An Ant Colony Optimisation and Tabu List Approach to the Detection of Gene-Gene Interactions in Genome-Wide Association Studies

Emmanuel Sapin, *College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, England*
Edward Keedwell, *College of Engineering, Mathematics and Physical Sciences, University of Exeter, Exeter, England*
and Timothy Frayling, *Exeter Medical School, University of Exeter, Exeter, England.*

Abstract

In this paper, a novel ant colony optimisation and tabu list approach for the discovery of gene-gene interactions in genome-wide association study data is proposed. The method is tested on a number of diseases drawn from the large established database, the Wellcome Trust Case Control Consortium which contains hundreds of thousands of small DNA changes known as single nucleotide polymorphisms. To analyse full scale genome-wide association study data, the standard ant colony optimisation algorithm has been adapted, with tournament path selection, a subset based approach, and tabu list included in the algorithm. These modifications, in addition to the use of a statistical test of significance of single nucleotide polymorphism interactions as a fitness function, greatly increase execution speeds and permit the discovery of combinations of single nucleotide polymorphisms that can discriminate cases and controls. The methodology is applied to several large-scale genome-wide association study disease datasets namely, inflammatory bowel disease, rheumatoid arthritis, type I diabetes and type II diabetes patients to discover putative gene-gene interactions in reasonable time on modest hardware.

Index Terms

Genome Wide Association Study, Ant Colony Optimisation, Single Nucleotide Polymorphism.

I. Introduction

The advent of the sequencing of the human genome in 2003 has created many opportunities for scientists to understand the associations between an individual's genome and the propensity for disease. Recent advances in sequencing techniques allow researchers to sequence the genomes of thousands of individuals and to compare genomes across a large cohort of subjects. Such studies, known as genome-wide association studies (GWAS), capture the small variations in genomes (known as single nucleotide polymorphisms (SNPs) [1]) among individuals and attempt to understand the association between these and the variation in phenotypic traits such as height, body mass index and the propensity to develop certain diseases. Associations between SNPs and a disease can be found by iteratively exploring the association of each SNP in turn, a computationally complex but feasible problem. The exploration of associations between more than one SNP and a disease is a much more computationally complex problem. So called gene-gene interactions can be investigated as an additive model where the effects of possessing two associated SNPs are simply added together or through other mechanisms such as epistasis where the individual (or main) effect of each SNP might be small but in combination, the effect is large [2].

The first reported GWAS were developed around 2007 to investigate the genetic basis of type II diabetes. Since then many other disease datasets have been created from large projects such as the Wellcome Trust Case Control Consortium (WTCCC) and UK Biobank. GWAS offer the potential to illuminate the genetic causes of diseases and provide an opportunity for early treatment and planning for patients leading to profound social and economic benefits.

The GWAS investigated here are real-world disease datasets taken from the WTCCC set. The diseases explored are the two types of diabetes type 1 diabetes (T1D) and type 2 diabetes (T2D), Inflammatory Bowel Disease (IBD) and Rheumatoid Arthritis (RA) and the heritability of these diseases has been the subject of numerous studies. Type II Diabetes (T2D), characterized by insulin resistance and affecting hundreds of millions of people worldwide [3], is studied in numerous GWAS [4], [5], [6], [7]. Type I Diabetes (T1D), a chronic autoimmune disorder with onset usually in childhood, is tackled by Vella et al. in these GWAS [8]. Rheumatoid arthritis, a chronic inflammatory disease characterized by the destruction of the synovial joints resulting in severe disability, is the subject of [9]. For inflammatory bowel disease, the pathogenic mechanisms are poorly understood and its heritability is studied in [10].

From a computational perspective, GWAS present a significant challenge as there are hundreds of thousands of SNPs (variables) per individual. In these datasets they are recorded for thousands of individuals creating a database of large proportions (almost 2.5bn elements in the experiments described later). Any computational approaches used for analysis therefore must be scalable in the face of these large-scale data. Many examples of GWAS data analysis exist in the literature that successfully demonstrate the association between a single SNP and the disease. When a SNP is strongly associated with a disease it is said to be one of the main effects in the dataset and the discovery of these single associations is computationally feasible with modern hardware. However the computational challenge increases markedly when the task is to find SNP combinations associated with a disease that demonstrate a significant gene-gene interaction.

There are known single associations for type II diabetes and traits such as height for instance, however, there is a considerable amount of missing heritability; for example only approximately 10% of variation in height can be explained by traditional single SNP GWAS. This missing heritability could be due to rare variants, or to combinations of SNPs (gene-gene interactions) which are beginning to be explored and increasingly becoming of interest. Standard GWAS analyses are carried out through full enumeration (e.g. the software package Plink: which can perform a range of basic, large-scale analyses in a computationally efficient manner [11]). With modern hardware, the association of hundreds of thousands of SNPs with a disease can be determined within reasonable computational time. However, when combinations (pairs, triplets and higher) are considered, the computational load becomes highly burdensome or completely intractable. This has led to a variety of approaches [12] for the discovery of gene-gene interactions that can be broadly divided into two groups, those that pre-screen SNPs for their association and exhaustively search the reduced dataset (known as the filter approach) and those that explore the entirety of the dataset through a heuristic technique (known as the wrapper approach). The filter method is often problematic for the discovery of epistasis as all SNPs or SNP combinations must be investigated during the filtering stage, leading to the exclusion of SNPs

with weak marginal effects (single associations) if only single SNPs are considered, or extremely high computation time if SNP combinations are considered. The wrapper approach, usually accomplished through a global search technique, is able to search the space of all combinations but cannot guarantee to find the best combinations within the dataset due to the exceptionally large search space and the greedy or stochastic nature of the algorithm.

The filter approach is investigated in numerous studies including [13], where a Bayesian partitioning model and a Markov chain Monte Carlo approach are used, and [14] in which a dimensionality reduction technique is used. In [15] a hierarchical learning algorithm to search for combinations of SNPs is investigated and in [16] a novel Bayesian graphical method, called BEAM3, is introduced for large-scale association mapping. Furthermore, an approach for genome-wide interaction analysis of case-control SNP data and quantitative traits, called INTERSNP, is presented in [17]. Among the filter approaches, one of the most popular tools for exploring gene-gene interactions in GWAS is BOOST [18]. It is a fast approach based on a noniterative method to approximate the likelihood ratio statistic and is able to search through all pair-wise combinations by using log-likelihood analysis.

Wrapper approaches include decision tree [19], neural networks [20], odds ratio [21] and filtering-based approaches [22] in addition to stochastic techniques such as ant colony optimisation (ACO) [23], [24], [25] that has been shown to be a promising technique. In our previous work, we have demonstrated how the ACO algorithm can be used to search for gene-gene interactions for type II diabetes [25] in a full set of GWAS data, comprising many SNPs and individuals without utilising expert knowledge.

The ACO technique is a strong candidate for this task as it has a natural fit with discrete optimisation problems, and with a modification to allow for the selection of subsets of variables and a highly configurable pheromone deposition rule, is well suited to the problem of finding gene-gene interactions in large data. The ACO algorithm has also been shown to deliver excellent results on discrete combinatorial test problems [26] and has been widely applied to real-world problems ranging from water distribution system optimisation [27] to robotics [28].

In this paper, inspired by [29], elements of tabu search were incorporated into the ACO approach in order to find gene-gene interactions. Tabu search [30], a local search method used for mathematical optimisation, searches the neighbourhood of solutions around the current solution, but is forbidden from moving to those solutions presented on the tabu list (often a list of solutions that were previously visited). This process ensures that the algorithm does not cycle among solutions and can be used to promote promising areas of the search space. In the approach described here, a tabu list is used to prevent the ACO algorithm from continually selecting SNPs that are associated with the main effects (individual SNP assocations) in the dataset. Tabu lists have been used in ACO approach since their inception where they were applied to the travelling salesman problem [31]. However, the lists are used in that application to remove visited cities from consideration in path selection during the optimisation, as opposed to removing single variables between optimisation runs as described here. In this application, SNPs that are associated with the main effects are included in the tabu list when they are detected. Once a SNP appears in the list it is not available for selection by the algorithm from that point on. This modification allows the ACO to concentrate on combinations of SNPs with smaller marginal effects which are therefore those more likely to yield epistatic interactions.

This paper presents an ACO approach for the analysis of full-scale GWAS data with the aim to find combinations of SNPs that have associations with T2D, T1D, RA and IBD across a population of thousands of individuals. The ACO algorithm incorporates pheromone trails and evaporation but is modified in several ways from a traditional method, with the inclusion of a subset-based pheromone deposition and tournament path selection. The following sections describe the methods, the experiments and the results. The last section concludes and summarizes the presented results.

II. METHOD

The ACO algorithm is run as a standard wrapper method for discovering gene-gene interactions. In this method, the algorithm selects N SNPs (in this case N=2) from the full dataset and evaluates the combination for their ability to discriminate between controls and cases within the dataset. The following subsections describe the specific ACO approach used.

