Journal of Human Genetics (2011) 56, 377–382
© 2011 The Japan Society of Human Genetics All rights reserved 1434-5161/11 $32.00
www.nature.com/jhg

ORIGINAL ARTICLE

Composite likelihood-based meta-analysis of breast cancer association studies

Ioannis Politopoulos[1], Jane Gibson[1], William Tapper[1], Sarah Ennis[1], Diana Eccles[2] and Andrew Collins[1]

For detecting low risk disease variants in genome-wide association panels, meta-analysis is a powerful strategy to increase power. We apply a composite likelihood-based method, which models association with disease in regions defined on a linkage disequilibrium map and combines the evidence across multiple genome-wide samples. This fixed region approach has the advantage that, as only one statistical test is made per region, there is no increased multiple testing penalty in higher marker density panels. Imputation of missing genotypes is also advantageous to increase coverage. Meta-analysis of three breast cancer data sets combines evidence from samples that show heterogeneity in phenotype and, particularly, in marker coverage. The *FGFR2* gene has the highest rank, consistent with previous analysis of one of these samples and supported by the small number of early-onset breast cancer cases included. The 8q24 breast cancer region also ranks highly and is supported by evidence from both early-onset and post-menopausal breast cancer samples. The *PIK3AP1* gene region is highlighted in this analysis as a strong candidate for further study.

Journal of Human Genetics (2011) **56,** 377–382; doi:10.1038/jhg.2011.23; published online 10 March 2011

Keywords: association mapping; breast cancer; composite likelihood; *FGFR2*; genotype imputation; genome-wide association; meta-analysis; 8q24

INTRODUCTION

Genome-wide association studies using large samples of cases and controls have been successful in identifying many genes involved in 'common' diseases. For breast cancer at least 20 genes or genomic regions have been found to be associated with the disease using these methods.[1–12] Rarer forms of variation such as disease variants in the *BRCA1* and *BRCA2* genes and moderate penetrance genes found by targeted candidate studies bring the total number of genes implicated in breast cancer to about 30. However, collectively these only account for ~30% of the disease genetic variance, suggesting many more genes remain to be discovered. There are several possible sources for the 'missing heritability' including genes missed because of low statistical power and uneven marker coverage, heterogeneity through sub-phenotypes associated with different genetic variants and rare genetic variation (as only common variants are screened by existing single nucleotide polymorphism (SNP) panels). One route to identifying new targets is through meta-analysis, which can increase the power to detect novel disease variation for further follow-up by combining data from independent samples. Zeggini *et al.*[13] describe meta-analysis of genome-wide association studies combining evidence for individual single SNPs using genotyped and imputed data. We develop and apply a composite likelihood-based method, which models information from multiple SNPs in a given genomic region and combines evidence across corresponding regions in independent data sets. Modelling association with disease on an underlying linkage disequilibrium (LD) map further increases power and resolution of mapping, compared with single SNP tests.[14,15] This approach also partitions the genome into regions, which contain equivalent levels of LD. As only one statistical test is made in each region, there is no increased multiple-testing penalty with greater marker density. The three data samples analysed here are from the Cancer Genetic Markers of Susceptibility Project (CGEMS),[1] which comprises post-menopausal breast cancer samples, a sample from the prospective study of outcomes in sporadic versus hereditary breast cancer (POSH) cohort of early-onset breast cancer,[4,16] and data from the Wellcome Trust Case Control Consortium (WTCCC).[17] The first two data sets have genome-wide SNPs whereas the latter comprises genotypes for non-synonymous coding SNPs only. The three samples therefore show a high degree of heterogeneity in both phenotype and marker coverage and thus present a challenge for meta-analysis. Previous findings from these data sets include primary evidence for association of the *FGFR2* gene with breast cancer,[1] whereas the POSH sample formed part of a larger study that determined novel breast cancer genes.[4] Our analysis examines the evidence for breast cancer genes through composite likelihood-based meta-analysis in these three samples.

[1]Genetic Epidemiology and Bioinformatics Research Group, Human Genetics Research Division, University of Southampton, School of Medicine, Southampton General Hospital, Hampshire, UK and [2]Cancer Sciences Division, University of Southampton, School of Medicine, Southampton General Hospital, Hampshire, UK
Correspondence: Professor A Collins, Human Genetics Research Division, Genetic Epidemiology and Bioinformatics Research Group, University of Southampton, School of Medicine, Duthie Building (808), Southampton General Hospital, Tremona Road, Southampton, Hampshire SO16 6YD, UK.
E-mail: arc@soton.ac.uk

Received 13 August 2010; revised 3 February 2011; accepted 6 February 2011; published online 10 March 2011

MATERIALS AND METHODS

Data preparation and quality control

We undertook the following procedures for the three data samples analysed in the meta-analysis:

