Genomic instability analysis in DNA from Papanicolaou test provides proof-of-principle early diagnosis of high-grade serous ovarian cancer

Lara Paracchini[1,2]†, Laura Mannarino[1,2]†, Chiara Romualdi[3]†, Riccardo Zadro[2], Luca Beltrame[2], Ilaria Fuso Nerini[2], Paolo Zola[4], Maria E. Laudani[4], Eva Pagano[5], Livia Giordano[5], Robert Fruscio[6], Fabio Landoni[6], Silvia Franceschi[7], Maria L. Dalessandro[2], Vincenzo Canzonieri[7,8], Luca Bocciolone[9], Domenica Lorusso[10], Cristina Bosetti[11], Francesco Raspagliesi[12], Isabella M. G. Garassino[13], the TOWARDS group‡, Maurizio D′Incalci[1,2]*, Sergio Marchini[2]

Late diagnosis and the lack of screening methods for early detection define high-grade serous ovarian cancer (HGSOC) as the gynecological malignancy with the highest mortality rate. In the work presented here, we investigated a retrospective and multicentric cohort of 250 archival Papanicolaou (Pap) test smears collected during routine gynecological screening. Samples were taken at different time points (from 1 month to 13.5 years before diagnosis) from 113 presymptomatic women who were subsequently diagnosed with HGSOC (pre-HGSOC) and from 77 healthy women. Genome instability was detected through low-pass whole-genome sequencing of DNA derived from Pap test samples in terms of copy number profile abnormality (CPA). CPA values of DNA extracted from Pap test samples from pre-HGSOC women were substantially higher than those in samples from healthy women. Consistently with the longitudinal analysis of clonal pathogenic *TP53* mutations, this assay could detect HGSOC presence up to 9 years before diagnosis. This finding confirms the continual shedding of tumor cells from fimbriae toward the endocervical canal, suggesting a new path for the early diagnosis of HGSOC. We integrated the CPA score into the EVA (early ovarian cancer) test, the sensitivity of which was 75% (95% CI, 64.97 to 85.79), the specificity 96% (95% CI, 88.35 to 100.00), and the accuracy 81%. This proof-of-principle study indicates that the early diagnosis of HGSOC is feasible through the analysis of genomic alterations in DNA from endocervical smears.

INTRODUCTION

Ovarian cancer is the most lethal gynecologic malignancy characterized by severe aggressiveness and dismal prognosis, with high-grade serous disease (HGSOC) being the most frequent and lethal histotype (*1*). More than 70% of patients diagnosed with HGSOC die within 5 years from diagnosis, the main reason being the delay in diagnosis because of the lack of specific symptoms in the early phases of the disease. The 5-year overall survival rate for stage I HGSOC is above 90%. In contrast, it is around 30% in patients with advanced disease when the tumor has already disseminated into the abdominal cavity (stage III) or even lower in patients who present with distant metastases (stage IV) (*2*, *3*). The high lethality and the large differences in the curability between stages make the early diagnosis of HGSOC potentially capable of improving survival.

Attempts over the years to develop early detection tests for HGSOC have been disappointing, generating skepticism about their overall feasibility. Several screening studies based on traditional methods of diagnosis, such as longitudinal determination of the plasma biomarker cancer antigen 125 (CA-125) alone or combined with transvaginal ultrasound imaging, failed to reduce HGSOC mortality (*4*, *5*). The most likely explanation for this failure is that the determination of such biomarkers is not sufficiently sensitive to reveal the initial phases of tumor development.

An effective strategy to detect HGSOC at a very early stage would benefit from the exploitation of recent insights into the natural history of HGSOC and its early genomic lesions. In most cases, HGSOC does not originate in the ovary but in the fallopian tubes as serous tubal intraepithelial carcinoma (STIC) (*6*, *7*). Evolutionary analysis shows that STIC presents some tumor-specific genomic alterations, such as clonal tumor protein p53 gene (*TP53*) mutation and aneuploidy, that are also found in both stage I (*8*) and metastatic HGSOC (*9*). Furthermore, recent evidence from large-scale pan-cancer studies suggests that, in addition to *TP53* mutations, loss of genomic material occurs early in HGSOC progression, leading to the hypothesis that these aberrant clones can be molecularly intercepted before they have fully developed their potential malignancy (*10*). Thus, a crucial issue for developing tools for the early

[1] Department of Biomedical Sciences, Humanitas University, Pieve Emanuele, Milan 20072, Italy. [2] Laboratory of Cancer Pharmacology, IRCCS Humanitas Research Hospital, Rozzano, Milan 20089, Italy. [3] Department of Biology, University of Padua, Padua 35121, Italy. [4] Department of Surgical Science, University of Turin, Turin 10126, Italy. [5] Unit of Clinical Epidemiology, Città della Salute e della Scienza Hospital, University of Turin and CPO Piemonte, Turin 10126, Italy. [6] Department of Obstetrics and Gynaecology, Università degli Studi Milano-Bicocca, Fondazione IRCCS San Gerardo dei Tintori, Monza 20900, Italy. [7] Centro di Riferimento Oncologico di Aviano IRCCS, Aviano, Pordenone 33081, Italy. [8] Department of Medical, Surgical and Health Sciences, University of Trieste, Trieste 34149, Italy. [9] Department of Obstetrics and Gynecology, IRCCS San Raffaele Hospital, Milan 20132, Italy. [10] Gynecologic Oncology Unit, Fondazione Policlinico Universitario A. Gemelli IRCCS, Rome 00168, Italy. [11] Unit of Cancer Epidemiology, Department of Oncology, Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Milan 20156, Italy. [12] Gynecological Oncology Unit, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan 20133, Italy. [13] Department of Medical Oncology and Hematology, IRCCS Humanitas Research Hospital, Rozzano, Milan 20089, Italy.
*Corresponding author. Email: maurizio.dincalci@hunimed.eu
†These authors contributed equally to this work.
‡Individual members of the TOWARDS group are listed at the end of the manuscript.

diagnosis of HGSOC is to select the optimal strategy to intercept such early molecular lesions.

Fallopian tubes are anatomically connected with the uterine cavity, and the passage of HGSOC cells from the STIC toward the cervical canal has been cytologically and molecularly demonstrated by many studies recently reviewed by Biskup *et al.* (*11*). Kinde *et al.* (*12*) provided the first demonstration that tumor DNA shed from ovarian cancer cells may be present in Papanicolaou (Pap) smear specimens. They analyzed a panel of 12 genes and found the same mutations in tumor tissue and liquid Pap smear specimens from the same patients in 41% of cases. Other studies reproduced similar findings on panels of 18 (*13*) or 8 genes (*14*), further corroborating the evidence that DNA from ovarian cancer cells can be detected in cervical smears. Further studies focused on *TP53* mutations common in HGSOC (*15*–*18*). By using ultrasensitive methods, the same tumor *TP53* mutations were found in endocervical swabs in more than 60% of cases and up to 80% in the lavage of the uterine cavity (*19*). *TP53* mutations were found even in DNA purified from Pap smear samples (pDNA) of patients at an early tumor stage. Arildsen *et al.* (*16*) reported mutated *TP53* in pDNA in two patients with stage IIA HGSOC. Paracchini *et al.* (*18*) found pathogenic *TP53* mutation in pDNA purified from archival Pap test smears taken up to 6 years before HGSOC diagnosis, a finding consistent with theoretical mathematical models predicting a temporal window of at least 5 to 6 years from the development of STIC to the initiation of HGSOC (*20*).

Despite these findings, several drawbacks limit the use of *TP53* aberrations as a suitable tool for the development of a diagnostic test. The workflows require a priori knowledge of the specific *TP53* mutation known to occur in the tumor. In a screening setting, the presence and genotype of tumors would not be known before evaluation. In addition, recently published data suggest that *TP53* is not a suitable biomarker to intercept early tumor samples because there is an abundant background of noncancer *TP53* mutations in normal tissues, which may confound cancer-specific signal detection (*21*). Therefore, the aim of the present study was to provide experimental evidence that analysis of somatic copy number alterations (SCNAs) in the pDNA, incorporated into a test that we have named EVA (Early oVArian cancer) test, can be used for the early diagnosis of HGSOC.

