Common shared genetic variation behind decreased risk of breast cancer in celiac disease

Emilio Ugalde-Morales, Jingmei Li, Keith Humphreys, Jonas F. Ludvigsson, Haomin Yang, Per Hall & Kamila Czene

There is epidemiologic evidence showing that women with celiac disease have reduced risk of later developing breast cancer, however, the etiology of this association is unclear. Here, we assess the extent of genetic overlap between the two diseases. Through analyses of summary statistics on densely genotyped immunogenic regions, we show a significant genetic correlation ($r = -0.17$, s.e. 0.05, $P < 0.001$) and overlap ($P_{\text{permuted}} < 0.001$) between celiac disease and breast cancer. Using individual-level genotype data from a Swedish cohort, we find higher genetic susceptibility to celiac disease summarized by polygenic risk scores to be associated with lower breast cancer risk (OR per-SD, 0.94, 95% CI 0.91 to 0.98). Common single nucleotide polymorphisms between the two diseases, with low $P$-values ($P_{\text{CD}} < 1.00E-05$, $P_{\text{BC}} \leq 0.05$), mapped onto genes enriched for immunoregulatory and apoptotic processes. Our results suggest that the link between breast cancer and celiac disease is due to a shared polygenic variation of immune related regions, uncovering pathways which might be important for their development.

Breast cancer risk has been reported to be consistently lower among celiac disease patients, ranging from being 10-15% lower in Nordic studies1–3 to as much as 50-80% lower in other European studies with smaller sample sizes5–8. Celiac disease is a lifelong gastrointestinal disease characterized by villous atrophy and inflammation in the small intestine9. It occurs in about 1% of the Caucasian population and is triggered by gluten exposure10. There is little in literature to clarify why a diagnosis of celiac disease confers protection against breast cancer. Explanations that have been forwarded include a lower body mass index and lower estrogen exposure both as a consequence of celiac disease; for example, characteristics secondary to undernutrition such as later menarche and earlier menopause commonly observed at high frequencies among women with celiac disease are associated with decreased risk of breast cancer2. A third hypothesis involves the role of immunogenic factors in breast cancer development and progression11. Inverse relationships have also been observed between breast cancer and other inflammatory disorders such as ulcerative colitis and rheumatoid arthritis, which is consistent with a possible involvement of the immune system in the etiological pathway of breast cancer2. Others have also shown that the interplay between hormonal and immune-related mechanisms can shape mammary tissue development12.

Both breast cancer and celiac disease have strong genetic components. Heritability is estimated to be between 25 to 31%13,14 and 68 to 75%15 for breast cancer and celiac disease, respectively. Heritability of complex diseases such as celiac disease and breast cancer is highly polygenic, which means that it is controlled not just by one gene, but rather, by multiple genes16. For example, genome-wide association studies (GWAS) interrogating upwards of ~200,000 single nucleotide polymorphisms (SNPs) have identified ~40 loci associated with celiac disease17, and more than twice as many (107 loci) for breast cancer18. Studies performed on different phenotypes have shown that certain genetic loci can be associated with seemingly distinct traits, otherwise known as pleiotropy19. It is also known that many diseases and traits exhibit significant coheritability20 and have shared genetic components21, and such genetic associations often reveal clues about novel mechanisms and pathways.

With the emergence of large-scale GWAS studies, it is timely to leverage on the data collected from international consortia to investigate possible causal trait relationships22 using clinical and epidemiological data as a guide23–25. Motivated by evidence of an association between celiac disease and breast cancer from epidemiologic
studies, the aims of this study are to test whether these findings could be due to shared genetic determinants, and to elucidate what mechanisms are potentially responsible for the common connection.

**Results**

**Inverse genetic correlation and genetic overlap between breast cancer and celiac disease.** Genetic correlation is indicative of shared genetic etiology. It refers to common genetic variation associated with a pair of phenotypic traits, assuming an additive model. Given the epidemiological association between breast cancer and celiac disease, as well as their strong genetic heritabilities, we used GWAS summary statistics for each disease to estimate genetic correlation. For breast cancer, data was obtained from the Collaborative Oncological Gene-Environment Study (COGS) consortia (http://www.cogseu.org/). The study includes individuals of European ancestry genotyped on a custom Illumina iSelect Array (iCOGS)26, which comprises 211,155 SNPs. iCOGS was designed to understand genetic susceptibility of three hormone related cancers: breast, ovarian, and prostate. As breast cancer is a heterogeneous disease, we also included two major subtypes of breast cancer based on estrogen receptor (ER) status. In the COGS, analyses had been conducted on 46,785 breast cancer cases and 42,892 controls to estimate breast cancer risk overall, 27,078 ER-positive cases and 42,111 controls for ER-positive breast cancer risk, and 7,333 ER-negative cases and 42,468 controls for ER-negative breast cancer risk27. Summary statistics for celiac disease (133,352 SNPs) were downloaded from the ImmunoBase (https://www.immunobase.org/), a web based resource focused on the genetics and genomics of immunologically related human diseases. Celiac disease data have been reported in a GWAS study by Trynka et al.23 on 12,041 celiac disease cases and 12,228 controls of European ancestry using the Illumina Infinium High-Density array (ImmunoChip), interrogating 195,806 SNPs and designed to target immune associated genome regions25. To improve the comparability between the two datasets, we used summary statistics for 173,301 SNPs in the iCOGS study imputed against the 1000 Genomes Project (1KG) March 2012 release reference panel28, which were also present on the ImmunoChip. The datasets were matched based on chromosome and SNP base pair positions, which resulted in 129,618 celiac disease SNPs used as input.

