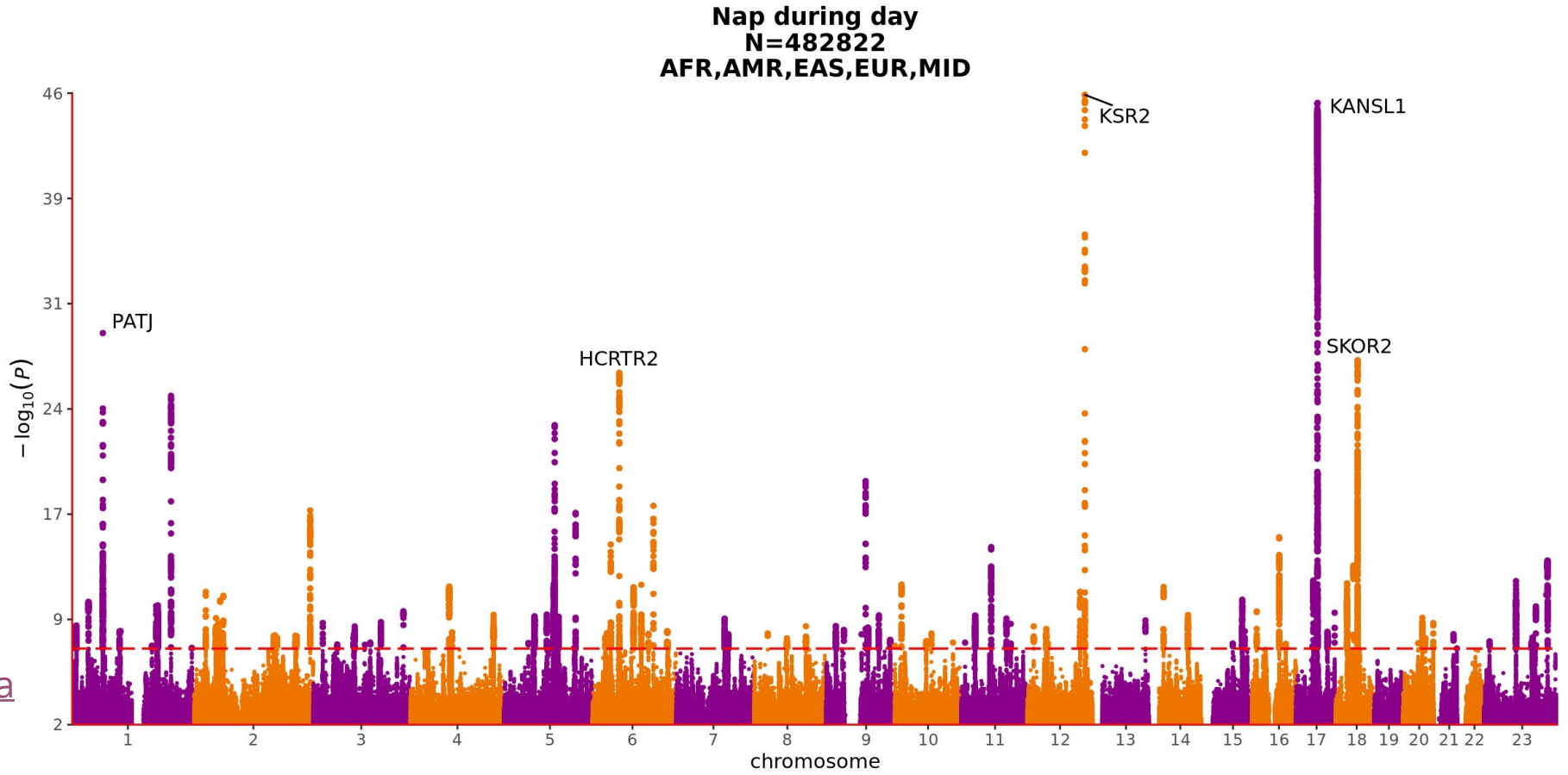
Intro to GWAS

BIOL 435/535: Bioinformatics

Feb 24th, 2022



@SbotGwa

ARTICLES
https://doi.org/10.1038/s41588-018-0147-3

nature genetics

Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals

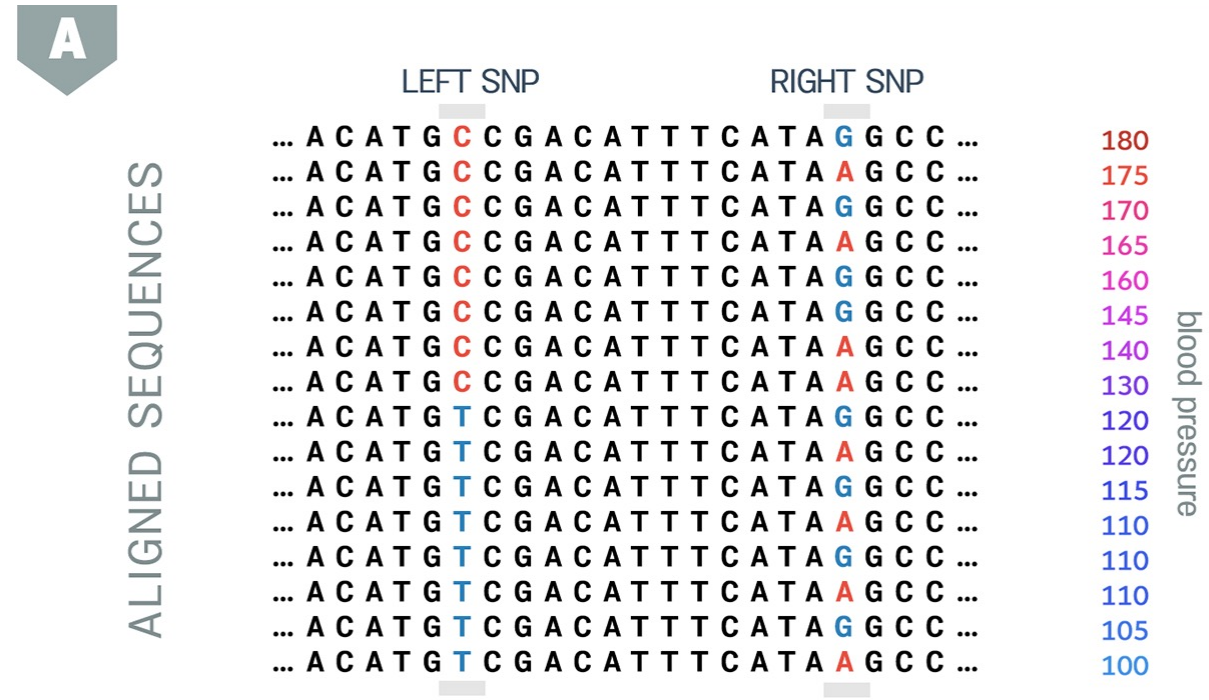
James J. Lee[1,58], Robbee Wedow[2,3,4,58], Aysu Okbay[5,6,58*], Edward Kong[7], Omeed Maghzian[7], Meghan Zacher[8], Tuan Anh Nguyen-Viet[9], Peter Bowers[7], Julia Sidorenko[10,11], Richard Karlsson Linnér[5,6,12], Mark Alan Fontana[9,13], Tushar Kundu[9], Chanwook Lee[7], Hui Li[7], Ruoxi Li[9], Rebecca Royer[9], Pascal N. Timshel[14,15], Raymond K. Walters[16,17], Emily A. Willoughby[1], Loïc Yengo[10], 23andMe Research Team[18], COGENT (Cognitive Genomics Consortium)[19], Social Science Genetic Association Consortium[18], Maris Alver[11], Yanchun Bao[20], David W. Clark[21], Felix R. Day[22], Nicholas A. Furlotte[23], Peter K. Joshi[21,24], Kathryn E. Kemper[10], Aaron Kleinman[23], Claudia Langenberg[22], Reedik Mägi[11], Joey W. Trampush[25,26], Shefali