Identification of candidate susceptibility genes in the rarer histotypes of epithelial ovarian cancer

Marina Pavanello

A thesis in fulfilment of the requirements for the degree of Doctor of Philosophy

School of Women's and Children's Health

Faculty of Medicine

May 2022

Abstract

Identification of ovarian cancer susceptibility genes has focussed on the most common histotype, the high-grade serous cases (HGSOC). Determining risk factors for the rarer, non-HGSOC histotypes (low-grade serous (LGSOC), endometrioid (ENOC), clear cell (CCOC), and mucinous (MOC)) has been challenging due to limited sample sizes. This thesis used resources from a large international collaboration to investigate these rare cases. Firstly, RNA expression data was used to identify grade 2 serous cases more likely to be LGSOC than HGSOC. Up to 14% of these cases could be re-classed as LGSOC (80%–95% accuracy). For the discovery of new candidate susceptibility genes, a pilot-study and a follow-up study with a larger number of samples were performed. For the pilot-study, germline whole-exome sequencing (WES) of 251 non-HGSOC was compared to public controls. Twenty-five new genes and seven HGSOC susceptibility genes were selected for targeted sequencing. Analysis was performed in 1,673 non-HGSC and 1,790 controls that passed QC. A significantly higher frequency of predicted protein truncating variants (PTV) in cases than controls was identified for *ERCC6* and *IL31RA*, although this was no longer significant after validation of variants. For the follow-up study, WES of 775 non-HGSOC were compared to public controls. Twenty-three new genes, including two identified in the pilot-study, and 11 known ovarian cancer susceptibility genes were selected for targeted sequencing. Histotype-specific analysis used 2,102 non-HGSOC cases, 1,456 HGSOC cases and 2,306 controls that passed QC. *PRKAA1* had a significantly higher frequency of PTV in all non-HGSOC cases than controls, as well as in ENOC cases alone. *IL31RA* and *ERCC6* were also significant when comparing ENOC and LGSOC to controls. However, these were not significant after validation of variants. Overall, a total of 2,405 non-HGSOC cases were compared to 3,121 controls. *IL31RA* had a borderline significantly higher frequency of validated or assumed validated PTV in non-HGSOC compared to controls (six in all non-HGSOC *vs* one in controls, p 0.048 and three in ENOC *vs* one in controls, p 0.043). Although not significant when adjusted for multiple testing, *IL31RA* could potentially be a new candidate gene for non-HGSOC, but sequencing in much larger numbers is needed. In addition, the frequency of PTV in the known susceptibility genes was determined in a large number of these rare cases, and this could improve advice on genetic risk for these patients.

Use of RNA expression data to classify serous ovarian tumours

1.1 Introduction

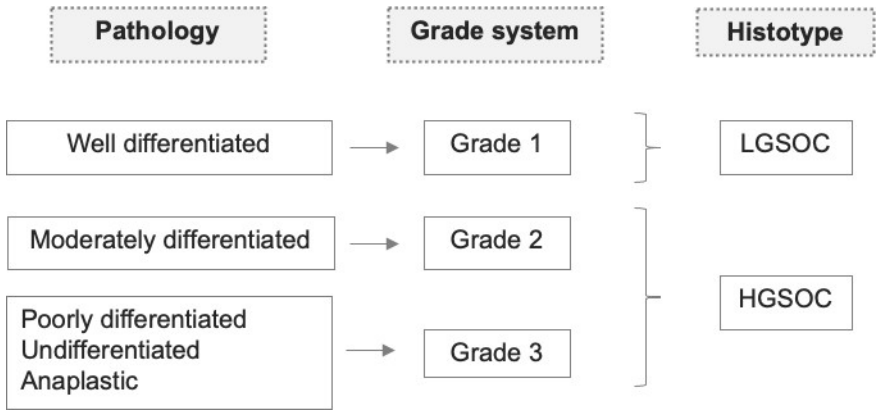
High- and low-grade serous carcinomas (HGSOC and LGSOC, respectively) are two very distinct ovarian cancer histotypes as previously described in the introduction of this thesis (Table 2.1). LGSOC is an extremely rare histotype and accounts for about 5% of all EOC while HGSOC accounts for approximately 68% of all the cases. Clinically, LGSOC cases are associated with a younger age of onset when compared to HGSOC cases; mean age at diagnosis for LGSOC cases is 55.5 years (about 7–10 years younger than patients with HGSOC) (152). They are often resistant to primary platinum-based chemotherapy unlike HGSOC that are typically sensitive (153). Despite chemoresistance in these patients, women diagnosed with LGSOC are still treated the same as women diagnosed with HGSOC – surgical debulking followed by a combination of platinum/taxane-based chemotherapy for six to eight cycles. Although being relatively resistant to treatment, LGSOC cases still have better survival than HGSOC cases. This might be due to relatively slow growth of the tumour and higher probability of the surgeon achieving no residual disease through primary surgery (154). However, when presented at an advanced stage after primary treatment, LGSOC patients have similarly poor outcomes as HGSOC patients (154,155).

Molecular changes also differ between these two histotypes. LGSOC typically contains RAS–MAPK pathway mutations (81), few DNA copy-number changes, lower rate of somatic mutation, and are usually wild-type *TP53* (82). Driver mutations in LGSOC are usually stable spatially and temporally (156) and, therefore, clinical trials have been assessing new treatments, such as RAS–MAPK pathway inhibitors, targeting these events. Novel therapeutic strategies using the BRAF inhibitor vemurafenib showed long-term partial response, CA125 reductions and symptom relief in $BRAF^{V600E}$-mutated LGSOC cases (88). In addition to somatic RAS–MAPK pathway mutations (such as *BRAF*, *KRAS*, and *NRAS* mutations), frameshift mutations in *NF1* might also be associated with deregulation of this pathway. Co-occurrence of *NRAS* and *EIF1AX* mutations found in other cancers, such as uveal melanoma and poorly differentiated thyroid cancer, were also seen in LGSOC (84). Additionally, almost every HGSOC case has mutant *TP53* (15) whereas LGSOC are characteristically *TP53* wild-type (157).

Table 2.1 Main differences between high- and low-grade serous ovarian carcinomas.

Feature	Low-grade serous ovarian cancer	High-grade serous ovarian cancer
Behaviour	Indolent	Highly aggressive
Pathology	Micro-papillary pattern	Papillary or solid growth pattern
Proliferative rate	Low	High
Response to treatment	Relatively chemoresistant	Initially chemosensitive following potential resistance
Prognosis	Good	Poor
Key molecular changes	*BRAF, KRAS, NRAS, PIK3CA*	*TP53, BRCA1/2*

In 2013, a formal statistical assessment was performed showing that a combination of TP53/CDKN2A protein expression can distinguish between low-grade and high-grade serous carcinomas (78). Previously, serous tumours were categorised by grade alone in the clinic, using the FIGO 3-tier grading system for serous ovarian cancer (Figure 2.1) (158). When classifying serous tumours from a 3 or 4 grade system into LGSOC or HGSOC, the grade 2 tumours are typically classed as HGSOC (82) (Figure 2.1). More recently, pathologists have changed the classification of these tumours into either high- or low-grade groups (20). However, in research data, it is still common to only have the grade data.



Note: studies using the 4 grade system also classify grade 4 tumours as HGSOC.

Figure 2.1 Schema of FIGO 3-tier grading system for serous ovarian cancer (Source: Baak JP et al. (1986)).

Robust diagnosis of different histotypes has become more important in the era of individualized therapy. It has been recently suggested that a subset of the grade 2 tumours might have been misclassified as high-grade and should be considered low-grade instead (82). These cases are being identified as *TP53* wild-type according to sequencing data and mutations in *BRAF*, *KRAS* or *NRAS* are present, which are commonly observed in the LGSOC cases (82). Sequencing performed on 11 grade 2 tumours identified that 18% of them had *NRAS* mutations and were *TP53* wild-type (82). Subsequent pathological review has found that these cases also presented low-grade features (82).

Currently, classification of HGSOC and LGSOC is usually done by assessing p53 immunohistochemistry (IHC) staining. It has been shown that p53 IHC has high sensitivity (0.96) and specificity (1.00) to discriminate between high- and low-grade serous carcinomas, and it is the most cost-effective method (159). However, it can miss some mutations. Alternatively, DNA copy number alterations could also be considered for classification as it has been shown that HGSOC tumours have a high degree of DNA copy number changes, while LGSOC normally do not have this feature (160). But DNA copy number is costly to perform on a large number of samples.

Although IHC might be the main tool to distinguish high- and low-grade serous tumours, it may not always be available in previously collected cohorts, and other assays could be considered when expert pathology review is not feasible on large numbers of cases. Gene expression profiling, for example, had been previously used to distinguish histological subtypes of ovarian cancer (161). This data is particular useful in a research setting where expression data may be already available. Gene expression has been extensively used in ovarian cancer research, leading to important findings, such as the identification of four distinct, and clinically relevant subtypes of HGSOC (C1 - high stromal response, C2 - high immune signature, C4 - low stromal response, and C5 - mesenchymal, low immune signature) (16). The identification of clinically relevant subtypes using gene expression profiling was observed in other solid tumours such as breast (162) and lung cancer (163).

It was extremely important for this thesis to correctly classify serous ovarian cancer histotypes given the rarity of LGSOC cases. Sample size is often a problem when finding rare variants for rare cancers, thus it was critical that a substantial proportion of low-grade serous was not being missed. To assess this issue, classification and prediction models were used in two large already available RNA expression datasets of serous ovarian cancer to identify which grade 2 cases could have been misclassified.

