Title

Targeted resequencing of the microRNAome and 3′UTRome reveals functional germline DNA variants with altered prevalence in epithelial ovarian cancer

Journal

Oncogene, 34(16)

ISSN

0950-9232

Authors

Chen, X
Paranjape, T
Stahlhut, C
et al.

Publication Date

2015-04-16

DOI

10.1038/onc.2014.117

Peer reviewed

Oncogene (2015) **34,** 2125–2137

www.nature.com/onc

ORIGINAL ARTICLE

Targeted resequencing of the microRNAome and 3′UTRome reveals functional germline DNA variants with altered prevalence in epithelial ovarian cancer

X Chen[1,2,9], T Paranjape[3,9], C Stahlhut[1], T McVeigh[4], F Keane[5], S Nallur[3], N Miller[4], M Kerin[4], Y Deng[6], X Yao[6], H Zhao[2,7,8], JB Weidhaas[3] and FJ Slack[1]

Ovarian cancer is a major cause of cancer deaths, yet there have been few known genetic risk factors identified, the best known of which are disruptions in protein coding sequences (*BRCA1* and *2)*. Recent findings indicate that there are powerful genetic markers of cancer risk outside of these regions, in the noncoding mRNA control regions. To identify additional cancer-associated, functional non-protein-coding sequence germline variants associated with ovarian cancer risk, we captured DNA regions corresponding to all validated human microRNAs and the 3′ untranslated regions (UTRs) of ~6000 cancer-associated genes from 31 ovarian cancer patients. Multiple single-nucleotide polymorphisms in the 3′UTR of the vascular endothelial growth factor receptor/*FLT1, E2F2* and *PCM1* oncogenes were highly enriched in ovarian cancer patients compared with the 1000 Genome Project. Sequenom validation in a case–control study (267 cases and 89 controls) confirmed a novel variant in the *PCM1* 3′UTR is significantly associated with ovarian cancer ($P = 0.0086$). This work identifies a potential new ovarian cancer locus and further confirms that cancer resequencing efforts should not ignore the study of noncoding regions of cancer patients.

Oncogene (2015) **34,** 2125–2137; doi:10.1038/onc.2014.117; published online 9 June 2014

INTRODUCTION

Ovarian cancer is the most lethal gynecological cancer.[1] The high death rate is primarily due to patients presenting with advanced disease due to vague symptoms that delay diagnosis and a lack of well-known risk factors. Although there is a familial-inherited risk component for ovarian cancer risk, historically very few genetic abnormalities identified (BRCA1- and *BRCA2*-coding sequence mutations[2]) have been associated with a meaningful increased risk for the disease. Such previously identified inherited mutations associated with cancer risk all reside in the protein-coding region of the DNA, and account for only 3% of all cancers (ACS, 2010). Attempts to find new meaningful inherited mutations have taken global approaches such as genome-wide association studies, but these studies have found exclusively non-functional variants that may only be associated with regions of DNA that harbor the functional variants,[3,4] resulting in only small effect sizes that are not clinically useful.[3,4] Although cancer genome resequencing projects have identified a few additional genetic alterations in individual patient tumors, the success of these research programs has been limited, possibly because of the rarity of mutations that result in such complex phenotypic changes as oncogenesis, or because of their focus on protein-coding sequences.[5–8]

MicroRNAs (miRNAs) provide a powerful new avenue to the discovery of functional genetic risk factors in cancer. MiRNAs have been found to be altered in all cancer types studied, including ovarian cancer.[9,10] Owing to the importance of miRNA functions in development and growth, as well as their ability to target hundreds of genes simultaneously, single miRNA disruptions can enhance oncogenesis and hence, mutations in miRNA genes, and in their binding sites in cancer genes, are proving powerful in cancer risk assessment.[11–13] A recent study of reported single-nucleotide polymorphisms (SNPs) in miRNAs found a relatively low level of sequence variation in functional regions of miRNAs.[11] Several such polymorphisms have been identified and appear to be deleterious in cancer, making them likely candidates for causal variants.[14–17,18] It has also been shown that there are genetic variations within the 3′ untranslated regions (3′UTRs) of cancer genes and in some cases, the variations specifically alter miRNA-binding sites.[15] The first discovered and best studied mutation of this class is a functional 3′UTR inherited mutation in *KRAS (*rs61764370), which has been shown to be a risk factor for multiple cancers, including ovarian cancer.[15,19,20]

Given the existence of relatively rare, functional variants in miRNAs and their binding sites in target genes, we chose to systematically sequence germline genomic DNA obtained from ovarian cancer patients to discover additional functional variants associated with cancer in the miRNA regions and 3′UTRs of cancer-related genes. Our workflow consisted of capturing these regions using NimbleGen's sequence capture technology using a custom developed hybridization array followed by

[1]Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, CT, USA; [2]Program in Computational Biology and Bioinformatics, Yale University School of Medicine, New Haven, CT, USA; [3]Department of Therapeutic Radiology, Yale University School of Medicine, New Haven, CT, USA; [4]Discipline of Surgery, National University of Ireland Galway and Galway University, Hospitals, Galway, Ireland; [5]Yale University School of Medicine, New Haven, CT, USA; [6]Yale Center for Analytical Sciences, Yale University School of Medicine, New Haven, CT, USA; [7]Department of Genetics, Yale University School of Medicine, New Haven, CT, USA and [8]Department of Biostatistics, Yale School of Public Health, Yale University, New Haven, CT, USA. Correspondence: Dr FJ Slack, Department of Molecular, Cellular and Developmental Biology, Yale University, 266 Whitney Avenue, PO Box 208103, New Haven, CT 06520, USA or Dr JB Weidhaas, Department of Therapeutic Radiology, Yale University School of Medicine, New Haven, CT 06520, USA.
E-mail: frank.slack@yale.edu or joanne.weidhaas@yale.edu
[9]These authors contributed equally to this work.
Received 6 March 2014; accepted 26 March 2014; published online 9 June 2014

high-throughput paired-end sequencing to identify genetic variations using individual genomic DNA samples from ovarian cancer patients. The sequencing data sets for our patients were of high quality and we applied stringent quality control and filtering to ensure the accuracy of variant identification. We next used a network-wide analysis to focus on those genes with variation in their sequence and their expression in ovarian tumors. Subsequently, a subset of the known and novel variants was validated using Sequenom technology in a case–control cohort. We have identified multiple novel and known variants both in miRNA genes as well as in the 3′UTR of cancer-related genes. Many of the variants in the 3′UTRs were also found to lie in target sites for miRNAs. A case–control validation of a subset of these mutations confirms significant enrichment of one of these variants in ovarian cancer patients. Our results demonstrate the existence of additional functional genetic variation located in the noncoding regions of the DNA that may help identify individuals at increased genetic risk for developing ovarian cancer.

RESULTS

Target enrichment and high-throughput sequencing of miRNA genes and 3′UTRs of cancer genes

We generated and analyzed targeted high-throughput sequencing data sets of ~700 miRNAs and 3′UTRs of ~6000 cancer-associated genes to pinpoint sequence variants associated with ovarian cancer (Figure 1a). The ovarian cancer population we studied comprised of 31 women of European descent, high-risk, ovarian cancer patients identified through the Yale Cancer Genetics Core, who were expected based on personal and family history to have a potential inherited cancer risk. Patients were selected to be without other known genetic lesions associated with ovarian cancer (OC) risk such as *BRCA* mutations[2] or the KRAS-variant (rs61764370),[15,20] in order to enrich for novel variants. Six samples known to carry the KRAS-variant were included as positive controls.

We used a target gene capture procedure to obtain sequences enriched in all 718 of the known human miRNA genes in miRBase 14 and 3′UTRs of cancer-associated genes from cancer patients. First, genomic DNA isolated from saliva or blood specimens was separately sheared to fragments compatible with Illumina/Solexa sequencing systems. Second, the individual genomic DNA was denatured and hybridized to a custom DNA NimbleGen array[21–23] carrying sequences complementary to the 718 known miRNA genes (miRBase 14) and 3′UTRs of 5424 genes (obtained from the list of cancer genes in The Cancer Genome Atlas: http://cancergenome.nih.gov/). Finally, the captured DNA was eluted and processed through standard Solexa sequencing.[24] We successfully performed gene capture, and large-scale paired-end sequencing with ~500X coverage per sample, and the majority of sequencing reads were mapped to the target

