The genetics of breast cancer: risk factors for disease

Andrew Collins
Ioannis Politopoulos

Genetic Epidemiology and Bioinformatics Research Group, Human Genetics Research Division, Southampton General Hospital, School of Medicine, University of Southampton, Southampton, UK

Abstract: The genetic factors known to be involved in breast cancer risk comprise about 30 genes. These include the high-penetrance early-onset breast cancer genes, BRCA1 and BRCA2, a number of rare cancer syndrome genes, and rare genes with more moderate penetrance. A larger group of common variants has more recently been identified through genome-wide association studies. Quite a number of these common variants are mapped to genomic regions without being firmly associated with specific genes. It is thought that most of these variants have gene regulatory functions, but their precise roles in disease susceptibility are not well understood. Common variants account for only a small percentage of the risk of disease because they have low penetrance. Collectively, the breast cancer genes identified to date contribute only ~30% of the familial risk. Therefore, there is much interest in accounting for the missing heritability, and possible sources include loss of information through ignoring phenotype heterogeneity (disease subtypes have genetic differences), gene–gene and gene–environment interaction, and rarer forms of variation. Identification of these rarer variations in coding regions is now feasible and cost effective through exome sequencing, which has already identified high-penetrance variants for some rare diseases. Targeting more ‘extreme’ breast cancer phenotypes, particularly cases with early-onset disease, a strong family history (not accounted for by BRCA mutations), and with specific tumor subtypes, provides a route to progress using next-generation sequencing methods.

Keywords: breast cancer, common and rare genetic variation, missing heritability, bioinformatics, exome sequencing

Introduction

Family history of breast cancer is known to be one of the strongest risk factors for this disease. For example, meta-analysis of familial breast cancer studies gives lifetime risk ratios of 1.80 in families with one affected first-degree relative, 2.93 in families with two affected relatives, and 3.90 in families with three affected relatives.1 Risk ratios are highest for cases at younger ages and, for a particular individual, are greater the younger their relative is diagnosed. The familial pattern of the disease provides clear evidence for the important role of genetic variation in determining risk. The identification of genetic factors involved in predisposition to breast cancer has been a topic of intensive study for more than 20 years. An important early breakthrough in the genetic dissection of the disease was linkage mapping, using breast cancer family data, of the BRCA1 and BRCA2 genes. Rare mutations in these genes confer high relative risks to carriers of 10- to 20-fold, corresponding to a 30%–60% risk by the age of 60 years, compared with 3% for the general population.4 These mutations account for ~16%–20%
of the familial risk of breast cancer in the general population.\textsuperscript{5,5}

In addition, there are a number of rare to very rare high-penetrance gene variants that underlie cancer syndromes and a few rare genes that have more moderate penetrance. Collectively, the rare genes found to date account for <25% of the familial risk. Recent studies have focused on the role of common genetic variation, through analysis of large samples of cases and controls tested for association at many thousands of single nucleotide polymorphism (SNP) markers. These studies have identified a number of common breast cancer genes and revealed new insights into the natural history of the disease. However, all these genes are low-penetrance variants that account for only a few percent of the familial risk. Because the bulk of the familial risk is unexplained by the genes identified thus far, research is focusing on identifying sources of the ‘missing’ heritability. This review considers what is known about the genetic basis of breast cancer and evaluates the clinical utility of the evidence, while emphasizing ongoing strategies to identify more of the genetic variation. New technologies, such as next-generation sequencing, and the development of novel bioinformatic approaches to analysis are at the forefront of this effort.

**Mendelian high-penetrance genes**

About 100 genes for genetic diseases showing Mendelian patterns of inheritance in families are known.\textsuperscript{6} These are invariably rare genes and associated with high relative risks. Most of the genes have been identified through linkage analysis of carefully selected families, followed by positional cloning. Within this category are the breast cancer \textit{BRCA1} and \textit{BRCA2} genes, which contain over 1000 mutations. Genetic screening for the spectrum of important mutations in these genes in high-risk families is well established. The \textit{BRCA1} ‘breast cancer 1 early-onset’ gene is involved in susceptibility to breast and ovarian cancer at a young age, and tumors can arise through somatic or germline mutations. Impaired or lost \textit{BRCA1} function underlies substantial genome instability including increases in the number of mutations, DNA breakage and chromatid exchanges, increased sensitivity to DNA damage, and defects in cell-cycle checkpoint functions. The role of \textit{BRCA1} in the DNA damage response is that of ‘caretaker’ or ‘master regulator’ in the genome.\textsuperscript{7,7}

