CS4220 Knowledge Discovery Methods in Bioinformatics Project

Project Title: An Integrated Method for Identifying Drug Resistance Associated Mutations in Bacteria

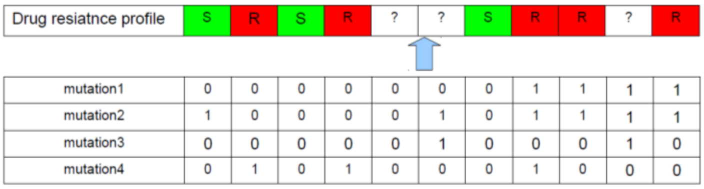
Name: Chayaporn Suphavilai (A0134669U)

Introduction

Antimicrobial resistance (AMR) threatens the effective prevention and treatment of an ever increasing range of infections caused by bacteria, parasites, viruses and fungi. Patients with infections caused by drug-resistant bacteria are generally at increased risk of worse clinical outcomes and death, and consume more healthcare resources than patients infected with the same bacteria that are not resistant [2]. For example, Tuberculosis (TB) is caused by bacteria (Mycobacterium tuberculosis) that most often affect the lungs. TB is spread from person to person through the air and is second only to HIV/AIDS as the greatest killer worldwide due to a single infectious agent [3]. Therefore, it is important to identify mutations that is responsible for drug resistance in bacteria.

Several mutations have been reported to associate with drug resistance mechanism in bacteria. In addition, large number of bacterial genomes being sequenced opens possibilities for using large scale computational approach. For example, GWAMAR system [6] uses whole genome comparative approaches to detect drug resistance mutations. GWAMAR system is a tool for identifying of drug resistance associated mutations based on whole genome comparative approaches. First, GWAMAR uses eCAMBer [7] to unify protein-coding gene annotation of bacterial strains and identify gene families. Next, it computes multiple alignment using MUSCLE [1] in order to generate information for phylogenetic tree construction and finding point mutations. Finally, it identify associations between the input phenotype (i.e. point mutation) and genotype (drug response of bacterial strains) data by computing several association scores.

The goal of the system is to computer association score between a given drug profile and a given mutation profile. Given a set of strains, a drug profile is a vector that indicates whether each strain is susceptible (S) or resistant (R) with the drug (Figure 1). An organism is called susceptible to a drug when the infection caused by it is likely to respond to treatment with this drug, at the recommended dosage. Resistant implies that the organism is expected not to respond to a given drug, irrespective of the dosage and of the location of the infection [4]. A mutation profile is a binary vector that indicates whether the mutation occur (1) or do not occur (0) on each strain (Figure 1). For a given drug profile and a mutation profile, five statistical scores can be computed by using GWAMAR are weighted support (ws), tree-generalized hypergeometric score (tgh), hypergeometric score (lh), odds ratio (or), and mutual information (mi). Out of the five scores, two phylogenetic tree-aware statistic scores (tgh and ws) achieve better performance than the tree ignorant statistic scores (lh, or, and mi). These results indicate that incorporating information from the phylogenetic tree improves that performance of the association prediction.



Drug resiatnce profile	S	R	S	R	?	?	S	R	R	?	R
mutation1	0	0	0	0	0	0	0	1	1	1	1
mutation2	1	0	0	0	0	1	0	1	1	1	1
mutation3	0	0	0	0	0	1	0	0	0	1	0
mutation4	0	1	0	1	0	0	0	1	0	0	0

Figure 1 Drug resistance profile and mutation profile [5]

Wozniak et al. present two case studies, Mtu173 and Mtu broad, on the M. tuberculosis. There are 88 drug-resistance associations classified as high-confidence in the Tuberculosis Drug Resistance Mutation Database (TBDReaMDB) and these associations have been used as a gold standard. In each case study, five association scores were computed for each drug resistance-mutation associations. The mutation that has higher chance to associate with drug resistance is on the top of the prediction list. In order to compare the accuracy of different scoring methods, PR curve for each case study has been generated (Figure 2).