RESULTS

Characteristics of HGSOC cases and of healthy women controls

A retrospective and multicentric collection of 250 archival Pap test smears withdrawn at different time points from 113 presymptomatic women who were subsequently diagnosed with HGSOC (pre-HGSOC) and from 77 healthy women (HW) collected during routine gynecological screening were selected for this study (table S1). The REMARK (Reporting Recommendations for Tumor Marker Prognostic Studies) diagram for patient selection is depicted in fig. S1. The entire collection was divided into two groups, named in the following as groups A (*n* = 62) and B (*n* = 51), balanced for clinical and pathological features (tables S2 and S3). Cases enrolled in group A were used to develop the EVA test on the basis of the analysis of genome-wide copy number instability as a biomarker of early cancer progression (Fig. 1). Cases enrolled in group B were used to validate, through analysis of clonal pathogenic *TP53* somatic variants, the previous findings that shedding of tumor cells from fimbriae to the uterine cavity is a continual and common event in cases who progress toward HGSOC (*18*). The median age at diagnosis was 60 and 61 years for groups A and B, respectively (range, 40 to 81 years for group A, and 42 to 80 for group B). The median age of the HW who underwent the Pap test was 43 years (range, 20 to 64 years). Because Pap test analysis is no longer recommended in postmenopausal women, it was difficult to collect Pap test smears from HW at a similar age as the pre-HGSOC women. Most cases were diagnosed as stage III/IV, according to the International Federation of Gynecology and Obstetrics (FIGO), with 87% in group A and 88% in group B. At the time of writing, all HW were still alive with no evidence of gynecological tumors or any other neoplastic or other gynecological inflammatory diseases. The time from the collection of archival samples to the time of HGSOC diagnosis ranged from 0 months to 13.6 years for group A and from 0 months to 9.5 years for group B. Within the entire collection of patients included in this study, 41 patients (36%) provided more than one archival sample, including 5 patients who provided four samples, 6 patients who provided three samples, and 30 patients who provided two Pap test samples. Seventy-two patients gave a single archival sample (tables S2 and S3).

Spectrum of SCNAs in pDNA

In contrast to single nucleotide variants, SCNAs are rarely found in normal tissues, although they are common in cancer (particularly in HGSOC) (*22*). This raises the question as to whether the detection of these genomic alterations could improve early diagnosis. To address this question, we used a shallow whole-genome sequencing (sWGS) approach to detect SCNA in DNA purified from tumor samples (tDNA) and in pDNA from pre-HGSOC women from group A (fig. S2). First, we evaluated the global SCNA traits in tDNA. Analysis with GISTIC (Genomic Identification of Significant Targets in Cancer) software confirmed the marked and heterogeneous SCNA profile of tDNA, with *3q26.31*/*8q24.12* and *16q21*/*12q24.13* being the genomic regions most frequently affected by gain or loss of genomic material, respectively (fig. S3, table S4). Then, to demonstrate the presence of DNA derived from tumor cells in Pap test smear, we investigated the presence of SCNA in the entire group of pDNA. Most pDNA from pre-HGSOC samples (*n* = 58, 89%; table S5) had detectable SCNA, involving the loss or gain of broad chromosomal regions. We thus questioned the possible pathological origin of such alterations by comparing the genome instability profiles of tDNA with those observed in matched pDNA from women with pre-HGSOC. Regions in each sample pair were considered to be in common if their reciprocal overlap was at least 15% of their length and with the same SCNA call (gain or loss). Common regions of genomic alteration were called in 75% (*n* = 51) of the sample pairs, whereas in the remaining 25% (*n* = 17), there was no overlap due to potentially insufficient sequencing resolution, sample quality, or intratumor heterogeneity in the mutational process (table S6). Among the overlapping regions, we found that 20 regions shared across more than 10 pre-HGSOC tDNA/pDNA pairs (table S7, A and B). Of these, eight regions with copy number loss belonged to chromosomes *16p* and *22q* (table S7A), and 12 regions with copy number gain were on chromosomes *6p* and *12p* (table S7B). Figure 2 depicts the copy number profiles derived from tDNA and three paired pDNAs taken at different time points from a representative pre-

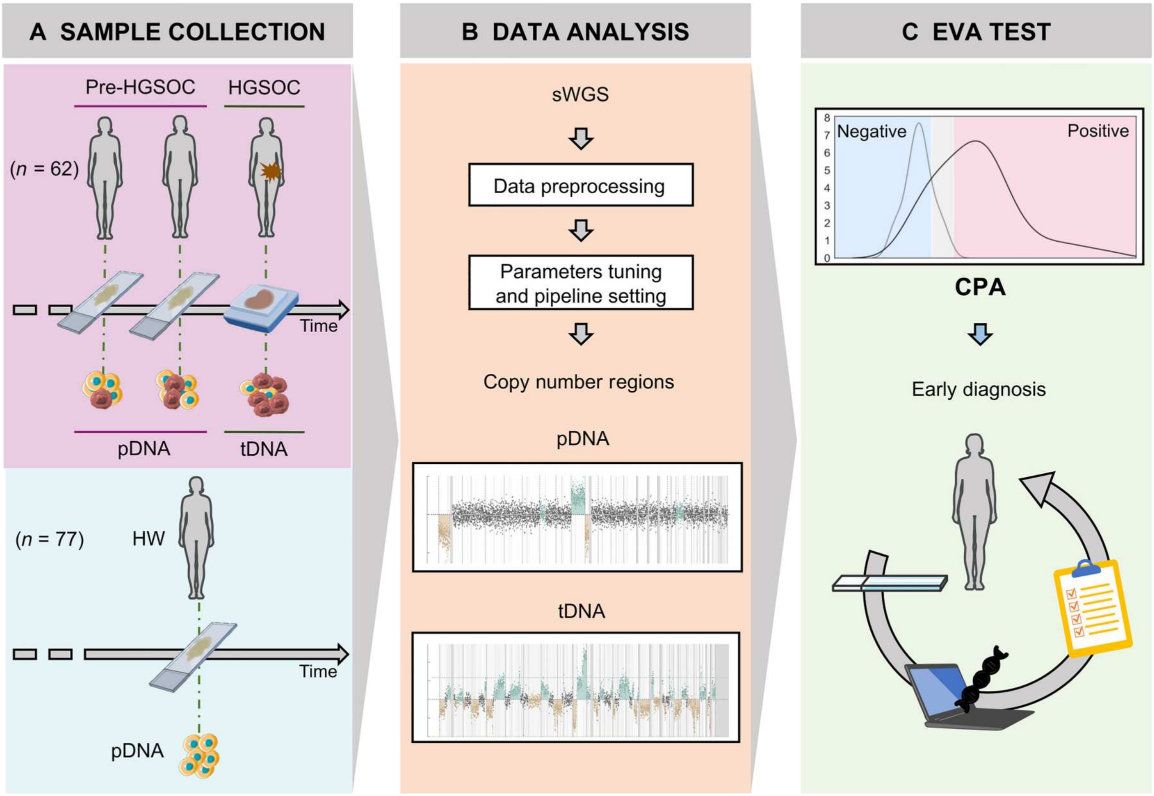

Fig. 1. Workflow of the EVA test. The workflow of this study can be divided into three sections. (**A**) Sample collection: A retrospective collection of FFPE tumor samples and matched Pap test smears from 62 patients with HGSOC was collected. Pap tests were collected at different time points before diagnosis. tDNA and pDNA were purified for downstream genomic instability analysis. As a control, pDNA was purified from 77 Pap test smears from HW. (**B**) Data analysis: sWGS sequencing was used to derive a global overview of genome instability in pDNA and tDNA. (**C**) EVA test: The CPA score was used to quantify the overall amount of genome affected by instability. The analysis identified two different thresholds that distribute the CPA scores into three distinct areas: negative for genomic alterations (blue area), area of uncertainty (gray area), and positive for genomic alterations (red area). HGSOC, high-grade serous ovarian cancer; pDNA, DNA from Pap test smears; tDNA, DNA from FFPE tumor tissues; HW, healthy women; sWGS, shallow whole-genome sequencing; CPA, copy number profile abnormality.

HGSOC woman (ID: 1240-11; P1 = 4.3 years; P3 = 6.2 years; P4 = 7.5 years). The tDNA was characterized by a high number of genomic aberrations across all chromosomes, whereas the pDNAs showed a prominently flat profile, a largely euploid genome, with few regions of DNA gain or loss. In these regions, the only loss on chromosome 16 (locus *p11.1*) was detectable in pDNAs at all time points, and in the tDNA sample (table S6 and fig. S4), whereas other regions with gain or loss, such as loss of chromosome 8 at P1, were private to each time point (Fig. 2). From a pathological standpoint, the loss of genomic material on chromosome 16 has been reported as an early genomic alteration in HGSOC (*10*). In conclusion, our data revealed that the genomic profile of pDNA from pre-HGSOC samples is characterized by both most SCNAs that make up the genome of HGSOC and pathogenic mutations in *TP53*. These abnormalities were seen in samples collected up to 6 years before diagnosis (*18*).

EVA test: Measuring the overall landscape of SCNA in pDNA

The data presented above suggest that the SCNA information derived from the pDNA of women who progress to malignancy displays generalized disorder across the genome that varies among samples and over time. We reasoned that a measure that summarizes the overall genomic complexity rather than the analysis of individual SCNA could be useful to allow discrimination of Pap tests between HW and women who progress to HGSOC. As a measure of the overall genomic instability, we used the copy number profile abnormality (CPA) score as previously published (*23*). The CPA provides a comprehensive quantification of unbalanced genomic traits: the higher the CPA value, the greater the genomic instability. We initially focused our analysis on those SCNA reported in the ovarian cohort of the Cancer Genome Atlas Ovarian Cancer (TCGA-OV) collection by restricting the analysis using a so-called "blacklist-A" as the hallmark of HGSOC (*n* = 80; tables S8 and S9).