Jensen et al\textsuperscript{10} isolated the large protein encoded by the \textit{BRCA2} gene and showed it to be a key mediator of homologous recombination. It is a crucial element in the DNA repair process which, if impaired through mutation, can lead to chromosome instability and cancer. It is known to mediate recombinational DNA repair by promoting assembly of RAD51 onto single-stranded DNA. This has a key role in catalyzing the invasion and exchange of homologous DNA sequences. Mutations in the \textit{BRCA2} gene may disrupt this mechanism and impair repair of DNA breaks, using homologous sequences from an intact homolog or sister chromatid, leading to errors in the repair process and chromosome instability.

\textit{BRCA1} and \textit{BRCA2} are likely to be the only major high-penetrance genes underlying breast cancer. Germline mutations in the \textit{TP53} gene cause Li–Fraumeni syndrome, a phenotype which includes early-onset breast cancer,\textsuperscript{11} but these mutations are far rarer. Both \textit{BRCA1} and \textit{BRCA2} genes were identified using linkage mapping in families, a method that has been successful in identifying many Mendelian disease genes. However, this strategy has contributed little to the study of more common or ‘complex’ forms of disease, mediated by genetic variants with reduced penetrance which may interact with environmental and other genetic factors. The complexity of this pattern of inheritance greatly reduces the power to detect genes through family-based studies.

**Rare cancer syndromes and rare moderate-penetrance genes**

There are a number of syndromes that include breast cancer as a component of the disease phenotype. Rare to uncommon mutations in the \textit{PTEN}\textsuperscript{12} and \textit{STK11}\textsuperscript{13} genes cause Cowden and Peutz–Jeghers syndromes, respectively, and both are associated with considerably increased breast cancer risk.\textsuperscript{14} The E-cadherin gene (\textit{CDH1}) encodes a cellular adhesion protein and is a powerful tumor suppressor of breast cancer.\textsuperscript{15} It is particularly implicated in invasive lobular breast carcinomas. \textit{RAD51C} is another gene involved in the recombinational repair of double-stranded DNA breaks. Rare germline mutations have been shown to confer increased risks of breast and ovarian cancer.\textsuperscript{16} Segregation in families follows Mendelian patterns, and the disease phenotype resembles that of \textit{BRCA1} and \textit{BRCA2} mutation carriers.

There are also a number of gene mutations associated with more moderate risks of breast cancer, which show marked departures from Mendelian patterns of inheritance. As a result, segregation of disease with the mutation may be unhelpful to confirm relationship with disease. Genes in this category include germline mutations in the ataxia-telangiectasia (\textit{ATM}) gene, which are associated with increased risk (\textapprox{}2.2-fold) of breast cancer in carriers of heterozygous mutations, with apparently higher risks below the age of 50 years.\textsuperscript{17} Other rare moderate-penetrance genes include heterozygous mutations in \textit{BRIP1} (encoding a \textit{BRCA1}-interacting protein) that
confers elevated risks of breast cancer and Fanconi anemia subtype FA-J for bi-allelic mutations. The partner and localizer of BRCA2 (PALB2) gene interacts with BRCA2, and mono-allelic mutations are involved in familial breast cancer, conferring a 2.3-fold risk. Mutations in BRCA2 are also known to underlie Fanconi anemia (subtype FA-D1), and bi-allelic mutations of PALB2 underlie the very similar Fanconi anemia subtype FA-N.18 Rare variants in the cell cycle checkpoint kinase 2 (CHEK2) gene are known to underlie an approximately twofold increase in risk of breast cancer. Products of this gene are involved in DNA damage repair, and mutations are found in 1%–2% of unselected women with breast cancer.19

