



## Research Paper

## Premalignant SOX2 overexpression in the fallopian tubes of ovarian cancer patients: Discovery and validation studies



Karin Hellner<sup>a,b,1</sup>, Fabrizio Miranda<sup>a,b,1</sup>, Donatien Fotso Chedom<sup>a,b,o,1</sup>, Sandra Herrero-Gonzalez<sup>a,b,1</sup>, Daniel M. Hayden<sup>c,1</sup>, Rick Tearle<sup>c,1</sup>, Mara Artibani<sup>a,b,d</sup>, Mohammad KaramiNejadRanjbar<sup>a,b</sup>, Ruth Williams<sup>d</sup>, Kezia Gaitskell<sup>a,b</sup>, Samar Elorbany<sup>a,b</sup>, Ruoyan Xu<sup>a,b</sup>, Alex Laios<sup>a,b</sup>, Petronela Buiga<sup>a,b</sup>, Karim Ahmed<sup>e</sup>, Sunanda Dhar<sup>f</sup>, Rebecca Yu Zhang<sup>c</sup>, Leticia Campo<sup>g</sup>, Kevin A. Myers<sup>g</sup>, María Lozano<sup>h</sup>, María Ruiz-Miró<sup>i</sup>, Sónia Gatiús<sup>j</sup>, Alba Mota<sup>k,l</sup>, Gema Moreno-Bueno<sup>k,l</sup>, Xavier Matias-Guiu<sup>j</sup>, Javier Benítez<sup>m,n</sup>, Lorna Witty<sup>o</sup>, Gil McVean<sup>o</sup>, Simon Leedham<sup>o</sup>, Ian Tomlinson<sup>o</sup>, Radoje Drmanac<sup>c,p</sup>, Jean-Baptiste Cazier<sup>g,q</sup>, Robert Klein<sup>c</sup>, Kevin Dunne<sup>c</sup>, Robert C. Bast Jr<sup>r</sup>, Stephen H. Kennedy<sup>b</sup>, Bassim Hassan<sup>s</sup>, Stefano Lise<sup>o</sup>, María José García<sup>m,n</sup>, Brock A. Peters<sup>c,p</sup>, Christopher Yau<sup>o,t</sup>, Tatjana Sauka-Spengler<sup>d</sup>, Ahmed Ashour Ahmed<sup>a,b,\*</sup>

<sup>a</sup> Ovarian Cancer Cell Laboratory, Weatherall Institute of Molecular Medicine, University of Oxford, Headington, Oxford OX3 9DS, UK

<sup>b</sup> Nuffield Department of Obstetrics and Gynaecology, University of Oxford, Women's Centre, John Radcliffe Hospital, Oxford OX3 9DU, UK

<sup>c</sup> Complete Genomics, Inc., 2071 Stierlin Ct., Mountain View, CA 94043, United States

<sup>d</sup> University of Oxford, Gene Regulatory Networks in Development and Disease Laboratory, Weatherall Institute of Molecular Medicine, Radcliffe Department of Medicine, Oxford OX3 9DS, UK

<sup>e</sup> Trinity College, University of Cambridge, Cambridge CB2 1TQ, UK

<sup>f</sup> Department of Histopathology, Oxford University Hospitals, Oxford OX3 9DU, UK

<sup>g</sup> Department of Oncology, University of Oxford, Old Road Campus Research Building, Roosevelt Drive, Oxford OX3 7DQ, UK

<sup>h</sup> Histopathology Core Unit, Spanish National Cancer Research Centre (CNIO), Madrid, Spain

<sup>i</sup> Biobank, Institut de Recerca Biomèdica Lleida (IRBLLEIDA), Lleida, Spain

<sup>j</sup> Department of Pathology and Molecular Genetics, Hospital Universitari Arnau de Vilanova, University of Lleida, IRBLLEIDA, Lleida, Spain

<sup>k</sup> Department of Biochemistry, Universidad Autónoma de Madrid (UAM), Instituto de Investigaciones Biomédicas "Alberto Sols" (CSIC-UAM), IdiPAZ, Madrid, Spain

<sup>l</sup> MD Anderson international Foundation, Madrid, Spain

<sup>m</sup> Human Genetics Group, Spanish National Cancer Research Center (CNIO), Madrid, Spain

<sup>n</sup> Biomedical Network Research Centre on Rare Diseases (CIBERER), Spain

<sup>o</sup> Wellcome Trust Centre for Human Genetics, NIHR Biomedical Research Centre, Roosevelt Drive, Oxford OX3 7BN, UK

<sup>p</sup> BGI-Shenzhen, Shenzhen, China

<sup>q</sup> Centre for Computational Biology, University of Birmingham, Edgbaston B15 2TT, UK

<sup>r</sup> Department of Experimental Therapeutics, M.D. Anderson Cancer Center, University of Texas, USA

<sup>s</sup> Tumor Growth Control Group, Sir William Dunn School of Pathology, South Parks Road, Oxford OX1 3RE, UK

<sup>t</sup> Department of Statistics, 1 South Parks Road, Oxford OX1 3TG, UK

## ARTICLE INFO

## Article history:

Received 31 May 2016

Received in revised form 30 June 2016

Accepted 30 June 2016

Available online 2 July 2016

## Keywords:

Ovarian cancer

Fallopian tube

BRCA mutations

SOX2

Screening

Precancer

## ABSTRACT

Current screening methods for ovarian cancer can only detect advanced disease. Earlier detection has proved difficult because the molecular precursors involved in the natural history of the disease are unknown. To identify early driver mutations in ovarian cancer cells, we used dense whole genome sequencing of micrometastases and microscopic residual disease collected at three time points over three years from a single patient during treatment for high-grade serous ovarian cancer (HGSOC). The functional and clinical significance of the identified mutations was examined using a combination of population-based whole genome sequencing, targeted deep sequencing, multi-center analysis of protein expression, loss of function experiments in an in-vivo reporter assay and mammalian models, and gain of function experiments in primary cultured fallopian tube epithelial (FTE) cells. We identified frequent mutations involving a 40 kb distal repressor region for the key stem cell differentiation gene SOX2. In the apparently normal FTE, the region was also mutated. This was associated with a profound increase in SOX2 expression ( $p < 2^{-16}$ ), which was not found in patients without cancer ( $n = 108$ ). Importantly, we show that SOX2 overexpression in FTE is nearly ubiquitous in patients with HGSOCs ( $n = 100$ ), and common in BRCA1-BRCA2 mutation carriers ( $n = 71$ ) who underwent prophylactic salpingo-oophorectomy. We propose

\* Corresponding author at: Ovarian Cancer Cell Laboratory, Weatherall Institute of Molecular Medicine, University of Oxford, Headington, Oxford OX3 9DS, UK.

E-mail address: [ahmed.ahmed@obs-gyn.ox.ac.uk](mailto:ahmed.ahmed@obs-gyn.ox.ac.uk) (A.A. Ahmed).

<sup>1</sup> These authors contributed equally to this work.

that the finding of SOX2 overexpression in FTE could be exploited to develop biomarkers for detecting disease at a premalignant stage, which would reduce mortality from this devastating disease.

© 2016 The Authors. Published by Elsevier B.V. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

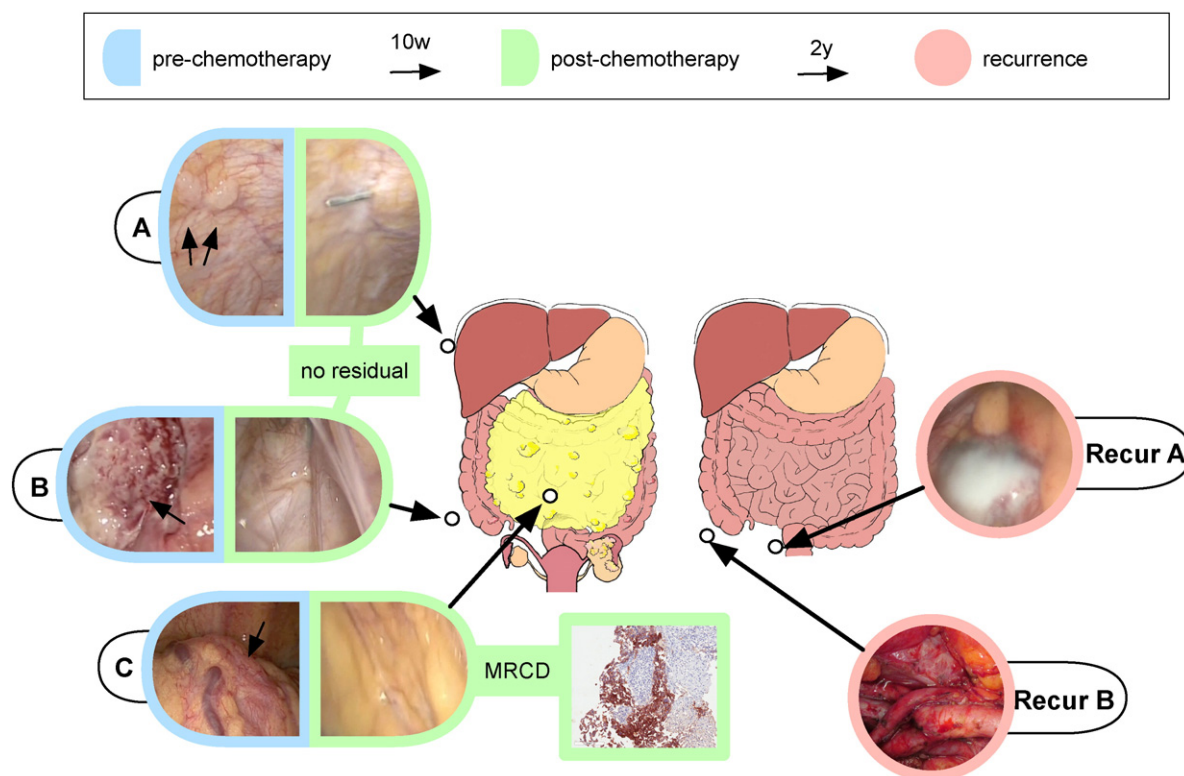
Because of late presentation and chemotherapy resistance ovarian cancer is the deadliest amongst all gynecological malignancies. Over 80% of ovarian tumors are high-grade serous ovarian cancers (HGSOCs), which is a particularly fatal type due to its initial asymptomatic but aggressive growth behavior. Current methods of detection have been successfully implemented for detection and possible reduction of mortality from ovarian cancer (Menon et al., 2015; Drescher et al., 2013). However, such methods are only capable of the detection of established invasive cancers. Understanding the natural history of the disease and the discovery of novel markers for detection at a premalignant stage will enable the effective control of ovarian cancer.

In this work, we prospectively analyzed the genomic composition of a single tumor over a three-year period to identify driver mutations that may have contributed to the initiation of the tumor. We identified non-coding mutations that cluster near genes involved in stem cell regulation. We established that one mutation is located in a previously unrecognized repressor element of *SOX2*, an important stem cell gene, and is associated with induction of *SOX2* expression. We demonstrate that the expansion of *SOX2*-expressing cells within the fallopian tube epithelium is a common feature of HGSOCs, a crucial finding that opens new avenues for early disease detection prior to clinical presentation.

