

#3007 DISCOVERY OF MULTIPLEX GENOMIC MARKERS FOR PREDICTING BREAST CANCER RECURRENCE

Cole C Harris, Exagen Diagnostics

Introduction

While primary tumor stage, tumor size, and lymph node status are predictive of distant recurrence, patients presenting tumors with similar pathological features may have widely varying outcomes. In particular these factors are of limited utility in distinguishing early stage breast cancer patients with good and poor prognosis. As a result, many more patients are subjected to adjuvant chemotherapy than stand to benefit from such treatment. With the larger goal of developing a clinically useful and practical prognostic test, we considered potential test implementation technologies as well as the availability of data for prognostic marker discovery. Fluorescence in situ hybridization (FISH) based assays are in wide use in cancer diagnostics, thus barriers to the use of a new prognostic FISH-based test would be minimal. However there is little publicly available high-resolution genomic data suitable for the discovery of prognostic FISH markers. To overcome this, we developed an integrated approach to DNA marker discovery that uses a concurrent global search for predictive patterns in both gene expression and array comparative genomic hybridization (CGH) data.

Data

Van't Veer et al (2002) addressed the question of identifying a gene expression profile correlating with prognosis in early stage, node-negative breast cancer. From genome-wide microarray data comprising 24,481 expression measurements, they identified a group of 70 genes useful in predicting prognosis. Comparison of 19 independent test sample expressions to the mean poor and good prognosis training sample profiles was shown to be a relatively accurate method of predicting prognosis. The 70-gene marker has been subsequently validated in larger studies (van de Vijver 2002), substantiating the prognostic information content of the original dataset.

Pollack (2002) demonstrated that, for tumors of the breast, for a significant proportion of the genome, gene expression is correlated with gene amplification. From DNA copy number and RNA gene expression measurements for 6095 genes across 37 tumor samples, they found that 12% of all gene expression variation is directly attributable to gene amplification. More so, for highly amplified regions, 62% of the genes located in these regions exhibit moderately or highly elevated expression.

Methods

We employed two distinct analyses:

Approach 1

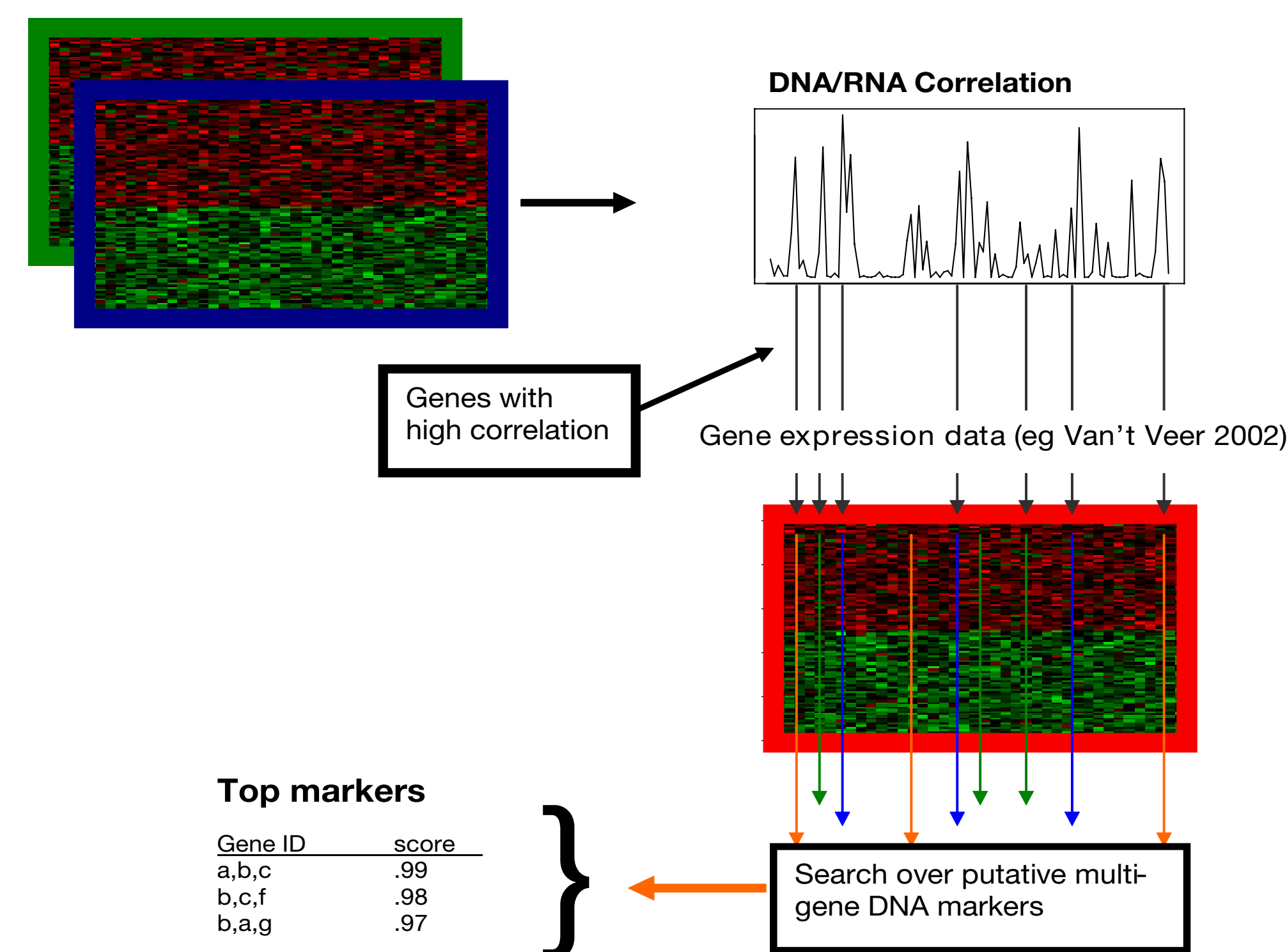
The Pollack study data provided a crucial link between gene expression and copy number variation. This link was used to compute a high-resolution genome-wide model of the contribution of copy number variation to gene expression. In a global search for multiplex gene expression patterns predictive of recurrence, this model was used to assess indirectly the likelihood of a corresponding gene-scale genomic aberration pattern (marker) in the array CGH data. A simplified version of this process is depicted in Figure 1.