**Common low-penetration breast cancer genes**

Genome-wide association studies (GWAS) use panels of up to a million or more SNPs to identify common gene variants in large case and control samples. GWAS have identified more than 100 such low-penetrance loci involved in cancer, including at least 17 related to breast cancer (Table 1). These variants have allele frequencies in the range 0.05–0.5, but they confer only small increases in disease risk.4 Because of the greatly reduced penetrance and strongly non-Mendelian patterns of inheritance, there is often considerable uncertainty about the exact underlying genetic mutation. Not only are the most strongly associated SNPs unlikely to be the causal sites (these are ‘tags’ selected to represent variation at many polymorphic sites that are not tested directly) but there also may be uncertainty about the gene involved. It has also been suggested that multiple rare variants create ‘synthetic association’ signals in a GWAS if they occur more often in association with a common tag SNP. This implies that causal variants could be many megabases away from variants detected in GWAS,20 although this scenario appears to be rare.21 Perhaps, one of the unexpected findings from these studies is a greater-than-anticipated role for noncoding variants in common diseases.22 From the analysis of population sequences,23 <30% of common variants associated with disease are annotated as, or in linkage disequilibrium with, nonsynonymous (coding) variation. This supports the view that many of the common disease variants have gene regulatory roles.

**Table 1** Known breast cancer susceptibility genes and regions

<table>
<thead>
<tr>
<th>Known gene/region</th>
<th>Location</th>
<th>Mapped by</th>
<th>Allele frequency</th>
<th>Known/possible function</th>
</tr>
</thead>
<tbody>
<tr>
<td>BRCA1</td>
<td>17q21</td>
<td>Linkage</td>
<td>Rare</td>
<td>DNA repair/genome stability</td>
</tr>
<tr>
<td>BRCA2</td>
<td>13q13.1</td>
<td>Linkage</td>
<td>Rare</td>
<td>Recombinational repair</td>
</tr>
<tr>
<td>TP53</td>
<td>17p13.1</td>
<td>Linkage</td>
<td>Rare</td>
<td>Li–Fraumeni syndrome, apoptosis</td>
</tr>
<tr>
<td>ATM</td>
<td>11q22.3</td>
<td>Candidate resequencing</td>
<td>Rare</td>
<td>DNA repair</td>
</tr>
<tr>
<td>BRI1</td>
<td>17q23.2</td>
<td>Candidate resequencing</td>
<td>Rare</td>
<td>DNA repair, associated with BRCA1</td>
</tr>
<tr>
<td>CHEK2</td>
<td>22q12.1</td>
<td>Candidate resequencing</td>
<td>Rare</td>
<td>DNA repair/cell cycle</td>
</tr>
<tr>
<td>PALB2</td>
<td>16p12.2</td>
<td>Candidate resequencing</td>
<td>Rare</td>
<td>Associated with BRCA2</td>
</tr>
<tr>
<td>RAD51C</td>
<td>17q22</td>
<td>Candidate resequencing</td>
<td>Rare</td>
<td>Homologous recombination repair</td>
</tr>
<tr>
<td>PTEN</td>
<td>10q23.3</td>
<td>Linkage</td>
<td>Rare</td>
<td>Cowden disease, cell signaling</td>
</tr>
<tr>
<td>STK11 (LKB1)</td>
<td>19p13.3</td>
<td>Linkage</td>
<td>Rare</td>
<td>Peutz–Jeghers syndrome, cell cycle arrest</td>
</tr>
<tr>
<td>CDH1</td>
<td>16q22.1</td>
<td>Linkage</td>
<td>Rare</td>
<td>Intercellular adhesion: lobular BC</td>
</tr>
<tr>
<td>FGFR2</td>
<td>10q26</td>
<td>GWAS</td>
<td>Common</td>
<td>Fibroblast growth factor receptor</td>
</tr>
<tr>
<td>TOX3 (TNRC9)/RBL2</td>
<td>16q12</td>
<td>GWAS</td>
<td>Common</td>
<td>Chromatin structure/cell cycle</td>
</tr>
<tr>
<td>MAP3K1</td>
<td>5q11.2</td>
<td>GWAS</td>
<td>Common</td>
<td>Cellular response to growth factors</td>
</tr>
<tr>
<td>LSP1</td>
<td>11p15.5</td>
<td>GWAS</td>
<td>Common</td>
<td>Neutrophil motility</td>
</tr>
<tr>
<td>8q24</td>
<td>8q24</td>
<td>GWAS</td>
<td>Common</td>
<td>Intergenic, enhancer of MYC proto-oncogene</td>
</tr>
<tr>
<td>2q35</td>
<td>2q35</td>
<td>GWAS</td>
<td>Common</td>
<td>Apoptosis</td>
</tr>
<tr>
<td>CASP8</td>
<td>2q33</td>
<td>GWAS</td>
<td>Common</td>
<td>Cell cycle control?</td>
</tr>
<tr>
<td>SLC4A7/NEK10?</td>
<td>3p24.1</td>
<td>GWAS</td>
<td>Common</td>
<td>Transport?</td>
</tr>
<tr>
<td>COXI1/STXB8?</td>
<td>17q22</td>
<td>GWAS</td>
<td>Common</td>
<td>Apoptosis</td>
</tr>
<tr>
<td>MRPS30?</td>
<td>5p12</td>
<td>GWAS</td>
<td>Common</td>
<td>Signaling/immune response?</td>
</tr>
<tr>
<td>NOTCH2/GFGR1B?</td>
<td>1p11.2</td>
<td>GWAS</td>
<td>Common</td>
<td>Homologous recombination repair</td>
</tr>
<tr>
<td>RAD51LI</td>
<td>14q24.1</td>
<td>GWAS</td>
<td>Common</td>
<td>Cyclin-dependent kinase inhibitors?</td>
</tr>
<tr>
<td>CDKN2A/CDKN2B?</td>
<td>9p21</td>
<td>GWAS</td>
<td>Common</td>
<td>Cell cycle control/fibroblast growth factors</td>
</tr>
<tr>
<td>MYEOV/CCNDL?</td>
<td>11q13</td>
<td>GWAS</td>
<td>Common</td>
<td>Zinc finger protein gene</td>
</tr>
<tr>
<td>ZNF365?</td>
<td>10q21.2</td>
<td>GWAS</td>
<td>Common</td>
<td>Helicase?</td>
</tr>
<tr>
<td>ANKRD16/FLXO18?</td>
<td>10p15.1</td>
<td>GWAS</td>
<td>Common</td>
<td>Regulates transcription factors?</td>
</tr>
<tr>
<td>ZMIZ1?</td>
<td>10q22.3</td>
<td>GWAS</td>
<td>Common</td>
<td></td>
</tr>
</tbody>
</table>