We saw that pDNA from HW had statistically different distributions of CPA (Mann-Whitney test, $P < 0.0001$) compared with the CPA derived from pre-HGSOC (Fig. 3A). These distributions enabled us to define three different intervals: (i) The first ($0 \leq \text{CPA} < 0.0887$) included pDNA from HW (*n* = 20, 80%) and pDNA from 11 pre-HGSOC (17%) with no SCNA. We arbitrarily defined this CPA range as "negative" for genomic alterations. (ii) The second interval ($0.0887 \leq \text{CPA} < 0.11207$) encompassed pDNA from HW (*n* = 5, 20%) and patients progressing to HGSOC (*n* = 10, 15%). We defined this range as an "area of uncertainty" (gray zone). (iii) The third interval ($\text{CPA} \geq 0.11207$) comprised pre-HGSOC pDNA (*n* = 44, 68%), the visual inspection of

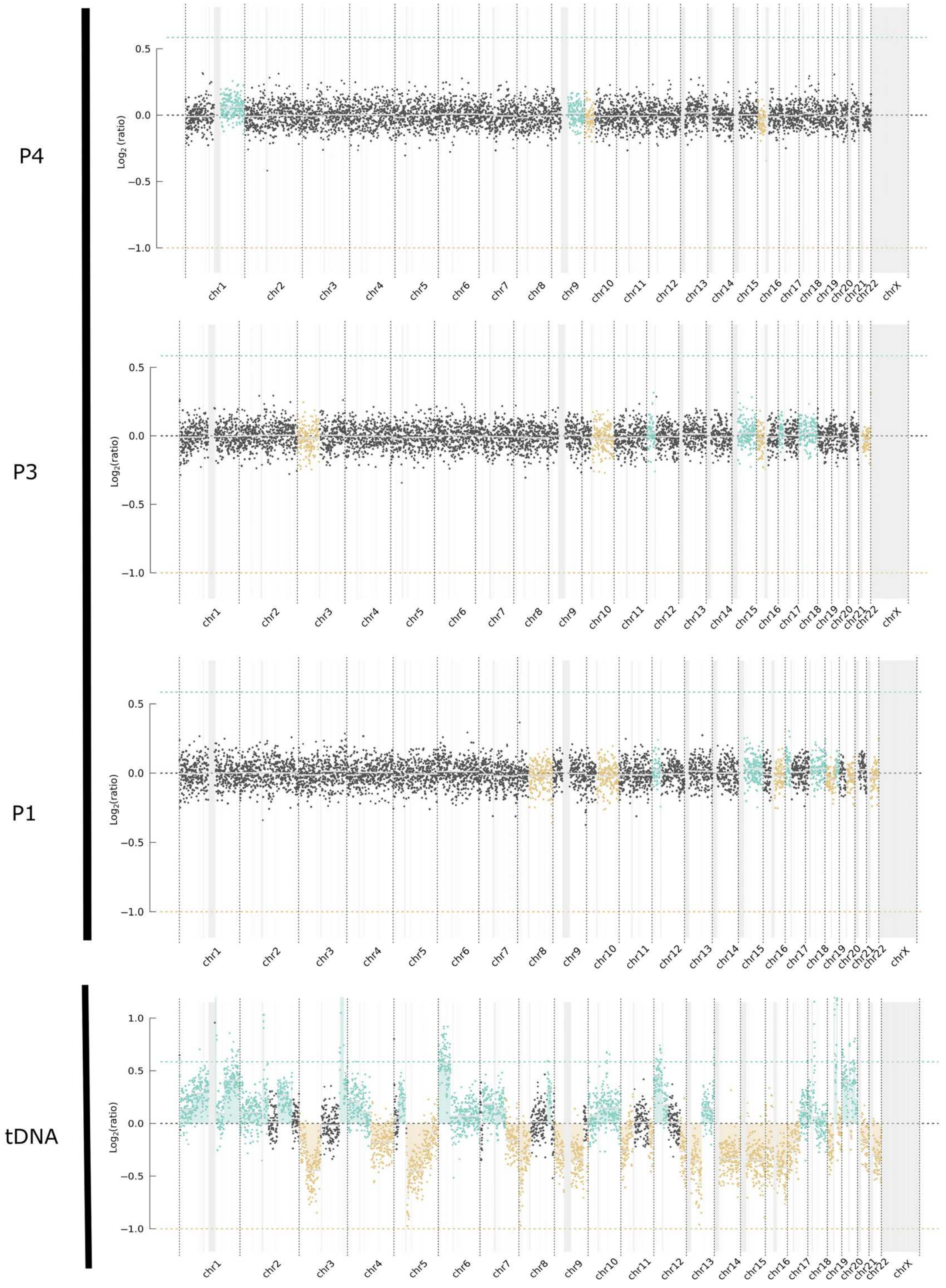

Fig. 2. Copy number profile comparison between tDNA and pDNA at different time points before diagnosis for a representative case. For patient ID 1240-11, the genome-wide somatic copy number alteration (SCNA) pattern was plotted for tDNA and matched Pap test smears (P1 = 4.3 years, P3 = 6.2 years, and P4 = 7.5 years before the diagnosis). The *x* axis and *y* axis represent the loci of 23 chromosomes and corresponding copy numbers in the $\log_2$ scale, respectively. Each black point represents a single 500-kbp bin, whereas horizontal white lines indicate copy number segments covering bins of expected equal copy number. Colored points signify regions of SCNA. Green dots refer to the gain of genomic material, whereas yellow refers to regions of losses of genomic material. Gray boxes indicate genomic regions not included in the analysis. The different $\log_2$ ratio between tDNA and pDNA is tumor fraction dependent.

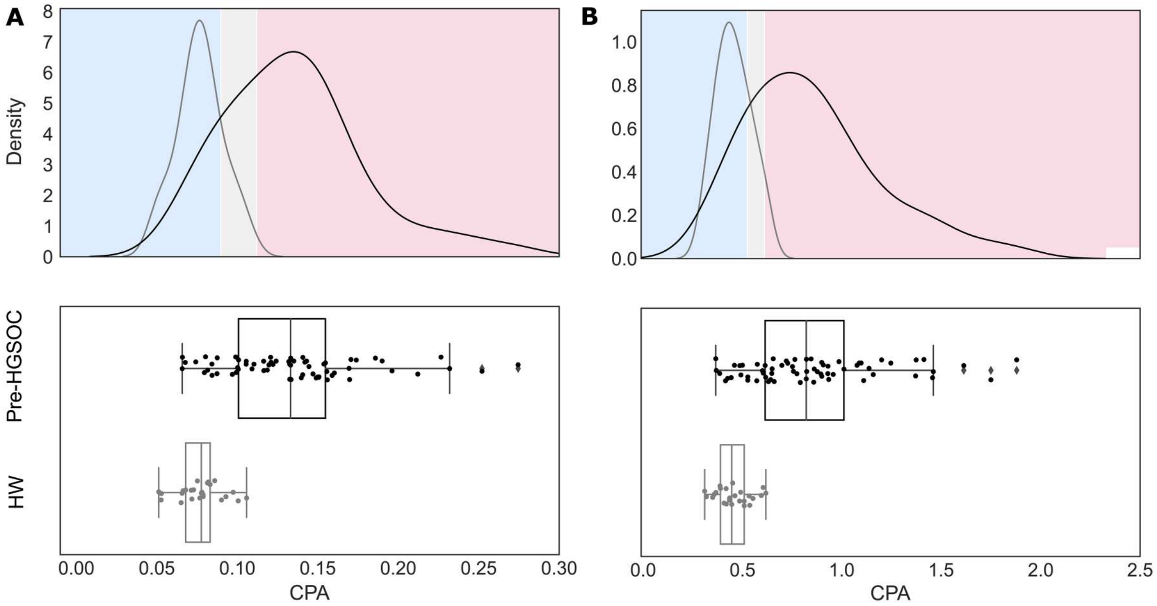

Fig. 3. Kernel density estimation (KDE) distribution and boxplots of the CPA values. Top: KDE curve distributions represent the CPA of the Pap test smear DNA (pDNA) from HW ($n = 25$, light line) and from presymptomatic women who had developed HGSOC (pre-HGSOC) ($n = 65$, marked line) . The *x* axis represents the CPA value, and the *y* axis represents the density. Colors in the background define the three different areas reported in the main text. Bottom: Boxplot reports the statistical distribution of the CPA values for pre-HGSOC cases in black and for HW in gray. (**A**) KDE distribution according to the chromosomal regions most frequently altered in HGSOC based on TCGA-OV: $0 \leq \text{CPA} < 0.0887$, blue zone or negative for genomic alterations. $0.0887 \leq \text{CPA} < 0.11207$, gray zone, as with uncertainty; $\text{CPA} \geq 0.11207$, red zone or positive for genomic alterations. (**B**) KDE distribution on the whole genome: $0 \leq \text{CPA} < 0.52372$, blue zone or negative for genomic alterations. $0.52372 \leq \text{CPA} < 0.61$ gray zone, as with uncertainty; $\text{CPA} \geq 0.61$, red zone or positive for genomic alterations.

which allowed clear detection of areas of aneuploidy. We defined this interval as "positive" for genomic alterations. Under these assumptions, the specificity was 100% as set by design, the sensitivity was 67.69% [95% confidence interval (CI), 56.38 to 78.99%], and the accuracy was 76.67%. To ensure the reproducibility of our model, we performed a second sequencing run on a subset of pDNA samples ($n = 10$) with comparable results (fig. S5).

