Human Molecular Genetics, 2010, Vol. 19, No. 12 **2507–2515**
doi:10.1093/hmg/ddq122
Advance Access published on March 23, 2010

Fine scale mapping of the breast cancer 16q12 locus

Miriam S. Udler[1,3,4], Shahana Ahmed[2,‡], Catherine S. Healey[2,‡], Kerstin Meyer[4,‡], Jeffrey Struewing[5,†], Melanie Maranian[2], Erika M. Kwon[3], Jinghui Zhang[5], Jonathan Tyrer[2], Eric Karlins[3], Radka Platte[2], Bolot Kalmyrzaev[2], Ed Dicks[2], Helen Field[2], Ana-Teresa Maia[4], Radhika Prathalingam[4], Andrew Teschendorff[4,6], Stewart McArthur[4], David R. Doody[7], Robert Luben[1], Carlos Caldas[4], Leslie Bernstein[8], Laurence K. Kolonel[9], Brian E. Henderson[10], Anna H. Wu[10], Loic Le Marchand[9], Giske Ursin[10,11], Michael F. Press[12], Annika Lindblom[13], Sara Margolin[13], Chen-Yang Shen[14], Show-Lin Yang[14], Chia-Ni Hsiung[14], Daehee Kang[15], Keun-Young Yoo[15], Dong-Young Noh[15], Sei-Hyun Ahn[16], Kathleen E. Malone[7], Christopher A. Haiman[10], Paul D. Pharoah[2], Bruce A.J. Ponder[2,4], Elaine A. Ostrander[3], Douglas F. Easton[1] and Alison M. Dunning[2,*]

[1]Department of Public Health and Primary Care and [2]Department of Oncology, University of Cambridge, Cambridge, UK, [3]Cancer Genetics Branch, NHGRI/NIH, Bethesda, MD, USA, [4]CRUK Cambridge Research Institute, Li Ka Shing Centre, Cambridge CB2 0RE, UK, [5]Laboratory of Population Genetics, National Cancer Institute/NIH, Bethesda, MD, USA, [6]Medical Genomics Group, UCL Cancer Institute, University College London, London, UK, [7]Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA, [8]Department of Population Sciences, City of Hope National Medical Center, Duarte, CA, USA, [9]Epidemiology Program, Cancer Research Center of Hawaii, University of Hawaii, Honolulu, HI, USA, [10]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA, USA, [11]Department of Nutrition, University of Oslo, Oslo, Norway, [12]Department of Pathology, Keck School of Medicine, University of Southern California/Norris Comprehensive Cancer Center, Los Angeles, CA, USA, [13]Department of Molecular Medicine and Surgery, Karolinska Institutet, Stockholm, Sweden, [14]Institute of Biomedical Sciences, Academia Sinica, Taipei, Taiwan, [15]Seoul National University College of Medicine, Seoul, Korea and [16]Ulsan University, Seoul, Korea

Received November 16, 2009; Revised March 11, 2010; Accepted March 17, 2010

Recent genome-wide association studies have identified a breast cancer susceptibility locus on 16q12 with an unknown biological basis. We used a set of single nucleotide polymorphism (SNP) markers to generate a fine-scale map and narrowed the region of association to a 133 kb DNA segment containing the largely uncharacterized hypothetical gene *LOC643714*, a short intergenic region and the 5′ end of *TOX3*. Re-sequencing this segment in European subjects identified 293 common polymorphisms, including a set of 26 highly correlated candidate causal variants. By evaluation of these SNPs in five breast cancer case–control studies involving more than 23 000 subjects from populations of European and Southeast Asian ancestry, all but 14 variants could be excluded at odds of <1:100. Most of the remaining variants lie in the intergenic region, which exhibits evolutionary conservation and open chromatin conformation, consistent with a regulatory function. African-American case–control studies exhibit a different pattern of association suggestive of an additional causative variant.

*To whom correspondence should be addressed at: Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK. Tel: +44 1223 740683; Fax: +44 1223 740147; Email: alison.dunning@srl.cam.ac.uk
†Present address: Division of Extramural Research, NIH/NHGRI, Bethesda, MD, USA.
‡Contributed equally to this manuscript.

INTRODUCTION

In our initial genome-wide association study (GWAS) (1) two single nucleotide polymorphisms (SNPs), rs12443621 and rs8051542, in *TOX3* were significantly associated with increased risk of breast cancer. *TOX3* (also called *TNRC9* or *CAGF9*) encodes a high mobility group box nuclear protein, involved in mediating calcium-dependent transcription (3). Increased expression of *TOX3* has been reported to predict breast cancer metastasis to bone (4). Frequent loss of heterozygosity of the chromosome 16q arm is observed in breast tumours; however, the location of the critical region of loss, containing a putative tumour suppressor gene, remains undefined (5). A second, independent GWAS (2), using a different SNP set, also found a significant association with a correlated SNP rs3803662, within *LOC643714*. This work presents the results of a strategy to identify the causative variant directly responsible for the observed associations. Towards this end we have pursued fine-scale mapping of the region using case–control studies of European, East and Southeast Asian and African-American descent. In addition, we have sought to determine whether candidate SNPs reside in regions of open chromatin conformation or are associated with differences in expression of genes close to the locus. Candidate SNPs were further evaluated with *in silico* examination of evolutionary sequence conservation and putative transcription factor binding sites.

RESULTS

The initial GWAS (1,2) identified three SNPs (rs8051542, rs12443621 and rs3803662) in 16q12 associated with increased risk of breast cancer. These SNPs resided in a 133 kb linkage disequilibrium (LD) block containing the 5′ end of *TOX3* gene and the entire hypothetical gene *LOC643714*. Initial refinement was performed on a larger 170 kb region which included this 133 kb block as well as two additional LD blocks that together contained the remainder of the *TOX3* gene. The LD blocks were delimited by inspection of D' plots using data from the CEU population of the International HapMap database (6). Nineteen SNPs were chosen to tag the 101 common SNPs in this region listed on the International HapMap database ($r^2 > 0.8$) using Haploview (7) (Fig. 1A). Eleven of these tags showed no association with breast cancer in 2165 cases and 2278 controls from the European SEARCH case–control study (Table 1, Supplementary Material, Table S1) leaving eight significantly associated SNPs (P-trend < 0.05) which tagged the 133 kb region (Fig. 1A). The strongest association was observed with tag SNP rs3803662, and none of the other tags, including the original GWAS hits, maintained a significant association after adjusting for this SNP. Analysis of rs3803662 and the other two SNPs identified in the GWAS, rs12443621 and rs8051542, in 21 860 cases and 22 578 controls by the Breast Cancer Association Consortium also showed that only rs3803662 was independently significant. Furthermore, haplotype analysis of rs3803662 with nine correlated SNPs ($D' \geq 0.5$) revealed multiple haplotypes, all carrying the minor allele of rs3803662 and all associated with an increased risk of breast cancer (Supplementary Material, Table S2). Taken together these analyses suggest strongly that the association is mediated through a causative variant, strongly correlated with SNP rs3803662, and thus common in European subjects.

