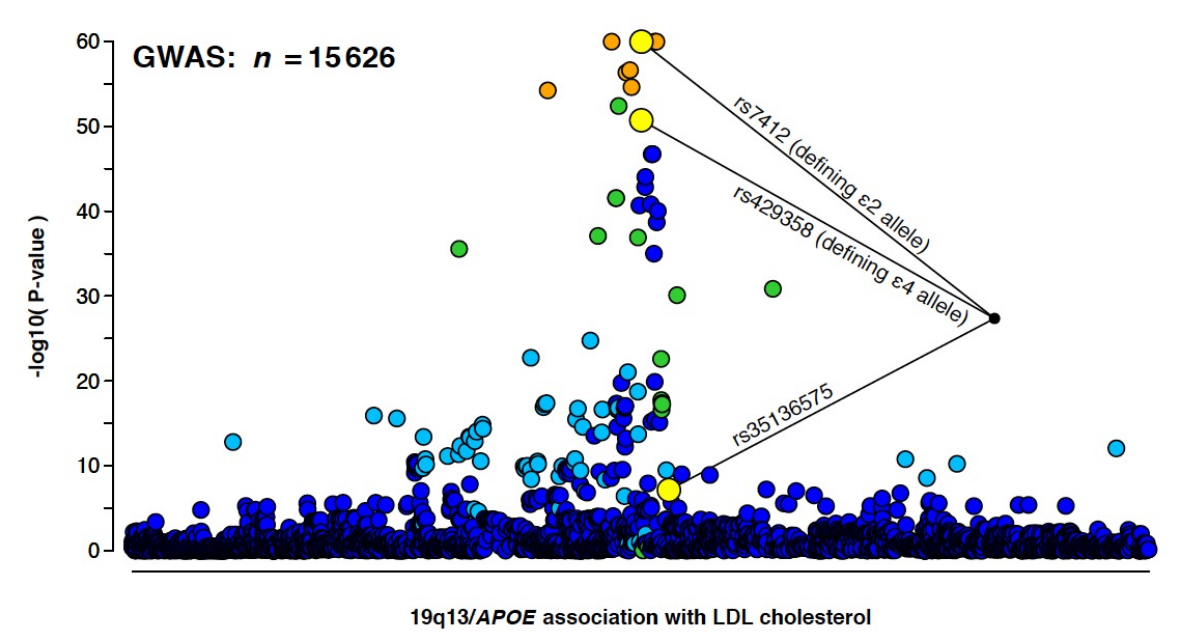
# Example: Variable selection in genetics




- Variants near each other are often highly correlated
  - |r|> 0.90 very common
- Which ones are causal and which are just passengers?

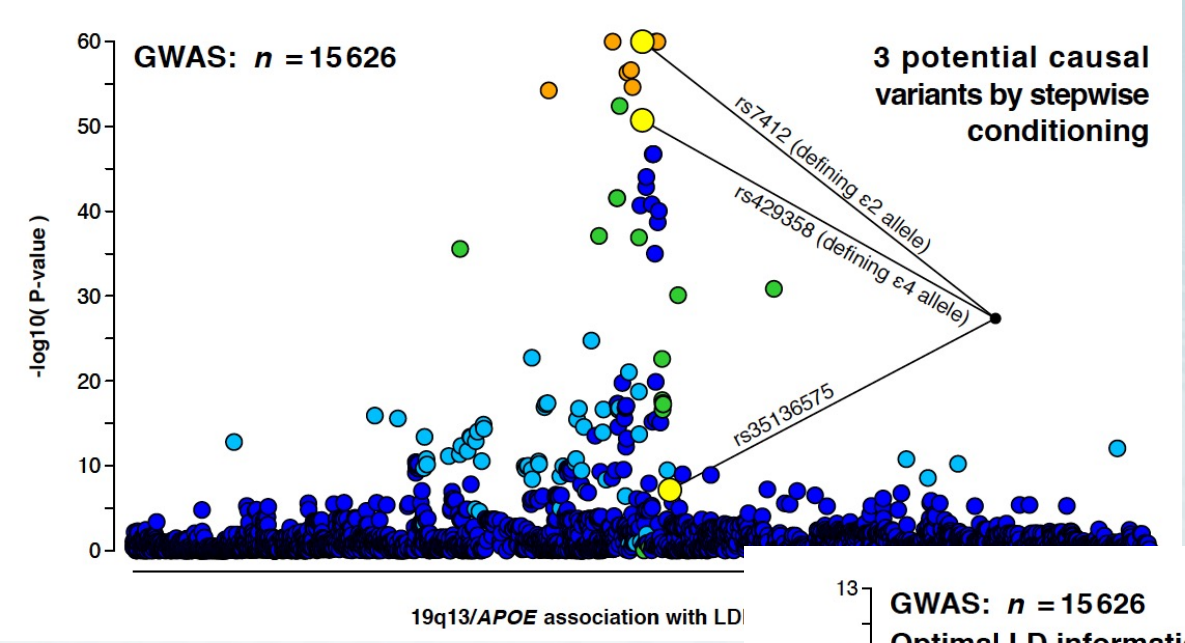
$n \sim 10^5$ (samples = individuals)

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| IND1 | M | A/A | A/C | A/C | C/C | A/A | A/C | A/C | C/C | C/C |
| IND2 | M | A/A | A/C | A/A | C/C | A/C | A/C | A/C | C/C | A/C |
| IND3 | M | A/C | C/C | A/C | C/C | A/A | C/C | C/C | C/C | C/C |
| IND4 | F | A/A | A/C | A/A | C/C | A/C | A/C | A/C | C/C | C/C |
| IND5 | F | A/A | C/C | A/C | C/C | A/A | A/A | C/C | A/A | A/C |
| IND6 | M | A/C | A/C | C/C | C/C | A/A | A/C | A/C | C/C | C/C |
| IND7 | F | A/A | A/A | A/C | C/C | A/A | C/C | A/C | C/C | A/C |
| IND8 | M | A/C | A/C | C/C | C/C | C/C | A/C | C/C | C/C | C/C |
| IND9 | F | A/A | A/A | A/C | C/C | A/A | A/C | C/C | A/C | C/C |

