# Risk Prediction for late-stage ovarian cancer by meta-analysis of 1,525 patient samples - Supplementary Information

Markus Riester, Wei Wei, Levi Waldron, Aedin Culhane, Lorenzo Trippa,
Esther Olivia, Sung-hoon Kim, Franziska Michor, Curtis Huttenhower,
Giovanni Parmigiani, Michael Birrer

January 16, 2014

Code and instructions to reproduce all results presented in this paper is available at `https://bitbucket.org/lima1/ovrc4_signew`.

## Contents

# 1 Leave-one-dataset-out cross-validation

Instead of arbitrarily setting aside some large and informative datasets for validation, we initially tested our models in a leave-one-dataset-out cross-validation procedure (Supplementary Figure S1). The goal of this procedure is to build the best possible model by using all major datasets for training, and to still obtain an accurate (in particular not over-optimistic) estimate of how well this model will perform when applied to independent data.

For the overall survival signature, we first selected the 6 largest datasets as training datasets. We excluded the small datasets ($< 75$) because small datasets often display unusual patient characteristics and because they have minor impact on the feature selection and are thus more valuable as test than as training datasets. Then we trained 6 different models and always excluded one dataset from training. These 6 models were then validated each time on the corresponding excluded datasets. This procedure thus results in one risk prediction per patient, which can then be used for model evaluation by comparing the risk prediction with the observed outcome. This is not an over-optimistic evaluation, because to calculate a particular patient's risk score, only information from different patient cohorts was utilized.

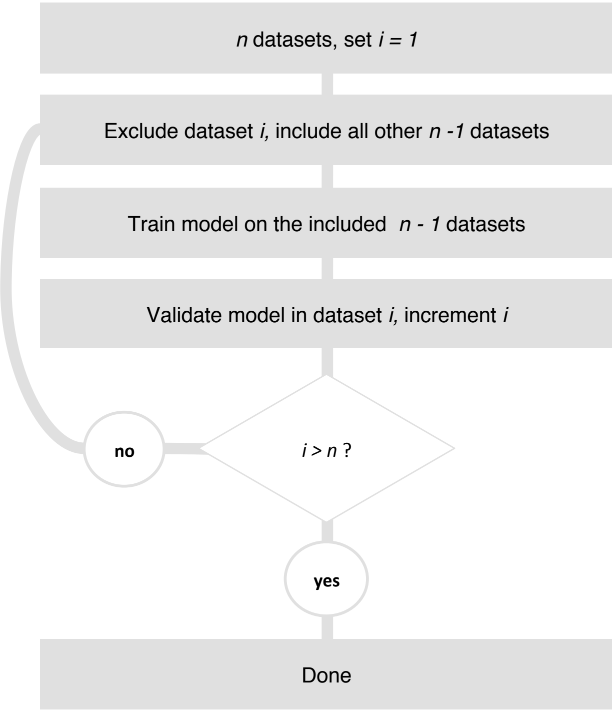

*n* datasets, set *i = 1*
Exclude dataset *i*, include all other *n -1* datasets
Train model on the included *n - 1* datasets
Validate model in dataset *i*, increment *i*
no
*i > n* ?
yes
Done

Figure S1: Flowchart leave-one-dataset-out cross-validation algorithm.

## 2 Fixed- vs. random-effects survival signature

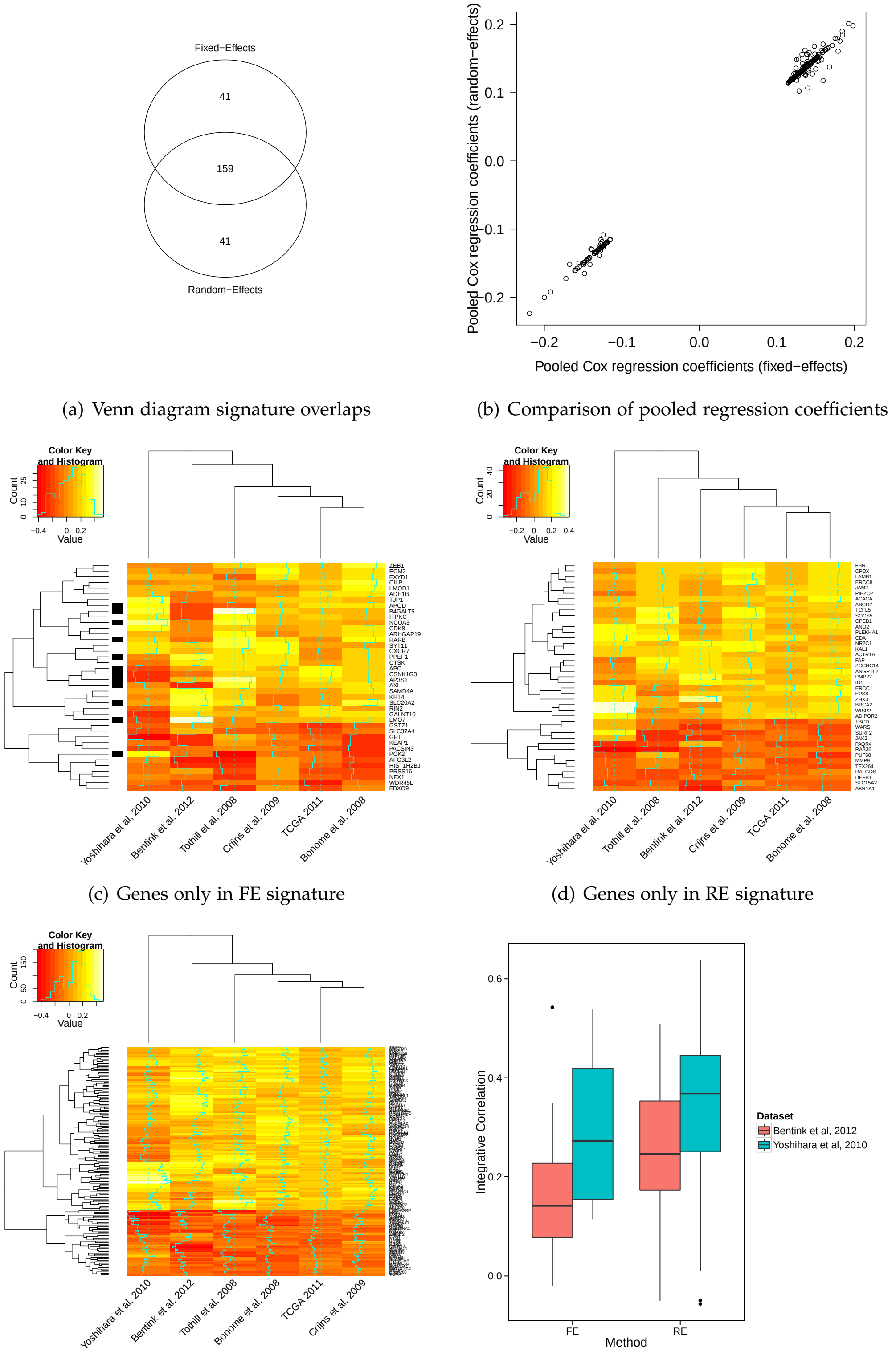
In this section, we explore how two different options for pooling the regression coefficients across datasets affect the final overall survival model. We compared a fixed-effects meta-analysis, in which coefficients are weighted by the inverse of their squared standard errors, with a random-effects model, which uses the restricted maximum-likelihood method, the default method in the `metafor` R package [27].

The overlap of the gene signatures obtained with a fixed- and a random-effects gene ranking is shown in Supplementary Figure 2(a). Out of the 200 genes in each signature, 159 were present in both signatures. The choice of the meta-analysis method had no marked impact on the pooled Cox regression coefficients; the 241 genes utilized in the two signatures had highly similar coefficients across the fixed- and random-effects meta-analysis (Supplementary Figure 2(b)). Supplementary Figure 2(c) to 2(e) show the Cox regression coefficients in all datasets, focussing on the genes present in only the fixed-effects signature, only the random-effects signature and present in both signatures, respectively (these cases correspond to the 3 sections of the Venn diagram). Genes with relatively higher heterogeneity across studies are marked in black in the heatmap. In the fixed-effects signature, 12 genes displayed statistically significant heterogeneity ($P < 0.05$, Q-Test [12]). When accounting for heterogeneity, probesets with heterogeneity drop in their ranks and are replaced in the random-effects signature with probesets displaying less heterogeneity (Supplementary Figure 2(d)).

The observed heteogeneity is almost exclusively caused by differences in the effects estimates from the Yoshihara 2010 and the Bentink et al. datasets. As a result, the random-effects analysis removed genes that have strong effect in most training datasets, but no effect or a discordant effect in either of these two studies. The random-effects model then replaced these genes with ones that have slightly weaker, but more consistent effect. This can be an advantage or a disadvantage depending on whether the heterogeneity is indicative of biological variation or of technology driven study-specific artifact.

To explore whether technical issues in these studies are the main sources of heterogeneity, we used Integrative Correlation Analysis [17]. A low integrative correlation of a gene indicates unusual pattern of expression correlations when compared to two reference datasets (Bonome and TCGA). We reported averaged results using once the Bonome and once the TCGA dataset as reference.

The analysis was done utilizing the `MergeMaid` [31] and `metaArray` [10] R packages with default Pearson Correlation. We then tested whether the genes with heterogeneity have lower integrative correlation, indicative of unusual expression, compared to the genes exclusively found in the random-effects signature (Supplementary Figure 2(f)). The analysis was performed independently for both the Yoshihara 2010 and the Bentink datasets. In both datasets, genes with heterogeneity tend to have lower integrative correlation than genes without, indicating that technical issues are at least partly a source of heterogeneity. The difference was however not statistically significant ($P = 0.1$, two-sided Student t-test).



(a) Venn diagram signature overlaps

(b) Comparison of pooled regression coefficients

(c) Genes only in FE signature

(d) Genes only in RE signature

(e) Genes in both FE and RE signature

(f) Integrative Correlation Analysis

Figure S2: Comparison of a fixed- (FE) and a random-effects (RE) ranking of genes. See main text for a description of the panels.

# 3 Survival gene signature size

In all our analyses, we fixed the gene signature size to 200 genes. This size was motivated by the fact that this size is sufficiently small to be practically useful in a clinical test and by the performance of validated and random signatures [28]. It was shown in Waldron et al. that smaller signatures tend to be less robust than large signatures. Furthermore, larger signatures allow the use of pathway enrichment tools, which are often useful for prioritizing genes for further experimental validation. Finally, because we plan to apply the models to a large cohort of FFPE specimens, we decided for a larger signature to provide some redundancy. This is important as we expect that some genes in the signatures will not be available after quality controlling.

Our algorithm weighs genes according their rank (see Methods). This means increasing the signature size is expected to have only limited influence on the prediction performance at some point as the weights of the genes decrease. Increasing the number of features in an ensemble classifier such as the compound covariate score is not prone to the overfitting as in multivariable regression, where one may optimize a model to perfectly explain the data. All tuning parameters - algorithmic details, training datasets and gene signature size - were determined beforehand to allow an unbiased use of the training data for testing purposes in our leave-one-dataset-out testing procedure.

In Supplementary Figure S3 we confirm that the signature size had only modest impact on the prediction accuracy in our algorithm, as long as the signature size was larger than 100 genes. In this figure, the prediction accuracy is reported with the C-Index metric. The C-Index is a pairwise comparison of patients, summarizing the fraction of pairs where the patient predicted to be at higher risk in fact has shorter survival. A C-Index of 0.5 would correspond to a random model, and a C-Index of 1.0 of to a perfect model. Such a perfect model would predict the correct order in which patients die. We chose this performance measure here instead of Hazard Ratios because it has an easy interpretation, is essentially parameter free and does not require a dichotomization of the prediction scores.

