Integrating gene expression and epidemiological data for the discovery of genetic interactions associated with cancer risk

Núria Bonifaci1, Eva Colas2, Jordi Serra-Musach1,3, Nazanin Karbalai4,5, Joan Brunet1, Antonio Gómez6, Manel Esteller6,8, Enrique Fernández-Taboada1, Antoni Berenguer9, Jaume Reventós1,10, Bertram Müller-M Bryso2,11, Laufey Amundadottir12, Eric J.Duell13 and Miquel Àngel Pujana1,4,6

1Breast Cancer and Systems Biology Unit, Translational Research Laboratory, Catalan Institute of Oncology (ICO), Bellvitge Institute for Biomedical Research (IDIBELL), L’Hospitalet del Llobregat, Barcelona 08908, Catalonia, Spain, 2Research Unit in Biomedicine and Translational and Pediatric Oncology, Vall d’Hebron Research Institute and Hospital, and Autonomous University of Barcelona, Barcelona 08035, Catalonia, Spain, 3Hereditary Cancer Programme, ICO, Girona Biomedical Research Institute (IDIBGI), Girona 17007, Catalonia, Spain, 4Statistical Genetics, Max Planck Institute of Psychiatry, Munich 80804, Germany, 5Munich Cluster for Systems Neurology (SyNergy), Munich 80804, Germany, 6Cancer Epigenetics and Bioinformatics Program (PBC), IDIBELL, L’Hospital del Llobregat, Barcelona 08908, Catalonia, Spain, 7Department of Physiological Sciences II, School of Medicine, University of Barcelona, Barcelona 08098, Catalonia, Spain, 8Unit of Biomarkers and Susceptibility, ICO, IDIBELL, L’Hospital del Llobregat, Barcelona 08908, Catalonia, Spain, 9Catalan Institution for Research and Advanced Studies (ICREA), Barcelona 08010, Catalonia, Spain, 10Unit of Biomarkers and Susceptibility, ICO, IDIBELL, L’Hospital del Llobregat, Barcelona 08908, Catalonia, Spain, 11Basic Sciences Department, International University of Catalonia, Barcelona 08017, Catalonia, Spain, 12Munich Cluster for Systems Neurology (SyNergy), Munich 80336, Germany, 13Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Bethesda, MD 20892, USA and 14Unit of Nutrition, Environment and Cancer, Cancer Epidemiology Program, ICO, IDIBELL, L’Hospital del Llobregat, Barcelona 08908, Catalonia, Spain

*To whom correspondence should be addressed. Tel: +34 932607463; Fax: +34 932607466; Email: mapujana@iconologia.net

Dozens of common genetic variants associated with cancer risk have been identified through genome-wide association studies (GWASs). However, these variants only explain a modest fraction of the heritability of disease. The missing heritability has been attributed to several factors, among them the existence of genetic interactions (G × G). Systematic screens for G × G in model organisms have revealed their fundamental influence in complex phenotypes. In this scenario, G × G overlap significantly with other types of gene and/or protein relationships. Here, by integrating predicted G × G from GWAS data and complex- and context-defined gene coexpression profiles, we provide evidence for G × G associated with cancer risk. G × G predicted from a breast cancer GWAS dataset identified significant overlaps [relative enrichments (REs) of 8–36%], empirical P values < 0.05 to 10^{-4} with complex (non-linear) gene coexpression in breast tumors. The use of gene or protein data not specific for breast cancer did not reveal overlaps. According to the predicted G × G, experimental assays demonstrated functional interplay between lipoma-preferred partner and transforming growth factor-β signaling in the MCF10A non-tumorigenic mammary epithelial cell model. Next, integration of pancreatic tumor gene expression profiles with pancreatic cancer G × G predicted from a GWAS corroborated the observations made for breast cancer risk (REs of 25–59%). The method presented here can potentially support the identification of genetic interactions associated with cancer risk, providing novel mechanistic hypotheses for carcinogenesis.

Introduction

Several genome-wide association studies (GWASs) have been completed that delineate the common genetic basis of cancer risk (1). The gene candidates identified in these studies have considerably expanded the biological knowledge of cancer etiology. These advances are being followed up by projects that aim to identify the corresponding genetic mutations and to improve cancer risk prediction. However, in most cases, the results of GWASs (in addition to complementary candidate approaches) have not yet identified the bulk of disease risk heritability. For example, to date, 79 low-penetrance loci have been identified in breast cancer, but together they account for only a modest percentage (~15%) of the familial relative risk. If moderately and highly penetrant mutations/genes are included, ~50% of the familial relative risk remains unexplained (2,3).