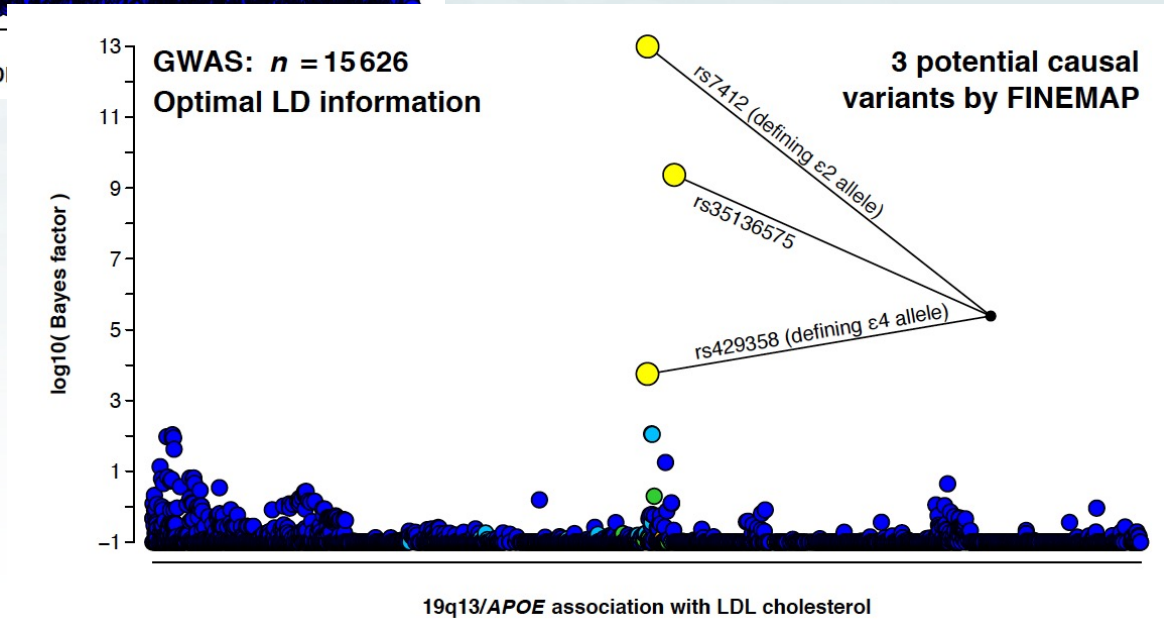
$p = 10^3 \ldots 10^4$ (predictors = variants)

C.Benner

# Variable selection = “fine-mapping”

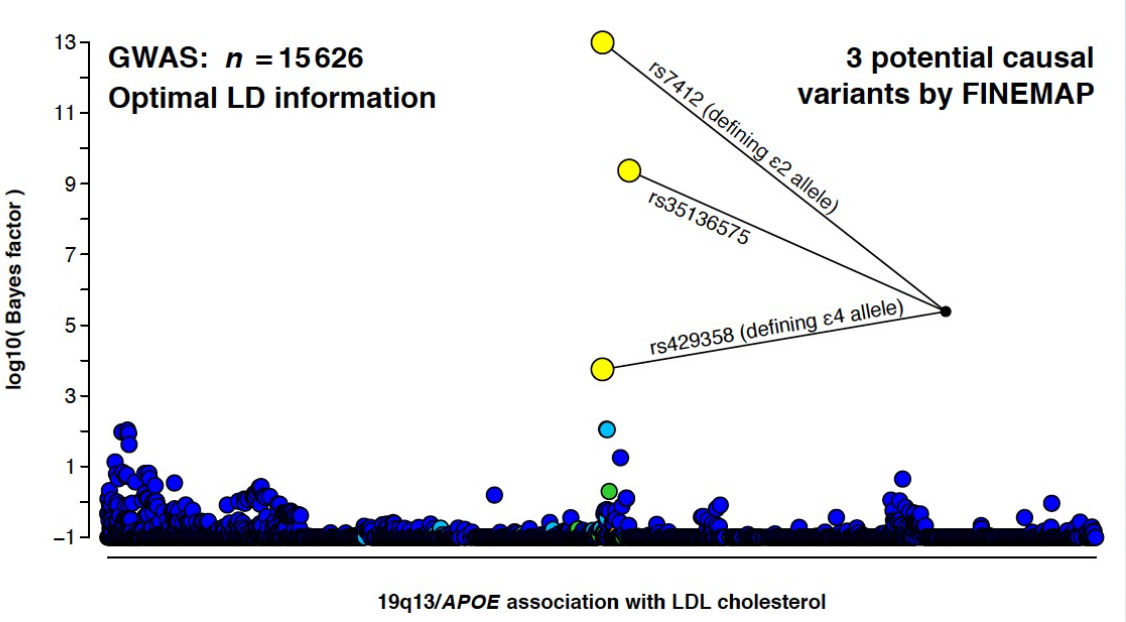



- Variants near each other are often highly correlated
- Which ones are causal and which are just passengers?




C.Benner

# Final results are probabilities




- What is the probability for each configuration of variants being causal?
- What is the probability for each variant being causal?

| rank | config | config_prob | config_log10bf |
|---|---|---|---|
| 1 | rs15,rs47 | 0.59 | 44.6 |
| 2 | rs15,rs42,rs47 | 0.02 | 44.9 |
| 3 | rs15,rs34,rs47 | 0.01 | 44.7 |

| index | snp | snp_prob | snp_log10bf |
|---|---|---|---|
| 15 | rs15 | 1.00 | 11.3 |
| 47 | rs47 | 1.00 | 10.6 |
| 42 | rs42 | 0.03 | -0.22 |

C.Benner

# Goal of probabilistic variable selection

To provide, for each predictor,

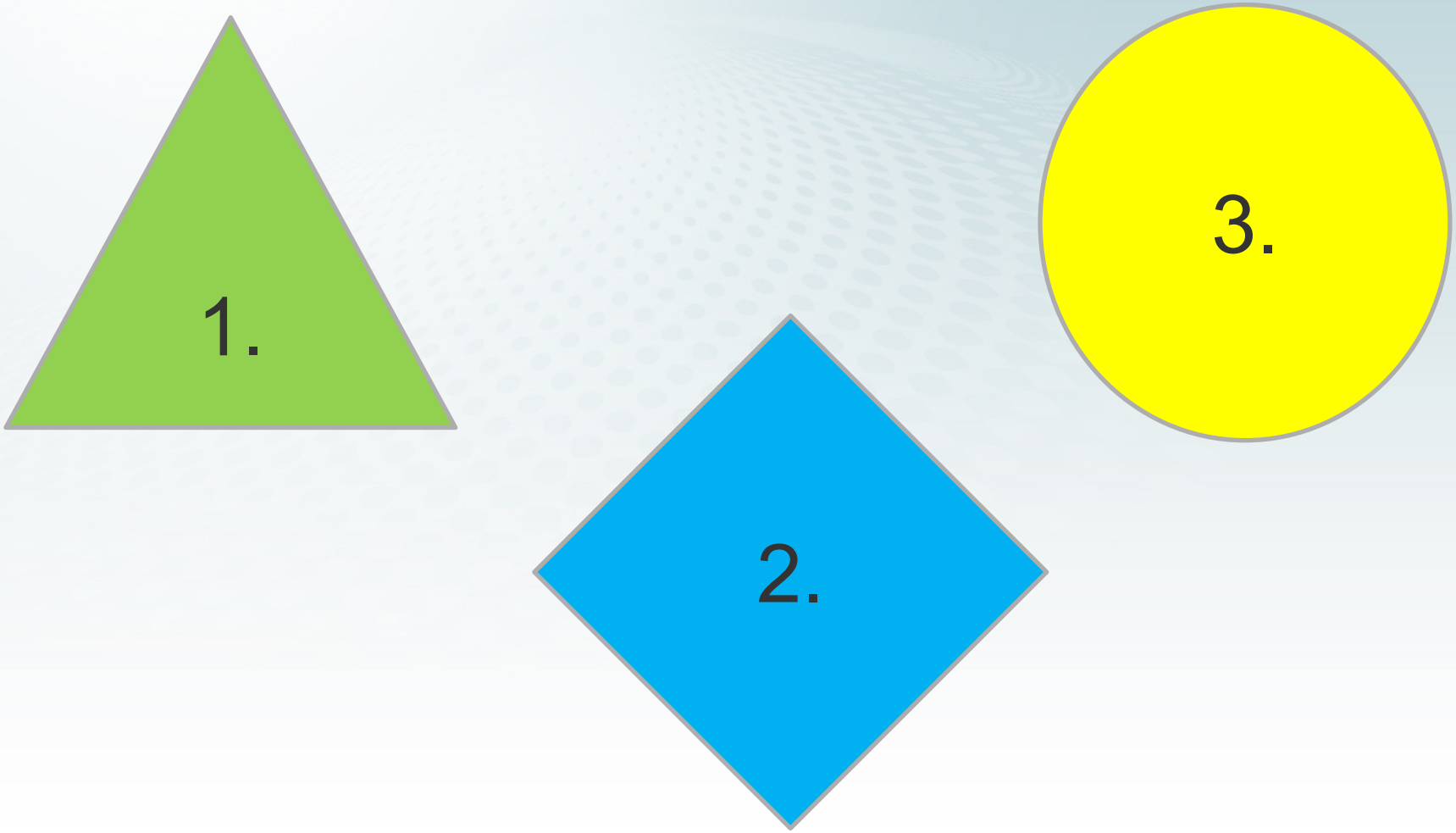
(1) a measure of being important ("causal")

