Association of Single Nucleotide Polymorphisms in Glycosylation Genes with Risk of Epithelial Ovarian Cancer

Thomas A. Sellers,[1] Yifan Huang,[1] Julie Cunningham,[2] Ellen L. Goode,[3] Rebecca Sutphen,[1] Robert A. Vierkant,[3] Linda E. Kelemen,[6] Zachary S. Fredericksen,[3] Mark Liebow,[4] V. Shane Pankratz,[3] Lynn C. Hartmann,[5] Jeff Myer,[2] Edwin S. Iversen, Jr.,[7] Joellen M. Schildkraut,[8] and Catherine Phelan[1]

[1]Division of Cancer Prevention and Control, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida; Departments of [2]Laboratory Medicine and Pathology, [3]Health Sciences Research, [4]Medicine, and [5]Medical Oncology, Mayo Clinic, Rochester, Minnesota; [6]Division of Population, Health and Information, Calgary, Alberta, Canada; and Departments of [7]Statistical Science and [8]Community and Family Medicine and the Duke Comprehensive Cancer Center, Duke University, Durham, North Carolina

Abstract

Studies suggest that underglycosylation of the cell membrane mucin MUC1 may be associated with epithelial ovarian cancer. We identified 26 genes involved in glycosylation and examined 93 single nucleotide polymorphisms (SNP) with a minor allele frequency of ≥0.05 in relation to incident ovarian cancer. Cases were ascertained at the Mayo Clinic, Rochester, MN (n = 396) or a 48-county region in North Carolina (Duke University; n = 534). Ovarian cancer–free controls (n = 1,037) were frequency matched to the cases on age, race, and residence. Subjects were interviewed to obtain data on risk factors and a sample of blood for DNA and genotyped using the Illumina GoldenGate assay. We excluded subjects and individual SNPs with genotype call rates of <90%. Data were analyzed using logistic regression, with adjustment for age and residence. We fitted dominant, log additive, and recessive genetic models. Among Caucasians, nine SNPs in eight genes were associated with risk at $P < 0.05$ under at least one genetic model before adjusting for multiple testing. A SNP in *GALNT1* (rs17647532) was the only one that remained statistically significant after Bonferroni adjustment for multiple testing but was not statistically significant in Hardy-Weinberg equilibrium among controls. Haplotype analyses revealed a global association of *GALNT1* with risk (P = 0.038, under a recessive genetic model), which largely reflected a decreased risk of one haplotype (0.10 frequency; odds ratio, 0.07; P = 0.01) compared with the most common haplotype (0.39 frequency). These results suggest that genetic polymorphisms in the glycoslyation process may be novel risk factors for ovarian cancer. (Cancer Epidemiol Biomarkers Prev 2008;17(2):397–404)

Introduction

Epithelial ovarian cancer has a recognized genetic component. There is an autosomal dominant pattern of ovarian or breast-ovarian cancer in some families (1, 2) with a high lifetime risk (3-5). Mutation screening of *BRCA1* and *BRCA2* in population-based series of ovarian cancer cases has shown that mutations in these genes collectively account for only a small percentage of cases (3, 6). Ovarian cancer also occurs in Lynch syndrome II, an autosomal dominant predisposition to cancers of the colon, endometrium, ovary, and other sites (7, 8). The genetic defects underlying Lynch syndrome II include the mismatch repair genes *MLH1, MSH2, PMS1, PMS2*, and *MSH6* (9, 10). The exact contribution of these genes to ovarian cancer risk is not known but is estimated to be lower than that of *BRCA1* and *BRCA2*. Collectively, these high-risk, high-penetrance genes are rare in the population and are estimated to account for 10% to 15% of ovarian cancer (11). Studies to identify other chromosomal regions that might harbor major genes for ovarian cancer risk provide little evidence for the existence of additional high-risk genes for ovarian cancer susceptibility (12). Thus, there is emerging consensus that most of the genetic component of ovarian cancer risk is due to genetic polymorphisms that confer low to moderate risk.

A common approach to identify risk variants is to rely on known biology to identify plausible candidate genes. The present report was motivated by two converging lines of evidence from molecular biology and epidemiology. The mucin glycoprotein MUC1 is frequently overexpressed in epithelial ovarian cancer (13) and found in circulation (14). Mucin glycoproteins have a protein backbone composed of repetitive domains rich in peptides that can serve as sites for O-linked glycoslyation. Multiple O-linked oligosaccharides bind to these sites and represent 50% to 80% of the total glycoprotein mass. Normal epithelial cells express a

Received 6/20/07; revised 11/13/07; accepted 11/20/07.

The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked *advertisement* in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

Requests for reprints: Thomas A. Sellers, Division of Cancer Prevention and Control, H. Lee Moffitt Cancer Center and Research Institute, Tampa, Florida. Phone: 813-632-1315. E-mail: sellerta@moffitt.usf.edu

doi:10.1158/1055-9965.EPI-07-0565

highly glycosylated form of MUC1 at low levels, but ovarian cancer cells express high levels of a hypoglycosylated form (15). Two recent studies provide evidence that links circulating levels of anti-MUC1 antibodies with risk factors for ovarian cancer. The first report (16) observed that oral contraceptive use, breast mastitis, bone fracture or osteoporosis, pelvic surgeries, nonuse of talc in genital hygiene, and (to a lesser extent) current smoking predicted anti-MUC1 antibody levels among 705 women selected as controls for an ovarian cancer study. A second report by the same group (17) on 721 controls found that early age at first birth, menstrual cycles longer than 30 days, and lifetime number of ovulatory cycles were inversely associated with anti-MUC1 antibodies.

