

# Imputation and meta-analysis in genome-wide association studies

Kalliope Panoutsopoulou  
Applied Statistical Genetics Team  
Wellcome Trust Sanger Institute  
[kp6@sanger.ac.uk](mailto:kp6@sanger.ac.uk)

# GWAS Principles

Case control pairs (or population cohorts)



Type for 500k-1M SNPs



Data Quality Control (QC)



Test association at each SNP with the trait of interest



Prioritise signals and seek replication

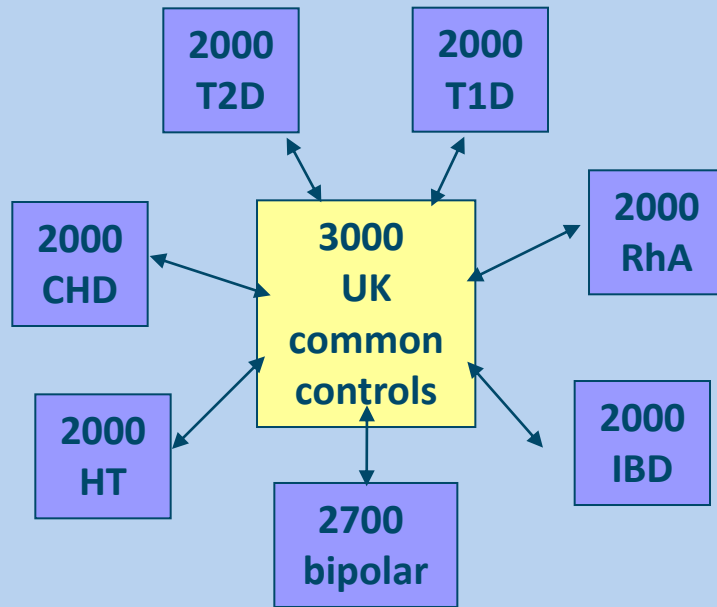


Combine your results with replication sets in a meta-analysis

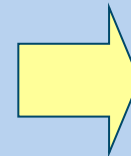


Establish (or not) association of SNP with disease

# Wellcome Trust Case Control Consortium GWAS Design



*Main study with national cases/controls*



500k Affymetrix

Data publically available

Vol 447 | 7 June 2007 | doi:10.1038/nature05911 nature

---

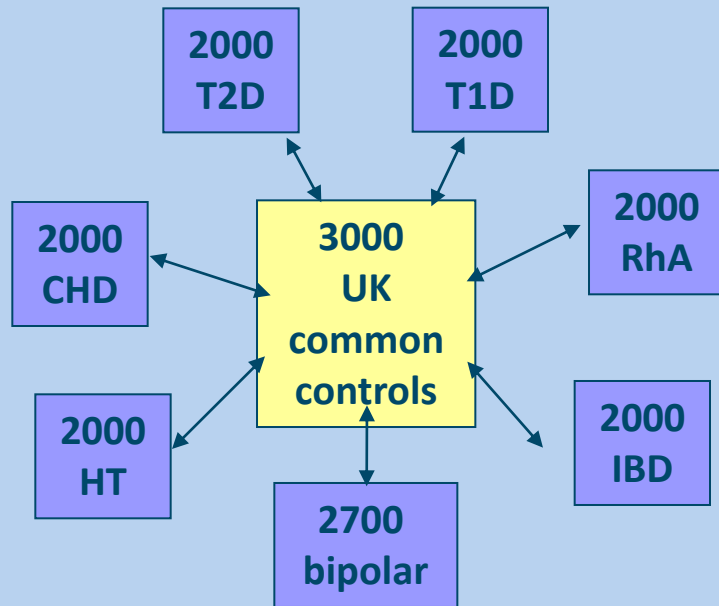
ARTICLES

---

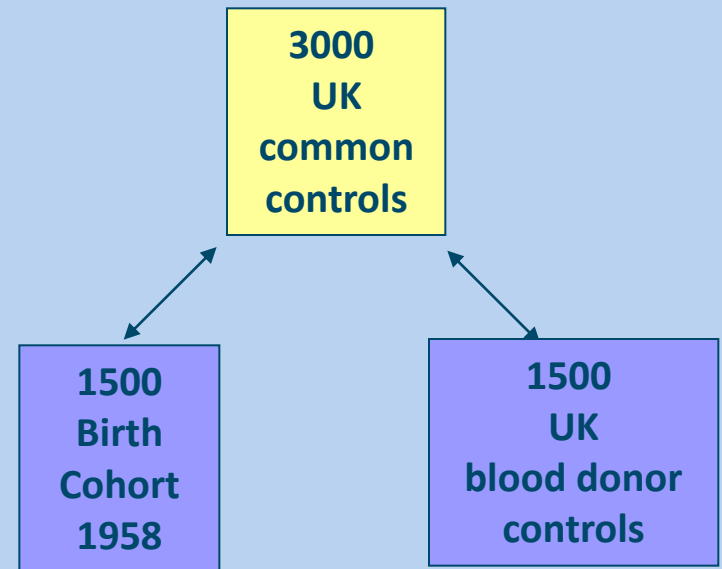
**Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls**

The Wellcome Trust Case Control Consortium\*

# WTCCC GWAS Design



*Main study with national cases/controls*



Nationally representative  
Unselected for disease phenotype  
Common controls  
Middle-aged – minimise survivor bias  
58BC also genotyped on Illumina 550k

Published Genome-Wide Associations through 12/2010,  
1212 published GWA at  $p \leq 5 \times 10^{-8}$  for 210 traits

2010 4th quarter



# The HapMap project

- Initiated in 2002 (20 groups in 6 countries)
- Aim: Characterise millions of sequence variants, their frequencies and the correlations between them in samples with ancestry from Africa, Asia and Europe