The 133 kb region was re-sequenced in 42–45 individuals of European ancestry from the Centre d'Etude du Polymorphisme Humain (CEPH) collection (https://cgwb.nci.nih.gov/). No structural rearrangements have been identified in this interval. Four hundred and twenty-three variants were identified, of which 245 had minor allele frequency (MAF) > 0.05. Twenty-five variants were well correlated ($r^2 > 0.5$) with the best tag SNP, rs3803662 and could thus be considered strong candidates for being the causative variant (Table 2, Supplementary Material, Table S3). Recently, the 1000 Genomes Project (http://www.1000genomes.org) released sequence data from 57 CEPH individuals. A comparison of the SNPs identified by this project and our own work is shown in Supplementary Material, Table S4. The 1000 Genomes Project identified 108 novel variants in this 133 kb region, bringing the total number of variants to 531, of which 293 had MAF ≥ 0.05. One of these newly discovered variants was also well correlated ($r^2 > 0.5$) with our best tag SNP, rs3803662, increasing the number of candidate causative variants to 26 (Fig. 1B, Table 2). All other confirmed SNPs identified by the 1000 Genomes Project which were correlated with rs3803662 ($0.2 \leq r^2 < 0.5$) were better tagged by one of our other tagSNPs (which were no longer significant after adjusting for rs3803662, as discussed above) and thus were not investigated further.

We hypothesized that the weaker LD between candidate causative variants in Asian and African-American populations (Fig. 2, Supplementary Material, Table S5) would increase the power to eliminate candidates, thereby improving resolution to locate causative alleles. Thus, we aimed to genotype these 26 candidates in 27 578 subjects from case–control studies of European, Asian and African-American ancestry (Supplementary Material, Fig. S1, Table S8 and Methods). Four of the variants could not be genotyped using high-throughput techniques, and so genotypes for these were determined by bidirectional sequencing of a subset of subjects, followed by imputation in the remaining European and Asian subjects (see Materials and Methods).

As anticipated, all 26 SNPs were significantly associated with breast cancer in the European studies (P-value $< 10^{-8}$). Sixteen of these 26 (SNPs A, B, G, J, L-Q, S-U, W-Z) were also significantly associated (P-value < 0.05) in the Asian studies and allelic risks were in the same direction (Supplementary Material, Fig. S1). In the African-American studies, however, the significantly associated SNPs (P-value < 0.05) had effects in the opposite direction—the risk allele in Europeans and Asians was protective in African-Americans and vice-versa (Supplementary Material, Fig. S1). A similar phenomenon was described in Stacey *et al*. (2) for SNP rs3803662 in a subset of the African-American samples from the Multiethnic Cohort (MEC) Study, also included in this present study. The opposite direction of effect for SNPs in African-Americans was further explored by haplotype analysis of the genotyped candidate SNPs in the three ethnic groups (Table 3). In Europeans there are just two common haplotypes: haplotype #1, containing all the non-risk associated alleles, and haplotype #2, containing all the risk alleles, which has 1.3-fold increased risk relative to haplotype #1. Asians exhibit the same

