Prospective evaluation of antibody response to *Streptococcus gallolyticus* and risk of colorectal cancer

Authors
Julia Butt*[1], Mazda Jenab*[2], Martina Willhauck-Fleckenstein[1], Angelika Michel[1], Michael Pawlita[1], Cecilie Kyrø[3], Anne Tjønneland[3], Marie-Christine Boutron-Ruault[4, 5], Franck Carbonnel[4, 5, 6], Gianluca Severi[4, 5], Rudolf Kaaks[7], Tilman Kühn[7], Heiner Boeing[8], Antonia Trichopoulou[9, 10], Carlo la Vecchia[9, 11], Anna Karakatsani[9, 12], Salvatore Panico[13], Rosario Tumino[14], Claudia Agnoli[15], Domenico Palli[16], Carlotta Sacerdote[17], Bas Bueno-de-Mesquita[18, 19, 20, 21], Elisabete Weiderpass[22, 23, 24, 25], Maria-José Sánchez[26, 27], Catalina Bonet Bonet[28], José María Huerta[27, 29], Eva Ardanaz[27, 30, 31], Kathryn Bradbury[32], Marc Gunter[2], Neil Murphy[2], Heinz Freisling[2], Elio Riboli[20], Kostas Tsilidis[20, 33], Dagfinn Aune[20], Tim Waterboer [#1], David Hughes [#34 +]
[*]Shared first authorship, [#]Shared last authorship, [+]Corresponding Author
Institutions addresses
[1]Division of Molecular Diagnostics of Oncogenic Infections, German Cancer Research Center (DKFZ), Heidelberg, Germany
[2]Nutritional Epidemiology Group, International Agency for Research on Cancer, Lyon, France
[3] Diet, Genes and Environment Unit, Danish Cancer Society Research Center, Copenhagen, Denmark
[4] CESP, INSERM U1018, Université Paris-Sud, UVSQ, Université Paris-Saclay, Villejuif Cedex, F-94805, France
[5] Institute Gustave Roussy, Villejuif, France
[6] Université Paris Sud and Gastroenterology Unit, Hopitaux Universitaires Paris Sud, CHU de Bicetre, AP-HP, Le Kremlin Bicetre, France
[7] Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany
[8] Department of Epidemiology, German Institute of Human Nutrition, Potsdam-Rehbrücke, Germany
[9] Hellenic Health Foundation, Athens, Greece
[10] WHO Collaborating Center for Nutrition and Health, Unit of Nutritional Epidemiology and Nutrition in Public Health, Department of Hygiene, Epidemiology and Medical Statistics, School of Medicine, National and Kapodistrian University of Athens, Greece
[11] Department of Clinical Sciences and Community Health, Universita degli Studi dei Milano
[12]2nd Pulmonary Medicine Department, School of Medicine, National and Kapodistrian University of Athens, "ATTIKON" University Hospital, Haidari, Greece
[13] Dipartamento di Medicina Clinica e Chirugia, Federico II University, Naples, Italy
[14] Cancer Registry and Histopathology Unit, 'Civic-M.P. Arezzo' Hospital, Ragusa, Italy
[15] Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori, Milano, Italy
[16]Cancer Risk Factors and Life-Style Epidemiology Unit, Cancer Research and Prevention Institute – ISPO, Florence, Italy
[17] Unit of Cancer Epidemiology, Città della Salute e della Scienza University-Hospital and Center for Cancer Prevention (CPO), Turin, Italy
[18] Department of Determinants of Chronic Diseases (DCD), National Institute for Public Health and Environment (RIVM), Bilthoven, Netherlands
[19] Department of Gastroenterology and Hepatology, University Medical Centre, Utrecht, Netherlands

[20] Department of Epidemiology and Biostatistics, School of Public Health, Imperial College London, London, UK
[21] Department of Social & Preventive Medicine, Faculty of Medicine, University of Malaya, Kuala Lumpur, Malaysia
[22]Department of Community Medicine, Faculty of Health Sciences, University of Tromsø, The Arctic University of Norway, Tromsø, Norway
[23]Department of Research, Cancer Registry of Norway, Institute of Population-Based Cancer Research, Oslo, Norway
[24] Department of Medical Epidemiology and Biostatistics, Karolinska Institut, Stockholm, Sweden
[25] Genetic Epidemiology Group, Folkhälsan Research Center, Helsinki, Finland
[26] Escuela Andaluza de Salud Pública. Instituto de Investigacion Biosanitaria ibs. Granada. Hospitales Universitarios de Granada/Universidad de Granada, Granada, Spain
[27] CIBER de Epidemiología y Salud Pública (CIBERESP), Madrid, Spain
[28] Unit of Nutrition and Cancer, Cancer Epidemiology Research Program, Institut Català d'Oncologia, Llobregat, Spain
[29] Department of Epidemiology, Murcia Regional Health Council, IMIB-Arrixaca, Murcia, Spain
[30] Navarra Public Health Institute, Pamplona, Spain
[31] IdiSNA, Navarra Institute for Health Research, Pamplona, Spain
[32] Cancer Epidemiology Unit, Nuffield Department of Population Health, University of Oxford, Oxford, UK
[33] Department of Hygiene and Epidemiology, University of Ioannina school of Medicine, Ioannina, Greece
[34] Cancer Biology and Therapeutics Group, UCD Conway Institute, University College Dublin, Dublin, Ireland

Running title

Serology of *Streptococcus gallolyticus* and incident colorectal cancer

Keywords

Colorectal neoplasms, *Streptococcus gallolyticus*, bacterial serology, antibodies, prospective cohort