Setia Verma[27], Yang Wu[10], Max Lam[28,29], Jing Hua Zhao[22], Zhili Zheng[10,30], Jason D. Boardman[2,3,4], Harry Campbell[21], Jeremy Freese[31], Kathleen Mullan Harris[32,33], Caroline Hayward[34], Pamela Herd[30,35], Meena Kumari[20], Todd Lencz[36,37,38], Jian'an Luan[22], Anil K. Malhotra[36,37,38], Andres Metspalu[11,39], Lili Milani[11], Ken K. Ong[22], John R. B. Perry[22], David J. Porteous[40], Marylyn D. Ritchie[27], Melissa C. Smart[21], Blair H. Smith[41,42], Joyce Y. Tung[23], Nicholas J. Wareham[22], James F. Wilson[21,24], Jonathan P. Beauchamp[43], Dalton C. Conley[44], Tõnu Esko[11], Steven F. Lehrer[45,46,47], Patrik K. E. Magnusson[48], Sven Oskarsson[49], Tune H. Pers[14,15], Matthew R. Robinson[10,50], Kevin Thom[51], Chelsea Watson[9], Christopher F. Chabris[52], Michelle N. Meyer[53], David I. Laibson[7], Jian Yang[10,54], Magnus Johannesson[55], Philipp D. Koellinger[5,6,12], Patrick Turley[16,17,59], Peter M. Visscher[10,54,59*], Daniel J. Benjamin[9,43,56,59*] and David Cesarini[47,51,57,59]

