nature genetics

Genome-wide association study identifies variants in the *ABO* locus associated with susceptibility to pancreatic cancer

Laufey Amundadottir[1,2,55*], Peter Kraft[3,4,55], Rachael Z Stolzenberg-Solomon[2,55], Charles S Fuchs[5,6,55], Gloria M Petersen[7], Alan A Arslan[8–10], H Bas Bueno-de-Mesquita[11], Myron Gross[12], Kathy Helzlsouer[13], Eric J Jacobs[14], Andrea LaCroix[15], Wei Zheng[16], Demetrius Albanes[2], William Bamlet[7], Christine D Berg[17], Franco Berrino[18], Sheila Bingham[19], Julie E Buring[20,21], Paige M Bracci[22], Federico Canzian[23], Françoise Clavel-Chapelon[24], Sandra Clipp[25], Michelle Cotterchio[26], Mariza de Andrade[7], Eric J Duell[27], John W Fox Jr[28], Steven Gallinger[29], J Michael Gaziano[30], Edward L Giovannucci[3,6,31], Michael Goggins[32], Carlos A González[33], Göran Hallmans[34], Susan E Hankinson[3,6], Manal Hassan[35], Elizabeth A Holly[22], David J Hunter[3,6], Amy Hutchinson[2,36], Rebecca Jackson[37], Kevin B Jacobs[2,36,38], Mazda Jenab[27], Rudolf Kaaks[23], Alison P Klein[39,40], Charles Kooperberg[15], Robert C Kurtz[41], Donghui Li[35], Shannon M Lynch[42], Margaret Mandelson[15,43], Robert R McWilliams[44], Julie B Mendelsohn[2], Dominique S Michaud[3,45], Sara H Olson[46], Kim Overvad[47], Alpa V Patel[14], Petra H M Peeters[45,48], Aleksandar Rajkovic[49], Elio Riboli[45], Harvey A Risch[50], Xiao-Ou Shu[16], Gilles Thomas[2], Geoffrey S Tobias[2], Dimitrios Trichopoulos[3,51], Stephen K Van Den Eeden[52], Jarmo Virtamo[53], Jean Wactawski-Wende[54], Brian M Wolpin[5,6], Herbert Yu[50], Kai Yu[2], Anne Zeleniuch-Jacquotte[9,10], Stephen J Chanock[1,2,55], Patricia Hartge[2,55] & Robert N Hoover[2,55]

We conducted a two-stage genome-wide association study of pancreatic cancer, a cancer with one of the lowest survival rates worldwide. We genotyped 558,542 SNPs in 1,896 individuals with pancreatic cancer and 1,939 controls drawn from 12 prospective cohorts plus one hospital-based case-control study. We conducted a combined analysis of these groups plus an additional 2,457 affected individuals and 2,654 controls from eight case-control studies, adjusting for study, sex, ancestry and five principal components. We identified an association between a locus on 9q34 and pancreatic cancer marked by the SNP rs505922 (combined $P = 5.37 \times 10^{-8}$; multiplicative per-allele odds ratio 1.20; 95% confidence interval 1.12–1.28). This SNP maps to the first intron of the *ABO* blood group gene. Our results are consistent with earlier epidemiologic evidence suggesting that people with blood group O may have a lower risk of pancreatic cancer than those with groups A or B.

Pancreatic cancer has one of the highest mortality rates of any cancer, with an estimated 5-year relative survival rate of <5% (refs. 1,2). Currently, there is not an effective screening test for this malignancy, and by the time of initial diagnosis, metastatic disease is commonly present. Established risk factors include a family history of pancreatic cancer, a medical history of type 2 diabetes and cigarette smoking[3]. Studies have also suggested an increased risk of pancreatic cancer within families with hereditary pancreatitis[4,5]. In addition, it has been estimated that a small proportion of pancreatic cancers are due to highly penetrant germline mutations[6]. Prior studies have suggested a genetic contribution to pancreatic cancer, but there has been limited success in replicating common variants reported to be associated with this disease. Here we report a genome-wide association study (GWAS) to identify common variants associated with pancreatic cancer.

We conducted a GWAS in 1,896 individuals with pancreatic cancer and 1,939 controls drawn from 12 prospective cohorts (the American Cancer Society Cancer Prevention Study II (ref. 7), the Alpha-Tocopherol, Beta-Carotene Cancer Prevention Study[8]; the European Prospective Investigation into Cancer and Nutrition[9]; CLUE II (ref. 10); the Health Professionals Follow-up Study[11]; the New York University Women's Health Study[12]; the Nurses' Health Study[11]; the Physicians' Health Study I (ref. 11); the Prostate, Lung, Colorectal and Ovarian Cancer Screening Trial[13]; the Shanghai Men's and Women's Health Study[14,15]; the Women's Health Initiative[16] and

*A full list of author affiliations appears at the end of the paper.

Received 25 March; accepted 8 July; published online 2 August 2009; doi:10.1038/ng.429

