Use of a novel non-parametric version of DEPTH to identify genomic regions associated with prostate cancer risk

Robert J. MacInnis[1,2], Daniel F. Schmidt[2], Enes Makalic[2], Gianluca Severi[3], Liesel M. FitzGerald[4], Matthias Reumann[5,6], Miroslaw K. Kapuscinski[2], Adam Kowalczyk[7,8], Zeyu Zhou[9], Benjamin W. Goudey[9], Guoqi Qian[8], Quang M. Bui[2], Daniel J. Park[10,11], Adam Freeman[2], Melissa C. Southey[10], Ali Amin Al Olama[12]; Zsofia Kote-Jarai[13] and Rosalind A Eeles[13,14] for the UKGPCS Collaborators[13]; John L. Hopper[2,15], Graham G. Giles[1,2]

None of the authors declare any conflicts of interest.

[1] Cancer Epidemiology Centre, Cancer Council Victoria, Melbourne, Australia;

[2] Centre For Epidemiology and Biostatistics, University of Melbourne, Melbourne, Australia;

[3] Human Genetics Foundation, Torino, Italy; Université Paris-Saclay, Univ. Paris-Sud, UVSQ, CESP, INSERM, Villejuif, France; Gustave Roussy, F-94805, Villejuif, France;

[4] Menzies Institute for Medical Research, University of Tasmania, Hobart, Tasmania

[5] IBM Research - Zurich, Switzerland

[6] UNU-MERIT (United Nations University – Maastricht Economic and Social Research Institute on Innovation and Technology) Maastricht University, Maastricht, The Netherlands

[7] Warsaw University of Technology, Warsaw, Poland

[8] School of Mathematics and Statistics, University of Melbourne, Parkville, Victoria, Australia

[9] IBM Research – Australia, Carlton, Australia.

[10] Genetic Epidemiology Laboratory, Department of Pathology, University of Melbourne, VIC, Australia.

[11] Melbourne Bioinformatics Platform, Victorian Life Sciences Computation Initiative, University of Melbourne, VIC, Australia

[12] Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, UK ;

[13] The Institute of Cancer Research, London, UK;

[14] The Royal Marsden NHS Foundation Trust, Surrey, UK

[15] Department of Epidemiology, School of Public Health and Institute of Health and Environment, Seoul National University, Seoul, South Korea

Running title: Use of DEPTH to identify prostate cancer genomic regions

Keywords: genome-wide association studies, machine learning algorithm, decision trees, single nucleotide polymorphism, prostate cancer.

Financial support: J. L. Hopper, G. Severi, E. Makalic, D. F. Schmidt, Q. M. Bui, G. Qian, D. J. Park, A. Kowalczyk received support from the National Health and Medical Research Council Australia project grant (1033452).

R. A. Eeles and Z. Kote-Jarai received support from Cancer Research UK (grant numbers C5047/A7357, C1287/A10118, C1287/A5260, C5047/A3354, C5047/A10692, C16913/A6135 and C16913/A6835), Prostate Research Campaign UK (now Prostate Cancer UK), The Institute of Cancer Research and The Everyman Campaign, The National Cancer Research Network UK, The National Cancer Research Institute (NCRI) UK. R. A. Eeles and Z. Kote-Jarai are grateful for support of NIHR funding to the NIHR Biomedical Research Centre at The Institute of Cancer Research and The Royal Marsden NHS Foundation Trust.

G. G. Giles, J. L. Hopper, M. C. Southey and G. Severi received support for the Prostate Cancer Research Program of Cancer Council Victoria from The National Health and Medical Research Council, Australia (126402, 209057, 251533, 396414, 450104, 504700, 504702, 504715, 623204, 940394, 614296), VicHealth, Cancer Council Victoria, The Prostate Cancer Foundation of Australia, The Whitten Foundation, PricewaterhouseCoopers, and Tattersall's.

J.L. Hopper is a Senior Principal Research Fellow of the National Health and Medical Research Council, Australia. M.C. Southey is a Senior Research Fellow of the National Health and Medical Research Council, Australia.

