Please cite this article in press as: Zhang et al., Integrating Genomic, Epigenomic, and Transcriptomic Features Reveals Modular Signatures Underlying Poor Prognosis in Ovarian Cancer, Cell Reports (2013), http://dx.doi.org/10.1016/j.celrep.2013.07.010

Cell Reports
Article

Cell PRESS
Open ACCESS

Integrating Genomic, Epigenomic, and Transcriptomic Features Reveals Modular Signatures Underlying Poor Prognosis in Ovarian Cancer

Wei Zhang,[1,4] Yi Liu,[1,3,4] Na Sun,[1] Dan Wang,[1,2] Jerome Boyd-Kirkup,[1] Xiaoyang Dou,[1] and Jing-Dong Jackie Han[1,*]
[1]Key Laboratory of Computational Biology, Chinese Academy of Sciences-Max Planck Partner Institute for Computational Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yue Yang Road, Shanghai 200031, China
[2]Graduate University of Chinese Academy of Sciences, Beijing 100049, China
[3]Beijing Key Laboratory of Traffic Data Analysis and Mining, School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044, China
[4]These authors contributed equally to this work
*Correspondence: jdhan@picb.ac.cn
http://dx.doi.org/10.1016/j.celrep.2013.07.010
This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial-No Derivative Works License, which permits non-commercial use, distribution, and reproduction in any medium, provided the original author and source are credited.

SUMMARY

Ovarian cancer has a poor prognosis, with different outcomes for different patients. The mechanism underlying this poor prognosis and heterogeneity is not well understood. We have developed an unbiased, adaptive clustering approach to integratively analyze ovarian cancer genome-wide gene expression, DNA methylation, microRNA expression, and copy number alteration profiles. We uncovered seven previously uncategorized subtypes of ovarian cancer that differ significantly in median survival time. We then developed an algorithm to uncover molecular signatures that distinguish cancer subtypes. Surprisingly, although the good-prognosis subtypes seem to have not been functionally selected, the poor-prognosis ones clearly have been. One subtype has an epithelial–mesenchymal transition signature and a cancer hallmark network, whereas the other two subtypes are enriched for a network centered on SRC and KRAS. Our results suggest molecular signatures that are highly predictive of clinical outcomes and spotlight "driver" genes that could be targeted by subtype-specific treatments.

INTRODUCTION

Ovarian cancer is the second most common and the most lethal gynecologic cancer, representing about 3% of all cancers diagnosed in females, with a median incidence at 63 years of age (Jemal et al., 2008; Kosary, 2007). Unlike many other cancers, the prognosis of ovarian cancer is poor and has not been significantly improved over recent decades, with a 5-year survival rate of around 47% (Johannes, 2010).

The underlying mechanisms for the poor prognosis of ovarian cancer are largely unknown. Despite treatment with similar surgery and adjuvant therapies, outcomes may be very different among different patients. A recent genome-wide association study (GWAS) identified a common variant at 19p13 associated with the survival time of ovarian cancer (Bolton et al., 2010). However, to date, many other GWAS studies have not been convincingly replicated.

The Cancer Genome Atlas (TCGA) has collected detailed clinical records and heterogeneous high-throughput data for more than 500 cases of ovarian serous cystadenocarcinoma, including gene expression, somatic mutation, promoter DNA methylation, microRNA (miRNA) expression, and copy number alteration (CNA) (Cancer Genome Atlas Research Network, 2011). From these data, the TCGA consortium found ten recurrent somatic mutations, including *TP53*, *BRCA1/2*, and *RB1*, in high-grade ovarian cancers and identified four transcriptional subtypes that were unrelated to prognosis (Cancer Genome Atlas Research Network, 2011). They showed that the mutual exclusivity of the *BRCA1/2* mutation and *CCNE1* amplification is related to ovarian cancer prognosis (Ciriello et al., 2012). Up to now, however, the question of whether ovarian cancer subtypes with different prognoses and distinct hallmark hazard factors exist has remained unclear. The TCGA data set gives us a unique opportunity to investigate whether combining all of these heterogeneous high-throughput data will allow us to (1) uncover hallmark hazard factors to distinguish subtypes with different prognoses, and (2) identify subtype-specific pathways that might explain such different prognoses.

Here, we developed an adaptive clustering algorithm based on the Bayesian Information Criterion (BIC) to automatically determine the optimal number of sample and feature clusters, and coupled this with deep clustering using our recently developed unsupervised "super k-means" algorithm on a combination of gene expression, DNA methylation, miRNA expression, and CNA data for ovarian cancer. With this approach, we were able

Please cite this article in press as: Zhang et al., Integrating Genomic, Epigenomic, and Transcriptomic Features Reveals Modular Signatures Underlying Poor Prognosis in Ovarian Cancer, Cell Reports (2013), http://dx.doi.org/10.1016/j.celrep.2013.07.010

Cell PRESS
Open ACCESS

Table 1. Numbers of Samples and Features of the Types of High-Throughput Data Used in this Study

Data Type	Platform	Number of Samples	Number of Features
mRNA expression	Agilent 244K Custom Gene Expression G4502A-07-1	513	17,436
DNA methylation	Illumina HumanMethylation27	513	12,854
miRNA expression	Agilent 8 × 15K Human miRNA-Specific Microarray	510	799
CNA of genes/miRNA	Agilent 1M	512	20,412/628

to de novo detect seven distinct subtypes with significantly different clinical outcomes. In addition, we developed an algorithm to systematically uncover molecular signatures that best distinguish the seven subtypes of ovarian cancer from the clustering result. The in-depth analysis of these molecular signatures not only broadens our current understanding of ovarian cancer but also sheds light on ways to achieve better diagnosis and treatment of this disease in the future.

RESULTS

Selection of Ovarian Cancer Hazard Factors

We first investigated whether we could identify hazard factors for ovarian serous cystadenocarcinoma. From the TCGA, we collected and preprocessed the clinical records and four types of high-throughput data of 513 patients together with their survival time from initial diagnoses to death, or to the last follow-up if they were still alive at the time of the TCGA study (Table 1; Experimental Procedures; Cancer Genome Atlas Research Network, 2011). All four types of data (messenger RNA [mRNA] and miRNA expression, promoter DNA methylation, and CNA) were available for 509 of the 513 samples. One sample lacked copy number data, and three other samples lacked miRNA expression data (Table 1). We used the 512 samples that had both gene expression and copy number data for analysis.

The Cox proportional hazard model (Cox regression model) is widely used with censored data to estimate the effect of different features on survival time (Andersen and Gill, 1982). To investigate which features are related to the prognosis of ovarian cancer, we first used the univariate Cox proportional hazard model to perform a regression analysis between each feature and the patients' survival time. In total, we selected 4,526 features as hazard factors ($p < 0.05$; Experimental Procedures), including 1,651 mRNA expression changes, 455 promoter DNA methylation changes, 140 miRNA expression changes, and CNAs of 2,191 genes and 89 miRNAs.

De Novo Characterization of Ovarian Cancer Subtypes by Molecular Signatures

To unbiasedly identify ovarian cancer subtypes with different types of features represented with equal probability, we performed adaptive clustering on the 512 samples to uncover subtypes of samples that showed different survival times using all of the above-described 4,526 features (hazard factors) simultaneously (Experimental Procedures). To reduce systematic variations among heterogeneous high-throughput data from different platforms, we normalized the 4,526 features against normal tissue controls (Experimental Procedures). Because there is no a priori knowledge about the number of ovarian cancer subtypes, we first applied the BIC (Schwarz, 1978) to these normalized data, which automatically determined the optimal number of clusters across the 512 samples to be seven, and across the 4,526 features to be 37 (Experimental Procedures; Figures S1A and S1B). Because BIC is mainly aimed at determining the right number of clusters and does not generate the most compact clustering result, we subsequently used the unsupervised super k-means clustering algorithm (Liu et al., 2013), which generates extremely compact clusters, to cleanly divide the samples into seven clusters, which we refer to hereafter as "subtypes," and similarly to divide the features into 37 clusters (Experimental Procedures; Figure 1A).

Since the data for the 512 samples were generated in 13 batches (Cancer Genome Atlas Research Network, 2011), we examined whether our subtype clustering result could have been due to batch effects. Fisher's exact tests showed that none of the seven subtypes were significantly enriched in any of the 13 batches (Experimental Procedures). They also confirmed that neither the selection of hazard factors nor the clustering approach was biased toward particular batches.

In order to examine the stability of the subtype clustering result, we performed 10-fold cross validation by randomly dividing the 512 samples into ten groups and hiding 10% of the samples each time to perform the super k-means clustering algorithm. The subtype classification was observed to be very stable, with the average sensitivity and precision both exceeding 87% (Figure 1D, empirical p value < 1e-04; Experimental Procedures). Remarkably, the accuracy stayed at the same level when in each round the remaining independent 10% of samples were assigned to the nearest cluster centers of the 90% (Figure 1E). This demonstrates the existence of robust molecular signatures that distinguish the de novo categorized ovarian cancer subtypes.

