# Supplementary Information

## Consensus on Molecular Subtypes of High-Grade Serous Ovarian Cancer

# Contents

**1 Datasets**

**2 Reproduction of Published HGSOC Subtype Classifiers**
- 2.1 Konecny et al., 2014
- 2.2 Verhaak et al., 2013 / TCGA 2011
- 2.3 Helland et al., 2011 / Tothill et al., 2008

**3 Reproduction of Subtype Clustering Methods**
- 3.1 TCGA
- 3.2 Konecny
- 3.3 Tothill

**4 Robustness of Subtypes**
- 4.1 Method: Prediction Strength
- 4.2 Results: Prediction Strength

**5 Pairwise Subtype Classifier Associations**
- 5.1 Helland vs Verhaak
- 5.2 Konecny vs Verhaak
- 5.3 Helland vs Konecny

**6 Survival Analysis**

# 1 Datasets

We analyzed data from the curatedOvarianData compendium and additional datasets from MetaGx-Ovarian, which composed of 15 whole-transcriptome datasets. We limited our analysis to tumors annotated as high-grade, late-stage serous ovarian cancers.

Table 1: Supplementary Table 1: 15 whole-transcriptome studies with at least 40 patients with late stage, high-grade serous histology from the curatedOvarianData compendium consisting of 1774 patients. 13 of these datasets provided 1581 patients with survival data.

| GEO Accession ID | Total number of samples, Number of samples with survival data: deceased (median survival months) | Microarray Platform | Number of Features |
|---|---|---|---|
| TCGA | 464 452: 239 (42.6) | Affymetrix HT Human Genome U133A | 12833 |
| GSE17260 | 43 43: 22 (29) | Agilent-012391 Whole Human Genome Oligo | 19596 |
| GSE14764 | 41 41: 13 (30) | Affymetrix HG-U133A | 12752 |
| GSE18520 | 53 53: 41 (21) | Affymetrix Human Genome U133 Plus 2.0 | 20282 |
| GSE26193 | 47 47: 39 (34) | Affymetrix Human Genome U133 Plus 2.0 | 20282 |
| PMID17290060 | 59 59: 36 (34) | Affymetrix HG-U133A | 12752 |
| GSE51088 | 85 84: 69 (43.6) | Agilent-012097 Human 1A Microarray (V2) G4110B | 15299 |
| GSE13876 | 98 98: 72 (22) | Operon human v3 $\sim$35K 70-mer two-color oligonucleotide microarrays | 13846 |
| GSE49997 | 132 122: 40 (23) | ABI Human Genome Survey Microarray Version | 216760 |
| E.MTAB.386 | 128 128: 73 (29.65) | Illumina humanRef-8 v2.0 expression beadchip | 10572 |
| GSE32062 | 129 129: 60 (40) | Agilent-014850 Whole Human Genome Microarray 4x44K G4112F | 19596 |
| GSE9891 | 142 140: 72 (28.5) | Affymetrix Human Genome U133 Plus 2.0 | 20282 |
| GSE26712 | 185 185: 129 (38.8) | Affymetrix Human Genome U133A Array | 12752 |
| GSE20565 | 89 (0) | Affymetrix Human Genome U133 Plus 2.0 | 20282 |
| GSE2109 | 79 (0) | Affymetrix HG-U133Plus | 220282 |

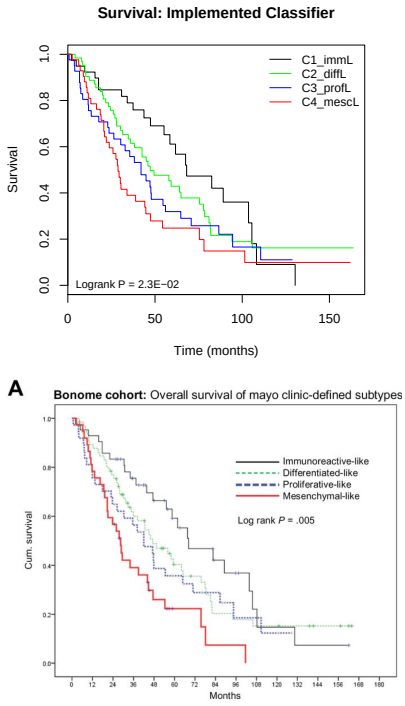
# 2 Reproduction of Published HGSOC Subtype Classifiers

## 2.1 Konecny et al., 2014

First, we implemented the subtype classifier by Konecny et al., 2014. Their subtype classifier uses a nearest-centroids approach with Spearman's correlation coefficient as the distance measure. The authors provided a list of 635 selected probe sets for classifying new cases. To allow cross-platform applicability, we implemented the subtype classifier using the 575 unique Entrez gene IDs corresponding to these probe sets (with the mean value taken for multiple probe sets mapping to the same gene ID). In their supplementary materials, Konecny et al. report their predicted subtypes on a validation dataset (Bonome et al.). To assess our implementation, we compared our predicted subtypes on the Bonome dataset. The contingency matrix and survival curves are given below, indicating a large degree of concordance between our implemented subtypes and the author's supplementary data. Overall, 96.7% of samples were classified identically between our implementation and the supplementary results.

| | Konecny Subtypes from Supplementary | | | |
|---|---|---|---|---|
| Implemented Konecny Subtypes | c1 | c2 | c3 | c4 |
| c1 | 39 | 0 | 0 | 0 |
| c2 | 0 | 63 | 0 | 0 |
| c3 | 0 | 3 | 37 | 0 |
| c4 | 3 | 0 | 0 | 37 |

Table 2: Contingency table showing strong concordance between using our implementation of the Konecny subtyping classifier and the predictions given in the supplementary materials of the Konecny manuscript. These predictions were made on the dataset of Bonome et al.