## 2. Materials and Methods

### 2.1. Overall Description of the Study Design

The clinical samples for this translational study were obtained from patients recruited to the Gynecological Oncology Targeted Therapy Study 01 (GO-Target-01) and the Oxford Ovarian Cancer Predict Chemotherapy Response Trial (OXO-PCR-01) under research ethics approval number 11-SC-0014 and 12-SC-0404, respectively. We performed intraoperative video recording to document sampling sites (Supplementary video). Strict standard operating procedures were used to diminish the risk of DNA cross-contamination during sample collection and processing. Whole genome sequencing (WGS) of laser capture microdissected tumor islets ( $n = 30$ , Supplementary Fig. S1A) and bulk tissue samples of a single HGSOC (patient study ID: 11152). We obtained WGS data from 39 samples from three independent data sets and a tumor recurrence set (Supplementary Fig. S1B). Samples were obtained from different locations before chemotherapy, after neoadjuvant chemotherapy and approximately two and half years later at the time of first recurrence (Fig. 1, Supplementary Table 1). Complete macroscopic clearance at all sites (Supplementary video) as well as microscopic clearance of the peritoneal implants at sites A and B were documented following chemotherapy. Microscopic residual chemoresistant disease



**Fig. 1.** An ovarian cancer model for investigating primary chemotherapy resistance. A diagram illustrating the sites from which the biopsies were obtained in patient 11152 and the corresponding intra-operative images of the biopsy sites. The sub-diaphragmatic peritoneum (site A), the para-cecal peritoneum (site B) and the omentum (site C) were sampled in the primary tumor. A para-rectal mass (Recur A) and a pelvic node (Recur B) were sampled at presentation of disease recurrence. Note the complete macroscopic resolution of the tumor following chemotherapy (also see Supplementary video). TP53 immunohistochemical staining of a tumor islet from MRCD is also presented.

(MRCD) was only detected at the omentum in position C. Sequencing of tumor islets was performed using long fragment read (LFR) WGS (Peters et al., 2012) of 30 tumor islets (typically comprising only tens of cancer cells). A table of variants was generated and filtered using “cgatools” and custom scripts. GREAT software (McLean et al., 2010) was used for gene ontology analysis. Targeted sequencing was carried out for validation of mutations using 100 bp paired-end sequencing and read mapping was performed using STAMPY software (Lunter and Goodson, 2011), whilst variants were identified using PLATYPUS (Rimmer et al., 2014). Digital PCR was performed using the Droplet Reader (BioRad) and analyzed with the QuantaLife software (BioRad). For the in-vivo studies, fertilized chicken eggs were electroporated with enhancer constructs and analyzed as previously described (Betancur et al., 2010; Simoes-Costa et al., 2012). Targeted deletion of the BB5 region was performed using Clustered Regularly Interspaced Short Palindromic Repeats (CRISPR) (Cong et al., 2013). Chromatin immunoprecipitation (ChIP) assays were carried out in primary fallopian tube epithelial cells (Cong et al., 2013) and a fraction of the products was used for either real-time PCR reactions or high throughput sequencing. Immunohistochemistry studies were carried out on the Leica Bond Max staining system including the appropriate positive and negative controls. Detailed materials and methods are available in the Supplementary methods section.

## 2.2. Translational Studies

### 2.2.1. Ethical Approval

Tumor islets and blood samples for whole genome sequencing were obtained from case 11152 who provided written consent to participate in the prospective biomarker validation study Gynecological Oncology Targeted Therapy Study 01 (GO-Target-01) under research ethics approval number 11/SC/0014. Targeted sequencing was performed on samples from patients who participated in the same study and patients who participated in the prospective Oxford Ovarian Cancer Predict Chemotherapy Response Trial (OXO-PCR-01), under research ethics approval number 12/SC/0404. Necessary informed consents from study participants were obtained as appropriate.

### 2.2.2. Clinical History and Sample Collection of Patient 11152

Patient 11152 presented with radiological evidence of at least stage IIIC ovarian cancer with evidence of multiple omental and peritoneal nodules, subcapsular splenic nodules and left paracardiac lymph node enlargement and elevated CA125 tumor marker. The patient had a diagnostic laparoscopy and biopsy to confirm the diagnosis of HGSOc. On the day of the procedure, tissue samples were retrieved, split in halves and either saved in cryovials and immersed in dry ice within minutes from obtaining the biopsy in theatre or formalin fixed for standard histological diagnosis. Research samples were stored in  $-80^{\circ}\text{C}$ . In addition, a blood sample was obtained in theatre and saved in sterile EDTA-collection tubes. The patient received three cycles of paclitaxel and carboplatin and had near complete response. Following ten weeks from the first laparoscopy, the patient underwent a diagnostic laparoscopy to evaluate chemotherapy response, prior to proceeding, in the same session, to standard debulking surgery. Guided by the first video-laparoscopy, samples were obtained from the same sites from which the initial biopsies were taken.

### 2.2.3. Precautions for Tissue Handling to Diminish the Risk of Cross Contamination of DNA

Samples were collected from  $-80^{\circ}\text{C}$  to a box of dry ice and transported within the same building to the cryostat room. Individual cryovials were obtained from the dry-ice box and immediately placed within a pre-cooled cryostat stage at  $-25^{\circ}\text{C}$ . The cryostat (CryoStar 70X, Thermo Scientific) was not previously used for routine diagnostic or research work. The cryostat stage and dissection board were cleaned by ethanol prior to use for individual samples, and gloves were changed

before cutting each new tissue biopsy. For each sample, a single dissection blade (MB DynaSharp Microtome Blade, Thermo Scientific) and single sterile forceps (that was cleaned, individually wrapped and autoclaved prior to use) were used. Samples were obtained from the cryovials and placed on a sterile dish and cut using disposal single-use blades to obtain a piece of tissue for processing. These tissue pieces were then placed on individual new dissection discs for snap freezing in optimal cutting temperature (OCT) compound (NEG-50, Richard-Allan Scientific) prior to microtome cutting.

### 2.2.4. Section Processing for Laser Capture Microdissection

The first tissue section was mounted onto regular glass slides (SuperFrost Plus, VWR International) for hematoxylin (Hematoxylin solution, Gill No. 3, Sigma) and eosin (Eosin Y solution, Sigma) staining (H&E), according to manufacturer's instructions, followed by six to ten sequential tissue sections at  $6\text{ }\mu\text{m}$  thickness onto polyethylene naphthalate membrane (PEN) glass slides (MembraneSlide 1.0 PEN, Zeiss) that were pre-exposed to UV light for 30 min. Slides were immediately stored at  $-80^{\circ}\text{C}$ . Nuclease-free technique was used throughout the procedure and buffers and alcohol solutions were cooled to  $4^{\circ}\text{C}$  and used fresh each time. Each H&E slide was reviewed by a gynecological oncology pathologist (SD) to confirm the presence of cancer cells and delineate their location. PEN slides were dipped in 50% ethanol for fixation and rinsed in  $\text{H}_2\text{O}$  to remove excess OCT compound. The slides were stained with cresyl violet (Sigma Aldrich) at a concentration of 0.1% (weight/volume) in 50% ethanol for 15 s, rinsed in 50% ethanol and immediately used for microdissection. For formalin-fixed, paraffin embedded (FFPE) material,  $6\text{ }\mu\text{m}$  sections were cut onto activated PEN slides and dried at  $56^{\circ}\text{C}$  overnight, then dewaxed in Xylene (Sigma) and rehydrated through graded alcohols to water, then briefly dipped in 1% methyl green (Sigma), washed in  $\text{H}_2\text{O}$ . The slides were dried at  $37^{\circ}\text{C}$  for 1 h and then used for microdissection. Laser capture microdissection was performed on a PALM Laser Microdissection System (Zeiss) and the cut tissue was catapulted into 200  $\mu\text{l}$  membrane caps (AdhesiveCap 200 opaque, Zeiss, Jena, Germany) and immediately stored on dry ice. Images of empty caps as well as target area in  $5\times$ ,  $10\times$  and  $20\times$  magnification were obtained prior to microdissection for documentation. To maintain DNA integrity, slides were kept on dry ice until microdissection and microdissection was performed for no longer than 45 min per slide. Caps were stored at  $-80^{\circ}\text{C}$  until sequencing.

### 2.2.5. Macrodissection of FTE

Serial unstained FFPE tissue sections ( $10\text{ }\mu\text{m}$ ) were aligned with the corresponding H&E stained section in which the location of the FTE was pen marked. With a sterile, single-use scalpel blade tip the FTE tissue was scrapped off the glass slide and mounted onto 200  $\mu\text{l}$  a membrane cap (AdhesiveCap 200 opaque, Zeiss, Jena, Germany). To avoid cross contamination, the slides and caps were placed in sterile tissue culture dishes. DNA extraction and amplification was performed using the REPLig FFPE kit (Qiagen). For macrodissection of the p53 signature from FFPE immuno-stained sections the slides were incubated overnight in Xylene at  $37^{\circ}\text{C}$ . The coverslip was lifted and the slide soaked in Xylene for 30 min, then washed in gradient ethanol. The tissue was macrodissected using a scalpel blade tip. DNA was extracted with the Arcturus® PicoPure® DNA isolation kit (Life technologies) and amplified using the REPLig mini kit (Qiagen) according to manufacturer's instructions. DNA was quantified as described below.

## 2.3. DNA Extraction

For DNA extraction from frozen tissues 10 to 20 scrolls of tissue were cut at  $60\text{ }\mu\text{m}$  thickness using a cryostat as described above. For FFPE samples, 10 to 20 scrolls of  $20\text{ }\mu\text{m}$  thick sections were dewaxed with Xylene and washed with 100% ethanol. DNA was extracted using the DNeasy blood and tissue kit (Qiagen) according to manufacturers instructions. DNA concentrations were quantified using the PicoGreen



(Quant-iT™ PicoGreen® dsDNA Assay Kit, Life technologies), A260/280 ratio and absorption spectra were generated using a spectrophotometer (Nanodrop ND-1000, Thermo Scientific) and broad range Qubit system (Invitrogen) and quality was checked using a 1% agarose E-gel (Invitrogen).

## 2.4. Sequencing

### 2.4.1. Targeted Sequencing

LFR sequencing and whole genome sequencing were performed as previously described (Dean et al., 2002; Drmanac et al., 2010). A summary of the clinical characteristics of patients for whom targeted sequencing was performed is presented in Supplementary Table 2. For sequencing the 2 Mb flanking the SOX2 gene, the online NimbleDesign tool (<http://www.nimblegen.com/products/nimbledesign/index.html>) was used to design capture primers (Roche). For sequencing of the test set the TruSeq® DNA HT Sample Preparation Kit (Illumina) was used to allow for multiplexing and captured using the Seq Cap EZ Choice Library (Roche). Both kits were used according to manufacturer's instructions. Library quality control was carried out using the broad range Qubit system (Invitrogen) and the 2200 TapeStation (Agilent). Sequencing was carried out on a HiSeq2500 (Illumina) using TruSeq® Rapid SBS 100 bp paired end sequencing. For sequencing the validation set of matched tumor/normal DNA pairs, a more focused library of capture primers was designed targeting ~1.6 Mb flanking SOX2 (chr3: 180,806,202–182,429,714). DNA from paired samples and ChIP products were fragmented (S2, Covaris) and sample libraries were constructed and multiplexed using Seq Cap EZ kit (Roche) according to manufacturer's instructions. Capture was performed as described above. Quality control was carried out on a 2100 Bioanalyzer (Agilent) and library concentrations were measured with the high sensitivity Qubit system (Invitrogen). Sequencing was performed on a MiSeq platform (Illumina) using v3 MiSeq chemistry. To increase coverage, the libraries for the blood sample and the microdissected tubal epithelium of case 11152 were recaptured and sequenced using v2 MiSeq chemistry, 100 bp paired end.

