Supplementary File 1

Detailed study design and REMARK

Since none of the candidate predictive markers (including HER2 - the target itself - and PIK3CA mutation) were predictive of benefit from trastuzumab in B-31 (data not shown), we did not have any clue as to which genes might be predictive of trastuzumab benefit or resistance when we designed the study. Therefore while nCounter or QRT-PCR would be a platform of choice for gene expression profiling of FFPE tumor blocks, we did not have any candidate genes to profile with such methods. Therefore while fully aware of the limitations of using microarray gene expression profiling using degraded RNA from paraffin blocks, we had to use it to identify candidate predictive genes.

Our previous experiences using microarray platforms for FFPE samples informed us that we can interrogate molecular subtypes such as ER or HER2 with microarrays in large cohort studies (Kim et al, *J Clin Oncol* 2012). However, due to assay to assay variability, it would be impossible to use microarrays as a diagnostic test for individual patients. Hence we simply used microarrays to identify potential candidate predictive markers for trastuzumab realizing that many of the identified genes would be false positive findings. We did not pursue building a predictive model using microarray data since it would be meaningless.

In designing nCounter assay, we took into consideration that microarray measurement might not be accurate enough and therefore miss potential predictive genes (false negative findings). Therefore we also included genes that have been described as prognostic genes in the literature.

One important detail is that since we selected candidate genes from microarray data based on data analysis using clinical data at the time of unblinding, and final analysis of nCounter assay discovery cohort was based on updated clinical data with median follow up of 7 years, gene assignments have changed and no longer make sense (for example many genes from nCounter assay would be predictive in microarray analyses with updated clinical data, although they were not originally selected by microarray analysis).   Therefore we decided not to report microarray data in this manuscript.

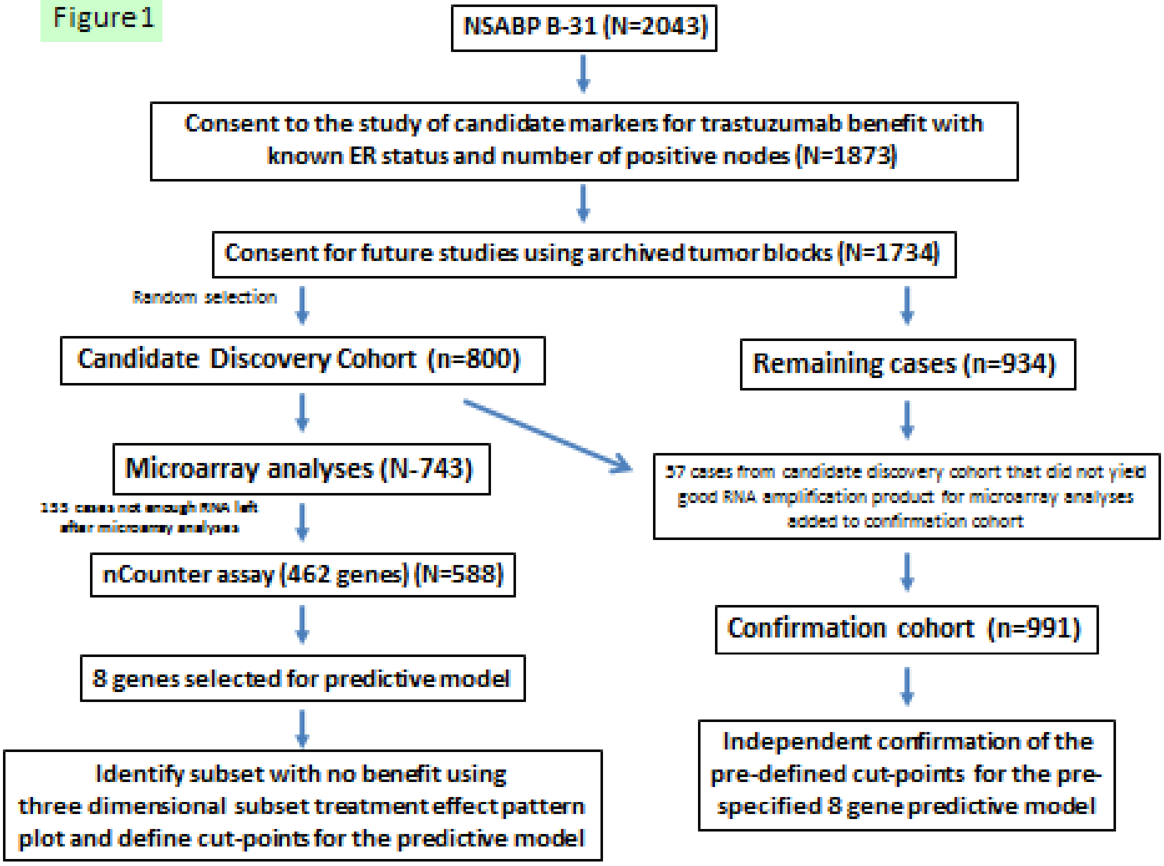
From 1734 cases that had tumor tissue with informed consent for future studies available, we randomly selected 800 cases for discovery cohort. Among those, 57 cases did not yield good RNA amplification product for microarray analyses, and therefore excluded from the discovery cohort. They were reassigned to the confirmation cohort.

