EJHG Open

European Journal of Human Genetics (2012) 20, 469–475
© 2012 Macmillan Publishers Limited All rights reserved 1018-4813/12
www.nature.com/ejhg

ARTICLE

A statistical method for region-based meta-analysis of genome-wide association studies in genetically diverse populations

Xu Wang[1], Xuanyao Liu[2], Xueling Sim[3], Haiyan Xu[3], Chiea-Chuen Khor[4], Rick Twee-Hee Ong[2,3], Wan-Ting Tay[5], Chen Suo[3], Wan-Ting Poh[1], Daniel Peng-Keat Ng[1], Jianjun Liu[4], Tin Aung[5,6], Kee-Seng Chia[1,2,3], Tien-Yin Wong[5,6,7,8], E-Shyong Tai[1,8] and Yik-Ying Teo*[,1,2,3,4,9]

Genome-wide association studies (GWAS) have become the preferred experimental design in exploring the genetic etiology of complex human traits and diseases. Standard SNP-based meta-analytic approaches have been utilized to integrate the results from multiple experiments. This fundamentally assumes that the patterns of linkage disequilibrium (LD) between the underlying causal variants and the directly genotyped SNPs are similar across the populations for the same SNPs to emerge with surrogate evidence of disease association. We introduce a novel strategy for assessing regional evidence of phenotypic association that explicitly incorporates the extent of LD in the region. This provides a natural framework for combining evidence from multi-ethnic studies of both dichotomous and quantitative traits that (i) accommodates different patterns of LD, (ii) integrates different genotyping platforms and (iii) allows for the presence of allelic heterogeneity between the populations. Our method can also be generalized to perform gene-based or pathway-based analyses. Applying this method on real GWAS data in type 2 diabetes (T2D) boosted the association evidence in regions well-established for T2D etiology in three diverse South-East Asian populations, as well as identified two novel gene regions and a biologically convincing pathway that are subsequently validated with data from the Wellcome Trust Case Control Consortium.

European Journal of Human Genetics (2012) **20,** 469–475; doi:10.1038/ejhg.2011.219; published online 30 November 2011

Keywords: genome-wide association studies; linkage disequilibrium; meta-analysis; pathway analysis

INTRODUCTION

Remarkable achievements have been made in large-scale genetic studies of common diseases and complex traits.[1,2] The identification of variants in the human genome that are convincingly associated with different phenotypes has mainly been carried out in individuals of European descent, although increasingly studies involving non-Caucasian samples from diverse population groups have been published or are currently being conducted. Genome-wide meta-analyses (GWMA) involving tens of thousands of samples have extended the success in allowing novel variants with smaller effect sizes to be discovered.[3–7] Despite these triumphs, these findings really account for only a small fraction of the total disease heritability,[8] suggesting undiscovered genetic mechanisms may be responsible or alternative methods to analyze these data may be necessary to address the missing heritability.

Current implementation of GWMA requires the same SNPs to display consistent evidence of phenotypic association across multiple populations. This implicitly assumes that across these populations, (i) the same causal variant is present; (ii) the linkage disequilibrium (LD) pattern between the causal variant and the assayed SNPs is similar and (iii) the effect sizes observed at the assayed SNPs are consistent.[9,10] Random effects methods for combining data across studies do not utilize information from neighboring SNPs that may present concurring evidence of disease association in different studies, and often have the tendency to weaken association signals.[9] SNP-based meta-analyses also require the same SNPs to be genotyped in all the populations, although this requirement can apparently be addressed by imputation strategies that effectively standardize the SNP content across different studies[11–13] (Figure 1). However, imputation does not always present an effective solution, particularly in the absence of appropriate population reference panels.[14,15]

Before recent whole-genome sequencing endeavors, SNP discoveries were predominantly made in populations of European ancestry.[16] This strong ascertainment bias has inadvertently skewed the SNPs surveyed in the International Hap-Map Project,[17] which consequently prejudiced the SNP content of commercial genotyping platforms to carry tagging SNPs that are liable to exhibit higher minor allele frequencies in European populations.[18] This means current genotyping arrays may be less optimal for non-European populations, resulting in attenuated association signals due to lower allele frequencies and weaker LD.[14,15]

[1]Department of Epidemiology and Public Health, National University of Singapore, Singapore, Singapore; [2]NUS Graduate School for Integrative Science and Engineering, National University of Singapore, Singapore, Singapore; [3]Centre for Molecular Epidemiology, National University of Singapore, Singapore, Singapore; [4]Genome Institute of Singapore, Agency for Science, Technology and Research, Singapore, Singapore; [5]Singapore Eye Research Institute, Singapore National Eye Centre, Singapore, Singapore; [6]Department of Ophthalmology, National University of Singapore, Singapore, Singapore; [7]Centre for Eye Research Australia, University of Melbourne, Melbourne, Australia; [8]Department of Medicine, National University of Singapore, Singapore, Singapore; [9]Department of Statistics and Applied Probability, National University of Singapore, Singapore, Singapore
*Correspondence: Professor Y-Y Teo, Department of Statistics and Applied Probability, Blk S16, Level 7, 6 Science Drive 2, Faculty of Science, National University of Singapore, Singapore 117546, Singapore. Tel: (65) 6516 2760; Fax: (65) 6872 3919; E-mail: statyy@nus.edu.sg

Received 16 August 2011; revised 24 October 2011; accepted 26 October 2011; published online 30 November 2011