CGEMS sample. The CGEMS sample (http://cgems.cancer.gov/) comprises data from 1145 post-menopausal breast cancer patients and 1142 controls genotyped with 555 148 SNPs. The downloaded sample excludes data that failed the original quality control (QC) employed by Hunter *et al.*[1] Using PLINK[18] we removed an additional 93 SNPs with inconsistent or ambiguous genomic locations, 8648 SNPs and one individual with $>10\%$ of genotypes missing, 53 615 SNPs with minor allele frequencies (MAF) <0.05 and 4308 SNPs with large deviations from Hardy–Weinberg (HW) ($\chi^2_1 > 10$) in the controls. Some SNPs failed QC for more than one of these reasons. After merging with the HapMap (phase 3) reference populations and undertaking multidimensional scaling cluster analysis, we identified and removed an additional four samples, which did not cluster strongly with the CEU (Western European origin) reference population, suggesting potential admixture. After QC we had a total of 498 786 SNPs typed in 1143 cases and 1139 controls.

We increased genotypic coverage by 98% through imputation of missing SNP genotypes, after merging CGEMS data with CEU data, and a total of 544 683 markers were imputed 'sufficiently' according to PLINK defaults. The default cut-offs designate reliably imputed SNPs and require there to be 'information content metric' values >0.8 and SNP genotype imputation for 90% or more individuals in the sample. Before combining with CGEMS, further QC removed 55 692 SNPs (488 991 imputed SNPs retained), of which 308 had $>10\%$ missing genotypes, 6800 showed significant HW deviation in the controls and 49 019 had MAF <0.05. Some SNPs failed QC for more than one of these reasons. The combined data set comprised genotypes for 987 777 SNPs (Supplementary Table 1).

POSH sample. Turnbull *et al.*[4] describe an analysis, which includes 308 POSH cases genotyped on an Illumina Infinium (Illumina, Inc, San Diego, CA, USA) 660k array forming part of a study of cases preferentially selected to have at least two affected first or second degree relatives. Most cases had been screened and found to be negative for germline mutations in the *BRCA1* and *BRCA2* genes. Data for 294 POSH cases and 580 030 SNPs were provided by the lead authors. Their QC procedures[4] resulted in the exclusion of 63 112 markers. In addition, a total of 14 individual samples were excluded because of apparent non-European ancestry (eight samples) and heterozygosity *P*-values $<10^{-5}$ (six samples). We used WTCCC phase 2 genotypic data from the European Genotype Archive (EGA) (http://www.ebi.ac.uk/ega/page.php?page=study&study=EGAS00000000028&cat=www.wtccc2.studies.xml&subcat=controls) as controls. Both 1958 birth cohort and UK National Blood Service controls (UNBS) Illumina 1.2M genotypic data sets were used in the analysis. During QC on the controls, using the exclusions lists supplied by the WTCCC, we removed 231 samples and 215 732 SNPs from the 1958 birth cohort and 236 samples and 214 848 SNPs from the UNBS data. From the original 1 155 595 markers genotyped across 2930 1958 birth cohort controls and 2737 NBS controls, first-stage QC yielded 939 863 SNPs in 2699 controls and 940 747 in 2,501 controls in the two samples, respectively.

We determined a common subset of 536 205 SNPs typed in both WTCCC controls and POSH cases genotypic data sets. Our standard QC in the combined data set identified an additional 3926 SNPs deviating from HW equilibrium ($\chi^2_1 > 10$) in the controls, 106 SNPs with $>10\%$ missing genotypes and 23 738 SNPs with MAF<0.05. The final data set comprised genotypic information for 280 cases, 5200 controls and 506 610 SNPs after removal of 2138 SNPs contained in the exclusion lists provided by the lead authors.

Using PLINK we then determined 535 110 sufficiently imputed SNPs, of which 17 387 were excluded because of significant HW deviation, 270 were excluded with $>10\%$ missing genotypes and 46 410 SNPs with MAF<0.05 (or excluded on more than one criteria). The final imputation-inclusive data set comprised 979 409 SNPs for 5200 WTCCC phase 2 controls and 280 POSH cases, suggesting an increase in genotypic coverage by imputation of $\sim$93%.

WTCCC sample. WTCCC phase 1 breast cancer data were obtained from the European Genotype Archive (http://www.ebi.ac.uk/ega/page.php?page=study&study=EGAS00000000024&cat=www.wtccc.studies.xml.ega2&subcat=BC). The aggregated genome-wide data set of genotypes for 15 436 SNPs across 1045 cases and 1476 controls was subjected to QC following the annotation files provided, which resulted in 2859 SNPs and 51 samples (41 cases and 10 controls) being removed. Marker exclusion was based on poor genotype call scores, high missing genotype rate, monomorphic SNPs and HW deviations. Sample exclusion was due to putative relatedness of individuals, questionable ancestry, missing genotypes and positive *BRCA2* testing. The final data set comprised of 12 577 SNPs, 1004 cases and 1466 controls.