While 743 cases from the discovery cohort were subjected to microarray analyses, we had data from only 588 cases from this cohort due to the reagent batch problem of nCounter assay. While nCounter platform allows robust gene expression profiling of degraded or fragmented RNA extracted from FFPE tissue samples, we found significant differences between two batch of reagents we custom ordered for discovery and confirmation set which could not be normalized using mathematical method. Therefore we had to re-assay entire discovery set using the second batch of reagents ordered for the confirmation set. During this process we had depleted extracted RNA from 155 cases from the original discovery cohort of 743 profiled with microarray and resultant final data available from 588 cases. Since those 155 cases had shallow tumor blocks, we did not want to deplete them by cutting more sections.

While this was a serious technical problem for nCounter assay when we performed them for discovery effort, it is no longer relevant since we can use synthetic targets of the 8 genes included in the model to normalize each case. Therefore we elected not to describe this problem in the paper in detail so as not to confuse readers.

We plan to use RNAseq for future studies to avoid the batch problems with nCounter assay that we encountered during this study.

Figure 1


NSABP B-31 (N=2043)
Consent to the study of candidate markers for trastuzumab benefit with known ER status and number of positive nodes (N=1873)
Consent for future studies using archived tumor blocks (N=1734)
Random selection
Candidate Discovery Cohort (n=800)
Remaining cases (n=934)
Microarray analyses (N-743)
57 cases from candidate discovery cohort that did not yield good RNA amplification product for microarray analyses added to confirmation cohort
155 cases not enough RNA left after microarray analyses
nCounter assay (462 genes) (N=588)
Confirmation cohort (n=991)
8 genes selected for predictive model
Independent confirmation of the pre-defined cut-points for the pre-specified 8 gene predictive model
Identify subset with no benefit using three dimensional subset treatment effect pattern plot and define cut-points for the predictive model

Supplementary File 2

3-D Subset Treatment Effect Pattern Plot (STEPP)

STEPP methodology is an exploratory tool for treatment by covariate interaction, which were developed by Bonetti and Gelber.[1-3] Originally, this approach only focused on one covariate, so we extended it for exploring two interaction effects simultaneously because we considered the treatment effect would be affected by both HER2 associated genes and ER associated genes.

For 3-D STEPP analysis, each subsequent subpopulation of 100 patients was formed by removing 50 patients with the lowest Covariate1 (in this study, PC1) values from the current sub-population and replacing them with the next 50 patients in the ordered list, while fixing 400 sub-population based on the ordered Covariate2 (in this study, PC2) values. Once the moving process based on Covariate 1 values were done, the next subpopulation based on Covariate 2 values were defined by removing 100 patients with the lowest Covariate 2 values from the current subpopulation and replacing them with the next 100 patients in the ordered list. These processes continued until all patients were included in at least one subpopulation. After the overlapping subpopulations were identified, the treatment effect was estimated within each subpopulation using the COX regression models adjusting for nodal status. Furthermore, this calculation was done again exchanging subpopulation setting Covariate1 for Covariate2 (thus, 400 patients were fixed based on Covariate2 values for consecutive 100 patients subpopulations based on Covariate2 values). 3-D STEPP analysis results are then shown graphically. All computational processes are provided as an SAS macro program.

References:

1. Bonetti M, Gelber RD: A graphical method to assess treatment-covariate interactions using the COX model on subsets of the data. Stat Med 19: 2595-2609, 2000.
2. Bonetti M, Gelber RD: Patterns of treatment effects in subsets of patients in clinical trials. Biostatistics 5: 465-481, 2004.
3. Bonetti M, Zahirieh D, Cole BF, et al: A small sample study of the STEPP approach to assessing treatment-covariate interactions in survival data. Stat Med 28: 1255-1268, 2009.

