AD_____

AWARD NUMBER: W81XWH-05-1-0204

TITLE: Identification, Characterization and Clinical Development of the New Generation of Breast Cancer Susceptibility Alleles

PRINCIPAL INVESTIGATOR: Nazneen Rahman, M.D., Ph.D.

CONTRACTING ORGANIZATION: The Institute of Cancer Research
London SW7 3RP; United Kingdom

REPORT DATE: March 2008

TYPE OF REPORT: Revised Annual

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE	2. REPORT TYPE	3. DATES COVERED
March 2008	Revised Annual	1 March 2007 – 29 February 2008

4. TITLE AND SUBTITLE

Identification, Characterization and Clinical Development of the New Generation of Breast Cancer Susceptibility Alleles

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-05-1-0204

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Nazneen Rahman, M.D., Ph.D.

E-Mail: nazneen.rahman@icr.ac.uk

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUM

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
The Institute of Cancer Research
London SW7 3RP, United Kingdom

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Breast cancer is a common disease in women but the causes are still largely unknown. There is considerable evidence to suggest that genetic factors play an important role in causing breast cancer, but the genes involved in the majority of breast cancers are currently unknown. Our aim is to identify genetic factors that increase the chance of breast cancer occurring. We have collected clinical information and samples from over 1500 breast cancer families. We compare the frequency of genetic factors in these cases with control women without breast cancer. Within the last few years we have used this new strategy to identify three new intermediate breast cancer predisposition genes, ATM, BRIP1 and PALB2, and have contributed to genome-wide association studies to identify six low-penetrance breast cancer susceptibility variants. Together these account for ~5% of excess risk of breast cancer. We are have also directly evaluated 15,000 coding genetic variants and aim to extend these studies to undertake whole genome resequencing in breast cancer cases in the near future.

15. SUBJECT TERMS
Breast cancer; genetic predisposition; cancer gene.

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			USAMRMC
U	U	U	UU	48	19b. TELEPHONE NUMBER *(include area code)*

Table of Contents

INTRODUCTION

Breast cancer is a common disease in women but the causes are still largely unknown. There is considerable evidence to suggest that genetic factors play an important role in causing breast cancer. In the last decade considerable progress has been made and two major breast cancer genes, *BRCA1* and *BRCA2*, have been identified (Rahman and Stratton, 1998). These genes carry a high risk of breast cancer but only account for a very small proportion of breast cancer families. Weaker genes are likely to be involved in the majority of familial breast cancers and some breast cancer cases without a family history of the disease, but relatively few have been identified (Antoniou and Easton, 2003; Meijers-Heijboer et al. 2002).

Our aim is to identify and characterize the genetic factors that increase the chance of breast cancer occurring. We have collected clinical information and samples from over 2000 breast cancer families. We first characterized these for the known breast cancer genes, *BRCA1* and *BRCA2*, with particular emphasis on clarifying the contribution and nature of large rearrangements of these genes, which have been identified in some familial breast cancer pedigrees and which are not identifiable by gene sequencing. We have then proceeded to try to identify new genes, by comparing the frequency of genetic factors in these cases with control women without breast cancer. Initially, we have been focusing on analyzing genes that we suspect may have a role in breast cancer, because they are related to known breast cancer genes. This has resulted in our identification of four, intermediate penetrance genes, *ATM*, *CHEK2*, *BRIP1*, *PALB2*, which confer risks of 2-3 fold (refs – references submitted with last years report). Within the last few years we have also been involved in large-scale international studies to use genome-wide tag SNP analyses of 100s of thousands of common variants in breast cancer cases and controls to identify variants associated with very low risks (<1.3 fold) (Easton et al 2007; Cox et al 2007; Stratton and Rahman 2008, see attached papers). We have

directly analysed 15,000 non-synonymous SNPs in 1000 familial breast cancer cases and 1500 controls. This did not provide conclusive evidence that any are associated with breast cancer, confirming the emerging impression that one cannot predict which variants are associated with breast cancer and that whole-genome (rather than targeted) strategies will be required to maximize the harvest of breast cancer susceptibility alleles (Wellcome Trust Case-control Consortium 2008, see attached paper).

Over the future course of the study we therefore plan to extend the genome-wide tag SNP approach in larger series to identify further common, low penetrance susceptibility alleles and we will use emerging whole genome-resequencing technologies to analyze every gene. If we find any variants that are more frequent in breast cancer cases than controls, it suggests that they may be involved in causing breast cancer. We will evaluate these variants in further cases and controls to prove an association with breast cancer and to define the risk and outcomes of carrying the genetic variant(s).

BODY

As part of the program of work we defined five tasks. The progress towards the tasks is outlined in detail below.

Task 1: Evaluate the contribution of BRCA1 and BRCA2 exonic deletions and duplications to breast cancer susceptibility.

We have undertaken analyses for genomic exonic deletions and duplications of *BRCA1* and *BRCA2* in 1500 familial breast cancer cases from separate pedigrees in which mutations of these genes have been excluded. We use a simple, cost-effective copy number analysis technique, multiplex ligation-dependent probe amplification (Schouten et al. 2002; Bunyan et al. 2004).

This analysis has resulted in the identification of genomic duplication / deletion abnormalities in ~ 4% breast cancer families.

Our analyses have demonstrated that:

- MLPA is a cheap, high-throughput and robust technique for copy-number variations, in most situations.
- MLPA should be undertaken in addition to sequencing in all breast cancer families.
- Certain probes show inter-assay variability. We have informed the manufacturers of this and the probes have been replaced.
- Single exon deletions must be further investigated and confirmed – firstly by sequencing to exclude a small exonic mutation under the probe, and if this is normal, by another copy-number assay such as quantitative PCR.
- The clinical features and risks of cancer are the same for families with genomic deletions / duplications as for intragenic mutations.

This strategy is being followed in diagnostic services throughout the UK and in many places internationally.

This project is now complete.

Task 2. Perform familial case-control analyses of in DNA repair genes in familial breast cancer cases, Months 1-36:

a) *Complete identification of coding SNPs by full gene screening of ~50 DNA repair genes in 96 non-BRCA1/2 familial breast cancer cases.*
b) *Analyse all non-synonymous coding SNPs identified in (a) in 500 additional non-BRCA1/2 familial breast cancer cases and 500 controls.*
c) *Analyse SNPs that show positive association with breast cancer in (b) in 10,000 unselected breast cancer cases and 10,000 controls.*

We have altered the design of our study to take advantage of technical improvements, more competitive pricing and an international consortium of ~ 30,000 cases and 30,000 controls (Breast Cancer Association Consortium, BCAC) that we are part of and that has been set-up to evaluate variants. This has allowed us to combine Tasks 2 and Task 4 (Identification genome-wide familial case-control analyses) as follows:

- We have identified 114 non-synonymous coding single nucleotide polymorphisms (SNPs) in DNA repair genes through our sequencing of DNA repair genes in 96 *BRCA1/2* negative cases Probes were successfully designed for 92 of these.
- We included these 92 probes in an array that also included 14,389 non-synonymous coding SNPs that were available from the databases.
- We analysed the 14471 SNPs in 864 familial breast cancer cases and 1498 controls. These results have identified a number of interesting candidates that we are now pursuing. The overall results of these analyses combined with similar analysis in three other diseases has now been published (Wellcome Trust Case-Control Consortium, 2008)

- We were part of a similar designed complementary study using genome-wide tag SNPs rather than non-synonymous SNPs (i.e. targeting common variation rather than potentially functional variants) in which successfully identified 5 new common breast cancer susceptibility variants. This study was undertaken using a 3 stage approach, initiating with a panel of 266,722 SNPs, selected to tag known common variants across the entire genome and genotyped in 408 breast cancer cases and 400; in the second stage second stage 12,711 SNPs were selected based on the significance of the difference in genotype frequency between cases and controls, genotyped in a further 3,990 invasive breast cancer cases and 3,916 controls; and in the third stage 30 of the most significant SNPs were tested in 22 additional case-control studies, comprising 21,860 cases of invasive breast cancer, 988 cases of carcinoma in situ and 22,578 controls. Five SNPS occurring within genes, or LD blocks containing genes, were identified with a combined significance level of $P < 10^{-7}$. Four of the SNPS identified occur in plausible causative genes (*FGFR2, TNRC9, MAP3K1 and LSP1*). Full results were published in Nature, (Easton et al. 2007, attached) and we provide here a summary of the per allele odds ratios by age at breast cancer diagnosis in stage 3, for the five independent SNPs reaching $p<10^{-7.}$

SNP rs number	<40	49-49	50-59	60+	p-trend
rs2981582	1.39 (1.23-1.56)	1.24 (1.16-1.34)	1.21 (1.15-1.28)	1.26 (1.20-1.32)	0.40
rs3803662	1.36 (1.16-1.60)	1.26 (1.16-1.36)	1.22 (1.15-1.29)	1.20 (1.14-1.26)	0.13
rs8051542	1.11 (0.97-1.27)	1.17 (1.08-1.27)	1.17 (1.11-1.24)	1.08 (1.03-1.14)	0.13
rs13281615	1.06 (0.91-1.23)	1.07 (0.99-1.16)	1.14 (1.08-1.20)	1.11 (1.06-1.17)	0.47
rs3817198	1.10 (0.96-1.27)	1.14 (1.05-1.24)	1.05 (1.00-1.11)	1.06 (1.01-1.11)	0.21

- As the common low penetrance breast cancer susceptibility variants appear to be embodied in non-coding rather coding variants we have altered the design of this experiment. We next plan to undertake a larger scale genome-wide tag SNP search (that we are leading) using 4000 familial breast cancer cases and 4000 controls. This is 10x larger than the Easton et al experiment and will be completed in the next 18 months.

Task 3. Characterise the histopathology and immunohistochemistry of familial breast cancer. Months 12-36:

a) *Perform detailed pathological review and immunohistochemical analysis of at least 150 non-BRCA1/2 familial breast cancers.*
b) *Compare pathology and immunohistochemistry of non-BRCA1/2 familial cancers, BRCA1 cancers, BRCA2 cancers and unselected breast cancers.*
c) *Define pathological / immunohistochemical characteristics of non-BRCA1/2 cancers which may allow stratification into subgroups that facilitate identification of underlying susceptibility alleles.*

- Within the last year we have identified three new breast cancer predisposition genes (see below). We are therefore focusing on obtaining and characterizing tumors from mutation carriers of these new genes.
- We are undertaking detailed pathology, immunohistochemistry and loss of heterozygosity analyses to define the tumor characteristics associated with the *ATM*, *BRIP1* and *PALB2* mutations.

Task 4. Perform genome-wide familial case-control analyses of non-synonymous coding SNPs, Months 12-48:

a) *Analyse ~30,000 non-synonymous coding SNPs (at least 1 from every gene) in 400 non-BRCA1/2 familial cases and 400 controls.*
a) *Evaluate top 5% (1500 SNPs) in 800 cases and 800 controls.*

We have undertaken the first phase of this task as outlined above under Task 2. We have been able to increase the size of the study at the same cost, greatly improving the power to detect true associations, due to methodological advancements. The results of the study are now published and we are focusing on analyses of tag SNPs as outlined above.
This project is complete.

Task 5. Identify low penetrance breast cancer susceptibility alleles, Months 36-60:

a) *Evaluate top 30-50 SNPs identified in Task 4 in 10,000 unselected breast cancer cases and 10,000 controls to identify which are truly associated with breast cancer and to determine the risks and phenotype in families and isolated breast cancer.*
b) *Evaluate novel breast cancer susceptibility alleles in BRCA1 / BRCA2 / CHEK2* 1100delC families to determine whether they modify or interact with these genes in breast cancer.*

- We have been undertaking an additional approach to identification of low penetrance breast cancer genes: mutational screening of candidate genes in familial case-control analysis. We have been focusing on DNA repair genes that interact with the known breast cancer genes. In 2006 we completed two of these studies which demonstrate that mutations in *ATM* and *BRIP1* (also known as *FANCJ*) are lower penetrance breast cancer

susceptibility alleles, ~doubling the breast cancer (Renwick et al. 2006; Seal et al. 2006 – papers sent in last years report).

- Through analyses of Fanconi anemia (part of my childhood cancer research) we identified that biallelic *PALB2* mutations cause a new subtype of Fanconi anemia FA-N, which is very similar to FA-D1 which is caused by biallelic *BRCA2* mutations (Reid et al 2007, see attached paper). This raised the possibility that monoallelic *PALB2* mutations might be associated with increased risk of breast cancer, which we were able to demonstrate using our familial case-control strategy (Rahman et al. 2007, paper sent in last years report).
- Over the last year several new DNA repair genes that are plausible breast cancer susceptibility genes have been identified. Moreover although our initial survey of DNA repair genes (started 5 years ago) was highly successful, identifying 4 new genes, it was underpowered analyzing 88 samples. We are therefore embarking on a new round of full gene screening of DNA repair genes to identify truncating variants that may be acting as rare, intermediate breast cancer susceptibility genes (summarized in Stratton and Rahman, 2008 see attached).
- We are also investigating how mutations in these genes interact with *BRCA1* and *BRCA2* mtuations by evaluating their prevalence in *BRCA1/2* mutation carriers

- **KEY RESEARCH ACCOMPLISHMENTS**

1) We have identified three, new, intermediate-penetrance breast cancer predisposition genes, *ATM*, *BRIP1* and *PALB2* (Renwick et al. 2006; Seal et al, 2006; Rahman et al. 2007).
2) We have analysed 14,471 non-synonymous coding SNPs in 864 familial BRCA1/2-negative breast cancer cases and 1498 controls (WTCCC and TASC, 2007).
3) We have participated in an international genome-wide tag SNP association study that has identified 5 low-penetrance breast cancer predisposition alleles (Easton et al 2007)
4) I was invited to write a perspective by the leading genetics journal (Nature Genetics) on the Emerging Landscape of Breast Cancer Susceptibility and the implications for other diseases, emphasizing our influential position in this arena. (Stratton and Rahman, 2008)

REPORTABLE OUTCOMES

We have published four research papers and one Perspective in Nature Genetics, one paper in Nature, a review in Human Molecular Genetics and a review in Oncogene.

CONCLUSION

We have had another exceptionally productive year. We have made substantial progress towards our goals and emerging technologies promise further advances and will allow us to considerably improve the power of the studies at similar cost. We are ensuring that our unique sample resources are being used for maximum benefit by participating in International consortia analyses as well as undertaking our own research. We anticipate that rest of the programme will proceed on course and are hopeful of further discoveries.

REFERENCES

Ahmed M and Rahman N (2006) ATM and breast cancer susceptibility Oncogene reviews 25:5906-5911

Antoniou AC and Easton DF (2003). Polygenic inheritance of breast cancer: Implications for design of association studies. Genet. Epidemiol. 25:190-202.

Cox A, Dunning AM, Garcia-Closas M, Balasubramanian S, Reed MWR, Pooley KA, Scollen S, Baynes C, Ponder BAJ, Chanock S, Lissowska J, Brinton L, Peplonska B, Southey MC, Hopper JL, McCredie MRE, Giles GG, Fletcher O, Johnson N, dos Santos Silva I, Gibson L, Bojesen SE, Nordestgaard BG, Axelsson CK, Torres D, Hamann U, Justenhoven C, Brauch H, Chang-Claude J, Kropp S, Risch A, Wang-Gohrke S, SchŸrmann P, Bogdanova N, Dšrk T, Fagerholm R, Aaltonen K, Blomqvist C, Nevanlinna H, Seal S, Renwick A, Stratton MR, Rahman N, Sangrajrang S, Hughes D, Odefrey F, Brennan P, Spurdle AB, Chenevix-Trench G, Beesley J, The Katherine Cuningham Foundation Consortium for Research into Familial Breast Cancer, Mannermaa A, Hartikainen J, Kataja V, Kosma V-M, Couch FJ, Olson JE, Goode EL, Broeks A, Schmidt MK, Hogervorst FBL, Van't Veer LJ, Kang D, Yoo K-Y, Noh D-Y, Ahn S-H, WedrŽn S, Hall P, Low Y-L, Liu J, Milne RL, Ribas G, Gonzalez-Neira A, Benitez J, Sigurdson AJ, Stredrick DL, Alexander BH, Struewing JP, Pharoah PDP and Douglas F. Easton, on behalf of the Breast Cancer Association Consortium (2007). A common coding variant in CASP8 is associated with breast cancer risk. Nature Genetics 39:352-358

Easton DF, Polley KA, Dunning AM, Pharaoh PD, Thompson D, Ballinger DG, Streuwing JP, Morrison J, Field H, Luben R, Wareham N, Ahmed S, Healey CS, Bowman R, The SEARCH collaborators, Luccarini C, Conroy D, Shah M, Munday H, Jordan C, Perkins B, West J, Redman K, Meyer KB, Haiman CA, Kolonel LK, Henderson BE, Le Marchand L, Brennan P, Sangrajrang S, Gaborieau V, Odefrey F, Shen CY, Wu PE, Wang HC, Eccles D, Evans DG, Peto J, Fletcher O, Johnson N, Seal S, Stratton MR, Rahman N, Chenevix-Trench G, Bojesen SE, Nordestgaard BG, Axelsson CK, Garcia-Closas M, Brinton L, Chanock S, Lissowska J, Peplonska B, Nevanlinna H, Fagerholm R, Eerola H, Kang D, Yoo KY, Noh DY, Ahn SH, Hunter DJ, Hankinson SE, Cox DG, Hall P, Wedren S, Liu J, Low YL, Bogdanova N, Schurmann P, Dork T, Tollenaar RA, Jacobi CE, Devilee P, Klijn JG, Sigurdson AJ, Doody MM, Alexander BH, Zhang J, Cox A, Brock IW, Macpherson G, Reed MW, Couch FJ, Goode EL, Olson JE, Meijers-Heijboer H, van den Ouweland A, Utterlinden A, Rivadeneira F, Milne RL, Ribas G, Gonzalez-Neira A, Benitex J, Hopper JL, McCredie M, Southey M, Giles GG, Schroen C, Justenhoven C, Brauch H, Hamann U, Ko YD, Spurdel AB, Beesley J, Chen X, IConFab, Aghmesheh M, Amor D, Andrews L, Antill Y, Armes J, Armitage S, Arnold L, Balleine R, Begley G, Beilby J, Bennett I, Bennett B, Berry G, Blackburn A, Brennan M, Brown M, Buckley M, Burke J, Butow P, Byron K, Callen D, Campbell I, Day NE, Cox DR, Ponder BA (2007). Genome-wide association study identifies novel breast cancer susceptibility loci. Nature 447:1087-1093

Meijers-Heijboer H, van den Ouweland A, Klijn J, Wasielewski M, de Snoo A, Oldenburg R, Hollestelle A, Houben M, Crepin E, van Veghel-Plandsoen M, Elstrodt F, van Duijn C, Bartels C, Meijers C, Schutte M, McGuffog L, Thompson D, Easton DF, Sodha N, Seal S, Barfoot R, Mangion J, Chang-Claude J, Eccles D, Eeles RA, Evans DG, Houlston RS, Murday VA, Narod S, Peretz T, Peto J, Phelan C, Zhang H, Szabo C, Devilee P, Goldgar D, Futreal PA, Nathanson K, Weber B, Rahman N, Stratton MR. Low-penetrance susceptibility to breast cancer due to CHEK2*1100delC in noncarriers of BRCA1 or BRCA2 mutations (2002). Nature Genetics 31:55-59.

Melchor L, Honrado E, Huang J, Alvarez S, Naylor TL, Garcia MJ, Osorio A, Blesa D, Stratton MR, Weber BL, Cigudosa JC, Rahman N, Nathanson KL, Benitez J (2007) Estrogen receptor status could modulate the genomic pattern in familial and sporadic breast cancer (2007) Clin Cancer Res 13:7305-7313

Rahman N and Stratton MR. Breast cancer susceptibility genes (1998). Annual Review of Genetics 32:95-121.

Rahman N, Seal S, Thompson D, Kelly P, Renwick A, Elliott A, Reid A, Spanova K, Barfoot R, Chagtai T, Jayatilake H, McGuffog L, Hanks S, Evans DG, Eccles D, The Breast Cancer Susceptibility Collaboration (UK), Easton DF and Stratton MR (2007). PALB2, which encodes a BRCA2 interacting protein, is a breast cancer susceptibility gene. Nature Genetics 39:165-167

Reid S, Schindler D, Hanenberg H, Barker K, Hanks S, Kalb R, Neveling K, Kelly P,Seal S, Freund M, Wurm M, Batish SD, Lach FP, Yetgin S, Neitzel H, Ariffin H, Tischkowitz M, Mathew CG, Auerbach AD, and Rahman N (2007) Biallelic mutations in PALB2, which encodes a BRCA2 interacting protein, cause Fanconi anemia subtype FA-N and predispose to childhood cancer. Nature Genetics 39:162-164

Renwick A, Thompson D, Seal S, Kelly P, Chagtai T, Ahmed M, North B, Jayatilake H, Barfoot R, McGuffog L, Evans DG, Eccles D, The Breast Cancer Susceptibility Collaboration (UK), Easton DF, Stratton MR and Rahman N (2006) ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. Nature Genetics 38:873-875

Schouten JP, McElgunn CJ, Waaijer R, Zwijnenburg D, Diepvens F, Pals G (2002). Relative quantification of 40 nucleic acid sequences by multiplex ligation-dependent probe amplification. Nucleic Acids Res. 30:e57.

Seal S, Thompson D, Renwick A, Elliott A, Kelly P, Barfoot R, Chagtai T, Jayatilaek H, Ahmed M, Spanova K, North B, McGuffog L, Evans DG, Eccles D, Easton DF, Stratton MR and Rahman N (2006) Truncating mutations in BRIP1 are low penetrance breast cancer susceptibility alleles. Nature Genetics 38:1239-1241

Stratton MR and Rahman N (2008) The emerging landscape of breast cancer susceptibility Nature Genetics 40:17-22

Wellcome Trust Case Control Consortium (WTCCC) and The Australo-Angle-American Spondylitis Consortium Association (TASC) scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants (2007) Nature Genetics 39:1329-1338

APPENDICES

Four publications – attached

ARTICLES

Genome-wide association study identifies novel breast cancer susceptibility loci

Douglas F. Easton[1], Karen A. Pooley[2], Alison M. Dunning[2], Paul D. P. Pharoah[2], Deborah Thompson[1], Dennis G. Ballinger[3], Jeffery P. Struewing[4], Jonathan Morrison[2], Helen Field[2], Robert Luben[5], Nicholas Wareham[5], Shahana Ahmed[2], Catherine S. Healey[2], Richard Bowman[6], the SEARCH collaborators[2]*, Kerstin B. Meyer[7], Christopher A. Haiman[8], Laurence K. Kolonel[9], Brian E. Henderson[8], Loic Le Marchand[9], Paul Brennan[10], Suleeporn Sangrajrang[11], Valerie Gaborieau[10], Fabrice Odefrey[10], Chen-Yang Shen[12], Pei-Ei Wu[12], Hui-Chun Wang[12], Diana Eccles[13], D. Gareth Evans[14], Julian Peto[15], Olivia Fletcher[16], Nichola Johnson[16], Sheila Seal[17], Michael R. Stratton[17,18], Nazneen Rahman[17], Georgia Chenevix-Trench[19], Stig E. Bojesen[20], Børge G. Nordestgaard[20], Christen K. Axelsson[21], Montserrat Garcia-Closas[22], Louise Brinton[22], Stephen Chanock[23], Jolanta Lissowska[24], Beata Peplonska[25], Heli Nevanlinna[26], Rainer Fagerholm[26], Hannaleena Eerola[26,27], Daehee Kang[28], Keun-Young Yoo[28,29], Dong-Young Noh[28], Sei-Hyun Ahn[30], David J. Hunter[31,32], Susan E. Hankinson[32], David G. Cox[31], Per Hall[33], Sara Wedren[33], Jianjun Liu[34], Yen-Ling Low[34], Natalia Bogdanova[35,36], Peter Schürmann[36], Thilo Dörk[36], Rob A. E. M. Tollenaar[37], Catharina E. Jacobi[38], Peter Devilee[39], Jan G. M. Klijn[40], Alice J. Sigurdson[41], Michele M. Doody[41], Bruce H. Alexander[42], Jinghui Zhang[4], Angela Cox[43], Ian W. Brock[43], Gordon MacPherson[43], Malcolm W. R. Reed[44], Fergus J. Couch[45], Ellen L. Goode[45], Janet E. Olson[45], Hanne Meijers-Heijboer[46,47], Ans van den Ouweland[47], André Uitterlinden[48], Fernando Rivadeneira[48], Roger L. Milne[49], Gloria Ribas[49], Anna Gonzalez-Neira[49], Javier Benitez[49], John L. Hopper[50], Margaret McCredie[51], Melissa Southey[50], Graham G. Giles[52], Chris Schroen[53], Christina Justenhoven[54], Hiltrud Brauch[54], Ute Hamann[55], Yon-Dschun Ko[56], Amanda B. Spurdle[19], Jonathan Beesley[19], Xiaoqing Chen[19], kConFab[57]*, AOCS Management Group[19,57]*, Arto Mannermaa[58,59], Veli-Matti Kosma[58,59], Vesa Kataja[58,60], Jaana Hartikainen[58,59], Nicholas E. Day[5], David R. Cox[3] & Bruce A. J. Ponder[2,7]

Breast cancer exhibits familial aggregation, consistent with variation in genetic susceptibility to the disease. Known susceptibility genes account for less than 25% of the familial risk of breast cancer, and the residual genetic variance is likely to be due to variants conferring more moderate risks. To identify further susceptibility alleles, we conducted a two-stage genome-wide association study in 4,398 breast cancer cases and 4,316 controls, followed by a third stage in which 30 single nucleotide polymorphisms (SNPs) were tested for confirmation in 21,860 cases and 22,578 controls from 22 studies. We used 227,876 SNPs that were estimated to correlate with 77% of known common SNPs in Europeans at $r^2 > 0.5$. SNPs in five novel independent loci exhibited strong and consistent evidence of association with breast cancer ($P < 10^{-7}$). Four of these contain plausible causative genes (*FGFR2*, *TNRC9*, *MAP3K1* and *LSP1*). At the second stage, 1,792 SNPs were significant at the $P < 0.05$ level compared with an estimated 1,343 that would be expected by chance, indicating that many additional common susceptibility alleles may be identifiable by this approach.

Breast cancer is about twice as common in the first-degree relatives of women with the disease as in the general population, consistent with variation in genetic susceptibility to the disease[1]. In the 1990s, two major susceptibility genes for breast cancer, *BRCA1* and *BRCA2*, were identified[2,3]. Inherited mutations in these genes lead to a high risk of breast and other cancers[4]. However, the majority of multiple case breast cancer families do not segregate mutations in these genes. Subsequent genetic linkage studies have failed to identify further major breast cancer genes[5]. These observations have led to the proposal that breast cancer susceptibility is largely 'polygenic': that is, susceptibility is conferred by a large number of loci, each with a small effect on breast cancer risk[6]. This model is consistent with the observed patterns of familial aggregation of breast cancer[7]. However, progress in identifying the relevant loci has been slow. As linkage studies lack power to detect alleles with moderate effects on risk, large case-control association studies are required. Such studies have identified variants in the DNA repair genes *CHEK2*, *ATM*, *BRIP1* and *PALB2* that confer an approximately twofold risk of breast cancer, but these variants are rare in the population[8–14]. A recent study has shown that a common coding variant in *CASP8* is associated with a moderate reduction in breast cancer risk[15]. After accounting for all the known breast cancer loci, more than 75% of the familial risk of the disease remains unexplained[16].

Recent technological advances have provided platforms that allow hundreds of thousands of SNPs to be analysed in association studies, thus providing a basis for identifying moderate risk alleles without

Affiliations of the above authors are given at the end of the paper.
*Lists of consortia participants and affiliations appear after author affiliations.

prior knowledge of position or function. It has been estimated that there are 7 million common SNPs in the human genome (with minor allele frequency, m.a.f., >5%)[17]. However, because recombination tends to occur at distinct 'hot-spots', neighbouring polymorphisms are often strongly correlated (in 'linkage disequilibrium', LD) with each other. The majority of common genetic variants can therefore be evaluated for association using a few hundred thousand SNPs as tags for all the other variants[18]. We aimed to identify further breast cancer susceptibility loci in a three-stage association study[19]. In the first stage, we used a panel of 266,722 SNPs, selected to tag known common variants across the entire genome[18]. These SNPs were genotyped in 408 breast cancer cases and 400 controls from the UK; data were analysed for 390 cases and 364 controls genotyped for ≥80% of the SNPs. The cases were selected to have a strong family history of breast cancer, equivalent to at least two affected female first-degree relatives, because such cases are more likely to carry susceptibility alleles[20]. Initally, we analysed 227,876 SNPs (85%) with genotypes on at least 80% of the subjects. We estimate that these SNPs are correlated with 58% of common SNPs in the HapMap CEPH/CEU (Utah residents with ancestry from northern and western Europe) samples at $r^2 > 0.8$, and 77% at $r^2 > 0.5$ (mean $r^2 = 0.75$; see Supplementary Fig. 1) (http://www.hapmap.org/)[21]. As expected, coverage was strongly related to m.a.f.: 70% of SNPs with m.a.f. > 10% were tagged at $r^2 > 0.8$, compared with 23% of SNPs with m.a.f. 5–10%. The main analyses were restricted to 205,586 SNPs that had a call rate of 90% and whose genotype distributions did not differ from Hardy–Weinberg equilibrium in controls (at $P < 10^{-5}$).

For the second stage we selected 12,711 SNPs, approximately 5% of those typed in stage 1, on the basis of the significance of the difference in genotype frequency between cases and controls. These SNPs were then genotyped in a further 3,990 invasive breast cancer cases and 3,916 controls from the SEARCH study, using a custom-designed oligonucleotide array. In the main analyses, we considered 10,405 SNPs with call rate of >95% that did not deviate from Hardy–Weinberg equilibrium in controls.

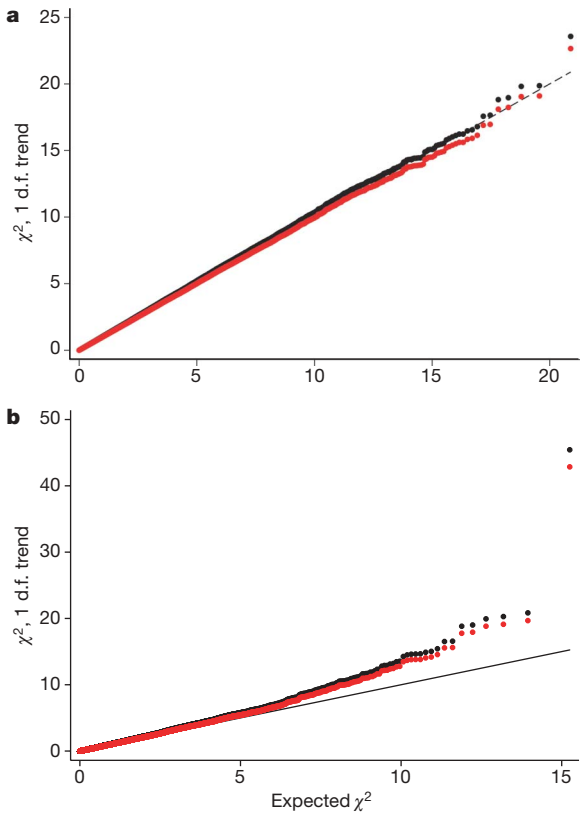

Figure 1 | Quantile–quantile plots for the test statistics (Cochran-Armitage 1 d.f. χ^2 trend tests) for stages 1 and 2. **a**, Stage 1; **b**, stage 2. Black dots are the uncorrected test statistics. Red dots are the statistics corrected by the genomic control method ($\lambda = 1.03$ for stage 1, $\lambda = 1.06$ for stage 2). Under the null hypothesis of no association at any locus, the points would be expected to follow the black line.

Comparison of the observed and expected distribution of test statistics showed some evidence for an inflation of the test statistics in both stage 1 (inflation factor $\lambda = 1.03$, 95% confidence interval (CI) 1.02–1.04) and stage 2 ($\lambda = 1.06$, 95% CI 1.04–1.12), based on the 90% least significant SNPs (Fig. 1). Possible explanations for this inflation include population stratification, cryptic relatedness among subjects, and differential genotype calling between cases and controls. There was evidence for an excess of low call rate SNPs among the most significant SNPs ($P < 0.01$) in stage 1, but not in stage 2, suggesting that some of this effect is a genotyping artefact (Supplementary Table 1). However, the inflation was still present among SNPs with call rate >99% in both cases and controls, possibly reflecting population substructure. We computed 1 degree of freedom (d.f.) association tests for each SNP, combining stages 1 and 2. After adjustment for this inflation by the genomic control method[22], we observed more associations than would have been expected by chance at $P < 0.05$ (Table 1). One SNP (dbSNP rs2981582) was significant at the $P < 10^{-7}$ level that has been proposed as appropriate for genome-wide studies[23].