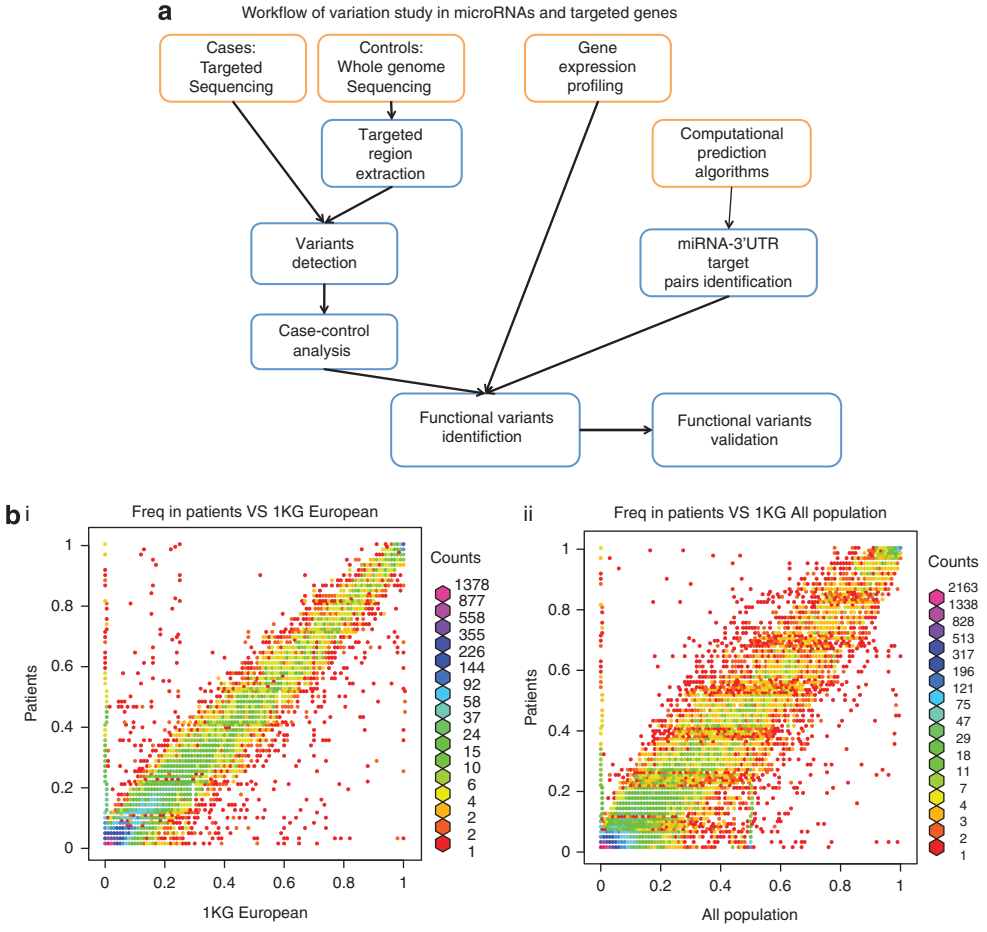

Figure 1. (**a**) Workflow for our integrated study of miRNAs and targeted genes. Targeted sequencing of miRNAs and 3′UTRs was performed on 31 ovarian cancer patients. These target regions were also extracted from the whole genome sequencing of the 1000 Genome project was used as controls to detect known and novel variants. To select functional and significant variants, we consider the allele frequency difference between cases and controls, differential expressed genes from gene expression profiling and miRNA-3′UTR-predicted target pairs. Finally, functional variants were validated by Sequenom in a larger corhort. (**b**) Allele frequency of known SNPs, patient samples versus the 1 KG database (European and all populations).

regions (TRs) by aligner BWA[25] (see Materials and methods, Supplementary Table 1).

Identification of known and novel genetic variants. Variations to the consensus genotype were called with Samtools (version 0.1.11).[26] To increase the quality of variant calls, the read depth threshold was set to 10. With this, each sample contained ~9000 SNPs passing the threshold compared with the reference genome hg18 downloaded from the University of California, Santa Cruz (UCSC) Genome Browser (http://genome.ucsc.edu/; Supplementary Table 2). Almost all of the samples have around 300 SNPs in pre-miRNA regions and 8500 other SNPs with a read depth of at least 10.

We successfully detected the KRAS-variant in the six cases where we had previously genotyped this SNP.[20] To determine whether SNPs identified in our patient samples corresponded to SNPs found in previous ovarian cancer studies, we compared our SNP list with the cancer-associated SNPs reported in 12 papers.[27–39] From a total of 94 SNPs (Supplementary Table 3), 28 were identified by our study as well. A few of them, such as rs10719 (RNASEN, *P*-value = 0.006), rs12889916 (SSTR1, *P*-value = 0.011), rs7499 (COL18A1, *P*-value = 0.028) and rs7869402 (TLR4, *P*-value = 0.039), have statistically different allele frequencies (*P*-value < 0.05) from normal populations present in the 1000 Genome Project (1 KG; Table 1), which indicated that our studies successfully identified certain known ovarian cancer-associated SNPs.

To determine how many novel SNPs were discovered in our sequencing data, the SNPs were compared with the dbSNP database (ver. 135)[40] and the 1000 Genomes Project (1 KG) (SNP calls are based on 1092 individuals from the 20101123 sequence and alignment release of the 1000 genomes project; 379 individuals from European population.).[41] We defined 'novel SNPs' as those not reported in dbSNP or 1 KG. We found approximately 50 novel SNPs in pre-miRNA regions and 900 novel SNPs in 3′UTRs per sample (Table 2a and Supplementary Table 4), which constitutes almost 5–10% of total SNPs identified from each patient, consistent with the identification rate of novel SNPs in other studies.[42] To determine the mutation properties of 3′UTR and miRNA regions, we compared our results with mutations in 1 KG by transition vs transversion (Ti/Tv rate) and substitution rate. We found that miRNAs have similar mutation properties to 1 KG, whereas 3′UTR regions are prone to have more transversions. A-G and T-C are frequent substitution types, but SNPs in our samples have more balanced substitutions between A and G and between T and C (Supplementary Figure 1), which further demonstrated that the sequencing of our samples was of high quality.

For known SNPs, we compared the allele frequency of SNPs with the frequency in 1 KG or dbSNP. The frequencies across all populations were obtained for all 24 834 known SNPs and the frequencies over the European population (which was the population used for this resequencing project) were obtained for 19 093 SNPs. The vast majority of SNPs have a very similar frequency between our patients and the 'normal' people from 1 KG, either all populations or European only, but 1 KG European population presents more similarity with the patients, as our patients are from this population (Figure 1b). However, certain SNPs presented with a much higher frequency in either our cases or the 1 KG controls, and are candidates to be associated with ovarian cancer (Figure 1b, Table 2b and Supplementary Table 5). For known SNPs in the 1 KG project, we performed Fisher's exact test on alleles in our patient samples and European population in 1 KG to test if the SNPs are associated with ovarian cancer. Within 19 093 known SNPs in the European population, 143 SNPs have a *P*-value less than 5×10^{-6}; within 3665 novel SNPs (the novel allele at least appears twice in patient samples), 724 SNPs have an allele frequency more than 10%. For example, known and novel SNPs in the 3′UTR of *HMGA1* (known SNP *P*-value = 2.28E-49; novel SNP allele frequency = 37.40%), an oncogene mis-expressed in ovarian cancer,[43] are enriched in the patient samples and are candidates for alleles with causal roles in ovarian cancer. Interestingly, *HMGA1* is also shown to have somatic mutations causally implicated in microfollicular thyroid adenoma and various

Table 1. Enrichment of previously described ovarian cancer SNPs

SNP ID	*Gene ID*	*Chr*	*Pos*	*Ref*	*Freq in DB*	*Alt*	*Freq*	P-*value*
rs10719	RNASEN	chr5	31437204	A	0.530	G	0.710	0.006
rs12889916	SSTR1	chr14	37749831	T	0.208	C	0.355	0.011
rs7499	COL18A1	chr21	45756756	G	0.465	A	0.323	0.028
rs7869402	TLR4	chr9	119517853	C	0.091	T	0.016	0.039
rs7957	TNFRSF10D	chr8	23049312	T	0.218	C	0.113	0.058
rs12245	KRAS	chr12	25249917	A	0.550	T	0.435	0.092
rs8065843	FLJ35220	chr17	76024941	G	0.686	T	0.790	0.095
rs4245739	MDM4	chr1	202785465	C	0.784	A	0.694	0.117
rs1126772	SPP1	chr4	89123210	A	0.188	G	0.258	0.188
rs9920	CAV1	chr7	115987328	T	0.044	C	0.081	0.199
rs895819	hsa-mir-27a	chr19	13808292	T	0.358	C	0.274	0.225
rs720014	DGCR8	chr22	18478882	T	0.214	C	0.274	0.272
rs12900401	SMAD3	chr15	65273644	C	0.037	T	0.065	0.294
rs17147016	UGT2A3	chr4	69829815	T	0.257	A	0.194	0.303
rs12010722	RPS6KA3	chrX	20080448	C	0.302	T	0.242	0.330
rs16869269	RRM2B	chr8	103288805	T	0.094	C	0.129	0.375
rs680	IGF2	chr11	2110210	T	0.717	C	0.774	0.391
rs2248718	ATP6V1C1	chr8	104151483	C	0.111	T	0.145	0.411
rs2910164	hsa-mir-146a	chr5	159844996	C	0.619	G	0.677	0.426
rs3757	DGCR8	chr22	18479331	G	0.212	A	0.258	0.431
rs17749202	WNT11	chr11	75575022	T	0.205	C	0.226	0.636
rs11169571	ATF1	chr12	49500032	T	0.342	C	0.323	0.788
rs2075993	E2F2	chr1	23708951	A	0.417	G	0.435	0.795
rs10900596	MDM4	chr1	202789080	T	0.573	C	0.597	0.795
rs12190214	ALDH5A1	chr6	24643187	C	0.072	A	0.081	0.801
rs6505162	hsa-mir-423	chr17	25468309	A	0.511	C	0.516	1.000
rs3917328	IL1R1	chr2	102160973	C	0.041	T	0.032	1.000

Table 2a. The top 20 of the list of novel SNPs residing in miRNA-binding sites ranked by allele frequency

Gene	*Chr*	*Pos*	*Ref*	*Alt*	*Freq*	*Targeted microRNA*
HS2ST1	chr1	87345928	G	AT	0.565	hsa-miR-3148
DDX17	chr22	37211649	T	A	0.516	hsa-miR-3145
UGT2B15	chr4	69547273	G	A	0.484	hsa-miR-545*, hsa-miR-376c
THOC4	chr17	77439146	G	A	0.484	hsa-miR-4311, hsa-miR-186, hsa-miR-3121
SLC2A12	chr6	134350481	A	G	0.452	hsa-miR-2054
SOX4	chr6	21705732	C	T	0.452	hsa-miR-186, hsa-miR-3133
NT5C2	chr10	104836975	A	G	0.419	hsa-miR-3128, hsa-miR-196a*
FABP7	chr6	123146836	C	T	0.387	hsa-miR-3163, hsa-miR-340, hsa-miR-452
CUL4A	chr13	112967273	C	T	0.387	hsa-miR-3148, hsa-miR-891a
CUL4A	chr13	112967272	T	AC	0.387	hsa-miR-3148, hsa-miR-891a
ARSJ	chr4	115042086	T	A	0.387	hsa-miR-3163, hsa-miR-142-5p, hsa-miR-873, hsa-miR-1252, hsa-miR-1286, hsa-miR-548c-3p, hsa-miR-548l
SNCA	*chr4*	*90865526*	*G*	*A*	*0.371*	*hsa-miR-4311*
GIGYF1	chr7	100117074	A	T	0.371	hsa-miR-4282, hsa-miR-548c-3p
RND3	chr2	151034403	C	A	0.371	hsa-miR-548p, hsa-miR-495, hsa-miR-376a*, hsa-miR-7-2*, hsa-miR-7-1*, hsa-miR-3065-5p, hsa-miR-3121, hsa-miR-410
HMGA1	*chr6*	*34321614*	*T*	*C*	*0.371*	*hsa-miR-495,hsa-miR-7-2*,hsa-miR-7-1*,hsa-miR-3065-5p*
MMP11	*chr22*	*22456467*	*T*	*A*	*0.355*	*hsa-miR-4264,hsa-miR-2053,hsa-miR-223,hsa-miR-4328,hsa-miR-500, hsa-miR-29b-1**
BASP1	chr5	17329847	A	C	0.355	hsa-miR-4307, hsa-miR-129-5p
ELAVL3	chr19	11423309	C	T	0.355	hsa-miR-511
MAP4K4	*chr2*	*101874911*	*G*	*A*	*0.323*	*hsa-miR-501-5p*
HMGA1	*chr6*	*34321585*	*C*	*T*	*0.323*	*hsa-miR-196a, hsa-miR-542-3p, hsa-miR-196b, hsa-miR-3148, hsa-miR-3125*

Abbreviations: Alt, alternative nucleotide; Chr, chromosome; Freq, allele frequency; Pos, genomic position; Ref, reference nucleotide.
Italic indicates genes differentially expressed in ovarian cancer. The entire list is shown in Supplementary Table 4.