Financial support

David J Hughes was supported by the Health Research Board of Ireland project grants HRA_PHS/2013/397 and HRA_PHS/2015/1142.
The coordination of EPIC is financially supported by the European Commission (DG-SANCO) and the International Agency for Research on Cancer. The national cohorts are supported by the Danish Cancer Society (Denmark); Ligue Contre le Cancer, Institut Gustave Roussy, Mutuelle Générale de l'Education Nationale, and Institut National de la Santé et de la Recherche Médicale (INSERM) (France); Deutsche Krebshilfe, Deutsches Krebsforschungszentrum, and Federal Ministry of Education and Research (Germany); the Hellenic Health Foundation (Greece); the Sicilian Government, AIRE ONLUS Ragusa, AVIS Ragusa, Associazione Italiana per la Ricerca sul Cancro-AIRC-Italy, and National Research Council (Italy); Dutch Ministry of Public Health, Welfare and Sports (VWS), Netherlands Cancer Registry (NKR), LK Research Funds, Dutch Prevention Funds, Dutch ZON (Zorg Onderzoek Nederland), World Cancer Research Fund (WCRF), and Statistics Netherlands (the Netherlands); Nordic Centre of Excellence programme on Food, Nutrition and Health. (Norway); Health Research Fund (FIS), PI13/00061 to Granada), Regional Governments of Andalucía, Asturias, Basque Country, Murcia (no. 6236) and Navarra, Regional Government of Asturias (Asturias, Spain), and ISCIII RETIC (RD06/0020) (Spain); Swedish Cancer Society, Swedish Scientific Council, and County Councils of

Skåne and Västerbotten (Sweden); Cancer Research UK (14136 to EPIC-Norfolk; C570/A16491 to EPIC-Oxford) and Medical Research Council (1000143 to EPIC-Norfolk) (United Kingdom); and a Girdlers' New Zealand Health Research Council Fellowship (to Dr. K.E. Bradbury). The funding sources had no influence on the design of the study; the collection, analysis, and interpretation of data; the writing of the report; or the decision to submit the paper for publication.

Address corresponding author

[+]David J Hughes,
Cancer Biology and Therapeutics Group,
UCD Conway Institute,
University College Dublin,
Dublin 4, Ireland
T: (+353 1) 716 6700
F: (+353 1) 716 6701
Email: david.hughes@ucd.ie

Conflict of interest disclosure

The authors declare no potential conflicts of interest.

Number of Words/Allowed number of words

Abstract	245/250
Introduction	452
Materials and Methods	829
Results	417
Discussion	1052
Sum	2750

Number of references

26

Number of tables and figures

5

Number of supplemental tables and figures

2

Abbreviations

CI – Confidence interval
CRC – Colorectal cancer
EPIC – European Prospective Investigation into Nutrition and Cancer
MFI – Median fluorescence intensity
OMP – Outer membrane protein
OR – Odds ratio
SGG – *Streptococcus gallolyticus* subspecies *gallolyticus*

Novelty and Impact

This study presents the serological analysis of an association of *Streptococcus gallolyticus* sub-species *gallolyticus* (SGG) with CRC in a prospective setting with samples from the large European Prospective Investigation into Nutrition and Cancer (EPIC) study. We found for the first time an association of antibody responses to SGG proteins with CRC risk already detectable with pre-diagnostic samples. The presented results support that SGG serology might provide a specific marker for risk of developing CRC.

Abstract

The gut microbiome is increasingly implicated in colorectal cancer (CRC) development. A subgroup of patients diagnosed with CRC show high antibody responses to *Streptococcus gallolyticus* subspecies *gallolyticus* (SGG). However, it is unclear whether the association is also present pre-diagnostically. We assessed the association of antibody responses to SGG proteins in pre-diagnostic serum samples with CRC risk in a case-control study nested within a prospective cohort.

Pre-diagnostic serum samples from 485 first incident CRC cases (mean time between blood draw and diagnosis 3.4 years) and 485 matched controls in the European Prospective Investigation into Nutrition and Cancer (EPIC) study were analyzed for antibody responses to eleven SGG proteins using multiplex serology. Odds ratios (OR) and 95 % confidence intervals (CI) were estimated using multivariable conditional logistic regression models.

Antibody positivity for any of the eleven SGG proteins was significantly associated with CRC risk with 56% positive controls compared to 63% positive cases (OR: 1.36, 95% CI: 1.04-1.77). Positivity for two or more proteins of a previously identified SGG 6-marker panel with greater CRC-specificity was also observed among 9% of controls compared to 17% of CRC cases, corresponding to a significantly increased CRC risk (OR: 2.17, 95% CI: 1.44-3.27).

In this prospective nested case-control study we observed a positive association between antibody responses to SGG and CRC development in serum samples taken pre-diagnostically. Further work is required to establish the possibly etiological significance of these observations and whether SGG serology may be applicable for CRC risk stratification.

Introduction

Colorectal cancer (CRC) is among the most frequently diagnosed cancers worldwide with an incidence of 746,000 new cases among men and 614,000 new cases among women in 2012[1]. Inflammation is thought to be among the major etiological risk factors for the development of CRC, and is a possible mechanism through which bacterial infections might contribute to carcinogenesis[2]. Changes in the local intestinal tissue can compromise the colonic barrier integrity resulting in a "leaky gut"[3]. Certain bacteria may opportunistically infect the intestinal tissue and potentially induce an immune response, although they usually act as commensals[4].

An interesting candidate in this respect might be the intestinal commensal *Streptococcus gallolyticus* subspecies *gallolyticus* (SGG), formerly known as *Streptococcus bovis* biotype I. In the 1970's it was found that infective endocarditis caused by bacteria belonging to the *S. bovis* complex[5-7], and later more specifically by the subspecies SGG[8], co-occurred with the presence of colorectal adenoma. A systematic review of CRC case series by Boleij et al. in 2011 showed that 60% of *S. bovis*-infected individuals in the reviewed studies had a concomitant colorectal adenoma/carcinoma and that SGG-infection was specifically responsible for this association compared to other *S. bovis* subtypes[9]. It is hypothesized that intestinal lesions are the entry port for the commensal SGG to the blood stream enabling the bacterium to turn pathogenic and cause bacteremia or endocarditis[10]. Antibodies against the infecting SGG may serve as markers for the presence of colorectal neoplasia. A significant association between SGG antibody response and presence of CRC has been observed in several studies, but to date these have been exclusively case-control designs with prevalent CRC cases[11-14].

