SIGNIFICANT VS PREDICTIVE

VARIABLE SELECTION AND PREDICTION

Mengying Sun

Apr. 20th 2016

Outline

- GWAS
- Missing Heritability
- Significant vs. Predictive Variables
- 3 Examples
- I Score
- Conclusions and comments

GWAS

- **G**enome-**W**ide **A**ssociation **S**tudy

An examination of common genetic variants in different individuals to see if any variant is associated with a trait.

$$\begin{array}{ccc} Trait\ Y & \quad & SNPs\ [X_1\ , \ldots X_j\ \ \ldots X_m] \\ \boxed{\begin{array}{c} y_1 \\ y_2 \\ \vdots \\ y_n \end{array}} & \quad & \boxed{\begin{array}{ccccc} x_{11} & \ldots & x_{j1} & \ldots & x_{m1} \\ \vdots & \ldots & x_{jk} & \ldots & \vdots \\ x_{1n} & \ldots & x_{jn} & \ldots & x_{mn} \end{array}} \\ n \times 1 & \quad & n \times m \end{array}$$

$j = 1,2, \ldots m$

$n = 5k$

$m = 500k, 1000k$

GWAS

- **G**enome-**W**ide **A**ssociation **S**tudy

An examination of common genetic variants in different individuals to see if any variant is associated with a trait.

$$\boldsymbol{Y} \quad \sim \quad \boldsymbol{X_j}$$

$$\underset{n \times 1}{\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}} \qquad \underset{n \times m}{\begin{bmatrix} x_{11} & \dots & x_{j1} & \dots & x_{m1} \\ \vdots & \dots & x_{jk} & \dots & \vdots \\ x_{1n} & \dots & x_{jn} & \dots & x_{mn} \end{bmatrix}}$$