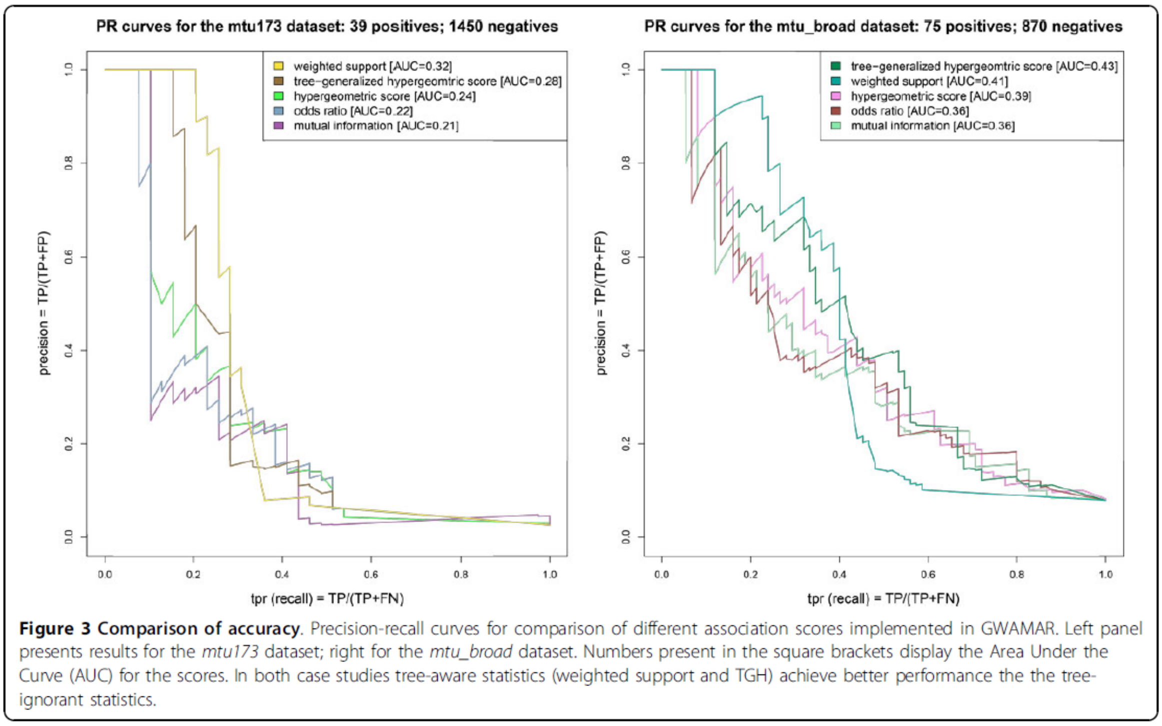


Figure 3 Comparison of accuracy. Precision-recall curves for comparison of different association scores implemented in GWAMAR. Left panel presents results for the *mtu173* dataset; right for the *mtu_broad* dataset. Numbers present in the square brackets display the Area Under the Curve (AUC) for the scores. In both case studies tree-aware statistics (weighted support and TGH) achieve better performance the the tree-ignorant statistics.

Figure 2 PR curves of the mtu173 and mtu_broad datasets [6]

Although several association scores have been calculated in order to identify drug resistance mutations, according to the comparison of accuracy (i.e. PR-curve) in [6] all individual methods have better than 50% precision only in the top half of their predictions. It is possible that a mutation in the top half of multiple methods are much more reliable than mutations in the top of an individual method. Therefore, in this project I propose an integrated method in order to take advantages of the high accuracy of the top half associations of all prediction methods.

Dataset

In this project, I focused on the mtu_broad dataset which obtained from GWAMAR website. This dataset is obtained from 1,398 strains of bacteria from Broad Institute database. The total number of mutations is 6,402. Note that these mutations are only from the sequences of genes of interest not the whole genome sequencing. In addition, I also downloaded the gold standard drug resistant mutations from the GWAMAR website. These gold standard drug resistant mutations are obtained from TB Drug Resistance Mutation Database (TBDReaMDB) which is a curated database for common mutations found for the major groups of anti-tuberculosis drugs. Six drugs considered in this project are Fluoroquinolones, Ethambutol, Isoniazid, Pyrazinamide, Rifampicin and Streptomycin.