We previously applied multiplex serology, a fluorescent bead-based high-throughput technology allowing serological typing of several antigens in one reaction[15], to analyze antibody responses to eleven SGG proteins in a German CRC case-control study we showed

that positivity to two or more proteins of a SGG 6-marker panel (Gallo0272, Gallo0748, Gallo1675, Gallo2018, Gallo2178 and Gallo2179) was associated with a 1.8-fold (95% CI: 1.07-3.06) increased risk for CRC (n=318) compared to controls (n=228)[16]. The 6-marker panel demonstrated a higher specificity for CRC risk compared to positivity towards any one of the eleven SGG proteins included in the multiplex serology panel. [16]. These previous findings were based on traditional case-control designs where blood samples were obtained post-diagnosis. It is currently unknown whether any antibody responses to SGG are associated with CRC development at various time points prior to diagnosis, i.e. whether SGG infection is merely a consequence of the disease or is in some way involved in CRC etiology[17].

In the current study, we evaluated whether antibody responses to SGG proteins, as measured by multiplex serology, in pre-diagnostic serum samples were associated with the risk of CRC, using serum samples of a case-control subset (485 cases and 485 matched controls) of participants of the European Prospective Investigation into Cancer and Nutrition (EPIC) study.

Methods

Study population, case ascertainment and control selection

EPIC is a multinational cohort to investigate the relation between diet, lifestyle and environmental factors with cancer incidence. A detailed description of study design has been published elsewhere[18]. Briefly, 521,468 participants, aged 35 to 70 years, were enrolled from 10 different European countries (Denmark, France, Greece, Germany, Italy, Netherlands, Norway, Spain, Sweden, United Kingdom) between 1992 and 2000.

Dietary and lifestyle data as well as biological samples, including serum, were collected at enrollment. The blood collection and processing protocols were standardized across the study centers and blood processing and separation was conducted prior to freezing. Serum samples

were stored at the International agency for research on cancer (IARC, Lyon, France) at -196°C. For multiplex serology analyses, serum samples were shipped on dry ice to the German Cancer Research Center, Heidelberg.

The nested CRC case-control study analyzed here included pre-diagnostic serum samples from 492 incident CRC cases (primary tumors, C18-C20 as by the 10th Revision of the International Statistical Classification of Diseases, Injury and Causes of Death) and 492 matched controls. Controls were selected by incidence density sampling from all cohort members alive and free of cancer at the time of matching. Cases and controls were matched 1:1 by age at blood collection (±6 month to ±2 years), sex, study center, time of the day at blood collection (±2 to 4 hours interval), fasting status at blood collection (<3/3-6 hours); among women by menopausal status, and among premenopausal women, by phase of menstrual cycle and hormone replacement therapy use at time of blood collection. After exclusion of 7 case-control pairs due to technical errors during detection, a total of 485 first incident CRC cases (colon n=432, rectum n=53) were identified that had a mean time between blood draw and diagnosis of 3.4 years (range 0.4 to 8.5 years)

SGG multiplex serology

Serum samples were analyzed for antibody responses against SGG in a final sample dilution of 1:1000 using multiplex serology. The method is described in detail elsewhere[15, 16]. Briefly, eleven SGG antigens (strain UCN34, Table 1) were bacterially expressed as recombinant GST-X-tag fusion proteins and each antigen was affinity-purified on a different bead set marked with a distinct internal fluorescent color (SeroMap, Luminex Corp., Austin, TX, USA). These differently loaded bead sets were mixed to form a suspension antigen array for serum presentation. A Luminex xMAP (Luminex Corp., Austin, TX, USA) analyzer identified the bead set and simultaneously quantified bound serum IgA, IgM and IgG antibodies by a reporter fluorescent conjugate, Streptavidin-R-Phycoerythrin. The level of

antibody response was given as median fluorescence intensity (MFI) on at least 100 beads per set. Net MFI values were generated by subtraction of bead-background and GST-tag background MFI values.

Due to lack of an appropriate serological gold standard for comparison with our assay, the cut-off definition for SGG antibody positivity was arbitrarily defined, as described elsewhere[16]. The distribution of antibody responses (MFI) to all eleven SGG proteins among controls was skewed towards low MFI, especially when compared to the outer membrane protein (OMP) of *Helicobacter pylori* (*H. pylori*), analyzed in the same experimental setting (Fig 1): the upper quartile of antibody responses does not exceed 100 MFI for any of the SGG antigens, whereas this antibody level was exceeded by 50% of the control sera to *H. pylori* OMP. Among controls, we compared the frequency of individuals with the highest antibody responses (upper 10^{th} percentile) to each protein with the frequency of individuals exceeding the same level of antibody response among cases. The technical minimum cut-off was 30 MFI (Table 1). Overall SGG positivity was defined as samples giving a high response to any of the eleven SGG proteins to allow for individual immune responses and infection with different strains. In a previous case-control study, we showed that refining the algorithm for overall SGG positivity to two or more proteins in a 6-marker panel (Gallo0272, Gallo0748, Gallo1675, Gallo2018, Gallo2178, Gallo2179) strengthened the association with CRC[16]. This algorithm was also applied here as a second definition for SGG positivity.

Statistical analyses

Risk factors for SGG positivity among controls were analyzed using Chi-squared-tests. We estimated the association of incident CRC with antibody responses to individual SGG proteins, positivity to any of the eleven SGG proteins, or 2 or more proteins of the 6-marker panel[16] using conditional logistic regression models to compute odds ratios (OR) and 95% confidence intervals (95% CI). A p-value below 0.05 was considered statistically significant.

