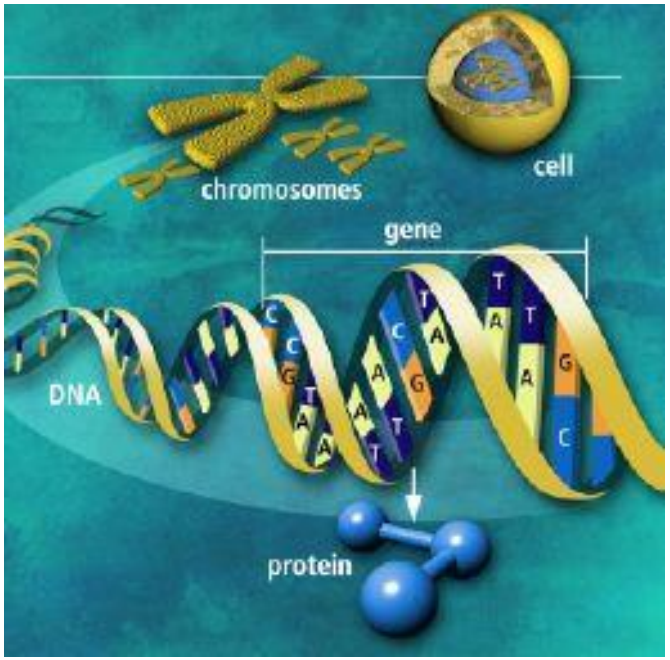
# Genome-wide association studies: in search of common and low frequency variants in complex traits

Ioanna Tachmazidou
Wellcome Trust Sanger Institute
it3@sanger.ac.uk

wellcome trust sanger institute

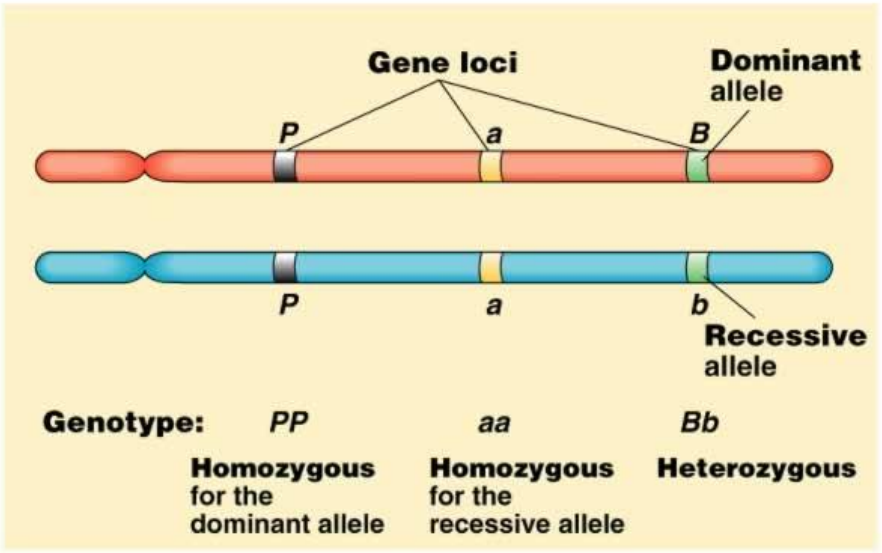
# The Book of Life

- A copy of all the DNA instructions used to make an organism.
- We have 2 copies of our genome packaged in 23 pairs of chromosomes, in the nucleus of each cell.
- DNA is made of combinations of four letters or nucleotide bases, which comprise the genetic “alphabet”.
- The order or sequence in which the A, C, T and G bases lie determines the meaning of the information encoded in DNA.
- Approximately, 3 billion letters of DNA make up the human genome.

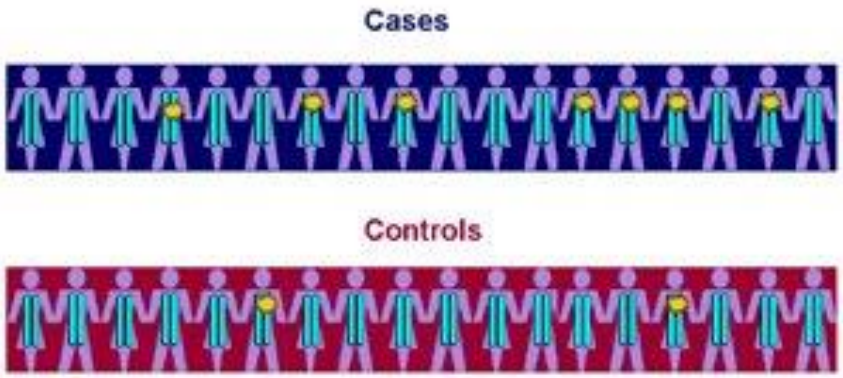



# Changes in our DNA

- There are always two copies of each gene, one from each parent.
- A gene locus can have different versions called **alleles**.
- The combination of alleles inherited from the parents is what gives rise to **genotypes**.
- Genotypes GG, Gg, gg at a locus in a population can be represented by 0,1,2 depending on the number of copies of allele g.
- The difference in a single nucleotide within and between populations is called **Single Nucleotide Polymorphism (SNP)**.
- The lowest allele frequency at a locus in a population is called **Minor Allele Frequency (MAF).**
- Some combinations of alleles in a population are seen more often than expected by chance. **Linkage Disequilibrium (LD)** is the non-random association of alleles at two or more loci.



# Association between a genetic variant and disease




**DD: variant homozygote**

**Dd: heterozygote**

**dd: common homozygote**

| Marker genotype | Affected | Unaffected | Total |
|---|---|---|---|
| DD | $n_{2A}$ | $n_{2U}$ | $n_{2.}$ |
| Dd | $n_{1A}$ | $n_{1U}$ | $n_{1.}$ |
| dd | $n_{0A}$ | $n_{0U}$ | $n_{0.}$ |
| Total | $n_{.A}$ | $n_{.U}$ | $n_{..}$ |

The odds ratio is the ratio of the odds of an event occurring in one group to the odds of it occurring in another group. Therefore the odds ratio for genotype DD relative to dd is:

$$\mathrm{OR}(DD : dd) = \frac{\text{odds of an individual with genotype DD carrying the disease}}{\text{odds of an individual with genotype dd carrying the disease}}$$

or else:

$$\mathrm{OR}(DD : dd) = \phi_{DD|dd} = \frac{n_{2A}/(n_{2A}+n_{2U})}{n_{2U}/(n_{2A}+n_{2U})} \Big/ \frac{n_{0A}/(n_{0A}+n_{0U})}{n_{0U}/(n_{0A}+n_{0U})} = \frac{n_{2A}/n_{2U}}{n_{0A}/n_{0U}}$$

Affected individual is $\phi_{DD|dd}$ times more likely to have marker genotype DD than dd.

Under the null hypothesis of no disease-marker association, the rows and columns of the contingency table are independent:

$$X^2 = \sum_{i=0,1,2} \sum_{j=A,U} \frac{(n_{ij} - \mathrm{E}(n_{ij}))^2}{\mathrm{E}(n_{ij})}, \text{ where } \mathrm{E}(n_{ij}) = \frac{n_{i.}n_{.j}}{n_{..}} \sim \chi^2_2$$

wellcome trust sanger institute

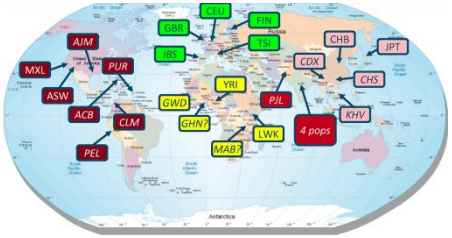
# What is a GWAS?

Genome-Wide Association Study – study scanning markers across genome (≈0.5M-2.5M) of many people (>2K) to find genetic variations associated with a particular disease or phenotype


## ➢ Tools

- Population-based studies (not family-based)
  - thousands of human subjects
- Detailed, annotated genome maps
  - Human genome project
- Encyclopedia of human genetic variation
  - HapMap, 1000 Genomes Project
- High-throughout genotyping platforms

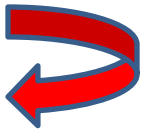

| Platforms |
| --- |
| Affymetrix 500k |
| Affymetrix 6.0 |
| Illumina 370k |
| Illumina 550k |
| Illumina 610k |
| Illumina 1M |
| Illumina 2.5M |

# GWAS Principles

**Case-control pairs or population cohorts**

**Type for 500K-2.5M SNPs**



**After imputation up to 40M SNPs**

**Data Quality Control**

**Obtain information about strength of association genome-wide**
**(within limits of sample size, allele frequency, LD etc.)**

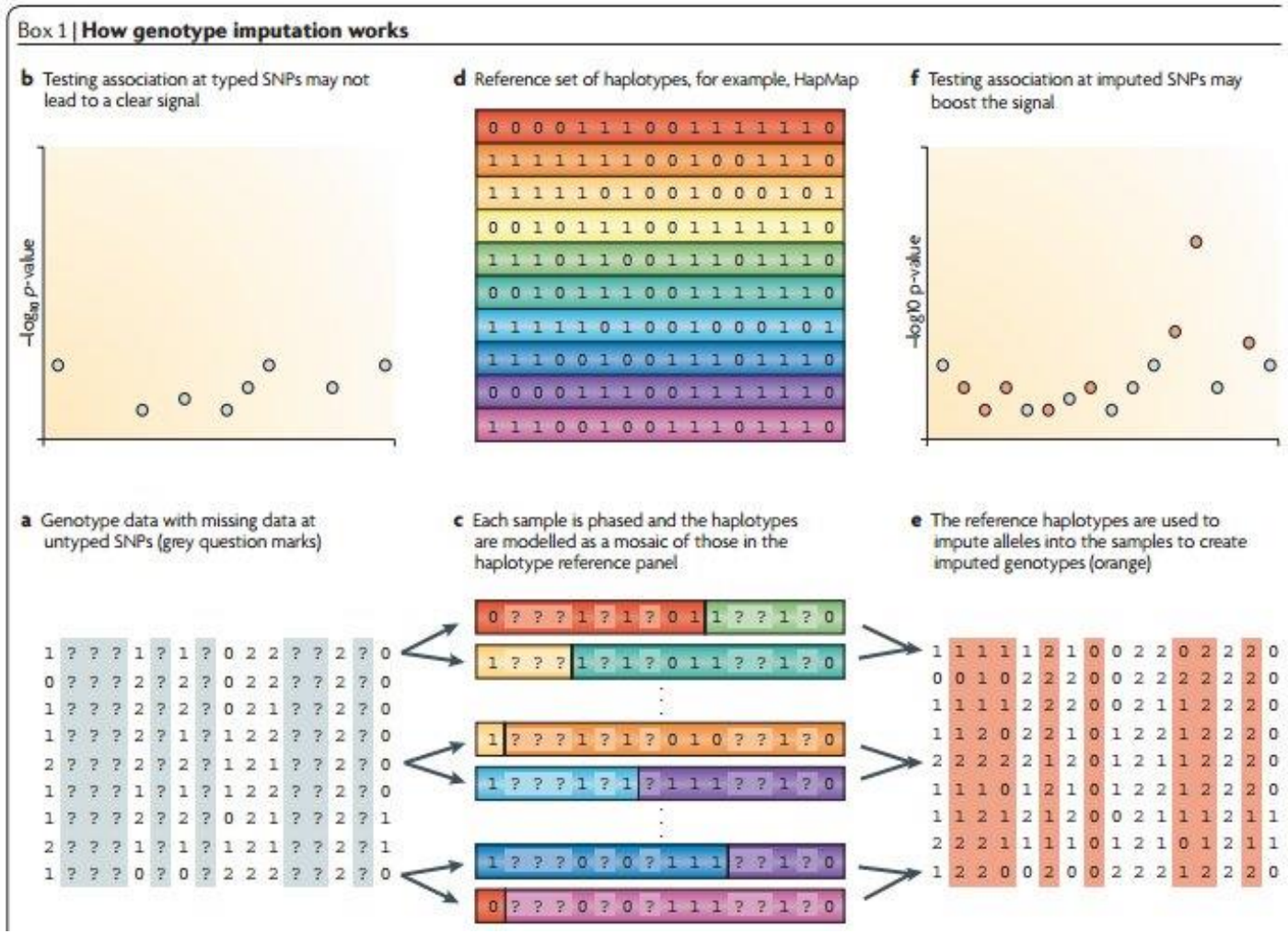
**Prioritise signals and seek replication**

**Establish association at the genome-wide significance level**
**(p-value < 5x10-8)**

# Genotype imputation

The process of predicting genotypes that are not directly genotyped in a dataset

- Allows you to directly test association at variants not genotyped or failed QC
- Facilitates the combination of results (meta-analysis) across cohorts that have used different chips

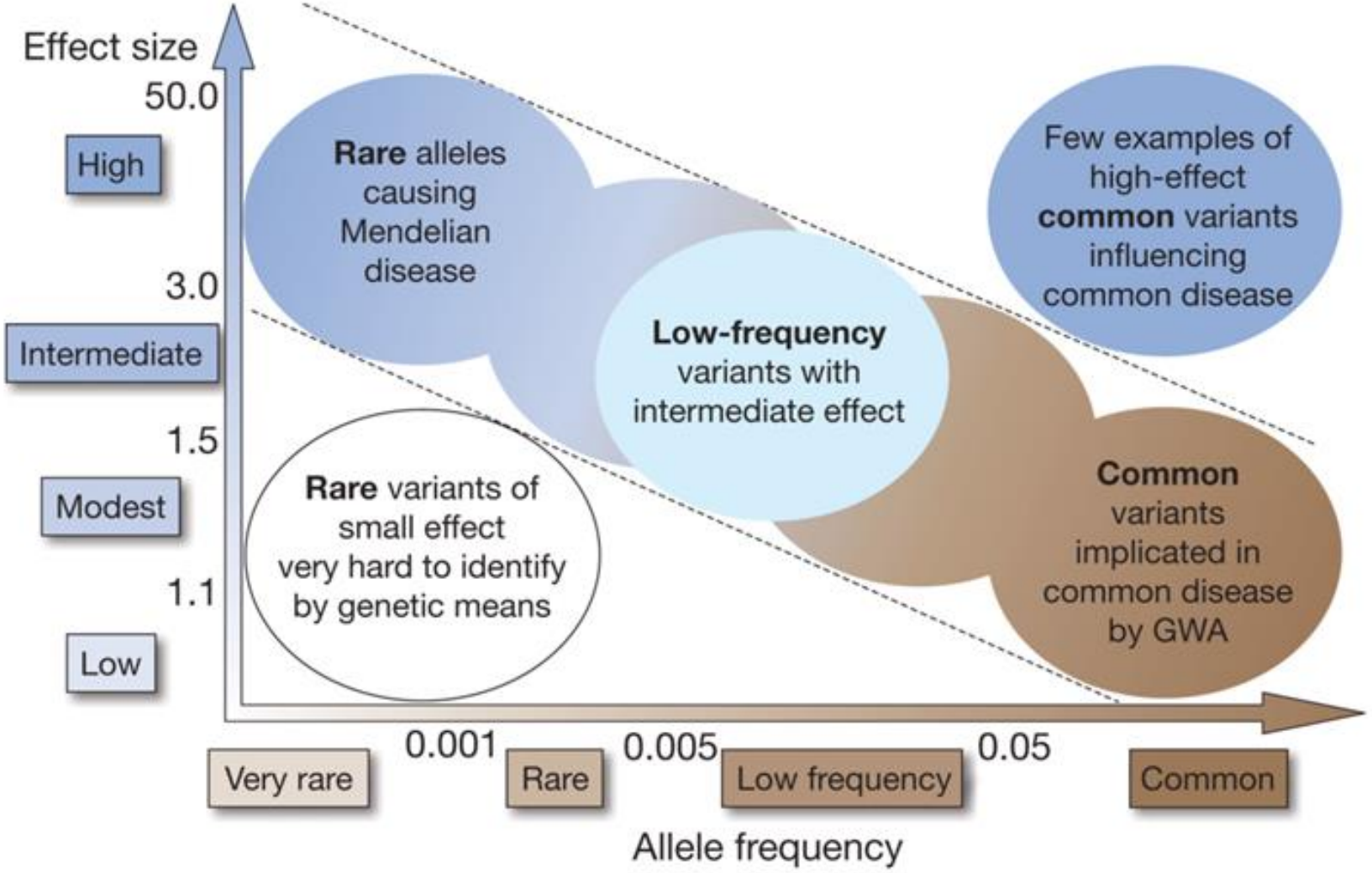



**Genotype uncertainty**

| Genotype | 0 | 1 | 2 |
|---|---|---|---|
| Probability | 0.01 | 0.18 | 0.81 |

Marchini and Howie, Nat. Genet. Rev., 2010

# Atlas of complex disease




Manolio et al., Nature, 2009

# Sample size matters

| MAF | N cases and controls (1:1) | | |
|---|---|---|---|
| | OR=2 | OR=1.4 | OR=1.2 |
| 0.40 | 680 | 2,000 | 10,000 |
| 0.05 | 2,500 | 13,000 | 46,000 |
| 0.01 | 11,000 | 50,000 | 220,000 |