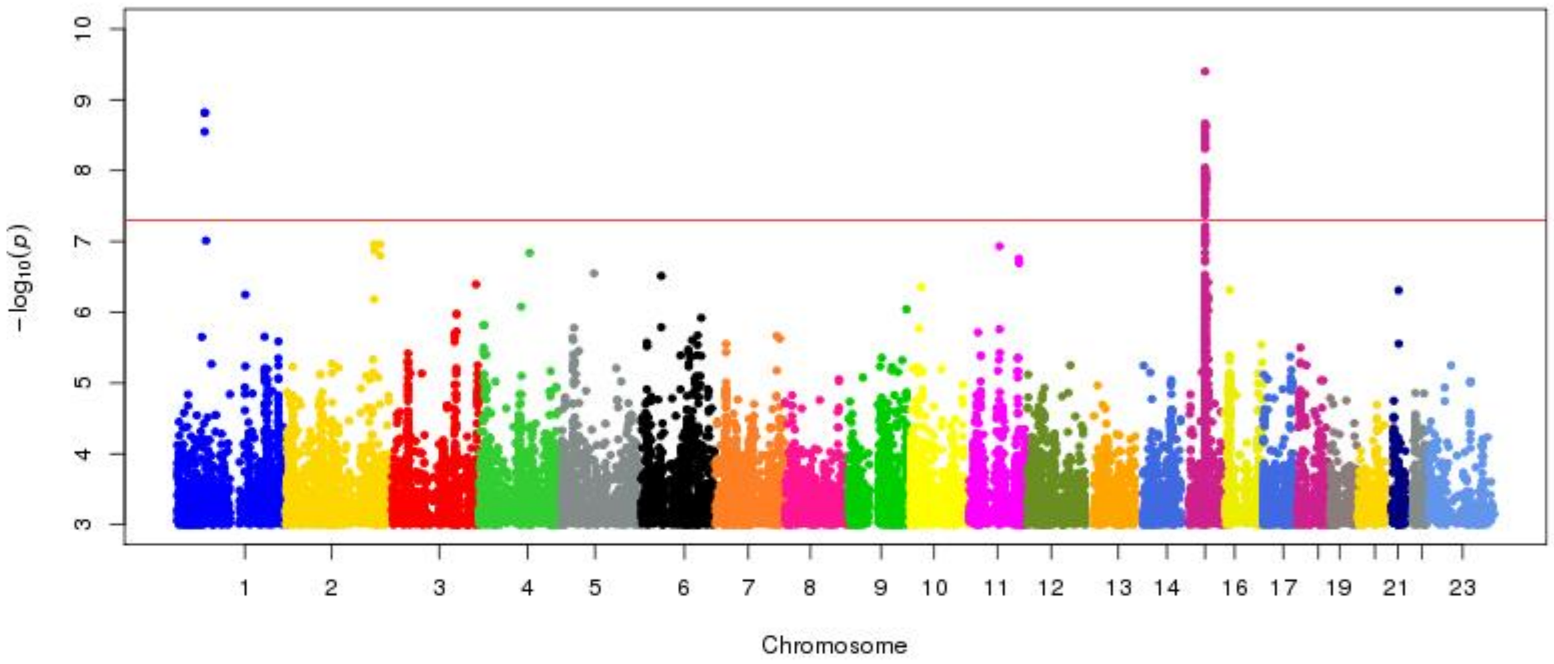
$j = 1,2, \dots m$

$n = 5k$

$m = 500k, 1000k$

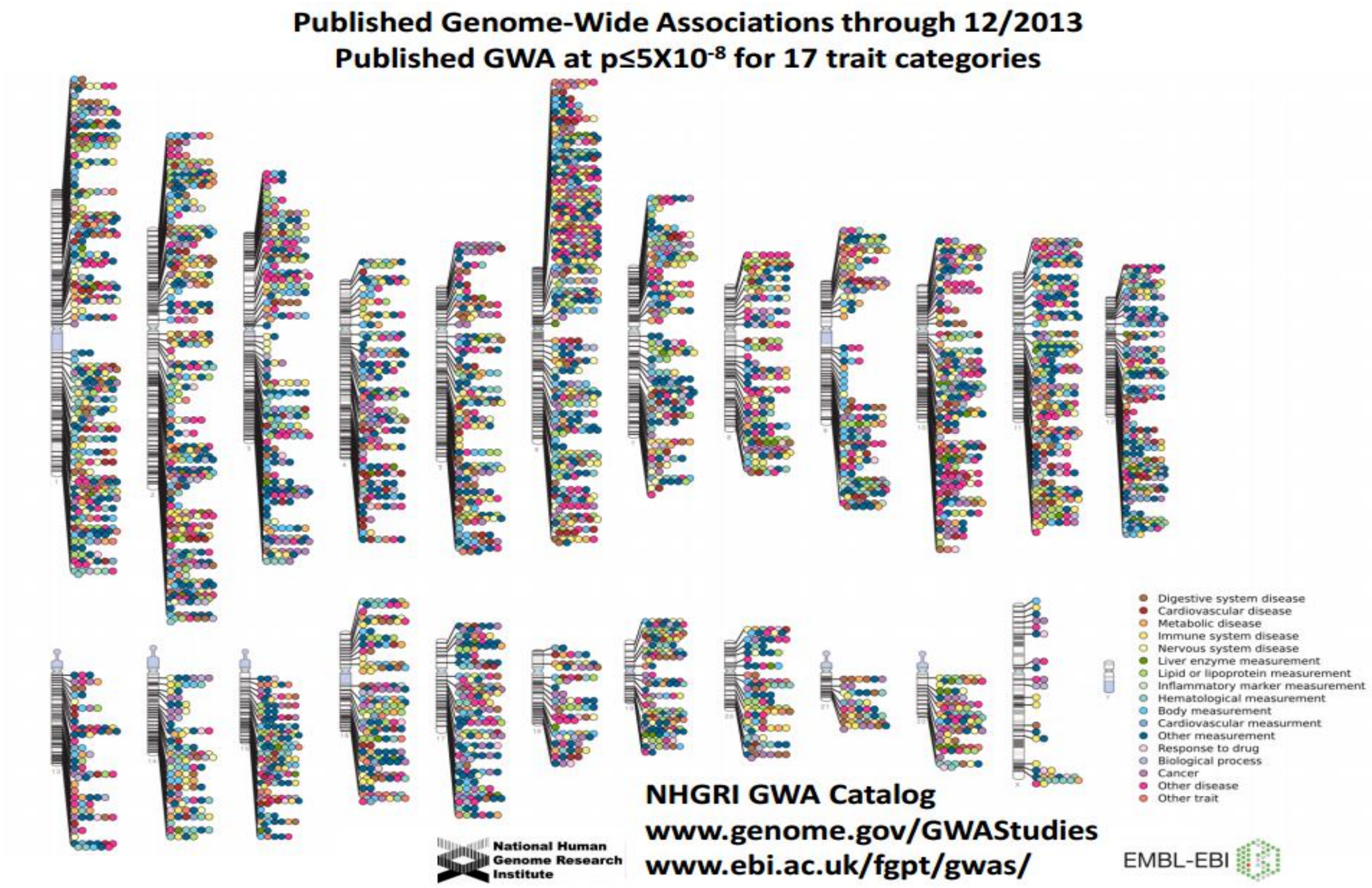
$$\boxed{P\ value = p_j}$$

Manhattan Plot



Styrkarsdottir, Unnur, Gudmar Thorleifsson, et al. 2014. “Severe Osteoarthritis of the Hand Associates with Common Variants within the ALDH1A2 Gene and with Rare Variants at 1p31.” *Nature Genetics* 46 (5): 498–502. doi:10.1038/ng.2957.

