Innovation, Sciences et Développement économique Canada
Office de la Propriété Intellectuelle du Canada

Innovation, Science and Economic Development Canada
Canadian Intellectual Property Office

CA 2889937 C 2020/12/29

(11)(21) 2 889 937

(12) BREVET CANADIEN
CANADIAN PATENT

(13) C

(86) Date de dépôt PCT/PCT Filing Date: 2013/10/17
(87) Date publication PCT/PCT Publication Date: 2014/05/08
(45) Date de délivrance/Issue Date: 2020/12/29
(85) Entrée phase nationale/National Entry: 2015/04/29
(86) N° demande PCT/PCT Application No.: US 2013/065342
(87) N° publication PCT/PCT Publication No.: 2014/070462
(30) Priorité/Priority: 2012/10/29 (US61/719,942)

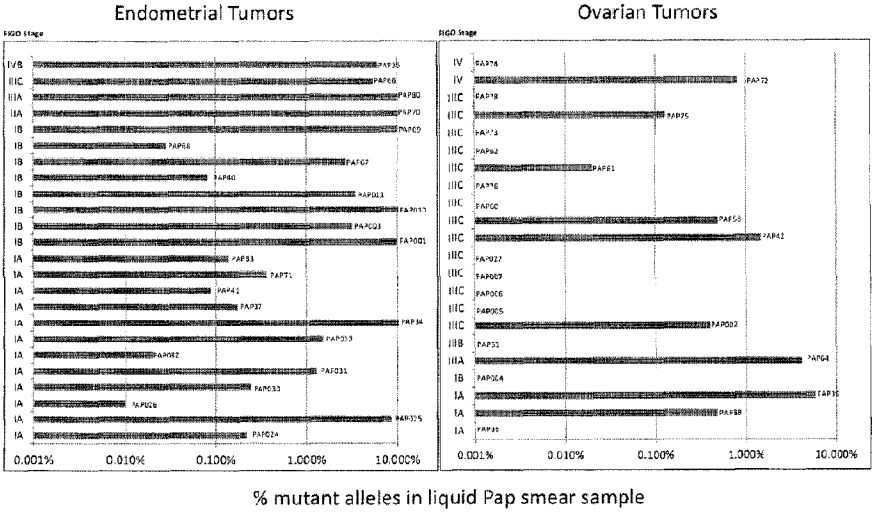
(51) Cl.Int./Int.Cl. *G01N 33/574* (2006.01),
G01N 33/68 (2006.01)

(72) Inventeurs/Inventors:
KINDE, ISAAC, US;
KINZLER, KENNETH W., US;
VOGELSTEIN, BERT, US;
PAPADOPOULOS, NICKOLAS, US;
DIAZ, LUIS, US;
BETTEGOWDA, CHETAN, US;
WANG, YUXUAN, US

(73) Propriétaire/Owner:
THE JOHNS HOPKINS UNIVERSITY, US

(74) Agent: TORYS LLP

(54) Titre : TEST DE PAPANICOLAOU POUR LES CANCERS DE L'OVAIRE ET DE L'ENDOMETRE
(54) Title: PAPANICOLAOU TEST FOR OVARIAN AND ENDOMETRIAL CANCERS



(57) Abrégé/Abstract:
The recently developed liquid-based Papanicolaou (Pap) smear allows not only cytologic evaluation but also collection of DNA for detection of HPV, the causative agent of cervical cancer. We tested these samples to detect somatic mutations present in rare tumor cells that might accumulate in the cervix once shed from endometrial and ovarian cancers. A panel of commonly mutated genes in endometrial and ovarian cancers was assembled and used to identify mutations in all 46 endometrial or cervical cancer tissue samples. We were able also able to identify the same mutations in the DNA from liquid Pap smears in 100% of endometrial cancers (24 of 24) and in 41% of ovarian cancers (9 of 22). We developed a sequence-based method to query mutations in 12 genes in a single liquid Pap smear without prior knowledge of the tumor's genotype.

(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(19) World Intellectual Property Organization
International Bureau

(43) International Publication Date
8 May 2014 (08.05.2014)

WIPO | PCT

(10) International Publication Number
WO 2014/070462 A1

(51) International Patent Classification:
G01N 33/574 (2006.01) *C12Q 1/68* (2006.01)
G01N 33/68 (2006.01)

(21) International Application Number:
PCT/US2013/065342

(22) International Filing Date:
17 October 2013 (17.10.2013)

(25) Filing Language: English

(26) Publication Language: English

(30) Priority Data:
61/719,942 29 October 2012 (29.10.2012) US

(71) Applicant: **THE JOHNS HOPKINS UNIVERSITY** [US/US]; 3400 N. Charles Street, Baltimore, MD 21218 (US).

(72) Inventors: **KINDE, Isaac**; 394 Mesa Verde Park, Beaumont, California 92223 (US). **KINZLER, Kenneth W.**; 616 Ponte Villas North, Baltimore, Maryland 21230 (US). **VOGELSTEIN, Bert**; 3700 Breton Way, Baltimore, Maryland 21208 (US). **PAPADOPOULOS, Nickolas**; 606 Horncrest Rd, Towson, Maryland 21204 (US). **DIAZ, Luis**; 5135 Crystal Springs, Ellicott City, Maryland 21043 (US). **BETTEGOWDA, Chetan**; 5239 Morning Dove Way, Perry Hall, Maryland 21128 (US). **WANG, Yuxuan**; 4004B Linkwood Road, Baltimore, Maryland 21210 (US).

(74) Agent: **KAGAN, Sarah A.**; Banner & Witcoff, Ltd., 1100 13th Street, N.W., Suite 1200, Washington, District of Columbia 20005-4051 (US).

(81) Designated States *(unless otherwise indicated, for every kind of national protection available)*: AE, AG, AL, AM, AO, AT, AU, AZ, BA, BB, BG, BH, BN, BR, BW, BY, BZ, CA, CH, CL, CN, CO, CR, CU, CZ, DE, DK, DM, DO, DZ, EC, EE, EG, ES, FI, GB, GD, GE, GH, GM, GT, HN, HR, HU, ID, IL, IN, IR, IS, JP, KE, KG, KN, KP, KR, KZ, LA, LC, LK, LR, LS, LT, LU, LY, MA, MD, ME, MG, MK, MN, MW, MX, MY, MZ, NA, NG, NI, NO, NZ, OM, PA, PE, PG, PH, PL, PT, QA, RO, RS, RU, RW, SA, SC, SD, SE, SG, SK, SL, SM, ST, SV, SY, TH, TJ, TM, TN, TR, TT, TZ, UA, UG, US, UZ, VC, VN, ZA, ZM, ZW.

(84) Designated States *(unless otherwise indicated, for every kind of regional protection available)*: ARIPO (BW, GH, GM, KE, LR, LS, MW, MZ, NA, RW, SD, SL, SZ, TZ, UG, ZM, ZW), Eurasian (AM, AZ, BY, KG, KZ, RU, TJ, TM), European (AL, AT, BE, BG, CH, CY, CZ, DE, DK, EE, ES, FI, FR, GB, GR, HR, HU, IE, IS, IT, LT, LU, LV, MC, MK, MT, NL, NO, PL, PT, RO, RS, SE, SI, SK, SM, TR), OAPI (BF, BJ, CF, CG, CI, CM, GA, GN, GQ, GW, KM, ML, MR, NE, SN, TD, TG).

Published:

— *with international search report (Art. 21(3))*

[Continued on next page]

(54) Title: PAPANICOLAOU TEST FOR OVARIAN AND ENDOMETRIAL CANCERS

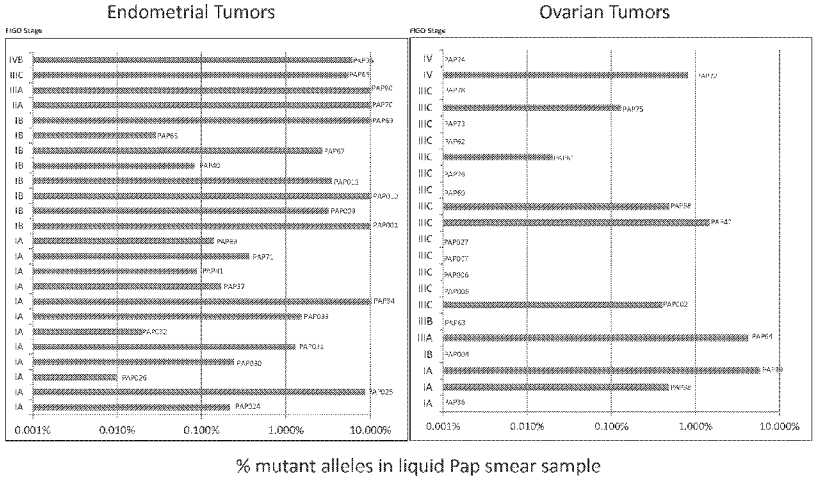


Fig. 3

(57) Abstract: The recently developed liquid-based Papanicolaou (Pap) smear allows not only cytologic evaluation but also collection of DNA for detection of HPV, the causative agent of cervical cancer. We tested these samples to detect somatic mutations present in rare tumor cells that might accumulate in the cervix once shed from endometrial and ovarian cancers. A panel of commonly mutated genes in endometrial and ovarian cancers was assembled and used to identify mutations in all 46 endometrial or cervical cancer tissue samples. We were able also able to identify the same mutations in the DNA from liquid Pap smears in 100% of endometrial cancers (24 of 24) and in 41% of ovarian cancers (9 of 22). We developed a sequence-based method to query mutations in 12 genes in a single liquid Pap smear without prior knowledge of the tumor's genotype.

WO 2014/070462 A1

WO 2014/070462 A1

— *before the expiration of the time limit for amending the claims and to be republished in the event of receipt of amendments (Rule 48.2(h))*

— *with sequence listing part of description (Rule 5.2(a))*

PAPANICOLAOU TEST FOR OVARIAN AND ENDOMETRIAL CANCERS

TECHNICAL FIELD OF THE INVENTION

[02] This invention is related to the area of cancer screening. In particular, it relates to ovarian and endometrial cancers.

BACKGROUND OF THE INVENTION

[03] Since the introduction of the Papanicolaou test, the incidence and mortality of cervical cancer in screened populations has been reduced by more than 75% (1, 2). In contrast, deaths from ovarian and endometrial cancers have not substantially decreased during that same time period. As a result, more than 69,000 women in the U.S. will be diagnosed with ovarian and endometrial cancer in 2012. Although endometrial cancer is more common than ovarian cancer, the latter is more lethal. In the U.S., approximately 15,000 and 8,000 women are expected to die each year from ovarian and endometrial cancers, respectively (Table 1). World-wide, over 200,000 deaths from these tumors are expected this year alone (3, 4).

[04] In an effort to replicate the success of cervical cancer screening, several approaches for the early detection of endometrial and ovarian cancers have been devised. For endometrial cancers, efforts have focused on cytology and transvaginal ultrasound (TVS). Cytology can indeed indicate a neoplasm within the uterus in some cases, albeit with low specificity (5). TVS is a noninvasive technique to measure the thickness of the endometrium based on the fact that endometria harboring a cancer are thicker than normal endometria (6). As with cytology, screening measurement of the endometrial thickness using TVS lacks sufficient specificity because benign lesions, such as polyps, can also result in a thickened endometrium. Accordingly, neither cytology nor TVS fulfills the requirements for a screening test (5, 7).

[05] Even greater efforts have been made to develop a screening test for ovarian cancer, using serum CA-125 levels and TVS. CA-125 is a high molecular weight transmembrane

glycoprotein expressed by coelomic- and Müllerian-derived epithelia that is elevated in a subset of ovarian cancer patients with early stage disease (8). The specificity of CA-125 is limited by the fact that it is also elevated in a variety of benign conditions, such as pelvic inflammatory disease, endometriosis and ovarian cysts (9). TVS can visualize the ovary but can only detect large tumors and cannot definitively distinguish benign from malignant tumors. Several clinical screening trials using serum CA-125 and TVS have been conducted but none has shown a survival benefit. In fact, some have shown an increase in morbidity compared to controls because false positive tests elicit further evaluation by laparoscopy or exploratory laparotomy (10-12).

[06] Accordingly, the U.S. Preventative Services Task Force, the American Cancer Society, the American Congress of Obstetricians and Gynecologists, as well as the National Comprehensive Cancer Network, do not recommend routine screening for endometrial or ovarian cancers in the general population. In fact, these organizations warn that "the potential harms outweigh the potential benefits" (13-16). An exception to this recommendation has been made for patients with a hereditary predisposition to ovarian cancer, such as those with germline mutations in a BRCA gene or those with Lynch syndrome. It is recommended that BRCA mutation carriers be screened every 6 months with TVS and serum CA-125, starting at a relatively early age. Screening guidelines for women with Lynch syndrome include annual endometrial sampling and TVS beginning between age 30 and 35 (15, 17).

[07] The mortality associated with undetected gynecologic malignancies has made the development of an effective screening tool a high priority. An important observation that inspired the current study is that asymptomatic women occasionally present with abnormal glandular cells (AGCs) detected in a cytology specimen as part of their routine cervical cancer screening procedure. Although AGCs are associated with premalignant or malignant disease in some cases (18-22), it is often difficult to distinguish the AGCs arising from endocervical, endometrial or ovarian cancer from one another or from more benign conditions. There is a continuing need in the art to detect these cancers at an earlier stage than done currently.

SUMMARY OF THE INVENTION

[08] According to one aspect of the invention a method is provided for detecting or monitoring endometrial or ovarian cancer. A liquid Pap smear of a patient is tested for a

genetic or epigenetic change in one or more genes, mRNAs, or proteins mutated in endometrial or ovarian cancer. Detection of the change indicates the presence of such a cancer in the patient.

[09] According to another aspect of the invention a method is provided for screening for endometrial and ovarian cancers. A liquid Pap smear is tested for one or more mutations in a gene, mRNA, or protein selected from the group consisting of *CTNNB1, EGFR, PI3KCA, PTEN, TP53, BRAF, KRAS, AKT1, NRAS, PPP2R1A, APC, FBXW7, ARID1A, CDKN2A, MLL2, RFF43,* and *FGFR2.* Detection of the mutation indicates the presence of such a cancer in the patient.

[10] Another aspect of the invention is a kit for testing a panel of genes in Pap smear samples for ovarian or endometrial cancers. The kit comprises at least 10 probes or at least 10 primer pairs. Each probe or primer comprises at least 15 nt of complementary sequence to one of the panel of genes. At least 10 different genes are interrogated. The panel is selected from the group consisting of *CTNNB1, EGFR, PI3KCA, PTEN, TP53, BRAF, KRAS, AKT1, NRAS, PPP2R1A, APC, FBXW7, ARID1A, CDKN2A, MLL2, RFF43,* and *FGFR2.*

[11] Still another aspect of the invention is a solid support comprising at least 10 attached probes. Each probe comprises at least 15 nt of complementary sequence to one of a panel of genes, wherein the panel is selected from the group consisting of *CTNNB1, EGFR, PI3KCA, PTEN, TP53, BRAF, KRAS, AKT1, NRAS, PPP2R1A, APC, FBXW7, ARID1A, CDKN2A, MLL2, RFF43,* and *FGFR2.*

[12] Another aspect of the invention is a solid support comprising at least 10 primers attached thereto. Each primer comprises at least 15 nt of complementary sequence to one of a panel of genes. The panel is selected from the group consisting of *CTNNB1, EGFR, PI3KCA, PTEN, TP53, BRAF, KRAS, AKT1, NRAS, PPP2R1A, APC, FBXW7, ARID1A, CDKN2A, MLL2, RFF43,* and *FGFR2.*

[13] These and other embodiments which will be apparent to those of skill in the art upon reading the specification provide the art with methods for assessing ovarian and endometrial cancers in a screening environment using samples that are already routinely collected.

BRIEF DESCRIPTION OF THE DRAWINGS

[14] Fig. 1. Schematic demonstrating the principle steps of the procedure described in this study. Tumors cells shed from ovarian or endometrial cancers are carried into the endocervical canal. These cells can be captured by the brush used for performing a routine Pap smear. The brush contents are transferred into a liquid fixative, from which DNA is isolated. Using next-generation sequencing, this DNA is queried for mutations that indicate the presence of a malignancy in the female reproductive tract.

[15] Fig. 2. Diagram of the assay used to simultaneously detect mutations in 12 different genes. A modification of the Safe-SeqS (Safe-Sequencing System) protocol, for simultaneous interrogation of multiple mutations in a single sample, is depicted. In the standard Safe-SeqS procedure, only one amplicon is assessed, while the new system is used to assess multiple amplicons from multiple genes at once.

[16] Fig. 3. Mutant allele fractions in Pap smear fluids. The fraction of mutant alleles from each of 46 pap smear fluids is depicted. The stage of each tumor is listed on the Y-axis. The X-axis demonstrates the % mutant allele fraction as determined by Safe-SeqS.

[17] Fig. 4. Heat map depicting the results of multiplex testing of 12 genes in Pap smear fluids. Each block on the y-axis represents a 30-bp block of sequence from the indicated gene. The 28 samples assessed (14 from women with cancer, 14 from normal women without cancer) are indicated on the x-axis. Mutations are indicated as colored blocks, with white indicating no mutation, yellow indicating a mutant fraction of 0.1% to 1%, orange indicate a mutant fraction of 1% to 10%, and red indicating a mutant fraction of >10%.

[18] Fig. 5 PRIOR ART. Essential elements of Safe-SeqS. In the first step, each fragment to be analyzed is assigned a unique identification (UID) DNA sequence (green or blue bars). In the second step, the uniquely tagged fragments are amplified, producing UID families, each member of which has the same UID. A supermutant is defined as a UID family in which ≥95% of family members have the same mutation.

[19] Fig. 6 PRIOR ART. Safe-SeqS with endogenous UIDs plus capture. The sequences of the ends of each fragment produced by random shearing (variously colored bars) serve as the unique identifiers (UIDs). These fragments are ligated to adapters (yellow and orange bars) so they can subsequently be amplified by PCR. One uniquely identifiable fragment is produced from each strand of the double-stranded template; only one strand is shown. Fragments of interest are captured on a solid phase containing oligonucleotides complementary to the sequences of interest. Following PCR amplification to produce UID families with primers containing 5′ "grafting" sequences (black and red bars), sequencing is performed and supermutants are defined as in Fig. 5.

[20] Fig. 7 PRIOR ART. Safe-SeqS with exogenous UIDs. DNA (sheared or unsheared) is amplified with a set of gene-specific primers. One of the primers has a random DNA sequence (e.g., a set of 14 Ns) that forms the unique identifier (UID) (variously colored bars), located 5′ to its gene-specific sequence, and both have sequences that permit universal amplification in the next step (yellow and orange bars). Two UID assignment cycles produce two fragments – each with a different UID – from each double-stranded template molecule, as shown. Subsequent PCR with universal primers, which also contain "grafting" sequences (black and red bars), produces UID families that are directly sequenced. Supermutants are defined as in the legend to Fig. 5.

[21] Fig. 8 PRIOR ART. Single-base substitutions identified by conventional (A) and Safe-SeqS (B) analysis. The exogenous UID strategy depicted in Fig. 7 was used to produce PCR fragments from the CTNNB1 gene of three normal, unrelated individuals. Mutation numbers represent one of 87 possible single-base substitutions (3 possible substitutions/base × 29 bases analyzed). These fragments were sequenced on an Illumina GA IIx instrument and analyzed in the conventional manner (A) or with Safe-SeqS (B). Safe-SeqS results are displayed on the same scale as conventional analysis for direct comparison; the inset is a magnified view. Note that most of the variants identified by conventional analysis are likely to represent sequencing errors, as indicated by their high frequency relative to Safe-SeqS and their consistency among unrelated samples.

DETAILED DESCRIPTION OF THE INVENTION

[25] The inventors have developed a test for detecting different cancers using samples that are already routinely collected for diagnosing uterine cancer and HPV (human papilloma virus) infection. Using a panel of genes, a high level of detection of both endometrial and ovarian cancers was achieved.

[26] Certain genes have been identified as mutated in a high proportion of endometrial and ovarian cancers. These include *CTNNB1, EGFR, PI3KCA, PTEN, TP53, BRAF, KRAS, AKT1, NRAS, PPP2R1A, APC, FBXW7, ARID1A, CDKN2A, MLL2, RFF43,* and *FGFR2.* The test can be performed on at least 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, or 17 of these genes. In addition, other genes can be added or substituted into the panel to achieve a higher rate of detection.

[27] Testing for a mutation may be done by analysis of nucleic acids, such as DNA or mRNA or cDNA. The nucleic acid analytes are isolated from cells or cell fragments found in the

liquid PAP smear sample. Suitable tests may include any hybridization or sequencing based assay. Analysis may also be performed on protein encoded by the genes in the panel. Any suitable test may be used including but not limited to mass spectrometry. Other suitable assays may include immunological assays, such as, immunoblotting, immunocytochemistry, immunoprecipitation, enzyme-linked immunosorbent assay (ELISA), radioimmunoassay (RIA), immunoradiometric assays (IRMA) and immunoenzymatic assays (IEMA), including sandwich assays using monoclonal or polyclonal antibodies.

[28] Genetic changes which can be detected are typically mutations such as deletions, insertions, duplications, substitutions (missense or nonsense mutations), rearrangements, etc. Such mutations can be detected *inter alia* by comparing to a wild type in another (non-tumor) tissue or fluid of an individual or by comparing to reference sequences, for example in databases. Mutations that are found in all tissues of an individual are germline mutations, whereas those that occur only in a single tissue are somatic mutations. Epigenetic changes can also be detected. These may be loss or gain of methylation at specific locations in specific genes, as well as histone modifications, including acetylation, ubiquitylation, phosphorylation and sumoylation.

[29] Tests may be done in a multiplex format, in which a single reaction pot is used to detect multiple analytes. Examples of such tests include amplifications using multiple primer sets, amplifications using universal primers, array based hybridization or amplification, emulsion based amplification.

[30] While probes and primers may be designed to interrogate particular mutations or particular portions of a gene, mRNA, or cDNA, these may not be separate entities. For example, probes and primers may be linked together to form a concatamer, or they may be linked to one or more solid supports, such as a bead or an array.

[31] Kits for use in the disclosed methods may include a carrier for the various components. The carrier can be a container or support, in the form of, *e.g.*, bag, box, tube, rack, and is optionally compartmentalized. The kit also includes various components useful in detecting mutations, using the above-discussed detection techniques. For example, the detection kit may include one or more oligonucleotides useful as primers for amplifying all or a portion of the target nucleic acids. The detection kit may also include one or more

oligonucleotide probes for hybridization to the target nucleic acids. Optionally the oligonucleotides are affixed to a solid support, *e.g.,* incorporated in a microarray included in the kit or supplied separately.

[32] Solid supports may contain one single primer or probe or antibody for detecting a single gene, protein, mRNA, or portion of a gene. A solid support may contain multiple primers, probes, or antibodies. They may be provided as a group which will interrogate mutations at least 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 15, 16, or 17 of the genes of the desired panel. The panel may be selected from or comprise *CTNNB1, EGFR, PI3KCA, PTEN, TP53, BRAF, KRAS, AKT1, NRAS, PPP2R1A, APC, FBXW7, ARID1A, CDKN2A, MLL2, RFF43,* and *FGFR2.*

[33] Primer pairs may be used to synthesize amplicons of various sizes. Amplicons may be for example from 50, 60, 75, 100, 125, 150, 200, 140, 180 bp in length. Amplicons may run up to 200, 250, 300, 400, 500, 750, 1000 bp in length, as examples. The size of the amplicon may be limited by the size and/or quality of the template retrieved from the liquid PAP smear. Probes and primers for use in the invention may contain a wild-type sequence or may contain a sequence of a particular mutant.