(Above) Survival curves the Bonome dataset using our implementation of the Konecny subtyping scheme. (Below) Corresponding Figure 3A from Konecny et al.

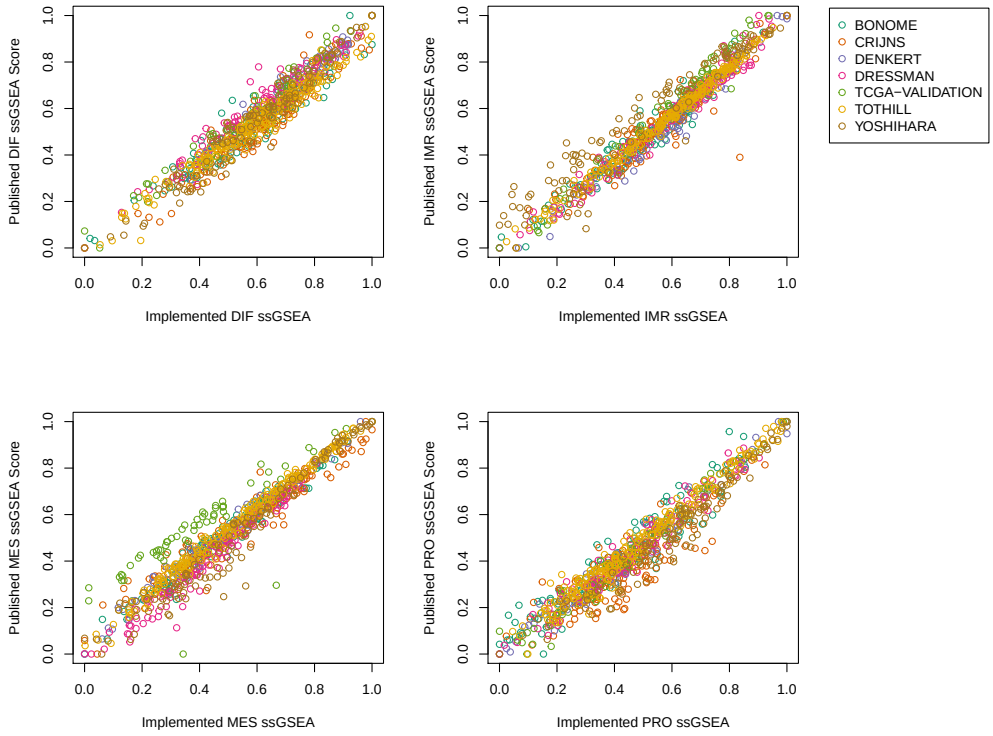
## 2.2 Verhaak et al., 2013 / TCGA 2011

Next, we implemented the subtype classification scheme given by Verhaak et al., 2013. The authors designed a classifier based on single-sample GSEA to classify samples into subtypes previously defined in TCGA, 2011. In their supplementary materials, the authors provide a list of four sets of gene symbols (100 total gene symbols), with each gene set associated with a subtype.

We implemented this subtype classifier using the provided gene sets and the ssGSEA implementation in `R` package `GSVA`. The parameters to the function `gsva` were: `method="ssgsea"`, and `tau=0.75`.

To assess our implementation, we compared our normalized ssGSEA scores with the scores in the validation set used in the original study. In their supplementary materials, Verhaak et al. provide their normalized ssGSEA scores for a validation set consisting of the datasets of Bonome, Crijns, Denkert, Dressman, Tothill, Yoshihara, and a subset of TCGA. This validation dataset consisted of 879 patients reported in their supplementary; we matched 865 patients from MetaGxOvarian data based on sample IDs.

We observed Pearson's correlation coefficients of 0.96, 0.97, 0.96, and 0.96 for subtypes DIF, IMR, MES, and PRO respectively.



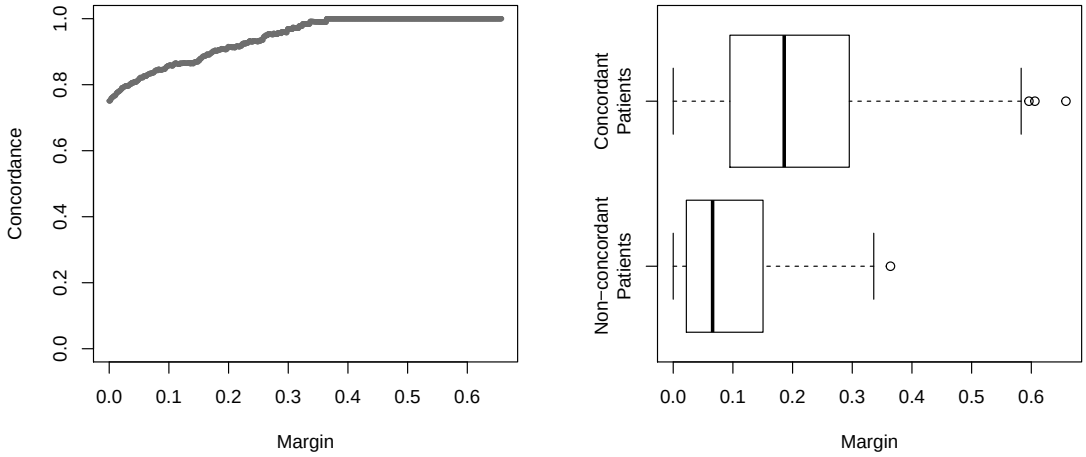

ssGSEA scores for the >800 patient validation set from our implementation (x axis) and the supplementary material (y axis) for each of the four subtypes.

