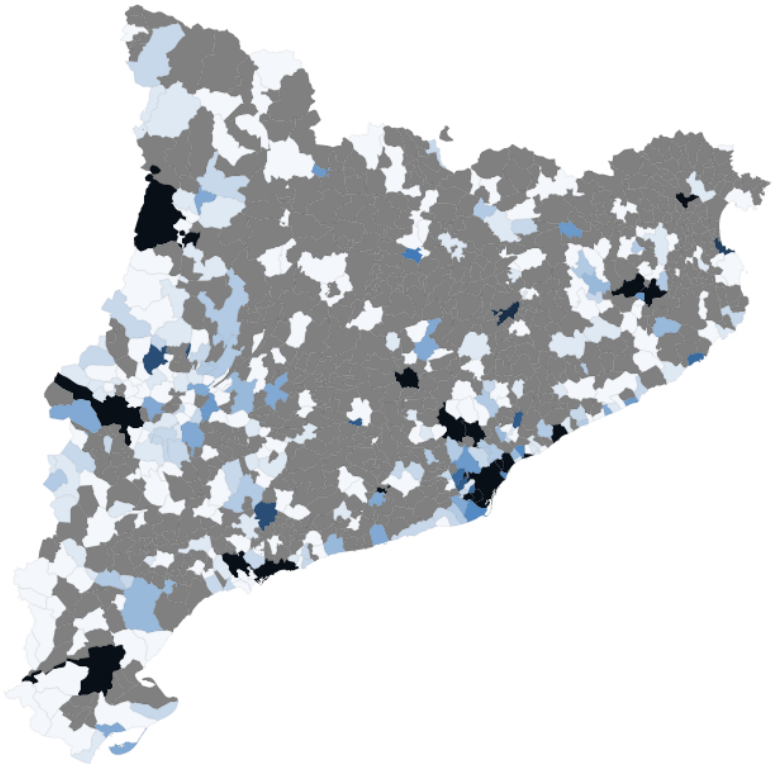
Random forests and phenome-wide association studies

Xavier Duran
GCAT Genomes for Life

BioinfoTalks
April 27th, 2016

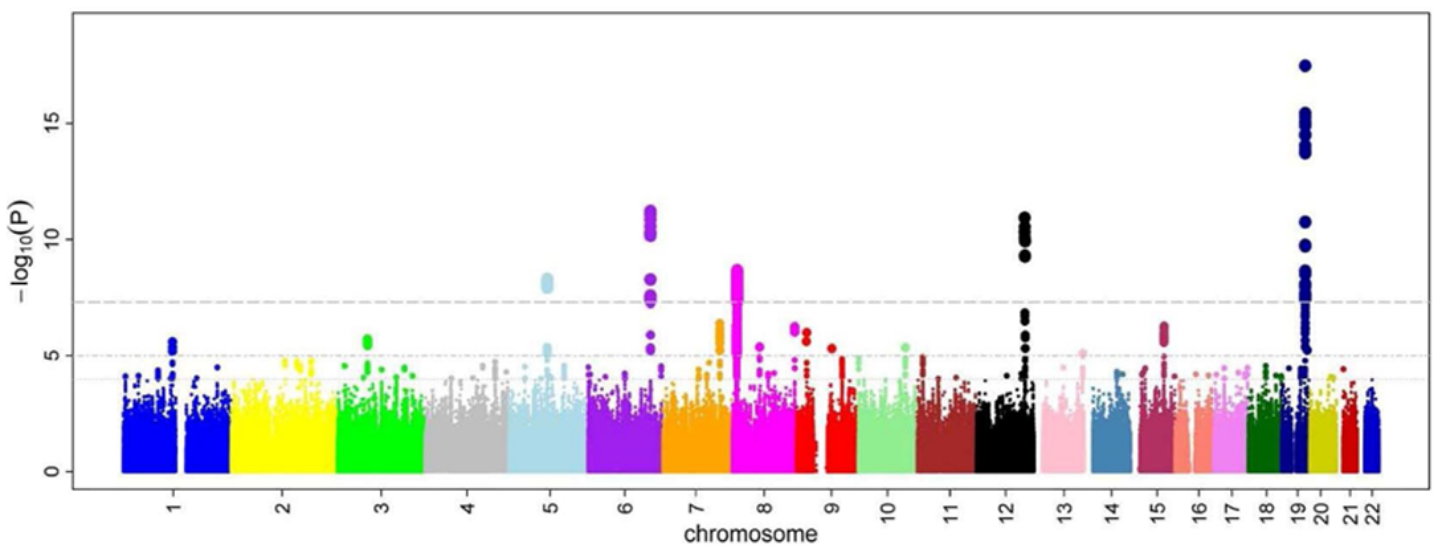
GCAT genotyping



- Longitudinal cohort study
- More than 10.000 participants
- Genotyping first 5.000 participants
- 1.2M SNPs
- Up to 7 to 10M SNPs after imputation

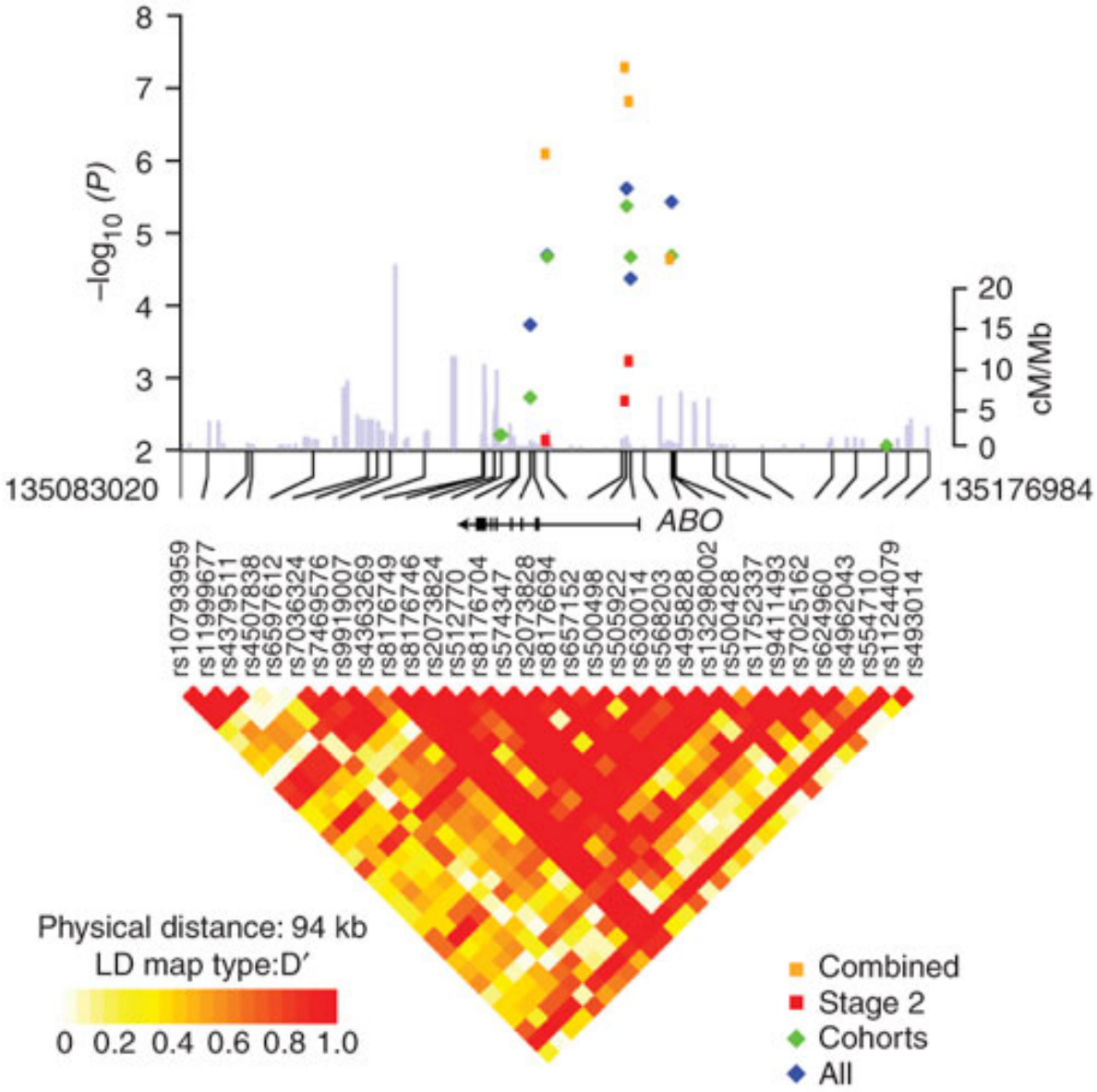
Genome-wide association studies (GWAS)

