Welcome to

MATH:7450 (22M:305) Topological Data Analysis

Office hours:
MWF 15:45 - 16:20 GMT (10:45 - 11:20 CDT),
M 2:00 - 3:00 am GMT (9pm - 10pm CDT)
and by appointment.

Office hours will be held in our online classroom (same URL for entering class).

I am also available via google+, skype, and in person at the University of Iowa.

www.math.uiowa.edu/~idarcy/AT/schedule.html

Aug 30	Download Mapper for Matlab Python Mapper Graphviz Web tool: Progression Analysis of Disease - PAD (includes Mapper) Ayasdi Iris, academic trial
	Additional readings: Topological Methods for the Analysis of High Dimensional Data Sets and 3D Object Recognition G. Singh , F. Memoli, G. Carlsson (2007) Topology and data, G Carlsson (2009) DNA MICROARRAY VIRTUAL LAB, youtube video Pearson Product-Moment Correlation

Application 1: breast cancer gene expression

Data: microarray gene expression data from 2 data sets, NKI and GSE2034

Distance: pearson correlation distance

Filters: (1) L-infinity centrality:

$$f(x) = \max\{d(x, p) : p \text{ in data set}\}$$

captures the structure of the points far removed from the center or norm.

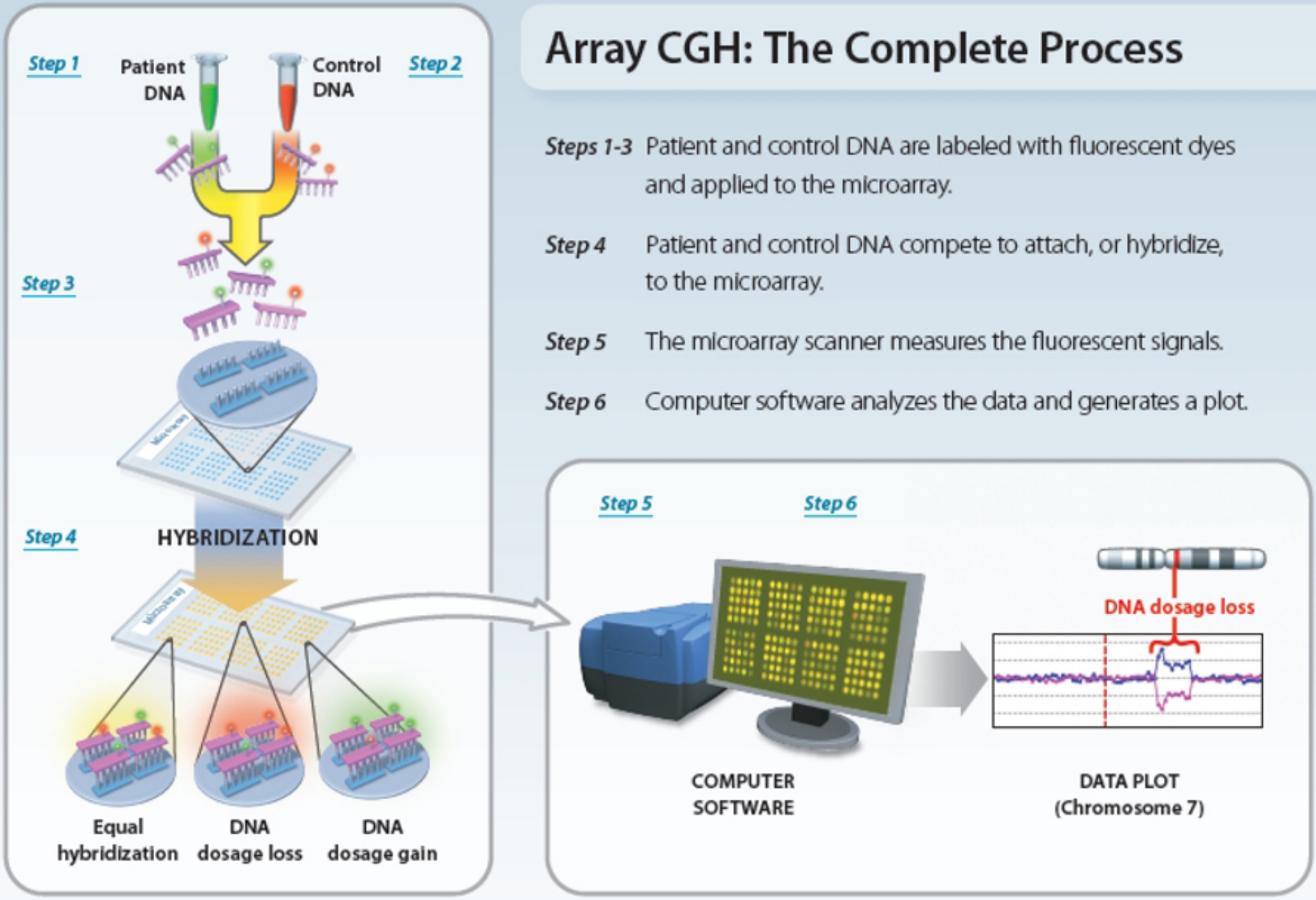
(2) NKI: survival vs. death

GSE2034: no relapse vs. relapse

Clustering: Single linkage.

http://www.nature.com/srep/2013/130207/srep01236/full/srep01236.html

Array CGH: The Complete Process



Steps 1-3 Patient and control DNA are labeled with fluorescent dyes and applied to the microarray.

Step 4 Patient and control DNA compete to attach, or hybridize, to the microarray.

Step 5 The microarray scanner measures the fluorescent signals.