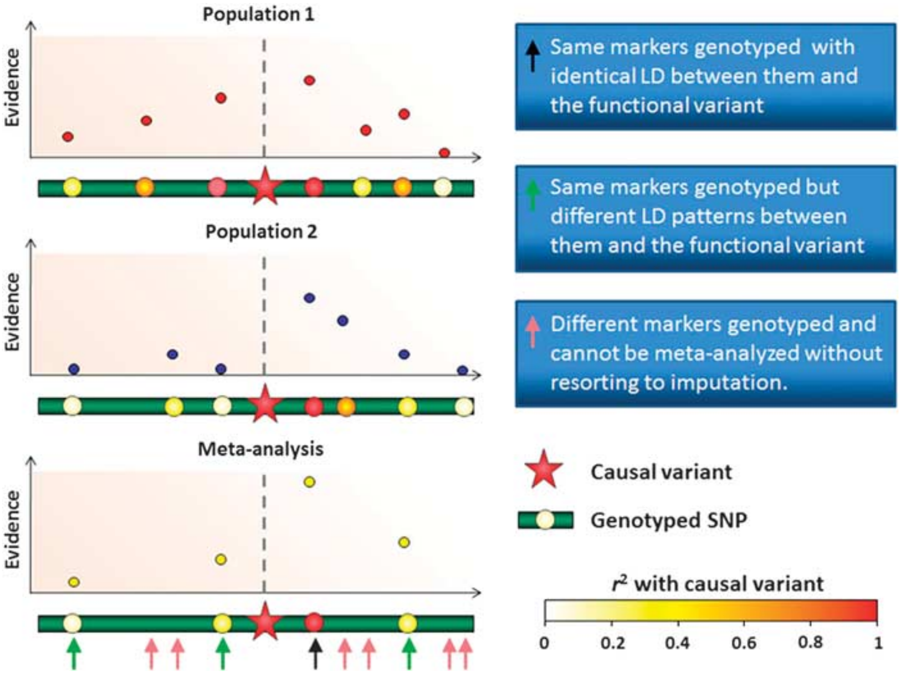
Population 1
Evidence
Population 2
Evidence
Meta-analysis
Evidence
Same markers genotyped with identical LD between them and the functional variant
Same markers genotyped but different LD patterns between them and the functional variant
Different markers genotyped and cannot be meta-analyzed without resorting to imputation.
Causal variant
Genotyped SNP
r^2 with causal variant
0
0.2
0.4
0.6
0.8
1

Figure 1 Illustration of the three scenarios in a meta-analysis, where the genotyped SNPs may be in different degree of LD with the unobserved causal variant (star): (i) the ideal situation where the same SNPs are genotyped in two studies, and the LD between them and the functional variant is identical in both populations (black arrow); (ii) a realistic situation where the same markers are genotyped in two studies, but different LD patterns exist between them and the functional variant (green arrow); (iii) realistic situation where different markers are genotyped in two studies, and cannot be meta-analysed without resorting to imputation (pink arrow). The LD between the causal variant and each SNP is represented in different color intensity ranging from white (low LD) to red (high LD).

The pursuit of evidence stronger than the genome-wide significance is thus more challenging, and larger sample sizes in non-European studies and meta-analyses of ethnically mixed populations are required to compensate for variations in LD patterns from European populations.[14,19] Instead of seeking individual variants that display convincing evidence of phenotypic association across multiple populations, a more realistic scenario is perhaps to look for genomic regions with consistent clustering of SNPs exhibiting moderate signals in these populations.

In this paper, we propose a novel paradigm for interrogating genetic data for disease association given either a dichotomous or a quantitative outcome. Our method works by quantifying the degree of over-representation of associated SNPs in a pre-defined genomic region, given a specific definition of statistical significance. Through an eigen-decomposition of the matrix measuring the LD between every possible pair of SNPs in the region, the effective number of independent SNPs as well as the number of independent SNPs exhibiting evidence of phenotypic association can be evaluated (Supplementary Figure S1). The regional evidence of phenotypic association is thus quantified as the extent of over-representation of independent associated SNPs against the effective total number of independent SNPs in the region. This approach can be applied in a genome-wide fashion by considering moving windows of a fixed length within a population. In addition, this presents a natural framework for integrating the results from multiple studies in a region-based genome-wide meta-analysis, where we can sum up the number of independent signals and independent SNPs in each region across the different studies, and to calculate a single regional *P*-value for this meta-analysis by quantifying the joint extent of over-representation. This framework also allows a straightforward extension to consider evidence across genes and biological pathways.

METHODS

Region-based analysis

Our region-based meta-analysis approach relies on the principle that when L^* independent hypotheses are tested at a statistical significance threshold of $\alpha\%$ (P_{crit}), on average we expect $\alpha L^*/100$ of these hypotheses to display statistical evidence more significant than $\alpha\%$ by chance. In the application within a genome-wide association study (GWAS), suppose there are 100 SNPs in a particular genomic window of 250 kb and the threshold for defining statistical significance has been set at 1%. Under the null hypothesis that none of the 100 SNPs are associated with the phenotype, we expect one SNP on average to exhibit a *P*-value that is <0.01 if the 100 SNPs are mutually independent. An over-representation of independent SNPs with *P*-value <0.01 in this genomic region thus corresponds to evidence that suggests the region is associated with the phenotype. However, the presence of LD implies the assumption of independence between the SNPs is unlikely to be valid.