Genome-wide association studies

- Identify SNPs that are associated with phenotype(s) of interest

H_O: Phenotype ⊥ SNP

H_A: Phenotype ~ SNP



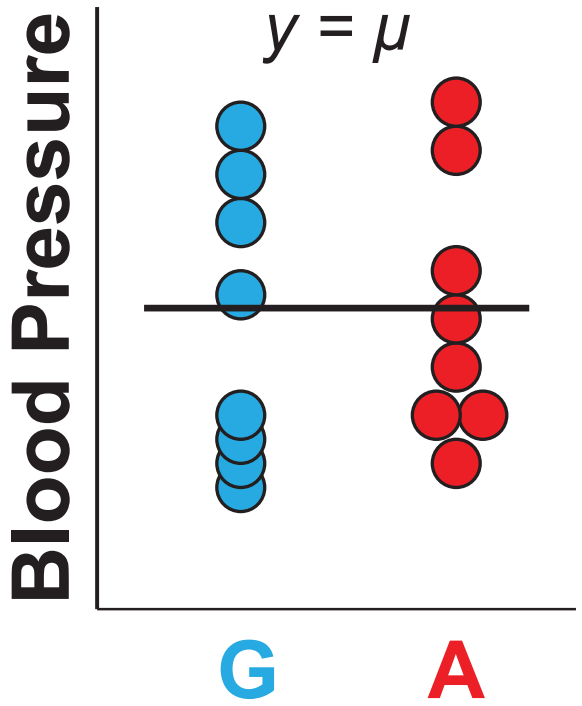
Sul et al 2018

Genome-wide association studies

- Identify SNPs that are associated with phenotype(s) of interest

H_O: Phenotype ⊥ SNP

H_A: Phenotype ~ SNP



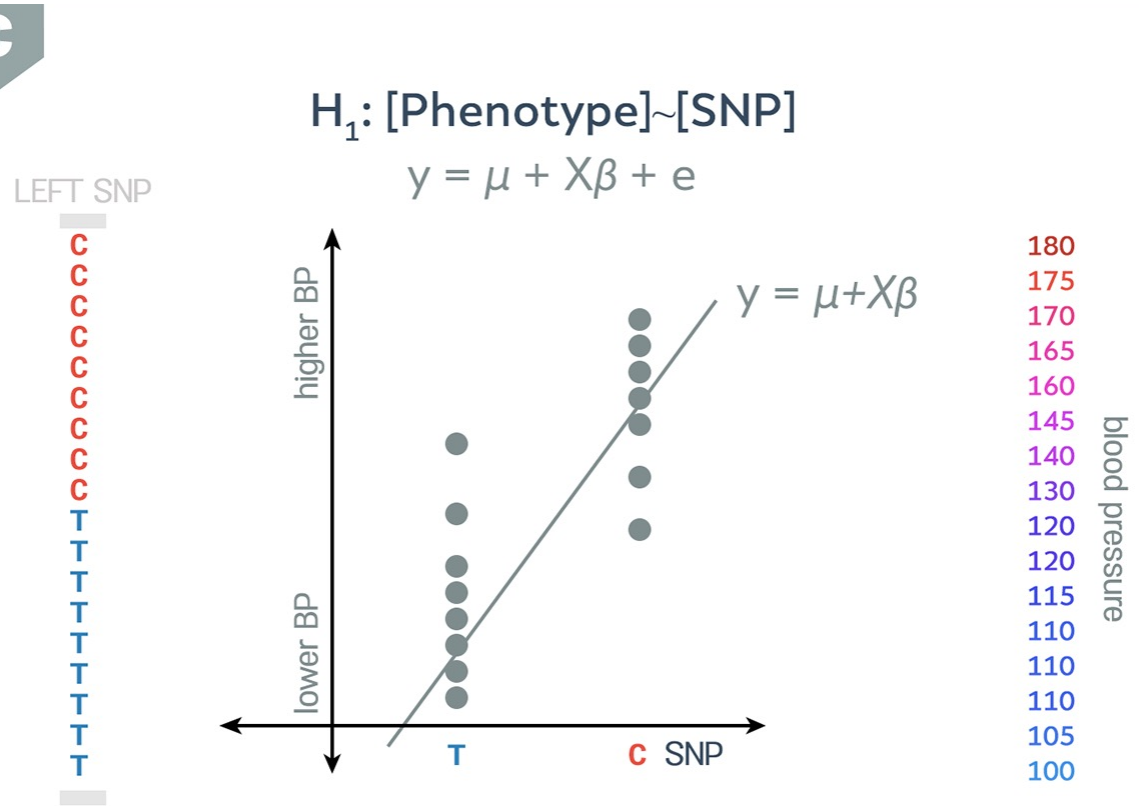
Sul et al 2018

Genome-wide association studies

- Identify SNPs that are associated with phenotype(s) of interest

