GWASs

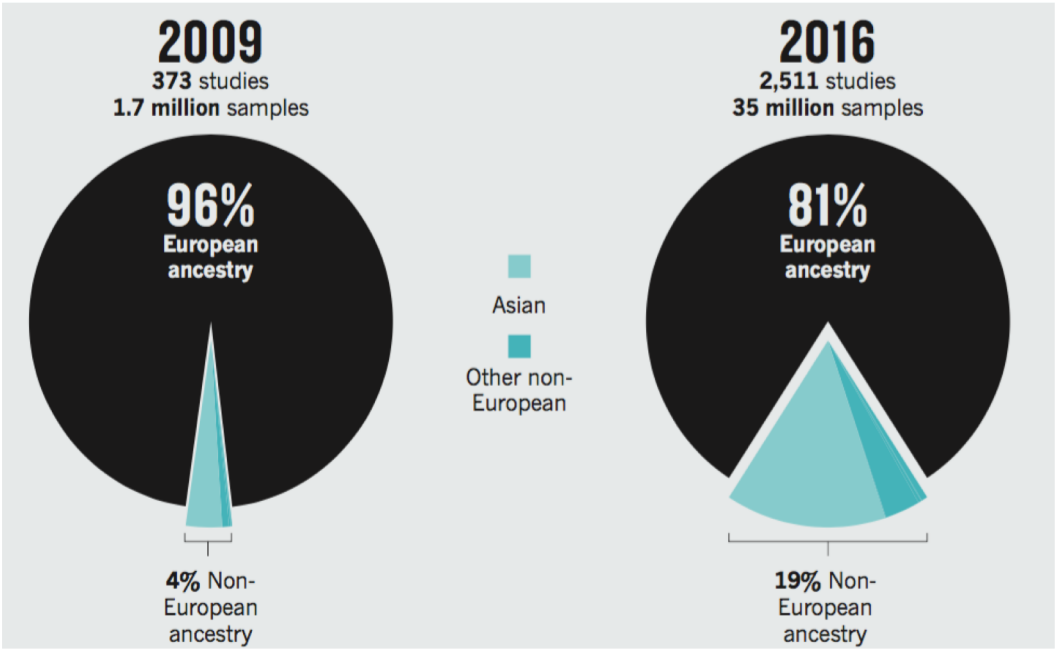
Genome-wide Association Studies

Paucity of non-European data in the field of genomics

CYRUS MCCRIMMON/DENVER POST/GETTY

Certain drugs may be less effective, or even unsafe, in some populations because of genetic differences.

Genomics is failing on diversity

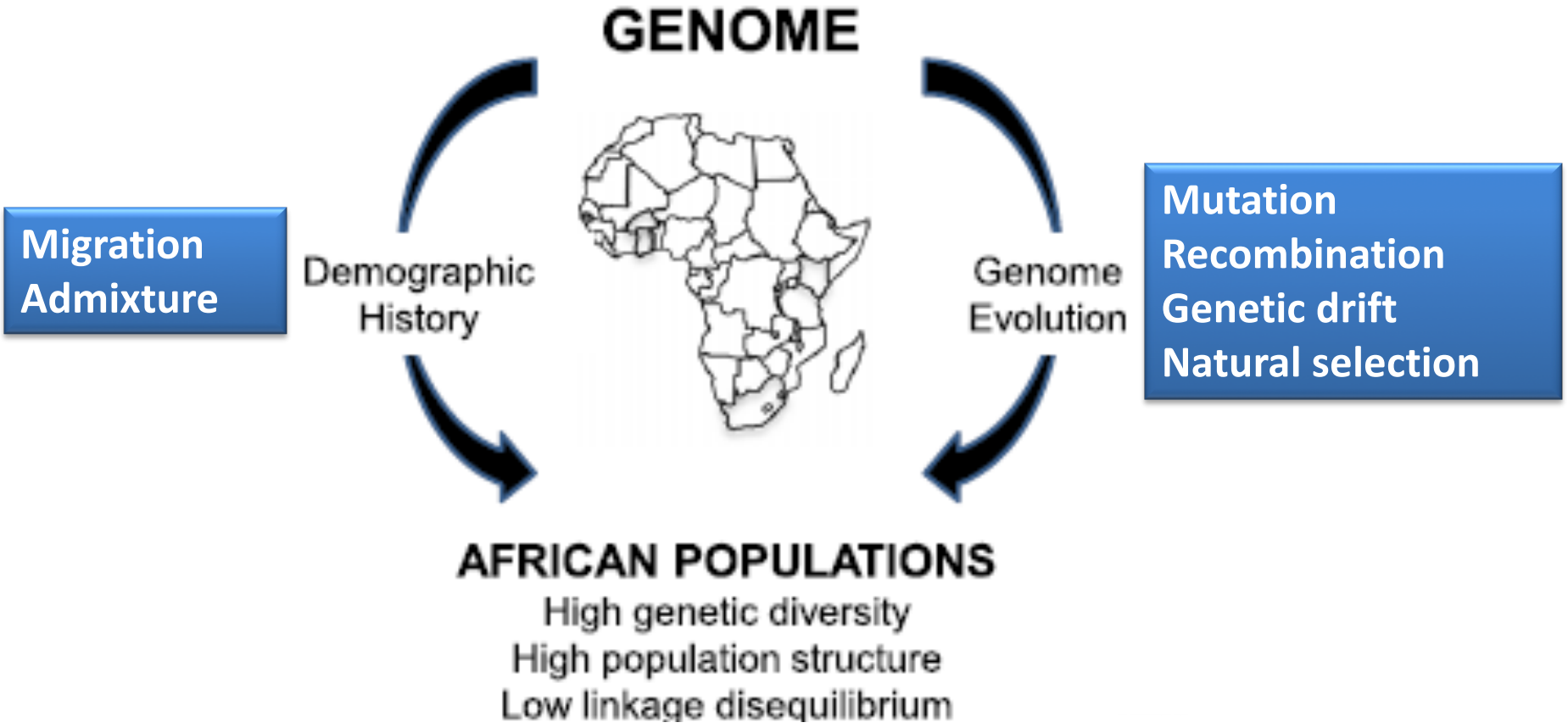


0.57% 3%

African Ancestry
(Majority African-American)

Popejoy and Fullerton.
Nature 538:161, Oct 2016

Shaping population genetic diversity



Migration
Admixture

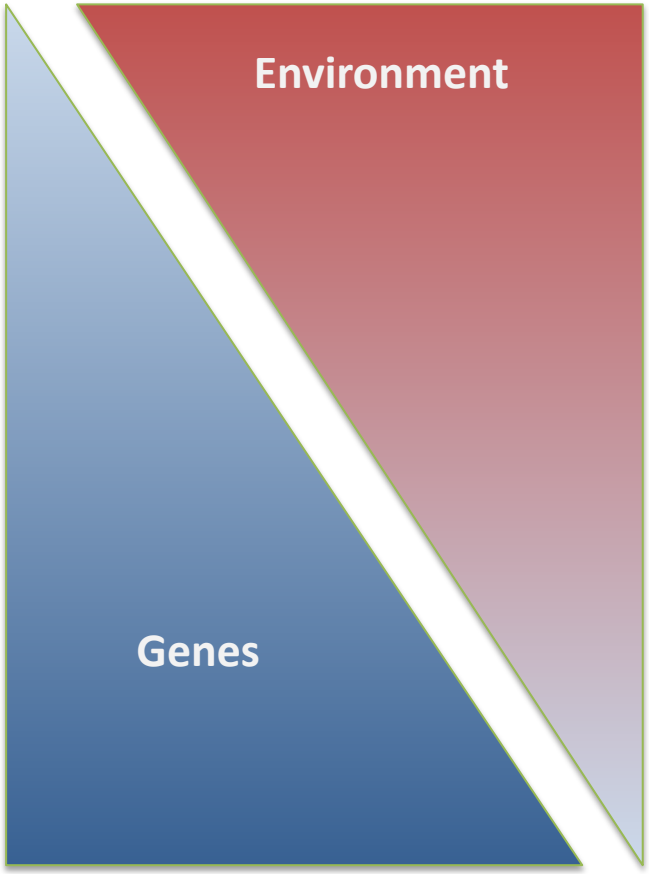
Mutation
Recombination
Genetic drift
Natural selection

Genetic variation contributes to susceptibility for complex traits and to the ability of a populations to adapt to a changing environment

Outline

- Complex traits
- GWAS principles
- Study designs
  - Calculating power
  - Considering population substructure
- African genome structure
- Replication
- Examples

Relative Contributions



Infectious disease (e.g. HIV and TB)

Obesity

Type 2 diabetes

Monogenic diseases (e.g. cystic fibrosis, familial hypercholesterolaemia)

Searching for genetic associations with complex traits

- Conditions that require both a **genetic risk** and **specific environmental triggers** before they manifest
- Complex traits are also referred to as multi-factorial traits, non-communicable diseases (NCD), chronic diseases
  - Examples: diabetes, stroke, asthma, kidney disease
  - Difficult to estimate heritability (genetic contribution to trait)
- Determining genetic risk for a complex trait

GWAS PRINCIPLES

Genome wide association studies (GWASs)

Purpose: To identify genetic associations to complex traits by using **genetic markers** throughout the genome

