SUPPLEMENTARY METHODS AND RESULTS

TABLE OF CONTENTS

A- OTTA Consortium Sample Description and Processing *3*

A.1 Sample Selection 3

A.2 Pathology Review 4

A.3 Additional Clinicopathological Data 5

A.4 NanoString Gene Selection 6

A.5 RNA Extractions 7

A.6 NanoString Processing 8

A.7 Quality Assurance, Batch Effect, and Normalization of the NanoString Data 9

Batch correction using control pools 11

Cross-site Controls 13

B: Subtype Labels Assignment to NanoString Data *15*

B.1 Transcriptome-Wide Data Inclusion 15

B.2 Two Independent Approaches 16

B.2.2 TCGA Approach 28

B.3 Platform Portability of the Array Classifier 30

B.4 Subtype Assignment in the NanoString data 32

Comparison of Consensus vs. Non-Consensus for Technical and Biological Variables 35

C: Development of a minimal predictive model on NanoString Data *42*

C.1 Data Breakdown for Downstream Analysis 42

C.2 Minimal Model Development 42

C.3 First Confirmation 48

C.4 Validation 49

C.5 Genes in the classifier 53

C.6 Gene Expression Distribution on Top Genes 55

D: Biological Correlates of Molecular Subtypes 58

D.1 Technical Variability and Potential Sources of Bias 58

D.1.1 Within CodeSet Replication 59

D.1.2 Cross-site Replication 60

D.1.3 Cross CodeSet Replication 60

D.2 Anatomical Site Variability 63

D.3 Biological Characterization of the Subtypes 64

D.3.1 Correlation with Clinical and Pathological Parameters 65

D.3.2 CD8 Tumor Infiltrating Lymphocyte Analysis 69

D.3.3 Survival analysis 71

E: Recommended Standard Operating Procedures for the Predictor of high-grade serous Ovarian carcinoma molecular subTYPE (PrOTYPE) 82

Starting material 82

RNA Extraction: 82

Deparaffinization of archival specimens: 82

Modified miRNeasy FFPE Protocol 83

NanoString Hybridization, Processing and Scanning 84

CDF File Setup and Naming Conventions: 86

References for Supplemental Appendix 88

Note: Larger supplemental tables are separate from this file

- *Table SA1*
- *Table SB1-SB5*
- *Table SC1 and SC7*

A- OTTA Consortium Sample Description and Processing

A.1 Sample Selection

Twenty participating studies across the Ovarian Tumor Tissue Analysis (OTTA) consortium from the UK, Europe, Australia, Canada and USA (Table SA1 and Figure SA1) contributed 4071 unique high-grade serous tubo-ovarian carcinoma (HGSOC) specimens for NanoString gene expression analysis (Tables SA1). The total number of samples processed was based on availability of HGSOC specimens from contributing OTTA consortia members and restricted to the available funding envelop for purchase of NanoString reagents. Data can be found under NCBI GEO dataset *[available on request]*

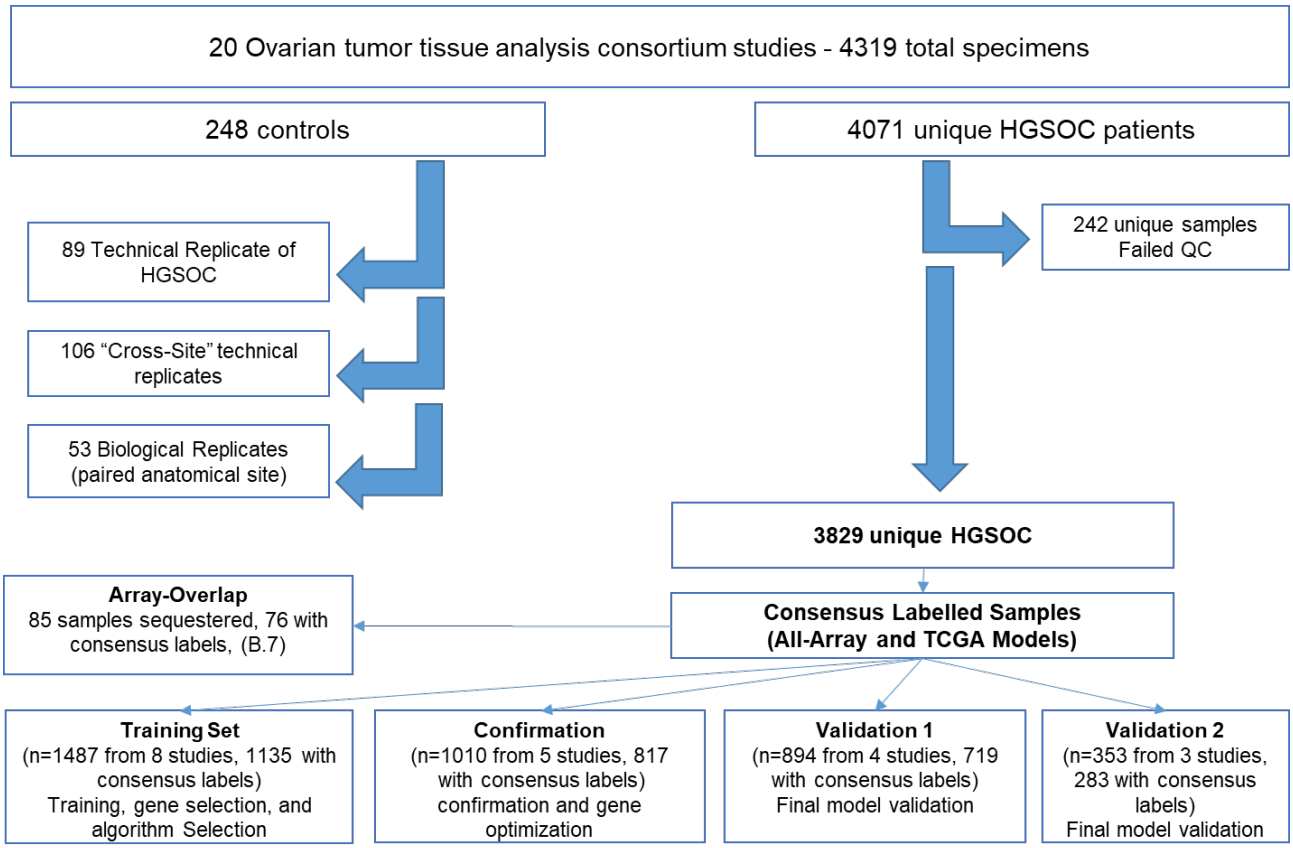

Figure SA1. Overview of processing of samples in from the ovarian tumor tissue analysis (OTTA) consortium. A total of 4319 specimens were run excluding reference pools. From this there were 4071 unique samples, 3829 of which passed quality control measures and were used for refinement of a minimal gene classifier and inference of biological associations (discussed in Supplement C). In addition, 248 specimens served as controls to specifically examine performance, reproducibility, and heterogeneity (in the context of biological replicates). Derivation of the minimal gene classifier is discussed below, whereas technical controls, biological replicates, heterogeneity and biological covariates are discussed in further detail in Supplement C. In addition to the samples represented in the workflow presented here, a subset of the unique samples was further replicated in two independently synthesized NanoString CodeSets and is discussed in further detail in section C.1.

A.2 Pathology Review

Each study provided, at minimum, one representative H&E slide for centralized pathology review and an archival formalin-fixed paraffin embedded tissue block for each case. Original (study provided) diagnosis was accepted if consistent with modern classification of HGSOC, for instance archival specimen of high-grade endometrioid were considered for inclusion as HGSOC - subject to pathologist review.(1,2) For 12% (468/3829 unique cases post-QC) of cases in sub-studies AOV, SEA, RTR and BRO (Table SA1), an H&E slide and cores from a tumor-enriched region of the matching FFPE tissue block was provided.

Inclusion/Exclusion criteria were samples from primary ovarian, fallopian tube, or primary peritoneal carcinoma; recurrences were excluded. Pathologist confirmed diagnosis of HGSOC, cellularity >20%. Information on interval debulking (i.e. neo-adjuvant chemotherapy; NACT) was limited, however given the span of years where most samples were collected it was presumed that few samples may have had NACT. That said, samples with pathology review suggesting chemotherapy effect consistent with NACT were excluded. No exclusions were done based on patient characteristics (age, stage, BRCA status). Contributing studies were asked to give preference to ovary, fallopian tube, or adnexal sites for specimens (when specifically recorded). We excluded distant metastasis (except as noted specifically re: omentum), and ascites.

Three processing centres conducted pathology review and subsequent experiments: (1) BC Cancer/University of British Columbia (BCCRC/UBC) in Vancouver, Canada; (2) University of Southern California (USC) in Los Angeles, USA; and (3) the Peter MacCallum Cancer Centre (PMC) in Melbourne, Australia. Each processing centre conducted pathology review with an expert gynecologic pathologist:

BC Cancer/UBC: Dr. Tayyebeh Mehrane Nazeran and Dr. C. Blake Gilks

USC: Dr Wafaa Elatre

PMC: Dr Mila Volchek

Pathologists were not blinded to the original diagnosis and were asked to judge whether specimens were consistent, based on morphology alone, with a diagnosis of HGSOC in the context of the 2014 WHO standards. (3) Pathologists also estimated the tumor cellularity and necrosis of each specimen, and marked slides for macrodissection-based enrichment (see A.4). This expert pathology review was designated as *gold standard* for histotype diagnosis, and any sample not judged to be HGSOC was omitted. At BCCRC/UBC, the centre that processed roughly half of samples, a two-level pathology review process was adopted: any sample that did not agree with its original diagnosis on first review (TMN) was re-reviewed by a second expert gyne-pathologist (CBG). In addition, samples that were part

of the NCO study (Table SA1) were reviewed by an expert gyne-pathologist at Duke University (Author: Rex Bentley) prior to being transferred to Vancouver, and therefore subject to two (sometimes three) rounds of expert review. Relevant variables including originating study and specimen ID, reviewing pathologist (and secondary pathologist where applicable), diagnosis, cellularity, and necrosis were recorded.

A.3 Additional Clinicopathological Data

We considered the following clinical and pathological variables where available:

- Age at diagnosis
- Stage: using the FIGO system(4) and dichotomized into low (stage 1 and 2) vs high (stage 3 and 4). The use of historical dataset precluded updating to the latest version of FIGO staging, however this will not impact the dichotomized data (3,5).
- Residual disease: dichotomized into no visible macroscopic disease vs any level of residual. In our conservative analysis, the latter category also included instances previously considered "optimal", and recorded in the OTTA consortium database as "< 1cm" (6). This limitation in our data annotation may have resulted in some patients that genuinely had no visible residual disease being included in the "any" level category.
- Tumor cellularity and necrosis in the NanoString-assayed specimen as determined during pathology review of a serial H&E section (see B.1).
- BRCA1/BRCA2 germline deleterious mutation status obtained from study specific clinical records and research-based testing. Methods varied depending on study and year of testing and were subjected to local verification (per noted citations) so as to ensure adequate distinction between known pathogenic germline mutation carriers and non-carriers. (7-9) (6,10)
- Race/ethnicity: as self-reported by women in participating studies.
- $CD8^+$ tumor infiltrating lymphocyte (TIL) counts: TIL was defined as the infiltrating lymphocytes within the ***epithelial compartment of tumors only***, derived from a previous reported (11). This measure assigned patients to categories based on the average number of $CD8^+$ TILs per high-powered field in tissue microarray cores: negative (none), low (1-2), moderate (3-19), and high (≥20).
- Anatomical site from which the tested HGSOC specimen was acquired. Adenxal specimens included ovary and/or tubal specimens or adnexal without additional detail. Specimens were presumed adnexal based on our request to contributing studies for treatment naïve, primary ovarian or tubal high-grade serous tubo-ovarian carcinoma specimens, but where a specific sampling site was not reported by the contributing study. Additional sites of sampling were accepted if defined: peritoneal specimens included all peritoneal sites, including peritoneal lymph nodes, unless omentum was specifically denoted. In the latter case omentum was specified as the site of sampling. Upper gynecological tract included specimens acquired from uterine and/or cervical sites where the diagnosis was still consistent with primary high-grade serous tubo-ovarian carcinoma. Lower gynecological tract included specimens acquired from the vagina.

Outcome data included overall survival and progression free survival where progression was determined by the treating physician. In both cases, the variable was calculated as time from diagnosis to time of event (death/progression) or censoring. Follow up that exceeded 10 years was left truncated

and right censored as at December 31st of the 10th year post diagnosis to minimize ascertainment bias and ensure non-informative censoring.

A.4 NanoString Gene Selection

513 genes were selected for the NanoString custom CodeSet and an additional 5 genes (*RPL19*, *ACTB*, *PGK1*, *SDHA*, and *POLR1B)*, were included as house-keeping genes for normalisation. 313 candidate genes were added. These were selected based on:

a) Association with prognosis or potential drug targets from the literature
b) residing within a 1 MB region of a potential survival GWAS hit $p<5x10^{-6}$ and showing a survival association in TCGA data.(12)
c) Utility in molecular subtype classification selected from:
    i. Differentially expressed genes by class (one vs all) using the Tothill and TCGA datasets independently, ranked by fold change with a minimum 2-fold change and (Benjamini-Hochberg corrected) p-value <0.01. Differential gene expression analysis using SAM(13) with the top ten highest fold change and lowest fold change genes for each class.
    ii. Genes selected from training and 10-fold cross validation on Tothill and TCGA datasets using linear discriminant analysis and recursive partitioning. Redundancy was introduced by removing genes selected in a first round and repeating the supervised analysis.
    iii. Manual review and curation of genes and pathways from the literature, specifically with reference to molecular subtypes. (12,14-17)
    iv. Genes selected for molecular subtype discrimination using a leave-one-out cross validation strategy as described in Leong *et al*.(18)
d) Ability to evaluate consistency with previous CodeSet analysis. Six additional genes (*TBP, GAPDH, KIF3B, GUSB, BMS1,* and *RPL41)*, were included to evaluate consistency with previous CodeSet analysis(18) but were not used in the normalization.
e) 200 genes selected from a meta-analysis of six transcriptome-wide microarray studies, including 1,516 tumors, to identify prognostic genes (Millstein et al. *submitted*).
f) "tagging" genes that represent the gene expression patterns of other genes that have correlated expression, identified using the methods in Rudd et al(19). For this study, we chose a threshold of 99% correlation observed in all four of the largest publicly-available HGSOC gene expression datasets(12,15,20). We determined that the genes selected (from *a-e* above) included 99%- correlated gene expression information for an additional 2,617 genes. We supplemented an additional 10 genes in order to maximize gene expression data representation in other parts of the transcriptome that were not already represented by previous selections. The 10 additional genes represent gene expression for 49 genes at our 99% correlation threshold.

Gene lists were combined for a total of 518 unique transcripts after removing redundancy, and including controls (housekeeping genes). Note: Of gene added from literature review (a), GWAS hits (b), and meta-analysis (e), 181 genes were unique from subtype selections noted above (c and d), however 89/181 were also significantly differentially expressed between subtypes (as defined above in "c) i" though not top ranking).

A.5 RNA Extractions

Standard operating procedures were established in advance and implemented at all sites performing sample preparation and NanoString reactions.

Tumor Content enrichment

Pathology reviewed, formalin-fixed, paraffin embedded (FFPE) tissue blocks without adjacent normal tissue (normal, non-tumor involved organ tissues, *BUT NOT* tumor associated stroma or tumor infiltrated tissues) or where normal tissue was less than ~10% were scrolled directly into sterile microfuge tubes, 3x 10um scrolls, without further enrichment.

Pathology reviewed, formalin-fixed, paraffin embedded (FFPE) tissue blocks with adjacent normal tissue (normal, non-tumor involved organ tissues, not tumor associated stroma) were sectioned, 3x 10um sections, and floated onto glass slides. Tumor (and infiltrating tumor associated stroma) was then manually macrodissected with a scalpel and placed into a sterile microfuge tube.

As noted above, specimens from a small number of studies provided up to 4 x 0.6 mm tissue cores selected from pathologist-marked tumor enriched regions of FFPE blocks. Relevant studies included: AOV, BRO and SEA. To increase the efficiency of RNA extraction from tissue cores, cores were frozen on dry ice, in microfuge tubes, and pulverized using (sterile) micro-pestles.

Deparaffinization

After tumor enrichment (above), all FFPE materials were deparaffinized using 1 ml xylene (per sample), and washed 1x in 95-100% ethanol to remove residual xylene. Samples were then allowed to dry for ~10 min at ambient temperature.

Tissue digestion & nucleic acid extraction

After deparaffinization, RNA was extracted using a modified protocol based on the Qiagen miRNeasy FFPE extraction kit (Qiagen) as outlined previously (21). Specific modifications were tissue digestion with proteinase K at 56°C for 45 min (instead of the recommended 15 min digest), then heating at 80°C for 15 min to reverse fixation cross-links. Optional DNAse 1 digest to remove contaminant DNA was also performed on all samples (10 min digest). Total RNA was eluted from miRNeasy micro columns in 20-50 ul of nuclease free, sterile ddH_2O and quantified on a NanoDrop spectrophotometer. RNA with

A260/280 ratio < 1.5 were excluded from experiments. All RNA was stored at -80°C until NanoString experiments were done.

A.6 NanoString Processing

As noted above for RNA extractions, all sites performing NanoString reactions followed standard operating procedures outlined in advance. Briefly, each day sites processed a maximum of 24 samples. Our standard operating procedure called for 500ng of total RNA, as measured from NanoDrop, combined with hybridization buffer and a custom NanoString reporter and capture CodeSet allowing hybridization for exactly 16 hours (short-hyb, 12 samples per day) or 20 hours (long-hyb, 12 samples per day) at 65°C in a pre-heated thermal cycler. Immediately at the end of the prescribed hybridization period samples were processed on an nCounter prep-station (NanoString) following standard procedures. Loaded cartridges (12 samples) were scanned at maximum resolution on an nCounter Digital Analyser (NanoString). The BC Cancer (Vancouver) site performed scanning on a Gen1 Digital Analyzer, while both USC (Los Angeles) and PMC (Melbourne, sometime denoted as AOC or Australian Ovarian Cancer study) performed scanning on Gen2 Digital Analyzers.

Relevant variables including processing date, operator, site, and hybridization time were recorded/embedded into specimen information (CDF) and data files (RCC). In addition to unique HGSOC samples a number of controls and sample replicates were run at all sites to enable evaluation of data quality (see also Fig SA1).

Reference Pools

To monitor for technical bias across sites and allow for cross-CodeSet comparisons (21), we ran 3 distinct control RNA pools. This reference-based normalization strategy is considered best practice for development of NanoString based clinical tests and is similar to the implementation already in use for Prosigna(22,23) and a number of other in development tests(24). Pools consisted of high-quality RNA from fresh-frozen ovarian cancer samples believed to be representative of all molecular subtypes and/or various ovarian cancer histotypes. Pools were assembled en-mass and aliquoted (5ul, 100ng total RNA) for single use without multiple freeze thaws at all sites. Control aliquots were stored at -80°C until ready for use and shipped on dry ice to all processing sites. Pool1 was run approximately every month at each site. Pool2 and Pool3 were run alternatingly, every other month, at each site.

Cross-Site Controls

In addition to control pools, a subset of 48 samples were run once at each of the three processing centres (144 individual run files created, 1 failed QC). The first 36/48 consisted of randomly selected high-grade serous ovarian carcinoma specimens, 12 from each processing centre. In addition, the Vancouver site selected 12 samples from non-High-grade serous histology samples (3 clear cell, 3 endometrioid, 3 low-grade serous, 3 mucinous). Aliquots of RNA chosen at each site were sent on dry

ice to the other two processing centres. Refer to B.6 below and C.1 for analysis of cross site reproducibility and inter-rater analysis.

Technical Replicate Samples

104 samples were chosen randomly and re-run at the same site with the same extractions of RNA, as a control against sample mix-up and data consistency. Similarly, as we migrated across two additional NanoString CodeSets samples were also replicated from the original RNA, in the second and third CodeSets. Due to yield of RNA from the initial extraction we were unable to run the same set across all three CodeSets, refer to supplemental section D.1 for analysis of replicates within and across CodeSets.

Paired Omentum Sampling

53 samples had paired samples from the adnexa and omentum sites. In these cases, the adnexal site was included in our "unique" HGSOC sample set while the paired other site was reserved to examine molecular subtype heterogeneity linked to specific anatomical sites.

A.7 Quality Assurance, Batch Effect, and Normalization of the NanoString Data

Raw data was assessed using several quality assurance (QA) metrics to measure imaging quality, oversaturation and overall signal to noise(21).

1. *Imaging quality controls*: Samples were flagged as imaging failures if the percentage of lane images FOV obtained was less than 75% of the requested number of fields.

2. *Linearity of the assa*y: Samples were flagged as linearity failures if spiked-in positive control probes at different concentrations had $R^2<0.95$.

3. *Detection of Smallest Positive Control*: Samples were flagged when the 0.5 fM positive control probe is smaller than 2 standard deviations from the mean of the negative controls probes.

4. *Sample Quality:*

    a. *% of Genes above Limit of Detection (LOD) of negative controls*: LOD is an upper bound of the background noise in the system, computed as two standard deviations above the mean of the spiked-in negative control probes. Samples below a 50% threshold were deemed of poor quality and considered failures.

    b. *Signal to noise ratio (S/N):* calculated as a ratio between the geometric mean of housekeeping genes and lower limit of detection: geometric mean/LOD. Samples with signal to noise ratio below a 170 threshold were deemed to be of unrecoverable poor quality and considered failures. Samples with S/N below 1000 (but > 170) were

considered to have low S/N (see also technical variability considered when labelling NanoString samples – section B.4)

Thresholds were set to maximize the number of samples of high quality included in the analysis. **Sample Quality** fails if either the **Limit of Detection** or **Signal to Noise** thresholds are not met.

5. Overall QC. This is an overall quality control flag which fails if any of the **Imaging**, **Linearity**, or **Smallest Positive Control** conditions fail.

A summary of the failures encountered is outlined in Table SA2. Samples are included in the analysis if they passed the overall QC and the sample quality. In Fig SA2 the sample quality is depicted. We do not observe any systematic failures by site.

Table SA2. Quality control failures encountered in the NanoString data. All percentages represent percent of total samples run. Percentages in the row of Total Samples run, represent the breakdown by site.*

	Total	AOC	USC	Vancouver
Total Samples Run	5258 (100%)	984 (18.7%)	1766 (33.6%)	2508 (47.7%)
Imaging Failures	11 (0.2%)	0 (0.0%)	7 (0.4%)	4 (0.2%)
Linearity Failures	15 (0.3%)	0 (0.0%)	15 (0.8%)	0 (0.0%)
Smallest PC Failures	1 (0.0%)	0 (0.0%)	1 (0.1%)	0 (0.0%)
Limit of Detection Failures	289 (5.5%)	70 (7.1%)	27 (1.5%)	192 (7.7%)
Signal to Noise Failures	327 (6.2%)	80 (8.1%)	34 (1.9%)	213 (8.5%)
Sample Quality Failures	346 (6.6%)	92 (9.3%)	35 (2.0%)	219 (8.7%)
Overall QC Failures	22 (0.4%)	0 (0.0%)	18 (1.0%)	4 (0.2%)

*Note the numbers of assays include biological and technical replicates, control specimens and non-HGSOC. Therefore the total numbers here are greater than the number of samples carried through in classifier design, biological correlate and technical control experiments.

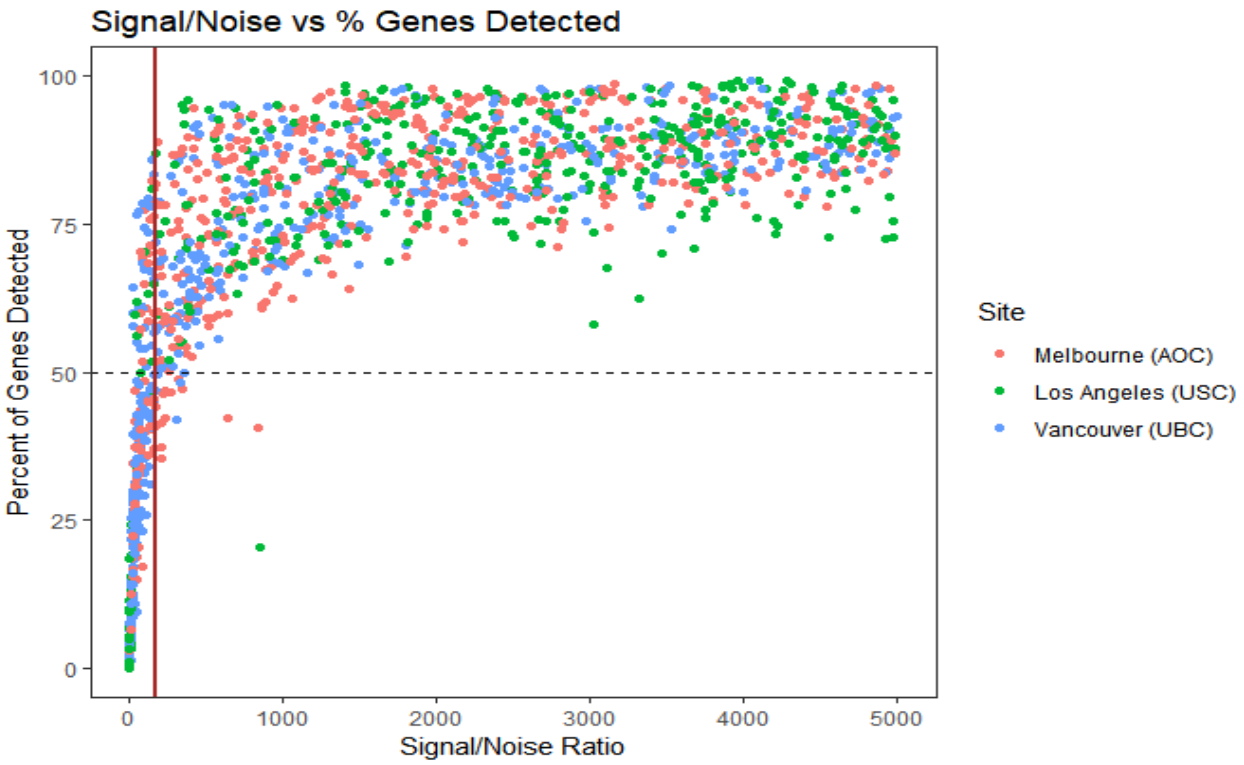

Figure SA2. Sample Quality was assessed by considering Signal to Noise (x-axis) and the % of Genes Detected above LOD (y-axis)

BATCH CORRECTION USING CONTROL POOLS

The reference sample methods described in Talhouk et al was used. Briefly, assuming two batches A and B. To calibrate samples with gene expressing X^B, that were run in batch B to samples with gene expression X^A, that were run in batch A

- Some number of reference samples (R) would be run in both batches A and B, resulting in expression R^A and R^B.
- To remove Batch Effect: X^B-R^B *and* X^A-R^A
- Or alternatively: $X^B + (R^A - R^B)$ would result in calibrating batch B to batch A.

As the same CodeSet was observed at all three sites, little difference was observed across sites; for consistency, everything was calibrated to the Vancouver batch. Fig SA3 depict the average log expression (base 2) of the housekeeping genes of the reference pools run at different sites, over time. We notice a slight degradation of the performance over time.

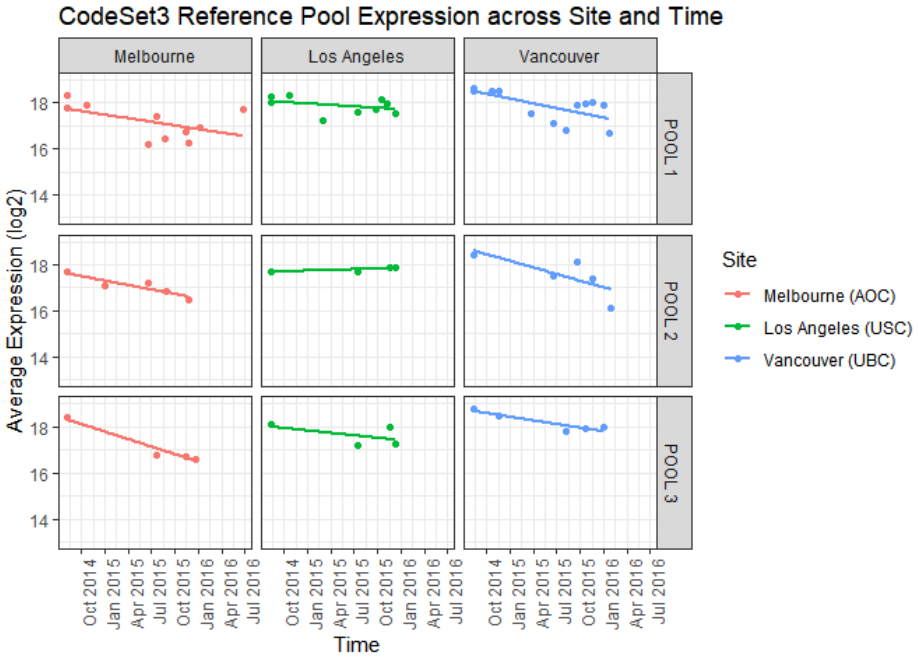

Figure SA3. Reference Pools Over Time

Cross-site Controls

We additionally wanted to confirm that data from biological samples was consistent across site. We used the following reliability measure: the coefficient of accuracy (Ca), squared Pearson's correlation coefficient (R2) and Lin's concordance correlation (Rc). We computed these measures gene-wise (histogram in Fig SA4) and for the average across all genes (Fig SA5).