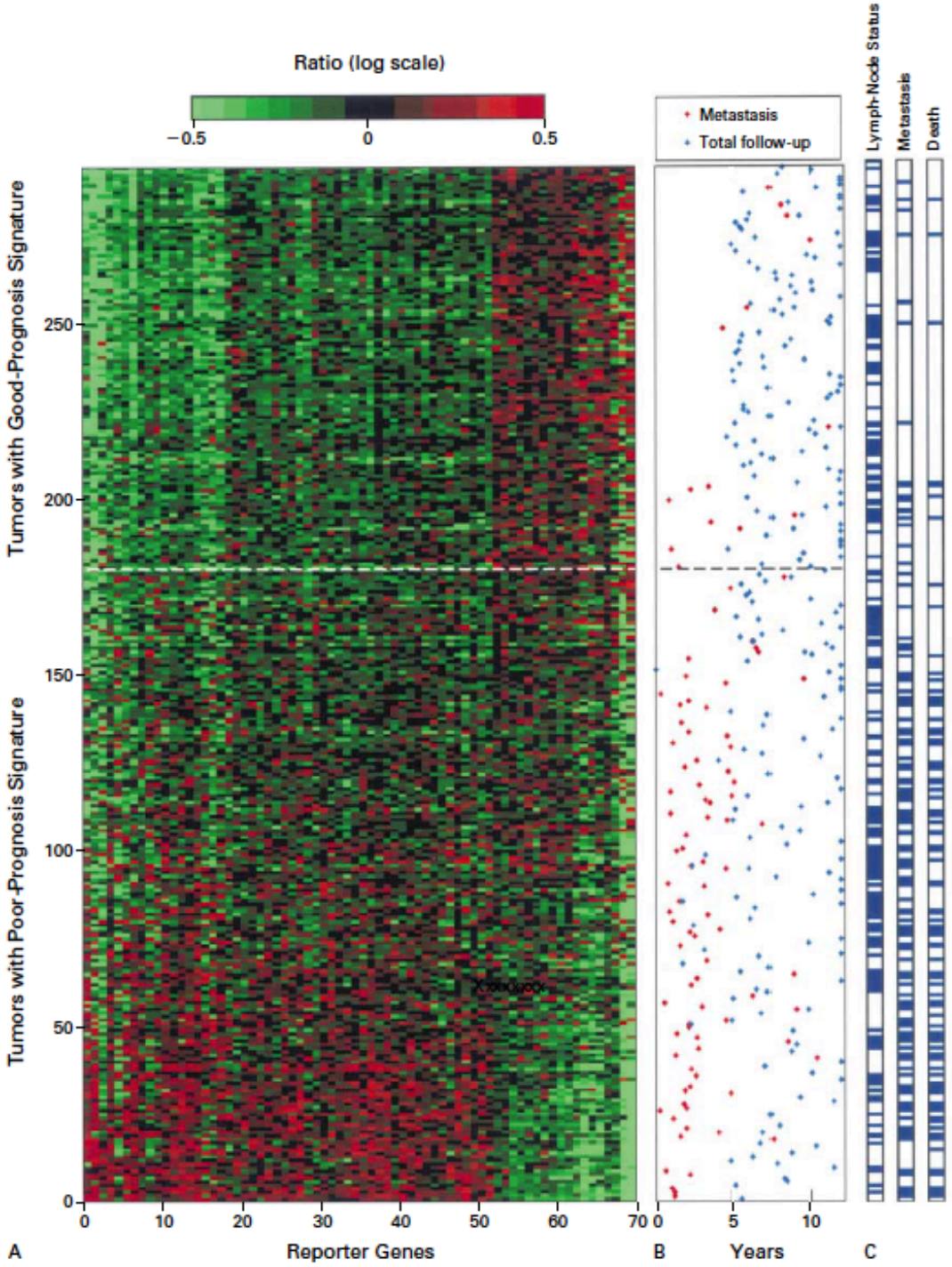
Step 6 Computer software analyzes the data and generates a plot.

www.nature.com/scitable/topicpage/microarray-based-comparative-genomic-hybridization-acgh-45432

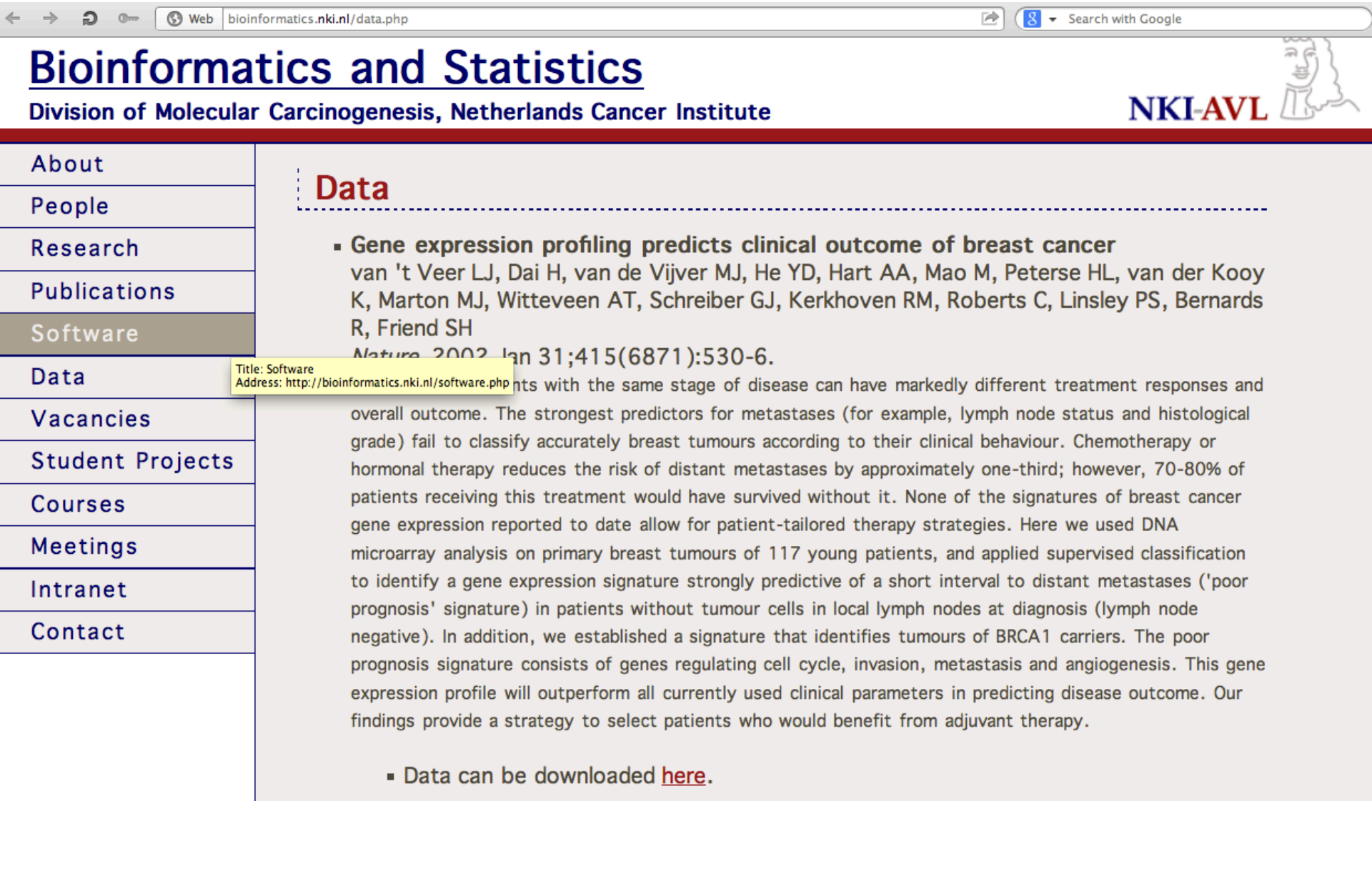
Gene expression profiling predicts clinical outcome of breast cancer

van 't Veer LJ, Dai H, van de Vijver MJ, He YD, Hart AA, Mao M, Peterse HL, van der Kooy K, Marton MJ, Witteveen AT, Schreiber GJ, Kerkhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH

Nature. 2002 Jan 31;415(6871):530-6.



http://bioinformatics.nki.nl/data.php



2 breast cancer data sets:

1.) NKI (2002):
gene expression levels of 24,000 from 272 tumors. Includes node-negative and node-positive patients, who had or had not received adjuvant systemic therapy. Also includes survival information.