We would like to acknowledge the NCRN nurses and Consultants for their work in the UKGPCS study. We thank all the patients who took part in this study.

Author for correspondence: Professor John L. Hopper, Centre for Epidemiology and Biostatistics, Melbourne School of Population and Global Health, The University of Melbourne, 207 Bouverie Street, Carlton, Victoria 3053, Australia

Tel: 61-3-8344-0697

Fax: 61-3-9349-5815

Email: j.hopper@unimelb.edu.au

Word count: 2803

Number of tables: 3

Number of figures: 2

ABSTRACT

Background: We have developed a GWAS analysis method called DEPTH (DEPendency of association on the number of Top Hits) to identify genomic regions potentially associated with disease by considering overlapping groups of contiguous markers (e.g. single nucleotide polymorphisms, SNPs) across the genome. DEPTH is a machine learning algorithm for feature ranking of ultra-high dimensional datasets, built from well-established statistical tools such as bootstrapping, penalised regression and decision trees. Unlike marginal regression, which considers each SNP individually, the key idea behind DEPTH is to rank groups of SNPs in terms of their joint strength of association with the outcome. Our aim was to compare the performance of DEPTH with that of standard logistic regression analysis.

Methods: We selected 1,854 prostate cancer cases and 1,894 controls from the UK for whom 541,129 SNPs were measured using the Illumina Infinium HumanHap550 array. Confirmation was sought using 4,152 cases and 2,874 controls, ascertained from the UK and Australia, for whom 211,155 SNPs were measured using the iCOGS Illumina Infinium array.

Results: From the DEPTH analysis we identified 14 regions associated with prostate cancer risk that had been reported previously; five of which would not have been identified by conventional logistic regression. We also identified 112 novel putative susceptibility regions.

Conclusions: DEPTH can reveal new risk-associated regions that would not have been identified using a conventional logistic regression analysis of individual SNPs.

Impact: This study demonstrates that the DEPTH algorithm could identify additional genetic susceptibility regions that merit further investigation.

INTRODUCTION

Conventional approaches to analysing Genome-Wide Association Studies (GWAS) have been based on considering each single nucleotide polymorphism (SNP) individually, and have identified at least 100 independent SNPs associated with prostate cancer risk.(1, 2) Difficulties encountered when analysing GWAS data this way include the large number of correlated risk-associated SNPs and the fact that disease-causing variants may not necessarily have been measured. Only about a third of the familial risk of prostate cancer can be explained by these genetic susceptibility markers discovered to date,(1) and there may be many more as yet unidentified risk-associated variants. While approaches such as using larger samples, meta-analyses, imputation and greater coverage arrays are likely to explain a greater proportion of the familial risk, the development and application of more complex statistical approaches to existing data could increase understanding of the genetic component of the risk for prostate cancer and potentially reduce the need for genotyping ever larger study samples, which is a major issue for most cancers.(3)

Recently, we developed a GWAS analysis method called DEPTH (DEPendency of association on the number of Top Hits) to identify regions potentially associated with disease by considering overlapping groups of contiguous markers across the genome.(4) DEPTH is a machine learning algorithm for feature ranking of ultra-high dimensional datasets. It is built from well-established statistical tools, such as bootstrapping, penalised regression and decision trees, which are utilised to determine which exposures in a dataset are associated with the outcome variable.(5-7) Unlike marginal regression, which considers each SNP individually, the key concept behind DEPTH is to rank groups of SNPs in terms of their joint strength of association with the outcome.

