GWAS – and next steps

Genome-wide association studies for complex traits: consensus, uncertainty and challenges

Mark I. McCarthy[*‡], *Gonçalo R. Abecasis*[§], *Lon R. Cardon*[*‖], *David B. Goldstein*[¶], *Julian Little*[#], *John P. A. Ioannidis*[**‡‡] *and Joel N. Hirschhorn*[§§‖‖¶¶]

GWAS

- Genome Wide Association Study
  - ascertain genotypes on some number of cases (as many as you can get) and some number of controls (as many as you can get)
  - for each locus in the genome test whether the alleles associate with disease state

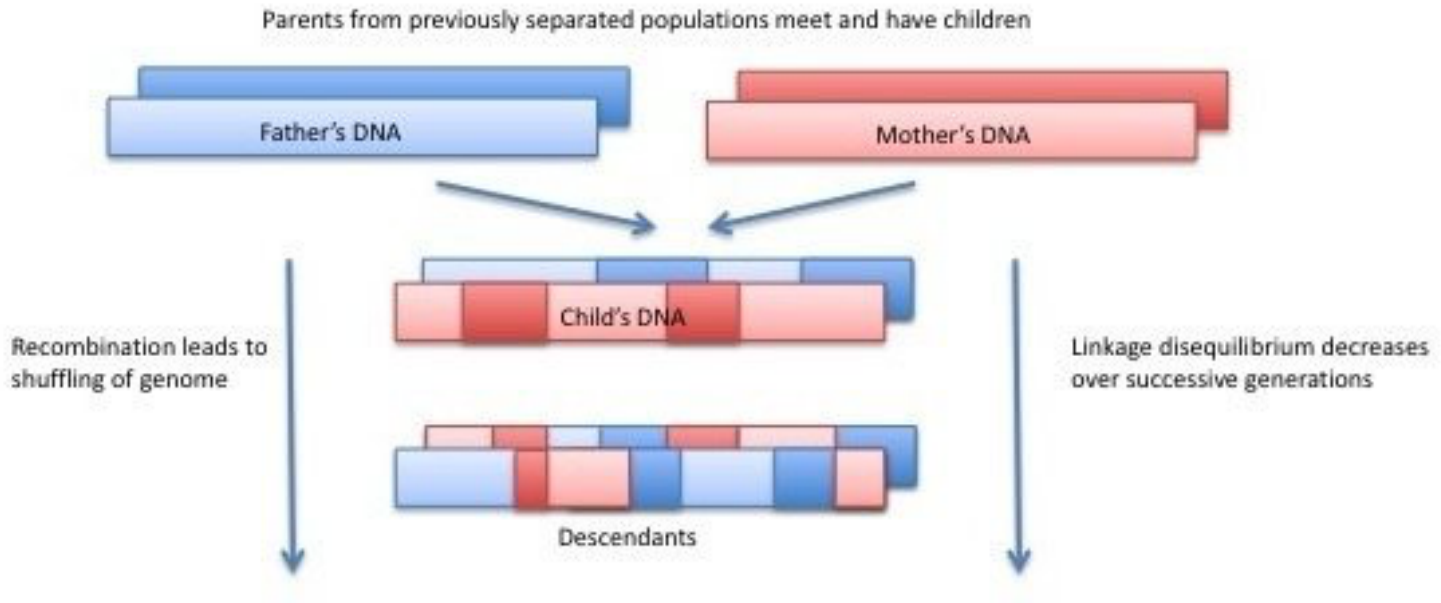
Genome-wide association studies for complex traits: consensus, uncertainty and challenges

Mark I. McCarthy[*‡], *Gonçalo R. Abecasis*[§], *Lon R. Cardon*[*||], *David B. Goldstein*[¶], *Julian Little*[#], *John P. A. Ioannidis*[**‡‡] *and Joel N. Hirschhorn*[§§||||¶¶]

Allele Frequency and Penetrance

- alleles vary in their frequency in different populations
  - many alleles have the same frequency in different populations, but some can vary substantially
  - have a look at dbSNP
- penetrance is a measurement of how likely it is, that a person with the risk allele will get the disease
  - highly penetrant alleles are ones where almost all people with the risk allele have the disease
  - CFTR – homozygous deletion/LoF is highly penetrant (close to 100%)

