Copy number signatures in cervical samples enable early detection of high-grade serous ovarian carcinoma

Laura Martin de la Fuente[1†], Srinivas Veerla[1†], Minerva X. Li[1], Guyuan Tang[1‡], Anna Ebbesson[1], Laura Mannarino[2,3], Lara Paracchini[2,3], Sergio Marchini[3], Maurizio D'Incalci[2,3], Anna Måsbäck[4], Susanne Malander[1], Päivi Kannisto[5], Ingrid Hedenfalk[1]*

[1]Division of Oncology, Department of Clinical Sciences Lund and Lund University Cancer Centre, Lund University and Skåne University Hospital Comprehensive Cancer Centre; Lund, Sweden.

[2]Department of Biomedical Sciences, Humanitas University; Pieve Emanuele, Milan, Italy.

[3]Laboratory of Cancer Pharmacology, IRCCS Humanitas Research Hospital; Rozzano, Milan, Italy.

[4]Department of Surgical Pathology, Division of Laboratory Medicine, Skåne University Hospital; Lund, Sweden.

[5]Department of Obstetrics and Gynaecology Lund, Skåne University Hospital, Lund University; Lund, Sweden.

[†]These authors contributed equally to this work

[‡]Current affiliation: Department of Medical Epidemiology and Biostatistics, Karolinska Institutet; Stockholm, Sweden.

*Corresponding author: Ingrid.Hedenfalk@med.lu.se.

Abstract

Ovarian cancer lacks effective screening methods, leading to late diagnosis and poor outcomes. Studies have detected ovarian carcinoma-derived somatic mutations in liquid biopsies, highlighting the potential for early detection. Since high-grade serous ovarian carcinoma (HGSC) is primarily driven by copy number alterations (CNAs), we analyzed CNAs from shallow whole genome sequencing (sWGS) in 212 cervical samples from 128 women (58 with HGSC, 22 prophylactically operated women with germline *BRCA1* or *BRCA2* mutations, and 48 benign controls). Samples were collected at surgery and up to 94 months prior. Using the machine-learning classifier SRIQ, we developed HCsig, a predictor for HGSC detection. HCsig identified HGSC in 79% of archival cervical samples, including 91% stage I-II (range: 0-27 months), and 77% stage III-IV (range: 0-65 months) cases. Detection rates were 90% and 76% for *BRCA1/2*-mutated and wildtype HGSC, respectively. Validation in 172 independent samples (range: 0-98 months) showed 76% sensitivity and 94% specificity (AUC=0.83), including high sensitivity for early-stage cancers. We show that HGSC-derived copy number features are detectable with high performance using our HCsig classifier in archival cervical samples collected several years before the development of symptoms and eventual diagnosis, including from early-stage disease. Combined, the high performance holds promise for future adaptation into a test for early detection. Secondary prevention based on earlier-stage detection addresses the unmet need for HGSC screening.

INTRODUCTION

Almost two-thirds of women diagnosed with ovarian cancer die from their disease. The main reason for the high mortality rate is the lack of effective screening methods and that most patients are diagnosed with advanced-stage disease (*1*). As women detected with localized disease have a survival rate >90%, considerable efforts are being invested in developing effective screening tools for early detection.

The majority of high-grade serous tubo-ovarian carcinomas (HGSC) are derived from precursor lesions arising from epithelial cells in the fimbriated end of the fallopian tubes (*2*). Pre-cancerous lesions, so-called serous tubal intraepithelial carcinomas (STICs) harbor clonal *TP53* mutations and display a histological appearance resembling HGSC (*3*), (*4*). Dysplastic cells harboring mutant *TP53* (p53 signatures) are believed to precede the STIC lesions (*5*), (*6*). These findings provide evidence for the possibility of finding tumor-driving mutations in asymptomatic women with early-stage HGSC (*7*). To this end, Kinde *et al.* showed that somatic mutations in DNA shed from ovarian cancers could be detected in standard liquid-based Pap test specimens collected at the time of diagnosis by massively parallel sequencing (*8*). More recently, we and others have reported the detection of tumor-specific *TP53* mutations in archival liquid-based cervical cytology samples obtained from asymptomatic women up to several years prior to the HGSC diagnosis using ultra-sensitive methods (*9*), (*10*). However, it has recently been shown that very low-frequency, cancer-like, age-related *TP53* mutations are found in healthy women, revealing the challenge of using mutations in *TP53* for the early detection of ovarian cancer (*11*), (*12*). Hence, while sensitivity has been the main constraint in previous studies, specificity has emerged as the major challenge for early detection using ultra-sensitive sequencing approaches.

HGSC is a prototypical tumor characterized by genomic instability, where mutational processes acting on cancer genomes drive copy number alterations that can be readily investigated by whole genome sequencing (WGS) (*13*). In fact, evolutionary analyses have reported that aneuploidy and genomic instability, in addition to *TP53* mutations, can occur as early in HGSC progression as in the STIC lesions (*14*). To address the possibility of detecting the earliest steps in the evolution of HGSC using an unbiased approach, we therefore applied low-pass/shallow WGS (sWGS) to assess the consequence of the earliest known molecular event in the development of HGSC, mutations in *TP53*, leading to genomic instability. In this study, we interrogated genome-wide copy number alterations (CNAs) in a discovery cohort of 309 samples, including 212 cervical samples collected at surgery and archival samples from routine screening for cervical cancer taken up to 94 months prior to surgery from HGSC patients, *BRCA1/2* mutation carriers undergoing risk-reducing salpingo-oophorectomy (RRSO) and women with benign gynecological conditions. We applied previously published copy number signatures derived from HGSC (*15*) as well as pan-cancer copy number signatures for chromosomal instability (*16*), (*17*) to explore the genomic landscapes in cervical samples. Using the unsupervised, machine learning-based clustering method SRIQ (*18*) we then derived novel copy number signatures (HCsig) from diagnostic and archival cervical samples. A prediction model was constructed based on centroids for HCsig, which was tested and validated in an independent cohort of 172 samples from HGSC and healthy women (**Fig 1**). Our findings demonstrate the feasibility of applying sWGS to standard cervical samples for simple, minimally-invasive, and accurate early detection of HGSC, with the potential for downstaging and improved survival.