Given the marked disease heterogeneity, we finally considered whether a priori selection of regions from TCGA-OV had introduced bias into the analysis that could hamper its future use as a biomarker for early disease detection. Thus, we repeated the entire set of analyses using a genome-wide rather than region-specific approach, excluding from the analysis only those genomic regions defined as constitutive SCNA, that is, those SCNA frequently occurring in the healthy population, "blacklist-B." The CPA distribution showed three different intervals: (i) $0 \leq \text{CPA} < 0.52372$, (ii) $0.52372 \leq \text{CPA} < 0.61$ (gray zone), and (iii) $\text{CPA} \geq 0.61$, in a similar fashion as our previous results with TCGA-OV (Fig. 3B and table S10). In this setting, the specificity, sensitivity, and accuracy were 96.00% (95% CI, 88.35 to 100.00), 75.38% (95% CI, 64.97 to 85.79), and 81.11%, respectively. Although CPA values increased approximately 10-fold as the calculation expanded to a larger genomic interval, the CPA distribution in HW and pre-HGSOC samples mirrored those described in the initial analysis (Fig. 3A). This finding demonstrates the feasibility of identifying regions of aneuploidy in pDNA despite highly variable SCNA profiles in the genome of HGSOC. This scenario may potentially allow the use of the EVA test for early disease detection.

Longitudinal analysis of CPA to track genomic instability evolution

Last, to systematically examine the evolution of genome complexity in the clinical history of each patient, the CPA values measured in pre-HGSOC pDNAs were plotted according to the time of archival Pap test collection. Figure 4 depicts for each woman with pre-HGSOC the evolution of CPA values (measured according to the genome-wide approach), across a wide range of times, ranging from very short (up to the time of diagnosis, $t = 0$) to 11 years before tumor onset. This shows a relatively broad period of chromosomal instability, indicating a variable timing of gain or loss of genomic material across different patients and from different Pap test samples taken from the same patient. Overall, we observed the following characterization of pre-HGSOC pDNA samples: 75.4% (49 of 65) showed the presence of an aneuploid genome (dark pink circles), 15.4% (10 of 65) showed the presence of a diploid genome (blue circles), and 9.2% (6 of 65) were "uncertain" (gray circles). The presence of an aneuploid genome can be detected by the CPA analysis up to 9 years before diagnosis (CPA values of 1.46 and 1.09 for samples 1240-2 and 1240-18, respectively; table S10). Focusing on the CPA values retrieved from samples collected between 2 and 5 years before diagnosis, we observed that 71% (15 of 21) were characterized by an aneuploid genome (dark pink circles), 14% (3 of 21) were classified as uncertain (gray circles), and 14% (3 of 21) had a diploid genome (blue circles). Comparable results were obtained using CPA values derived from a targeted selection from TCGA-OV (fig. S6). Consistently with the results obtained by the CPA analysis, we have obtained further longitudinal evidence that *TP53* pathogenic variants can be detected in endocervical swabs up

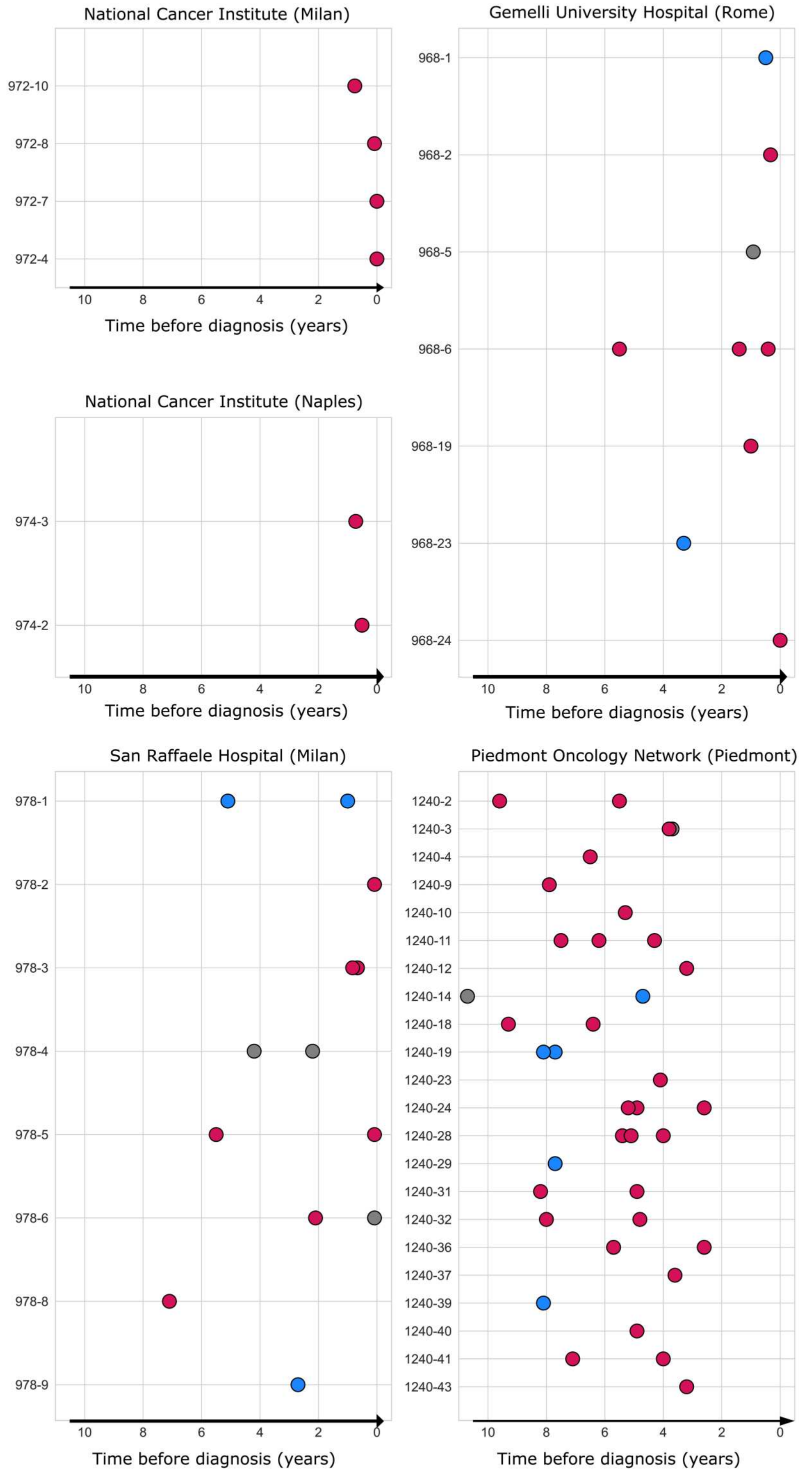

Fig. 4. CPA distribution in longitudinal Pap test smears calculated on the whole genome. pDNA samples are represented with circles and colored according to their CPA value: blue (negative), gray (uncertain), or dark pink (positive). Samples are ordered along the *x* axis representing the time (years) before diagnosis (time = 0 on the right).

to 9 years before the diagnosis (cohort B; fig. S7 and tables S11 and S12).

DISCUSSION

In the present study, we demonstrated that the early detection of HGSOC is potentially achievable through SCNA analysis of DNA extracted from archival Pap test smears of pre-HGSOC women. The analysis of cell swabs for the early detection of HGSOC has been investigated in the past because the procedure involved is noninvasive and well tolerated. Furthermore, cervical cell swabs offer the potential for population-wide screening. They can be easily incorporated into routine gynecological examinations, rendering them accessible to many women. This accessibility could lead to increased early detection rates and ultimately improve survival rates for ovarian cancer.

In the work described here, we focused on two early molecular events occurring during HGSOC development: *TP53* mutations and genomic instability. The results presented here are consistent with recent evidence from large-scale pan-cancer studies suggesting that in many cancer types, genomic alterations are present before the disease is detectable and may possibly contribute to disease development (*10*).

Prior analysis of the *TP53* mutations present in the tumor of each patient made it possible to verify that the same mutations found in the tumors were present in the pDNA taken from the same patients before HGSOC diagnosis. Earlier work in our laboratory showed that specific tumor *TP53* mutations were detectable in ~70% of cases up to 6 years before the diagnosis by applying the very sensitive droplet digital polymerase chain reaction (ddPCR) approach (*18*). We confirmed these findings in a larger multicentric collection of patients, including samples taken 9 years before diagnosis. Two main limitations hamper the suitability of analyzing *TP53* mutations for the early detection of HGSOC. First, this approach requires prior knowledge of the *TP53* mutations present in the tumor. Second, low-frequency somatic pathogenic mutations of *TP53* are unlikely to be a suitable diagnostic biomarker because they are often detectable in normal tissues under physiological conditions as abnormal hematopoiesis or physiological somatic mosaicism, and they increase with age (*15*, *21*, *24*). Thus, the determination of *TP53* mutations would not discriminate the presence of a tumor with sufficient specificity.