The SAS **TDSTEPPPLOT** Macro

Noriko Tanaka

July 18, 2011

Description

%TDSTEPPplot is a SAS macro that visually examines the interaction effect of two continuous variables and treatment on failure time with 3D plots, applying COX proportional hazard model. This method is an extension of STEPP analysis which was originally proposed by Bonetti and Gelber (2000).[1]

Invocation and details

In order to run this macro, you may need to include the following in your SAS program where you save the file 3dstepp.sas such as:

%include “c: \program file\mysasfiles\tdsteppmacro.sas” ;

Then execute the macro TDSTEPPplot.

An example macro call is:

```
options nonotes;

                    %TDSTEPPplot(ds=data1, var1=var1, var2=var2,
                    outds=outsm, rr1=300, rr2=400, r1=50, r2=100, cov=age,
                    trt=treatment, time=surv,
                        cens=censor, cind=1, maxhr=1.5);

  quit;
options notes;
```

Definition of macro variables:

<Parameters for the dataset>

DSN: name of the SAS data set containing survival times, status, and covariates

<Parameters for the variables>

Var1: continuous variable name of interest

Var2: another continuous variable name of interest

time: survival time

cens: event status indicator variable

icens: censoring status indicator variable value (ex. 1)

COVS: list of covariates, separated by blanks. Covariates must be continuous or dummy variables.

<Parameters for STEPP analysis>

Rr1: the largest number of subjects in common among consecutive subpopulations for variable 1.

Rr2: the number of subjects in each subpopulation for variable 1. (rr2>rr1)

R1: the largest number of subjects in common among consecutive subpopulations for variable 2.

R2: the number of subjects in each subpopulation for variable 2. (r2>r1)

<Parameters for the outputs>

Outds: name of the SAS dataset to create a new output dataset for 3D plot.

Maxhr: maximum value of Hazard ratio (Z axis) for the 3-D plot.

References:

1. Bonetti M, Gelber RD: A graphical method to assess treatment-covariate interactions using the COX model on subsets of the data. Stat Med 19: 2595-2609, 2000.

2. Bonetti M, Gelber RD: Patterns of treatment effects in subsets of patients in clinical trials. Biostatistics 5: 465-481, 2004.

3. Bonetti M, Zahirieh D, Cole BF, et al: A small sample study of the STEPP approach to assessing treatment-covariate interactions in survival data. Stat Med 28: 1255-1268, 2009.