In order to evaluate the effective number of 'independent' SNPs in each genomic region, we perform an eigen-decomposition of the $L \times L$ symmetric correlation matrix M between the L SNPs with entry m_{ij} denoting the LD in directional r^2 between the ith and the jth SNP, where the direction is determined by the sign of D'. Here we assume the minor allele frequencies of all L SNPs are at least 1%. The resulting eigenvectors effectively represent mutually independent contributions in explaining the variance in the correlation matrix, and each eigenvector is given as a linear combination of SNPs that are in at least some degree of LD. The SNP loadings of each eigenvector measure the extent each SNP contributes to the eigenvector, and the relative loadings between the SNPs for each eigenvector provide a surrogate for the degree of correlation between the SNPs. The L eigenvectors thus represent independent sources of information from all the SNPs in the window, and the number of eigenvectors N_{total} that cumulatively accounts for $\tau\%$ of the variance can be determined as $\text{argmin}_l \sum_{i=1}{}^{l} \lambda_i \geq \tau L\%$, for $1 \leq l \leq L$ and where λ_i represents the eigenvalue corresponding to the ith eigenvector $\boldsymbol{e}_i$. Let $\boldsymbol{w}$ denote a vector of length L with the w_i entry corresponding to one if the observed *P*-value for the ith SNP is $<P_{\text{crit}}$, and zero otherwise. Suppose

e_{ij} denote the *j*th component in the *i*th eigenvector, then the corresponding component in the *i*th scaled eigenvector represented by $\mathbf{e'}_i$ is $|e_{ij}|/\sum_{j=1}{}^{L}|e_{ij}|$. The effective number of independent SNPs that exhibit *P*-value $<P_{\text{crit}}$ is thus calculated as

$$N_{hit} = \sum_{i=1}^{N_{total}} w * e_i^{'}.$$

The regional evidence for the extent of over-representation of SNPs with *P*-values $<P_{\text{crit}}$ is calculated as the upper-tailed *P*-value of the exact Binomial test for observing N_{hit} out of N_{total} SNPs when the success probability is given as P_{crit}. However, as N_{hit} can be a non-integer, we estimate the *P*-value associated with N_{hit} by linear interpolating between the *P*-values obtained for the floor and ceiling integer values of N_{hit}, or equivalently $P_{\text{floor}}+(P_{\text{ceiling}}-P_{\text{floor}})\times(N_{\text{hit}}-\lfloor N_{\text{hit}}\rfloor)$, where P_{floor} and P_{ceiling} denote the *P*-values associated with $\lfloor N_{\text{hit}}\rfloor$ (the floor integer value of N_{hit}) and $\lceil N_{\text{hit}}\rceil$ (the ceiling integer value of N_{hit}), respectively. Our simulation study suggests this procedure is conservative and tends to underestimate the evidence (Supplementary Figure S2), thus the true evidence is likely to be more significant than the resultant linear-interpolated *P*-value.

A region-based meta-analysis across *K* independent populations can be performed by calculating the corresponding $N_{\text{hit}}{}^{(k)}$ and $N_{\text{total}}{}^{(k)}$ from each population *k* in the same genomic window. The cumulative evidence across the *K* populations will then be quantified by the upper tailed *P*-value of the exact Binomial test for observing $\sum_k N_{\text{hit}}{}^{(k)}$out of$\sum_k N_{\text{total}}{}^{(k)}$ SNPs when the success probability is given as P_{crit}.

Type 2 diabetes (T2D) data sets

We applied our region-based meta-analysis approach to combine the evidence from three separate genome-wide surveys of type 2 diabetes in the Chinese, South-East Asian Malays and Asian Indians from Singapore. Results from each individual survey and the SNP-based meta-analysis have been reported elsewhere.[20] Briefly, the Chinese GWAS examined 2010 cases and 1945 controls (post-QC) that were typed on a mixture of Illumina (San Diego, CA, USA) 610 (1082 cases/1006 controls) and Illumina1M arrays (928 cases/939 controls). The corresponding numbers for the Malay and Indian GWAS were 794 cases/1240 controls and 977 cases/1169 controls, and these were all genotyped on the Illumina 610 arrays. A genome-wide region-based meta-analysis was first performed between the Chinese data that were genotyped on the two arrays to yield a single set of findings for the Chinese experiment. The three experiments for the different population groups were used as discovery cohorts for a region-based meta-analysis with a window size of 250 kb and a sliding gap of 50 kb such that two consecutive windows have a 200-kb overlap. We also performed a gene-based meta-analysis across 30 037 genes identified from the hg18 version of the TransMap UCSC gene mapping, with each window spanning a 100-kb flanking buffer from the start and end coordinates of each gene. A pathway analysis was also performed for 212 pathways in the KEGG database[21–23] (http://www.genome.jp/kegg/pathway.html). Each gene (inclusive of a 25-kb flanking buffer) in a particular pathway was considered as a distinct window, except for genes within 50 kb of each other, which we merged as one discrete window. The intra-population evidence for a pathway was calculated from the summation of the effective number of independent significant and total SNPs across the windows. The *P*-value threshold (P_{crit}) was set at 0.01. We identified any genomic region that exhibited *P*-value <0.001 in at least two populations from the region-based and gene-based analyses. This is an additional criteria to ensure that at least two populations are contributing to the observed signals, given the fundamental strategy of our approach is to identify genomic regions that are associated with the outcome in multiple populations. We excluded any regions that are known to carry copy number changes as estimation of LD is likely to be inaccurate in these regions. For the pathway analysis, we identified a pathway that exhibited *P*-value <0.05 in at least two populations. To avoid artificial signals of disease association that were the results of erroneous genotype calling, genotyping quality was visually ascertained in each cohort for every SNP located in the discovered regions from the region-based analysis. To validate the findings, similar analyses were performed on the type 2 diabetes data from Phase 1 of the Wellcome Trust Case Control Consortium (WTCC).[19] Calculation of LD in each of the discovery and validation cohorts was performed with 500 control samples from the respective study.

Software implementation

The method described in this paper is implemented in three separate C++ programs: (i) regionalP for performing genome-wide region-based analysis; (ii) regionalP-gene for performing gene-based analysis; and (iii) regionalP-pathway for performing pathway-based analysis. The programs are available from http://www.statgen.nus.edu.sg/~software/regionalP.html.

Descriptions of the set up for the simulations, along with additional methods and analyses are available in the Supplementary Material online.