We hypothesized that polymorphisms in genes encoding glycosylation enzymes may be risk factors for epithelial ovarian cancer. The current study compared the frequencies of genotypes and selected haplotypes between women recently diagnosed with epithelial ovarian cancer and a group of control women without ovarian cancer. The findings may have important implications for our understanding of the pathogenesis of this important malignancy.

Materials and Methods

Study Population. Details of the study design have been presented previously (18). Briefly, the subjects recruited for this research project were identified through two institutions: Duke University and Mayo Clinic. The study is ongoing, with the current results based on recruitment from June 1999 to March 2006. The protocol was approved by the institutional review board at each institution, and all study subjects provided written informed consent. Although there were several differences in the specific study designs used at each site, the overall inclusion-exclusion criteria were similar. Cases were Caucasian or African American women ascertained within 1 year of a diagnosis of histologically confirmed primary epithelial ovarian cancer, either borderline or invasive. The lower age limit was 20 years at both sites, but Mayo had no upper age limit, whereas Duke excluded women ages >74 years.

The Mayo ascertainment began in January 2000 and was clinic based. The catchment area was limited to a six-state region that represents >85% of all ovarian cancer cases seen there: Minnesota, Iowa, Wisconsin, Illinois, North Dakota, and South Dakota. We selected clinic-based controls from women seeking general medical evaluation, frequency matched to cases on age (5-year age category), race, and state of residence. Potential controls were excluded if they had a history of ovarian cancer or oophorectomy. Response rates for those invited to participate at the Mayo site were 83% for cases and 74% for controls.

The Duke study is population based with a rapid case ascertainment network covering a 48-county region of North Carolina. Recruitment has been ongoing since May 1, 1999. List-assisted random digit dialing and Health Care Financing Administration roster methods were used to identify control subjects. Controls were frequency matched to the cases on the basis of race (Black versus non-Black), age (5-year age categories), and residence. The response rate was 71% among eligible cases (87% among cases who could be contacted) and 64% among the controls.

Risk Factor Data Collection. Information on known and suspected ovarian cancer risk factors and demographic data were collected through in-person interviews. Mayo cases were interviewed by telephone if discharged from the hospital before contact about the study. Similar questionnaires were used at each institution. Information collected included race/ethnicity, menstrual and reproductive history, use of exogenous hormones, medical and surgical history, height and weight 1 year before the interview, use of tobacco, education level, and family history of breast or ovarian cancer in first- or second-degree relatives.

Collection and Processing of Biospecimens. Genomic DNA was obtained from cases and controls in one of two ways. At Duke, venipuncture was done at the conclusion of the interview. At Mayo, participants had an extra vial of blood drawn in the course of their scheduled medical care. DNA was extracted from fresh peripheral blood using the Gentra AutoPure LS Purgene salting out methodology (Gentra). Due to limited quantity of available DNA from Duke subjects, we did whole-genome amplification (WGA) on all Duke samples (534 cases and 568 controls) using the REPLI-G protocol (Qiagen), with 200 ng genomic DNA. DNA concentrations were adjusted to 50 ng/μL and verified using PicoGreen dsDNA Quantitation kit (Molecular Probes). The samples were bar-coded to ensure accurate and reliable sample processing and storage.

Selection of Candidate Genes and Single Nucleotide Polymorphisms. Genes involved in the glycosylation process were identified through several sources, including peer-reviewed published literature (14, 19) and the Cancer Genome Anatomy Project Biokarta and Kegg pathway databases. A total of 26 genes were selected: *FUT10, FUT3, FUT5, FUT6, FUT7, FUT8, FUT9, GALNT1, GALNT14, GALNT2, GALNT3, GALNT5, GALNT6, GALNT7, MGAT2, MGAT3, MGAT5, POFUT1, SIAT1, SIAT4A, ST3GAL3, ST3GAL4, ST3GAL5, ST6GALNAC5, ST8SIA2*, and *ST8SIA5*.

For each gene, chromosome and protein attributes were selected and the data mined from the Ensembl database version 34 (Biomart) using the gene reference sequence identification number (RefSeq ID) and the approved gene symbol from HUGO or Entrez Gene. The chromosomal location on build 35 and strand (forward or reverse) were provided to Illumina. Illumina verified chromosomal coordinates. We requested all single nucleotide polymorphisms (SNP) within each gene as well as up to 10 kb in the 5′ and 3′ flanking regions and all nonsynonymous SNPs with a minor allele frequency (MAF) > 0.05 and Illumina Design Score > 0.6. The Illumina Assay Design Tool database includes all SNP data contained in the public domain, filtering out SNPs that are not suitable for the Illumina platform, such as insertions/deletions, tri- and tetra-allelic SNPs, and SNPs that are not uniquely localized.

Genotyping Methods and Quality Control. Ninety-six glycosylation SNPs were included in an Illumina GoldenGate assay for 1,536 SNPs. Ninety-three percent of the SNPs had SNP_scores greater than 0.6 and none

had scores less than 0.4. 2368 samples (1,086 genomic and 1,282 WGA; 250 ng each) were genotyped on 1,967 unique subjects following the Illumina protocol, including 930 cases (396 genomic and 534 WGA) and 1,037 controls (469 genomic and 568 WGA). Genotype calls were made using the Genotyping module of BeadStudio 2 software. Genomic and WGA DNA were analyzed separately, as WGA DNA clustered differently from genomic DNA. For this project, samples with GenCall scores below 0.25 and/or call rates below 90%, and SNPs with GenCall scores below 0.4 or call rates below 90%, were failed. For genomic DNA, 1,492 SNPs passed this initial quality assurance cutoff, whereas 1,435 passed in the WGA DNA set.