Notes: ? refers to ‘possible’ gene or function in the breast cancer context. There is uncertainty about the exact genes and their functional roles in breast cancer.

Abbreviation: BC, breast cancer; GWAS, genome-wide association studies.

The Application of Clinical Genetics downloaded from https://www.dovepress.com/ by 54.70.40.11 on 09-Apr-2019 For personal use only.
Powered by TCPDF (www.tcpdf.org)
Among the set of well-established common susceptibility genes are variants in intron 2 of the FGFR2 gene,\textsuperscript{24} which, among the common variants, are likely to make one of the larger contributions to relative risk, at least for postmenopausal disease. Easton et al\textsuperscript{25} found that the rs2981582 SNP (allele frequency 0.38) contributes odds ratios of 1.23 and 1.63 for heterozygote and homozygote genotypes, respectively. The FGFR2 gene encodes a fibroblast growth factor (FGF) receptor. FGFs and their corresponding receptors are involved in regulation of the proliferation, survival, migration, and differentiation of cells. The considerable importance of FGF signaling in a range of tumor types is now becoming recognized.\textsuperscript{26} SNPs within intron 2 are involved in FGFR2 upregulation, and aberrant signaling activation induces proliferation and survival of tumor cells.\textsuperscript{27} The identification of this gene, which was unanticipated as a cancer gene, has prompted research into related genes and their potential roles in cancer. Other FGFs (eg, FGF-8) appear to be involved in breast cancer cell growth through stimulation of cell cycle and prevention of cell death.\textsuperscript{28}

