Statistical Methods for High Dimensional Biology
STAT/BIOF/GSAT 540

Human Genetics & GWAS

Keegan Korthauer
24 March 2021

With slide contributions from Sara Mostafavi

Announcements

- Analysis Assignment due Monday (29 March)

- Next week's lectures (29 and 31 March - our last! 😭) will both be **synchronous**
  - Guest lecturer **Dr. Yongjin Park** will be back with us to talk about causal inference in genomics on Monday
  - Guest lecturer **Dr. Jessica Dennis** will be with us to talk about polygenic risk scores and phenome-wide association studies on Wednesday

- Project presentations will take place during the last 3 class sessions (all synchronous)
  - Reminder that these Zoom sessions will be recorded (only made available to registered students in the course through canvas)
  - Peer review assignments for individual reports will be announced next week

Learning objectives

- Describe properties of genetic variants in terms of their prevalence in the population, and the likelihood of its effect using relevant terminology (e.g. **SNV vs SNP**, **allele**, **penetrance**)
- Explain the "**Common Disease-Common Variant**" hypothesis and its implications
- Understand the purpose of Genome-Wide Association Studies (**GWAS**)
- List two commonly used **statistical** approaches for GWAS analysis and describe advantages / disadvantages of each
- Understand the main mathematical ideas behind **Chi square testing** and **logistic regression**

Outline

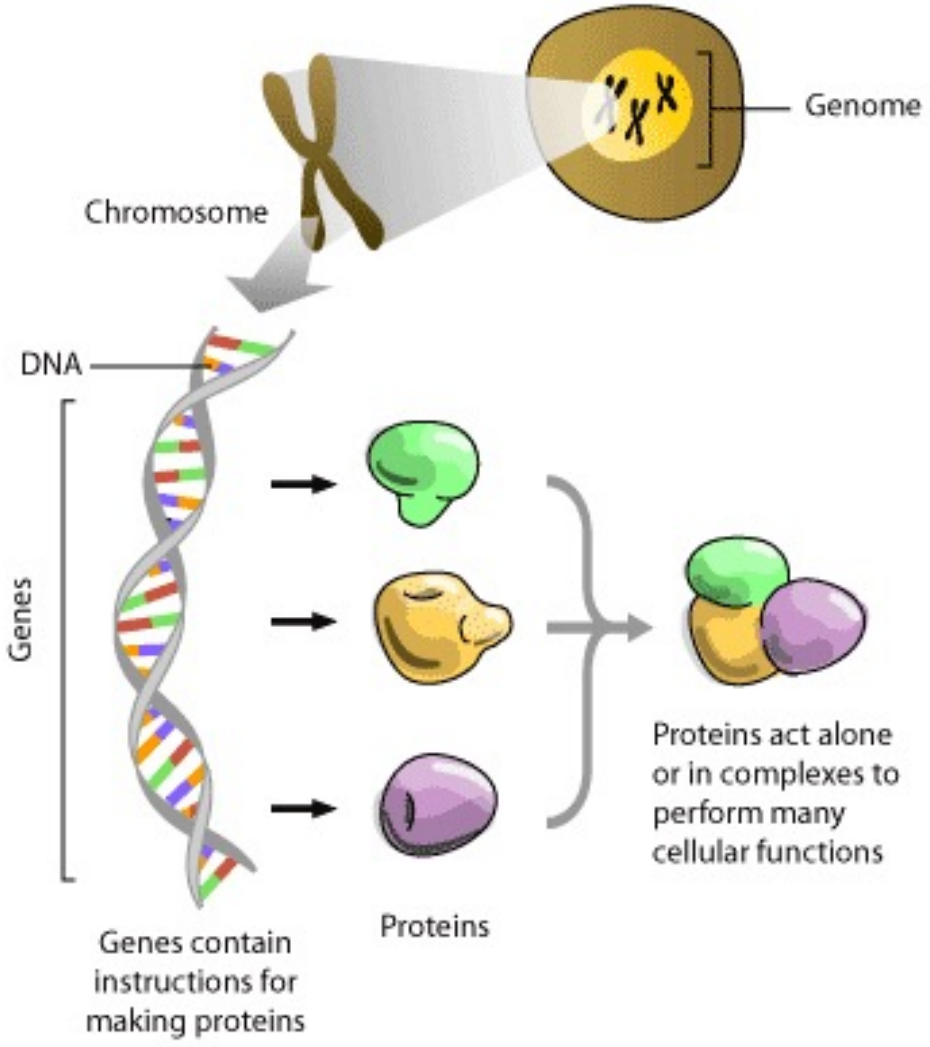
- **Terminology and fundamentals**
- Human genetics and disease
- GWAS – genome-wide association studies
  - Statistical testing: chi-square test/logistic regression

Human genetic differences are common

Many of these differences are **heritable** (passed on from parents to offspring)

image source: https://www.broadinstitute.org/medical-population-genetics

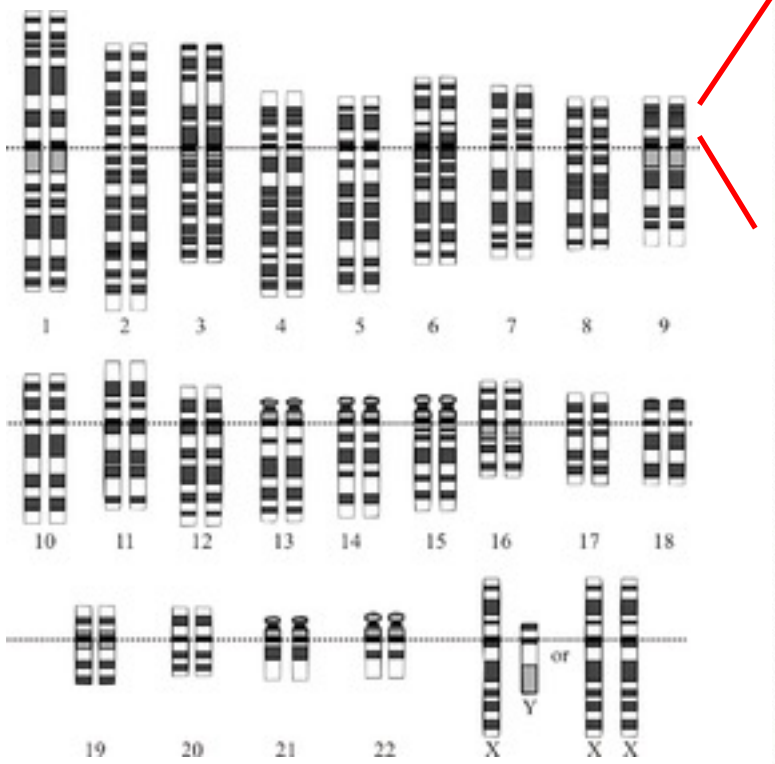
Recall: definitions



- Every (nucleated) cell has a complete copy of the **genome** (DNA)
- **Gene**: a segment of DNA that has a "function"
- **Genotype**: genetic makeup of an individual
- **Phenotype**: observable trait
- **Genetic association**: relationship between genotype and phenotype

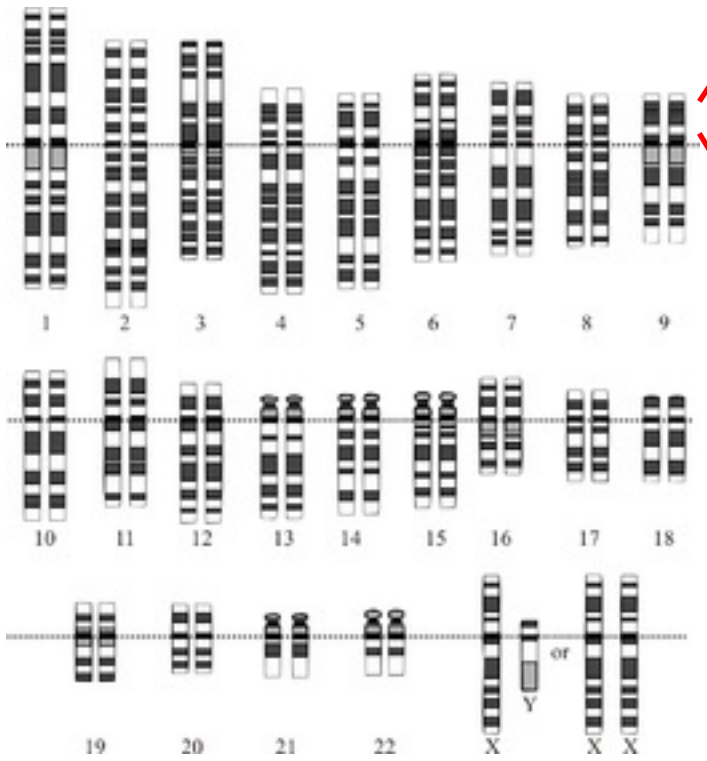
image source: https://www.scq.ubc.ca/the-human-genome-project-the-impact-of-genome-sequencing-technology-on-human-health/

