Supplementary file

Whole-exome Sequencing Combined with Functional Genomics Reveals Novel Candidate Driver Cancer Genes in Endometrial Cancer

Han Liang, Lydia W.T. Cheung, Jie Li, Zhenlin Ju, Shuangxing Yu, Katherine Stemke-Hale, Turgut Dogruluk, Yiling Lu, Xiuping Liu, Chao Gu, Wei Guo, Steven E. Scherer, Hannah Carter, Shannon N. Westin, Mary Dyer, Roeland Verhaak, Fan Zhang, Rachel Karchin, Chang-gong Liu, Karen H. Lu, Russell R. Broaddus, Kenneth L. Scott, Bryan T. Hennessy, Gordon B. Mills

Summary

Supplementary Methods

Exome sequencing

Agilent SureSelect™ Human All Exon Kit (Agilent P/N G3362A) for human all exon enrichment and SOLiD fragment library construction kit (Applied Biosystems, P/N 4443471) were used for sequencing library preparation following the provided protocols. The SureSelect™ Human All Exon kit design covers 1.22% of human genomic regions, approximately 38 Mb corresponding to the NCBI Consensus CDS database (CCDS). In brief, 14 endometrial cancer genomic DNA and their paired normal samples were quantitated and qualified individually by nanodrop-1000 (Nanodrop Technologies) and Agilent Bioanalyzer 2100 (Agilent Technologies). Two µg of intact genomic DNA of each sample in 100 µl low TE buffer was fragmented by Covaris S2 into target peak size of 100-150 base pairs (bp). The purification of fragmented genomic DNA was performed using Purelink PCR purification kit (Invitrogen). The fragmented genomic DNA was end repaired using both T4 DNA polymerase and Klenow DNA polymerase at room temperature for 30 min. After purification from end repair mixes, the DNA fragments were ligated with P1 and P2 adaptors on both ends at room temperature for 15 min. Then, a size selection of approximate 200 bp DNA fragments was performed by electrophoresis using E-gel SizeSelect 2% gel (Invitrogen, P/N G661002). The size-selected DNA library (150-200 bp) was amplified by nick translation performed on PCR 9700 thermocycler in 12 cycles using SureSelect pre-capture primers. The PCR products were quantified by Agilent bioanalyzer 2100 DNA 1000 assay. The amplified library demonstrated a peak of size around 200-250 bp.

In the experiment, 500 ng of individual library prepared was hybridized with exome capture library for 24 h at 65˚C according to the manufacturer's instructions (Agilent SureSelect Target Enrichment system V1.0). After hybridization, captured targets were enriched by pulling

down the biotinylated probe/target hybrids with streptavidin-coated magnetic beads (Dynal magnetic beads, Invitrogen) and purifying the targets with Qiagen MinElute PCR purification column. Finally, the enriched targeted-DNA libraries were further amplified by post hybridization PCR in 12 cycles and purified by PureLink PCR purification kit. The final library was quantitated by Agilent Bioanalyzer 2100 high sensitivity DNA assay.

The captured exome library was emulsified and amplified individually by emulsion PCR. Full length template beads were enriched and modified for deposition. The 14 endometrial tumors and paired normal libraries were sequenced in 50 nucleotide (nt) single tags in quad chambers of SOLiD™ V3.0. The tumor samples were sequenced in two quads per sample; and the matched normal samples were sequenced in one quad per sample.

Detection of somatic alternations

Supplementary Fig. 1 shows the overall computational pipeline for detecting somatic mutations. For each sample, the sequenced reads of 50 nt in color-space from SOLiD™ fragment were primarily analyzed with Applied Biosystems SOLiD™ System Bioscope (version 1.21) software package. The reads were first mapped to the unmasked human reference genome (hg19, http://genome.ucsc.edu) with default parameters. For targeted regions, single nucleotide variations (SNVs) in tumor samples were called using diBayes module with medium stringency. As an initial quality evaluation, the percentage of inferred SNVs already present in dbSNP (132) was used as an index of SNV calling accuracy. Among the 14 tumor samples, that percentage in one sample (86.1%) was substantially lower than that in the other 13 samples (from 91.1% to 93.9%, a typical range in exome-sequencing literature (Berger *et al.* 2010). Further analysis on tumor purity based on the contrast of reference allele and novel allele frequency at somatic

mutation and polymorphism positions suggested that this sample had the lowest tumor content. Therefore, the tumor sample was excluded from subsequent analysis. SNVs in normal samples were called using diBayes module with low stringency. To reduce false positives in final somatic mutations, additional filters were applied to tumor SNVs: (i) coverage ≥15×; (ii) mapping quality of novel alleles (QV) ≥11; (iii) the number of novel allele starts ≥8×; and (iv) *P*-value for SNV calling = 0. The cut-off value for each parameter was selected based on its effect on the fraction of tumor SNVs that were also called as SNVs in a normal sample, as shown in Supplementary Fig. 2. Accordingly, nucleotide positions with coverage of ≥15× in a tumor and ≥8× in the matched normal were defined as callable positions. To detect somatic mutations, tumor SNVs were filtered by removing those (i) also called as SNVs in any normal sample; (ii) already present in dbSNP (132)(Sherry *et al.* 2001) and (iii) common SNPs in the 1000 Genomes Project (1000 Genome Project Consortium 2010). Any tumor SNVs already in the COSMIC database (Bamford *et al.* 2004) were retained. Small indels (deletions up to 11 bp and insertions up to 3 bp) were called with Find Small Indels Tools in Bioscope with default parameters. Somatic small indels were identified if a tumor indel (i) had a coverage of ≥ 8× in the matched normal; (ii) not called as an indel in any normal sample; and (iii) not present in dbSNP (132). The functional annotation of somatic mutations or indels was performed with ANNOVAR (Wang *et al.* 2010).