Other low-penetrance variants that have been identified through GWAS include CASP8 (caspase 8), which encodes an apoptotic enzyme.\textsuperscript{29} The variant rs1045485 is protective, contributing odds ratios of 0.89 and 0.74 for heterozygotes and rare homozygotes, respectively. Recently, variants in CASP8 have been shown to alter risks (in a protective direction) in individuals with a family history of breast cancer.\textsuperscript{30}

Breast tumors are classified according to whether they have receptor proteins that bind to estrogen and progesterone. Such cells are termed ER+ and PR+ and require estrogen and progesterone to grow. Conversely, ER− and PR− tumors lack the protein that allows the hormones to bind. Tumor classifications influence the choice of treatment regimes for the patient. A further classification arises through tumors that overexpress the human epidermal growth factor receptor 2 (HER2) gene, which are termed HER2+ (conversely, HER2−). The triple-negative subtypes are ER−, PR−, and HER2− and are characterized by aggressive tumors and reduced range of effective treatment options. Several common gene variants are more strongly associated with specific cancer subtypes. These include the TOX3 gene, formerly called TNRC9 in which variant rs3803662 contributes a 1.64-fold homozygote risk, specifically in ER+ cancer.\textsuperscript{31} This gene encodes a high-mobility group chromatin-associated protein and increased expression is implicated in bone metastasis.\textsuperscript{32} Fine mapping has shown that hypothesized susceptibility variants lie in an intergenic region consistent with a gene regulatory function.\textsuperscript{33} These authors note there remains uncertainty as to whether the causal variant is actually involved in the regulation of the nearby retinoblastoma-like gene 2 (RB2L2) gene, which is involved in cell cycle regulation, given gene expression evidence.

The mitogen-activate protein kinase (MAP3K1) breast cancer gene\textsuperscript{25} is a member of the Ras/Raf/MEK/ERK signaling pathway (as is FGFR2) and is involved in regulating transcription of a number of cancer genes. MAP3K1 has been found to be more strongly associated with ER+ and PR+ tumors than ER−/PR− subtypes. There is also a stronger association with HER2+ tumors.\textsuperscript{34}

The LSP1 gene was identified as a breast cancer susceptibility locus by Easton et al,\textsuperscript{25} who identified an SNP within intron 10 as the most strongly associated. LSP1 encodes lymphocyte-specific protein 1, which is an F-actin binding cytoskeletal protein. The same study also identified a breast cancer variant in the 8q24 region containing no known genes. This region is also associated with prostate cancer.\textsuperscript{35} Stacey et al\textsuperscript{31} identified a SNP on 2q35, a region with no known genes, as associated with breast cancer in Icelandic patients with ER+ breast cancer. Milne et al\textsuperscript{36} also found an association with ER− disease, although there was a stronger signal for ER+. Other breast cancer associations include signals on 3p24, potentially relating to the genes SLC4A7 or NEK10, and on 17q22, perhaps related to COX11. These SNPs contribute odds ratios of 1.11 and 0.97 for heterozygote and homozygote genotypes, respectively.\textsuperscript{37} Additionally, a common variant close to MRPS30 on 5p12 was found to confer higher risk of ER+ disease.\textsuperscript{38} Turnbull et al\textsuperscript{39} described five new associations on chromosomes 9, 10 (three regions), and 11. Two further signals reported by Thomas et al\textsuperscript{40} include a SNP in the pericentromeric part of chromosome 1, within a region containing NOTCH2 and FCGR1B, and a signal associated with another double-strand break repair gene (RAD51L1) on 14q24.1. There is evidence that the chromosome 1 locus is more strongly associated with ER+ disease.

Considerable additional follow-up investigation will be required to establish the relationships between many of the SNPs and the actual causal variant(s) and to further elucidate the role in disease for many of these common genes.

