Twin Research and Human Genetics
Volume 20 ▮ Number 1 ▮ pp. 1–9 ▮ © The Author(s) 2017 ▮ doi:10.1017/thg.2016.100

A Novel Approach for Pathway Analysis of GWAS Data Highlights Role of BMP Signaling and Muscle Cell Differentiation in Colorectal Cancer Susceptibility

Aniket Mishra and Stuart MacGregor
Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO) and the Colorectal Cancer Family Registry (CCFR), and Statistical Genetics, QIMR Berghofer Medical Research Institute, Brisbane, Queensland, Australia

Genome-wide association studies (GWAS) have revolutionized the field of gene mapping. As the GWAS field matures, it is becoming clear that for many complex traits, a proportion of the missing heritability is attributable to common variants of individually small effect. Detecting these small effects individually can be difficult, and statistical power would be increased if relevant variants could be grouped together for testing. Here, we propose a VEGAS2Pathway approach that aggregates association strength of individual markers into pre-specified biological pathways. It accounts for gene size and linkage disequilibrium between markers using simulations from the multivariate normal distribution. Pathway size is taken into account via a resampling approach. Importantly, since the approach only requires summary data, the method can easily be applied in all GWASs, including meta-analysis, singleton-based, family-based, and DNA-pooling-based designs. This approach is implemented in a user-friendly web page https://vegas2.qimrberghofer.edu.au and a command line tool. The web implementation uses gene-sets from the gene ontology (GO), curated gene-sets from MSigDB (containing canonical pathways and gene-sets from BIOCARTA, REACTOME, KEGG databases), PANTHER, and pathway commons databases, enabling analysis of a wide range of complex traits. We applied this method on a colorectal cancer GWAS meta-analysis data set (10,934 cases, 12,328 controls) from the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO). We report statistically significant enrichment of association signal for the 'BMP signaling' and 'muscle cell differentiation' pathways, suggesting a possible role for these pathways onto the risk of colorectal cancer.

■ **Keywords:** GWAS, VEGAS2Pathway, VEGAS, pathway analysis, colorectal cancer

Genome-wide association studies (GWASs) have substantially improved our understanding of the genetic basis of various complex human phenotypes. GWAS typically tests the association of each common (minor allele frequency >0.01 or 0.05) single nucleotide polymorphism (SNP) with a trait of interest. To control false positives, frequently only SNPs with association p value less than the genome-wide significance threshold (p value $< 5 \times 10^{-8}$) are reported, with relatively little attention paid toward the remaining SNPs. Since relatively few SNPs will reach genome-wide significance in a given study, it may not be clear that a subset of them act within a particular pathway. By also considering additional sub-threshold SNPs, then it may be possible to (1) prioritize SNPs for follow-up based on the pathway they lie in, and (2) to better define the pathways involved in the trait, leading to insight into the underlying molecular mechanisms.

Pathway-centric approaches that test the association of the combined effects of variants in a set of biologically or functionally related genes are becoming increasingly popular as a complementary method to GWAS (Pers et al., 2015; Wang et al., 2009; Wang et al., 2010). Pathway-based association strategies have several advantages, but further methodological development is required. In the past few years, several reviews have been published discussing the issues related to pathway-based association tests (Robinson et al., 2014; Wang et al., 2010; 2011). Some of the potential

RECEIVED 13 September 2016; ACCEPTED 17 November 2016

ADDRESS FOR CORRESPONDENCE: Aniket Mishra, Statistical Genetics, QIMR Berghofer Medical Research Institute, 300 Herston Rd, Brisbane QLD 4006, Australia. E-mail: aniket.mishra@qimrberghofer.edu.au

https://doi.org/10.1017/thg.2016.100 Published online by Cambridge University Press

problems are as follows: (1) inadequate modeling of the linkage disequilibrium (LD) pattern within a gene, and (2) gene-set size and the number of SNPs in a gene may influence the significance of a gene-set. Here, we report VEGAS2Pathway, a versatile pathway-based approach for GWAS data, and demonstrate through simulations how it appropriately accounts for LD between SNPs, gene size, and pathway size. We report its implementation in a user-friendly web page and in command line software.

We applied our method on colorectal cancer GWAS summary data obtained from the Genetics and Epidemiology of Colorectal Cancer Consortium (GECCO; Peters et al., 2013) and Colorectal Cancer Family Registry (total sample size: 10,934 cases and 12,328 controls). Fifty variants so far reported are associated with colorectal cancer susceptibility (Peters et al., 2015). Together, these variants explain only small proportion of the heritability of colorectal cancer (Jiao et al., 2014).

Materials and Methods

VEGAS2Pathway Strategy

VEGAS2Pathway is a two-step pathway analysis strategy. First, it calculates the gene-based test statistics for all genes using the VEGAS (VErsatile Gene-based Association Study) approach (Liu et al., 2010; Mishra & MacGregor, 2015), which accounts for the LD between the SNPs within a gene through simulation. Second, for each of a set of pre-specified gene-sets, the relevant gene-based results are carried forward to compute a pathway-based test. To give users a pathway analysis platform useful for a wide range of complex traits, we incorporated gene-sets from the Gene Ontology (GO; Gene Ontology Consortium, 2008) curated gene-sets from MSigDB (Liberzon et al., 2011; containing canonical pathways and gene-sets from BIOCARTA, REACTOME, KEGG databases), PANTHER (Thomas et al., 2003), and pathway commons (Cerami et al., 2011) databases. We filtered these gene-sets to include only those with size between 10 and 1,000 genes. Overall, there are 6,212 gene-sets, including 18,399 genes with 511,336 annotations. To compute the pathway-based test, the relevant gene-based p values were first converted to upper-tail χ^2 statistics with one degree of freedom, before summing. This process of summing is potentially an informative approach for polygenic traits because each gene contributes to the pathway results concomitant with strength of evidence for association at that gene. Contrast this with the hypergeometric-based pathway approaches (Lee et al., 2012; Pers et al., 2015), where genes are either ‘in’ or ‘out’ of the pathway, based on them exceeding a particular p value. Generating accurate significance levels for very strongly associated genes is difficult because the simulation derived gene-based p value can be zero (in the default scenario this means 0 of the 1×10^6 simulation replicates exceeded the result for the real data). In such cases, we allocate a p value of 1×10^{-6} for such genes. This approach may underestimate the importance of such genes but has the advantage of ensuring the resultant pathway results are not driven entirely by one or a few genes.