# 4 Comparison with the TCGA signatures

We compared our signature to the TCGA signature [5]. To apply the TCGA signature across microarray platforms, we matched the 193 probe sets in the signature to 185 unique gene symbols [22] used in curatedOvarianData. We first reproduced the reported performance of this model in the three TCGA test datasets [3, 9, 25] to ensure the correctness of our model implementation (Supplementary Figure S4). Among the 200 genes in our signature, 17 overlapped with the TCGA signature ($P < 0.001$). The p-value of this overlap between our and the TCGA signature was calculated with the hypergeometric distribution, using the number of genes common to all training datasets as background.

We used standard bootstrap to assess the statistical significance of the difference in Hazard Ratios (HR) between our signature and TCGA's. We considered all cohorts together, excluding TCGA, resulting in a total of 1031 patients where direct comparison to the TCGA signature could be made. We generated 10,000 bootstrap replicates using standard sampling with replacement from this pool of 1031 patients. From these replicates we estimated confidence intervals for the individual HRs and their difference. We reported p-value corresponding to the fraction of bootstrap replicates with higher HR in TCGA. Note that the confidence interval of the difference in HR may exclude zero (thus resulting in a statistically significant difference) even though the two confidence intervals of the individual signatures are overlapping. We further calculated the improvement in C-Index of our model compared to TCGA and found only a moderately, not statistically significant improvement from 0.605 for TCGA to 0.615 for our model.

Since the C-Index is widely known to be relatively insensitive to prediction improvements [18], we tried to estimate the required improvement in Hazard Ratio to achieve significance on the C-Index scale. To this end, we simulated better versions of our model with statistically significantly improved C-Index. We utilized a greedy Monte Carlo optimization (start with our risk scores; randomly select a patient; sample a new risk score for this patient from our risk score distribution; accept the change if the Cox model is

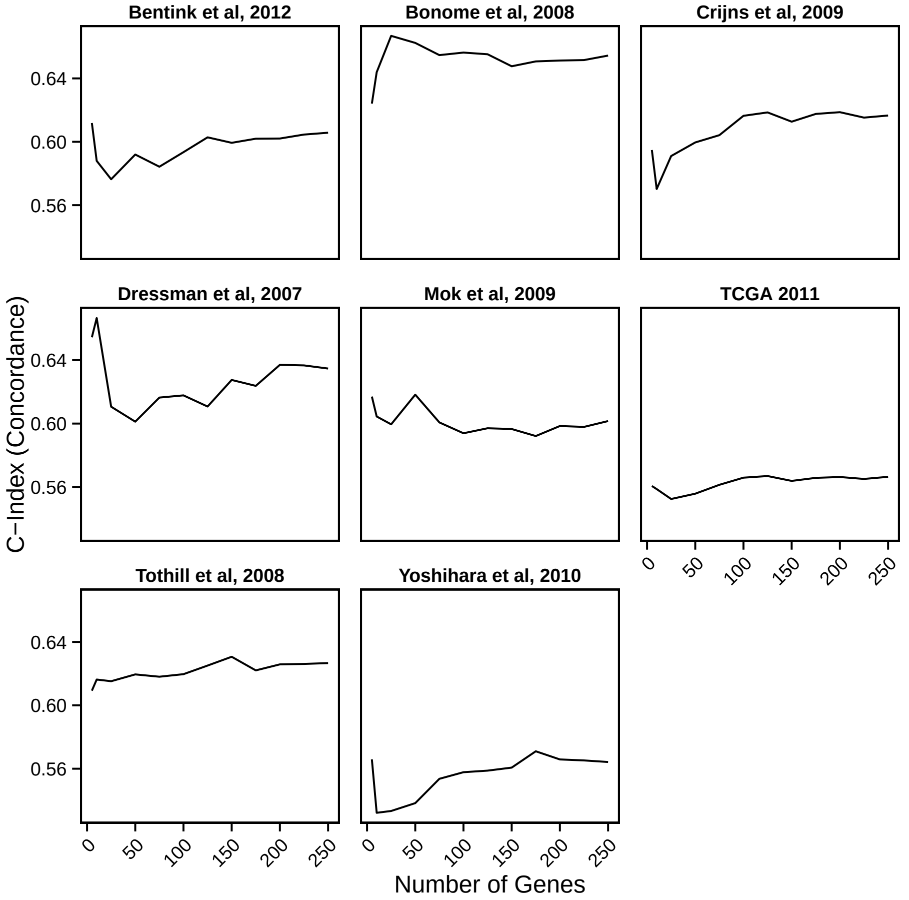


Figure S3: We used for all our gene signatures a fixed gene signature size of 200 genes. Here we show the influence of this cutoff on the prediction concordance of the overall survival signature. Each point represents the prediction concordance of a model with $x$ genes in the corresponding dataset that was trained using the remaining datasets only.

improved; stop when the C-Index improvement over TCGA achieves significance; repeat this simulation 25 times). We observe that in these simulations, an additional HR improvement of 0.27 over the 2.19 of our model (to 2.46, compared to 1.83 for TCGA, see Figure 4 in the main paper) would be needed to achieve a statistically significant improvement in C-Index. Note this simulation assumes independence of risk scores (patients with similar expression profile would have similar risk scores, however), but the main purpose of this simulation is to illustrate the relationship of the scales of HR and C-Index.

The TCGA project recently published an improved gene signature called CLOVAR [26]. We could map 87 of the 100 genes of the signature to probesets used in curatedOvarianData. 21 of these genes overlapped with our gene signature ($P < 0.001$).

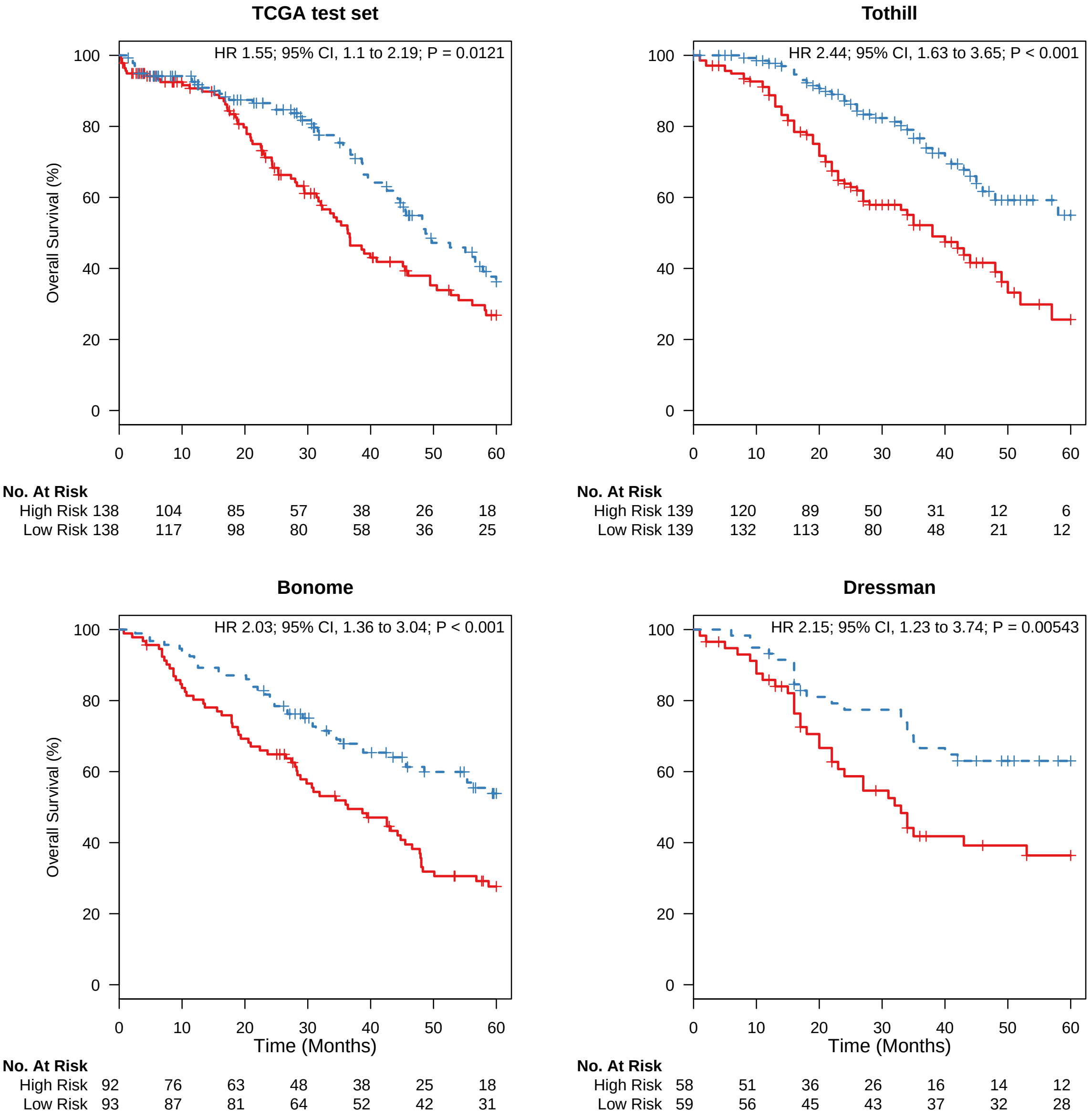


Figure S4: TCGA model applied to the author test sets shown in Figure 2c of the TCGA paper [5]. The identical results show that our implementation of the TCGA model is correct. Red survival curves correspond to high risk patients, blue curves to low risk patients.

# 5 Comparison with the CLOVAR multivariable model

We classified all datasets by TCGA subtype using a single sample GSEA variant as implemented in the GSVA package. Subtype specific gene sets were first identified with the limma package in the TCGA data using the official TCGA subtype labels[1]. Final subtype scores were obtained by subtracting the ssGSEA scores for the down-regulated gene sets from the up-regulated genes. We again used default gene set sizes of 200 genes for each subtype (100 up- and 100 down-regulated genes per subtype). Applied back to the TCGA data, this approach classified 89.2% of the samples correctly. Note that TCGA used unified expression measures obtained from multiple platforms for training, not the Affymetrix data we used in our meta-analysis. In Supplementary Figure S5, we show an association of subtype with overall survival in all datasets except TCGA consistent with the report of Verhaak et al. [26]. The immunoreactive subtype had in both TCGA and the remaining datasets the best prognosis. Poor survival was in general observed for samples classified as mesenchymal.

We then tested for consistent overlaps with the Australian Ovarian Cancer Study Group (AOCS) subtypes as published by Tothill et al. [25]. All mesenchymal samples were assigned to the AOCS cluster c1, which had poor prognosis in the Tothill data (Supplementary Figures S5C and S6). Most immunoreactive samples were assigned to the c2 cluster and c2 samples had consistent with our results better outcome in the Tothill dataset.

Verhaak et al. provided ssGSEA scores for various public datasets in their Supplement and we compared these scores with the scores of our implementation. We found a high correlation of scores ($\rho = 0.88$, Supplementary Figure S5D). Note that a perfect correlation is not expected, since we used a different pre-processing protocol of the public data and utilized different probesets.

Verhaak et al. proposed a multivariable model including tumor stage, debulking status, BRCA1/2 mutation status and ssGSEA scores for the immunoreactive and mesenchymal subtypes. BRCA1/2 mutation status was unavailable for our validation datasets. As the survival association of the ssGSEA scores were discovered in most of our validation datasets, we tested a multivariable model using the CLOVAR patient stratifications in high- and low-risk (based on the median of CLOVAR risk scores in all datasets except the validation data), tumor stage, debulking status and ssGSEA scores for all 4 subtypes. This model was then 5-fold cross-validated using all datasets combined. This model performed very similarily to the one proposed by the authors using only 2 ssGSEA scores, but did not require any biased feature selection.

The Pearson correlation of the risk scores of our meta-analysis signature and both TCGA signatures (for Verhaak et al. based on the CLOVAR signature only, not the multivariable predictions) are shown in Supplementary Figure S7. In this figure, we further compare the models with the random-effects variant of our model, in which the univariate Cox regression coefficients were pooled with a random-effects model as implemented in the `metafor` package. The predictions from the random-effects model were very similar compared to the default fixed-effects model (Pearson correlation > 0.99).