80% power to detect an effect at $p=5x10^{-8}$

# Principles of meta-analysis

Synthesis of different datasets to obtain a summary based on evidence from the combined data

Increases power by increasing sample size

Facilitated by imputation, which enables the combination of data across different genotyping platforms





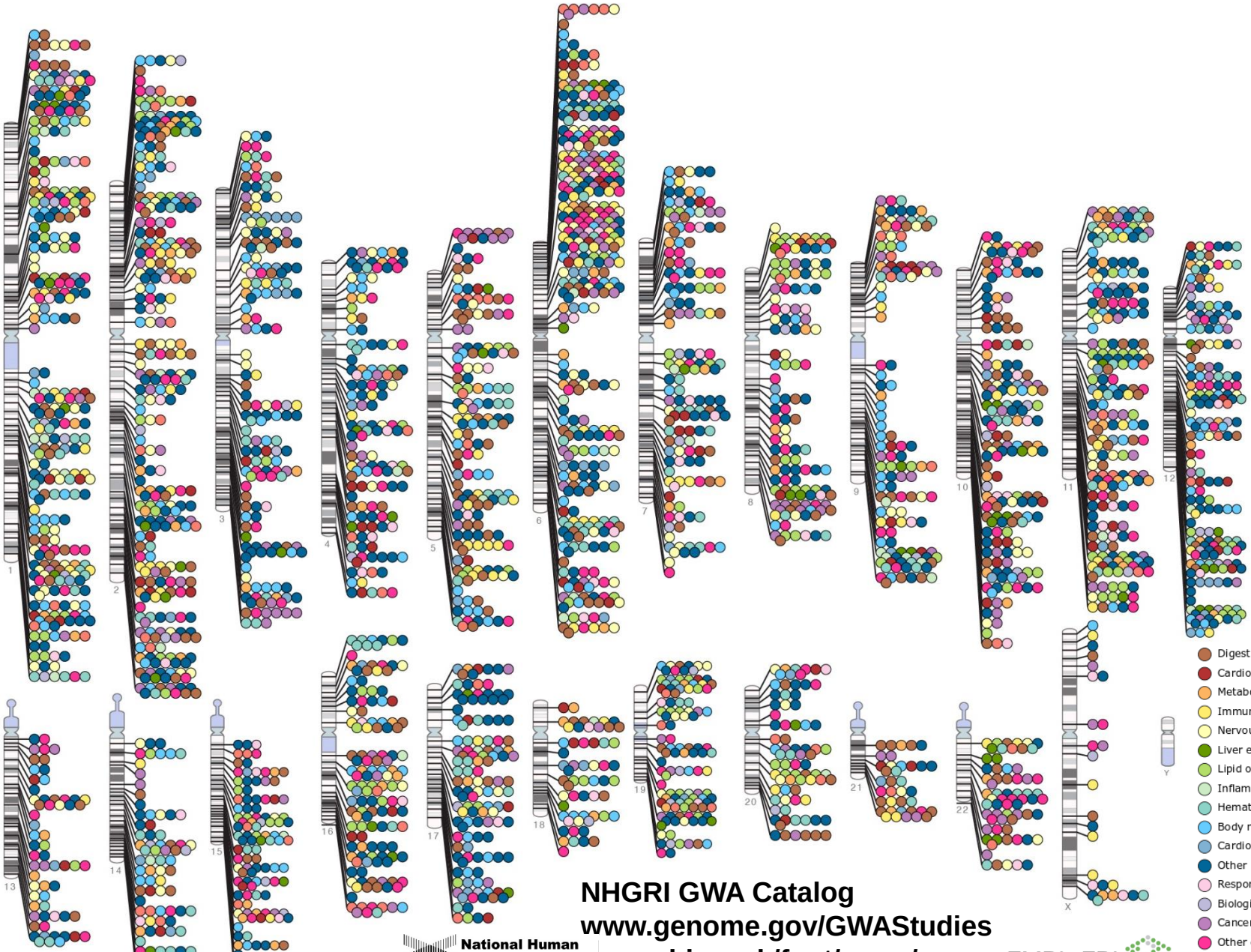
DIAbetes Genetics
Replication And Meta-analysis

# Sample size vs inverse variance based meta-analysis

| | Sample size based | Inverse variance based |
|---|---|---|
| Inputs | $N_i$ - sample size for study $i$<br>$P_i$ – $P$-value for study $i$<br>$\Delta_i$ - direction of effect for study $i$ | $\beta_i$- effect size estimate for study $i$<br>$se_i$ - standard error for study $i$ |
| Intermediate Statistics | $Z_i = \Phi^{-1}(P_i/2) * \text{sign}(\Delta_i)$<br>$w_i = \sqrt{N_i}$ | $w_i = 1/\text{SE}_i^2$<br>$se = \sqrt{1/\sum_i w_i}$<br>$\beta = \sum_i \beta_i w_i / \sum_i w_i$ |
| Overall Z-Score | $Z = \frac{\sum_i Z_i w_i}{\sqrt{\sum_i w_i^2}}$ | $Z = \beta/SE$ |
| Overall $P$-value | $P = 2\Phi(\lvert -Z \rvert)$ | |

- Fixed versus random effects meta-analysis
- Must have independent set of effect sizes
- Larger studies should carry more weight
- Weight each effect size by the inverse variance

# Published Genome-Wide Associations through 12/2013
## Published GWA at p≤5X10$^{-8}$ for 17 trait categories




**NHGRI GWA Catalog**
**www.genome.gov/GWAStudies**
**www.ebi.ac.uk/fgpt/gwas/**

National Human Genome Research Institute

EMBL-EBI

# Missing heritability in complex traits

- Interactions
- Structural variation
- Epigenetics and environment
- Thousands of very small effects
- Large phenotype-genotype heterogeneity
- Low frequency (0.01<MAF<0.05) and rare variants (MAF<0.01)

Evidence already exists that rare variants associate with disease
Role of rare variants in complex disease is poorly characterized
Chip-based GWAS do not access low frequencies well
Rare variants do not impute well

# Next generation Whole Genome Sequencing (WGS)



- Generates millions of short reads inexpensively, but with relatively high error rates
- Relies on redundant sequencing of each base to distinguish sequencing errors from true genetic variants
- To achieve high accuracy at rarer sites requires high average depth
- WGS a large number of samples with low depth more powerful than a small number of samples with high depth

# UK10K: 10,000 UK Genomes


10.4M GBP strategic award grant by the Wellcome Trust in 2010
164 researchers from 51 institutions
10 times deeper information than 1000 Genomes Project
Find almost all variants with MAF > 0.1%
4,000 cohort samples WGS at ~6x depth

2,000 ALSPAC (The Avon Longitudinal Study of Parents and Children, Bristol University)

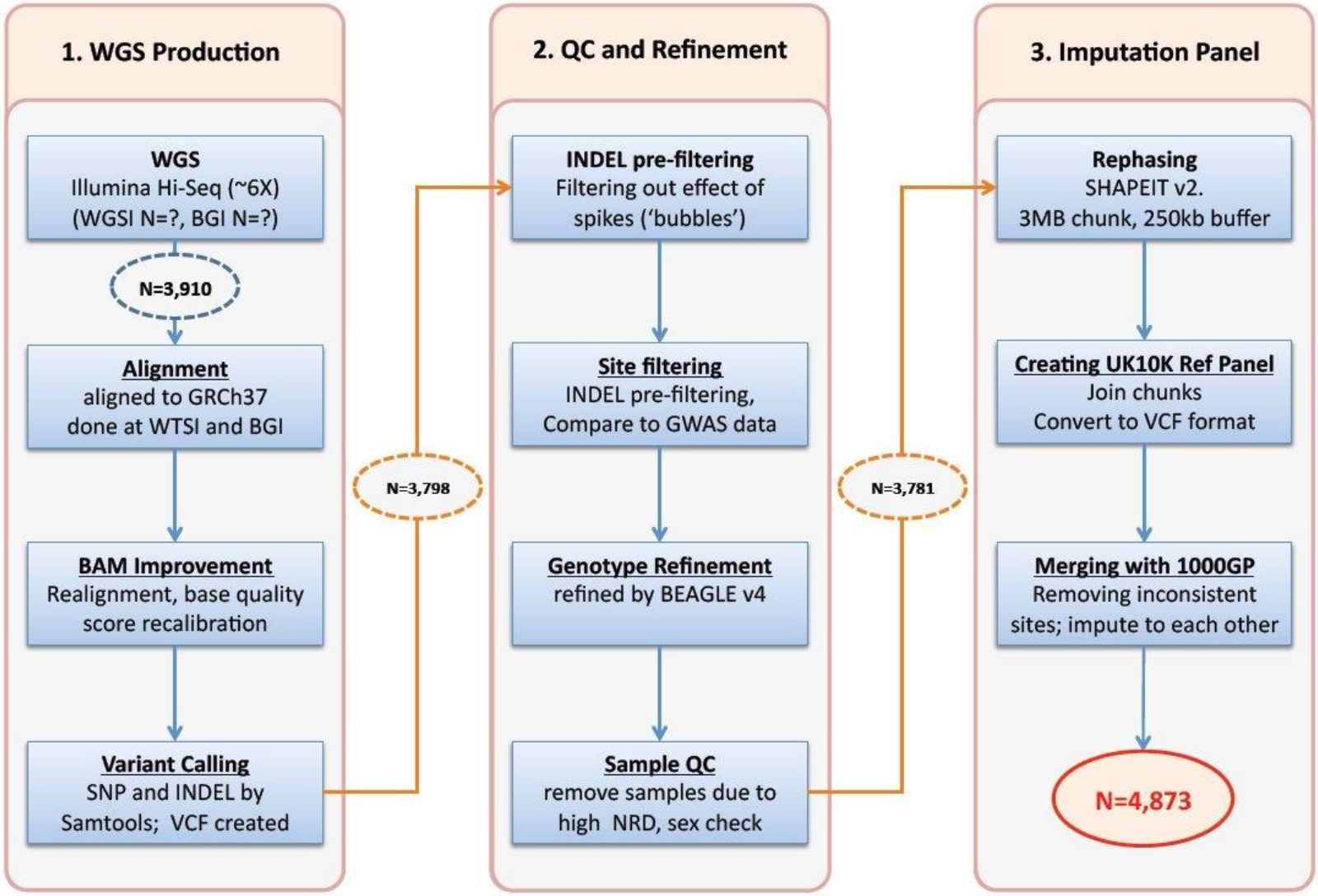
- Children/adolescents (18 yrs.)
- Males and females
- Geographically restricted (Avon health district, around Bristol)

2,000 TwinsUK (Identical and non-identical Twins, Department of Twin Research, Kings College London)

- Adults (median age 46 yrs.)
- All females
- UK-wide origin
- One twin per pair

- Deep genetic and phenotype coverage (clinical, questionnaire, molecular)
- 50 core phenotypes

# Production pipeline

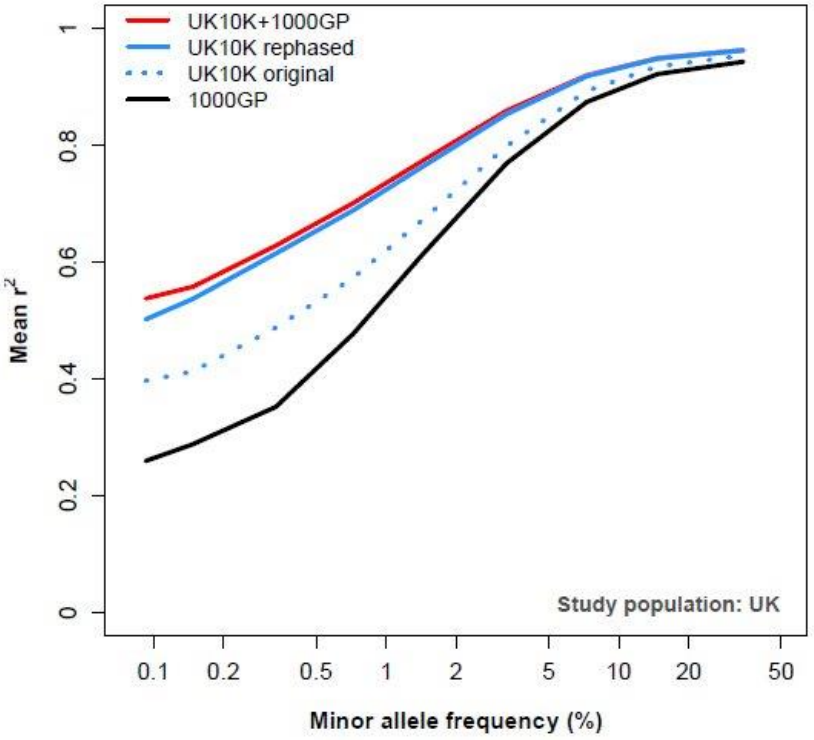


# The UK10K imputation panel

**Table 1 | Descriptives for the UK10K and 1000GP reference panels used for imputation.**

| | UK10K | 1000GP(Phase 1 v3) | Combined | Overlap |
|---|---|---|---|---|
| N samples (% European) | 3,781 (100%) | 1,092 (34.7%) | 4,873 | — |
| N total sites in final release | 45,492,035 | 39,527,072 | — | |
| N total sites after filtering* | 26,032,603 | 32,449,428 | 42,359,694 | 16,122,337 |
| **Autosome SNPs** | 23,411,635 | 29,797,220 | 38,238,102 | 14,970,753 |
| **Autosome INDELs** | 1,698,262 | 1,370,819 | 2,407,858 | 661,223 |
| **Chr X SNPs** | 858,380 | 1,223,328 | 1,612,230 | 469,478 |
| **Chr X INDELs** | 64,326 | 58,061 | 101,504 | 20,883 |

*For UK10K, the following sites were excluded: 18,180,633 singletons that do not exist in 1000GP, 1,064,168 multi-allelic sites and 214,631 mis-matched alleles sites. For 1000GP, the following sites were excluded: 7,053,246 singletons that do not exist in UK10K, 23,932 sites with a SNP and an INDEL at the same position and 443 within large structural deletions. The bold indicates that these four categories of variants are subsets of the N total sites after filtering.




