Human Molecular Genetics, 2022, Vol. 31, 15, 2483–2497
https://doi.org/10.1093/hmg/ddac031
Advance access publication date: 14 February 2022
Original Article

Genome-wide association study and functional follow-up identify 14q12 as a candidate risk locus for cervical cancer

Dhanya Ramachandran[1], Joe Dennis[2], Laura Fachal[2], Peter Schürmann[1], Kristine Bousset[1], Fabienne Hülse[1], Qianqian Mao[1], Yingying Wang[1], Matthias Jentschke[1], Gerd Böhmer[3], Hans-Georg Strauß[4], Christine Hirchenhain[5], Monika Schmidmayr[6], Florian Müller[7], Ingo Runnebaum[8], Alexander Hein[9], Frederik Stübs[9], Martin Koch[9], Matthias Ruebner[9], Matthias W. Beckmann[9], Peter A. Fasching[9], Alexander Luyten[10,11], Matthias Dürst[9], Peter Hillemanns[1], Douglas F. Easton[2,12] and Thilo Dörk[1,*]

[1]Department of Gynaecology, Comprehensive Cancer Center, Hannover Medical School, Hannover 30625, Germany
[2]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK
[3]IZD Hannover, Hannover, Germany
[4]Department of Gynaecology, University Clinics, Martin-Luther University, Halle-Wittenberg, Germany
[5]Department of Gynaecology, Clinics Carl Gustav Carus, University of Dresden, Dresden, Germany
[6]Department of Gynaecology, Technische Universität München, Munich, Germany
[7]Martin-Luther Hospital, Charite University, Berlin, Germany
[8]Department of Gynaecology, Jena University Hospital, Friedrich-Schiller-University Jena, Jena 07747, Germany
[9]Department of Gynaecology and Obstetrics, Comprehensive Cancer Center Erlangen-EMN, Erlangen University Hospital, Friedrich Alexander University of Erlangen–Nuremberg (FAU), Erlangen 91054, Germany
[10]Dysplasia Unit, Department of Gynaecology and Obstetrics, Mare Klinikum, Kronshagen, Germany
[11]Department of Gynaecology, Wolfsburg Hospital, Wolfsburg, Germany
[12]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge CB1 8RN, UK
[*]To whom correspondence should be addressed at: Gynaecology Research Unit (OE6411), Hannover Medical School, Carl-Neuberg-Str. 1, 30625 Hannover, Germany. Tel: +49 5115326075; Fax: +49 5115326081; Email: Doerk.Thilo@mh-hannover.de

Abstract

Cervical cancer is among the leading causes of cancer-related death in females worldwide. Infection by human papillomavirus (HPV) is an established risk factor for cancer development. However, genetic factors contributing to disease risk remain largely unknown. We report on a genome-wide association study (GWAS) on 375 German cervical cancer patients and 866 healthy controls, followed by a replication study comprising 658 patients with invasive cervical cancer, 1361 with cervical dysplasia and 841 healthy controls. Functional validation was performed for the top GWAS variant on chromosome 14q12 (rs225902, close to *PRKD1*). After bioinformatic annotation and *in silico* predictions, we performed transcript analysis in a cervical tissue series of 317 samples and demonstrate rs225902 as an expression quantitative trait locus (eQTL) for *FOXG1* and two tightly co-regulated long non-coding RNAs at this genomic region, *CTD-2251F13* (*lnc-PRKD1-1*) and *CTD-2503I6* (*lnc-FOXG1-6*). We also show allele-specific effects of the 14q12 variants via luciferase assays. We propose a combined effect of genotype, HPV status and gene expression at this locus on cervical cancer progression. Taken together, this work uncovers a potential candidate locus with regulatory functions and contributes to the understanding of genetic susceptibility to cervical cancer.

Introduction

Cervical cancer is the third most common cancer in females worldwide and is estimated to be the third leading cause of cancer death in females aged 15–44 years in Germany (1,2). Infection by high risk subtypes of human papillomavirus (HPV 16, 18, 33, 52 and 58, among others) is known to be necessary but not sufficient to cause cervical cancer development (3–7), and several environmental factors such as smoking and co-infections are known to increase risk. (1)

Family studies have demonstrated familial aggregation of cervical cancer consistent with a significant heritable component, with estimates of heritability ranging from 27 to 36% (8–11). Genome-wide association studies (GWASs) have so far identified consistent genomic regions associated with cervical cancer on chromosomes 6p21.32-33 (the human leukocyte antigen locus), 5p15.3 (*CLPTM1L*/*TERT*), 17q12 (*GSDMB*) and 2q14 (*PAX8*) (12–16) (reviewed in 17). These variants, however, may only explain a small proportion of cervical cancer heritability (18,19). Further GWASs and meta-analyses have proposed additional variants for cervical cancer risk (15,20–23), as well as for prognosis after neoadjuvant chemotherapy (24), but these have yet to be replicated.

The large majority of GWASs variants having regulatory function were discovered in the non-coding genome

Received: July 28, 2021. **Revised:** January 21, 2022. **Accepted:** January 24, 2022

(25–28). These may act as modulators of gene expression via disturbed transcription factor (TF) binding motifs, miRNA binding sites, long-range interactions, chromatin marks or long non-coding RNAs (lncRNAs) (25). The functional characterization of GWAS signals as molecular trait loci can help to interpret their role in the etiology of disease (29). However, the functional roles of most of the hitherto known cervical cancer risk loci have not yet been studied in detail.

Here, we performed a two-stage GWAS, in which the discovery stage included invasive cervical cancer (ICC) and cancer-free controls, followed by a larger replication stage that included cervical cancer and dysplasia cases versus healthy controls. Post GWAS, expression quantitative trait loci (eQTL) analysis in cervical tissue samples and luciferase analyses were also performed. We propose a novel locus on chromosome 14q12 as a potential contributor to cervical cancer risk.

Results

GWAS results

A total of 375 ICC and 866 female population controls were genotyped using the Illumina Oncoarray (30). The study design is depicted in Fig. 1A. After quality control (QC) measures and imputation, dosages for 363 cases and 861 controls were available for association analysis. There was little evidence for inflation in the association statistics (Fig. 1B, genomic inflation factor (λ) = 1.085). The strongest signal was for a variant on chromosome 14q12 (rs225902, $P = 1.1 \times 10E-7$, nearest gene *PRKD1*; Fig. 1C). This signal appeared robust when adjusted with different numbers of principal components (PCs; *P*-values between $6.4 \times 10E-7$ and $7.9 \times 10E-8$). A list of all variants at $P < 5 \times 10E-6$ from the GWAS summary statistics and details of wet-lab validation are provided in Supplementary Material, Table S1.

Replication genotyping and association analyses

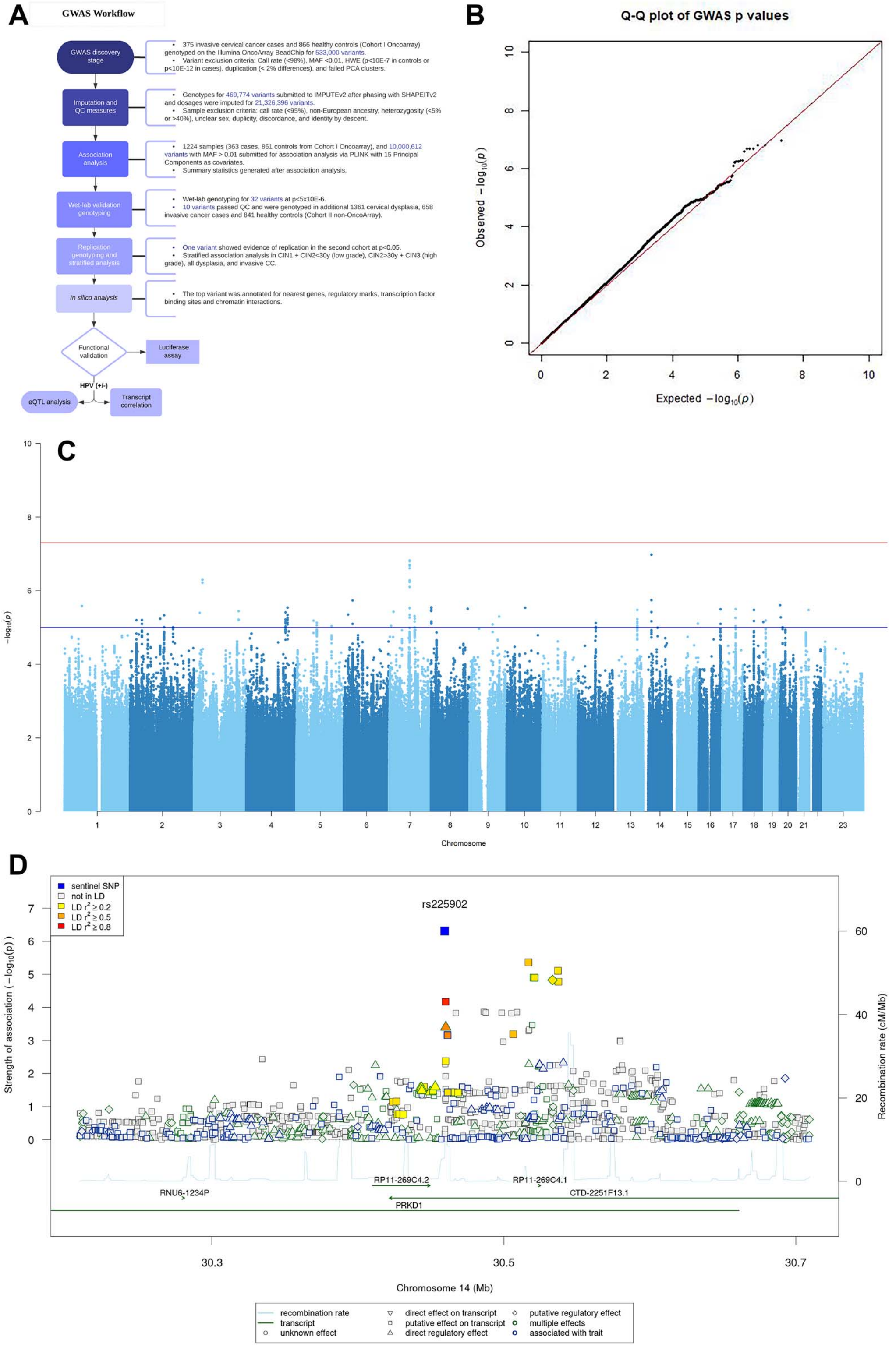
We validated the genotype results from the Oncoarray GWAS using mainly SNPtype assays for variants at $P < 5 \times 10E-6$, because most of the top signals had been imputed (Fig. 1A, see Materials and Methods for details) in 363 cases and 861 controls. Not all samples were successfully wet-lab genotyped for each variant; therefore, the total numbers genotyped vary from single nucleotide polymorphism (SNP) to SNP. After exclusion of imputed variants that failed this wet-lab validation through direct genotyping, 10 variants were taken for replication genotyping to the second stage, which included a further 2019 cervical dysplasia or invasive cancer cases and 841 healthy female control samples (see Materials and Methods). Only the top variant, rs225902, showed evidence of replication [odds ratio (OR) = 1.19, 95% confidence interval (CI) = 1.01–1.40, $P = 0.042$] in 362 cases and 854 controls in the second stage (Table 1). In the combined analysis of the SNPtype genotyping results, unadjusted with PCs, rs225902 showed evidence of association with ICC (OR = 1.33, 95%CI = 1.20–1.63, $P = 1.6 \times 10E-5$) and with overall cervical disease (OR = 1.35, 95%CI 1.18–1.55, $P_{adj} = 9.7 \times 10E-5$) (Table 1). Meta-analysis between Stages I (adjusted with PCs) and Stage II (unadjusted with PCs) for rs225902 under a fixed-effects model yielded similar results, showing a combined OR of 1.35 (95%CI 1.18–1.55, $P_{adj} = 9.66 \times 10E-5$) and indicating some modest heterogeneity between the discovery and the replication cohorts (I^2 = 0.85, $P = 0.01$) (Supplementary Material, Fig. S1), perhaps explainable by the lower number of invasive cases in the replication cohort. There was an indication of increasing OR with increasing severity of disease from low-grade dysplasia (CIN1+ $\text{CIN2}_{<30\text{years}}$) (OR = 1.15, 95%CI = 0.87–1.51, $P = 0.32$), to high-grade dysplasia ($\text{CIN2}_{\geq 30\text{years}}$ + CIN3) (OR = 1.26, 95%CI = 1.08–1.47, $P = 3.0 \times 10E-3$), to ICC (OR = 1.40, 95%CI = 1.20–1.63, $P = 1.4 \times 10E-5$) (Table 1). When cases were stratified for histology, variant rs225902 was similarly associated with adenocarcinomas (rs225902: OR = 1.62, 95%CI = 1.23–2.14, $P = 0.001$) and with squamous cell carcinoma (rs225902: OR = 1.35, 95%CI = 1.13–2.61, $P = 0.001$). The association appeared stable when adjusted for age (rs225902: OR = 1.33, 95%CI = 1.11–2.61, $P = 0.002$ for invasive cases; rs225902: OR = 1.25, 95%CI = 1.09–1.43, $P = 0.002$ overall).

