Genetics 101

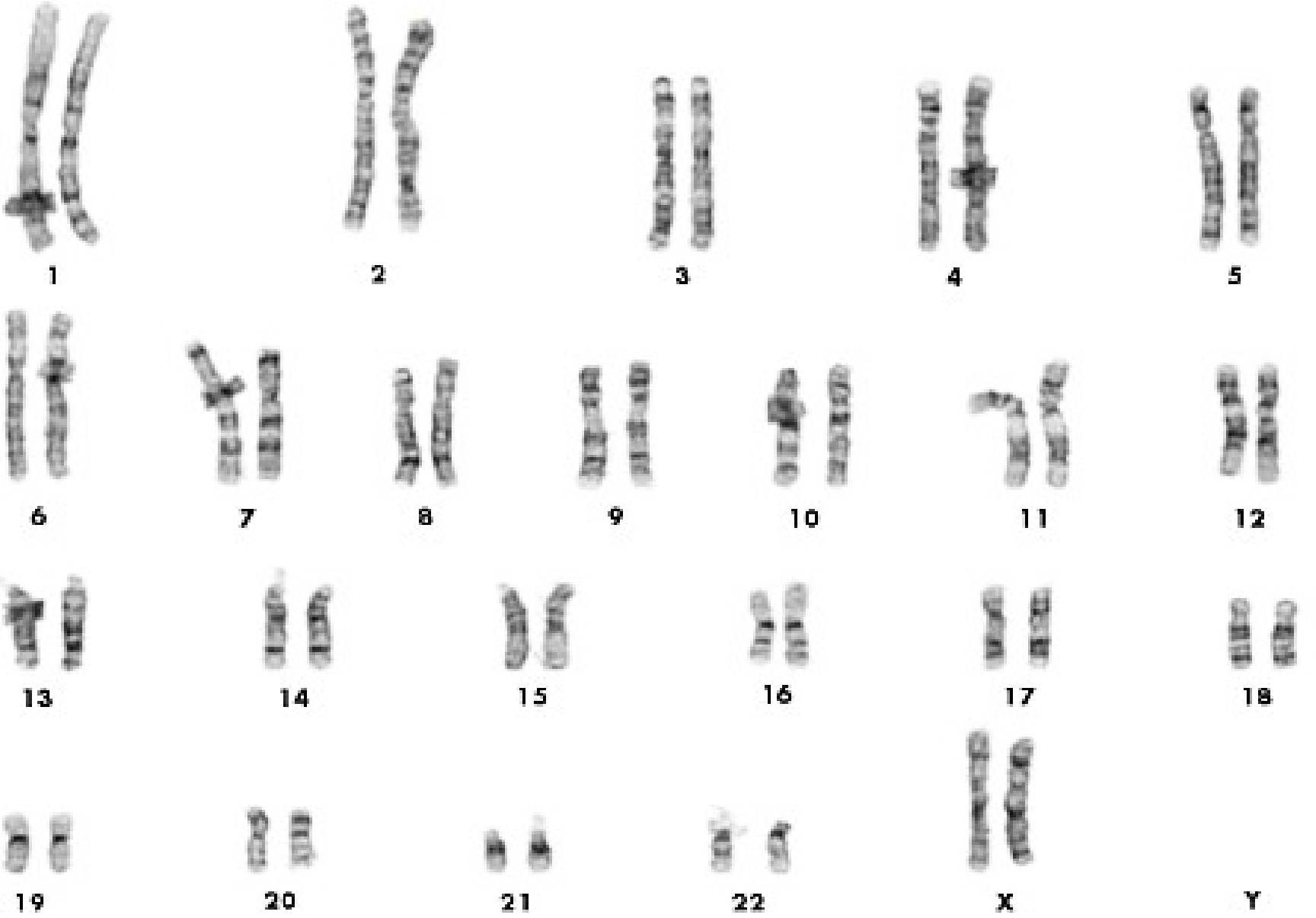
Tatiana Foroud, Ph.D.
Joe C. Christian Professor
Chair, Department of Medical and Molecular Genetics
Indiana University School of Medicine

Overview

- Basic terminology
- Autosomal dominant (early onset) AD
- APOE and late onset AD
- Genomewide association study (GWAS)
- Sequencing

Basic Genetics

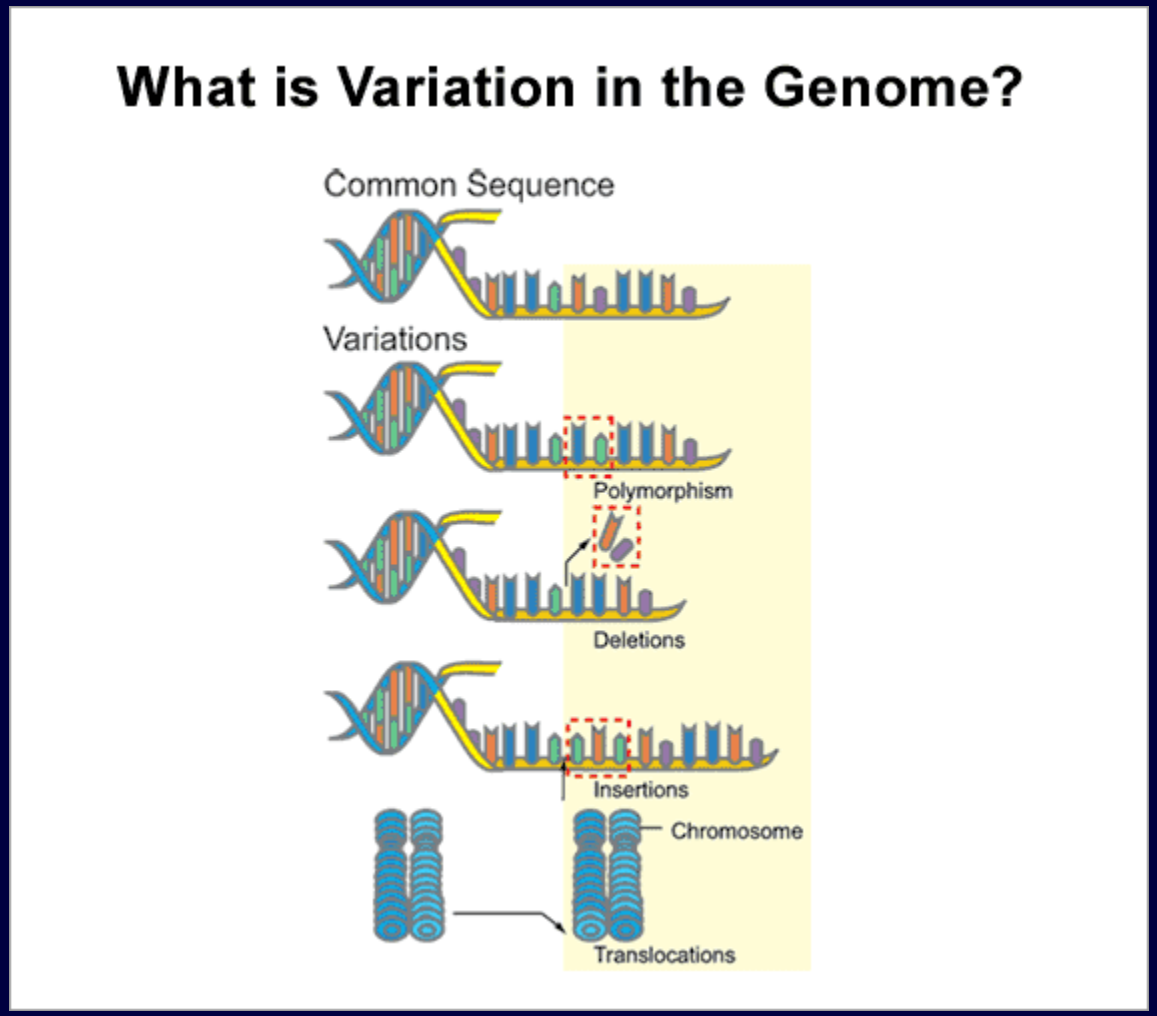
23 Pairs of Chromosomes



DNA Variation

- DNA is organized as a string of nucleotides (A, C, T and G)
- Changes in the DNA sequence can occur
  - Such changes might include replacing a nucleotide with another (A for G, or anything else)
  - Such changes can mean removing or adding in a nucleotide (called a deletion or an insertion)