[34] In one embodiment, the test can be performed using samples that are collected over time. The test results can be compared for quantitative or qualitative changes. Such analysis can be used after a potentially curative therapy, such as surgery.

[35] Georgios Papanicolaou published his seminal work, entitled "Diagnosis of Uterine Cancer by the Vaginal Smear," in 1943 (*31*). At that time, he suggested that endocervical sampling could, *in theory,* be used to detect not only cervical cancers but also other cancers arising in the female reproductive tract, including endometrial carcinomas. The research reported here moves us much closer to that goal. In honor of Papanicolaou's pioneering contribution to the field of early cancer detection, we have named the approach described herein as the "PapGene" test.

[36] One of the most important developments over the last several years is the recognition that all human cancers are the result of mutations in a limited set of genes and an even more limited set of pathways through which these genes act (*32*). The whole-exome sequencing data we present, combined with previous genome-wide studies, provide a

striking example of the common genetic features of cancer (Fig. 6, Table 2). Through the analysis of particular regions of only 12 genes (Fig. 11, table S5), we could detect at least one driver mutation in the vast majority of nine different gynecologic cancers (Fig. 5, Table 1). Though several of these 12 genes were tumor suppressors, and therefore difficult to therapeutically target, knowledge of their mutational patterns provides unprecedented opportunities for cancer diagnostics.

[37] The most important finding in this paper is that significant amounts of cells or cell fragments from endometrial and ovarian cancers are present in the cervix and can be detected through molecular genetic approaches. Detection of malignant cells from endometrial and ovarian carcinomas in cervical cytology specimens is relatively uncommon. Microscopic examination cannot always distinguish them from one another, from cervical carcinomas, or from more benign conditions. Our study showed that 100% of endometrial cancers (n=24), even those of low grade, and 41% of ovarian cancers (n=22), shed cells into the cervix that could be detected from specimens collected as part of routine Pap smears. This finding, in conjunction with technical advances allowing the reliable detection of mutations present in only a very small fraction of DNA templates, provided the foundation for the PapGene test.

[38] This study demonstrates the value of sensitive endocervical DNA testing but there are many issues that need to be addressed before optimal clinical use is achieved. The test, even in its current format, appears to be promising for screening endometrial cancer, as the data in Fig. 3 show that even the lowest stage endometrial cancers could be detected through the analysis of DNA in Pap smear fluid through Safe-SeqS. However, only 41% of ovarian cancers could be detected in Pap smears even when the mutations in their tumors were known. In eight of the nine Pap smears from ovarian cancer patients that contained detectable mutations, the mutant allele fractions were >0.1% and therefore within the range currently detectable by PapGene testing (Fig. 9, table S3). Further improvements in the technology could increase the technical sensitivity of the PapGene test and allow it to detect more ovarian cancers. Other strategies to increase this sensitivity involve physical maneuvers, such as massaging the adnexal region during the pelvic examination or by performing the PapGene test at specified times during the menstrual cycle. Development of an improved method of collection may also be able to improve sensitivity. The current liquid specimen is designed for the detection of cervical cancer and as such utilizes a brush that collects cells from the ectocervix and only

minimally penetrates the endocervical canal. A small cannula that can be introduced into the endometrial cavity similar to the pipelle endometrial biopsy instrument could theoretically obtain a more enriched sample of cells coming from the endometrium, fallopian tube and ovary (*33*).

[39] The high sensitivity and the quantitative nature of the PapGene test also opens the possibility of utilizing it to monitor the response to hormonal agents (*e.g.,* progestins) used to treat young women with low risk endometrial cancers. Some of these women choose to preserve fertility, undergoing medical therapy rather than hysterectomy (*34*). The detection of pre-symptomatic ovarian cancers, even if advanced, could also be of benefit. Although not entirely analogous, it has been demonstrated that one of the most important prognostic indicators for ovarian cancer is the amount of residual disease after surgical debulking. Initially, debulking was considered optimal if the residual tumor was less than 2 cm. Subsequently, the threshold was reduced to 1 cm and surgeons now attempt to remove any visible tumor. With each improvement in surgical debulking, survival has lengthened (*35*). A small volume of tumor is likely to be more sensitive to cytotoxic chemotherapy than the large, bulky disease typical of symptomatic high-grade serous carcinoma.

[40] An essential aspect of the screening approach described here is that it should be relatively inexpensive and easily incorporated into the pelvic examination. Evaluation of HPV DNA is already part of routine Pap smear testing because HPV analysis increases the test's sensitivity (*36, 37*). The DNA purification component of the PapGene test is identical to that used for HPV, so this component is clearly feasible. The preparation of DNA, multiplex amplification, and the retail cost of the sequencing component of the PapGene test can also be performed at a cost comparable to a routine HPV test in the U.S. today. Note that the increased sensitivity provided by the Safe-SeqS component of the PapGene test (see Example 6) can be implemented on any massively parallel sequencing instrument, not just those manufactured by Illumina. With the reduction in the cost of massively parallel sequencing expected in the future, PapGene testing should become even less expensive.

[41] There are millions of Pap smear tests performed annually in the U.S.. Could PapGene testing be performed on such a large number of specimens? We believe so, because the entire DNA purification and amplification process can be automated, just as it is for HPV

testing. Though it may now seem unrealistic to have millions of these sophisticated sequence-based tests performed every year, it would undoubtedly have seemed unrealistic to have widespread, conventional Pap smear testing performed when Papanicolaou published his original paper (*31*). Even today, when many cervical cytology specimens are screened using automated technologies, a significant percentage require evaluation by a skilled cyotpathologist. In contrast, the analysis of PapGene testing is done completely *in silico* and the read-out of the test is objective and quantitative.

[42] In sum, PapGene testing has the capacity to increase the utility of conventional cytology screening through the unambiguous detection of endometrial and ovarian carcinomas. In addition to the analysis of much larger numbers of patients with and without various types of endometrial, ovarian, and fallopian tube cancers, the next step in this line of research is to include genes altered in cervical cancer as well as HPV amplicons in the multiplexed Safe-SeqS assay (Fig. 11, table S5). These additions will provide information that could be valuable for the management of patients with the early stages of cervical neoplasia, as HPV positivity alone is not specific for the detection of cervical cancer and its precursor lesions, particularly in young, sexually active women who frequently harbor HPV infections in the absence of neoplasia.

[43] The above disclosure generally describes the present invention.

A more complete understanding can be obtained by reference to the following specific examples which are provided herein for purposes of illustration only, and are not intended to limit the scope of the invention.

EXAMPLE 1

[44] We reasoned that more sophisticated molecular methods might be able to detect the presence of cancer cells in endocervical specimens at higher sensitivities and specificities than possible with conventional methods. In particular, we hypothesized that somatic mutations characteristic of endometrial and ovarian cancers would be found in the DNA purified from routine liquid-based Pap smears (henceforth denoted as "Pap smears"; Fig. 1). Unlike cytologically abnormal cells, such oncogenic DNA mutations are specific, clonal markers of neoplasia that should be absent in non-neoplastic cells. However, we did not know if such DNA would indeed be present in endocervical specimens, and we

did not know if they would be present in a sufficient amount to detect them. The experiments described here were carried out to test our hypothesis.

[45] There were four components to this study: I. Determination of the somatic mutations typically present in endometrial and ovarian cancers; II. Identification of at least one mutation in the tumors of 46 patients with these cancers; III. Determination of whether the mutations identified in these tumors could also be detected in Pap smears from the same patients; and IV. Development of a technology that could directly assess cells from Pap smears for mutations commonly found in endometrial or ovarian cancers.

EXAMPLE 2

Prevalence of somatically mutated genes in endometrial and ovarian cancers.

[46] There are five major histopathologic subtypes of ovarian cancers. The most prevalent subtype is high grade serous (60% of total), followed by endometrioid (15%), clear cell (10%), and low-grade serous carcinoma (8%) (Table 1). Genome-wide studies have identified the most commonly mutated genes among the most prevalent ovarian cancer subtypes (Table 2) (*23-25*).

[47] Such comprehensive studies have not yet been reported for the endometrioid and mucinous subtypes, collectively representing ~20% of ovarian cancer cases (Table 1). However, commonly mutated genes in the endometrioid and mucinous subtypes have been reported (*26*). In aggregate, the most commonly mutated gene in epithelial ovarian cancers was *TP53*, which was mutated in 69% of these cancers (Table 2). Other highly mutated genes included *ARID1A*, *BRAF*, *CTNNB1*, *KRAS*, *PIK3CA*, and *PPP2R1A* (Table 2).

[48] Among endometrial cancers, the endometrioid subtype is by far the most common, representing 85% of the total (Table 1). Because cancers of this subtype are so frequent and have not been analyzed at a genome-wide level, we evaluated them through whole-exome sequencing. The DNA purified from 22 sporadic endometrioid carcinomas, as well as from matched non-neoplastic tissues, was used to generate 44 libraries suitable for massively parallel sequencing. The clinical aspects of the patients and histopathologic features of the tumors are listed in table S1. Though the examination of 22 cancers cannot provide a comprehensive genome landscape of a tumor type, it is

adequate for diagnostic purposes - as these only require the identification of the most frequently mutated genes.

[49] Among the 44 libraries, the average coverage of each base in the targeted region was 149.1 with 88.4% of targeted bases represented by at least ten reads. Using stringent criteria for the identification of somatic mutations (as described in Materials and Methods), the sequencing data clearly demarcated the tumors into two groups: ten cancers (termed the N Group, for *n*on-highly mutated) harbored <100 somatic mutations per tumor (median 32, range 7 to 50), while 12 cancers (termed the H Group, for *h*ighly mutated) harbored >100 somatic mutations per tumor (median 674, range 164 to 4,629) (Fig. 7, table S1).

[50] The high number of mutations in the Group H tumors was consistent with a deficiency in DNA repair. Eight of the 12 Group H tumors had microsatellite instability (MSI-H, table S1), supporting this conjecture. Moreover, six of the Group H tumors contained somatic mutations in the mismatch repair genes *MSH2* or *MSH6*, while none of the Group N cancers contained mutations in mismatch repair genes. Mismatch repair deficiency is known to be common among endometrial cancers and these tumors occur in 19-71% of women with inherited mutations of mismatch repair genes (i.e., patients with the Hereditary Nonpolyposis Colorectal Cancer) (*27*).

[51] 12,795 somatic mutations were identified in the 22 cancers. The most commonly mutated genes included the PIK3 pathway genes *PTEN* and *PIK3CA (28),* the APC pathway genes *APC* and *CTNNB1*, the fibroblast growth factor receptor *FGFR2*, the adapter protein *FBXW7*, and the chromatin-modifying genes *ARID1A* and *MLL2* (Table 2). Genes in these pathways were mutated in both Group N and H tumors. Our results are consistent with prior studies of endometrioid endometrial cancer that had evaluated small numbers of genes, though mutations in *FBXW7*, *MLL2* and *APC* had not been appreciated to occur as frequently as we found them. It was also interesting that few*TP53* mutations (5%) were found in these endometrial cancers (Table 2), a finding also consistent with prior studies.

[52] Papillary serous carcinomas of the endometrium account for 10-15% of endometrial cancers, and a recent genome-wide sequencing study of this tumor subtype has been published (*29*). The most common mutations in this subtype are listed in Table 2. The

least common subtype of endometrial cancers is clear cell carcinomas, which occur in <5%. Genes reported to be mutated in these cancers were garnered from the literature (Table 2).

EXAMPLE 3

Identification of mutations in tumor tissues

[53] We acquired tumors from 46 cancer patients in whom Pap smears were available. These included 24 patients with endometrial cancers and 22 with ovarian cancers; clinical and histopathologic features are listed in table S3.

[54] Somatic mutations in the 46 tumors were identified through whole-exome sequencing as described above or through targeted sequencing of genes frequently mutated in the most common subtypes of ovarian or endometrial cancer (Table 2). Enrichment for these genes was achieved using a custom solid phase capture assay comprised of oligonucleotides ("capture probes") complementary to a panel of gene regions of interest. For the oncogenes, we only targeted their commonly mutated exons, whereas we targeted the entire coding regions of the tumor suppressor genes.

[55] Illumina DNA sequencing libraries were generated from tumors and their matched non-neoplastic tissues, then captured with the assay described above. Following amplification by PCR, four to eight captured DNA libraries were sequenced per lane on an Illumina GA IIx instrument. In each of the 46 cases, we identified at least one somatic mutation (table S3) that was confirmed by an independent assay, as described below.

EXAMPLE 4

Identification of somatic mutations in Pap smears

[56] In the liquid-based Pap smear technique in routine use today, the clinician inserts a small brush into the endocervical canal during a pelvic exam and rotates the brush so that it dislodges and adheres to loosely attached cells or cell fragments. The brush is then placed in a vial of fixative solution (*e.g.*, ThinPrep). Some of the liquid from the vial is used to prepare a slide for cytological analysis or for purification of HPV DNA. In our study, an aliquot of the DNA purified from the liquid was used to assess for the presence of DNA from the cancers of the 46 patients described above. Preliminary studies showed that the

fixed cells or cell fragments in the liquid, pelleted by centrifugation at 1,000 g for five minutes, contained >95% of the total DNA in the vial. We therefore purified DNA from the cell pellets when the amount of available liquid was greater than 3 mL (as occurs with some liquid-based Pap smear kits) and, for convenience, purified DNA from both the liquid and cells when smaller amounts of liquid were in the kit. In all cases, the purified DNA was of relatively high molecular weight (95% >5 kb). The average amount of DNA recovered from the 46 Pap smears was 49.3 ± 74.4 ng/ml (table S3).

[57] We anticipated that, if present at all, the amount of DNA derived from neoplastic cells in the Pap smear fluid would be relatively small compared to the DNA derived from normal cells brushed from the endocervical canal. This necessitated the use of an analytic technique that could reliably identify a rare population of mutant alleles among a great excess of wild-type alleles. A modification of one of the Safe-SeqS (Safe-Sequencing System) procedures described in (*30*) was designed for this purpose (Fig. 2).

[58] In brief, a limited number of PCR cycles was performed with a set of gene-specific primers. One of the primers contained 14 degenerate N bases (equal probability of being an A, C, G, or T) located 5' to its gene-specific sequence, and both primers contained sequences that permitted universal amplification in the next step. The 14 N's formed unique identifiers (UID) for each original template molecule. Subsequent PCR products generated with universal primers were purified and sequenced on an Illumina MiSeq instrument. If a mutation preexisted in a template molecule, that mutation should be present in every daughter molecule containing that UID, and such mutations are called "supermutants" (*30*). Mutations not occurring in the original templates, such as those occurring during the amplification steps or through errors in base calling, should not give rise to supermutants. The Safe-SeqS approach used here is capable of detecting 1 mutant template among 5,000 to 1,000,000 wild-type templates, depending on the amplicon and the position within the amplicon that is queried (*30*).

[59] We designed Safe-SeqS primers (table S4) to detect at least one mutation from each of the 46 patients described in table S3. In the 24 Pap smears from patients with endometrial cancers, the mutation present in the tumor was identified in every case (100%). The median fraction of mutant alleles was 2.7%, and ranged from 0.01% to 78% (Fig. 3 and table S3). Amplifications of DNA from non-neoplastic tissues were used as negative controls in these experiments to define the detection limits of each queried

mutation. In all cases, the fraction of mutant alleles was significantly different from the background mutation levels determined from the negative controls ($P<0.001$, binomial test). There was no obvious correlation between the fraction of mutant alleles and the histopathologic subtype or the stage of the cancer (Fig. 3 and table S3).

[60] In two endometrial cancer cases, two mutations found in the tumor DNA were evaluated in the Pap smears (table S3). In both cases, the mutations were identified in DNA from the Pap smear (table S3). Moreover, the ratios between the mutant allele fractions of the two mutations in the Pap smears were correlated with those of the corresponding tumor samples. For example, in the Pap smear of case PAP 083 the mutant allele fractions for the *CTNNB1* and *PIK3CA* mutations were 0.143% and 0.064%, respectively - a ratio of 2.2 (=0.14% to 0.064%). In the primary tumor from PAP 083, the corresponding ratio was 2.0 (79.5% to 39.5%).

[61] Similar analysis of Pap smear DNA from ovarian cancer patients revealed detectable mutations in nine of the 22 patients (41%). The fraction of mutant alleles was smaller than in endometrial cancers (median of 0.49%, range 0.021% to 5.9%; see Fig. 3 and table S3). All but one of the cases with detectable mutations were epithelial tumors; the exception was a dysgerminoma, a malignant germ cell tumor of the ovary (table S3). As with endometrial cancers, there was no statistically significant correlation between the fraction of mutant alleles and histopathologic criteria. However, most ovarian cancers are detected only at an advanced stage, and this was reflected in the patients available in our cohort.

EXAMPLE 5

A genetic test for screening purposes

[62] The results described above document that mutant DNA molecules from most endometrial cancers and some ovarian cancers can be found in routinely collected Pap smears. However, in all 46 cases depicted in Fig. 3, a specific mutation was known to occur in the tumor, and an assay was subsequently designed to determine whether that mutation was also present in the corresponding Pap smears. In a screening setting, there obviously would be no known tumor prior to the test. We therefore designed a prototype test based on Safe-SeqS that could be used in a screening setting (Fig. 2).

[63] This multiplexed approach included 50 primer pairs that amplified segments of 241 to 296 bp containing frequently mutated regions of DNA. The regions to be amplified were chosen from the results described in Section I and included exons from *APC, AKT1, BRAF, CTNNB1, EGFR, FBXW7, KRAS, PIK3CA, PPP2R1A, PTEN, and TP53.* In control experiments, 46 of the 50 amplicons were shown to provide information on a minimum of 2,500 templates; the number of templates sequenced can be determined directly from SafeSeqS-based sequencing (Fig. 2). Given the accuracy of SafeSeqS, this number was adequate to comfortably detect mutations existing in >0.1% of template molecules (*30*). The regions covered by these 46 amplicons (table S5), encompassing 10,257 bp, were predicted to be able to detect at least one mutation in >90% of either endometrial or ovarian cancers.

[64] This test was applied to Pap smears of 14 cases - twelve endometrial and two ovarian - as well as 14 Pap smears collected from normal women. The 14 cancer cases were arbitrarily chosen from those which had mutant allele fractions >0.1% (table S3) and therefore above the detection limit of the multiplexed assay. In all 14 Pap smears from women with cancer, the mutation expected to be present (table S3) was identified (Fig. 4 and table S6). The fraction of mutant alleles in the multiplexed test was similar to that observed in the original analysis of the same samples using only one Safe-SeqS primer pair per amplicon (table S3 and table S6). Importantly, no mutations were detected in the 14 Pap smears from women without cancer (Fig. 4; see Materials and Methods).

EXAMPLE 6

Materials and Methods

Patient Samples

[65] All samples for this study were obtained using 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 was collected for each case. All histopathology was centrally re-reviewed by board-certified pathologists. Staging was based on 2009 FIGO criteria (38).

[66] Fresh-frozen tissue specimens of surgically resected neoplasms of the ovary and endometrium were analyzed by frozen section to assess 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.

[67] Formalin-fixed paraffin embedded (FFPE) tissue samples were assessed by a board-certified pathologist (Propath LLC, Dallas, TX) for tumor cellularity and to demarcate area of high tumor cellularity. Tumor tissue 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.

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

[69] Liquid-based Pap smears were collected using cervical brushes and transport medium from Digene HC2 DNA Collection Device (Qiagen) or ThinPrep 2000 System (Hologic) and stored using the manufacturer's recommendations.

[70] Unless otherwise indicated, all patient-related values are reported as mean ± 1 standard deviation.

DNA Extraction

[71] DNA was purified from tumor and normal tissue as well as liquid-based Pap Smears using 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) following the manufacturer's protocol. DNA was purified from Pap smear liquids by adding five volumes of RLTM buffer when the amount of liquid was less than 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 qPCR, employing the primers and conditions previously described (39).

Microsatellite instability testing

[72] Microsatellite instability was detected using 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. Following 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 of Illumina DNA libraries and capture for exomic sequencing

[73] 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 using 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 prevent them from later self-ligation; a corresponding "T" on the 3' end of adaptor molecules provided the complementary overhang. Following 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.

[74] 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 (SEQ ID NO: 1), where the six base pair "XXXXXX" denotes one of 12 sample-specific indexes).

Targeted gene enrichment

[75] Targeted gene enrichment was performed by modifications of previously described methods (40, 41). In brief, targeted regions of selected oncogenes and tumor suppressor genes were synthesized as oligonucleotide probes by Agilent Technologies. Probes of 36 bases were designed to capture both the plus and the minus strand of the DNA and had a 33-base overlap. The oligonucleotides were cleaved from the chip by incubating with 3

mL of 35% ammonium hydroxide at room temperate 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' (SEQ ID NO: 2) and 5'-GACCGCGACTCCAG*C-3' (SEQ ID NO: 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 with a MinElute Purification Column (Qiagen). The PCR products were ligated to form concatamers as described (40).

[76] The major difference between the protocol described in (40, 41) and the one used in the present study involved the amplification of the ligated PCR products and the solid phase capture method. The modifications were as follows: 50 ng of ligated PCR product was amplified using the REPLI-g Midi Kit (Qiagen) with the addition of 2.5 nmol Biotin-dUTP (Roche) in a 27.5 μL reaction. The reaction was incubated at 30oC for 16 hours, the polymerase was inactivated at 65oC for 3 mins. The amplified probes were purified with QiaQuick PCR Purification Columns (Qiagen). For capture, 4-5 μg of library DNA was incubated with 1 μg of the prepared probes in a hybridization mixture as previously described(40). The biotinylated probes and captured library sequences were subsequently purified using 500 μg Dynabeads® MyOne Streptavidin (Invitrogen). After washing as per the manufacturer's recommendations, the captured sequences were eluted with 0.1 M NaOH and then neutralized with 1M Tris-HCl (pH 7.5). Neutralized DNA was desalted and concentrated using a QIAquick MinElute Column (Qiagen) in 20 μL. The elute 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 10s, 60°C for 30 s, 72°C for 30 s; and 72°C for 5 min. The amplified pool containing enriched target sequences was purified using an Agencourt AMPure XP system (Beckman) and quantified using a 2100 Bioanalyzer (Agilent).

Next-generation sequencing and somatic mutation identification

[77] After capture of targeted sequences, paired-end sequencing using 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 using the ELANDv2 algorithm in the CASAVA 1.6 software (Illumina). Known polymorphisms recorded in dbSNP were removed from the analysis. Identification of high confidence mutations was performed as described previously (24).

Assessment of low-frequency mutations

[78] Primer Design. We attempted to design primer pairs to detect mutations in the 46 cancers described in the text. Primers were designed as described (30), using Primer3.(42) Sixty percent of the primers amplified the expected fragments; in the other 40%, a second or third set of primers had to be designed to reduce primer dimers or non-specific amplification.