The hypothesis was that a subset of grade 2 serous tumours was LGSOC cases, and these could be identified from RNA expression data. As previously mentioned,

molecular classification has been shown to provide accurate stratification of histotypes in many cancers including ovarian cancer (16). Therefore, the aim of this expression data analysis was to classify grade 2 tumours as HGSOC or LGSOC, and subsequently to identify additional misclassified LGSOC cases that may be suitable for the following sequencing studies.

1.2 Materials and methods

1.2.1 Samples

RNA expression data was generated using nCounter technology from NanoString. Formalin-Fixed Paraffin-Embedded (FFPE) ovarian tumours from 32 studies were included. The studies were part of the Ovarian Tumour Tissue Analysis Consortium (OTTA). RNA expression data was made available for the purpose of this thesis from OTTA collaborators as two different datasets. Dataset 1 had RNA expression of 513 genes for 43 LGSOC cases and 2,936 HGSOC cases (489 grade 2 and 2,447 grade 3/4). All cases were classified by grade by a gynaecologic pathologist (Table 2.2). Dataset 2 had RNA expression of 335 genes (218 genes overlapping with dataset 1) for 1,194 HGSOC cases (166 grade 2, and 1,028 grade 3/4) (Table 2.3). In total, gene expression data was available for 655 grade 2 cases, of these 134 had suitable consent for WES (20%) and 414 for targeted sequencing (63%). Six studies, indicated in Table 2.2 and 2.3, used a 4-grade system and the grade 4 cases were combined with grade 3 as grade 3/4 throughout this chapter. Datasets 1 and 2 had not been published by the time of these analyses.

Table 2.2 Serous samples in dataset 1 categorized by study.

Study ID	Study name	Grade 1	Grade 2	Grade 3/4	Total by study
AOC^	Australia Ovarian Cancer Study	2	94	451	547
DOV^	Diseases of the Ovary and their Evaluation Study	16	62	418	496
NCO	North Carolina Ovarian Cancer Study	13	126	210	349
MAY1^	Mayo Clinic Ovarian Cancer Case Control Study	0	5	336	341
SEA	UK Studies of Epidemiology and Risk Factors in Cancer Heredity Ovarian Cancer Study	0	3	234	237
VAN	British Columbia Cancer Agency	5	18	179	202
SRF	SCOTROC 4 randomized trial study	0	30	107	137

Study ID	Study name	Grade 1	Grade 2	Grade 3/4	Total by study
WMH	Westmead Hospital	1	31	96	128
TRI	TRIO University of California, Los Angeles	4	23	93	120
UKO	UK Ovarian Cancer Population Study	1	13	97	111
TRT	TORONTO	0	21	61	82
POC	Poland Ovarian Cancer Study	0	28	34	62
POL	NCI Ovarian Case-Control Study in Poland	0	20	24	44
USC^	University of Southern California	0	7	30	37
HAW^	Hawaii Ovarian Cancer Study	0	0	35	35
LAX	Women's Cancer Research Institute (Cedars-Sinai Medical Center)	0	2	25	27
MAY2	Mayo Clinic Ovarian Cancer Study	0	0	14	14
GER	German Ovarian Cancer Study	0	5	2	7
AOV	OVarian Types study	1	1	1	3
Total by grade		**43**	**489**	**2447**	**2979**

^ Studies with grade 3 and 4 cases combined

Table 2.3 Serous samples in dataset 2 divided by study.

Study ID	Study name	Grade 2	Grade 3/4	Total by study
MAY1^	Mayo Clinic Ovarian Cancer Case Control Study	5	132	137
WAG	Western Australia Gynaecological Oncology Biobank	0	129	129
BEL	Belgian Ovarian Cancer Study (BCOS)	0	117	117
NEC	New England Case-Control Study	16	94	110
BAV	Bavarian Ovarian Cancer Cases and Controls	14	95	109
MAL	Danish Malignant Ovarian Tumor Study	31	48	79
SWE^	Sweden	19	59	78
JGO	Japanese Gynecologic Oncology Study	15	51	66
PVD	Pelvic Mass Study	31	28	59
COH	City of Hope - Omics and Ovarian Cancer	3	53	56
LAX	Women's Cancer Research Institute (Cedars-Sinai Medical Center)	2	51	53

Study ID	Study name	Grade 2	Grade 3/4	Total by study
COE	Gynecologic Cancer Center of Excellence/HJF INOVA	0	52	52
BRZ	Brazil Gynecologic Tumor Bank (BRZ) study	19	25	44
WMH	Westmead Hospital	5	38	43
HOP	Hormones and Ovarian Cancer Prediction	4	30	34
USC^	University of Southern California	2	15	17
CAL	Calgary HGSC chemoresistant	0	11	11
Total by grade		**166**	**1028**	**1194**

^ Studies with grade 3 and 4 cases combined

1.2.2 RNA expression data

The nCounter NanoString technology was used for quantifying RNA expression of samples in datasets 1 and 2. This technology is based on single-molecule imaging of color-coded barcodes bound to target-specific probes – see Figure 2.2. The target-probe complex contains the target-specific probes which carries the fluorescent barcode and hybridizes directly to the target molecule.

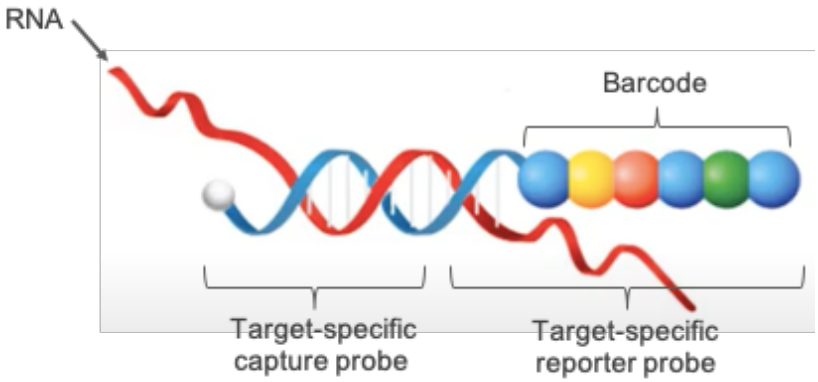


Figure 2.2 Target-probe complex illustration. Source: NanoString Technologies website: https://nanostring.com/products/ncounter-analysis-system/ncounter-systems-overview/

The data was normalised to five housekeeping genes (164). The data generated is in “relative counts” which is then interpreted as relative changes across the samples. The genes were selected for prognosis and HGSOC molecular subtype classification (165).

1.2.3 Gene expression analysis models

To identify which subset of grade 2 tumours could be potentially reclassified as LGSOC based on RNA expression data, classification and prediction modelling was performed following the pipeline below (Figure 2.3). Initially, classification models in dataset 1 were used to distinguish expression in grade 1 from grade 3/4 cases. Two sets of analysis were considered: (1) all-studies; and (2) only studies that initially had grade 1 tumours in their cohort (reduced studies). Analysis using a reduced set of studies was performed to lessen errors due to the imbalanced dataset. Gene lists generated from these models were used to predict which subset of the grade 2 cases were more likely to be LGSOC (Figure 2.3a).

Identification of LGSOC cases within the grade 2 cases subset from dataset 2 was more complex as no grade 1 cases were included. Genes that overlapped between dataset 1 and 2 were identified, then a list of genes that distinguished grade 1 from grade 3/4 in dataset 1 were found, using only this panel of overlapped genes. This trained model was then used to predict which grade 2 cases in dataset 1 were more likely to be LGSOC (Figure 2.3b). The same 2 sets of studies were analysed (all-studies and reduced studies). Finally, the best trained model overall, was selected to predict which grade 2 cases in dataset 2 were more likely to be LGSOC instead of HGSOC (Figure 2.3c).

1.2.3.1 Classification model using Random forest

Classification models were performed by a supervised learning algorithm called Random forest. Performed in R Studio (Version 1.2.5001), this algorithm combines predictions made by multiple decision trees. The best classification model was selected based on the Out-of-Bag (OOB) error rate and the class error of the minority group, thus error of grade 1 cases. These errors were calculated based on the misclassified points of the prediction trees in the training set, divided by the total number of observations.

Random forest was further used to rank the importance of gene expressions to best distinguish between grade 1 and grade 3/4 tumours. This rank used the mean decrease accuracy score as the criterion for gene expression selection. The mean decrease accuracy score estimates the loss in prediction performance when a specific gene expression is omitted from the training set. Error rates of the top 20, 40 and 60 most important genes, as well as the error using all genes, were analysed. Models that presented the smallest error rate were selected for the prediction analyses.

