SUPPLEMENTARY MATERIAL

Molecular stratification of endometrioid ovarian carcinomas predicts clinical outcome

Robert L Hollis[1,a], John P Thomson[1,a], Barbara Stanley[1,a], Michael Churchman[1], Alison M Meynert[2], Tzyvia Rye[1], Clare Bartos[1], Yasushi Iida[1,4], Ian Croy[1], Melanie Mackean[3], Fiona Nussey[3], Aikou Okamoto[4], Colin A Semple[2], Charlie Gourley[1,b] and C. Simon Herrington[1,b]

[1]Nicola Murray Centre for Ovarian Cancer Research, Cancer Research UK Edinburgh Centre, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, UK

[2]MRC Human Genetics Unit, MRC Institute of Genetics and Molecular Medicine, University of Edinburgh, UK

[3]Edinburgh Cancer Centre, Western General Hospital, Edinburgh, UK.

[4]The Jikei University School of Medicine, Tokyo, Japan

SUPPLEMENTARY MATERIALS AND METHODS

SECTION 1: IMMUNOHISTOCHEMISTRY

1A. Immunohistochemistry for WT1

Immunohistochemistry (IHC) for WT1 was performed on the Leica Bond III Autostainer using protocol F. WT1 IHC used 1:1000 dilution anti-human WT1 monoclonal mouse antibody clone 6F-H2 (DAKO).

Samples with any WT1 nuclear staining in tumour cells were recorded as WT1 positive and those with complete absence of staining as WT1 negative. Positive nuclear staining of vascular endothelial cells served as internal controls.

1B. Immunohistochemistry for CK7 and CK20

CK7 staining was performed using a 1:100 dilution of the monoclonal mouse CK7 antibody (Leica, Clone RN7, HIER1 – 20 minutes). A WT1 positive high grade serous ovarian carcinoma tissue section was used as a positive control. CK7 staining was considered positive with any positive nuclear staining of tumour cells.

CK20 staining was performed using a 1:50 dilution of the monoclonal mouse CK20 antibody (Leica, clone KS20.8, HIER1-20 minutes). Normal stomach tissue was used as a positive control. CK20 staining was considered positive with any positive nuclear staining of tumour cells.

SECTION 2: WHOLE EXOME SEQUENCING

2A. Generation of sequence libraries and Exome sequencing

Libraries were prepared from each DNA sample using the Illumina TruSeq Exome Library Prep kit (#FC-150-1002 - Illumina) according to the provided protocol using modifications for working with FFPE sourced material.

200ng of DNA was end-repaired to remove 3' and 5' overhangs, and fragment length was optimised using sample purification beads. A single 'A' nucleotide was added to the 3' ends of the blunt fragments to prevent them from ligating to another during the subsequent adapter ligation reaction, and a corresponding single 'T' nucleotide on the 3' end of the adapter provided a complementary overhang for ligating the adapter to the fragment. Multiple indexing adapters were then ligated to the ends of the ds cDNA to prepare them for hybridisation onto a flow cell, before 12 cycles of PCR were used to selectively enrich those DNA fragments that had adapter molecules on both ends and amplify the amount of DNA in the library suitable for sequencing. Libraries were quantified using the Qubit 2.0 Fluorometer and the Qubit DNA HS assay (#Q32854 - ThermoFisher) and the size distribution of fragments was assessed using the Agilent Bioanalyser with the DNA HS Kit (#5067-4626 - Agilent).

DNA libraries containing unique indexes were combined in pools of 6, and then target regions of the DNA were bound with capture probes. Streptavidin Magnetic Beads were then used to capture probes hybridised to the targeted regions of interest and a series of washes removed nonspecific binding from the beads. This process was repeated to ensure high specificity of the captured regions. Captured enriched library was then purified before 8 cycles of PCR amplification and a final purification step to remove unwanted products.

Exome-captured sequencing library pools were quantified using the Qubit 2.0 Fluorometer and the Qubit DNA HS assay (#Q32854 - ThermoFisher) and the size distribution of fragments was assessed using the Agilent Bioanalyser with the DNA HS Kit (#5067-4626 - Agilent). Fragment size and quantity measurements were used to calculate molarity for each library pool.

Sequencing was performed using the NextSeq 500/550 High-Output v2 (150 cycle) Kit (# FC-404-2002) on the NextSeq 550 platform (Illumina Inc, #SY-415-1002).

2C. Mapping of sequenced reads

Base calling and quality scoring was conducted using the tool FASTQC. Data was then processed with a python toolkit providing pipelines for fully automated high throughput sequencing analysis (bcbio-nextgen - see https://github.com/bcbio/bcbio-nextgen for full documentation and informatic

pipelines). Raw sequence data was mapped to the hg38 genome build (FASTQ files) using the Burrows–Wheeler alignment algorithm 0.7.17 [1].

2C. Variant calling

Somatic variant calling was carried out on mapped BAM files using a majority vote from three variant caller algorithms; VarDict [2] (REF), mutect2 [3] (REF), freebayes [4] (REF) .Filtering for C>T (FFPE artifacts) and G>T (oxidation artifacts) was applied using GATK (CollectSequencingArtifactMetrics and FilterByOrientationBias). Resulting VCF files were then analysed in R using the maftools package (https://bioconductor.org/packages/release/bioc/html/maftools.html. Datasets were also filtered to remove common population variants by comparing to the 1000 Genomes reference datasets (1000 genomes phase 1 snp and indel dataset; http://www.internationalgenome.org/) and the Exome Aggregation consortium (EXAC) reference datasets (ExAC.0.3.GRCh38 : http://exac.broadinstitute.org/)).

Variants which are not predicted to result in causal mutations were filtered using the Polymorphism Phenotyping (PolyPhen) [5] and Sorting Intolerant from Tolerant (SIFT) [6] prediction tools as well as being cross referenced to the NCBI ClinVar database [7] which aggregates pathogenicity reports associated with genomic variants.

Filtering was applied to define high impact mutations where the variant allele frequency of a given mutation was > 10% across regions with a minimum read coverage of 20X.

Microsatellite instability scores were assigned based on the number of INDELS detected in a given sample, with data taken from the VCF files. Transitions and transversions were calculated using the titv function in maftools. This function classifies SNPs into Transitions and transversions and returns a list of summarized tables in various ways. Summarized data was visualised as a boxplot showing overall distribution of six different conversions and as a stacked barplot showing fraction of conversions in each sample. For more information on the R package maftools see the notes at https://bioconductor.org/packages/release/bioc/vignettes/maftools/inst/doc/maftools.html

SECTION 2: ONCOGENIC PATHWAY ANALYSIS

In order to determine the major oncogenic pathways alerted across the tumour samples we used the OncogenicPathways function in the R package maftools. This highlighted PIK-AKT, WNT, RAS and NOTCH pathways as major altered networks. We then visualised mutations across individual members of these pathways using the PlotIncogenicPathways function in maftools which draws an oncoplot for each sample in a given pathway (supplemental figure S7). To generate figure 4A we collapsed all samples containing at least 1 mutation in a pathway instead of showing all genes.