- HapMap phase 1  
30 CEU trios, 30 YRI trios, 45 CHB + 45 JPT  
~1.2 M SNPs
- HapMap phase 2  
~ 3.1 million SNPs
- HapMap phase 3 additional populations

# Imputation

Genotype imputation is the process of predicting (imputing) genotypes that are not directly genotyped in a sample of individuals

Imputation is typically performed by combining a *reference panel* of individuals genotyped at a dense set of polymorphic sites (SNPs) with a *study sample* genotyped at a subset of these sites

*Reference panel:* HapMap and/or 1000 Genomes and/or your own

*Study sample:* GWAS, a smaller region of interest

# Imputation

The shared ancestry of chromosomes in a population results in haplotype stretches which are shared by different individuals.

Making use of these shared haplotype stretches, and thereby accounting for the correlation of alleles at nearby markers (linkage disequilibrium, LD), statistical algorithms can make inferences about unobserved alleles.

To estimate a missing allele at a specific SNP on a haplotype, these algorithms compare flanking markers with those from other haplotypes in the sample to find appropriate “template” or reference haplotypes to inform an estimate of the missing allele.

# Genotype Imputation

Reference set of haplotypes e.g. HapMap

0000111001111110

1111101001000101

0000111001111110

1110110011101110

0010111001111110

The reference haplotypes are then used to impute alleles into the sample to create imputed genotypes

1???1?1?022??2?0

0???1?1?011??1?0

1???1?1?011??1?0

1111121002202220

Genotype data of an individual with missing data at untyped SNPs (?)