RESULTS

Power and false-positive rates

We compared our method for regional analysis against standard SNP-based analyses with (i) only the genotyped SNPs or with (ii) the full set of SNPs after imputing against reference panels from phase 2 of the HapMap (HapMap2). In the meta-analysis combining the results from all three populations, the power of the region-based strategy was similar to that from a meta-analysis of the imputed SNPs (Figure 2). This was significantly higher than the power from the meta-analysis of only the genotyped SNPs. The false-positive rates of all three meta-analytic approaches were $<5\%$, although the region-based approach had a near-zero false-positive rate when we imposed an additional restriction requiring at least two populations to exhibit *P*-values of <0.001 in the same region (Supplementary Table S1 online). At a genome-wide significance of 10^{-8}, this additional restriction resulted in only a marginal decrease in statistical power, although this decrease was more substantial at less stringent significance thresholds. Investigating the sensitivity of our method by the allelic spectrum of the simulated causal variants in CEU, we observed the region-based approach was less powerful in identifying low-frequency causal variants (MAF of causal variant $\leq 5\%$) but was marginally more powerful for common causal variants (MAF of causal variant $>5\%$, see Figure 2).

We also explored the performance of the three approaches in the presence of allelic heterogeneity, defined as having different causal variants in the same gene or genomic location across different populations. Specifically, we performed another series of simulations assuming two different causal variants in CEU and JPT+CHB, while allowing YRI to carry either of the two possible causal variants. Our simulation explicitly selected causal variants that are at least 20 kb away but within 50 kb of each other. The region-based approach significantly outperformed both SNP-based approaches in the meta-analyses across CEU and JPT+CHB, particularly at lower Type I errors and when LD between the two causal variants is low (Figure 3). When the LD between the two causal variants is high ($r^2>0.8$), there is almost no difference in the results of the SNP-based meta-analyses of all three populations at higher Type I errors as compared with the power observed in our earlier simulations with only one causal variant. This is reassuring since we expect the two causal variants to behave as effectively a single variant when the LD is high. However, the low power experienced by the SNP-based methods in the presence of two separate causal variants reflects the inadequacy of SNP-based approaches for integrating data across diverse populations, and the greatest merit of the region-based approach is in the presence of allelic heterogeneity between populations where the different causal variants are in weak and non-existent LD.

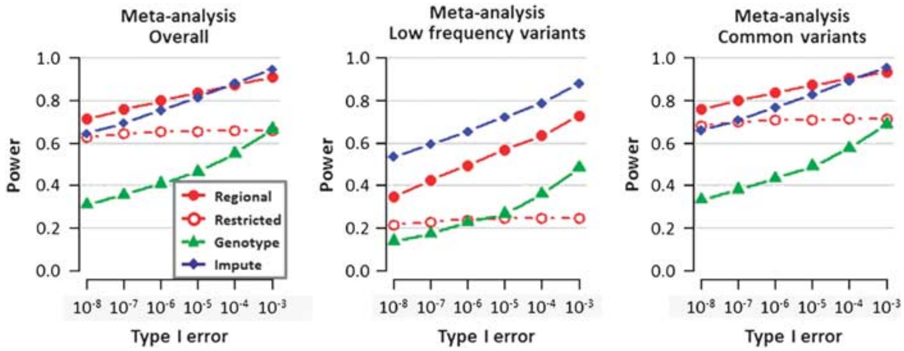

Figure 2 Power comparisons of the different methods for the meta-analysis across all three Hapmap populations. Simulations were performed with HAPGEN (Wellcome Trust Centre for Human Genetics, Oxford, UK) assuming a causal variant that was present in all HapMap phase 2 panels with a multiplicative allelic relative risk of 1.5. The case–control genotype data were subsequently thinned to the SNP content of Affymetrix 500K (CEU simulations), Illumina 1M (JPT+CHB simulations) and Affymetrix 6.0 (Santa Clara, CA, USA) (YRI simulations). We calculate the power when only the genotyped SNPs were considered (green triangles), and when we performed region-based analyses of 100 kb regions in each of the three populations (red circles). Imputation was performed with population-specific haplotypes to recover the SNPs removed from the thinning (except for the causal SNP), and a SNP-based analysis was performed on this denser set of imputed and genotyped SNPs (blue diamonds). The SNP-based meta-analyses considered either the genotyped SNPs present across all three platforms only (green triangles) or across the denser set of imputed and genotyped SNPs common to all three populations (blue diamonds). The region-based meta-analysis was performed without restriction (red circles), and with the restriction that at least two populations display region-based *P*-value <0.001 (red open circles).