Table 2b. The top 20 of the list of known SNPs residing in miRNA-binding sites ranked by *P*-value

Gene	*Chr*	*Pos*	*Ref*	*Alt*	*Db freq*	*Freq*	P-*value*	*Targeted microRNA*
IL18	*chr11*	*111519362*	*C*	*G*	*0.98*	*0.032*	*4.68E−74*	*hsa-miR-1178, hsa-miR-505, hsa-miR-4253, hsa-miR-1226*, hsa-miR-4260*
INPP5B	chr1	38100173	C	T	1	0.258	5.02E−62	hsa-miR-34c-5p, hsa-miR-34a, hsa-miR-449b, hsa-miR-449a
EIF3A	chr10	120785256	G	T	1	0.355	2.06E−52	hsa-miR-200c, hsa-miR-23b, hsa-miR-130a*, hsa-miR-23a, hsa-miR-371-5p
HMGA1	*chr6*	*34321274*	*T*	*C*	*1*	*0.387*	*2.28E−49*	*hsa-miR-4297*
ESRRA	chr11	63840689	A	G	1	0.452	1.76E−43	hsa-miR-600, hsa-miR-148b*, hsa-miR-627, hsa-let-7a-2*, hsa-miR-4294, hsa-miR-593, hsa-let-7g*, hsa-miR-493*, hsa-miR-924, hsa-miR-3121
MTFMT	chr15	63081112	G	A	1	0.484	1.25E−40	hsa-miR-548d-3p, hsa-miR-1323, hsa-miR-548x, hsa-miR-548o
BCKDHB	chr6	81110585	C	T	0.82	0.032	2.42E−38	hsa-miR-4253, hsa-miR-612, hsa-miR-654-5p, hsa-miR-1285, hsa-miR-762, hsa-miR-541
IDO2	chr8	39992635	T	C	1	0.532	1.89E−36	hsa-miR-4307, hsa-miR-183*, hsa-miR-548c-3p, hsa-miR-551b*, hsa-miR-570
CACNB2	*chr10*	*18870675*	*T*	*C*	*0.77*	*0.032*	*5.43E−33*	*hsa-miR-552*
PPP1R14B	*chr11*	*63768801*	*G*	*C*	*1*	*0.597*	*4.72E−31*	*hsa-miR-1228*, hsa-miR-886-5p, hsa-miR-3144-5p*
MAD2L1	chr4	121200662	T	C	0.71	0.016	1.63E−29	hsa-miR-3074, hsa-miR-181c, hsa-miR-625*, hsa-miR-144*, hsa-miR-181a, hsa-miR-410, hsa-miR-181b
ZNF28	chr19	57993057	T	C	0.91	0.290	5.79E−28	hsa-miR-3165
RAB7L1	chr1	204004404	T	C	0.91	0.290	5.79E−28	hsa-miR-541*, hsa-miR-1976
PSPH	chr7	56046588	T	C	0.74	0.065	4.53E−27	hsa-miR-105*
FGFR2	*chr10*	*123231486*	*T*	*C*	*1*	*0.677*	*1.50E−24*	*hsa-miR-152, hsa-miR-764, hsa-miR-552, hsa-miR-148a*
ALPK1	chr4	113582830	C	T	0.65	0.032	1.71E−23	hsa-miR-219-2-3p, hsa-miR-216a
AGPS	chr2	178113035	A	G	0.81	0.194	4.19E−23	hsa-miR-656, hsa-miR-410
VHL	chr3	10168683	T	G	0.7	0.081	1.13E−22	hsa-miR-4284, hsa-miR-484
ZNF665	chr19	58359449	C	T	0.6	0.016	5.48E−22	hsa-miR-125a-3p
TLX3	chr5	170671441	G	A	0.71	0.113	7.40E−21	hsa-miR-578, hsa-miR-525-3p, hsa-miR-103-2*

Italic ones are genes differentially expressed in ovarian cancer. The entire list is shown in Supplementary Table 5.

benign mesenchymal tumors in the COSMIC database.[44] Besides *HMGA1*, two other genes, *FGFR2* and *TLX3* with known SNPs near the top of our list, have causally implicated somatic mutations in non-small cell lung cancer (NSCLC), endometrial cancer (*FGFR2*) and T-cell acute lymphocytic leukemia (*TLX3*), respectively.

GATK[45] UnifiedGenotyper was used to further comfirm the SNPs called by Samtools. Among 24 834 known SNPs called by Samtools, 22 483 (90%) were called as SNPs and 492 (~2%) were called as indels in GATK; among 13 030 novel SNPs called by Samtools, 3421 (26%) were called as SNPs and 2255 (17%) were

called as indels in GATK (Supplementary Tables 4). This showed that variant calling programs have very high concordance on known SNP calling, but they have lower reliability in novel variant calling. We noticed that some SNPs have very small *P*-values comparing cases to controls. We suspect that this might be due to mapping and/or SNP calling biases and errors from either 1 KG or our own patient samples, or due to poor matching between cancer patients and 1 KG controls.

MiRNA target site prediction and gene expression analysis. We predicted the targets of all human miRNAs in miRBase v14 (including 5p, 3p and star miRNAs) by miRanda[46] and TargetScan,[47] and then compared targets with SNPs called from our samples. Of 24 834 distinct, known SNPs, 14 084 of them are within predicted miRNA-binding sites; of 13 030 distinct, novel SNPs, 7023 of them are within miRNA complementary sites. Almost half of SNPs are at putative miRNA complementary sites. To further evaluate the function of these SNPs, we obtained a list of differentially expressed genes in ovarian cancer from three independent gene microarray studies.[48–50] These studies identified 259, 2048 and 568 differentially expressed genes, respectively (77, 726 and 144 genes were examined in this study). Within them, the numbers of genes having SNPs (for novel ones, we only considered those with more than one alternative allele in cases) from our patient samples are 65 (357 SNPs), 642 (3921 SNPs) and 109 (575 SNPs; Table 3), respectively. To check if the SNPs are enriched in differentially expressed genes compared with all sequenced cancer genes, we considered the average number of SNPs per gene. The fold changes of the average number of SNPs per gene are 1.02, 1.18 and 1.00 for three gene sets compared with all sequenced genes. So the SNPs are comparatively enriched in the largest differentially expressed gene set; however, they are not enriched in the other two.

In Table 4, we show the top differentially expressed genes identified by at least two microarray studies, and the genes that have SNPs at miRNA target sites (Supplementary Table 6a for differentially expressed genes with SNPs, Supplementary Table 6b for the SNP list in genes identified by at least two differential expression microarrays). We note some very important cancer-associated genes in this table, including the oncogenes *FLT1* (*P*-value = 4.36E-13, encoding the vascular endothelial growth factor

Table 3. SNPs in differentially expressed genes

Authors	*No. of patients*	*Microarray platforms*	*No. of DE genes*	*No. of genes having SNPs*	*No. of genes having SNPs in miRNA target*
Chien *et al.*[50]	10	cDNA-DASL	259	65	54
Bowen *et al.*[49]	12	Oligonucleotides	2048	642	590
Ramakrishna *et al.*[48]	68	Oligonucleotides	568	109	98

Abbreviations: cDNA-DASL, cDNA-mediated annealing, selection, extension, and ligation; DE, differentially expressed; miRNA, microRNA; SNP, single-nucleotide polymorphism. For novel SNPs, we only considered those that have more than one alternative alleles in patient samples.

Table 4. SNPs in differentially expressed genes (identified in at least two studies) having SNPs in miRNA targets

(a) Known SNPs (top 10 in the list shown in Supplementary Table 6a), ranked by P-value

Gene	*Chr*	*Pos*	*Ref*	*Alt*	*Db freq*	*Freq*	P-*value*	*Targeted miRNAs*
FLT1	chr13	27840450	T	C	0.97	0.677	4.36E−13	hsa-miR-664*
E2F2	chr1	23706198	G	T	0.53	0.097	6.07E−12	hsa-miR-4278
GNAS	chr20	56919207	C	T	0.01	0.081	1.69E−03	hsa-miR-105*,hsa-miR-876-5p,hsa-miR-4273
SKIL	chr3	171593223	T	A	0.88	0.742	4.90E−03	hsa-miR-140-3p
BIRC5	chr17	73731801	T	C	0.4	0.258	3.01E−02	hsa-miR-936
FLT1	chr13	27859409	T	A	0.01	0.048	4.39E−02	hsa-miR-660
BIRC5	chr17	73733023	T	C	0.68	0.548	4.83E−02	hsa-miR-4325
FLT1	chr13	27858593	T	A	0.5	0.371	6.37E−02	hsa-miR-1285
FLT1	chr13	27860757	A	T	0.53	0.403	6.39E−02	hsa-miR-548a-3p, hsa-miR-582-3p, hsa-miR-553, hsa-miR-548e, hsa-miR-223*
BIRC5	chr17	73732965	A	G	0.69	0.581	8.85E−02	hsa-miR-764, hsa-miR-3127