Under a polygenic model and assuming a non-competitive test, larger pathways will typically be more significant than smaller pathways. Since we believe the genetic architecture for many complex traits will be polygenic, a non-competitive test is unlikely to provide useful results (as large pathways will be much more likely to be significant than small ones due to the accumulation of polygenes). Instead, we implemented a competitive test in which each pathway is benchmarked against the ‘typical’ pathway of the same size. Specifically, we corrected for pathway size bias by adopting a resampling approach where the same number of genes as present in a pathway are repeatedly drawn at random from all set of genes used in the study and summed. The empirical p value of association for a pathway is calculated using following formula:

$$\text{Emp } P = \frac{\sum_{1}^{N} I\left(\chi^{2*} \geq \chi^{2}\right) + 1}{N + 1},$$

where N is the number of resamples performed. The χ^{2*} is the summed χ^2 statistics computed per resample, which is compared against the observed summed χ^2 for pathway under consideration. *I()* is an indicator function. Figure 1 describes the schematic representation of the VEGAS2Pathway strategy.

Simulation to Establish Properties of VEGAS2Pathway

We applied VEGAS2Pathway on 1,000 simulated quantitative phenotypes with standard normal distribution $N(0, 1)$. We extracted the top-pathway test statistics from 1,000 simulations to allow estimation of a significance threshold (assessment of how differently the results for the 6,212 pathways behave relative to the situation where, for example, there were 6,212 completely independent pathways). We also used the set of 1,000 simulations to assess the correlation between gene size and gene p value, and between pathway size and pathway p value.

Application on GWAS Summary Data

We applied VEGAS2Pathway approach on colorectal cancer GWAS summary data obtained from GECCO (Peters et al., 2013) and the Colorectal Cancer Family Registry. This meta-analysis of GWASs was performed across 11 studies totaling 10,934 cases and 12,328 controls of European ancestry. Sample selection and genotyping methods for participating studies have been previously described (Peters et al., 2013). Please refer to Supplementary Text 1 for brief description on participating studies and GWAS methodology.

We used GWAS meta-analysis summary data to perform the pathway analysis. The 1000 Genomes phase 1 European data was used as a reference for VEGAS2 (Mishra & MacGregor, 2015) joint-SNP gene-based analysis. The

https://doi.org/10.1017/thg.2016.100 Published online by Cambridge University Press

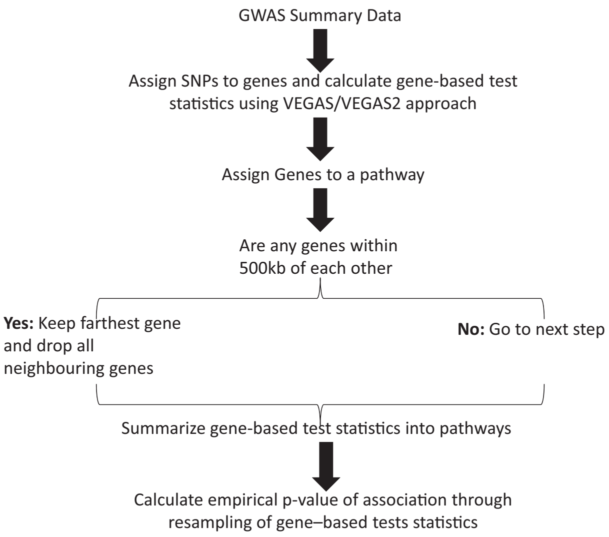
GWAS Summary Data
Assign SNPs to genes and calculate gene-based test statistics using VEGAS/VEGAS2 approach
Assign Genes to a pathway
Are any genes within 500kb of each other
Yes: Keep farthest gene and drop all neighbouring genes
No: Go to next step
Summarize gene-based test statistics into pathways
Calculate empirical p-value of association through resampling of gene–based tests statistics

FIGURE 1

Schematic representation of VEGAS2Pathway strategy: After reading in the two-column (SNP id and GWAS *p* value) text file, the gene-based test statistics are calculated using the VEGAS approach. These genes are then assigned to pathways. If a pathway contains a gene within another gene then the smaller gene is dropped. In another scenario, if a pathway contains overlapping or genes less than 500 kb away then only one gene (chosen randomly) is used to represent the association signal of the region while all nearby genes are dropped. In this way, VEGAS2Pathway ensures that all gene-based test statistics assigned to a pathway are at least 500 kb away from each other. The gene-based test statistics are then summed to compute a statistic for each pathway. An empirical *p* value for each pathway is computed by comparing the summed set score against that of replicates of the same size obtained through resampling of gene-based test statistics.

50 kb gene-boundary option was used to include the cis-regulatory variants. The computational burden for VEGAS2 increases dramatically with an increase in the number of SNPs per gene, and hence for a gene containing more than 1,000 variants it successively prunes the list of variants with r^2 criteria of 0.99, 0.90, 0.70, and 0.50. After each pruning interval, VEGAS2 checks the number of pruned SNPs. If the number of pruned SNPs is less than 1,000, then VEGAS2 uses the pruned SNPs from that interval to perform gene-based analysis; otherwise, it iteratively applies an increasingly stringent r^2 criterion on all SNPs in the gene. After applying a pruning criteria of $r^2 = 0.50$, it uses all pruned SNPs for analysis, irrespective of the number. Once the gene-based test statistics were calculated, we performed pathway analyses using the resampling approach.