“Genome: bought the book; hard to read.” –Eric Lander



CATGACGTCGCGGACAACCCAGAATTGTCTTGAGCGATGGTAAGATCTAACCTCACTGCCGGGGGAGGCTCATAC
CTGGGGCTTTACTGATGTCATACCGTCTTGCACGGGGATAGAATGACGGTGCCCGTGTCTGCTTGCCTCGAAGCA
ATTTTCTGAAAGTTACAGACTTCGATTAAAAAGATCGGACTGCGCGTGGGCCCGGAGAGACATGCGTGGTAGTCA
TTTTTCGACGTGTCAAGGACTCAAGGGAATAGTTTGGCGGGAGCGTTACAGCTTCAATTCCCAAAGGTCGCAAGA
CGATAAAATTCAACTACTGGTTTCGGCCTAATAGGTCACGTTTTATGTGAAATAGAGGGGAACCGGCTCCCAAAT
CCCTGGGTGTTCTATGATAAGTCCTGCTTTATAACACGGGGCGGTTAGGTTAAATGACTCTTCTATCTTATGGTG
ATCCAAGCGCCCGCTAATTCTGTTCTGTTAATGTTCATACCAATACTCACATCACATTAGATCAAAGGATCCCCG
AGCCCAGTCGCAAGGGTCTGCTGCTGTTGTCGACGCCTCATGTTACTCCTGGAATCTACCTGCCCTCCCCTCACC
GGTTAAGGCGTGTGATCGACGATGCAGGTATACATCGGCTCGGACCTACAGTGGTCGATCGACTGGCTACTGGCT
TCGCGGTTCGGCGCGTAGTTGAGTGCGATAACCCAACCGGTGGCAAGTAGCAAGAAGACCTACCTGGGTCACCTT
AGACAACCTAACTAATAGTCTCTAACGGGGAATTACCTTTACCAGTCTCATGCCTCCAATATATCTGCACCGCTT
CAATGATATCGCCCACAGAAAGTAGGGTCTCAGGTATCGCATACGCCGCGCCCGGGTCCCAGCTACGCTCAGGAC
GACAGTAGAGAGCTATTGTGTAATTCAGGCTCAGCATTCATCGACCTTTCCTGTTGTGAATATTGTGCTAATGCA
TCTCGTCCGTAACGATCTGGGGGGCAAAACCGAATATCCGTATTCTCGTCCTACGGGTCCACAATGAGAAAGTCC
TGCGCGTGATCGTCAGTTAAGTTAAATTAATTCAGGCTACGGTAAACTTGTAGTGAGCTAAGAATCACGGGAATC
ACGGGTTCGCTACAGATGAACTGAATTTATACACGGACAACTCATCGCCCATTTGGGCGTGGGCACCGCAGATCA
AAAGTGGCAGATTAGGAGTGCTTGATCAGGTTAGCAGGTGGACTGTATCCAACAGCGCATCAAACTTCAATAAAT
CCAAAGCGTTGTAGTGGTCTAAGCACCCCTGAACAGTGGCGCCCATCGTTAGCGTAGTACAACCCTTCCCCCTTG
AGGTGCGACATGGGGCCAGTTAGCCTGCCCTATATCCCTTGCACACGTTCAATAAGAGGGGCTCTACAGCGCCGC
TTTTTAAATTAGGATGCCGACCCCATCATTGGTAACTGTATGTTCATAGATATTTCTTCAGGAGTAATAGCGACA
AGCTGACACGCAAGGGTCAACAATAATTTCTACTATCACCCCGCTGAACGACTGTCTTTGCAAGAACCAACTGGG
CTTAGATTCGCGTCCTAACGTAGTGAGGGCCGAGTCATATCATAGATCAGGCATGAGAAACCGACGTCGAGTCTA
CACACGAGTTGTAAACAACTTGATTGCTATACTGTAGCTACCGCAAGGATCTCCTACATCAAAGACTACTGGGCG
ATCTGGATCCGAGTCAGAAATACGAGTTAATGCAAATTTACGTAGACCGGTGAAAACACGTGCCATGGGTTGCGT
AGACCGTAGTCAGAAGTGTGGCGCGCTATTCGTACCGAACCGGTGGAGTATACAGAATTGCTCTTCTACGACGTA
AGGAGCTCGGTCCCCAATGCACGCCAAAAAAGGAATAAAGTATTCAAACTGCGCATGGTCCCTCCGCCGGTGGCA
CTATTATCCATCCGAACGTTGAACCTACTTCCTCGGCTTATGCTGTCCTCAACAGTATCGCTTATGAATCGCATG
CGGCTGTGGATCTTAACGGCCACATTCTTAATTCCGACCGATCACCGATCGCCTTTCCTCGCTGGTACAATGAGT
ACTAAGTTATCCAGATCAAGGTTTGAACGGACTCGTATGACATGTGTGACTGAACCCGGGAGGAAATGCAGAGAA
CTGTTTCAAGGCCTCTGCTTTGGTATCACTCAATATATTCAGACCAGACAAGTGGCAAAATTTCGTGCGCCTCTC
CTAGGTATTCACGCAACCGTCGTAACATGCACTAAGGATAACTAGCGCCAGGGGGGCATACTAGGTCCCGGAGCT
AAAGACTACCCTATGGATTCCTTGGAGCGGGGACAATGCAGACCGGTTACGACACAATTATCGGGATCGTCTAGA
GGTATTATTAGCAAGACAATAAAGGACATTGCACAGAGACTTATTAGAATTCAACAAACAGGATCATATCATGCG
GTGTTGGGTCGGGCAAGTCCCCGAAGCTCGGCCAAAAGATTCGCCATGGAACCGTCTGGTCCTGTTAGCGTGTAC
GCCTGCTCCTGTTCCGGGTACCATAGATAGACTGAGATTGCGTCAAAAAATTGCGGCGAAAATAGAGGGGCTCCT
TGTAGAAATACCAGACTGGGGAATTTAAGCGCTTTCCACTATCTGAGCGACTAAACATCAACAAATGCGTCTACT
CGAATCCGCAGTAGGCAATTACAACCTGGTTCAGATCACTGGTTAATCAGGGATGTCTTCATAAGATTATACTTG
CCCCGACGCGACAGCTCTTCAAGGGGCCGATTTTTGGACTTCAGATACGCTAGAATTTAAAGGGTCTCTTACACC
TGCTGCGGCCTGCAGGGACCCCTAGAACTTGCCGCCTACTTGTCTCAGTCTAATAACGCGCGAAGCCGTGGGGCA
CGTGACCTTAAGTCGCAGAGCGAGTGATGAATTTGGGACGCTAATATGGGTGAATAGAGACTTATATCATCAGGG

“Genome: bought the book; hard to read.” –Eric Lander



CATGACGTCGCGGACAACCCAGAATTGTCTTGAGCGATGGTAAGATCTAACCTCACTGCCGGGGGAGGCTCATAC
CTGGGGCTTTACTGATGTCATACCGTCTTGCACGGGGATAGAATGACGGTGCCCGTGTCTGCTTGCCTCGAAGCA
ATTTTCTGAAAGTTACAGACTTCGATTAAAAAGATCGGACTGCCGCGTGGGCCCGGAGAGACATGCGTGGTAGTCA
TTTTTCGACGTGTCAAGGACTCAAGGGAATAGTTTGGCGGGAGCGTTACAGCTTCAATTCCCAAAGGTCGCAAGA
CGATAAAATTCAACTACTGGTTTCGGCCTAATAGGTCACGTTTTATGTGAAATAGAGGGGAACCGGCTCCCAAAT
CCCTGGGTGTTCTATGATAAGTCCTGCTTTATAACACGGGGCGGTTAGGTTAAATGACTCTTCTATCTTATGGTG
ATCCAAGCGCCCGCTAATTCTGTTCTGTTAATGTTCATACCAATACTCACATCACATTAGATCAAAGGATCCCCG
AGCCCAGTCGCAAGGGTCTGCTGCTGTTGTCGACGCCTCATGTTACTCCTGGAATCTACCTGCCCTCCCCTCACC
GGTTAAGGCGTGTGATCGACGATGCAGGTATACATCGGCTCGGACCTACAGTGGTCGATCGACTGGCTACTGGCT
TCGCGGTTCGGCGCGTAGTTGAGTGCGATAACCCAACCGGTGGCAAGTAGCAAGAAGACCTACCTGGGTCACCTT
AGACAACCTAACTAATAGTCTCTAACGGGGAATTACCTTTACCAGTCTCATGCCTCCAATATATCTGCACCGCTT
CAATGATATCGCCCACAGAAAGTAGGGTCTCAGGTATCGCATACGCCGCGCCCGGGTCCCAGCTACGCTCAGGAC
GACAGTAGAGAGCTATTGTGTAATTCAGGCTCAGCATTCATCGACCTTTCCTGTTGTGAATATTGTGCTAATGCA
TCTCGTCCGTAACGATCTGGGGGGCAAAACCGAATATCCGTATTCTCGTCCTACGGGTCCACAATGAGAAAGTCC
TGCGCGTGATCGTCAGTTAAGTTAAATTAATTCAGGCTACGGTAAACTTGTAGTGAGCTAAGAATCACGGGAATC
ACGGGTTCGCTACAGATGAACTGAATTTATACACGGACAACTCATCGCCCATTTGGGCGTGGGCACCGCAGATCA
AAAGTGGCAGATTAGGAGTGCTTGATCAGGTTAGCAGGTGGACTGTATCCAACAGCGCATCAAACTTCAATAAAT
CCAAAGCGTTGTAGTGGTCTAAGCACCCCTGAACAGTGGCGCCCATCGTTAGCGTAGTACAACCCTTCCCCCTTG
AGGTGCGACATGGGGCCAGTTAGCCTGCCCTATATCCCTTGCACACGTTCAATAAGAGGGGCTCTACAGCGCCGC
TTTTTAAATTAGGATGCCGACCCCATCATTGGTAACTGTATGTTCATAGATATTTCTTCAGGAGTAATAGCGACA
AGCTGACACGCAAGGGTCAACAATAATTTCTACTATCACCCCGCTGAACGACTGTCTTTGCAAGAACCAACTGGG
CTTAGATTCGCGTCCTAACGTAGTGAGGGCCGAGTCATATCATAGATCAGGCATGAGAAACCGACGTCGAGTCTA
CACACGAGTTGTAAACAACTTGATTGCTATACTGTAGCTACCGCAAGGATCTCCTACATCAAAGACTACTGGGCG
ATCTGGATCCGAGTCAGAAATACGAGTTAATGCAAATTTACGTAGACCGGTGAAAACACGTGCCATGGGTTGCGT
AGACCGTAGTCAGAAGTGTGGCGCGCTATTCGTACCGAACCGGTGGAGTATACAGAATTGCTCTTCTACGACGTA
AGGAGCTCGGTCCCCAATGCACGCCAAAAAAGGAATAAAGTATTCAAACTGCGCATGGTCCCTCCGCCGGTGGCA
CTATTATCCATCCGAACGTTGAACCTACTTCCTCGGCTTATGCTGTCCTCAACAGTATCGCTTATGAATCGCATG
CGGCTGTGGATCTTAACGGCCACATTCTTAATTCCGACCGATCACCGATCGCCTTTCCTCGCTGGTACAATGAGT
ACTAAGTTATCCAGATCAAGGTTTGAACGGACTCGTATGACATGTGTGACTGAACCCGGGAGGAAATGCAGAGAA
CTGTTTCAAGGCCTCTGCTTTGGTATCACTCAATATATTCAGACCAGACAAGTGGCAAAATTTCGTGCGCCTCTC
CTAGGTATTCACGCAACCGTCGTAACATGCACTAAGGATAACTAGCGCCAGGGGGGCATACTAGGTCCCGGAGCT
AAAGACTACCCTATGGATTCCTTGGAGCGGGGACAATGCAGACCGGTTACGACACAATTATCGGGATCGTCTAGA
GGTATTATTAGCAAGACAATAAAGGACATTGCACAGAGACTTATTAGAATTCAACAAACAGGATCATATCATGCG
GTGTTGGGTCGGGCAAGTCCCCGAAGCTCGGCCAAAAGATTCGCCATGGAACCGTCTGGTCCTGTTAGCGTGTAC
GCCTGCTCCTGTTCCGGGTACCATAGATAGACTGAGATTGCGTCAAAAAATTGCGGCGAAAATAGAGGGGCTCCT
TGTAGAAATACCAGACTGGGGAATTTAAGCGCTTTCCACTATCTGAGCGACTAAACATCAACAAATGCGTCTACT
CGAATCCGCAGTAGGCAATTACAACCTGGTTCAGATCACTGGTTAATCAGGGATGTCTTCATAAGATTATACTTG
CCCCGACGCGACAGCTCTTCAAGGGGCCGATTTTTGGACTTCAGATACGCTAGAATTTAAAGGGTCTCTTACACC
TGCTGCGGCCTGCAGGGACCCCTAGAACTTGCCGCCTACTTGTCTCAGTCTAATAACGCGCGAAGCCGTGGGGCA
CGTGACCTTAAGTCGCAGAGCGAGTGATGAATTTGGGACGCTAATATGGGTGAATAGAGACTTATATCATCAGGG

gene

“Genome: bought the book; hard to read.” –Eric Lander

gene

regulatory region (e.g. enhancer)