A second, partially correlated variant rs225957 ($r^2 = 0.34$) did not replicate in the second stage (first stage OR 1.42, $P = 8.9 \times 10E-5$; second stage OR 1.09, $P = 0.199$) (Supplementary Material, Table S2). There was, however, some evidence of a stronger association with the haplotype consisting of the rare alleles of both rs225902 and rs225957 (OR = 1.39, 95%CI = 1.21–1.59, $P = 1.1 \times 10E-6$) (Table 2).

Bioinformatic annotation for top loci

We annotated genes ±2500 bp from rs225902 using SNiPA v3.4 (Fig. 1D). The genomic context of *PRKD1* and lncRNAs *CTD-2251F13* and *CTD-2503I6*, as well as the variants rs225902 and rs225957, was visualized using the Ensemble browser GrCh37 (Fig. 2A). We further mined two putative *PRKD1* promoter sequences, PRKD1_1 and PRKD1_2 (chr14: 30,397,035–30,397,094 and chr14: 30,396,824–30,396,883) (Fig. 2B), based on evidence from Eukaryotic promoter database (EPDnew) v006. A putative promoter covering both regions was then chosen for luciferase assays.

A circos plot showing chromatin interactions (−250/+500 bp of transcription start site) and eQTL information was generated for the top GWAS variant rs225902 using FUMA SNP2GENE, showing that this variant has chromatin interactions with and is an eQTL for *PRKD1* and lncRNA *CTD-2251F13* (in red) (Fig. 2C). Further annotation via HaploReg v4.2, SNPNexus and RegulomeDBv2 indicated several chromatin marks, eQTLs, deleteriousness and motifs changed for the variants of interest (rs225902 and rs225957) (Supplementary Material, Table S3).



- 375 invasive cervical cancer cases and 866 healthy controls (Cohort I Oncoarray) genotyped on the Illumina OncoArray BeadChip for 533,000 variants.
- Variant exclusion criteria: Call rate (<98%), MAF <0.01, HWE (p<10E-7 in controls or p<10E-12 in cases), duplication (< 2% differences), and failed PCA clusters.

- Genotypes for 469,774 variants submitted to IMPUTEv2 after phasing with SHAPEITv2 and dosages were imputed for 21,326,396 variants.
- Sample exclusion criteria: call rate (<95%), non-European ancestry, heterozygosity (<5% or >40%), unclear sex, duplicity, discordance, and identity by descent.

- 1224 samples (363 cases, 861 controls from Cohort I Oncoarray), and 10,000,612 variants with MAF > 0.01 submitted for association analysis via PLINK with 15 Principal Components as covariates.
- Summary statistics generated after association analysis.

- Wet-lab genotyping for 32 variants at p<5x10E-6.
- 10 variants passed QC and were genotyped in additional 1361 cervical dysplasia, 658 invasive cancer cases and 841 healthy controls (Cohort II non-OncoArray).

- One variant showed evidence of replication in the second cohort at p<0.05.
- Stratified association analysis in CIN1 + CIN2<30y (low grade), CIN2>30y + CIN3 (high grade), all dysplasia, and invasive CC.

- The top variant was annotated for nearest genes, regulatory marks, transcription factor binding sites and chromatin interactions.

Figure 1. GWAS workflow and outcomes. (**A**) Schematic diagram of the study design. (**B**) A quantile–quantile plot of the expected and observed −log10 *P* values on the *X*-axis and *Y*-axis, respectively. The red line indicates an ideal population with no underlying sub-structure. (**C**) Manhattan plot depicting −log10 *P* values after regression analysis on the *Y*-axis and chromosome and base pair position on the *X*-axis. A red line indicates genome-wide significance (GWS) at *P* = 5x10E-8 and the dark blue line is at *P* = 1 × 10E-5. Chromosomes are coloured alternatively in light and dark blue. (**D**) SNiPA regional association plot ±2500 bp near the variant rs225902 on 14q12.

Table 1. Stratified analysis of the top variant at chromosome 14, rs225902, after genotyping in the Cervigen cohort, in Stage I (Oncoarray) and Stage II (non-Oncoarray), and overall combined analysis of the two cohorts

Stratum	Stage I				Stage II				Combined			
	n_{Cases}	$n_{Controls}$	OR (95% CI)	P	n_{Cases}	$n_{Controls}$	OR (95% CI)	P	n_{Cases}	$n_{Controls}$	OR (95% CI)	P
Low-grade dysplasia	0	854	NA	NA	234	813	1.08 (0.82; 1.44)	0.579	234	1667	1.15 (0.87; 1.51)	0.317
High-grade dysplasia	0	854	NA	NA	1114	813	1.19 (0.99; 1.43)	0.059	1114	1667	1.26 (1.08; 1.47)	0.003
Invasive	362	854	1.75 (1.38; 2.22)	4.7 × 10E-6	654	813	1.21 (0.99; 1.47)	0.061	1016	1667	1.40 (1.20; 1.63)	1.4 × 10E-5
Invasive and high-grade dysplasia	362	854	1.75 (1.38; 2.22)	4.7 × 10E-6	1768	813	1.20 (1.02; 1.42)	0.033	2130	1667	1.33 (1.17; 1.51)	1.6 × 10E-5
All cases	362	854	1.75 (1.38; 2.22)	4.7 × 10E-6	2002	813	1.19 (1.01; 1.40)	0.042	2364	1667	1.31 (1.15; 1.49)	2.9 × 10E-5

Cervical intraepithelial neoplasia (CIN) was differentiated into low-grade (CIN1 + CIN2 < 30 years) and high-grade (CIN2 ≥ 30 years + CIN3) groups. ICC was further combined with high-grade dysplasia, followed by a joint analysis over all cases. Indicated are the number of cases (n_{Cases}), number of controls ($n_{Controls}$), OR with CI for the minor allele and P-values (P) generated after stratified logistic regression analyses restricted to the disease subtype.

TF binding changes predicted via MEME suite v5.3.3 tool TOMTOM, and ConSite showed that the rare allele of rs225902 (A) was predicted to alter a MYC binding site (ACCACATGGGA) (Fig. 2D). However, this position in the motif is not fully conserved, and chromatin immunoprecipitation experiments using MYC antibody for the region surrounding rs225902 in heterozygous HeLa cells, followed by Sanger sequencing, showed no significant allele specificity of MYC binding (Supplementary Material, Fig. S2), suggesting that other TFs may contribute to allele-specific effects (Supplementary Material, Table S3).

Promoter and enhancer studies via luciferase assay

The *PRKD1* putative promoter (chr14: 30,395,806–30,397, 607; hg19) was cloned into the pGL3 basic vector (Promega) and tested for promoter activity in HeLa cells via luciferase assays. The pGL3(SV40) promoter was strongly active in HeLa cells, as compared to the pGL3 basic construct ($P < 0.001$), and therefore taken as positive control for all experiments. The putative promoter tested for *PRKD1* was very weak in HeLa cells ($P = 0.3$) (Fig. 3A), so that the putative enhancer elements containing the 14q12 variants rs225902 and rs225957 were tested on the pGL3(SV40) promoter. The sequence containing the common allele ‘G/C’ of rs225902 showed an allele-specific repressive effect on the pGL3(SV40) promoter in HeLa cells ($P = 0.05$), and the repression was uplifted when the rare allele ‘A/T’ was present. Similarly, the rare allele ‘A/T’ of the linked variant rs225957 also showed an allele-specific removal of repression ($P = 0.02$). Combinations of elements containing these variants remained repressive ($P = 0.12$, 0.0002, <0.0001, 0.002). These findings linked the risk alleles at the 14q12 locus and their surrounding sequence elements with altered promoter activity in luciferase assays (Fig. 3B).

Gene expression, effect of HPV, correlation and eQTL analysis

For four candidate genes in the vicinity of the GWAS locus on chromosome 14q12 (*FOXG1, PRKD1* and the lncRNAs *CTD-2251F13* and *CTD-2503I6*), transcript analysis was performed in 317 cervical tissue samples. The levels of lncRNAs *CTD-2251F13* and *CTD-2503I6* were found to be highly correlated with each other (Pearson’s $R = 0.956$, $P = 2.06 \times 10E{-}143$, $N = 267$) and also correlated with transcript levels of *FOXG1*, while all three genes showed a weaker correlation with *PRKD1* at this locus (Fig. 4A). The correlation between the genes was independent of HPV status (Supplementary Material, Table S4a and S4b). The lncRNA transcript *CTD-2251F13* and *FOXG1* were upregulated in HPV+ tissue specimens ($P = 0.025$ and 0.013, respectively) (Fig. 4B). *PRKD1* expression was detected at low levels only in a minor proportion of the cervical tissues samples (105 out of 289) and in more HPV negative samples as compared to HPV positive samples ($P = 0.002$). This was consistent in HPV+ lesion+ samples versus HPV− lesion− samples ($P = 0.017$).