The individual is phased and the haplotypes are modelled as a mosaic of those in the haplotype panel

# SNP tagging-based approaches

**PLINK**

(<http://pngu.mgh.harvard.edu/~purcell/plink/>)

**SNPMSTAT**

(<http://www.bios.unc.edu/~lin/software/SNPMStat/>)

**UNPHASED**

(<http://www.mrc-bsu.cam.ac.uk/personal/frank/software/unphased/>)

**TUNA**

(<http://www.stat.uchicago.edu/~wen/tuna/>)

# Hidden Markov Model-based

**IMPUTE**

(<http://mathgen.stats.ox.ac.uk/impute/impute.html>)

**MACH**

(<http://www.sph.umich.edu/csg/abecasis/MACH/index.html>)

**BEAGLE**

(<http://faculty.washington.edu/browning/beagle/beagle.html>)

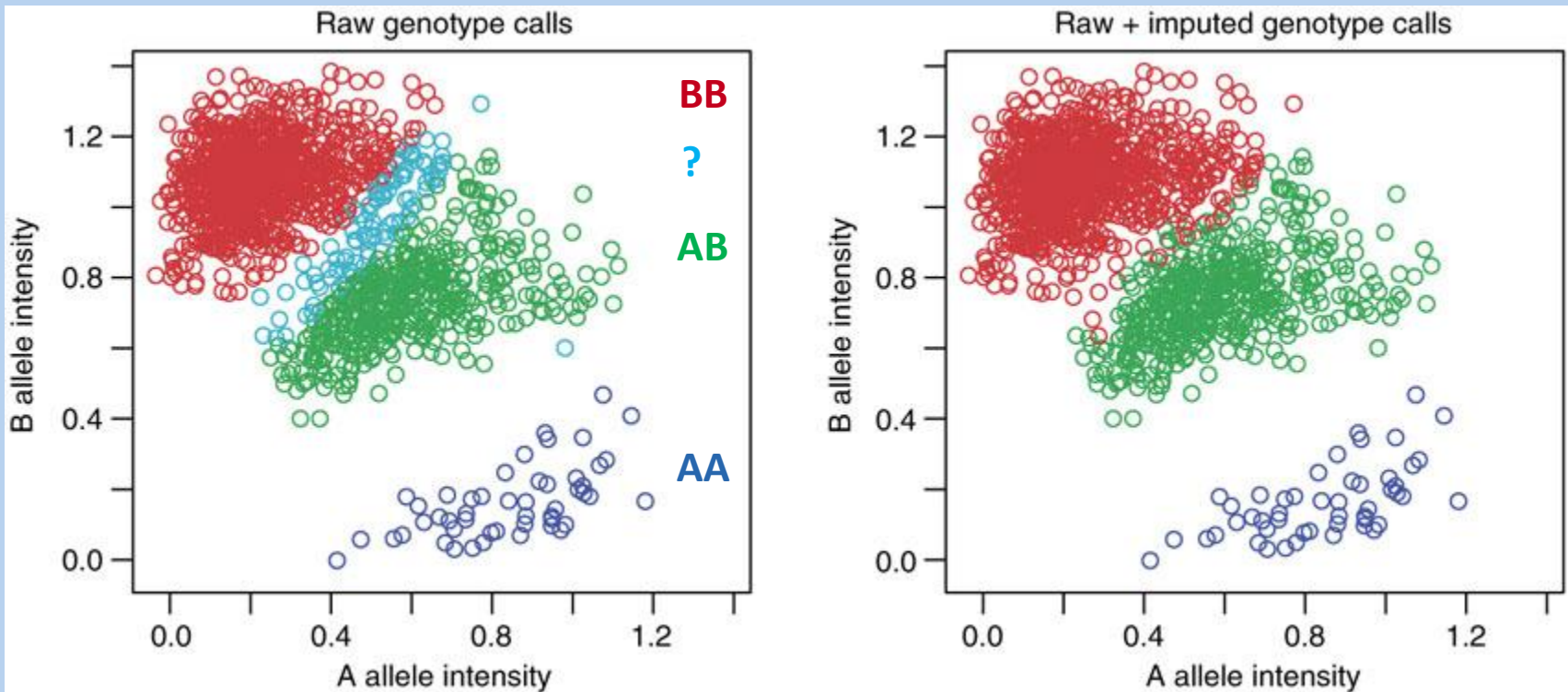
**fastPHASE**

(<http://stephenslab.uchicago.edu/software.html>)