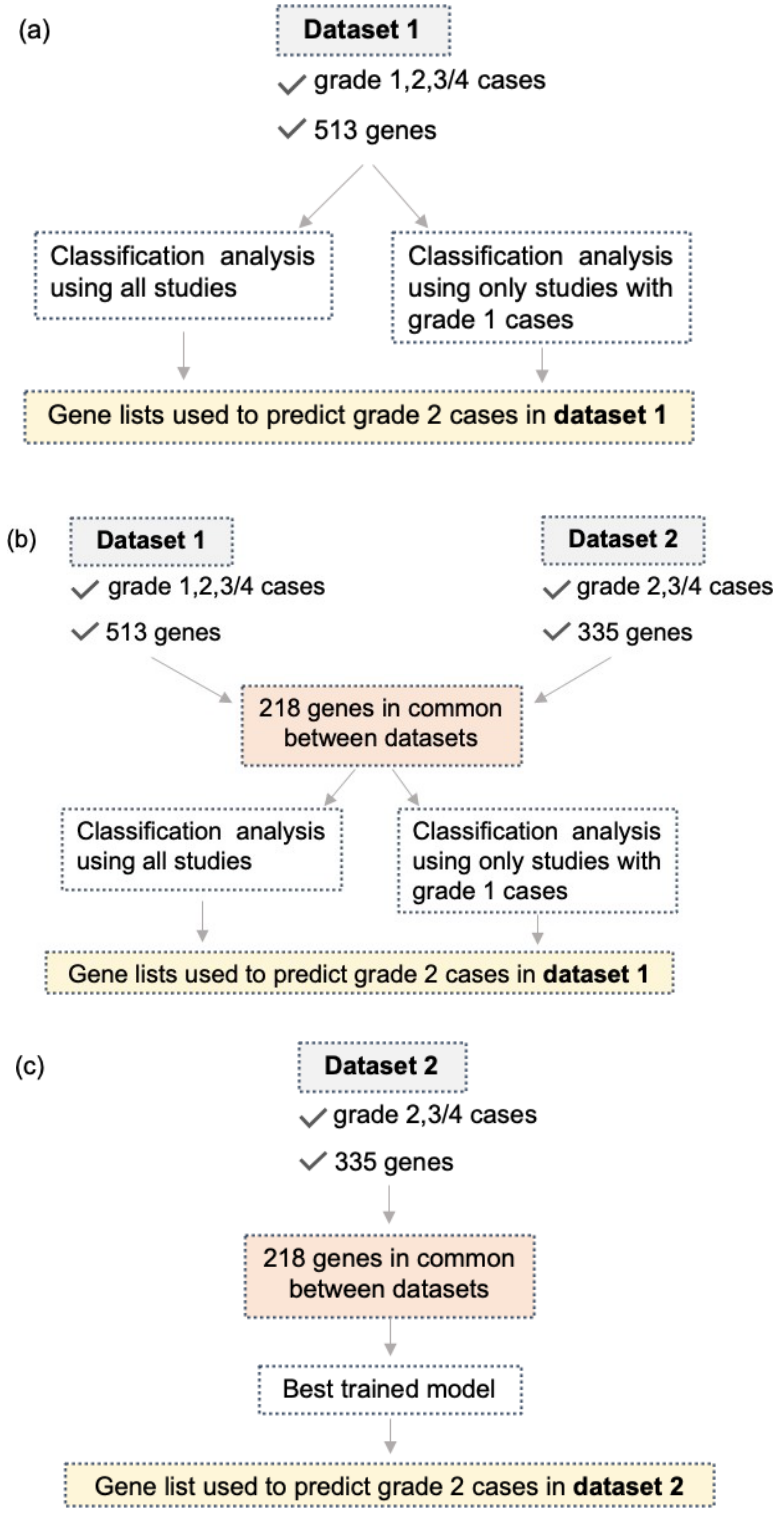


Figure 2.3 Classification and prediction analyses pipeline used to reclassify grade 2 serous ovarian tumours. (Source: R Studio – version 1.2.5001).

1.2.3.2 Prediction model

The function named *Predict method for Linear Model Fits* from R Studio was subsequently used to predict which of the grade 2 samples were more similar to the grade 1 samples, and which were more similar to the grade 3/4 samples based on their gene expression data. This allowed the identification of grade 2 tumours that were more likely to be LGSOC instead of HGSOC.

1.3 Results

1.3.1 Classification of grade 2 samples in dataset 1 using a panel of 513 genes

1.3.1.1 Classification analysis

Dataset 1 had RNA expression data available for 513 genes in total. For the all-studies analysis, a total of 43 grade 1 cases and 2,447 grade 3/4 cases were included (19 studies). For the reduced set of studies, a total of 40 grade 1 cases and 1,351 grade 3/4 cases were included in the analysis (5 studies). Data were input into the Random forest model and a total of 500 multiple trees were performed to distinguish grade 1 from grade 3/4 cases in each analysis. Random forest outputs are shown in Table 2.4. 2,490 and 1,391 observations were identified in the all-studies and in the reduced studies analyses, respectively. Six out of 43 cases were misclassified in the grade 1 class for the all-studies analysis, whereas seven out of 40 cases were misclassified in the grade 1 class for the reduced studies analysis. For the grade 3/4 class, 122 out of 2,447 cases were misclassified in the all-studies analysis, and 49 out of 1,351 cases were misclassified in the reduced studies analysis (Table 2.4). The OOB error rates were estimated to be 5.14% using all-studies (6 + 122 / 2,490 = 5.14%) and 4.03% using reduced studies (7 + 49 / 1,391 = 4.03%) (Table 2.4).

Table 2.4 Random forest classification using 513 genes expression in dataset 1.

	All-studies				*Reduced studies*			
	Grade 1	**Grade 3/4**	**Class error**	**OOB error**	**Grade 1**	**Grade 3/4**	**Class error**	**OOB error**
Grade 1	37	6	14%	5.14%	33	7	17%	4.03%
Grade 3/4	122	2,325	5%		49	1,302	4%	
Total	159	2,331	-		82	1,309	-	

Abbreviations: OOB - Out-of-Bag

An importance table was subsequently generated. The genes were sorted by the mean decrease accuracy score to rank which genes were the most important to distinguish grade 1 from grade 3/4 cases. Random forest was then similarly performed using 3 new sets of genes ranked by the importance table: the 20 most important genes, the 40 most important genes and the 60 most important genes (Table 2.5). For the all-studies analysis, the smallest error was observed in classifying grade 1 cases, as well as the smallest OOB error, when the top 40 most important genes were used (Table 2.5). For the reduced studies analysis, the top 60 most important genes were selected based on the smallest OOB error and second smallest minor class error (Table 2.5). Lists of the top 10 ranked genes for each analysis are shown in Table 2.6. Table 2.6 includes each mean decrease accuracy value and the average expression of each gene in the grade 1, as well as in the grade 3/4 cases. Seven genes (*CD302, ANKRD1, VSIG4, PDE6D, ERGIC3, MAK* and *PTGS2*) overlapped between the most important genes of all-studies and reduced studies. Full lists of the most important genes used to distinguish grade 1 from grade 3/4 cases in each analysis are given in Supplementary Table 2.1.

Table 2.5 Random forest classification using the top 20, top 40 and top 60 most important genes in dataset 1.

	All-studies				*Reduced studies*			
	Grade 1	**Grade 3/4**	**Class error**	**OOB error**	**Grade 1**	**Grade 3/4**	**Class error**	**OOB error**
				Top 20 genes				
Grade 1	35	8	18%		36	4	10%	
Grade 3/4	138	2,309	6%	5.86%	54	1,297	4%	4.17%
Total	173	2,317	-		90	1,301	-	
				Top 40 genes				
Grade 1	36	7	**16%**		34	6	15%	
Grade 3/4	115	2,332	5%	**4.90%**	43	1,308	3%	3.52%
Total	151	2,339	-		77	1,314	-	
				Top 60 genes				
Grade 1	35	8	18%		35	5	**12%**	
Grade 3/4	124	2,323	5%	5.30%	36	1,315	2%	**2.95%**
Total	159	2,331	-		71	1,320	-	

Abbreviations: OOB - Out-of-Bag

Table 2.6 List of the 10 most important genes used to distinguish between grade 1 and grade 3/4 cases in dataset 1.

Rank	*All-studies*				*Reduced studies*			
	Genes	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**	**Genes**	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**
1	*GCNT3*	2.36	-5.45	-6.60	*PSRC1*	3.29	-5.06	-3.70
2	*MRE11A*	2.32	-3.57	-3.52	*PAX8*	3.15	-1.25	-2.23
3	*FANCG*	2.25	-5.15	-4.62	*RICTOR*	2.59	-4.05	-4.02
4	*GK5*	2.20	-4.04	-3.71	*ISG15*	2.55	-3.97	-1.57
5	*KIF4A*	2.09	-6.56	-5.64	*BIRC5*	2.38	-4.65	-2.72
6	*KIF24*	2.03	-5.33	-4.90	*FOXJ1*	2.05	-2.11	-3.47
7	*FOXRED2*	1.99	-4.47	-4.10	*PAX2*	2.01	-5.25	-7.24
8	*LRRC15*	1.96	-4.77	-5.55	*EGFL6*	2.00	-6.61	-5.19
9	*BAX*	1.91	-1.27	-1.87	*BABAM1*	1.92	-2.19	-1.26
10	*EPB41L3*	1.87	-3.45	-3.61	*RSPO1*	1.92	-3.79	-4.77

Abbreviations: Avg exp – Average of expression; MDA: mean decrease accuracy

1.3.1.2 Prediction analysis

The selected panels of the most important genes that best distinguished grade 1 from grade 3/4 cases in dataset 1 were then used as the model object on the grade 2 cases to predict which were more likely to be LGSOC instead of HGSOC. Dataset 1 had a total of 489 grade 2 samples and according to the prediction model, using all- studies, 69 of them (14.1%) were reclassified as LGSOC and 420 (85.9%) were classed as HGSOC (Table 2.7). When the prediction model using reduced studies was implemented, 57 of 489 grade 2 cases (12%) were reclassified as LGSOC and 432 (88%) were classified as HGSOC (Table 2.7). Comparing the samples' predictions between all-studies and the reduced set of studies, a total of 451 grade 2 samples (92%) had the same prediction in which 44 (9.8%) were samples reclassified as LGSOC and 407 (90.2%) were samples classed as HGSOC (Table 2.7). Only 38 cases (8%) had a discordant classification.

Table 2.7 Predictions concordance between all-studies *versus* reduced studies analyses with the 513 genes expression panel.