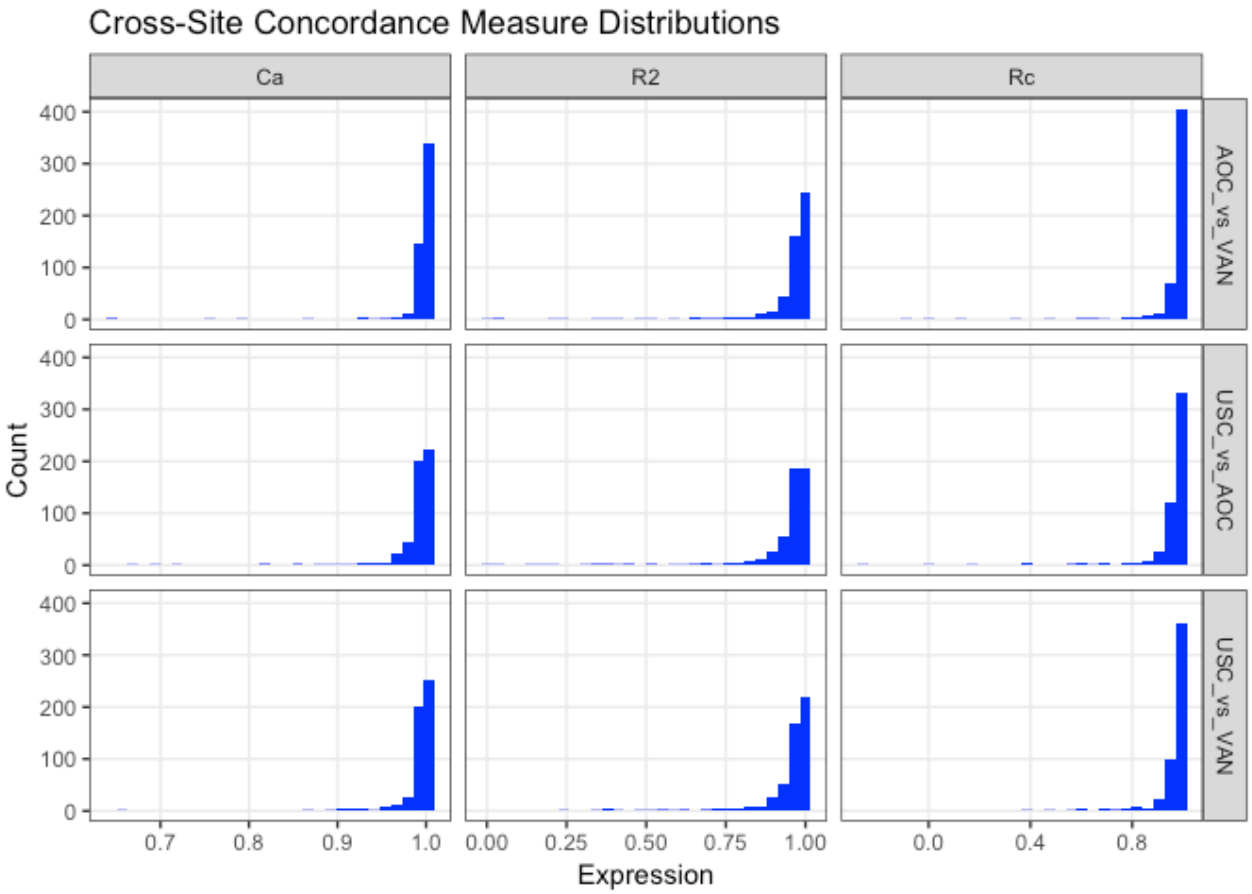

Figure SA4. Data Consistency of Cross-Site Controls across sites. Average of gene-wise reliability measures, when comparing samples that were run cross-site.

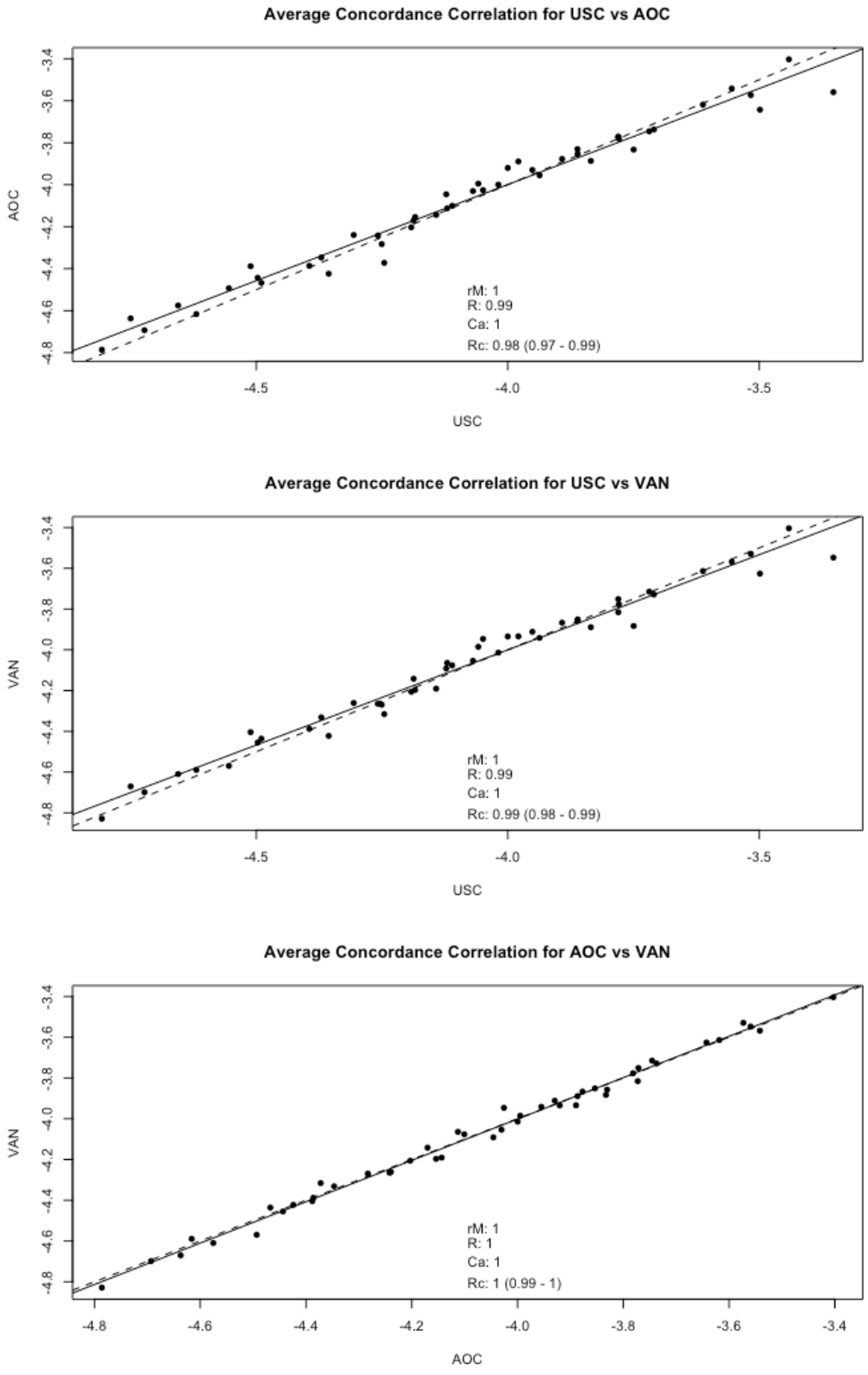

Figure SA5. Pairwise plot of cross-site controls comparing the expression levels of identical samples across different sites.

B: Subtype Labels Assignment to NanoString Data

B.1 Transcriptome-Wide Data Inclusion

We included all high-grade serous ovarian carcinoma (HGSOC) studies for which whole-transcriptome gene expression data are available (Table SB1 and Fig SB1). The histomorphological classification of HGSOC has changed considerably over the past two decades. Since none of the existing considered datasets defined histology based on the 2014 WHO standards(3) (at the time of their publications), we applied the following previously validated criteria(2) to restrict to HGSOC specimens:

- *Presentation:* Primary ovarian carcinoma, including adnexal sites, fallopian tubes or peritoneal, excluding recurrences, ascites, and metastasis to the ovary. If site was unknown, it was presumed to be a primary ovarian cancer.
- *Treatment:* All specimens were presumed to be chemotherapy naïve. Samples showing any indication of exposure to neoadjuvant treatment were excluded. We presumed that first line therapy was consistently platinum +/- taxane and excluded any exploratory therapies.
- *Histological diagnosis:*
    - Serous of either grade 2 or grade 3 **and** stage 3 or 4. If grade was missing and stage was 3 or 4, samples were included. Similarly, if stage was missing, but grade was 2 or 3, samples were also included. If both stage and grade were unknown, samples were excluded unless authors referred to those samples in the text as high grade and/or high stage.
    - Endometrioid histology was included only if grade 3 **and** stage 3 or 4. Unknown grade or stage were excluded.
    - Undifferentiated histology and carcinosarcomas were included if grade 2 or 3 **and** stage 2 or 3. If grade was missing and stage was 3 or 4, samples were included. Similarly, if stage was missing, but grade was 2 or 3, samples were also included. If both stage and grade were unknown, samples were excluded.
    - Mixed histology was included only if the dominant component was identified as serous. If the dominant component was unknown or non-serous, then samples were excluded.
    - All other histologies were excluded.
    - If samples were re-reviewed (post-publication) by an expert pathologist following 2014 standards and were deemed to be HGSOC then they were included, likewise whose reviewed and deemed non-HGSOC were excluded. This applied to samples in the Tothill *et al.* study and a subset of “p53” wildtype samples from the TCGA study(25).
    - If sequencing(25) or IHC data for p53 were available and were not consistent with p53 alterations, then samples were excluded.

    - Samples that clustered in the C3 or C6 groups (only relevant to the Tothill et al(15) study) were presumed to be non-HGSOC

- *Duplicates:* samples that were run in multiple studies were included only once.
- *Sample size:* Studies were included if they had a minimum of n=40 tumors that met all the criteria noted above.

B.2 Two Independent Approaches

We adopted two approaches, performed in parallel by two independent teams. The purpose of this was to mitigate possible downstream effect of batch correction and to ensure a robust final clustering. Prior to data processing, and after curation of samples, data was split into four groups that were processed independently (see also Fig SB1).

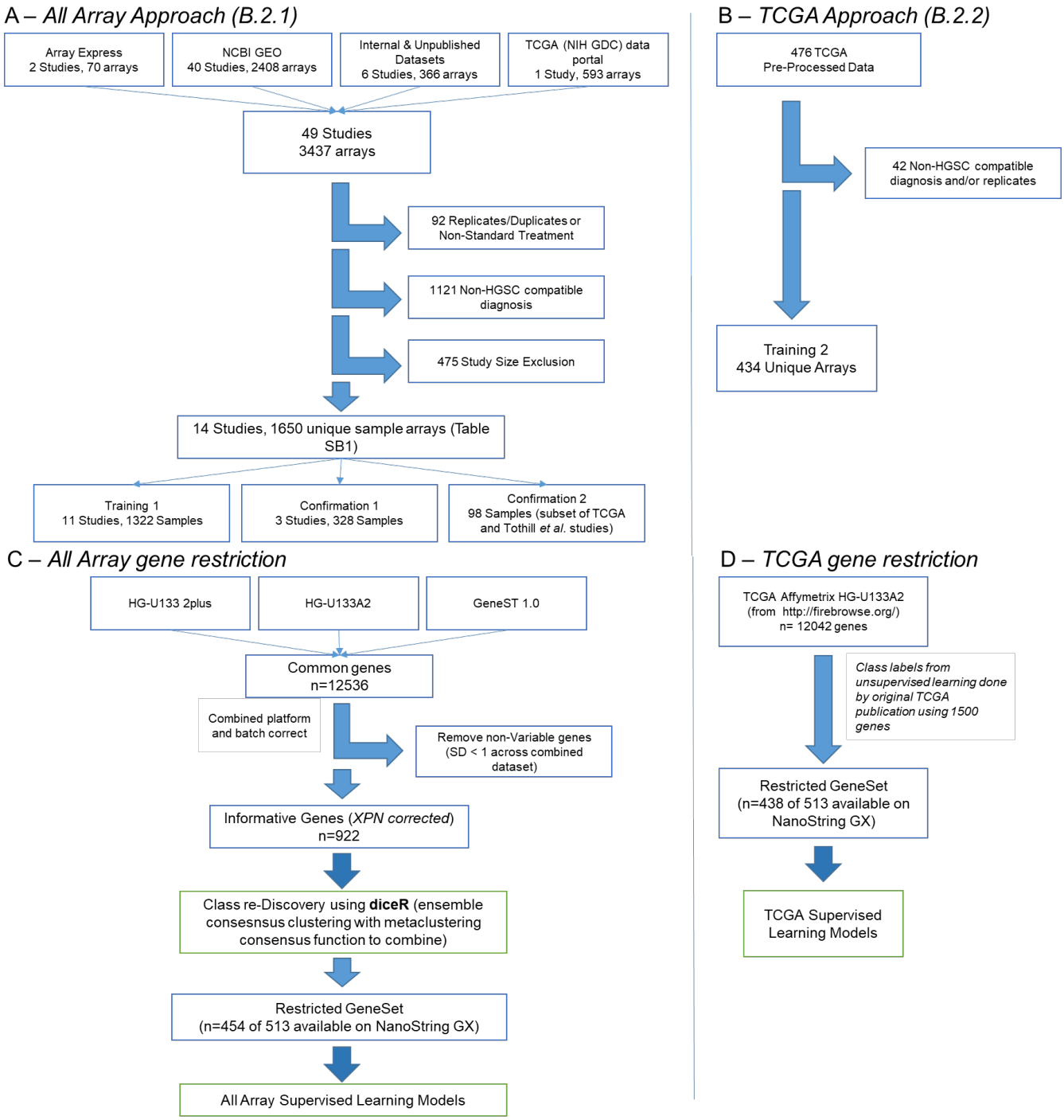

Figure SB1. Upper panels define gene expression array study sample selection and curation workflow (A &B), while lower panels (C & D) define the gene restriction workflow relevant for class re-discovery (*All Array* strategy only) and derivation of *de facto* standard for gene-expression based subtypes (further described below). **A.** Outlines the process for the All Array strategy, additional detail on included studies can be found in Table SB1. **B.** Outlines the TCGA strategy. Note the TCGA cohort is used in both parallel strategies with the *All Array Approach*

taking data from 593 Affymetrix U133A2 arrays (before applying including/exclusion criteria), and the *TCGA Approach* taking expression array data from the Broad Institute data portal (Affymetrix U133A2). The same inclusion/exclusion criteria (section B.1) were applied to all samples regardless of approach. **C.** Outlines the total number of genes considered that were common to array platforms used in the *All Array Approach*, the filtering applied before clustering for class-discovery, and gene restriction for supervised model development. **D.** likewise described the number of genes available for the *TCGA Approach* from the Broad Institute data portal (Affymetrix U133A2) and restrictions applied for supervised model development. No class discovery step is undertaken in the *TCGA strategy* as class labels are used from the class discovery exercise undertaken in the Cancer Genome Atlas Consortia study.(12)

In the approach referred to as the "*All Array Approach*", we used:

1. *Training 1:* samples and data from multiple studies committed as the training set for the *All Array strategy* (n=1322).
2. *Confirmation 1:* samples from multiple Affymetrix array datasets, excluded from any training sets (above; n=328).
3. *Confirmation 2*: a subset of samples for which both Affymetrix array and NanoString data were available. This subset of samples was excluded from the training sets noted above and was used to ensure cross-platform portability for both the *All Array* and *TCGA strategies* (n=98).

In the other approach, referred to as the "*TCGA Approach*", we used:

1. *Training 2:* samples and pre-processed data accessed directly from the Broad Institute data portal (Affymetrix U133A2; http://firebrowse.org/?cohort=OV&download_dialog=true)(12), curated following rules outlined above (B.1; Fig SB1). This dataset was committed as the training set for the *TCGA strategy* (n=411).
2. *Confirmation 1 & 2:* (described above).

B.2.1 All Array Approach

Normalization

raw data from gene expression microarrays of ovarian cancers on Affymetrix platforms (HG-U133A2, HG-U133 Plus 2, Human Gene 1.0 ST) were collected from Gene Expression Omnibus or ArrayExpress (Table SB1). Samples were processed in separate groups to minimize cross-group influences. Within each group, frozen robust multi-array analysis (fRMA) normalization were performed using R version 3.1.2 (2014-10-31) and Bioconductor packages (affy version 1.42.3) for 3'IVT arrays (HG-U133 series), or Affymetrix Power Tools version 1.15.2 for human exon array (human gene 1.0 ST)(14,26). We retained 12536 probes common across platforms. Probe matching of the Affymetrix platforms were based on "U133PlusVsHuGene_BestMatch" annotation from Affymetrix.

Batch Effects Correction

Batch Effect (BE) correction was performed using two known batch effects removal methods: empirical Bayes (EB) and cross-platform normalization (XPN) using R package inSilicoMerging(27). This resulted in two datasets, one EB corrected and the other XPN corrected, for each group that were carried forward for analysis.

Unsupervised learning for de novo class re-discovery

Class discovery was undertaken using an ensemble unsupervised training approach on Training 1, using the two batch effect correction approaches (XPN and EB) independently. This was done using the CRAN library diceR described in Chiu and Talhouk(28). Briefly, for each dataset, genes with standard deviation greater than 1 were retained and the dataset standardized gene-wise to have mean 0, variance 1.

One thousand subsamples of the data were generated by sampling without replacement, each time selecting 80% of the cases. The following nine unsupervised algorithms were repeatedly used on each subsample to assign clusters: Non-negative matrix factorization using Kullback-Leibler divergence (NMF_Lee)(29), Non-negative matrix factorization using Euclidean distance (NMF_Brunet)(30), k-means clustering using Euclidean distance (KM_Eucl), k-means clustering using Spearman distance (KM_Spear), k-means clustering using Manhattan distance (KM_Manh), partition around medoids clustering using Euclidean distance (PAM_Eucl), partition around medoids clustering using Spearman distance (PAM_Spear)(31), partition around medoids clustering using Manhattan distance (PAM_Manh), block clustering using latent model (BLOCK)(32). We excluded hierarchical clustering algorithms because those have been known to perform poorly in high dimensions(33).

The k-modes(34) consensus function was used to combine the clustering to result in a final consensus cluster assignment. Principal Components Analysis (PCA) and Principal Variance Components Analysis (PVCA)(35) were used to measure the impact of batch and assess sources of variability in the corrected data. Figure SB2 shows that for Training 1 both batch correction methods eliminate the variability associated with study. The XPN batch effect correction method performed nominally better in our dataset and was selected for onward analysis.

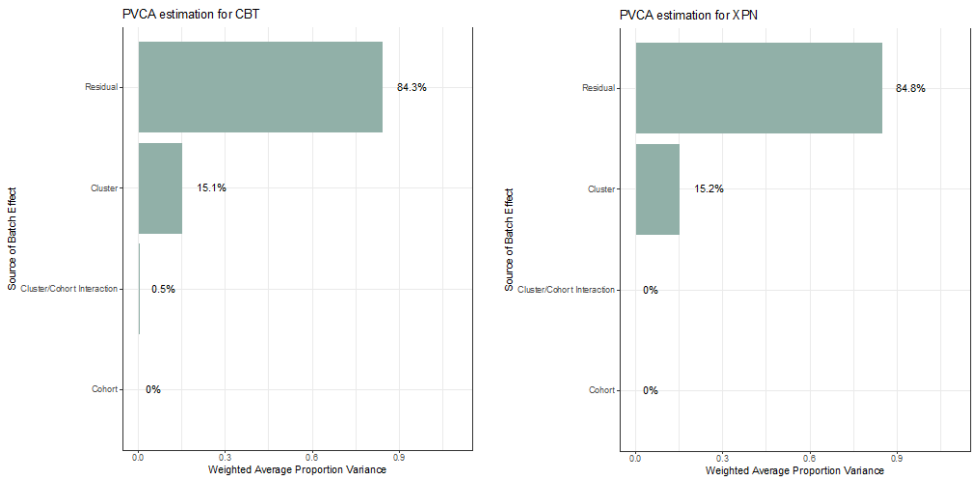

Figure SB2. Principal Variance Components Analysis (PVCA) analysis illustrating sources of variance explained by the different component. The percentages represent the variability explained by each factor and first

Pathway Analysis

Biological differences between molecular subtypes was examined using Significance Analysis of Microarrays (SAM) analysis performed in R (version 3.5.0) using RStudio (version 1.1.453) and the package samr(13). The package was run using Shiny through GitHub using the instructions on the website (https://github.com/MikeJSeo/SAM). Analysis was performed as a one vs. all analysis where each subtype was compared to the other three grouped together. The following variables in SAM analysis were set as below:

Data type = Array

Response Type = Two class unpaired

Analysis = Standard (genes)

Test statistic = T-statistic

Median centre the arrays = No

Are data in log scale (base 2) = Yes

Estimator of s0 factor for denominator = Automatic

Number of permutations = 1000

K-nearest Neighbours Imputer = 10

Random seed = 1234567

When saving results, the delta value was manually entered to whichever value gave a median FDR of <0.05 as shown in the delta table results. All other options when saving results were left as default. Delta values used:

Subtype 1 Analysis delta = 0.47

Subtype 2 Analysis delta = 0.32

Subtype 3 Analysis delta = 0.05

Subtype 4 Analysis delta = 0.16

For each subtype, two lists of genes were generated, based on whether they were overexpressed ("Significant Positive" in tables SB2-SB5), or under expressed ("Significant Negative") relative to the other three subtypes, with statistical significance defined as an FDR q-value cut off of 0.05 (Tables SB2-SB5 for subtypes 1-4 respectively). For example, for subtype 2, the "Significant positive" gene list represents all genes overexpressed in the subtype labelled as "2", compared to all other subtypes (1, 3, and 4), and "Significant Negative" genes are genes that are under expressed in that subtype compared to all others. To examine biological pathways that are enriched in each subtype the statistically significant genes from each subtype analysis were used for gene set analysis using the Molecular Signature Database (MSigDB 6.1, Broad Institute, http://software.broadinstitute.org/gsea/msigdb/index.jsp) to compute the overlap of overexpressed and under expressed genes(36-38). This was done separately for overexpressed (positive) and under expressed (negative) genes. For this analysis genes were compared against the "Canonical Pathways" subset (C2:CP), with the top 20 pathway-based gene sets returned (ranked by FDR; $q<0.05$). The biological pathways over- and underexpressed in each subtype were compared back to previously published studies reporting on molecular subtyping and their gene expression patterns to label our clustered subtypes. Values of specific genes previously reported in subtypes were also examined in the centroid data to confirm if specific genes were expressed higher in specific subtypes.

Subtype 1 = Immunoreactive (aka. C2 or C2.IMM)

Subtype 2 = Differentiated (aka. C4 or C4.DIF)

Subtype 3 = Proliferative (aka. C5 or C5.PRO)

Subtype 4 = Mesenchymal (aka. C1 or C1.MES)

Subtype 1 has a similar biology to the previously reported Immunoreactive subtype, with 12 of the top 20 upregulated pathways being immune related or containing immune related genes. These include genes involved in the Immune System, Interferon Signaling, Cytokine Signaling in the Immune System, Interferon Alpha/Beta Signaling, Chemokine receptors bind chemokines, and Interferon Gamma signaling. Genes with the highest fold-change included *CCL5*, *CXCL10*, *TAP1*, *CXCL9*, and *CXCL11*, all of which are immune genes that have previously been shown as strongly expressed in the immunogenic subtype. As few as ten transcription factor regulators may control immune gene

expression in the stroma of Immunoreactive tumours(39). Two of these master regulators are present in our data set (*HCLS1* and *IRF7*) and showed statistically significant up-regulation in Subtype 1.

Subtype 2 best represents the Differentiated subtype which has been previously characterised by markers of differentiation and ovarian/fallopian physiology. This subtype showed the highest expression of mucins *MUC1* and *MUC16* and enrichment of the pathway Genes involved in O-linked glycosylation of mucins, and the highest expression of the secretory fallopian tube marker *SLPI*, previously linked to the differentiated subtype. Cell cycle gene *CDKN2A* and transcription factor *MYC* were overexpressed in the Differentiated (12,15,17) subtype and are similarly overexpressed here. Other pathways and genes previously associated with the Differentiated subtype (12,15,17) such as pro-apoptotic genes, other ovarian physiology genes, cell cycle, and double strand break repair were not enriched in this analysis but are due to the genes not being represented in the array data set. The HER2 pathway (including *ERBB3*, *ERBB4*, *ESR1* and *IL6R*) are statistically significantly overexpressed in this subtype, along with nine solute carrier (SLC) genes, enriching the Transport of small molecules and SLC mediated transmembrane transport pathways.

One study(40) has suggested that the TCGA Differentiated/Tothill C4 subtype is best represented by two distinct subtypes they refer to as Anti-mesenchymal (characterized by statistically significant down regulation of extracellular matrix (ECM) genes including *COL11A1*, *POSTN*, *VCAN*, *DCN*, *ZEB1* and *FAP*, and up regulation of oxidative phosphorylation, peroxisome, and butanoate metabolism pathways) and a second subtype that is more closely related to the Differentiated subtype (characterized by increased expression of *COLEC11*, *STAR*, *ITGB4*, *MGLL*, *MLPH*, *DEFB1* and up regulation of Ribosome and Cytochrome P450 pathways). Subtype 2 in the current study has features of both these groups, with statistically significant decreases in ECM genes and pathways including those associated with anti-mesenchymal, and strong up regulation of *COLEC11*, *STAR*, *MGLL*, *MLPH*, and *DEFB1*. Enrichment of the pathways associated with the two subtypes was not observed but may be because a large number of the genes involved in these pathways were not in the data set. However, eight genes from these pathways were quantified and were statistically significantly upregulated in Subtype 2 (*ATP6V1B1*, *ACSL5*, *ACSM3*, *ALDH3B2*, *PIPOX*, *MOAB*, *GSTA2*, and *UQCRH*). This shows that the biology of these two subtypes may be present in Subtype 2 in this study.

Subtype 3 closely matches the biology of the previously reported Proliferative/ Tothill C5 subtype(12,15,17). Several of the top upregulated pathways were enriched for WNT signaling genes, and included two WNT Signaling Pathways, which have been associated with the Proliferative subtype in several previous studies. Several transcription factors were also overexpressed as previously described (including *HMGA2*, *TOX*, *TCF7L1*, and *SOX11*), some homeobox genes (*HOXB6*, *HOXB7*, *HOXC6*) and the stem cell marker *LGR5*. This subtype also has the lowest expression of ovarian

markers *MUC1* and *MUC16* and kallikreins (*KLK5*, *KLK6*, *KLK7*, *KLK8*, *KLK10*, *KLK11*) confirming the suggestion of a de-differentiated phenotype. Cadherin *CDH2* and *CDH3* are also overexpressed. Double strand break and homologous recombination repair have been reported to be increased in the Proliferative subtype. Many of these genes were not in this data set but one gene previously reported (*RAD51AP1*) was statistically significantly increased in this subtype.

Eleven of the top 20 pathways enriched for down regulated genes include immune genes indicating this subtype shares the feature of the Proliferative subtype of low immune expression. The top significant down regulated genes are related to antigen presentation including many HLA/MHC genes.

The TCGA analysis(41) reported high proliferative markers *MCM2* and *PCNA* in the Proliferative subtype. Another study of molecular subtypes (42) did not replicate this proliferative signature and suggested mesenchymal may be a better descriptor for this subtype. Our data do not include *PNCA* and *MCM2* so can't we distinguish this in our subtype, however other cell cycle genes including *CCND1*, *CCND2*, and *CCNE2* are overexpressed in this subtype.

Subtype 4 has statistically significant enrichment for stromal genes and pathways that matches it to the TCGA Mesenchymal/Tothill C1 subtype(12,17). All of the top 20 pathways upregulated in Subtype 4 are related to and enriched with extracellular matrix related genes, and includes Core Matrisome, Extracellular matrix glycoproteins, Extracellular matrix regulators, Focal Adhesion, Extracellular matrix organization, Integrins, Collagen formation, and Signaling by PDGF. These contain many collagen genes, MMPs, remodeling genes such as PLAU, Fibroblast- and Platelet Derived- Growth factors. Several markers of activated cancer associated fibroblasts (CAF) are statistically significantly upregulated in this subtype including *FAP* and *ACTA2* possibly suggesting a high CAF population in the tumor microenvironment. Epithelial to Mesenchymal Transition (EMT) markers are also upregulated including *SNAI2*, *ZEB1*, and *TWIST1*. Five genes from the Insulin like growth factor (IGF) pathway are also significantly upregulated (*IGF1*, *IGF2*, *IGFBP3*, *IGFBP4*, and *IGFBP5*).

There was some up regulation of immune genes in this subtype with two immune pathways enriched in the top 20; Complement and Coagulation Cascades and Cytokine Cytokine-Receptor Interactions. Some antigen presenting genes are also statistically significantly upregulated; however, the immune signature is not as strong as the stromal signature, or the immune signature in Subtype 1. This is consistent with the previous finding of Tothill *et al*. that the Mesenchymal (C1) subtype has some elevated immune pathways (15).

Alongside the master regulators of the Immunoreactive subtype, Zhang *et al*. (39) also identified 6 master regulator transcription factors controlling the stromal signature in the Mesenchymal subtype.

Four of these (*AEBP1*, *HOPX*, *SNAI2*, and *ZEB1*) were quantified in our data set and highly upregulated in Subtype 4.

Down regulated genes in Subtype 4 show enrichment for cell cycle and cell division pathways. The top 20 down regulated pathways include Kegg Cell Cycle, Reactome Cell Cycle, Meiotic Recombination, Meiotic Synapse, Meiosis, Telomere Maintenance, and RNA transcription. Most of these pathways are enriched by significant down regulation of cell cycle genes (*CCNA1*, *CCNE1*, *CCNE2*, and *CDKN2A*) and 7 histone genes (*HIST1H2AC*, *HIST1H2AE*, *HIST1H2BC*, *HIST1H2BD*, *HIST1H3H*, *HIST2H2AA3*, and *HIST2H2BE*).