A. Algorithm

The basic ant colony optimisation approach for searching for combinations of SNPs that can discriminate between controls and cases within a GWAS dataset is as follows. An ant (agent) selects a combination of SNPs from the dataset randomly with a bias towards SNPs with the greatest pheromone value. The fitness (discrimination capability) of the chosen combination of the two SNPs $snp1$ and $snp2$ is calculated and the corresponding pheromone value P($snp1$+$snp2$) is deposited on each SNP. This is repeated for a population of ants and then all pheromone values are evaporated by applying a uniform multiplier (< 1.0) across the SNPs.

The algorithm can be described as follows:

1: Initialise pheromone on each SNP to $initpheromone$;
2: **repeat**
3: **for all** $nbant$ ants **do**
4: Select two SNPs via tournament selection (see subsection below);
5: Calculate the fitness of the combination;
6: **end for**
7: Update the pheromone of the two SNPs with the best fitness;
8: **for all** SNPs **do**
9: Evaporate the pheromone;
10: **end for**
11: **until** the end of the execution

where $nbant$ is the number of ants of the algorithm and $initpheromone$ is the initial value of the amount of pheromone for each ant that was experimentally chosen as 100.

The algorithm described above and in Figure 1 is a somewhat standard ACO algorithm. However, certain novel adaptations are required to configure the algorithm for use with the gene-gene interaction problem on real data.

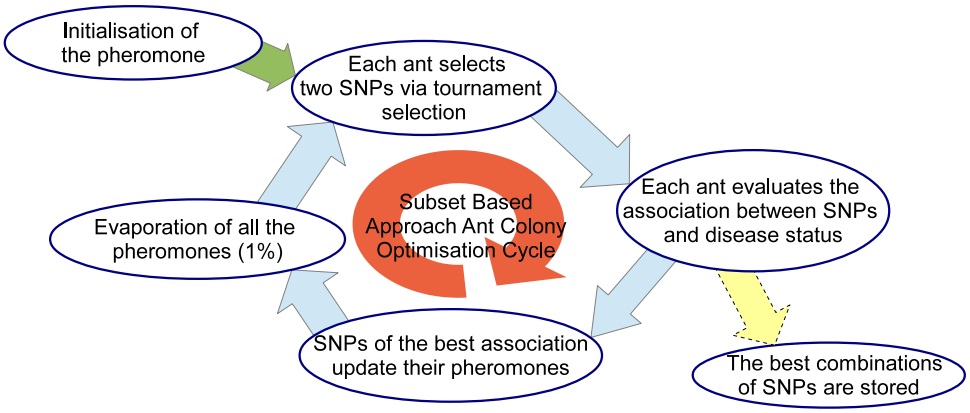

Fig. 1. Overview of the ant colony optimisation approach for searching for combinations of SNPs.

1) Selection of SNPs: Selecting SNPs for new combinations based on their pheromone values can be computationally expensive in this ACO algorithm due to the high number of SNPs (e.g. 400,000). In standard ACO implementations, a structure similar to a biased roulette wheel is used to make the path choice that is both stochastic and biased towards the path with the greatest pheromone. However, this approach breaks down when the number of variables is high, as the case here. Therefore we make use of tournament selection to achieve stochastic path selection as it has proven to be better for high dimensional problems [32]. In the tournament-based approach, a number of SNPs (nbt) are randomly selected from the possible set to form a tournament and the SNP with the highest amount of pheromone among them is selected as part of this solution. The tournament has many of the same properties of the roulette wheel approach, in that it enables a balance between selecting paths randomly with lower amounts of pheromone and biasing the search towards those with high pheromone. The size of the tournament clearly has an effect on this balance between the exploration and exploitation capability of the algorithm. The setting of this parameter is investigated in the later experimental sections.

2) Genotypes and Logical Combinations: Each SNP is comprised of one of three genotypes, common homozygous (CH) (e.g. CC), rare homozygous (RH) (e.g. GG) and heterozygous (H) (e.g. CG). There are a number of standard models of interaction between genes, for example the additive model states that the effect of one gene adds to the effect of another, in contrast in epistasis, SNPs have relatively small individual effects but the combined effect of SNPs is large. This usually implements the AND model of combination [23], individuals will have one genotype (e.g. RH) AND another genotype (e.g. H) if they are to be included in the positive group. However, when considering the details of the combination at the genotype level a number of different possibilities present themselves. In this work, the algorithm can explore a number of combination types providing it with greater expressive power. Initially all logical boolean operations between two genotypes were considered by the algorithm, but analysis revealed that this can be reduced down to the following four combinations that encompass all real-world possibilities.

In the approach described here, the following logical interactions between two SNPs are considered:

- An individual is positive if and only if the first SNP takes a specific value and the second SNP takes a specific value. (AND)
- An individual is positive if and only if the first SNP or the second SNP takes their specific values. (OR)
- An individual is positive if and only if the first SNP takes a specific value and the second SNP does not take a specific value. (AND NOT)
- An individual is positive if and only if exactly one of the two SNPs takes a specific value and the other SNP does not take a specific value. (XOR)

From the above list it can be seen that this extension allows the algorithm to search for more sophisticated interactions between genotypes than the standard AND relationship. During the search process, when two SNPs are selected by the ACO approach, their genotypes are investigated using all of these four logical combinations, from which the best is selected to be used as the fitness of the combination.

3) Fitness Function: The fitness function must represent the discriminatory ability of a combination between control individuals and cases. SNPs in combination with high fitness values will receive more pheromone and are therefore more likely to be selected for new combinations. Thus this function leads the search process of the algorithm and is therefore a key aspect of the algorithm. The fitness function is based on standard statistical measures that are implemented on a binary classification of controls and cases in genome wide association studies [16], [23] and are described below.

The efficacy of two SNPs $snp1$ and $snp2$ in discriminating between the two classes is evaluated by the numbers of positive (p) and negative (n) individuals among the cases (D_p and D_n) and controls (C_p and C_n), where the determination of positive and negative individuals is achieved through the use of the logical combination rules described previously, and according to the following process.

1: Initialise two 4 by 4 tables $T_{Controls}$ and T_{Cases};
2: **for all** controls **do**
3: Increment $T_{Controls}[Value\_of\_snp1][Value\_of\_snp2]$ by 1;
4: **end for**
5: **for all** cases **do**
6: Increment $T_{Cases}[Value\_of\_snp1][Value\_of\_snp2]$ by 1;
7: **end for**
8: C_p (C_n) $\leftarrow$ sum of cells of $T_{Controls}$ for which the combination is true (false);
9: D_p (D_n) $\leftarrow$ sum of cells of T_{Cases} for which the combination is true (false);

The complexity of this calculation is $\mathcal{O}(n)$ where n is the total number of individuals (controls and cases). This is important because an ACO run may require over a million fitness evaluations, the main computational load of the algorithm is devoted to the evaluation of the fitness function and therefore this must be as efficient as possible.

To calculate fitness, Pearson's chi-squared test on a binary classification of controls and cases is used. The four values C_p, C_n, D_p and D_n are used to calculate the expected values $E.C_p$, $E.C_n$, $E.D_p$ and $E.D_n$The chi-squared

statistic $X^2_{snp1,snp2}(v1, v2)$ is given by the formula:

$$X^2_{snp1,snp2}(v1, v2) = \frac{(E.D_p - D_p)^2}{E.D_p} + \frac{(E.D_n - D_n)^2}{E.D_n} + \frac{(E.C_p - C_p)^2}{E.C_p} + \frac{(E.C_n - C_n)^2}{E.C_n} \quad (1)$$

As described previously, there are three possible values CH, H and RH for $v1$ and again three possible values for $v2$. Therefore there are 9 (3×3) different chi-squared values for $snp1$ and $snp2$ and the largest of these is selected as the fitness function value $f(snp1, snp2)$ of the combination of the two SNPs $snp1$ and $snp2$.

$$f(snp1, snp2) = max\{X^2_{snp1,snp2}(v1, v2)\} \text{ such that } (v1, v2) \in \{\text{CH, H, RH}\}^2 \quad (2)$$

From the Pearson's chi-squared the p-value (probability of achieving this result through chance) of the association can be calculated.

4) Updating pheromone: At each generation of the algorithm, each of the $nbant$ ants selects two SNPs to test their combination. The amount of pheromone of the two SNPs contained in the combination with the highest fitness are updated. For the two pheromone levels the following is applied:

$$P(snp1) \leftarrow P(snp1) + f(snp1, snp2) \quad (3)$$

$$P(snp2) \leftarrow P(snp2) + f(snp1, snp2) \quad (4)$$

B. Memory Management

The database used is composed of samples of the genome of approximately 2,000 individuals (Cases) with the disease (1,999 for T2D, 2,000 for T1D, 2,005 for IBD and 1,999 for RA) and 3,004 control samples. Each sample is composed of 490,294 SNPs and due to the diploid nature of the human genome each SNP consists of two alleles (two among Adenine (A), Cytosine (C), Guanine (G) and Thymine (T)) leading to three possible genotypes described above. Additionally, due to the sequencing of the genome, a genotype can be unknown and therefore a fourth possible value of 'unknown' exists for a SNP.