Numerous factors or modeling approaches can explain the problem of ‘missing heritability’ (4–6). Notably, the recent meta-analysis of several breast cancer GWAS revealed an excess of significant association signals (not reaching genome-wide significance) that suggests that >1000 loci are involved in susceptibility, each of which exerts a very small effect (7). This modeling did not take into account genetic interactions (G × G), which have also been suggested to explain part of the missing heritability (4–6). In this regard, the identification of interactions could potentially improve the accuracy of risk models and improve cancer prevention (8,9). Several methods have been developed for exhaustive searching of statistical interactions in data from GWASs (10,11). These analyses (limited to two locus interactions) are time consuming but computationally achievable. However, the vast number of loci pairs raises the issue of multiple testing, which limits the identification of true interactions based on only statistical terms. In addition, the translation of the statistical findings to biological interactions or models is unclear (12) and potentially complex (13).

Systematic analyses in model organisms have shown that, in many cases, a given phenotype is explained not simply by additive allele effects but also by G × G (or epistasis in statistical terms; i.e. deviation from additivity for a quantitative phenotype by the effect of a genetic variant or mutation in a different locus) (14). Importantly, studies in yeast with ~6000 annotated genes have predicted the existence of hundreds of thousands of G × G (15). It could therefore be hypothesized that G × G are of similar biological relevance in humans (9). Their relevance is also based on the identification of synthetic lethal interactions for specific mutations in cancer (16). However, the methodology to systematically screen for mammalian G × G has only recently been described (17,18). On the basis of previous evidence that G × G inform about other types of molecular or functional relationships between genes and/or proteins (genes/proteins), we hypothesize that a genome-wide integrative strategy could help to discover G × G associated with cancer risk.

Materials and methods

Genetic data and G × G analysis

The National Cancer Institute has conducted GWASs to identify common genetic variants and the corresponding candidate genes associated with cancer risk, which included breast (19) and pancreatic (20,21) cancer. For breast cancer, the initial GWAS by the Cancer Genetic Markers of Susceptibility initiative was designed to identify variants with a significant marginal effect in postmenopausal women. The study involved 1145 invasive postmenopausal breast cancer cases and 1142 matched controls from the Nurses’ Health Study. The GWAS data was obtained upon approval of a Data Access Request to dbGAP (http://cancer.gov/data/). Missing genotypes were imputed using the MACH software (22). The GWAS data for pancreatic cancer was also obtained upon approval of a Data Access Request to dbGAP and analyzed for specific variants (i.e. variants selected by their identifier and/or location in...
a specific gene locus, rather than randomly selected from a given gene rank), gene pair bins (bins defined from gene pairs ranked according to their complex coexpression in tumors; each bin corresponds to 20 gene pairs, starting from the highest coexpression value), and G × G using the dbGAP PLINK (http://purl.org/mbi/darray/purcell/plink/) (23) file. This file was used to prevent potential differences in reprocessing the original GWAS data and contained 914 cases and 1027 controls. The whole genome screen for breast cancer G × G was carried out using EPIBLASTER (24). A two-stage analytical process was implemented: first, all pairwise single nucleotide polymorphism (SNP) combinations (considering only allelic informative and gene-centered mapped SNPs) were assessed for the Pearson’s correlation coefficient (PCC) difference between cases and controls; second, the likelihood ratio test of the logistic regression (LR) was applied to those subsets of SNP pairs deemed significant in the previous stage. Using simulated and real data sets, the method was shown previously to conduct a search for G × G that was unbiased to the marginal loci effects and captured most of the real G × G (24).

The quality controls for the use of SNP data in this analysis were: minor allele frequency > 0.05 and P value cutoff of 10−5 for the Hardy–Weinberg equilibrium test. For linking SNPs to genes, each SNP was assigned to a specific gene locus if the variant mapped to a region ±10kb from the corresponding genomic structure (first and last exon), using the ENSEMBL human genome release 57. Since the analysis required unambiguous SNP gene correspondences, the SNPs that overlapped with two or more gene loci were excluded. The pairs of SNPs with some evidence of linkage disequilibrium (LD; r2 > 0.2) were also excluded from the analysis.