Several quality-control procedures were established. For the genomic DNA, we included eight replicates of a CEPH family trio from the Coriell Institute (mother, father, and child) and replicates of an additional three standard DNAs in each 96-well plate. The replicate and inheritance data were used to review and refine clustering. In addition, two samples per 96-well plate were blindly duplicated (n = 20). Among the WGA samples, there were 88 replicates (same WGA preparation) and 15 replicate WGA (separate WGA preparation) samples genotyped. In addition, 124 of the WGA samples had sufficient genomic DNA to allow genotyping and thus a means to monitor the concordance between genomic and WGA DNA genotype calls. Of these, two had an unusually high number of discrepancies between the genome and WGA'd samples due to sampling errors and were therefore excluded. These genomic DNAs were analyzed along with the 1,086 genomic DNAs to generate genotype calls.

We attempted to genotype 2,058 subjects on 96 SNPs in 26 glycosylation genes. Of these subjects, we excluded 87 with poor clustering and therefore failed for every SNP, 2 with low call rates (<95% of SNPs successfully genotyped) and 2 with an unusually high number of discrepancies between genomic DNA genotype results and WGA results. Of the 1,536 SNPs targeted for genotyping, we excluded 44, which failed completely (or nearly completely), 6 because of low call rates (<95%), and 7 with no genetic variability (that is, monomorphic), resulting in a total of 1,480 SNPs. Of the 57 SNPs that failed, 3 were in glycosylation genes. The exact number of SNPs for analysis varied slightly per study site: 1,421 had results of sufficient quality for both Mayo and Duke subjects, 51 were successfully genotyped for Mayo subjects but failed for Duke subjects, and 7 were successfully genotyped for Duke subjects but had either low call rates or no genetic variability for the Mayo samples. Genotypes at these 58 problematic loci were coded as missing. This resulted in a final sample size of 1,967 subjects and 93 SNPs in glycosylation genes.

Statistical Methods. Before analysis, we determined descriptive statistics using frequencies and percents for categorical variables and means and SDs for continuous variables. The distributions of covariates were compared across study site and case status using ANOVA methods for continuous variables and χ^2 tests for categorical variables. SNP genotype frequencies among the controls were tested for Hardy-Weinberg equilibrium (HWE) using χ^2 goodness-of-fit tests.

We first did SNP-specific analyses to examine the main effects of the SNPs on risk of ovarian cancer. We fitted log additive, dominant, and recessive models for each individual SNP. Unconditional logistic regression models were fitted using the R function glm (http://www.r-project.org) to estimate odds ratios (OR) and corresponding 95% confidence intervals (95% CI) between genotypes and case status. Wald χ^2 tests were calculated using the R function ANOVA to obtain P values for SNP effects. Due to differences in genotype frequencies by race, separate logistic regression models were fitted for Caucasian and non-Caucasian subjects. All models were adjusted for the design variables of geographic area/study site and age group. We also adjusted for several nongenetic risk factors: body mass index (BMI), months of hormone replacement therapy (HRT) use, months of oral contraceptive use, and parity/age first birth combination. Although none were traditional confounding factors, results were similar regardless whether they were in the model. In addition, we carried out sensitivity analyses by excluding the borderline cases in the Caucasian samples. Individual SNP associations with risk were compared before and after exclusion of borderline cases.

We next examined the association of haplotypes in each glycosylation gene with ovarian cancer status. These analyses were restricted to Caucasian subjects due to the relatively low number of enrolled non-Caucasian subjects and done using the statistical program Haplo.stats (20). For each gene with multiple SNPs, we first used the function Haplo.score to estimate haplotypes and test global significance, adjusting for the covariates mentioned above. If the global P value for haplotype association was smaller than 0.05, we examined individual haplotypes by comparing the risk of ovarian cancer associated with each inferred haplotype to the risk associated with the highest estimated frequency haplotype. The estimated OR and the corresponding P value were obtained using the function Haplo.glm with adjustment for the covariates. Rare haplotypes were pooled into a single category to reduce the burden of sparse table cells.

In addition to performing single SNP and haplotype analyses, we assessed potential modifying effects of selected demographic and clinical variables by fitting a series of interaction models. The modifying effects of the following variables were considered based on published association with MUC1 antibody levels: ever smoked, BMI (median split), live birth (0 versus $\geq$1), ever used oral contraceptive, ever used HRT, and menopausal status (premenopausal and postmenopausal). These analyses were also restricted to Caucasian subjects. To avoid further the possibility of sparse table cells, all the SNPs were first modeled as dichotomous variables based on the presence (one or two copies) or absence (zero copies) of the variant alleles. Two SNPs (rs3828139 and rs37460) with significant recessive main effect and high MAF were also modeled as dichotomous variables based on the presence or absence of the common alleles. Similarly, each environmental variable was dichotomized based on either a median split of the distribution among controls or a pooling of categories. For each interaction between the nongenetic variables and SNPs, a logistic regression model was fitted with the corresponding SNP, the corresponding environmental

variable, the interaction term, and other potentially confounding covariates. Based on the models, P values for testing the interaction effects were obtained using Wald χ^2 tests. For each interaction with $P < 0.05$, we obtained OR and 95% CI estimates for individual strata of SNP and environmental variable level.

Results

The distribution of nongenetic risk factors by study site and case-control status is summarized in Table 1. Cases tended to be more obese, less likely to use oral contraceptive pills, have lower parity, and more likely to report a family history of ovarian cancer. Case-control differences were generally similar across sites, with the exception of HRT use, with Duke cases having higher exposure levels than the other groups.