# Application of Imputation to GWAS

- **Impute sporadic missing genotypes**
- **Impute genotypes at untyped markers by using a reference panel**
  - **Fine-mapping**
  - **Combine results of 2 or more studies genotyped on different platforms**

# Impute missing data



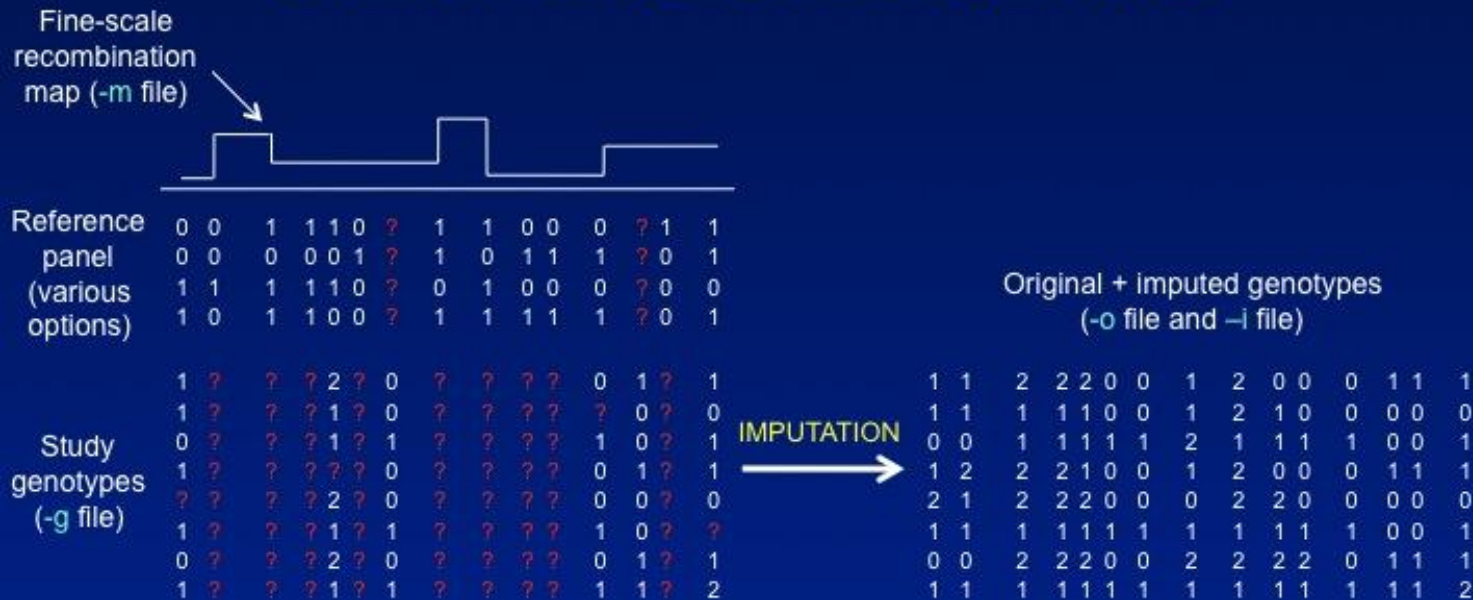
Taken from Marchini *et al.* (2007) Nat.Genet.