In the third stage, to establish whether any SNPs were definitely associated with risk, we tested 30 of the most significant SNPs in 22 additional case-control studies, comprising 21,860 cases of invasive breast cancer, 988 cases of carcinoma *in situ* (CIS) and 22,578 controls (Supplementary Table 2). Six SNPs showed associations in stage 3 that were significant at $P \leq 10^{-5}$ with effects in the same direction as in stages 1 and 2 (Table 2, Supplementary Table 3, and Fig. 2). All these SNPs reached a combined significance level of $P < 10^{-7}$ (ranging from 2×10^{-76} to 3×10^{-9}). Of these six SNPs, five were within genes or LD blocks containing genes. SNP rs2981582 lies in intron 2 of *FGFR2* (also known as *CEK3*), which encodes the fibroblast growth factor receptor 2. SNPs rs12443621 and rs8051542 are both located in an LD block containing the 5′ end of *TNRC9* (also known as *TOX3*), a gene of uncertain function containing a tri-nucleotide repeat motif, as well as the hypothetical gene, *LOC643714*. SNP rs889312 lies in an LD block of approximately 280 kb that contains *MAP3K1* (also known as *MEKK*), which encodes the signalling protein mitogen-activated protein kinase kinase kinase 1, in addition to two other genes: *MGC33648* and *MIER3*. SNP rs3817198 lies in intron 10 of *LSP1* (also known as *WP43*), encoding lymphocyte-specific protein 1, an F-actin bundling cytoskeletal protein expressed in haematopoietic and endothelial cells. A further SNP, rs2107425, located just 110 kilobases (kb) from rs3817198, was also identified (overall $P = 0.00002$). rs2107425 is within the *H19* gene, an imprinted maternally expressed untranslated messenger RNA closely involved in regulation of the insulin growth factor gene, *IGF2*. In stage 3, however, rs2107425 was only weakly significant after adjustment for rs3817198 by logistic regression ($P = 0.06$). This suggests that the association with breast cancer risk may be driven by variants in *LSP1* rather than in *H19*. The sixth SNP reaching a combined $P < 10^{-7}$ was rs13281615, which lies on 8q. It is correlated with SNPs in a 110 kb LD block that contains no known

Table 1 | Number of significant associations after stage 2

Level of significance	Observed	Observed adjusted*	Expected	Ratio
0.01–0.05	1,239	1,162	934.3	1.24
0.001–0.01	574	517	347.6	1.49
0.0001–0.001	112	88	53.3	1.65
0.00001–0.0001	16	12	7.0	1.71
<0.00001	15	13	0.96	13.5
All $P < 0.05$	1,956	1,792	1,343.2	1.33

Observed numbers of SNPs associated with breast cancer after stage 2, by level of significance, before and after adjustment for population stratification, and expected numbers under the null hypothesis of no association.
* Adjusted for inflation of the test statistic by the genomic control method.

Table 2 | Summary of results for eleven SNPs selected for stage 3 that showed evidence of an association with breast cancer

rs Number	Gene	Position*	m.a.f.†	Per allele OR (95% CI)	HetOR (95% CI)	HomOR (95% CI)	*P*-trend		
							Stages 1 and 2	Stage3	Combined
rs2981582	*FGFR2*	10q 123342307	0.38 *(0.30)*	1.26 (1.23–1.30)	1.23 (1.18–1.28)	1.63 (1.53–1.72)	4×10^{-16}	5×10^{-62}	2×10^{-76}
rs12443621	*TNRC9/ LOC643714*	16q 51105538	0.46 *(0.60)*	1.11 (1.08–1.14)	1.14 (1.09–1.20)	1.23 (1.17–1.30)	10^{-7}	9×10^{-14}	2×10^{-19}
rs8051542	*TNRC9/ LOC643714*	16q 51091668	0.44 *(0.20)*	1.09 (1.06–1.13)	1.10 (1.05–1.16)	1.19 (1.12–1.27)	4×10^{-6}	4×10^{-8}	10^{-12}
rs889312	*MAP3K1*	5q 56067641	0.28 *(0.54)*	1.13 (1.10–1.16)	1.13 (1.09–1.18)	1.27 (1.19–1.36)	4×10^{-6}	3×10^{-15}	7×10^{-20}
rs3817198	*LSP1*	11p 1865582	0.30 *(0.14)*	1.07 (1.04–1.11)	1.06 (1.02–1.11)	1.17 (1.08–1.25)	8×10^{-6}	10^{-5}	3×10^{-9}
rs2107425	*H19*	11p 1977651	0.31 *(0.44)*	0.96 (0.93–0.99)	0.94 (0.90–0.98)	0.95 (0.89–1.01)	7×10^{-6}	0.01	2×10^{-5}
rs13281615		8q 128424800	0.40 *(0.56)*	1.08 (1.05–1.11)	1.06 (1.01–1.11)	1.18 (1.10–1.25)	2×10^{-7}	6×10^{-7}	5×10^{-12}
rs981782		5p 45321475	0.47 *(0.37)*	0.96 (0.93–0.99)	0.96 (0.92–1.01)	0.92 (0.87–0.97)	8×10^{-5}	0.003	9×10^{-6}
rs30099		5q 52454339	0.08 *(0.39)*	1.05 (1.01–1.10)	1.06 (1.00–1.11)	1.09 (0.96–1.24)	0.003	0.02	0.001
rs4666451		2p 19150424	0.41 *(0.04)*	0.97 (0.94–1.00)	0.98 (0.93–1.02)	0.93 (0.87–0.99)	5×10^{-6}	0.04	6×10^{-5}
rs3803662‡	*TNRC9/ LOC643714*	16q 51143842	0.25 *(0.60)*	1.20 (1.16–1.24)	1.23 (1.18–1.29)	1.39 (1.26–1.45)	3×10^{-12}	10^{-26}	10^{-36}

OR, odds ratio; HetOR, odds ratio in heterozygotes; HomOR, odds ratio in rare homozygotes (relative to common homozygotes); CI, confidence interval.
* Build 36.2 position.
† Minor allele frequency in SEARCH (UK) study. Combined allele frequency from three Asian studies in italics.
‡ rs3803662 was not part of the initial tag SNP set but identified as a result of fine-scale mapping of the *TNRC9/LOC643714* locus and typed in the stage 2 and stage 3 sets (but not the stage 1 set).

genes. The basis of this association therefore remains obscure. This SNP is approximately 130 kb proximal to rs1447295, 60 kb proximal to rs6983267 and 230 kb distal to rs16901979, recently shown to be associated with prostate cancer[24–26].

In addition to the seven SNPs described above, there was evidence of association among the remaining 23 SNPs (global $P = 0.001$ in stage 3). In particular, three SNPs showed some evidence of association in stage 3 ($P < 0.05$, in each case in the same direction as in stages 1 and 2; Table 2). SNPs rs981782 and rs30099 both lie in the centromeric region of chromosome 5. rs4666451 lies on 2p, a region for which some evidence of linkage to breast cancer in families has been reported[5]. The 20 other SNPs showed no evidence of association in stage 3 (global $P = 0.11$), suggesting that most of these associations from stages 1 and 2 were false positives.

FGFR2

The most significantly associated SNP, rs2981582, lies within a 25 kb LD block almost entirely within intron 2 of *FGFR2*. We found no evidence of association with SNPs elsewhere in the gene (Fig. 3a). In an attempt to identify a causal variant, we first identified the 19 common variants (m.a.f. > 0.05) in this block from HapMap CEU data. These were tagged ($r^2 > 0.8$) by 7 SNPs including rs2981582. The additional tag SNPs were genotyped in the SEARCH study cases and controls. Multiple logistic regression analysis of these variants found no additional evidence for association after adjusting for rs2981582. Haplotype analysis of these 7 SNPs indicated that multiple haplotypes carrying the minor (*a*) allele of rs2981582 were associated with an increased risk of breast cancer, implying that the association was being driven by rs2981582 itself or a variant strongly correlated with it (Supplementary Table 4).

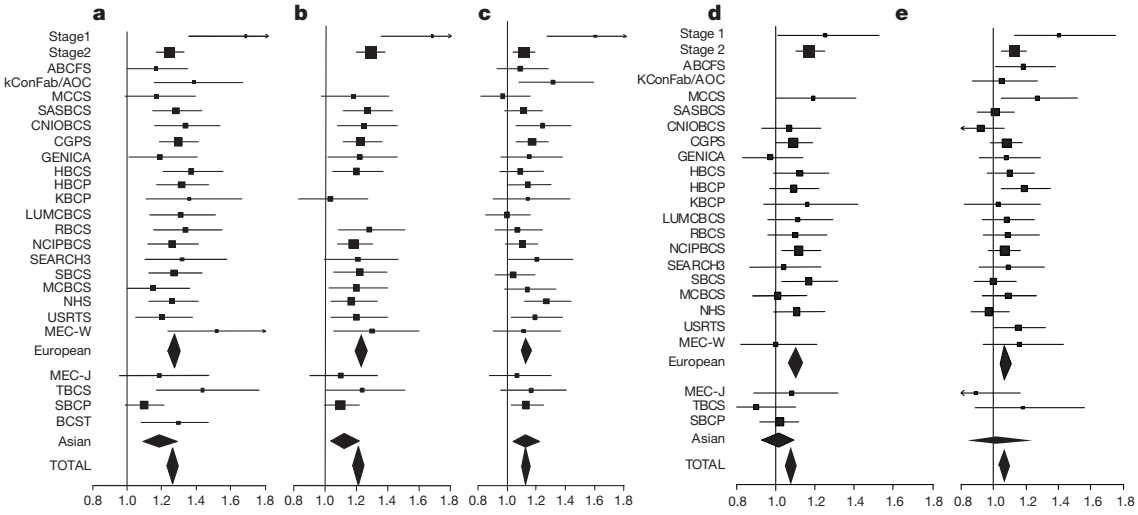

Figure 2 | Forest plots of the per-allele odds ratios for each of the five SNPs reaching genome-wide significance. a, rs2981582; **b**, rs3803662; **c**, rs889312; **d**, rs13281615; and **e**, rs3817198. The *x*-axis gives the per-allele odds ratio. Each row represents one study (see Supplementary Table 2), with summary odds ratios for all European and all Asian studies, and all studies combined. The area of the square for each study is proportional to the inverse of the variance of the estimate. Horizontal lines represent 95% confidence intervals. Diamonds represent the summary odds ratios, with 95% confidence intervals, based on the stage 3 studies only.

Resequencing of this region in 45 subjects of European origin identified 29 variants that were strongly correlated with rs2981582 ($r^2 > 0.6$) (http://cgwb.nci.nih.gov; Fig. 3b and Supplementary Tables 5–8). A subset of 14 variants tagged 27 of these in European ($r^2 > 0.95$) and Asian (Korean) samples ($r^2 > 0.86$). Two variants could not be genotyped reliably. This new tagging set was then genotyped in SEARCH and 3 studies from Asian populations; the Asian studies were included because the LD is weaker, providing greater power to resolve the causal variant (Fig. 3b, left panel). The strongest association was found with rs7895676. On the assumption that there is a single disease-causing allele, we calculated a likelihood for each variant. 21 SNPs (including rs2981582) had a likelihood ratio of <1/100 relative to rs7895676, indicating that none of these are likely to be the causal variant (Supplementary Table 8). Six variants were too strongly correlated for their individual effects to be separated using a genetic epidemiological approach. Functional assays will be required to determine which is causally related to breast cancer risk.

Intron 2 of *FGFR2* shows a high degree of conservation in mammals, and contains several putative transcription-factor binding sites (http://genomequebec.mcgill.ca/PReMod)[27], some of which lie in close proximity to the relevant SNPs. We therefore speculate that the association with breast cancer risk is mediated through regulation of *FGFR2* expression. Of possible relevance is that only three of these variants (rs10736303, rs2981578 and rs35054928) are within sequences conserved across all placental mammals (Fig. 3c and Supplementary Table 8). Of these, the disease associated allele of rs10736303 generates a putative oestrogen receptor (ER) binding site. rs35054928 lies immediately adjacent to a perfect POU domain protein octamer (Oct) binding site. However, multiple splice variants have been reported in *FGFR2*, and differential splicing might provide an alternative mechanism for the association. FGFR2 is a receptor tyrosine kinase that is amplified and overexpressed in 5–10% of breast tumours[28–30]. Somatic missense mutations of *FGFR2* that are likely to be implicated in cancer development have also been demonstrated in primary tumours and cell lines of multiple tumour types (http://www.sanger.ac.uk/genetics/CGP/cosmic/)[30,31].

TNRC9/LOC643714 locus

As two SNPs in the *TNRC9/LOC643714* locus, rs12443621 and rs8051542, both showed convincing evidence of association, we further evaluated this region by genotyping, in the SEARCH set, an additional 19 SNPs tagging 101 common variants within the entire *TNRC9* and *LOC643714* genes, based on the HapMap CEU data. SNPs tagging the coding region of *TNRC9* showed no evidence of association. The strongest association was observed with rs3803662, a synonymous coding SNP of *LOC643714* that lies 8 kb upstream of *TNRC9*. This SNP was therefore genotyped in the stage 3 set (Table 2). Logistic regression analysis indicated that rs3803662 exhibited a stronger association with disease than other SNPs, and the associations with other SNPs were non-significant after adjustment for rs3803662. These results suggest

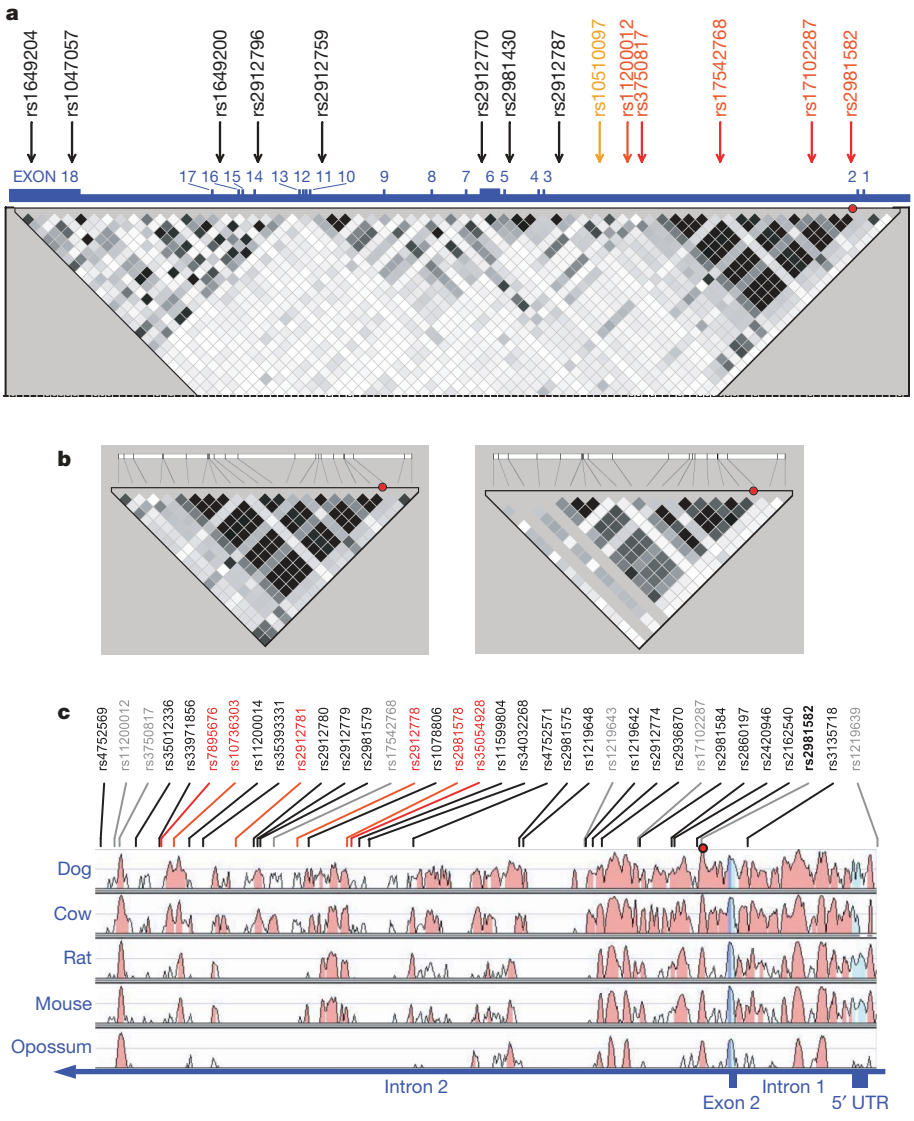

Figure 3 | The *FGFR2* locus. a, Map of the whole *FGFR2* gene, viewed relative to common SNPs on HapMap. The gene is 126 kb long and in reverse 3′–5′ orientation on chromosome 10. Exon positions are illustrated with respect to the 67 SNPs with m.a.f. > 5% in HapMap CEU (therefore the map is not to physical scale). Numbered SNPs are those tested in the genome-wide study. SNPs in black were not significant in stage 1. Those in red were significant at $P < 0.0001$ after stage 2. rs10510097 (orange) was significant in stage 1, but failed quality control in stage 2 owing to deviation from Hardy–Weinberg equilibrium. Squares indicate pairwise r^2 on a greyscale (black = 1, white = 0). Red circle indicates rs2981582. **b**, Resequenced 32 kb region, shown relative to SNPs in CEU with m.a.f. > 5%, showing pairwise LD for SNPs in HapMap CEU (left panel) and JPT/CHB (right panel). Red circle indicates rs2981582, shown in bold black. **c**, Sequence conservation of 32 kb region in five species, relative to human sequence (http://pipeline.lbl.gov/methods.shtml)[35]. Red circle indicates rs2981582. SNPs in grey are those used in the initial tagging of known common HapMap SNPs within the block. SNPs in black are correlated with rs2981582 with $r^2 > 0.6$ in European samples. Six SNPs in red were those consistent with being the causative variant on the basis of the genetic data (not excluded at odds of 100:1 relative to the SNP with the strongest association, rs7895676).

that the causal variant is closely correlated with rs3803662. Four SNPs in the HapMap CEU data (rs17271951, rs1362548, rs3095604 and rs4784227) that span *LOC643714* and the 5′ regulatory regions of *TNRC9* are strongly correlated with rs3803662, and it therefore remains unclear in which gene the causative variant lies. *TNRC9* contains a putative HMG (high mobility group) box motif, suggesting that it might act as a transcription factor.

Pattern of risks

We assessed in more detail, in the stage 3 data, the pattern of the risks associated with the five independent SNPs that reached an overall $P < 10^{-7}$: rs2981582 (*FGFR2*), rs3803662 (*TNRC9/LOC643714*), rs889312 (*MAP3K1*), rs13281615 (8q) and rs3817198 (*LSP1*). For each of these five SNPs, the minor allele in Europeans was associated with an increased risk of breast cancer in a dose-dependent manner, with a higher risk of breast cancer in homozygous than in heterozygous carriers. Simple dominant and recessive models could be rejected for each SNP (all $P = 0.02$ or less). There was a marked difference in allele frequencies between populations, with the risk-associated alleles of rs8051542, rs889312 and rs13281615 being the major allele in Asian populations. The per allele odds ratio associated with rs2981582 was significantly smaller, though still elevated, in the Asian versus European populations ($P = 0.04$ for difference in odds ratio). This difference is consistent with the hypothesis that rs2981582 is not the functional variant at the *FGFR2* locus, and was not seen for SNPs exhibiting stronger evidence in the fine-scale mapping. No other evidence for heterogeneity in the per-allele odds ratio among studies was observed (Fig. 2).

Three of the SNPs (rs2981582, rs3803662 and rs889312) also showed evidence of association with breast CIS (Supplementary Table 9). For rs2981582 and rs3803662, the estimated odds ratios were greater for a diagnosis of breast cancer before age 40 years, but the trends by age were not statistically significant (Supplementary Table 10). There was evidence of an association with family history of breast cancer for three SNPs: for rs2981582 ($P = 0.02$), rs3803662 ($P = 0.03$) and rs13281615 ($P = 0.05$), the susceptibility allele was commoner in women with a first-degree relative with the disease than in those without (Supplementary Table 11). rs2981582 was also associated with bilaterality ($P = 0.02$). The associations with family history and bilaterality are to be expected for susceptibility loci, and are similar to previous observations for alleles in *CHEK2* and *ATM* (refs 10, 12, 14).

Discussion

This study has identified five novel breast cancer susceptibility loci, and demonstrated conclusively that some of the variation in breast cancer risk is due to common alleles. None of the loci we identified had been previously reported in association studies. Most previously identified breast cancer susceptibility genes are involved in DNA repair, and many association studies in breast cancer have concentrated on genes in DNA repair and sex hormone synthesis and metabolism pathways. None of the associations reported here appear to relate to genes in these pathways. It is notable that three of the five loci contain genes related to control of cell growth or to cell signalling, but only one (*FGFR2*) had a clear prior relevance to breast cancer. These results should, therefore, open up new avenues for basic research.

Our results emphasize the critical importance of study size in genetic association studies. It is notable that none of the confirmed associations reached genome-wide significance after stage 1 and only one reached this level after stage 2. As most common cancers have similar familial relative risks to breast cancer, it is likely that similarly large studies will be required to identify common alleles for other cancers. The fine-scale mapping of the *FGFR2* locus demonstrates that, even with a clear association, identification of the causative variant can be extremely problematic. However, the use of studies from multiple populations with different patterns of LD can substantially reduce the number of variants that need to be subjected to functional analysis.

As these susceptibility alleles are very common, a high proportion of the general population are carriers of at-risk genotypes. For example, approximately 14% of the UK population and 19% of UK breast cancer cases are homozygous for the rare allele at rs2981582. On the other hand, the increased risks associated with these alleles are relatively small—on the basis of UK population rates, the estimated breast cancer risk by age 70 years for rare homozygotes at rs2981582 is 10.5%, compared to 6.7% in heterozygotes and 5.5% in common homozygotes. At this stage, it is unlikely that these SNPs will be appropriate for predictive genetic testing, either alone or in combination with each other. However, as further susceptibility alleles are identified, a combination of such alleles together with other breast cancer risk factors may become sufficiently predictive to be important clinically.

On the basis of the relative risk estimates from stage 3, and assuming that the five most significant loci interact multiplicatively on disease risk, these loci explain an estimated 3.6% of the excess familial risk of breast cancer. On the basis of our staged design and the estimated distribution of linkage disequilibrium between the typed SNPs and those in HapMap, we estimate that the power to identify the five most significant associations at $P < 10^{-7}$ (rs2981582, rs3803662, rs889312, rs13281615 and rs3817198) was 93%, 71%, 25%, 3% and 1% respectively. These estimates are uncertain, notably because the true coverage of HapMap SNPs is unknown. Nevertheless, these calculations indicate that the power to detect the two strongest associations was high, and suggest that there are likely to be few other common variants with a similar effect on variation in breast cancer risk to rs2981582. In contrast, the low power to detect rs13281615 and rs3817198 suggests that these variants may represent a much larger class of loci, each explaining of the order of 0.1% of the familial risk of breast cancer. An example of such a locus is provided by *CASP8* D302H, which showed strong evidence of association in a previous large study[15]. This SNP was tested in stage 1, but the association was missed because it did not reach the threshold for testing in stage 2. The excess of associations after stage 2 is also consistent with the existence of many such loci. In addition, because the coverage for SNPs with m.a.f. $< 10\%$ was low, many low frequency alleles may have been missed. The detection of further susceptibility loci will require genome-wide studies with more complete coverage and using larger numbers of cases and controls, together with the combination of results across multiple studies. The present study demonstrates that common susceptibility loci can be reliably identified, and that they may together explain an appreciable fraction of the genetic variance in breast cancer risk.

METHODS SUMMARY

Cases for stage 1 were identified through clinical genetics centres in the UK and a national study of bilateral breast cancer. Cases in stage 2 were drawn from a population-based study of breast cancer (SEARCH)[32]. Controls for stages 2 and 3 were drawn from EPIC-Norfolk, a population-based study of diet and cancer[33].

Cases and controls for stage 3 were identified through case-control studies in Europe, North America, South-East Asia and Australia participating in the Breast Cancer Association Consortium (Supplementary Table 2)[34].

Genotyping for stages 1 and 2 was conducted using high-density oligonucleotide microarrays. For the main analyses, we excluded samples called on ≤80% of SNPs in either stage. We also excluded SNPs that achieved a call rate of ≤90% in stage 1 and ≤95% in stage 2, and SNPs whose frequency deviated from Hardy–Weinberg equilibrium in controls at $P < 0.00001$. Genotyping for stage 3, and for the fine-scale mapping of the *FGFR2* locus, was conducted using either a 5′ nuclease assay (Taqman, Applied Biosystems) or MALDI-TOF mass spectrometry using the Sequenom iPLEX system. For each centre, we excluded any sample called on ≤80% of SNPs, and any SNP with a call rate of ≤95% or a deviation from Hardy–Weinberg equilibrium in controls at $P < 0.00001$. Tests of association were 1 d.f. Cochran-Armitage tests, stratified for stage, centre and ethnic group (European or Asian). Odds ratios for each SNP were estimated using stratified logistic regression, using the stage 3 data only.

Full Methods and any associated references are available in the online version of the paper at www.nature.com/nature.

Received 9 February; accepted 30 April 2007.
Published online 27 May 2007; corrected 28 June 2007 (details online).

1. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: Collaborative reanalysis of individual data from 52 epidemiological

studies including 58 209 women with breast cancer and 101 986 women without the disease. *Lancet* **358,** 1389–1399 (2001).
2. Miki, Y. *et al.* A strong candidate for the breast and ovarian-cancer susceptibility gene BRCA1. *Science* **266,** 66–71 (1994).
3. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene *BRCA2*. *Nature* **378,** 789–792 (1995).
4. Antoniou, A. *et al.* Average risks of breast and ovarian cancer associated with mutations in *BRCA1* or *BRCA2* detected in case series unselected for family history: A combined analysis of 22 studies. *Am. J. Hum. Genet.* **72,** 1117–1130 (2003).
5. Smith, P. *et al.* A genome wide linkage search for breast cancer susceptibility genes. *Genes Chromosom. Cancer* **45,** 646–655 (2006).
6. Pharoah, P. D. P. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nature Genet.* **31,** 33–36 (2002).
7. Antoniou, A. C., Pharoah, P. D. P., Smith, P. & Easton, D. F. The BOADICEA model of genetic susceptibility to breast and ovarian cancer. *Br. J. Cancer* **91,** 1580–1590 (2004).
8. Rahman, N. *et al. PALB2*, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nature Genet.* **39,** 165–167 (2007).
9. Thompson, D. *et al.* Cancer risks and mortality in heterozygous ATM mutation carriers. *J. Natl Cancer Inst.* **97,** 813–822 (2005).
10. Meijers-Heijboer, H. *et al.* Low-penetrance susceptibility to breast cancer due to *CHEK2**1100delC in noncarriers of *BRCA1* or *BRCA2* mutations. *Nature Genet.* **31,** 55–59 (2002).
11. Erkko, H. *et al.* A recurrent mutation in *PALB2* in Finnish cancer families. *Nature* **446,** 316–319 (2007).
12. Renwick, A. *et al. ATM* mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nature Genet.* **38,** 873–875 (2006).
13. Seal, S. *et al.* Truncating mutations in the Fanconi anemia J gene *BRIP1* are low-penetrance breast cancer susceptibility alleles. *Nature Genet.* **38,** 1239–1241 (2006).
14. The CHEK2 Breast Cancer Case-Control Consortium. CHEK2*1100delC and susceptibility to breast cancer: A collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from ten studies. *Am. J. Hum. Genet.* **74,** 1175–1182 (2004).
15. Cox, A. *et al.* A common coding variant in *CASP8* is associated with breast cancer risk. *Nature Genetics* 39, 352–358 (2007); corrigendum **39,** 688 (2007).
16. Easton, D. F. How many more breast cancer predisposition genes are there? *Breast Cancer Res.* **1,** 1–4 (1999).
17. Kruglyak, L. & Nickerson, D. A. Variation is the spice of life. *Nature Genet.* **27,** 234–236 (2001).
18. Hinds, D. A. *et al.* Whole-genome patterns of common DNA variation in three human populations. *Science* **307,** 1072–1079 (2005).
19. Satagopan, J. M., Verbel, D. A., Venkatraman, E. S., Offit, K. E. & Begg, C. B. Two-stage designs for gene-disease association studies. *Biometrics* **58,** 163–170 (2002).
20. Antoniou, A. C. & Easton, D. F. Polygenic inheritance of breast cancer: Implications for design of association studies. *Genet. Epidemiol.* **25,** 190–202 (2003).
21. Altshuler, D. *et al.* A haplotype map of the human genome. *Nature* **437,** 1299–1320 (2005).
22. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55,** 997–1004 (1999).
23. Thomas, D. C., Haile, R. W. & Duggan, D. Recent developments in genomewide association scans: A workshop summary and review. *Am. J. Hum. Genet.* **77,** 337–345 (2005).
24. Amundadottir, L. T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nature Genet.* **38,** 652–658 (2006).
25. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nature Genet.* **39,** 645–649 (2007).
26. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nature Genet.* **39,** 631–637 (2007).
27. Ferretti, V. *et al.* PReMod: a database of genome-wide mammalian cis-regulatory module predictions. *Nucleic Acids Res.* **35,** D122–D126 (2007).
28. Moffa, A. B., Tannheimer, S. L. & Ethier, S. P. Transforming potential of alternatively spliced variants of fibroblast growth factor receptor 2 in human mammary epithelial cells. *Mol. Cancer Res.* **2,** 643–652 (2004).
29. Adnane, J. *et al.* Bek and Flg, 2 receptors to members of the Fgf family, are amplified in subsets of human breast cancers. *Oncogene* **6,** 659–663 (1991).
30. Jang, J. H., Shin, K. H. & Park, J. G. Mutations in fibroblast growth factor receptor 2 and fibroblast growth factor receptor 3 genes associated with human gastric and colorectal cancers. *Cancer Res.* **61,** 3541–3543 (2001).
31. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446,** 153–158 (2007).
32. Lesueur, F. *et al.* Allelic association of the human homologue of the mouse modifier *Ptprj* with breast cancer. *Hum. Mol. Genet.* **14,** 2349–2356 (2005).
33. Day, N. *et al.* EPIC-Norfolk: Study design and characteristics of the cohort. *Br. J. Cancer* **80,** 95–103 (1999).
34. Breast Cancer Association Consortium. Commonly studied SNPs and breast cancer: Negative results from 12,000 – 32,000 cases and controls from the Breast Cancer Association Consortium. *J. Natl Cancer Inst.* **98,** 1382–1396 (2006).
35. Hubbard, T. *et al.* The Ensembl genome database project. *Nucleic Acids Res.* **30,** 38–41 (2002).

Supplementary Information is linked to the online version of the paper at www.nature.com/nature.

Acknowledgements The authors thank the women who took part in this research, and all the funders and support staff who made this study possible. The principal funding for this study was provided by Cancer Research UK. Detailed acknowledgements are provided in Supplementary Information.

Author Contributions D.F.E., A.M.D., P.D.P.P., D.R.C. and B.A.J.P. designed the study and obtained financial support. D.G.B. and D.R.C. directed the genotyping of stages 1 and 2. D.F.E. and D.T. conducted the statistical analysis. K.A.P. and A.M.D. coordinated the genotyping for stage 3 and the fine-scale mapping of the *FGFR2* and *TNRC9* loci. J.P.S. and J.Z. performed resequencing and analysis of the *FGFR2* locus. K.A.P., S.A., C.S.H., R.B., C.A.H., L.K.K., B.E.H., L.L.M., P.B., S.S., V.G., F.O., C-Y. S., P-E.W. and H-C.W. conducted genotyping for the fine-scale mapping. R.L., J.M., H.F. and K.B.M. provided bioinformatics support. D.E., D.G.E., J.P., O.F., N.J., S.S., M.R.S. and N.R. coordinated the studies used in stage 1. N.W. and N.E.D. coordinated the EPIC study used in stages 1 and 2. The remaining authors coordinated the studies in stage 3 and undertook genotyping in those studies. D.F.E. drafted the manuscript, with substantial contributions from K.A.P., A.M.D., P.D.P.P. and B.A.J.P. All authors contributed to the final paper.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Correspondence and requests for materials should be addressed to D.F.E. (d.easton@srl.cam.ac.uk).