- The UK10K panel was combined with the 1000GP panel to produce the UK10K+1000GP panel.
- Imputation accuracy improvement at rare and low-frequency variants.
- UK10K+1000GP yielded a larger number of high confidence imputed variants.

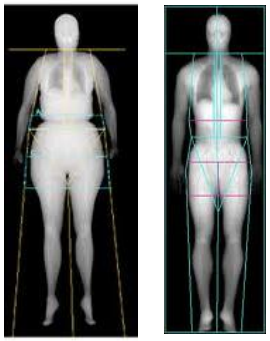
Huang, J. et al. *Nat Commun* 2015

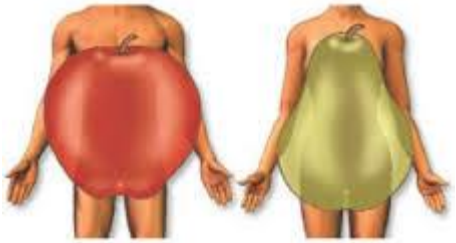
# Focus on body shape and composition

BMI
Weight
Height

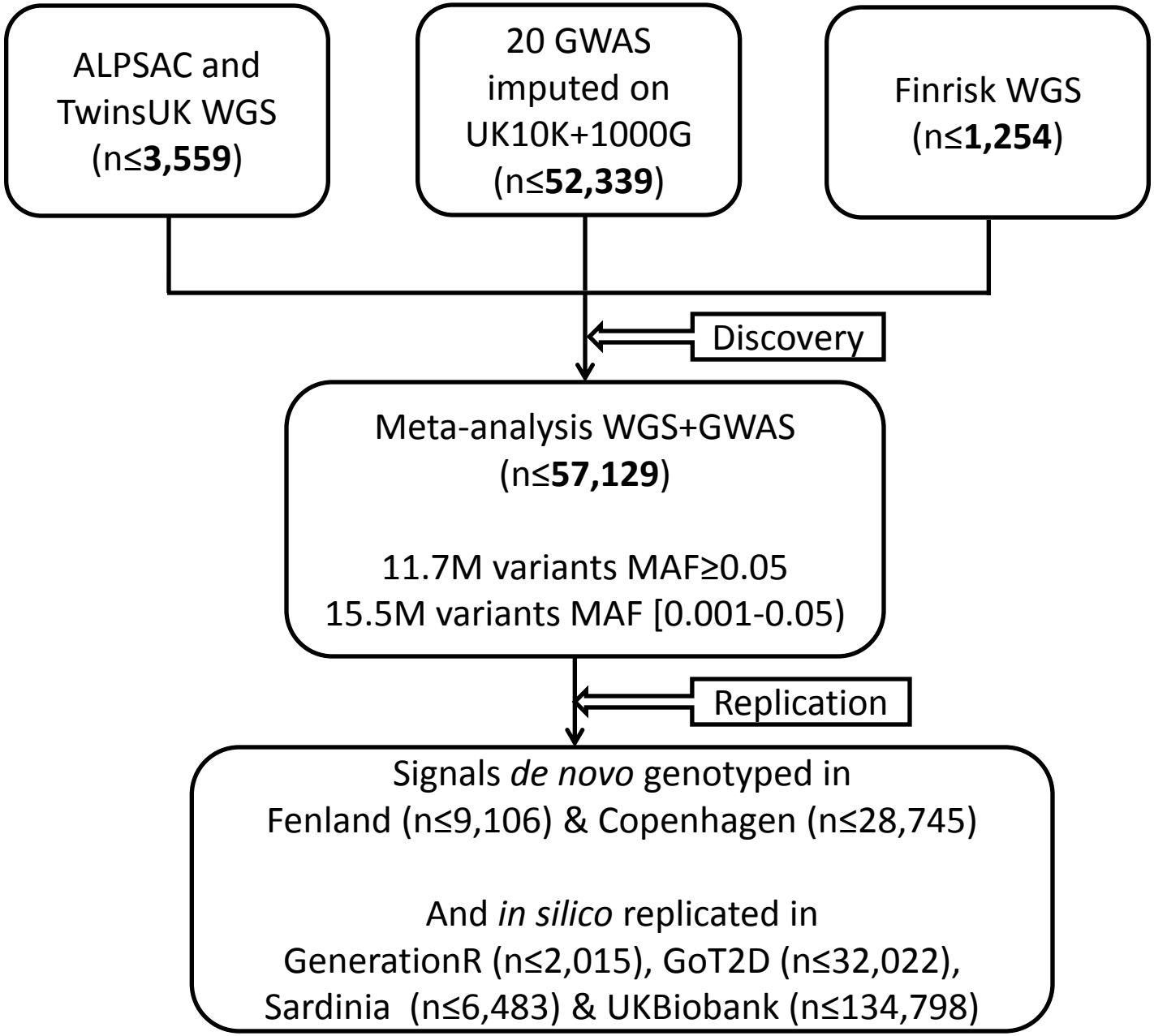
Total Fat Mass
Total Lean Mass
Trunk Fat Mass

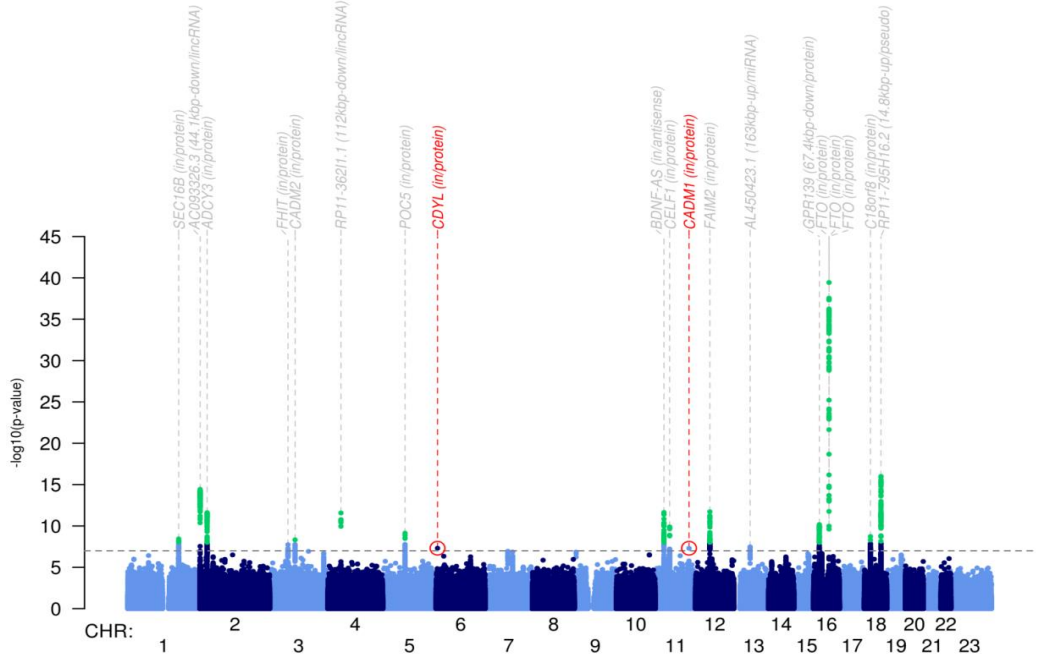
WHR BMI-adjusted
Waist BMI-adjusted
Hip BMI-adjusted
WHR
Waist
Hip

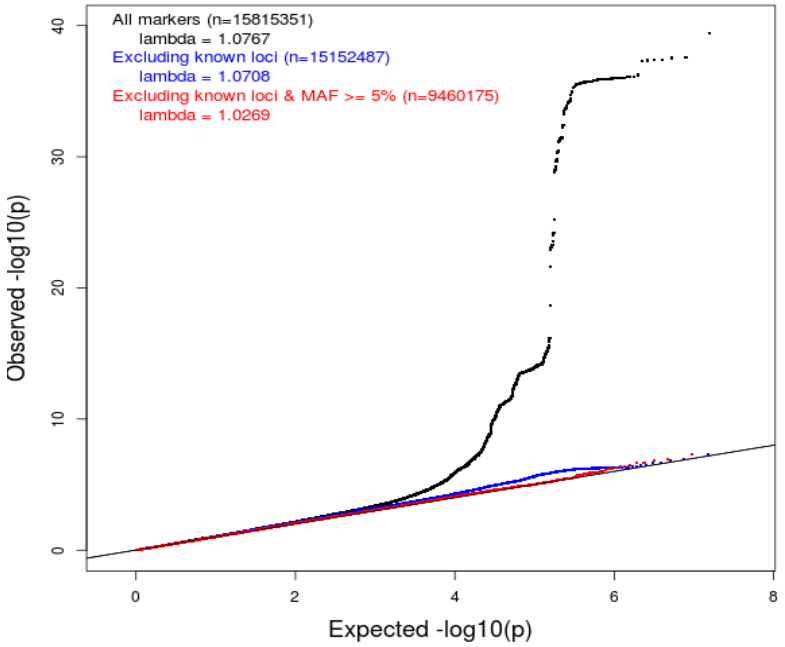



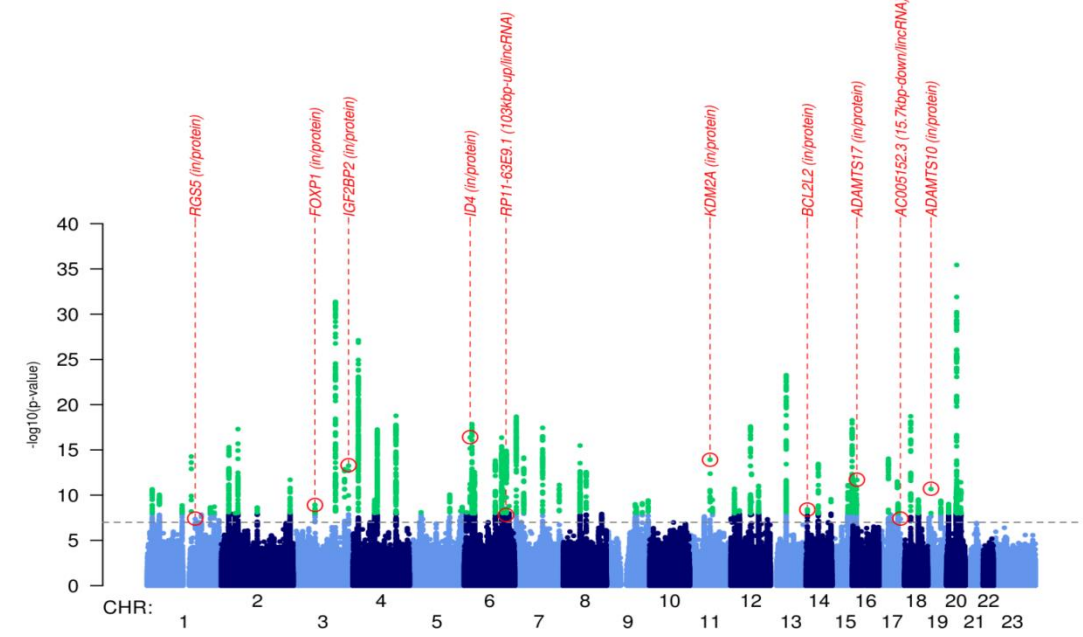


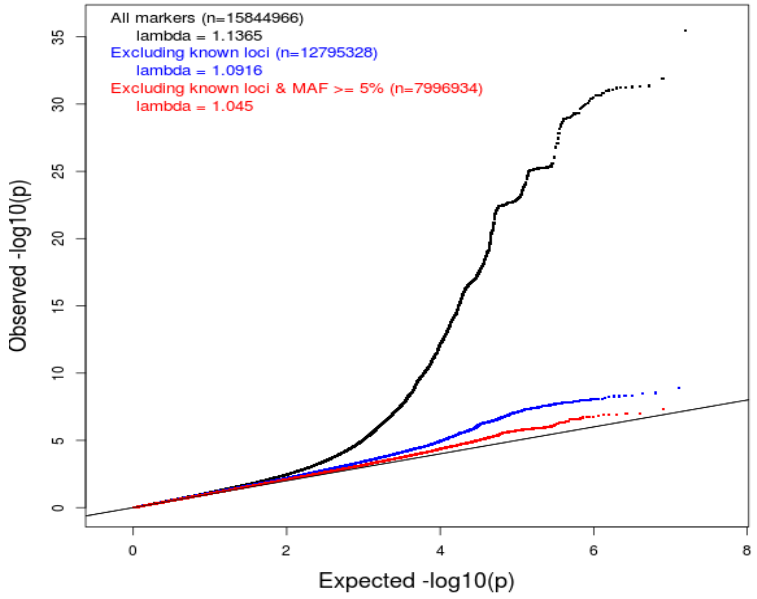
# Study design for single marker tests




SEC16B (in/protein)
AC093326.3 (44.1kbp-down/lincRNA)
ADCY3 (in/protein)
FHIT (in/protein)
CADM2 (in/protein)
RP11-362I1.1 (112kbp-down/lincRNA)
POC5 (in/protein)
CDYL (in/protein)
BDNF-AS (in/antisense)
CELF1 (in/protein)
CADM1 (in/protein)
FAIM2 (in/protein)
AL450423.1 (163kbp-up/miRNA)
GPR139 (67.4kbp-down/protein)
FTO (in/protein)
FTO (in/protein)
FTO (in/protein)
C18orf8 (in/protein)
RP11-795H16.2 (14.8kbp-up/pseudo)
-log10(p-value)
CHR:

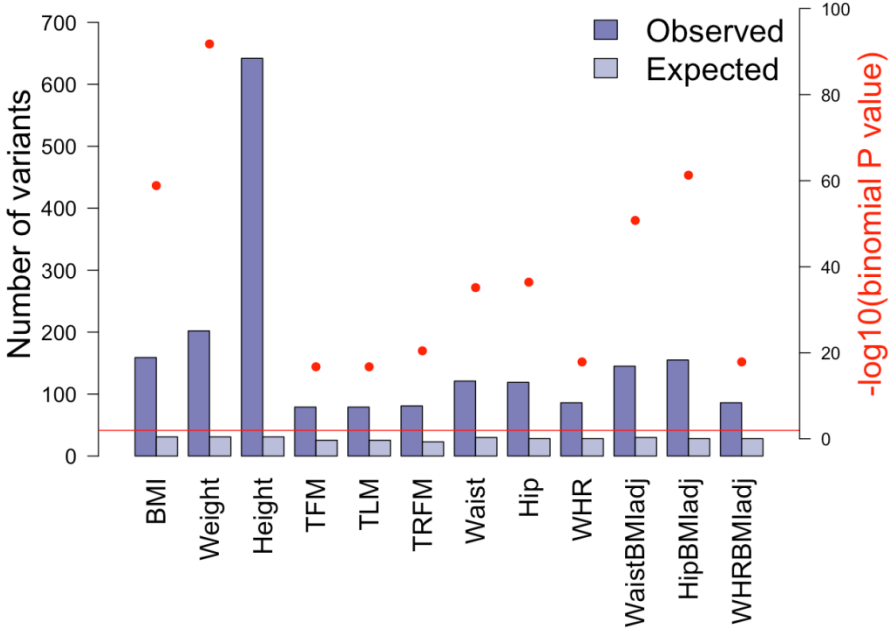

All markers (n=15815351)
lambda = 1.0767
Excluding known loci (n=15152487)
lambda = 1.0708
Excluding known loci & MAF >= 5% (n=9460175)
lambda = 1.0269
Observed -log10(p)
Expected -log10(p)


