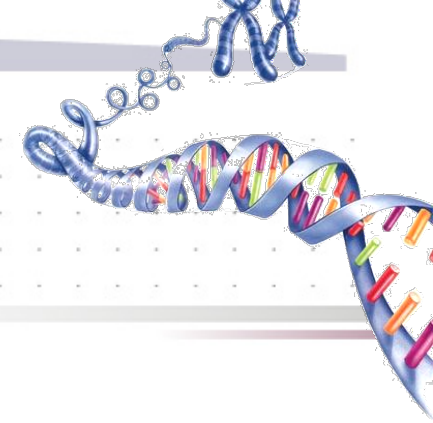


*Comparison of Five Commonly Used  
Gene-Gene Interaction Detecting  
Methods in Schizophrenia*

Chung-Keng Hsieh and Guan-Hua Huang

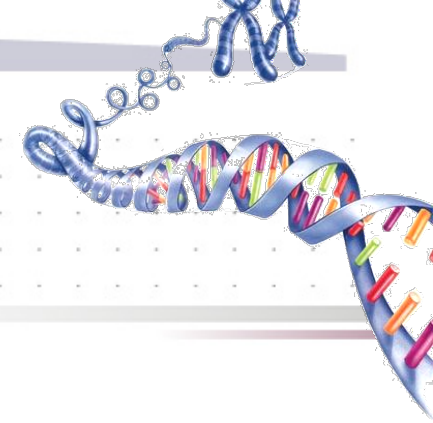
Institute of Statistics

National Chiao Tung University



# Outline

- ◆ INTRODUCTION
- ◆ METHODOLOGY
  - ◆ Study population
  - ◆ Preliminary analyses
  - ◆ Study design
  - ◆ Methods
  - ◆ Cross validation
- ◆ RESULTS
- ◆ CONCLUSION

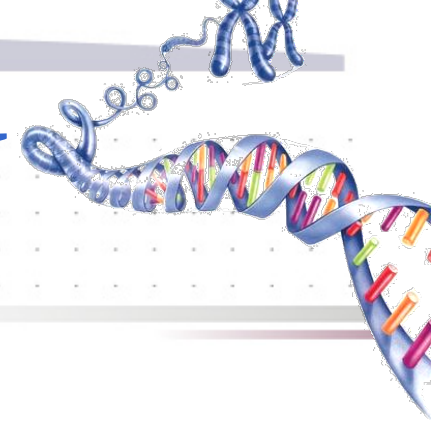


# Outline

## ◆ INTRODUCTION

- ◆ METHODOLOGY
  - ◆ Study population
  - ◆ Preliminary analyses
  - ◆ Study design
  - ◆ Methods
  - ◆ Cross validation
- ◆ RESULTS
- ◆ CONCLUSION

