Review

High-throughput genomic technology in research and clinical management of breast cancer

Evolving landscape of genetic epidemiological studies

Yen-Ling Low[1], Sara Wedrén[2] and Jianjun Liu[1]

[1]Population Genetics, Genome Institute of Singapore, Singapore
[2]Department of Medical Epidemiology and Biostatistics, Karolinska Institutet, Stockholm, Sweden

Corresponding author: Jianjun Liu, liuj3@gis.a-star.edu.sg

Published: 28 June 2006
This article is online at http://breast-cancer-research.com/content/8/3/209

Breast Cancer Research 2006, **8:**209 (doi:10.1186/bcr1511)

Abstract

Candidate polymorphism-based genetic epidemiological studies have yielded little success in the search for low-penetrance breast cancer susceptibility genes. The lack of progress is partially due to insufficient coverage of genomic regions with genetic markers, as well as economic constraints, limiting both the number of genetic targets and the number of individuals being studied. Recent rapid advances in high-throughput genotyping technology and our understanding of genetic variation patterns across the human genome are now revolutionizing the way in which genetic epidemiological studies are being designed and conducted. Genetic epidemiological studies are quickly progressing from candidate gene studies to comprehensive pathway investigation and, further, to genomic epidemiological studies where the whole human genome is being interrogated to identify susceptibility alleles. This paper reviews the evolving approaches in the search for low-penetrance breast cancer susceptibility gene variants and discusses their potential promises and pitfalls.

Introduction

Family history is a well-established risk factor for breast cancer. Breast cancer risk is typically increased by two- to three-fold in first-degree relatives of affected individuals. Mutations in high-penetrance cancer susceptibility genes such as *BRCA1* and *BRCA2* account for less than 20% of the excess familial risk [1]. The remaining familial risk is likely to be explained by a polygenic model where breast cancer susceptibility is conferred by a large number of low-penetrance alleles. The risk conferred by each of these alleles may be small but these alleles may combine additively or multiplicatively to affect breast cancer susceptibility substantially [2]. Rare, high-penetrance susceptibility alleles have been successfully mapped using family-based linkage studies. Further progress in the search for genetic determinants of breast cancer likely lies in the identification of the large number of low-penetrance cancer susceptibility alleles by population-based genetic association studies.

Numerous genetic association studies on breast cancer have been published but results have been equivocal, partly due to shortcomings in study design [3]. The past few years have witnessed rapid advances in high-throughput technologies for genotyping analysis as well as in our understanding of genetic variation patterns across the human genome. These advances have empowered researchers to improve the design of genetic epidemiological studies, especially the way in which genetic variation is captured. In this short review, we will focus on the recent developments in high-throughput technologies for genotyping analysis and their impact on genetic epidemiological studies of breast cancer, addressing both their promises and pitfalls.

Candidate polymorphism analysis

The genetic association studies published on breast cancer from the 1990s onwards have typically compared the allelic and/or genotypic frequencies of selected polymorphisms between breast cancer cases and controls. These studies aimed to find polymorphisms that may be directly related to breast cancer risk as causal variants or indirectly related to breast cancer risk due to being in linkage disequilibrium (LD) with causal variants. These studies typically start with the selection of candidate genes based on current biological understanding of their potential role in breast cancer carcinogenesis. Then a small number of polymorphisms are selected in these genes and genotyped. Polymorphism selection has usually been based on isolated reports of a polymorphism's potential functional effect, such as coding variants, and/or its feasibility to be successfully genotyped at that time.

LD = linkage disequilibrium; SNP = single nucleotide polymorphism.