Genome-wide imputation in QC-clean WTCCC data, after merging with the CEU data (111 individuals and 1 615 203 SNPs), identified 36 587 sufficiently imputed SNPs. Our standard QC in the combined data set identified 28 SNPs with $>10\%$ missing genotypes, 1829 SNPs with significant HW deviation and 7052 SNPs with MAF<0.05. The final imputation-inclusive data set comprised 40 300 SNPs, 1004 cases and 1466 controls.

Composite likelihood mapping

We used the CHROMSCAN program,[19] which models association between disease and SNP markers in a chromosome region to compute a maximum composite likelihood location, *S*, for a disease variant, a standard error for that location, a 95% confidence interval and a *P*-value. The program incorporates the underlying LD structure in the region as a LD unit (LDU) map,[20] which represents regions of strong LD (blocks) as plateaus and recombination hot-spots as steps when plotted against the kilobase map. Gene mapping on the LDU map has been shown to increase power and accuracy.[21] The LDU maps were made from the CEU sample (HapMap phase II and build 36 of the human genome sequence). CHROMSCAN establishes significance for a region through a permutation test, which employs a large number of replicates for which the disease phenotype is randomised by shuffling. Computing probabilities for the test statistic based on the null *P*-value distribution avoids distortions (inflation and deflation) in the *P*-value distribution. The program generates an information matrix from which information weights, *W*, for location *S* are obtained along with a standard error. CHROMSCAN analysis was performed in fixed regions of four LD units, which facilitates meta-analysis and provides reasonable coverage (on average) of each region (>30 SNPs in a $\sim$550 000 SNP scan, given that there are $\sim$60 000 LDUs in the CEU genome[22,23]).

Meta-analysis

For combining information across the CGEMS, POSH and WTCCC samples, we examined Fisher's combined probability test (CPT),[24] the Z-transform test (ZTT),[25] and the weighted *Z*-test (WZT).[26] Whitlock's study[26] shows that the WZT has greater power and precision than the CPT and ZTT in simulated data. We used the CPT to combine permutation-based *P*-values from the three samples (or fewer if a sample contained no information for a given region) in corresponding four LDU regions as: $\chi^2_F = -2\sum_{i=1}^{k} \ln P_i$, where *k* is the number of samples and χ^2_F has a χ^2 distribution with 2*k* degrees of freedom. χ^2 were converted to the corresponding probability and χ^2_1 using the appropriate functions from the *gsl* library (http://www.gnu.org/software/gsl/manual/html_node/The-Chi_002dsquared-Distribution.html). The ZTT first converts *P*-values to the corresponding (signed) standard normal deviates Z. *Z*-scores were obtained from the permutation-based probabilities from each sample using the *gsl* library as above. The combined *Z*-score (Z_S) is obtained as: $Z_S = \sum_{i=1}^{k} z_i/\sqrt{k}$, where Z_S has a standard normal distribution. We computed the corresponding combined *P*-values and the corresponding χ^2_1 using the *gsl* library. The WZT is a weighted version of the ZTT. We weighted each sample by information *W*, from the composite likelihood model as above, and obtained $Z_W = \sum_{i=1}^{k} W_i Z_i / \sqrt{\sum_{i=1}^{k} W_i^2}$.

We evaluated statistical heterogeneity from pooling evidence across samples following Tapper *et al.*[27] which computes $\sum W_i(\hat{S} - S_i)^2$, the heterogeneity χ^2 with *k*-1 degrees freedom (χ^2_{k-1}) where $\hat{S} = \sum W_i S_i / \sum W_i$ and S_i is the maximum composite likelihood LDU location from the i^{th} sample with data in the LDU region under consideration.

RESULTS

The meta-analysis combines association tests in four-LDU regions across CGEMS (14 340 regions with data), POSH (13 908 regions with data) and WTCCC (3679 regions with data) samples. Although the WTCCC sample comprises far fewer SNPs than the other samples (Supplementary Table 1) and only covers non-synonymous SNPs, it provides additional information that can be exploited in composite likelihood-based meta-analysis. Turnbull *et al.*[4] in their Table 1 detail 13 loci confirmed as associated with breast cancer through association studies (both genome-wide and candidate gene based). We examined evidence for each of these regions in our meta-analysis (Table 1). The *FGFR2* locus is known to contribute one of the largest effect sizes among the common susceptibility loci detected so far. The gene was identified in the CGEMS sample by Hunter *et al.*[1] who recognised it as a risk factor for sporadic postmenopausal breast cancer. It is noteworthy that the evidence is supported by the small number (280) of POSH individuals, indicating it may have roles in both early-onset and post-menopausal breast cancer. The 95% confidence interval for the location of the associated variant for CGEMS and POSH samples spans 26 kb, which corresponds to intron 2 of the gene (Table 2). Intron 2 is not tagged by the non-synonymous SNP panel from the WTCCC genotypes so this sample provides no additional evidence. The combined meta-analysis χ^2_1 of 25.75 (ZTT) makes this the highest ranked region and the only region achieving genome-wide significance, P=0.006, after a conservative Bonferroni correction for 14 340 tested regions.