(b) Novel SNPs. Top nine (SNP appearing at least twice) in the list shown in Supplementary Table 6b , ranked by allele frequency

Gene	*Chr*	*Pos*	*Ref*	*Alt*	*Freq*	*miRanda*
FLT1	chr13	27772664	A	C	0.081	hsa-miR-603
PTX3	chr3	158643893	G	T	0.065	hsa-miR-4307,hsa-miR-452*,hsa-miR-335*,hsa-miR-340, hsa-miR-190b,hsa-miR-33a*,hsa-miR-567,hsa-miR-190
FLT1	chr13	27861284	A	T	0.048	hsa-miR-218
FLT1	chr13	27859961	G	A	0.048	hsa-miR-548u
FLT1	chr13	27860592	T	G	0.032	hsa-miR-135a
INPPL1	chr11	71627495	G	T	0.032	hsa-miR-205
E2F2	chr1	23706194	A	T	0.032	hsa-miR-4278
DNMT3B	chr20	30860677	A	T	0.032	hsa-miR-569, hsa-miR-935, hsa-miR-145, hsa-miR-590-3p, hsa-miR-4282
SYNE1	chr6	152484598	T	C	0.032	hsa-miR-485-3p, hsa-let-7a-2*, hsa-miR-181d, hsa-miR-511, hsa-miR-655, hsa-miR-889, hsa-let-7g*, hsa-miR-493*, hsa-miR-1183, hsa-miR-2054, hsa-miR-548c-3p, hsa-miR-181b

Abbreviations: miRNA, microRNA; SNP, single-nucleotide polymorphism.

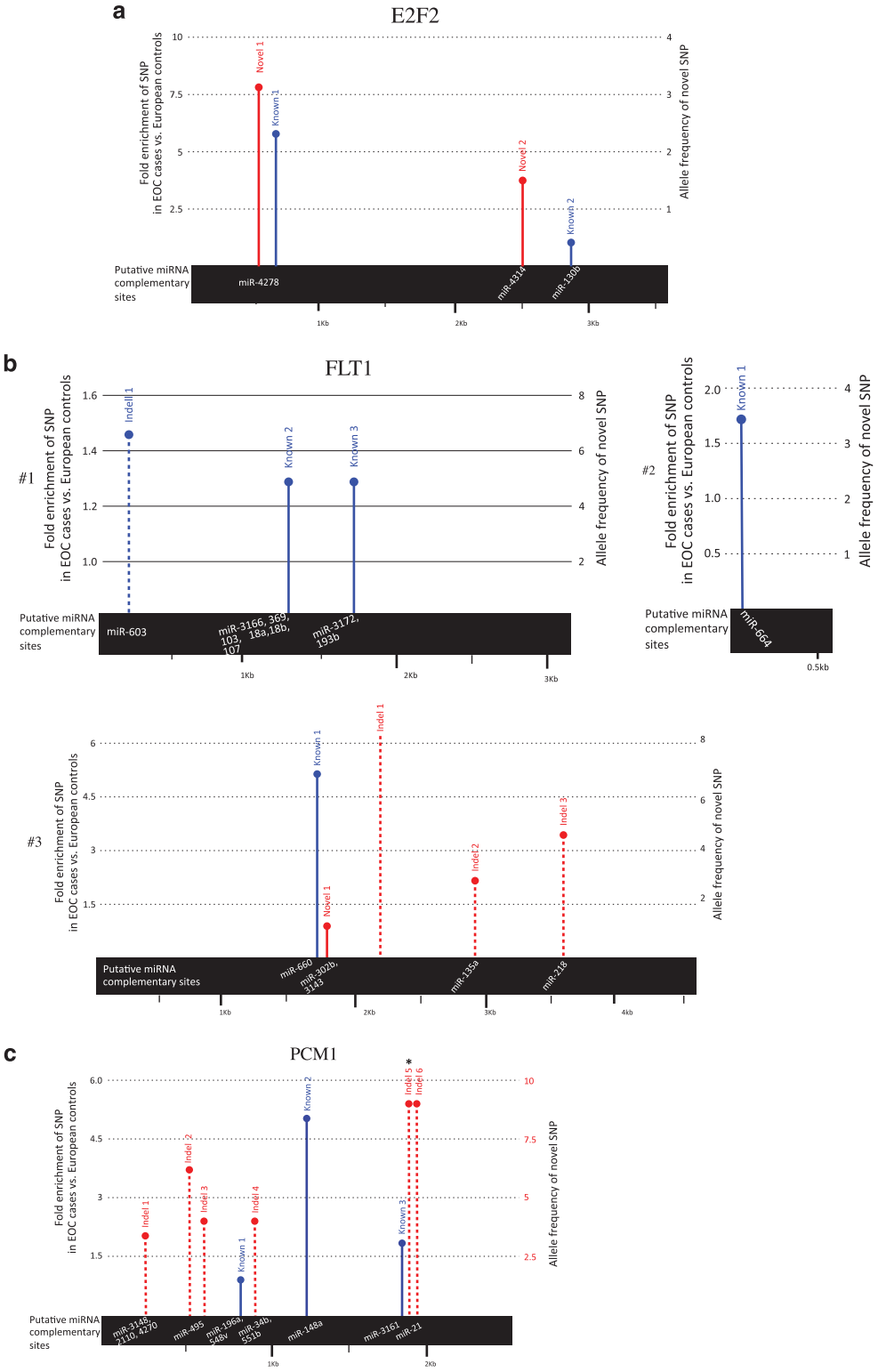

Figure 2. The SNPs and indels identified in our EOC cases. All novel SNPs and indels identified in EOC cases and only those known variants that were found to be enriched in EOC cases compared with reference European population controls from the 1000 Genomes Project were mapped to their relative positions in the 3′UTR of (**a**) *E2F2*, (**b**) *FLT1* (three separate 3′UTR transcripts) and (**c**) *PCM1* genes, respectively. The solid line represents a SNP, dashed line represents an indel, red indicates novel SNP/indel, blue is known SNP/indel and the height of the bar represents the fold enrichment of known SNPs/indels in our EOC cases vs reference European population controls from the 1000 Genomes Project or the allele frequency of novel SNPs in cases. The bar with a circular end represents SNPs within high confidence miRNA complementary sites, which are shown in the black bar. The asterix indicates the novel variant in the *PCM1* 3′UTR found to be significantly enriched in EOC cases.

receptor 1 tyrosine kinase), and *E2F2* (*P*-value = 6.07E-12, encoding a key transcription factor responsible for the G1-S transition). In fact, we found multiple known and novel SNPs as well as variants within the E2F2 and *FLT1* 3′UTRs that were highly enriched in our cancer patients (Figures 2a and b). Another gene, *PCM1,* with multiple known and novel variants with relatively high allele frequency in cancer patients, was also in the list (Figure 2c). The genes in this category were also screened for evidence of somatic mutations in cancers. The third gene in the list, *GNAS*, was found to contain mutations affecting pituitary adenoma.

Gene ontology analysis of differentially expressed SNP-containing genes putatively targeted by miRNAs revealed that the putative ovarian cancer-related genes involve kinase activity, nucleoside binding, metabolic process and cell cycle control. Kyoto Encyclopedia of Genes and Genomes (KEGG) pathway analysis showed the genes function in pathways in cancer (adjusted *P*-value = 5.528^{-13}), cell cycle (adjusted *P*-value = 8.045^{-5}) and small-cell lung cancer (adjusted *P*-value = 8.892^{-5}; Table 5).

Validation of candidate variants with Sequenom and Sanger sequencing. We tested 2 known and 19 novel SNPs (Supplementary Table 7) by Sequenom for validation in a prospectively collected clinically completely annotated case–control study ($n = 356$, 267 cases and 89 controls) by criteria as described in the Materials and methods section. Interestingly, most of our novel SNPs were detected in the control group, showing that 1 KG project may be under-reporting these variants, possibly due to low coverage. Univariate analysis was performed with χ^2 test and logistic regression, and then multivariate logistic regression was performed with age, and cancer-related mutation status (KRAS-variant and *BRCA*) as covariates to test whether the SNP allele is significantly associated with disease. From the 21 SNPs, one novel variant in gene *PCM1* (chr8 17931372) was significantly associated with ovarian cancer patients (χ^2 *P*-value of 0.0086). Multivariate analysis results further confirmed the association between this biomarker and disease risk, after adjusting age and *KRAS*/BRCA mutation status (Table 6, Figure 2c).

To further test the variant within a sample subgroup without known risk factors for ovarian cancer such as *BRCA* and KRAS-variant, we performed univariate and multivariate analyses for samples that do not carry *BRCA* mutations or the KRAS-variant. The *PCM1* variant is even more significantly associated with ovarian cancer patients in this subgroup (Fisher's exact *P*-value of 0.0014; Table 6C). Sanger sequencing of 256 samples (142 cases and 114 controls) was used to corroborate the data from the Sequenom validation (Supplementary Figure 2). This allele was also found in three of four ovarian cancer cell lines sequenced. CaOV3 and BG1 are homozygous for the ATTT insertion, IGROV1 is homozygous for the ATTT deletion and SKOV3 is heterozygous for the ATTT insertion.

To also evaluate if this variant may be a risk for other cancers, such as breast cancer, we subsequently evaluated this variant in a separate cohort (Supplementary Table 8) of prospectively collected clinically completed annotated Irish breast cancer cases ($n = 377$) and controls ($n = 372$), but did not find any significant association with breast cancer overall or with any specific molecular subtype (χ^2 *P*-value = 0.434; Supplementary Table 8). This finding further demonstrates the specific association of this novel variant with ovarian cancer.

We next evaluated the association of the *PCM1* variant with resistance to platinum chemotherapy and overall survival in ovarian cancer patients. The *PCM1* variant was not found to be associated with platinum resistance as analyzed by univariate and multivariate models (adjusted for histology, grade and stage, Supplementary Table 9a and b). In the overall survival analysis, both stage and histology (serous vs other types) were statistically significant predictors of differences in survival by univariate analysis. After adjusting for stage, age and histology, the effect of the *PCM1* variant on overall survival was not statistically significant ($P = 0.078$ Table 7). Next, we tested whether this novel variant was associated with a particular tumor type, grade, stage or histolgy and found that the variant was most prevalent in malignant mix mullerian tumors (Supplementary Table 10).