Currently, there are two implementations of the DEPTH algorithm: (i) an IBM BlueGene/Q supercomputer version which is written in C/C++ and uses the Eigen library for numerical computing (parametric version), and (ii) a MATLAB implementation (non-parametric version). The IBM BlueGene/Q version of DEPTH is intended to be used for datasets where the sample size (N) and the number of predictors (P) are extremely large (e.g., N > 50,000 or P > 1,000,000). The MATLAB version, which is preferred for smaller datasets as it can be run on a commodity PC, is based on decision trees, a nonparametric technique commonly used in regression and classification problems. Decision trees are estimated using minimum message length, a well-established information theoretic approach to model selection and parameter estimation. The DEPTH algorithm grows decision trees using a sliding window of SNPs and provides a measure of association for each window of SNPs under consideration. The statistic of association is equivalent to the Bayesian posterior log-odds in favour of association, which is driven by the ability of the SNPs to discriminate between cases and controls. One can, of course, investigate the directions of associations for individual SNPs by examining the estimates for SNPs in a window from a fitted model. However, because the model fits marginal and joint interaction effects between the SNPs within the window, the direction of association for a given SNP could depend on other SNPs in the region. The size of the sliding window can be defined in terms of genetic distance (e.g., 100 Kb) or a fixed number of variants (e.g., 100 SNPs) as arguments in the main command line.

Our aim was to compare conventional logistic regression GWAS analysis with the MATLAB implementation of the DEPTH algorithm using two previously analysed prostate cancer case-control datasets.

MATERIALS AND METHODS

Two previously analysed datasets were used in this analysis, a Stage 1 GWAS of the UK Genetic Prostate Cancer Study (UKGPCS; henceforth referred to as the UK GWAS dataset) and a Stage 2 custom array UK and Australian dataset (referred to as the iCOGS dataset). All study participants were self-reported as Caucasian and gave written informed consent. Both studies were approved by the appropriate national ethics committees.

For the UK GWAS dataset, a total of 1,854 prostate cancer cases and 1,894 controls were selected from the UKGPCS.(8) Cases were diagnosed at or before the age of 60 years or had a first- or second-degree relative with prostate cancer. Controls were men aged 50 years or older with a PSA of <0.5ng/ml, frequency matched to the geographical distribution of the cases. A total of 541,129 SNPs were genotyped using the Illumina Infinium HumanHap550 array (version 1). This dataset is described in detail elsewhere.(9)

For the iCOGS dataset, a total of 4,544 cases and 3,376 controls were selected from the UK and Australia. The 2,859 prostate cancer cases and 2,193 controls from the UK were selected from the UKGPCS and were not participants in the aforementioned UK GWAS dataset.(10) Blood DNA was collected from prostate cancer cases aged ≤ 60 years at diagnosis across the UK and from a systematic series of cases attending the prostate cancer clinic at The Royal Marsden NHS Foundation Trust. Diagnosis was confirmed from medical record or death certificate, 60% were clinically detected. Controls with normal PSA levels (<3ng/ml) were selected from the same GP register and five-year age band as the cases. The remaining 1,685 cases and 1,183 controls were selected from the Early Onset Prostate Cancer Study (EOPCS), the Risk Factors for Prostate Cancer Study (RFPCS), and the Melbourne Collaborative Cohort Study (MCCS). These studies are described in detail elsewhere.(10-15) A total of 211,155 SNPs were genotyped using the iCOGS chip.(10) As there was notable ethnic heterogeneity, we selected 4,152 cases and 2,874 controls for further analyses based on

inspection of scatter plots from principal components, where the first eigenvalue was between -1 and 0 and the second eigenvalue was less than 0.5 (see Supplementary Figure 1).

We excluded SNPs with a call rate <95%, minor allele frequency for controls <1%, or exhibiting distributions strongly departed from that expected under Hardy-Weinberg equilibrium (P<0.00001). SNPs added to the iCOGS array for fine mapping (see (10) for details) were also excluded, but otherwise SNPs were not pruned based on linkage disequilibrium. This left 508,932 (UK GWAS dataset) and 173,524 (iCOGS dataset) SNPs available for analysis. Of the 100 SNPs that have been identified as being associated with prostate cancer susceptibility((1), 67 (UK GWAS dataset) and 68 (iCOGS dataset) were directly genotyped in both datasets.