From these normalized ssGSEA scores, subtype classification may be performed by assigning a patient according to the highest ssGSEA score. Overall, this method produces a concordance of 86.59% of patients classified identically between our implementation and supplementary results.

We sought to investigate whether the discrepancies may be explained as patients whose expression profiles are marginal cases, e.g. expression profiles that are similar to two different subtypes. We investigated the relationship between classification accuracy and the margin, defined as the difference between the top-scoring ssGSEA score and second-top scoring ssGSEA score (using the published ssGSEA normalized scores).

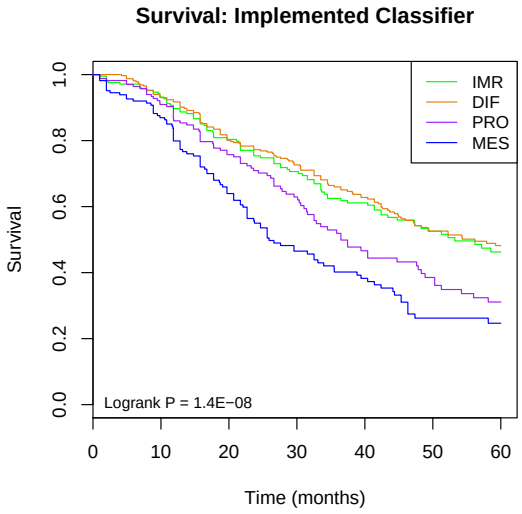
| | Verhaak Subtypes from Supplementary | | | |
|---|---|---|---|---|
| Implemented Verhaak Subtypes | DIF | IMR | MES | PRO |
| DIF | 265 | 28 | 23 | 8 |
| IMR | 7 | 187 | 13 | 3 |
| MES | 2 | 2 | 156 | 4 |
| PRO | 2 | 13 | 11 | 141 |

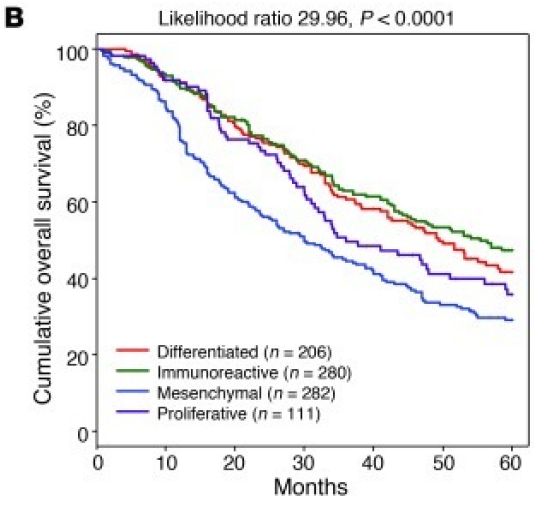
Table 3: Contingency table showing concordance using our implementation of the Verhaak subtyping classifier and the predictions given in the supplementary of the Verhaak manuscript. The predictions for both implementations were made on the combined >800 sample dataset by taking the max ssGSEA subtype score.




Left: the condordance of patients classified above a given margin threshold. Cases with a large difference between top ssGSEA subtype scores are more likely to be classified concordantly. Right: The margin values of patients with concordant vs non-concordant classification between our implementation and published subtypes. Patient classified differently between our implementation and the published subtypes had significantly lower margin values (one-sided Wilcoxon rank-sum p = 1.6E-28).

We performed survival analysis with using the clinical annotations of the validation datasets, and observed similar survival curves as figure 2B in Verhaak et al.




The survival curves produced by our implemented classifier (top) appear similar to Figure 2B in the original publication (bottom).

