Letter to the Editor

SNPs associated with molecular subtypes of breast cancer: On the usefulness of stratified Genome-wide Association Studies (GWAS) in the identification of novel susceptibility loci

Recent developments in whole-genome Single Nucleotide Polymorphism (SNP) analysis have spurred a number of Genome-Wide Association Studies (GWAS) in breast, prostate, and up-coming testicular and bowel cancer (Easton et al., 2007; Eeles et al., 2008). Until now the genetic variants that have been revealed by these studies are responsible for comparatively small increases in relative risk of disease and it is unlikely that highly predictive genetic variants will emerge from subsequent studies. This may be seen by those arguing that a change in strategy is called for as evidence of the weakness of these studies. But is this a fair inference? Hardly any higher relative risk and more penetrant alleles should have been expected from these studies in the first place; had these alleles existed they would have been found through the extensive linkage studies of the past 20 years in family studies of sufficient size (Oldenburg et al., 2007; Balmain et al., 2003). This holds true even if one takes into consideration the multi-stage design of GWAS studies that is aimed at "amplifying" the hereditary effect by using individuals with a highly familial component of the disease as a starting point. Experiments such as the whole-genome association analyses funded by Cancer Research UK (Easton et al., 2007; Eeles et al., 2008) constitute a brave and important exploratory step. It is the kind of experiment one has little chance of knowing whether it is worth doing before it is actually done. The experiments that have already been made represent an unprecedented effort invaluable for the scientific community given the amount of money and the effort made by the projects' principle investigators to bring together so many research groups in order to collect sufficient number of patients for such large multistage studies.

The accumulation of a critical mass of data is of crucial importance. Large amounts of data have been generated already, the money is well spent, and it is of huge importance that this data is available to the scientific community so that it can be analysed in various novel ways. There are two alternative ways in which we can proceed from here: we can either (a) increase the power of these studies by including more and more samples in both the case and the control groups or (b) we can elaborate on the data analysis concentrating on clinical and molecular subtypes for each of the diseases studied. The enlarged GWAS studies are bound to confirm the finding(s) of the original studies, however, it is doubtful whether they will reveal novel susceptibility loci, particularly if the data are analysed in the same way as before. The main focus of data analysis in GWAS studies so far has been on reproducibility of findings using classical hazard ratio analysis, which means that markers with extremely low p-values have been favoured. This requirement for certainty when studying the effects of 300,000 and 500,000 putative factors for susceptibility (corresponding to the size of the SNP arrays used) has caused investigators to report only on susceptibility alleles of universal impact for all individuals, rather than high penetrance alleles in a discreet subgroup of patients, which remain insignificant in the overall pool. Using this type of statistical analysis on GWAS studies can be compared to attempting to describe quantum mechanics using tools from classical Newtonian physics. Therefore we strongly advocate proceeding in the alternative direction: first identifying discreet molecular and clinical subtypes for each disease and subsequently elaborating on the data analysis including probability models, pattern recognition and Bayesian neuronal networking models. Our personal experience is that there is a common scepticism towards these approaches rooted on a "what you do not see cannot be there" attitude, and that it may be rather difficult to publish work that employs these approaches.

A further criticism of the rationale behind GWAS studies is based on the argument that inferring an individual's cancer risk from a limited number of alleles is probably a flawed proposition. The opponents argue that given that the majority of such genes will be inherited randomly, the predisposing alleles will tend to be balanced out by those that do not predispose to the condition (i.e. 'the "good" alleles will tend to balance out the "bad" ones'). Therefore the distribution/shape

of the risk profile would tend to be so narrow as to be fairly ineffective for the great majority of people. This is an argument we disagree with. Indeed, it is unlikely that common susceptibility markers will be found for the total population or ''the great majority of people'', but it is of equal health benefit to identify the susceptibility markers for different subpopulations at risk. That the ''good'' and the ''bad'' alleles would balance each other does not mean that there is no point in identifying them by well designed epidemiological studies. It is necessary to perform stratified analysis and identify markers specific for groups of individuals at risk, sharing common familial background or developing common cancer phenotype. This can be illustrated with our own studies of breast cancer. Recently, GWAS analysis of breast cancer revealed SNPs in five novel genes associated to susceptibility: *TNRC9*, *FGFR2*, *MAP3K1*, *H19* and *LSP1* (Easton et al., 2007). The results were confirmed for *FGFR2* and *TNRC9* in two independent studies (Hunter et al., 2007; Stacey et al., 2007). While there is little doubt that the novel susceptibility markers identified by such highly powered studies are genuine, the mechanism by which they cause the susceptibility remains unravelled. Pooling of such a large amount of cases leads inevitably to conceal the various histological and clinico-pathological subtypes, suggesting that the observed genes are either of universal importance for breast cancer development, or are associated with a subgroup which dominates the overall pool or with any subgroup but with an association sufficiently strong to dominate the overall result. Breast cancer patients can be divided into five distinct molecular subtypes based on their expression profiles (Perou et al., 2000). The existence of these five subgroups, Luminal A, Luminal B, Basal-like, ErbB2[+], and Normal-like, has been confirmed in independent data sets (Sorlie et al., 2003) and they are associated with different clinical outcome (Sorlie et al., 2001). If the probability to develop a given subclass of breast cancer is genetically determined we might expect to find that the newly discovered susceptibility genes (Easton et al., 2007) are differentially expressed in the various tumor subclasses, and that their transcription is regulated *in cis* by SNPs within them. With this in mind, we retrieved the mRNA expression data of *TNRC9*, *FGFR2*, *MAP3K1*, *H19* and *LSP1* from 112 breast tumors representing all five subclasses (Sorlie et al., 2003). Recently we reported that significantly different mRNA levels between the subtypes were found for all the five genes by ANOVA analysis (Nordgard et al., 2007).