Differences in genotype frequencies between cases and controls were compared by race and study site for each of the 93 SNPs in the 26 genes investigated. Caucasian and non-Caucasian subjects had significantly different allele frequencies for some of the SNPs, so all analyses were stratified by race. There were no differences in allele frequency by study site, so data from Mayo and Duke were combined for analysis. After fitting dominant,

Table 1. Distribution of nongenetic risk factors by study site and ovarian cancer case-control status

Variable	Mayo		Duke	
	Cases (n = 396)	Controls (n = 469)	Cases (n = 534)	Controls (n = 568)
Age, mean (SD)	59.8 (13.3)	60.1 (13.0)	54.0 (11.5)	54.7 (12.2)
Race, n (%)				
Caucasian	385 (97.2)	462 (98.5)	444 (83.3)	479 (84.3)
African American	11 (2.8)	7 (1.5)	89 (16.7)	88 (15.7)
Highest education achieved,* n (%)				
No diploma	25 (6.9)	19 (4.3)	53 (9.9)	69 (12.1)
High school diploma	136 (37.4)	117 (26.4)	153 (28.7)	149 (26.2)
Post-high school education	203 (55.8)	307 (69.3)	327 (61.4)	350 (61.6)
Smoking status,† n (%)				
Never	233 (63.7)	285 (64.6)	290 (54.4)	282 (49.6)
Former	101 (27.26)	132 (29.9)	177 (33.2)	181 (31.9)
Current	32 (8.7)	24 (5.4)	66 (12.4)	105 (18.5)
Pack-years cigarettes smoked, n (%)				
None	233 (64.9)	285 (68.3)	297 (57.6)	291 (53.5)
≤20	71 (19.8)	84 (20.1)	130 (25.2)	148 (27.2)
>20	55 (15.3)	48 (11.5)	89 (17.2)	105 (19.3)
BMI,* mean (SD)	28.2 (6.2)	26.9 (5.6)	28.3 (7.3)	27.6 (6.5)
Age at menarche (y), n (%)				
<12	55 (18.7)	68 (15.8)	130 (24.4)	118 (20.8)
12	77 (26.2)	100 (23.2)	153 (28.8)	166 (29.2)
13	79 (26.9)	126 (29.2)	134 (25.2)	161 (28.3)
14+	83 (28.2)	137 (31.8)	115 (21.6)	123 (21.7)
Parity/age first birth combo,‡ n (%)				
Nulliparous	70 (18.3)	66 (15)	113 (21.2)	73 (12.9)
1-2, ≤20 y	29 (7.6)	25 (5.7)	73 (13.7)	69 (12.1)
1-2, >20 y	103 (26.9)	131 (29.8)	193 (36.2)	233 (41.0)
3+, ≤20 y	73 (19.1)	64 (14.5)	81 (15.2)	93 (16.4)
3+, >20 y	108 (28.2)	154 (35.0)	73 (13.7)	100 (17.6)
Problem getting pregnant,† n (%)				
No	306 (80.3)	359 (84.9)	384 (72.2)	440 (77.6)
Yes	75 (19.7)	64 (15.1)	148 (27.8)	127 (22.4)
Postmenopausal, n (%)				
No	113 (29.8)	109 (24.7)	140 (28.3)	183 (33.0)
Yes	266 (70.2)	333 (75.3)	354 (71.7)	372 (67.0)
Oral contraceptive use (mo),* n (%)				
Never	176 (46.6)	166 (38.4)	182 (34.7)	181 (32.2)
1-48	98 (26.5)	92 (21.3)	158 (30.2)	160 (28.5)
48+	96 (25.9)	174 (40.3)	184 (35.1)	221 (39.3)
HRT use (mo),‡ n (%)				
Never	240 (63.8)	248 (58.6)	196 (37.7)	349 (63.0)
1-60	64 (17.1)	80 (18.9)	207 (39.8)	109 (19.7)
60+	72 (19.1)	95 (22.5)	117 (22.5)	96 (17.3)
Family history ovarian cancer,*,† n (%)				
No	333 (86.7)	411 (92.6)	492 (92.1)	543 (95.6)
Yes	51 (13.3)	33 (7.4)	42 (7.9)	25 (4.4)
Family history ovarian cancer or breast cancer, n (%)				
No	217 (56.5)	255 (57.4)	338 (63.3)	378 (66.5)
Yes	167 (43.5)	189 (42.6)	196 (36.7)	190 (33.5)

*Case-control differences at Mayo <0.01.
†Case-control differences at Duke <0.05.
‡Case-control differences at Duke <0.01.

Table 2. Glycosylation gene SNPs significantly associated with ovarian cancer risk among Caucasian subjects