Published GWA



Published Genome-Wide Associations through 12/2013

Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories

NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

Hindorff LA, MacArthur J, Morales J, Junkins HA, Hall PN, Klemm AK, and Manolio TA. A Catalog of Published Genome-Wide Association Studies. www.genome.gov/gwastudies

Validation

$Y\ -\ trait$

$X\ -\ significant\ variable\ set$

$$Y = X\beta + \epsilon$$

$$\hat{\beta} \rightarrow \hat{y}$$

$$Y \sim \hat{y}$$

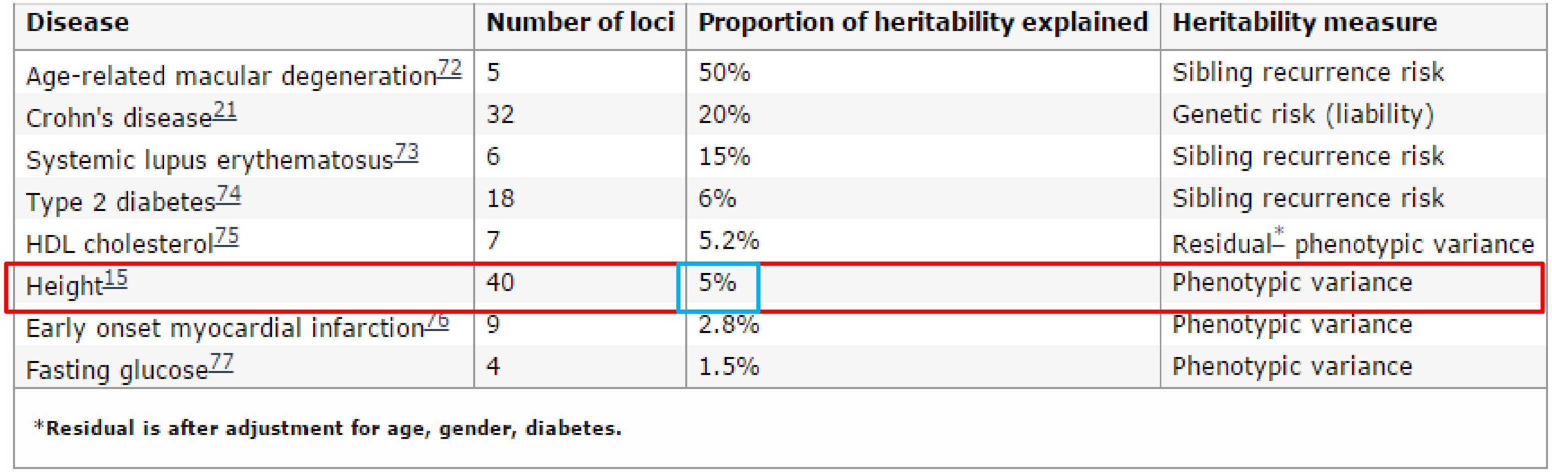
$r^2\ -\ coefficient\ of\ determination$

$cor(y, \hat{y}) - correlation\ between\ y\ \&\ \hat{y}$

Problem !!!

Missing Heritability

- Height is 80 – 90% heritable
- In GWAS, 40 genetic variants turned up associated with the height difference
- However, the variants accounted for only 5% of height’s heritability



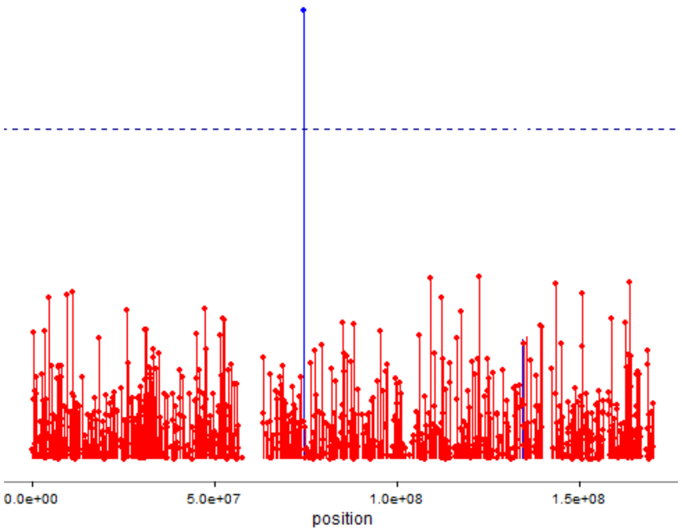
Disease	Number of loci	Proportion of heritability explained	Heritability measure
Age-related macular degeneration[72]	5	50%	Sibling recurrence risk
Crohn's disease[21]	32	20%	Genetic risk (liability)
Systemic lupus erythematosus[73]	6	15%	Sibling recurrence risk
Type 2 diabetes[74]	18	6%	Sibling recurrence risk
HDL cholesterol[75]	7	5.2%	Residual[*] phenotypic variance
Height[15]	40	5%	Phenotypic variance
Early onset myocardial infarction[76]	9	2.8%	Phenotypic variance
Fasting glucose[77]	4	1.5%	Phenotypic variance

*Residual is after adjustment for age, gender, diabetes.