### 2.4.2. Sanger Sequencing

For DNA sequencing using dye-terminator method, TP53 exon 8 was amplified and sequenced using primers TP53-forward 'GGGTGCAGTTA TGCCTCAGATT' and TP53-reverse 'CGGCATTTTGAGTGTAGACTGG' as previously described (Ahmed et al., 2010). SOX2 BB5 was amplified and sequenced using the BB5-forward 'CACCATGTGAATCATCTCG' and BB5-reverse 'ACCAGGTGTCCGAGAGTACG' primers. PCR was performed using the high fidelity DNA Phusion polymerase (NEB) as per manufacturer's instructions. Sequencing was performed for the rare variants identified in patients (Supplementary Table 3) using the primers listed in Supplementary Table 4.

### 2.4.3. Digital Droplet PCR

Digital droplet polymerase chain reaction (PCR) was performed on duplicate samples. Primers 5833217\_F; 5'-ACCTACTAGACCCCA GGCAAG-3' and 5833217\_R; 5'-GGCGCAGGAGGAGACC-3' were used to amplify a 60 bp amplicon containing the BB5 nucleotide and either detected using 5833217\_V; 5'-CCTGGGACCCAAACC-3' VIC-labeled probe for wild type or 5833217\_M; 5'-CTGGCACCCAAACC-3' FAM-labeled probe for mutant amplicons (TaqMan® SNP Genotyping Assays, custom design, Roche Molecular Systems). TP53 mutation was quantified using primers 22410689\_F; 5'-CTGTGCGCCGGTCTCT-3' and 22410689\_R; 5'-TGGGACGGAACAGCTTGTAG-3' to amplify a 64 bp amplicon and detected using 22410689\_V 5'-TGCGTGTGTGCTG-3' VIC-tagged probe for wild type and 22410689\_M; 5'-TGCGTGTGTGCTG-3' FAM-tagged probe for mutant amplicons. Reactions were prepared using droplet digital PCR Super Mix (BioRad) and standard PCR performed according to manufacturer's instructions. Amplification events were detected with a digital PCR plate reader

(QX100 Droplet Reader, BioRad) and data was analyzed using the QuantaSoft Software (Version 1.3.2.0, BioRad). Average droplet count was 11,728 per sample. Samples with <7000 droplets were excluded from the analysis.

## 2.5. Sequencing Analysis

### 2.5.1. Whole Genome Sequencing

Reads were mapped to the reference genome (GRCh37) and variants were called by local de novo assembly as previously described (Carnevali et al., 2012). A table of variants across the various tumor and normal genome assemblies was generated and filtered using the Complete Genomics cgatools program and custom scripts. Sequence data were deposited in the European Genome-Phenome Archive (EGA) with the following accession number: EGAS00001001909.

### 2.5.2. Performing Gene Ontology Enrichment Analysis

The functional prediction of *cis*-regulatory regions was performed using the Genomic Region Enrichment of Annotation Tool (GREAT) as previously described (McLean et al., 2010). In brief, GREAT assigns regulatory domains for each gene that consists of a basal domain (5 kb upstream and 1 kb downstream of transcription start site [TSS]) plus extension of up to 1 Mb, but not beyond 1 Mb, in both directions to the nearest gene basal domain. The enrichment for a particular biological process gene ontology (which consists of a number of genes) was computed by obtaining the ratio of the fraction of foreground (FG) variants (i.e. the ancestor variants) that mapped to genes of a gene ontology  $x$  to the fraction of background (BG) variants (e.g. progeny variants) that mapped to the same gene ontology. To be regarded as a gene ontology hit, we required that a gene ontology had a false discovery rate "Q value" of <0.01 and a number of genes supporting a particular ontology of more than ten. The BG was defined as either the progeny variants (variants present in two or more tumor sites), the whole genome or the number of germline variants as indicated in the text. Permutation analysis was based on the assumption that random samples from the background of equal size to the FG should not give higher enrichment than the one observed for the true FG. To test this, we obtained 10,000 samples of 750 variants from the BG and computed the enrichment for each and counted the number of times in which that enrichment was higher than the one observed by the true FG and presented the result ( $p$ -value) as the fraction of the count from the total number of random samples. The number of genes that GREAT assigned to the gene ontology "stem cell differentiation" was 73 and this was the basis of the analysis described above. We repeated the above analysis using the total number of human genes (283) assigned to the same gene ontology "GO:0048863" at the Gene Ontology data base (<http://amigo.geneontology.org/cgi-bin/amigo/go.cgi>) and by using a subset of genes ( $n = 81$ ) that had an experimentally verified link to stem cell differentiation by selecting the human genes that satisfied any of the following terms: IDA; inferred from direct assay, IEP; inferred from expression pattern, IGI; inferred from genetic interactions, IMP; inferred from mutant phenotype.

### 2.5.3. Analysis of Targeted Sequencing Data

Reads were mapped to the reference genome (GRCh37) using STAMPY software (Lunter and Goodson, 2011) and variants were identified using an in-house developed software, PLATYPUS (Rimmer et al., 2014). Downstream analysis was performed using in-house developed scripts. There were a total of 37,291 variants identified by sequencing. Further analysis focused on single nucleotide substitutions in cancer samples ( $n = 18,456$  in 33 samples). To identify high quality rare variants, the analysis was restricted to variants that were present in less than five samples, were not called in the 1000 Genomes Project (Abecasis et al., 2012), and had a high quality score (i.e. flagged as "pass" or "allele bias" by PLATYPUS). Only 861 variants met these criteria and these are shown in Supplementary Table 5. In order to

compute whether the observed rare variants displayed regional enrichment in the ovarian cancer set we adopted a simulation approach. First, we obtained minor allele frequencies (MAF) for common SNPs (defined as having MAF > 5%) from the 1000 Genomes Project in the 2 Mb region flanking SOX2. We then simulated genotypes, based on those minor allele frequencies, for samples consisting of 33 individuals (we ignore linkage effects) and counted the number of heterozygotes or non-reference homozygotes observed across the 33 simulated profiles. This process was repeated 1000 times in order to obtain an average profile of common SNP density across the region. We then examined 40 kb windows, spaced at overlapping intervals of 100 bp, and counted the number of rare variants observed across the 33 ovarian cancers. For each window, we then computed the test statistic (observed number of rare variants in ovarian cancers/expected number of common SNPs). In order to assess whether the calculated ratios were significantly enriched, we randomly sampled 100 subsets of 33 individuals from the welllderly (well elderly individuals) dataset and computed the same statistic for each subset by counting rare variants falling in any particular window. We then computed a *p*-value from the order-statistic of the observed test statistic in the ovarian cancers compared to those of the welllderly data.

In addition, the 861 variants were annotated using data from the Encyclopedia of DNA Elements (ENCODE) (Bernstein et al., 2012) to identify those that were within regulatory regions as reported by digital genomic footprint (Neph et al., 2012) downloaded from ([ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration\\_data\\_jan2011/byDataType/footprints/jan2011/](ftp://ftp.ebi.ac.uk/pub/databases/ensembl/encode/integration_data_jan2011/byDataType/footprints/jan2011/)), by DNaseI hypersensitivity (Thurman et al., 2012) downloaded from (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegDnaseClustered/>) or by ChIP-Seq (Yip et al., 2012) experiments downloaded from (<http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeRegTfbsClustered/>). In order to assess the degree of local enrichment of variants we first computed the global rate of variant occurrence over the 2 Mb region (113 variants/2,000,000 bp). We then considered a series of 40 kb overlapping windows spaced at 100 bp intervals spanning the 2 Mb. For each window, we counted the number of variants observed in the tumors and computed the probability of observing (at least) this many variants under the null hypothesis of uniform variant occurrence. We assumed that, under the null hypothesis, the number of observed variants follows a Poisson distribution with a rate parameter given by the global rate over the region.

#### 2.5.4. Analysis of ChIP-sequence Tracks

Following the ENCODE guidelines (Landt et al., 2012), read quality was first assessed using FastQC (v.0.10.1; <http://www.bioinformatics.babraham.ac.uk>). Low quality reads were filtered by using Trim galore (v 0.3.1; <http://www.bioinformatics.babraham.ac.uk>) for paired reads. The 3' and 5' adapters were trimmed using the Cutadapt program implemented in Trim galore. Only reads with a length > 60 bp after trimming were retained. Reads were then mapped against the human reference genome (build hg19) using bowtie (Langmead and Salzberg, 2012). Only non-duplicated aligned reads were retained. Multiple alignment of read was not allowed. To identify enriched regions, peaks were called on the immune-precipitated samples versus their input controls using Model-Based Analysis of ChIP-seq 2 (Feng et al., 2012) (MACS2 version 2.0.10) with a *p*-value threshold of 0.001 and default parameters. Signal track was built following the guideline at <https://github.com/taoliu/MACS/wiki/Build-Signal-Track>. Exploratory analysis and generation of figures was performed in R using custom scripts.

#### 2.5.5. Motif Analysis for the BB5 Variant

The motif UW.Motif.0169 overlapping the BB5 variant was discovered from digital footprints data. The logo plot in Supplementary Fig. S6A was generated using the webLogo3 software (<http://weblogo.threeplusone.com/>).

### 2.6. Fallopian Tubes Primary Epithelial Cell Culture

Patients scheduled to undergo surgical procedures provided written consent, prior to surgery, agreeing to participate in the study. The infundibular region of the fallopian tube was isolated, dissected and opened to reveal the lumen. Fallopian tubes were incubated in 15 ml conical tubes containing 0.5% trypsin and 0.1% DNaseI in MEM for 1 h at 37 °C, with shaking. The supernatant, containing the epithelial cells, was removed and mixed with 10% FBS in DMEM. Fallopian tube epithelial cells (FTEC) were centrifuged and plated in 10% DMEM. The purity of epithelial cells was checked by immunofluorescence. Optional purification using CD326 microbeads (Miltenyi Biotech) was if further purification was required.

#### 2.7. Viral Transduction

pLx302-SOX2 was generated using Gateway system (Invitrogen) according to the manufacturer's instructions. SOX2 was cloned into pDONOR221 (Invitrogen) from pMXs-Sox2 (Addgene) plasmid by BP Gateway clonase (Invitrogen) to generate entry clones. Using LR Gateway clonase (Invitrogen) SOX2 was cloned into the destination vector pLx302 (a gift from Dr. David Root (Yang et al., 2011); Addgene plasmid 25896). Packaging cells (HEK-293T) were co-transfected with the packaging vector p8.91, the envelope plasmid pMDG (gifts from Dr. Didier Trono, University of Geneva, Switzerland) and pLX302-SOX2 as previously described (Besnier et al., 2002). Following 72 h the HEK293-T medium containing the virus was collected, filtered through a 0.45 µm Minisart NML Syringe Filter (Sartorius) and stored at –80 °C for later transduction.

#### 2.8. Cloning, Mutant Generation and Chicken Embryo Transfection

The 1 kb regions flanking the BB5 SNP was cloned from Human genomic blood DNA by PCR amplification using Phusion high fidelity polymerase (NEB) according to the manufacturer's instructions using the following primers BB5\_F; 'TTTTTTCGTCTCgccaggTTACTCCAATA TGAGAGATAAGAGCA' and BB5\_R; 'TTTTTTCGTCTCcaacagCGCTCACACG GTGATTAGAA', Test1\_F; 'TTTTTTCGTCTCgccaggCATTACTGGCAGCTGA GGGG' and Test1\_R; 'TTTTTTCGTCTCcaacagTGATTTTCCTGGGCAGACA', Test2\_F; 'TTTTTTCGTCTCgccaggCTACTAGACCCAGGCAAGG' and Test2\_R; 'TTTTTTCGTCTCcaacagTGATTTTCCTGGGCAGACA', Test3\_F; 'TTTTTTCGTCTCgccaggTCCCTGTTCTCACTCTCT' and Test3\_R; 'TTTTTTCGTCTCcaacagTGATTTTCCTGGGCAGACA'. Products were cloned into a ptk citrine – BsmBI vector. Restriction digestion of the PCR products and the vector was performed using BsmBI enzyme (NEB) and the digested product was inserted into the vector using a T4 ligase (NEB). Point mutants were generated using QuikChange® II Site-Directed Mutagenesis Kit (Stratagene) as per manufacturer's instructions using the citrine vector containing each region and the primer sets SOX2\_BB5mut\_F; 'ACCTGGGCCTGGCACCCAAACCCTT' and SOX2\_BB5mut\_R 'AAGGGTTTGGGTGCCAGGCCAGCT'. DNA sequences of all cloned PCR products were verified with direct sequencing.