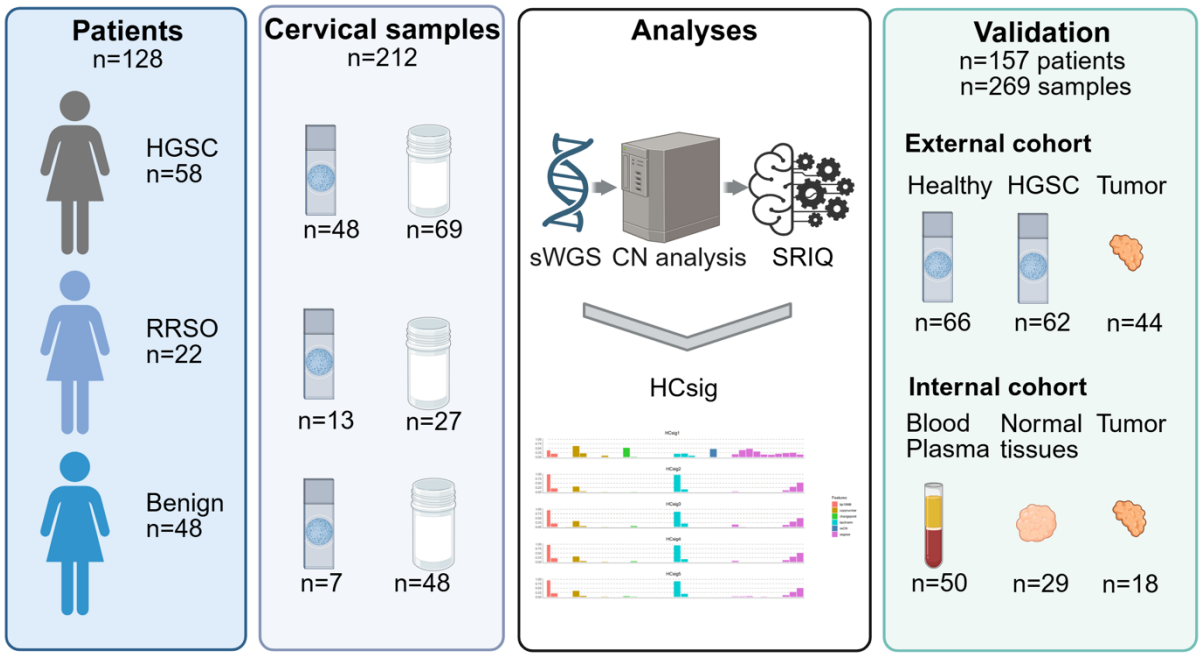

Fig. 1. Overview of patients, samples and workflow for HCsig. Liquid- and slide based cervical samples from women with HGSC, *BRCA1/2* mutation carriers undergoing risk-reducing salpingo-oophorectomy (RRSO) and women with benign gynecological conditions were used for sWGS. Cervical samples were from 0-94 months before surgery. The random forest machine learning-based clustering method SRIQ was applied to define HGSC-derived copy number signatures, HCsig 1-5. Validation was performed using an independent cohort of women with HGSC and healthy controls (external cohort), as well as normal tissues, tumors and blood/plasma from the discovery cohort (internal cohort). Created using BioRender.com.

RESULTS

sWGS and cancer-derived copy number signatures in cervical samples, tumors and normal tissues

The discovery cohort consisted of cervical samples (slide- and liquid-based, collected at diagnosis, and up to 8 years before diagnosis), tumors, blood/plasma and normal tissues from a previously published prospective cohort (*9*), and a retrospective cohort (total n=58 HGSC, 22

RRSO and 48 with benign gynecological conditions). The patient cohorts and sample types are shown in **Table 1.** Information on *BRCA1/2* mutation status, FIGO stage and CA-125 levels was retrieved from patient records. The majority of the HGSC patients were diagnosed with advanced-stage disease (47/58, 81%). We applied sWGS to all samples, aiming at a coverage of approximately 1X (range 0.4-2.2X). Although the DNA quality from slides was poorer than the liquid-based cervical samples, the results of the library preparation and sequencing were satisfactory. The average number of mapped reads was $49.2x10^6$, $25.4x10^6$, $19.8x10^6$, $41.7x10^6$, $40.6x10^6$, $57.9x10^6$ for tumor, blood, plasma, normal tissues, slide-based, and liquid-based cervical samples, respectively (range: $27.2x10^6$-$63.4x10^6$, $18.1x10^6$-$41.8x10^6$, $16.5x10^6$-$24.8x10^6$, $32.1x10^6$-$64.1x10^6$, $12.7x10^6$-$55.3x10^6$, $34.0x10^6$-$105.0x10^6$). Various fixed bin sizes were explored based on sequencing quality, with a final selection of 50kb for all sample types. The generation of relative and absolute copy numbers, and the assessment of copy number signatures, was performed as outlined in the Methods section.

Signature similarity exposure profiles from three previously published collections of cancer-derived copy number signatures (*15*), (*16*), (*17*) were evaluated to identify copy number patterns across sample types, including cervical samples (diagnostic samples from the time of surgery and archival samples collected during routine screening for cervical cancer), tumor and normal tissue, blood and plasma from women with HGSC, women at high risk of developing HGSC due to germline *BRCA1/2* mutations and women with benign gynecological conditions (**Table S1**).

Table 1. Patient and sample information

		HGSC	RRSO	Benign	Total
Patients (n)		**58**	**22**	**48**	**128**
Age	Range (years)	43-77	36-73	23-86	
	Median (years)	64	45	66	
BRCA1/2	Mutant (n)	10*	22	0	**32**
	Wildtype (n)	42	0	0	**42**
	Not tested (n)	6	0	48	**54**
Stage	Advanced (FIGO III and IV) (n)	47	N/A	N/A	
	Localized (FIGO I and II) (n)	11	N/A	N/A	
CA-125	Range (U/mL)	7.4-11,666†	†	4.1-1,207†	
	Median (U/mL)	925	N/A	35	
Samples (n)		**173**	**54**	**82**	**309**
	Cervical samples (n)	**117**	**40**	**55**	**212**
	Diagnostic (n)	31	17	55	103
	Archival (n)	86	23	0	109
	Additional samples (n)	**56**	**14**	**27**	**97**
	Tumor tissue (n)	18	N/A	N/A	18
	Normal tissue‡ (n)	9	7	13	29
	Blood (n)	18	4	7	29
	Plasma (n)	11	3	7	21

HGSC, high-grade serous ovarian cancer; RRSO, risk-reducing salpingo-oophorectomy

*, n=6 germline *BRCA1* mutations, 1 germline *BRCA2* mutation, 3 somatic *BRCA1* mutations

†, missing CA-125 values: n=2 HGSC, 22 RRSO, 12 benign

‡, endometrium, ovary, uterus

N/A, not applicable

The HGSC-derived CN signatures (*15*) more readily captured CNAs in cervical samples from HGSC patients in our cohort compared to the pan-cancer signatures (**Fig. S1**). Overall, HGSC and pan-cancer signatures defined by a high number of breakpoints, short to medium-sized, clustered and oscillating copy number changes, proposed to be caused by impaired homologous recombination, replication stress and/or chromothripsis, were commonly observed in tumors and cervical samples from patients with HGSC. In addition, the allele-specific panConusig signatures also captured LOH events in both tumors and a subset of cervical samples.

Collation and exploration of HGSC-derived copy number signatures using cervical samples

The previously reported copy number signatures were derived from advanced-stage cancers and may therefore lack the sensitivity required for capturing the early stages of HGSC. With the aim of developing a screening test for the early detection of HGSC using cervical samples, we decided to use the cervical samples in our discovery cohort to generate copy number signatures indicative of HGSC based on the six fundamental copy number features previously reported as hallmarks of genomic aberrations (*15*), including breakage-fusion-bridge cycles, chromothripsis and tandem duplication, and found in both HGSC tumor tissue and cervical samples in our cohort.

Cervical samples from the HGSC group (n=117) and healthy women (n=55 benign and 40 RRSO) were included in the analysis. To generate HGSC-derived Cervical copy number signatures (HCsig), we applied the unsupervised, random forest machine learning-based clustering method SRIQ (*18*) to discover clusters of samples in our data and to identify the copy number features defining these clusters. SRIQ identified five discrete copy number clusters (signatures), with the majority of the HGSC cervical samples being enriched for HCsig1. The heatmap in **Fig. 2A** displays the five clusters and their defining features in each sample, and the bar plot in **Fig. 2B** shows the weights of the defining components in the signatures, denoted HCsig1-5.