- Case-control approach
- Identify genetic variants linked to disease risk or a trait
- Test genotype frequency
- Regression modelling (linear, logistic) + covariates (sex, age, ethnicity)
- Multiple comparison, assumption of independence
- Bonferroni correction
- Threshold significance 5×10^{-8}



Genome-wide association studies (GWAS)

Candidate genes



Genome-wide association studies (GWAS)

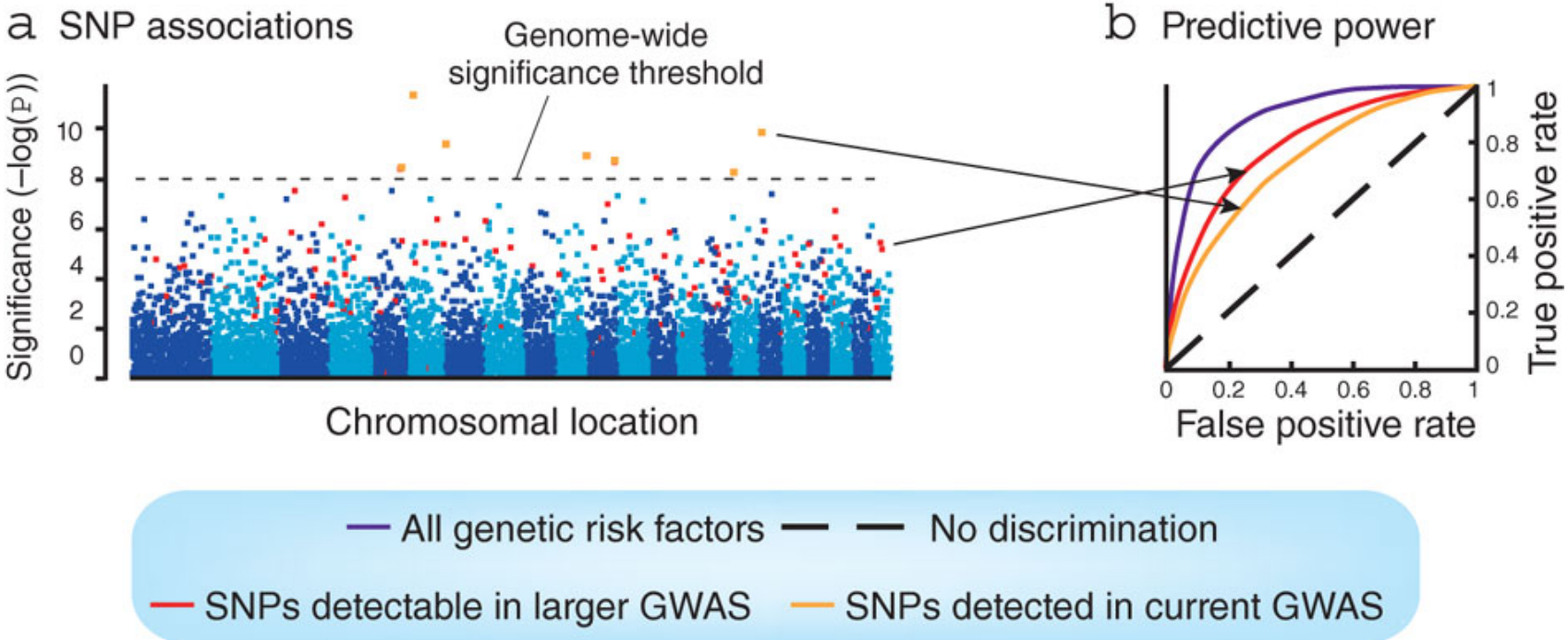
Success and limitations

- GWAS Catalog

Genome-wide association studies (GWAS)

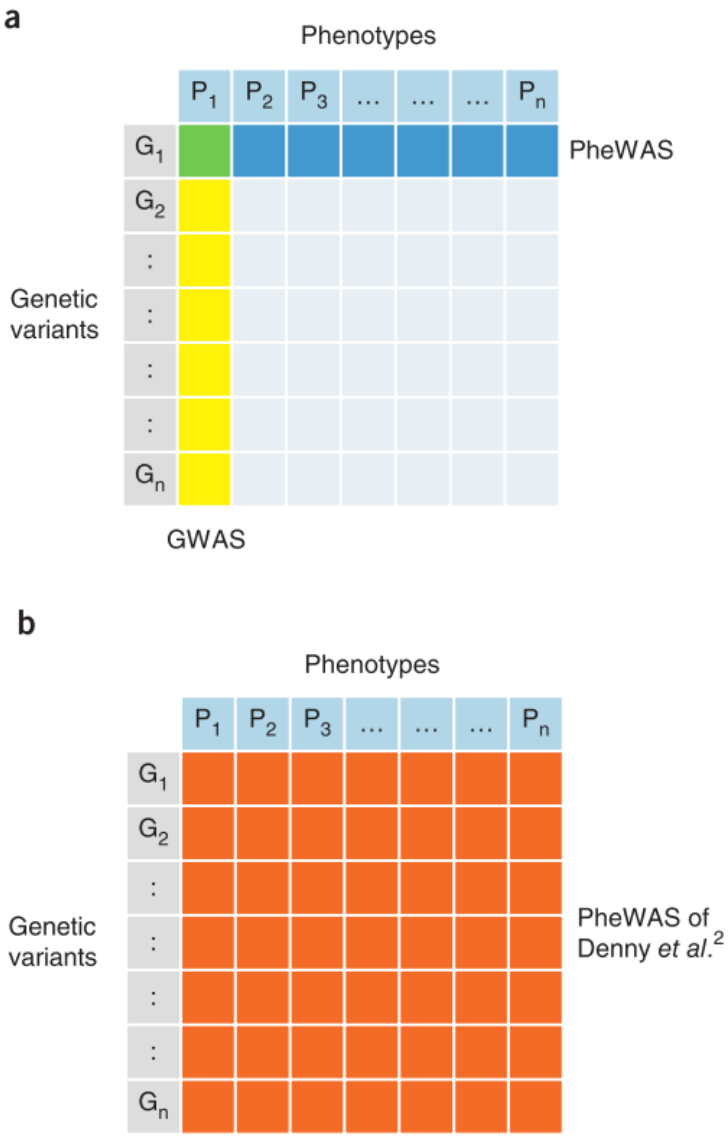
Success and limitations

- Single SNP association studies explain a small part of disease heritability
- The success depends on both biological and statistical reasons



Phenome-wide association studies (PheWAS)

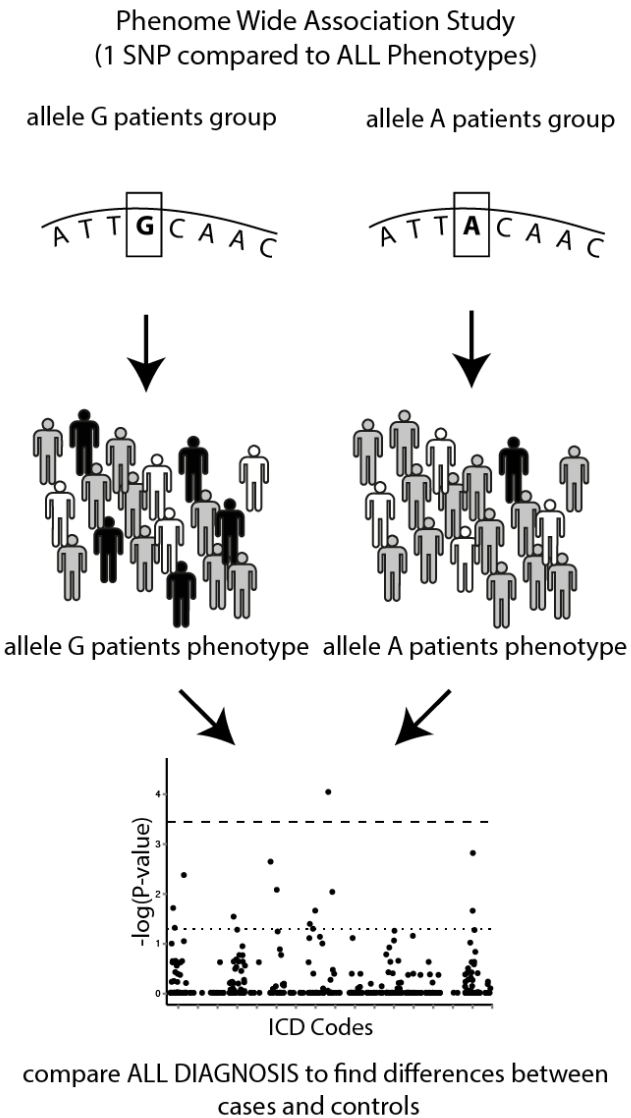
An alternative approach



Scan all the phenotypes of all this patients to find systematic associations between this mutation and all the phenotypes.