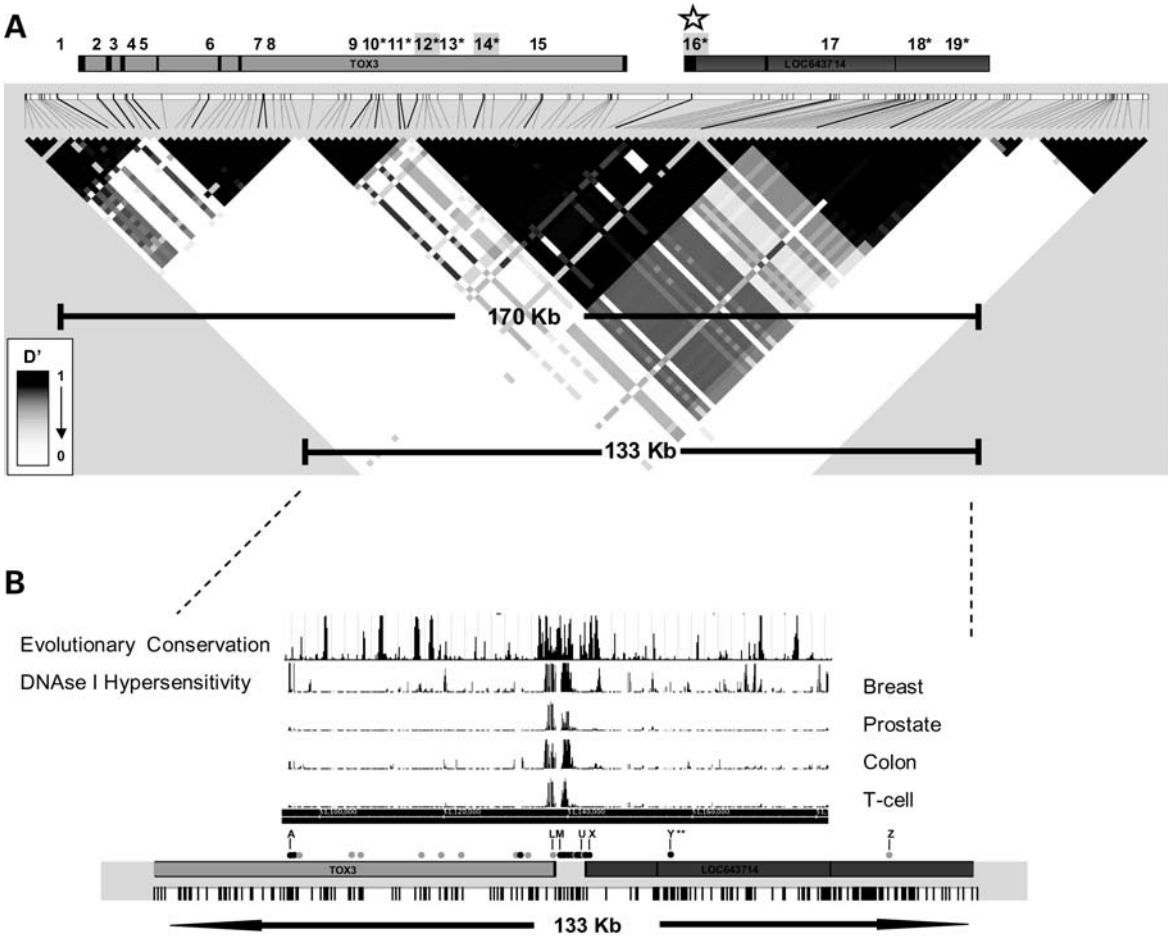

Figure 1. (**A**) LD blocks (D' in greyscale) showing the 19 variants selected to tag *TOX3* and *LOC643714*. Highlighted SNPs are those identified in GWAS by Easton *et al*. (1) and Stacey *et al*. (2). TagSNPs significantly associated with breast cancer risk are marked with asterisks. The most strongly associated SNP (16, rs3803662) is indicated by the star. After adjustment for rs3803662, associations of all other SNPs were non-significant (Refer to Table 1 for corresponding rs numbers). The narrowed 133 kb region was re-sequenced in individuals of European ancestry. (**B**) Re-sequenced 133 kb region showing the 26 variants (dots) strongly correlated with rs3803662. Black dots indicate those remaining as potential causative variants. Grey dots indicate those excluded (SNPs coded as A—Z from left to right. Refer to Table 2 for corresponding rs numbers). The most significant SNP (Y, rs4784227) from the combined analysis of European and Asian studies is marked with two asterisks. Evolutionary conservation of 17 placental mammalian species was extracted from the UCSC genome browser. DNase I hypersensitivity profiles for breast cancer (MCF-7, PMC42 and MDA231), prostate (LnCapC4b and RWPE-1), colon cancer (HCT116) and T-cells indicate DNA in open chromatin conformation by vertical lines. Gaps mark regions that could not be tiled due to repetitive sequences.

two haplotypes, with similar risks, as well as two further common haplotypes, #5 and #6. In contrast in African-Americans, haplotype #2, which is relatively uncommon, appears to be associated with similar risk to haplotype #1, although the confidence limits on this are wide. The second most common African-American haplotype is #5, which is associated with a clear reduction in risk relative to #1. Taken together, these results are consistent with the presence of a single, common causative variant in both the European and Asian populations, but suggest a different pattern of association in African-Americans. This different pattern in African-Americans meant these data could not be used to refine the mapping of the putative causative variant in Europeans and Asians.

On the basis of the assumption that there is a single disease-causing allele in European and Asian subjects, the likelihood of each candidate being causative was estimated (Table 2, Supplementary Material, Table S3). We computed a likelihood ratio for each candidate compared with the most strongly associated SNP rs4784227. Twelve of the variants had likelihood ratios 100-fold worse than rs4784227and so could reasonably be excluded from further consideration. The 14 remaining SNPs span three distinct, potentially functional, genetic elements. Three candidates (A, B and J) reside within intron 1 of *TOX3*; eight (M, N, O, P, Q, S, T and U) are clustered within a 3 kb segment of the intergenic region; and the remaining three lie within *LOC643714*: (W) is a synonymous change in a Ser residue encoded by putative exon 4, (X) is in putative intron 3 and the current strongest candidate (Y, rs4784227) is in putative intron 2 (Fig. 1B and Table 2).

We have used several approaches to further elucidate the likely functionality of the remaining candidates. We examined the chromatin conformation around each candidate using hypersensitivity to DNase I digestion in breast, prostate and colon cell lines as well as in primary human T-cells. There is a region of open chromatin conformation in the

Table 1. Breast cancer associations of 19 SNPs tagging TOX3 and LOC643714

SNP	rs number	Per allele OR (95% CI)	*P*-trend
1	rs7188855	1.00 (0.92–1.08)	9×10^{-1}
2	rs1116205	1.02 (0.99–1.11)	7×10^{-1}
3	rs2075236	1.02 (0.99–1.12)	7×10^{-1}
4	rs1420542	0.99 (0.89–1.10)	9×10^{-1}
5	rs3095611	1.00 (0.92–1.10)	1.0
6	rs11647305	1.00 (0.92–1.09)	1.0
7	rs4784217	1.03 (0.95–1.12)	5×10^{-1}
8	rs16951204	0.96 (0.85–1.08)	5×10^{-1}
9	rs7188610	1.07 (0.95–1.20)	3×10^{-1}
10	rs9926539	0.86 (0.78–0.94)	9×10^{-4}
11	rs1111481	0.91 (0.84–0.99)	3×10^{-2}
12	rs8051542	1.15 (1.05–1.25)	2×10^{-3}
13	rs4784220	1.16 (1.07–1.27)	6×10^{-4}
14	rs12443621	1.16 (1.06–1.26)	7×10^{-4}
15	rs1420531	0.92 (0.83–1.02)	1×10^{-1}
16	**rs3803662**	**1.30 (1.18–1.42)**	$\mathbf{4 \times 10^{-8}}$
17	rs9708611	1.15 (0.97–1.36)	1×10^{-1}
18	rs3112623	1.12 (1.03–1.22)	7×10^{-3}
19	rs12922061	1.21 (1.10–1.33)	2.0×10^{-4}

All SNPs were genotyped in 2165 cases and 2278 controls from the SEARCH study (Set 01). See Supplementary Material, Table S1 for full data. The most significant association is shown in bold.

non-coding region between *TOX3* and *LOC643714* (Fig. 1B and Supplementary Material, Table S3), and examination of this region at higher resolution indicates that five candidates (SNPs M, N, O, P and Q) are in highly accessible chromatin (Supplementary Material, Fig. S2). The DNA sequence of this intergenic region is also highly conserved across mammalian species (Fig. 1B). Furthermore bioinformatic analysis using PReMod (http://genomequebec.mcgill.ca/PReMod/) identified two regulatory modules overlapping three of our candidates (SNPs M, N & O, Supplementary Material, Fig. S2) in open chromatin conformation as well as a third overlying SNPs W and X, but not in open chromatin. We also used multiple search algorithms to predict transcription factor binding sites containing the candidate causative SNPs (see Materials and Methods). Three of the predicted transcription factors binding to sites overlying SNPs N, O and Q also have the potential to interact with transcription factors predicted to be part of the two PReMod regulatory modules in open chromatin conformation (Supplementary Material, Table S6). The risk associated (T) allele of our current strongest candidate after genetic mapping, SNP Y (rs4784227), is predicted to create a C/EBP alpha binding site; however, this SNP resides in closed chromatin conformation.

A plausible hypothesis from the above findings is that the causative variant(s) may regulate gene expression. We first explored the possibility that the risk associated SNPs may alter expression levels of either *TOX3* or *LOC643714*. *TOX3* is expressed in both normal breast and breast tumour cells at similar levels (data not shown). We therefore examined the association between tag SNP rs3803662 genotype and *TOX3* mRNA levels in 38 normal breast samples and in 77 breast tumours, and found no significant associations (regression *P*-trend = 0.83 and 0.66, respectively). Two predicted transcripts of *LOC643714* (Ensembl:ENSESTT00000054674 and ENSESTT00000054675) have been detected, at very low levels, in the breast cancer cell line SUM190PT (data not shown), but levels were negligible in both normal breast and breast tumours, so that similar association studies with LOC643714 mRNA were not possible. Thus, as yet, we have no convincing evidence that either of these genes are regulated by the putative breast cancer association variant within this locus (Table 4).