Fertilized chicken (*Gallus gallus*) eggs (Henry Stewart & Co. Ltd, Louth, UK) were incubated at 37 to 38 °C for approximately 20 to 24 h prior to electroporation. In the UK no license is required at the embryo stage used for this work. The entire epiblast of stage four chicken embryos (staged according to Hamburger and Hamilton (Hamburger and Hamilton, 1992)) was electroporated with enhancer constructs, cultured using modified new culture (Sauka-Spengler and Barenbaum, 2008) and analyzed as previously described (Betancur et al., 2010; Simoes-Costa et al., 2012) (*n* = 10 to 15 per group). Neither randomization nor blinding was carried out. Non-vital or damaged embryos were excluded from further analysis. Immunofluorescence for detection of endogenous Sox2 expression and co-localization with citrine expression was carried out as previously described (Betancur et al., 2010; Simoes-Costa et al., 2012) using anti-Sox2 antibody (ab97959, Abcam) or anti-GFP antibody (MAB3580, Millipore) and detected using alexa-488 conjugated and alexa-568 conjugated secondary antibodies

(Invitrogen) and observed using an inverted confocal microscope (LSA 510 META, Zeiss).

### 2.9. Chromatin Immunoprecipitation (ChIP) Assay

ChIP was performed on FTEC and HEK293T cells (purchased from ATCC, mycoplasma tested) using the commercially available ChIP-IT express enzymatic kit (Active Motif) according to manufacturer's instructions and the following antibodies: H3K27ac (C15410174 lot: A.7071-001P; Diagenode) and EZH2 (C15410039; lot: 003; Diagenode). Briefly, one million cells were cross-linked using formaldehyde 1% for 10 min and the reaction was stopped by adding glycine for 5 min at room temperature. Cells were then washed twice with ice cold PBS containing  $1 \times$  PMSF, and then resuspended in cell lysis buffer containing  $1 \times$  protease inhibitor cocktail and PMSF and rotated at 4 °C for 30 min. Nuclei were extracted using a Dounce homogenizer (pestle B) and treated with an enzymatic shearing cocktail at 37 °C for 20 min. Digested chromatin was immunoprecipitated at 4 °C overnight by adding protein G magnetic beads and rabbit IgG (cat. N. 12-370; lot. N. 1969696; Millipore) as a negative control. Following washes, reverse cross-linking and proteinase K treatment the DNA was purified using the Chromatin IP DNA purification kit (Active Motif). Sequencing of ChIP material was performed as described above. Real-time PCR was performed in a 96-well plate qPCR machine (ABI 7000, Applied Biosystems). Reactions were set up using SybrGreen (Applied Biosystems) according to the manufacturer's protocols using 1 µl of each ChIP reaction or 1: 100 to 200 dilution of the input fraction. The  $\Delta\Delta C_t$  method was used for quantification and calculations performed according to ChIP-qPCR data analysis instructions (Supper-Array, Bioscience Corporation). The following primers for the BB5 region were used: 5833217\_F 5'-ACCTACTAGACCCAGGCAAG-3' and 5833217\_R 5'-GGCGCAGGAGGAGACC-3'. For EZH2 inhibition, cell were treated for 96 h with 5 µM of UNC1999 (Sigma) before chromatin extraction.

### 2.10. CRISPR Vector Construction, Cell Culture and Transfection

Vector px330 was used as previously described (Cong et al., 2013). A pair of oligonucleotides (CRISPR-BB5-F; 5'-CACCGAGGGTTTGG GTCCCAGGCC-3' and CRISPR-BB5-R; 5'-AAACGGGCTGGGACCC AAACCTC-3') encompassing the BB5 nucleotide (underlined) and extending up to a protospacer adjacent motif [AGG] (not shown) were annealed, phosphorylated, and ligated to the linearized vector.

HEK293 cells (ATCC) were maintained in DMEM (Invitrogen) supplemented with 10% fetal bovine serum and 100 U/ml penicillin/streptomycin and incubated at 37 °C and 5% CO<sub>2</sub>. HEK293 cells were transfected with the construct px330-BB5 using FUGENE HD (Promega) according to the manufacturer's instructions. After two weeks, the DNA for each clone was extracted using DNeasy blood and tissue kit (Qiagen) and the region flanking the BB5 nucleotide was amplified using the following pair of primers: SH-BB5-01-F; 5'-TCCAATATGAGAGATAA GAGCAC-3' and SH-BB5-01-R; 5'-GCTGAAAAGACCAACTTAAAC-3'. HEK293T cells were transfected and two days later single cells were plated in 96-well plates and harvested for DNA collection on the indicated dates. The number of positive clones carrying bi-allelic mutations (nominated as homozygous), the number of clones carrying mono-allelic mutation (nominated as heterozygous) and clones with no deletions (nominated as negative) are reported in Supplementary Table 6.

### 2.11. Immunofluorescence

FTEC or SKOV3 cells were fixed with 4% (w/v) formaldehyde in PBS for 4 min and then permeabilized overnight with 100% ethanol at -20 °C. Cells were then blocked with 1.5% BSA in TBS for 1 h and incubated with primary mouse antibody against pan-cytokeratin (ab7753, Abcam), primary rabbit antibody against MYC (9402S, Cell Signaling Technology), primary rabbit antibody against SOX2 (3579, Cell

signaling Technology), primary mouse antibody against SOX2 (4900, Cell Signaling Technology) or primary rabbit antibody against PAX8 (10336, Proteintech) for 1 h. Cells were washed using washing buffer (TBS + Triton 0.2% + SDS 0.04%) and then incubated with donkey anti-mouse or donkey anti-rabbit conjugated with Alexa Fluor 488 or Alexa Fluor 546 secondary antibodies (Invitrogen), respectively. After three washes cells were mounted using DAPI-containing mounting media (Vector Laboratories).

### 2.12. Immunohistochemistry

A summary of patients from both groups (discovery set and validation sets) for whom SOX2 IHC was performed is presented in Supplementary Table 7. Power calculations were performed to determine the minimum specimen number for the validation set based on the data from the discovery set (mean % for benign: 3, s.d.: 4; mean % for HGSOs: 14, s.d.: 10, difference between means: 11%) assuming a desire for 90% power and a 5% significance level. To identify the difference in 3+ staining intensity for the above scenario we estimated a group size of ten cases.

Tissue sections of 2.5 µm or 4 µm thickness were cut from FFPE tumor or control samples. Automated staining was carried out with the Leica Bond Max autostainer (Leica Microsystems) or Autostainer plus Link 48 (Dako®). In short, antigen retrieval at 100 °C for 20 min was followed by primary antibody incubation with the rabbit anti-SOX2 (3579, Cell Signaling Technology) or IgG control for up to 1 h then detection using the BOND™ Polymer Refine Detection System (DS9800, Leica Biosystems) as per manufacturer's instructions. Double staining was achieved using the BOND™ Polymer Refine Detection System and the BOND™ Polymer Refine Red Detection System (DS9390, Leica Biosystems) according to the manufacturer's instructions. Stained slides were scanned at 20× and 40× magnification on the Aperio slide scanner (Aperio) or Mirax scan (Zeiss) fitted with a 40×/0.95 Plan Apochromat objective lens (Zeiss) for the second validation set (CNIO; Spanish National Cancer Research Centre). The ImageScope software (v11.2.0.780, Aperio) was used for quantification of nuclear staining for the discovery set and the first validation set. For scoring of SOX2 positivity in fallopian tubes, only the tubal epithelium was marked by using the "negative pen tool" to exclude stroma. The marked FTE was analyzed with the program algorithm "nuclear v9" which scored the staining of all nuclei within the marked area and assigned scores of 0 for not detectable signal, +1 for weak staining, +2 for moderate staining and +3 for strong staining. Nuclear positivity was confirmed on selected areas using the "deconvolution" algorithm to subtract SOX2 from underlying hematoxylin staining. The diagnosis and clinical details are provided in Supplementary Table 7. For scoring of SOX2 expression in HGSOs tumor foci were marked and benign tissue excluded and the procedure completed as described above. For scoring the second validation set (CNIO), the AxioVision 4.6.3 software (Zeiss) pattern recognition training software was used. Algorithm parameters were set by using the AxioVision Wizard so that threshold and filter conditions (size, diameter and perimeter of different spots) were established to train the application in a noise reduction method that determines the different positive activation signals: 0 for no signal, +1 for weak staining, +2 for moderate staining and +3 for strong staining. Immunohistochemistry for other proteins was conducted in a similar manner to the method described above using anti-p53 (Clone DO-7, Dako) and anti-WT1 (Clone WT1-562, Leica) antibodies, anti-TUBB4 (T7941, Sigma), anti-MYC (ab32072, Abcam), anti-EZH2 (Clone 6A10, Leica).

## 3. Results

### 3.1. Laparoscopy-guided Prospective Multi-region Sampling in an Ovarian Cancer Patient

We set out to discover early (root) driver mutations in a single high-grade serous ovarian tumor that may have contributed to its genesis.



We hypothesized that such early mutations would be present in all cancer cells irrespective of their location, time of presentation or biological behavior (e.g. chemotherapy response). To account for these three variables, we utilized intraoperative video recording to document the site of sampling before and after chemotherapy and at recurrence and to obtain a direct visual evaluation of chemotherapy response. LFR WGS technology enabled the accurate sequencing of very low (picogram) quantities of DNA (Peters et al., 2015) which is not currently possible by standard WGS or other methods of long fragment sequencing (Kuleshov et al., 2014). Since post-chemotherapy samples only had microscopic residual disease, using this technique on laser-capture micro-dissected cancer islets was essential.

We identified 750 unique mutations that were present in practically all tumor islet samples in our discovery set ( $\geq 90\%$  of tumor islets) irrespective of space, tumor site or biological behavior. These mutations were extensively confirmed using standard WGS and targeted sequencing (Supplementary Fig. S2A). The variants included two synonymous mutations, eight non-synonymous mutations in protein-coding regions, and 740 mutations in non-coding regions (Supplementary Table 8). One mutation, detected in exon 8 of *TP53*, was shared between all tumor islets but was not present in normal stroma samples from patient 11152. This result was consistent with the previous finding by Ahmed et al. (2010) and Anon (2011) of ubiquitous *TP53* mutations in HGSOCS and has, therefore, validated our approach of discovering potential driver mutations.

### 3.2. Non-coding Mutations Cluster at Potential cis Regulatory Elements of Genetic Drivers of Stem Cell Differentiation

We mapped the 750 mutations to the nearest transcription start sites of genes and performed pathway-enrichment analyses of the identified genes as previously described (McLean et al., 2010). This analysis revealed five significantly enriched ontologies of biological processes (Supplementary Fig. S2B, Supplementary Tables 9 and 10) that were all related to embryonic and stem cell differentiation. Furthermore, the stem cell differentiation ontology was supported by 23 mutations that mapped to 15 genes (e.g., *SOX2*, *PAX7*, *WNT7A*). This finding was confirmed by several iterations of analysis (Supplementary Fig. S3) and multiple rounds of further sequencing (Supplementary Fig. S1A). The identification of *SOX2*, a key driver of stem cell differentiation (Takahashi and Yamanaka, 2006) that was recently implicated in skin cancer tumor initiation (Boumahdi et al., 2014), was particularly notable in our analysis because of the proximity of a total number of six mutations to the *SOX2* gene locus. We termed the nucleotides, in which the mutations mapping near *SOX2* occurred, BB1 to BB6.