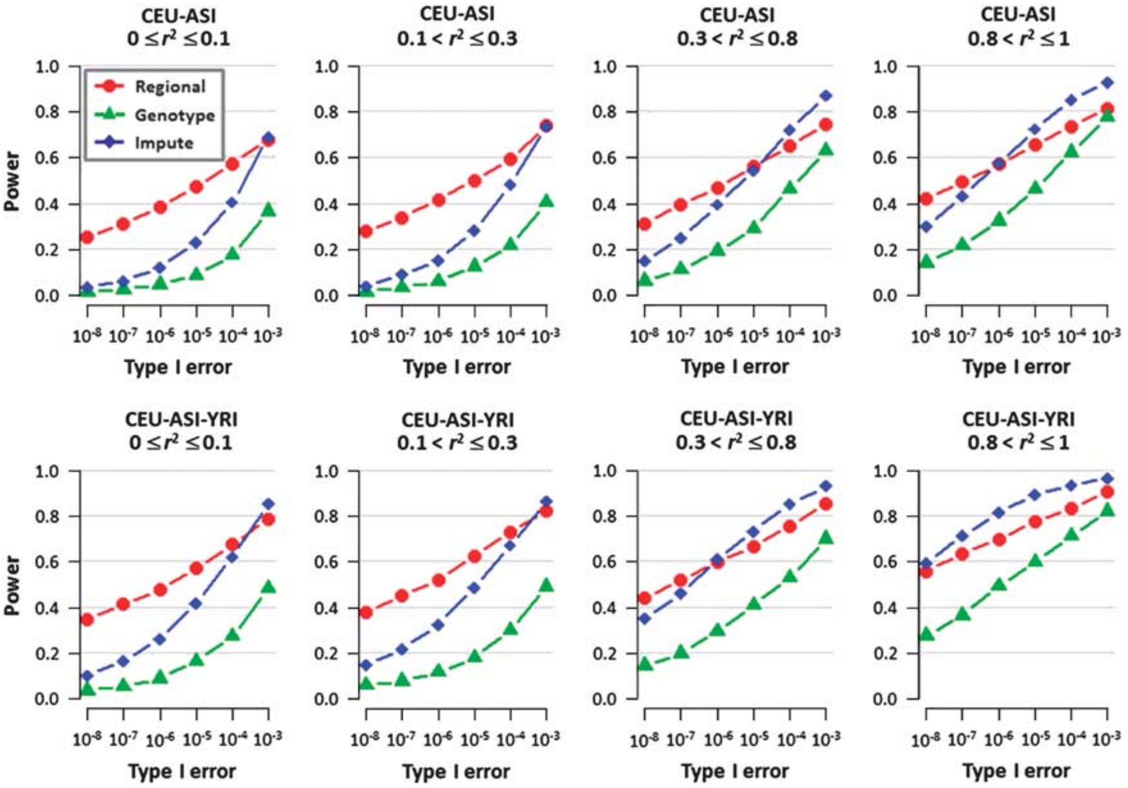

Figure 3 Power comparisons of the different methods for meta-analysis in the presence of allelic heterogeneity. A different causal variant was selected in CEU and JPT+CHB, respectively, while either of the two causal variants was equally likely to be present in the YRI simulations. The two causal variants are located at least 20 kb away but are not >50 kb apart, and have minor allele frequencies of at least 10% in all three HapMap populations. The case–control genotype data simulated from HAPGEN were subsequently thinned to the SNP content of Affymetrix 500K (CEU), Illumina 1M (JPT+CHB) and Affymetrix 6.0 (YRI). We calculated the power when only the CEU and JPT+CHB populations were combined (top row), and when all three HapMap panels were combined (bottom row), investigating the performance of the meta-analysis across the SNPs on all three arrays (green triangles), and for the region-based meta-analysis considering 250 kb regions (red circles). Imputation was performed with population-specific haplotypes to recover the SNPs removed from the thinning, and a SNP-based meta-analysis was performed on this denser set of imputed and genotyped SNPs common to all three populations (blue diamonds). We binned the 3000 pairs of causal variants according to the LD between the two SNPs into four groups: (i) $0 \leq r^2 \leq 0.1$; (ii) $0.1 < r^2 \leq 0.3$; (iii) $0.3 < r^2 \leq 0.8$; (iv) $0.8 < r^2 \leq 1$.

Application to T2D data

We applied our method to perform region-based, gene-based and pathway-based meta-analyses in three independent genome-wide studies of type 2 diabetes (T2D) involving the Chinese, Malays and Asian Indians in Singapore. This was performed across all the autosomal chromosomes within each of the three GWAS in a hypothesis-generating fashion, where for the region-based analyses we considered sliding windows of 250 kb each with a sliding distance of 50 kb such that every pair of consecutive windows overlapped by 200 kb. About half of the Chinese samples were genotyped on the Illumina 1M array, while the remaining half of the Chinese, Malay and Indian samples were genotyped on the Illumina 610 array. Results of the SNP-based meta-analyses using both the genotyped SNPs and the imputed SNPs have been reported elsewhere.[20] Briefly, none of the SNPs achieved genome-wide significance in the meta-analyses, although variants in *CDKAL1* and *HHEX/IDE/KIF11* displayed moderate evidence of T2D association in at least two of the three populations. In particular, variants in *CDKAL1* were found against a genomic background exhibiting substantial LD variations between the populations.[24]

The genome-wide meta-analysis with our region-based method identified five regions exhibiting $P<0.001$ in at least two of the three populations (Table 1). Other than the region on chromosome 6 that encompassed *CDKAL1*, the other four regions did not emerge in the SNP-based meta-analyses[20] (Supplementary Table S2 online). In the replication experiment with the WTCCC data, two of these five regions displayed strong evidence of regional association ($P<10^{-4}$) in the case–control T2D GWAS, which included the stretch on chromosome 6 encompassing *CDKAL1* and the region on chromosome 3 between 21.73 and 22.13 Mb that encompassed *ZNF659*. Suggestive corroborative evidences ($P<0.05$) from WTCCC were also seen in the region on chromosome 2 that spanned the *STK39* gene and the region on chromosome 14 containing the genes *GNG2* and *NID2*. There was no evidence of regional association in the WTCCC for the remaining region on chromosome 20 spanning *STX16* and *NPEPL1*.

Remarkably, all five regions have been previously implicated in diabetes, obesity or other cardiovascular biomarkers. The convincing signal for the region encompassing *CDKAL1* is consistent with established findings for T2D,[25–30] while *ZNF659* has been associated with young-onset type 2 diabetes in the American Indians.[31] The *STK39* gene has been consistently reported to harbor variants implicated in hypertension and in obesity and diabetes-related rodent quantitative trait loci.[32] Previous pathway analysis has identified the G-protein *GNG2* to be associated with type 1 diabetes,[33] suggesting a serotonin modulating mechanism that is similarly relevant in the etiology of type 2 diabetes. Variants in *STX16* have also been reported to significantly slow the reversal of insulin-stimulated glucose transport,[34,35] a biological mechanism that is highly relevant to T2D.