## 2.3 Helland et al., 2011 / Tothill et al., 2008

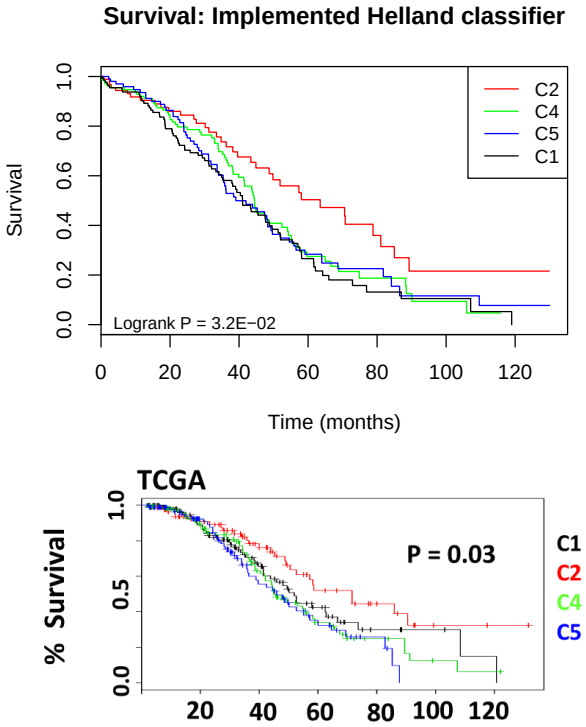
Next, we implemented the subtype classifier of Helland et al., 2011. The same group as the Tothill et al. study implemented a different classifier for their previously-described subtypes. They identified a gene list for each of their four previously-defined high-grade serous ovarian carcinoma subtypes. Using a method described in another study for breast cancer classification (Lim et al., Nat. Med. 2009), they trained a set of weights for each gene list. Classification was performed by taking a linear combination of weights and expression levels for each gene list, normalizing the scores, and classifying according to the highest-scoring subtype.

Using their published gene list and weights from the supplementary text, we implemented their subtype classifier and applied it to the TCGA dataset. The authors kindly provided a spreadsheet listing their classifier's labels on the TCGA dataset. Overall, 93.35% of samples were classified identically between the authors' implementation and ours.

| | Original Helland Subtypes | | | |
|---|---|---|---|---|
| Implemented Helland Subtypes | C1 | C2 | C4 | C5 |
| C1 | 125 | 2 | 3 | 4 |
| C2 | 1 | 87 | 4 | 0 |
| C4 | 1 | 10 | 122 | 1 |
| C5 | 2 | 2 | 1 | 101 |

Table 4: Contingency table showing concordance of our implementation and and the predictions given by the table provided by Helland et al. Predictions were made on the TCGA dataset. Note that subtypes C3 and C6 were excluded in the original study since they are associated with non-HGS ovarian tumours.

We performed survival analysis on the TCGA dataset.



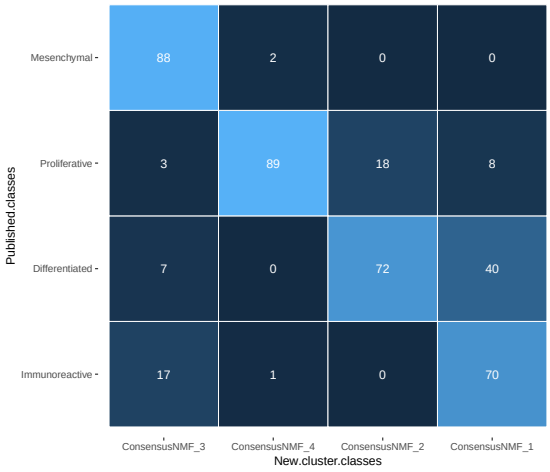

(Above) Survival curves the TCGA dataset using our implementation of the Helland subtyping scheme.
(Below) Corresponding survival plot from Figure 1B from Helland et al.

# 3 Reproduction of Subtype Clustering Methods

## 3.1 TCGA

We contacted the authors to acquire the original gene list used for clustering. Using `R` package `NMF`, we ran NMF with 100 iterations, and used hierarchical clustering on the co-membership matrix to define consensus cluster groupings.

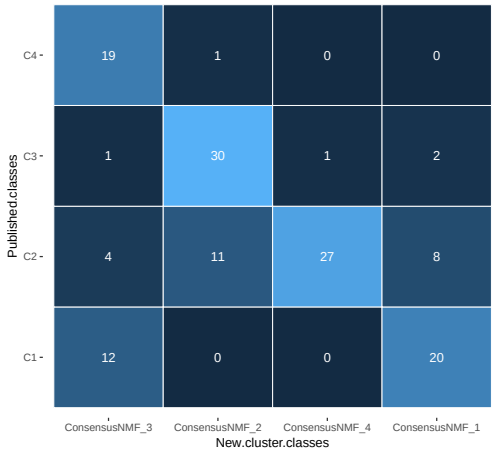
| | ConsensusNMF_3 | ConsensusNMF_4 | ConsensusNMF_2 | ConsensusNMF_1 |
|---|---|---|---|---|
| Mesenchymal | 88 | 2 | 0 | 0 |
| Proliferative | 3 | 89 | 18 | 8 |
| Differentiated | 7 | 0 | 72 | 40 |
| Immunoreactive | 17 | 1 | 0 | 70 |



## 3.2 Konecny

Konecny et al. performed clustering by first taking a subset of the top 2500 probesets by median absolute deviation (MAD), then used non-negative matrix factorization. We matched these probeset names to Entrez IDs from data from GEO, and performed clustering on the series matrix from GEO. In order to ensure all expression values were positive, all expression quantities were increased by the absolute value of the smallest (most negative) value. We ran NMF with 100 iterations, and used hierarchical clustering on the co-membership matrix to define consensus cluster groupings.

| | ConsensusNMF_3 | ConsensusNMF_2 | ConsensusNMF_4 | ConsensusNMF_1 |
|---|---|---|---|---|
| C4 | 19 | 1 | 0 | 0 |
| C3 | 1 | 30 | 1 | 2 |
| C2 | 4 | 11 | 27 | 8 |
| C1 | 12 | 0 | 0 | 20 |



## 3.3 Tothill

The dataset of Tothill et al. (2008) consisted of 285 patients, of which 142 had late-stage, high-grade serious ovarian cancer. On their full dataset (n = 285), they performed the following clustering procedure: probes were selected if at least one sample had an expression above 7.0, and global variance was above 0.5. They performed a form of consensus k-means clustering by performing k-means clustering 1000 times, identifyied a "robust" set of samples that co-clustered consistently, then used diagonal LDA and k-nearest neighbours to classify remaining samples.

