Introduction

Breast cancer is a complex disease with a well-established genetic component (1–4). Different genetic methodologies have been used in the last 20 years to identify multiple susceptibility loci that vary in population frequency, relative risk, and potential functional role. For example, familial linkage and positional cloning studies found that rare mutations in the genes BRCA1, BRCA2, TP53, and PTEN are associated with a 10-fold to 20-fold increase in breast cancer risk (5–8). Variations in these genes are thought to contribute to breast cancer susceptibility through different cellular mechanisms. Whereas BRCA1 and BRCA2 belong to the DNA repair mechanism in cells (9, 10), TP53 and PTEN are tumor suppressor genes that participate in processes related to cell cycle control and cell proliferation (11, 12). Further interrogation of candidate genes associated with these cellular processes has led to the discovery of additional rare genetic variants conferring moderate relative risks (2–3-fold) of breast cancer (13–16).

Recently, genome-wide association studies (GWAS) have become a key paradigm in genetic studies of complex diseases. These studies are particularly powerful in identifying common-low penetrance risk alleles. Indeed, several novel markers with low (<2) relative risk for breast cancer have been detected by this approach (17–21). However, these reported loci are only those that reached a stringent statistical “genome-wide” significance criterion, whereas less attention has been given to the other hundreds of thousands of genotyped markers. Using new methods to study existing GWAS data may provide additional biological insights and highlight new candidate loci. To this end, a pathway-based approach is particularly appealing. This method examines whether the cumulative contribution of genes with a common biological denominator is greater than expected by chance. This
This study has recently been applied to GWAS of several non-cancer complex diseases (22–25).

In this study, we applied pathway analysis to the breast cancer GWAS of the Cancer Genetic Markers of Susceptibility (CGEMS) project of the National Cancer Institute (NCI). This study has recently identified a significant association of the breast cancer complex diseases (22–25).

We identified a significant association of single nucleotide polymorphisms (SNPs) in the FGFR2 gene with breast cancer susceptibility (18). We used the modified gene-set enrichment analysis of Wang and colleagues (22) to identify an excess of genotype-phenotype association signals in pathways from different resources. Finally, we examined whether the excess of association signals in different pathways is driven by the same subset of genes comprising a common biological module. These analyses illuminated three pathways and one canonical cascade that are possibly involved in genetic susceptibility to breast cancer.

Materials and Methods

Pathway construction

We collected pathway data from three widely used resources: the BioCarta pathway database (26), the Kyoto Encyclopedia of Genes and Genomes (KEGG; 27), and the NCI Protein Interaction Database (PID; ref. 28). Genes belonging to these pathways were associated with SNPs included in the Illumina Sentrix HumanHap550 genotyping BeadChip that was used in the CGEMS GWAS. SNPs were mapped to genes if they were located within a genomic region encompassing 20 kb 5' upstream and 10 kb 3' downstream of the gene coding region (National Center for Biotechnology Information human genome build 36.3). Because FGFR2 is highly associated with breast cancer in CGEMS (18), we excluded all SNPs mapped to this gene from our analysis. Finally, we restricted our analysis to pathways with 10 to 100 genes to alleviate the multiple testing problem by avoiding testing too narrowly or too broadly defined functional categories. A complete list of the studied pathways is available in Supplementary Table S1.

Gene-set enrichment analysis

The CGEMS breast cancer GWAS includes 1,145 postmenopausal women of European ancestry with invasive breast cancer and 1,142 controls. Associations between SNPs and breast cancer were determined using the Cochran-Armitage test for trend (29). Each gene and breast cancer were determined using the Cochran-Armitage test for trend (29). Each gene was assigned the number of genes in our data set) was assigned the total number of genes common to both pathways (A and B) and calculated it as follows:

$$ES_S = max_{S \subseteq \mathcal{X}} \left\{ \sum_{G \in \mathcal{S} \cap \mathcal{S}_j} \frac{r_{G,S_j}}{W_{S}} - \sum_{G \in \mathcal{S} \cap \mathcal{S}_j} \frac{1}{N - N_{\mathcal{S}_j}} \right\}$$