H_O: Phenotype ⊥ SNP

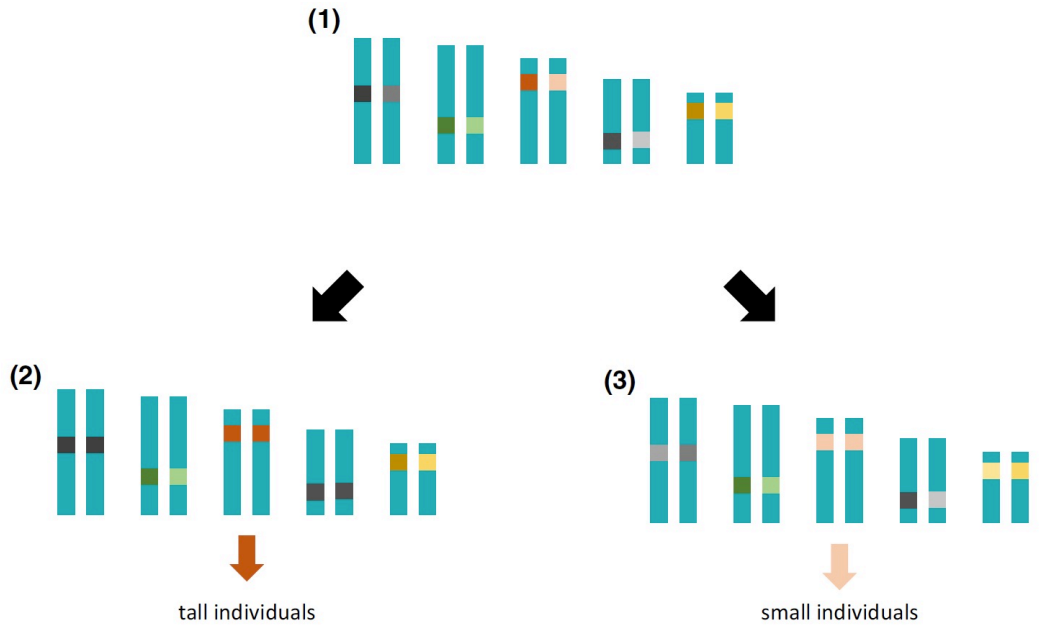
H_A: Phenotype ~ SNP



Sul et al 2018

SNP associations may not be due to causal effect

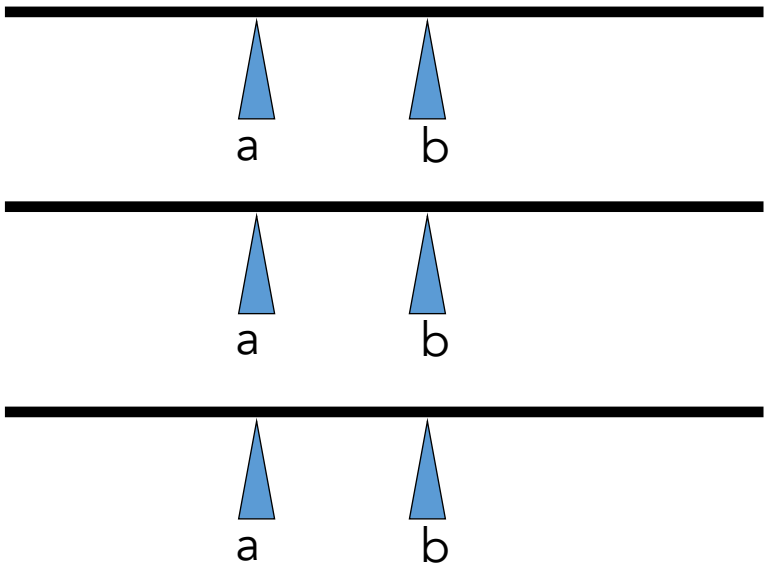
Population structure (differential relatedness among individuals) means that SNPs resulting from Isolation by Descent (IBD) will be associated with the phenotype of interest



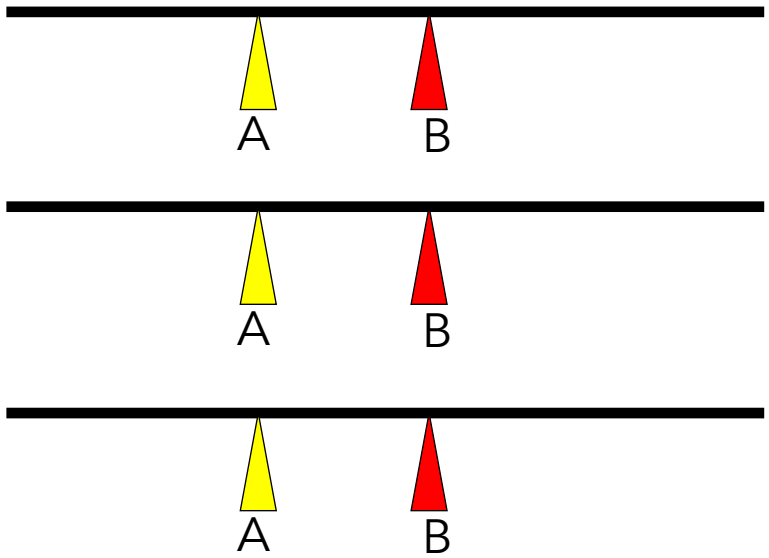
Santure and Garant 2018

Population structure causes linkage disequilibrium (LD)