Working on the principle that regulatory variants may alter expression of distant genes in *cis* (8), we tested whether the 16q12 locus altered expression of surrounding genes (both confirmed and hypothetical) using expression data, where available, from lymphocytes (8,9) and breast tumours (Supplementary Material, Table S7). Expression of both *TOX3* and *LOC643714* is negligible in lymphocytes so could not be reliably assessed using these data. Of the 11 genes lying within 1 Mb of SNP rs3803662 with appropriate expression data, significant association with genotype was observed only with mRNA levels from *RBL2* (Retinoblastoma-like gene 2, Supplementary Material, Fig. S3 and Table S7). Dose-dependent associations of the breast cancer risk allele with increasing levels of *RBL2* mRNA were observed in lymphocytes from 210 HapMap subjects (*P*-trend = 0.01). TagSNPs across the 16q12 locus show moderate correlations between their associations with breast cancer risk and associations with RBL2 expression in HapMap samples (Supplementary Material, Fig. S4 and Table S8). However, no similar significant association of rs3803662 and RBL2 levels was observed in 77 breast tumours (*P*-value = 0.8) although this tumour set had limited power to detect such an association. *RBL2* is a member of the retinoblastoma (Rb) gene family (10) is involved in cell cycle regulation and is frequently deleted in breast tumours (11).

DISCUSSION

This study serves to illustrate the complexities of identifying causal disease-susceptibility variants, even within a locus with very clear evidence of association. Using genetic epidemiology, we have been able to reduce the 293 common variants (MAF $\geq$ 0.05) found by re-sequencing to 14 strong candidates. Larger Asian case–control studies, when available, may eliminate more. However, four of these candidates (N, P, T and U) are too strongly correlated ($r^2 > 0.96$) in both European and Asian studies, to be eliminated by epidemiological studies. The African-American data, with a different pattern of association, further adds to this complexity. It also remains possible that the causal variant we are seeking was not detected during re-sequencing.

The pattern of association in African-Americans, markedly different from that in Europeans and Asians, is puzzling. It is possible that the observed inverted allelic effect is a chance finding due to a lack of power. Indeed, the African-American studies are the smallest studies utilized in this analysis, and the ORs for SNPs differed across studies, as may be seen in Supplementary Material, Figure S1. Furthermore, African-Americans are of mixed ethnicity, and it has not been possible to assess the ancestral composition of the study subjects.

Table 2. Likelihood ratios for 26 variants identified as candidate causative variants

	rs number	Alleles[a]	r^2 in CEPH individuals[b]	Log[10] likelihood ratio[c] Europeans	Asians	Combined likelihood ratio[c]
A	**rs17271951**	**T/C**	**0.89**	**0.96**	**0.58**	**35**
B[d]	**rs35668161**	**C/A**	**0.81**	**1.02**	**0.36**	**25**
C[d]	rs12600239	C/T	0.86	0.29	3.17	2863
D[d]	rs7500427	G/A	0.87	0.17	3.28	2811
E	rs9936081	G/A	0.85	−0.91	3.58	471
F	rs1345388	T/C	0.86	−0.79	3.67	765
G	rs12918816	G/A	0.85	−0.92	3.12	159
H	rs1362548	G/C	0.89	−0.37	3.28	809
I	rs9921569	T/C	0.93	−0.68	3.22	346
J	**rs35850695**	**G/A**	**0.90**	**0.48**	**−0.14**	**2**
K	rs4784223	A/G	0.90	−0.30	3.22	835
L	rs8045285	A/G	1.0	−0.28	3.07	610
M	**rs12930156**	**C/T**	**0.74**	**−0.63**	**1.36**	**5**
N	**rs3095604**	**G/C**	**0.86**	**−1.17**	**1.54**	**2**
O[d]	**rs45538731**[e]	**16/15bps**	**1.0**	**−0.88**	**1.39**	**3**
P	**rs28463809**	**G/T**	**1.0**	**−1.26**	**1.56**	**2**
Q	**rs4784226**	**C/T**	**0.60**	**0.42**	**−0.24**	**2**
R	rs45465998	G/A	0.31[f]	−0.97	3.30	215
S	**rs45482301**	**−/A**	**0.67**	**−1.12**	**2.96**	**68**
T	**rs3095606**	**A/G**	**0.95**	**−0.86**	**1.31**	**3**
U	**rs3095607**	**T/G**	**0.95**	**−0.84**	**1.29**	**3**
V	rs3112578	A/G	0.66	−0.83	3.23	251
W	**rs3803662**	**C/T**	**1.0**	**−0.19**	**1.55**	**23**
X	**rs3803661**	**G/A**	**1.0**[g]	**−0.81**	**1.46**	**4**
Y	**rs4784227**	**C/T**	**0.68**	**0.00**	**0.00**	**1**
Z	rs12922061	C/T	0.45	6.34	−0.37	943 798

[a]Protective allele/risk allele in Europeans; G/C SNPs are listed for forward strand.
[b]r^2 for each SNP calculated with rs3803662 from re-sequencing data of 42–45 CEPH individuals.
[c]Log[10] likelihood ratio for each candidate compared with the most strongly associated SNP rs4784227, see Materials and Methods. Analyses stratified by two European and six Asian study populations. Variants <100-fold less significant than rs4784227, and thus still candidates for causative variant, are shown in bold.
[d]SNPs genotyped by bidirectional sequencing in a subset of individuals and imputed.
[e]Alternative rs numbers: rs45492607 and rs1362549.
[f]SNP rs45465998 was included, despite being difficult to genotype, because rare homozygotes were shared with rs3803662.
[g]r^2 value for SNP rs3803661 calculated from 57 CEPH individuals in the 1000 Genomes Project.