The data for approximately 5,000 individuals were stored in 'oxstat' and 'plink' formats [33] on a normal hard drive and required more than an hour to open and to read these files for each disease with an Intel® Core™ i7-2600 CPU @3.40GHz processor.

An ACO run may require over a million fitness evaluations and so even a small improvement in complexity of the fitness function will have a large impact on performance and clearly, a function that requires the searching of a database on disk will lead to run times orders of magnitude longer than one in which it is stored in RAM. However, with more than 2 billion elements to represent (5,000 individuals $\times$ 490,294 SNPs), each SNP cannot

be represented by more than 1-2 bytes in memory. The representation below implements a lossless compression of the data that enables it to be kept in memory when considering whole genome analysis.

C. Representation

In order to keep the database in memory when considering whole genome analysis, a SNP is encoded as follows:

Unknown: 0

Common Homozygous (CH): 1

Heterozygous (H): 2

Rare Homozygous (RH): 3

By raising the SNP number to the power of the encoding above, four SNPs from an individual are encoded in one byte (4^4 = 256).

By using this lossless compression, the size of the database is reduced four-fold and crucially, enables the dataset to be stored entirely in RAM on a standard PC with at least 1GB of RAM. Furthermore, files using this representation that are created on the hard drive can be read and opened in less than one and a half minutes using the hardware described above.

Although extra processing is required to compress and decompress the data, the benefit from holding all data in memory with no paging to disk easily outweighs this disadvantage. The compression method also provides scalability in the presence of a greater number of individuals or SNPs, a likely scenario in this field with larger databases already on the horizon.

III. Experimental Setup

Experimentation has been conducted using the modified ant colony approach on four real-world genome wide association datasets taken from the Wellcome Trust Case Control Consortium, each consisting of approximately 500,000 SNPs (variables) and 5 000 individuals (records). The following subsections discuss the SNP exclusion criteria used and the experimentation conducted to determine the best parameters for exploration and exploitation of the dataset by the ACO algorithm. This experimentation focused on the size of tournament in the tournament selector and the resulting coverage of SNPs in the dataset during an algorithm run. Furthermore, the use of permutation testing to determine benchmark p-values is also explored.

A. Exclusion Criteria

A variety of exclusion criteria are required in GWAS datasets before processing can begin. Readers are referred to the GWAS literature for more information on these criteria [34], [23]. The SNPs kept are those meeting the following standard conditions in the 3,004 control samples:

- HWE Exact Test $> 5.7 \times 10^{-7}$, minor allele frequency > 1 % and studywise missing data proportion < 5%.
- Studywise minor allele frequency > 5% OR studywise missing data proportion < 1%.
- 58C versus NBS 1dfTT p-value $> 5.7 \times 10^{-7}$ and 58C versus NBS 2dfGT p-value $> 5.7 \times 10^{-7}$.

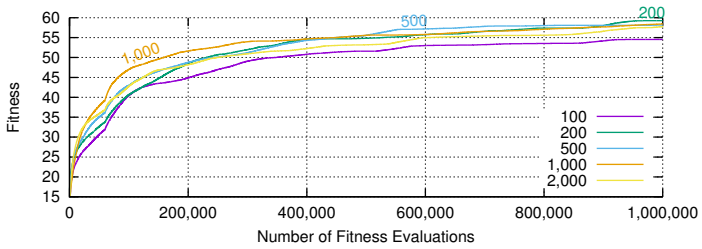

Fig. 2. Evolution of the average best fitness over 10 runs of the algorithm with 50 ants and 100, 200, 500, 1,000 and 2,000 as a tournament size.

There are three conditions in the approximately 2,000 samples of genomes of individuals with the disease: HWE Exact Test $> 5.7 \times 10^{-7}$, studywise missing data proportion $< 5\%$ and minor allele frequency $> 1\%$.

The remaining data after the application of these criteria contain 395,711 SNPs for T2D, 395,602 SNPs for T1D, 396,093 SNPs for IBD and 395,862 SNPs for RA.

B. Parameters

Stochastic search algorithms, such as ACO, often have a set of associated parameter values that must be set before experimentation can begin. The number of ants in a population, pheromone evaporation rate, and pheromone deposition in ACO will all have an effect on the way in which the algorithm runs. In the following experimentation we focused on the tournament size for path selection, as this is a novel element of the algorithm, and also the number of ants in the population. The pheromone evaporation rate was kept constant at 1% and the pheromone deposition was simply the fitness provided by the fitness function described above without transformation.

- $nbant$: Number of ants of the algorithm.
- nbt: Number of SNPs in the tournament of the selection process.

An investigation into these parameters was conducted to determine the effect of changing the population size and tournament size on the execution of the algorithm. Inspired by [32], the algorithm was run with the values 50 and 200 for the number of ants and the values 100, 200, 500, 1,000 and 2,000 for the tournament size. Ten algorithm runs were conducted for each combination of values of these two parameters on the type II diabetes dataset taken from the Wellcome Trust Case Control Consortium database. The highest fitness values that have been found during each run are stored and an average of these is then computed. The variation of the highest fitness was considered against the number of function evaluations, where the stopping criterion is set to a maximum of 100,000 such function evaluations. The results of the algorithm with 50 ants and 200 ants are described in the following subsections.

C. Algorithm Results

1) 50 Ants: The average of the highest fitness values was computed and shown in Figure 2 for 50 ants.

The overall best fitness from these runs is for a tournament size of 200. It is worth noting that before the first 30,000 evaluations of the fitness function, the highest fitness is for a tournament size of 2,000 items, then between

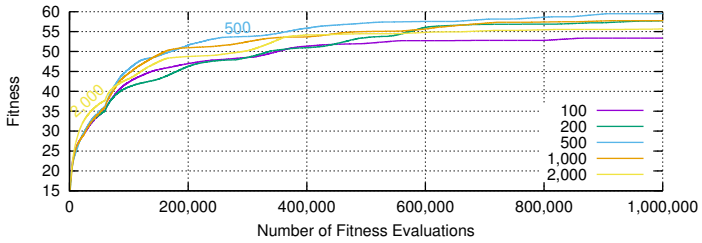

Fig. 3. Evolution of the average best fitness over 10 runs of the algorithm with 200 ants and 100, 200, 500, 1,000 and 2,000 as a tournament size.

30,000 and 490,000 the highest fitness is for a tournament size of 1,000, then between 490,000 and 900,000 the highest fitness is for a tournament size of 500 and 200 thereafter. This can be explained as larger tournament sizes will result in more exploitation and less exploration leading to better performance initially, but earlier convergence. There would appear to be too much exploitation without enough exploration for 2,000, 1,000 and 500 items in the tournament. For a population of 50 ants, 200 items in the tournament appears to be a reasonable setting as it achieves the best performance in these runs.

2) 200 Ants: An average of the highest fitness values is shown in Figure 3 for 200 ants.

For 200 ants the overall best fitness achieved by the algorithm is for a tournament size of 2,000 at the beginning, reducing to 500 items after 1,000,000 evaluations of the fitness function.

The 50 and 200 ants experiments have shown that population size and tournament size have an effect on one another as expected. However, each of the population sizes behaves relatively consistently, achieving a chi-squared value of just under 60 when configured with the correct tournament size. As expected, the smallest tournament sizes (i.e. 100) and largest tournaments (i.e. 2,000) perform comparatively poorly indicating slow and early algorithm convergence respectively. The best range for tournaments appears to be between 200 and 1,000, a ratio of just 0.05-0.25% with respect to the number of variables, far smaller than would be expected in EA tournament selection where a tournament size of 10% of the population is the norm. Clearly, the extent to which the algorithm exploits and explores can be tuned by use of the tournament size parameter.

3) Dataset Coverage: A further key question regarding these parameter settings is the explorative capability of the algorithm and in particular the extent to which the dataset of almost 400,000 SNPs is covered by the algorithm. To this end, Figures 4 (population of 50 ants) and 5 (population of 200 ants) show the dataset coverage of the algorithm as it progresses for differing tournament sizes. Unsurprisingly, the tournament of size 100 explores more of the space than any other, but as seen in the previous subsection, this comes at the expense of the discovery of good combinations within reasonable time.

However, the exploration is not improved greatly for a tournament size of 100 over the preferred figure of 200 for this number of ants as shown in Figure 4.