Sub-population 1

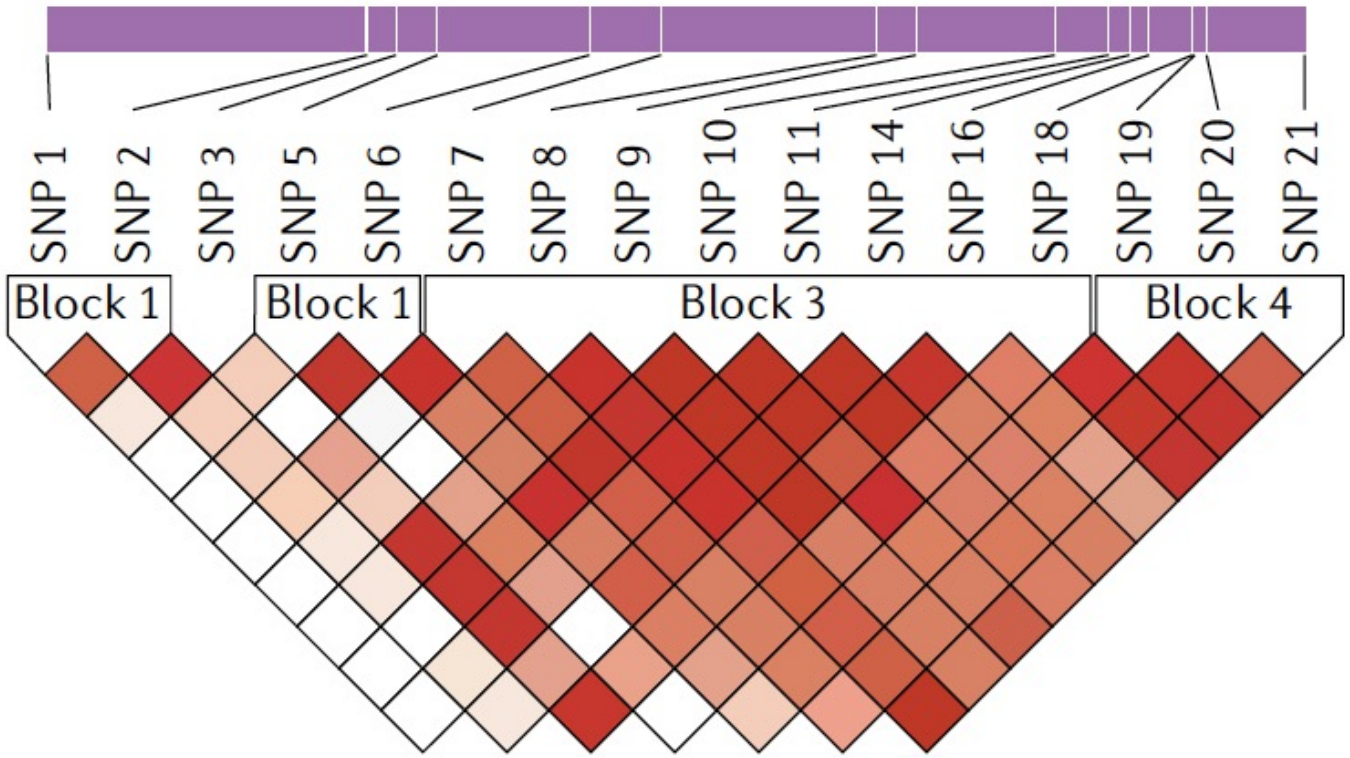


Sub-population 2

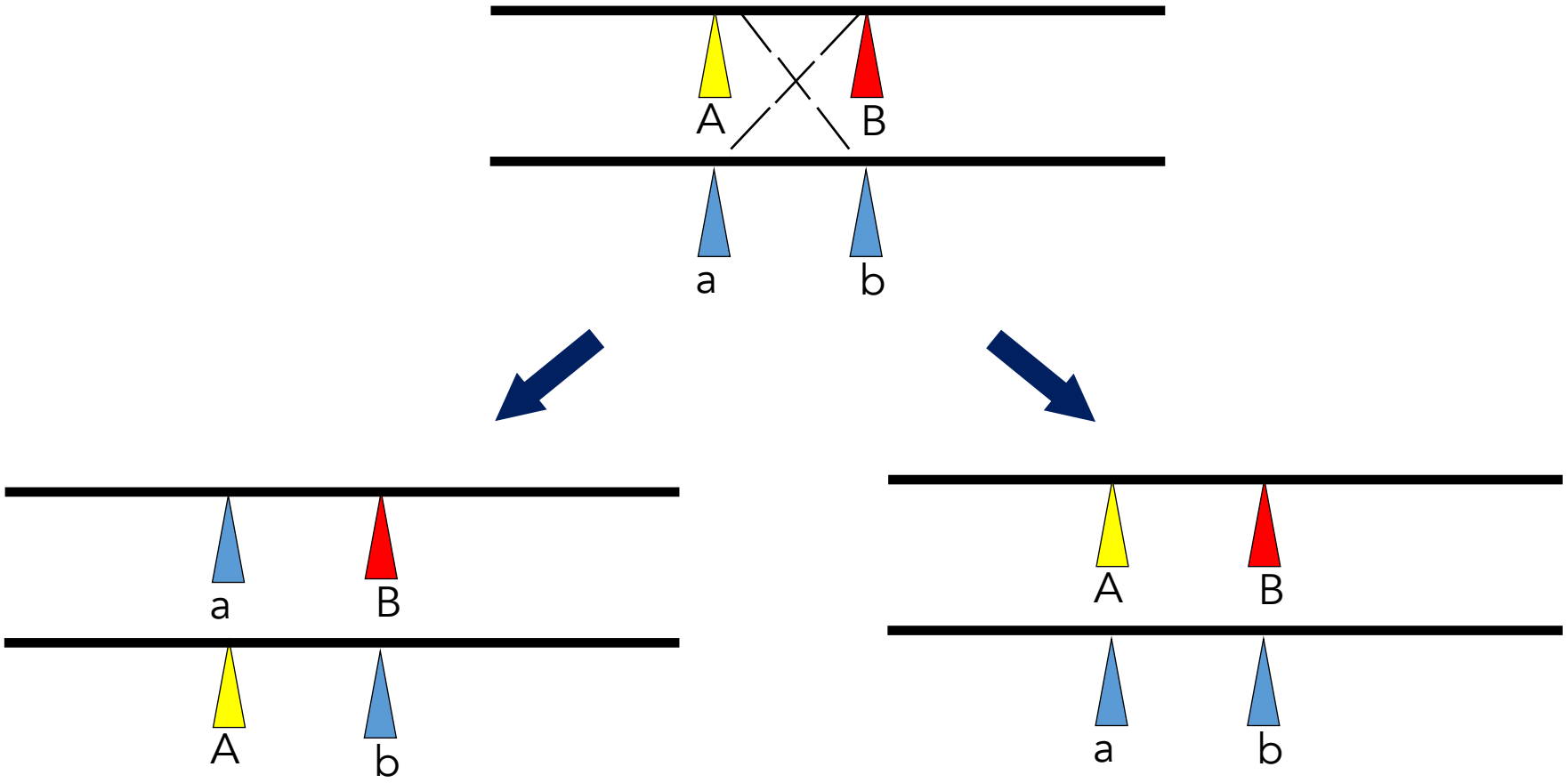


Population structure causes linkage disequilibrium (LD)