Precision and sensitivity estimation of mutation calling

To estimate the accuracy of our mutation calling method, 97 non-silent somatic mutation sites were selected based on their potential biological interest for Sequenom MASSarray validation. Sequenom assays were designed with the AssayDesign software (version 3.0). The assay design was successful for 95 of the 97 sites. Primers were purchased from Sigma (Houston, TX, USA)

and pooled as indicted by the AssayDesign software. The genomic DNA around the putative SNV was amplified in a multiplex PCR reactions using Qiagen Hotstar Mastermix supplemented with 1.5 mM $MgCl_2$. Unincorporated primers were removed by Shrimp Alkaline Phosphatase (SAP) digestion at 37C for 40 min, then the SAP was heat inactivated at 85°C for 5 min. Thermosequenase (Sequenom) was used for primer extension reactions according to Sequenom's standard protocol. All samples were run in duplicate and visually inspected using the Sequenom Typer 3.4 and Typer 4.0 software. In-house software was used to determine whether the putative SNV was confirmed (i.e., the base change was present in the tumor but not the matched normal sample). Of the 95 sites tested, 77 mutations were validated (13 SNPs and 5 wild type), yielding an estimated true positive rate of 81%. Sensitivity is more challenging to assess than precision, owing to the lack of a set of "ground truth" mutations. Re-sequencing of nine genes (*CTNNB1, PIK3CA, ARID1A, FRAP1, PIK3CG, PIK3R5, RAB11F1P5, RPS6KC1 and TYRO3)* in the tumor samples was performed with Sanger sequencing at the Human Genome Sequencing Center at the Baylor College of Medicine (Houston, Texas) in order to search for potentially missed somatic mutations. Sequenom analysis was performed for hot spot mutations in *PIK3CA* to provide additional confidence. Each novel (non-silent) tumor SNV in these genes was further genotyped by Sequenom in both tumor and the matched normal samples to determine whether it was a true somatic mutation. With this approach, there were 15 somatic mutations identified in the callable positions of these genes in total, and 12 were also detected by SOLiD, yielding an estimated sensitivity of 80%.

References

Bamford, S., Dawson, E., Forbes, S., Clements, J., Pettett, R., Dogan, A., Flanagan, A., Teague, J., Futreal, P.A., Stratton, M.R. et al. 2004. The COSMIC (Catalogue of Somatic Mutations in Cancer) database and website. *Br J Cancer* **91:** 355-358.

Berger, M.F., Levin, J.Z., Vijayendran, K., Sivachenko, A., Adiconis, X., Maguire, J., Johnson, L.A., Robinson, J., Verhaak, R.G., Sougnez, C. et al. 2010. Integrative analysis of the melanoma transcriptome. *Genome Res* **20:** 413-427.

1000 Genome Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* **467:** 1061-1073.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29:** 308-311.

Wang, K., Li, M., and Hakonarson, H. 2010. ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Res* **38:** e164.

Supplementary Table 1 Clinical characteristics of patients providing samples for exome sequencing

	Endometrial Carcinoma					Mismatch Repair
Sample ID	**Histotype**	**Grade**	**Stage**	**Age**	**Ethnicity**	**IHC**
E27	Endometrioid & serous	3	IIIc	66	Caucasian	Intact positive
E28	Endometrioid	3	IIIc	83	Caucasian	Intact positive
E35	Endometrioid & serous	2	IIIa	71	Caucasian	Intact positive
E58	Endometrioid	2	IIa	46	Caucasian	Intact positive
E62	Endometrioid	2	IIIc	71	Caucasian	MSH2/MSH6 negative
E70	Endometrioid	2	IIb	71	Hispanic	Intact positive
E82	Endometrioid	2	Ic	64	Hispanic	Intact positive
E99	Endometrioid	2	IIIa	41	Caucasian	Intact positive
E101	Endometrioid	2	IVb	65	Caucasian	Intact positive
E114	Endometrioid	2	Ib	73	Caucasian	Intact positive
E161	Endometrioid	2	IVb	53	African	Intact positive
E170	Endometrioid	2	Ib	45	Hispanic	Intact positive
E172	Endometrioid	2	Ia	42	Hispanic	MLH1 negative

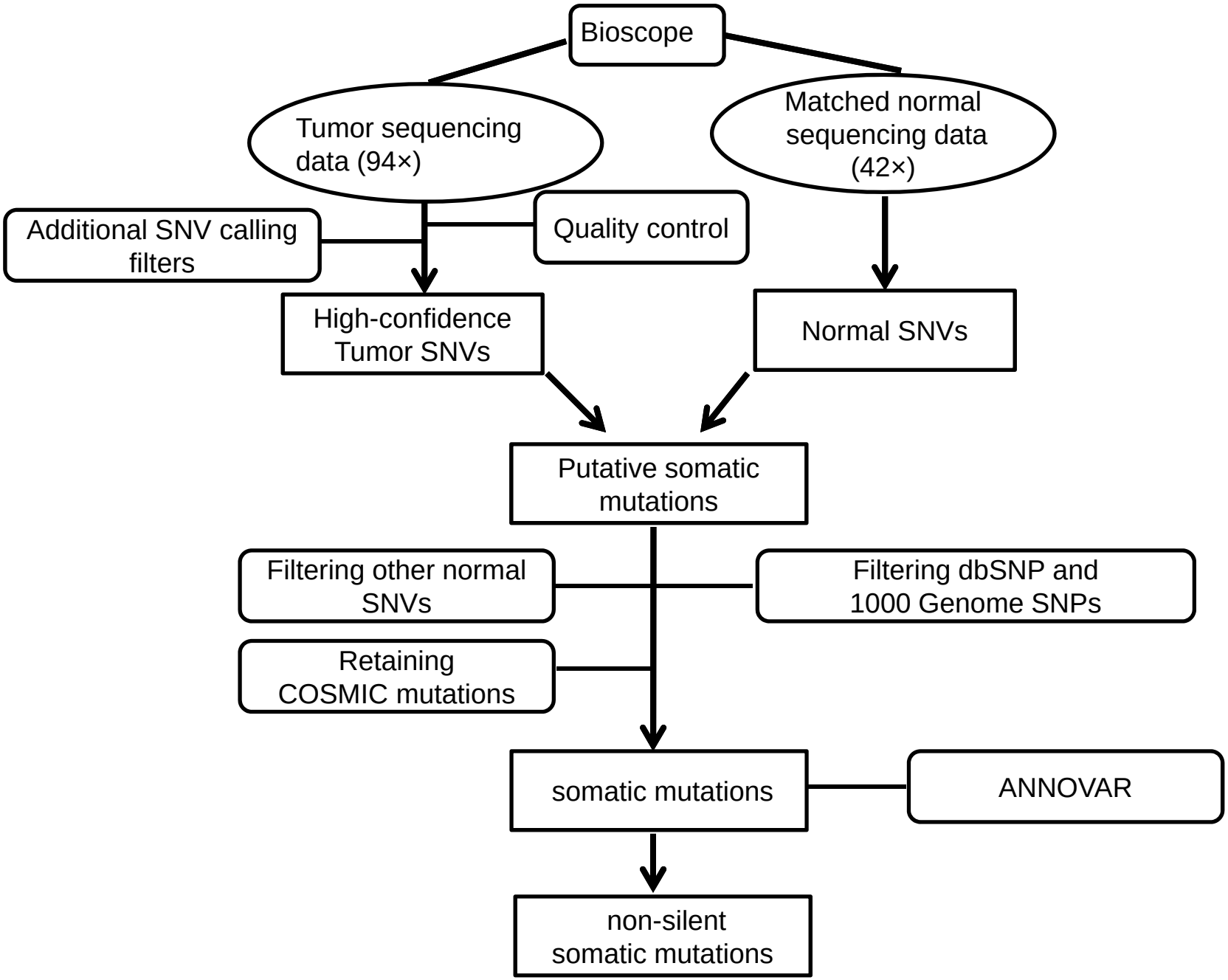
Supplementary Table 3 Top biological pathways enriched in the mutated genes