4) Summary: The extent to which the algorithm can cover the dataset and explore combinations is important in determining the level of exploration and exploitation within the algorithm. The goal of these experiments is to

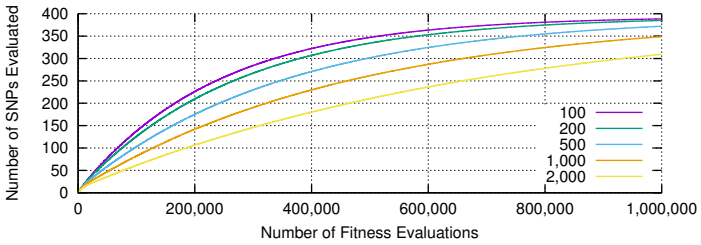

Fig. 4. Evolution of the number of SNPs evaluated at least once over 10 runs of the algorithm with 50 ants and 100, 200, 500, 1,000 and 2,000 as a tournament size.

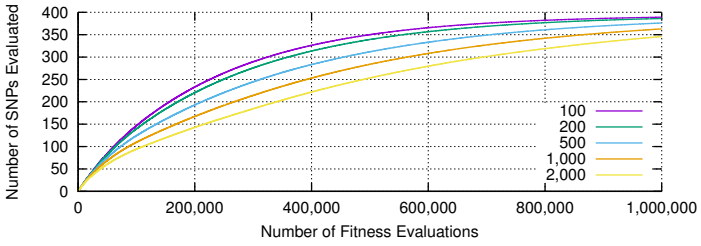

Fig. 5. Evolution of the number of SNPs evaluated at least once over 10 runs of the algorithm with 200 ants and 100, 200, 500, 1,000 and 2,000 as a tournament size.

determine parameter settings that ensure good coverage, but also some eventual convergence on a set of likely SNPs within the computation time available. The results in Figures 2 and 3 clearly show the number of ants and the size of the tournament are not independent in this approach [25] for these GWAS. In this experimentation, the fitness is the highest for 50 ants and a tournament size of 200 and they are the chosen parameters of the algorithm for further experiments.

D. Permutation Tests

A key aspect of the analysis of GWAS data is that any discovered combination should be validated to determine the likelihood that it could exist by chance and lead to a type I error [35]. As the database consists of hundreds of thousands of SNPs and many statistical tests are conducted by the algorithm, a permutation test is a common approach to estimating the p-values of associations that can be expected purely by chance. The permutation test establishes the baseline p-values that arise by chance and so provides a benchmark for the ACO results on unshuffled real data.

To determine these p-values, the ACO algorithm is run on shuffled databases. 1,999 individuals among the 5,003 individuals (Cases + Controls) are randomly chosen to be the individuals with the disease (Cases) while the remaining individuals are those without the disease (Controls). The process is repeated 200,000 times and the algorithm is run for each set of shuffled data and the best p-value of each run is stored. There are 2,000 best p-values (1% of 200,000) lower than 2×10^{-11}. Any result obtained with the method on real data with a p-value lower than 2×10^{-11} has less than 1% chance to exist by chance and therefore be due to type I error.

TABLE I
BEST COMBINATIONS OF SNPS DISCOVERED BY IDENTIFIER (RS NUMBER). CHROMOSOME NUMBER AND GENE NAME (WHERE APPLICABLE) IN PARENTHESES.

DISEASE	COMBINATION	p-VALUE
T2D	rs9508846(13,hCG_1815504)=AA AND rs7901695(10,TCF7L2)=CC	8×10^{-15}
T2D	rs11196205(10,TCF7L2)=GG OR rs10992923(9)=GA	3×10^{-15}
T2D	rs7077039(10,TCF7L2)=TT XOR rs9783382(11)=GG	6×10^{-14}
T2D	rs9508846(13,hCG_1815504)≠GA AND rs7901695(10,TCF7L2)=CC	4×10^{-16}
IBD	rs12242030(10)=AA AND rs17116117(11,HTR3B)=CC	2×10^{-28}
IBD	rs10210302(2,ATG16L1)=CC OR rs17116117(11,HTR3B)=TC	1×10^{-30}
IBD	rs7382225(6)=GG XOR rs17116117(11,HTR3B)=TC	3×10^{-27}
IBD	rs2076756(16,NOD2)≠GG AND rs17116117(11,HTR3B)=CC	4×10^{-31}
T1D	rs3805006(3,ITPR1)=CT AND rs9270986(6,NCBI36)=AA	7×10^{-256}
T1D	rs9273363(6,HLA-DQB1)=GG OR rs3805006(3,ITPR1)=CC	7×10^{-246}
T1D	rs9273363(6,HLA-DQB1)=TT XOR rs7859401(9,ZNF367)=CC	6×10^{-247}
T1D	rs3805006(3,ITPR1)≠CC AND rs9270986(6,NCBI36)=AA	9×10^{-293}
RA	rs17104722(14)=CC AND rs4718582(7,TYW1)=CC	5×10^{-103}
RA	rs4718582(7,TYW1)=AG OR rs2076533(6)=AA	8×10^{-101}
RA	rs4718582(7,TYW1)=AA XOR rs7295430(12)=AA	6×10^{-87}
RA	rs9268403(6)≠TT AND rs4718582(7,TYW1)=TT	3×10^{-104}

IV. GWAS RESULTS

In Table I, the best results of an algorithm run over 20,000 generations with 50 and 200 ants as a tournament size are presented for each disease. With these parameters, a generation of the algorithm requires an average of 0.13 seconds, meaning an optimisation run requires an average of 43 minutes and 33 seconds on the hardware described earlier.

The ACO algorithm found good results exceeding the permutation test threshold for combinations of two SNPs concerning T2D, T1D, IBD and RA. This table demonstrates the efficacy of the ACO approach in discovering SNP combinations with low p-values across a range of diseases. Associations where linkage disequilibrium (LD) is expected to be involved (i.e. where SNPs are close together on the genome and are correlated) have been removed, and all the interactions described above have SNPs on different chromosomes, eliminating the possibility of LD. The p-values vary widely among diseases, indicating the difference in strength of the underlying main effect in each disease. An additional interesting point is that there are a variety of logical combinations represented, from these results it certainly does not appear that gene-gene interactions must be confined to 'and'-type relationships.

For T2D, all the best combinations contain the SNP rs7901695 that is in the gene TCF7L2 and is well known to be associated with T2D [36]. This demonstrates that the ACO approach is able to find SNPs that have been associated with this disease in the literature.

For IBD, the ACO algorithm found combinations of SNPs with a p-value around 10^{-30} that contain the SNP with rs17116117 that is in gene HTR3B that is a major determinant of serotonin-receptor function [37]. This SNP or those close to it drive this effect and confirm the findings of previous GWAS.

For T1D, rs9270986 and rs9273363 lead the results and are in the HLA region that contains many genes involved

in the immune system's recognition and the latter is also known to be highly associated with type I diabetes [34].

For RA, the results are driven by the main effect of rs4718582, this SNP is in the gene TYW1 that is the human homolog of a yeast gene essential for Wybutosine synthesis [38].

Table I shows the impact of the combination of the SNPs and the very low p-values. For each disease, combinations of two SNPs that can discriminate patients from controls have been discovered by the ACO algorithm. For these diseases, the above results show that the convergence of the algorithm is driven by one or two SNPs that are the main effect of the associations (i.e. with the exception of RA, that they have been previously identified in single association studies to be highly associated with the disease). This demonstrates that the algorithm is capable of discovering biologically plausible associations from the data, but in many cases, one SNP is providing the main effect and the effect of the interaction is rather small. An ACO variant incorporating a tabu list was therefore implemented to tackle this phenomenon and is the subject of the next section.

V. ACO-TABU METHOD

This method is based on the ant colony optimisation method described earlier in section II. The modified method removes the main effects from the search as they are discovered and allows the ACO algorithm to concentrate on combinations of SNPs with smaller marginal effects. In detail, the ACO algorithm runs $numgen$ generations, the SNP $snp1$ with the largest amount of pheromone is identified and all combinations of $snp1$ and all remaining SNPs in the dataset are calculated. The combination with the highest chi-squared value is recorded and $snp1$ is removed from the dataset for further combinations. The ACO-Tabu method can be described by the following algorithm.

1: **repeat**
2: Run the ACO algorithm $numgen$ generations;
3: Identify SNP with highest amount pheromone as $snp1$;
4: Calculate all the combinations of $snp1$ and ;
5: Record the best combination;
6: Remove $snp1$ from the dataset;
7: **until** end of the run

The number of generations $numgen$ between the removal of SNPs has been experimentally chosen to be 1,000.

A. Results

Figure 6 shows a typical run of the execution of the ACO-Tabu hybrid, removing the most significant SNP at every 1,000 generations. As would be expected, performance drops for a time, before climbing to another peak. Inevitably over time, the overall fitness drops as more top SNPs are deleted. The first four SNPs removed from the database are present in the TCF7L2 gene and the fifth is located in the well-known FTO gene.