All four subtypes had upregulated pathways that were enriched with genes related to extracellular matrix genes. Two pathways in particular are present in all four subtypes; NABA Matrisome and NABA Matrisome Associated. These pathways include a broad range of genes that are linked to the extracellular matrix and have a high degree of overlapping genes(43). The enrichment of these pathways in some of the subtypes are driven by genes unique to the subtype, e.g. in Subtype 2 (Differentiated) these pathways contain genes like mucins that are not seen in the other subtypes. However, some subtypes such as Subtype 1 (Immunoreactive) and 4 (Mesenchymal) share similar upregulated genes including cytokines, collagens and matrix metalloproteinases. This suggests that while a strong up regulation of stromal and extracellular matrix genes is a key feature of the Mesenchymal subtype, all four subtypes have some unique differential expression of ECM genes, and some overlap of ECM gene expression, particularly the subtypes matching the Immunoreactive and Mesenchymal subtypes.

Outcome by re-established molecular subtypes

Using outcome data available from the originating studies compiled in CISOVDB (26) and updated outcome data from the ovarian TCGA cohort(12,44) we re-examined overall and progression free survival (Fig SB3). Overall survival data reflected a similar pattern to previous reports with C2.IMM and C4.DIF having the best outcomes and C1.MES having the worst. Differences were less pronounced in disease-free survival (progression-free survival) but the overall pattern remained consistent.

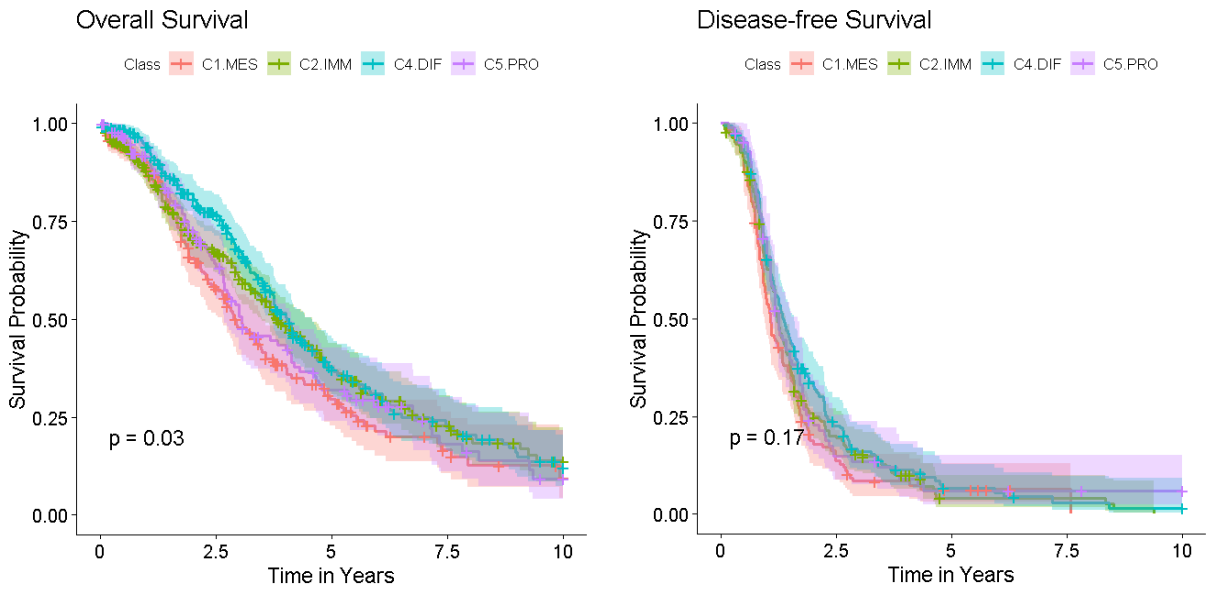

Figure SB3. Overall and disease-free (progression free) survival for all samples in the All Array strategy, segregated in the de-novo re-discovered molecular subtypes.

Supervised Learning

Using the data from Training 1 only (see Fig SB1), and treating the labels assigned based on pathway analysis in the previous step as gold standard, we restricted the gene set to include only those genes that were evaluated in our NanoString custom CodeSet (n=454/513 possible genes on the NanoString platform).

The genes were additionally filtered to include only the top variable genes (minimum standard deviation of 0.5). Furthermore, to ensure the cross-platform portability of the final model, data in Affymetrix were "standardized" sample-wise by subtracting the average expression level of the housekeeping genes: *ACTB*, *POLR1B*, *SDHA*, *RPL19, PGK1.* This means that for each sample the expression level is no longer an absolute measure, but rather is the expression level relative to housekeeping genes; this was done to facilitate future portability to NanoString where samples are normalized in this manner.

Several algorithms were used for model training: linear discriminant analysis (LDA), random forest (RF), prediction analysis for microarrays (PAM), multinomial regression with L1 penalty with embedded feature selection (MLR_lasso), multinomial regression with L2 penalty where coefficients are shrunk

towards 0 but not set at 0 (MLR_ridge), K-nearest neighbours (KNN), adaptive boosting using decision trees (adaboost), naïve bayes (nbayes), and support vector machines (SVM).

Internal Validation

A bootstrap approach was used for internal validation and to correct for over-optimism, whereby n observations of size equal to the training dataset (i.e. group 1) were sampled with replacement. From this, a model was built using those observations and evaluated on the "out of bag" samples, i.e. the samples that were not selected as part of the bootstrap subset(45). This was repeated 500 times. Hyperparameters, tuning parameters and supervised training algorithms are fit within each bootstrap iteration before any out of bag prediction and evaluation.

These models each resulted in a "probability measure" that assigned the class based on the highest probability. The obtained predictions are compared to the gold standard label which was assigned via the de novo clustering in the previous step (see A.3). The top 5 performing algorithms are ranked below (Table SB6).

Table SB6. Measures of accuracy of all generated models from the All Array approach in internal validation. For all metrics, a higher score indicates increased accuracy.

Strategy	Algorithm	Metrics (median; 95% CI)				
		Overall Accuracy	F1-score C1.MES	F1-score C2.IMM	F1-score C4.DIF	F1-score C5.PRO
All Array	SVM	0.87 (0.85-0.89)	0.89 (0.86-0.92)	0.86 (0.82-0.88)	0.88 (0.85-0.91)	0.84 (0.79-0.88)
All Array	MLR_ridge	0.85 (0.83-0.88)	0.89 (0.86-0.93)	0.83 (0.80-0.86)	0.85 (0.82-0.89)	0.83 (0.78-0.88)
All Array	Adaboost	0.85 (0.82-0.87)	0.89 (0.85-0.92)	0.84 (0.80-0.87)	0.85 (0.82-0.89)	0.79 (0.72-0.85)
All Array	MLR_lasso	0.85 (0.82-0.87)	0.90 (0.86-0.93)	0.83 (0.80-0.86)	0.85 (0.82-0.88)	0.82 (0.77-0.87)
All Array	RF	0.85 (0.83-0.87)	0.89 (0.85-0.92)	0.84 (0.81-0.87)	0.86 (0.83-0.89)	0.80 (0.73-0.86)

The top five ranking algorithms above (adaboost, random forest, lasso, SVM and ridge regression) were followed up with external validation.

External Validation

We tested the top five algorithms (adaboost, random forest, lasso, SVM and ridge regression) on independent, XPN batch-corrected data -- Confirmation 1 (defined in A2). We used previously defined knowledge to assess each model by first assigning molecular subtype labels using ssGSEA and the gene-set signatures published for CLOVAR(17) and the Australian ovarian cancer study (AOCS) dataset(15,16). Using these signatures in ssGSEA we assigned a molecular subtype label to

Confirmation 1 (see Fig SB1), that was concordant for both CLOVAR and AOCS signatures, for 241/328 samples. We then used these labelled samples to evaluate predictions from our five all-array models. Performance was highly concordant across the board with accuracy ranging from 0.747 to 0.784. As this information suggested all models performed similarly, we did not lock down on these criteria, but decided to examine consistency from genome-wide expression array to our target NanoString platform, in Confirmation 2, to lock down a final model for the All Array strategy.

Second External Validation

To lock down a final model we examined the performance of array-derived models on a subset of samples that existed in both the array and NanoString datasets and overlapped with the Tothill and TCGA publications (defined as Confirmation 2, Fig SB1). Here we found that the Adaboost model was ranked highly when comparing performance to published labels (adaboost: 0.91; lasso: 0.89; ridge: 0.89; random forest: 0.88; SVM: 0.93). Similar results were obtained for pre-processed NanoString data (Adaboost: 0.79; lasso: 0.81; ridge: 0.81; random forest: 0.74; SVM: 0.26; see also section B.5 for NanoString pre-processing details) except for SVM, for which performance can be poor without tuning hyperparameters(46). While most models performed comparably, we felt the robust ensemble paradigm, and otherwise proven performance on difficult to classify specimens(47), was sufficient to justify lock down as our chosen model.

Table SB7. Contingency table showing the relationship between all-array k-modes clustering and TCGA training (published) labels (n=434).

A - Confusion Matrix

	TCGA Published Labels			
All-Array K-modes Clustering Labels	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	79	6	2	4
C2.IMM	36	76	19	13
C4.DIF	2	9	101	23
C5.PRO	3	1	3	57

B - Overall Metrics

Metric		P value
accuracy	0.72 (0.68 - 0.76)	< 0.001
kappa	0.63 (0.57 - 0.68)	

C – By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy
C1.MES	0.66	0.96	0.87	0.88	0.75	0.21	0.81
C2.IMM	0.83	0.80	0.53	0.94	0.64	0.33	0.81

C4.DIF	0.81	0.89	0.75	0.92	0.78	0.31	0.85
C5.PRO	0.59	0.98	0.89	0.89	0.71	0.15	0.78

Software and Computational Resources

R version 3.5.1 (2018-07-02) was used on a Beowulf-style cluster with 500 compute nodes, each with 12 cores and 48 GB RAM, for a total of 6,000 cores (12,000 threads) and 20 TB RAM. All nodes are running CentOS 5.4 and are using open-source Sun Grid Engine 6.2u5 as their scheduler. Each node has 160 GB local temp space, as well as access over an InfiniBand 40Gbps network to a dedicated storage system running GPFS. The file system has 700 TB usable scratch space and is served by 8 IBM x3850 servers. The code is available on GitHub: https://github.com/AlineTalhouk/PrOType.

B.2.2 TCGA Approach

Normalization

The normalized and batch corrected data for 476 unique Ovarian Cancer samples were obtained directly from the TCGA data portal (http://firebrowse.org/?cohort=OV&download_dialog=true)(12). From this, we removed samples that were suspected non-HGSOC (25) leaving 434 unique array specimens for training (Training 2 – see Fig SB1).

Batch Effects Correction

The data from the TCGA portal was already batch-corrected and normalized according to a reference sample approach and combined from gene-expression array platforms as outlined in their original study (12042 genes)(12). As such, no further pre-processing of the data was needed.

Class Labels

Class labels were NOT derived de novo and were obtained directly from TCGA data portal.

Supervised Learning

The gene expression data for all the genes common to the OTTA NanoString platform and the TCGA dataset (n=438/513 possible genes on the NanoString platform) were used to build the classification model. The expression data were standardized gene-wise to have zero mean and unit variance (note: see also supplement B, standardization was also applied in the NanoString expression data when applying models derived from the *TCGA strategy*).

Several classification algorithms known to work efficiently with gene expression data based classification including Diagonal Linear discriminant analysis (DLDA), K-Nearest Neighbor (KNN), Support Vector Machines (SVM), Gradient Boosting Machines (GBM) and Random Forests (RF) were explored to predict the subtype label assigned by the TCGA project.

Internal Validation

A different internal validity approach was used here. The accuracies of the different classifiers were estimated using a k-fold cross-validation approach(48). In k-fold cross-validation, the original sample is randomly partitioned into k equal sized subsamples.

Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining k − 1 subsamples are used as training data.

The predicted labels are then compared to that of the TCGA original labels. The accuracy of the classification models was estimated using 10-fold cross validation. Results are outlined in Table SA6.

Table SB8. Measures of accuracy of all generated models from the TCGA approach in internal validation. For all metrics, a higher score indicates increased accuracy.

Strategy	Algorithm	Metrics (median; 95% CI)				
		Overall Accuracy	F1-score C1.MES	F1-score C2.IMM	F1-score C4.DIF	F1-score C5.PRO
TCGA	DLDA	0.87 (0.84-0.90)	0.94	0.82	0.85	0.86
TCGA	KNN	0.88 (0.84-0.91)	0.92	0.85	0.86	0.86
TCGA	SVM	0.87 (0.83-0.90)	0.90	0.85	0.86	0.85
TCGA	GBM	0.86 (0.83-0.91)	0.90	0.86	0.86	0.85
TCGA	RF	0.87 (0.83-0.90)	0.90	0.85	0.86	0.85

Among the classification algorithms tested, Random Forest (RF) and SVM provided the best performance metrics. RF was chosen due to processing speed advantage.

External Validation

The dataset from Tothill *et al.*(15) with published subtype labels were used as the external dataset to assess the classifier accuracy. The RMA normalized data and log2 transformed expression data was downloaded from GEO (GSE9899). Gene expression data were then standardized to have zero mean unit variance. The transformed data was then used as an input to obtain the subtype labels of the samples. The published labels of the samples were then used to compute the accuracy of the classifier.

Table SB9. Contingency table showing the relationship between the TCGA strategy predicted labels (row) and the published labels (column) (n=215).

A - Confusion Matrix

	AOCS Published Labels			
AOCS Predicted Labels	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	76	3	0	1

C2.IMM	4	42	0	1
C4.DIF	0	5	46	1
C5.PRO	3	0	0	33

B - Overall Metrics

Metric		P value
accuracy	0.92 (0.87 - 0.95)	< 0.001
kappa	0.88 (0.83 - 0.94)	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy
C1.MES	0.92	0.97	0.95	0.95	0.93	0.37	0.94
C2.IMM	0.84	0.97	0.89	0.95	0.87	0.22	0.90
C4.DIF	1.00	0.96	0.88	1.00	0.94	0.24	0.98
C5.PRO	0.92	0.98	0.92	0.98	0.92	0.17	0.95

Software and Computational Resources

The supervised approach using the TCGA dataset as the learning set was also developed using R version 3.5.1. The code is available on GitHub: https://github.com/AlineTalhouk/PrOType.

B.3 Platform Portability of the Array Classifier

Using the final locked-down models from each of the two strategies (All array and TCGA; see supplemental A.5), and the data overlapping the NanoString platform (see section A.1 and Fig SA1), we compared the subtype predictions generated from the array data and the NanoString data to assess model portability across platforms. Table SB10 and SB11 compares the cross-platform predictions from the All-array and the TCGA strategies, respectively. When we consider, the consensus label from both models obtained in NanoString with a small subset of cases (the overlap set; Fig SA1) the agreement is excellent (94% accuracy and 0.92 kappa; Table SB12).

Table SB10. All Array model portability comparing predictions on NanoString to predictions on Array.

A - Confusion Matrix

	Predicted on Array			
Predicted on NanoString	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	21	2	1	3
C2.IMM	2	13	4	0
C4.DIF	2	5	20	2
C5.PRO	3	2	1	16

B - Overall Metrics

Metric		P value
accuracy	0.72 (0.62 - 0.81)	< 0.001
kappa	0.63 (0.51 – 0.75)	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy
C1.MES	0.75	0.91	0.78	0.90	0.76	0.28	0.83
C2.IMM	0.59	0.92	0.68	0.88	0.63	0.20	0.76
C4.DIF	0.77	0.87	0.69	0.91	0.73	0.30	0.82
C5.PRO	0.76	0.92	0.73	0.93	0.74	0.23	0.84

Table SB11. TCGA model portability comparing predictions on a subset of samples from the Tothill et al. dataset performed on Affymetrix Array data versus those done on NanoString data.

A - Confusion Matrix

	Predicted on Array			
Predicted on NanoString	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	21	2	1	2
C2.IMM	1	7	2	1
C4.DIF	1	7	19	1
C5.PRO	0	1	1	18

B - Overall Metrics

Metric		P value
accuracy	0.76 (0.66 - 0.85)	< 0.001
kappa	0.68 (0.56 – 0.80)	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy
C1.MES	0.91	0.92	0.81	0.97	0.86	0.31	0.92
C2.IMM	0.41	0.94	0.64	0.86	0.50	0.13	0.68
C4.DIF	0.83	0.85	0.67	0.93	0.75	0.33	0.84
C5.PRO	0.82	0.97	0.90	0.94	0.86	0.24	0.89

Table SB12. Comparing consensus samples predictions with published labels.

(note: 76 samples from array data had consensus labels, CL, however only 67 of these had published labels in original studies(12,15))

A - Confusion Matrix

Consensus Labels	Published Labels			
	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	20	1	0	0
C2.IMM	0	10	1	0
C4.DIF	0	1	17	1
C5.PRO	**0**	**0**	**0**	16

B - Overall Metrics

Metric		P value
accuracy	0.94 (0.85 - 0.98)	< 0.001
kappa	**0.92 (0.84 - 1)**	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy
C1.MES	1.00	0.98	0.95	1.00	0.98	0.31	0.99
C2.IMM	0.83	0.98	0.91	0.96	0.87	0.16	0.91
C4.DIF	0.94	0.96	0.89	0.98	0.92	0.28	0.95
C5.PRO	0.94	1.00	1.00	0.98	0.97	0.24	0.97

B.4 Subtype Assignment in the NanoString data

Agreement between the TCGA and All array models

The locked down models from each of two strategies, All array and TCGA models are used to assign subtypes to all of the NanoString data that passed QC. Since no gold standard exists, the predictions from the two models were compared to one another (Table SB13). Agreement between models was 79.1% and we took this consensus to be the *de facto* standard for downstream design of a minimal-gene classifier.

Table SB13. Comparison of subtypes predictions from the all array model and the TCGA model.

A - Confusion Matrix

Predicted Labels-All-array model	Predicted Labels - TCGA model			
	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	922	94	17	44
C2.IMM	149	619	148	41
C4.DIF	0	170	923	37

C5.PRO	21	15	63	566

B - Overall Metrics

Metric		P value
accuracy	0.79 (0.78 - 0.8)	< 0.001
kappa	0.72 (0.7 – 0.74)	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy
C1.MES	0.84	0.94	0.86	0.94	0.85	0.28	0.89
C2.IMM	0.69	0.88	0.65	0.90	0.67	0.25	0.79
C4.DIF	0.80	0.92	0.82	0.92	0.81	0.30	0.86
C5.PRO	0.82	0.97	0.85	0.96	0.84	0.17	0.90

Predictive Entropy

For each sample, each of the two models resulted in four probabilities that quantify the likelihood of the sample to belong to each subtype. Subtype assignment is by default the subtype with the highest probability. This is not indicative of how certain the assignment of the subtype with the highest probability is in comparison to the other subtypes, therefore we quantified and compared each of the model's uncertainty using predictive entropy.(49) The idea of entropy is rooted in information theory and is used to quantify the expected "surprise". Entropy is lowest when the probability for one class is 100% and highest when the probabilities are uniform among the four classes. The formula for entropy is defined as: $H(X|Y) = -\sum_{i,j} p(x_i, y_j) \log \frac{p(x_i, y_j)}{p(y_j)}$

Where $p(x_i, y_j)$ is the probability that $X = x_i$ and $Y = y_j$. For a four-class prediction model, and using log base 2, entropy will vary between 0, when the model predicts a single class with 100% probability (and the other classes with 0 probability), and 2 when each of the classes are assigned equally 25% probability. In Fig SB4, we depict the entropy from predictions generated from the TCGA and All array models on the entire NanoString dataset and the samples where there was disagreement in class assignment. Within each model (all-array and TCGA) the predictive entropy was significantly different among samples that were assigned the same label by the two models in comparison to those that had different labels (P value < 0.0001 in both cases using the Mann-Whitney U Test).

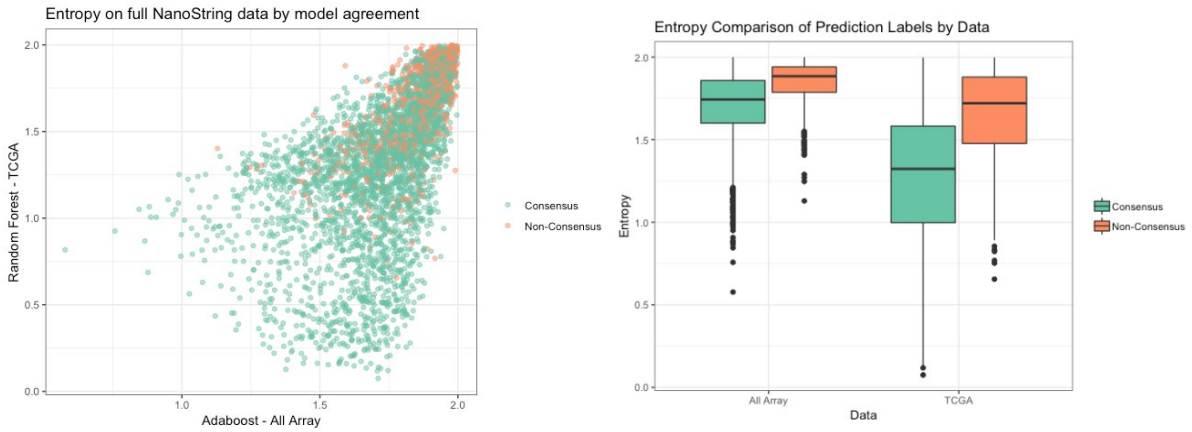

Figure SB4. Entropy from the two models on the entire NanoString data highlighting the points where the two models agree/disagree

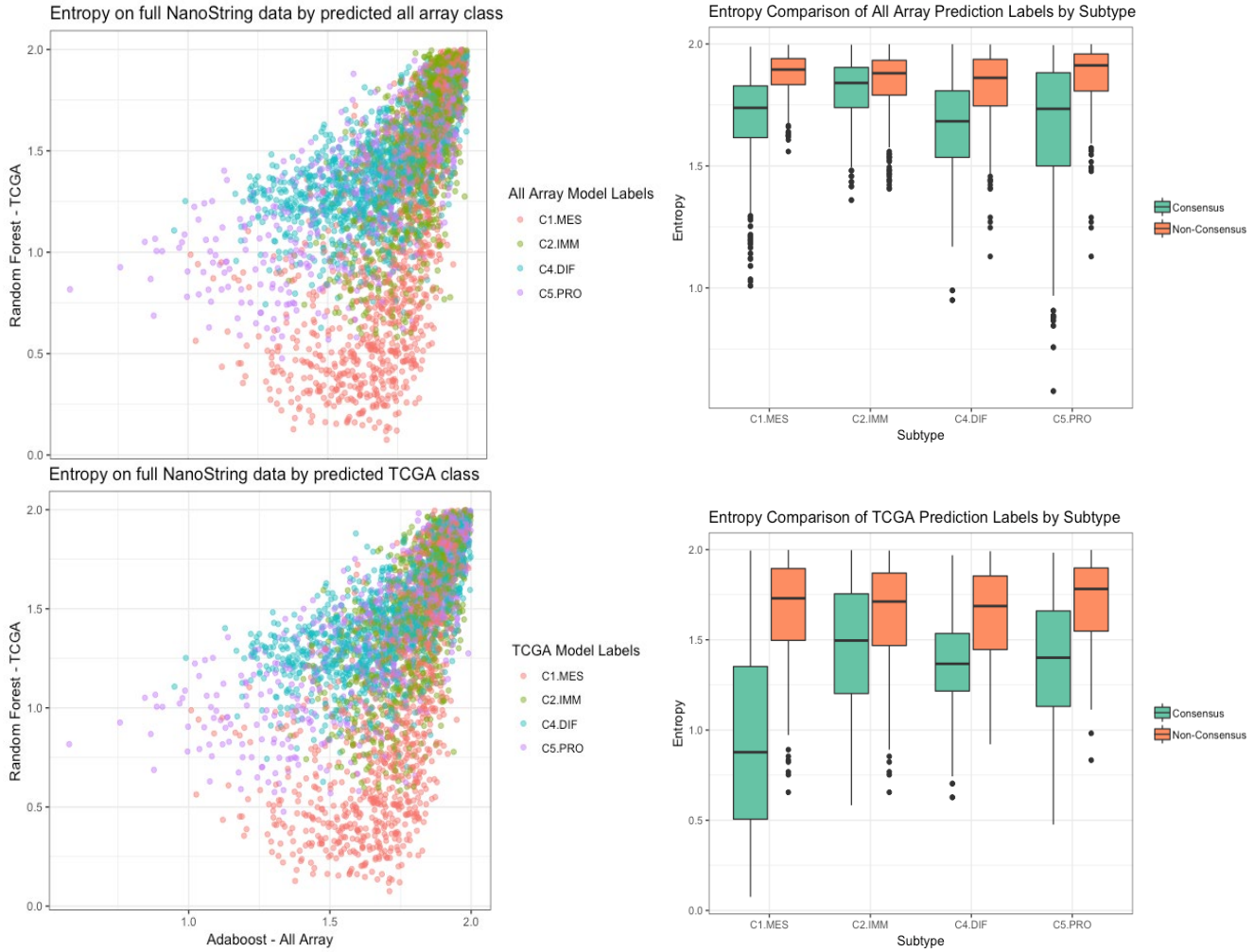

Figure SB5. Entropy from the two models on the entire NanoString data highlighting the four subtypes as classified by the All-array and TCGA model

In Fig SB5, we compare the TCGA and All array entropy across subtypes. It is evident that for both methods it was more difficult to classify the C2.IMM subtype. The C1.MES subtype had lower entropy especially within the TCGA modeling approach.

Technical Variability between consensus and non-consensus labelled samples

We examined potential sources of technical variability between consensus and non-consensus labelled specimen. Only signal-to-noise ratio (below ratio of 1000) was significantly different between consensus and non-consensus samples. This is consistent with lower quality of RNA input and hence inaccuracy of gene expression measurements in a subset of sample. Nonetheless only a relatively small fraction of non-consensus samples had lower S/N (<1000) and low S/N was also present in a subset of consensus labelled samples (5% of CL vs 7.5% of non-CL; see table below). Overall this suggested that technical variability due to platform processing was responsible for only a very small percentage of disagreements between *All Array* and *TCGA* final models.

COMPARISON OF CONSENSUS VS. NON-CONSENSUS FOR TECHNICAL AND BIOLOGICAL VARIABLES

Cohort characteristics for all cases by sample label

Variable	Levels	CL	non-CL	Total	PValue
	N (%)	2667 (78.7%)	720 (21.3%)	3387 (100.0%)	OneWay_Test
Age at Diagnosis	Mean (sd)	60.2 (10.5)	59.8 (10.6)	60.1 (10.5)	0.3862
	Median (IQR)	60 (53 - 67)	60 (52 - 67)	60 (53 - 67)	
	Missing	50	15	65	
					PearsonChi_square
Cellularity	0-20	22 (0.9%)	10 (1.5%)	32 (1.0%)	0.1287
	21-40	126 (5.0%)	26 (3.8%)	152 (4.8%)	
	41-60	340 (13.5%)	88 (12.9%)	428 (13.4%)	
	61-80	963 (38.4%)	290 (42.5%)	1253 (39.2%)	
	81-100	1060 (42.2%)	268 (39.3%)	1328 (41.6%)	
	Missing	156	38	194	

Necrosis	none	636 (30.3%)	153 (26.0%)	789 (29.4%)	0.0839
	<=20%	1313 (62.6%)	386 (65.5%)	1699 (63.3%)	
	>20%	147 (7.0%)	50 (8.5%)	197 (7.3%)	
	Missing	571	131	702	
RNA Absorbance	High	1071 (96.8%)	290 (96.3%)	1361 (96.7%)	0.8096
	Low	35 (3.2%)	11 (3.7%)	46 (3.3%)	
	Missing	1561	419	1980	
Hybridization Window	Long	1165 (49.4%)	323 (50.9%)	1488 (49.7%)	0.5306
	Short	1195 (50.6%)	312 (49.1%)	1507 (50.3%)	
	Missing	307	85	392	
Percent of Genes above Limit of Detection	High	2554 (95.8%)	690 (95.8%)	3244 (95.8%)	1.0000
	Low	113 (4.2%)	30 (4.2%)	143 (4.2%)	
Signal to Noise Ratio	High	2533 (95.0%)	666 (92.5%)	3199 (94.4%)	0.0130
	Low	134 (5.0%)	54 (7.5%)	188 (5.6%)	

Survival

For each model, both *All Array* and *TCGA*, we calculated the overall (OS) and progression-free survival (PFS) once the models had been applied to NanoString data. Survival differences were significant between molecular subtypes using both models and reflective of previous reports, showing C2.IMM and C4.DIF with the best survival and C1.MES having the worst (Figures SB6 and SB7). The same pattern was also observed with the consensus between each approach (Figure SB8). Differences in survival between subtypes were also significant in samples that were discordant between models, however samples in this category did not follow the typical pattern (Figure SB9 and SB10).