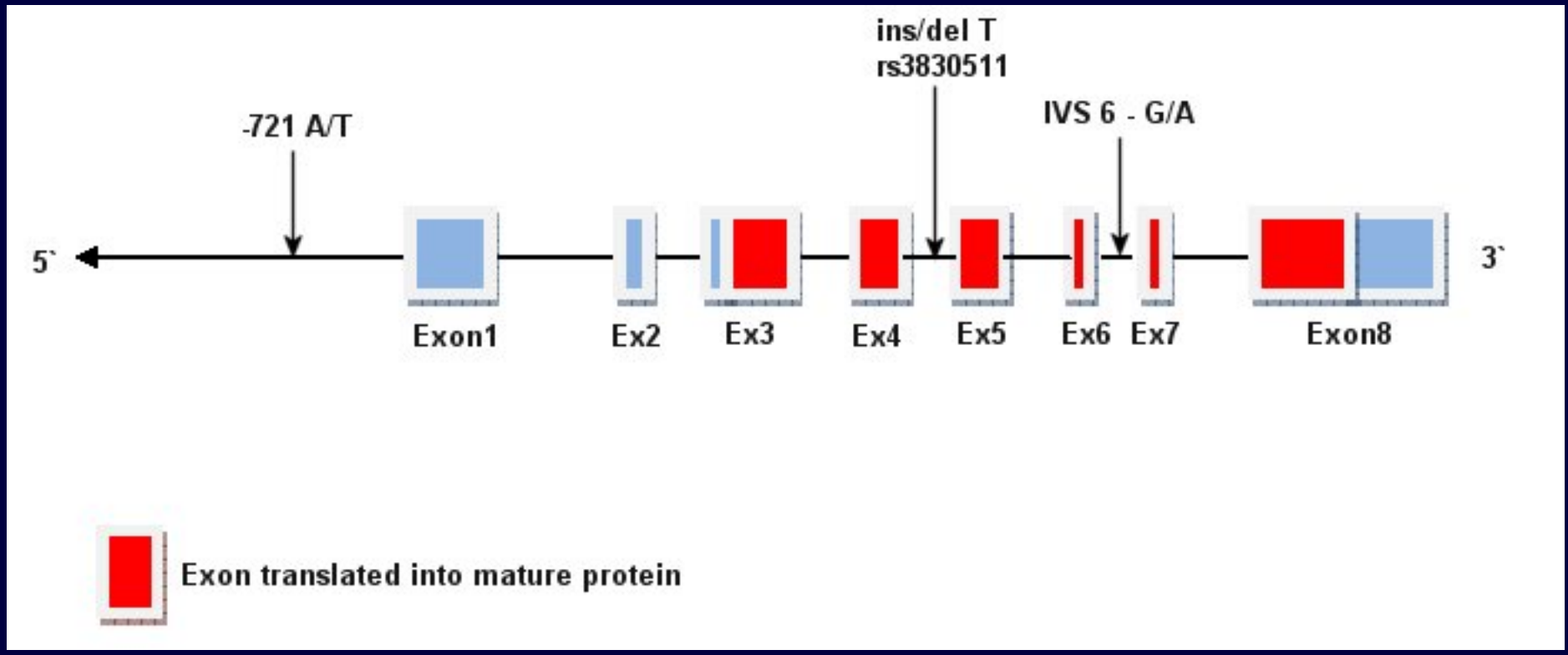
Types of DNA variation



At a particular DNA position

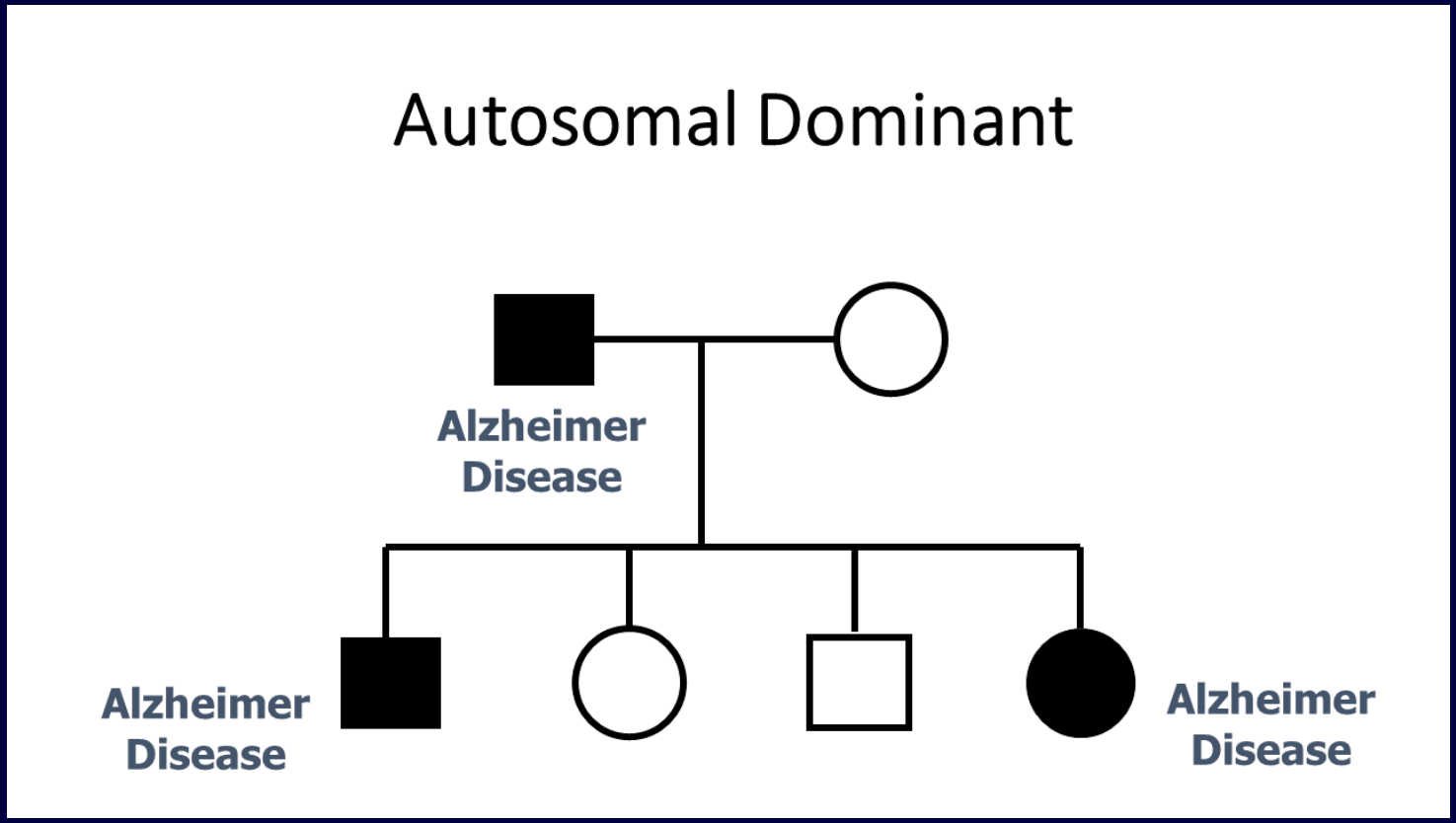
- Allele
  - A, T, G, C
- Genotype – combination of 2 alleles on the 2 members of the chromosome pair
  - A/A
  - A/T
  - T/T
- Single nucleotide polymorphism (SNP)
  - Specific position that has 2 possible nucleotides that can be found at that position
  - Called typically by an rs number

Gene Structure



Genetics of Alzheimer Disease

Early Onset Alzheimer Disease



- Onset of disease may be very young (20-60 years)

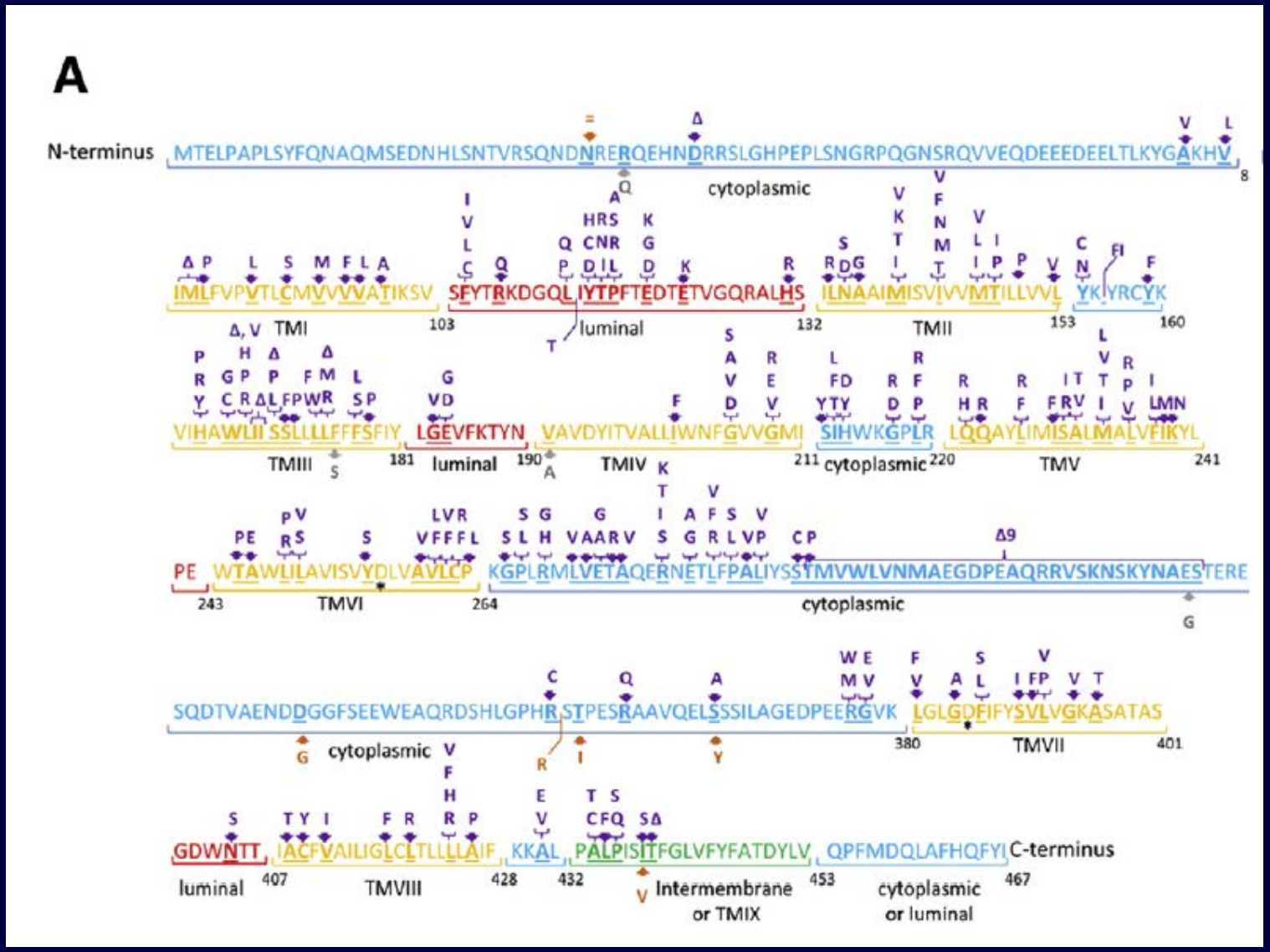
Early Onset AD Genetics

Changes in the DNA sequence of one of these genes can result in Alzheimer disease

- Presenilin 1 (PS1)
- Presenilin 2 (PS2)
- Amyloid precursor protein (APP)

Change in DNA sequence that can cause disease is often called a ‘mutation’

Presenilin 1



DIAN: Dominantly Inherited Alzheimer Network

- Recruiting families with early onset AD who have a PS1, PS2 or APP mutation
  - Family members complete a detailed evaluation
  - Return each year for follow-up

What has early onset AD told us?