Macro Program

```
%macro stepp(r1=, r2=, ds=, var=, cov=, trt=, time=, cens=, cind= );

%let window=%eval(&r2-&r1);
proc means data=&ds;
      var &var;
      output out=outds n=n;

run;

data outds;set outds;
      k=int(n/&window);
      call symput("k",trim(put(k,best.)));
      call symput("obsn",trim(put(n,best.)));

run;

proc rank data=&ds out=&ds;
      var &var;
      ranks rank;

run;

%do i=1 %to &k;

      %let f=%eval(1+&window*(&i.-1));
      %let l=%eval(&f+&r2);

      %if &i<&k %then %do;
            data data&i; set &ds;
                  if &f=< rank<=&l;

%end;

run;
      %if &i=&k %then %do;
```

```
             data data&i; set &ds;
                   if &f=< rank;

%end;

run;

       proc means data=data&i;
             var &var;
             output out=out&i median=med;

       run;

       data out&i; set out&i;
             call symput("median",trim(put(med,best.)));

       run;

       proc phreg data=data&i;
             model &time*&cens(&cind)=&TRT &cov /rl;
             Hazardratio &TRT;
             ods output HazardRatios =hr&i;

       run;

             data hr&i; set
             hr&i; i=&i;
             median=&median;

       run;

 %end;

 data hr&var; set %do s=1 %to &k; hr&s %end;; run;

 %mend;

                    %macro TDSTEPP(ds=, var2=, var1=, rr1=, rr2=, r1=, r2=,
                    cov=, trt=, time=, cens=, cind= );

 data &ds;set &ds; drop rank:; run;

 %let window1=%eval(&rr2-&rr1);
 proc means data=&ds;
       var &var1;
       output out=outds1 n=n;

 run;

 data outds1;set outds1;
       kk=int(n/&window1);
       call symput("kk",trim(put(kk,best.)));
       call symput("nall",trim(put(n,best.)));

 run;

 proc rank data=&ds out=&ds;
       var &var1;
       ranks rank1;

 run;

 %do q=1 %to &kk;
```

```
      %let f1=%eval(1+&window1.*(&q.-1));
      %let l1=%eval(&f1+&rr2.);

      %if &q<&kk %then %do;
            data d&q; set &ds; if &f1=< rank1<=&l1; run;

      %end;
      %if &q=&kk %then %do;
            data d&q; set &ds; if &f1=< rank1; run;

      %end;

      proc means data=d&q;
            var &var1;
            output out=out1_&q median=med;

      run;

      data out1_&q; set out1_&q;
            call symput("median1",trim(put(med,best.)));

      run;

              %stepp(r1=&r1, r2=&r2, ds=d&q, var=&var2, cov=&cov,
              trt=&trt, time=&time, cens=&cens, cind=&cind  );

      data hrr&q; set hr&var2;
            q=&q;
            &var1=&median1;
            rename median=&var2;

      run;

%end;

data hrall&var1; set %do t=1 %to &kk; hrr&t %end;; run;

%mend;

                      %macro TDSTEPPplot(ds=, var1=, var2=, outds=, rr1=,
                      rr2=, r1=, r2=, cov=, trt=, time=, cens=, cind= ,
                      maxhr= );

ods listing close;
%TDSTEPP(ds=&ds, var2=&var2, var1=&var1, rr1=&rr1, rr2=&rr2, r1=&r1, r2=&r2,
           cov=&cov, trt=&trt, time=&time, cens=&cens, cind=&cind );

 quit;

            %TDSTEPP(ds=&ds, var2=&var1, var1=&var2, rr1=&rr1, rr2=&rr2,
            r1=&r1, r2=&r2, cov=&cov, trt=&trt, time=&time, cens=&cens,
            cind=&cind );

ods listing;

data hrall; set hrall&var1 hrall&var2;run;

proc means data=hrall;
      var &var1;
      output out=out1 max=max1 min=min1;

run;
```

```
data out1;
      set out1;
      call symput("max1",trim(put(max1,best.)));
      call symput("min1",trim(put(min1,best.)));

run;

proc means data=hrall;
      var &var2;
      output out=out2 max=max2 min=min2;

run;

data out2; set out2;
      call symput("max2",trim(put(max2,best.)));
call symput("min2",trim(put(min2,best.)));
run;
proc g3grid data=hrall out=&outds;
   grid &var1*&var2=HazardRatio / spline smooth=.2
                axis1=&min1. to &max1. by 0.5
                axis2=&min2. to &max2. by 0.5;

run;
goptions reset=all border ;
axis3 order=(0 to &maxhr by 0.1) label=none;
proc g3d data=&outds;
   plot &var1*&var2=HazardRatio / rotate=60 grid zaxis=axis3 zticknum=14
                                   zmin=0 zmax=1.5;

run;
quit;
%mend;
```

Supplementary File 3

Model building from the candidate discovery cohort

For prediction model building with the data from the discovery cohort (N=588), we attempted two approaches.

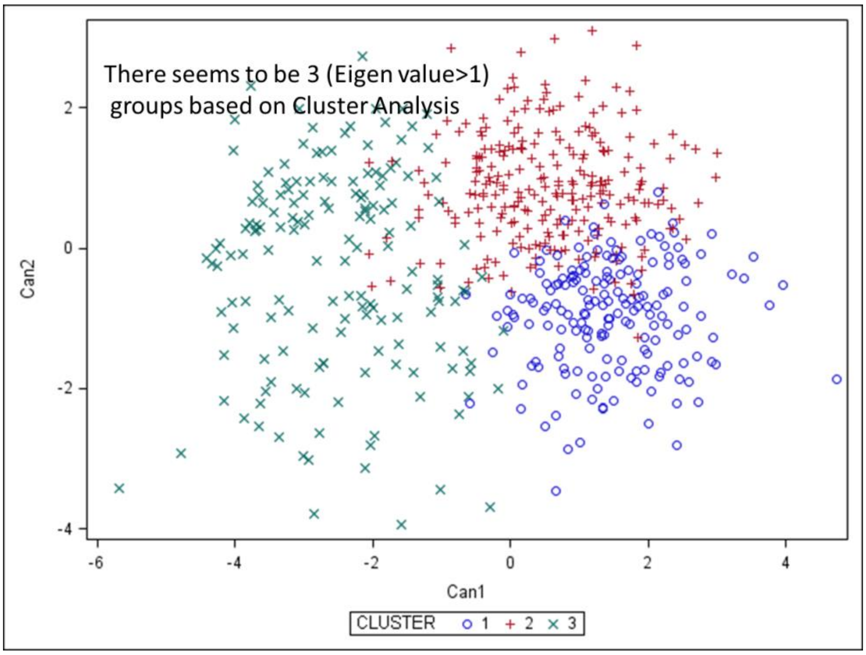
In the first approach, we attempted to select candidate genes solely according to treatment-by-gene expression interaction p values from COX model applying 10-fold Jack-knives calculated using discovery cohort(N=588).