Over 100 runs of the algorithm of 5,000 generations each, the number of times each SNP that has been found at least once is shown in Figure 7. A SNP in the gene TCF7L2 is found in every one of the hundred runs. Clearly,

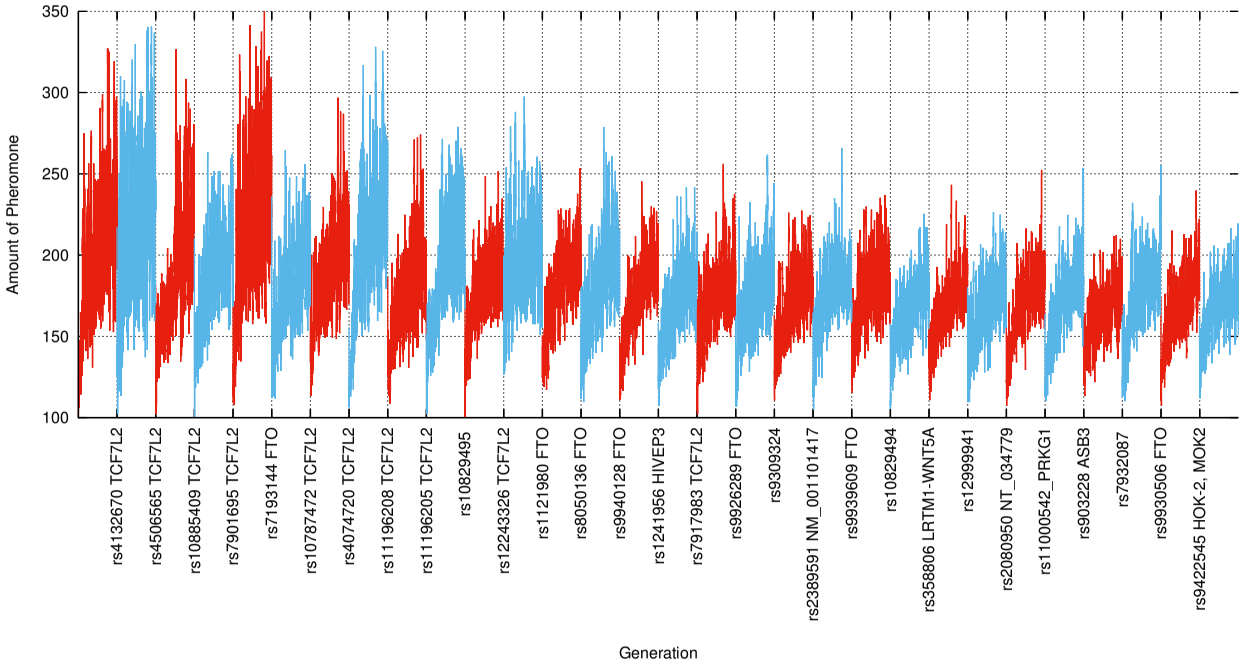

Fig. 6. Evolution of the highest amount of pheromone over 30,000 generations. Every 1,000 generations, the SNP with the highest amount of pheromone can be seen.

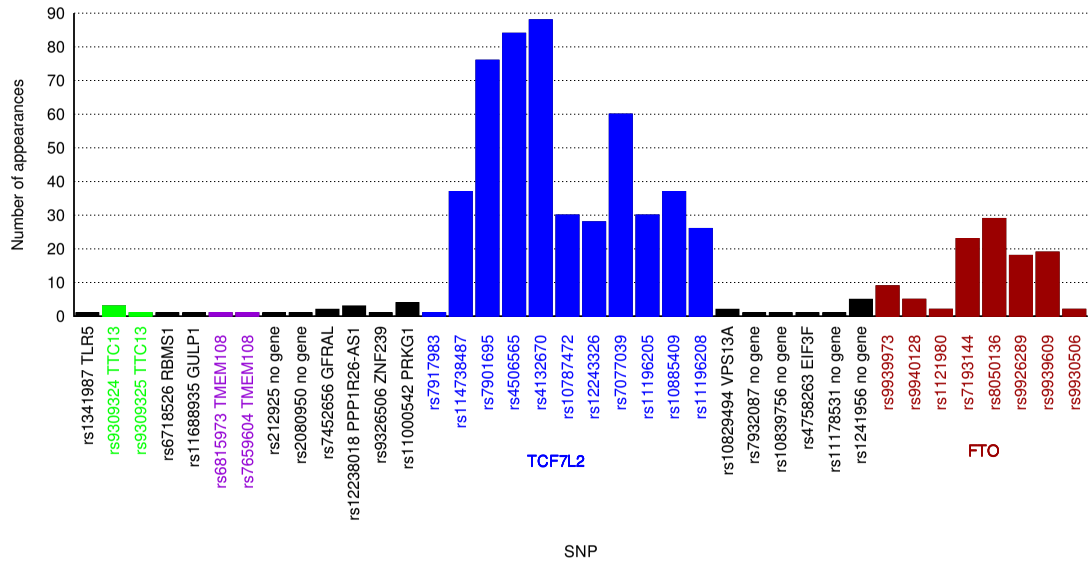

Fig. 7. Number of times each SNP was found over 100 runs of the algorithm and 5,000 generations.

the best two known signals, TLCF7 and FTO, dominate this figure. However, there are some other promising genes identified by the algorithm that could represent new avenues for investigation.

A sample of the best combinations from this algorithm on the type II diabetes dataset can be seen in Table II. Encouragingly, despite removing the main effects, the system is still able to discover gene-gene interactions with good p-values for type II diabetes. Due to the removal of the main effects, the resulting combinations are less

well-known in the literature and so it is more difficult to verify their biological plausibility. However some of the SNPs from Table II, other than the SNPs in the genes FTO and TCF7L2, are known to be associated with T2D and are identified below.

The SNP with rs1481279 (combination X) is described as the most notable signal contribution to T2D predisposition outside known loci [39]. The SNP with rs17139608 (combination IX) is known to be associated with BMI in the entire population-based full-heritage Pima Indian sample [40]. In [41], interesting signals for epistasis are searched for and the strongest evidence for epistasis that is discovered is the combination of rs1935683 and rs11196205 (combination V). The SNP with rs11742692 (combination VI) is in the gene ARL15 that influences Adiponectin, a protein inversely associated with risk of T2D mellitus [42]. The SNP with rs7767391 (combination IV) in the gene CDKAL1 was identified to be significantly associated with T2D [43]. The SNP with rs713129 (combination VII) is in the gene SOCS6 and it has been shown that constitutive expression of SOCS6 protein in retinal neurons may improve glucose metabolism [44].

TABLE II
SAMPLE OF THE BEST COMBINATIONS OF SNPS DESCRIBED AS RS-NUMBER (CHROMOSOME, GENE OR INTERGENIC REGION(IGR)) AND THEIR P-VALUE THAT WERE DISCOVERED.

NUMBER	COMBINATION	p-VALUE
I	rs4506565(10, TCF7L2)=AA AND rs2578050(10, IGR)=AA	1×10^{-16}
II	rs210357(14, IGR)=AA AND rs4506565(10, TCF7L2)=AA	1×10^{-16}
III	rs4132670(10, TCF7L2)=AA AND rs210357(14, IGR)=AA	1×10^{-16}
IV	rs7901695(10, TCF7L2)=CC AND rs7767391(6, CDKAL1)=CC	1×10^{-15}
V	rs1935683(6, RFPL4B)=CC AND rs11196205(10, TCF7L2)=GG	6×10^{-15}
VI	rs11742692(5, ARL15)=CC AND rs8050136(16, FTO)=AA	8×10^{-15}
VII	rs7077039(10, TCF7L2)=TT AND rs713129(18, SOCS6)=CC	2×10^{-14}
VIII	rs7193144(16, FTO)=AA AND rs255761(5, ARL15)=GG	6×10^{-14}
IX	rs9309325(2, IGR)=TT AND rs17139608(16, A2BP1)=GG	5×10^{-12}
X	rs349586(5, IGR)=GA AND rs1481279(4, SLC9B2)=TT	4×10^{-11}
XI	rs765534(11, IGR)=AC AND rs4765066(12, IGR)=CT	4×10^{-11}

B. Method comparisons

Firstly, a comparison is made between the ACO-Tabu approach and a Monte Carlo approach that consists of the generation and testing of random pairs of SNPs. As expected, a Monte Carlo method on the type II diabetes dataset does not perform well over one million generated pairs, the average p-value yielded is 3.5×10^{-7} and the best is 1.9×10^{-13}. The ACO-Tabu search algorithm is compared here with other popular algorithms designed to search for gene-gene interaction in GWAS.