Canonical Pathways	*P*-value	FDR	Ratio	Molecules
Integrin Signaling	0.00017	0.023	0.076	PIK3CA,PIK3R1,ITGA8,BCAR1,ITGAL,PTEN,PTK2,MYLK, TLN2, PIK3CG,CAPN9,PIK3R2,ITGA7,CTTN,NEDD9,ACTN1
Complement System	0.00028	0.023	0.18	C9,C8B,C5,C6,C8A,CR2
Angiopoietin Signaling	0.00046	0.023	0.10	PTK2,PIK3CA,ANGPT2,FOXO1,DOK2,PIK3CG,PIK3R1,PIK3R2
Lymphotoxin β Receptor Signaling	0.00077	0.024	0.12	PIK3CA,PIK3CG,PIK3R1,TRAF4,TRAF5,PIK3R2,EP300
FAK Signaling	0.00078	0.024	0.091	PTK2,PIK3CA,TLN2,PIK3CG,PIK3R1,CAPN9,PIK3R2,BCAR1, PTEN
PTEN Signaling	0.00095	0.024	0.083	PTK2,PIK3CA,FOXO1,PIK3CG,PIK3R1,PIK3R2,IGF2R,BCAR1,P DGFRB,PTEN
Crosstalk between Dendritic Cells and Natural Killer Cells	0.00098	0.024	0.098	KIR3DL1,KIR3DL3/LOC100133046,KIR2DL3,TLN2,CD209, ITGAL,PVRL2,HLA-C,CAMK2G
NF-κB Activation by Viruses	0.0010	0.024	0.10	PIK3CA,RIPK1,PIK3CG,PIK3R1,MAP3K1,PIK3R2,ITGAL,CR2
Sphingosine-1-phosphate Signaling	0.0011	0.024	0.086	PTK2,PIK3CA,PLCE1,PIK3CG,PIK3R1,ADCY1,PIK3R2,ADCY8, CASP5,PDGFRB
Leptin Signaling in Obesity	0.0012	0.025	0.10	PIK3CA,PLCE1,FOXO1,PIK3CG,PIK3R1,ADCY1,PIK3R2,ADCY8
SAPK/JNK Signaling	0.0013	0.025	0.089	MAP4K3,TP53,PIK3CA,RIPK1,DUSP10,PIK3CG,PIK3R1, MAP3K1,PIK3R2
TR/RXR Activation	0.0028	0.046	0.090	PIK3CA,COL6A3,NCOA6,PIK3CG,PIK3R1,PIK3R2,SYT12,EP300
Acute Phase Response Signaling	0.0028	0.046	0.070	PIK3CA,HPX,FN1,RIPK1,ITIH2,PIK3CG,PIK3R1,MAP3K1,C9,C5, PIK3R2,AGT

Supplementary Table 4 List of 30 mutated genes selected for Ba/F3 functional assays