Some more terminology and statistics

- 3 billion (3b) bases in human DNA
- A **variant** is a "base" that is different between individuals
- ~50% from each parent
- ~60 "new/acquired" variants compared to parents[1]
- ~0.6% different between individuals from same population[2]

[1] https://www.nature.com/articles/ng.862 (2011)
[2] https://www.nature.com/articles/nature15393 (2015)

Alleles

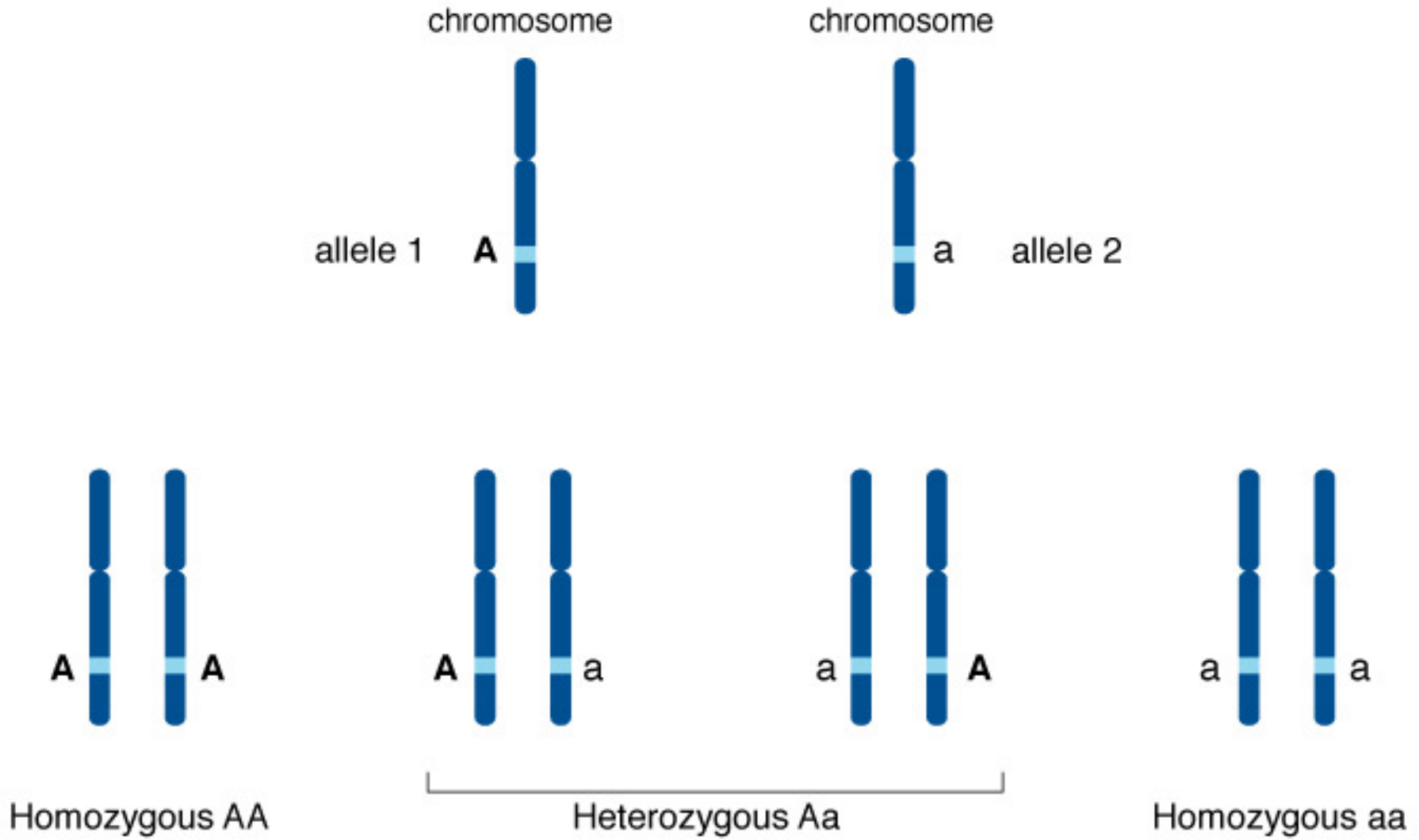


image source: https://rarediseases.info.nih.gov/GlossaryDescription/9/0

Review: SNPs vs SNVs

- **Common** vs **rare** variants

- Single nucleotide polymorphisms (**SNPs**) vs single nucleotide variants (**SNVs**)
  - Common: more than 1% in population
  - Rare: less than 1% in population

- Other types of variation: structural variation, insertion/deletion, duplication

Outline

- Terminology and fundamentals
- Human genetics and disease
- GWAS – genome-wide association studies
  - Statistical testing: chi-square test/logistic regression

Genetic studies

- **Cross-species**: which genetic differences result in observed differences between different species (e.g. phylogenetic)

- **Within-species**: which genetic differences result in observed differences within a species (e.g. traits/phenotypes)

- Today: a particular type of within-species comparisons called Genome-wide association studies

How does genetic disease arise?

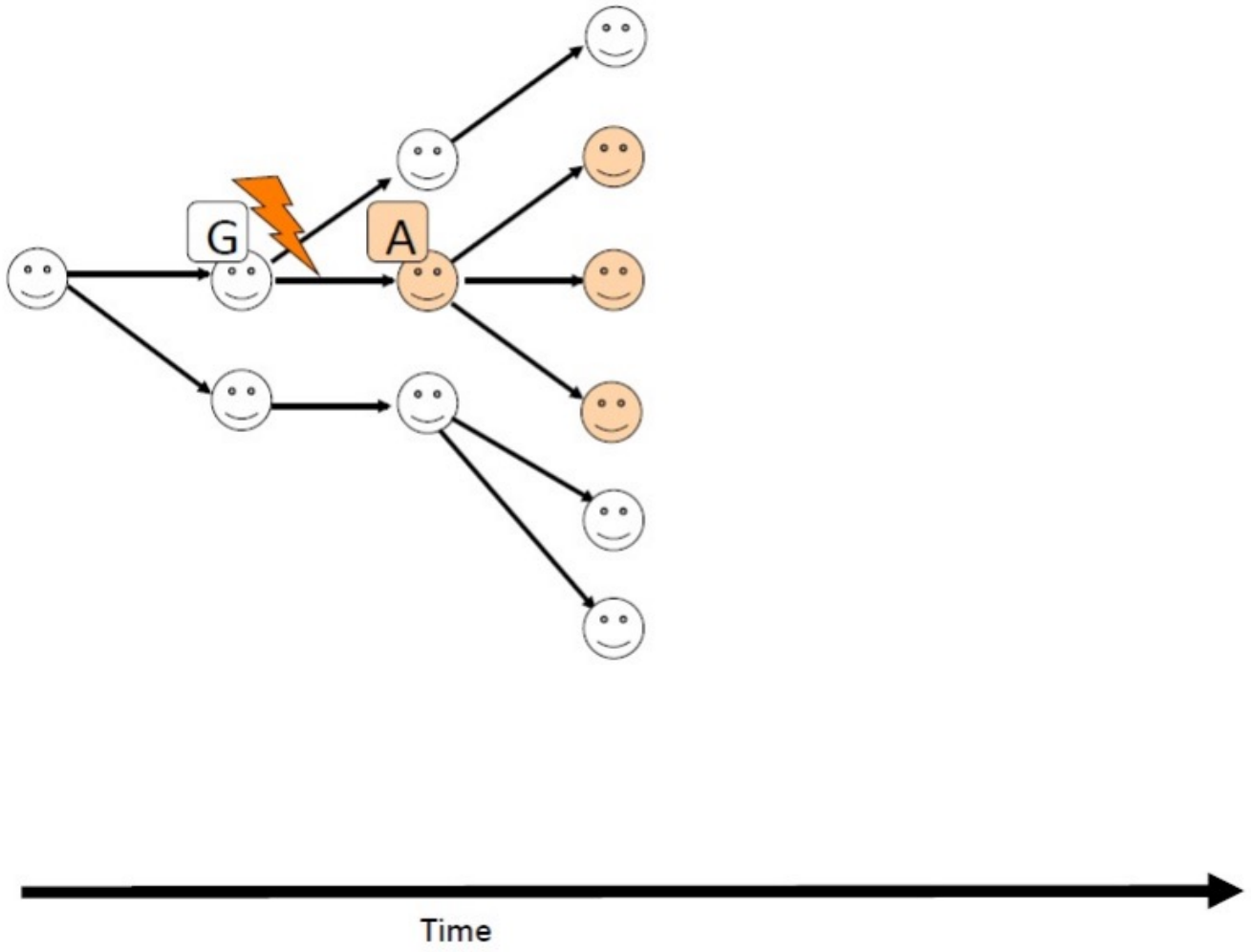


Image source: Itsik Pe'er, Columbia

How does genetic disease arise?