2.) GSE203414 (2005)
expression of 22,000 transcripts from total RNA of frozen tumour samples from 286 lymph-node-negative patients who had not received adjuvant systemic treatment. Also includes time to relapse information.

http://www.nature.com/srep/2013/130207/srep01236/full/srep01236.html

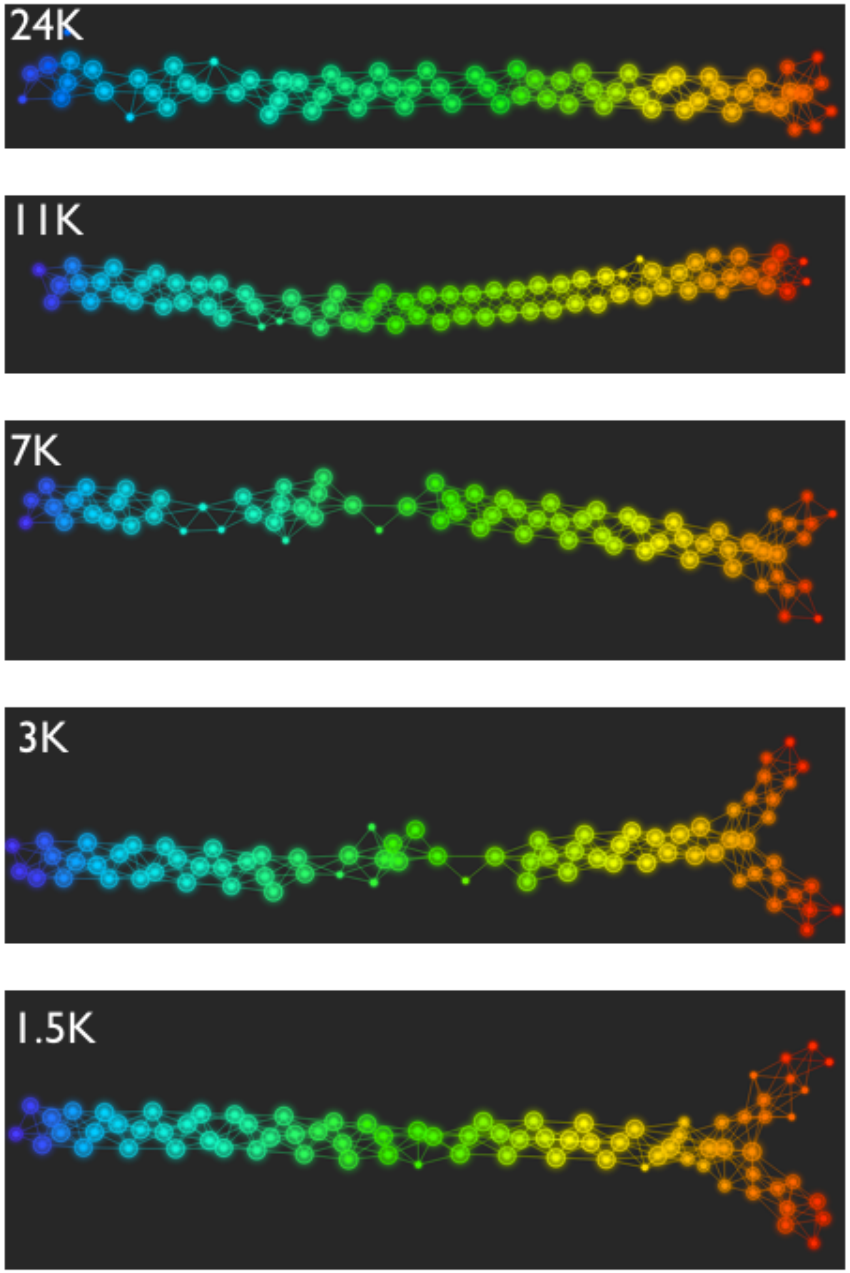


Fig. S1. Shape of the data becomes more distinct as the analysis columns are restricted to the top varying genes.

24K: all the genes on the microarray were used in the analysis;

11K: 10,731 top most varying genes were used in the analysis;

7K: 6.688 top most varying genes were used in the analysis;

3K: 3212 top most varying genes were used in the analysis;

1.5K: 1553 top most varying genes were used in the analysis.

Graphs colored by the L-infinity centrality values. Red: high; Blue: low

http://www.nature.com/srep/2013/130207/srep01236/full/srep01236.html

Comparison of our results with those of Van de Vijver and colleagues is difficult because of differences in patients, techniques, and materials used.

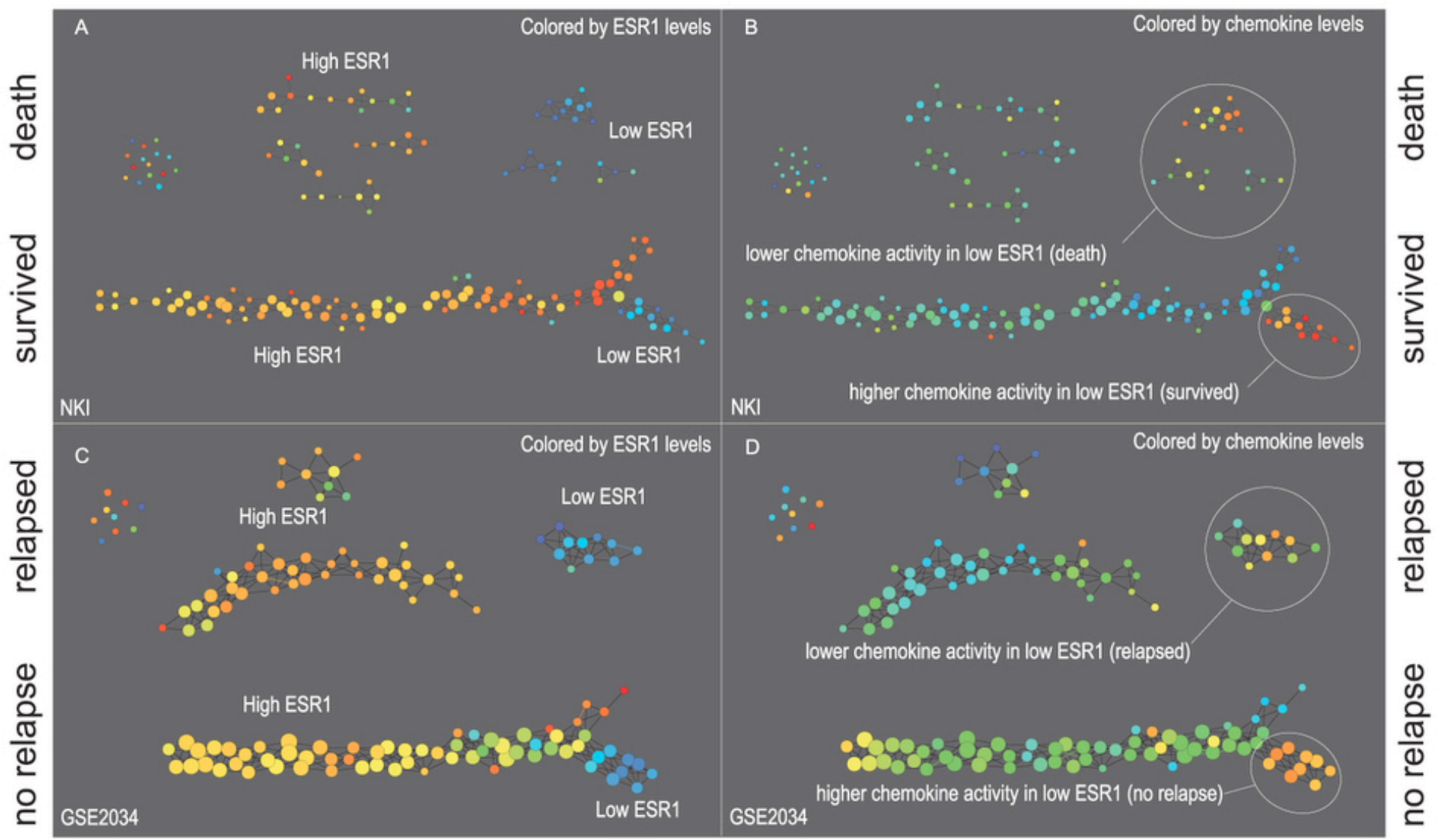
Their study included node-negative and node-positive patients, who had or had not received adjuvant systemic therapy, and only women younger than 53 years.

