**Supplementary Methods, Tables and Figures**

Content

[Supplementary Material and Methods 2](#_Toc19806322)

[Table S1: Association results for 22 GBC cases and 7637 healthy controls among 337,000 unrelated individuals of white British ancestry available at UK Biobank. 6](#_Toc19806323)

[Table S2: D-prime and r-squared values among the three recently identified common GBC risk variants in Chilean† and European control individuals. 7](#_Toc19806324)

[Table S3: Haplotype association results for the three recently identified common GBC risk variants in Chileans† and Europeans. 8](#_Toc19806325)

[Table S4: Used genotyping arrays and available genotype data. 9](#_Toc19806326)

[Table S5: Genotype call rates and IMPUTE2 info scores. 10](#_Toc19806327)

[Figure S6: Scree plot from a principal component analysis of Chilean (A) and European (B) genome-wide genotype data. 11](#_Toc19806328)

[Figure S7: Comparison between the estimated per-allele odds ratios considering five (x-axis) and twenty (y-axis) principal components. 12](#_Toc19806329)

# Supplementary Material and Methods

***Study design, genotyping and ethical approval***

The present multicentre, population-based, candidate variant association study included 255 GBC cases and 2042 controls from a Chilean retrospective study, and 108 GBC cases and 181 controls from a consortium of large European prospective cohorts. Cases were diagnosed with GBC according to the International Classification of Diseases for Oncology Version 3, site code C23. Patients diagnosed with distinct biliary duct malignancies were not included in the study. The majority of GBC patients (78%) showed biliary pain, underwent an echography that revealed gallstones, received cholecystectomy and were diagnosed with incidental GBC by a pathologist after histopathological gallbladder inspection. A minority of the patients were directly diagnosed with GBC without a previous cholecystectomy. Chilean controls were Chileans without reported GBC and included individuals affected by gallstone disease. Additional details on the recruitment strategy and on the demographic characteristic for the Chilean controls can be found in our recent publication.[1] European study participants were recruited over the past 30 years by eight large European prospective cohorts: the Esther study, the Estonian Genome Project, LifeLines, the European prospective investigation into cancer and nutrition, the Swedish Twin Registry, the National FINRISK Study, the Nord-Trøndelag Health Study and the Study of Health in Pomerania. The European controls were healthy participants without a reported cancer history.

Genotype data were obtained using various GWAS arrays (**Supporting Table S4**). In addition to the three investigated GBC risk variants, genome-wide genotype data was used to infer the underlying population structure and to estimate individual ancestry components. Intentional duplicates, samples with sex information inconsistent with genotype data, arrays with more than 5% missing genotypes, familial cases and related individuals, cases with missing age at diagnosis, and controls with missing age at interview were excluded.

Ethics approval for Chilean samples was obtained from the Medical Faculty of the Universidad de Chile (approval #123-2012), and from Universidad de Tarapacá and University College London as previously described in Ruiz-Linares et al.[2] For European samples ethical approval was attained by the responsible ethical committees: IARC Ethics Committee (IEC, #16-23), Ethics Commitee of the University of Tartu (#223/T-10), THL Biobank (Finrisk, #BB2016\_32), REK (HUNT, #2016/1222), Lifelines (#OV16\_0371), Forschungsverbund Community Medicine (#SHIP/2016/137/M+D) and EPN (TwinGene, #2016/2:11). All participants provided written informed consent prior to participation. Ethics approvals, material/data transfer agreements, and the structured questionnaires applied to our volunteers are available upon request.

***Reference individuals for ancestry estimation***

Reference individuals were used to infer individual ancestry proportions for our Chileans. Surrogates of African and European ancestry were 87 Yorubans in Ibadan, Nigeria, and 80 Utah residents with Northern and Western European ancestry from the 1000 Genome Project.[2] Nine Mapuche and nine Aymara individuals were selected to represent the two largest indigenous peoples in Chile based on the three following criteria: four grandparental Mapuche or Aymara surnames, estimated Native American proportion of at least 74% for Mapuche and at least 99% for Aymara reference individuals, and mitochondrial DNA haplogroups consistent with Mapuche (haplogroup C or D) or Aymara (haplogroup B) descent.[1]