RGS5 (in/protein)
FOXP1 (in/protein)
IGF2BP2 (in/protein)
ID4 (in/protein)
RP11-63E9.1 (103kbp-up/lincRNA)
KDM2A (in/protein)
BCL2L2 (in/protein)
ADAMTS17 (in/protein)
AC005152.3 (15.7kbp-down/lincRNA)
ADAMTS10 (in/protein)
-log10(p-value)
CHR:

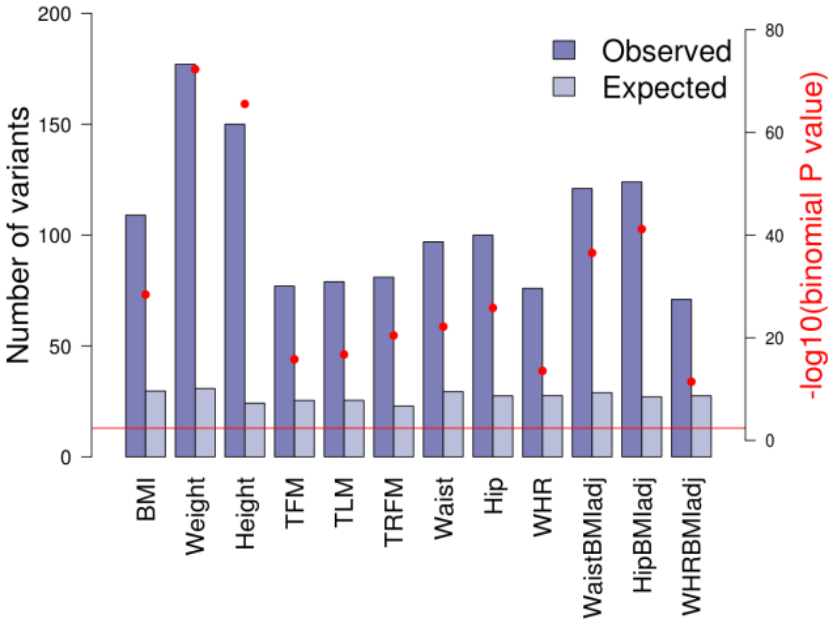

All markers (n=15844966)
lambda = 1.1365
Excluding known loci (n=12795328)
lambda = 1.0916
Excluding known loci & MAF >= 5% (n=7996934)
lambda = 1.045
Observed -log10(p)
Expected -log10(p)

# Enrichment in discovery meta-analysis

Using independent variants (r2<0.2) with MAF ≥ 0.1%

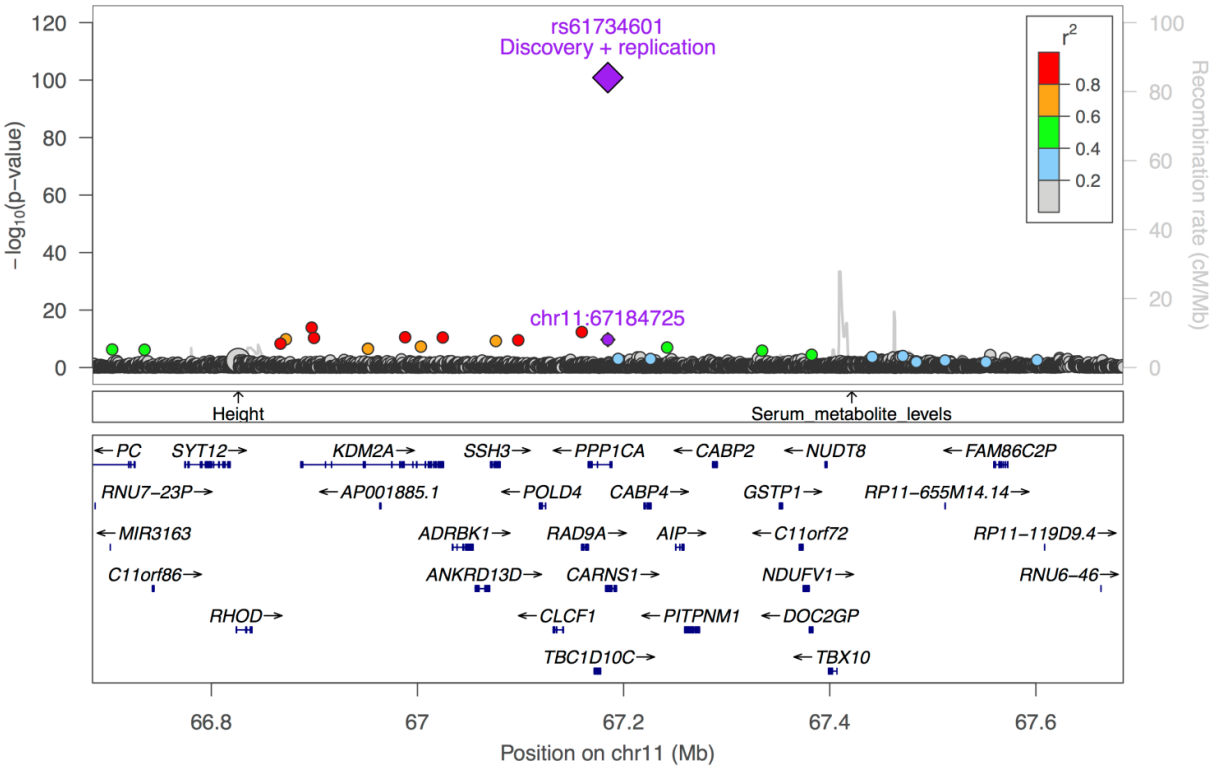



And after excluding previously known loci (±500 kb)



# Novel height locus on chr11

- rs61734601 (stage 1 and 2 EAF 8.2%, beta=-0.113, *P*=1.38x10$^{-101}$) falls in a gene dense region.
- It is located in the intron of *PPP1CA* and a non-coding exon of *CARNS1*, but is reported as significantly associated with expression of *RAD9A,* a DNA repair gene 20kb downstream, in several different tissues.
- DNA repair genes have previously been linked to growth disorders.
- rs61734601 is in high LD ($r^2$=0.82) with rs553917782, a 6-nucleotide insertion 10bp upstream of *RAD9A*.
- The 8 following nucleotides are conserved and occur near the centre of a DNase hypersensitivity peak that coincides with nucleosome depletion in multiple tissues, indicating likely transcription factor binding.



# Rare variant methodology

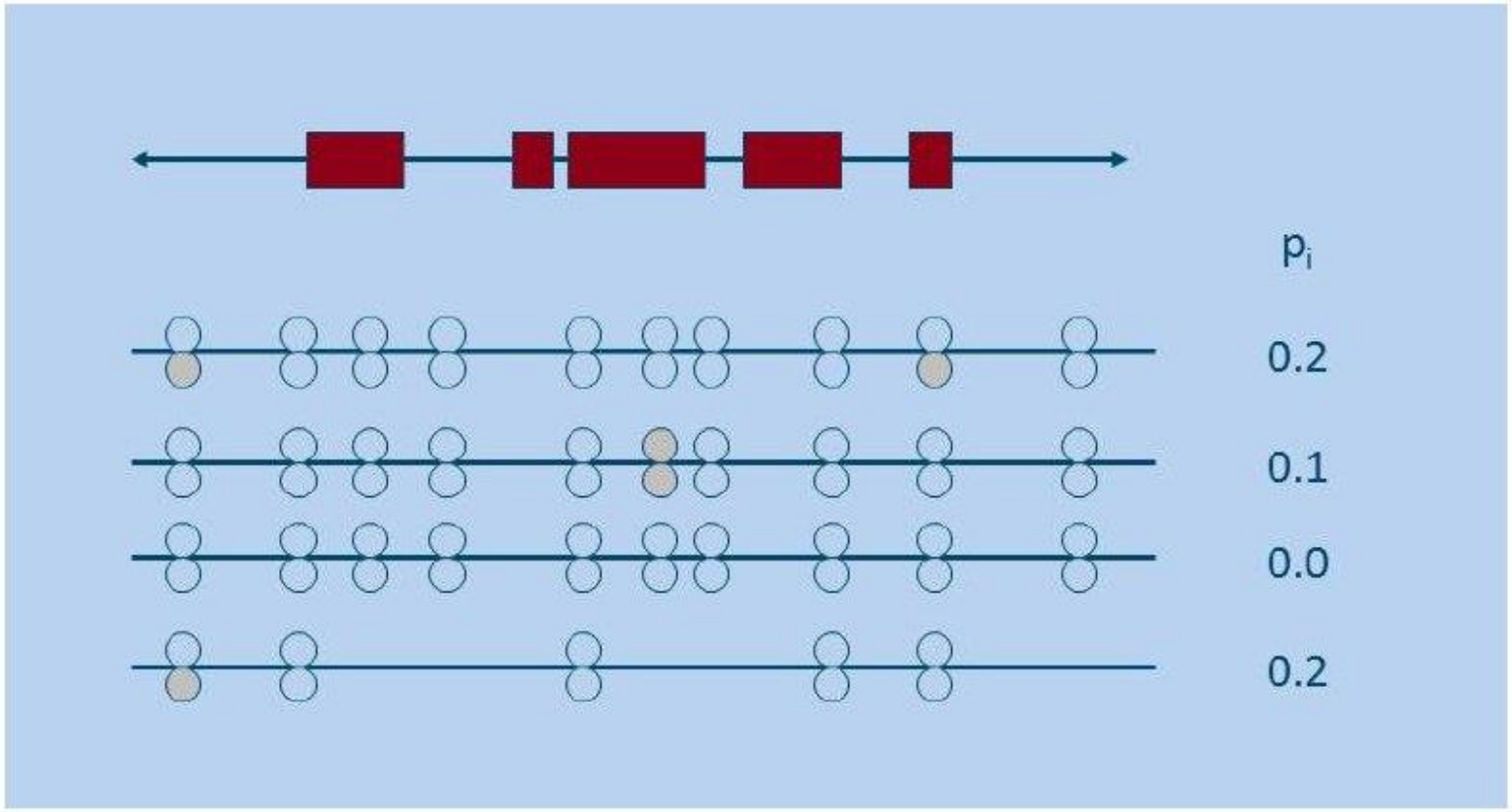
Single-point analysis of rare variants is under-powered.
An alternative is to use methods that combine information across multiple variant sites within a region.

| Reference | Method |
|---|---|
| Morgenthaler and Thilly, Mut Res 2007 | Cohort allelic sums test |
| Li and Leal, AJHG 2008 | Unweighted collapsing: presence/absence of rare va |
| Madsen and Browning, PLoS Gen 2009 | Weighted sum test |
| Morris and Zeggini, Gen Epi 2009 | Unweighted collapsing: proportion of rare variants |
| Mukhopadhyay et al, Gen Epi 2010 | Unweighted kernel-based association test |
| Han and Pan, Human Heredity 2010 | Data adaptive sum test |
| Hoffman et al, PLoS ONE 2010 | Step-up collapsing |
| Lawrence et al, BMC Bioinform 2010 | Unweighted collapsing: presence/absence of rare va |
| Price et al, AJHG 2010 | Variable threshold collapsing |
| Zawistowski et al, AJHG 2010 | Collapsing: cumulative minor allele counts |
| Bhatia et al, PLoS Comput Biol 2010 | Subset Selection |
| King et al, PLoS Gen 2010 | Linear mixed model for pooled association testing |
| Liu et al, PLoS Gen 2010 | Kernel-based adaptive cluster |
| Li et al, AJHG 2010 | Weighted Haplotype and Imputation-Based Tests |
| Garner, Gen Epi 2010 | Hidden Markov model |
| Zhou et al, Pac Symp Biocomput 2011 | Ridge regression |
| Neale et al, PLoS Gen 2011 | C-alpha |
| Wu et al, AJHG 2011 | Sequencing kernel association test |
| Ionita-Laza et al, PLoS Gen 2011 | Weighted collapsing test |
| Lin and Tang, AJHG 2011 | Weighted collapsing counts in a regression framewo |
| Asimit et al, Human Heredity 2012 | Weighted collapsing: proportion |
| Asimit et al, Human Heredity 2012 | Weighted kernel-based test |

wellcome trust sanger institute

# Collapsing approach

Morris & Zeggini, Genet Epidemiol. 2010




$$y_i = \alpha + \lambda \frac{r_i}{m_i} + \beta \boldsymbol{x}_i + \epsilon_i$$

# Sequence Kernel Association Test (SKAT)

Wu et al, *AJHG* 2011

- A multiple regression model allows for each variant to have its own direction and magnitude of effect or no effect

$$y_i = \alpha_0 + \boldsymbol{\alpha} \boldsymbol{X}_i + \boldsymbol{\beta} \boldsymbol{G}_i + \epsilon_i,$$

where $\boldsymbol{X}_i = (X_{i1}, ..., X_{im})$ and $\boldsymbol{G}_i = (G_{i1}, ..., G_{ip})$ denote covariates and genotypes respectively for subject $i$ across $p$ SNPs and $m$ covariates, and $\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)$ and $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)$ denote their regression coefficients

- Assume $\beta_j \sim$ Distribution$(0, \omega_j \tau)$
- $H_0 : \boldsymbol{\beta} = \mathbf{0} \Rightarrow H_0 : \tau = 0 \Rightarrow$ Variance-component score statistic $Q$
  - $Q$ only requires fitting the null model
  - $Q$ is based on the weighted linear kernel function $K(\cdot, \cdot)$, where $K(\boldsymbol{G}_i, \boldsymbol{G}_{i'}) = \sum_{j=1}^{p} \omega_j G_{ij} G_{i'j}$ measures the genetic similarity between subjects $i$ and $i'$
  - Fast as $Q = \sum_{j=1}^{p} \omega_j S_j^2$, where $S_j$ is the score statistic for testing the marginal effect of marker $j$

# Choice of weights

- To allow rare variants to have larger effects than common variants use $\sqrt{\omega_j} = \mathrm{Beta}(\mathrm{MAF}_j; \alpha_1, \alpha_2)$ with $0 < \alpha_1 \leq 1$ and $\alpha_2 \geq 1$
- Use $\alpha_1 = 1$ and $\alpha_2 = 25$ to increase the weight of rare variants and put decent nonzero weights for variants with MAF 1%-5%
- All variants are weighted equally ($\omega_j = 1$) if $\alpha_1 = \alpha_2 = 1$
- To put almost zero weight for MAF$>$ 1% use $\alpha_1 = \alpha_2 = 0.5 \Rightarrow \sqrt{\omega_j} = 1/\sqrt{\mathrm{MAF}_j(1 - \mathrm{MAF}_j)}$
- When all $\omega_j = 1$, case/control outcome and no covariates, SKAT is equivalent to C-alpha (Neale et al, PLoS Genet 2011)
- Weights estimated from PolyPhen scores or other bioinformatics tools possible

wellcome trust sanger institute

# Optimal unified approach (SKAT-O)

Lee et al, *AJHG* 2012

$$Q_k(\rho) = (1-\rho)Q_{k,\mathrm{SKAT}} + \rho\, Q_{k,\mathrm{Burden}}, \quad Q_{k,SKAT} = \sum_{j=1}^{m_k} w_{kj}^2 S_{kj}^2, \quad Q_{k,Burden} = \left(\sum_{j=1}^{m_k} w_{kj} S_{kj}\right)^2.$$

- The unified test reduces to SKAT when ρ = 0 and to the burden test when ρ = 1.
- Use an adaptive procedure SKAT-O to find an optimal ρ to maximize power.

$$Q_{optimal} = \min_{0\le\rho\le 1} p_\rho, \qquad 0 = \rho_1 < \rho_2 < \ldots < \rho_b = 1,$$

- For large samples and for given ρ, each test statistic can be decomposed into a mixture of two random variables, one asymptotically follows a chi-square distribution with one DF, and the other can be asymptotically approximated to a mixture of chi-square distributions.