Top rank genes are listed below.

genesymbol	cv_support	mean p-value	max p-value	min p-value
FLOT2	100	0.0025	0.0054	0.0002
UNC119	100	0.0049	0.01	0.0008
TUBB2C	100	0.0051	0.0136	0.0008
XYLT1	100	0.0054	0.0131	0.0016
CA12	100	0.0059	0.0269	0.0007
GATA3	100	0.007	0.0154	0.001
GTF3C2	90	0.0078	0.0509	0.0003
SLC39A14	100	0.0088	0.0223	0.0014
FTH1	100	0.0145	0.0347	0.0024
SUPT6H	100	0.0155	0.0385	0.0013
ACVR1B	100	0.0156	0.0349	0.0041
DKFZP434A	90	0.0166	0.0533	0.005
ILF2	90	0.0188	0.0825	0.0012
DNAJC4	90	0.0194	0.0591	0.0056
ABHD2	100	0.02	0.0477	0.002
ZACN	100	0.0214	0.0476	0.0093
TPBG	90	0.0239	0.0976	0.0041
FAM84B	100	0.0242	0.0396	0.0034
SPDEF	90	0.0243	0.0562	0.0042
DAD1	80	0.0277	0.0808	0.0074
CASC3	80	0.0297	0.1148	0.0039
MYADM	90	0.03	0.0535	0.0044
PTTG1	90	0.0316	0.1292	0.0079
UHMK1	80	0.0329	0.0827	0.0059
TMBIM6	60	0.0346	0.0666	0.0059
THOP1	80	0.0348	0.0911	0.006
ANGPTL2	90	0.0364	0.0863	0.0058
ISOC1	80	0.0366	0.139	0.005
TMSB10	90	0.0379	0.086	0.0131
PIK3CA	90	0.0388	0.2252	0.0056
SLC7A2	70	0.0401	0.107	0.0097
ORC6L	60	0.0407	0.1022	0.0088
SPP1	60	0.0408	0.0607	0.0116
CD9	60	0.0411	0.0881	0.0083
PCK2	70	0.0426	0.095	0.009
CEACAM1	70	0.0433	0.097	0.0125
RPL21	60	0.0437	0.0896	0.0159
C17orf37	70	0.0442	0.1008	0.0084
KHSRP	70	0.0458	0.1119	0.016
RASSF7	70	0.0462	0.1588	0.0111
RPL34	70	0.0477	0.1475	0.0127
ERBB2	60	0.0485	0.1114	0.0064
RPL23A	60	0.0489	0.1281	0.0116
NUF2	60	0.0497	0.1363	0.0083
EGFR	50	0.0516	0.0997	0.0122
ENPP1	60	0.0525	0.1375	0.0126
ZNF609	70	0.0528	0.0949	0.0138
NLK	60	0.0542	0.1148	0.007
IGF1R	30	0.0593	0.0954	0.0112
L3MBTL2	80	0.0603	0.2704	0.0089
LOXL3	50	0.0612	0.1314	0.0336
PRR3	60	0.0648	0.1645	0.0038
C9orf58	40	0.0657	0.1155	0.0106
B4GALT1	50	0.0665	0.1483	0.026
TBX21	60	0.0676	0.2062	0.0203
FBXW11	50	0.0682	0.1752	0.014
MTCH2	50	0.0687	0.1844	0.0097
ZNF124	40	0.0701	0.2389	0.0297
KRT7	50	0.0714	0.1418	0.024
IGKV1-5	10	0.0914	0.1663	0.0157
KLHL25	40	0.0949	0.2731	0.0235

Among the 44 genes with lowest p-values for interaction, 10 were shared by Affymetrix and Agilent microarray data when they were analyzed using the updated clinical data. Those genes are highlighted in yellow in the above table.