Manolio, Teri A., Francis S. Collins, Nancy J. Cox, David B. Goldstein, Lucia A. Hindorff, David J. Hunter, Mark I. McCarthy, et al. 2009. “Finding the Missing Heritability of Complex Diseases.” *Nature* 461 (7265): 747–53. doi:10.1038/nature08494.

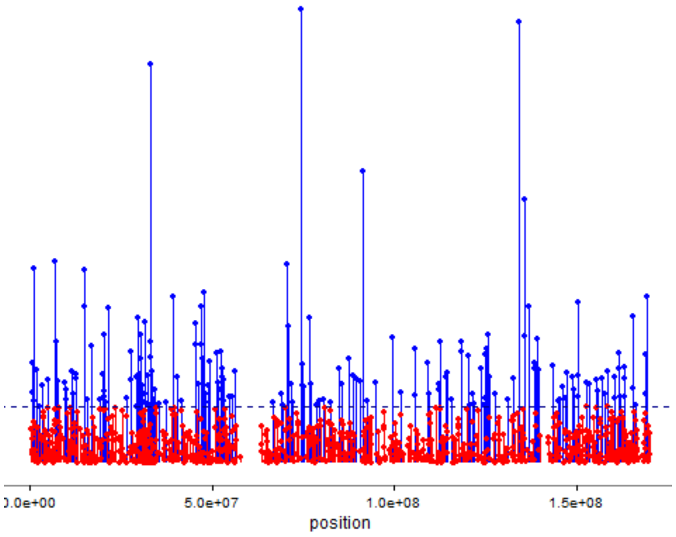
Missing Heritability

- Single locus



🙂

- Complex traits



🙃

EPI 855, Simulated data

Missing Heritability

- The way we select markers is WRONG ?

Maher, Brendan. 2008. “Personal Genomes: The Case of the Missing Heritability.” *Nature News* 456 (7218): 18–21. doi:10.1038/456018a

Paper

- Why significant variables aren’t automatically good predictors

Lo, Adeline, Herman Chernoff, Tian Zheng, and Shaw-Hwa Lo. 2015.
Proceedings of the National Academy of Sciences 112 (45): 13892–97. doi:10.1073/pnas.1518285112

PNAS

Why significant variables aren’t automatically good predictors

Adeline Lo[a], Herman Chernoff[b,1], Tian Zheng[c], and Shaw-Hwa Lo[c,1]

[a]Department of Political Science, University of California, San Diego, La Jolla, CA 92093; [b]Department of Statistics, Harvard University, Cambridge, MA 02138; and [c]Department of Statistics, Columbia University, New York, NY 10027

Contributed by Herman Chernoff, September 17, 2015 (sent for review December 15, 2014)

Thus far, genome-wide association studies (GWAS) have been disappointing in the inability of investigators to use the results of identified, statistically significant variants in complex diseases to

From the scientist’s point of view there are two basic problems, complicated by the large size of the data set. These are variable selection and prediction. For variable selection, we wish to find a

Significant vs. Predictive variables

- **Key Difference**

Different properties of their underlying distributions

$variable\ set\colon\ X$

f_0: The distribution of X among control

f_1: The distribution of X among cases

- **Variable selection (Significance)**

$H_0\colon f_0 = f_1 \quad \text{vs} \quad H_1\colon f_0 \neq f_1$

- **Prediction**

$x, observed\ value\ of\ X\ for\ a\ single\ individual, belongs\ to\ f_0\ or\ f_1$

Significant vs. Predictive variables

- **Variable selection (Significance)**

$H_0: f_0 = f_1 \quad \text{vs} \quad H_1: f_0 \neq f_1$

$$T_n$$

$$t_n = T_n(x)$$

$$\mathrm{P}(\mathrm{T_n} \ \geq t_n \mid \mathrm{H_0})$$

Reject $\mathrm{H_0}$ if $\mathrm{P}(\mathrm{T_n} \ \geq t_n \mid \mathrm{H_0})$ is sufficiently small

Significant vs. Predictive variables

- **Prediction**

x

$if\ f_0(x) < f_1(x), x\ is\ more\ likely\ from\ f_1$

$if\ f_1(x) < f_0(x), x\ is\ more\ likely\ from\ f_0$

Likelihood rate

$$prediction\ rate = 0.5 \sum_x \max(f_0(x)\,, f_1(x))$$

Significant vs. Predictive variables

- **Variable selection (Significance)**

$$H_0: f_0 = f_1 \quad \text{vs} \quad H_1: f_0 \neq f_1$$

$$\text{Reject } \mathrm{H}_0 \text{ if } \mathrm{P}(\mathrm{T}_\mathrm{n} \geq t_n \mid \mathrm{H}_0) \text{ is sufficiently small}$$

