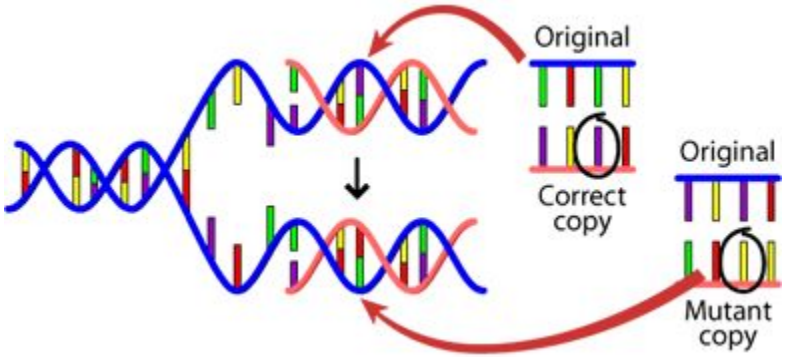
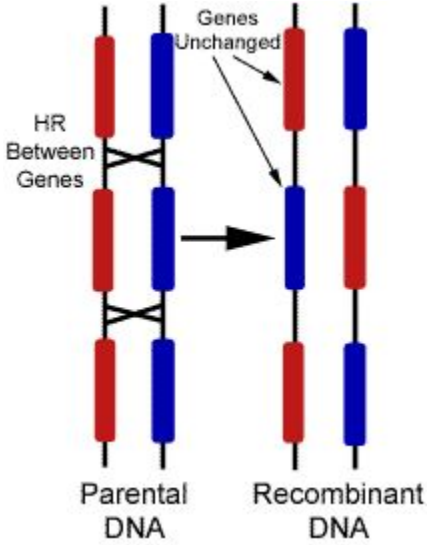
Week 05: Genetic variation & Quantitative genetics

- Genome-wide association studies
  - Regularized linear regression
  - Polygenic risk score
  - Statistical inference, P-values, & Multiple hypothesis testing

Genetic variation





Single Nucleotide Polymorphisms (SNPs)

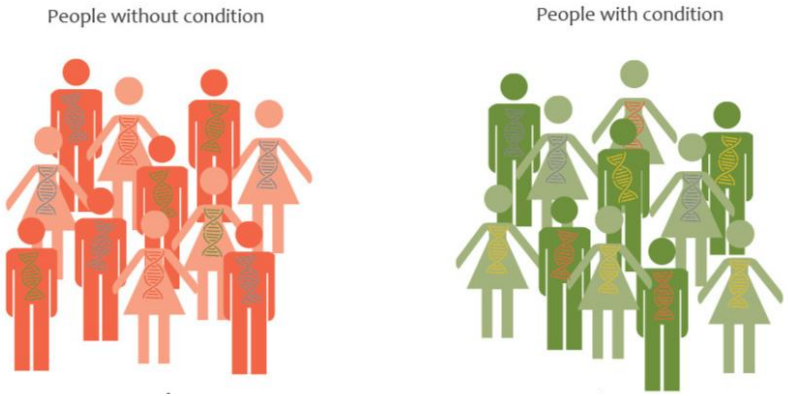
Insertions

Deletions

Copy Number Variants (CNVs)