Author affiliations: [1]CR-UK Genetic Epidemiology Unit, Department of Public Health and Primary Care and, [2]Department of Oncology, University of Cambridge, Cambridge CB1 8RN, UK. [3]Perlegen Sciences, Inc., 2021 Stierlin Court, Mountain View, California 94043, USA. [4]Laboratory of Population Genetics, US National Cancer Institute, Bethesda, Maryland 20892, USA. [5]EPIC, Department of Public Health and Primary Care, University of Cambridge, Cambridge CB1 8RN, UK. [6]MRC Dunn Clinical Nutrition Centre, Cambridge CB2 0XY, UK. [7]Cancer Research UK Cambridge Research Institute, Cambridge CB2 0RE, UK. [8]Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California 90033, USA. [9]Epidemiology Program, Cancer Research Center of Hawaii, University of Hawaii, Honolulu, Hawaii 96813, USA. [10]International Agency for Research on Cancer, 150 Cours Albert Thomas, Lyon 69008, France. [11]National Cancer Institute, Bangkok 10400, Thailand. [12]Institute of Biomedical Sciences, Academia Sinica, Taipei 11529, Taiwan. [13]Wessex Clinical Genetics Service, Princess Anne Hospital, Southampton SO16 5YA, UK. [14]Regional Genetic Service, St Mary's Hospital, Manchester M13 0JH, UK. [15]London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK, and Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK. [16]Breakthrough Breast Cancer Research Centre, London SW3 6JB, UK. [17]Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK. [18]Cancer Genome Project, Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [19]Queensland Institute of Medical Research, Brisbane, Queensland 4006, Australia. [20]Departments of Clinical Biochemistry and [21]Breast Surgery, Herlev and Bispebjerg University Hospitals, University of Copenhagen, DK-2730 Herlev, Denmark. [22]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland 20852, USA. [23]Advanced Technology Center, National Cancer Institute, Gaithersburg, Maryland 20877, USA. [24]Cancer Center and M. Sklodowska-Curie Institute of Oncology, Warsaw 02781, Poland. [25]Nofer Institute of Occupational Medicine, Lodz 90950, Poland. [26]Departments of Obstetrics and Gynecology, and [27]Department of Oncology, Helsinki University Central Hospital, Helsinki 00029, Finland. [28]Seoul National University College of Medicine, Seoul 151-742, Korea. [29]National Cancer Center, Goyang 411-769, Korea. [30]Ulsan University College of Medicine, Ulsan 680-749, Korea. [31]Program in Molecular and Genetic Epidemiology, Harvard School of Public Health, 677 Huntington Ave., Boston, Massachusetts 02115, USA. [32]Channing Laboratory, Brigham and Women's Hospital and Harvard Medical School, 181 Longwood Ave., Boston, Massachusetts 02115, USA. [33]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, Stockholm SE-171 77, Sweden. [34]Population Genetics, Genome Institute of Singapore, 60 Biopolis Street, Singapore 138672, Republic of Singapore. [35]Department of Radiation Oncology and [36]Department of Gynecology and Obstetrics, Hannover Medical School, D-30625 Hannover, Germany. [37]Department of Surgery and [38]Department of Medical Decision Making and [39]Departments of Human Genetics and Pathology, Leiden University Medical Center, Albinusdreef 2, 2333 ZA Leiden, the Netherlands. [40]Family Cancer Clinic, Department of Medical Oncology, Erasmus MC-Daniel den Hoed Cancer Center, Groene Hilledijk 301, 3075 EA Rotterdam, the Netherlands. [41]Radiation Epidemiology Branch, Division of Cancer Epidemiology and Genetics, National Cancer Institute, NIH, DHHS, Bethesda, Maryland 20892, USA. [42]Environmental Health Sciences, University of Minnesota, Minneapolis, Minnesota 55455, USA. [43]Institute for Cancer Studies and [44]Academic Unit of Surgical Oncology, Sheffield University Medical School, Sheffield S10 2RX, UK. [45]Mayo Clinic College of Medicine, Rochester, Minnesota 55905, USA. [46]VU University Medical Center, 1007 MB Amsterdam, the Netherlands. [47]Department of Clinical Genetics and [48]Internal Medicine, Erasmus University, Rotterdam NL-3015-GE, the Netherlands. [49]Spanish National Cancer Centre (CNIO), Madrid E-28029, Spain. [50]Centre for Molecular, Environmental, Genetic and Analytical Epidemiology, University of Melbourne, Carlton, Victoria 3053, Australia. [51]Department of Preventive and Social Medicine, University of Otago, Dunedin 9001, New Zealand. [52]Cancer Epidemiology Centre, Cancer Council Victoria, Carlton, Victoria 3053, Australia. [53]Genetic Epidemiology

Laboratory, Department of Pathology, University of Melbourne, Parkville, Victoria 3052, Australia. [54]Dr. Margarete Fischer-Bosch-Institute of Clinical Pharamcology, 70376 Stuttgart and University of Tuebingen, 72074 Tuebingen, Germany. [55]Deutsches Krebsforschungszentrum, Heidelberg 69120, Germany. [56]Evangelische Kliniken Bonn gGmbH Johanniter Krankenhaus, 53113 Bonn, Germany. [57]Peter MacCallum Cancer Centre, Melbourne, Victoria 3002, Australia. [58]Insitute of Clinical Medicine, Pathology and Forensic Medicine, University of Kuopio, Kuopio FIN-70210, Finland. [59]Departments of Oncology and Pathology, University Hospital of Kuopio, Kuopio FIN-70211, Finland. [60]Department of Oncology, Vaasa Central Hospital, Vaasa 65130, Finland.

The SEARCH collaborators Craig Luccarini[1], Don Conroy[1], Mitul Shah[1], Hannah Munday[1], Clare Jordan[1], Barbara Perkins[1], Judy West[1], Karen Redman[1] & Kristy Driver[1]. **kConFab** Morteza Aghmesheh[2], David Amor[3], Lesley Andrews[4], Yoland Antill[5], Jane Armes[6], Shane Armitage[7], Leanne Arnold[7], Rosemary Balleine[8], Glenn Begley[9], John Beilby[10], Ian Bennett[11], Barbara Bennett[4], Geoffrey Berry[12], Anneke Blackburn[13], Meagan Brennan[14], Melissa Brown[15], Michael Buckley[16], Jo Burke[17], Phyllis Butow[18], Keith Byron[19], David Callen[20], Ian Campbell[21], Georgia Chenevix-Trench[22], Christine Clarke[23], Alison Colley[24], Dick Cotton[25], Jisheng Cui[26], Bronwyn Culling[27], Margaret Cummings[28], Sarah-Jane Dawson[5], Joanne Dixon[29], Alexander Dobrovic[30], Tracy Dudding[31], Ted Edkins[32], Maurice Eisenbruch[33], Gelareh Farshid[34], Susan Fawcett[35], Michael Field[36], Frank Firgaira[37], Jean Fleming[38], John Forbes[39], Michael Friedlander[40], Clara Gaff[41], Mac Gardner[41], Mike Gattas[42], Peter George[43], Graham Giles[44], Grantley Gill[45], Jack Goldblatt[46], Sian Greening[47], Scott Grist[37], Eric Haan[48], Marion Harris[49], Stewart Hart[50], Nick Hayward[22], John Hopper[51], Evelyn Humphrey[17], Mark Jenkins[52], Alison Jones[7], Rick Kefford[53], Judy Kirk[54], James Kollias[55], Sergey Kovalenko[56], Sunil Lakhani[57], Jennifer Leary[54], Jacqueline Lim[58], Geoff Lindeman[59], Lara Lipton[60], Liz Lobb[61], Mariette Maclurcan[62], Graham Mann[23], Deborah Marsh[63], Margaret McCredie[64], Michael McKay[49], Sue Anne McLachlan[65], Bettina Meiser[4], Roger Milne[26], Gillian Mitchell[49], Beth Newman[66], Imelda O'Loughlin[67], Richard Osborne[51], Lester Peters[68], Kelly Phillips[5], Melanie Price[62], Jeanne Reeve[69], Tony Reeve[70], Robert Richards[71], Gina Rinehart[72], Bridget Robinson[73], Barney Rudzki[74], Elizabeth Salisbury[75], Joe Sambrook[21], Christobel Saunders[76], Clare Scott[5], Elizabeth Scott[77], Rodney Scott[31], Ram Seshadri[37], Andrew Shelling[78], Melissa Southey[26], Amanda Spurdle[22], Graeme Suthers[48], Donna Taylor[79], Christopher Tennant[58], Heather Thorne[21], Sharron Townshend[46], Kathy Tucker[4], Janet Tyler[4], Deon Venter[80], Jane Visvader[81], Ian Walpole[46], Robin Ward[82], Paul Waring[30], Bev Warner[83], Graham Warren[67], Elizabeth Watson[67], Rachael Williams[84], Judy Wilson[85], Ingrid Winship[69] & Mary Ann Young[49]. **AOCS Management Group** David Bowtell[86], Adele Green[22], Anna deFazio[87], Georgia Chenevix-Trench[22], Dorota Gertig[51] & Penny Webb[22].

Consortia affiliations: [1]Department of Oncology, University of Cambridge, Cambridge CB1 8RN, UK. [2]Oncology Research Centre, Prince of Wales Hospital, Randwick, New South Wales 2031, Australia. [3]Genetic Health Services Victoria, Royal Children's Hospital, Melbourne, Victoria 3050, Australia. [4]Hereditary Cancer Clinic, Prince of Wales Hospital, Randwick, New South Wales 2031, Australia. [5]Department of Haematology and Medical Oncology, Peter MacCallum Cancer Centre, St Andrews Place, East Melbourne, Victoria 3002, Australia. [6]Anatomical Pathology, Royal Women's Hospital, Carlton, Victoria 3053, Australia. [7]Molecular Genetics Laboratory, Royal Brisbane and Women's Hospital, Herston, Queensland 4029, Australia. [8]Departments of Translational and Medical Oncology, Westmead Hospital, Westmead, New South Wales 2145, Australia. [9]Cancer Biology Laboratory, TVW Institute for Child Health Research, Subiaco, Western Australia 6008, Australia. [10]Pathology Centre, Queen Elizabeth Medical Centre, Nedlands, Western Australia 6009, Australia. [11]Silverton Place, 101 Wickham Terrace, Brisbane, Queensland 4000, Australia. [12]Department of Public Health and Community Medicine, University of Sydney, Sydney, New South Wales 2006, Australia. [13]John Curtin School of Medical Research, Australian National University, PO Box 334, Canberra, Australian Capital Territory 2601, Australia. [14]NSW Breast Cancer Institute, PO Box 143, Westmead, New South Wales 2145, Australia. [15]Department of Biochemistry, University of Queensland, St. Lucia, Queensland 4072, USA. [16]Molecular and Cytogenetics Unit, Prince of Wales Hospital, Randwick, New South Wales 2031, Australia. [17]Royal Hobart Hospital, GPO Box 1061L, Hobart, Tasmania 7001, Australia. [18]Medical Psychology Unit, Royal Prince Alfred Hospital, Camperdown, New South Wales 2204, Australia. [19]Australian Genome Research Facility, Walter & Eliza Hall Medical Research Institute, Royal Melbourne Hospital, Parkville, Victoria 3050, Australia. [20]Dame Roma Mitchell Cancer Research Laboratories, University of Adelaide/ Hanson Institute, PO Box 14, Rundle Mall, South Australia 5000, Australia. [21]Peter MacCallum Cancer Centre, St Andrew's Place, East Melbourne, Victoria 3002, Australia. [22]Queensland Institute of Medical Research, Herston, Queensland 4006, Australia. [23]Westmead Institute for Cancer Research, University of Sydney, Westmead Hospital ,Westmead, New South Wales 2145, Australia. [24]Department of Clinical Genetics, Liverpool Health Service, PO Box 103, Liverpool, New South Wales 2170, Australia. [25]Mutation Research Centre, St Vincent's Hospital, Victoria Parade, Fitzroy, Victoria 3065, Australia. [26]Centre for Genetic Epidemiology, The University of Melbourne, Level 2 723 Swanston Street, Carlton, Victoria 3053, Australia. [27]Molecular and Clinical Genetics, Level 1 Building 65, Royal Prince Alfred Hospital, Camperdown, New South Wales 2050, Australia. [28]Department of Pathology, University of Queensland Medical School, Herston, New South Wales 4006, Australia. [29]Central Regional Genetic Services, Wellington Hospital, Private bag 7902, Wellington South 6039, New Zealand. [30]Molecular Department of Pathology, Peter MacCallum Cancer Centre, St Andrew's Place, East Melbourne, Victoria 3002, Australia. [31]Hunter Genetics, Hunter Area Health Service, Waratah, New South Wales 2310, Australia. [32]Clinical Chemistry, Princess Margret Hospital for Children, Box D184, Perth, Western Australia 6001, Australia. [33]Department of Multicultural Health, University of Sydney, New South Wales 2052, Australia. [34]Tissue Pathology, Institute of Medical & Veterinary Science, Adelaide, South Australia 5000, Australia. [35]Family Cancer Clinic, Monash Medical Centre, Clayton, Victoria 3168, Australia. [36]Faculty of Medicine, Royal North Shore Hospital, Vindin House, St Leonards, New South Wales 2065, Australia. [37]Department of Haematology, Flinders Medical Centre, Bedford Park, South Australia 5042, Australia. [38]Eskitis Institute of Cell & Molecular Therapies, School of Biomolecular and Biomedical Sciences, Griffith University, Nathan, Queensland 4111, Australia. [39]Surgical Oncology, University of Newcastle, Newcastle Mater Hospital, Waratah, New South Wales 2298, Australia. [40]Department of Medical Oncology, Prince of Wales Hospital, Randwick, New South Wales 2031, Australia. [41]Victorian Clinical Genetics Service, Royal Melbourne Hospital, Parkville, Victoria 3052, Australia. [42]Queensland Clinical Genetic Service, Royal Children's Hospital, Bramston Terrace, Herston, Queensland 4020, Australia. [43]Clinical Biochemistry Unit, Canterbury Health Labs, PO Box 151, Christchurch 8140, New Zealand. [44]Cancer Epidemiology Centre, The Cancer Council Victoria, 1 Rathdowne Street, Carlton, Victoria 3053, Australia. [45]Department of Surgery, Royal Adelaide Hospital, Adelaide, South Australia 5000, Australia. [46]Genetic Services of WA, King Edward Memorial Hospital, 374 Bagot Road, Subiaco, Western Australia 6008, Australia. [47]Wollongong Hereditary Cancer Clinic, Wollongong Public Hospital, Private Mail Bag 8808, South Coast Mail Centre, New South Wales 2521, Australia. [48]Department of Medical Genetics, Women's and Children's Hospital, North Adelaide, South Australia 5006, Australia. [49]Familial Cancer Clinic, Peter MacCallum Cancer Centre, St Andrew's Place, East Melbourne, Victoria 3002, Australia. [50]Breast and Ovarian Cancer Genetics, Monash Medical Centre, 871 Centre Road, Bentleigh East, Victoria 3165, Australia. [51]Centre for Molecular Environmental, Genetic & Analytic Epidemiology, University of Melbourne, Melbourne, Victoria 3010, Australia. [52]School of Population Health, The University of Melbourne, 723 Swanston Street, Carlton, Victoria 3053, Australia. [53]Medical Oncology, Westmead Hospital, Westmead, New South Wales 2145, Australia. [54]Familial Cancer Service, Department of Medicine, Westmead Hospital, Westmead, New South Wales 2145, Australia. [55]Breast Endocrine and Surgical Unit, Royal Adelaide Hospital, North Terrace, South Australia 5000, Australia. [56]Molecular Pathology Department, Southern Cross Pathology, Monash Medical Centre, Clayton, Victoria 3168, Australia. [57]Molecular and Cellular Pathology, The University of Queensland, Herston, Queensland 4006, Australia. [58]Department of Psychological Medicine, Royal North Shore Hospital, St Leonards, New South Wales 2065, Australia. [59]Breast Cancer Laboratory, Walter and Eliza Hall Institute, PO Royal Melbourne Hospital, Parkville, Victoria 3050, Australia. [60]Medical Oncology and Clinical Haematology Unit, Western Hospital, Footscray, Victoria 3011, Australia. [61]WA Centre for Cancer, Edith Cowan University, Churchlands, Western Australia 6018, Australia. [62]Department of Psychological Medicine, University of Sydney, New South Wales 2006, Australia. [63]Kolling Institute of Medical Research, Royal North Shore Hospital, St Leonards, New South Wales 2065, Australia. [64]Cancer Epidemiology Research Unit, NSW Cancer Council, 153 Dowling Street, Woolloomooloo, New South Wales 2011, Australia. [65]Department of Oncology, St Vincent's Hospital, 41 Victoria Parade, Fitzroy, Victoria 3065, Australia. [66]School of Public Health, Queensland University of Technology, Victoria Park, Kelvin Grove, Queensland 4059, Australia. [67]St Vincent's Breast Clinic, PO Box 4751, Toowoomba, Queensland 4350, Australia. [68]Radiation Oncology, Peter MacCallum Cancer Centre, St Andrew's Place, East Melbourne, Victoria 3002, Australia. [69]Genetic Services, Auckland Hospital, Private Bag 92024, Auckland 1142, New Zealand. [70]Cancer Genetics Laboratory, University of Otago, PO Box 56, Dunedin 9054, New Zealand. [71]Department of Cytogenetics and Molecular Genetics, Women and Children's Hospital, Adelaide, South Australia 5006, Australia. [72]Hancock Family Breast Cancer Foundation, PO Locked Bag 2, West Perth, Western Australia 6005, Australia. [73]Oncology Service, Christchurch Hospital, Private Bag 4710, Christchurch 8140, New Zealand. [74]Molecular Pathology Institute of Medical and Veterinary Science, Frome Road, Adelaide, South Australia 5000, Australia. [75]Section of Cytology, Institute of Clinical Pathology and Medical Research, Westmead Hospital, Westmead, New South Wales 2145, Australia. [76]School of Surgery and Pathology, QE11 Medical Centre, M block 2nd Floor, Nedlands, Western Australia 6907, Australia. [77]South View Clinic, Suite 13, Level 3 South Street, Kogarah, New South Wales 2217, Australia. [78]Department of Obstetrics and Gynaecology, University of Auckland, Private Bag 92019, Auckland 1142, New Zealand. [79]Department of Radiology, Royal Perth Hospital, Box X2213, Perth 6011, Western Australia, Australia. [80]Murdoch Institute, Royal Children's Hospital, Parkville, Victoria 3050, Australia. [81]Molecular Genetics of Cancer Division, Walter & Eliza Hall Medical Research Institute, Royal Melbourne Hospital, Parkville, Victoria 3050, Australia. [82]Department of Medical Oncology, St Vincents Hospital, Darlinghurst, New South Wales 2010, Australia. [83]Cabrini Hospital, 183 Wattletree Road, Malvern, Victoria 3144, Australia. [84]Family Cancer Clinic, St Vincent's Hospital, Darlinghurst, New South Wales 2010, Australia. [85]Medical Psychology Research Unit, Royal North Shore Hospital, St Leonards, New South Wales 2065, Australia. [86]Cancer Genomics & Biochemistry Laboratory, Peter MacCallum Cancer Centre, St Andrew's Place, East Melbourne, Victoria 3002, Australia. [87]Obstetrics & Gynaecology, Westmead Hospital, University of Sydney, New South Wales 2006, Australia.

METHODS

Subjects. Cases in stage 1 were identified through clinical genetics centres in Cambridge ($n = 91$), Manchester (96) and Southampton (136), and a national study of bilateral breast cancer (85). Cases were women diagnosed with invasive breast cancer under the age of 60 years who had a family history score of at least 2, where the score was computed as the total number of first-degree relatives plus half the number of second-degree relatives affected with breast cancer. The score for women with bilateral breast cancer was increased by 1, so that women were eligible if they were diagnosed with bilateral breast cancer and had one affected first-degree relative. Cases known to carry a *BRCA1* or *BRCA2* mutation were excluded. Controls were selected from the EPIC-Norfolk study, a population-based cohort study of diet and cancer based in Norfolk, East Anglia, UK[33]. Controls were chosen to be women aged over 50 years and free of cancer at the time of entry. Genotyping was attempted on 408 cases, plus 32 duplicate case samples, and 400 controls. For the analysis in Table 1, 54 samples with genotype call rates <80% were excluded, so the final analyses were based on 390 cases and 364 controls. The minimum genotype call rate for the remaining samples was 89%. The overall genotype discordance rate between duplicate samples in stage 1 was 0.01%.

For stage 2, invasive breast cancer cases were drawn from SEARCH, a population-based study of cancer in East Anglia[32]. Controls were women selected from the EPIC-Norfolk study, as previously described[33]. Eighty-eight subjects who were also genotyped in stage 1, and 35 controls who subsequently developed breast cancer and were also in the case series, were excluded from the analysis, leaving 3,990 breast cancer cases and 3,916 controls, plus five duplicates. The overall rate of discordance of genotypes between duplicate samples in stage 2 was 0.008%.

Twenty-one additional studies were included in stage 3 (see Supplementary Table 2). These studies participated through the Breast Cancer Association Consortium, an ongoing collaboration among investigators conducting case-control association studies in breast cancer[15,33]. All studies provided information on disease status (invasive breast cancer, carcinoma *in situ* or control), age at diagnosis/observation, ethnic group, first-degree family history of breast cancer and bilaterality of breast cancer. One further study (Breast Cancer Study of Taiwan) was included in the fine-scale mapping of the *FGFR2* locus.

Genotyping. For stage 1, genotyping was performed on 200 ng DNA that was first subjected to whole genome amplification using Multiple Displacement Amplification (MDA)[36]. Samples were then genotyped for a set of 266,732 SNPs using high-density oligonucleotide, photolithographic microarrays at Perlegen Sciences. For stage 2, genotyping was performed using 2.5 μg genomic DNA. These samples were genotyped for a set of 13,023 SNPs selected on the basis of the stage 1 results, using a custom designed oligonucleotide array. For both stages, each SNP was interrogated by 24 25-mer oligonucleotide probes synthesized by photolithography on a glass substrate. The 24 features comprise 4 sets of 6 features interrogating the neighbourhoods of SNP reference and alternative alleles on forward and reference strands. Each allele and strand is represented by five offsets: −2, −1, 0, 1 and 2 indicating the position of the SNP within the 25-mer, with zero being at the thirteenth base. At offset 0 a quartet was tiled, which included the perfect match to reference and alternative SNP alleles, and the two remaining nucleotides as mismatch probes. When possible, the mismatch features were selected as a purine nucleotide substitution for a purine perfect match nucleotide and a pyrimidine nucleotide substitution for a pyrimidine perfect match nucleotide. Thus, each strand and allele tiling consisted of 6 features comprising five perfect match probes and one mismatch.

Individual genotypes were determined by clustering all SNP scans in the two-dimensional space defined by reference and alternative trimmed mean intensities, corrected for background. Allele frequencies were approximated using the intensities collected from the high-density oligonucleotide arrays. An SNP's allele frequency, p, was estimated as the ratio of the relative amount of the DNA with reference allele to the total amount of DNA. The $\hat{p}$ value was computed from the trimmed mean intensities of perfect match features, after subtracting a measure of background computed from trimmed means of intensities of mismatch features. The trimmed mean disregarded the highest and the lowest intensity from the five perfect match intensities before computing the arithmetic mean. For the mismatch features, the trimmed mean is the individual intensity of the specified mismatch feature.

The genotype clustering procedure was an iterative algorithm developed as a combination of K-means and constrained multiple linear regressions. The K-means at each step re-evaluated the cluster membership representing distinct diploid genotypes. The multiple linear regressions minimized the variance in $\hat{p}$ within each cluster while optimizing the regression lines' common intersect. The common intersect defined a measure of common background that was used to adjust the allele frequencies for the next step of K-means. The K-means and multiple linear regression steps were iterated until the cluster membership and background estimates converged. The best number of clusters was selected by maximizing the total likelihood over the possible cluster counts of 1, 2 and 3 (representing the combinations of the three possible diploid genotypes). The total likelihood was composed of data likelihood and model likelihood. The data likelihood was determined using a normal mixture model for the distribution of $\hat{p}$ around the cluster means. The model likelihood was calculated using a prior distribution of expected cluster positions, resulting in optimal $\hat{p}$ positions of 0.8 for the homozygous reference cluster, 0.5 for the heterozygous cluster and 0.2 for the homozygous alternative cluster.

A genotyping quality metric was compiled for each genotype from 15 input metrics that described the quality of the SNP and the genotype. The genotyping quality metric correlated with a probability of having a discordant call between the Perlegen platform and outside genotyping platforms (that is, non-Perlegen HapMap project genotypes). A system of 10 bootstrap aggregated regression trees was trained using an independent data set of concordance data between Perlegen genotypes and HapMap project genotypes. The trained predictor was then used to predict the genotyping quality for each of the genotypes in this data set. Genotypes with quality scores of less than 7 were discarded. Data were analysed for 227,876 SNPs in stage 1 and 12,026 (of 13,023 selected) in stage 2, for which the call rate was >80%.

The 12,711 SNPs for stage 2 were primarily selected on the basis of a 1 d.f. Cochran-Armitage trend test (11,809, all with $P < 0.052$). We also included 826 SNPs with $P < 0.01$ testing for the difference in frequency of either homozygote between cases and controls (that is, assuming either a dominant or recessive model) and 76 SNPs that achieved $P < 0.01$ on a Cochran-Armitage test, weighting individuals by their family history score as above.

For the main analyses, we discarded SNPs with a call rate <90% in stage 1 and 95% in stage 2, and SNPs with a deviation from Hardy–Weinberg equilibrium significant at $P < 0.00001$ in either stage, leaving 205,586 SNPs in stage 1 and 10,621 SNPs in stage 2.

The 30 SNPs included in the stage 3 analyses were initially selected on the basis of a combined analysis of stage 1 and stage 2. We included all SNPs achieving a combined $P < 0.00002$ (based on either the Cochran-Armitage or 2 d.f. test, see below). Following re-evaluation of the stage 2 genotyping by 5′ nuclease assay (Taqman, Applied Biosystems) using the ABI PRISM 7900HT (Applied Biosystems), and exclusion of some samples, 16 of these SNPs were significant at $P < 0.00002$ and 24 at $P < 0.0002$ (Supplementary Table 3). One additional SNP, rs3803662, was added as a result of fine-scale mapping of the *TNRC9*/*LOC643714* locus.

The 31 stage 3 SNPs were genotyped in 22 studies (including cases and controls from SEARCH not used in stage 2, together with 21 other studies). For 18 of the studies, genotyping was performed by 5′ nuclease assay (Taqman) using the ABI PRISM 7900HT or 7500 Sequence Detection Systems according to manufacturer's instructions. Primers and probes were supplied directly by Applied Biosystems (http://www.appliedbiosystems.com/) as Assays-by-Design. All assays were carried out in 384-well or 96-well format, with each plate including negative controls (with no DNA). Duplicate genotypes were provided for at least 2% of samples in each study. For three studies, SNPs were genotyped using matrix assisted laser desorption/ionization time of flight mass spectrometry (MALDI-TOF MS) for the determination of allele-specific primer extension products using Sequenom's MassARRAY system and iPLEX technology. The design of oligonucleotides was carried out according to the guidelines of Sequenom and performed using MassARRAY Assay Design software (version 1.0). Multiplex PCR amplification of amplicons containing SNPs of interest was performed using Qiagen HotStart Taq Polymerase on a Perkin Elmer GeneAmp 2400 thermal cycler (MJ Research) with 5 ng genomic DNA. Primer extension reactions were carried out according to manufacturer's instructions for iPLEX chemistry. Assay data were analysed using Sequenom TYPER software (version 3.0). One study used both the Taqman and MALDI-TOF MS approaches. The SNPs genotyped in stage 3 were also regenotyped in the stage 2 samples using Taqman; these genotype calls were used in the overall analyses (Table 2, Supplementary Table 3, and Fig. 2).

We eliminated any sample that could not be scored on 20% of the SNPs attempted. We also removed data for any centre/SNP combination for which the call rate was less than 90%. In any instances where the call rate was 90–95%, the clustering of genotype calls was re-evaluated by an independent observer to determine whether the clustering was sufficiently clear for inclusion. We also eliminated all the data for a given SNP/centre where the reproducibility in duplicate samples was <97%, or where there was marked deviation from Hardy–Weinberg equilibrium in the controls ($P < 0.00001$).

Fine-scale mapping of *FGFR2*. Initial tagging of the associated region was done by identifying all SNPs with an m.a.f. > 5% in the HapMap CEPH/CEU set (Utah residents with ancestry from northern and western Europe). We then selected 7 SNPs (in addition to rs2981582) that tagged these variants with a

pairwise $r^2 > 0.8$, using the program Tagger (http://www.broad.mit.edu/mpg/tagger/)[37]. To identify additional common variants within the 32.5 kb region of linkage around the associated SNP, we resequenced 45 lymphocyte DNA samples from a subset of European subjects also genotyped by HapMap and other publicly available data sets. Seventy overlapping PCR amplicons were designed from positions 123317613 to 123348192 of chromosome 10 (average amplicon size 650 bp, 160 bp 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.nih.gov). Eighty-six per cent of the nucleotides across the region could be scored for polymorphisms in at least 80% of subjects. This set gave a >97% probability of detecting a variant with an m.a.f. > 5%. One hundred and seventeen variants were identified, including 27 present in dbSNP but without individual genotype information in European subjects, and an additional 46 not in dbSNP. Individual genotype information was then compared and merged with publicly available genotypes from Caucasian subjects (HapMap release 21 for 60 CEU parents, 22 European subjects from the Environmental Genome Project (EGP) resequencing effort (http://egp.gs.washington.edu/data/fgfr2/), and 24 European subjects from Perlegen (retrieved through http://gvs.gs.washington.edu/GVS)). There were 2 discrepancies among 389 genotype calls among subjects in common between our resequencing effort and EGP or Perlegen data, and 10 out of 926 compared to HapMap genotypes.

On the basis of these data, we identified 28 SNPs correlated with rs2981582 with $r^2 > 0.6$. We then attempted to genotype these 28 SNPs, plus rs2981582, in a subset of 80 controls from SEARCH and 84 controls from the Seoul Breast Cancer Study. Twenty-two of the variants were genotyped using Taqman. Four further variants (rs34032268, rs2912778, rs2912781 and rs7895676), which were not amenable to Taqman, were genotyped by Pyrosequencing (Biotage; http://www.biotagebio.com/). Assays were designed using Pyrosequencing Assay Design Software 1.0. The remaining 2 SNPs (rs35393331 and rs33971856) could not be genotyped using either technology and were excluded from further analyses. We cannot therefore comment on their likelihood of being the causal variant. Using these data, we selected tagging sets of 11 SNPs for UK subjects and 14 SNPs for Korean subjects (including rs2981582), such that each of the remaining variants was correlated with a tagging SNP with $r^2 > 0.95$ in the UK study or $r^2 > 0.86$ in the Korean study. After genotyping the 11 tag SNPs in SEARCH, two of these SNPs (rs4752569 and rs35012336) showed strong evidence against being the causative variant and were not considered further. The remaining 12 tag SNPs from the Korean subset were then genotyped in the samples from the IARC-Thai Breast Cancer Study, the Breast Cancer Study in Taiwan and the Multi-Ethnic Cohort (MEC), by Taqman.

Statistical methods. The primary test used for each SNP was a Cochran-Armitage 1 d.f. score test for association between disease status and allele dose. In the combined analysis, we performed a stratified Cochran-Armitage test. Stage 1 was given a weight of 4 in this analysis (corresponding to a weight of 2 in the score statistic), to allow for the expected greater effect size given the inclusion of cases with a family history. In the stage 3 analyses, each study was treated as a separate stratum, except for the MEC, in which the European American and Japanese American subgroups were treated as separate strata. For all studies except the MEC, individuals from a minor ethnic group for that study were excluded. Per-allele and genotype-specific odds ratios, and confidence intervals, were estimated using logistic regression, adjusting for the same strata. The summary odds ratios in Fig. 2 are based on the data from the stage 3 studies only, to avoid the bias inherent in estimates from the stage 1 and 2 data for SNPs exhibiting an association (the so called 'winner's curse'). The effects of genotype on family history of breast cancer (first degree yes/no) and bilaterality were examined by treating these variables as outcomes in a stratified Cochran-Armitage test.

To assess the global significance of the SNPs in stage 3, we computed the sum of the χ^2 trend statistics (excluding the 6 SNPs reaching genome-wide significance, plus rs2107425 as it was in LD with rs3817198) over those SNPs (17 of 23) for which the estimated odds ratios in stage 3 were in the same direction as the combined stage 1/stage 2[38]. Under the null hypothesis of no association, the asymptotic distribution of this statistic is χ^2 with n degrees of freedom, where n has a binomial distribution with parameters 23 and 1/2. The significance of this statistic was then assessed by computing a weighted sum of the tails of the relevant χ^2 distributions.