microarray platforms used in the studies differ—Affymetrix and Agilent.

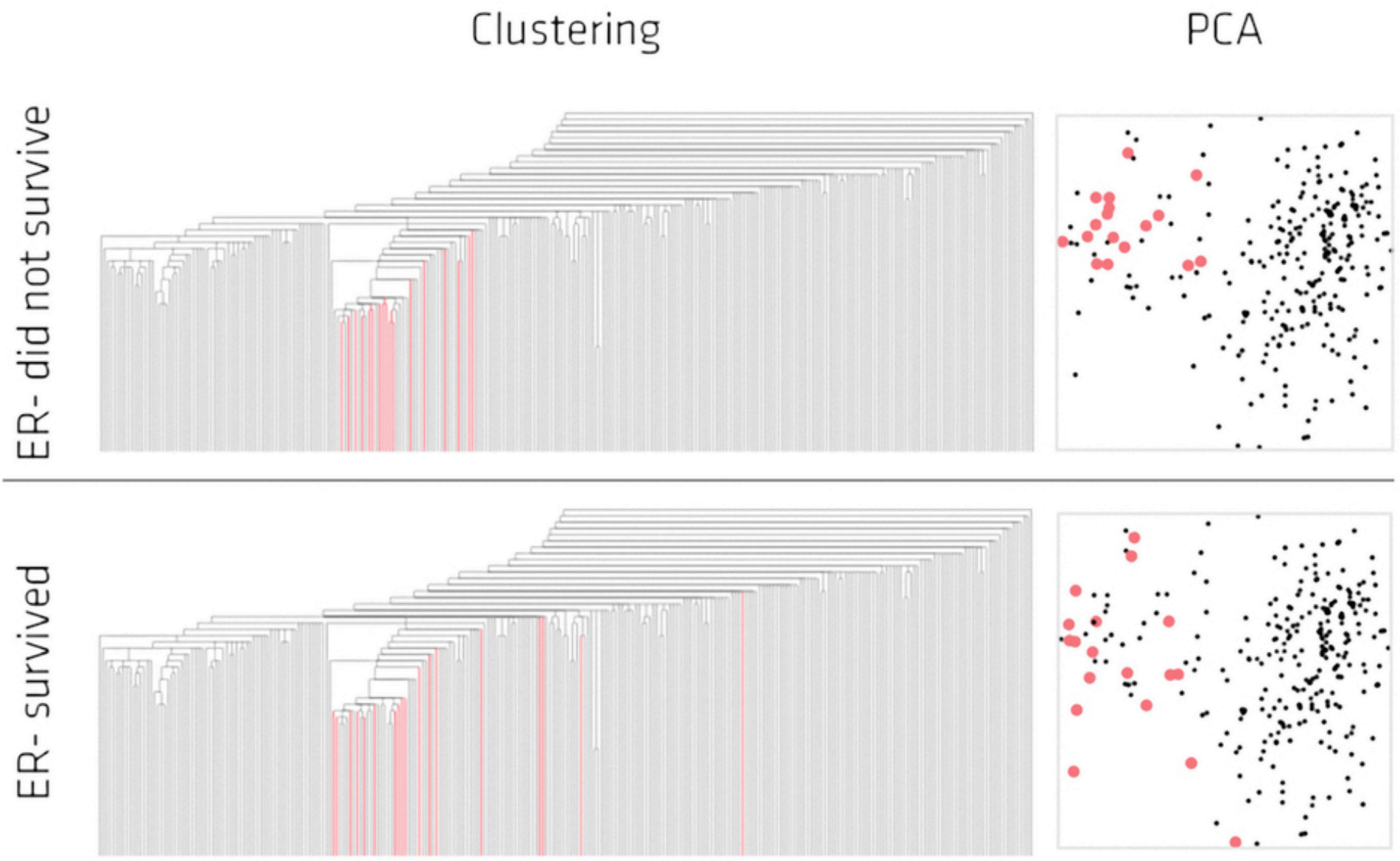
Of the 70 genes in the study by van't Veer and co-workers, 48 are present on the Affymetrix U133a array, whereas only 38 of our 76 genes are present on the Agilent array. There is a three-gene overlap between the two signatures (cyclin E2, origin recognition complex, and TNF superfamily protein).

Despite the apparent difference, both signatures included genes that identified several common pathways that might be involved in tumour recurrence. This finding supports the idea that although there might be redundancy in gene members, effective signatures could be required to include representation of specific pathways.

From: Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer, Yixin Wang et al, The Lancet, Volume 365, Issue 9460, 19–25 February 2005, Pages 671–679, http://www.sciencedirect.com/science/article/pii/S0140673605179471