Phenome-wide association studies (PheWAS)

Study design



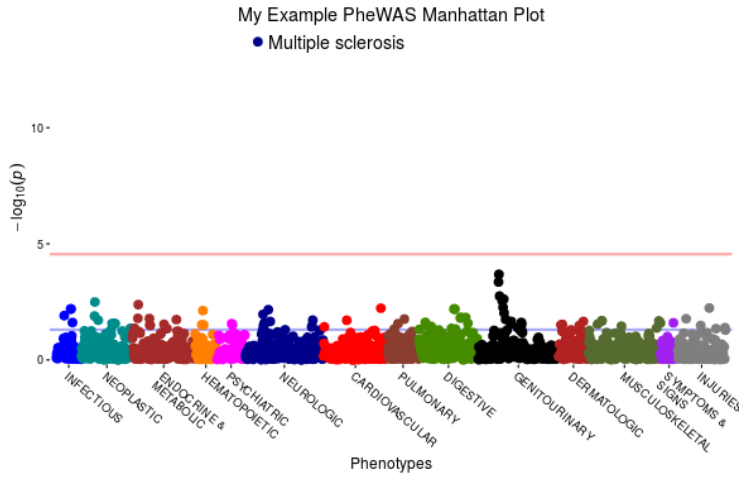
- Hypothesis-free: only assumes a relationship
- Mendelian Randomization
- Direction of inference, from exposure to outcome
- Systematic examination of variants of special interest
- Environmental exposures
- Unknown comorbidities
- Adjustment for multiple testing (Bonferroni, false discovery rate)

Phenome-wide association studies (PheWAS)

R package

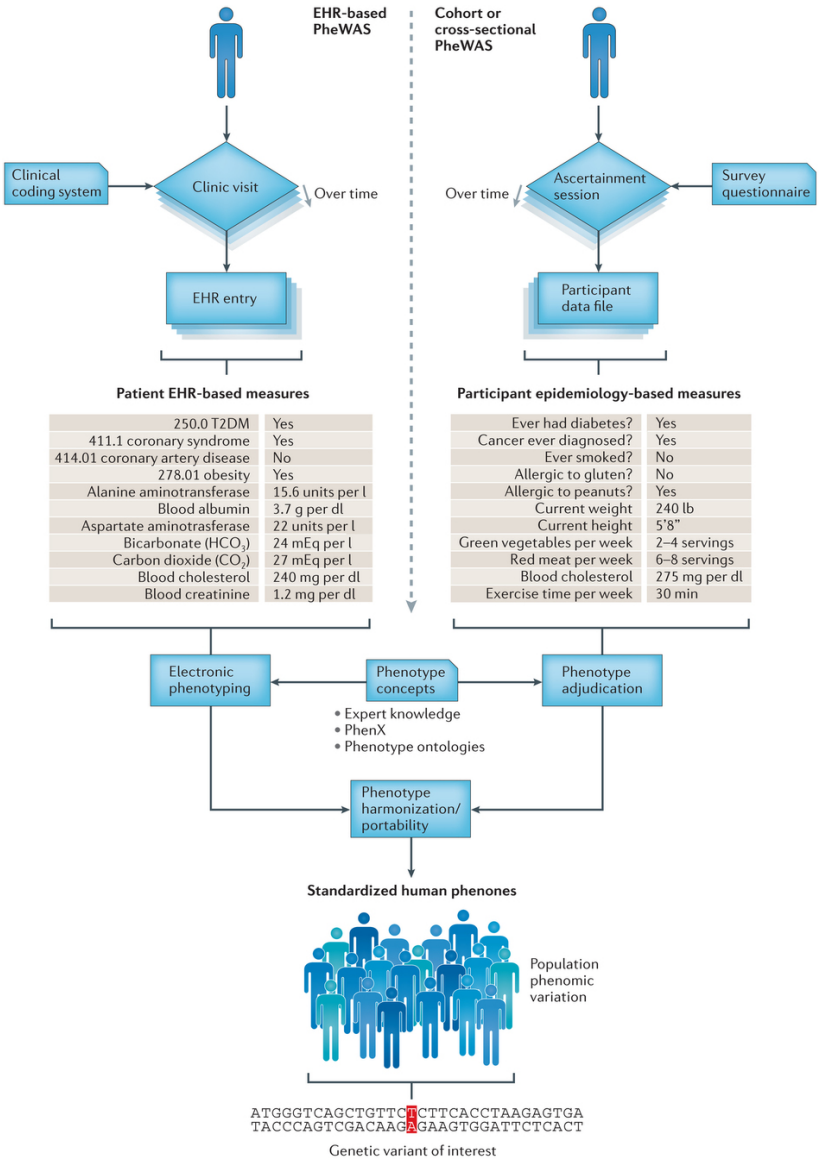
```r
library(PheWAS)
install.packages("devtools")
library(devtools)
install_github("PheWAS/PheWAS")

result <- phewas(phenotypes = diseases,
                 genotypes = genotypes,
                 covariates = csv.phenotypes[, c("id","gender", "age", "ethnicity")],
                 significance.threshold=c("bonferroni"))
phewasManhattan(result,
                annotate.angle=0,
                title="Metabolic disease PheWAS Manhattan Plot",
                annotate.phenotype = TRUE,
                annotate.snp = TRUE)
```



Phenome-wide association studies (PheWAS)