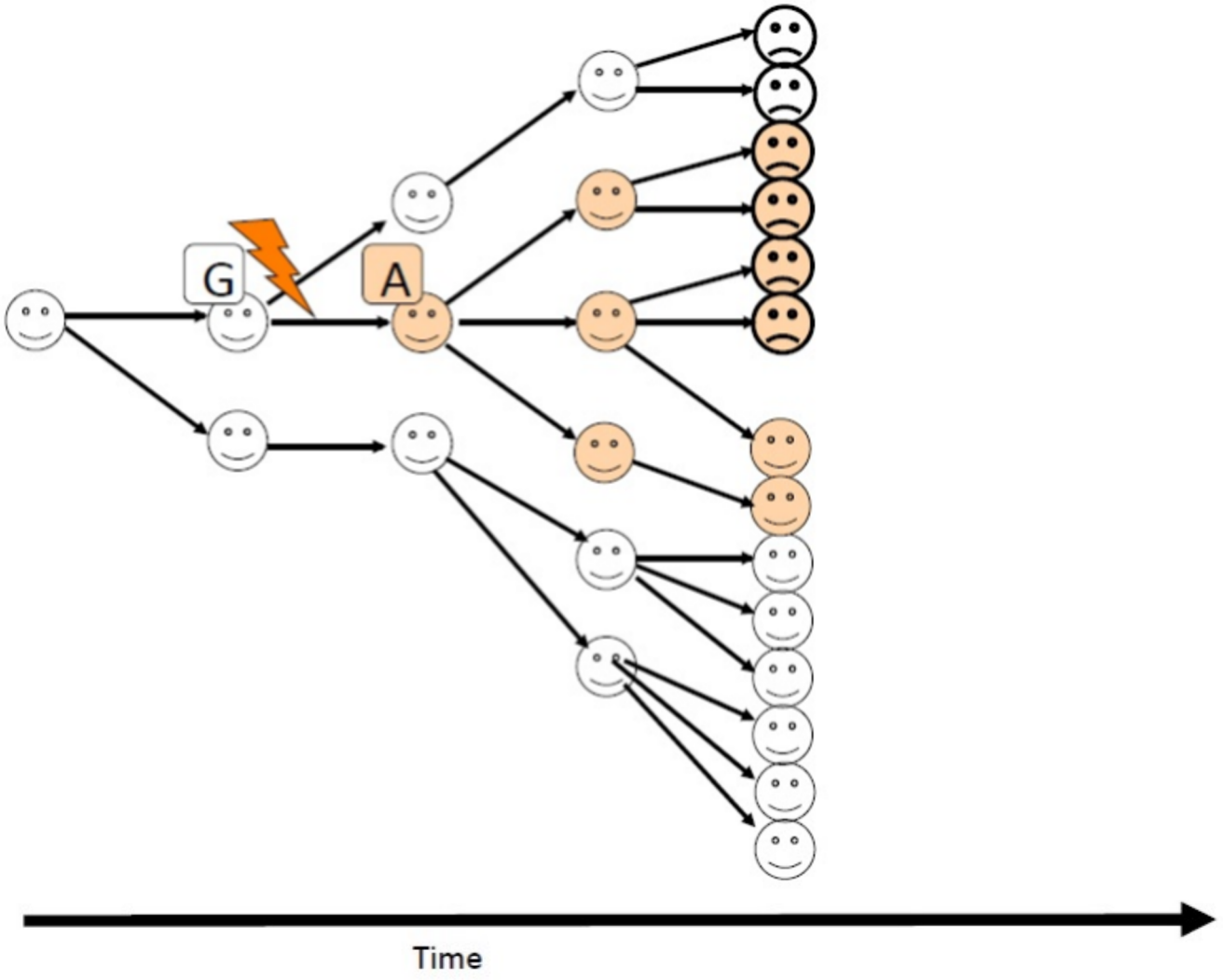
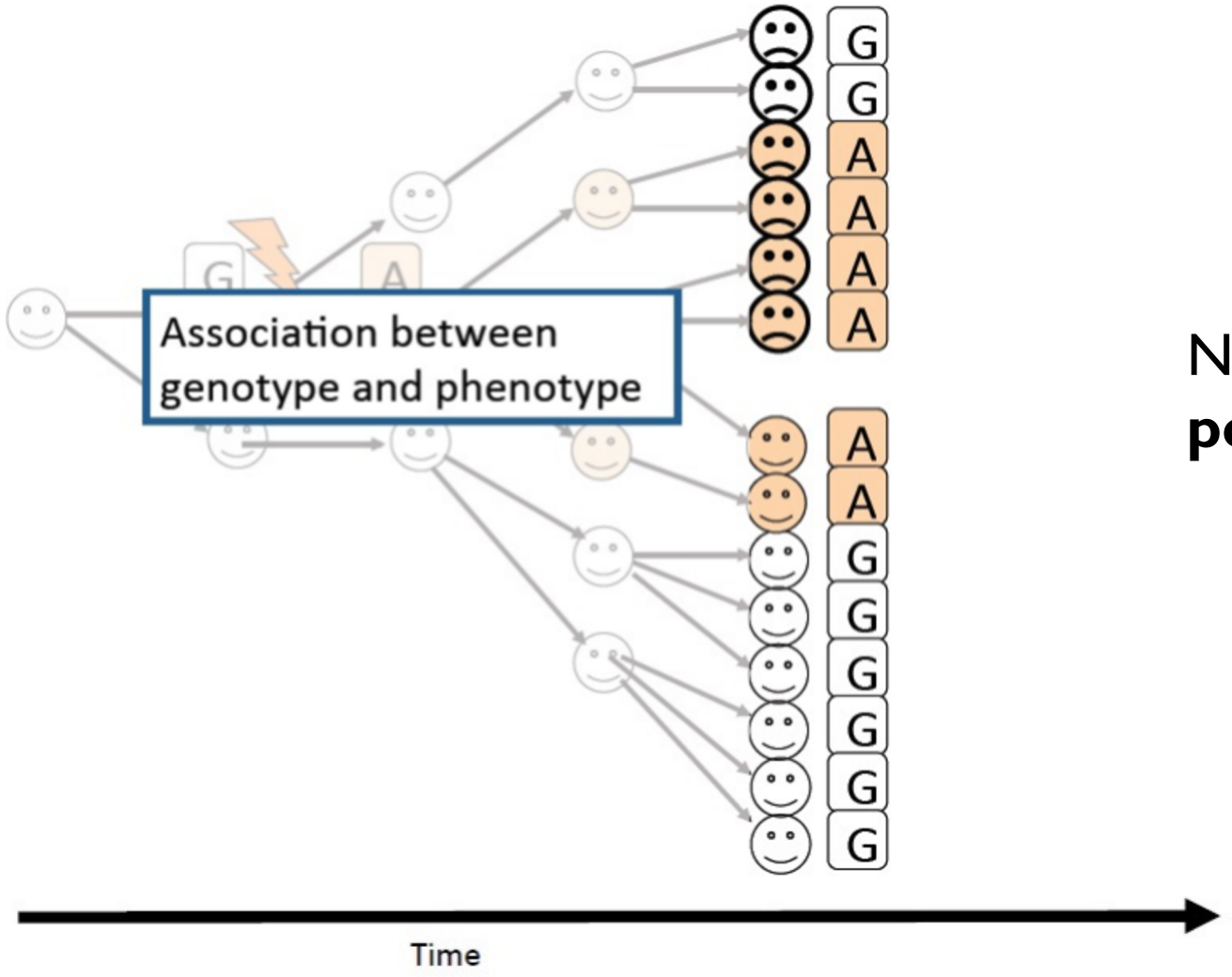


Image source: Itsik Pe'er, Columbia

How does genetic disease arise?



Note: not perfect **penetrance**

Image source: Itsik Pe'er, Columbia

Terminology so far:

- Base
- Variant
- SNP
- SNV
- Allele
- Genotype
- Penetrance

Heritable disease/phenotype/traits

- Some examples:
  - TID
  - Schizophrenia
  - Multiple Sclerosis
  - Sickle Cell Anemia
  - Tay–Sachs disease

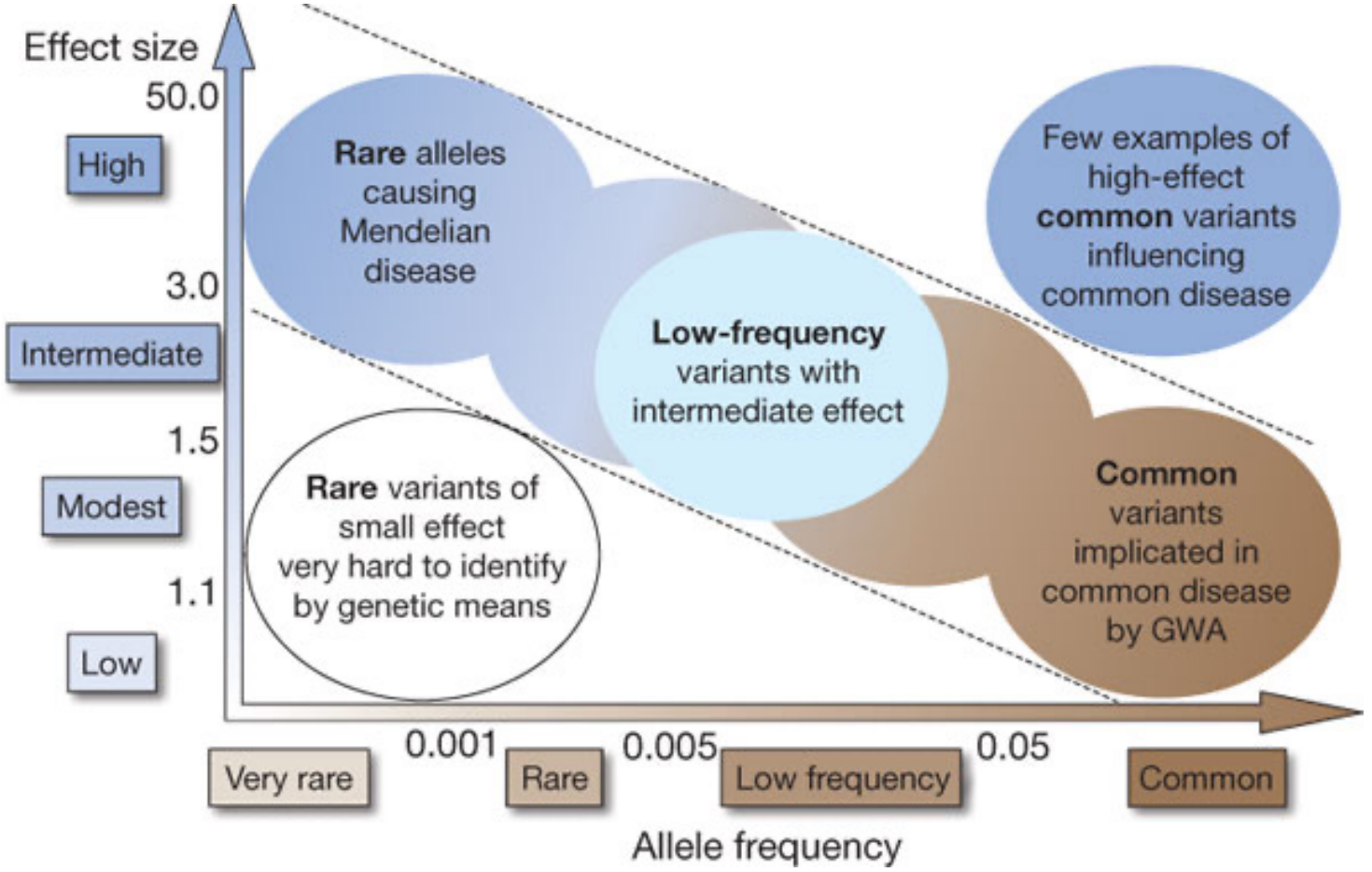
Common disease-common variant hypothesis

