

Supplementary Materials for

Evaluation of DNA from the Papanicolaou Test to Detect Ovarian and Endometrial Cancers

Isaac Kinde, Chetan Bettegowda, Yuxuan Wang, Jian Wu, Nishant Agrawal, Ie-Ming Shih, Robert Kurman, Fanny Dao, Douglas A. Levine, Robert Giuntoli, Richard Roden, James R. Eshleman, Jesus Paula Carvalho, Suely Kazue Nagahashi Marie, Nickolas Papadopoulos, Kenneth W. Kinzler, Bert Vogelstein,* Luis A. Diaz Jr.*

*To whom correspondence should be addressed. E-mail: bertvog@gmail.com (B.V.); ldiaz1@jhmi.edu (L.A.D.)

Published 9 January 2013, *Sci. Transl. Med.* 5, 167ra4 (2013)
DOI: 10.1126/scitranslmed.3004952

This PDF file includes:

Materials and Methods

Other Supplementary Material for this manuscript includes the following:
(available at www.sciencetranslationalmedicine.org/cgi/content/full/5/167/167ra4/DC1)

Table S1 (Microsoft Excel format). Summary characteristics of endometrial cancers (endometrioid subtype) studied by whole-exome sequencing.
Table S2 (Microsoft Excel format). Mutations identified by whole-exome sequencing in 22 endometrioid endometrial cancers.
Table S3 (Microsoft Excel format). Clinical characteristics and mutations assessed in Pap specimens.
Table S4 (Microsoft Excel format). Primers used to assess individual mutations in Pap specimens.
Table S5 (Microsoft Excel format). Primers used to simultaneously assess 12 genes in Pap specimens with the multiplexed Safe-SeqS strategy.
Table S6 (Microsoft Excel format). Mutations identified in Pap specimens through simultaneous assessment of 12 genes with the multiplexed Safe-SeqS strategy.

Supplementary Materials and Methods

Patient Samples

All samples for this study were obtained according to protocols approved by the Institutional Review Boards of The Johns Hopkins Medical Institutions (Baltimore, MD), Memorial Sloan Kettering Cancer Center (New York, NY), University of Sao Paulo (Sao Paulo, Brazil), and ILSbio, LLC (Chestertown, MD). Demographic, clinical and pathologic staging data were collected for each case. All histopathology was centrally re-reviewed by board-certified pathologists. Staging was based on 2009 FIGO criteria (46).

Fresh-frozen tissue specimens of surgically resected neoplasms of the ovary and endometrium were assessed for neoplastic cellularity by a board-certified pathologist. Serial frozen sections were used to guide the trimming of Optimal Cutting Temperature (OCT) compound embedded frozen tissue blocks to enrich the fraction of neoplastic cells for DNA extraction.

Formalin-fixed paraffin embedded (FFPE) tissue samples were assessed by a board-certified pathologist (Propath LLC, Dallas, TX) for tumor cellularity and to demarcate areas of high tumor cellularity. Tumor tissues from serial 10 micron sections on slides from the original tumor block were macrodissected with a razorblade to enrich the fraction of neoplastic cells for DNA extraction.

The source of normal DNA was matched whole blood or non-neoplastic normal adjacent tissue.

Liquid-based Pap smear specimens were collected with cervical brushes and transport medium from a Digene HC2 DNA Collection Device (Qiagen) or ThinPrep 2000 System (Hologic) and stored according to the manufacturer's recommendations.

Unless otherwise indicated, all patient-related values are reported as means $\pm$ 1 standard deviation.

DNA Extraction

DNA was purified from tumor and normal tissue, as well as Pap specimens, with an AllPrep kit (Qiagen) according to the manufacturer's instructions. DNA was purified from tumor tissue by adding 3 mL RLTM buffer (Qiagen) and then binding to an AllPrep DNA column (Qiagen) according to the manufacturer's protocol. DNA was purified from Pap specimens by adding five volumes of RLTM buffer when the amount of liquid was $\leq$ 3 mL. When the amount of liquid was $>$ 3 mL, the cells and cell fragments were pelleted at 1,000 x g for five minutes and the pellets were dissolved in 3 mL RLTM buffer. DNA was quantified in all cases with quantitative PCR, employing the primers and conditions previously described (47).

Microsatellite instability testing

Microsatellite instability was detected with the MSI Analysis System (Promega), containing five mononucleotide repeats (BAT-25, BAT-26, NR-21, NR-24 and MONO-

27) and two pentanucleotide repeat loci, per the manufacturer's instructions. After amplification, the fluorescent PCR products were sized on an Applied Biosystems 3130 capillary electrophoresis instrument (Invitrogen). Tumor samples were designated as follows: MSI-high if two or more mononucleotides varied in length compared to the germline DNA; MSI-low if only one locus varied; and microsatellite stable (MSS) if there was no variation compared to the germline. Pentanucleotide loci confirmed identity in all cases.