where $$W_S = \sum_{G \in \mathcal{S} \cap \mathcal{S}_j} |r_{G,S_j}|$$ and $$N_{\mathcal{S}_j}$$ is the number of genes in a pathway. This statistic reflects the relative overrepresentation of genotype-phenotype association signals in a particular set of genes (30). We used a permutation procedure (10,000 permutations, permuting the case-control status) to assess the significance of the pathway-based ES. This procedure generates a null distribution for the ES of each pathway based on its real genotype data and therefore accounts for the variable SNP count and the intrinsic correlation between SNPs within the different genes. Finally, a normalized enrichment score (NES) was calculated for each pathway in the observed and permuted data to allow a direct comparison of pathways of different sizes as suggested by Wang and colleagues (22). The calculation of the NES for each pathway, $$NES^*_p$$, relied on the ES, and the mean and SD of $$ES^*_p$$ in all permutations for a given pathway ($$S$$).

$$NES_S = \frac{ES_S - mean(ES^*_S)}{SD(ES^*_S)}$$

These data were then used to calculate a false discovery rate (FDR; ref. 31) that controls for the proportion of expected false-positive findings to be under a certain threshold. For a pathway with a NES value of $$NES^*_p$$, the FDR is calculated as:

$$FDR = \frac{\% \text{ of all } (S, \text{per}) \text{ with } NES^*_p \geq NES^*_p \% \text{ of all observed } S \text{ with } NES^*_p \geq NES^*_p}{\% \text{ of all observed } S \text{ with } NES^*_p \geq NES^*_p}$$

Pathways with FDR < 0.2 were considered noteworthy.

Overlap

To explore the relationships between pathways in our database, we defined the overlap between the set of genes in pathway A and the set of genes in pathway B as the percentage of genes common to both pathways (A and B) and calculated it as follows:

$$\text{Overlap}(\%) = \frac{\#(A \cap B)}{\#(A \cup B)} \times 100\%$$

Next, we used a hierarchical clustering algorithm to assemble pathways with similar gene content. This algorithm iterates over all pathways, finds the most similar pair of pathways, and groups them together. This process was repeated until all pathways are grouped in one cluster in a hierarchical manner. To determine clusters of significantly related pathways, we used the same guidelines of the confidence interval algorithm originally developed for linkage disequilibrium block determination in the genome described by Gabriel and colleagues (32). In short, clusters were defined as containing significantly related pathways if they included at least 5 pathways, and if ≥95% of the pairwise overlap scores were greater than 9.6% (the 95th percentile of all pairwise overlap scores in our database).

Finally, we applied the adaptive rank-truncated product (ARTP) method (33) using the 10,000 permuted data sets to assess whether pathways within the defined clusters had higher ES than expected by chance. This method assesses the null hypothesis that a cluster of pathways does not contain excess of pathways with high ES. Denote the ordered $$P$$ values of the ES of pathways belonging to a particular cluster as ($$p_1 \leq \ldots \leq p_k$$), with $$p_1$$ being the smallest $$P$$ value. To test
for the global null hypothesis, the ARTP procedure calculate the RTP statistics \( W_k = \prod_{i=1}^{k} p_i \) over all possible truncation points \( (k) \), and assess the significance of the best RTP using the permutation data.

**Results**

Overall, 421 pathways and 3,962 genes (164/996, 155/3,088, and 102/1,317 pathways/genes from BioCarta, KEGG, and PID, respectively) were included in our study. These were represented by 69,525 of the 528,173 SNPs (13.2%) that were originally genotyped in this GWAS. A significant variation was seen between pathway resources. More than 90% of the pathways from BioCarta in our database had <30 genes, whereas pathways from KEGG and PID were generally larger and more variable in size (mean gene count, 44 and 36; SD, 24 and 16, respectively). This variation between databases was evident even in pathways nominally representing the same biological processes. For example, the mammalian target of rapamycin signaling pathway included 19, 24, and 49 genes in BioCarta, PID, and KEGG, respectively. Interestingly, pathways from BioCarta tended to be ranked higher (higher ES) than KEGG pathways but not than PID pathways (Kruskal-Wallis test, \( P = 0.013 \) and \( P = 0.151 \), respectively). No rank differences were seen between pathways from PID and KEGG.