- **Rare disease**: very rare variants ("mutations") in one or a very small number of loci/genes.

- **Common disease**: influenced by variants that are common in the population (e.g., SNPs)
  - Also implies that the **effect size** for each variant should be small and that multiple variants must be at play

- The same study design will not be successful for both types of diseases
  - Implies that family studies unlikely to be successful and need to assess disease association in population studies

Common disease-common variant hypothesis



Genetic study considerations:

- Feasibility of identifying genetic variants by risk allele
- Frequency and strength of genetic effect (odds ratio)

image source: TA Manolio *et al. Nature* **461**, 747-753 (2009) doi:10.1038/nature08494

Heritable human traits

- **Rare/simple ("Mendelian") disease**
  - One or small number of loci
  - Inheritance pattern is clear (e.g. recessive loss/gain of function, autosomal dominant, X-linked)
  - Qualitative/discrete differences

- **Complex/common disease**
  - Multiple genetic factors contribute
  - Mendelian ratios are not applicable; need a different method for studying them
  - Environmental factors are also at play
  - Genetic inheritance is not simple to understand/predict

Why do we want to look at genetics vs gene expression?

Outline

- Terminology and fundamentals
- Human genetics and disease
- GWAS – genome-wide association studies
  - Statistical testing: chi-square test/logistic regression

What is a GWAS?

- **Genome-Wide Association Study** – study interrogating the relationship between genome-wide genetic variation and a phenotype.
- GWAS: experimental design + type of data
- Characteristics
  - Case control study design
  - Large number of samples (n >10K)
  - Standardized QC, and Imputation
  - Standardized statistical tests and multiple testing correction
  - Much of the data is 'negative'

A little bit of history

- 2001: a draft of the human genome sequence becomes available
- 2001: The international SNP Map Working Group publishes a SNP Map of 1.42 million SNPs that contained all SNPs identified so far
- 2005: HapMap project Phase I starts:
  - Genotype at least one common SNP (MFAF>5%) every 5Kb across 270 individuals
  - Geographical diversity
  - 1.3 million SNPs
- 2012: 1,000 Genomes project completed
- 2018: 100,000 Genomes project completed