[79] Sequencing Library Preparation. Templates were amplified as described previously (30), with modifications that will be described in full elsewhere. 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) - using two to four cycles of amplicon-specific PCR (UID assignment PCR cycles, see Fig. 2). While 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 - only the forward primer contained the UID, positioned between the 5' universal tag and the 3' gene-specific sequences (four N's were included in the reverse primer to facilitate sequencing done on paired-end libraries) (table S4). The UID assignment PCR cycles included Phusion Hot Start II (Thermo Scientific) in a 50 μL reaction containing 1X Phusion HF buffer, 0.25 mM dNTPs, 0.5 μM each of forward (containing 14 N's) and reverse primers, and 2 U of polymerase. Carryover of residual UID-containing primers to the second-round amplification, which can complicate template quantification (30), was minimized through exonuclease digestion at 370C to degrade unincorporated primers and subsequent purification with AMPure XP beads (Beckman) and elution in 10 μL TE (10 mM Tris-HCl, 1 mM EDTA, pH 8.0).

[80] The eluted templates were amplified in a second-round PCR using primers containing the grafting sequences necessary for hybridization to the Illumina GA IIx flow cell at their 5' ends (Fig. 2) and two terminal 3' phosphorothioates to protect them from residual exonuclease activity(30). 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(30). 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 polymerase in a total of 50 μL. After an initial heat activation step at 980 C for 2 minutes, twenty-three cycles of PCR were performed using 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 per sample with the primers described in table S5. The PCR products were purified using AMPure XP beads and used directly for sequencing on either the Illumina MiSeq or GA IIx instruments, with equivalent results.

[81] Data Analysis. High quality sequence reads were analyzed as previously described.(30) Briefly, 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 UID. 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 using a custom script (available from the authors upon request). Artifactual mutations – introduced during the sample preparation and/or sequencing steps – were eliminated by requiring that >50% of reads sharing the same UID contained the identical mutation (a "supermutant;" see 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 equaling the mean frequency of all supermutants plus six standard deviations of the mean. Only supermutants exceeding this threshold were designated as mutations and reported in Fig. 4 and table S6.

References

1. M. Arbyn, A. Anttila, J. Jordan, G. Ronco, U. Schenck, N. Segnan, H. Wiener, A. Herbert, L. von Karsa, European Guidelines for Quality Assurance in Cervical Cancer Screening. Second edition--summary document. *Ann Oncol* 21, 448-458 (2010).

2. R. M. DeMay, *Practical principles of cytopathology*. (ASCP Press, Chicago, 2010), pp. xi, 402 p.

3. F. Bray, J. S. Ren, E. Masuyer, J. Ferlay, Global estimates of cancer prevalence for 27 sites in the adult population in 2008. *Int J Cancer*, (2012).

4. J. Ferlay, H.R. Shin, F. Bray, D. Forman, C. Mathers, D.M. Parkin, *GLOBOCAN 2008 v2.0, Cancer Incidence and Mortality Worldwide* (IARC CancerBase No. 10, Lyon, France, 2010).

5. S. B. Sams, H. S. Currens, S. S. Raab, Liquid-based Papanicolaou tests in endometrial carcinoma diagnosis. Performance, error root cause analysis, and quality improvement. *Am J Clin Pathol* 137, 248-254 (2012).

6. P. Smith, O. Bakos, G. Heimer, U. Ulmsten, Transvaginal ultrasound for identifying endometrial abnormality. *Acta Obstet Gynecol Scand* 70, 591-594 (1991).

7. H. Mitchell, G. Giles, G. Medley, Accuracy and survival benefit of cytological prediction of endometrial carcinoma on routine cervical smears. *Int J Gynecol Pathol* 12, 34-40 (1993).

8. K. J. Carlson, S. J. Skates, D. E. Singer, Screening for ovarian cancer. *Ann Intern Med* 121, 124-132 (1994).

9. H. Meden, A. Fattahi-Meibodi, CA 125 in benign gynecological conditions. *Int J Biol Markers* 13, 231-237 (1998).

10. S. S. Buys, E. Partridge, A. Black, C. C. Johnson, L. Lamerato, C. Isaacs, D. J. Reding, R. T. Greenlee, L. A. Yokochi, B. Kessel, E. D. Crawford, T. R. Church, G. L. Andriole, J. L. Weissfeld, M. N. Fouad, D. Chia, B. O'Brien, L. R. Ragard, J. D. Clapp, J. M. Rathmell, T. L. Riley, P. Hartge, P. F. Pinsky, C. S. Zhu, G. Izmirlian, B. S. Kramer, A. B. Miller, J. L. Xu, P. C. Prorok, J. K. Gohagan, C. D. Berg, Effect of screening on ovarian cancer mortality: the Prostate, Lung, Colorectal and Ovarian (PLCO) Cancer Screening Randomized Controlled Trial. *JAMA* 305, 2295-2303 (2011).

11. ACOG Practice Bulletin. Clinical Management Guidelines for Obstetrician-Gynecologists. Number 60, March 2005. Pregestational diabetes mellitus. *Obstet Gynecol* 105, 675-685 (2005).

12. E. Partridge, A. R. Kreimer, R. T. Greenlee, C. Williams, J. L. Xu, T. R. Church, B. Kessel, C. C. Johnson, J. L. Weissfeld, C. Isaacs, G. L. Andriole, S. Ogden, L. R. Ragard, S. S. Buys, Results from four rounds of ovarian cancer screening in a randomized trial. *Obstet Gynecol* 113, 775-782 (2009).

13. American Cancer Society. Detailed guide: ovarian cancer — can ovarian cancer be found early? (Available at http://www.cancer.org/Cancer/OvarianCancer/DetailedGuide/ovarian-cancer-detection).

14. Screening for ovarian cancer: recommendation statement. U.S. Preventive Services Task Force. *Am Fam Physician* 71, 759-762 (2005).

15. ACOG Committee Opinion: number 280, December 2002. The role of the generalist obstetrician-gynecologist in the early detection of ovarian cancer. *Obstet Gynecol* 100, 1413-1416 (2002).

16. National Comprehensive Cancer Network Practice Guidelines in Oncology: ovarian cancer and genetic screening. (Available at http://www.nccn.org/professionals/physician_gls/PDF/genetics_screening.pdf).

17. N. M. Lindor, G. M. Petersen, D. W. Hadley, A. Y. Kinney, S. Miesfeldt, K. H. Lu, P. Lynch, W. Burke, N. Press, Recommendations for the care of individuals with an inherited predisposition to Lynch syndrome: a systematic review. *JAMA* 296, 1507-1517 (2006).

18. J. P. Marques, L. B. Costa, A. P. Pinto, A. F. Lima, M. E. Duarte, A. P. Barbosa, P. L. Medeiros, Atypical glandular cells and cervical cancer: systematic review. *Rev Assoc Med Bras* 57, 234-238 (2011).

19. R. P. Insinga, A. G. Glass, B. B. Rush, Diagnoses and outcomes in cervical cancer screening: a population-based study. *Am J Obstet Gynecol* 191, 105-113 (2004).

20. K. E. Sharpless, P. F. Schnatz, S. Mandavilli, J. F. Greene, J. I. Sorosky, Dysplasia associated with atypical glandular cells on cervical cytology. *Obstet Gynecol* 105, 494-500 (2005).

21. C. P. DeSimone, M. E. Day, M. M. Tovar, C. S. Dietrich, 3rd, M. L. Eastham, S. C. Modesitt, Rate of pathology from atypical glandular cell Pap tests classified by the Bethesda 2001 nomenclature. *Obstet Gynecol* 107, 1285-1291 (2006).

22. C. S. Geier, M. Wilson, W. Creasman, Clinical evaluation of atypical glandular cells of undetermined significance. *Am J Obstet Gynecol* 184, 64-69 (2001).

23. Integrated genomic analyses of ovarian carcinoma. *Nature* 474, 609-615 (2011).

24. S. Jones, T. L. Wang, M. Shih Ie, T. L. Mao, K. Nakayama, R. Roden, R. Glas, D. Slamon, L. A. Diaz, Jr., B. Vogelstein, K. W. Kinzler, V. E. Velculescu, N. Papadopoulos, Frequent mutations of chromatin remodeling gene ARID1A in ovarian clear cell carcinoma. *Science* 330, 228-231 (2010).

25. S. Jones, T. L. Wang, R. J. Kurman, K. Nakayama, V. E. Velculescu, B. Vogelstein, K. W. Kinzler, N. Papadopoulos, M. Shih Ie, Low-grade serous carcinomas of the ovary contain very few point mutations. *J Pathol* 226, 413-420 (2012).

26. S. A. Forbes, N. Bindal, S. Bamford, C. Cole, C. Y. Kok, D. Beare, M. Jia, R. Shepherd, K. Leung, A. Menzies, J. W. Teague, P. J. Campbell, M. R. Stratton, P. A. Futreal, COSMIC:

mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res* 39, D945-950 (2011).

27. E. Barrow, L. Robinson, W. Alduaij, A. Shenton, T. Clancy, F. Lalloo, J. Hill, D. G. Evans, Cumulative lifetime incidence of extracolonic cancers in Lynch syndrome: a report of 121 families with proven mutations. *Clin Genet* 75, 141-149 (2009).

28. K. Oda, D. Stokoe, Y. Taketani, F. McCormick, High frequency of coexistent mutations of PIK3CA and PTEN genes in endometrial carcinoma. *Cancer Res* 65, 10669-10673 (2005).

29. E. Kuhn, R. C. Wu, B. Guan, G. Wu, J. Zhang, Y. Wang, L. Song, X. Yuan, L. Wei, R. B. Roden, K. T. Kuo, K. Nakayama, B. Clarke, P. Shaw, N. Olvera, R. J. Kurman, D. A. Levine, T. L. Wang, I. M. Shih, Identification of Molecular Pathway Aberrations in Uterine Serous Carcinoma by Genome-wide Analyses. *J Natl Cancer Inst*, (2012).

30. I. Kinde, J. Wu, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci U S A* 108, 9530-9535 (2011). See Appendix.

31. H. F. Traut, G. N. Papanicolaou, Cancer of the Uterus: The Vaginal Smear in Its Diagnosis. *Cal West Med* 59, 121-122 (1943).

32. B. Vogelstein, K. W. Kinzler, Cancer genes and the pathways they control. *Nat Med* 10, 789-799 (2004).

33. J. M. Cooper, M. L. Erickson, Endometrial sampling techniques in the diagnosis of abnormal uterine bleeding. *Obstet Gynecol Clin North Am* 27, 235-244 (2000).

34. C. C. Gunderson, A. N. Fader, K. A. Carson, R. E. Bristow, Oncologic and reproductive outcomes with progestin therapy in women with endometrial hyperplasia and grade 1 adenocarcinoma: a systematic review. *Gynecol Oncol* 125, 477-482 (2012).

35. R. E. Bristow, R. S. Tomacruz, D. K. Armstrong, E. L. Trimble, F. J. Montz, Survival effect of maximal cytoreductive surgery for advanced ovarian carcinoma during the platinum era: a meta-analysis. *J Clin Oncol* 20, 1248-1259 (2002).

36. M. H. Mayrand, E. Duarte-Franco, I. Rodrigues, S. D. Walter, J. Hanley, A. Ferenczy, S. Ratnam, F. Coutlee, E. L. Franco, Human papillomavirus DNA versus Papanicolaou screening tests for cervical cancer. *N Engl J Med* 357, 1579-1588 (2007).

37. P. Naucler, W. Ryd, S. Tornberg, A. Strand, G. Wadell, K. Elfgren, T. Radberg, B. Strander, B. Johansson, O. Forslund, B. G. Hansson, E. Rylander, J. Dillner, Human papillomavirus and Papanicolaou tests to screen for cervical cancer. *N Engl J Med* 357, 1589-1597 (2007).

38. S. Pecorelli, Revised FIGO staging for carcinoma of the vulva, cervix, and endometrium. *Int J Gynaecol* Obstet 105, 103-104 (2009).

39. C. Rago, D. L. Huso, F. Diehl, B. Karim, G. Liu, N. Papadopoulos, Y. Samuels, V. E. Velculescu, B. Vogelstein, K. W. Kinzler, L. A. Diaz, Jr., Serial assessment of human tumor burdens in mice by the analysis of circulating DNA. *Cancer Res* 67, 9364-9370 (2007).

40. J. Wu, H. Matthaei, A. Maitra, M. Dal Molin, L. D. Wood, J. R. Eshleman, M. Goggins, M. I. Canto, R. D. Schulick, B. H. Edil, C. L. Wolfgang, A. P. Klein, L. A. Diaz, Jr., P. J. Allen, C. M. Schmidt, K. W. Kinzler, N. Papadopoulos, R. H. Hruban, B. Vogelstein, Recurrent GNAS mutations define an unexpected pathway for pancreatic cyst development. *Sci Transl Med* 3, 92ra66 (2011).

41. J. He, J. Wu, Y. Jiao, N. Wagner-Johnston, R. F. Ambinder, L. A. Diaz, Jr., K. W. Kinzler, B. Vogelstein, N. Papadopoulos, IgH gene rearrangements as plasma biomarkers in Non- Hodgkin's lymphoma patients. *Oncotarget* 2, 178-185 (2011).

42. S. Rozen, H. Skaletsky, Primer3 on the WWW for general users and for biologist programmers. Methods *Mol Biol* 132, 365-386 (2000).

43. N. Howlader, A.M. Noone, M. Krapcho, N. Neyman, R. Aminou, S. F. Altekruse, C.L. Kosary, J. Ruhl, Z. Tatalovich, H. Cho, A. Mariotto, M. P. Eisner, D. R. Lewis, H. S. Chen, E. J. Feuer, K. A. Cronin, *SEER Cancer Statistics Review, 1975-2009* (National Cancer Institute. Bethesda, MD, 2012).

44. A. Malpica, M. T. Deavers, K. Lu, D. C. Bodurka, E. N. Atkinson, D. M. Gershenson, E. G. Silva, Grading ovarian serous carcinoma using a two-tier system. *Am J Surg Pathol* **28**, 496-504 (2004).

45, L.A.G. Ries, J.L. Young, G.E. Keel, M.P. Eisner, Y.D. Lin, M-J. Horner, S*EER Survival Monograph:* Cancer *Survival Among Adults: US SEER Program, 1988-2001, Patient and Tumor Characteristics* (NIH Pub. No. 07-6215. National Cancer Institute, Bethesda, MD, 2007).

46. C. A. Hamilton, M. K. Cheung, K. Osann, L. Chen, N. N. Teng, T. A. Longacre, M. A. Powell, M. R. Hendrickson, D. S. Kapp, J. K. Chan, Uterine papillary serous and clear cell carcinomas predict for poorer survival compared to grade 3 endometrioid corpus cancers. *Br J Cancer* **94**, 642-646 (2006).

DESCRIPTION OF TABLES

Table 1. Epidemiology of Ovarian and Endometrial Tumors. The estimated numbers of new cases and deaths in the U.S. from the major subtypes of ovarian and endometrial cancers are listed.

Table 2. Genetic Characteristics of Ovarian and Endometrial Cancers. The frequencies of the commonly mutated genes in ovarian and endometrial cancers are listed.

Table Sl. Endometrial Cancers (Endometrioid Subtype) Studied by Whole-exome Sequencing. The summary characteristics of the 22 cancers used for exome sequencing are listed.

Table S3. Mutations Assessed in Pap Smears. Clinical characteristics of the 46 tumor samples are listed, along with the mutation identified in each case and the fraction of mutant alleles identified in the Pap smears.

Table S4. Primers Used to Assess Individual Mutations in Pap Smears. The sequences of the forward and reverse primers used to test each mutation via Safe-SeqS are listed in pairs (SEQ ID NO: 4-99, respectively).

Table S5. Primers Used to Simultaneously Assess 12 Genes in Pap Smears. The sequences of the forward and reverse primers for each tested region are listed in pairs (SEQ ID NO: 100-191, respectively).

Table S6. Mutations Identified in Pap Smears through Simultaneous Assessment of 12 Genes. The fraction of mutant alleles identified in the Pap smears using this approach is listed, along with the precise mutations identified.

Table 1 - Epidemiology of Ovarian and Endometrial Tumors

Tissue	Type	Subtype	Fraction of total	Estimated New Cases in U.S., 2012	Estimated Deaths in U.S., 2012	5-year Survival	Reference No.
Ovarian	Epithelial	High-grade Serous	60%	13,368	9,224	9%	43, 44
		Endometrioid	15%	3,342	2,373	71%	43, 45
		Clear Cell	10%	2,228	1,381	62%	43, 45
		Low-grade Serous	8%	1,782	1,105	40%	43, 44
		Mucinous	2%	446	348	65%	43, 45
		Other	5%	1,114	722	N/A	43, 44, 45
Endometrial	Type I: Endometrioid	Endometrioid	85%	40,060	5,180	91%	43, 45
	Type II: Non-Endometrioid	Papillary Serous	10%	4,713	2,194	45%	43, 45
		Clear Cell	5%	2,357	650	68%	43, 46

Table 2 - Genetic Characteristics of Ovarian and Endometrial Cancers

Tissue	Type	Subtype	Common Mutations (Frequency)	Reference No.
Ovarian	Epithelial	High-grade Serous	TP53 (96%)	23
		Endometrioid	TP53 (68%)	26
			ARID1A (30%)	26
			CTNNB1 (26%)	26
			PTEN (17%)	26
			PIK3CA (15%)	26
			KRAS (10%)	26
			PPP2R1A (11%)	26
			CDKN2A (12%)	26
			BRAF (8%)	26
		Clear Cell	ARID1A (57%)	24
			PIK3CA (40%)	24
			PPP2R1A (7.1%)	24
			KRAS (4.7%)	24
		Low-grade Serous	BRAF (38%)	25
			KRAS (19%)	25
		Mucinous	TP53 (56%)	26
			KRAS (40%)	26
			PPP2R1A (33%)	26
			CDKN2A (16%)	26
			PTEN (11%)	26

Tissue	Type	Subtype	Common Mutations (Frequency)	Reference No.
Endometrial	Type I: Endometriod	Endometrioid	PTEN (64%)	Current study
			PIK3CA (59%)	Current study
			ARID1A (55%)	Current study
			CTNNB1 (32%)	Current study
			MLL2 (32%)	Current study
			FBXW7 (27%)	Current study
			RNF43 (27%)	Current study
			APC (23%)	Current study
			FGFR2 (18%)	Current study
			KRAS (9%)	Current study
			PIK3R1 (9%)	Current study
			EGFR (14%)	Current study
			AKT1 (5%)	Current study
			NRAS (5%)	Current study
			TP53 (5%)	Current study
	Type II: Non-Endometrioid	Papillary serous	TP53 (81.6%)	29
			PIK3CA (24%)	29
			FBXW7 (19.7%)	29
			PPP2R1A (18.4%)	29
		Clear Cell	TP53 (45%)	26
			PPP2R1A (33%)	26
			PIK3CA (29%)	26
			PTEN (13%)	26
			PIK3R1 (9%)	26
			KRAS (5%)	26

Table S1: Endometrial Cancers (Endometrioid Subtype) Studied by Whole-exome Sequencing

Tumor ID	Age	Stage (FIGO)	Pathologic Stage (TNM class)	Number of Mutations	Microsatellite Stability Status*
PAP 003	53	IB	T1bN0M0	847	MSS
PAP 010	73	IB	T1bN0M1	29	MSS
PAP 011	58	IB	T1bN0M2	579	MSI-H
PAP 024	56	IA	T1aN0	7	MSS
PAP 026	86	IA	T1a	769	MSI-H
PAP 030	73	IA	T1aNx	49	MSS
PAP 031	61	IA	T1aNx	41	MSS
PAP 032	82	IA	T1aNx	9	MSS
PAP 033	68	IA	T1aNX	34	MSS
PAP 034	55	IA	T1aN0	454	MSI-H
PAP 043	55	IB	T1BN0MX	26	MSS
PAP 045	57	IB	T1BN0MX	4629	MSS
PAP 046	44	IIA	T2ANXMX	40	MSS
PAP 047	53	IA	T1AN0MX	1767	MSS
PAP 048	62	IIIC	T2AN1MX	394	MSI-H
PAP 049	45	IIB	T2BN0MX	20	MSS
PAP 050	39	IB	T1BN0MX	50	MSS
PAP 052	70	IVB	T1AN1M1	164	MSI-H
PAP 053	66	IB	T1BN0MX	1102	MSI-H
PAP 054	73	IA	T1ANXMX	413	MSI-H
PAP 055	61	IA	T1AN0MX	1195	MSS
PAP 057	59	IIB	T2BN0MX	176	MSI-H

* MSI-H: micosatellite unstable; MSS: microsatellite stable. See Materials and Methods