Gene	SNP	Case/control (829/941)			MAF*	$P^{\dagger}$	Model[‡]	OR (95% CI)	$P^{\S}$
FUT3	rs2306969	*AA*	*AG*	*GG*					
	Combined	496/534	289/356	42/49	0.23	0.30	D	0.79 (0.64-0.97)	0.024
	Mayo only	224/259	143/185	17/16	0.23	0.013		0.85 (0.63-1.15)	
	Duke only	272/275	146/171	25/33	0.23	0.37		0.76 (0.57-1.02)	
FUT7‖	rs10732706	*TT*	*TC*	*CC*					
	Mayo only	172/189	170/213	43/60	0.35	1.0	A	0.80 (0.64-1.00)	0.049
GALNT1	rs17647532	*TT*	*TC*	*CC*					
	Combined	678/757	150/162	1/20	0.10	0.002	R	0.07 (0.01-0.53)	0.00017
	Mayo only	314/378	70/71	1/12	0.10	0.0003		0.14 (0.02-1.14)	
	Duke only	364/379	80/91	0/8	0.10	0.35		NA	
GALNT2‖	rs3213495	*AA*	*AG*	*GG*					
	Mayo only	313/331	64/105	4/8	0.12	0.92	A	0.62 (0.44-0.87)	0.0053
GALNT6	rs907352	*GG*	*GC*	*CC*					
	Combined	635/723	172/205	22/12	0.13	0.55	R	2.38 (1.09-5.19)	0.024
	Mayo only	286/349	88/109	11/4	0.13	0.15		3.61 (0.96-13.5)	
	Duke only	349/374	84/96	11/8	0.12	0.52		1.72 (0.63-4.72)	
GALNT7	rs934358	*CC*	*CG*	*GG*					
	Combined	522/548	280/342	27/50	0.22	0.72	A	0.80 (0.67-0.95)	0.010
	Mayo only	252/274	118/165	15/23	0.21	0.78		0.70 (0.54-0.92)	
	Duke only	270/274	162/177	12/27	0.23	0.82		0.84 (0.66-1.07)	
MGAT5	rs1257187	*AA*	*AG*	*GG*					
	Combined	571/684	230/243	27/14	0.16	0.14	A	1.27 (1.05-1.55)	0.016
	Mayo only	265/334	110/122	10/6	0.16	0.16		1.21 (0.90-1.64)	
	Duke only	306/350	120/121	17/8	0.16	0.50		1.32 (1.01-1.74)	
ST3GAL3	rs3828139	*TT*	*TC*	*CC*					
	Combined	243/238	423/469	160/230	0.47	0.97	R	0.77 (0.60-0.98)	0.035
	Mayo only	122/116	187/230	76/116	0.47	0.16		0.78 (0.54-1.12)	
	Duke only	121/122	236/239	84/114	0.48	0.50		0.75 (0.53-1.07)	
	rs37460	*GG*	*GC*	*CC*					
	Combined	237/242	426/462	164/233	0.48	0.67	R	0.78 (0.61-0.99)	0.048
	Mayo only	120/119	188/227	77/116	0.47	0.71		0.80 (0.56-1.15)	
	Duke only	117/123	238/235	87/117	0.48	0.82		0.77 (0.55-1.09)	

*MAF estimated using both cases and controls.
† P for testing departure from HWE among controls.
‡ Genetic models (A, additive; D, dominant; R, recessive) with adjustment for age, geographic region, BMI, HRT use, oral contraceptive use, and parity/age at first birth.
§ P for testing the genetic effects before multiplicity adjustment.
‖ SNPs rs10732706 and rs3213495 genotyping in the Duke samples did not pass the quality control.

recessive, and additive models and adjusting for age and the covariates listed as a footnote in Table 2, statistically significant main effects before adjusting for multiple testing were identified in 8 of the 26 genes studied and 9 of the 93 SNPs (Table 2). One gene (*ST3GAL3*) had 2 SNPs associated with risk. It is interesting to note that 7 of the 9 SNPs were inversely associated with risk and that only 1 of the 9 was based on a dominant genetic model. Only one SNP (rs17647532 in *GALNT1*) was not in HWE ($P = 0.002$), but this was limited to controls from Mayo ($P = 0.0003$), not Duke ($P = 0.35$), although the estimate of the MAF was 0.10 at both sites. Table 2 also presents the results for Mayo and Duke samples separately.

Given the large number of statistical tests (93 SNPs ×3 genetic models), one needs to be cautious about interpretation. Therefore, we used the Bonferroni method to adjust for multiple tests and keep the overall type I error rate under 5%. Only the *GALNT1* SNP (rs17647532) remained statistically significant. This SNP was rare (MAF of 0.10); therefore, the results are based on few homozygous carriers (1 case and 20 controls). Given the lack of HWE for this SNP among the Mayo controls, we did *post hoc* analyses by study site to see if this influenced the findings. The only case homozygous for the minor allele was ascertained at Mayo, which meant the association could not be estimated among the Duke subjects. However, the combined OR (95% CI) of 0.07 was somewhat attenuated among the Mayo subset [0.14 (0.02-1.14)]. This suggests that the overall result was not affected by the absence of HWE at Mayo. The SNP is located in the promoter/regulatory region in *GALNT1*. We applied bioinformatic tools to help interpret the plausibility of this finding. In particular, we ran the program FASTSNP (21) to predict the potential functional significance of the polymorphism. The SNP was categorized to have low to medium risk, indicating that it may affect the level, location, or timing of gene expression. Because there is some evidence that borderline ovarian cancers may be etiologically different from invasive epithelial cancers, we did *post hoc* sensitivity analyses after excluding cases with borderline tumors. Results were unchanged with one exception: the two SNPs in *ST3GAL3* both became more strongly and significantly inversely associated with risk (rs3828139 OR, 0.72; P without multiplicity adjustment = 0.010; rs37460 OR, 0.71; P without multiplicity adjustment = 0.0069).

The analysis of Caucasian subjects continued with haplotype analyses for genes that had more than one SNP genotyped. Based on the global score statistic, only *GALNT1* was significantly associated with risk

Table 3. Haplotype analysis of *GALNT1* and ovarian cancer risk among Caucasian subjects

Haplotype	Frequency	OR	*P*
TTTAT	0.39	Reference	—
TTCCT	0.35	0.84	0.25
ACTAC	0.10	0.07	0.01
TTTCT	0.09	1.12	0.85
ACTAT	0.07	1.85	0.28

NOTE: Adjusted for age, geographic region, BMI, HRT use, oral contraceptive use, and parity/age at first birth. Based on SNPs rs6507133 (*T*/*A*), rs11663626 (*T*/*C*), rs556736 (*T*/*C*), rs607498 (*A*/*C*), and rs17647532 (*T*/*C*).