Comparison with other Pathway Analysis Approaches

We analyzed the top five pathways identified by VEGAS2Pathway, using the pathway analysis approaches implemented in the INRICH (Lee et al., 2012) and MAGENTA (Segre et al., 2010) software. We used the same gene-pathway annotations as used in VEGAS2Pathway analysis to perform pathway analysis using INRICH and MAGENTA. The INRICH program is based on the hypergeometric pathway analysis approach that compares the number of significant versus non-significant genes in a pathway, and control for type 1 error through permutation (Lee et al., 2012). We created the interval file using plink (with criteria: –clump-p1 1×10^{-4} –clump-p2 0.05 –clump-r2 0.5 –clump-range-border 20). The 1000 Genomes European reference data were used to compute pair-wise LD. We ran INRICH with default settings except the number of permutations parameter (-*r*) was changed to 1,000,000. The MAGENTA program is an extension of gene-set enrichment analysis (GSEA), a widely used approach for pathway analysis of gene expression data. It is based on a weighted Kolmogorov–Smirnov-like running sum statistic algorithm, which tests whether a gene-set is more enriched with highly ranked gene scores than would be expected by chance (Segre et al., 2010). We ran MAGENTA using default settings.

We also tested whether VEGAS2Pathway implicates the top pathways obtained using MAGENTA and INRICH programs. We performed pathway analysis on colorectal cancer GWAS data considering aforementioned 6,213 gene-sets using INRICH and MAGENTA. For INRICH, we first performed pathway analysis using default settings and re-analyzed the top 10 pathways by changing the number of permutations parameter (-*r*) to 1,000,000. We used default settings to perform pathway analysis using MAGENTA.

Results

Establishing Type 1 Error Cut-Off

The pathway annotations used for VEGAS2pathway analysis comprises the pathways from different sources, including hierarchical gene-sets from GO, curated gene-sets from MSigDB, PANTHER, and pathway commons. Hence, there is a frequent overlap in the set of genes included in different pathways. Since these pathways are not independent, application of traditional, multiple testing methods such as Bonferroni correction, Šidák correction, and false discovery rate (FDR) procedure would be overly conservative. To estimate the significance threshold correcting for multiple association tests performed for these overlapping pathways, we applied the VEGAS2Pathway approach on GWAS summary files from 1,000 simulated quantitative phenotypes with standard normally distributed trait values. Figure 2A shows the distribution of -log10 *p* values in the top pathway for 1,000 replicates of simulated phenotypes. The simulation-derived 95% empirical significance threshold for VEGAS2Pathway association test, taking into account the multiple testing of 6,213 correlated pathways, was 1.00×10^{-5} (50 of 1,000 replicates exceeded this). By using this empirical threshold, we are correcting for 5,000 effectively independent tests (since $0.05/5000 = 1.00 \times 10^{-5}$). In comparison, if we were to Bonferroni correct for all

https://doi.org/10.1017/thg.2016.100 Published online by Cambridge University Press