- We want variable selection **with estimates of uncertainty** that elastic net does not give (by default)

(2) by accounting for correlation between the predictors

- Elastic net gives only an optimum but does not output other possible solutions that could include some of the highly correlated predictors instead of the chosen ones. We want a longer list of most probable configurations.

# Three pieces of efficient variable selection

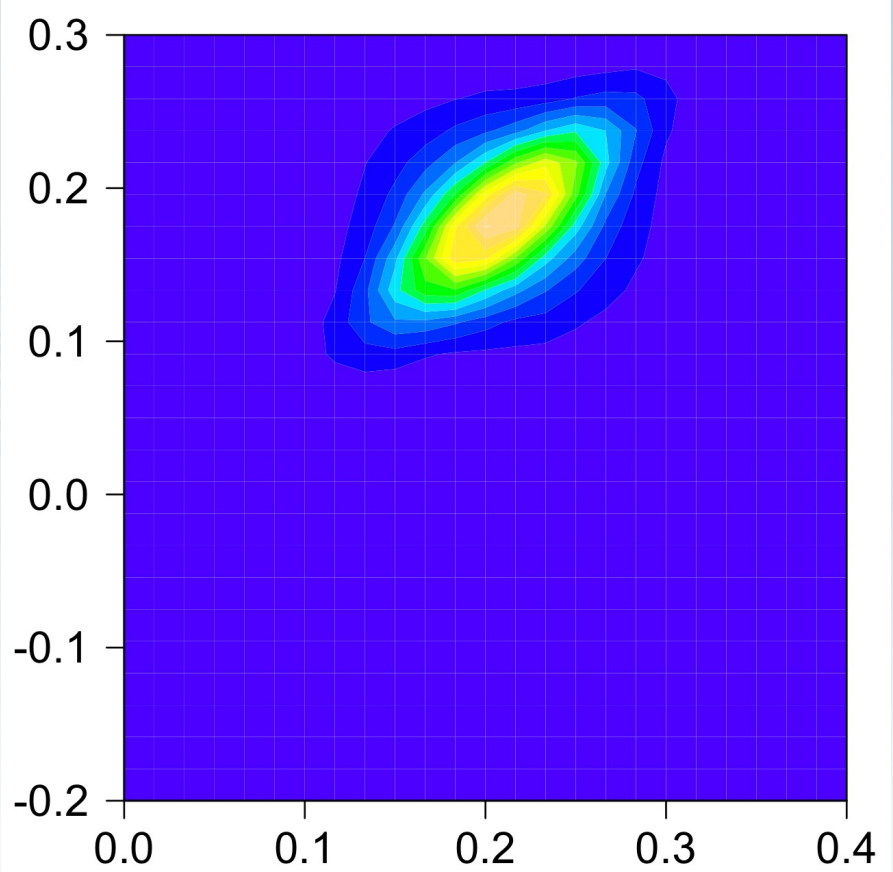


# Univariate “betas”

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times 1}{\mathbf{X}_\ell}\ \underset{1\times 1}{\beta_\ell} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

$$\widehat{\beta_\ell} = (\mathbf{X}_\ell^T \mathbf{X}_\ell)^{-1} \mathbf{X}_\ell{}^T \mathbf{y}$$

These are simple regression coeffs between one predictor and outcome. Computed efficiently as in Exercise 1.5.



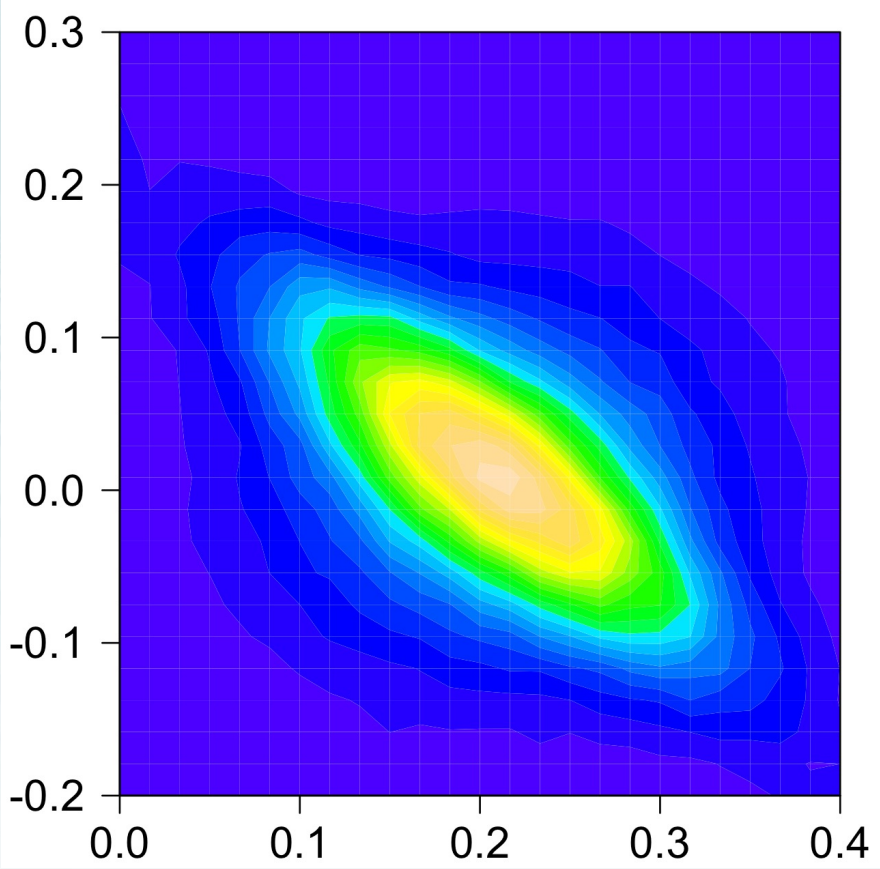

- Cor($X_1$, $X_2$) = 0.85
- $X_1$ has effect 0.2, $X_2$ is null (=0)