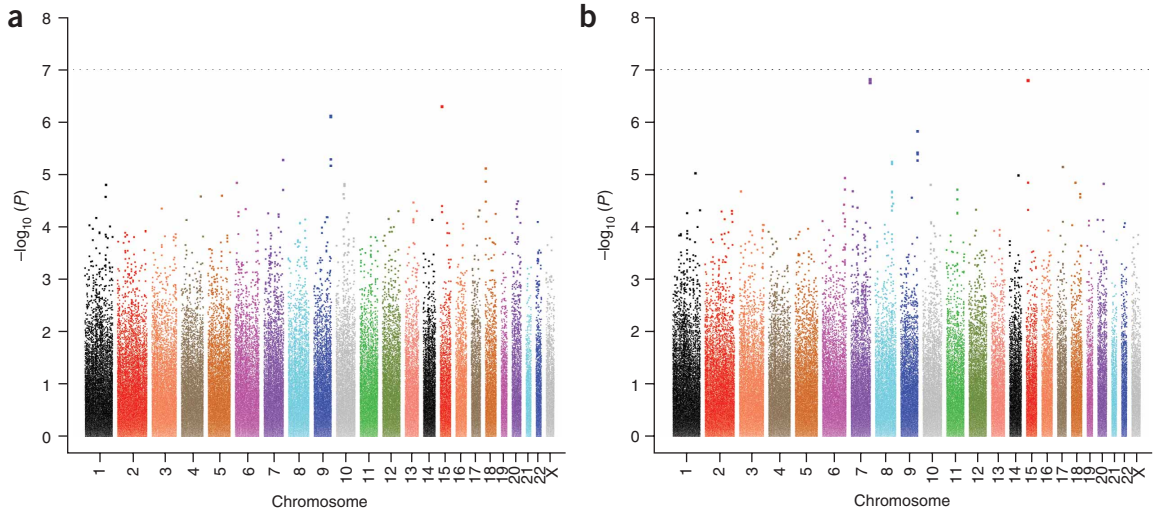

Figure 1 Manhattan plot of the *P* values in the pancreatic cancer GWAS. (**a**) Association with pancreatic cancer for the entire GWAS (12 cohort studies and the Mayo case-control study; Online Methods). (**b**) Results of the GWAS including only the 12 cohorts studies. Association was assessed using unconditional logistic regression adjusted for study, arm, age, sex, ancestry and the top five principal components of the population stratification analysis. The *x* axis represents chromosomal locations and the *y* axis shows *P* values on a logarithmic scale.

the Women's Health Study[17]; **Supplementary Table 1**) plus one hospital-based case-control study (the Mayo Clinic Molecular Epidemiology of Pancreatic Cancer Study[18]). Eight case-control studies participated in the independent rapid follow-up, known as a fast-track replication phase, of 2,457 cases and 2,654 controls (the University of Toronto[19], University of California San Francisco[20], the Johns Hopkins University, M.D. Anderson Cancer Center[21], PACIFIC Study of Group Health and Northern California Kaiser Permanente, Memorial Sloan-Kettering Cancer Center[22], Yale University[23], and the Mayo Clinic Molecular Epidemiology of Pancreatic Cancer Study[18]; **Supplementary Table 2**).

After quality control of genotypes using the HumanHap500 chip, 558,542 SNPs were available for analysis. We fit a logistic regression model for genotype trend effects (1 degree of freedom (d.f.)), adjusting for study, age, sex, ancestry and the top five principal components of population stratification (Online Methods). The quantile-quantile plot did not demonstrate a systematic deviation from the expected distribution, minimizing the likelihood of systematic genotype error or bias due to underlying population substructure (**Supplementary Fig. 1**). The results of the GWAS are shown in **Figure 1a**. Because of the potential for survivor bias in case-control studies owing to rapid mortality, we also analyzed the GWAS for the cohort studies alone (that is, excluding the Mayo participants) (**Fig. 1b**).

We conducted a fast track replication of SNPs from three regions (9q34, 7q36 and a gene desert on 15q14) in eight case-control studies (four hospital-based and four population-based). At least two SNPs per region ranked among the lowest 25 *P* values in the initial GWAS: (i) rs505922, rs495828, rs657152 and rs630014 (ranked 2, 3, 8 and 17) on 9q34, which includes the *ABO* gene, (ii) rs167020, rs172310 and rs288746 (ranked 6, 10 and 89) on 7q36, which includes *SHH* (sonic hedgehog homolog), and (iii) rs8028529, rs4130461 and rs4459505 (ranked 1, 18 and 26) in the gene desert on 15q14 (**Table 1**).

In a combined analysis of individuals of European background[24], the strongest association with pancreatic cancer below the threshold for genome-wide significance[25] was for a locus marked by rs505922 on chromosome 9q34, located within the first intron of *ABO*, a well-described blood group gene ($P = 5.37 \times 10^{-8}$, trend model; heterozygous odds ratio (OR_{Het}) = 1.20, 95% confidence interval (c.i.) 1.12–1.28; homozygous odds ratio (OR_{Hom}) = 1.44, 95% c.i. 1.26–1.63). We observed a comparable result when we included all ethnic groups in stage 1 ($P = 2.61 \times 10^{-8}$; **Supplementary Table 3**). In the case-control replication set, we genotyped a second SNP, rs687621 ($r^2 = 1$ with rs505922 in HapMap CEU and $r^2 = 0.91$ in stage 2 controls), located 12 kb centromeric in intron 2; this confirmed the signal at the locus ($P = 1.57 \times 10^{-4}$ in the stage 2 case-control studies only). In the combined analysis, we observed a comparably strong signal for rs630014 ($P = 1.58 \times 10^{-7}$; $OR_{Het} = 0.85$, $OR_{Hom} = 0.71$), which resides within 500 bp of rs505922 and is in linkage disequilibrium (LD) ($r^2 = 0.52$ in HapMap CEU and 0.40 in PanScan GWAS European controls). After adjusting for rs505922, none of the remaining SNPs in *ABO* was significant at $P < 0.01$. The SNPs reside in a haplotype block that encompasses the proximal promoter and introns 1 and 2 (**Fig. 2**).

Blood groups were first described by Karl Landsteiner in 1900, but the structure of the ABO antigens and their biosynthesis remained elusive until after 1950. The *ABO* gene encodes a glycosyltransferase that catalyzes the transfer of carbohydrates to the H antigen, forming the antigenic structure of the ABO blood groups. The proteins encoded by the A and B alleles of *ABO* differ minimally in amino acid sequence but catalyze the transfer of different carbohydrates (N-acetylgalactosamine or galactose) onto the H antigen to form the A or B antigens. Individuals with the O blood group do not produce either the A or B antigens because of a single-base deletion.

Our findings are notable because multiple studies, mainly from the 1950s and 1960s, reported an association between ABO blood type and gastrointestinal cancers, most strongly for gastric cancer but also for pancreatic cancer[26,27]. The protective allele T for rs505922 is in complete LD ($r^2 = 1.0$) with the O allele of the *ABO* locus, consistent with earlier reports showing increased risk of gastric and pancreatic cancer for individuals of the A and B blood groups. It is plausible that the single-base deletion that generates the O blood group underlies the association signal, but further mapping and laboratory work will be required to determine which variant(s) account for the observed association.

Genetic variation in the first intron of the *ABO* gene has also been associated with circulating levels of serum tumor necrosis factor alpha

(TNFα)[28], circulating levels of soluble intracellular adhesion molecule 1 (sICAM-1)[29] and plasma levels of alkaline phosphatase[30]. Although higher TNFα levels are associated with the common allele of rs505922, which is protective for pancreatic cancer in our study, the data concerning the relationship between blood groups and TNFα levels are inconsistent[28]. Furthermore, this region could be important for regulating circulating levels of sICAM-1, as rs507666 and rs505922 (located 170 bp apart) were recently reported to be associated with the amount of circulating ICAM-1 (ref. 29). In addition, SNPs in the *ABO* locus, including rs657152, have been associated with plasma levels of liver-derived alkaline phosphatase[30]. Last, ABO antigen expression is altered in primary and metastatic pancreatic cancers compared with normal pancreatic tissues[31].