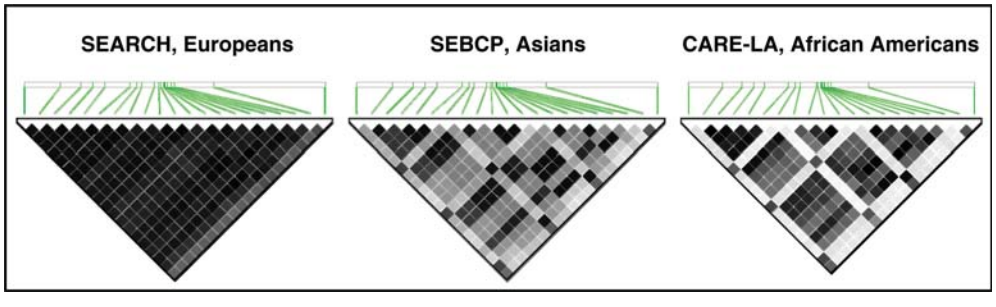

Figure 2. Haplotype blocks of candidate SNPs genotyped by TaqMan in three study populations. Twenty-one SNPs all highly associated with breast cancer risk were genotyped in the SEARCH UK study as well as in the Seoul Breast Cancer Project. All but rs8045285 were genotyped in the CARE study, with the data from African-Americans within the Los Angeles study centre depicted here. See Supplementary Material, Table S5 for a complete list of r^2 values for the variants in each population.

It is thus possible that admixture has influenced these results, causing SNP allele frequencies in cases to differ from those in controls simply because the proportions of European and African ancestries differ between cases and controls. Our findings could also be explained by the existence of an additional risk variant, carried on a subset of haplotype #1 in the African-American population in addition to a variant shared by all populations. This possibility is supported by the fact that four of the remaining variants (A, J, Q and Y), including the current strongest candidate (Y, rs4784227), share the same risk allele in all three ethnic groups, although these associations did not reach statistical significance in the African-American case–control studies. Therefore, analysis of additional African and African-American studies will be necessary to clarify our findings. If these findings can be replicated, further re-sequencing and genotyping in African-American studies will be required to determine whether an additional risk variant underlies the different pattern of association.

Table 3. Haplotype-specific breast cancer risks by ethnic group

SEARCH Europeans	Haplotype frequency Controls	Cases	OR (95% CI)	*P*-value
1. 00000000**00000000000000**	0.71	0.65	–	–
2. 11111111**11111111111111**	0.17	0.22	1.31 (1.21–1.41)	1×10^{-12}
3. 11111111**11111111111110**	0.03	0.04	1.30 (1.10–1.53)	0.002
4. 00000000**00000000000001**	0.02	0.02	0.95 (0.75–1.20)	0.7
11111111111000110110**0**	0.01	0.02	1.21 (0.95–1.55)	0.1
Rare pooled			1.27 (1.11–1.47)	0.0007
Asians[a]				
1. **0**0000**0**0000**0**0000000**0**	0.37	0.34	–	–
2. **1**1111**1**1111**11**1111111**1**	0.18	0.22	1.26 (1.15–1.38)	1×10^{-6}
5. **0**1111**0**1111**0**1111111**0**	0.19	0.18	0.98 (0.89–1.08)	0.7
6. **0**0000**0**0111**00**0110**1**0**0**	0.13	0.14	1.14 (1.03–1.27)	0.01
4. **0**0000**0**0000**00**0000000**1**	0.03	0.02	0.95 (0.75–1.21)	0.7
00011**1**1111**11**111111**1**11	0.03	0.03	1.14 (0.92–1.40)	0.2
00011**1**1111**11**111111**1**10	0.02	0.02	1.11 (0.82–1.51)	0.5
3. **1**1111**1**1111**11**11111**1**0	0.01	0.01	0.89 (0.67–1.19)	0.4
0000000011**00**01101**1**01	0.01	0.01	1.47 (0.97–2.23)	0.07
Rare pooled			0.85 (0.75–0.97)	0.01
African-Americans[b]				
1. **0**0000000**00000000**0_**0**0	0.45	0.49	–	–
5. 01111**0**1111**0**111111_**0**0	0.30	0.26	0.79 (0.69–0.91)	0.0008
2. 11111111**11111**111111_**1**1	0.07	0.07	0.96 (0.75–1.23)	0.7
000000000110000101_00	0.05	0.04	0.77 (0.57–1.04)	0.09
000000100000000000_00	0.02	0.03	1.39 (0.91–2.11)	0.1
000010001110001101_00	0.02	0.02	0.98 (0.65–1.49)	0.9
6. 00000000**1110001**101_**0**0	0.02	0.02	1.34 (0.86–2.09)	0.2
000000001110000101_00	0.01	0.01	0.58 (0.32–1.04)	0.07
000000001110111111_00	0.01	0.01	0.75 (0.43–1.33)	0.3
011111011110111111_01	0.01	0.01	1.29 (0.65–2.56)	0.5
3. 11111111**11111**111111_**1**0	0.01	0.01	2.25 (0.98–5.18)	0.06
Rare pooled			1.29 (0.96–1.72)	0.09

16q21 Haplotypes of 21 candidate SNPs in 3 populations.
Haplotypes derived from 21 candidate causal variants. Order; **rs17271951**, rs9936081, rs1345388, rs12918816, rs1362548, rs9921569, **rs35650695**, rs4784223, **rs12930156**, **rs3095604, rs28463809, rs4784226**, rs45465998, **rs45482301, rs3095606, rs3095607**, rs3112578, **rs3803662, rs3803661, rs4784227**, rs12922061 (bold SNPs remain after LR analysis, Table 2).
0 represents the common/protective allele and 1 the minor/risk allele in Europeans. Numbered haplotypes discussed in the text.
Analyses were stratified by study centre. Asians studies: SEBCS, MEC-J, SBCS, LAABC and TBCS. African-American studies: CARE-5 Cities, CARE-LA, MEC-AA and LIFE.
[a]Alleles underlined were not included in haplotype analysis because not genotyped in MEC-J, SBCS, LAABC and TBCS. Haplotypes based on subset of SNPs were related to the full haplotype based on SEBCS data.
[b]SNP rs3803661 not genotyped in African-American studies.

The 14 remaining variants are all strong candidates for being causally important for breast cancer risk. We have explored these SNPs further using analysis of (i) chromatin conformation, (ii) evolutionary conservation and (iii) transcription factor binding site motifs. It is important to note that although rs4784227, in *LOC643714*, is the most significant SNP of those we tested, the other 13 remaining candidates could not be excluded at 100:1 odds, and any one of these may be the causative variant that we are seeking. The three analyses, listed above, are not definitive but they hint that the causative variant could be one of the five candidates (SNPs M, N, O, P and Q) located within open chromatin in the conserved, intergenic region.

We have, additionally, used the currently available data to search for evidence that this locus may regulate expression levels of neighbouring genes. In this analysis, we focused on the association of gene mRNA levels with SNP rs3803662, the best initial tagSNP of the locus and strongly correlated ($r^2 > 0.8$) with the other remaining candidates in Europeans (It is unlikely that any of the other 13 SNP associations with expression would differ substantially from those observed for rs3803662 in breast tissues from European subjects. Association of rs4784227 with *RBL2* expression, in the HapMap lymphoblastic cell lines, demonstrated similar findings to rs3803662—data not shown). These limited data have raised the intriguing hypothesis that this breast cancer locus might act via regulation of the *RBL2* gene. However, this will need confirmation in larger datasets when they become available, as the risk of a false-positive finding is quite high, given the number of mRNAs examined in the existing small sample sets. An alternative hypothesis—that this locus regulates *TOX3* and/or *LOC643714* would be highly plausible, but this is not apparent from breast tissue or lymphocyte expression levels, perhaps because the relevant transcript or time-point was not examined. It also remains possible that this locus could regulate more distant genes in *cis* or even in *trans*.