Table S3. Mutations Assessed in Pap Smears

Case #	Age	Tissue	Subtype	Clinical Stage (FIGO)	Race/ Ethnicity	DNA recovered from Pap smear fluid (ug)	Mutated Gene Name	Mutated Gene ID	Nucleotide (genomic)*	Transcript	Nucleotide (transcript)	Amino Acid (protein)	Mutation Type	Fraction of mutant alleles in Pap smear fluid
PAP 001	**45**	**Endometrial**	**Endometrioid**	**IB**	**White**	**30.8**	**PIK3CA**	**ENSG00000121879**	**g.chr3:180434779A>G**	**CCDS43171.1**	**c.3140A>G**	**p.H1047R**	**Missense**	**10.00%**
PAP 002	39	Ovarian	Papillary serous	IIIC	White	3.65	TP53	ENSG00000141510	g.chr17:7519215A>T	CCDS11118.1	c.440T>A	p.V147D	Missense	0.40%
PAP 003	**53**	**Endometrial**	**Endometrioid**	**IB**	**White**	**5.2**	**APC**	**ENSG00000134982**	**g.chr5:112203538C>T**	**CCDS4107.1**	**c.4348C>T**	**p.R1450X**	**Missense**	**3.20%**
PAP 004	38	Ovarian	Papillary serous	IB	White	0.36	TP53	ENSG00000141510	g.chr17:7519154delC	CCDS11118.1	c.501delG	Frameshift	Indel	<0.005%
PAP 005	**59**	**Ovarian**	**Borderline papillary serous cystadenocarcinoma**	**IIIC**	**White**	**7.6**	**NF1**	**ENSG00000196712**	**g.chr17:26552610insC**	**CCDS42292.1**	**c.1241insC**	**Frameshift**	**Indel**	**<0.005%**
PAP 006	57	Ovarian	Papillary serous	IIIC	White	8.2	TP53	ENSG00000141510	g.chr17:7517824C>T	CCDS11118.1	c.839G>A	p.R280K	Missense	<0.005%
PAP 007	**63**	**Ovarian**	**Papillary serous**	**IIIC**	**White**	**6.8**	**TP53**	**ENSG00000141510**	**g.chr17:7517783C>A**	**CCDS11118.1**	**c.880G>T**	**p.E294X**	**Nonsense**	**<0.005%**
PAP 010	73	Endometrial	Endometrioid	IB	White	8.6	FBXW7	ENSG00000109670	g.chr4:153466817G>A	CCDS3777.1	c.1435C>T	p.R479X	Nonsense	15.00%
PAP 011	**58**	**Endometrial**	**Endometrioid**	**IB**	**White**	**11.2**	**KRAS**	**ENSG00000133703**	**g.chr12:25289551C>G**	**CCDS8703.1**	**c.35G>C**	**p.G12A**	**Missense**	**3.50%**
PAP 024	56	Endometrial	Endometrioid with squamous differentiation	IA	White	5.6	CTNNB1	ENSG00000168036	g.chr3:41241108G>T	CCDS2694.1	c.101G>T	p.G34V	Missense	0.22%
PAP 025	**75**	**Endometrial**	**Serous carcinoma**	**IA**	**Black**	**0.74**	**TP53**	**ENSG00000141510**	**g.chr17:7520119_7520120delGG**	**CCDS11118.1**	**c.292_293delCC**	**Frameshift**	**Indel**	**8.70%**
PAP 026	86	Endometrial	Endometrioid carcinoma	IA	White	17.64	PTEN	ENSG00000171862	g.chr10:89710752G>A	CCDS31238.1	c.923G>A	p.R308H	Missense	0.01%
PAP 027	**65**	**Ovarian**	**Papillary serous**	**IIIC**	**White**	**18.6**	**SETD2**	**ENSG00000181555**	**g.chr3:47138125G>A**	**NM_014159**	**c.805C>T**	**p.Q269X**	**Nonsense**	**<0.005%**
PAP 030	73	Endometrial	Endometrioid	IA	White	17.58	MSH6	ENSG00000116062	g.chr2:47880779G>A	CCDS1836.1	c.2153G>A	p.S718N	Missense	0.24%
PAP 031	**61**	**Endometrial**	**Endometrioid**	**IA**	**Asian**	**101.4**	**CTNNB1**	**ENSG00000168036**	**g.chr3:41241117C>A**	**CCDS2694.1**	**c.110C>A**	**p.S37Y**	**Missense**	**1.30%**
PAP 032	82	Endometrial	Endometrioid	IA	White	2.218	PTEN	ENSG00000171862	g.chr10:89682884C>G	CCDS31238.1	c.388C>G	p.R130G	Missense	0.02%
PAP 033	**68**	**Endometrial**	**Endometrioid with squamous differentiation**	**IA**	**White**	**11.3**	**KRAS**	**ENSG00000133703**	**g.chr12:25289551C>A**	**CCDS8703.1**	**c.35G>T**	**p.G12V**	**Missense**	**1.53%**
PAP 34	55	Endometrial	Endometrioid	IA	White	11.24	CTNNB1	ENSG00000168036	g.chr3:41241107G>A	CCDS2694.1	c.100G>A	p. G34R	Missense	20.12%
PAP 35	**50**	**Endometrial**	**Endometrioid**	**IVB**	**White**	**0.48**	**KRAS**	**ENSG00000133703**	**g.chr12:25289551C>T**	**CCDS8703.1**	**c.35G>A**	**p.G12D**	**Missense**	**6.14%**
PAP 36	57	Ovarian	Endometrioid with squamous differentiation	IA	White	2.47	PIK3CA	ENSG00000121879	g.chr3:180418789A>T	CCDS43171.1	c.1637A>T	p.Q546L	Missense	<0.005%
PAP 37	**64**	**Endometrial**	**Endometrioid**	**IA**	**Asian**	**10.3**	**PIK3CA**	**ENSG00000121879**	**g.chr3:180434712A>G**	**CCDS43171.1**	**c.3073A>G**	**p.T1025A**	**Missense**	**0.17%**
PAP 38	47	Ovarian	Papillary serous	IA	Asian	5.2	TP53	ENSG00000141510	g.chr17:7518915T>C	CCDS11118.1	c.659A>G	p.Y220C	Missense	0.48%
PAP 39	**73**	**Ovarian**	**Papillary serous**	**IA**	**Asian**	**9.0**	**TP53**	**ENSG00000141510**	**g.chr17:7518244delG**	**CCDS11118.1**	**c.871delC**	**Frameshift**	**Indel**	**5.90%**
PAP 40	57	Endometrial	Endometrioid	IB	Asian	7.7	CTNNB1	ENSG00000168036	g.chr3:41241141C>A	CCDS2694.1	c.134C>A	p.S45Y	Missense	0.08%
PAP 41	**66**	**Endometrial**	**Papillary serous**	**IA**	**Asian**	**5.7**	**FGFR2**	**ENSG00000066468**	**g.chr10:123269667G>C**	**CCDS7620.2**	**c.755C>G**	**p.S252W**	**Missense**	**0.13%**
PAP 41	**66**	**Endometrial**	**Papillary serous**	**IA**	**Asian**	**5.7**	**PTEN**	**ENSG00000171862**	**g.chr10:89682902T>C**	**CCDS31238.1**	**c.406T>C**	**p.C136R**	**Missense**	**0.09%**
PAP 42	45	Ovarian	Adenosquamous Carcinoma	IIIC	White	14.3	TP53	ENSG00000141510	g.chr17:7520074insA	CCDS11118.1	c.339insT	Frameshift	Indel	1.47%
PAP 58	**62**	**Ovarian**	**Papillary serous**	**IIIC**	**White**	**14.468**	**PIK3CA**	**ENSG00000121879**	**g.chr3:180399630G>A**	**CCDS43171.1**	**c.323G>A**	**p.R108H**	**Missense**	**0.49%**
PAP 60	69	Ovarian	Papillary serous	IIIC	White	9.312	TP53	ENSG00000141510	g.chr17:7519131C>T	CCDS11118.1	c.524G>A	p.R175H	Missense	<0.05%
PAP 61	**70**	**Ovarian**	**Papillary serous**	**IIIC**	**White**	**5.872**	**TP53**	**ENSG00000141510**	**g.chr17:7517846G>A**	**CCDS11118.1**	**c.817C>T**	**p.R273C**	**Missense**	**0.02%**
PAP 62	55	Ovarian	Papillary serous	IIIC	White	1.444	TP53	ENSG00000141510	g.chr17:7520213_7520216delGCAT	CCDS11118.1	c.196_199delATGC	Frameshift	Indel	<0.005%
PAP 63	**66**	**Ovarian**	**Papillary serous**	**IIIB**	**White**	**7.376**	**TP53**	**ENSG00000141510**	**g.chr17:7519131C>T**	**CCDS11118.1**	**c.524G>A**	**p.R175H**	**Missense**	**<0.005%**
PAP 64	50	Ovarian	Dysgerminoma	IIIA	Asian	5.6	TP53	ENSG00000141510	g.chr17:7518233T>G	CCDS11118.1	c.773A>C	p.E258A	Missense	4.20%
PAP 66	**70**	**Endometrial**	**Endometrioid Adenocarcinoma, Villoglandular Type**	**IIIC**	**Asian**	**7.2**	**TP53**	**ENSG00000141510**	**g.chr17:7518960T>C**	**CCDS11118.1**	**c.614A>G**	**p.Y205C**	**Missense**	**5.48%**
PAP 67	61	Endometrial	Endometrioid	IB	Asian	9.3	TP53	ENSG00000141510	g.chr17:7518937G>A	CCDS11118.1	c.637C>T	p.R213X	Nonsense	2.70%
PAP 68	**63**	**Endometrial**	**Endometrioid Adenocarcinoma with Mucinous Differentiation**	**IB**	**Asian**	**8.6**	**PIK3CA**	**ENSG00000121879**	**g.chr3:180410674T>C**	**CCDS43171.1**	**c.1258T>C**	**p.C420R**	**Missense**	**0.03%**
PAP 69	51	Endometrial	Endometrioid	IB	Asian	7.4	TP53	ENSG00000141510	g.chr17:7520084G>A	CCDS11118.1	c.328C>T	p.R110C	Missense	14.70%
PAP 70	**49**	**Endometrial**	**Endometrioid**	**IIA**	**Asian**	**1.6**	**PIK3CA**	**ENSG00000121879**	**g.chr3:180399570G>A**	**CCDS43171.1**	**c.263G>A**	**p.R88Q**	**Missense**	**29.60%**
PAP 71	57	Endometrial	Endometrioid	IA	White	4.42	PIK3CA	ENSG00000121879	g.chr3:180434779A>G	CCDS43171.1	C.3140A>G	p.H1047R	Missense	0.37%
PAP 72	**64**	**Ovarian**	**Papillary serous**	**IV**	**White**	**4.24**	**TP53**	**ENSG00000141510**	**g.chr17:7518272C>T**	**CCDS11118.1**	**c.734G>A**	**p.G245D**	**Missense**	**0.81%**
PAP 73	64	Ovarian	Papillary serous	IIIC	White	4.62	TP53	ENSG00000141510	g.chr17:7519167T>C	CCDS11118.1	c.488A>G	p.Y163C	Missense	<0.005%
PAP 74	**73**	**Ovarian**	**Papillary serous**	**IV**	**White**	**8.28**	**TP53**	**ENSG00000141510**	**g.chr17:7519131C>T**	**CCDS11118.1**	**c.524G>A**	**p.R175H**	**Missense**	**<0.05%**
PAP 75	54	Ovarian	Papillary serous	IIIC	White	7.42	PIK3CA	ENSG00000121879	g.chr3:180434768G>A	CCDS43171.1	c.3129G>A	p.M1043I		0.13%
PAP 76	**53**	**Ovarian**	**Papillary serous**	**IIIC**	**White**	**9.38**	**TP53**	**ENSG00000141510**	**g.chr17:7519275G>A**	**CCDS11118.1**	**c.380C>T**	**p.S127F**	**Missense**	**<0.005%**
PAP 78	53	Ovarian	Papillary serous	IIIC	White	1.36	TP53	ENSG00000141510	g.chr17:7519131C>T	CCDS11118.1	c.524G>A	p.R175H	Missense	<0.005%
PAP 80	**94**	**Endometrial**	**Papillary serous**	**IIIA**	**Black**	**7.78**	**TP53**	**ENSG00000141510**	**g.chr17:7519008delG**	**CCDS11118.1**	**c.566delC**	**Frameshift**	**Indel**	**79.80%**
PAP 83	51	Endometrial	Endometrioid	IA	Black	4.5	PIK3CA	ENSG00000121879	g.chr3:180434779A>G	CCDS43171.1	C.3140A>G	p.H1047R	Missense	0.14%
PAP 83	51	Endometrial	Endometrioid	IA	Black	4.5	CTNNB1	ENSG00000168036	g.chr3:41241105C>A	CCDS2694.1	c.98C>A	p.S33Y	Missense	0.07%

*Coordinates refer to the human reference genome hg18 release (NCBI 36.1, March 2006).

Table S4. Primers Used to Assess Individual Mutations in Pap Smears

Sample#	Gene	Mutation	Amplified Genomic Fragment Size (bp)	Forward Primer Sequence	Reverse Primer Sequence
PAP 001	PIK3CA	H1047R	82	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNGATCCAATCCATTTTTGTTGTCCAG	cacacaggaaacagctatgaccatgTGAGCAAGAGGCTTTGGAGT
PAP 002	TP53	V147D	93	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNGGCCAAGACCTGCCCTG	cacacaggaaacagctatgaccatgTGCTGTGACTGCTTGTAGATGG
PAP 003	APC	R1450X	93	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNCCACCTCCTCAAACAGCTCAA	cacacaggaaacagctatgaccatgTGCAGCTTGCTTAGGTCCACT
PAP 004	TP53	Q167fs*(1 base deletion of C)	85	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNTGGCCATCTACAAGCAGTCA	cacacaggaaacagctatgaccatgNNNNTCACCATCGCTATCTGAGCA
PAP 005	NF1	L414fs*(1 base insertion of C)	74	cgacgtaaaacgacggccagtNNNNNNNNNNNNNCATTGGTGATGATTCGATGG	cacacaggaaacagctatgaccatgCTGCCTGGCTCAGAATTCAC
PAP 006	TP53	R280K	90	cgacgtaaaacgacggccagtNNNNNNNNNNNNNCCCTTTCTTGCGGAGATTC	cacacaggaaacagctatgaccatgCTACTGGGACGGAACAGCTT
PAP 007	TP53	E294X	87	cgacgtaaaacgacggccagtNNNNNNNNNNNNNGGAAGAGAATCTCCGCAAGA	cacacaggaaacagctatgaccatgGCTTCTTGTCCTGCTTGCTT
PAP 010	FBXW7	R479X	82	cgacgtaaaacgacggccagtNNNNNNNNNNNNNGGCCTGTCTCAATATCCCAAA	cacacaggaaacagctatgaccatgTTGTTTTTCTGTTTCTCCCTCTG
PAP 011	KRAS	G12A	92	cgacgtaaaacgacggccagtNNNNNNNNNNNNNCAAGGCACTCTTGCCTACG	cacacaggaaacagctatgaccatgCATTTTCATTATTTTTATTATAAGGCCTG
PAP 024	CTNNB1	G34V	72	cgacgtaaaacgacggccagtNNNNNNNNNNNNNCTGTGGTAGTGGCACCAGAA	cacacaggaaacagctatgaccatgNNNNAAGCGGCTGTTAGTCACTGG
PAP 025	TP53	P98fs*(2 base deletion of CC)	91	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNGGCCCCTGTCATCTTCTGT	cacacaggaaacagctatgaccatgGACTTGGCTGTCCCAGAATG
PAP 026	PTEN	R308H	99	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNCAAGAAATCGATAGCATTTGCAG	cacacaggaaacagctatgaccatgTTTATTTGCTTTGTCAAGATCATTTTT
PAP 027	SETD2	Q269X	85	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNAGGAAATATCTGCTTGCTCATTCAA	cacacaggaaacagctatgaccatgGAAGCAGATACTAAGCAGGACACTATATC
PAP 030	MSH6	S718N	76	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNTTCCCTTGGATTCTGACACA	cacacaggaaacagctatgaccatgAGCACCATTCGTTGATAGGC
PAP 031	CTNNB1	S37Y	85	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNCACTGGCAGCAACAGTCTTACC	cacacaggaaacagctatgaccatgGATTGCCTTTACCACTCAGAGAAG
PAP 032	PTEN	R130G	71	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNttgcagcaattcactgtaaagc	cacacaggaaacagctatgaccatgccgatgtaataaatatgcacatatcattac
PAP 033	KRAS	G12V	92	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNCAAGGCACTCTTGCCTACG	cacacaggaaacagctatgaccatgCATTTTCATTATTTTTATTATAAGGCCTG
PAP 034	CTNNB1	G34R	85	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNCACTGGCAGCAACAGTCTTACC	cacacaggaaacagctatgaccatgGATTGCCTTTACCACTCAGAGAAG
PAP 035	KRAS	G12D	92	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNCAAGGCACTCTTGCCTACG	cacacaggaaacagctatgaccatgCATTTTCATTATTTTTATTATAAGGCCTG
PAP 036	PIK3CA	Q546L	84	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNAGCTCAAAGCAATTTCTACACGA	cacacaggaaacagctatgaccatgNNNNGCACTTACCTGTGACTCCATAGAA
PAP 037	PIK3CA	T1025A	70	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNtcttttgatgacattgcatacattc	cacacaggaaacagctatgaccatgNNNNactccaaagcctcttGCTCA
PAP 038	TP53	Y220C	83	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNcagttgcaaaccagacctca	cacacaggaaacagctatgaccatgtgtggagtatttggatgacagaaa
PAP 039	TP53	I254fs*(1 base deletion of C)	76	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNgtggcaagtggctcctga	cacacaggaaacagctatgaccatgNNNNCATGGGCGGCATGAAC
PAP 040	CTNNB1	S45Y	90	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNactggcagcaacagtcttacct	cacacaggaaacagctatgaccatgNNNNCCTCAGGATTGCCTTTACCa
PAP 041	FGFR2	S252W	71	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNagtccggcttGGAGGATG	cacacaggaaacagctatgaccatgtccccactcctcctttctc
	PTEN	C136R	78	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNNggaaagGGACGAACTGGTG	cacacaggaaacagctatgaccatgNNNNtagggcctctgtgccttta
PAP 042	TP53	F113fs*(1 base insertion of T)	80	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNtcttctgtcccttcccagaa	cacacaggaaacagctatgaccatgNNNNgacttGGCTGTCCCAGAATG
PAP 058	PIK3CA	R108H	82	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNtcatACCAATTTCTCGATTGAGG	cacacaggaaacagctatgaccatgNNNNcggcttttcaacccttttt
PAP 060	TP53	R175H	85	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNTGGCCATCTACAAGCAGTCA	cacacaggaaacagctatgaccatgNNNNTCACCATCGCTATCTGAGCA
PAP 061	TP53	R273C	76	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNggGACGGAACAGCTTTGAG	cacacaggaaacagctatgaccatgNNNNgcggagattctcttcctctgt
PAP 062	TP53	M66fs*(4 base deletion of GCAT)	77	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNCGGTGTAGGAGCTGCTGGT	cacacaggaaacagctatgaccatgNNNNACCCAGGTCCAGATGAAGC
PAP 063	TP53	R175H	85	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNTGGCCATCTACAAGCAGTCA	cacacaggaaacagctatgaccatgNNNNTCACCATCGCTATCTGAGCA
PAP 064	TP53	E258A	76	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNgtggcaagtggctcctga	cacacaggaaacagctatgaccatgNNNNCATGGGCGGCATGAAC
PAP 066	TP53	Y205C	74	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNCGAAAAGTGTTTCTGTCATCCA	cacacaggaaacagctatgaccatgNNNNGCCCCTCCTCAGCATCTTAT
PAP 067	TP53	R213X	83	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNcagttgcaaaccagacctca	cacacaggaaacagctatgaccatgNNNNtgtggagtatttggatgacagaaa
PAP 068	PIK3CA	C420R	90	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNTTATATTTCCCCATGCCAATG	cacacaggaaacagctatgaccatgNNNNggtgttttgaaatgtgttttataatttaga
PAP 069	TP53	R110C	80	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNtcttctgtcccttcccagaa	cacacaggaaacagctatgaccatgNNNNgacttGGCTGTCCCAGAATG
PAP 070	PIK3CA	R88Q	66	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNGAAAAAGCCGAAGGTCACAA	cacacaggaaacagctatgaccatgNNNNCTCAAGAAGCAGAAAGGGAAGA
PAP 071	PIK3CA	H1047R	82	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNGATCCAATCCATTTTTGTTGTCCAG	cacacaggaaacagctatgaccatgNNNNTGAGCAAGAGGCTTTGGAGT
PAP 072	TP53	G245D	74	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNTGTGATGATGGTGAGGATGg	cacacaggaaacagctatgaccatgNNNNtccacTACAACTACATGTGTAACAGTTC
PAP 073	TP53	Y163C	84	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNCTCCGTCATGTGCTGTGACT	cacacaggaaacagctatgaccatgNNNNcagctgtgggttgattcca
PAP 074	TP53	R175H	85	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNTGGCCATCTACAAGCAGTCA	cacacaggaaacagctatgaccatgNNNNTCACCATCGCTATCTGAGCA
PAP 075	PIK3CA	M1043I	79	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNccaatccatttttgttgtcca	cacacaggaaacagctatgaccatgNNNNTGAGCAAGAGGCTTTGGAGT
PAP 076	TP53	S127F	82	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNctgcacagggcaggtctt	cacacaggaaacagctatgaccatgNNNNctgtctcctcctcttcctaca
PAP 078	TP53	R175H	85	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNTGGCCATCTACAAGCAGTCA	cacacaggaaacagctatgaccatgNNNNTCACCATCGCTATCTGAGCA
PAP 080	PIK3CA	A189fs*(1 base deletion of C)	78	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNcacgcaaATTTCCTTCCACT	cacacaggaaacagctatgaccatgNNNNggcctctgattcctcactgat
PAP 083	PIK3CA	H1047R	82	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNGATCCAATCCATTTTTGTTGTCCAG	cacacaggaaacagctatgaccatgNNNNTGAGCAAGAGGCTTTGGAGT
	CTNNB1	S33Y	90	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNactggcagcaacagtcttacct	cacacaggaaacagctatgaccatgNNNNCCTCAGGATTGCCTTTACCa