For the fine-scale mapping of the *FGFR2* locus, we first derived haplotype frequencies using the haplo.stats package in S-plus[39], separately for the European and Asian populations, using data from the case-control studies on whom the tag SNPs were typed plus the 164 control individuals on whom all SNPs were typed. These were used to impute genotype probabilities for each identified SNP in each individual. We then used an EM algorithm to fit a logistic regression model assuming that each SNP in turn was the causal variant, allowing for uncertainty in the genotypes of untyped SNPs, and hence to determine the likelihood that each SNP was the causal variant.

Coverage of the stage 1 tagging set was estimated using HapMap phase II as a reference. We based estimates on 2,116,183 SNPs with an m.a.f. of >5% in the CEU population. Of the SNPs successfully genotyped in stage 1, 187,663 were also on HapMap. For those SNPs not on HapMap, we identified 'surrogate' SNPs that were in perfect LD based on genotyping of 24 Caucasians by Perlegen Sciences (269,203 SNPs)[18]. To estimate coverage, we determined the best pairwise r^2 for each HapMap SNP and each tag SNP or a surrogate SNP, using the HapMap CEU data. This coverage was summarized in terms of the distribution of r^2 by allele frequency in 10 categories.

To estimate the power to detect each of the associations found, we computed the non-centrality parameter for the test statistic at each stage, based on the per-allele relative risk, allele frequency and r^2. This was used to estimate the power for a given r^2, based on a simulated trivariate normal distribution for the score statistics after each stage to allow for the correlations in the test statistics. We assumed a cut-off of $P < 0.05$ for stage 1, $P < 0.00002$ for stage 2 and $P < 10^{-7}$ for stage 3 (the first is slightly conservative, as more SNPs than this were actually taken forward). The overall power was obtained by averaging the power estimates for each r^2 over the distribution of r^2 obtained from the HapMap data, applicable to a SNP of that frequency.

The expected number of significant associations after stage 2 (Table 1) was calculated using a bivariate normal distribution for the joint distribution of the (weighted) Cochran-Armitage score statistics after stage 1 and after both stages, using a correlation of 0.525 between the two statistics (reflecting the weighted sizes of the two studies). These calculations were based on the 205,586 SNPs reaching the required quality control in stage 1. Of these, 11,313 reached a $P < 0.05$, of which 7,405 (65.5%) were successfully genotyped to the required quality control in stage 2. Thus the expected number reaching a given significance level with good quality control was calculated from the total number expected to reach this level × 65.5%. We adjusted the variances of the test statistics, separately for stages 1 and 2, using the genomic control method[22]. The adjustment factor, λ, was estimated from the median of the smallest 90% of the test statistics for SNPs typed in that stage, divided by the predicted median for the smallest 90% of a sample of χ^2_1 distributions (that is, the 45% percentile of a χ^2_1 distribution, 0.375).

36. Dean, F. B. *et al.* Comprehensive human genome amplification using multiple displacement amplification. *Proc. Natl Acad. Sci. USA* **99**, 5261–5266 (2002).
37. de Bakker, P. I. W. *et al.* Efficiency and power in genetic association studies. *Nature Genet.* **37**, 1217–1223 (2005).
38. Tyrer, J., Pharoah, P. D. P. & Easton, D. F. The admixture maximum likelihood test: A novel experiment-wise test of association between disease and multiple SNPs. *Genet. Epidemiol.* **30**, 636–643 (2006).
39. Schaid, D. J., Rowland, C. M., Tines, D. E., Jacobson, R. M. & Poland, G. A. Score tests for association between traits and haplotypes when linkage phase is ambiguous. *Am. J. Hum. Genet.* **70**, 425–434 (2002).

nature genetics

Association scan of 14,500 nonsynonymous SNPs in four diseases identifies autoimmunity variants

Wellcome Trust Case Control Consortium[1] & The Australo-Anglo-American Spondylitis Consortium[1]

We have genotyped 14,436 nonsynonymous SNPs (nsSNPs) and 897 major histocompatibility complex (MHC) tag SNPs from 1,000 independent cases of ankylosing spondylitis (AS), autoimmune thyroid disease (AITD), multiple sclerosis (MS) and breast cancer (BC). Comparing these data against a common control dataset derived from 1,500 randomly selected healthy British individuals, we report initial association and independent replication in a North American sample of two new loci related to ankylosing spondylitis, *ARTS1* and *IL23R*, and confirmation of the previously reported association of AITD with *TSHR* and *FCRL3*. These findings, enabled in part by increased statistical power resulting from the expansion of the control reference group to include individuals from the other disease groups, highlight notable new possibilities for autoimmune regulation and suggest that *IL23R* may be a common susceptibility factor for the major 'seronegative' diseases.

Genome-wide association scans are currently revealing a number of new genetic variants for common diseases[1–11]. We have recently completed the largest and most comprehensive scan conducted to date, involving genome-wide association studies of 2,000 individuals from each of seven common disease cohorts and 3,000 common control individuals using a dense panel of >500,000 markers[12]. In parallel with this scan, we conducted a study of 5,500 independent individuals with a genome-wide set of nonsynonymous coding variants, an approach that has recently yielded new findings about type 1 diabetes and Crohn's disease and that has been proposed as an efficient complementary approach to whole-genome scans[13–15]. Here we report several new replicated associations in our scan of nsSNPs in 1,500 shared controls and 1,000 individuals from each of four different diseases: ankylosing spondylitis, AITD (of which all had Graves' disease), breast cancer and multiple sclerosis.

RESULTS

Initial genotyping was carried out with a custom-made Infinium array (Illumina) and involved 14,436 nsSNPs (assays were synthesized for 16,078 nsSNPs). At the inception of the study, this comprised the complete set of experimentally validated nsSNPs with minor allele frequency (MAF) > 1% in western European samples. In addition, because three of the diseases were of autoimmune etiology, we also typed a dense set of 897 SNPs throughout the MHC that, together with 348 nsSNPs in this region, provided comprehensive tag SNP coverage ($r^2 \geq 0.8$ with all SNPs in ref. 16). Finally, 103 SNPs were typed in pigmentation genes specifically designed to differentiate between population groups. Similar to those from previous studies, our data revealed that detailed assessment of initial data is critical to the process of association inference, as biases in genotype calling lead to inflation of false-positive rates[12,17]. This inflation is exaggerated in nsSNP data, because nsSNPs tend to have lower allele frequencies than otherwise anonymous genomic SNPs, and genotype calling is often most difficult for rare alleles. If only cursory filtering had been applied in the present case, numerous false-positives would have emerged (**Supplementary Figs. 1–4** online). **Table 1** shows the total number of SNPs and individuals remaining after genotype and sample quality control procedures (see Methods).

Association with the MHC

The strongest associations observed in the study were between SNPs in the MHC region and the three autoimmune diseases studied—ankylosing spondylitis, AITD and MS—with P values of $<10^{-20}$ for each disease (**Fig. 1**). No association of the MHC was seen with breast cancer ($P > 10^{-4}$ across the region). For each of the autoimmune diseases, the maximum signal was centered around the known HLA-associated genes (for example, those encoding HLA-B in ankylosing spondylitis, HLA-DRB1 in MS and the MHC class I and class II molecules in AITD), but in all cases, it extended far beyond the specific associated haplotype(s). For example, in ankylosing spondylitis, association was observed at $P < 10^{-20}$ across ~1.5 Mb. Given the well-known strong effect of HLA-B27 variant on the probability of developing ankylosing spondylitis (odds ratio 100–200 in most populations), the extent of this association signal reflects that with such large effects, even very distant SNPs in modest linkage disequilibrium (LD) will show indirect evidence for association. Strong signals like these may also cloud the evidence for additional HLA loci[18]. Disentangling similar patterns of association within the MHC has proven extremely challenging in the past and will be addressed in future studies of these data. Here we focus specifically on the nsSNP results.

[1]The complete lists of participants and affiliations appear at the end of the article. Correspondence should be addressed to L.R.C. (lcardon@fhcrc.org) or D.M.E. (davide@well.ox.ac.uk).

Received 17 July; accepted 17 September; published online 21 October 2007; doi:10.1038/ng.2007.17

Table 1 Number of individuals and SNPs tested in each cohort

	Cohort				
	AS	AITD	BC	MS	58C
Males	610	138	0	271	732
Females	312	762	1,004	704	734
Number of SNPs genotyped	15,436	15,436	15,436	15,436	15,436
SNPs with low GC score	783	816	771	802	796
SNPs with low genotyping	133	206	124	218	186
Monomorphic SNPs	1,842	1,829	1,854	1,810	1,687
SNPs with HW $P < 10^{-7}$[a]	129	74	104	97	132
Differences in missing rate $P < 10^{-4}$	51	101	172	309	n/a
'Manual' exclusions	33	33	33	33	33
Total number of SNPs tested	12,701	12,572	12,577	12,374	

[a]Only SNPs with HW $P < 10^{-7}$ in the 1958 birth cohort (58C) control group were excluded from analyses.

Association with nsSNPs

A major advantage of the Wellcome Trust Case Control Consortium (WTCCC) design is the availability of multiple disease cohorts that are similar in terms of ancestry and that have been typed on the same genetic markers[12,17]. Assuming that each disease has at least some unique genetic loci, we hypothesized that combining the other three case groups with the controls for the 1958 birth cohort (58C)[19] would increase power to detect association. For each disease, we therefore conducted two primary analyses: first, we tested nsSNP associations for each disease against the controls in the 58C; and second, we tested the same associations for each disease against an expanded reference group comprising the combined cases from the other three disease groups plus individuals from the 58C. A similar set of analyses was conducted for each of the autoimmune disorders against a reference group comprising 58C controls and individuals with breast cancer, but the results were very similar to those for the fully expanded groups, so here we describe the larger sample (**Supplementary Table 1** online). In addition, because it is possible that different autoimmune diseases share similar genetic etiologies, we also compared a combined ankylosing spondylitis, AITD and MS group (immune cases) against the combined set of individuals with breast cancer and 58C controls. All of our analyses are reported without regard to specific treatment of population structure, as the degree of structure in our final genotype data is not severe (Genomic Control[20] $\lambda = 1.07$–1.13 in the 58C-only datasets; $\lambda = 1.03$–1.06 in the expanded reference group comparisons; **Table 2**), consistent with our recent findings from 17,000 UK individuals involving the same controls[12].

nsSNP association results (excluding the MHC region) for each of the four disease groups against the 58C controls are shown in **Figure 2** and **Table 3**. Two SNPs on chromosome 5 reached a high level of statistical significance for ankylosing spondylitis (rs27044: $P = 1.0 \times 10^{-6}$; rs30187: $P = 3.0 \times 10^{-6}$). This level of significance exceeds the 10^{-5}–10^{-6} thresholds advocated for gene-based scans[21], as well as the oft-used Bonferroni correction at $P < 0.05$ (see refs. 12,21 for a discussion of genome-wide association significance). Both of these markers reside in the gene *ARTS1* (*ERAAP*, *ERAP1*), which encodes a type II integral transmembrane aminopeptidase with diverse immunological functions. Four additional SNPs show significance at $P < 10^{-4}$, with an increasing number of possible associations at more modest significance levels. Several of the more strongly associated SNPs, and others in the same genes, have been previously associated with these particular diseases, and for yet others there exists functional evidence of involvement in these particular conditions. Among these are SNPs in *FCRL3* and *FCRL5* in the case of AITD, *IL23R* in the case of ankylosing spondylitis, *MEL18* in the case of breast cancer and *IL7R* for MS. The complete list of single-marker association results is provided in **Supplementary Table 1**.

The results of analyses involving the expanded reference group are presented in **Supplementary Figure 5** online and **Supplementary**

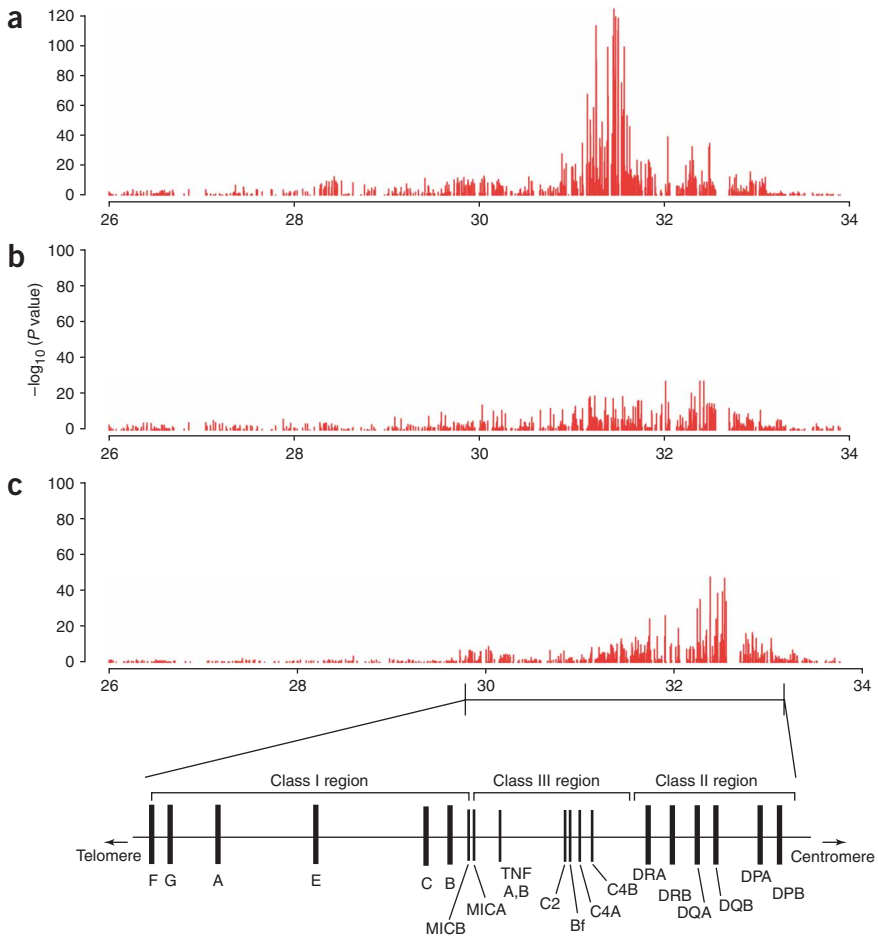

Figure 1 Minus $\log_{10}$ *P* values for the Armitage test of trend for MHC association with ankylosing spondylitis (**a**), autoimmune thyroid disease (**b**) and multiple sclerosis (**c**). Note in particular how evidence for association extends along very long regions of the MHC, reflecting statistical power to detect association even when linkage disequilibrium amongst SNPs is relatively low or when there exists the possibility of multiple disease-predisposing loci.

Table 2 Estimates of λ for single and combined cohorts

		λ
Single cohort	AS cases versus 58C	1.07
	AITD cases versus 58C	1.12
	BC cases versus 58C	1.13
	MS cases versus 58C	1.12
Mixed cohorts	AS cases versus all others	1.03
	AITD cases versus all others	1.05
	BC cases versus all others	1.04
	MS cases versus all others	1.06
	IMMUNE cases versus BC and 58C	1.04

Table 1. Many of the SNPs that showed moderate to strong evidence for association in the initial analysis had substantially greater significance when the larger reference group was used. Notably, these included the SNPs rs27044 ($P = 4.0 \times 10^{-8}$) and rs30187 ($P = 2.1 \times 10^{-7}$) in *ARTS1*, as well as several other variants in this gene. A second SNP, rs7302230 in the gene encoding calsyntenin-3 on chromosome 12, showed substantially stronger evidence for association in the expanded reference group analysis ($P = 5.3 \times 10^{-7}$) relative to the 58C-only results ($P = 1.1 \times 10^{-4}$). Results of the expanded group also showed elevated results for several SNPs that did not appear exceptional in the original (non-combined) analyses, including SNPs in several candidate genes such as those encoding sialoadhesin[22] and complement receptor 1 for ankylosing spondylitis, *PIK3R2* for MS, and *C8B*, *IL17R* and *TYK2* in the combined autoimmune disease analysis. SNP rs3783941 in the gene *TSHR*, encoding the thyroid-stimulating hormone receptor, emerged as among the most significant in the expanded reference group analyses of AITD ($P = 2.1 \times 10^{-5}$). Several polymorphisms in *TSHR* have previously been associated with Graves' disease[23,24]. This known association did not reach even the modest significance level of 10^{-3} in the original analyses, but the addition of 3,000 further reference samples delineated it from the background noise and further supports the original independent report.

ARTS1 association confirmed in an independent cohort

To validate the most exceptional findings from the initial study, we genotyped the *ARTS1*, *CLSTN3* and *LNPEP* SNPs in 471 independent ankylosing spondylitis cases (**Table 4**) and 625 new controls (all self identified North American Caucasian). The data strongly suggest that the *ARTS1* association is genuine. All *ARTS1* nsSNPs revealed independent replication in the same direction of effect, with replication significance levels ranging from 4.7×10^{-4} to 5.1×10^{-5}. When combined with the original samples, the results showed strong evidence for association with ankylosing spondylitis ($P = 1.2 \times 10^{-8}$ to 3.4×10^{-10}). The population attributable risk[25] contributed by the most strongly associated marker in the North American dataset (rs2287987) was 26%.

Association was also confirmed with marker rs2303138 in the *LNPEP* gene, which lies 127 kb 3′ of *ARTS1*. This marker was in strong LD with *ARTS1* markers ($D' = 1$, rs27044–rs2303138). We tested the interdependence of the *ARTS1* and *LNPEP* associations using conditional logistic regression. The remaining association at *LNPEP* was weak after controlling for *ARTS1* ($P = 0.01$), whereas the association at *ARTS1* remained strong after controlling for *LNPEP* ($P = 2.7 \times 10^{-6}$), suggesting that the *LNPEP* association may only be secondary to LD, with a true association at *ARTS1*.

No association was seen with *CLSTN3* in the confirmation set. The US controls showed the same allele frequency as the UK controls (5%), but the allele frequency in the US cases was less than that of the UK cases (6% versus 8%), suggesting no association in the US samples and substantially reducing the significance of the combined data. Calystenin-3 is a postsynaptic neuronal membrane protein and is an unlikely candidate for involvement in inflammatory arthritis. The failure to replicate this association suggests that our replication sample size was insufficient to detect the modest effect or that it was a false positive in the initial scan.

IL23R variants confer risk of ankylosing spondylitis

The *IL23R* variant rs11209026, although not notable in the initial nsSNP scan ($P = 1.7 \times 10^{-3}$), was of particular interest, as it has recently been associated with both Crohn's disease[26,27] and psoriasis[28], conditions that commonly co-occur with ankylosing spondylitis. To better define this association, seven additional SNPs in *IL23R* were genotyped in the same 1,000 British ankylosing spondylitis cases and 1,500 58C controls as well as the North American Caucasian replication samples (**Table 4**). In the WTCCC dataset, we observed strong association in seven of eight genotyped SNPs ($P \leq 0.008$, including the original nsSNP rs11209026), with the strongest association at rs11209032 ($P = 2.0 \times 10^{-6}$). In the replication dataset, we noted association with all genotyped SNPs ($P \leq 0.04$), with peak association with marker rs10489629 ($P = 4.2 \times 10^{-5}$). In the combined dataset,

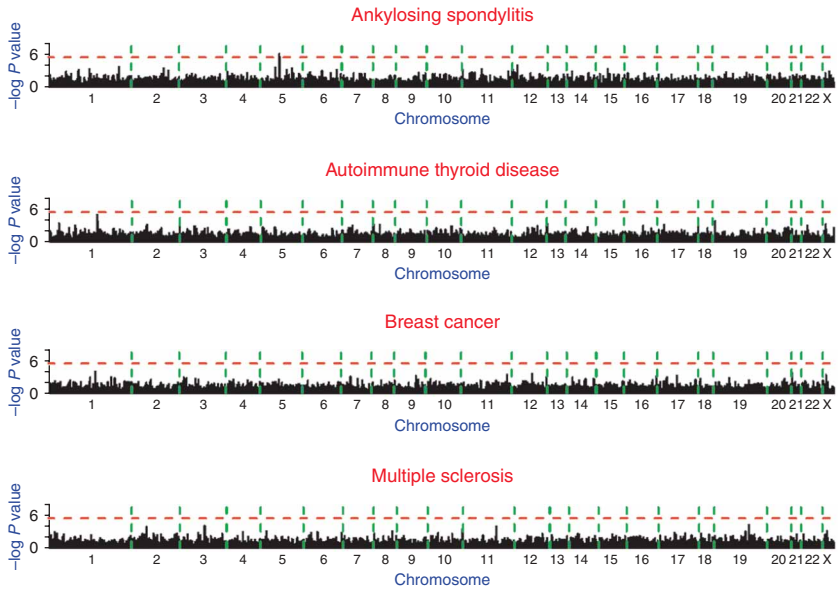

Figure 2 Minus $\log_{10}$ *P* values for the Armitage test of trend for genome-wide association scans of ankylosing spondylitis, autoimmune thyroid disease, breast cancer and multiple sclerosis. The spacing between SNPs on the plot is uniform and does not reflect distances between the SNPs. The vertical dashed lines reflect chromosomal boundaries. The horizontal dashed lines display the cutoff of $P = 10^{-6}$. Note that SNPs within the MHC are not included in this diagram.

the strongest association observed was with SNP rs11209032 (odds ratio 1.3, 95% confidence interval 1.2–1.4, $P = 7.5 \times 10^{-9}$). The attributable risk for this marker in the replication cohort is 9%. Conditional logistic regression analyses did not indicate a single primary disease-associated marker; residual association remained after we controlled for association at the remaining SNPs. Considering only individuals with ankylosing spondylitis who self-reported as not having inflammatory bowel disease ($n = 1{,}066$) the association remained strong and was still strongest at rs11209032 ($P = 6.9 \times 10^{-7}$), indicating that there is a primary association with ankylosing spondylitis and that the observed association was not due to coexistent clinical inflammatory bowel disease.

In contrast to the pleiotropic effects of *IL23R*, the *ARTS1* association evidence seems confined to ankylosing spondylitis. We genotyped

Table 3 nsSNPs outside the MHC that meet a point-wise significance level of $P < 10^{-3}$ for the Cochran-Armitage test for trend

Disease	SNP	Chromosome	Position (bp)	MAF	OR	χ^2	*P* value	Gene
AS	rs696698	1	74777462	0.04	1.84	11.13	8.5×10^{-4}	*C1orf173*
	rs10494217	1	119181230	0.17	0.77	11.62	6.5×10^{-4}	*TBX15*
	rs2294851	1	206966279	0.13	0.73	13.55	2.3×10^{-4}	*HHAT*
	rs8192556	2	182368504	0.01	0.45	12.24	4.7×10^{-4}	*NEUROD1*
	rs16876657	5	78645930	0.02	3.10	13.05	3.0×10^{-4}	*JMY*
	rs27044	5	96144608	0.34	1.40	23.90	1.0×10^{-6}	*ARTS-1*
	rs17482078	5	96144622	0.17	0.76	13.55	2.3×10^{-4}	*ARTS-1*
	rs10050860	5	96147966	0.18	0.75	14.87	1.1×10^{-4}	*ARTS-1*
	rs30187	5	96150086	0.40	1.33	21.82	3.0×10^{-6}	*ARTS-1*
	rs2287987	5	96155291	0.18	0.75	14.31	1.6×10^{-4}	*ARTS-1*
	rs2303138	5	96376466	0.10	1.58	19.41	1.1×10^{-5}	*LNPEP*
	rs11750814	5	137528564	0.16	0.77	10.99	9.1×10^{-4}	*BRD8*
	rs11959820	5	149192703	0.02	0.49	12.41	4.3×10^{-4}	*PPARGC1B*
	rs907609	11	1813846	0.13	0.76	10.91	9.5×10^{-4}	*SYT8*
	rs3740691	11	47144987	0.29	0.80	11.86	5.7×10^{-4}	*ZNF289*
	rs11062385	12	297836	0.24	0.79	11.82	5.9×10^{-4}	*JARID1A*
	rs7302230	12	7179699	0.08	1.57	14.97	1.1×10^{-4}	*CLSTN3*
AITD	rs10916769	1	20408244	0.17	0.76	12.10	5.0×10^{-4}	*FLJ32784*
	rs6427384	1	154321955	0.18	1.43	18.97	1.3×10^{-5}	*FCRL5*
	rs2012199	1	154322098	0.17	1.35	13.18	2.8×10^{-4}	*FCRL5*
	rs6679793	1	154327170	0.22	1.33	14.69	1.3×10^{-4}	*FCRL5*
	rs7522061	1	154481463	0.47	1.25	13.78	2.1×10^{-4}	*FCRL3*
	rs1047911	2	74611433	0.15	1.34	11.24	8.0×10^{-4}	*MRPL53*
	rs7578199	2	241912838	0.26	1.26	11.53	6.9×10^{-4}	*HDLBP*
	rs3748140	8	9036429	0.00	0.28	11.44	7.2×10^{-4}	*PPP1R3B*
	rs1048101	8	26683945	0.42	0.82	10.98	9.2×10^{-4}	*ADRA1A*
	rs7975069	12	132389146	0.30	0.80	12.06	5.2×10^{-4}	*ZNF268*
	rs2271233	17	6644845	0.07	0.94	11.32	7.7×10^{-4}	*TEKT1*
	rs2856966	18	897710	0.19	0.76	14.00	1.8×10^{-4}	*ADCYAP1*
	rs7250822	19	2206311	0.04	1.97	13.83	2.0×10^{-4}	*AMH*
	rs2230018	23	44685331	0.14	1.41	11.55	6.8×10^{-4}	*UTX*
BC	rs4255378	1	151919300	0.48	1.25	14.70	1.3×10^{-4}	*MUC1*
	rs2107732	7	44851218	0.10	1.40	10.96	9.3×10^{-4}	*CCM2*
	rs4986790	9	117554856	0.07	1.54	11.46	7.1×10^{-4}	*TLR4*
	rs2285374	11	118457383	0.38	0.82	12.25	4.7×10^{-4}	*VPS11*
	rs7313899	12	54231386	0.03	2.10	13.02	3.1×10^{-4}	*OR6C4*
	rs2879097	17	34143085	0.20	0.78	11.73	6.1×10^{-4}	*MEL18*
	rs2822558	21	14593715	0.13	0.73	13.87	2.0×10^{-4}	*ABCC13*
	rs2230018	23	44685331	0.14	1.40	12.14	4.9×10^{-4}	*UTX*
MS	rs17009792	2	74400978	0.02	0.44	14.41	1.5×10^{-4}	*SLC4A5*
	rs1132200	3	120633526	0.15	0.73	15.22	9.6×10^{-5}	*FLJ10902*
	rs6897932	5	35910332	0.23	0.80	11.04	8.9×10^{-4}	*IL7R*
	rs6470147	8	124517985	0.36	1.23	10.92	9.5×10^{-4}	*FLJ10204*
	rs3818511	10	134309378	0.24	1.28	12.84	3.4×10^{-4}	*INPP5A*
	rs11574422	11	67970565	0.02	2.82	14.64	1.3×10^{-4}	*LRP5*
	rs388706	19	49110533	0.48	1.22	11.19	8.2×10^{-4}	*ZNF45*
	rs1800437	19	50873232	0.17	0.74	16.11	6.0×10^{-5}	*GIPR*
	rs2281868	23	69451484	0.50	1.26	11.38	7.4×10^{-4}	*SAP102*

Table 4 Ankylosing spondylitis replication results

		UK cases				US cases				All cases			
Gene	SNP	Case MAF	Control MAF	OR	*P* value	Case MAF	Control MAF	OR	*P* value	Case MAF	Control MAF	OR	*P* value
ARTS1	rs27044	0.34	0.27	1.40	1.0×10^{-6}	–	–	–	–	–	–	–	–
ARTS1	rs17482078	0.17	0.22	0.76	2.3×10^{-4}	0.15	0.21	0.65	5.1×10^{-5}	0.16	0.22	0.70	1.2×10^{-8}
ARTS1	rs10050860	0.18	0.23	0.75	1.2×10^{-4}	0.15	0.22	0.66	8.8×10^{-5}	0.17	0.22	0.71	7.6×10^{-9}
ARTS1	rs30187	0.40	0.33	1.33	3.0×10^{-6}	0.41	0.35	1.30	0.00047	0.41	0.34	1.40	3.4×10^{-10}
ARTS1	rs2287987	0.18	0.22	0.75	1.6×10^{-4}	0.15	0.21	0.66	8.4×10^{-5}	0.17	0.22	0.71	1.0×10^{-8}
LNPEP	rs2303138	0.10	0.07	1.58	1.1×10^{-5}	0.11	0.09	1.40	0.018	0.11	0.07	1.48	1.1×10^{-6}
CLSTN3	rs7302230	0.08	0.05	1.57	1.1×10^{-4}	0.06	0.05	1.10	0.56	0.07	0.05	1.30	0.0039
IL23R	rs11209026	0.04	0.06	0.63	0.0017	0.038	0.06	0.63	0.014	0.04	0.06	0.63	4.0×10^{-6}
IL23R	rs1004819	0.35	0.30	1.20	0.0013	0.35	0.30	1.30	0.0045	0.35	0.30	1.20	1.1×10^{-5}
IL23R	rs10489629	0.43	0.45	0.90	0.062	0.39	0.47	0.72	4.2×10^{-5}	0.41	0.46	0.83	0.00011
IL23R	rs11465804	0.04	0.06	0.67	0.0019	0.049	0.06	0.68	0.04	0.04	0.06	0.68	0.0002
IL23R	rs1343151	0.30	0.34	0.85	0.0077	0.29	0.36	0.71	6.7×10^{-5}	0.30	0.34	0.80	1.0×10^{-5}
IL23R	rs10889677	0.36	0.31	1.20	0.00066	0.37	0.29	1.40	4.7×10^{-5}	0.36	0.31	1.30	1.3×10^{-6}
IL23R	rs11209032	0.38	0.32	1.30	2.0×10^{-6}	0.38	0.32	1.30	0.0013	0.38	0.32	1.30	7.5×10^{-9}
IL23R	rs1495965	0.49	0.44	1.20	0.0021	0.50	0.43	1.40	0.00019	0.49	0.44	1.20	3.1×10^{-6}

the five ankylosing spondylitis–associated SNPs in 755 British Crohn's disease and 1,011 ulcerative colitis cases and 633 healthy controls. No association was seen with either ulcerative colitis or Crohn's disease (Armitage trend $P > 0.4$ for all markers).

FCRL3 confirmed in AITD pathogenesis

In addition to the ankylosing spondylitis replications, we attempted to confirm and extend the *FCRL3* association in AITD. The SNP rs7522061 in the *FCRL3* gene was recently reported to be associated with AITD[29] and two other autoimmune diseases, rheumatoid arthritis and systemic lupus erythematosus[30]. Our initial association evidence ($P = 2.1 \times 10^{-4}$) likely reflects the signal of the originally detected polymorphism, because the level of LD is high across this gene. In fact, the entire 1q21–q23 region (which includes another gene, *FCRL5*, flagged in our scan) has also been implicated in several autoimmune diseases, including psoriasis and multiple sclerosis[31,32].

On the basis of the original findings on 1q21–q23, the original cohort was increased from 1,000 to 2,500 Graves disease cases, and we used 2,500 controls from the 58C control set. We selected eight SNPs that tagged the *FCRL3* and *FCRL5* gene regions and typed them in all 5,000 samples using an alternative genotyping platform. SNP rs3761959, which tags rs7522061 and rs7528684 (previously associated with rheumatoid arthritis and Graves' disease), was associated with Graves' disease in this extended cohort (**Table 5**), confirming the original result. In total, three of the seven *FCRL3* SNPs showed some evidence for association ($P < 0.05$), with SNP rs11264798 showing the strongest association of the tag SNPs ($P = 4.0 \times 10^{-3}$). SNP rs6667109 in *FCRL5*, which tagged SNPs rs6427384, rs2012199 and rs6679793, all found to be weakly associated in the original study, showed little evidence of association in this extended cohort.