Gene and protein data analyses

Whole genome expression data for primary breast (25–27) (NKI-295 data set and Gene Expression Omnibus reference GSE2034), colorectal (Gene Expression Omnibus reference GSE14333, ref. 28) and pancreatic (Gene Expression Omnibus reference GSE69294, ref. 29) tumors was analyzed using the preprocessed and normalized values. The NKI-295 breast tumors data set contained 69 estrogen receptor α (ERα)-negative and 22 ERα-positive tumors. The colorectal and pancreatic datasets included 290 and 91 tumors, respectively. The PCCs were computed in R software and the mutual information (MI) was estimated using the ARACNE approach that applies a Gaussian kernel estimator (30). No pruning of MI-based edges (directed at specifically identified transcription factor–target interactions) was performed. Release #7 of the Human Protein Reference Database (31) was used, which contains 9461 proteins and 37 081 interactions that mainly represent experimentally demonstrated interactions compiled through literature curation. The high-confidence interactions dataset was derived from the integration of diverse data and contained 7401 proteins and 20 614 interactions (32).

Gene Ontology analyses

The Gene Ontology (GO) Biological Processes term annotations were downloaded from the Open Biological Ontologies release 2012/06 (MySQL version). GO terms were assigned to gene symbols after record linkage in which regular expression searches were required. Genes annotated at level 5 or lower in the hierarchy were assigned to level 4, but those also occurring at level 3 were excluded. Homodimers and gene pairs where both members share a GO annotation were also excluded. Only those term pairs with a frequency ≥15% in the test set were evaluated. The test sets were 173 gene pairs from the 205 predicted G × G in breast cancer, and 82 gene pairs from the significant (P < 0.05) G × G identified in the 14 highest-ranked bins for pancreatic cancer. Significance was assessed by comparing the observed frequency of each term–term interaction in the test set with the null distribution obtained by randomly selecting equivalent gene pairs (1000 sets of similar gene pair size to the test set) from the top 0.25% MI values in the breast cancer setting, or from the top 10 000 gene pairs (according to their MI values) in the pancreatic cancer setting.

Cell culture and short hairpin RNAs

The MCF10A cells were obtained from the American Type Culture Collection, cultured in HuMEC (Invitrogen) media supplemented (hereafter ‘supplemented media’, in contrast to ‘non-supplemented’) with HuMEC Supplement and Bovine Pituitary Extract (Life Technologies) and used with <10 passages (from the initial American Type Culture Collection vial) for all assays. The short hairpin RNA (shRNA) used for depletion of lipoma-preferred partner (LPP) expression was the validated MISSION catalog TRCN0000301828 (Sigma–Aldrich). The lentiviral packaging, envelope, control and green fluorescent protein expression plasmids (psPAX2, pMD2.G, non-hairpin plLKO.1, scrambled-plLKO.1 and pWPT-GFP) were purchased from Addgene. Production and collection of lentiviral particles followed a modified Addgene protocol. Initial viral titers > 5 × 105/ml were confirmed by Lenti-X GoStix (Clontech) and supernatants were then concentrated by ultracentrifugation or Lenti-X Concentrator (Clontech) and stored at ~80°C. Concentrated viral supernatants were titrated for optimal inhibition of the target. Cells were infected with viral supernatants in the presence of 8 µg/ml polybrene and, after 48 h, incubated with puromycin to select stable populations of MCF10A control (shRNA control) or LPP-depleted cells (shRNA–LPP).

Proliferation, wound-healing and spheroids assays

Cells (5 × 104) transduced and selected for shRNA control or shRNA–LPP were plated in triplicate in 96-well plates with complete medium. After 24 h, adherent cells were cultured with supplemented and non-supplemented media in the presence or absence of transforming growth factor-β1 (TGFβ1) (100 pM in 1x cell culture media). Cell proliferation was measured at the time of replacing media at 24, 48 and 72 h, and using the CellTiter 96 AQueous One Solution Cell Proliferation Assay (Promega). For wound-healing assays, cells transduced and selected for shRNA control or shRNA–LPP were plated at confluence in duplicate on 24-well plates and incubated overnight. A straight line was then gently performed at the bottom of the dish. Cells were washed and incubated in non-supplemented media. After 24 h, transmission images were captured for each cell line using a FSX100 microscope (Olympus). Images were analyzed using the ImageJ software (Wright Cell Imaging Facility). Initial and final wound area (mm2) were the variables used to calculate the wound closure percentage. For spheroids assays, cells (5 × 104) transduced and selected for shRNA control or shRNA–LPP were plated in 20 µl drops of supplemented media and allowed to grow in suspension. After 4-6 days, transmission images were captured using a FSX100 microscope (Olympus). This experiment was performed with a minimum of 20 drops per cell line.