Genetic correlation between breast cancer and celiac disease was analyzed using LD score regression (LDSC)29 to model effect size estimates for immunogenic SNPs in both diseases. Therefore, rather than studying cross-correlation, we are studying the role of celiac disease risk loci in breast cancer susceptibility. Given the observed reduced risk of breast cancer in celiac disease patients, we would expect an inverse genetic correlation. After LDSC filtering procedure and merge with reference LD scores (Supplementary Table 1), genetic correlation analyses included 45,451, 45,451 and 45,447 matching SNPs between the celiac disease and breast cancer overall, ER-positive and ER-negative breast cancer datasets, respectively. Significant inverse genetic correlations (r) were found between celiac disease and overall breast cancer (r = −0.17, s.e. 0.05, P = 0.0005) and ER-positive breast cancer (r = −0.15, s.e. 0.06, P = 0.01), but not for ER-negative breast cancer (r = −0.03, s.e. 0.07, P = 0.71) (Fig. 1).

Further interrogation of shared common genetic components between the two diseases was carried out using SNP effect concordant analysis (SECA)29, where SNP effect size estimates were tested for concordant or discordant effects, analogous to genetic correlation tested with LDSC. Additionally, SECA also assesses the extent of genetic overlap (enrichment of overlapping SNPs between the two traits with low P-values). For each dataset pair comparison, SECA aligned and selected 15,365, 15,400, 15,428 independent SNPs that are common between the celiac disease and overall, ER-positive, and ER-negative breast cancer datasets, respectively (Supplementary Table 2).

In the primary analyses which are summarized in Supplementary Figure 1, genetic overlap (defined as excess of celiac disease SNPs in overlap with breast cancer datasets) was significant between celiac disease and breast cancer overall (P_{PT-permuted} = <0.001), ER-positive (P_{PT-permuted} < 0.001), and ER-negative (P_{PT-permuted} < 0.001). SNP effect discordance (inverse correlation) between celiac disease and breast cancer overall (P_{PT-permuted} = 0.059), ER-positive (P_{PT-permuted} = 1.000), and ER-negative (P_{PT-permuted} = 0.278) did not reach significance. In order to determine size of the association, SECA identified a subset of overlapping SNPs yielding the most significant correlation, namely, minimum concordance or discordance. This analysis was carried out studying the association between celiac disease and breast cancer overall (OR_{PT-min} 0.60, 95% CI 0.44–0.82, P_{PT-min} = 0.001), ER-positive (OR_{PT-min} OR, 0.86, 95%CI, 0.74–1.00 P_{PT-min} = 0.05), and ER-negative (OR_{PT-min} OR, 0.73, 95%CI, 0.57–0.95 P_{PT-min} = 0.02) (Table 1). After adjusting for multiple testing, only the association for overall breast cancer (which showed discordance) remained significant (P_{PT-min-permuted} = 0.022).

**Higher celiac disease genetic susceptibility associated with decreased breast cancer risk.** In a third approach, we tested whether top SNPs for celiac disease (CD-SNPs) could predict breast cancer status in a group of women. Genetic susceptibility to celiac disease was summarized using polygenic risk scores (celiac-PRS). In essence, celiac-PRS accounts for the genetic susceptibility of an individual based on the risk allele load weighted by the SNP effect sizes reported by the celiac disease GWAS, under an additive model. Celiac-PRS was analyzed as an exposure variable in a case-control study comprised of 5,002 breast cancer cases and 5,433 controls from the pKARMA cohort. Celiac-PRS was found to be inversely associated with overall and ER-positive breast cancer risk in a dose-dependent manner (P-trend < 0.02) (Table 2). Celiac-PRS based on 199 genome-wide significant CD-SNPs was associated with 6% lower risk of overall (OR_{per-SD} 0.94, 95% CI 0.91 to 0.98, P = 0.002) and ER-positive breast cancer (OR_{per-SD} 0.94, 95% CI 0.90 to 0.98, P = 0.004), and 2% for ER-negative breast cancer (OR_{per-SD} 0.98, 95% CI 0.90 to 1.06, P = 0.54). The risk was 13% lower in individuals with the highest genetic susceptibility to celiac disease (4th celiac-PRS quartile compared to 1st quartile) for both overall and ER-positive breast cancer risk (overall: OR_{Q4} 0.87, 95% CI 0.78 to 0.97, P = 0.016; ER-positive: OR_{Q4} 0.87, 95% CI 0.77 to 0.98, P = 0.022). The risk was up to 17% lower in women with highest susceptibility when the celiac-PRS included controls from the pKARMA cohort. Celiac-PRS was found to be inversely associated with overall and ER-positive breast cancer risk (overall: OR_{Q4} 0.83, 95% CI 0.75 to 0.93, P = 0.001; ER-positive: OR_{Q4} 0.83, 95% CI 0.74 to 0.93, P = 0.002). As expected under a polygenic model, including more CD-SNPs in the profiles improved the strength of the association with breast cancer risk (yielding...
smaller $P$-values) (Fig. 2). In a case-only study, no significant difference was observed within tumor subgroups defined by ER-status, lymph node involvement status, HER2 status, tumor grade, and tumor size (Supplementary Table 3).