(P = 0.038). There were five measured SNPs in *GALNT1* (rs6507133, rs11663626, rs556736, rs607498, and rs17647532) and they were in LD with r^2 values ranging from 0.06 to 1.0. Although 32 haplotypes were possible, we observed only 5 common haplotypes (frequency > 0.05); their frequencies and individual haplotype effects are presented in Table 3. Because the individual SNP analysis revealed a SNP (rs17647573) with significant recessive effect, we fitted the haplotype genetic model to be a recessive model. The result of the global score test was largely the reflection of the *ACTAC* haplotype (0.10 frequency) that was associated with decreased risk (OR, 0.07; P = 0.01) compared with the most common haplotype *TTTAT* (0.39 frequency), consistent with an untyped causal allele or the haplotype.

SNP-specific results based on the African American subjects are shown in Table 4. A total of 10 SNPs in 7 genes had multivariate-adjusted P values smaller than 0.05 from any of the three genetic models. All of the SNPs were in HWE. Three of the 7 genes were also associated with risk among the Caucasians (*GALNT1, MGAT5*, and *ST3GAL3*), but only 1 of the SNPs was in common: rs1257189 in *MGAT5*. Most of the SNPs (7 of 10) were associated with increased risk, and most often (6 of 10) this was under an additive genetic model. None of these associations were statistically significant after adjustment for multiple comparisons.

Exploratory models were then fitted to the Caucasian subjects to test gene × environment interactions (technically SNP × environment interactions). We included a fixed set of nongenetic variables based on the epidemiologic literature linking them with MUC1 antibodies (16, 17): BMI, oral contraceptive use, HRT use, menopause status, parity, and smoking. Rather than fitting interactions under all genetic models, this exercise was limited to a dominant genetic effect for all 93 SNPs plus 2 recessive models (based on main effect results). There were 18 interactions significant at the 0.05 level (very close to the expected value of 20 based on chance alone). Because we conducted 570 tests (93 SNPs × 6 environmental factors + 2 SNPs under other genetic models × 6 environmental factors), the P value cutoff based on the Bonferroni method is 0.05/570 or 0.000088. None of the interactions exceeded this threshold (data not shown).

Discussion

This study sought to evaluate whether inherited variation in genes involved in the glycosylation of MUC1 were potential risk factors for ovarian cancer. The analysis focused on 93 SNPs in 26 genes and considered different genetic models, haplotypes, and gene × environment interaction. Before adjustment for multiple comparisons, statistically significant main effects were identified among Caucasians in 6 of the 26 genes and 9 of the 93 SNPs and another 10 genes had at least one statistically

Table 4. Glycosylation gene SNPs significantly associated with ovarian cancer risk among African American subjects

Gene	SNP	Case/control (100/96)			MAF*	P †	Model ‡	OR (95% CI)	P§
FUT5	rs8108862	*CC*	*CT*	*TT*					
		67/49	29/42	2/1	0.20	0.01	D	0.47 (0.22-1.00)	0.049
FUT9	rs1325078	*GG*	*GC*	*CC*					
		23/31	53/45	24/20	0.47	0.62	D	2.29 (1.06-4.95)	0.032
GALNT1	rs6507133	*TT*	*TA*	*AA*					
		47/46	41/41	12/8	0.31	0.79	A	1.88 (1.09-3.25)	0.021
	rs11663626	*TT*	*TC*	*CC*					
		47/46	41/39	12/8	0.31	0.95	A	1.95 (1.12-3.41)	0.016
MGAT5	rs1257189	*GG*	*GA*	*AA*					
		28/29	45/49	27/18	0.47	0.74	R	2.81 (1.16-6.80)	0.020
	rs1257196	*GG*	*GA*	*AA*					
		39/29	41/48	20/18	0.42	0.81	R	2.26 (1.02-4.98)	0.040
ST3GAL3	rs3011217	*CC*	*CT*	*TT*					
		50/39	41/45	9/12	0.33	0.86	A	2.12 (1.21-3.71)	0.0068
ST8SIA2	rs3759917	*TT*	*TG*	*GG*					
		30/20	48/46	22/30	0.50	0.76	A	0.56 (0.33-0.93)	0.022
ST8SIA5	rs3889927	*CC*	*CT*	*TT*					
		66/54	28/35	6/7	0.23	0.69	A	0.54 (0.30-0.98)	0.039
	rs3897629	*AA*	*AT*	*TT*					
		46/45	42/47	11/4	0.31	0.05	A	1.90 (1.02-3.52)	0.037

*MAF estimated using both cases and controls.
† P for testing departure from HWE among controls.
‡ Genetic models (A, additive; D, dominant; R, recessive) with adjustment for age, geographic region, BMI, HRT use, oral contraceptive use, and parity/age at first birth.
§ P for testing the genetic effects before multiplicity adjustment.

significant interaction effect with environment. After adjustment for multiple comparisons, one SNP in *GALNT1* remained statistically significant and inversely associated with risk. Seven genes were significantly associated with risk among African Americans: three genes (*GALNT1, MGAT5*, and *ST3GAL3*) and one SNP (rs1257189 in *MGAT5*) were consistent with the results among the Caucasians, but the results did not hold after correction for multiple testing. To our knowledge, this is the first study of glycosylation genes as risk factors for ovarian cancer and the early results merit further investigation.

Interpretation of the current results is limited by incomplete information regarding the biological significance of the SNPs studied. During the selection of the genes and SNPs, we targeted exons and regulatory regions but in some instances necessarily included intronic SNPs to ensure good coverage of the candidate gene. We also selected every cSNP reported in the public databases, but unfortunately none of the 23 cSNPs passed the GoldenGate assay review at Illumina and were dropped. Thus, we were left with many SNPs that had no known biological function. Another issue is the absence of HWE among the Mayo controls for the candidate gene found to show the strongest association with ovarian cancer risk (*GALNT1*).