We present results from the DEPTH algorithm using a 100 Kb sliding window of SNPs from which the posterior log-odds in favour of association for each window was calculated. Note, the posterior log-odds in favour of association are not directly comparable with P values derived from logistic regression. The null distribution was empirically estimated using 1,000 bootstrap iterations. The maximum 95th percentile of the null distribution equalled approximately 1.0 across the genome for both datasets. To reduce the possibility of identifying false-positive regions, we defined a "risk-associated region" in the UK GWAS dataset as having a peak with a magnitude of at least 1 unit above the 95th percentile of the null distribution, extending from both sides of the peak until the signal drops below the null distribution; see discussion about this below. Regions were deemed confirmed by the same criteria in the iCOGS dataset (i.e., greater than 1 unit above the 95th percentile of the null distribution).

Figure 1 shows a sample of the output generated from chromosome 8 using the DEPTH software. In this example, only the region at the 23.0-23.6 Mb position (shown with black

shading) was classified as risk-associated as the signal was greater than 1 unit above the 95th percentile of the null distribution. Peaks such as the one at 20.0-20.2 (shown with only light grey shading) were less than 1 unit above the 95th percentile of the null distribution and thus were not considered to be risk-associated. Signals that were less than 0 are denoted as 0 in the output for ease of visual interpretation.

Conventional logistic regression analyses were computed for each SNP using the software package PLINK v1.9 (http://pngu.mgh.harvard.edu/purcell/plink/).(16) As in an earlier publication that analysed the UK GWAS dataset,(9) all SNPs with a $P < 10^{-6}$ based on a 1 degree of freedom trend test were deemed significant, while all SNPs with a $P < 0.002$ and in the same direction (based on a Bonferroni adjustment for 50 SNPs) in the iCOGS dataset were deemed to be confirmed.

RESULTS

From the DEPTH analysis of the UK GWAS dataset, we identified 137 prostate cancer risk-associated regions with maximum posterior log-odds in favour of association greater than 1 unit above the 95th percentile of the null distribution. Twenty-five of these regions contained 33 of the previously identified 100 independent prostate cancer susceptibility SNPs. The remaining 112 regions that were not confirmed by this criterion in the iCOGS dataset represent potential novel susceptibility regions (results not shown). The number of measured SNPs within the 137 regions depended on whether they were previously identified prostate cancer susceptibility regions or not. For example, across the 25 regions that contained at least one previously identified susceptibility SNP, there was an average of 65 and 101 genotyped SNPs per region for the UK GWAS and iCOGS datasets, respectively. On the other hand, for

the remaining 112 regions, the average number of SNPs genotyped per region was 48 for the UK GWAS dataset, but only 16 for the iCOGS chip.

Of the 137 susceptibility regions identified from the DEPTH analysis of the UK GWAS dataset, we confirmed 14 from the DEPTH analysis of the iCOGS dataset (Table 1). All 14 confirmed regions contained at least one previously identified prostate cancer susceptibility SNP. Table 1 shows that four of these regions (#2, #3, #6, #9) did not contain any SNPs with $P < 10^{-6}$ when analysed using standard logistic regression. Three of these regions (#2, #6, #9) were subsequently identified in a third-stage analysis involving an additional 16,229 cases and 14,821 controls from 21 studies (17), while region #3 was identified using 25,074 prostate cancer cases and 24,272 controls from the international PRACTICAL Consortium.(10)

After performing conventional logistic regression analyses of the UK GWAS dataset, we identified 50 SNPs that were significant at the $P < 10^{-6}$ level (Table 2). We found confirmatory evidence ($P < 0.002$) for 40 of the 44 SNPs that were genotyped in the iCOGS dataset (the six SNPs that were not genotyped were all located in the 8q24 region). These 40 SNPs were located in 11 regions, and these regions were also confirmed by DEPTH analyses as being risk-associated. Two of the four SNPs that were not confirmed by logistic regression analyses (rs2660753 and rs2659056) were located very close to at least one other SNP that was confirmed by logistic regression; therefore, we considered that these two regions were confirmed by logistic regression. The remaining two SNPs (rs9364554 and rs902774), however, did not have any other confirmed SNPs located nearby. We found, that the region encompassing rs9364554 on chromosome 6 was confirmed by DEPTH analyses as being risk-associated. This region, therefore, presents an additional region to the four regions identified from the DEPTH analyses that logistic regression would not have found. On the other hand, we found no confirmatory evidence using either analysis method for the region

on chromosome 12 that contained the SNP rs902774, but this may be due to the disease characteristics of the iCOGS dataset as this SNP was originally identified from an analysis of 2891 advanced prostate cases and 4592 controls of European ancestry.(18)