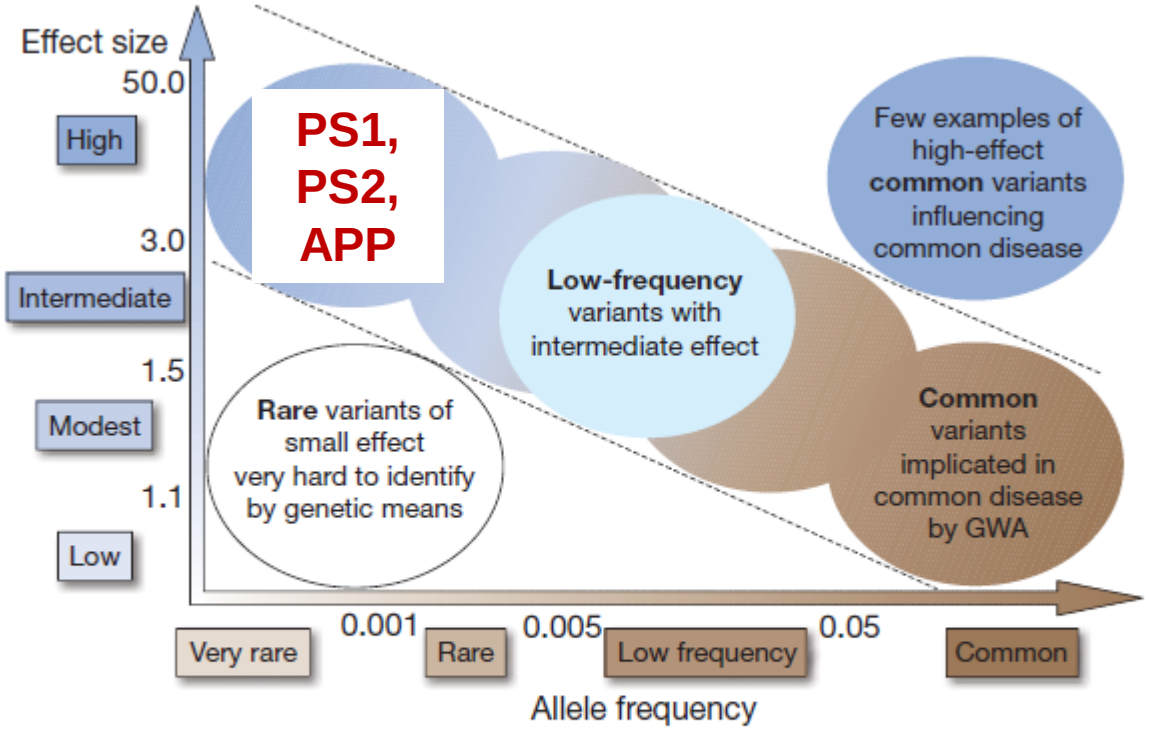
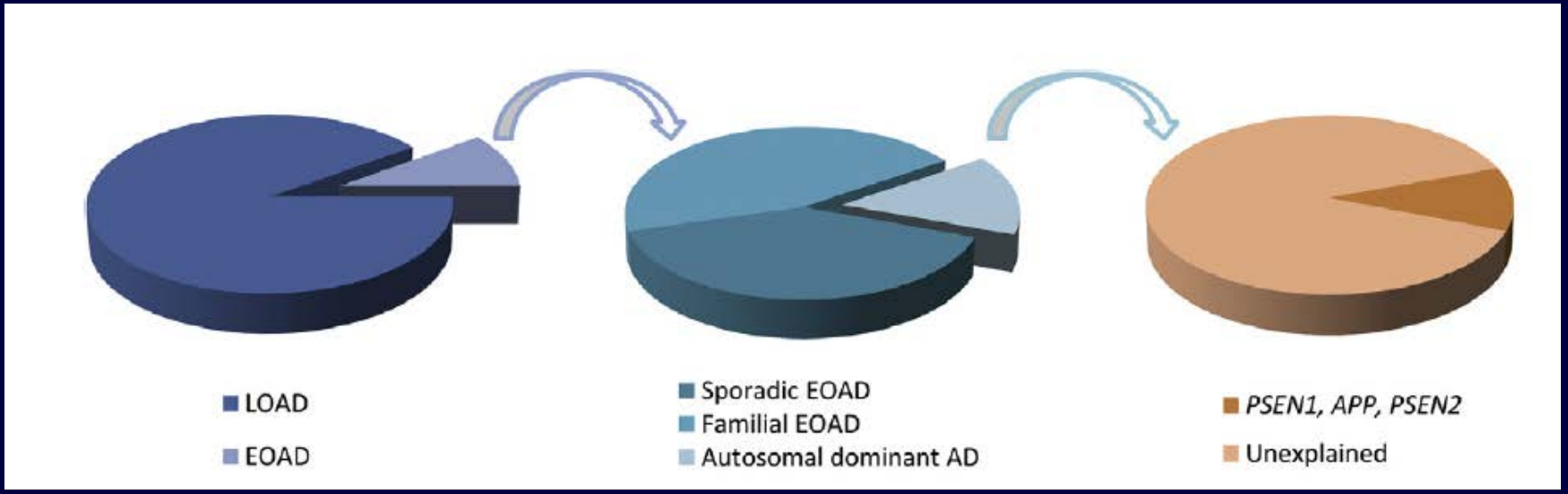


Figure 1 | Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Adapted from ref. 42.

Manolio et al, 2009

Alzheimer Disease



Cacace et al, 2016

Late Onset Alzheimer Disease

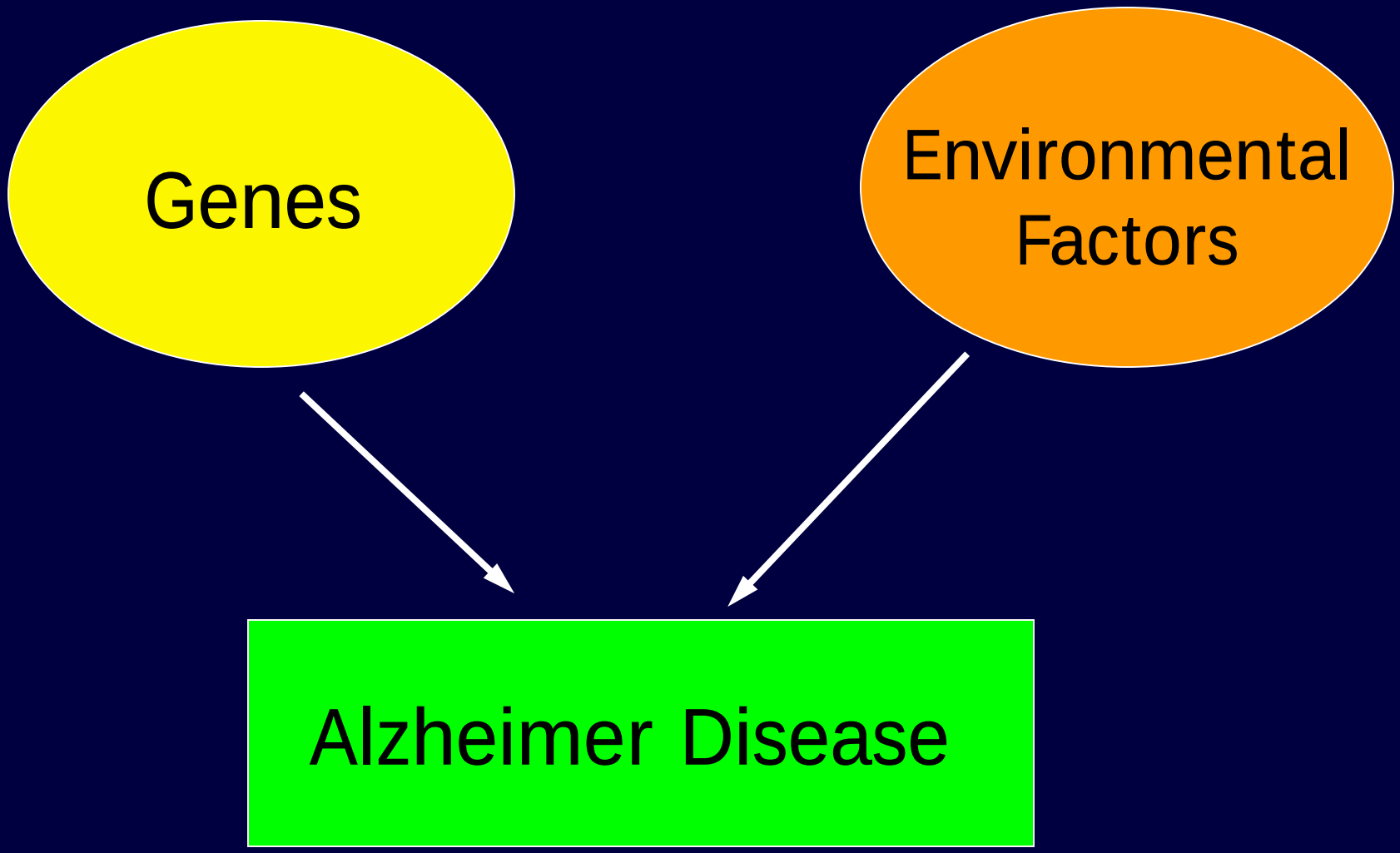
Late Onset Alzheimer Disease

- Families with late onset AD typically do not have a clear pattern of inheritance
- It is thought that rather than a single gene ‘causing’ AD as it does in early onset AD, in late onset AD there are likely to be many genes involved

NIA-LOAD Family Study

- Late Onset Alzheimer Disease (LOAD) study started in the ADCs over a decade ago
  - ADCs recruited multiplex late onset AD families
  - Many have been expanded to include a third generation
  - Very widely used in genetic studies