# Rare Variant Meta-Analysis tests (metaSKAT-O)

Lee et al, *AJHG* 2013

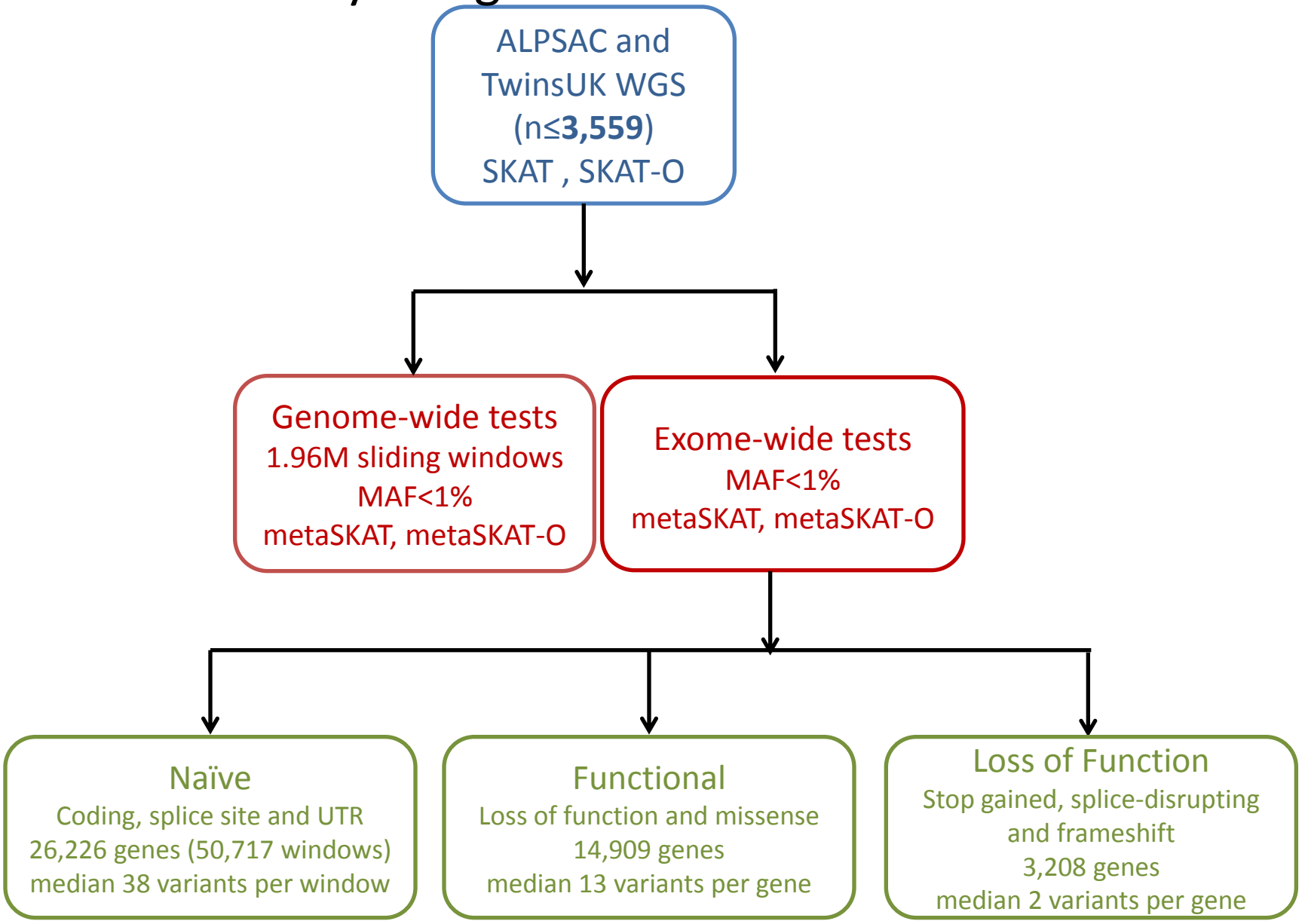
- Uses summary statistics.
- Same power as joint analysis.
- Corresponds to a fixed and random effects meta-analysis model.

$$Q_{\text{hom-meta}}(\rho) = (1-\rho)Q_{\text{hom-meta-SKAT}} + \rho Q_{\text{meta-Burden}},$$

where

$$Q_{\text{hom-meta-SKAT}} = \sum_{j=1}^{m}\left(\sum_{k=1}^{K}\omega_{kj}S_{kj}\right)^2, \quad Q_{\text{meta-Burden}} = \left(\sum_{j=1}^{m}\sum_{k=1}^{K}\omega_{kj}S_{kj}\right)^2$$

# Study design for rare variants tests



# Genome-wide significance thresholds

Xu et al. *Genetic Epidemiology* 2014

## Problem

- The AF spectrum of low frequency variants is very different from common ones
- Rare variation is usually jointly analysed in a series of genomic windows or regions

## Estimate the effective number of independent tests

- Based on correlations between all tests (Li et al. *Hum Genet* 2012)

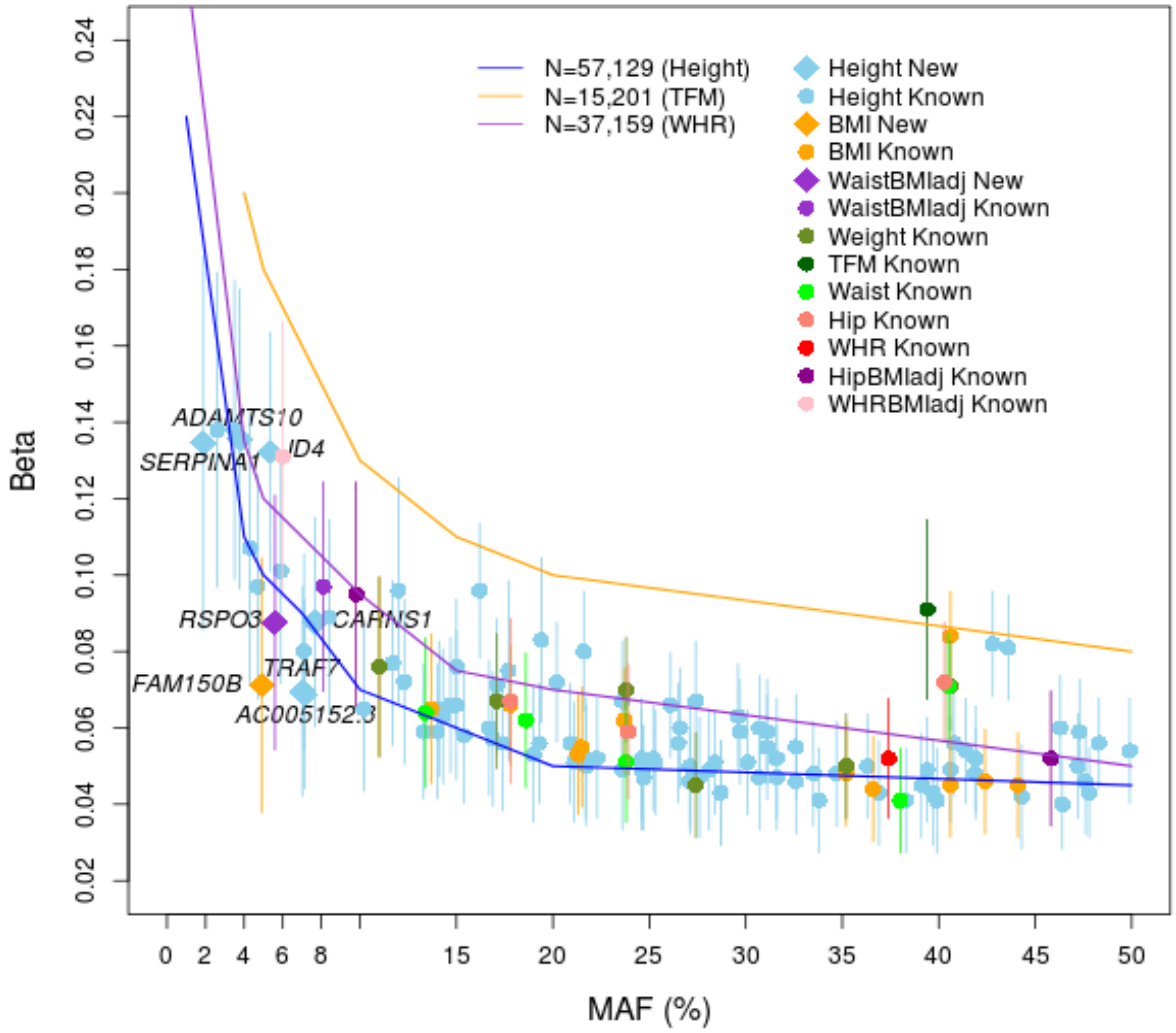
$$m_e = m - \sum_{i=1}^{m} (I(\lambda_i > 1)(\lambda_i - 1)).$$

- Using simulation to describe the behaviour of the minimal p-value across Regions under the null

## Significance thresholds

Between $2.5\times10^{-8}$ and $8\times10^{-8}$ for window-based testing.
Between $0.6\times10^{-8}$ and $1.5\times10^{-8}$ for a combined strategy of single-SNP tests and rare variant testing using a sliding-window test strategy.

# Power to detect association in discovery



# UK10K anthropometry effort

- Largest-scale association testing of low frequency variants with anthropometric traits to date.

- Newly identified associations at variants at the lower end of the frequency spectrum, not captured by the HapMap reference panel.
  - Demonstrates the power of imputation based on WGS haplotype sets.

- Discovery of 2 novel coding variants robustly associated with height in genes implicated in syndromic disorders (*SERPIN1A* and *ADAMTS10*), demonstrate genetic overlap between monogenic and polygenic anthropometric traits.

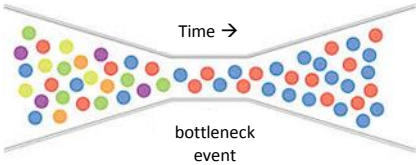
- Even though well-powered to detect them, we find no evidence of low frequency variants with strong effect sizes for anthropometric traits.

- Increasing sample size and sequencing depth, and building large reference panels to facilitate accurate imputation of SNVs is likely to identify further potentially functional low frequency and rare variants underpinning the genetic architecture of medically-relevant human complex traits.

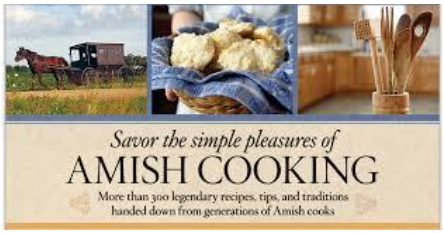
# Population isolates

The study of rare variants can be empowered by focusing on isolated populations.

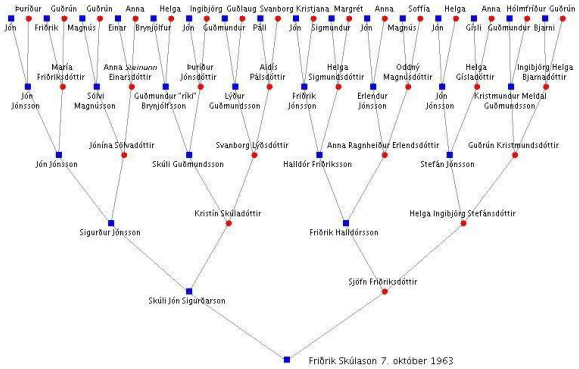
- Some rare variation is lost due to bottleneck effects , but others may have increased in frequency




- Linkage disequilibrium tends to be extended
- Homogeneous environment




- Deep information on genealogy



# HELIC: Hellenic isolated cohorts

- HELIC-MANOLIS (Minoan Isolates)
- Mylopotamos villages, Crete, Greece
- Geographically isolated
- N~4,500 of which 1,600 collected

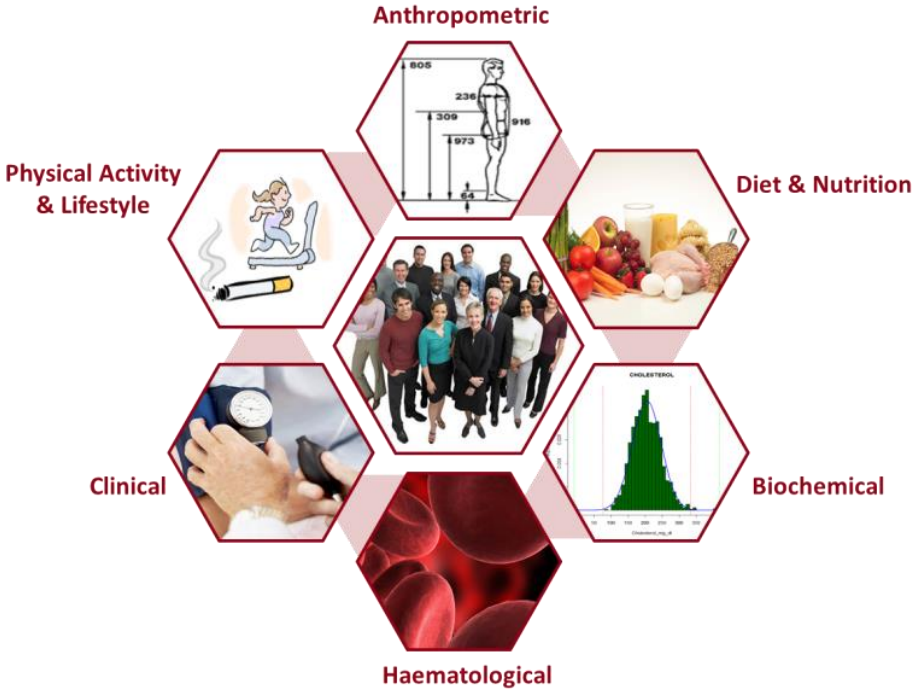
- Deeply phenotyped
- High fat content diet
- High rates of longevity
- Low rates of metabolic disease complications
- Ability to recontact individuals


www.helic.org

# HELIC: Hellenic isolated cohorts

- HELIC-Pomak
- Pomak villages, Xanthi, Greece
- Geographically isolated
- Religiously isolated
- N~11,000 of which 2,000 collected