The comparison with methods such as BEAM3 [16] is difficult as BEAM3 cannot run on a dataset of the size used here, namely 400,000 SNPs and 5,000 individuals. In [16], this algorithm was run on each chromosome individually to select 3,809 SNPs from different chromosomes and subsequently BEAM3 was run on these SNPs. The ACO-Tabu search algorithm explored all combinations when running on only 3,809 SNPs and can run on much larger datasets as shown above. Due to the size of the dataset, the ACO-Tabu, in our experimentation, discovered the

TABLE III
CALCULATION OF THE COMPUTATION TIME OF BOOST, ACO-TABU ALGORITHM, PLINK, INTERSNP. PLINK IS TESTED WITH THE FAST-EPISTASIS OPTION AND WITHOUT THE CASE-ONLY OPTION. THE TIMINGS OF BOOST, PLINK AND INTERSNP ARE CARRIED OUT ON A 3.0 GHZ CPU WITH 4GB MEMORY AND ACO-TABU ALGORITHM ON A 3.4 GHZ CPU USING 2GB MEMORY.

DATA SIZE	BOOST	ACO-TABU ALGORITHM	PLINK	INTERSNP
5,000 INDIVIDUALS AND 1,000 SNPs	<2s	25s	106s	160s
5,000 INDIVIDUALS AND 5,000 SNPs	42s	625s	2,703s	4,277s
5,000 INDIVIDUALS AND 10,000 SNPs	170s	2,500s	10,915s	15,805s

same set of gene-gene interactions as BEAM3 and therefore the ACO-Tabu search algorithm performed equivalently to BEAM3, in terms of results discovered, on this small dataset.

In [18], a fast approach to detecting gene-gene interactions in GWAS is presented, called BOOST. In this work, the computation time of BOOST is compared with that of Plink [11] and INTERSNP [17] on small dataset sizes (5,000 individuals and 1,000, 5,000 and 10,000 SNPs). On these smaller sizes of data the ACO-Tabu search algorithm can store all of the SNPs in the tabu list and therefore can test all the pairs of SNPs in computation times shown in Table III (The 10,000 first SNPs of chromosome 1 of the T2D dataset were used for this experiment).

The ACO-Tabu algorithm is faster than Plink and INTERSNP under these conditions, however, BOOST is faster than the ACO-Tabu approach on these data sizes. Clearly BOOST is a benchmark comparison here for the discovery of gene-gene interactions from this type of data, particularly as it appears to be able to discover interactions from large-scale data in a reasonable time frame. However, as a filter approach, BOOST will always be susceptible to longer runtimes resulting from increase in the number of SNPs in the data and the number of SNPs considered in a gene-gene interaction and some studies have found that BOOST can yield very high run times [45]. Although the ACO algorithm would also require more resources to operate on larger datasets, the link between runtimes and dataset size is not as fixed as it is with the filter approaches. In addition, the ACO approach described here is also capable of searching the space of possible logical interactions between SNPs and is not reliant on a single model. Finally, on the type II diabetes dataset of the WTCCC with BOOST, the authors of [18] did not find non-trivial interactions whereas the ACO-Tabu algorithm apparently discovered a number of these as described in the previous section. Nevertheless, it is interesting to discover how many of the gene-gene interactions discovered by BOOST could be discovered by the ACO algorithm, despite it being a stochastic approach, and thus the following experiment was carried out.

In this further experiment, the ability for the ACO-Tabu approach to find the best SNP interactions in approximately the same computational time as BOOST was investigated. The following experiment was conducted:

- Randomly select 100,000 SNPs in the database of T2D.
- Randomly select 400 individuals in the controls and 400 individuals in the cases of the database of T2D.
- Run BOOST to find the 100 best associations within these data, which took 125 minutes on our machine with the executable file provided in [46].
- Run the ACO approach with 200 as a tournament size with these data using tabu list of sizes 200, 100 and

TABLE IV
PERCENTAGE OF BEST INTERACTIONS DISCOVERED BY THE ACO-TABU APPROACH IN APPROXIMATELY EQUAL COMPUTATION TIME OVER 100 INDEPENDENT RUNS.

	TABU LIST OF SIZE 50	TABU LIST OF SIZE 100	TABU LIST OF SIZE 200
FOR THE 100 BEST INTERACTIONS	77.1%	84.9%	94%
FOR THE 50 BEST INTERACTIONS	78.8%	88.6%	96.1%
FOR THE 10 BEST INTERACTIONS	93.6%	99.2%	100%

50, with each run in a similar timeframe with the definition of SNP interactions taken from [18].

This size of data was chosen to be able to perform several runs in parallel and is also the size of the simulated data sets used in BOOST [46]. As the version of BOOST that can be downloaded in [46] does not consider unknown values, they have been converted into the value of the closest SNP for the same individual in the dataset.

This experiment was performed 50 times and the average percentage of best associations that were discovered for various sizes of tabu lists (50, 100 and 200) are presented in Table IV.

This clearly shows that the ACO algorithm is able to find a large proportion of the best signals within a sizeable database and lends confidence to the notion that it is searching these larger databases effectively. This is particularly the case for the runs of a tabu list of 200 which correctly identifies 100% of the top 10 interactions identified by BOOST and over 90% of the top 100. Additional experiments were performed to determine how long the ACO algorithm requires to find the top 10 interactions and over 90% of the top 100 interactions. Over ten runs of a tabu list of 200, the longest the ACO algorithm needed to find the top 10 interactions is 47 minutes and to discover over 90% of the top 100 interactions identified by BOOST is only 78 minutes.

The additional capability of the ACO algorithm to search larger databases, larger numbers of interacting genes and more sophisticated interactions between SNPs is not tested here.

VI. CONCLUSIONS AND FURTHER WORK

An ACO-Tabu list approach to the problem of discovering combinations of SNPs from large-scale GWAS data that can discriminate various diseases has been described.

The algorithm has been adapted so as to be scalable to the size of dataset both in terms of its memory requirements through the use of a byte-wise representation of genomes and through the use of a tournament path selection to greatly increase execution speeds. Due to a robust approach and these novel modifications, the ACO algorithm is able to operate on full-scale GWAS data and this is, to the best of our knowledge, the first time that an ACO method has been successfully applied to such data over a range of diseases.

Combinations of two SNPs that can discriminate inflammatory bowel disease, rheumatoid arthritis, type I diabetes and type II diabetes patients from controls have been discovered by the approach. The ACO algorithm has been able to find some of the strongest statistical signals in the dataset and has also found SNPs that have a known biological relationship to the diseases. The investigation of logical variations has shown that these provide the algorithm with

greater power to express the relationship between two or more SNPs. In particular, the NOT operator which allows the system to exclude one genotype in a SNP and include the others is an important logical distinction.

For the combinations that were discovered, in many cases, one SNP provided the main effect and the contribution of the gene-gene interaction is rather small. An ACO variant incorporating a tabu list has therefore been implemented to tackle this phenomenon. The ACO-Tabu hybrid allows the algorithm to investigate interactions between genes once the main effects have been removed, an important modification that allows the ACO algorithm to provide gene-gene associations that could generate new knowledge in the field. Further work is required to examine these relationships in more detail and to determine if they have biological plausibility in addition to statistical significance.

The approach has been compared to some of the most popular tools for exploring gene-gene interactions in data, in Plink, BOOST and INTERSNP and has been found to be competitive in terms of computational complexity and the quality of interactions discovered. This is in addition to the ability to process large datasets, investigate varying logical combinations and higher order gene-gene interactions that the ACO approach brings.

Although some of the discovered SNPs do not at present have a known biological function, it is this discovery of plausible known information and targets for further investigation that make the approach a promising addition to the GWAS toolbox.

The algorithm is also able to discover higher order combinations of SNPs (e.g. 3+ SNPs, not shown) that may not be possible using existing methods and further work is required to assess the statistical and biological meaning of these larger gene-gene interactions.

ACKNOWLEDGMENT

The work contained in this paper was supported by an EPSRC First Grant (EP/J007439/1).

This study makes use of data generated by the Wellcome Trust Case Control Consortium. A full list of the investigators who contributed to the generation of the data is available from http://www.wtccc.org.uk. Funding for the project was provided by the Wellcome Trust under award 076113.

REFERENCES

[1] S.T. Sherry, M. Ward, and K. Sirotkin, ”dbSNP - database for single nucleotide polymorphisms and other classes of minor genetic variation,” *Genome Research*, vol. 9, no. 8, pp. 677-679, 1999.

[2] S. Wright, ”Genic and organismic selection,” *Evolution*, vol. 34, No. 5, pp. 825–843, 1980.

[3] M. Kasuga, ”Insulin resistance and pancreatic beta cell failure,” *J Clin Invest*, vol. 116, pp. 1756-1760, 2006.

[4] S.S. Rich, ”Mapping genes in diabetes. Genetic epidemiological perspective,” *Diabetes.*, vol. 39, no. 11, pp. 1315–1319, 1990.