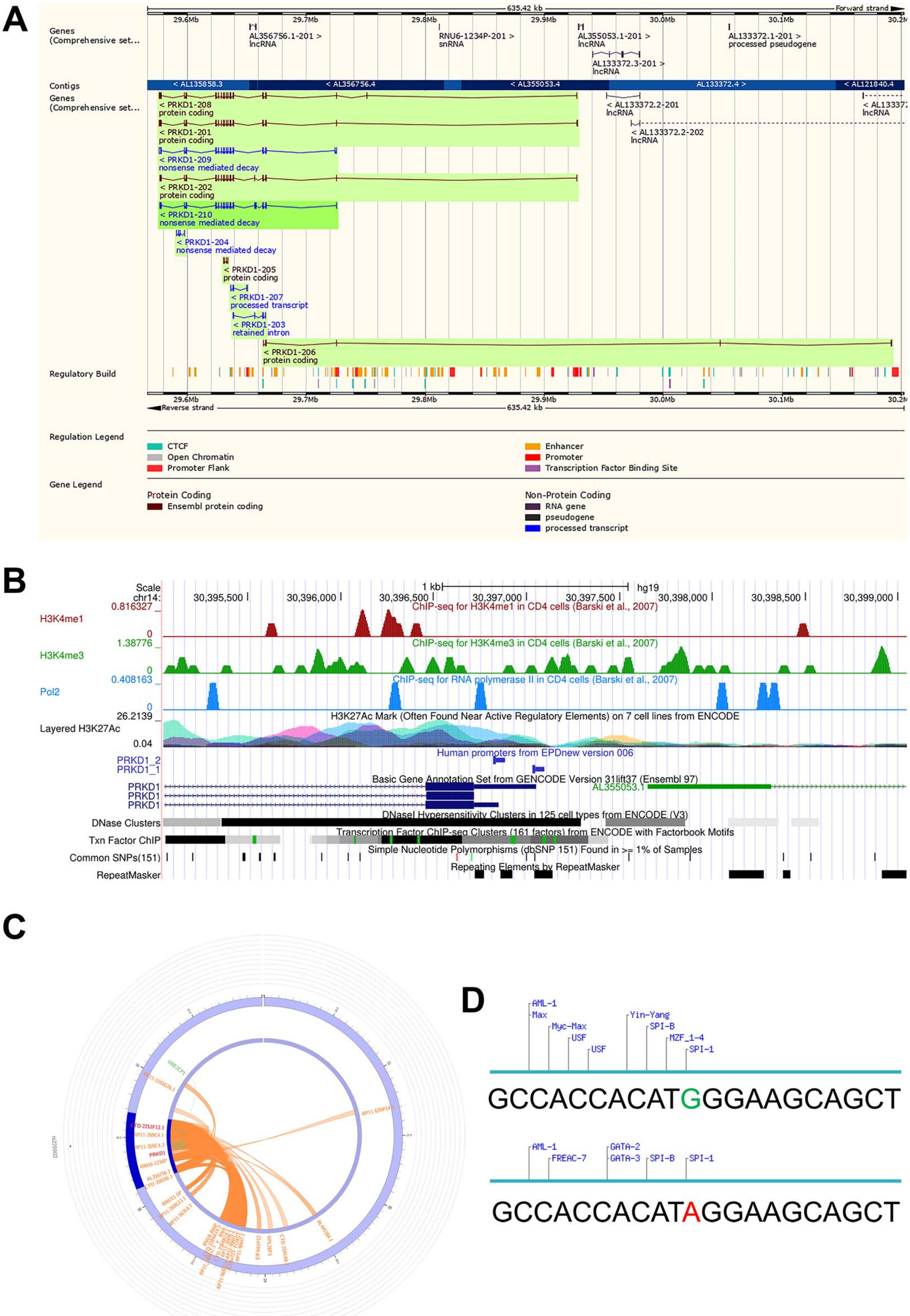

A
Genes (Comprehensive set...
635.42 kb
Forward strand
29.6Mb
29.7Mb
29.8Mb
29.9Mb
30.0Mb
30.1Mb
AL356756.1-201 > lncRNA
RNU6-1234P-201 > snRNA
AL355053.1-201 > lncRNA
AL133372.3-201 > lncRNA
AL133372.1-201 > processed pseudogene
Contigs
< AL135858.3
< AL356756.4
< AL355053.4
AL133372.4 >
< AL121840.4
< PRKD1-208 protein coding
< PRKD1-201 protein coding
< PRKD1-209 nonsense mediated decay
< PRKD1-202 protein coding
< PRKD1-210 nonsense mediated decay
< PRKD1-204 nonsense mediated decay
< PRKD1-205 protein coding
< PRKD1-207 processed transcript
< PRKD1-203 retained intron
< PRKD1-206 protein coding
< AL133372.2-201 lncRNA
< AL133372.2-202 lncRNA
Regulatory Build
Reverse strand
Regulation Legend
CTCF
Open Chromatin
Promoter Flank
Enhancer
Promoter
Transcription Factor Binding Site
Gene Legend
Protein Coding
Ensembl protein coding
Non-Protein Coding
RNA gene
pseudogene
processed transcript
B
Scale
chr14:
1 kb
hg19
30,395,500
30,396,000
30,396,500
30,397,000
30,397,500
30,398,000
30,398,500
30,399,000
H3K4me1
0.816327
ChIP-seq for H3K4me1 in CD4 cells (Barski et al., 2007)
H3K4me3
1.38776
ChIP-seq for H3K4me3 in CD4 cells (Barski et al., 2007)
Pol2
0.408163
ChIP-seq for RNA polymerase II in CD4 cells (Barski et al., 2007)
Layered H3K27Ac
26.2139
0.04
H3K27Ac Mark (Often Found Near Active Regulatory Elements) on 7 cell lines from ENCODE
Human promoters from EPDnew version 006
PRKD1_2
PRKD1_1
Basic Gene Annotation Set from GENCODE Version 31lift37 (Ensembl 97)
PRKD1
AL355053.1
DNase Clusters
DNaseI Hypersensitivity Clusters in 125 cell types from ENCODE (V3)
Txn Factor ChIP
Transcription Factor ChIP-seq Clusters (161 factors) from ENCODE with Factorbook Motifs
Common SNPs(151)
Simple Nucleotide Polymorphisms (dbSNP 151) Found in >= 1% of Samples
RepeatMasker
Repeating Elements by RepeatMasker
C
PRKD1
D
AML-1
Max
Myc-Max
USF
USF
Yin-Yang
SPI-B
MZF_1-4
SPI-1
GCCACCACATGGGAAGCAGCT
AML-1
FREAC-7
GATA-2
GATA-3
SPI-B
SPI-1
GCCACCACATAGGAAGCAGCT

Figure 2. Annotation of the top GWAS variant rs225902. (**A**) Ensembl browser GrCh37 snapshot of the chromosome 14 region surrounding the variant in detail, with transcript annotation and regulatory marks. (**B**) Screenshot of the Eukaryotic Promoter database (EPD) view at UCSC browser (hg19) showing putative *PRKD1* promoters (PRKD1_1 and PRKD1_2), along with regulatory marks in the region (H3K4Me1, H3K4me3, Pol2, H3K27Ac, DNaseI hypersensitivity sites, Transcription Factor ChIP-seq, Common SNPs v151 and repeat sequences). (**C**) Circos plot generated in FUMA shows various levels of information at rs225902 locus in chromosome 14. The outer layer is the GWAS *P* value for the SNP rs225902, the next layer shows the position in the genomic context with darker blue indicating identified risk loci, next are genes with known chromatin interactions with the variant in orange, known eQTLs in green, and genes having evidence for both interactions with the variant, in red. (**D**) CONSITE prediction of TF binding motifs ±20 bp near the variant rs225902 with the common allele ‘G’ and the rare allele ‘A’.

Table 2. Haplotype analysis of the variants rs225902 (C>T) and rs225957 (C>T) after genotyping in Stage I (Oncoarray) and Stage II (non-Oncoarray), as well as overall combined analysis

Haplotype	Stage I			Stage II			Overall		
	Frequency	OR (95% CI)	P	Frequency	OR (95% CI)	P	Frequency	OR (95% CI)	P
CC	0.622	0.64 (0.54; 0.77)	9.09×10^{-7}	0.651	0.94 (0.84; 1.06)	0.2861	0.642	0.89 (0.81; 0.98)	0.0114
CT	0.236	1.19 (0.97; 1.46)	0.086	0.201	0.96 (0.84; 1.10)	0.5724	0.212	0.96 (0.87; 1.07)	0.5188
TT	0.111	1.67 (1.29; 2.20)	8.22×10^{-5}	0.134	1.26 (1.06; 1.50)	0.0061	0.127	1.39 (1.21; 1.59)	1.14×10^{-6}
TC	0.031	1.72 (1.05; 2.79)	0.0172	0.013	0.67 (0.42; 1.09)	0.0807	0.019	0.85 (0.61; 1.17)	0.2893

Frequency of the genotype in the cohorts is indicated, OR with 95% CI and *P*-values (p) generated after Fisher's chi-square test.