Genome wide association studies (GWAS)

Purpose: To identify genetic associations to complex traits by using **genetic markers** throughout the genome

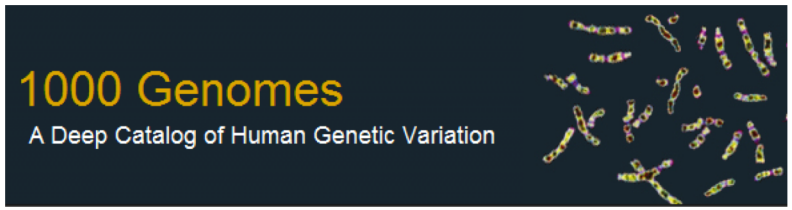
Genome wide association studies (GWAS)

Purpose: To identify genetic associations to complex traits by using **genetic markers** throughout the genome

Understanding patterns of human genome sequence variation

HapMap

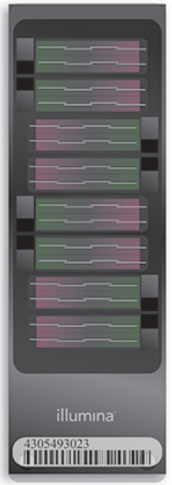
1000 Genomes



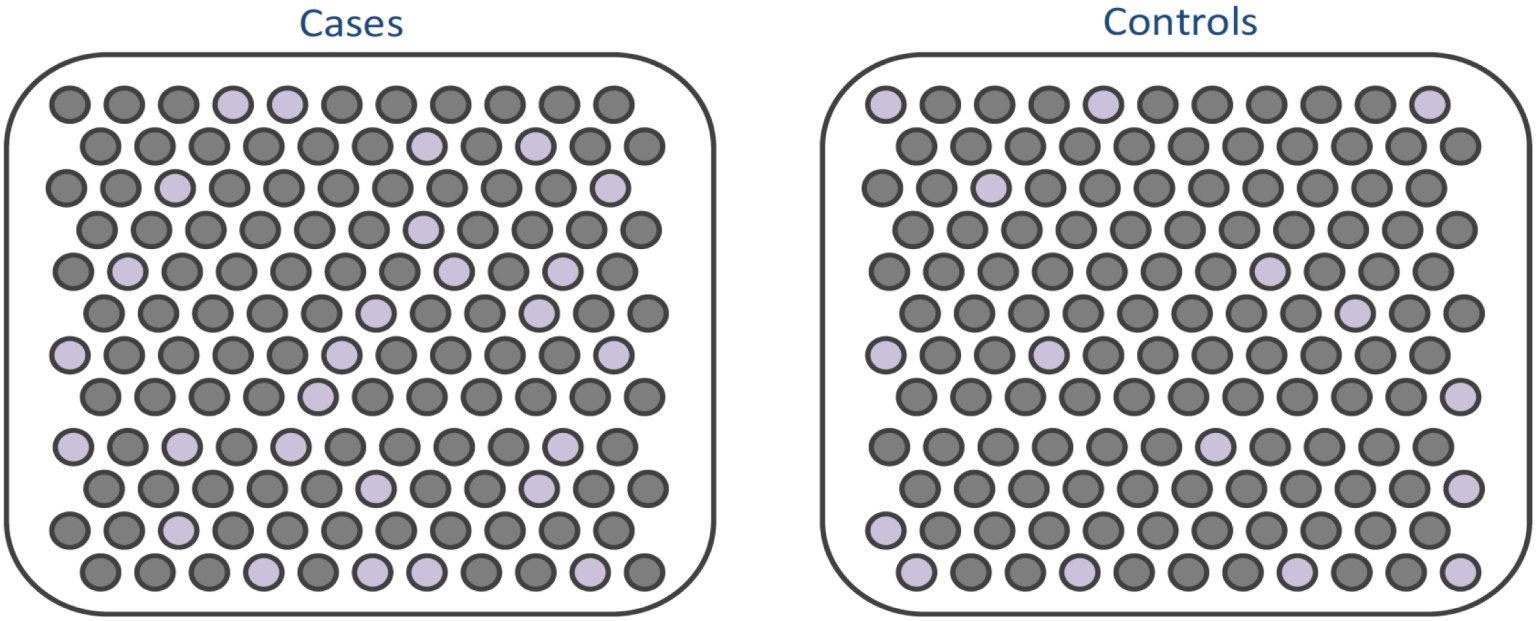
High throughput genotyping technologies

Illumina

Affymetrix



Association between genetic variants and disease



Variant allele frequency 25%

Variant allele frequency 15%

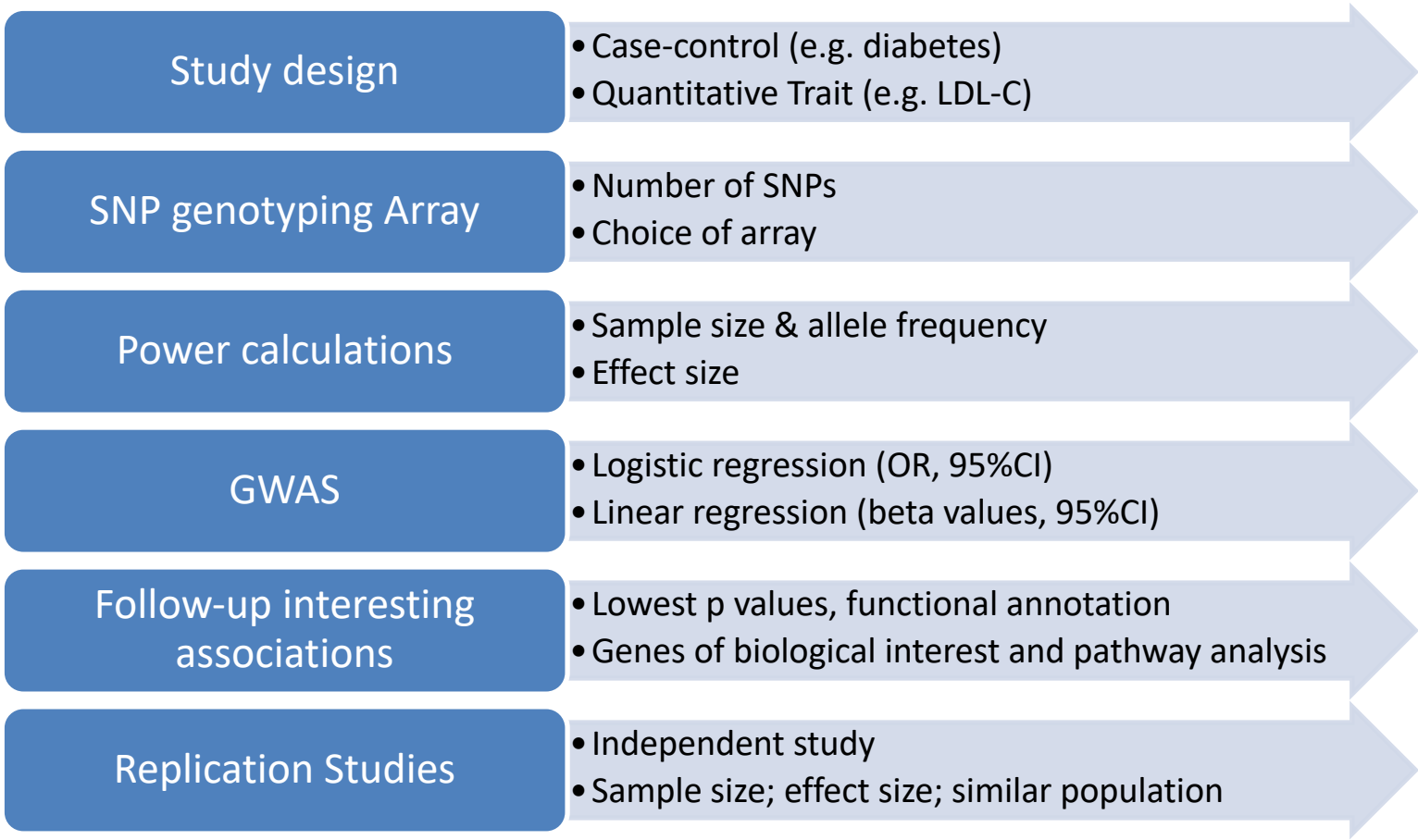
Allelic association – single SNV

Genotype association – Model for mode of inheritance in terms of the genotype effect (recessive, dominant, codominant, multiplicative)

GWAS by Linkage Disequilibrium

Stage of analysis	European pops	African pops
Detecting association	High LD increases chance of detecting associations	Low LD reduces likelihood
Replicating association	Good chance of replicating even if causal variant not typed	Reduced likelihood unless causal variant directly typed
Localising causal variant	Can be difficult because of high LD	May be easier because of low LD

GWAS workflow



GWAS Catalogue

The NHGRI-EBI Catalogue of published genome-wide association studies

As of 2018-05-29, the GWAS Catalogue contains 3395 publications and 62156 unique SNP-trait associations. GWAS Catalogue data is currently mapped to Genome Assembly GRCh38.p12 and dbSNP Build 150.