Linkage disequilibrium



Recombination can break down LD,
but it takes time, secondary contact

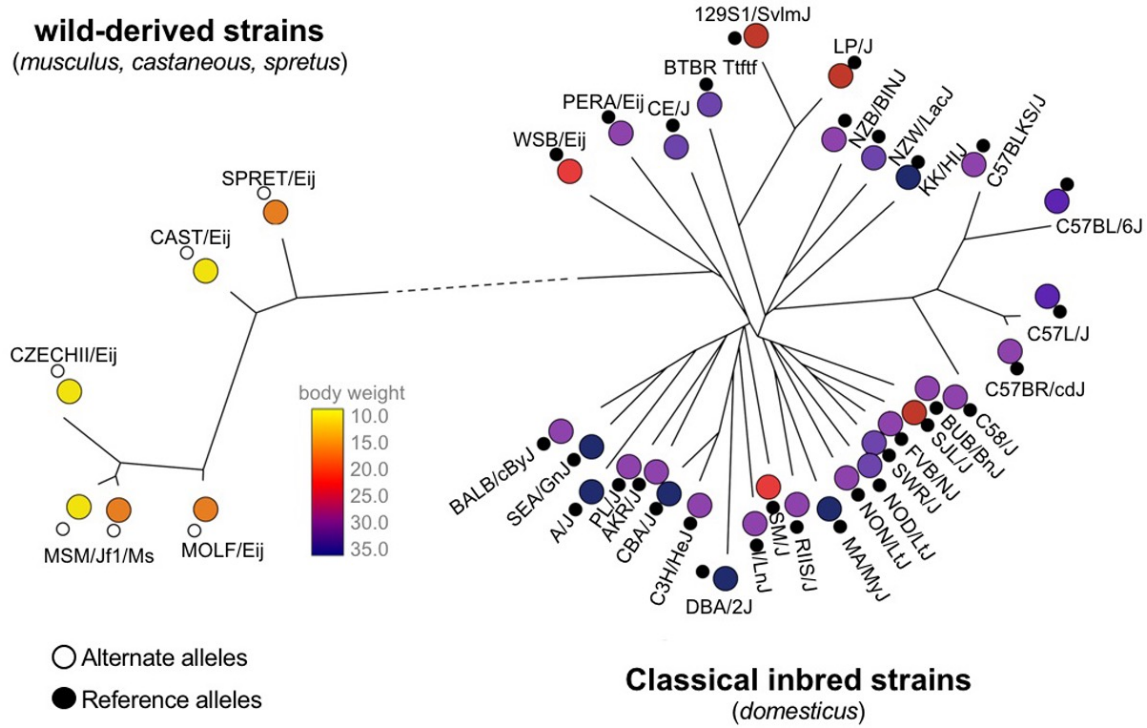


Recombinant alleles

Parental alleles

SNP associations may not be due to causal effect

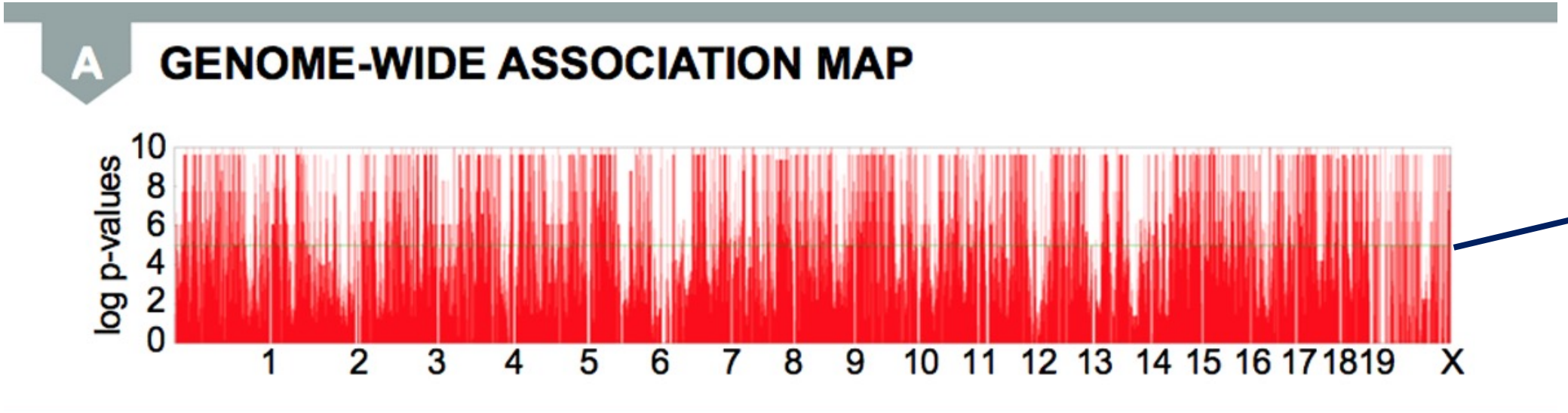
Population structure (differential relatedness among individuals) means that SNPs resulting from Isolation by Descent (IBD) will be associated with the phenotype of interest



Sul et al 2018

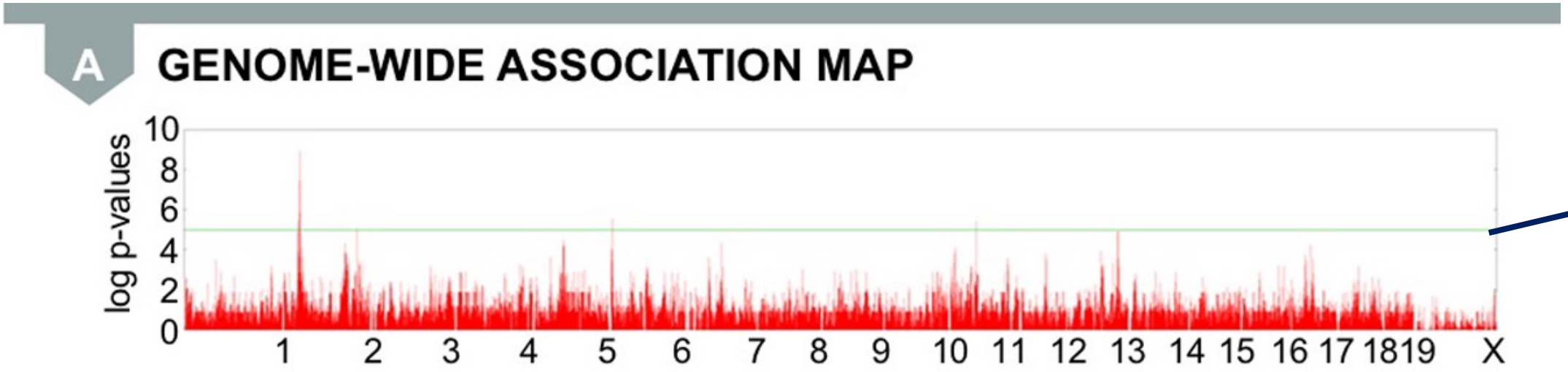
SNP associations may not be due to causal effect

Population structure (differential relatedness among individuals) means that SNPs resulting from Isolation by Descent (IBD) will be associated with the phenotype of interest