We then designed an algorithm to systematically uncover the molecular signatures for each of the seven ovarian cancer subtypes from the clustering results (Extended Experimental Procedures) and in this way identified 18 of the most significant subtype-association signature feature clusters (Extended Results). We found that using just two features from each of the selected 18 clusters (36 features in total) was sufficient to achieve high subtype-identification accuracy (81%), indicating that the 36 signature features are a good candidate marker set for ovarian cancer classification (Extended Results; Figures S1C and S1D; Tables S1 and S2).

Our algorithm identified distinct features for six of the seven subtypes. These include a deletion of chr6 (p24.1-p12.1) in subtype 1, upregulation of three gene expression clusters in subtype 2, amplification of multiple chromosomes in subtypes 4–6, and deletion of chr19 (q13.2-q13.43) in subtypes 6 and 7 (Figures 1A and 1B).

Remarkably, patients with the seven subtypes of ovarian cancer differ significantly in their survival time. The median survival times of the subtypes with the shortest and longest prognoses

are 2.36 and 4.72 years, respectively, representing a difference of 2.0-fold (log-rank test, p = 3.16e-04; Figures 1B and 1C). The 164 patients (32.0%) in subtypes 2, 4, and 5 have a poor prognosis, with a median survival time of <3 years, whereas the 213 patients (41.6%) in subtypes 3 and 7 have a better (good) prognosis, with a median survival time of >4.5 years (Figures 1B and 1C). Therefore, in this way, the majority of patients (73.6%) can be successfully assigned into either high- or low-risk groups using our adaptive clustering approach.

In addition to the above subtypes, patients in subtypes 1 and 6 have a median survival time in between those of the high- and low-risk groups (Figures 1B and 1C). Interestingly, subtype 6 has both the signature of a multiple chromosome amplification similar to that of subtypes 4 and 5, which have a poor prognosis, and the signature of a chr19 (q13.2-q13.43) deletion similar to that of subtype 7, which has a good prognosis. Because subtype 6 shows a moderate survival time compared with the high- and low-risk groups (Figures 1A–1C), these signatures may oppose each other in survival. However, it remains formally possible that the two signatures of this subtype might reflect a combination of tumor cells from two different origins with different alterations in their genomes.

Finally, it is worth noting that these poor-prognosis subtypes could not be correctly identified by transcriptomic data alone, even when prognosis-related genes were used for clustering (Figures S1E and S1F).

Functional Analysis of Signature Feature Clusters

When we examined the Gene Ontology (GO) terms (Ashburner et al., 2000) and Kyoto Encyclopedia of Genes and Genomes (KEGG) (Kanehisa et al., 2010) pathways enriched among the subtype signatures, we found that genes that are specifically upregulated in subtype 2 are enriched for many functions related to tumorigenesis, such as cell adhesion, angiogenesis, transforming growth factor β (TGF-β) binding, and positive regulation of cell proliferation (Table 2). Among them, the most significantly enriched was the "cell adhesion" function (Table 2), which is related to metastasis. This may well explain the poor prognosis of subtype 2 (Figures 1B and 1C). Clinical records revealed that the 70 patients in subtype 2 showed higher pathological tumor stages (stage IIIC or IV) compared with other patients (Wilcoxon rank sum test, p = 8.995e-3). The higher pathological stages, stages III and IV, indicate metastasis, with IV showing more extensive metastasis than III. Interestingly, when we compared the 70 patients in subtype 2 with the 81 patients of tumor stage IV (the highest stage), we found that their median survival times were similar, whereas the 5-year survival rate of subtype 2 was even worse than that of tumor stage IV (Figure 2). This indicates that the subtype 2 features we found at the molecular level are at least a comparable, or possibly even better, indication of poor prognosis than the clinical observation of distant metastasis.

The epithelial–mesenchymal transition (EMT) and an embryonic stem cell (ESC)-like gene expression signature have been associated with malignancy and metastasis in human tumors (Ben-Porath et al., 2008; Polyak and Weinberg, 2009). Through a literature cocitation analysis, we found that the frequency of literature cocitation of the term "epithelial mesenchymal transition" or "embryonic stem cells" with the genes specifically upregulated in subtype 2 was significantly higher than random expectation (empirical p value < 0.001 for both terms; Experimental Procedures). In fact, the genes that are specifically upregulated in subtype 2 are enriched in the key regulators and signaling pathways of the EMT, such as SNAI1, TWIST1, and the TGF-β pathway, as summarized by Polyak and Weinberg (2009) (Figure 3). Moreover, the genes that are specifically upregulated in subtype 2 are also enriched in NOS (NANOG, OCT4, and SOX2) targets that were previously identified in a chromatin immunoprecipitation (ChIP)-chip experiment in human ESCs (hESCs) (Boyer et al., 2005; Figure 3), even though the expression of *OCT4*, *NANOG*, *SOX2*, *MYC*, and *KLF4* themselves is not increased in subtype 2. This indicates that the tumor cells in this subtype contain at least a partial hESC-like gene expression signature. Taken together, these findings support the hypothesis that an EMT and hESC-like gene expression signature may have key roles in cancer metastasis and poor prognosis (Polyak and Weinberg, 2009).

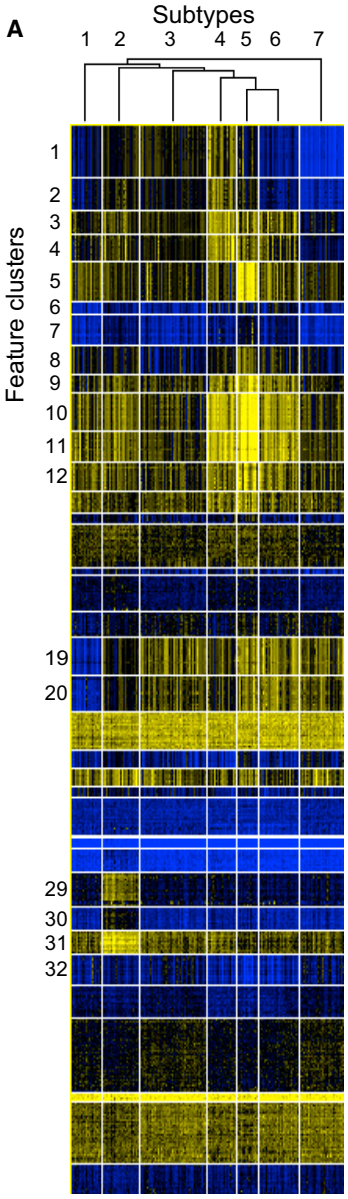
Intriguingly, we did not find more somatic mutations among the subtype 2 samples on the subtype 2 signature genes. This prompted us to investigate whether any nonmutational or epigenetic mechanisms are involved in maintaining the upregulation of subtype 2 signature genes. Given the requirements for histone demethylase KDM5A and the insulin growth factor 1 (IGF-1) signaling pathway for drug resistance in cancer stem cells (CSCs) (Sharma et al., 2010), we also examined the expression, CNA, and promoter DNA methylation status of these genes across the subtypes. The copy number and expression of *KDM5A*, which are highly correlated across the 512 samples (Pearson correlation coefficient [PCC] = 0.73), are relatively low in good-prognosis subtypes (subtypes 3 and 7) and high in other subtypes. *IGF1* is specifically upregulated in subtype 2 (the metastasis subtype) and downregulated in other subtypes (Wilcoxon rank sum test, p = 2.997e-16). Similarly, *IGFBP3* is also upregulated in subtype 2 compared with the other subtypes (Wilcoxon rank sum test, p = 3.717e-7). However, the classical CSC markers *CD133* and *CD44* are generally repressed in all subtypes compared with normal tissue controls (Wilcoxon rank sum test, p = 4.212e-3 for *CD133* and p = 6.619e-6 for *CD44*; Figure 3). Therefore, although they lack classical CSC markers, in general, the ovarian cancer samples with a poor prognosis possess a recently discovered drug-resistant CSC feature: high expression of *KDM5A* and *IGF1*. Indeed, the poor-prognosis subtypes (subtypes 2, 4, and 5) have the worst outcomes in response to drug treatment and are more likely to be resistant to cancer drug treatments than other subtypes (Figure S2; Tables S3, S4, and S5).