		All-studies		
		Predicted LGSOC	Predicted HGSOC	Total
Reduced studies	Predicted LGSOC	**44**	25	69 (14%)
	Predicted HGSOC	13	**407**	420 (86%)
	Total	57 (12%)	432 (88%)	489

Abbreviations: LGSOC – low-grade serous ovarian cancer; HGSOC – high-grade serous ovarian cancer

1.3.2 Classification of grade 2 samples in dataset 1 using panel of 218 overlapped genes

1.3.2.1 Classification analysis

Dataset 2 had no grade 1 tumours, so a second model was trained on dataset 1 using only genes occurring in both two datasets. A total of 218 genes occurred in common between both datasets 1 and 2. Classification using Random forest and prediction models were similarly performed on dataset 1 (all-studies and reduced studies) using this overlapped panel of 218 overlapped genes.

In the all-studies analysis, the classification using the whole set of 218 genes best distinguished grade 1 from grade 3/4 cases (OOB error: 3.78%; grade 1 class error: 12%) and in the reduced studies analysis, the top 60 most important genes provided the best classification result (OOB error: 2.59%; grade 1 class error: 12%) (Table 2.8).

Table 2.8 Random forest classification using the most important genes and all overlapped set of 218 genes in dataset 1.

	All-studies				*Reduced studies*			
	Grade 1	**Grade 3/4**	**Class error**	**OOB error**	**Grade 1**	**Grade 3/4**	**Class error**	**OOB error**
	Top 20 genes							
Grade 1	34	9	21%	6.67%	36	4	10%	3.88%
Grade 3/4	157	2,290	6%		50	1,301	4%	
Total	191	2,299	-		86	1,305	-	
	Top 40 genes							
Grade 1	33	10	23%	5.58%	32	8	20%	3.74%
Grade 3/4	129	2,318	5%		44	1,307	3%	

	All-studies				*Reduced studies*			
	Grade 1	**Grade 3/4**	**Class error**	**OOB error**	**Grade 1**	**Grade 3/4**	**Class error**	**OOB error**
Total	162	2,328	-		76	1,315	-	
Top 60 genes								
Grade 1	34	9	21%		35	5	**12%**	
Grade 3/4	113	2,334	5%	4.90%	31	1,320	3%	**2.59%**
Total	147	2,343	-		66	1,325	-	
All 218 overlapped genes								
Grade 1	38	5	**12%**		35	5	12%	
Grade 3/4	89	2,358	4%	**3.78%**	44	1,307	3%	3.52%
Total	127	2,363	-		79	1,312	-	

Abbreviations: OOB - Out-of-Bag

Table 2.9 shows only the first 10 most important overlapped genes that were used to distinguish grade 1 from grade 3/4 tumours in all and reduced sets of studies. The full tables with the 218 overlapped genes (best error in the all-studies analysis) as well as the top 60 most important genes (best error in the reduced studies analysis) are given in Supplementary Table 2.2.

In the all-studies analysis, 19 genes overlapped with the previous list of the 40 most important genes using the full set of 513 genes expression (*ANKRD1, MAK, PPP2R4, VSIG4, CRISPLD2, LUM, CTNNBL1, POSTN, EPB41L3, MAP4K3, B4GALT5, ESD, CTLA4, FOXRED2, LRRC15, ASRGL1, MRE11A, GMPR* and *IDO1*). In the reduced studies analysis, 11 genes overlapped with the previous list of the 60 most important genes (*MAK, BAALC, TLR4, CCL5, LRRC15, CX3CR1, CD302, TMEM45A, PAX8, ANKRD1* and *TSHR*).

Table 2.9 List of the 10 most important overlapped genes used to distinguish grade 1 from grade 3/4 cases in dataset 1.

Rank	*All-studies*				*Reduced studies*			
	Genes	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**	**Genes**	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**
1	*MITF*	2.86	-5.55	-6.08	*PAX8*	4.08	-1.25	-2.23
2	*ANKRD1*	2.81	-5.08	-7.50	*MDM2*	2.77	-2.65	-3.38
3	*MDM2*	2.57	-2.67	-3.33	*ANKRD1*	2.70	-5.14	-7.32
4	*CDH1*	2.56	-0.85	-1.26	*KIF1A*	2.67	-6.03	-3.87
5	*FOXJ1*	2.47	-2.17	-3.42	*CD74*	2.60	2.52	2.13

Rank	*All-studies*				*Reduced studies*			
	Genes	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**	**Genes**	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**
6	*NOTCH3*	2.41	-2.64	-2.06	*ADAMDEC1*	2.59	-7.01	-5.71
7	*PAX8*	2.40	-1.31	-2.24	*EZR*	2.40	0.00	-0.77
8	*MAK*	2.39	-5.19	-6.47	*MRPS27*	2.39	-2.46	-3.21
9	*ATP5A1*	2.29	-0.20	-0.12	*INHBA*	2.36	-4.28	-4.64
10	*CD8A*	2.27	-5.15	-5.18	*CDH1*	2.28	-0.83	-1.25

Abbreviations: Avg exp – Average of expression; MDA – mean decrease accuracy

1.3.2.2 Prediction analysis

For the all-studies analysis, the selected panel of 218 genes was used as the new model object for the prediction of grade 2 samples. Out of the 489 grade 2 cases, 60 cases (12%) were reclassified as LGSOC and 429 (88%) were classed as HGSOC (Table 2.10). For the reduced studies analysis, using the selected top 60 genes, a similar classification was observed: 57 grade 2 cases (12%) were reclassified as LGSOC, and 432 cases (88%) were classed as HGSOC (Table 2.10). A comparison between the all-studies prediction *versus* reduced studies prediction was performed. Out of all 489 grade 2 cases, 462 had the same prediction (94.5%) in which 45 of these samples were reclassified as LGSOC (9.7%) and 417 were classed as HGSOC (90.3%). Only 27 predictions were discordant (5.5%) (Table 2.10).

Table 2.10 Predictions concordance between all-studies *versus* reduced studies analyses using 218 overlapped genes expression panel.

		All-studies		
		Predicted LGSOC	Predicted HGSOC	Total
Reduced studies	Predicted LGSOC	**45**	12	57 (12%)
	Predicted HGSOC	15	**417**	432 (88%)
	Total	60 (12%)	429 (88%)	489

Abbreviations: LGSOC – low-grade serous ovarian cancer; HGSOC – high-grade serous ovarian cancer

1.3.3 Concordance test comparing predictions in dataset 1

Additional concordance tests were performed using all-studies and reduced studies analyses comparing the 513 panel of genes *versus* 218 overlapped panel of genes. These comparisons were performed to confirm if the different panel of genes were predicting the same samples as either LGSOC or HGSOC. The results are given

in Table 2.11 and Table 2.12 by study to evaluate if there was any heterogeneity or systematic issue with any particular study.

Overall concordance rate was 94% between the 2 panel of genes for both all and reduced set of studies analyses. Concordance ranged from 80% to 100% between studies. All studies showed high concordance of predictions using either the panel of 513 genes or the panel of 218 overlapped genes. Studies that had a concordance rate between approximately 80% to 85% had less than 10 samples, however, which might have resulted in an inaccurate concordance rate (Tables 2.11 and 2.12).

Table 2.11 Concordance rates by study using all-studies analyses.

Studies	Concordant HGSOC pred	Concordant LGSOC pred	Discordant pred	Total	Concordance rate (%)
NCO	111	9	6	126	95
AOC	78	9	7	94	93
DOV	52	9	1	62	98
WMH	30	1	0	31	100
SRF	27	1	2	30	93
POC	17	7	4	28	86
TRI	16	5	2	23	91
TRT	18	2	1	21	95
POL	18	1	1	20	95
VAN	17		1	18	94
UKO	10	1	2	13	85
USC	6	1	0	7	100
MAY	2	2	1	5	80
GER	4		1	5	80
SEA	2	1	0	3	100
LAX	1	1	0	2	100
AOV	1		0	1	100
Total	**410**	**50**	**29**	**489**	**94**

Abbreviations: HGSOC – high-grade serous ovarian cancer; LGSOC – low-grade serous ovarian cancer; pred - prediction

Table 2.12 Concordance rates by studies using reduced studies analyses.

Studies	Concordant HGSOC pred	Concordant LGSOC pred	Discordant pred	Total	Concordance rate (%)
NCO	112	7	7	126	94
AOC	81	6	7	94	93

Studies	Concordant HGSOC pred	Concordant LGSOC pred	Discordant pred	Total	Concordance rate (%)
DOV	51	8	3	62	95
WMH	30	1	0	31	100
SRF	26	2	2	30	93
POC	20	6	2	28	93
TRI	16	4	3	23	87
TRT	16	3	2	21	90
POL	19	1	0	20	100
VAN	16	0	2	18	89
UKO	12	1	0	13	100
USC	6	0	1	7	86
MAY	2	2	1	5	80
GER	5	0	0	5	100
SEA	3	0	0	3	100
LAX	1	1	0	2	100
AOV	1	0	0	1	100
Total	**417**	**42**	**30**	**489**	**94**

Abbreviations: HGSOC – high-grade serous ovarian cancer; LGSOC – low-grade serous ovarian cancer; pred - prediction

1.3.4 Prediction of grade 2 samples in dataset 2

Predictions were finally performed in the second dataset. Best trained models using the overlapped set of genes between all-studies as well as reduced studies were used to predict 166 grade 2 cases from dataset 2. Nine grade 2 cases (5%) were reclassified as LGSOC and 157 (95%) classified as HGSOC using the all-studies trained model (Table 2.13). Likewise, 10 grade 2 samples (6%) were reclassified as LGSOC and 156 (94%) were classed as HGSOC using the reduced studies trained model (Table 2.13).

Table 2.13 Grade 2 cases predictions performed in dataset 2.