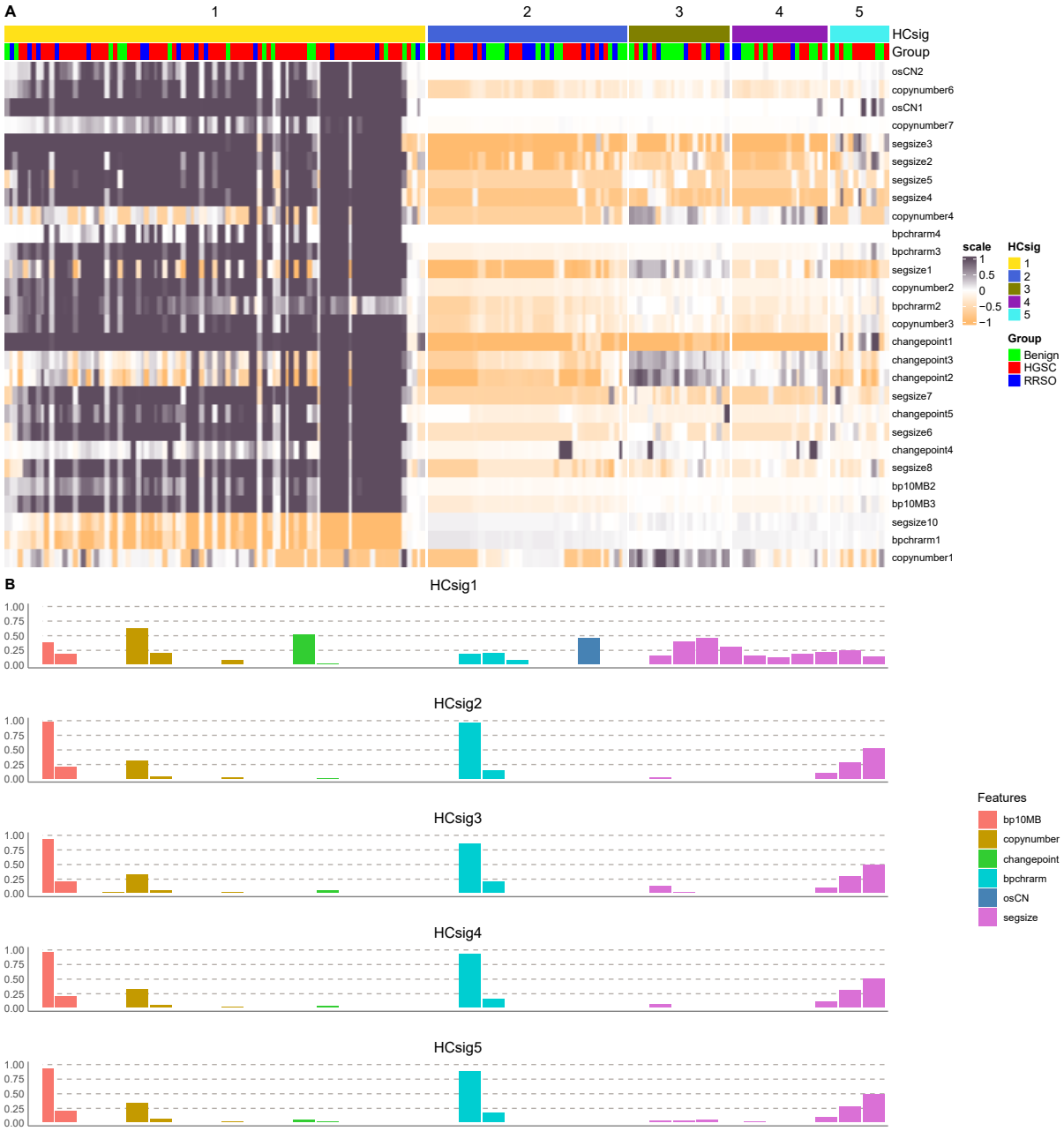

Fig. 2. Copy number features underlying the HCsig classifier in cervical samples. (A). SRIQ-based clustering and heatmap of cervical samples from patients with HGSC (n=117), benign gynecological disease (n=55) and who had RRSO (n=40). Samples are clustered on the x-axis, with the HCsig clusters 1-5 and patient groups shown above. The copy number features are shown on the y-axis. **(B).** Bar plot displaying the distributions and weights of the defining copy number components for HCsig 1-5.

The majority (67/117, 57%) of the HGSC cervical samples clustered together and exhibited enrichment for HCsig1, defined by highly variable segment sizes, oscillating copy number events, one to two breakpoints per chromosome arm, breaks evenly distributed across the genome, single copy number change points and copy number two. Notably, HCsig1 shared many features with the HGSG-derived CN signatures s3, s6 and s7 related to HR deficiency and failure of cell cycle control (*15*), and was identified as a descriptor for HGSC in the HGSC cervical samples in our cohort. HCsig2-5 demonstrated few defining copy number features and limited differences; these signatures were mainly defined by large segment sizes and few breakpoints. The majority of the cervical samples from women with benign conditions and from women who had undergone RRSO were enriched for HCsig2-5, suggesting that large segment sizes and the occurrence of few breakpoints are not tumor-specific features. Taken together, samples enriched for HCsig1 (*i.e.* if HCsig1 was the most highly enriched signature, or, alternatively, if HCsig1 was the second-most enriched signature with a delta <0.05 compared to the highest signature, see Methods) were considered as HCpositive (**Fig. 3**).

We next examined whether our sWGS approach could be used to capture HGSC in archival cervical samples, mirroring the pre-diagnostic/pre-symptomatic setting. The HCsig clusters were converted into centroids and used to perform a binary centroid-based prediction analysis (HCpositive or HCnegative). In total, 46/58 (79%) women with HGSC had at least one positive cervical sample before surgery. Liquid and slide-based cervical samples showed similar distributions of HCsig1-5 and similar fractions of HCpositive and HCnegative samples, indicating limited impact of sample processing (**Fig. S2**). Importantly, 91% (10/11) of women with stage I-II HGSC had at least one positive cervical sample before surgery (range: 0-27 months), while 77% (36/47) of women with advanced stage disease (stage III-IV) had at least

one positive cervical sample before surgery (range: 0-65 months). Notably, for women with *BRCA1/2*-mutated tumors (germline or somatic), 90% (9/10) were detectable in at least one pre-diagnostic cervical sample (range: 0-54 months), compared to 76% (32/42) of the *BRCA1/2* wildtype tumors (range: 0-51 months). Among healthy women, 20/70 (29%) had a false positive cervical sample; 11 of these women were *BRCA1/2* germline mutation carriers who had an RRSO, one of whom displayed STICs according to the pathology reports (**Fig. 3**).