The evidence for other breast cancer genes (Table 1) is variable and reflects the relatively low power of fairly small, heterogeneous and incompletely genotyped samples for detecting low risk variants. However, it is noteworthy that the 8q24 breast cancer region has the fifth highest rank (ZTT χ^2_1 14.35) using combined evidence from CGEMS and POSH samples (but not WTCCC as this is a 'gene desert'[28]). There is evidence, therefore, that this region may be involved in both early-onset and post-menopausal breast cancer. The 8q24 region has well established associations with prostate, breast and colorectal cancer.[8] It has been shown that the effects of a number of risk alleles in the region are cancer site specific. Easton *et al.*[2] first reported an association between rs13281615 (128.42 Mb) (Figure 1) and breast cancer and subsequently Fletcher *et al.*[8] reported a protective effect at rs13254738 (128.17 Mb) with limited evidence for interaction between the two. The LDU map of the region (Figure 1) and the likelihood surface for the POSH data (Figure 2) shows that the 95% confidence interval (128.38–128.44 Mb) (Table 2) includes rs13281615. However, rs13254738 lies outside of the four LDU region and that region was not identified as high ranking in the meta-analysis. As pointed out by Fletcher *et al.*[8] the relatively low power of studies undertaken so far suggests that 8q24 may contain several additional breast cancer susceptibility loci. Composite likelihood evidence from the CGEMS sample (Table 1) places the peak at 128.497 Mb (Figure 1) within the prostate cancer 'region 3',[8] although the most significant single SNP in the region is rs10447995 at 128.427 Mb, much closer to, and in the same LD block as rs1328165 and the peak from the POSH sample. Multiple signals reflect the existence of a cluster of, possibly independently acting, associated variants with heterogeneous influences on disease phenotype, which may impact, for example, on differences in age of onset.

The *COX11* gene region ranks one hundred and seventh (ZTT) and the WTCCC study provides contributory information in the meta-analysis. The most significant non-synonymous WTCCC SNP in this region is rs7222197, which is in the *STXBP4* gene, adjacent to *COX11*. However this presumably reflects LD across this region as the SNP considered to be most strongly associated with risk is correlated with elevated levels of cytochrome C assembly protein 11 (*COX11*) and not with altered expression at *STXBP4*.[29] However, causal relationships between this effect and breast cancer predisposition have not been determined.

Of the 13 known breast cancer genes/regions listed in Table 1 the highest χ^2_1, suggesting a more powerful test, is achieved by the ZTT in six cases, by the WZT in five cases and by the CPT in two cases. Consistent with this pattern, the highest sum of χ^2 is for the ZTT (69.02) whereas the lowest is for the WZT (63.6). Given this empirical

Table 1 Results for known breast cancer genes and regions

			CGEMS		*POSH*		*WTCCC*				
Locus	*Chromosome*	*LDU range*	N[a]	χ^2_1	N[a]	χ^2_1	N[a]	χ^2_1	*CPT*χ^2_1	*WZT*χ^2_1	*ZTT*χ^2_1 (*rank*[b])
FGFR2	10q26.13	2376–2380	47	16.37	45	10.93	0	—	24.18	24.51	25.75 (1)
8q24	8q24	2256–2260	79	10.55	82	5.45	0	—	13.30	6.09	14.35 (5)
COX11	17q23.2	1200–1204	117	5.68	119	0.82	17	4.02	6.31	7.84	6.99 (107)[c]
MAP3K1	5q11.2	1144–1148	132	3.70	138	2.35	27	0.75	3.41	2.32	4.24 (502)
2q35	2q35	3608–3612	65	3.11	60	3.77	0	—	4.93	5.59	5.84 (210)
5p12	5p12	1028–1032	333	5.91	321	0.75	0	—	4.41	5.99	4.17 (529)
1p11.2	1p11.2	2372–2376	46	3.05	50	0.27	18	2.05	2.26	0.55	2.62 (1362)
TOX3	16q12.1	980–984	45	0.25	48	4.95	0	—	3.01	4.85	2.39 (1551)
6q25.1	6q25.1	2692–2696	114	3.03	116	0.09	20	0.87	1.24	2.18	1.20
RAD51L1	14q24.1	1016–1020	134	1.05	135	0.76	0	—	0.80	1.13	1.14
LSP1	11p15.5	44–48	52	0.02	55	3.20	2	0.08	0.63	2.43	0.24
SLC4A7	3p22	768–772	79	0.89	95	0.64	9	0.00	0.16	0.14	0.09[c]
CASP8	2q33.1	3344–3348	119	1.52	119	0.00	33	0.02	0.08	0.02	0.00
Total	—	—	—	—	—	—	—	—	64.72	63.63	69.02