**The genetic basis of breast cancer subphenotypes**

Analysis of breast cancer as a single phenotype is becoming less typical as genetic differences between disease subtypes
are more clearly established. Increased power to detect genetic variants is expected using patients belonging to genetically more homogeneous subgroups, rather than analyzing more heterogeneous groupings. There is evidence that many breast cancer GWAS studies have been enriched with ER+ cases because ER positivity is found in a higher proportion of the later-onset (usually postmenopausal) cases used in most of these studies. For this reason, ER+ disease is better characterized genetically than ER− disease. For example, Stacey et al39 identified two SNPs on chromosome 5p12 that confer risk preferentially for ER+ tumors. Garcia-Closas et al41 showed that variants in FGFR2 are more strongly related to ER+ than ER− (and also more strongly associated with PR+, low tumor grade, and lymph node-positive tumors). The breast cancer association in the 8q24 region is significantly stronger for ER+, PR+, and low-grade tumors. Reeves et al42 examined risk odds ratios for low-penetrate breast cancer genes in a sample of more than 10,000 cases and controls in relation to ER+ and ER− classification, for bilateral and unilateral disease, and for lobular versus ductal tumors. They noted higher odds ratios for ER+ disease for FGFR2 and TCN9, compared with ER− disease, greater association with bilateral, compared with unilateral, and for lobular disease compared with ductal disease in the 2q region. Using a polygenic risk score, based on seven breast cancer SNPs, the estimated cumulative incidence in the top fifth of the score distribution for ER+ disease is 7.4% compared with only 1.4% for ER− disease. Since the polygenic risk score is substantially more strongly predictive for ER+ disease, there is a strong case for more thorough evaluation of the genetic basis of the ER− subtype.

Triple-negative breast cancers are associated with poor prognosis due to aggressive tumor behavior and poor response to chemotherapy.43 After screening 2301 triple-negative cases and 3949 controls, Antoniou et al44 identified five SNPs on 19p13 that modify risk in BRCA1 mutation carriers and are specifically associated with triple-negative breast cancer. Additional phenotypic subtypes which are currently being interrogated genetically include differences in susceptibility variants between racial groups and in response to treatment and prognosis.

Genetic risk factors for breast cancer: clinical applications

Mutations in the BRCA1 and BRCA2 genes are rare but underlie severe and early-onset forms of the disease. Screening for mutations in women with a strong family history, usually linked to BRCA mutations, determines individual risks for this early-onset form of disease. However, most patients (~95%) do not show clear-cut family histories of early- or later-onset disease. The role of more common breast cancer variation in risk prediction is far less well established. Pharoah et al45 determined a multiplicative model using the 12 most significant common variants to define individual relative risks in the range 0.4-fold to fourfold compared with the general population. Given that there is a 12% population lifetime risk, deleterious common mutations contribute a 24%–36% lifetime risk, which may be high enough to instigate earlier and more intensive screening for common genetic forms of the disease. Gulcher and Stefansson46 point out that some women classified as at average risk would be reclassified as at higher risk based on their profile of common breast cancer variation. Similarly, some women might be reclassified as having lower-than-average risk based on their common gene profile. Risk estimates might be more reliably determined by multiplying risks from the genetic profiles with independent risks from conventional measures, such as family history, age at menarche, and pregnancy history. Successful application of common breast cancer gene profiles in clinical practice would have potential benefits by facilitating earlier diagnosis, reduced costs, less intensive therapeutic intervention, and disease management in the longer term.

As understanding of the genetic basis of breast cancer increases, further refinement in genetic risk models can be expected. The different genetic basis of tumor subtypes is a clear example of where refinement might take place as genetic profiles become predictive of tumor characteristics. At this stage, it is already well established that women with, or at higher risk for, ER+ cancer are good candidates for treatment with tamoxifen or raloxifene that specifically targets ER+ disease.

Finding the missing heritability

The breast cancer genes identified thus far explain only about 30% of the heritability, which is the proportion of the phenotypic variance that can be attributed to genetic variation. There are several possible sources for the missing genes, and this is a subject of intense argument and ongoing research.