# INTRODUCTION

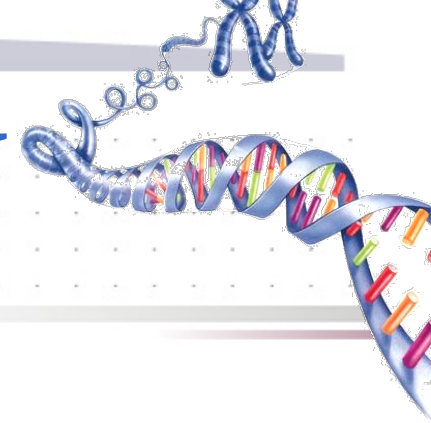


- ◆ Single-locus methods
- ◆ Gene-gene interaction

◆ Genotype data → **SNP**

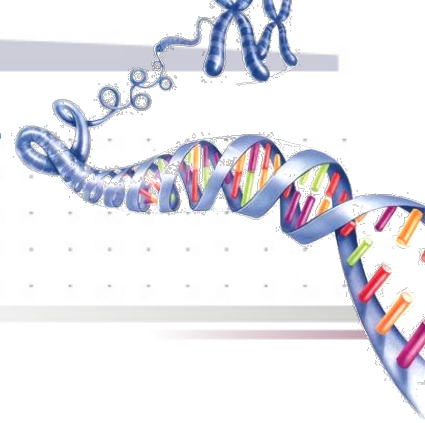
◆ Haplotype data → **Haplotype Block**

# INTRODUCTION

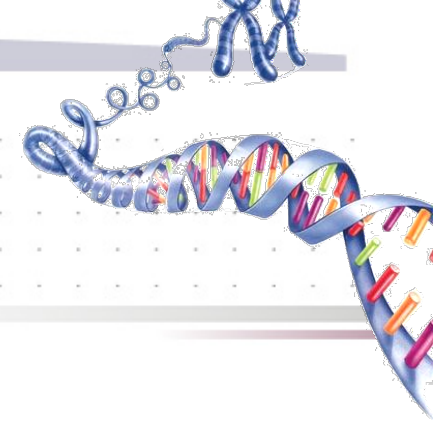


- ◆ In the present study:
  - ◆ Assessed the importance of gene-gene interactions on schizophrenia risk
  
- ◆ Data:
  - ◆ 65 SNPs from 5 candidate genes
  - ◆ 514 cases and 376 controls

# INTRODUCTION



- ◆ Five commonly used gene-gene interaction detecting methods
- ◆ Cross validation



# Outline

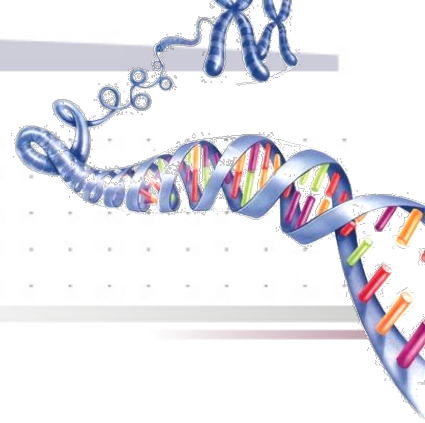
## ◆ INTRODUCTION

# ◆ METHODOLOGY

- ◆ **Study population**
- ◆ Preliminary analyses
- ◆ Study design
- ◆ Methods
- ◆ Cross validation

## ◆ RESULTS

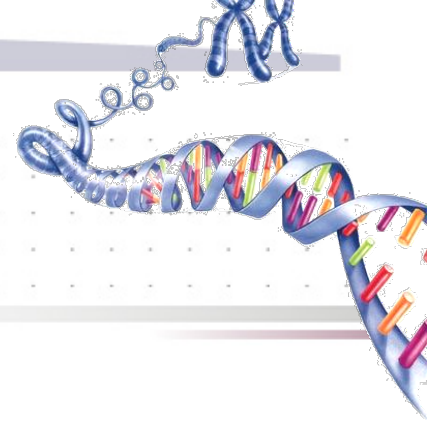
## ◆ CONCLUSION



# Study population

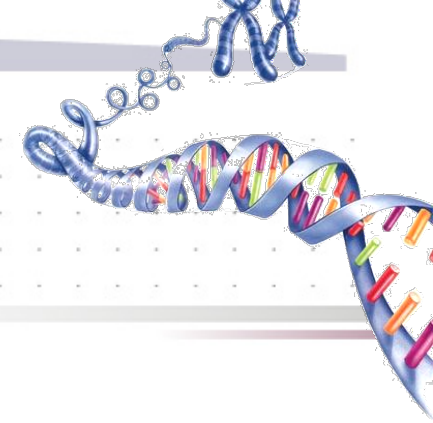
- ◆ Schizophrenia dataset
  - ◆ Data collection was based on TSLs program
  
- ◆ Genotyping of markers on **5** candidate genes:
  - ◆ ***DISC1, NRG1, DAO, G72 and CACNG2***





# Study population

- ◆ **514** schizophrenia cases and **376** controls
- ◆ Total **65** SNPs in five candidate genes