Statistical models were first run conditioned on the matching factors, and subsequently with multivariable adjustment for the following variables: level of education attainment, BMI, smoking status and level of alcohol consumption [g/day] at baseline assessment. Missing observations in these variables were included in the model as individual category to save statistical power. The resulting risk estimates did not substantially differ from those calculated without further adjustment (supplementary table S1). Sensitivity analyses were carried out excluding cases diagnosed within two years after blood draw to assess the potential for reverse causation.

Explorative sub-group analyses were conducted by sex, age at blood draw applying interaction analyses, as well as by anatomical sub-site.

All statistical analyses were performed with SAS version 9.4 (SAS Institute).

Results

Study characteristics and risk factors for SGG positivity

There were no significant differences between cases and controls in any of the baseline characteristics (Table 2).

The comparison of SGG positive versus negative control subjects did not identify any major determinants of SGG positivity (Supplementary table S2).

Association of antibody responses to SGG with CRC risk

The risk of developing CRC was significantly increased with positivity to any of the eleven SGG proteins (OR: 1.36, 95% CI: 1.04-1.77), and also positivity to individual SGG proteins Gallo0272 (OR: 1.59, 95% CI: 1.06-2.40), Gallo0748 (OR: 1.49, 95% CI: 1.02-2.16) and Gallo2178 (OR: 3.01, 95% CI: 1.49-6.08) (Table 3). Positivity for two or more proteins of the

previously identified 6-marker panel (Gallo0272, Gallo0748, Gallo1675, Gallo2018, Gallo2178 and Gallo2179)[16] was also significantly associated with increased CRC risk (OR: 2.17, 95% CI: 1.44-3.27) with 9% positive controls compared to 17% positive cases.

To assess the potential impact of reverse causation, we performed a sensitivity analysis excluding those cases diagnosed within 2 years after blood draw and their respective controls (Table 3). The association for positivity to any of the eleven SGG proteins (OR: 1.38, 95% CI: 1.02-1.87) as well as positivity to two or more proteins of the 6-marker panel (OR: 2.07, 95% CI: 1.29-3.31) with CRC risk was generally unaltered. Positivity to individual proteins Gallo0272 (OR: 1.87, 95% CI: 1.15-3.05) and Gallo2178 (OR: 3.28, 95% CI: 1.25-8.57) retained statistical significance while Gallo0748 lost significance but with little change in the magnitude of the risk estimate (OR: 1.40, 95% CI: 0.90-2.18).

Explorative subgroup analyses

Positivity for two or more proteins of the 6-marker panel was associated with only a minor fraction of CRC cases (17%). We assessed whether particular subgroups showed different risk associations for CRC. Analyses stratified by age at blood draw and sex did not reveal any statistically significant difference between the subgroups.

Separate analyses by colon or rectal sub-site showed different associations (Fig 2). Positivity to two or more proteins of the 6-marker panel was associated with a 10-fold increased risk of rectal cancer (95% CI: 1.05-95.78) and a much lower, but also statistically significant, near two-fold higher risk for colon cancer (OR: 1.96, 95% CI: 1.28-3.00). However, it is important to note that the number of rectal cancers was small (n=53) resulting in wide confidence intervals and imprecision of the risk estimate.

Discussion

In this CRC case-control study nested within the prospective multinational EPIC cohort we found that antibody responses to SGG proteins, in particular to two or more proteins seropositive among a 6-marker panel (Gallo0272, Gallo0748, Gallo1675, Gallo2018, Gallo2178 and Gallo2179) were significantly associated with risk of developing CRC.

These findings replicate and expand previous findings from two case-control studies with CRC cases from Spain (multicenter case-control study (MCC Spain))[14] and an independent German study[16]. In MCC Spain, an association of prevalent CRC with antibody responses to SGG protein Gallo2178 alone and Gallo2178 in combination with Gallo2179 was found[14]. In the German case-control study, the SGG multiplex serology panel was extended to eleven SGG proteins. Positivity to any of these proteins was associated with prevalent CRC. Seropositivity for at least two proteins from a 6-marker panel subset (Gallo0272, Gallo0748, Gallo1675, Gallo2018, Gallo2178 and Gallo2179) was more specifically associated with CRC (19% SGG positives) compared to controls (11% of SGG-positives)[16].

It is currently unknown whether SGG infects colon tissue before or after initiation of tumor development. However, it is hypothesized that the commensal SGG enters the bloodstream through a leaky epithelium, arising due to various environmental exposures, or along the processes of CRC development[4]. This hypothesis is supported by observations showing the presence of SGG already in early colorectal lesions, including polyps and adenoma[11, 12, 16, 19]. Here, we offer the first prospective observational evidence to support early involvement of SGG in colon carcinogenesis by showing that antibody responses to SGG were more frequently present in subjects who later developed CRC even more than two years after blood draw than those who remained disease-free during the same time-frame. The natural history of CRC is characterized by the progressive development of neoplasia of the colon mucosa and can take up to 10-15 years from an initial polyp to tumor diagnosis. Therefore, it is likely that

a number of individuals in this study, who developed CRC, already had a precancerous lesion at the time of recruitment into the cohort, but were undiagnosed and likely asymptomatic.