***Genetic principal component analysis and estimation of ancestry proportions***

Non-autosomal polymorphisms, variants with a missing call rate over 5% and variants with a minor allele frequency (MAF) under 5% in controls were excluded. After linkage disequilibrium (LD) pruning at r2 higher than 0.1, the Chilean dataset comprised more than 45,000 variants and the European dataset more than 25,000 variants. Subsequent genetic principal component analyses (PCA) were conducted using the EIGENSTRAT function available at popgen.dk/software/index.php/Rscripts.[3] The ADMIXTURE software was used for supervised estimation of individual African, European, Mapuche, and Aymara ancestry components relying on the above-mentioned reference individuals.[4]

***Imputation analyses for gallbladder cancer susceptibility variants***

The three investigated variants were not directly genotyped in all samples and, as a result, the overall genotyping call rate was high (99%) for rs4148808 and rs1709837 but low (69%) for rs1558375 (**Supporting Table S5**). After exclusion of polymorphisms with call rates under 95%, MAFs under 0.5% in controls, and adenine-thymine or guanine-cytosine alleles to avoid strand flipping errors, the strand of our genotype data was aligned with the 1000 Genomes Project data. Study genotypes were pre-phased with the SHAPEIT software version 2.12.[5] Missing genotypes on chromosome 7 between 87 Mb and 88 Mb (GRCH38 built) were separately imputed for studies with consistent sets of variants using the IMPUTE2 software version 2.3.2 with version 3 of the 1000 Genomes Project data as the reference set.[2, 6] For the National FINRISK study, genotype data imputed based on a large imputation panel of 2,690 high-coverage whole-genome-sequenced and 5,092 whole-exome-sequenced Finnish genomes were available.

***Sensitivity analyses***

In agreement with the reported recombination blocks for individuals of Indian descent, all three variants showed D-prime values over 0.9 with similar LD estimates for Chilean and European controls (**Supporting Table S2**). To examine the effect of linkage disequilibrium (LD) patterns between the three investigated SNPs, an alternative logistic regression analysis was applied, considering the count of high-risk A-A-A-haplotypes as main explanatory variable. Results are displayed in **Supporting Table S3**.

In consistency with the publication by Mhatre et al., the first five principal components were used for adjustment of potential population stratification. Others and we have demonstrated that the genetic admixture of Chileans is well captured by the four used reference populations corresponding to the first three principal components.[1, 7] Novembre et al. have shown that the first two principal components mirror the geographic origin of Europeans and are sufficient to control the genomic inflation in simulated GWAS to a minimum.[8] Moreover, scree plots for both Chileans and Europeans revealed a rapid decline in the genetic variance captured by the first three principal components (**Supporting Figure S6**). However, in order to completely avoid hidden population structure, sensitivity analyses were conducted using the first twenty principal components for stratification adjustment. Results are shown in **Supporting Figure S7**.

***Software***

All association analyses were carried out with the SNPTEST software version 2.5.2 with the exception of the likelihood ratio test and the haplotype association analyses which were conducted using SAS version 9.4.[9]

# Table S1: Association results for 22 GBC cases and 7637 healthy controls among 337,000 unrelated individuals of white British ancestry available at UK Biobank.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **SNP ID** | **RAF in controls** | **OR** | **95%** | **CI** | **p-value†** |
| **rs1558375** | 0.82 | 0.98 | 0.46 | 2.14 | 0.93 |
| **rs4148808** | 0.86 | 1.25 | 0.43 | 2.80 | 0.57 |
| **rs17209837** | 0.86 | 1.26 | 0.45 | 2.81 | 0.55 |

†The p-values were provided by the Neale Lab (<http://www.nealelab.is/blog/2017/7/19/rapid-gwas-of-thousands-of-phenotypes-for-337000-samples-in-the-uk-biobank>) from a least-squares linear model predicting the phenotype with an additive genotype coding, with sex and the first 10 principal components from the UK Biobank sample QC file as covariates.

SNP=Single nucleotide polymorphism, ID=identification, RAF=risk allele frequency (risk allele=Adenin), OR=unadjusted per-allele odds ratio from allele counts in GBC cases and controls (own calculations).

