Search for new breast cancer susceptibility genes

Oldenburg, R.A.

CHAPTER 6

DOES THE BRCAX GENE EXIST? FUTURE OUTLOOK

Genetic research aimed at the identification of new breast cancer susceptibility genes is at an interesting crossroad. On the one hand, the existence of extended kindred's with many cases of (early-onset) breast cancer, in which a role for *BRCA1* and *BRCA2* has been excluded with high certainty, strongly suggests that there are still *BRCA1/2*-like genes to be found.[75] On the other hand, the absence of significant linkage signals in a set of 149 non-*BRCA1/2* breast cancer families indicates that if such a locus exists, it is unlikely to explain a major proportion of non-*BRCA1/2* families.[235] Are further 'classical' linkage studies therefore futile?

Before dismissing linkage analysis entirely, one should realize that the linkage results published to date do not permit a formal exclusion of the possibility that there are multiple genes causing breast cancer risks comparable to *BRCA1* or *BRCA2*, i.e., *BRCA3*, *BRCA4*, etc., but that their individual mutation frequencies are so low that each will explain no more than 10% of the families under study. The statistical power required to significantly resolve that kind of genetic heterogeneity was not achieved by any of the published linkage studies to date. Even in the largest study of 149 families, only 24 families had four or more cases of breast cancer diagnosed under 50 and 74 had four or more cases diagnosed before the age of 60 (ref. 235 and D. Easton, personal communication). For comparison's sake, previous studies addressing genetic heterogeneity analyzed more than 200 such families.[75,224] If four or five of those were in fact linked to a hypothetical *BRCA3* locus, one would have to be extremely fortunate with the informativity of the genotyped markers and patients to detect a significant linkage peak. For this reason, the Breast Cancer Linkage Consortium is now undertaking a study with the aim to obtain linkage data on at least 250 breast cancer families. Even though it might appear as if classical linkage approaches are running out of steam with regards to their potential to detect new breast cancer susceptibility loci, our claim on chromosome 9 proves that it's still possible to identify new breast cancer risk loci with classical linkage when genetic heterogeneity is reduced. In our case we selected families from the Dutch population, which is known to harbor many founder mutations for different diseases and therefore this group is, to an extent, genetically distinct. In addition, studies from different populations all show different susceptibility loci.[229,231,236,264] So, these loci might reflect population specific

effects. Our obtained linkage result suggests a region on chromosome 9 that may harbor a novel breast cancer susceptibility gene. However, if such gene exist in this region it's likely that it will account for only a limited fraction of the non-*BRCA1/2* families, especially in populations other than the Dutch.

Nowadays other new bioinformatics tools are being developed to improve linkage power. A promising option is to integrate chromosome segregation data with data obtained from functional genomic approaches such as large-scale, high-throughput molecular profiling technologies.[265]

Molecular profiling of global gene expression has already enabled the subclassification of breast cancer into prognostically relevant subgroups,[42,266] and has demonstrated to be capable of identifying cases who are very likely to be *BRCA1* gene carriers.[42,267] Similar results have been obtained by using array-CGH with tumor DNA.[255,256] On this basis, one could hypothesize that breast tumors with the same genetic etiology (because of a shared familial predisposition) will also be more likely to share a molecular signature. This could provide means to eliminate phenocopies from the linkage analysis, or to assign a liability to each patient as to how much her tumor resembles a typical *BRCAX*-related cancer. This will require sufficient resolution both in terms of number of probe sets, and the number of cases and controls to be analyzed, because initial results seem to suggest that *BRCAX*-related tumors resemble sporadic breast tumors in terms of somatic genetic and immunohistochemical make-up.[187,258,260] However, when comparing BRCAx-breast tumor array-CGH profiles with profiles of control samples we showed that there are multiple regions of differential gains and losses. Unfortunately, the dendrogram of BRCAx tumors indicated that there are no major branch points to suggest obvious distinct BRCAx CGH subtypes. Interesting though was the observation that when sporadic and BRCAx tumors were clustered together a non-random distribution occurred. Both BRCAx and sporadic tumors remained clustered in just eight sub clusters, five of which contained only BRCAx tumors and three clusters contained all sporadic cases plus one BRCAx tumor (van Beers/Oldenburg submitted, *Cancer Research*). Although, it was not possible to distinguish different groups within the BRCAX tumors at this point, it deserves further exploration.