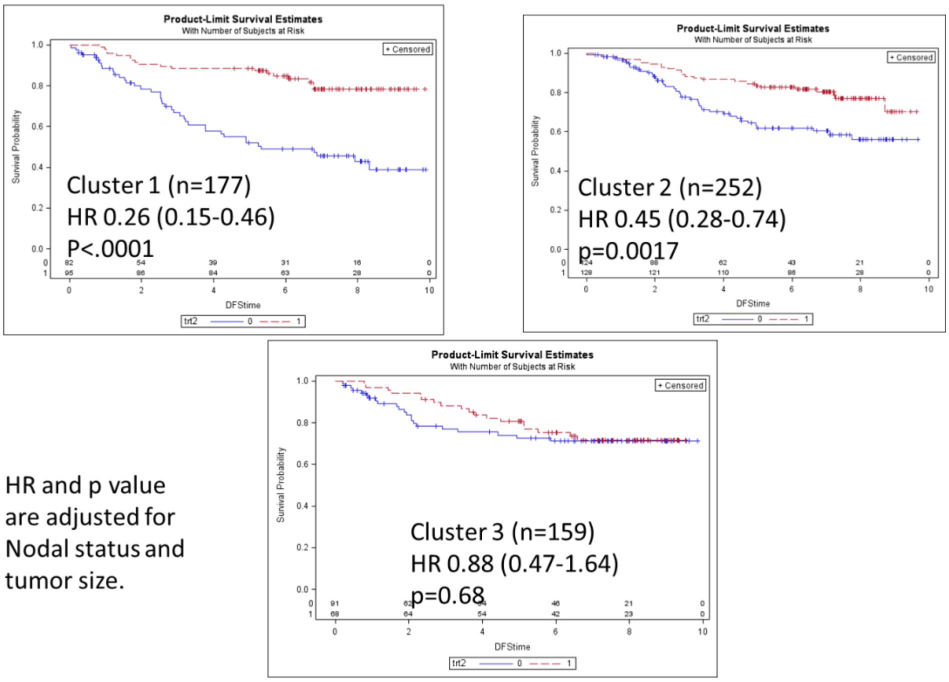
Using cluster analysis using those 10 selected genes, we found three sub-groups in the di scovery samples with differential benefit from trastuzumab.



There seems to be 3 (Eigen value>1) groups based on Cluster Analysis

KM-plots for each subgroup are shown below.

KM plot by cluster



Cluster 1 (n=177)
HR 0.26 (0.15-0.46)
P<.0001

Cluster 2 (n=252)
HR 0.45 (0.28-0.74)
p=0.0017

Cluster 3 (n=159)
HR 0.88 (0.47-1.64)
p=0.68

HR and p value are adjusted for Nodal status and tumor size.

While these genes could be used to identify subsets with differential benefit from trastuzumab, even the group with least benefit had a hazard ratio of 0.88 (0.47-1.64) with p value of 0.68. Since the control group in the latter subset had high enough recurrence rate, we thought that not using trastuzumab based on this model with hazard ratio of 0.88 would not be clinically justified. Similar efforts using different combination of genes selected by statistical criteria failed to yield clinically meaningful subsets. Therefore, we abandoned this approach.

The second approach was based on biological and clinical knowledge. We tried to build a prediction model using genes that are associated with ERBB2 (HER2) and ESR1 (estrogen receptor) mRNAs. The logic behind this approach was as follows;1) since many significant predictive genes were ER or ERBB2 associated genes as listed in the table above, 2) subset treatment pattern plots of these two genes showed trends for interesting non-linear interaction with trastuzumab, 3) we already knew that ERBB2 is the target for trastuzumab and ER status influenced response to neoadjuvant trastuzumab, and 4) IGFR1, which was a published candidate trastuzumab resistant marker was associated with ESR1.

We chose not to use ERBB2 and ESR1 alone to build the predictive model since we could not use the combination of the two genes to readily identify a subset with hazard ratio over 1.