[5] E. Zeggini, M.N. Weedon, C.M. Lindgren, T.M. Frayling, K.S. Elliott, H. Lango, N.J. Timpson, J.R. Perry, N.W. Rayner, R.M. Freathy, J.C. Barrett, B. Shields, A.P. Morris, S. Ellard, C.J. Groves, L.W. Harries, J.L. Marchini, K.R. Owen, B. Knight, L.R. Cardon, M. Walker, G.A. Hitman, A.D. Morris, A.S. Doney, Wellcome Trust Case Control Consortium, M.I. McCarthy, and A.T. Hattersley, ”Replication of genome-wide association signals in UK samples reveals risk loci for type 2 diabetes,” *Science* (New York, N.Y.), vol. 316, no. 5829, pp. 1336–1341, 2007. PUBMED: 17463249; DOI: 10.1126/science.1142364

[6] A.L. Gloyn, M.N. Weedon, K.R. Owen, M.J. Turner, B.A. Knight, G. Hitman, M. Walker, J.C. Levy, M. Sampson, S. Halford, M.I. McCarthy, A.T. Hattersle, and T.M. Frayling, ”Large-scale association studies of variants in genes encoding the pancreatic beta-cell KATP channel subunits Kir6.2 (KCNJ11) and SUR1 (ABCC8) confirm that the KCNJ11 E23K variant is associated with type 2 diabetes,” *Diabetes*, vol. 52, pp. 568–572, 2003.

[7] S.F. Grant, G. Thorleifsson, I. Reynisdottir, R. Benediktsson, A. Manolescu, J. Sainz, A. Helgason, H. Stefansson, V. Emilsson, A. Helgadottir, U. Styrkarsdottir, K.P. Magnusson, G.B. Walters, E. Palsdottir, T. Jonsdottir, T. Gudmundsdottir, A. Gylfason, J. Saemundsdottir, R.L. Wilensky, M.P. Reilly, D.J. Rader, Y. Bagger, C. Christiansen, V. Gudnason, G. Sigurdsson, U. Thorsteinsdottir, J.R. Gulcher, A. Kong, and K. Stefansson, "Variant of transcription factor 7-like 2 (TCF7L2) gene confers risk of type 2 diabetes," *Nat Genet*, vol. 38, no. 3, pp. 320–323, 2006.

[8] A. Vella *et al.*, "Localization of a type 1 diabetes locus in the il2ra/cd25 region by use of tag single-nucleotide polymorphisms," *Am. J. Hum. Genet.*, vol. 76, pp. 773–779, 2005.

[9] D. Jawaheer, M.F. Seldin, C.I. Amos, W.V. Chen, R. Shigeta, J. Monteiro, M. Kern, L.A. Criswell, S. Albani, J.L. Nelson, D.O. Clegg, R. Pope, H.W. Jr. Schroeder, S.L. Jr. Bridges, D.S. Pisetsky, R. Ward, D.L. Kastner, R.L. Wilder, T. Pincus, L.F. Callahan, D. Flemming, M.H. Wener, and P.K. Gregersen, "A genome wide screen in multiplex rheumatoid arthritis families suggests genetic overlap with other autoimmune diseases," *Am. J. Hum. Genet.*, vol 68, pp. 927–936, 2001.

[10] C. Tysk, E. Lindberg, G. Järnerot, and B. Flodérus–myrhed, "Ulcerative-colitis and crohns-disease in an unselected population of monozygotic and dizygotic twins. A study of heritability and the influence of smoking," *Gut*, vol. 29, pp. 990-996, 1988.

[11] S. Purcell, B. Neale, K. Todd-Brown, L. Thomas, M.A. Ferreira, D. Bender, J. Maller, P. Sklar, P.I. de Bakker, M.J. Daly, and P.C. Sham, "PLINK: A tool set for whole-genome association and population-based linkage analyses," *Am. J. Hum. Genet.*, vol. 81, pp. 559-575, 2007.

[12] J.H. Moore and W.C. White, "Exploiting knowledge in genetic programming for genome-wide genetic analysis," *Lecture Note in Computer Science, Parallel Problem Solving from Nature - PPSN IX*, vol. 4193, pp. 969–977, 2006.

[13] Y. Zhang and J. Liu, "Bayesian inference of epistatic interactions in case-control studies," *Nat. Genet.*, vol. 39, pp. 1167–1173, 2007.

[14] S. Oh, J. Lee, M.S. Kwon, B. Weir, K. Ha, and T. Park, "A novel method to identify high order gene-gene interactions in genome-wide association studies: Gene-based MDR," *BMC Bioinformatics*, vol. 13(Suppl 9), S5, 2012.

[15] X. Wan, C. Yang, Q. Yang, H. Xue, N. Tang, and W. Yu, "MegaSNPHunter: A learning approach to detect disease predisposition SNPs and high level interactions in genome wide association study," *BMC Bioinformatics*, vol. 10, no. 13, 2009.

[16] Y. Zhang, "A novel bayesian graphical model for genome-wide multi-snp association mapping," *Genet Epi*, vol. 36, pp. 36–37, 2011.

[17] C. Herold, M. Steffens, F.F. Brockschmidt, M.P. Baur, and T. Becker. "INTERSNP: Genome-wide interaction analysis guided by a priori information," *Bioinformatics*, vol. 25, pp. 3275-3281, 2009.

[18] W. Xiang, C. Yang, Q. Yang, H. Xue, X. Fan, N.L.S. Tang, and W. Yu, "BOOST: A fast approach to detecting gene-gene interactions in genome-wide case-control studies," *AJHG*, vol. 87, no. 3, pp. 325-340, 2010.

[19] A. Jiang, T. Wanwan, X. Wu, and W. Fu, "A random forest approach to the detection of epistatic interactions in case-control studies," *BMC Bioinformatics*, vol. 10 Suppl 1, S65, 2009.

[20] F. Gunther, N. Wawro, and K. Bammann, "Neural networks for modeling gene-gene interactions in association studies," *BMC Genetics*, vol. 10, no. 87., 2009.

[21] L.Y. Chuang, M.C. Lin, and C.H. Yang, "Improved branch and bound algorithm for detecting SNP-SNP interactions in breast cancer," *Journal of Clinical Bioinformatics*, vol. 3, no. 4, 2013.

[22] C. Yang, Z. He, X. Wan, Q. Yang, H. Xue, and W. Yu. "SNPHarvester: a filtering-based approach for detecting epistatic interactions in genome-wide association studies," *Bioinformatics*, vol. 25, no. 4, pp. 504–511, 2009.

[23] J. Christmas, E. Keedwell, TM. Frayling, and JRB. Perry. "Ant colony optimisation to identify genetic variant association with type 2 diabetes," *Information Sciences*, vol. 181, pp. 1609–1622, 2011.

[24] C. Greene, B. White, and J. Moore, "Ant colony optimization for genome-wide genetic analysis," *Lecture Notes in Computer Science, Ant Colony Optimization and Swarm Intelligence*, vol. 5217, pp. 37–47, 2008.

[25] E. Sapin, E.C. Keedwell, and T. Frayling, "Subset-based ACO for genome wide association study: Discovery of promising combinations," *Proc. of 15 Annual Conference on Genetic and Evolutionary Computation*, pp. 295–302, 2013.

[26] M. Dorigo and G.D. Caro, "The ant colony optimization meta-heuristic," In *New Ideas in Optimization*, pp. 11–32. McGraw-Hill, 1999.

[27] A. Zecchin, H.R. Maier, A.R. Simpson, A., M. Leonard, and J.B. Nixon. "Ant colony optimization applied to water distribution system design: Comparative study of five algorithms," *J. Water Resour. Plann. Manage.*, vol. 133, no. 1, pp. 87–92, 2007.

[28] M.M. Mohamad, "Articulated robots motion planning using foraging ant strategy," *Jurnal Teknologi Maklumat*, vol. 20, no. 4, pp 163–181, 2008.

[29] T. Kaji, "Approach by ant tabu agents for traveling salesman problem," *Proc. of IEEE International Conference on Systems, Man, and Cybernetics*, vol. 5, pp. 3429–3434, 2001.

[30] F. Glover, "Future paths for integer programming and links to artificial intelligence," *Computers and Operations Research*, vol. 13, no. 5, pp. 533-549.

[31] T. Stützle and M. Dorigo, "ACO algorithms for the traveling salesman problem," *In K. Miettinen, M. Makela, P. Neittaanmaki and J. Periaux (eds.)* John Wiley & Sons, 1999.

[32] E. Sapin and E.C. Keedwell, "T-ACO - tournament ant colony optimisation for high-dimensional problems," *Proc. of 4th International Joint Conference on Computational Intelligence*, pp. 81-86, 2012.

[33] Wellcome Trust Case Control Consortium, June 2013 http://www.wtccc.org.uk/.