SECTION 4: TUMOUR COMPLEXITY SCORING

Variant allele frequencies (VAF) densities across all genes were plotted for each sample to assess genomic complexity (supplementary materials section 3); low complexity specimens, with a single driver event and associated outgrowth, were anticipated to display a single VAF peak. Conversely, highly complex tumours with multiple driver events, branched evolution and cell population expansion, would demonstrate multiple VAF peaks. Analysis was carried out using the inferHeterogeneity function in the R package maftools [8, 9] (supplementary materials section 3). Resulting MATH scores represent the width of the VAF distribution. In previous studies higher MATH scores are found to be associated with poor outcome [10]. See https://bioconductor.org/packages/release/bioc/vignettes/maftools/inst/doc/maftools.html for more information on the R package maftools.

SECTION 5: COPY NUMBER ANALYSIS

See https://github.com/wwcrc/geneCN for full documentation and informatic pipelines

SUPPLEMENTAL FIGURES

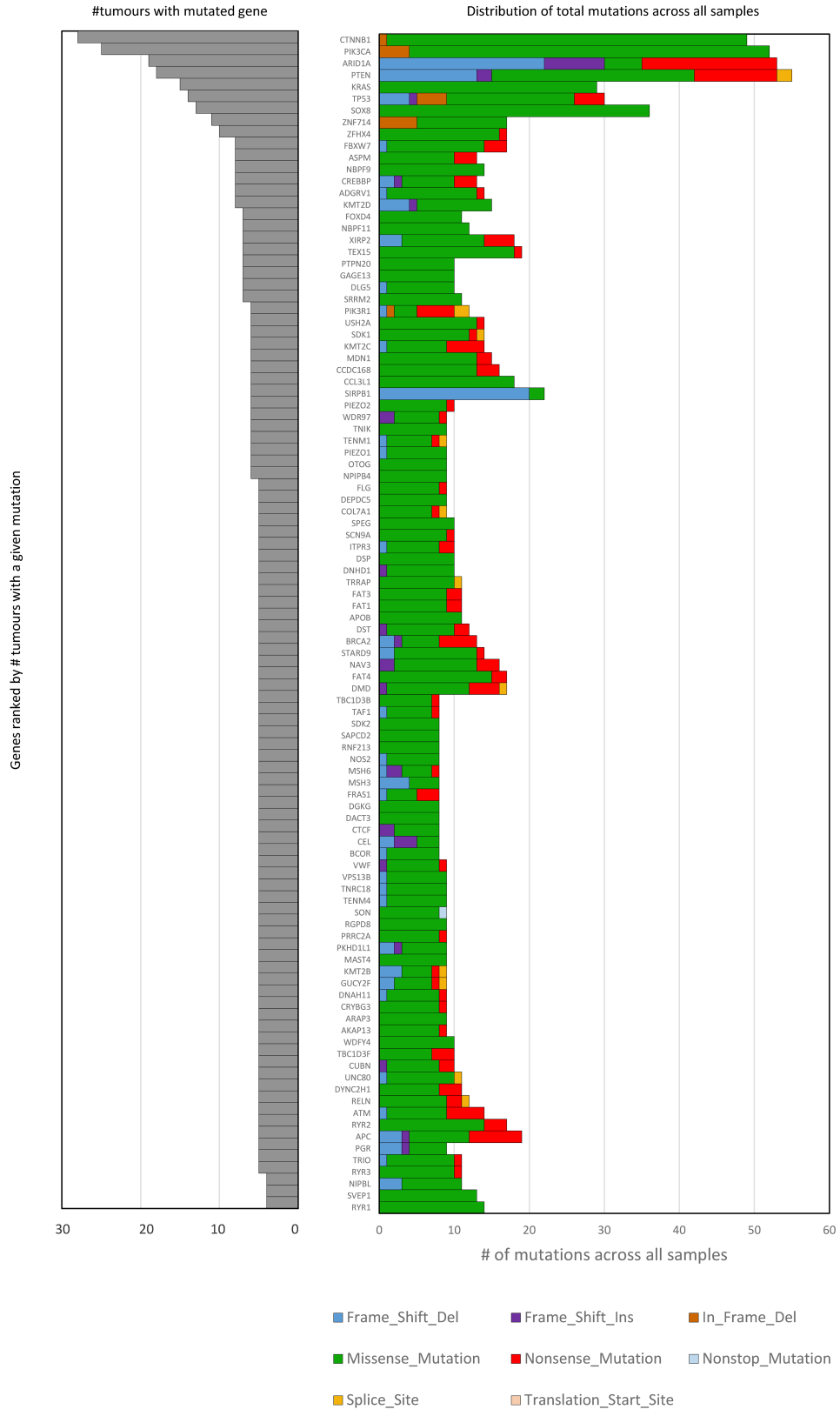

Figure S1. Plot of the 100 most frequently mutated genes from SNV analysis. Left: Bar plot of the number of tumours containing each gene mutation. Right: stacked bar plot showing the distribution of mutation classes for each gene. Data ranked by tumour mutation count.