- Duplications & deletions

Complex traits and diseases

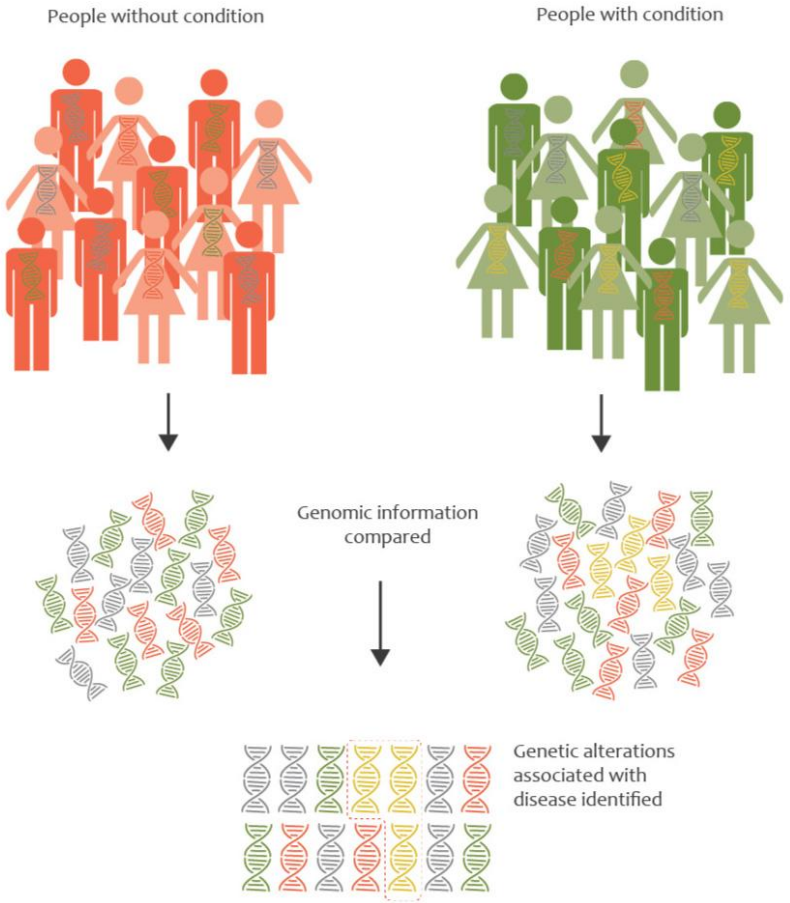


What factors contribute to a particular trait or the risk of getting a particular disease?

- Genetic factors (numerous)
- Other biological factors: age, sex, ethnicity
- Environmental factors (e.g. geography, nutrition)
- Interaction between genome and environment
    - Phenotypic Variation = G + E + GxE

How do you quantify how much the genome actually contributes?

Genome-wide Association Study (GWAS)



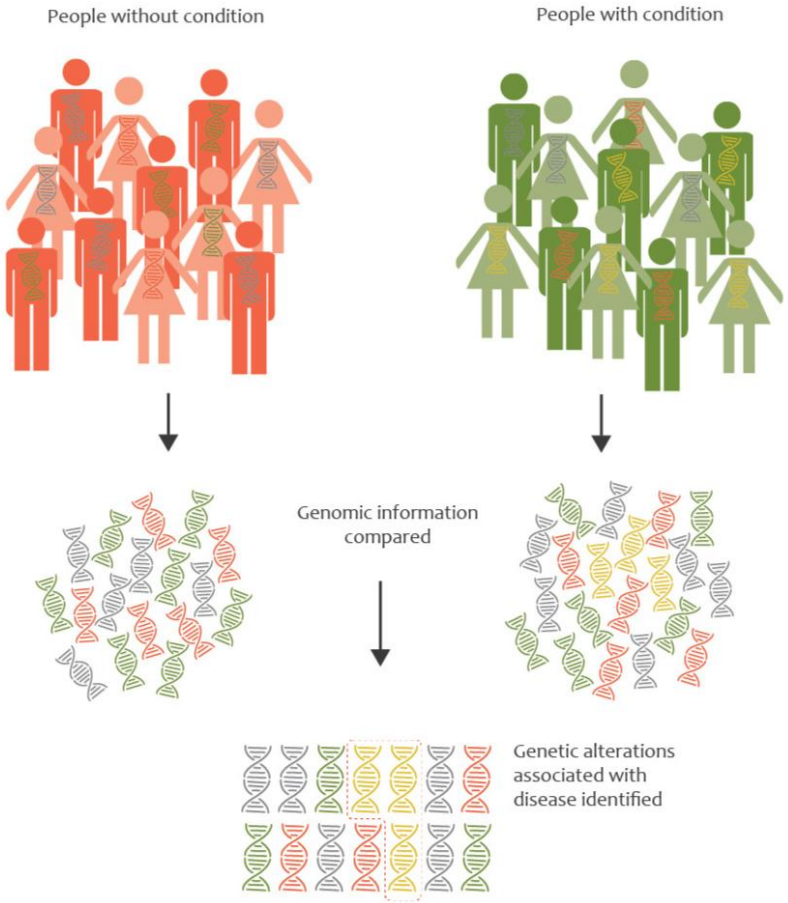
Expensive to sequence entire genome.