In eQTL analysis in all cervical epithelial complementary DNA (cDNA) samples, rs225902 was found to be an eQTL for *FOXG1* and the two lncRNAs *CTD-2251F13* and *CTD-2503I6* ($P_{\text{ANOVA}} = 0.01$, 0.005, 0.03, respectively) (Fig. 4C); these associations remained significant in HPV negative lesion negative tissues (Supplementary Material, Fig. S3a). In HPV negative tissues, rs225902 also showed suggestive evidence of being an eQTL for the main *PRKD1* transcript ($P_{\text{T-test}} = 0.047$) (Fig. 4D). In haplotype analysis in HPV negative tissues, the rare homozygous genotype (TT) for the linked variants rs225902 and rs225957 was found to be an eQTL associated with increased levels for *FOXG1* ($P = 0.02$), *CTD-2251F13* ($P = 0.01$) and *CTD-2503I6* ($P = 0.01$) (Supplementary Material, Table S5).

lncRNA annotation, target prediction and transcript analysis

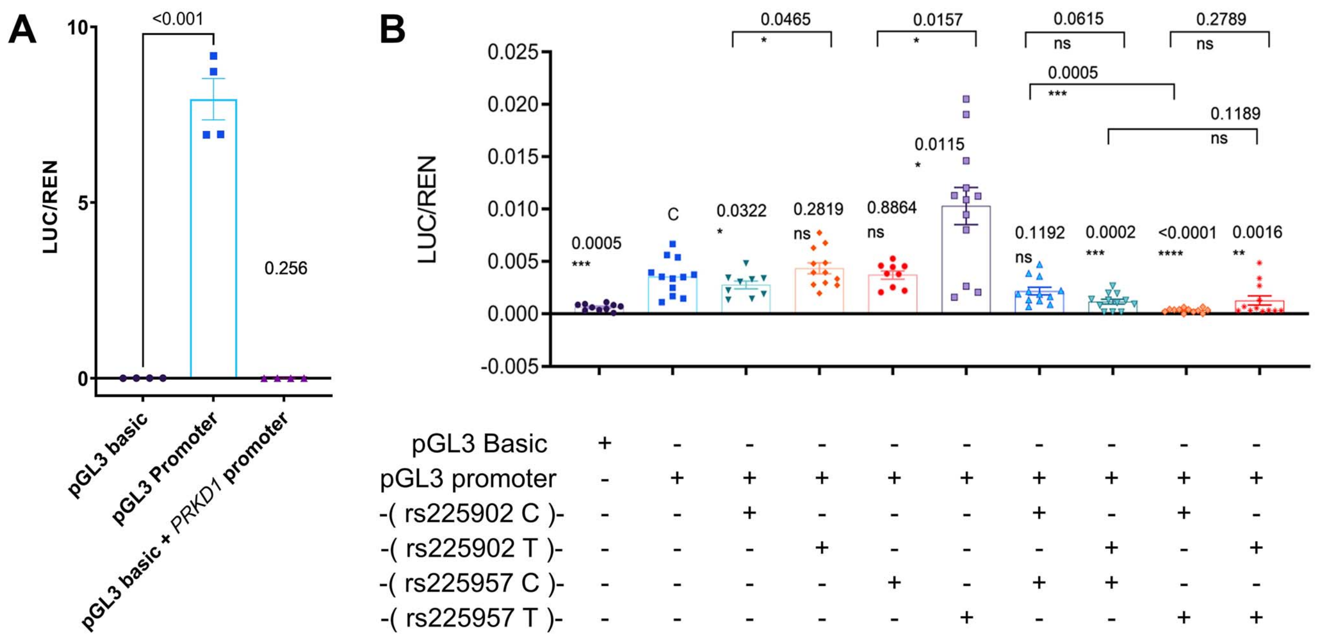
The lncRNA candidates *CTD-2251F13* and *CTD-2503I6* at 14q12 were annotated by NONCODE and LNCipedia (Table 3). The FASTA sequences of the three RNA isoforms encoded by *CTD-2251F13* and the single transcript of *CTD-2503I6* were submitted to CPC2, as well as the reverse sequences, in order to investigate their coding potential, and all transcripts were predicted to be non-coding (Table 4). Both lncRNAs were found to be abundant in the nucleus in HeLa cells by real time quantitative reverse transcription polymerase chain reaction (qRT-PCR) after cellular fractionation (Fig. 5A).

Further, cervical tissue samples were divided into two groups based on the upper and lower quartile expression values of either lncRNA, and gene expression was tested in HPV negative lesion negative tissues between the two groups. The levels of *PRKD1* were found to be higher in *CTD-2251F13*^{low} and *CTD-2503I6*^{low} samples ($P = 0.02$ and 0.05, respectively) (Fig. 5B). Conversely, *CTD-2251F13* and *CTD-2503I6* levels were found to be higher in *PRKD1*^{low} samples ($P = 0.04$ and 0.03, respectively) (Fig. 5C). Taken together, *CTD-2251F13, CTD-2503I6* and *FOXG1* showed a particularly strong correlation, tended to be inversely correlated with *PRKD1* levels, were specifically associated with the rs225902 genotype and were upregulated in HPV-positive tissues (Fig. 5D).

Discussion

While a significant genetic component for cervical cancer is predicted by family-based studies (8,9), only few genome-wide significant loci have been identified, thus far. Our present GWAS and follow-up studies have investigated candidate risk regions that were newly identified in a discovery set and then validated in a larger replication case–control series from the same population. Only one signal, rs225902 on chromosome 14q12, could be replicated at a marginal level of significance and remained associated with cervical cancer overall at $P \sim 10\text{E}{-5}$, with effect sizes increasing with the degree of disease severity. This variant had not emerged from recent GWAS analyses of the UK and Finnish Biobank studies ($P = 0.07$ in the UKBB and $P = 0.64$ in FinnGen [see UK Biobank (Cervical cancer GWAS with female controls only, https://github.com/Nealelab/UK_Biobank_GWAS file: 20001_1041.gwas.imputed_v3.female), and FinnGen freeze 5 (https://r5.finngen.fi/)]. However, the linked variant at chromosome 14q12, rs225957, showed mild evidence for replication in the cervical cancer GWAS from Rashkin *et al.* ($P = 0.047$, OR = 1.04, GWAS Catalog accession ID: GCST90011816) (14). Nevertheless, in regard of the sub-genome-wide significance and lack of replication in most other GWAS, it is premature to postulate that rs225902 or rs225957 are bona fide cervical cancer risk loci. Population stratification or differences in subtypes, given the small effect size in low-grade dysplasia, could partially explain this heterogeneity between studies. The proposal of rs225902 as a candidate genomic risk factor for cervical cancer therefore will need further replication.

Despite the lack of genome-wide significance, our subsequent functional studies provided evidence for this locus being implicated in cervical cancer. Luciferase reporter assays in HeLa cells indicated that the correlated variants rs225902 and rs225957 are located in two repressive elements and that the risk alleles specifically alleviate this repression. Our transcript analyses in cervical tissue specimens indicated a tightly correlated regulation of genes at the GWAS-identified risk locus. Among the genes on chromosome 14q12, rs225902 was a *cis*-eQTL for *FOXG1* encoding a transcriptional repressor that has already been implicated in cervical cancer (31). *PRKD1* encodes protein kinase D (formerly protein kinase Cμ), a known regulator of mitogenic pathways, histone deacetylation and invasiveness of cancer cells (32,33). The nuclear lncRNAs *CTD-2251F13* and *CTD-2503I6* (also termed *lnc-PRKD1-1* and *lnc-FOXG1-6*, respectively) themselves have not previously been associated with



pGL3 Basic	+	-	-	-	-	-	-	-	-	-
pGL3 promoter	-	+	+	+	+	+	+	+	+	+
-(rs225902 C)-	-	-	+	-	-	-	+	-	+	-
-(rs225902 T)-	-	-	-	+	-	-	-	+	-	+
-(rs225957 C)-	-	-	-	-	+	-	+	+	-	-
-(rs225957 T)-	-	-	-	-	-	+	-	-	+	+

Figure 3. Promoter and enhancer analysis via luciferase assays. (**A**) Comparison of the putative *PRKD1* promoter and pGL3 SV40 promoter to pGL3 basic construct [*P* values after two-sided *t*-test are indicated with pGL3 basic as control (denoted by 'C'), and error bars indicate ±standard error of the mean (SEM)], in HeLa cells. $N_{\text{Biological replicates}} = 2$; $N_{\text{Technical replicates}} = 2$. (**B**) Comparison of constructs containing enhancer sequences with ancestral (WT, rs225902 'G/C', rs225957 'G/C') and minor alleles (SNP, rs225902 'A/T', rs225957 'A/T') of rs225902 and rs225957 and their combinations to the pGL3 SV40 promoter in HeLa cells [two-sided *t*-test *P* values are indicated with pGL3 Promoter as control ('C') and also between selected bars indicated by a line, error bars indicate ±SEM]. $N_{\text{Biological replicates}} = 4$; $N_{\text{Technical replicates}} = 3$.

cancer. However, the lncRNA *CTD-2251F13* (*lnc-PRKD1-1*) appears to affect survival outcomes in the Cancer Genome Atlas cervical squamous cell carcinoma and endocervical adenocarcinoma (TCGA CESC) cohort, as visualized by TANRIC (log rank *P* value = 0.0047) (34). Additionally, the variants rs225902 and rs225957 have been associated with mucinous ovarian cancer (rs225902, *P* = 0.007, trait ID: ieu-a-1232) and cancers of the urinary tract or kidney (rs225957, *P* = 0.0002, trait ID: ukb-d-C_URINARY_TRACT and *P* = 0.0006, trait ID: ukb-b-1316, respectively) in MR-base (35). Given our evidence that the variants are eQTLs for *CTD-2251F13* and *CTD-2503I6*, the non-coding RNAs may be important for these cancers as well and thus add to the list of cancer-relevant lncRNAs (36).

In summary, we have proposed a candidate risk region for cervical cancer that has been identified from a novel GWAS and whose critical roles are supported by gene expression studies in cervical cancer cells. Further studies will be needed to reveal how this knowledge can be translated to improve cervical cancer risk prediction and management. It is likely that several further genetic risk variants exist that remain to be uncovered in future GWAS approaches for cervical cancer, a disease that results from viral and host genome interactions.

Materials and Methods

Patient material

The German Cervigen case–control series has previously been described in candidate gene association studies (37,38). Nine hospitals in Hannover, Wolfsburg, Jena, Erlangen, Dresden, Halle, Munich, Berlin and Bad Münder contributed 5 ml peripheral venous ethylenediamine tetraacetic acid (EDTA) blood from 1033 cases with ICC and 1361 cases with cervical dysplasia. At the same time, a total of 1707 healthy female controls were provided by the clinics in Hannover and Erlangen. In the current study, a total of 4101 samples were genotyped. The HPV status of patients in this study has been described previously (37,38). Genomic DNA was extracted from the blood samples via a standard phenol–chloroform extraction method.

A smaller cohort of 317 healthy participants (without invasive cancer) contributed methanol-fixed cervical tissue smears, from which RNA was extracted and a cDNA biobank was established. Genomic DNA was extracted from these samples as well. HPV positivity and lesion status for this smaller cohort have been described elsewhere (37,38).

SNP calling, quality control, imputation and PLINK analysis

For the GWAS (Stage I), DNA from 375 females with ICC and 866 healthy controls was selected, for genotyping on the Illumina OncoArray BeadChip. This array includes ~533 000 variants, of which approximately half were selected as a GWAS backbone so that the large majority of common variants are correlated with a backbone variant; the remaining variants were selected to cover regions associated with other cancers in more detail, and other variants of relevance to cancer or cancer-associated phenotypes, as described elsewhere (30).