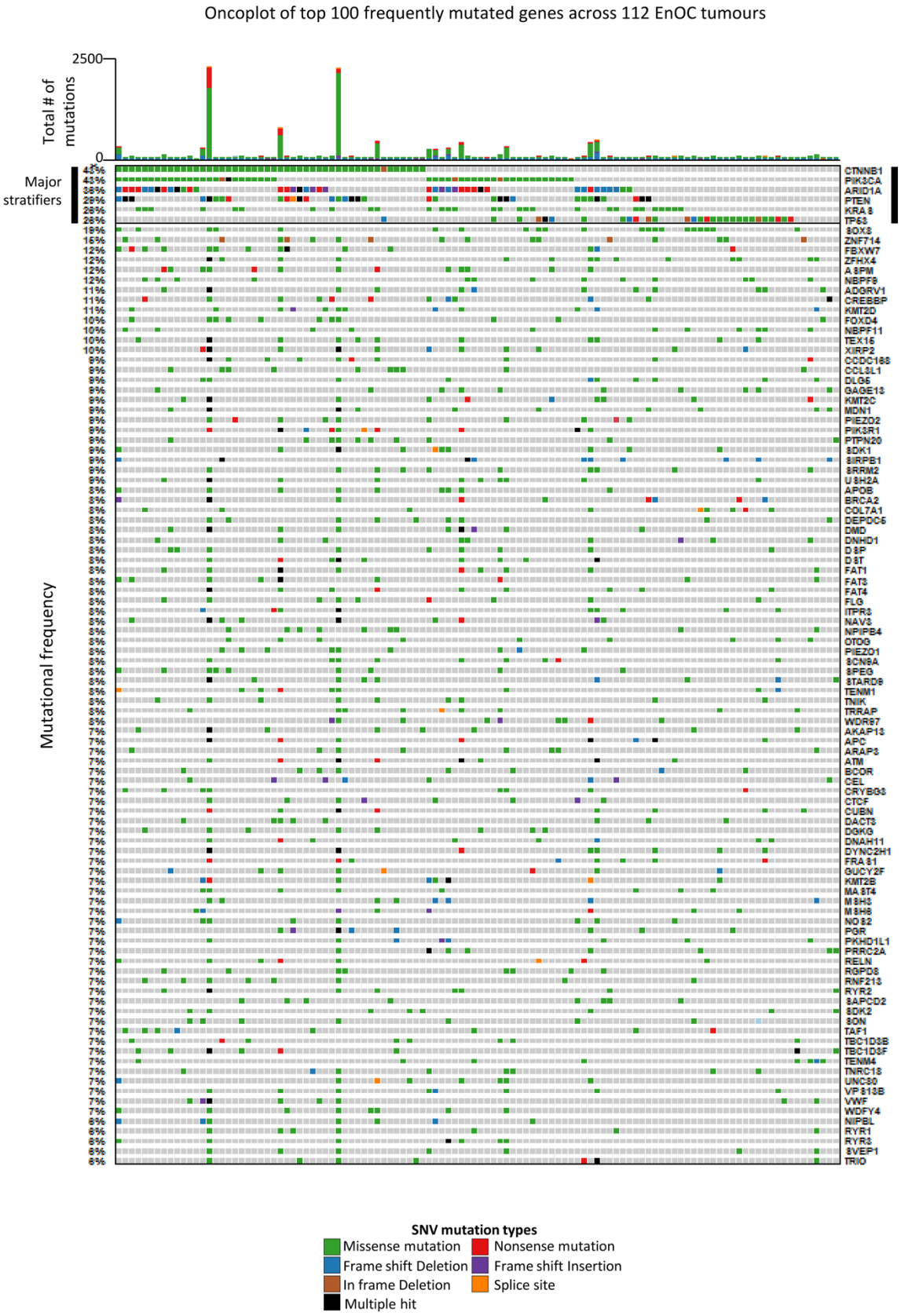

Figure S2. Oncoplot for the 100 most frequently mutated genes from SNV analysis. Total number of mutations per tumour are plotted above. Colour code defines mutation type. Grey denotes no mutation. Percentages on the left indicate % of samples with a given mutation.

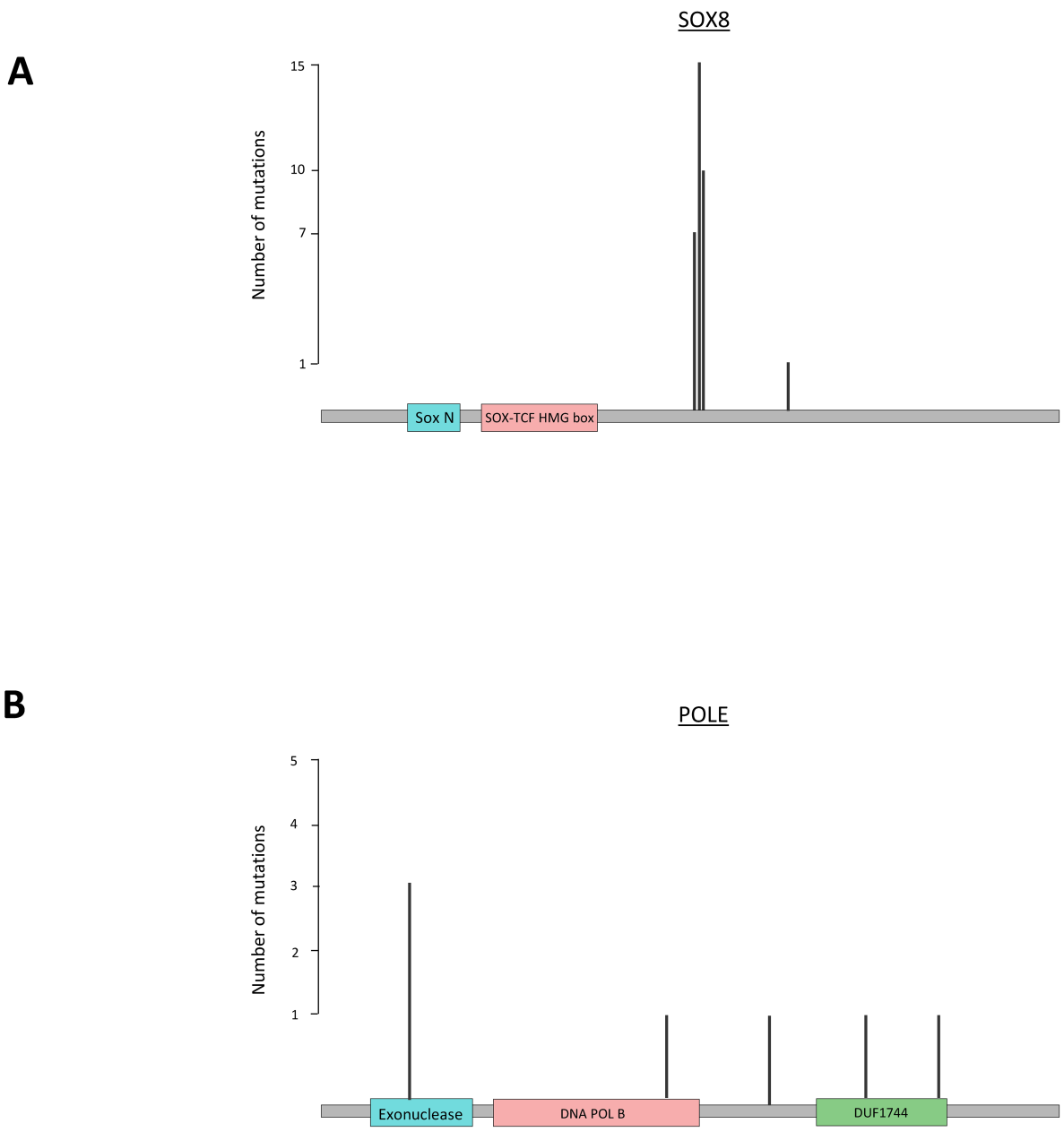

Figure S3. Lollipop plot of location of variants within the *SOX8* (**A**) and POLE (**B**) genes, with known protein coding domains highlighted.