# Outline

## ◆ INTRODUCTION

# ◆ METHODOLOGY

- ◆ Study population
- ◆ **Preliminary analyses**
- ◆ Study design
- ◆ Methods
- ◆ Cross validation

## ◆ RESULTS

## ◆ CONCLUSION

# Preliminary analyses



## ◆ Data quality control:

### ◆ exclude SNP if

- ◆ HWE p value  $< 0.001$
- ◆ missing genotypes  $> 25\%$  (SNP call rate  $< 75\%$ )
- ◆ MAF is less than 1%

### ◆ exclude individuals if

- ◆ percentage of missing SNPs  $> 50\%$

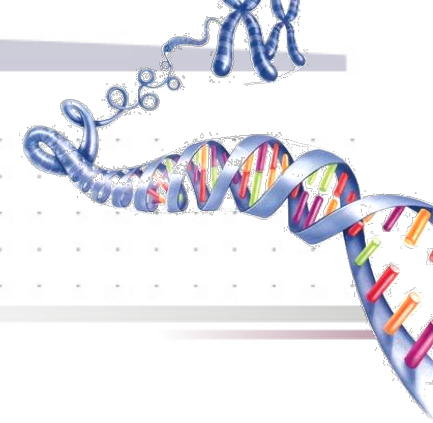
## ◆ After filtering data

- ◆ **55** SNPs
- ◆ **889** individuals (513 cases / 376 controls).

# Preliminary analyses



- ◆ Missing data imputation:
  - ◆ Imputation: replacing missing genotypes with predicted values that are based on the observed genotypes at neighboring SNPs.
- ◆ We implement data imputation by using the **MDR Data Tool** software
  - ◆ It will perform a simple frequency-based imputation.



# Outline

## ◆ INTRODUCTION

# ◆ METHODOLOGY

- ◆ Study population
- ◆ Preliminary analyses
- ◆ **Study design**
- ◆ Methods
- ◆ Cross validation

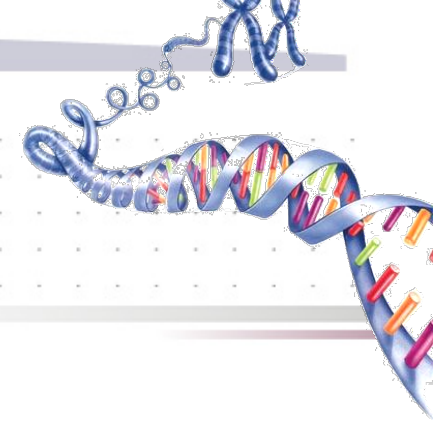
## ◆ RESULTS

## ◆ CONCLUSION



# Study design

- ◆ The data was analyzed by two strategies:
  - ◆ use the original genotype-based data
    - ◆ 55 SNPs
  - ◆ use the haplotype-based data
    - ◆ 10 Haplotype block + 29 SNPs
  
- ◆ In haplotype-based study, we use the **Haploview** software to define haplotype block and use the **PHASE** software to estimate individual's haplotype



# Outline

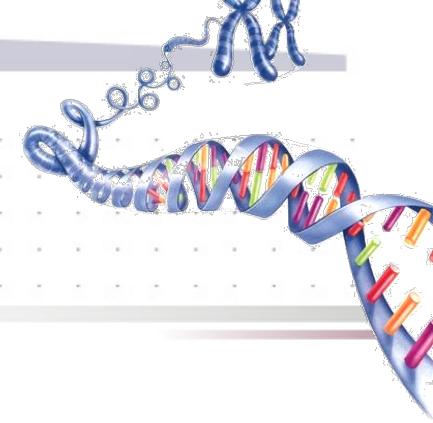
## ◆ INTRODUCTION

# ◆ METHODOLOGY

- ◆ Study population
- ◆ Preliminary analyses
- ◆ Study design
- ◆ **Methods**
- ◆ Cross validation

## ◆ RESULTS

## ◆ CONCLUSION



# Methods

◆ Chi-square test



◆ Logistic regression model (LRM)



