-1.797	-1.329		-1.087
signal transducer and activator of transcription	-0.2855	-0.6135	2.59	-0.0555	-0.4895	-0.3215	-1.224	-1.718	-1.387	-0.4765	-0.7565	-1.143	-0.8755			-0.9545
gamma-aminobutyric acid (GABA) A receptor	3.044		-1.498	0.076	0.153	-0.766	-0.789			-1.485	-0.69		-0.823	-0.104		-0.235
EphB6 AA609284			0.6365	-1.062	-0.5295	-0.1345		-0.6565				-0.0415	-0.0885	0.0535		-0.3235
secretory leukocyte protease inhib	-2.088	-1.806	-1.596	0.434	-1.378	-1.269	-0.849	-1.961	-2.645	-3.187	-1.637		-0.996	-1.568	0.538	-1.344
aldo-keto reductase family 1, member C1 (dih	0.83	0.835	-0.435	1.743	1.173	-0.558	-1.21			-1.547		-0.834	0.712	0.104		-0.296
latrophilin W74533	-1.28	0.216	-0.322	-0.467	-0.563	0.111	0.383	-0.648	-0.95	-1.333	-0.903		0.469	0		0.274
echinoderm microtubule associated	-0.0045	-0.1755	-0.1055	0.1595	-0.2585	-0.5155	-1.05	-0.3895	-0.1395	-0.8335	-0.0445	-0.4845	0.0045	-0.7135	0.1435	-0.2125
epidermal growth factor receptor (erythroblast	-0.676	-1.527	-0.203	-1.07	-1.157	-0.979	-1.085			-2.181	-1.547	-0.782	-0.768			0.432
ERO1-like (S. cerevisiae) AA18680	0.3395	0.4075	-0.6115	-0.3415		-0.4095	-0.8285	-0.4075	-0.4405	-0.4075	-1.335	-0.2325	-0.7075	-0.7125		-0.3035
**hypothetical protein FLJ20624 R9	-0.232	-0.341		-0.153	-0.446	0.838	-1.658	0.466	-0.553	-0.909	-0.199	-0.662	-0.334	0.371	-0.029	0.617
forkhead box D1 AA069132	-1.192	-0.07	-0.666	-0.596	0.02	0.388	-0.54	0.11		-1.016	-1.28	-1.077	-0.051	-0.477	-0.561	0.796
met proto-oncogene (hepatocyte growth factor receptor) AA410591				-1.151	-0.373	0.228		0.322		-0.654	-0.23	0.176	0.74	0.446		0.177
ESTs, Weakly similar to TRHY_HUMAN TRICH		-0.137	-0.378	-0.339	0.263	0.618	0.02	0.518		-1.656	1.242	0.885	0.19	0.389		-0.164
ESTs AA149250				1.847	-0.295	0.293		-0.215		-1.641	-0.868	-0.33	-0.395	0.21		-0.185
Homo sapiens mRNA; cDNA DKFZp564O2364	3.135	0.0215	0.3145	0.7915	-0.2245	1.343	-0.0505			-0.1155	1.125		-0.1125	-0.5255		-0.3445
hypothetical protein FLJ10337 AA1	-0.6325	0.8375	0.2155	-0.7935	-0.1815	-0.3315	-0.4795	-0.6435	-0.6095	-0.4435	0.2225	0.0585	-0.3355	0.2375		0.1205
integral membrane protein 3 AA034	-0.998	-0.313	0.52	-1.106	-0.005	-0.618	-0.901	-0.847	-2.533	-2.828	-1.247	-0.647	-1.005	-1.651	-0.119	-0.465



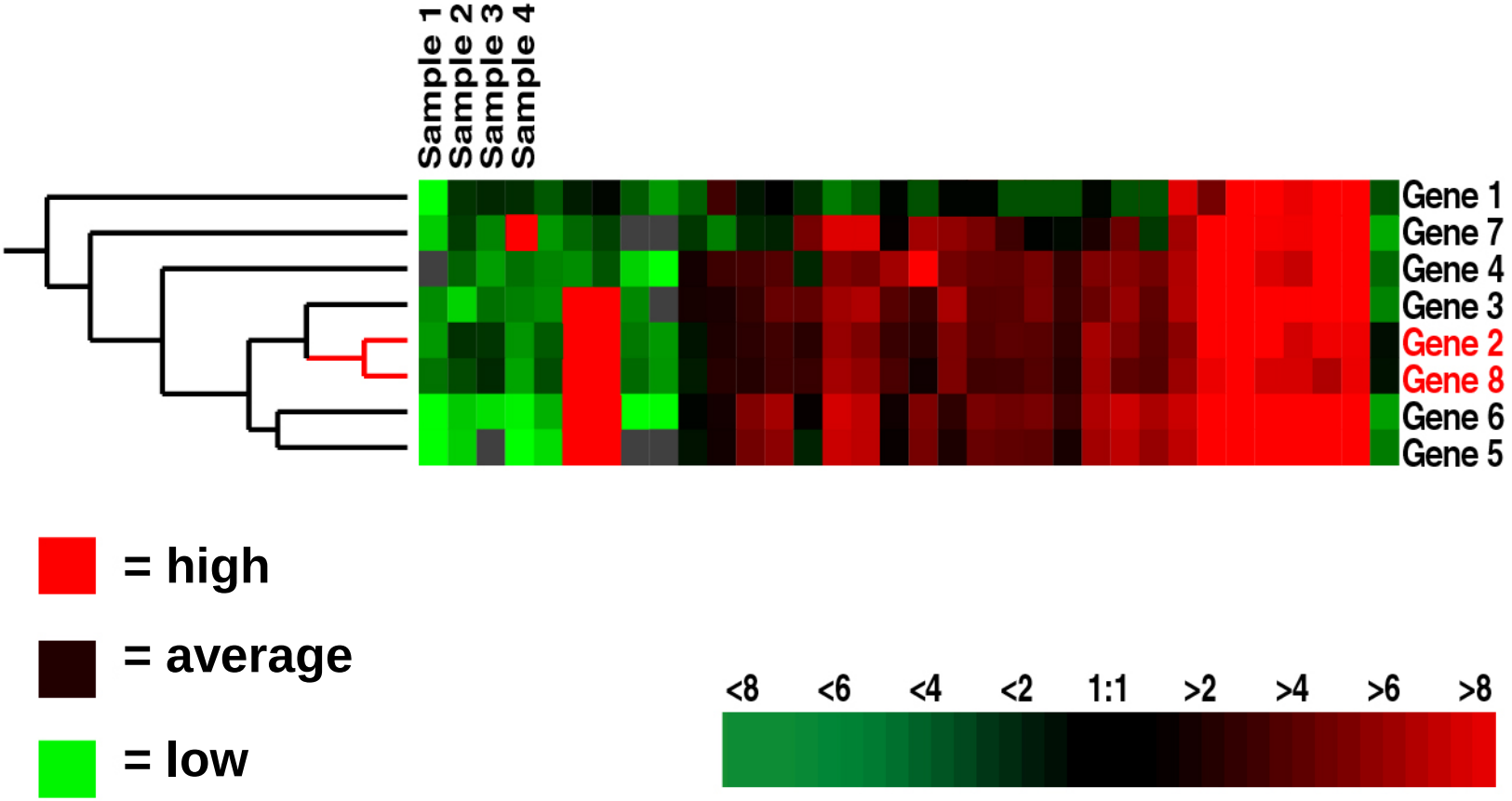


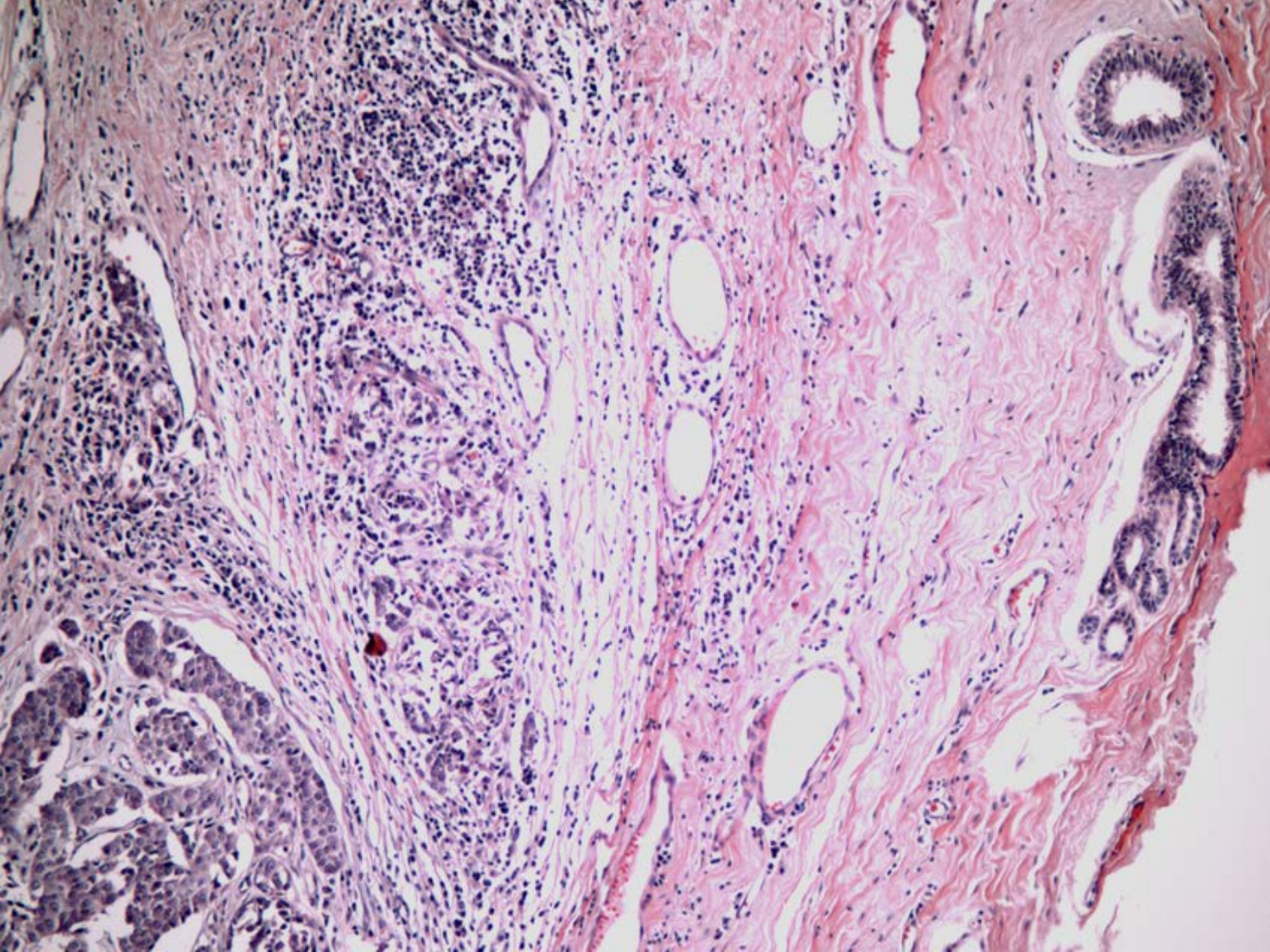


Cluster analysis and display of genome-wide expression patterns

MICHAEL B. EISEN, PAUL T. SPELLMAN, PATRICK O. BROWN, AND DAVID BOTSTEIN

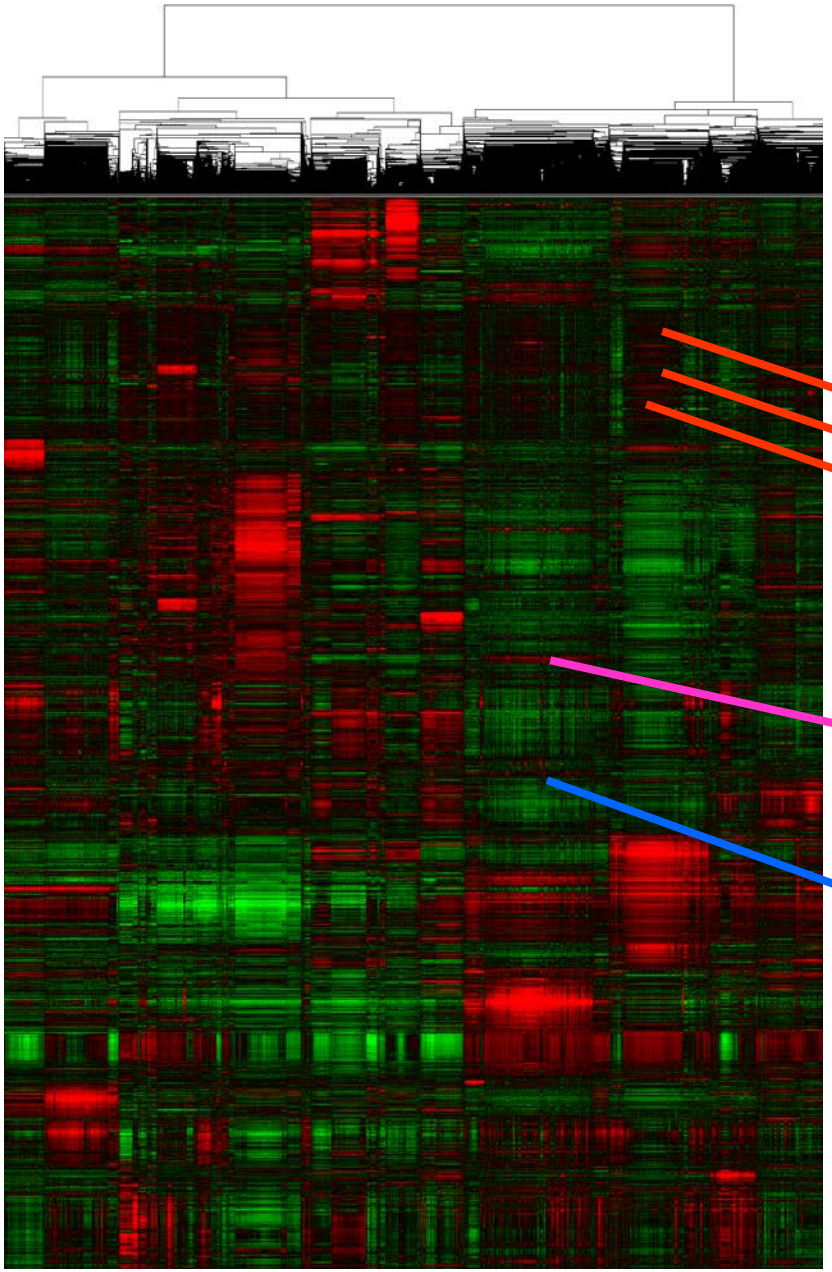
Proc.Natl. Acad.Sci.USA Vol. 95, pp. 14863-14868, December 1998