# Table S2: D-prime and r-squared values among the three recently identified common GBC risk variants in Chilean† and European control individuals.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Population** |  | **SNP ID** | **D-prime** | **r-squared** |
| **rs4148808** | **rs17209837** | **rs4148808** | **rs17209837** |
| Chilean |  | **rs1558375** | 0.95 | 0.92 | 0.84 | 0.81 |
| 0% to 100% |  | **rs4148808** |  | 0.96 |  | 0.96 |
| Mapuche proportion |  |  |  |  |  |  |
| Chilean  |  | **rs1558375** | 0.96 | 0.93 | 0.89 | 0.93 |
| 42% to 100% |  | **rs4148808** |  | 0.98 |  | 0.97 |
| Mapuche proportion |  |  |  |  |  |  |
| Chilean  |  | **rs1558375** | 0.96 | 0.93 | 0.86 | 0.83 |
| 36% to 41% |  | **rs4148808** |  | 0.98 |  | 0.96 |
| Mapuche proportion |  |  |  |  |  |  |
| Chilean |  | **rs1558375** | 0.94 | 0.90 | 0.83 | 0.80 |
| 29% to 35% |  | **rs4148808** |  | 0.97 |  | 0.96 |
| Mapuche proportion |  |  |  |  |  |  |
| Chilean  |  | **rs1558375** | 0.95 | 0.90 | 0.78 | 0.75 |
| 0% to 28% |  | **rs4148808** |  | 0.95 |  | 0.94 |
| Mapuche proportion |  |  |  |  |  |  |
|  |  | **rs1558375** | 1.00 | 1.00 | 0.88 | 0.87 |
| European |  | **rs4148808** |  | 1.00 |  | 0.99 |
|  |  |  |  |  |  |  |

†Chilean results are additionally stratified according to the proportions of Mapuche ancestry, with each group containing one quarter of the Chilean individuals.

SNP=Single-nucleotide polymorphism, ID=identification, D-prime and r-squared between GBC risk variants in controls

# Table S3: Haplotype association results for the three recently identified common GBC risk variants in Chileans† and Europeans. By leveraging on the LD information, haplotype analyses revealed a per-haplotype OR of 1.50 (95% CI 1.12–2.00) for all Chileans and 1.82 (95% CI 1.14–2.91) for Chileans with 42%–100% Mapuche ancestry. By contrast, no per-haplotype association was found for Chileans with less than 35% Mapuche ancestry or Europeans.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Population** | **Mapuche****proportion** | **Proportion of** **A-A-A A-A-A diplotypes** | **OR‡** | **95%** | **CI** | **Trend p-value** |
| **Cases** | **Controls** |
| Chileans | - | 66.7% | 59.9% | **1.50** | 1.12 | 2.00 | 0.006 |
| Chileans | 42%-100% | 67.6% | 54.6% | **1.82** | 1.14 | 2.91 | 0.01 |
| Chileans | 36%-41% | 69.8% | 63.4% | 1.77 | 0.98 | 3.21 | 0.06 |
| Chileans | 29%-35% | 63.0% | 62.3% | 1.13 | 0.60 | 2.12 | 0.72 |
| Chileans | 0%-28% | 63.0% | 59.0% | 0.94 | 0.37 | 2.39 | 0.89 |
| Europeans | - | 72.2% | 67.4% | 1.08 | 0.66 | 1.77 | 0.75 |

†Chilean analyses are also stratified by the estimated proportion of Mapuche ancestry, with each group containing one quarter of the individuals.

‡Bold type denotes associated 95% CI that do not include 1.

OR=per-A-A-A-haplotype odds ratio adjusted for age, sex and first five principal components, CI=confidence interval.

# Table S4: Used genotyping arrays and available genotype data.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Population** | **Study/cohort** | **Cases** | **Controls** | **Genotyping array / available genotype data** |
| **Chilean** | Identification of biomarkers for the personalized prevention and therapy of gallbladder cancer, phase I | 20 | 0 | Illumina Infinium OmniExpressExome-8  |
| Identification of biomarkers for the personalized prevention and therapy of gallbladder cancer, phase II | 235 | 306 | Illumina Infinium Global Screening Array-24 v1.0  |
| Chilean subset of the Consortium for the Analysis of the Diversity and Evolution of Latin America (CANDELA) | 0 | 1736 | Illumina Human610-Quad BeadChip |
| **European** | Esther† | 8 | 8 | Illumina Infinium Global Screening Array-24 v1.0  |
| Estonian Genome Project | 5 | 4 |
| LifeLines‡ | 1 | 5 |
| European prospective investigation into cancer and nutrition (EPIC) | 59 | 57 |
| Swedish Twin Registry | 10 | 49 |
| National FINRISK Study | 6 | 7 | Pre-phased (Eagle) and imputed (IMPUTE2) genotype data |
| Nord-Trøndelag Health Study | 14 | 15 | Phased haplotypes  |
| Study of Health in Pomerania | 5 | 22 | Affymetrix Genome-Wide Human SNP Array 6.0 |
| 0 | 12 | Illumina Infinium Global Screening Array-24  |
| 0 | 2 | Illumina Infinium Omni2.5Exome-8 Kit |