A Functional Interaction Network for the Signature Genes

We next investigated the relationship among the signature genes of each subtype using a functional interaction network. To enrich for the "driver" genes (i.e., those genes in which changes are directly reflected at the gene expression level), we first removed the genes that were identified by genomic or epigenomic signatures (CNA or DNA methylations) that are not consistent with

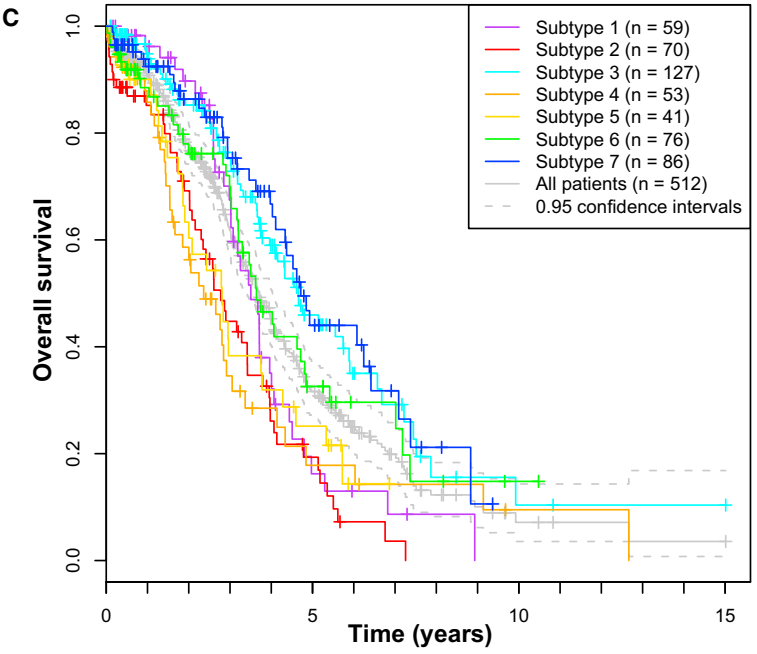
Please cite this article in press as: Zhang et al., Integrating Genomic, Epigenomic, and Transcriptomic Features Reveals Modular Signatures Underlying Poor Prognosis in Ovarian Cancer, Cell Reports (2013), http://dx.doi.org/10.1016/j.celrep.2013.07.010

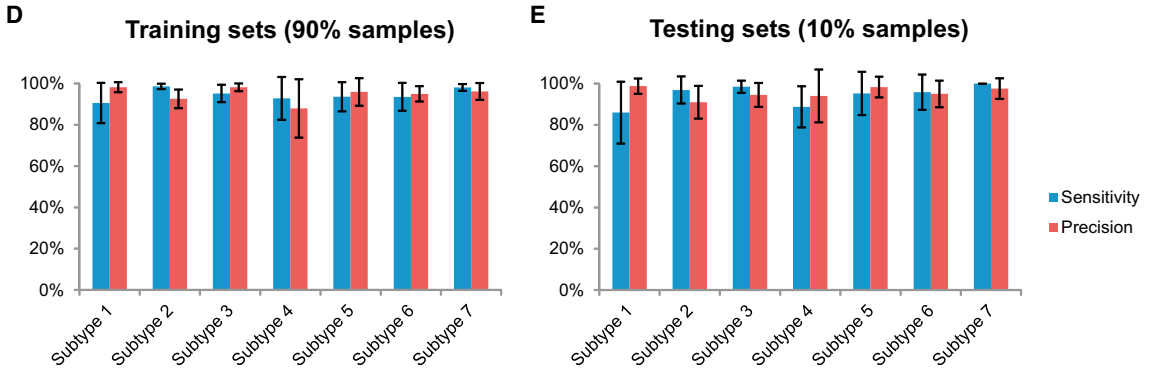
Cell PRESS
Open ACCESS



B

			Subtypes						
			1	2	3	4	5	6	7
Number			59	70	127	53	41	76	86
Median survival time (Years)			3.50	2.78	4.65	2.36	2.81	3.64	4.72
Feature Clusters	1, 2	CNA: chr19 (q13.2-q13.43)						Del	Del
	3, 4	CNA: chr19 (q12- q13.2)				Amp			
	5	CNA: chr2 (p16.2-q11.2)					Amp		
	6, 7	CNA: chr6 (q24.2-q27) Chr17 (q11.2-q21.31)	Del		Del				Del
	8	CNA: chr12 (q21.1-q23.3)					Amp		
	9 - 12	CNA: chr20 (p13-q13.2) CNA: chr12 (p13.33-p11.22)				Amp	Amp	Amp	
	19, 20	CNA: chr6 (p24.1- p12.1)	Del						
	29 - 31	Expression: Cell adhesion		Up-regulation					
	32	Promoter DNA methylation				Hypo-methylation			





(legend on next page)

Table 2. Enriched GO Terms of the Subtype-2-Specific Upregulated Genes

Term	p Value	Fold Enriched	Symbols
GO:0007155 | cell adhesion	3.55E-08	3.67	*SPON2, COL16A1, SPOCK1, COL6A6, SVEP1, SLAMF7, THBS2, THBS1, CTGF, WISP1, CYR61, ADAM12, ITGA5, CD93, LAMB1, COL3A1, ECM2, CDH11, PDPN, CLDN11, AEBP1, CCL11, ITGB1, KAL1, COL8A2, COL8A1, ANTXR1, SIRPA, FN1, CD36, COL5A1*
GO:0019838 | growth factor binding	4.19E-07	9.19	*THBS1, CTGF, WISP1, TGFBR1, CYR61, IL18R1, COL3A1, HTRA1, KDR, TGFBR2, CD36, COL5A1*
GO:0040011 | locomotion	9.56E-07	4.33	*PLAUR, WWP1, ARID5B, CTGF, TGFBR1, CYR61, FAP, ETS1, ITGA5, CXCL14, SBDS, C5AR1, SCG2, TWIST1, AGTR1, CCL11, ITGB1, KAL1, FN1, COL5A1, RNASE2*
GO:0001525 | angiogenesis	7.33E-06	7.84	*THBS1, FGF1, CTGF, ANXA2, KDR, SCG2, BMP4, COL8A2, COL8A1, FN1, ELK3*
GO:0050431 | TGF-β binding	1.24E-04	32.48	*THBS1, TGFBR1, TGFBR2, CD36*
GO:0008284 | positive regulation of cell proliferation	1.89E-03	3.01	*FGF1, TGFBR1, PRRX1, LAMB1, IGF1, BNC1, FABP4, KDR, GAS1, SCG2, TGFBR2, BMP4, KRT6A, CCL11, ADRA2A*
GO:0005520 | IGF binding	2.14E-03	13.53	*CTGF, WISP1, CYR61, HTRA1*

gene expression changes (PCC < 0.3 for CNAs, PCC > −0.3 for promoter DNA methylations; Figure S3; Extended Experimental Procedures). We then also removed the genes with inconsistent direction of expression changes within the featured subtype (with dominant direction in <60% samples; Figure 1B). To distinguish the driver events from the "passengers," we further focused on the signature genes that were within cancer-hallmark-related pathways (annotated according to Hanahan and Weinberg, 2000, 2011), ESC pluripotency signaling pathways (http://www.cellsignal.com/reference/pathway/ESC_pluripotency.html), multi-drug-resistance factors (http://www.biocarta.com/pathfiles/h_mrppathway.asp), or EMT signaling pathways (Polyak and Weinberg, 2009; Experimental Procedures). Finally, we mapped the signature genes present in the hallmark pathways to an accurately curated human functional protein interaction network to study their functional relationships (Wu et al., 2010; Experimental Procedures).

The resulting network was observed to be highly modular, with the genes in similar pathways clustered together (Figure 4B). In particular, the signature genes that were upregulated in subtype 2 or amplified in subtypes 4–6 were largely enclosed by four network modules (Figure 4; see the yellow nodes/signature for cluster 2 and the blue nodes/signature for clusters 4–6). Interestingly, the signature genes for the subtypes with relatively good prognosis did not show up in the hallmark network or cluster into tightly connected modules, indicating that poor prognosis, but not good prognosis, is associated with functional selections at the network level.

In this network, we found that the genes that were specifically upregulated in subtype 2 predominantly participated in three modules corresponding to the functions "focal adhesion and extracellular matrix (ECM)-receptor interaction," "TGF-β signaling pathway," and "Wnt signaling pathway," which are all related to EMT and metastasis (Figure 4). Remarkably, the network module of "proliferation circuits (mTOR, ErbB, and mitogen-activated protein kinase [MAPK] signaling pathways)" is enriched for genes specifically amplified in subtypes 4–6, including several key regulators in these pathways, such as *KRAS*, *SRC*, *PLCG1*, and *AKT2* (Figure 4). Another key transcription factor, *E2F1*, which regulates the cell cycle, was also specifically amplified in these three subtypes. The network model suggests that in these subtypes, multiple signaling pathways related to cell proliferation and the cell cycle might be constitutively activated through genomic amplification of these key regulators, which in turn would accelerate the division and proliferation of tumor cells, thus contributing to the poor prognosis of subtypes 4 and 5.

Validation of Subtype 2 Classification by Independent Sample Sets

To determine whether our subtype-specific molecular signatures derived from the TCGA data are also applicable to a completely independent set of samples, we collected the expression

Figure 1. De Novo Categorization of Ovarian Cancer Subtypes

(A) Super k-means clustering of the 512 TCGA ovarian cancer samples (columns) using the 4,526 features that are significantly associated with prognosis (rows). The normalized value of each feature is indicated by color intensity, with yellow/blue representing higher/lower expression, copy number, and promoter DNA compared with the controls. The subtype-specific signature clusters are labeled at the left side of the heatmap.