Methods

For each case study the mutation rank of each scoring methods will be generated by using GWAMAR and compared. If the overlap between the top mutations all rank is small, then we expected to see that an integrated method has much better performance. This is because low overlap indicates that the results of different scoring methods are complement to each other. The following section describes different methods for integrating the score/rank of the five different association scores.

Average Rank

After all association scores have been calculated for each drug resistance-mutation association, we rank the score in descending order. The higher score implies higher possibility that the mutation is associated with drug resistance phenotype. Note that for tree-generalized hypergeometric score (tgh) and hypergeometric score (lh), $-log(p - value)$ is used as a score.

Integrated p-value

There are two p-values, p_{tgh} and p_{lh}, that can be used to integrate because the current version of GWAMAR do not provide p-values of the association for other types of scores (or, mi, ws). The integrated p-value is calculated and transformed by negative log. Therefore, the higher score, the higher possibility that the mutation is associated with drug resistance phenotype.

$$-\log(p_{tgh}) \pm \log(p_{lh}) = -(\log(\mathrm{p_{tgh}}) + \log(p_{lh})) = -\log(p_{tgh}p_{lh})$$

Summation of scores

This method sum up all the raw association scores of the five types of scores calculated from GWAMAR. In addition, we also calculate the summation of the tree based association scores (or+mi+lh) and the summation of the non-tree based association scores (ws+tgh).

Summation of normalized scores

All raw association scores are scaled to [0,1] where 0 is the minimum score and 1 is the maximum score. The summation is performed as the previous method.

Union

This methods union the top 10 associations from all five types of scores, from all tree based scores, and from all non-tree based scores.

Interpretation of the integrative scores

According to the PR curve in Figure 2, mutations in the top rank of the individual method have better precision. If we combine the scores or rankings of the top associations of those methods, it is possible that we can obtain better results than the individual methods. Several methods can be used to integrate the scores or rankings of the individual methods.

First, the simple way is to average the rank obtained from all five individual methods. This method maintains the associations which are in the top rank of all individual methods in the top rank of the integrated method. This method can have problem when there are too many associations and the range of the score is narrow because large number of associations can have the same scores and the ranking might not make sense. However, in the mtu_broad data set, the number of association is not large, so the rankings are still useful.

Second, the summation of scores from all five individual methods can be considered as an integrated score. Because all scores have the same interpretation i.e. the higher score the higher probability that the mutations are associated with drug resistance phenotype. The summation of the raw scores might not be a good integrated score because the scales of the five association scores are quite different. Therefore, the summation of the normalized score can also be used. The five association scores are scaled to 0 to 1 where 0 is the minimum score and 1 is the maximum score.

Finally, since Figure 2 shows that the top associations of all methods are reliable, so the union of the top rank, i.e. top 10, can be more reliable than the top associations of the individual methods.

Validation

The drug resistance mutations from the TBDReaMDB are used to validate the association list obtained from the integrated methods. The gold standard associations are 88 high confidence associations from the TBDReaMDB. Among the 88 high confidence associations 74 are also in the 6,402 mutations in the mtu_broad dataset, so the 74 associations will be the positives. Two validation methods are used to validate the results of the integrated methods. First, the coverage of the top 100 association are calculated and plotted. The x axis is the ranking of the associations and the y axis is the percentage of the associations that are positive. The second validation methods is calculating AUC of the PR curves. The negatives are randomly chosen from the identified putative associations and the number of negatives equals to the total number of mutations present in genes which has at least one high-confidence mutation. For mtu_broad dataset, the number of negative is 870. Because the negatives are randomly chosen, the AUC might not be the same for each run. Therefore, I randomly select 870 putative associations to be negatives and calculate the AUC for 1,000 times. Then the boxplot of the AUC can be used to compare the performance of the integrated methods and the individual methods.

Results

For a given drug profile and a mutation profile, five association scores are computed by using GWAMAR. The five association scores are weighted support (ws), tree-generalized hypergeometric score (tgh), hypergeometric score (lh), odds ratio (or), and mutual information (mi). Out of the five scores, two are phylogenetic tree based association scores (tgh and ws) and three are non-tree based association scores (lh, or, and mi). There are five integrated scores: average ranking (avg), integrated p-value (lh+tgh), summation of normalized non tree-based score (mi+or+lh (norm)), summation of all scores (sum), summation of all normalized scores (sum (norm)).