Linkage Disequilibrium



https://www.quora.com/What-is-linkage-equilibrium

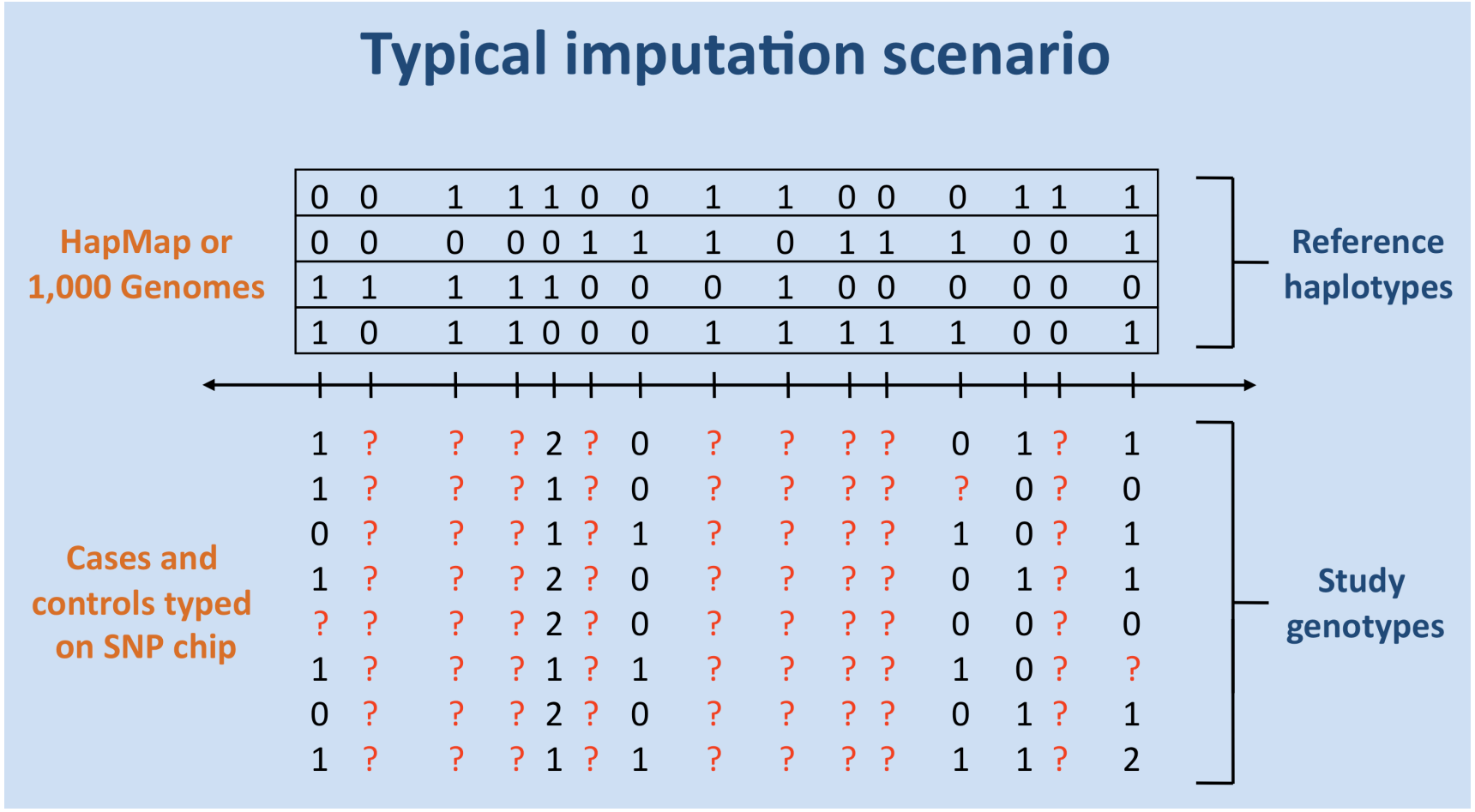
LD

- SNPs/variants that are close to each other tend to be highly correlated
- this is the basis for imputation (next slide)
- LD is smallest (the LD blocks are shortest) in African populations, and tend to be longer in more recent populations (eg Europeans)
- there are many measures (r-squared – just correlation, or D')