[34] Wellcome Trust Case Control Consortium, "Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls," *Nature*, vol. 447, pp. 661–678, 2007.

[35] K. F. Manly, D. Nettleton, and J.T. Gene Hwang, "Genomics, prior probability, and statistical tests of multiple hypotheses," *Genome Res.*, vol. 14, no. 6, pp. 997–1001, Jun. 2004.

[36] T. Jin and L. Liu, "The Wnt signaling pathway effector TCF7L2 and type 2 diabetes mellitus," *Mol. Endocrinol.*, vol. 22, no. 11, pp. 2383–2392, 2008.

[37] P.A. Davies, M. Pistis, M.C. Hanna, J.A. Peters, J.J. Lambert, T.G. Hales, and E.F. Kirkness, "The 5-HT3B subunit is a major determinant of serotonin-receptor function," *Nature.*, vol. 397, no. 6717, pp. 359–363, Jan. 1999.

[38] T. Suzuki and A. Noma, "Ribonucleome analysis identified enzyme genes responsible for wybutosine synthesis," *Nucleic Acids Symp Ser (Oxf).*, vol. 50, pp. 65–66, 2006.

[39] B.F. Voight, L.J. Scott, V. Steinthorsdottir, A.P. Morris, C. Dina, R.P. Welch, E. Zeggini, C. Huth, Y.S. Aulchenko, G. Thorleifsson, L.J. McCulloch, T. Ferreira, H. Grallert, N. Amin, C.J. Willer, S. Raychaudhuri, S.A. McCarroll, C. Langenberg, O.M. Hofmann, J. Dupuis, L. Qi, A.V. Segrè, M. van Hoek, P. Navarro, K. Ardlie, B. Balkau, R. Benediktsson, A.J. Bennett, R. Blagieva, E. Boerwinkle, L.L. Bonnycastle, K Bengtsson Boström, B. Bravenboer, S. Bumpstead, N.P. Burtt, G. Charpentier, P.S. Chines, M. Cornelis, D.J. Couper, G. Crawford, A.S. Doney, K.S. Elliott, A.L. Elliott, M.R. Erdos, C.S. Fox, C.S. Franklin, M. Ganser, C. Gieger, N. Grarup, T. Green, S. Griffin, C.J. Groves, C. Guiducci, S. Hadjadj, N. Hassanali, C. Herder, B. Isomaa, A.U. Jackson, P.R. Johnson, T. Jørgensen, W.H. Kao, N. Klopp, A. Kong, P. Kraft, J. Kuusisto, T. Lauritzen, M. Li, A. Lieverse, C.M. Lindgren, V. Lyssenko, M. Marre, T. Meitinger, K. Midthjell, M.A. Morken, N. Narisu, P. Nilsson, K.R. Owen, F. Payne, J.R. Perry, A.K. Petersen, C. Platou, C. Proença, I Prokopenko, W. Rathmann, N.W. Rayner, N.R. Robertson, G. Rocheleau, M. Roden, M.J. Sampson, R. Saxena, B.M. Shields, P. Shrader, G. Sigurdsson, T. Sparsø, K. Strassburger, H.M. Stringham, Q. Sun, A.J. Swift, B. Thorand, J. Tichet, T. Tuomi, R.M. van Dam, T.W. van Haeften, T. van Herpt, J.V. van Vliet–Ostaptchouk, G.B. Walters, M.N. Weedon, C. Wijmenga, J. Witteman, R.N. Bergman, S. Cauchi, F.S. Collins, A.L. Gloyn, U. Gyllensten, T. Hansen, W.A. Hide, G.A. Hitman, A. Hofman, D.J. Hunter, K. Hveem, M. Laakso, K.L. Mohlke, A.D. Morris, C.N. Palmer, P.P. Pramstaller, I. Rudan, E. Sijbrands, L.D. Stein, J. Tuomilehto, A. Uitterlinden, M. Walker, N.J. Wareham, R.M. Watanabe, G.R. Abecasis, B.O. Boehm, H. Campbell, M.J. Daly, A.T. Hattersley, F.B. Hu, J.B. Meigs, J.S. Pankow, O. Pedersen, H.E. Wichmann, I. Barroso, J.C. Florez, T.M. Frayling, L. Groop, R. Sladek, U. Thorsteinsdottir, J.F. Wilson, T. Illig, P. Froguel, C.M. van Duijn, K. Stefansson, D. Altshuler, M. Boehnke, M.I. McCarthy, MAGIC investigators, GIANT Consortium. "Twelve type 2 diabetes susceptibility loci identified through large-scale association analysis," *Nat Genet.*, vol. 43, no. 4, pp. 579-589, Jul. 2011.

[40] L. Ma, R.L. Hanson, M.T. Traurig, Y.L. Muller, B.P. Kaur, J.M. Perez, D. Meyre, M. Fu, A. Körner, P.W. Franks, W. Kiess, S. Kobes, W.C. Knowler, P. Kovacs, P. Froguel, A.R. Shuldiner, C. Bogardus, and L.J. Baier, "Evaluation of A2BP1 as an obesity gene," *Diabetes.*, vol. 59, no. 11, pp. 2837–2845, Nov. 2010.

[41] J.T. Bell, N.J. Timpson, N.W. Rayner, E. Zeggini, T.M. Frayling, A.T. Hattersley, A.P. Morris, and M.I. McCarthy, "Genome-wide association scan allowing for epistasis in type 2 diabetes," *Annals of human genetics*, vol. 75, no. 6, pp. 10–19, 2011.

[42] J.B. Richards, D. Waterworth, S. O'Rahilly, M.F. Hivert, R.J. Loos, J.R. Perry, T. Tanaka, N.J. Timpson, R.K. Semple, N. Soranzo, K. Song, N. Rocha, E. Grundberg, J. Dupuis, J.C. Florez, C. Langenberg, I Prokopenko, R. Saxena, R. Sladek, Y. Aulchenko, D. Evans, G. Waeber, J. Erdmann, M.S. Burnett, N. Sattar, J. Devaney, C. Willenborg, A. Hingorani, J.C. Witteman, P. Vollenweider, B. Glaser, C. Hengstenberg, L. Ferrucci, D. Melzer, K. Stark, J. Deanfield, J. Winogradow, M. Grassl, A.S. Hall, J.M. Egan, J.R. Thompson, S.L. Ricketts, I.R. König, W. Reinhard, S. Grundy, H.E. Wichmann, P. Barter, R. Mahley, Y.A. Kesaniemi, D.J. Rader, M.P. Reilly, S.E. Epstein, A.F. Stewart, C.M. Van Duijn, H. Schunkert, K. Burling, P. Deloukas, T. Pastinen, N.J. Samani, R. McPherson, G. Davey Smith, T.M. Frayling, N.J. Wareham,

J.B. Meigs, V. Mooser, T.D. Spector, GIANT Consortium. "A genome-wide association study reveals variants in ARL15 that influence adiponectin levels," *PLOS GENET.*, vol. 5, no. 12, 2010.

[43] V. Steinthorsdottir, G. Thorleifsson, I. Reynisdottir, R. Benediktsson, T. Jonsdottir, G.B. Walters, U. Styrkarsdottir, S. Gretarsdottir, V. Emilsson, S. Ghosh, A. Baker, S. Snorradottir, H. Bjarnason, M.C. Ng, T. Hansen, Y. Bagger, R.L. Wilensky, M.P. Reilly, A. Adeyemo, Y. Chen, J. Zhou, V. Gudnason, G. Chen, H. Huang, K. Lashley, A. Doumatey, W.Y. So, R.C. Ma, G. Andersen, K. Borch-Johnsen, T. Jorgensen, J.V. van Vliet–Ostaptchouk, M.H. Hofker, C. Wijmenga, C. Christiansen, D.J. Rader, C. Rotimi, J.C. Chan, O. Pedersen, G. Sigurdsson, J.R. Gulcher, U. Thorsteinsdottir, A. Kong, K. Stefansson. "A variant in CDKAL1 influences insulin response and risk of type 2 diabetes," *Nat Genet.*, vol. 39, no. 6, pp. 770–775, 2007.

[44] L. Xuebin, M. G. Mameza, Y. S. Lee, C. I. Eseonu, C.R. Yu, J.J.K. Derwent, and C.E. Egwuagu, "Suppressors of cytokine-signaling proteins induce insulin resistance in the retina and promote survival of retinal cells," *Diabetes*, vol. 57, no. 6, pp. 1651–1658, 2008.

[45] B. Goudey, D. Rawlinson, Q. Wang, F. Shi, H. Ferra, R.M. Campbell, L. Stern, M.T. Inouye, C.S. Ong, and A. Kowalczyk, "GWIS - model-free, fast and exhaustive search for epistatic interactions in case-control GWAS," *BMC Genomics*, 14(Suppl 3), S10, 2013.

[46] http://bioinformatics.ust.hk/BOOST.html Jul. 2014.