Preparation and sequence capture of Illumina DNA libraries

Overview. Preparation of Illumina genomic DNA libraries for exomic and targeted DNA captures was performed according to the manufacturer's recommendations. Briefly, 1-3 µg of genomic DNA was used for library preparation with the TruSeqDNA Sample Preparation Kit (Illumina). The DNA was acoustically sheared (Covaris) to a target size of ~200 bp. The fragments were subsequently end-repaired to convert overhangs into blunt ends. A single "A" nucleotide was then added to the 3' ends of blunt fragments to later prevent them from self-ligation; a corresponding "T" on the 3' end of adaptor molecules provided the complementary overhang. After ligation to adaptors, the library was amplified with 8-14 cycles of PCR to ensure yields of 0.5 and 4 µg for exomic and targeted gene captures, respectively.

Exomic enrichment. Exomic capture was performed with the SureSelect Human Exome Kit V 4.0 (Agilent) according to the manufacturer's protocol with the addition of TruSeq index-specific blocks in the hybridization mixture
(AGATCGGAAGAGCACACGTCTGAACTCCAGTCAC-XXXXXX-

ATCTCGTATGCCGTCTTCTGCTTGT, where the six base “XXXXXX” region denotes one of 12 sample-specific indexes).

Targeted gene enrichment. Targeted gene enrichment was performed by modifications of previously described methods (48, 49). In brief, targeted regions of selected oncogenes and tumor suppressor genes were synthesized as oligonucleotide probes by Agilent Technologies. Probes of 36 bases, tiled in 3-base intervals (that is, 33-base overlaps), were designed to capture both the plus and the minus strands of the DNA. The oligonucleotides were cleaved from the chip by incubating with 3 mL of 35% ammonium hydroxide at room temperature for five hours. The solution was transferred to two 2 mL tubes, dried under vacuum, and redissolved in 400 µL of ribonuclease (RNase)- and deoxyribonuclease (DNase)-free water. Five microliters of the solution was used for PCR amplification with primers complementary to the 12-base sequence common to all probes: 5'-TGATCCCGCGACGA*C-3' and 5'-GACCGCGACTCCAG*C-3', with * indicating a phosphorothioate bond. The PCR products were purified with a MinElute Purification Column (Qiagen), end-repaired with End-IT DNA End-Repair Kit (Epicentre), and then purified again with a MinElute Purification Column. The PCR products were ligated to form concatamers as described (48). The major differences between the protocol previously described (48, 49) and the one used in the present study involved the amplification of the ligated PCR products and the solid phase capture method as noted below. Biotinylated capture probes were prepared by amplifying 50 ng of ligated PCR products with the REPLI-g Midi Kit (Qiagen) supplemented with 2.5 nmol Biotin-16-dUTP (Roche) in a 27.5 µL reaction incubated at 30°C for 16 hours. After inactivating the polymerase by incubating at 65°C for three mins, the biotinylated

capture probes were purified with QIAquick PCR Purification Columns. For capture, 4-5 µg of library DNA was incubated with 1 µg of the prepared probes in a hybridization mixture as previously described (48). The biotinylated capture probes and captured library sequences were subsequently purified with 500 µg Dynabeads MyOne Streptavidin C1 (Invitrogen). After washing as per the manufacturer's recommendations, the captured sequences were eluted with 0.1 M NaOH and then neutralized with 1 M Tris-HCl (pH 7.5). Neutralized DNA was desalted and concentrated with a MinElute Purification Column using a 20 µL elution volume. The eluate was amplified in a 100 µL Phusion Hot Start II (Thermo Scientific) reaction containing 1X Phusion HF buffer, 0.25 mM dNTPs, 0.5 µM each forward and reverse TruSeq primers, and 2 U polymerase with the following cycling conditions: 98°C for 30 s; 14 cycles of 98°C for 10 s, 60°C for 30 s, and 72°C for 30 s; and 72°C for 5 min. The amplified pool containing enriched target sequences was purified with an Agencourt AMPure XP system (Beckman) and quantified with a 2100 Bioanalyzer (Agilent).

Massively parallel sequencing and somatic mutation identification

After capture of targeted sequences, paired-end sequencing with an Illumina GA IIx Genome Analyzer provided 2 x 75 base reads from each fragment. The sequence tags that passed filtering were aligned to the human genome reference sequence (hg18) and subsequent variant-calling analysis was performed with the ELANDv2 algorithm in the CASAVA 1.6 software (Illumina). Known polymorphisms recorded in dbSNP Build 130 (50) were removed from the analysis. Identification of high confidence mutations was performed as described previously (30).