(B) Systematically categorized ovarian cancer subtype classification feature clusters. These clusters are derived from (A). The number of patients, median survival time of each subtype, and the major data type for each feature cluster are also listed.

(C) Overall survival probabilities (y axis) are plotted against the survival times (years, x axis) of the seven ovarian cancer subtypes. Samples with censored survival data are indicated by a "+" at their censoring time (last follow-up). The survival curve of all 512 patients as well as the 0.95 confidence intervals are also shown for comparison. Overall survival is defined as in the Cancer Genome Atlas Research Network (2011), i.e., as the interval from the date of initial surgical resection to the date of death or last follow-up.

(D and E) The average sensitivity and precision for each subtype on the training (D) or testing set samples (E) are compared with the original subtypes in all 512 samples of (A) (used as benchmarks here) in ten trails of clustering with 10% of the samples left out in each trial. Error bars indicate the SDs of ten trails.

See also Figure S1 and Tables S1 and S2.

Cell PRESS
Open ACCESS

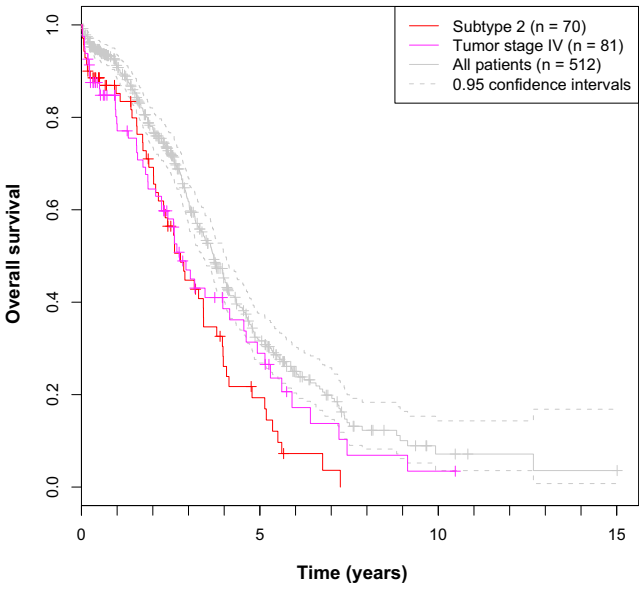


Figure 2. Comparison of Survival Curves between Subtype 2 and Stage IV Ovarian Cancer Patients

Overall survival probabilities (y axis) are plotted against the survival times (years, x axis) of subtype 2 or stage IV ovarian cancer patients. Samples with censored survival data are indicated by a "+" at their censoring time (last follow-up). The survival curve of all 512 patients with 0.95 confidence intervals is also shown for comparison.

profiles of 696 high-grade serous ovarian carcinoma samples used in a recent prognostic study (Verhaak et al., 2013) from five independent data sets (Bonome et al., 2008; Crijns et al., 2009; Denkert et al., 2009; Tothill et al., 2008; Yoshihara et al., 2010). Then, we used k-means clustering to group samples from each of the five independent data sets into two clusters based on the expression profiles of subtype 2 signature genes (note that only 170–214 out of the 222 signature genes are quantified in the five data sets). We found that most of our predicted subtype 2 samples were indeed classified as a "mesenchymal" (MES) subtype in the previous prognostic study (Verhaak et al., 2013), and only a few of them fall into the "immunoreactive" (IMR) subtype (Figure 5A). Almost no samples in subtype 2 fall into the other two subtypes classified by Verhaak et al. (2013). Moreover, patients in the predicted subtype 2 cluster also have a significantly shorter survival time (log-rank test, p = 1.11e-05) than the other patients, which independently confirmed that subtype 2 has a poor prognosis. Specifically, the median survival times for patients in the predicted subtype 2 group and the other group are 30 versus 47.4 months, respectively, representing a 1.58-fold difference (Figure 5B), which is even larger than the difference we found in the TCGA data set (1.36-fold, 33.36 versus 45.36 months, respectively; Figures 1B and 1C). Furthermore, we also examined the relationship among all available survival-related factors with the two-group partition of patients from the five independent data sets. Patients in the predicted subtype 2 cluster are associated with a high rate of platinum resistance and have more advanced tumor stages (Wilcoxon rank sum test, p = 1.993e-6 and 1.422e-3, respectively), and are also more likely to relapse (log-rank test, p = 4.2e-03; Figure 5E), whereas factors such as tumor grade, tumor residue disease, and patients' age at initial diagnosis do not differ greatly between the groups (Figures 5C and 5D). The association between subtype 2 and more advanced tumor stages is consistent with our observations from the TCGA data set, and once again suggests a possible link between subtype 2 signatures and metastasis.

Since subtype 2 is associated with advanced tumor stages (Figure 5C), we tested whether this is the main cause for the poor prognosis of patients in this subtype by comparing the prognoses of subtype 2 and non-subtype 2 patients from the five independent data sets after stratifying for tumor stage. Remarkably, within tumor stage III, which comprises the vast majority of patients (85.5%), subtype 2 patients still show significantly shorter survival times than the rest of the patients at stage III (log-rank test, p = 2.92e-04, Figure 5F). The other two stages (stages II and IV), however, do not show significant differences in prognosis between subtype 2 and other patients (Figure 5F), which might simply be due to the small number of patients in the two stages and/or the influence of other factors, such as surgical and drug interventions. Similarly, in the TCGA data set, we also observed differences in prognosis for subtype 2 versus other patients after stratifying for tumor stage factor. Within either tumor stage IIIC or IV of the TCGA samples, subtype 2 patients showed a significantly shorter survival time than the rest of the patients at the same stage (Figure S4A; log-rank test, p = 4.92e-03 and 2.36e-02, respectively).

However, samples of one stage are frequently classified into many different subtypes and have very different survival times, indicating that the clinical diagnosis of stage has a much lower resolution than our subtyping (Figure S4B).

To conclude, the significant difference in prognosis between subtype 2 patients and other patients in stage III indicates that our subtype 2's association with short survival time cannot be simply attributed to the factor of tumor stages. Given the fact that subtype-2-specific upregulated genes are enriched for metastasis-related functions (e.g., cell adhesion and EMT), these genes provide additional resolution for identifying the molecular features of this metastatic disease (Figures 5F and S4), and thus provide extra information for more effective targeted treatment in the future.

DISCUSSION

In this study, we have established an analysis framework to systematically, automatically, and unbiasedly uncover signature features and pinpoint the potential driver genes of each subtype. We first combined high-throughput genomic, epigenomic, and transcriptomic data and used the Cox proportional hazard model to select features associated with patient survival time in ovarian cancer. We then developed an adaptive clustering algorithm to automatically determine the optimal number of sample and feature clusters based on the BIC. Following this, we deep clustered samples/features using our recently developed super k-means algorithm. This approach led to the de novo categorization of seven distinct ovarian cancer subtypes with significantly different clinical outcomes.

The signatures identified include a metastasis, EMT, and partial hESC-like gene expression signature (Ben-Porath et al.,

A

Gene set	P-value	Fold enriched	Symbols
Epithelial–mesenchymal transition	4.20E-07	8.15	*TGFBR1, SNAI2, ITGA5, FZD1, EDNRA, TGFBR2, TWIST1, SNAI1, WNT4, ITGB1*
NOS targets	2.14E-02	2.79	*THBS2, KDR, CTGF, MST150, SFRP2, OLFML3*
Oct4 targets	2.45E-02	2.29	*THBS2, KDR, CTGF, MST150, SFRP2, GAP43, GAS1, OLFML3*

B

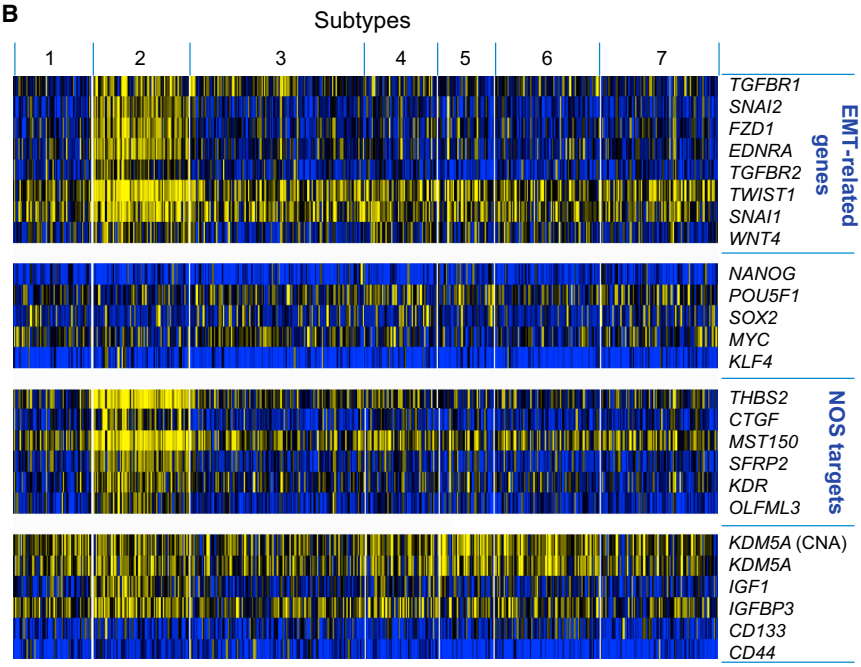


Figure 3. EMT and hESC-Related Genes, Drug-Resistance Signatures, and CSC Markers in Ovarian Cancer Subtypes

(A) Enriched EMT- and hESC-related genes specifically upregulated in subtype 2.