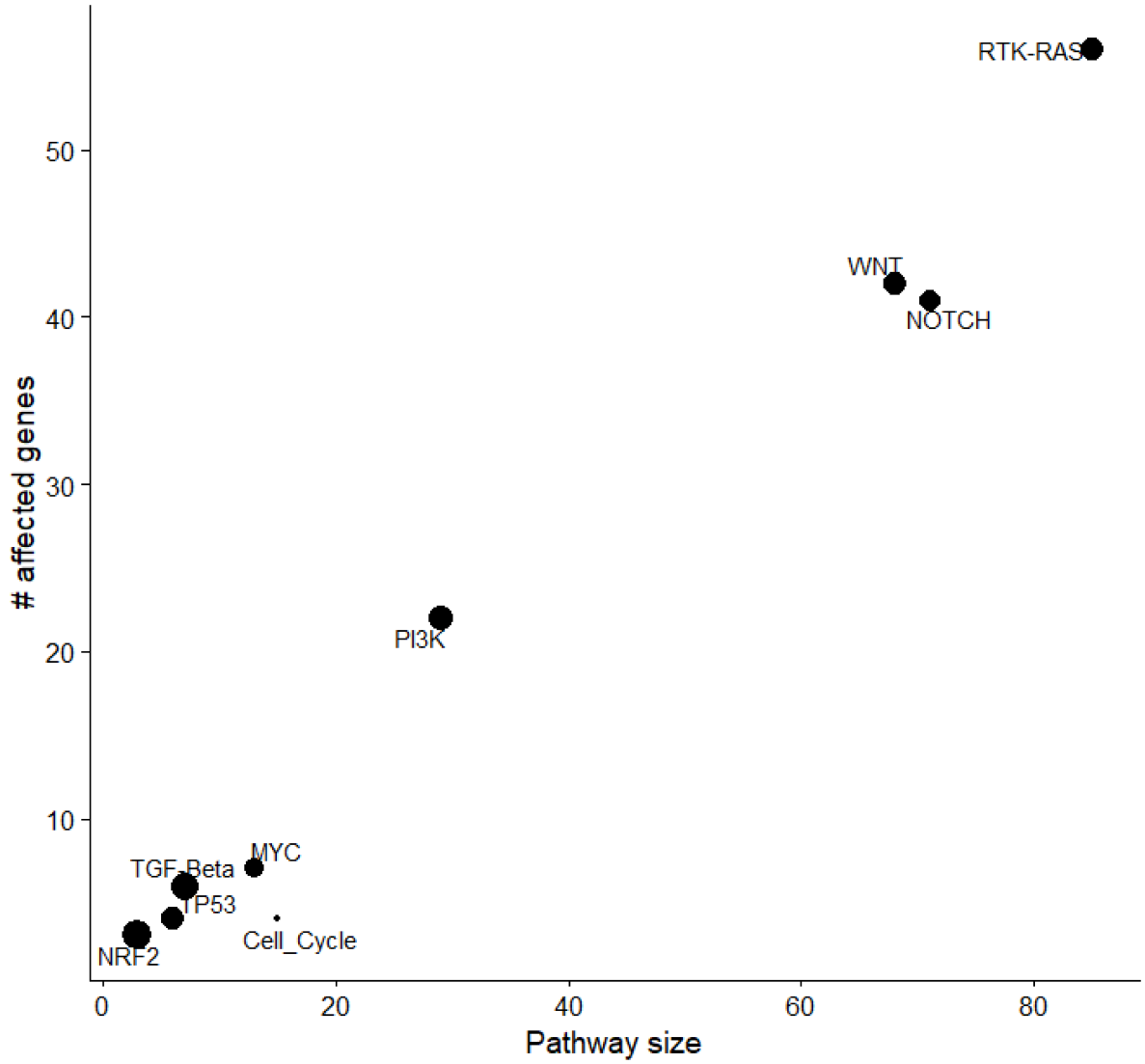

Figure S4. Scatter plot of the number of genes altered in oncogenic pathways vs total pathway size.

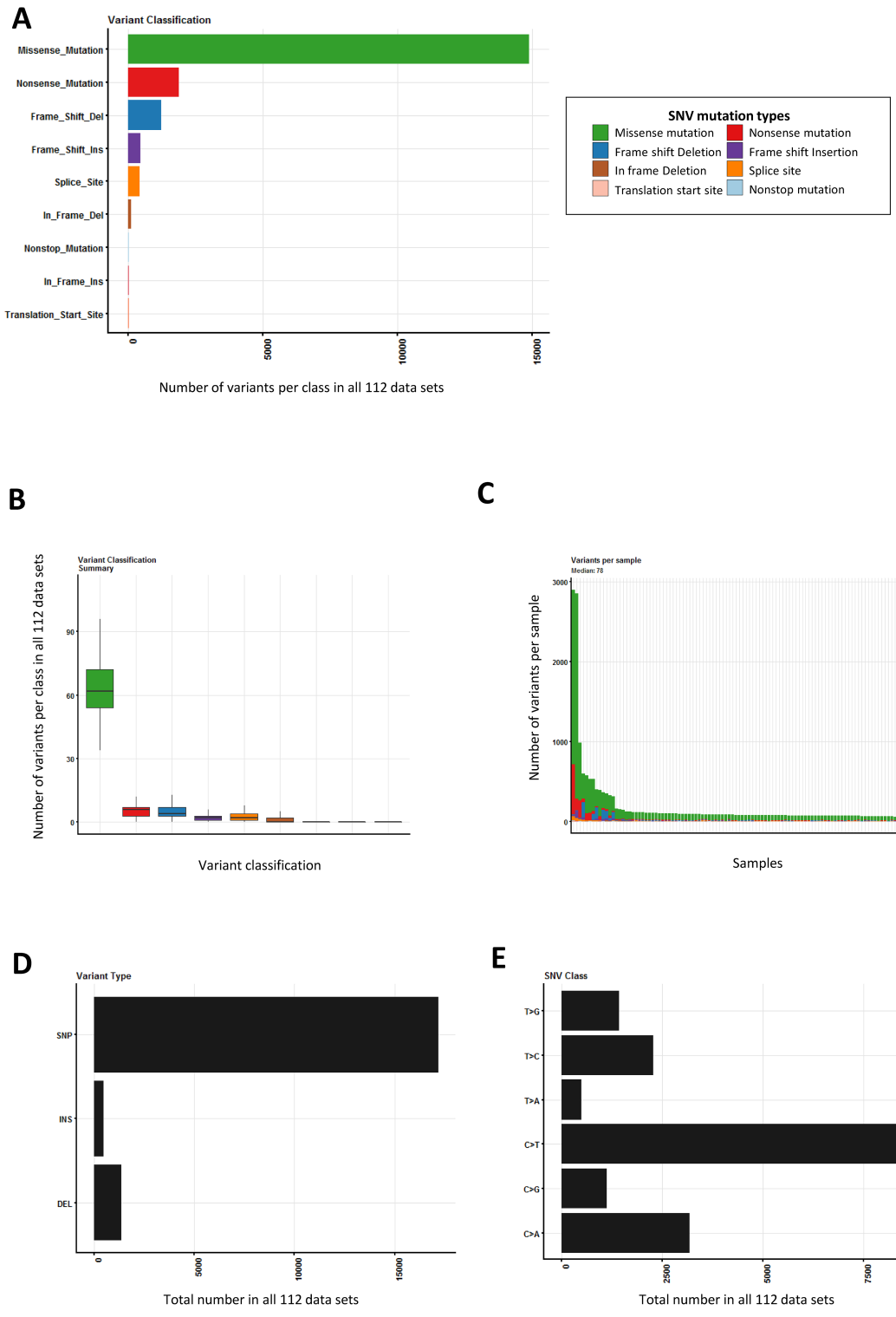

Figure S5. Whole exome variant call summary statistics. S5A-C share the colour key for variant type (box, top right) **A.** Plot of the total number of variant types across all of the 112 samples. **B**. Box plot of the number of variants per sample for each of the classifications. **C.** stacked plot of total variant count per sample containing variant type information. **D.** Plot of number of single nucleotide polymorphisms (SNP), insertions (INS) and deletions (DEL) present in all of the 112 tumours. **E.** Summary of the single nucleotide variant (SNV) base changes across the 112 tumours.