For these reasons, in the current study, we chose to focus on genomic instability instead of low-frequency somatic *TP53* mutations. Patterns of genomic instability have been previously suggested to be present in the early stages of HGSOC development (*8*, *10*). Taking advantage of the fact that, in most cases, HGSOC develops initially in the fallopian tubes, shedding of tumor cells in this locus is eminently plausible, which renders their detection in the cervical swabs feasible. The presence of detectable tumor material in the cervix is conceivable in the early stages of neoplasia when a few layers of cancer cells are present in the fallopian tube. The detection of tumor DNA within the DNA from endocervical swabs supports this hypothesis. An essential element for the robust quantification of genomic instability is the appropriate method of its measurement.

Consistent with the literature, our findings highlight that a simple model of cancer progression is difficult to establish for HGSOC. With the exception of pathogenic mutations in the *TP53* gene, the evolutionary path of individual tumors is highly variable, with loss of genomic material on chromosomes *17q*, *16q*, and *22q* or gain on chromosomes *3p26* and *8q24* as the most frequent events, although they are not common to all cases. For this reason, the EVA test has been developed exploiting the CPA score, which reflects the overall amount of SCNA present in the tumor genome rather than in single genomic regions. The CPA score offers a number of advantages. First, it has been used successfully in a setting with a low presence of tumor DNA (circulating tumor DNA) in patients with lung carcinoma (*23*). Second, it obviates the requirement of an inordinate amount of fine-tuning when tumor DNA is present at low concentrations; the measurement is based on *z*-scores that allow a precise detection of instability, assuming the presence of a suitable control reference. Third, the use of a single measurement bypasses the issue of low resolution when tumor DNA is scarce, causing problems in identifying the precise regions where SCNA occurs (*25*). Fourth, CPA analysis allows the global measurement of genomic instability without prior knowledge of the precise genomic alterations that characterize the tumor genome, thus making the "agnostic" analysis of tumor DNA feasible. Last, CPA is derived from a low-pass whole-genome approach, which is now a fast and cheap sequencing protocol that can be easily implemented in a clinical analytical laboratory.

The use of a numerical measure implies the selection of acceptable thresholds to maximize the number of true positives and minimize or eliminate, if possible, the number of false positives. To this aim, we defined two CPA thresholds that include a "gray zone" in which the presence of genomic alterations is considered uncertain. Furthermore, we adopted a conservative approach to ensure high confidence in calling a sample positive for genomic alterations at the cost of registering more false negatives. We marked all pre-HGSOC pDNA samples in the gray zone as false negatives, although at least some of them were probably true negatives. By applying this stringent approach, we achieved a sensitivity of 75% and a specificity of 96%. We surmise that this sensitivity is satisfactory, considering that the archival Pap test collection was not originally meant to be used for DNA analysis, which implies that the sampling procedure and storage conditions were probably suboptimal for the aims of this study. A fraction of false negatives might be inevitable considering that, according to some recent pathological reports, probably not all HGSOCs derive from the fallopian tube (*26*). We proffer the contention that the EVA test may well be suitable for the early detection of HGSOC. It might possibly also be applied to the early detection of the fraction of endometrial cancers characterized by genomic instability (*27*). Exploration of this possibility requires specific studies.

We are aware that our study has the following limitations: (i) It is retrospective with a limited sample size; (ii) samples were not collected at standardized time points, limiting the possibility to evaluate the EVA test performance over time; and (iii) there is a difference in median age between pre-HGSOC and HW. It is known that SCNA accumulate with age (*28*, *29*). In the work presented here, we have minimized the impact of these genomic variations by removing the constitutive copy number variations from our assay. Nevertheless, we consider the data sufficiently convincing to warrant prospective clinical investigations aimed at verifying whether the longitudinal analysis of CPA in Pap test smears renders the prediction of HGSOC possible.

The EVA test based on sWGS is cheap and easily integratable in already ongoing mass screening programs for the early detection of cervical cancer. The incidence of ovarian cancer in the general population is very low, around 20 cases per 100,000 women. To demonstrate that the test benefits the survival of patients with HGSOC, a prospective study has to involve necessarily a very large population of women and a long follow-up time with longitudinal sampling. The specificity of the EVA test at 96% means that there might be 4000 false-positive cases in every 100,000 women tested. However, the relative simplicity and cost-effectiveness of the procedure render serial analyses from the same patient, such as every 6 months or every year, highly feasible. This would reduce the number of false positives, probably 160 cases after the second sampling and 6.4 after the third. It might be propitious to apply the test initially to women with germline mutations in *BRCA* genes who have a high probability of presenting with HGSOC. Now, a large fraction of these women opt to undergo prophylactic surgery consisting of salpingectomy with or without ovariectomy (*30*). This patient group might provide an opportunity to verify whether the analysis of CPA in Pap test swabs conducted before the operation can predict the presence of a STIC or in situ HGSOC in the fallopian tube.

In conclusion, our study provides the basis for a new approach to the early detection of HGSOC based on the assessment of genomic instability patterns of DNA extracted from cervical smears. Because the low survival of patients with HGSOC is usually related to the delay in diagnosis, we believe that the application of the approach proposed here may have a marked impact on mortality from this neoplasm.

MATERIALS AND METHODS

Study design

A multicentric, retrospective cohort of 113 patients with a diagnosis of HGSOC, who underwent primary surgical treatment from 2008 to 2021, was selected from eight independent Italian hospitals (fig. S1 and tables S2 and S3). Being a proof-of-principle retrospective study, no proper sample size was calculated, but study sample was chosen on convenience basis taking all samples available. All patients enrolled had at least one Pap test smear performed during cervical cancer screening before the diagnosis (range, 0 months to 13.6 years) (table S1). The Pap test samples were collected between 2003 and 2021 and were cytologically negative for dysplasia or any other malignant neoplasms. For 112 of 113 enrolled patients, a formalin-fixed paraffin-embedded (FFPE) primary tumor biopsy was also available. To fulfill the minimum sample size (n = 40) requested by WisecondorX for the construction of the panel of normals (PoN) (see "Bioinformatics analysis" section, Materials and Methods, and the Supplementary Materials), 77 Pap test smears collected between 2015 and 2022 during routine gynecological screening from women with no evidence of gynecological tumors or any other neoplastic or gynecological inflammatory disease were used as a control set.

Samples used for this study were collected in accordance with the institutional review boards and with all current national and European laws and regulations (ordinance of the Ministry of Health 17 December 2004, Gazzetta Ufficiale n. 43, 22 February 2005), including the Good Clinical Practice (GCP) Rules (legislative decree 24 June 2003 n. 211). The study was performed following the principles of the Declaration of Helsinki, and written informed consent was obtained from each participant.

DNA extraction library preparation and sequencing

tDNA and pDNA were extracted and purified from both FFPE tumor samples and Pap test smears using Maxwell RCS DNA FFPE kit (Promega). DNA concentrations were determined by Qubit high-sensitivity DNA assay (Thermo Fisher Scientific), and DNA quality was assessed by Tape Station 4200 System (Agilent Technologies) (table S13). Ten nanograms up to 150 ng of purified tDNA or pDNA was used to construct WGS libraries according to standard protocols (KAPA Hyper Plus kit, Roche) and barcoded on NextSeq550 (Illumina) to achieve at least 10 million reads per sample (sWGS, median coverage 0.5×); targeted next-generation sequencing libraries to analyze *TP53* mutational landscape were prepared and sequenced at 2500× coverage on NextSeq550 according to the manufacturer's protocol (SeqCap EZ HyperCap Workflow, Roche) as previously reported (*8*, *13*). Details are reported in the Supplementary Materials.

Bioinformatics analysis

The entire workflow used for sWGS analysis is summarized in fig. S2 and detailed in the Supplementary Materials. Specifically, raw reads were preprocessed by trimming with fastp software (*31*) and then aligned to the reference genome (hg38 build) with BWA (*32*), and duplicate reads were removed. WisecondorX (*33*) was used to call aberrant regions in pDNA and tDNA samples compared with a reference PoN. Raw aligned reads were divided into 500-kbp (kilo–base pair) bins, quality-filtered, and segmented with circular binary segmentation (*34*). Then, z-scores were calculated for each segment, and segments were called aberrant if their absolute value was greater than 5 (tDNA) or 3 (pDNA). The value of the CPA for each sample was calculated as described by Raman *et al.* (*33*).

The procedures used to select the PoN (fig. S8), the calculation of overlap between pDNA and tDNA, and the stability analysis of CPA are detailed in the Supplementary Materials. In particular, for the definition of the CPA threshold, a set of samples from HW (n = 77) was divided into two groups consisting of 52 (68%) and 25 (32%) samples, respectively. The first was used as PoN, and the latter was used as validation set. The analysis was done either including or excluding predefined genomic regions as detailed in the Supplementary Materials. The software was run with customized pipelines built with Nextflow (*35*) and with the tools provided by the nf-core framework (*36*). Code is provided at 10.5281/zenodo.10013056.