STUDY DESIGN

GWAS

- Usually non-hypothesis based (exploratory research)
- Examine genetic **associations** (genetic markers e.g. SNPs throughout the genome) **with a phenotype**
- Look for highly significant associations (multiple testing problem)
- Associations seldom causal factors (in linkage disequilibrium – they segregate together) (direct vs indirect association)
- Sample size:
  - **Small:** Will miss important genetic determinants that have a minor effect on the phenotype
  - **Large:** Will be powered to detect small effects

Study designs

- **Quantitative traits:** Association across a continuous phenotype spectrum (e.g. Height, lipid levels, blood pressure)
  - Effect measured as a beta value
- **Case-Control studies:** Groups of individuals dichotomised. Cases vs Controls (e.g. Diabetes, hypertension). Need discrete cut off points for cases.
  - Effect measured as an Odds Ratio (OR) e.g. for allelic association
  - OR=1 no effect
  - OR=1.1 small effect
  - OR=2 larger effect
  - OR<1 variant lowers risk

Genetic architecture of complex traits

- Percentage of phenotype variation explained by genetic susceptibility
- Variant frequency
  - Common variants
  - Rare variants
- Linkage disequilibrium
  - Direct of indirect association
- Contribution to the trait
  - Small effects (many contributing variants)
  - Large (fewer contributing variants each explaining more of the phenotype
  - Combination of a few core high effect variants and many small effect variants

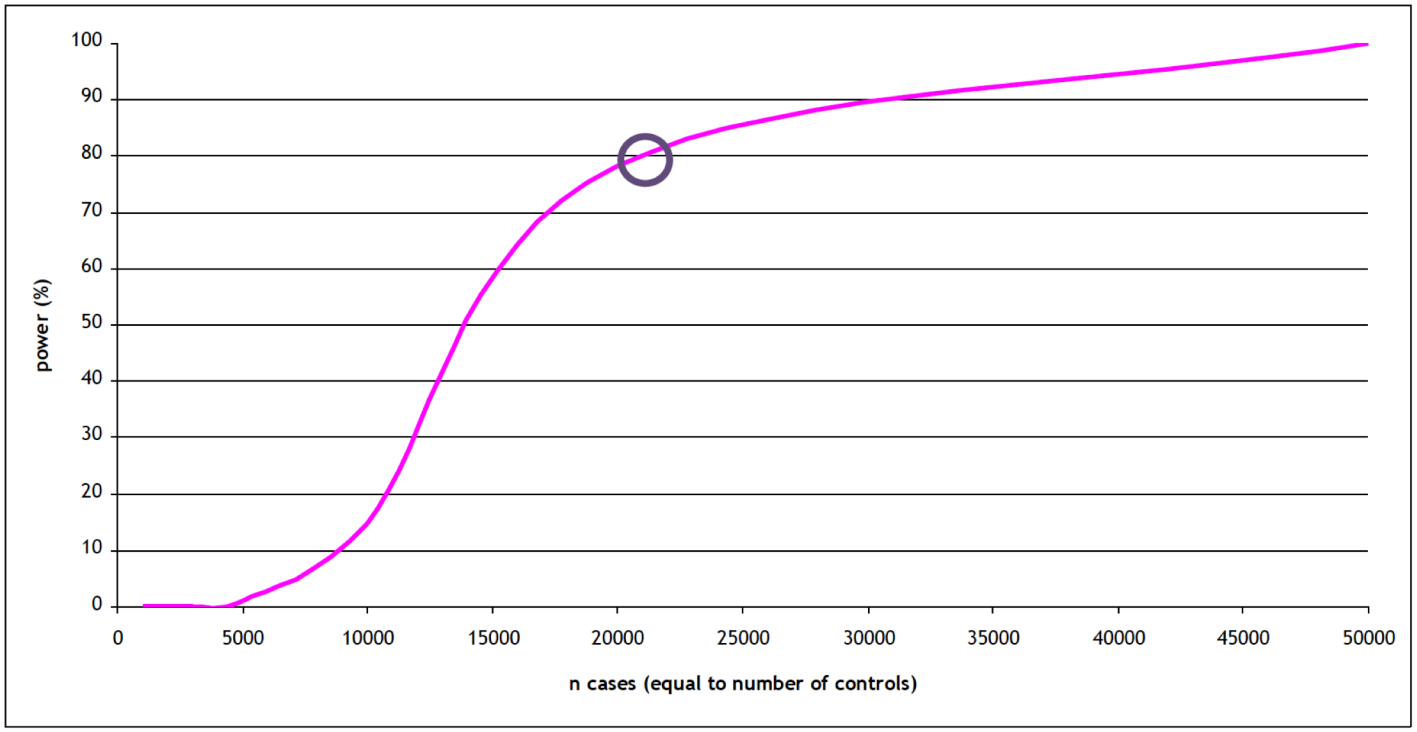
Calculating the power of a GWAS

- Power: The statistical likelihood (probability) of detecting genuine associations
- Several factors influence power
  - **Sample size**
  - **Effect size (Odds ratio for case-control studies)**
  - **Allele frequency**
- You would like your study to be at least 80% powered to detect an association

Power: Case – Control Study

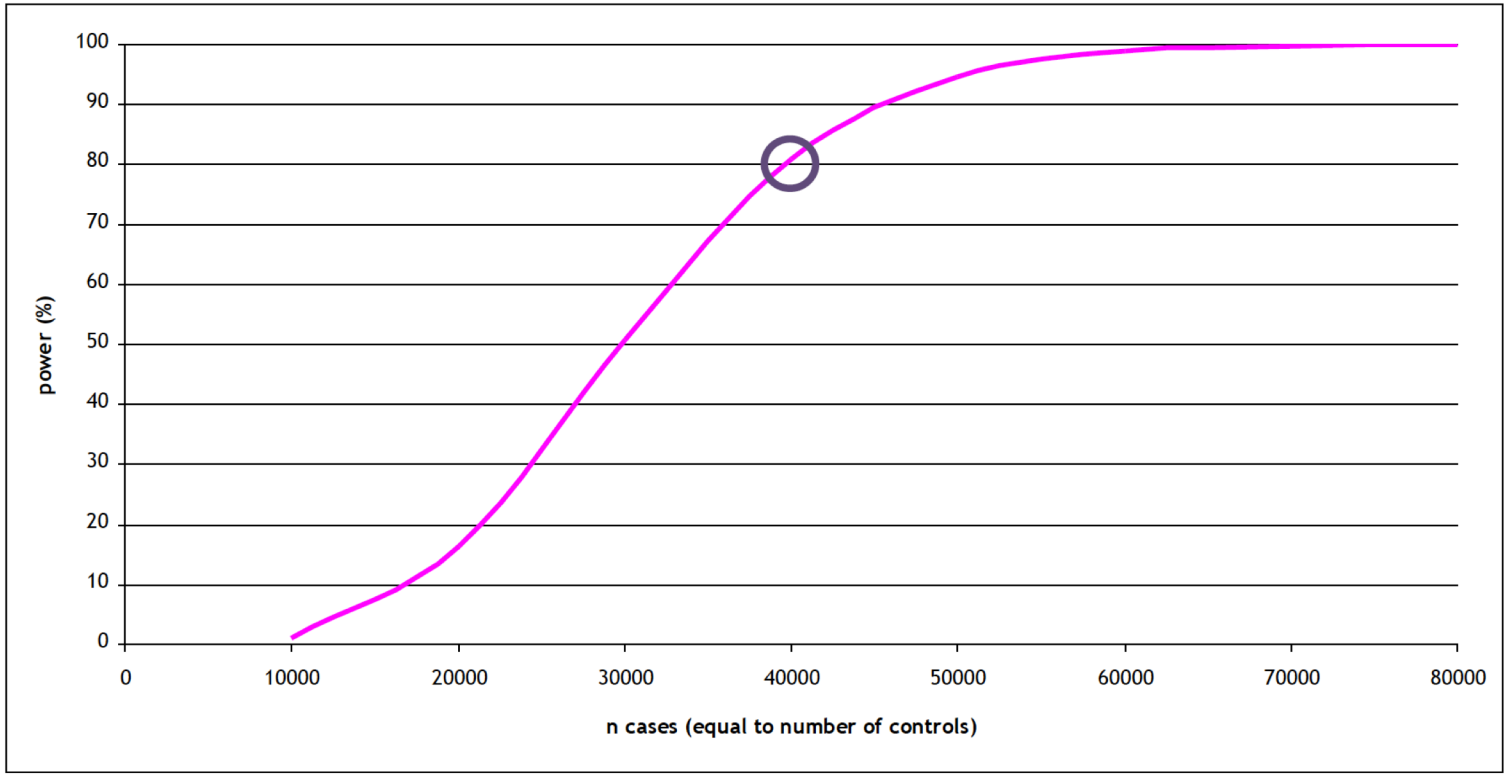
Sample size matters

Power to detect association ($p=5x10^{-8}$) at a variant with risk allele frequency 0.30 and allelic OR 1.10



Slide from Ele Zeggini – June 2018

Power to detect association ($p=5x10^{-8}$) at a variant with risk allele frequency 0.005 and allelic OR 1.50