Although we have no data on CRC screening to estimate the numbers with existing polyps, it is likely to be comparable to other European population studies, such as for Germany where the detection rate of non-advanced and advanced adenoma was 22.3% and 9.0%, respectively, among males and 14.9% and 5.2%, respectively, among females above age 55 years[20]. As only a minority of adenomas progress to cancer, a similar proportion of the controls would also be expected to have some form of colorectal adenoma at blood draw that had not progressed to malignant disease by the end of follow-up. Thus, the finding that antibody responses to SGG appear prior to cancer diagnosis raises the question whether SGG infection is a potential etiological factor in the transition of an adenoma to malignant disease and whether its detection could help stratify the risk for tumor progression from a precancerous lesion. However, we were unable to directly address this question within the limitations of our study. Studies by Abdulamir et al. found pro-inflammatory cytokine profiles in human CRC tissue positive for SGG DNA and support the hypothesis of an involvement of SGG in tumor progression[11, 21]. A recent study comprehensively showed that SGG promotes proliferation of colon cancer cells in vitro and tumor development in a mouse model overall supporting a role of SGG in colonic tumorigenesis[22]. Our observations will hopefully stimulate further epidemiological studies with CRC screening data and mechanistic investigations of the potential SGG induced transformation of benign polyps to more advanced disease states.

The antigens selected for SGG multiplex serology include proteins predicted to be present at the cell wall of the bacterium or to be secreted[23, 24]. Pilus proteins Gallo2178 and Gallo2179, both included in the 6-marker panel, were previously shown to be potential virulence factors in endocarditis and for infection of colon tumor tissue by mediating attachment to collagen in tissue[10, 25]. Functions of the other proteins had been so far only predicted by sequence

comparison to proteins of other bacteria and include enzymatic (Gallo0112A/B, Gallo0748, Gallo0933, Gallo2018) as well as adhesion functions (Gallo0272, Gallo0577, Gallo1570). The function of Gallo1675 is unknown[26]. Future studies should focus on this 6-marker panel as it is a stronger marker for CRC risk than being positive to any of the eleven proteins included in the multiplex serology (OR: 2.17 vs OR: 1.36, respectively).

Stratification by age and sex did not reveal statistically significant differences. However, the small sub-group sample sizes may have limited the analysis. Secondary sub-group analysis by anatomical sub-site suggested a stronger cancer risk association for the rectum versus the colon with SGG antibody responses. This observation is highly interesting and warrants further investigation, but is limited due to small number of rectal cancer cases (n=53) included in the present analysis. The disparity between the number of colon and rectal cancer cases analyzed in this study are due to limited availability of biological samples for the required laboratory analyses in this sub-set of EPIC CRC cases.

Key advantages of this study are its prospective setting, multi-center design and the use of a detailed, validated biomarker approach to assess SGG exposures. A main limitation is the small sample size, being based on a subset of CRC cases in the EPIC cohort with available biological samples for the required SGG biomarker analyses. Furthermore, the SGG exposures assessed here reflect levels at recruitment into the cohort upon blood collection and so may not pertain to longer term exposures. An additional potential limitation applicable to all observational studies is the possibility for residual or uncontrolled confounding. Although, the EPIC data have been very well measured and validated, the possibility of residual confounding cannot ever be wholly discounted. Uncontrolled confounding is unlikely because the multivariate adjusted model presented here addressed a large number of potentially important confounding variables. Nevertheless, in the absence of further confirmation of these

findings from a larger series of CRC cases from EPIC or from other prospective cohorts, caution in the interpretation of the findings is necessary.

In conclusion, this study provides the first exploration in a prospective setting of the association between SGG infection and risk of CRC development. Our observations indicate a positive association of antibody responses to SGG proteins with CRC risk, taking into account other important confounding factors. SGG infection, possibly acquired through lifestyle exposures leading to colonic epithelial barrier dysfunction, may be an important etiological component of CRC development. Thus, antibody responses to SGG proteins may be indicative for individuals at increased risk for developing CRC.

Acknowledgments

We thank Ute Koch, Monika Oppenländer and Claudia Brandel for excellent technical assistance with the serological measurements.

References

1. Ferlay J SI, Ervik M, Dikshit R, Eser S, Mathers C, Rebelo M, Parkin DM, Forman D, Bray, F.GLOBOCAN 2012 v1.0, .Lyon, France: International Agency for Research on Cancer;. Cancer Incidence and Mortality Worldwide: IARC CancerBase No. 11 [Internet], vol. 2016, 2013.

2. Brenner H, Kloor M, Pox CP. Colorectal cancer. *Lancet* 2014;**383**: 1490-502.

3. Kong SY, Tran HQ, Gewirtz AT, McKeown-Eyssen G, Fedirko V, Romieu I, Tjonneland A, Olsen A, Overvad K, Boutron-Ruault MC, Bastide N, Affret A, et al. Serum Endotoxins and Flagellin and Risk of Colorectal Cancer in the European Prospective Investigation into Cancer and Nutrition (EPIC) Cohort. *Cancer epidemiology, biomarkers & prevention : a publication of the American Association for Cancer Research, cosponsored by the American Society of Preventive Oncology* 2016;**25**: 291-301.

4. Tjalsma H, Boleij A, Marchesi JR, Dutilh BE. A bacterial driver-passenger model for colorectal cancer: beyond the usual suspects. *Nature reviews Microbiology* 2012;**10**: 575-82.

5. Klein RS, Recco RA, Catalano MT, Edberg SC, Casey JI, Steigbigel NH. Association of Streptococcus bovis with carcinoma of the colon. *The New England journal of medicine* 1977;**297**: 800-2.

6. Murray HW, Roberts RB. Streptococcus bovis bacteremia and underlying gastrointestinal disease. *Archives of internal medicine* 1978;**138**: 1097-9.

7. Noble CJ, Uttley AH, Falk RH, Richardson PJ. Streptococcus bovis endocarditis and colonic cancer. *Lancet* 1978;**1**: 766.

8. Ruoff KL, Miller SI, Garner CV, Ferraro MJ, Calderwood SB. Bacteremia with Streptococcus bovis and Streptococcus salivarius: clinical correlates of more accurate identification of isolates. *J Clin Microbiol* 1989;**27**: 305-8.

9. Boleij A, van Gelder MM, Swinkels DW, Tjalsma H. Clinical Importance of Streptococcus gallolyticus infection among colorectal cancer patients: systematic review and meta-analysis. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 2011;**53**: 870-8.