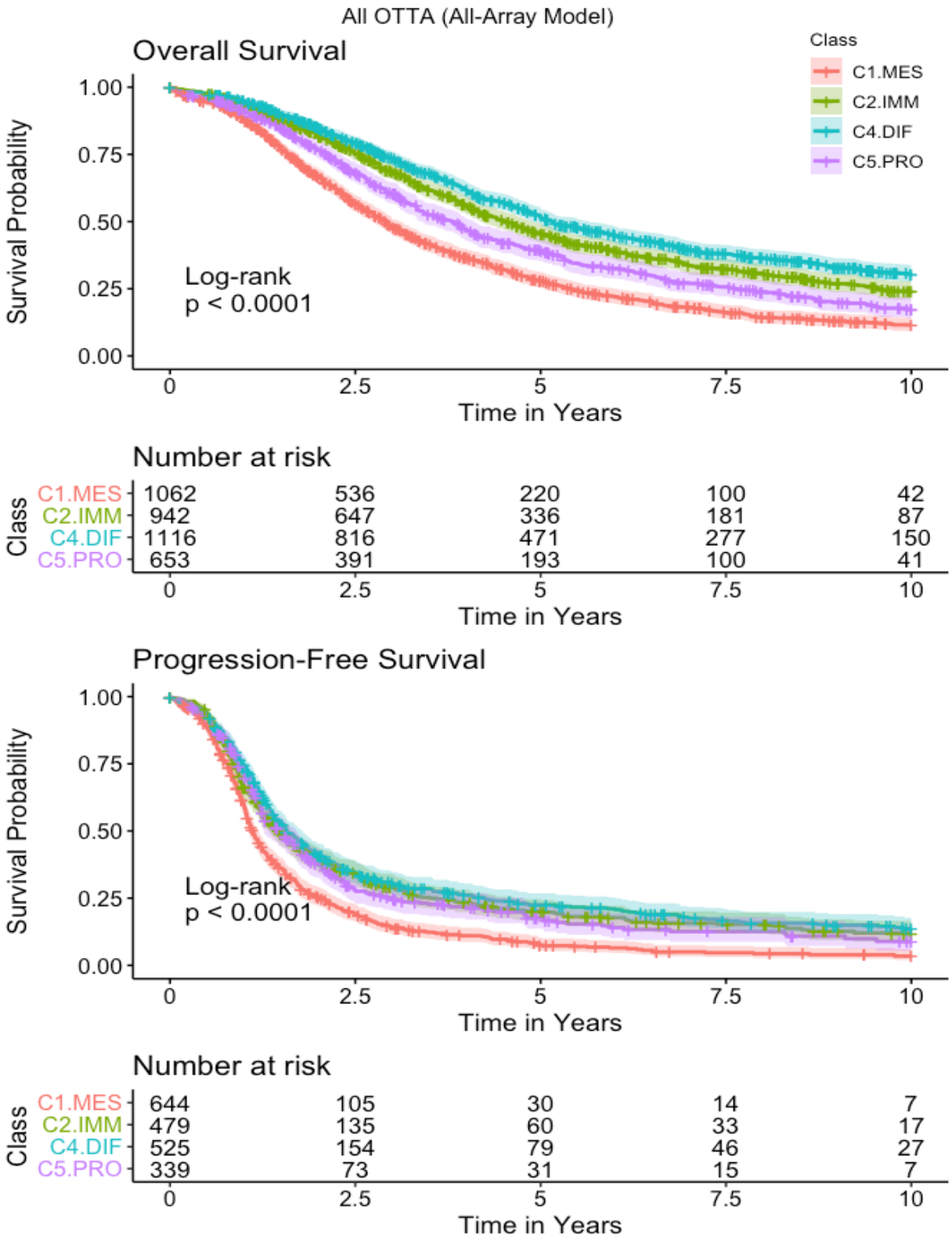

Figure SB6. Kaplan-Meier survival curves of predicted subtypes using the all-array model

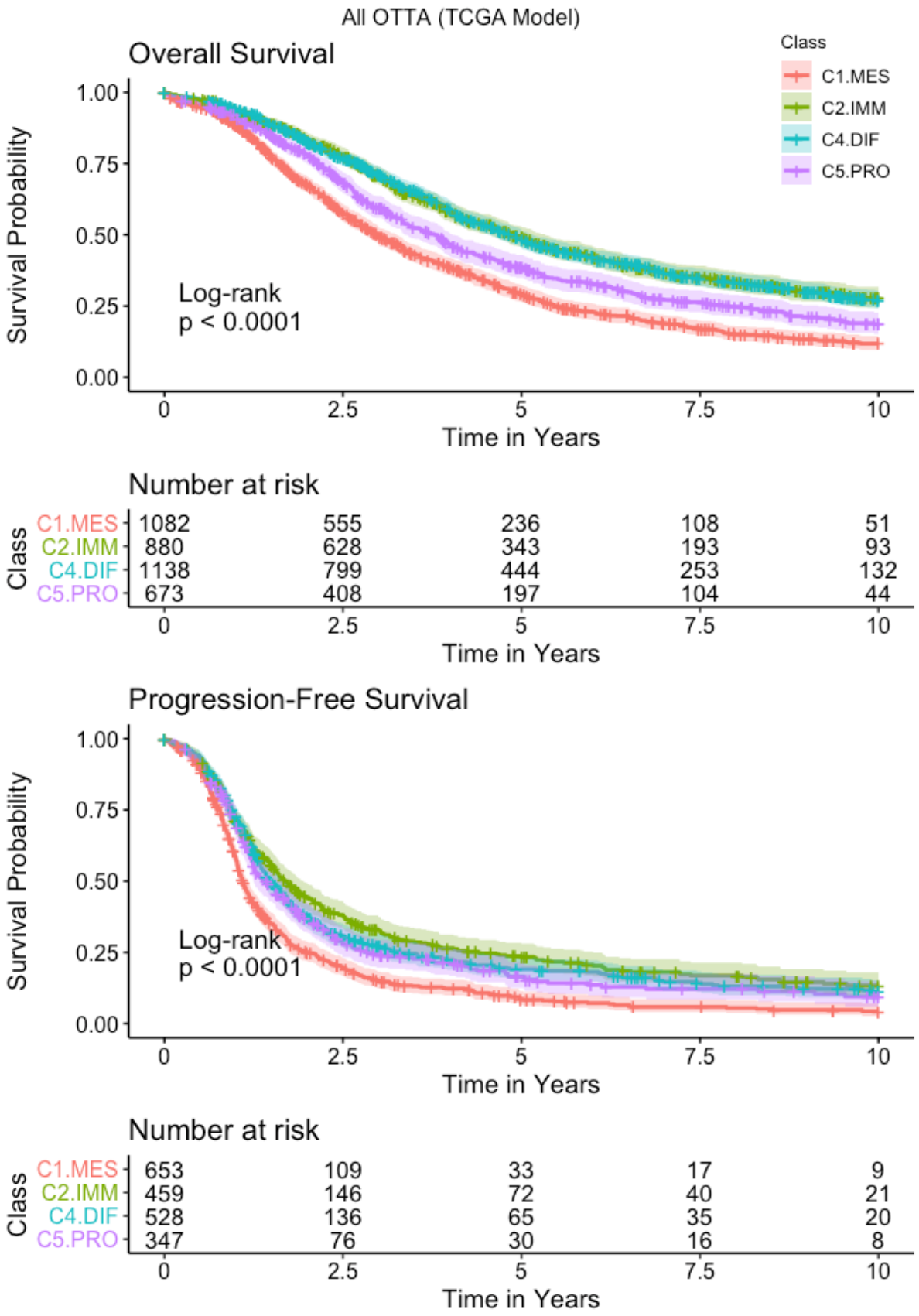

Figure SB7. Kaplan-Meier survival curves of predicted subtypes using the TCGA model

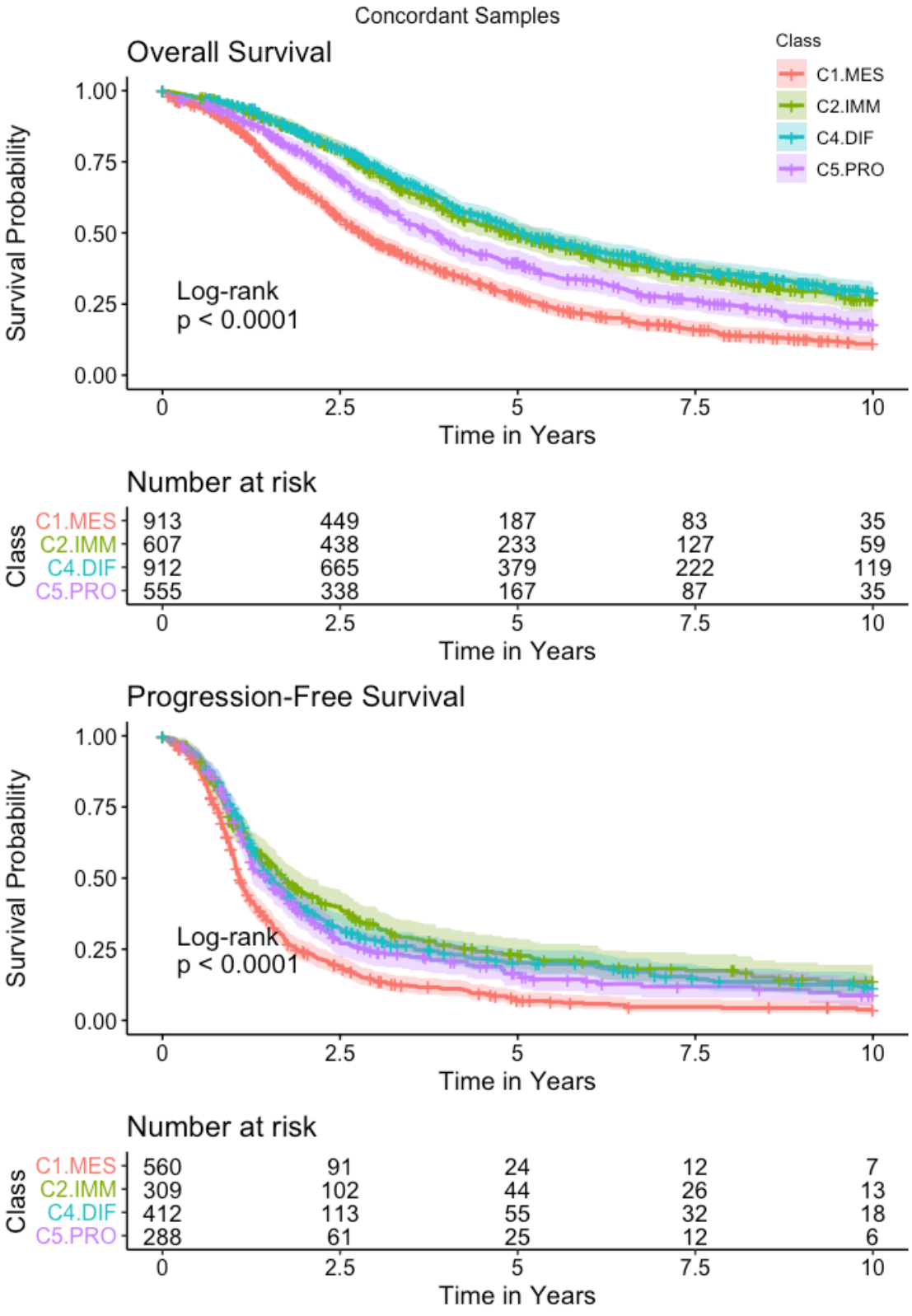

Figure SB8. Kaplan-Meier survival curves of predicted subtypes where the all-array and the TCGA models are in agreement

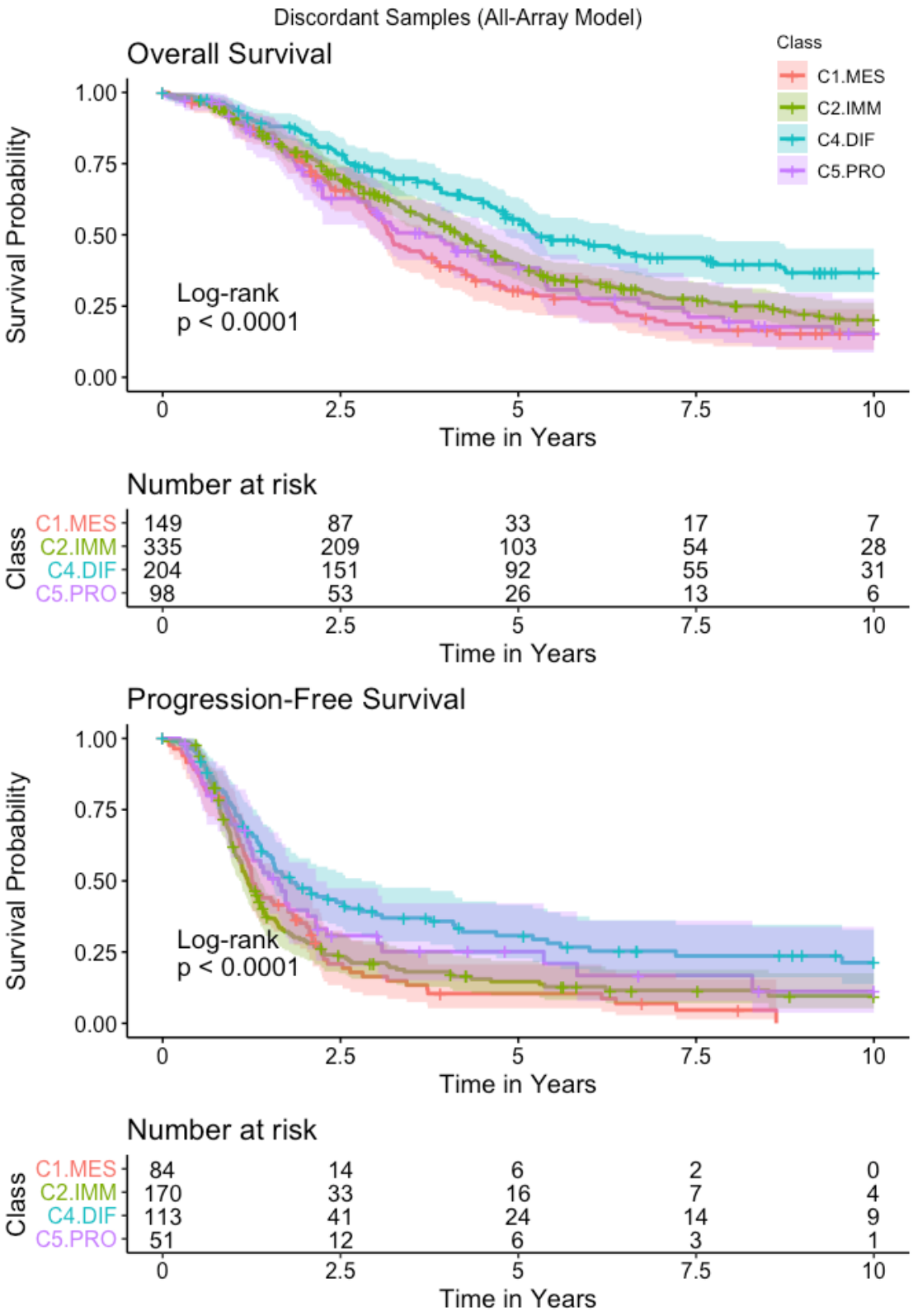

Figure SB9. Kaplan-Meier survival curves of predicted subtypes where the two models are in disagreement (labelled according to the all-array model)

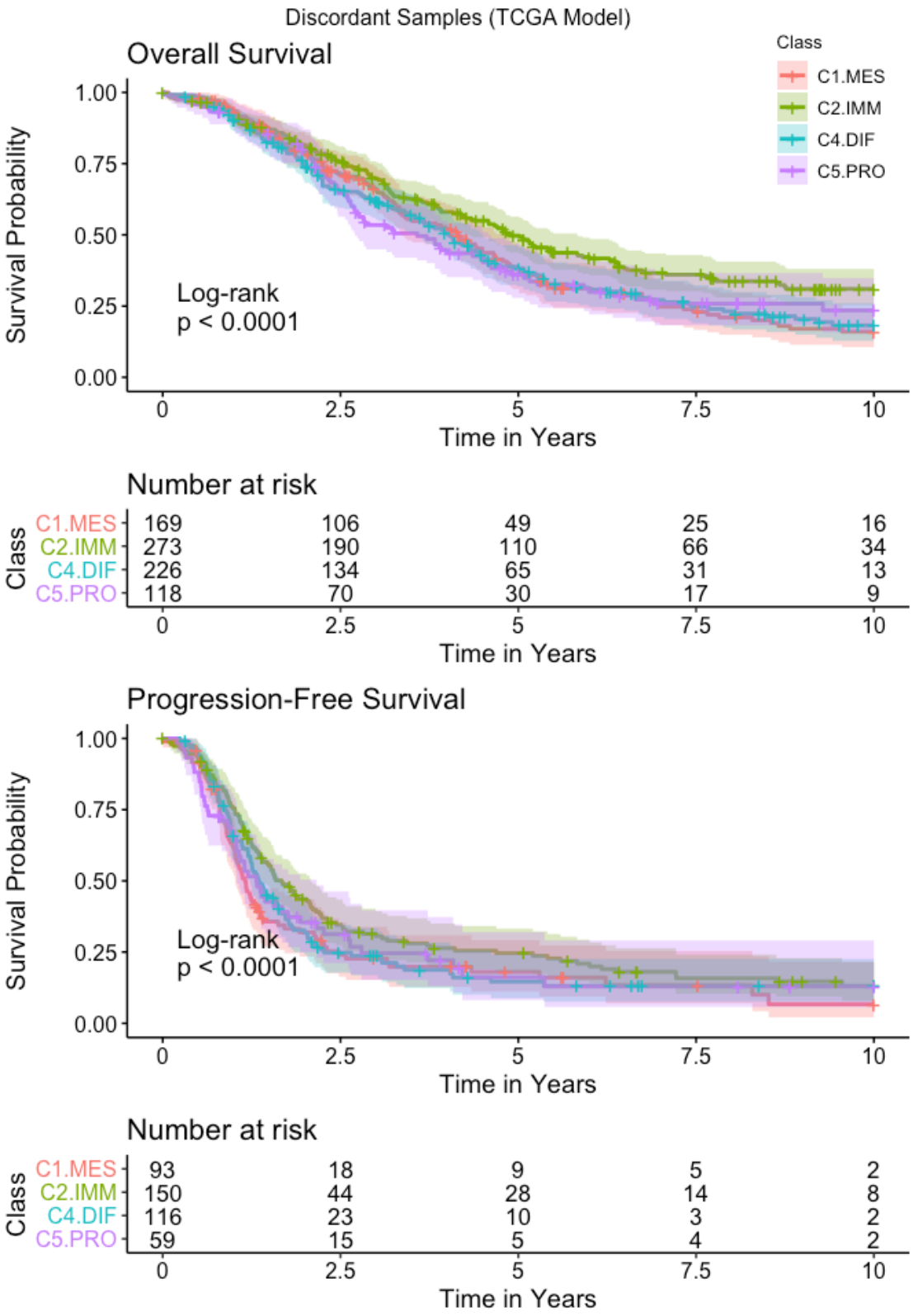

Figure SB10. Kaplan-Meier survival curves of predicted subtypes where the two models are in disagreement (labelled according to the TCGA model

C: Development of a minimal predictive model on NanoString Data

C.1 Data Breakdown for Downstream Analysis

We followed a series of best-practice guidelines to enable the training, confirmation, and validation of an optimized minimal gene prediction model(50). The NanoString data was split into 4 groups by studies, as outlined in Fig SA1, with the consensus labels derived from agreement between our two array-based strategies as the gold standard (79.1% agreement across all NanoString samples). Training Group 1, the largest group, was composed of 8 studies 1191 consensus labelled sample. This was used to train a minimal gene list model using the consensus labels (described below). Three subsequent validations were performed including confirmation and further optimization of the number of genes in Confirmation 1 (5 studies, 837 consensus labelled samples), and final validations in Validation 1 (4 studies, 719 consensus labelled samples) and Validation 2 (3 studies, 283 consensus labelled samples). The latter deliberately left as an independent group due to its composition of clinical trial studies.

C.2 Minimal Model Development

In order to train a minimal gene classifier, using the training data, consisting of 1135 samples with consensus labels, from 8 studies, we adopted a leave one study out validation scheme (see Fig SC1), where samples from each study are predicted using a model that was built using all the samples from all the other seven studies as illustrated in each row of in the figure below. This created 8 subsets of the data.

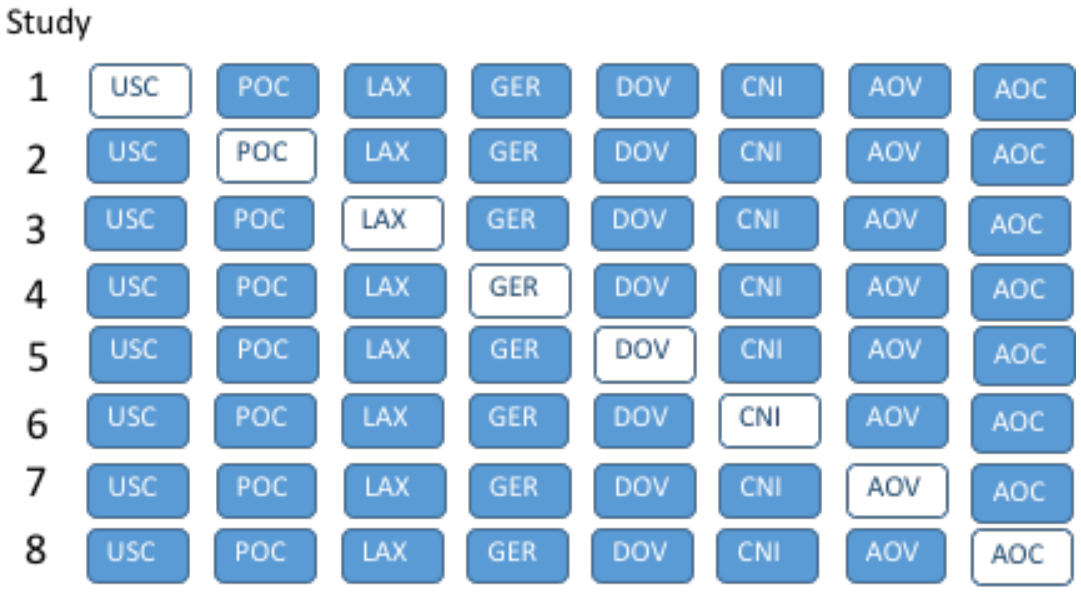

Figure SC1. Schematic diagram explaining the experimental design to derive the optimal gene list model

Within each partition, corresponding to a row in Fig SC1:

1. We resampled with replacement 500 times, using bootstrap approach, excluding the validation (left-out) study with white background.

    (i) On each of the 500 resamplings, we trained a model using all the genes and three different algorithms: Random Forest, Lasso, Adaboost.

    (ii) For each algorithm and each of the 500 resamplings, we computed a gene importance score by sorting the Gini feature importance, for Random Forest and Adaboost, and counting the genes with non-zero coefficients for Lasso.

2. We obtained a final ranking, by considering the proportion of times a gene is included in the top 100 most important genes (as ranked in 1 (ii) above). This resulted in 3 ranked lists, one for each algorithm (Table SC1).

3. Lastly, we obtained predictions of the left-out study (with the white background in Fig SC1) by going down each of the ranked lists from 2 (above), and fitting models of increasing complexity by increasing the numbers of genes (i.e. iterating over a sequence between 4 and 100, increasing in steps of 5 genes).

The aggregate accuracy was computed by concatenating the prediction of studies when in the left-out position (as described above) and comparing them with the consensus labels (Fig SC2 and SC3). Results show a clear advantage to the Lasso and Random forest algorithms. The Lasso and Random Forest algorithms both more or less plateau around ~ 30 genes whereas the Adaboost algorithm is significantly lower.

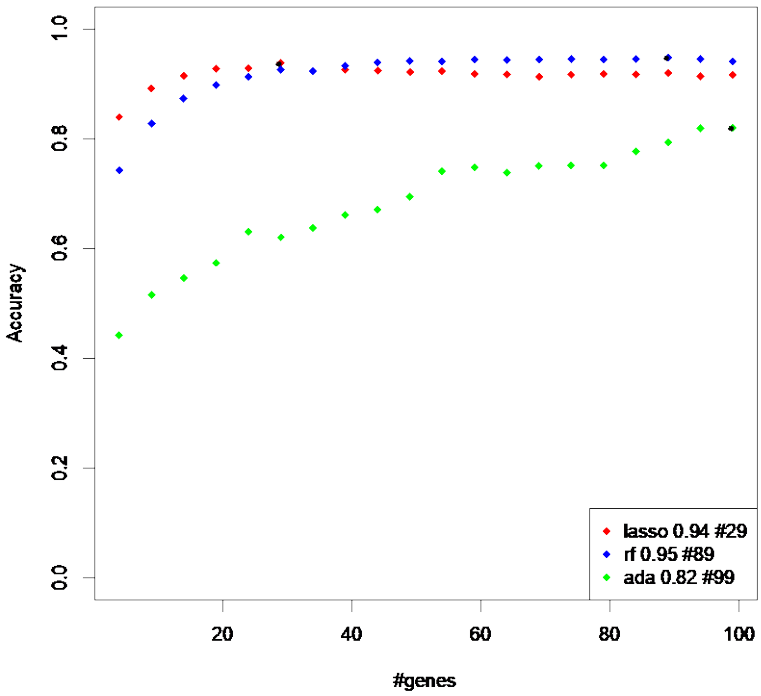

Figure SC2. The aggregate accuracy (for all samples in all studies) obtained by increasing number of genes and using the top n genes from each frequency list computed above, where n varied from 4 to 100. Note that the top n genes from each study are not necessarily the same.

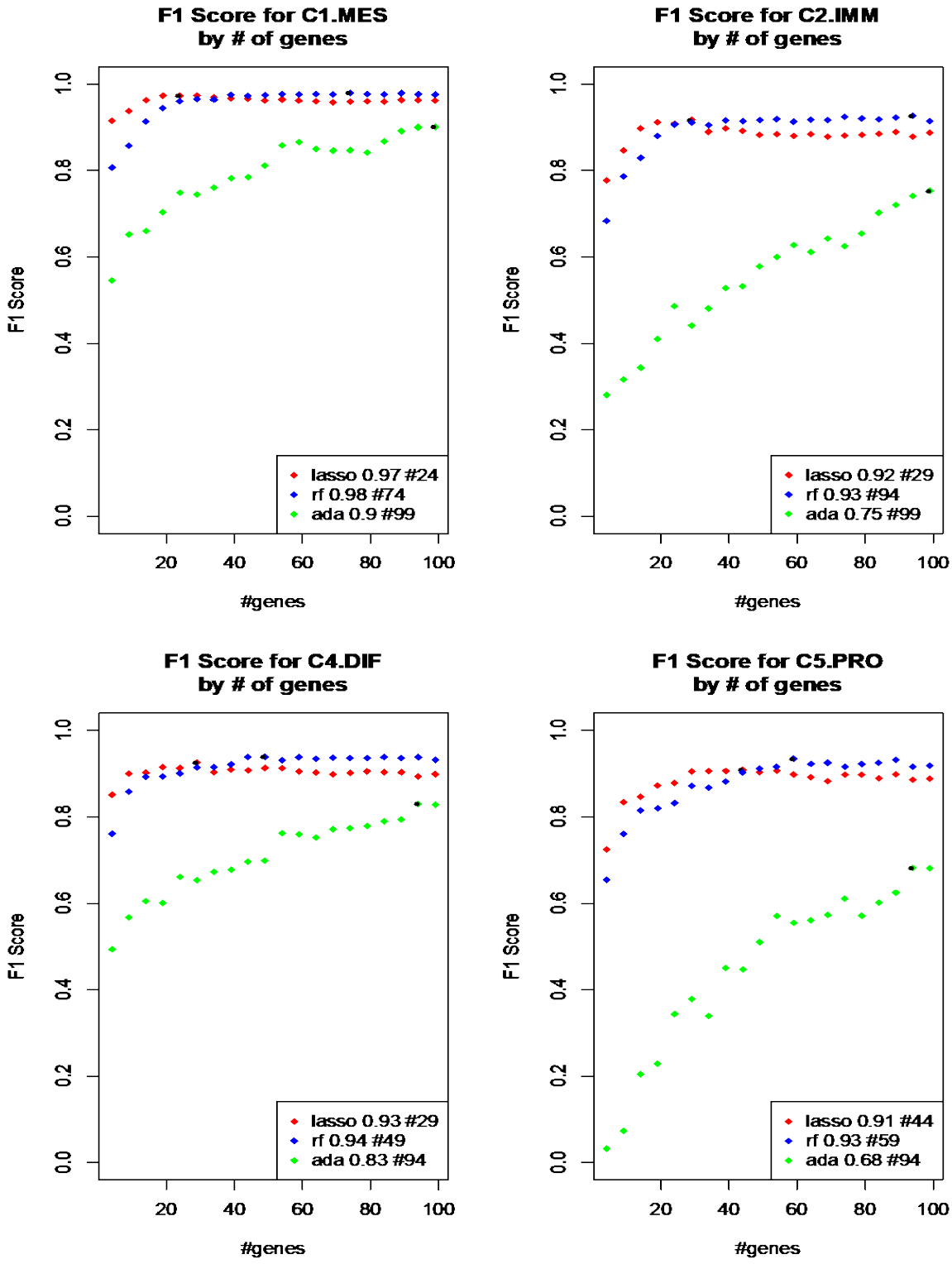

Figure SC3. The aggregate F1- score by subtype, describing a combined measure of sensitivity and positive predictive value, with increasing number of genes.