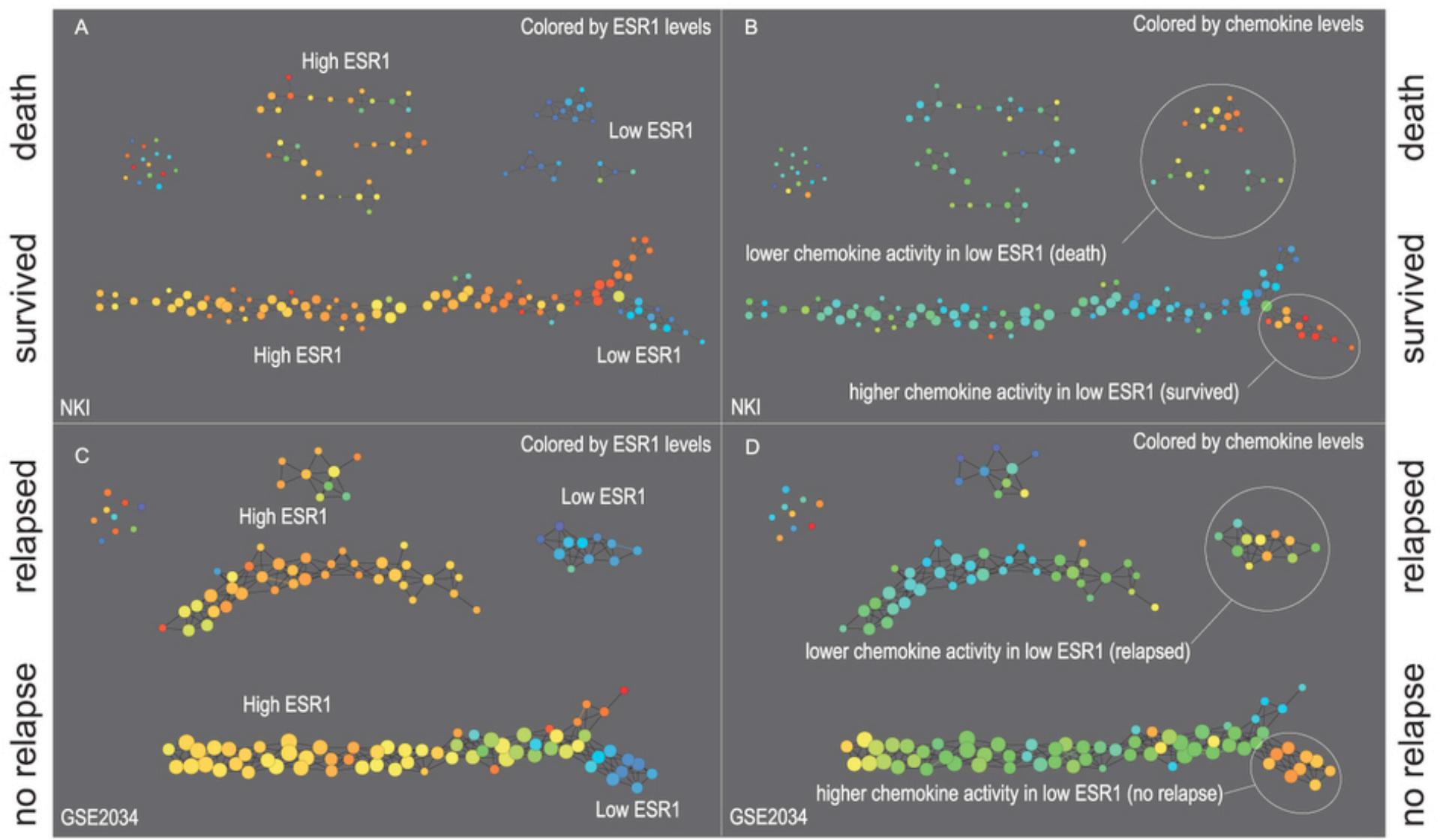


Two filter functions, L-Infinity centrality and survival or relapse were used to generate the networks. The top half of panels A and B are the networks of patients who didn't survive, the bottom half are the patients who survived. Panels C and D are similar to panels A and B except that one of the filters is relapse instead of survival. Panels A and C are colored by the average expression of the ESR1 gene. Panels B and D are colored by the average expression of the genes in the KEGG chemokine pathway. Metric: Correlation; Lens: L-Infinity Centrality (Resolution 70, Gain 3.0x, Equalized) and Event Death (Resolution 30, Gain 3.0x). Color bar: red: high values, blue: low values.



Highlighted in red are the lowERNS (top panel) and the lowERHS (bottom panel) patient sub-groups.

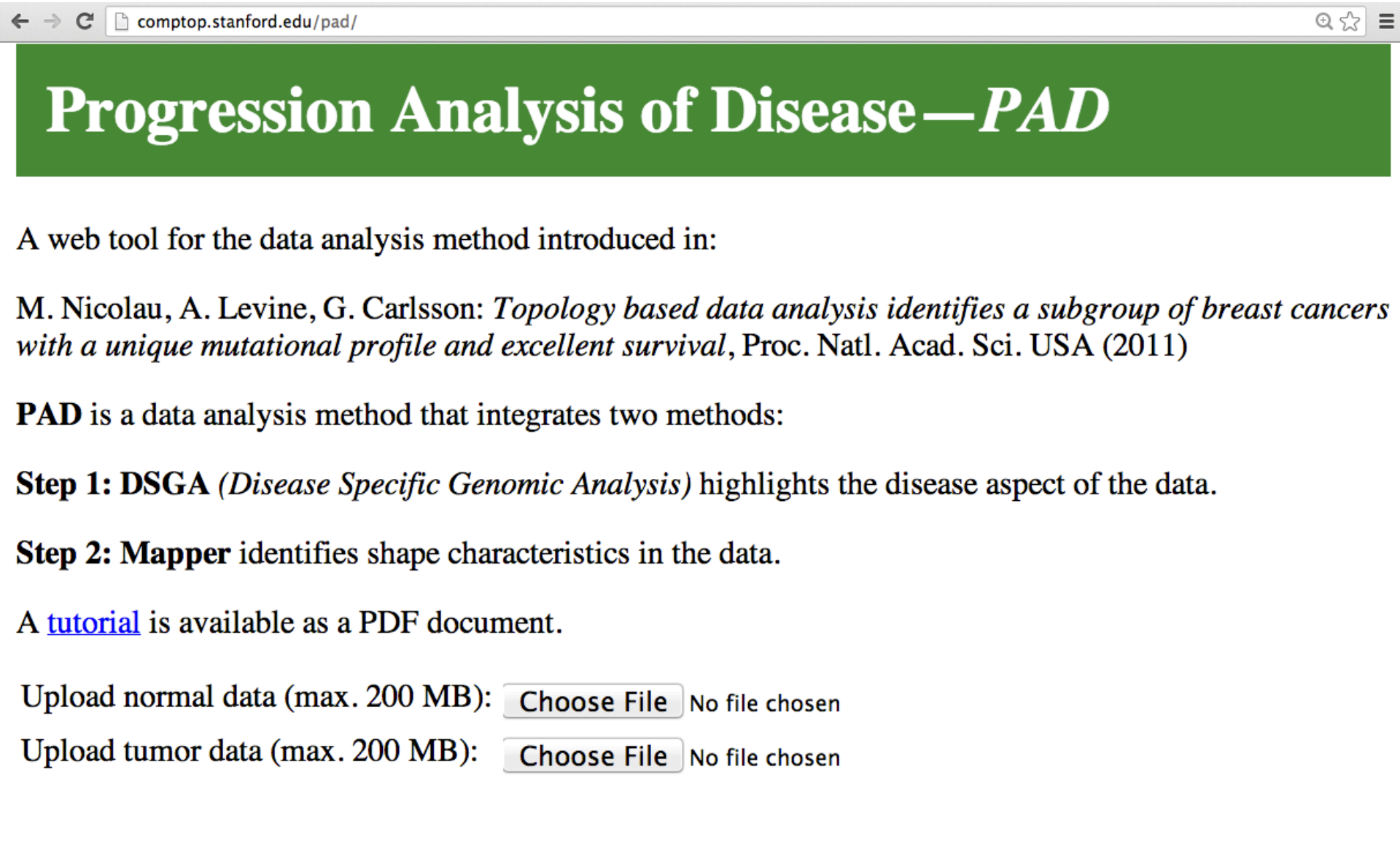
http://www.nature.com/srep/2013/130207/srep01236/full/srep01236.html