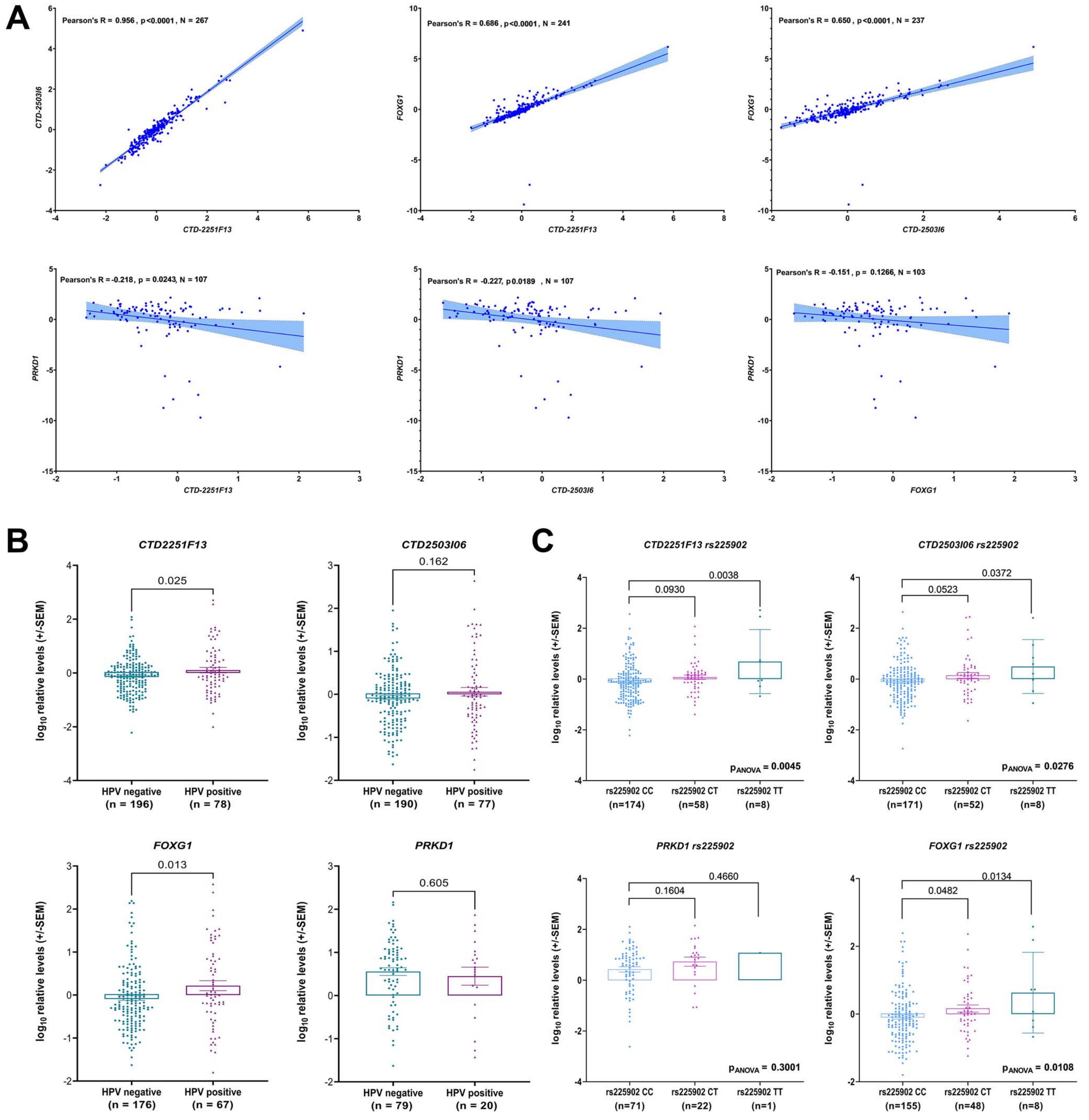


Figure 4. Gene transcript correlations, effect of HPV and eQTL analysis in cervical epithelial tissue cohort. (**A**) Correlation of transcript levels of *CTD2251F13* and *CTD2503I6*, *FOXG1* and *PRKD1* [indicated are Pearson *R* values, *P* values and number of samples (*n*)]. (**B**) Log10 relative mean levels (±SEM) of various transcripts at the chromosome 14 locus were associated with the HPV status of samples (*P* values after t-test between HPV positive and negative groups). (**C**) Log10-transformed relative mean levels (±SEM) of transcripts, in all samples were tested for association with the genotype of the variant rs225902 under the allelic model of inheritance shows *cis*-eQTLs. (**D**) Log10-transformed relative mean levels (±SEM) of transcripts in samples without HPV infection were tested for association with the genotype of the variant rs225902 under the allelic model of inheritance shows *cis*-eQTLs. *P* values shown are from two-sided *t*-test between groups, unless otherwise indicated as *P* values after ANOVA (for 3 groups).

Variants were filtered out by low call rate [<95% for common SNPs; <98% for SNPs with minor allele frequency (MAF) <0.01], Hardy–Weinberg equilibrium (HWE) testing ($P < 10^{-7}$ in controls or $P < 10^{-12}$ in cases), comparison of duplicate samples (>2% differences) and failed cluster plots. (30,38) SNPs with a different genotype MAF from the 1000 Genomes Project (1KG/1000G) or those not linked to 1KG SNPs were excluded, and genotypes for 469 774 SNPs were submitted for imputation to IMPUTEv2 after phasing with SHAPEITv2. Imputed dosages were obtained for 21 326 396 variants. The numbers of SNP exclusions after each step of filtering are detailed in Supplementary Material, Table S6.

We followed the same pipeline for QC as described previously for the OncoArray (30,39). Samples were excluded based on low genotype call rate (<95%), non-European

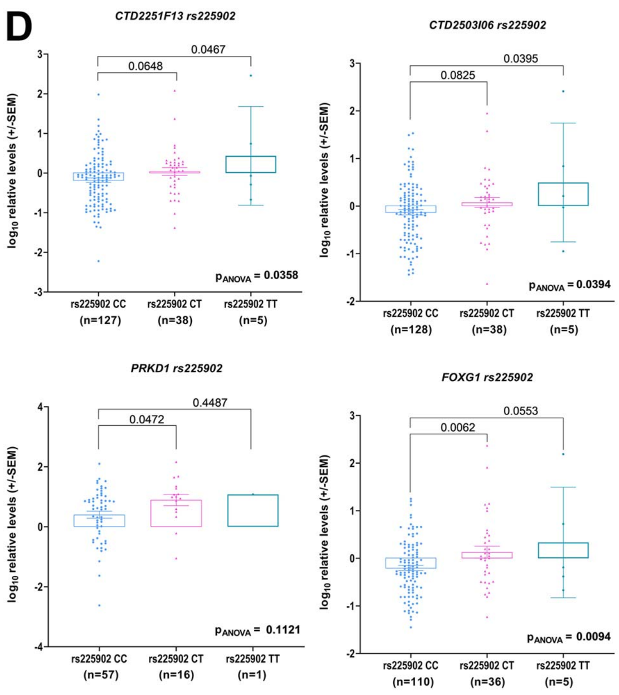


Figure 4. Continued.

ancestry based on PC analysis, excess heterozygosity (<5% or >40%) or sex chromosome anomalies. After QC, 1224 samples remained (363 cases, 861 controls). The dosages were analysed for association with cervical cancer via PLINK v1.90b4.9 64-bit (13 October 2017) using the first 15 ancestry informative PCs, derived by the R-package FastPop (http://sourceforge.net/projects/fastpop/), as covariates in order to reduce false positives arising from underlying ancestry-related inflation. (40–42)

The observed −log10 P values from the GWAS summary statistics (10 000 612 variants with non-NA P-values, imputation $r^2 > 0.3$ and MAF > 0.01) were plotted in a quantile–quantile plot after regression analysis, and the genomic inflation factor (GIF λ) was calculated to check for population sub-structure (43). Manhattan plots were used to visualize the GWAS summary statistics, and regional association plots were generated using SNiPAv4.3 for loci of interest, to visualize the nearest genes and underlying linkage disequilibrium (LD) with variants in the vicinity. The plots were generated in R v3.4 and v4.0 (package: qqman). References and links to all software and databases used are listed in Supplementary Material, File S1.

SNP genotyping

The 1224 samples (363 cases, 861 controls) that had been genotyped on the GWAS Oncoarray (30) (Stage I) were again genotyped for variants at $P < 5 \times 10E{-}6$, in order to validate the results of the Illumina genotyping (Fig. 1A). This wet-lab validation was deemed necessary because most of the top signals had been imputed. When several variants were found at the same region, LDlink was used to determine LD and the variant with the lowest P-value was taken. These variants were then selected for genotyping via Fluidigm SNPType assays (PN 68000098Q1), restriction fragment length polymorphism or TaqMan assays (Thermo Fisher Scientific, USA, MAN0009593) (Supplementary Material, Table S1). When the assay designs failed for a SNP, a proxy within two orders of magnitude was drawn from the GWAS data, and LDproxy (as part of LDlink) was used to confirm the correlation ($R^2 \geq 0.3$) with the substitute in 1000G EUR data.

Allele-specific probes were labelled with FAM® or HEX® dyes in case of Fluidigm SNPType assays and with FAM® and VIC® in TaqMan Assays. Two samples without template served as negative controls. Samples failing in more than 25% of assays were removed from further analysis. At first, the top variants at $P < 5 \times 10E{-}6$ were validated by genotyping the samples that were on the Oncoarray (cohort I, $n_{\text{cases}} = 363$, $n_{\text{controls}} = 861$, after QC) and passed HWE testing. The variants, which were concordant and validated in this first cohort, were additionally genotyped in the remaining samples of the Cervigen cohort (cohort II, non-Oncoarray, $n = 2860$) (Fig. 1A). By contrast with the first stage which included only invasive cases and controls, this second cohort comprised 1361 cervical dysplasia cases (130 CIN1, 221 CIN2, 1010 CIN3), 658 cases with ICC and 841 healthy female controls.

After genotyping in the second cohort, logistic regression analysis was performed in STATA12. ORs, P-values and 95% CIs under an additive model were generated from cohort I (Oncoarray, adjusted with PCs) and cohort II (non-Oncoarray, unadjusted), as well as after combined analysis for association with overall cervical disease. The combined analysis was done on the wet-lab genotypes obtained with the same platform (usually Fluidigm Biomark). Stratified analysis was performed in STATA12 with case–control status as the outcome variable and genotype as the predictor variable. For this, the cases were stratified into low-grade dysplasia, high-grade dysplasia or invasive cases, and the invasive cases were stratified into the main histological subgroups squamous cell carcinoma or adenocarcinoma. Genotype distributions of cases within each stratum were compared with the genotype distribution among all controls. The generated Z scores were transformed into P values. A meta-analysis was also performed between stage I and II samples using STATA12, taking into account the adjustment with PCs in the first stage, and Forest plots were generated via MetaGenyo.