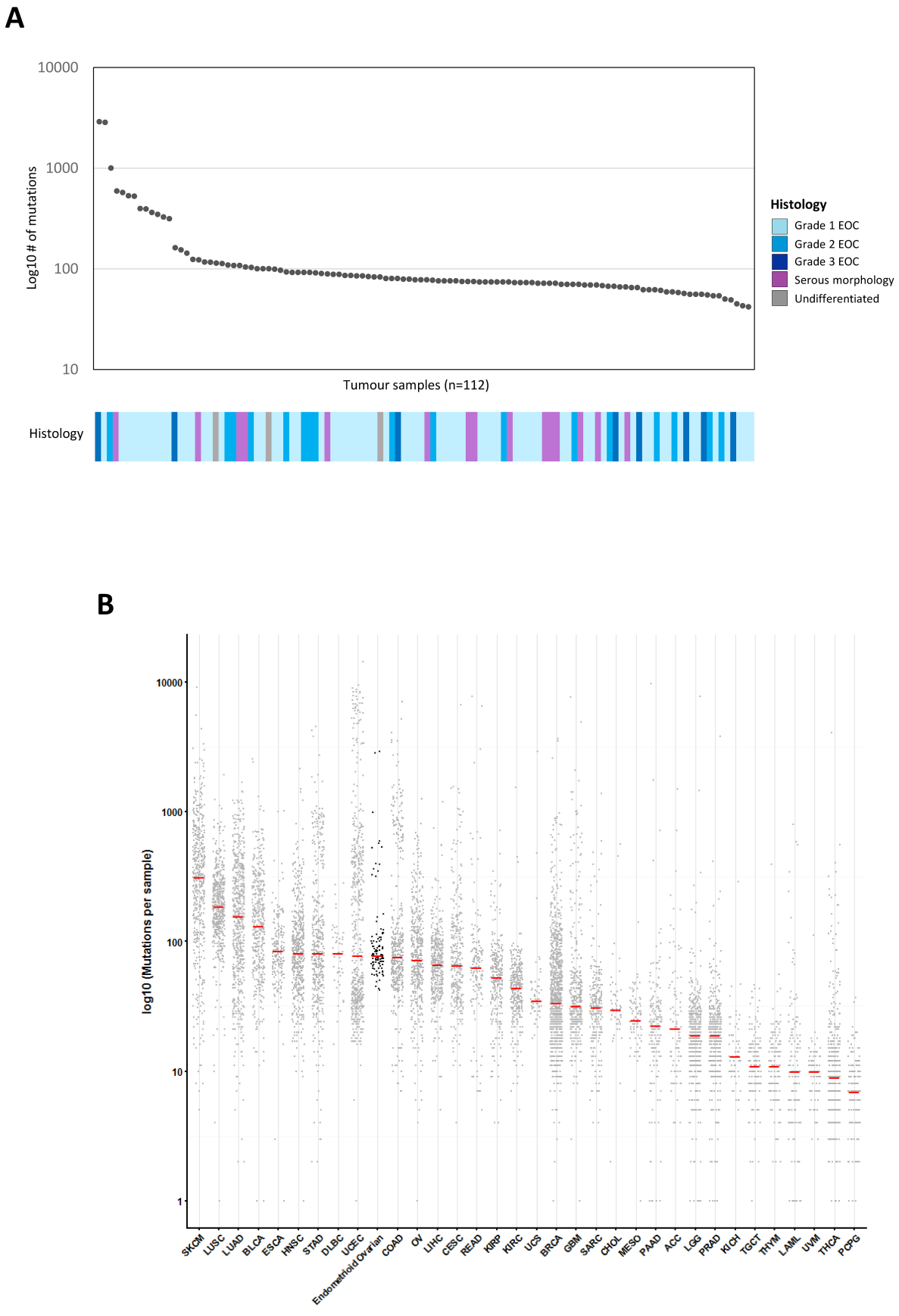

Figure S6. A. Summary of total SNV counts across all 112 tumours, ranked by number **B**. Plot of mutational load in the endometroid carcinoma samples in this study against 33 TCGA landmark cohort

datasets. Individual dots represent each sample in a given study (grey = TCGA, black = our study). Red bar denotes median mutation count. SKCM: Skin Cutaneous Melanoma, LUSC: Lung squamous cell carcinoma, LUAD: Lung adenocarcinoma, BLCA: Bladder Urothelial Carcinoma, ESCA: Esophageal carcinoma, HNSC: Head and Neck squamous cell carcinoma, STAD: Stomach adenocarcinoma, DLBC: Lymphoid Neoplasm Diffuse Large B-cell Lymphoma, UCEC: Uterine Corpus Endometrial Carcinoma, COAD: Colon adenocarcinoma, OV: Ovarian serous cystadenocarcinoma, LIHC: Liver hepatocellular carcinoma, CESC: Cervical squamous cell carcinoma and endocervical adenocarcinoma, READ: Rectum adenocarcinoma, KIRP: Kidney renal papillary cell carcinoma, KIRC: Kidney renal clear cell carcinoma, UCS: Uterine Carcinosarcoma, BRCA: Breast invasive carcinoma, GBM: Glioblastoma multiforme, SARC: Sarcoma, CHOL: Cholangiocarcinoma, MESO: Mesothelioma, PAAD: Pancreatic adenocarcinoma, ACC: Adrenocortical carcinoma, LGG: Brain Lower Grade Glioma, PRAD: Prostate adenocarcinoma, KICH: Kidney Chromophobe, TGCT: Testicular Germ Cell Tumors, THYM: Thymoma, LAML : Acute Myeloid Leukemia, UVM: Uveal Melanoma, THCA: Thyroid carcinoma, PCPG: Pheochromocytoma and Paraganglioma .

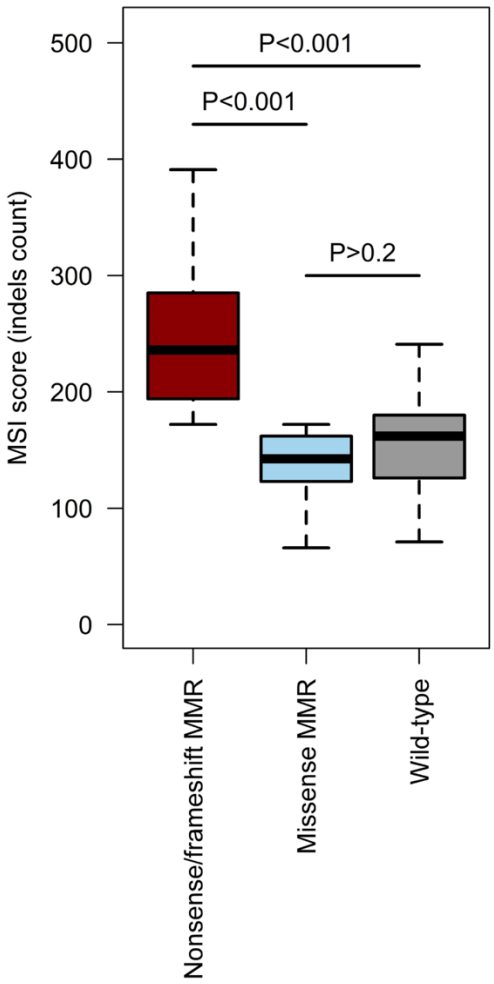

Figure S7. Box plot of MSI score (number of InDels in a given tumour) split by MMR mutation. Grey: MMR gene wt, orange: MMR gene missense, blue: MMR gene nonsense/frameshift