DISCUSSION

The scale of GWMA with diverse European and non-European populations is expected to increase markedly given the popularity of genome-wide designs in studying the genetic etiology of common diseases and complex traits. This, however, increases the challenge of accommodating varying patterns of LD that may exist between genetically diverse populations, which can compromise the ability to reproduce the association signals from surrogate markers that are correlated to the unobserved functional polymorphisms. We have introduced an alternative strategy for combining the evidence across different populations that is robust to dissimilar patterns of LD surrounding a bona fide association signal. The approach is applicable to both case–control studies or in association studies of quantitative traits. Our method has also been shown to perform comparably to imputation-based meta-analysis, except it relies on available genotype

Table 1 Results of the region-based meta-analysis for type 2 diabetes

			Discovery – Single population				*Discovery – Combined*			*Validation from WTCCC1*					
Chromosome	*Start*[a] *(top window)*	*End*[a] *(top window)*	*Pop*[b]	*# Hits*[c]	*# SNP*[d]	P	*# Hits*[c]	*# SNP*[d]	P	*Start*[e]	*End*[e]	*# Hits*[c]	*# SNP*[d]	P	*Gene*
2	168 408 674 (168 458 674)	168 858 674 (168 708 674)	**C**	**4.9**	**43**	**1.49×10^{-4}**	11.2	90	1.60×10^{-9}	168 758 674 (168 808 674)	169 208 674 (169 058 674)	2.2	21	1.47×10^{-2}	STK39
			M	**6.3**	**23**	**6.03×10^{-8}**									
			I	0	24	1									
3	21 736 044 (21 786 044)	22 136 044 (22 036 044)	**C**	**6.2**	**103**	**4.91×10^{-4}**	13.6	221	2.55×10^{-7}	21 186 044 (21 286 044)	21 636 044 (21 536 044)	4.5	25	6.06×10^{-5}	ZNF659
			M	**7.3**	**58**	**1.36×10^{-6}**									
			I	0	60	1									
6	20 594 609 (20 594 609)	20 894 609 (20 844 609)	**C**	**4.1**	**41**	**6.56×10^{-4}**	10.1	107	9.68×10^{-7}	20 494 609 (20 594 609)	21 044 609 (20 844 609)	7.3	20	5.03×10^{-10}	CDKAL1
			M	0	28	1									
			I	**6**	**39**	**2.42×10^{-6}**									
14	51 355 752 (51 455 752)	51 755 572 (51 705 752)	C	3	67	2.91×10^{-2}	19.7	146	2.92×10^{-16}	50 955 752 (51 105 752)	51 405 752 (51 355 752)	2.5	30	1.99×10^{-2}	GNG2, NID2
			M	**7.5**	**34**	**2.24×10^{-8}**									
			I	**9.2**	**45**	**5.03×10^{-10}**									
20	56 559 795 (56 609 795)	56 859 795 (56 859 795)	**C**	**5**	**75**	**9.22×10^{-4}**	11.1	173	1.56×10^{-6}	56 609 795	56 859 795	0.9	25	0.294	STX16, NPEPL1
			M	0	46	1									
			I	**6.1**	**52**	**1.20×10^{-5}**									

Genomic regions identified by the region-based analysis, with the discovery mechanism based on three genome-wide association studies conducted in Chinese, Malays and Asian Indians in Singapore. Validation of the regions that emerged was performed on the type 2 diabetes case–control study from Phase 1 of the Wellcome Trust Case–Control Consortium (WTCCC).

[a] The start and end positions of the genomic region containing consecutive windows with $P<0.001$ in at least two of the populations (in bold). The start and end positions of the top 250 kb window are shown in brackets. Subsequent columns show the evidence for the discovery populations in the top window.

[b] The three discovery populations abbreviated: C, SP2 Chinese; M, SiMES Malays; I, SINDI Indians.

[c] Effective number of independent SNPs with $P<0.01$ after accounting for LD.

[d] Effective number of independent SNPs across the region after accounting for LD.

[e] The start and end positions of the genomic region containing consecutive windows with evidence of validation (defined as $P<0.05$), with the start and end positions of the top 250 kb being shown in brackets. Subsequent columns show the evidence for WTCCC1 in the top window. For regions without any 250 kb windows displaying $P<0.05$, the best window in that region is shown instead.

information from the experiment without requiring additional reference data from appropriately matched populations. In the presence of allelic heterogeneity, our approach outperforms both SNP-based approaches using either genotyped or imputed SNPs. The application of the region-based method to three genome-wide surveys in T2D resulted in the discovery of novel and established regions that are subsequently validated with data from the WTCCC.

The region-based approach relies on the elegant application of the concept of statistical significance in evaluating a genomic region for evidence of trait association. For example, under the null hypothesis that the region is independent of the phenotype, we expect 5% of the SNPs to be statistically significant by chance when adopting a *P*-value threshold of 5%, if indeed all the SNPs in this region are mutually independent. If this assumption of mutual independence is true, an over-representation of statistically significant SNPs in this region constitutes evidence that this region is associated with the phenotype, with the extent of over-representation indicating the strength of the evidence. This is analogous to the use of 5×10^{-8} as the definition of genome-wide significance for assessing the likely authenticity of single markers. LD between the SNPs can confound the measurement of over-representation, as this can either inflate the number of significant signals, which increases false positives, or produce an inflated estimate of the total number of SNPs, which decreases statistical power. The eigen-decomposition of the LD matrix allows the effective number of independent SNPs to be estimated and consequently, also the effective number of independent association signals that are statistically significant. By surveying the same genomic region across different independent populations, the same statistical framework can be extended to consolidate the evidence from multiple populations, simply by summing the effective number of independent SNPs and signals across these populations and assessing the evidence for an over-representation of significant signals. This provides a simple but, yet, effective solution to combining the results from experiments that use different genotyping platforms. By searching for the same regions rather than the same SNPs to emerge in the different GWAS, inter-population variation in LD patterns between the assayed SNPs and the causal variant is expected to have lesser impact on the sensitivity of our approach.