10. Boleij A, Muytjens CM, Bukhari SI, Cayet N, Glaser P, Hermans PW, Swinkels DW, Bolhuis A, Tjalsma H. Novel clues on the specific association of Streptococcus gallolyticus subsp gallolyticus with colorectal cancer. *J Infect Dis* 2011;**203**: 1101-9.

11. Abdulamir AS, Hafidh RR, Mahdi LK, Al-jeboori T, Abubaker F. Investigation into the controversial association of Streptococcus gallolyticus with colorectal cancer and adenoma. *BMC cancer* 2009;**9**: 403.

12. Garza-Gonzalez E, Rios M, Bosques-Padilla FJ, Francois F, Cho I, Gonzalez GM, Perez-Perez GI. Immune response against Streptococcus gallolyticus in patients with adenomatous polyps in colon. *International journal of cancer Journal international du cancer* 2012;**131**: 2294-9.

13. Boleij A, Roelofs R, Schaeps RM, Schulin T, Glaser P, Swinkels DW, Kato I, Tjalsma H. Increased exposure to bacterial antigen RpL7/L12 in early stage colorectal cancer patients. *Cancer* 2010;**116**: 4014-22.

14. Butt J, Romero-Hernandez B, Perez-Gomez B, Willhauck-Fleckenstein M, Holzinger D, Martin V, Moreno V, Linares C, Dierssen-Sotos T, Barricarte A, Tardon A, Altzibar JM, et al. Association of Streptococcus gallolyticus subspecies gallolyticus with colorectal cancer: Serological evidence. *International journal of cancer Journal international du cancer* 2016;**138**: 1670-9.

15. Waterboer T, Sehr P, Michael KM, Franceschi S, Nieland JD, Joos TO, Templin MF, Pawlita M. Multiplex human papillomavirus serology based on in situ-purified glutathione s-transferase fusion proteins. *Clinical chemistry* 2005;**51**: 1845-53.

16. Butt J, Werner S, Willhauck-Fleckenstein M, Michel A, Waterboer T, Zornig I, Boleij A, Dramsi S, Brenner H, Pawlita M. Serology of Streptococcus gallolyticus subspecies gallolyticus and its association with colorectal cancer and precursors. *International journal of cancer Journal international du cancer* 2017.

17. Mai V, Morris JG, Jr. Need for prospective cohort studies to establish human gut microbiome contributions to disease risk. *Journal of the National Cancer Institute* 2013;**105**: 1850-1.

18. Riboli E, Hunt KJ, Slimani N, Ferrari P, Norat T, Fahey M, Charrondiere UR, Hemon B, Casagrande C, Vignat J, Overvad K, Tjonneland A, et al. European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public health nutrition* 2002;**5**: 1113-24.

19. Paritsky M, Pastukh N, Brodsky D, Isakovich N, Peretz A. Association of Streptococcus bovis presence in colonic content with advanced colonic lesion. *World journal of gastroenterology : WJG* 2015;**21**: 5663-7.

20. Brenner H, Altenhofen L, Kretschmann J, Rosch T, Pox C, Stock C, Hoffmeister M. Trends in Adenoma Detection Rates During the First 10 Years of the German Screening Colonoscopy Program. *Gastroenterology* 2015;**149**: 356-66 e1.

21. Abdulamir AS, Hafidh RR, Bakar FA. Molecular detection, quantification, and isolation of Streptococcus gallolyticus bacteria colonizing colorectal tumors: inflammation-driven potential of carcinogenesis via IL-1, COX-2, and IL-8. *Molecular cancer* 2010;**9**: 249.

22. Kumar R, Herold JL, Schady D, Davis J, Kopetz S, Martinez-Moczygemba M, Murray BE, Han F, Li Y, Callaway E, Chapkin RS, Dashwood WM, et al. Streptococcus gallolyticus subsp. gallolyticus promotes colorectal tumor development. *PLoS pathogens* 2017;**13**: e1006440.

23. Rusniok C, Couve E, Da Cunha V, El Gana R, Zidane N, Bouchier C, Poyart C, Leclercq R, Trieu-Cuot P, Glaser P. Genome sequence of Streptococcus gallolyticus: insights into its adaptation to the bovine rumen and its ability to cause endocarditis. *Journal of bacteriology* 2010;**192**: 2266-76.

24. Sillanpaa J, Nallapareddy SR, Qin X, Singh KV, Muzny DM, Kovar CL, Nazareth LV, Gibbs RA, Ferraro MJ, Steckelberg JM, Weinstock GM, Murray BE. A collagen-binding adhesin, Acb, and ten other putative MSCRAMM and pilus family proteins of Streptococcus gallolyticus subsp. gallolyticus (Streptococcus bovis Group, biotype I). *Journal of bacteriology* 2009;**191**: 6643-53.

329 25. Danne C, Entenza JM, Mallet A, Briandet R, Debarbouille M, Nato F, Glaser P, Jouvion G,
330 Moreillon P, Trieu-Cuot P, Dramsi S. Molecular characterization of a Streptococcus gallolyticus
331 genomic island encoding a pilus involved in endocarditis. *J Infect Dis* 2011;**204**: 1960-70.
332 26. Hinse D, Vollmer T, Ruckert C, Blom J, Kalinowski J, Knabbe C, Dreier J. Complete genome
333 and comparative analysis of Streptococcus gallolyticus subsp. gallolyticus, an emerging pathogen of
334 infective endocarditis. *BMC genomics* 2011;**12**: 400.