Samples was stratified by cancer severity, into low-grade dysplasia [CIN1 + CIN2 cases at age <30 years ($\text{CIN2}_{<30}$)], high-grade dysplasia [CIN2 cases at age ≥30 years ($\text{CIN2}_{\geq 30}$) and CIN3] and ICC. A combined analysis was also performed for CIN2 at age ≥30, CIN3 and invasive cases.

Two-sided P-values below 0.05 were considered confirmatory evidence in the second stage and indicative for

Table 3. Annotation and features of the lncRNA *CTD-2251F13* and *CTD-2503I6* via NONCODE and LNCipedia

Feature	*CTD-2251F13*	*CTD-2503I6*
LNCipedia transcript ID	lnc-PRKD1-1:4	lnc-FOXG1-6:17
LNCipedia gene ID	lnc-PRKD1-1	lnc-FOXG1-6
Ensembl Gene ID	ENSG00000248975	ENSG00000257120
Ensembl Transcript ID	ENST00000549360	ENST00000550941
Gene Location (GrCh37)	chr14: 30421603-30 766 249	chr14: 30,122,015-30,127,122
Strand	Minus	Plus
Class	Intronic	Antisense
Sequence Ontology term	Sense intronic ncRNA	Antisense lncRNA
Transcript size	768 bp	551 bp
Exons	3	2
Number of transcripts	3	1

All transcripts	AL133372.2-201 (ENST00000548124.1—922 bp);
	AL133372.2-202 (ENST00000549360.1—768 bp);
	AL133372.2-203 (ENST00000508469.2—233 bp);
		AL356756.1 (ENST00000550941.1);

Alternative gene names	ENSG00000248975; CTD-2251F13.1; OTTHUMG00000170491.1; AL133372.2	ENSG00000257120.1; CTD-2503I6.1; OTTHUMG00000170489.1; AL356756.1
Alternative transcript names	ENST00000549360.1; CTD-2251F13.1-002; OTTHUMT00000409381.1; NONHSAT036216	ENST00000550941.1; CTD-2503I6.1-001; OTTHUMT00000409378.1; NONHSAT036209

Table 4. Sequence-based (non-) coding potential of the lncRNA transcripts calculated by CPC2

Sequence ID	Strand	Label	Coding probability	Peptide length (aa)	Fickett score	Isoelectric point	ORF integrity
CTD-2251F13.1-201	Forward	Noncoding	0.00590617	21	0.41452	0.41452	Complete
CTD-2251F13.1-201	Reverse	Noncoding	0.00800117	12	0.4106	0.4106	Complete
CTD-2251F13.1-202	Forward	Noncoding	0.133846	101	0.34133	0.34133	Incomplete
CTD-2251F13.1-202	Reverse	Noncoding	0.0627548	75	0.34564	0.34564	Complete
CTD-2251F13.1-203	Forward	Noncoding	0.0802353	76	0.36306	0.36306	Complete
CTD-2251F13.1-203	Reverse	Noncoding	0.00337841	18	0.37955	0.37955	Complete
CTD-2503I06.1	Forward	Noncoding	0.0659486	60	0.42321	0.42321	Complete
CTD-2503I06.1	Reverse	Noncoding	0.0237596	43	0.41716	0.41716	Complete

Shown are the label, coding probability, putative peptide length, Fickett Testcode score, putative isoelectric point and ORF integrity.

functional follow-up, and two-sided *P*-values below 0.005 were considered significant in the sub-group analyses. References and links to all software and databases used are listed in Supplementary Material, File S1.

Bioinformatic annotation of top loci

The top loci from the GWAS were analysed in further detail with an array of bioinformatic tools. Variation Viewer, Ensembl browser and UCSC Genome Browser were used to visualize each SNP in the genomic context. For variant annotation, SNPNexus provided comprehensive information with UCSC and RefSeq annotations for adjacent genes, coding/non-coding status, MAF (based on 1000G EUR), changes in chromatin marks (repression, activation, modulation) in cell lines, implication in previous GWAS or other diseases or traits, and evidence for deleteriousness (CADD score) (Supplementary Material, Tables S3 and S8).

For variants of interest, a comprehensive circos plot was created within FUMA, integrating GWAS *P*-value, putative locus, and chromatin and eQTL interactions. Ten bases up and downstream of the variant were submitted to three webtools to predict and score allele-specific TF binding: ConSite, TOMTOM (MEMESuite v5.3.3) which was mapped against the HOCOMOCOv11_full_HUMAN database of motifs, and SNP select from SNP2TFBS. The two sequences containing the wildtype or rare allele were compared and changes to the binding of TFs were noted.

Publicly available data sources were queried via webtools such as PanCanQTL, and HaploRegv4.2 for evidence for eQTL and FUMA for known chromatin interactions. A list of rsIDs of the top variants was submitted to HaploRegv4.2, and RegulomeDBv2 to identify regulatory features on or in the vicinity of the SNP. HaploRegv4.2 also provided a comprehensive list of linked SNP and evidences for motifs altered, adjacent gene, as well as whether the SNP has activating or repressing marks in any cells (Supplementary Material, Table S3). Chromatin (enhancer/repressor/promotor) marks provided evidence for designing luciferase assays (Supplementary Material, Table S3). References and links to all software and databases used are listed in Supplementary Material, File S1.

Transcript analysis

RNA was extracted from the cervical epithelial tissue cohort and reverse transcribed into cDNA as described previously. (37,38) Custom primer-probe assays (IDT) and Fluidigm® DeltaGene assays were designed for genes of

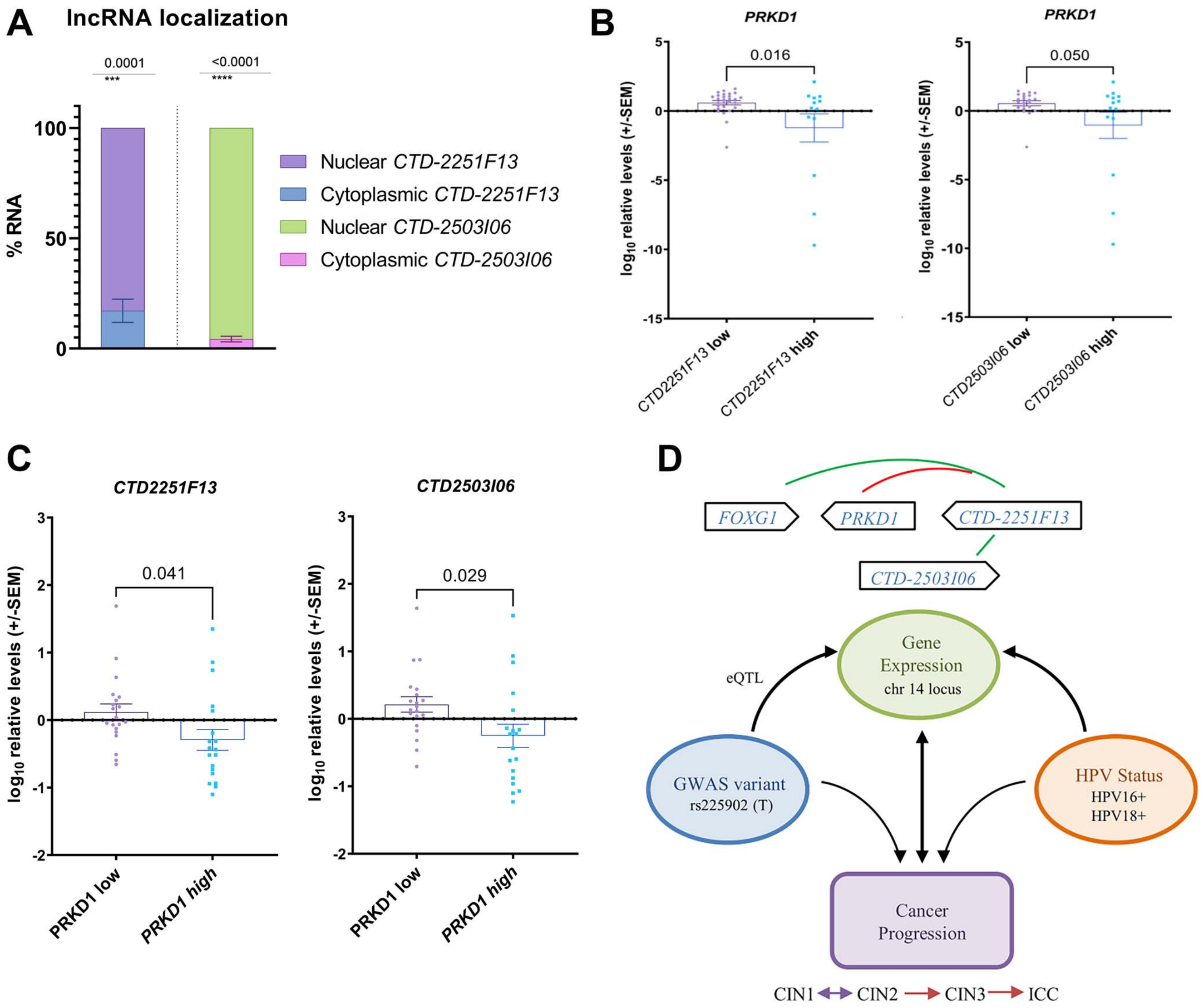


Figure 5. LncRNA localization and expression studies. (**A**) Localization of the lncRNAs (*CTD-2251F13* on the left and *CTD-2503I6* on the right) in cellular compartments of HeLa cells quantified by qRT-PCR after sub-cellular fractionation and RNA isolation. *P* values indicate *t* tests between the nuclear and cytoplasmic abundance. (**B**) Effect of high versus low expression quartiles of lncRNA *CTD-2251F13* and *CTD-2503I6* on *PRKD1* transcript levels. *P* values shown are after *t* test between groups, error bars ±SEM. (**C**) Effect of high versus low expression quartiles of *PRKD1* on transcript levels of *CTD-2251F13* and *CTD-2503I6*. *P* values shown are after *t* test between groups, error bars ±SEM. (**D**) Genes at chromosome 14 are highly correlated or are influenced by HPV status or rs225902 genotype (eQTL). Combined effect of genotype, HPV and gene expression at the chromosome 14 locus contributes to CC progression.

interest and housekeeping genes (Supplementary Material, Table S7a and S7b). qRT-PCR reactions were performed in technical duplicates on a BioMark HD real-time PCR instrument (Fluidigm). qBASE+ (Biogazelle, Belgium) and GeNorm were used to check the stability of housekeeping genes and calculate relative gene quantities taking *B2M* and *RPL13A* as housekeeping controls.

Outliers were identified and excluded with the ROUT method (1% FDR) on GraphPad Prism v9.0. Association of gene expression and HPV status was tested. SNP–gene eQTL combinations were tested for association between relative gene levels in cDNA and genotype in the corresponding gDNA sample under the allelic model of inheritance. Pairwise Pearson correlation coefficients (*R*) were calculated between genes overall, and after stratification based on HPV and lesion status. Two-sided *t*-test was performed to compare two groups, whereas ANOVA was performed to compare three or more groups. Multiple testing correction was applied to identify noteworthy *P* value thresholds.