# Multiple regression "lambdas"

$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times 1}{\boldsymbol{\lambda}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

$$\hat{\boldsymbol{\lambda}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

These are the direct effects and account for other predictors but are complicated to compute for large p, especially when p > n.




- Cor($X_1$, $X_2$) = 0.85
- $X_1$ has effect 0.2, $X_2$ is null (=0)

# Betas and lambdas

Assuming standardized predictors

$$\beta = \frac{1}{n}(\mathbf{X}^T\mathbf{X})\lambda = \mathbf{R}\lambda$$

where **R** is pairwise correlation matrix of predictors.

$$\beta = \begin{bmatrix} 1 & 0.85 \\ 0.85 & 1 \end{bmatrix} \begin{bmatrix} 0.2 \\ 0 \end{bmatrix} = \begin{bmatrix} 0.2 \\ 0.17 \end{bmatrix}$$

# Summary data

› Computation for multiple regression model is possible using **summary data**: univariate z-scores and correlation matrix of predictors (R matrix)

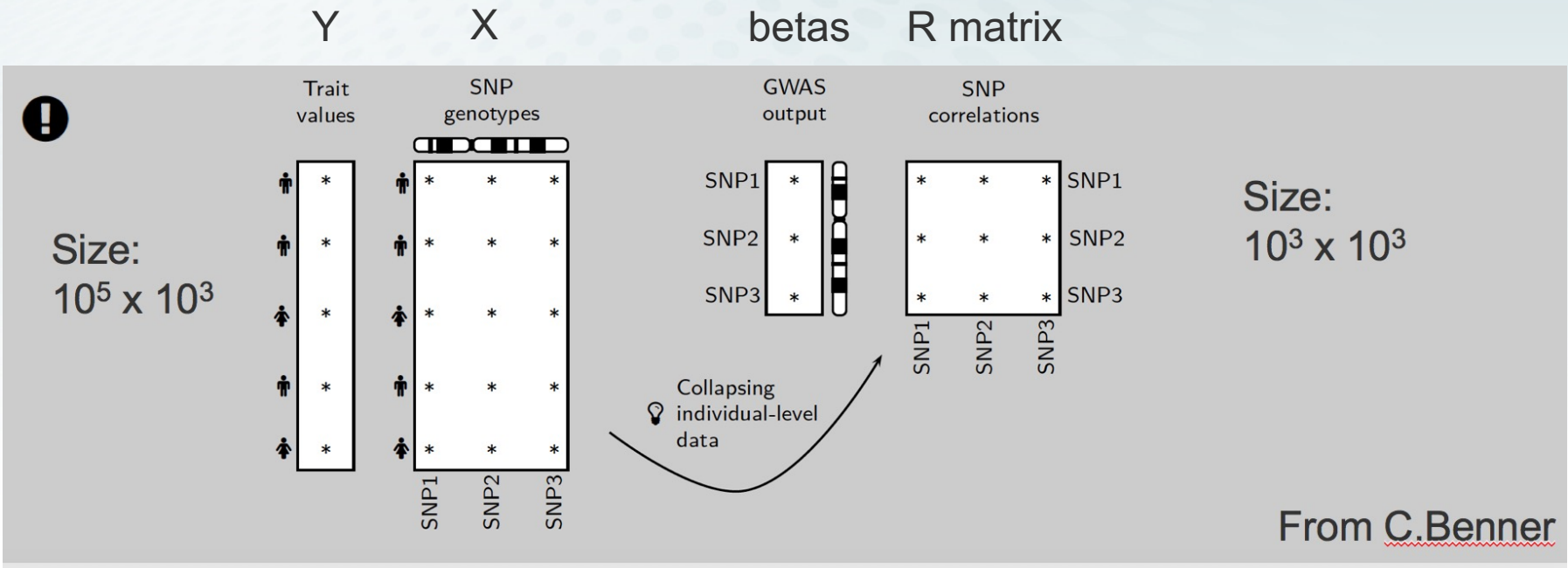
$$\underset{n\times 1}{\mathbf{Y}} = \underset{n\times p}{\mathbf{X}}\ \underset{p\times 1}{\boldsymbol{\lambda}} + \underset{n\times 1}{\boldsymbol{\varepsilon}}$$

$$\hat{\boldsymbol{\lambda}} = \underbrace{(\mathbf{X}^T\mathbf{X})^{-1}}_{n\mathbf{R}}\ \underbrace{\mathbf{X}^T\mathbf{Y}}_{\sqrt{n}\sigma_\varepsilon \mathbf{z}}$$

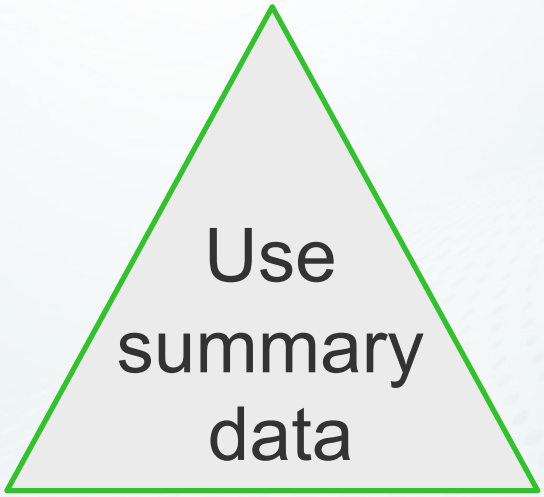
**R,** predictors' correlations
**Z,** univariate z-scores

# Use summary data to make multiple regression possible from univariate results

- Working with less data but with full information
  - For p=1,000 and n=100,000, data reduction is **100 fold**