[1] `https://tcga-data.nci.nih.gov/tcgafiles/ftp_auth/distro_ftpusers/anonymous/other/publications/ov_exp/TCGA_489_UE.k4.txt`

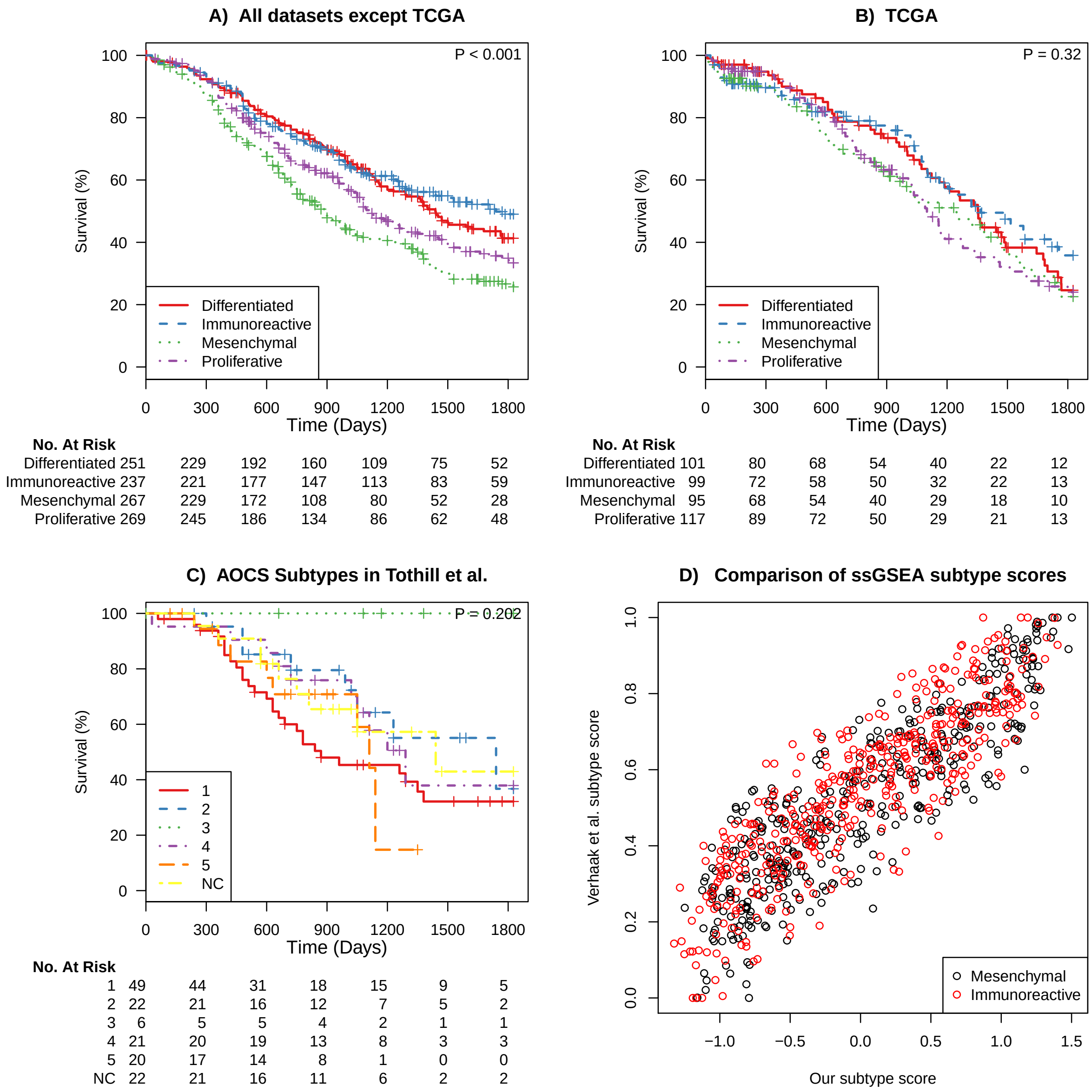


Figure S5: Association of subtype and overall survival. The model proposed by Verhaak et al. [26] utilizes a subtype score and the purpose of this figure is to demonstrates the correctness of our subtype model implementation. (A) All training and validation datasets excluding TCGA. Note the very similar survival curves in the corresponding figure 2B in Verhaak et al. [26]. (B) Stratification of TCGA samples by subtype. (C) Kaplan-Meier curves of the subtypes proposed by the Australian Ovarian Cancer Study Group (AOCS) in Tothill et al. [25]. This analysis corresponds to Figure 5B of the Tothill study, with the difference that here we show only the late-stage, high-grade, serous tumors used in our meta-analysis. (D) This scatterplot shows the high correlation of both the Mesenchymal and Immunoreactive ssGSEA subtype scores from our implementation and the official scores we extracted from the Verhaak et al. Supplement. The Mesenchymal and Immunoreactive scores are utilized in the Verhaak et al. model. Shown are the scores for the Bonome, Dressman, Tothill and Yoshihara samples.

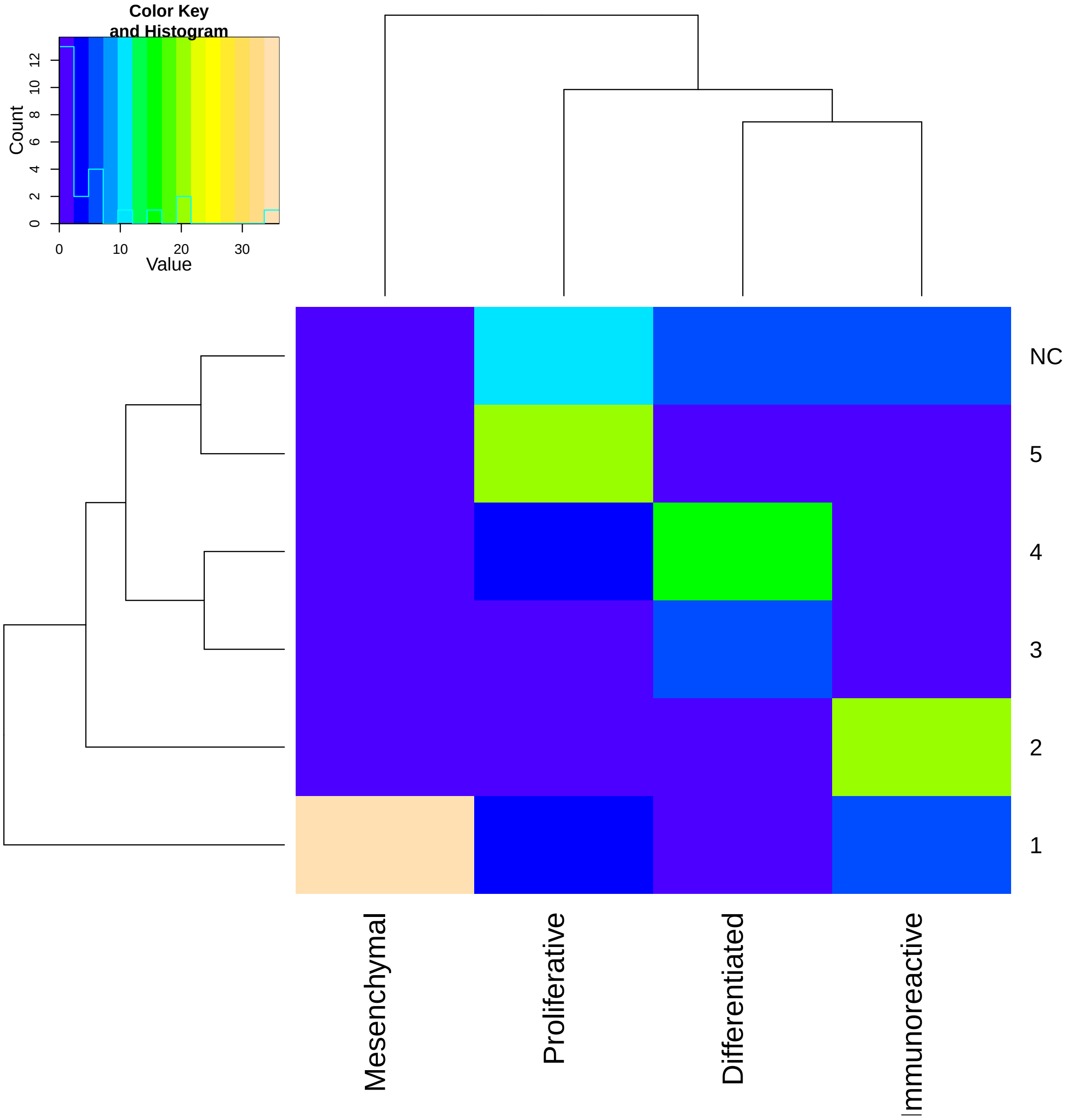


Figure S6: Comparison AOCS and TCGA subtypes in Tothill et al. [25]. Here we compare our classification of Tothill samples in TCGA subtypes with the official Tothill et al. clusters. The colors as indicated in the legend on the upper left corner visualize the number of patients in the pairwise AOCS (rows) and TCGA (columns) subtype combinations. For example, 35 patients of the AOCS cluster c1 were classified as Mesenchymal.

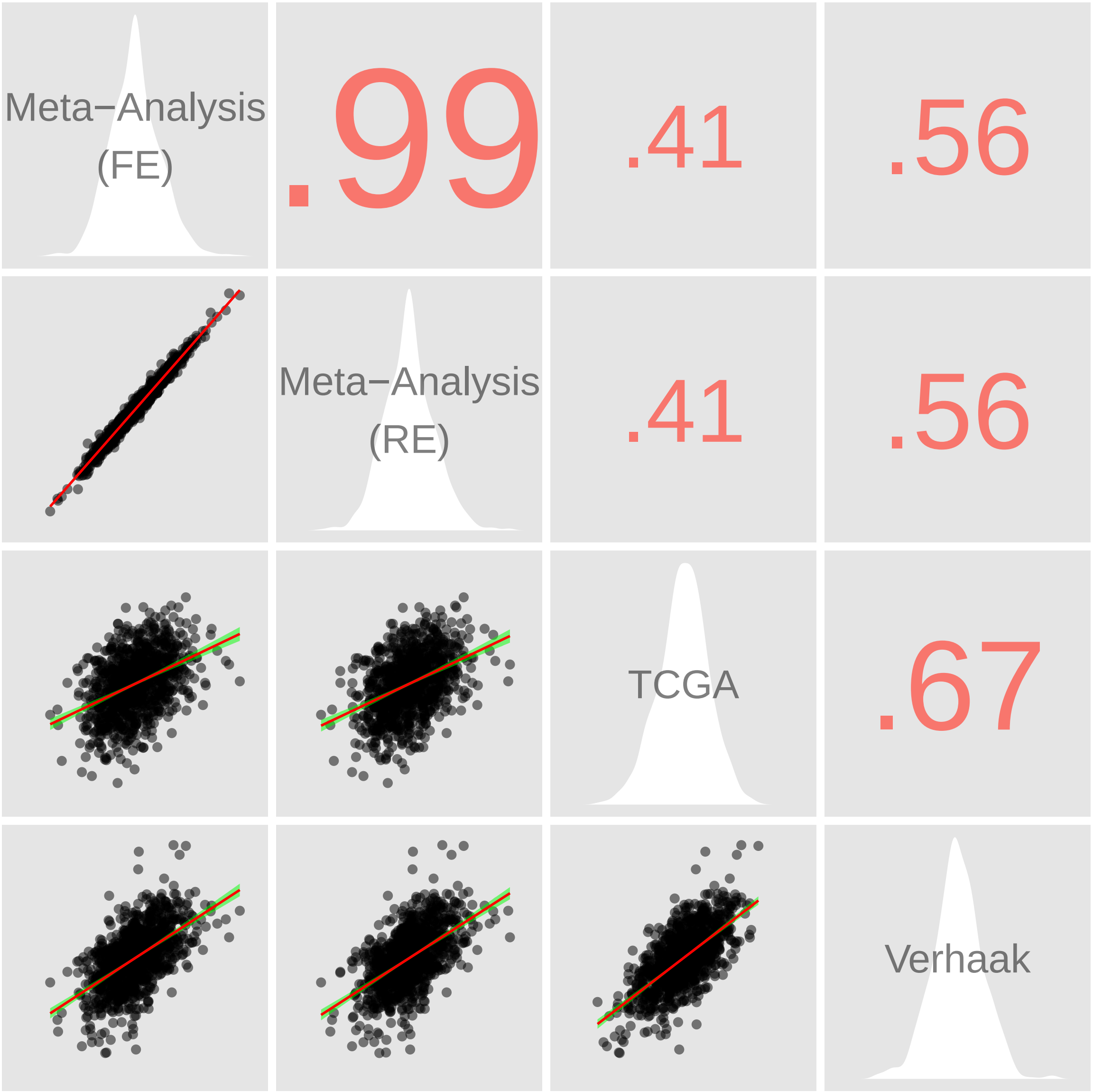


Figure S7: Pairwise Pearson correlation of the gene signature risk scores for the Meta-Analysis, TCGA and Verhaak et al. signatures. For the Meta-Analysis signature, we show risk scores correlation obtained with our default fixed-effects (FE) meta-analysis and with a random-effects model (RE). Numbers in the upper-right, tringular half of the matrix are the Pearson correlation coefficients, which were all statistically significant ($P < 0.05$). Pairwise scatterplots of the risk scores are shown in the lower-left half and the risk score histograms are shown on the matrix diagonal.