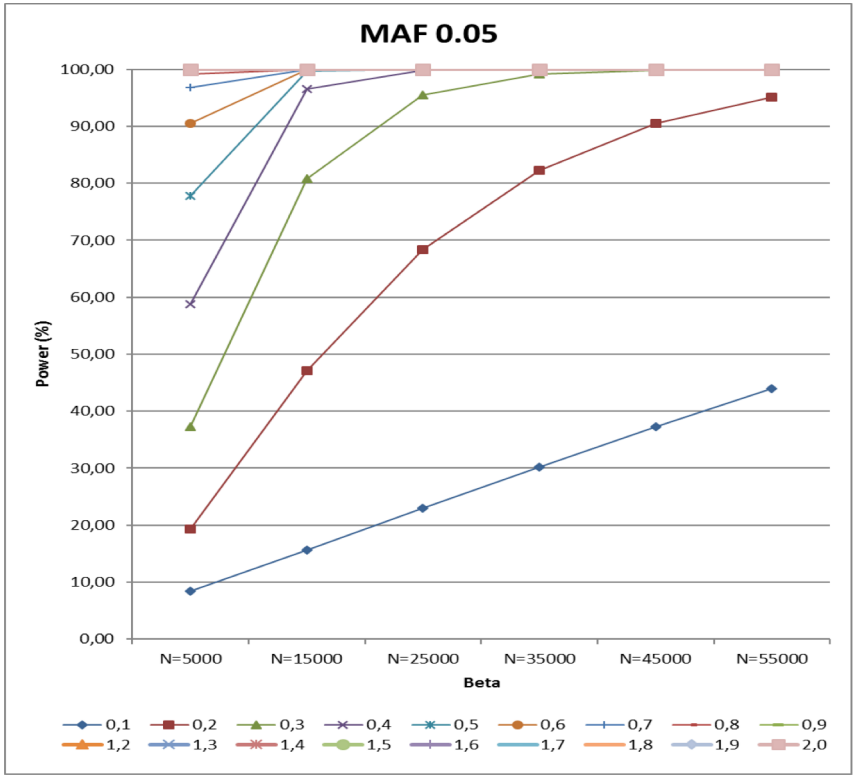


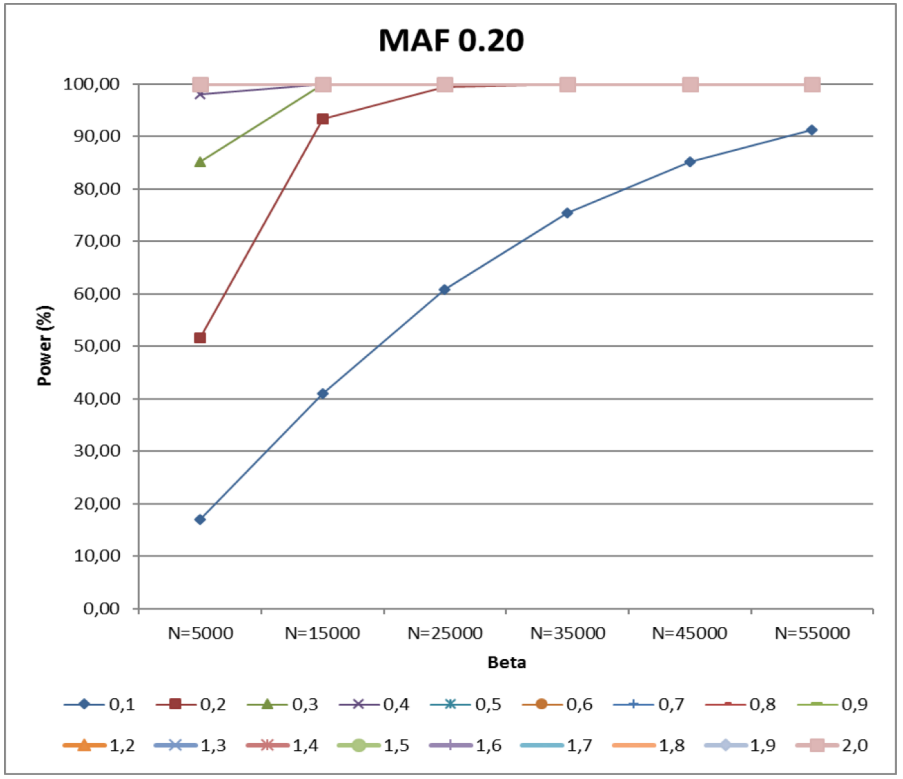
Slide from Ele Zeggini – June 2018

Study Power - Quantitative Traits

Important factors:

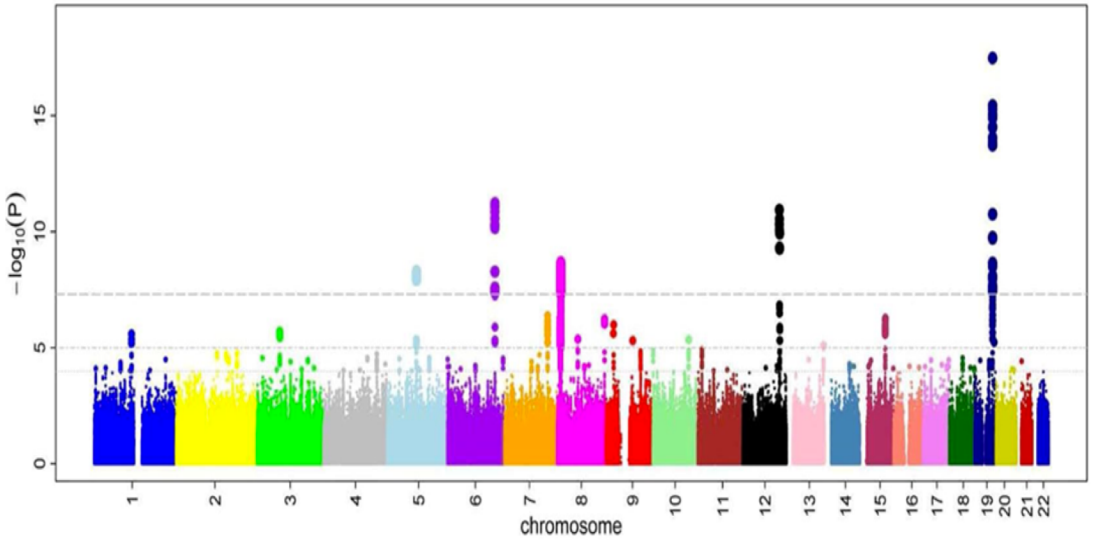
1. Sample size 2. Allele frequency and 3. Effect size