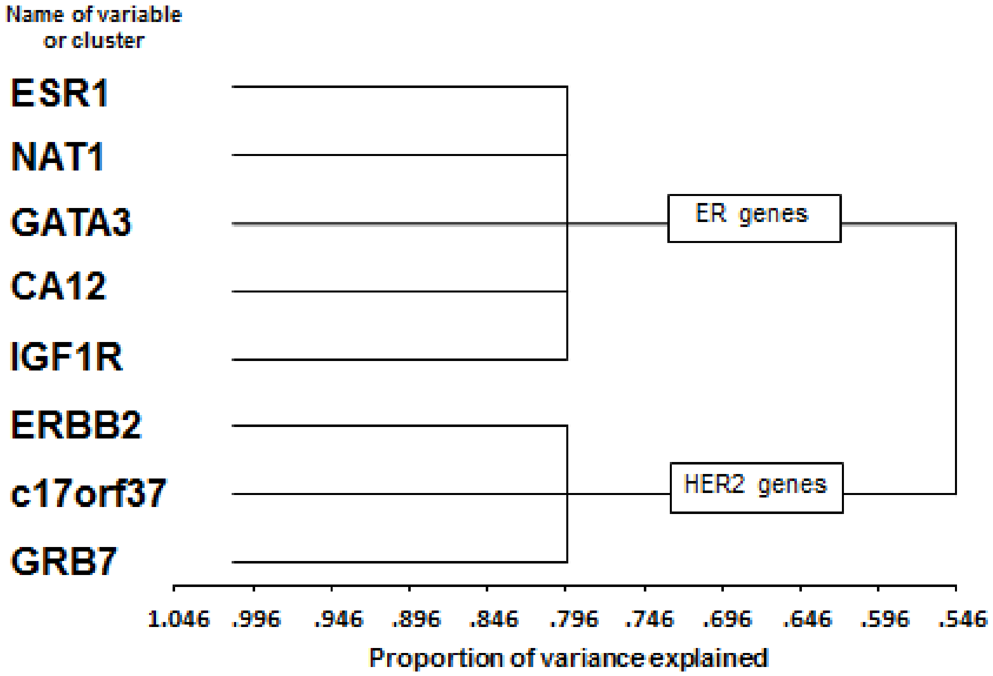
First, we chose the candidates based on the spearman's correlation coefficients between each genes and ERBB2 or ESR1 expression calculated using the discovery set.

Among the genes correlated with ERBB2 or ESR1 with spearman's correlation coefficient >0.7, we selected 8 genes with interaction p-value (min p-value in the above table) <0.1 as highlighted in yellow.

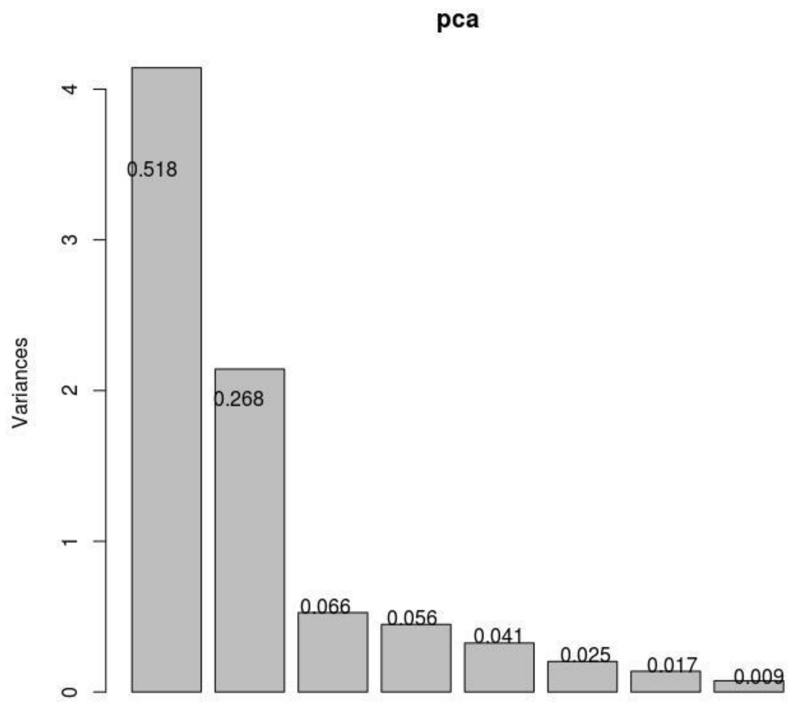
Gene Symbol	**Correlation with *ERBB2***	**Minimum Interaction P Value**
ERBB2	1	0.025
GRB7	0.912	0.06
C17orf37	0.833	0.0003
KRT7	0.498	0.047
TMEM45B	0.453	0.29
ORMDL3	0.448	0.076
C1orf93	0.427	0.1
SPDEF	0.4	0.013
VEGFA	0.395	0.24
FGFR4	0.347	0.35

Gene Symbol	Correlation with *ESR1*	Minimum Interaction P Value
ESR1	1	0.064
TBC1D9	0.757	0.49
CA12	0.733	0.0024
IGF1R	0.731	0.042
GATA3	0.727	0.0036
THSD4	0.727	0.12
NAT1	0.701	0.075
SLC39A6	0.685	0.21
SCUBE2	0.637	0.47
SIAH2	0.632	0.19