To test if the 3′UTR variant could affect regulation of the *PCM1* gene, we subcloned 1741 bp of the *PCM1* 3′UTR downstream of luciferase in a reporter plasmid. This region of the 3′UTR showed 1.5-fold repression of reporter gene expression compared with an empty vector control in the CaOV3 ovarian cancer cell line, and 6.2-fold repression in the MCF-7 breast cancer cell line (Figure 3a). This repression was not significantly affected by the presence of the ATTT variant insertion in the 3′UTR. To focus on a potential regulatory role for the ATTT insertion, we focused on the region of the 3′UTR immediately upstream and downstream of the location of the ATTT insertion. Accordingly, we generated luciferase reporters with an insert containing from 342 nt upstream to 249 nt downstream of the position of the variant, either containing the insertion (PCM1 SF1 ATTT Variant) or lacking it (PCM1 SF1). Expression of the reporter lacking the insertion was significantly upregulated relative to empty vector in CaOV3 ovarian cancer cells (Figure 3b). This upregulation was also observed in MCF-7 breast cancer cells (Figure 3b), HeLa cervical cancer cells and A549 lung

Table 5. GO and KEGG pathway analysis for differentially expressed genes with SNPs and putative miRNA targets

Term	P-*value*	*Benjamini*
GO molecular function		
GO:0004672 ~ protein kinase activity	1.61E−10	1.52E−07
GO:0004713 ~ protein tyrosine kinase activity	1.71E−07	8.10E−05
GO:0001883 ~ purine nucleoside binding	2.99E−07	9.43E−05
GO:0030554 ~ adenyl nucleotide binding	4.05E−07	9.57E−05
GO:0001882 ~ nucleoside binding	4.37E−07	8.26E−05
GO biological process		
GO:0006796 ~ phosphate metabolic process	6.73E−12	2.30E−08
GO:0006793 ~ phosphorus metabolic process	6.73E−12	2.30E−08
GO:0042127 ~ regulation of cell proliferation	1.12E−09	1.91E−06
GO:0022403 ~ cell cycle phase	4.51E−09	5.14E−06
GO:0009725 ~ response to hormone stimulus	5.94E−09	5.08E−06
GO cellular component		
GO:0031012 ~ extracellular matrix	5.41E−06	0.002361673
GO:0000793 ~ condensed chromosome	8.42E−06	0.001838134
GO:0005604 ~ basement membrane	1.32E−05	0.00192431
GO:0005578 ~ proteinaceous extracellular matrix	1.70E−05	0.001851468
GO:0005829 ~ cytosol	7.08E−05	0.006167834
KEGG pathway		
hsa05200:Pathways in cancer	5.53E−13	1.00E−10
hsa04110:Cell cycle	8.05E−05	0.007254812
hsa05222:Small-cell lung cancer	8.89E−05	0.005350669
hsa00330:Arginine and proline metabolism	0.001255698	0.055269935
hsa04510:Focal adhesion	0.001294948	0.045824355

Abbreviations: GO, gene ontology; miRNA, microRNA; SNP, single-nucleotide polymorphism. Only top five terms in each category are shown.

cancer cells (Supplementary Figure 3). Notably, reporter upregulation is partially (CaOV3, HeLa) or completely (MCF-7, A549) lost in reporters containing the ATTT insertion (Figure 3b, Supplementary Figure 1). These results indicate that the ATTT insertion has the potential to cause mis-regulation of *PCM1* expression, and may constitute a 'functional' variant.

Table 6. The *PCM1* variant is associated with ovarian cancer using a recessive model

(a) Univariate analysis by χ^2 test and logisitic regression for all samples

Table of case by PCM1

Case	*PCM1*			P-*value*
	0+1	2	Total	
Case	202 71.89%	65 86.67%	267	0.0086
Control	79 28.11%	10 13.33%	89	
Total	281	75	356	

Analysis of maximum likelihood stimates

Parameter		*DF*	*Estimate*	*Standard error*	*Wald* χ^2	*Pr*$>\chi^2$
Intercept		1	1.4053	0.1823	59.3983	$<$.0001
PCM1	ATTT insertion	1	0.4665	0.1823	6.5449	0.0105

Odds ratio estimates

Effect	*Point estimate*	*95% Wald confidence limits*	
PCM1 2 vs 0+1	2.542	1.244	5.195

(b) Multivariate analysis for all samples controlled for *KRAS*/BRCA mutation status and age

Analysis of maximum likelihood estimates

Parameter		*DF*	*Estimate*	*Standard error*	*Wald* χ^2	*Pr*$>\chi^2$
Intercept		1	2.6208	0.7100	13.6244	0.0002
PCM1	ATTT insertion	1	0.4461	0.1834	5.9161	0.0150
KV or BRCA mutation	0	1	0.0318	0.1530	0.0431	0.8356
Age		1	−0.0202	0.0110	3.4010	0.0652

Odds ratio estimates

Effect	*Point estimate*	*95% Wald confidence limits*	
PCM1 2 vs 0+1	2.441	1.189	5.009
KV or_BRCA_mutation 0 vs 1	1.066	0.585	1.941
Age	0.980	0.959	1.001

(c) Univariate analysis by Fisher's exact test and logisitic regression, and multivariate analysis for samples without *KRAS*/BRCA mutation controlled for age

Models (samples w/o KRAS/BRCA)	*Fisher's exact test*	*Univariate logistic regression*	*Multivariate logistic regression*
P-value	0.0014	0.0033	0.0034
OR	3.790	3.805	3.804
95% CI	1.529–11.307	1.677–10.261	1.672–10.277

Abbreviations: CI, confidence interval; DF, degrees of freedom; OR, odds ratio. 0: homozygous ATTT deletion, 1: heterozygous ATTT insertion, 2: homozygous ATTT insertion.

Table 7. The *PCM1* variant is not associated with overall survival in ovarian cancer

Type 3 tests

Effect	*DF*	*Wald* χ^2	*Pr*$>\chi^2$
Age	1	5.0475	0.0247
Stage_yale__X_means_	3	8.7872	0.0323
PCM1 variant	1	3.1046	0.0781
Histo	1	0.0836	0.7725

Analysis of maximum likelihood estimates

Parameter	*DF*	*Parameter estimate*	*Standard error*	χ^2	*Pr*$>\chi^2$	*Hazard ratio*	*Label*
Age	1	0.02907	0.01294	5.0475	0.0247	1.029	Age
Stage_yale__X_means_1	1	0.11009	0.74131	0.0221	0.8819	1.116	Stage yale (X means neoadjuvant) 1
Stage_yale__X_means_2	1	−2.02514	0.75154	7.2611	0.0070	0.132	Stage yale (X means neoadjuvant) 2
Stage_yale__X_means_3	1	−0.42677	0.25765	2.7436	0.0976	0.653	Stage yale (X means neoadjuvant) 3
PCM1 ATTT insertion	1	0.52038	0.29534	3.1046	0.0781	1.683	rs17931372 ATTT insertion
Histo other	1	−0.16139	0.55834	0.0836	0.7725	0.851	Histo Other

Multivariate analysis adjusted for age, stage and histology.

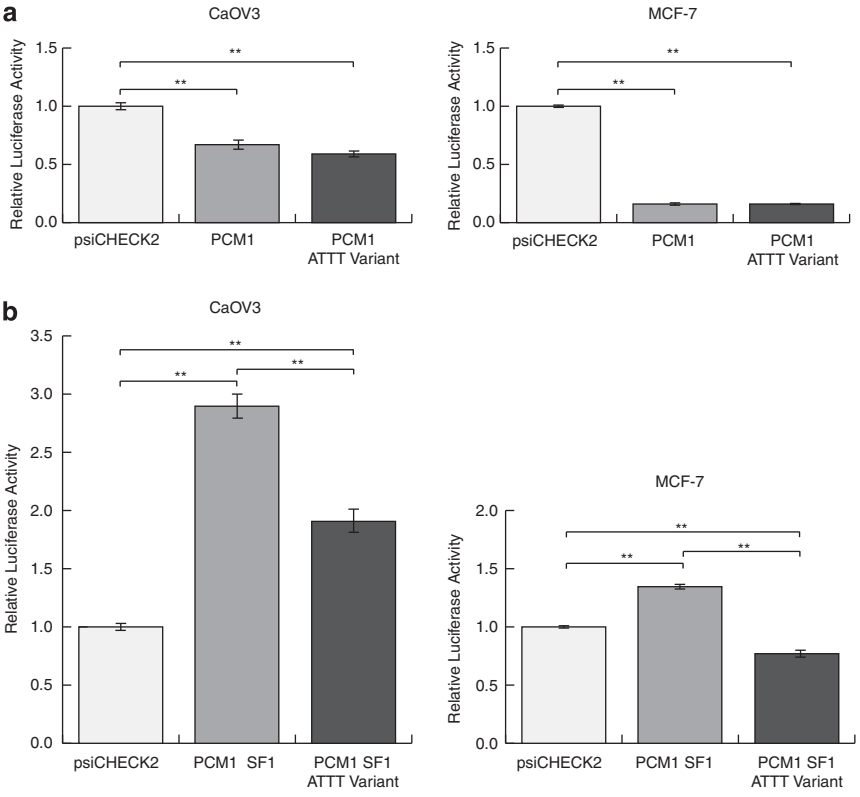

Figure 3. (**a**) The ATTT insertion mediates differential regulation of the *PCM1* 3′UTR. A luciferase reporter construct containing the *PCM1* 3′UTR (PCM1) is significantly repressed in CaOV3 and MCF-7 cells, relative to empty vector (psiCHECK2). This repression is maintained in a luciferase reporter containing the *PCM1* 3′UTR variant containing the ATTT insertion (PCM1 ATTT variant). (**b**) Luciferase reporters containing ~600 nt of the *PCM1* 3′UTR comprising the region flanking the position of the ATTT insertion were generated, either lacking the ATTT insertion (PCM1 SF1) or containing the insertion (PCM1 SF1 ATTT variant). PCM1 SF1 is significantly upregulated relative to empty vector (psiCHECK2) in CaOV3 and MCF-7 cells. This repression is partially lost in CaOV3 cells and completely lost in MCF-7 cells in PCM1 SF1 ATTT variant. Plotted: mean ± s.d.; $n = 3$; $^{**}P < 0.01$, Student's *t*-test.