The results were consistent when analyses were restricted to the Duke subjects, careful inspection of the Illumina results provided no indication of genotype errors, and the number of SNPs that violated HWE (n = 7) is consistent with expectations based on chance. Moreover, one must also consider the biological importance of O-linked glycosylation of proteins begins with the addition of a single GalNAc monosaccharide to a serine or threonine residue on the polypeptide. Attachment is catalyzed by a UDP-*N*-α-D-galactosamine: polypeptide *N*-acetylgalactosaminyltransferase (ppGalNAc-T). During glycosylation, additional glycosyltransferases are responsible for catalyzing other types of glycosidic linkage—an extensive family of up to 24 ppGalNAc-T's has been described (19).

Despite the incomplete data on functional significance of the actual SNPs studied, the possibility that genes involved in glycosylation may be risk factors for ovarian cancer could nonetheless contribute to our understanding of the biology of this disease. Although there is a substantial literature that expression of mucins in epithelial ovarian cancer have diagnostic and prognostic value (22, 23) and may represent therapeutic targets (24), the basis for their altered expression remains unknown. There is now emerging evidence that glycans in general (25), and MUC1 in particular (22) may play a role in host defense against pathogenic molecules. Because GALNT1 is known to participate in the glycoslyation of MUC1 (26), and because MUC1 is involved in defense against pathogenic agents, it is tempting to speculate that genetic polymorphisms that encode enzymes involved in glycosylation may reflect interindividual differences to respond to an infectious agent. This may partly explain the observation that tubal ligation is inversely associated with ovarian cancer risk (27-30). Although this association has been hypothesized to reflect tubal ligation effects on hormonal mechanisms (31) or inflammation (32), the current findings lend some credence to the hypothesis raised by Walhberg (33) of an infectious etiology to ovarian cancer.

A critical issue for studies of this type for which there is no consistent agreement among the scientific community is adjustment for multiple testing. Exacerbating the debate for the current study was the *a priori* decision to fit three genetic models to the data. This turned out to be potentially important, as among the Caucasians only 1 of the 9 statistically significant SNPs (before adjustment for multiple comparisons) was based on a dominant genetic model and 6 of 10 statistically significant SNPs among the African Americans were based on an additive model. Regardless, there were several strategies and methods one could have adopted that merit some discussion. For example, adjustment could be based at a gene level or at a SNP level, with the former being a more relaxed stringency in which fewer degrees of freedom are spent. In this report, we opted for adjustment based on the number of SNPs examined (n = 93), with further correction for the fact that we examined the association of these SNPs with ovarian cancer risk as dominant, additive, or recessive effects. This led to an even greater stringency on the significance level. Although this is the most conservative approach to protect against false-positive results, there may be some true-positive findings that are being downplayed but still merit consideration in future studies.

Strengths of the study include the high participation rates, relatively large sample size, the careful inclusion/exclusion criteria, the tight quality control on the genotype data, and data on nongenetic risk factors that were comparable across sites and available for adjustment in statistical models. Despite these strengths, interpretation of the findings should consider that there were too few African American subjects for stable estimates of effect. Genotyping included a mixture of genomic and amplified DNA. Although stringent call rates were applied, some SNPs and subjects did drop out from analysis.

This decreased sample size slightly, but the dropout rates did not vary by case-control status and the concordance of native to WGA DNA was 99.16%, so there is little reason to believe that this introduced any sort of bias. The Mayo ascertainment was not population based, but the proportion of in-state cases was similar for cases and controls (52% and 58%, respectively). Moreover, residence was adjusted for in the multivariate analyses, so any imbalances were accounted for statistically. Finally, although the results may provide support for an infectious component of ovarian cancer, we did not have data on tubal ligation, which could have been instructive. Identification of the genes involved in glycosylation relied on a review of the existing literature and consultation with experts on MUC1 biology. Although we were thorough in our approach, there may be additional genes involved in glycosylation that are relevant to MUC1.

In summary, this large case-control study of epithelial ovarian cancer provides some evidence to support the role of interindividual differences in glycosylation genes as risk factors for the disease. More detailed investigation is warranted with a larger and more ethnically diverse study sample, greater coverage in genetic variation, stratification by histologic subtype, and inclusion of additional nongenetic covariates that might help further elucidate the actual pathogenesis.