Candidate immune response genes and pathways underlying the association between celiac disease and breast cancer. To identify celiac disease genes and molecular pathways involved in breast cancer susceptibility, we performed enrichment analysis using Data-driven Expression Prioritized Integration for Complex Traits (DEPICT)\(^3\). The method assumes that the selected loci surpass the genome-wide significant threshold ($P < 5.00E-08$). However, since genome-wide significant variants did not determine the genetic overlap between breast cancer and celiac disease, we changed the threshold to include variants with moderate signals. From the 15,365 independent SNPs defined in the SECA analysis, we selected SNPs which had suggestive association with celiac disease ($P_{CD} < 1.00E-05$). Given that the genetic overlap was limited to immune-related genomic regions genotyped for celiac disease, we ranked overlapping SNPs by their evidence of association (by $P$-values) relative to the overlap with breast cancer to reduce the possibility that the prioritized genes could be due to chance (sensitivity analysis). SNPs were divided into two SNP subsets, $P_{BC} \leq 0.05$ and $P_{BC} > 0.05$.
In agreement with epidemiological studies showing lower breast cancer risk in celiac disease patients, we found an inverse genetic association between the two diseases using three different methods. There was no evidence that the association between genetic susceptibility to celiac disease and lower risk of breast cancer differed by tumor characteristics. We also prioritized apoptotic and immune-related genes that could represent important etiological factors underlying the reduced risk of breast cancer in celiac disease.

In spite of the strong heritability of breast cancer and celiac disease, the etiological role played by their genetic components remains to be uncovered. Both prospective cohort studies and randomized clinical trials have found that breastfeeding duration and age at gluten introduction may be less important than previously thought for the etiology of celiac disease. Instead, it seems that genetic factors determine the risk of celiac disease. For breast cancer, immune response factors have been shown to be important factors associated with breast cancer prognosis, and potentially to breast cancer susceptibility. As for genetic markers of breast cancer, studies on immune response candidate genes have identified few single risk alleles for specific populations.

| Profiles (quartile range) | BC-Overall | ER-positive (n = 3,804) | ER-negative (n = 695) |
|--------------------------|------------|------------------------|----------------------|
|                          | OR 95%CI   | P-value  | OR 95%CI  | P-value |
| GWAS significant (P_CD < 5E-08) – 199 SNPs |
| Q1 (-0.0992 to -0.065)   | 1.00       | 1.00     | 1.00     |
| Q2 (-0.065 to -0.0554)   | 1.02       | 0.09     | 1.00     |
| Q3 (-0.0554 to -0.0366)  | 0.99       | 0.96     | 1.14     |
| Q4 (-0.0366 to 0.0694)   | 0.87       | 0.87     | 0.96     |
| Continuous variable      | 0.94       | 0.94     | 0.98     |
| OR > 1, 95%CI            | 1.00       | 1.00     | 1.00     |
| P-value                  | 1.00       | 1.00     | 0.90     |
| OR < 1, 95%CI            | 1.00       | 1.00     | 0.90     |
| P-value                  | 1.00       | 1.00     | 0.90     |

Table 2. Association of celiac-PRS profiles with breast cancer risk. Breast cancer risk association with celiac-PRS profiles including CD-SNPs with P-value less than four significance thresholds [P_CD < 5E-08, P_CD < 1E-05, P_CD < 0.01, and P_CD < 0.05]. Celiac-PRS quartiles (Q1–Q4) were defined based on PRS distribution in controls. Celiac-PRSs as continuous variables expressed per 1 standard deviation.

**Discussion**

In agreement with epidemiological studies showing lower breast cancer risk in celiac disease patients, we found an inverse genetic association between the two diseases using three different methods. There was no evidence that the association between genetic susceptibility to celiac disease and lower risk of breast cancer differed by tumor characteristics. We also prioritized apoptotic and immune-related genes that could represent important etiological factors underlying the reduced risk of breast cancer in celiac disease.

In spite of the strong heritability of breast cancer and celiac disease, the etiological role played by their genetic components remains to be uncovered. Both prospective cohort studies and randomized clinical trials have found that breastfeeding duration and age at gluten introduction may be less important than previously thought for the etiology of celiac disease. Instead, it seems that genetic factors determine the risk of celiac disease. For breast cancer, immune response factors have been shown to be important factors associated with prognosis, and potentially to breast cancer susceptibility. As for genetic markers of breast cancer, studies on immune response candidate genes have identified few single risk alleles for specific populations.
In our study, inverse genetic correlation between breast cancer and celiac disease were consistent across different methods, indicating that the link between the two diseases is a result of shared immunogenetic components. Through LDSC and SECA methods which use all SNPs with available summary statistics for pair of traits, we found significant genetic correlation and overlap between breast cancer and celiac disease. While LDSC uses all common SNPs between the two diseases to estimate genetic correlation ($r = -0.17$), SECA identified inverse correlation in the most significant subset of SNPs ($OR = 0.60$), indicating the presence of allelic effects that increases risk for one disease and decreases risk for the other. The fact that genetic discordance (negative correlation) was not found significant in the SECA primary analyses as compared to the LD score regression, could be due to loss of power and related to the different approaches they use to deal with LD structure (see methods). In a third approach, genetic correlation was estimated by first summarizing the per-individual allelic load using a polygenic risk score, and then regressing the effect in a case control set up. We found 6 up to 17% lower risk to breast cancer to be comparable to the 10–15% decreased risk reported in Nordic epidemiological studies. Given shared environmental exposures which could mediate the association, we considered GWAS summary statistics data for body mass index (BMI) from the GIANT consortium. The average BMI has been shown to be lower in celiac
disease patients, while low BMI is also associated with lower risk of breast cancer. However, we did not find any indication of a genetic correlation between BMI and breast cancer or celiac disease ($P = 0.23$ and $P = 0.79$ respectively, data not shown).

The involvement of the immune system is typically associated with ER-negative disease. Lymphocytic infiltration has been reported as a favorable prognostic factor for ER-negative$^{45}$ and triple-negative breast cancer$^{49}$. ER-negative breast cancer is characterized by a stronger immunogenic component, which could be proposed as the underlying link with celiac disease. Still, we did not observe genetic correlation between celiac disease and ER-negative breast cancer, most probably due to the low statistical power as a consequence of a smaller sample size for the ER-negative datasets.