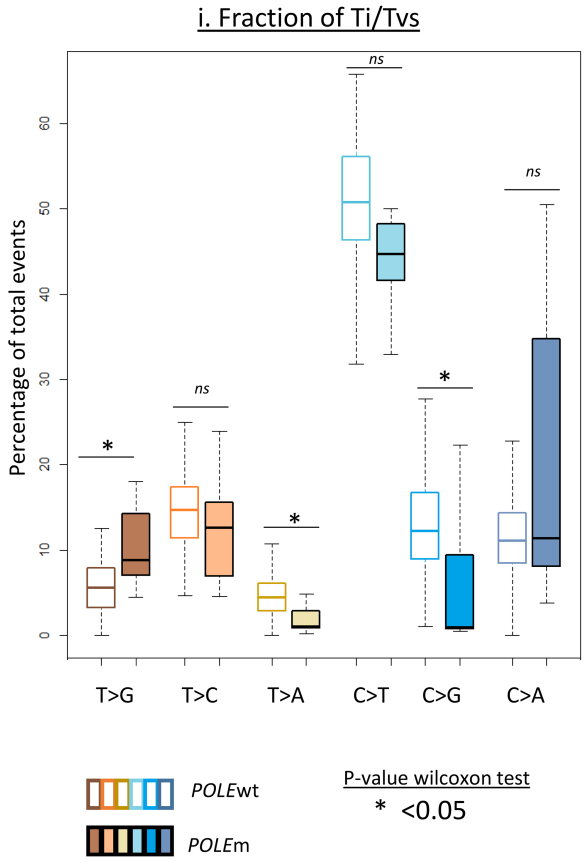

Figure S8 Boxplots of Ti/Tv fractions between *POLE*wt and *POLE*m tumours.

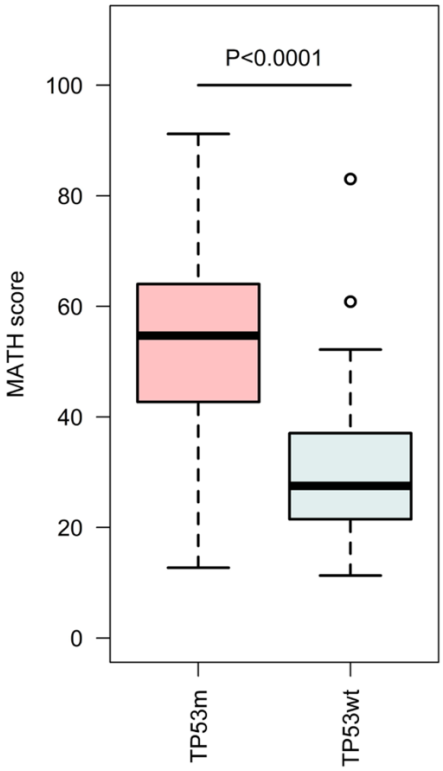

Figure S9. Boxplot displaying genomic complexity in *TP53*m and *TP53*wt tumours.

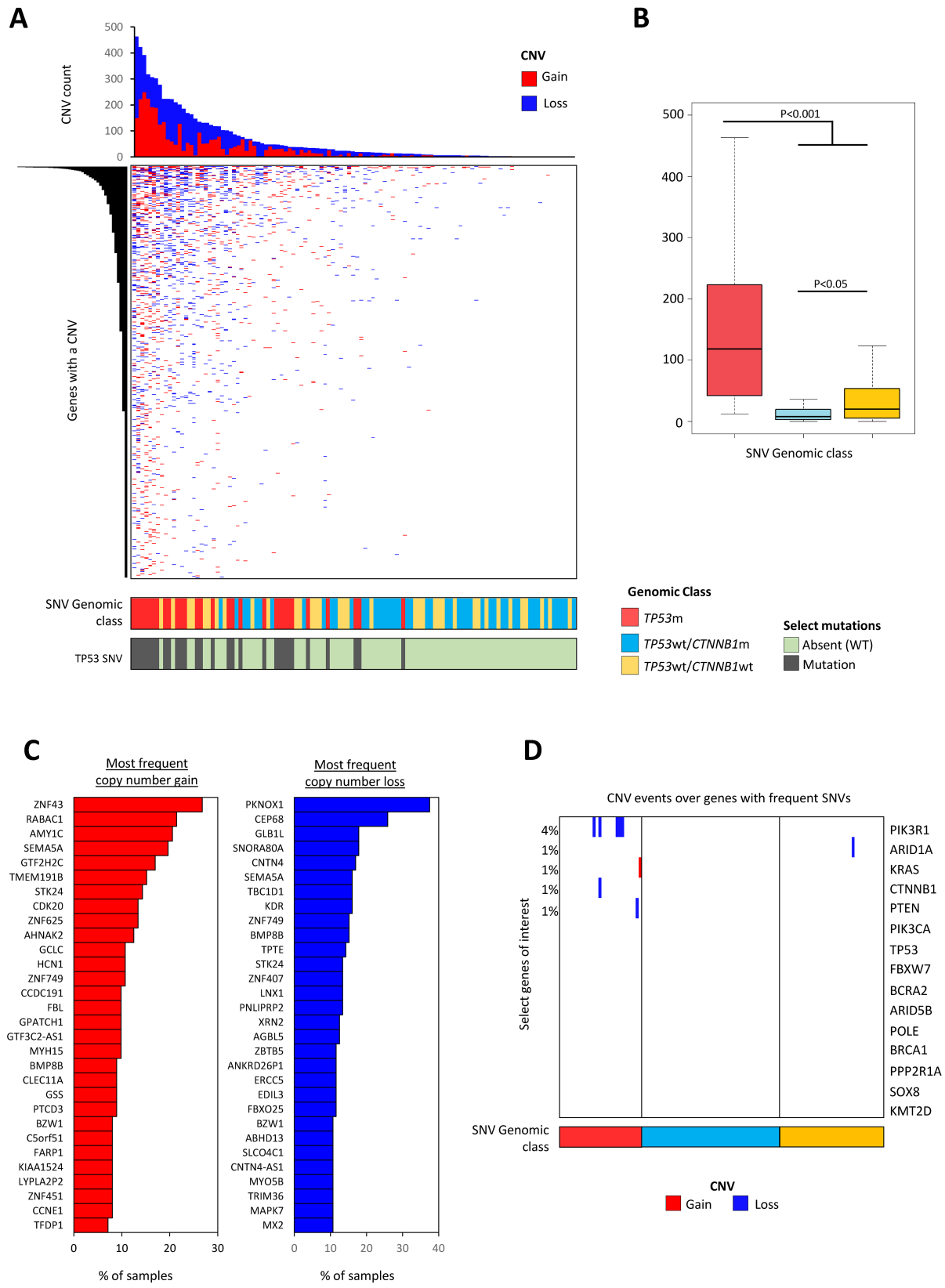

Figure S10. Analysis of single copy number alterations (CNVs) reveals distinct landscape of alterations in *TP53* mutant tumours. **A.** Plot of copy number changes across genes containing at least one CNV, displaying gain (red) or loss (blue). Samples ranked by total copy number count (histogram at the top).