EHR-linked epidemiological study & biobank

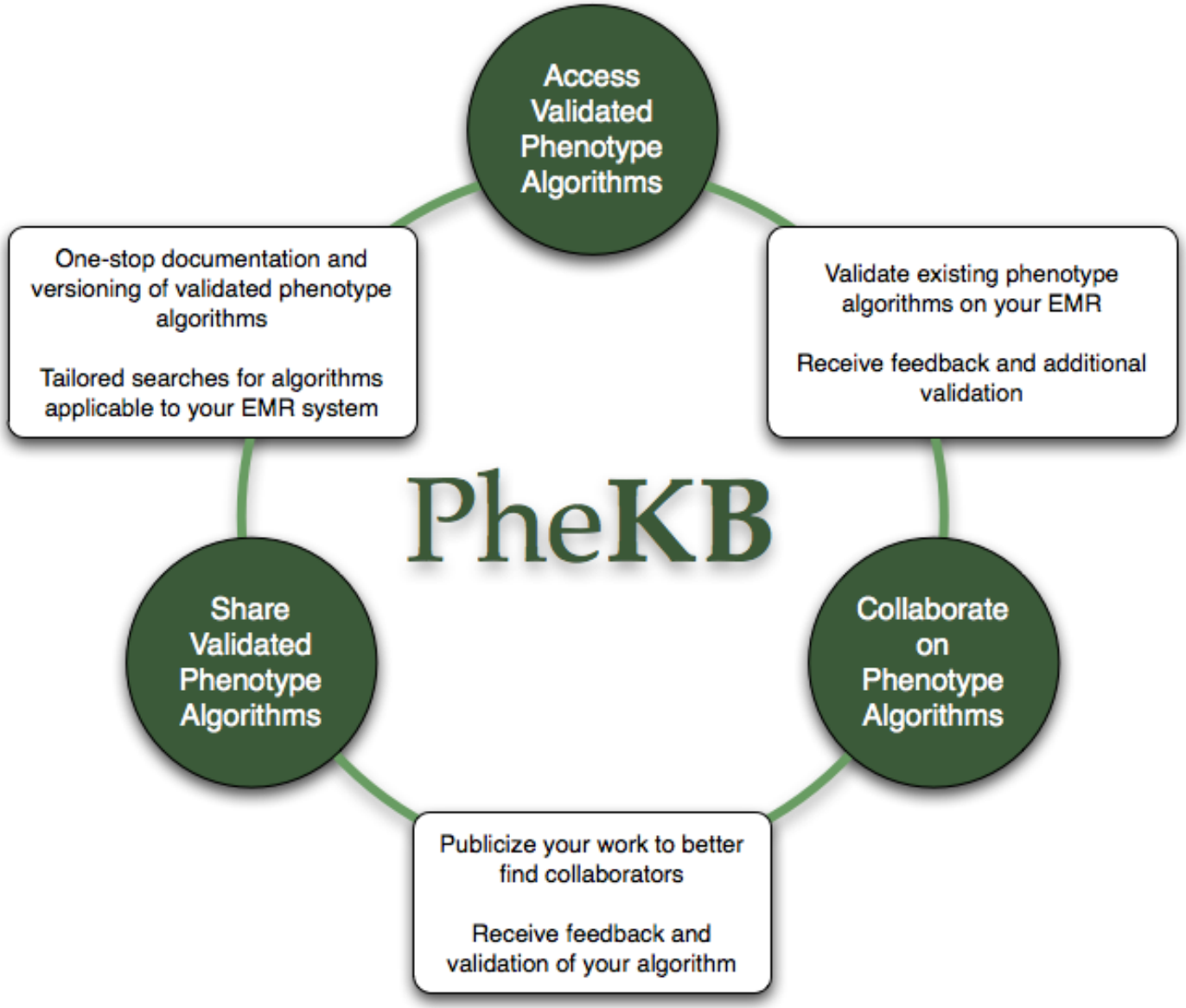


Nature Reviews | Genetics

- Electronic Health Records phenotyping
- ICD9-10
- EMERGE Network

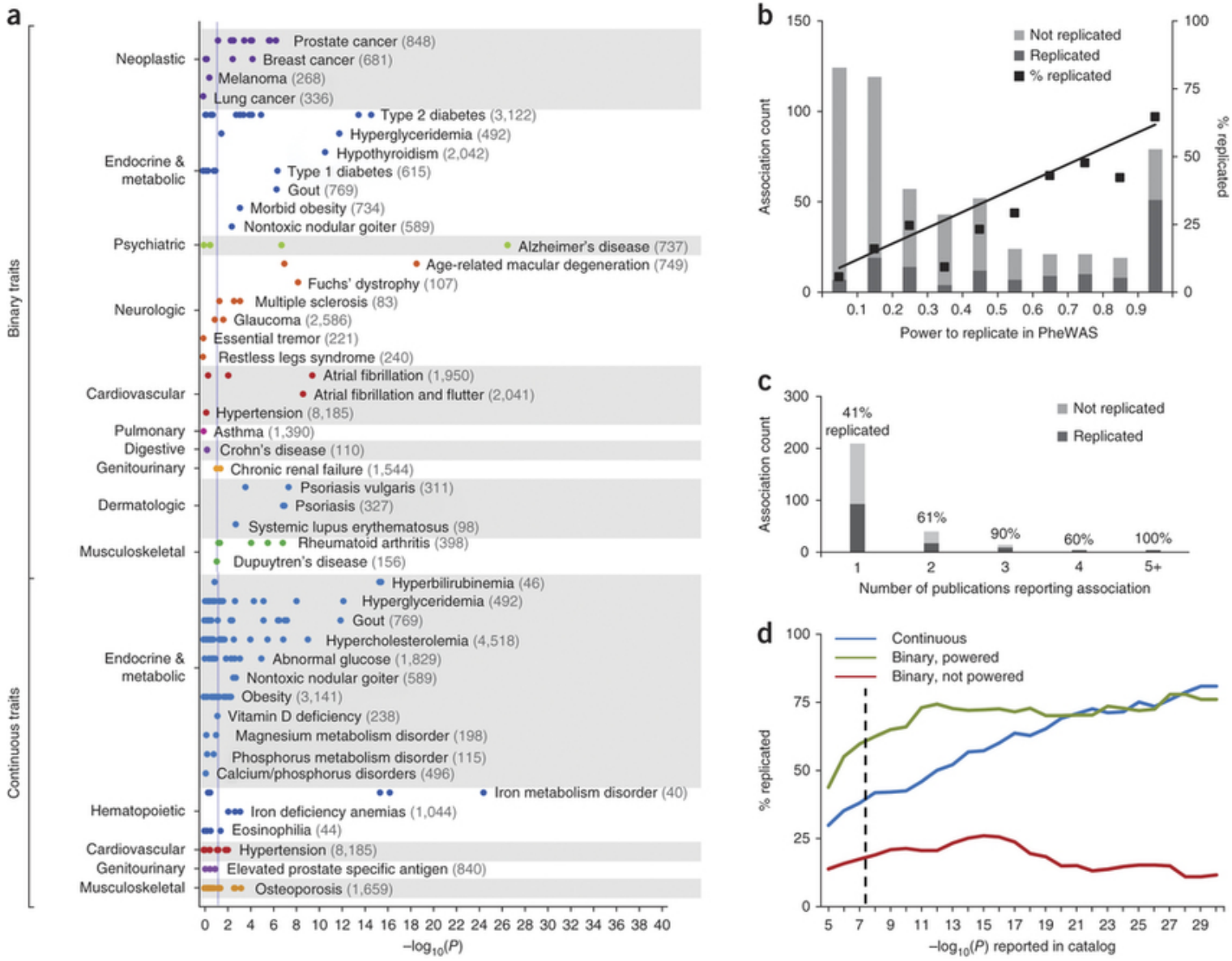
Phenome-wide association studies (PheWAS)

Electronic phenotyping



Phenome-wide association studies (PheWAS)

PheWAS catalog



Phenome-wide association studies (PheWAS)