Since complex traits such as breast cancer and celiac disease involve the deregulation of multiple interrelated biological processes, it may be informative to isolate which genes are important in the etiology of both diseases. In our functional enrichment results, we found genes involved in relevant mechanisms to be implicated in developmental and immunoregulatory processes. The most significantly prioritized genes were: Mesoderm Induction Early Response 1 (MIER1), C-src Tyrosine Kinase (CSK), Secretary Carrier Membrane Protein 2 (SCAMP2), and Interleukin 12 Alpha (IL12A). MIER1 codes for proteins with transcriptional repressive function and has been found upregulated in human breast carcinoma cell lines and tumors$^{44}$. By interaction with transcription factors and chromatin modifiers such as ER-alpha and histone deacetylase inhibitor HDAC1/2$^{45,46}$, MIER1-alpha inhibits estrogen dependent growth and lack nucleus internalization during breast cancer progression$^{47}$. Thus, is possible that genetic variants could affect MIER1 protein interactions and migration mechanisms necessary to exert its function, explaining the high expression level seen in breast cancer cells as a compensatory response. CSK gene codes for the human cytosolic non-receptor tyrosine kinase protein, which regulate different transcription signals implicated in cell growth, differentiation, migration and immune response processes. CSK inactivates the sarcoma (Src) family kinases which otherwise would lead to T-Cell antigen specific response by phosphorylating zeta chain T-cell receptor (TCR), which has been found downregulated for different cancer types, autoimmune disease and chronic inflammation$^{48}$. IL12A codes for a cytokine with important effects on the regulation of the immune and inflammatory responses$^{49}$ and has been considered for cancer immunotherapy$^{50}$. IL112A loci selected through genetic population factors have been associated with celiac disease and other autoimmune diseases$^{51}$. A query in a pathway catalog (http://pathcards.genecards.org; accessed on November 15, 2016) showed that other significantly prioritized genes (SEMA7A, ITGA4, IL18RAP, CD24, TAGA, TNFRSF, ZFP36L1) are classified on immune-related signaling pathways with important immunomodulatory$^{52,53}$ and apoptotic functions$^{54}$. Overall, this suggests that a complex network of signaling pathways play an important role in the regulation of the immune response and surveillance. Disruption of this network could lead to autoimmune responses, or to changes in mammary microenvironment elements predisposing to cancer immune evasion, in line with cumulative evidence highlighting the relevance of host immunity and genomic alterations in the disease heterogeneity and for tailoring therapeutic interventions$^{55}$. It could be hypothesized that while some of the overlapping variants might be involved in heightened immune responses, they could at the same time increase immunosurveillance against carcinogenic processes in breast tissue, thereby reducing breast cancer risk. Our findings might guide future studies that can help to understand the role played by the immune system in breast cancer susceptibility.

The main strength of our study is the use of reliable summary statistics from large multicenter GWAS consortia for both diseases, and the leverage of epidemiological association to estimate genetic correlation and immune-related genetic susceptibility to breast cancer. By using different polygenic approaches and prior biological knowledge, we could detect novel associations. It is notable that the shared genetic component between celiac disease and breast cancer was not driven by strong signals (e.g. SNPs surpassing stringent GWAS threshold), but rather determined by several weaker signals, namely ‘suggestive variants’. Although this type of variants are typically not the ones identified in conventional genome-wide or candidate gene association studies, they may still be indicative of biological importance. A notable limitation is that the custom SNP chips used by the respective consortia target different regions of the genome, which reduces comparability, even when imputation of breast cancer genotypes is used. We also explored the use of methods such as Direct Imputation of Summary Statistics (DIST)$^{56}$ on celiac disease dataset, which increased the number of common variants with good quality imputation (INFO > 0.9), to approximately 500 K. However, the use of this method did not improve the comparability between the two diseases. After LD-based pruning as performed in SECA analysis, ~14 K independent SNP remained for comparison. Despite its restrictions, the ImmunoChip provides information on the most important genetic components of celiac disease (mainly at both HLA and no-HLA regions) and therefore can be used to highlight an otherwise undermined immunogenic role in breast cancer susceptibility. Our analysis should not be regarded as a full genome assessment of the genetic overlap between the two diseases, but rather as an assessment of the shared genetic variation of immune-related regions. If we had had access to celiac disease individual-level genotype data, genetic correlation analysis using other robust methods such as the GCTA-GREML$^{57}$ would have been possible. It is however unclear whether imputation based on raw data could allow for a more comprehensive comparison of the genetic variation between the two diseases. It is also possible that deeper genome coverage could improve the assessment of the genetic overlap and facilitate the identification of common causal variants.

In summary, we show evidence of a shared genetic component underlying the link between the two diseases at immune-related regions. The protective effect associated to higher load of celiac disease genetic susceptibility, summarized by the celiac polygenic risk score in a Swedish cohort, suggest that a less responsive immune system is implicated in the predisposition to breast cancer. While considering that our analyses were constrained by the immune-related genomic coverage, we used functional annotation analyses to identify genetic loci known to be involved in the complex regulation of the innate immune response which are likely to underlie common etiological basis between the two diseases. Replication of our findings and refined analysis related to disease subtypes will require larger samples sizes and better genotype data. Functional analyses integrating other layers of Omics data...
will be helpful to identify and validate specific mechanisms underlying breast cancer development, and possibly shed light on breast cancer prevention and treatment strategies.