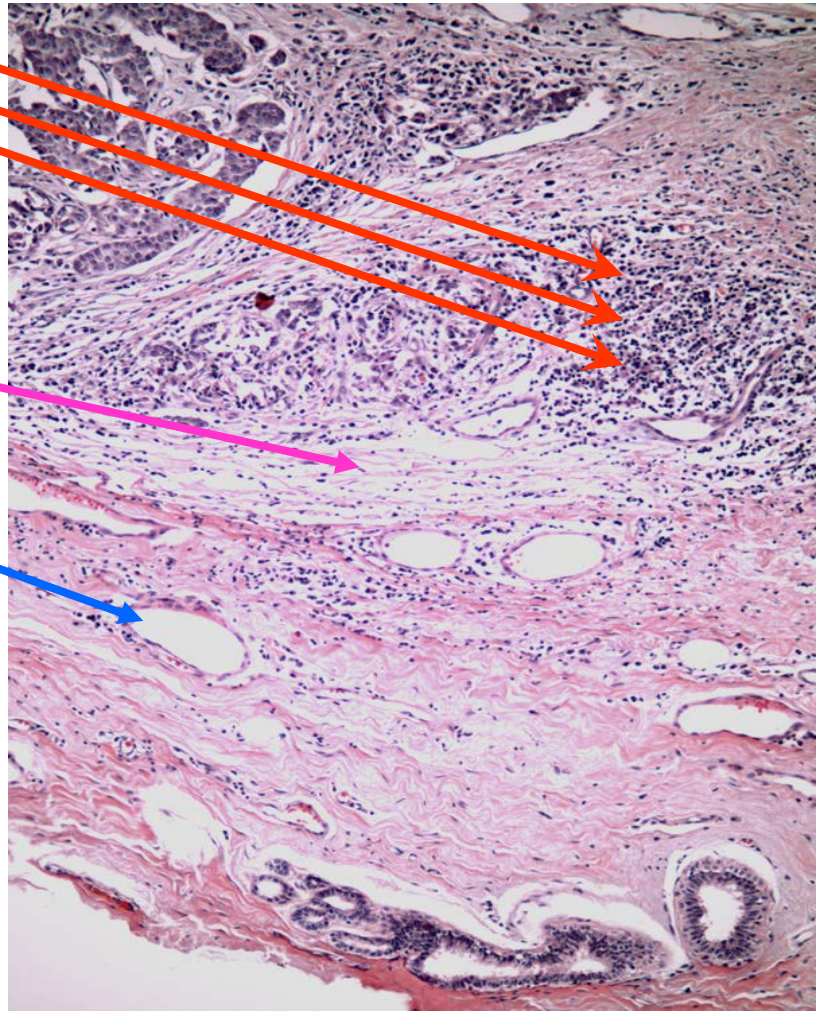


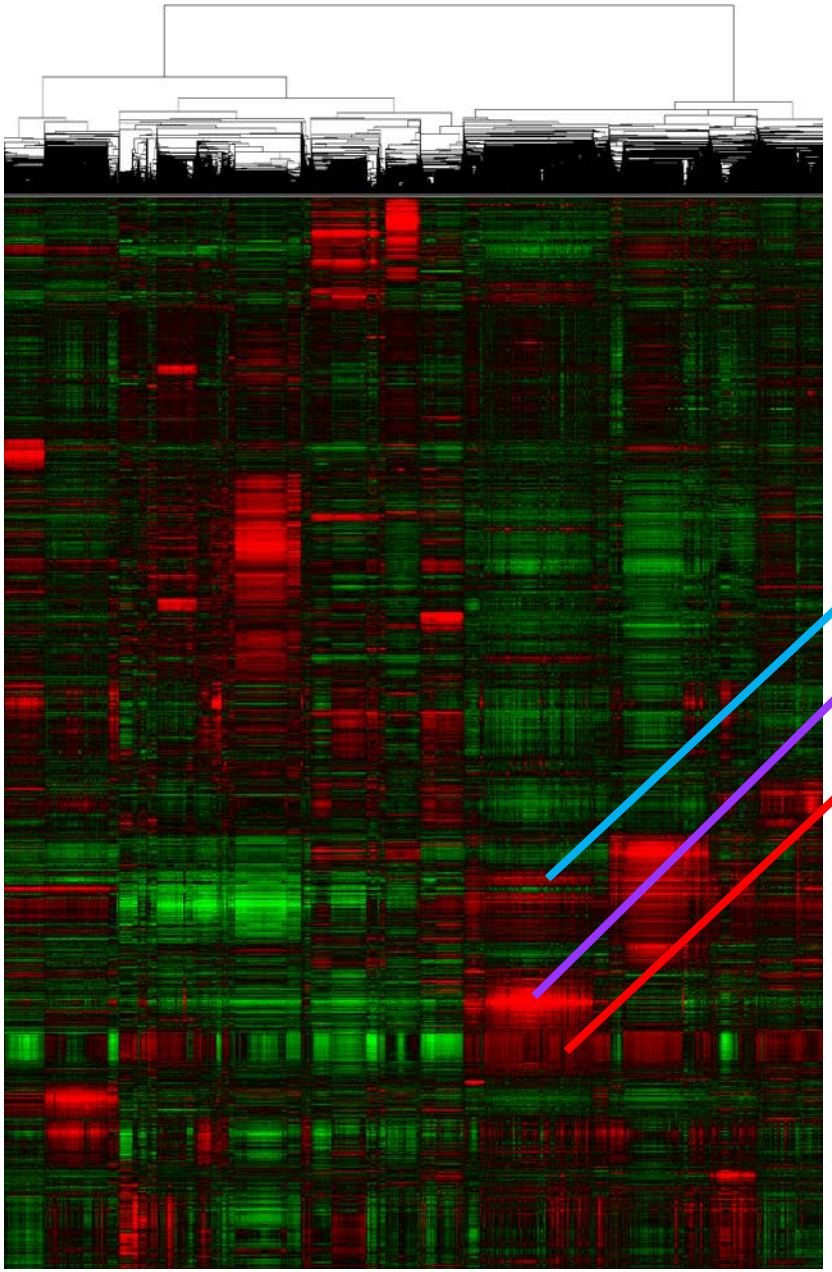


B-cells
T-cells
Macrophages
Fibroblasts
Endothelial Cells

Unsupervised cluster analysis of 10,000 grossly dissected human tumors

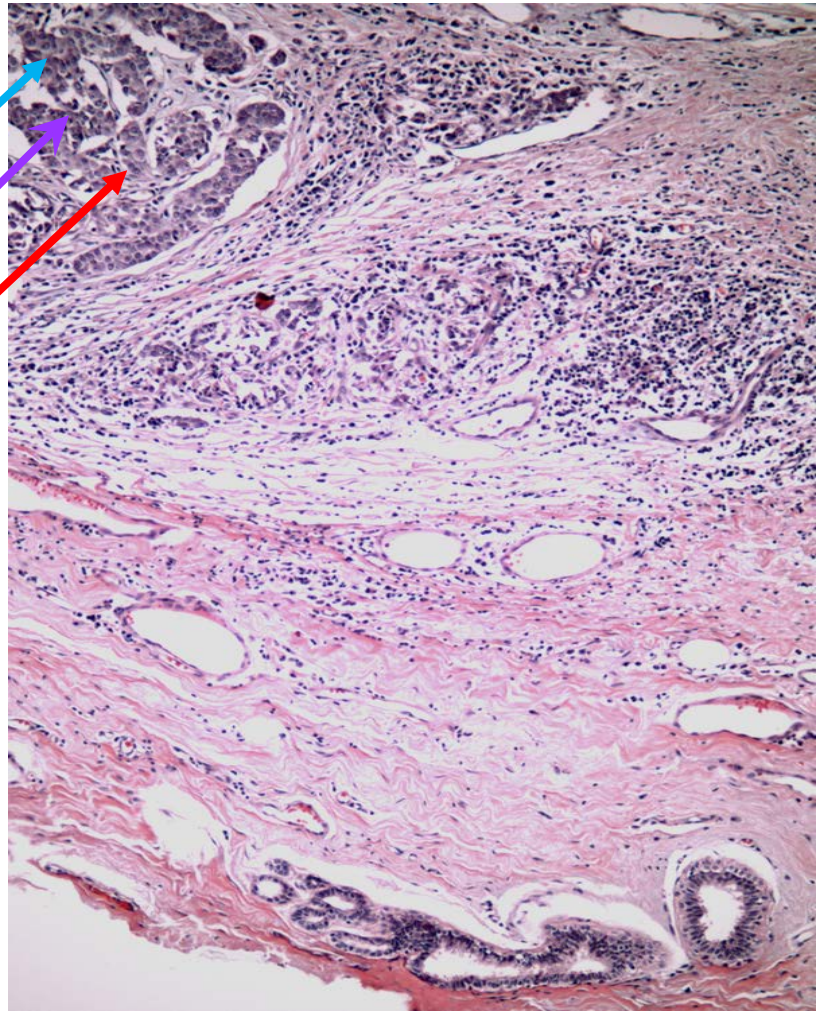






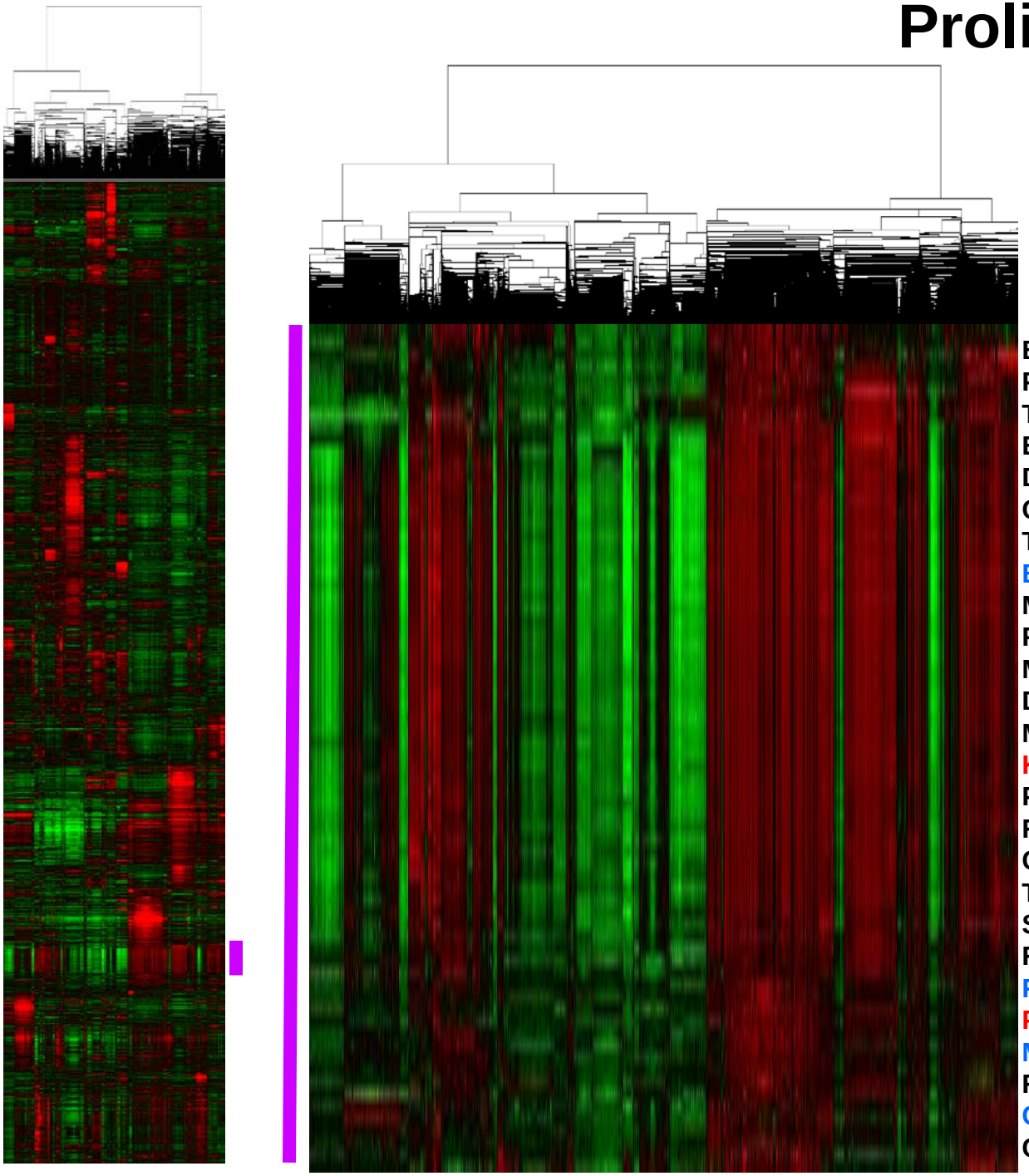
B-cells
T-cells
Macrophages
Fibroblasts
Endothelial Cells

Luminal/ER+ genes
HER2 amplicon
Proliferation



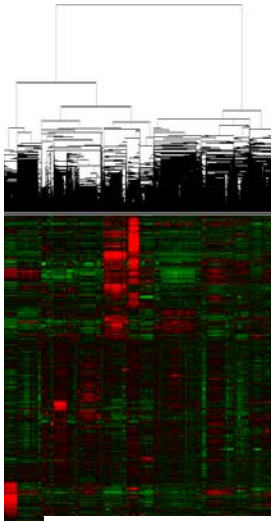
Proliferation Gene Set

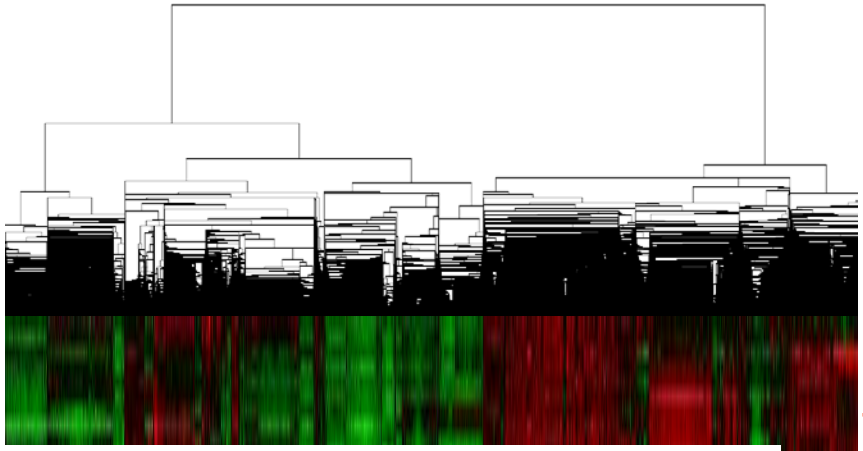
Whitfield et al.
Common Markers of Proliferation
Nature Reviews Cancer
Vol. 6; 99-106 (2006)
PMID:12058064

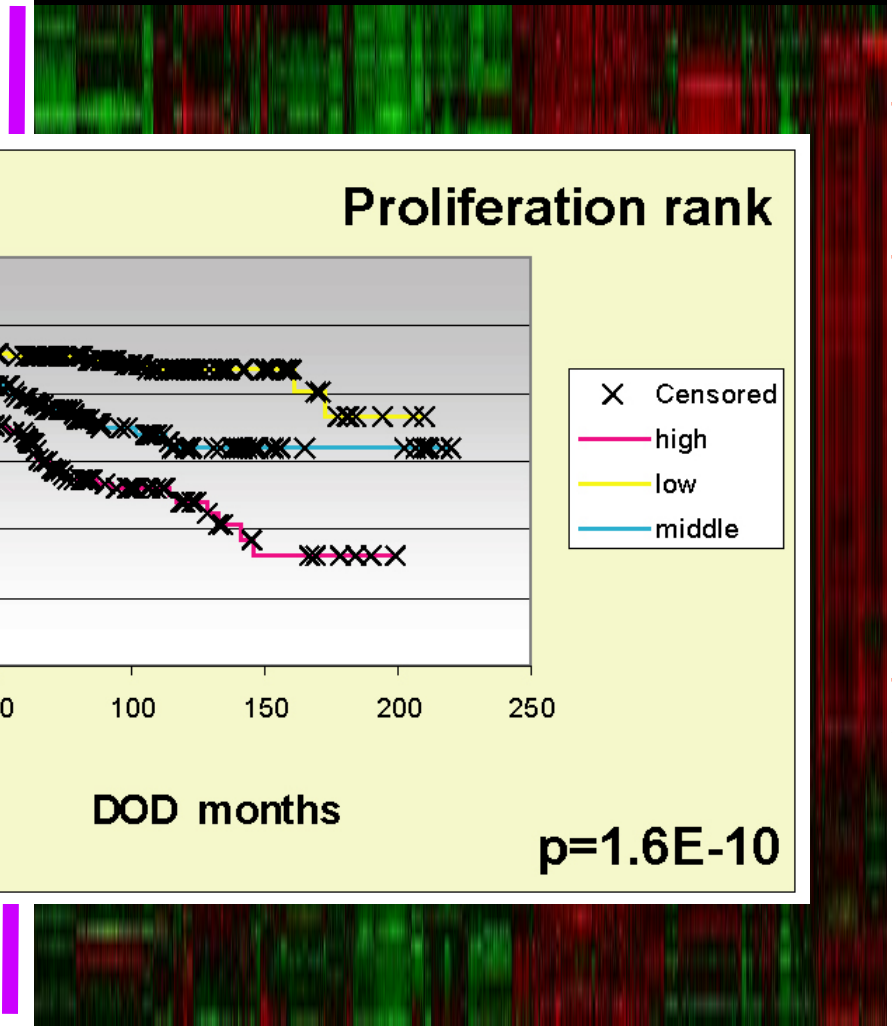


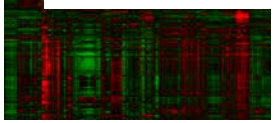
BUB1
PLK1
Thymidylate Synthetase
EZH2
DNA Polymerase alpha
Cyclin A2, B2, E1
Tubulin
BRCA1, 2
MCM2, 3, 5, 6, 7, 8, 10
Forkhead Box M1
MAD2
Dihydrofolate reductase
MYBL2
Ki-67
PTTG1
Replication Factor C
CENPA, E, F, H
TOP2a
STK6/15
RAD51
FANCA
PCNA
MSH2
Ribonucleotide reductase
CHEK1
CDC1, 2, 7, 8, 20, 25