Pleiotropy

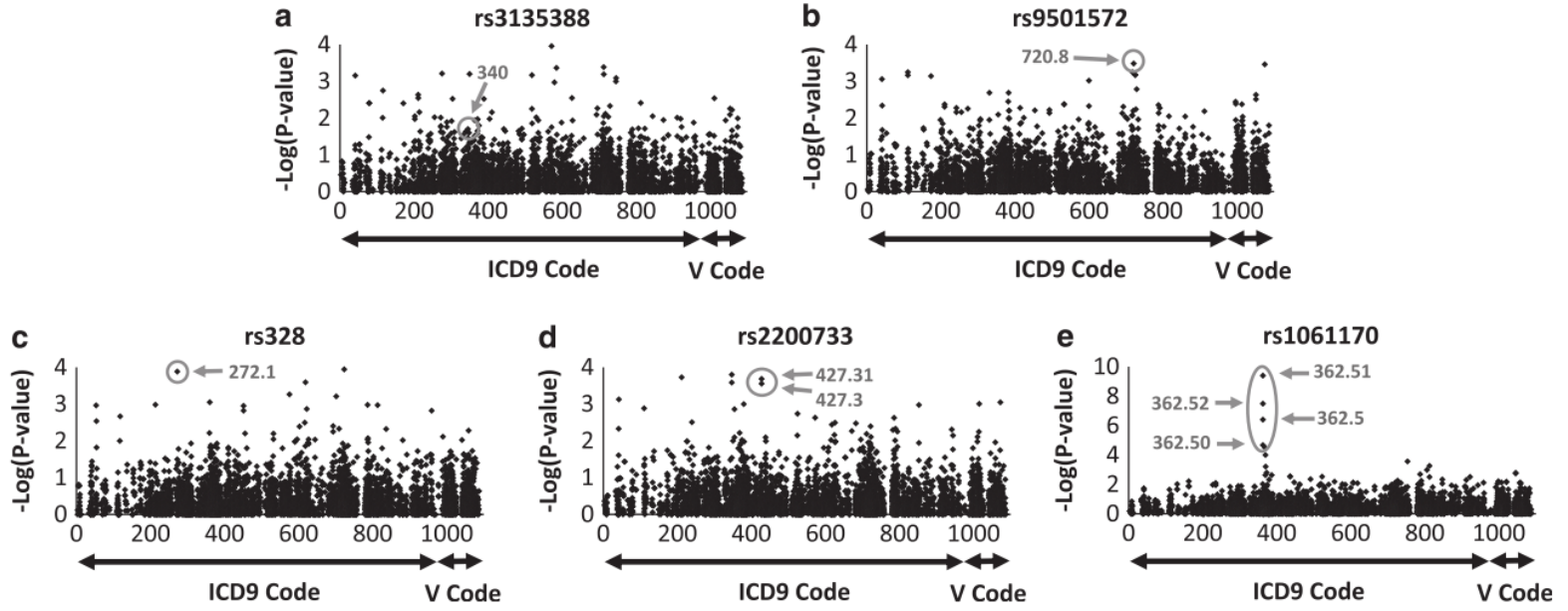
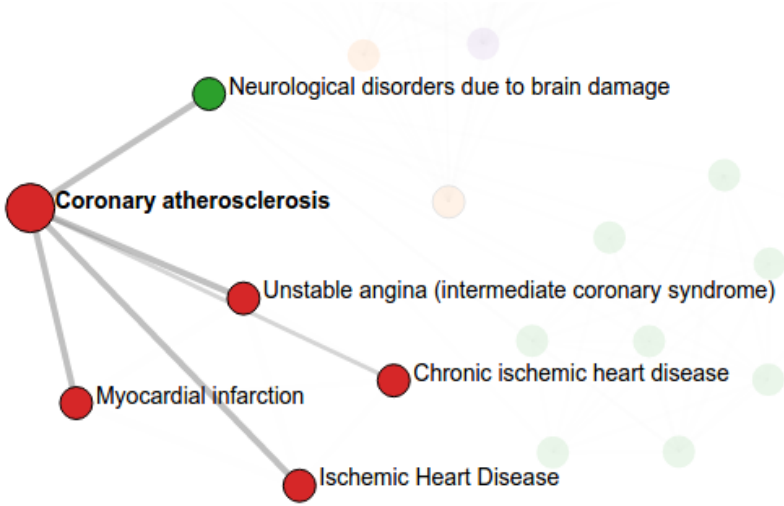


Figure 1 Manhattan plots of unadjusted −log10 (*P*-values) for the 4841 ICD9 and V codes that define the phenome. Highlighted are association results for (**a**) multiple sclerosis (ICD9 340) for rs3135388, (**b**) other inflammatory spondylopathies (ICD9 720.8) for rs9501572, (**c**) pure hyperglyceridemia (ICD9 272.1) for rs328, (**d**) atrial fibrillation (ICD9 427.31 and 427.3) for rs2200733, and (**e**) age-related macular degeneration (AMD) (ICD9 362.50, 362.51, 362.52, and 362.5) for rs1061170.

- Shared mechanism or biological pathway
- Novel drug targets
- Drug repositioning

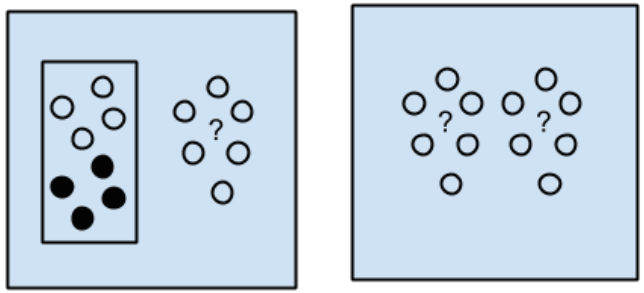
Phenome-wide association studies (PheWAS)

Phenotypes co-association network



- Each node represents a phenotype
- The color represents the clinical category of the phenotype
- The weight of the link depends on the number of co-association in the different analyses

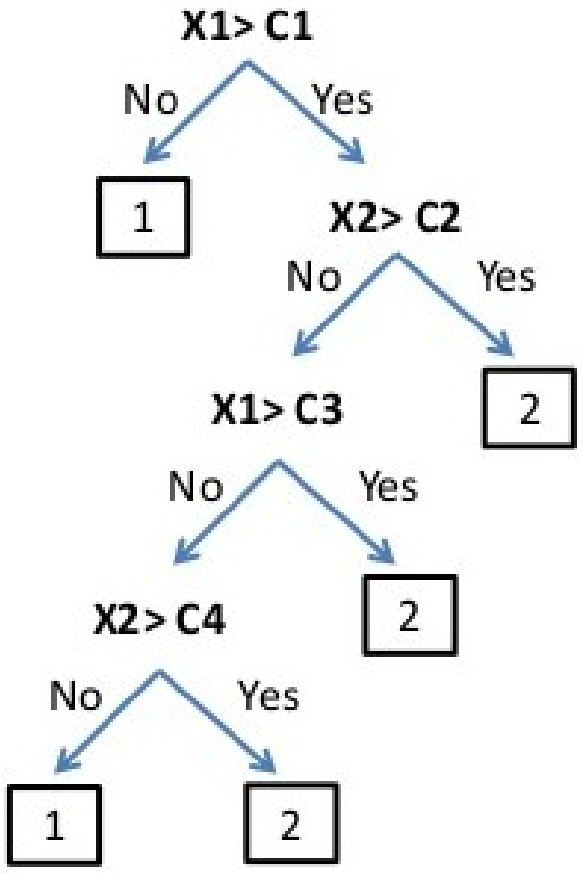
Machine learning valuable alternatives

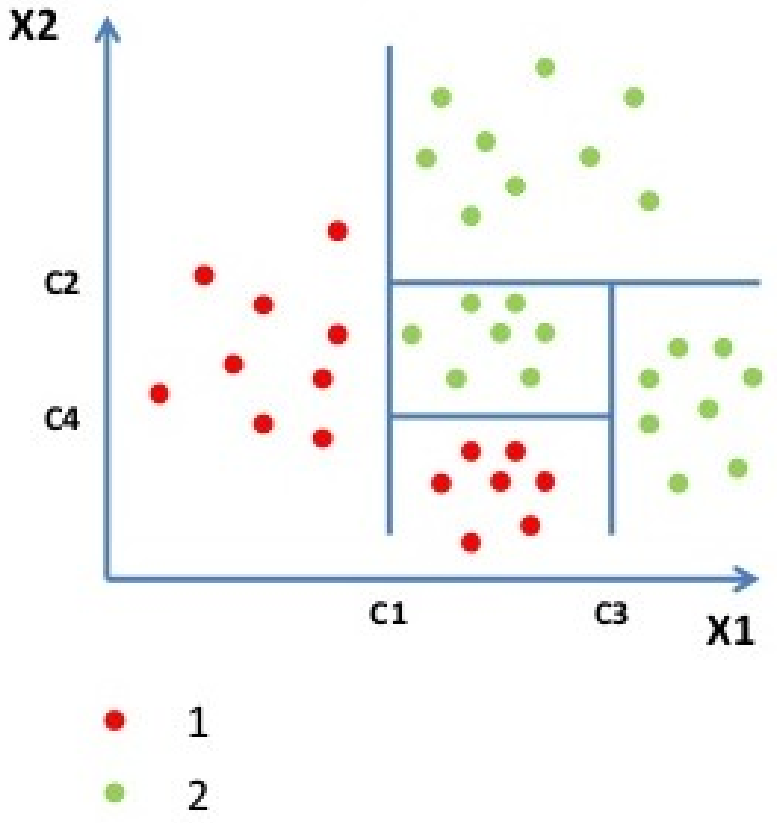


- Learn from known data (model and hypothesis generation)
- Make predictions about unknown data