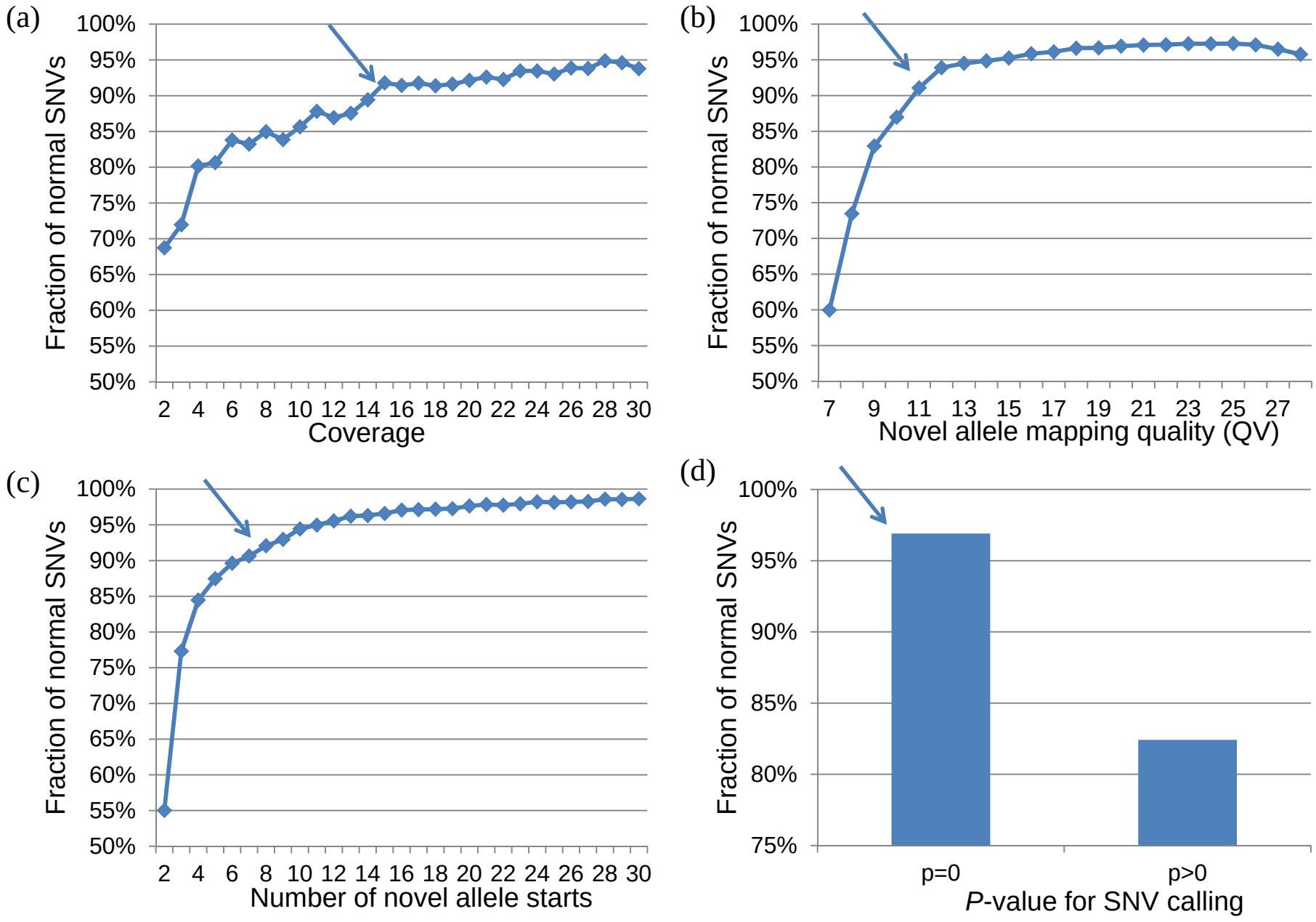
Gene	Multi-hits	CHASM	FI score	Cancer role	Description	Family
GRM8	1		high		glutamate receptor, metabotropic 8	G-protein coupled receptor
AMDHD1	1				amidohydrolase domain containing 1	enzyme
CHRM5	1				cholinergic receptor, muscarinic 5	G-protein coupled receptor
EPHA2	1				EPH receptor A2	kinase
KMO	1				kynurenine 3-monooxygenase (kynurenine 3-hydroxylase)	enzyme
NOS1AP	1				nitric oxide synthase 1 (neuronal) adaptor protein	other
PDE4DIP	1				phosphodiesterase 4D interacting protein	enzyme
ARID1A	1				AT rich interactive domain 1A (SWI-like)	transcription regulator
RPS6KC1	1				ribosomal protein S6 kinase	kinase
ACCN1		1			amiloride-sensitive cation channel 1, neuronal	other
CDK17		1			cyclin-dependent kinase 17	kinase
FCRLA		1			Fc receptor-like A	other
INHBA		1			inhibin, beta A	growth factor
LATS2		1			LATS, large tumor suppressor, homolog 2 (Drosophila)	kinase
MAP3K1		1			mitogen-activated protein kinase kinase kinase 1	kinase
PHKA2		1			phosphorylase kinase, alpha 2 (liver)	kinase
PHKG1		1			phosphorylase kinase, gamma 1 (muscle)	kinase
TTLL5		1			tubulin tyrosine ligase-like family, member 5	enzyme
WWP2		1			WW domain containing E3 ubiquitin protein ligase 2	enzyme
EP300			high		E1A binding protein p300	transcription regulator
ERBB3			high		v-erb-b2 erythroblastic leukemia viral oncogene homolog 3	kinase
FLAD1			high		FAD1 flavin adenine dinucleotide synthetase homolog	enzyme
PTK2			high		PTK2 protein tyrosine kinase 2	kinase
TRPS1			high		trichorhinophalangeal syndrome I	transcription regulator
WNT11			high		wingless-type MMTV integration site family, member 11	other
AKTIP				interacting with AKT	AKT interacting protein	other
C1orf198				novel	chromosome 1 open reading frame 198	other
IGFBP3				putative tumor suppressor	insulin-like growth factor binding protein 3	other
MFHAS1				putative oncogene	malignant fibrous histiocytoma amplified sequence 1	other
MTUS1				putative tumor suppressor	microtubule associated tumor suppressor 1	other

Supplementary Fig. 1 The analytic pipeline for detecting somatic mutations in the exomes of endometrial tumors