Abbreviations: CPT, combined probability test; LDU, linkage disequilibrium unit; POSH, prospective study of outcomes in sporadic versus hereditary breast cancer; SNP, single nucleotide polymorphism; WTCCC, Wellcome Trust Case Control Consortium; WZT, weighted *Z*-test; ZTT, Z-transform test.
[a] Total number of SNPs typed and imputed in the region.
[b] Ranks for *Z*-test among 14 340 regions.
[c] Regions significant by heterogeneity test.

Table 2 The *FGFR2* and 8q24 regions: association with post-menopausal and early-onset breast cancer

Data	N *SNPs*	*Location S (LDU, Mb)*	*95% CI, Mb (difference, Kb)*	χ^2_3 *(P value)*	*msSNP (Mb)*
FGFR2					
CGEMS	47	2379.7, 123.342	123.316–123.342 (26)	22.47 (0.000052)	rs2420946 (123.341)
POSH	45	2379.7, 123.342	123.316–123.342 (26)	16.38 (0.000946)	rs2420946 (123.341)
8q24					
CGEMS	79	2259.7, 128.497	128.470–128.497 (27)	15.95 (0.001159)	rs10447995 (128.427)
POSH	82	2257.7, 128.413	128.381–128.439 (58)	9.88 (0.019578)	rs622556 (128.402)

Abbreviations: CI, confidence interval; ms, most significant; LDU, linkage disequilibrium unit; POSH, prospective study of outcomes in sporadic versus hereditary breast cancer; SNP, single nucleotide polymorphism.

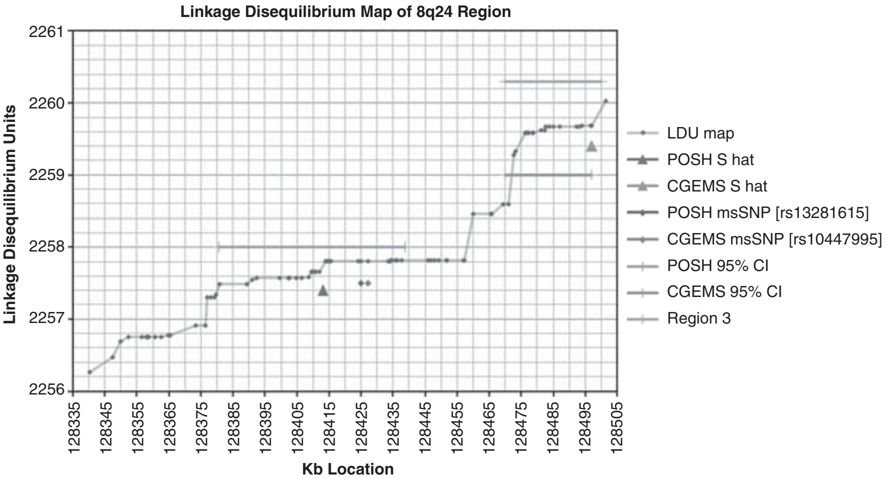

Figure 1 Linkage disequilibrium (LD) map of the 8q24 region. The LD map locations of the most significant (ms) SNPs are shown along with the maximum composite likelihood locations (S) for CGEMS and the prospective study of outcomes in sporadic versus hereditary breast cancer (POSH), and the 95% confidence intervals for the locations. The evidence from the POSH data maps to the same LD block as the rs13281615 variant previously implicated in breast cancer, whereas the CGEMS evidence maps to nearby 'region 3', but within the same four LD unit region, implicated in prostate cancer.