Decision trees

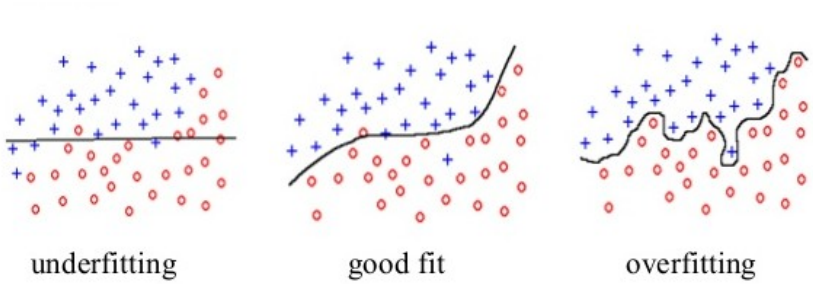
Building a tree

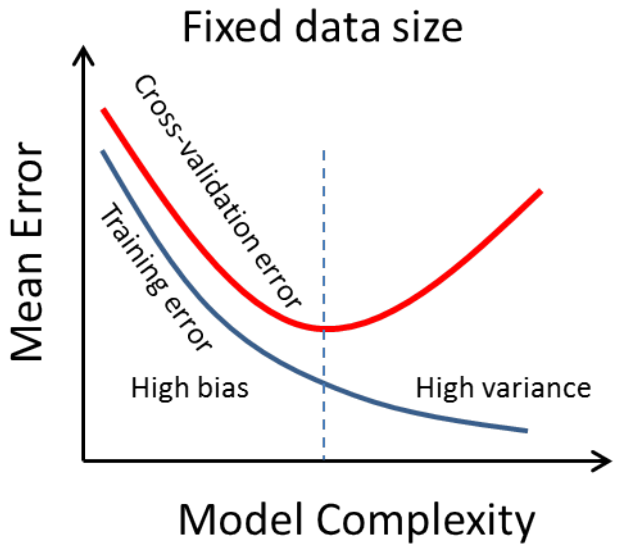




Decision trees

Problems

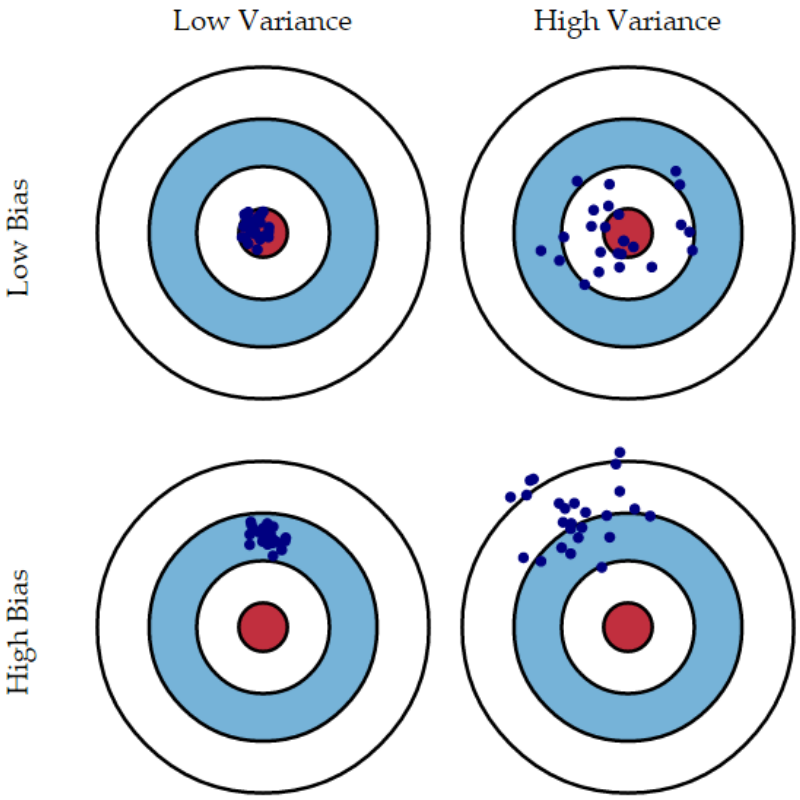




- Memorizing data: signal and noise
- Overfitting
- Poor generalization

Decision trees

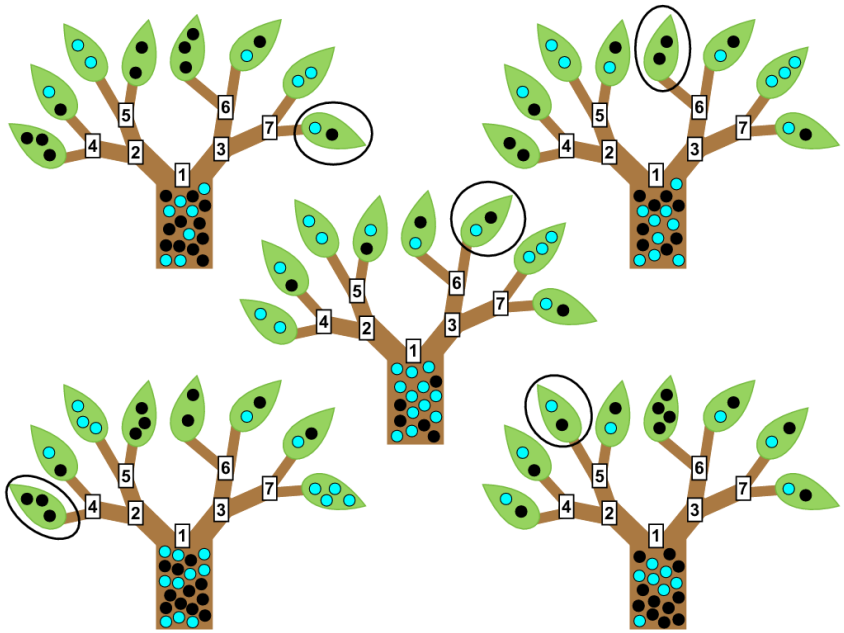
Bias-variance tradeoff



- Decision trees have low bias but high variance

Random Forests

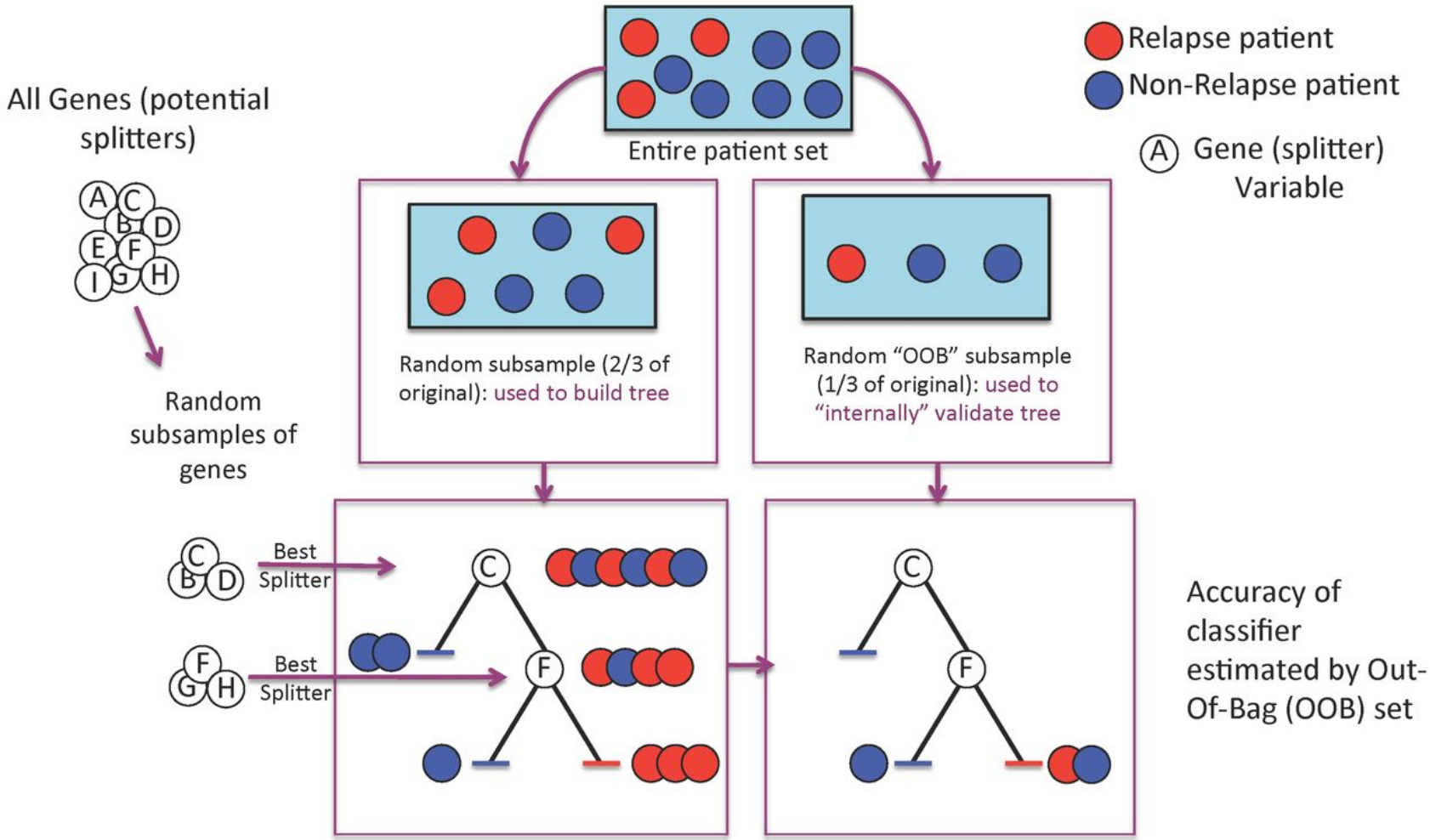
An ensemble of classification trees



- Collection of trees
- Non-deterministic using a two-stage randomization procedure
- Decorrelate trees
- Low variance