Statistical analysis

Comparisons were done with the Mann-Whitney test as implemented in the SciPy Python package. To calculate specificity, sensitivity, and accuracy, samples (n = 25 from HW, and n = 65 from pre-HGSOC) were considered positive when above the CPA threshold; otherwise, they were considered negative. Sensitivity, specificity, and accuracy were calculated as follows:

$$\text{Sensitivity} = \frac{\text{True positive}}{\text{True positive} + \text{False negative}}$$

$$\text{Specificity} = \frac{\text{True negative}}{\text{True negative} + \text{False positive}}$$

$$\text{Accuracy} = \frac{\text{True positive} + \text{True negative}}{\text{Total number of cases}}$$

Receiver operating characteristic (ROC) curves were calculated with the roc_curve method implemented in the scikit-learn Python package. Confidence intervals (CIs) were calculated following binomial distribution.

REFERENCES AND NOTES

1. J. Ferlay, I. Soerjomataram, R. Dikshit, S. Eser, C. Mathers, M. Rebelo, D. M. Parkin, D. Forman, F. Bray, Cancer incidence and mortality worldwide: Sources, methods and major patterns in GLOBOCAN 2012. *Int. J. Cancer* **136**, E359–E386 (2015).
2. S. Lheureux, C. Gourley, I. Vergote, A. M. Oza, Epithelial ovarian cancer. *Lancet* **393**, 1240–1253 (2019).
3. D. D. Bowtell, S. Böhm, A. A. Ahmed, P.-J. Aspuria, R. C. Bast, V. Beral, J. S. Berek, M. J. Birrer, S. Blagden, M. A. Bookman, J. D. Brenton, K. B. Chiappinelli, F. C. Martins, G. Coukos, R. Drapkin, R. Edmondson, C. Fotopoulou, H. Gabra, J. Galon, C. Gourley, V. Heong, D. G. Huntsman, M. Iwanicki, B. Y. Karlan, A. Kaye, E. Lengyel, D. A. Levine, K. H. Lu, I. A. McNeish, U. Menon, S. A. Narod, B. H. Nelson, K. P. Nephew, P. Pharoah, D. J. Powell, P. Ramos, I. L. Romero, C. L. Scott, A. K. Sood, E. A. Stronach, F. R. Balkwill, Rethinking ovarian cancer II: Reducing mortality from high-grade serous ovarian cancer. *Nat. Rev. Cancer* **15**, 668–679 (2015).
4. U. Menon, A. Gentry-Maharaj, M. Burnell, N. Singh, A. Ryan, C. Karpinskyj, G. Carlino, J. Taylor, S. K. Massingham, M. Raikou, J. K. Kalsi, R. Woolas, R. Manchanda, R. Arora, L. Casey, A. Dawnay, S. Dobbs, S. Leeson, T. Mould, M. W. Seif, A. Sharma, K. Williamson, Y. Liu, L. Fallowfield, A. J. McGuire, S. Campbell, S. J. Skates, I. J. Jacobs, M. Parmar, Ovarian cancer population screening and mortality after long-term follow-up in the UK Collaborative Trial of Ovarian Cancer Screening (UKCTOCS): A randomised controlled trial. *Lancet* **397**, 2182–2193 (2021).
5. S. S. Buys, E. Partridge, M. H. Greene, P. C. Prorok, D. Reding, T. L. Riley, P. Hartge, R. M. Fagerstrom, L. R. Ragard, D. Chia, G. Izmirlian, M. Fouad, C. C. Johnson, J. K. Gohagan, PLCO Project Team, Ovarian cancer screening in the Prostate, Lung, Colorectal and Ovarian (PLCO) cancer screening trial: Findings from the initial screen of a randomized trial. *Am. J. Obstet. Gynecol.* **193**, 1630–1639 (2005).
6. S. I. Labidi-Galy, E. Papp, D. Hallberg, N. Niknafs, V. Adleff, M. Noe, R. Bhattacharya, M. Novak, S. Jones, J. Phallen, C. A. Hruban, M. S. Hirsch, D. I. Lin, L. Schwartz, C. L. Maire, J.-C. Tille, M. Bowden, A. Ayhan, L. D. Wood, R. B. Scharpf, R. Kurman, T.-L. Wang, I.-M. Shih, R. Karchin, R. Drapkin, V. E. Velculescu, High grade serous ovarian carcinomas originate in the fallopian tube. *Nat. Commun.* **8**, 1093 (2017).
7. T. R. Soong, B. E. Howitt, A. Miron, N. S. Horowitz, F. Campbell, C. M. Feltmate, M. G. Muto, R. S. Berkowitz, M. R. Nucci, W. Xian, C. P. Crum, Evidence for lineage continuity between early serous proliferations (ESPs) in the fallopian tube and disseminated high-grade serous carcinomas. *J. Pathol.* **246**, 344–351 (2018).
8. C. Pesenti, L. Beltrame, A. Velle, R. Fruscio, M. Jaconi, F. Borella, F. M. Cribiù, E. Calura, L. V. Venturini, D. Lenoci, F. Agostinis, D. Katsaros, N. Panini, T. Bianchi, F. Landoni, M. Miozzo, M. D'Incalci, J. D. Brenton, C. Romualdi, S. Marchini, Copy number alterations in stage I epithelial ovarian cancer highlight three genomic patterns associated with prognosis. *Eur. J. Cancer* **171**, 85–95 (2022).
9. M. A. Eckert, S. Pan, K. M. Hernandez, R. M. Loth, J. Andrade, S. L. Volchenboum, P. Faber, A. Montag, R. Lastra, M. E. Peter, S. D. Yamada, E. Lengyel, Genomics of ovarian cancer progression reveals diverse metastatic trajectories including intraepithelial metastasis to the fallopian tube. *Cancer Discov.* **6**, 1342–1351 (2016).
10. M. Gerstung, C. Jolly, I. Leshchiner, S. C. Dentro, S. Gonzalez, D. Rosebrock, T. J. Mitchell, Y. Rubanova, P. Anur, K. Yu, M. Tarabichi, A. Deshwar, J. Wintersinger, K. Kleinheinz, I. Vázquez-García, K. Haase, L. Jerman, S. Sengupta, G. Macintyre, S. Malikic, N. Donmez, D. G. Livitz, M. Cmero, J. Demeulemeester, S. Schumacher, Y. Fan, X. Yao, J. Lee, M. Schlesner, P. C. Boutros, D. D. Bowtell, H. Zhu, G. Getz, M. Imielinski, R. Beroukhim, S. C. Sahinalp, Y. Ji, M. Peifer, F. Markowetz, V. Mustonen, K. Yuan, W. Wang, Q. D. Morris, PCAWG Evolution & Heterogeneity Working Group, P. T. Spellman, D. C. Wedge, P. Van Loo, PCAWG Consortium, The evolutionary history of 2,658 cancers. 578, *Nature*, 122–Nat128 (2020).
11. E. Biskup, R. S. Wils, C. Hogdall, E. Hogdall, Prospects of improving early ovarian cancer diagnosis using cervical cell swabs. *Anticancer Res.* **42**, 1–12 (2022).
12. I. Kinde, C. Bettegowda, Y. Wang, J. Wu, N. Agrawal, I.-M. Shih, R. Kurman, F. Dao, D. A. Levine, R. Giuntoli, R. Roden, J. R. Eshleman, J. P. Carvalho, S. K. N. Marie, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, L. A. Diaz, Evaluation of DNA from the Papanicolaou test to detect ovarian and endometrial cancers. *Sci. Transl. Med.* **5**, 167ra4 (2013).
13. Y. Wang, L. Li, C. Douville, J. D. Cohen, T.-T. Yen, I. Kinde, K. Sundfelt, S. K. Kjær, R. H. Hruban, I.-M. Shih, T.-L. Wang, R. J. Kurman, S. Springer, J. Ptak, M. Popoli, J. Schaefer, N. Silliman, L. Dobbyn, E. J. Tanner, A. Angarita, M. Lycke, K. Jochumsen, B. Afsari, L. Danilova, D. A. Levine, K. Jardon, X. Zeng, J. Arseneau, L. Fu, L. A. Diaz, R. Karchin, C. Tomasetti, K. W. Kinzler, B. Vogelstein, A. N. Fader, L. Gilbert, N. Papadopoulos, Evaluation of liquid from the Papanicolaou test and other liquid biopsies for the detection of endometrial and ovarian cancers. *Sci. Transl. Med.* **10**, eaap8793 (2018).
14. X. Jiang, W. Li, J. Yang, S. Wang, D. Cao, M. Yu, K. Shen, J. Bai, Y. Gao, Identification of somatic mutations in papanicolaou smear dna and plasma circulating cell-free DNA for detection of endometrial and epithelial ovarian cancers: A pilot study. *Front. Oncol.* **10**, 582546 (2020).
15. J. D. Krimmel-Morrison, T. S. Ghezelayagh, S. Lian, Y. Zhang, J. Fredrickson, D. Nachmanson, K. T. Baker, M. R. Radke, E. Hun, B. M. Norquist, M. J. Emond, E. M. Swisher, R. A. Risques, Characterization of TP53 mutations in Pap test DNA of women with and without serous ovarian carcinoma. *Gynecol. Oncol.* **156**, 407–414 (2020).
16. N. S. Arildsen, L. Martin de la Fuente, A. Måsbäck, S. Malander, O. Forslund, P. Kannisto, I. Hedenfalk, Detecting TP53 mutations in diagnostic and archival liquid-based Pap samples from ovarian cancer patients using an ultra-sensitive ddPCR method. *Sci. Rep.* **9**, 15506 (2019).
17. L. Paracchini, L. Mannarino, I. Craparotta, C. Romualdi, R. Fruscio, T. Grassi, V. Fotia, G. Caratti, P. Perego, E. Calura, L. Clivio, M. D'Incalci, L. Beltrame, S. Marchini, Regional and temporal heterogeneity of epithelial ovarian cancer tumor biopsies: Implications for therapeutic strategies. *Oncotarget* **5**, 2404–2417 (2016).
18. L. Paracchini, C. Pesenti, M. Delle Marchette, L. Beltrame, T. Bianchi, T. Grassi, A. Buda, F. Landoni, L. Ceppi, C. Bosetti, M. Paderno, M. Adorni, D. Vicini, P. Perego, B. E. Leone, M. D'Incalci, S. Marchini, R. Fruscio, Detection of *TP53* clonal variants in papanicolaou test samples collected up to 6 years prior to high-grade serous epithelial ovarian cancer diagnosis. *JAMA Netw. Open* **3**, e207566 (2020).
19. E. Maritschnegg, Y. Wang, N. Pecha, R. Horvat, E. Van Nieuwenhuysen, I. Vergote, F. Heitz, J. Sehouli, I. Kinde, L. A. Diaz, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, P. Speiser, R. Zeillinger, Lavage of the uterine cavity for molecular detection of Müllerian duct carcinomas: A proof-of-concept study. *J. Clin. Oncol.* **33**, 4293–4300 (2015).
20. N.-Y. Y. Wu, C. Fang, H.-S. Huang, J. Wang, T.-Y. Chu, Natural history of ovarian high-grade serous carcinoma from time effects of ovulation inhibition and progesterone clearance of p53-defective lesions. *Mod. Pathol.* **33**, 29–37 (2020).
21. J. J. Salk, K. Loubet-Senear, E. Maritschnegg, C. C. Valentine, L. N. Williams, J. E. Higgins, R. Horvat, A. Vanderstichele, D. Nachmanson, K. T. Baker, M. J. Emond, E. Loter, M. Tretiakova, T. Soussi, L. A. Loeb, R. Zeillinger, P. Speiser, R. A. Risques, Ultra-sensitive TP53 sequencing for cancer detection reveals progressive clonal selection in normal tissue over a century of human lifespan. *Cell Rep.* **28**, 132–144.e3 (2019).
22. S. Killcoyne, A. Yusuf, R. C. Fitzgerald, Genomic instability signals offer diagnostic possibility in early cancer detection. *Trends Genet.* **37**, 966–972 (2021).
23. L. Raman, M. Van der Linden, K. Van der Eecken, K. Vermaelen, I. Demedts, V. Surmont, U. Himpe, F. Dedeurwaerdere, L. Ferdinande, Y. Lievens, K. Claes, B. Menten, J. Van Dorpe,