DISCUSSION

Using DEPTH analysis we identified 14 regions associated with prostate cancer risk that had been reported previously; five of which would not have been identified using conventional logistic regression on these datasets. We also identified 112 novel putative susceptibility regions that were not identified using logistic regression.

As the iCOGS chip was developed as a custom genotyping array, the design focused on previously known risk loci and did not include a GWAS backbone. We were, therefore, unable to confirm any of the 112 novel risk-associated regions detected by DEPTH, primarily due to insufficient numbers of iCOGS array SNPs in those regions. While using imputation could be a solution, it does not provide independent measures of SNPs. Increasing SNP density increases the chance of discovering associations, but generally will result in larger stretches of the same signal on the DEPTH plot, and has little effect on the ranking process or the generation of the empirical null distribution. A future version of DEPTH will incorporate imputed SNPs and be used to test whether imputed SNPs improve risk loci detection compared with using only measured SNPs.

While DEPTH presents a new approach to analysing GWAS data, the statistical techniques that underlie this methodology are well established. DEPTH is a fusion of ideas, which share a common goal towards analysing genomic data. It is designed to run in a parallel environment and exploits the correlation structure within the data. It is very flexible and can

accommodate different models (additive, dominant, recessive) and window sizes (based on number of SNPs or base pairs) to suit most analytical situations. In addition, the non-parametric version is straightforward to implement and does not require supercomputing facilities to complete analyses in a timely manner. We also plan to implement continuous phenotypes in future papers.

At present, the non-parametric version of DEPTH does not allow for principal components adjustments. We intend to implement this feature in a future version of DEPTH. While ethnic background of the UK participants was fairly homogeneous, this was not the case for the Australian participants who predominantly included Australian born men of northern European background, but also included southern European migrants. In sensitivity analyses, we observed that *P* values from the logistic regression analyses were similar when using the restricted iCOGS dataset compared with the full iCOGS dataset after adjustment for principal components (results not shown). While it is preferable to utilise the full dataset, the similar results from the sensitivity analyses suggest it is unlikely that the results from our non-parametric DEPTH analyses would change appreciably after adjustment for principal components, but further work is needed in this regard.

The conventional approach for identifying individual susceptibility SNPs involves using a "Bonferroni adjusted" *p*-value threshold to classify observed associations as being "significant". These thresholds are deliberately chosen to be highly conservative in order to minimise false positives. It should be noted that any choice of threshold, not matter how it is made, is essentially arbitrary. Here, where consideration is about strength of signal across a region (not statistical significance of a single marker), we used simulations to determine the empirical null distribution of the code lengths as a guide to selecting an appropriate threshold for "significance". To accurately estimate extreme percentiles, though, is computationally expensive due to the requirement for very large number of simulations to be run. This is

particularly apparent if the number of SNPs in the window size is large because the computational burden depends more on the density of the SNP array than the number of cases and controls. To circumvent these issues, we used the 95% percentile of the empirical null distribution as an initial choice of threshold T, and increased this base threshold by some quantity $\delta \geq 0$ (we chose δ to equal 1 in the above analyses) to obtain a more conservative threshold without the requirement for excessive simulations. Another advantage of our approach is that the threshold chosen this way still retains a clear Bayesian interpretation: the quantity $\exp(-T-\delta)$ is approximately equally to the Bayes factor required to reject the null hypothesis. This can be used to guide the choice of $T+\delta$ based on the particular aim of the analysis, which in this paper is discovering regions worthy of further investigation; e.g. by sequencing or fine mapping.