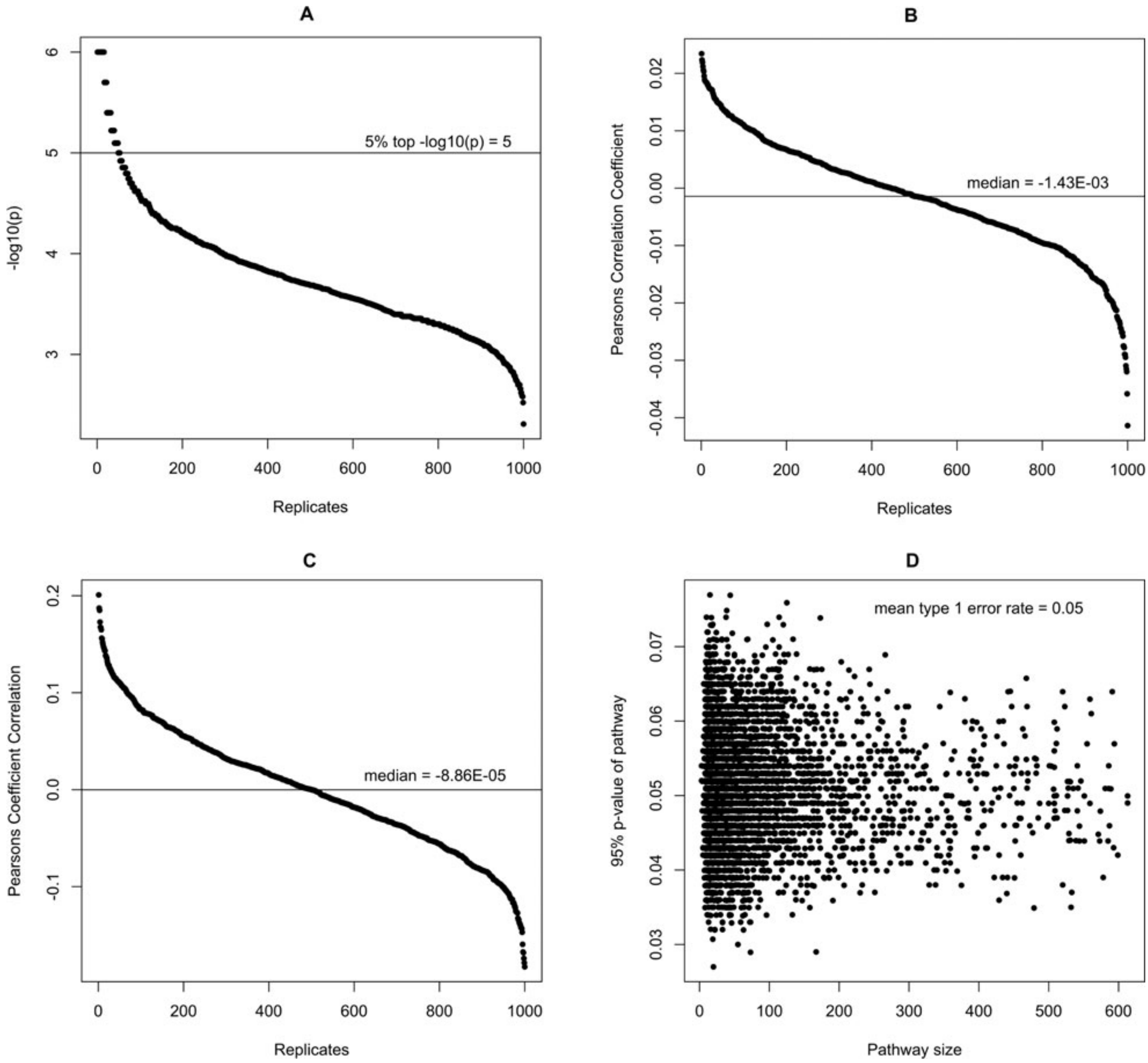

FIGURE 2

Simulation results. (A) The distribution of absolute log10 *p* values of 1,000 replicates. The line represent 5% interval of top -log10 *p* values. (B) The distribution of Pearson's correlation between gene-based *p* value and number of SNPs in the gene. (C) The distribution of Pearson's correlation between pathway-based *p* value and number of SNPs in the pathway. (D) Pathway type 1 error rate versus pathway size.

6,213 pathways, our threshold for significance would be $0.05/6213 = 8.02 \times 10^{-6}$.

The effect of both gene size and pathway size are important to consider in pathway analysis. If there is a positive correlation between a gene's size and its gene-based *p* value, then pathways with an excess of bigger genes in them will repeatedly appear as significant in pathway analysis. Similarly, problems may also occur if pathway size is correlated to pathway significance. VEGAS2Pathway uses the VEGAS approach to perform gene-based analysis. In addition to accounting for LD between SNPs, we expect that the VEGAS approach will also deal appropriately with gene size. To account for pathway size, VEGAS2Pathway compares each gene-set with multiple resamples of the sets with the same number of genes, an approach that should deal appropriately with different pathway sizes. To assess performance of VEGAS2Pathway in practice, we calculated the Pearson's correlation coefficient between gene size and gene-based *p* value, and pathway size and empirical *p* value of pathway association. Figures 2B and 2C show the correlation distribution between and gene-size and gene-based *p* value, and pathway-size and pathway-based *p* value respectively in 1,000 replicates. The correlation is typically close to zero, both between gene size and gene-based

https://doi.org/10.1017/thg.2016.100 Published online by Cambridge University Press

TABLE 1
Top Five Pathways from Pathway Analysis of Colorectal Cancer GWAS Summary Data

Pathway ID	Pathway size (# of genes)	Empirical *p* value	Empirical *p* values after removing genes with genome-wide significant SNPs
GO:0060393_Regulation_of_pathway-restrictedSMAD_protein_phosphorylation	18	$<1.00 \times 10^{-6}$	1.14×10^{-3}
PID_BMPPATHWAY	40	2.00×10^{-6}	4.80×10^{-3}
GO:0042692_Muscle_cell_differentiation	86	8.00×10^{-6}	7.80×10^{-5}
Panther_TGF-beta_signaling_pathway	57	3.00×10^{-5}	2.70×10^{-3}
GO:0048729_Tissue_morphogenesis	156	4.40×10^{-5}	6.60×10^{-3}

p value (median -1.43 × 10^{-3}, with 90% of simulation replicates in a -0.01 to +0.01 range), and between pathway size and pathway-based *p* value (median 2.16 × 10^{-5}, with 90% of simulation replicates in a -0.08 to +0.08 range). Furthermore, we checked how the type 1 error rate varied by pathway size. Figure 2D shows that the type 1 error rate is independent of the pathway size and mean type 1 error rate is preserved to 0.05.

Application of VEGAS2Pathway on Colorectal Cancer GWAS Summary Data

We performed pathway analysis on colorectal cancer GWAS summary data using the VEGAS2Pathway approach. Three pathways reached the genome-wide, pathway-based significant *p* value of less than 1.00 × 10^{-5}, which are GO:0060393_regulation_of_pathway-restricted_SMAD_protein_phosphorylation (pathway-based *p* value < 1.00 × 10^{-6}), PID_BMPPATHWAY (pathway-based *p* value = 2.00 × 10^{-6}) and GO:0042692_muscle_cell_differentiation (pathway-based *p* value = 8.00 × 10^{-6}). Overall, the top five pathways pointed toward the Sma/Mad-related protein (SMAD), bone morphogenetic protein (BMP), and transforming growth factor beta (TGFβ) signaling and muscle cell differentiation pathways. Table 1 shows the top five pathways observed in pathway analysis.

The GO:0060393_regulation_of_pathway-restricted_SMAD_protein_phosphorylation pathway consists of 18 genes, of which six show a gene-based *p* value less than .05. Sixteen of the 18 genes in this pathway contain a top-SNP with a *p* value less than .05 (refer to Supplementary Table 1 for gene-based test statistics of genes in the GO:0060393_regulation_of_pathway-restricted_SMAD_protein_phosphorylation pathway). The *SMAD7* and *BMP4* genes are the largest contributors to the result for this pathway. After removing all genes with top SNPs showing *p* value less 5.00 × 10^{-8}, the pathway still shows evidence for an association (pathway-based *p* value = 1.14 × 10^{-3}), showing that sub-genome-wide threshold genes make an important contribution to this pathway result. Germline mutations in *BMPR1A* (gene-based *p* value = 3.77 × 10^{-4}, top SNP rs71503853 *p* value = 2.91 × 10^{-6}) have been reported to be associated with juvenile polyposis syndrome (JPS; Yamaguchi et al., 2014; OMIM 174900), which is a risk factor for colorectal cancer (Brosens et al., 2007; Howe et al., 1998). Another gene, *SMAD6* (gene-based *p* value = 5.00 × 10^{-3}, top SNP rs3809571 *p* value = 9.70 × 10^{-5}), is involved in lung cancer cell growth and survival (Jeon et al., 2008). Two further genes with gene-based *p* values less than .05 are *BMP2* and *BMP7*. *BMP2* is located 342 kb away from the colorectal cancer associated 38 kb LD region at 20p12.3 (Study et al., 2008). Overexpression of *BMP2* is a risk factor for survival of non-small cell lung cancer patients (Chu et al., 2014). *BMP7* influences proliferation, migration and invasion of lung (Liu et al., 2012), and breast cancer cells (Alarmo et al., 2009).