DISCUSSION

Our scan of nsSNPs has identified and validated two new genes (*ARTS1* and *IL23R*) associated with ankylosing spondylitis, confirmed and extended markers in the *TSHR* and *FCRL3* genes that have previously been associated with AITD, and provided a dense set of association data for AITD, ankylosing spondylitis and MS across the MHC region. The challenge now is to design functional studies that

Table 5 Autoimmune thyroid disease replication results

		Replication cohort				Combined cohort			
Gene	SNP	Case MAF	Control MAF	OR	*P* value	Case MAF	Control MAF	OR	*P* value
FCRL3	rs3761959[a]	0.48	0.45	0.87	0.013	0.49	0.45	0.87	9.4×10^{-3}
FCRL3	rs11264794	0.42	0.45	1.10	0.079	0.42	0.46	1.12	0.013
FCRL3	rs11264793	0.27	0.24	0.87	0.029	0.26	0.24	0.90	0.044
FCRL3	rs11264798	0.44	0.49	1.18	4.0×10^{-3}	0.44	0.49	1.22	1.6×10^{-5}
FCRL3	rs10489678	0.19	0.20	1.04	0.58	0.20	0.20	1.04	0.43
FCRL3	rs6691569	0.28	0.29	1.02	0.75	0.29	0.29	1.00	0.93
FCRL3	rs2282284	0.062	0.058	0.92	0.015	0.062	0.058	0.93	0.47
FCRL5	rs6667109	0.17	0.16	0.93	0.38	0.18	0.15	0.85	7.7×10^{-2}

[a]This SNP tags the SNP rs7522061, which was flagged as associated with AITD in the WTCCC screen ($P = 2.1 \times 10^{-4}$).

will reveal how variation in these genes translates into physiological processes that influence disease risk.

From a functional perspective, *ARTS1* and *IL23R* represent excellent biological candidates for association with ankylosing spondylitis. The protein ARTS1 has two known functions, either of which may explain the association. First, within the endoplasmic reticulum, ARTS1 is involved in trimming peptides to the optimal length for MHC class I presentation[33,34]. Ankylosing spondylitis is primarily an HLA class I–mediated autoimmune disease[35], with >90% of cases carrying the HLA-B27 allele. How HLA-B27 increases risk of developing ankylosing spondylitis is unknown, but if the association of *ARTS1* with the disease relates to effects of ARTS1 on peptide presentation, this relationship would inform research into the mechanism underlying the association of HLA-B27 with ankylosing spondylitis. Second, ARTS1 cleaves cell surface receptors for the pro-inflammatory cytokines IL-1 (IL-1R2)[36], IL-6 (IL-6Rα)[37] and TNF (TNFR1)[38], thereby downregulating their signaling. Genetic variants that alter the functioning of ARTS1 could therefore have pro-inflammatory effects through this mechanism.

In addition to their association with ankylosing spondylitis, polymorphisms in *IL23R* have been recently documented in Crohn's disease[26,27] and psoriasis[28], suggesting that this gene is a common susceptibility factor for the major 'seronegative' diseases, at least partially explaining their co-occurrence. IL-23R is a key factor in the regulation of a newly defined effector T-cell subset, T_H17 cells. T_H17 cells were originally identified as a distinct subset of T-cells expressing high levels of the pro-inflammatory cytokine IL-17 in response to stimulation, in addition to IL-1, IL-6, TNFα, IL-22 and IL-25 (IL-17E). IL-23 has been shown to be important in the mouse models of experimental autoimmune encephalomyelitis[39], collagen-induced arthritis[40] and inflammatory bowel disease[41], but it has not been studied in ankylosing spondylitis, either in human or other animal models of the disease. These studies show that blocking IL-23 reduces inflammation in these models, suggesting that the *IL23R* variants associated with disease are pro-inflammatory. Successful treatment of Crohn's disease has been reported with anti–IL-12p40 antibodies, which block both IL-12 and IL-23, as these cytokines share the IL-12p40 chain[42]. No functional studies of *IL23R* variants have been reported to date, and it is unclear to what extent findings in studies targeting IL-23 can be generalized to mechanisms by which *IL23R* variation affects disease susceptibility. Our genetic findings provide notable insight into the etiopathogenesis of ankylosing spondylitis and suggest that treatments targeting IL-23 may prove effective for this condition, but clearly much more needs to be understood about the mechanism underlying the observed association.

Despite the successful identification of the *ARTS1* and *IL23R* genes, it is likely either that additional real associations are present in our data but were overlooked because of their modest effect sizes, or that our focus on non-synonymous coding changes led us to miss real loci. The issue of limited statistical power is emphasized in studies of nonsynonymous coding changes, which have a greater number of rare variants than other genetic variants and thus will require even larger sample sizes unless the effect sizes are larger. Other analytical approaches, such as assessing evidence for association between clusters of rare variants rather than individual loci, may prove highly informative in this regard[43], but most of the nsSNPs available in this study exist either by themselves in each gene or with one or two others, which precludes these assessments (**Supplementary Fig. 6** online). In our analyses, *ARTS1* was the only locus showing exceptional statistical significance in the scan of 1,000 cases and 1,500 controls, thus emphasizing the need for greater statistical power. We increased power by expanding the controls, or 'reference set,' to include some or all of the other disease samples. When we did so, *ARTS1* showed even stronger association evidence, the *IL23R* SNPs increased to a level that began to delineate them from background noise, and the AITD/*TSHR* confirmation emerged. This demonstration of increased statistical power through the combination of multiple datasets is timely, given the international impetus to make genotype data available to the scientific community. Future investigations will be needed to assess the power versus confounding effects and the statistical corrections needed to combine more heterogeneous samples from broader sampling regions.

These results also highlight the question of how much information may be missed by focusing on coding SNPs rather than searching more broadly over the genome at large. This question is relevant because the tradeoff between SNP panel and sample size selection is a salient factor in the design of every genome-wide study. In the HapMap data[44], a substantial portion of the common nonsynonymous variation in our nsSNP set is captured by available genome-wide panels (about 65% of common (MAF > 5%) nsSNPs in the Illumina Human NS-12 Beadchip are tagged with an $r^2 > 0.8$ using the Affymetrix 500 K chip, rising to 90% in the Illumina HumanHap300, which includes almost all of the nsSNPs from the NS-12 Beadchip). The four primary associated variants flagged in our study (that is, in *ARTS1*, *IL23R*, *TSHR* and *FCRL3*) would have been detected using any of the genome-wide panels, because either the markers themselves or a SNP in high LD with them ($r^2 \geq 0.78$) are present on the genome-wide chips. This LD relationship also emphasizes the fact that observing an association with a nsSNP does not necessarily imply that the nsSNP is causal, as it may be indirectly associated with other genetic variants in or outside the gene. Given this high degree of overlap, the continuously increasing coverage of many available genotyping products and concomitant pressures to decrease assay costs, these data suggest that future gene-centric scans will be efficiently subsumed by the more encompassing and less hypothesis-driven genome-wide SNP panels.

METHODS

Subjects. Individuals included in the study were self-identified as white and of European ancestry and came from mainland UK (England, Scotland and Wales, but not Northern Ireland). The 1,500 control samples were from the British 1958 Birth Cohort (58C, also known as the National Child Development Study), which included all the births in England, Wales and Scotland that occurred during 1 week in 1958. Recruitment details and diagnostic criteria for each of the four case groups, as well as for the North American AS replication cohort and the 58C are further described in the **Supplementary Methods** online.

Sample quality assurance and control genome-wide identity by state (IBS) sharing was calculated for each pair of individuals in the combined sample of cohorts to identify first- and second-degree relatives whose data might contaminate the study. One subject from any pair of individuals who shared <400 genotypes IBS = 0 and/or >80% alleles IBS (that is, the individual with the most missing genotypes) was removed from all subsequent analyses. To identify individuals who might have ancestries other than Western European, we merged each of our cohorts with the 60 western European (CEU) founder, 60 Nigerian (YRI) founder, and 90 Japanese (JPT) and Han Chinese (CHB) individuals from the International HapMap Project[44]. We calculated genome-wide IBD distances for each pair of individuals (that is, 1 minus average IBS sharing) on those markers shared between HapMap and our nonsynonymous panel, and then used the multidimensional scaling option in R to generate a two dimensional plot based upon individuals' scores on the first two principal coordinates from this analysis (**Supplementary Fig. 2**). Any WTCCC sample that was not present in the main cluster with the CEU individuals was excluded from subsequent analyses. Finally, any individual with >10%

of genotypes missing was removed from the analysis. The number of individuals remaining after these quality control measures were applied is shown in **Table 1**.

Genotyping. We genotyped a total of 14,436 nsSNPs across the genome on all case and control samples. Because three of the diseases were of autoimmune etiology, we also typed an additional 897 SNPs within the MHC region, as well as 103 SNPs in pigmentation genes specifically designed to differentiate between population groups. SNP genotyping was performed with the Infinium I assay (Illumina), which is based on allele-specific primer extension (ASPE) and the use of a single fluorochrome. The assay requires ~250 ng of genomic DNA, which is first subjected to a round of isothermal amplification that generates a 'high-complexity' representation of the genome with most loci represented at usable amounts. There are two allele-specific probes (50-mers) per SNP, each on a different bead type; each bead type is present on the array an average of 30 times (and a minimum of 5 times), allowing for multiple independent measurements. We processed six samples per array. Clustering was carried out with the GenCall software version 6.2.0.4, which assigns a quality score to each locus and an individual genotype confidence score that is based on the distance of a genotype from the center of the nearest cluster. First, we removed samples with more than 50% of loci having a quality score below 0.7 and then all loci with a quality score below 0.2. After clustering, we applied two additional filtering criteria: (i) we omitted individual genotypes with a genotype confidence score <0.15 and (ii) we removed any SNP for which more than 20% of samples had genotype confidence scores <0.15. The above criteria were designed to optimize genotype accuracy and minimize uncalled genotypes.

Statistical analysis markers that were monomorphic in both case and control samples, SNPs with $>10\%$ missing genotypes and SNPs with differences in the amount of missing data between cases and controls ($P < 10^{-4}$ as assessed by χ^2 test) were excluded from all analyses involving that case group only. In addition, any marker that failed an exact test of Hardy-Weinberg equilibrium in controls ($P < 10^{-7}$) was excluded from all analyses[45].

Cochran-Armitage tests for trend[46] were conducted using the PLINK program[47]. For the present analyses, we used the significance thresholds of $P < 10^{-4}$–10^{-6}, as suggested for gene-based scans with stronger prior probabilities than scans of anonymous markers[21]. In the present context, the lower thresholds are similar to Bonferroni significance levels (Bonferroni-corrected $P = 0.05$ corresponds to nominal $P = 3 \times 10^{-6}$). The conditional logistic regression analyses involving the *LNPEP* and *ARTS1* SNPs were carried out using Purcell's WHAP program[48].

We manually rechecked the genotype calls of every nsSNP with an asymptotic significance level of $P < 10^{-3}$ by inspecting raw signal intensity values and their corresponding automated genotype calls. Notably, this flagged an additional 33 markers with clear problems in genotype calling, which were subsequently excluded from all analyses (**Supplementary Fig. 4**). These results indicate that this genotyping platform generally yields highly accurate genotypes, but errors do occur and can be distributed nonrandomly between cases and controls despite stringent quality control procedures. It is imperative to check the clustering of the most significant SNPs to ensure that evidence for associations is not a result of genotyping error.

Although great lengths were taken to ensure that our samples were as homogenous as possible in terms of genetic ancestry, even subtle population substructure can substantially influence tests of association in large genome-wide analyses involving thousands of individuals[49]. We therefore calculated the genomic-control inflation factor, λ (ref. 20), for each case-control sample as well as in the analyses where we combined the other case groups with the control individuals (**Table 2**). In general, values for λ were small (~1.1), indicating a small degree of substructure in UK samples that induces only a slight inflation of the test statistic under the null hypothesis, consistent with the results from our companion paper[12]. We therefore present uncorrected results in all analyses reported.

Consent was granted from ethical review boards of the institutions with which the participants were affiliated, and informed consent was obtained from the individuals involved in the WTCCC. Individual-level data from this study will be widely available through the Consortium's Data Access Committee (http://www.wtccc.org.uk).

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

We would like to thank the Wellcome Trust for supporting this study, and all the individuals and controls who participated in this study.

AITD: We thank the collection coordinators, J. Carr-Smith and all contributors to the AITD national DNA collection of index cases and family members from centres including Birmingham, Bournemouth, Cambridge, Cardiff, Exeter, Leeds, Newcastle and Sheffield. Principal leads for the AITD UK national collection are S.C. Gough (Birmingham), S.H.S. Pearce (Newcastle), B. Vaidya (Exeter), J.H. Lazarus (Cardiff), A. Allahabadia (Sheffield), M. Armitage (Bournemouth), P.J. Grant (Leeds) and V.K. Chatterjee (Cambridge).

Ankylosing spondylitis: We thank the Arthritis Research Campaign (UK). MAB is funded by the National Health and Medical Research Council (Australia). TASC is funded by the National Institute of Arthritis and Musculoskeletal and Skin Diseases grants 1PO1-052915-01,RO1 AR046208 and RO1-AR048465, as well as by University of Texas at Houston CTSA grant UL1RR024148, Cedars-Sinai GCRC grant MO1-RR00425, The Rosalind Russell Center for Arthritis Research at The University of California San Francisco, and the Intramural Research Program, National Institute of Arthritis and Musculoskeletal and Skin Diseases, US National Institutes of Health. We thank R. Jin for technical assistance and L. Diekman, L. Guthrie, F. Lin and S. Morgan for their study coordination.

Breast cancer: The breast cancer samples were clinically and molecularly curated with the assistance of A. Renwick, A. Hall, A. Elliot, H. Jayatilake, T. Chagtai, R. Barfoot, P. Kelly and K. Spanova. Our research is supported by United States Army Medical Research and Material Command grant no. W81XWH-05-1-0204, The Institute of Cancer Research and Cancer Research UK.

MS: Our work has been supported by the Wellcome Trust (grant ref. 057097), the Medical Research Council (UK) (grant ref. G0000648), the Multiple Sclerosis Society of Great Britain and Northern Ireland (grant ref. 730/02) and the National Institutes of Health (USA) (grant ref. 049477). A.G. is a postdoctoral fellow of the Research Foundation–Flanders (FWO–Vlaanderen).

L. Galver and P. Ng at Illumina and J. Morrison at the Sanger Institute contributed in the design of the nsSNP array. We thank the DNA team of the JDRF/WT DIL and T. Dibling, C. Hind and D. Simpkin at the Sanger Institute for carrying out the genotyping. We also thank S. Bingham and the WTCCC Inflammatory Bowel Disease group for genotyping the *ARTS1* markers in their replication samples.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Rioux, J.D. *et al.* Genome-wide association study identifies new susceptibility loci for Crohn disease and implicates autophagy in disease pathogenesis. *Nat. Genet.* **39**, 596–604 (2007).
2. Sladek, R. *et al.* A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* **445**, 881–885 (2007).
3. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
4. Libioulle, C. *et al.* Novel Crohn disease locus identified by genome-wide association maps to a gene desert on 5p13.1 and modulates expression of PTGER4. *PLoS Genet.* **3**, e58 (2007).
5. Zanke, B.W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39**, 989–994 (2007).
6. Haiman, C.A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* **39**, 638–644 (2007).
7. Gudmundsson, J. *et al.* Two variants on chromosome 17 confer prostate cancer risk, and the one in TCF2 protects against type 2 diabetes. *Nat. Genet.* **39**, 977–983 (2007).
8. Moffatt, M.F. *et al.* Genetic variants regulating ORMDL3 expression contribute to the risk of childhood asthma. *Nature* **448**, 470–473 (2007).
9. Zeggini, E. *et al.* Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* **316**, 1336–1341 (2007).
10. Scott, L.J. *et al.* A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* **316**, 1341–1345 (2007).
11. Saxena, R. *et al.* Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* **316**, 1331–1336 (2007).
12. WTCCC. Genome-wide association studies of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–683 (2007).
13. Hampe, J. *et al.* A genome-wide association scan of nonsynonymous SNPs identifies a susceptibility variant for Crohn disease in ATG16L1. *Nat. Genet.* **39**, 207–211 (2007).
14. Jorgenson, E. & Witte, J.S. Coverage and power in genomewide association studies. *Am. J. Hum. Genet.* **78**, 884–888 (2006).
15. Smyth, D.J. *et al.* A genome-wide association study of nonsynonymous SNPs identifies a type 1 diabetes locus in the interferon-induced helicase (IFIH1) region. *Nat. Genet.* **38**, 617–619 (2006).

16. Miretti, M.M. *et al.* A high-resolution linkage-disequilibrium map of the human major histocompatibility complex and first generation of tag single-nucleotide polymorphisms. *Am. J. Hum. Genet.* **76**, 634–646 (2005).
17. Clayton, D.G. *et al.* Population structure, differential bias and genomic control in a large-scale, case-control association study. *Nat. Genet.* **37**, 1243–1246 (2005).
18. Sims, A.M. *et al.* Non-B27 MHC associations of ankylosing spondylitis. *Genes Immun.* **8**, 115–123 (2007).
19. McGinnis, R., Shifman, S. & Darvasi, A. Power and efficiency of the TDT and case-control design for association scans. *Behav. Genet.* **32**, 135–144 (2002).
20. Devlin, B. & Roeder, K. Genomic control for association studies. *Biometrics* **55**, 997–1004 (1999).
21. Thomas, D.C. & Clayton, D.G. Betting odds and genetic associations. *J. Natl. Cancer Inst.* **96**, 421–423 (2004).
22. Jiang, H.R. *et al.* Sialoadhesin promotes the inflammatory response in experimental autoimmune uveoretinitis. *J. Immunol.* **177**, 2258–2264 (2006).
23. Dechairo, B.M. *et al.* Association of the TSHR gene with Graves' disease: the first disease specific locus. *Eur. J. Hum. Genet.* **13**, 1223–1230 (2005).
24. Hiratani, H. *et al.* Multiple SNPs in intron 7 of thyrotropin receptor are associated with Graves' disease. *J. Clin. Endocrinol. Metab.* **90**, 2898–2903 (2005).
25. Miettinen, O.S. Proportion of disease caused or prevented by a given exposure, trait or intervention. *Am. J. Epidemiol.* **99**, 325–332 (1974).
26. Duerr, R.H. *et al.* A Genome-Wide Association Study Identifies IL23R as an Inflammatory Bowel Disease Gene. *Science* (2006).
27. Tremelling, M. *et al.* IL23R variation determines susceptibility but not disease phenotype in inflammatory bowel disease. *Gastroenterology* **132**, 1657–1664 (2007).
28. Cargill, M. *et al.* A large-scale genetic association study confirms IL12B and leads to the identification of IL23R as psoriasis-risk genes. *Am. J. Hum. Genet.* **80**, 273–290 (2007).
29. Simmonds, M.J. *et al.* Contribution of single nucleotide polymorphisms within FCRL3 and MAP3K7IP2 to the pathogenesis of Graves' disease. *J. Clin. Endocrinol. Metab.* **91**, 1056–1061 (2006).
30. Kochi, Y. *et al.* A functional variant in FCRL3, encoding Fc receptor-like 3, is associated with rheumatoid arthritis and several autoimmunities. *Nat. Genet.* **37**, 478–485 (2005).
31. Capon, F. *et al.* Fine mapping of the PSORS4 psoriasis susceptibility region on chromosome 1q21. *J. Invest. Dermatol.* **116**, 728–730 (2001).
32. Dai, K.Z. *et al.* The T cell regulator gene SH2D2A contributes to the genetic susceptibility of multiple sclerosis. *Genes Immun.* **2**, 263–268 (2001).
33. Chang, S.C., Momburg, F., Bhutani, N. & Goldberg, A.L. The ER aminopeptidase, ERAP1, trims precursors to lengths of MHC class I peptides by a "molecular ruler" mechanism. *Proc. Natl. Acad. Sci. USA* **102**, 17107–17112 (2005).
34. Saveanu, L. *et al.* Concerted peptide trimming by human ERAP1 and ERAP2 aminopeptidase complexes in the endoplasmic reticulum. *Nat. Immunol.* **6**, 689–697 (2005).
35. Brown, M.A. *et al.* HLA class I associations of ankylosing spondylitis in the white population in the United Kingdom. *Ann. Rheum. Dis.* **55**, 268–270 (1996).
36. Cui, X., Rouhani, F.N., Hawari, F. & Levine, S.J. Shedding of the type II IL-1 decoy receptor requires a multifunctional aminopeptidase, aminopeptidase regulator of TNF receptor type 1 shedding. *J. Immunol.* **171**, 6814–6819 (2003).
37. Cui, X., Rouhani, F.N., Hawari, F. & Levine, S.J. An aminopeptidase, ARTS-1, is required for interleukin-6 receptor shedding. *J. Biol. Chem.* **278**, 28677–28685 (2003).
38. Cui, X. *et al.* Identification of ARTS-1 as a novel TNFR1-binding protein that promotes TNFR1 ectodomain shedding. *J. Clin. Invest.* **110**, 515–526 (2002).
39. Cua, D.J. *et al.* Interleukin-23 rather than interleukin-12 is the critical cytokine for autoimmune inflammation of the brain. *Nature* **421**, 744–748 (2003).
40. Murphy, C.A. *et al.* Divergent pro- and antiinflammatory roles for IL-23 and IL-12 in joint autoimmune inflammation. *J. Exp. Med.* **198**, 1951–1957 (2003).
41. Hue, S. *et al.* Interleukin-23 drives innate and T cell-mediated intestinal inflammation. *J. Exp. Med.* **203**, 2473–2483 (2006).
42. Mannon, P.J. *et al.* Anti-interleukin-12 antibody for active Crohn's disease. *N. Engl. J. Med.* **351**, 2069–2079 (2004).
43. Cohen, J.C. *et al.* Multiple rare alleles contribute to low plasma levels of HDL cholesterol. *Science* **305**, 869–872 (2004).
44. The International HapMap Consortium. A haplotype map of the human genome. *Nature* **437**, 1299–1320 (2005).
45. Wigginton, J.E., Cutler, D.J. & Abecasis, G.R. A note on exact tests of Hardy-Weinberg equilibrium. *Am. J. Hum. Genet.* **76**, 887–893 (2005).
46. Armitage, P. Test for linear trend in proportions and frequencies. *Biometrics* **11**, 375–386 (1955).
47. Purcell, S. *et al.* PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *Am. J. Hum. Genet.* **81**, 559–575 (2007).
48. Purcell, S., Daly, M.J. & Sham, P.C. WHAP: haplotype-based association analysis. *Bioinformatics* **23**, 255–256 (2007).
49. Marchini, J., Cardon, L.R., Phillips, M.S. & Donnelly, P. The effects of human population structure on large genetic association studies. *Nat. Genet.* **36**, 512–517 (2004).

The complete list of authors is as follows:

Wellcome Trust Case Control Consortium

Management Committee: Paul R Burton[1], David G Clayton[2], Lon R Cardon[3,55], Nick Craddock[4], Panos Deloukas[5], Audrey Duncanson[6], Dominic P Kwiatkowski[3,5], Mark I McCarthy[3,7], Willem H Ouwehand[8,9], Nilesh J Samani[10], John A Todd[2], Peter Donnelly (Chair)[11]
Analysis Committee: Jeffrey C Barrett[3], Paul R Burton[1], Dan Davison[11], Peter Donnelly[11], Doug Easton[12], David M Evans[3], Hin-Tak Leung[2], Jonathan L Marchini[11], Andrew P Morris[3], Chris CA Spencer[11], Martin D Tobin[1], Lon R Cardon[3,55], David G Clayton[2]
UK Blood Services & University of Cambridge Controls: Antony P Attwood[5,8], James P Boorman[8,9], Barbara Cant[8], Ursula Everson[13], Judith M Hussey[14], Jennifer D Jolley[8], Alexandra S Knight[8], Kerstin Koch[8], Elizabeth Meech[15], Sarah Nutland[2], Christopher V Prowse[16], Helen E Stevens[2], Niall C Taylor[8], Graham R Walters[17], Neil M Walker[2], Nicholas A Watkins[8,9], Thilo Winzer[8], John A Todd[2], Willem H Ouwehand[8,9]
1958 Birth Cohort Controls: Richard W Jones[18], Wendy L McArdle[18], Susan M Ring[18], David P Strachan[19], Marcus Pembrey[18,20]
Bipolar Disorder (Aberdeen): Gerome Breen[21], David St Clair[21]; (Birmingham): Sian Caesar[22], Katharine Gordon-Smith[22,23], Lisa Jones[22]; (Cardiff): Christine Fraser[23], Elaine K Green[23], Detelina Grozeva[23], Marian L Hamshere[23], Peter A Holmans[23], Ian R Jones[23], George Kirov[23], Valentina Moskivina[23], Ivan Nikolov[23], Michael C O'Donovan[23], Michael J Owen[23], Nick Craddock[23]; (London): David A Collier[24], Amanda Elkin[24], Anne Farmer[24], Richard Williamson[24], Peter McGuffin[24]; (Newcastle): Allan H Young[25], I Nicol Ferrier[25]
Coronary Artery Disease (Leeds): Stephen G Ball[26], Anthony J Balmforth[26], Jennifer H Barrett[26], Timothy D Bishop[26], Mark M Iles[26], Azhar Maqbool[26], Nadira Yuldasheva[26], Alistair S Hall[26]; (Leicester): Peter S Braund[10], Paul R Burton[1], Richard J Dixon[10], Massimo Mangino[10], Suzanne Stevens[10], Martin D Tobin[1], John R Thompson[1], Nilesh J Samani[10]
Crohn's Disease (Cambridge): Francesca Bredin[27], Mark Tremelling[27], Miles Parkes[27]; (Edinburgh): Hazel Drummond[28], Charles W Lees[28], Elaine R Nimmo[28], Jack Satsangi[28]; (London): Sheila A Fisher[29], Alastair Forbes[30], Cathryn M Lewis[29], Clive M Onnie[29], Natalie J Prescott[29], Jeremy Sanderson[31], Christopher G Matthew[29]; (Newcastle): Jamie Barbour[32], M Khalid Mohiuddin[32], Catherine E Todhunter[32], John C Mansfield[32]; (Oxford): Tariq Ahmad[33], Fraser R Cummings[33], Derek P Jewell[33]
Hypertension (Aberdeen): John Webster[34]; (Cambridge): Morris J Brown[35], David G Clayton[2]; (Evry, France): Mark G Lathrop[36]; (Glasgow): John Connell[37], Anna Dominiczak[37]; (Leicester): Nilesh J Samani[10]; (London): Carolina A Braga Marcano[38], Beverley Burke[38], Richard Dobson[38], Johannie Gungadoo[38], Kate L Lee[38], Patricia B Munroe[38], Stephen J Newhouse[38], Abiodun Onipinla[38], Chris Wallace[38], Mingzhan Xue[38], Mark Caulfield[38]; (Oxford): Martin Farrall[39]
Rheumatoid Arthritis: Anne Barton[40], The Biologics in RA Genetics and Genomics Study Syndicate (BRAGGS) Steering Committee*, Ian N Bruce[40], Hannah Donovan[40], Steve Eyre[40], Paul D Gilbert[40], Samantha L Hilder[40], Anne M Hinks[40], Sally L John[40], Catherine Potter[40], Alan J Silman[40], Deborah PM Symmons[40], Wendy Thomson[40], Jane Worthington[40]
Type 1 Diabetes: David G Clayton[2], David B Dunger[2,41], Sarah Nutland[2], Helen E Stevens[2], Neil M Walker[2], Barry Widmer[2,41], John A Todd[2]
Type 2 Diabetes (Exeter): Timothy M Frayling[42,43], Rachel M Freathy[42,43], Hana Lango[42,43], John R B Perry[42,43], Beverley M Shields[43], Michael N Weedon[42,43], Andrew T Hattersley[42,43]; (London): Graham A Hitman[44]; (Newcastle): Mark Walker[45]; (Oxford): Kate S Elliott[3,7], Christopher J Groves[7], Cecilia M Lindgren[3,7], Nigel W Rayner[3,7], Nicolas J Timpson[3,46], Eleftheria Zeggini[3,7], Mark I McCarthy[3,7]

Tuberculosis (Gambia): Melanie Newport[47], Giorgio Sirugo[47], (Oxford): Emily Lyons[3], Fredrik Vannberg[3], Adrian V S Hill[3]
Ankylosing Spondylitis: Linda A Bradbury[48], Claire Farrar[49], Jennifer J Pointon[49], Paul Wordsworth[49], Matthew A Brown[48,49]
Autoimmune Thyroid Disease: Jayne A Franklyn[50], Joanne M Heward[50], Matthew J Simmonds[50], Stephen CL Gough[50]
Breast Cancer: Sheila Seal[51], Breast Cancer Susceptibility Collaboration (UK)*, Michael R Stratton[51,52], Nazneen Rahman[51]
Multiple Sclerosis: Maria Ban[53], An Goris[53], Stephen J Sawcer[53], Alastair Compston[53]
Gambian Controls (Gambia): David Conway[47], Muminatou Jallow[47], Melanie Newport[47], Giorgio Sirugo[47]; (Oxford): Kirk A Rockett[3], Dominic P Kwiatkowski[3,5]
DNA, Genotyping, Data QC and Informatics (Wellcome Trust Sanger Institute, Hinxton): Suzannah J Bumpstead[5], Amy Chaney[5], Kate Downes[2,5], Mohammed JR Ghori[5], Rhian Gwilliam[5], Sarah E Hunt[5], Michael Inouye[5], Andrew Keniry[5], Emma King[5], Ralph McGinnis[5], Simon Potter[5], Rathi Ravindrarajah[5], Pamela Whittaker[5], Claire Widden[5], David Withers[5], Panos Deloukas[5]; (Cambridge): Hin-Tak Leung[2], Sarah Nutland[2], Helen E Stevens[2], Neil M Walker[2], John A Todd[2]
Statistics (Cambridge): Doug Easton[12], David G Clayton[2]; (Leicester): Paul R Burton[1], Martin D Tobin[1]; (Oxford): Jeffrey C Barrett[3], David M Evans[3], Andrew P Morris[3], Lon R Cardon[3,55]; (Oxford): Niall J Cardin[11], Dan Davison[11], Teresa Ferreira[11], Joanne Pereira-Gale[11], Ingeleif B Hallgrimsdóttir[11], Bryan N Howie[11], Jonathan L Marchini[11], Chris CA Spencer[11], Zhan Su[11], Yik Ying Teo[3,11], Damjan Vukcevic[11], Peter Donnelly[11]
Principal Investigators: David Bentley[5,54], Matthew A Brown[48,49], Lon R Cardon[3,55], Mark Caulfield[38], David G Clayton[2], Alastair Compston[53], Nick Craddock[23], Panos Deloukas[5], Peter Donnelly[11], Martin Farrall[39], Stephen CL Gough[50], Alistair S Hall[26], Andrew T Hattersley[42,43], Adrian V S Hill[3], Dominic P Kwiatkowski[3,5], Christopher G Matthew[29], Mark I McCarthy[3,7], Willem H Ouwehand[8,9], Miles Parkes[27], Marcus Pembrey[18,20], Nazneen Rahman[51], Nilesh J Samani[10], Michael R Stratton[51,52], John A Todd[2], Jane Worthington[40]
AITD Replication Group: Sarah L Mitchell[50], Paul R Newby[50], Oliver J Brand[50], Jackie Carr-Smith[50], Simon H S Pearce[56], Stephen C L Gough[50]
IL23R replication: R McGinnis[5], A Keniry[5], P Deloukas[5], TASC.
The Australo-Anglo-American Spondylitis Consortium (TASC): John D Reveille[57], Xiaodong Zhou[57], Linda A Bradbury[58], Anne-Marie Sims[58], Alison Dowling[58], Jacqueline Taylor[58], Tracy Doan[58], Lon R Cardon[55,59], John C Davis[60], Jennifer J Pointon[61], Laurie Savage[62], Michael M Ward[63], Thomas L Learch[64], Michael H Weisman[65], Paul Wordsworth[61], Matthew A Brown[58,61]