Of the 421 examined pathways, 21 were significantly enriched with association signals at the \( P < 0.05 \) level (Table 1). Of these, “syndecan-1–mediated signaling events” from the PID database (Fig. 1A; \( P_{ES} = 0.00053 \), “signaling of hepatocyte growth factor receptor” from BioCarta (Fig. 1B; \( P_{ES} = 0.00063 \), and “growth hormone signaling” from BioCarta (Fig. 1C; \( P_{ES} = 0.00084 \)) had distinctly smaller \( P \) values than the rest of the pathways, with a noteworthy FDR value of 0.118. For brevity, we will further refer to these three pathways as “syndecan-1 signaling,” “c-Met signaling,” and “growth hormone signaling,” respectively.

Further examination of the gene content of these three pathways revealed some overlap. For example, the gene mitogen-activated protein kinase 3 (MAPK3) is involved in all three pathways but does not seem to contribute to their significant ES due to its nonsignificant association with breast cancer risk in this study (\( P = 0.35 \)). In contrast, the gene hepatocyte growth factor (HGF) was the strongest associated gene in the syndecan-1 signaling and c-Met signaling pathways (\( P_{\text{trend}} = 0.0007 \), for rs975360) and therefore a major contributor for their high ES. Another two genes [MAPK1 and met proto-oncogene (hepatocyte growth factor receptor; MET)] were shared by these two pathways. Similarly, six other genes [growth factor receptor-bound protein 2 (GRB2); v-Ha-ras Harvey rat sarcoma viral oncogene homologue (HRAS1); mitogen-activated protein kinase kinase 1 (MAP2K1); phosphoinositide-3-kinase, catalytic,