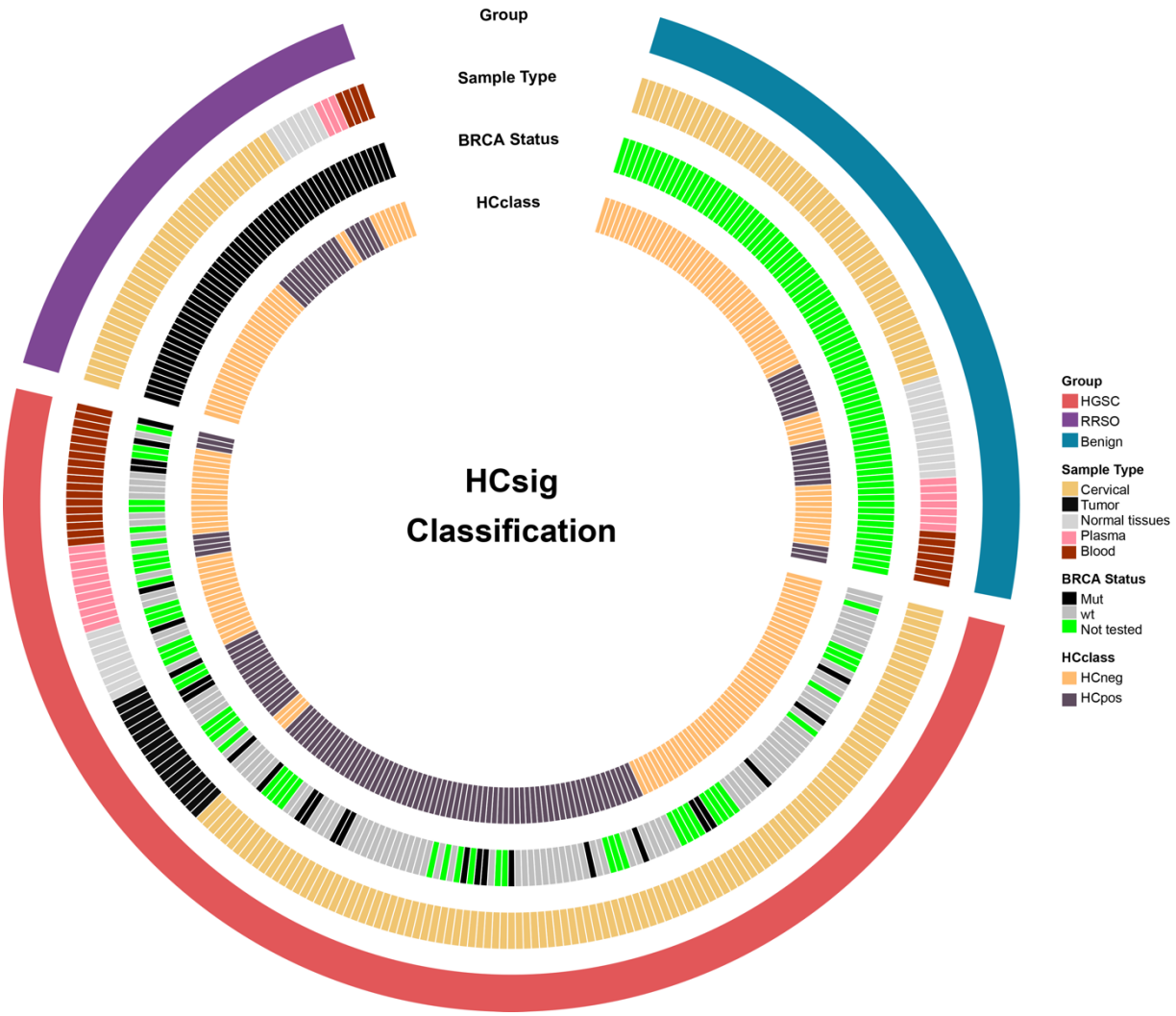

Fig. 3. HCsig classification in cervical and non-cervical samples. Samples (sample type: cervical, tumor tissue (only HGSC), normal tissues, plasma and blood, respectively) from women with HGSC, benign gynecological

conditions and women who had RRSO (group) are displayed in the plot. Annotations include *BRCA1/2* mutation status and HCsig class (HCpositive or HCnegative).

To further explore the distribution and origin of the features identified in the HCsig signatures, we investigated the exposures across different sample types using cosine similarity, clearly demonstrating enrichment of the HCpositive class in the majority of tumor tissue samples (15/18, 83%), demonstrating the HGSC origin of HCsig1. Only 3/18 (16%) of blood and 4/11 (36%) plasma samples from HGSC patients were HCpositive, indicating lower sensitivity in peripheral samples compared to cervical samples (**Fig. 3**). When including all sample types from all patient groups in the analysis, the sensitivity of the HCsig classifier was 86%, the specificity 79% and the accuracy 82%. The normal samples (blood, plasma and normal tissues) displayed similar exposures to all HCsig signatures; *i.e.* no enrichment of tumor-like features. There were no indications in the pathology reports specifying the presence of malignant features in tissues from women with an HCpositive sample from the benign or RRSO groups. However, the fallopian tube tissue from one of the women who had undergone an RRSO was found to display a STIC lesion. Taken together, the most distinguishing features captured in HCsig1 (HCpositive) in cervical samples from HGSC patients were also more prevalent in tumor tissue compared to normal tissues.

Modelling pre-symptomatic detection: analysis of HCsig in longitudinal cervical samples from HGSC patients

To explore the ability of the HCsig classifier to capture copy number features from HGSC in cervical samples collected during routine gynecological examination of pre-

symptomatic women and who were later diagnosed with HGSC, we applied the HCsig classifier to archival cervical samples to simulate the pre-symptomatic setting. Women later diagnosed with HGSC, who had ≥2 cervical samples taken at different timepoints were included in the analysis (n=20 patients) and samples were ordered according to the time of sampling in relation to the time of diagnosis (range: 0-94 months).

The HCsig classification of cervical samples from women diagnosed with HGSC sorted by time is shown in **Fig. 4**. Enrichment for the HCpositive class was observed in both diagnostic and archival cervical samples from women with HGSC. Seventeen of 20 patients with HGSC (85%) had at least one HCpositive cervical sample before surgery (up to 65 months before diagnosis). Enrichment for the HCpositive class in archival samples obtained up to >5 years before diagnosis supports the notion that copy number features captured from sWGS data may enable early detection of HGSC in pre-symptomatic women.

In summary, these findings show that complex CNAs resembling those observed in HGSC and captured using HCsig can be detected in archival cervical samples obtained from asymptomatic women up to several months, and even years, prior to the onset of symptoms and subsequent diagnosis.

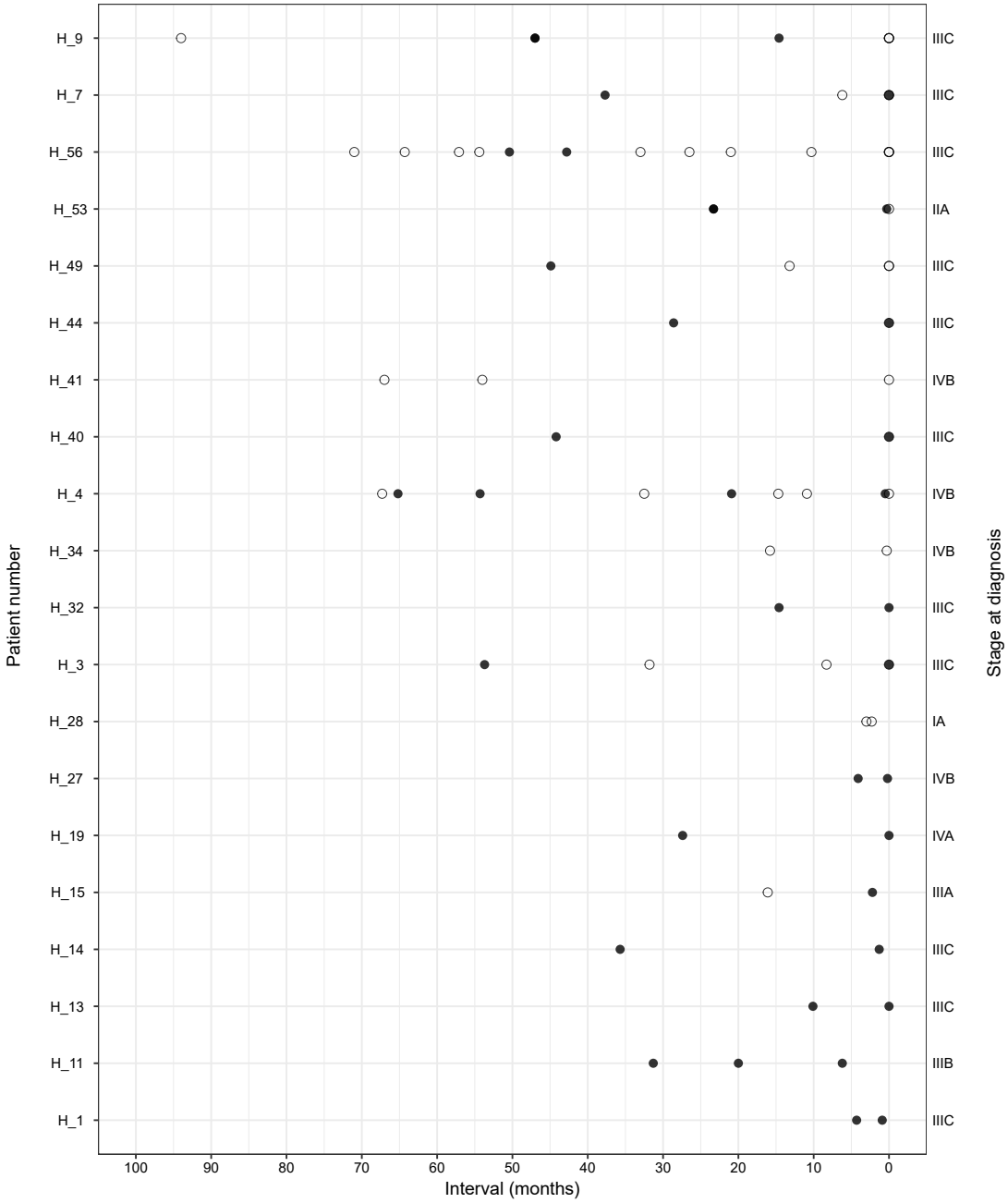

Fig. 4. HCsig classification in longitudinal cervical samples from women with HGSC. The dot plot shows the classification of cervical samples from women with HGSC and ≥2 cervical samples available. The x-axis shows the time in months before diagnosis (right to left). Filled circles represent HCpositive samples and open circles represent HCnegative samples. FIGO stage at diagnosis of HGSC is indicated for each patient.