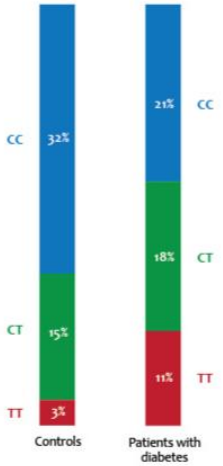
Focus on only a small part of the genome (SNPs) that are common and might contribute to variation.

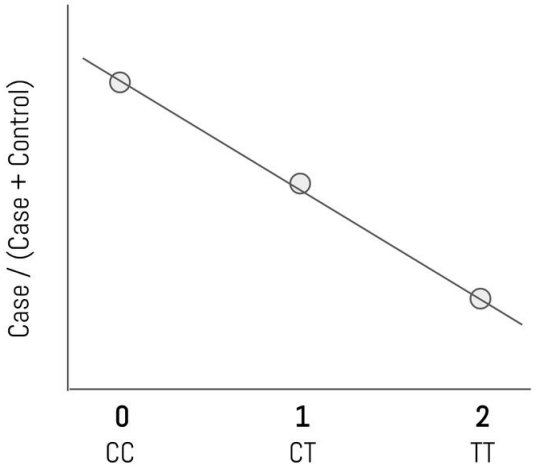
- About 5–10 million SNPs in the human genome.
- Use a SNP array – a small chip that has DNA probes that is complementary to regions in the genome that have SNPs.

Visscher (2017) AJHG; @genomicsedu

Genome-wide Association Study (GWAS)