Genotyping

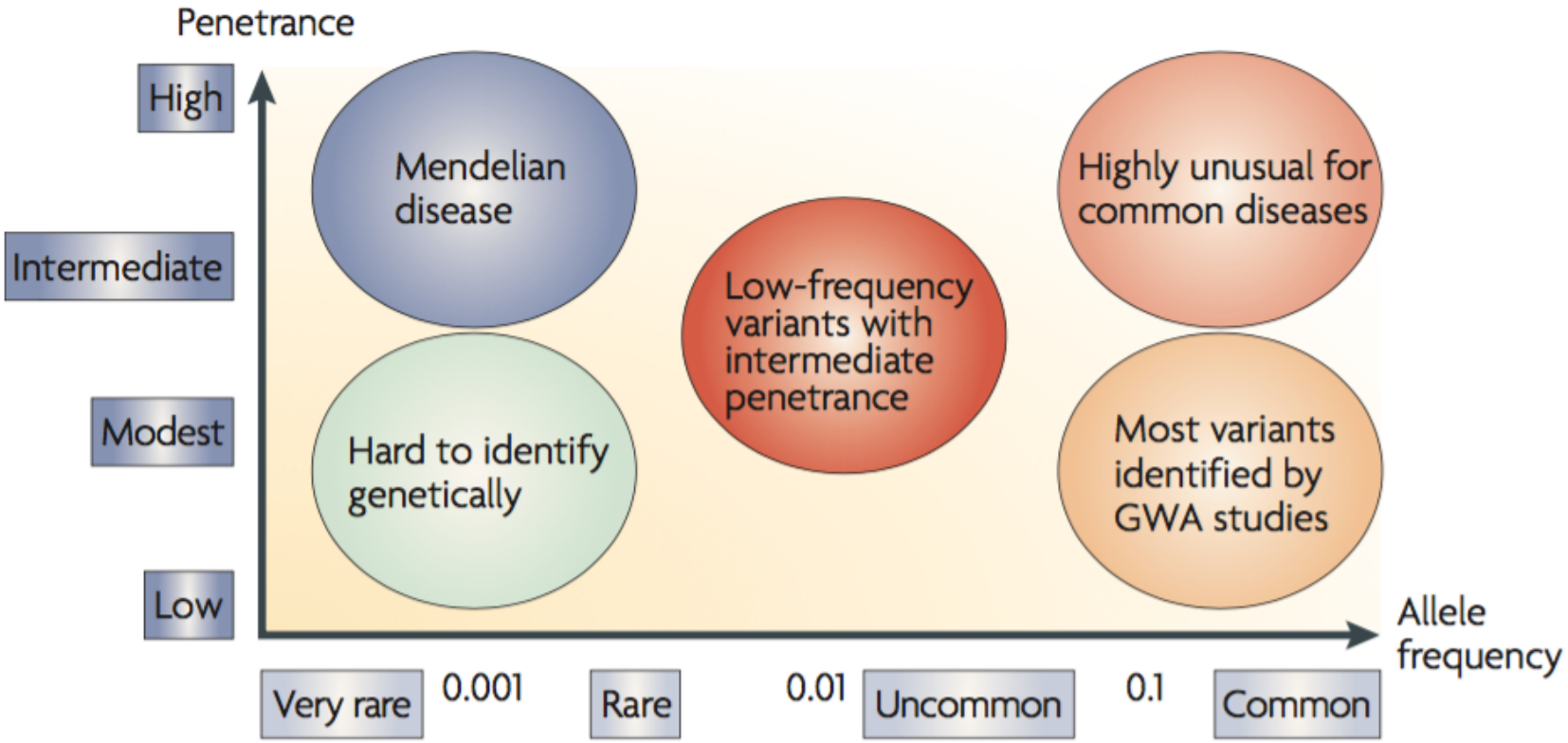
- most common approach is to use some form of genotyping array followed by imputation
- genotype imputation is generally very accurate (and the accuracy can be assessed)
- relies on the fact that we inherit DNA in fairly long blocks and the fact that there are relatively few mutations per live birth (estimates around 7)
- one generates an imputation panel by sequencing some number (UK10K, TopMed etc) of individuals then using those sequence data to impute the genotypes for people who you only know their