The "$1,000 genome" is here

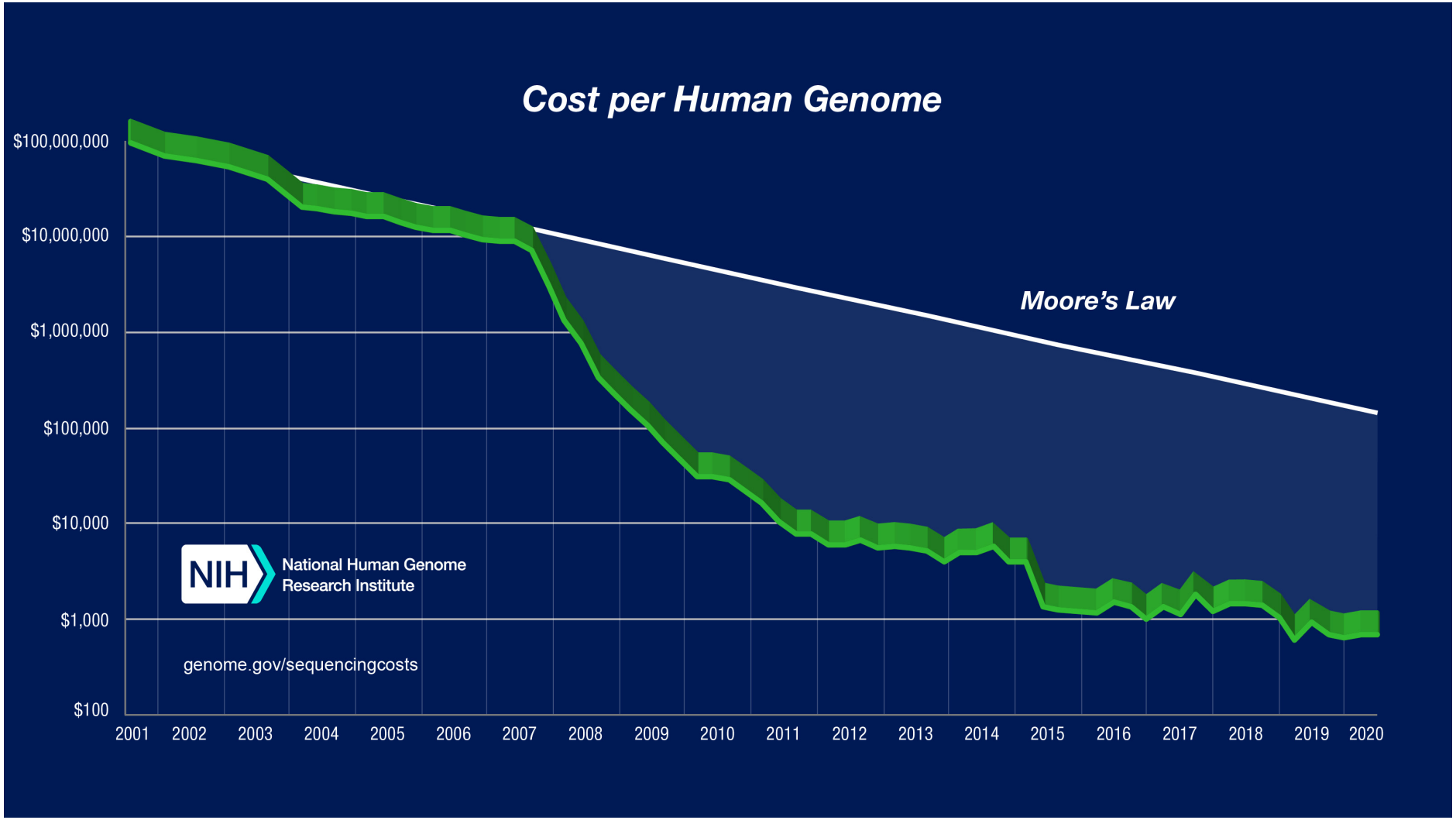


image source: www.genome.gov/sequencingcosts

Primary tool of GWAS: Genotyping arrays

- Platforms: Illumina and Affymetrix
- ~1M – 5M "genotyped" SNPs per array

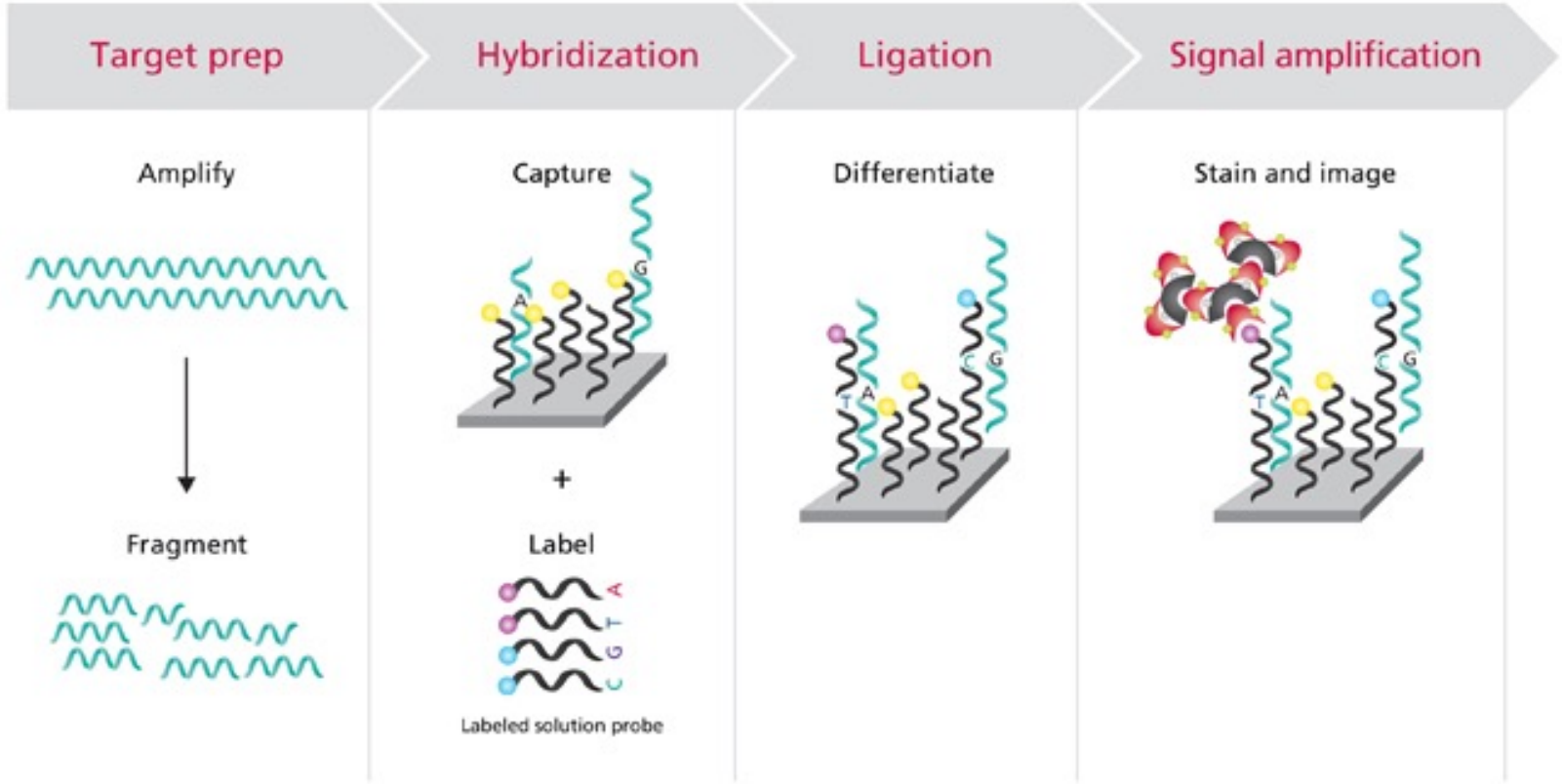
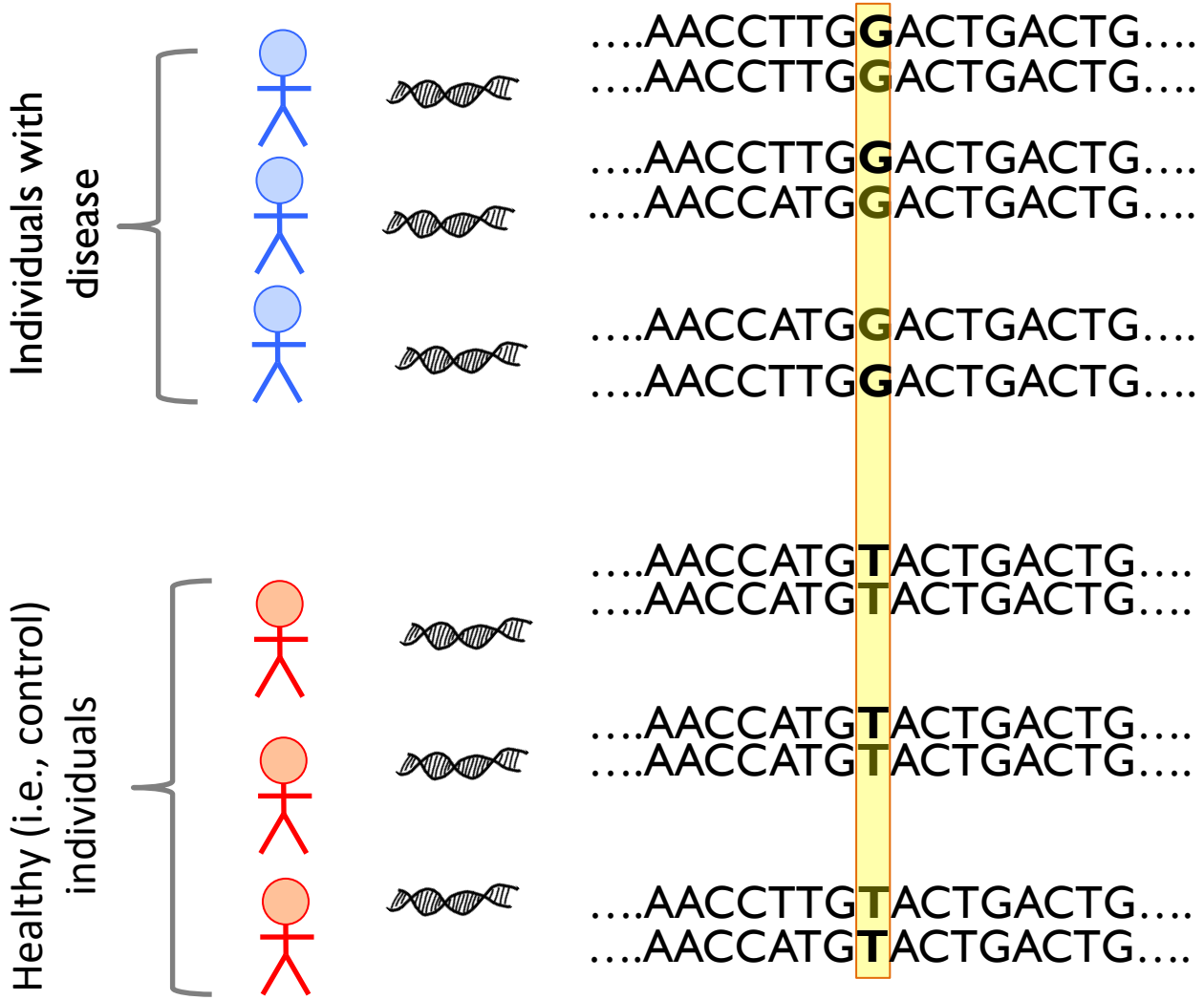


image source: https://www.affymetrix.com/products_services/axiom_custom/axiom_mydesign.affx

Welcome to the era of GWAS!



Genotyping:
measure a large set of pre-determined SNPs

GWAS studies' amazing success since 2005

- Prior to invention of GWAS, there was very little success in identifying genetic sources of common heritable disease
- GWAS studies have identified hundreds to thousands of genetic risk loci for complex diseases like SCZ, MDD, T2D

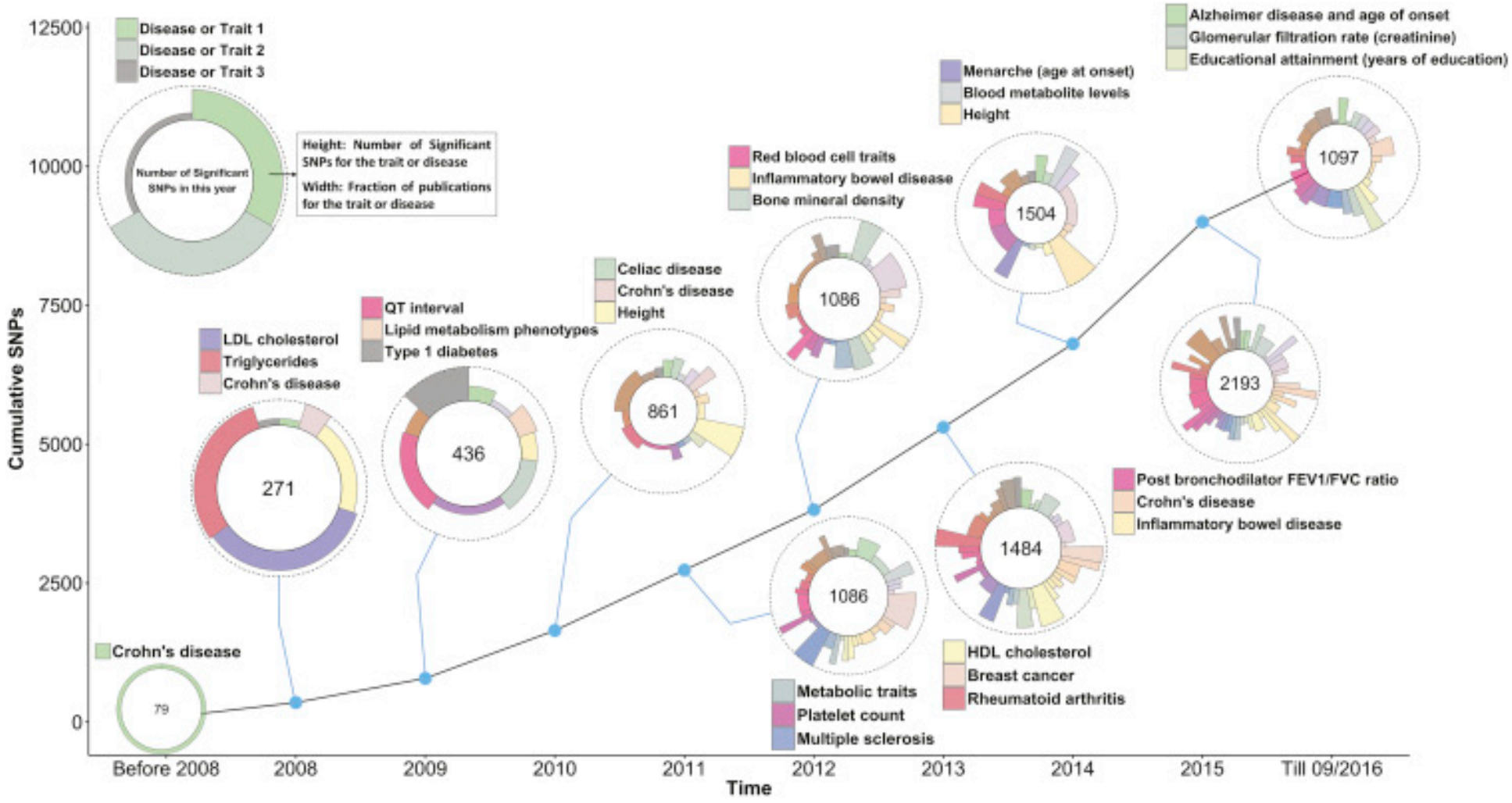

AJHG
Volume 101, Issue 1, 6 July 2017, Pages 5-22
Cell PRESS

Review

10 Years of GWAS Discovery: Biology, Function, and Translation

Peter M. Visscher [1, 2], Naomi R. Wray [1, 2], Qian Zhang [1], Pamela Sklar [3], Mark I. McCarthy [4, 5, 6], Matthew A. Brown [7], Jian Yang [1, 2]