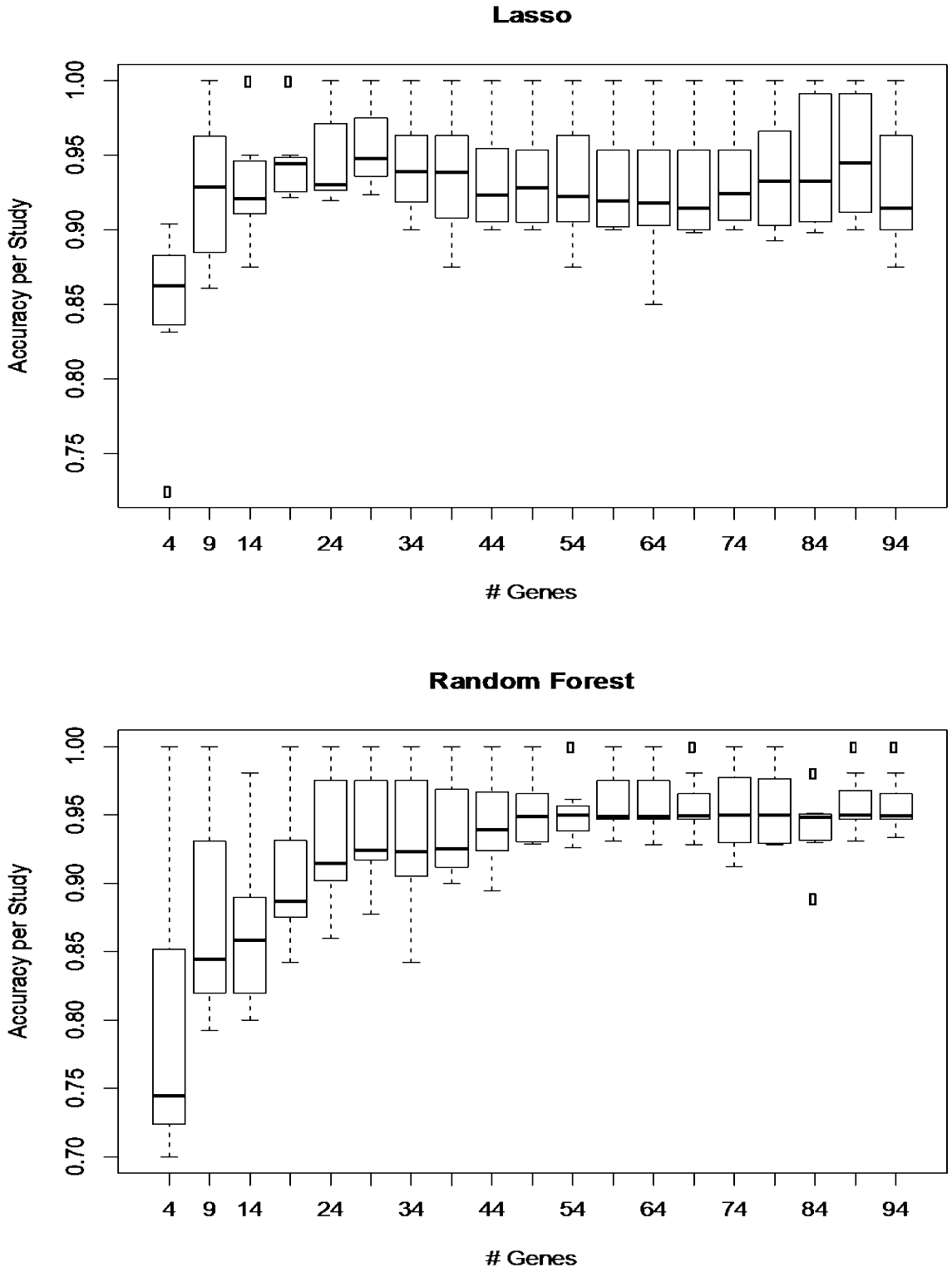

Figure SC4. Boxplots of the prediction accuracy by study using the Lasso and the Random forest algorithm. Each point in the boxplot corresponds to the individual study prediction (when left-out).

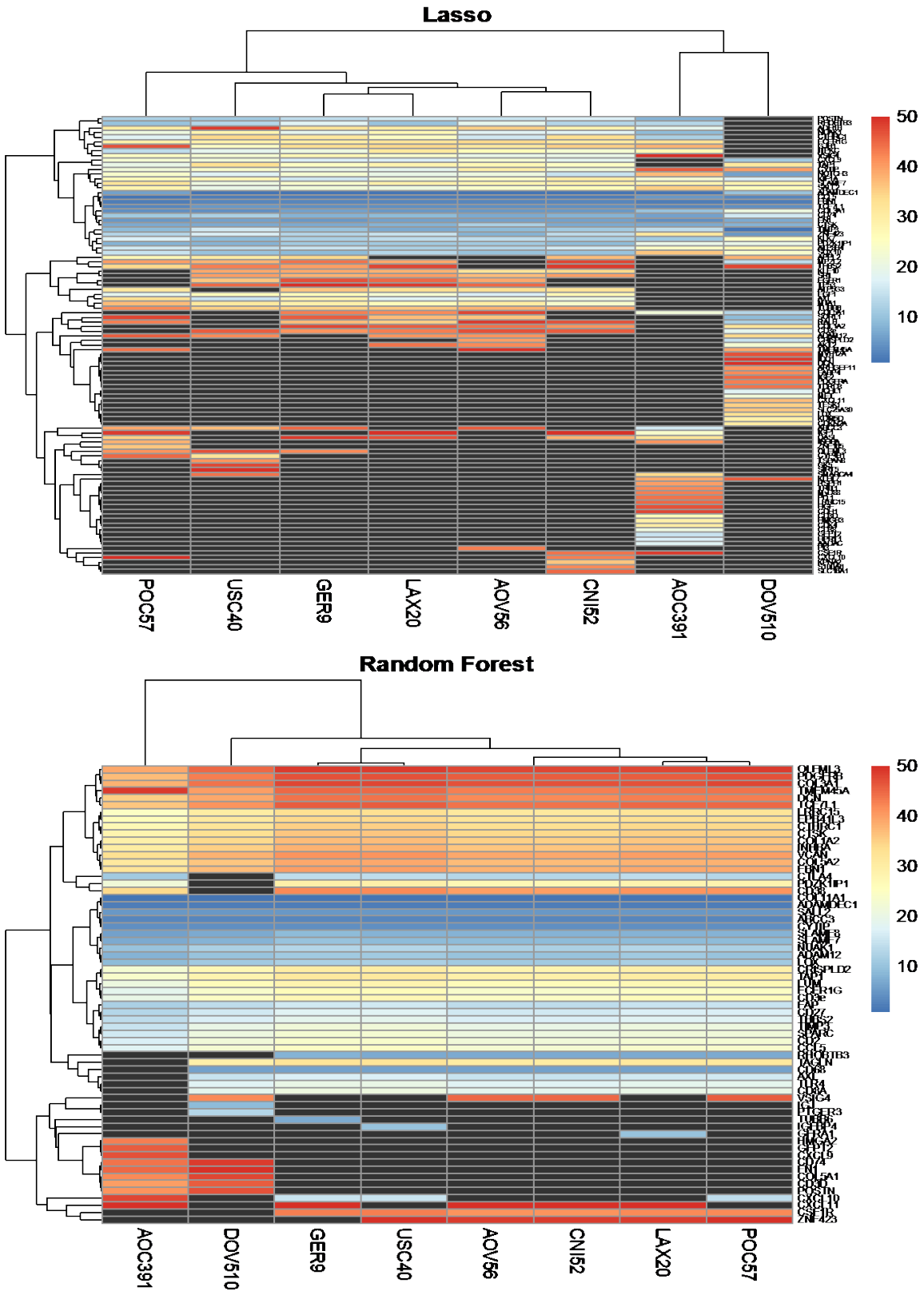

Figure SC5. Heatmap depicting the importance rank of the top 50 ranking genes from each study iteration, indicates that random forest showed a more stable ranking across predicted studies in comparison with Lasso.

Fig SC4 illustrates prediction accuracy by study with the Lasso and the Random forest algorithms. The variation in prediction accuracy across studies using the Lasso algorithm is generally constant as a function of increasing number of genes. In contrast, the Random Forest algorithm showed a much smaller variability across study especially as the number of genes used for modeling exceeded ~40.

Random Forest showed a more stable ranking across predicted studies in comparison with Lasso (Fig SC5). As such, the Random Forest algorithm was selected and an overall ranking was obtained by averaging the proportions for each gene across study iterations then sorting by the Random Forest ranking and then by Lasso to break ties.

C.3 First Confirmation

Using the final overall gene list ranking obtained in the previous step, we fit models of increasing complexity on the entire training set (group 1) by considering the top 40-78 genes from list by increasing the sequence one gene at a time. These models were then used to predict the confirmation set (group 2). The prediction accuracy in this group was excellent, even when using a small number of genes. The accuracy was negligibly increasing with the number of genes used in the model varying between 95% and 97% (Fig SC6).

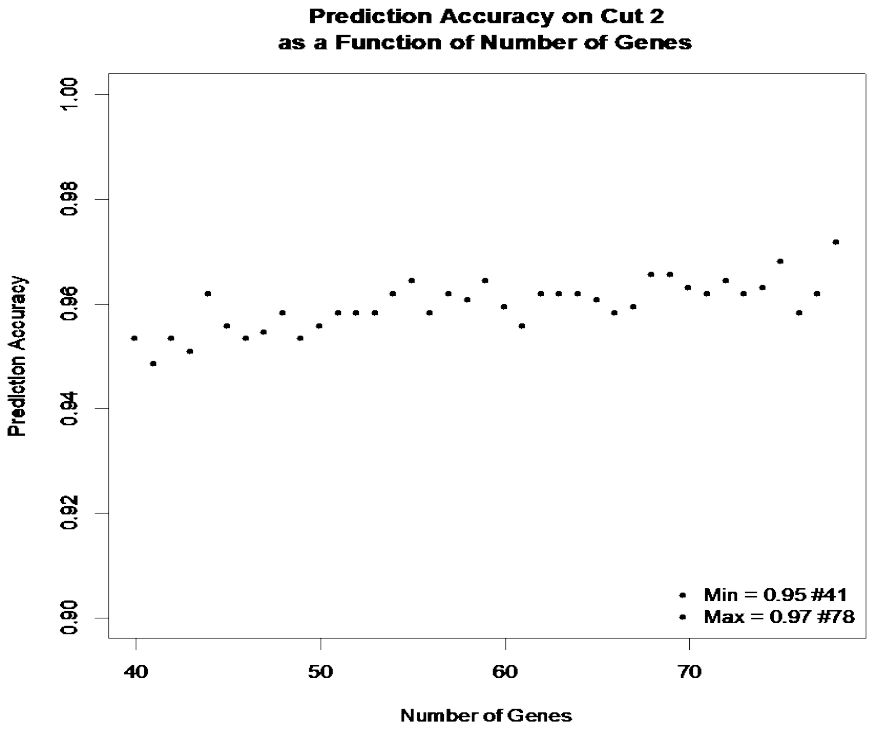

Figure SC6. Overall prediction accuracy in group 2

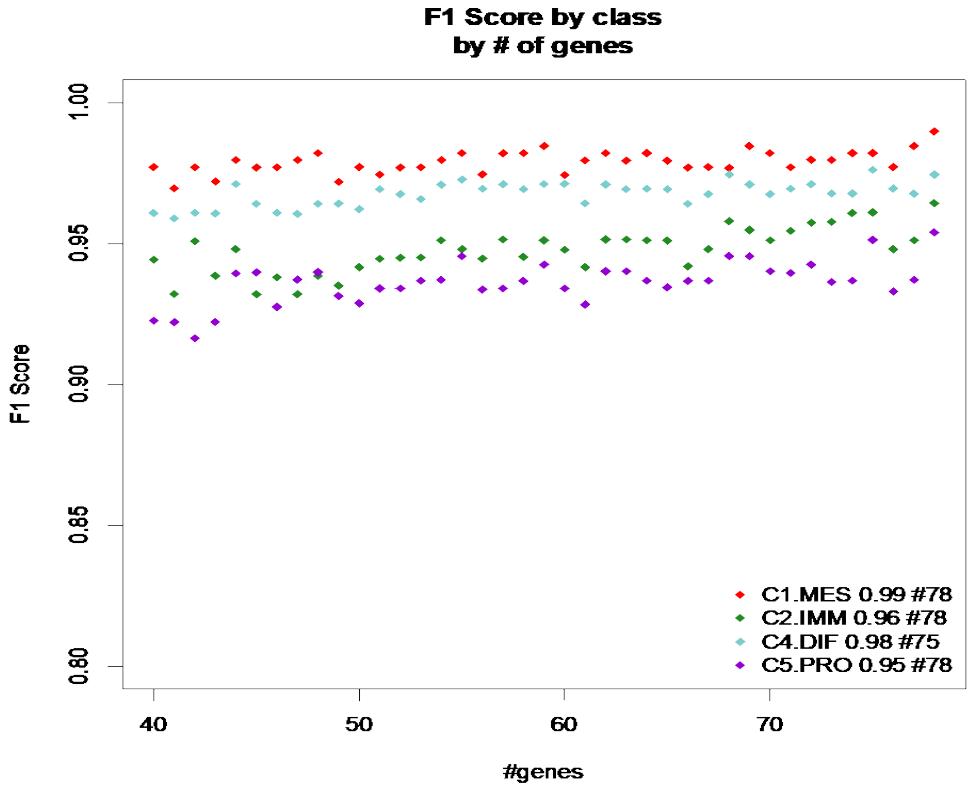

Figure SC7. F1-score in group 2 by subtype

The F1-score indicating the reliability of prediction by class mirrored the same pattern (Fig SC7) with C1.MES and C4.DIF having the highest prediction accuracy and C2.IMM and C5.PRO were lowest. All accuracy measures were above 95%.

Limited gains can be obtained by adding more genes. We pragmatically decided that 55 genes would be a safe cut-off option, to balance good accuracy and redundancy of genes to allow for potential probe-specific failures, without adding extensive complexity or cost to the final assay.

C.4 Validation

We selected the top 55 genes after removing one gene (CTHRC1) for which there was no probe available in the TCGA array platform. The omission of this gene was done to allow backward compatibility to this Affymetrix array platform and was tested to ensure its omission does not impact accuracy.

Tables SC2-SC6 compare the predictions from the final model with the consensus labels of the training set (group 1), the confirmation set (group 2) and the two validation sets (groups 3 and 4) and the set that overlaps the array respectively. Predictions remain very strong using the minimal gene set.

Samples that were not assigned a consensus label were predicted using the final model and the entropy compared with those sampled that had consensus. Figure SB6 indicated that on average those samples had significantly higher entropy in comparison with samples that had consensus labels (P value < 0.001, Mann-Whitney U Test).

Table SC2. Comparison of final model predictions with consensus labels from the training set (group 1)

A - Confusion Matrix

	Consensus Label			
Predicted with Final Model	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	391	0	0	0
C2.IMM	0	256	0	0
C4.DIF	0	0	319	0
C5.PRO	0	0	0	169

B - Overall Metrics

Metric		P value
accuracy	1 (1 - 1)	< 0.001
kappa	1 (1 – 1)	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy
C1.MES	1	1	1	1	1	0.34	1
C2.IMM	1	1	1	1	1	0.23	1
C4.DIF	1	1	1	1	1	0.28	1
C5.PRO	1	1	1	1	1	0.15	1

Table SC3. Comparison of final model predictions with consensus labels from the confirmation set (group 2)

A - Confusion Matrix

	Consensus Labels			
Predicted with Final Model	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	191	4	0	0
C2.IMM	1	147	1	6
C4.DIF	0	2	285	10
C5.PRO	1	1	4	164

B - Overall Metrics

Metric		P value
accuracy	0.96 (0.95 - 0.98)	< 0.001
kappa	0.95 (0.93 – 0.97)	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy

C1.MES	0.99	0.99	0.98	1.00	0.98	0.24	0.99
C2.IMM	0.95	0.99	0.95	0.99	0.95	0.19	0.97
C4.DIF	0.98	0.98	0.96	0.99	0.97	0.36	0.98
C5.PRO	0.91	0.99	0.96	0.98	0.94	0.21	0.95

Table SC4. Comparison of final model predictions with consensus labels from the first validation set (group 3)

A - Confusion Matrix

	Consensus Labels			
Predicted with Final Model	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	211	4	0	4
C2.IMM	3	145	8	2
C4.DIF	0	1	210	7
C5.PRO	1	0	4	119

B - Overall Metrics

Metric		P value
accuracy	0.95 (0.93 - 0.97)	< 0.001
kappa	0.94 (0.91 – 0.96)	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy
C1.MES	0.98	0.98	0.96	0.99	0.97	0.30	0.98
C2.IMM	0.97	0.98	0.92	0.99	0.94	0.22	0.97
C4.DIF	0.95	0.98	0.96	0.98	0.95	0.30	0.96
C5.PRO	0.90	0.99	0.96	0.98	0.93	0.17	0.95

Table SC5. Comparison of final model predictions with consensus labels from the second validation set (group 4)

A - Confusion Matrix

	Consensus Labels			
Predicted with Final Model	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	97	2	0	1
C2.IMM	1	47	1	3

C4.DIF	0	2	67	5
C5.PRO	0	0	1	56

B - Overall Metrics

Metric		P value
accuracy	0.94 (0.91 - 0.97)	< 0.001
kappa	0.92 (0.89 – 0.96)	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy
C1.MES	0.99	0.98	0.97	0.99	0.98	0.35	0.99
C2.IMM	0.92	0.98	0.90	0.98	0.91	0.18	0.95
C4.DIF	0.97	0.97	0.91	0.99	0.94	0.26	0.97
C5.PRO	0.86	1.00	0.98	0.96	0.92	0.20	0.93

Table SC6. Comparison of final model predictions with consensus labels from the overlap set

A - Confusion Matrix

	Consensus Labels			
Predicted with Final Model	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	25	0	0	0
C2.IMM	0	8	0	0
C4.DIF	0	0	23	0
C5.PRO	0	0	0	20

B - Overall Metrics

Metric		P value
accuracy	1 (0.95 - 1)	< 0.001
kappa	1 (1 – 1)	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy
C1.MES	1	1	1	1	1	0.33	1
C2.IMM	1	1	1	1	1	0.11	1
C4.DIF	1	1	1	1	1	0.30	1
C5.PRO	1	1	1	1	1	0.26	1

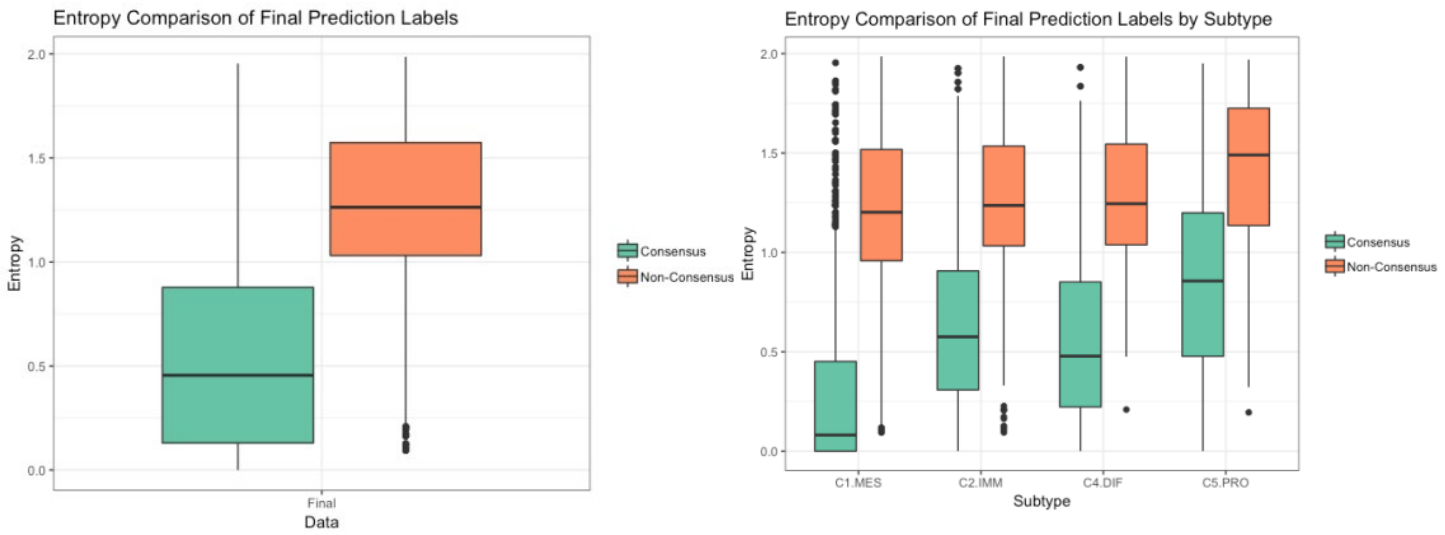

Figure SC8. Predictive entropy of the final model by consensus and non-consensus labels and by subtype.

C.5 Genes in the classifier

The final 55 genes (Table SC7) retained representation from many of the pathways originally identified as enriched (12,14-17,42,51). The majority of genes and pathway they are representative of could be grouped into 5 major clusters using functional annotation clustering (DAVID 6.7(52,53) and major annotation databases: BIOCARTA, KEGG_PATHWAY, PANTHER_MF_ALL, PANTHER_PATHWAY, and REACTOME_PATHWAY (Figure SC9 A). Herein we could observe pathway clusters dominated by extracellular matrix components, immune function and immune cell markers, surface receptors and kinases, chemokine and cytokine signalling, and cell morphology and angiogenesis related genes. NetworkAnalyst(54,55) visualization of Reactome(56,57) annotated genes also suggests the many of the represented pathways are interlinked (Figure SC9 B).

Figure SC9. Pathways represented in the minimal gene classifier represent much of the same biology as network and pathway analysis has previously revealed from genome-wide studies. (A) Major represented functional gene clusters in the final 55-gene classifier as defined by functional annotation clustering using the DAVID 6.7 database. (B) Visualization of Reactome annotated genes in a network diagram generated though NetworkAnalyst.

C.6 Gene Expression Distribution on Top Genes

NanoString

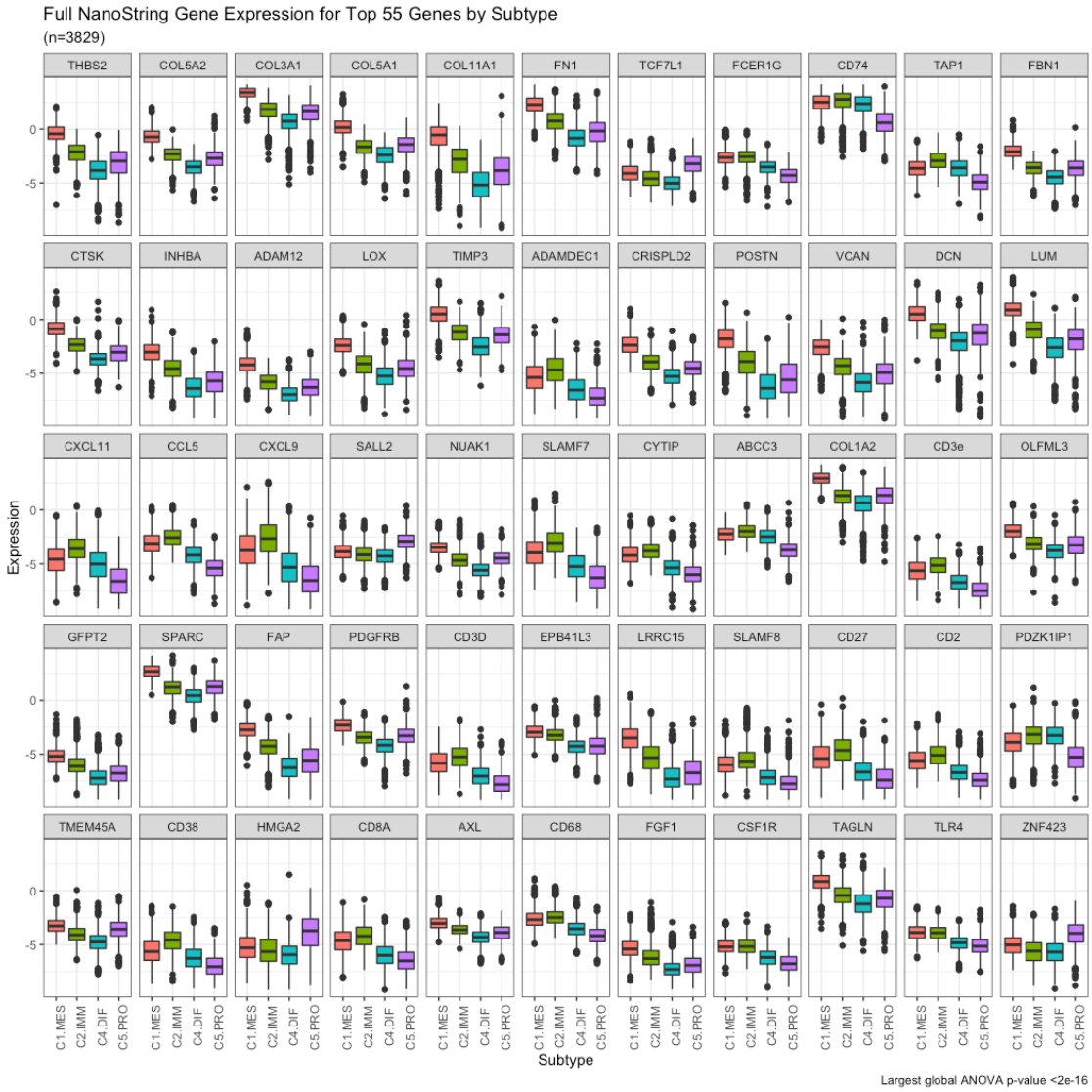

Figure SC10. Gene expression distribution on top genes in NanoString Data

All-Array

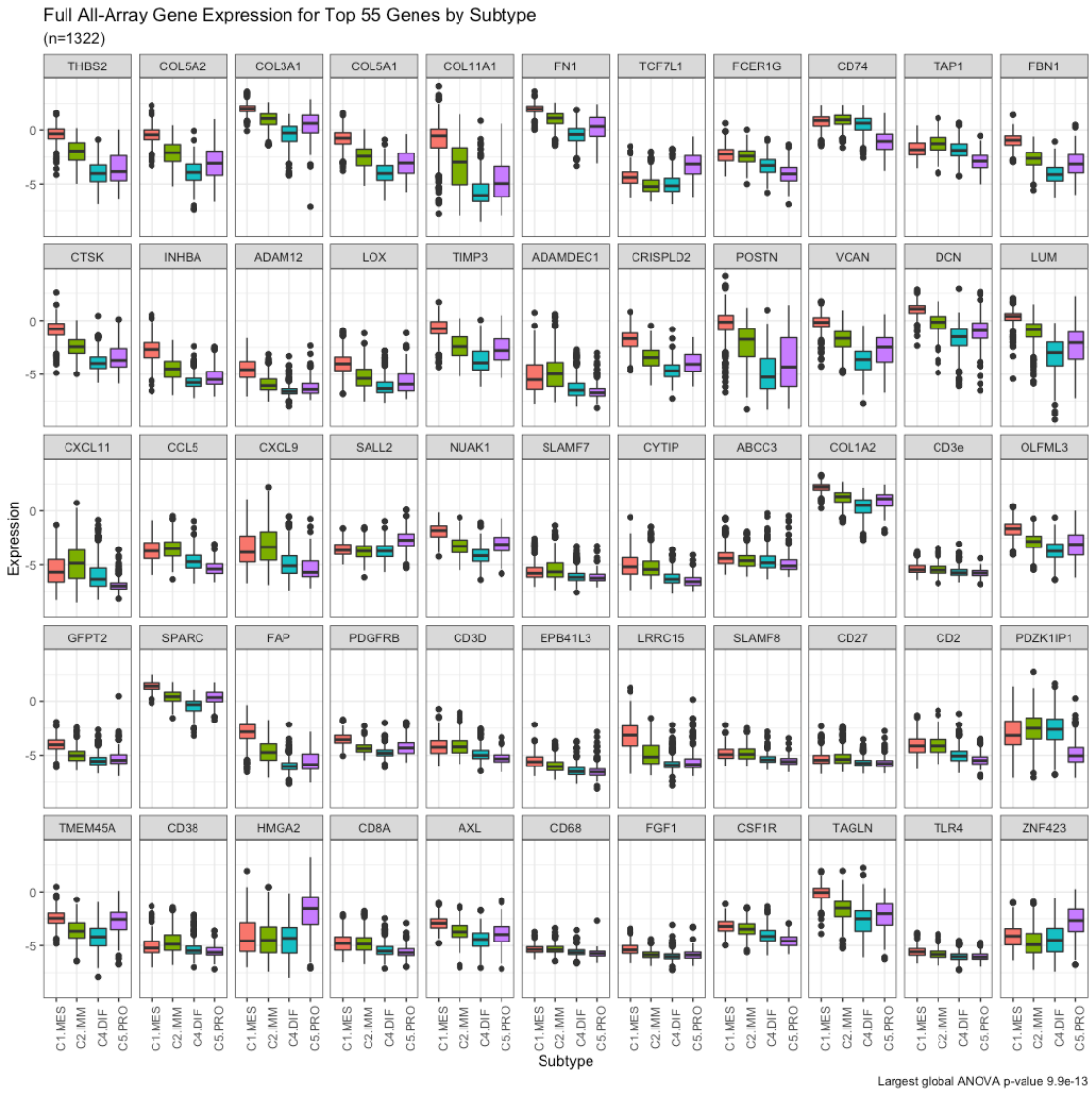

Figure SC11. Gene expression distribution on top genes in All-Array Data

TCGA

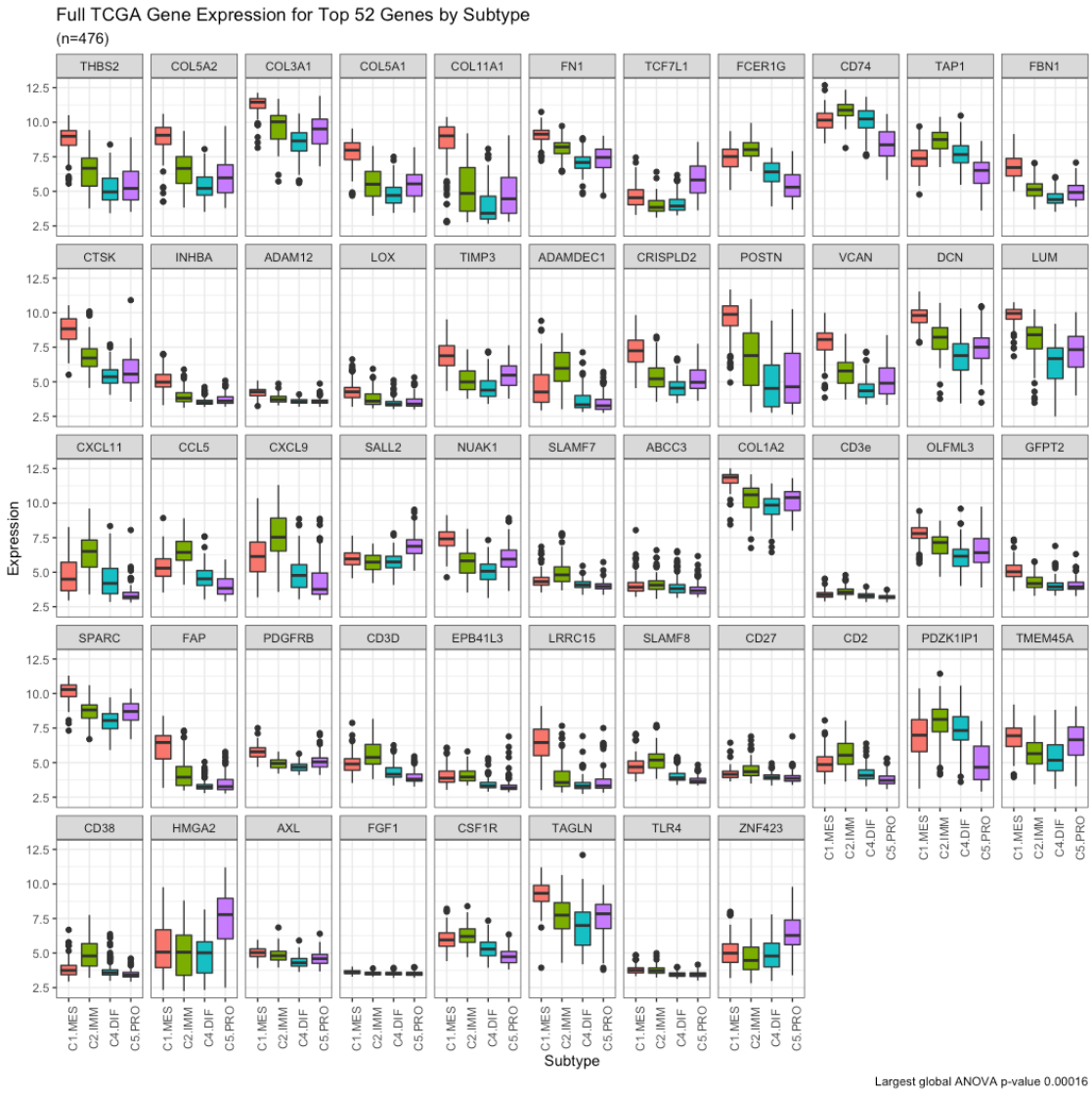

Figure SC12. Gene expression distribution on top genes in TCGA Data

D: Biological Correlates of Molecular Subtypes

D.1 Technical Variability and Potential Sources of Bias

Previous studies of NanoString data have suggested the platform is robust to most technical variability "within CodeSet"(21,22,58-60) and the reference-based strategy we have employed has been previously shown to correct for substantial processing bias across CodeSets, similar to the strategy currently employed in the NanoString Prosigna assay(22). Nonetheless, given the scale of our study a number of controls, replicates and sub-analyses was performed to address technical variability that may have been attributable to differences in processing at contributing sites, technical bias, real biological heterogeneity, and cross-CodeSet variability. Control subsets are also described in section B.5 with reproducibility of overall data across sites described in section B.6. The following sections will address if any remaining variability, post normalization to reference pools, affected prediction of HGSOC molecular subtypes in the final model. Experimental cohorts used in technical and biological control experiments, and replication within and across NanoString CodeSets is described in Fig SD1.