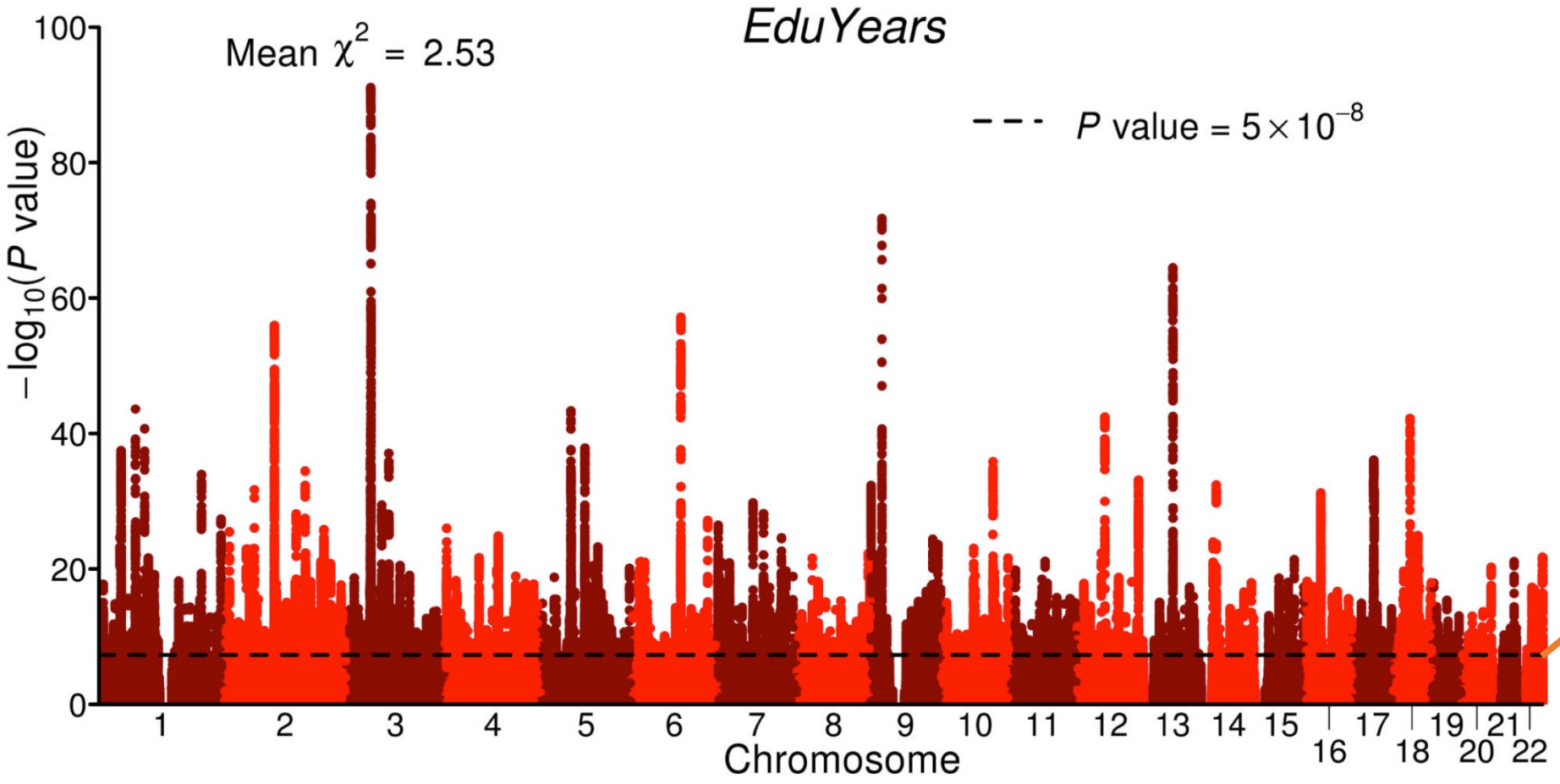
Sul et al 2018

Accounting for population structure can remove spuriously associated SNPs



Sul et al 2018

Doing GWAS with only European populations is a problem

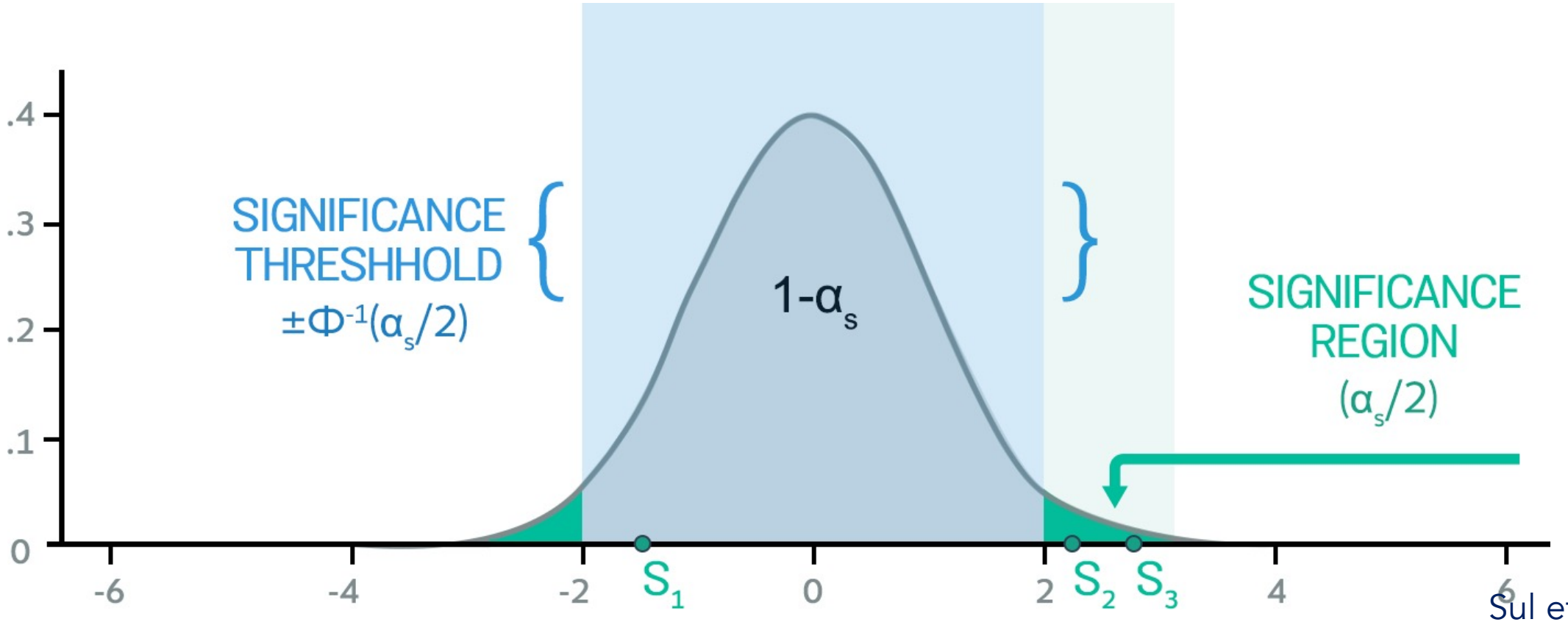


Significance threshold

Lee et al 2018

Determining Statistical Significance

False Discovery Rate (Benjamini & Hochberg 1996)



Sul et al 2018

SNP databases

dbSNP (NCBI)

OMIM (Online Mendelian Inheritance in Man)

HapMap Project

1000 Genomes Project

Performing GWAS – Linking Genotype to Phenotype

Case-control studies

- Compare SNP frequencies in cases vs. controls

Performing GWAS – Linking Genotype to Phenotype

Case-control studies

Bulk segregant analysis

- Separate phenotypes into bins, compare SNP frequencies across bins

Performing GWAS – Linking Genotype to Phenotype

Case-control studies

Bulk segregant analysis

Natural variation

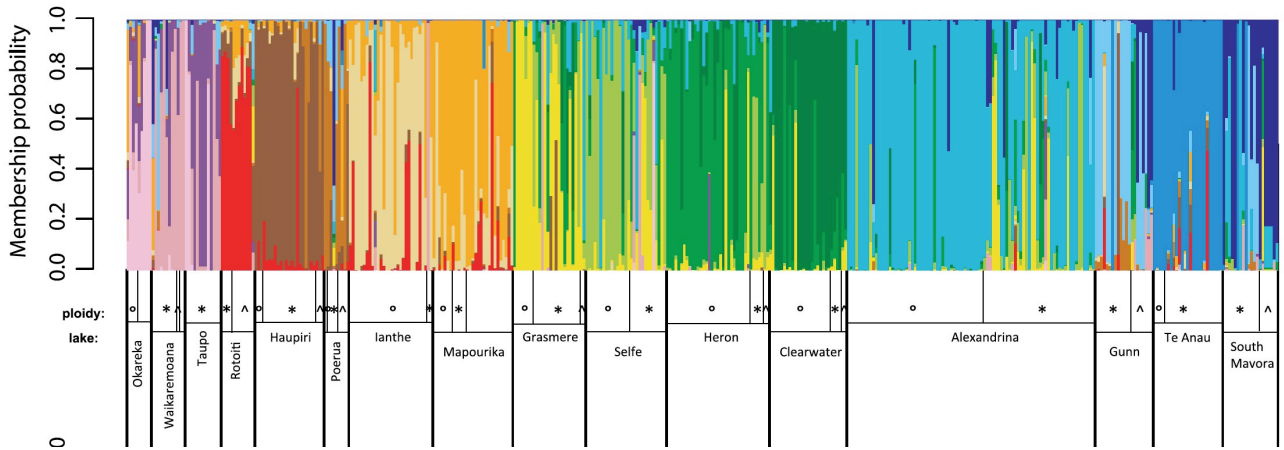