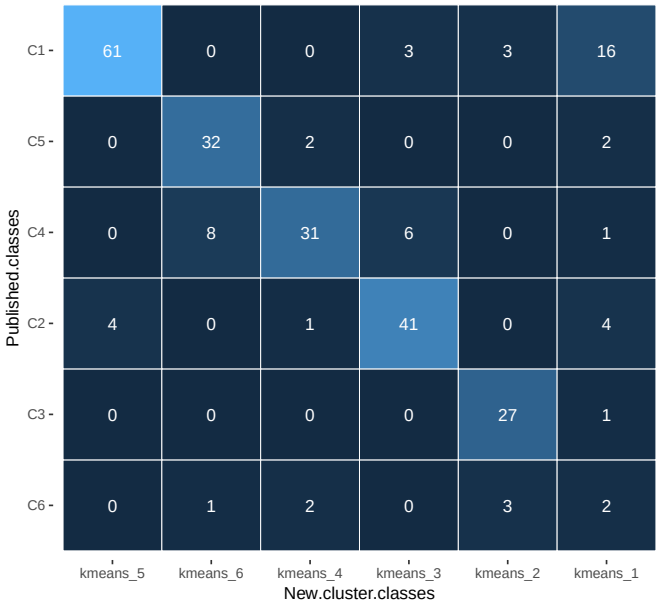
Using their full dataset, we filtered to probes with at least one sample with expression above 7.0, and variance above 0.5. We performed k-means clustering $B = 1000$ times, and used `R` package `clue` to perform consensus clustering using the criterion of Dimitriadou et al. (2002):

$$C_{\text{consensus}} = \min_{C \in \mathcal{C}} \left\{ \sum_{b=1}^{B} d(C, C_b)^2 \right\}$$

where $\mathcal{C}$ is the set of all possible clusterings, $d$ is the Euclidean distance, and $\{C_1, C_2, \ldots, C_B\}$ are the k-means clusterings.

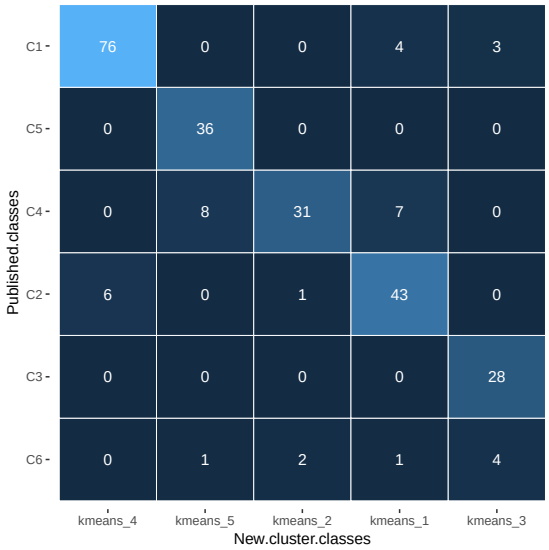
Despite the difference between the consensus strategies, the cluster results appear to be similar:

| | kmeans_5 | kmeans_6 | kmeans_4 | kmeans_3 | kmeans_2 | kmeans_1 |
|---|---|---|---|---|---|---|
| C1 | 61 | 0 | 0 | 3 | 3 | 16 |
| C5 | 0 | 32 | 2 | 0 | 0 | 2 |
| C4 | 0 | 8 | 31 | 6 | 0 | 1 |
| C2 | 4 | 0 | 1 | 41 | 0 | 4 |
| C3 | 0 | 0 | 0 | 0 | 27 | 1 |
| C6 | 0 | 1 | 2 | 0 | 3 | 2 |



Since it appears that this implementation of k-means clustering is not capturing C6, we performed consensus k-means clustering with k = 5:

| | kmeans_4 | kmeans_5 | kmeans_2 | kmeans_1 | kmeans_3 |
|---|---|---|---|---|---|
| C1 | 76 | 0 | 0 | 4 | 3 |
| C5 | 0 | 36 | 0 | 0 | 0 |
| C4 | 0 | 8 | 31 | 7 | 0 |
| C2 | 6 | 0 | 1 | 43 | 0 |
| C3 | 0 | 0 | 0 | 0 | 28 |
| C6 | 0 | 1 | 2 | 1 | 4 |



# 4 Robustness of Subtypes

## 4.1 Method: Prediction Strength

The discovery of molecular subtypes of HGSOC requires two steps: a clustering step, and a classification step. We sought to address the question of whether the methods used to define molecular subtypes are robust. In cluster analysis, robustness is a measure of how Tibshirani and Walther (2005) suggest Prediction Strength as a statistic for cluster validation. Prediction Strength is a measure of the similarity between pairwise co-memberships of a validation dataset from class labels assigned by (1) a clustering algorithm and (2) a classification algorithm trained on another dataset.