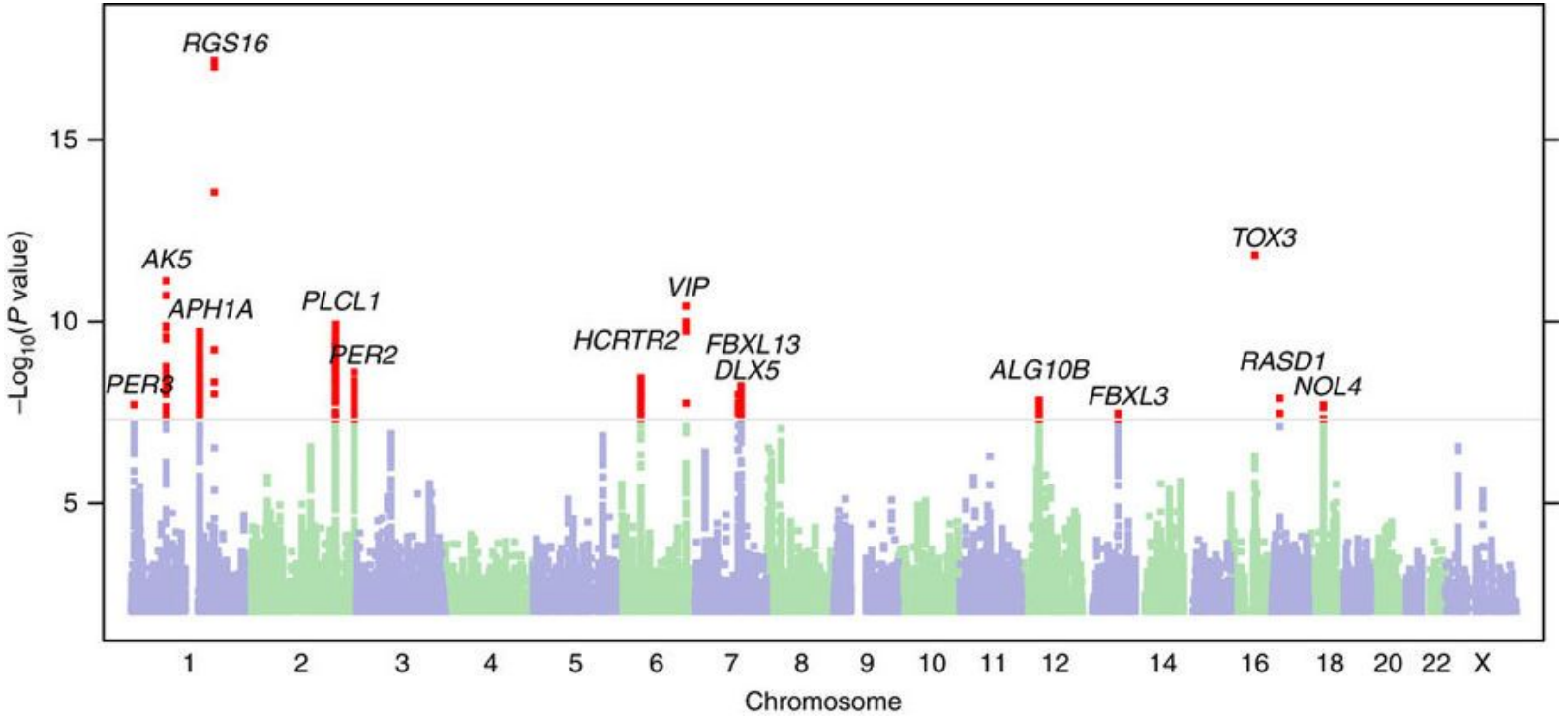
A C/T SNP from a hypothetical GWAS for type 2 diabetes

- Increase in freq of T allele in patients w/ diabetes compared to controls.
- We know where this SNP is on the genome → study surrounding sequence