(B) Heatmap of EMT- and hESC-related genes, drug-resistance signatures, and CSC markers in ovarian cancer subtypes. The normalized value of each feature compared with the normal tissue controls is indicated by color intensity, with yellow/blue representing high/low expression or copy number. All data used in this figure are expression values except for the row marked "*KDM5A* (CNA)," which shows the copy number changes of *KDM5A*.

See also Figure S2 and Tables S3, S4, and S5.

2008; Polyak and Weinberg, 2009) for subtype 2, and a recently reported drug-resistance signature (Sharma et al., 2010) for poor-prognosis subtypes (subtypes 2, 4, and 5). In contrast to Ciriello et al.'s hypothesis about the mutual exclusivity in network modules (Ciriello et al., 2012), we found that the signature genes of a subtype tend to be concentrated in the same modules in our network, indicating their potential synergistic effects. In particular, the genes that are upregulated in the three poor-prognosis subtypes via transcriptomic and genomic alterations are highly interconnected in the network through pathways related to metastasis, EMT, and cell proliferation. Our automated analysis framework could easily be applied to similar studies of other types of cancer.

In the original TCGA study, using ~1,500 differentially expressed genes, researchers identified four ovarian cancer subtypes; however, these subtypes did not differ significantly in survival time (Cancer Genome Atlas Research Network, 2011). Nevertheless, the TCGA investigators did select a 193-gene expression signature associated with survival, and a recent analysis revealed a 100-gene signature associated with prognosis (Verhaak et al., 2013). Although the latter signature does overlap significantly with our combined seven subtype signatures, the biggest difference is that previous studies used supervised approaches to build a prognosis predictor to maximize the difference in survival time between poor and good groups, whereas our approach uses unsupervised learning to identify intrinsic molecular subtypes of ovarian cancers de novo. Using our approach, we observed a difference in prognosis between different subtypes based on their distinct molecular signatures, but we did not set out to maximize this difference at the beginning of the analysis. Therefore, our approach offers a more unbiased molecular signature and mechanism to develop personalized treatment compared with the approach of Verhaak et al. (2013), which masked the intrinsic distinction of molecular programs among subtypes with similar prognosis status. As a result, the patients in Verhaak et al.'s study who had similar prognoses were deprived of the chance of being differentially treated by targeting their distinct molecular hallmarks.

A more recent study also used a clustering approach to identify five ovarian cancer subtypes based on the gene expression profiles of 1,538 tumor samples, and further used a small hairpin RNA (shRNA) library to screen for molecular targets that are essential for cell growth in a subtype of poor prognosis (Tan et al., 2013). Similarly to the study by Verhaak et al. (2013), that study was based solely on expression profiles, whereas our approach combines genomic, epigenomic, and transcriptomic data, and identified five (out of seven) subtypes with clear CNA signatures. Interestingly, Tan et al. (2013) also identified a poor-prognosis subtype characterized by "mesenchymal" genes that overlap significantly with our subtype 2 signature genes (Fisher's exact test, $p < 2.2e-16$). This again implies the universal existence of this signature in most sample collections. However, the poor-prognosis subtypes 4 and 5, which show amplification of the proliferation-related genes centered on KRAS and SRC, were identified only by our approach.

From a methodological perspective, Tan et al. (2013) used consensus clustering (CC) (Monti et al., 2005) to detect five

Please cite this article in press as: Zhang et al., Integrating Genomic, Epigenomic, and Transcriptomic Features Reveals Modular Signatures Underlying Poor Prognosis in Ovarian Cancer, Cell Reports (2013), http://dx.doi.org/10.1016/j.celrep.2013.07.010

Cell PRESS
Open ACCESS

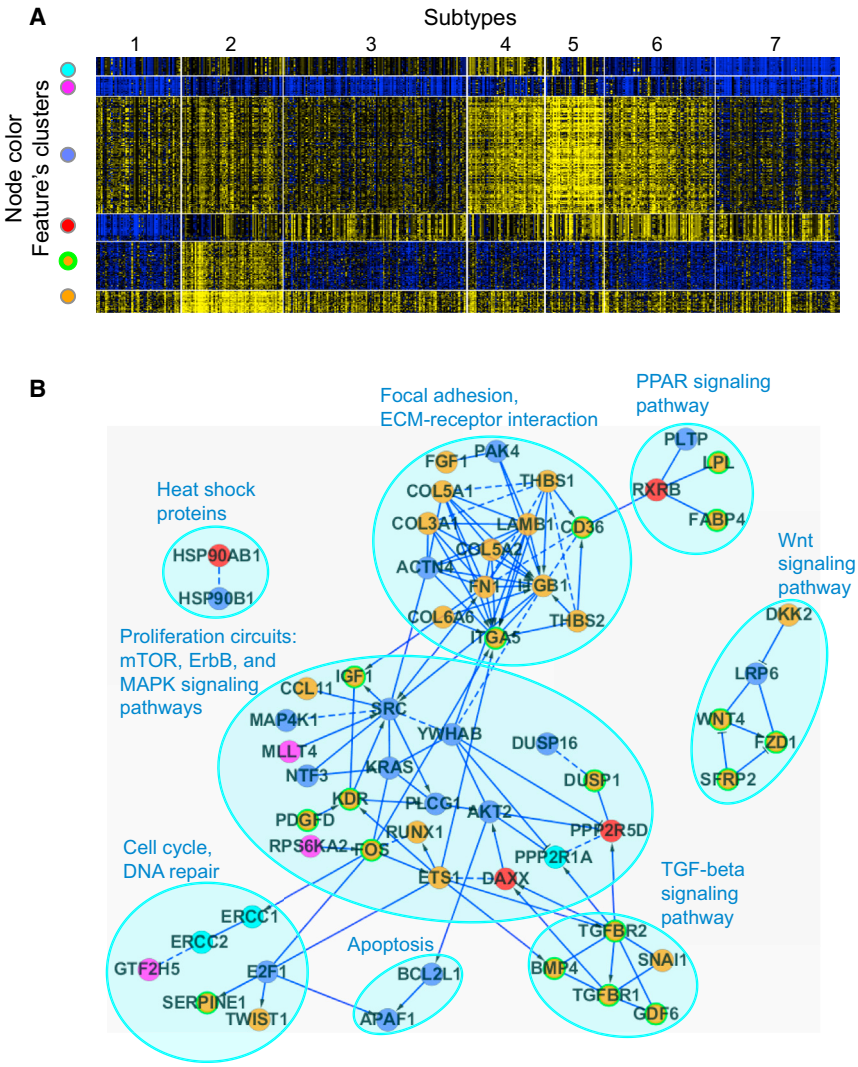


Figure 4. Functional Interaction Network of Ovarian Cancer Subtypes

(A) Heatmap of subtype-specific features in the seven subtypes of ovarian cancer samples. The normalized value of each feature is indicated by color intensity as described in Figures 1A and 3.

(B) The functional interaction network among the genes in the subtype-specific signatures was constructed based on the network of Wu et al. (2010). The cluster assignment of a gene is indicated by the node color on the left side of (A), with red representing chr6 (p24.1-p12.1) deletion in subtype 1; orange representing expression upregulation in subtype 2; orange with a green border representing expression upregulation in subtype 2 and downregulation in other subtypes; cyan representing chr19 (q13.2-q13.43) deletion mainly in subtype 7; blue representing chr20 (p13-q13.2), chr19 (q12-q13.2), or chr12 (p13.33-p11.22, q21.1-q23.3) amplification in subtypes 4, 5, and 6; and magenta representing chr6 (q24.2-q27) deletion in subtypes 1, 3, and 7. The types of interactions are indicated by the edge style, with solid lines representing manually curated pathways, and dashed lines representing predicted interactions. The network modules were manually dissected based on the network structure and gene functions, with their most concordant functions labeled above them.

See also Figure S3.

tumor subtypes (clusters) based on gene expression data. It is worth noting that CC is only able to identify a small number of clusters (usually fewer than ten) within data due to its limited discriminative power. As a result, it can be used to decide the number of tumor subtypes, but not the number of molecular signature clusters (which is often much larger). In contrast, our adaptive clustering approach can be used for either task, enabling a comprehensive characterization of the subtype-feature relationships (as shown by the biclustering result in Figure 1A).