DEPTH is a discovery tool with the ability to reveal risk-associated regions that complements other approaches. Confirmation cannot be sought only by testing for disease associations of individual SNPs using independent data sets. Rather it should involve more nuanced approaches to detecting susceptibility regions, including DEPTH analyses of other data sets, burden tests of candidate regions, family-based linkage analyses, and targeted sequencing. Moreover, DEPTH signals could be due to one or more rare variants that are not necessarily observed in other studies. The genetic architecture of cancers is obviously more complex than the current highly conservative GWAS analysis paradigm based on testing for independent associations of common SNPs

In summary, we have presented a new GWAS analytical method and shown that it is able to detect risk-associated regions that would otherwise be missed using conventional regression approaches that consider each SNP individually. From our study of two existing prostate cancer datasets, we have identified and confirmed 14 regions that have been previously reported to be associated with prostate cancer risk, five of which would not have been

identified using the conventional approach that considers each SNP individually. This study demonstrates that the DEPTH algorithm can be applied to existing and future datasets to identify additional genetic susceptibility regions that merit further investigation.

REFERENCES

1. Amin Al Olama A, Kote-Jarai Z, Berndt SI, Conti DV, Schumacher F, Han Y, et al. A meta-analysis of 87,040 individuals identifies 23 new susceptibility loci for prostate cancer. Nat Genet. 2014;46:1103-9.

2. Eeles R, Goh C, Castro E, Bancroft E, Guy M, Al Olama AA, et al. The genetic epidemiology of prostate cancer and its clinical implications. Nat Rev Urol. 2014;11:18-31.

3. Camastra F, Di Taranto MD, Staiano A. Statistical and Computational Methods for Genetic Diseases: An Overview. Comput Math Methods Med. 2015;2015:954598.

4. Makalic E, Schmidt DF, Hopper JL. DEPTH: A Novel Algorithm for Feature Ranking with Application to Genome-Wide Association Studies. In: Cranefield S, Abhaya N, editors. AI 2013: Advances in Artificial Intelligence. Cham, Switzerland: Springer International Publishing; 2013. p. 80-5.

5. Wallace CS. Statistical and Inductive Inference by Minimum Message Length: Springer; 2005.

6. Wallace CS, Patrick JD. Coding Decision Trees. Machine Learning. 1993;11:7-22.

7. Breiman L, Friedman JH, Olshen RA, Stone CJ. Classification and regression trees. . Monterey, CA.: Wadsworth & Brooks/Cole Advanced Books & Software; 1984.

8. Eeles RA. Genetic predisposition to prostate cancer. Prostate Cancer Prostatic Dis. 1999;2:9-15.

9. Eeles RA, Kote-Jarai Z, Giles GG, Olama AA, Guy M, Jugurnauth SK, et al. Multiple newly identified loci associated with prostate cancer susceptibility. Nat Genet. 2008;40:316-21.

10. Eeles RA, Olama AA, Benlloch S, Saunders EJ, Leongamornlert DA, Tymrakiewicz M, et al. Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array. Nat Genet. 2013;45:385-91.

11. Giles GG, Severi G, McCredie MR, English DR, Johnson W, Hopper JL, et al. Smoking and prostate cancer: findings from an Australian case-control study. Ann Oncol. 2001;12:761-5.

12. Giles GG, Severi G, Sinclair R, English DR, McCredie MR, Johnson W, et al. Androgenetic alopecia and prostate cancer: findings from an Australian case-control study. Cancer Epidemiol Biomarkers Prev. 2002;11:549-53.

13. MacInnis RJ, English DR, Gertig DM, Hopper JL, Giles GG. Body size and composition and prostate cancer risk. Cancer Epidemiol Biomarkers Prev. 2003;12:1417-21.

14. Severi G, Giles GG, Southey MC, Tesoriero A, Tilley W, Neufing P, et al. ELAC2/HPC2 polymorphisms, prostate-specific antigen levels, and prostate cancer. J Natl Cancer Inst. 2003;95:818-24.