DISCUSSION

In this work, we took a unique approach to identify genetic variations associated with ovarian cancer risk. We discovered a new variant in the *PCM1* 3′UTR that has been missed by prior efforts, is likely functional and is signficantly enriched in ovarian cancer patients, suggesting that it represents a potential new ovarian cancer risk loci. *PCM1,* pericentriolar material 1, is a centrosomal protein that shows dynamic distribution during the

cell cycle and exhibits a distinct cell cycle-dependent association with the centrosome complex.[51] Expression of the *PCM1* gene and its intracellular sublocalization are altered in papillary thyroid carcinoma.[52] Abberations of the gene have also been associated with atypical chronic myeloid leukemia and T-cell lymphoma.[51–54] Furthermore, *PCM1* has been shown to be differentially expressed in ovarian cancer patients,[49,55,56] supporting the validity of our finding.

Based on strong evidence that 3′UTRs and miRNAs have a critical role in oncogenesis, our study was a hypothesis-driven investigation of these regions through sequencing of the non-protein-coding regions/3′UTRs of ~6000 cancer genes and ~700 validated miRNA genes in 31 ovarian cancer patients. We further focused only on those genes with known varied expression in ovarian cancer, and applied bioinformatics to identify variants in predicted miRNA-binding sites. We found a significant number of novel variants not previously identified in any existing database. We further validated both the existance of a group of these variants as well as their enrichment in ovarian cancer using a case–control cohort and the Sequenom platform.

Our findings indicate that there are potentially numerous additional potential inherited markers of ovarian cancer risk in these previously poorly explored regions of the genome.[57,58] In particular, our functional variant in the 3′UTR of the known cancer gene *PCM1* was enriched in ovarian cancer patients and even more significantly enriched in patients without other known genetic risks for ovarian cancer. Although *PCM1* is known to be mis-expressed in ovarian cancer, this is the first report of a 3′UTR DNA variant in *PCM1* that could alter its expression. In addition, we identified variants in other important oncogenes, such as *FLT1*, which encodes a member of the vascular endothelial growth factor receptor family of receptor tyrosine kinases. This protein has an important role in angiogenesis, a key hallmark of ovarian cancer,[59] and is a target of ovarian cancer therapy.[60] We will continue to further validate the variants identified in our study, to both confirm their existence (considering the possibility of false positives of variant calling algorithms), their enrichment in ovarian cancer patients, as well as their function and association with clinical factors.

Although compared with large genome-wide association study data sets, our sample sets are small, we used hypothesis-directed investigation of specific regions of the genome. Our validation study set confirmed the existance of many of these novel variants, as well as their enrichment in ovarian cancer patients as well as biological function. Since existing data on 3′UTR variants indicates that their function often leads to altered outcome in cancer patients, it is critical to use well clinically annotated data sets, without ascertainment bias, as is found in most larger genome-wide association study-based data sets, to both discover and validate them. That said, further validation studies in the appropriate, clinically well annotated and non-biased data sets will further confirm our findings.

Our data further confirm the paradigm that the non-protein-coding regions of the genome must be included in resequencing projects and in DNA screening to better predict individual cancer risk, and that functional 3′UTR variants must be confirmed on appropriately data sets. Such approaches as described in this comprehensive analysis of variation in miRNA genes and 3′UTR regions have the potential to identify new markers that predict risk in diseases beyond ovarian cancer.

MATERIALS AND METHODS

Ethics statement

Complete clinical data and DNA from women diagnosed with EOC were included from different institutions as described previously[20] under individual International Review Board approvals.

Patient samples. The data for ovarian cancer patients are described in Supplementary Table 11. Patients were drawn from a study described previously.[20] Importantly, patients in this study were all prospectively collected, with complete clinical annotation, avoid selection bias as frequently found with other patient cohorts. The number of sequenced individuals is within an acceptable range used previously to obtain significant results.[61–63] Meanwhile, in our study, we used these 31 patients identify variants of interest for a larger case–control validation. Samples 2, 3, 8,11, 20 and 31 were known to carry the KRAS-variant and were included as positive controls. The data for Irish breast cancer cases are described in Supplementary Table 8. The controls were normal healthy subjects of Irish decent with ages ranging between 60 and 98 years with the median age being 70 years.

Evaluating 3′UTR and miRNA gene resequencing using NimbleGen Sequence Capture Arrays Gene selection. The total of all 718 human miRNAs from miRBase v14, a searchable database of published miRNA sequences and annotation,[64] was selected for resequencing. Cancer-related genes were obtained from the cancer gene list of The Cancer Genome Atlas. The coordinates of these 718 miRNA genes and 5437 3′UTRs covering a total 9 681 943 bp TR of interest were identified and submitted to Roche Diagnostics (Indianapolis, IN, USA) for custom array design using the 2.1 M (2.1 M probes) HX1 NimbleGen sequence capture array. The coordinates included the sequence of the pre-miRNAs plus an additional 200 bp of flanking sequences in order to cover the regulatory sequences as well as aid with efficient capture. Regions for 45 miRNA genes could not be covered on the NimbleGen array (Supplementary Table 12).

Sample preparation. The genomic DNA samples were quantified on a Nanodrop, and analyzed for quality and purity by gel electrophoresis. Genomic DNA was separately sheared to fragments compatible with the Solexa sequencing system. Next the individual genomic DNA was denatured and hybridized to our custom NimbleGen DNA array carrying sequences complementary to the 3′UTRs and miRNA genes of interest. Finally, the captured DNA was eluted and processed through standard high-throughput sequencing on an Illumina platform at the core sequencing laboratory at the Yale Center for Genome Analysis. Individual samples were run per well of the flow cell along with a standard positive control using a read length of 74 bp. All of these samples were sequenced by paired-end sequencing.

NimbleGen individual data analysis. Individual targeted sequencing reads were mapped to the reference genome using Burrows-Wheeler Aligner (BWA).[25] Each of the patient samples had more than 60 million sequenced reads (Supplementary Table 1). As expected, all of the samples had a high percentage of mappable reads. Approximately 95% of reads could be mapped to the reference human genome when three mismatches but no gaps were allowed. For paired-end sequencing reads, we also excluded the reads mapped with clipping before subsequent analyses. The number of reads in TRs and the average of coverage were determined. It has been claimed that 30X coverage is sufficient to identify SNPs from resequencing data.[26] Our results showed that our samples have a high percentage of reads mapping to TRs and have excellent coverage, about 500-fold (Supplementary Table 1). Most of the TRs were covered with a sufficient number of reads; a few positions or genes were not well captured by NimbleGen technology.

Consensus genotype calling and SNP association test. To infer which alleles are represented at a certain position according to the aligned reads, we performed consensus genotype calling. Based on the fact that sequencing data have errors and biases and the human is a diploid species, the consensus genotype should distinguish real heterozygous alleles from those resulting from errors and biases. A widely used method, Samtools,[26] calls the consensus genotype with a Bayesian model that incorporates correlated errors and diploid sampling. Samtools was able to achieve high sensitivity for our individual sequencing data. We used default settings to call the SNPs and then applied filtering of raw SNP calls with parameters -d8 -D10000 -11e-5 -20 -41e-7. Then, to further control the false positive rates, we discarded SNPs less than 4 bp away from a potential indel (called from gapped mapping results) or covered by less than 10 reads. If there were more than two SNPs in a 10-bp window, we discarded them all. As in our subsequent Sanger sequencing effort we discovered that one of the variants initially called as a SNP by Samtools was infact an indel, we used the GATK[45] to confirm the SNP calling results from Samtools. We preprocessed the reads mapped to the human genome by indel realignment, base quality recalibration and duplicates removal. Then the processed reads were subjected to UnifiedGenotyper in GATK to call

variants on all the patient samples simutaneously. We only retained SNPs and indels without missing calls from any sample. The GATK calling results are present in Supplementary Table 4. For known SNPs in the 1 KG project, we performed Fisher's exact test on alleles in patient samples and the European population in 1 KG to test if SNPs are associated with ovarian cancer. For novel SNPs, we only considered SNPs those appear at least twice in the patient samples, as there is the possibility for false positive calls in novel SNPs, especially when only one allele is called as novel SNP. However, novel SNPs (not found in 1 KG or dbSNP) may contain a large number of false positives because of region complexity or sequence composition, so further investigation is needed to validate the existence of these putative novel SNPs.

MiRNA target prediction and gene expression analysis. miRNA targets were predicted by two computational prediction programs—miRanda[46] and TargetScan,[47] with default settings. The genomic locations of putative binding sites were obtained from the prediction results and then matched with SNPs identified in the patients' samples. Three independent gene expression studies were used to extract differentially expressed genes in ovarian cancer. The gene lists were obtained from results of microarray analysis in three original papers. Gene ontology analysis and KEGG pathway analysis were conducted on differentially expressed genes with SNPs and putative miRNA targets with DAVID bioinformatics resources.[65,66]

SNP selection and genotyping for validation set. Twenty-one (19 novel and 2 known) SNPs with evidence for association with ovarian cancer were tested by Sequenom genotyping in a case–control study. SNPs were selected based on the minor allele frequency of $\geqslant 8\%$ in ovarian cancer patients, with a predicted miRNA-binding site at the SNP location and primer compatibility in multiplex genotyping on the Sequenom platform. Additionally for the known SNPs they had to be enriched in the patients with Fisher Exact test *P*-value less than 1×10^{-4} and a differential fold change of $\geqslant 2$ compared with the European controls in the 1 KG database. *FLT1* and *E2F* alleles failed the Sequenom primer design. SNPs were genotyped using the Sequenom MassArray system (Sequenom, San Diego, CA, USA) according to the manufacturer's instructions using 10 ng of genomic DNA. The validation cohort consisted of 267 ovarian cancer patients and 89 controls, mostly caucasians.[67] The controls were healthy subjects without a prior history of any cancer. Appropriate positive and negative control samples were included on the plates along with the samples to ensure genotyping accuracy. In every case, Sequenom successfully identified the correct allele (data not shown).