Shallow whole-genome sequencing of plasma cell-free DNA accurately differentiates small from non-small cell lung carcinoma. *Genome Med.* **12**, 35 (2020).

24. J. D. Krimmel, M. W. Schmitt, M. I. Harrell, K. J. Agnew, S. R. Kennedy, M. J. Emond, L. A. Loeb, E. M. Swisher, R. A. Risques, Ultra-deep sequencing detects ovarian cancer cells in peritoneal fluid and reveals somatic TP53 mutations in noncancerous tissues. *Proc. Natl. Acad. Sci. U.S.A.* **113**, 6005–6010 (2016).
25. V. A. Adalsteinsson, G. Ha, S. S. Freeman, A. D. Choudhury, D. G. Stover, H. A. Parsons, G. Gydush, S. C. Reed, D. Rotem, J. Rhoades, D. Loginov, D. Livitz, D. Rosebrock, I. Leshchiner, J. Kim, C. Stewart, M. Rosenberg, J. M. Francis, C.-Z. Zhang, O. Cohen, C. Oh, H. Ding, P. Polak, M. Lloyd, S. Mahmud, K. Helvie, M. S. Merrill, R. A. Santiago, E. P. O'Connor, S. H. Jeong, R. Leeson, R. M. Barry, J. F. Kramkowski, Z. Zhang, L. Polacek, J. G. Lohr, M. Schleicher, E. Lipscomb, A. Saltzman, N. M. Oliver, L. Marini, A. G. Waks, L. C. Harshman, S. M. Tolaney, E. M. Van Allen, E. P. Winer, N. U. Lin, M. Nakabayashi, M.-E. Taplin, C. M. Johannessen, L. A. Garraway, T. R. Golub, J. S. Boehm, N. Wagle, G. Getz, J. C. Love, M. Meyerson, Scalable whole-exome sequencing of cell-free DNA reveals high concordance with metastatic tumors. *Nat. Commun.* **8**, 1324 (2017).
26. S. Zhang, I. Dolgalev, T. Zhang, H. Ran, D. A. Levine, B. G. Neel, Both fallopian tube and ovarian surface epithelium are cells-of-origin for high-grade serous ovarian carcinoma. *Nat. Commun.* **10**, 5367 (2019).
27. Cancer Genome Atlas Research Network, C. Kandoth, N. Schultz, A. D. Cherniack, R. Akbani, Y. Liu, H. Shen, A. G. Robertson, I. Pashtan, R. Shen, C. C. Benz, C. Yau, P. W. Laird, L. Ding, W. Zhang, G. B. Mills, R. Kucherlapati, E. R. Mardis, D. A. Levine, Integrated genomic characterization of endometrial carcinoma. *Nature* **497**, 67–73 (2013).
28. A. J. Iafrate, L. Feuk, M. N. Rivera, M. L. Listewnik, P. K. Donahoe, Y. Qi, S. W. Scherer, C. Lee, Detection of large-scale variation in the human genome. *Nat. Genet.* **36**, 949–951 (2004).
29. J. R. MacDonald, R. Ziman, R. K. C. Yuen, L. Feuk, S. W. Scherer, The Database of Genomic Variants: A curated collection of structural variation in the human genome. *Nucleic Acids Res.* **42**, D986–D992 (2014).
30. I. A. S. Stroot, J. Brouwer, J. Bart, H. Hollema, D. J. Stommel-Jenner, M. M. Wagner, H. C. van Doorn, J. A. de Hullu, K. N. Gaarenstroom, M. Beurden, L. R. C. W. van Lonkhuijzen, B. F. M. Slangen, R. P. Zweemer, E. B. G. Garcia, M. G. E. M. Ausems, I. A. Boere, K. van Engelen, C. J. van Asperen, M. K. Schmidt, M. R. Wevers, G. H. de Bock, M. J. E. Mourits; HEBON Investigators, High-grade serous carcinoma at risk-reducing salpingo-oophorectomy in asymptomatic carriers of *BRCA1/2* pathogenic variants: Prevalence and clinical factors. *J. Clin. Oncol.* **41**, 2523–2535 (2023).
31. S. Chen, Y. Zhou, Y. Chen, J. Gu, fastp: An ultra-fast all-in-one FASTQ preprocessor. *Bioinformatics* **34**, i884–i890 (2018).
32. H. Li, R. Durbin, Fast and accurate long-read alignment with Burrows-Wheeler transform. *Bioinformatics* **26**, 589–595 (2010).
33. L. Raman, A. Dheedene, M. De Smet, J. Van Dorpe, B. Menten, WisecondorX: Improved copy number detection for routine shallow whole-genome sequencing. *Nucleic Acids Res.* **47**, 1605–1614 (2019).
34. A. B. Olshen, E. S. Venkatraman, R. Lucito, M. Wigler, Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics* **5**, 557–572 (2004).
35. P. Di Tommaso, M. Chatzou, E. W. Floden, P. P. Barja, E. Palumbo, C. Notredame, Nextflow enables reproducible computational workflows. *Nat. Biotechnol.* **35**, 316–319 (2017).
36. P. A. Ewels, A. Peltzer, S. Fillinger, H. Patel, J. Alneberg, A. Wilm, M. U. Garcia, P. Di Tommaso, S. Nahnsen, The nf-core framework for community-curated bioinformatics pipelines. *Nat. Biotechnol.* **38**, 276–278 (2020).
37. A. Tarasov, A. J. Vilella, E. Cuppen, I. J. Nijman, P. Prins, Sambamba: Fast processing of NGS alignment formats. *Bioinformatics* **31**, 2032–2034 (2015).
38. A. McKenna, M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis, A. Kernytsky, K. Garimella, D. Altshuler, S. Gabriel, M. Daly, M. A. DePristo, The genome analysis toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Res.* **20**, 1297–1303 (2010).
39. K. Cibulskis, M. S. Lawrence, S. L. Carter, A. Sivachenko, D. Jaffe, C. Sougnez, S. Gabriel, M. Meyerson, E. S. Lander, G. Getz, Sensitive detection of somatic point mutations in impure and heterogeneous cancer samples. *Nat. Biotechnol.* **31**, 213–219 (2013).
40. Z. Lai, A. Markovets, M. Ahdesmaki, B. Chapman, O. Hofmann, R. McEwen, J. Johnson, B. Dougherty, J. C. Barrett, J. R. Dry, VarDict: A novel and versatile variant caller for next-generation sequencing in cancer research. *Nucleic Acids Res.* **44**, e108 (2016).
41. H. Li, A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics* **27**, 2987–2993 (2011).
42. W. McLaren, L. Gil, S. E. Hunt, H. S. Riat, G. R. S. Ritchie, A. Thormann, P. Flicek, F. Cunningham, The ensembl variant effect predictor. *Genome Biol.* **17**, 122 (2016).
43. K. Eilbeck, S. E. Lewis, Sequence ontology annotation guide. *Comp. Funct. Genomics* **5**, 642–647 (2004).
44. K. J. Karczewski, L. C. Francioli, G. Tiao, B. B. Cummings, J. Alföldi, Q. Wang, R. L. Collins, K. M. Laricchia, A. Ganna, D. P. Birnbaum, L. D. Gauthier, H. Brand, M. Solomonson, N. A. Watts, D. Rhodes, M. Singer-Berk, E. M. England, E. G. Seaby, J. A. Kosmicki, R. K. Walters, K. Tashman, Y. Farjoun, E. Banks, T. Poterba, A. Wang, C. Seed, N. Whiffin, J. X. Chong, K. E. Samocha, E. Pierce-Hoffman, Z. Zappala, A. H. O'Donnell-Luria, E. V. Minikel, B. Weisburd, M. Lek, J. S. Ware, C. Vittal, I. M. Armean, L. Bergelson, K. Cibulskis, K. M. Connolly, M. Covarrubias, S. Donnelly, S. Ferriera, S. Gabriel, J. Gentry, N. Gupta, T. Jeandet, D. Kaplan, C. Llanwarne, R. Munshi, S. Novod, N. Petrillo, D. Roazen, V. Ruano-Rubio, A. Saltzman, M. Schleicher, J. Soto, K. Tibbetts, C. Tolonen, G. Wade, M. E. Talkowski; Genome Aggregation Database Consortium, B. M. Neale, M. J. Daly, D. G. MacArthur, The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
45. 1000 Genomes Project Consortium, A. Auton, L. D. Brooks, R. M. Durbin, E. P. Garrison, H. M. Kang, J. O. Korbel, J. L. Marchini, S. McCarthy, G. A. McVean, G. R. Abecasis, A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
46. M. J. Landrum, J. M. Lee, M. Benson, G. R. Brown, C. Chao, S. Chitipiralla, B. Gu, J. Hart, D. Hoffman, W. Jang, K. Karapetyan, K. Katz, C. Liu, Z. Maddipatla, A. Malheiro, K. McDaniel, M. Ovetsky, G. Riley, G. Zhou, J. B. Holmes, B. L. Kattman, D. R. Maglott, ClinVar: Improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
47. D. Tamborero, C. Rubio-Perez, J. Deu-Pons, M. P. Schroeder, A. Vivancos, A. Rovira, I. Tusquets, J. Albanell, J. Rodon, J. Tabernero, C. de Torres, R. Dienstmann, A. Gonzalez-Perez, N. Lopez-Bigas, Cancer genome interpreter annotates the biological and clinical relevance of tumor alterations. *Genome Med.* **10**, 25 (2018).
48. H. Li, Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv: 1303.3997 [quant-ph] (16 March 2013).
49. C. H. Mermel, S. E. Schumacher, B. Hill, M. L. Meyerson, R. Beroukhim, G. Getz, GISTIC2.0 facilitates sensitive and confident localization of the targets of focal somatic copy-number alteration in human cancers. *Genome Biol.* **12**, R41 (2011).
50. A. R. Quinlan, I. M. Hall, BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics* **26**, 841–842 (2010).
51. H. Zhao, Z. Sun, J. Wang, H. Huang, J.-P. Kocher, L. Wang, CrossMap: A versatile tool for coordinate conversion between genome assemblies. *Bioinformatics* **30**, 1006–1007 (2014).
52. R. Straver, E. A. Sistermans, H. Holstege, A. Visser, C. B. M. Oudejans, M. J. T. Reinders, WISECONDOR: Detection of fetal aberrations from shallow sequencing maternal plasma based on a within-sample comparison scheme. *Nucleic Acids Res.* **42**, e31 (2014).
53. R. K. Dale, B. S. Pedersen, A. R. Quinlan, Pybedtools: A flexible Python library for manipulating genomic datasets and annotations. *Bioinformatics* **27**, 3423–3424 (2011).
54. J. D. Hunter, Matplotlib: A 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).
55. M. L. Waskom, Seaborn: Statistical data visualization. *J. Open Source Softw.* **6**, 3021 (2021).