Proliferation Gene Set









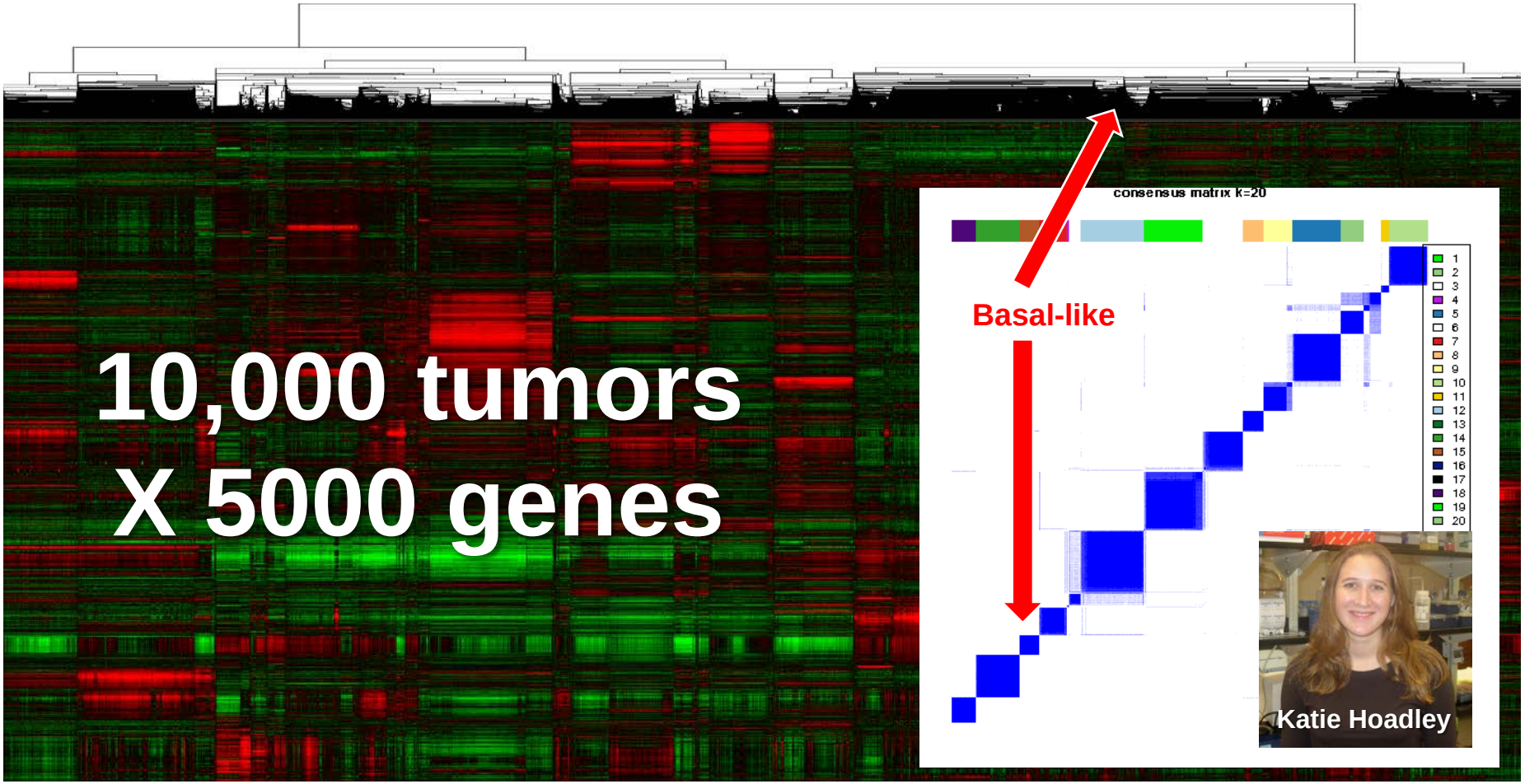
Chemotherapeutic drug targets = taxanes, anthracyclines, cytoxan, cis/carboplatin, capecitabene, gemcitabene, methotrexate, and more

BUB1
PLK1
Thymidylate Synthetase = 5FU
EZH2
DNA Polymerase alpha
Cyclin A2, B2, E1
Tubulin = taxanes
BRCA1, 2
MCM2, 3, 5, 6, 7, 8, 10
Forkhead Box M1
MAD2
Dihydrofolate reductase = methotrexate
MYBL2
Ki-67
PTTG1
Replication Factor C
CENPA, E, F, H
TOP2a = doxorubicin, etoposide
STK6/15
RAD51
FANCA
PCNA
MSH2
Ribonucleotide reductase = hydroxyurea
CHEK1
CDC1, 2, 7, 8, 20, 25

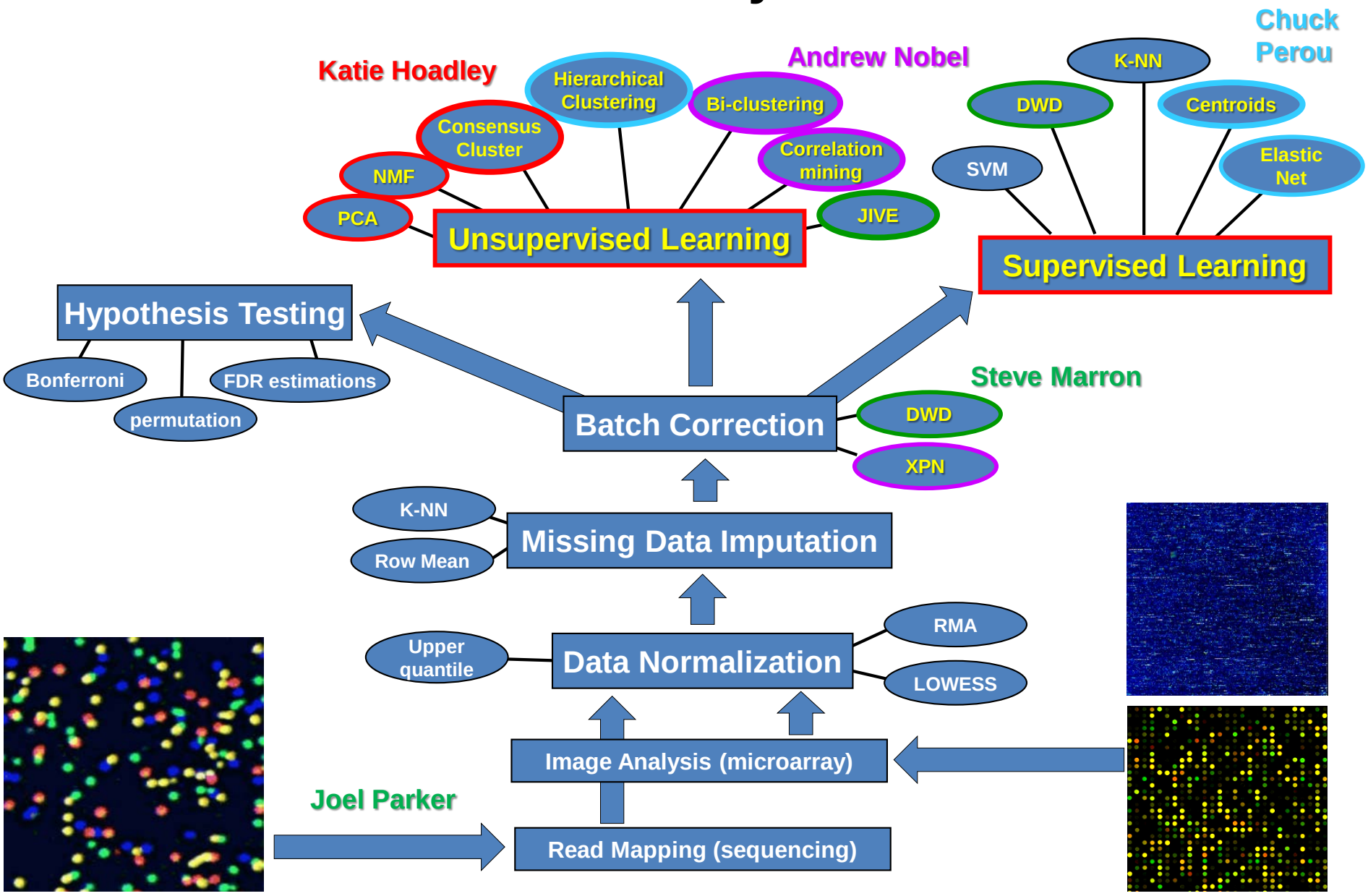
TCGA Final Production Totals from UNC

9906 tumors mRNA-sequenced and uploaded to the CGHub public repository (https://cghub.ucsc.edu/), which represents >100 trillion bases

33 tumor types studied including breast, bladder, colon, rectum, head & neck, renal clear cell & chromophobe, lung squamous & adenocarcinoma, melanoma, ovarian, glioblastoma, prostate, endometrial, thyroid, and many others



A Complex Picture of Microarray/Gene Expression Data Analysis



Advice for Interpreting Genomic Analysis Results (or Big Data Results in General)

1. **Does the selected feature set make biological sense and/or have a common annotation theme?**

2. **What was the False-Discovery Rate?**

3. **Was there a training set and an independent test set?**

4. **Has the new feature(s) been tested in a multivariable analysis with previously known important features?**

Note = All these questions are true for ANY multi-analyte assay/result, and also true for ANY single analyte assay/result

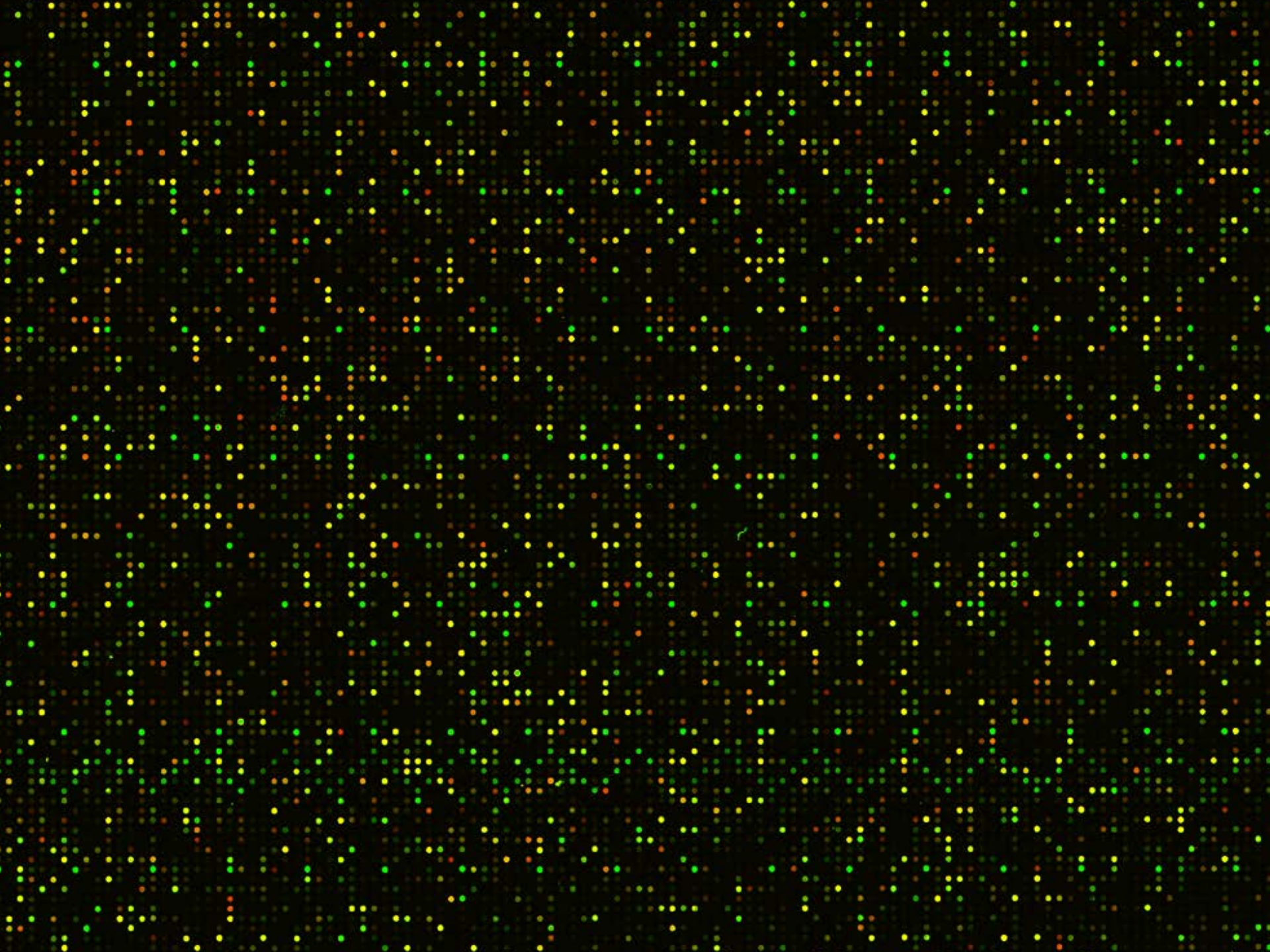
Current Gene Expression Technologies

1. DNA microarrays
   - Moderate precision covering ALL genes, fast, and inexpensive
     - long oligos (60-70mers/Agilent/Nimblegen) – 1-2 oligos/gene
     - short oligos (~25mers/Affymetrix) – 5-10 oligos/gene
     - cDNA arrays (100-1000bp PCR products/Stanford) – 1-2 cDNA clones/gene

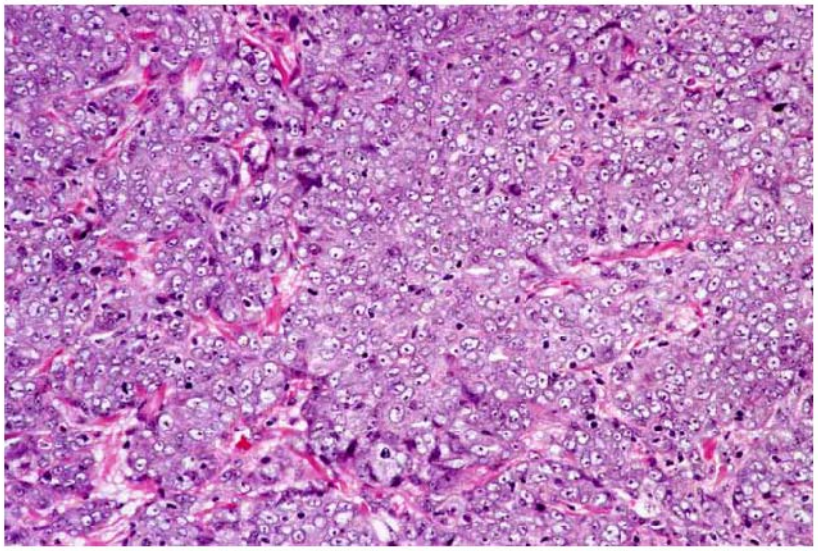
2. Quantitative RT-PCR
   - high precision, large dynamic range, fast, and inexpensive
   - smaller number of genes assayed (compatible with Formalin-fixed, Paraffin Embedded Materials (FFPE), which is the medical standard of care)

3. Nanostring nCounter
   - high precision, large dynamic range, fast, and moderate expense
   - single mRNA molecule counting capabilities (compatible with FFPE)

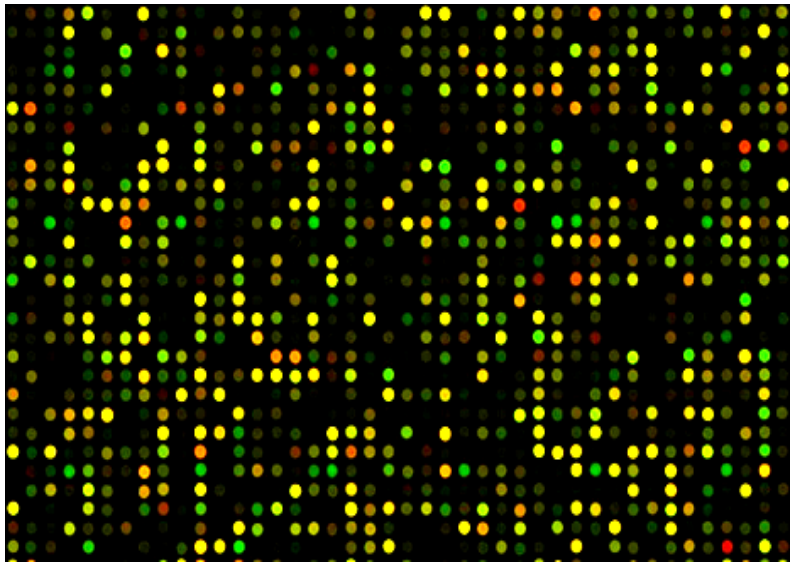
4. Illumina-based RNA-sequencing
   - Gene expression via Massively Parallel Sequencing (MPS)
   - high precision, large dynamic range, not fast, and expensive
   - compatible with FFPE