◆ Bayesian epistasis association mapping (BEAM) algorithm



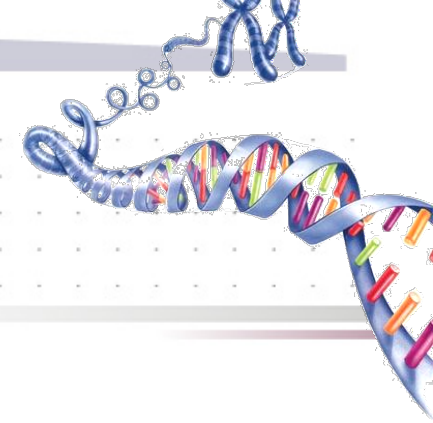
◆ Classification and regression trees (CART)



◆ Multifactor dimensionality reduction (MDR) method







# Outline

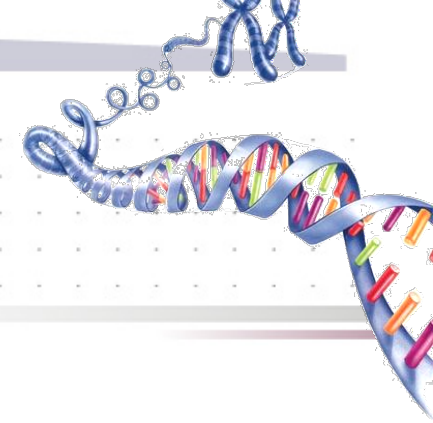
## ◆ INTRODUCTION

# ◆ METHODOLOGY

- ◆ Study population
- ◆ Preliminary analyses
- ◆ Study design
- ◆ Methods
- ◆ **Cross validation**

## ◆ RESULTS

## ◆ CONCLUSION



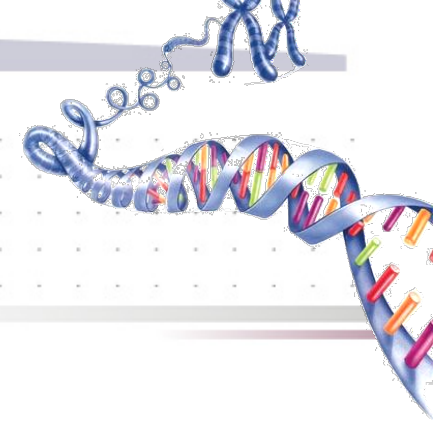
# Cross Validation

- ◆ We want to compare the abilities of prediction in these five methods
- ◆ We randomly divided our genotype-based data into training set and testing set.
  - ◆ The sample size of training set doubles that of testing set.
- ◆ We repeat this procedure 100 times to create 100 dataset



# Cross Validation

- ◆ For each CV, we apply the five methods to the training set and get the best model for one-way, two-way, and three-way interaction.
- ◆ We use the training set to build a prediction rule for the best model
  - ◆ Like MDR, we compute the case-control ratio for each genotype combination
  - ◆ While the prediction rule is built, we can calculate the prediction error



# Outline

- ◆ INTRODUCTION
- ◆ METHODOLOGY
  - ◆ Study population
  - ◆ Preliminary analyses
  - ◆ Study design
  - ◆ Methods
  - ◆ Cross validation

## ◆ RESULTS

- ◆ CONCLUSION

# RESULTS



Table 2.a. Single marker effects detected by the five methods in genotype-based data

rank	Chisq	LRM	BEAM	CART	MDR
1	<i>rsDAO_13</i>	<i>rsDAO_13</i>	<i>rsDAO_13</i>	<i>rsDAO_7</i>	<i>rsDAO_7</i>
2	<i>rsDAO_7</i>	<i>rsDAO_7</i>	<i>rsDAO_7</i>		<i>rsDAO_6</i>
3	<i>rsDAO_6</i>	<i>rsDAO_6</i>	<i>rsNRG1_6</i>		<i>rsNRG1_6</i>
4	<i>rsNRG1_6</i>	<i>rsNRG1_6</i>	<i>rsCACNG2_3</i>		<i>rsDAO_13</i>
5	<i>rsDISC1_38</i>	<i>rsDISC1_38</i>	<i>rsDISC1_38</i>		<i>rsDAO_8</i>