# 6 Comparison with clinical characteristics

In the main text, we presented a Kaplan-Meier analysis of our risk stratifications (Figure 4a) and of a multivariable model utilizing these stratification together with debulking status (optimal vs. suboptimal) and tumor stage (III vs. IV), the most important and commonly available characteristics with known survival association. In Supplementary Table S2, we provide the Cox regression coefficients (the log-transformed Hazard Ratios) for the multivariable models. For datasets with available age at time of diagnosis, we included age in the model.

# 7 Survival signature pathway analysis

Pathway analysis was conducted using the Pathway studio 7.1 program (Ariadne Genomics). All 200 genes in the debulking signature and the overall survival signature were input into the program for analysis respectively. Pathway enrichment analysis (build-in algorithm) was used to identify statistically significant pathways within the signature. A p-value less than 0.05 was considered significant, indicating the associated pathway identification was unlikely to be resulted from chance.

Our survival signature did not reveal dominant pathways, unlike the debulking signature. This is likely related to the more complex and diverse biological determinant factors for patient overall survival. Nevertheless, 25 genes in our 200-gene signature can be enriched through TGF-$\beta$ pathway, suggesting its poor prognostic impact on patient overall survival (Supplementary Figure S8). Secondly, 9 genes can be linked to PDGF signaling, in which the overexpression of both receptor (PDGFA and PDGFB) and ligand (PDGFD) indicates poor outcome. The TAF activation in the OS signature, marked by COL11A1 [25], is also noticed, considering the synergistic effects of TFG-$\beta$ and PDGF in TAF activation [7]. Functionally, 11 genes have been demonstrated to be involved in EMT (SNAI2, ZEB1, BMI1, C13ORF15, EDNRA/ET-1, PDGFD, SERPINE1/PAI-1, PDPN, CUX1, SERPINA1 and SPDEF), which is closely related to tumor metastasis and chemoresistance to portend poor prognosis in ovarian cancer. Recently, the emerging roles of PDGF signaling (especially via PDGFD) in EMT and cancer stem cell maintenance during tumorigenesis have been described [8, 30]. In addition, the endothelin/ET-1 pathway also positively enhance the expression of SNAI2, thus EMT [20, 19].

# 8 Meta-analysis is superior to single study training

The training sample size was positively correlated with accuracy of patient risk stratifications, up to the maximum training sample sizes of 1,250 (Supplementary Figure S9). This finding demonstrated that (i) a meta-analysis of microarray datasets even from different platforms is superior compared to single study training and (ii) our leave-one-dataset-out approach, in which we removed training datasets to obtain unbiased estimates of the final signature's HR, is not an over-optimistic estimate, because removing training datasets as expected made the signature worse.

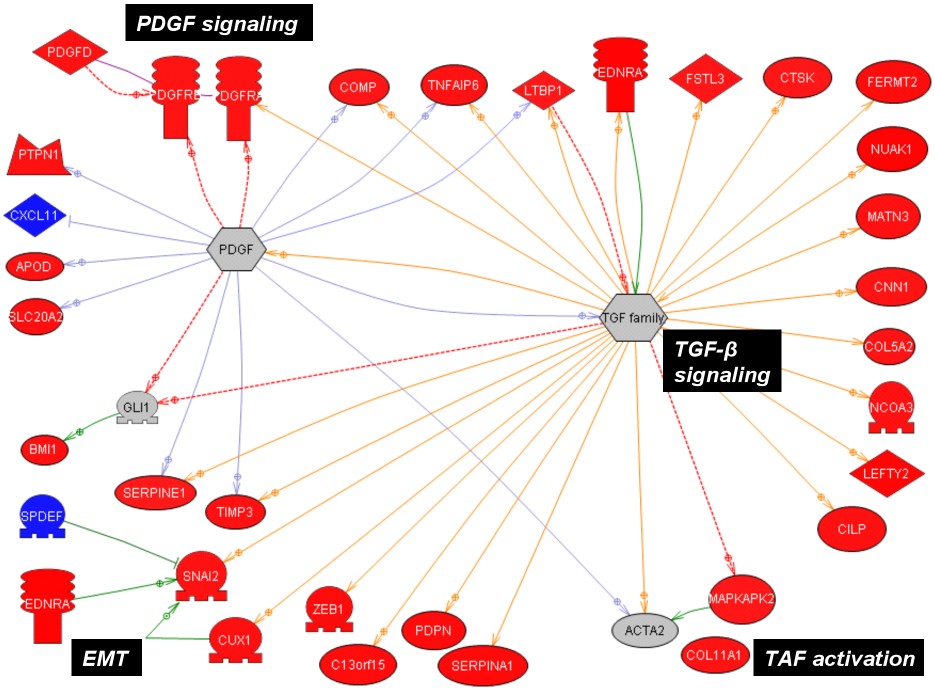


Figure S8: Pathway analysis of the OS signature. A gene is labeled in red when it portends poor prognosis. Conversely, genes indicate favorable outcome are labeled in blue. Red broken arrows: direct stimulatory modification; Orange arrows: TGF-$\beta$ signaling activated transcriptional regulations; Blue solid arrows: PDGF signaling activated transcriptional regulations; Green arrows: other direct regulations.

| | Bentink 2012 | Bonome 2008 | Crijns 2009 | Dressman 2007 | Gillet 2012 | Konstantin. 2010 | Mok 2009 | Tothill 2008 | Yosh. 2010 | Yosh. 2012 |
|---|---|---|---|---|---|---|---|---|---|---|
| Gene Signature (Low- vs high-risk) | $-0.54^{*}$ | $-0.55^{**}$ | $-0.59^{**}$ | $-0.82^{*}$ | $-0.81^{**}$ | $-0.78$ | $-0.69^{*}$ | $-0.78^{**}$ | $-0.57$ | $-0.79^{*}$ |
| | (0.24) | (0.19) | (0.20) | (0.35) | (0.29) | (0.44) | (0.32) | (0.27) | (0.35) | (0.33) |
| Debulking (subopt. vs. opt.) | 0.18 | $0.38^{*}$ | | 0.28 | $1.14^{***}$ | | | $-0.29$ | $0.79^{*}$ | 0.36 |
| | (0.27) | (0.19) | | (0.36) | (0.32) | | | (0.29) | (0.37) | (0.35) |
| Stage (IV vs III) | 0.42 | 0.27 | | 0.37 | $-0.14$ | | | $1.50^{**}$ | $-0.13$ | 0.37 |
| | (0.31) | (0.22) | | (0.50) | (0.37) | | | (0.46) | (0.42) | (0.33) |
| Age | 0.01 | $0.03^{**}$ | 0.01 | | 0.01 | | | 0.01 | | |
| | (0.01) | (0.01) | (0.01) | | (0.01) | | | (0.01) | | |
| Num. events | 71 | 127 | 113 | 36 | 54 | 23 | 41 | 63 | 38 | 40 |
| Num. obs. | 124 | 182 | 157 | 58 | 93 | 42 | 53 | 121 | 84 | 91 |
| Missings | 3 | 3 | 0 | 1 | 0 | 0 | 0 | 19 | 0 | 0 |
| PH test | 0.57 | 0.05 | 0.75 | 0.03 | 0.20 | 0.59 | 0.36 | 0.45 | 0.58 | 0.03 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table S2: Cox regression coefficients for multivariable models utilizing the gene signature risk stratifications and the available characteristics with known overall survival association when provided by the authors. Numbers in brackets are standard errors. The row *PH Test* reports the p-value indicating whether the proportional hazards assumption of the model is violated [11].

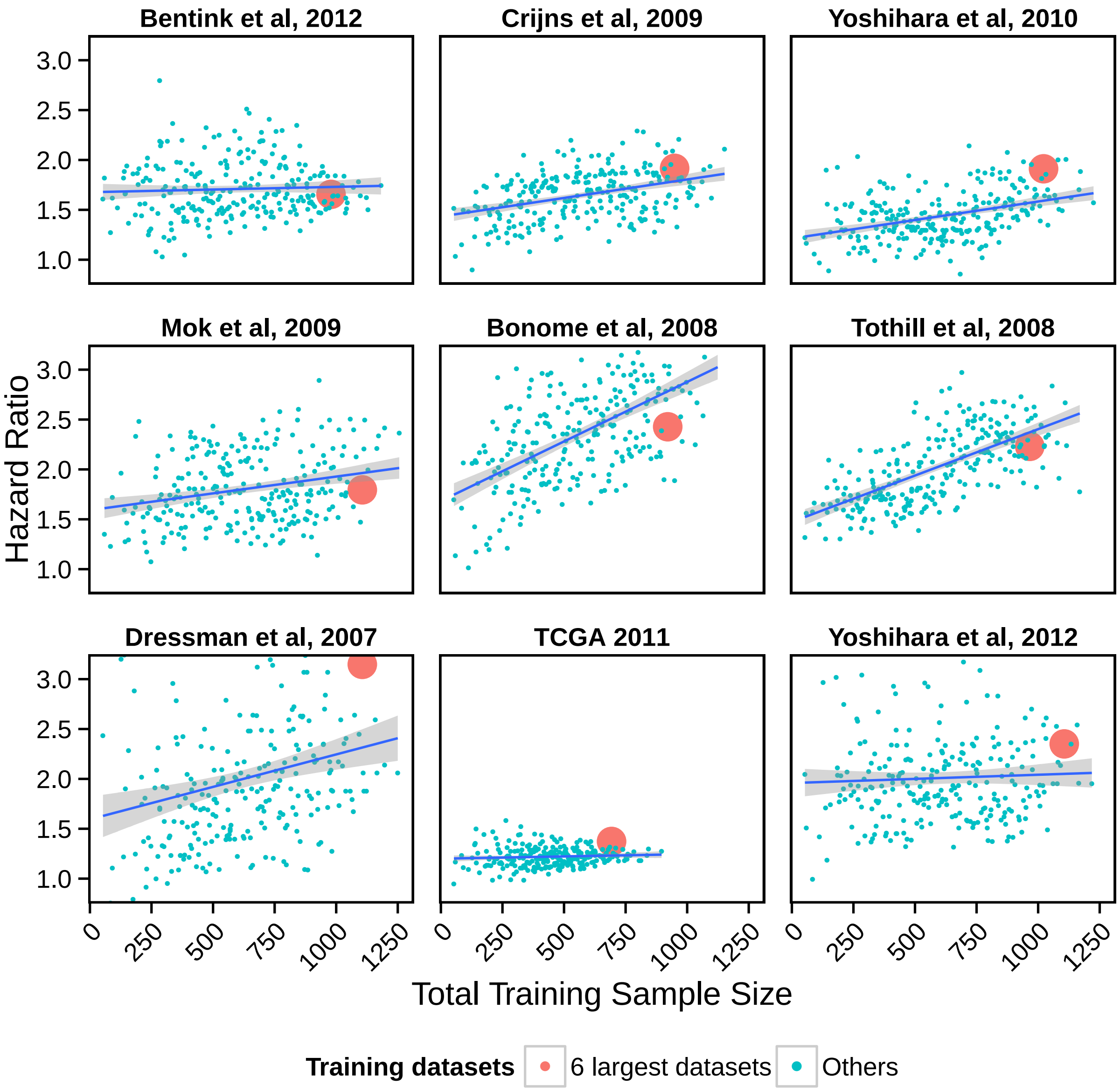


Figure S9: Prediction accuracy as a function of training sample sizes. This plot shows the improvement of predictions when training samples sizes were increased. For each of the 9 shown datasets, 255 different models were trained using the remaining 8 datasets only. These 255 ($2^8 - 1$) models correspond to all possible combinations of the 8 remaining datasets. Each point in the plot represents a training dataset combination and the combination's total sample size is shown on the x-axis, its Hazard Ratio (HR) in the validation data on the y-axis. The results of our training datasets, the 6 large studies published before March 2012, are marked with a red dot. This analysis showed that increasing the sample size via meta-analysis typically increased the model HRs. TCGA data was further identified as a difficult validation dataset.