### 3.3. The BB5 Mutation Occurred at the Pre-neoplastic Lesion of the Tumor and Marked a Region That Was Frequently Mutated in HGSOCS

We next performed deep-targeted sequencing of the 2 Mb region flanking the *SOX2* gene locus and BB1 to BB6 on tumor samples from 33 patients with HGSOCS, including case 11152 (Table 1, Supplementary Table 2), and identified 861 single nucleotide polymorphisms (SNPs) in this region (Supplementary Table 5) that were not previously reported in the 1000 Genomes Project (median = 21, range = 11 to 97 per cancer sample). Comparing the frequency of occurrence of these SNPs with those previously identified in the 1000 Genomes Project or with SNPs identified from whole genome sequencing analysis of a new set of 597 healthy elderly volunteers combined with deep whole genome re-sequencing of DNA from 436 individuals from the 1000 Genomes Project identified a 40 kb region flanking the BB5 nucleotide (Fig. 2A), that was significantly enriched in rare SNPs ( $p < 0.01$ , permutation analysis). These mutations significantly targeted biochemically active sites (implying that they were active regulatory elements) as determined by mining the ENCODE data for DNase hypersensitivity or transcription

**Table 1**

Patient baseline data for targeted sequencing analysis. The clinical details of patients who donated samples used for targeted sequencing are presented in this table.

Patients for targeted sequencing	49
Age at surgery	
Mean (median)	65.3 (66)
Min–max	40–82
<60 years	20.4%
>60 years	79.6%
Site of primary disease	
Ovary	29 (59%)
Tube	10 (20%)
Peritoneum	8 (16%)
Other <sup>a</sup>	2 (4%)
Tumor status	
Primary	19 (39%)
IDS	15 (31%)
Post-NACT	12 (24%)
Recurrence	3 (6%)
Biopsy site	
Ovary	17
Peritoneum <sup>b</sup>	15
Omentum	17
Other <sup>c</sup>	4
Normal tissue or blood <sup>d</sup>	24
Tumor stage <sup>e</sup>	
I–II	10 (20%)
III–IV	39 (80%)

Data are  $n$  (%). Disease stages I and II indicate early disease, stages III and IV indicate advanced disease. NACT = neoadjuvant chemotherapy. IDS = interval debulking surgery (e.g. after three or four cycles of NACT).

<sup>a</sup> Ambivalent histology includes one case of tubal or ovarian origin and a case of ovarian or endometrial origin.

<sup>b</sup> Peritoneum includes diaphragmatic peritoneum and small bowel serosa and mesentery.

<sup>c</sup> Other biopsy sites include spleen, parasacral tumor and falciform ligament.

<sup>d</sup> Normal tissue includes primary cell lines generated from normal fallopian tube epithelium.

<sup>e</sup> For staging we used the FIGO classification valid at time of initial diagnosis.

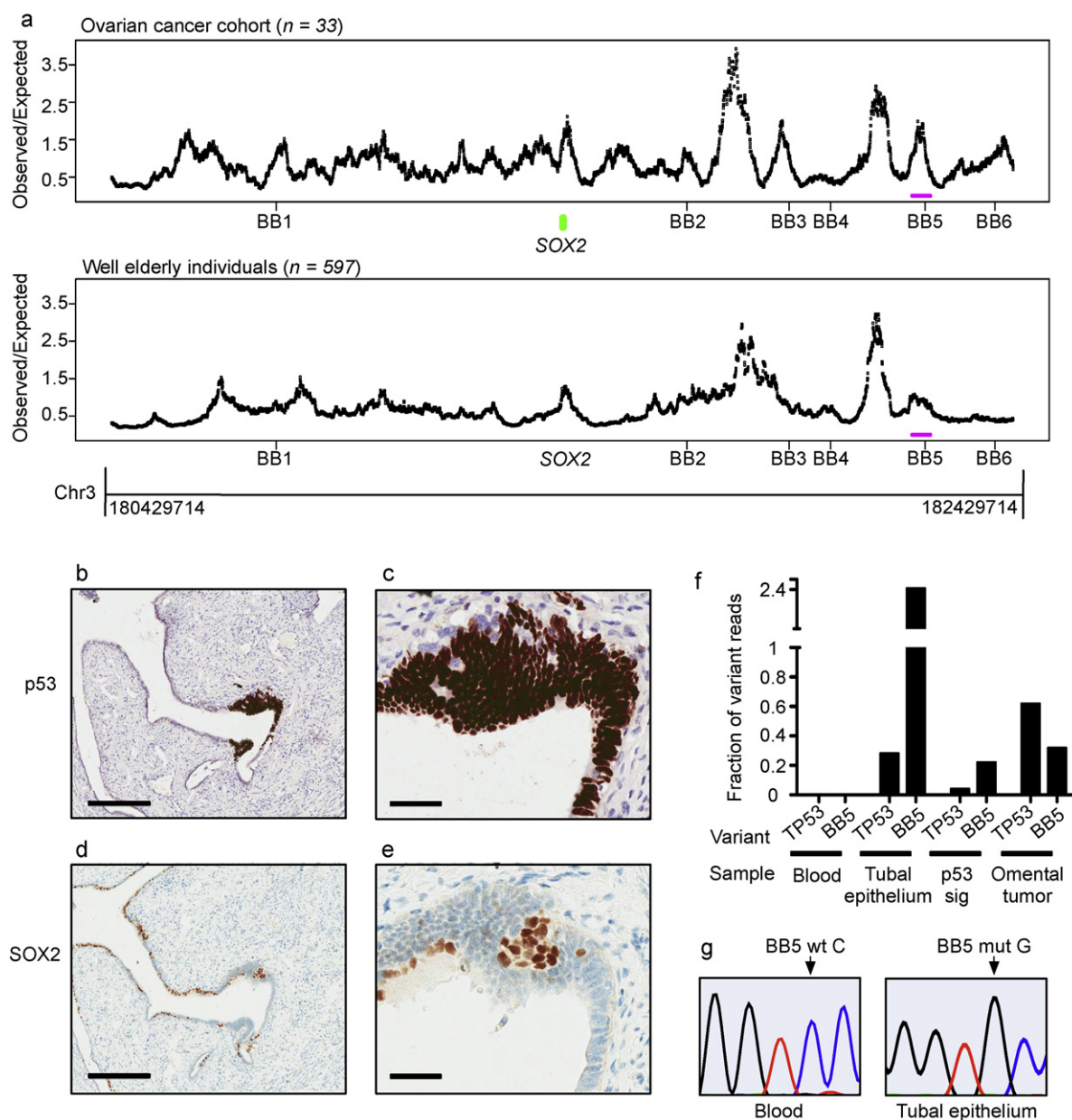
factor binding activity (Supplementary Fig. S4). We therefore chose to characterize this 40 kb region, referred to as the BB5 region, further.

Sequencing results from tumor and normal tissue revealed that the BB5 region (chr3: 182,189,714–182,229,714) included 21 single nucleotide substitutions in HGSOCS from 16 patients (48.5% of all patients), of which nine variants from seven patients (21.2%) were somatic mutations (Supplementary Table 3). Independent validation by deep-targeted sequencing of a 1.6 Mb region flanking *SOX2* in tumor-normal tissue pairs of 16 additional patients with HGSOCS identified two further mutations in the BB5 region in two patients.

Examining the fallopian tube as a potential tissue of origin of HGSOCS (Crum, 2009; Lee et al., 2007; Kurman and Shih, 2011; Karst et al., 2011) led to the identification of the p53 signature (Lee et al., 2007), indicated by p53 nuclear overexpression in fallopian tube epithelium (FTE), in patient 11152 (Fig. 2b–c). A subpopulation of the cells at the signature also expressed high levels of nuclear *SOX2* (Fig. 2d–e). Using a combination of deep-targeted sequencing, droplet-digital PCR and Sanger sequencing, we confirmed that both, the BB5 nucleotide and *TP53*, were already mutated in the p53 signature and the adjacent FTE (Fig. 2f–g). Importantly, in two additional patients for whom DNA extraction from the paraffin-embedded FTE was possible, we showed that the mutations occurring in the fully established HGSOCS were also present in the FTE (Supplementary Fig. S5).

### 3.4. Expansion of FTE Cells Strongly Expressing *SOX2* is a Feature of HGSOCS

The profound increase in the number of *SOX2*-expressing cells in the FTE of our index patient was in sharp contrast to the rare *SOX2* expression in the FTE of patients with benign conditions. Paradoxically, *SOX2*



**Fig. 2.** A 40 kb region is frequently mutated in HGSOs and consists of a distal SOX2 repressor. **a.** To determine whether any of the six mutations (BB1 to BB6) mapping near the SOX2 gene locus marked regions that were sites of frequent occurrence of private variants or mutations in HGSOs, we performed deep-targeted sequencing of the 2 Mb region flanking SOX2 on 33 HGSOs (Supplementary Table 2). A total of 861 single nucleotide substitutions (Supplementary Table 5) were identified that were not previously reported in the 1000 Genomes Project (median = 21, range = 11 to 97). Because functionally important genomic regions tend to be significantly less susceptible to genomic variation within a population, we determined whether the identified rare variants accumulated in specific areas within the 2 Mb region that were less susceptible to genomic alterations on a population scale. To test this hypothesis, we constructed overlapping moving windows of 40 kb size and compared the observed frequency of rare mutations (not previously described in the 1000 Genomes Project) in our group of patients. The expected frequency of SNPs in the same windows was based on 1000 sets of simulated cohorts of 33 individuals from the previously reported 1000 Genomes Project data. Upper panel: shown is the ratio of the observed number of variants in 40 kb “moving” windows in the cancer set to the expected number of variants in the equivalent windows based on 1000 permutations of simulated 1000 Genomes Project data. Based on this analysis, a peak observed/expected ratio (enrichment statistic) was identified in a 40 kb region flanking the BB5 nucleotide referred to as the BB5 region. Lower panel: To test whether this observation was higher than what would be expected by chance, we sequenced germline DNA from 597 healthy elderly volunteers and sequenced germline DNA from 436 individuals from the 1000 Genomes Project at higher depth. We then identified rare variants in the elderly set and repeated the above analysis. Comparing the enrichment statistic in the BB5 region in the cancer set to that obtained from 100 permutations of 33 individuals from the elderly set confirmed the significant enrichment of rare variants in the ovarian cancer set ( $p < 0.01$ ). **b–e.** Immunohistochemical staining of p53 and SOX2 at low (**b** and **d**, scale bar = 300  $\mu$ m) and high (**c** and **e**, scale bar = 50  $\mu$ m) magnification in normal FTE and the p53 signature in patient 11152. Note the strong focal p53 staining at the multi-layered epithelium (p53 signature). **f.** The fraction of mutant reads relative to the total number of reads of the BB5 and TP53 mutations in germline DNA, the FTE, the p53 signature (p53 sig) and the tumor of the index case ( $n = 4$  repeats). **g.** Sequencing trace indicating the BB5 mutation.

expression in tumor cells was almost absent (Supplementary Figs. S5 and S6). The increase in SOX2 expression in potential precursor cells is particularly intriguing since it is one of only four factors required for acquisition of stem cell properties by terminally differentiated cells (Takahashi and Yamanaka, 2006) with important implication for acquisition of pluripotency, the ability of a cell to differentiate into different cell types, and susceptibility to tumor formation.