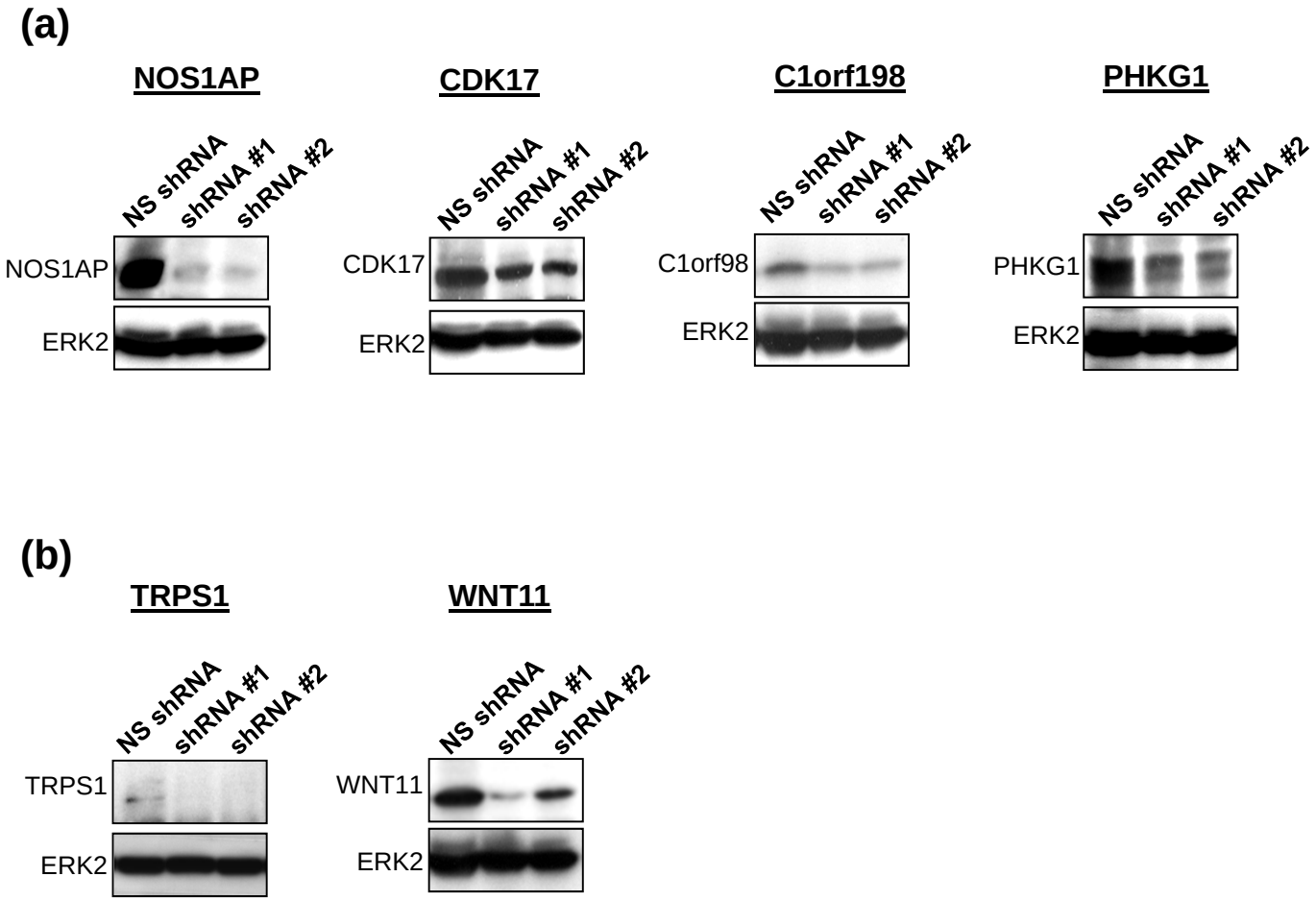


Supplementary Fig. 2 Optimization of various parameters for tumor SNV calling

Additional filters were applied to boost tumor SNV calling accuracy; and the fraction of tumor SNV positions that are called as normal SNVs as an index for optimization.



Supplementary Fig. 3 Western blots showing shRNA knock-down efficiency
The membranes were reprobed with ERK2 to confirm equal loading.