**Methods**

**Genetic correlation and overlap tests using GWAS summary statistics.** Genetic correlation was estimated using the cross-trait LD score regression (LDSC) software (v1.0.0) on matching SNPs surpassing LDSC filter procedure. Given that imputation quality correlates with LD score, HapMap3 SNPs with European MAF > 1% (wa_hm3.snplist) were filtered with the -merge-allele flag. LDSC defines genetic correlation between a pair of traits as the genetic covariance normalized by heritabilities on each phenotype accounted by the genotyped variants (SNPs) across the genome. Genetic covariance is estimated under a model where standardized genotype effects sizes are treated as random and is, in practice, estimated by regressing z-scores on sample size weighted linkage disequilibrium (LD) scores. In our analyses, we only included SNPs with reliable LD scores available in the in-software file (wa_hm3.snplist), comprised of 1,217,311 SNPs with pre-computed LD scores estimated from European-ancestry samples in the 1 KG reference panel (see online Methods in Bulik-Sullivan et al.58), using the ‘merge-alleles’ flag.

SNP Effect Concordance Analysis (SECA) was used to test genetic overlap and SNP effect direction (analogous to genetic correlation tested by LDSC above). For each pair of datasets (celiac disease against overall, ER-positive, and ER-negative breast cancer), matching SNP were selected through SECA filtering and alignment procedures. Independent (index) SNPs were selected by SECA through a two-step ‘merge-value informed’ LD-clumping procedure (first round: pairwise LD $r^2 > 0.1$ within 1 Mb windows; second round: pairwise LD $r^2 > 0.1$ within 10 Mb windows) based on 1 KG v3 CEU (637 rsIDs; MAF > 1%). Following SECA scripts, pleiotropy tests were performed for each dataset pair on 144 subsets defined by combinations of 12 × 12 P-value thresholds on breast cancer ($P_{BC}$) and celiac datasets ($P_{CD}$), that is $[P_{BC}, P_{CD}] = \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. Genetic overlap was analyzed using binomial tests (BT) to determine whether there is an excess of lower P-values, the ‘overlap’ null probability (e.g. ‘expected proportion’) is defined as the estimated using the cross-trait LD score regression (LDSC) software (v1.0.0) on matching SNPs surpassing LDSC filter procedure. Given that imputation quality correlates with LD score, HapMap3 SNPs with European MAF > 1% (wa_hm3.snplist) were filtered with the -merge-allele flag. LDSC defines genetic correlation between a pair of traits as the genetic covariance normalized by heritabilities on each phenotype accounted by the genotyped variants (SNPs) across the genome. Genetic covariance is estimated under a model where standardized genotype effects sizes are treated as random and is, in practice, estimated by regressing z-scores on sample size weighted linkage disequilibrium (LD) scores. In our analyses, we only included SNPs with reliable LD scores available in the in-software file (wa_hm3.snplist), comprised of 1,217,311 SNPs with pre-computed LD scores estimated from European-ancestry samples in the 1 KG reference panel (see online Methods in Bulik-Sullivan et al.58), using the ‘merge-alleles’ flag.

LDSC and SECA differ in their approaches to evaluating whether pairs of phenotypes have a shared genetic basis. LDSC is model based and treats genotype effects as random, whilst SECA is based on a fixed effects basis. LDSC filter procedure. Given that imputation quality correlates with LD score, HapMap3 SNPs with European MAF > 1% (wa_hm3.snplist) were filtered with the -merge-allele flag. LDSC defines genetic correlation between a pair of traits as the genetic covariance normalized by heritabilities on each phenotype accounted by the genotyped variants (SNPs) across the genome. Genetic covariance is estimated under a model where standardized genotype effects sizes are treated as random and is, in practice, estimated by regressing z-scores on sample size weighted linkage disequilibrium (LD) scores. In our analyses, we only included SNPs with reliable LD scores available in the in-software file (wa_hm3.snplist), comprised of 1,217,311 SNPs with pre-computed LD scores estimated from European-ancestry samples in the 1 KG reference panel (see online Methods in Bulik-Sullivan et al.58), using the ‘merge-alleles’ flag.

SNP Effect Concordance Analysis (SECA) was used to test genetic overlap and SNP effect direction (analogous to genetic correlation tested by LDSC above). For each pair of datasets (celiac disease against overall, ER-positive, and ER-negative breast cancer), matching SNP were selected through SECA filtering and alignment procedures. Independent (index) SNPs were selected by SECA through a two-step ‘merge-value informed’ LD-clumping procedure (first round: pairwise LD $r^2 > 0.1$ within 1 Mb windows; second round: pairwise LD $r^2 > 0.1$ within 10 Mb windows) based on 1 KG v3 CEU (637 rsIDs; MAF > 1%). Following SECA scripts, pleiotropy tests were performed for each dataset pair on 144 subsets defined by combinations of 12 × 12 P-value thresholds on breast cancer ($P_{BC}$) and celiac datasets ($P_{CD}$), that is $[P_{BC}, P_{CD}] = \{0.01, 0.05, 0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$. Genetic overlap was analyzed using binomial tests (BT) to determine whether there is an excess of lower P-values, the ‘overlap’ null probability (e.g.‘expected proportion’) is defined as the estimated using the cross-trait LD score regression (LDSC) software (v1.0.0) on matching SNPs surpassing LDSC filter procedure. Given that imputation quality correlates with LD score, HapMap3 SNPs with European MAF > 1% (wa_hm3.snplist) were filtered with the -merge-allele flag. LDSC defines genetic correlation between a pair of traits as the genetic covariance normalized by heritabilities on each phenotype accounted by the genotyped variants (SNPs) across the genome. Genetic covariance is estimated under a model where standardized genotype effects sizes are treated as random and is, in practice, estimated by regressing z-scores on sample size weighted linkage disequilibrium (LD) scores. In our analyses, we only included SNPs with reliable LD scores available in the in-software file (wa_hm3.snplist), comprised of 1,217,311 SNPs with pre-computed LD scores estimated from European-ancestry samples in the 1 KG reference panel (see online Methods in Bulik-Sullivan et al.58), using the ‘merge-alleles’ flag.