Imputation



http://mathgen.stats.ox.ac.uk/impute/impute_v2.html

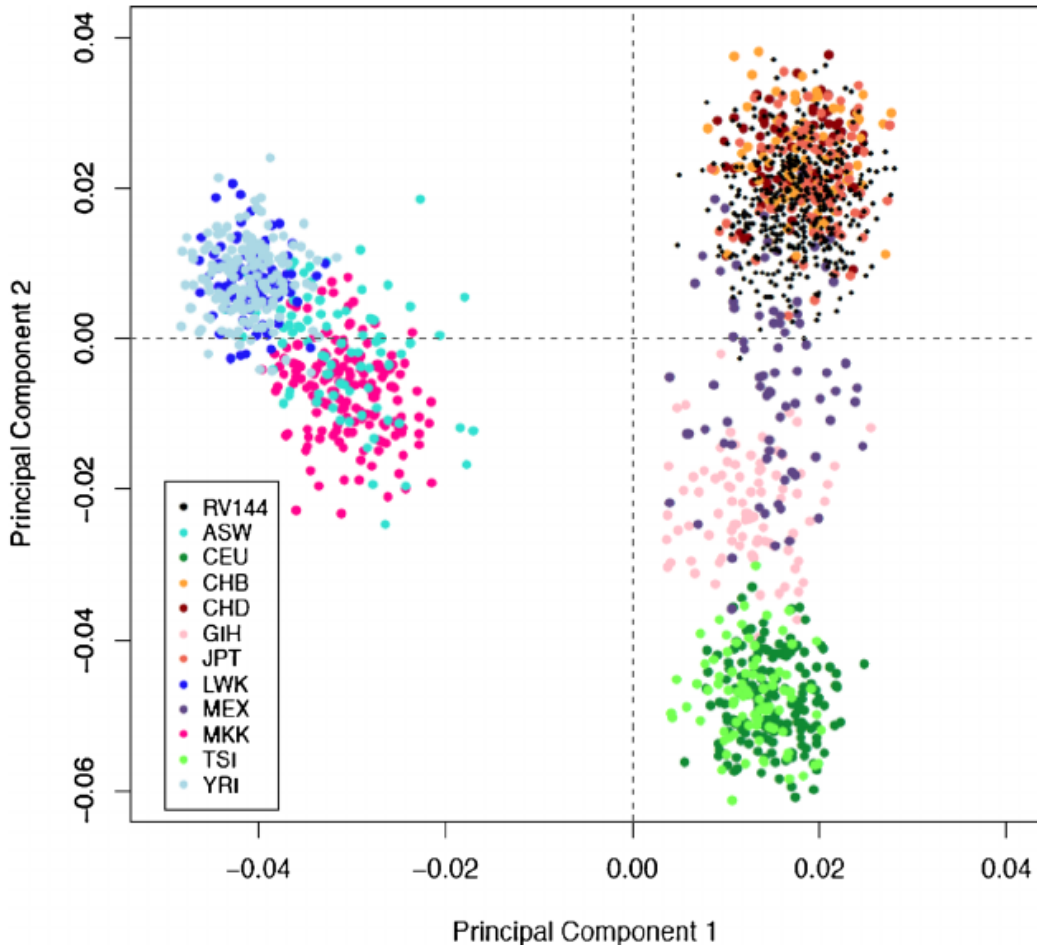
Variants and Effects



Population Structure

- different human populations have different genetic structure
  - alleles at different frequencies
- this can adversely effect GWAS analysis
  - typically we adjust for PCs
  - typically we carry out different GWAS for different populations

PCA plot of Genotypes



HLA class I, KIR, and genome-wide SNP diversity in the RV144 Thai phase 3 HIV vaccine clinical trial