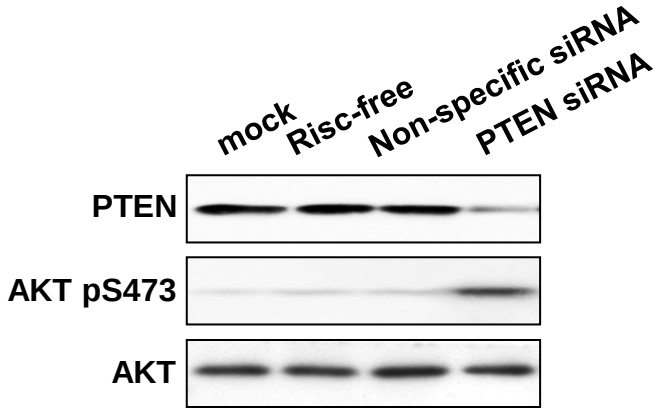


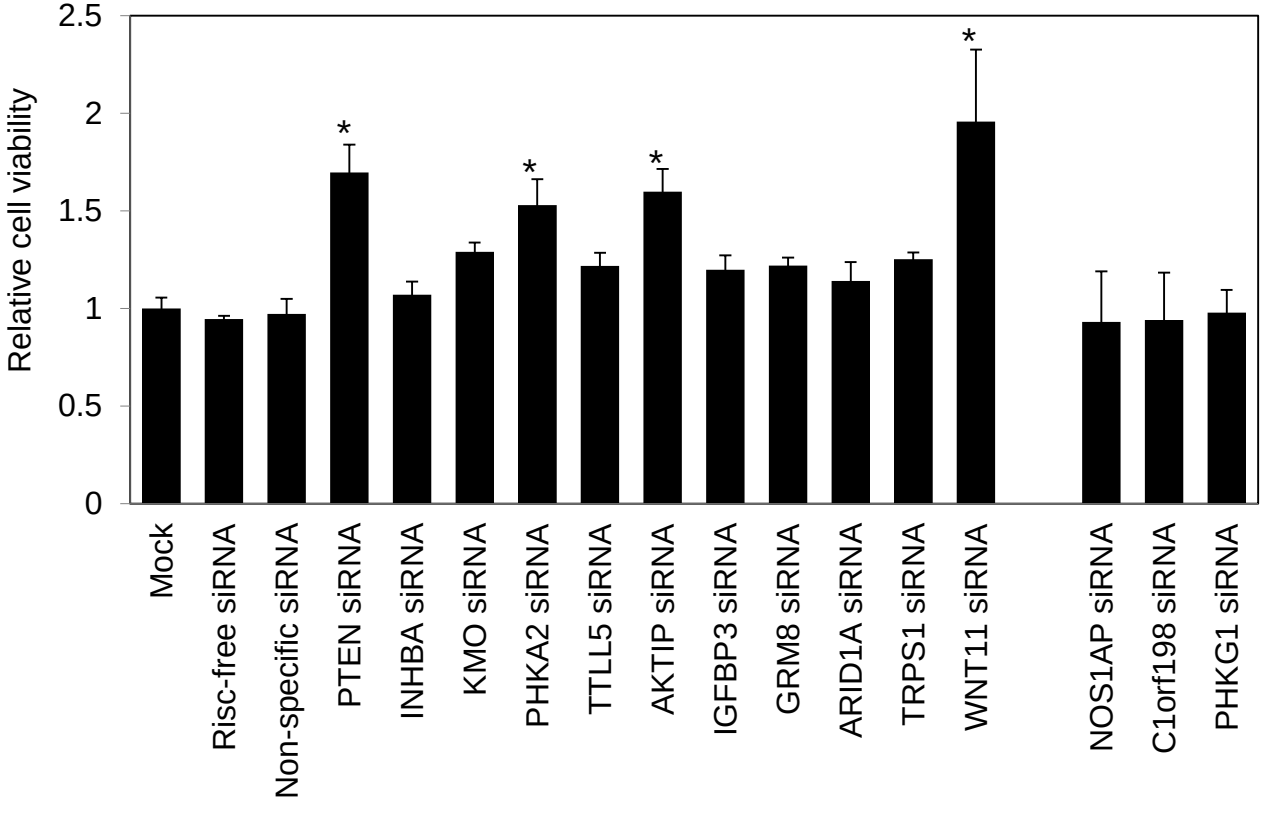
Supplementary Fig. 4 Functional effect of candidate driver cancer genes by siRNA-mediated gene silencing in three endometrial cancer cell lines

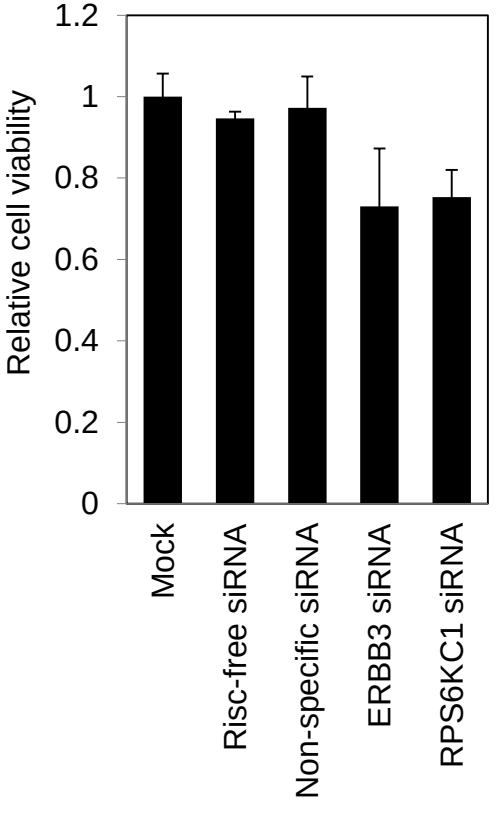
(a) EFE184 cells were transfected with siRNAs targeting the indicated genes. Mock, risc-free siRNA and non-specific shRNA served as controls. Efficacy of PTEN siRNA on AKT phosphorylation was determined by Western blotting. Cells transfected with indicated siRNAs were assayed for cell viability 5 days post-transfection. (b) Similarly, SK-UT-2 cells were transfected with the siRNAs but transfected cells were assayed for cell viability 6 days post-transfection, except the dataset for IGFBP3 which was measured 4 days after transfection. (c) Transfected SNG-II cells were harvested for cell viability assay 5 days post-transfection, except the dataset for IGFBP3 which was measured 6 days after transfection. Cell viability relative to mock transfected cells was shown. * $P < 0.05$, compared with mock control.