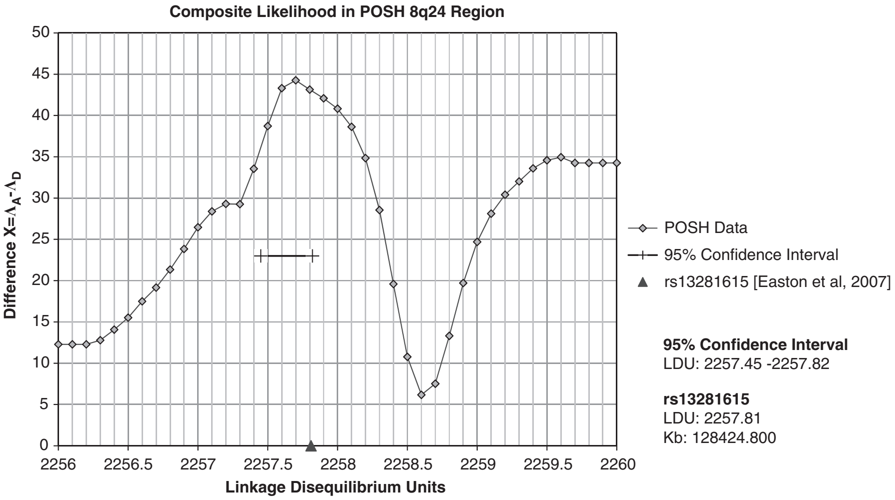

Figure 2 Composite likelihood in the prospective study of outcomes in sporadic versus hereditary breast cancer (POSH) 8q24 region. The test statistic difference *X* in (composite) likelihood for null and disease association models in the 8q24 genomic region. The location of the rs13281615 variant previously implicated in breast cancer is within the 95% confidence interval suggesting a role for this region in early-onset disease.

support for the ZTT we list the 10 highest ranked regions on the basis of this test (Table 3). The highest ranking regions identify the *FGFR2* region as most significant. Also represented is the region containing the *PIK3AP1* gene (χ^2_1 13.06). This gene is known to be associated with a key carcinogenesis pathway and has been found to be upregulated in the peripheral blood of breast cancer patients. Expression of this gene

Table 3 Highest ranked regions (ZTT) in meta-analysis (CGEMS, POSH and WTCCC)

Chromosome	*LDU region*	*CPT* χ_1^2	*ZTT* χ_1^2	*WZT* χ_1^2	*msSNP/gene*
10	2376–2380	24.18	25.75	24.51	rs2420946/*FGFR2*[a]
12	460–464	14.73	14.86[b]	16.40	rs11055780
17	1636–1640	16.88	14.64	18.29	rs745232
5	2760–2764	13.16	14.52[b]	13.57	rs7711971
8	2256–2260	13.30	14.35	6.09	rs10447995[a]
15	640–644	13.01	13.80	14.47	rs2437948/*FBN1*
17	1296–1300	12.23	13.43	13.66	rs8073257
10	1912–1916	11.45	13.06	5.37	rs563654/*PIK3AP1*
12	988–992	10.98	13.04	11.77	rs11168740/*ADCY6*
18	548–552	11.76	13.01	12.86	rs636173/*IMPA2*

Abbreviations: CPT, combined probability test; LDU, linkage disequilibrium unit; ms, most significant; POSH, prospective study of outcomes in sporadic versus hereditary breast cancer; SNP, single nucleotide polymorphism; WTCCC, Wellcome Trust Case Control Consortium; WZT, weighted *Z*-test; ZTT, Z-transform test.
[a] Established breast cancer gene/region.
[b] Regions significant by heterogeneity test.

has been used as part of a set of profiles as a molecular predictor of breast cancer.[30]

There is evidence for significant statistical heterogeneity for the *COX11* and *SLC4A7* breast cancer regions (Table 1) and for two of the top-ranked regions (Table 3). Among a number of likely sources for heterogeneity, evidenced by combining statistics from these samples, are variations in marker coverage and informativeness, differences in breast cancer phenotype, and the impact of multiple (or different) association signals within a region. Increasing sample sizes, more consistent marker coverage, more refined breast cancer phenotypes and fine mapping (including mapping within smaller LDU windows in samples with higher marker density) will reduce the impact of these sources of heterogeneity in future.

DISCUSSION

Composite likelihood-based meta-analysis in discrete regions defined on an underlying LD map has a number of advantages over single SNP based approaches for combining evidence. Model fitting combines evidence across a number of SNPs giving a point estimate along with a confidence interval in the region of interest. Within fixed regions increasing marker density, including that achieved by imputation of genotypes, does not increase the multiple testing penalty. Combination of *P*-values across regions using the ZTT enables meta-analysis of samples that have heterogeneous phenotypes (such as early and late-onset disease in the POSH and CGEMS samples respectively) and widely differing marker coverage profiles (such as the WTCCC compared with CGEMS and POSH data sets). The empirical evidence from known breast cancer gene regions (Table 1) marginally supports the use of the ZTT rather than the weighted test (WZT). The Fisher test (CPT) clearly lacks power as noted by Whitlock.[26] The weighted *Z*-test favoured by that author is likely to be the most powerful where reliable weights are available. The apparent modest superiority of the ZTT over the WZT in our study may reflect instability in the weights where the completeness of marker coverage and marker information content is particularly variable in these heterogeneous samples. There is also a statistical argument, pointed out by Whitlock,[26] that the *P*-values are already weighted by sample size when using the ZTT.[31] The empirical evidence supports the use of the ZTT in composite likelihood-based meta-analysis but examination of the weighted scores may identify further potential candidates for follow up.