Western blotting and antibodies

Whole cell extracts from cultures transduced and selected for shRNA control or shRNA–LPP, in the presence or absence of supplemented media and TGFβ1, were prepared and used for western blotting as described elsewhere (33). The primary antibodies used for blotting were rabbit anti-ITGAS (dilution 1:500, #AB1949; Millipore), anti-ITGB1 (dilution 1:2500, #610467; BD Pharmingen), anti-LPP (dilution 1:100, #00325-05; immunoGlobe) and anti-TUBA (dilution 1:2000, #2125; Cell Signaling).

Results

G × G overlap with complex context-defined gene coexpression

To compute G × G from a GWAS, we used data from the Cancer Genetic Markers of Susceptibility initiative in breast cancer (19) and applied a two-step analytical process (24). This strategy was implemented because it allows for fast and exhaustive computation of G × G. The first step consisted in calculating the difference in PCCs between controls and cases across all informative SNP pairs (Figure 1). The SNPs were those that were informative but also mapped to an annotated gene locus (i.e. a known gene). In the second step, significant SNP pairs (PCC difference > 10−5) were analyzed by LR. Thus, ~390 000 SNP pairs with no evidence of LD and corresponding to ~339 000 gene pairs were analyzed at this stage. Next, from the distribution of P LR values, a threshold P LR ≤ 10−6 was defined, which yielded a set of 39 417 gene pairs (Figure 1). This set of predicted G × G (i.e. G × G potentially associated with risk of breast cancer in the general population) was subsequently assessed for overlap with known gene/protein relationships (Figure 1).

Integrative analyses in yeast have shown that genome-wide experimentally identified G × G overlap with protein–protein interactions to a degree that is significantly higher than expected by chance, of 10–20% of the known protein–protein interactions (34). We therefore examined whether the predicted G × G overlap with human protein–protein interactions. No significant overlap relative to what would be expected at random was identified using either a compiled dataset from the literature (31) or a high-confidence subset (32); indeed, only nine literature-compiled protein–protein interactions were found to be in common with the predicted G × G (data not shown).

Next, as G × G may also overlap significantly with gene coexpression (34), expression data from a large series of breast tumors (25) was analyzed. First, a standard measure of coexpression (i.e. PCC) was computed. In this analysis, all possible microarray probe pairs were evaluated and the maximum PCC value was then selected for each gene pair. By ranking the gene pairs according to their PCCs and assessing the top bins (starting at the top 0.1% of PCCs, which corresponded to 200 744 gene pairs), no significant overlap was observed.
with the predicted G × G ($P_{\text{empirical}} = 0.73$ using 1000 permutations of the 39 417 G × G); indeed, an opposite trend (i.e. under enrichment) was observed for the first bin (Figure 2A).

As G × G represent complex genetic relationships, we next used the MI measure to assess the overlap with non-linear expression relationships. Notably, using the same breast tumor dataset, the top 0.1 and 0.5% of MI-ranked gene pairs overlapped significantly with the predicted G × G (Figure 2B and Supplementary Table S1, available at Carcinogenesis Online). In the top 0.1%, 54 gene pairs were in common, whereas in the top 0.5%, 205 gene pairs were in common, which corresponded to relative enrichments (REs) of 32% (1.32 ratio relative to random, $P_{\text{empirical}} = 0.019$) and 16% (1.16 ratio, $P_{\text{empirical}} = 0.015$), respectively (Figure 2B and Supplementary Table S1, available at Carcinogenesis Online). When the top 0.5% PCCs were excluded from these MI bins, a suggestion of higher enrichment was observed in the first bin: RE of 43%, $P_{\text{empirical}} = 0.009$, which corresponded to 43 gene pairs in common (Figure 2C and Supplementary Table S1, available at Carcinogenesis Online).