Table S5. Primers Used to Simultaneously Assess 12 Genes in Pap Smears

Set	Gene and Codons Covered by Amplicon	Amplicon Length (bp)	Forward Primer Sequence	Reverse Primer sequence
	CTNNB1 005 To 064	251	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNAAAGTAACATTTCCAATCTACTAATGCT	cacacaggaaacagctatgaccatgNNNNtgagaaaatccctgttcccac
	EGFR 697 to 728	264	cgacgtaaaacgacggccagtNNNNNNNNNNNNNNTGGAGCCTCTTACACCCAGT	cacacaggaaacagctatgaccatgNNNNAAAAACACTGGAGTTTCCCAAA
	PIK3CA 0023 to 88	241	cgacgtaaaacgacggccagtNNNNNNNNNNNNNgaaaAAGCCGAAGGTCACAA	cacacaggaaacagctatgaccatgNNNNATGCCCCCAAGAATCCTAGT
#1	PTEN 001 to 026	257	cgacgtaaaacgacggccagtNNNNNNNNNNNNNcatccgtctactcccacgtt	cacacaggaaacagctatgaccatgNNNNatcagctaccgccaagtcc
	PTEN 085 to 123	261	cgacgtaaaacgacggccagtNNNNNNNNNNNNNTTCGTCCCTTTCCAGCTTTA	cacacaggaaacagctatgaccatgNNNNggaatccagtgtttctttttaaatacct
	TP53 126 to 186	259	cgacgtaaaacgacggccagtNNNNNNNNNNNNNaccagccctgtcgtctctc	cacacaggaaacagctatgaccatgNNNNgccctgactttcaactctgtct
	TP53 308 To 331	259	cgacgtaaaacgacggccagtNNNNNNNNNNNNNgcctcagattcacttttatcacct	cacacaggaaacagctatgaccatgNNNNaccaggagccattgtctttg
	BRAF 439 to 475	251	cgacgtaaaacgacggccagtNNNNNNNNNNNNNTCACCACATTACATACTTACCATGC	cacacaggaaacagctatgaccatgNNNNaaggggatctcttcctgtatcc
	EGFR 738 to 761	279	cgacgtaaaacgacggccagtNNNNNNNNNNNNNTCTGGATCCCAGAAGGTGAG	cacacaggaaacagctatgaccatgNNNNGGCCAGTGCTGTCTCTAAGG
	KRAS 001 to 021	281	cgacgtaaaacgacggccagtNNNNNNNNNNNNNGTCCACAAAATGATTCTGAATTAGC	cacacaggaaacagctatgaccatgNNNNACGATACACGTCTGCAGTCAAC
#2	PIK3CA 0062 to 0117	269	cgacgtaaaacgacggccagtNNNNNNNNNNNNNACAGAAATATTTTAGAAAGGGACAACA	cacacaggaaacagctatgaccatgNNNNAGAAAATACCCCCTCCATCAAC
	PIK3CA 0980 to 1049	272	cgacgtaaaacgacggccagtNNNNNNNNNNNNNGATCCAATCCATTTTTGTTGTCCAG	cacacaggaaacagctatgaccatgNNNNTCCAAACTGACCAAACTGTTCTTA
	PTEN 100 to 164	294	cgacgtaaaacgacggccagtNNNNNNNNNNNNNgaaacccaaaatctgttttcca	cacacaggaaacagctatgaccatgNNNNGACCATAACCCACCACAGCTA
	TP53 186 to 224	251	cgacgtaaaacgacggccagtNNNNNNNNNNNNNctcctcccagagaccccagt	cacacaggaaacagctatgaccatgNNNNcatgagcgctgctcagatag
	TP53 332 To 366	263	cgacgtaaaacgacggccagtNNNNNNNNNNNNNcctaggaaggcaggggagt	cacacaggaaacagctatgaccatgNNNNtgcatgttgcttttgtaccg
	AKT1 017 to 058	279	cgacgtaaaacgacggccagtNNNNNNNNNNNNNaccacccgcacgtctgtag	cacacaggaaacagctatgaccatgNNNNagccagtgcttgttgcttg
	EGFR 762 to 823	251	cgacgtaaaacgacggccagtNNNNNNNNNNNNNCACACTGACGTGCCTCTCC	cacacaggaaacagctatgaccatgNNNNttatctccccctccccgtatc
	NRAS 001 to 014	275	cgacgtaaaacgacggccagtNNNNNNNNNNNNNGATTGTCAGTGCGCTTTTCC	cacacaggaaacagctatgaccatgNNNNgctaaggatgggggttgcta
	PIK3CA 0272 to 347	279	cgacgtaaaacgacggccagtNNNNNNNNNNNNNTGACTTTACCTTATCAATGTCTCGAA	cacacaggaaacagctatgaccatgNNNNGCTCGCCCCCTTAATCTCT
#3	PPP2R1A 176 to 217	277	cgacgtaaaacgacggccagtNNNNNNNNNNNNNGGTACTTCCGGAACCTGTGC	cacacaggaaacagctatgaccatgNNNNccgagtcctagggagaggag
	PTEN 056 to 070	250	cgacgtaaaacgacggccagtNNNNNNNNNNNNNttgttaatggtggcttttgttt	cacacaggaaacagctatgaccatgNNNNaaatgatctaacaatgctcttggac
	PTEN 165 to 211	265	cgacgtaaaacgacggccagtNNNNNNNNNNNNNcatggaaggatgagaatttcaag	cacacaggaaacagctatgaccatgNNNNtggctacgacccagttacca
	TP53 033 to 102	256	cgacgtaaaacgacggccagtNNNNNNNNNNNNNAAACCGTAGCTGCCCTGGTA	cacacaggaaacagctatgaccatgNNNNtgactgctcttttcacccatc
	TP53 225 to 260	257	cgacgtaaaacgacggccagtNNNNNNNNNNNNNtcatcttgggcctgtgttatc	cacacaggaaacagctatgaccatgNNNNgatgagaggtggatgggtagtagt
	EGFR 830 to 875	258	cgacgtaaaacgacggccagtNNNNNNNNNNNNNTTCAGGGCATGAACTACTTGG	cacacaggaaacagctatgaccatgNNNNatcctcccctgcatgtgtta
	KRAS 038 to 061	260	cgacgtaaaacgacggccagtNNNNNNNNNNNNNTCCCTCATTGCACTGTACTCC	cacacaggaaacagctatgaccatgNNNNggtgcttagtggccatttgt
	NRAS 038 to 062	250	cgacgtaaaacgacggccagtNNNNNNNNNNNNNTGGTCTCTCATGGCACTGTACT	cacacaggaaacagctatgaccatgNNNNattagcaatttgagggacaaacc
#4	PTEN 071 to 084	250	cgacgtaaaacgacggccagtNNNNNNNNNNNNNtgtatctcactcgataatctggatgac	cacacaggaaacagctatgaccatgNNNNtgtcacattataaagattcaggcaat
	PTEN 213 to 267	271	cgacgtaaaacgacggccagtNNNNNNNNNNNNNagtttgacagttaaaggcatttcc	cacacaggaaacagctatgaccatgNNNNtgtcttattttggatatttctccc
	TP53 063 to 125	276	cgacgtaaaacgacggccagtNNNNNNNNNNNNNGAAGACCCAGGTCCAGATGA	cacacaggaaacagctatgaccatgNNNNagaaatgcaggggatacg
	TP53 262 To 306	280	cgacgtaaaacgacggccagtNNNNNNNNNNNNNgaggcataactgcaccctg	cacacaggaaacagctatgaccatgNNNNgggagtagatggagcctggt
	APC 1254 to 1328	269	cgacgtaaaacgacggccagtNNNNNNNNNNNNNtgctggatttggttctaggg	cacacaggaaacagctatgaccatgNNNNtgccacttgcaaagtttcttc
	APC 1403 to 1476	264	cgacgtaaaacgacggccagtNNNNNNNNNNNNNggaagaacctggaccctctg	cacacaggaaacagctatgaccatgNNNNttttgagagtcgttcgattgc
#5	APC 1537 to 1617	289	cgacgtaaaacgacggccagtNNNNNNNNNNNNNtgcaacctgttttgtgatgg	cacacaggaaacagctatgaccatgNNNNaggaaaatgacaatgggaatga
	FBXW7 474 to 548	296	cgacgtaaaacgacggccagtNNNNNNNNNNNNNtgattcatcaggagagcatttaag	cacacaggaaacagctatgaccatgNNNNttgtttttctgtttctccctctg
	FBXW7 558 to 618	289	cgacgtaaaacgacggccagtNNNNNNNNNNNNNGATGGTATCCATGTGGTGAGTG	cacacaggaaacagctatgaccatgNNNNtttgtgatgctaaggctccat
	TP53 001 to 24	254	cgacgtaaaacgacggccagtNNNNNNNNNNNNNactgccttccgggtcact	cacacaggaaacagctatgaccatgNNNNagcccaaccctgtccttac
	TP53 368 To 393	265	cgacgtaaaacgacggccagtNNNNNNNNNNNNNgaggctgtcagtggggaac	cacacaggaaacagctatgaccatgNNNNaacatatttgcatggggtgtg
	APC 1326 to 1405	283	cgacgtaaaacgacggccagtNNNNNNNNNNNNNccactgcatggttcactctg	cacacaggaaacagctatgaccatgNNNNatcctgtgagcgaagttcca
	APC 1475 to 1545	264	cgacgtaaaacgacggccagtNNNNNNNNNNNNNtgcctctttctctggttttca	cacacaggaaacagctatgaccatgNNNNggacctaagcaagctgcagtaa
	FBXW7 243 to 287	257	cgacgtaaaacgacggccagtNNNNNNNNNNNNNacacccaatgaagaatgtaattgat	cacacaggaaacagctatgaccatgNNNNggttgtgtgtagatgtgagttttcc
#6	FBXW7 329 to 374	276	cgacgtaaaacgacggccagtNNNNNNNNNNNNNttctgttacattgtgcagagttca	cacacaggaaacagctatgaccatgNNNNtgggttttgagcagagagatgg
	FBXW7 413 to 472	266	cgacgtaaaacgacggccagtNNNNNNNNNNNNNggaagaagtcccaaccatga	cacacaggaaacagctatgaccatgNNNNtcactttttcctttctacccaaaag
	FBXW7 549 to 593	259	cgacgtaaaacgacggccagtNNNNNNNNNNNNNGAATCTGCATTCCCAGAGACA	cacacaggaaacagctatgaccatgNNNNcctgtttcccatcctcttcc
	KRAS 134 to 150	267	cgacgtaaaacgacggccagtNNNNNNNNNNNNNGACACAAAACAGGCTCAGGAC	cacacaggaaacagctatgaccatgNNNNagaaaccaaagccaaaagca
	TP53 026 to 32	253	cgacgtaaaacgacggccagtNNNNNNNNNNNNNccatgggactgactttctgc	cacacaggaaacagctatgaccatgNNNNtcatctggacctgggtcttc

CA 02889937 2017-02-23

Table S6. Mutations Identified in Pap Smears through Simultaneous Assessment of 12 Genes

Case #	Tumor Type	Mutated Gene Name	Mutated Gene ID	Nucleotide (genomic)*	Transcript	Nucleotide (transcript)*	Amino Acid (protein)	Mutation Type	Fraction of mutant alleles in Pap smear fluid
PAP 001	Endometrial	KRAS	ENSG00000133703	g.chr12:25289551C>G	CCDS8703.1	c.35G>C	p.G12A	Missense	12.51%
		PIK3CA	ENSG00000121879	g.chr3:180434779A>G	CCDS43171.1	c.3140A>G	p.H1047R	Missense	5.74%
PAP 003	Endometrial	PIK3CA	ENSG00000121879	g.chr3:180399570G>A	CCDS43171.1	c.263G>A	p.R88Q	Missense	11.60%
		APC	ENSG00000134982	g.chr5:112203538C>T	CCDS4107.1	c.4348C>T	p.R1450X	Missense	12.50%
		PTEN	ENSG00000171862	g.chr10:89614243A>C	CCDS31238.1	c.38A>C	p.K13T	Missense	12.38%
PAP 010	Endometrial	FBXW7	ENSG00000109670	g.chr4:153466817G>A	CCDS3777.1	c.1435C>T	p.R479X	Nonsense	20.00%
PAP 011	Endometrial	KRAS	ENSG00000133703	g.chr12:25289551C>G	CCDS8703.1	c.35G>C	p.G12A	Missense	3.23%
PAP 025	Endometrial	TP53	ENSG00000141510	g.chr17:7520119_7520120delGG	CCDS11118.1	c.292_293delCC	Frameshift	Indel	10.42%
PAP 033	Endometrial	PIK3CA	ENSG00000121879	g.chr3:180404243T>G	CCDS43171.1	c.1031T>G	p.V344G	Missense	1.22%
		KRAS	ENSG00000133703	g.chr12:25289551C>A	CCDS8703.1	c.35G>T	p.G12V	Missense	1.13%
		PTEN	ENSG00000171862	g.chr10:89707637insC	CCDS31238.1	c.682insC	Frameshift	Indel	0.87%
PAP 34	Endometrial	PIK3CA	ENSG00000121879	g.chr3:180399630G>A	CCDS43171.1	c.323G>A	p.R108H	Missense	22.78%
		CTNNB1	ENSG00000168036	g.chr3:41241107G>A	CCDS2694.1	c.100G>A	p. G34R	Missense	18.41%
		PTEN	ENSG00000171862	g.chr10:89707750delA	CCDS31238.1	c.795delA	Frameshift	Indel	13.28%
		PIK3CA	ENSG00000121879	g.chr3:180399554T>G	CCDS43171.1	c.247T>G	p.F83V	Missense	4.49%
		KRAS	ENSG00000133703	g.chr12:25289551C>T	CCDS8703.1	c.35G>A	p.G12D	Missense	0.92%
PAP 35	Endometrial	KRAS	ENSG00000133703	g.chr12:25289551C>T	CCDS8703.1	c.35G>A	p.G12D	Missense	5.83%
		PIK3CA	ENSG00000121879	g.chr3:180399548G>A	CCDS43171.1	c.241G>A	p.E81K	Missense	5.32%
		PTEN	ENSG00000171862	g.chr10:89675287T>C	CCDS31238.1	c.202T>C	p.Y68H	Missense	4.73%
PAP 39	Ovarian	TP53	ENSG00000141510	g.chr17:7518244delG	CCDS11118.1	c.871delC	Frameshift	Indel	0.73%
PAP 67	Endometrial	TP53	ENSG00000141510	g.chr17:7518937G>A	CCDS11118.1	c.637C>T	p.R213X	Nonsense	2.31%
PAP 69	Endometrial	TP53	ENSG00000141510	g.chr17:7520084G>A	CCDS11118.1	c.328C>T	p.R110C	Missense	19.23%
		TP53	ENSG00000141510	g.chr17:7517747G>A	CCDS11118.1	c.916C>T	p.R306X	Nonsense	13.60%
PAP 70	Endometrial	PIK3CA	ENSG00000121879	g.chr3:180399570G>A	CCDS43171.1	c.263G>A	p.R88Q	Missense	28.05%
PAP 71	Endometrial	KRAS	ENSG00000133703	g.chr12:25289551C>T	CCDS8703.1	c.35G>A	p.G12D	Missense	0.39%
		PIK3CA	ENSG00000121879	g.chr3:180434779A>G	CCDS43171.1	c.3140A>G	p.H1047R	Missense	0.31%
PAP 72	Ovarian	TP53	ENSG00000141510	g.chr17:7518272C>T	CCDS11118.1	c.734G>A	p.G245D	Missense	0.54%

*Coordinates refer to the human reference genome hg18 release (NCBI 36.1, March 2006).

APPENDIX

DETECTION AND QUANTIFICATION OF RARE MUTATIONS WITH MASSIVELY PARALLEL SEQUENCING (PRIOR ART)

Kinde I, Wu J, Papadopoulos N, Kinzler KW, Vogelstein B (2011) Detection and quantification of rare mutations with massively parallel sequencing. *Proc Natl Acad Sci USA* 108:9530–9535.

[82] The identification of mutations that are present in a small fraction of DNA templates is essential for progress in several areas of biomedical research. Although massively parallel sequencing instruments are in principle well suited to this task, the error rates in such instruments are generally too high to allow confident identification of rare variants. We here describe an approach that can substantially increase the sensitivity of massively parallel sequencing instruments for this purpose. The keys to this approach, called the Safe-Sequencing System ("Safe-SeqS"), are (i) assignment of a unique identifier (UID) to each template molecule, (ii) amplification of each uniquely tagged template molecule to create UID families, and (iii) redundant sequencing of the amplification products. PCR fragments with the same UID are considered mutant ("supermutants") only if ≥95% of them contain the identical mutation. We illustrate the utility of this approach for determining the fidelity of a polymerase, the accuracy of oligonucleotides synthesized in vitro, and the prevalence of mutations in the nuclear and mitochondrial genomes of normal cells.

[83] Genetic mutations underlie many aspects of life and death – through evolution and disease, respectively. Accordingly, their measurement is critical to several fields of research. Luria and Delbrück's classic fluctuation analysis is a prototypic example of the insights into biological processes that can be gained simply by counting the number of mutations in carefully controlled experiments (A1). Counting de novo mutations in humans, not present in their parents, has similarly led to new insights into the rate at which our species can evolve (A2, A3). Similarly, counting genetic or epigenetic changes in tumors can inform fundamental issues in cancer biology (A4). Mutations lie at the core of current problems in managing patients with viral diseases such as AIDS and hepatitis by virtue of the drug resistance they can cause (A5, A6). Detection of such mutations, particularly at a stage before their becoming dominant in the population, will likely be essential to optimize

therapy. Detection of donor DNA in the blood of organ transplant patients is an important indicator of graft rejection and detection of fetal DNA in maternal plasma can be used for prenatal diagnosis in a noninvasive fashion (A7, A8). In neoplastic diseases, which are all driven by somatic mutations, the applications of rare mutant detection are manifold; they can be used to help identify residual disease at surgical margins or in lymph nodes, to follow the course of therapy when assessed in plasma, and to identify patients with early, surgically curable disease when evaluated in stool, sputum, plasma, and other bodily fluids (A9, A10, A11).

[84] These examples highlight the importance of identifying rare mutations for both basic and clinical research. Accordingly, innovative ways to assess them have been devised over the years. The first methods involved biologic assays based on prototrophy, resistance to viral infection or drugs, or biochemical assays (A1, A12, A13, A14, A15, A16, A17, A18). Molecular cloning and sequencing provided a new dimension to the field, as they allowed the type of mutation, rather than simply its presence, to be identified (A19, A20, A21, A22, A23, A24). Some of the most powerful of these newer methods are based on digital PCR, in which individual molecules are assessed one by one (A25). Digital PCR is conceptually identical to the analysis of individual clones of bacteria, cells, or virus, but is performed entirely in vitro with defined, inanimate reagents. Several implementations of digital PCR have been described, including the analysis of molecules arrayed in multiwell plates, in polonies, in microfluidic devices, and in water-in-oil emulsions (A25, A26, A27, A28, A29, A30). In each of these technologies, mutant templates are identified through their binding to oligonucleotides specific for the potentially mutant base.

[85] Massively parallel sequencing represents a particularly powerful form of digital PCR in that hundreds of millions of template molecules can be analyzed one by one. It has the advantage over conventional digital PCR methods in that multiple bases can be queried sequentially and easily in an automated fashion. However, massively parallel sequencing cannot generally be used to detect rare variants because of the high error rate associated with the sequencing process. For example, with the commonly used Illumina sequencing instruments, this error rate varies from ~1% (A31, A32) to ~0.05% (A33, A34), depending on factors such as the read length (A35), use of improved base-calling algorithms (A36,

A37, A38), and the type of variants detected (A39). Some of these errors presumably result from mutations introduced during template preparation, during the preamplification steps required for library preparation, and during further solid-phase amplification on the instrument itself. Other errors are due to base misincorporation during sequencing and basecalling errors. Advances in base calling can enhance confidence (e.g., refs. A36–A39), but instrument-based errors are still limiting, particularly in clinical samples wherein the mutation prevalence can be $\leq$0.01% (A11). In the work described herein, we show how templates can be prepared and the sequencing data obtained from them more reliably interpreted, so that relatively rare mutations can be identified with commercially available instruments.

RESULTS

Overview

[86] Our approach, called the Safe-Sequencing System ("Safe-SeqS"), involves two basic steps (Fig. 5). The first is the assignment of a unique identifier (UID) to each DNA template molecule to be analyzed. The second is the amplification of each uniquely tagged template, so that many daughter molecules with the identical sequence are generated (defined as a UID family). If a mutation preexisted in the template molecule used for amplification, that mutation should be present in every daughter molecule containing that UID (barring any subsequent replication or sequencing errors). A UID family in which at least 95% of family members have the identical mutation is called a "supermutant". Mutations not occurring in the original templates, such as those occurring during the amplification steps or through errors in base calling, should not give rise to supermutants.

Endogenous UIDs

[87] UIDs, sometimes called barcodes or indexes, can be assigned to nucleic acid fragments using a variety of methods. These methods include the introduction of exogenous sequences through PCR (A40, A41) or ligation (A42, A43). Even more simply, randomly sheared genomic DNA inherently contains UIDs consisting of the sequences of the two ends of each sheared fragment (Fig. 6). Paired-end sequencing of these fragments yields UID families that can be analyzed as described above. To use such endogenous UIDs in

APPENDIX

Safe-SeqS, we used two separate approaches: one designed to evaluate many genes simultaneously and the other designed to evaluate a single gene fragment in depth (Fig. 6).

[88] For the evaluation of multiple genes, we ligated standard Illumina sequencing adapters to the ends of sheared DNA fragments to produce a standard sequencing library and then captured genes of interest on a solid phase (A44). In this experiment, a library made from the DNA of ~15,000 normal cells was used, and 2,594 bp from six genes were targeted for capture. After excluding known single-nucleotide polymorphisms, 25,563 apparent mutations, corresponding to 2.4×10^{-4} mutations/bp, were also identified (Table A1). On the basis of previous analyses of mutation rates in human cells, at least 90% of these apparent mutations were likely to represent mutations introduced during template and library preparation or base-calling errors. Note that the error rate determined here (2.4×10^{-4} mutations/bp) is considerably lower than usually reported in experiments using the Illumina instrument because we used very stringent criteria for base calling.

[89] With Safe-SeqS analysis of the same data, we determined that 69,505 original template molecules were assessed in this experiment (i.e., 69,505 UID families, with an average of 40 members per family, were identified) (Table A1). All of the polymorphic variants identified by conventional analysis were also identified by Safe-SeqS. However, only eight supermutants were observed among these families, corresponding to 3.5×10^{-6} mutations/bp. Thus, Safe-SeqS decreased the presumptive sequencing errors by at least 70-fold.

[90] A strategy using endogenous UIDs was also used to reduce false-positive mutations upon deep sequencing of a single region of interest. In this case, a library prepared as described above from ~1,750 normal cells was used as template for inverse PCR using primers complementary to a gene of interest, so the PCR products could be directly used for sequencing. With conventional analysis, an average of 2.3×10^{-4} mutations/bp were observed, similar to that observed in the capture experiment (Table A1). Given that only 1,057 independent molecules from normal cells were assessed in this experiment, as determined through Safe-SeqS analysis, all mutations observed with conventional analysis

likely represented false positives (Table A1). With Safe-SeqS analysis of the same data, no supermutants were identified at any position.

Table A1. Safe-SeqS with endogenous UIDs

	Capture	Inverse PCR
Conventional analysis		
High-quality base pairs	106,958,863	1,041,346,645
Mean high-quality base pairs read depth	38,620x	2,0885,600x
Mutations identified	25,563	234,352
Mutations/bp	2.4*E*-04	2.3*E*-04
Safe-SeqS analysis		
High-quality base pairs	106,958,863	1,041,346,645
Mean high-quality base pairs read depth	38,620x	2,085,600x
UID families	69,505	1,057
Average no. of members/UID family	40	21,688
Median no. of members/UID family	19	4
Supermutants identified	8	0
Supermutants/bp	3.5*E*-06	0.0

Exogenous UIDs

[91] Although the results described above show that Safe-SeqS can increase the reliability of massively parallel sequencing, the number of different molecules that can be examined using endogenous UIDs is limited. For fragments sheared to an average size of 150 bp (range 125-175), 36-base paired-end sequencing can evaluate a maximum of ~7,200 different molecules containing a specific mutation (2 reads × 2 orientations × 36 bases/read × 50-base variation on either end of the fragment). In practice, the actual number of UIDs is smaller because the shearing process is not entirely random.

[92] To make more efficient use of the original templates, we developed a Safe-SeqS strategy that used a minimum number of enzymatic steps. This strategy also permitted the use of degraded or damaged DNA, such as found in clinical specimens or after bisulfite treatment for the examination of cytosine methylation (A45). As depicted in Fig. 7, this strategy employs two sets of PCR primers. The first set is synthesized with standard

phosphoramidite precursors and contained sequences complementary to the gene of interest on the 3′ end and different tails at the 5′ ends of both the forward and reverse primers. The different tails allowed universal amplification in the next step. Finally, there was a stretch of 12-14 random nucleotides between the tail and the sequence-specific nucleotides in the forward primer (A40). The random nucleotides form the UIDs. An equivalent way to assign UIDs to fragments, not used in this study, would employ 10,000 forward primers and 10,000 reverse primers synthesized on a microarray. Each of these 20,000 primers would have gene-specific primers at their 3′ ends and one of 10,000 specific, predetermined, nonoverlapping UID sequences at their 5′ ends, allowing for 10^8 [i.e., $(10^4)^2$] possible UID combinations. In either case, two cycles of PCR are performed with the primers and a high-fidelity polymerase, producing a uniquely tagged, double-stranded DNA fragment from each of the two strands of each original template molecule (Fig. 7). The residual, unused UID assignment primers are removed by digestion with a single strand-specific exonuclease, without further purification, and two new primers are added. The new primers, complementary to the tails introduced in the UID assignment cycles, contain grafting sequences at their 5′ ends, permitting solid-phase amplification on the Illumina instrument, and phosphorothioate residues at their 3′ ends to make them resistant to any remaining exonuclease. Following 25 additional cycles of PCR, the products are loaded on the Illumina instrument. As shown below, this strategy allowed us to evaluate the majority of input fragments and was used for several illustrative experiments.

Analysis of DNA Polymerase Fidelity

[93] Measurement of the error rates of DNA polymerases is essential for their characterization and dictates the situations in which these enzymes can be used. We chose to measure the error rate of Phusion polymerase, as this polymerase has one of the lowest reported error frequencies of any commercially available enzyme and therefore poses a particular challenge for an in vitro-based approach. We first amplified a single human DNA template molecule, comprising a segment of an arbitrarily chosen human gene, through 19 rounds of PCR. The PCR products from these amplifications, in their entirety, were used as templates for Safe-SeqS as described in Fig. 7. In seven independent experiments of this

type, the number of UID families identified by sequencing was 624,678 ± 421,274, which is consistent with an amplification efficiency of 92 ± 9.6% per round of PCR.

[94] The error rate of Phusion polymerase, estimated through cloning of PCR products encoding β-galactosidase in plasmid vectors and transformation into bacteria, is reported by the manufacturer to be 4.4×10^{-7} errors/bp/PCR cycle. Even with very high-stringency base calling, conventional analysis of the Illumina sequencing data revealed an apparent error rate of 9.1×10^{-6} errors/bp/PCR cycle, more than an order of magnitude higher than the reported Phusion polymerase error rate (Table A2, polymerase fidelity). In contrast, Safe-SeqS of the same data revealed an error rate of 4.5×10^{-7} errors/bp/PCR cycle, nearly identical to that measured for Phusion polymerase in biological assays (Table A2, polymerase fidelity). The vast majority (>99%) of these errors were single-base substitutions, consistent with previous data on the mutation spectra created by other prokaryotic DNA polymerases (A15, A46, A47).