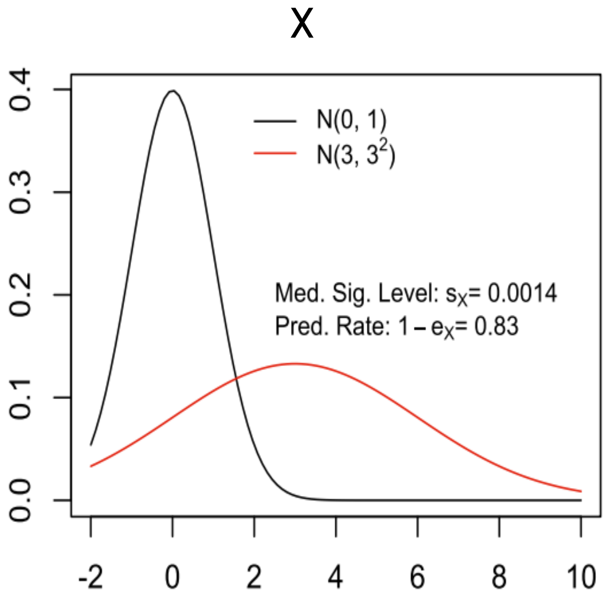
- **Prediction**

$$prediction\ rate = 0.5 \sum_x \max(f_0(x), f_1(x))$$

The former uses assumptions on, but no knowledge of, the EXACT distributions of the variables, whereas the latter, requires knowledge of both f_0 and f_1, which is the distribution of X

Example 1

- **Single Observation**



$H_0 : X \sim N\ (0, 1)$

$H_1 : X \sim N\ (3, 3^2)$

- Classifying the state of an individual (H/D) who yields the the observation X

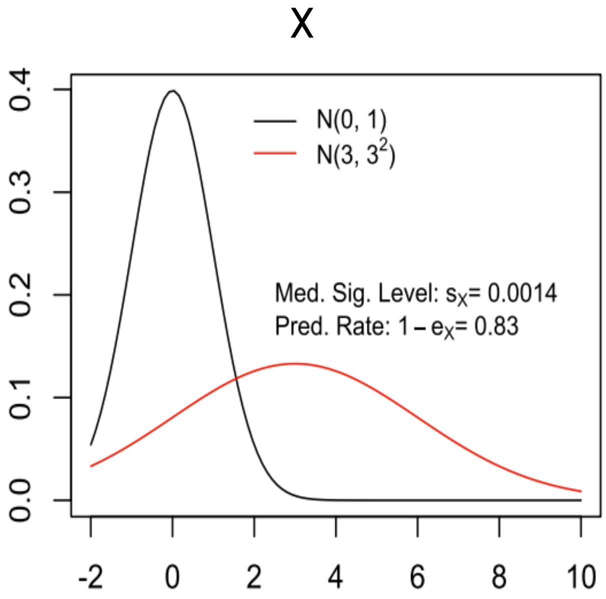
e_x

	T	F
T	√	e (c, H0)
F	e (c, H1)	√

- Reject H0 when likelihood ratio is large
- X is random, for every ratio c, 2*2
- choose c minimize the average of e(c,H0)+e(c,H1)