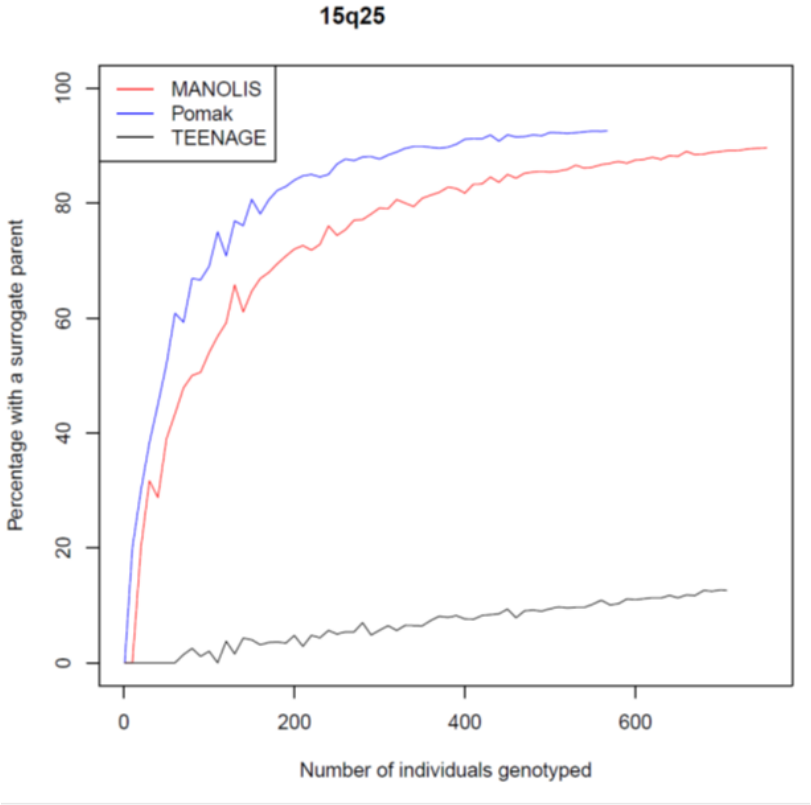
- Deeply phenotyped
- High levels of metabolic disease
- Ability to recontact individuals




www.helic.org

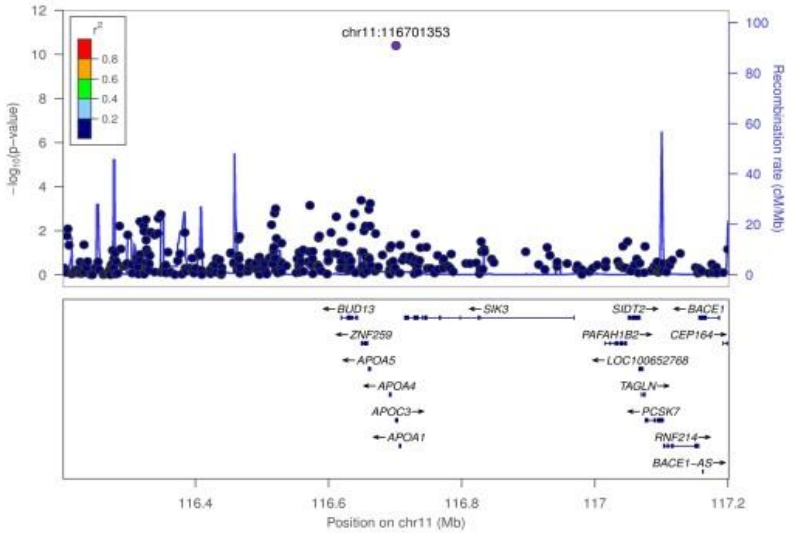
# HELIC overview

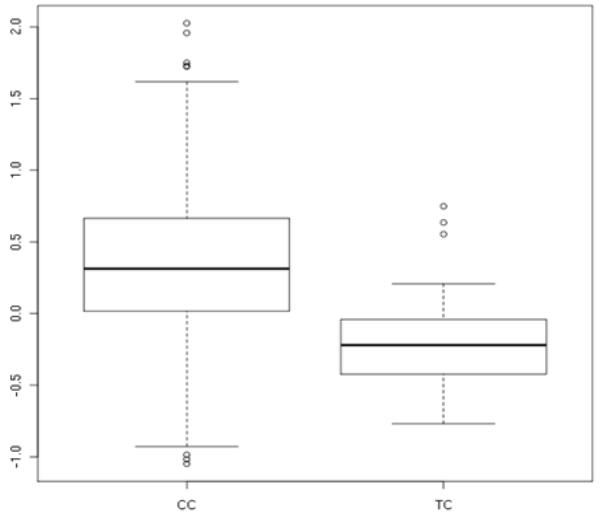
- ~3,500 samples with genome-wide association scan and exome chip data
- ~250 MANOLIS samples with whole genome sequencing at 4x
- ~2,500 samples with whole genome sequencing at 1x




Panoutsopoulou *et al.*, *Nature communications* 2014


# R19X *APOC3* cardio-protective variant association with lipid levels - exome chip data



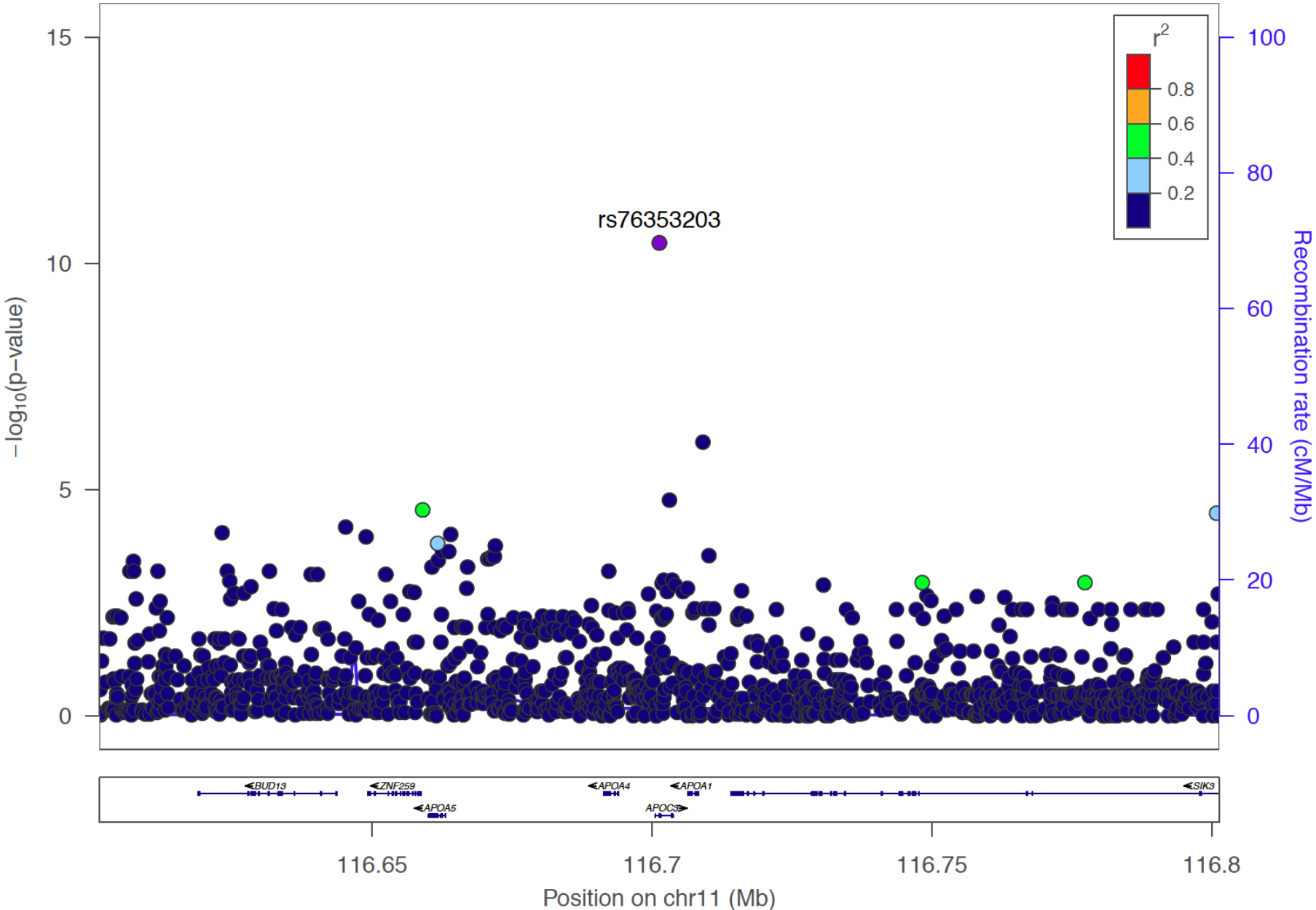




- Mylopotamos villages (n=1256, MAF 2%, $p=10^{-11}$)
- Meta-analysis across Mylopotamos villages and the Amish: $p=10^{-31}$, total n=2700
- Detection of this effect would have required 67,000 Europeans (MAF 0.05%)
- Exemplifies the value of population isolates and generalizability of findings

Tachmazidou et al, Nature Communications, 2013; Pollin et al, Science, 2008

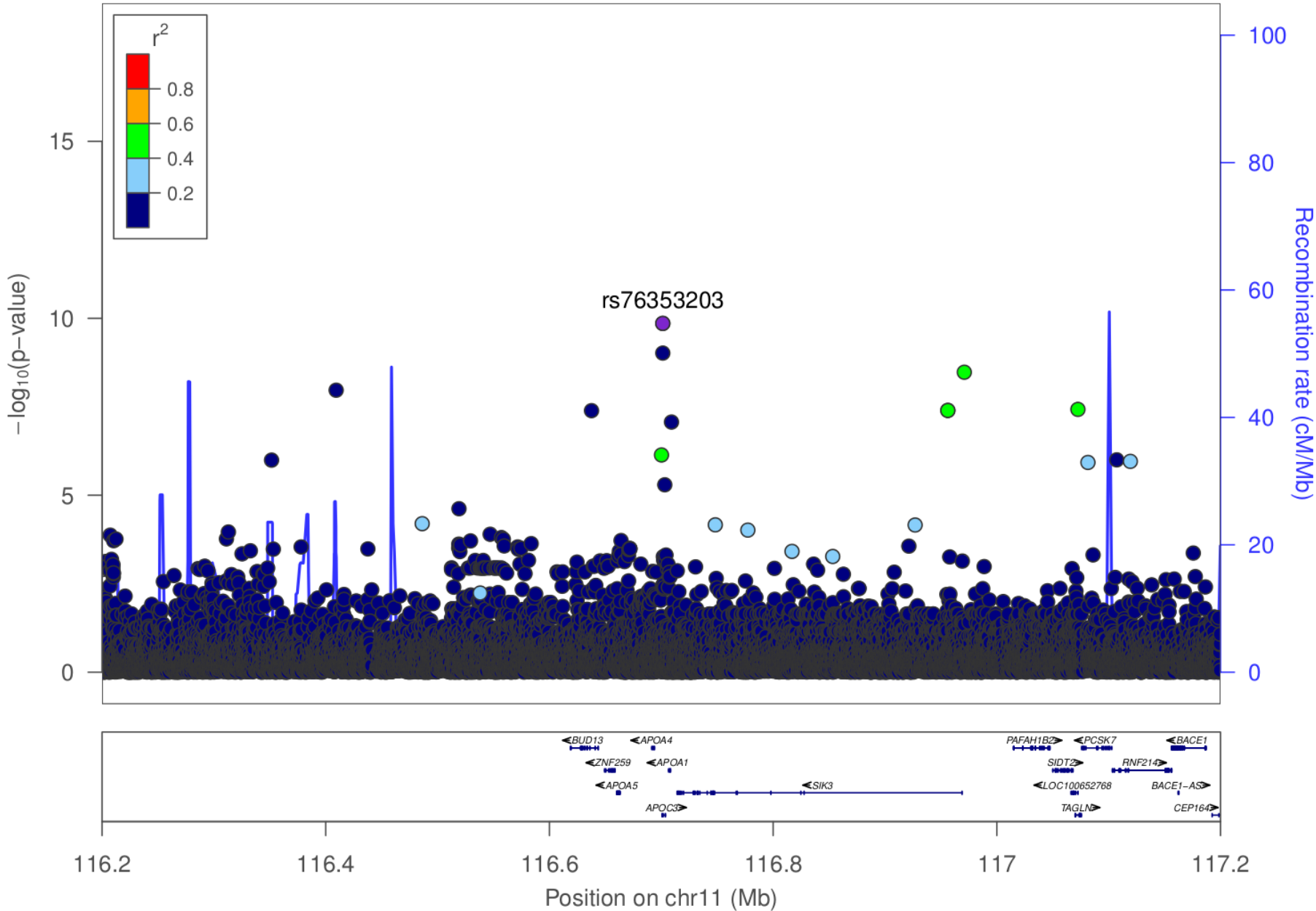
# R19X variant in *APOC3* - imputed genome-wide data