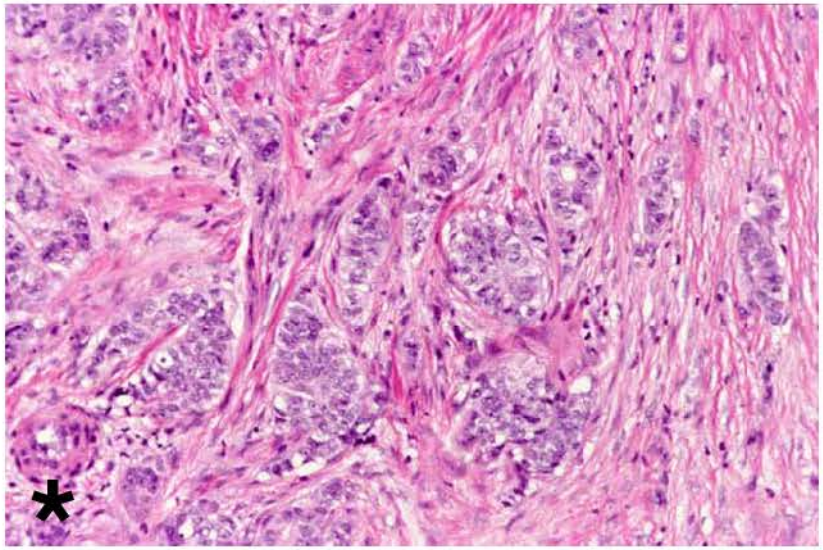


Breast Tumor, Grade III, Estrogen and Progesterone Receptor Positive, HER2-negative T3 tumors (diameter >5cm), N1 (lymph node metastasis positive),

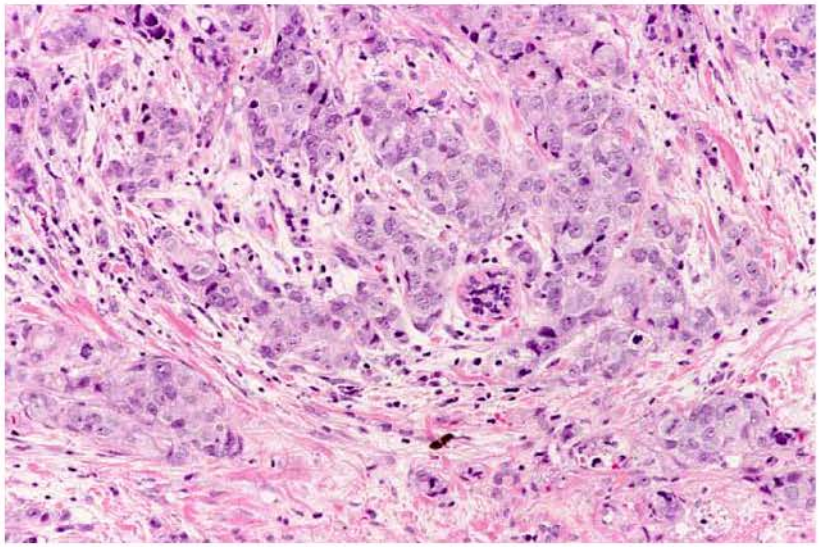


Died after 9 months





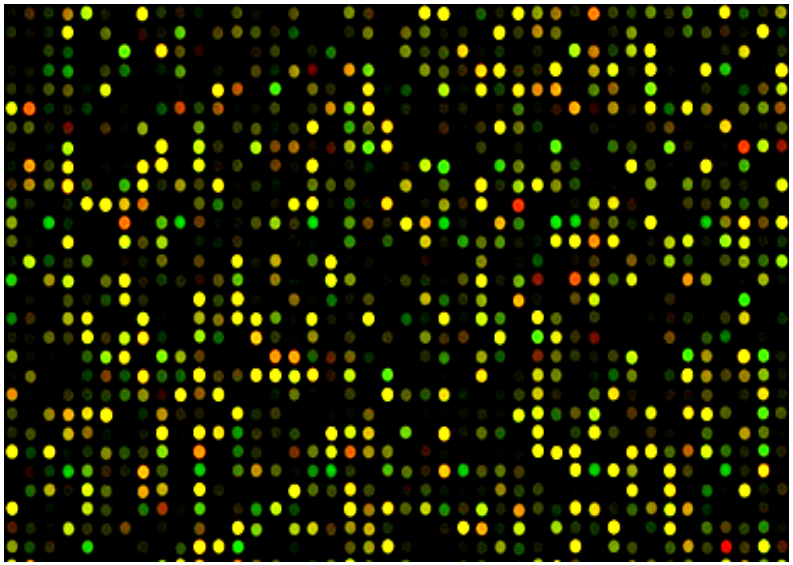
Died after 17 months

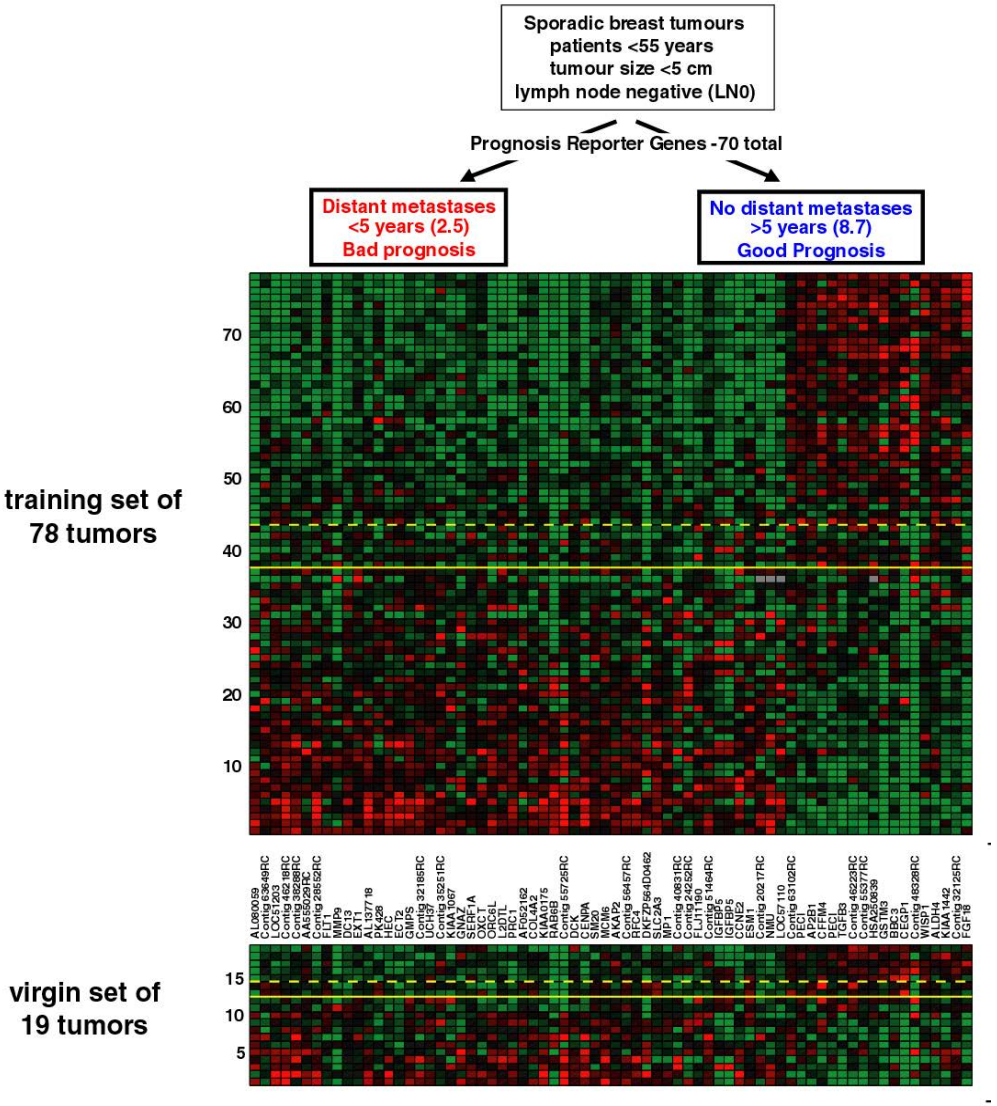


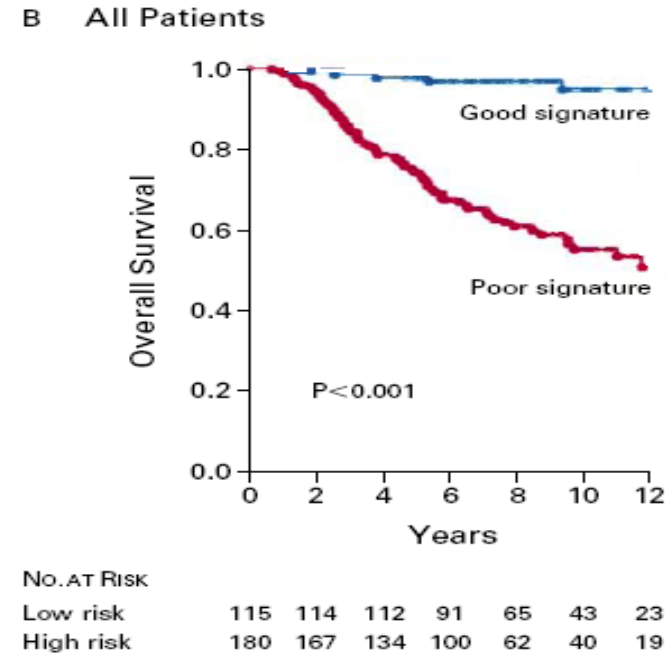
Alive for 96 months

Agendia MammaPrint Assay

(Agilent DNA Microarray)







Van't Veer et al. Nature 415:530-536, 2002
Van de Viiver et al. N Engl J Med. 347:1999-2009, 2002.

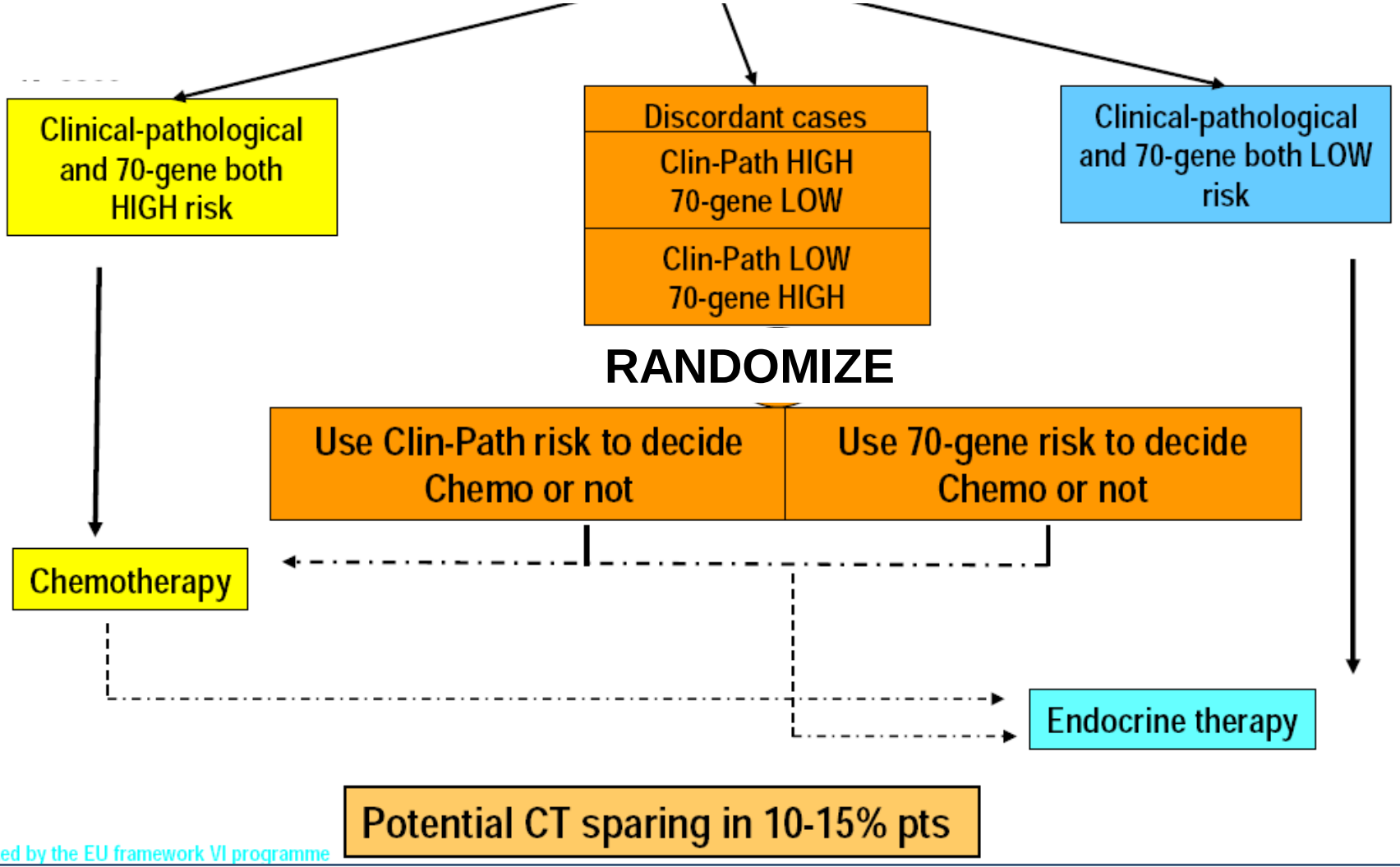
EORTC Trial 10041 (BIG 3-04), MIND-ACT

Total accrual (updated on 10/14/2009):

enrollment completed = ~6000 patients

http://www.eortc.be/services/unit/mindact/MINDACT_websiteii.asp

A prospective randomized study comparing the 70-gene signature with the common clinical-pathological criteria in selecting patients for adjuvant chemotherapy in breast cancer with 0-3 positive nodes.

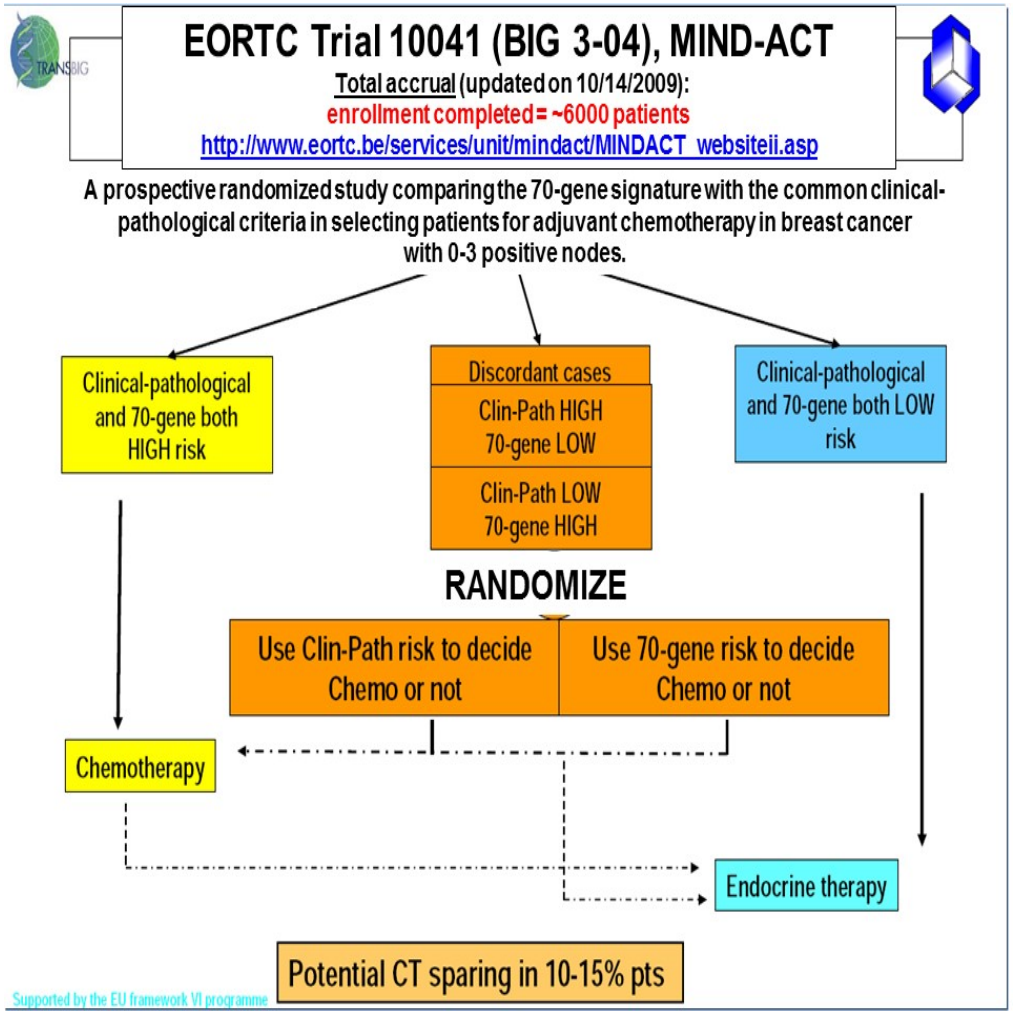


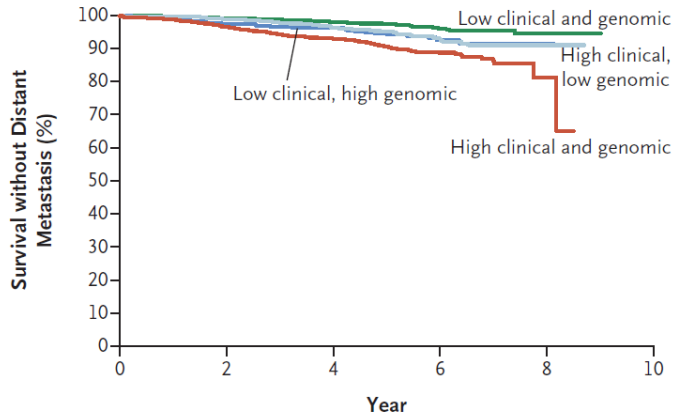
Supported by the EU framework VI programme

The NEW ENGLAND JOURNAL of MEDICINE
ESTABLISHED IN 1812 AUGUST 25, 2016 VOL. 375 NO. 8

70-Gene Signature as an Aid to Treatment Decisions in Early-Stage Breast Cancer

F. Cardoso, L.J. van't Veer, J. Bogaerts, L. Slaets, G. Viale, S. Delaloge, J.-Y. Pierga, E. Brain, S. Causeret, M. DeLorenzi, A.M. Glas, V. Golfinopoulos, T. Goulioti, S. Knox, E. Matos, B. Meulemans, P.A. Neijenhuis, U. Nitz, R. Passalacqua, P. Ravdin, I.T. Rubio, M. Saghatchian, T.J. Smilde, C. Sotiriou, L. Stork, C. Straehle, G. Thomas, A.M. Thompson, J.M. van der Hoeven, P. Vuylsteke, R. Bernards, K. Tryfonidis, E. Rutgers, and M. Piccart, for the MINDACT Investigators*





No. at Risk

Low clinical and genomic risk	2745	2628	2331	735	33
Low clinical, high genomic risk	592	550	484	136	2
High clinical, low genomic risk	1550	1457	1317	311	9
High clinical and genomic risk	1806	1689	1462	395	11

Figure 3. Survival without Distant Metastasis in the Four Risk Groups.
The analysis includes all enrolled patients, and the risk groups are based on corrected risk. The time-to-event curves were estimated by means of the Kaplan–Meier method.