**Robustness of the overlap between G × G and context-defined gene coexpression**

As we observed an asymptotic distribution of $P_{LR}$ values (Figure 1) and it was computationally unfeasible to analyze ranks of billions of gene pairs, a threshold $P_{LR} \leq 10^{-6}$ was initially used. Nonetheless, a significant enrichment was also revealed at a higher threshold ($P_{LR} \leq 10^{-5}$): with the top 0.5% MI-ranked gene pairs, 1322 were found in common with the predicted G × G, which corresponded to a RE of 16% and $P_{\text{empirical}} = 1.1 \times 10^{-4}$. At this threshold, the top 0.5% PCC-ranked gene pairs showed some suggestion of enrichment, although with a lower magnitude: 1002 pairs in common, RE of 5% and $P_{\text{empirical}} = 0.043$. Moreover, the top 5% of MI-ranked gene pairs was also found to be significantly enriched, but, as expected, with a lower magnitude than the top 0.5%: 10 044 pairs in common, RE = 8% and $P_{\text{empirical}} = 1.0 \times 10^{-4}$.

![Fig. 1. Analytical strategy for the identification of G × G associated with breast cancer risk. The numbers of SNPs in the original GWAS dataset and in subsequent steps are shown. Also shown are the number of gene pairs at relevant analytical steps and the distribution of $P_{LR}$ values. The selection of 39 417 gene pairs representing predicted G × G is subsequently evaluated for their overlap with other gene/protein relationships.](image1)

![Fig. 2. Overlap between predicted breast cancer G × G and gene coexpression. (A) Overlap assessment with gene pairs ranked according to the highest PCCs computed from breast tumors (from top 0.1% to top 30% of pairs). The y-axis shows the REs. (B) Overlap assessment with gene pairs ranked according to the highest MIs computed from breast tumors (from top 0.1% to top 30% of pairs). The significant bins with the number of overlapping gene pairs are marked (red dots). (C) Overlap assessment with gene pairs ranked according to the highest MIs and subtracting those gene pairs included in the top 0.5% of PCCs from breast tumors. (D) Overlap assessment with gene pairs ranked according to the highest MIs computed from colorectal tumors.](image2)
The Cancer Genetic Markers of Susceptibility study was centered on sporadic postmenopausal breast cancer, hence most of the enrolled cases had developed tumors that were ERα-positive (19). Consequently, the explained risk was mainly for ERα-positive and not ERα-negative breast cancer. The MIs were therefore computed separately for ERα-positive and ERα-negative tumors and then examined for their overlap with the predicted $G \times G$ ($P_{LR} \leq 10^{-6}$). Having defined the top 0.1% of MI-ranked gene pairs in ERα-positive and ERα-negative tumors, the REs were 36 and −5%, $P_{\text{empirical}} = 0.041$ and 0.56, respectively. This enrichment for ERα-positive cases, which was slightly (but not significantly) higher than observed in the full tumor dataset (36 versus 32%), corresponded to 34 gene pairs in common.

Using another large breast cancer expression dataset (27), a similar enrichment to the above was revealed for the top MI-ranked gene pairs: with the $P_{LR} \leq 10^{-6}$ threshold, the REs of the top 0.5 and 5% MI-ranked gene pairs were 15 and 13%, $P_{\text{empirical}} = 0.081$ and 0.045, respectively. Although the RE estimation for the top 0.1% was similar (12%), it was not significant ($P_{\text{empirical}} = 0.27$) probably because the number of gene pairs contained in this set was relatively low ($n = 175,614$). Taking the 205 G × G predicted from the analysis of the first breast cancer dataset, the overlap with this second dataset was: 10 pairs at the top 0.1% of MIs; 24 at the top 0.5%; 30 at the top 1% and 43 at the top 5% (Supplementary Table S2, available at Carcinogenesis Online). Although this level of overlap was significant ($P_{\text{hypergeometric}} = 2.4 \times 10^{-6}$), the difference for the microarray platforms used in these studies may contribute to a substantial proportion of false-negative pairs.

Next, the overlap was assessed for complex gene coexpression in a distinct epithelial neoplasm, colorectal cancer (28). Using an expression dataset of a similar size to the breast cancer studies, no evidence of overlap was obtained at any MI threshold and with $P_{LR} \leq 10^{-6}$; the RE estimation for the top 0.1% of MI-ranked gene pairs, which also contained a similar number of gene pairs to the above study, was −1% (Figure 2D and Supplementary Table S1, available at Carcinogenesis Online). Therefore, $G \times G$ associated with risk of a given cancer type might only be predicted on the basis of complex gene expression relationships in the specific condition.