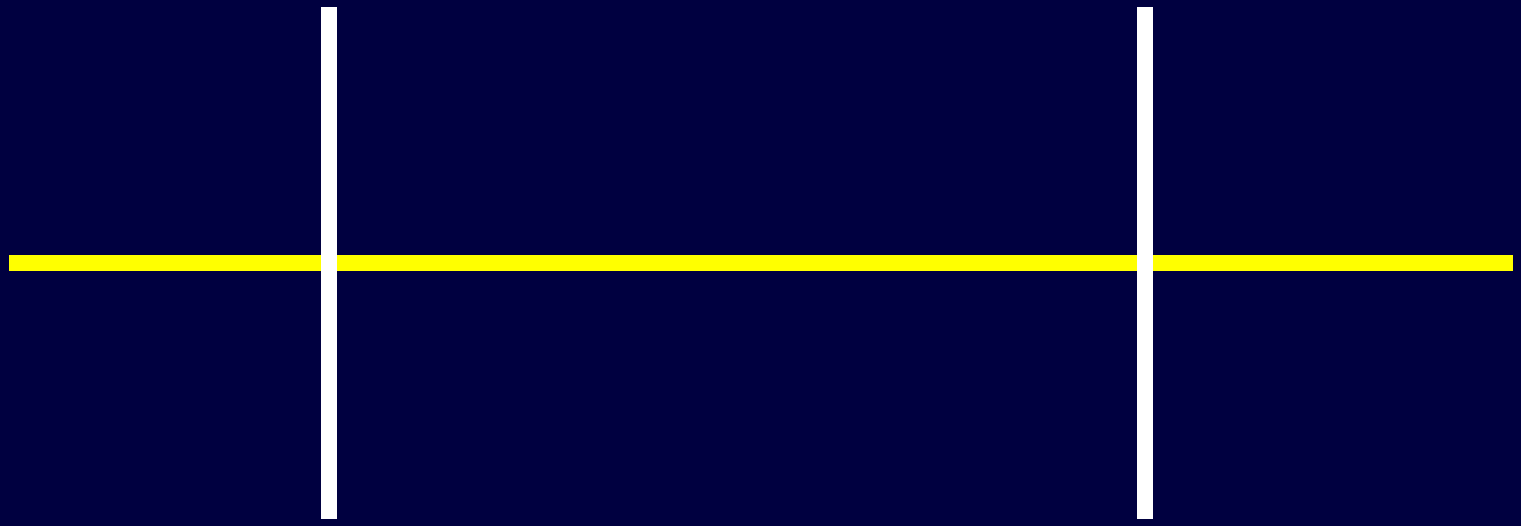
Alzheimer Disease



Apolipoprotein E (APOE) Susceptibility Gene

- Has 3 major forms:
  - APOE2, APOE3, APOE4
- Risk factor for AD
- Smaller effect on disease risk as compared to mutations in PS1, PS2, APP

APOE
Consists of Results from 2 SNPs



rs429358

T or C
Nucleotide
(allele)

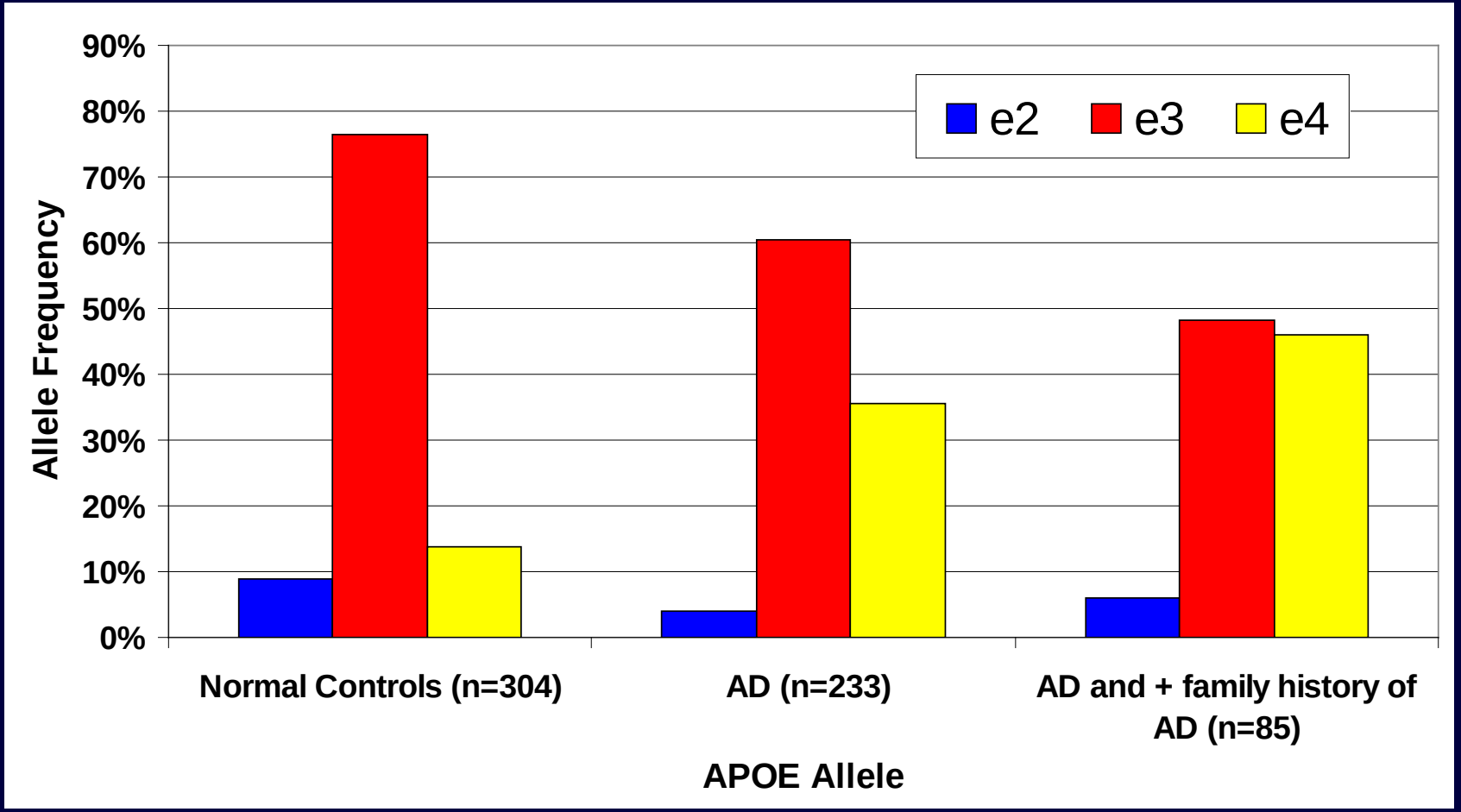
rs7412

T or C
Nucleotide
(allele)

APOE Genotype

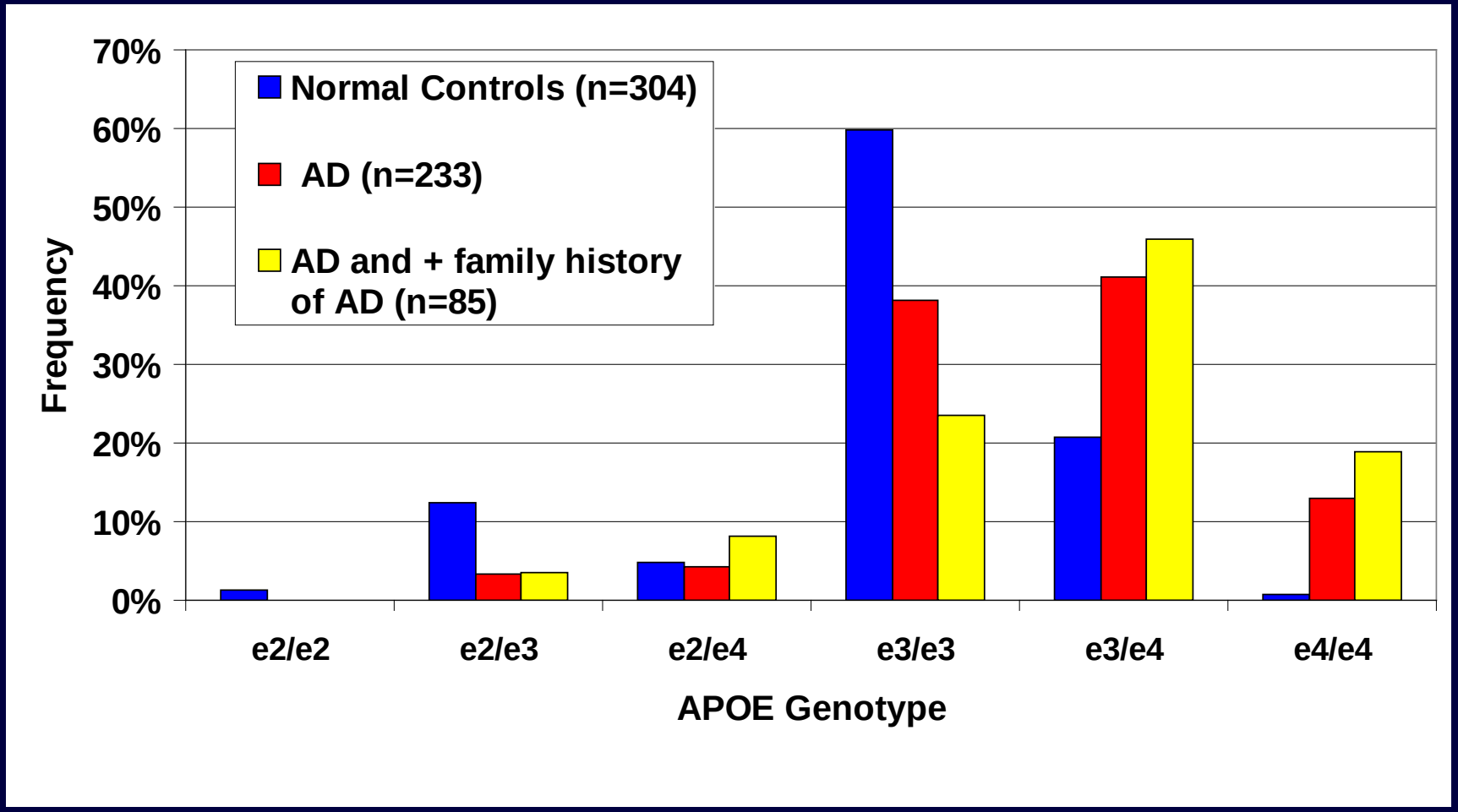
	rs429358 Allele 1	rs429358 Allele 2	rs7412 Allele 1	rs7412 Allele 2
E2 E2	T	T	T	T
E3 E3	T	T	C	C
E4 E4	C	C	C	C
E2 E3	T	T	C	T
E3 E4	C	T	C	C
E2 E4	C	T	C	T

APOE Allele Frequencies in Controls and Individuals with AD



Jarvik et al, 1996

APOE Genotypes in Controls and Individuals with AD



Jarvik et al, 1996

NCRAD and APOE

- NCRAD generates APOE genotype for all DNA samples sent to NCRAD from the ADCs once they match NACC data
  - APOE genotype is generated at LGC Genomics (previously at Prevention Genetics)
  - Samples are sent in batches by NCRAD 3-4 times/year
  - Results of APOE genotyping are made available by NACC on the ADC NACC website

What have we learned about AD?