Gilly *et al* Human Molecular Genetics 2016



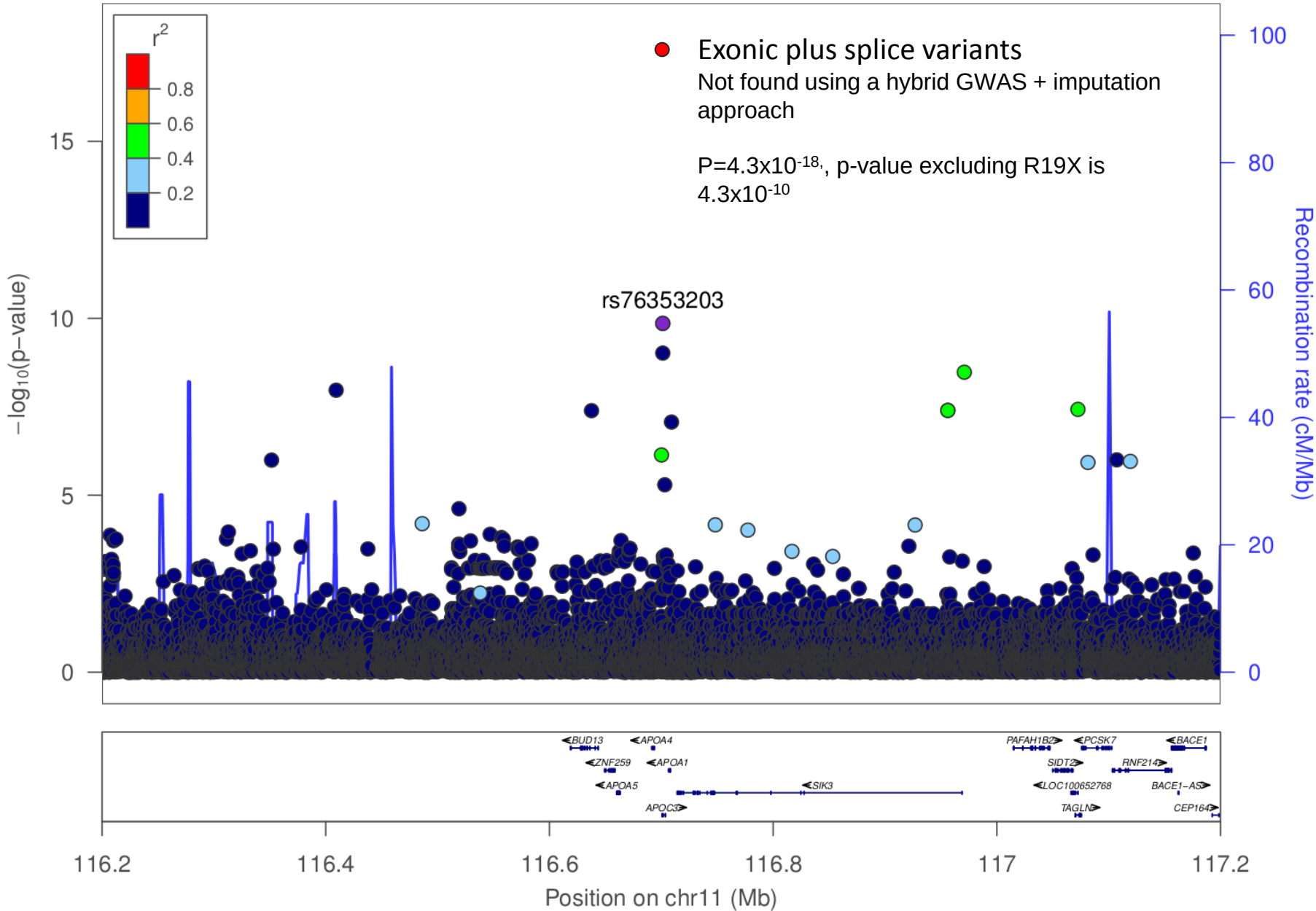
# R19X variant in *APOC3* - 1x WGS data

Gilly *et al* Human Molecular Genetics 2016



# R19X variant in *APOC3* - 1x WGS data

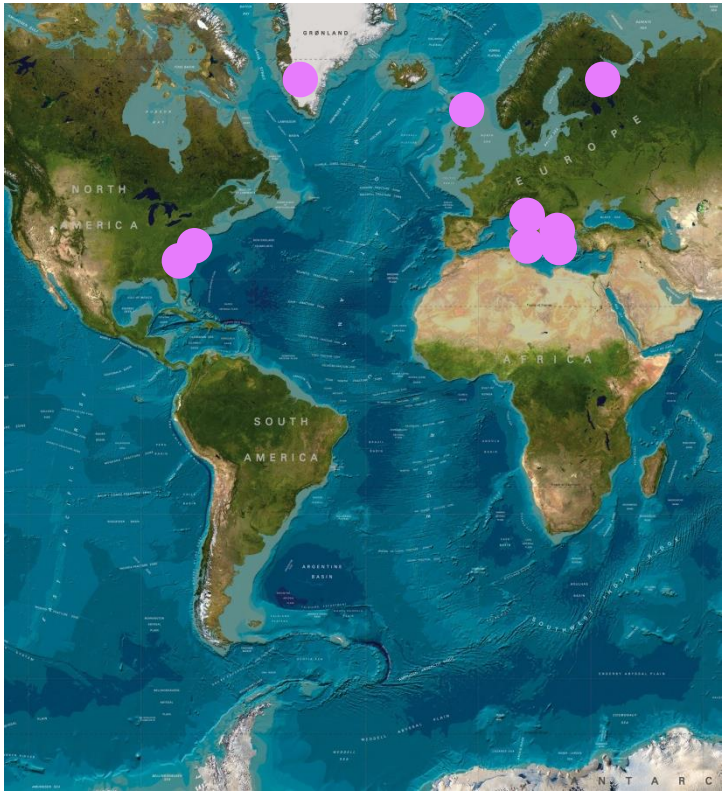
Gilly *et al* Human Molecular Genetics 2016



# Network of isolated population cohorts

Over 15 well-phenotyped cohorts, including founder populations in:

- Greece (Pomak and Mylopotamos villages),
- Finland (general Finnish population cohorts and Northern Finland sub-isolates),
- Italy (Carlantino, Val Borbera and Friuli Venezia Giulia villages, Sardinia),
- UK (Orkney islands),
- USA (Amish, Ashkenazi Jewish),
- Greenland



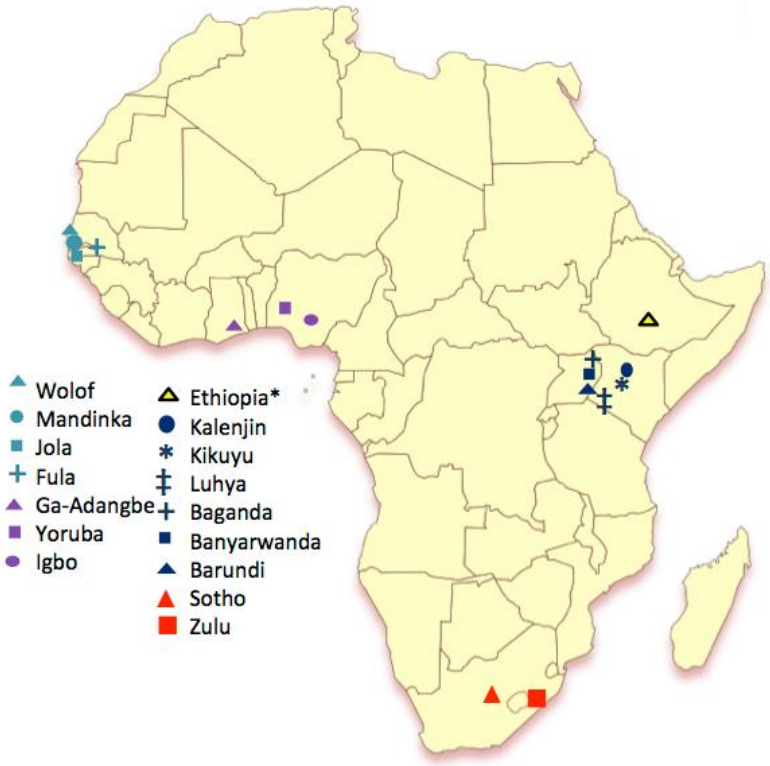

Emerging international WGS isolates consortium currently amassing in excess of 30,000 samples

# The African Genome Variation Project

Gurdasani D *et al. Nature 2015*

- A framework to help build genomic expertise and resources in Africa, and to drive forward genomic research
- 1,481 individuals across 18 ethnolinguistic groups with 2.5M genotype data
- 320 individuals (Ethiopia, South Africa, Uganda) with 4x WGS

**West:**
**The Gambia:** Jola, Fula, Mandinka, Wolof
**West-Central:**
**Nigeria:** Yoruba, Igbo
**Ghana:** Ga-Adangbe
**South:**
**South Africa:** Zulu, Sotho
**East:**
**Kenya:** Luhya, Kalenjin, Kikuyu
**Uganda:** Baganda, Barundi, Banyarwanda
**Ethiopia:** Amhara, Oromo, Somali



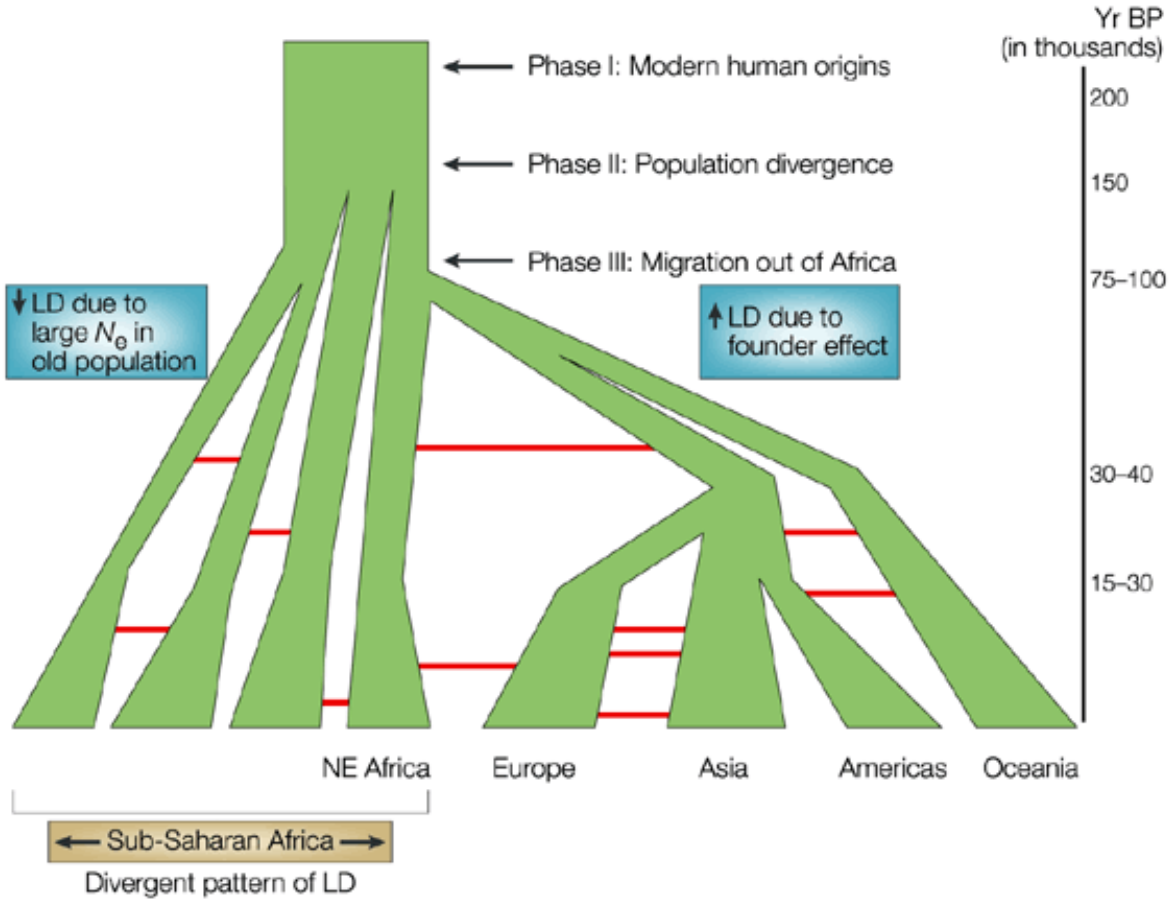

CRGGH

APCDR

MalariaGEN
GENOMIC EPIDEMIOLOGY NETWORK

# Challenge
# Pronounced genetic diversity across ethnic groups

# Challenge
# Low levels of correlation between genetic variants

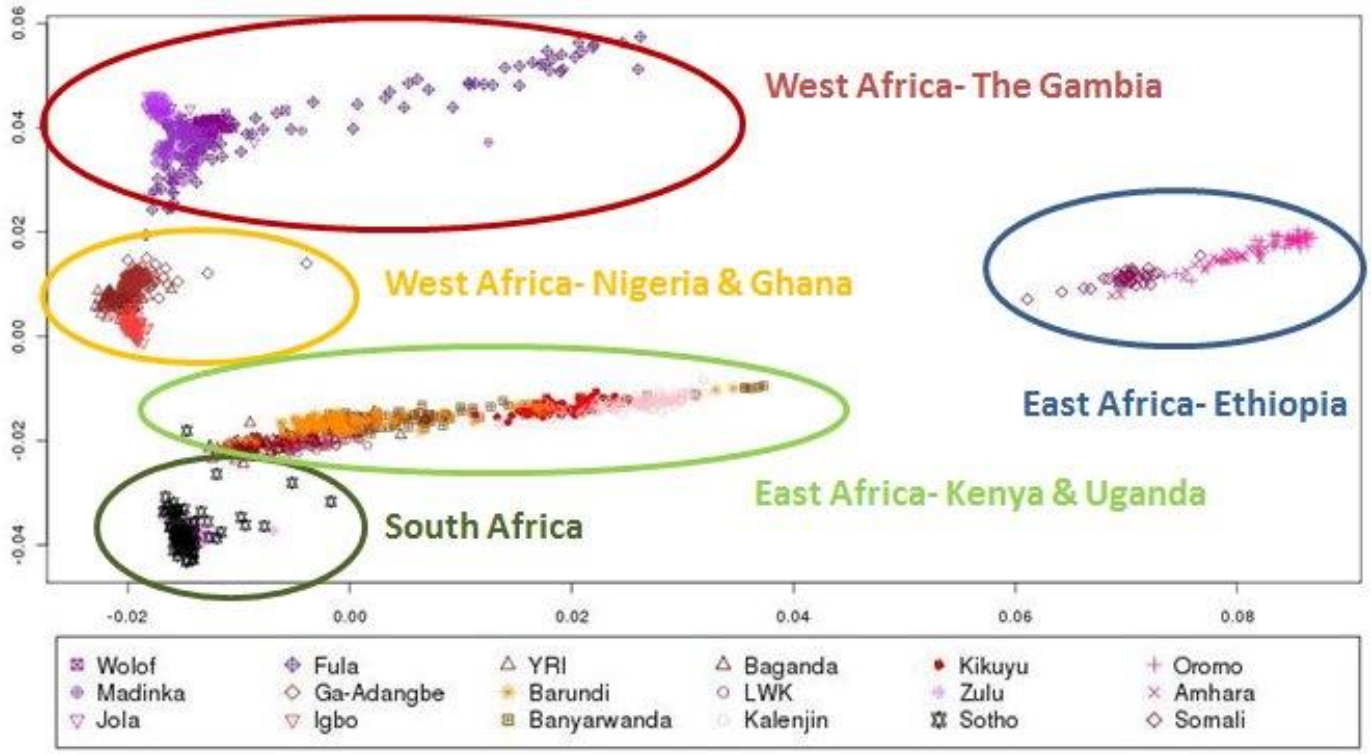



Tishkoff & Williams

Nature Reviews | Genetics

Ancestral African populations have maintained a large and subdivided population structure.
Disadvantage: need denser arrays
Advantage: fine mapping

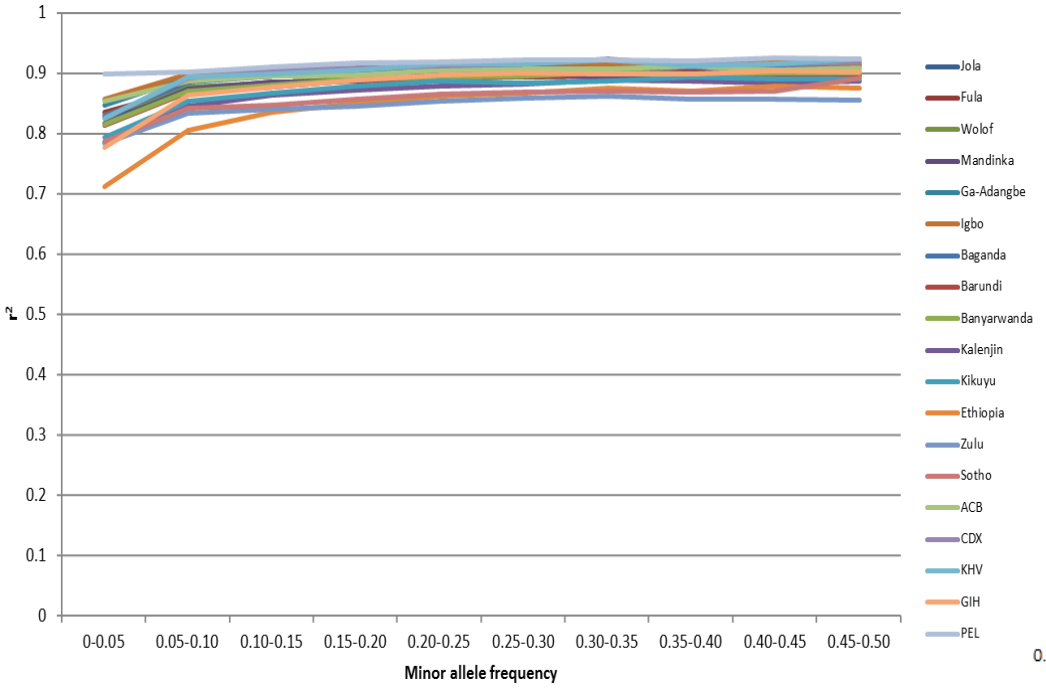
# How genetically diverse are African populations?