[95] Safe-SeqS also allowed a determination of the total number of distinct mutational events and an estimation of PCR cycle in which the mutation occurred. There were 19 cycles of PCR performed in wells containing a single template molecule in these experiments. If a polymerase error occurred in cycle 19, there would be only one supermutant produced (from the strand containing the mutation). If the error occurred in cycle 18, there should be two supermutants (derived from the mutant strands produced in cycle 19), etc. Accordingly, the cycle in which the error occurred is related to the number of supermutants containing that error. The data from seven independent experiments demonstrate a relatively consistent number of observed total polymerase errors ($2.2 \pm 1.1 \times 10^{-6}$ distinct mutations/bp), in reasonable agreement with the number expected from simulations ($1.5 \pm 0.21 \times 10^{-6}$ distinct mutations/bp). The data also show a highly variable timing of occurrence of polymerase errors among experiments, as predicted from classic fluctuation analysis (A1). This kind of information is difficult to derive using conventional analysis of the same next-generation sequencing data, in part because of the prohibitively high apparent mutation rate noted above.

APPENDIX

Table A2. Safe-SeqS with exogenous UIDs

	Mean	SD
Polymerase fidelity		
Conventional analysis of seven replicates		
High-quality base pairs	996,855,791	64,030,757
Total mutations identified	198,638	22,515
Mutations/bp	2.0*E*-04	1.7*E*-05
Calculated Phusion error rate (errors/bp/cycle)	9.1*E*-06	7.7*E*-07
Safe-SeqS analysis of seven replicates		
High-quality base pairs	996,855,791	64,030,757
UID families	624	421,274
Members/UID family	107	122
Total supermutants identified	197	143
Supermutants/bp	9.9*E*-06	2.3*E*-06
Calculated Phusion error rate (errors/bp/cycle)	4.5*E*-07	1.0*E*-07
CTNNB1 mutations in DNA from normal human cells		
Conventional analysis of three individuals		
High-quality base pairs	559,334,774	66,600,749
Total mutations identified	118,488	11,357
Mutations/bp	2.1*E*-04	1.6*E*-05
Safe-SeqS analysis of three individuals		
High-quality base pairs	559,334,774	66,600,749
UID families	374,553	263,105
Members/UID family	68	38
Total supermutants identified	99	78
Supermutants/bp	9.0*E*-06	3.1*E*-06
Mitochondrial mutations in DNA from normal human cells		
Conventional analysis of seven individuals		
High-quality base pairs	147,673,456	54,308,546
Total mutations identified	30,599	12,970
Mutations/bp	2.1*E*-04	9.4*E*-05
Safe-SeqS analysis of seven individuals		
High-quality base pairs	147,673,456	54,308,546
UID families	515,600	89,985
Members/UID family	15	6
Total supermutants identified	135	61
Supermutants/bp	1.4*E*-05	6.8*E*-06

APPENDIX

Analysis of Oligonucleotide Composition

[96] A small number of mistakes during the synthesis of oligonucleotides from phosphoramidite precursors are tolerable for most applications, such as routine PCR or cloning. However, for synthetic biology, wherein many oligonucleotides must be joined together, such mistakes present a major obstacle to success. Clever strategies for making the gene construction process more efficient have been devised (A48, A49), but all such strategies would benefit from more accurate synthesis of the oligonucleotides themselves. Determining the number of errors in synthesized oligonucleotides is difficult because the fraction of oligonucleotides containing errors can be lower than the sensitivity of conventional next-generation sequencing analyses.

[97] To determine whether Safe-SeqS could be used for this determination, we used standard phosphoramidite chemistry to synthesize an oligonucleotide containing 31 bases that were designed to be identical to that analyzed in the polymerase fidelity experiment described above. In the synthetic oligonucleotide, the 31 bases were surrounded by sequences complementary to primers that could be used for the UID assignment steps of Safe-SeqS (Fig. 7). By performing Safe-SeqS on ~300,000 oligonucleotide templates, we found that there were $8.9 \pm 0.28 \times 10^{-4}$ supermutants/bp and that these errors occurred throughout the sequence of the oligonucleotides. The oligonucleotides contained a large number of insertion and deletion errors, representing 8.2 ± 0.63% and 25 ± 1.5% of the total supermutants, respectively. Importantly, both the position and the nature of the errors were highly reproducible among seven independent replicates of this experiment performed on the same batch of oligonucleotides. This nature and distribution of errors had little in common with that of the errors produced by Phusion polymerase, which were distributed in the expected stochastic pattern among replicate experiments. The number of errors in the oligonucleotides synthesized with phosphoramidites was ~60 times higher than that in the equivalent products synthesized by Phusion polymerase. These data, in toto, indicate that the vast majority of errors in the former were generated during their synthesis rather than during the Safe-SeqS procedure.

[98] Does Safe-SeqS preserve the ratio of mutant:normal sequences in the original templates? To address this question, we synthesized two 31-base oligonucleotides of identical

sequence with the exception of nucleotide 15 (50:50 C/G instead of T) and mixed them at nominal mutant/normal fractions of 3.3% and 0.33%. Through Safe-SeqS analysis of the oligonucleotide mixtures, we found that the ratios were 2.8% and 0.27%, respectively. We conclude that the UID assignment and amplification procedures used in Safe-SeqS do not greatly alter the proportion of variant sequences and thereby provide a reliable estimate of that proportion when unknown. This conclusion is also supported by the reproducibility of variant fractions when analyzed in independent Safe-SeqS experiments.

Analysis of DNA Sequences from Normal Human Cells

[99] The exogenous UID strategy (Fig. 7) was then used to determine the prevalence of rare mutations in a small region of the *CTNNB1* gene isolated from ~100,000 normal human cells from three unrelated individuals. Through comparison with the number of UID families obtained in the Safe-SeqS experiments (Table A2, *CTNNB1* mutations in DNA from normal human cells), we calculated that the majority (78 ± 9.8%) of the input fragments were converted into UID families. There was an average of 68 members/UID family, easily fulfilling the required redundancy for Safe-SeqS. Conventional analysis of the Illumina sequencing data revealed an average of 118,488 ± 11,357 mutations among the ~560 Mb of sequence analyzed per sample, corresponding to an apparent mutation prevalence of $2.1 \pm 0.16 \times 10^{-4}$ mutations/bp (Table A2, *CTNNB1* mutations in DNA from normal human cells). Only an average of 99 ± 78 supermutants were observed in the Safe-SeqS analysis. The vast majority (>99%) of supermutants were single-base substitutions and the calculated mutation rate was $9.0 \pm 3.1 \times 10^{-6}$ mutations/bp. Safe-SeqS thereby reduced the apparent frequency of mutations in genomic DNA by at least 24-fold (Fig. 8).

[100] We applied the identical strategy to a short segment of mitochondrial DNA isolated from ~1,000 cells from each of seven unrelated individuals. Conventional analysis of the Illumina sequencing libraries produced with the Safe-SeqS procedure (Fig. 7) revealed an average of 30,599 ± 12,970 mutations among the ~150 Mb of sequence analyzed per sample, corresponding to an apparent mutation prevalence of $2.1 \pm 0.94 \times 10^{-4}$ mutations/bp (Table A2, mitochondrial mutations in DNA from normal human cells). Only 135 ± 61 supermutants were observed in the Safe- SeqS analysis. As with the *CTNNB1* gene, the vast majority of mutations were single-base substitutions, although occasional

single-base deletions were also observed. The calculated mutation rate in the analyzed segment of mtDNA was $1.4 \pm 0.68 \times 10^{-5}$ mutations/bp (Table A2, mitochondrial mutations in DNA from normal human cells). Thus, Safe-SeqS thereby reduced the apparent frequency of mutations in mitochondrial DNA by at least 15-fold.

DISCUSSION

[101] The results described above demonstrate that the Safe-SeqS approach can substantially improve the accuracy of massively parallel sequencing (Tables A1 and A2). It can be implemented through either endogenous or exogenously introduced UIDs and can be applied to virtually any sample preparation workflow or sequencing platform. As demonstrated here, the approach can easily be used to identify rare mutants in a population of DNA templates, to measure polymerase error rates, and to judge the reliability of oligonucleotide syntheses. One of the advantages of the strategy is that it yields the number of templates analyzed as well as the fraction of templates containing variant bases. Previously described in vitro methods for the detection of small numbers of template molecules (e.g., refs. A29 and A50) allow the fraction of mutant templates to be determined but cannot determine the number of mutant and normal templates in the original sample.

[102] It is of interest to compare Safe-SeqS to other approaches for reducing errors in next-generation sequencing. As mentioned in the Introduction, sophisticated algorithms to increase the accuracy of base calling have been developed (e.g., refs. A36, A37, A38, A39). These improved base calling algorithms can certainly reduce false-positive calls, but their effectiveness is still limited by artifactual mutations occurring during the PCR steps required for library preparation as well as by any residual base-calling errors. For example, the algorithm used in the current study used very stringent criteria for base calling and was applied to short read lengths, but was still unable to reduce the error rate to less than an average of 2.0×10^{-4} errors/bp. This error frequency is at least as low as those reported with other algorithms. To improve sensitivity further, these base-calling improvements can be used together with Safe-SeqS. Travers et al. describe another powerful strategy for reducing errors (A51). With this technology, both strands of each template molecule are sequenced redundantly after a number of preparative enzymatic steps. However, this approach can be performed only on a specific instrument. Moreover, for many clinical

applications, there are relatively few template molecules in the initial sample and evaluation of nearly all of them is required to obtain the requisite sensitivity. The approach described here with exogenously introduced UIDs (Fig. 7) fulfills this requirement by coupling the UID assignment step with a subsequent amplification in which few molecules are lost. Our endogenous UID approaches (Fig. 6) and the one described by Travers et al. are not ideally suited for this purpose because of the inevitable losses of template molecules during the ligation and other preparative steps.

[103] How do we know that the mutations identified by conventional analyses in the current study represent artifacts rather than true mutations in the original templates? Strong evidence supporting this is provided by the observation that the mutation prevalence in all but one experiment was similar: 2.0×10^{-4} - 2.4×10^{-4} mutations/bp (Tables A1 and A2). The exception was the experiment with oligonucleotides synthesized from phosphoramidites, in which the error of the synthetic process was apparently higher than the error rate of conventional Illumina analysis when used with stringent base-calling criteria. In contrast, the mutation prevalence of Safe-SeqS varied much more, from 0.0 to 1.4×10^{-5} mutations/bp, depending on the template and experiment. Moreover, the mutation prevalence measured by Safe-SeqS in the most controlled experiment, in which polymerase fidelity was measured (Table A2, polymerase fidelity), was almost identical to that predicted from previous experiments in which polymerase fidelity was measured by biological assays. Our measurements of mutation prevalence in the DNA from normal cells are consistent with some previous experimental data. However, estimates of these prevalences vary widely and may depend on cell type and sequence analyzed. We therefore cannot be certain that the relatively low number of mutations revealed by Safe-SeqS represented errors occurring during the sequencing process rather than true mutations present in the original DNA templates.

[104] Like all techniques, Safe-SeqS has limitations. For example, we have demonstrated that the exogenous UIDs strategy can be used to analyze a single amplicon in depth. This technology may not be applicable to situations wherein multiple amplicons must be analyzed from a sample containing a limited number of templates. Multiplexing in the UID assignment cycles (Fig. 7) may provide a solution to this challenge. A second limitation is

that the efficiency of amplification in the UID assignment cycles is critical for the success of the method. Clinical samples can contain inhibitors that reduce the efficiency of this step. This problem can presumably be overcome by performing more than two cycles in the UID assignment PCR step (Fig. 7), although this would complicate the determination of the number of templates analyzed. The specificity of Safe-SeqS is currently limited by the fidelity of the polymerase used in the UID assignment PCR step, i.e., 8.8×10^{-7} mutations/bp in its current implementation with two cycles. Increasing the number of cycles in the UID assignment PCR step to five would decrease the overall specificity to $\sim 2 \times 10^{-6}$ mutations/bp. However, this specificity can be increased by requiring more than one supermutant for mutation identification – the probability of introducing the same artifactual mutation twice or three times would be exceedingly low [$(2 \times 10^{-6})^2$ or $(2 \times 10^{-6})^3$, respectively]. In sum, there are several simple ways to vary the Safe-SeqS procedure and analysis to realize the needs of specific experiments.

[105] Luria and Delbrück, in their classic paper in 1943, wrote that their "prediction cannot be verified directly, because what we observe, when we count the number of resistant bacteria in a culture, is not the number of mutations which have occurred but the number of resistant bacteria which have arisen by multiplication of those which mutated, the amount of multiplication depending on how far back the mutation occurred" (ref. A1, p. 495). The Safe-SeqS procedure described here can verify such predictions because the number as well as the time of occurrence of each mutation can be estimated from the data, as noted in the experiments on polymerase fidelity. In addition to templates generated by polymerases in vitro, the same approach can be applied to DNA from bacteria, viruses, and mammalian cells. We therefore expect that this strategy will provide definitive answers to a variety of important biomedical questions.

MATERIALS AND METHODS

Endogenous UIDs

[106] To expose endogenous UIDs, DNA was fragmented to an average size of ~200 bp by acoustic shearing (Covaris) and then end-repaired, A-tailed, and ligated to Y-shaped adapters according to standard Illumina protocols. DNA was captured (A44) with a filter

containing 2,594 nt corresponding to six cancer genes. For the inverse PCR experiments, we ligated custom adapters (IDT) instead of standard Y-shaped Illumina adapters to sheared cellular DNA. Inverse PCR was performed using *KRAS* forward and reverse primers that both contained grafting sequences for hybridization to the Illumina GA IIx flow cell.

Exogenous UIDs

[107] Each strand of each template molecule was encoded with a 12- or 14-base UID using two cycles of amplicon-specific PCR, as described in the text and Fig. 7. The amplicon-specific primers both contained universal tag sequences at their 5′ ends for a later amplification step. The UIDs constituted 12 or 14 random nucleotide sequences appended to the 5′ end of the forward amplicon-specific primers. Following two cycles of PCR for UID assignment, the products were digested with a single-strand DNA-specific nuclease. Primers complementary to the introduced universal tags and containing 3′-terminal phosphorothioates were added and 25 additional cycles of PCR were performed.

Sequencing

[108] Sequencing of all of the libraries described above was performed using an Illumina GA IIx instrument as specified by the manufacturer. High-quality reads were grouped into UID families on the basis of their endogenous or exogenous UIDs. Only UID families with two or more members were considered.

REFERENCES

A1. Luria SE, Delbrück M (1943) Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* 28:491–511.

A2. Roach JC, et al. (2010) Analysis of genetic inheritance in a family quartet by wholegenome sequencing. *Science* 328:636–639.

A3. Durbin RM, et al.; 1000 Genomes Project Consortium (2010) A map of human genome variation from population-scale sequencing. *Nature* 467:1061–1073.

A4. Shibata D (2011) Mutation and epigenetic molecular clocks in cancer. *Carcinogenesis* 32:123–128.

A5. McMahon MA, et al. (2007) The HBV drug entecavir - effects on HIV-1 replication and resistance. *N Engl J Med* 356:2614–2621.

A6. Eastman PS, et al. (1998) Maternal viral genotypic zidovudine resistance and infrequent failure of zidovudine therapy to prevent perinatal transmission of human immunodeficiency virus type 1 in pediatric AIDS Clinical Trials Group Protocol 076. *J Infect Dis* 177:557–564.

A7. Chiu RW, et al. (2008) Noninvasive prenatal diagnosis of fetal chromosomal aneuploidy by massively parallel genomic sequencing of DNA in maternal plasma. *Proc Natl Acad Sci USA* 105:20458–20463.

A8. Fan HC, Blumenfeld YJ, Chitkara U, Hudgins L, Quake SR (2008) Noninvasive diagnosis of fetal aneuploidy by shotgun sequencing DNA from maternal blood. *Proc Natl Acad Sci USA* 105:16266–16271.

A9. Hoque MO, et al. (2003) High-throughput molecular analysis of urine sediment for the detection of bladder cancer by high-density single-nucleotide polymorphism array. *Cancer Res* 63:5723–5726.

A10. Thunnissen FB (2003) Sputum examination for early detection of lung cancer. *J Clin Pathol* 56:805–810.

A11. Diehl F, et al. (2008) Analysis of mutations in DNA isolated from plasma and stool of colorectal cancer patients. *Gastroenterology* 135:489–498.

A12. Barnes WM (1992) The fidelity of Taq polymerase catalyzing PCR is improved by an Nterminal deletion. *Gene* 112:29–35.

A13. Araten DJ, et al. (2005) A quantitative measurement of the human somatic mutation rate. *Cancer Res* 65:8111–8117.

A14. Campbell F, Appleton MA, Shields CJ, Williams GT (1998) No difference in stem cell somatic mutation between the background mucosa of right- and left-sided sporadic colorectal carcinomas. *J Pathol* 186:31–35.

A15. Tindall KR, Kunkel TA (1988) Fidelity of DNA synthesis by the Thermus aquaticus DNA polymerase. *Biochemistry* 27:6008–6013.

A16. Kunkel TA (1985) The mutational specificity of DNA polymerase-beta during in vitro DNA synthesis. Production of frameshift, base substitution, and deletion mutations. *J Biol Chem* 260:5787–5796.

APPENDIX

A17. van Dongen JJ, Wolvers-Tettero IL (1991) Analysis of immunoglobulin and T cell receptor genes. Part II: Possibilities and limitations in the diagnosis and management of lymphoproliferative diseases and related disorders. *Clin Chim Acta* 198:93–174.

A18. Grist SA, McCarron M, Kutlaca A, Turner DR, Morley AA (1992) In vivo human somatic mutation: Frequency and spectrum with age. *Mutat Res* 266:189–196.

A19. Liu Q, Sommer SS (2004) Detection of extremely rare alleles by bidirectional pyrophosphorolysis-activated polymerization allele-specific amplification (Bi-PAP-A): Measurement of mutation load in mammalian tissues. *Biotechniques* 36:156–166.

A20. Monnat RJ, Jr., Loeb LA (1985) Nucleotide sequence preservation of human mitochondrial DNA. *Proc Natl Acad Sci USA* 82:2895–2899.

A21. Shi C, et al. (2004) LigAmp for sensitive detection of single-nucleotide differences. *Nat Methods* 1:141–147.

A22. Keohavong P, Thilly WG (1989) Fidelity of DNA polymerases in DNA amplification. *Proc Natl Acad Sci USA* 86:9253–9257.

A23. Sidransky D, et al. (1991) Identification of p53 gene mutations in bladder cancers and urine samples. *Science* 252:706–709.

A24. Bielas JH, Loeb LA (2005) Quantification of random genomic mutations. *Nat Methods* 2:285–290.

A25. Vogelstein B, Kinzler KW (1999) Digital PCR. *Proc Natl Acad Sci USA* 96:9236–9241.

A26. Mitra RD, et al. (2003) Digital genotyping and haplotyping with polymerase colonies. *Proc Natl Acad Sci USA* 100:5926–5931.

A27. Chetverina HV, Samatov TR, Ugarov VI, Chetverin AB (2002) Molecular colony diagnostics: Detection and quantitation of viral nucleic acids by in-gel PCR. *Biotechniques* 33:150–152, 154, 156.

A28. Zimmermann BG, et al. (2008) Digital PCR: A powerful new tool for noninvasive prenatal diagnosis? *Prenat Diagn* 28:1087–1093.

A29. Dressman D, Yan H, Traverso G, Kinzler KW, Vogelstein B (2003) Transforming single DNA molecules into fluorescent magnetic particles for detection and enumeration of genetic variations. *Proc Natl Acad Sci USA* 100:8817–8822.

A30. Ottesen EA, Hong JW, Quake SR, Leadbetter JR (2006) Microfluidic digital PCR enables multigene analysis of individual environmental bacteria. *Science* 314:1464–1467.

A31. Quail MA, et al. (2008) A large genome center's improvements to the Illumina sequencing system. *Nat Methods* 5:1005–1010.

A32. Nazarian R, et al. (2010) Melanomas acquire resistance to B-RAF(V600E) inhibition by RTK or N-RAS upregulation. *Nature* 468:973–977.

A33. He Y, et al. (2010) Heteroplasmic mitochondrial DNA mutations in normal and tumour cells. *Nature* 464:610–614.

A34. Gore A, et al. (2011) Somatic coding mutations in human induced pluripotent stem cells. *Nature* 471:63–67.

A35. Dohm JC, Lottaz C, Borodina T, Himmelbauer H (2008) Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Res* 36:e105.

A36. Erlich Y, Mitra PP, delaBastide M, McCombie WR, Hannon GJ (2008) Alta-Cyclic: A selfoptimizing base caller for next-generation sequencing. *Nat Methods* 5:679–682.

A37. Rougemont J, et al. (2008) Probabilistic base calling of Solexa sequencing data. *BMC Bioinformatics* 9:431.

A38. Druley TE, et al. (2009) Quantification of rare allelic variants from pooled genomic DNA. *Nat Methods* 6:263–265.

A39. Vallania FL, et al. (2010) High-throughput discovery of rare insertions and deletions in large cohorts. *Genome Res* 20:1711–1718.

A40. McCloskey ML, Stöger R, Hansen RS, Laird CD (2007) Encoding PCR products with batch-stamps and barcodes. *Biochem Genet* 45:761–767.

A41. Parameswaran P, et al. (2007) A pyrosequencing-tailored nucleotide barcode design unveils opportunities for large-scale sample multiplexing. *Nucleic Acids Res* 35:e130.

A42. Craig DW, et al. (2008) Identification of genetic variants using bar-coded multiplexed sequencing. *Nat Methods* 5:887–893.

A43. Miner BE, Stöger RJ, Burden AF, Laird CD, Hansen RS (2004) Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. *Nucleic Acids Res* 32:e135.

A44. Herman DS, et al. (2009) Filter-based hybridization capture of subgenomes enables resequencing and copy-number detection. *Nat Methods* 6:507–510.

A45. Jones PA, Baylin SB (2007) The epigenomics of cancer. *Cell* 128:683–692.

A46. de Boer JG, Ripley LS (1988) An in vitro assay for frameshift mutations: Hotspots for deletions of 1 bp by Klenow-fragment polymerase share a consensus DNA sequence. *Genetics* 118:181–191.

A47. Eckert KA, Kunkel TA (1990) High fidelity DNA synthesis by the Thermus aquaticus DNA polymerase. *Nucleic Acids Res* 18:3739–3744.

A48. Kosuri S, et al. (2010) Scalable gene synthesis by selective amplification of DNA pools from high-fidelity microchips. *Nat Biotechnol* 28:1295–1299.

A49. Matzas M, et al. (2010) High-fidelity gene synthesis by retrieval of sequence-verified DNA identified using high-throughput pyrosequencing. *Nat Biotechnol* 28: 1291–1294.

A50. Li J, et al. (2008) Replacing PCR with COLD-PCR enriches variant DNA sequences and redefines the sensitivity of genetic testing. *Nat Med* 14:579–584.

A51. Eid J, et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* 323:133–138.