<table>
<thead>
<tr>
<th>Rank</th>
<th>Source</th>
<th>Pathway name</th>
<th>No. of genes</th>
<th>( P_{ES}^\dagger )</th>
<th>FDR†</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>PID</td>
<td>Syndecan-1-mediated signaling events</td>
<td>13</td>
<td>0.00053</td>
<td>0.118</td>
</tr>
<tr>
<td>2</td>
<td>BioCarta</td>
<td>Signaling of hepatocyte growth factor receptor</td>
<td>33</td>
<td>0.00063</td>
<td>0.118</td>
</tr>
<tr>
<td>3</td>
<td>BioCarta</td>
<td>Growth hormone signaling</td>
<td>22</td>
<td>0.00084</td>
<td>0.118</td>
</tr>
<tr>
<td>4</td>
<td>BioCarta</td>
<td>Insulin signaling</td>
<td>19</td>
<td>0.00642</td>
<td>0.442</td>
</tr>
<tr>
<td>5</td>
<td>KEGG</td>
<td>Basal cell carcinoma</td>
<td>53</td>
<td>0.00674</td>
<td>0.487</td>
</tr>
<tr>
<td>6</td>
<td>BioCarta</td>
<td>Eicosanoid metabolism</td>
<td>19</td>
<td>0.00695</td>
<td>0.515</td>
</tr>
<tr>
<td>7</td>
<td>PID</td>
<td>Signaling events mediated by stem cell factor receptor (c-Kit)</td>
<td>50</td>
<td>0.00779</td>
<td>0.515</td>
</tr>
<tr>
<td>8</td>
<td>KEGG</td>
<td>Tryptophan metabolism</td>
<td>57</td>
<td>0.01516</td>
<td>0.664</td>
</tr>
<tr>
<td>9</td>
<td>BioCarta</td>
<td>ADP-ribosylation factor</td>
<td>16</td>
<td>0.01800</td>
<td>0.842</td>
</tr>
<tr>
<td>10</td>
<td>PID</td>
<td>Fibroblast growth factor signaling pathway</td>
<td>52</td>
<td>0.02189</td>
<td>0.871</td>
</tr>
<tr>
<td>11</td>
<td>KEGG</td>
<td>ErbB signaling</td>
<td>86</td>
<td>0.02379</td>
<td>0.871</td>
</tr>
<tr>
<td>12</td>
<td>BioCarta</td>
<td>ALK in cardiac myocytes</td>
<td>33</td>
<td>0.02747</td>
<td>0.871</td>
</tr>
<tr>
<td>13</td>
<td>BioCarta</td>
<td>Interleukin-2 receptor β-chain in T-cell activation</td>
<td>34</td>
<td>0.03389</td>
<td>0.881</td>
</tr>
<tr>
<td>14</td>
<td>BioCarta</td>
<td>Intrinsic prothrombin activation</td>
<td>16</td>
<td>0.03600</td>
<td>0.881</td>
</tr>
<tr>
<td>15</td>
<td>KEGG</td>
<td>Streptomyacin biosynthesis</td>
<td>10</td>
<td>0.03600</td>
<td>0.881</td>
</tr>
<tr>
<td>16</td>
<td>BioCarta</td>
<td>Angiotensin II–mediated activation of JNK pathway via Pyk2-dependent signaling</td>
<td>27</td>
<td>0.03663</td>
<td>0.881</td>
</tr>
<tr>
<td>17</td>
<td>BioCarta</td>
<td>Sprouty regulation of tyrosine kinase signals</td>
<td>15</td>
<td>0.04095</td>
<td>0.881</td>
</tr>
<tr>
<td>18</td>
<td>KEGG</td>
<td>Maturity onset diabetes of the young</td>
<td>24</td>
<td>0.04547</td>
<td>0.881</td>
</tr>
<tr>
<td>19</td>
<td>BioCarta</td>
<td>Regulation of hematopoiesis by cytokines</td>
<td>15</td>
<td>0.04568</td>
<td>0.881</td>
</tr>
<tr>
<td>20</td>
<td>KEGG</td>
<td>Phosphatidylinositol signaling system</td>
<td>75</td>
<td>0.04674</td>
<td>0.881</td>
</tr>
<tr>
<td>21</td>
<td>KEGG</td>
<td>Melanogenesis</td>
<td>98</td>
<td>0.04747</td>
<td>0.881</td>
</tr>
</tbody>
</table>

*\( P \) values are computed for the pathway ES using 10,000 permutations.
†FDR for the 421 pathways examined in this study.
β-polypeptide (PIK3CB); v-raf-1 murine leukemia viral oncogene homologue (RAF1); and son of sevenless homologue 1 (SOS1) were shared by the c-Met signaling and growth hormone signaling pathways.

In light of the considerable overlap between the three most significant pathways in our study, we hypothesized that common biological modules may underlie the enrichment signals in multiple biological pathways. To explore this hypothesis, we used hierarchical clustering to identify pathways with overlapping genes likely to constitute common biological process. Consequently, 16 clusters of highly related pathways were characterized (Fig. 2A). Of these, cluster #1 (Fig. 2B) was significantly more likely to include pathways with high ES than expected by chance (ARTP $P = 0.0045$, FDR = 0.07). Further examination of this cluster revealed that it contained two of three of the top three pathways (signaling of hepatocyte growth factor receptor and growth hormone signaling) and was dominated by the genes GRB2, SOS1, HRAS, RAF1, MAP2K1, and MAPK3, which are major components of the RAS/RAF/MAPK canonical signaling cascade (Fig. 1B and C).
To assess the power of our analysis to identify true susceptibility pathways, we constructed an artificial positive control pathway from 11 genes that were previously associated with breast cancer risk in different GWAS of postmenopausal women of European ancestry (see details in Supplemental Table S2). Applying the pathway analysis to this pseudo-pathway yielded a $P_{ES} = 0.0126$, which ranked it 8th out of all pathways in the study. It is noteworthy that most of the individual genes in the pseudo-pathway had unremarkable significance (Supplemental Table S2, last column) in the CGEMS study itself. Thus, the high rank of this artificial positive control pathway shows the ability of the gene-set enrichment analysis to detect pathways containing multiple genes that individually have weak association signals and may be missed by standard single-SNP analysis of GWAS data.

