Many common diseases, including most cancers (1), may have a component of inheritance that is due to low-penetrance, common genetic variants. However, identifying these variants has been a challenge. In this issue of the Journal, the Breast Cancer Association Consortium reports an impressive collaborative effort—the large-scale evaluation of 16 single nucleotide polymorphisms (SNPs) that have been previously hypothesized to modify susceptibility to breast cancer (2). Despite accumulated genotype information on 12,000–32,000 breast cancer case and control subjects for each SNP, results are largely null. No SNP shows an association with \( P < 0.0031 \), the threshold required after adjusting for 16 tests. Five variants with \( P \) values between .009 and .06 require further testing, but the odds ratios (ORs) are already hovering at around 0.9–1.1. For the other 11 SNPs, results are even more conclusively null. Extensive subgroup analyses largely reinforce these conclusions.

These data add to the tribulations of the “candidate gene” approach that has guided complex disease genetics for over a decade. In this approach, biologic pathways thought to be relevant to specific diseases were scrutinized for gene variants modulating disease risk. Postulated associations were then often extended to other phenotypes, with or without much rationale for the extension (3). The gene–disease association literature grew exponentially. In 2005 alone, 194 original research articles were published that probed gene–disease associations for breast cancer (4); I selected every 10th article (n = 19) for perusal. Fifteen of these articles claimed associations overall, in subgroups, or for specific outcomes. The parade of claimed associated genes in this tiny sample is already impressive: HER2, IL10, NCOA3, TGFBRI, TGFB1, ESR2, HFE, IGF-I, ESR1, AR, CHEK2, PAI-1, XRCC1, HSMH2, SULT1A1, and IFNG. If all these claimed associations are real, a 10% sample of the published genetic association research in a single year alone seemingly suffices to explain all that causes breast cancer: the total attributable fraction from this small sample of associations already reaches close to 100%.

Is this an apotheosis of data dredging? Even in my small sample of 19 articles, one comes across an association that is statistically significant only in the sub-sub-subgroup of postmenopausal women who have at least three pregnancies and also have no wild-type allele (5); a polymorphism with statistically significantly decreased risk for early-stage breast cancer but increased risk for advanced-stage disease (6); another increasing the risk, especially for grade 3 tumors (7); a marker with 13 variants, of which one shows a statistically significant association versus all others combined, while hundreds of different groupings are conceivable (8); polymorphisms that have no statistically significant associations on their own but do in one of their many constructed haplotypes (9); joint effects of polymorphisms of different genes acting in obvious (10) or not-so-obvious (11) pathways; associations that are not even tested statistically (12); and so forth.

In addition to the 16 SNPs examined by the Breast Cancer Association Consortium, meta-analyses of published data are available for another 16 candidate gene variants in breast cancer (Table 1) (8, 13–19). When compared with the odds ratios in the first studies published on each of these 16 variants, the odds ratios in the meta-analyses typically converge to the null (Fig. 1), as shown also in genetic association studies in other diseases (20–22). For three SNPs, meta-analyses still reached \( P < 0.05 \) (i.e., the odds ratios were statistically significantly different from 1.0), but all three associations are implausible. For H-ras-I rare alleles, early studies used erroneous allele calling methods (23, 24); recent larger studies have found no association (19). The borderline association of GSTT1 is justly dismissed by the meta-analysis authors (17); the miniscule effect is driven by very small studies. Finally, in the case of IGF-I, a polymorphic marker was categorized arbitrarily in two categories (8). More recent studies claim effects in the opposite direction or with different categorizations (25, 26), and opposing effects might cancel out in an updated meta-analysis.

Overall, a sobering picture arises across the 32 candidate associations. They may well all be false (27); even if a couple of associations survive further testing, the picture will not change much. Clearly, we should be cautious about evidence on “significant” genetic associations based on several hundred or even several thousand subjects. Conclusive evidence may require sample sizes in the range of 20,000; for uncommon variants, even evidence from such large studies is tentative (28).