	Prediction	Total
All studies	Predicted LGSOC	9 (5.4%)
	Predicted HGSOC	157 (94.6%)
	Total	**166 (100%)**
Reduced studies	Predicted LGSOC	10 (6%)
	Predicted HGSOC	156 (94%)
	Total	**166 (100%)**

Abbreviations: HGSOC – high-grade serous ovarian cancer; LGSOC – low-grade serous ovarian cancer

1.3.5 Prediction models' accuracy

Towards the end of this analysis, *TP53* IHC data became available for a subset of serous cases in both dataset 1 and 2. The IHC data was then used to check the accuracy of the RNA expression prediction. The accuracy was calculated dividing the proportion of true positives and true negatives by the total number of cases that had IHC data available. Importantly, a large number of tumours lacked this data, so there were limitations to this benchmarking.

The IHC data were available to 173 of the 489 grade 2 samples in dataset 1 – 30 of the samples had no mutation and 143 had mutant *TP53*. Firstly, accuracy was calculated using the prediction generated from the entire panel of 513 genes expression. In the all-studies analysis, out of the 30 predicted LGSOC, 24 were *TP53* wild-type and 6 were mutant *TP53*; thus prediction accuracy was 80% (Table 2.14). Similarly, out of the 143 predicted HGSOC, 131 were mutant *TP53* and 12 were *TP53* wild-type, resulting in an accuracy of 91.6% (Table 2.14). In the reduced studies analysis, out of 22 predicted LGSOC, 21 had no *TP53* mutations and 1 was mutant (95.5% accuracy), and out of 151 predicted HGSOC, 136 cases were mutant and 15 were wild-type (accuracy 90.1%) – see Table 2.14. Better LGSOC prediction accuracy was observed in the reduced studies analysis, and similar accuracy was observed for the HGSOC predictions

Table 2.14 Prediction accuracy in dataset 1 – 513 genes expression analysis.

	Prediction	*TP53* wild-type	Mutant *TP53*	NA	Total	Accuracy*
All-studies	Predicted LGSOC	24	6	39	69 (14.1%)	80%
	Predicted HGSOC	12	131	277	420 (85.9%)	91.6%
	Total	**36**	**137**	**316**	**489 (100%)**	**-**
Reduced studies	Predicted LGSOC	21	1	35	57 (11.7%)	95.5%
	Predicted HGSOC	15	136	281	432 (88.3%)	90.1%
	Total	**36**	**137**	**316**	**489 (100%)**	**-**

Abbreviations: NA – not available; HGSOC – high-grade serous ovarian cancer; LGSOC – low-grade serous ovarian cancer. * Accuracy calculated based on *TP53* IHC data.

Secondly, accuracy was calculated using the prediction generated from the overlapped panel of 218 genes expression. Twenty-five predicted LGSOC had *TP53* data available of which 21 were *TP53* wild-type and 4 were mutant *TP53*, thus accuracy was 84% (Table 2.15). Likewise, 148 predicted HGSOC had IHC data available, of which 133 were mutant *TP53* and 15 were *TP53* wild-type, so accuracy was 90% (Table 2.15).

Accuracy was slightly higher in the reduced studies analysis – LGSOC and HGSOC predictions had 91% and 90% accuracy, respectively.

Table 2.15 Prediction accuracy in dataset 1 – 218 genes expression analysis.

	Prediction	*TP53* wild-type	Mutant *TP53*	NA	Total	Accuracy*
All studies	Predicted LGSOC	21	4	35	60 (12.3%)	84%
	Predicted HGSOC	15	133	281	429 (87.7%)	89.9%
	Total	**36**	**137**	**316**	**489 (100%)**	**-**
Reduced studies	Predicted LGSOC	21	2	34	57 (11.7%)	91.3%
	Predicted HGSOC	15	135	282	432 (88.3%)	90%
	Total	**36**	**137**	**316**	**489 (100%)**	**-**

Abbreviations: NA – not available; LGSOC – low-grade serous ovarian cancer; HGSOC – high-grade serous ovarian cancer. * Accuracy calculated based on *TP53* IHC data.

The predictions in dataset 2 were not benchmarked given that *TP53* IHC data were missing for the majority of the cases (only 37% of the cases had IHC available). In summary, based on the different models generated by Random forest using dataset 1 expression data, approximately 12% to 14% of the grade 2 cases should be LGSOC instead of HGSOC. Accuracy ranged from 80 to 95% when RNA expression data was benchmarked with *TP53* immunohistochemistry. The reduced set of studies showed better accuracy when predicting the misclassified LGSOC using either the full set of 513 genes expression or the genes in common between dataset 1 and 2 (218 overlapped genes) (Figure 2.4).

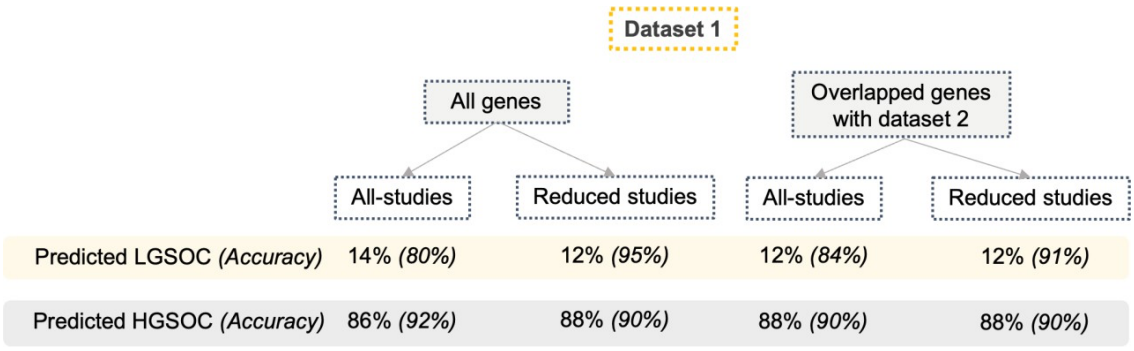


Figure 2.4 Flowchart summarising the prediction results and its accuracy from the different Random forest models used in dataset 1.

1.4 Discussion

Analysis using all 513 genes showed that the reduced studies modelling resulted in better LGSOC prediction accuracy than the all-studies modelling (all-studies: 14% of grade 2 was predicted as LGSOC with 80% accuracy *versus* reduced studies: 12% reclassified samples with 95% accuracy). However, accuracy was slightly better for the HGSOC prediction (all-studies: 86% HGSOC with 92% accuracy *versus* reduced studies: 88% HGSOC with 90% accuracy). Of these predicted samples, 44 were concordant LGSOC samples and 407 were concordant HGSOC between the all-studies and reduced studies models. Analysis using only the overlapped set of 218 genes showed that the reduced studies modelling also resulted in better LGSOC prediction accuracy (all-studies: 12% of grade 2 predicted as LGSOC with 84% accuracy *versus* 12% reclassified samples with 95% accuracy for the reduced studies). On the other hand, HGSOC prediction accuracy remained the same (all-studies and reduced studies: 88% predicted HGSOC with 90% accuracy). Within these predictions, 45 predicted LGSOC and 417 predicted HGSOC samples were concordant between the all-studies and reduced studies models. Overall, 92% of the 489 grade 2 samples were concordant between the all-studies *versus* the reduced studies analysis using all 513 genes, and 94% were concordant using the overlapped-set of 218 genes.

Tests comparing the two sets of panels of genes also showed high concordance rates (513 genes *versus* 218 overlapped genes). The overall concordance rate was 94% between these two panels of genes in both all and reduced sets of studies. Concordance was divided by studies to assess any systematic issue with a particular study, but all-studies showed high concordance of predictions using either the panel of 513 genes or the panel of 218 overlapped genes (concordance between studies ranged from 80 to 100%).

The use of bioinformatics models allowed the identification of a classifier that could potentially distinguish between low- and high-grade serous ovarian cancer based on RNA expression data. Two models were used for this analysis: a classification model that selected the best panel of genes to differentiate LGSOC (grade 1) from HGSOC (grade 3/4), and a prediction model that classified which serous grade 2 tumours were likely to be LGSOC or HGSOC. As previously described, around 12% to 14% of grade 2 cases in dataset 1 were reclassified as LGSOC. However, fewer grade 2 cases were predicted to be LGSOC in dataset 2: around 5% to 6% of the cases. This could be due to improvements in the pathology classification, such as better and more IHC markers, better pathologists, as well as the addition of molecular analysis that has expanded the pathology field.

The high accuracy of the prediction models when benchmarking gene expression with the gold standard IHC suggested that misclassified LGSOC could be identified in

the context of these data. However, this analysis had some limitations. Most of the samples did not have IHC data, so a subset of the samples was used to determine the accuracy of the classifications. Additionally, although IHC is the current gold-standard methodology to discriminate between high and low-grade serous cases, it can still miss some mutations (159). In addition, pathology classification changes over time as previously shown by Kommoss et al. (2016). The researchers had an experienced gynaecopathologist (pathologist A) reviewing pathology specimens for translational research purposes in 2002. The same pathologist (pathologist A), and another gynaecopathologist (pathologist B) reassessed the same cohort in 2016. Pathologist A reported the same histotype diagnosis in only 54% of cases, and there was a very high degree of interobserver reproducibility for the 2016 review (98%) (166). For this thesis, a pathology review of only the predicted LGSOC cases was performed. These cases were all found to have a LGSOC morphology, therefore confirming the initial hypothesis that some grade 2 cases are misclassified as HGSOC when they are in fact LGSOC cases. The pathology review also confirmed that the LGSOC predictions were accurate. Overall, conventional features would be enough to distinguish true LGSOC from HGSOC. But this needs to be performed by the right pathologist, since the cases used for this analysis were recently reviewed as HGSOC for the NanoString project.