Cell culture

The spontaneously immortalized human cervical epithelial adenocarcinoma HeLa cells (ATCC® CCL-2™, certified by the European collection of cell cultures (ECACC) to be mycoplasma free) were cultured in DMEM-high glucose medium (Gibco, US), supplemented with 10% FBS, Penicillin (1 U/ml medium) and Streptomycin (0.1 mg/ml medium) (Biowest, France). Cells were washed with phosphate-buffered saline (1X PBS), trypsinized with Trypsin/EDTA solution (Biochrom, UK) and counted in a Neubauer chamber after adding Trypan blue solution to count live cells.

Chromatin immunoprecipitation

3×10^6 HeLa cells (heterozygous for the SNP rs225902) were harvested and crosslinked with 1% formaldehyde for 10 min and neutralized with 2.5 M glycine for 10 min at room temperature. The lysate was then sonicated using the s220 Covaris system (peak power 175 W, duty factor 10%, cycles per burst 200, temperature 4°C, and water level 12, for 240 s) to obtain DNA fragments between 100 and 300 bp. Chromatin immunoprecipitation experiments were conducted using the MAGnify™ Chromatin Immunoprecipitation System (Thermo Scientific, PN# MAN0001631), as per manufacturer's instructions. C-Myc antibody (CST #9402) and Rabbit IgG (Thermo Scientific) were used for immunoprecipitation, followed by precipitation and purification of DNA. The region of interest was amplified by PCR (Primers: 5′ GATGGACCAGACTCCCAATC 3′ and 5′ GCTGAGCAGAGTTCTCACTAG 3′) and sequenced using BigDye® Sanger sequencing on a SeqStudio Genetic Analyzer (Applied Biosystems, US). The sequencing chromatograms were visualized using CodonCode aligner software and peaks were quantitatively analysed by QSVanalyzer. The experiments were performed in seven biological replicates. Statistical analysis was performed in GraphPad Prism v9.0. References and links to all software and databases used are listed in Supplementary Material, File S1.

Luciferase construct designs/plasmids

Putative promoters of *PRKD1* (chr14: 30,395,806-30,397,607; hg19) were identified via the eukaryotic promoter database (EPDnew v006). A composite promoter spanning two suggested promoters in EPD was designed. The promoter sequences were synthesized and cloned into the pGL3 Basic vector (Promega, US, E1751) by GenScript (Supplementary Material, Table S9).

For the variant rs225902 (C/T), 784 bases upstream and 217 bases downstream were cloned as the putative enhancer/repressor element downstream of the SV40 promoter driven *luciferin* gene into the pGL3 Promoter vector (Promega, E1761), on the basis of experimental evidence for regulatory activity. For the second variant on chromosome 14, rs225957 (C/T), 500 bases upstream and downstream of the variant were synthesized and cloned downstream of the pGL3 promoter-driven *luciferin* gene by GenScript.

Combinations of the DNA elements containing these two variants were also synthesized and cloned into the pGL3 Promoter vector to create four further constructs (rs225902C + rs225957C, rs225902C + rs225957T, rs225902T + rs225957C and rs225902T + rs225957T).

The synthesized plasmids were grown in high efficiency 10-beta Competent *Escherichia coli* (New England Biolabs) and isolated via endotoxin-free plasmid DNA purification kit NucleoBond® Xtra Midi EF according to the manufacturer's instructions (Macherey-Nagel, Germany, Protocol-at-a-glance Rev.06). The plasmid concentrations were determined by NanoDrop8000 spectrophotometer (Thermo Scientific) and the constructs were confirmed by restriction digestion and agarose gel electrophoresis.

HeLa cells were transfected in technical triplicates in 24 well plates by Lipofectamine3000 (Invitrogen, MAN0009872) with 300 ng of plasmid of interest, along with 50 ng pRLTK plasmid (Promega E2241) containing *Renilla* luciferase, and pUC19 (Plasmid of University of California 19, New England Biolabs, US, N3041) for molar equivalency (44). Cells were grown in Opti-MEM reduced serum medium (Gibco) for 24 h and then Dual-Glo luciferase assay system (Promega, E2920) and the Glomax plate reader (Promega) were employed as per manufacturer's instructions for luminescence detection. Each experiment was performed at least three times with appropriate controls, and after background subtraction, GraphPad Prism v9.0 was used for statistical analysis.

LncRNA analysis and localization

The lncRNAs at chromosome 14 (*CTD-2251F13* and *CTD-2503I6*) were annotated by LNCipedia and the transcripts were visualized in Ensembl browser. The forward and reverse FASTA sequences of the three transcripts of *CTD-2251F13* (*201*, *202 and 203*) and the single transcript of *CTD-2503I6* were submitted to CPC2 for calculating the coding potential of these transcripts.

RNA fractionation into cytoplasmic and nuclear fractions was performed as described previously, in order to localize the lncRNA in HeLa cells (45). Trizol (PeqGold TriFast™, VWR, US) was added into each of the fractions in diethyl pyrocarbonate (DEPC)-treated tubes and placed on ice. 200 μl chloroform was added, the mix was incubated for 5 min on ice and centrifuged for 15 min at 16100*g* at 4°C. The upper aqueous phase was mixed with an equal volume of isopropanol and incubated for 10 min on ice and centrifuged for 15 min at 16100*g* at 4°C. The pellet was washed twice with 75% uvasol, mixed and centrifuged as before. The pellet was dissolved in sterile DEPC water and total RNA concentration was measured by a Nanodrop 8000 spectrophotometer. One microgram of RNA was reverse transcribed into cDNA as previously described. lncRNA transcript levels were measured by qRT-PCR assays (IDT, USA, Supplementary Material, Table S7a) on a CFX384 thermocycler (BioRad, US) or Rotor-Gene 6000 real-time PCR machine (Qiagen, Germany) and normalized to the housekeeping genes (*B2M* and *RPL13A*). The log2 normalized values were statistically tested using GraphPad Prism v9.0. The experiments were performed at least three times (biological replicates) and the qRT-PCR was performed with at least two technical replicates. References and links to all software and databases used are listed in Supplementary Material, File S1.

Relative gene quantities and Pearson's correlation coefficients were calculated for the candidate genes in the cervical epithelial tissues as mentioned above, and statistical analysis was carried out in GraphPad Prism v9.0. The transcript levels were tested between HPV positive and negative tissues and also as eQTLs with the variant rs225902. In specific analysis, expression of the *PRKD1* was compared between samples in the upper and lower quartiles of expression of either lncRNA by *t*-test (two-sided). Similarly, the levels of both lncRNAs were compared between samples belonging to the lower and upper quartiles of *PRKD1* expression levels. References and links to all software and databases used are listed in Supplementary Material, File S1.

Statistical analysis

Statistical software and methods used for association analysis, logistic regression, stratified analysis and haplotype testing are described under the relevant subsections. In transcript analysis, eQTL and luciferase experiments, outliers were excluded with the ROUT method (1% FDR) on GraphPad Prism v9.0. Pairwise Pearson correlation coefficients (*R*) were calculated and reported with *P* values (*P*) and number of data points (*n*) indicated. Biological and technical replicates are indicated in figure legends. Two-sided *t*-test was performed to compare two groups, whereas ANOVA was performed to compare three or more groups. Multiple testing correction was used to define noteworthy *P* value thresholds.

Supplementary Material

Supplementary Material is available at *HMG* online.

Acknowledgements

We cordially thank all the patients, volunteers and clinicians who participated in sampling and documentation. We specifically thank Ivonne Bünting, Britta Wieland and Natalia Bogdanova for their technical support; Rainer Niedenthal for access to the Glomax plate reader; Colin Davenport for maintaining the MHH high speed computing cluster; Eda Celik for her assistance in documentation; Anja Thümmler, Tina Ruider and the German Colposcopy Study Group for their support in patient recruitment and Juliane Hoyer, Arif Ekici and Julia Enßen for their support in DNA extraction. We are grateful to Thomas Schulz and Michael Heuser for helpful discussion. We would like to acknowledge the OncoArray consortium and Genome Quebec for genome-wide genotyping and variant calling.

Conflict of Interest statement. Peter A. Fasching received grants from BioNTech and Cepheid, and personal fees from Novartis, Roche, Pfizer, Daiichi-Sankyo, Eisai, AstraZeneca, Merck Sharp & Dohme, Pierre Fabre, SeaGen, Hexal, Agenida and Lilly during the conduct of our study. This funding had no connection to the study presented here and did not influence our study design and results. None of the other author have disclosures. The authors declare that there are no conflicts of interest in the writing, preparing, or finalizing of our study.

Ethics declarations

This study was approved by the Ethics committee of Hannover Medical School (Vote No. 441) and the samples as well as data used were in accordance with German medical council regulations.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request. The GWAS summary statistics will be uploaded to the EBI GWAS catalogue.

Code availability

All the software, databases and webtools used in this study are detailed in Supplementary Material, File S1.

Funding

Bruno and Helene Jöster Foundation; Hannover Biomedical Research School (HBRS) and the MD/PhD program ‘Molecular Medicine’ (to D.R.); a research stipend from the Shanghai New Medical Union Hospital (to Q.M.); a research stipend from the Xuzhou Medical University (to Y.W.).

References

1. Bruni, L., Albero, G., Serrano, B., Mena, M., Collado, J. J., Gómez, D., Muñoz, J., Bosch, F. X., de Sanjosé, S. *ICO/IARC Information Centre on HPV and Cancer (HPV Information Centre)*. Human Papillomavirus and Related Diseases in the World. Summary Report 22 October 2021. https://hpvcentre.net/statistics/reports/XWX.pdf (Accessed on 7 December 2021).
2. Bray, F., Ferlay, J., Soerjomataram, I., Siegel, R.L., Torre, L.A. and Jemal, A. (2018) Global cancer statistics 2018: GLOBOCAN estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.*, **68**, 394–424.
3. Bonde, J., Bottari, F., Parvu, V., Pedersen, H., Yanson, K., Iacobone, A.D., Kodsi, S., Landoni, F., Vaughan, L., Ejegod, D.M. and Sandri, M.T. (2019) Bayesian analysis of baseline risk of CIN2 and ≥CIN3 by HPV genotype in a European referral cohort. *Int. J. Cancer*, **145**, 1033–1041.
4. The Cancer Genome Atlas Research Network (2017) Integrated genomic and molecular characterization of cervical cancer. *Nature*, **543**, 378–384.
5. Tjalma, W. (2018) HPV negative cervical cancers and primary HPV screening. *Facts Views Vis Obgyn*, **10**, 107–113.
6. Kaliff, M., Karlsson, M.G., Sorbe, B., Bohr Mordhorst, L., Helenius, G. and Lillsunde-Larsson, G. (2019) HPV-negative Tumors in a Swedish Cohort of Cervical Cancer. *Int. J. Gynecol. Pathol.*, **39**, 279–288.
7. Ruiz, F.J., Sundaresan, A., Zhang, J., Pedamallu, C.S., Halle, M.K., Srinivasasainagendra, V., Zhang, J., Muhammad, N., Stanley, J., Markovina, S. *et al.* (2021) Genomic characterization and therapeutic targeting of HPV undetected cervical carcinomas. *Cancers*, **13**, 4551.