Discussion

Three pathways and one signaling cascade are highlighted in this study. The top ranked pathway in this study, syndecan-1 signaling, contains 13 genes involved in different cellular processes that are mediated by syndecan-1 (SDC1). This gene encodes a transmembrane heparan sulfate proteoglycan that mediates signal transduction cascades leading to cell proliferation, cell migration, and cell adhesion processes following interactions with extracellular matrix proteins. There are multiple lines of evidence for a potential role of syndecan-1 in breast cancer development. For example, altered syndecan-1 expression has been detected in several different tumor types and has been linked with unfavorable breast cancer prognosis (34). Additionally, expression of
syndecan-1 by stromal fibroblasts has been shown to promote breast carcinoma growth in vivo and stimulates tumor angiogenesis (35). In this GWAS, SDC1 was only moderately associated with breast cancer susceptibility ($P_{\text{trend}} = 0.019$ for rs7563245); however, it is a key mediator of different pathways illuminated in this study (e.g., c-Met signaling and fibroblast growth factor signaling; see discussion below). Therefore, it can modulate breast cancer susceptibility through different biological mechanisms.

The second highest ranked pathway in our analysis, c-Met signaling, consists of 33 genes participating in signal transduction mechanisms induced by the tyrosine kinase proto-oncogene c-Met (MET). Stimulation of the c-Met pathway can lead to several different cellular processes related to tumor growth and progression, such as proliferation, enhanced cell motility, invasion, and apoptosis (36). Moreover, both c-Met and its ligand, hepatocyte growth factor, have been shown to be dysregulated and correlated with poor prognosis in a number of human malignancies, including breast cancer (37). Consequently, this pathway has served as an important therapeutic target for human cancers, particularly through the development of a small molecule that inhibit the c-Met/hepatocyte growth factor–dependent signaling (37). Notably, coexpression of SDC1 and MET, the key players in the top two ranked pathways in our study, have been established as a marker signature associated with poor prognostic factors in ductal carcinoma in situ of the breast (38).

The third ranked pathway in our study, growth hormone signaling, contains 22 genes participating in cellular mechanisms induced by either growth hormone or insulin receptors. These two receptors as well as the insulin-like growth factor (IGF) receptor are all transmembrane tyrosine kinase receptors that induce cell growth and proliferation. Alteration in the activity of these receptors or their related pathways may lead to hyperplasia and eventually to the development of a tumor (39). Naturally, the growth hormone signaling pathway had considerable overlap with both the IGF-1 signaling and insulin signaling pathways from BioCarta (32% and 41%, respectively); however, only the latter had a significant ES in our study ($P_{\text{ES}} = 0.0064$). Examining the genes of these three closely related pathways revealed that what differentiates their respective ES is a SNP in the insulin receptor gene (INSR) with a small $P$ value ($P_{\text{trend}} = 0.0019$ for rs12460755) that is absent from the IGF-1 signaling pathway.

Interestingly, the fibroblast growth factor signaling pathway was ranked 10th in our study (Table 1; $P_{\text{ES}} = 0.0219$) despite the exclusion of the FGFR2 SNPs from our analyses (see Materials and Methods). Adding the FGFR2 SNPs to this pathway improved its enrichment signal ($P_{\text{ES}} = 0.0053$) and ranked it 4th in our study. This finding, in combination with the FGFR2 signal in CEAGESM (18) and elsewhere (17), suggests that variations in other genes involved in FGFR2 signaling may modulate breast cancer susceptibility.