Example 1

- **Single Observation**



$H_0: X \sim N\ (0, 1)$

$H_1: X \sim N\ (3, 3^2)$

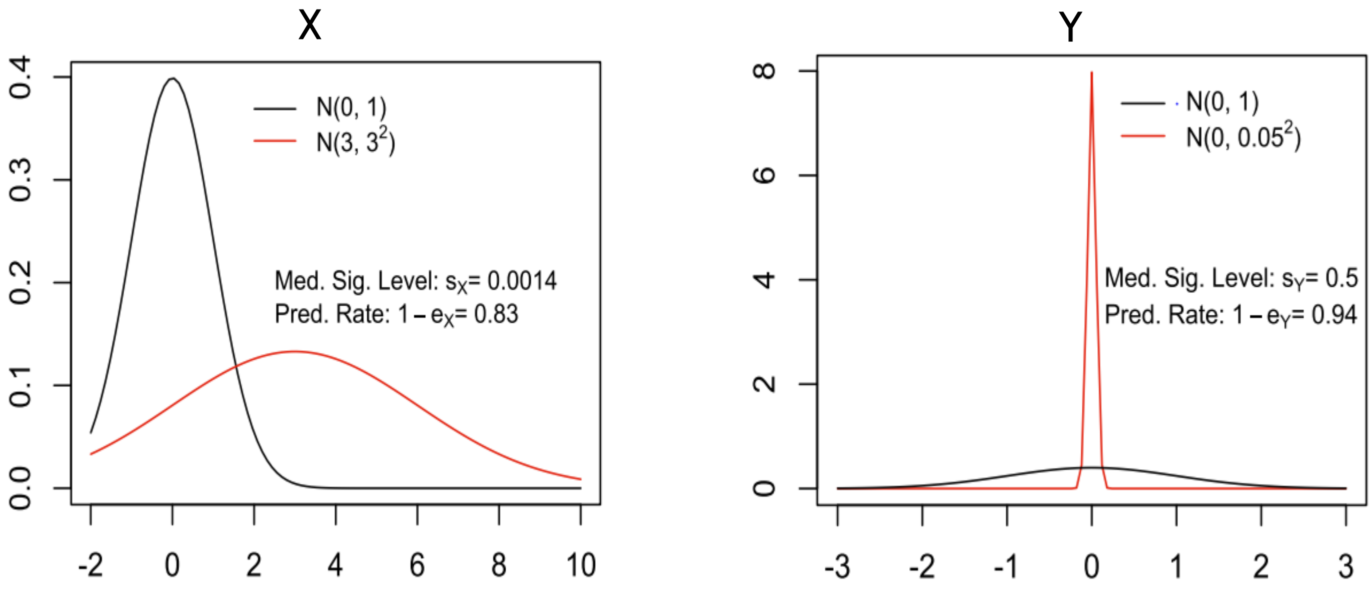
- Classifying the state of an individual (H/D) who yields the the observation X

s_x

- Large value of x will favor to reject H0
- For each observed x, P-value, choose the “median” value as significance level

Example 1

- **Single Observation**



- Scientist’s decision: whether observe X or Y
- Prefer X based on significance
- Prefer Y based on predictivity

Example 1

- **Conclusion**

  - It happens that a variable that is more significant serves not as good as another less significant variable in prediction, based on the different distribution of those two variables.

- **Some doubts about Example 1**

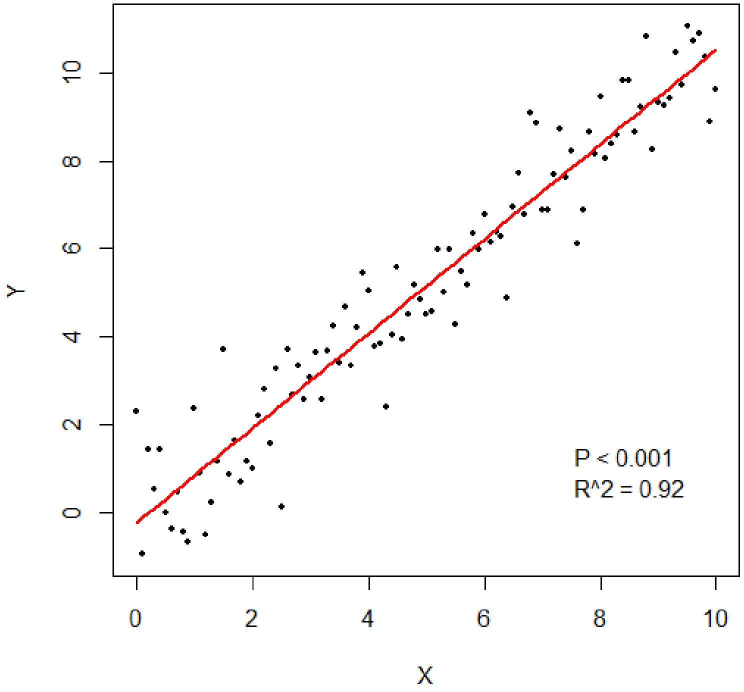
  - Was expecting more extreme case (0.83 vs 0.94)
  - Choice of measures of significance and predictability (ex and sx) not commonly used which is biggest problem throughout the paper
  - Confusing and absurd