Assessment of low-frequency mutations

Primer Design. We attempted to design primer pairs to detect mutations in the 46 cancers described in the text. Primers were designed as described previously (37) with Primer3 (51). Sixty percent of the primer pairs amplified the expected fragments; in the other 40%, a second or third set of primer pairs had to be designed to reduce primer dimers or non-specific amplification.

Sequencing Library Preparation. Templates were prepared for sequencing as described previously (37), with the modifications noted below that facilitated the amplification of multiple gene regions in a single well of a 96-well PCR plate. In brief, each strand of each template molecule was encoded with a 14 base unique identifier (UID) – comprised of degenerate “N” bases (equal probability of being an “A,” “C,” “G,” or “T”) - with two to four cycles of amplicon-specific PCR (“UID assignment PCR cycles,” see Fig. 2). Both forward and reverse gene-specific primers contained universal tag sequences at their 5' ends, providing the primer binding sites for the second-round amplification, but only the forward primer contained the UID, which was positioned between the 5' universal tag and the 3' gene-specific sequences (Fig. 2). Four “N” bases were additionally included in the reverse primer to facilitate sequencing analysis of paired-end libraries (table S4). The UID assignment PCR cycles were performed on 66 ng of DNA in a 50 μ L reaction containing 1X Phusion HF buffer, 0.25 mM dNTPs, 0.5 μ M each of forward (containing the 14 “N” bases) and reverse primers, and 2 U of Phusion Hot Start II Polymerase (Thermo Scientific). Carryover of residual UID-containing primers to the second-round amplification, which can complicate

template quantification (37), was minimized through a 15 min exonuclease digestion at 37⁰C to degrade unincorporated primers. In Kinde *et al.* (37), Exonuclease-I (Enzymatics) was chosen to eliminate the residual UID-containing primers, however we found that a different exonuclease – RecJ_f (New England Biolabs) – followed by purification with AMPure XP beads (Beckman) and elution in 10 µL of TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0), more extensively removed the UID-containing primers and yielded more robust amplification products. The eluted templates were amplified in a second-round PCR with primers containing the grafting sequences necessary for hybridization to the Illumina GA IIx flow cell at their 5' ends (Fig. 2). The reverse amplification primer additionally contained an index sequence between the 5' grafting and 3' universal tag sequences to enable the PCR products from multiple individuals to be simultaneously analyzed in the same flow cell compartment of the sequencer (37). The second-round amplification reactions contained 1X Phusion HF buffer, 0.25 mM dNTPs, 0.5 µM each of forward and reverse primers, and 2 U of Phusion Hot Start II Polymerase in a total of 50 µL. After an initial heat activation step at 98⁰ C for 2 minutes, 23 cycles of PCR were performed with the following cycling conditions: 98⁰C for 10 s, 65⁰C for 15 s, and 72⁰C for 15 s. The multiplexed assay was performed in similar fashion utilizing six independent amplifications – each containing 66 ng of DNA (that is, ~400 ng total) – per sample with the primers described in table S5. The PCR products were purified with AMPure XP beads and used directly for sequencing on either Illumina MiSeq or GA IIx instruments, with equivalent results.

Data Analysis. High quality sequence reads were analyzed as previously described (37). Briefly, we selected reads that contained high quality basecalls in their UID region

(that is, the first 14 cycles) by utilizing the quality scores generated by the sequencing instrument, which indicate the probability that an individual base call was made in error (52). Reads in which each of the 14 bases comprising the UID (representing one original template strand; see Fig. 2) had a quality score ≥ 15 were grouped by their UIDs and only the UIDs supported by more than one read were retained for further analysis. The template-specific portion of the reads that contained the sequence of an expected amplification primer was matched to a reference sequence set with a custom script (available from the authors upon request). Artifactual mutations – introduced during the sample preparation or sequencing steps – were reduced by requiring that $>50\%$ of reads sharing the same UID contained the identical mutation (a “supermutant,” Fig. 2). For the 46 assays querying a single amplicon, we required that the fraction of mutant alleles was significantly different from the background mutation levels determined from a negative control ($P < 0.001$, binomial test). As mutations are not known *a priori* in a screening environment, we used a more agnostic metric to detect mutations in the multiplexed assay. A threshold supermutant frequency was defined for each sample as the sum of the mean frequency of supermutants – including only those with average quality scores ≥ 25 and excluding those at positions with relatively high-frequency variants (such as SNPs or any mutation present at proportions $\geq 1\%$) – and six standard deviations of the mean. Only supermutants exceeding this threshold and representing at least two template molecules were designated as mutations and reported in Fig. 4 and table S6.