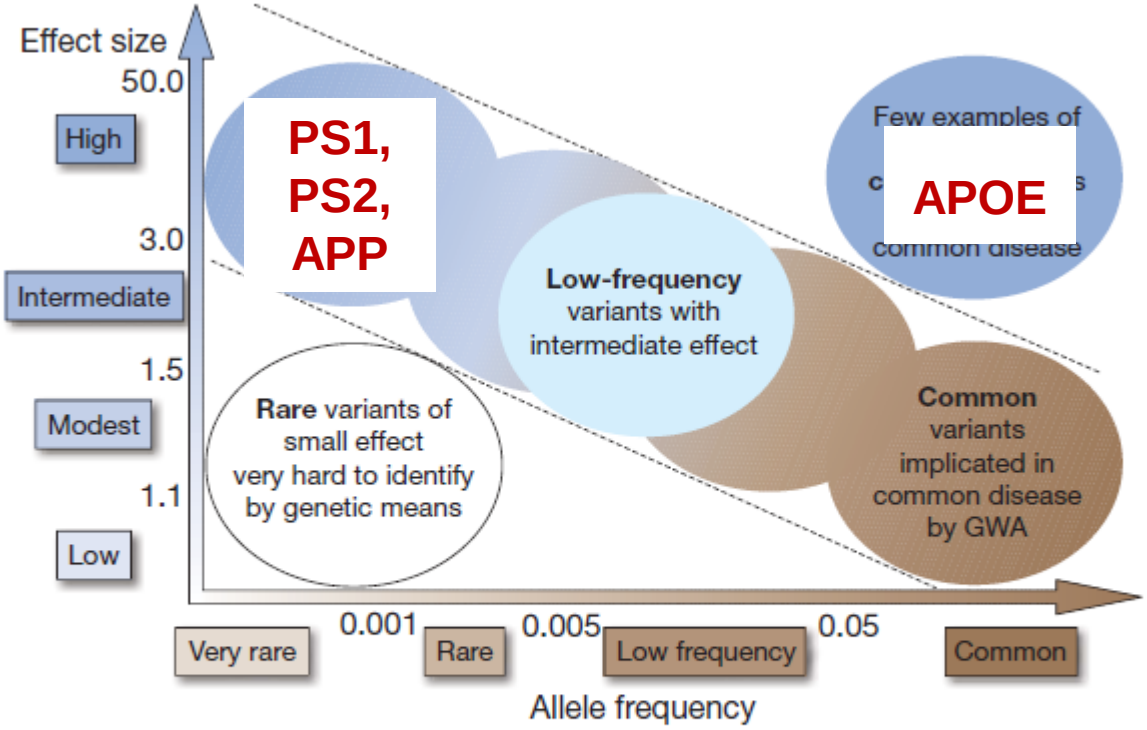


Figure 1 | Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Adapted from ref. 42.

Manolio et al, 2009

Genome Wide Association Studies (GWAS)

- Genome wide association studies are intended to provide dense coverage of the whole genome
- Dense coverage allows the detection of genes (alleles) associated with phenotypes including disease risk and therapeutic effect

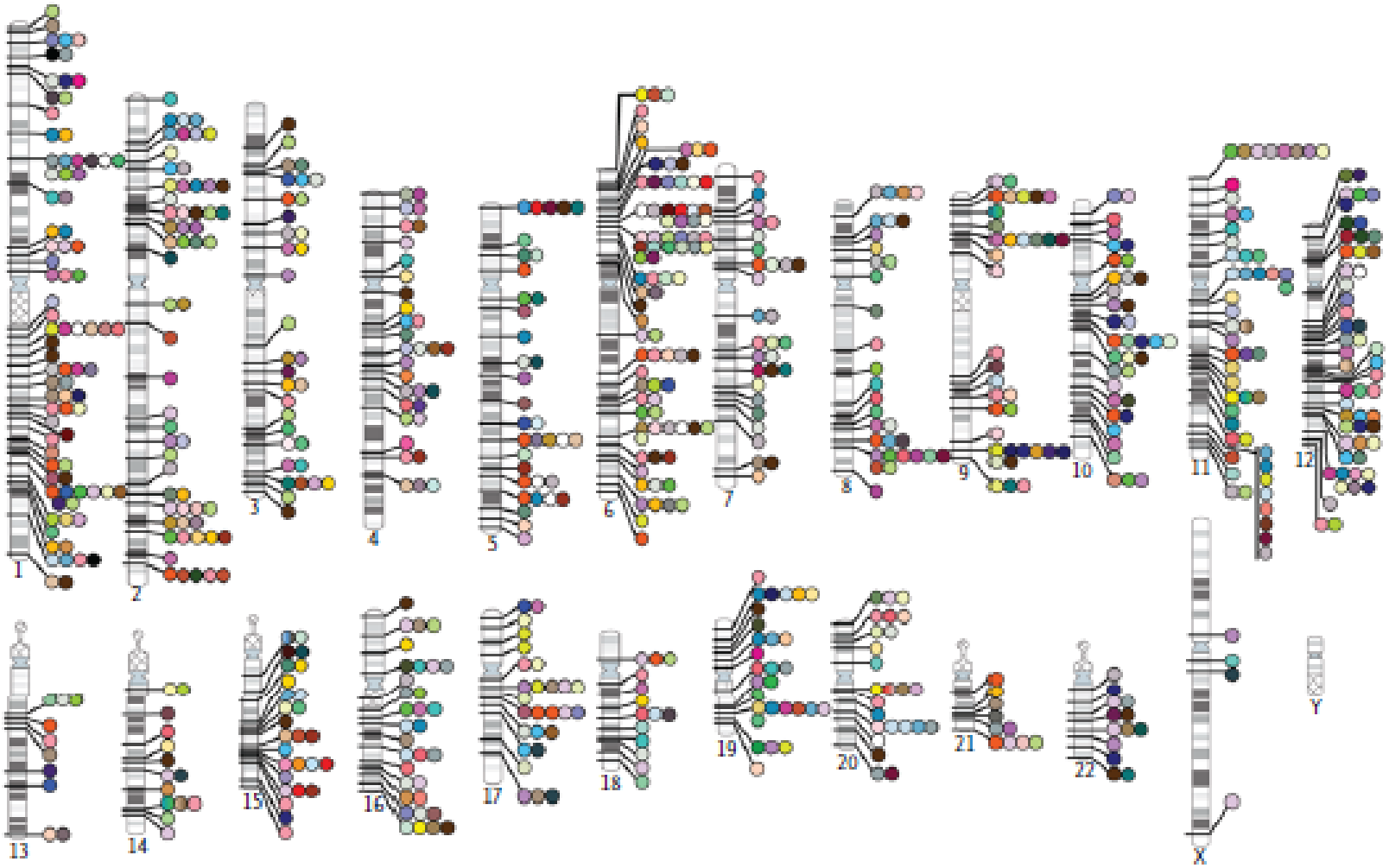


Figure 3. Genomewide Associations Reported through March 2010.

Circles indicate the chromosomal location of nearly 800 single-nucleotide polymorphisms (SNPs) significantly associated ($P<5\times10^{-8}$) with a disease or trait and reported in the literature (545 studies published through March 2010 yielded the associations depicted). Each disease type or trait is coded by color. Adapted from the National Human Genome Research Institute.[4]

Genome Wide Association Studies

Test millions of SNPs throughout the genome

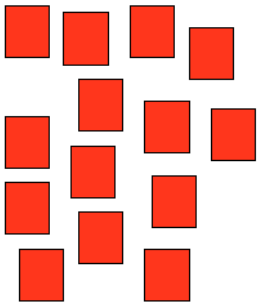
Compare frequency of SNP alleles in two groups

Compare frequency of SNP genotypes in two groups

GWAS Study Design

Selection of cases

Cases

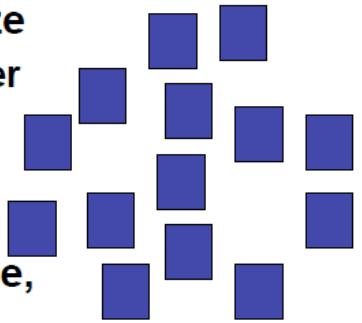


- Potential criteria to enrich genetic effect size
  - More severely affected individuals
  - Require other family member to have disease
  - Younger age-of-disease onset

Selection of controls

- Potential criteria to enrich genetic effect size
  - Low risk of disease rather than population-based samples
- Matched to cases on age, sex, demographics