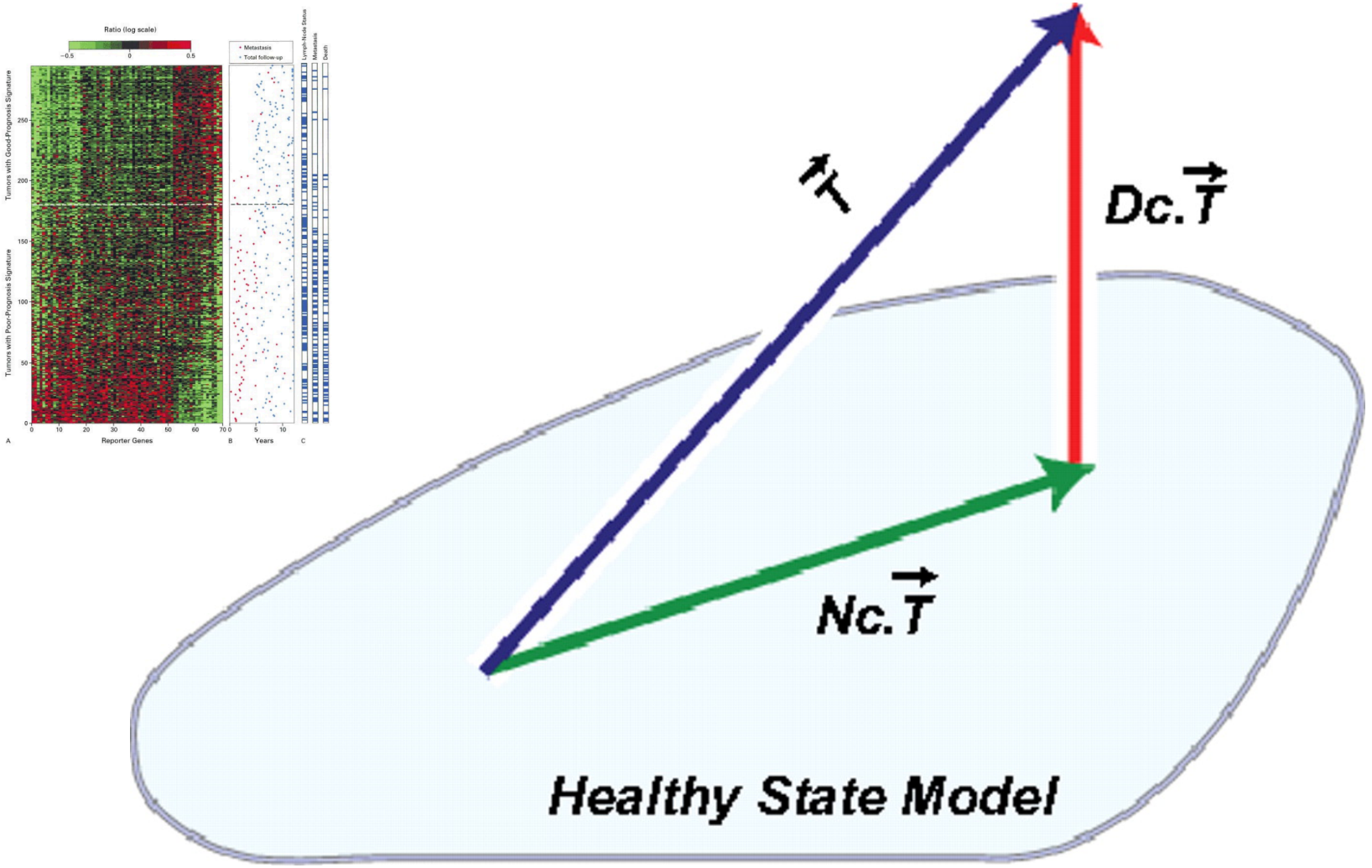
Two filter functions, L-Infinity centrality and survival or relapse were used to generate the networks. The top half of panels A and B are the networks of patients who didn't survive, the bottom half are the patients who survived. Panels C and D are similar to panels A and B except that one of the filters is relapse instead of survival. Panels A and C are colored by the average expression of the ESR1 gene. Panels B and D are colored by the average expression of the genes in the KEGG chemokine pathway. Metric: Correlation; Lens: L-Infinity Centrality (Resolution 70, Gain 3.0x, Equalized) and Event Death (Resolution 30, Gain 3.0x). Color bar: red: high values, blue: low values.