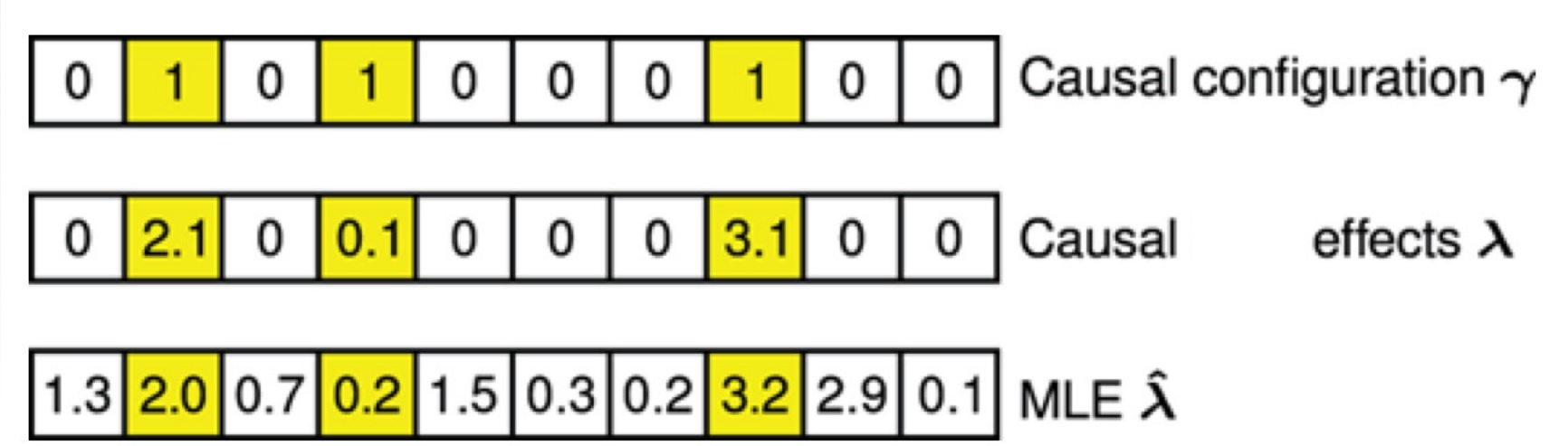


# Three pieces of efficient variable selection




# Assumption: true configuration is sparse

- Joint MLE (or ridge regression) of all predictors is not our final answer to variable selection since it does not lead to sparse solutions
- Bayesian answer:
  - Define a prior probabilities for configurations
  - Define a prior distribution for regression coefficients of a configuration
  - Integrate (prior x likelihood) leading to *marginal likelihood* for the configuration




"causal effects" = "direct effects" = "multiple regression coefficients"

# Bayesian model for variable selection

› Define a configuration $\gamma$ as a binary vector over predictors

$$\gamma = \begin{array}{|c|c|c|c|c|c|c|c|} \hline 0 & 0 & 1 & 0 & 0 & 0 & 1 & 0 \\ \hline \end{array}$$

› This configuration represents model where predictors 3 and 7 are allowed to have non-zero effects while the other predictors have effect size 0

# Bayesian model for variable selection

› Define a configuration $\gamma$ as a binary vector for predictors


› In total there are $2^p$ configurations on p predictors, but we will assume that only **sparse** configurations are plausible, say those with < 10 non-zero predictors

- This is similar idea to LASSO that sets many coefficients to 0

› Ultimate goal is to compute probability for each configuration, given the observed data

- This is much more challenging than the LASSO optimization

# Bayesian model for variable selection

› Define a configuration $\gamma$ as a binary vector for predictors

› Each non-zero predictor picks its effect from N(0, $s^2$)
- This is the slab part of the spike and slab prior
- This is similar prior as in ridge regression but now the model is sparse which is different from ridge regression

$$p(\boldsymbol{\lambda}|\boldsymbol{\gamma}) = \mathcal{N}\left(\mathbf{0}, s^2\boldsymbol{\Delta}_{\boldsymbol{\gamma}}\right)$$

Causal configuration $\gamma$

| 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
|---|---|---|---|---|---|---|---|---|---|

$$\boldsymbol{\Delta}_{\boldsymbol{\gamma}} = \text{diag}(\boldsymbol{\gamma}) = \begin{bmatrix} 1 & & & \\ & 0 & & \\ & & \ddots & \\ & & & 0 \end{bmatrix}$$

# Bayesian model for variable selection

› Define causal configuration $\gamma$ as a binary vector for predictors

› Each non-zero predictor picks its effect from $N(0, s^2)$

› For each configuration, compute the Bayes factor (BF), i.e., how well the configuration explains the data relative to the null model

$$\mathrm{BF}_\gamma = \frac{P(\mathrm{DATA}|\gamma)}{P(\mathrm{DATA}|\mathrm{NULL})}$$

› How to compute the numerator?

# Marginal likelihood for a configuration

$$\mathcal{L}(\gamma) = \int p(\boldsymbol{y}|\boldsymbol{\lambda}, \mathbf{X})\ p(\boldsymbol{\lambda}|\gamma)\, d\boldsymbol{\lambda}$$

(Likelihood x prior)

$$= \int \mathcal{N}\left(\widehat{\boldsymbol{\lambda}}|\boldsymbol{\lambda}, \sigma^2(\mathbf{X}^T\mathbf{X})^{-1}\right) \mathcal{N}\left(\boldsymbol{\lambda}|0, s_\lambda^2\sigma^2\boldsymbol{\Delta}_\gamma\right) d\boldsymbol{\lambda}$$

$$= \mathcal{N}\left(\widehat{\boldsymbol{\lambda}}|0, \sigma^2(n\mathbf{R})^{-1} + s_\lambda^2\sigma^2\boldsymbol{\Delta}_\gamma\right)$$

$$= \mathcal{N}\left(\widehat{\boldsymbol{z}}|0, \mathbf{R} + \mathbf{R}\boldsymbol{\Delta}_\gamma^*\mathbf{R}\right)$$

Depends on X and Y only through Summary data Z and R!

$$\widehat{\boldsymbol{z}} = \widehat{\boldsymbol{\beta}}/\mathrm{SE}_\beta = \frac{\sqrt{n}}{\sigma}\widehat{\boldsymbol{\beta}} \text{ and } \boldsymbol{\Delta}_\gamma^* = s_\lambda^2\boldsymbol{\Delta}_\gamma$$

# Marginal likelihood for a configuration

$$\mathcal{L}(\gamma) = \mathcal{N}\left(\hat{\boldsymbol{z}} | 0, \mathbf{R} + \mathbf{R}\boldsymbol{\Delta}^{*}_{\gamma}\mathbf{R}\right)$$

+ Depends on data only through univariate summary statistics and correlation matrix R, i.e., **summary statistics**
  Thus, we do not need access to original X and Y!
- Dimension is p, the number of predictors that can be 10,000s, which makes evaluation of many configurations impossible since each config requires decomposition of a pxp matrix and this is $O(p^3)$

# Using only causal predictors

Consider configuration $\gamma$
Divide predictors into causal ( C ) and non-causal ( N )