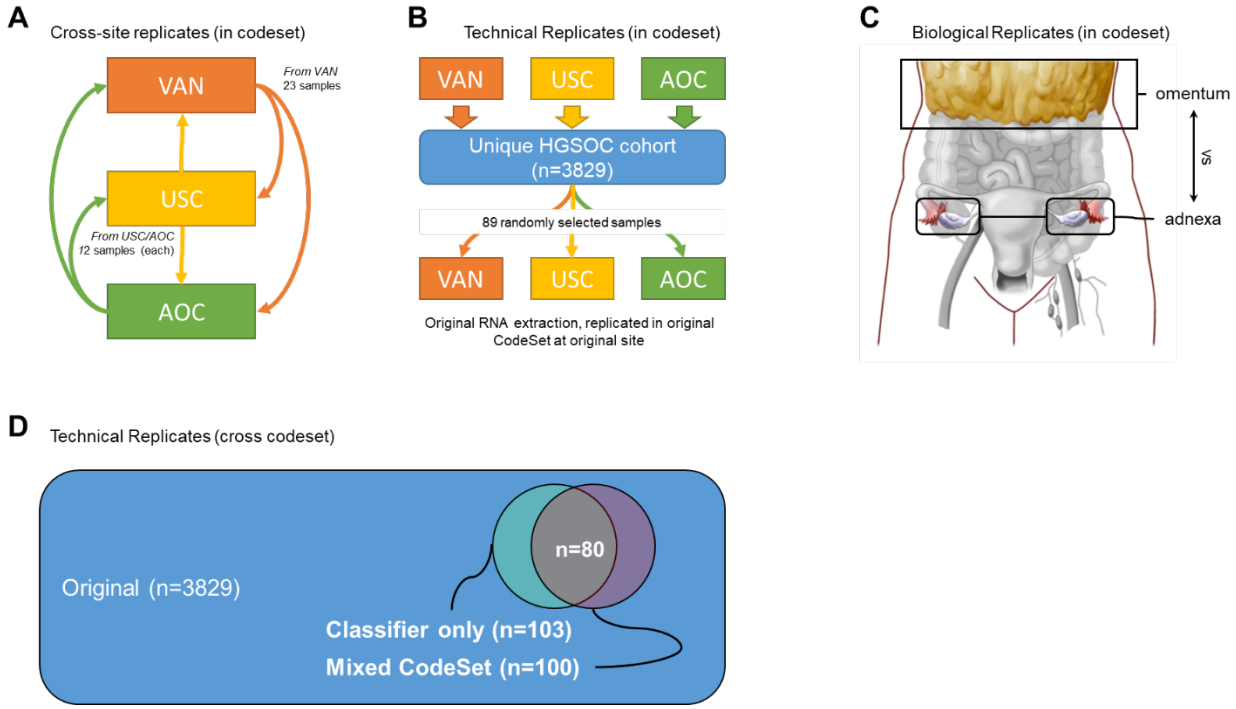

Figure SD1. Schematic representation of control experiments using technical or biological replicates to investigate potential sources of bias. (A) Illustrates our strategy of repeating sample across sites to ensure NanoString processing did not affect data output. Here each site (VAN = University of British Columbia, Vancouver; USC = University of Southern California, Los Angeles, and AOC = Australian Ovarian Cancer study group, Peter MacCallum Cancer Centre, Melbourne) selected samples at random and sent these to be rerun at the other two sites. 23 samples were selected from Vancouver (VAN) and 12 from each of AOC (Melbourne) and USC (Los Angeles) sites. Of these 47 samples, 141 output data files were generated and 35 are represented in the unique clinical data, 70 are replicates of these 35 from other run sites, and the remaining 36 are triplicate (12 samples run at each site) control only specimens. See also Figure SA1. (B) Illustrates controls for repeated measures using 89 technical replicates selected from the total pool of HGSOC samples. Samples were then re-run at their original co-ordinating site. (C) Illustrates our strategy to investigate biological bias related to the anatomical source of tissue being assayed. Here we examined 53 paired specimens taken from the same patient at the initial debulking surgery and compared molecular subtype prediction from adnexal sourced tissue to omental sourced tissue. While we attempted to source

the majority of our tissue from the site of origin (ovary/tube – adnexa) the omentum was the second most common source.

D.1.1 Within CodeSet Replication

We randomly selected 89 samples to repeat, to ensure consistency within CodeSets. Table SD1 shows very good within site replicability with overall accuracy at 94% and kappa at 0.92, 95% CI (0.86 – 0.99). Because these samples were selected at random they varied in their entropy. We compared the entropy upon replication (Fig SD2) and note there was disagreement, with the entropy larger as would be expected.

Table SD1. Comparison of within site replicates on 89 samples

A - Confusion Matrix

	Final Originals			
Predicted Replicates	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	21	0	0	0
C2.IMM	0	17	3	0
C4.DIF	0	0	26	1
C5.PRO	0	0	1	20

B - Overall Metrics

Metric		P value
accuracy	0.94 (0.87 – 0.98)	< 0.001
kappa	0.92 (0.86 – 0.99)	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy
C1.MES	1.00	1.00	1.00	1.00	1.00	0.24	1.00
C2.IMM	1.00	0.96	0.85	1.00	0.92	0.22	0.98
C4.DIF	0.87	0.98	0.96	0.94	0.91	0.30	0.92
C5.PRO	0.95	0.99	0.95	0.99	0.95	0.24	0.97

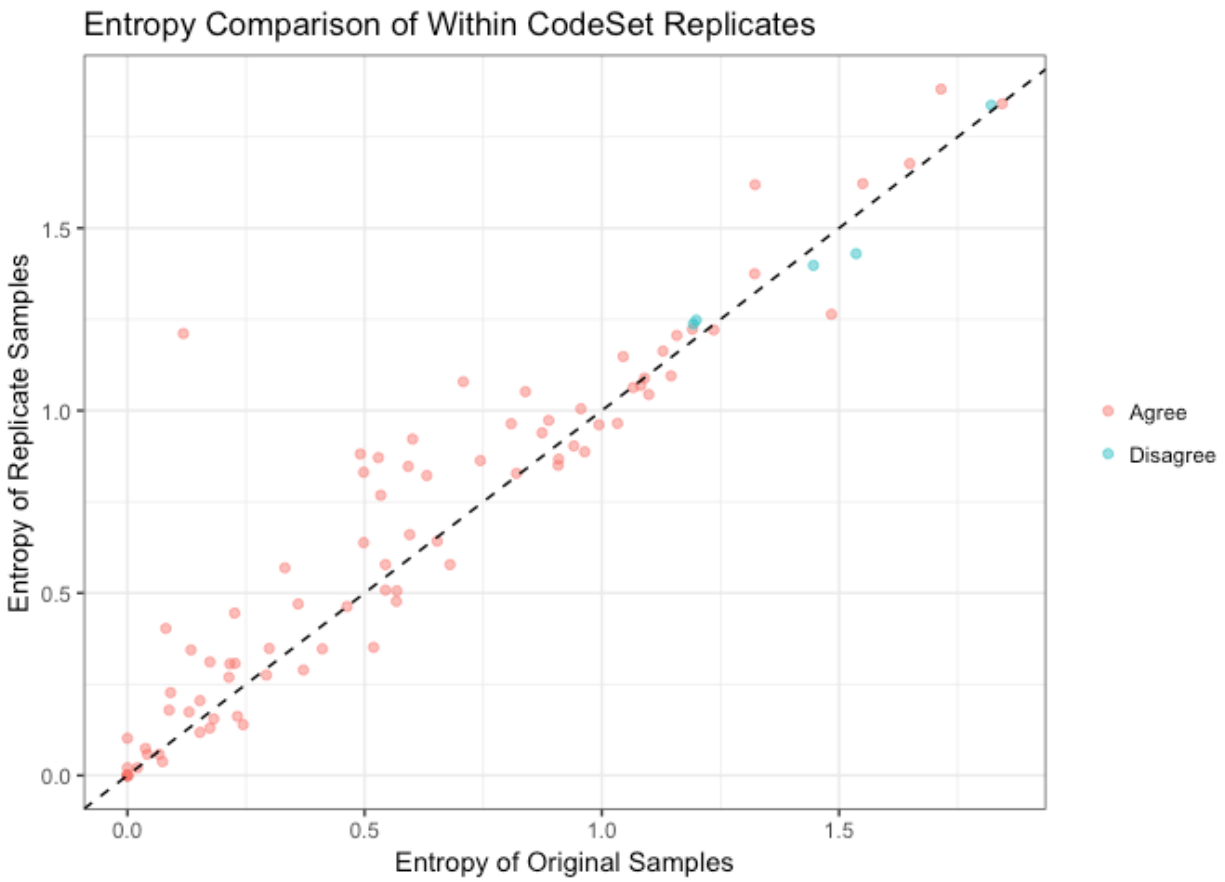

Figure SD2. Entropy of within CodeSet replicates. The blue dots indicate where classification disagree upon replication. The dashed line is the identity line.

D.1.2 Cross-site Replication

36 samples were randomly selected and repeated at the three different sites; a single replicate failed QC leaving 35 samples for analysis (see also Fig SD1). Classification results matched 100% across the three sites.

D.1.3 Cross CodeSet Replication

In order to evaluate consistent accuracy across CodeSets and our reference-based normalization, single-sample processing strategy, we replicated a number of samples across two additional CodeSets (Tables SD2 – SD4). One CodeSet, the Mixed CodeSet, contained probes for our 55 classifer genes, controls, and other probes unrelated to classification. The other CodeSet contained exclusively our 55 classifier gene probes and controls: “Classifier Only CodeSet” (see Fig SD1 - D). Accuracy was high in all pairwise comparisons of the three CodeSets, and the addition of additional genes did not impact prediction. This suggests in a research context that our classifier could be effectively combined with other research parameters without negative effect.

Table SD2. Comparison of predictions from the "OTTA original" CodeSet with the "OTTA Classifier Only" CodeSet (n=103)

A - Confusion Matrix

	Predicted Original Labels			
Predicted Classifier Only CodeSet Labels	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	37	0	1	0
C2.IMM	0	22	1	0
C4.DIF	1	0	26	0
C5.PRO	0	0	0	15

B - Overall Metrics

Metric		P value
accuracy	0.97 (0.92 - 0.99)	< 0.001
kappa	0.96 (0.91 - 1)	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy
C1.MES	0.97	0.98	0.97	0.98	0.97	0.37	0.98
C2.IMM	1.00	0.99	0.96	1.00	0.98	0.22	0.99
C4.DIF	0.93	0.99	0.96	0.97	0.95	0.26	0.96
C5.PRO	1.00	1.00	1.00	1.00	1.00	0.15	1.00

Table SD3. Comparison of predictions from the "OTTA original" CodeSet with the "OTTA Mixed" CodeSet (n=100)

A - Confusion Matrix

	Predicted Original Labels			
Predicted Mixed Codeset Replicate Labels	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	25	1	0	0
C2.IMM	0	20	0	0

C4.DIF	1	0	34	0
C5.PRO	0	0	1	18

B - Overall Metrics

Metric		P value
accuracy	0.97 (0.91 - 0.99)	< 0.001
kappa	0.96 (0.91 - 1)	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy
C1.MES	0.96	0.99	0.96	0.99	0.96	0.26	0.97
C2.IMM	0.95	1.00	1.00	0.99	0.98	0.20	0.98
C4.DIF	0.97	0.98	0.97	0.98	0.97	0.35	0.98
C5.PRO	1.00	0.99	0.95	1.00	0.97	0.19	0.99

Table SD4. Comparison of predictions from the "OTTA Mixed" CodeSet with the "OTTA Classifier Only" CodeSet (n=80)

A - Confusion Matrix

	Predicted Mixed CodeSet Labels			
Predicted Classifier Only CodeSet Labels	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	20	0	0	1
C2.IMM	1	15	1	0
C4.DIF	0	0	27	0
C5.PRO	0	0	0	15

B - Overall Metrics

Metric		P value
accuracy	0.96 (0.89 - 0.99)	< 0.001
kappa	0.95 (0.89 - 1.01)	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy

C1.MES	0.95	0.98	0.95	0.98	0.95	0.26	0.97
C2.IMM	1.00	0.97	0.88	1.00	0.94	0.21	0.98
C4.DIF	0.96	1.00	1.00	0.98	0.98	0.34	0.98
C5.PRO	0.94	1.00	1.00	0.98	0.97	0.19	0.97

Excellent agreement was observed between samples that were evaluated in the 78 samples (Fig SD1 – D) that were predicted in all three CodeSets (Fleiss' Kappa 0.95; P value <0.001). We also calculated Fleiss' Kappa for the 98 unique samples that were repeated in more than one CodeSet with missing values, ignoring repeats. In this case Fleiss' Kappa was a little weaker at 0.78 but still significant (P value <0.001).

D.2 Anatomical Site Variability

There were 53 samples (LAX n=25, VAN n=21, BRO n=7) that had paired adnexal and omentum samples which were both evaluated (Table SD5). We noted that 44/53 (83%) samples that were distributed across the subtypes when specimens were obtained from the adnexal became C1.MES when classified from specimens obtained from the omentum.

In examination of our clinical series of unique HGSOC samples we were also able to obtain anatomical sampling site information on 2182 samples (Table SD6), and the rest were presumed adnexal. This presumption appears valid as the comparison of subtype distribution with known adnexal sites was not significant (Table SD7, Chi-Squared P value=0.09).

The vast majority of samples were adnexal (n=1740; defined as ovary and/or tube, or adnexal without further detail), however the second most common sampling site was the omentum (n=355). A bias in C1.MES subtype prediction was significant (Table SD7, Chi-Squared P value <0.001) in the omentum sampling: 71.5% C1.MES as opposed to 24.6% C1.MES from known adnexal samplings (Table SD6). Specimens were presumed adnexal based on our request to contributing studies for treatment naïve, primary ovarian or tubal high-grade serous tubo-ovarian carcinoma specimens, but where a specific sampling site was not reported by the contributing study. Additional sites of sampling were accepted if defined: peritoneal specimens included all peritoneal sites, including peritoneal lymph nodes, unless omentum was specifically denoted. In the latter case omentum was specified as the site of sampling. Upper gynecological tract included specimens acquired from uterine and/or cervical sites where the diagnosis was still consistent with primary high-grade serous tubo-ovarian carcinoma. Lower gynecological tract included specimens acquired from the vagina.

Table SD5. Comparison of anatomical site, 53 matched pairs adnexa and omentum

A - Confusion Matrix

	Final Adnexal			
Predicted Omentum	C1.MES	C2.IMM	C4.DIF	C5.PRO
C1.MES	14	7	15	8

C2.IMM	0	2	3	3
C4.DIF	0	1	0	0
C5.PRO	0	0	0	0

B - Overall Metrics

Metric		P value
accuracy	0.3 (0.18 - 0.44)	0.76
kappa	0.06 (-0.1 - 0.23)	

C - By-Class Metrics

Class	Sensitivity	Specificity	PPV	NPV	F1	Detection Prevalence	Balanced Accuracy
C1.MES	1.0	0.23	0.32	1.00	0.48	0.83	0.62
C2.IMM	0.2	0.86	0.25	0.82	0.22	0.15	0.53
C4.DIF	0.0	0.97	0.00	0.65	NA	0.02	0.49
C5.PRO	0.0	1.00	NA	0.79	NA	0.00	0.50

Table SD6. Molecular subtype by anatomical sampling site

	ADNEXAL	PRESUMED ADNEXAL	OMENTUM	LOWER GENITAL TRACK	UPPER GENITAL TRACK	PERITONEAL
C1.MES	429 (24.7%)	394 (23.9%)	256 (72.1%)	0 (0.0%)	8 (50.0%)	32 (45.7%)
C2.IMM	447 (25.7%)	389 (23.6%)	69 (19.4%)	1 (100.0%)	2 (12.5%)	16 (22.9%)
C4.DIF	550 (31.6%)	574 (34.9%)	23 (6.5%)	0 (0.0%)	2 (12.5%)	12 (17.1%)
C5.PRO	314 (18.0%)	290 (17.6%)	7 (2.0%)	0 (0.0%)	4 (25.0%)	10 (14.3%)
Total	1740 (45.4%)	1647 (43.0%)	355 (9.3%)	1 (0.0%)	16 (0.4%)	70 (1.8%)

Table SD7. Comparison of subtype distribution by anatomical site

Comparison	Statistic	P Value	df	Method
Adnexal vs Omentum	325.6	<0.0001	3	Pearson's Chi-squared test
Adnexal vs Presumed Adnexal	4.428	0.2188	3	Pearson's Chi-squared test
Adnexal vs Other	21.79	<0.0001	3	Pearson's Chi-squared test

D.3 Biological Characterization of the Subtypes

D.3.1 Correlation with Clinical and Pathological Parameters

The distribution of subtypes within pathogenic BRCA mutation status was relatively consistent (Table SD8). C1.MES and C4.DIF were the majority classes whether a wildtype or mutation was observed.

In known and presumed adnexal sites, the distribution of year of diagnosis was similar across all subtypes (Fig SD3).

We looked for univariable associations between subtype and clinicopathological parameters in known and presumed adnexal sites (Table SD9) and known adnexal sites only (Table SD10). Age at diagnosis, stage, residual disease, cellularity, necrosis, and CD8 were significant for both data sources (P value <0.001). Race was only significant when considering both known and presumed adnexal sites (P value=0.03).

Table SD8. Molecular subtype by pathogenic BRCA mutation.

Levels	all wildtypes	pathogenic BRCA1 mutation	pathogenic BRCA2 mutation	pathogenic BRCA1/BRCA2/NOS mutation
C1.MES	202 (30.1%)	45 (34.1%)	22 (30.6%)	68 (33.0%)
C2.IMM	151 (22.5%)	28 (21.2%)	13 (18.1%)	41 (19.9%)
C4.DIF	205 (30.6%)	50 (37.9%)	25 (34.7%)	76 (36.9%)
C5.PRO	113 (16.8%)	9 (6.8%)	12 (16.7%)	21 (10.2%)

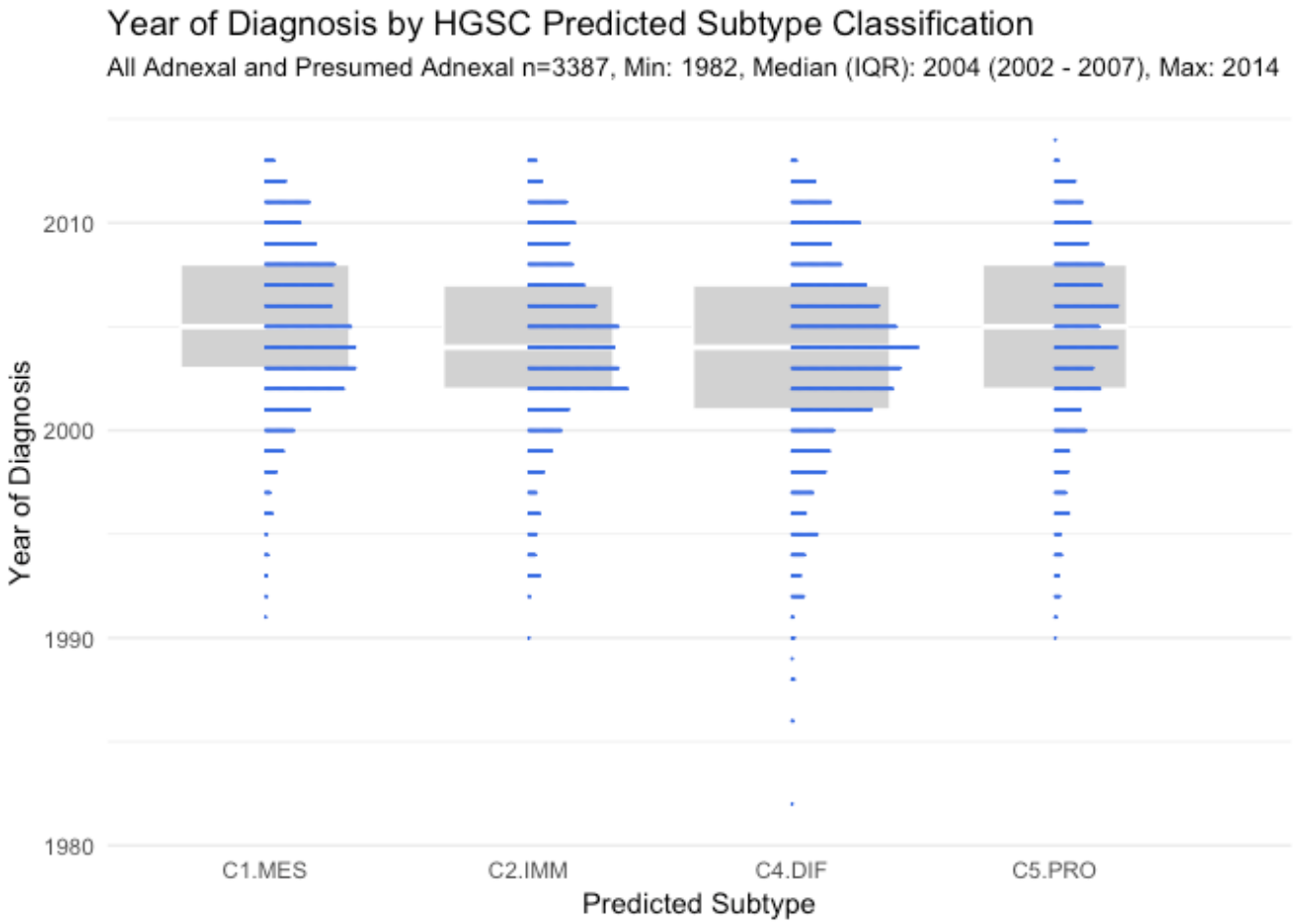

Figure SD3. Boxplot of year of diagnosis by subtype

Table SD9. Cohort characteristics for known and presumed adnexal sites by subtype

	C1.MES	C2.IMM	C4.DIF	C5.PRO	Total	PValue
N (%)	823 (24.3%)	836 (24.7%)	1124 (33.2%)	604 (17.8%)	3387 (100.0%)	
Age at Diagnosis						
Mean (sd)	61 (10.4)	60.4 (10.6)	57.8 (10.2)	62.7 (10.3)	60.1 (10.5)	< 0.0001
Median (IQR)	62 (54 - 68)	60 (53 - 68)	58 (50 - 65)	63 (56 - 70)	60 (53 - 67)	
Missing	13	15	22	15	65	
Stage						
Low	52 (6.5%)	164 (20.7%)	281 (26.4%)	95 (16.4%)	592 (18.3%)	< 0.0001
High	746 (93.5%)	627 (79.3%)	783 (73.6%)	484 (83.6%)	2640 (81.7%)	
Missing	25	45	60	25	155	
Residual Disease						

None	100 (27.5%)	143 (39.7%)	210 (43.4%)	116 (41.4%)	569 (38.2%)	< 0.0001
Any	264 (72.5%)	217 (60.3%)	274 (56.6%)	164 (58.6%)	919 (61.8%)	
Missing	459	476	640	324	1899	
Cellularity						
0-20	15 (1.9%)	10 (1.3%)	1 (0.1%)	6 (1.1%)	32 (1.0%)	< 0.0001
21-40	85 (10.7%)	33 (4.1%)	17 (1.6%)	17 (3.0%)	152 (4.8%)	
41-60	187 (23.6%)	129 (16.2%)	63 (6.1%)	49 (8.6%)	428 (13.4%)	
61-80	312 (39.3%)	329 (41.3%)	410 (39.7%)	202 (35.6%)	1253 (39.2%)	
81-100	195 (24.6%)	296 (37.1%)	543 (52.5%)	294 (51.8%)	1328 (41.6%)	
Missing	29	39	90	36	194	
Necrosis						
None	216 (31.7%)	178 (26.1%)	261 (30.5%)	134 (28.8%)	789 (29.4%)	0.0010
<=20%	432 (63.4%)	431 (63.2%)	542 (63.3%)	294 (63.1%)	1699 (63.3%)	
>20%	33 (4.8%)	73 (10.7%)	53 (6.2%)	38 (8.2%)	197 (7.3%)	
Missing	142	154	268	138	702	
BRCA1/BRCA2						
Wildtype	153 (79.7%)	134 (79.3%)	201 (74.2%)	111 (84.1%)	599 (78.4%)	0.1518
BRCA1	26 (13.5%)	24 (14.2%)	47 (17.3%)	9 (6.8%)	106 (13.9%)	
BRCA2	13 (6.8%)	11 (6.5%)	23 (8.5%)	12 (9.1%)	59 (7.7%)	
Missing	631	667	853	472	2623	
Race						
White	515 (85.8%)	493 (81.4%)	669 (81.3%)	354 (86.1%)	2031 (83.2%)	0.0523
Hispanic	82 (13.7%)	106 (17.5%)	150 (18.2%)	56 (13.6%)	394 (16.1%)	
Other	3 (0.5%)	7 (1.2%)	4 (0.5%)	1 (0.2%)	15 (0.6%)	
Missing	223	230	301	193	947	
CD8						
None	86 (21.2%)	19 (4.8%)	69 (11.6%)	132 (45.2%)	306 (18.1%)	<0.0001
Low	52 (12.8%)	19 (4.8%)	99 (16.6%)	65 (22.3%)	235 (13.9%)	
Med	210 (51.9%)	187 (47.1%)	297 (49.7%)	88 (30.1%)	782 (46.2%)	
High	57 (14.1%)	172 (43.3%)	132 (22.1%)	7 (2.4%)	368 (21.8%)	
Missing	418	439	527	312	1696	
Anatomical Site						
Adnexal	429 (52.1%)	447 (53.5%)	550 (48.9%)	314 (52.0%)	1740 (51.4%)	0.2188
Presumed adnexal	394 (47.9%)	389 (46.5%)	574 (51.1%)	290 (48.0%)	1647 (48.6%)	

Table SD10. Cohort characteristics for known adnexal sites by subtype

	C1.MES	C2.IMM	C4.DIF	C5.PRO	Total	PValue
N (%)	429 (24.7%)	447 (25.7%)	550 (31.6%)	314 (18.0%)	1740 (100.0%)	
Age at Diagnosis						
Mean (sd)	59.9 (9.4)	60.1 (10.6)	58.3 (10)	62.3 (9.9)	59.9 (10.1)	< 0.0001
Median (IQR)	61 (54 - 66)	60 (52 - 68)	58 (51 - 65)	63 (56 - 69)	60 (53 - 67)	
Missing	12	12	22	14	60	
Stage						
Low	28 (6.8%)	89 (21.0%)	141 (27.3%)	54 (18.2%)	312 (18.9%)	< 0.0001
High	383 (93.2%)	335 (79.0%)	376 (72.7%)	242 (81.8%)	1336 (81.1%)	
Missing	18	23	33	18	92	
Residual Disease						
None	26 (21.7%)	70 (40.7%)	100 (44.4%)	51 (40.8%)	247 (38.5%)	0.0004
Any	94 (78.3%)	102 (59.3%)	125 (55.6%)	74 (59.2%)	395 (61.5%)	
Missing	309	275	325	189	1098	
Cellularity						
0-20	7 (1.6%)	4 (0.9%)	0 (0.0%)	1 (0.3%)	12 (0.7%)	<0.0001
21-40	64 (15.0%)	18 (4.0%)	8 (1.5%)	11 (3.5%)	101 (5.8%)	
41-60	106 (24.8%)	82 (18.3%)	39 (7.1%)	33 (10.5%)	260 (15.0%)	
61-80	165 (38.6%)	179 (40.0%)	242 (44.1%)	109 (34.8%)	695 (40.0%)	
81-100	86 (20.1%)	164 (36.7%)	260 (47.4%)	159 (50.8%)	669 (38.5%)	
Missing	1	0	1	1	3	
Necrosis						
None	154 (39.6%)	115 (28.2%)	185 (36.1%)	97 (34.6%)	551 (34.7%)	0.0030
<=20%	215 (55.3%)	248 (60.8%)	295 (57.5%)	164 (58.6%)	922 (58.0%)	
>20%	20 (5.1%)	45 (11.0%)	33 (6.4%)	19 (6.8%)	117 (7.4%)	
Missing	40	39	37	34	150	
BRCA1/BRCA2						
Wildtype	68 (74.7%)	70 (78.7%)	77 (74.8%)	47 (87.0%)	262 (77.7%)	0.6937
BRCA1	16 (17.6%)	14 (15.7%)	18 (17.5%)	5 (9.3%)	53 (15.7%)	
BRCA2	7 (7.7%)	5 (5.6%)	8 (7.8%)	2 (3.7%)	22 (6.5%)	
Missing	338	358	447	260	1403	
Race						
White	260 (83.9%)	241 (81.1%)	304 (84.2%)	155 (83.8%)	960 (83.3%)	0.5907
Hispanic	48 (15.5%)	54 (18.2%)	57 (15.8%)	30 (16.2%)	189 (16.4%)	
Other	2 (0.6%)	2 (0.7%)	0 (0.0%)	0 (0.0%)	4 (0.3%)	
Missing	119	150	189	129	587	
CD8						
None	60 (20.1%)	12 (4.3%)	38 (10.5%)	89 (47.6%)	199 (17.7%)	< 0.0001
Low	35 (11.7%)	6 (2.2%)	55 (15.2%)	43 (23.0%)	139 (12.4%)	
Med	160 (53.5%)	134 (48.6%)	187 (51.5%)	54 (28.9%)	535 (47.6%)	
High	44 (14.7%)	124 (44.9%)	83 (22.9%)	1 (0.5%)	252 (22.4%)	
Missing	130	171	187	127	615	

We observe more events and longer follow-up for overall survival than progression-free survival (Table SD11).