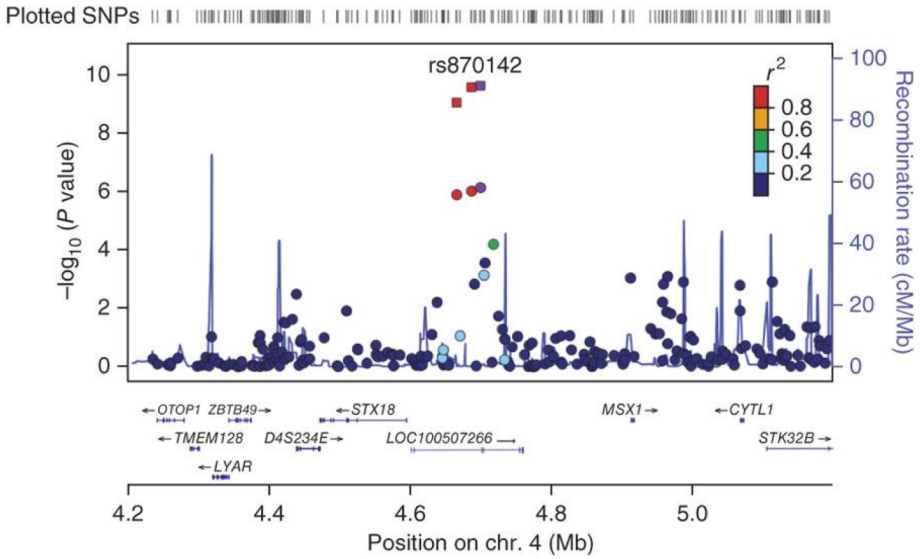


Visualisation of GWAS outcomes

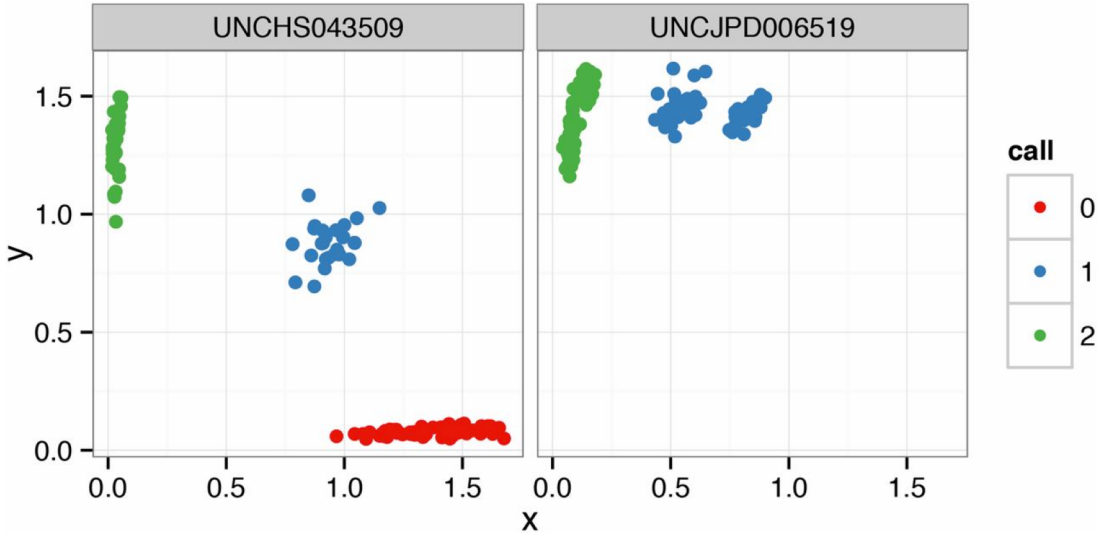
- Manhattan Plots



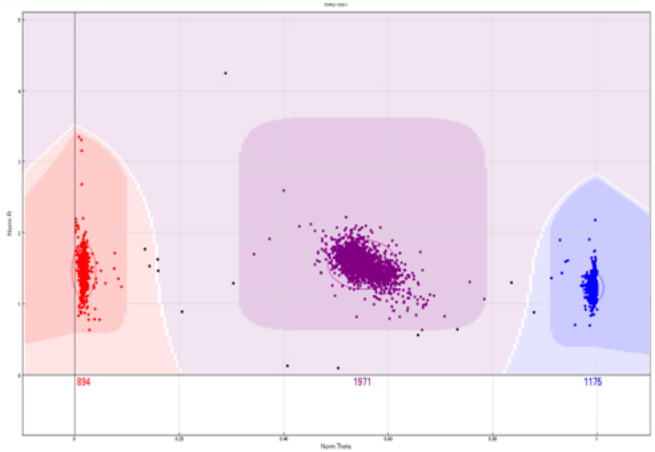
- Locus zoom plots



Ensure that associated variant has good genotype clustering (Illumina)



rs401681 – Illumina 550K



What expectations could we have?

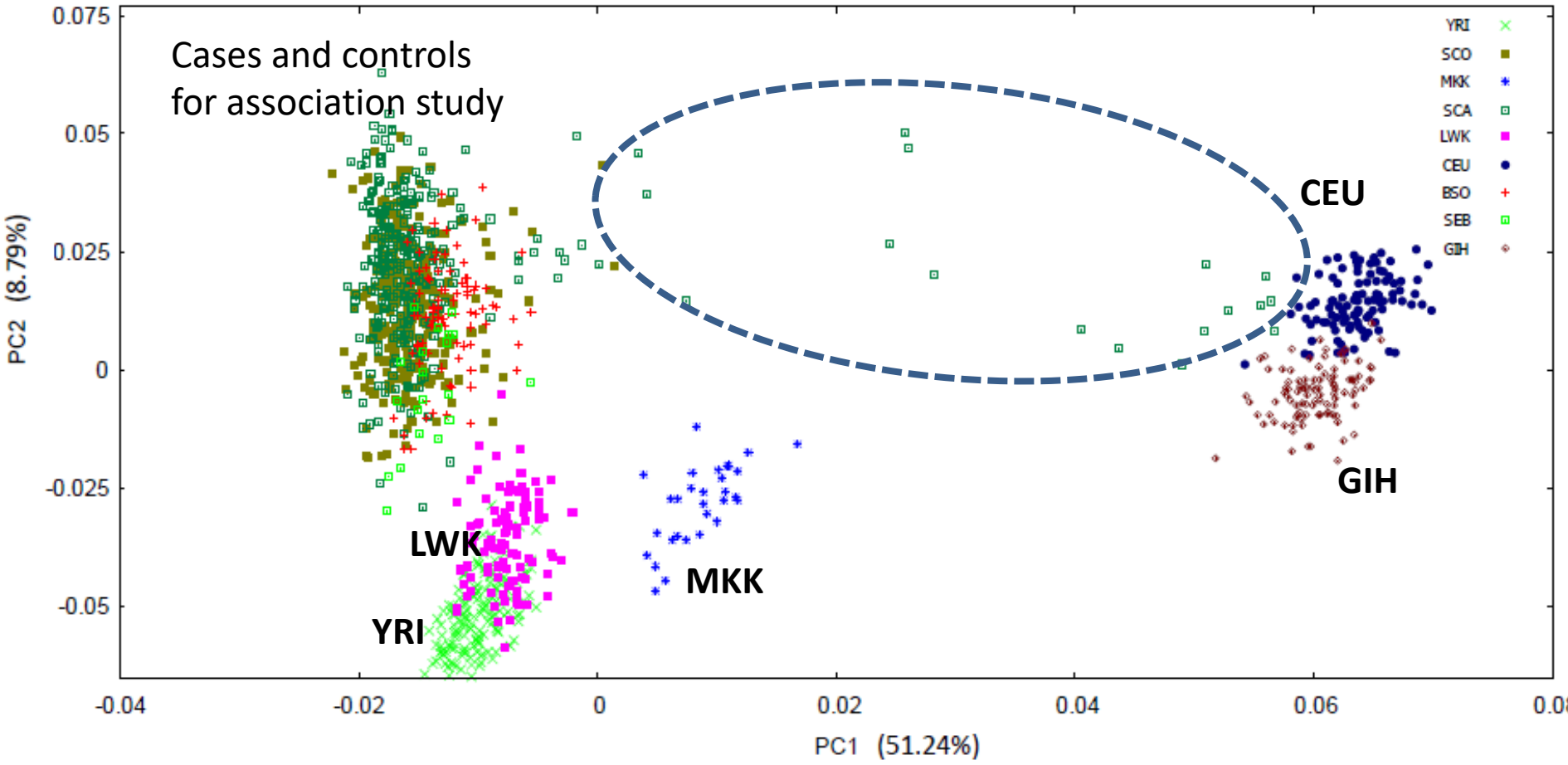
- With modest sample size we can examine previous associations and test for similar effects
- With modest sample size we can detect novel large effect associations
- With large sample sizes we can discover novel modest to small effect associations
- Need to explore insights into the biology and impact that African studies can bring to science

AFRICAN GENOME STRUCTURE

How do we take into consideration ethnic differences in genetic association studies and why does it matter?

- Population stratification
- Same SNPs can have different effect sizes in different populations (or associate in one, but not another population)
- Most GWAS arrays have more common SNPs for European populations (H3Africa array more African appropriate)
- High false discovery rate if you do not correct for population structure
- Different chromosomal backgrounds in a study population can influence ability to detect associations
- Imputation is a handy tool to extend data
- Advantages to studying African populations

Population stratification

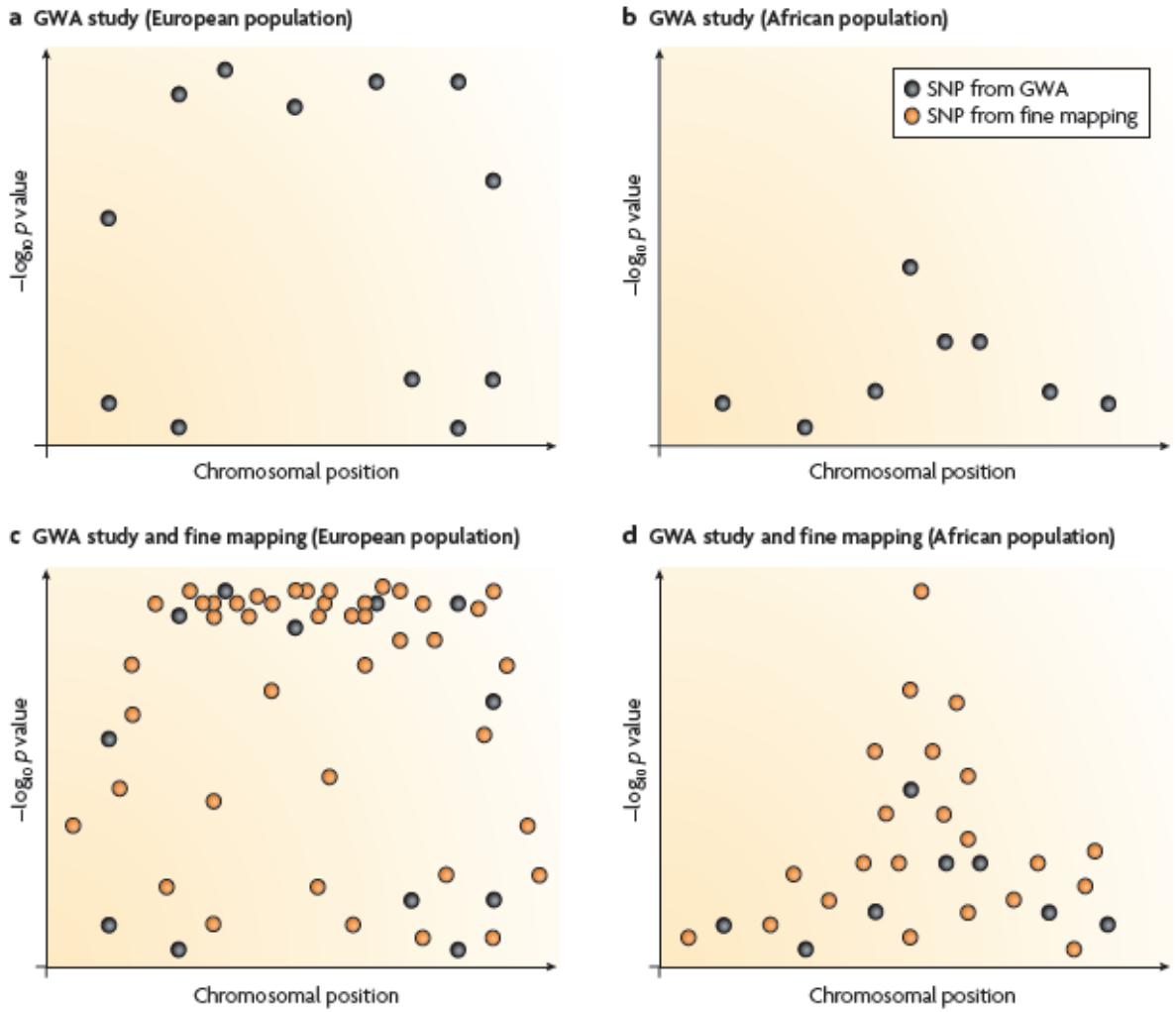


Advantage of studying African populations

European | **African**

GWAS

Fine mapping



African populations have lower linkage disequilibrium (LD) and higher haplotype diversity