In this search process, both the performance of the marker in predicting recurrence from gene expression data, and the expected variability of that performance on the corresponding DNA copy number data were used to rank markers. Permutation testing was employed to compute the statistical significance of these markers. Genes appearing frequently in the top-ranked significant markers were deemed prognostic.

Approach 2

From a previous publication (Sorlie et al 2001) prognosis data, in the form of DFS (disease free survival), were available for 26 of the 37 Pollack study samples. In this subset, we sought to directly identify gene-scale genomic copy number regions correlating with survival. Permutation testing was employed to determine the statistical significance of these genomic markers.

Figure 1



Results

From the integrated analysis of the van't Veer and Pollack datasets, we identified nine genes whose RNA gene expression is prognostic, and whose RNA expression is highly correlated with DNA copy number. These nine genes are highlighted in aqua in Figure 2 depicting the genomic regions of the identified prognostic regions.

DFS was available for a subset of the Pollack breast cancer DNA amplification data. This prognosis data was not used in the identification of the nine genes, and thus can be used to test the prognostic accuracy of the nine genes.

Of the 26 samples for which survival data are available, 18 were collected from patients with a poor prognosis (less than 5 years DFS), and 8 from patients with a good prognosis (greater than 5 years DFS).

Due to the small number of samples for which DFS was available, we limited our analysis to one and two gene combinations from the previously discovered nine genes.