# 9 Comparison with the Berchuck 2004 debulking signature

We generated a gene signature for suboptimal debulking surgery and validated this signature by *leave-one-dataset-out* cross-validation. The meta-analysis summary is shown in Supplementary Figure S10 as ROC curves.

We compared our results to the only published model predicting debulking success [2] we were aware of, with the corresponding ROC curves shown in Supplementary Figure S11. Because Berchuck et al. did not provide the coefficients of their model, we subtracted the average expression of genes down-regulated in suboptimal from the average of up-regulated genes. The authors further did not specify the exact probe sets and provided Unigene or Genbank accession numbers for only a subset of genes in their signature. We manually tried to identify current HGNC symbols for the genes. We could map 21 of the 32 genes utilized in their model to probes used in curatedOvarianData. None of these 21 genes overlapped with our two signatures.

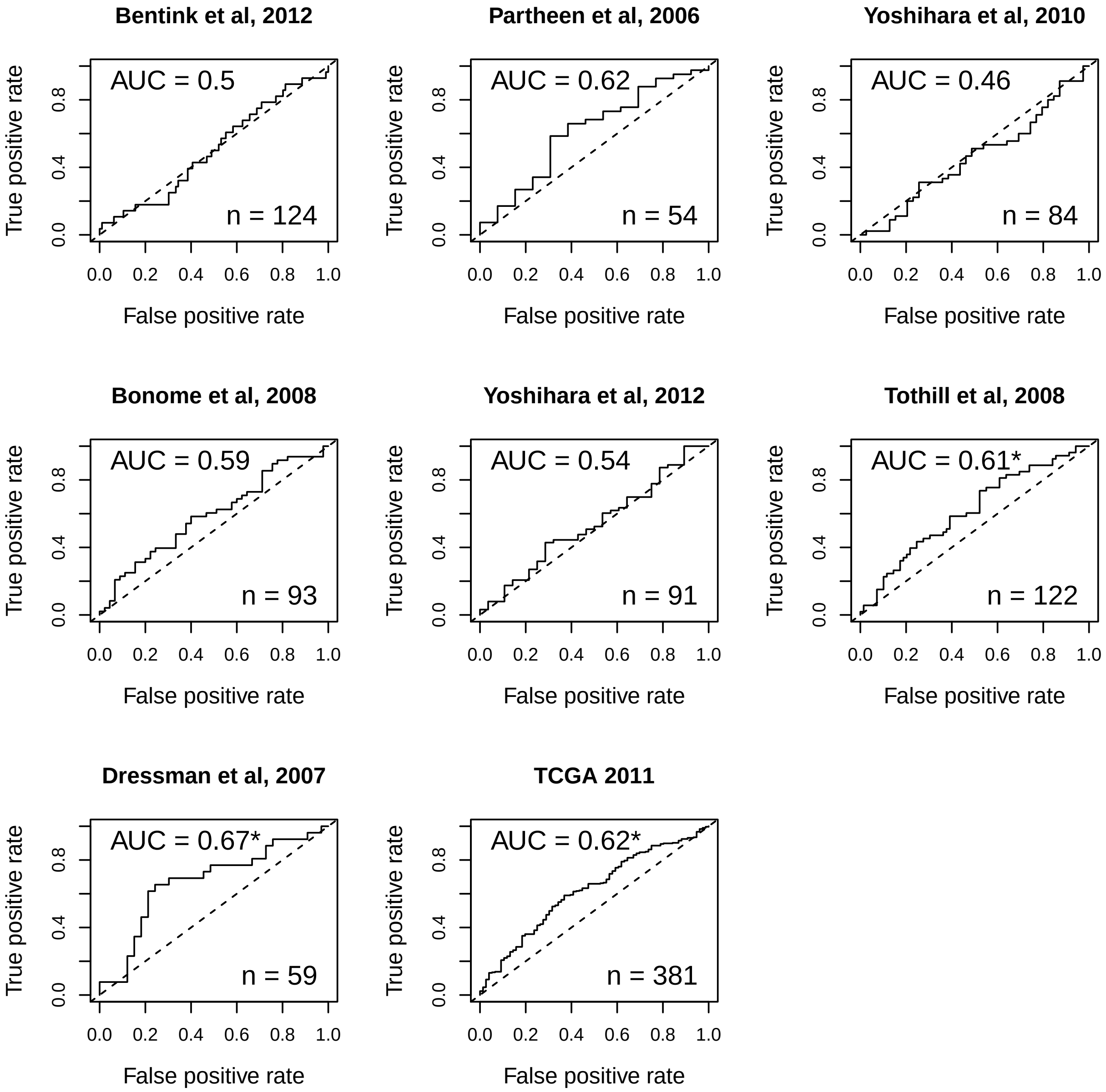
In Supplementary Table S3, we show the logistic regression coefficients of the gene signature risk scores adjusted for FIGO stage (III vs. IV).

|  | Meta-Analysis | Berchuck et al. 2004 |
|---|---|---|
| Intercept | $-1.12$ | $-1.33^{*}$ |
|  | $(0.60)$ | $(0.62)$ |
| Gene Signature | $0.01^{**}$ | $0.03$ |
|  | $(0.00)$ | $(0.02)$ |
| Stage (III vs IV) | $0.49^{**}$ | $0.57^{**}$ |
|  | $(0.19)$ | $(0.20)$ |
| Num. obs. | 1004 | 946 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table S3: Prediction of debulking status. The table lists the regression of our leave-one-dataset-out cross-validated meta-analysis debulking gene signature signature and the signature published by Berchuck et al. [2]. The predictions were adjusted for tumor stage. Numbers in brackets are standard errors.

ROC curves for public microarray data of our top-ranked hit *POSTN* are shown in Supplementary Figure S12.



Bentink et al, 2012

Partheen et al, 2006

Yoshihara et al, 2010

Bonome et al, 2008

Yoshihara et al, 2012

Tothill et al, 2008

Dressman et al, 2007

TCGA 2011

Figure S10: Prediction of suboptimally debulked tumors in a leave-one-dataset-out cross-validation. The prediction model calculates for each sample a score. The higher the score, the higher the probability the tumor will be not optimally debulkable. For each dataset, the model is trained using only the remaining datasets. ROC curves visualize the true and false positive rates as a function of the probability cutoffs. Datasets with sample size smaller than 75 were not used for training. AUCs statistically significantly ($P < 0.05$) larger than 0.5 are marked with an asterisk.

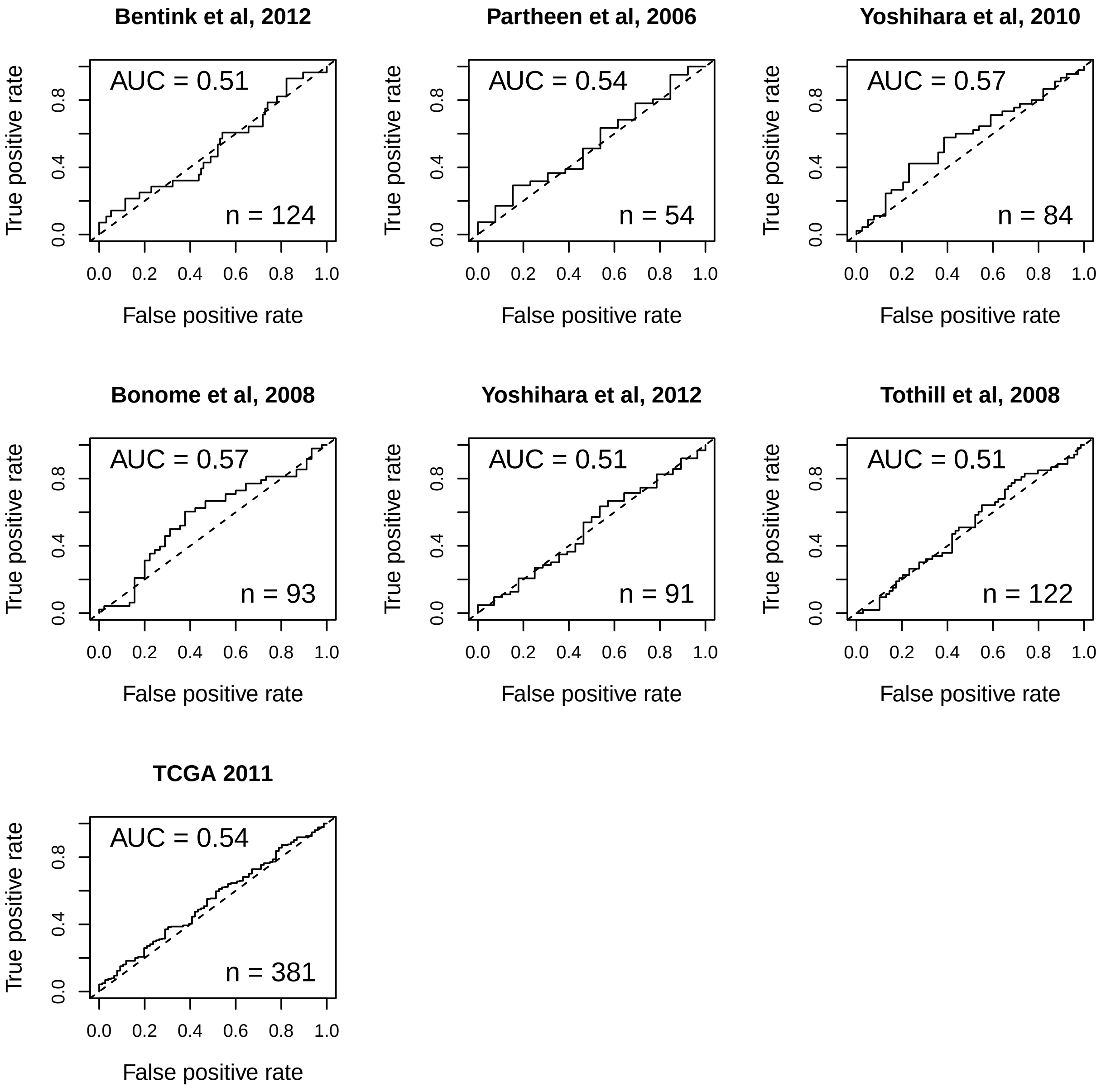


Figure S11: Prediction of debulking status with the Berchuck et al. signature [2] as in Supplementary Figure S10. The Dressman data was excluded because a subset of Dressman samples was used for training.