Statistical analysis. Patient charateristics were presented using descriptive statistics. χ^2 test and univariate logistic regression were performed to investigate the assocation between SNP with ovarian cancer. Odds ratio was calculated with 95% confidence interval. Multivariate logistic regression model was built to estimate the SNP association with ovarian cancer by taking into account age and *KRAS*/BRCA mutation status. χ^2 test was then used to estimate the association between the PCM1 variant with resistance to platinum chemotherapy, followed by a multivariate logistic regression model adjusted for histology, grade and stage. Log-rank test and Cox proportional-hazards model were used to test the significance of PCM1 in the prediction of overall survival time.

Sanger sequencing. A genomic fragment of ~623 bp surrounding the novel variant identified in *PCM1* 3′UTR was PCR amplified using the forward and reverse primers, PCM1-PCR-F: 5′-TTCCCTGCGAGGACATT TAC-3′ and PCM1-PCR-R: 5′- GGCCAGCTCATTATTTTAGGC-3′ from genomic DNA using KOD hot start polymerase. The PCR product was then verified on by agarose gel electrophoresis and ~70 ng of T-SAP-*Exo*I-digested product was sequenced using the above primers as well as a second set of nested primers; P-SEQ-F: 5′- CGGAGTTCTTATCCAGGTGCT-3′ and P-SEQ-R: 5′-TGAATGCCTAACCCTTCAGC-3′. This variant in *PCM1* was initially called as a SNP by Samtools, but after Sanger sequencing it was identified to be a 4-nt ATTT insertion corresponding to the reference allele (version hg18).

Luciferase reporter construct generation. The 3′UTR of *PCM1* was amplified from human genomic DNA by PCR using primers PCM1-UTR-F: 5′-atgcagCTCGAGgcccatccattaggccagtc-3′ and PCM1-UTR-R: 5′-cagattGCG GCCGCctcaacctgcataaagttctctct-3′. The PCR product was isolated and cloned between the *Xho*I-*Not*I sites of psiCHECK2 (Promega, Madison, WI, USA). This product contained the ATTT insertion, yielding psiCHECK-PCM1 ATTT variant. To create psiCHECK-PCM1, the ATTT insertion was deleted by primer extension mutagenesis with psiCHECK-PCM1 ATTT variant as template. The 5′ fragment was generated using primers PCM1-UTR-F and PCM-Del-RA: 5′-GATTAATAGCAGCTGTAACACCAGTCAAGCAATTTTGAT AAGG-3′. The 3′ fragment was generated using primers PCM-Del-FB: 5′-C CTTATCAAAATTGCTTGACTGGTGTTACAGCTGCTATTAATC-3′ and PCM1-UTR-R. The full-length 3′UTR was obtained by PCR with primers PCM1-UTR-F and PCM1-UTR-R in the presence of the 5′ and 3′ fragments as templates. This product was cloned between the *Xho*I-*Not*I sites of psiCHECK2 to produce psiCHECK-PCM1.

To generate reporters flanking the region ~300 nt upstream and downstream of the ATTT insertion in PCM1, this region of the *PCM1* 3′UTR was amplified from psiCHECK-PCM1 or psiCHECK-PCM1 ATTT variant to generate psiCHECK-PCM1 SF1 and psiCHECK-PCM1 SF1 ATTT variant, respectively. The PCR product was generated with primers PCM1SF-F: 5′-atgcagCTCGAGccctgcgaggacatttactg-3′ and PCM1-UTR-R: 5′-cagattGCG GCCGCctcaacctgcataaagttctctct-3′. The PCR products were isolated and cloned between the *Xho*I-*Not*I sites of psiCHECK2 to yield the final reporters.

Luciferase assays. Twenty-four hours before transfection, CaOV3 cells were seeded in antibiotic-free media in 12-well plates, at a density of 30 000 cells/well. One hundred nanogram of reporter was transfected using Lipofectamine 2000 (Invitrogen, Carlsbad, CA, USA), according to the manufacturer's instructions. Twenty-four hours after transfection, luciferase activity was assayed using the Dual Luciferase Reporter Assay (Promega) and a Glomax-Multi+ Plate Reader (Promega), following the manufacturer's instructions. After subtracting background measurements, *Renilla* luciferase activity intensity (IRluc) was normalized over firefly luciferase activity (IFluc). The fold change expression of the reporters relative to psiCHECK2 was calculated as: (IRluc *Reporter* /IFluc *Reporter*)/(IRluc *psiCHECK2*/IFluc *psiCHECK2*). The assay was repeated least three times for each reporter and the *P*-value was calculated by Student's *t*-test.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

XC was supported by Lo Graduate Fellowship for Excellence in Stem Cell Research and a fellowship from the China Scholars Council. TP was supported by a donation from William Hyman, Yale College Class of 1980, in memory of Barbara Skydel. We thank I Tikhonova, M Mahajan and S Mane at the YCGA for performing the Nimblegen enrichment and Sequenom arrays. We thank M State for critical reading of this manuscript. This work was supported by a grant from an anonymous foundation.

REFERENCES

1 Agarwal R, Kaye SB. Ovarian cancer: strategies for overcoming resistance to chemotherapy. *Nat Rev Cancer* 2003; **3**: 502–516.
2 Despierre E, Lambrechts D, Neven P, Amant F, Lambrechts S, Vergote I. The molecular genetic basis of ovarian cancer and its roadmap towards a better treatment. *Gynecol Oncol* 2010; **117**: 358–365.
3 Manolio TA, Collins FS, Cox NJ, Goldstein DB, Hindorff LA, Hunter DJ *et al.* Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
4 Hindorff LA, Sethupathy P, Junkins HA, Ramos EM, Mehta JP, Collins FS *et al.* Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proc Natl Acad Sci USA* 2009; **106**: 9362–9367.
5 Mardis ER, Ding L, Dooling DJ, Larson DE, McLellan MD, Chen K *et al.* Recurring mutations found by sequencing an acute myeloid leukemia genome. *New Engl J Med* 2009; **361**: 1058–1066.
6 Jones S, Hruban RH, Kamiyama M, Borges M, Zhang X, Parsons DW *et al.* Exomic sequencing identifies PALB2 as a pancreatic cancer susceptibility gene. *Science* 2009; **324**: 217.
7 Yan H, Parsons DW, Jin G, McLendon R, Rasheed BA, Yuan W *et al.* IDH1 and IDH2 mutations in gliomas. *New Engl J Med* 2009; **360**: 765–773.
8 Collins FS, Barker AD. Mapping the cancer genome. Pinpointing the genes involved in cancer will help chart a new course across the complex landscape of human malignancies. *Sci Am* 2007; **296**: 50–57.
9 Iorio MV, Visone R, Di Leva G, Donati V, Petrocca F, Casalini P *et al.* MicroRNA signatures in human ovarian cancer. *Cancer Res* 2007; **67**: 8699–8707.
10 Volinia S, Calin GA, Liu CG, Ambs S, Cimmino A, Petrocca F *et al.* A microRNA expression signature of human solid tumors defines cancer gene targets. *Proc Natl Acad Sci USA* 2006; **103**: 2257–2261.