Validation of the HCsig classification model for HGSC using samples from an independent dataset

Finally, to examine the diagnostic potential of our classification model, we performed a validation using an independent dataset with a comparable number of samples and patients, including archival cervical samples from women with HGSC and healthy women. The time from the collection of cervical samples to the time of HGSC diagnosis ranged from 0 months to 13.6 years (*19*). sWGS data (1-5X) from 172 samples were preprocessed and analyzed using QDNASeq, followed by Rascal and calculation of copy number features as described above. Bin sizes were selected based on sequencing depth and quality, with a 50 kb bin size used for cervical and tumor samples and 500 kb used for the pool of normals (PoN). The SRIQ-derived HCsig clusters were converted into centroids and used to perform a binary centroid-based prediction analysis (HGSC or healthy donors). Cross-tabulation was performed, showing the correct classification of 81/106 HGSC and 62/66 non-HGSC samples. Among women with HGSC and at least one cervical sample, 43/61 (70%) women had at least one HCpositive cervical sample (up to 98 months before diagnosis). In all, only four false positive and 25 false negative samples were found, yielding a specificity of 94% (95% CI: 0.90, 0.98) and sensitivity of 76% (95% CI: 0.70, 0.83). A receiver operating characteristic curve yielded an accuracy of AUC=0.83 (95% CI: 0.78, 0.89) (**Fig. 5**). Notably, HCsig detected a majority of the early-stage HGSC cases (5/7) in the cervical samples from the external cohort. None of the four false positive cervical samples from healthy women was reported to have a positive CPA genome score (Copy number Profile Abnormality >0.61).

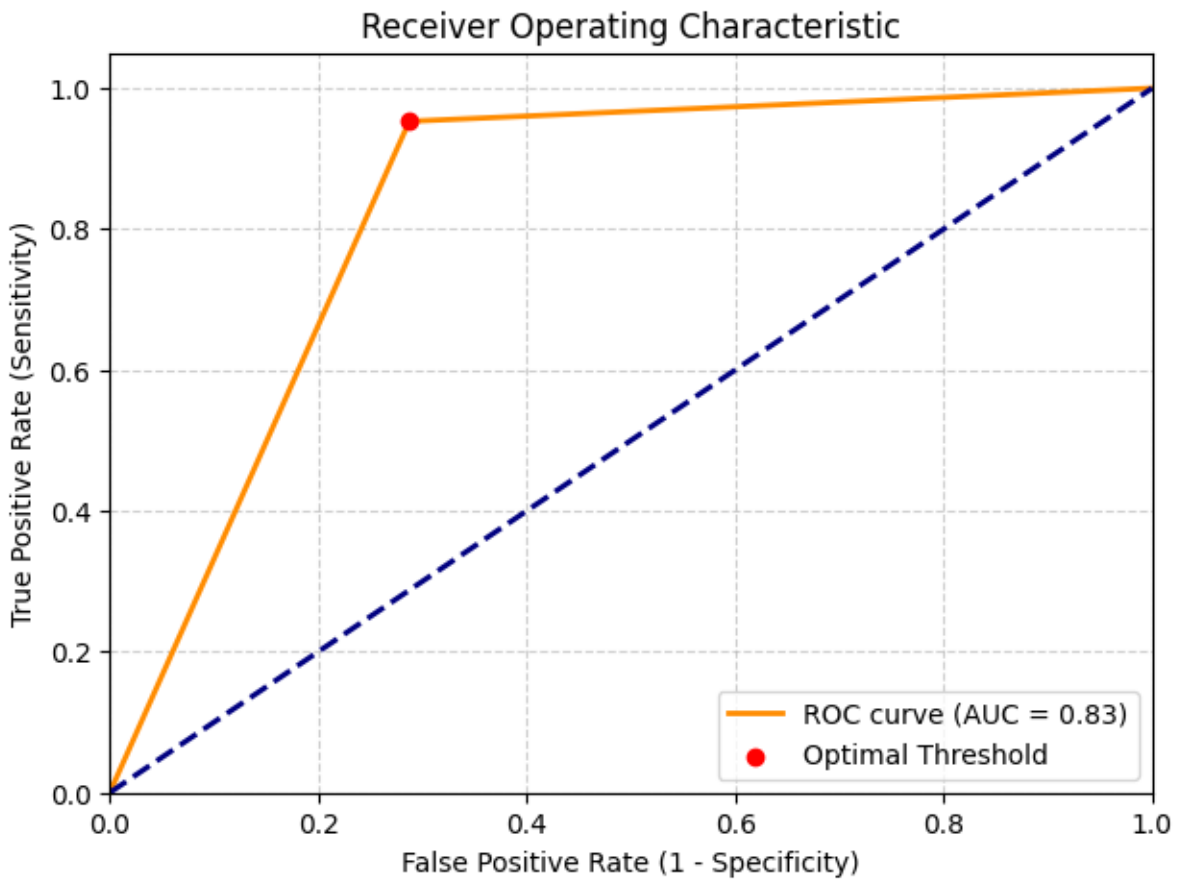

Fig. 5. Validation of HCsig in the independent cohort. The receiver-operating curve (ROC) shows the sensitivity and specificity of HCsig in the independent cohort from Paracchini *et al.* (*19*) with healthy women and women with HGSC.

DISCUSSION

In the present study we have demonstrated that the early detection of genomically aberrant HGSC-derived DNA in cervical samples collected from asymptomatic women is feasible. Increased early detection rates may enable earlier-stage diagnosis and hence improve survival rates for women with the most common type of ovarian cancer, HGSC. A screening tool for early detection of HGSC is an urgent unmet medical need and is required to be both highly sensitive and specific. Based on the premise that HGSC-derived cells and DNA can be shed into

the cervix due to the anatomical connection via the uterine cavity and given evidence of genomic instability patterns in early HGSC development [7], we developed HCsig, a model leveraging copy number signatures from sWGS data for the early detection of HGSC in conventional cervical samples (*7*). While ultra-deep sequencing approaches have illustrated the possibility of highly sensitive detection of cancer-derived mutant *TP53* (*11*), (*20*), the occurrence of somatic clonal evolution linked to ageing limits its utility as a diagnostic test.

We show that sWGS combined with machine learning to characterize genomic features prototypical of HGSC and readily captured in cervical samples is both sensitive and specific, despite the low abundance of HGSC-derived DNA in the cervical samples. This approach provides a quantitative and qualitative measure of CNAs, presenting an improvement over previous gene-centric methods (*8*), (*21*), (*22*). Sampling in close proximity to the tumor mass contributes to the accuracy of this approach, supported by the higher fraction of true positive cervical samples vs. plasma samples from HGSC patients.

While retrospective in nature, the access to archival cervical samples allowed for an unbiased investigation of CNAs in non-symptomatic women up to several years before their HGSC diagnosis (*7*). The inability to correctly classify a subset of cervical samples from women with HGSC may be due to the closeness in time to surgery, involving multiple samplings during work-up, resulting in inferior sample quality or insufficient tumor-derived DNA. Another possible explanation may be blockage of the fallopian tubes in advanced-stage disease. Importantly, the performance of our model was validated in an independent cohort with high accuracy, suggesting that it is generalizable across populations and laboratories.