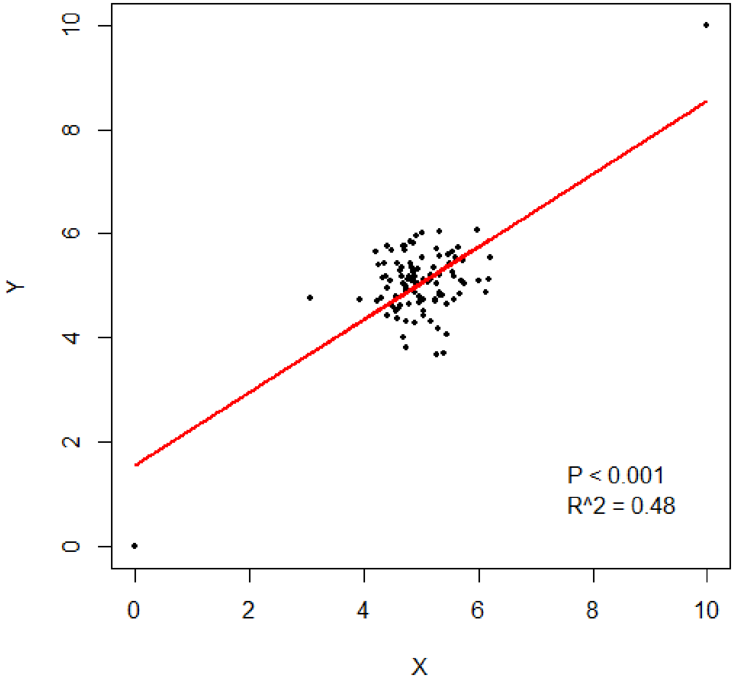
What we can derive from the concept

- **Does distribution of X really affects the discrepancy between significance and predictivity?**

What we can derive from the concept

- **An Extreme Case** $y = x + \varepsilon, \quad \varepsilon \sim N\ (0,1)$





What we can derive from the concept

- **Significance**

$$E(y\ |Xi) \sim N(X_i\beta, \sigma^2)$$

- **Prediction**

X	0	1	2
MAF=p	$(1-p)^2$	$2p(1-p)$	p^2
$EX = 2p$		$\mathrm{Var}(X) = 2p(1-p)$	

$$\begin{aligned} Var(\hat{y}) &= Var(X\beta + \varepsilon) \\ &= \beta^2 Var(X) \\ &= \underbrace{2p(1-p)}_{\text{distrn of x}}\underbrace{\beta^2}_{\text{effects}} \end{aligned}$$

Example 2

- **1 SNP**

  - Outcome variable is the case or control
  - Explanatory variable is the reading on 1 SNP for each of 500 cases and 500 controls
  - Significance level using X^2 test
  - Prediction rate by calculating the proportion of correct classification among the 1000 individuals

Example 2

- **1 SNP**

Predictive variable sets

genotype	0	1	2
Cases	.274	.500	.226
Controls	.226	.500	.274

Pred. rate =**0.524** Median P-value=6x10^{-2}

Median I score = **1.544**

1 SNP

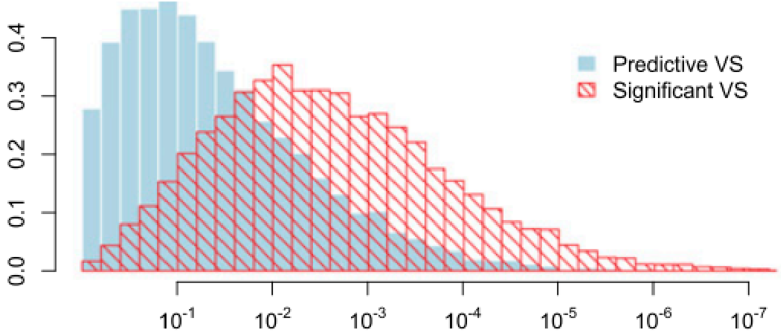
Significant variable sets

genotype	0	1	2
Cases	0.01	0.19	0.80
Controls	0.04	0.16	0.80

Pred. rate =0.515 **Median P-value=3.5x10^{-3}**

Median I score = 0.632

Distribution of P-values from Chi-squared test. Plotted on –log$_{10}$ scale.



Example 2

Example 2

- **2 SNPs**

Predictive variable sets

genotype	00	01	02	10	11	12	20	21	22
Cases	.1	.125	.1	.125	.1	.125	.1	.125	.1
Controls	.125	.1	.125	.1	.125	.1	.125	.1	.125

Pred. rate = **0.56** Median P-value=$3\text{x}10^{-2}$

2 SNPs

⇔

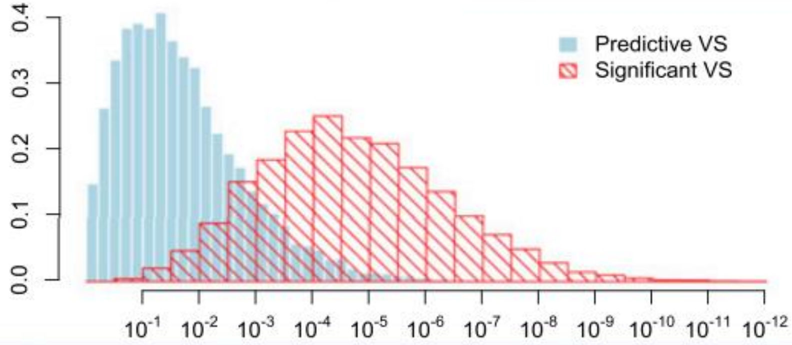
Significant variable sets

genotype	00	01	02	10	11	12	20	21	22
Cases	.01	.052	.134	.134	.134	.134	.134	.134	.134
Controls	.052	.01	.134	.134	.134	.134	.134	.134	.134

Pred. rate = 0.52 **Median P-value=$2.8\text{x}10^{-5}$**

Median I score = **1.854**

Distribution of P-values from Chi-squared test. Plotted on $-\log_{10}$ scale.