Genomic Microarray-based platforms (and NGS) can currently assay for

1. **Gene Expression: typically thought of as mRNA/protein coding genes = ~25,000 genes**

2. **MicroRNAs: important class of small regulatory RNA molecules = ~1500**

3. **Long Non-Coding RNAs (lncRNAs): are mRNAs with PolyA tails and exons, but do not code for proteins = 40,000 - 50,000**

4. **DNA Methylation = 450,000 methylation sites**

5. **Germline DNA Polymorphisms (>1,000,000 SNPs) and Tumor DNA Copy Number Changes**

Current Gene Expression Technologies

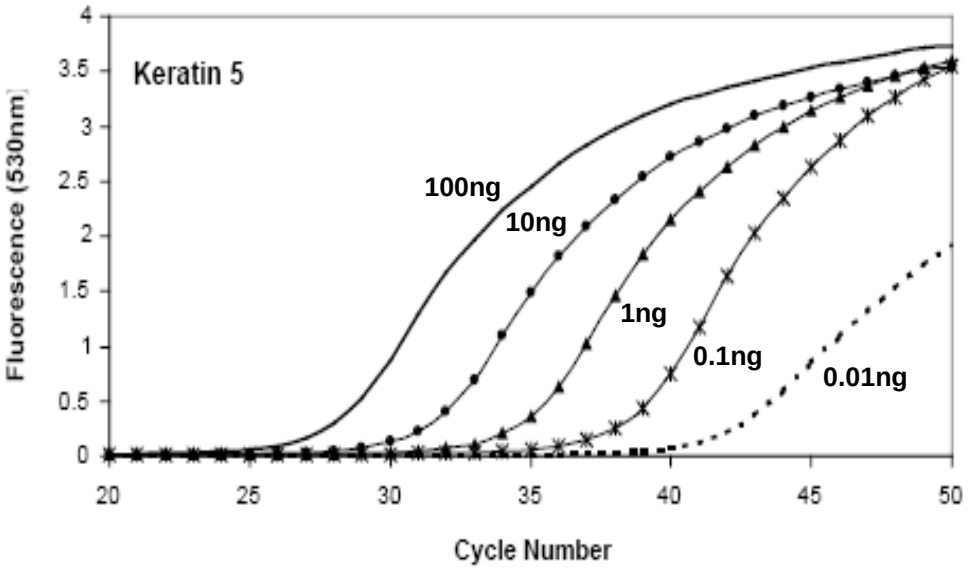
1. DNA microarrays
   - Moderate precision covering ALL genes, fast, and inexpensive
     - long oligos (60-70mers/Agilent/Nimblegen) – 1-2 oligos/gene
     - short oligos (~25mers/Affymetrix) – 5-10 oligos/gene
     - cDNA arrays (100-1000bp PCR products/Stanford) – 1-2 cDNA clones/gene

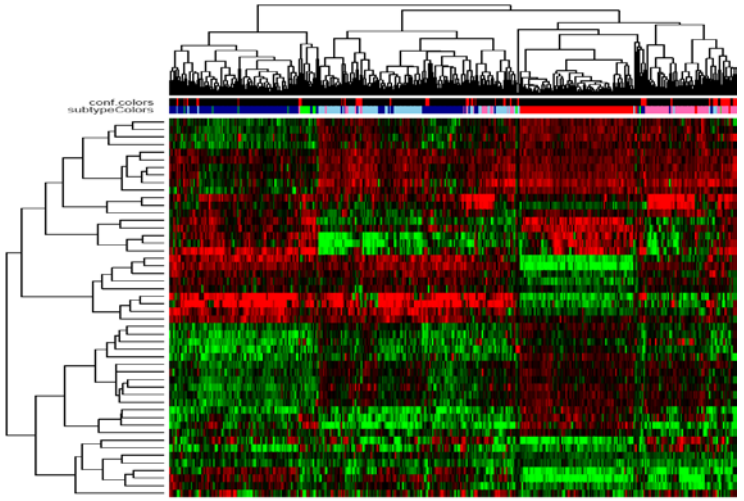
2. Quantitative RT-PCR
   - high precision, large dynamic range, fast, and inexpensive
   - smaller number of genes assayed (compatible with Formalin-fixed, Paraffin Embedded Materials (FFPE), which is the medical standard of care)

3. Nanostring nCounter
   - high precision, large dynamic range, fast, and moderate expense
   - single mRNA molecule counting capabilities (compatible with FFPE)

4. Illumina-based RNA-sequencing
   - Gene expression via Massively Parallel Sequencing (MPS)
   - high precision, large dynamic range, not fast, and expensive
   - compatible with FFPE

Quantitative RT-PCR

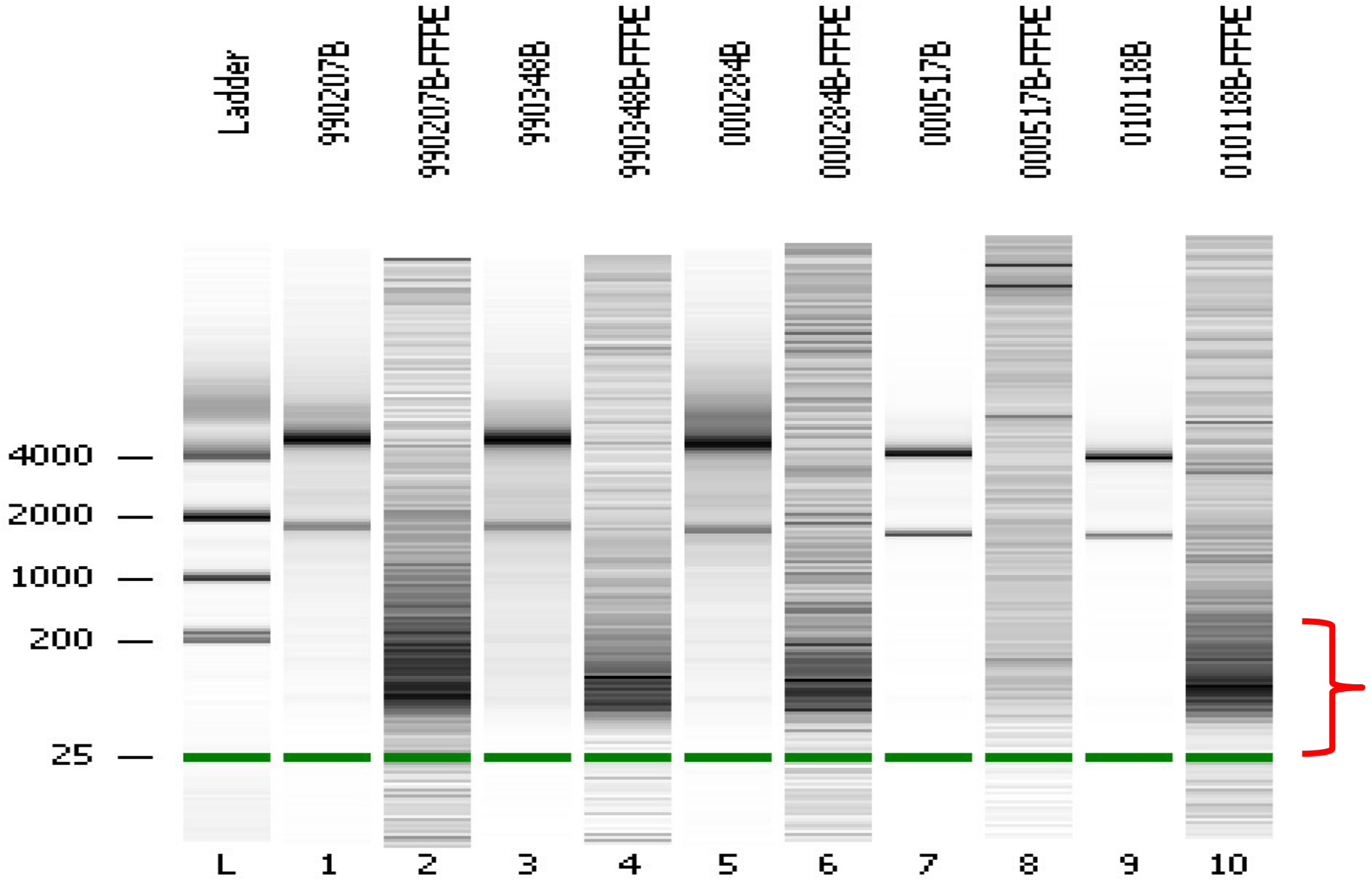




ABI 7900HT Fast Real-Time PCR System

Roche Lightcycler 480

Formalin-Fixed Paraffin Embedded (FFPE) RNAs pose a unique challenge for nucleic acid based studies



A Multi-gene Assay to Predict Recurrence of Tamoxifen-Treated Node-Negative Breast Cancer.
Paik et al., The New England Journal of Medicine, 351:2817-26 (2004)

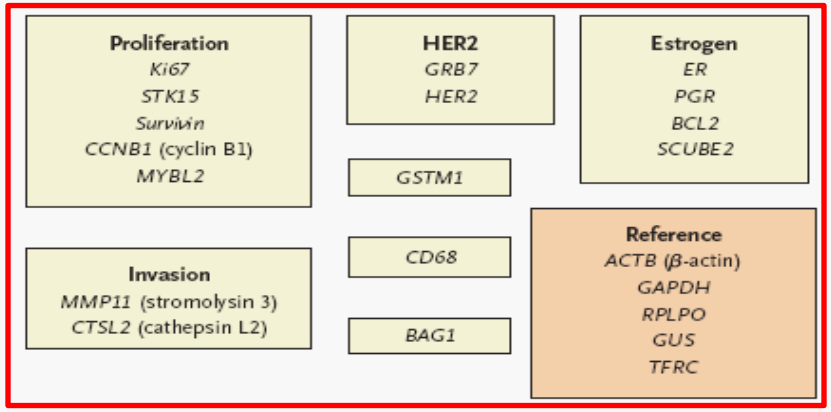


Figure 1. Panel of 21 Genes and the Recurrence-Score Algorithm.

The recurrence score on a scale from 0 to 100 is derived from the reference-normalized expression measurements in four steps. First, expression for each gene is normalized relative to the expression of the five reference genes (*ACTB* [the gene encoding β-actin], *GAPDH*, *GUS*, *RPLPO*, and *TFRC*). Reference-normalized expression measurements range from 0 to 15, with a 1-unit increase reflecting approximately a doubling of RNA. Genes are grouped on the basis of function, correlated expression, or both. Second, the *GRB7*, *ER*, proliferation, and invasion group scores are calculated from individual gene-expression measurements, as follows: *GRB7* group score = $0.9 \times GRB7 + 0.1 \times HER2$ (if the result is less than 8, then the *GRB7* group score is considered 8); *ER* group score = $(0.8 \times ER + 1.2 \times PGR + BCL2 + SCUBE2) \div 4$; proliferation group score = $Survivin + KI67 + MYBL2 + CCNB1$ [the gene encoding cyclin B1] $+ STK15) \div 5$ (if the result is less than 6.5, then the proliferation group score is considered 6.5); and invasion group score= (*CTSL2* [the gene encoding cathepsin L2] + *MMP11* [the gene encoding stromolysin 3]) ÷2. The unscaled recurrence score (RS_U) is calculated with the use of coefficients that are predefined on the basis of regression analysis of gene expression and recurrence in the three training studies[24-26]: $RS_U = +0.47 \times GRB7$ group score $- 0.34 \times ER$ group score $+ 1.04 \times$ proliferation group score $+ 0.10 \times$ invasion group score $+ 0.05 \times CD68 - 0.08 \times GSTM1 - 0.07 \times BAG1$. A plus sign indicates that increased expression is associated with an increased risk of recurrence, and a minus sign indicates that increased expression is associated with a decreased risk of recurrence. Fourth, the recurrence score (RS) is rescaled from the unscaled recurrence score, as follows: $RS = 0$ if $RS_U < 0$; $RS = 20 \times (RS_U - 6.7)$ if $0 \le RS_U \le 100$; and $RS = 100$ if $RS_U > 100$.

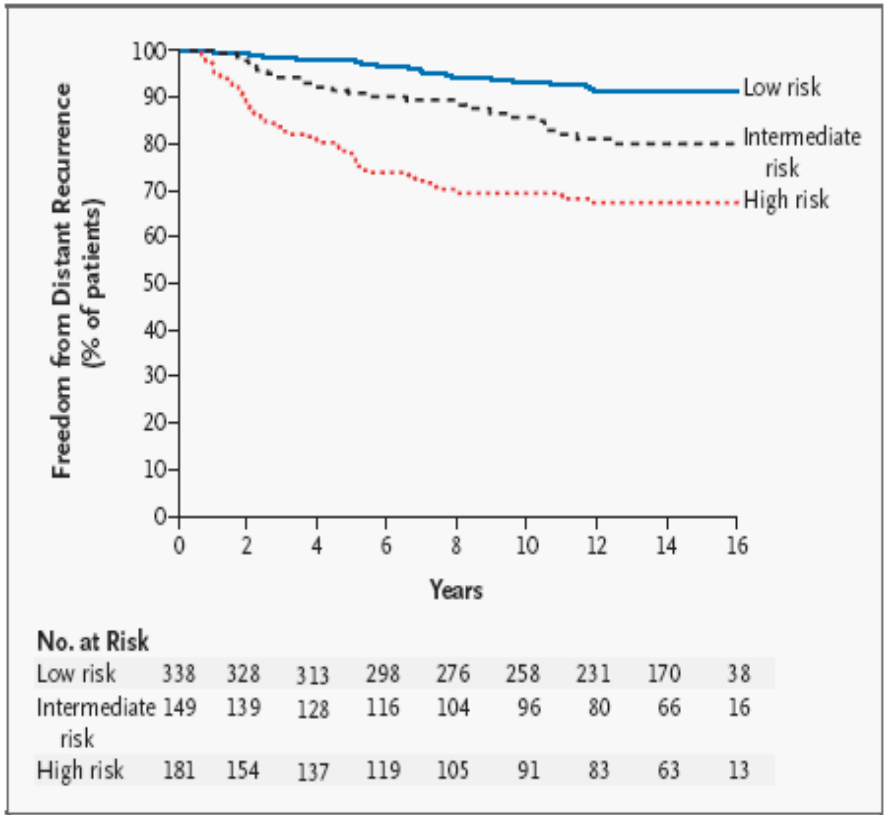


Figure 2. Likelihood of Distant Recurrence, According to Recurrence-Score Categories.

A low risk was defined as a recurrence score of less than 18, an intermediate risk as a score of 18 or higher but less than 31, and a high risk as a score of 31 or higher. There were 28 recurrences in the low-risk group, 25 in the intermediate-risk group, and 56 in the high-risk group. The difference among the groups is significant ($P<0.001$).