The second significant pathway PID_BMPPATHWAY is an expert-curated pathway provided by the pathway interaction database (Schaefer et al., 2009; downloaded from MSigDB). This pathway consists of 40 genes, including 7 (6 genes with gene-based *p* value < .05) of the 18 genes of GO:0060393_regulation_of_pathway-restricted_SMAD_protein_phosphorylation pathway. In addition to the six overlapping genes with GO:0060393_regulation_of_pathway-restricted_SMAD_protein_phosphorylation pathway, there are three more genes in PID_ BMPPATHWAY showing a gene-based *p* value less than .05: *GREM1* (gene-based *p* value = 3.60 × 10^{-5}), *TAB2* (gene-based *p* value = 4.83 × 10^{-5}), and *SMAD5* (gene-based *p* value = 0.011). Refer to Supplementary Table 2 for gene-based test statistics of all genes in the PID_BMPPATHWAY. Variants in the *GREM1* are associated with the colorectal cancer. The PID_BMPPATHWAY shows evidence of association (pathway-based *p* value = 4.80 × 10^{-3}) after removal of *GREM1* and other genes with top SNP *p* value less than 5.00 × 10^{-8}, suggesting the contribution of sub-genome-wide threshold genes.

The last significant pathway is GO:0042692_ muscle_cell_differentiation, consisting of 86 genes of which 18 show a gene-based *p* value less than .05 (refer to Supplementary Table 3). Interestingly, apart from *BMP4* none of the other genes showed strong evidence for an association at a gene-based level, with much of the signal

https://doi.org/10.1017/thg.2016.100 Published online by Cambridge University Press

TABLE 2

Top Five Pathways from VEGAS2Pathway Analysis of Colorectal Cancer GWAS Summary Data and their Test Statistics Using the INRICH and MAGENTA Approaches

Pathway ID	Pathway size	VEGAS2Pathway *p* value	INRICH *p* value	MAGENTA *p* value
GO:0060393_Regulation_of_pathway-restrictedSMAD_protein_phosphorylation	18	$<1.00 \times 10^{-6}$	6.00×10^{-6}	1.17×10^{-2}
PID_BMPPATHWAY	40	2.00×10^{-6}	7.00×10^{-6}	1.31×10^{-2}
GO:0042692_Muscle_cell_differentiation	86	8.00×10^{-6}	1.75×10^{-4}	3.60×10^{-3}
Panther_TGF-beta_signaling_pathway	61	3.00×10^{-5}	8.81×10^{-3}	6.00×10^{-4}
GO:0048729_Tissue_morphogenesis	156	4.40×10^{-5}	3.61×10^{-4}	5.20×10^{-5}

deriving from the combined effect of many genes with moderate effect sizes. As a result, the significance of GO:0042692_muscle_cell_differentiation was only slightly altered by removing genes, which were individually significant (those with top SNPs with p value less 5.00×10^{-8}) from the analysis (pathway-based p value $= 7.80 \times 10^{-5}$).

Comparisons with other Pathway Analysis Approaches

To see if these pathways are also implicated by other approaches, we performed pathway analysis using INRICH and MAGENTA software. We used the same gene-pathway annotations as used for VEGAS2pathway to make association statistics comparable across approaches. Table 2 shows the association p-values for our reported top five pathways using the INRICH and MAGENTA approaches.

All top five pathways observed for colorectal cancer using VEGAS2pathway show association p values less than .05 using both INRICH and MAGENTA. In each case, the association p values obtained using INRICH and MAGENTA were less significant than the ones obtained using VEGAS2Pathway. These results are in line with the observations by Evangelou et al. (2012), who reported that the Fisher's method approach (in which combined test statistics of all variants in a pathway is compared against combined test statistics of permuted gene-sets of same size, similar to the VEGAS2Pathway approach) is likely to be a more powerful competitive pathway analysis approach than the hypergeometric and GSEA approaches.

INRICH, a hypergeometric approach, by default uses 1.00×10^{-4} as a threshold for a variant to be called significant, hence it will be underpowered to detect a pathway containing many variants with association p value greater than 1.00×10^{-4}. On the contrary, VEGAS2Pathway uses information of all associations from GWAS analysis to perform pathway analysis. Hence, VEGAS2Pathway outperforms INRICH to identify GO:0042692_muscle_cell_differentiation pathway, which contains many moderately associated genes (refer to Supplementary Table 3).

Both VEGAS2Pathway and MAGENTA are two-step, pathway analysis methods in which a gene-based association test statistics are first calculated to test the association of a pathway. MAGENTA considers only the top SNP within a gene-boundary to define gene's association test statistics, whereas VEGAS2Pathway combines association test statistics of all SNPs within a gene boundary to calculate each gene's association test statistics. Hence, VEGAS2Pathway would be more powerful than MAGENTA to identify pathways containing genes with many independently associated variants. The variation in results from VEGAS2Pathway and MAGENTA might also be due to inherent differences between hypotheses tested. As mentioned previously, MAGENTA is based on the GSEA approach, which tests whether the gene-set is enriched with highly ranked genes than expected by chance, whereas VEGAS2Pathway tests whether the combined association test statistics of genes within a pathway is significantly different from the ones obtained from randomly generated gene-sets.

We further tested whether VEGAS2Pathway also implicates the top pathways for colorectal cancer obtained using INRICH and MAGENTA. Supplementary Tables 4 and 5 show top pathways for colorectal cancer obtained using INRICH and MAGENTA respectively, and their VEGAS2Pathway association p values.