8. Magnusson, P.K.E., Lichtenstein, P. and Gyllensten, U.B. (2000) Heritability of cervical tumours. *Int. J. Cancer*, **88**, 698–701.
9. Zoodsma, M., Sijmons, R.H., de Vries, E.G. and Zee, A.G. (2004) Familial cervical cancer: case reports, review and clinical implications. *Hered. Cancer Clin. Pract.*, **2**, 99.
10. Brown, M.A. and Leo, P.J. (2019) Genetic susceptibility to cervical neoplasia. *Papillomavirus Res.*, **7**, 132–134.
11. Hemminki, K. and Vaittinen, P. (1998) Familial risks in in situ cancers from the family-cancer database. *Cancer Epidemiol. Biomark. Prev.*, **7**, 865–868.
12. Chen, D., Juko-Pecirep, I., Hammer, J., Ivansson, E., Enroth, S., Gustavsson, I., Feuk, L., Magnusson, P.K., McKay, J.D., Wilander, E. and Gyllensten, U. (2013) Genome-wide association study of susceptibility loci for cervical cancer. *J. Natl. Cancer Inst.*, **105**, 624–633.
13. Chen, D., Enroth, S., Liu, H., Sun, Y., Wang, H., Yu, M., Deng, L., Xu, S. and Gyllensten, U. (2016) Pooled analysis of genome-wide association studies of cervical intraepithelial neoplasia 3 (CIN3) identifies a new susceptibility locus. *Oncotarget*, **7**, 42216–42224.
14. Rashkin, S.R., Graff, R.E., Kachuri, L., Thai, K.K., Alexeeff, S.E., Blatchins, M.A., Cavazos, T.B., Corley, D.A., Emami, N.C., Hoffman, J.D. *et al.* (2020) Pan-cancer study detects genetic risk variants and shared genetic basis in two large cohorts. *Nat. Commun.*, **11**(1), 4423.
15. Bowden, S.J., Bodinier, B., Kalliala, I., Zuber, V., Vuckovic, D., Doulgeraki, T., Whitaker, M.D., Wielscher, M., Cartwright, R., Tsilidis, K.K. *et al.* (2021) Genetic variation in cervical preinvasive and invasive disease: a genome-wide association study. *Lancet Oncol.*, **22**, 548–557.
16. Leo, P.J., Madeleine, M.M., Wang, S., Schwartz, S.M., Newell, F., Pettersson-Kymmer, U., Hemminki, K., Hallmans, G., Tiews, S., Steinberg, W. *et al.* (2017) Defining the genetic susceptibility to cervical neoplasia—a genome-wide association study. *PLoS Genet.*, **13**, 1–20.
17. Ramachandran, D. and Dörk, T. (2021) Genomic risk factors for cervical cancer. *Cancer*, **13**(20), 5137.
18. Chen, D. and Gyllensten, U. (2015) Lessons and implications from association studies and post-GWAS analyses of cervical cancer. *Trends Genet.*, **31**, 41–54.
19. Chen, D., Cui, T., Ek, W.E., Liu, H., Wang, H. and Gyllensten, U. (2015) Analysis of the genetic architecture of susceptibility to cervical cancer indicates that common SNPs explain a large proportion of the heritability. *Carcinogenesis*, **36**, 992–998.
20. Miura, K., Mishima, H., Kinoshita, A., Hayashida, C., Abe, S., Tokunaga, K., Masuzaki, H. and Yoshiura, K. (2014) Genome-wide association study of HPV-associated cervical cancer in Japanese women. *J. Med. Virol.*, **86**, 1153–1158.
21. Takeuchi, F., Kukimoto, I., Li, Z., Li, S., Li, N., Hu, Z., Takahashi, A., Inoue, S., Yokoi, S., Chen, J. *et al.* (2019) Genome-wide association study of cervical cancer suggests a role for ARRDC3 gene in human papillomavirus infection. *Hum. Mol. Genet.*, **28**, 341–348.
22. Shi, Y., Li, L., Hu, Z., Li, S., Wang, S., Liu, J., Wu, C., He, L., Zhou, J., Li, Z. *et al.* (2013) A genome-wide association study identifies two new cervical cancer susceptibility loci at 4q12 and 17q12. *Nat. Genet.*, **45**, 918–922.
23. Koel, M., Võsa, U., Lepamets, M., Laivuori, H., Lemmelä, S., Daly, M., Estonian Biobank Research Team, FinnGen, Palta, P., Mägi, R. and Laisk, T. (2021) GWAS meta-analysis and gene expression data link reproductive tract development, immune response and cellular proliferation/apoptosis with cervical cancer and clarify overlap with other cervical phenotypes. *medRxiv*, 2021.06.18.21259075. doi: https://doi.org/10.1101/2021.06.18.21259075.
24. Li, X., Huang, K., Zhang, Q., Zhou, J., Sun, H., Tang, F., Zhou, H., Hu, T., Wang, S., Jia, Y. *et al.* (2017) Genome-wide association study identifies four SNPs associated with response to platinum-based neoadjuvant chemotherapy for cervical cancer. *Sci. Rep.*, **7**, 1–7.
25. Gallagher, M.D. and Chen-Plotkin, A.S. (2018) The post-GWAS era: from association to function. *Am. J. Hum. Genet.*, **102**, 717–730.
26. Edwards, S.L., Beesley, J., French, J.D. and Dunning, A.M. (2013) Beyond GWASs: illuminating the dark road from association to function. *Am. J. Hum. Genet.*, **93**, 779–797.
27. Gondro, C., van der Werf, J. and Hayes, B. (2017) Genome-Wide Association Studies and Genomic Prediction. *Genome-Wide Association Studies and Genomic Prediction*; Humana Press.
28. Freedman, M.L., Monteiro, A.N.A., Gayther, S.A., Coetzee, G.A., Risch, A., Plass, C., Casey, G., De Biasi, M., Carlson, C., Duggan, D. *et al.* (2011) Principles for the post-GWAS functional characterization of cancer risk loci. *Nat. Genet.*, **43**, 513–518.
29. Vandiedonck, C. (2018) Genetic association of molecular traits: a help to identify causative variants in complex diseases. *Clin. Genet.*, **93**, 520–532.
30. Amos, C.I., Dennis, J., Wang, Z., Byun, J., Schumacher, F.R., Gayther, S.A., Casey, G., Hunter, D.J., Sellers, T.A., Gruber, S.B. *et al.* (2017) The Oncoarray consortium: a network for understanding the genetic architecture of common cancers. *Cancer Epidemiol. Biomark. Prev.*, **26**, 126–135.
31. Zeng, F., Xue, M., Xiao, T., Li, Y., Xiao, S., Jiang, B. and Ren, C. (2016) MiR-200b promotes the cell proliferation and metastasis of cervical cancer by inhibiting FOXG1. *Biomed. Pharmacother.*, **79**, 294–301.
32. Rozengurt, E., Rey, O. and Waldron, R.T. (2005) Protein kinase D signaling. *J. Biol. Chem.*, **280**, 13205–13208.
33. Roy, A., Ye, J., Deng, F. and Wang, Q. (2017) Protein kinase D signaling in cancer: a friend or foe? *Biochim. Biophys. Acta, Gen. Subj.*, **1868**, 283–294.
34. Li, J., Han, L., Roebuck, P., Diao, L., Liu, L., Yuan, Y., Weinstein, J.N. and Liang, H. TANRIC: an interactive open platform to explore the function of lncRNAs in cancer. *Cancer Res.*, **75**, 3728–3737.
35. Hemani, G., Zheng, J., Elsworth, B., Wade, K.H., Haberland, V., Baird, D., Laurin, C., Burgess, S., Bowden, J., Langdon, R. *et al.* (2018) The MR-base platform supports systematic causal inference across the human phenome. *elife*, **7**, 1–29.
36. Bartonicek, N., Maag, J.L.V. and Dinger, M.E. (2016) Long noncoding RNAs in cancer: mechanisms of action and technological advancements. *Mol. Cancer*, **15**, 1–10.
37. Ramachandran, D., Schürmann, P., Mao, Q., Wang, Y., Bretschneider, L.M., Speith, L.M., Hülse, F., Enßen, J., Bousset, K., Jentschke, M. *et al.* (2020) Association of genomic variants at the human leukocyte antigen locus with cervical cancer risk, HPV status and gene expression levels. *Int. J. Cancer*, **147**, 2458–2468.
38. Ramachandran, D., Wang, Y., Schürmann, P., Hülse, F., Mao, Q., Jentschke, M., Böhmer, G., Strauß, H.G., Hirchenhain, C., Schmidmayr, M. *et al.* (2021) Association of genomic variants at PAX8 and PBX2 with cervical cancer risk. *Int. J. Cancer*, **149**, 893–900.
39. Michailidou, K., Lindström, S., Dennis, J., Beesley, J., Hui, S., Kar, S., Lemaçon, A., Soucy, P., Glubb, D., Rostamianfar, A. *et al.* (2017) Association analysis identifies 65 new breast cancer risk loci. *Nature*, **551**, 92–94.
40. Price, A.L., Patterson, N.J., Plenge, R.M., Weinblatt, M.E., Shadick, N.A. and Reich, D. (2006) Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.*, **38**, 904–909.

41. Zhang, F., Wang, Y. and Deng, H.W. (2008) Comparison of population-based association study methods correcting for population stratification. *PLoS One*, **3**, 1–7.
42. Wellcome Trust Case Control Consortium (2007) Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature*, **447**, 661–678.
43. Voorman, A., Lumley, T., McKnight, B. *et al.* (2011) Behavior of QQ-plots and genomic control in studies of gene-environment interaction. *PLoS One*, **6**, e19416.
44. Helbig, S., Wockner, L., Bouendeu, A., Hille-Betz, U., McCue, K., French, J.D., Edwards, S.L., Pickett, H.A., Reddel, R.R., Chenevix-Trench, G. *et al.* (2017) Functional dissection of breast cancer risk-associated TERT promoter variants. *Oncotarget*, **8**, 67203–67217.
45. Cabianca, D.S., Casa, V., Bodega, B., Xynos, A., Ginelli, E., Tanaka, Y. and Gabellini, D. (2012) A long ncRNA links copy number variation to a polycomb/trithorax epigenetic switch in fshd muscular dystrophy. *Cell*, **149**, 819–831.