Teo et al. (2010) Nature Reviews, Genetics

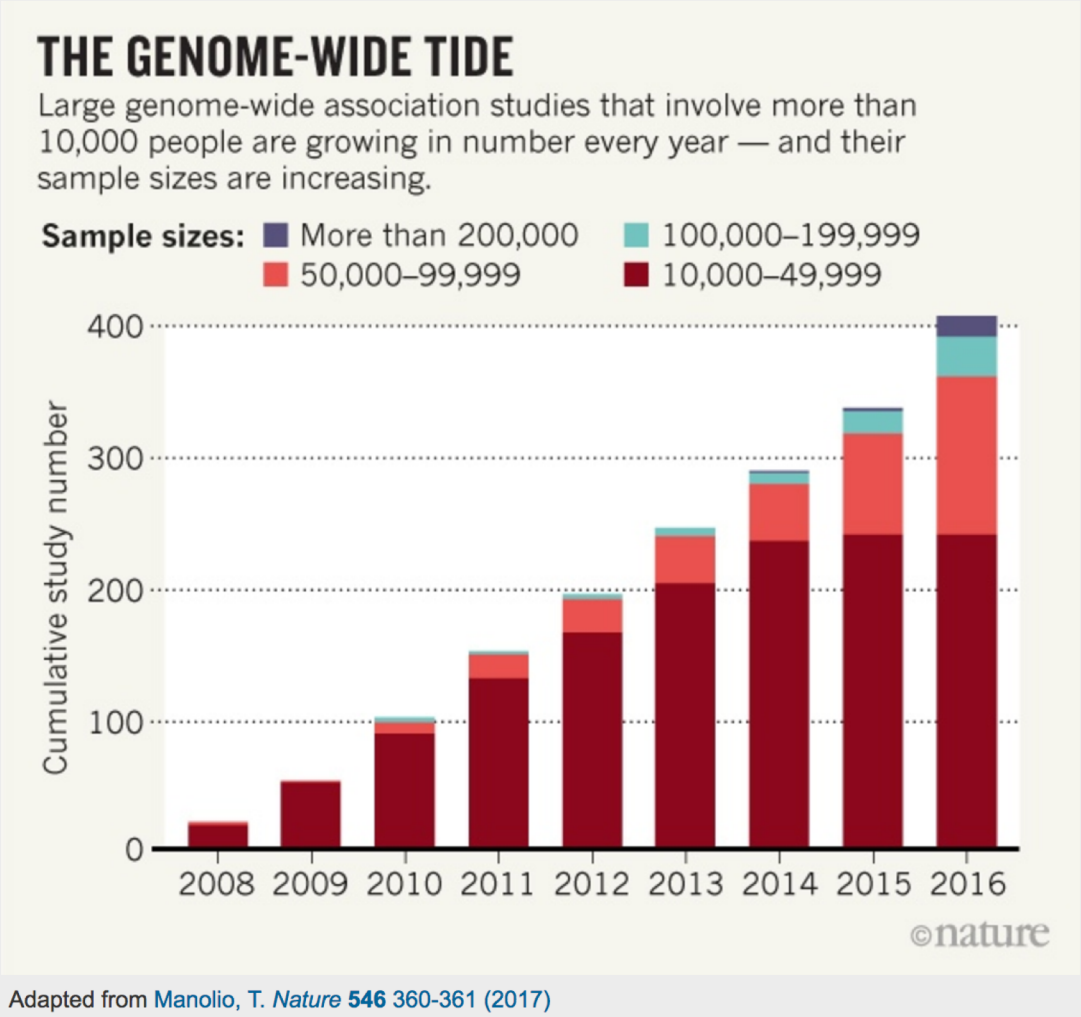
REPLICATION

Replication

- Additional independent sample (dataset)
- Sufficient size
- Identical phenotype
- Look for similar effect
  - Same genomic region not necessarily same SNP (examine LD Block)
  - Same direction

Teri Manolio (Nature 2017)

"…as with many things in science, the more we know, the more we have to learn."



Adapted from Manolio, T. *Nature* **546** 360-361 (2017)

GWAS

1. Explain only a small fraction of the heritability of a trait
2. Most associations in regions of the genome with no known function

Nature 546, 360–361
(15 June 2017)
doi:10.1038/546360a

- Modest study – association with large effect variant – obesity among Samoans
- Very large study – UK Biobank and blood pressure
- Integrating the science – review on biology of obesity

EXAMPLES

Example of modest genomic study of obesity with a high effect variant and biological insights

Nature Genetics (Sept 2016) 48(9): 1049-1054

A thrifty variant in *CREBRF* strongly influences body mass index in Samoans

Ryan L Minster[#1], Nicola L Hawley[#2], Chi-Ting Su[#1,12], Guangyun Sun[#3], Erin E Kershaw[4], Hong Cheng[3], Olive D Buhule[5,12], Jerome Lin[1], Muagututi'a Sefuiva Reupena[6], Satupa'itea Viali[7], John Tuitele[8], Take Naseri[9], Zsolt Urban[1,14], Ranjan Deka[3,14], Daniel E Weeks[1,5,14], and Stephen T McGarvey[10,11,14]

Founder effect in Samoa for high levels of obesity

GWAS: 3,072 participants in discovery study
2,102 in replication study

Highly associated *GREBRF* missense variant p.Arg457Gln ($p=1.4 \times 10^{-20}$) in meta-analysis

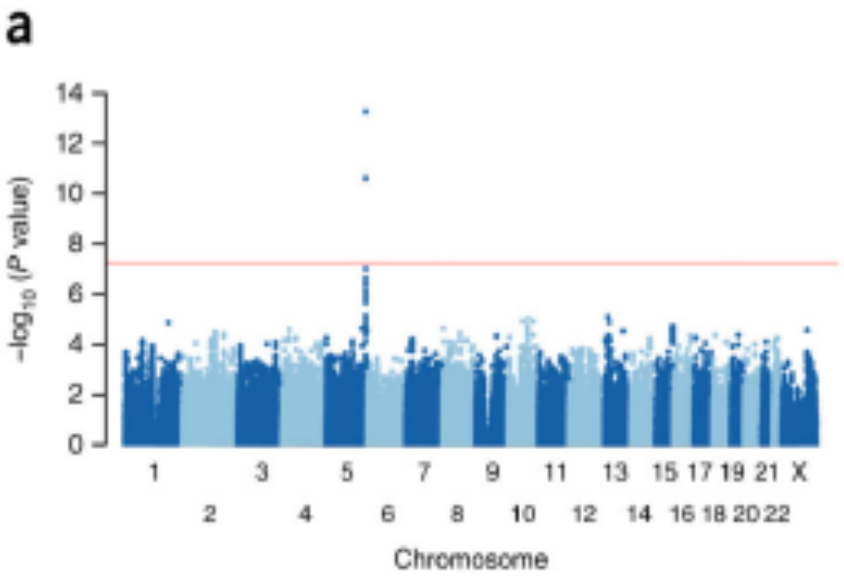
Frequency of associated variant 0.295 in Samoa (very rare elsewhere)