LDSC and SECA differ in their approaches to evaluating whether pairs of phenotypes have a shared genetic basis. LDSC is model based and treats genotype effects as random, whilst SECA is based on a fixed effects basis. LDSC filter procedure. Given that imputation quality correlates with LD score, HapMap3 SNPs with European MAF > 1% (wa_hm3.snplist) were filtered with the -merge-allele flag. LDSC defines genetic correlation between a pair of traits as the genetic covariance normalized by heritabilities on each phenotype accounted by the genotyped variants (SNPs) across the genome. Genetic covariance is estimated under a model where standardized genotype effects sizes are treated as random and is, in practice, estimated by regressing z-scores on sample size weighted linkage disequilibrium (LD) scores. In our analyses, we only included SNPs with reliable LD scores available in the in-software file (wa_hm3.snplist), comprised of 1,217,311 SNPs with pre-computed LD scores estimated from European-ancestry samples in the 1 KG reference panel (see online Methods in Bulik-Sullivan et al.58), using the ‘merge-alleles’ flag.

**Association of celiac-PRS with breast cancer risk.** We constructed celiac disease PRS profiles (celiac-PRS) using individual level genotype data for subjects in the pKARMA study genotyped as part of the iCOGS initiative. pKARMA is made up of 5,002 invasive breast cancer cases (from the Linne-Brost 1 (Libro1) study) and 5,433 controls (from the Karolinska Mammography Project for Risk Prediction of Breast Cancer (KARMA)63). Libro1 consists of female primary breast cancer cases diagnosed in Stockholm between January 2001 and December 2008 identified via the Regional Cancer register65. Tumor characteristics were retrieved from the Stockholm-Gotland Regional Breast Cancer quality registry65. ER status was recorded as negative or positive, determined by radioimmunoassay or immunohistochemistry. Tumor size was categorized as < 20, 20–40 and > 40 in diameter (mm). Human epidermal growth factor receptor 2 (HER2) status, assessed by IHC/immunocytochemistry and confirmed by fluorescence in situ hybridization analysis if protein levels from IHC/immunocytochemistry showed 2+ or 3+, was recorded in the register as positive or negative. Lymph node involvement status was dichotomized (No/Yes). Registry information was essentially complete (98%) for tumor size and lymph node status, but with more missing data for ER status (80% complete). Grade was available from 2004 onward, with 93% completeness. Controls were breast cancer-free participants recruited between 2010 and 2011 from Helsingborg and Stockholm in Sweden, a subset of the KARMA study. All participants had been genotyped on the iCOGS array in accordance with relevant guidelines as described previously26 and missing genotypes were imputed using 1 KG (phase 1 integrated variant set release v3) in the National Center of Biotechnology Information build 37 [hg19 coordinates]. Each participant gave informed consent and this study has been approved by the ethical review board at Karolinska Institutet.

Celiac-PRS profiles for each individual were generated by summing the number of celiac disease risk allele copies weighted by effect estimates reported on the GWAS study by Trynka et al.17, using a scoring routine in the PLINK program (version 1.9b3893. We computed four celiac-PRS profiles based on subsets of independent ($r^2 > 0.2$) celiac disease SNPs defined by different P-value thresholds [$P_{CD} < 5E-08 (n = 199)$, $P_{CD} < 1E-05 (n = 276)$, $P_{CD} < 0.01 (n = 1,284)$, and $P_{CD} < 0.05 (n = 3,803)$].
Statistical analyses were performed in R (version 3.2.4). Unconditional logistic regressions were used to estimate ORs and corresponding 95% CI interval for association of celiac-PRS with overall, ER-negative, and ER positive breast cancer risk. PRS profiles were tested as both a continuous variable per standard deviation (per-SD) and a categorical variable defined by quartiles (based on PRS distribution in breast cancer controls), with the lowest quartile as the reference. We also investigated whether celiac-PRS differentially influences breast cancer tumor characteristics in a case-only study: ER status, lymph node involvement, and HER2-status were tested as binary outcomes using binominal logistic regressions; tumor grade and tumor size were modeled as categorical variables using multinominal logistic regressions (with the "nnet" R package).

**Enrichment analysis on top-overlapping SNPs.** To aid in the biological interpretation of top (lowest P)-values overlapping SNPs between celiac disease and breast cancer risk overall, we performed SNP enrichment analysis using DEPICT (version 1r1194)30, an integrative tool that systematically prioritizes the most likely causal genes at associated loci and highlights enriched pathways based on a pre-computed probability of gene set membership across 14,461 reconstituted gene sets. Loci within base pairs 25,000,000–35,000,000 on chromosome 6 are excluded due to the heightened LD seen on this major histocompatibility region.