For rapidly fatal conditions, case-control studies are prone to distortion because they include survivors disproportionately. For variants unrelated to survival, case-control data are suitable for discovery and replication of risk-related markers. However, for variants related to survival, case-control studies yield biased estimates of

Table 1 Association of SNPs on chromosomes 9q34, 7q36 and 15q14 with risk of pancreatic cancer

Marker[a], alleles[b], chromosome[c], location[c] and gene[d]	Subset[e]	MAF		Number of participants		χ^{2f}	*P* value[f]	OR_{Het} (95% c.i.)	OR_{Hom} (95% c.i.)[g]
		Controls	Cases	Controls	Cases				
rs505922 (T, C)	Stage 1 (cohorts)	0.357	0.417	1,462	1,406	21.11	4.33×10^{-6}	1.29 (1.16–1.43)	1.66 (1.33–2.05)
9q34	Stage 1 (all)	0.357	0.411	1,805	1,771	22.18	2.48×10^{-6}	1.26 (1.14–1.39)	1.59 (1.31–1.92)
135139050	Stage 2	0.343	0.375	2,127	2,120	9.50	2.06×10^{-3}	1.15 (1.05–1.26)	1.32 (1.11–1.58)
ABO	Stage 1 + 2	0.349	0.392	3,932	3,891	29.58	5.37×10^{-8}	1.20 (1.12–1.28)	1.44 (1.26–1.63)
rs495828 (G, T)	Stage 1 (cohorts)	0.192	0.236	1,423	1,362	18.11	2.08×10^{-5}	1.35 (1.18–1.55)	1.82 (1.38–2.41)
9q34	Stage 1 (all)	0.194	0.236	1,755	1,717	21.37	3.78×10^{-6}	1.34 (1.18–1.51)	1.79 (1.40–2.29)
135144688	Stage 2	0.223	0.238	1,786	1,718	2.10	1.47×10^{-1}	1.09 (0.97–1.21)	1.18 (0.94–1.47)
ABO	Stage 1 + 2	0.209	0.237	3,541	3,435	17.93	2.30×10^{-5}	1.19 (1.10–1.30)	1.43 (1.21–1.68)
rs657152 (G, T)	Stage 1 (cohorts)	0.380	0.437	1,463	1,408	18.05	2.15×10^{-5}	1.26 (1.13–1.40)	1.59 (1.28–1.97)
9q34	Stage 1 (all)	0.380	0.430	1,806	1,773	18.13	2.06×10^{-5}	1.23 (1.12–1.35)	1.51 (1.25–1.83)
135129086	Stage 2	0.374	0.404	1,791	1,729	7.24	7.13×10^{-3}	1.14 (1.04–1.26)	1.30 (1.07–1.58)
ABO	Stage 1 + 2	0.377	0.417	3,597	3,502	24.29	8.28×10^{-7}	1.19 (1.11–1.27)	1.41 (1.23–1.61)
rs630014 (C, T)	Stage 1 (cohorts)	0.475	0.421	1,463	1,408	18.04	2.16×10^{-5}	0.80 (0.72–0.88)	0.63 (0.51–0.78)
9q34	Stage 1 (all)	0.473	0.427	1,805	1,773	16.74	4.28×10^{-5}	0.82 (0.75–0.90)	0.67 (0.56–0.81)
135139543	Stage 2	0.479	0.441	2,196	2,118	11.83	5.84×10^{-4}	0.86 (0.79–0.94)	0.74 (0.63–0.88)
ABO	Stage 1 + 2	0.477	0.435	4,001	3,891	27.49	1.58×10^{-7}	0.85 (0.79–0.90)	0.71 (0.63–0.81)
rs167020 (G, A)	Stage 1 (cohorts)	0.250	0.313	1,462	1,408	27.28	1.76×10^{-7}	1.37 (1.22–1.54)	1.88 (1.48–2.38)
7q36	Stage 1 (all)	0.259	0.307	1,805	1,773	20.06	7.52×10^{-6}	1.27 (1.15–1.41)	1.62 (1.31–2.00)
155312494	Stage 2	0.278	0.294	1,802	1,734	2.39	1.22×10^{-1}	1.09 (0.98–1.20)	1.18 (0.96–1.45)
SHH	Stage 1 + 2	0.269	0.301	3,607	3,507	18.12	2.07×10^{-5}	1.17 (1.09–1.26)	1.38 (1.19–1.60)
rs172310 (C, A)	Stage 1 (cohorts)	0.272	0.336	1,454	1,399	27.02	2.01×10^{-7}	1.36 (1.21–1.53)	1.85 (1.47–2.34)
7q36	Stage 1 (all)	0.282	0.329	1,796	1,763	17.43	2.98×10^{-5}	1.25 (1.12–1.38)	1.56 (1.26–1.92)
155308388	Stage 2	0.305	0.323	1,768	1,699	2.80	9.45×10^{-2}	1.09 (0.99–1.21)	1.19 (0.97–1.46)
SHH	Stage 1 + 2	0.293	0.326	3,564	3,462	17.04	3.66×10^{-5}	1.17 (1.08–1.25)	1.36 (1.17–1.57)
rs288746 (T, C)	Stage 1 (cohorts)	0.109	0.144	1,458	1,403	14.57	1.35×10^{-4}	1.37 (1.16–1.61)	1.87 (1.36–2.59)
7q36	Stage 1 (all)	0.114	0.138	1,800	1,768	8.08	4.48×10^{-3}	1.23 (1.07–1.42)	1.52 (1.14–2.02)
155299433	Stage 2	0.116	0.128	1,805	1,735	2.59	1.08×10^{-1}	1.12 (0.97–1.30)	1.26 (0.95–1.68)
SHH	Stage 1 + 2	0.115	0.133	3,605	3,503	10.14	1.45×10^{-3}	1.18 (1.07–1.30)	1.39 (1.13–1.70)
rs8028529 (T, C)	Stage 1 (cohorts)	0.198	0.255	1,457	1,404	25.92	3.55×10^{-7}	1.38 (1.22–1.56)	1.91 (1.49–2.45)
15q14	Stage 1 (all)	0.202	0.249	1,800	1,768	23.13	1.51×10^{-6}	1.31 (1.17–1.47)	1.72 (1.38–2.15)
34441889	Stage 2	0.231	0.229	1,800	1,736	0.02	8.92×10^{-1}	0.99 (0.89–1.11)	0.98 (0.79–1.23)
None	Stage 1 + 2	0.217	0.239	3,600	3,504	11.12	8.53×10^{-4}	1.14 (1.06–1.24)	1.31 (1.12–1.53)
rs4130461 (G, T)	Stage 1 (cohorts)	0.224	0.273	1,463	1,408	18.71	1.53×10^{-5}	1.31 (1.16–1.47)	1.70 (1.34–2.17)
15q14	Stage 1 (all)	0.231	0.272	1,806	1,773	16.64	4.52×10^{-5}	1.25 (1.12–1.39)	1.56 (1.26–1.94)
34439130	Stage 2	0.256	0.250	1,802	1,736	0.39	5.32×10^{-1}	0.97 (0.87–1.08)	0.93 (0.75–1.16)
None	Stage 1 + 2	0.243	0.261	3,608	3,509	6.15	1.32×10^{-2}	1.10 (1.02–1.19)	1.21 (1.04–1.41)
rs4459505 (G, A)	Stage 1 (cohorts)	0.177	0.218	1,455	1,402	15.51	8.21×10^{-5}	1.30 (1.14–1.49)	1.70 (1.30–2.21)
15q14	Stage 1 (all)	0.178	0.214	1,796	1,765	14.92	1.12×10^{-4}	1.26 (1.12–1.42)	1.59 (1.26–2.01)
34443314	Stage 2	0.196	0.198	1,803	1,737	0.08	7.81×10^{-1}	1.02 (0.90–1.14)	1.03 (0.82–1.31)
None	Stage 1 + 2	0.187	0.206	3,599	3,502	8.52	3.51×10^{-3}	1.13 (1.04–1.23)	1.28 (1.08–1.51)