HGSC is characterized by a low frequency of recurrent oncogenic mutations, few recurrent copy number alterations and highly complex genomic profiles, which pose challenges

for screening, diagnosis and classification of the disease. Novel approaches to early detection have focused on identifying the ubiquitous *TP53* mutations (*9*), (*10*) or a small panel of cancer-related genes (*23*), (*22*) in plasma or in samples of cervical or uterine origin. The main constraint of these mutation-based screening studies has been the lack of sensitivity, while specificity has recently emerged as the key challenge due to ultra-sensitive methods capturing clonal evolution of cancer-like age-related *TP53* mutations in healthy women and normal tissues (*11*). However, unlike mutations, copy number alterations are rarely present in normal tissues and may thus present a robust foundation for creating a test able to distinguish early cancer from benign or normal tissue. A quantitative analysis of genomic instability was recently proposed for early detection of ovarian cancer in cervical samples collected up to several years before diagnosis, with promising results (*19*). In the present study, we used an unsupervised approach, including mixture modelling and cosine similarities to extract both quantitative and qualitative genomic features known to occur in HGSC (*15*) for each cervical sample to generate a prediction model based on distinct copy number features of HGSC. Importantly, the method is tumor mutation agnostic, able to capture early-stage disease, relatively inexpensive, and requires only small amounts of tumor-derived DNA. We provide proof-of-concept that CNAs found in archival cervical samples can be used as a biomarker to distinguish between HGSC and benign gynecological conditions, up to several years before the onset of symptoms and diagnosis. Of note, some women in the study underwent frequent sampling due to suspected cervical cell atypia, but these samples were generally called as HCnegative by the classifier, hence not confounding the results. Importantly, both advanced-stage and localized HGSC were detectable using our classifier, supporting the sensitivity and specificity of HCsig and the reported early genomic evolution of HGSC from STIC lesions (*7*). The detection of early-stage cancers,

involving growth only on the ovaries and/or fallopian tubes, supports the robustness of the classifier and provides a promising approach for early detection with the possibility of increasing curability and reducing mortality. This supports that sampling in close proximity to the tumor holds superior sensitivity over *e.g.* plasma and reflects the tubal origin of HGSC. Enrichment of copy number signatures indicative of mutational processes described in HGSC (*15*) and other epithelial cancers (*16*), (*17*), including HRD and tandem duplications support the HGSC-derived origin in the cervical samples in our study. In line with our findings, using a combination of whole-genome cell-free DNA (cfDNA) fragmentome and protein biomarker (CA-125 and HE4) analysis, Medina *et al.* recently reported a high accuracy for the detection of stage I-IV HGSC (*24*). Exploring our HCsig classifier in combination with fragmentomes in cervical samples and conventional protein biomarkers may improve the performance further.

A strength of our study was the inclusion of women with germline *BRCA1/2* mutations at high risk of developing HGSC, who underwent RRSO and hence a thorough pathology review. While no invasive cancers were detected, STIC lesions were noted for one of 13 women. Of note, 13/22 (59%) false positive cervical samples were from these *BRCA1/2* mutation carriers, suggesting that *BRCA1/2* heterozygous cells may instigate carcinogenesis in the fallopian tubes due to potential haploinsufficiency. This possibility is supported by reports of *BRCA1* haploinsufficiency causing genomic instability in mammary epithelial cells (*25*). Further refinement of HCsig in a larger cohort of *BRCA1/2* mutation carriers is required to establish a cut-off for distinguishing between the *BRCA1/2* heterozygous state and complete loss of function. A recent report showed that *TP53* mutation burden in cervical cytology samples increased with age in *BRCA1/2* germline mutation carriers with HGSC (*20*), leading to greater genomic instability. This finding further supports the potential of risk assessment based on

cumulative genomic instability using *e.g.* HCsig to aid in decision-making regarding the timing of RRSO for high-risk women.

To summarize, we have shown that HGSC-derived copy number signatures can be captured in cervical samples collected up to several years prior to diagnosis from women later diagnosed with HGSC with high sensitivity and specificity. The results were validated in an independent cohort, strengthening the findings further and supporting the robustness and generalizability of the method. Our findings indicate that early detection is possible in asymptomatic women, opening up the possibility for the development of a copy number-based tool for early diagnosis. HCsig may be applied in a screening setting for early detection in high-risk individuals, and as a diagnostic tool in symptomatic women. Further characterization of the signatures and a large, prospective study should pinpoint which components to include in a diagnostic tool for early detection and increased survival.

MATERIALS AND METHODS

Study design

This study included 128 patients and 309 samples: 173 samples from 58 women diagnosed with HGSC, 82 samples from 48 women diagnosed with benign gynecological conditions and 54 samples from 22 *BRCA1/2* mutation carriers who underwent RRSO. Clinical information and sample types by patient group are presented in **Table 1**.

Two cohorts were combined for the present study; one prospective cohort where diagnostic cervical samples were collected from 187 women who underwent RRSO or had surgery due to adnexal masses at the Gynecological Department in Lund, Skåne University

Hospital, Sweden (2015-2017) (*9*) and one cohort where 59 patients with HGSC or RRSO surgery (2017-2020), with accessible archival cervical samples, were retrospectively identified. To be eligible for this study, the availability of archival samples was also required from the patients in the first cohort. A flowchart of the inclusion and exclusion of patients is shown in **Fig. S3**. The study was approved by the Ethics Committee at Lund University, Sweden.

Cervical samples

Archival cervical samples from cervical cancer screening were obtained from the Pathology Department at Skåne University Hospital, Sweden, while fresh-frozen tissue, blood, and plasma samples were available from the regional biobank for gynecological conditions in the southern Swedish healthcare region. Liquid-based cervical cytology samples were preserved in either DNAgard® (Sigma-Aldrich), ThinPrep® solution (Hologic) or SurePath® solution (BD Biosciences). Diagnostic samples were collected at surgery, and archival samples were from any time before surgery (range: 0 to 94 months).

In total, we analyzed 117 cervical samples from the HGSC group: 31 diagnostic cervical samples (from 19 patients), and 86 archival samples (from 52 patients). For the RRSO group, we analyzed 40 cervical samples: 17 samples were collected at surgery (from 12 patients), and 23 were archival (from 14 patients). Fifty-five diagnostic cervical samples were from women with benign gynecological conditions (48 women). Twenty-five patients (20 HGSC and five RRSO) had samples available from multiple time points. In addition, blood, plasma, tumor and normal tissues were also collected, for a total of 309 samples from 128 patients.

Sample processing and library preparation

The cover glass or film was removed from cervical cytology slides using xylene or acetone, respectively, and DNA was extracted with the Maxwell RSC DNA FFPE kit (Promega). Liquid-based cervical samples were extracted with the QIAamp DNA Micro kit (Qiagen). For the additional samples, DNA was extracted using the AllPrep DNA/RNA Mini kit (fresh-frozen tissue) or QIAmp DNA Blood Midi kit (blood and plasma).

DNA concentrations were measured by Qubit fluorometric assays and quality-checked by Illumina TruSeq FFPE Library Prep QC Kit (Illumina). Sequencing libraries were prepared from 1-10 ng DNA using the QIAseq FX DNA library prep kit (Qiagen). qPCR with EvaGreen DNA-binding dye was used to monitor amplification (4-12 cycles). Low pass/shallow whole genome sequencing (sWGS) was performed on the NovaSeq 6000 platform (Illumina) at the Center for Translational Genomics (CTG) in Lund, Sweden. The target sequencing depth was >1X for all sample types, except blood and plasma, which were sequenced at 0.5X. Approximately 15% of the extracted cervical samples were not sequenced due to poor DNA quality and/or quantity.

Data analysis workflow

A data analysis workflow was assembled to extract copy number features from all samples. We calculated similarities with previously published cancer-derived copy number signatures and generated novel copy number signatures from HGSC in cervical samples using sWGS data. Demultiplexed FASTQ files from paired-end sWGS data were used as input. The workflow contains copy number analysis to generate relative copy number profiles and solutions for cellularity and ploidy, as well as the manually curated segmentation bin sizes for each

sample. Absolute copy number profiles were generated and used to explore HGSC-derived copy number signatures (CNsignatures) (*15*), pan-cancer chromosomal instability copy number signatures (CIN signatures) (*16*), and allele-specific copy number signatures (panConusig) (*17*). For the construction of the HGSC-specific cervical copy number prediction model (HCsig), we used the absolute copy number profiles of the cervical samples from patients with and without HGSC, performed unsupervised SRIQ clustering (*18*) to generate copy number-based signatures and subsequently a prediction model, which was then tested and validated using an independent dataset from Paracchini *et al.* (*19*).