Moving from family-based linkage studies to population-based genetic association analysis causes a shift from microsatellite markers to single nucleotide polymorphisms (SNPs) as the leading marker for genetic analysis. Microsatellite markers have been extremely useful in mapping causal genetic variants in family pedigrees and have been successfully used to identify high-penetrance genes, as in the case of *BRCA1* [4]. But microsatellite markers are less efficient in population-based genetic association analysis and have rarely been used in the search for low-penetrance alleles using unrelated subjects [5,6], partly due to their relatively high mutation rate and complex mutation patterns. Compared to microsatellite markers, SNPs are stable, more abundant, associated with lower genotyping error, easier to automate and thus cheaper in terms of cost and labor. The availability of detailed information on LD patterns of SNPs has also enabled genetic variation to be captured more effectively using SNPs. Hence, SNPs have increasingly dominated the field of population-based genetic association studies in breast cancer. Examples of genes investigated using candidate SNPs include the steroid hormone metabolism genes (*CYP17*, *CYP19*, *COMT*, *SHBG*), estrogen-signaling genes (*ESR1*, *ESR2*), carcinogen metabolism genes (*CYP1A1*, *NAT1*, *NAT2*, *GSTM1*) and DNA repair genes (*XRCC1-3*, *ATM*) [7-9]. Although being commonly termed candidate gene analysis, such studies can at most qualify as candidate polymorphism analysis since only a very small number of polymorphisms within each gene were evaluated and these cannot be assumed to represent the whole gene, especially if the gene is large.

Despite huge efforts being invested in population-based genetic association studies of breast cancer, the outcome has not been satisfactory. The low throughput and high cost of genotyping analysis has constrained investigators to studying only a few polymorphisms within a few candidate genes in a limited number of samples. Positive results have been rare and often not replicated in subsequent studies. It is possible that the generally negative findings of past studies may be due to a true absence of risk alleles of moderate to high effect for breast cancer. But given both poor coverage and inadequate power of past studies, causal alleles are likely to be missed even if they exist. Hence negative results of such studies could not be used as evidence to rule out the role of a particular gene in breast cancer risk. To illustrate the problem of inadequate power, a systematic review of genetic association studies of breast cancer found 46 case-control studies published between 1983 and July 1998. Most studies were small, with the median number of cases and controls combined being 391 (range 58 to 1,431). From power calculations, a study of 315 cases and 315 controls will be needed to detect a risk allele with a frequency of 20% conferring a relative risk of 2.5 with 90% power at the 5% significance level. Only 10 out of 46 studies met these criteria [8]. It has been further argued that to reduce false positives arising from multiple testing, a significance level of 10^{-4} should be used for candidate gene studies. Then a study of approximately 1,000 cases and 1,000 controls will be needed to detect a susceptibility allele with a frequency of 20% conferring a relative risk of 1.5 [10]. Few candidate polymorphism studies in breast cancer have managed to fulfill such criteria. In summary, limited progress has been made by such candidate polymorphism-based genetic epidemiological studies in identifying low-penetrance risk alleles for breast cancer.

Recent developments in high-throughput genotyping technology

The rapid development of high-throughput technology for SNP genotyping over the past few years has resulted in a wide variety of SNP genotyping platforms now available for use, each with unique features. On platforms such as the Illumina BeadArray™ and the Affymetrix GeneChip® array systems, up to thousands of SNPs can be analyzed simultaneously (i.e., multiplexed) in each sample. These have dramatically increased the throughput of genotyping and brought down the genotyping cost per SNP. Such platforms are well suited for large-scale screening studies where thousands of SNPs are analyzed in a fair number of samples. However, due to their high level of multiplexing, total cost, and sometimes lengthy process of initial assay development, these platforms become unwieldy in studies where only a moderate number of SNPs needs to be analyzed. For such studies, Sequenom's MassARRAY® system is one of the better choices as it only requires up to 29 SNPs for each multiplexing assay and requires short assay development time by investigators themselves. Such systems provide greater flexibility and efficiency for investigators to carry out either medium-size studies that target a moderate number of candidate genes or follow-up studies where a limited number of positive findings from initial large-scale screening studies are further investigated in large samples. In situations where only single or a very limited number of SNPs need to be analyzed in a large number of samples (e.g., in confirmation studies), methods such as TaqMan® and Pyrosequencing™ assays are more suitable. Such systems can only genotype very few SNPs at a time but are very robust and efficient. A summary of the main features of some of the main genotyping platforms available for custom SNPs is shown in Table 1. A detailed discussion of SNP genotyping technology is beyond the scope of this review but has been reviewed elsewhere [11-13].