Another intriguing possibility is to exploit gene expression patterns in normal cells of cases (such as lymphocytes or skin fibroblasts). This method relies on the fact that messenger RNA's with premature stop codons (i.e., nonsense mutations, which constitute a frequent class of mutations underlying inherited disease) are efficiently degraded by the conserved nonsense-mediated decay pathway. The number of genes

displaying consistently lower expression in normal cells of familial cases versus controls can be substantially reduced by comparing their genomic location to chromosomal segments shared among family-members identical-by-descent. An example of this approach is the recent identification of *AIP* as a susceptibility gene for pituitary adenoma in a Finnish founder population.[268]

Also, the discovery of micro-RNA's opens many new doors in cancer research (see the excellent review by G.A. Calin and C.M. Croce[269]). Micro-RNA's (miRNA) are a family of 19-25-nucleotide non-coding small RNA's that function as gene regulators and are involved in crucial biological processes, including development, apoptosis, proliferation and differentiation through pairing with target messenger RNA's (mRNA) of protein-coding genes. Perfect pairing of miRNAs to mRNAs leads to degradation of the mRNA, whereas less strict sequence complementarity results in translational repression. Recently, it has been shown that miRNAs can function either as tumor suppressors or oncogenes and the genomic abnormalities found to influence the activity of miRNAs are the same as those previously described for protein-coding genes, such as chromosomal rearrangements, genomic amplification or deletions and mutations. The role of miRNAs in tumor formation is strongly supported by the observation that the coding sequence of 50% of the known miRNAs, frequently situated in introns of coding genes, are located inside or close to fragile sites and minimal regions of LOH, minimal regions of amplification and common breakpoints associated with cancer. In addition, miRNA expression profiles showed that miRNAs are differentially expressed in normal and tumor samples (the expression in tumor tissue seems generally lower than in normal tissue) and that their expression fingerprints correlate with clinical and biological characteristics of tumors. In breast cancer the expression profile of a set of 15 miRNAs correctly predicted the nature of the breast cancer sample analyzed with 100% accuracy. Furthermore the expression of miRNAs was correlated with specific breast cancer pathological features such as estrogen-receptor status, tumor stage, vascular invasion, proliferation index and clinical features such as prognosis and response to therapy. Unfortunately, to date no attempt has been made to use miRNA expression profiles for subclassification of the heterogeneous group of familial breast cancer. It might be worthwhile to explore this possibility.

Obviously, the success of all of these approaches is dependent on the amount and quality of the information from the pedigrees under study and the availability of biological samples from the patients. The analysis of gene expression profiles of tumors still relies heavily on the availability of frozen tissue samples. Logistically, these

are particularly difficult to obtain from multiple affected family-members, which explains why so few of such studies have appeared in the published literature.

Notwithstanding these developments, attention is now shifting rapidly toward the whole genome association studies in population-based breast cancer cases and controls.[270] In a typical association study, the frequency of a genetic variant in affected individuals (cases) is compared to that in individuals without the disease (controls).[271,272] Allelic association is present when the distribution of genotypes differs in cases and controls. Most association studies are based on candidate genes that encode proteins thought to be involved in carcinogenesis, such as those involved in apoptosis, cell-cycle control, or DNA repair. Within candidate genes, variants for which a functional connotation can be inferred are preferably tested for association with disease. In this way, variants in *TGFβ1* and *CASP8* have been identified as breast cancer susceptibility alleles (see chapter 2, sections 3.2.2. and 3.2.3.).

Association studies are the only alternative to family-based linkage analyses for detecting alleles that confer low to moderate disease risks.[273,274] Even though risks are low, such alleles have the potential of explaining a substantial proportion of disease heredity, depending on their population frequency. There are several examples of common variants that contribute to common diseases,[275,276] but none of them could have been detected by linkage analysis. For example, the P12A variant in the *PPARG*-gene, which affects the risk of type 2 diabetes, would only be detected using linkage studies of over one million affected sib pairs.[277]