It should also be noted that not all large-scale evidence is equal. One might wish to have single studies with tens of thousands of cases and controls. However, such studies are currently lacking. Therefore, we need to join forces and pieces of evidence from many teams. To date, meta-analyses have generally used the published literature. However, publication and selective reporting bias may lead to spurious conclusions (29–31). A consortium approach, such as the one adopted by the Breast Cancer Association Consortium, can help avoid these biases (32–34). Investigators working in the same field harmonize their data; they procure all collected information, published or unpublished, on specific associations under study; and they analyze the information with common rules.

How often would a consortium get different results than literature-based meta-analyses? We need more data to answer this question (35, 36). For example, for the AURKA F31H polymorphism, two meta-analyses of published data found statistically significant associations without any between-study heterogeneity (37, 38). However, a subsequent updated meta-analysis showed...
an association of only borderline statistical significance, and there was extensive between-study heterogeneity ($P<.001$) (39). The Breast Cancer Association Consortium shows absolutely no overall effect ($P = .57$) and only modest heterogeneity ($P = .09$). A subgroup genetic effect cannot be excluded (e.g., limited to early–age onset cancer), but it is more likely that the early meta-analyses found statistically significant results simply because of various biases.

To further minimize errors, consortia of investigators should perform genotyping prospectively and with stringent quality control procedures (40). This process has already been used in some fields in cancer and beyond (34,41,42) and should be further encouraged in cancer genomics. Combining data should be anticipated in the study design phase. This collaborative approach becomes even more attractive with the recent advent of genome-wide association studies (43,44). With hypothesis-free testing of hundreds of thousands of genetic variants, small biases may yield thousands of highly statistically significant but spurious signals (45,46), burying the true ones. Hundreds of variants identified by genome-wide screening may be tested by each replicating team. If the extent of this multiplicity is not transparent, some associations with highly statistically significant results may still be spurious, despite large-scale evidence. Replicating one association is not the same when there are 1000 replication candidates as when there are 32 replication candidates. As multiplicity increases, sample sizes in the range of 50,000 may eventually be required to validate associations derived from multiple-stage screening with multiple testing in each stage.

Obtaining such large-scale evidence of high quality is not an easy task. Of course, some of the genetic effects may be large and large studies would not be needed. The first major success of genome-wide approaches—the identification of a genetic

Table 1. Published meta-analyses of gene–disease associations for susceptibility to breast cancer*