Absolute copy number analysis

Raw sequencing data were demultiplexed using Illumina's bcl2fastq software to generate FASTQ files, which were subsequently aligned to the reference genome (GRCh37) using BWA (v0.7.17-r1188). The alignment files were processed with Samtools (v1.13) to convert SAM to BAM format, followed by duplicate marking using PicardTools (v3.3.0), and GATK software (v4.1) was applied for base quality score recalibration (BQSR).

Relative copy number profiles were generated using a modified version of QDNAseq. Custom bin annotations for the GRCh37 reference genome were created to optimize the analysis for a 100bp paired-end sequencing setup, following the package's recommended workflow. A new control dataset, comprising 58 human genomes from the 1000 Genomes Project Phase 3, was incorporated to enhance analytical accuracy.

The segmented relative copy number profiles, derived under manually optimized bin sizes for the 309 samples, were processed using the Rascal R package. Rascal employs the ACE

(Absolute Copy-number Estimation) method to scale relative copy numbers to absolute values. To refine the results, we applied an additional model adjustment using mutant allele frequencies (MAFs) for *TP53*, previously quantified via ddPCR (IBSAFE®, Saga Diagnostics) in three patients (*9*).

Integration with published copy number signatures

The outputs generated by Rascal were utilized to derive sample-by-component matrices for HGSC-specific copy number signatures (*15*) and pan-cancer chromosomal instability (CIN) signatures (*16*). We performed a cosine similarity analysis to identify signatures with the most similar patterns in our dataset. This compared the sample-by-component matrices with the corresponding signature-by-component definition matrices, resulting in sample-by-signature matrices containing cosine similarity values.

For cases with matched blood samples, we applied panConusig (*17*), which leverages allele-specific copy number profiles. Prior to calculating these profiles, we used a modified workflow of Battenberg v2.2.9 on the BAM files to generate haplotype-phased data. The workflow was adapted to include only the initial steps: counting the major and minor allele frequencies and determining phased states. The haplotype-phased allele frequency data were subsequently processed with the shallow-coverage version of ASCAT (ASCAT.sc v0.1) to calculate allele-specific copy number profiles.

Construction of HGSC-derived copy number signatures (HCsig) using cervical samples

Absolute copy number profiles in cervical samples from women with HGSC were used to extract genome-wide copy number features. These features included segment sizes (segsize), breakpoint counts per 10Mb (bp10MB), lengths of oscillating copy number segment chains (osCN), copy number change points (changepoint), segment copy numbers (copynumber), and breakpoint counts per chromosome arm (bpchrarm). Each feature was further decomposed into distinct components using mixture modeling, as described in (*15*), resulting in 36 unique components. These components formed the basis of a sample-by-component matrix. This matrix was then used as input for the SRIQ algorithm (*18*), an unsupervised clustering approach that integrates principles from the RandomForest machine learning classifier, quality thresholding, and k-nearest neighbor clustering. The aim of the SRIQ method is to delineate a core cluster of samples or components sharing common patterns without the need for *a priori* knowledge of the data or a predetermined number of clusters. As a result of this analysis, distinct clusters were successfully identified. Subsequently, mean values of components within each cluster were calculated to formulate individual cluster centroid profiles and were used to create High-grade serous cancer Cervical copy number signature profiles (HCsig).

Validation of the centroid-based prediction tool HCsig in independent data

The HCsig signature profiles were used to develop and validate a prediction model for identifying HGSC in an independent cohort comprising 172 samples. This cohort included 44 biopsy samples, and 128 cervical samples collected from presymptomatic women, with 62 samples from individuals diagnosed with HGSC and 66 from healthy controls [19]. The downloaded BAM files, which were processed using genome reference hg38, were run through our analysis pipeline to produce a sample-by-component matrix. Each sample was compared to

the HCsig centroids using Spearman correlation. Samples were then annotated with the HCsig signature showing the highest Spearman correlation value. Samples with close similarity to two centroids, one of which being HCsig1 (D<0.05 between HCsig1 and any other HCsig) were considered as HCpositive. A custom Java-based tool (see Code Availability section) was developed for this analysis.

The model's performance was evaluated using a receiver operating characteristic (ROC) curve, with the area under the curve (AUC) providing an overall measure of predictive accuracy. Additionally, confusion matrix metrics, including sensitivity, specificity, and accuracy, were calculated to assess the model's ability to classify samples accurately as originating from HGSC patients or healthy controls. All performance analyses, including calculations and visualizations, were conducted using Python's scikit-learn package.

Statistical analyses

Chi-square tests were used for categorical variables. P<0.05 was considered statistically significant. Confidence intervals (95% CI) were calculated using the Normal Approximation method.

List of Supplementary Materials

Supplementary Figures S1-S3

Supplementary Table S1

References

1. F. Bray, M. Laversanne, H. Sung, J. Ferlay, R. L. Siegel, I. Soerjomataram, A. Jemal, Global cancer statistics 2022: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J Clin* **74**, 229-263 (2024).
2. C. P. Crum, R. Drapkin, A. Miron, T. A. Ince, M. Muto, D. W. Kindelberger, Y. Lee, The distal fallopian tube: a new model for pelvic serous carcinogenesis. *Curr Opin Obstet Gynecol* **19**, 3-9 (2007).
3. D. W. Kindelberger, Y. Lee, A. Miron, M. S. Hirsch, C. Feltmate, F. Medeiros, M. J. Callahan, E. O. Garner, R. W. Gordon, C. Birch, R. S. Berkowitz, M. G. Muto, C. P. Crum, Intraepithelial carcinoma of the fimbria and pelvic serous carcinoma: Evidence for a causal relationship. *Am J Surg Pathol* **31**, 161-169 (2007).
4. Y. Lee, A. Miron, R. Drapkin, M. R. Nucci, F. Medeiros, A. Saleemuddin, J. Garber, C. Birch, H. Mou, R. W. Gordon, D. W. Cramer, F. D. McKeon, C. P. Crum, A candidate precursor to serous carcinoma that originates in the distal fallopian tube. *J Pathol* **211**, 26-35 (2007).
5. J. M. Piek, P. J. van Diest, R. P. Zweemer, J. W. Jansen, R. J. Poort-Keesom, F. H. Menko, J. J. Gille, A. P. Jongsma, G. Pals, P. Kenemans, R. H. Verheijen, Dysplastic changes in prophylactically removed Fallopian tubes of women predisposed to developing ovarian cancer. *J Pathol* **195**, 451-456 (2001).
6. A. Finch, P. Shaw, B. Rosen, J. Murphy, S. A. Narod, T. J. Colgan, Clinical and pathologic findings of prophylactic salpingo-oophorectomies in 159 BRCA1 and BRCA2 carriers. *Gynecol Oncol* **100**, 58-64 (2006).

7. 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).

8. I. Kinde, C. Bettegowda, Y. Wang, J. Wu, N. Agrawal, M. Shih Ie, R. Kurman, F. Dao, D. A. Levine, R. Giuntoli, R. Roden, J. R. Eshleman, J. P. Carvalho, S. K. Marie, N. Papadopoulos, K. W. Kinzler, B. Vogelstein, L. A. Diaz, Jr., Evaluation of DNA from the Papanicolaou test to detect ovarian and endometrial cancers. *Sci Transl Med* **5**, 167ra164 (2013).

9. N. S. Arildsen, L. Martin de la Fuente, A. Masback, 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).

10. 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).

11. 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 e133 (2019).