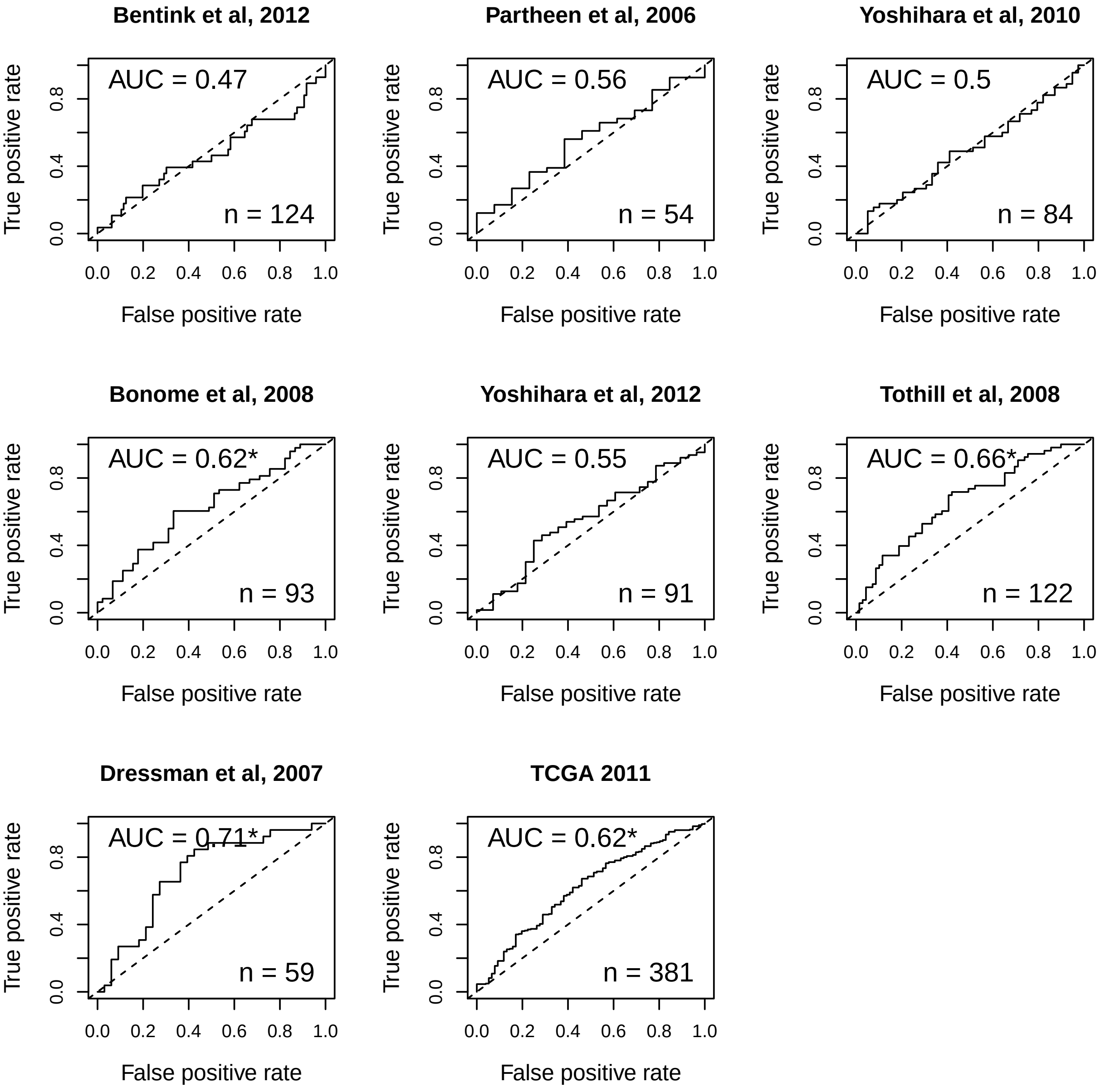


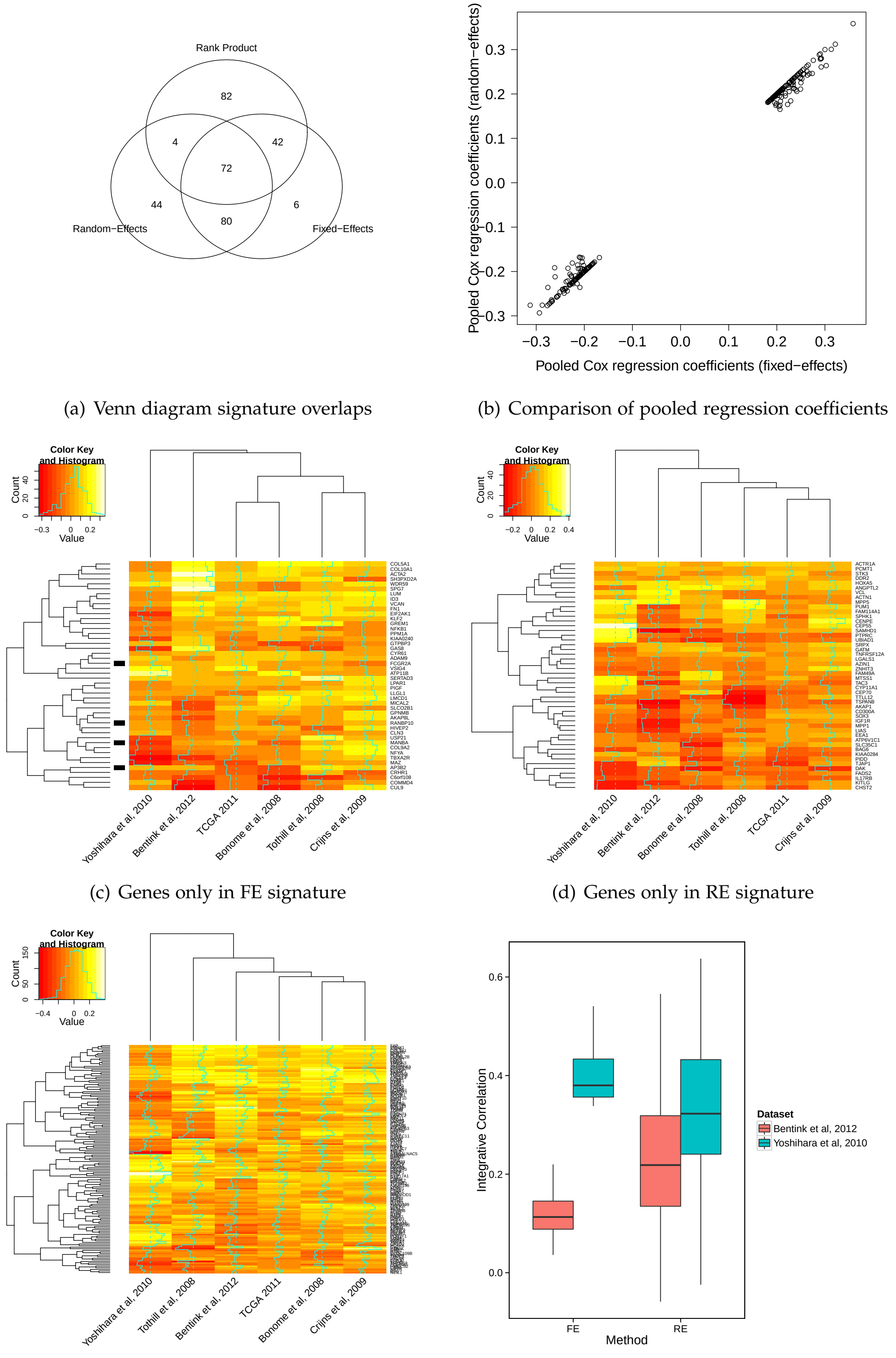
Figure S12: Prediction of debulking status with the POSTN expression alone as in Supplementary Figure S10.

# 10 Fixed- vs. random-effects debulking signature

As in section 2, we examined the overlap of the debulking gene signatures obtained with a fixed- and a random-effects gene ranking (Supplementary Figure 13(a)). We further compared both fixed-effects and random-effects signature with a signature obtained with the Rank Products method [4], a frequently used alternative microarray meta-analysis method. The Rank Products method currently does not support censored outcome and could not be compared to the overall survival signatures. The Rank Products signature was more similar to the fixed-effects (overlap of 114 genes) than to the random-effects signature (overlap of 76 genes), most likely because Rank Product also weights datasets strictly according their sample size.

We observed a lower number of genes with marked heterogeneity in the fixed-effects signature (4 genes) compared to the overall survival signature. The Yoshihara 2010 and the Bentink et al. datasets were again the most common source of heterogeneity, albeit less clear than for the overall survival signature (Supplementary Figures 13(c)-13(e)).

Integrative correlation analysis [17] showed that there were no systematic differences between the genes with statistically significant heterogeneity and the ones without (Supplementary Figure 13(f)). We expect, however, that the largest datasets (most importantly the Bonome dataset published from our lab and the TCGA samples) are similar to our validation samples in terms of surgical procedures and pathologic review. Accounting for heterogeneity with a random-effects model may overly weight smaller studies for which we have only limited surgery information.



(a) Venn diagram signature overlaps

(b) Comparison of pooled regression coefficients

(c) Genes only in FE signature

(d) Genes only in RE signature

(e) Genes in both FE and RE signature

(f) Integrative Correlation Analysis

Figure S13: Comparison of a fixed- (FE) and a random-effects (RE) ranking of genes associated with debulking status. See main text as well as Supplementary Figure S2 and accompanying text for description of the panels.

## 11 Debulking gene signature size

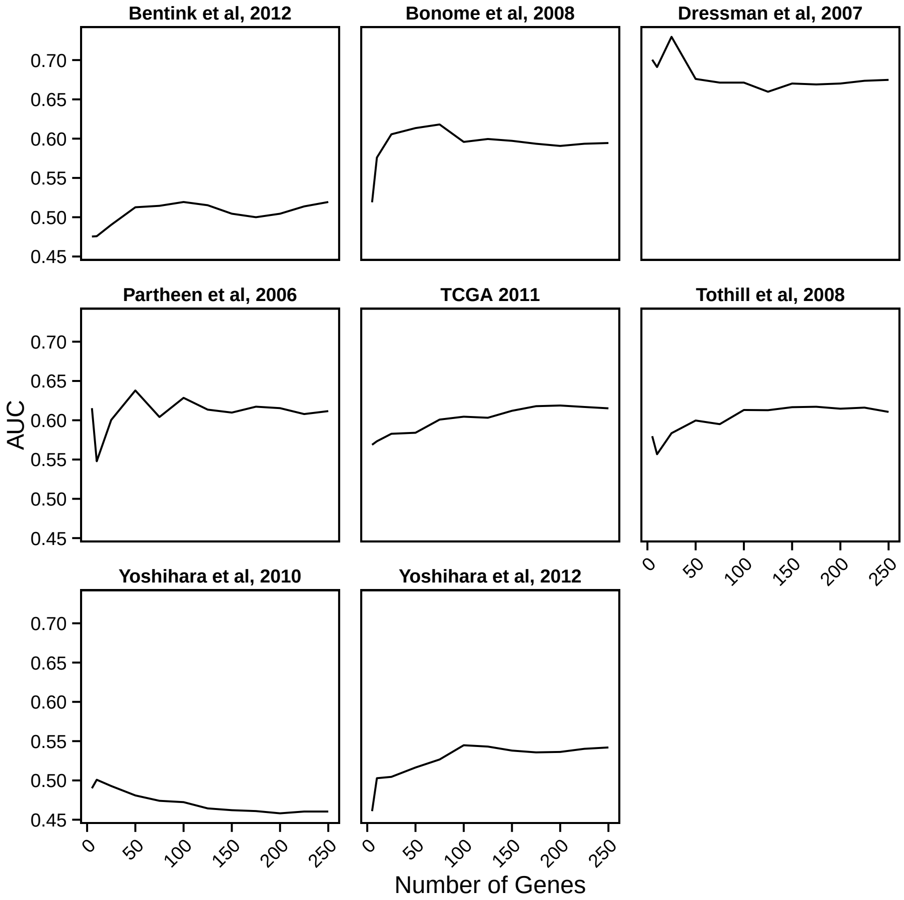


Figure S14: We used for all our gene signatures a fixed gene signature size of 200 genes. Here we show the influence of this cutoff on the prediction accuracy of the debulking signature. Each point represents the AUC of a model with $x$ genes in the corresponding dataset that was trained using the remaining datasets only. For the Bonome et al. dataset, this shows the leave-one-dataset-out cross-validated performance in the 93 samples used for training. See main text for the AUCs of the Bonome validation Affymetrix and qRT-PCR data.