Our combined evidence thus indicates a likely gene-regulatory function for this locus, but the gene or genes under regulation are not easily identified. Other tests

Table 4. Association of SNP rs3803662 with TOX3 and RBL2 mRNA levels

Gene	Probe used for mRNA detection (Illumina)	Increase in mRNA per T allele of rs3803662 in: 77 Breast tumours	270 HapMap lymphocytes (Meta-analysis of 60 CEU parents, 60 YRI parents, 45 CHB and 45 JPT)	38 Normal breast samples
TOX3	GI_29747038	0.76 (0.7)	not expressed in lymphocytes	1.28 (0.8)
RBL2	GI_21361291	−0.01 (0.8)	0.06 (0.01)	not tested

Evaluated by linear regression. All values are effect (*P*-value).
T allele = risk associated allele in European and Asian studies.

of function will be required to evaluate the 14 variants that remain candidates after exhaustive evaluation by epidemiological studies.

MATERIALS AND METHODS

Study populations

Initial associations were detected in the SEARCH breast cancer study, a population-based study in East Anglia (12). Eight additional studies were included in this fine-scale mapping work, all containing cases diagnosed with invasive breast cancer and cancer-free controls (see Supplementary Material, Methods). Briefly, there were 7536/7710 cases/controls of European ethnicity from the SEARCH (6704/6840) and KARBAC (832/870) studies; 4268/3868 cases/controls of Asian ethnicity from the MEC (Japanese 447/394), LAABC (Japanese 447/394, Chinese 263/375, Filipino 304/297), SEBCS (2159/1548) and TBCS (920/940) studies; and 1323/1398 controls/cases of African-American ethnicity from the CARE-5 Cities (452/435), MEC (428/654) and LIFE combined with CARE-LA (518/234) studies.

Re-sequencing

In order to create a full catalogue of common SNPs (MAF $\geq$ 0.5), DNA samples from CEPH individuals were sequenced across the 133 kb region of linkage. Two hundred and sixty-four overlapping PCR amplicons were designed from positions 51 074 000 to 51 206 999 of chromosome 16 (average amplicon size 660 pb, 160 pb overlap). M13-tagged PCR products were bidirectionally sequenced using Big Dye 3.0 (Applied Biosystems) and processed using automated trace analysis through the Cancer Genome Workbench (cgwb.nci.gov). The sequencing was done in two stages, with 108 kb in the first stage sequenced on 45 CEPH subjects and 25 kb in the second on 42 of the same individuals. In the first stage, 67% of nucleotides across the region could be scored for polymorphisms in at least 80% of subjects. In the second stage, 93% of nucleotides could be scored for polymorphisms in at least 80% of 43 subjects. This gave a >98% probability of detecting a variant with an MAF > 5%. A total of 423 variants were identified in this region with 245 at MAF $\geq$ 5%. SNP data from 57 CEPH individuals sequenced through the 1000 Genomes Project (http://www.1000genomes.org), identified an additional 48 variants with MAF $\geq$ 5%. A comparison of the SNPs identified by these two different sources is shown in Supplementary Material, Table S4.

Genotyping

For 22 of the 26 candidate SNPs, genotyping was performed in individual centres by TaqMan 5′ nuclease assay on 10 ng template DNA in a 384 well format containing No Template Controls and duplicate samples in each plate to ensure quality control. Genotypes were determined using the ABI PRISM 7900HT Sequence Detection System according to the manufacturer's instructions. Primers and probes were obtained from Applied Biosystems (http://www.appliedbiosystems.com/) as Assays-by-Design (distributed by one centre).

For four variants, which could not be assayed with TaqMan, genotyping was performed via bidirectional sequencing. These four variants included an insertion/deletions (rs45538731 [O], also known as rs45492607 and rs1362549) and three SNPs that failed to design as Assays-by-Design (rs35668161 [B], rs12600239 [C] and rs7500427 [D]). PCR conditions were: 95°C for 10 min, 40 cycles of 94°C for 30 s, 60°C (or 64°C) for 30 s, 72°C for 45 s and a final extension step at 72°C for 10 min. For up to 46 SEARCH samples and 69 Korean samples, 5–10 ng genomic DNA were used in a 5–10 μl PCR reactions. PCR products were treated using the ExoSAP-IT method (USB Corporation, Cleveland, OH, USA) and bidirectional sequencing performed using BigDye Terminator v3.1 Cycle Sequencing Kits (Applied Biosystems, Foster City, CA, USA) according to the manufacturer's instructions. Samples were run on a 3100 Genetic Analyzer (ABI) using Run 3100 Data Collection v2.0 and Sequencing Analysis Version 5.1.1 software.

DNase I hypersensitivity

Three breast cell lines, MCF-7, PMC42 and MDA231, two prostate lines, LnCapC4b and RWPE-1 and one colon cancer cell line, HCT116, were obtained from the Cambridge Research Institute (CRI) culture collection. All cell lines were maintained in RPMI with 10% foetal calf serum, except RWPE-1 which was maintained in keratinocyte serum free medium (Gibco, UK) supplemented with 0.05 mg/ml bovine pituitary extract and 5 ng/ml epidermal growth factor (both from Sigma, UK). Primary human T-cells were isolated from filters obtained from the Cambridge Blood Transfusion Service (REC reference number: 04/Q0108/21). Filters were flushed and cells purified using MACS® Separators (Miltenyi Biotec) and cultured for 24 h in RPMI medium supplemented with 20% foetal calf serum and 2% PHA-M (Sigma, UK). Cells were harvested while in the exponential growth phase. DNase I hypersensitivity experiments were carried out as described in Follows *et al*. (13) with amendments as published

(14). The array data were corrected using Loess normalization and analyzed by ACME (15), using a 95% cut-off and a sliding window size of 500 bp by the CRI Computational Biology group. For each cell line at least two hybridizations were carried out. Results obtained with cell lines from the same tissue were averaged. The data was visualized using the Affymetrix Integrated Genome Browser.