Table SD11. Median Follow-up Time in Years and Events by Final Subtype

	Class	Observation Time	Censoring Time	Reverse KM	Events
OS	All Cases	2.9	2.5	8.1	2533
	C1.MES	2.3	2.1	7.7	830
	C2.IMM	3.4	2.7	8.1	572
	C4.DIF	3.5	2.8	8.5	696
	C5.PRO	2.8	2.4	7.9	435
PFS	All Cases	1.2	1.1	6.5	1710
	C1.MES	1.0	1.0	5.5	613
	C2.IMM	1.3	1.1	7.3	397
	C4.DIF	1.4	1.2	6.7	437
	C5.PRO	1.2	1.1	5.1	263

D.3.2 CD8 Tumor Infiltrating Lymphocyte Analysis

Previous studies have correlated molecular subtype with tumour infiltrating lymphocytes (TIL), in particular C2.IMM being the most enriched for TIL and C5.PRO being TIL-deficient(14,15). To corroborate the relationships between molecular subtype and immune response we examined CD8+ TIL in a subset of 1839 HGSOC tumors with existing data(11). In that previous study, $CD8^+$ TIL data were summarized as four tiers of scores according to TMA staining results (0=none, 1=weak, 2=moderate, and 3=high), and higher CD8 TIL score were significantly associated with better outcome. This finding provided the largest validation, as well as histotype-specific analysis, of previous observations related to improved outcome and increased immune cell infiltration(11).

In a sensitivity analysis (Table SD12), we compared CD8-TIL score distributions with consensus (n=1443), and predicted subtyping results (n=1839). We found statistically significant associations of CD8 TIL data with consensus subtyping ($p<0.001$) and final subtyping data ($p<0.001$), according to chi-square test. Specifically, more than 45% of HGSOC tumors with high $CD8^+$ TIL levels were C2.IMM

subtype, validating previous findings, yet less than 6% of CD8 negative tumors were C2.IMM. In contrast, more than 40% of negative CD8 tumors were classified as C5.PRO subtype; less than 2% of the CD8 high tumors were classified as C5.PRO (Table SD12). Similar to previous reports the C4.DIF group had the second highest levels of immune infiltrate with the group representing ~35% of high CD8-TIL tumours, while this substantially lower (~10% less) than the proportion of C2.IMM tumours it was more than double the next highest subtype (C1.MES at ~16%). Suggesting this group is also likely influenced by immune infiltration.

Table SD12. Distribution of HGSOC Molecular Subtype by CD8+ TIL Level, according to consensus subtyping (n=1443) and predicted subtyping data (n=1839)

CD8 \ CONSENSUS SUBTYPE	**C1.MES**	**C2.IMM**	**C4.DIF**	**C5.PRO**	**P-VALUE (CHI SQUARE)**
0 = Negative	87 (29.90%)	16 (5.50%)	56 (19.24%)	132 (45.36%)	p<0.001
1 = Low	72 (32.29%)	9 (4.04%)	81 (36.32%)	61 (27.35%)	
2 = Moderate	198 (31.13%)	134 (21.07%)	229 (36.01%)	75 (11.79%)	
3 = High	48 (16.38%)	134 (45.73%)	106 (36.18%)	5 (1.71%)	
CD8 SCORE PREDICTED SUBTYPE	**C1.MES**	**C2.IMM**	**C4.DIF**	**C5.PRO**	**P-VALUE (CHI SQUARE)**
0 = Negative	94 (29.28%)	21 (6.54%)	69 (21.50%)	137 (42.68%)	p<0.001
1 = Low	79 (29.48%)	21 (7.84%)	103 (38.43%)	65 (24.25%)	
2 = Moderate	245 (29.20%)	200 (23.84%)	304 (36.23%)	90 (10.73%)	
3 = High	94 (29.28%)	21 (6.54%)	69 (21.50%)	137 (42.68%)	

Similar to above, we conducted multivariate cox-regression accounting for age, molecular subtypes, stage and CD8 together, and found that the extent of $CD8^+$ TIL levels remained an independent predictor of overall survival (Table SD13 and SD14) and elevated TIL was associated with improved survival in all subtypes. In addition, our model further suggested that subtype, independent of $CD8^+$ TIL infiltration, was also prognostic. For example, molecular subtype was independent associated with OS and PFS in multivariable analysis and suggested C2.IMM, C4.DIF, and C5.PRO all had superior outcome to C1.MES even without $CD8^+$ TIL infiltration (Table SD13 and SD14; second column in each, ref “no CD8 TIL”). This may suggest underlying molecular subtype features and $CD8^+$ TIL mediated immunity play joint roles in determine patients’ prognosis and improved outcomes in C2.IMM and C4.DIF cannot be fully attributed to $CD8^+$ TIL infiltration.

D.3.3 Survival analysis

We also wanted to see how the subtypes from the two locked-down models correlated with survival. When comparing the final model subtype predictions between all adnexal sites and known adnexal sites only, the Kaplan-Meier survival curves appear slightly more compact in the latter, for both overall survival (OS) and progression-free survival (PFS) (Fig SD4-SD5). We censored survival time at 10 years so patients with negligible differences at long follow-up times are not considered.

Multivariable survival analysis was conducted for OS and PFS for all adnexal sites and known adnexal sites only (Table SD13-SD16), and for complete cases (Table SD17-SD20). Hazard ratios are shown with significance at 5% level indicated by an asterisk. Because of prominent missingness in some covariates, we compared different models using a “forward-selection” style approach. Final predicted subtypes, age at diagnosis, and stage were included in all models. We then added CD8, residual disease, and BRCA mutation to the model one by one, in order of increasing number of missing observations. Subtype is significant in all models that have only age and stage. In OS, subtype is no longer significant when residual disease enters the model. In PFS, the same is true when CD8 enters the model. The proportional hazards assumption is met for the full model with all covariates, tested using scaled Schoenfeld residuals.

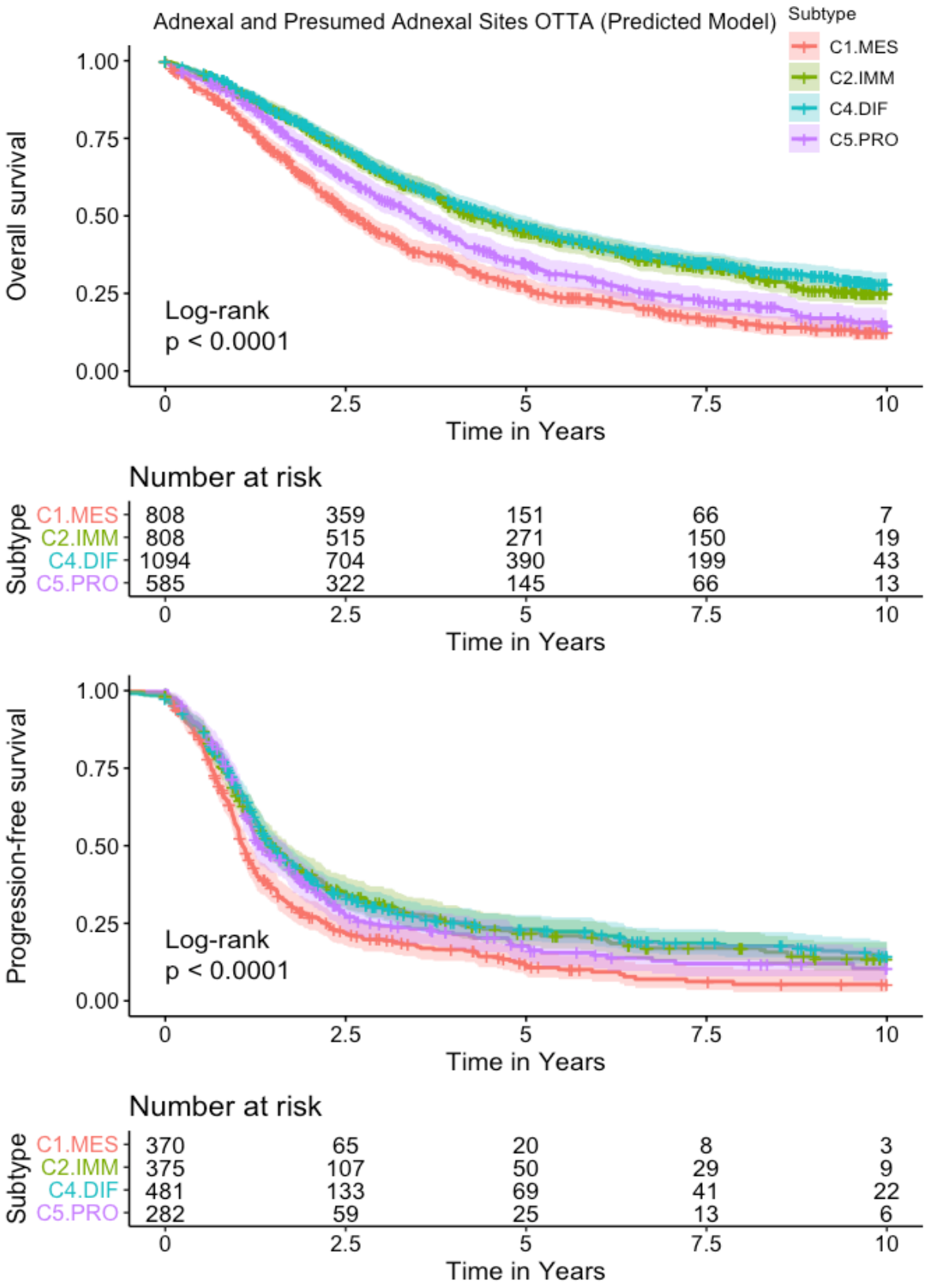

Figure SD4. Kaplan-Meier survival curves of predicted subtypes using final model on all adnexal and presumed adnexal sites

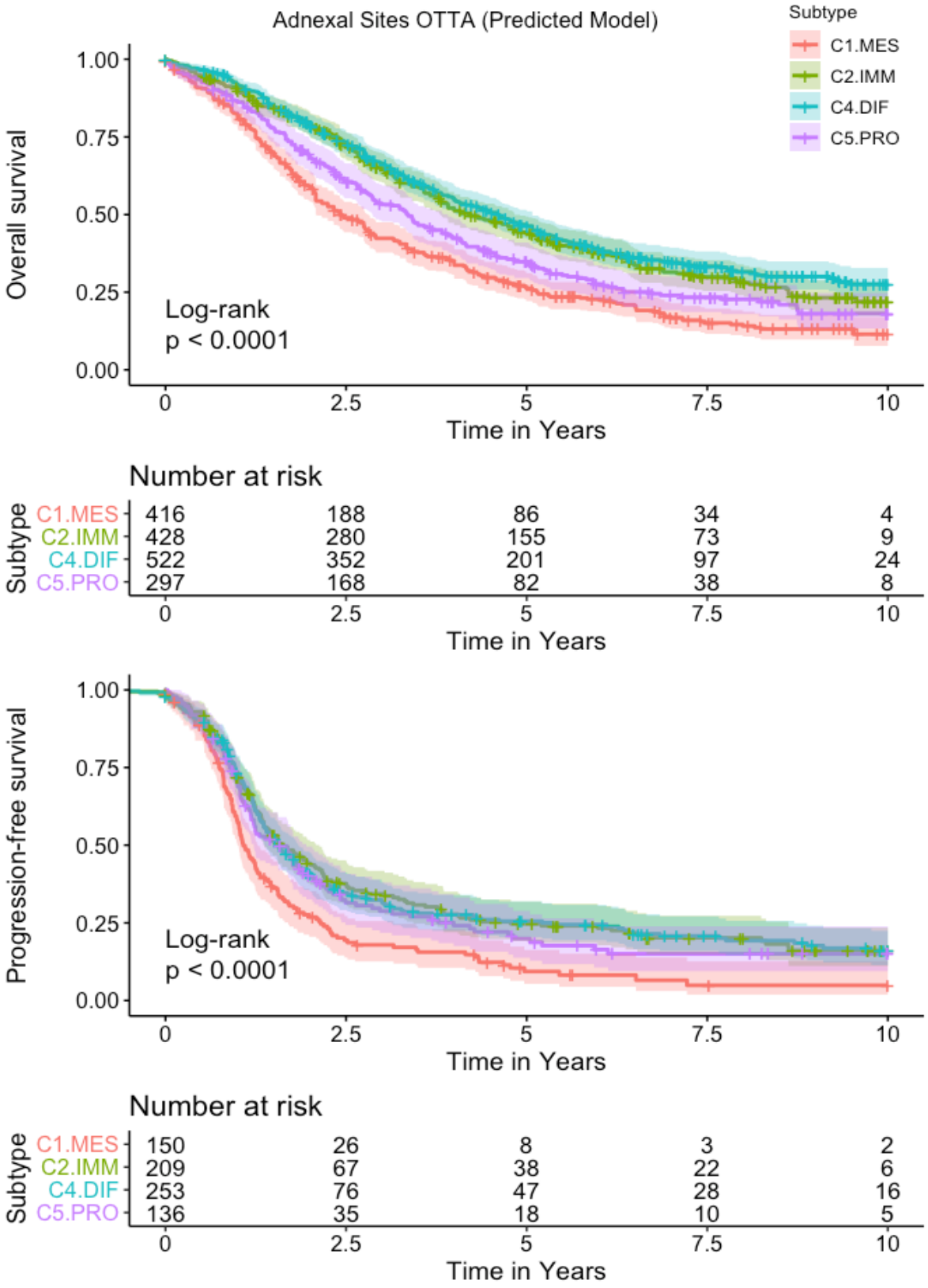

Figure SD5. Kaplan-Meier survival curves of predicted subtypes using final model on known adnexal sites

Table SD13. Multivariable survival analysis of overall survival, for known and presumed adnexal site

# of events / n	2137 / 3203	1154 / 1650	424 / 643	143 / 213
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Final Subtype (ref C2.IMM)				
C4.DIF	1.04 (0.92-1.17)*	0.93 (0.79-1.09)*	1.1 (0.83-1.44)	0.88 (0.52-1.48)
C5.PRO	1.21 (1.06-1.38)	0.91 (0.75-1.12)	1.23 (0.88-1.71)	0.88 (0.46-1.69)
C1.MES	1.41 (1.25-1.59)	1.25 (1.05-1.49)	1.33 (0.98-1.79)	0.99 (0.6-1.64)
Age	1.02 (1-1.03)*	1.03 (1.01-1.05)*	0.97 (0.94-1)	0.95 (0.89-1.01)
Stage (ref low)				
high	3.12 (2.7-3.61)*	3.5 (2.85-4.3)*	2.22 (1.55-3.18)*	2.03 (0.98-4.2)*
CD8 (ref none)				
low		1 (0.82-1.22)*	1.05 (0.75-1.47)	1.03 (0.56-1.91)*
med		0.82 (0.7-0.97)	0.87 (0.65-1.16)	0.65 (0.38-1.11)
high		0.65 (0.53-0.8)	0.77 (0.53-1.1)	0.37 (0.18-0.75)
Residual Disease (ref none)				
any			1.72 (1.37-2.17)*	2.07 (1.37-3.13)*
BRCA1/2 (ref wt)				
BRCA1				0.95 (0.56-1.62)*
BRCA2				0.22 (0.09-0.52)

Table SD14. Multivariable survival analysis of Progression-Free Survival, Known and Presumed Adnexal Sites

# of events / n	1138 / 1471	525 / 656	448 / 570	152 / 184
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Final Subtype (ref C2.IMM)				
C4.DIF	1.05 (0.9-1.24)*	1.01 (0.79-1.3)	1.16 (0.88-1.52)	1.04 (0.6-1.82)
C5.PRO	1.09 (0.91-1.31)	0.96 (0.71-1.29)	1.27 (0.92-1.77)	0.77 (0.4-1.5)
C1.MES	1.3 (1.1-1.54)	1.2 (0.92-1.58)	1.16 (0.86-1.56)	0.74 (0.44-1.25)
Age	1.18 (1.14-1.22)*	1.15 (1.1-1.21)*	1.12 (1.07-1.18)*	1.06 (0.97-1.16)
Stage (ref low)				
high	3.22 (2.57-4.02)*	3.32 (2.39-4.62)*	2.67 (1.86-3.84)*	3.86 (1.85-8.06)*
CD8 (ref none)				
low		1.17 (0.87-1.59)	1.26 (0.9-1.77)	1.24 (0.65-2.34)
med		0.98 (0.76-1.27)	1.06 (0.79-1.42)	0.86 (0.49-1.51)
high		0.76 (0.55-1.05)	0.98 (0.68-1.4)	0.84 (0.42-1.69)
Residual Disease (ref none)				
any			1.72 (1.39-2.13)*	1.83 (1.24-2.7)*
BRCA1/2 (ref wt)				
BRCA1				0.71 (0.41-1.21)
BRCA2				0.66 (0.34-1.26)

Table SD15. Multivariable survival analysis of Overall Survival, Known Adnexal Sites

# of events / n	1146 / 1628	813 / 1110	238 / 333	79 / 107
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Final Subtype (ref C2.IMM)				
C4.DIF	0.98 (0.84-1.15)*	0.9 (0.74-1.09)*	0.92 (0.63-1.32)	0.67 (0.32-1.4)
C5.PRO	1.16 (0.97-1.38)	0.91 (0.71-1.17)	1.09 (0.69-1.73)	0.57 (0.22-1.47)
C1.MES	1.38 (1.18-1.62)	1.2 (0.98-1.46)	1.09 (0.71-1.69)	0.69 (0.33-1.48)
Age	1.01 (0.99-1.03)	1.02 (0.99-1.05)	0.96 (0.92-1)	0.9 (0.81-1)
Stage (ref low)				
high	3.86 (3.12-4.77)*	4.11 (3.14-5.38)*	2.52 (1.55-4.1)*	2.18 (0.84-5.65)
CD8 (ref none)				
low		1.07 (0.83-1.37)*	1.11 (0.69-1.77)	0.88 (0.32-2.46)
med		0.78 (0.63-0.95)	0.72 (0.48-1.08)	0.71 (0.32-1.6)
high		0.71 (0.56-0.92)	0.72 (0.43-1.22)	0.26 (0.08-0.8)
Residual Disease (ref none)				
any			1.83 (1.33-2.54)*	3.84 (2.06-7.15)*
BRCA1/2 (ref wt)				
BRCA1				0.81 (0.39-1.68)*
BRCA2				0.13 (0.04-0.43)

Table SD16. Multivariable survival analysis of progression-free survival, known adnexal sites

# of events / n	570 / 718	301 / 369	253 / 317	84 / 99
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Final Subtype (ref C2.IMM)				
C4.DIF	1.12 (0.9-1.4)*	0.97 (0.7-1.34)	1.02 (0.7-1.47)	0.76 (0.34-1.71)
C5.PRO	1.17 (0.91-1.51)	0.87 (0.57-1.33)	1.04 (0.65-1.67)	0.35 (0.13-1)
C1.MES	1.52 (1.19-1.94)	1.27 (0.87-1.87)	1.14 (0.74-1.77)	0.68 (0.3-1.55)
Age	1.17 (1.12-1.22)*	1.14 (1.08-1.21)*	1.12 (1.06-1.18)*	1.07 (0.94-1.22)
Stage (ref low)				
high	3.15 (2.36-4.19)*	2.85 (1.91-4.25)*	2.22 (1.42-3.46)*	2.85 (1.1-7.4)*
CD8 (ref none)				
low		1.08 (0.71-1.64)*	0.95 (0.59-1.54)	1.47 (0.52-4.13)
med		0.76 (0.53-1.08)	0.71 (0.47-1.06)	0.66 (0.3-1.47)
high		0.52 (0.33-0.82)	0.58 (0.34-0.98)	0.31 (0.1-0.99)
Residual Disease (ref none)				
any			1.99 (1.46-2.7)*	2.67 (1.46-4.89)*
BRCA1/2 (ref wt)				
BRCA1				0.53 (0.25-1.12)*
BRCA2				0.39 (0.16-0.96)

Table SD17. Multivariable survival analysis of overall survival, complete cases with known or presumed adnexal sites

# of events / n	143 / 213	143 / 213	143 / 213	143 / 213
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Final Subtype (ref C2.IMM)				
C4.DIF	0.9 (0.55-1.48)	0.76 (0.46-1.27)	0.82 (0.49-1.38)	0.88 (0.52-1.48)
C5.PRO	1.03 (0.59-1.8)	0.76 (0.41-1.43)	0.95 (0.5-1.82)	0.88 (0.46-1.69)
C1.MES	1.16 (0.73-1.85)	0.96 (0.59-1.58)	0.86 (0.52-1.42)	0.99 (0.6-1.64)
Age	0.99 (0.93-1.05)	0.98 (0.92-1.05)	0.97 (0.91-1.04)	0.95 (0.89-1.01)
Stage (ref low)				
high	2.89 (1.45-5.75)*	2.65 (1.32-5.31)*	1.99 (0.97-4.07)*	2.03 (0.98-4.2)*
CD8 (ref none)				
low		1.08 (0.59-1.98)	1.18 (0.64-2.17)	1.03 (0.56-1.91)*
med		0.77 (0.46-1.3)	0.8 (0.47-1.34)	0.65 (0.38-1.11)
high		0.5 (0.25-0.99)	0.49 (0.24-0.98)	0.37 (0.18-0.75)
Residual Disease (ref none)				
any			1.93 (1.29-2.88)*	2.07 (1.37-3.13)*
BRCA1/2 (ref wt)				
BRCA1				0.95 (0.56-1.62)*
BRCA2				0.22 (0.09-0.52)

Table SD18. Multivariable survival analysis of progression-free survival, complete cases with known or presumed adnexal sites

# of events / n	152 / 184	152 / 184	152 / 184	152 / 184
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Final Subtype (ref C2.IMM)				
C4.DIF	0.95 (0.57-1.59)	0.9 (0.52-1.55)	0.99 (0.57-1.72)	1.04 (0.6-1.82)
C5.PRO	0.7 (0.39-1.24)	0.63 (0.33-1.2)	0.79 (0.41-1.54)	0.77 (0.4-1.5)
C1.MES	0.82 (0.51-1.33)	0.79 (0.47-1.32)	0.71 (0.42-1.21)	0.74 (0.44-1.25)
Age	1.1 (1.01-1.2)*	1.1 (1-1.19)*	1.07 (0.98-1.17)	1.06 (0.97-1.16)
Stage (ref low)				
high	4.87 (2.42-9.8)*	4.61 (2.26-9.43)*	3.64 (1.75-7.6)*	3.86 (1.85-8.06)*
CD8 (ref none)				
low		1.17 (0.62-2.22)	1.21 (0.64-2.28)	1.24 (0.65-2.34)
med		0.87 (0.5-1.52)	0.87 (0.5-1.51)	0.86 (0.49-1.51)
high		0.84 (0.42-1.68)	0.85 (0.42-1.7)	0.84 (0.42-1.69)
Residual Disease (ref none)				
any			1.86 (1.26-2.74)*	1.83 (1.24-2.7)*
BRCA1/2 (ref wt)				
BRCA1				0.71 (0.41-1.21)
BRCA2				0.66 (0.34-1.26)

Table SD19. Multivariable survival analysis of Overall Survival, Known Adnexal Sites, Complete Cases

# of events / n	79 / 107	79 / 107	79 / 107	79 / 107
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Final Subtype (ref C2.IMM)				
C4.DIF	0.82 (0.43-1.58)	0.65 (0.32-1.33)	0.65 (0.31-1.35)	0.67 (0.32-1.4)
C5.PRO	0.79 (0.36-1.71)	0.59 (0.24-1.43)	0.75 (0.3-1.85)	0.57 (0.22-1.47)
C1.MES	0.94 (0.48-1.83)	0.74 (0.35-1.57)	0.51 (0.23-1.1)	0.69 (0.33-1.48)
Age	1.01 (0.92-1.1)	1 (0.91-1.09)	0.96 (0.88-1.06)	0.9 (0.81-1)
Stage (ref low)				
high	3.08 (1.3-7.31)*	2.67 (1.09-6.55)*	1.61 (0.64-4.06)	2.18 (0.84-5.65)
CD8 (ref none)				
low		1.13 (0.43-2.96)	1.17 (0.45-3.05)	0.88 (0.32-2.46)
med		0.88 (0.42-1.85)	1.01 (0.48-2.14)	0.71 (0.32-1.6)
high		0.49 (0.16-1.48)	0.43 (0.14-1.33)	0.26 (0.08-0.8)
Residual Disease (ref none)				
any			3.21 (1.78-5.81)*	3.84 (2.06-7.15)*
BRCA1/2 (ref wt)				
BRCA1				0.81 (0.39-1.68)*
BRCA2				0.13 (0.04-0.43)

Table SD20. Multivariable survival analysis of Progression-Free Survival, Known Adnexal Sites, Complete Cases

# of events / n	84 / 99	84 / 99	84 / 99	84 / 99
	HR (95% CI)	HR (95% CI)	HR (95% CI)	HR (95% CI)
Final Subtype (ref C2.IMM)				
C4.DIF	0.86 (0.42-1.75)	0.64 (0.29-1.4)	0.69 (0.31-1.53)	0.76 (0.34-1.71)
C5.PRO	0.54 (0.23-1.27)	0.34 (0.12-0.9)	0.4 (0.14-1.1)	0.35 (0.13-1)
C1.MES	0.95 (0.46-1.96)	0.71 (0.32-1.58)	0.54 (0.24-1.23)	0.68 (0.3-1.55)
Age	1.17 (1.04-1.32)*	1.16 (1.03-1.31)*	1.11 (0.98-1.26)	1.07 (0.94-1.22)
Stage (ref low)				
high	4.26 (1.75-10.38)*	3.61 (1.43-9.11)*	2.44 (0.94-6.34)	2.85 (1.1-7.4)*
CD8 (ref none)				
low		1.33 (0.48-3.68)	1.42 (0.52-3.92)	1.47 (0.52-4.13)
med		0.7 (0.32-1.53)	0.76 (0.35-1.65)	0.66 (0.3-1.47)
high		0.4 (0.13-1.25)	0.36 (0.11-1.15)	0.31 (0.1-0.99)
Residual Disease (ref none)				
any			2.63 (1.46-4.75)*	2.67 (1.46-4.89)*
BRCA1/2 (ref wt)				
BRCA1				0.53 (0.25-1.12)*
BRCA2				0.39 (0.16-0.96)

E: Recommended Standard Operating Procedures for the Predictor of high-grade serous Ovarian carcinoma molecular subtype (PrOTYPE)

Starting material

is typically archival formalin-fixed (10% neutral buffered formalin) paraffin embedded (FFPE) tumour tissues 3 x 10 um sections, or approximate equivalent, that can be (a) scrolled directly into a tube, (b) scrapped from the surface of a glass slide, or (c) macrodissected from the surface of a glass slide. It is assumed the face size, or macrodissected face, for each section is ≥1cm^2 and minimal tumour content in the sampled area is > 20% (with invasive stroma but without adjacent normal organ tissue).