11 Saunders MA, Liang H, Wen-Hsiung L. Human polymorphism at microRNAs and microRNA target sites. *Proc Natl Acad Sci USA* 2007; **104**: 3300–3305.
12 Blitzblau RC, Weidhaas JB. MicroRNA binding-site polymorphisms as potential biomarkers of cancer risk. *Mol Diagn Ther* 2010; **14**: 335–342.
13 Ryan BM, Robles AI, Harris CC. Genetic variation in microRNA networks: the implications for cancer research. *Nat Rev Cancer* 2010; **10**: 389–402.
14 Calin GA, Ferracin M, Cimmino A, Di Leva G, Shimizu M, Wojcik SE *et al.* A MicroRNA signature associated with prognosis and progression in chronic lymphocytic leukemia. *New Engl J Med* 2005; **353**: 1793–1801.
15 Chin LJ, Ratner E, Leng S, Zhai R, Nallur S, Babar I *et al.* A SNP in a let-7 microRNA complementary site in the KRAS 3' untranslated region increases non-small cell lung cancer risk. *Cancer Res* 2008; **68**: 8535–8540.
16 Shen J, DiCioccio R, Odunsi K, Lele SB, Zhao H. Novel genetic variants in miR-191 gene and familial ovarian cancer. *BMC Cancer* 2010; **10**: 47.
17 Shen J, Ambrosone CB, DiCioccio RA, Odunsi K, Lele SB, Zhao H. A functional polymorphism in the miR-146a gene and age of familial breast/ovarian cancer diagnosis. *Carcinogenesis* 2008; **29**: 1963–1966.
18 Hoffman AE, Zheng T, Yi C, Leaderer D, Weidhaas J, Slack F *et al.* microRNA miR-196a-2 and breast cancer: a genetic and epigenetic association study and functional analysis. *Cancer Res* 2009; **69**: 5970–5977.
19 Paranjape T, Heneghan H, Lindner R, Keane FK, Hoffman A, Hollestelle A *et al.* A 3′-untranslated region KRAS variant and triple-negative breast cancer: a case-control and genetic analysis. *Lancet Oncol* 2011; **12**: 377–386.
20 Ratner E, Lu L, Boeke M, Barnett R, Nallur S, Chin LJ *et al.* A KRAS-variant in ovarian cancer acts as a genetic marker of cancer risk. *Cancer Res* 2010; **70**: 6509–6515.
21 Hodges E, Xuan Z, Balija V, Kramer M, Molla MN, Smith SW *et al.* Genome-wide in situ exon capture for selective resequencing. *Nat Genet* 2007; **39**: 1522–1527.
22 Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, Zwick ME. Microarray-based genomic selection for high-throughput resequencing. *Nat Methods* 2007; **4**: 907–909.
23 Albert TJ, Molla MN, Muzny DM, Nazareth L, Wheeler D, Song X *et al.* Direct selection of human genomic loci by microarray hybridization. *Nat Methods* 2007; **4**: 903–905.
24 Kato M, de Lencastre A, Pincus Z, Slack FJ. Dynamic expression of small non-coding RNAs, including novel microRNAs and piRNAs/21U-RNAs, during Caenorhabditis elegans development. *Genome Biol* 2009; **10**: R54.
25 Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics* 2009; **25**: 1754–1760.
26 Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Res* 2008; **18**: 1851–1858.
27 Liang D, Meyer L, Chang DW, Lin J, Pu X, Ye Y *et al.* Genetic variants in MicroRNA biosynthesis pathways and binding sites modify ovarian cancer risk, survival, and treatment response. *Cancer Res* 2010; **70**: 9765–9776.
28 Li Y, Liang J, Kang S, Dong Z, Wang N, Xing H *et al.* E-cadherin gene polymorphisms and haplotype associated with the occurrence of epithelial ovarian cancer in Chinese. *Gynecol Oncol* 2008; **108**: 409–414.
29 Pongsavee M, Yamkamon V, Dakeng S, Oc P, Smith DR, Saunders GF *et al.* The BRCA1 3'-UTR: 5711+421T/T_5711+1286T/T genotype is a possible breast and ovarian cancer risk factor. *Genet Test Mol Biomarkers* 2009; **13**: 307–317.
30 Terry KL, Vitonis AF, Hernandez D, Lurie G, Song H, Ramus SJ *et al.* A polymorphism in the GALNT2 gene and ovarian cancer risk in four population based case-control studies. *Int J Mol Epidemiol Genet* 2010; **1**: 272–277.
31 Doherty JA, Rossing MA, Cushing-Haugen KL, Chen C, Van Den Berg DJ, Wu AH *et al.* ESR1/SYNE1 polymorphism and invasive epithelial ovarian cancer risk: an Ovarian Cancer Association Consortium study. *Cancer Epidemiol Biomarkers Prev* 2010; **19**: 245–250.
32 Pastrello C, Polesel J, Della Puppa L, Viel A, Maestro R. Association between hsa-mir-146a genotype and tumor age-of-onset in BRCA1/BRCA2-negative familial breast and ovarian cancer patients. *Carcinogenesis* 2010; **31**: 2124–2126.
33 Wynendaele J, Bohnke A, Leucci E, Nielsen SJ, Lambertz I, Hammer S *et al.* An illegitimate microRNA target site within the 3′ UTR of MDM4 affects ovarian cancer progression and chemosensitivity. *Cancer Res* 2010; **70**: 9641–9649.
34 Pearce CL, Doherty JA, Van Den Berg DJ, Moysich K, Hsu C, Cushing-Haugen KL *et al.* Genetic variation in insulin-like growth factor 2 may play a role in ovarian cancer risk. *Hum Mol Genet* 2011; **20**: 2263–2272.
35 Batra J, Nagle CM, O'Mara T, Higgins M, Dong Y, Tan OL *et al.* A Kallikrein 15 (KLK15) single nucleotide polymorphism located close to a novel exon shows evidence of association with poor ovarian cancer survival. *BMC Cancer* 2011; **11**: 119.
36 Permuth-Wey J, Kim D, Tsai YY, Lin HY, Chen YA, Barnholtz-Sloan J *et al.* LIN28B polymorphisms influence susceptibility to epithelial ovarian cancer. *Cancer Res* 2011; **71**: 3896–3903.
37 Lurie G, Wilkens LR, Thompson PJ, Shvetsov YB, Matsuno RK, Carney ME *et al.* Estrogen receptor beta rs1271572 polymorphism and invasive ovarian carcinoma risk: pooled analysis within the Ovarian Cancer Association Consortium. *PLoS ONE* 2011; **6**: e20703.
38 Peethambaram P, Fridley BL, Vierkant RA, Larson MC, Kalli KR, Elliott EA *et al.* Polymorphisms in ABCB1 and ERCC2 associated with ovarian cancer outcome. *Int J Mol Epidemiol Genet* 2011; **2**: 185–195.
39 Kontorovich T, Levy A, Korostishevsky M, Nir U, Friedman E. Single nucleotide polymorphisms in miRNA binding sites and miRNA genes as breast/ovarian cancer risk modifiers in Jewish high-risk women. *Int J Cancer* 2010; **127**: 589–597.
40 Smigielski EM, Sirotkin K, Ward M, Sherry ST. dbSNP: a database of single nucleotide polymorphisms. *Nucleic Acids Res* 2000; **28**: 352–355.
41 Durbin RM, Abecasis GR, Altshuler DL, Auton A, Brooks LD, Gibbs RA *et al.* A map of human genome variation from population-scale sequencing. *Nature* 2010; **467**: 1061–1073.
42 Choi M, Scholl UI, Ji W, Liu T, Tikhonova IR, Zumbo P *et al.* Genetic diagnosis by whole exome capture and massively parallel DNA sequencing. *Proc Natl Acad Sci USA* 2009; **106**: 19096–19101.
43 Masciullo V, Baldassarre G, Pentimalli F, Berlingieri MT, Boccia A, Chiappetta G *et al.* HMGA1 protein over-expression is a frequent feature of epithelial ovarian carcinomas. *Carcinogenesis* 2003; **24**: 1191–1198.
44 Bamford S, Dawson E, Forbes S, Clements J, Pettett R, Dogan A *et al.* The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer* 2004; **91**: 355–358.
45 McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A *et al.* The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res* 2010; **20**: 1297–1303.
46 John B, Enright AJ, Aravin A, Tuschl T, Sander C, Marks DS. Human MicroRNA targets. *PLoS Biol* 2004; **2**: e363.
47 Lewis BP, Shih IH, Jones-Rhoades MW, Bartel DP, Burge CB. Prediction of mammalian microRNA targets. *Cell* 2003; **115**: 787–798.
48 Ramakrishna M, Williams LH, Boyle SE, Bearfoot JL, Sridhar A, Speed TP *et al.* Identification of candidate growth promoting genes in ovarian cancer through integrated copy number and expression analysis. *PLoS ONE* 2010; **5**: e9983.
49 Bowen NJ, Walker LD, Matyunina LV, Logani S, Totten KA, Benigno BB *et al.* Gene expression profiling supports the hypothesis that human ovarian surface epithelia are multipotent and capable of serving as ovarian cancer initiating cells. *BMC Med Genomics* 2009; **2**: 71.
50 Chien J, Fan JB, Bell DA, April C, Klotzle B, Ota T *et al.* Analysis of gene expression in stage I serous tumors identifies critical pathways altered in ovarian cancer. *Gynecol Oncol* 2009; **114**: 3–11.
51 Corvi R, Berger N, Balczon R, Romeo G. RET/PCM-1: a novel fusion gene in papillary thyroid carcinoma. *Oncogene* 2000; **19**: 4236–4242.
52 Balczon R, Bao L, PCM-1 Zimmer WE. A 228-kD centrosome autoantigen with a distinct cell cycle distribution. *J Cell Biol* 1994; **124**: 783–793.
53 Adelaide J, Perot C, Gelsi-Boyer V, Pautas C, Murati A, Copie-Bergman C *et al.* A t(8;9) translocation with PCM1-JAK2 fusion in a patient with T-cell lymphoma. *Leukemia* 2006; **20**: 536–537.
54 Bousquet M, Quelen C, De Mas V, Duchayne E, Roquefeuil B, Delsol G *et al.* The t(8;9)(p22;p24) translocation in atypical chronic myeloid leukaemia yields a new PCM1-JAK2 fusion gene. *Oncogene* 2005; **24**: 7248–7252.
55 Armes JE, Hammet F, de Silva M, Ciciulla J, Ramus SJ, Soo WK *et al.* Candidate tumor-suppressor genes on chromosome arm 8p in early-onset and high-grade breast cancers. *Oncogene* 2004; **23**: 5697–5702.
56 Pils D, Horak P, Gleiss A, Sax C, Fabjani G, Moebus VJ *et al.* Five genes from chromosomal band 8p22 are significantly down-regulated in ovarian carcinoma: N33 and EFA6R have a potential impact on overall survival. *Cancer* 2005; **104**: 2417–2429.
57 Ramsingh G, Koboldt DC, Trissal M, Chiappinelli KB, Wylie T, Koul S *et al.* Complete characterization of the microRNAome in a patient with acute myeloid leukemia. *Blood* 2010; **116**: 5316–5326.
58 Parsons DW, Li M, Zhang X, Jones S, Leary RJ, Lin JC *et al.* The genetic landscape of the childhood cancer medulloblastoma. *Science* 2011; **331**: 435–439.
59 Markman M. Antiangiogenic drugs in ovarian cancer. *Expert Opin Pharmacother* 2009; **10**: 2269–2277.
60 Kumaran GC, Jayson GC, Clamp AR. Antiangiogenic drugs in ovarian cancer. *Br J Cancer* 2009; **100**: 1–7.
61 Wang L, Tsutsumi S, Kawaguchi T, Nagasaki K, Tatsuno K, Yamamoto S *et al.* Whole-exome sequencing of human pancreatic cancers and characterization of genomic instability caused by MLH1 haploinsufficiency and complete deficiency. *Genome Res* 2012; **22**: 208–219.
62 Cromer MK, Starker LF, Choi M, Udelsman R, Nelson-Williams C, Lifton RP *et al.* Identification of somatic mutations in parathyroid tumors using whole-exome sequencing. *J Clin Endocrinol Metab* 2012; **97**: E1774–E1781.
63 Liu P, Morrison C, Wang L, Xiong D, Vedell P, Cui P *et al.* Identification of somatic mutations in non-small cell lung carcinomas using whole-exome sequencing. *Carcinogenesis* 2012; **33**: 1270–1276.

64 Griffiths-Jones S, Grocock RJ, van Dongen S, Bateman A, Enright AJ. miRBase: microRNA sequences, targets and gene nomenclature. *Nucleic Acids Res* 2006; **34** (Database issue): D140–D144.

65 Huang da W, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nat Protoc* 2009; **4**: 44–57.

66 Huang da W, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res* 2009; **37**: 1–13.

67 Ratner ES, Keane FK, Lindner R, Tassi RA, Paranjape T, Glasgow M *et al.* A KRAS variant is a biomarker of poor outcome, platinum chemotherapy resistance and a potential target for therapy in ovarian cancer. *Oncogene* 2012; **31**: 4559–4566.

Supplementary Information accompanies this paper on the Oncogene website (http://www.nature.com/onc)