References

1. Miki Y, Swensen J, Shattuck-Eidens D, et al. A strong candidate for the breast and ovarian cancer susceptibility gene BRCA1. Science 1994;266:66–71.
2. Wooster R, Bignell G, Lancaster J, et al. Identification of the breast cancer susceptibility gene BRCA2. Nature 1995;378:789–92.
3. Antoniou A, Pharoah PD, Narod S, et al. Average risks of breast and ovarian cancer associated with BRCA1 or BRCA2 mutations detected in case series unselected for family history: a combined analysis of 22 studies. Am J Hum Genet 2003;72:1117–30.
4. Ford D, Easton DF, Stratton M, et al. Genetic heterogeneity and penetrance analysis of the BRCA1and BRCA2 genes in breast cancer families. The Breast Cancer Linkage Consortium. Am J Hum Genet 1998;62:676–89.
5. Ford D, Easton DF, Bishop DT, Narod SA, Goldgar DE. Risks of cancer in BRCA1-mutation carriers. Breast Cancer Linkage Consortium. Lancet 1994;343:692–5.
6. Risch HA, McLaughlin JR, Cole DE, et al. A. Prevalence and penetrance of germline BRCA1 and BRCA2 mutations in a population series of 649 women with ovarian cancer. Am J Hum Genet 2001; 68:700–10.
7. Lynch HT, Conway T, Lynch J. Hereditary ovarian cancer. Pedigree studies. Part II. Cancer Genet Cytogenet 1991;53:161–83.
8. Lynch HT, Lemon S, Lynch J, Casey MJ. Hereditary gynecologic cancer. Cancer Treat Res 1998;95:1–102.
9. Marra G, Boland CR. Hereditary nonpolyposis colorectal cancer: the syndrome, the genes, and historical perspectives. J Natl Cancer Inst 1995;87:1114–25.
10. Boyd J, Rubin SC. Hereditary ovarian cancer: molecular genetics and clinical implications. Gynecol Oncol 1997;64:196–206.
11. Chen S, Iversen ES, Friebel T, et al. Characterization of BRCA1 and BRCA2 mutations in a large United States sample. J Clin Oncol 2006; 24:863–71.
12. Sekine M, Nagata H, Tsuji S, et al. Localization of a novel susceptibility gene for familial ovarian cancer to chromosome 3p22-25. Hum Mol Genet 2001;10:1421–9.
13. Giuntoli RL, Rodriguez GC, Whitaker RS, Dodge R, Voynow JA. Mucin gene expression in ovarian cancers. Cancer Res 1998;58: 5546–50.
14. Baldus SE, Engelmann K, Hanisch FG. MUC1 and the MUCs: a family of human mucins with impact in cancer biology. Crit Rev Clin Lab Sci 2004;41:189–231.
15. Ho SB, Niehans GA, Lyftogt C, et al. Heterogeneity of mucin gene expression in normal and neoplastic tissues. Cancer Res 1993;53: 641–51.
16. Cramer DW, Titus-Ernstoff L, McKolanis JR, et al. Conditions associated with antibodies against the tumor-associated antigen MUC1 and their relationship to risk for ovarian cancer. Cancer Epidemiol Biomarkers Prev 2005;14:1125–31.
17. Terry KL, Titus-Ernstoff L, McKolanis JR, Welch WR, Finn OJ, Cramer DW. Incessant ovulation, mucin 1 immunity, and risk for ovarian cancer. Cancer Epidemiol Biomarkers Prev 2007;16:30–5.
18. Sellers TA, Schildkraut JM, Pankratz VS, et al. Estrogen bioactivation, genetic polymorphisms, and ovarian cancer. Cancer Epidemiol Biomarkers Prev 2005;14:2536–43.
19. Ten Hagen KG, Fritz TA, Tabak LA. All in the family: the UDP-GalNAc:polypeptide *N*-acetylgalactosaminyltransferases. Glycobiology 2003;13:1–16R.
20. Schaid DJ, Rowland CM, Tines DE, Jacobson RM, Poland GA. Score tests for association between traits and haplotypes when linkage phase is ambiguous. Am J Hum Genet 2002;70:425–34.
21. Yuan HY, Chiou JJ, Tseng WH, et al. FASTSNP: an always up-to-date and extendable service for SNP function analysis and prioritization. Nucleic Acids Res 2006;34:W635–41.
22. Brayman M, Thathiah A, Carson DD. MUC1: a multifunctional cell surface component of reproductive tissue epithelia. Reprod Biol Endocrinol 2004;2:4.
23. Chauhan SC, Singh AP, Ruiz F, et al. Aberrant expression of MUC4 in ovarian carcinoma: diagnostic significance alone and in combination with MUC1 and MUC16 (CA125). Mod Pathol 2006; 19:1386–94.
24. Gendler SJ. MUC1, the renaissance molecule. J Mammary Gland Biol Neoplasia 2001;6:339–53.
25. Hooper LV, Gordon JI. Glycans as legislators of host-microbial interactions: spanning the spectrum from symbiosis to pathogenicity. Glycobiology 2001;11:1–10R.
26. Wandall HH, Hassan H, Mirgorodskaya E, et al. Substrate specificities of three members of the human UDP-*N*-acetyl-α-D-galactosamine:polypeptide *N*- acetylgalactosaminyltransferase family, GalNAc-T1, -T2, and -T3. J Biol Chem 1997;272:23503–14.
27. Mori M, Harabuchi I, Miyake H, Casagrande JT, Henderson BE, Ross RK. Reproductive, genetic, and dietary risk factors for ovarian cancer. Am J Epidemiol 1988;128:771–7.
28. Hankinson SE, Hunter DJ, Colditz GA, et al. Tubal ligation, hysterectomy, and risk of ovarian cancer. A prospective study. JAMA 1993;270:2813–8.
29. Pati S, Cullins V. Female sterilization. Evidence. Obstet Gynecol Clin North Am 2000;27:859–99.
30. Whittemore AS, Wu ML, Paffenbarger RS, Jr., et al. Personal and environmental characteristics related to epithelial ovarian cancer. II. Exposures to talcum powder, tobacco, alcohol, and coffee. Am J Epidemiol 1988;128:1228–40.
31. Rosenblatt KA, Thomas DB. Reduced risk of ovarian cancer in women with a tubal ligation or hysterectomy. The World Health Organization Collaborative Study of Neoplasia and Steroid Contraceptives. Cancer Epidemiol Biomarkers Prev 1996;5:933–5.
32. Ness RB, Grisso JA, Cottreau C, et al. Factors related to inflammation of the ovarian epithelium and risk of ovarian cancer. Epidemiology 2000;11:111–7.
33. Wahlberg C. Tubal ligation, hysterectomy, and risk of ovarian cancer. JAMA 1994;271:1236; author reply 1236–7.