Table 2.b. Single marker effects detected by the five methods in haplotype-based data

rank	Chisq	LRM	BEAM	MDR
1	<i>DAO_block1</i>	<i>DAO_block1</i>	<i>DAO_block1</i>	<i>DAO_block1</i>
2	<i>G72_block2</i>	<i>G72_block2</i>	<i>CACNG2_block2</i>	<i>rsNRG1_6</i>
3	<i>rsNRG1_6</i>	<i>rsNRG1_6</i>		<i>DISC1_block4</i>
4	<i>CACNG2_block2</i>	<i>CACNG2_block2</i>		<i>DISC1_block2</i>
5	<i>rsDISC1_38</i>	<i>rsDISC1_38</i>		<i>G72_block2</i>

# RESULTS

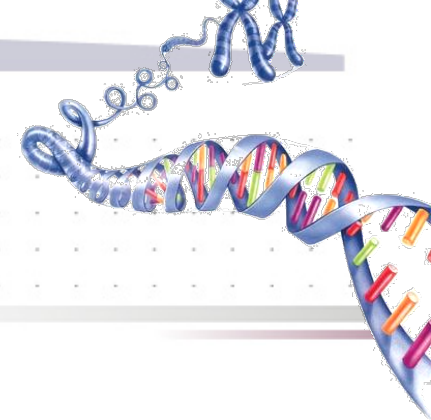
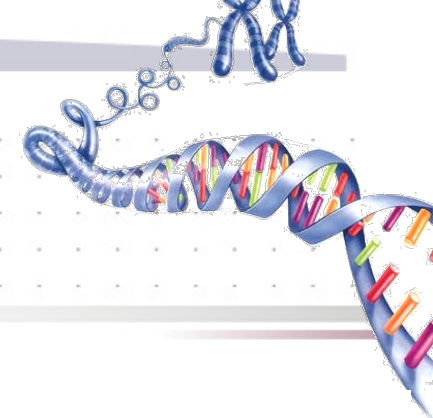


Table 3.a. Two-way interaction detected by the five methods in genotype-based data

rank	Chisq	LRM	BEAM	CART	MDR
1	<i>rsDAO_6 rsDAO_7</i>	<i>rsDAO_6 rsDAO_7</i>	<i>rsDISC1_E_7</i> <i>rsDISC1_E_4</i>	<i>rsDAO_7 rsDAO_8</i>	<i>rsNRG1_14 rsG72_16</i>
2	<i>rsNRG1_6 rsDAO_6</i>	<i>rsDAO_7 rsDAO_8</i>		<i>rsDAO_6 rsDAO_7</i>	<i>rsNRG1_6 rsDAO_6</i>
3	<i>rsNRG1_6 rsDAO_7</i>	<i>rsDAO_6 rsDAO_8</i>			<i>rsDISC1_3 rsDAO_7</i>
4	<i>rsDAO_7 rsDAO_13</i>	<i>rsDISC1_20 rsNRG1_6</i>			<i>rsDISC1_16 rsNRG1_6</i>
5	<i>rsDAO_6 rsDAO_13</i>	<i>rsDISC1_16 rsDISC1_20</i>			<i>rsDAO_6 rsDAO_7</i>