What is claimed is:

1. A method for detecting endometrial, fallopian tubal, or ovarian neoplasia or cancer, comprising:

testing a liquid Pap specimen collected from a human subject for a genetic or epigenetic change in one or more nucleic acids mutated in endometrial, fallopian tubal, or ovarian neoplasia or cancer; wherein the one or more nucleic acids are selected from the group consisting of: CTNNB1 , EGFR, PIK3CA, PTEN, TP53, BRAF, KRAS, AKT1 , NRAS, PPP2R1A, APC, FBXW7, ARID1A, CDKN2A, MLL2, RNF43, and FGFR2; and wherein the step of testing is performed by increasing the sensitivity of massively parallel sequencing instruments with an error reduction technique that allows for the detection of rare mutant alleles in a range of 1 mutant template among up to 1,000,000 wild-type templates, wherein the error reduction technique comprises:

assignment of a unique identifier (UID) to each nucleic acid;

amplification of each uniquely tagged nucleic acid to create UID families; and

redundant sequencing of the amplified nucleic acid.

2. The method of claim 1 wherein the change is a substitution mutation.

3. The method of claim 1 wherein the change is a rearrangement.

4. The method of claim 1 wherein the change is a deletion.

5. The method of claim 1 wherein the change is a loss or gain of methylation.

6. The method of claim 1 wherein the change is determined with respect to the bulk of the nucleic acids present in the liquid Pap specimen.

7. A method for detecting endometrial, fallopian tubal, or ovarian neoplasia or cancer, comprising:

Date Reçue/Date Received 2020-06-24

testing a liquid Pap specimen collected from the gynecologic tract of a human subject for one or more mutations in one or more nucleic acids selected from the group consisting of CTNNB1 , EGFR, PIK3CA, PTEN, TP53, BRAF, KRAS, AKT1 , NRAS, PPP2R1A, APC, FBXW7, ARID1A, CDKN2A, MLL2, RNF43, and FGFR2; wherein the step of testing is performed by increasing the sensitivity of massively parallel sequencing instruments with an error reduction technique that allows for the detection of rare mutant alleles in a range of 1 mutant template among up to 1,000,000 wild-type templates, wherein the error reduction technique comprises:

assignment of a unique identifier (UID) to each nucleic acid;

amplification of each uniquely tagged nucleic acid to create UID families; and

redundant sequencing of the amplified nucleic acid.

8. The method of claim 7 wherein the step of testing is performed on at least 3 of said nucleic acids.

9. The method of claim 7 wherein the step of testing is performed on at least 5 of said nucleic acids.

10. The method of claim 7 wherein the step of testing is performed on at least 7 of said nucleic acids.

11. The method of claim 7 wherein the step of testing is performed on at least 9 of said nucleic acids.

12. The method of claim 7 wherein the step of testing is performed on at least 11 of said nucleic acids.

13. The method of claim 7 wherein the step of testing is performed on at least 12 of said nucleic acids.

14. The method of claim 7 wherein the step of testing is performed in a multiplex assay.

Date Reçue/Date Received 2020-06-24

15. The method of claim 7 wherein the step of testing is repeated over time.

16. The method of claim 7 wherein the liquid Pap specimen is collected after surgical debulking of an ovarian tumor.

17. A kit for testing a panel of genes in Pap specimens for ovarian, fallopian tubal, or endometrial neoplasms or cancers, the kit comprising at least 3 probes or at least 3 primer pairs, wherein each probe or each primer of the primer pair in the kit comprises at least 15 nt of complementary sequence to one of a panel of genes, wherein the panel is complementary to at least 3 different genes, wherein the panel is selected from the group consisting of CTNNB1 , EGFR, PIK3CA, PTEN, TP53, BRAF, KRAS, AKT 1, NRAS, PPP2R1A, APC, FBXW7, ARID1A, CDKN2A, MLL2, RNF43, and FGFR2, wherein the probes are biotinylated or attached to beads, and wherein at least one member of each primer pair is attached to a bead.

18. The kit of claim 17 which comprises probes and wherein the probes are attached to a solid support.

19. The kit of claim 17 which comprises primer pairs, wherein the primer pairs prime synthesis of a nucleic acid of between 240 and 300 bp.

20. The kit of claim 17 which comprises primer pairs, wherein the primer pairs prime synthesis of a nucleic acid of between 200 and 325 bp.

21. The kit of claim 17 which comprises primer pairs, wherein the primer pairs prime synthesis of a nucleic acid of between 60 and 1000 bp.

22. The kit of claim 21 wherein at least one primer from each primer pair is attached to a solid support.

Date Reçue/Date Received 2020-06-24

23. The kit of claim 17 wherein the probe or primer comprises at least 20 nt of complementary sequence to one of the panel of genes.

24.A solid support comprising at least 3 probes attached thereto, wherein each probe on the solid support comprises at least 15 nt of complementary sequence to one of a panel of genes, wherein the panel is selected from the group consisting of CTNNB 1, EGFR, PIK3CA, PTEN, TP53, BRAF, KRAS, AKT1 , NRAS, PPP2R1A, APC, FBXW7, ARID 1A, CDKN2A, MLL2, RNF43, and FGFR2, wherein the panel is complementary to at least 3 different genes.

25. The method of claim 1 wherein the liquid Pap specimen is collected from the cervix.

26. The method of claim 7 wherein the liquid Pap specimen is collected from the cervix.

27. The method of claim 7 wherein the step of testing is performed by increasing the sensitivity of massively parallel sequencing instruments with an error reduction technique that allows for the detection of rare mutant alleles in a range of 1 mutant template among 5,000 to 1,000,000 wild-type templates.

28. The method of claim 1 wherein the step of testing is performed by increasing the sensitivity of massively parallel sequencing instruments with an error reduction technique that allows for the detection of 1 mutant template among 5,000 to 1,000,000 wild-type templates.

29. The method of claim 1 wherein the nucleic acid is selected from the group consisting of: AKT1, NRAS, APC, MLL2, RNF43, and FGFR2.

30. The method of claim 7 wherein the nucleic acid is selected from the group consisting of: AKT1, NRAS, APC, MLL2, RNF43, and FGFR2.

31. The kit of claim 17 wherein the panel is selected from the group consisting of: AKT1, NRAS, APC, MLL2, RNF43, and FGFR2.

Date Reçue/Date Received 2020-06-24

32. The solid support of claim 24 wherein the panel is selected from the group consisting of: AKT1, NRAS, APC, MLL2, RNF43, and FGFR2.

33. The method of claim 1, wherein the liquid Pap specimen comprises cells or cell fragments from endometrial or ovarian cancer.

34. The method of claim 1, wherein the ovarian cancer is chosen from a high-grade serous, endometrioid, clear cell, low-grade serous, or mucinous ovarian tumor.

35. The method of claim 1, wherein the endometrial cancer is a Type I endometrioid or a Type II non-endometrioid cancer.

36. The method of claim 7, wherein the liquid Pap specimen comprises cells or cell fragments from endometrial or ovarian cancer.

37. The method of claim 7, wherein the ovarian cancer is chosen from a high-grade serous, endometrioid, clear cell, low-grade serous, or mucinous ovarian tumor.

38. The method of claim 7, wherein the endometrial cancer is a Type I endometrioid or a Type II non-endometrioid cancer.

39. A method for detecting endometrial, fallopian tubal, or ovarian neoplasia or cancer, comprising:

testing an endocervical or an endometrial sample comprising cells or fragments from the endometrium, fallopian tube and ovary for a genetic or epigenetic change in one or more nucleic acids mutated in endometrial, fallopian tubal, or ovarian neoplasia or cancer; wherein the one or more nucleic acids are selected from the group consisting of: CTNNB1 , EGFR, PIK3CA, PTEN, TP53, BRAF, KRAS, AKT1 , NRAS, PPP2R1A, APC, FBXW7, ARID1A, CDKN2A, MLL2, RNF43, and FGFR2; and wherein the step of testing is performed by increasing the sensitivity of massively parallel sequencing instruments with an error reduction technique that

Date Reçue/Date Received 2020-06-24

allows for the detection of rare mutant alleles in a range of 1 mutant template among up to 1,000,000 wild-type templates, wherein the error reduction technique comprises:

assignment of a unique identifier (UID) to each nucleic acid;

amplification of each uniquely tagged nucleic acid to create UID families; and

redundant sequencing of the amplified nucleic acid.

40. The method of claim 39, wherein the change is a substitution mutation; a rearrangement; a deletion; or a loss or gain of methylation.

41. A method for detecting endometrial, fallopian tubal, or ovarian neoplasia or cancer, comprising:

testing an endocervical or an endometrial sample comprising cells or fragments from the endometrium, fallopian tube and ovary for one or more mutations in one or more nucleic acids selected from the group consisting of CTNNB1 , EGFR, PIK3CA, PTEN, TP53, BRAF, KRAS, AKT1 , NRAS, PPP2R1A, APC, FBXW7, ARID1A, CDKN2A, MLL2, RNF43, and FGFR2; wherein the step of testing is performed by increasing the sensitivity of massively parallel sequencing instruments with an error reduction technique that allows for the detection of rare mutant alleles in a range of 1 mutant template among up to 1,000,000 wild-type templates, wherein the error reduction technique comprises:

assignment of a unique identifier (UID) to each nucleic acid;

amplification of each uniquely tagged nucleic acid to create UID families; and

redundant sequencing of the amplified nucleic acid.

42. The method of claim 41, wherein the step of testing is performed on at least 3, at least 5, at least 7, at least 9, at least 11, or at least 12 of said nucleic acids.

43. The method of claim 41, wherein the step of testing is performed in a multiplex assay.

44. The method of claim 41, wherein the step of testing is repeated over time.

Date Reçue/Date Received 2020-06-24

45. A kit for testing a panel of genes in Pap specimens for ovarian, fallopian tubal, or endometrial neoplasms or cancers, the kit comprising at least 3 probes or at least 3 primer pairs, wherein each probe or each primer of the primer pair in the kit comprises at least 15 nt of complementary sequence to one of a panel of genes, wherein the panel is complementary to at least 3 different genes, wherein the panel is selected from the group consisting of CTNNB1 , EGFR, PIK3CA, PTEN, TP53, BRAF, KRAS, AKT 1, NRAS, PPP2R1A, APC, FBXW7, ARID1A, CDKN2A, MLL2, RNF43, and FGFR2.

46. The kit of claim 45, wherein the probes are biotinylated or attached to a solid support.

47. The kit of claim 46, wherein the solid support is a bead or an array.

48. The kit of claim 45, wherein at least one member of each primer pair is attached to a solid support.

49. The kit of claim 48, wherein the solid support is a bead or an array.

50. The kit of claim 45, wherein each primer comprises a 5' universal tag for universal amplification.

51. The kit of claim 50, wherein one primer of each primer pair comprises a unique identifier interposed between the 5' universal tag and the at least 15 nt of complementary sequence.

52. The kit of claim 51, wherein the unique identifier comprises 14 degenerate N bases.

Date Reçue/Date Received 2020-06-24

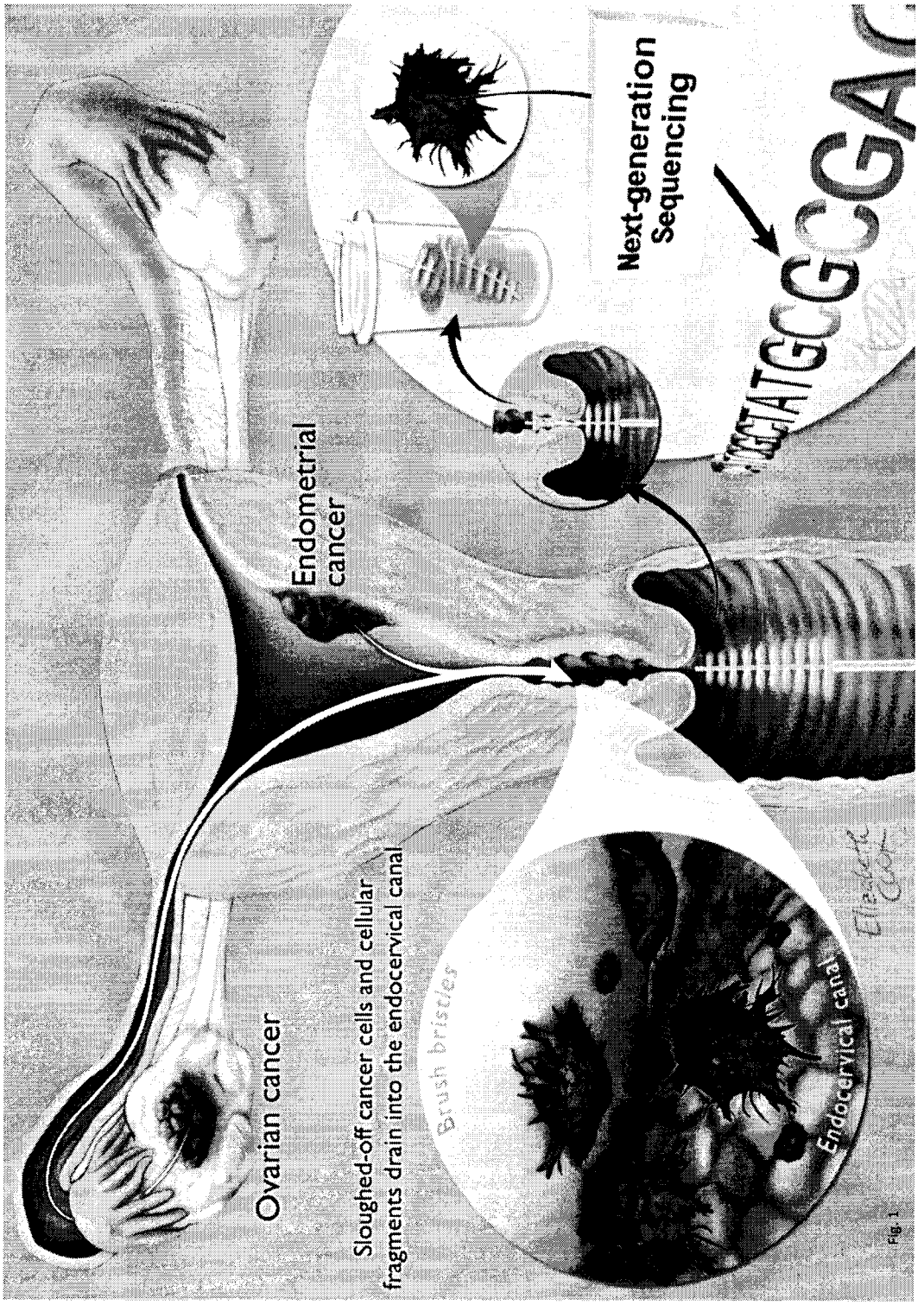

Ovarian cancer
Endometrial cancer
Sloughed-off cancer cells and cellular fragments drain into the endocervical canal
Brush bristles
Endocervical canal
Next-generation Sequencing
Fig. 1

Fig. 1

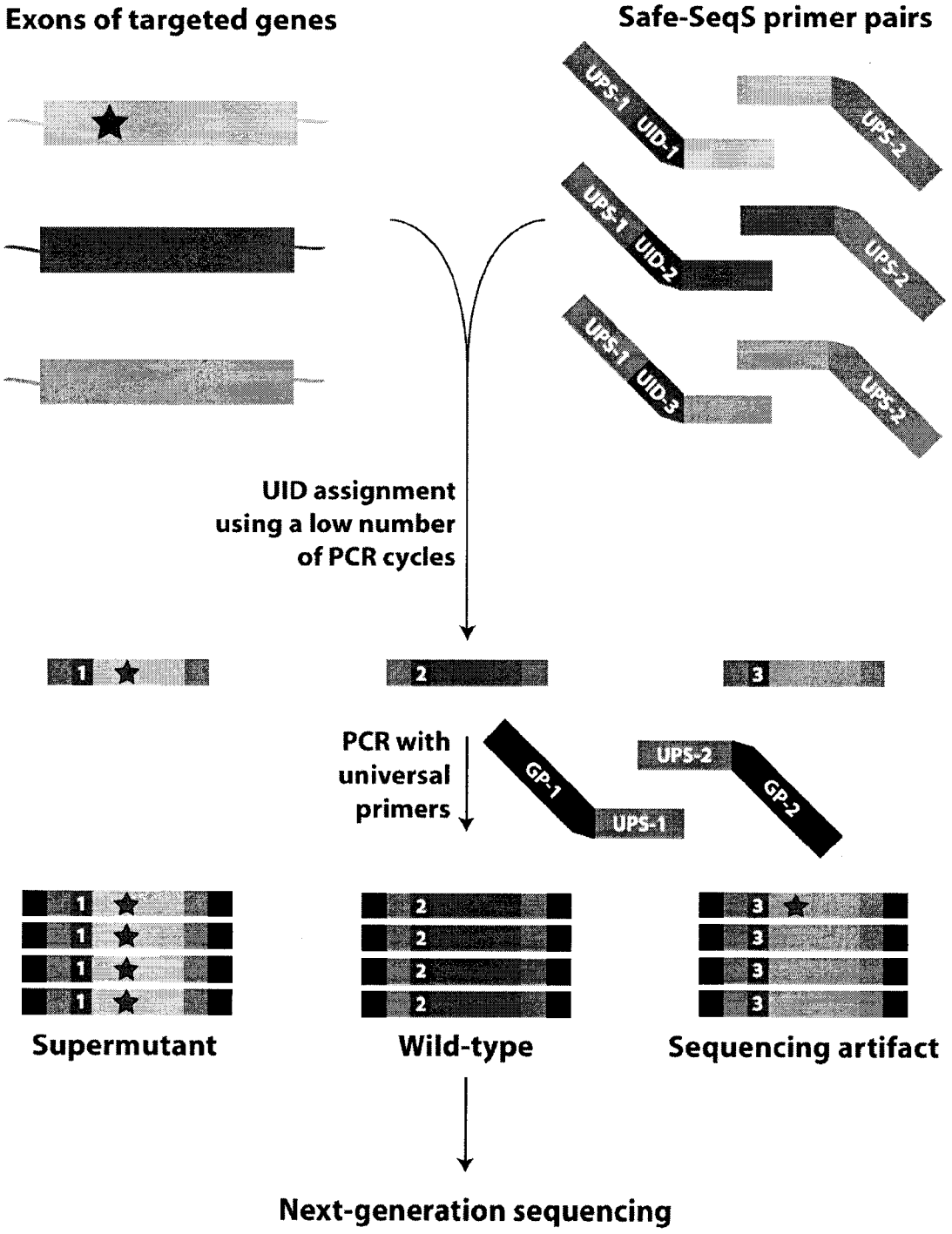

Exons of targeted genes
Safe-SeqS primer pairs
UPS-1
UID-1
UPS-2
UPS-1
UID-2
UPS-2
UPS-1
UID-3
UPS-2
UID assignment
using a low number
of PCR cycles
1
2
3
PCR with
universal
primers
GP-1
UPS-1
UPS-2
GP-2
Supermutant
Wild-type
Sequencing artifact
Next-generation sequencing

Fig. 2

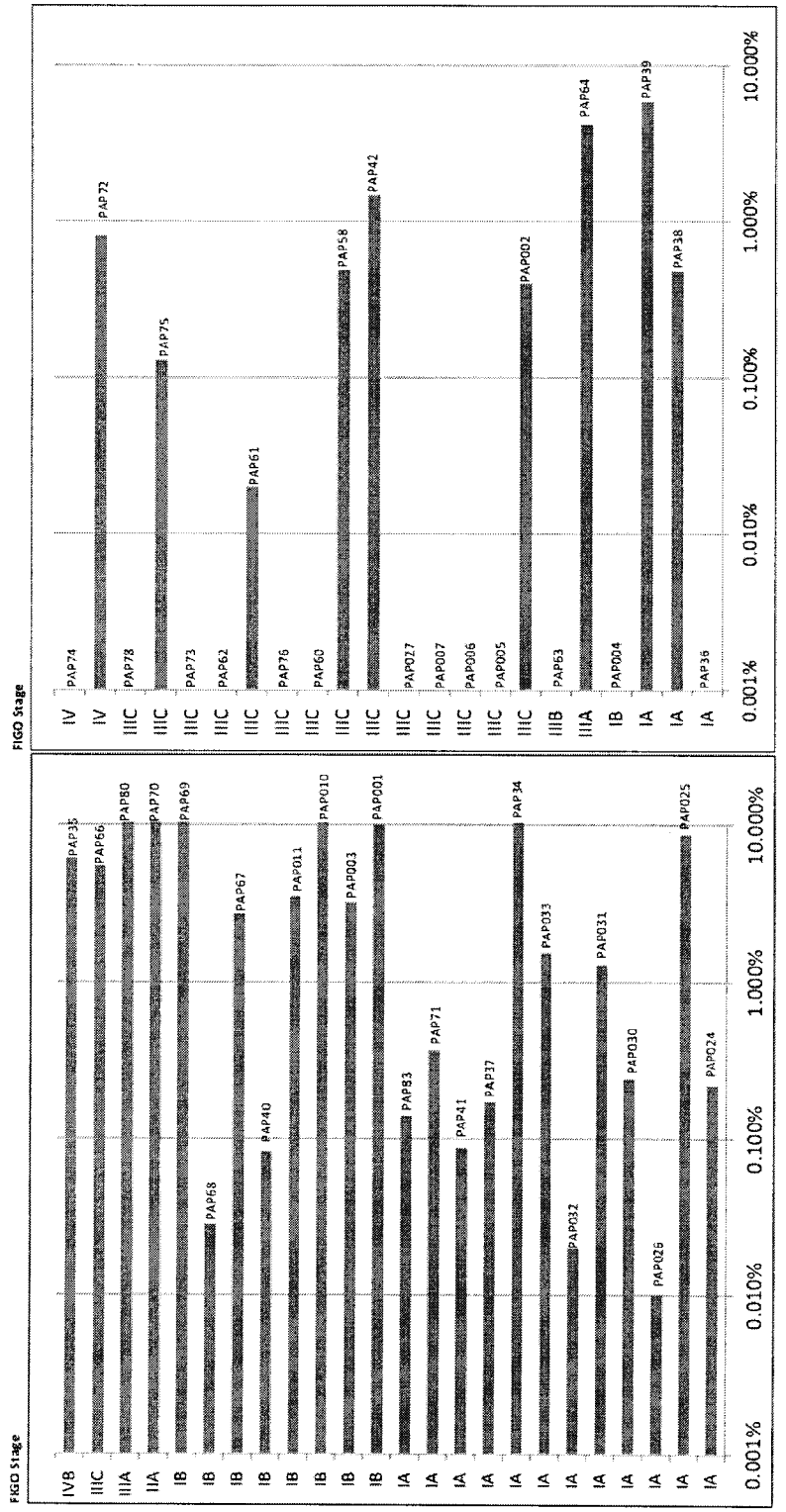

Endometrial Tumors
FIGO Stage
IVB PAP35
IIIC PAP66
IIIA PAP80
IIA PAP70
IB PAP69
IB PAP68
IB PAP67
IB PAP40
IB PAP011
IB PAP010
IB PAP003
IB PAP001
IA PAP83
IA PAP71
IA PAP41
IA PAP37
IA PAP34
IA PAP033
IA PAP032
IA PAP031
IA PAP030
IA PAP026
IA PAP025
IA PAP024
0.001% 0.010% 0.100% 1.000% 10.000%
Ovarian Tumors
FIGO Stage
IV PAP74
IV PAP72
IIIC PAP78
IIIC PAP75
IIIC PAP73
IIIC PAP62
IIIC PAP61
IIIC PAP76
IIIC PAP60
IIIC PAP58
IIIC PAP42
IIIC PAP027
IIIC PAP007
IIIC PAP006
IIIC PAP005
IIIC PAP002
IIIB PAP63
IIIA PAP64
IB PAP004
IA PAP39
IA PAP38
IA PAP36
0.001% 0.010% 0.100% 1.000% 10.000%
% mutant alleles in liquid Pap smear sample