# 12 IHC and qRT-PCR validation of genes associated with surgery outcome

## 12.1 Gene selection criteria

qRT-PCR was used to validate as best as possible the signature because the technique is more quantitative than arrays. Therefore, the genes were selected based upon the proposed biology and focused upon the identified pathways. Except for PTCH1, which was selected to verify the contribution of hyperactivated Hedgehog pathway, a potent metastasis and EMT inducer [14], all genes selected were involved related to TGF-$\beta$ signaling (increased TGFBR2 as one of the potential modulators, Figure 5). In addition to their direct oncogenic roles, POSTN and FAP have been demonstrated as markers of the activation of tumor associated fibroblasts, which play a definite role in tumor growth and metastasis [7]. Although less details were known about the function of CXCL14, NUAK1 and TNFAIP6, all three genes have been shown as positive regulators of tumor metastasis [29, 6, 21]. Additionally, NUAK1 mediates TGF-$\beta$ induced hypoxia tolerance [24], which is critical for both metastatic initiation and the adaption of micrometastases. Upregulated CXCL14 upon TAF activation further enhances the potential of TAFs to support tumor metastasis and angiogenesis [1]. Furthermore, the expression of CXCL14, NUAK1 and TNFAIP6 all predicts poor prognosis as revealed by our meta-analysis, indicating the potential importance of these genes in ovarian tumor biology. Finally, all selected genes encode secreted proteins or kinases and are thus ‘targetable’. We expect the validation will contribute the design of novel targeted therapeutic regime to improve the rate of optimal debulking.

For IHC candidate selection, POSTN and CXCL14 are the two genes ranked the highest by fold-change and FDR in the debulking signature for which there were commercially available antibodies which had been used in published studies. In addition to these, we selected an additional protein target which could be used as evidence of pathway activation. Phospho-Smad2/3 was selected to confirm the TGF-$\beta$ hyperactivation in tumors which could not be optimally debulked.

## 12.2 IHC Scoring

Scoring of the IHC level of each protein was conducted by determining the percentage and intensity of positive cells in three different areas at 100x magnification for each tissue core. First, the percentage of positive cells in each section was scored with a 5-point scale: 0 for <5%, 1 for 5-25%, 2 for 26- 50%, 3 for 50-75%, and 4 for over 75%. Second, the intensity of positive signal was scored with a 3-point scale: 1 for weak staining, 2 for moderate staining, and 3 for intense staining. The weighed score of staining intensity of each section was obtained by multiplying the percentage score by the intensity score (the maximum weighed score is 12). Since each sample was represented by two tissue cores, the average score was calculated for the IHC intensity of each sample (listed in Table S1). A categorical system was then applied as follows: Class 0 (-) score 0-3; Class 1 (+) score 4-6; Class 2 (++) score 7-9; Class 3 (+++) score 9-12. The IHC was scored by two individuals including one pathologist for independent IHC scoring. The scoring was done blinded to the clinical data.

## 12.3 Multivariable Models

As regression coefficients obtained from microarray data are not directly translatable to qRT-PCR or IHC measurements, we only used the signs of the coefficients from the microarray-based signature. This means that all genes were equally weighted, with the expression levels of down-regulated genes in suboptimal subtracted from the ones of up-regulated genes. Group sizes for patient stratification in high-, medium- and low-risk corresponded to the numbers of suboptimal and optimal tumors for high- and low-risk, respectively. The 33% of high-risk samples with lowest risk and the 33% of low-risk samples with highest risk were then classified as medium-risk.

In Supplementary Table S4, we show that POSTN, pSmad2/3 and CXCL14 are predictors debulking surgery outcome independent of tumor grade and stage. In the multivariable models with all three proteins, both adjusted and unadjusted for stage and grade (shown in columns 1 and 2 of Supplementary Table S4), inclusion of CXCL14 could not further improve the model.

We further validated selected genes by qRT-PCR in 78 Bonome samples not used for training. In Figure S15, we show the Spearman correlation of the qRT-PCR expression values and the Affymetrix signal intensities.

| | IHC | IHC adj. | POSTN adj. | pSmad2/3 adj. | CXCL14 adj. |
|---|---|---|---|---|---|
| (Intercept) | $-6.35^{***}$ | $-21.64^{***}$ | $-15.94^{***}$ | $-20.56^{***}$ | $-14.39^{***}$ |
| | (1.01) | (4.72) | (4.11) | (4.21) | (3.73) |
| POSTN | $0.26^{**}$ | $0.29^{**}$ | $0.44^{***}$ | | |
| | (0.08) | (0.10) | (0.08) | | |
| pSmad23 | $0.41^{**}$ | $0.56^{***}$ | | $0.75^{***}$ | |
| | (0.12) | (0.16) | | (0.14) | |
| CXCL14 | 0.21 | 0.21 | | | $0.52^{***}$ |
| | (0.11) | (0.14) | | | (0.11) |
| Grade | | 2.32 | 1.88 | $3.10^{**}$ | 1.32 |
| | | (1.22) | (1.17) | (1.14) | (1.10) |
| Stage | | $2.25^{***}$ | $2.07^{***}$ | $1.80^{***}$ | $1.87^{***}$ |
| | | (0.61) | (0.54) | (0.51) | (0.48) |
| Num. obs. | 177 | 168 | 168 | 169 | 169 |

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table S4: Multivariable prediction of debulking status. Shown are multivariable models based on IHC staining only and models adjusted (adj.) for tumor stage (III vs. IV) and grade (2 vs. 3). Numbers in brackets are standard errors.

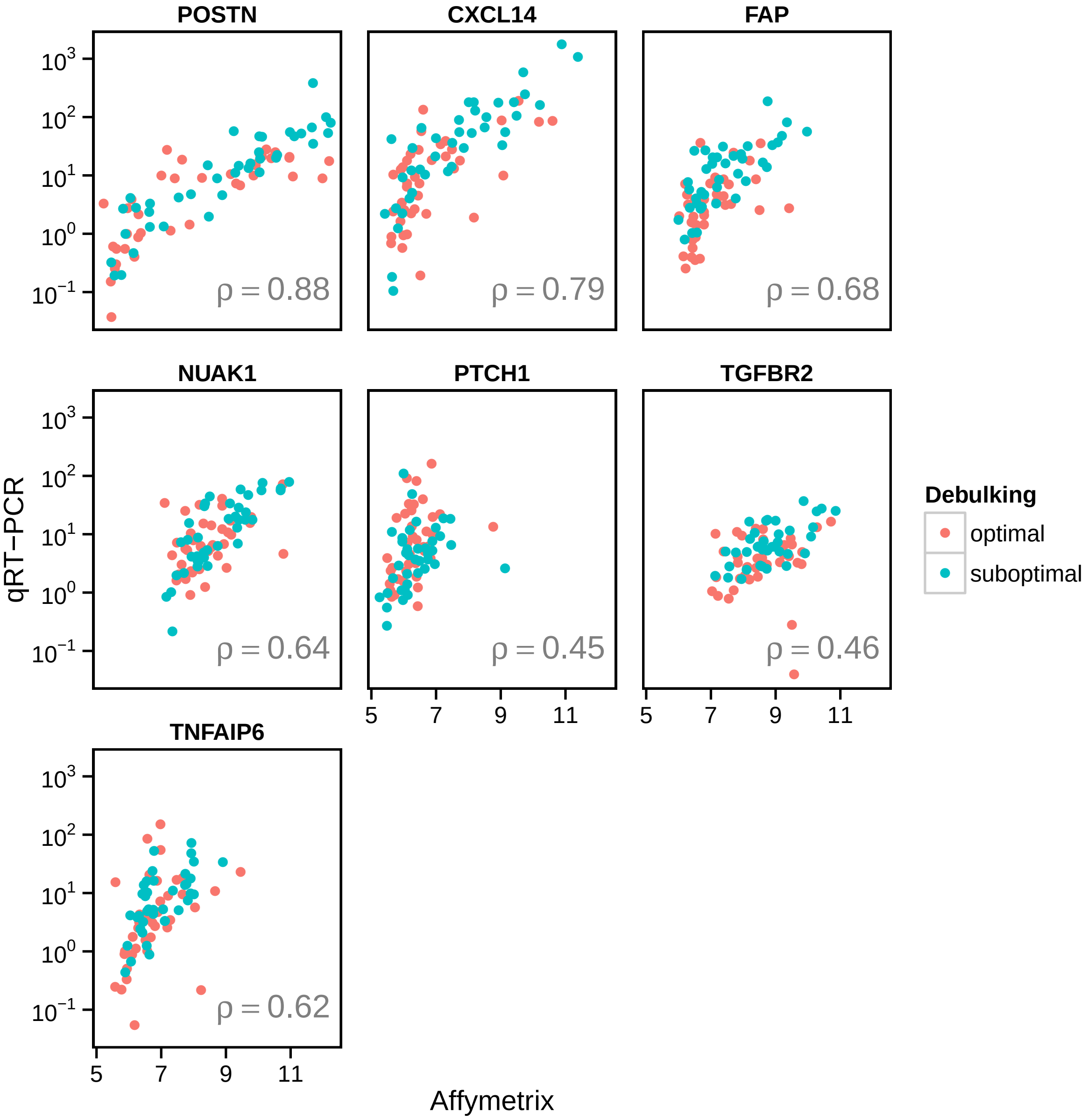


Figure S15: Validation of selected genes by qRT-PCR in the Bonome validation dataset, a subset of 78 samples (39 optimal and 39 suboptimal tumors). Points represent patients, the x-axis shows the Affymetrix fRMA normalized intensities, the y-axis the qRT-PCR level. The Spearman correlation of platforms was statistically highly significant for all genes ($P < 0.001$).

# 13 Role of POSTN in ovarian cancer

The rationale underlying the debulking signature is that the biology of the tumor contributes the ability to cytoreduce a patient. Even with large tumor burdens, the frequent confinement of ovarian cancer to the abdominal cavity makes it amenable to debulking. Therefore, optimal debulking is more likely to be achieved in cancers that are inherently less disseminative. Enhanced tumor metastasis may increase the chance of surgically inaccessible tumors and thus lower the rate of optimal debulking. The debulking signature developed in this study reveals the co-activation of TGF-$\beta$, Ras/MAPK/EGR-1 and Hedgehog pathways as well as stromal (TAF) activation in suboptimal debulked tumors. These changes may directly contribute to malignant phenotypes with elevated EMT, extracellular matrix remodeling and pro-angiogenic activity to facilitate tumor dissemination (especially parenchymal metastases) and thereby render the hindrance for surgical resection. POSTN is known to positively drive each step of tumor metastasis (migration, invasion, angiogenesis, formation of the metastatic niche, and the growth of the micrometastases).

POSTN is a secreted cell adhesion protein belonging to the superfamily of TGF-$\beta$-inducible proteins [13]. Periostin binding to the integrins ($\alpha$v$\beta$3, $\alpha$v$\beta$5, and $\alpha$6$\beta$4) directly or indirectly (by recruiting RTKs such as EGFR) activates the Akt- and FAK-mediated signaling pathways, leading to increased cell survival, angiogenesis, migration, invasion, and importantly, epithelial-mesenchymal transition of carcinoma cells [16]. In addition to the metastatic initiation, POSTN is also important in the maintenance of metastatic niche for the efficient colonization, growing and vascularization of micrometastases, a speed-limit step for tumor dissemination into second sites [23, 15]. POSTN is frequently overexpressed in ovarian cancer. The therapeutic potential of POSTN has been demonstrated in an orthotropic mouse xenograft model for ovarian cancer, through a neutralizing antibody mediated substantial reduction of tumor burden and the number of tumor foci [32].