Median I score = 1.676

Example 3

Example 2

- **Conclusion**

  - There is overlap between two variables set which means that to some extent significant variable serves good prediction ability and those highly predictive variable are also tend to be significant. However that is NOT necessary;
  - In fact, large proportion of the predictive variables are not significant, vice versa.
  - As the number of SNPs increase (complexity increase), the overlap become smaller and smaller
  - But all results above based on their way of calculating the so-called “prediction rate” as measurement of prediction ability, which is a shortcoming of the paper.

I score

- I score

$$I = \sum_{j=1}^{m1} \frac{n_j}{n} \frac{\left(\bar{Y}_j - \bar{Y}\right)^2}{s^2/n_j} = \frac{\sum_{j=1}^{m_1} n_j^2 \left(\bar{Y}_j - \bar{Y}\right)^2}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}$$

$Y_i: Y\ value\ of\ ith\ individual$

$Y_j: mean\ values\ of\ Y\ in\ cell\ j$

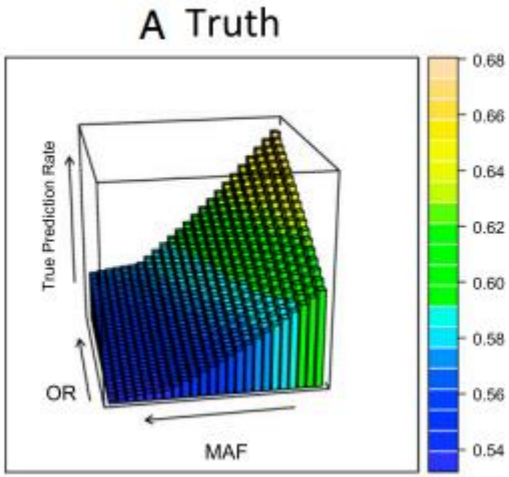
$\bar{Y}:\text{mean of all Ys}$

$m_1: number\ of\ \ cells$

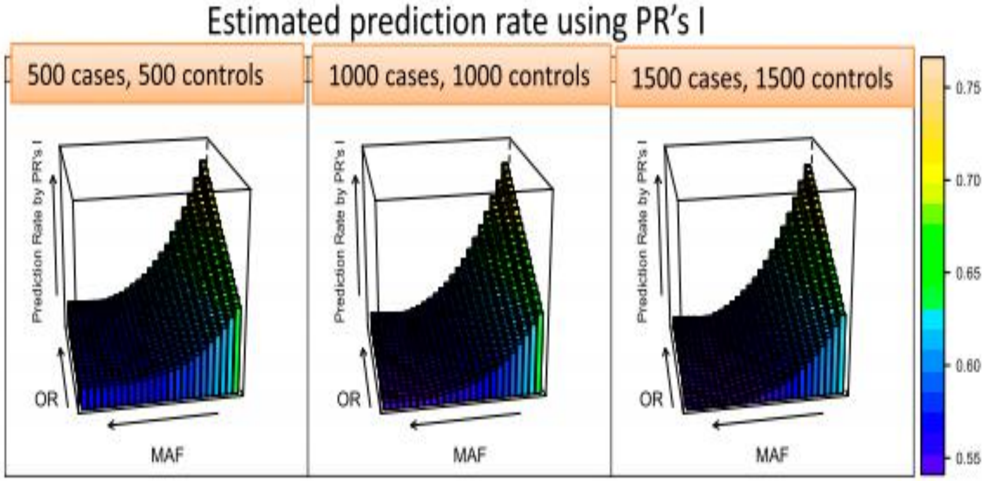
- Weighted average of X^2s
- Proportion of between-group variance divide by the total variance
- The bigger the I score, the better prediction ability
- Applied to small groups of

Example 3

- 6 SNPs



A Truth



Estimated prediction rate using PR's I

500 cases, 500 controls

1000 cases, 1000 controls

1500 cases, 1500 controls

Conclusion and Comments

- Give us a sight into how the distribution of a variable will effect the significance and prediction ability.

- **Based on their way of measuring significance and prediction ability, which is not common(consistent) and powerful in real study**

- Establish a I score which to some extent useful in identifying whether a variable is predictive regardless of significance

- **Can only be used to small group of variables thus limited to be applied to general big genetic data when number of SNPs is huge**