An important extension to the gene-set enrichment analysis is the clustering analysis that highlighted the RAS/RAF/MAPK canonical signaling cascade as the common denominator of pathways associated with breast cancer risk in this study. This cascade plays an essential role in transmitting extracellular signals from growth factors to promote the growth, proliferation, differentiation, and survival of cells, and modification in its activity has been linked to multiple human malignancies (40). Notably, this cascade plays an important role in all three top pathways in this study. It is an integral component in the signaling of hepatocyte growth factor receptor and growth hormone signaling pathways, and a transducer for signals initiated by the syndecan-1 pathway.

An important limitation of the pathway-based approach for GWAS analysis is the incomplete annotation of the human genome. At present, the function of many human genes is unknown and therefore they cannot be assigned to known pathways. Moreover, susceptibility loci in intergenic regions are also not included in a study of this kind. As a result, when using this approach, only a small portion of the human genome variation can be studied and therefore it should only be used as a supplementary method to the standard single-SNP analysis of GWAS. Additionally, there is no gold standard or pathway definition, and different databases have different guidelines for their pathway construction and curation. Consequently, the gene content of pathways representing the same biological process may vary substantially between different databases, and this may have a major impact on the sensitivity and specificity of this approach. We aimed at minimizing this effect by selecting pathways from three commonly used and manually curate resources. Still, considerable differences were observed between similar pathways from different resources. For example, the c-Met signaling pathway from BioCarta contained 33 genes and was ranked second in our analysis, whereas a pathway with the same name and presumably the same cellular function in the PID database included 52 genes and was ranked only 69th. Although 30 of the 33 genes in the BioCarta pathway were also included in the PID pathway, the remaining 22 genes likely attenuated the signal in the PID pathway. These differences emphasize the importance of further synthesizing the results of such pathway-based approach. The clustering analysis we applied in this study is one way to do so, as it helps in finding the common genes underlying enrichment signals in multiple pathways and allows one to focus on a limited number of candidate genes. Other methods aiming to improve the characterization of pathways and their overlap might be useful.

A second limitation of our study is that it does not include validation of the results using a completely independent data set. The experiment we conducted using a positive control pathway with known breast cancer susceptibility genes provides us validation that our approach has high power to detect pathways containing multiple weakly associated susceptibility genes. Therefore, the top ranked pathways in our study should be prime targets for future analyses in independent data sets.

In conclusion, our results suggest that genetic alterations associated with the top three pathways and one canonical signaling cascade in our study may contribute to breast cancer susceptibility. Ultimately, additional studies would be needed to confirm and further explore the genetic variations underlying the association of these pathways with...
breast cancer. Moreover, this study highlights the potential insight that could be gained by using a pathway-based approach as a complimentary method to the primary single-SNP analysis of GWAS. Particularly, the organization of multiple association signals according to underlying biological processes may improve our understanding of the cellular mechanisms underlying this too common malignancy.

Disclosure of Potential Conflicts of Interest

No potential conflicts of interest were disclosed.

References


Acknowledgments

This study was partly supported by the Gene Environment Initiative program of the National Institute of Health.

Grant Support

Intramural Research Program of the NIH/NCI and the Gene Environment Initiative program of the NII.

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

Received 12/10/2009; revised 02/19/2010; accepted 03/13/2010; published OnlineFirst 05/11/2010.
Pathway Analysis of Breast Cancer Genome-Wide Association Study Highlights Three Pathways and One Canonical Signaling Cascade

Idan Menashe, Dennis Maeder, Montserrat Garcia-Closas, et al.

Cancer Res  Published OnlineFirst May 11, 2010.

Updated version  Access the most recent version of this article at: doi:10.1158/0008-5472.CAN-09-4502

Supplementary Material  Access the most recent supplemental material at: http://cancerres.aacrjournals.org/content/suppl/2010/05/10/0008-5472.CAN-09-4502.DC1

E-mail alerts  Sign up to receive free email-alerts related to this article or journal.

Reprints and Subscriptions  To order reprints of this article or to subscribe to the journal, contact the AACR Publications Department at pubs@aacr.org.

Permissions  To request permission to re-use all or part of this article, contact the AACR Publications Department at permissions@aacr.org.