Mutations in the currently known high risk breast cancer genes are common in families with a large number of cases of breast and/or ovarian cancer,[75] but they have been estimated to explain at best 20-25% of the overall excess familial risk[278] and less than 5% of the total breast cancer incidence.[71] The proportion of breast cancer that can be attributed to genetic factors is not clear, but several studies have suggested it to be much larger than 5%. A large twin study has estimated that up to 30% of all breast cancer has a genetic basis,[279] while a study of the incidence of contralateral breast cancer has even suggested that the majority of all breast cancer occurs in a small minority of women who are susceptible for it.[280] It is unlikely that further *BRCA1/2*-like genes, if they are detected, will be capable of accounting for these attributable risks, because their allele frequencies are already predicted to be rare. More common alleles with moderate effects could do so, but it is not clear how many of such alleles exist and how much of the genetic predisposition to breast cancer can be attributed to them. Assuming relative risks in the order of 1.3 – 1.5, the remainder of excess risk could equally well be explained by a few hundred common variants

(with frequencies of >1%) or thousands of rare variants.[270] Such a polygenic model has in fact been supported by segregation analyses in non-*BRCA1/2* families (chapter 2, section 4.2). Under this model, many low to moderate risk cancer susceptibility genes cause breast cancer predisposition, together with environmental risk factors, in a multiplicative or additive way, with no single gene accounting for a large fraction of the familial aggregation. Individuals carrying few such alleles would be at lower or equal population risk compared to those carrying multiple such alleles. Conversely, there will also be alleles that protect against the development of cancer (such as *CASP8*, chapter 2, section 3.2.3).

The major problem haunting association studies is the lack of reproducibility by other, independent studies.[57,58] Thus, most studies are too small and probable report false positive results due to chance (type 1 error), which depends on the level of significance used. Unfortunately, the levels of significance appropriate in other contexts (P=0.05 or P=0.01) can be highly misleading in association studies.[270] By using more stringent levels of statistical significance this false positive rate can be reduced. Alternatively failure to confirm associations might be the result of heterogeneity in risk between populations due to for example interacting lifestyle and environmental factors. Also, strikingly little research has been performed on combinations of polymorphisms. It is still possible that polymorphisms not associated with breast cancer when studied separately, are associated with breast cancer when studied in combination with other polymorphisms. For example, recently a significant trend in risk with increasing numbers of variant alleles for 25 SNP's in *BRCA1*, *BRCA2*, *ATM*, *TP53* and *CHEK2* was observed[281] whereas common polymorphic variants in these genes separately are unlikely to increase breast cancer risk.[282]

In addition, a lack of association of a candidate SNP does not necessarily rule out the presence of another important variant in the same gene. For any given gene of interest, there might be tens or even hundreds of different sequence variants.

A large genome-wide association study, which would involve millions of SNP's with the use of stringent significance levels, would be ideal to identify common breast cancer susceptibility genes. However, the number of cases to be genotyped depends on the allele frequency of the variant and the disease risk conferred by it. For relative risks in the order of 1.5, allele frequencies in the range 10% – 40% will require at least a thousand cases and controls to be genotyped[270,273] in order to have 90% power to detect associations at a significance level of 10^{-4}. Much larger numbers are required for allele frequencies in the 1% – 10% range. This is presently not feasible at the going costs per genotype per sample. Fortunately, it is not necessary to genotype all pos-

sible SNP's to detect an association, because the alleles of SNP's that are physically close to each other tend to be correlated with each other (tag SNP mapping). This phenomenon is called linkage disequilibrium (LD).[283-285] The ability of one SNP to report on another depends on the strength of LD between them. The general consensus is that an $r^2 > 0.8$ (r^2 is a measure of correlation between a pair of variables[273]) is sufficient for a tag SNP to obtain a good coverage of untyped common SNP's. It has been estimated that 200,000 – 500,000 tagging SNP's will be needed to adequately tag all SNP's with a minor allele frequency of 5% or more.[286,287]

Accumulating the data necessary to choose such SNP's is one of the main goals of the human HapMap project.[287,288] The existence of LD can also be exploited to examine candidate genes by haplotype analysis, whereby the haplotypes are defined by a set of tagging SNP's. Any common variant in the gene that increases disease risk will then be detected as an increase of the particular haplotype on which this variant resides. This approach has been used to investigate a possible involvement of common variation in *CYP19* and *BRCA2* genes.[289,290]