Results from the unconditional logistic regression of the genotypes generated in the initial GWAS and the follow-up studies in a total of 3,891 individuals with pancreatic cancer and 4,001 controls. The analysis was adjusted for age in 10-year categories, sex, study, arm, ancestry and five principal components of population stratification. OR, odds ratio; Het, heterozygous; Hom, homozygous for minor allele.
[a]NCBI dbSNP identifier. [b]Major allele, minor allele. [c]Chromosome and NCBI Human Genome Build 36 location. [d]Gene neighborhood within 20 kb upstream and 10 kb downstream of SNP. [e]Stage 1: initial GWAS; stage 2: replication. [f]1-d.f. score test. [g]Estimate assuming multiplicative odds model.

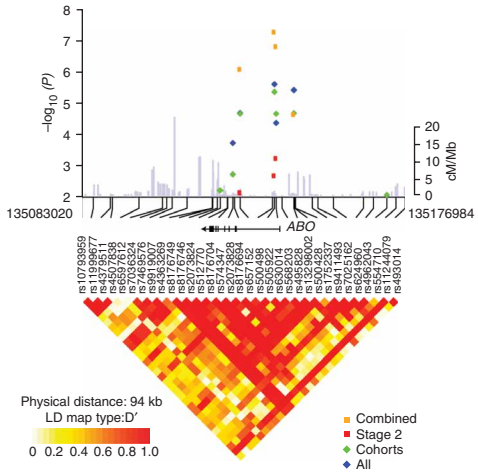

Figure 2 Association and linkage disequilibrium plot of the 9q34 locus. Association results are shown for all GWAS studies (blue diamonds), GWAS cohorts (green diamonds), replication studies (red circles) and all studies combined (yellow circles). Overlaid on the association panel is a plot of estimated recombination rates (cM/Mb) across the region from HapMap Phase II. The LD plot shows a region of chromosome 9 marked by SNPs, rs505922 and rs630014 ($r^2 = 0.52$ in HapMap CEU and 0.40 in PanScan European control individuals) and bounded by SNPs between chromosome 9 135083020 and 135176984 (NCBI Human Genome Build 36). Linkage disequilibrium (LD) is depicted for SNPs with mean allele frequency (MAF) >5% within PanScan. rs505922 and rs630014 are located in the first intron of the *ABO* gene, shown above the LD plot. Only SNPs genotyped in both the GWAS and 'fast track' replication are shown.

the association with pancreatic cancer risk. *ABO* variants seem to be unrelated to survival and show strong, similar signals in both cohort and case-control data.

We observed a genome-wide association with *SHH* ($P = 1.76 \times 10^{-7}$) among cohorts that was not replicated in the follow-up in case-control studies ($P = 0.12$), raising three possibilities: that the cohort finding is due to chance, that *SHH* is related to both survival and to risk or that the SNPs failed to replicate because of chance (**Table 1**). Because there is substantial evidence that *SHH* has a role in pancreatic carcinogenesis, further work is required to investigate this region[32].

Pancreatic cancer is among the deadliest of cancers, with mortality rates approaching incidence rates[1]. As there are few known risk factors, improved diagnostics and a finer understanding of the molecular pathogenesis are urgently needed. Our findings have identified the contribution of genetic variation in the *ABO* locus of 9q34 to pancreatic carcinogenesis, a finding that supports an epidemiologic observation first made a half-century ago and recently confirmed[33]. We are now conducting a GWAS in the eight studies from stage 2 of this study and anticipate that this will bring the identification of additional loci associated to pancreatic cancer. The discovery of additional genetic risk variants for this highly lethal cancer could contribute to improvements in risk stratification, prevention, early detection and therapeutic approaches to pancreatic cancer.

METHODS

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

Data access. The Cancer Genetic Markers of Susceptibility (CGEMS) data portal provides access to data for 558,542 SNPs in 3,835 individuals. Investigators from certified scientific institutions may access the portal after approval of their submitted Data Access Request.

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

ACKNOWLEDGMENTS

The authors gratefully acknowledge the energy and contribution of our late colleague, Robert Welch. Additional acknowledgments are found in the **Supplementary Note**.