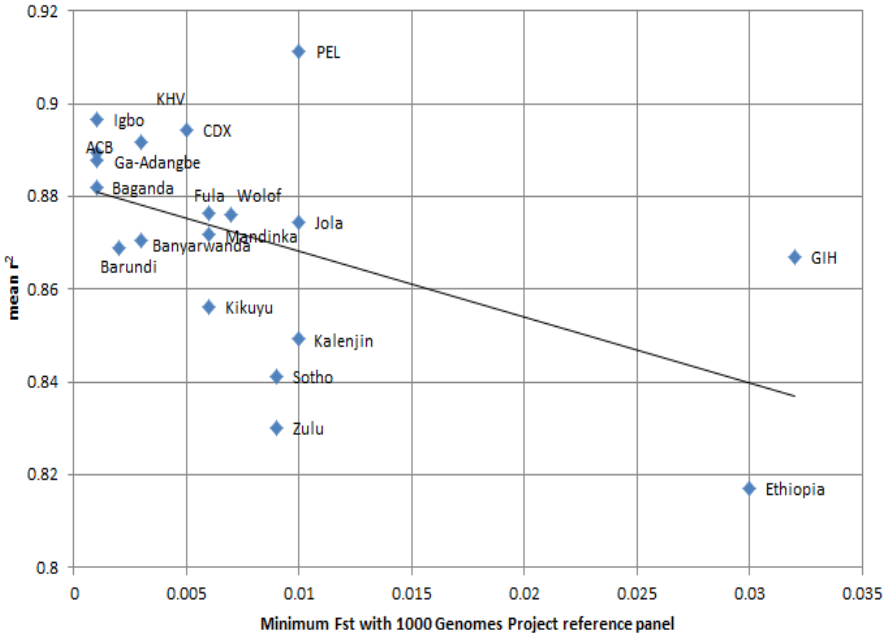



PC2 (7.6%)

PC1 (21%)

# Utility of existing reference panels for imputation in SSA



# Implications for genetic association studies

Not so good news:

- Large numbers of monomorphic and rare/low frequency variants on the genotyping array
  - For GWAS, 1.1-1.36M sites instead of 2.5M
- High levels of redundancy
  - 16-35% of variants have perfect proxies
- Low proportion of common variation captured
  - ~60-70% of common variation captured with $r^2$=0.8

Good news:

- Array serves as a good scaffold for imputing common variants in African populations using existing imputation panels

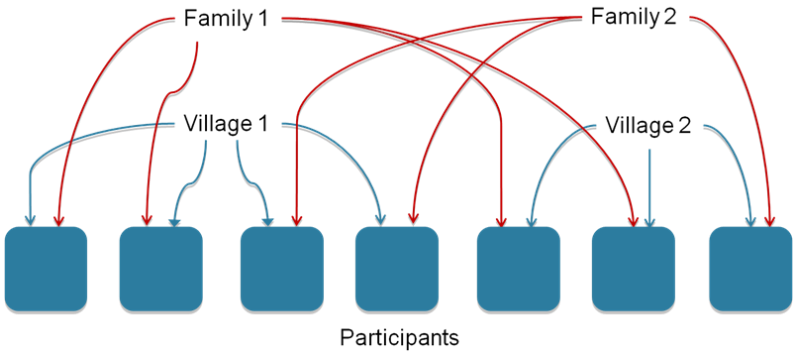
Admixture, population substructure, pronounced allelic diversity, low levels of LD, greater haplotype diversity have implications for the design of large-scale genomic studies within and among SSA populations

# African Genome Variation Project

- Provides basic framework for genetic studies in Africa
- Underpins the design of next-generation experiments
- Helps identify analytical challenges and develop statistical genetics methods to address them
- Generates a valuable resource for the scientific community
- Promotes collaboration and synergies among contributing parties
- Committed to building partnerships and research programmes that enable researchers in developing countries to share in the benefits of genomic research

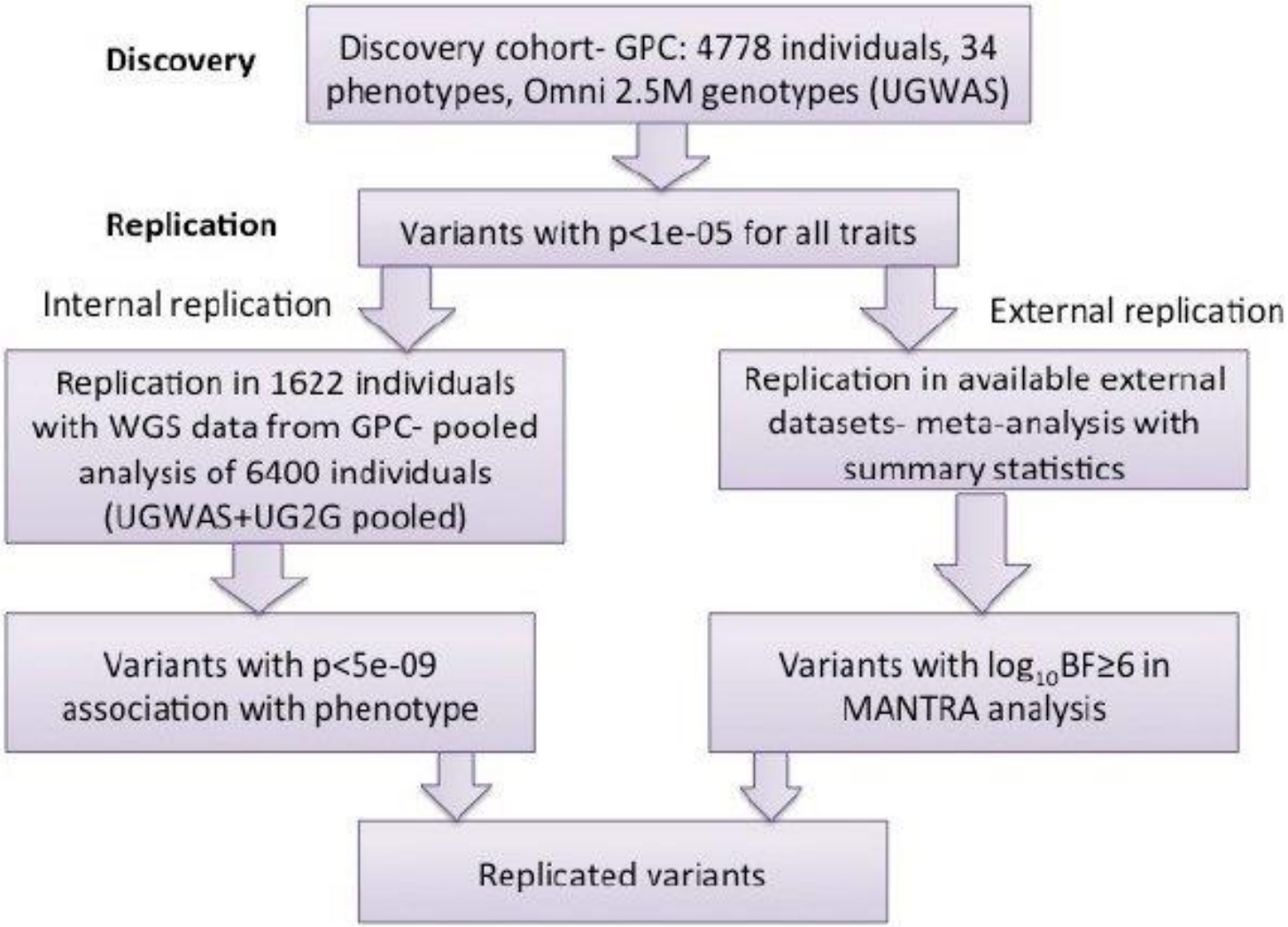
# Large-scale GWAS in an Ugandan cohort (2015-today)

- ~7000 individuals from the General Population Cohort
- (Asiki et al, IJE, 2013)
  - 2000 with whole genome sequence 4x
  - 4778 with Omni 2.5M genotypes
  - 50 phenotypic traits: hematological, anthropometric, blood pressure, metabolic, liver function and infectious disease traits


APCDR

# Schematic of Uganda GWAS discovery and replication



# Trans-ethnic meta-analysis

MANTRA, Andrew Morris *Genetic Epidemiology* 2011

- Allows populations from the same ethnic group to be more homogeneous than those that are more distantly related.
- Bayesian partition model that clusters populations according to their similarity in terms of relatedness (shared ancestry).
- Bayes factor in evidence of association, posterior probability of allelic effect, posterior probability of heterogeneity via MCMC.
- Hybrid meta-analysis, incorporating both fixed (within cluster) and random (between clusters) effects. Bayesian implementation of fixed-effects and random-effects meta-analysis.
- Improves power and resolution of fine-mapping, when heterogeneity in allelic effects is well represented by the prior Bayesian partition model.

# GWAS is a powerful tool

- Successful study design for identifying robust genetic associations with common disease
- Careful collection and quality-checking is essential to avoid errors
  - Phenotype misclassification
  - Population stratification
  - Traditional confounders
- Genetic effects of common variants are mostly moderate/small and require very large sample sizes to identify with certainty
  - Meta-analysis of GWAS improves the power of detecting and validating such associations
- GWAS only identifies regions of association
  - Causal variants identified by fine-mapping and targeted resequencing experiments
- Discovery of a genetic locus has important implications on its own
  - May highlight biological pathways and thus give insights into developing new therapeutics
- Population isolates can empower locus discovery
- Both discovery and fine mapping can be empowered by studying heterogeneous populations

# Future perspective

- Imputation based on the Haplotype Reference Consortium
  - A European haplotype map of over 50,000 haplotypes by combining together many low-coverage sequencing studies (1x-12x)
  - Next-generation resource for rare variant imputation into GWAS
  - Provides substantial increase over 1000 Genomes Phase 3 imputation

- Whole-genome sequencing-based meta-analysis consortia
  - Burden tests and meta-analysis of burden tests
  - How to best define regions

- The 100,000 Genomes Project
  - Genomics England, established in 2013, a company owned by the UK Department of Health
  - Linking to e-health records, sample size + interesting statistics
  - Marks the beginnings of a UK genomics industry and the start of a personalised medical service

- GWAS of non-European descent
  - The African Genome Variation Project
  - Large-scale GWAS in a Ugandan cohort
  - Trans-ethnic fine mapping

# Acknowledgements

**Principal Applicants**
Leena Peltonen, Wellcome Trust Sanger Institute
Richard Durbin, Wellcome Trust Sanger Institute

**Co-applicants**
Jeffrey Barrett, Wellcome Trust Sanger Institute
Inês Barroso, Wellcome Trust Sanger Institute
George Davey-Smith, University of Bristol
Ismaa Sadaf Farooqi, University of Cambridge
Matthew Hurles, Wellcome Trust Sanger Institute
Stephen O'Rahilly, University of Cambridge
Aarno Palotie, Wellcome Trust Sanger Institute
Nicole Soranzo, Wellcome Trust Sanger Institute
Tim Spector, King's College London
Nic Timpson, University of Bristol
Eleftheria Zeggini, Wellcome Trust Sanger Institute


**Named collaborators**
Phil Beales, University College London
Jamie Bentham, University of Oxford
Shoumo Bhattacharya, University of Oxford
Patrick Bolton, King's College London
Gerome Breen, King's College London
Krishnan Chatterjee, University of Cambridge
Laura K Curran, King's College London
Anne Farmer, King's College London
David Fitzpatrick, Edinburgh University
Daniel Geschwind, UCLA, USA
Steve Humphries, University College London
Jouko Lonnqvist, National Public Health Institute, Finland
Peter McGuffin, King's College London
Lucy Raymond, University of Cambridge
David Savage, University of Cambridge
Peter Scambler, University College London
Robert Semple, University of Cambridge
David St Clair, University of Aberdeen
Lennart von Wendt, University of Helsinki, Finland

UK 10K

# www.helic.org

Field teams
**Alina Farmaki** (Coordinator)
Emmanouil Tsafantakis (Director of Anogia Health Centre)
Maria Karaleftheri (Director of Echinos Health Centre)
Erica Daoutidou
Chrysoula Kiagiadaki
**George Dedoussis**
Ioanna Ntalla
plus numerous nurses, physicians, students

Sample logistics
Sarah Edkins
Emma Gray


Genotyping
Robert Andrews
Siobhan Whitehead
Doug Simpkin
Hannah Blackburn


Informatics
Josh Randall
Martin Pollard

Sequencing
John Burton
Danielle Walker
Sara Widaa
Jonathan Bailey

Analysis team
**Ioanna Tachmazidou**
**Loz Southam**
**Arthur Gilly**
**Kallia Panoutsopoulou**
**Kostas Hatzikotoulas**
**Graham Ritchie**
Denise Xifara
Angela Matchan
Will Rayner
Yuan Chen
Jeremy Schwartzentruber
Loukas Moutsianas
Yali Xue
Chris Tyler-Smith
Gil McVean
Enza Colonna

Research administration
Anja Kolb-Kokocinski
Carol Smee
Danielle Walker

University of Maryland
Toni Pollin
Jeff O'Connell
Laura Yerges-Armstrong

Harokopio University of Athens, Greece

The MANOLIS cohort is named in honour of Manolis Giannakakis, 1978-2010

African Partnership for Chronic Disease Research (APCDR)
Centre for Research on Genomics and Global Health (CRGGH)
Malaria Genomic Epidemiology Network (MalariaGEN)
Wellcome Trust Sanger Institute (WTSI)
1000 Genomes Project





**Manj Sandhu**
**Eleftheria Zeggini**
**Charles Rotimi**
**Fasil Ayele**
**Deepti Gurdasani**
**Tommy Carstensen**
**Chris Tyler-Smith**
**Luca Pagani**
**Yali Xue**
**Daniel Shriner**
**Cristina Pomilla**
**Kostantinos Hatzikotoulas**
**Jennifer Asimit**

Lucas Amanga-Etego
Joel M Francis
Asiki Gershim
Ramadhani Hashim
Katja Kivinen
Dominic Kwiatkowski
Georgina Murphy
Carolyne Ndila
Rebecca Nsubuga
Theo Papamarkou
Kirk Rockett
Ousmane Toure
Liz Young
Ananyo Choudhury
Segun Fatumo
Savita Karthikeyan

**APCDR, CRGGH: Collection Centre PIs**

**Uganda**
Anatoli Kamali
Janet Seely
Pontiano Kaleebu

**South Africa**
Ayesha Motala
Fraser Pirie
Michele Ramsey

**Ghana/Kenya/Nigeria**
Adebowale Adeyemo
Albert Amoah
Clement Adebamowo
Johnnie Oli

**Ethiopia**
Neil Bradman
Rosemary Ekong Endashaw Bekele
Tamiru Oljira
Ephrem Mekonen