Effect size: 1.36 to 1.45 kg/m^2 (BMI) per copy of the risk allele

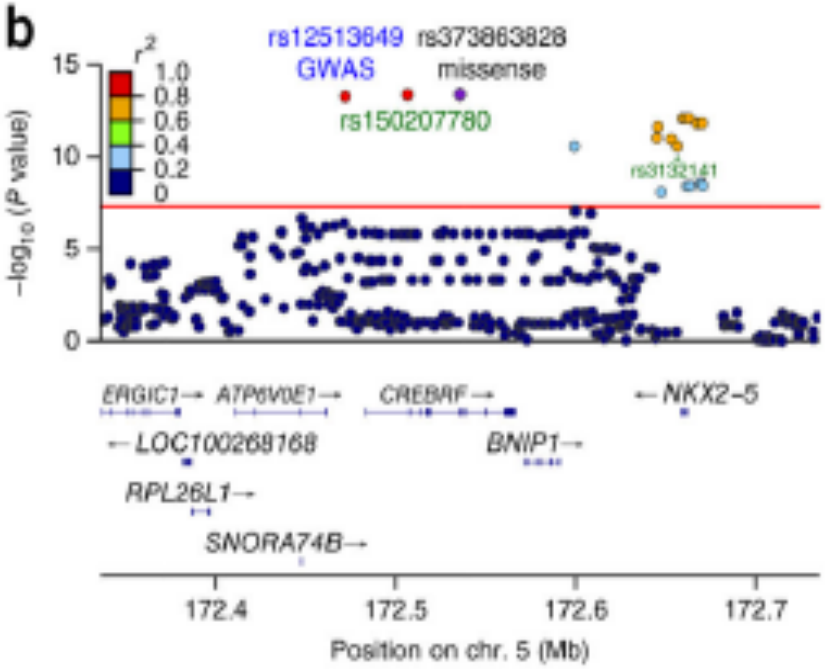
Biological insight: Adipose cell model shows that when the variant is overexpressed, it selectively decreases energy use and increases fat storage

Conclusion: Supports selection of the allele through the "thrifty" variant hypothesis

GWAS



Manhattan Plot

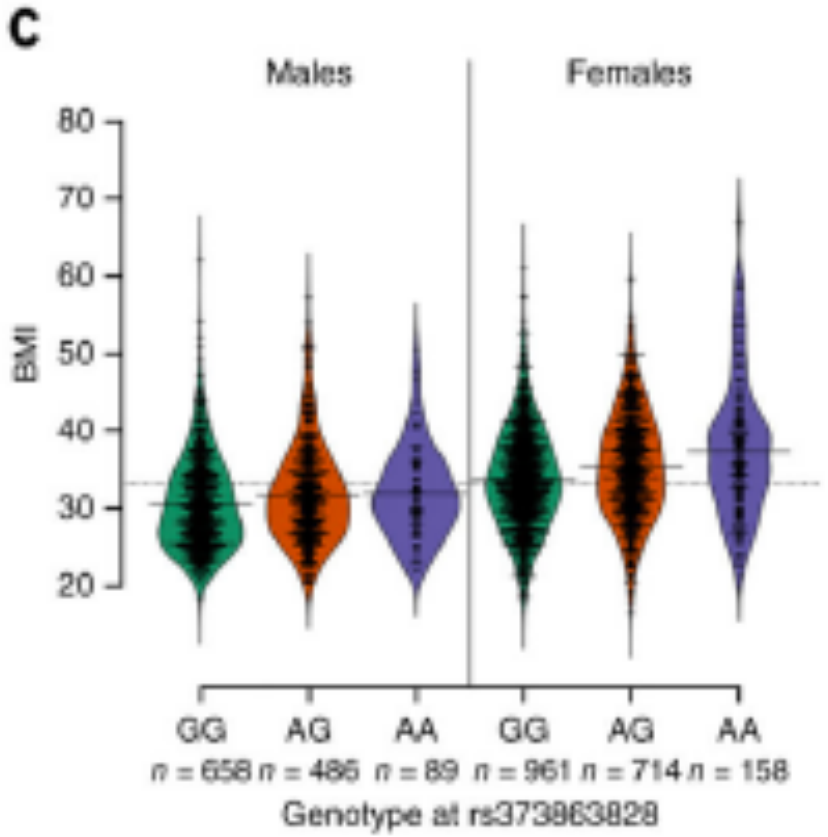


Following Imputation

GREBRF missense variant p.Arg457Gln ($p=1.4x\ 10^{-20}$)

Nature Genetics. (Sept 2016) 48(9): 1049-1054

Effect on obesity



Sex-specific effects are important

How predictive are the alleles?

Novel insights into obesity

Data from discovery cohort: 1233 men and 1833 women

Nature Genetics. (Sept 2016) 48(9): 1049-1054

Example of a large genomic study of a complex trait and the nature of the results

Genome-wide association analysis identifies novel blood pressure loci and offers biological insights into cardiovascular risk

NATURE GENETICS VOLUME 49 | NUMBER 3 | MARCH 2017

Pages 403-415 and then online methods

Helen R Warren[1,2,60], Evangelos Evangelou[3,4,60], Claudia P Cabrera[1,2,60], He Gao[3,5,60], Meixia Ren[1,2,60], Borbala Mifsud[1,60], Ioanna Ntalla[1], Praveen Surendran[6], Chunyu Liu[7–9], James P Cook[10], Aldi T Kraja[11], Fotios Drenos[12,13], Marie Loh[3,14], Niek Verweij[15–18], Jonathan Marten[19], Ibrahim Karaman[3], Marcelo P Segura Lepe[3,20], Paul F O'Reilly[21], Joanne Knight[22], Harold Snieder[23], Norihiro Kato[24], Jiang He[25], E Shyong Tai[26,27], M Abdullah Said[15], David Porteous[28], Maris Alver[29], Neil Poulter[30], Martin Farrall[31], Ron T Gansevoort[32], Sandosh Padmanabhan[33], Reedik Mägi[29], Alice Stanton[34], John Connell[35], Stephan J L Bakker[36], Andres Metspalu[29], Denis C Shields[37], Simon Thom[38], Morris Brown[1,2], Peter Sever[38], Tõnu Esko[16,29], Caroline Hayward[19], Pim van der Harst[15], Danish Saleheen[39–41], Rajiv Chowdhury[6], John C Chambers[3,42–44], Daniel I Chasman[45,46], Aravinda Chakravarti[47], Christopher Newton-Cheh[16–18], Cecilia M Lindgren[16,48,49], Daniel Levy[7,9], Jaspal S Kooner[43,50,51], Bernard Keavney[52,53], Maciej Tomaszewski[52,53], Nilesh J Samani[54,55], Joanna M M Howson[6], Martin D Tobin[56], Patricia B Munroe[1,2], Georg B Ehret[47,57], Louise V Wain[56], The International Consortium of Blood Pressure (ICBP) 1000G Analyses[58], The CHD Exome+ Consortium[59], The ExomeBP Consortium[59], The T2D-GENES Consortium[59], The GoT2DGenes Consortium[59], The Cohorts for Heart and Ageing Research in Genome Epidemiology (CHARGE) BP Exome Consortium[59], The International Genomics of Blood Pressure (iGEN-BP) Consortium[59], Michael R Barnes[1,2,60], Ioanna Tzoulaki[3–5,60], Mark J Caulfield[1,2,60] & Paul Elliott[3,5,60] for The UK Biobank CardioMetabolic Consortium BP working group

The International Consortium of Blood Pressure (ICBP) 1000G Analyses

Louise V Wain[56], Ahmad Vaez[23,61], Rick Jansen[62], Roby Joehanes[9,63], Peter J van der Most[23], A Mesut Erzurumluoglu[56], Paul O'Reilly[21], Claudia P Cabrera[1,2], Helen R Warren[1,2], Lynda M Rose[45], Germaine C Verwoert[64], Jouke-Jan Hottenga[65], Rona J Strawbridge[66,67], Tonu Esko[29,68,69], Dan E Arking[47], Shih-Jen Hwang[70,71], Xiuqing Guo[72], Zoltan Kutalik[73,74], Stella Trompet[75,76], Nick Shrine[56], Alexander Teumer[77,78], Janina S Ried[79], Joshua C Bis[80], Albert V Smith[81,82], Najaf Amin[83], Ilja M Nolte[23], Leo-Pekka Lyytikäinen[84,85], Anubha Mahajan[48], Nicholas J Wareham[86], Edith Hofer[87,88], Peter K Joshi[89], Kati Kristiansson[90], Michela Traglia[91], Aki S Havulinna[90], Anuj Goel[48,92], Mike A Nalls[93,94], Siim Sõber[95], Dragana Vuckovic[96,97], Jian'an Luan[86],

Study design

- GWAS with quantitative trait – blood pressure (BP) – three measures
  - Systolic blood pressure (SBP)
  - Diastolic blood pressure (DBP)
  - Pulse pressure (PP)
- Not with hypertension (>140/90 mmHG)
- BP strong, heritable and modifiable driver of risk for stroke and coronary artery disease
- To date many associated loci (120) – common variants and small effects

Discovery cohort: ~140 000 people from UK Biobank with at least 2 sitting BP measurements.