Visscher (2017) AJHG; @genomicsedu

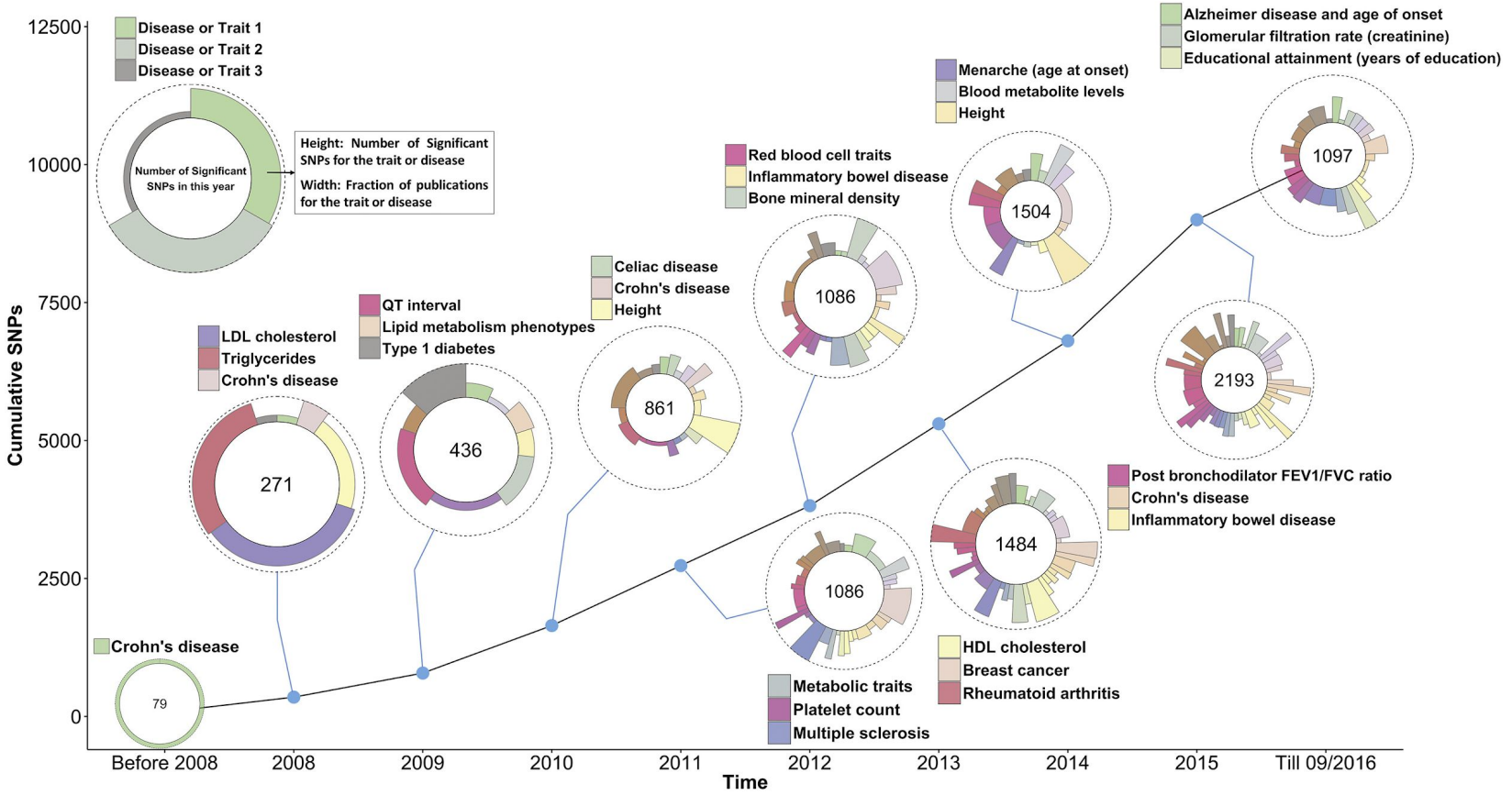
GWAS

GWAS of 89,283 individuals identifies genetic variants associated with… being a morning person!