Gene Expression and Benefit of Chemotherapy in Women With Node-Negative, Estrogen Receptor-Positive Breast Cancer

Paik et al., Journal of Clinical Oncology, 24: 1-12 (2006).

NSABP B-20 (651 tumor FFPE samples)

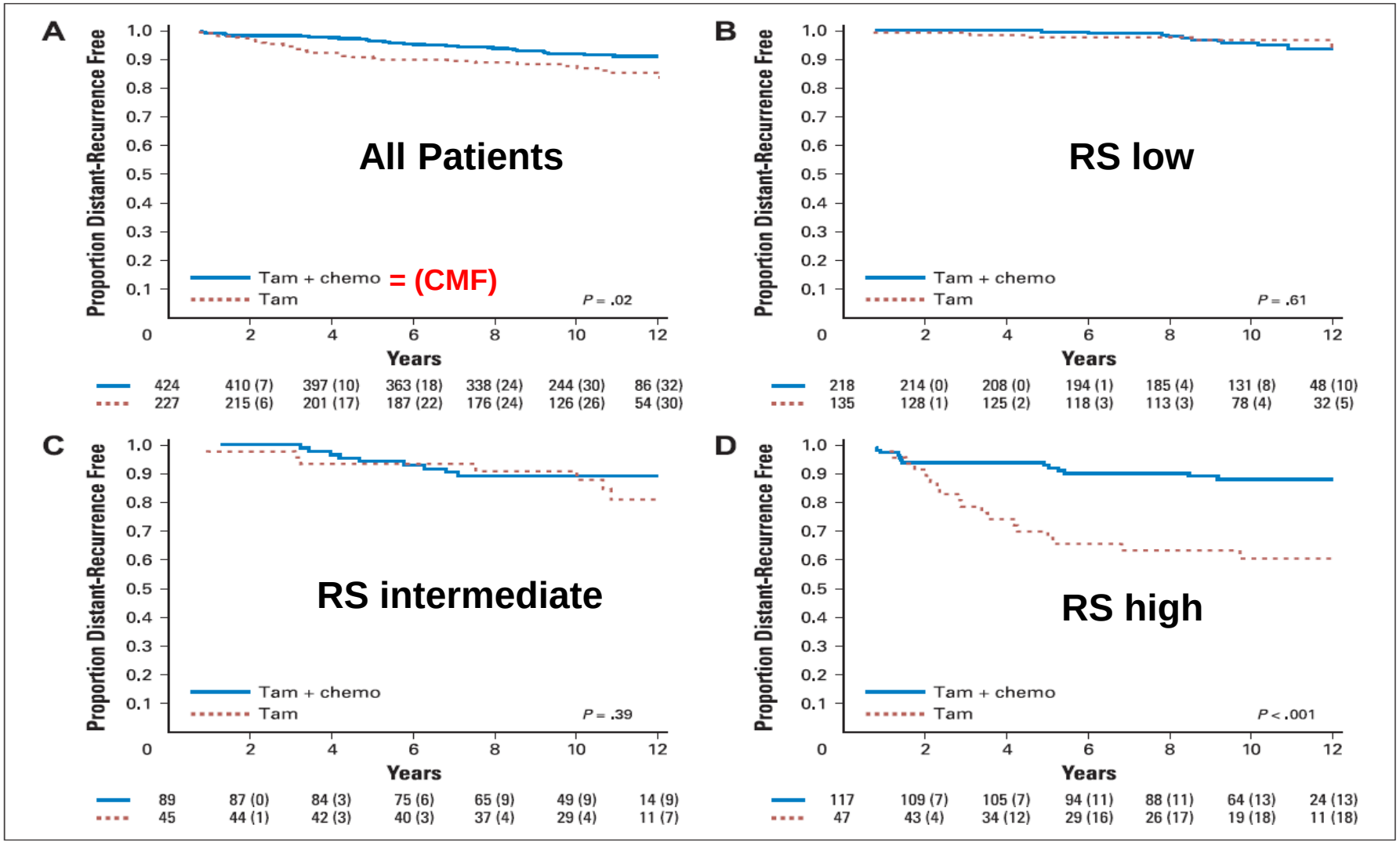
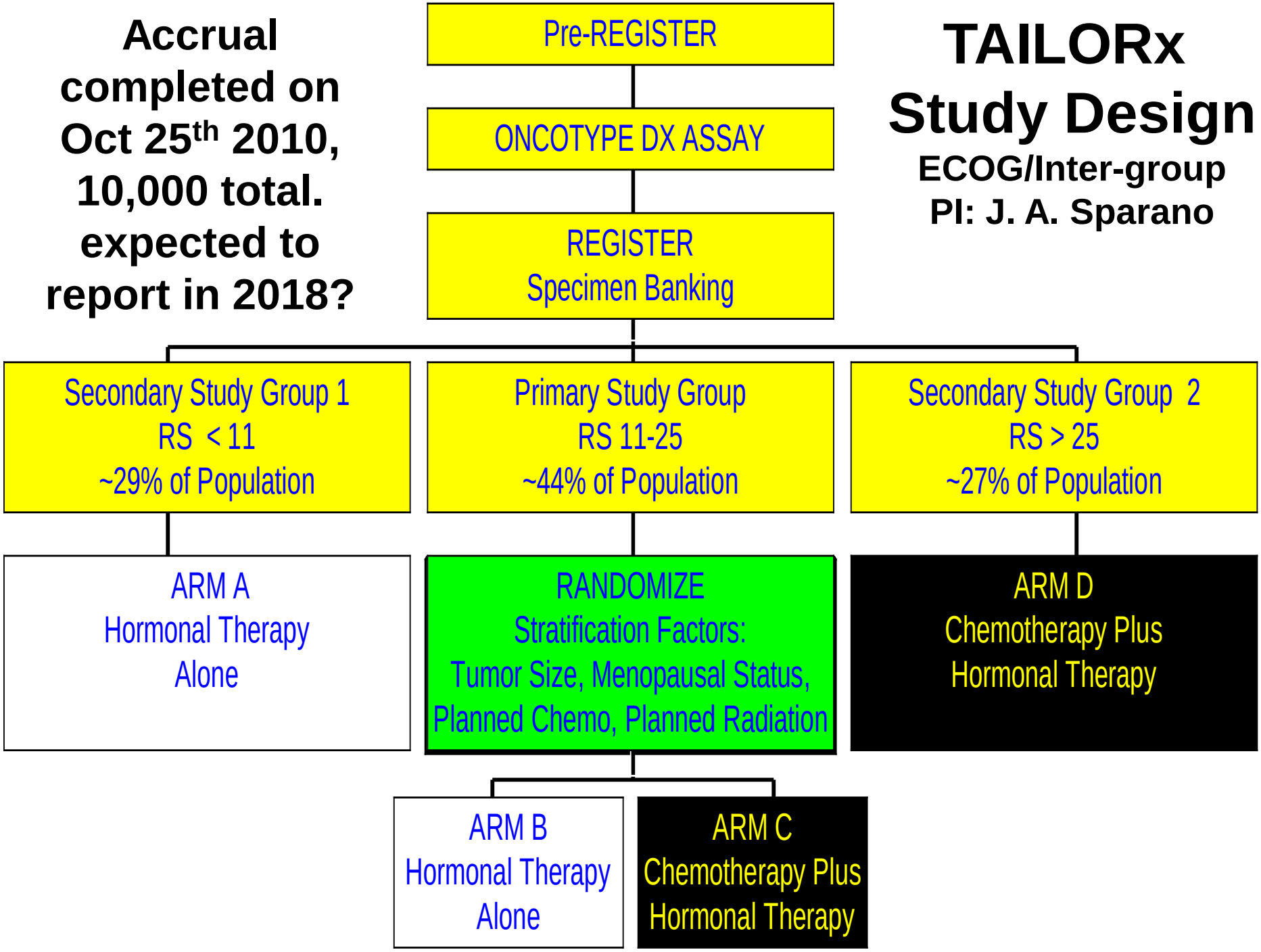


Fig 2. Kaplan-Meier plots for distant recurrence comparing treatment with tamoxifen (Tam) alone versus treatment with tamoxifen plus chemotherapy (Tam + chemo). (A) All patients; (B) low risk (recurrence score [RS] < 18); (C) intermediate risk (RS 18-30); (D) high risk (RS ≥ 31). The number of patients at risk and the number of distant recurrences (in parentheses) are provided below each part of the figure.

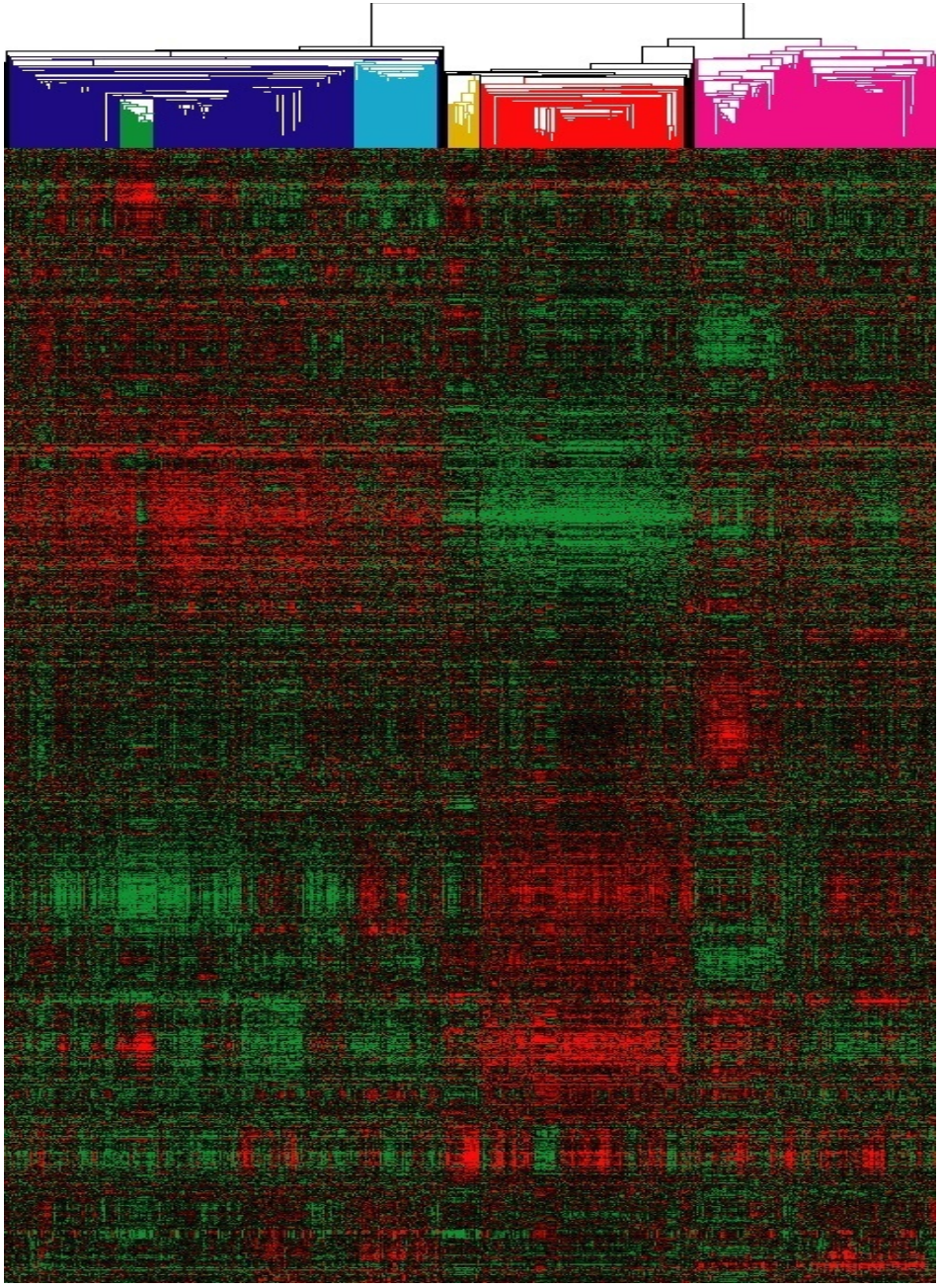
TAILORx Study Design

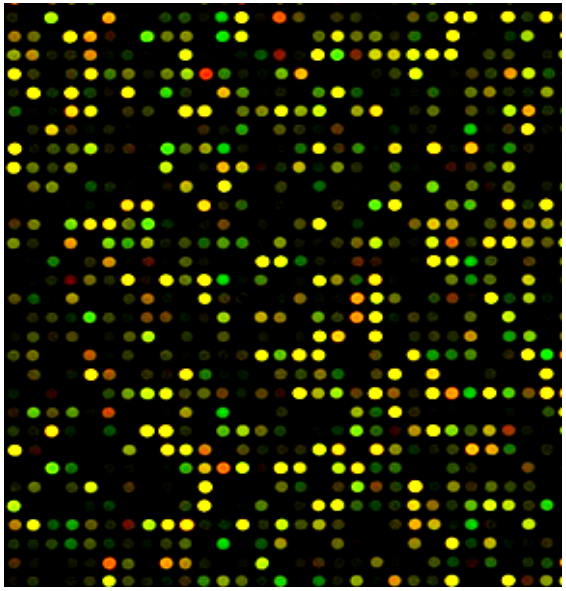
ECOG/Inter-group
PI: J. A. Sparano

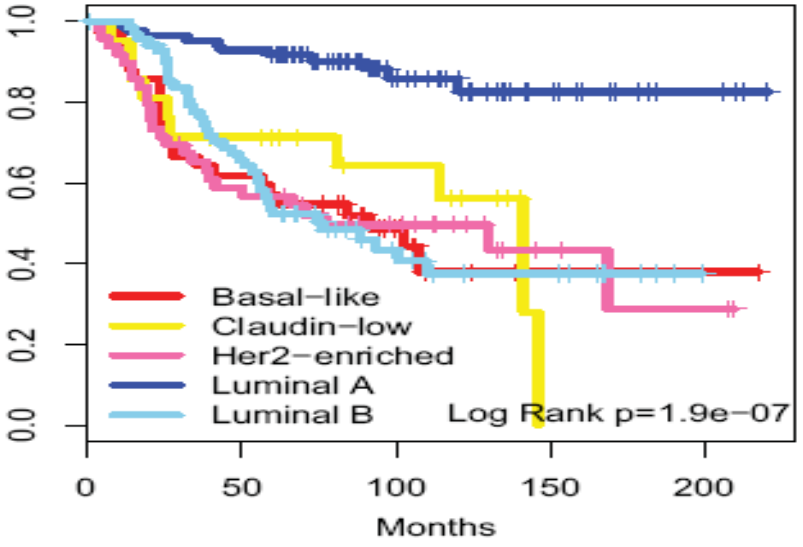
Accrual completed on Oct 25th 2010, 10,000 total. expected to report in 2018?