Analysis: Single variant linear regression under an additive model

Number of SNVs after imputation: ~9.8 million SNVs with MAF >0.1

P value cut off: $p<10^{-6}$

Replication: 240 loci
Cohorts: 2 large BP consortia
Criteria: $p<5 \times 10^{-8}$ for genome-wide significant of combined replication and discovery set (meta-analysis)
Cut off for replication only of previous associations and same direction $p<0.01$

Results

107 loci validated at $p < 5 \times 10^{-8}$
- 32/107 were novel associations
- 75/107 in other study on BP using UK Biobank
- 53/75 were validated for the first time

Associations with 107 loci
- 24 primarily with SBP
- 41 primarily with DBP
- 42 primarily with PP

All together: Validated loci increased percentage of trait variance explained by about 1% (e.g. increased to 3.56% for SBP)

Genetic Risk Score (GRS)

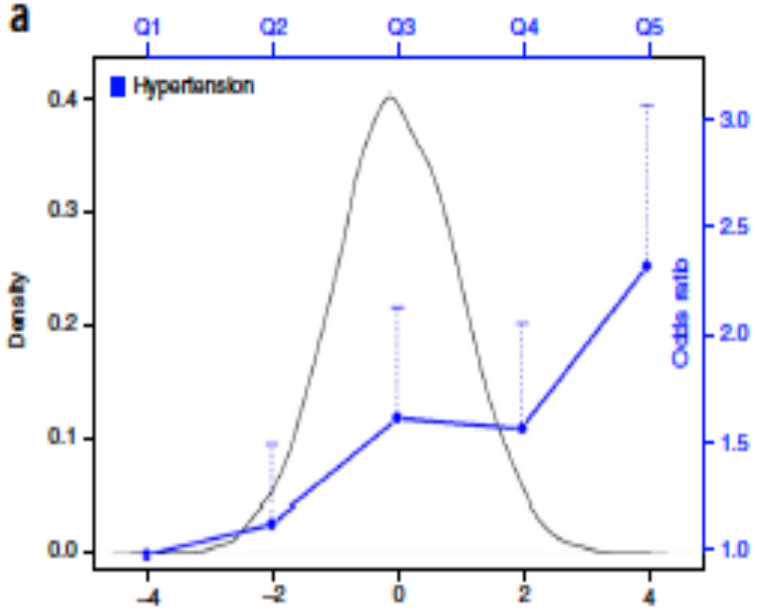
Divided individuals into Quintiles
(Q1, Q2, Q3, Q4, Q5) (only >50 year olds)

Odds ratio (OR) for association with hypertension (sex adjusted)

For individuals >50 years, adjusted for sex, the highest quintile had a SBP 9.3 mmHg higher than those in the lowest quintile.

OR was 2.32 for hypertension (HT) compared to the lowest quintile (HT > 140/90)

a



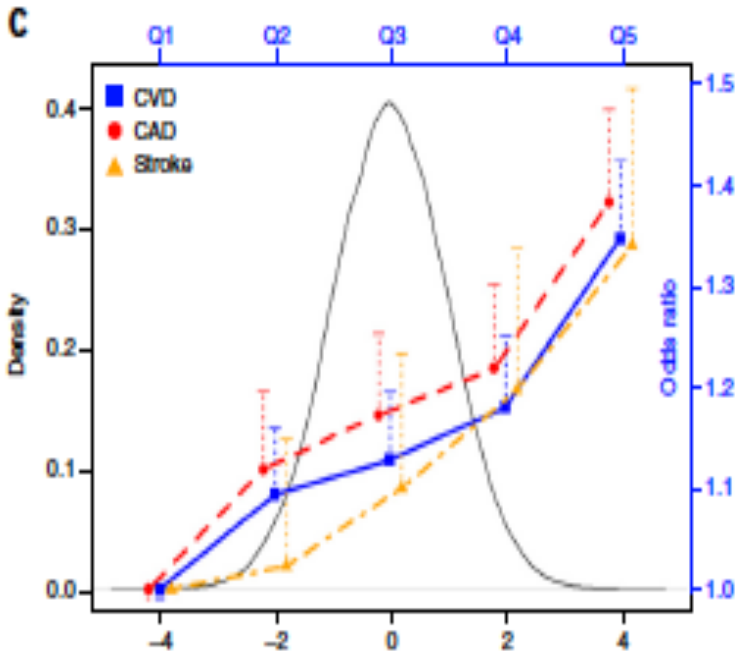
b

(mm Hg)	Q1	Q2	Q3	Q4	Q5
SBP	136.4 (17.2)	138.7 (16.5)	142.5 (18.8)	142.5 (19.4)	145.7 (19.7)
DBP	82.9 (9.9)	84.3 (10.3)	85.9 (11.1)	86.0 (10.9)	87.6 (11.9)
PP	53.4 (11.6)	54.4 (11.3)	56.7 (12.3)	56.5 (12.4)	58.2 (12.4)

Genetic Risk Score for CVD, CAD and Stroke

Modest ORs for difference between lowest and highest quintiles

Stroke OR = 1.34
CAD OR = 1.35



d

(count)	Q1	Q2	Q3	Q4	Q5
CVD	2,462	2,689	2,759	2,874	3,229
CAD	1,783	1,996	2,078	2,154	2,417
Stroke	581	597	640	695	776

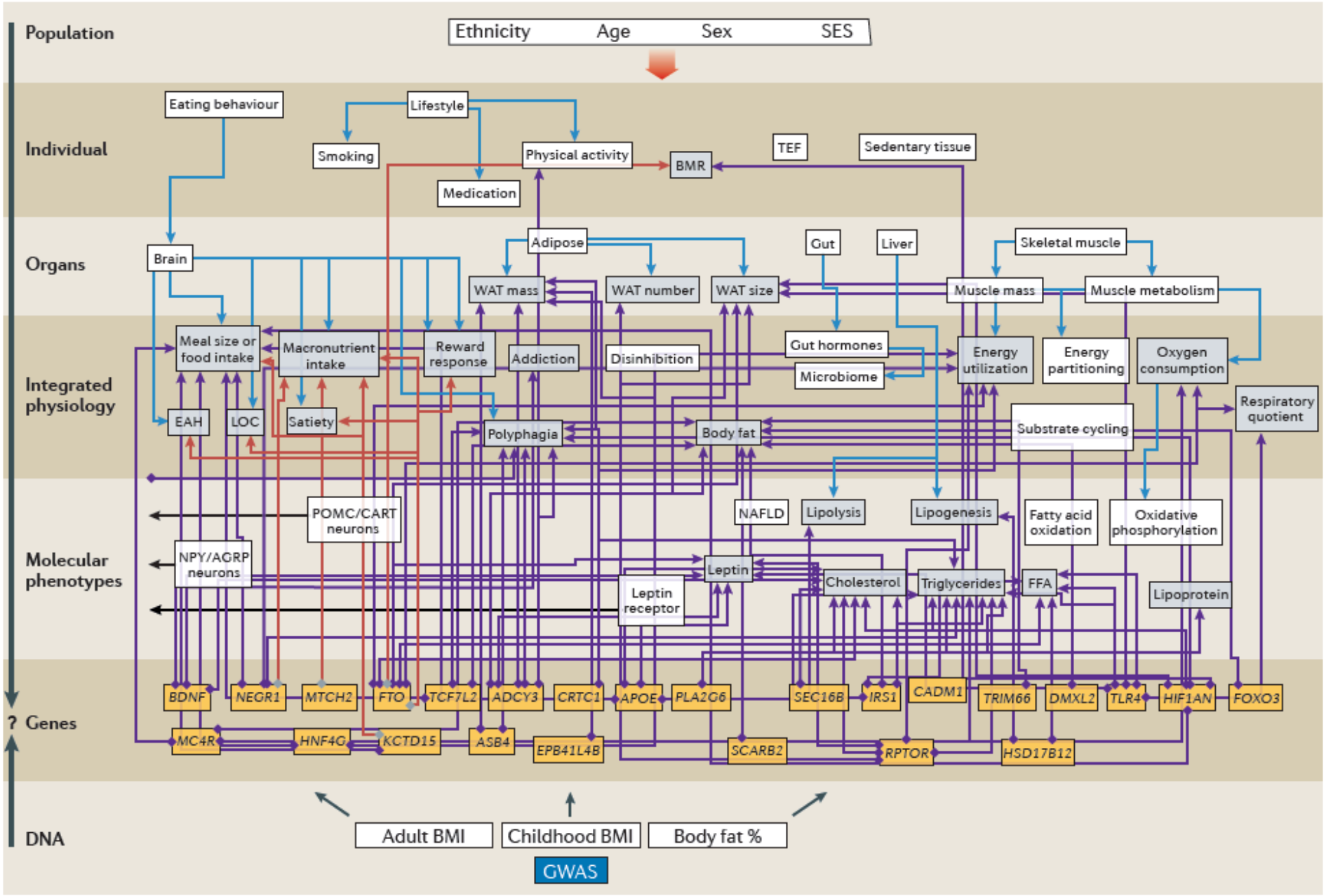
Review that unpacks the biology of obesity – taking several steps further than genetic association

DISEASE MECHANISMS

Nature Reviews Genetics (Dec 2017) 18:713 - 748

Convergence between biological, behavioural and genetic determinants of obesity

Sujoy Ghosh[1] and Claude Bouchard[2]



Nature Reviews Genetics (Dec 2017) 18:713 - 748

Conclusions

- Genetic associations for complex multi-factorial traits is **complex**
- NCDs are caused by **genetic risk variants and environmental effects**
- If **heritability** is high it should be possible to find genetic risk variants that explain a lot of the phenotype variability
- Genome-wide association studies are used to find genetic risk variants but often highly associated loci have small effects on the phenotype (they are therefore individually **not good predictors on the phenotype**)
- There are two main **study designs**:
  - Case : Control studies (e.g. diabetes vs healthy controls)
  - Quantitative traits (e.g. lipid levels – spectrum of low to high)
- **Outcome of a GWAS**
  - Genetic association (not disease causality)
  - Most often the associated variants are not the actual variants that contribute to the trait (they are proxies through LD)
  - Functional analysis and biological insights