We next compared SOX2 expression in the nuclei of normal FTE of patients with benign conditions and patients with endometrial cancer (age-matched cases) and the normal FTE and corresponding ovarian tumors from patients diagnosed with HGSO (Table 2, Supplementary Table 7). Automated image analysis of these samples demonstrated that the median SOX2 expression (intensity score of 3+) in the normal appearing FTE from patients with HGSOs was 17-fold higher than that



**Table 2**

Patient characteristics for immunohistochemistry studies. The clinical details of patients who donated the samples used for SOX2 immunohistochemistry are presented.

Patients for IHC staining <i>n</i> = 279	Benign	HGSOC	BRCA mutation carrier
Patient cohort	108	100	71
Discovery cohort	16	22 <sup>a</sup>	
Validation cohort	92	78	71
Age at surgery			
Mean (median)	48.0 (48)	62.7 (63)	50.7 (48)
Min–max	28–81	32–94	29–81
Site of primary disease			
Ovary		79 (79%)	
Tube		10 (10%)	
Peritoneum		5 (5%)	
Other <sup>b</sup>		6 (6%)	
Tumor status			
Primary		60 (60%)	
IDS		15 (15%)	
Post-NACT		19 (19%)	
Tumor stage <sup>c</sup>			
I–II		26 (26%)	
III–IV		69 (69%)	

Data are *n* (%) or mean (median), unless otherwise indicated. Disease stages I and II indicate early disease, stages III and IV indicate advanced disease. IDS = Interval debulking surgery (e.g. after three or four cycles of NACT). NACT = neoadjuvant chemotherapy.

<sup>a</sup> These patients were also included in targeted sequencing study.

<sup>b</sup> Ambivalent histology cohort includes four cases of ovarian or tubal origin and two cases of ovarian or peritoneal origin.

<sup>c</sup> For staging we used the FIGO classification valid at time of initial diagnosis.

in the normal FTE from patients with benign conditions ( $p < 2^{-16}$ , one-way analysis of variance [ANOVA] followed by Tukey's test). The median SOX2 expression significantly decreased in the corresponding, fully established ovarian tumors ( $p < 2^{-16}$ , one-way ANOVA followed by Tukey's test, Fig. 3a–b). We validated this observation on an independent set of fallopian tubes from 88 women with either HGSOCs ( $n = 42$ ) or benign gynecological conditions ( $n = 46$ ) and confirmed that a significant expansion of cells strongly expressing SOX2 was present in the benign FTE of women with HGSOCs ( $p < 2^{-16}$ , one-way ANOVA followed by Tukey's test, Fig. 3c). The magnitude of SOX2 overexpression was not significantly influenced by tumor stage (early versus late) or by whether or not the patient had chemotherapy prior to surgical excision (data not shown). In addition, fitting a binomial general linear model on the entire dataset revealed that strong SOX2 expression accurately predicted whether the fallopian tube was from an HGSOC patient or from a patient with benign pathology ( $p = 9.95^{-8}$ , logistic regression model). Fitting a receiver operating characteristic (ROC) curve to the data from the logistic regression analysis confirmed the high predictive power of strong SOX2 expression (Fig. 3d, area under the curve [AUC] = 0.87). This ubiquitous feature of the FTE of patients with HGSOCs has not been previously described. Furthermore, only eight out of 28 cases stained for p53 also showed a p53 signature indicating that SOX2 overexpression may occur prior to p53 dysregulation.

To test whether the expansion of SOX2-expressing cells occurred prior to HGSOC development, we analyzed the fallopian tubes of 48 women at high risk of developing ovarian cancer because they were BRCA1 or BRCA2 gene mutation carriers and therefore underwent prophylactic salpingo-oophorectomy (Fig. 3c, Table 2). This analysis showed that a significant expansion of cells strongly expressing SOX2 was present in the benign FTE of women who were BRCA1 or BRCA2 mutation carriers ( $p < 0.001$ , one-way ANOVA followed by Tukey's test).

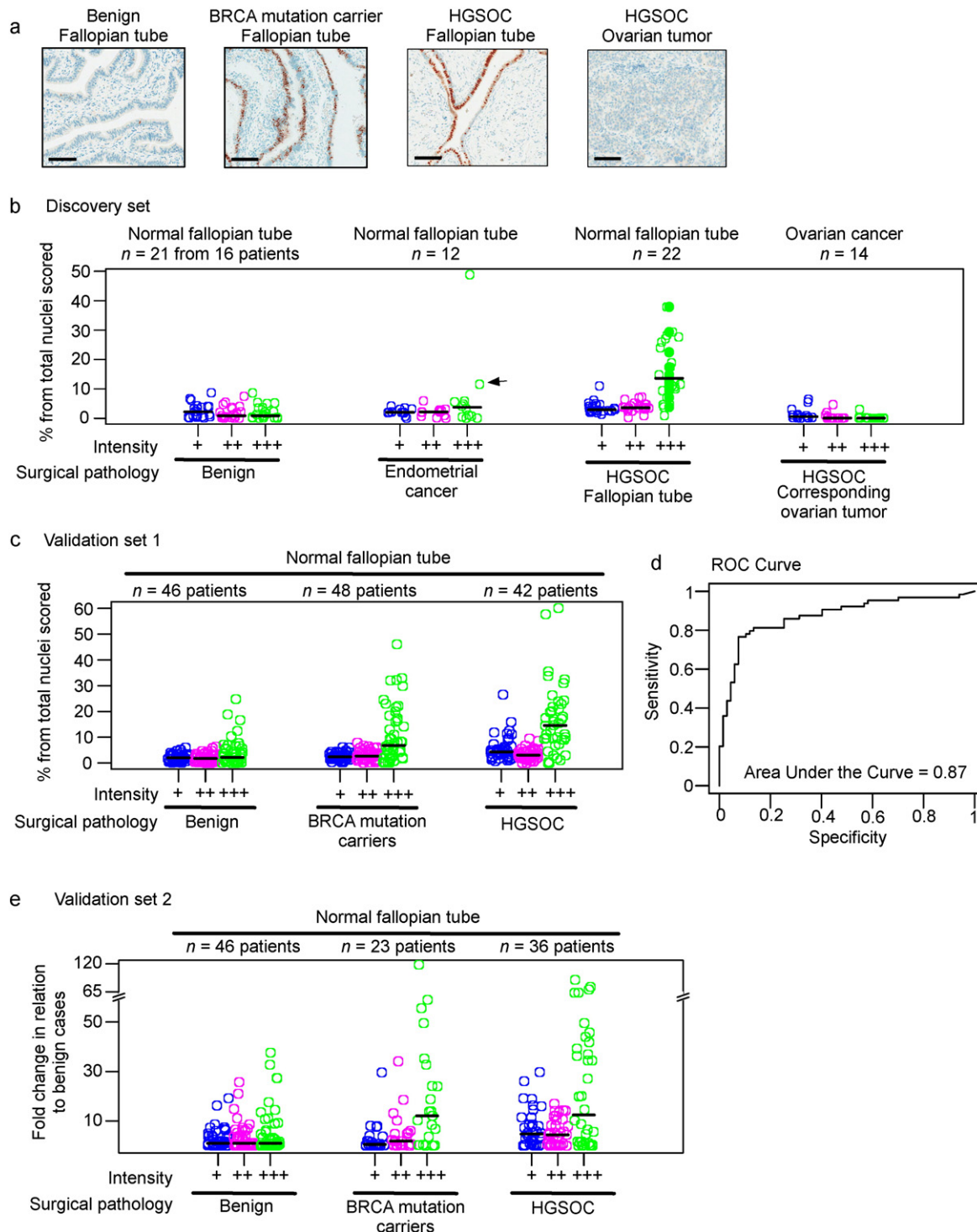
Finally, to confirm the above findings, we repeated the analysis on an additional cohort of 98 patients from three independent institutes (Fig. 3e, Table 2 and Supplementary Table 7). This analysis further confirmed that a significant expansion of cells strongly expressing SOX2 occurred in the benign FTE of women who are BRCA1 or BRCA2 mutation carriers and in women who had HGSOCs ( $p = 0.018$  and  $p < 0.001$ , respectively,

one-way ANOVA followed by Tukey's test). Expression of SOX2 in the normal FTE of BRCA gene mutation carriers was independent of age and whether the mutation was present in the BRCA1 or BRCA2 (data not shown). History of prior breast cancer was available for 43 cases who were either BRCA1 or BRCA2 mutation carriers. Of those only 15 patients had history of prior breast cancer. There was no significant difference in SOX2 expression between cases that either had or did not have breast cancer.

### 3.5. The BB5 Region Is a Repressor of SOX2 Expression

Specific sequences of non-coding regions in the genome are capable of regulating the expression of neighboring genes. Several lines of evidence strongly suggested that the BB5 region acted as a repressor of SOX2 expression. First, the BB5 mutation occurred in a 12-nucleotide DNA sequence (Supplementary Fig. S7A) that matched the previously reported regulatory motif model UW.0169 that was discovered by genomic DNaseI footprinting (Neph et al., 2012). We used an in-vivo reporter assay to test whether this motif changed the expression of a fluorescent protein and, therefore, was potentially functional (Supplementary Fig. S7B–E). Such standard reporter assays are traditionally performed in cultured cells. In this study we conducted the assay in live chicken embryos to gain additional information about the time and location of the activity of the BB5 region in regulating fluorescent protein (the reporter) expression. This analysis confirmed that the motif consisting of the BB5 nucleotide acted as a repressor and that mutation of the BB5 nucleotide significantly relieved this repressor activity (Supplementary Fig. S7B–E). Importantly, the protein fluorescence partially co-localized with cells that strongly expressed endogenous Sox2 in the chicken embryo and, therefore, implied that the element was plausible for regulating Sox2 expression. Second, we performed CRISPR-Cas9-mediated genome deletion, a method to remove a number of nucleotides from a target sequence in the genome, and deleted multiple essential nucleotides included in the motif model UW.0169 (Supplementary Fig. S7F) (Ran et al., 2013) in a human cell model. This induced a significant reduction in H3K27 acetylation, an indicator of open chromatin, and a significant and sustained increase in SOX2 expression in human cells (Supplementary Fig. S7G, Supplementary Table 6). Thus, this unequivocally confirmed that the element acted as a repressor of SOX2 expression in a human cell model. Third, mining the ENCODE data for potential transcription factor binding at this element revealed several factors that are known to be involved in regulating the stem cell state, such as YY1, NANOG, OCT4 and EZH2 (Supplementary Fig. S7F). Since EZH2, the catalytic subunit of the Polycomb repressor complex 2, is known to be involved in repressing gene expression by inducing H3K27 trimethylation at regulatory elements, we performed ChIP-sequencing on primary cultured fallopian tube epithelial cells to test whether EZH2 occupied the BB5 element or other elements that were sites of somatic mutations in other HGSOCs. This analysis confirmed that EZH2 occupied the BB5 element and several others (Supplementary Fig. S8A, Supplementary Table 11). In addition, the enzymatic inhibition of EZH2 resulted in a significant reduction in H3K27 tri-methylation (Supplementary Fig. S8B) indicating that the localization of the EZH2 at the BB5 element was indeed functional. In summary, the above data confirmed that the BB5 region is a repressor of SOX2 expression and, when mutated, releases SOX2 repression.