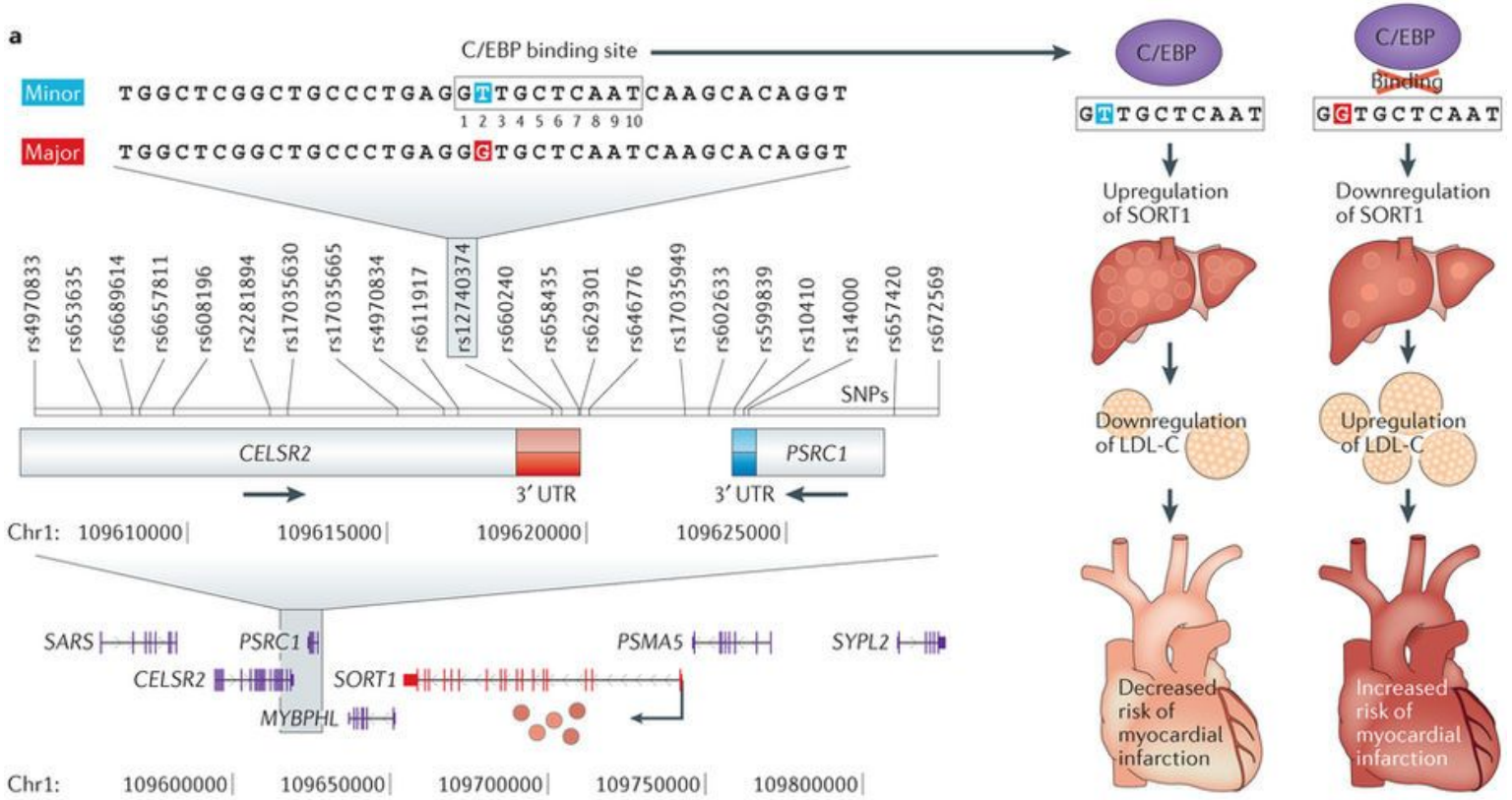
Hu (2016) Nat. Comm.

GWAS – Timeline of discoveries



Visscher (2017) AJHG

GWAS – Examples



Albert & Kruglyak (2015) Nat. Rev. Genet.

Statistical analysis of genome-wide association

- Description of the problem: cases, features
- Lasso: Regularized linear regression
  - Loss function: L1 vs. L2
  - Regularization (parameter: **λ**)
- Lasso is an example of “feature selection”

$$\hat{\beta}^{\text{lasso}} = \underset{\beta}{\operatorname{argmin}} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}$$

Statistical analysis of genome-wide association

- Solving lasso with the least-angle regression algorithm
- If a non-zero coefficient hits zero, remove it from the active set of predictors and recompute the joint direction.

Algorithm 3.2 *Least Angle Regression.*

1. Standardize the predictors to have mean zero and unit norm. Start with the residual $\mathbf{r} = \mathbf{y} - \bar{\mathbf{y}}$, $\beta_1, \beta_2, \ldots, \beta_p = 0$.
2. Find the predictor $\mathbf{x}_j$ most correlated with $\mathbf{r}$.
3. Move β_j from 0 towards its least-squares coefficient $\langle \mathbf{x}_j, \mathbf{r} \rangle$, until some other competitor $\mathbf{x}_k$ has as much correlation with the current residual as does $\mathbf{x}_j$.
4. Move β_j and β_k in the direction defined by their joint least squares coefficient of the current residual on $(\mathbf{x}_j, \mathbf{x}_k)$, until some other competitor $\mathbf{x}_l$ has as much correlation with the current residual.
5. Continue in this way until all p predictors have been entered. After $\min(N - 1, p)$ steps, we arrive at the full least-squares solution.

Elements of Statistical Learning