# References

[1] M Augsten, C Hägglöf, E Olsson, C Stolz, P Tsagozis, T Levchenko, M J Frederick, A Borg, P Micke, L Egevad, and A Ostman. CXCL14 is an autocrine growth factor for fibroblasts and acts as a multimodal stimulator of prostate tumor growth. *Proc Natl Acad Sci U S A*, 106(9):3414–3419, Mar 2009.

[2] A Berchuck, E S Iversen, J M Lancaster, H K Dressman, M West, J R Nevins, and J R Marks. Prediction of optimal versus suboptimal cytoreduction of advanced-stage serous ovarian cancer with the use of microarrays. *Am J Obstet Gynecol*, 190(4):910–925, Apr 2004.

[3] T Bonome, D A Levine, J Shih, M Randonovich, C A Pise-Masison, F Bogomolniy, L Ozbun, J Brady, J C Barrett, J Boyd, and M J Birrer. A gene signature predicting for survival in suboptimally debulked patients with ovarian cancer. *Cancer Res*, 68(13):5478–5486, Jul 2008.

[4] R Breitling, P Armengaud, A Amtmann, and P Herzyk. Rank products: a simple, yet powerful, new method to detect differentially regulated genes in replicated microarray experiments. *FEBS Lett*, 573(1-3):83–92, Aug 2004.

[5] Cancer Genome Atlas Research Network. Integrated genomic analyses of ovarian carcinoma. *Nature*, 474(7353):609–615, Jun 2011.

[6] X Z Chang, J Yu, H Y Liu, R H Dong, and X C Cao. ARK5 is associated with the invasive and metastatic potential of human breast cancer cells. *J Cancer Res Clin Oncol*, 138(2):247–254, Feb 2012.

[7] P Cirri and P Chiarugi. Cancer-associated-fibroblasts and tumour cells: a diabolic liaison driving cancer progression. *Cancer Metastasis Rev*, 31(1-2):195–208, Jun 2012.

[8] E Devarajan, Y H Song, S Krishnappa, and E Alt. Epithelial-mesenchymal transition in breast cancer lines is mediated through PDGF-D released by tissue-resident stem cells. *Int J Cancer*, 131(5):1023–1031, Sep 2012.

[9] H K Dressman, A Berchuck, G Chan, J Zhai, A Bild, R Sayer, J Cragun, J Clarke, R S Whitaker, L Li, J Gray, J Marks, G S Ginsburg, A Potti, M West, J R Nevins, and J M Lancaster. An integrated genomic-based approach to individualized treatment of patients with advanced-stage ovarian cancer. *J Clin Oncol*, 25(5):517–525, Feb 2007.

[10] Debashis Ghosh and Hyungwon Choi. *metaArray: Integration of Microarray Data for Meta-analysis*. R package version 1.38.0.

[11] P M Grambsch and T M Therneau. Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, 81(3):515–526, 1994.

[12] LV Hedges and I Olkin. *Statistical Methods for Meta-Analysis*. Academic Press, San Diego, CA, 1985.

[13] K Horiuchi, N Amizuka, S Takeshita, H Takamatsu, M Katsuura, H Ozawa, Y Toyama, L F Bonewald, and A Kudo. Identification and characterization of a novel protein, periostin, with restricted expression to periosteum and periodontal ligament and increased expression by transforming growth factor beta. *J Bone Miner Res*, 14(7):1239–1249, Jul 1999.

[14] Y Li, M Y Maitah, A Ahmad, D Kong, B Bao, and F H Sarkar. Targeting the hedgehog signaling pathway for cancer therapy. *Expert Opin Ther Targets*, 16(1):49–66, Jan 2012.

[15] I Malanchi, A Santamaria-Martínez, E Susanto, H Peng, H A Lehr, J F Delaloye, and J Huelsken. Interactions between cancer stem cells and their niche govern metastatic colonization. *Nature*, 481(7379):85–89, Jan 2012.

[16] L Morra and H Moch. Periostin expression and epithelial-mesenchymal transition in cancer: a review and an update. *Virchows Arch*, 459(5):465–475, Nov 2011.

[17] G Parmigiani, E S Garrett-Mayer, R Anbazhagan, and E Gabrielson. A cross-study comparison of gene expression studies for the molecular classification of lung cancer. *Clin Cancer Res*, 10(9):2922–2927, May 2004.

[18] Margaret Sullivan Pepe, Kathleen F. Kerr, Gary Longton, and Zheyu Wang. Testing for improvement in prediction model performance. *Statistics in medicine*, 32(9):1467–1482, Apr 2013.

[19] L Rosanò, R Cianfrocca, F Spinella, V Di Castro, M R Nicotra, A Lucidi, G Ferrandina, P G Natali, and A Bagnato. Acquisition of chemoresistance and EMT phenotype is linked with activation of the endothelin a receptor pathway in ovarian carcinoma cells. *Clin Cancer Res*, 17(8):2350–2360, Apr 2011.

[20] L Rosanò, F Spinella, and A Bagnato. The importance of endothelin axis in initiation, progression, and therapy of ovarian cancer. *Am J Physiol Regul Integr Comp Physiol*, 299(2):395–404, Aug 2010.

[21] C S Schuetz, M Bonin, S E Clare, K Nieselt, K Sotlar, M Walter, T Fehm, E Solomayer, O Riess, D Wallwiener, R Kurek, and H J Neubauer. Progression-specific genes identified by expression profiling of matched ductal carcinomas in situ and invasive breast tumors, combining laser capture microdissection and oligonucleotide microarray analysis. *Cancer Res*, 66(10):5278–5286, May 2006.

[22] R L Seal, S M Gordon, M J Lush, M W Wright, and E A Bruford. genenames.org: the HGNC resources in 2011. *Nucleic Acids Res*, 39(Database issue):514–519, Jan 2011.

[23] J Soikkeli, P Podlasz, M Yin, P Nummela, T Jahkola, S Virolainen, L Krogerus, P Heikkilä, K von Smitten, O Saksela, and E Hölttä. Metastatic outgrowth encompasses COL-I, FN1, and POSTN up-regulation and assembly to fibrillar networks regulating cell adhesion, migration, and growth. *Am J Pathol*, 177(1):387–403, Jul 2010.

[24] A Suzuki, G Kusakai, Y Shimojo, J Chen, T Ogura, M Kobayashi, and H Esumi. Involvement of transforming growth factor-beta 1 signaling in hypoxia-induced tolerance to glucose starvation. *J Biol Chem*, 280(36):31557–31563, Sep 2005.

[25] R W Tothill, A V Tinker, J George, R Brown, S B Fox, S Lade, D S Johnson, M K Trivett, D Etemadmoghadam, B Locandro, N Traficante, S Fereday, J A Hung, Y E Chiew, I Haviv, Australian Ovarian Cancer Study Group, D Gertig, A DeFazio, and D D Bowtell. Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. *Clin Cancer Res*, 14(16):5198–5208, Aug 2008.

[26] R G Verhaak, P Tamayo, J Y Yang, D Hubbard, H Zhang, C J Creighton, S Fereday, M Lawrence, S L Carter, C H Mermel, A D Kostic, D Etemadmoghadam, G Saksena, K Cibulskis, S Duraisamy, K Levanon, C Sougnez, A Tsherniak, S Gomez, R Onofrio, S Gabriel, L Chin, N Zhang, P T Spellman, Y Zhang, R Akbani, K A Hoadley, A Kahn, M Köbel, D Huntsman, R A Soslow, A Defazio, M J Birrer, J W Gray, J N Weinstein, D D Bowtell, R Drapkin, J P Mesirov, G Getz, D A Levine, and M Meyerson. Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. *J Clin Invest*, 123(1):517–525, Jan 2013.

[27] Viechtbauer W. Bias and Efficiency of Meta-Analytic Variance Estimators in the Random-Effects Model. *Journal of Educational and Behavioral Statistics*, 30(3):261–293, 2005.

[28] L Waldron, B Haibe-Kains, AC Culhane, M Riester, J Ding, XV Wang, M Ahmadifar, S Tyekucheva, C Bernau, T Risch, B Ganzfried, C Huttenhower, M Birrer, and Parmigiani G. A comparative meta-analysis of prognostic gene signatures for late-stage ovarian cancer. *JNCI*, 2014.

[29] M N Wente, C Mayer, M M Gaida, C W Michalski, T Giese, F Bergmann, N A Giese, M W Büchler, and H Friess. CXCL14 expression and potential function in pancreatic cancer. *Cancer Lett*, 259(2):209–217, Feb 2008.

[30] Q Wu, X Hou, J Xia, X Qian, L Miele, F H Sarkar, and Z Wang. Emerging roles of PDGF-D in EMT progression during tumorigenesis. *Cancer Treat Rev*, 39(6):640–646, Oct 2013.

[31] Xiaogang Zhong, Leslie Cope, Elizabeth Garrett, and Giovanni Parmigiani. *MergeMaid: Merge Maid*, 2007. R package version 2.32.0.

[32] M Zhu, R E Saxton, L Ramos, D D Chang, B Y Karlan, J C Gasson, and D J Slamon. Neutralizing monoclonal antibody to periostin inhibits ovarian tumor growth and metastasis. *Mol Cancer Ther*, 10(8):1500–1508, Aug 2011.

# A Session Info

- R version 3.0.2 (2013-09-25), `x86_64-apple-darwin10.8.0`
- Locale: `C`
- Base packages: base, datasets, grDevices, graphics, grid, methods, parallel, splines, stats, utils
- Other packages: AnnotationDbi 1.24.0, Biobase 2.22.0, BiocGenerics 0.8.0, DBI 0.2-7, Formula 1.1-1, GSEABase 1.24.0, GSVA 1.10.1, Hmisc 3.13-0, KernSmooth 2.23-10, LeviRmisc 0.16.4, MASS 7.3-29, Matrix 1.1-1.1, MergeMaid 2.34.0, RColorBrewer 1.0-5, ROCR 1.0-5, RSQLite 0.11.4, RankProd 2.34.0, affy 1.40.0, annotate 1.40.0, bioDist 1.34.0, biomaRt 2.18.0, caTools 1.16, car 2.0-19, cluster 1.14.4, codetools 0.2-8, curatedOvarianData 1.0.5, cvTools 0.3.2, exactRankTests 0.8-27, ez 4.2-2, gdata 2.13.2, genefilter 1.44.0, ggplot2 0.9.3.1, gplots 2.12.1, graph 1.40.1, gtools 3.1.1, hgu133a.db 2.10.1, impute 1.36.0, knitr 1.5, lattice 0.20-24, limma 3.18.7, lme4 1.0-5, lpSolve 5.6.7, maxstat 0.7-18, mclust 4.2, memoise 0.1, metaArray 1.40.0, metafor 1.9-2, mgcv 1.7-27, mvtnorm 0.9-9996, nlme 3.1-113, nnet 7.3-7, org.Hs.eg.db 2.10.1, pROC 1.6.0.1, penalized 0.9-42, plyr 1.8, prodlim 1.3.7, qvalue 1.36.0, reshape2 1.2.2, rmeta 2.16, robustbase 0.9-10, sampling 2.6, scales 0.2.3, stringr 0.6.2, survC1 1.0-2, survHD 0.5.0, survIDINRI 1.1-1, survcomp 1.12.0, survival 2.37-4, texreg 1.30, xtable 1.7-1
- Loaded via a namespace (and not attached): BiocInstaller 1.12.0, Biostrings 2.30.1, GEOquery 2.28.0, GenomicRanges 1.14.4, IRanges 1.20.6, RCurl 1.95-4.1, Rcpp 0.10.6, SuppDists 1.1-9.1, XML 3.95-0.2, XVector 0.2.0, affyio 1.30.0, bitops 1.0-6, bootstrap 2012.04-1, colorspace 1.2-4, dichromat 2.0-0, digest 0.6.4, evaluate 0.5.1, formatR 0.10, genomes 2.8.0, gtable 0.1.2, highr 0.3, labeling 0.2, minqa 1.2.2, multtest 2.18.0, munsell 0.4.2, preprocessCore 1.24.0, proto 0.3-10, stats4 3.0.2, survivalROC 1.0.3, tcltk 3.0.2, tools 3.0.2, zlibbioc 1.8.0