Random forests

Algorithm



Random forests

Hyperparameters

- Number of trees
- Number of selected variables per node ($\sqrt{M}$)
- Impurity measure (best split)
- Maximum depth of the tree before terminating into a prediction

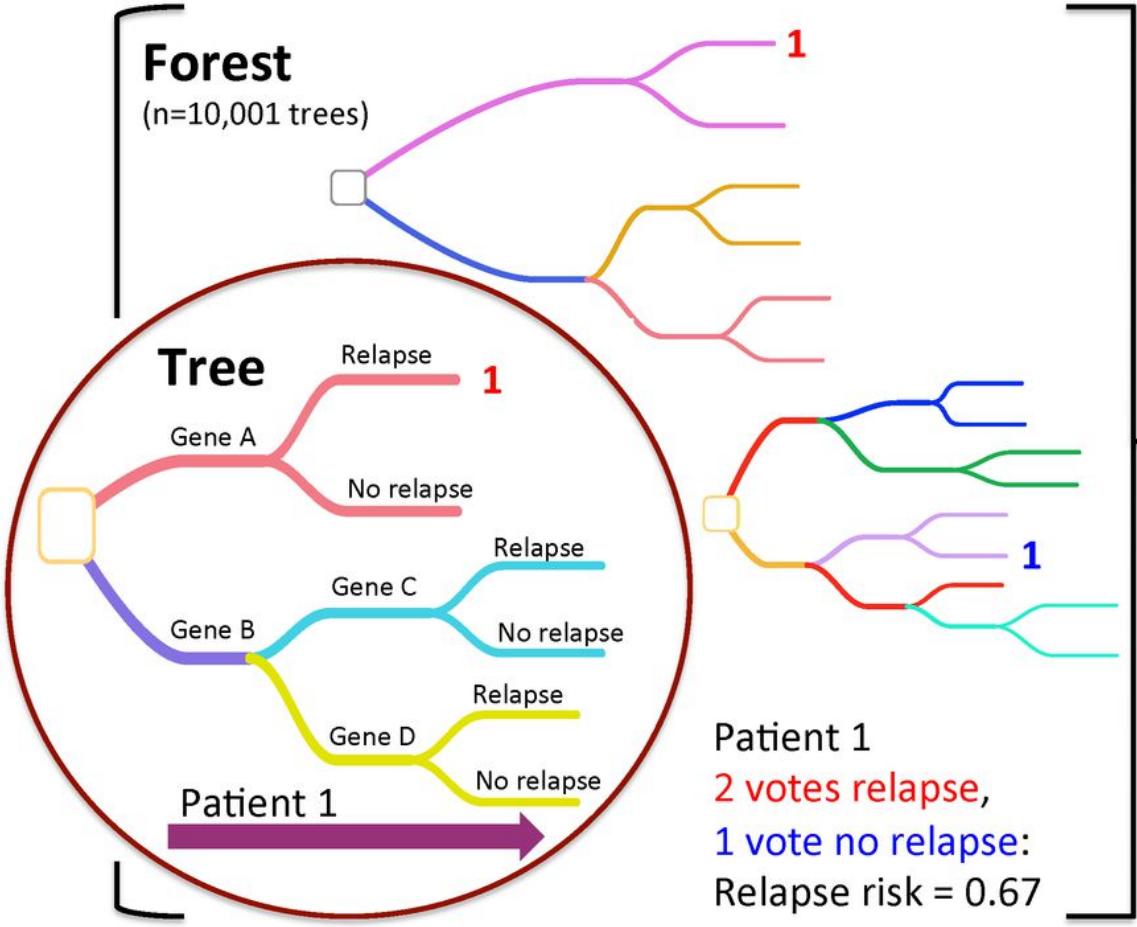
Random forests

Association studies with Random Forests

- Smallest possible set of genes that can still achieve good predictive performance
- Well suited for microarray data
- Can be used when there are many more variables than observations
- Good predictive performance even when most predictive variables are noise
- Incorporates interactions among predictor variables

Random forests

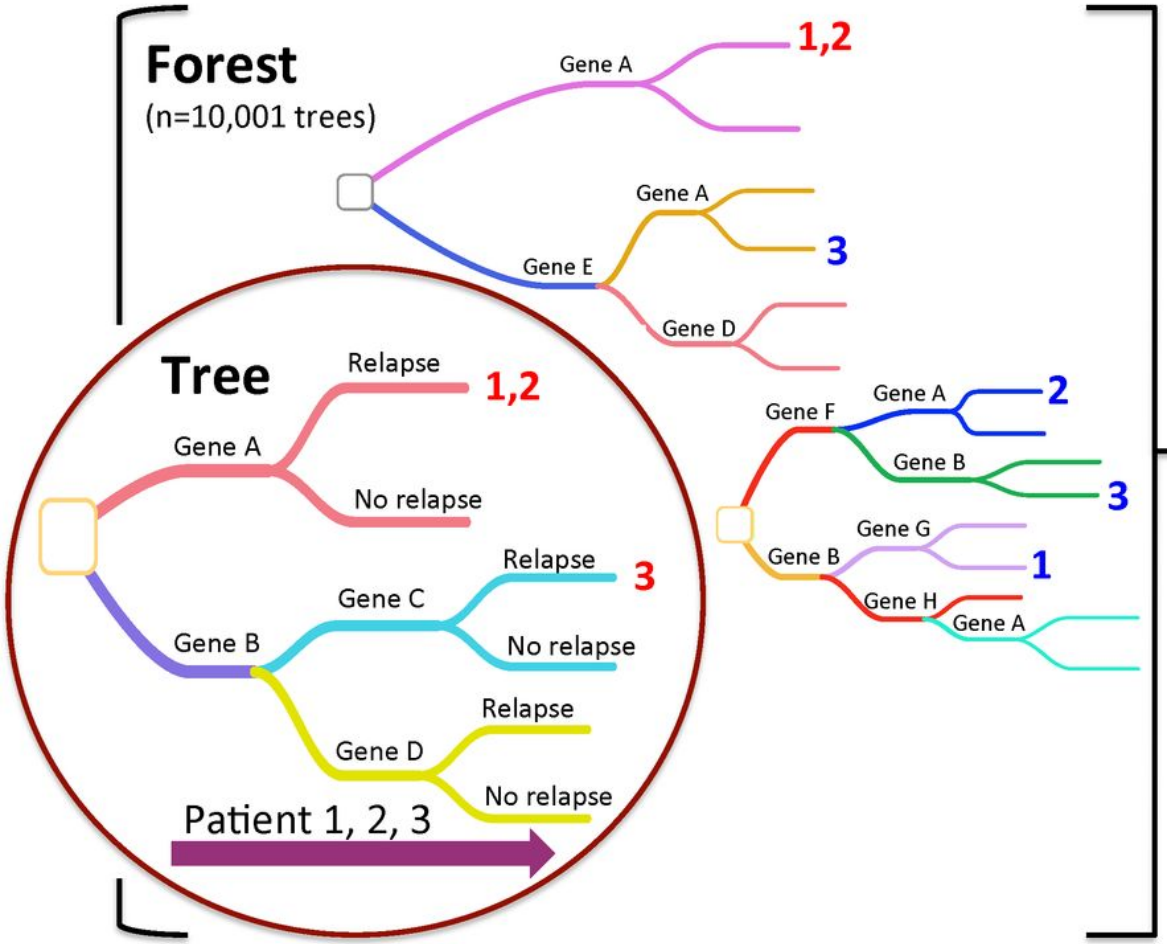
Classify new samples



- **Creates a forest of many binary decision trees**
- Each Patient traverses each tree until it reaches a terminal node
- At the terminal node each tree casts a vote (eg. “relapse”); the proportion of relapse votes from all votes is that patient’s predicted relapse risk