Another recent competitive pathway analysis approach (DEPICT) was not included in this comparison, since its current implementation does not allow analysis using user specified gene-pathway annotations.

Web-Based Implementation of VEGAS2Pathway

We implemented VEGAS2Pathway in VEGAS2 web page https://vegas2.qimrberghofer.edu.au/ and a command line tool. Users can select VEGAS2pathway analysis in the options box of the VEGAS2 web page to perform the pathway analysis. The web-based version is easy to use and requires only a two-column GWAS summary file with 'rsID' and 'p-value.' The user manual and scripts can also be downloaded from VEGAS2 webpage.

Discussion

We present a novel but simple competitive approach for pathway analysis of GWAS data. Our approach differs from existing approaches in a number of ways. Gene-based calculation in ALIGATOR (Holmans et al., 2009) is based on the assumption that LD within genes is constant, with

https://doi.org/10.1017/thg.2016.100 Published online by Cambridge University Press

the authors highlighting the difficulty in accounting for variable LD without resorting to computationally intensive permutation-based methods. A computationally efficient simulation approach (VEGAS) is used here to calculate gene-based test statistics that account for variable LD patterns within a gene and give results similar to those from a permutation approach. The hypergeometric approach (implemented in DEPICT and INRICH) requires users to prespecify a significance threshold for SNPs in the pathway test, with less significant SNPs ignored. In contrast, our approach does not use a pre-specified inclusion threshold for SNPs/genes; instead, all genes are included, with weighting determined by the evidence for each gene from a gene-based test. Approaches such as GenGen and MAGENTA consider only a top SNP within a gene boundary to calculate gene-based test statistics. Only including one SNP per gene has the potential to discard relevant information if there are multiple independent signals in a gene. As just one example, Yang et al. (2012) reported that among genes with SNPs associated with height, many had more than one associated SNP. Our default gene-based test includes all SNPs within a gene; hence, it would show improved power to detect pathway that comprises genes with multiple independent risk variants.

In addition to the adjustment of major confounders for pathway analysis of GWAS data such as LD between variants, gene size, and pathway size, the implementation of VEGAS2Pathway is easy to use. The implementation does not depend upon commercial software (as is the case with MAGENTA, which requires a commercial MATLAB licence); rather, a free user-friendly web-based implementation is provided. Our approach does not require much preprocessing of input files (as the case in INRICH) a two-column text file with rsID and p value is a sufficient input to perform default analysis. The web implementation ignores variants without rsID, whereas the command line tool allow user to incorporate variants without rsIDs in their analysis (providing the user has a data set available to estimate the relevant pair-wise LD values).

We incorporated both manually curated pathways from MSigDB (containing canonical pathways and gene-sets from BIOCRATA, REACTOME, KEGG databases), PANTHER, and pathway commons, and computationally predicted pathways from the GO database. We have provided a simulation derived multiple testing cut-off since the traditional multiple testing corrections are over-conservative for non-independent pathway tests. Including pathways from different sources should ensure our approach is applicable to a wide range of complex traits. Moreover, the VEGAS2Pathway command line implementation allows the user to specify other gene pathway annotations, giving more flexibility for analysis.

Our pathway analysis results suggest the enrichment of association signals for risk of colorectal cancer in BMP-SMAD signaling and muscle cell differentiation pathways. BMP-SMAD signaling is involved in cellular differentiation (Kobayashi et al., 2005) and apoptosis (Guha et al., 2002). Interestingly, different genes encoding regulators of BMP signaling such as the BMP antagonist (*GREM1*), BMPs (*BMP2*, *BMP4* and *BMP7*), kinase (*TAB1*), BMP receptor (*BMPR1A*), and inhibitory SMADs (*SMAD6* and *SMAD7*) are driving the association of the PID_BMPPATHWAY. Recently, the BMP-SMAD1 signaling pathway was shown to stabilize tumor suppressor *p53*, suggesting its loss of function in tumorogenesis (Chau et al., 2012). Another study reported the inactivation of BMP signaling in majority of sporadic colorectal cancer patients (Kodach et al., 2008).

Our study is the first to report an association of the muscle cell differentiation pathway with colorectal cancer. Many genes moderately associated with colorectal cancer drive the association of muscle cell differentiation pathway including *NRG1*, *CAPN2*, *NOS1*, *FGF10*, *SOX15,* and *ATG5*. Several lines of evidence suggest biological plausibility for the role of this pathway in tumorigenesis. The high expression of the transmembrane *NRG1* and NRG1/HER3 signaling in the tumor mesenchymal stem cells is associated with poor colorectal cancer prognosis and colorectal cancer cell progression, respectively (De Boeck et al., 2013). *CAPN2*-dependent IκBα degradation is responsible for secondary resistance of colorectal cancer cells to the CPT-11 (irinotecan) anti-cancer therapy (Fenouille et al., 2012). Expression of *NOS1* is negatively associated with patient response to adoptive T-cell anti-cancer therapy (Liu et al., 2014). *FGF10* is reported to induce cell migration and invasion in pancreatic cancer cells through *FGF10/ FGFR2* signaling (Nomura et al., 2008). *SOX15* was reported to show tumor suppressive effects in pancreatic cancer (Thu et al., 2014). *ATG5* is strongly downregulated in colorectal cancer patients (Cho et al., 2012). In aggregate, these results suggest that the variants residing in genes involved in muscle cell differentiation pathway might play an important role in colorectal cancer risk.