12. 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).
13. L. B. Alexandrov, J. Kim, N. J. Haradhvala, M. N. Huang, A. W. Tian Ng, Y. Wu, A. Boot, K. R. Covington, D. A. Gordenin, E. N. Bergstrom, S. M. A. Islam, N. Lopez-Bigas, L. J. Klimczak, J. R. McPherson, S. Morganella, R. Sabarinathan, D. A. Wheeler, V. Mustonen, P. M. S. W. Group, G. Getz, S. G. Rozen, M. R. Stratton, P. Consortium, The repertoire of mutational signatures in human cancer. *Nature* **578**, 94-101 (2020).
14. 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).
15. G. Macintyre, T. E. Goranova, D. De Silva, D. Ennis, A. M. Piskorz, M. Eldridge, D. Sie, L. A. Lewsley, A. Hanif, C. Wilson, S. Dowson, R. M. Glasspool, M. Lockley, E. Brockbank, A. Montes, A. Walther, S. Sundar, R. Edmondson, G. D. Hall, A. Clamp, C. Gourley, M. Hall, C. Fotopoulou, H. Gabra, J. Paul, A. Supernat, D. Millan, A. Hoyle, G. Bryson, C. Nourse, L. Mincarelli, L. N. Sanchez, B. Ylstra, M. Jimenez-Linan, L. Moore, O. Hofmann, F. Markowetz, I. A. McNeish, J. D. Brenton, Copy number signatures and mutational processes in ovarian carcinoma. *Nat Genet* **50**, 1262-1270 (2018).

16. R. M. Drews, B. Hernando, M. Tarabichi, K. Haase, T. Lesluyes, P. S. Smith, L. Morrill Gavarro, D. L. Couturier, L. Liu, M. Schneider, J. D. Brenton, P. Van Loo, G. Macintyre, F. Markowetz, A pan-cancer compendium of chromosomal instability. *Nature* **606**, 976-983 (2022).

17. C. D. Steele, A. Abbasi, S. M. A. Islam, A. L. Bowes, A. Khandekar, K. Haase, S. Hames-Fathi, D. Ajayi, A. Verfaillie, P. Dhami, A. McLatchie, M. Lechner, N. Light, A. Shlien, D. Malkin, A. Feber, P. Proszek, T. Lesluyes, F. Mertens, A. M. Flanagan, M. Tarabichi, P. Van Loo, L. B. Alexandrov, N. Pillay, Signatures of copy number alterations in human cancer. *Nature* **606**, 984-991 (2022).

18. J. Karlstrom, M. Aine, J. Staaf, S. Veerla, SRIQ clustering: A fusion of Random Forest, QT clustering, and KNN concepts. *Comput Struct Biotechnol J* **20**, 1567-1579 (2022).

19. L. Paracchini, L. Mannarino, C. Romualdi, R. Zadro, L. Beltrame, I. Fuso Nerini, P. Zola, M. E. Laudani, E. Pagano, L. Giordano, R. Fruscio, F. Landoni, S. Franceschi, M. L. Dalessandro, V. Canzonieri, L. Bocciolone, D. Lorusso, C. Bosetti, F. Raspagliesi, I. M. G. Garassino, T. group, M. D'Incalci, S. Marchini, Genomic instability analysis in DNA from Papanicolaou test provides proof-of-principle early diagnosis of high-grade serous ovarian cancer. *Sci Transl Med* **15**, eadi2556 (2023).

20. T. S. Ghezelayagh, B. F. Kohrn, J. Fredrickson, J. D. Krimmel-Morrison, E. Latorre-Esteves, X. R. Tee, M. R. Radke, E. Manhardt, B. M. Norquist, R. Katz, E. M. Swisher, R. A. Risques, TP53 somatic evolution in cervical liquid-based cytology and blood from individuals with and without ovarian cancer and BRCA1 or BRCA2 germline mutations. *Oncogene* **43**, 2421-2430 (2024).

21. E. Maritschnegg, Y. Wang, N. Pecha, R. Horvat, E. Van Nieuwenhuysen, I. Vergote, F. Heitz, J. Sehouli, I. Kinde, L. A. Diaz, Jr., N. Papadopoulos, K. W. Kinzler, B. Vogelstein, P. Speiser, R. Zeillinger, Lavage of the Uterine Cavity for Molecular Detection of Mullerian Duct Carcinomas: A Proof-of-Concept Study. *J Clin Oncol* **33**, 4293-4300. (2015).

22. Y. Wang, L. Li, C. Douville, J. D. Cohen, T. T. Yen, I. Kinde, K. Sundfelt, S. K. Kjaer, 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, Jr., 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**, (2018).

23. J. Phallen, M. Sausen, V. Adleff, A. Leal, C. Hruban, J. White, V. Anagnostou, J. Fiksel, S. Cristiano, E. Papp, S. Speir, T. Reinert, M. W. Orntoft, B. D. Woodward, D. Murphy, S. Parpart-Li, D. Riley, M. Nesselbush, N. Sengamalay, A. Georgiadis, Q. K. Li, M. R. Madsen, F. V. Mortensen, J. Huiskens, C. Punt, N. van Grieken, R. Fijneman, G. Meijer, H. Husain, R. B. Scharpf, L. A. Diaz, Jr., S. Jones, S. Angiuoli, T. Orntoft, H. J. Nielsen, C. L. Andersen, V. E. Velculescu, Direct detection of early-stage cancers using circulating tumor DNA. *Sci Transl Med* **9**, (2017).

24. J. E. Medina, A. V. Annapragada, P. Lof, S. Short, A. L. Bartolomucci, D. Mathios, S. Koul, N. Niknafs, M. Noe, Z. H. Foda, D. C. Bruhm, C. Hruban, N. A. Vulpescu, E. Jung, R. Dua, J. V. Canzoniero, S. Cristiano, V. Adleff, H. Symecko, D. van den Broek, L. J. Sokoll, S. B. Baylin, M. F. Press, D. J. Slamon, G. E. Konecny, C. Therkildsen, B.

Carvalho, G. A. Meijer, C. L. Andersen, S. M. Domchek, R. Drapkin, R. B. Scharpf, J. Phallen, C. A. R. Lok, V. E. Velculescu, Early detection of ovarian cancer using cell-free DNA fragmentomes and protein biomarkers. *Cancer Discov* **15**, 105-118 (2024).

25. M. Sedic, A. Skibinski, N. Brown, M. Gallardo, P. Mulligan, P. Martinez, P. J. Keller, E. Glover, A. L. Richardson, J. Cowan, A. E. Toland, K. Ravichandran, H. Riethman, S. P. Naber, A. M. Naar, M. A. Blasco, P. W. Hinds, C. Kuperwasser, Haploinsufficiency for BRCA1 leads to cell-type-specific genomic instability and premature senescence. *Nat Commun* **6**, 7505 (2015).

Acknowledgments

The authors wish to thank all patients who participated in the study, RCC South for providing infrastructure for the collection and biobanking of samples, and Clinical Genomics Lund, SciLifeLab, and Center for Translational Genomics (CTG) Lund University for sequencing services. We thank J-M Jönsson for critical review and feedback on the manuscript.

Funding

The Swedish Cancer Society grants 21 1684 Pj and 24 3839 Pj (IH)

The Sjöberg Foundation grant 2023-639 (IH)

The Berta Kamprad Foundation grant FBKS-2023-36 (IH)

The Cancer and Allergy Foundation grants 10381, 10672 and 11002 (IH)

The King Gustaf V Jubilee Foundation grant 197011 (IH)

Governmental funding of clinical research within the national health services (ALF) grant 40615 (IH)

Fondazione Alessandra Bono and AIRC grants IG 2024 ID30972 and ID30381 (MD, SM)

Author contributions

Conceptualization: LMF, PK, IH

Resources: LMF, LP, LM, SM, MD, AM, SM, PK, IH

Data curation: LMF, SV, ML, AE, IH

Methodology: SV, ML, GT, AE, IH

Investigation: LMF, SV, AE, IH

Visualization: SV, AE

Funding acquisition: SM, MD, IH

Project administration: IH

Supervision: SV, IH

Writing – original draft: IH

Writing – review & editing: all authors

Competing interests

The authors declare that they have no competing interests.

Data Availability

The data presented in this paper contain sensitive information that cannot be shared openly. Work on depositing the data in FEGA Sweden has been initiated. FEGA Sweden is a national node of the Federated European Genome-phenome Archive (FEGA), which allows data to be shared under controlled access. The datasets in FEGA Sweden are findable through the European Genome-phenome Archive web portal (https://ega-archive.org).

Code Availability

All analyses were performed using open-source software such as R, and Java programming language. The pipeline and other source codes used in this study can be found at: https://github.com/IngridHLab/HCsig and https://github.com/IngridHLab/CentriodClassification.