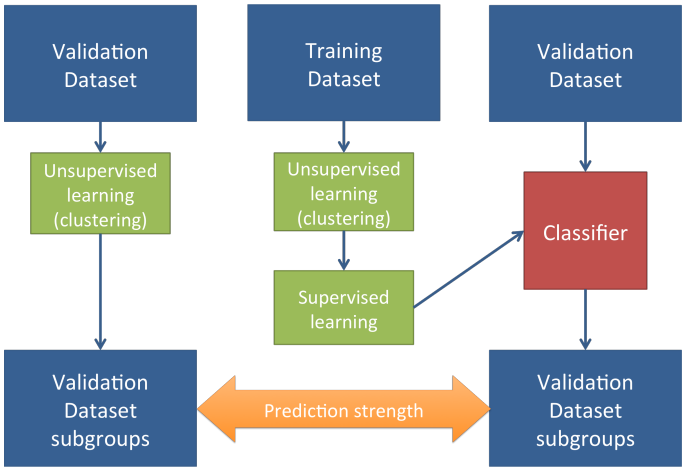



Figure 1: Flowchart of the Prediction Strength statistic.

## 4.2 Results: Prediction Strength

Each dataset was clustered according to our implementation of the clustering algorithms and gene sets of Konecny, TCGA, and Tothill. Each dataset was also classified using our implementation of the corresponding classification algorithms of Konecny, TCGA/Verhaak, and Tothill/Helland. This produced two sets of subtype labels for each validation dataset, from which we computed prediction strength.

We performed each clustering algorithm 100 times for each dataset, producing 100 prediction strength estimates per dataset. In the boxplot below, each data point represents the mean estimated prediction strength for a dataset under a given subtyping clustering/classification scheme.

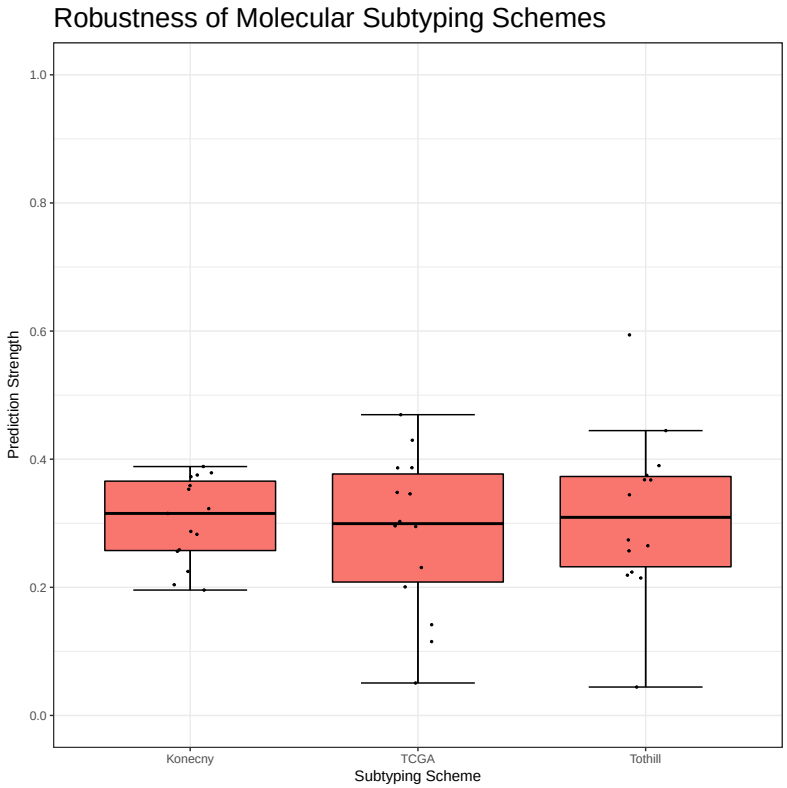



Figure 2: Prediction Strength of each clustering/classification algorithm pair across datasets.