Supplementary Fig. 4

(a) EFE184

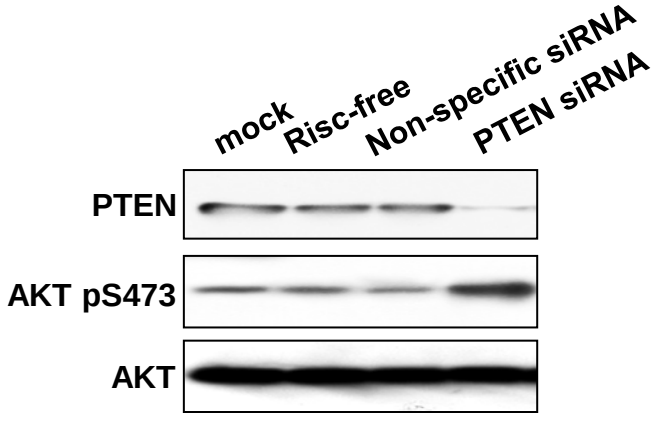


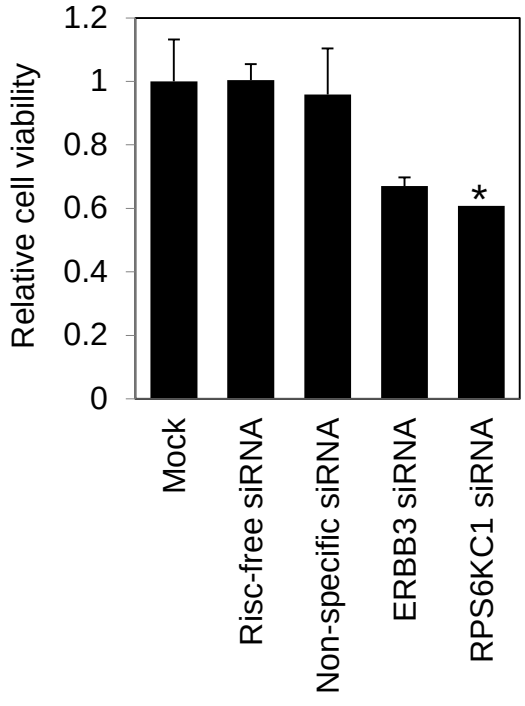


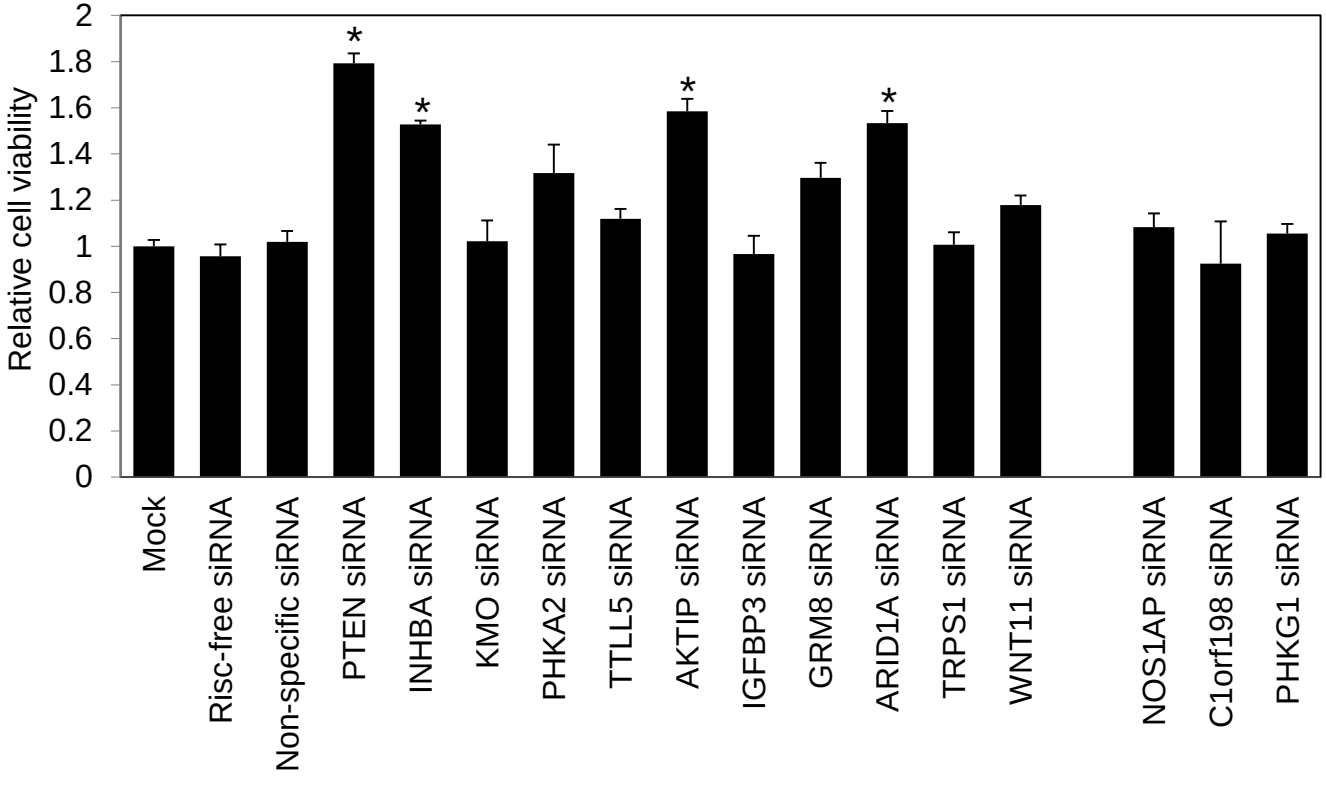


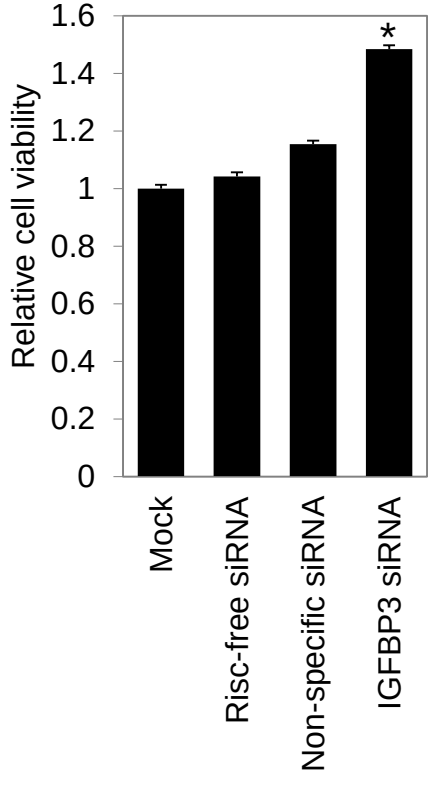
Supplementary Fig. 4

(b) SK-UT2