One feature of our method is the ability to sharpen the association evidence in regions containing multiple weak signals across different ethnic groups. These signals may be weaker as a result of SNP ascertainment biases in the design of genotyping arrays, resulting in weaker LD between the assayed SNPs and the causal variants. The current definition of genome-wide significance excludes many potential signals to be considered in a bid to protect against the abundance of false discoveries that is associated with testing in excess of a million hypotheses. This poses a significant challenge to genome-wide studies and GWMA in populations with short LD, such as African populations,[15,36] as it is less likely for variants to be in sufficient LD to exhibit statistical evidence stronger than the stringent threshold. Furthermore, the greater genetic diversity that is common of such populations means it is not immediately straightforward to compensate for the lower LD by increasing the effective sample size through a meta-analysis of several populations. Our method thus provides a viable solution within a sound statistical framework to exploit and combine the evidence from SNPs that are weakly associated with the phenotype.

The application of analytical methods that investigate regions in the genome rather than relying on individual SNPs is not a new concept. Neither is implementing a statistical strategy to estimate the effective number of independent association tests in the presence of LD. Numerous approaches have in fact been introduced to address the issue of multiple testing in the presence of correlated SNPs.[37–43] However, these methods either assign the most significant SNP-based evidence as the statistical evidence for the set of loci,[39] or do not explicitly incorporate the association evidence in adjusting for the effective number of tests.[38,40–43] A recent region-based approach adopted a more sophisticated approach that borrows information from surrounding SNPs, although it tends to rely on heuristic measures such as the proximity to specific genomic features (eg, known genes, evolutionarily conserved regions and haplotype blocks) for defining SNP clusters.[44] In our opinion, the imputation frameworks that MACH[12] and IMPUTE[13] are built on provide a more natural way to incorporate information from surrounding SNPs without relying on pre-defined features that may not adequately account for the correlation between SNPs. We thus benchmarked our method against the performance of the imputation-based approach, which has become the strategy of choice in recent genome-wide studies. More importantly, neither of the previous region-based approaches provide a natural solution to integrate the evidence across multiple genome-wide studies in a meta-analysis, nor adequately manage the complexity due to allelic heterogeneity.

We have proposed a novel and powerful strategy for querying the genome for genotype–phenotype associations that realistically manages the challenges imposed by the fundamental design of genome-wide studies and in combining several such studies from diverse populations. We envisage this approach has the potential to be further developed for burden-related tests of rare or low-frequency variants across multiple heterogeneous populations, which is an emerging issue given the increasing popularity of exome-sequencing experiments across numerous traits.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ACKNOWLEDGEMENTS

This project acknowledges the support of the Yong Loo Lin School of Medicine from the National University of Singapore, the National Medical Research Council (NMRC) 0796/2003, Singapore and the Biomedical Research Council (BMRC), 09/1/35/19/616, Singapore. The study used data generated by the WTCCC, the International HapMap Consortium and the Singapore Genome Variation Project. A full list of the investigators who contributed to the generation of the WTCCC data are available from http://www.wtccc.org.uk. Funding for the WTCCC project was provided by the Wellcome Trust under award 076113. YYT acknowledges support from the National Research Foundation, NRF-RF-2010-05, Singapore.

AUTHOR CONTRIBUTIONS

YYT conceived, designed and directed the experiment; YYT, XW and CCK wrote the paper; XW, XL and WTP implemented the C++ software; XW, HX, XS, CS, RTHO, CCK and YYT analyzed the data; DPKN, JL, TA, KSC, EST and TYW contributed samples.

1 Donnelly P: Progress and challenges in genome-wide association studies in humans. *Nature* 2008; **456**: 728–731.
2 McCarthy MI, Abecasis GR, Cardon LR *et al*: Genome-wide association studies for complex traits: consensus, uncertainty and challenges. *Nat Rev Genet* 2008; **9**: 356–369.
3 Kathiresan S, Willer CJ, Peloso GM *et al*: Common variants at 30 loci contribute to polygenic dyslipidemia. *Nat Genet* 2009; **41**: 56–65.
4 Lango Allen H, Estrada K, Lettre G *et al*: Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* 2010; **467**: 832–838.
5 Lindgren CM, Heid IM, Randall JC *et al*: Genome-wide association scan meta-analysis identifies three Loci influencing adiposity and fat distribution. *PLoS Genet* 2009; **5**: e1000508.