AUTHOR CONTRIBUTIONS

L.A., P.K., R.Z.S.-S., C.S.F., G.M.P., K.B.J., S.M.L., J.B.M., G.S.T., S.J.C., P.H. and R.N.H. organized and designed the study. L.A., A.H., K.B.J., G.T. and S.J.C. supervised genotyping of samples. L.A., P.K., R.Z.S.-S., C.S.F., K.B.J., C.K., K.Y., S.J.C., P.H. and R.N.H. contributed to the design and execution of statistical analysis. LA., S.J.C., P.H. and R.N.H. wrote the first draft of the manuscript. R.Z.S.-S., C.S.F., G.M.P., A.A.A., H.B.B.-d.-M., M.G., K.H., E.J.J., A.L., W.Z., D.A., W.B., C.D.B., F.B., S.B., J.E.B., P.M.B., F.C., F.C.-C., S.C., M.C., M.d.A., E.J.D., J.W.F., S.G., J.M.G., E.L.G., M.G., C.A.G., G.H., S.E.H., M.H., E.A.H., D.J.H., R.J., M.J., R.K., A.P.K., C.K., R.C.K., D.L., M.M., R.R.M., D.S.M., S.H.O., K.O., A.V.P., P.H.M.P., A.R., E.R., H.A.R., X.-O.S., D.T., S.K.V.D.E., J.V., J.W.-W., B.M.W., H.Y. and A.Z.-J. conducted the epidemiologic studies and contributed samples to the PanScan GWAS and/or replication. All authors contributed to the writing of the manuscript.

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

1. Jemal, A. *et al.* Cancer statistics, 2008. *CA Cancer J. Clin.* **58**, 71–96 (2008).
2. Ferlay, J., Bray, F., Pisani, P. & Parkin, D.M. GLOBOCAN 2002: Cancer Incidence, Mortality and Prevalence Worldwide. *IARC CancerBase* **No. 5**, IARC*Press*, Lyon (2004).
3. Anderson, K.E., Mack, T. & Silverman, D. Cancer of the pancreas. in *Cancer Epidemiology and Prevention* (ed. Schottenfeld, D. & Fraumeni, J.J.) (Oxford Univ. Press, New York, 2006).
4. Lowenfels, A.B. *et al.* Hereditary pancreatitis and the risk of pancreatic cancer. International Hereditary Pancreatitis Study Group. *J. Natl. Cancer Inst.* **89**, 442–446 (1997).
5. Castleman, B. Case records of the Massachusetts General Hospital. *N. Engl. J. Med.* **286**, 1353–1359 (1972).
6. Klein, A.P. *et al.* Prospective risk of pancreatic cancer in familial pancreatic cancer kindreds. *Cancer Res.* **64**, 2634–2638 (2004).
7. Calle, E.E. *et al.* The American Cancer Society Cancer Prevention Study II Nutrition Cohort: rationale, study design, and baseline characteristics. *Cancer* **94**, 2490–2501 (2002).
8. The ATBC Cancer Prevention Study Group. The alpha-tocopherol, beta-carotene lung cancer prevention study: design, methods, participant characteristics, and compliance. *Ann. Epidemiol.* **4**, 1–10 (1994).
9. Riboli, E. *et al.* European Prospective Investigation into Cancer and Nutrition (EPIC): study populations and data collection. *Public Health Nutr.* **5**, 1113–1124 (2002).
10. Gallicchio, L. *et al.* Single nucleotide polymorphisms in inflammation-related genes and mortality in a community-based cohort in Washington County, Maryland. *Am. J. Epidemiol.* **167**, 807–813 (2008).
11. Wolpin, B.M. *et al.* Circulating insulin-like growth factor binding protein-1 and the risk of pancreatic cancer. *Cancer Res.* **67**, 7923–7928 (2007).
12. Zeleniuch-Jacquotte, A. *et al.* Postmenopausal levels of sex hormones and risk of breast carcinoma in situ: results of a prospective study. *Int. J. Cancer* **114**, 323–327 (2005).
13. Hayes, R.B. *et al.* Methods for etiologic and early marker investigations in the PLCO trial. *Mutat. Res.* **592**, 147–154 (2005).
14. Xu, W.H. *et al.* Joint effect of cigarette smoking and alcohol consumption on mortality. *Prev. Med.* **45**, 313–319 (2007).
15. Zheng, W. *et al.* The Shanghai Women's Health Study: rationale, study design, and baseline characteristics. *Am. J. Epidemiol.* **162**, 1123–1131 (2005).
16. Anderson, G.L. *et al.* Implementation of the Women's Health Initiative study design. *Ann. Epidemiol.* **13**, S5–S17 (2003).
17. Rexrode, K.M., Lee, I.M., Cook, N.R., Hennekens, C.H. & Buring, J.E. Baseline characteristics of participants in the Women's Health Study. *J. Womens Health Gend. Based Med.* **9**, 19–27 (2000).
18. McWilliams, R.R. *et al.* Polymorphisms in DNA repair genes, smoking, and pancreatic adenocarcinoma risk. *Cancer Res.* **68**, 4928–4935 (2008).
19. Eppel, A., Cotterchio, M. & Gallinger, S. Allergies are associated with reduced pancreas cancer risk: A population-based case-control study in Ontario, Canada. *Int. J. Cancer* **121**, 2241–2245 (2007).
20. Duell, E.J. *et al.* Detecting pathway-based gene-gene and gene-environment interactions in pancreatic cancer. *Cancer Epidemiol. Biomarkers Prev.* **17**, 1470–1479 (2008).