Fig. 3

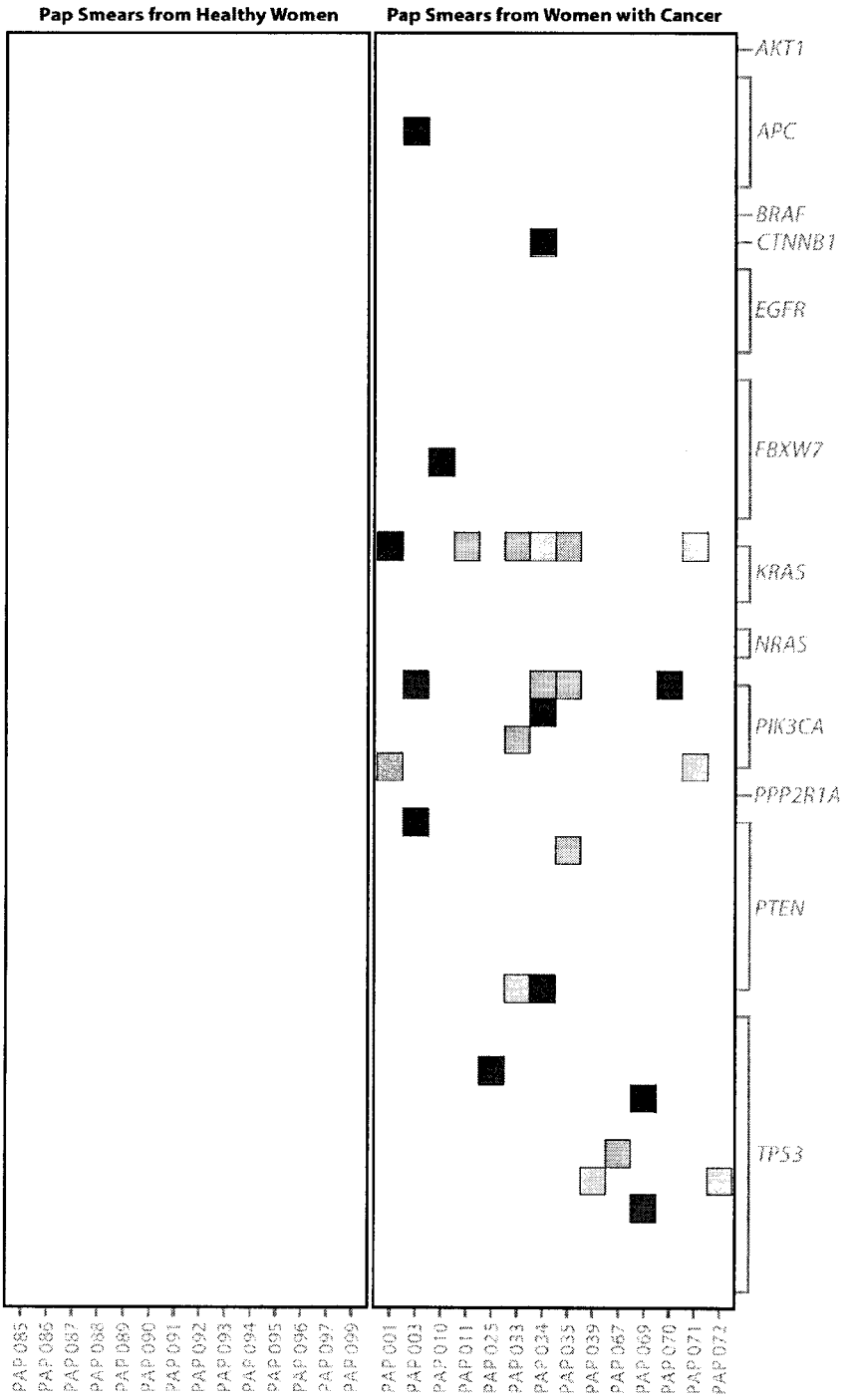
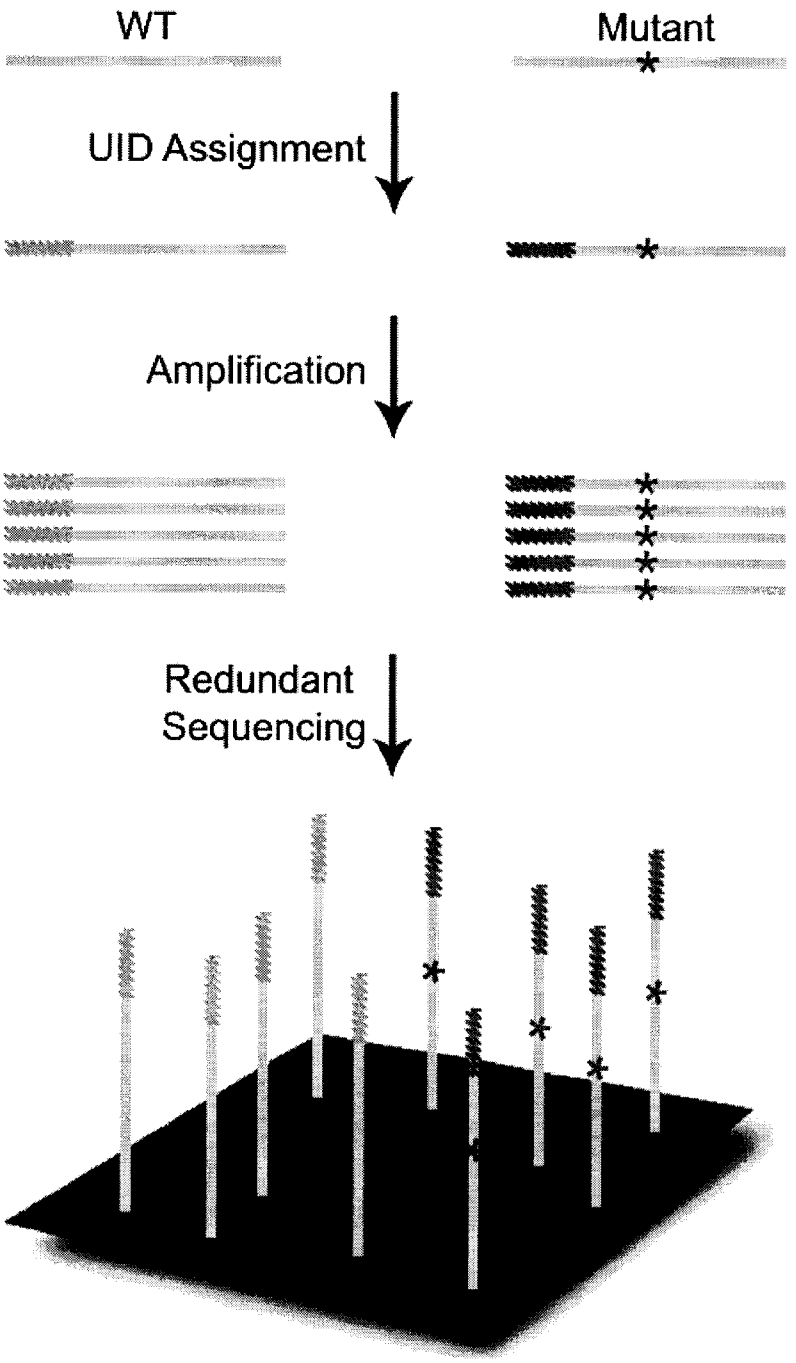
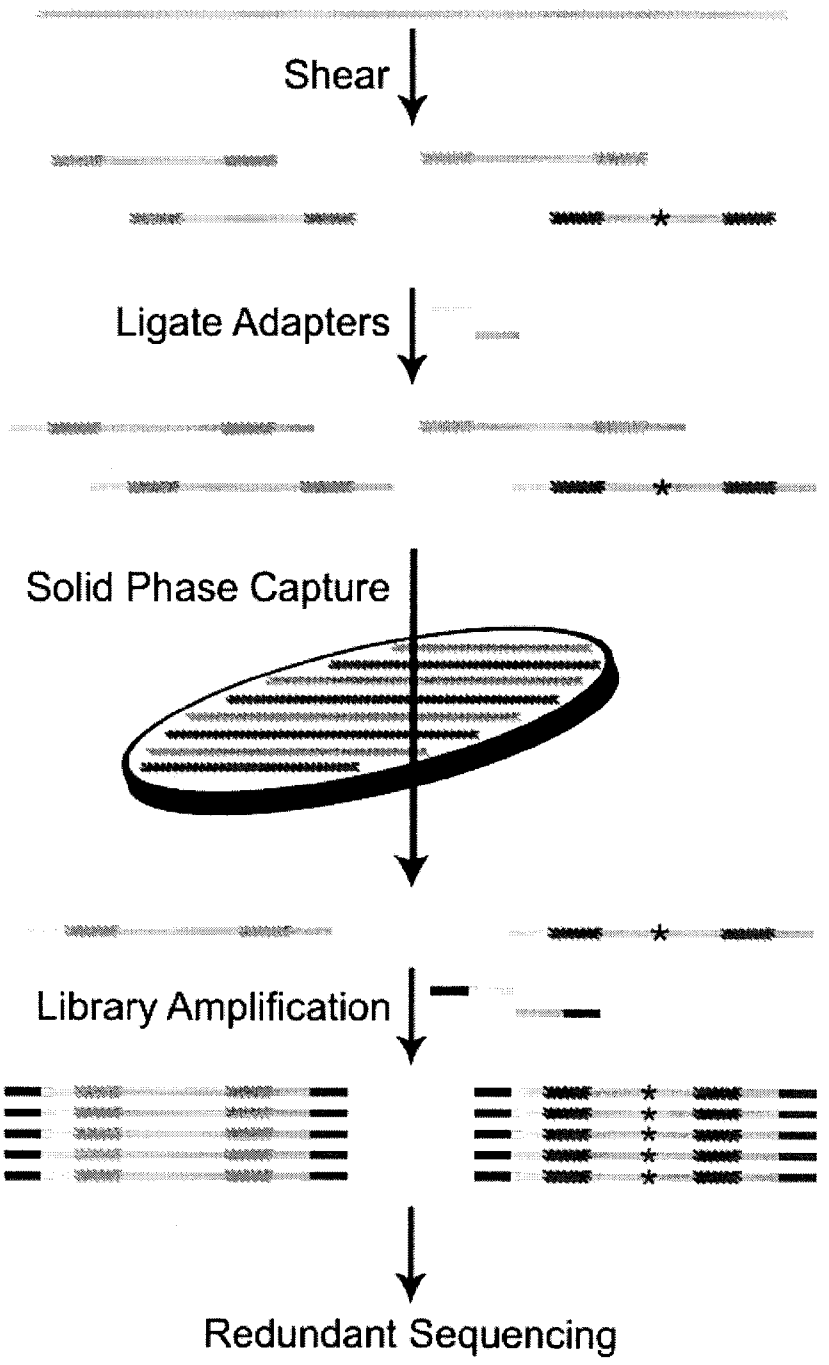

Pap Smears from Healthy Women
Pap Smears from Women with Cancer
AKT1
APC
BRAF
CTNNB1
EGFR
FBXW7
KRAS
NRAS
PIK3CA
PPP2R1A
PTEN
TP53
PAP085
PAP086
PAP087
PAP088
PAP089
PAP090
PAP091
PAP092
PAP093
PAP094
PAP095
PAP096
PAP097
PAP099
PAP001
PAP003
PAP010
PAP011
PAP025
PAP033
PAP034
PAP035
PAP039
PAP067
PAP069
PAP070
PAP071
PAP072

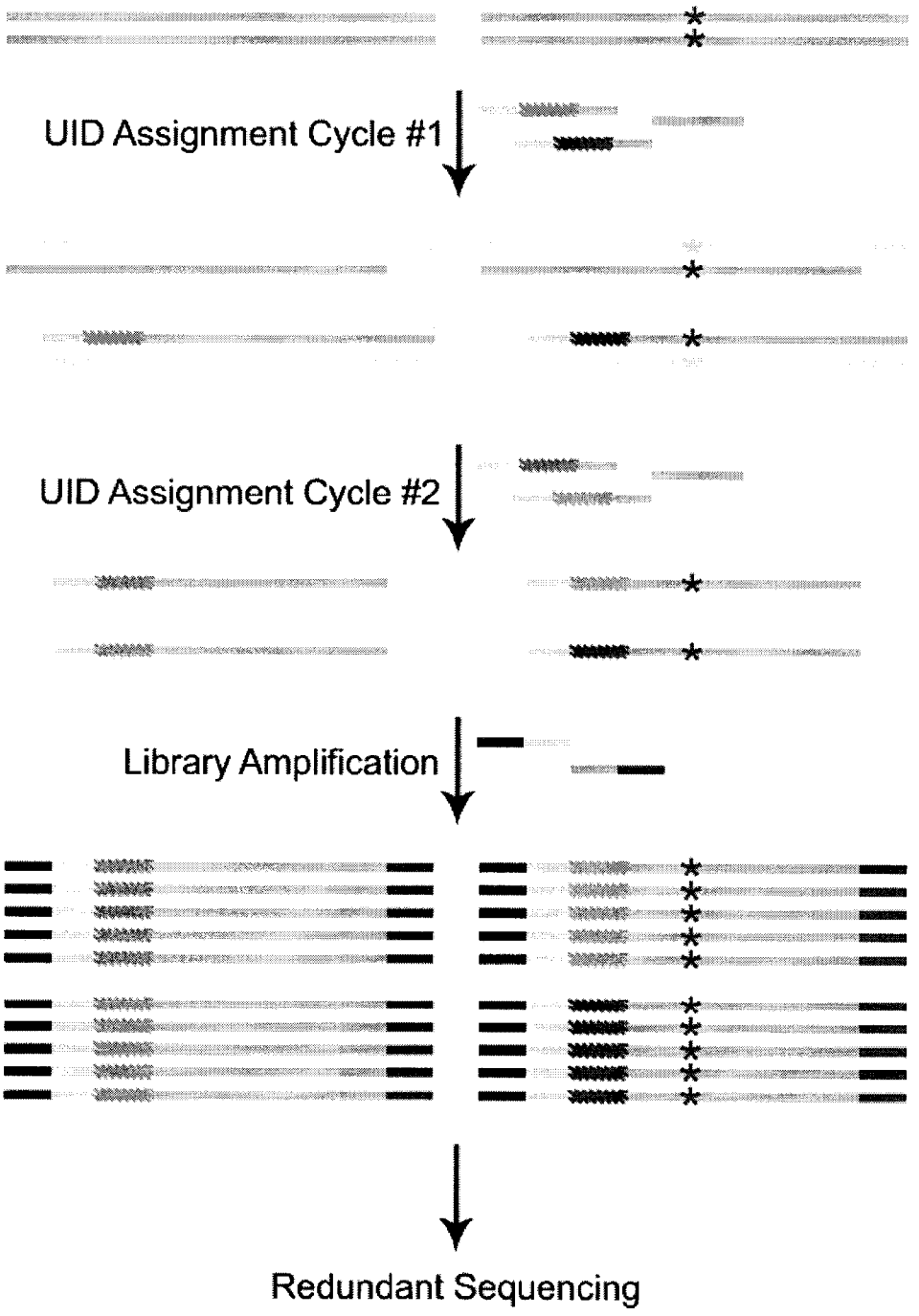
Fig. 4



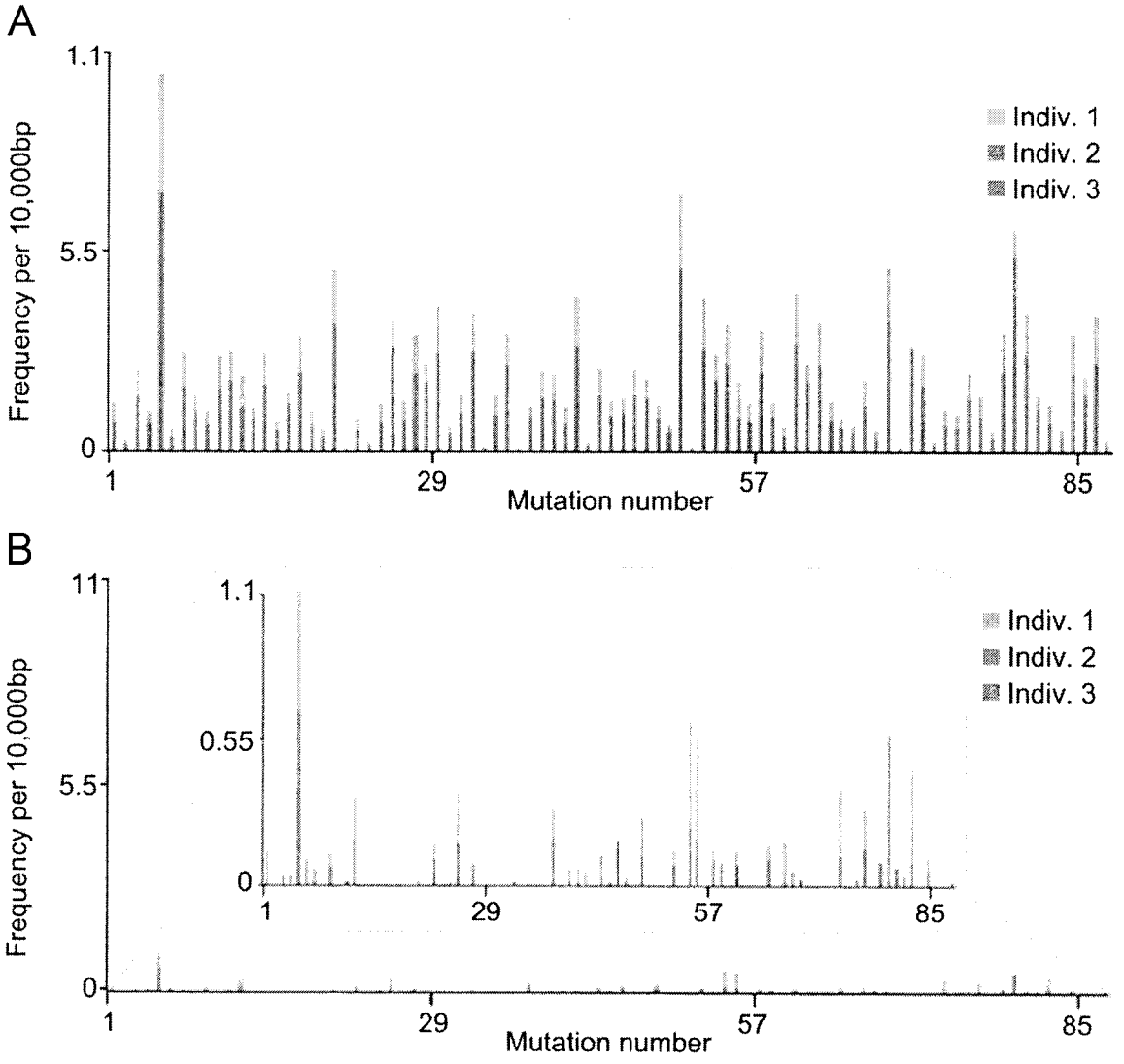
PRIOR ART
Fig. 5



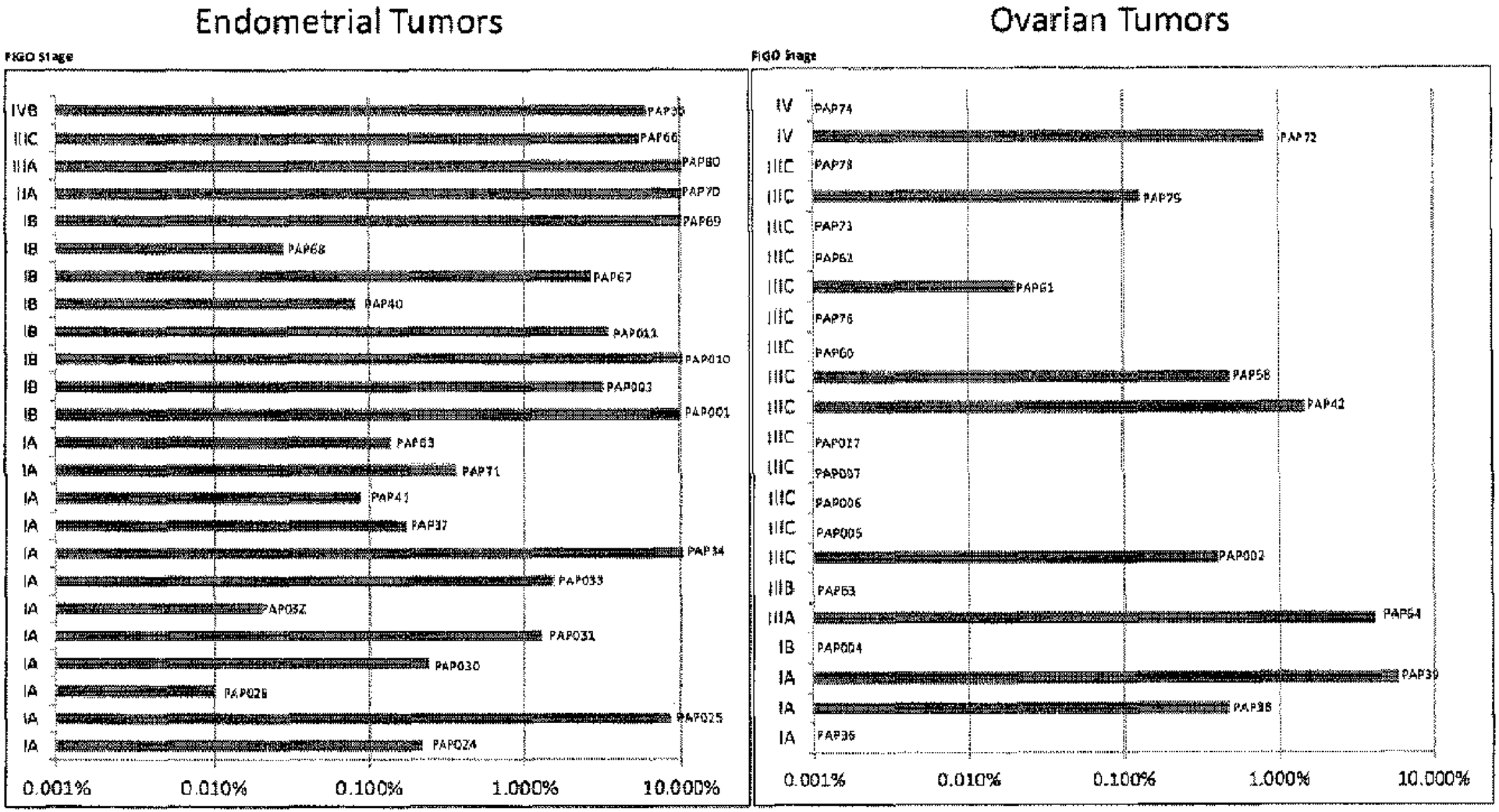
PRIOR ART
Fig. 6



PRIOR ART
Fig. 7



PRIOR ART
Fig. 8



% mutant alleles in liquid Pap smear sample