335

Supplementary Table S1: Antibody responses to SGG proteins and protein combinations in relation to CRC risk in a nested case-control study within EPIC

	Positive n (%)		Unadjusted model[1]		Adjusted model [2]	
	Controls n=485	Cases n=485	OR	95% CI	OR	95% CI
Gallo0112A	33 (7)	37 (8)	1.14	0.69-1.90	1.09	0.64-1.84
Gallo0112B	28 (6)	26 (5)	0.93	0.54-1.60	0.96	0.55-1.67
Gallo0272	47 (10)	67 (14)	1.49	1.00-2.21	1.59	1.06-2.40
Gallo0577	47 (10)	49 (10)	1.05	0.69-1.59	1.03	0.67-1.59
Gallo0748	50 (10)	74 (15)	1.51	1.05-2.18	1.49	1.02-2.16
Gallo0933	49 (10)	44 (9)	0.89	0.58-1.36	0.92	0.59-1.43
Gallo1570	47 (10)	52 (11)	1.13	0.73-1.74	1.13	0.72-1.76
Gallo1675	48 (10)	51 (11)	1.07	0.70-1.63	1.08	0.70-1.67
Gallo2018	47 (10)	54 (11)	1.16	0.77-1.74	1.24	0.81-1.89
Gallo2178	12 (2)	31 (6)	2.58	1.33-5.03	3.01	1.49-6.08
Gallo2179	47 (10)	64 (13)	1.43	0.95-2.14	1.48	0.97-2.24
Any SGG protein	273 (56)	306 (63)	1.32	1.02-1.71	1.36	1.04-1.77
$\geq$ 2 of 6-marker panel[3]	45 (9)	83 (17)	2.03	1.37-3.01	2.17	1.44-3.27

[1]Conditional logistic regression model conditioned on the matching factors; [2] Model 1 with further adjustment for BMI, highest level of education attainment, smoking status and alcohol intake at baseline as categorical variables, missings in the variables considered as individual category; [3]Gallo0272, Gallo0748, Gallo1675, Gallo2018, Gallo2178, Gallo2179

Supplementary Table S2: Comparison of SGG negative and positive individuals for demographic and other risk factors among controls.

		Any *S. gallolyticus* protein			≥2 of 6-marker panel[2]		
		neg (n=212)	pos (n=273)	p-value[1]	neg (n=440)	pos (n=45)	p-value[1]
Sex	female	103 (49)	144 (53)		225 (51)	22 (49)	
	male	109 (51)	129 (47)	0.363	215 (49)	23 (51)	0.774
Age at blood draw, years	37-55	44 (21)	76 (28)		106 (24)	14 (31)	
	56-60	55 (26)	69 (25)		113 (26)	11 (24)	
	61-77	113 (53)	128 (47)	0.180	221 (50)	20 (44)	0.573
	mean (range)	60 (39-77)	59 (37-75)		60 (37-77)	59 (37-74)	
Country	France	4 (2)	7 (3)		9 (2)	2 (4)	
	Italy	47 (22)	54 (20)		95 (22)	6 (13)	
	Spain	37 (17)	45 (16)		75 (17)	7 (16)	
	United Kingdom	60 (28)	74 (27)		119 (27)	15 (33)	
	The Netherlands	29 (14)	41 (59)		61 (14)	9 (20)	
	Greece	3 (1)	8 (3)		11 (3)	0 (0)	
	Germany	32 (15)	55 (58)	0.907	70 (16)	6 (13)	0.497
Education	≤primary school	92 (45)	120 (46)		196 (46)	16 (39)	
	technical/professional	54 (26)	61 (23)		103 (24)	12 (29)	
	≥secondary school	60 (29)	82 (31)	0.736	129 (30)	13 (32)	0.663
	missing	6	10		12	4	
BMI	<25	76 (36)	91 (33)		151 (34)	16 (36)	
	25-29.9	95 (45)	143 (52)		218 (50)	20 (44)	
	≥30	41 (19)	39 (14)	0.177	71 (16)	9 (20)	0.739
Smoking status	never	94 (45)	140 (51)		212 (48)	22 (49)	
	former	73 (35)	81 (30)		136 (31)	18 (40)	
	current	44 (21)	51 (19)	0.316	90 (21)	5 (11)	0.238
	missing	1	1		2	0	
Alcohol intake at baseline (g/day)	<6	100 (47)	129 (47)		209 (48)	20 (44)	
	6-20	54 (25)	73 (27)		118 (27)	9 (20)	
	>20	58 (27)	71 (26)	0.925	113 (26)	16 (36)	0.316

[1]Pearson's Chi-Square-test; [2]Gallo0272, Gallo0748, Gallo1675, Gallo2018, Gallo2178 and Gallo2179

Tables:

Table 1: Antigens included in SGG (strain UCN34) multiplex serology and antigen specific cut-offs.

Name	Putative function	Antigen specific cut-off (MFI)
Gallo0112A	Fructan hydrolase N-terminus	30
Gallo0112B	Fructan hydrolase C-terminus	30
Gallo0272*	Glucan binding protein C domain	192
Gallo0577	Cell-wall protein with CnaB domain	185
Gallo0748*	Cell-envelope proteinase A	96
Gallo0933	Tannase	175
Gallo1570	*Pil2* pilus subunit	185
Gallo1675*	Cell wall protein of unknown function	36
Gallo2018*	Putative cell wall protein involved in bacteriocin synthesis	95
Gallo2178*	*Pil1* pilus subunit (major pilin)	30
Gallo2179*	*Pil1* pilus subunit (collagen-binding domain)	118

* antigens included in 6-marker panel

Formatted Table
Formatted: English (U.S.)
Formatted: Left

Table 2: Baseline characteristics of the CRC case-control study nested within EPIC