This meta-analysis supports existing evidence for at least two known breast cancer genes and regions (*FGFR2* and 8q24), despite the relatively small number of samples included and heterogeneous marker coverage in the three data sets. The application of this approach to combination of evidence from larger samples and for defined breast cancer sub-types may be useful to further characterise the genetic basis of breast cancer and contribute to the identification of some of the 'missing heritability'.

The WTCCC non-synonymous SNP panel has, understandably, limited genome coverage even after imputation of SNPs. Perhaps remarkably, meta-analysis of 1200 SNPs known to be associated with diseases, found that for 40% of SNPs there was no association with known exonic sequences.[32] There are at least five associated breast cancer non-genic regions and other variants within genes are known to be intronic (for example, the *FGFR2* association).

Although only the *FGFR2* gene achieves genome-wide significance in the combined sample after correction for the number of regions tested, we note that the high ranked regions include the *PIK3AP1* gene, which is a promising candidate for further study. We also present evidence of association for *FGFR2* genes and 8q24 regions with early-onset disease, despite the relatively small number of early-onset (POSH) cases studied. Early-onset cancers include a greater proportion of estrogen receptor negative (ER−) tumours but most genome-wide association studies undertaken so far have focussed on later onset disease and have had greater power to detect genes associated with ER+ tumours.[9] The evidence suggests that the associations at genes such as *FGFR2* and *8q24* are stronger for ER+ tumours and there is reportedly greater *FGFR2* expression in ER+ cell lines. Genetic analysis of larger samples of early-onset cases, stratified by tumour sub-types, is essential to fully comprehend the heterogeneity in phenotype-genotype associations and the degree to which early and late-onset disease may have different genetic backgrounds.

ACKNOWLEDGEMENTS

The work was funded by the Breast Cancer Campaign.

1 Hunter, D. J., Kraft, P., Jacobs, K. B., Cox, D. G., Yeager, M., Hankinson, S. E. *et al.* A genome-wide association study identifies alleles in FGFR2 associated with risk of sporadic postmenopausal breast cancer. *Nat. Genet.* **39,** 870–874 (2007).

2 Easton, D. F., Pooley, K. A., Dunning, A. M., Pharoah, P. D., Thompson, D., Ballinger, D. G. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447,** 1087–1093 (2007).

3 Gold, B., Kirchhoff, T., Stefanov, S., Lautenberger, J., Viale, A., Garber, J. *et al.* Genome-wide association study provides evidence for a breast cancer risk locus at 6q22.33. *Proc. Natl Acad. Sci.* **105,** 4340–4345 (2008).

4 Turnbull, C., Ahmed, S., Morrison, J., Pernet, D., Renwick, A., Maranian, M. *et al.* Genome-wide association study identifies five new breast cancer susceptibility loci. *Nat. Genet.* **42,** 504–507 (2010).

5 Stacey, S. N., Manolescu, A., Sulem, P., Rafnar, T. & Gudmundsson, J. Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.* **39,** 865–869 (2007).

6 Ahmed, S., Thomas, G., Ghoussaini, M., Healey, C. S., Humphreys, M. K., Platte, R. *et al.* Newly discovered breast cancer susceptibility loci on 3p24 and 17q23.2. *Nat. Genet.* **41,** 585–590 (2009).

7 Cox, A., Dunning, A. M., Garcia-Closas, M., Balasubramanian, S. & Reed, M. W. A common coding variant in CASP8 is associated with breast cancer risk. *Nat. Genet.* **39,** 352–358 (2007).

8 Fletcher, O., Johnson, N., Gibson, L., Coupland, B., Fraser, A., Leonard, A. *et al.* Association of genetic variants at 8q24 with breast cancer risk. *Cancer Epidemiol. Biomarkers Prev.* **17,** 702–705 (2008).

9 Garcia-Closas, M. & Chanock, S. Genetic susceptibility loci for breast cancer by estrogen receptor status. *Clin. Cancer Res.* **14,** 8000–8009 (2008).

10 Stacey, S. N., Manolescu, A., Sulem, P., Thorlacius, S. & Gudjonsson, S. A. Common variants on chromosome 5p12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.* **40,** 703–706 (2008).

11 Thomas, G., Jacobs, K. B., Kraft, P., Yeager, M., Wacholder, S., Cox, D. G. *et al.* A multistage genome-wide association study in breast cancer identifies two new risk alleles at 1p11.2 and 14q24.1 (RAD51L1). *Nat. Genet.* **41,** 579–584 (2009).