The technological limit of genotyping analysis has been further challenged by the recent release of ultra high-throughput systems from Illumina and Affymetrix. Innovative multiplexing chemistry allows these systems to analyze between approximately 317,000 SNPs (Illumina's Sentrix® humanHap300 beadchip and Infinium™ II assay) and 500,000 SNPs (Affymetrix's GeneChip® Mapping 500K Array) in a single experiment. Both systems are of fixed contents, meaning that all the SNPs for analysis have been

Table 1

Main features of some custom SNP genotyping platforms available

	Assay design	Multiplexing capability	Throughput (no. of samples per 8 hour working day)	Cost per genotype (at maximal multiplexing)[a]	Type of study design that platform is suitable for
TaqMan®	By manufacturer or flexible design by investigator	No	Up to 10,000+	>US$0.30	Small number of SNPs, large sample size
Pyrosequencing™	Flexible design by investigator	From 1 to 3	Up to 4,000+	>US$0.30	Small number of SNPs, large sample size
Sequenom MassARRAY®	Flexible design by investigator	From 1 to 29-plex	Up to 3,000+	US$0.05-0.10	Moderate number of SNPs, small/moderate sample size
Illumina BeadArray™ (GoldenGate® Assay)	By manufacturer	From 96 to 1,536-plex	Up to 192	<US$0.05	Large number of SNPs, small/moderate/large sample size
Affymetrix GeneChip®	By manufacturer	From 1,500 to 20,000-plex	Up to 16	<US$0.05	Large number of SNPs, small/moderate/large sample size

[a]The estimates are heavily influenced by the size of study, that is, large studies will enjoy more efficient usage of reagents and volume-based price discount from manufacturers than medium and small studies. SNP, single nucleotide polymorphism.

pre-selected by the manufacturers. While Illumina's SNP selection is based on the available information on allele frequency and the LD pattern of the human genome from the HapMap project, Affymetrix's SNP selection is generally random and mainly based on the SNPs' feasibility to be genotyped. By driving down the genotyping cost below US$0.01 per SNP, such systems have transformed whole-genome association analysis into reality.

The technological advancements in genotyping analysis, coupled with the extensive collection of validated SNPs and knowledge of LD patterns across the human genome from the HapMap project, have transformed the landscape of genetic epidemiological studies. These advancements have allowed us to progress from the investigation of candidate polymorphisms to truly comprehensive candidate gene and whole-genome studies.

Comprehensive candidate gene study using the haplotype tagging approach

Knowledge of LD patterns across different genes has given rise to the haplotype tagging approach as an efficient way of conducting comprehensive candidate gene studies. Due to the extensive non-independence between SNPs and the limited haplotype diversity within regions of strong LD (LD blocks) in the human genome, only a subset of selected SNPs, instead of all variants, needs to be analyzed to capture the majority of common genetic variation within such blocks. With an average LD block size of between 11 and 22 kb and assuming 3 to 5 haplotypes per block, it has been estimated that around 300,000 to 1,000,000 well-chosen tagging SNPs (in non-African and African samples, respectively) would be required to capture the 10 million SNPs that are thought to exist [14]. Equipped with large sample sizes and efficient coverage of all genetic variation within candidate genes, current genetic epidemiological studies are expected to stand a good chance of detecting susceptibility alleles with moderate effects, if they exist. While current genetic association studies are being geared up to a comprehensive coverage of common variants and are thus greatly enhancing the confidence of a negative result, it will be difficult to assertively exclude the role of a candidate gene purely based on the results of LD mapping. Although there is general agreement on the merits of using the haplotype tagging approach in genetic association studies, there are pitfalls [15] and active discussions are still ongoing on several issues, including optimizing tagging SNP selection [16,17] and haplotype construction [18], as well as statistical analysis of such SNP/haplotype data to study disease associations [19].

Genetic association studies on breast cancer that have used haplotype tagging SNPs for candidate gene analysis are starting to appear in the literature. Some examples of genes studied in this manner include *CYP19* [20], *HSD17B1* [21], *EMSY* [22] and *CHEK2* [23], and more results are expected in the near future. Currently, published studies have focused on assessing genetic variation within single candidate genes, but more efforts will be needed to evaluate entire biological pathways or gene families. Genes often work together as part of complex biological pathways. Selecting a single candidate gene within a pathway for genetic epidemiological investigation is likely to be over-simplistic. Instead, genetic variability of entire biological pathways, for example, the estrogen metabolism pathway, should be investigated to evaluate potential association with disease. Although it is no longer technologically challenging to capture most, if not all, of the

common genetic variation within a biological pathway using the haplotype tagging approach, the method for data analysis is not straightforward. Locus-by-locus analysis can detect SNPs associated with moderate main effects. But this method of analysis will become less effective in situations where breast cancer susceptibility is attributed to a fair number of alleles, each of which is only associated with a weak effect (below the threshold for detection) or in situations where susceptibility is attributed to the interaction of multiple SNPs, each with negligible effect. Therefore, success of comprehensive candidate gene studies will rely substantially on the development of new statistical methods for evaluating the cumulative effect of whole biological pathways on susceptibility to breast cancer.