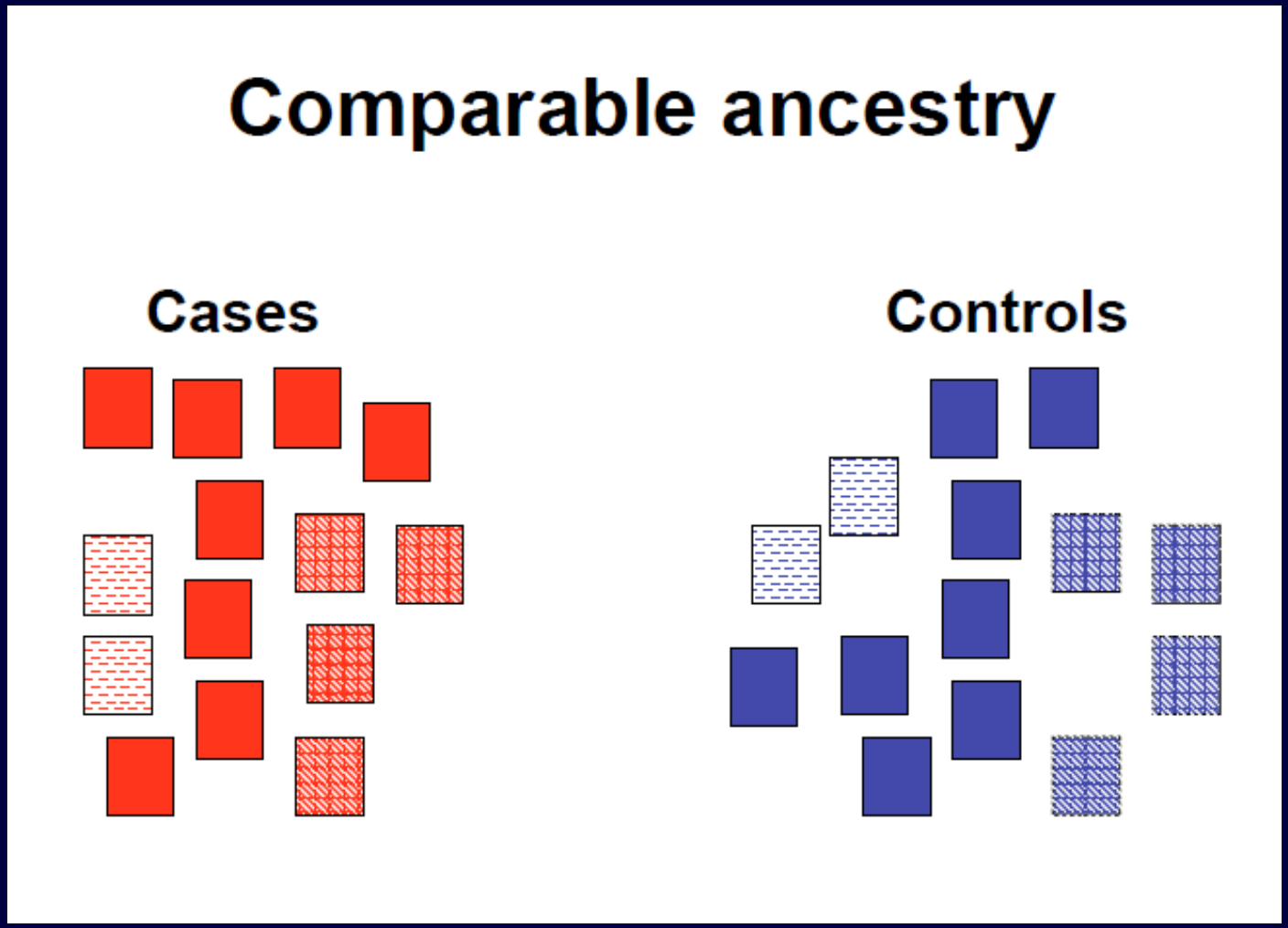
Controls



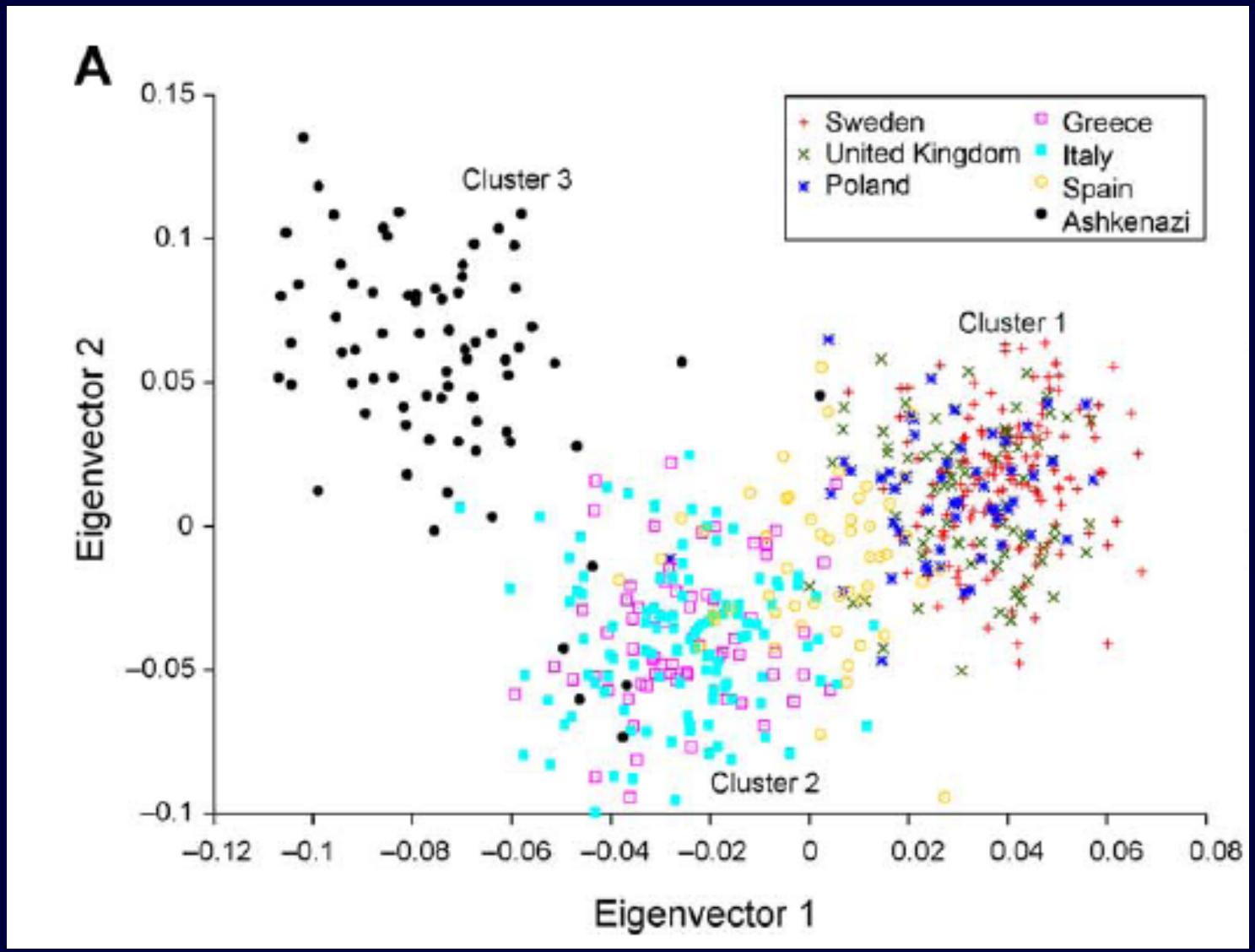
AD cases

Healthy controls

GWAS Study Design



Sample Stratification



Price et al, PLOS Genetics, e236, 2008

Genomewide Association Study (Manhattan Plot)

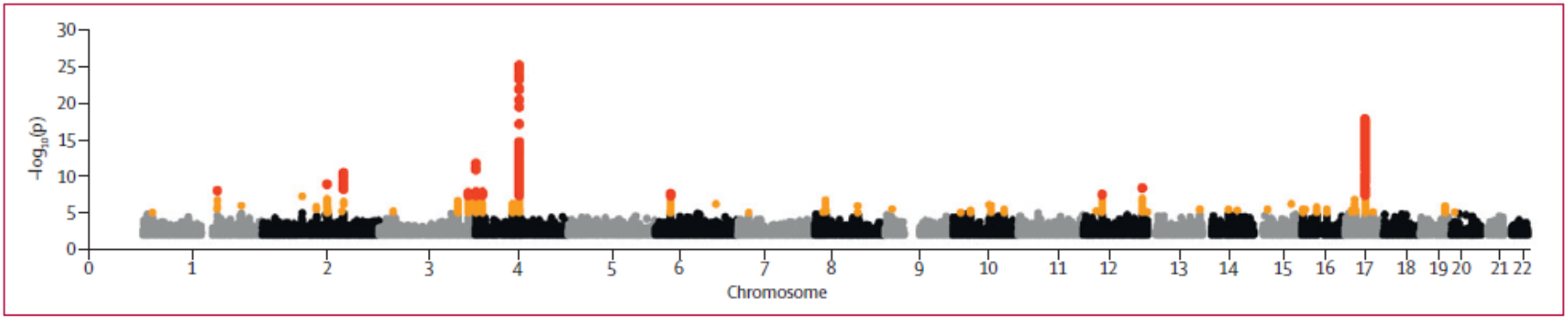


Figure 1: **Manhattan plot of Parkinson's disease associations for all SNPs in the discovery phase**
p values from fixed-effects meta-analysis for 7 689 524 SNPs successfully imputed or genotyped in at least two individual datasets. Genomic inflation factor=1·035. Red points=SNPs with $p<5\times10^{-8}$. Orange points=SNPs with p values ranging from less than 1×10^{-5} to 5×10^{-8}. Regions containing red points were followed up in replication analyses. SNP=single nucleotide polymorphism.