Further understanding of the molecular mechanisms underlying the tumor subclassification and the role of these novel susceptibility genes play in developing these subtypes is necessary. One known strong factor underlying the subclassification is the estrogen receptor (ER). The relationship between the ER and its ligand, oestradiol, and the enzymes that synthesize it is not well understood. In a previous study we found that the expression of a set of nine mRNA transcripts of members of the oestradiol metabolic and signalling pathways, including CYP19 (aromatase), was significantly different in the five molecular subtypes (Kristensen et al., 2005). We have previously reported a SNP in the 3′-untranslated area of CYP19 mRNA to be associated to higher mRNA levels and tumor size in locally advanced breast cancers (stage III/IV) (Kristensen et al., 2000). The distribution of the different alleles of this SNP is also significantly different in the five molecular subtypes (Fig. 1). Since the induction and promotion of carcinogenesis is attributed to ER receptor mediated stimulation of proliferation of pre-malignant breast epithelial cells, the amount of produced oestrogen and its availability at the time of this stimulation is likely to be of importance for the further development of the disease, and may ultimately reflect in the tumor classes observed by gene expression profiling. This illustrates the necessity to conduct stratified SNP-disease association studies and stratification of patients by their molecular subtypes may give much more power to the classical case control studies, and genes of no or borderline overall significance may be highly penetrant for certain subclasses, and therefore identifiable.
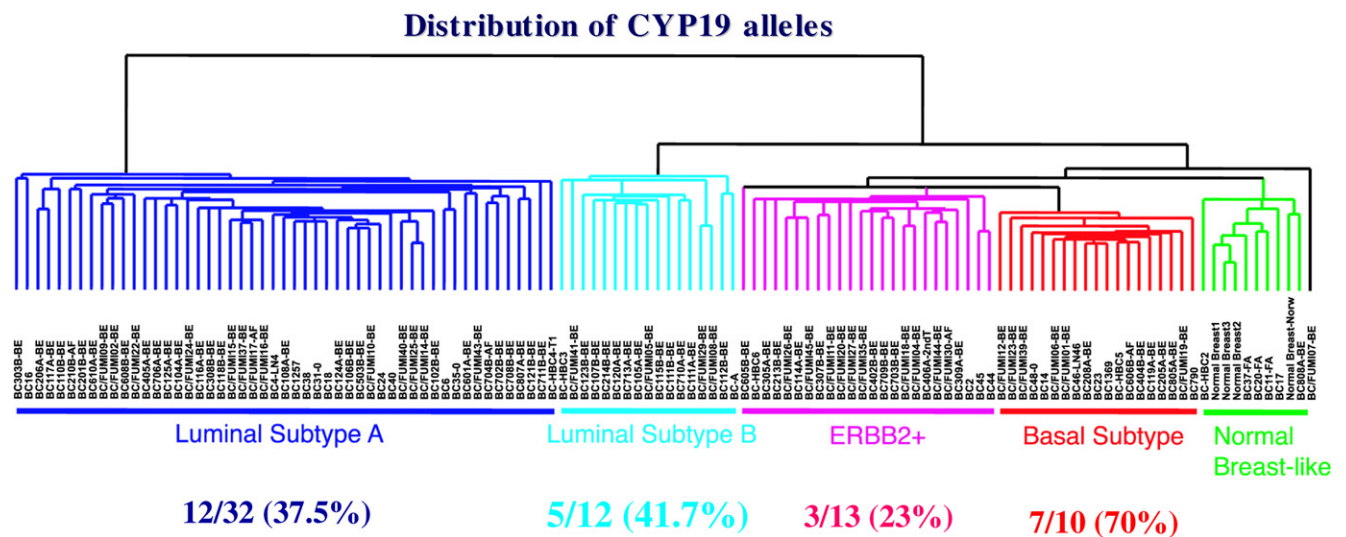

Figure 1 – Distribution of the TT genotypes (homozygous for the variant allele) in the molecular subclasses of breast cancer defined by whole-genome mRNA expression. TT homozygotes are significantly more frequent in patients developing basal-like tumors ($p = 0.05$).