Genomic epidemiological studies

The success of candidate gene studies, whether based on single genes or whole pathways, is constrained by our current biological understanding of breast carcinogenesis. Since breast carcinogenesis is a complex and still only partially understood process, it is likely that many important genes are overlooked in candidate gene studies. Such a limitation can only be overcome by genomic epidemiological studies where no prior biological hypotheses are assumed and the entire human genome is targeted for identifying genetic variation associated with breast cancer susceptibility. Several research groups have embarked on whole genome association studies in breast cancer but no results have been published yet. The use of whole genome scans in genetic association studies is still in its infancy. Design issues for genome-wide association studies are still evolving and have been reviewed elsewhere [24,25].

Although promising, genome-wide association studies bring about major challenges in regard to data analysis. Genetic epidemiological studies have conventionally been designed in such a way that a relatively small number of potential risk factors (both genetic and non-genetic) are evaluated in a much larger number of samples. Locus-by-locus approaches for statistical analysis are well developed for such designs to evaluate the main effect of a genetic variant and simple interactions between genetic variants. In contrast, genome-wide association studies are expected to involve analysis of hundreds of thousands of SNPs in several hundred (or thousand) samples. This means that the number of testing targets will be far greater than the number of samples, which is unfavorable for a conventional locus-by-locus statistical analysis approach. This issue has already emerged when attempting to extend the candidate gene approach to studying multiple genes in a pathway but will become greatly compounded in the whole genome analysis. By performing a locus-by-locus test on each of the hundreds of thousands of SNPs in a moderate sample size, a large number of false positive findings are expected to be generated in addition to the expected small number of true positive results. Because the true risk alleles are likely to be associated with moderate effects, the true positive association results are by no means guaranteed to enjoy stronger statistical evidence than the false positive ones. Although Bonferroni correction or false discovery rate can be used to control the adverse effect of multiple testing and reduce the false positive rate, they cannot improve the power for detection. As a means of validating initial positive findings, a two-stage design may be used in which a large number of potential positive findings from the initial genome-wide analysis are tested in a much bigger sample. But the efficiency of such a design still needs to be proven by real studies. Hypothesis-free attempts to identify interactions among genetic variants at the genomic level will be even more challenging, due to the immense number of tests involved. Initial simulation analysis has demonstrated the feasibility of performing genome-wide interaction analysis [26], but more will need to be done to verify its efficiency.

Future directions

Looking ahead, the technical barriers to genotyping are unlikely to be a limiting factor. Future breakthroughs in the search for breast cancer susceptibility genes will probably hinge heavily on devising novel data analysis strategies to make sense out of the vast amount of data generated. Although still speculative, novel statistical and/or mathematical approaches that allow the incorporation of the information of biological network and genomic structure will likely champion the field of data analysis.

With the vast amount of data generated from high throughput genotyping, many genetic association findings are expected. Replication will be needed and functional verification will need to be conducted to identify true causal alleles. Efforts to devise efficient methods for functional validation would accelerate the accumulation of well-founded evidence. Despite all the promises held by genome-wide association studies, if such studies are not handled properly, large numbers of false positive results will be generated and published. This will result in a significant drain in resources invested in studies with slim prior probabilities of yielding significant findings, which would slow down the search for breast cancer susceptibility genes. Recognizing the promises and the pitfalls of such genomic approaches, efforts are already underway to coordinate genetic association studies to build a roadmap for efficient and effective human genomic epidemiology [27].