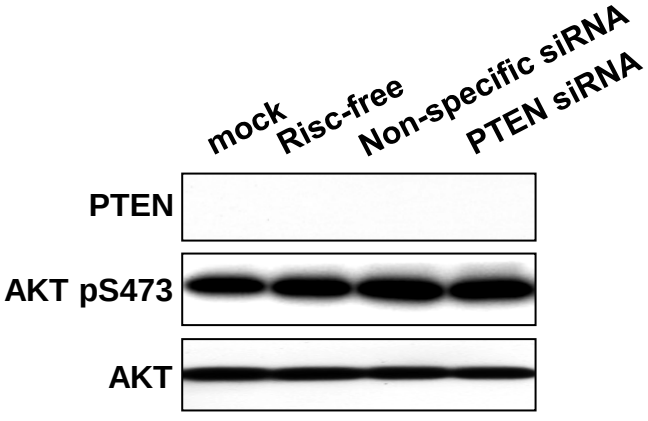


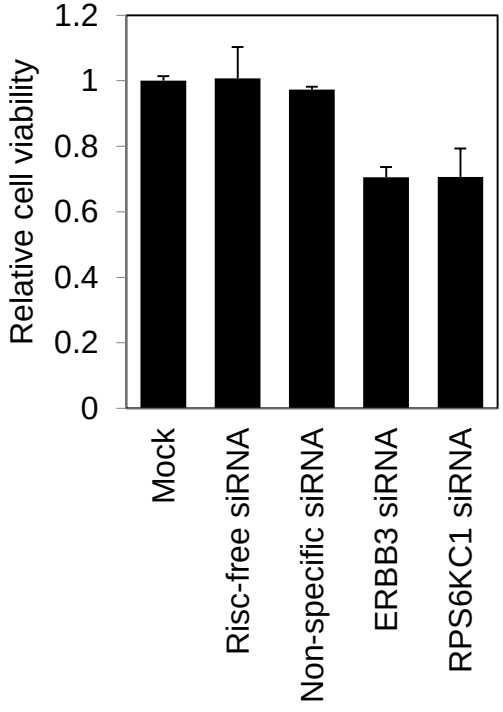


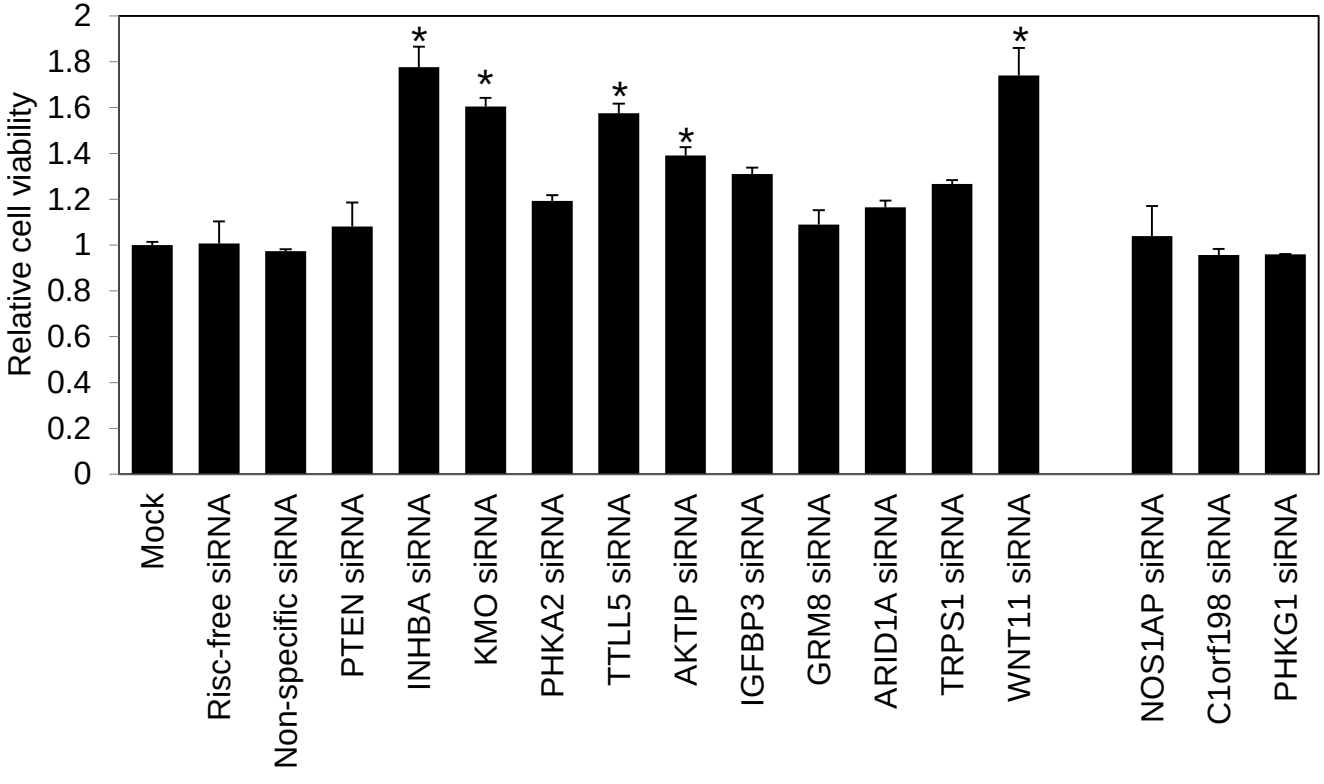


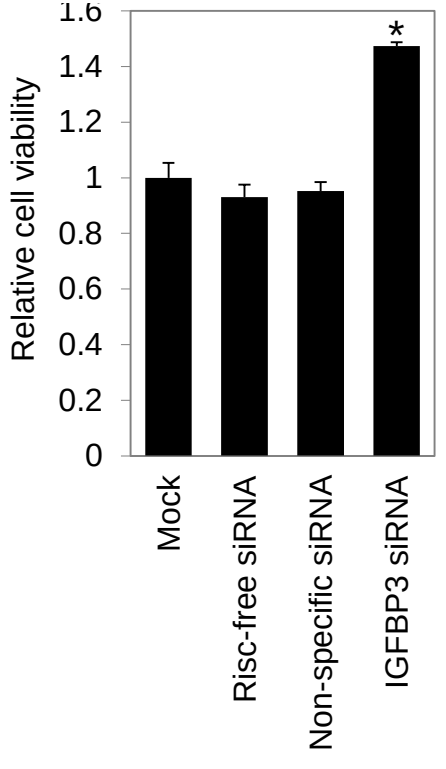
Supplementary Fig. 4

(c) SNG-II









Supplementary Fig. 5 Mutation distribution along *ARID1A* gene