B. Boxplot of number of CNV events in each of the SNV defined genomic classes. **C.** Plot of the top 30 most frequently altered genes as defined through copy number change for gain (red) or loss (blue). **D.** Plot of copy number changes over the 15 genes highlighted from SNV mutational analysis in Figure 3A, displaying gain (red) or loss (blue). Samples ordered by SNV defined genomic class and genes by frequency of CNV event. Percentages on the left indicate % of total samples with a given copy number change.

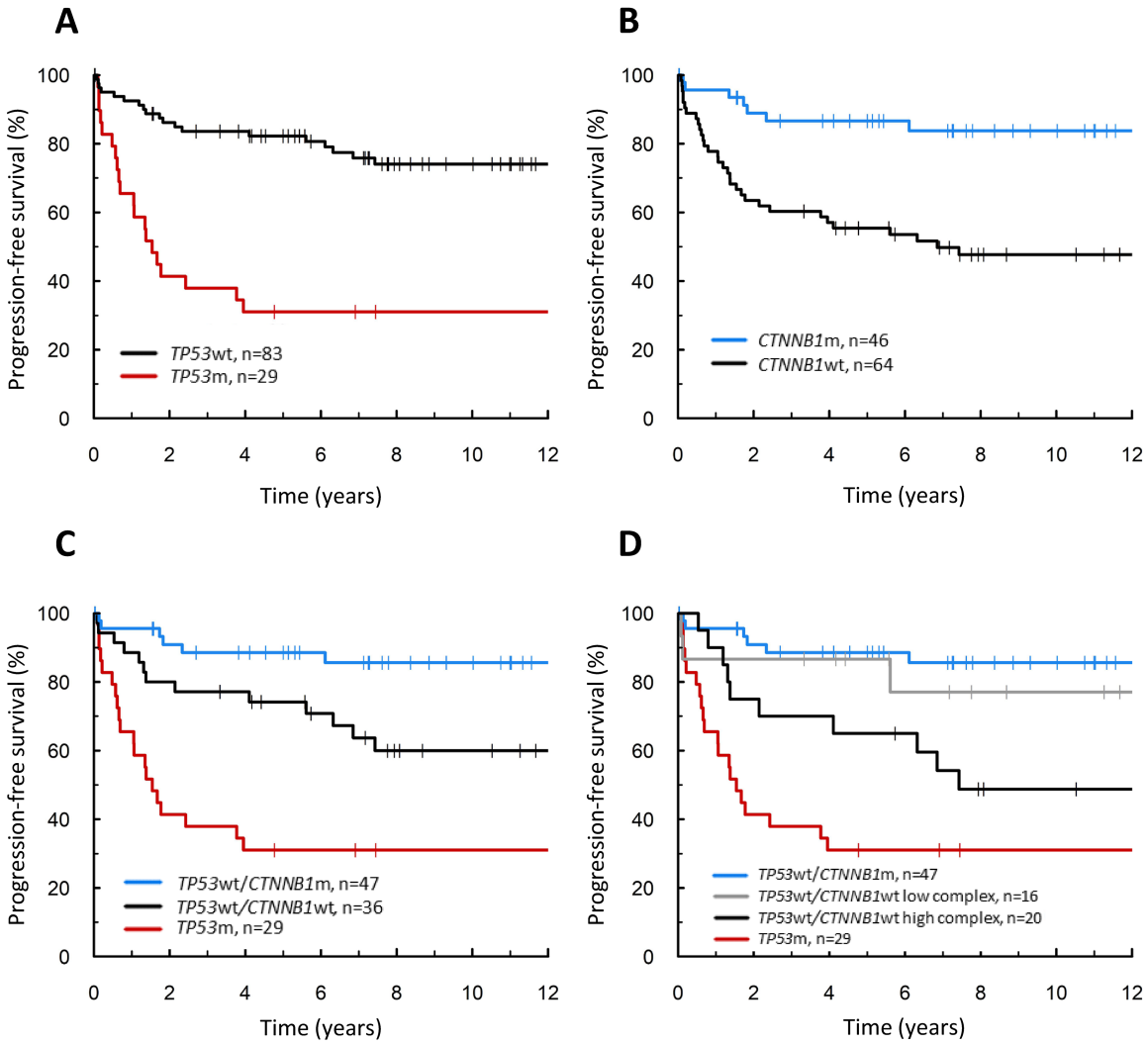

Figure S11. Progression-free survival of molecularly-defined EnOC subtypes. (A) By *TP53*m, (B) by *CTNNB1*m, (C) by *TP53*m and *CTNNB1*m; (D) using the PRTISTINE algorithm.

SUPPLEMENTAL TABLES

Table S1. *TP53* mutation status vs discrete variant allele frequency peak counts

	*TP53*m		*TP53*wt	
VAF peaks	n	%	n	%
1	4	13.8	45	54.2
2	17	58.6	34	41.0
3	8	27.6	4	4.8
total	29		83	
Chisq P<0.001; 1 peaks vs >1 peak				

M, mutant; wt, wild-type; VAF, variant allele frequency

Table S2. Univariable analysis of survival

			HR	Lower 95% CI	Upper 95% CI	P	P adj.
TP53	DSS	mutant	4.43	2.27	8.64	<0.001	<0.001
		wild-type	-	-	-	-	-
	PFS	mutant	4.46	2.37	8.42	<0.001	<0.001
		wild-type	-	-	-	-	-
CTNNB1	DSS	mutant	0.23	0.10	0.56	0.001	0.010
		wild-type	-	-	-	-	-
	PFS	mutant	0.24	0.11	0.55	<0.001	0.005
		wild-type	-	-	-	-	-
PIK3CA	DSS	mutant	0.76	0.38	1.51	0.439	1.00
		wild-type	-	-	-	-	-
	PFS	mutant	0.67	0.34	1.3	0.230	1.00
		wild-type	-	-	-	-	-
ARID1A	DSS	mutant	0.48	0.22	1.06	0.069	0.552
		wild-type	-	-	-	-	-
	PFS	mutant	0.72	0.36	1.42	0.341	1.00
		wild-type	-	-	-	-	-
PTEN	DSS	mutant	0.48	0.2	1.15	0.098	0.7832
		wild-type	-	-	-	-	-
	PFS	mutant	0.59	0.27	1.28	0.178	1.00
		wild-type	-	-	-	-	-
KRAS	DSS	mutant	0.48	0.2	1.15	0.099	0.794
		wild-type	-	-	-	-	-
	PFS	mutant	0.44	0.18	1.04	0.062	0.495
		wild-type	-	-	-	-	-
MMR	DSS	mutant	0.43	0.13	1.42	0.167	1.00
		wild-type	-	-	-	-	-
	PFS	mutant	0.65	0.25	1.65	0.362	1.00
		wild-type	-	-	-	-	-
POLE	DSS	mutant	0.38	0.05	2.76	0.337	1.00
		wild-type	-	-	-	-	-
	PFS	mutant	0.73	0.18	3.03	0.663	1.00
		wild-type	-	-	-	-	-

DSS, disease-specific survival. PFS, progression-free survival; HR, hazard ratio; CI confidence interval

Table S3. Clinicopathological features of EnOC subtypes defined by the PRISTINE algorithm.