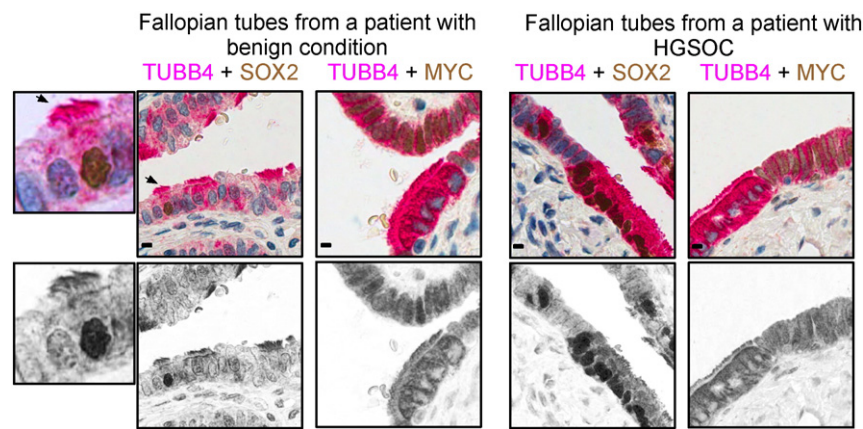
As previously mentioned, SOX2 is part of the four essential stem cells transcription factors (SOX2, OCT4, c-Myc, KLF4), also known as Yamanaka factors (Takahashi et al., 2007), which are highly expressed in embryonic stem cells and that are sufficient to induce pluripotency in human somatic cells. We, therefore, investigated the expression of the Yamanaka factors in the FTE and observed a mutually exclusive expression pattern of SOX2 and MYC. In the FTE of benign cases, MYC appeared to be exclusively and ubiquitously expressed in non-ciliated cells whilst SOX2 was only expressed at low levels in a minor population of



**Fig. 3.** Overexpression of SOX2 is a biomarker for pre-neoplastic detection of HGSOC. **a.** Representative SOX2 immunohistochemistry images for the FTE of the indicated samples are presented. Scale bars = 100  $\mu$ m. **b.** Percentages (y-axis), median percentage (horizontal black bars) and intensity (x-axis) of SOX2 staining in the normal FTE of women with benign conditions, endometrial cancer or HGSOCs and in the paired HGSOC tumors. Solid circles represent the FTE from cancers harboring rare variants and mutations in the BB5 region. The black arrow indicates a case of high-grade serous endometrial cancer. **c.** Power calculations were used to determine the required case number for the validation set based on the data from the discovery set (see Supplementary methods). Data from an independent set of 88 cases as well as an additional 48 BRCA1 or BRCA2 mutation carriers who underwent prophylactic excision of the fallopian tubes are presented. **d.** The receiver operating characteristics (ROC) curve is presented for the combined data presented in panels b and c. **e.** Data from a further independent set of cases from a second institute (CNIO) are presented. Data are presented as fold change relative to benign cases. A y-axis break was added to facilitate comparison with the data presented in b and c.

ciliated cells (Fig. 4). In contrast, in the FTE from HGSOC cases MYC maintained a similar expression pattern in non-ciliated cells, whereas SOX2 was strongly expressed in the majority of ciliated cells (Fig. 4, Supplementary Fig. S9A–B). Furthermore, the analysis of 209 HGSOCs revealed that MYC over-expression was accompanied by the loss of

SOX2 expression and vice versa ( $p < 0.001$ , Fisher Exact test, Supplementary Fig. S9C). Thus, SOX2 may repress MYC in differentiating FTE cells and in a fraction of SOX2-expressing tumors. Consistent with our proposed role of SOX2 in FTE cells, its ectopic expression significantly reduced expression of MYC and PAX8, an important transcription factor



**Fig. 4.** SOX2 represses MYC and PAX8 expression at the secretory-ciliated FTE cell interface. Representative immunohistochemistry images of FTE double stained with the indicated antibodies ( $n = 10$ ). TUBB4 was used as a marker for ciliated cells. Scale bars = 10  $\mu$ m. Arrows indicate cells that are presented in higher magnifications.

during embryogenesis of Müllerian tissues, in primary cultured FTE cells and SKOv3 ovarian cancer cells to undetectable levels by immunostaining (Supplementary Fig. S10). These findings are consistent with the known role of SOX2 in inducing pluripotency and may have important implications in understanding the genesis of epithelial ovarian cancers.

#### 4. Discussion

Less than a third of ovarian cancer patients are diagnosed at an early, hence potentially curable, stage. The currently available detection options are patient driven symptom awareness and screening using combination of a serum marker (CA125) and imaging by ultrasound methods. Advances have been made to divert from the single threshold rule for CA125 to more complex algorithms in order to improve performance and facilitate early treatment. Jacobs et al. (2016) recently reported results of the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS), a study in which over 200,000 postmenopausal women were enrolled to assess whether screening is effective for improving early ovarian cancer diagnosis and survival. Results indicated a limited mortality reduction in the primary analysis (Jacobs et al., 2016). This, again, underscores the need for effective biomarkers for early detection.

Recent work strongly suggests that the FTE is a common site of a large fraction of high-grade serous pelvic cancers (Perets et al., 2013). Inducing mutations in *TP53*, *BRCA1-BRCA2*, and *PTEN* in secretory cells in the FTE of mice resulted in the development of HGSOCs (Perets et al., 2013). However, there continues to be a need for novel molecular markers of premalignant transformation that could be used for early detection. Current screening methods only enable the detection of established invasive ovarian cancers. Mathematical modeling of tumor occurrence in *BRCA1-BRCA2* mutations carriers suggests that tumors remain occult for up to four years (Brown and Palmer, 2009). This duration represents an unutilized opportunity for premalignant or early stage disease detection because of the limitations of existing markers. In addition, patients with a family history of ovarian cancer who don't have affected living relatives are not eligible for *BRCA1-BRCA2* mutation testing. Furthermore, in many cases of familial ovarian malignancies *BRCA* testing is negative. Subsequently, these patients are ineligible for risk-reducing salpingo-oophorectomy.

Lack of specificity of *TP53* overexpression (the so called p53 signature) in the fallopian tube (Lee et al., 2007) undermines its potential use in early detection. In contrast, we have shown that SOX2 overexpression has high specificity and sensitivity as evidenced by the high AUC in our ROC analysis. In addition, unlike the p53 signature which only involves a limited number of cells, SOX2 overexpression occupies a large surface area of the fallopian tube. This makes SOX2 overexpression amenable for detection. Safe screening methods that exploit such

feature would need to be developed in both *BRCA* mutation carriers and patients with HGSOC. The use of cell-penetrating peptides or antibodies that specifically recognize SOX2 could be labeled with PET imaging tracers to enable early, non-invasive detection in patients. In addition, fallopian tube catheterization for obtaining cytobrush specimens for quantifying SOX2 expression is also possible. Such methods could be tested in clinical trials for early detection in young women with *BRCA1-BRCA2* mutations who do not wish to undergo prophylactic excision of both tubes and ovaries. Finally, as SOX2 is a potent transcription factor, it is highly likely that its aberrant expression is associated with co-expression of secreted proteins. The future discovery of such proteins could enable biomarker screening using blood sampling.

In summary, in this study we demonstrated that SOX2 overexpression occurs in a fraction of women with *BRCA1* and *BRCA2* mutations prior to ovarian cancer initiation and in the majority of patients with HGSOCs irrespective of tumor stage. These findings could be exploited for filling the current gap in early detection strategies for ovarian cancer. We believe that this is the first report of the expansion of SOX2-expressing cells in the FTE of HGSOCs. This finding has important implications, as it provides a potentially powerful tool for screening for HGSOCs. Utilizing our findings as potential biomarker should take high priority.

Supplementary data to this article can be found online at <http://dx.doi.org/10.1016/j.ebiom.2016.06.048>.

#### Funding Sources

This work is funded by the Medical Research Council (H8RSRS00), Ovarian Cancer Action (HER00070), the Oxford Biomedical Research Centre, the National Institute for Health Research (HJRWAC05) and the Experimental Cancer Medicine Centre. MJG is recipient of a research contract from the Instituto de Salud Carlos III of the Ministerio Español de Sanidad y Consumo (Miguel Servet tipo II Program, CP11 13-00047). C.Y. acknowledges the support of an MRC New Investigator Research Grant (Ref No. MR-L001411-1) and the Wellcome Trust Core Award Grant Number 090532-Z-09-Z.

#### Conflicts of Interests

R.T., B.A.P., R.D., D.M.H., R.Y.Z., R.K., and K.D. own stock in Complete Genomics, Inc. All other authors have no conflict of interest to declare.

#### Contributions

A.A.A. conceived the project. All other authors designed and conducted experiments. A.A.A., K.H., F.M., S.H.G., and C.Y. wrote the manuscript. All authors contributed to the final manuscript.



## Acknowledgements

We thank the Gynecological Oncology Multidisciplinary Team at Oxford, the Early Phase Clinical Trials Unit and the Oncology Clinical Trials Office at Oxford for help in recruiting patients to clinical trials. We thank the High-Throughput Genomics Group at the Wellcome Trust Centre for Human Genetics (funded by Wellcome Trust grant reference 090532-Z-09-Z and MRC Hub grant G0900747 91070) for their help with targeted sequencing and Stephen Hague for assistance with droplet digital PCR. We thank the Oxford Centre for Histopathology Research for their technical work and Hayley Davis for assisting with laser capture microdissection. We would like to acknowledge the support of all Complete Genomics employees who worked on the generation of libraries, reagents, and whole genome sequencing data for this project. We thank the CNIO Histopathology Core Unit for technical work. We thank the biobanks in CNIO (Ref. B.000848), MD Anderson (Ref. B.0000745) and IRBLleida (Ref. B.0000682) and the Spanish National Biobank Network Platform (Instituto de Salud Carlos III) for coordinating sample and clinical data retrieval.