<table>
<thead>
<tr>
<th>Reference</th>
<th>Gene and variant</th>
<th>Case subjects</th>
<th>Control subjects</th>
<th>Contrast</th>
<th>Odds ratio (95% confidence interval)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(13)</td>
<td>MTHFR C677T</td>
<td>5476</td>
<td>7336</td>
<td>Recessive</td>
<td>1.06 (0.91 to 1.24)</td>
</tr>
<tr>
<td>(13)</td>
<td>MTHFR A1298C</td>
<td>3768</td>
<td>5276</td>
<td>Recessive</td>
<td>1.06 (0.84 to 1.35)</td>
</tr>
<tr>
<td>(14)</td>
<td>X RCC1 R144W</td>
<td>5752</td>
<td>6050</td>
<td>Dominant</td>
<td>1.00 (0.80 to 1.20)</td>
</tr>
<tr>
<td>(15)</td>
<td>CYP1A1 T3801C</td>
<td>2484</td>
<td>2609</td>
<td>CC vs. TT</td>
<td>0.97 (0.52 to 1.80)</td>
</tr>
<tr>
<td>(15)</td>
<td>CYP1A1 Thr491Asp</td>
<td>2245</td>
<td>1139</td>
<td>AspAsp vs. ThrThr</td>
<td>0.95 (0.20 to 4.49)</td>
</tr>
<tr>
<td>(16)</td>
<td>CYP1B1 Ala119Ser</td>
<td>3969</td>
<td>3723</td>
<td>AlaSer vs. AlaAla</td>
<td>1.07 (0.91 to 1.25)</td>
</tr>
<tr>
<td>(16)</td>
<td>CYP1B1 Leu432Val</td>
<td>5712</td>
<td>5107</td>
<td>SerSer vs. AlaAla</td>
<td>0.98 (0.81 to 1.19)</td>
</tr>
<tr>
<td>(16)</td>
<td>CYP1B1 Asn453Ser</td>
<td>2165</td>
<td>2010</td>
<td>ValVal vs. LeuLeu</td>
<td>1.06 (0.90 to 1.26)</td>
</tr>
<tr>
<td>(16)</td>
<td>COMT Val/Met</td>
<td>8286</td>
<td>7344</td>
<td>ValMet vs. ValVal</td>
<td>0.96 (0.83 to 1.12)</td>
</tr>
<tr>
<td>(8)</td>
<td>IGF-I (CA)</td>
<td>1331</td>
<td>1478</td>
<td>(CA)19 carriers</td>
<td>1.22 (1.06 to 1.41)</td>
</tr>
<tr>
<td>(17)</td>
<td>GSTM1 null</td>
<td>5950</td>
<td>6601</td>
<td>Null carriers</td>
<td>1.05 (0.98 to 1.13)</td>
</tr>
<tr>
<td>(17)</td>
<td>GSTT1 null</td>
<td>4873</td>
<td>5245</td>
<td>Null carriers</td>
<td>1.11 (1.01 to 1.22)</td>
</tr>
<tr>
<td>(17)</td>
<td>GSTP1 Ile105Val</td>
<td>2136</td>
<td>2282</td>
<td>Recessive</td>
<td>1.04 (0.87 to 1.25)</td>
</tr>
<tr>
<td>(18)</td>
<td>CYP17 Mspl</td>
<td>4227</td>
<td>4730</td>
<td>A2A2 vs. A1A1</td>
<td>1.05 (0.87 to 1.21)</td>
</tr>
<tr>
<td>(19)</td>
<td>H-ras-1 rare alleles</td>
<td>694</td>
<td>937</td>
<td>Rare allele carriers</td>
<td>2.70 (2.10 to 3.40)</td>
</tr>
</tbody>
</table>

*Meta-analyses based on published data for common genetic variants that were not addressed by the Breast Cancer Association Consortium article (2). Whenever more than one meta-analysis is available for the same association, only the most comprehensive was used. Only associations with gene variants that have a frequency exceeding 1% among healthy controls were considered. The list has been compiled based on searches of PubMed and the HuGENet Published Literature Database (4). For associations in which the meta-analysis addressed separately the contrasts of homozygotes versus wild-type homozygotes and heterozygotes versus wild-type homozygotes, both contrasts are presented; for the other meta-analyses, one major contrast is presented.
determinant of age-related macular degeneration—pertained to an odds ratio of 7 (47); fewer than 100 subjects sufficed to detect the association. However, this is the exception (48–51). For breast cancer in particular, genome-wide association studies are ongoing; common genetic variants with large odds ratios have not yet been revealed. Probably, we will soon have a new wave of claimed associations, with much lower $P$ values but also with more complex multiple testing backgrounds.

The near future will be fascinating. Discoveries will be made, and for some diseases they may be a torrent (52). For other diseases, the yield may be lower. Uncommon mutations will be difficult to unearth through screening of common variants. For breast cancer, the greatest successes to date pertain to uncommon mutations, such as $BRC A1$, $BRC A2$, and $CHEK2$ (53, 54). Moreover, some common diseases may end up having largely “private” epidemiology (55), whereby most individuals get the disease from unique or uncommon constellations of genetic and acquired risk factors. Making progress on the genetics side would require progress on environmental components. “Private” epidemiology may be a common theme for many common cancers.

The problem of replicating associations is not limited to genetic epidemiology. Indeed, genetics has sensitized us to many problems that affect epidemiologic investigation at large. Genomics have advanced at speeds unimaginable for traditional epidemiology, for which measurement methods remain relatively primitive. However, even in genomics, advances in measurement platforms should be complemented by transparently designed and reported collaborative large-scale evidence with high-quality data. If common genetic variants modulate susceptibility to common diseases such as breast cancer, we now have excellent tools to dissect them. We should just do so correctly.

**References**