# 5 Pairwise Subtype Classifier Associations

## 5.1 Helland vs Verhaak

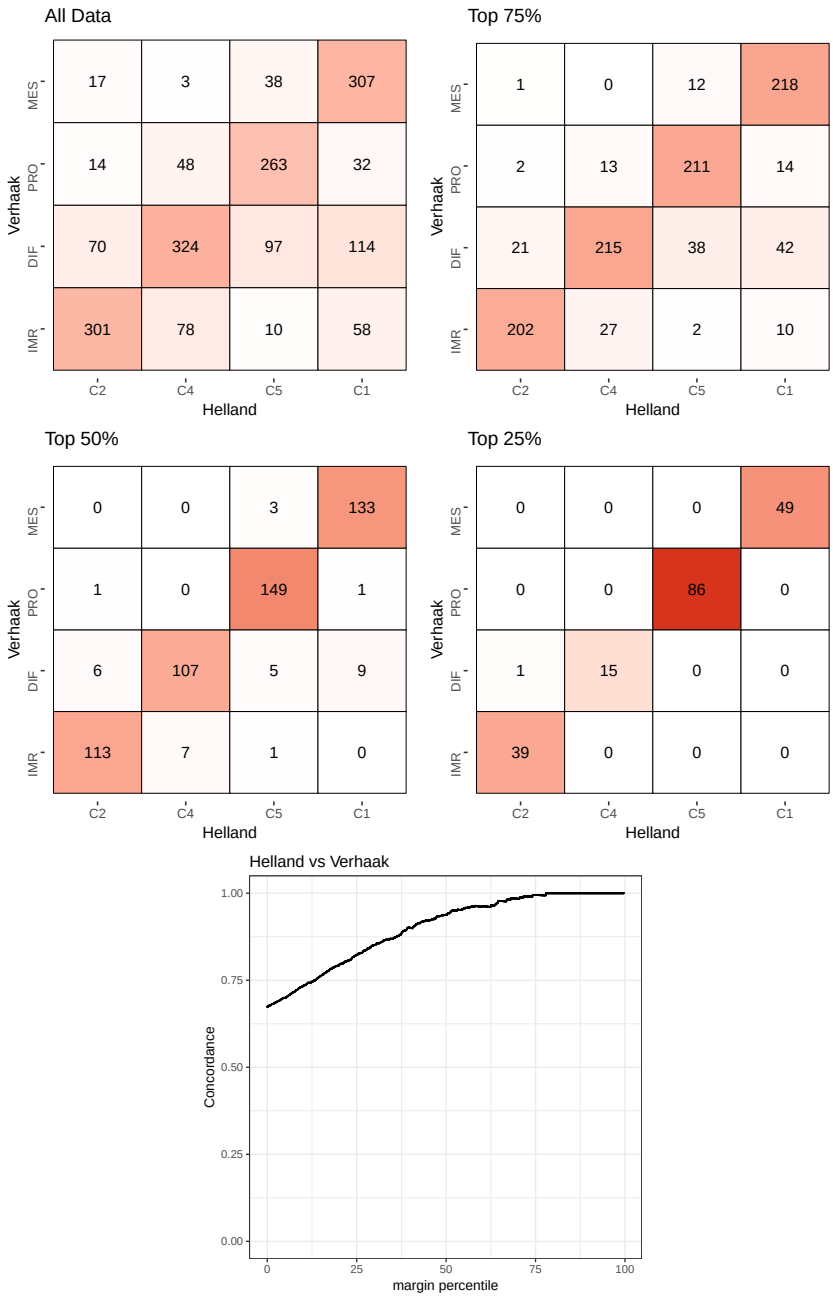



Figure 3: Pairwise subtype association between the Helland and Verhaak classifiers. Each classifier produces a real-valued subtype score per patient; from this score, a margin value can be defined as the difference between the top two subtype scores. We assessed subtype association, considering only patients for whom the margin values are in the top 75%, top 50%, and top 25% of both classifiers. Using these margin value cutoffs, we observed an increase in between-classifier concordance.