Table 3.b. Two-way interaction detected by the five methods in haplotype-based data

rank	Chisq	LRM	BEAM	MDR
1	<i>rsNRG1_6 G72_block2</i>	<i>rsDISC1_E_7 G72_block2</i>	No two-way interaction detected	<i>DISC1_block3 DAO_block1</i>
2	<i>DAO_block1 G72_block2</i>	<i>rsNRG1_6 CACNG2_block2</i>		<i>DISC1_block1 DAO_block1</i>
3	<i>G72_block2 CACNG2_block2</i>	<i>rsDISC1_E_7 rsCACNG2_3</i>		<i>DAO_block1 G72_block1</i>
4	<i>rsNRG1_6 DAO_block1</i>	<i>G72_block2 CACNG2_block2</i>		<i>DISC1_block4 DAO_block1</i>
5	<i>rsNRG1_6 CACNG2_block2</i>	<i>rsDISC1_38 CACNG2_block2</i>		<i>DISC1_block5 DAO_block1</i>



# RESULTS

Table 4.a. Three-way interaction detected by the five methods in genotype-based data

rank	Chisq	LRM	BEAM	CART	MDR
1	<i>rsDAO_6</i>	<i>rsDISC1_16</i>	<i>No three-way interaction detected</i>	<i>rsDISC1_E_7</i>	<i>rsNRG1_6</i>
	<i>rsDAO_7</i>	<i>rsNRG1_6</i>		<i>rsDAO_6</i>	<i>rsDAO_6</i>
	<i>rsDAO_13</i>	<i>rsDAO_6</i>		<i>rsDAO_7</i>	<i>rsG72_16</i>
2	<i>rsNRG1_6</i>	<i>rsDISC1_38</i>			<i>rsDISC1_12</i>
	<i>rsDAO_6</i>	<i>rsDAO_7</i>			<i>rsNRG1_6</i>
	<i>rsDAO_7</i>	<i>rsDAO_13</i>			<i>rsCACNG2_3</i>
3	<i>rsNRG1_6</i>	<i>rsDISC1_16</i>			<i>rsNRG1_6</i>
	<i>rsDAO_7</i>	<i>rsNRG1_6</i>			<i>rsNRG1_14</i>
	<i>rsDAO_13</i>	<i>rsCACNG2_3</i>			<i>rsG72_16</i>
4	<i>rsNRG1_6</i>	<i>rsNRG1_6</i>			<i>rsDISC1_16</i>
	<i>rsDAO_6</i>	<i>rsDAO_6</i>			<i>rsNRG1_6</i>
	<i>rsDAO_13</i>	<i>rsDAO_13</i>			<i>rsDAO_6</i>
5		<i>rsNRG1_6</i>			<i>rsNRG1_6</i>
		<i>rsDAO_7</i>			<i>rsDAO_6</i>
		<i>rsDAO_13</i>			<i>rsCACNG2_3</i>

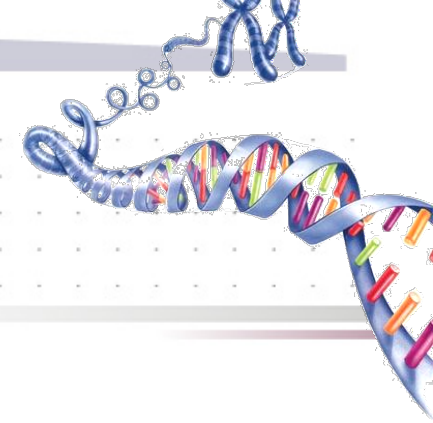


# RESULTS

Table 4.b. Three-way interaction detected by the five methods in haplotype-based data

rank	Chisq	BEAM	MDR
1	<i>G72_block2</i> <i>rsNRG1_6</i> <i>CACNG2_block2</i>	<i>No three-way interaction detected</i>	<i>DISC1_block1</i> <i>DISC1_block3</i> <i>DAO_block1</i>
2	<i>DAO_block1</i> <i>G72_block2</i> <i>rsNRG1_6</i>		<i>DISC1_block1</i> <i>DAO_block1</i> <i>G72_block1</i>
3	<i>DAO_block1</i> <i>rsNRG1_6</i> <i>CACNG2_block2</i>		<i>DISC1_block1</i> <i>DISC1_block4</i> <i>DAO_block1</i>
4	<i>DAO_block1</i> <i>G72_block2</i> <i>CACNG2_block2</i>		<i>DISC1_block3</i> <i>DISC1_block4</i> <i>DAO_block1</i>
5			<i>DISC1_block2</i> <i>DISC1_block4</i> <i>DAO_block1</i>