NanoString technology has many advantages due to the fact that direct count of native RNA system increases robustness and simplifies raw data analysis. However, in the clinical setting, it is still more likely that immunohistochemistry would be chosen over RNA expression. Nevertheless, in the context of this project, a very large number of serous tumours had their RNA expression already measured by NanoString for many genes within the OTTA tumour expression project, which allowed these analyses to be performed. As this classifier showed high accuracy with *TP53* IHC data, gene expression could potentially replace pathology review to save time and eliminate the need for expert pathology review in retrospective studies that had RNA available for extraction.

The aim of this analysis was to evaluate whether gene expression data could identify grade 2 serous cases that were likely to be LGSOC, so these additional cases could be included in the following sequencing projects in an effort to increase the sample size. The high accuracy of the prediction models suggested that misclassified LGSOC cases could be identified, therefore these potential additional samples were evaluated. The initial hypothesis was that approximately 20% of the grade 2 cases would be reclassified as LGSOC based on previous data (82). It was found that approximately 12% to 14% might be misclassifications, a slightly lower frequency of misclassifications than expected. In total, there was gene expression data available for 655 grade 2 cases – 414 samples were from studies with suitable sequencing consent (63%). Overall, 67

grade 2 samples were reclassified as LGSOC by both prediction models, but only 11 of these cases had suitable WES sequencing consent, and 44 had suitable targeted sequencing consent. In the design of the WES analysis, more than 200 LGSOC cases were planned to be included, thus the addition of 11 would not significantly improve the sample size. For the targeted sequencing run, it was estimated that ~300 LGSOC cases would be included in the pilot-study and more than 350 would be included in the follow-up study which again indicated that the addition of 44 cases would not improve the statistical analysis. In addition, the COVID-19 restrictions in Australia and overseas had a major impact on the logistics of the downstream sequencing projects. The many restrictions affected if collaborating overseas institutions could prepare and ship samples when needed.

In conclusion, classification using NanoString RNA profiling data could be used to distinguish which grade 2 serous ovarian cancers are LGSOC or HGSOC in the research setting. This is not currently feasible or needed clinically as is still more likely that immunohistochemistry would be chosen over RNA expression. Due to the small number of eligible grade 2 cases predicted to be LGSOC, these potentially misclassified samples will not be included in the following sequencing projects despite the positive results from these analyses.

2 Appendices

Supplementary Table 2.1. Full lists of the most important genes used to distinguish grade 1 from grade 3/4 cases in dataset 1 using 513 genes.

Rank	*All studies*				*Reduced studies*			
	Genes	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**	**Genes**	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**
1	*GCNT3*	2.36	-5.45	-6.6	*PSRC1*	3.29	-5.06	-3.7
2	*MRE11A*	2.32	-3.57	-3.52	*PAX8*	3.15	-1.25	-2.23
3	*FANCG*	2.25	-5.15	-4.62	*RICTOR*	2.59	-4.05	-4.02
4	*GK5*	2.2	-4.04	-3.71	*ISG15*	2.55	-3.97	-1.57
5	*KIF4A*	2.09	-6.56	-5.64	*BIRC5*	2.38	-4.65	-2.72
6	*KIF24*	2.03	-5.33	-4.9	*FOXJ1*	2.05	-2.11	-3.47
7	*FOXRED2*	1.99	-4.47	-4.1	*PAX2*	2.01	-5.25	-7.24
8	*LRRC15*	1.96	-4.77	-5.55	*EGFL6*	2	-6.61	-5.19
9	*BAX*	1.91	-1.27	-1.87	*BABAM1*	1.92	-2.19	-1.26
10	*EPB41L3*	1.87	-3.45	-3.61	*RSPO1*	1.92	-3.79	-4.77
11	*IDO1*	1.84	-4.17	-3.76	*CD302*	1.87	-3.77	-4.43
12	*PPP2R4*	1.82	-3.32	-3.36	*HBB*	1.85	-3.7	-4.59
13	*IFT88*	1.8	-3.41	-4.34	*EPB41L3*	1.83	-3.44	-3.57
14	*ESD*	1.73	-2.04	-2.64	*ANKRD1*	1.78	-5.14	-7.32
15	*S100A6*	1.73	2.97	1.96	*ANKRA2*	1.78	-4.28	-5.15
16	*CTLA4*	1.71	-5.8	-6	*MITF*	1.77	-5.49	-6.07
17	*BRAF*	1.71	-2.66	-2.69	*VSIG4*	1.76	-3.38	-3.38
18	*CTNNBL1*	1.71	-3.56	-3.38	*PDE6D*	1.74	-5.1	-5.07
19	*GPR64*	1.69	-2.82	-3.92	*KRAS*	1.73	-2.76	-2.33
20	*DHX35*	1.69	-5.1	-4.91	*SRI*	1.71	-1.07	-1.04
21	*STC1*	1.69	-4.85	-5.54	*KLK7*	1.7	-3.06	-2.66
22	*VSIG4*	1.68	-3.37	-3.54	*ERGIC3*	1.67	-1.34	-1.31
23	*ERGIC3*	1.65	-1.34	-1.24	*TSC1*	1.64	-3.14	-3.5
24	*PDE6D*	1.62	-5.09	-5.05	*MAK*	1.64	-5.09	-6.35
25	*CRISPLD2*	1.61	-3.45	-3.99	*PSME2*	1.61	-0.69	-0.06
26	*MAP4K3*	1.6	-2.59	-2.62	*CDK9*	1.6	-2.78	-3.1

Rank	*All studies*				*Reduced studies*			
	Genes	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**	**Genes**	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**
27	*RNF114*	1.6	-2.76	-2.5	*BAALC*	1.59	-5.18	-6.07
28	*ANKRD1*	1.6	-5.08	-7.5	*FOXC2*	1.58	-6.27	-7.13
29	*LUM*	1.58	-0.54	-1.16	*TSHR*	1.57	-8.53	-7.99
30	*FHIT*	1.56	-5.79	-6.15	*PTGS2*	1.55	-5.08	-5.9
31	*CEACAM5*	1.54	-7.05	-7.72	*FAM126B*	1.54	-3.68	-3.6
32	*GMPR*	1.52	-2.84	-2.35	*S100A6*	1.53	3	2.01
33	*HIST1H2BG*	1.49	-1.83	-1.07	*CD8A*	1.53	-5.11	-4.97
34	*ASRGL1*	1.49	-1.86	-2.28	*MyD88*	1.52	-2.39	-2.3
35	*PLAC4*	1.49	-7.8	-8.53	*FOXRED2*	1.52	-4.45	-4.05
36	*MROH5*	1.48	-9.35	-9.25	*FKSG2*	1.48	-8.16	-8.25
37	*POSTN*	1.46	-4	-4.35	*MEST*	1.48	-2.52	-2.13
38	*MAK*	1.45	-5.19	-6.47	*TMEM45A*	1.47	-3.87	-3.77
39	*CAMK1*	1.42	-4.62	-4.89	*COL11A1*	1.46	-3.15	-2.71
40	*B4GALT5*	1.41	-1.97	-1.81	*CCL5*	1.45	-3.88	-3.5
41	-	-	-	-	*KIF4A*	1.45	-6.55	-5.6
42	-	-	-	-	*KIF3B*	1.44	-3.57	-3.61
43	-	-	-	-	*GFPT2*	1.44	-5.38	-6.05
44	-	-	-	-	*FOLR1*	1.44	-2.43	-1.42
45	-	-	-	-	*CX3CR1*	1.43	-4.55	-5.09
46	-	-	-	-	*PEX3*	1.43	-3.09	-3.22
47	-	-	-	-	*PIK3CA*	1.43	-3.22	-2.87
48	-	-	-	-	*SLC25A19*	1.42	-5.2	-5.19
49	-	-	-	-	*PTGS1*	1.42	-2.9	-2.22
50	-	-	-	-	*PBX1*	1.42	-1.17	-1.06
51	-	-	-	-	*FOXP3*	1.41	-6.8	-6.87
52	-	-	-	-	*FAM58A*	1.41	-4	-3.81
53	-	-	-	-	*PJA2*	1.41	-1.4	-2.04
54	-	-	-	-	*KIF24*	1.41	-5.25	-4.83
55	-	-	-	-	*CTLA4*	1.4	-5.73	-5.81
56	-	-	-	-	*PKM2*	1.4	1.01	1.21
57	-	-	-	-	*TLR4*	1.4	-4.06	-4.26

Rank	*All studies*				*Reduced studies*			
	Genes	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**	**Genes**	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**
58	-	-	-	-	*CCNE1*	1.39	-5.14	-3.03
59	-	-	-	-	*LRRC15*	1.39	-4.78	-5.32
60	-	-	-	-	*BRAF*	1.38	-2.63	-2.69

Supplementary Table 2.2. Full lists of the most important genes used to distinguish grade 1 from grade 3/4 cases in dataset 1 using 218 overlapped genes