## 5.2 Konecny vs Verhaak

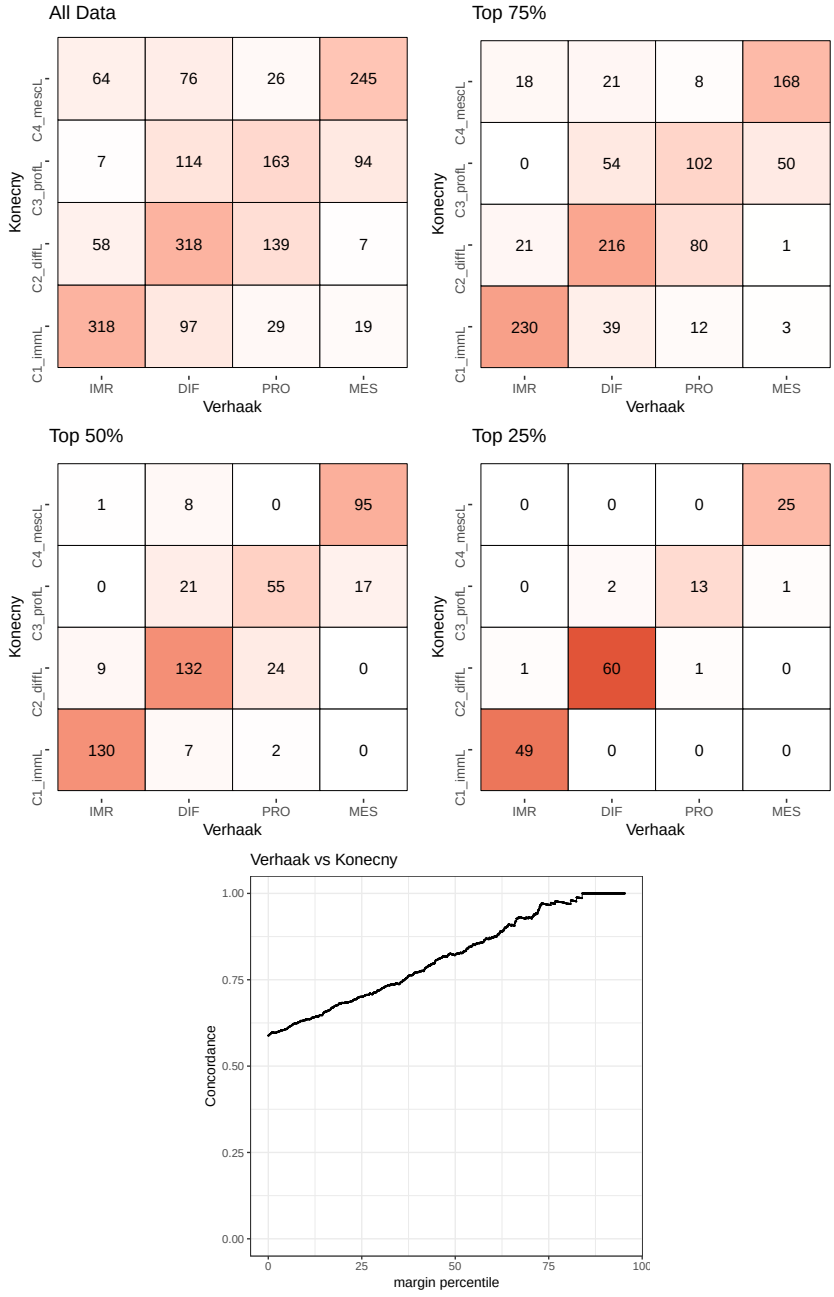



Figure 4: Pairwise subtype association between the Konecny and Verhaak classifiers. Each classifier produces a real-valued subtype score per patient; from this score, a margin value can be defined as the difference between the top two subtype scores. We assessed subtype association, considering only patients for whom the margin values are in the top 75%, top 50%, and top 25% of both classifiers. Using these margin value cutoffs, we observed an increase in between-classifier concordance.

## 5.3 Helland vs Konecny

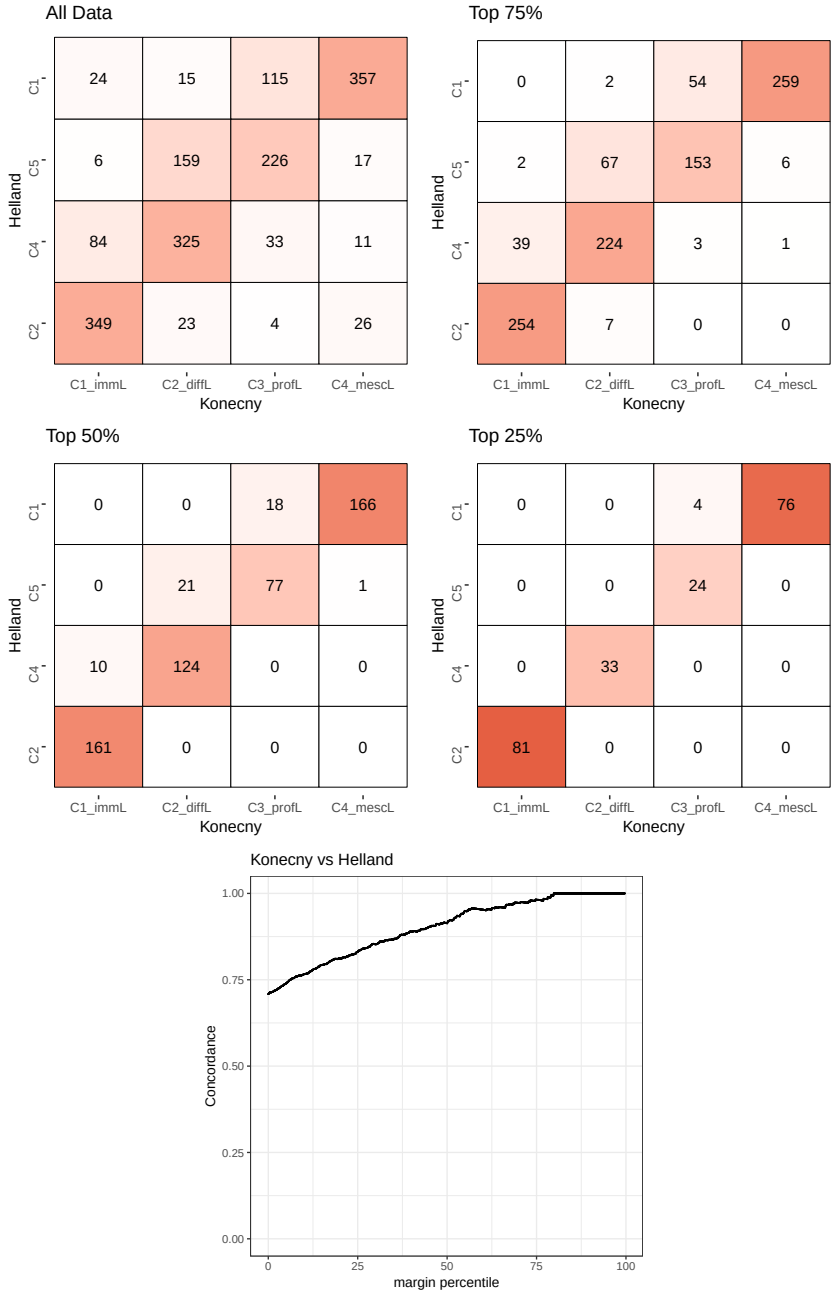



Figure 5: Pairwise subtype association between the Helland and Konecny classifiers. Each classifier produces a real-valued subtype score per patient; from this score, a margin value can be defined as the difference between the top two subtype scores. We assessed subtype association, considering only patients for whom the margin values are in the top 75%, top 50%, and top 25% of both classifiers. Using these margin value cutoffs, we observed an increase in between-classifier concordance.

# 6 Survival Analysis