Figure 3A shows the high confidence association coverage of the top 100 associations of the individual methods. The x axis is the number of associations and the y axis is the percent coverage (of the 74 positives). Weighted support (ws) performs the best because its top 100 associations cover more positives. Figure 3B shows the association coverage of the top 100 associations of the individual methods where the y axis is the percent coverage (of the 212 associations in TBDReaMDB). In this case, tree-generalized hypergeometric score (tgh) is also good as the ws. From Figure 3A-B, it is easy to see that when we consider the top 100 associations the information from the phylogenetic tree improves the performance of the association prediction methods, tgh and ws.

Figure 3C shows the high confidence association coverage of the top 100 associations of the integrated methods (color) and the individual methods (gray). From the figure, if we only consider the top 100 associations, the integrated methods cannot perform better than ws. Among the integrative methods, average ranking, is the best for top 50 associations. Figure 3D shows the association coverage of the top 100 associations of the integrated methods. From this figure, we can see that the average ranking perform the best and comparable to tgh and ws.

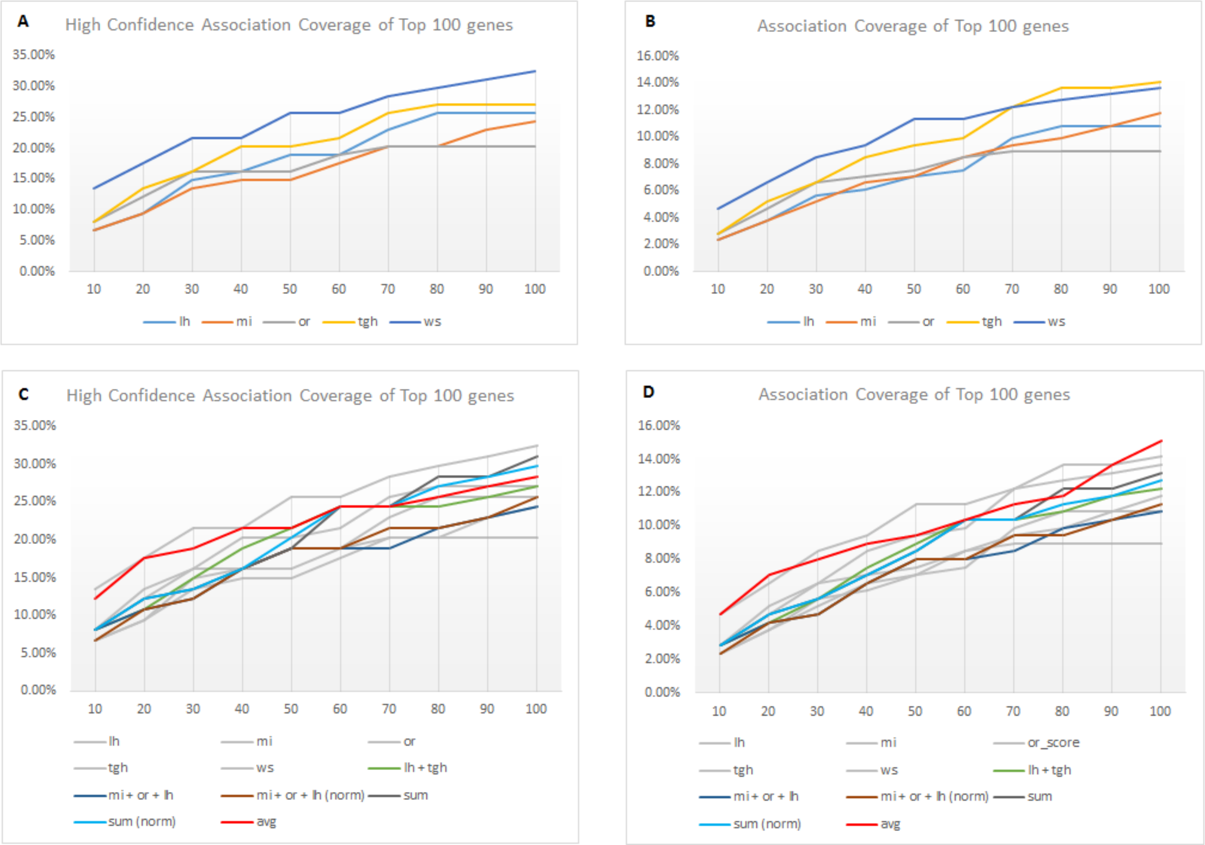


Figure 3 Coverage of the top 100 associations

According to the positives and negatives defined in the Validation section, Figure 4 shows the boxplot of the AUC of 1,000 PR curves of each association score. The gray boxes are the AUC of the individual methods and the blue boxes are the AUC of the integrated methods. We can see that the integrated p-values (lh+tgh), average ranking, and summations are as good as or better than tgh. The average ranking has the best AUC median 0.4328. The summation and the summation of the normalized score have 0.4303 and 0.4284 respectively. These three integrated methods have slightly better AUC then the tgh (0.4280) which is the best method among the individual methods. This results suggested that the integrated scores are slightly better than the individual scores. In addition, we can also see from the AUC boxplot that the score normalization can slightly improves the performance of the association prediction.

Moreover, one might interesting only the top 50 or top 100 predicted associations. Figure 5-6 shows the AUC of the top 50 and top 100 predicted associations of both individual and integrated methods. The ws has the best AUC (0.649) follow by average ranking (0.642), tgh (0.544) and or (0.553). This suggest that when we consider only top 100 predicted associations, ws perform the best. When we consider top 50 predicted associations, Figure 6 shows that average ranking has the best AUC (0.759) follow by ws (0.746).

When we consider only the 74 positive associations (occur in both putative mutations and TBDReaMDB), 45 associations have score > 0 for all five individual scores and 20 association have scores > 0 for four of the five individual scores (Figure 7). This suggests that it might be possible to use the frequency (number of scores that predict a given association). However, if we consider all 6,402 associations, 3,344 also have scores > 0 for all five individual scores (Figure 8). This suggests that we cannot just use the number of scores predicting a given association to perform the prediction because there will be too many false positives.