**References**

1. Askling, J. et al. Cancer incidence in a population-based cohort of individuals hospitalized with celiac disease or dermatitis herpetiformis. *Gastroenterology* **123**, 1428–35 (2002).
2. Ludvigsson, J. F., West, J., Ekborn, A. & Stephansson, O. Reduced risk of breast, endometrial and ovarian cancer in women with celiac disease. *Int J Cancer* **131**, E244–50 (2012).
3. Hemminki, K. et al. Effect of autoimmune diseases on risk and survival in female cancers. *Gynecol Oncol* **127**, 180–5 (2012).
4. Viljamaa, M. et al. Malignancies and mortality in patients with coeliac disease and dermatitis herpetiformis: 30-year population-based study. *Dig Liver Dis* **38**, 374–80 (2006).
5. Card, T. R., West, J. & Holmes, G. K. Risk of malignancy in diagnosed coeliac disease: a 24-year prospective, population-based, cohort study. *Aliment Pharmacol Ther* **20**, 769–75 (2004).
6. West, J., Logan, R. F., Smith, C. J., Hubbard, R. B. & Card, T. R. Malignancy and mortality in people with coeliac disease: population based cohort study. *BMJ* **329**, 716–9 (2004).
7. Goldacre, M. J., Wotton, C. J., Yateas, D., Seagratt, V. & Jewell, D. Cancer in patients with ulcerative colitis, Crohn's disease and coeliac disease: record linkage study. *Eur J Gastroenterol Hepatol* **20**, 297–304 (2008).
8. Silano, M. et al. Delayed diagnosis of coeliac disease increases cancer risk. *BMJ Gastroenterol* **7**, 8 (2007).
9. Ludvigsson, J. et al. The Oslo definitions for coeliac disease and related terms. *Gut* **62**, 43–52 (2013).
10. Maki, M. et al. Prevalence of celiac disease among children in Finland. *New England Journal of Medicine* **348**, 2317–2524 (2003).
11. Jang, X. & Shapiro, D. J. The immune system and inflammation in breast cancer. *Mol Cell Endocrinol* **382**, 673–82 (2014).
12. Need, E. F., Atashgarian, V., Inqman, W. V. & Dasari, P. Hormonal regulation of the immune microenvironment in the mammary gland. *J Mammary Gland Biol Neoplasia* **19**, 229–39 (2014).
13. Czene, K., Lichtenstein, P. & Hemminki, K. Environmental and heritable causes of cancer among 9.6 million individuals in the Swedish family-cancer database. *International Journal of Cancer* **99**, 260–266 (2002).
14. Mucci, L. A. et al. Familial Risk and Heritability of Cancer Among Twins in Nordic Countries. *Jama-Journal of the American Medical Association* **315**, 68–76 (2016).
15. Kuja-Halkola, R. et al. Heritability of non-HLA genetics in coeliac disease: a population-based study in 107,000 twins. *Gut* (2016).
16. Visscher, P. M., Brown, M. A., McCarthy, M. I. & Yang, J. Five years of GWAS discovery. *Am J Hum Genet* **90**, 7–24 (2012).
17. Trynka, G. et al. Dense genotyping identifies and localizes multiple common and rare variant association signals in celiac disease. *Nat Genet* **43**, 1193–201 (2011).
18. Michailidou, K. et al. Genome-wide association analysis of more than 120,000 individuals identifies 15 new susceptibility loci for breast cancer. *Nature Genetics* **47**, 373–12U17 (2015).
19. Soloviiev, N., Cotsapas, C., Lee, P. H., Purcell, S. M. & Smoller, J. W. Pleiotropy in complex traits: challenges and strategies. *Nat Rev Genet* **14**, 483–95 (2013).
20. Pendergrass, S. A. et al. Phenome-wide association study (PheWAS) for detection of pleiotropy within the Population Architecture using Genomics and Epidemiology (PAGE) Network. *PLoS Genet* **9**, e1003807 (2013).
21. Sivakumaran, S. et al. Abundant pleiotropy in human complex diseases and traits. *Am J Hum Genet* **89**, 607–18 (2011).
22. Gratten, J. & Visscher, P. M. Genetic pleiotropy in complex traits and diseases: implications for genomic medicine. *Genome Med* **8**, 79 (2016).
23. Traylor, M. et al. Shared genetic contribution to Ischaemic Stroke and Alzheimer's Disease. *Ann Neurol* (2016).
24. H. F. Zhuang, Q. S. & Shen, L. Genetic overlap between type 2 diabetes and major depressive disorder identified by bioinformatics analysis. *Oncotarget* **7**, 17410–17414 (2016).
25. Clarke, T. K. et al. Investigating shared aetiology between type 2 diabetes and major depressive disorder in a population based cohort. *Am J Med Genet B Neuropsychiatr Genet* (2016).
26. Michailidou, K. et al. Large-scale genotyping identifies 41 new loci associated with breast cancer risk, *Nat Genet* **45**, 535–61, 361e1–2 (2013).
27. Cortes, A. & Brown, M. A. Promise and pitfalls of the Immunochip. *Arthritis Res Ther* **13**, 101 (2011).
28. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat Genet* **47**, 1236–41 (2015).
29. Nyholt, D. R. SECA: SNP effect concordance analysis using genome-wide association summary results. *Bioinformatics* **30**, 2086–8 (2014).
30. Pers, T. H. et al. Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications* **6** (2015).
31. Welander, A., Tjernberg, A. R., Montgomery, S. M., Ludvigsson, J. & Ludvigsson, J. F. Infectious Disease and Risk of Later Celiac Disease in Childhood. *Pediatrics* **125**, E530–E536 (2010).
32. Stordal, K., White, R. A. & Eggesbo, M. Early Feeding and Risk of Celiac Disease in a Prospective Birth Cohort. *Pediatrics* **132**, E1202–E1209 (2013).
33. A. Aronsson, C. A. et al. Age at Gluten Introduction and Risk of Celiac Disease. *Pediatrics* **135**, 239–245 (2015).
34. Vriezinga, S. L. et al. Randomized Feeding Intervention in Infants at High Risk for Celiac Disease. *New England Journal of Medicine* **371**, 1304–1315 (2014).