21. Hassan, M.M. *et al.* Risk factors for pancreatic cancer: case-control study. *Am. J. Gastroenterol.* **102**, 2696–2707 (2007).
22. Olson, S.H. *et al.* Allergies, variants in IL-4 and IL-4R alpha genes, and risk of pancreatic cancer. *Cancer Detect. Prev.* **31**, 345–351 (2007).
23. Risch, H.A. Etiology of pancreatic cancer, with a hypothesis concerning the role of N-nitroso compounds and excess gastric acidity. *J. Natl. Cancer Inst.* **95**, 948–960 (2003).
24. Skol, A.D., Scott, L.J., Abecasis, G.R. & Boehnke, M. Joint analysis is more efficient than replication-based analysis for two-stage genome-wide association studies. *Nat. Genet.* **38**, 209–213 (2006).
25. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
26. Aird, I., Bentall, H.H. & Roberts, J.A. A relationship between cancer of stomach and the ABO blood groups. *BMJ* **1**, 799–801 (1953).
27. Marcus, D.M. The ABO and Lewis blood-group system. Immunochemistry, genetics and relation to human disease. *N. Engl. J. Med.* **280**, 994–1006 (1969).
28. Melzer, D. *et al.* A genome-wide association study identifies protein quantitative trait loci (pQTLs). *PLoS Genet.* **4**, e1000072 (2008).
29. Paré, G. *et al.* Novel association of ABO histo-blood group antigen with soluble ICAM-1: results of a genome-wide association study of 6,578 women. *PLoS Genet.* **4**, e1000118 (2008).
30. Yuan, X. *et al.* Population-based genome-wide association studies reveal six loci influencing plasma levels of liver enzymes. *Am. J. Hum. Genet.* **83**, 520–528 (2008).
31. Itzkowitz, S.H. *et al.* Cancer-associated alterations of blood group antigen expression in the human pancreas. *J. Natl. Cancer Inst.* **79**, 425–434 (1987).
32. Berman, D.M. *et al.* Widespread requirement for Hedgehog ligand stimulation in growth of digestive tract tumours. *Nature* **425**, 846–851 (2003).
33. Wolpin, B.M. *et al.* ABO blood group and the risk of pancreatic cancer. *J. Natl. Cancer Inst.* **101**, 424–431 (2009).

[1]Laboratory of Translational Genomics, Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, USA. [2]Division of Cancer Epidemiology and Genetics, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, USA. [3]Department of Epidemiology and [4]Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts, USA. [5]Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, Massachusetts, USA. [6]Channing Laboratory, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [7]Department of Health Sciences Research, College of Medicine, Mayo Clinic, Rochester, Minnesota, USA. [8]Department of Obstetrics and Gynecology and [9]Department of Environmental Medicine, New York University School of Medicine, New York, New York, USA. [10]New York University Cancer Institute, New York, New York, USA. [11]National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands, and Department of Gastroenterology and Hepatology, University Medical Center Utrecht, Utrecht, The Netherlands. [12]Department of Laboratory Medicine/Pathology, School of Medicine, University of Minnesota, Minneapolis, Minnesota, USA. [13]Prevention and Research Center, Mercy Medical Center, Baltimore, Maryland, USA. [14]Department of Epidemiology and Surveillance Research, American Cancer Society, Atlanta, Georgia, USA. [15]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, Seattle, Washington, USA. [16]Department of Medicine and Vanderbilt-Ingram Cancer Center, Vanderbilt University, Nashville, Tennessee, USA. [17]Division of Cancer Prevention, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, USA. [18]Etiological Epidemiology and Prevention Unit, Fondazione IRCCS Istituto Nazionale dei Tumori di Milano, Milan, Italy. [19]MRC Dunn Human Nutrition Unit, University of Cambridge, Cambridge, UK. [20]Divisions of Preventive Medicine and Aging, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA. [21]Department of Ambulatory Care and Prevention, Harvard Medical School, Boston, Massachusetts, USA. [22]Department of Epidemiology & Biostatistics, University of California San Francisco, San Francisco, California, USA. [23]Division of Cancer Epidemiology, German Cancer Research Center (DKFZ), Heidelberg, Germany. [24]Institut National de la Santé et de la Recherche Médicale (INSERM) and Institut Gustave Roussy, Villejuif, France. [25]The Johns Hopkins Bloomberg School of Public Health, George W. Comstock Center for Public Health Research and Prevention, Hagerstown, Maryland, USA. [26]Cancer Care Ontario and Dalla Lana School of Public Health, University of Toronto, Ontario, Canada. [27]International Agency for Research on Cancer, Lyon, France. [28]College of Human Medicine, Michigan State University, East Lansing, Michigan, USA. [29]Samuel Lunenfeld Research Institute, Mount Sinai Hospital, Toronto, Ontario, Canada. [30]Physicians' Health Study, Divisions of Aging, Cardiovascular Medicine and Preventive Medicine, Department of Medicine, Brigham and Women's Hospital and Harvard Medical School, Boston, Massachusetts, USA, and Massachusetts Veterans Epidemiology Research and Information Center, Veterans Affairs Boston Healthcare System, Boston, Massachusetts, USA. [31]Department of Nutrition, Harvard School of Public Health, Boston, Massachusetts, USA. [32]Departments of Oncology, Pathology and Medicine, The Sol Goldman Pancreatic Research Center, The Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [33]Unit of Nutrition, Environment and Cancer, Cancer Epidemiology Research Programme, Catalan Institute of Oncology (ICO), Barcelona, Spain. [34]Department of Public Health and Clinical Medicine, Nutritional Research, Umeå University, Umeå, Sweden. [35]Department of Gastrointestinal Medical Oncology, University of Texas M.D. Anderson Cancer Center, Houston, Texas, USA. [36]Core Genotyping Facility, Advanced Technology Program, SAIC-Frederick, Inc., NCI-Frederick, Frederick, Maryland, USA. [37]Division of Endocrinology, Diabetes and Metabolism, Department of Internal Medicine and Center for Clinical and Translational Science, Ohio State University, Columbus, Ohio, USA. [38]Bioinformed, LLC, Gaithersburg, Maryland, USA. [39]Department of Oncology, The Johns Hopkins University School of Medicine, Baltimore, Maryland, USA. [40]Department of Epidemiology, The Johns Hopkins Bloomberg School of Public Health, The Sol Goldman Pancreatic Research Center, The Johns Hopkins Medical Institutions, Baltimore, Maryland, USA. [41]Department of Medicine, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [42]Epidemiology and Genetics Research Program, Division of Cancer Control and Population Sciences, National Cancer Institute, National Institutes of Health, Department of Health and Human Services, Bethesda, Maryland, USA. [43]Group Health Center for Health Studies, Seattle, Washington, USA. [44]Department of Oncology, College of Medicine, Mayo Clinic, Rochester, Minnesota, USA. [45]Division of Epidemiology, Public Health and Primary Care, Imperial College London, London, UK. [46]Department of Epidemiology and Biostatistics, Memorial Sloan-Kettering Cancer Center, New York, New York, USA. [47]Department of Cardiology and Department of Clinical Epidemiology, Aalborg Hospital, Aarhus University Hospital, Aalborg, Denmark. [48]Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, The Netherlands. [49]Department of Obstetrics and Gynecology, Baylor College of Medicine, Houston, Texas, USA. [50]Yale University School of Public Health, New Haven, Connecticut, USA. [51]Department of Hygiene and Epidemiology, University of Athens Medical School, Athens, Greece. [52]Division of Research, Kaiser Permanente, Northern California Region, Oakland, California, USA. [53]Department of Chronic Disease Prevention, National Institute for Health and Welfare, Helsinki, Finland. [54]Department of Social and Preventive Medicine, University at Buffalo, State University of New York, Buffalo, New York, USA. [55]These authors contributed equally to this work. Correspondence should be addressed to S.J.C. (chanocks@mail.nih.gov).