Transcription factor binding sites searches

Searches were performed for the 100 base-pair surrounding sequences of the remaining candidate causal SNPs in following transcription factor motif search engines: AliBaba 2.1 (16) (http://darwin.nmsu.edu/~molb470/fall2003/Projects/solorz/aliBaba_2_1.htm), TFSEARCH (17) (http://www.cbrc.jp/research/db/TFSEARCH.html), Genomatix (18) (http://www.genomatix.de/). Scores >0.85 using at least two search engines were considered indicative of genuine prediction. Potential regulatory modules were identified using a fourth program, PReMod (19,20) (http://genomequebec.mcgill.ca/PReMod), which predicts regulatory regions on the basis of the clustering of TF binding sites. The potential of SNP binding proteins to interact with these modules was assessed using BioGRID, General Repository for Interaction Datasets (http://www.thebiogrid.org).

Expression analyses

DNA from 77 breast tumours from the Nottingham City Cohort (21), and 38 normal breast samples were genotyped for SNP rs3803662 using fluorescent 5′ exonuclease assay (TaqMan, Applied Biosystems). Normal breast tissue was collected at the Addenbroke's Hospital, from women undergoing aesthetic surgery, for reasons not related to cancer. The samples were analysed by a histopathologist, to ensure that they were free of displasia. Ethical approval was obtained for the collection and research use of all blood and breast samples used in this study.

Analysis of comparative *TOX3* expression was performed on total RNA from a subset of 11 breast tumour and 12 normal breast samples. cDNA was prepared with the TaqMan Reverse Transcription Reagents kit (Applied Biosystems) using random hexamers, according to the manufacturer's instructions. Expression levels were determined using TaqMan Gene Expression assay Hs00300355_m1 (Applied Biosystems) for *TOX3* and primers specific to *LOC643714* predicted transcript ENSESTT00000054674 (Ensembl database) with SYBRgreen mix (ABI), and normalized to two different housekeeping genes. All samples were run in triplicate.

Associations between *TOX3* expression and rs3803662 genotype were assessed using linear regression. Expression levels of the 38 normal breast samples were determined using TaqMan Gene Expression assay Hs00300355_m1 (Applied Biosystems) for *TOX3*. Microarray expression data for the breast tumours were available using the Illumina platform (22). For analyses involving breast tumours, we incorporated in the regression model a covariate for copy number based on array-comparative genome hybridization data (21) using the CGH probe closest to each gene expression probe location.

Analyses of the relationship between SNP genotype and gene expression were also analyzed with publicly available expression data generated from Epstein-Barr virus-transformed lymphoblastoid cell lines (8,9).

Statistical methods

Each of the 22 SNPs genotyped by TaqMan was assessed for association with disease status using a likelihood ratio test. Subjects missing more than 25% of the genotyped variants were excluded from analyses. Per-allele odds ratios (ORs) and confidence intervals (CIs) were estimated by logistic regression stratified by study centre using Intercooled Stata version 8.2. Some SNPs were not genotyped by all study centres (see Supplementary Material, Fig. S1), and genotypes of these SNPs were imputed (see below).

Sampling weights were developed for the CARE-5 Cities data to account for the non-random selection of subjects from the study population described in greater detail in (23). Estimates were very similar using both weighted and unweighted analysis (data not shown). For simplicity and consistency with the fine-scale mapping analyses, unweighted analyses are presented in the main text.

Haplotype frequencies were estimated using the haplo.stats package in S-plus (24), separately for the European and Asian populations, using the data from the case–control studies on whom the tagSNPs plus the 115 individuals on whom all SNPs were typed. The haplotype frequencies were used to impute genotype probabilities for each SNP in each individual. An Expectation-maximization (EM) algorithm was then used to fit a logistic regression model allowing uncertainty in the genotypes of the untyped SNPs and assuming that each SNP in turn was the causal variant. Thus, we calculated the likelihood that each SNP was the causal variant (1).

Haplotype analysis was conducted using an in-house program based on the TagSNPs program (25). Breast cancer risk was assessed for common haplotypes (frequencies >0.01) composed of the 20 SNPs genotyped by TaqMan. Rare haplotypes were pooled. Haplotype frequencies and subject-specific expected haplotype indicators were calculated separately for each study using the EM algorithm to account for the haplotype uncertainty given the unphased genotype data. Subjects missing $>50\%$ of genotype data were excluded from the analysis. Logistic regression was used to generate haplotype specific risks with respect to the baseline, chosen as the most common haplotype.

SUPPLEMENTARY MATERIAL

Supplementary Material is available at *HMG* online.

ACKNOWLEDGEMENTS

The authors thank all the women who participated in this research as well as Caroline Baynes, Don Conroy, Craig Luccarini, Hannah Munday and Mitul Shah for work within the SEARCH study.

Conflict of Interest statement. None declared.

FUNDING

The genotyping and analysis of this study, and the conduct of the SEARCH study, was funded by Cancer Research UK grants (C20/A3084, C1287/A10118, C490/A11021, C8197/A10123, C1287/A7497, C8197/A10865) and COGS EU FP7 Health-F2-2009-223175. A.M.D., P.D.P. and H.F. were supported by Cancer Research UK. This research was supported in part by the Intramural Research Programs of the National Cancer Institute and National Human Genome Research Institute, NIH, U.S. Department of Health and Human Services. M.U. was supported by the NIH-Oxford/Cambridge PhD program. The MEC Study was supported by the US National Cancer Institute (CA 54281, CA 63464, CA132839). The LAABC study was supported by the California Breast Cancer Research Program (1RB-0287, 3PB-0102, 5PB-008, 10PB-0098). The CARE study was supported by the National Institute of Child Health and Human Development, with additional support from the National Cancer Institute, through contracts with Emory University (N01 HD 3-3168), Fred Hutchinson Cancer Research Center (N01 HD 2-3166), Karmanos Cancer Institute at Wayne State University (N01 HD 3-3174), University of Pennsylvania (N01 HD 3-3176), University of Southern California (N01 HD 3-3175) and through an intra-agency agreement with the Centers for Disease Control and Prevention (Y01 HD 7022). General support through SEER contracts [N01-PC-67006 (Atlanta), N01-CN-65064 (Detroit), N01-CN-67010 (Los Angeles) and N01-CN-0532 (Seattle)] are also acknowledged. B.A.J.P. is Li Ka Shing Professor of Oncology and we acknowledge Hutchison Whampoa Limited. KARBAC thank the Swedish Cancer Society, The Jubilee Foundation, and Bert von Kantzow foundation. The Seoul Breast Cancer Project is supported by the Ministry of Health & Welfare, ROK (01-PJ3-PG6-01GN07-0004).