***See Supplementary Note for details.**

Affiliations for participants: [1]Genetic Epidemiology Group, Department of Health Sciences, University of Leicester, Adrian Building, University Road, Leicester LE1 7RH, UK. [2]Juvenile Diabetes Research Foundation/Wellcome Trust Diabetes and Inflammation Laboratory, Department of Medical Genetics, Cambridge Institute for Medical Research, University of Cambridge, Wellcome Trust/MRC Building, Cambridge CB2 0XY, UK. [3]Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive, Oxford OX3 7BN, UK. [4]Department of Psychological Medicine, Henry Wellcome Building, School of Medicine, Cardiff University, Heath Park, Cardiff CF14 4XN, UK. [5]The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [6]The Wellcome Trust, Gibbs Building, 215 Euston Road, London NW1 2BE, UK. [7]Oxford Centre for Diabetes, Endocrinology and Medicine, University of Oxford, Churchill Hospital, Oxford OX3 7LJ, UK. [8]Department of Haematology, University of Cambridge, Long Road, Cambridge CB2 2PT, UK. [9]National Health Service Blood and Transplant, Cambridge Centre, Long Road, Cambridge CB2 2PT, UK. [10]Department of Cardiovascular Sciences, University of Leicester, Glenfield Hospital, Groby Road, Leicester LE3 9QP, UK. [11]Department of Statistics, University of Oxford, 1 South Parks Road, Oxford OX1 3TG, UK. [12]Cancer Research UK Genetic Epidemiology Unit, Strangeways Research Laboratory, Worts Causeway, Cambridge CB1 8RN, UK. [13]National Health Service Blood and Transplant, Sheffield Centre, Longley Lane, Sheffield S5 7JN, UK. [14]National Health Service Blood and Transplant, Brentwood Centre, Crescent Drive, Brentwood CM15 8DP, UK. [15]The Welsh Blood Service, Ely Valley Road, Talbot Green, Pontyclun CF72 9WB, UK. [16]The Scottish National Blood Transfusion Service, Ellen's Glen Road, Edinburgh EH17 7QT, UK. [17]National Health Service Blood and Transplant, Southampton Centre, Coxford Road, Southampton SO16 5AF, UK. [18]Avon Longitudinal Study of Parents and Children, University of Bristol, 24 Tyndall Avenue, Bristol BS8 1TQ, UK. [19]Division of Community Health Services, St. George's University of London, Crammer Terrace, London SW17 0RE, UK. [20]Institute of Child Health, University College London, 30 Guilford St., London WC1N 1EH, UK. [21]University of Aberdeen, Institute of Medical Sciences, Foresterhill, Aberdeen AB25 2ZD, UK. [22]Department of Psychiatry, Division of Neuroscience, Birmingham University, Birmingham B15 2QZ, UK. [23]Department of Psychological Medicine, Henry Wellcome Building, School of Medicine, Cardiff Univeristy, Heath Park, Cardiff CF14 4XN, UK. [24]SGDP, The Institute of Psychiatry, King's College London, De Crespigny Park, Denmark Hill, London SE5 8AF, UK. [25]School of Neurology, Neurobiology and Psychiatry, Royal Victoria Infirmary, Queen Victoria Road, Newcastle upon Tyne NE1 4LP, UK. [26]LIGHT and LIMM Research Institutes, Faculty of Medicine and Health, University of Leeds, Leeds LS1 3EX, UK. [27]IBD Research Group, Addenbrooke's Hospital, University of Cambridge, Cambridge CB2 2QQ, UK. [28]Gastrointestinal Unit, School of Molecular and Clinical Medicine, University of Edinburgh, Western General Hospital, Edinburgh EH4 2XU UK. [29]Department of Medical & Molecular Genetics, King's College London School of Medicine, 8th Floor Guy's Hospital, London SE1 9RT, UK. [30]Institute for Digestive Diseases, University College London Hospitals Trust, London NW1 2BU, UK. [31]Department of Gastroenterology, Guy's and St. Thomas' NHS Foundation Trust, London SE1 7EH, UK. [32]Department of Gastroenterology & Hepatology, University of Newcastle upon Tyne, Royal Victoria Infirmary, Newcastle upon Tyne NE1 4LP, UK. [33]Gastroenterology Unit, Radcliffe Infirmary, University of Oxford, Oxford OX2 6HE, UK. [34]Medicine and Therapeutics, Aberdeen Royal Infirmary, Foresterhill, Aberdeen, Grampian AB9 2ZB, UK. [35]Clinical Pharmacology Unit and the Diabetes and Inflammation Laboratory, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK. [36]Centre National de Genotypage, 2 Rue Gaston Cremieux, Evry, Paris 91057, France. [37]BHF Glasgow Cardiovascular Research Centre, University of Glasgow, 126 University Place, Glasgow G12 8TA, UK. [38]Clinical Pharmacology and Barts and The London Genome Centre, William Harvey Research Institute, Barts and The London, Queen Mary's School of Medicine, Charterhouse Square, London EC1M 6BQ, UK. [39]Cardiovascular Medicine, University of Oxford, Wellcome Trust Centre for Human Genetics, Roosevelt Drive, Oxford OX3 7BN, UK. [40]arc Epidemiology Research Unit, University of Manchester, Stopford Building, Oxford Rd, Manchester M13 9PT, UK. [41]Department of Paediatrics, University of Cambridge, Addenbrooke's Hospital, Cambridge, CB2 2QQ, UK. [42]Genetics of Complex Traits, Institute of Biomedical and Clinical Science, Peninsula Medical School, Magdalen Road, Exeter EX1 2LU, UK. [43]Diabetes Genetics, Institute of Biomedical and Clinical Science, Peninsula Medical School, Barrack Road, Exeter EX2 5DU UK. [44]Centre for Diabetes and Metabolic Medicine, Barts and The London, Royal London Hospital, Whitechapel, London E1 1BB, UK. [45]Diabetes Research Group, School of Clinical Medical Sciences, Newcastle University, Framlington Place, Newcastle upon Tyne NE2 4HH, UK. [46]The MRC Centre for Causal Analyses in Translational Epidemiology, Bristol University, Canynge Hall, Whiteladies Rd., Bristol BS2 8PR, UK. [47]MRC Laboratories, Fajara, The Gambia. [48]Diamantina Institute for Cancer, Immunology and Metabolic Medicine, Princess Alexandra Hospital, University of Queensland, Woolloongabba, Queensland 4102, Australia. [49]Botnar Research Centre, University of Oxford, Headington, Oxford OX3 7BN, UK. [50]Department of Medicine, Division of Medical Sciences, Institute of Biomedical Research, University of Birmingham, Edgbaston, Birmingham B15 2TT, UK. [51]Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton SM2 5NG, UK. [52]Cancer Genome Project, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge CB10 1SA, UK. [53]Department of Clinical Neurosciences, University of Cambridge, Addenbrooke's Hospital, Hills Road, Cambridge CB2 2QQ, UK. [54]Illumina Cambridge, Chesterford Research Park, Little Chesterford, NR Saffron Walden, Essex CB10 1XL, UK. [55]Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, Washington 98109, USA. [56]University of Newcastle, Institute for Human Genetics, Centre for Life, Central Parkway, Newcastle upon Tyne NE1 3BZ, UK. [57]Rheumatology and Clinical Immunogenetics, University of Texas-Houston Medical School, Houston, Texas 77030, USA. [58]Diamantina Institute for Cancer, Immunology and Metabolic Medicine, University of Queensland, Brisbane 4072, Australia. [59]Statistical Genetics, Wellcome Trust Centre for Human Genetics, Oxford OX37BN, UK. [60]Department of Rheumatology, University of California, San Francisco, California 94143, USA. [61]Botnar Research Centre, University of Oxford, Oxford OX37BN, UK. [62]The Spondylitis Association of America, Sherman Oaks, California 91403, USA. [63]National Institute of Arthritis and Musculoskeletal and Skin Diseases, National Institutes of Health, Bethesda, Maryland 20892, USA. [64]Department of Radiology and [65]Department of Medicine/Rheumatology, Cedars-Sinai Medical Center, Los Angeles, California 90048, USA.

nature genetics

A common coding variant in *CASP8* is associated with breast cancer risk

Angela Cox[1,30], Alison M Dunning[2,30], Montserrat Garcia-Closas[3,30], Sabapathy Balasubramanian[1], Malcolm W R Reed[1], Karen A Pooley[2], Serena Scollen[2], Caroline Baynes[2], Bruce A J Ponder[2], Stephen Chanock[3], Jolanta Lissowska[4], Louise Brinton[3], Beata Peplonska[5], Melissa C Southey[6], John L Hopper[6], Margaret R E McCredie[7], Graham G Giles[8], Olivia Fletcher[9], Nichola Johnson[9], Isabel dos Santos Silva[9], Lorna Gibson[9], Stig E Bojesen[10], Børge G Nordestgaard[10], Christen K Axelsson[10], Diana Torres[11], Ute Hamann[11], Christina Justenhoven[12], Hiltrud Brauch[12], Jenny Chang-Claude[13], Silke Kropp[13], Angela Risch[13], Shan Wang-Gohrke[14], Peter Schürmann[15], Natalia Bogdanova[16], Thilo Dörk[15], Rainer Fagerholm[17], Kirsimari Aaltonen[17,18], Carl Blomqvist[18], Heli Nevanlinna[17], Sheila Seal[19], Anthony Renwick[19], Michael R Stratton[19], Nazneen Rahman[19], Suleeporn Sangrajrang[20], David Hughes[21], Fabrice Odefrey[21], Paul Brennan[21], Amanda B Spurdle[22], Georgia Chenevix-Trench[22], The Kathleen Cunningham Foundation Consortium for Research into Familial Breast Cancer, Jonathan Beesley[22], Arto Mannermaa[23], Jaana Hartikainen[23], Vesa Kataja[23], Veli-Matti Kosma[23], Fergus J Couch[24], Janet E Olson[24], Ellen L Goode[24], Annegien Broeks[25], Marjanka K Schmidt[25], Frans B L Hogervorst[25], Laura J Van't Veer[25], Daehee Kang[26], Keun-Young Yoo[26,27], Dong-Young Noh[26], Sei-Hyun Ahn[28], Sara Wedrén[29], Per Hall[29], Yen-Ling Low[30], Jianjun Liu[30], Roger L Milne[31], Gloria Ribas[31], Anna Gonzalez-Neira[31], Javier Benitez[31], Alice J Sigurdson[32], Denise L Stredrick[32], Bruce H Alexander[32], Jeffery P Struewing[32], Paul D P Pharoah[2] & Douglas F Easton[2], on behalf of the Breast Cancer Association Consortium

The Breast Cancer Association Consortium (BCAC) has been established to conduct combined case-control analyses with augmented statistical power to try to confirm putative genetic associations with breast cancer. We genotyped nine SNPs for which there was some prior evidence of an association with breast cancer: *CASP8* D302H (rs1045485), *IGFBP3* –202 C→A (rs2854744), *SOD2* V16A (rs1799725), *TGFB1* L10P (rs1982073), *ATM* S49C (rs1800054), *ADH1B* 3′ UTR A→G (rs1042026), *CDKN1A* S31R (rs1801270), *ICAM5* V301I (rs1056538) and *NUMA1* A794G (rs3750913). We included data from 9–15 studies, comprising 11,391–18,290 cases and 14,753–22,670 controls. We found evidence of an association with breast cancer for *CASP8* D302H (with odds ratios (OR) of 0.89 (95% confidence interval (c.i.): 0.85–0.94) and 0.74 (95% c.i.: 0.62–0.87) for heterozygotes and rare homozygotes, respectively, compared with common homozygotes; $P_{trend} = 1.1 \times 10^{-7}$) and weaker evidence for *TGFB1* L10P (OR = 1.07 (95% c.i.: 1.02–1.13) and 1.16 (95% c.i.: 1.08–1.25), respectively; $P_{trend} = 2.8 \times 10^{-5}$). These results demonstrate that common breast cancer susceptibility alleles with small effects on risk can be identified, given sufficiently powerful studies.

[1]Sheffield University Medical School, Sheffield S10 2RX, UK. [2]Department of Oncology and Department of Public Health and Primary Care, University of Cambridge, Cambridge CB2 1TN, UK. [3]Division of Cancer Epidemiology and Genetics, National Cancer Institute, Rockville, Maryland 20852, USA, and Core Genotyping Facility, Advanced Technology Center, National Cancer Institute, Gaithersburg, Maryland 20892, USA. [4]Cancer Center and M. Sklodowska-Curie Institute of Oncology, 02-781 Warsaw, Poland. [5]Nofer Institute of Occupational Medicine, 90-950 Lodz, Poland. [6]University of Melbourne, Melbourne, Victoria 3010 Australia. [7]University of Otago, Dunedin, New Zealand. [8]Cancer Epidemiology Centre, The Cancer Council Victoria, Melbourne, Victoria 3053, Australia. [9]The Breakthrough Breast Cancer Research Centre, The Institute of Cancer Research, London SW3 6JB, UK, and London School of Hygiene and Tropical Medicine, London WC1E 7HT, UK. [10]Department of Clinical Biochemistry, and Department of Breast Surgery, Herlev University Hospital, University of Copenhagen, 2730 Herlev, Denmark. [11]Deutsches Krebsforschungszentrum, 69120 Heidelberg, Germany. [12]Dr. Margarete Fischer Bosch Institute of Clinical Pharmacology, D-70376 Stuttgart, Germany, and University of Tübingen, 72074 Tübingen, Germany. [13]German Cancer Research Center, 69120 Heidelberg, Germany. [14]University of Ulm, 89069 Ulm, Germany. [15]Department of Gynecology and Obstetrics and [16]Department of Radiation Oncology, Hannover Medical School, 30625 Hannover, Germany. [17]Departments of Obstetrics and Gynecology and [18]Department of Oncology, Helsinki University Central Hospital, FIN-00290 Helsinki, Finland. [19]Section of Cancer Genetics, Institute of Cancer Research, Sutton, Surrey SM2 5NG, UK. [20]National Cancer Institute, 10400 Bangkok, Thailand. [21]International Agency for Research on Cancer, 69372 Lyon, France. [22]Queensland Institute of Medical Research, Brisbane, Queensland 4029, Australia. [23]Institute of Clinical Medicine, Pathology and Forensic Medicine, University of Kuopio, FI-70211 Kuopio, Finland, and Departments of Oncology and Pathology, University Hospital of Kuopio, FI-70211 Kuopio, Finland. [24]Mayo Clinic College of Medicine, Rochester, Minnesota 55905, USA. [25]Netherlands Cancer Institute, Departments of Experimental Therapy, Epidemiology and Molecular Pathology, 1066 CX Amsterdam, The Netherlands. [26]Seoul National University College of Medicine, Seoul 151-742, Korea. [27]National Cancer Center, Goyang 411769, Korea. [28]Department of Surgery, Ulsan University College of Medicine, Ulsan 680-749, Korea. [29]Department of Medical Epidemiology and Biostatistics, Karolinska Institute, SE-171 77 Stockholm, Sweden. [30]Population Genetics, Genome Institute of Singapore, Singapore 138672. [31]Spanish National Cancer Research Centre, (CNIO), E-28029 Madrid, Spain. [32]National Cancer Institute, US National Institutes of Health (NIH), Department of Health and Human Services, Bethesda, Maryland 20892, USA, and University of Minnesota, Division of Environmental Health Sciences, Minneapolis, Minnesota 55455, USA. [33]These authors contributed equally to this work. Correspondence should be addressed to A.C. (a.cox@shef.ac.uk).

Received 18 August 2006; accepted 17 January 2007; published online 11 February 2007; doi:10.1038/ng1981

Rare, high-penetrance germline mutations in genes such as *BRCA1* or *BRCA2* account for less than 25% of the familial risk of breast cancer, and much of the remaining variation in genetic risk is likely to be explained by combinations of more common, lower-penetrance variants[1]. To date, case-control studies have generally focused on the investigation of putative functional candidate gene variants to attempt to identify low-penetrance susceptibility variants. However, individual studies often have only enough statistical power to detect effects of the order of 1.5 or more, depending on the frequency of the variant[2], and thus collaborative studies are needed in order to achieve the sample sizes necessary to detect more modest effects. The Breast Cancer Association Consortium (BCAC) was established in 2005 to facilitate such collaborative studies in breast cancer. The consortium currently comprises over 20 international collaborating research groups, with a potential combined sample size of up to 30,000 cases and 30,000 controls. The first combined data analysis carried out by the consortium involved 16 SNPs that had been investigated in at least three independent studies with at least 10,000 genotyped subjects in total[3]. Members of the consortium then carried out further genotyping for four of these SNPs that showed borderline evidence of associations with risk: caspase-8 (*CASP8*) D302H (rs1045485), insulin-like growth factor binding protein 3 (*IGFBP3*) –202 C→A (rs2854744), manganese superoxide dismutase (*SOD2* or *MnSOD*) V16A (rs1799725) and transforming growth factor beta (*TGFB1*) L10P (rs1982073), in order to confirm or refute these results. In addition, the BCAC examined five other SNPs for which there was published or unpublished evidence of an association: ataxia telangiectasia mutated (*ATM*) S49C (rs1800054)[4,5], class I alcohol dehydrogenase 1B (*ADH1B*, formerly called *ADH2*) 3′UTR A→G (rs1042026) (P.D.P.P. *et al.*, unpublished data), cyclin-dependent kinase inhibitor 1A (*CDKN1A*) S31R (rs1801270) (P.D.P.P. *et al.* and A.C. *et al.*, unpublished data), intercellular adhesion molecule 5 (*ICAM5*) V301I (rs1056538)[6] and nuclear mitotic apparatus protein (*NUMA1*) A794G (rs3750913)[7].

Details of the 20 studies contributing data to this report are shown in **Supplementary Table 1** online. Apart from two studies in Asian populations, cases and controls were selected from populations of predominantly European ancestry, all with high breast cancer incidence rates (age-standardized rates ranging from 42.6 per 100,000 to 99.4 per 100,000 (ref. 8)).

Two of the nine SNPs evaluated showed significant associations with invasive breast cancer: *CASP8* D302H and *TGFB1* L10P. Caspase-8 is an important initiator of apoptosis (programmed cell death) and is activated by external death signals and in response to DNA damage[9]. Two previous studies suggested that the D302H polymorphism in *CASP8* (rs1045485), which results in an aspartic acid to histidine substitution, could reduce breast cancer risk[10,11].

Our analysis of 16,423 cases and 17,109 controls from 14 studies showed convincing evidence for a protective effect in an allele dose–dependent manner ($P_{trend} = 1.1 \times 10^{-7}$, per allele odds ratio (OR) = 0.88 (with 95% confidence interval (c.i.) of 0.84–0.92); **Table 1** and **Fig. 1a**). The result remained significant after excluding the initial positive result from the Sheffield Breast Cancer Study[10] ($P_{trend} = 1 \times 10^{-6}$), and there was no evidence of between-study heterogeneity ($P = 0.97$). We found no evidence that the ORs varied

Table 1 Summary odds ratios and 95% confidence intervals for nine polymorphisms and breast cancer risk

SNP	No. of studies	No. of controls	No. of cases	MAF	Between-study heterogeneity[a]	Test for association[a]	Trend test[a]	Analysis model	Per-allele OR (95% c.i.)[b]	Heterozygote OR (95% c.i.)[b]	Rare homozygote OR (95% c.i.)[b]
ADH1B 3′ UTR A→G rs1042026	9	15,570	11,391	0.29	0.35	0.044	0.54	Fixed effects	0.99 (0.95, 1.03)	0.94 (0.89, 1.00)	1.04 (0.95, 1.14)
								Random effects	0.99 (0.95, 1.04)	0.99 (0.90, 1.10)	1.04 (0.95, 1.14)
CASP8 D302H rs1045485	14	17,109	16,423	0.13	0.97	5.7×10^{-7}	1.1×10^{-7}	Fixed effects	0.88 (0.84, 0.92)	0.89 (0.85, 0.94)	0.74 (0.62, 0.87)
								Random effects	0.88 (0.84, 0.92)	0.89 (0.85, 0.94)	0.73 (0.60, 0.90)
CDKN1A S31R rs1801270	15	22,670	18,290	0.072	0.009	0.55	0.28	Fixed effects	1.03 (0.98, 1.09)[c]	1.03 (0.97, 1.10)	1.07 (0.86, 1.33)[c]
								Random effects	1.02 (0.93, 1.11)[c]	1.04 (0.93, 1.09)	1.20 (0.82, 1.76)[c]
ICAM5 V301I rs1056538	15	22,229	17,687	0.39	0.58	0.58	0.78	Fixed effects	1.00 (0.97, 1.03)	1.02 (0.98, 1.07)	1.00 (0.94, 1.06)
								Random effects	1.00 (0.97, 1.03)	1.02 (0.97, 1.08)	0.99 (0.93, 1.06)
IGFBP3 –202C→A rs2854744	10	17,926	13,101	0.45	0.72	0.051	0.046	Fixed effects	0.97 (0.94, 1.00)	1.00 (0.94, 1.05)	0.93 (0.87, 0.99)
								Random effects	0.97 (0.93, 1.00)	1.00 (0.94, 1.05)	0.92 (0.86, 0.99)
SOD2 V16A rs1799725	13	21,349	16, 273	0.50	0.016	0.13	0.31	Fixed effects	0.98 (0.96, 1.01)	1.02 (0.97, 1.08)	0.97 (0.91, 1.03)
								Random effects	0.98 (0.94, 1.03)	1.02 (0.97, 1.08)	0.96 (0.88, 1.06)
TGFB1 L10P rs1982073	11	15,109	12,946	0.38	0.68	1.5×10^{-4}	2.8×10^{-5}	Fixed effects	1.08 (1.04, 1.11)	1.07 (1.02, 1.13)	1.16 (1.08, 1.25)
								Random effects	1.08 (1.04, 1.11)	1.07 (1.02, 1.13)	1.16 (1.08, 1.25)
ATM S49C rs1800054	12	19,488	15,905	0.012	0.27	0.08[d]		Fixed effects		1.13[d] (0.99, 1.30)	
								Random effects		1.13[d] (0.96, 1.32)	
NUMA1 A794G rs3750913	13	18,320	14,642	0.028	0.029	0.52[d]		Fixed effects		1.03[d] (0.94, 1.14)	
								Random effects		1.03[d] (0.90, 1.19)	

MAF: Minor allele frequency in the control sample.
[a]*P* values. The test of association and trend test are 2 d.f. and 1 d.f. LRT, respectively. [b]Reference group: common homozygotes. [c]Analyses excluded three studies (Helsinki Breast Cancer Study, Mayo Clinic Breast Cancer Study and USRT) because no homozygous variants were observed among cases or controls. [d]Heterozygote and homozygote variant genotypes were combined because of small number of women with the homozygote variant genotype.

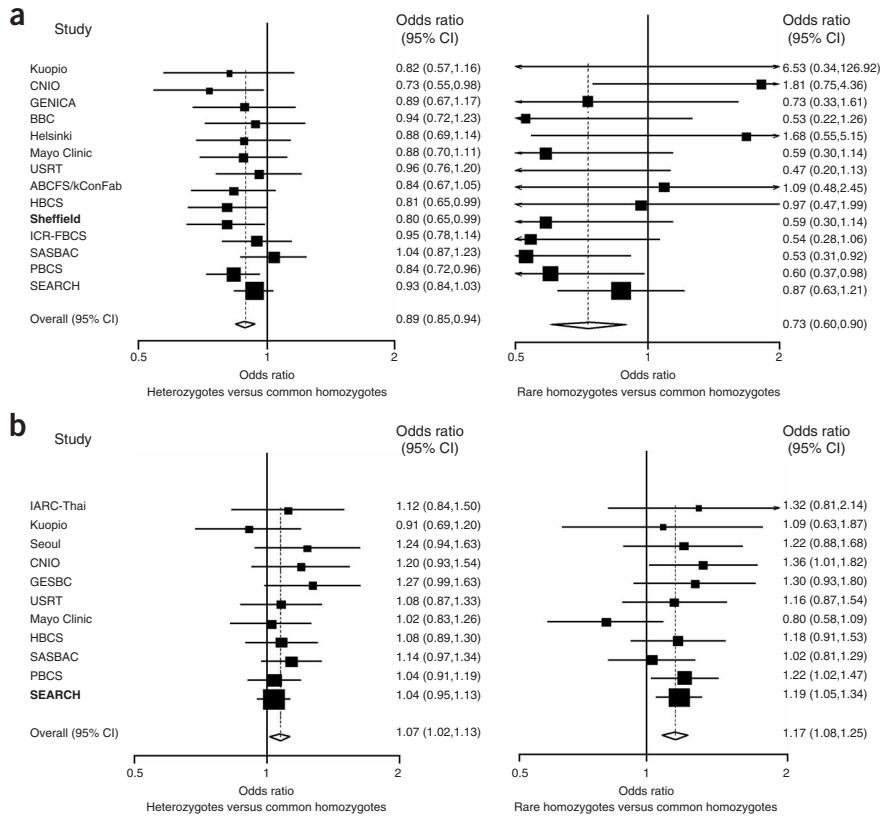

Figure 1 Genotype-specific OR and 95% c.i. by study. (**a**) *CASP8* D302H (rs1045485). (**b**) *TGFB1* L10P (rs1982073). Common homozygotes are the reference group. The initial study is indicated in bold. Studies are weighted and ranked according to the inverse of the variance of the log OR estimate for the heterozygotes.

with age, estrogen receptor or progesterone receptor status, grade, stage or histopathological subtype (**Table 2**). The ORs for ductal carcinoma *in situ* (DCIS) tumors were similar to that for invasive breast cancer. We saw no evidence of a stronger association in women with a history of breast cancer in first-degree female relatives, such as has been observed for other susceptibility alleles in *ATM* and *CHEK2* (refs. 12,13) (per-allele OR for *CASP8* D302H = 0.87 (95% c.i.: 0.82–0.91), 0.98 (95% c.i.: 0.89–1.07) and 0.90 (95% c.i.: 0.79–1.01) for zero, one and two or more first-degree relatives, respectively). An association with family history would be expected under a polygenic model with multiplicative effects at different loci, and this result may therefore suggest a different pattern of interaction with other susceptibility alleles. Of note, this site was not polymorphic in Korean, Han Chinese or Japanese women (D.K. *et al.*, unpublished data, http://www.hapmap.org/). The functional consequences of the aspartic acid-to-histidine substitution are not yet known, and further experiments are required to establish whether D302H itself, or another variant in strong linkage disequilibrium with it, is causative. Although this SNP was identified through a candidate gene approach, the association achieved a significance level close to that required for genome-wide studies[14].

Transforming growth factor-β (TGF-β) is a polypeptide cytokine that, *inter alia*, regulates normal mammary gland development and function by activating the TGF-β signaling pathway (reviewed in ref. 15). There is a dual-role model for the action of TGF-β in which it is thought to inhibit the development of early benign tumors, but once somatic oncogenic mutations have destroyed the normal tumor suppressor action of TGF-β, it then promotes tumor invasion and metastasis[15,16]. Our analysis of the L10P variant (rs1982073) in the *TGFB1* signal peptide showed a significant dose-dependent association of the proline-encoding allele with increased risk of invasive breast cancer based on analyses of data from 11 studies comprising 12,946 cases and 15,109 controls ($P_{trend} = 2.8 \times 10^{-5}$, per-allele OR = 1.08, (95% c.i.: 1.04–1.11); **Table 1** and **Fig. 1b**). This result remained significant after exclusion of the initial result from the Studies of Epidemiology and Risk Factors in Cancer Heredity (SEARCH)[17] ($P_{trend} = 8.0 \times 10^{-4}$), with no evidence of between-study heterogeneity ($P = 0.68$).

The proline variant of *TGFB1* has been associated with higher circulating levels of acid-activatable TGF-β[18] and increased rates of TGF-β secretion in *in vitro* transfection experiments[17]. From the dual-role model, it has been suggested that the proline (rapid TGF-β secretion) variant should be associated with a reduced risk of *in situ* tumors but an increased risk of invasive cancer. This study had insufficient cases with ductal carcinoma *in situ* (DCIS) to detect a significant differential risk ($n = 328$), but the estimated ORs for DCIS were consistent with a protective effect (**Table 3**). As might be predicted by a polygenic model, the ORs were greatest in those under 40 and closer to unity in older age groups, although this trend was not significant at the $P = 0.05$ level (**Table 3**). The ORs did not vary substantially by stage, grade or estrogen receptor status of the tumor. However, the significant association of the proline variant was confined to individuals with progesterone receptor–negative (rather than progesterone receptor–positive) tumors ($P = 0.017$; **Table 3**).

The findings of previously published studies, which have not subsequently been subsumed into the BCAC, have been contradictory or null[19–24]. A meta-analysis of the BCAC data together with the published studies (the latter totaling 4,021 cases and 8,253 controls) showed much weaker evidence for an increase in risk of the rare allele (per-allele OR = 1.04 (95% c.i.: 1.01–1.07), $P_{trend} = 0.012$). Differences in case selection or characteristics between studies could contribute to the discrepancy with the published results. The BCAC data may be more reliable, as it should be less susceptible to any publication bias. However, despite the size of our study and the relatively high level of significance, we cannot rule out the possibility that the *TGFB1* L10P association we found is a false positive result.

We observed borderline evidence of associations for two additional SNPs. The data suggest a recessive association for a promoter SNP in *IGFBP3* (–202C→A, rs2854744), (OR = 0.93 (95% c.i.: 0.87–0.99), $P_{trend} = 0.046$, **Table 1**). Two of the three previously published studies are included in the current analysis[25,26]; one previous null report is not included[27]. IGFBP3 is the principal binding protein regulating the activity of insulin-like growth factor 1 (IGF1), a circulating peptide hormone and growth factor for breast and other tissues. The A allele of the 202C→A SNP has been repeatedly shown to be associated with

Table 2 Subgroup analysis for *CASP8* D302H and breast cancer risk

	Category	No. of cases	Test for association[b]	Heterozygotes[c] OR	Heterozygotes[c] 95% c.i.	Rare homozygotes[c] OR	Rare homozygotes[c] 95% c.i.	Heterogeneity test[d]
Age group, years[a]	<40	1,737	0.038	0.75	(0.60, 0.94)	1.16	(0.56, 2.40)	0.61
	40–49	3,962	0.0024	0.86	(0.76, 0.98)	0.55	(0.36, 0.85)	
	50–59	5,309	0.26	0.93	(0.84, 1.02)	0.91	(0.67, 1.23)	
	≥60	5,065	0.0058	0.89	(0.81, 0.98)	0.70	(0.51, 0.95)	
ER status	+	5,846	0.0042	0.89	(0.82, 0.96)	0.83	(0.65, 1.06)	0.24
	–	1,776	0.46	0.95	(0.84, 1.07)	0.82	(0.55, 1.24)	
PR status	+	3,416	0.024	0.90	(0.81, 0.99)	0.74	(0.53, 1.04)	0.82
	–	1,838	0.087	0.87	(0.76, 0.99)	0.94	(0.64, 1.40)	
Stage	I	3,591	0.31	0.95	(0.87, 1.05)	0.82	(0.59, 1.13)	0.32
	II	2,952	0.063	0.88	(0.79, 0.98)	0.93	(0.67, 1.31)	
	III/IV	288	0.82	0.91	(0.68, 1.23)	0.88	(0.32, 2.40)	
Grade	1	1,924	0.41	0.93	(0.83, 1.05)	0.86	(0.58, 1.28)	0.44
	2	4,229	0.026	0.90	(0.83, 0.98)	0.80	(0.61, 1.07)	
	3	2,731	0.017	0.88	(0.80, 0.98)	0.74	(0.52, 1.04)	
Histopathology	Ductal	7,629	0.0002	0.87	(0.81, 0.93)	0.85	(0.68, 1.07)	0.93
	Lobular	1,504	0.047	0.92	(0.80, 1.05)	0.59	(0.35, 0.98)	
DCIS		456	0.42	0.86	(0.68, 1.09)	0.86	(0.40, 1.84)	

ER, estrogen receptor; PR, progesterone receptor.
[a]Age in years at diagnosis (cases) or interview (controls). [b]LRT, 2 d.f. [c]Reference group: common homozygotes. [d]*P* value for case-only LRT of between-subgroup heterogeneity.

increased circulating IGFBP3 levels[27,28]. However, the role of plasma IGFBP3 levels in breast cancer risk remains uncertain. Our data are consistent with the hypothesis that higher circulating levels of IGFBP3 are protective, but even the current large investigation has insufficient power to detect a recessive association with this allele at more than borderline levels of significance. *ADH1B* 3′ UTR A→G (rs1042026) also yielded a borderline significant association ($P = 0.044$). However, the heterozygote and homozygote genotypic associations were in opposite directions (**Table 1**), they were not consistent across studies and they were not seen under the random effects model (**Table 1**, **Supplementary Tables 2** and **3** online). Given that there is no biological rationale for such an observation, it is highly likely that the heterozygote association is due to chance.

ATM S49C (rs1800054) was not significantly associated with overall breast cancer risk. However, the c.i. did not exclude a modest association, and this SNP increased the risk of progesterone receptor–positive breast cancer (OR = 1.48 (95% c.i.: 1.08–2.04) under a dominant model (**Supplementary Table 4** online). For the remaining four SNPs (*CDKN1A* S31R, *ICAM5* V301I, *SOD2* V16A and *NUMA1* A794G), there was no evidence of an association with breast cancer (**Table 1** and **Supplementary Fig. 1** online). There was some evidence for heterogeneity between studies for *CDKN1A* S31R ($P = 0.009$), *NUMA1* A794G ($P = 0.029$) and *SOD2* V16A ($P = 0.016$), but all ORs and 95% confidence intervals were virtually unchanged using a random effects model to allow for heterogeneity (**Table 1**). When we removed the only study of *CDKN1A* S31R in Asian women (International Agency for Research on Cancer-Thailand Study (IARC-Thai)), summary estimates from the remaining 14 studies in women of predominately European ancestry suggested a recessive association for this SNP (OR = 1.37 (95% c.i.: 1.04–1.81) comparing rare homozygotes with common homozygotes; $P = 0.051$). OR estimates for the other two SNPs were similar in the two studies in Asian countries, and we found no clear explanation for the observed heterogeneity. Confidence intervals for summary ORs, particularly from random effects models, did not exclude modest associations for these SNPs (**Table 1**). We did not observe any additional modification of genotype associations with breast cancer risk by age, estrogen receptor or progesterone receptor tumor status and did not find any significant associations for DCIS tumors (**Supplementary Tables 4–7** online).