ONLINE METHODS

Study participants. Participants in stage 1 of the GWAS were drawn from 12 cohort studies and one case-control study (**Supplementary Table 1**) in the Pancreatic Cancer Cohort Consortium Genome-Wide Association Study (PanScan1) and are part of a larger international consortium, the National Cancer Institute–sponsored Cohort Consortium. They include the Alpha-Tocopherol Beta-Carotene Cancer Prevention Study (ATBC)[8], CLUE II (ref. 10), the American Cancer Society Cancer Prevention Study II (CPS-II)[7], the European Prospective Investigation into Cancer and Nutrition (EPIC, comprising cohorts from Denmark, France, Germany, the UK, Greece, Italy, the Netherlands, Spain and Sweden)[9], the Health Professionals Follow-up Study (HPFS)[11], Nurses' Health Study (NHS)[11], the New York University Women's Health Study (NYUWHS)[12], the Physicians' Health Study I (PHS I)[11], the Prostate, Lung, Colorectal, and Ovarian Cancer Screening Trial (PLCO)[13], the Shanghai Men's and Women's Health Study (SMWHS)[14,15], the Women's Health Initiative (WHI)[16], and the Women's Health Study (WHS)[17]. Each cohort that participated in PanScan had a defined population from whom blood or buccal cells were collected before the diagnosis of pancreatic cancer. Incident primary pancreatic adenocarcinoma cases were identified by self-report with subsequent medical record review, linkage with a cancer registry or both. Cases were defined as primary adenocarcinoma of the exocrine pancreas (ICD-O-3 code C250–C259). Non-exocrine pancreatic tumors (histology type 8150, 8151, 8153, 8155 and 8240) were excluded.

We identified 1,770 incident cases among the cohorts as part of a nested case-control study. We selected an equal number of controls within their respective cohort. One control was matched per case based on calendar year of birth (±5 years), sex, broad category of race and ethnicity and DNA (blood or buccal cell). Each control was alive and free of pancreatic cancer on the calendar date that his or her matching case was diagnosed with pancreatic cancer. The NHS, HPFS, WHS and PHS cohorts also matched cases and controls based on smoking status. Four hundred individuals with pancreatic adenocarcinoma and 400 controls were included from the Mayo Clinic Molecular Epidemiology of Pancreatic Cancer Study[18]. The Molecular Epidemiology of Pancreatic Cancer study was initiated in 2000 and used an 'ultrarapid' case ascertainment system in which >95% of patients at the Mayo Clinic from Minnesota, Iowa, and Wisconsin who were suspected to have pancreatic cancer were approached. Among those with pancreatic cancer, 72% provided consent and a blood sample. Clinic controls were drawn from patients seeking general medical care and were frequency matched to cases on age, race, gender, and area of residence.

Eight case-control studies from the PANC4 consortium participated in a replication of promising SNPs from the initial scan: University of Toronto[19], University of California San Francisco[20], Johns Hopkins University, M.D. Anderson Cancer Center[21], PACIFIC Study of Group Health and Northern California Kaiser Permanente, Memorial Sloan-Kettering Cancer Center[22] and Yale University[23] and distinct cases and controls from the Mayo Clinic Molecular Epidemiology of Pancreatic Cancer Study[18] (**Supplementary Table 2**).

Each participating study obtained informed consent from study participants and approval from its Institutional Review Board. Each cohort study and the Mayo Clinic case-control study obtained Institutional Review Board certification, permitting data sharing in accordance with the US National Institutes of Health (NIH) Policy for Sharing of Data Obtained in NIH-Supported or -Conducted Genome-Wide Association Studies.

Genotyping and quality control. We selected 4,063 DNA samples (including 311 from buccal cells) for genotyping (representing 3,932 individuals). One hundred twenty-nine DNA samples were analyzed in duplicate.

Owing to the multitude of studies of varying sample sizes in PanScan, we compared the results of genotype clustering to verify goodness of fit, detect genotype discordances and monitor potential cluster heterogeneity. The genotype models evaluated included (i) default cluster definitions provided by Illumina, (ii) clusters estimated from each study separately, (iii) clusters estimated from each study separately using samples with >98% completion rates, calling the 'low-completion' samples using those cluster models, (iv) clusters estimated from all studies together using all samples, (v) clusters estimated from all studies together using samples with >98% completion rates, then calling the low-completion samples using those cluster models and (vi) clusters estimated from each study separately using samples with >98% completion rates, followed by grouping and reclustering studies that showed similar cluster metrics. Genotypes for low-completion samples were called using the corresponding cluster model. On the basis of completion rates and low discordance between known duplicate samples, the most rigorous clustering methods were (iii), (v) and (vi). Model (v) was chosen on the basis of parsimony.

We attempted 561,466 SNP genotype assays on the 4,063 DNA samples using the Human Hap500 Infinium Assay (Illumina). Samples with <98% completion after the second attempt were subsequently excluded. SNP assays with call rates <90% were excluded. We observed an average discordance rate of 0.017% for 139 pairs of duplicate DNA assays (including 129 plated duplicate samples).

Deviation from Hardy-Weinberg proportions were tested[34] in control samples (with CEU ancestry >0.80 by STRUCTURE) of each study (**Supplementary Fig. 2**). No SNPs were excluded from analysis, as the tests for association are valid in the presence of departure from Hardy-Weinberg proportions.

Some participants with valid genotypes were excluded because of (i) unanticipated interstudy duplicates ($n = 14$), (ii) completion rates <98% ($n = 219$ samples corresponding to 74 participants), (iii) unexpected within-study duplicate ($n = 1$) and (iv) ineligible samples ($n = 8$). The final count of participants for the stage 1 association analysis was 1,896 cases and 1,939 controls (**Supplementary Table 4**).