- Multivariate regression of continuous distributions vs. SNP frequencies

Performing GWAS – Key considerations

Problems with the population:

- Hardy-Weinberg Equilibrium
- Population structure

  individuals non-randomly related to other individuals
- Linkage Disequilibrium (LD)
- Missing data



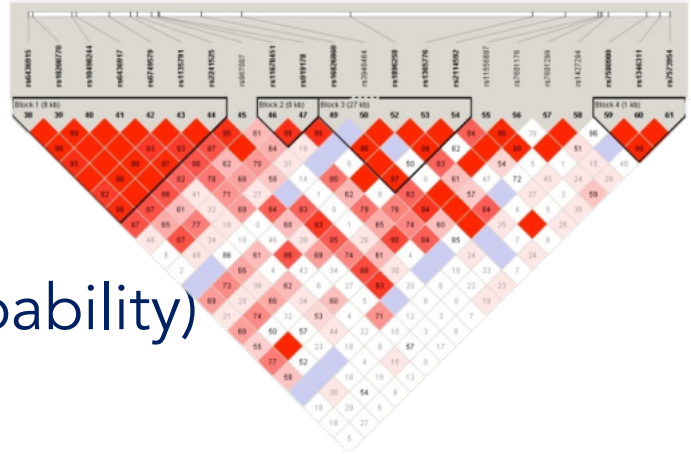
Performing GWAS – Key considerations

Problems with the population:

- Hardy-Weinberg Equilibrium
- Population structure
- Linkage Disequilibrium (LD) –

  Non-random association of two alleles (i.e., > 0.5 probability)

- Missing data



Performing GWAS – Key considerations

Problems with the population:

Problems with the sample:

- Sample size
- Phenotype selection

Performing GWAS – Key considerations

Problems with the population:

Problems with the sample:

Problems with statistical analyses:

- Correcting for multiple comparisons
- False positives
- P-hacking