Acknowledgments: We are grateful to A. Gescher (University of Leicester, UK) for critical revision and editing of the manuscript. **Funding:** This was supported by the Alessandra Bono Foundation and Alliance Against Cancer (ACC) grant RCR-2021-23671213 to M.D.I., with additional support from the Fondazione AIRC per la Ricerca sul Cancro grant IG-19997 to S.M. and IG 21837 to C.R. **Author contributions:** This study was conceptualized and designed by L. P., L.M., L. Beltrame, D.L., M.D.I., C.R., and S.M. Data and specimen collection was completed by P. Z., M.E.L., E.P., L.G., R.F., F.L., S.F., V.C., L. Bocciolone, D.L., F.R., I.M.G.G., and the TOWARDS group. This collaborative network was built by M.D.I. and S.M., maintained by G. Ciccone, and managed by I.F.N. DNA extraction, sequencing, and data validation were done by L.P. and M.L.D. Computational analysis of sequencing data, as well as visual presentation, was designed and performed by L.M., R.Z., and L. Beltrame. Statistical planning was a task of C.B. and C.R. The study was supervised by M.D.I., C.R., and S.M. The manuscript was written by L.P., L.M., R.Z., L. Beltrame, P.Z., C.B., M.D.I., C.R., and S.M. Editing was completed by all authors. **Competing interests:** The authors declare that they have no competing interests. **Data and materials availability:** All data associated with this study are present in the paper or the Supplementary Materials. sWGS data are available in the European Genome-Phenome Archive under the study EGAS00001007084 and at 10.5281/zenodo.10013056.

TOWARDS group collaborators

In addition to the TOWARDS group members who are authors (Maurizio D'Incalci[1,2], Sergio Marchini[2], Paolo Zola[4], Maria E. Laudani[4], Eva Pagano[5], Livia Giordano[5], Robert Fruscio[6], Fabio Landoni[6], Silvia Franceschi[7], Vincenzo Canzonieri[7,8], Luca Bocciolone[9], Domenica Lorusso[10], Francesco Raspagliesi[12], Isabella M. G. Garassino[13]), the following investigators are collaborators who have contributed to sample and data collection:

Tommaso Grassi[6], Tommaso Bianchi[6], Giulia Cursano[2], Giorgia Mangili[9], Giovanni Scambia[10], Claudia Marchetti[10], Renzo Boldorini[14], Giovanni De Rosa[15], Annamaria Ferrero[16], Elda Feyles[17], Margherita Goia[18], Claudia Manini[19], Renzo Orlassino[20], Daniela Surico[21], Marco Volante[22], Stefano Greggi[23], Marta Jaconi[24], Camillo Di Bella[24], Domenico Vitobello[25], Carla di Loreto[26,27], Stefano Pizzolitto[27], Fabrizio Zanconati[28], Giovannino Ciccone[29], Paola Armaroli[29], Cristina

Larato[29], Raffaella Rizzolo[29]

[14]Division of Pathology, University of Eastern Piedmont, Novara 28100, Italy. [15]Pathology Unit, Ordine Mauriziano Hospital, Turin 10128, Italy. [16]Academic Department of Obstetrics and Gynecology, Mauriziano Hospital, Turin 10128, Italy. [17]Pathology Unit, Cardinal Massaia Hospital, Asti 14100, Italy. [18]Pathology Unit, Città della Salute e della Scienza Hospital, Turin 10126, Italy. [19]Pathology Unit, San Giovanni Bosco Hospital, Turin 10154, Italy. [20]S.C. Pathology Unit, Ivrea General Hospital, Ivrea 10015, Turin, Italy. [21]Department of Pathology, Maggiore Hospital, University School of Medicine Amedeo Avogadro, Novara 28100, Italy. [22]Department of Oncology, University of Turin at San Luigi Hospital, Orbassano 10043, Turin, Italy. [23]Gynecologic Oncology Unit, Istituto Nazionale Tumori, IRCCS - Fondazione G.Pascale, Naples 80131, Italy. [24]Department of Pathology, Università degli Studi Milano-Bicocca, Fondazione IRCCS San Gerardo dei Tintori, Monza 20900, Italy. [25]Unit of Obstetrics and Gynecology, IRCCS Humanitas Research Hospital, Rozzano 20089, Milan, Italy. [26]Department of Medicine (DAME), University of Udine, Udine 33100, Italy. [27]Unit of Pathology, Santa Maria della Misericordia Academic Medical Centre, Udine 33100, Italy. [28]Pathology Unit, Ospedale di Cattinara, Azienda Sanitaria Universitaria Giuliano Isontina, Trieste 34149, Italy. [29]Unit of Clinical Epidemiology, AOU Città della Salute e della Scienza di Torino and CPO Piemonte, Turin 10126, Italy.