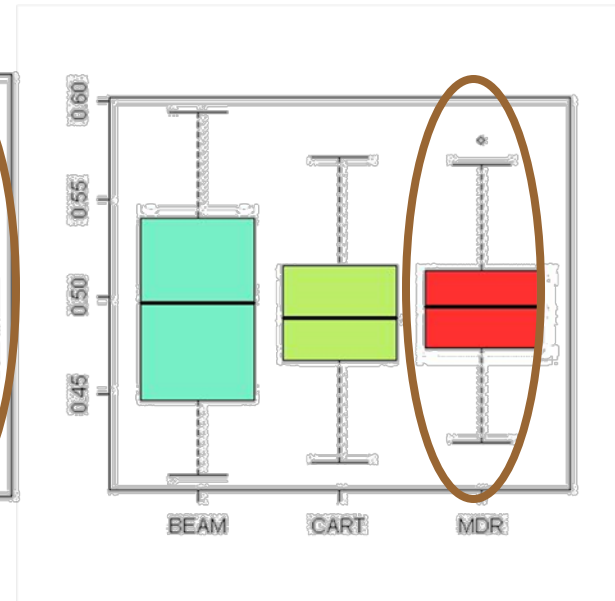
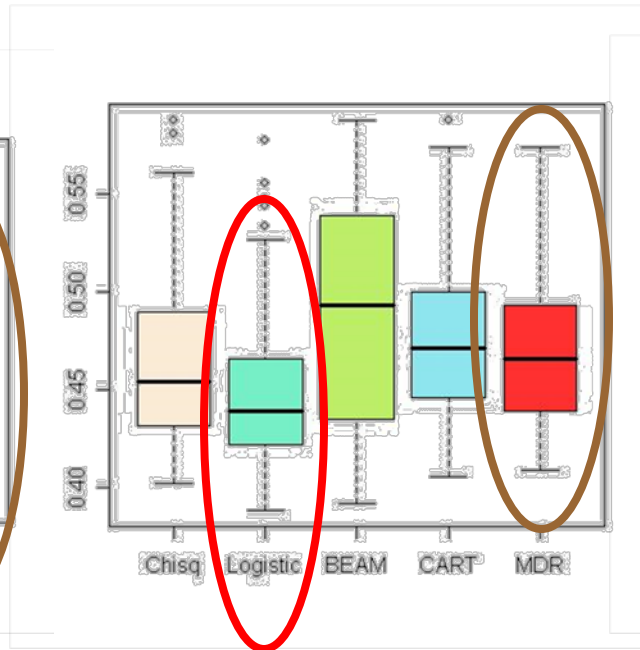
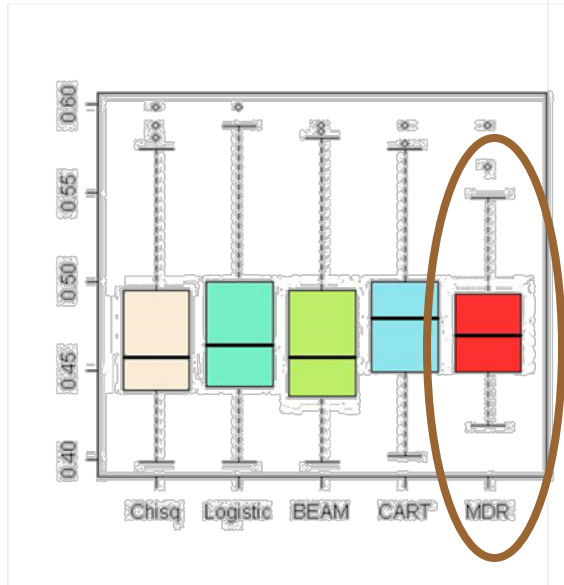
# RESULTS