**Biological processes in the $G \times G$ associated with breast cancer risk**

No enrichment in significant marginal effects was observed in the set of SNPs involved in the predicted $G \times G$ (Supplementary Table S3, available at Carcinogenesis Online). However, the 205 gene pairs included four candidate genes identified in GWASs for breast cancer: $FTO$, $ITPR1$, $PDE4D$ and $TGF\beta 2$ (Supplementary Table S3, available at Carcinogenesis Online). It was predicted that variants in $ITPR1$ and $BNC2$ interact to confer increased risk ($Z$ score = 4.90, $P_{LR} = 7.32 \times 10^{-7}$), and variation in $BNC2$ has previously been associated with ovarian cancer in a GWAS (35). Interestingly, these variants in $BNC2$ are not in LD based on HapMap Caucasians data ($r^2 = 0.01$; the interacting variant (rs717267) is at <2 kb from the 3′-exon of $BNC2$, whereas the marginal effect was detected in the 5′-region (35).

Next, an analysis of GO Biological Processes term annotations was performed to define the functional profile of the predicted $G \times G$. Of the 205 gene pairs depicted above, 173 contained annotations for both members. Subsequently, using 1000 randomly selected equivalent gene pair sets, a network of significant (false discovery rate < 1%) term interactions was obtained (Figure 3). In this network, node size was proportional to the number of genes annotated with the corresponding term. Thus, the most frequent terms in the predicted $G \times G$ corresponded to interacting genes involved in metabolic or biosynthetic processes (Figure 3).

**Biological insight from the predicted $G \times G$**

$G \times G$ have the potential to uncover functional relationships within or between biological processes and/or signaling pathways (36). In addition to the GWAS-based candidates mentioned above, the 205 gene set contained functional candidates previously linked to breast cancer risk. Among these, and also included in the top 5% MI-ranked pairs from the second breast cancer dataset, variation in $SMAD3$ has been associated with breast cancer risk in $BRCA2$ mutation carriers (37). $SMAD3$—together with other $SMAD$ family members—is a critical signal transducer and transcriptional regulatory downstream of $TGF\beta R1$ (38). In our study, an interaction was predicted between $SMAD3$ rs2289263, which is not in LD with the risk variant ($r^2 < 0.2$), and LPP rs4689690 (Supplementary Table S3, available at Carcinogenesis Online). LPP is a nuclear envelope protein involved in cell adhesion and motility, and transcriptional regulation (39,40).

Therefore, to assess the prediction for breast carcinogenesis, cellular alterations upon depletion of LPP and/or modeling of $TGF\beta 1$ signaling were assessed using the non-tumorigenic MCF10A mammary epithelial cell line. Notably, simultaneous depletion of LPP and incubation with $TGF\beta 1$ (without the presence of other stimuli) increased cellular proliferation relative to the corresponding single perturbations (48 h timepoint, two-tailed $t$-test $P$ values < 0.01; Figure 4A). In the control assays, $TGF\beta 1$ produced an antiproliferative effect as shown by a diminished proliferation rate (Figure 4A). Regarding the potential invasiveness, depletion of LPP impaired the formation of cellular spheroids (Figure 4B). Consistent with this observation, depletion of LPP increased cell migratory capacity in a wound-healing assay (Figure 4C). Moreover, the expression of the integrin receptors $\beta 1$ and $\alpha 5$ was modulated and, in particular, $\alpha 5$ increased significantly upon simultaneous depletion of LPP and incubation with $TGF\beta 1$ (Figure 4D). Therefore, while depletion of LPP alone may provide a protumorigenic phenotype by increasing migration and impairing

---

**Fig. 3.** Network of GO biological process terms linked to breast cancer $G \times G$. The nodes represent GO terms (identifiers are shown) and an edge links two terms if the term–term interaction is overrepresented (false discovery rate < 1%) in the predicted $G \times G$ set.
differentiation, simultaneous activation of TGFβ1 signaling substantially enhances cellular proliferation, which provides a mechanistic hypothesis for the predicted G × G.