> *Protocol modification for FFPE tissue cores*: If FFPE tissue cores are to be used, pulverising the cores can increase deparaffinization and proteinase K digest efficiency resulting in higher nucleic acid yields. This modification should be undertaken prior to deparaffinization.
>
> Typically 1-5 cores (up to 1mm diameter) are frozen on dry ice in a microfuge tube. Cores are then crushed manually using a microfuge pestle (e.g. VWR Catalogue #47750-354 – *or similar*). Multiple rounds of re-freezing and crushing may be needed to fully pulverize the tissue depending on the size and nature of the FFPE tissue cores.

RNA Extraction:

Protocol below is modified from the *Qiagen miRNeasy FFPE kit (Catalogue # 217504).* Other extraction methods/kits may provide suitable results, however, they have not been tested.

Deparaffinization of archival specimens:

Xylene based deparaffinization is the standardized method described below, however the remainder of the extraction protocol is compatible with other methods and our studies have used either xylene or melting (see appendix from Qiagen miRNeasy FFPE kit). If large numbers of sections are used, trimming of excess paraffin and/or adding additional solvent washes during deparaffinization may be required.

Briefly:

1. Add 1 ml Xylene to each microfuge tube, containing specimen to be extracted. Vortex vigorously for 10-15s, and centrifuge at full speed (~13500 rpm) for 2 min.
2. Carefully remove the supernatant by pipetting without disturbing the pellet.
3. Add 1 ml ethanol (96–100%) to the pellet, mix by vortexing, and centrifuge at full speed for 2 min.

4. Carefully remove the supernatant by pipetting without disturbing the pellet. Carefully remove any residual ethanol using a fine pipet tip.
5. Keep the lid open & speedvac samples for 5-10 min at ambient temperature or until all residual ethanol has evaporated.
6. Continue directly to the modified Qiagen Protocol below.

Modified miRNeasy FFPE Protocol

All reagent buffer references below are to components of the Qiagen miRNeasy FFPE kit (catalogue #217504). ***Before starting*** *shelf life of reagents, storage conditions and preparation of stock reagents noted in the Qiagen miRNeasy FFPE kit must be adhered to for this modified protocol.*

A. Prepare DNase I stock solution by dissolving the lyophilized DNase I (1500 Kunitz units) in 550 µl RNase-free water (provided). To avoid loss of DNase I, do not open the vial. Inject RNase-free water into the vial using an RNase-free needle and syringe. Mix gently by inverting the vial. Do not vortex.

B. Prepare Buffer RPE by adding 4 volumes (44 ml) ethanol (96–100%) to the bottle containing 11 ml Buffer RPE concentrate. Before starting the procedure, mix reconstituted Buffer RPE by shaking.

C. Turn on and set two dry-heat blocks: one at 56˚C and one at 80˚C

1. LABEL ALL TUBES (including: 1 set of 2ml tubes, 1 set of 1.5ml tubes, all extraction columns, and final elution 1.5 ml microfuge tubes.)
2. Resuspend deparaffinized tissue pellet in 150ul of buffer PKD, close the tube tightly and mix by vigorously inverting the tube 3-5 times (or by vortexing). Briefly spin down tube to collect.
3. Add 10 µl proteinase K to the sample and mix gently by pipetting up and down.
4. Incubate at 56°C for 45 min, then at 80°C for 15 min.
5. Incubate on ice for 3 min. Then centrifuge for 15 min at 20,000 x g (13,500 rpm).
6. Transfer supernatant to a new 2ml microcentrifuge tube taking care not to disturb the pellet.
7. To the supernatant that has been transferred to the new 2ml microcentrifuge tube, add 16ul DNase Booster Buffer and 10 µl DNase I stock solution. Mix by inverting the tube. Centrifuge briefly to collect residual liquid from the sides of the tube.
8. Incubate at room temperature for 10 min.
9. Add 320 µl Buffer RBC and mix the lysate thoroughly by pipetting.
10. Add 1120 µl ethanol (100%) to the sample, and mix well by pipetting. Do not centrifuge. Proceed immediately to the next step. *Precipitates may be visible, this does not affect the procedure.*
11. Transfer 700 µl of the sample to an RNeasy MinElute spin column placed in a 2 ml collection tube (supplied w/ Qiagen Kit). Close the lid gently, and centrifuge for 15 seconds at ≥8000 x *g* (≥10,000 rpm). Discard the flow-through and use a new collection tube in the next step.

12. Repeat step 11 until the entire lysate/ethanol sample mix has passed through the RNeasy MinElute spin column. Use a new collection tube in the next step.
13. Add 500 µl Buffer RPE to the RNeasy MinElute spin column. Close the lid gently, and centrifuge for 15 s at ≥8000 x g (≥10,000 rpm). Discard the flow-through and use a new collection tube in the next step.
14. Add 500 µl Buffer RPE to the RNeasy MinElute spin column (in a new collection tube). Close the lid gently, and centrifuge for 2 min at ≥8000 x g (≥10,000 rpm) to wash the spin column membrane. Discard the collection tube with the flow-through.
15. Place the RNeasy MinElute spin column in a new 2 ml collection tube. Open the lid of the spin column, and centrifuge at full speed for 5 min (*note: when centrifuging with open collection tube it is recommended to leave an empty space between tubes and orient the lid to trail behind the column during the spin*). Discard the collection tube with the flow-through.
16. Place the RNeasy MinElute spin column in a new 1.5 ml collection tube. Add 30 µl RNase-free water directly to the spin column membrane. Close the lid gently, allow incubation at room temperature ~1-2 min then centrifuge for 1 min at full speed to elute the RNA.
17. Evaluate the concentration of the specimen by NanoDrop spectrophotometer. Specimens with absorbance A260:A280 ratio < 1.5 are not recommended for downstream application.
18. Store RNA at -80°C until ready for use.

NanoString Hybridization, Processing and Scanning

Protocol below is modified from the NanoString gene expression "XT" chemistry protocol. Details are provided for Gen1 and Gen2/Flex NanoString prep stations and digital analysers. ***Refer to Supplemental Table SC7*** for CodeSet targets and design details, any changes in the CodeSet designs of individual probes, used for housekeeping or prediction, is not supported and may yield non-interpretable results.

Sample Input – 500 ng total RNA as measured by NanoDrop following the recommended archival RNA extraction procedure outlined above.

Store reagent components according to manufacturer recommendations. Note that reagents will come pre-aliquoted for preparation of experiments in multiples of 12, repeated freeze-thawing of reagents for processing samples outside of multiples of 12 is not recommended.

Use only CodeSet reagents and consumables provided with the NanoString gene expression XT or Legacy CodeSet and Master Kits in prep station and digital analyser instruments.

1. Prepare a sample sheet with all specimen and order they will be placed in 12-strip tubes (multiple sets of 12 may be processed at the same time). Keep this sheet as a reference when setting up your hybridization and sample information (CDF) files.
2. Label NanoString Master Kit 12-well strip tubes and caps (0.2 ml thin-walled PCR tubes) with permanent marker. *Optional: cut each 12-well strip tube and strip-caps in two equal*

halves of 6-tubes/caps (this will permit the use of 8-tube benchtop micro-centrifuges for spinning reagents in your strip-tube).

3. Thaw RNA on ice and dilute samples (if necessary) with nuclease-free sterile water to a concentration of 100 ng/ul (500 ng total RNA in 5 ul volume). Use of DEPC treated water is not recommended.
4. Following exact manufacturer instruction for NanoString XT gene expression assays and using plasticware/disposables provided in NanoString custom CodeSet and master-kit sample processing kits:
   4.1. Mix NanoString custom CodeSet Reporter reagent with Master Kit hybridization buffer sufficient for the number of samples being processed at a ratio of 5ul of hybridization buffer per 3ul of Reporter CodeSet.

   If mixing directly in the provided custom CodeSet aliquots from NanoString then add 70ul of hybridization buffer to 42ul of reporter provided in a 12-reaction reporter CodeSet tube (excess reagent is provided). Mix by flicking and spin down briefly.

   4.2. In each tube of the NanoString provided 12-well strip tube add 8ul of reporter/hybridization buffer from step 4.1. The pipette tip may be re-used for aliquoting this master mix in clean tubes.
   4.3. Next, add 5ul of total RNA (500 ng at 100ng/ul prepared in step 3) to each strip tube following the order from your sample setup sheet. A new pipette tip must be used for each specimen.
   4.4. To each tube then add 2 ul of NanoString custom CodeSet capture reagent mix. A new pipette tip must be used for each specimen.
   4.5. Seal all tubes, verify labels, and mix reagents by flicking tubes. Spin down contents briefly on a low-speed benchtop micro-centrifuge with a 0.2ml PCR strip-tube adaptor.
   4.6. Program and pre-heat thermal cycler for 65C with a heated lid at 70C (non-programmable heated lids up to 110C are acceptable).
   4.7. Place strip tubes in a thermal cycler and incubate at 65C for 16 hours or 20 hours. Program the thermal cycler to ramp-down to 4C after the incubation period.
   4.8. Specimen may hold at 4C for no longer than 24hrs. It is recommended that samples are processed on the NanoString Prep Station immediately after hybridization.
5. Automated Sample Prep on the nCounter Prep Station
   5.1. Remove wash plates and cartridge from freezers and allow to equilibrate to room temperature 20-30 minutes <u>before opening packaging</u> and setting up the prep station.
   5.2. Spin down wash plates per instructions on the NanoString prep station to ensure capture beads and reagents are not stuck to foil covers.
   5.3. Enter detail on number of samples to be processed on the NanoString nCounter prep station. If high-sensitivity mode is available on your system this should be selected.
   5.4. Follow the instructions on the prep station and place specimens in 12-strip tube, wash plate reagents, cartridge and plasticware in the indicated deck positions.
   5.5. Close the prep station cover and start the instrument.
   5.6. When sample prep program is complete, remove and seal the cartridge immediately using the Master Kit adhesive covers. Label cartridges on the top surface with caution

not to block the light path for cartridge lanes. Scanning on the digital analyser immediately after processing is recommended.

6. Scanning on the nCounter Digital Analyzer
   6.1. It is recommended that sample information be entered using pre-filled CDF files (see notes below on **CDF File Setup and Naming Conventions**). This information should be transferred to the NanoString Digital Analyzer in advance of sample processing.
   6.2. Place the sealed cartridge into 1 of the digital analyser positions, ensure it is aligned per manufacturers specification and close the lid.
      6.2.1. If you are using a Gen1 digital analyser cartridges must have oil applied before placing in the digital analyser. Follow NanoString guidelines for application of oil droplets to the under-surface of the cartridge.
   6.3. Follow the onscreen menus to select you pre-filled CDF sample processing file.
      6.3.1. Ensure the correct CodeSet file (RLF) and MAX FOV has been selected in your sample file (CDF) or manual setup: Gen1 this is 1155 FOV, Gen2/Flex is 555 FOV.
   6.4. Start the NanoString Digital Analyzer.
   6.5. Cartridges may be added mid-run following manufacturers guidelines.
   6.6. Remove and store scanned cartridges at 4C following scanning and until you have run quality control checks to verify scanning. Sealed cartridges may be stored up to 5 days at 4C for batch scanning or re-scanning in case of scan failures.
7. Specimen Analysis can be performed using the PrOTYPE online tool or code provided in this manuscript (see RESOURCES – below). You will need:
   7.1. Your HGSOC sample data (RCC file or files) on which you wish to predict molecular subtypes.
   7.2. Reference data from POOL1, POOL2, and POOL3 (RCC files), multiple runs of references are acceptable. References MUST have been run on the exact same CodeSet (synthesis lot) as the HGSOC data file(s).

To obtain aliquots of POOL1, POOL2, and POOL3 contact the corresponding authors. Use of other reference samples and/or re-use of reference sample data from non-identical CodeSet/synthesis lots may render prediction data non-interpretable.

CDF File Setup and Naming Conventions:

The following sample naming conventions have been adopted in our code and web tools. Deviation from these recommendations may cause our code to mis-interpret sample or reference input information and yield non-interpretable data.

Samples and cartridge names can contain letters, numbers and the following symbols only: @ ^ + - # % (~ &)

The following fields of the standard NanoString CDF sample information file must be filled out by the user and transferred to the Digital Analyser. This information will be embedded in the output datafile (RCC) and used for sample identification and quality assurance in processing tools.

Users must fill out values for:

1. **CartridgeID** We recommend following a nomenclature that identifies the project, site, and sequential run number. However, the cartridge ID can be any alphanumeric string do not add spaces or underscore characters. A cartridge ID structure identifying the project, sequential run number is recommended.
2. **SampleID** This should be specimen identifier and may be alphanumeric, with NO spaces, or underscores (dashes are ok).
    a. **Naming and running of Reference control pool RNA samples:** There are three control pool RNA samples that are required for normalization. These references should be named exactly as provided: POOL1, POOL2, and POOL3 (not case sensitive). Any experimental CodeSet wanting to run the PrOTYPE tool will require at least one of each reference pool to have been run in the SAME CodeSet (same probes and same synthesis lot) as the sample(s) being analyzed.
3. **Owner** This field was used to identify the operator for a given run.
4. **Comment** (optional) can contain any string. In our studies, this field was used to identify the hyb conditions for the given sample (e.g. “short-hyb” or “long-hyb”).
5. **Date** use ONLY format YYYYMMDD
6. **FovCount** use maximum fields 1155 for Gen1 scanner, 555 for Gen2/flex scanner
7. **GeneRLF** identifies the CodeSet and synthesis, please input the RLF name exactly as provided by NanoString. (*typically follows a format of “NAME_IDCODE”*)

Data File Name: The CDF file will define how the scanner names the output data file (.RCC) for each sample. This will follow the format:

ScanDate_CartridgeID_SampleID_Lane.RCC

Resources**:**

Code: https://github.com/AlineTalhouk/PrOType

PrOType Web tool: https://ovcare.shinyapps.io/PrOType/

REFERENCES FOR SUPPLEMENTAL APPENDIX

1. Kommoss S, Gilks CB, du Bois A, Kommoss F. Ovarian carcinoma diagnosis: the clinical impact of 15 years of change. British journal of cancer **2016**;115(8):993-9 doi 10.1038/bjc.2016.273.
2. Peres LC, Cushing-Haugen KL, Anglesio M, Wicklund K, Bentley R, Berchuck A*, et al.* Histotype classification of ovarian carcinoma: A comparison of approaches. Gynecologic oncology **2018** doi 10.1016/j.ygyno.2018.08.016.
3. Kurman RJ, Carcangiu ML, Herrington CS, Young RH, editors. WHO Classification of Tumours of Female Reproductive Organs. 4 ed. Volume 6. Lyon, France: IARC-WHO Press; 2014. 316 p.
4. Heintz AP, Odicino F, Maisonneuve P, Quinn MA, Benedet JL, Creasman WT*, et al.* Carcinoma of the ovary. FIGO 26th Annual Report on the Results of Treatment in Gynecological Cancer. International journal of gynaecology and obstetrics: the official organ of the International Federation of Gynaecology and Obstetrics **2006**;95 Suppl 1:S161-92 doi 10.1016/S0020-7292(06)60033-7.
5. Prat J, Oncology FCoG. FIGO's staging classification for cancer of the ovary, fallopian tube, and peritoneum: abridged republication. J Gynecol Oncol **2015**;26(2):87-9 doi 10.3802/jgo.2015.26.2.87.
6. Sieh W, Kobel M, Longacre TA, Bowtell DD, deFazio A, Goodman MT*, et al.* Hormone-receptor expression and ovarian cancer survival: an Ovarian Tumor Tissue Analysis consortium study. Lancet Oncol **2013**;14(9):853-62 doi 10.1016/S1470-2045(13)70253-5.
7. Song H, Cicek MS, Dicks E, Harrington P, Ramus SJ, Cunningham JM*, et al.* The contribution of deleterious germline mutations in BRCA1, BRCA2 and the mismatch repair genes to ovarian cancer in the population. Human molecular genetics **2014**;23(17):4703-9 doi 10.1093/hmg/ddu172.
8. Soegaard M, Kjaer SK, Cox M, Wozniak E, Hogdall E, Hogdall C*, et al.* BRCA1 and BRCA2 mutation prevalence and clinical characteristics of a population-based series of ovarian cancer cases from Denmark. Clin Cancer Res **2008**;14(12):3761-7 doi 10.1158/1078-0432.CCR-07-4806.
9. Alsop K, Fereday S, Meldrum C, deFazio A, Emmanuel C, George J*, et al.* BRCA mutation frequency and patterns of treatment response in BRCA mutation-positive women with ovarian cancer: a report from the Australian Ovarian Cancer Study Group. J Clin Oncol **2012**;30(21):2654-63 doi 10.1200/JCO.2011.39.8545.
10. Rambau PF, Vierkant RA, Intermaggio MP, Kelemen LE, Goodman MT, Herpel E*, et al.* Association of p16 expression with prognosis varies across ovarian carcinoma histotypes: an Ovarian Tumor Tissue Analysis consortium study. J Pathol Clin Res **2018** doi 10.1002/cjp2.109.
11. Ovarian Tumor Tissue Analysis C, Goode EL, Block MS, Kalli KR, Vierkant RA, Chen W*, et al.* Dose-Response Association of CD8+ Tumor-Infiltrating Lymphocytes and Survival

Time in High-Grade Serous Ovarian Cancer. JAMA Oncol **2017**;3(12):e173290 doi 10.1001/jamaoncol.2017.3290.
12. Cancer Genome Atlas Research N. Integrated genomic analyses of ovarian carcinoma. Nature **2011**;474(7353):609-15 doi 10.1038/nature10166.
13. Tusher VG, Tibshirani R, Chu G. Significance analysis of microarrays applied to the ionizing radiation response. Proceedings of the National Academy of Sciences of the United States of America **2001**;98(9):5116-21.
14. Tan TZ, Miow QH, Huang RY, Wong MK, Ye J, Lau JA*, et al.* Functional genomics identifies five distinct molecular subtypes with clinical relevance and pathways for growth control in epithelial ovarian cancer. EMBO molecular medicine **2013**;5(7):983-98 doi 10.1002/emmm.201201823.
15. Tothill RW, Tinker AV, George J, Brown R, Fox SB, Lade S*, et al.* Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clin Cancer Res **2008**;14(16):5198-208 doi 10.1158/1078-0432.CCR-08-0196.
16. Helland A, Anglesio MS, George J, Cowin PA, Johnstone CN, House CM*, et al.* Deregulation of MYCN, LIN28B and LET7 in a molecular subtype of aggressive high-grade serous ovarian cancers. PLoS One **2011**;6(4):e18064 doi 10.1371/journal.pone.0018064.
17. Verhaak RG, Tamayo P, Yang JY, Hubbard D, Zhang H, Creighton CJ*, et al.* Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. J Clin Invest **2013**;123(1):517-25 doi 10.1172/JCI65833.
18. Leong HS, Galletta L, Etemadmoghadam D, George J, Australian Ovarian Cancer S, Kobel M*, et al.* Efficient molecular subtype classification of high-grade serous ovarian cancer. The Journal of pathology **2015**;236(3):272-7 doi 10.1002/path.4536.
19. Rudd J, Zelaya RA, Demidenko E, Goode EL, Greene CS, Doherty JA. Leveraging global gene expression patterns to predict expression of unmeasured genes. BMC Genomics **2015**;16:1065 doi 10.1186/s12864-015-2250-5.
20. Yoshihara K, Tsunoda T, Shigemizu D, Fujiwara H, Hatae M, Fujiwara H*, et al.* High-risk ovarian cancer based on 126-gene expression signature is uniquely characterized by downregulation of antigen presentation pathway. Clin Cancer Res **2012**;18(5):1374-85 doi 10.1158/1078-0432.CCR-11-2725.
21. Talhouk A, Kommoss S, Mackenzie R, Cheung M, Leung S, Chiu DS*, et al.* Single-Patient Molecular Testing with NanoString nCounter Data Using a Reference-Based Strategy for Batch Effect Correction. PLoS One **2016**;11(4):e0153844 doi 10.1371/journal.pone.0153844.
22. Nielsen T, Wallden B, Schaper C, Ferree S, Liu S, Gao D*, et al.* Analytical validation of the PAM50-based Prosigna Breast Cancer Prognostic Gene Signature Assay and nCounter Analysis System using formalin-fixed paraffin-embedded breast tumor specimens. BMC Cancer **2014**;14:177 doi 10.1186/1471-2407-14-177.
23. Wallden B, Storhoff J, Nielsen T, Dowidar N, Schaper C, Ferree S*, et al.* Development and verification of the PAM50-based Prosigna breast cancer gene signature assay. BMC Med Genomics **2015**;8:54 doi 10.1186/s12920-015-0129-6.
24. Scott DW, Abrisqueta P, Wright GW, Slack GW, Mottok A, Villa D*, et al.* New Molecular Assay for the Proliferation Signature in Mantle Cell Lymphoma Applicable to Formalin-

Fixed Paraffin-Embedded Biopsies. J Clin Oncol **2017**;35(15):1668-77 doi 10.1200/JCO.2016.70.7901.
25. Wong KK, Izaguirre DI, Kwan SY, King ER, Deavers MT, Sood AK*, et al.* Poor survival with wild-type TP53 ovarian cancer? Gynecologic oncology **2013**;130(3):565-9 doi 10.1016/j.ygyno.2013.06.016.
26. Tan TZ, Yang H, Ye J, Low J, Choolani M, Tan DS*, et al.* CSIOVDB: a microarray gene expression database of epithelial ovarian cancer subtype. Oncotarget **2015**;6(41):43843-52 doi 10.18632/oncotarget.5983.
27. Taminau J, Meganck S, Lazar C, Steenhoff D, Coletta A, Molter C*, et al.* Unlocking the potential of publicly available microarray data using inSilicoDb and inSilicoMerging R/Bioconductor packages. BMC bioinformatics **2012**;13:335 doi 10.1186/1471-2105-13-335.
28. Chiu DS, Talhouk A. diceR: an R package for class discovery using an ensemble driven approach. BMC bioinformatics **2018**;19(11):1-4 doi 10.1186/s12859-017-1996-y.
29. Lee DD, Seung HS. Algorithms for non-negative matrix factorization. 2001. p 556-62.
30. Brunet JP, Tamayo P, Golub TR, Mesirov JP. Metagenes and molecular pattern discovery using matrix factorization. Proceedings of the National Academy of Sciences of the United States of America **2004**;101(12):4164-9 doi 10.1073/pnas.0308531101.
31. Kaufman L, Rousseeuw P. Statistical Data Analysis Based on the L1 Norm and Related Methods. In: Dodge Y, editor. Clustering by means of medoids. Amsterdam: North-Holland; 1987. p 405-16.
32. Govaert G, Nadif M. Block clustering with Bernoulli mixture models: Comparison of different approaches. Computational Statistics & Data Analysis **2008**;52(6):3233-45.
33. Steinbach M, Ertöz L, Kumar V. The challenges of clustering high dimensional data. New directions in statistical physics: Springer; 2004. p 273-309.
34. Huang Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. Research Issues on Data Mining and Knowledge Discovery1997. p 1-8.
35. Li J, Bushel PR, Chu TM, Wolfinger RD. Principal variance components analysis: Estimating batch effects in microarray gene expression data. Batch Effects and Noise in Microarray Experiments: Sources and Solutions **2009**:141-54.
36. Liberzon A, Birger C, Thorvaldsdottir H, Ghandi M, Mesirov JP, Tamayo P. The Molecular Signatures Database (MSigDB) hallmark gene set collection. Cell Syst **2015**;1(6):417-25 doi 10.1016/j.cels.2015.12.004.
37. Liberzon A, Subramanian A, Pinchback R, Thorvaldsdottir H, Tamayo P, Mesirov JP. Molecular signatures database (MSigDB) 3.0. Bioinformatics (Oxford, England) **2011**;27(12):1739-40 doi 10.1093/bioinformatics/btr260.
38. Subramanian A, Tamayo P, Mootha VK, Mukherjee S, Ebert BL, Gillette MA*, et al.* Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences of the United States of America **2005**;102(43):15545-50.
39. Zhang S, Jing Y, Zhang M, Zhang Z, Ma P, Peng H*, et al.* Stroma-associated master regulators of molecular subtypes predict patient prognosis in ovarian cancer. Sci Rep **2015**;5:16066 doi 10.1038/srep16066.

40. Wang C, Armasu SM, Kalli KR, Maurer MJ, Heinzen EP, Keeney GL*, et al.* Pooled Clustering of High-Grade Serous Ovarian Cancer Gene Expression Leads to Novel Consensus Subtypes Associated with Survival and Surgical Outcomes. Clin Cancer Res **2017**;23(15):4077-85 doi 10.1158/1078-0432.ccr-17-0246.
41. Integrated genomic analyses of ovarian carcinoma. Nature **2011**;474(7353):609-15.
42. Konecny GE, Wang C, Hamidi H, Winterhoff B, Kalli KR, Dering J*, et al.* Prognostic and therapeutic relevance of molecular subtypes in high-grade serous ovarian cancer. J Natl Cancer Inst **2014**;106(10) doi 10.1093/jnci/dju249.
43. Naba A, Clauser KR, Ding H, Whittaker CA, Carr SA, Hynes RO. The extracellular matrix: Tools and insights for the "omics" era. Matrix Biol **2016**;49:10-24 doi 10.1016/j.matbio.2015.06.003.
44. Liu J, Lichtenberg T, Hoadley KA, Poisson LM, Lazar AJ, Cherniack AD*, et al.* An Integrated TCGA Pan-Cancer Clinical Data Resource to Drive High-Quality Survival Outcome Analytics. Cell **2018**;173(2):400-16 e11 doi 10.1016/j.cell.2018.02.052.
45. Breiman L. Out-of-bag estimation. Berkeley, CA: University of California Berkeley; 1996. 1-13 p.
46. Rojas-Domínguez A, Padierna LC, Valadez JMC, Puga-Soberanes HJ, Fraire HJ. Optimal Hyper-Parameter Tuning of SVM Classifiers With Application to Medical Diagnosis. IEEE Access **2018**;6:7164-76.
47. Afolabi LT, Saeed F, Hashim H, Petinrin OO. Ensemble learning method for the prediction of new bioactive molecules. PloS one **2018**;13(1):e0189538.
48. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection.
49. MacKay DJC. Information Theory, Inference, and Learning Algorithms. Cambridge, UK: Cambridge University Press; 2003.
50. Institute of Medicine. EVOLUTION OF TRANSLATIONAL OMICS Lessons Learned and the Path Forward. Micheel CM, Nass SJ, Omenn GS, editors. Washington, DC: The National Academies Press; 2012.
51. Chen GM, Kannan L, Geistlinger L, Kofia V, Safikhani Z, Gendoo DMA*, et al.* Consensus on Molecular Subtypes of High-Grade Serous Ovarian Carcinoma. Clin Cancer Res **2018**;24(20):5037-47 doi 10.1158/1078-0432.CCR-18-0784.
52. Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc **2009**;4(1):44-57 doi nprot.2008.211 [pii]

10.1038/nprot.2008.211.
53. Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC*, et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome biology **2003**;4(5):P3.
54. Xia J, Benner MJ, Hancock RE. NetworkAnalyst--integrative approaches for protein-protein interaction network analysis and visual exploration. Nucleic acids research **2014**;42(Web Server issue):W167-74 doi 10.1093/nar/gku443.

55. Xia J, Gill EE, Hancock RE. NetworkAnalyst for statistical, visual and network-based meta-analysis of gene expression data. Nat Protoc **2015**;10(6):823-44 doi 10.1038/nprot.2015.052.
56. Fabregat A, Jupe S, Matthews L, Sidiropoulos K, Gillespie M, Garapati P*, et al.* The Reactome Pathway Knowledgebase. Nucleic acids research **2018**;46(D1):D649-D55 doi 10.1093/nar/gkx1132.
57. Croft D, O'Kelly G, Wu G, Haw R, Gillespie M, Matthews L*, et al.* Reactome: a database of reactions, pathways and biological processes. Nucleic acids research **2011**;39(Database issue):D691-7 doi 10.1093/nar/gkq1018.
58. Germain PL, Vitriolo A, Adamo A, Laise P, Das V, Testa G. RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. Nucleic acids research **2016**;44(11):5054-67 doi 10.1093/nar/gkw448.
59. Veldman-Jones MH, Brant R, Rooney C, Geh C, Emery H, Harbron CG*, et al.* Evaluating Robustness and Sensitivity of the NanoString Technologies nCounter Platform to Enable Multiplexed Gene Expression Analysis of Clinical Samples. Cancer research **2015**;75(13):2587-93 doi 10.1158/0008-5472.CAN-15-0262.
60. Malkov VA, Serikawa KA, Balantac N, Watters J, Geiss G, Mashadi-Hossein A*, et al.* Multiplexed measurements of gene signatures in different analytes using the Nanostring nCounter Assay System. BMC Res Notes **2009**;2:80 doi 1756-0500-2-80 [pii]

10.1186/1756-0500-2-80.