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_C \\ \boldsymbol{z}_N \end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix} \mathbf{R}_{CC} & \mathbf{R}_{CN} \\ \mathbf{R}_{NC} & \mathbf{R}_{NN} \end{bmatrix}$$

# Using only causal variants

Consider configuration $\gamma$
Divide predictors into causal ( C ) and non-causal ( N )

$$\boldsymbol{z} = \begin{bmatrix} \boldsymbol{z}_C \\ \boldsymbol{z}_N \end{bmatrix} \qquad \mathbf{R} = \begin{bmatrix} \mathbf{R}_{CC} & \mathbf{R}_{CN} \\ \mathbf{R}_{NC} & \mathbf{R}_{NN} \end{bmatrix}$$

Cond. distr of $\mathbf{Z_N} \mid \mathbf{Z_C}$ is the same for configuration $\gamma$ as it is for null model !

$$\begin{aligned} \mathrm{BF}(\gamma : \mathrm{NULL}) &= \frac{\mathcal{N}(\boldsymbol{z}|0,\, \mathbf{R} + \mathbf{R}\,\boldsymbol{\Delta}^*_\gamma\,\mathbf{R})}{\mathcal{N}(\boldsymbol{z}|0,\, \mathbf{R})} \\ &= \frac{\mathcal{N}(\boldsymbol{z}_C|0,\, \mathbf{R}_{CC} + \mathbf{R}_{CC}\,\boldsymbol{\Delta}^*_{\mathbf{CC}}\,\mathbf{R}_{CC})\,\mathcal{N}(\boldsymbol{z}_N|\boldsymbol{z}_C)}{\mathcal{N}(\boldsymbol{z}_C|0,\, \mathbf{R}_{CC})\,\mathcal{N}(\boldsymbol{z}_N|\boldsymbol{z}_C)} \\ &= \frac{\mathcal{N}(\boldsymbol{z}_C|0,\, \mathbf{R}_{CC} + \mathbf{R}_{CC}\,\boldsymbol{\Delta}^*_{\mathbf{CC}}\,\mathbf{R}_{CC})}{\mathcal{N}(\boldsymbol{z}_C|0,\, \mathbf{R}_{CC})} \end{aligned}$$

Benner et al. 2016


This derivation holds for small effects, general case in Benner et al. 2018 www.biorxiv.org/content/10.1101/318618v1.full.pdf

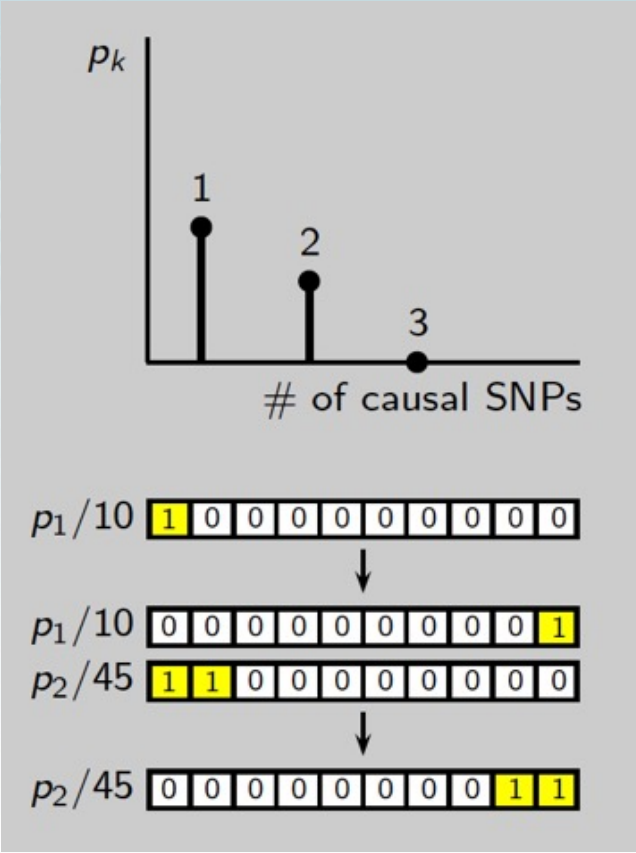
# Bayesian model for variable selection

- Define causal configuration $\gamma$ as a binary vector for predictors
- Each non-zero predictor picks its effect from $N(0, s^2)$
- For each configuration compute the Bayes factor (BF), i.e., how well the configuration explains the data relative to the null model
- By combining BFs with prior probabilities of configurations we get the posterior probabilities

$$p_\gamma = P(\gamma|\mathrm{DATA}) \propto \mathrm{prior}_\gamma \times \mathrm{BF}_\gamma$$

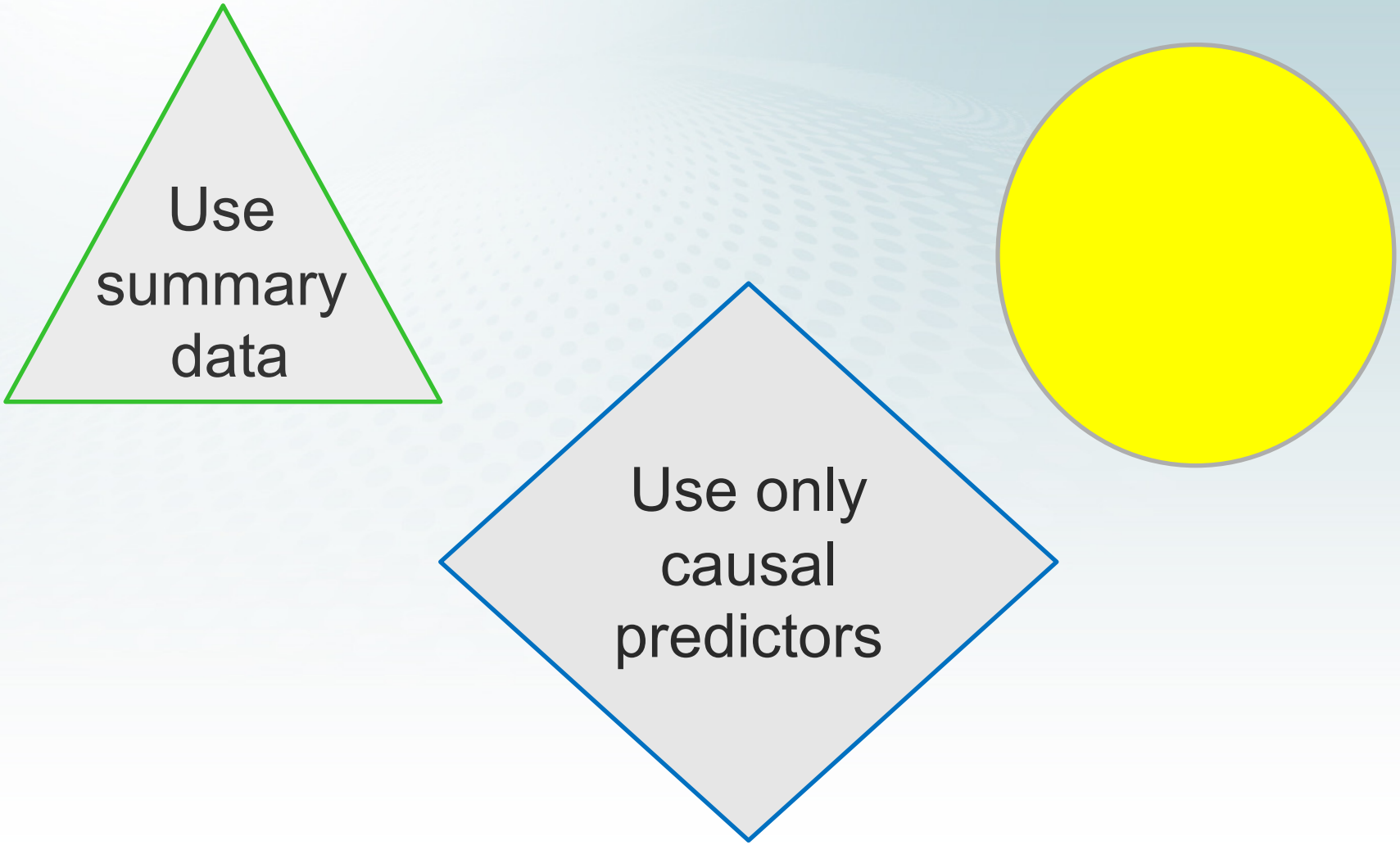
# Prior on configurations to enforce sparsity