REFERENCES

1. Easton, D.F., Pooley, K.A., Dunning, A.M., Pharoah, P.D., Thompson, D., Ballinger, D.G., Struewing, J.P., Morrison, J., Field, H., Luben, R. *et al.* (2007) Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature*, **447**, 1087–1093.
2. Stacey, S.N., Manolescu, A., Sulem, P., Rafnar, T., Gudmundsson, J., Gudjonsson, S.A., Masson, G., Jakobsdottir, M., Thorlacius, S., Helgason, A. *et al.* (2007) Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.*, **39**, 865–869.
3. Yuan, S.H., Qiu, Z. and Ghosh, A. (2009) TOX3 regulates calcium-dependent transcription in neurons. *Proc. Natl. Acad. Sci. USA*, **106**, 2909–2914.
4. Smid, M., Wang, Y., Klijn, J.G., Sieuwerts, A.M., Zhang, Y., Atkins, D., Martens, J.W. and Foekens, J.A. (2006) Genes associated with breast cancer metastatic to bone. *J. Clin. Oncol.*, **24**, 2261–2267.
5. Rakha, E.A., Green, A.R., Powe, D.G., Roylance, R. and Ellis, I.O. (2006) Chromosome 16 tumor-suppressor genes in breast cancer. *Genes Chromosomes Cancer*, **45**, 527–535.
6. (2003) The International HapMap Project. *Nature*, **426**, 789–796.
7. Barrett, J.C., Fry, B., Maller, J. and Daly, M.J. (2005) Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, **21**, 263–265.
8. Stranger, B.E., Forrest, M.S., Clark, A.G., Minichiello, M.J., Deutsch, S., Lyle, R., Hunt, S., Kahl, B., Antonarakis, S.E., Tavare, S. *et al.* (2005) Genome-wide associations of gene expression variation in humans. *PLoS Genet.*, **1**, e78.
9. Dixon, A.L., Liang, L., Moffatt, M.F., Chen, W., Heath, S., Wong, K.C., Taylor, J., Burnett, E., Gut, I., Farrall, M. *et al.* (2007) A genome-wide association study of global gene expression. *Nat. Genet.*, **39**, 1202–1207.
10. Mayol, X., Grana, X., Baldi, A., Sang, N., Hu, Q. and Giordano, A. (1993) Cloning of a new member of the retinoblastoma gene family (pRb2) which binds to the E1A transforming domain. *Oncogene*, **8**, 2561–2566.
11. Naylor, T.L., Greshock, J., Wang, Y., Colligon, T., Yu, Q.C., Clemmer, V., Zaks, T.Z. and Weber, B.L. (2005) High resolution genomic analysis of sporadic breast cancer using array-based comparative genomic hybridization. *Breast Cancer Res.*, **7**, R1186–R1198.
12. Lesueur, F., Pharoah, P.D., Laing, S., Ahmed, S., Jordan, C., Smith, P.L., Luben, R., Wareham, N.J., Easton, D.F., Dunning, A.M. *et al.* (2005) Allelic association of the human homologue of the mouse modifier Ptprj with breast cancer. *Hum. Mol. Genet.*, **14**, 2349–2356.
13. Follows, G.A., Janes, M.E., Vallier, L., Green, A.R. and Gottgens, B. (2007) Real-time PCR mapping of DNaseI-hypersensitive sites using a novel ligation-mediated amplification technique. *Nucleic Acids Res.*, **35**, e56.
14. Udler, M.S., Meyer, K.B., Pooley, K.A., Karlins, E., Struewing, J.P., Zhang, J., Doody, D.R., MacArthur, S., Tyrer, J., Pharoah, P.D. *et al.* (2009) FGFR2 variants and breast cancer risk: fine-scale mapping using African American studies and analysis of chromatin conformation. *Hum. Mol. Genet.*, **18**, 1692–1703.
15. Scacheri, P.C., Crawford, G.E. and Davis, S. (2006) Statistics for ChIP-chip and DNase hypersensitivity experiments on NimbleGen arrays. *Methods Enzymol.*, **411**, 270–282.
16. Grabe, N. (2002) AliBaba2: context specific identification of transcription factor binding sites. *In Silico Biol.*, **2**, S1–S15.
17. Heinemeyer, T., Wingender, E., Reuter, I., Hermjakob, H., Kel, A.E., Kel, O.V., Ignatieva, E.V., Ananko, E.A., Podkolodnaya, O.A., Kolpakov, F.A. *et al.* (1998) Databases on transcriptional regulation: TRANSFAC, TRRD and COMPEL. *Nucleic Acids Res.*, **26**, 362–367.
18. Cartharius, K., Frech, K., Grote, K., Klocke, B., Haltmeier, M., Klingenhoff, A., Frisch, M., Bayerlein, M. and Werner, T. (2005) MatInspector and beyond: promoter analysis based on transcription factor binding sites. *Bioinformatics*, **21**, 2933–2942.
19. Blanchette, M., Bataille, A.R., Chen, X., Poitras, C., Laganiere, J., Lefebvre, C., Deblois, G., Giguere, V., Ferretti, V., Bergeron, D. *et al.* (2006) Genome-wide computational prediction of transcriptional regulatory modules reveals new insights into human gene expression. *Genome Res.*, **16**, 656–668.
20. Ferretti, V., Poitras, C., Bergeron, D., Coulombe, B., Robert, F. and Blanchette, M. (2007) PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res.*, **35**, D122–D126.
21. Chin, S.F., Teschendorff, A.E., Marioni, J.C., Wang, Y., Barbosa-Morais, N.L., Thorne, N.P., Costa, J.L., Pinder, S.E., van de Wiel, M.A., Green, A.R. *et al.* (2007) High-resolution aCGH and expression profiling identifies a novel genomic subtype of ER negative breast cancer. *Genome Biol.*, **8**, R215.
22. Blenkiron, C., Goldstein, L.D., Thorne, N.P., Spiteri, I., Chin, S.F., Dunning, M.J., Barbosa-Morais, N.L., Teschendorff, A.E., Green, A.R., Ellis, I.O. *et al.* (2007) MicroRNA expression profiling of human breast cancer identifies new markers of tumor subtype. *Genome Biol.*, **8**, R214.
23. Malone, K.E., Daling, J.R., Doody, D.R., Hsu, L., Bernstein, L., Coates, R.J., Marchbanks, P.A., Simon, M.S., McDonald, J.A., Norman, S.A. *et al.* (2006) Prevalence and predictors of BRCA1 and BRCA2 mutations in a population-based study of breast cancer in white and black American women ages 35 to 64 years. *Cancer Res.*, **66**, 8297–8308.
24. Schaid, D.J., Rowland, C.M., Tines, D.E., Jacobson, R.M. and Poland, G.A. (2002) Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.*, **70**, 425–434.
25. Stram, D.O., Haiman, C.A., Hirschhorn, J.N., Altshuler, D., Kolonel, L.N., Henderson, B.E. and Pike, M.C. (2003) Choosing haplotype-tagging SNPS based on unphased genotype data using a preliminary sample of unrelated subjects with an example from the Multiethnic Cohort Study. *Hum. Hered.*, **55**, 27–36.