Undetected common variation

GWAS using SNPs target only high-frequency alleles, and risk alleles found through these methods all have frequencies well in excess of 0.05.22 Even within this common allele ‘window’, the SNP panels provide incomplete genome coverage, due in part to technical limitations of the genotyping platforms, but mainly due to cost, which places reliance on tagging SNPs (using a SNP in linkage disequilibrium with many others to
represents or tags a specific haplotype). Such an approach is cost effective but loses information. Furthermore, these platforms are relatively enriched for nonsynonymous coding SNPs (cSNPs), so the coverage of synonymous cSNPs and noncoding SNPs is incomplete. Given that common disease variants include a higher proportion of regulatory SNPs, which lie outside coding regions, it is likely that important common variation has been missed by the GWAS undertaken thus far. Because effect sizes of common variants are low, very large samples of cases and controls are required for effective GWAS. Many as yet undetected common variants will have increasingly small effects on risk as variants with larger effect sizes will have already been detected through the completed GWAS. The largest study to date of common variation underlying a complex trait is the analysis of the genetic basis of height. Allen et al²⁸ tested data from 183,727 individuals and identified hundreds of common genetic variants in at least 180 loci that account for ~10% of the phenotypic variation in height. They estimated that as yet unidentified common variation (with similar effect sizes to those already found) will eventually account for ~20% of the heritable variation, but detecting these would require a sample size of 500,000 individuals. Importantly, they concluded that many genetic loci underlying variation in height show allelic heterogeneity suggesting that as yet unidentified causal variants will map to the loci already identified in GWAS. These missing variants are likely to span the allele frequency spectrum, including rare variants with higher penetrance, but the remaining low-penetrance variants can only be detected by ever-larger GWAS.

**Structural variation**

Structural variation, such as copy number variants (CNVs), which are not well tagged by SNPs in current arrays, may be a source of missing heritability in breast cancer. There is evidence that at least the common CNVs are in strong linkage disequilibrium with common SNPs genotyped in GWAS and hence may be adequately ‘tagged’ by existing panels.⁴⁰ Significant associations with rare CNVs (frequency range 0.2%–1%) have been identified for a number of neuropsychiatric traits, such as autism, epilepsy, and mental retardation,⁴⁰ although no CNVs have been convincingly associated with cancer phenotypes thus far.⁴⁰

**Gene–gene and gene–environment interaction**

Other possibilities include interaction effects between genes and between genes and environment. Exploring such scenarios presents analytical challenges and there is relatively limited evidence for an important role for interaction thus far. Ritchie et al⁴¹ modeled data for 10 SNPs in the genes COMT, CYP1A1, CYP1B1, GSTM1, and GSTT1. They identified an interaction between all the genes that were significantly associated with increased risk for sporadic breast cancer. Briollais et al⁴² also identified SNP–SNP interactions associated with breast cancer, including an interaction between XPD and IL10 genes as the most significant two-way interaction. Travis et al⁴³ examined the relationship between environmental variables, such as reproductive, behavioral, and anthropometric factors, with low-penetrance breast cancer genes. After allowing for multiple testing, they observed no evidence for increased breast cancer risk arising through gene–environment interaction in their sample of 7610 women. Because of the potentially huge number of statistical tests in such comparisons, obtaining a large enough sample to have power to demonstrate an effect can be difficult. Furthermore, confirmatory studies, along with functional analyses of the biological pathways involved, are essential to fully comprehend the importance of putative gene–gene and gene–environment interactions.⁴⁴ Moore et al⁴⁵ argued that the information gleaned from GWAS data collected thus far has been limited by failure to integrate existing knowledge about disease pathology: the ‘single SNP’ analysis approach ignores the genomic and environmental context. They recommend enhanced bioinformatic approaches to develop a holistic approach that recognizes the full complexity of gene–phenotype, gene–gene, and gene–environment interactions.