In addition, it has been proposed that the power to detect associations may be increased by genotyping familial cases rather than sporadic population-based cases,[291,292] an effect which was indeed observed for the *CHEK2*1100delC* variant.[175,176] It seems therefore the most efficient to perform a whole-genome scan for association in a small sample of cases that are enriched for susceptibility. These could be familial cases or early-onset bilateral breast cancer cases, but one could also select cases enriched for other risk factors with a strong genetic component such as breast density[29] or cellular radio sensitivity.[293]

The recent publication by the Breast Cancer Association Consortium proved the success of this strategie.[294] In the first stage a panel of 266,722 SNP's (selected to tag known common variants across the entire genome) was genotyped in 408 breast cancer cases with a strong family history of breast cancer and 400 controls. In the second stage 12,711SNP's (approximately 5% of those typed in stage one) were selected on the basis of the significance of the difference in genotype frequency between cases and controls and genotyped in 3,990 invasive breast cancer cases and 3,916 controls. In the third stage 30 of the most significant SNP's were tested in 21,860 cases and 22,578 controls. This resulted in five novel loci strongly associated with breast cancer with a significance level ranging from 2×10^{-76} to 3×10^{-9} of which four contain plausible causative genes; *FGFR2, TNRC9, MAP3K1* and *LSP1*. The five SNP's that reached an overall P-value $<10^{-7}$ showed an increased breast cancer risk of the minor allele in a dose dependent manner, with higher risk of breast cancer in homozygous

than in heterozygous carriers. It is notable that none of the confirmed associations reached genome wide significance after stage 1 and only one reached this level after stage 2, emphasizing the critical importance of study size in genetic association studies.

This study has also demonstrated conclusively that some of the variation in breast cancer risk is due to common alleles, as these five identified susceptibility alleles are very common. So, a high proportion of the general population is carrier of at risk genotypes. For example 14% of the UK population are homozygous for the rare allele. However, the increased risks associated with these alleles are relatively small. On the basis of UK population rates, the estimated breast cancer risk by age 70 years for rare homozygote's at the SNP in *FGFR2* is 10.5%, compared to 6.7% in heterozygote's and 5.5% in common homozygote's. It is likely that there are still other common variants to be identified as *CASP8* D302H, which showed strong evidence of association in a previous large study[172] was missed, because it did not reach the threshold for testing in stage 2. Also the excess of association ($P<0.05$) after stage 2 is consistent with the existence of many such loci. In addition, because the coverage for SNP's with minor allele frequency's <10 % was low, many low frequency alleles have probable been missed. How much of the overall familial risk these alleles will be able to explain remains to be seen. It has been argued on the basis of evolutionary arguments that the role of rare alleles (i.e., frequencies <<1%) in causing late-onset disease such as cancer could be substantial.[295] Detecting this class of variants by current genetic approaches is impossible. Probable it will require genome-wide studies with more complete coverage (perhaps total genome sequencing) and using much larger number of (familial) cases and controls. Over the next decade, progress with the identification of common low risk variants will teach us how substantial this fraction is.

The proposed polygenetic model would not only be capable of explaining large but rare autosomal dominant-like familial clusters of (early-onset) breast cancer, but could also explain substantial proportions of the total breast cancer incidence.[296] It has been estimated that, should we be able to characterize all the relevant risk factors in all women of a given population, 50% of all breast cancer would occur in 12% of women with the highest risk profile.[296] From a health care perspective, the identification of these risk factors is therefore of great practical importance. Not only to define the cancer risk for women and their family members in order to make adequate decisions on surveillance and preventive strategies, but also for the development of gene targeted therapy. For example PARP [poly(ADP-ribose) polymerase] inhibitors may represent a novel way of selectively targeting BRCA2- or p53-deficient

breast cancer cells. Probably due to the additional inhibition of PARP activity an increase of unrepaired DNA damage occurs, causing a shift from DNA repair to apoptosis.[297-300]

Due to the low risk of the newly identified genes and SNP's and the existence of more low risk alleles to be identified it is too early to include these in predictive genetic testing at this stage. However, as further susceptibility alleles are identified over the next years, combinations of such alleles together with other breast cancer risk factors may become sufficiently predictive to be important clinically. And ideally, a chip with all risk alleles for predictive genetic testing is constructed in the near future to test women at risk.