Table 5. Average prediction error across 100 CVs

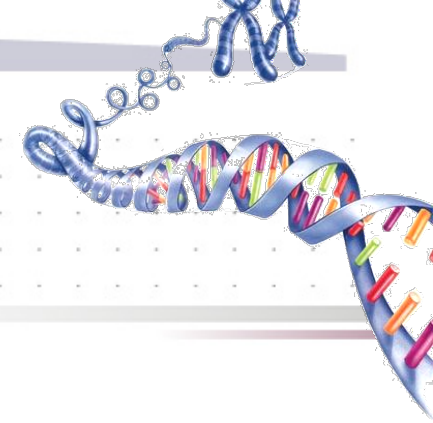
	Chisq	LRM	BEAM	CART	MDR
one-way	0.471283784	0.476047297	<b>0.471148649</b>	0.486824324	0.473783784
two-way	0.464207618	<b>0.448881209</b>	0.488123798	0.477674915	0.470942832
three-way			0.495776846	<b>0.491696159</b>	0.494607021

Two green arrows point downwards from the CART and MDR columns, indicating a downward trend in error for those models across the different cross-validation scenarios.

# RESULTS



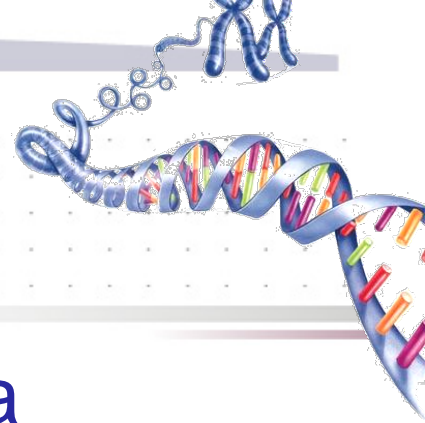
## Box-plot of prediction error



# Outline

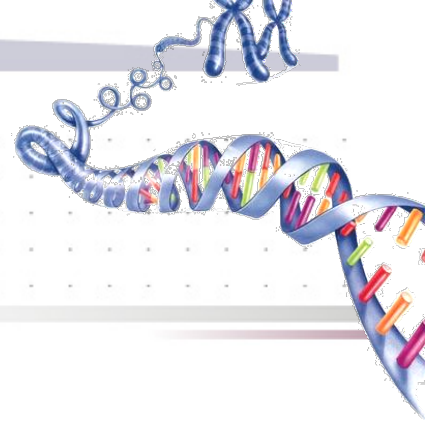
- ◆ INTRODUCTION
- ◆ METHODOLOGY
  - ◆ Study population
  - ◆ Preliminary analyses
  - ◆ Study design
  - ◆ Methods
  - ◆ Cross validation
- ◆ RESULTS

◆ CONCLUSION



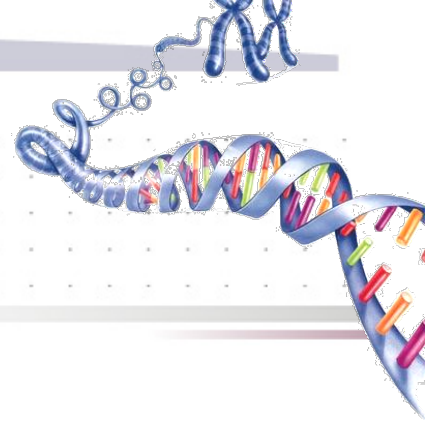
# CONCLUSION

- ◆ Our aim of this study is to propose a methodological issue in detecting gene-gene interaction
- ◆ We chose five commonly used methods and apply them to a schizophrenia data



# CONCLUSION

- ◆ we find that SNPs *rsDAO\_13* and *rsDAO\_7* have strong main effect
- ◆ SNPs *rsDAO\_6*, *rsDAO\_7*, and *rsG72\_16* have strong gene-gene interaction effects
- ◆ **LRM** shows the best predictive ability in our data



**THANK YOU!**