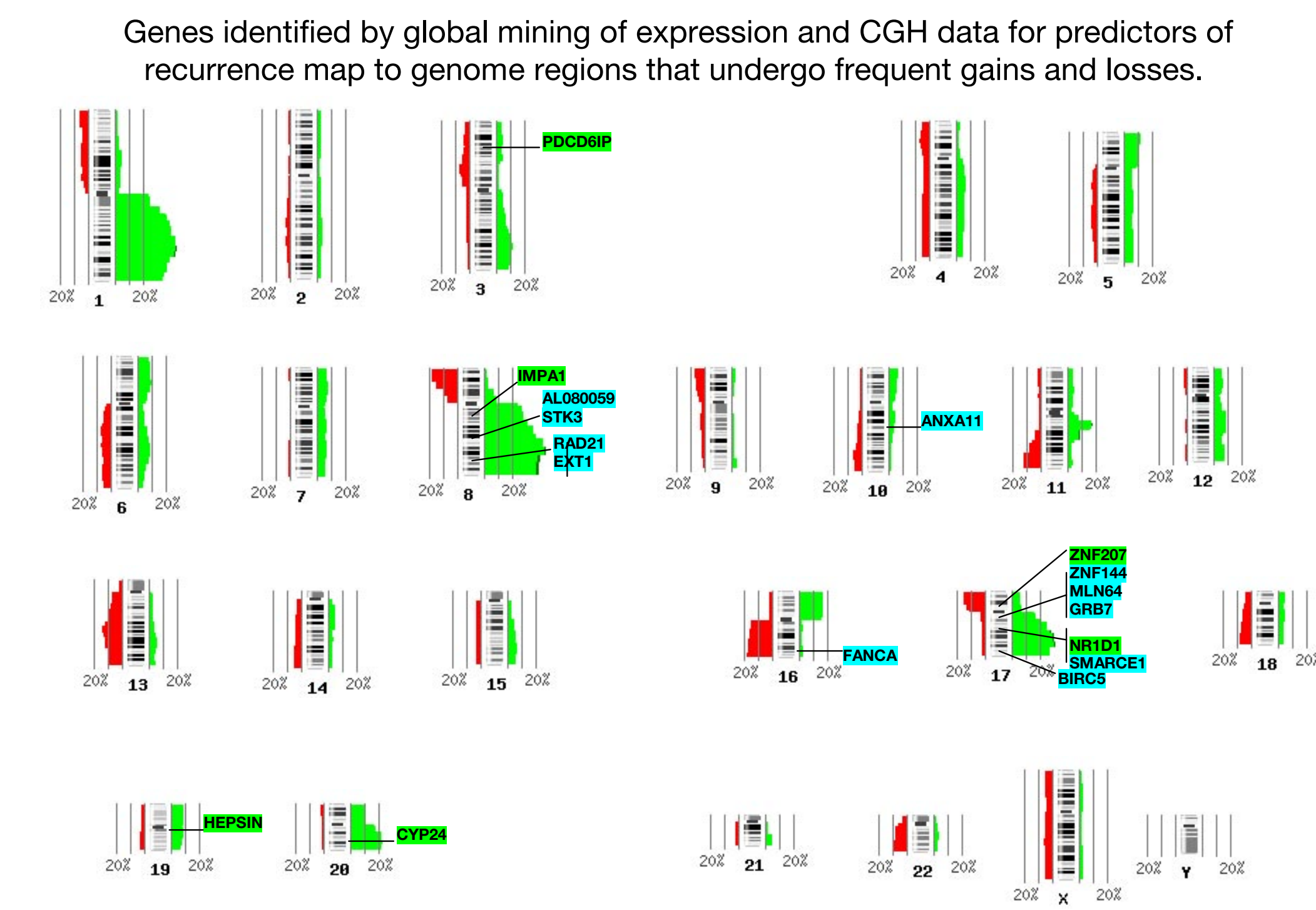
We found that the most informative gene (AL080059), as determined solely from the van't Veer gene expression data, does map to a region of high correlation between gene expression and DNA amplification, and is 73% accurate in predicting prognosis in the DNA amplification data. A combination of two (AL080059, ANXA11) of the top nine informative genes that map to such regions achieved an accuracy of 89% in leave-one-out cross validation on the independent prognosis data.

An additional two genes, GRB7 and MLN64, were identified that exhibited highly variable DNA copy number and prognostic RNA expression. These two genes are also highlighted in aqua in Figure 2.

Using only the subset of the Pollack DNA data for which DFS was available, and additional five genes were identified as significantly prognostic.

These are highlighted in green in Fig. 2. One additional gene, PDCD6IP, also highlighted in green, exhibited a significant association with hormone status.

Figure 2



Progenetix, M. Baudis
www.progenetix.com

Conclusions

In a comprehensive approach to marker discovery, integration of disparate data types can result in a relatively data-efficient technique for the identification of markers. In this example, the integration of gene expression data, clinical outcomes, and a model of the relationship between gene expression and genomic aberrations, has provided a pathway for the discovery of prognostic DNA markers.

Based on our results, we have conducted a FISH-based 308 sample validation study of the prognostic significance of the 17 markers. (Harris et al, SABCS 2004, poster 3008).

References

Pollack JR, Sorlie T, Perou CM, Rees CA, Jeffrey SS, Lonning PE, Tibshirani R, Botstein D, Borresen-Dale AL, Brown PO. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. Proc Natl Acad Sci U S A. 2002 Oct 1;99(20):12963-8. Epub 2002 Sep 24.

Sorlie T, Perou CM, Tibshirani R, Aas T, Geisler S, Johnsen H, Hastie T, Eisen MB, van de Rijn M, Jeffrey SS, Thorsen T, Quist H, Matese JC, Brown PO, Botstein D, Eystein Lonning P, Borresen-Dale AL. Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications. Proc Natl Acad Sci U S A. 2001 Sep 11;98(19):10869-74.

van de Vijver MJ, He YD, van't Veer LJ, Dai H, Hart AA, Voskuil DW, Schreiber GJ, Peterse JL, Roberts C, Marton MJ, Parrish M, Atsma D, Witteveen A, Glas A, Delahaye L, van der Velde T, Bartelink H, Rodenhuis S, Rutgers ET, Friend SH, Bernards R. A gene-expression signature as a predictor of survival in breast cancer. N Engl J Med. 2002 Dec 19;347(25):1999-2009.

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, Kerckhoven RM, Roberts C, Linsley PS, Bernards R, Friend SH. Gene expression profiling predicts clinical outcome of breast cancer. Nature. 2002 Jan 31;415(6871):530-6.