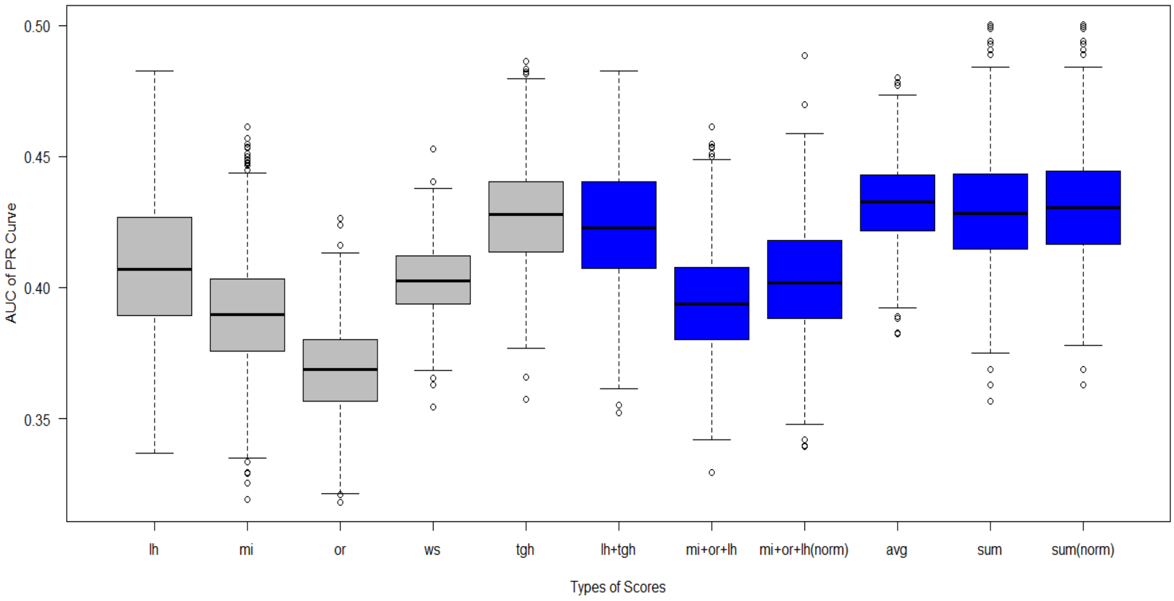


Figure 4 AUC of 1,000 PR curves of association scores

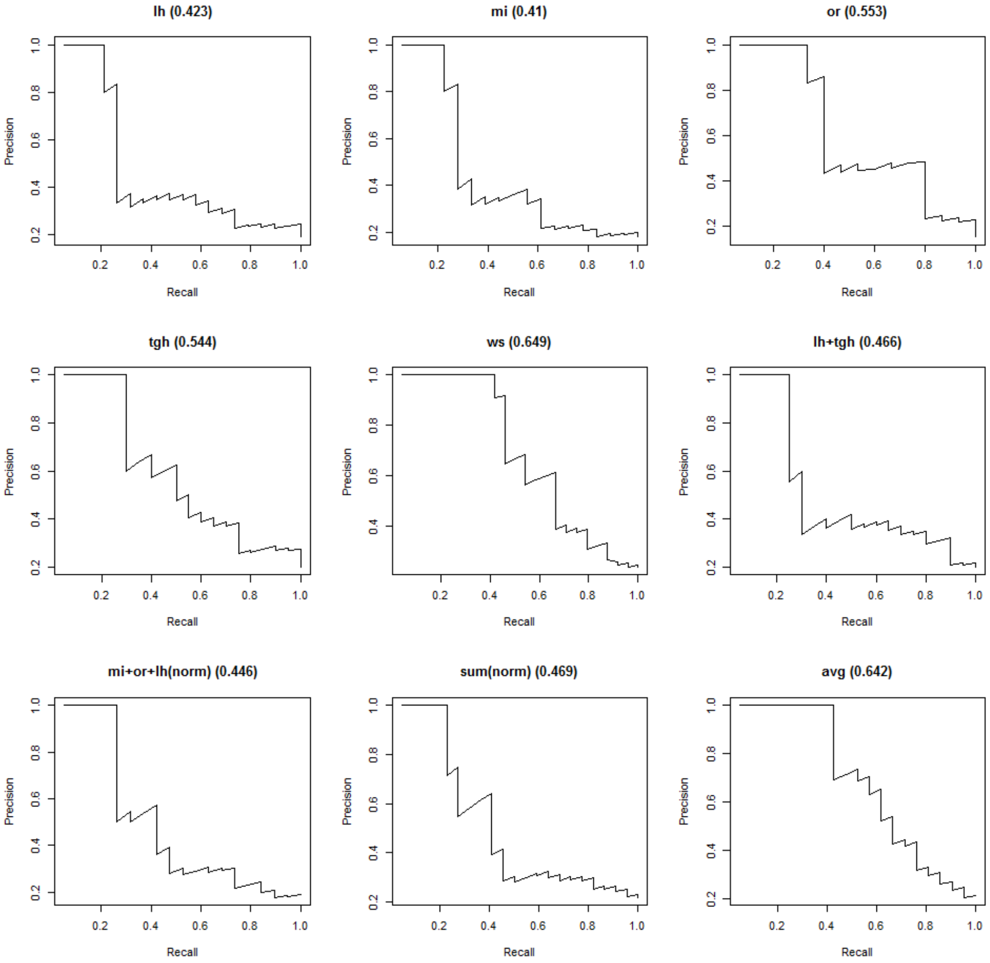


Figure 5 PR curve of top 100 associations of association scores

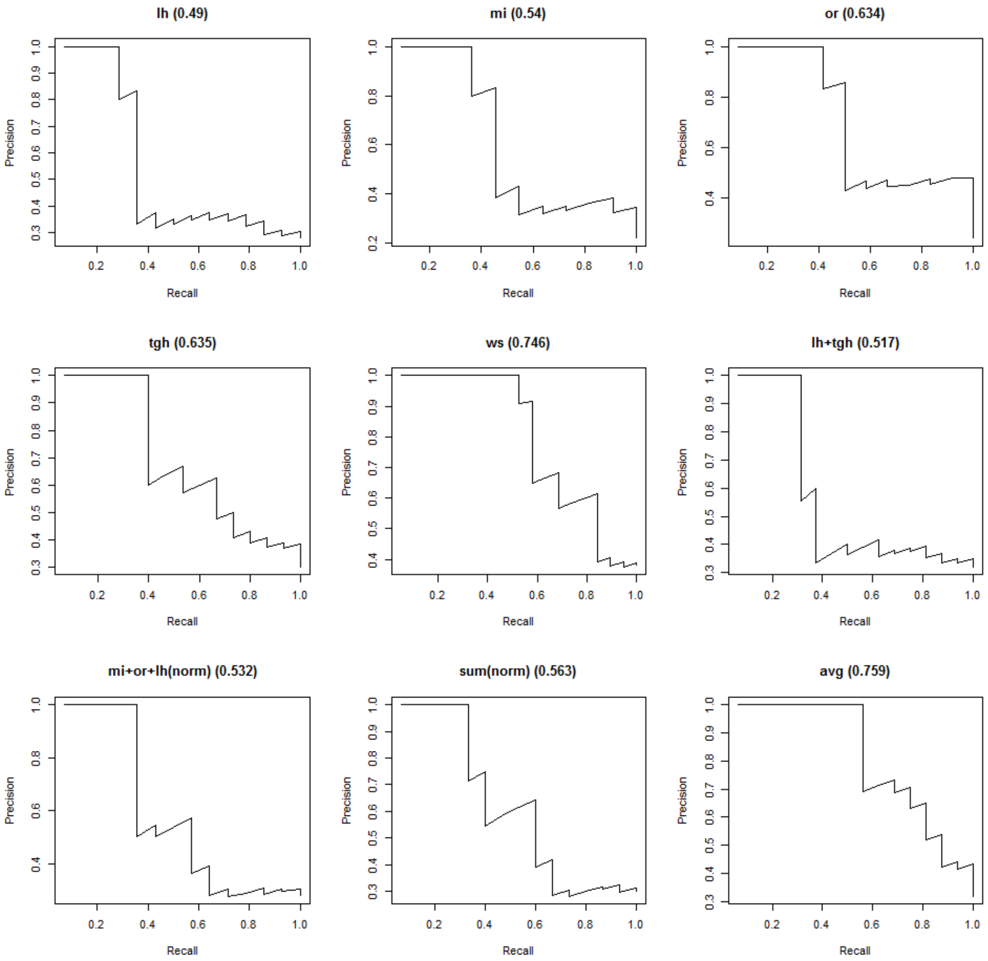
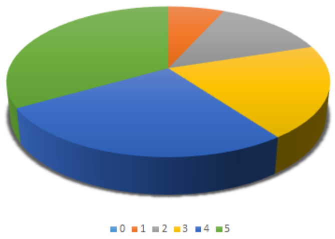
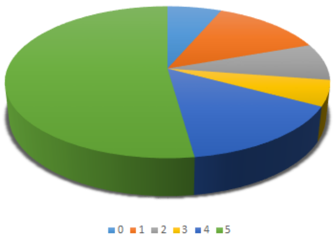


Figure 6 PR curve of top 50 associations of association scores



# scores	# high confidence associations
0	3
1	0
2	6
3	0
4	20
5	45

Figure 7 Frequency of scores predicting the 74 high confidence associations



# scores	# putative associations
0	410
1	840
2	505
3	348
4	955
5	3344

Figure 8 Frequency of scores of putative associations

Table 1 Frequency of the union of the top 10 associations of all 5 individual scores

High Confidence Associations		**Not High Confidence Associations**	
# scores	**# associations**	**# scores**	**# associations**
1	4	1	4
2	2	2	2
3	0	3	3
4	1	4	0
5	4	5	0

Now instead of consider all 6,402 associations, we consider the union of the top 10 associations of all five individual scores. The total number of associations is 21 (11 is the high confidence associations and 10 is not the high confidence associations). According to Table 1, there are 4 high confidence associations that are in the top 10 of all 5 individual methods and 1 high confidence associations that are in the top 10 of 4 individual methods. In contrast, there is no non-high confidence associations that are in at least 4 individual methods. This result suggested that the frequency (number of scores) can also be used to predict the drug resistant-mutation associations. In addition, from Table 2-3 it is interesting that when we consider the union of top 25 and top 50 of all five individual scores, if an association is in top 25 or 50 of all five individual scores, then it is a high confidence association.