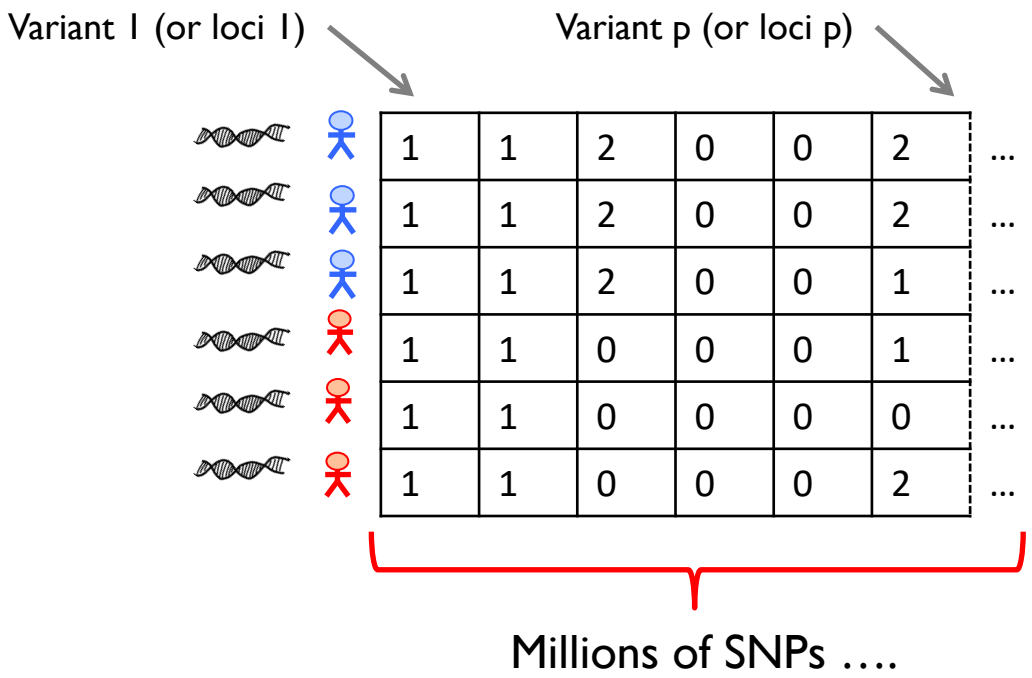
GWAS studies' amazing success since 2005



In Practice

Most SNPs are bialleleic (major and minor allele)

Data representation: count the number of minor alleles at each site



a) Identifying susceptibility SNPs

b) Risk prediction

Out of scope

- QC process
  - Sample quality
  - Sex inconsistency
  - HW concordance
  - Base calling quality
  - Genomic inflation/population stratification (later)

- Imputation of additional SNPs
  - LD: linkage disequilibrium
  - Typically impute ~5M additional SNPs

UNIT

Quality Control Procedures for Genome-Wide Association Studies

Stephen Turner, Loren L. Armstrong, Yuki Bradford, Christopher S. Carlson, Dana C. Crawford, Andrew T. Crenshaw, Mariza de Andrade, Kimberly F. Doheny, Jonathan L. Haines ... **See all authors**

Published: 22 July 2012

Fast and accurate genotype imputation in genome-wide association studies through pre-phasing

Bryan Howie, Christian Fuchsberger, Matthew Stephens, Jonathan Marchini & Gonçalo R Abecasis

Nature Genetics **44**, 955–959(2012) | Cite this article

2214 Accesses | **1028** Citations | **24** Altmetric | Metrics

Study design and statistics in GWAS

Study design:

- Case/control design (binary disease/healthy outcome)
- Some studies also look at continuous outcomes (e.g. BMI, height)
- Genotype at several million SNPs

Statistics:

- Chi-Square test
- Logistic regression to compute log odds
- One SNP at a time

Why not just use ANOVA/linear regression?

- Linear regression models E(Y) as a **linear** function X
  - What is Y (response) in a case/control GWAS?

- Recall the **assumptions**:
  - normality
  - constant variance

Contingency tables

- We are interested in the relationship between two categorical variables with *k* and *m* levels
- We want to know whether the probability of one being in case/control class is statistically dependent on the allele
- **Contingency table**: count the number of subjects in combination of levels from each category

	Case	Control
Minor allele (Allele A)	a	b
Major allele (Allele G)	c	d

Contingency tables

- We are interested in the relationship between two categorical variables with *k* and *m* levels
- We want to know whether the probability of one being in case/control class is statistically dependent on the allele
- **Contingency table**: count the number of subjects in combination of levels from each category

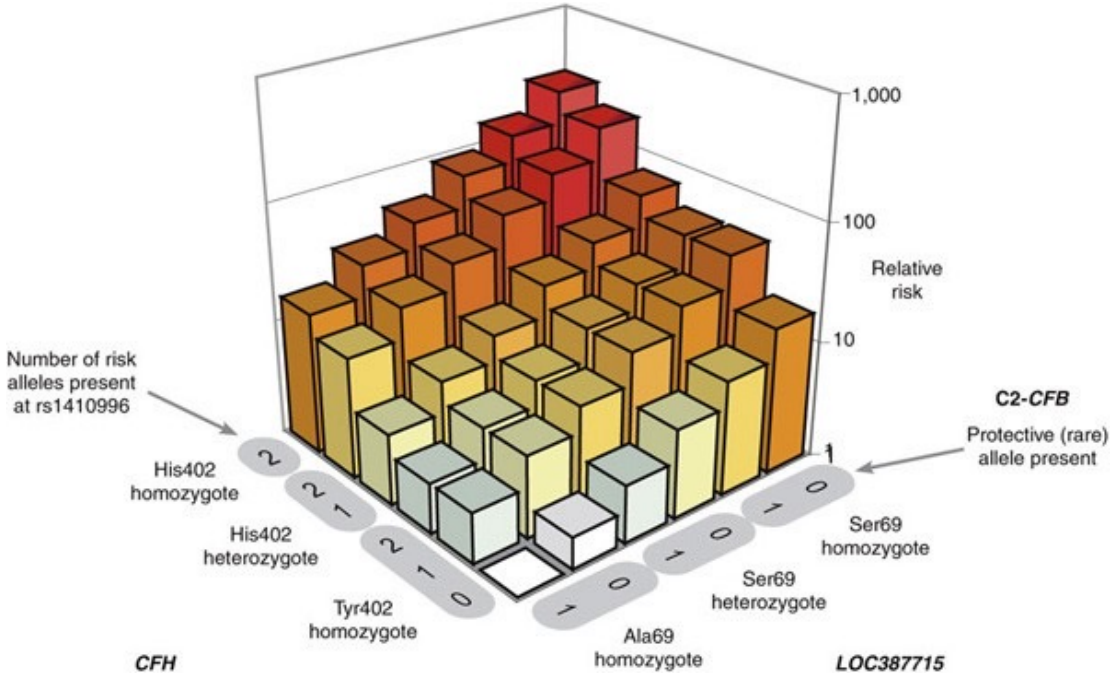
	Case	Control	**Row total**
Minor allele (A)	a	b	**a + b**
Major allele (G)	c	d	**c + d**
Column total	**a + c**	**b + d**	**a + b + c + d**

Example: Age related macular degeneration

image source: neoretina.com/blog/a-simple-guide-to-understanding-age-related-macular-degeneration/

Relative risk plotted as a function of the genetic load of the 5 variants that influence the risk of AMD

Maller et al. (2006): https://www.nature.com/articles/ng1873



- 2004: Little known about the cause of AMD
- 2006: GWAS discovers 3 genes (5 common variants) that explain >50% of risk

Example: contingency table from AMD GWAS

	Case	Control	**Row total**
Allele A	1522	670	**2192**
Allele G	954	1198	**2152**
Column total	**2476**	**1868**	**4344**

The probabilities of each alleles in cases and controls looks different:

P(A | Case) =1522 / 2476 ~ 61%

P(A | Control) = 670 / 1868 ~ 35%

Contingency tables- Chi-Square test

	Case	Control	**Row total**
Allele A	1522	670	**2192**
Allele G	954	1198	**2152**
Column total	**2476**	**1868**	**4344**

Chi-square (χ^2) statistic: sum of squared differences between observed value in each cell and its expected value

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(\text{Observed}_{ij} - \text{Expected}_{ij})^2}{\text{Expected}_{ij}}$$

Distribution of χ^2 statistic is known from theory: under the null, it follows a χ^2 distribution with (k-1)(m-1) df

Expected counts

Intuition: conditional on row and column totals, what do we expect under the null (i.e. if there's no association)?

	Case	Control	**Row total**
Allele A			**2192**
Allele G			**2152**
Column total	**2476**	**1868**	**4344**

2192 / 4344 = 50%

2476 / 4344 = 57%

If 57% of samples are Cases, and 50% of samples have allele A, how many Cases do we expect to have allele A under the null?

Expected counts

	Case	Control	**Row total**
Allele A	R1 x C1 / N	R1 x C2 / N	**R1 = 2192**
Allele G	R2 x C1 / N	R2 x C2 / N	**R2 = 2152**
Column total	**C1 = 2476**	**C2 = 1868**	**N = 4344**

Expected counts

Observed	Case	Control	**Row total**
Allele A	1522	670	**2192**
Allele G	954	1198	**2152**
Column total	**2476**	**1868**	**4344**

Expected	Case	Control	**Row total**
Allele A	1249	943	**2192**
Allele G	1227	925	**2152**
Column total	**2476**	**1868**	**4344**

Chi-square statistic

Observed	Case	Control
Allele A	1522	670
Allele G	954	1198

Expected	Case	Control
Allele A	1249	943
Allele G	1227	925

$$\chi^2 = \sum_{i=1}^{k}\sum_{j=1}^{m} \frac{(\text{Observed}_{ij} - \text{Expected}_{ij})^2}{\text{Expected}_{ij}}$$

$$\chi^2 = 279.2$$

$$df = (2 - 1) * (2 - 1) = 1$$

$$p - value = 1.2 \times 10^{-62}$$

Note that R returns `0` for `1-pchisq(279.2, 1)`
(See also `chisq.test()`)

Chi-square test

- Simple – one loci at a time test
- Rigid – doesn't naturally allow investigating more than two variables at a time…

Logistic regression

- A **Generalized Linear Model (GLM)** for a binary outcome (e.g. case-control)
- Allows for inclusion of quantitative and/or categorical variables, or (nonlinear) function of your variables (e.g., log(x))