**Undetected rare variation**

Searching for rarer variants with larger effect sizes is likely to be a successful strategy for identifying more of the missing heritability. Rare variants have not been screened by GWAS, so this source of novel genetic variation is largely unexplored. Rare variants may contribute odds ratios in the range 2–5, compared with common variants that typically have odds ratios <1.5.⁴⁶ Targets for ongoing and future studies include low-frequency variants with minor allele frequencies in the 0.3%–5% range. Studies exploiting next-generation sequencing include Johansen et al,⁴⁷ who tested association with high triglyceride levels and resequenced loci previously identified as containing common variation. They found approximately twice as many rare coding genetic variants associated with high triglycerides located within the same genes. Determining the extent to which low-frequency and rare causal variants are colocated within breast cancer loci already identified in GWA studies.
depends on future sequencing efforts targeting these well-established genes. There is clearly a strong case to examine susceptibility variants over the full allele frequency range.\textsuperscript{16} Next-generation sequencing for the analysis of breast cancer exomes (the ∼30 Mb of sequence within protein-coding exons) of patients with early onset and a strong family history, which are negative for known highly penetrant rare mutations (\textit{BRCA1}, \textit{BRCA2}, and \textit{TP53}), are likely to be informative. Using exome sequencing, the full complement of, for example, SNPs and insertion–deletion polymorphisms can be characterized in every sample. Support for this strategy comes from the identification of rare highly penetrant mutations in the \textit{RAD51C} gene.\textsuperscript{16} For less completely penetrant variants, there are, however, many more difficulties in assessing the significance of the variation identified. The 1000 genomes project\textsuperscript{23} provides reference sequence for studying relationships between phenotype and genotype. The pilot phase describes exon-targeted sequencing of 697 individuals and whole genome sequencing of 179 individuals. The study determined the location and frequency of 15 million SNPs, 1 million insertion/deletion polymorphisms, and 20,000 structural variants. Comparing the pattern of variation identified in disease samples with this catalog of ‘normal’ variation is a crucial step in the process of determining the disease significance of any variants found.

Scale-up to sequence exomes of much larger samples of cases to investigate variants with intermediate and lower penetrance has not been undertaken thus far. Large samples will be required, and genetic heterogeneity, combined with the huge volume of (mostly unimportant) variation uncovered, poses extreme challenges for data interpretation even given knowledge about ‘background’ variation provided by the 1000 genomes project. In these cases, assessment of potential functional roles for variants found requires integration of information on gene expression profiles and other sources of transcriptome data and implementation of bioinformatic approaches to predict functional effects. Exome sequencing has its limitations. Apart from the fact that only exons are screened and that much important variation is known to reside outside these regions, there is limited information on structural variation, such as CNVs. Whole genome sequencing will generate a complete catalog of the variation, but many issues concerning the management, analysis, and interpretation of the huge volumes of data generated are not yet resolved.

**Conclusion**

In recent years, ∼30 genes and gene regions have been confirmed as containing variants underlying susceptibility to breast cancer. The majority of recent discoveries have been low-penetrance common variants identified through GWAS with SNPs. Disease risks associated with these SNPs are low, typically much <1.5-fold. In many cases, the causal variant is unknown, and the associated marker is only in linkage disequilibrium with the actual site. In the majority of the cases, the role of these variants in causing disease is also unknown, but ongoing study is revealing novel insights into breast cancer biology. Areas of intensive research include investigation of the genetic basis of disease subtypes, for which there appear to be marked genetic differences, the impact of genetic variation on prognosis and on response to treatment. Despite the huge amount of work undertaken thus far, ∼70% of the disease heritability remains unexplained. The common, low-penetrance variants identified through GWAS have contributed only a small proportion of this missing heritability. Aside from rare variants in the \textit{BRCA1} and \textit{BRCA2} genes and a small number of other rare genes that show approximately Mendelian patterns of inheritance, the majority of breast cancer genes found contribute little toward the prediction of individual disease risk. A thorough understanding of the biological role of the variation detected is some way off, and much more detailed functional and bioinformatic analysis is required for further progress. In the meantime, analysis of breast cancer exomes to identify SNPs and insertion–deletion polymorphisms will provide important insights by providing the first opportunity to examine rarer forms of variation in coding regions. This strategy will be effective for variants with higher penetrance, but where penetrance is toward the lower end of the spectrum, interpretation of the roles of numerous rarer variants will present new challenges for bioinformatic and functional assays. Once these problems are resolved, exome and whole genome sequencing strategies are likely to offer the best opportunity to identify additional breast cancer genetic risk factors. The identification of these genes is the crucial first step in fully comprehending the biology of disease and moving toward individualized treatments.

**Disclosure**

The authors report no conflicts of interest in this work.

**References**