The clinical observation-based classification of ovarian cancer includes cancer staging (IA–IV), which assesses the extent of cancer spreading, and cancer grading (G1–G3), which measures the differentiation of cancer cells. For the TCGA data, both the cancer-grade- and cancer-stage-based subtype classifications are not exclusively correlated with our seven subtypes, except for subtype 2, which is significantly associated with advanced tumor stages (IIIC and IV; Wilcoxon rank sum test, p = 8.995e-3). Compared with the molecular-signature-based classifications, these clinical classifications have much lower resolution for predicting prognosis, for two main reasons: (1) Most of the patients are assigned to a single advanced grade (86.2% patients are assigned to G3) or to a single advanced stage (71.2% patients are assigned to IIIC), with a difference between the shortest- and longest-prognosis cancer grades or tumor stages of 1.3- or 1.7-fold (log-rank test, p = 3.77e-02 or 8.71e-04 when comparing tumor grade G3 versus G1+G2 or tumor stage IV versus IA–IIIB). However, with our adaptive clustering approach, 32.0% and 41.6% patients are assigned to high- and low-risk groups, respectively, with a more significant difference of 2.0-fold between the shortest- and longest-prognosis subtypes (log-rank test, p = 3.16e-04; Figures 1B and 1C). (2) Samples of one stage are frequently classified into many different subtypes and have very different survival times. In the case of stage IIIC samples, they can be classified into all seven subtypes and show significant differences in survival (Figure S4B). Therefore, the molecular-signature-based subtype classification has its own merit beyond classifications based on clinical observations.

Our network also provides hints for designing specific treatment protocols for different ovarian cancer subtypes. For example, the metastasis and EMT of subtype 2 cases conceivably could be attenuated by known anticancer drugs that target the upregulated ECM receptor ITGB1 (volociximab) or the vascular endothelial growth factor (VEGF) receptor KDR (sunitinib and sorafenib). Additionally, drugs that are not currently

used in clinical treatments may also be worth trialing for subtype 2. According to Sharma et al. (2010), despite a lack of inhibitors to KDM5A, trichostatin A (TSA), an inhibitor to the KDM5A binding partner histone deacetylases (HDACs), can suppress the drug-dependent expansion of drug-resistant CSCs. We therefore postulate that TSA and other HDAC inhibitors can be used to effectively treat the ovarian subtypes 2, 4, and 5, and that AEW541, an inhibitor to IGF1R can be used in addition to TSA to treat subtype 2 to prevent chemotherapy drug resistance, thereby improving the prognosis of these subtypes and ovarian cancers in general. In addition to HDAC inhibitors and IGF1R inhibitors to treat drug-resistant CSCs, an inhibitor of TGF-β receptors, such as SB-431542 (Halder et al., 2005) or the antibody OMP-18R5, which targets the Wnt receptor Frizzled (Gurney et al., 2012), could also be considered to treat subtype 2 patients, since the TGF-β and Wnt signaling pathways are both of great importance in the EMT. Finally, the poor-prognosis subtypes 4 and 5 may also be treated using inhibitors of the oncogene SRC (dasatinib), apoptotic regulator BCR2L1 (navitoclax), or heat shock protein HSP90B1 (retaspimycin and ganetespib).

On the other hand, we note an interesting observation for subtype 7, which shows the longest median survival time. Within the DNA regions specifically deleted in subtype 7, there are only three known cancer-related genes: a protein phosphatase (*PPP2R1A*), which can negatively regulate cell growth and division, and two genes related to DNA repair (*ERCC1* and *ERCC2*; Figure 4). It was previously reported that the expression level of *ERCC1* is negatively correlated with the response to platinum-based chemotherapy (Steffensen et al., 2008; Vella et al., 2011). This agrees well with our observations (Table S5), as patients in subtype 4 with upregulated *ERCC1* were more likely to be platinum resistant, whereas patients in subtype 7 were not.

Although we treated each type of feature (DNA methylation, expression, and CNA) equally in our analysis, DNA methylation did not appear to be an important signature for subtype classification. One reason for this could be the low coverage of DNA methylation data. The Illumina 27k bead array covers no more than 0.1% of the total CpGs in the human genome, and the correlation between DNA methylation and gene expression is difficult to observe when using the methylation state of only a few CpGs to represent the DNA methylation state of a gene (there are on average two CpGs per promoter). With deeper whole genome-wide DNA methylation detection approaches gradually being adopted, we expect that some of the gene expression changes (e.g., changes in subtype 2) will be further explained by epigenetic marks.

EXPERIMENTAL PROCEDURES

Data Preprocessing

All data were downloaded from the TCGA (Cancer Genome Atlas Research Network, 2011). We used all 512 ovarian cancer samples for further analysis in which the patients' survival times or censor times were known and we had at least the expression and copy number data for the coding genes. The copy number segmentation data were mapped to the chromosomal positions of genes and miRNAs to quantitatively measure CNAs.

We used the clinically verified normal ovarian tissue data as controls, including the gene expression, DNA methylation, and miRNA expression data for eight samples, and the copy number data for 127 samples.

The TCGA also has intragenic somatic mutation data, but we did not use them for analysis because they cover only 63.3% (324) of the 512 samples and cannot be easily quantified.

Selection of Hazard Factors Using the Cox Proportional Hazard Model

We used the "coxph" program in the R statistics software to fit a univariate Cox proportional hazard model (Andersen and Gill, 1982) between each feature and the survival time of the patients. Samples with missing value(s) were excluded from the analysis. We then used the Wald test, likelihood ratio test, and chi-square statistics score test to filter the features. Only the features that passed the cutoff of $p < 0.05$ in all three tests were considered to be related to survival time and were selected as hazard factors for further analysis.

Adaptive Clustering for De Novo Subtype Identification

We started our clustering analysis with 4,526 features (hazard factors) selected by Cox regression. One of the 513 samples was not used for the clustering because it lacked copy number data, which account for 50.4% of all hazard factors. For each feature of the other 512 samples, missing values were filled by the median of nonmissing values and then the values were normalized as follows: $Value' = (Value - Median_{controls}) / STD_{patients}$, where $Value'$ is a vector of normalized values, $Value$ is a vector of raw values, $Median_{controls}$ is the median of normal tissue controls, and $STD_{patients}$ is the SD of all 512 patients.

In the adaptive clustering step, we used the BIC to determine the optimal number of clusters among the 512 samples or the 4,526 features (Schwarz, 1978). Because the BIC is mainly aimed at selecting the right number of clusters that will best balance the complexity of the clustering model and its fitness to data, it does not generate the most compact clustering result in the sense of least-squared deviations of the data samples to their corresponding cluster centers. As a result, after the adaptive clustering step, we further employed the super k-means algorithm to generate the desired compact clustering result for subsequent analysis (Arthur and Vassilvitskii, 2007; Hartigan and Wong, 1979) (see "The Adaptive Clustering Algorithm" in Extended Experimental Procedures for more details). When running the super k-means algorithm, we chose Euclidian distance to calculate the distances between the data and the cluster centers, and ran the algorithm 1,000 times to obtain the best clustering result with the lowest sum of squared distances from each point to its nearest cluster center.

The feature clusters and sample subtypes were further clustered using the hierarchical clustering algorithm based on the PCC between the centers of the super k-means clusters.

10-Fold Cross Validation

We performed 10-fold cross validation by randomly dividing the 512 samples into ten groups and hiding 10% of the samples each time to perform the super k-means clustering algorithm. The seven subtypes clustered by 90% of the samples were compared with the original all-sample clustering result, which was used as a benchmark to calculate the precision and sensitivity. The sensitivity and precision were defined as $Sensitivity = true\ positive/(true\ positive + false\ negative)$ and $Precision = true\ positive/(true\ positive + false\ positive)$. We then performed 10,000 permutations by randomly assigning the samples to seven subtypes in order to calculate the empirical p values of the sensitivity and precision. Then the 10% remaining samples or features were assigned to the nearest cluster centers (Euclidian distance) of the other 90% of samples as an independent test.

Testing Batch Effects

To examine whether batch effects affected the selection of hazard factors and the clustering result, we performed Fisher's exact test with Benjamini-Hochberg correction for each subtype versus each batch.