- Specify probability $p_k$ that there are k non-zero predictors
- Divide that probability equally between all configurations having k non-zero predictors
- This prior could be learned from data or remain as an ad-hoc choice




C. Benner

# Three pieces of efficient fine-mapping

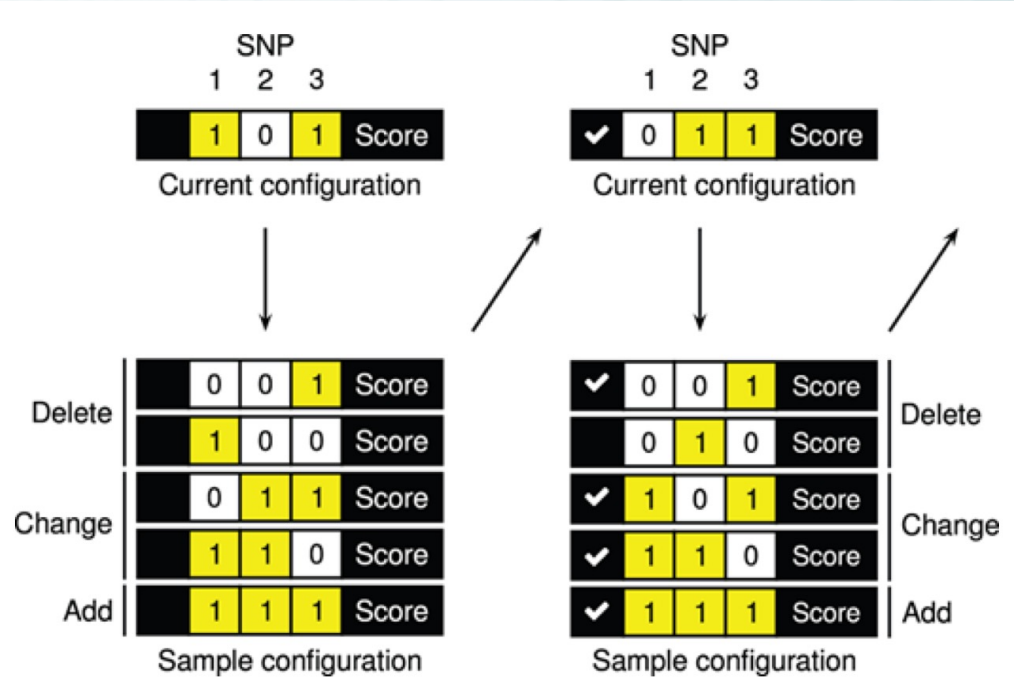


# Space of causal configurations is huge, $2^p$

- Best subset algorithms evaluate every configuration
  - Can allow at most 3 causal predictors when 1000s of predictors are available
  - Experimenting with genetic data: On average only about 100 configs out of 70,000,000 already covered 95% of posterior in setting: p=750 , 5 causal predictors (Benner et al. 2016)
    - Can be different in some other application fields!