Using complex gene coexpression evidence to predict G × G

The results above suggest that evidence based on complex- and context-defined gene coexpression patterns can be used to predict G × G associated with cancer risk. Thus, in a reverse strategy, we first analyzed gene expression profiles in pancreatic tumors (29) and then integrated the results with data from a pancreatic cancer GWAS (20,21). With the gene pairs ranked according to their MIs, 14 bins were defined, each of which contained 20 pairs (from the highest MI value to the 5000th value). Next, G × G using all unlinked ($r^2 < 0.2$) SNPs in a pair were computed and the number of significant associations was evaluated at two thresholds: $P_{LR} < 0.05$ and $P_{LR} < 0.01$. This analysis revealed more G × G than expected by chance: an average of 5.99% ($P_{WRT} = 0.003$ for the null hypothesis of ≤5% across the 14 bins) and an average of 1.41% ($P_{WRT} = 0.028$ for the null hypothesis of ≤1% across the 14 bins) of the SNP pairs showed $P_{LR} < 0.05$ and $P_{LR} < 0.01$, respectively (Figure 5).

Two control analyses were carried out to assess the identification of excess of G × G nominally significant for pancreatic cancer risk. An analogous bin analysis was carried out, but in this case the lowest 5000 MIs were used (i.e. non-significant gene coexpression). The results of this analysis did not detect significant G × G over the thresholds: $P_{WRT} = 0.35$ and 0.83 for the 5 and 1% thresholds, respectively. In addition, the REs between the 14 top and bottom bins were 25 and 59% for the $P_{LR} < 0.05$ and $P_{LR} < 0.01$, respectively, which appeared to be consistent with the enrichments shown above for breast cancer. Conversely, no enrichment was identified when using the 14 top bins but basing the ranking exclusively on PCCs: $P_{WRT} = 0.64$ and 0.39 for the 5 and 1% thresholds, respectively.

The significant G × G for pancreatic cancer risk in the top MI-based ranked bin included the following gene pairs: AQP8-FFG, LOC402251-LOC442270, DLX5-MAFP5 and ABCC8-PCP4 (Supplementary Table S4, available at Carcinogenesis Online). Two of these genes, DLX5 and PCP4, have been functionally linked to
axon guidance pathway contributes to pancreatic carcinogenesis (29). Next, GO term enrichment analyses indicated (false discovery rate < 1%) frequent involvement of metabolic and biosynthetic processes, but also indicated mechanistic differences relative to breast cancer $G \times G$ (i.e. the involvement of genes in developmental processes; Figure 6).

Discussion

The identification of human $G \times G$ has the potential to add fundamental knowledge to our understanding of the genetic basis, molecular mechanisms and biological processes/signaling pathways involved in carcinogenesis (5,8,9,15). Although there are several well-established analytical strategies, the large (and continually increasing) number of genetic variants makes genome-wide $G \times G$ analyses highly time consuming, and it remains difficult to interpret the results from a biological perspective. This study introduces an integrative genomics strategy that can potentially support the identification of statistically significant $G \times G$ associated with cancer risk. In designing this study, it seemed reasonable to assume that there are $G \times G$ associated with cancer risk and, critically, that they can be identified by integrating different types of gene/protein relationships. The degree of overlap between genome-wide gene/protein relationships has been evaluated, and clearly established, in diverse studies in model organisms. Although human conditions should not be an exception, the lack of large-scale human $G \times G$ datasets has hampered similar integrative analyses. In addition, while experimental methodologies to systematically identify mammalian $G \times G$ have recently been developed (17,18), the results of our study suggest that context-specific studies must also be carried out. Thus, $G \times G$ associated with cancer risk may only be confidently identified when gene/protein relationships related to the specific cancer type/subtype are analyzed. In this regard, the lack of overlap with protein–protein interactions may be due to the fact that this type of evidence is typically not tissue- or cell type-specific. In addition, the human protein–protein interactions known to date do not represent the complete space of interactions occurring in cells. In fact, the gene expression analysis probably covers a larger fraction of all potential gene pairs.

The results of this study may lead to the genetic analysis of specific $G \times G$ in breast and pancreatic cancer. Although the enrichments shown may be considered relatively low (maximum of 36% for breast cancer and 59% for pancreatic cancer), the integration of additional gene/protein relationships could potentially improve the predictions. From the evaluation of the overlap for the 205 gene pairs between the two breast cancer expression datasets analyzed, it could be presumed that the conclusions of this study are limited by the characteristics of each dataset. In addition, the study may be limited by the relatively small sample size of the GWAS datasets analyzed and by the required assumption that a given SNP pair corresponds to a unique gene pair defined by the genomic location of the SNPs; however, it is frequently observed that the functional effects of low-penetrance mutations can implicate genes located dozens or hundreds of kilobases away (43). Integration of independent GWAS data could help to provide better $G \times G$ predictions. Moreover, the EPICBLASTER algorithm might not capture all possible forms of epistasis described in the literature (24).