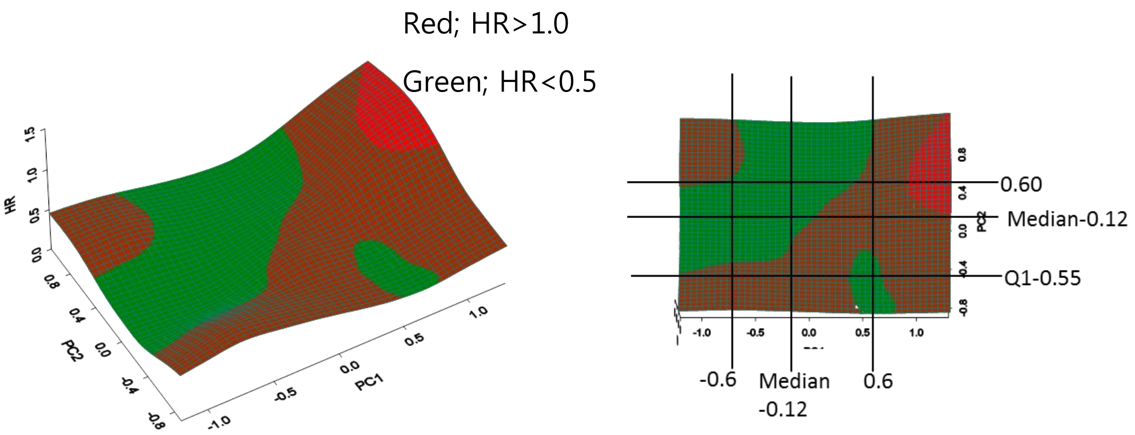
Correlation structure of the 8 selected genes is shown below.



In a principal component analysis, first two principal components of these genes accounted for 78.6% of the total variance.



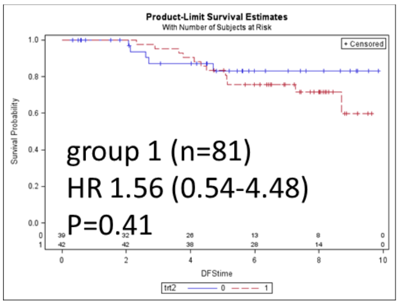
While the interaction between principal components of these 8 genes did not show linear interaction with trastuzumab, Three Dimensional Subset Treat Effect Pattern Plot (TDSTEPP) showed subset with hazard ratio of over 1 and we decided to pursue building predictive model based on these genes.

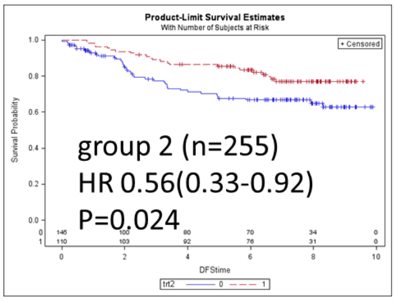


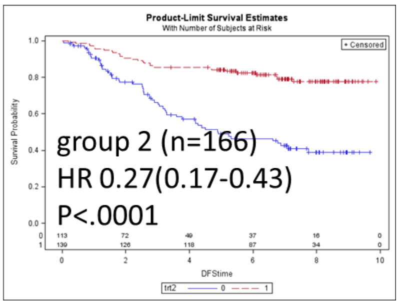
Based on the above bird's eye view of the TDSTEPP, we developed the algorithm to divide three groups (1; non-benefit group, 2; may benefit group, 3; should benefit group) as follows;

```
if factor1>0.6 and factor2>0.1 then group=1;
if -0.12<factor1 and factor2<=0.1 then group=2;
if factor1<=-0.12 and factor2<=-0.55 then group=2;
if factor1<=-0.6 and factor2>=0.6 then group=2;
if -0.12<factor1<=0.6 and 0.1<factor2<=0.6 and factor2<=factor1+0.22 then group=2;
if -0.12<factor1<=0.6 and 0.1<factor2<=0.6 and factor2>factor1+0.22 then group=3;
if -0.6<factor1<=0.6 and factor2>=0.6 then group=3;
if factor1<=-0.12 and -0.55<factor2<0.6 then group=3;
```

The above cut-points description was reduced to what was described in the manuscript by eliminating cut point description for group 2 (intermediate benefit group) since they can be identified as the remaining cases after identifying high and no benefit group, to make it less complex.







As shown above, Kaplan Meier plots of discovery set generated using the above model indicated that we may be able to identify a small subset with no benefit from trastuzumab. Furthermore due to their excellent baseline prognosis without tratuzumab, a molecular test to identify these patients may have a clinical utility.

Based on the above findings, we decided to pursue biology based prediction model for further confirmation.