Assessment of population structure was performed with STRUCTURE[35] by seeding the genotypes from the PanScan studies with the reference HapMap genotypes (based on Build 22 for HapMap II with MAF >5% in any of three HapMap populations)[36]. A set of 9,405 SNPs with $r^2 < 0.004$ was selected for this analysis[37–39]. A total of 59 participants (29 cases and 30 controls) were estimated to be of admixed origin with <80% similarity to CEU. No participants were excluded based on results from STRUCTURE but were assigned to the following categories for adjustment in the association analysis: European if CEU admixture portion was >80%, Asian if JPT/HCB admixture portion was >80% and 'other' if admixture with no one continental group was >80% (**Supplementary Fig. 3**). African American ancestry was defined based on self-report, with similarity to YRI ranging from 41% to 96%.

A principal component analysis (PCA) of DNA samples in this study (excluding inferred sib and half-sib pairs) was performed with EIGENSTRAT[40]. Five principal components were effective[41] for distinguishing significant population groups and were included as quantitative covariates to correct for genetic admixture.

Genotype data for the full scan was used to identify 144 participants with 60%–99% identity by state as potential relatives. Two sets of SNPs with pairwise $r^2 < 0.004$ were selected separately for Asian (13,905 SNPs) and non-Asian studies (9,405 SNPs) and run on PREST[42] to identify five unexpected full-sib pairs and two unexpected half-sib pairs (seven cases and seven controls), who were excluded from PCA but included in the association analysis.

TaqMan assays (ABI) were designed and optimized for ten SNPs in the three notable regions as well as for a technical replica assay for rs505922 (rs687621) because this SNP could not be optimized (96% genotype concordance with HapMap samples) as per SNP500Cancer.

For the fast-track replication study, 5,845 samples were genotyped, including 180 duplicate DNA samples for quality control purposes. Genotyping was performed using a multiplex integrated fluidic technology (Fluidigm Biomark) and individual TaqMan assays (ABI). During the follow-up replication genotyping, the opportunity arose to conduct a GWAS with the Illumina Infinium 610Quad. Because the same SNPs would be later genotyped, we completed genotyping only for the top ten ranked SNPs (a second GWAS is ongoing). Consequently, genotyping of some of the samples for the fast-track replication was performed with low quantities of DNA (reserving sufficient DNA for the GWAS). Sample completion ranged from 28.90% to 99.40% per study, and genotype completion rates per locus ranged from 57.7% to 99.8%. Overall genotype concordance between duplicate samples was 96.52%, indicating the reliability of the current fast-track replication results. Discordant genotypes between duplicates were set to 'missing'. A small proportion (0.2%) of samples genotyped in stage 2 were excluded, as they were unanticipated interstudy or intrastudy duplicates or had incomplete covariate data. The

corresponding Infinium cluster plots for the ten SNPs are shown in **Supplementary Figure 4**.

Association analysis. All association analyses were conducted using logistic regression, adjusted for age (in 10-year categories), sex, study, arm (for WHI, intervention versus observation), ancestry and five principal components of genetic structure. Each SNP genotype was coded as a count of minor alleles, with the exception of X-linked SNPs among males, which were coded as '2' if the participant carried the minor allele and '0' if he carried the major allele[25]. This log-linear odds model has near-optimal power across a wide range of alternative hypotheses, the main exception being rare recessive variants, for which we have limited power regardless of genotype coding[43]. A score test was performed on all genetic parameters in each model to determine statistical significance, with 1 d.f.

We analyzed each study separately and conducted two analyses pooling multiple studies: the first included all cohorts ('cohorts'); the second included all studies ('all'). We assessed heterogeneity in genetic effects across study using the Q and I^2 statistics[44].

We selected for replication ten SNPs from the three most notable regions from the GWAS (based on two or more SNPs per region ranking in the top 25 SNPs) based on the results of the two pooled analyses. We tested association between pancreatic cancer and the replication SNPs by fitting logistic regression models and testing the estimated genetic effects using the GLU software package. We analyzed each study separately in addition to pooling all eight studies. Models were adjusted for age in 10-year intervals as well as for sex, self-reported race and study. Genotypes were coded as counts of minor alleles (1-d.f. trend test). Combined single-SNP analyses pooled stage 1 and stage 2 data sets and adjusted for study, arm, age, sex, race and top five principal components of population stratification. In stage 2 studies, principal components could not be calculated and were set to 0. **Supplementary Table 5** provides the results of the stage 2 association analysis.

Data analysis and management used GLU (Genotyping Library and Utilities version 1.0), a suite of tools available as an open-source application for management, storage and analysis of GWAS data.

URLs. CGEMS portal, http://cgems.cancer.gov/; Core Genotyping Facility, http://cgf.nci.nih.gov/; EIGENSTRAT, http://genepath.med.harvard.edu/~reich/EIGENSTRAT.htm; GLU, http://code.google.com/p/glu-genetics/; SNP500Cancer, http://snp500cancer.nci.nih.gov/; STRUCTURE, http://pritch.bsd.uchicago.edu/structure.html; Tagzilla, http://tagzilla.nci.nih.gov/; PANC4, http://panc4.org.

34. 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).
35. Pritchard, J.K., Stephens, M. & Donnelly, P. Inference of population structure using multilocus genotype data. *Genetics* **155**, 945–959 (2000).
36. Frazer, K.A. *et al*. A second generation human haplotype map of over 3.1 million SNPs. *Nature* **449**, 851–861 (2007).
37. Thomas, G. *et al.* Multiple loci identified in a genome-wide association study of prostate cancer. *Nat. Genet.* **40**, 310–315 (2008).
38. 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).
39. Yu, K. *et al*. Population substructure and control selection in genome-wide association studies. *PLoS One* **3**, e2551 (2008).
40. Price, A.L. *et al*. Principal components analysis corrects for stratification in genome-wide association studies. *Nat. Genet.* **38**, 904–909 (2006).
41. Patterson, N., Price, A.L. & Reich, D. Population structure and eigenanalysis. *PLoS Genet.* **2**, e190 (2006).
42. Sun, L., Wilder, K. & McPeek, M.S. Enhanced pedigree error detection. *Hum. Hered.* **54**, 99–110 (2002).
43. Lettre, G., Lange, C. & Hirschhorn, J.N. Genetic model testing and statistical power in population-based association studies of quantitative traits. *Genet. Epidemiol.* **31**, 358–362 (2007).
44. Higgins, J.P. & Thompson, S.G. Quantifying heterogeneity in a meta-analysis. *Stat. Med.* **21**, 1539–1558 (2002).