The predicted $G \times G$ including candidate genes previously linked to cancer risk may help to further delineate the mechanisms of carcinogenesis. Furthermore, since some of the proposed $G \times G$ involve non-correlated variants relative to the marginal effect, they can potentially unveil mutations linked to differential effects. Following on from the predicted $G \times G$ between $LPP$ and $SMAD3$, the identification of a signaling interplay between $LPP$ and TGFβ1 in a non-tumorigenic mammary model provides a mechanistic hypothesis centered on altered epithelial cell proliferation and differentiation (38). $LPP$ has been found to be highly expressed in normal luminal mammary cells (44), which are typically ER-positive, and coexpressed in breast tumors with a regulator of mammary cell differentiation (44,45). In this scenario, there is evidence for an expression quantitative trait locus in rs2289263 for $SMAD3$ (46), which could provide a hypothesis for the interaction with $LPP$; there is no published evidence for an expression quantitative trait locus in rs4686980 (or for rs28615981 in LD) but these $LPP$ variants appear to map within a c-FOS binding region identified by chromatin immunoprecipitation in the ENCODE project (47). $SMAD3$ and c-FOS have been shown to cooperate in promoting TGFβ signaling (48) and, therefore, this cooperation might regulate $LPP$ function/levels. Importantly, a recent study has identified LPP as a key regulator of TGFβ-induced migration and invasion in HER2-overexpressing breast cancer (49). Our study expands on this observation by proposing that perturbation of LPP–TGFβ signaling promotes the initial stage of breast carcinogenesis. At the level of the biological processes overrepresented in the predicted $G \times G$ sets for breast and pancreatic cancer risks, the common identification of metabolic and biosynthetic processes might be explained by their role in buffering phenotypic variability (36,50). Other identified processes, such as defense response for breast cancer risk and cell development for pancreatic cancer risk, might be related to tissue specificity and could be used to integrate additional gene/protein relationships for prioritizing $G \times G$. Together, our study proposes a method that may help to further decipher the genetic basis of cancer risk.

Fig. 6. GO Biological Process terms linked to predicted pancreatic cancer $G \times G$. The nodes represent GO terms (identifiers are shown) and an edge links two terms if the term–term interaction is overrepresented (false discovery rate < 1%) in the predicted $G \times G$ set. This test set corresponded to the significant $G \times G$ represented in Figure 5 (top 14 gene pair bins, $p_{14} < 0.05$).
Conclusions
Here, based on the premise that genes/proteins act coordinately across biological levels, we undertook an integrative study in order to predict G × G associated with cancer risk. Our study was centered on breast and pancreatic cancer and the results show that G × G associated with risk may be partially supported on the basis of complex gene coexpression in the specific cancer type. The requirement of complex (i.e., non-linear) coexpression in a defined cancer setting is consistent with the intricate nature of epistasis and the molecular specificities of carcinogenesis. The predicted G × G provide novel hypotheses for the functional interplay between biological processes in carcinogenesis. The knowledge generated by this study may stimulate new research toward a better understanding of the genetic basis of cancer risk.

Supplementary material
Supplementary Tables S1–S4 can be found at http://carcin.oxfordjournals.org/

Funding
Fundación Eugenio Rodríguez Pascual (2012-13); Fundación Ramón Areces (XV Enfermedades Raras); Fundación Roses Contra el Cancer (2012); Generalitat de Catalunya (2009-SGR283); Red Temática de Investigación Colaborativa en Cáncer (12/0036/0008); Spanish Ministry of Health, Instituto Salud Carlos III, Fondo de Investigación Sanitaria (12/01528).

Acknowledgements
We wish to thank all study participants for their valuable contribution and the National Cancer Institute NIH initiatives for making available the genome-wide association studies data.

Conflict of Interest Statement: None declared.

References
2. Couch,F.J.; et al.; kConFab Investigators; SWE-BRCA; Ontario Cancer Genetics Network; HEBON; EMBRACE; GEMO Study Collaborators; BCFR; CIMBA. (2013) Genome-wide association study in BRCA1 mutation carriers identifies novel loci associated with breast and ovarian cancer risk. PLoS Genet., 9, e1003212.

Received July 8, 2013; revised November 7, 2013; accepted November 27, 2013