In the current analysis, we assigned variants that lie within 50 kb on either side of a gene's transcription site to compute its association p value. This choice of boundary might have ignored distantly located risk variants; we used this selection criterion to strike a balance between inclusions of possible cis-regulatory variants and maintaining specificity of a gene. The VEGAS2Pathway implementation allows user to specify the gene-boundary to compute gene-based p value. Furthermore, to note SNP assignment based on a wide boundary leads to inclusion of many nonrisk variants, which can dilute the association signal of a gene. Assigning SNPs based on their regulatory relationship to genes is an important future direction. Similarly, future work would ideally expand this approach to include rare variants for gene- and pathway-based analyses. Further to add, our current work did not explore all aspects of VEGAS2 gene-based tests such as top-SNP and top percentage tests, but this study was performed considering default

https://doi.org/10.1017/thg.2016.100 Published online by Cambridge University Press

all variant test. Recently, Wojcik et al. (2015) compared 21 different methods for gene- and pathway-level analysis of GWAS data and reported that the VEGAS approach considering the top 10% of SNP association p values performs better in terms of both sensitivity and control of type 1 error rate. In this work, we focused on the performance of three methods — VEGAS2Pathway, INRICH, and MAGENTA — on real data from colorectal cancer GWAS; further comparison of VEGAS2Pathway approach with other available methods under different simulated scenarios will be an important extension of current work.

In summary, we report the VEGAS2Pathway approach for pathway analysis of GWAS summary data. It accounts for LD between SNPs within a gene, and between neighboring genes, gene size, and pathway size. The current version of VEGAS2Pathway uses computationally predicted gene ontology pathways and expert-curated pathways from the MSigDB, PANTHER, and pathway commons databases. This approach is implemented in both a user-friendly web page and a unix command line perl script. We applied our method on colorectal cancer GWAS summary data and found evidence that genes involved in the BMP signaling and muscle cell differentiation pathways might play a role in the development of colorectal cancer.

Acknowledgments

We thank Dr Matthew Law for useful comments on the manuscript. We thank Xiaping Lin and Jonathan Davies from QIMR Berghofer IT for assistance with the web application. For the majority of work, AM was supported by an ANZ Trustees PhD scholarship. SM is supported by Australian National Health and Research Council and Australian Research Council fellowships. None of the authors have any conflicts of interest to report.

For description on the funding and acknowledgements for GECCO, please refer to Supplementary text 2.

Supplementary Material

To view supplementary material for this article, please visit https://doi.org/10.1017/thg.2016.100.

References

Alarmo, E. L., Parssinen, J., Ketolainen, J. M., Savinainen, K., Karhu, R., & Kallioniemi, A. (2009). BMP7 influences proliferation, migration, and invasion of breast cancer cells. *Cancer Letters, 275*, 35–43.

Brosens, L. A., van Hattem, A., Hylind, L. M., Iacobuzio-Donahue, C., Romans, K. E., Axilbund, J., … Giardiello, F. M. (2007). Risk of colorectal cancer in juvenile polyposis. *Gut, 56*, 965–967.

Cerami, E. G., Gross, B. E., Demir, E., Rodchenkov, I., Babur, O., Anwar, N., … Sander, C. (2011). Pathway Commons, a web resource for biological pathway data. *Nucleic Acids Research, 39*, D685–690.

Chau, J. F., Jia, D., Wang, Z., Liu, Z., Hu, Y., Zhang, X., … Li, B. (2012). A crucial role for bone morphogenetic protein-Smad1 signalling in the DNA damage response. *Nature Communications, 3*, 836.

Cho, D. H., Jo, Y. K., Kim, S. C., Park, I. J., & Kim, J. C. (2012). Down-regulated expression of ATG5 in colorectal cancer. *Anticancer Research, 32*, 4091–4096.

Chu, H., Luo, H., Wang, H., Chen, X., Li, P., Bai, Y., … Zhang, G. (2014). Silencing BMP-2 expression inhibits A549 and H460 cell proliferation and migration. *Diagnostic Pathology, 9*, 123.

De Boeck, A., Pauwels, P., Hensen, K., Rummens, J. L., Westbroek, W., Hendrix, A., … De Wever, O. (2013). Bone marrow-derived mesenchymal stem cells promote colorectal cancer progression through paracrine neuregulin 1/HER3 signalling. *Gut, 62*, 550–560.

Evangelou, M., Rendon, A., Ouwehand, W. H., Wernisch, L., & Dudbridge, F. (2012). Comparison of methods for competitive tests of pathway analysis. *PLoS One, 7*, e41018.

Fenouille, N., Grosso, S., Yunchao, S., Mary, D., Pontier-Bres, R., Imbert, V., … Lagadec, P. (2012). Calpain 2-dependent IkappaBalpha degradation mediates CPT-11 secondary resistance in colorectal cancer xenografts. *Journal of Pathology, 227*, 118–129.

Gene Ontology Consortium. (2008). The gene ontology project in 2008. *Nucleic Acids Research, 36*, D440–444.

Guha, U., Gomes, W. A., Kobayashi, T., Pestell, R. G., & Kessler, J. A. (2002). In vivo evidence that BMP signaling is necessary for apoptosis in the mouse limb. *Developmental Biology, 249*, 108–120.

Holmans, P., Green, E. K., Pahwa, J. S., Ferreira, M. A., Purcell, S. M., Sklar, P., … Craddock, N. (2009). Gene ontology analysis of GWA study data sets provides insights into the biology of bipolar disorder. *American Journal of Human Genetics, 85*, 13–24.

Howe, J. R., Mitros, F. A., & Summers, R. W. (1998). The risk of gastrointestinal carcinoma in familial juvenile polyposis. *Annals of Surgical Oncology, 5*, 751–756.

Jeon, H. S., Dracheva, T., Yang, S. H., Meerzaman, D., Fukuoka, J., Shakoori, A., … Jen, J. (2008). SMAD6 contributes to patient survival in non-small cell lung cancer and its knockdown reestablishes TGF-beta homeostasis in lung cancer cells. *Cancer Research, 68*, 9686–9692.

Jiao, S., Peters, U., Berndt, S., Brenner, H., Butterbach, K., Caan, B. J., … Hsu, L. (2014). Estimating the heritability of colorectal cancer. *Human Molecular Genetics, 23*, 3898–3905.