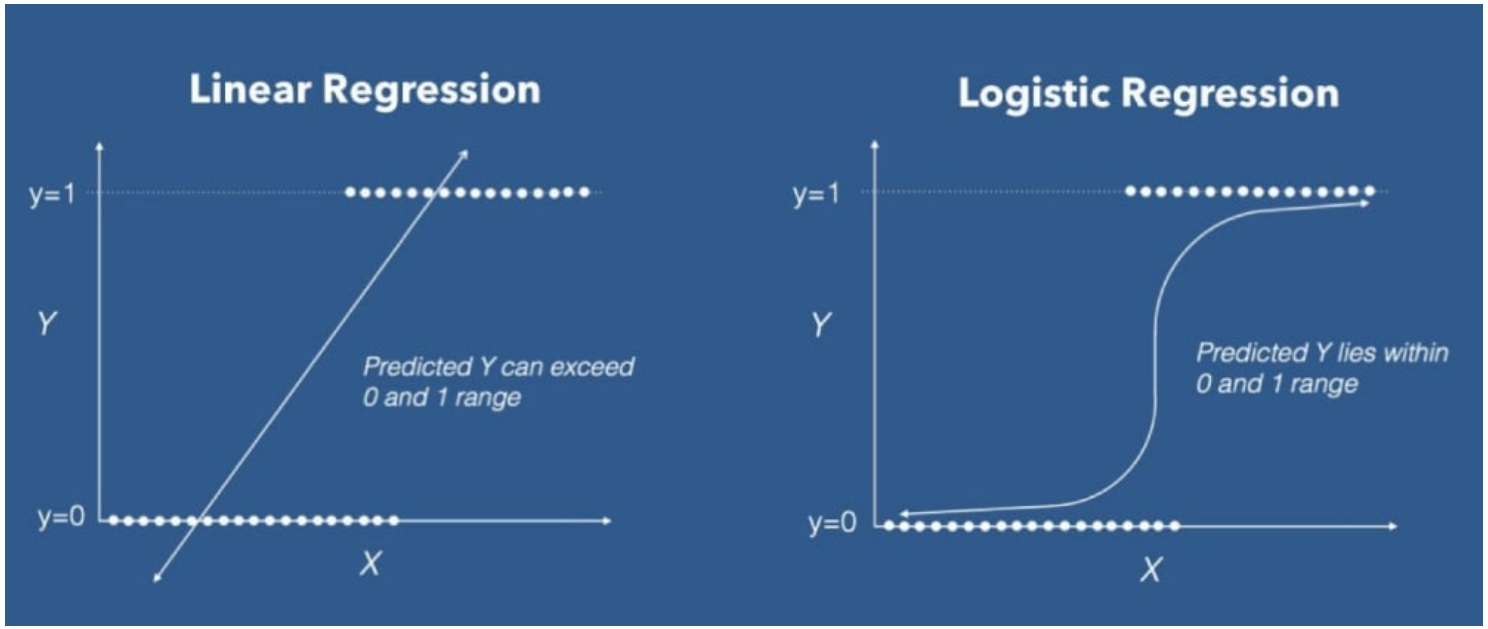


image source: https://dev.to/adityaberi8/logistic-regression-eg1

Logistic regression

Instead of modeling E(Y), we are modeling logit(E(Y)):

$$\log\left(\frac{E(Y)}{1-E(Y)}\right) = \log\left(\frac{p(Y=1)}{1-p(Y=1)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

logit = Log odds of "success"

Odds of "success"

- **Odds**: ratio of probability of success to probability of failure

$$\text{odds}(Y = 1) = \frac{p(Y = 1)}{p(Y = 0)} = \frac{p(Y = 1)}{1 - p(Y = 1)}$$

- If odds = 1, probability of success and failure are **equal**
  - we model log odds because this quantity is **symmetric about 0**

- If success rate is 0.9, then odds = 0.9/0.1 = 9 (i.e. for every nine successes there is one failure)
  - If we observe 25% success rate, what's the odd of success?
  - What about for an 75% success rate?
  - What are the log odds for both of these cases?

Parameters in logistic regression

- Hypothesis tests: Wald test for individual parameters, Likelihood ratio test for nested models
- Linear combination of parameters gives the log odds of success
- Intercept: log odds of success when all explanatory variables are 0
- One unit change in explanatory variable x_p changes log odds by $\hat{\beta}_p$
  - Interpret parameters by converting back to odds or probability scale:

$$\text{odds}(Y=1) = \frac{p(Y=1)}{1-p(Y=1)} = e^{\hat{\beta}_0+\hat{\beta}_1 x_1 +\cdots+ \hat{\beta}_p x_p}$$

$$p(Y=1) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 x_1 +\cdots+ \hat{\beta}_p x_p}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 x_1 +\cdots+ \hat{\beta}_p x_p}}$$

Fitting the logistic regression model

- Recall for OLS the parameters are the **Maximum Likelihood Estimates (MLE)**
  - Maximize the (normal) likelihood of the data given parameters
  - Equivalent to minimizing sum of squared errors
- Similarly for logistic regression, we want to maximize the likelihood of data given parameters (equivalent to maximizing log-likelihood)

$$L(\boldsymbol{Y}|\boldsymbol{x},\boldsymbol{\beta}) = \prod_{i=1}^{n} p(Y_i = 1)^{Y_i}(1 - p(Y_i = 1))^{1-Y_i}$$

$$\ell(\boldsymbol{Y}|\boldsymbol{x},\boldsymbol{\beta}) = \sum_{i=1}^{n} (Y_i \log\big(p(Y_i = 1)\big) + (1 - Y_i)\log(1 - p(Y_i = 1)))$$

Fitting the logistic regression model

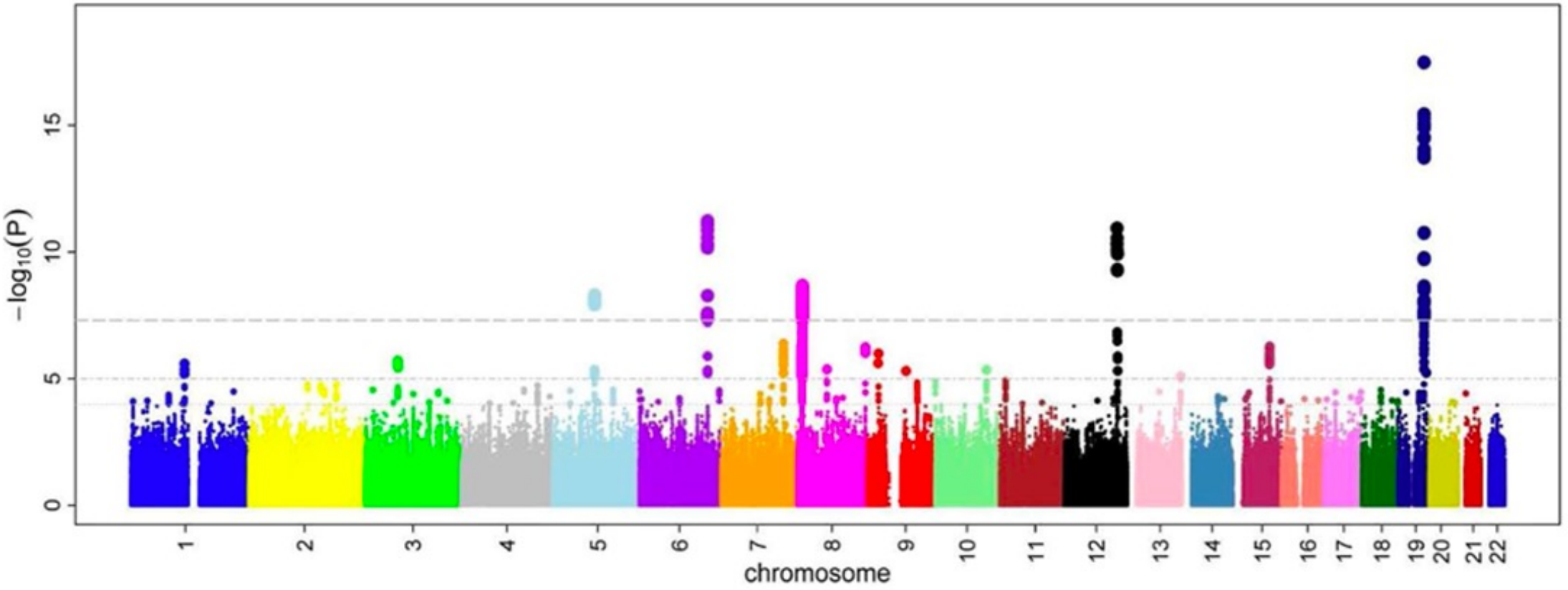
Continued... (plugging in $\mathrm{p}(Y_i = 1) = \frac{e^{\beta_0+\beta_1 x_{i1} \ +\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1} \ +\cdots+\beta_p x_{ip}}}$)

$$\ell(\boldsymbol{Y}|\boldsymbol{x},\boldsymbol{\beta}) = \sum_{i=1}^{n} Y_i \log\left(\frac{e^{\beta_0+\beta_1 x_{i1} \ +\cdots+\beta_p x_{ip}}}{1+e^{\beta_0+\beta_1 x_{i1} \ +\cdots+\beta_p x_{ip}}}\right) + \sum_{i=1}^{n} (1-Y_i)\log\left(\frac{1}{1+e^{\beta_0+\beta_1 x_{i1} \ +\cdots+\beta_p x_{ip}}}\right)$$

$$= \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 x_{i1} \ + \cdots + \beta_p x_{ip}) - \sum_{i=1}^{n} \log\left(1+e^{\beta_0+\beta_1 x_{i1} \ +\cdots+\beta_p x_{ip}}\right)$$

Solving for $\widehat{\boldsymbol{\beta}}$ that maximizes $\ell(\boldsymbol{Y}|\boldsymbol{x},\boldsymbol{\beta})$ requires numerical methods (e.g., coordinate descent)

Manhattan plot



Significance in GWAS

- For determining which SNPs are "significant", it is conventional to assess **genome-wide significance**
- Typically, multiple correction is done with Bonferroni adjustment
  - recall from lecture 9 that this controls the FWER
  - recall that this is **conservative** (prob at least one error)
- Convention to use a p-value threshold of 5×10^{-8} (corresponds to Bonferroni adjustment for 1M tests)
  - corresponds to estimates of the number of **independent** SNPs in human
- Recent research explores the use of FDR correction in GWAS
  - e.g. Brzyski et al. 2017

Summary of GWAS

- GWAS typically implies a study design and data type
  - Most commonly case-control and genotyping at several million SNPs
- Standardized steps for QC, and statistical analysis
  - Chi-square test/logistic regression
  - One SNP at a time analysis
  - Multiple testing correction (Bonferroni is the standard)
- Results: summary statistics; visualized with Manhattan plots
- Guest lectures next week:
  - Considering multiple genes at a time (polygenic risk scores)
  - Phenome-wide association studies
  - Causal inference