Apart from genetic factors, environmental and lifestyle factors also play a substantial role in affecting breast cancer risk [28-30]. Low penetrance genes most likely act in concert with lifestyle and other environmental factors to affect breast cancer risk. The subtle effects of some genetic variants may be magnified and only become detectable in the presence of certain exposures. Failure to take into account these external factors may hinder the search for breast cancer susceptibility gene variants. For example, the associations between

Box 1

Glossary of terms

Genetic epidemiology	The study of the relationship between variation in specific genes and disease risk
Genomic epidemiology	The study of the relationship between variation across the entire human genome and disease risk
Haplotype	A set of closely linked alleles that tend to be inherited together
HapMap project	A multi-country effort to identify and catalog common genetic variants in humans and work out their haplotype structures
Linkage disequilibrium	The phenomenon that alleles physically close to each other tend to be correlated and are co-inherited as a block of DNA segment
Microsatellite	A type of DNA sequence variation where there is tandem repetition of a short DNA sequence (usually two to four nucleotides)
Penetrance	Probability that a deleterious gene variant will actually result in disease
Polymorphism	Variation in DNA sequence among individuals
Single nucleotide polymorphism	A type of DNA sequence variation in which a single nucleotide (A, T, C, or G) in the genome sequence is altered

polymorphisms in DNA repair genes and breast cancer risk were only detectable in women with a high intake of folate and carotenoids [31,32]. Studies of such gene-environment interactions will not only help in the search for low-penetrance gene variants affecting breast cancer risk, but can also uncover ways by which risk may be modified.

Finally, it deserves to be mentioned that no amount of genetic, technological or statistical sophistication can compensate for a badly devised study. Sound epidemiological design remains fundamental in order to obtain valid and reproducible genomic epidemiological results. Sufficient numbers of carefully defined cases and appropriately chosen controls with accurate information about potential confounders and effect modifiers are needed. Ideally such study samples will be derived from large prospective studies.

This article is part of a review series on *High-throughput genomic technology in research and clinical management of breast cancer*, edited by Yudi Pawitan and Per Hall.

Other articles in the series can be found online at http://breast-cancer-research.com/articles/review-series.asp?series=bcr_Genomic

Competing interests

The authors declare that they have no competing interests.

References

1. Anglian Breast Cancer Study Group: **Prevalence and penetrance of BRCA1 and BRCA2 mutations in a population-based series of breast cancer cases. Anglian Breast Cancer Study Group.** *Br J Cancer* 2000, **83:**1301-1308.
2. Pharoah PD, Antoniou A, Bobrow M, Zimmern RL, Easton DF, Ponder BA: **Polygenic susceptibility to breast cancer and implications for prevention.** *Nat Genet* 2002, **31:**33-36.
3. Colhoun HM, McKeigue PM, Davey Smith G: **Problems of reporting genetic associations with complex outcomes.** *Lancet* 2003, **361:**865-872.
4. Hall JM, Lee MK, Newman B, Morrow JE, Anderson LA, Huey B, King MC: **Linkage of early-onset familial breast cancer to chromosome 17q21.** *Science* 1990, **250:**1684-1689.
5. Risch NJ: **Searching for genetic determinants in the new millennium.** *Nature* 2000, **405:**847-856.
6. Savage SA, Chanock SJ: **Using germ-line genetic variation to investigate and treat cancer.** *Drug Discov Today* 2004, **9:**610-618.
7. Kang D: **Genetic polymorphisms and cancer susceptibility of breast cancer in Korean women.** *J Biochem Mol Biol* 2003, **36:** 28-34.
8. Dunning AM, Healey CS, Pharoah PD, Teare MD, Ponder BA, Easton DF: **A systematic review of genetic polymorphisms and breast cancer risk.** *Cancer Epidemiol Biomarkers Prev* 1999, **8:** 843-854.
9. Coughlin SS, Piper M: **Genetic polymorphisms and risk of breast cancer.** *Cancer Epidemiol Biomarkers Prev* 1999, **8:** 1023-1032.
10. Pharoah PD, Dunning AM, Ponder BA, Easton DF: **Association studies for finding cancer-susceptibility genetic variants.** *Nat Rev Cancer* 2004, **4:**850-860.
11. Engle LJ, Simpson CL, Landers JE: **Using high-throughput SNP technologies to study cancer.** *Oncogene* 2006, **25:**1594-1601.
12. Sobrino B, Brion M, Carracedo A: **SNPs in forensic genetics: a review on SNP typing methodologies.** *Forensic Sci Int* 2005, **154:**181-194.
13. Syvanen AC: **Accessing genetic variation: genotyping single nucleotide polymorphisms.** *Nat Rev Genet* 2001, **2:**930-942.
14. Gabriel SB, Schaffner SF, Nguyen H, Moore JM, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, *et al*: **The structure of haplotype blocks in the human genome.** *Science* 2002, **296:**2225-2229.
15. Terwilliger JD, Hiekkalinna T: **An utter refutation of the "Fundamental Theorem of the HapMap".** *Eur J Hum Genet* 2006, **14:** 426-437.
16. Stram DO: **Software for tag single nucleotide polymorphism selection.** *Hum Genomics* 2005, **2:**144-151.
17. Ke X, Miretti MM, Broxholme J, Hunt S, Beck S, Bentley DR, Deloukas P, Cardon LR: **A comparison of tagging methods and their tagging space.** *Hum Mol Genet* 2005, **14:**2757-2767.
18. Adkins RM: **Comparison of the accuracy of methods of computational haplotype inference using a large empirical dataset.** *BMC Genet* 2004, **5:**22.
19. Clayton D, Chapman J, Cooper J: **Use of unphased multilocus genotype data in indirect association studies.** *Genet Epidemiol* 2004, **27:**415-428.
20. Haiman CA, Stram DO, Pike MC, Kolonel LN, Burtt NP, Altshuler D, Hirschhorn J, Henderson BE: **A comprehensive haplotype analysis of CYP19 and breast cancer risk: the Multiethnic Cohort.** *Hum Mol Genet* 2003, **12:**2679-2692.
21. Feigelson HS, Cox DG, Cann HM, Wacholder S, Kaaks R, Henderson BE, Albanes D, Altshuler D, Berglund G, Berrino F, *et al*: **Haplotype analysis of the HSD17B1 gene and risk of breast cancer: A comprehensive approach to multicenter analyses of prospective cohort studies.** *Cancer Res* 2006, **66:**2468-2475.
22. Benusiglio PR, Lesueur F, Luccarini C, McIntosh J, Luben RN, Smith P, Dunning A, Easton DF, Ponder BA, Pharoah PD: **Common variation in EMSY and risk of breast and ovarian**