15. Severi G, Morris HA, MacInnis RJ, English DR, Tilley W, Hopper JL, et al. Circulating steroid hormones and the risk of prostate cancer. Cancer Epidemiol Biomarkers Prev. 2006;15:86-91.

16. Purcell S, Neale B, Todd-Brown K, Thomas L, Ferreira MA, Bender D, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. Am J Hum Genet. 2007;81:559-75.

17. Eeles RA, Kote-Jarai Z, Al Olama AA, Giles GG, Guy M, Severi G, et al. Identification of seven new prostate cancer susceptibility loci through a genome-wide association study. Nat Genet. 2009;41:1116-21.

18. Schumacher FR, Berndt SI, Siddiq A, Jacobs KB, Wang Z, Lindstrom S, et al. Genome-wide association study identifies new prostate cancer susceptibility loci. Hum Mol Genet. 2011;20:3867-75.

Table 1. Summary results for regions identified using DEPTH from the UK GWAS dataset and confirmed with the iCOGS dataset.

DEPTH Region #	Chr	Build 37 Position	No. known PCa SNPs	DEPTH[a] UKGWAS	Min P value UKGWAS	DEPTH[a] iCOGS	Min P value iCOGS
1	3	87.0-87.2	1	3.8	1.2×10^{-07}	2.9	5.5×10^{-07}
2	4	95.4-95.7	2	1.1	1.9×10^{-04}	1.9	1.6×10^{-06}
3	6	153.3-153.5	1	1.0	1.5×10^{-02}	1.4	2.2×10^{-05}
4	6	160.5-161.0	1	2.9	9.8×10^{-07}	3.5	2.3×10^{-08}
5	7	97.6-97.9	1	4.4	1.3×10^{-08}	3.0	4.4×10^{-06}
6	8	23.1-23.6	2	1.3	5.3×10^{-06}	1.6	1.1×10^{-05}
7	8	127.7-128.6	6	16.2	7.8×10^{-17}	13.2	1.2×10^{-14}
8	10	51.5-51.6	1	20.2	2.1×10^{-23}	8.5	9.4×10^{-13}
9	11	2.1-2.3	1	2.4	1.7×10^{-05}	4.7	7.7×10^{-09}
10	11	68.8-69.1	1	2.7	2.2×10^{-07}	15.9	9.5×10^{-20}
11	17	36.0-36.2	2	7.4	1.3×10^{-12}	10.8	7.0×10^{-16}
12	17	68.8-69.3	1	4.5	5.8×10^{-07}	2.2	8.4×10^{-07}
13	19	51.2-51.5	1	16.7	4.9×10^{-20}	8.2	2.9×10^{-12}
14	X	51.0-51.8	1	4.8	2.4×10^{-08}	6.5	5.6×10^{-06}

[a] Measured in terms of posterior log-odds in favour of association

Table 2. Summary results for the 50 SNPs selected from the UK GWAS dataset with P $<10^{-6}$.