Rank	*All studies*				*Reduced studies*			
	Genes	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**	**Genes**	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**
1	*MITF*	2.86	-5.55	-6.08	*PAX8*	4.08	-1.25	-2.23
2	*ANKRD1*	2.81	-5.08	-7.50	*MDM2*	2.77	-2.65	-3.38
3	*MDM2*	2.57	-2.67	-3.33	*ANKRD1*	2.70	-5.14	-7.32
4	*CDH1*	2.56	-0.85	-1.26	*KIF1A*	2.67	-6.03	-3.87
5	*FOXJ1*	2.47	-2.17	-3.42	*CD74*	2.60	2.52	2.13
6	*NOTCH3*	2.41	-2.64	-2.06	*ADAMDEC1*	2.59	-7.01	-5.71
7	*PAX8*	2.4	-1.31	-2.24	*EZR*	2.40	0.00	-0.77
8	*MAK*	2.39	-5.19	-6.47	*MRPS27*	2.39	-2.46	-3.21
9	*ATP5A1*	2.29	-0.20	-0.12	*INHBA*	2.36	-4.28	-4.64
10	*CD8A*	2.27	-5.15	-5.18	*CDH1*	2.28	-0.83	-1.25
11	*TCF7L1*	2.14	-4.23	-4.36	*OPA1*	2.21	-2.71	-2.40
12	*PTH2R*	2.10	-6.99	-5.15	*BAALC*	2.05	-5.18	-6.07
13	*GFRA1*	2.09	-5.87	-6.50	*POSTN*	2.03	-4.07	-4.12
14	*ADH1B*	2.09	-4.31	-5.38	*TCF7L1*	2.01	-4.24	-4.36
15	*RBMS3*	2.01	-4.06	-5.35	*WWP1*	1.99	-2.47	-2.59
16	*ABCB1*	2.01	-5.16	-5.73	*AR*	1.99	-3.48	-4.05
17	*IGHM*	2.00	-4.76	-3.10	*CD302*	1.98	-3.77	-4.43
18	*KLHL7*	1.92	-4.32	-4.42	*LRRC15*	1.97	-4.78	-5.32
19	*NTRK2*	1.88	-4.74	-5.56	*HIST1H2AM*	1.90	-1.54	-0.23
20	*USP8*	1.81	-2.51	-2.64	*SLAMF7*	1.90	-4.89	-4.23
21	*SPARC*	1.81	1.74	1.38	*KLHL7*	1.88	-4.28	-4.39
22	*CTSK*	1.69	-1.81	-2.45	*SOX17*	1.87	-1.63	-1.49
23	*PPP2R4*	1.67	-3.32	-3.36	*ATP5A1*	1.87	-0.22	-0.08
24	*ENOX1*	1.66	-4.79	-5.69	*FAP*	1.85	-4.20	-4.45
25	*TIMP3*	1.64	-0.88	-1.15	*MAK*	1.82	-5.09	-6.35
26	*CYTIP*	1.60	-4.56	-4.73	*IGKC*	1.78	-1.52	0.84
27	*CD74*	1.59	2.52	2.12	*NUAK2*	1.76	-2.73	-2.77
28	*VSIG4*	1.59	-3.37	-3.54	*HIF1A*	1.74	-0.92	-1.31
29	*HSP90AA1*	1.59	0.74	0.64	*DUSP4*	1.72	-2.79	-5.23

Rank	*All studies*				*Reduced studies*			
	Genes	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**	**Genes**	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**
30	*SLAMF7*	1.57	-4.98	-4.39	*ZNF423*	1.70	-4.57	-5.25
31	*RAD51C*	1.53	-3.19	-3.27	*CPNE1*	1.67	-1.56	-1.49
32	*CD38*	1.50	-6.10	-5.72	*TESK1*	1.66	-4.58	-4.43
33	*RNASEL*	1.49	-3.87	-4.45	*COL3A1*	1.66	2.45	2.07
34	*SLAMF8*	1.48	-6.13	-6.38	*SCGB1D2*	1.65	-5.33	-4.92
35	*HNF1B*	1.45	-5.39	-6.47	*APC*	1.63	-4.09	-4.48
36	*SMO*	1.43	-3.64	-4.10	*CCL5*	1.52	-3.88	-3.50
37	*PGRB*	1.43	-6.02	-7.13	*TMEM45A*	1.51	-3.87	-3.77
38	*CD27*	1.41	-5.86	-5.79	*UQCC*	1.49	-3.54	-3.30
39	*STK16*	1.39	-3.30	-3.34	*TUBB6*	1.47	-2.27	-2.51
40	*EZR*	1.38	-0.03	-0.82	*AKT1S1*	1.47	-3.52	-3.64
41	*LOX*	1.36	-3.97	-4.07	*CYP4B1*	1.47	-3.74	-2.78
42	*NBN*	1.34	-2.68	-2.62	*DUSP1*	1.47	1.31	-0.45
43	*CD3D*	1.33	-6.26	-6.41	*FN1*	1.39	0.66	0.66
44	*ESR1*	1.31	-1.69	-2.09	*TLR4*	1.38	-4.06	-4.26
45	*PTGER3*	1.29	-4.84	-5.30	*UCHL1*	1.33	-5.21	-3.80
46	*TMEM45A*	1.29	-3.86	-3.94	*NOTCH3*	1.33	-2.60	-2.02
47	*FOXP3*	1.29	-6.83	-7.07	*CXCL14*	1.33	-2.68	-3.29
48	*RARRES1*	1.27	-3.17	-3.26	*TSHR*	1.33	-8.53	-7.99
49	*TSHR*	1.23	-8.54	-8.01	*DCN*	1.31	-0.04	-0.94
50	*ADAM12*	1.22	-5.58	-5.79	*CRISPLD2*	1.27	-3.41	-3.80
51	*INHBA*	1.21	-4.29	-4.81	*CX3CR1*	1.26	-4.55	-5.09
52	*TESK1*	1.18	-4.61	-4.44	*CD2*	1.26	-5.98	-5.75
53	*SCGB1D2*	1.18	-5.27	-5.14	*WT1*	1.25	-1.35	-1.08
54	*FAP*	1.17	-4.18	-4.67	*GJB1*	1.24	-3.65	-4.51
55	*CRISPLD2*	1.15	-3.45	-3.99	*NBN*	1.23	-2.71	-2.80
56	*ZNF423*	1.13	-4.58	-5.26	*IDO1*	1.23	-4.15	-3.71
57	*GJB1*	1.12	-3.63	-4.54	*CTNNBL1*	1.23	-3.56	-3.43
58	*FCER1G*	1.09	-3.38	-3.20	*SNRPA1*	1.20	-2.73	-2.45
59	*IL22*	1.07	-8.97	-9.56	*ESR2*	1.18	-4.27	-5.55
60	*PDGFRB*	1.07	-2.76	-3.40	*AXL*	1.13	-3.17	-3.65

Rank	*All studies*				*Reduced studies*			
	Genes	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**	**Genes**	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**
61	*CD68*	1.07	-3.18	-3.11	-	-	-	-
62	*AR*	1.07	-3.50	-4.10	-	-	-	-
63	*GUSB*	1.07	-2.34	-2.01	-	-	-	-
64	*HMGA2*	1.06	-5.79	-5.26	-	-	-	-
65	*LUM*	1.05	-0.54	-1.16	-	-	-	-
66	*TAGLN*	1.04	0.09	-0.34	-	-	-	-
67	*FOLR1*	1.04	-2.37	-1.45	-	-	-	-
68	*CTNNBL1*	1.02	-3.56	-3.38	-	-	-	-
69	*WT1*	1.02	-1.33	-1.13	-	-	-	-
70	*MYOD1*	1.02	-7.89	-7.40	-	-	-	-
71	*THBS2*	1.01	-1.74	-2.35	-	-	-	-
72	*CAV1*	1.01	-3.97	-4.76	-	-	-	-
73	*COL1A2*	0.99	2.21	1.51	-	-	-	-
74	*COL5A2*	0.99	-1.98	-2.30	-	-	-	-
75	*NUAK2*	0.98	-2.72	-2.80	-	-	-	-
76	*PCDH9*	0.97	-6.10	-6.67	-	-	-	-
77	*SMARCA4*	0.95	-2.25	-2.24	-	-	-	-
78	*APBB2*	0.94	-2.74	-3.22	-	-	-	-
79	*POSTN*	0.94	-4.00	-4.35	-	-	-	-
80	*EPB41L3*	0.92	-3.45	-3.61	-	-	-	-
81	*NUAK1*	0.90	-4.06	-4.57	-	-	-	-
82	*SRI*	0.89	-1.07	-1.04	-	-	-	-
83	*CXCL11*	0.88	-6.27	-4.79	-	-	-	-
84	*IGJ*	0.88	-4.54	-3.44	-	-	-	-
85	*RB1*	0.87	-2.91	-3.43	-	-	-	-
86	*SALL2*	0.87	-3.80	-3.92	-	-	-	-
87	*MAL*	0.86	-4.80	-3.95	-	-	-	-
88	*YWHAB*	0.86	-0.80	-0.62	-	-	-	-
89	*ESR2*	0.83	-4.38	-5.52	-	-	-	-
90	*ANXA4*	0.78	-1.93	-2.43	-	-	-	-
91	*ZNF165*	0.78	-4.57	-4.46	-	-	-	-