12 Zheng, W., Long, J., Gao, Y- T., Li, C., Zheng, Y., Xiang, Y. B. *et al.* Genome-wide association study identifies a new breast cancer susceptibility locus at 6q25.1. *Nat. Genet.* **41,** 324–328 (2009).

13 Zeggini, E., Scott, L. J., Saxena, R., Voight, B. F., Marchini, J. L., Hu, T. *et al.* Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat. Genet.* **40,** 638–645 (2008).

14 Maniatis, N., Collins, A. & Morton, N. E. Effects of single SNPs, haplotypes, and whole-genome LD maps on accuracy of association mapping. *Genet. Epidemiol.* **31,** 179–188 (2007).

15 Politopoulos, I., Gibson, J., Tapper, W., Ennis, S., Eccles, D. & Collins, A. Genome-wide association of breast cancer: composite likelihood with imputed genotypes. *Eur. J. Hum. Genet.* **19,** 194–199 (2011).

16 Eccles, D., Gerty, S., Simmonds, P., Hammond, V., Ennis, S., Altman, D. G. *et al.* Prospective study of outcomes in sporadic versus hereditary breast cancer (POSH): study protocol. *BMC Cancer* **7,** 160 (2007).

17 Wellcome Trust Case Control Consortium (WTCCC), Australo-Anglo-American Spondylitis Consortium (TASC). Association scan of 14 500 nonsynonymous SNPs in four diseases identifies autoimmunity variants. *Nat. Genet.* **39,** 1329–1337 (2007).

18 Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A., Bender, D. *et al.* PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am. J. Hum. Genet.* **81,** 559–575 (2007).

19 Collins, A. & Lau, W. CHROMSCAN: genome-wide association using a linkage disequilibrium map. *J. Hum. Genet.* **53,** 121–126 (2008).

20 Maniatis, N., Collins, A., Xu, C- F., McCarthy, L. C., Hewett, D. R., Tapper, W. *et al.* The first linkage disequilibrium (LD) maps: delineation of hot and cold blocks by diplotype analysis. *Proc. Natl Acad. Sci. USA* **99,** 2228–2233 (2002).

21 Gibson, J., Tapper, W., Cox, D., Zhang, W., Pfeufer, A., Gieger, C. *et al.* A multimetric approach to analysis of genome-wide association by single markers and composite likelihood. *Proc. Natl Acad. Sci. USA* **105,** 2592–2597 (2008).

22 Lau, W., Kuo, T- Y., Tapper, W., Cox, S. & Collins, A. Exploiting large scale computing to construct high resolution linkage disequilibrium maps of the human genome. *Bioinformatics* **23,** 517–519 (2007).

23 Tapper, W., Collins, A., Gibson, J., Maniatis, N., Ennis, S. & Morton, N. E. A map of the human genome in linkage disequilibrium units. *Proc. Natl Acad. Sci. USA* **102,** 11835–11839 (2005).

24 Fisher, R., Immer, F. & Tedin, O. The genetical interpretation of statistics of the third degree in the study of quantitative inheritance. *Genetics* **17,** 107–124 (1932).

25 Stouffer, S. A., Suchman, E. A., DeVinney, L. C., Star, S. A. & Williams, R. M. J. *Adjustment during Army Life: The American Soldier*, vol. 1Princeton University Press: Princeton, (1949).

26 Whitlock, M. C. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J. Evol. Biol.* **18,** 1368–1373 (2005).

27 Tapper, W., Collins, A. & Morton, N. Mapping a gene for rheumatoid arthritis on chromosome 18q21. *BMC Proc.* **1,** S18 (2007).

28 Wasserman, N. F., Aneas, I. & Nobrega, M. A. An 8q24 gene desert variant associated with prostate cancer risk confers differential *in vivo* activity to a MYC enhancer. *Genome Res.* **20,** 1191–1197 (2010).

29 Couch, F. & Wang, X. Genome-wide association studies identify new breast cancer susceptibility genes. *Curr. Breast Cancer Rep.* **1,** 131–138 (2009).

30 Blanchard, E. M., Domhan, S., Ma, L., Schwager, C., Ambika, S., Martin, L. A. *et al.* Peripheral blood transcriptomics-based molecular predictors of breast cancer. *J. Clin. Oncol.* **28,** e21018 (2010).

31 Becker, B Combining significance levels. In: Cooper H, Hedges L (eds). *The handbook of research synthesis*, Russell Sage: New York, pp 15–230 (1994).

32 Visel, A., Rubin, E. M. & Pennacchio, L. A. Genomic views of distant-acting enhancers. *Nature* **461,** 199–205 (2009).

Supplementary Information accompanies the paper on Journal of Human Genetics website (http://www.nature.com/jhg)