35. Lionetti, E. et al. Introduction of Gluten, HLA Status, and the Risk of Celiac Disease in Children. *New England Journal of Medicine* **371**, 1295–1303 (2014).
36. Liu, E. et al. Risk of Pediatric Celiac Disease According to HLA Haplotype and Country. *New England Journal of Medicine* **371**, 42–49 (2014).
37. Garcia-Martinez, E. et al. Tumor-infiltrating immune cell profiles and their change after neoadjuvant chemotherapy predict response and prognosis of breast cancer. *Breast Cancer Research* **16** (2014).
38. Strayer, D. R., Carter, W. A. & Brodsky, I. Familial occurrence of breast cancer is associated with reduced natural killer cytotoxicity. *Breast Cancer Res Treat* 7, 87–92 (1986).
39. Quan, L. et al. Cytokine and cytokine receptor genes of the adaptive immune response are differentially associated with breast cancer risk in American women of African and European ancestry. *International Journal of Cancer* 134, 1408–1421 (2014).
40. Zeng, J., Fang, Y. & Li, P. Y. FAS-1377 A/G polymorphism in breast cancer: a meta-analysis. *Tumor Biology* 35, 2575–2581 (2014).
41. Locke, A. E. et al. Genetic studies of body mass index yield new insights for obesity biology. *Nature* 518, 197–206 (2015).
42. Mahmoud, S. M. et al. Tumor-infiltrating CD8+ lymphocytes predict clinical outcome in breast cancer. *J Clin Oncol* 29, 1949–55 (2011).
43. Ibrahim, E. M., Al-Foheidi, M. E., Al-Mansour, M. M. & Kazed, G. A. The prognostic value of tumor-infiltrating lymphocytes in triple-negative breast cancer: a meta-analysis. *Breast Cancer Research and Treatment* 148, 467–476 (2014).
44. Paterno, G. D. et al. Molecular cloning of human e1 RNA and its differential expression in breast tumours and tumour-derived cell lines. *Gene* 222, 77–82 (1998).
45. Ding, Z. H., Gillespie, I. L. & Paterno, G. D. Human MI-ER1 alpha and beta function as transcriptional repressors by recruitment of histone deacetylase 1 to their conserved ELM2 domain. *Molecular and Cellular Biology* 23, 250–258 (2003).
46. Li, S. N., Paterno, G. D. & Gillespie, I. L. Nuclear Localization of the Transcriptional Regulator MIER1 alpha Requires Interaction with HDAC1 in Breast Cancer Cells. *Plos One* 8 (2013).
47. McCarthy, P. L. et al. Changes in subcellular localisation of MI-ER1 alpha, a novel oestrogen receptor-alpha interacting protein, is associated with breast cancer progression. *Br J Cancer* 99, 639–46 (2008).
48. Eleftheriadis, T., Antoniadis, G., Liakopoulos, V. & Kortsaris, A. T-Cell zeta chain expression, phosphorylation and degradation and their role in T-cell signal transduction and immune response regulation in health and disease. *Current Signal Transduction Therapy* 1, 191–208 (2006).
49. Trinchieri, G. et al. Natural killer cell stimulatory factor (NKSF) or interleukin-12 is a key regulator of immune response and inflammation. *Prog Growth Factor Res* 4, 355–68 (1992).
50. Luen, S., Virassamy, B., Savas, P., Salgado, R. & Loi, S. The genomic landscape of breast cancer and its interaction with host immunity. *Breast* 29, 241–250 (2016).
51. Lee, S. H., Yang, J., Goddard, M. E., Visscher, P. M. & Wray, N. R. Estimation of pleiotropy between complex diseases using single-nucleotide polymorphism-derived genomic relationships and restricted maximum likelihood. *Bioinformatics* 28, 2540–2542 (2012).
52. Bulik-Sullivan, B. K. et al. LD Score regression distinguishes confounding from polygenicity in genome-wide association studies. *Nat Genet* 47, 291–5 (2015).
53. Strayer, D. R., Carter, W. A. & Brodsky, I. Familial occurrence of breast cancer is associated with reduced natural killer cytotoxicity. *Breast Cancer Res Treat* 7, 87–92 (1986).
54. Strayer, D. R., Carter, W. A. & Brodsky, I. Familial occurrence of breast cancer is associated with reduced natural killer cytotoxicity. *Breast Cancer Res Treat* 7, 87–92 (1986).
55. Strayer, D. R., Carter, W. A. & Brodsky, I. Familial occurrence of breast cancer is associated with reduced natural killer cytotoxicity. *Breast Cancer Res Treat* 7, 87–92 (1986).
56. Strayer, D. R., Carter, W. A. & Brodsky, I. Familial occurrence of breast cancer is associated with reduced natural killer cytotoxicity. *Breast Cancer Res Treat* 7, 87–92 (1986).
57. Strayer, D. R., Carter, W. A. & Brodsky, I. Familial occurrence of breast cancer is associated with reduced natural killer cytotoxicity. *Breast Cancer Res Treat* 7, 87–92 (1986).
58. Strayer, D. R., Carter, W. A. & Brodsky, I. Familial occurrence of breast cancer is associated with reduced natural killer cytotoxicity. *Breast Cancer Res Treat* 7, 87–92 (1986).
59. Strayer, D. R., Carter, W. A. & Brodsky, I. Familial occurrence of breast cancer is associated with reduced natural killer cytotoxicity. *Breast Cancer Res Treat* 7, 87–92 (1986).
60. Strayer, D. R., Carter, W. A. & Brodsky, I. Familial occurrence of breast cancer is associated with reduced natural killer cytotoxicity. *Breast Cancer Res Treat* 7, 87–92 (1986).
Author Contributions
K.C. and J.L. conceived and directed the research; E.U.-M. performed data analysis and together with J.L. wrote the manuscript. K.H., H.Y., and J.F.L. provided support in the data analysis. K.C. and P.H. are responsible for the pKARMA cohort. All authors participated in the interpretation of the results and production of the final manuscript.

Additional Information
Supplementary information accompanies this paper at doi:10.1038/s41598-017-06287-9

Competing Interests: The authors declare that they have no competing interests.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2017