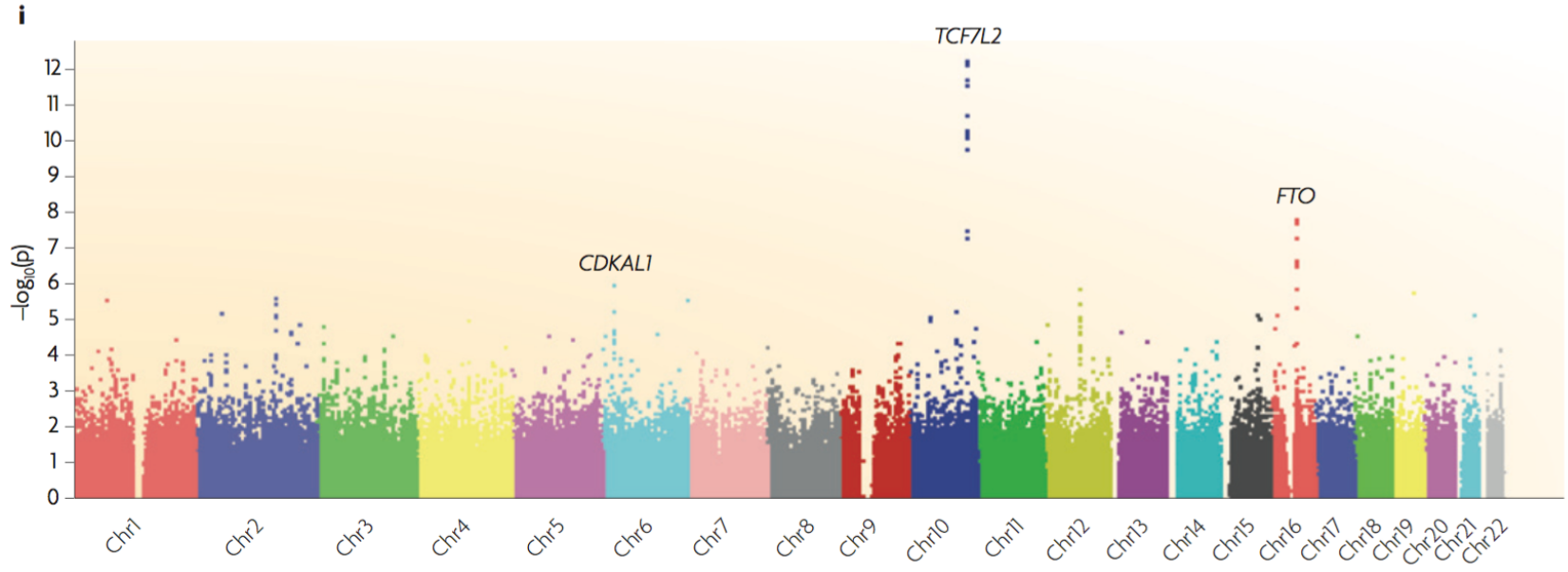
GWAS

- humans are diploid so at any locus (assuming bi-allelic) they can be AA, AB or BB (where A and B are one of ACTG)
  - often we use dose of the minor allele (0, 1, 2)
  - humans are tri-allelic and at some sites all four alleles are present, but here we assume only biallelic
  - for this lecture we will ignore the issues of structural variation (indels, inversions and so on)

GWAS

- typically we model the data using a logistic regression (disease/no disease)
- we include age, sex, some number of principal components (first 5?), chip/platform
- then we compute estimated ORs and p-values for each SNP (either the genotyped ones or the imputed ones)
- a very arbitrary value of 5x10-8 is used for genome wide significance
- finally a Manhatten plot is produced

Manhatten plot



Conditional Analysis

- the “usual practice”, until quite recently was to exclude the region around significant variants
- this was due to concerns about LD structure, but that does not seem to be a well founded concern
- now, once a hit is found, it is more common to simply add it as a covariate in the model and then repeat the GWAS computations, looking for additional hits
  - the additional hits tend to be lower frequency and LD is generally close to 0 (experience)

Some issues/concerns

- the power to detect a signal is:
  - a function of the prevalence of the disease (eg number of cases)
  - the minor allele frequency (closer to 1/2 more power)
  - the penetrance
- common alleles typically have small effects (ORs close to 1)

Interesting Bioc intro

- https://onlinelibrary.wiley.com/doi/full/10.1002/sim.6605

Functional Consequences

- a GWAS hit simply gives you a location in the genome that associates with the probability of disease
- that hit is rarely (at least with tools today) the functional or causal variant
- to go from a GWAS hit to a functional variant is referred to as the fine mapping problem

eQTL

- an eQTL is an expression quantitative trait locus
- essentially it is a genetic variant that associates with gene expression
- this is a potentially functional relationship
  - very low expression may be equivalent to loss of function
  - very high expression may be equivalent to gain of function

eQTL

- for eQTLs the response is gene expression in some units (hopefully biologically interpretable like TPM)
- GTex is a resource
- https://www.gtexportal.org/home/