Chr	Marker	Build 37 Position	UKGWAS P-Logistic	iCOGS P-Logistic	DEPTH Region #[a]
3	rs2660753	87110674	1.2×10^{-07}	9.7×10^{-02}	1
3	rs17023900	87134800	3.8×10^{-07}	5.0×10^{-04}	1
6	rs9364554	160833664	9.8×10^{-07}	3.3×10^{-02}	4
7	rs705308	97695363	5.1×10^{-08}	5.8×10^{-06}	5
7	rs6465654	97786282	8.0×10^{-08}	6.8×10^{-06}	5
7	rs6465657	97816327	1.3×10^{-08}	7.4×10^{-06}	5
8	rs12543663	127924659	9.7×10^{-07}	6.1×10^{-06}	7
8	rs1016343	128093297	1.9×10^{-08}	3.5×10^{-11}	7
8	rs16901966	128110252	4.8×10^{-08}	Not genotyped	7
8	rs16901970	128112715	4.8×10^{-08}	Not genotyped	7
8	rs10505483	128125195	6.9×10^{-08}	Not genotyped	7
8	rs7817677	128125504	1.0×10^{-07}	Not genotyped	7
8	rs6983267	128413305	1.2×10^{-13}	5.4×10^{-13}	7
8	rs7837328	128423127	2.2×10^{-08}	2.3×10^{-09}	7
8	rs7014346	128424792	1.5×10^{-08}	1.5×10^{-09}	7
8	rs1447293	128472320	1.7×10^{-07}	8.3×10^{-05}	7
8	rs921146	128475185	2.3×10^{-08}	1.2×10^{-07}	7
8	rs1447295	128485038	2.8×10^{-16}	1.2×10^{-12}	7
8	rs4242382	128517573	1.5×10^{-16}	4.7×10^{-14}	7
8	rs4242384	128518554	7.8×10^{-17}	Not genotyped	7
8	rs7017300	128525268	3.0×10^{-11}	3.2×10^{-09}	7
8	rs11988857	128531873	3.1×10^{-13}	9.5×10^{-10}	7
8	rs9656816	128534654	4.6×10^{-14}	1.5×10^{-09}	7
8	rs7837688	128539360	1.6×10^{-16}	Not genotyped	7
10	rs2611512	51515534	3.0×10^{-11}	6.8×10^{-07}	8
10	rs3123078	51524971	7.9×10^{-15}	6.6×10^{-09}	8
10	rs7920517	51532621	9.0×10^{-13}	8.0×10^{-09}	8
10	rs11006207	51538176	7.2×10^{-13}	6.8×10^{-09}	8
10	rs10993994	51549496	2.1×10^{-23}	9.4×10^{-13}	8
11	rs7931342	68994497	2.6×10^{-07}	2.4×10^{-14}	10
11	rs10896449	68994667	2.2×10^{-07}	8.7×10^{-15}	10
11	rs10896450	69008114	3.5×10^{-07}	7.7×10^{-15}	10
11	rs12799883	69010651	2.8×10^{-07}	5.2×10^{-15}	10
12	rs902774	53273904	2.3×10^{-07}	4.6×10^{-02}	-
17	rs3744763	36090885	2.0×10^{-07}	1.9×10^{-05}	11
17	rs7501939	36101156	1.3×10^{-12}	8.5×10^{-09}	11
17	rs3760511	36106313	4.0×10^{-08}	1.3×10^{-10}	11
17	rs1859962	69108753	5.8×10^{-07}	4.6×10^{-06}	12
17	rs9889335	69115146	6.2×10^{-07}	3.4×10^{-06}	12
19	rs2659056	51335943	1.4×10^{-07}	5.4×10^{-03}	13
19	rs266849	51349090	1.7×10^{-16}	5.5×10^{-06}	13
19	rs266870	51351934	2.4×10^{-09}	6.3×10^{-04}	13
19	rs1058205	51363398	4.9×10^{-20}	1.1×10^{-07}	13
19	rs2735839	51364623	7.9×10^{-20}	2.3×10^{-07}	13

X	rs4907790	51197711	1.0×10^{-06}	5.8×10^{-05}	14
X	rs1327301	51210057	1.2×10^{-07}	5.6×10^{-06}	14
X	rs5945572	51229683	1.0×10^{-07}	1.3×10^{-05}	14
X	rs5945619	51241672	2.4×10^{-08}	2.8×10^{-05}	14
X	rs1419040	51352035	1.9×10^{-07}	3.7×10^{-04}	14
X	rs5991735	51552884	1.4×10^{-07}	2.5×10^{-04}	14

[a] Regions identified using DEPTH from the UK GWAS dataset and confirmed with the iCOGS dataset, see Table 1.

LEGENDS TO FIGURES

Figure 1. Sample output obtained from the DEPTH software using the UK GWAS dataset. The solid line represents the raw signal, the light grey shading represents signal above the 95th percentile of the null distribution, and the black shaded area represents the "risk-associated region" (i.e., 1 unit above the 95th percentile of the null distribution). Position was based on SNP Build 37/hg19 coordinates.