Table 2 Frequency of the union of the top 25 associations of all 5 individual scores

High Confidence Associations		**Not High Confidence Associations**	
# scores	**# associations**	**# scores**	**# associations**
1	7	1	29
2	3	2	9
3	1	3	6
4	0	4	1
5	8	5	0

Table 3 Frequency of the union of the top 50 associations of all 5 individual scores

High Confidence Associations		**Not High Confidence Associations**	
# scores	**# associations**	**# scores**	**# associations**
1	8	1	70
2	6	2	20
3	2	3	15
4	0	4	6
5	9	5	0

For the union of the top 50 associations of all 5 individual scores, there are 136 associations in total. Therefore, we can consider the frequency (1 to 5) as a score and plot the coverage as in Figure 3. However, the coverage in this case is not better than the average ranking.

Discussion

Several methods have been proposed to identify drug resistant mutations by providing an association score of drug resistant phenotype and a mutations. The gold of this project is to find a method to integrate the existing association scores in order to get a better prediction. In this project, I propose methods to integrate results from five different association scores computed by GWAMAR systems. Two phylogenetic tree-aware statistic scores, weighted support (ws) and tree-generalized hypergeometric score (tgh), perform the best among the five individual methods.

The average ranking and the summation of normalized scores perform the best among the proposed integrated methods and are better than the individual methods in some experiments. The average ranking also has the maximum median of AUC of 1,000 among all association scores. In addition, the summation and the summation of normalized scores also have the median of AUC greater than all individual scores. These results suggest that by integrating all the individual score together, the performance of the prediction slightly increase.

When we only consider the top 50 and top 100 associations of each individual score, the weighted support has the highest AUC, 0.746 and 0.649, respectively. The best integrated methods is the average ranking (AUC = 0.759 and 0.642). For the top 50 associations, the average ranking has the highest AUC, although it is just slightly higher than the weighted support. This suggests that if we only consider to top rank of the proposed integrated scores, the results are not better than the individual scores.

In addition, I union the top 10, top 25, and top 50 associations of all 5 individual scores together. Interestingly, I found that if an association is in all top 10-50 of all individual scores, then it is a high confidence association in TBDReaMDB which is a gold standard used in this project. These results suggest that the associations that is in the top rank of all types of association scores are very reliable. However, for the results of the union, the recovery rate is still low. Among the union of the top 50 associations, only 9 associations are in all the top 50 rankings.

To sum up, this project confirmed that the associations that are in top rank of several methods are more reliable. Although the results of the integrated scores are only slightly better than the individual scores, the results suggest that we need more information to help identifying the drug resistance mutations. Weighted support (ws) and tree-generalized hypergeometric score (tgh) already incorporated the information from phylogenetic tree. However, there could be other possible mechanisms effect the drug sensitivity. Incorporate gene expression and epigenomic data might be helpful for increase recovery rate.

Finally, the high confidence association obtained from TBDReaMDB is not a real gold standard. There could be several mutations identified by both individual and integrated methods that are indeed associated with drug resistance phenotype, but the mutations have not been well studied yet, so they are in the TBDReaMDB.

References

[1] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic acids research, 32(5):1792{1797, 2004.

[2] World Health Organization. Antimicrobial resistance, 2014.

[3] World Health Organization. Tuberculosis, 2014.

[4] Jozef Vandepitte, J Verhaegen, Kraesten Engbaek, Patrick Rohner, Peter Piot, CC Heuck, et al. Basic laboratory procedures in clinical bacteriology. Number Ed. 2. World Health Organization, 2003.

[5] Michal Wozniak, Jerzy Tiuryn, and Limsoon Wong. An approach to identifying drug resistance associated mutations in bacterial strains. BMC genomics, 13(Suppl 7):S23, 2012.

[6] Michal Wozniak, Jerzy Tiuryn, and Limsoon Wong. Gwamar: Genome-wide assessment of mutations associated with drug resistance in bacteria. BMC genomics, 15(Suppl 10):S10, 2014.

[7] MichalWozniak, LimsoonWong, and Jerzy Tiuryn. ECamber: efficient support for large-scale comparative analysis of multiple bacterial strains. BMC bioinformatics, 15(1):65, 2014.