We estimate that the *CASP8* D302H and *TGFB1* L10P variants may account for approximately 0.3% and 0.2% of the excess familial risk of breast cancer, respectively, in populations of European ancestry. These data are the strongest evidence to date for common breast cancer susceptibility alleles, and they demonstrate the value of large consortia in identifying these variants.

METHODS

Subjects. Twenty breast cancer case-control studies contributed data to these analyses. A summary of the individual studies is given in **Supplementary Table 1**. All but two comprise subjects of predominantly European descent. Seven of the studies used population-based case ascertainment, nine ascertained cases from hospital-based series and one from a cohort. Five studies specifically included cases with a strong family history and/or bilateral cases. All studies were approved by the appropriate local Institutional Review Board or Research Ethics Committee, and informed consent was obtained from all

Table 3 Subgroup analysis for *TGFB1* L10P and breast cancer risk

	Category	No. of cases	Test for association[b]	Heterozygotes[c] OR	Heterozygotes[c] 95% c.i.	Rare homozygotes[c] OR	Rare homozygotes[c] 95% c.i.	Heterogeneity test[d]
Age group, years[a]	<40	1,123	0.09	1.27	(1.01, 1.60)	1.29	(0.94, 1.76)	0.32
	40–49	3,502	0.15	1.05	(0.93, 1.19)	1.19	(1.00, 1.41)	
	50–59	4,145	0.07	1.08	(0.98, 1.18)	1.16	(1.02, 1.32)	
	≥60	3,808	0.52	1.06	(0.96, 1.16)	1.03	(0.90, 1.18)	
ER status	+	4,571	0.04	1.01	(0.94, 1.09)	1.14	(1.03, 1.27)	0.59
	–	1,398	0.09	1.11	(0.98, 1.25)	1.19	(1.00, 1.42)	
PR status	+	2,473	0.87	0.98	(0.89, 1.09)	1.01	(0.88, 1.17)	0.017
	–	1,318	0.01	1.15	(1.01, 1.31)	1.31	(1.09, 1.57)	
Stage	I	3,175	0.15	1.05	(0.96, 1.14)	1.13	(1.00, 1.28)	0.42
	II	2,762	0.041	1.04	(0.95, 1.14)	1.19	(1.04, 1.35)	
	III/IV	222	0.21	1.15	(0.86, 1.55)	1.43	(0.97, 2.13)	
Grade	1	1,527	0.21	1.02	(0.91, 1.15)	1.16	(0.98, 1.36)	0.35
	2	3,374	0.0096	1.02	(0.93, 1.11)	1.19	(1.06, 1.34)	
	3	2,092	0.0051	1.14	(1.03, 1.26)	1.24	(1.08, 1.43)	
Histopathology	Ductal	6,643	0.0001	1.03	(0.96, 1.10)	1.22	(1.11, 1.33)	0.30
	Lobular	1,236	0.42	1.09	(0.96, 1.24)	1.03	(0.85, 1.24)	
	DCIS	328	0.61	0.89	(0.70, 1.13)	0.90	(0.63, 1.27)	

ER, estrogen receptor; PR, progesterone receptor.
[a]Age in years at diagnosis (cases) or interview (controls). [b]LRT, 2 d.f. [c]Reference group: common homozygotes. [d]*P* value for case-only LRT of between-subgroup heterogeneity.

subjects (for the Netherlands Cancer Institute Study, an approved coding procedure was used; see ref. 17 in **Supplementary Table 1**).

Genotyping. Primers and probes used for TaqMan assays are listed in **Supplementary Table 8** online; alternative assay methods were used by some studies (**Supplementary Table 1**). Genotyping quality control was tested using duplicate DNA samples within studies and SNP assays. For all SNPs, >99% concordant results were obtained. Studies using DNA from lymphocytes on the TaqMan and MALDI-TOF MS platforms obtained genotype calls in >96% of samples tested. A minority of studies that used DNA from paraffin blocks or buccal cells or other genotyping platforms had lower completion rates. Quality control data for each SNP are shown in **Supplementary Table 9** online.

Statistical methods. Deviation of the genotype frequencies in the controls from those expected under Hardy-Weinberg Equilibrium (HWE) was assessed by χ^2 tests (1 degree of freedom (d.f.)), for each study separately. The main test of the null hypothesis of no association (with invasive breast cancer; that is, excluding DCIS) was a likelihood ratio test (LRT) (2 d.f.) comparing a model that included terms for genotype and study with a model including only a term for study, and a trend test (1d.f.) that included a single parameter for allele dose. Genotype-specific risks for each SNP were estimated as ORs for the heterozygote and rare homozygote genotypes with the common homozygote as the baseline category using unconditional logistic regression. We also estimated a per-allele risk under a multiplicative codominant genetic model by fitting the number of rare alleles carried as an ordinal covariate.

Genotype counts from individual studies are given in **Supplementary Table 2** online, and study-specific ORs are given in **Supplementary Table 3** online. We tested for heterogeneity between study strata by comparing logistic regression models with and without a genotype × study interaction term using a likelihood ratio test. Data were also analyzed using a random-effects model to allow for heterogeneity.

We estimated category-specific risks by comparing the genotype distribution of cases and controls within each category (for age) or between each case category and all controls (for other variables) (**Tables 2** and **3** and **Supplementary Tables 4–7**). To investigate the effects of age, subjects were separated into four categories (under 40, 40–49, 50–59 and 60+) according to age at diagnosis (cases) or interview (controls). Family history categories were (i) no family history of breast cancer, (ii) one first-degree relative with breast cancer and (iii) two or more first-degree relatives with breast cancers or bilateral breast cancer cases. Estrogen receptor and progesterone receptor status were categorized as positive or negative; tumor grade as 1, 2 or 3; and stage as I, II or III/IV. Histopathology categories were ductal and lobular. Individuals with DCIS were defined as not having had invasive breast cancer up to and including the time of diagnosis of DCIS. Category-specific data were not available for all subjects; the number of cases with data available for the relevant variables is indicated in **Tables 2** and **3** and **Supplementary Tables 4–7**.

We tested for interaction between genotype and other variables (age at diagnosis, family history, estrogen receptor status, progesterone receptor status, grade, stage and histopathological subtype) using a cases-only design. This approach is more powerful than standard case-control methods for detecting interaction[29]. Polytomous logistic regression was used to compare genotype frequencies in the different subgroups of each category stratified by study (**Tables 2** and **3** and **Supplementary Tables 4–7**). The other variables and the number of rare alleles carried were fitted as ordinal covariates and a LRT (1 d.f.) then used to compare a model that included terms for genotype and study with a model including only a term for study.

The relative risk to daughters of an affected individual attributable to a given SNP was calculated using the formula

$$\lambda^* = \frac{p(pr_2+qr_1)^2+q(pr_1+q)^2}{[p^2r_2+2pqr_1+q^2]^2}$$

where p is the population frequency of the minor allele, $q = 1 - p$, and r_1 and r_2 are the relative risks (estimated as OR) for heterozygotes and rare homozygotes, relative to common homozygotes. The proportion of the familial risk attributable to the SNP was then calculated as $\log(\lambda^*)/\log(\lambda_0)$, where λ_0 is the overall familial relative risk to offspring estimated from epidemiological studies (this formula assumes a multiplicative interaction between the SNP of interest and the other susceptibility alleles). λ_0 was assumed to be 1.8 (ref. 30).

Note: Supplementary information is available on the Nature Genetics website.

ACKNOWLEDGMENTS

The authors thank all participants for taking part in this research. The Sheffield Breast Cancer Study was supported by Yorkshire Cancer Research and the Breast Cancer Campaign. We thank S. Higham, H. Cramp, D. Connley, I. Brock, G. MacPherson, N. Bhattacharyya and M. Meuth for their contribution to this study. SEARCH was funded by Cancer Research-UK (CR-UK). P.D.P.P. is a Senior Clincal Research Fellow, and D.F.E. is a Principal Research Fellow of CR-UK. The Polish Breast Cancer Study was funded by Intramural Research Funds of the US National Cancer Institute. The authors thank N. Szeszenia-Dabrowska of the Nofer Institute of Occupational Medicine and W. Zatonski of the M.Sklodowska-Curie Institute of Oncology and Cancer Center for their contribution to the Polish Breast Cancer Study. The Australian Breast Cancer Family Study (ABCFS) was funded by the Australian National Health and Medical Research Council (NHMRC), the Victorian Health Promotion Foundation, the New South Wales Cancer Council, and, as part of the Breast Cancer Family Registry, by the US National Cancer Institute (RFA # CA-95-003). The content of this manuscript does not necessarily reflect the views or policies of the National Cancer Institute or any of the collaborating centers in the Cancer Family Registries (CFRs), nor does mention of trade names, commercial products or organizations imply endorsement by the US government or the CFR Centers. The Kathleen Cuningham Foundation Consortium for Research into Familial Breast Cancer (kConFab) is supported by the National Breast Cancer Foundation, the NHMRC of Australia and the Cancer Councils of Queensland, New South Wales, Western Australia, South Australia and Victoria. We thank kConFab nurses; the staff of the Familial Cancer Clinics; H. Thorne, L. Williams, D. Surace, L. Tarcova, E. Niedermayr, S. Picken, H. Holland, G. Dite and X. Chen for their contribution to the ABCFS and kConFab studies; and the Clinical Follow-Up Study of kConFab (funded by the NHMRC grants 145684 and 288704) for supplying some data. The genotyping and analysis were supported by grants from the NHMRC. A.B.S. is funded by an NHMRC Career Development Award, and G.C.-T. and J.L.H. are NHMRC Principal and Senior Principal Research Fellows, respectively. The British Breast Cancer study and the Mammography Oestrogens and Growth Factors study are funded by CR-UK and Breakthrough Breast Cancer. The Copenhagen Breast Cancer Study and The Copenhagen City Heart Study were supported by Chief Physician Johan Boserup and Lise Boserup Fund, the Danish Medical Research Council and Copenhagen County. The Gene-Environment Interaction and Breast Cancer in Germany (GENICA) study was supported by the German Human Genome Project and funded by the German Federal Ministry of Education and Research (BMBF) (grants 01KW9975/5, 01KW9976/8, 01KW9977/0 and 01KW0114). Genotyping analyses were supported by Deutsches Krebsforschungszentrum, Heidelberg and the Robert Bosch Foundation of Medical Research (Stuttgart). Y. Ko was involved in the design of the GENICA study and was responsible for patient recruitment and collection of clinical data. B. Pesch was involved in the design of the GENICA study and responsible for recruitment of the study subjects as well as collection of epidemiological data. The Genetic Epidemiology Study of Breast Cancer by Age 50 was supported by the Deutsche Krebshilfe e.V. (project number 70492) and the genotyping in part by the state of Baden-Württemberg through the Medical Faculty of the University of Ulm (P.685). We thank U. Eilber, M. Rohrbacher and T. Koehler for their technical support. The Hannover Breast Cancer Study was supported by an intramural grant of Hannover Medical School. N.B. was supported by a fellowship of the German Research Foundation (DO 761/2-1). We acknowledge the technical assistance of M. Haidukiewicz in DNA sample preparation and the initial contributions of P. Yamini to the ARMS assay for the ATM*S49C variant. We thank C. Sohn, A. Scharf, P. Hillemanns, M. Bremer and J. Karstens for their support in terms of infrastructure and patient samples. The Helsinki Breast Cancer Study was supported by The Academy of Finland (project 110663), Helsinki University Central Hospital Research Funds, The Sigrid Juselius Foundation and The Finnish Cancer Society. We thank Research Nurse N. Puolakka for help with the sample and data collection. The Institute of Cancer Research Familial Breast Cancer Study (ICR_FBCS) is supported by Cancer Research UK. The families are recruited by the Breast Cancer Susceptibility Collaboration (UK). The controls are from the British 1958 Birth Cohort DNA collection funded by the Medical Research Council grant G0000934 and the Wellcome Trust grant 068545/Z/02. We thank S. Wiangnon (Khon Kaen University) and P. Boffetta (IARC) for their contributions to the IARC-Thai study and thank V. Gaborieau (IARC) for statistical support for this study. The Kuopio Breast Cancer Project was supported by special Government Funding to Kuopio University Hospital (grant 5654113) and by the Cancer Fund of North Savo. We are grateful to E. Myöhänen for technical assistance. The Mayo Clinic Breast Cancer Study was supported by US National Institutes of Health grants CA82267 and P50 CA116201 and the US medical research and materiel command breast

cancer IDEA award W81XWH-04-1-0588. ELG is a Fraternal Order of the Eagles Cancer Research Fellow. The Netherlands Cancer Institute thanks L. Braaf, R. Pruntel and R. Tollenaar (Leiden University Medical Center) and other project members of the 'Clinical outcome of breast cancer in BRCA1/2 carriers' study, and we are grateful for funding by the Dutch Cancer Society and the Dutch National Genomics Initiative. The Singapore and Swedish Breast Cancer Study (SASBAC) was supported by funding from the Agency for Science, Technology and Research of Singapore (A*STAR), the US National Institute of Health (NIH) and the Susan G. Komen Breast Cancer Foundation. The Seoul Breast Cancer Study was supported by a grant from the National Research and Development Program for Cancer Control, Ministry of Health and Welfare, Republic of Korea (0620410-1). The Spanish National Cancer Centre study was supported by the Genome Spain Foundation. We thank E. Gonzalez and C. Alonso for their technical support. The US Radiologic Technologist (USRT) study was supported in part by the Intramural Research Program of the Division of Cancer Epidemiology and Genetics and Center for Cancer Research, National Cancer Institute, US National Institutes of Health. The authors are grateful to the radiologic technologists who participated in the USRT Study; J. Reid of the American Registry of Radiologic Technologists for continued support of this study; D. Kampa and A. Iwan of the University of Minnesota for data collection and study coordination; C. McClure of Research Triangle International for tracing and data management; L. Bowen of Information Management Services for biomedical computing and M. Pineda of the Laboratory of Population Genetics for genotyping assistance.

AUTHOR CONTRIBUTIONS

A.C., A.M.D. and M.G.-C. contributed equally to the writing of this manuscript. Data analysis was carried out by P.D.P.P. and M.G.-C. and the project was coordinated by D.F.E. The Sheffield Breast Cancer Study was designed and initiated by M.W.R.R., and genotyping and data management were coordinated by A.C. S.B. was responsible for patient recruitment, collection and validation of clinicopathological data and assisted in revising the manuscript. The SEARCH study was initiated by B.A.J.P. and is managed by P.D.P.P., D.F.E. and B.A.J.P. Genotyping was coordinated by A.M.D., K.A.P. and C. Baynes carried out genotyping within SEARCH and provided reagents, protocols and technical advice to BCAC members. S. Scollen carried out genotyping and assisted in drafting the manuscript. M.G.-C., L.B., J. Lissowska and B.P. initiated the Polish Breast Cancer Study and participated in the study design as well as in data and biological specimen collection. M.G.-C. was responsible for the overall supervision of the study, quality control, and genotyping. S.C. contributed to genotyping. J.L.H., M.R.E.McC., G.G.G and M.C.S. devised and designed the ABCFS study, were responsible for the recruitment of subjects and collection of samples and critically reviewed the manuscript. A.B.S. participated in the study design, genotyping design, supervision and quality control, data management and critical review of the manuscript for the kConFab study. J. Beesley was involved in genotyping design, implementation, quality control and critical review of the manuscript. G.C.-T. was involved in the study design, funding, project supervision and critical review of manuscript. O.F., N.J., L.G. and I.d.S.S. all participated in the design of the British Breast Cancer Study and collection of the samples. N.J. was responsible for the genotyping and O.F. and L.G. were responsible for coding and cleaning of data. S.E.B., B.G.N. and C.K.A. all initiated the Copenhagen Breast Cancer Study and designed the concept. B.G.N. secured funding and provided administrative support. B.G.N. provided controls, and C.K.A. provided patients. S.E.B. directed the molecular analyses and collected and assembled the data. S.E.B., B.G.N. and C.K.A. all revised the manuscript. U.H. initiated the GENICA study, designed the concept and secured funding. She was responsible for molecular analyses and was involved in revising the manuscript. H.B. initiated and coordinated the GENICA study, designed concepts and secured funding. She is responsible for the conduct of molecular analyses and participated in the revision of the manuscript. C.J. participated in the GENICA study, particularly in the molecular design and analyses, and was involved in revising the manuscript. D.T. participated in the GENICA study, particularly in the molecular design and analyses. J.C.-C. initiated and designed the Genetic Epidemiology Study of Breast Cancer by Age 50, secured funding and participated in revising the manuscript. S.K. was responsible for data collection and analysis and participated in revising the manuscript. A. Risch was responsible for the molecular analyses of IGFBP3 and participated in revising the manuscript. S.W.-G. initiated and conducted the molecular analyses for *SOD2* and participated in revising the manuscript. N.B. and P.S. contributed to the molecular design of the Hannover Breast Cancer Study, performed TaqMan assays and ARMS analyses of the HBCS samples and participated in revising the manuscript. T.D. coordinated the Hannover Breast Cancer Study and took part in the project design, experimental work, data evaluation and critical review of the manuscript. R.F. and K.A. coordinated the genotyping, data collection and management, C. Blomquist was responsible for the recruitment of the patients and H.N. initiated and coordinated the Helsinki Breast Cancer Study. M.R.S. and N.R. were responsible for the design of the ICR_FBCS study. S. Seal and A. Renwick undertook the genotyping. The IARC-Thai study was jointly designed by S. Sangrajrang and P.B. S. Sangrajrang coordinated the recruitment of participants and collection of biological samples. The genotyping was coordinated by D.H. and F.O. A.M., V.K. and V.-M.K. participated in the study design for the Kuopio Breast Cancer Project. V.K. coordinated the data collection and clinical data update and contributed to the funding. V.-M.K. was responsible for the histological analyses, revised the manuscript and contributed to the funding. A.M. was responsible for the molecular analyses and was involved in the revision of the manuscript. J.H. contributed to the genotyping and participated in the revision of the manuscript. F.J.C. participated in the Mayo Clinic Breast Cancer Study design, was responsible for the genotyping and was involved in revising the manuscript. J.E.O. oversaw biological sample and data collection and assisted in revising the manuscript, E.L.G. assisted in manuscript revision. A.B., M.K.S., F. B.L.H. and L.J.V.V. were responsible for and/or contributed to project initiation and data collection, performance and interpretation of SNP array test and data cleaning and formatting for Dutch patients (contribution of the Netherlands Cancer Institute) and were involved in revision of the manuscript. P.H., S.W., Y.-L.L. and J. Liu contributed in the securing of funding, study design, genotyping and revision of the manuscript for the SASBAC study. D.K. designed the Seoul Breast Cancer Study. K.-Y.Y. participated in risk factor analysis. D.-Y.N. and S.-H.A. provided biological samples. R.L.M. participated in the CNIO study design, was responsible for data cleaning and formatting and participated in revising the manuscript. G.R. and A.G.-N. were responsible for genotyping and participated in the revision of the manuscript. J. Benitez was responsible for the design of the CNIO study and the securing of funding (Genome Spain grant) and was involved in revising the manuscript. A.J.S. participated in the US Radiologic Technologist (USRT) study design, oversaw biologic sample and data collection and assisted in revising the manuscript. D.L.S. performed the genotyping and participated in the revision of the manuscript. B.H.A. participated in USRT study design, oversaw sample and data collection and assisted in revising the manuscript. J.P.S. participated in the USRT study design, was responsible for the genotyping and was involved in revising the manuscript.

COMPETING INTERESTS STATEMENT

The authors declare that they have no competing financial interests.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Pharoah, P.D. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.* **31**, 33–36 (2002).
2. Colhoun, H.M., McKeigue, P.M. & Davey Smith, G. Problems of reporting genetic associations with complex outcomes. *Lancet* **361**, 865–872 (2003).
3. The Breast Cancer Association Consortium. Commonly studied SNPs and breast cancer: negative results from 12,000 - 32,000 cases and controls from the Breast Cancer Association Consortium. *J. Natl. Cancer Inst.* **98**, 1382–1396 (2006).
4. Stredrick, D.L. *et al.* The ATM missense mutation p.Ser49Cys (c.146C>G) and the risk of breast cancer. *Hum. Mutat.* **27**, 538–544 (2006).
5. Buchholz, T.A. *et al.* A Ser49Cys variant in the ataxia telangiectasia, mutated, gene that is more common in patients with breast carcinoma compared with population controls. *Cancer* **100**, 1345–1351 (2004).
6. Kammerer, S. *et al.* Large-scale association study identifies ICAM gene region as breast and prostate cancer susceptibility locus. *Cancer Res.* **64**, 8906–8910 (2004).
7. Kammerer, S. *et al.* Association of the NuMA region on chromosome 11q13 with breast cancer susceptibility. *Proc. Natl. Acad. Sci. USA* **102**, 2004–2009 (2005).
8. Ferlay, J., Bray, F., Pisani, P. & Parkin, D.M. *GLOBOCAN 2002: Cancer Incidence, Mortality and Prevalence Worldwide* (IARCPress, Lyon, 2004).
9. Hengartner, M.O. The biochemistry of apoptosis. *Nature* **407**, 770–776 (2000).
10. MacPherson, G. *et al.* Association of a common variant of the CASP8 gene with reduced risk of breast cancer. *J. Natl. Cancer Inst.* **96**, 1866–1869 (2004).
11. Frank, B. *et al.* Re: Association of a common variant of the CASP8 gene with reduced risk of breast cancer. *J. Natl. Cancer Inst.* **97**, 1012 (2005).
12. The CHEK2 Breast Cancer Case-Control Consortium. CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. *Am. J. Hum. Genet.* **74**, 1175–1182 (2004).
13. Renwick, A. *et al.* ATM mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat. Genet.* **38**, 873–875 (2006).
14. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
15. Bierie, B. & Moses, H.L. Tumour microenvironment: TGFbeta: the molecular Jekyll and Hyde of cancer. *Nat. Rev. Cancer* **6**, 506–520 (2006).

16. Derynck, R., Akhurst, R.J. & Balmain, A. TGF-beta signaling in tumor suppression and cancer progression. *Nat. Genet.* **29**, 117–129 (2001).
17. Dunning, A.M. *et al.* A transforming growth factorbeta1 signal peptide variant increases secretion in vitro and is associated with increased incidence of invasive breast cancer. *Cancer Res.* **63**, 2610–2615 (2003).
18. Yokota, M., Ichihara, S., Lin, T.L., Nakashima, N. & Yamada, Y. Association of a T29→C polymorphism of the transforming growth factor-beta1 gene with genetic susceptibility to myocardial infarction in Japanese. *Circulation* **101**, 2783–2787 (2000).
19. Ziv, E., Cauley, J., Morin, P.A., Saiz, R. & Browner, W.S. Association between the T29→C polymorphism in the transforming growth factor beta1 gene and breast cancer among elderly white women: the study of osteoporotic fractures. *J. Am. Med. Assoc.* **285**, 2859–2863 (2001).
20. Krippl, P. *et al.* The L10P polymorphism of the transforming growth factor-beta 1 gene is not associated with breast cancer risk. *Cancer Lett.* **201**, 181–184 (2003).
21. Le Marchand, L. *et al.* T29C polymorphism in the transforming growth factor beta1 gene and postmenopausal breast cancer risk: the multiethnic cohort study. *Cancer Epidemiol. Biomarkers Prev.* **13**, 412–415 (2004).
22. Kaklamani, V.G. *et al.* Combined genetic assessment of transforming growth factor-beta signaling pathway variants may predict breast cancer risk. *Cancer Res.* **65**, 3454–3461 (2005).
23. Shin, A., Shu, X.O., Cai, Q., Gao, Y.T. & Zheng, W. Genetic polymorphisms of the transforming growth factor-beta1 gene and breast cancer risk: a possible dual role at different cancer stages. *Cancer Epidemiol. Biomarkers Prev.* **14**, 1567–1570 (2005).
24. Feigelson, H.S. *et al.* Transforming growth factor beta receptor type I and transforming growth factor beta1 polymorphisms are not associated with postmenopausal breast cancer. *Cancer Epidemiol. Biomarkers Prev.* **15**, 1236–1237 (2006).
25. Rohrbacher, M., Risch, A., Kropp, S. & Chang-Claude, J. The A(-336C) insulin-like growth factor binding protein-3 promoter polymorphism is not a modulator of breast cancer risk in Caucasian women. *Cancer Epidemiol. Biomarkers Prev.* **14**, 289–290 (2005).
26. Al-Zahrani, A. *et al.* IGF1 and IGFBP3 tagging polymorphisms are associated with circulating levels of IGF1, IGFBP3 and risk of breast cancer. *Hum. Mol. Genet.* **15**, 1–10 (2006).
27. Schernhammer, E.S., Hankinson, S.E., Hunter, D.J., Blouin, M.J. & Pollak, M.N. Polymorphic variation at the -202 locus in IGFBP3: Influence on serum levels of insulin-like growth factors, interaction with plasma retinol and vitamin D and breast cancer risk. *Int. J. Cancer* **107**, 60–64 (2003).
28. Ren, Z. *et al.* Genetic polymorphisms in the IGFBP3 gene: association with breast cancer risk and blood IGFBP-3 protein levels among Chinese women. *Cancer Epidemiol. Biomarkers Prev.* **13**, 1290–1295 (2004).
29. Piegorsch, W.W., Weinberg, C.R. & Taylor, J.A. Non-hierarchical logistic models and case-only designs for assessing susceptibility in population-based case-control studies. *Stat. Med.* **13**, 153–162 (1994).
30. Collaborative Group on Hormonal Factors in Breast Cancer. Familial breast cancer: collaborative reanalysis of individual data from 52 epidemiological studies including 58,209 women with breast cancer and 101,986 women without the disease. *Lancet* **358**, 1389–1399 (2001).

A common coding variant in *CASP8* is associated with breast cancer risk

Angela Cox, Alison M Dunning, Montserrat Garcia-Closas, Sabapathy Balasubramanian, Malcolm W R Reed, Karen A Pooley, Serena Scollen, Caroline Baynes, Bruce A J Ponder, Stephen Chanock, Jolanta Lissowska, Louise Brinton, Beata Peplonska, Melissa C Southey, John L Hopper, Margaret R E McCredie, Graham G Giles, Olivia Fletcher, Nichola Johnson, Isabel dos Santos Silva, Lorna Gibson, Stig E Bojesen, Børge G Nordestgaard, Christen K Axelsson, Diana Torres, Ute Hamann, Christina Justenhoven, Hiltrud Brauch, Jenny Chang-Claude, Silke Kropp, Angela Risch, Shan Wang-Gohrke, Peter Schürmann, Natalia Bogdanova, Thilo Dörk, Rainer Fagerholm, Kirsimari Aaltonen, Carl Blomqvist, Heli Nevanlinna, Sheila Seal, Anthony Renwick, Michael R Stratton, Nazneen Rahman, Suleeporn Sangrajrang, David Hughes, Fabrice Odefrey, Paul Brennan, Amanda B Spurdle, Georgia Chenevix-Trench, The Kathleen Cunningham Foundation Consortium for Research into Familial Breast Cancer, Jonathan Beesley, Arto Mannermaa, Jaana Hartikainen, Vesa Kataja, Veli-Matti Kosma, Fergus J Couch, Janet E Olson, Ellen L Goode, Annegien Broeks, Marjanka K Schmidt, Frans B L Hogervorst, Laura J Van't Veer, Daehee Kang, Keun-Young Yoo, Dong-Young Noh, Sei-Hyun Ahn, Sara Wedrén, Per Hall, Yen-Ling Low, Jianjun Liu, Roger L Milne, Gloria Ribas, Anna Gonzalez-Neira, Javier Benitez, Alice J Sigurdson, Denise L Stredrick, Bruce H Alexander, Jeffery P Struewing, Paul D P Pharoah & Douglas F Easton, on behalf of the Breast Cancer Association Consortium

The affiliations of Angela Cox, Alison M. Dunning and Montserrat Garcia-Closas are incorrect in this PDF, but will be corrected soon.

PERSPECTIVE

nature genetics

The emerging landscape of breast cancer susceptibility

Michael R Stratton & Nazneen Rahman

The genetic basis of inherited predisposition to breast cancer has been assiduously investigated for the past two decades and has been the subject of several recent discoveries. Three reasonably well-defined classes of breast cancer susceptibility alleles with different levels of risk and prevalence in the population have become apparent: rare high-penetrance alleles, rare moderate-penetrance alleles and common low-penetrance alleles. The contribution of each component to breast cancer predisposition is still to be fully explored, as are the phenotypic characteristics of the cancers associated with them, the ways in which they interact, much of their biology and their clinical utility. These recent advances herald a new chapter in the exploration of susceptibility to breast cancer and are likely to provide insights relevant to other common, heterogeneous diseases.

In most Western populations, approximately one in ten women develop breast cancer. Epidemiological studies have shown that first-degree female relatives of women with breast cancer are at approximately two-fold risk of developing the disease compared to the general population[1]. Although, in principle, this could be attributable to shared environmental or genetic factors, or both, twin studies indicate that most of the excess familial risk is due to inherited predisposition[2].

Rare high-penetrance breast cancer susceptibility genes

Major advances in understanding breast cancer susceptibility were made in the last decade of the twentieth century through genetic linkage mapping and positional cloning of two major predisposition genes, *BRCA1* and *BRCA2* (refs. 3–6). Disease-causing variants in *BRCA1* and *BRCA2* confer a high risk of breast cancer, approximately 10- to 20-fold relative risk. This translates into a 30–60% risk by age 60, compared to 3% in the general population. The relative risks are higher for early-onset breast cancers, and there are also elevated risks of ovarian and other cancers[7,8]. Disease-causing mutations in *BRCA1* and *BRCA2* result in inactivation of the encoded proteins, generally by causing premature protein truncation or nonsense-mediated RNA decay. There is population variation in mutation prevalence, but mutations are infrequent in most populations. Approximately 1 in 1,000 individuals in the UK are heterozygous mutation carriers of each gene, and there are numerous different mutations, each of which is very rare[9,10]. Cancer predisposition is transmitted as an autosomal dominant trait in families harboring mutations. However, at the cellular level, *BRCA1* and *BRCA2* act as recessive cancer genes, with mutations converted to homozygosity in the cancers which they cause, usually through loss of the wild-type allele. Several years of biological investigation have firmly implicated BRCA1 and BRCA2 in double-strand DNA break repair[11].

Mutations in *BRCA1* and *BRCA2* account for ~16% of the familial risk of breast cancer[9,10]. Germline mutations in *TP53* cause Li-Fraumeni syndrome, which includes a high risk of breast and other cancers, but these mutations are very rare and hence account for a much smaller proportion of the familial risk. Cancer predisposition syndromes due to mutations in *PTEN* (Cowden syndrome), *STK11* (Peutz-Jeghers syndrome) and *CDH1* are also associated with elevated risks of breast cancer, although the cancer risks and prevalence of mutations in these genes are not well defined. It is unlikely that mutations in all six of these genes together account for more than 20% of the familial risk of the disease[12,13]. Genome-wide linkage analyses using large numbers of families without mutations in *BRCA1* or *BRCA2* have not mapped additional susceptibility loci[14]. Although this does not completely exclude the existence of further high-penetrance breast cancer susceptibility genes, it strongly suggests that, if they exist, they account for a very small fraction of familial risk. So, how can the remaining ~80% of the familial risk of breast cancer be explained?

A new harvest of breast cancer susceptibility alleles has recently emerged through two distinct strategies: direct interrogation of genes believed to be strong candidates, which has led to the identification of rare moderate-penetrance alleles[15–19], and genome-wide tag SNP association studies, which have identified common low-penetrance alleles[20–22] (**Box 1**). We have considered these two new classes separately and in distinction to the rare high-penetrance genes discussed previously. It is possible that the differences among these classes may, at least in part, be attributable to the methods employed in their identification, and further discoveries may render the boundaries among them less distinct. Nevertheless, they currently provide a useful basis for considering the genetic landscape of breast cancer susceptibility.