Mapper Software

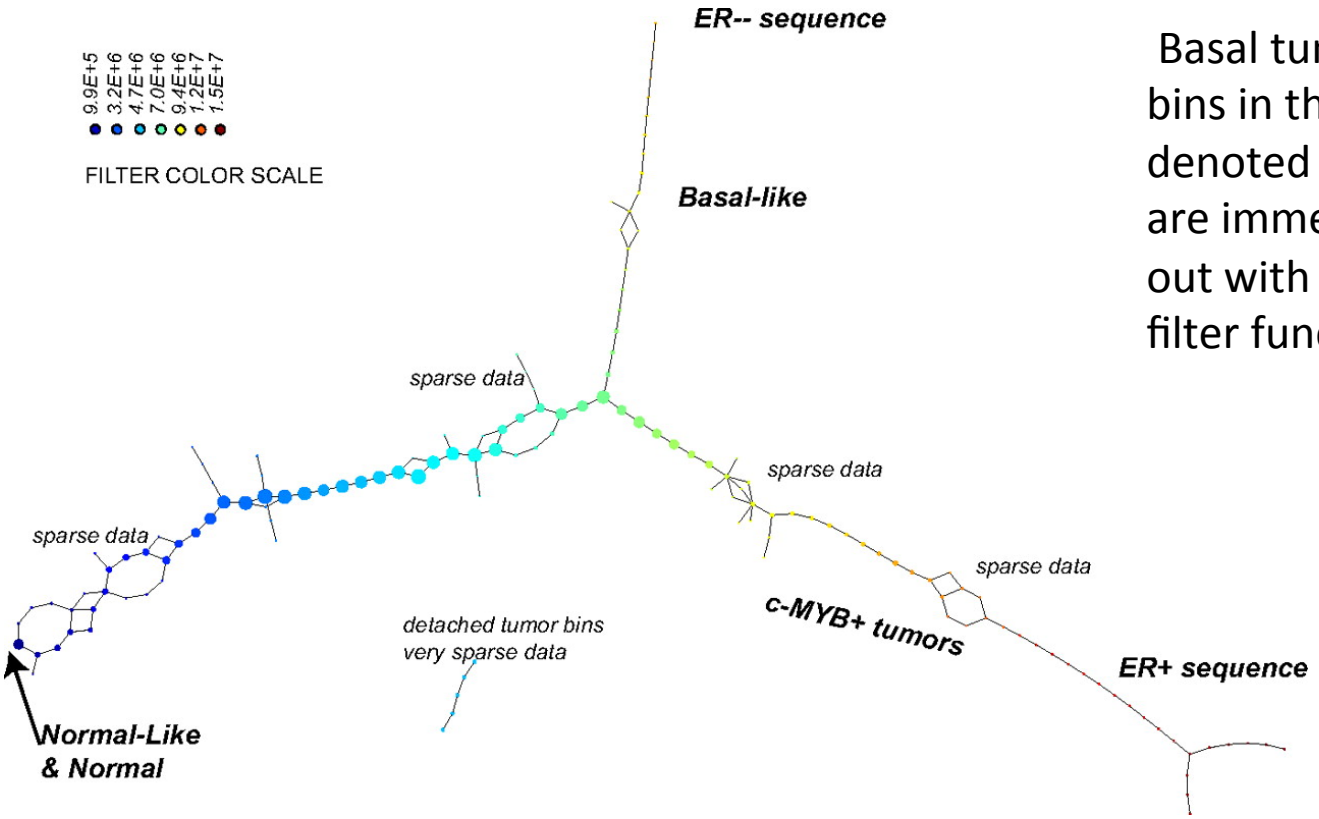
http://comptop.stanford.edu/pad/



DSGA decomposition of the original tumor vector into the Normal component its linear models fit onto the Healthy State Model and the Disease component vector of residuals.



Nicolau M et al. PNAS 2011;108:7265-7270



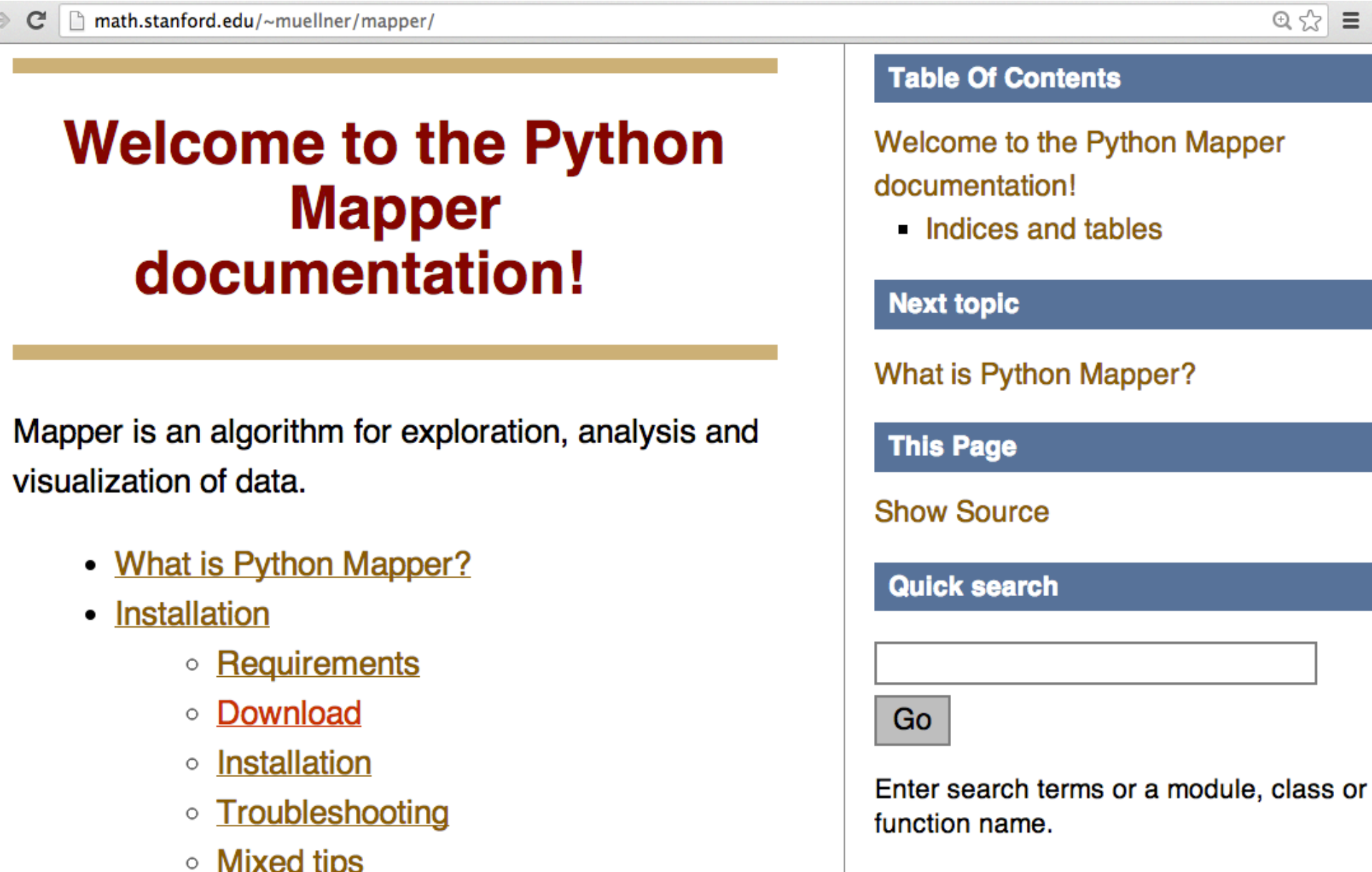
Basal tumors occupy most of the bins in the tumor sequence denoted as ER− sequence. They are immediately visible and stand out with large value (red) in the filter function

Normal tissue samples all fall in the same bin together with 15 additional ER+ tumors.