Rank	*All studies*				*Reduced studies*			
	Genes	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**	**Genes**	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**
92	*MAP4K3*	0.76	-2.59	-2.62	-	-	-	-
93	*B4GALT5*	0.73	-1.97	-1.81	-	-	-	-
94	*KRT6*	0.70	-4.78	-5.11	-	-	-	-
95	*NF1*	0.70	-1.87	-2.22	-	-	-	-
96	*ADAMDEC1*	0.70	-7.01	-5.77	-	-	-	-
97	*IGFBP4*	0.68	0.70	-0.54	-	-	-	-
98	*CCL5*	0.66	-3.90	-3.64	-	-	-	-
99	*UCP2*	0.65	-2.72	-2.34	-	-	-	-
100	*SEMA6A*	0.64	-3.66	-4.33	-	-	-	-
101	*FAM58A*	0.64	-4.00	-3.82	-	-	-	-
102	*AKT1S1*	0.64	-3.57	-3.70	-	-	-	-
103	*WWP1*	0.64	-2.47	-2.65	-	-	-	-
104	*FN1*	0.63	0.72	0.53	-	-	-	-
105	*ESD*	0.62	-2.04	-2.64	-	-	-	-
106	*FGF1*	0.60	-5.78	-6.32	-	-	-	-
107	*PDZK1IP1*	0.58	-3.35	-3.80	-	-	-	-
108	*ENPP1*	0.57	-4.31	-5.12	-	-	-	-
109	*CXCL14*	0.57	-2.69	-3.37	-	-	-	-
110	*TSPAN8*	0.56	-5.08	-5.77	-	-	-	-
111	*FABP4*	0.53	-4.74	-5.75	-	-	-	-
112	*VCAN*	0.51	-3.68	-4.59	-	-	-	-
113	*PI3*	0.48	-5.21	-4.43	-	-	-	-
114	*RHOBTB3*	0.44	-3.19	-3.94	-	-	-	-
115	*SNRPA1*	0.43	-2.74	-2.48	-	-	-	-
116	*COL5A1*	0.42	-0.65	-1.40	-	-	-	-
117	*IGKC*	0.41	-1.54	0.52	-	-	-	-
118	*C19orf12*	0.39	-4.57	-4.64	-	-	-	-
119	*APC*	0.37	-4.12	-4.46	-	-	-	-
120	*CD302*	0.37	-3.84	-4.43	-	-	-	-
121	*GMNN*	0.37	-5.28	-4.28	-	-	-	-
122	*PCK2*	0.35	-4.11	-3.84	-	-	-	-

Rank	*All studies*				*Reduced studies*			
	Genes	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**	**Genes**	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**
123	*E2F6*	0.35	-3.26	-3.30	-	-	-	-
124	*RAD51B*	0.32	-3.82	-4.10	-	-	-	-
125	*SOX17*	0.29	-1.63	-1.56	-	-	-	-
126	*DCN*	0.29	-0.08	-1.01	-	-	-	-
127	*SHPRH*	0.25	-3.07	-3.43	-	-	-	-
128	*CTLA4*	0.24	-5.80	-6.00	-	-	-	-
129	*SERPINE1*	0.21	-1.34	-2.24	-	-	-	-
130	*WDR91*	0.20	-3.45	-3.25	-	-	-	-
131	*TP53*	0.17	-1.97	-2.59	-	-	-	-
132	*HIST1H2AM*	0.13	-1.50	-0.22	-	-	-	-
133	*CXCL10*	0.12	-5.72	-4.04	-	-	-	-
134	*FGFR1*	0.12	-1.53	-1.97	-	-	-	-
135	*APPL2*	0.09	-3.60	-3.86	-	-	-	-
136	*OASL*	0.05	-6.48	-5.44	-	-	-	-
137	*UQCC*	0.02	-3.55	-3.34	-	-	-	-
138	*OR1G1*	0.00	-8.80	-9.03	-	-	-	-
139	*GFPT2*	0.00	-5.39	-6.24	-	-	-	-
140	*CXCL9*	-0.03	-5.26	-4.39	-	-	-	-
141	*SPTLC2*	-0.05	-3.01	-3.09	-	-	-	-
142	*PIK3CA*	-0.06	-3.23	-2.93	-	-	-	-
143	*CD2*	-0.07	-6.03	-6.00	-	-	-	-
144	*OLFML3*	-0.11	-2.69	-3.02	-	-	-	-
145	*IGFBP2*	-0.12	-1.01	-0.36	-	-	-	-
146	*CPNE1*	-0.14	-1.56	-1.47	-	-	-	-
147	*CYP4B1*	-0.15	-3.62	-2.84	-	-	-	-
148	*PAX2*	-0.15	-5.34	-7.26	-	-	-	-
149	*PGRA*	-0.16	-3.78	-4.94	-	-	-	-
150	*TAP1*	-0.18	-4.31	-3.67	-	-	-	-
151	*GTF2H5*	-0.23	-3.03	-3.20	-	-	-	-
152	*CRABP2*	-0.23	-2.00	-0.45	-	-	-	-
153	*AXL*	-0.23	-3.20	-3.70	-	-	-	-

Rank	*All studies*				*Reduced studies*			
	Genes	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**	**Genes**	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**
154	*OPA1*	-0.25	-2.71	-2.35	-	-	-	-
155	*CX3CR1*	-0.26	-4.57	-5.23	-	-	-	-
156	*FOXRED2*	-0.27	-4.47	-4.10	-	-	-	-
157	*ZFHX4*	-0.32	-4.49	-5.48	-	-	-	-
158	*MYC*	-0.34	-0.64	-1.16	-	-	-	-
159	*IGF1*	-0.36	-3.19	-3.20	-	-	-	-
160	*SUPT6H*	-0.38	-2.60	-2.70	-	-	-	-
161	*DUSP4*	-0.40	-2.88	-5.17	-	-	-	-
162	*LRRC15*	-0.40	-4.77	-5.55	-	-	-	-
163	*GALNT6*	-0.45	-3.77	-3.54	-	-	-	-
164	*ASRGL1*	-0.46	-1.86	-2.28	-	-	-	-
165	*DAB2*	-0.49	-2.04	-2.59	-	-	-	-
166	*MRE11A*	-0.50	-3.57	-3.52	-	-	-	-
167	*COL11A1*	-0.54	-3.09	-3.12	-	-	-	-
168	*TRIM27*	-0.54	-2.93	-2.96	-	-	-	-
169	*CD3e*	-0.58	-6.08	-6.10	-	-	-	-
170	*MYCN*	-0.58	-5.05	-4.85	-	-	-	-
171	*ABCC2*	-0.61	-6.31	-7.02	-	-	-	-
172	*FEN1*	-0.71	-5.68	-5.12	-	-	-	-
173	*MDM4*	-0.71	-2.14	-2.43	-	-	-	-
174	*NUCB2*	-0.71	-2.90	-2.68	-	-	-	-
175	*CDK6*	-0.80	-3.49	-3.95	-	-	-	-
176	*CDKN3*	-0.82	-6.07	-4.12	-	-	-	-
177	*GMPR*	-0.83	-2.84	-2.35	-	-	-	-
178	*TUBB6*	-0.83	-2.31	-2.58	-	-	-	-
179	*COL3A1*	-0.84	2.47	1.92	-	-	-	-
180	*HIF1A*	-0.89	-0.91	-1.26	-	-	-	-
181	*HBB*	-0.91	-3.80	-4.77	-	-	-	-
182	*TLR4*	-0.92	-4.09	-4.37	-	-	-	-
183	*KDM5D*	-0.93	-8.65	-8.49	-	-	-	-
184	*RAD50*	-0.95	-3.34	-3.96	-	-	-	-

Rank	*All studies*				*Reduced studies*			
	Genes	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**	**Genes**	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**
185	*DUSP1*	-0.98	1.17	-0.47	-	-	-	-
186	*IGF2*	-1.02	-2.42	-2.55	-	-	-	-
187	*COBL*	-1.02	-3.73	-4.60	-	-	-	-
188	*FBN1*	-1.03	-2.85	-3.48	-	-	-	-
189	*BNIP3L*	-1.10	-1.42	-2.02	-	-	-	-
190	*MAP2K4*	-1.11	-3.20	-3.73	-	-	-	-
191	*ABCE1*	-1.15	-2.45	-2.67	-	-	-	-
192	*HIST1H2BD*	-1.22	-1.19	-0.47	-	-	-	-
193	*BRCA2*	-1.26	-5.99	-5.30	-	-	-	-
194	*SVIL*	-1.33	-1.84	-2.35	-	-	-	-
195	*CTHRC1*	-1.37	-1.48	-1.73	-	-	-	-
196	*KIFC1*	-1.43	-4.53	-2.68	-	-	-	-
197	*SAC3D1*	-1.45	-4.31	-3.79	-	-	-	-
198	*BAALC*	-1.45	-5.26	-6.23	-	-	-	-
199	*CSF1R*	-1.48	-5.39	-5.72	-	-	-	-
200	*IDO1*	-1.49	-4.17	-3.76	-	-	-	-
201	*BRCA1*	-1.56	-4.90	-4.86	-	-	-	-
202	*MRPS27*	-1.57	-2.47	-3.24	-	-	-	-
203	*CCNE1*	-1.57	-5.14	-3.07	-	-	-	-
204	*MEST*	-1.59	-2.58	-2.18	-	-	-	-
205	*TBX2*	-1.60	-2.80	-3.63	-	-	-	-
206	*MINPP1*	-1.67	-3.72	-4.12	-	-	-	-
207	*KIF1A*	-1.69	-6.09	-3.89	-	-	-	-
208	*PLK2*	-1.71	-0.78	-2.65	-	-	-	-
209	*SCGB2A1*	-1.76	-4.04	-3.83	-	-	-	-
210	*LPAR3*	-1.85	-3.86	-2.34	-	-	-	-
211	*ABCC3*	-1.87	-1.87	-2.55	-	-	-	-
212	*PTEN*	-2.13	-1.67	-2.10	-	-	-	-
213	*PARP4*	-2.25	-1.02	-1.46	-	-	-	-
214	*RASA1*	-2.29	-2.31	-2.83	-	-	-	-
215	*MCM3*	-2.34	-2.81	-2.12	-	-	-	-

Rank	*All studies*				*Reduced studies*			
	Genes	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**	**Genes**	**MDA**	**Avg exp grade 1**	**Avg exp grade 3/4**
216	*DNAJC9*	-2.39	-3.07	-2.62	-	-	-	-
217	*UCHL1*	-2.55	-5.20	-3.79	-	-	-	-
218	*CDKN2A*	-3.40	-4.80	-2.82	-	-	-	-