Rare moderate-penetrance breast cancer susceptibility genes

The candidacy of the breast cancer susceptibility genes recently identified through direct interrogation for disease-causing mutations has been

Michael R. Stratton and Nazneen Rahman are at the Section of Cancer Genetics, Institute of Cancer Research, 15 Cotswold Road, Sutton, Surrey SM2 5NG, UK. Michael R. Stratton is in the Cancer Genome Project, The Wellcome Trust Sanger Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SA, UK. e-mail: nazneen.rahman@icr.ac.uk

Published online 27 December 2007; doi:10.1038/ng.2007.53

Box 1 Classes and key features of known breast cancer susceptibility alleles

High-penetrance breast cancer susceptibility genes

Examples: *BRCA1*, *BRCA2*, *TP53*

- **Risk variants:** Multiple, different mutations that predominantly cause protein truncation
- **Frequency:** Rare (population carrier frequency ≤0.1%)
- **Risk of breast cancer:** 10- to 20-fold relative risk
- **Primary strategy for identification:** Genome-wide linkage and positional cloning

Moderate-penetrance breast cancer susceptibility genes

Examples: *ATM*, *BRIP1*, *CHEK2*, *PALB2*

- **Risk variants:** Multiple, different mutations that predominantly cause protein truncation
- **Frequency:** Rare (population carrier frequency ≤0.6%)
- **Risk of breast cancer:** two- to fourfold relative risk
- **Primary strategy for identification:** Direct interrogation of candidate genes for coding variants in large, genetically enriched breast cancer case series and controls

Low-penetrance breast cancer susceptibility alleles

Examples: rs2981582 (*FGFR2*, 10q), rs3803662 (*TNRC9* (recently renamed *TOX3*), 16q), rs889312 (*MAP3K1*, 5q), rs3817198 (*LSP1*, 11p), rs13281615 (8q), rs13387042 (2q), rs1045485 (*CASP8*_D302H)

- **Risk variants:** Single-nucleotide polymorphisms that are causal or in linkage disequilibrium with the causal variant(s). May occur in noncoding, nongenic regions.
- **Frequency:** Common (population frequency 5–50%)
- **Risk of breast cancer**: up to ~1.25-fold (heterozygous) or 1.65-fold (homozygous) relative risk
- **Primary strategy for identification:** Genome-wide association studies of hundreds of thousands of SNPs in large breast cancer case-control series

based primarily on involvement of the encoded proteins in biological pathways that include BRCA1 and BRCA2. To date, this strategy has identified at least four genes: *CHEK2*, *ATM*, *BRIP1* and *PALB2* (refs. 15–19). CHEK2 is a checkpoint kinase involved in DNA repair that directly modulates the activities of p53 and BRCA1 by phosphorylation[23]. *ATM* also encodes a checkpoint kinase that has key functions in DNA repair, and which also phosphorylates p53 and BRCA1 (ref. 24). BRIP1 (also known as BACH1) was discovered as a binding partner of BRCA1 and is implicated in some BRCA1 activities relating to DNA repair[25]. PALB2 was discovered as a protein associated with BRCA2 (ref. 26). The patterns of susceptibility associated with these four genes have many features in common.

In *CHEK2*, *ATM*, *BRIP1* and *PALB2*, most of the disease-causing mutations result in premature protein truncation or nonsense-mediated RNA decay through nonsense codons or translational frameshifts. A small proportion is likely to be rare missense variants that disrupt critical functions. In each of the four genes, there are multiple different pathogenic mutations, each of which is generally very rare. Disease-causing mutations in each gene are found in less than 1% of the UK population: ~0.6% are heterozygous carriers of *CHEK2* mutations (a single mutation, *CHEK2**1100delC, accounts for most of these), ~0.4% are heterozygous carriers of *ATM* mutations and ~0.1% or fewer are heterozygous carriers of *BRIP1* or *PALB2* mutations[15–18,27]. The prevalence of mutations in most other populations is currently less well characterized, although it is noteworthy that founder mutations in *CHEK2* and *PALB2* in Finland allowed independent identification of the association of these genes with breast cancer[19,28].

Overall, with respect to their effect on protein function, their prevalence in the population and their biological consequences, disease-causing mutations in *CHEK2*, *ATM*, *BRIP1* and *PALB2* bear many similarities to disease-causing mutations in *BRCA1* and *BRCA2*. Where they differ is in the risks of breast cancer they confer. Although there is currently some imprecision in the risk estimates, it is clear that mutations in *CHEK2*, *ATM*, *BRIP1* and *PALB2* confer less elevated risks of breast cancer (about two- to threefold, with confidence intervals ranging from 1.2 to 3.9) than mutations in *BRCA1* or *BRCA2* (10- to 20-fold)[15–18,27]. Carriers of moderate-penetrance mutant alleles therefore have approximately a 6–10% risk of developing breast cancer by age 60, compared to ~3% in the general population. For each gene, it is possible that there is risk heterogeneity, with some variants conferring greater risks than others (as is the case for *BRCA1* and *BRCA2* mutations), but there are currently few persuasive examples of this. Because *CHEK2*, *ATM*, *BRIP1* and *PALB2* mutations confer a smaller increased risk of breast cancer than *BRCA1* and *BRCA2* mutations, and their disease-causing mutations are uncommon, each of these moderate-risk genes makes a relatively small contribution to the overall familial risk of breast cancer. Current estimates suggest that mutations in the four genes together account for 2.3% of the familial risk of breast cancer, compared to 16% for *BRCA1* and *BRCA2* together[9,10,12,15].

Features of rare moderate-penetrance susceptibility genes

Despite the many similarities of *CHEK2*, *ATM*, *BRIP1* and *PALB2* to *BRCA1* and *BRCA2*, the lower breast cancer risk conferred by mutations in the former group leads to some uncomfortable departures from familiar genetic patterns. For example, in breast cancer–affected families carrying *BRCA1* or *BRCA2* mutations, the mutation and disease status usually track together, although even in this context the occasional sporadic 'phenocopy' is encountered. However, when the breast cancer risks associated with a particular allele are only two- to threefold, disease-causing mutations often do not segregate with the disease. This is because most mutation carriers do not actually develop breast cancer, because the sporadic rate of breast cancer is high, and because familial breast cancer clusters not associated with mutations in *BRCA1* or *BRCA2* probably reflect chance aggregations of susceptibility alleles in multiple different genes. As a consequence, segregation of the disease with the mutation, which is one of the tests a new disease susceptibility gene is routinely subjected to, is generally unhelpful for confirmation of lower-penetrance alleles. If sufficient multiply sampled breast cancer–affected families with mutations are analyzed, it should be possible to formally show that the mutation segregates with the disease more frequently than

would occur simply by chance. Thus far, however, sufficient families have only been available to show this for *CHEK2* (ref. 16).

Similarly, the familiar pattern of loss of the wild-type allele in cancers, which is generally associated with high-penetrance autosomal dominant cancer genes that operate in a recessive fashion in cancer cells, may be less apparent when sought in the context of lower-penetrance susceptibility alleles. Given the predominant pattern of inactivating disease-causing mutations, it is mechanistically plausible that *CHEK2*, *ATM*, *BRIP1* and *PALB2* behave in a fashion similar to *BRCA1* and *BRCA2* and show somatic loss of the wild-type allele in the cancers they cause. However, to demonstrate this pattern may require analysis of a substantial number of tumors, because only about half of breast cancers in individuals with a mutation in a cancer susceptibility gene conferring a twofold risk arise because of the mutation—the remainder would have occurred anyway. Allelic loss in cancers not due to the mutation will follow the pattern present in sporadic cancers for that locus, and will target the wild-type and mutant alleles equally. Thus, it may be necessary to analyze a large series of breast cancers from mutation carriers before meaningful, statistically robust data on loss of the wild-type allele can be obtained.

Elucidation of the phenotypes associated with heterozygous mutations in *CHEK2*, *ATM*, *BRIP1* and *PALB2* will also be hindered by the considerations discussed above, compounded by the rarity of disease-causing alleles. At this stage, strong evidence does not exist for a higher risk of early-onset breast cancer, but most studies have had insufficient power to demonstrate it. The risks of other cancers, and the histological phenotypes of the breast cancers associated with mutations in these genes, are uncertain and may require large-scale collaborative initiatives to generate sufficient numbers.

Phenotypes associated with biallelic mutations

Mutations in high- and moderate-penetrance breast cancer genes confer an elevated risk of breast cancer in monoallelic (heterozygous) carriers. However, individuals with biallelic (homozygous or compound heterozygous) mutations in some of these genes have a different phenotype, often manifesting during childhood. This is exemplified by *ATM*, which was initially discovered by positional cloning of the gene underlying ataxia telangiectasia, an autosomal recessive condition characterized by loss of cerebellar Purkinje cells, immune deficiency and cancer predisposition[29]. Several epidemiological studies over the past two decades have shown that heterozygous (monoallelic) female carriers of ataxia telangiectasia–causing *ATM* mutations are at elevated risk of breast cancer, and molecular confirmation of this association was finally reported last year[17,30].

Similarly, in 2002, it was shown that biallelic *BRCA2* mutations cause a rare subgroup of Fanconi anemia, subtype FA-D1 (ref. 31). Fanconi anemia is a genetically heterogeneous, recessive, chromosomal instability disorder characterized by growth retardation, skeletal abnormalities, bone marrow failure, cancer predisposition and cellular hypersensitivity to DNA cross-linking agents. FA-D1 is a distinctive subtype associated with severe disease and a high risk of childhood solid tumors such as Wilms tumor, medulloblastoma and glioma that occur rarely in classic Fanconi anemia[32]. Subsequently, it was shown that biallelic mutations in *BRIP1* and *PALB2* also cause rare subgroups of Fanconi anemia (FA-J and FA-N, respectively)[33–36]. The phenotype of FA-N, resulting from biallelic *PALB2* mutations, is characterized by severe disease and a high risk of childhood solid tumors and is virtually identical to that of FA-D1, presumably reflecting the close functional relationship between BRCA2 and PALB2 (refs. 32,34). However, FA-J, caused by biallelic *BRIP1* mutations, results in the classic Fanconi anemia phenotype and has not been associated with childhood solid tumors[33,36]. It is possible that biallelic mutations in additional breast cancer susceptibility genes are responsible for other Fanconi anemia subtypes. However, both epidemiological and molecular analyses suggest that only a subset of Fanconi anemia genes are breast cancer susceptibility genes[37]. The factors that determine whether a Fanconi anemia gene is also a breast cancer predisposition gene are not known.

There is no known phenotype associated with biallelic mutations in *CHEK2* or *BRCA1*. One individual homozygous for *CHEK2**1100delC has been reported and was healthy until developing colorectal cancer at 52 years[38]. Conversely, although more than a decade has elapsed since *BRCA1* was identified, no confirmed *BRCA1* biallelic mutation carrier has been reported. It is conceivable that biallelic *BRCA1* mutations cause a rare syndrome yet to be attributed to this gene, are embryonic lethal or (perhaps less likely) are not associated with any distinctive phenotype.

Common low-penetrance breast cancer susceptibility alleles

A third component of the landscape of breast cancer susceptibility has been the subject of speculation for years, but has only just begun to surface. It is comprised of common alleles that confer very small increases in risk (common low-penetrance alleles). The currently known susceptibility alleles of this type have been discovered through association studies, either targeted at individual genes on the basis of biological candidacy or, more recently, through genome-wide tag SNP searches. In the past, numerous associations were proposed from targeted association studies involving relatively small numbers of cases and controls. Most of these have not been confirmed when evaluated on additional series, and such observations have acquired a certain notoriety and disrepute. Progress in this area of breast cancer research has depended, at least in part, on the formation of multigroup collaborations that combine data from very large numbers of cases and controls from many different locations and ethnic groups. These combined sets of tens of thousands of cases and controls provide substantial power to detect small effects and can obviate problems and limitations intrinsic to individual series[39].

Only a small number of statistically unimpeachable, common low-penetrance breast cancer susceptibility alleles have thus far been reported and confirmed in different populations[20–22]. For the purposes of this review, we focus on seven for which there is strong evidence and that can serve to illustrate at least the outlines of the emerging landscape[20–22,40]. However, these are unlikely to represent all the patterns that will be found in future studies.

Five of the seven confirmed breast cancer risk alleles are within regions of linkage disequilibrium that cover known protein-coding genes. The genes in these regions include *CASP8* (encoding caspase 8, a member of the cysteine-aspartic acid protease family whose sequential activation has a central role in the execution of apoptosis), *FGFR2* (encoding fibroblast growth factor receptor 2), *TNRC9* (recently renamed *TOX3*, encoding a protein with a putative high-mobility-group motif suggesting that it might act as a transcription factor), *MAP3K1* (encoding mitogen-activated protein kinase kinase kinase 1, a protein likely involved in growth signaling) and *LSP1* (encoding lymphocyte-specific protein 1, an intracellular F-actin binding protein). Some of these regions of linkage disequilibrium contain other genes, and it is conceivable that the functional associations are related to these rather than to the genes cited above, or perhaps to other, currently cryptic, genetic elements. Two of the seven susceptibility loci are on 8q and 2q, in regions with no known protein-coding genes[20–22,40].

The increased risks of breast cancer conferred by these seven susceptibility alleles are small. The relative risks of breast cancer associated with carrying a single copy of each risk allele range from 1.07 to 1.26, with the *FGFR2* and 2q susceptibility alleles at the high end of this spectrum. The population prevalence of each risk allele is high, however, ranging from 28% to 87%. Interestingly, for some of these loci, the higher-risk

allele is the more common. Because the predisposing alleles are common, despite the low risks they confer, their contribution to the familial risk of breast cancer is relatively substantial. The six loci characterized by Easton *et al.* and Cox *et al.* are estimated to account for 3.9% of the familial risk of breast cancer in European populations[20,40].

It is likely that there are very few, if any, additional common low-penetrance susceptibility alleles that make contributions to the familial risk of breast cancer as substantial as those in *FGFR2* or the locus on 2q. However, there is evidence for the existence of many, perhaps hundreds of, yet-to-be-discovered common susceptibility alleles with smaller effects[20]. Therefore, a sizeable proportion of the genetic architecture of breast cancer susceptibility may be embodied in a multitude of common susceptibility alleles, each of which accounts for a very small fraction of the familial risk.

Features of common low-penetrance susceptibility alleles

The disease-causing variants underlying these recently reported associations may not be easily identifiable, because the primary association is with a sentinel, reporter SNP that is often in tight linkage disequilibrium with many nearby variants. Even if the disease-causing variant is ultimately identified, it may not be obvious which gene(s) mediates its biological effects. Despite these complications and the limited number of common low-penetrance breast cancer susceptibility alleles thus far identified, some incipient trends and patterns may be emerging.

First, common low-penetrance breast cancer risk variants frequently reside in noncoding regions of the genome. For example, the susceptibility variant in *FGFR2* is within an intron of the gene. Moreover, the susceptibility variants on 2q and 8q are both several tens of kilobases away from the nearest protein-coding genes. Of particular interest is the locus on 8q, which is in close proximity to different linkage disequilibrium blocks that contain alleles predisposing to prostate cancer and colorectal cancer[41–47]. It seems unlikely that this physical clustering is simply coincidence. Nevertheless, it remains to be seen whether these associations are mediated by a related biological mechanism.

Second, the mechanism of action of at least some common low-risk breast cancer–predisposing loci may be through activation of growth-promoting genes, in contrast to the inactivation of DNA repair genes that characterizes known rare high- and moderate-risk genes. For example, somatically acquired missense mutations, amplification and overexpression of *FGFR2* are well documented in human cancer and result in overactivity of the protein[48,49]. Furthermore, the gene closest to the breast, prostate and colorectal cancer risk variants on 8q, remarkably, is *MYC*, which is commonly amplified or overexpressed through chromosomal rearrangement in many types of cancer. Assuming that the predisposing variants at these loci are exerting their effects through FGFR2 and MYC (which is by no means certain), our current understanding of these genes would predict that the susceptibility alleles increase the activity of the encoded proteins. However, most of the currently mapped common low-penetrance loci are anonymous or have functions previously unrelated to cancer development, and they therefore may lead us into previously uncharted areas of cancer biology.

Third, in contrast to the rare high-penetrance and moderate-penetrance genes, homozygosity for a common low-penetrance susceptibility variant does not usually confer a distinct phenotype. Instead, homozygotes are phenotypically normal, but have an increased breast cancer risk that seems to be approximately the product of the risk for heterozygotes. Exploration of the histological phenotypes of cancers associated with common low-penetrance alleles is in its infancy, although at least some of these alleles seem to be particularly associated with estrogen receptor–positive breast cancers, in contrast to *BRCA1* mutations, which are strongly associated with estrogen receptor–negative tumors[22,50].

Identification of further breast cancer susceptibility genes

The recent discoveries described here have together exposed a clearer picture of the genetic architecture of breast cancer susceptibility. *BRCA1* and *BRCA2* are likely to be the only major high-penetrance breast cancer susceptibility genes, and together with other rare, high-penetrance genes, they account for approximately 20% of the familial risk of disease. The remaining susceptibility is therefore due to genes conferring more modest increases in risk. *CHEK2*, *ATM*, *BRIP1* and *PALB2* are breast cancer susceptibility genes that bear many biological similarities to *BRCA1* and *BRCA2* but confer a breast cancer relative risk of two- to fourfold. They represent the current paradigms for a second class of rare moderate-penetrance risk alleles, but it would not be surprising if other such genes exist.

As disease-causing mutations in these genes do not generally result in large pedigrees with multiple breast cancer cases, further susceptibility genes of this class will not easily be mapped by genetic linkage analysis. Moreover, because the disease-causing alleles are uncommon, it is unlikely that they will be detected by association studies. Therefore, the most effective strategy to detect this class of gene is likely to remain the systematic screening of entire genes for potential disease-causing variants (usually truncating mutations) in series of breast cancer cases compared to controls. Because the breast cancer risks conferred by these variants are only two- to fourfold and the risk alleles are rare, the numbers of subjects required in these studies are large, rendering the analyses laborious by current technology. The problem can, to some extent, be mitigated by using familial rather than population-based breast cancer cases, as even lower-penetrance breast cancer susceptibility alleles are usually enriched in familial breast cancer cases compared to nonfamilial series. Use of population isolates with founder mutations of higher prevalence than is typical of outbred populations can also empower gene identification studies[19]. Such studies in Finnish breast cancer cases have provided suggestive data that *RAD50* may be a moderate-penetrance breast cancer predisposition gene, although the rarity of truncating mutations precluded confirmation of an association with breast cancer in UK families[51,52]. It is difficult to predict how many more rare moderate-penetrance genes exist, how much breast cancer susceptibility is accounted for by this component of the landscape or whether this pattern of susceptibility will extend beyond genes involved in DNA repair. Furthermore, the resequencing studies required for their identification are currently restricted to limited sets of candidate genes. However, with the likely advent of genome-wide resequencing of constitutional DNA, further exploration of this class of susceptibility allele should be possible.

Finally, the floodgates seem to be opening for the set of common low-penetrance alleles that confer risks of 1.3-fold or less. Although the current state of knowledge is sketchy, we can at least now be sure that they exist and that they show biological differences from the rare high-penetrance and rare moderate-penetrance genes. Only a small proportion of the familial risk of breast cancer is thus far explained by well-supported examples of this class of susceptibility allele. However, it is possible that a substantial proportion of the still unexplained (>70%) familial risk may be due to large numbers of similar variants with smaller effects. Further studies should yield additional variants in this class, although even with existing large-scale collaborations, sufficient samples may not yet be available to conclusively identify many variants with weak effects.

Are there other areas of the landscape to be explored? An intriguing feature is the apparent discontinuity of breast cancer risks among the three currently defined groups of susceptibility alleles. Mutations in *BRCA1* and *BRCA2* confer 10- to 20-fold relative risks of breast cancer, the rare moderate-penetrance genes confer relative risks of 2- to 4-fold and the common low-penetrance alleles confer relative risks less than

1.3-fold. Whether this pattern reflects a genuine biological stratification or an ascertainment artifact compounded by the limited number of known alleles remains to be seen.

It is also plausible that rare, nontruncating variants contribute to the genetic architecture of breast cancer susceptibility, given that rare truncating and common nontruncating variants are already known to be important. Investigating the role of rare nontruncating variants will, however, be challenging; their rarity will severely hamper detection through association studies, and it is very difficult to distinguish pathogenic nontruncating variants a priori from the plethora of innocuous rare variants.

Interactions between breast cancer susceptibility alleles

The available data suggest that many familial breast cancer clusters are likely to be due to the coincidence of multiple, lower-risk breast cancer susceptibility alleles[13,53]. This raises the question of the manner in which each breast cancer susceptibility allele in such clusters interacts with the others. The evidence for the common low-penetrance variants seems to indicate that, in general, they interact with each other multiplicatively[20,22]. Investigation of the breast cancer risks conferred by *CHEK2**1100delC, however, showed that the pattern of multiplicative interaction does not always apply. Although *CHEK2**1100delC confers an approximately twofold risk of breast cancer in most genetic backgrounds, it does not seem to confer an elevated breast cancer risk in carriers of *BRCA1* or *BRCA2* mutations[16]. Understanding that the proteins encoded by these genes lie in the same biological pathways provides a simple but credible explanation. In this example, abrogation of functions of these pathways by an inactivating mutation of *BRCA1*, *BRCA2* or *CHEK2* confers breast cancer susceptibility. However, if the relevant function is already abolished by a *BRCA1* or *BRCA2* mutation, an inactivating mutation in *CHEK2* will not confer an additional breast cancer risk. Because CHEK2 is known to phosphorylate and regulate BRCA1 and is involved elsewhere in double-strand DNA break repair, this notion has a reasonably solid foundation in our current understanding of these pathways[11,23].

It is currently unknown how common susceptibility alleles interact with rare susceptibility variants, though it is likely that relevant data will be forthcoming in the near future. Exploration of interactions among breast cancer risk alleles and nongenetic factors, such as hormonal profiles and environmental exposures, is also in its infancy, and will be vital in building a comprehensive picture of the underlying causes of familial clustering of the disease.

Clinical utility

Diagnostic testing for mutations in *BRCA1* and *BRCA2* has been routine clinical practice in many countries for several years. It facilitates risk estimation and implementation of cancer prevention strategies and increasingly has the potential to influence cancer therapy[54,55]. Management interventions in breast cancer–affected families without *BRCA1* or *BRCA2* mutations have inevitably been more limited, as less information has been available for risk evaluation. The identification of new susceptibility alleles may offer the potential for improved care in such families: for example, if combinations of alleles alter the risk category of an individual such that screening or prophylactic interventions might be considered. However, clinical testing of the new generation of susceptibility genes will need to be undertaken carefully and cautiously, and more detailed information on the associated risks and interactions will first be required. Implementing routine testing of a large number of different susceptibility alleles in a substantial set of genes will also require careful deliberation, as it may generate considerable technical and economic burdens for clinical diagnostic services.

Future challenges

These recent advances have underscored the complexity of breast cancer susceptibility, revealing at least three different strata in the genetic architecture of the disease: rare high-penetrance alleles, rare moderate-penetrance alleles and common low-penetrance alleles. It is likely that these categories of susceptibility alleles are germane to many other complex conditions. However, their exploration remains demanding, particularly as the identification of alleles underlying each class requires different strategies and technologies. Moreover, despite the remarkable progress made in the last year, most of the familial risk of breast cancer remains unexplained, highlighting the need for ongoing efforts to expand our view of the emerging landscape of breast cancer susceptibility.

ACKNOWLEDGMENTS

We are grateful to C. Turnbull and R. Scott for their critical reading of the manuscript and helpful comments.

Published online at http://www.nature.com/naturegenetics
Reprints and permissions information is available online at http://npg.nature.com/reprintsandpermissions

1. Collaborative Group on Hormonal Factors in Breast Cancer. Breast cancer and breastfeeding: collaborative reanalysis of individual data from 47 epidemiological studies in 30 countries, including 50,302 women with breast cancer and 96,973 women without the disease. *Lancet* **360**, 187–195 (2002).
2. Peto, J. & Mack, T.M. High constant incidence in twins and other relatives of women with breast cancer. *Nat. Genet.* **26**, 411–414 (2000).
3. Hall, J.M. *et al.* Linkage of early-onset familial breast cancer to chromosome 17q21. *Science* **250**, 1684–1689 (1990).
4. Wooster, R. *et al.* Localization of a breast cancer susceptibility gene, *BRCA2*, to chromosome 13q12–13. *Science* **265**, 2088–2090 (1994).
5. Wooster, R. *et al.* Identification of the breast cancer susceptibility gene *BRCA2*. *Nature* **378**, 789–792 (1995).
6. Miki, Y. *et al.* A strong candidate for the breast and ovarian cancer susceptibility gene *BRCA1*. *Science* **266**, 66–71 (1994).
7. Thompson, D. & Easton, D.F. Cancer incidence in *BRCA1* mutation carriers. *J. Natl. Cancer Inst.* **94**, 1358–1365 (2002).
8. The Breast Cancer Linkage Consortium. Cancer risks in *BRCA2* mutation carriers. *J. Natl. Cancer Inst.* **91**, 1310–1316 (1999).
9. Anglian Breast Cancer Study Group. Prevalence and penetrance of *BRCA1* and *BRCA2* mutations in a population-based series of breast cancer cases. *Br. J. Cancer* **83**, 1301–1308 (2000).
10. Peto, J. *et al.* Prevalence of *BRCA1* and *BRCA2* gene mutations in patients with early-onset breast cancer. *J. Natl. Cancer Inst.* **91**, 943–949 (1999).
11. Gudmundsdottir, K. & Ashworth, A. The roles of BRCA1 and BRCA2 and associated proteins in the maintenance of genomic stability. *Oncogene* **25**, 5864–5874 (2006).
12. Thompson, D. & Easton, D. The genetic epidemiology of breast cancer genes. *J. Mammary Gland Biol. Neoplasia* **9**, 221–236 (2004).
13. Antoniou, A.C. & Easton, D.F. Models of genetic susceptibility to breast cancer. *Oncogene* **25**, 5898–5905 (2006).
14. Smith, P. *et al.* A genome wide linkage search for breast cancer susceptibility genes. *Genes Chromosom. Cancer* **45**, 646–655 (2006).
15. Rahman, N. *et al. PALB2*, which encodes a BRCA2-interacting protein, is a breast cancer susceptibility gene. *Nat. Genet.* **39**, 165–167 (2007).
16. Meijers-Heijboer, H. *et al.* Low-penetrance susceptibility to breast cancer due to CHEK2(*)1100delC in noncarriers of *BRCA1* or *BRCA2* mutations. *Nat. Genet.* **31**, 55–59 (2002).
17. Renwick, A. *et al. ATM* mutations that cause ataxia-telangiectasia are breast cancer susceptibility alleles. *Nat. Genet.* **38**, 873–875 (2006).
18. Seal, S. *et al.* Truncating mutations in the Fanconi anemia J gene *BRIP1* are low-penetrance breast cancer susceptibility alleles. *Nat. Genet.* **38**, 1239–1241 (2006).
19. Erkko, H. *et al.* A recurrent mutation in *PALB2* in Finnish cancer families. *Nature* **446**, 316–319 (2007).
20. Easton, D.F. *et al.* Genome-wide association study identifies novel breast cancer susceptibility loci. *Nature* **447**, 1087–1093 (2007).
21. Hunter, D.J. *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).
22. Stacey, S.N. *et al.* Common variants on chromosomes 2q35 and 16q12 confer susceptibility to estrogen receptor-positive breast cancer. *Nat. Genet.* **39**, 865–869 (2007).
23. Ahn, J., Urist, M. & Prives, C. The Chk2 protein kinase. *DNA Repair (Amst.)* **3**, 1039–1047 (2004).
24. Shiloh, Y. The ATM-mediated DNA-damage response: taking shape. *Trends Biochem. Sci.* **31**, 402–410 (2006).
25. Peng, M., Litman, R., Jin, Z., Fong, G. & Cantor, S.B. BACH1 is a DNA repair protein supporting BRCA1 damage response. *Oncogene* **25**, 2245–2253 (2006).
26. Xia, B. *et al.* Control of BRCA2 cellular and clinical functions by a nuclear partner, PALB2. *Mol. Cell* **22**, 719–729 (2006).

27. CHEK2 Breast Cancer Case-Control Consortium. CHEK2*1100delC and susceptibility to breast cancer: a collaborative analysis involving 10,860 breast cancer cases and 9,065 controls from 10 studies. *Am. J. Hum. Genet.* **74**, 1175–1182 (2004).
28. Vahteristo, P. *et al.* A *CHEK2* genetic variant contributing to a substantial fraction of familial breast cancer. *Am. J. Hum. Genet.* **71**, 432–438 (2002).
29. Savitsky, K. *et al.* A single ataxia telangiectasia gene with a product similar to PI-3 kinase. *Science* **268**, 1749–1753 (1995).
30. Thompson, D. *et al.* Cancer risks and mortality in heterozygous *ATM* mutation carriers. *J. Natl. Cancer Inst.* **97**, 813–822 (2005).
31. Howlett, N.G. *et al.* Biallelic inactivation of *BRCA2* in Fanconi anemia. *Science* **297**, 606–609 (2002).
32. Reid, S. *et al.* Biallelic *BRCA2* mutations are associated with multiple malignancies in childhood including familial Wilms tumour. *J. Med. Genet.* **42**, 147–151 (2005).
33. Levitus, M. *et al.* The DNA helicase BRIP1 is defective in Fanconi anemia complementation group *J. Nat. Genet.* **37**, 934–935 (2005).
34. Reid, S. *et al.* Biallelic mutations in *PALB2* cause Fanconi anemia subtype FA-N and predispose to childhood cancer. *Nat. Genet.* **39**, 162–164 (2007).
35. Xia, B. *et al.* Fanconi anemia is associated with a defect in the BRCA2 partner PALB2. *Nat. Genet.* **39**, 159–161 (2007).
36. Levran, O. *et al.* The BRCA1-interacting helicase BRIP1 is deficient in Fanconi anemia. *Nat. Genet.* **37**, 931–933 (2005).
37. Seal, S. *et al.* Evaluation of Fanconi anemia genes in familial breast cancer predisposition. *Cancer Res.* **63**, 8596–8599 (2003).
38. van Puijenbroek, M. *et al.* Homozygosity for a CHEK2*1100delC mutation identified in familial colorectal cancer does not lead to a severe clinical phenotype. *J. Pathol.* **206**, 198–204 (2005).
39. Breast Cancer Association Consortium. Commonly studied single-nucleotide polymorphisms and breast cancer: results from the Breast Cancer Association Consortium. *J. Natl. Cancer Inst.* **98**, 1382–1396 (2006).
40. Cox, A. *et al.* A common coding variant in CASP8 is associated with breast cancer risk. *Nat. Genet.* **39**, 352–358 (2007).
41. Amundadottir, L.T. *et al.* A common variant associated with prostate cancer in European and African populations. *Nat. Genet.* **38**, 652–658 (2006).
42. Gudmundsson, J. *et al.* Genome-wide association study identifies a second prostate cancer susceptibility variant at 8q24. *Nat. Genet.* **39**, 631–637 (2007).
43. Haiman, C.A. *et al.* Multiple regions within 8q24 independently affect risk for prostate cancer. *Nat. Genet.* **39**, 638–644 (2007).
44. Yeager, M. *et al.* Genome-wide association study of prostate cancer identifies a second risk locus at 8q24. *Nat. Genet.* **39**, 645–649 (2007).
45. Haiman, C.A. *et al.* A common genetic risk factor for colorectal and prostate cancer. *Nat. Genet.* **39**, 954–956 (2007).
46. Tomlinson, I. *et al.* A genome-wide association scan of tag SNPs identifies a susceptibility variant for colorectal cancer at 8q24.21. *Nat. Genet.* **39**, 984–988 (2007).
47. Zanke, B.W. *et al.* Genome-wide association scan identifies a colorectal cancer susceptibility locus on chromosome 8q24. *Nat. Genet.* **39**, 989–994 (2007).
48. Pollock, P.M. *et al.* Frequent activating *FGFR2* mutations in endometrial carcinomas parallel germline mutations associated with craniosynostosis and skeletal dysplasia syndromes. *Oncogene* **26**, 7158–7162 (2007).
49. Greenman, C. *et al.* Patterns of somatic mutation in human cancer genomes. *Nature* **446**, 153–158 (2007).
50. Honrado, E., Benitez, J. & Palacios, J. Histopathology of BRCA1- and BRCA2-associated breast cancer. *Crit. Rev. Oncol. Hematol.* **59**, 27–39 (2006).
51. Heikkinen, K. *et al. RAD50* and *NBS1* are breast cancer susceptibility genes associated with genomic instability. *Carcinogenesis* **27**, 1593–1599 (2006).
52. Tommiska, J. *et al.* Evaluation of *RAD50* in familial breast cancer predisposition. *Int. J. Cancer* **118**, 2911–2916 (2006).
53. Pharoah, P.D. *et al.* Polygenic susceptibility to breast cancer and implications for prevention. *Nat. Genet.* **31**, 33–36 (2002).
54. Farmer, H. *et al.* Targeting the DNA repair defect in BRCA mutant cells as a therapeutic strategy. *Nature* **434**, 917–921 (2005).
55. Domchek, S.M. & Weber, B.L. Clinical management of *BRCA1* and *BRCA2* mutation carriers. *Oncogene* **25**, 5825–5831 (2006).

SUPPORTING DATA

None