IPDGC, 2011

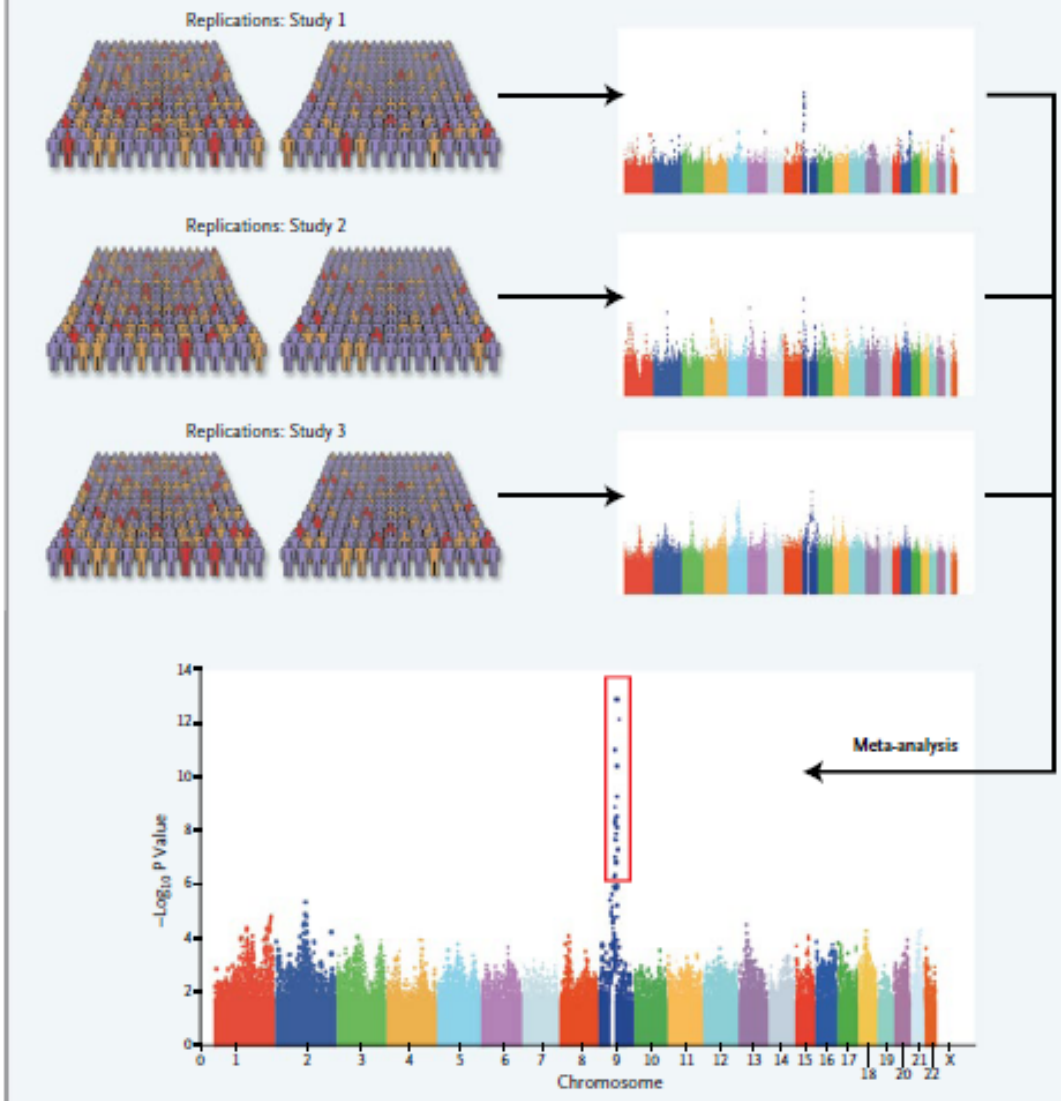


Figure 2. Meta-Analysis of Genomewide Association Studies.

The results of genomewide association studies can be evaluated in a meta-analysis, which combines the results of multiple studies to improve the power for detecting associations. In this example, the results of three studies, none of which may show genomewide significance individually, are combined in a meta-analysis to reveal a strong, significant signal on chromosome 9.

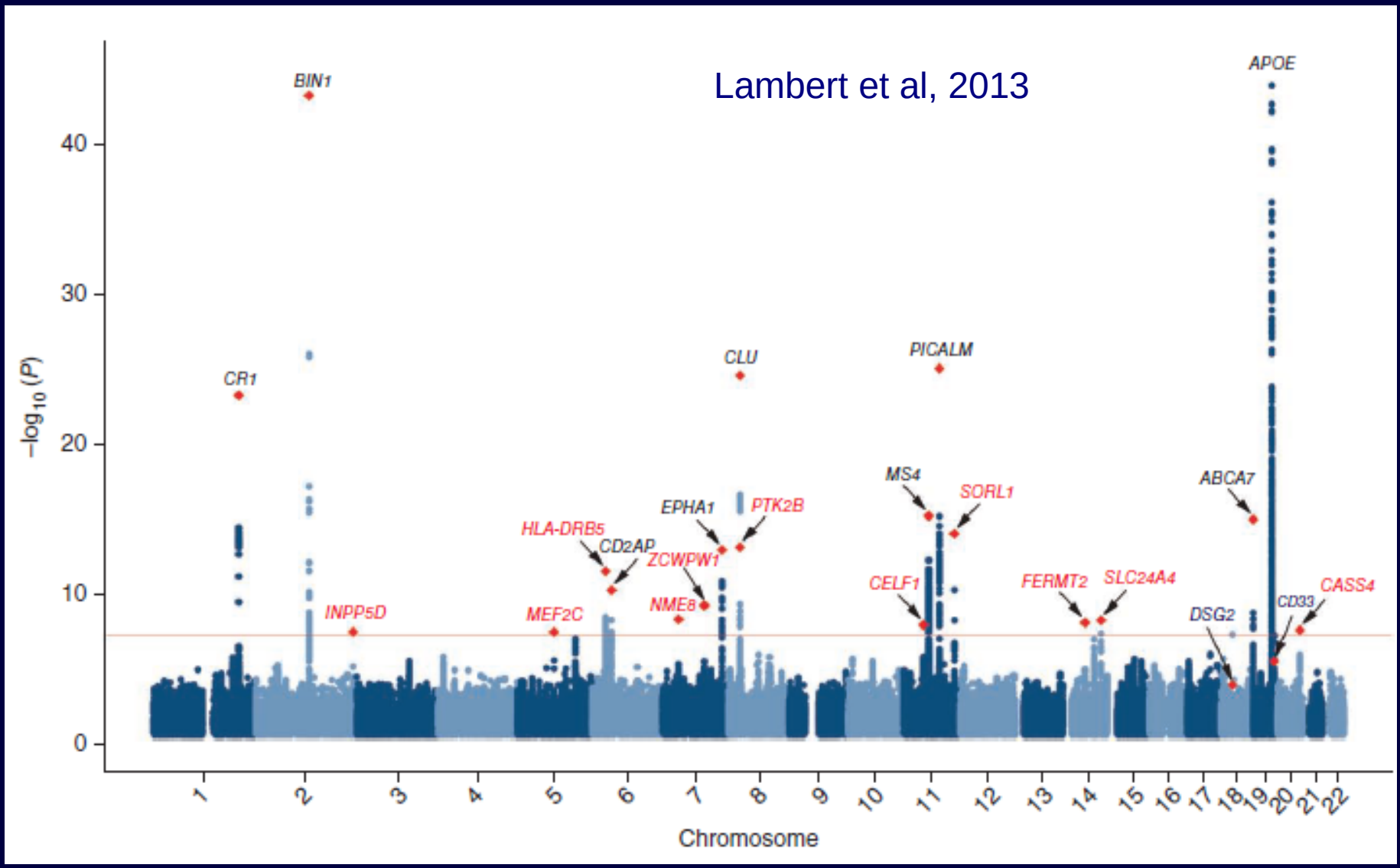
Meta-Analysis

A challenge in GWAS is being certain what you have identified is real

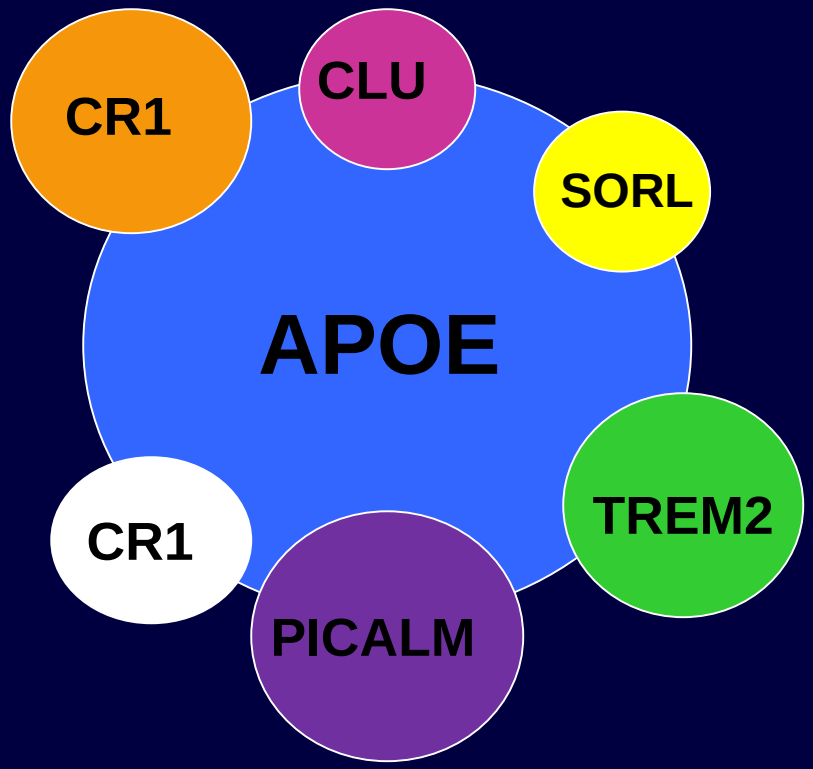
Manolio, 2010

GWAS in Alzheimer Disease

Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for Alzheimer's disease



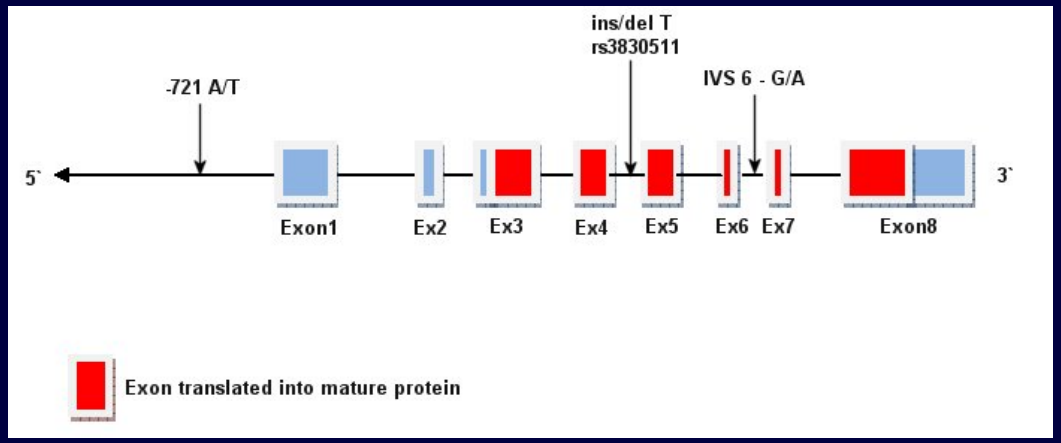
Additional Genes Important in Late Onset AD



- In total, more than 20 genes have been identified that may play a role in the risk of Alzheimer disease
- They may work together in various combinations

GWAS vs. Exome Chip

- GWAS is focused on ‘common’ SNPs
  - Largely found in regions outside the exons
- Exome chip is focused on SNPs in exons
  - Tend to be SNPs that are not ‘common’
- Initially ran exome chip alone; now exome chip SNPs included with GWAS (combined chip)



ADCs and GWAS Data

- Alzheimer Disease Genetics Consortium (ADGC) generates GWAS data using ADC samples
  - Focus has been on AD cases and controls
  - GWAS is run in rounds (Rounds 1-8 finished)
- Data is returned by the ADGC to NACC
  - NACC posts GWAS data for each ADC to the NACC website
  - NACC posts exome chip data (alone or with GWAS)
  - File type: plink (can’t use excel – too many SNPs

What have we learned about AD?