6 Voight BF, Scott LJ, Steinthorsdottir V *et al*: Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis. *Nat Genet* 2010; **42**: 579–589.
7 Zeggini E, Scott LJ, Saxena R *et al*: Meta-analysis of genome-wide association data and large-scale replication identifies additional susceptibility loci for type 2 diabetes. *Nat Genet* 2008; **40**: 638–645.
8 Manolio TA, Collins FS, Cox NJ *et al*: Finding the missing heritability of complex diseases. *Nature* 2009; **461**: 747–753.
9 Stephens M, Balding DJ: Bayesian statistical methods for genetic association studies. *Nat Rev Genet* 2009; **10**: 681–690.
10 Teo YY, Ong RT, Sim X, Tai ES, Chia KS: Identifying candidate causal variants via trans-population fine-mapping. *Genet Epidemiol* 2010; **34**: 653–664.
11 Browning SR, Browning BL: Rapid and accurate haplotype phasing and missing-data inference for whole-genome association studies by use of localized haplotype clustering. *Am J Hum Genet* 2007; **81**: 1084–1097.
12 Li Y, Willer CJ, Ding J, Scheet P, Abecasis GR: MaCH: using sequence and genotype data to estimate haplotypes and unobserved genotypes. *Genet Epidemiol* 2010; **34**: 816–834.
13 Marchini J, Howie B, Myers S, McVean G, Donnelly P: A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat Genet* 2007; **39**: 906–913.
14 Jallow M, Teo YY, Small KS *et al*: Genome-wide and fine-resolution association analysis of malaria in West Africa. *Nat Genet* 2009; **41**: 657–665.
15 Teo YY, Small KS, Kwiatkowski DP: Methodological challenges of genome-wide association analysis in Africa. *Nat Rev Genet* 2010; **11**: 149–160.
16 Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R: Ascertainment bias in studies of human genome-wide polymorphism. *Genome Res* 2005; **15**: 1496–1502.
17 Frazer KA, Ballinger DG, Cox DR *et al*: A second generation human haplotype map of over 3.1 million SNPs. *Nature* 2007; **449**: 851–861.
18 Rosenberg NA, Huang L, Jewett EM, Szpiech ZA, Jankovic I, Boehnke M: Genome-wide association studies in diverse populations. *Nat Rev Genet* 2010; **11**: 356–366.
19 Wellcome Trust Case Control Consortium: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 2007; **447**: 661–678.
20 Sim X, Ong RT, Suo C *et al*: Transferability of type 2 diabetes implicated Loci in multi-ethnic cohorts from southeast Asia. *PLoS Genet* 2011; **7**: e1001363.
21 Kanehisa M, Goto S, Furumichi M, Tanabe M, Hirakawa M: KEGG for representation and analysis of molecular networks involving diseases and drugs. *Nucleic Acids Res* 2010; **38**: D355–D360.
22 Kanehisa M, Goto S: KEGG: kyoto encyclopedia of genes and genomes. *Nucleic Acids Res* 2000; **28**: 27–30.
23 Kanehisa M, Goto S, Hattori M *et al*: From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Res* 2006; **34**: D354–D357.
24 Teo YY, Sim X, Ong RT *et al*: Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. *Genome Res* 2009; **19**: 2154–2162.
25 Han X, Luo Y, Ren Q *et al*: Implication of genetic variants near SLC30A8, HHEX, CDKAL1, CDKN2A/B, IGF2BP2, FTO, TCF2, KCNQ1, and WFS1 in type 2 diabetes in a Chinese population. *BMC Med Genet* 2010; **11**: 81.
26 Saxena R, Voight BF, Lyssenko V *et al*: Genome-wide association analysis identifies loci for type 2 diabetes and triglyceride levels. *Science* 2007; **316**: 1331–1336.
27 Scott LJ, Mohlke KL, Bonnycastle LL *et al*: A genome-wide association study of type 2 diabetes in Finns detects multiple susceptibility variants. *Science* 2007; **316**: 1341–1345.
28 Sladek R, Rocheleau G, Rung J *et al*: A genome-wide association study identifies novel risk loci for type 2 diabetes. *Nature* 2007; **445**: 881–885.
29 Takeuchi F, Serizawa M, Yamamoto K *et al*: Confirmation of multiple risk Loci and genetic impacts by a genome-wide association study of type 2 diabetes in the Japanese population. *Diabetes* 2009; **58**: 1690–1699.
30 Zeggini E, Weedon MN, Lindgren CM *et al*: Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes. *Science* 2007; **316**: 1336–1341.
31 Hanson RL, Bogardus C, Duggan D *et al*: A search for variants associated with young-onset type 2 diabetes in American Indians in a 100K genotyping array. *Diabetes* 2007; **56**: 3045–3052.
32 Wang Y, O'Connell JR, McArdle PF *et al*: From the cover: whole-genome association study identifies STK39 as a hypertension susceptibility gene. *Proc Natl Acad Sci USA* 2009; **106**: 226–231.
33 Torkamani A, Topol EJ, Schork NJ: Pathway analysis of seven common diseases assessed by genome-wide association. *Genomics* 2008; **92**: 265–272.
34 Perera HK, Clarke M, Morris NJ, Hong W, Chamberlain LH, Gould GW: Syntaxin 6 regulates Glut4 trafficking in 3T3-L1 adipocytes. *Mol Biol Cell* 2003; **14**: 2946–2958.
35 Smith EN, Chen W, Kahonen M *et al*: Longitudinal genome-wide association of cardiovascular disease risk factors in the Bogalusa Heart Study. *PLoS Genet* 2010; **6**: pii e1001094.
36 Campbell MC, Tishkoff SA: African genetic diversity: implications for human demographic history, modern human origins, and complex disease mapping. *Annu Rev Genomics Hum Genet* 2008; **9**: 403–433.
37 Cheverud JM: A simple correction for multiple comparisons in interval mapping genome scans. *Heredity* 2001; **87**: 52–58.
38 Nyholt DR: A simple correction for multiple testing for single-nucleotide polymorphisms in linkage disequilibrium with each other. *Am J Hum Genet* 2004; **74**: 765–769.
39 Lin DY: An efficient Monte Carlo approach to assessing statistical significance in genomic studies. *Bioinformatics* 2005; **21**: 781–787.
40 Moskvina V, Schmidt KM: On multiple-testing correction in genome-wide association studies. *Genet Epidemiol* 2008; **32**: 567–573.
41 Pan W: Asymptotic tests of association with multiple SNPs in linkage disequilibrium. *Genet Epidemiol* 2009; **33**: 497–507.
42 Li J, Ji L: Adjusting multiple testing in multilocus analyses using the eigenvalues of a correlation matrix. *Heredity* 2005; **95**: 221–227.
43 Mukhopadhyay I, Feingold E, Weeks DE, Thalamuthu A: Association tests using kernel-based measures of multi-locus genotype similarity between individuals. *Genet Epidemiol* 2010; **34**: 213–221.
44 Wu MC, Kraft P, Epstein MP *et al*: Powerful SNP-set analysis for case-control genome-wide association studies. *Am J Hum Genet* 2010; **86**: 929–942.

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