	*TP53*m		*TP53*wt/*CTNNB1*m		*TP53*wt/*CTNNB1*m low complex		*TP53*wt/*CTNNB1*m high complex	
	n	%	n	%	n	%	n	%
Cases	29		47		16		20	
Concurrent endometrial ca.	1	3.4	9	19.1	4	25.0	5	25.0
Age median	61	32-79	57	28-88	57	37-75	57	37-75
FIGO stage								
I	8	27.6	24	53.3	6	37.5	9	45.0
II	7	24.1	17	37.8	6	37.5	9	45.0
III	8	27.6	3	6.7	3	18.8	1	5.0
IV	6	20.7	1	2.2	1	6.3	1	5.0
NA	0		2		0		0	
RD following debulking								
Zero macroscopic RD	15	55.6	41	91.1	10	71.4	16	84.2
Macroscopic RD	12	44.4	4	8.9	4	28.6	3	15.8
NA	2		2		2		1	

RD, residual disease; m, mutant; wt, wild-type; NA, not available

Table S4. Multivariable disease-specific survival analysis by *TP53* mutation status

DSS		mHR	mHR low CI	mHR high CI	P
TP53	*TP53*m	2.62	1.09	6.25	0.031
	*TP53*wt	-	-	-	-
FIGO stage at diagnosis	I/II	0.2	0.08	0.5	<0.001
	III/IV	-	-	-	-
RD following debulking	Zero macroscopic RD	0.21	0.08	0.54	0.001
	Macroscopic RD	-	-	-	-
Diagnosis period	1980s	-	-	-	-
	1990s	0.66	0.25	1.74	0.401
	2000s	0.37	0.13	1.1	0.074
	2010s	0.49	0.1	2.55	0.399
Age	years	1.02	0.98	1.05	0.369

DSS, disease-specific survival; mHR, multivariable hazard ratio; m, mutant; wt, wild-type; RD, residual disease

Table S5. Multivariable progression-free survival analysis by *TP53* mutation status

PFS		mHR	mHR low CI	mHR high CI	P
TP53	*TP53*m	2.84	1.27	6.31	0.011
	*TP53*wt	-	-	-	-
FIGO stage at diagnosis	I/II	0.18	0.07	0.47	<0.001
	III/IV	-	-	-	-
RD following debulking	Zero macroscopic RD	0.3	0.12	0.75	0.01
	Macroscopic RD	-	-	-	-
Diagnosis period	1980s	-	-	-	-
	1990s	0.57	0.22	1.48	0.248
	2000s	0.32	0.11	0.88	0.028
	2010s	0.55	0.15	1.99	0.36
Age	years	1.02	0.98	1.05	0.327

PFS, progression-free survival; mHR, multivariable hazard ratio; m, mutant; wt, wild-type; RD, residual disease

Table S6. Multivariable DSS analysis of *CTNNB1*m status

DSS		mHR	mHR low CI	mHR high CI	P
CTNNB1	*CTNNB1*m	0.31	0.12	0.81	0.017
	*CTNNB1*wt	-	-	-	-
FIGO stage at diagnosis	I/II	0.12	0.05	0.33	<0.001
	III/IV	-	-	-	-
RD following debulking	Zero macroscopic RD	0.32	0.12	0.86	0.023
	Macroscopic RD	-	-	-	-
Diagnosis period	1980s	-	-	-	-
	1990s	0.65	0.26	1.66	0.37
	2000s	0.29	0.1	0.89	0.03
	2010s	0.48	0.1	2.47	0.383
Age	years	1.02	0.99	1.06	0.154

DSS, disease-specific survival; mHR, multivariable hazard ratio; m, mutant; wt, wild-type; RD, residual disease

Table S7. Multivariable PFS analysis of *CTNNB1*m status

PFS		mHR	mHR low CI	mHR high CI	P
CTNNB1	*CTNNB1*m	0.29	0.12	0.69	0.006
	*CTNNB1*wt	-	-	-	-
FIGO stage at diagnosis	I/II	0.11	0.04	0.29	<0.001
	III/IV	-	-	-	-
RD following debulking	Zero macroscopic RD	0.43	0.17	1.11	0.08
	Macroscopic RD	-	-	-	-
Diagnosis period	1980s	-	-	-	-
	1990s	0.58	0.23	1.43	0.234
	2000s	0.25	0.09	0.72	0.01
	2010s	0.57	0.16	2.04	0.391
Age	years	1.02	0.99	1.06	0.153

PFS, progression-free survival; mHR, multivariable hazard ratio; m, mutant; wt, wild-type; RD, residual disease

Table S8. Impact of genomic complexity on survival outcome

			mHR	mHR low CI	mHR high CI	mHR P
VAF peaks	DSS	1 peak	-	-	-	-
		2+ peaks	2.45	1.15	5.21	0.02
	PFS	1 peak	-	-	-	-
		2+ peaks	2.28	1.14	4.56	0.0202
MATH score	DSS	score	1.03	1.02	1.05	<0.0001
	PFS	score	1.03	1.02	1.05	<0.0001

mHR, multivariable hazard ratio; VAF, variant allele frequency; MATH, mutant allele tumour heterogeneity

Supplementary references

[1] Li H, Durbin R. Fast and accurate long-read alignment with Burrows-Wheeler transform. Bioinformatics (Oxford, England). 2010;26:589-95.
[2] Lai Z, Markovets A, Ahdesmaki M, Chapman B, Hofmann O, McEwen R, et al. VarDict: a novel and versatile variant caller for next-generation sequencing in cancer research. Nucleic acids research. 2016;44:e108.
[3] Cibulskis K, Lawrence MS, Carter SL, Sivachenko A, Jaffe D, Sougnez C, et al. Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. Nature biotechnology. 2013;31:213-9.
[4] Garrison E, Marth G. Haplotype-based variant detection from short-read sequencing. arXiv preprint. 2012;arXiv:1207.3907 [q-bio.GN].
[5] Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, et al. A method and server for predicting damaging missense mutations. Nature methods. 2010;7:248-9.
[6] Ng PC, Henikoff S. SIFT: Predicting amino acid changes that affect protein function. Nucleic acids research. 2003;31:3812-4.
[7] Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, et al. ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic acids research. 2014;42:D980-5.
[8] Mayakonda A, Lin DC, Assenov Y, Plass C, Koeffler HP. Maftools: efficient and comprehensive analysis of somatic variants in cancer. Genome research. 2018;28:1747-56.
[9] Mroz EA, Rocco JW. MATH, a novel measure of intratumor genetic heterogeneity, is high in poor-outcome classes of head and neck squamous cell carcinoma. Oral oncology. 2013;49:211-5.
[10] Muller PA, Trinidad AG, Timpson P, Morton JP, Zanivan S, van den Berghe PV, et al. Mutant p53 enhances MET trafficking and signalling to drive cell scattering and invasion. Oncogene. 2013;32:1252-65.