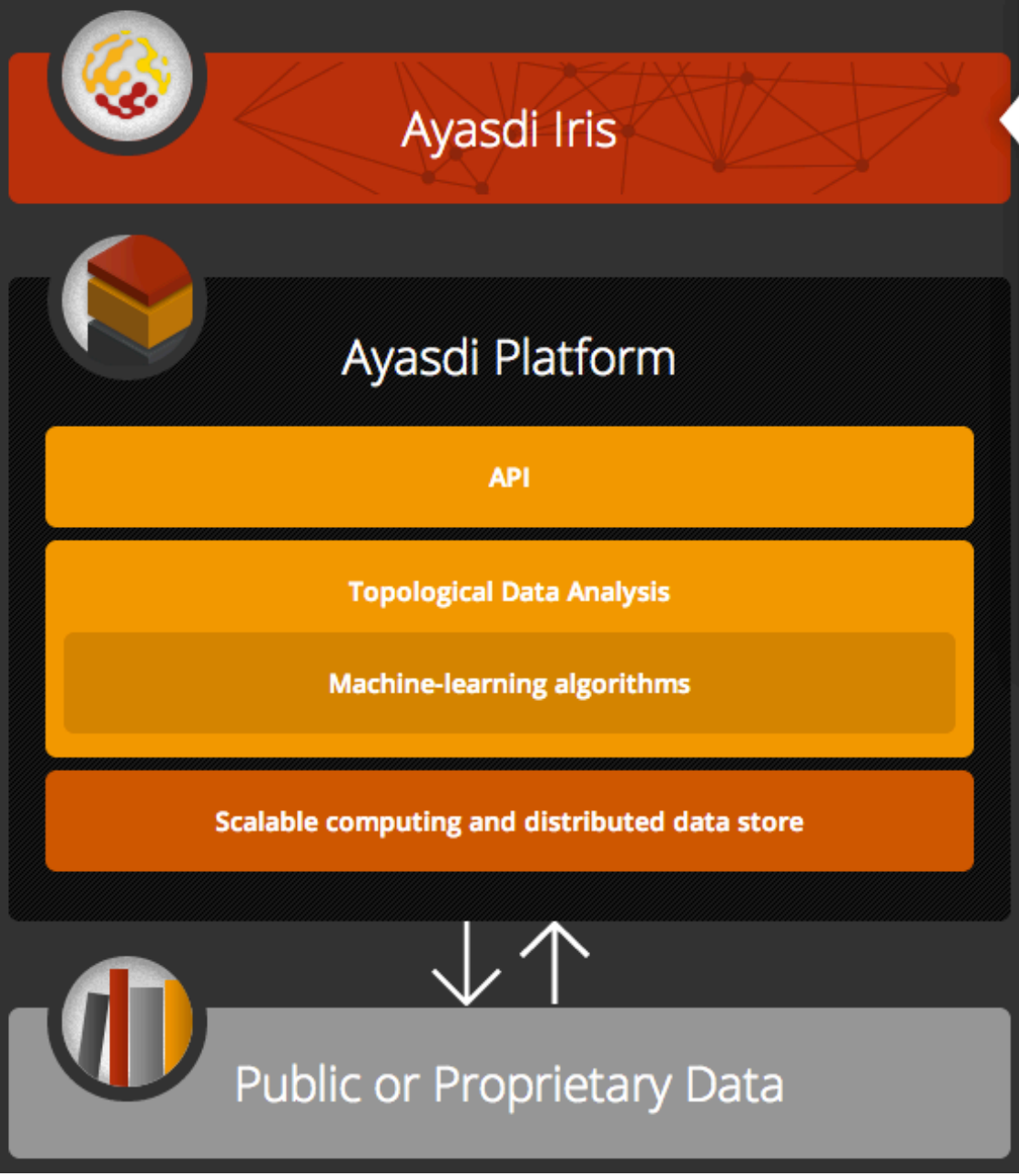
The known group of her2+ tumors is not yet visible, owing to the well-understood problem that only a small number of genes (on 17q) identify it, making them mathematically less visible, despite the fact that the small number of coordinates (17q genes) are biologically important.

A long tumor sequence on the graph, the ER+ sequence showing large deviation from normal, is visible, as defined by the filter. This tumor sequence also consists of ER+ tumors, but unlike the first (blue) group of tumors, these are distinct from normal tissue in that the value of the filter function—the Lp magnitudes of the tumor vectors in these bins—is very large.

http://math.stanford.edu/~muellner/mapper/



www.ayasdi.com/product/



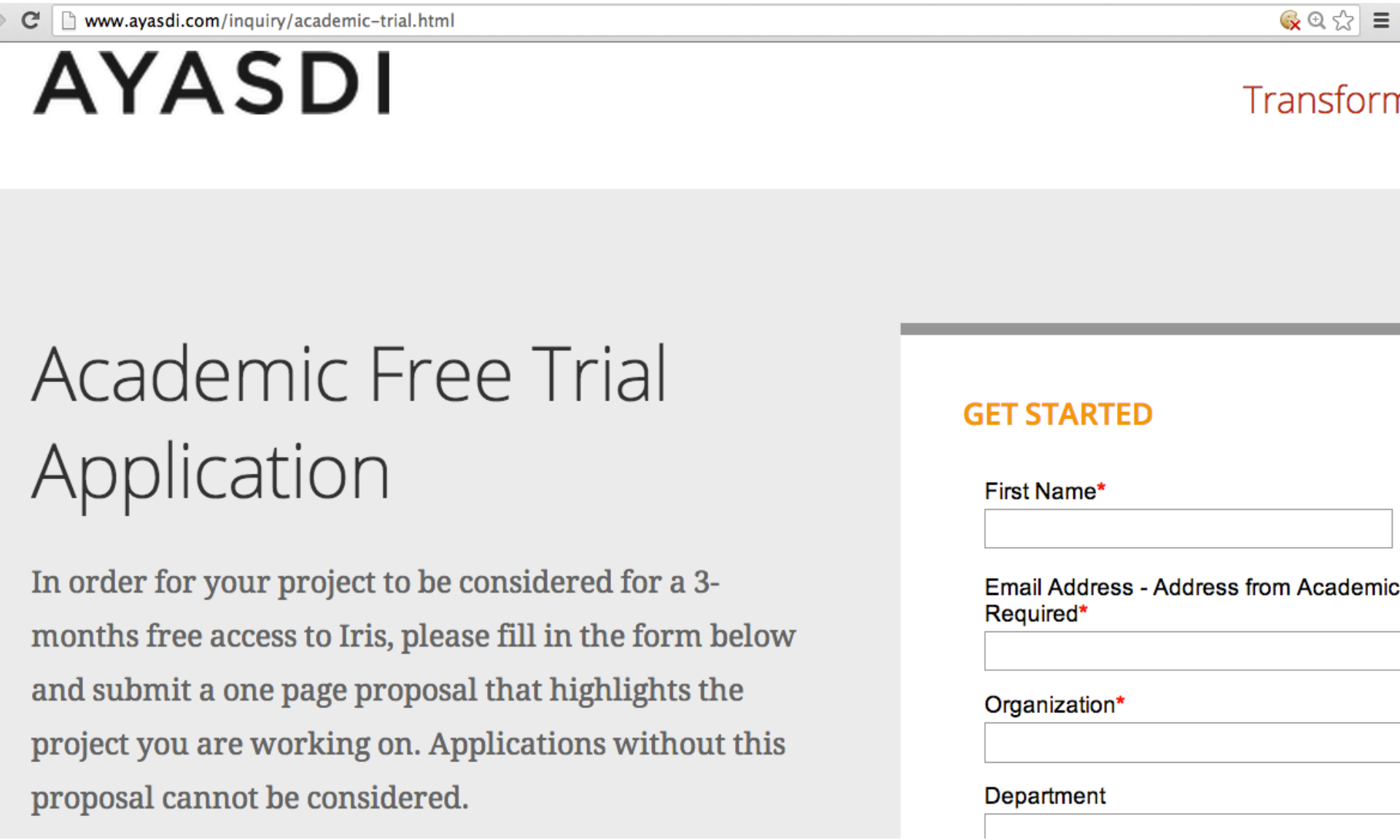
Ayasdi Iris

Ayasdi Iris is the world's first Insight Discovery solution that uses TDA to highlight the underlying geometric shapes of your data and allowing for real-time interaction to produce immediate insights.

Ayasdi Iris is offered as a multi-tenant cloud or as an on-premise solution, and is applicable for any organization needing to discover insights from complex data.

www.ayasdi.com/product/

http://www.ayasdi.com/inquiry/academic-trial.html



Matlab version demonstration