Kobayashi, T., Lyons, K. M., McMahon, A. P., & Kronenberg, H. M. (2005). BMP signaling stimulates cellular differentiation at multiple steps during cartilage development. *Proceedings of the National Academy of Sciences of the United States of America, 102*, 18023–18027.

Kodach, L. L., Wiercinska, E., de Miranda, N. F., Bleuming, S. A., Musler, A. R., Peppelenbosch, M. P., … Hardwick, J. C. (2008). The bone morphogenetic protein pathway is inactivated in the majority of sporadic colorectal cancers. *Gastroenterology, 134*, 1332–1341.

Lee, P. H., O'Dushlaine, C., Thomas, B., & Purcell, S. M. (2012). INRICH: interval-based enrichment analysis for genome-wide association studies. *Bioinformatics, 28*, 1797–1799.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdottir, H., Tamayo, P., & Mesirov, J. P. (2011). Molecular signatures database (MSigDB) 3.0. *Bioinformatics, 27*, 1739–1740.

Liu, J. Z., McRae, A. F., Nyholt, D. R., Medland, S. E., Wray, N. R., Brown, K. M., … MacGregor, S. (2010). A versatile gene-based test for genome-wide association studies. *American Journal of Human Genetics, 87*, 139–145.

Liu, Q., Tomei, S., Ascierto, M. L., De Giorgi, V., Bedognetti, D., Dai, C., … Marincola, F. M. (2014). Melanoma NOS1 expression promotes dysfunctional IFN signaling. *Journal of Clinical Investigation, 124*, 2147–2159.

Liu, Y., Chen, J., Yang, Y., Zhang, L., & Jiang, W. G. (2012). Muolecular impact of bone morphogenetic protein 7, on lung cancer cells and its clinical significance. *International Journal of Molecular Medicine, 29*, 1016–1024.

Mishra, A., & MacGregor, S. (2015). VEGAS2: Software for more flexible gene-based testing. *Twin Research and Human Genetics, 18*, 86–91.

Nomura, S., Yoshitomi, H., Takano, S., Shida, T., Kobayashi, S., Ohtsuka, M., … Miyazaki, M. (2008). FGF10/FGFR2 signal induces cell migration and invasion in pancreatic cancer. *British Journal of Cancer, 99*, 305–313.

Pers, T. H., Karjalainen, J. M., Chan, Y., Westra, H. J., Wood, A. R., Yang, J., … Franke, L. (2015). Biological interpretation of genome-wide association studies using predicted gene functions. *Nature Communications, 6*, 5890.

Peters, U., Bien, S., & Zubair, N. (2015). Genetic architecture of colorectal cancer. *Gut, 64*, 1623–1636

Peters, U., Jiao, S., Schumacher, F. R., Hutter, C. M., Aragaki, A. K., Baron, J. A., … Colon Cancer Family Registry and the Genetics and Epidemiology of Colorectal Cancer Consortium, . (2013). Identification of genetic susceptibility loci for colorectal tumors in a genome-wide meta-analysis. *Gastroenterology, 144*, 799–807.e24.

Robinson, M. R., Wray, N. R., & Visscher, P. M. (2014). Explaining additional genetic variation in complex traits. *Trends in Genetics, 30*, 124–132.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., & Buetow, K. H. (2009). PID: The pathway interaction database. *Nucleic Acids Research, 37*, D674–679.

Segre, A. V., DIAGRAM Consortium, MAGIC investigators Groop, L., Mootha, V. K., Daly, M. J., & Altshuler, D. (2010). Common inherited variation in mitochondrial genes is not enriched for associations with type 2 diabetes or related glycemic traits. *PLoS Genet, 6*, e1001058.

Study, C., Houlston, R. S., Webb, E., Broderick, P., Pittman, A. M., Di Bernardo, M. C., … Dunlop, M. G. (2008). Meta-analysis of genome-wide association data identifies four new susceptibility loci for colorectal cancer. *Nature Genetics, 40*, 1426–1435.

Thomas, P. D., Campbell, M. J., Kejariwal, A., Mi, H., Karlak, B., Daverman, R., … Narechania, A. (2003). PANTHER: a library of protein families and subfamilies indexed by function. *Genome Research, 13*, 2129–2141.

Thu, K. L., Radulovich, N., Becker-Santos, D. D., Pikor, L. A., Pusic, A., Lockwood, W. W., … Tsao, M. S. (2014). SOX15 is a candidate tumor suppressor in pancreatic cancer with a potential role in Wnt/beta-catenin signaling. *Oncogene, 33*, 279–288.

Wang, K., Li, M., & Hakonarson, H. (2010). Analysing biological pathways in genome-wide association studies. *Nature Reviews Genetics, 11*, 843–854.

Wang, K., Zhang, H., Kugathasan, S., Annese, V., Bradfield, J. P., Russell, R. K., … Hakonarson, H. (2009). Diverse genome-wide association studies associate the IL12/IL23 pathway with Crohn disease. *American Journal of Human Genetics, 84*, 399–405.

Wang, L., Jia, P., Wolfinger, R. D., Chen, X., & Zhao, Z. (2011). Gene set analysis of genome-wide association studies: Methodological issues and perspectives. *Genomics, 98*, 1–8.

Wojcik, G. L., Kao, W. H., & Duggal, P. (2015). Relative performance of gene- and pathway-level methods as secondary analyses for genome-wide association studies. *BMC Genetics, 16*, 34.

Yamaguchi, J., Nagayama, S., Chino, A., Sakata, A., Yamamoto, N., Sato, Y., … Arai, M. (2014). Identification of coding exon 3 duplication in the BMPR1A gene in a patient with juvenile polyposis syndrome. *Japanese Journal of Clinical Oncology, 44*, 1004–1008.

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., Genetic Investigation of, A. T. C., Replication, D. I. G., … Visscher, P. M. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Genetics, 44*, 369–375.

https://doi.org/10.1017/thg.2016.100 Published online by Cambridge University Press