GO, KEGG, EMT, and hESC-Related Gene Sets Annotation Enrichment

Annotation enrichments were calculated as described previously (Xia et al., 2006). Briefly, the enrichment of GO terms (http://www.geneontology.org/), KEGG pathways (http://www.genome.jp/kegg/), EMT signaling pathways (Polyak and Weinberg, 2009), and hESC-associated gene sets (Ben-Porath et al., 2008) were calculated by Fisher's exact test on the systematically detected

Cell PRESS
Open ACCESS

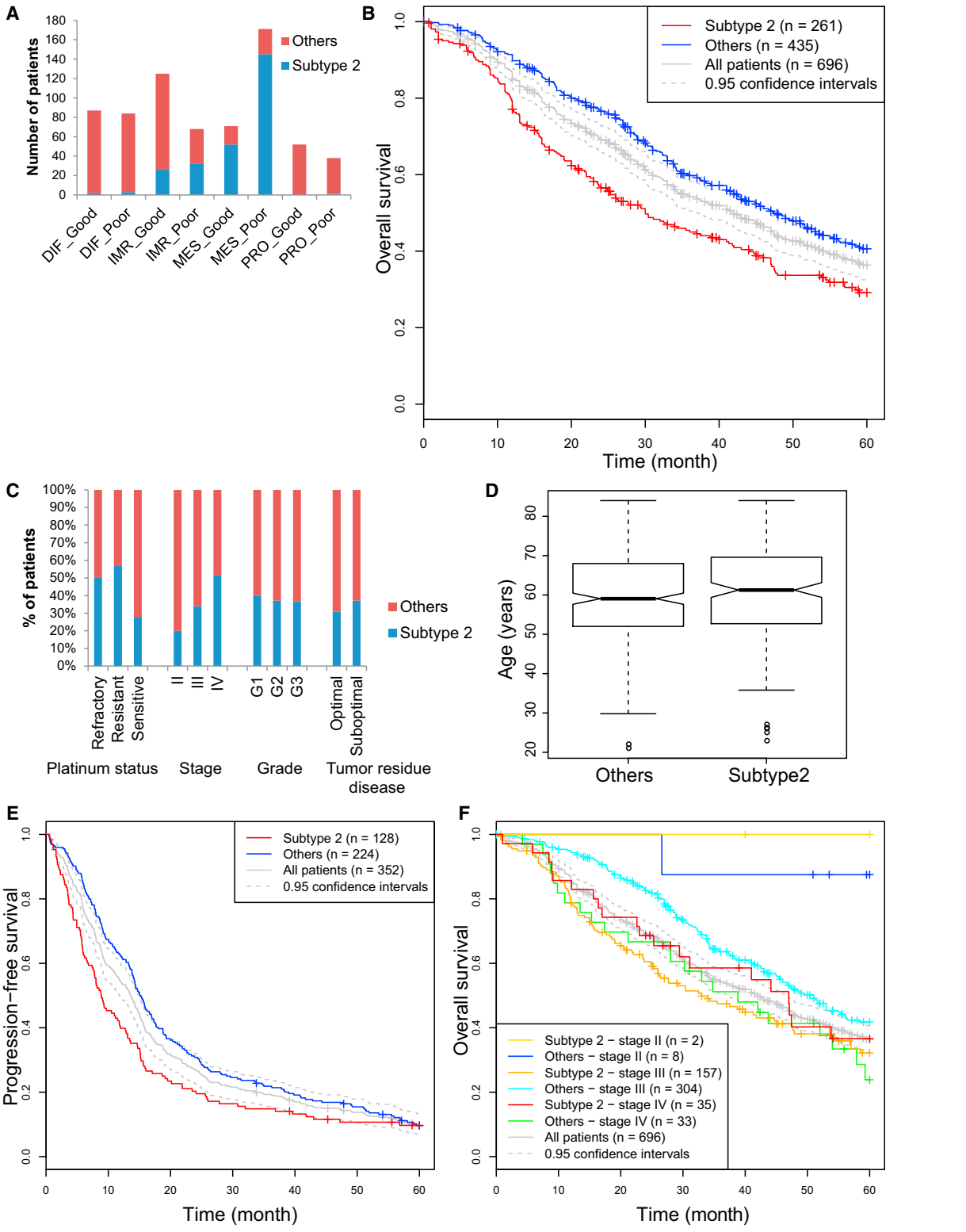


Figure 5. Validation of the Subtype 2 Classification Using an Independent Sample Set

(A) Number of overlapping patients between the predicted subtype 2 and the subtypes classified in a previous study (Verhaak et al., 2013), with DIF, IMR, MES, and PRO representing differentiated, immunoreactive, mesenchymal, and proliferative subtypes, respectively.

(B) Overall survival probabilities (y axis) are plotted against the survival times (months, x axis) of predicted subtype 2 or other patients. The survival curve of all 696 patients with 0.95 confidence intervals is also shown for comparison.

(legend continued on next page)

signature features of each subtype. A Benjamini-Hochberg-corrected false discovery rate (FDR) ≤ 0.05 was used to determine the enriched functions.

Literature Cocitations

Literature cocitations were calculated as described previously (Liu et al., 2013).

Building the Cancer Knowledge Base

The knowledge base of cancer-related genes was manually summarized based on Hanahan and Weinberg's reviews on the hallmarks of cancer (Hanahan and Weinberg, 2000, 2011). It contains the genes in the following 25 KEGG pathways and GO terms: (1) pathways in cancer, (2) MAPK signaling pathway, (3) mTOR signaling pathway, (4) ErbB signaling pathway, (5) Jak-STAT signaling pathway, (6) cytokine-cytokine receptor interaction, (7) cell cycle, (8) PPAR signaling pathway, (9) TGF-β signaling pathway, (10) apoptosis, (11) telomere maintenance, (12) VEGF signaling pathway, (13) Wnt signaling pathway, (14) ECM-receptor interaction, (15) adherens junction, (16) focal adhesion, (17) p53 signaling pathway, (18) base excision repair, (19) mismatch repair, (20) nucleotide excision repair, (21) inflammatory response, (22) glycolysis/gluconeogenesis, (23) T cell receptor signaling pathway, (24) B cell receptor signaling pathway, and (25) natural killer cell mediated cytotoxicity. It also contains the genes in ESC pluripotency signaling pathways (http://www.cellsignal.com/reference/pathway/ESC_pluripotency.html), multi-drug-resistance factors (http://www.biocarta.com/pathfiles/h_mrppathway.asp), and EMT signaling pathways (Polyak and Weinberg, 2009).

Constructing the Network

We used the human functional protein interaction network constructed by Wu et al. (2010) as a template to construct the subnetwork among the ovarian cancer subtype-specific hallmark genes. Wu et al.'s network template consists of manually curated interactions (MSKCC Cancer Cell Map [http://cancer.cellmap.org]; NCI-Nature Pathway Interaction Database [http://pid.nci.nih.gov]; KEGG [Kanehisa et al., 2004]; BioCarta [http://www.biocarta.com/genes/index.asp]; Reactome [Vastrik et al., 2007]; TRED [Jiang et al., 2007]; and pantherdb [Mi et al., 2007]) and predicted interactions derived from noncurated sources.

Additional methodologies we developed in this study are described in the Extended Experimental Procedures.

SUPPLEMENTAL INFORMATION

Supplemental Information includes Extended Results, Extended Experimental Procedures, four figures, and four tables and can be found with this article online at http://dx.doi.org/10.1016/j.celrep.2013.07.010.

WEB RESOURCES

The URLs for data presented herein are as follows:

BioCarta - Pathways, http://www.biocarta.com/genes/index.asp
BioCarta - Pathways: Multi-Drug Resistance Factors, http://www.biocarta.com/pathfiles/h_mrppathway.asp
Cell Signaling Technology, http://www.cellsignal.com/reference/pathway/ESC_pluripotency.html
Gene Ontology, http://www.geneontology.org
KEGG, http://www.genome.jp/kegg
MSKCC Cancer Cell Map, http://cancer.cellmap.org
NCI-Nature Pathway Interaction Database, http://pid.nci.nih.gov

AUTHOR CONTRIBUTIONS

J.-D.J.H. conceived the study. W.Z. performed the study and processed and analyzed the data. Y.L. and W.Z. developed the algorithm for uncovering the molecular signatures of cancer subtypes. Y.L. and N.S. developed the adaptive clustering algorithm. W.Z., J.-D.J.H., Y.L., D.W., and J.B.-K. analyzed the data and interpreted the results. X.D. helped W.Z. in processing parts of the data. All authors contributed to preparation of the manuscript.

ACKNOWLEDGMENTS

This project was funded by grants from the National Natural Science Foundation of China (NSFC; grants 31210103916 and 91019019), Chinese Ministry of Science and Technology (grant 2011CB504206), Chinese Academy of Sciences (CAS; grants KSCX2-EW-R-02, KSCX2-EW-J-15, and YZ201243), Stem Cell Leading Project (XDA01010303), and Shanghai Academic Leader Project (11XD1405700) to J.-D.J.H., and from Beijing Jiaotong University (K12RC00090 to Y.L.). J.B.-K. holds a CAS Fellowship for Young International Scientists (2011Y1SB05) and acknowledges support from the NSFC Fund for Young International Scientists (grant 31150110469).

Received: March 2, 2013
Revised: May 18, 2013
Accepted: July 9, 2013
Published: August 8, 2013

REFERENCES

Andersen, P., and Gill, R. (1982). Cox's regression model for counting processes, a large sample study. Ann. Stat. *10*, 1100–1120.

Arthur, D., and Vassilvitskii, S. (2007). k-means++: the advantages of careful seeding. Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms, 1027–1035.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.; The Gene Ontology Consortium. (2000). Gene ontology: tool for the unification of biology. Nat. Genet. *25*, 25–29.