Random forests

Ranking variables

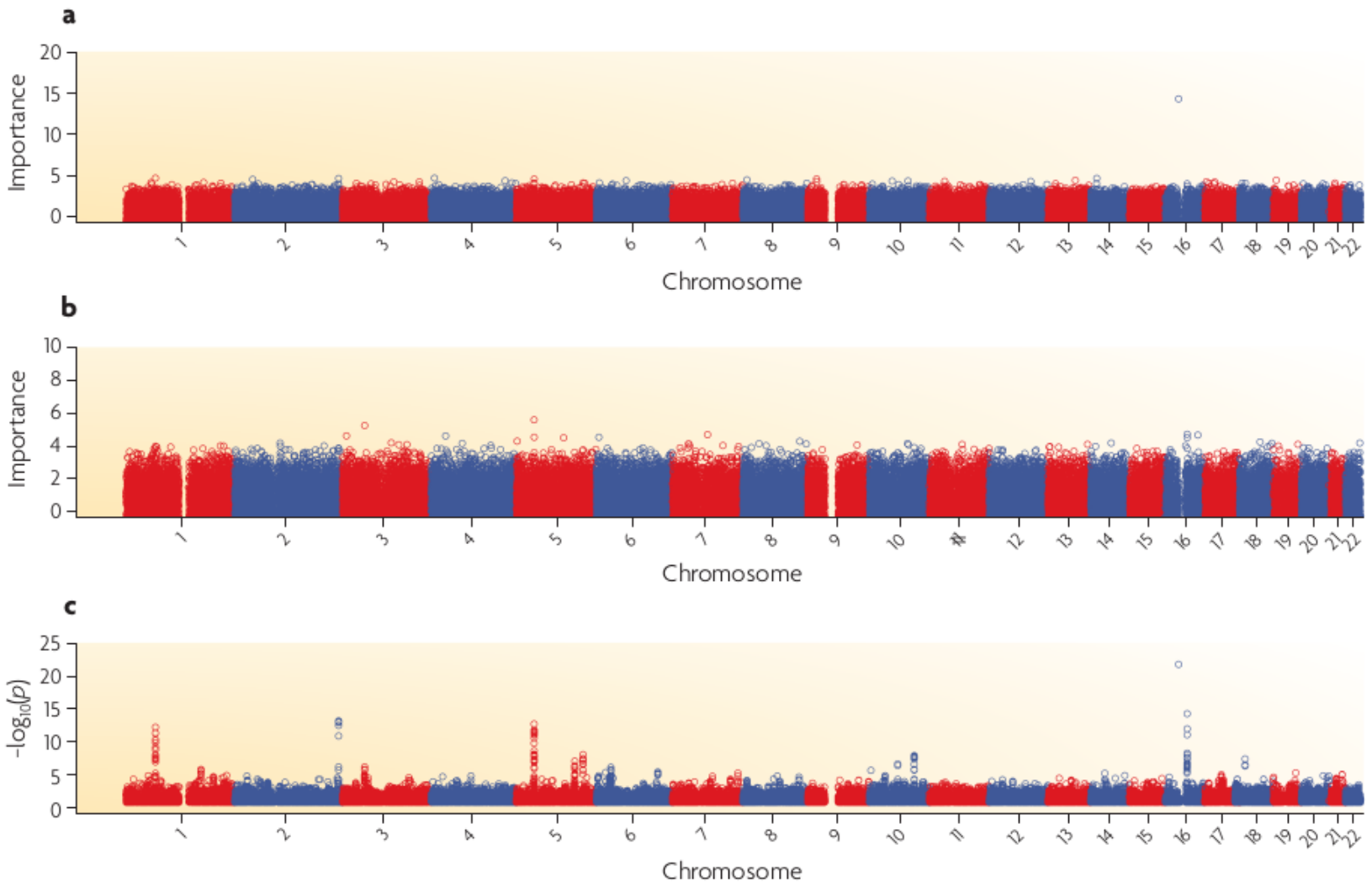


- The more often a gene is chosen as a splitter variable, the higher its "Variable Importance" – This can be used to prioritize which genes to select for an assay with limited gene measurements

Gene	Var. Imp.
Gene A	0.67
Gene B	0.20
Gene D	0.13
...	...

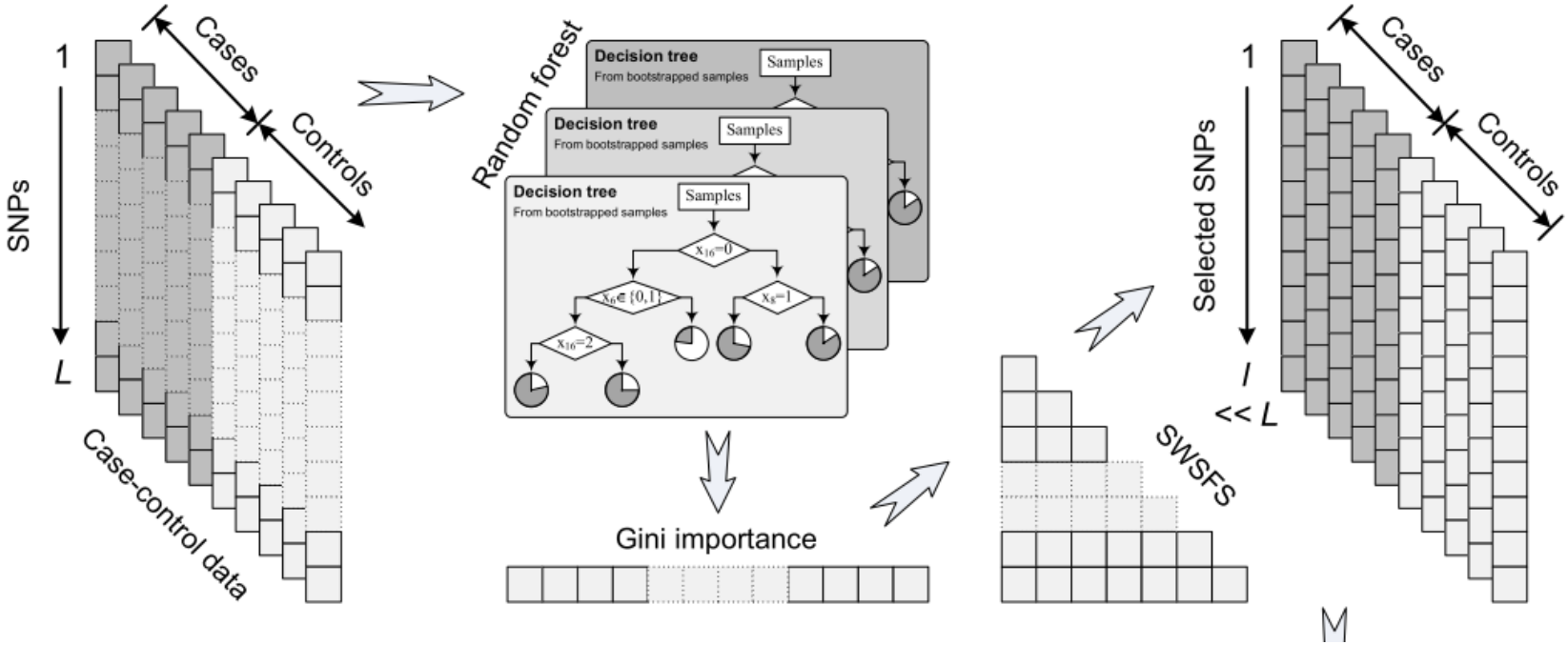
Random forests

Manhattan plot



Random forests

Genomic profiling



Summary

- Curse of dimensionality
- It will be worse with NGS!
- N << #variables
- Need to new exploratory, hypothesis-free methods

Thanks!