cancer: a case-control study using HapMap tagging SNPs. *BMC Cancer* 2005, **5:**81.
23. Einarsdottir K, Humphreys K, Bonnard C, Palmgren J, Iles M, Sjolander A, Li Y, Chia KS, Liu E, Hall P, *et al*: **Linkage disequilibrium mapping of CHEK2: Common variation and breast cancer risk.** *PLOS Medicine* 2006, in press.
24. Hirschhorn JN, Daly MJ: **Genome-wide association studies for common diseases and complex traits.** *Nat Rev Genet* 2005, **6:** 95-108.
25. de Bakker PI, Yelensky R, Pe'er I, Gabriel SB, Daly MJ, Altshuler D: **Efficiency and power in genetic association studies.** *Nat Genet* 2005, **37:**1217-1223.
26. Marchini J, Donnelly P, Cardon LR: **Genome-wide strategies for detecting multiple loci that influence complex diseases.** *Nat Genet* 2005, **37:**413-417.
27. Ioannidis JP, Gwinn M, Little J, Higgins JP, Bernstein JL, Boffetta P, Bondy M, Bray MS, Brenchley PE, Buffler PA, *et al*: **A road map for efficient and reliable human genome epidemiology.** *Nat Genet* 2006, **38:**3-5.
28. Colditz GA, Hankinson SE: **The Nurses' Health Study: lifestyle and health among women.** *Nat Rev Cancer* 2005, **5:**388-396.
29. Coyle YM: **The effect of environment on breast cancer risk.** *Breast Cancer Res Treat* 2004, **84:**273-288.
30. Tsubura A, Uehara N, Kiyozuka Y, Shikata N: **Dietary factors modifying breast cancer risk and relation to time of intake.** *J Mammary Gland Biol Neoplasia* 2005, **10:**87-100.
31. Han J, Hankinson SE, Ranu H, De Vivo I, Hunter DJ: **Polymorphisms in DNA double-strand break repair genes and breast cancer risk in the Nurses' Health Study.** *Carcinogenesis* 2004, **25:**189-195.
32. Han J, Hankinson SE, Zhang SM, De Vivo I, Hunter DJ: **Interaction between genetic variations in DNA repair genes and plasma folate on breast cancer risk.** *Cancer Epidemiol Biomarkers Prev* 2004, **13:**520-524.