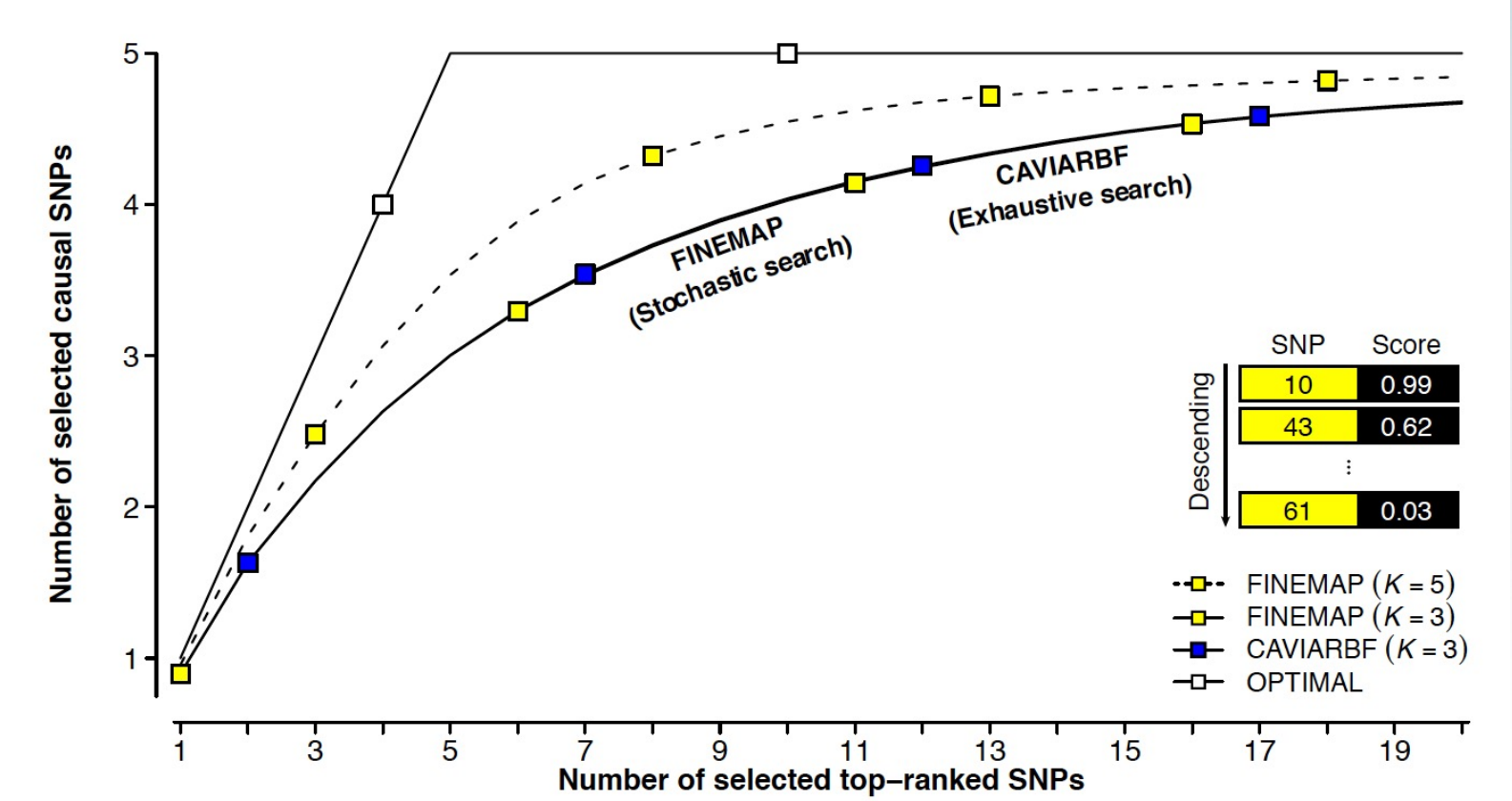
# Shotgun stochastic search algorithm

- Collect configurations from a high probability region using Shotgun stochastic search (Hans et al. 2007)
  - Memorize BFs of all those configurations seen during the search
  - Stop once not much new probability mass is found
  - Renormalize posteriors with respect to the configurations visited



# Example: FINEMAP software

- Simulations with p=1500 of which 5 are truly non-zero
  - FINEMAP runs in a few seconds
  - Enumeration is impossible in practice



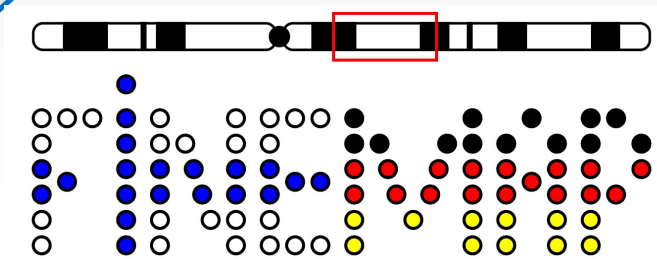

Benner et al. 2016

# Three pieces of efficient variable selection

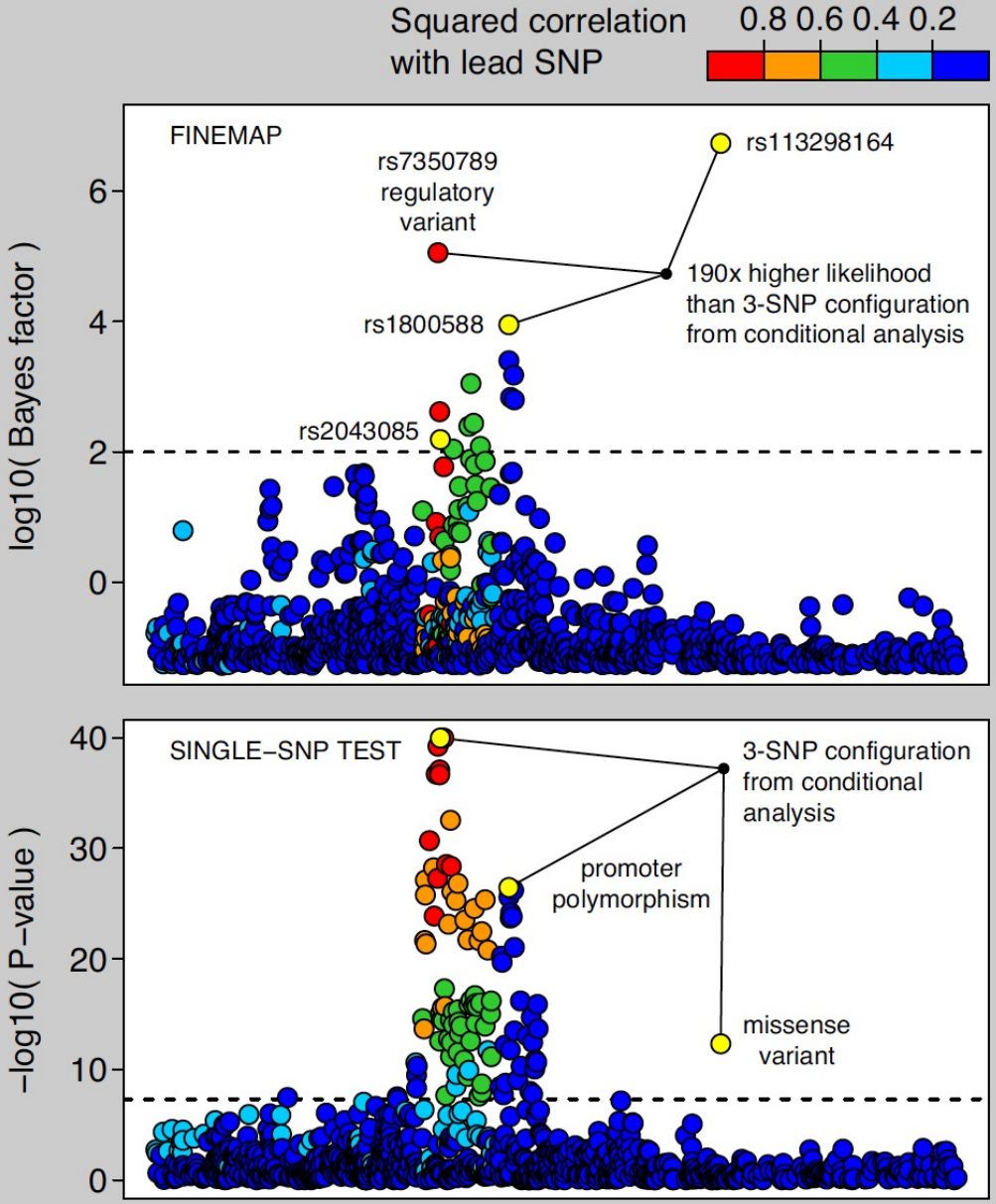
Use summary data

Use only causal predictors

Use efficient search algorithm



# 15q21/*LIPC* association with HDL cholesterol




6 Mb region
8612 variants

Benner et al. 2016

Surakka et al.
Nat. Genet. 2015

NATIONAL INSTITUTE FOR HEALTH AND WELFARE
FINLAND

FINRISK STUDY
20000 indivduals

# Acknowledgements

Christian Benner ( www.finemap.me )

FIMM

ACADEMY OF FINLAND

NATIONAL INSTITUTE
FOR HEALTH AND WELFARE