†Ongoing population-based cohort study set up with the aim of improving the prevention, early detection, and treatment of chronic diseases in older adults

‡Longitudinal cohort study with over 167,000 participants of the northern parts of the Netherlands over a period of 30 years

# Table S5: Genotype call rates and IMPUTE2 info scores.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Population** | **Single and combined studies used for respective imputation** | **Number of samples used for imputation** | **Call rate (%)** | **Impute2 info score** |
| **rs1558375** | **rs4148808** | **rs17209837** | **rs1558375** | **rs4148808** | **rs17209837** |
| **Chilean** | Identification of biomarkers for the personalized prevention and therapy of gallbladder cancer, phase II | 541 | 0.0  | 99.6 | 100.0 | 0.999 | 1.000 | 1.000 |
| Identification of biomarkers for the personalized prevention and therapy of gallbladder cancer, phase I and Chilean subset of CANDELA | 1756 | 99.7 | 99.7 | 99.9 |
| **European** | Esther, Estonian Genome Project, LifeLines, EPIC and the Swedish Twin Registry | 206 | 0.0 | 99.3 | 100.0 | 0.992 | 1.000 | 1.000 |
| National FINRISK Study† | 13 | 100.0 | 100.0 | 100.0 |
| Nord-Trøndelag Health Study | 29 | 100.0 | 100.0 | 100.0 |
| Study of Health in Pomerania | 41 | 0.0 | 34.2 | 34.2 |

†Genotype data from the FINRISK study were available as IMPUTE2 imputed genotypes

# Figure S6: Scree plot from a principal component analysis of Chilean (A) and European (B) genome-wide genotype data.

|  |
| --- |
|  |
|  |

# Figure S7: Comparison between the estimated per-allele odds ratios considering five (x-axis) and twenty (y-axis) principal components.



For each considered (sub)population, per-allele odds ratios from the trend test which corrects for population structure with five principal components (x-axis) are plotted against per-allele odds ratios from a test which adjusts for twenty principal components. Candidate variants at which the per-allele odds ratios are affected by the correction of further principal components appear as points off the diagonal. Such differences are seen for all (sub)populations, but are minor and act in both directions. It was just confirmed that the first five principal components were sufficient to adjust for population stratification in the association analyses.

**References:**

[1] Lorenzo Bermejo J, Boekstegers F, Gonzalez Silos R, Marcelain K, Baez Benavides P, Barahona Ponce C, et al. Subtypes of Native American ancestry and leading causes of death: Mapuche ancestry-specific associations with gallbladder cancer risk in Chile. PLoS genetics. 2017;13:e1006756.

[2] Genomes Project C, Auton A, Brooks LD, Durbin RM, Garrison EP, Kang HM, et al. A global reference for human genetic variation. Nature. 2015;526:68-74.

[3] Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. Principal components analysis corrects for stratification in genome-wide association studies. Nature genetics. 2006;38:904-9.

[4] Alexander DH, Novembre J, Lange K. Fast model-based estimation of ancestry in unrelated individuals. Genome research. 2009;19:1655-64.

[5] O'Connell J, Gurdasani D, Delaneau O, Pirastu N, Ulivi S, Cocca M, et al. A general approach for haplotype phasing across the full spectrum of relatedness. PLoS genetics. 2014;10:e1004234.

[6] Howie BN, Donnelly P, Marchini J. A flexible and accurate genotype imputation method for the next generation of genome-wide association studies. PLoS genetics. 2009;5:e1000529.

[7] Eyheramendy S, Martinez FI, Manevy F, Vial C, Repetto GM. Genetic structure characterization of Chileans reflects historical immigration patterns. Nature communications. 2015;6:6472.

[8] Novembre J, Johnson T, Bryc K, Kutalik Z, Boyko AR, Auton A, et al. Genes mirror geography within Europe. Nature. 2008;456:98-101.

[9] Marchini J, Howie B. Genotype imputation for genome-wide association studies. Nature reviews Genetics. 2010;11:499-511.