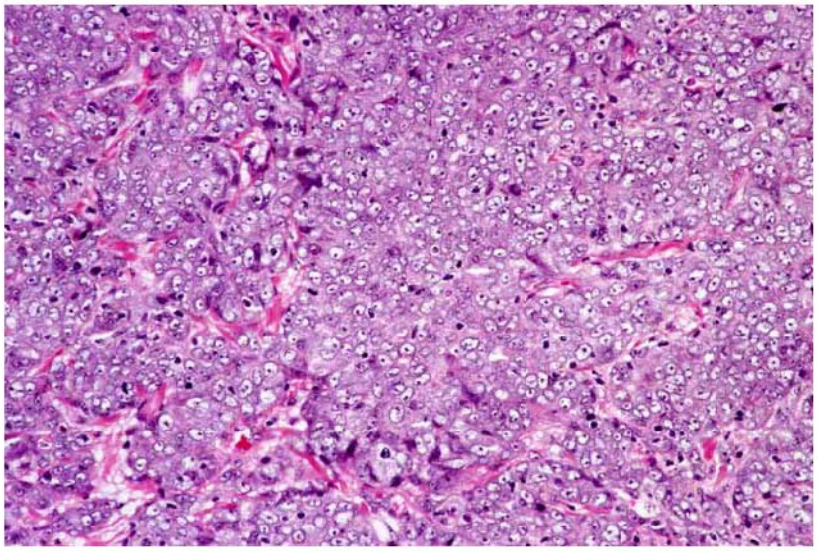
Intrinsic Subtypes of Breast Cancer



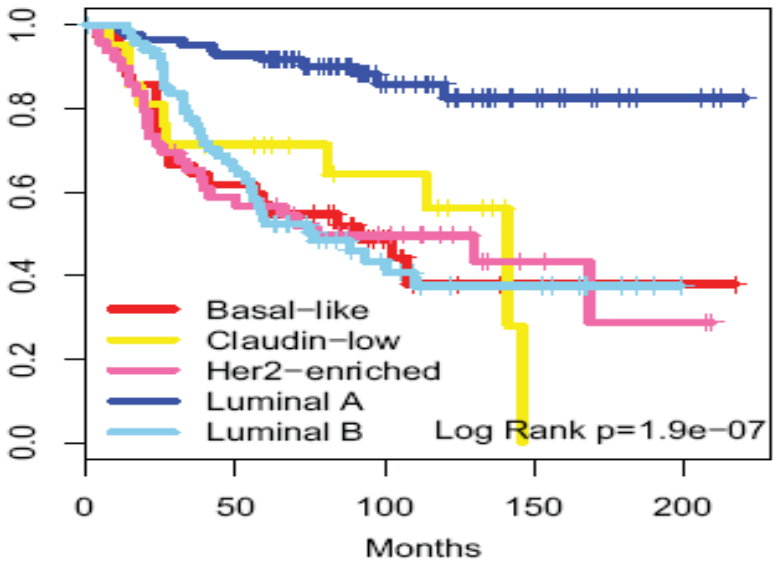


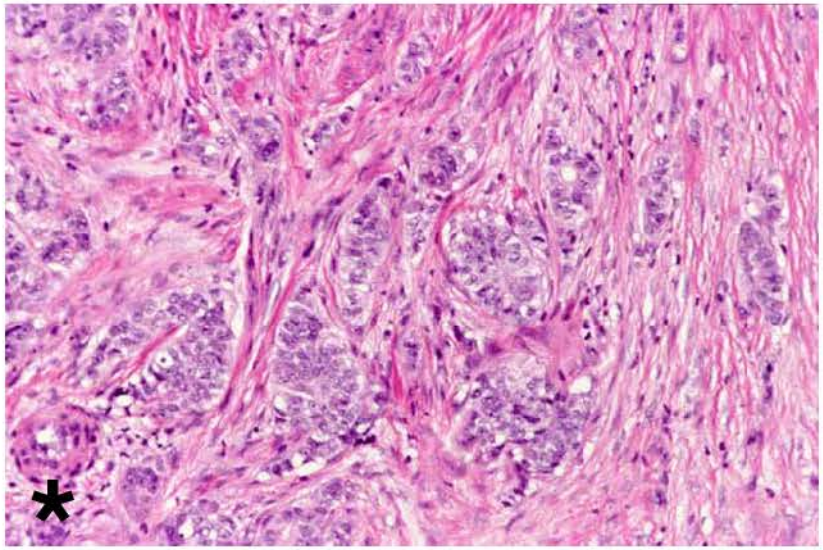


Breast Tumor, Grade III, Estrogen and Progesterone Receptor Positive, HER2-negative
T3 tumors (diameter >5cm), N1 (lymph node metastasis positive),
All received neoadjuvant chemotherapy and adjuvant tamoxifen

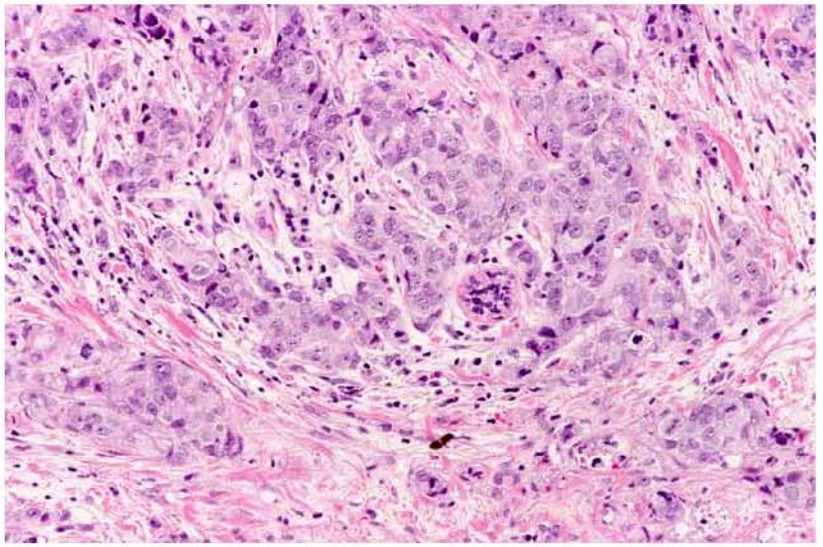


Died after 9 months: LumB





Died after 17 months: Basal-like



Alive for 96 months: LumA

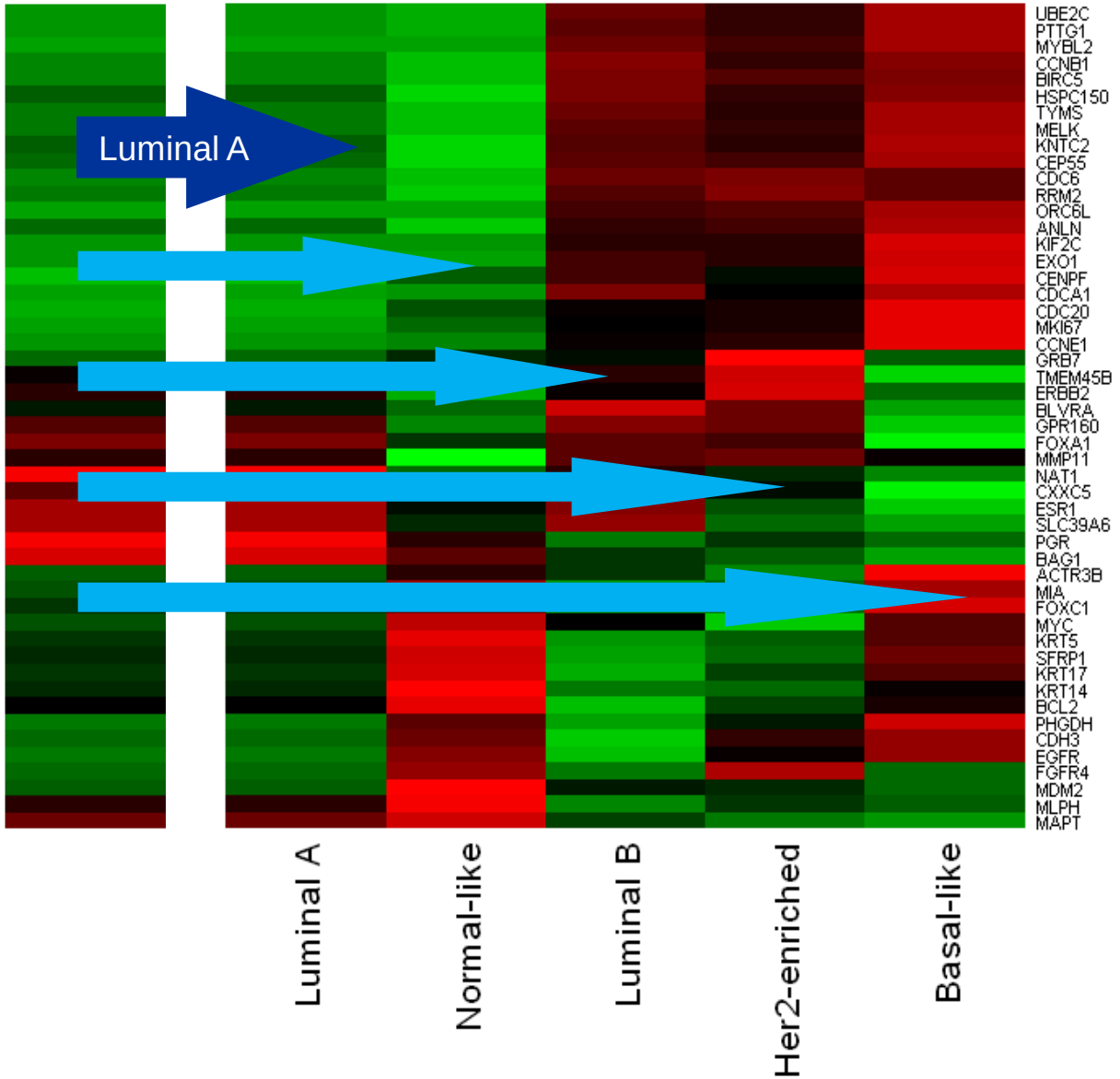
Supervised Risk Predictor of Breast Cancer Based on Intrinsic Subtypes

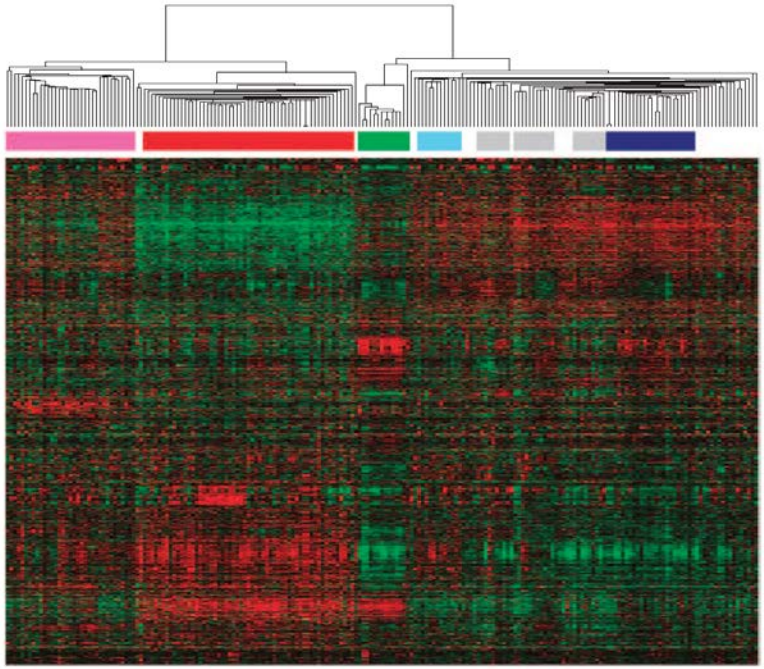
Parker et al., J Clinical Oncology; 27:1160-1167 2009 (PMID:19204204)

1. **The qRT-PCR assay consists of 50 genes and 5 centroids (provided at https://genome.unc.edu/)**

2. **The Cross Validation (CV) classification concordance of the 50 genes by qRT-PCR compared with 2000 genes by microarray was 93%**

3. **The assay works using RNA from FFPE materials or fresh frozen tissues**





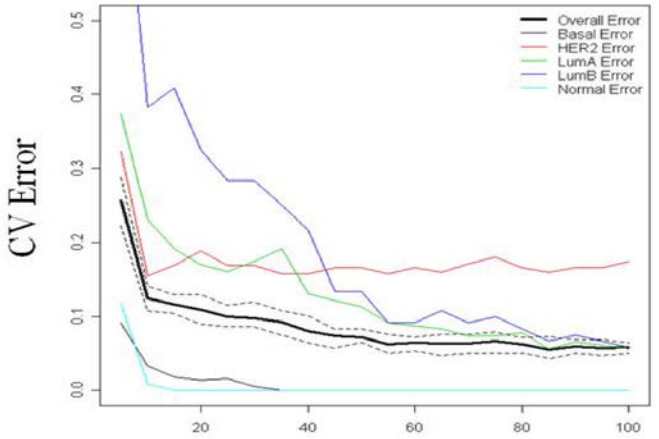
Step 1: cluster analysis using ~2000 gene intrinsic list, which was three previous intrinsic lists combined

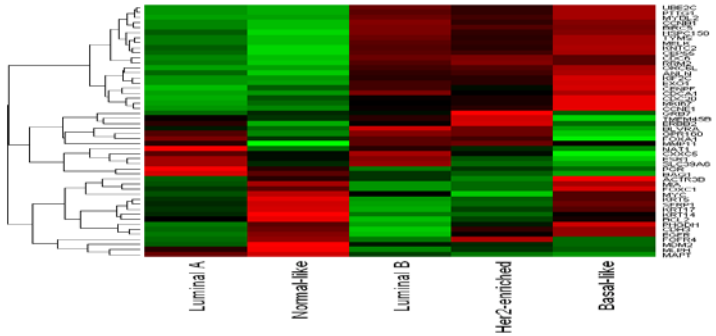
Statistical Significance of Clustering for High-Dimension, Low-Sample Size Data

Yufeng Liu, David Neil Hayes, Andrew Nobel and J. S. Marron
Journal of the American Statistical Association
Vol. 103, No. 483 (Sep., 2008), pp. 1281-1293

Published by: Taylor & Francis, Ltd. on behalf of the American Statistical Association
Stable URL: http://www.jstor.org/stable/27640161
Page Count: 13

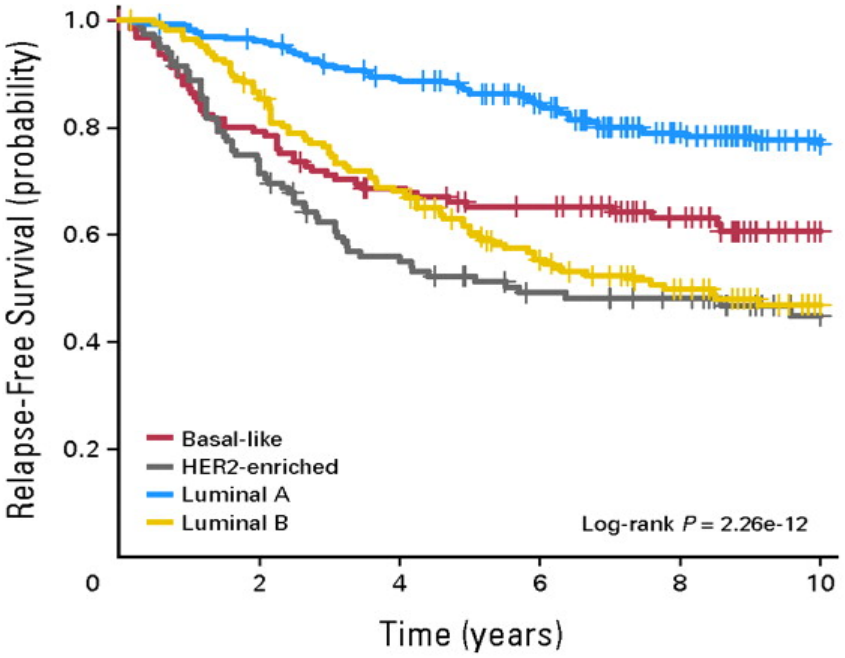
Step 2: use "SigClust" to objectively identify significant groups/subtypes, and the prototypical tumors of each group/subtype





Step 4: use Nearest Centroid Predictor with 5 subtypes and 50 genes (PAM50), by qRT-PCR with FFPE derived RNA

Step 3: use cross validation classification concordance testing to shrink the gene list from ~2000 genes to 50 genes, using "ClaNC" algorithm to balance gene selection (equal # of genes per class)



710 node negative breast cancer patients who received no systemic adjuvant therapy as a test set (pure prognosis)

Table 2. Models of Relapse Free Survival (untreated)

Model	A		B		C	
Variable	Hazard Ratio	p-value	Hazard Ratio	p-value	Hazard Ratio	p-value
Basal-like*	1.33	0.330	***1.79***	***0.030***	1.58	0.066
HER2-enriched*	***2.53***	***0.00012***	***3.25***	***<0.0001***	***2.90***	***<0.0001***
Luminal B*	***2.43***	***<0.0001***	***2.88***	***<0.0001***	***2.54***	***<0.0001***
ER Status~	0.83	0.38	0.83	0.34	0.83	0.32
Tumor Size†	***1.36***	***0.034***	***1.43***	***0.012***	***1.57***	***0.001***
Node Status‡	***1.75***	***0.035***	***1.72***	***0.041***	-	-
Histologic Grade^	***1.40***	***0.0042***	-	-	-	-
Full vs Subtype≈		***<0.0001***		***<0.0001***		***<0.0001***
Full vs Clinical¥		***<0.0001***		***<0.0001***		***<0.0001***

Parker et al. JCO 2009 (PMID:19204204)

Prognostic Risk Classification Strategy (ROR)

- **Similarity to the subtypes are used as variables in the prognostic model where the outcome is Risk of Recurrence (ROR):**

$$\text{(Model 1) ROR-S} = \beta_{1*}\text{Basal} + \beta_{2*}\text{HER2} + \beta_{3*}\text{LumA} + \beta_{4*}\text{LumB}$$

$$\text{(Model 2) ROR-T} = \beta_{1*}\text{Basal} + \beta_{2*}\text{HER2} + \beta_{3*}\text{LumA} + \beta_{4*}\text{LumB} + \beta_{5*}\text{Size}$$

$$\text{(Model 3) ROR-PT} = \beta_{1*}\text{Basal} + \beta_{2*}\text{HER2} + \beta_{3*}\text{LumA} + \beta_{4*}\text{LumB} + \beta_{5*}\text{Size} + \beta_{6*}\text{Proliferation}$$