## References

- Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., McVean, G.A., 2012. An integrated map of genetic variation from 1,092 human genomes. *Nature* 491, 56–65.
- Ahmed, A.A., Etemadmoghadam, D., Temple, J., Lynch, A.G., Riad, M., Sharma, R., Stewart, C., Feraday, S., CALDAS, C., Defazio, A., Bowtell, D., Brenton, J.D., 2010. Driver mutations in TP53 are ubiquitous in high grade serous carcinoma of the ovary. *J. Pathol.* 221, 49–56.
- Anon, 2011. Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609–615.
- Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., Snyder, M., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature* 489, 57–74.
- Besnier, C., Takeuchi, Y., Towers, G., 2002. Restriction of lentivirus in monkeys. *Proc. Natl. Acad. Sci. U. S. A.* 99, 11920–11925.
- Betancur, P., Bronner-Fraser, M., Sauka-Spengler, T., 2010. Genomic code for Sox10 activation reveals a key regulatory enhancer for cranial neural crest. *Proc. Natl. Acad. Sci. U. S. A.* 107, 3570–3575.
- Boumahdi, S., Driessens, G., Lapouge, G., Rorive, S., Nassar, D., Le Mercier, M., Delatte, B., Cawu, A., Lenglez, S., Nkusi, E., Brohee, S., Salmon, I., Dubois, C., Del Marmol, V., Fuks, F., Beck, B., Blanpain, C., 2014. SOX2 controls tumour initiation and cancer stem-cell functions in squamous-cell carcinoma. *Nature* 511, 246–250.
- Brown, P.O., Palmer, C., 2009. The preclinical natural history of serous ovarian cancer: defining the target for early detection. *PLoS Med.* 6, e1000114.
- Carnevali, P., Baccash, J., Halpern, A.L., Nazarenko, I., Nilsen, G.B., Pant, K.P., Ebert, J.C., Brownley, A., Morenzoni, M., Karpinchyk, V., Martin, B., Ballinger, D.G., Drmanac, R., 2012. Computational techniques for human genome resequencing using mated gapped reads. *J. Comput. Biol.* 19, 279–292.
- Cong, L., Ran, F.A., Cox, D., Lin, S., Barretto, R., Habib, N., Hsu, P.D., Wu, X., Jiang, W., Marraffini, L.A., Zhang, F., 2013. Multiplex genome engineering using CRISPR/Cas systems. *Science* 339, 819–823.
- Crum, C.P., 2009. Intercepting pelvic cancer in the distal fallopian tube: theories and realities. *Mol. Oncol.* 3, 165–170.
- Dean, F.B., Hosono, S., Fang, L., Wu, X., Faruqi, A.F., Bray-Ward, P., Sun, Z., Zong, Q., Du, Y., Du, J., Driscoll, M., Song, W., Kingsmore, S.F., Egholm, M., Lasken, R.S., 2002. Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl. Acad. Sci. U. S. A.* 99, 5261–5266.
- Drescher, C.W., Shah, C., Thorpe, J., O'Brian, K., Anderson, G.L., Berg, C.D., Urban, N., McIntosh, M.W., 2013. Longitudinal screening algorithm that incorporates change over time in CA125 levels identifies ovarian cancer earlier than a single-threshold rule. *J. Clin. Oncol.* 31, 387–392.
- Drmanac, R., Sparks, A.B., Callow, M.J., Halpern, A.L., Burns, N.L., Kermani, B.G., Carnevali, P., Nazarenko, I., Nilsen, G.B., Yeung, G., Dahl, F., Fernandez, A., Staker, B., Pant, K.P., Baccash, J., Borcharding, A.P., Brownley, A., Cedeno, R., Chen, L., Chernikoff, D., Cheung, A., Chirita, R., Curson, B., Ebert, J.C., Hacker, C.R., Hartlage, R., Hauser, B., Huang, S., Jiang, Y., Karpinchyk, V., Koenig, M., KONG, C., Landers, T., Le, C., Liu, J., McBride, C.E., Morenzoni, M., Morey, R.E., Mutch, K., Perazich, H., Perry, K., Peters, B.A., Peterson, J., Pethiyagoda, C.L., Pothuraju, K., Richter, C., Rosenbaum, A.M., Roy, S., Shafro, J., Sharanovich, U., Shannon, K.W., Sheppy, C.G., Sun, M., Thakuria, J.V., Tran, A., Vu, D., Zaranek, A.W., Wu, X., Drmanac, S., Oliphant, A.R., Banyai, W.C., Martin, B., Ballinger, D.G., Church, G.M., Reid, C.A., 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327, 78–81.
- Feng, J., Liu, T., Qin, B., Zhang, Y., Liu, X.S., 2012. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* 7, 1728–1740.
- Hamburger, V., Hamilton, H.L., 1992. A series of normal stages in the development of the chick embryo. 1951. *Dev. Dyn.* 195, 231–272.
- Jacobs, I.J., Menon, U., Ryan, A., Gentry-Maharaj, A., Burnell, M., Kalsi, J.K., Amso, N.N., Apostolidou, S., Benjamin, E., Cruickshank, D., Crump, D.N., Davies, S.K., Dawnay, A., Dobbs, S., Fletcher, G., Ford, J., Godfrey, K., Gunu, R., Habib, M., Hallett, R., Herod, J., Jenkins, H., Karpinskyj, C., Leeson, S., Lewis, S.J., Liston, W.R., Lopes, A., Mould, T., Murdoch, J., Oram, D., Rabideau, D.J., Reynolds, K., Scott, I., Seif, M.W., Sharma, A., Singh, N., Taylor, J., Warburton, F., Widschwendter, M., Williamson, K., Woolas, R., Fallowfield, L., McGuire, A.J., Campbell, S., Parmar, M., Skates, S.J., 2016. Ovarian cancer screening and mortality in the UK Collaborative Trial of Ovarian Cancer Screening (UKTOCS): a randomised controlled trial. *Lancet* 387, 945–956.
- Karst, A.M., Levanon, K., Drapkin, R., 2011. Modeling high-grade serous ovarian carcinogenesis from the fallopian tube. *Proc. Natl. Acad. Sci. U. S. A.* 108, 7547–7552.
- Kuleshov, V., Xie, D., Chen, R., Pushkarev, D., Ma, Z., Blauwkamp, T., Kertesz, M., Snyder, M., 2014. Whole-genome haplotyping using long reads and statistical methods. *Nat. Biotechnol.* 32, 261–266.
- Kurman, R.J., Shih, Ie, M., 2011. Molecular pathogenesis and extraovarian origin of epithelial ovarian cancer—shifting the paradigm. *Hum. Pathol.* 42, 918–931.
- Landt, S.G., Marinov, G.K., Kundaje, A., Kheradpour, P., Pauli, F., Batzoglou, S., Bernstein, B.E., Bickel, P., Brown, J.B., Cayting, P., Chen, Y., Desalvo, G., Epstein, C., Fisher-Aylor, K.I., Euskirchen, G., Gerstein, M., Gertz, J., Hartemink, A.J., Hoffman, M.M., Iyer, V.R., Jung, Y.L., Karmakar, S., Kellis, M., Kharchenko, P.V., Li, Q., Liu, T., Liu, X.S., Ma, L., Milosavljevic, A., Myers, R.M., Park, P.J., Pazin, M.J., Perry, M.D., Raha, D., Reddy, T.E., Rozowsky, J., Shores, N., Sidow, A., Slattery, M., Stamatiouyannopoulos, J.A., Tolstorukov, M.Y., White, K.P., Xi, S., Farnham, P.J., Lieb, J.D., Wold, B.J., Snyder, M., 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome Res.* 22, 1813–1831.
- Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* 9, 357–359.
- Lee, Y., Miron, A., Drapkin, R., Nucci, M.R., Medeiros, F., Saleemuddin, A., Garber, J., Birch, C., Mou, H., Gordon, R.W., Cramer, D.W., McKeon, F.D., Crum, C.P., 2007. A candidate precursor to serous carcinoma that originates in the distal fallopian tube. *J. Pathol.* 211, 26–35.
- Lunter, G., Goodson, M., 2011. Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Res.* 21, 936–939.
- McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schafer, B.T., Lowe, C.B., Wenger, A.M., Bejerano, G., 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.* 28, 495–501.
- Menon, U., Ryan, A., Kalsi, J., Gentry-Maharaj, A., Dawnay, A., Habib, M., Apostolidou, S., Singh, N., Benjamin, E., Burnell, M., Davies, S., Sharma, A., Gunu, R., Godfrey, K., Lopes, A., Oram, D., Herod, J., Williamson, K., Seif, M.W., Jenkins, H., Mould, T., Woolas, R., Murdoch, J.B., Dobbs, S., Amso, N.N., Leeson, S., Cruickshank, D., Scott, I., Fallowfield, L., Widschwendter, M., Reynolds, K., McGuire, A., Campbell, S., Parmar, M., Skates, S.J., Jacobs, I., 2015. Risk algorithm using serial biomarker measurements doubles the number of screen-detected cancers compared with a single-threshold rule in the united kingdom collaborative trial of ovarian cancer screening. *J. Clin. Oncol.* 33, 2062–2071.
- Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., Maurano, M.T., Humbert, R., Rynes, E., Wang, H., Vong, S., Lee, K., Bates, D., Diegel, M., Roach, V., Dunn, D., Neri, J., Schafer, A., Hansen, R.S., Kutayavin, T., Giste, E., Weaver, M., Canfield, T., Sabo, P., Zhang, M., Balasundaram, G., Byron, R., Maccoss, M.J., Akey, J.M., Bender, M.A., Groudine, M., Kaul, R., Stamatiouyannopoulos, J.A., 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. *Nature* 489, 83–90.
- Perets, R., Wyant, G.A., Muto, K.W., Bijron, J.G., Poole, B.B., Chin, K.T., Chen, J.Y., Ohman, A.W., Stepule, C.D., Kwak, S., Karst, A.M., Hirsch, M.S., Setlur, S.R., Crum, C.P., Dinulescu, D.M., DRAPKIN, R., 2013. Transformation of the fallopian tube secretory epithelium leads to high-grade serous ovarian cancer in Brca1;Pten models. *Cancer Cell* 24, 751–765.
- Peters, B.A., Kermani, B.G., Sparks, A.B., Alferov, O., Hong, P., Alexeev, A., Jiang, Y., Dahl, F., Tang, Y.T., Haas, J., Robasky, K., Zaranek, A.W., Lee, J.H., Ball, M.P., Peterson, J.E., Perazich, H., Yeung, G., Liu, J., Chen, L., Kennemer, M.L., Pothuraju, K., Konvicka, K., Tsoukpo-Sitnikow, M., Pant, K.P., Ebert, J.C., Nilsen, G.B., Baccash, J., Halpern, A.L., Church, G.M., Drmanac, R., 2012. Accurate whole-genome sequencing and haplotyping from 10 to 20 human cells. *Nature* 487, 190–195.
- Peters, B.A., Kermani, B.G., Alferov, O., Agarwal, M.R., McElwain, M.A., Gulbahce, N., Hayden, D.M., Tang, Y.T., Zhang, R.Y., Tearnle, R., Crain, B., Prates, R., Berkeley, A., Munne, S., Drmanac, R., 2015. Detection and phasing of single base de novo mutations in biopsies from human in vitro fertilized embryos by advanced whole-genome sequencing. *Genome Res.* 25, 426–434.
- Ran, F.A., Hsu, P.D., Wright, J., Agarwala, V., Scott, D.A., Zhang, F., 2013. Genome engineering using the CRISPR-Cas9 system. *Nat. Protoc.* 8, 2281–2308.
- Rimmer, A., Phan, H., Mathieson, I., Iqbal, Z., Twigg, S.R., Consortium, W. G. S., Wilkie, A.O., McVean, G., Lunter, G., 2014. Integrating mapping-, assembly- and haplotype-based approaches for calling variants in clinical sequencing applications. *Nat. Genet.* 46, 912–918.
- Sauka-Spengler, T., Barembaum, M., 2008. Gain- and loss-of-function approaches in the chick embryo. *Methods Cell Biol.* 87, 237–256.
- Simoes-Costa, M.S., McKeown, S.J., Tan-Cabugao, J., Sauka-Spengler, T., Bronner, M.E., 2012. Dynamic and differential regulation of stem cell factor FoxD3 in the neural crest is encrypted in the genome. *PLoS Genet.* 8, e1003142.
- Takahashi, K., Yamanaka, S., 2006. Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors. *Cell* 126, 663–676.
- Takahashi, K., Tanabe, K., Ohnuki, M., Narita, M., Ichisaka, T., Tomoda, K., Yamanaka, S., 2007. Induction of pluripotent stem cells from adult human fibroblasts by defined factors. *Cell* 131, 861–872.
- Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., John, S., Sandstrom, R., Bates, D., Boatman, L.K., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutayavin, T., Lajoie, B., Lee, B.K., Lee, K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Yan,

- Y., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Stamatoyannopoulos, G., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Akey, J.M., Sabo, P.J., Kaul, R., Furey, T.S., Dekker, J., Crawford, G.E., Stamatoyannopoulos, J.A., 2012. The accessible chromatin landscape of the human genome. *Nature* 489, 75–82.
- Yang, X., Boehm, J.S., Salehi-Ashtiani, K., Hao, T., Shen, Y., Lubonja, R., Thomas, S.R., Alkan, O., Bhimdi, T., Green, T.M., Johannessen, C.M., Silver, S.J., Nguyen, C., Murray, R.R., Hieronymus, H., Balcha, D., Fan, C., Lin, C., Ghamsari, L., Vidal, M., Hahn, W.C., Hill, D.E., Root, D.E., 2011. A public genome-scale lentiviral expression library of human ORFs. *Nat. Methods* 8, 659–661.
- Yip, K.Y., Cheng, C., Bhardwaj, N., Brown, J.B., Leng, J., Kundaje, A., Rozowsky, J., Birney, E., Bickel, P., Snyder, M., Gerstein, M., 2012. Classification of human genomic regions based on experimentally determined binding sites of more than 100 transcription-related factors. *Genome Biol.* 13, R48.