Ben-Porath, I., Thomson, M.W., Carey, V.J., Ge, R., Bell, G.W., Regev, A., and Weinberg, R.A. (2008). An embryonic stem cell-like gene expression signature in poorly differentiated aggressive human tumors. Nat. Genet. *40*, 499–507.

Bolton, K.L., Tyrer, J., Song, H., Ramus, S.J., Notaridou, M., Jones, C., Sher, T., Gentry-Maharaj, A., Wozniak, E., Tsai, Y.Y., et al.; Australian Ovarian Cancer Study Group; Australian Cancer Study (Ovarian Cancer); Ovarian Cancer Association Consortium. (2010). Common variants at 19p13 are associated with susceptibility to ovarian cancer. Nat. Genet. *42*, 880–884.

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

Boyer, L.A., Lee, T.I., Cole, M.F., Johnstone, S.E., Levine, S.S., Zucker, J.P., Guenther, M.G., Kumar, R.M., Murray, H.L., Jenner, R.G., et al. (2005). Core transcriptional regulatory circuitry in human embryonic stem cells. Cell *122*, 947–956.

Ciriello, G., Cerami, E., Sander, C., and Schultz, N. (2012). Mutual exclusivity analysis identifies oncogenic network modules. Genome Res. *22*, 398–406.

(C) Percentage of platinum status, stage, grade, and tumor residue disease of ovarian cancer patients in predicted subtype 2 and other subtypes.
(D) Boxplot of the age at initial diagnosis of ovarian cancer patients in predicted subtype 2 and other subtypes. The ends of whiskers represent 1.5 interquartile ranges.
(E) Progression-free survival probabilities (y axis) are plotted against the survival times (months, x axis) of predicted subtype 2 or other patients.
(F) Overall survival probabilities (y axis) are plotted against the survival time (months, x axis) of predicted subtype 2 or other patients within each tumor stage separately.
See also Figure S4.

Crijns, A.P., Fehrmann, R.S., de Jong, S., Gerbens, F., Meersma, G.J., Klip, H.G., Hollema, H., Hofstra, R.M., te Meerman, G.J., de Vries, E.G., and van der Zee, A.G. (2009). Survival-related profile, pathways, and transcription factors in ovarian cancer. PLoS Med. *6*, e24.

Denkert, C., Budczies, J., Darb-Esfahani, S., Györffy, B., Sehouli, J., Könsgen, D., Zeillinger, R., Weichert, W., Noske, A., Buckendahl, A.C., et al. (2009). A prognostic gene expression index in ovarian cancer—validation across different independent data sets. J. Pathol. *218*, 273–280.

Gurney, A., Axelrod, F., Bond, C.J., Cain, J., Chartier, C., Donigan, L., Fischer, M., Chaudhari, A., Ji, M., Kapoun, A.M., et al. (2012). Wnt pathway inhibition via the targeting of Frizzled receptors results in decreased growth and tumorigenicity of human tumors. Proc. Natl. Acad. Sci. USA *109*, 11717–11722.

Halder, S.K., Beauchamp, R.D., and Datta, P.K. (2005). A specific inhibitor of TGF-beta receptor kinase, SB-431542, as a potent antitumor agent for human cancers. Neoplasia *7*, 509–521.

Hanahan, D., and Weinberg, R.A. (2000). The hallmarks of cancer. Cell *100*, 57–70.

Hanahan, D., and Weinberg, R.A. (2011). Hallmarks of cancer: the next generation. Cell *144*, 646–674.

Hartigan, J.A., and Wong, M.A. (1979). Algorithm AS 136: a k-means clustering algorithm. Appl. Stat. *28*, 100–108.

Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T., and Thun, M.J. (2008). Cancer statistics, 2008. CA Cancer J. Clin. *58*, 71–96.

Jiang, C., Xuan, Z., Zhao, F., and Zhang, M.Q. (2007). TRED: a transcriptional regulatory element database, new entries and other development. Nucleic Acids Res. *35*(Database issue), D137–D140.

Johannes, L. (2010). Test to help determine if ovarian masses are cancer. The Wall Street Journal, March 9, 2010.

Kanehisa, M., Goto, S., Kawashima, S., Okuno, Y., and Hattori, M. (2004). The KEGG resource for deciphering the genome. Nucleic Acids Res. *32*(Database issue), D277–D280.

Kanehisa, M., Goto, S., Furumichi, M., Tanabe, M., and Hirakawa, M. (2010). KEGG for representation and analysis of molecular networks involving diseases and drugs. Nucleic Acids Res. *38*(Database issue), D355–D360.

Kosary, C.L. (2007). Chapter 16: Cancer of the Ovary. NIH Pub. No. 07-6215 (Bethesda, MD: NIH).

Liu, Y., Qiao, N., Zhu, S., Su, M., Sun, N., Boyd-Kirkup, J., and Han, J.D. (2013). A novel Bayesian network inference algorithm for integrative analysis of heterogeneous deep sequencing data. Cell Res. *23*, 440–443.

Mi, H., Guo, N., Kejariwal, A., and Thomas, P.D. (2007). PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways. Nucleic Acids Res. *35*(Database issue), D247–D252.

Monti, S., Savage, K.J., Kutok, J.L., Feuerhake, F., Kurtin, P., Mihm, M., Wu, B., Pasqualucci, L., Neuberg, D., Aguiar, R.C., et al. (2005). Molecular profiling of diffuse large B-cell lymphoma identifies robust subtypes including one characterized by host inflammatory response. Blood *105*, 1851–1861.

Polyak, K., and Weinberg, R.A. (2009). Transitions between epithelial and mesenchymal states: acquisition of malignant and stem cell traits. Nat. Rev. Cancer *9*, 265–273.

Schwarz, G.E. (1978). Estimating the dimension of a model. Ann. Stat. *6*, 461–464.

Sharma, S.V., Lee, D.Y., Li, B., Quinlan, M.P., Takahashi, F., Maheswaran, S., McDermott, U., Azizian, N., Zou, L., Fischbach, M.A., et al. (2010). A chromatin-mediated reversible drug-tolerant state in cancer cell subpopulations. Cell *141*, 69–80.

Steffensen, K.D., Waldstrøm, M., Jeppesen, U., Brandslund, I., and Jakobsen, A. (2008). Prediction of response to chemotherapy by ERCC1 immunohistochemistry and ERCC1 polymorphism in ovarian cancer. Int. J. Gynecol. Cancer *18*, 702–710.

Tan, T.Z., Miow, Q.H., Huang, R.Y., Wong, M.K., Ye, J., Lau, J.A., Wu, M.C., Bin Abdul Hadi, L.H., Soong, R., Choolani, M., et al. (2013). Functional genomics identifies five distinct molecular subtypes with clinical relevance and pathways for growth control in epithelial ovarian cancer. EMBO Mol. Med. *5*, 1051–1066.

Cancer Genome Atlas Research Network. (2011). Integrated genomic analyses of ovarian carcinoma. Nature *474*, 609–615.

Tothill, R.W., Tinker, A.V., George, J., Brown, R., Fox, S.B., Lade, S., Johnson, D.S., Trivett, M.K., Etemadmoghadam, D., Locandro, B., et al.; Australian Ovarian Cancer Study Group. (2008). Novel molecular subtypes of serous and endometrioid ovarian cancer linked to clinical outcome. Clin. Cancer Res. *14*, 5198–5208.

Vastrik, I., D'Eustachio, P., Schmidt, E., Gopinath, G., Croft, D., de Bono, B., Gillespie, M., Jassal, B., Lewis, S., Matthews, L., et al. (2007). Reactome: a knowledge base of biologic pathways and processes. Genome Biol. *8*, R39.

Vella, N., Aiello, M., Russo, A.E., Scalisi, A., Spandidos, D.A., Toffoli, G., Sorio, R., Libra, M., and Stivala, F. (2011). 'Genetic profiling' and ovarian cancer therapy. Mol. Med. Rep. *4*, 771–777.

Verhaak, R.G., Tamayo, P., Yang, J.Y., Hubbard, D., Zhang, H., Creighton, C.J., Fereday, S., Lawrence, M., Carter, S.L., Mermel, C.H., et al.; Cancer Genome Atlas Research Network. (2013). Prognostically relevant gene signatures of high-grade serous ovarian carcinoma. J. Clin. Invest. *123*, 517–525.

Wu, G., Feng, X., and Stein, L. (2010). A human functional protein interaction network and its application to cancer data analysis. Genome Biol. *11*, R53.

Xia, K., Xue, H., Dong, D., Zhu, S., Wang, J., Zhang, Q., Hou, L., Chen, H., Tao, R., Huang, Z., et al. (2006). Identification of the proliferation/differentiation switch in the cellular network of multicellular organisms. PLoS Comput. Biol. *2*, e145.

Yoshihara, K., Tajima, A., Yahata, T., Kodama, S., Fujiwara, H., Suzuki, M., Onishi, Y., Hatae, M., Sueyoshi, K., Fujiwara, H., et al. (2010). Gene expression profile for predicting survival in advanced-stage serous ovarian cancer across two independent datasets. PLoS ONE *5*, e9615.