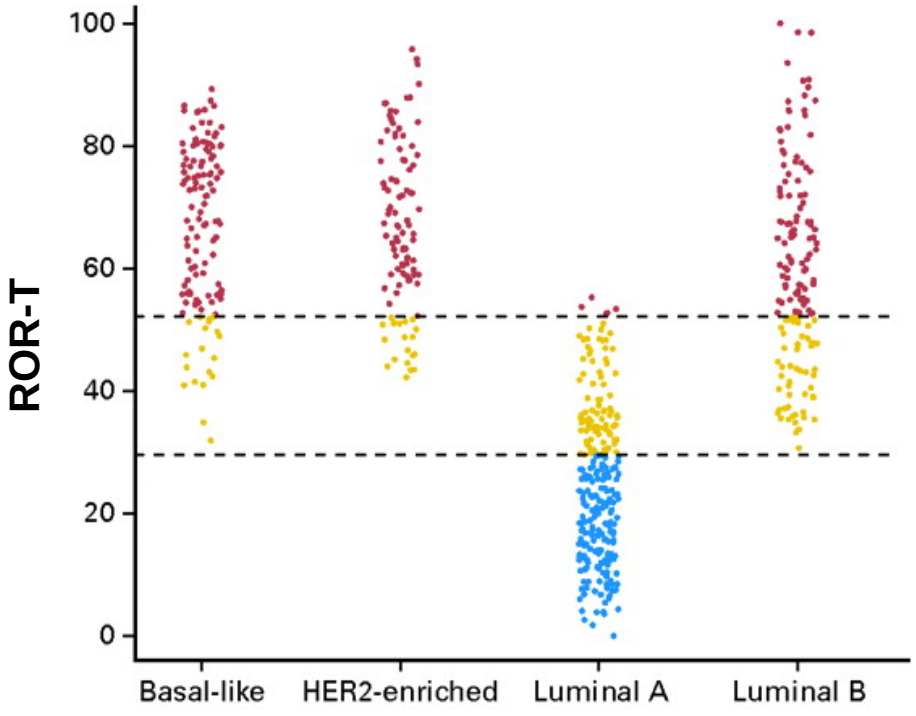
- **Weights (β) for each variable are learned from a training data set using a Cox proportional hazards model with Ridge Regression[1]**

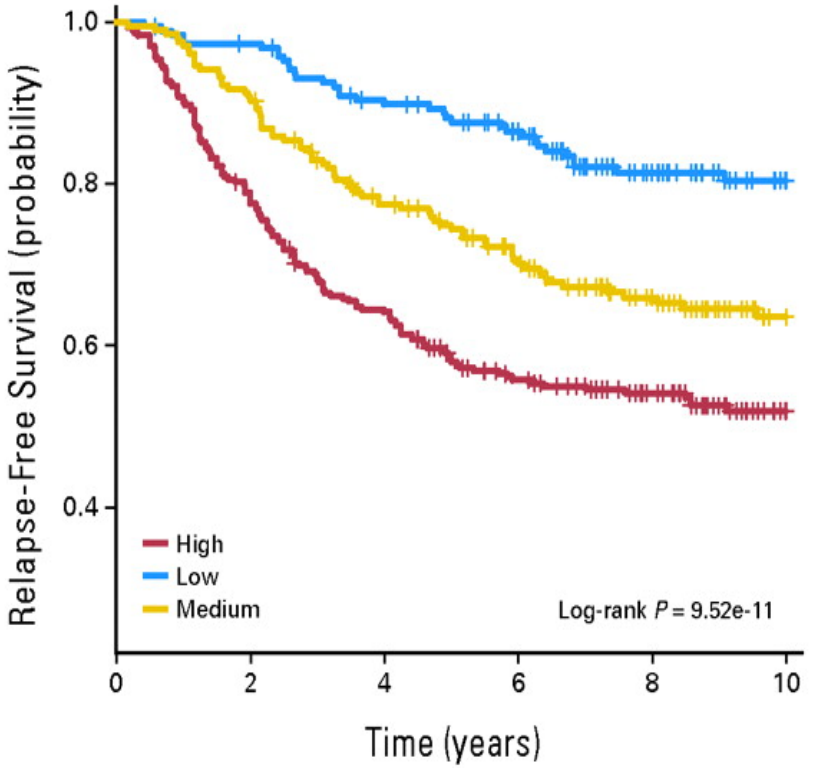
- **The weighted sum is assigned as the ROR score for a test case and a threshold may be applied for class assignment**

[1]**Ridge regression with Cox model: Tibshirani, Statistics in Medicine, 1997 (PMID:9044528) and Bovelstad et al. Bioinformatics 2007 (PMID:17553857)**

Parker et al. JCO 2009 (PMID:19204204)

Prognostic Risk Classification Strategy (ROR)



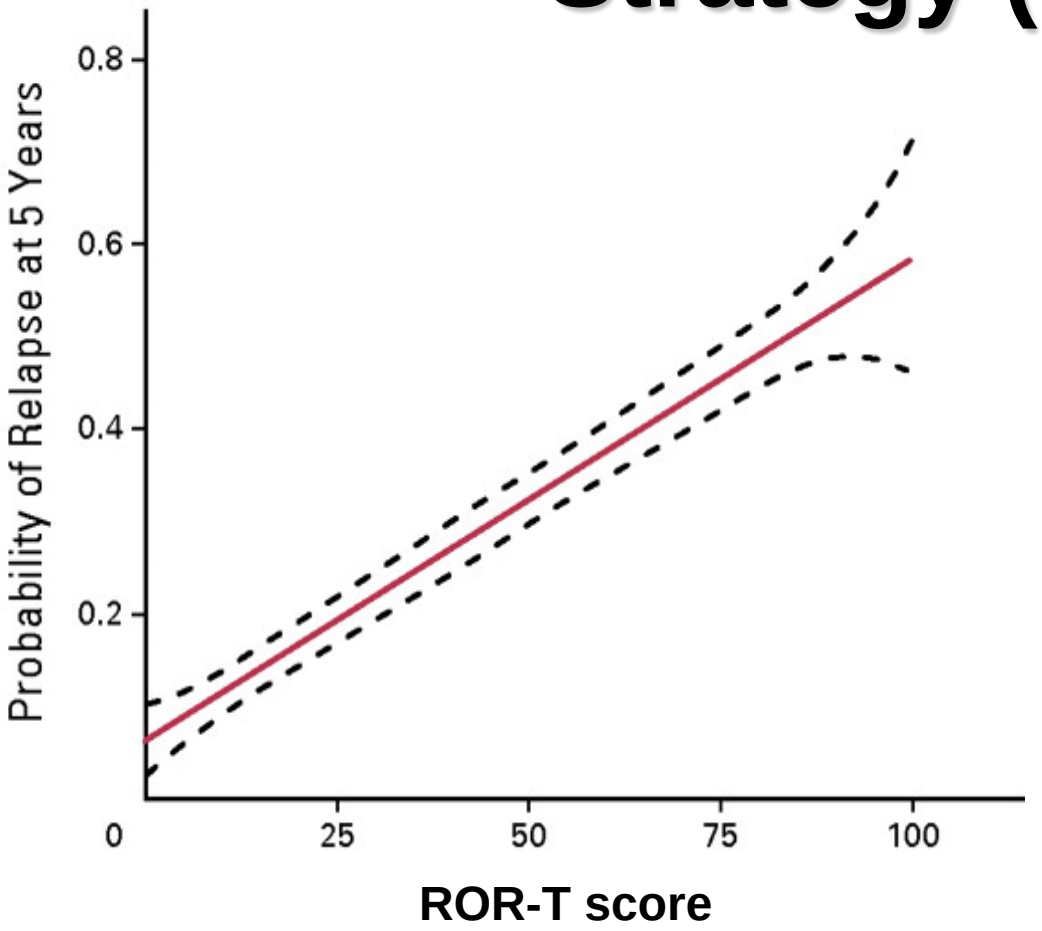


N=558 no adjuvant systemic therapy and node negative test cases

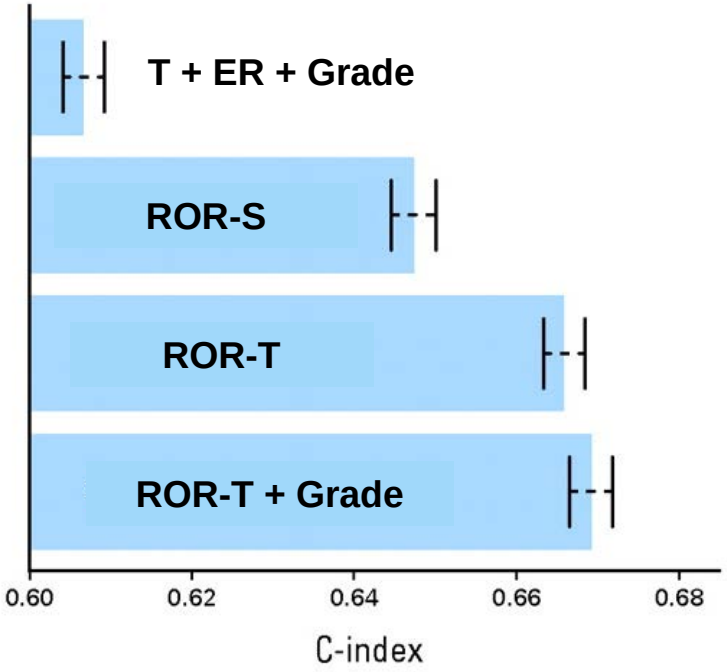
ROR-T thresholds determined from training cases

Parker et al. JCO 2009 (PMID:19204204)

Prognostic Risk Classification Strategy (ROR)



N=558 no adjuvant systemic therapy and node negative test cases



C-index: Harrell et al., JAMA, 1982 (PMID:7069920)

"The c-index is the proportion of all pairs of subjects whose survival time can be ordered such that the subject with the higher predicted survival is the one who survived longer" (taken from Harrell, Regression Modeling Strategies, Springer Series in Statistics).

Parker et al. JCO 2009 (PMID:19204204)

Current Gene Expression Technologies

1. DNA microarrays
   - Moderate precision covering ALL genes, fast, and inexpensive
     - long oligos (60-70mers/Agilent/Nimblegen) – 1-2 oligos/gene
     - short oligos (~25mers/Affymetrix) – 5-10 oligos/gene
     - cDNA arrays (100-1000bp PCR products/Stanford) – 1-2 cDNA clones/gene

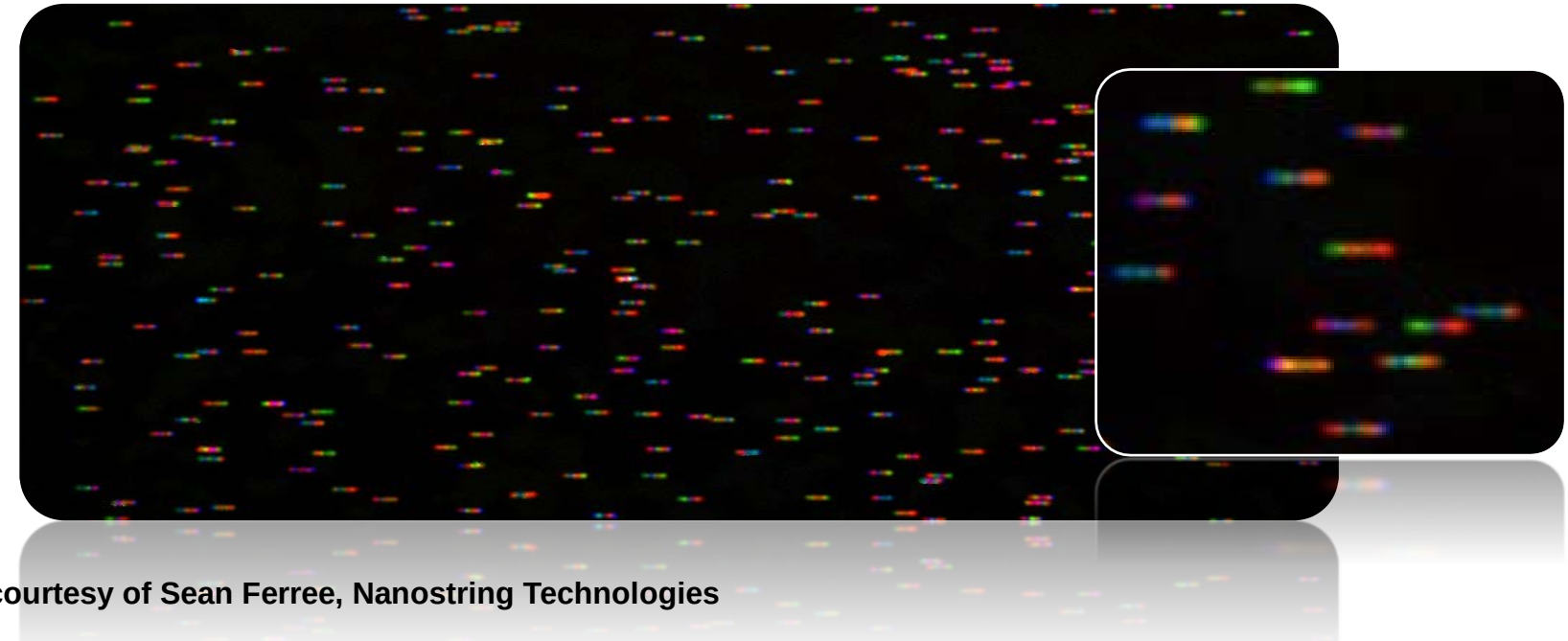
2. Quantitative RT-PCR
   - high precision, large dynamic range, fast, and inexpensive
   - smaller number of genes assayed (compatible with Formalin-fixed, Paraffin Embedded Materials (FFPE), which is the medical standard of care)

3. Nanostring nCounter
   - high precision, large dynamic range, fast, and moderate expense
   - single mRNA molecule counting capabilities (compatible with FFPE)

4. Illumina-based RNA-sequencing
   - Gene expression via Massively Parallel Sequencing (MPS)
   - high precision, large dynamic range, not fast, and expensive
   - compatible with FFPE

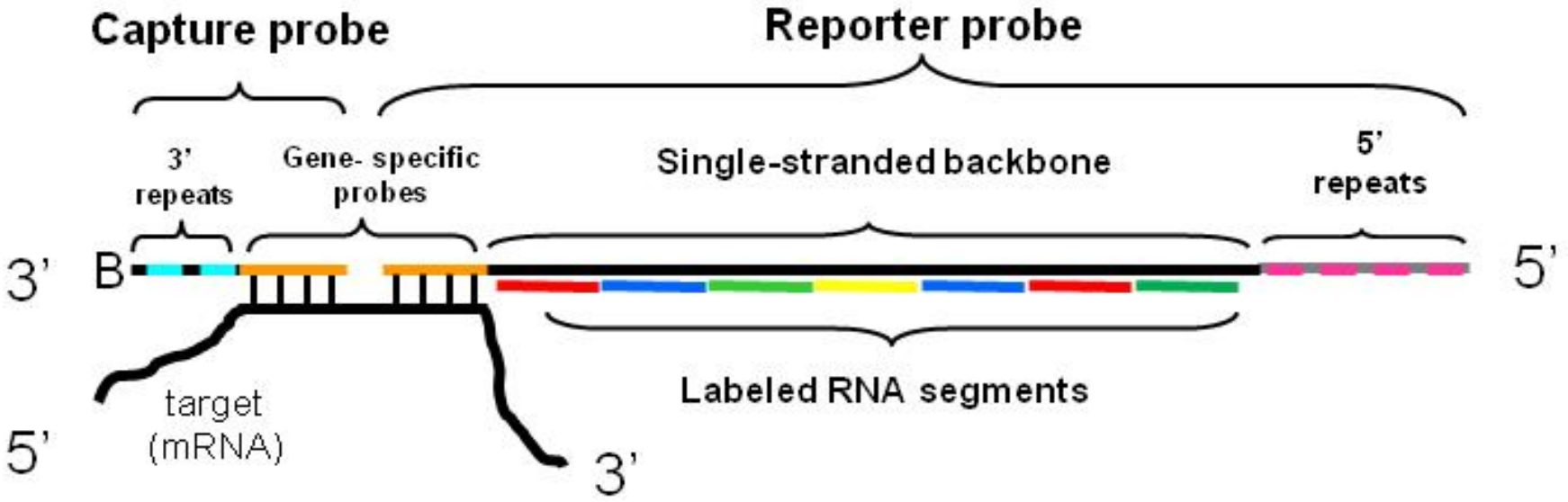
Scientific Reasons for PAM50 platform change to NanoString nCounter (~2010)

- **Easier to use than PCR or microarrays**
- **Fast and simple workflow**
- **Compatible with a distributed test and prefabricated kits (IVD)**
- **Compatible with RNA coming from FFPE materials**



slides courtesy of Sean Ferree, Nanostring Technologies

Schematic of gene detection probe system



1. Capture and Reporter probes are mixed with RNAs for hybridization
2. Gene specific oligos are covalently attached by ligation reaction
3. RNA segments are annealed to surface via 3' repeats and imaged
4. 1-800 genes can be assayed per sample

Nat Biotechnol. March 26(3):317-25, 2008 (PMID:18278033)