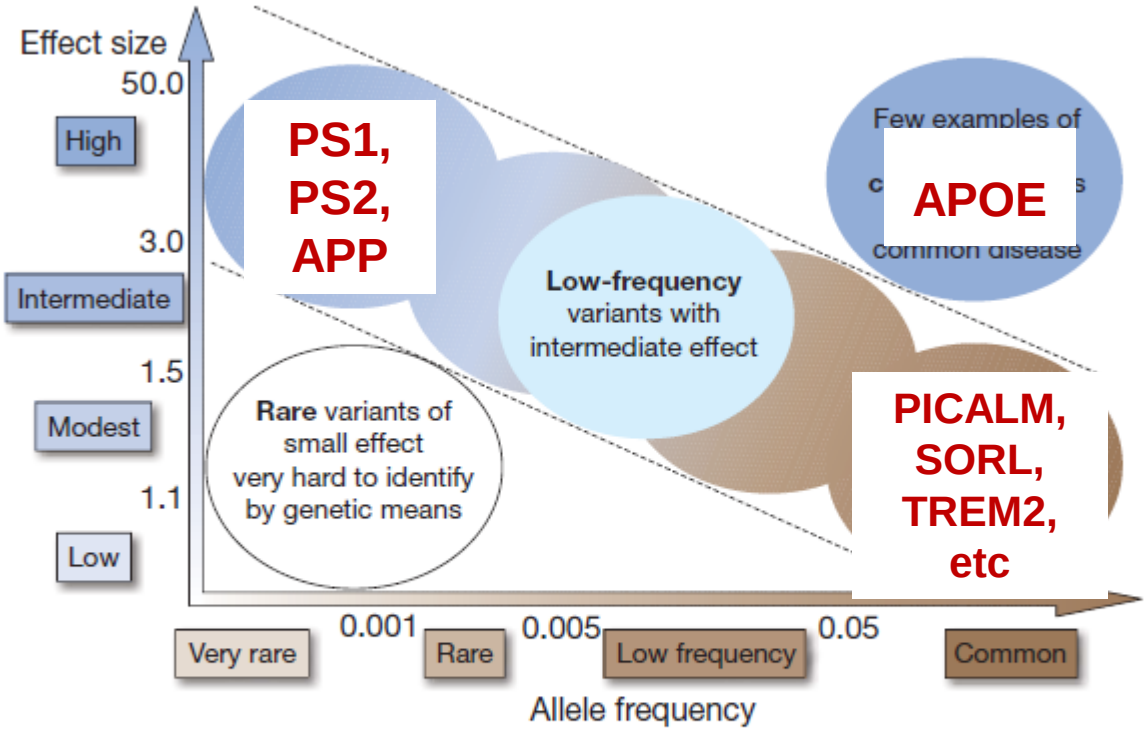
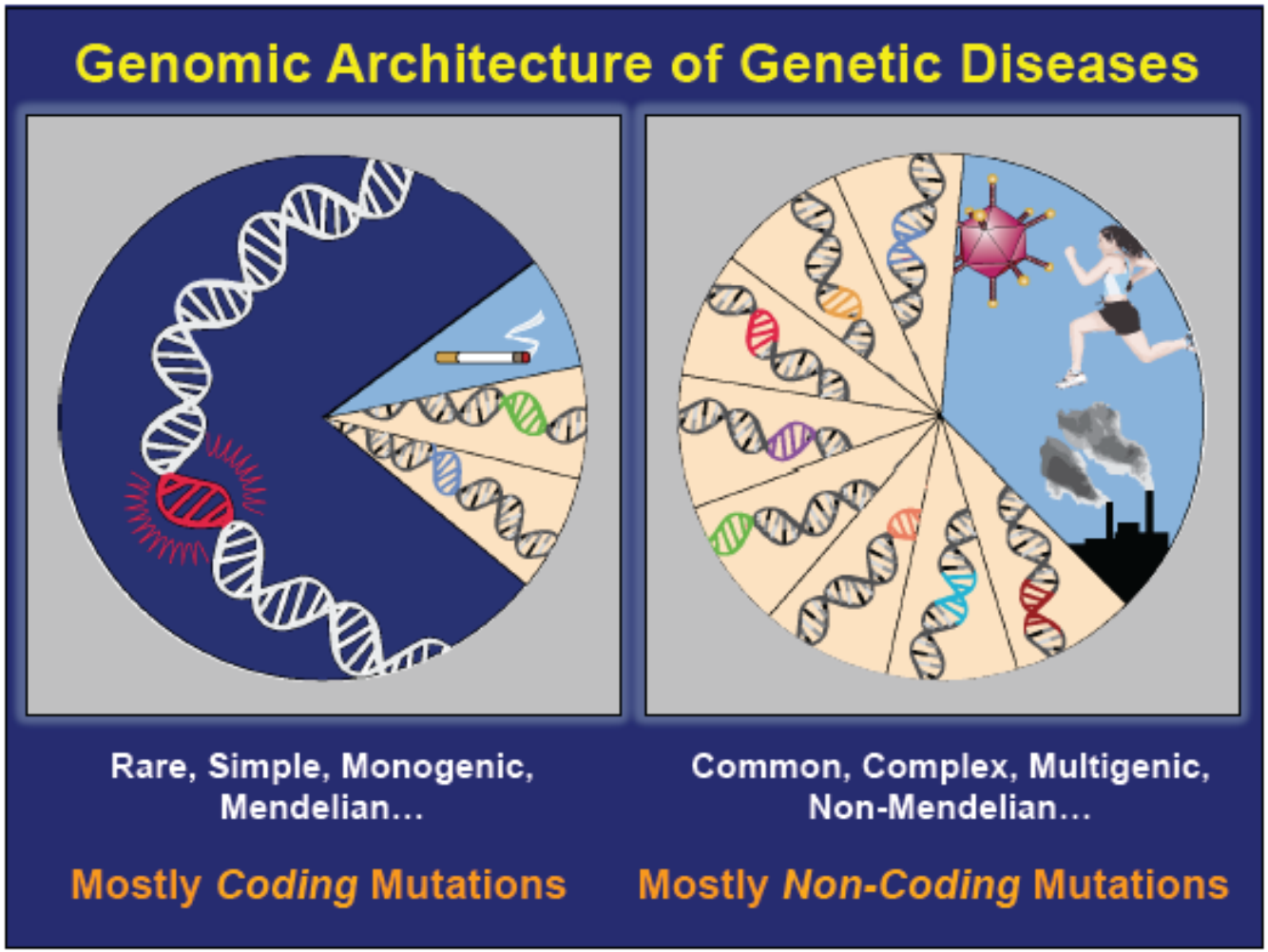


Figure 1 | Feasibility of identifying genetic variants by risk allele frequency and strength of genetic effect (odds ratio). Most emphasis and interest lies in identifying associations with characteristics shown within diagonal dotted lines. Adapted from ref. 42.

Manolio et al, 2009

Genomic Architecture of Disease



Sequencing

- Sequencing is not a new idea
  - PS1, APP, PS2 all found by sequencing nearly 3 decades ago
  - The human genome was sequenced more than a decade ago
- What is new is how we can sequence
  - New technologies and methods make it faster and easier
  - No longer focus on sequencing a single gene, now we can sequence the entire exome or genome

	Whole-genome sequencing (WGS)	Exome sequencing
Cost	Still costly, but decreasing rapidly	Reduced cost is a tenth to a third of WGS
Technical	No capture step, automatable	Capture step, technical bias
Variation	Uncovers all genetic and and genomic variation (SNVs and CNVs) Discovery of functional coding and noncoding variation ~3.5 million variants	Focuses on ~1% of the genome Limited to coding and splice-site variants in annotated genes ~20,000 variants
Disease	Suitable for mendelian and complex trait gene identification, as well as sporadic phenotypes caused by *de novo* SNVs or CNVs	Good for highly penetrant mendelian disease gene identification

Figure 3

A comparison of the weaknesses and strengths of whole-genome sequencing (WGS) and exome sequencing approaches for disease-gene identification. Abbreviations: CNVs, copy-number variants; SNVs, simple nucleotide variants.

Gonzaga-Jauregui, Lupski and Gibbs, Ann Rev Med 63: 35-61, 2012.

ADSP Sequence Data

- Generated in high quality research laboratories funded by NIH
- Intended for research purposes only
- Not intended to be returned to subjects

ADSP Sequence Data

- Data from the Discovery Phase is available through dbGaP (NIAGADS)
- ADCs can request WES from subjects in their ADC

- Data not available through NACC, different process

Genomic Data Sharing Policy

- Created to ensure the broad and responsible sharing of genomic research data.
- Effective January 25, 2015
- Applies to all NIH-funded research
- https://gds.nih.gov/06researchers1.html
  - Recommendations for researchers including guidance for consent documents

Kelley Faber