		Controls (n=485) n (%)	Cases (n=485) n (%)
Sex	female	247 (51)	247 (51)
	male	238 (49)	238 (49)
Age at blood draw, years	37-55	120 (25)	121 (25)
	56-60	124 (25)	122 (25)
	61-77	241 (50)	242 (50)
	Mean (range)	60 (37-77)	59 (37-77)
Country	France	11 (2)	11 (2)
	Italy	101 (21)	101 (21)
	Spain	82 (17)	82 (17)
	United Kingdom	134 (28)	134 (28)
	The Netherlands	70 (14)	70 (14)
	Greece	11 (2)	11 (2)
	Germany	76 (16)	76 (16)
Education	≤primary school	212 (45)	215 (46)
	Technical/professional	115 (25)	95 (21)
	>=secondary school	142 (30)	153 (33)
	missing	16	22
BMI	<25	167 (34)	160 (33)
	25-29.9	238 (49)	220 (45)
	≥30	80 (16)	105 (22)
Smoking status	never	234 (48)	202 (42)
	former	154 (32)	183 (38)
	current	95 (20)	96 (20)
	missing	2	4
Alcohol intake at baseline (g/day)	<6	229 (47)	213 (44)
	6-20	127 (26)	127 (26)
	>20	129 (27)	144 (30)
	missing	0	1

Table 2: Comparison of SGG negative and positive individuals for demographic and other risk factors among controls.

		Any *S. gallolyticus* protein			≥2 of 6 marker panel[2]		
		neg (n=212)	pos (n=273)	p-value[1]	neg (n=440)	pos (n=45)	p-value[1]
Sex	female	103 (49)	144 (53)		225 (51)	22 (49)	
	male	109 (51)	129 (47)	0.363	215 (49)	23 (51)	0.774
Age at blood draw, years	37-55	44 (21)	76 (28)		106 (24)	14 (31)	
	56-60	55 (26)	69 (25)		113 (26)	11 (24)	
	61-77	113 (53)	128 (47)	0.180	221 (50)	20 (44)	0.573
	mean (range)	60 (39-77)	59 (37-75)		60 (37-77)	59 (37-74)	
Country	France	4 (2)	7 (3)		9 (2)	2 (4)	
	Italy	47 (22)	54 (20)		95 (22)	6 (13)	
	Spain	37 (17)	45 (16)		75 (17)	7 (16)	
	United Kingdom	60 (28)	74 (27)		119 (27)	15 (33)	
	The Netherlands	29 (14)	41 (59)		61 (14)	9 (20)	
	Greece	3 (1)	8 (3)		11 (3)	0 (0)	
	Germany	32 (15)	55 (58)	0.907	70 (16)	6 (13)	0.497
Education	≤primary school	92 (45)	120 (46)		196 (46)	16 (39)	
	technical/professional	54 (26)	61 (23)		103 (24)	12 (29)	
	≥secondary school	60 (29)	82 (31)	0.736	129 (30)	13 (32)	0.663
	missing	6	10		12	4	
BMI	<25	76 (36)	91 (33)		151 (34)	16 (36)	
	25-29.9	95 (45)	143 (52)		218 (50)	20 (44)	
	≥30	41 (19)	39 (14)	0.177	71 (16)	9 (20)	0.739
Smoking status	never	94 (45)	140 (51)		212 (48)	22 (49)	
	former	73 (35)	81 (30)		136 (31)	18 (40)	
	current	44 (21)	51 (19)	0.316	90 (21)	5 (11)	0.238
	missing	1	1		2	0	
Alcohol intake at baseline (g/day)	<6	100 (47)	129 (47)		209 (48)	20 (44)	
	6-20	54 (25)	73 (27)		118 (27)	9 (20)	
	>20	58 (27)	71 (26)	0.925	113 (26)	16 (36)	0.316

[1]Pearson's Chi-Square test; [2]Gallo0272, Gallo0748, Gallo1675, Gallo2018, Gallo2178 and Gallo2179

Table 43: Antibody responses to SGG proteins in relation to CRC incidence in a nested case-control study within EPIC

	All				Diagnosed more than 2 years after blood draw			
	Positive n (%)				Positive n (%)			
	Controls n=485	Cases n=485	OR[1]	95% CI	Controls n=355	Cases n=355	OR[1]	95% CI
Gallo0112A	33 (7)	37 (8)	1.09	0.64-1.84	22 (6)	23 (6)	1.08	0.55-2.11
Gallo0112B	28 (6)	26 (5)	0.96	0.55-1.67	15 (4)	16 (5)	1.13	0.55-2.32
Gallo0272	47 (10)	67 (14)	1.59	1.06-2.40	32 (9)	51 (14)	1.87	1.15-3.05
Gallo0577	47 (10)	49 (10)	1.03	0.67-1.59	34 (10)	36 (10)	1.05	0.64-1.72
Gallo0748	50 (10)	74 (15)	1.49	1.02-2.16	37 (10)	51 (14)	1.40	0.90-2.18
Gallo0933	49 (10)	44 (9)	0.92	0.59-1.43	37 (10)	38 (11)	1.05	0.64-1.73
Gallo1570	47 (10)	52 (11)	1.13	0.72-1.76	36 (10)	41 (12)	1.19	0.72-1.96
Gallo1675	48 (10)	51 (11)	1.08	0.70-1.67	38 (11)	39 (11)	1.09	0.67-1.76
Gallo2018	47 (10)	54 (11)	1.24	0.81-1.89	38 (11)	43 (12)	1.22	0.77-1.95
Gallo2178	12 (2)	31 (6)	3.01	1.49-6.08	7 (2)	17 (5)	3.28	1.25-8.57
Gallo2179	47 (10)	64 (13)	1.48	0.97-2.24	34 (10)	44 (12)	1.47	0.90-2.40
Any SGG protein	273 (56)	306 (63)	1.36	1.04-1.77	201 (57)	224 (63)	1.38	1.02-1.87
≥2 of 6-marker panel[2]	45 (9)	83 (17)	2.17	1.44-3.27	36 (10)	60 (17)	2.07	1.29-3.31

[1]Conditional logistic regression model with multivariable adjustment for BMI, education, smoking and alcohol intake at baseline; [2]Gallo0272, Gallo0748, Gallo1675, Gallo2018, Gallo2178, Gallo2179