# Impute missing data

Individual	SNP 1	SNP 2	SNP 3	SNP 4
1	GG	AC	--	AT
2	GG	--	CC	AA
N	TT	AA	GG	AT

# Impute data at untyped SNP

## Overview of imputation using IMPUTE2



This figure over-simplifies what IMPUTE2 does. The output for each genotype is actually a probability distribution:

Genotype	0	1	2
Probability	0.01	0.18	0.81

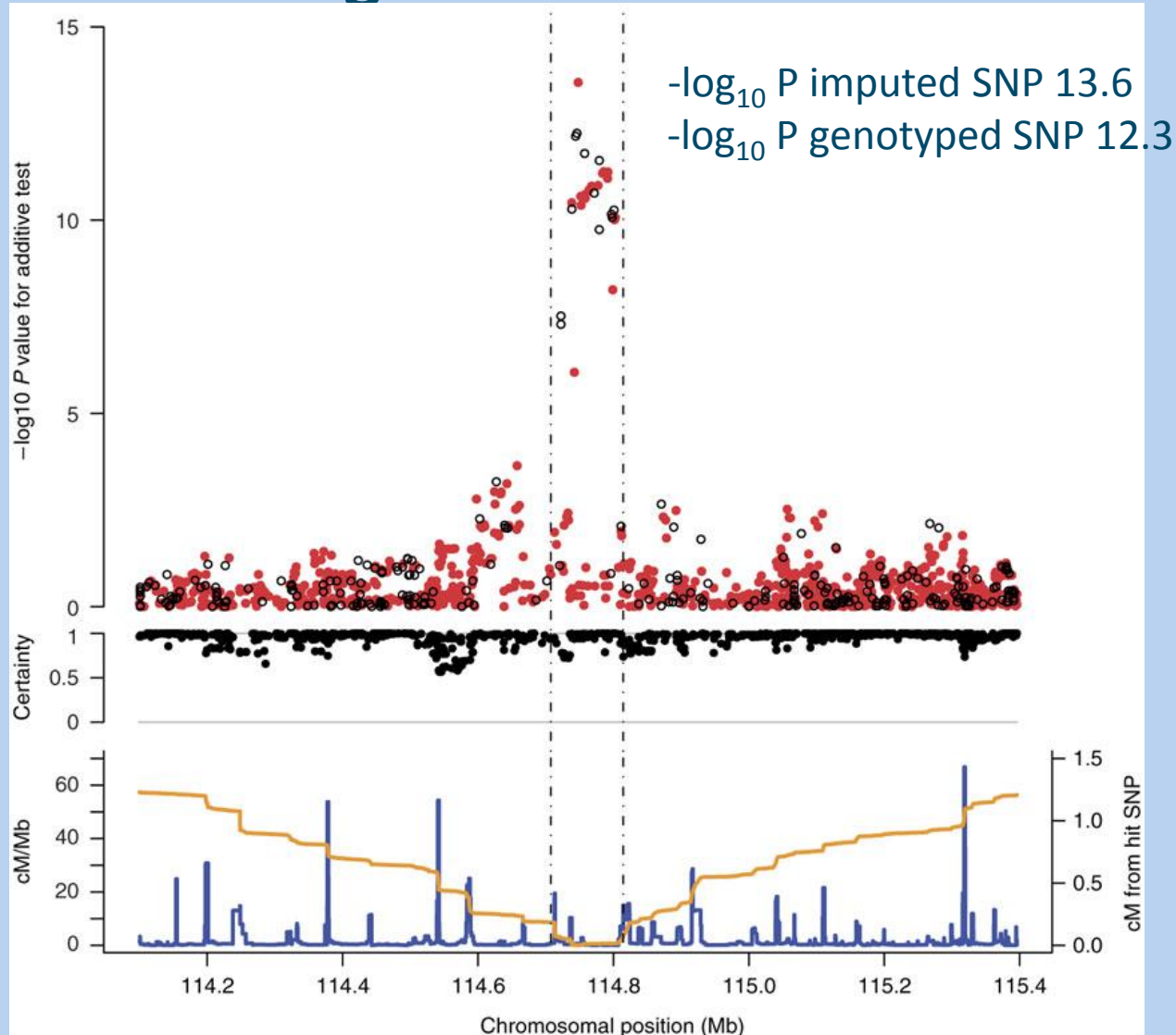
This captures the uncertainty in the prediction.

Howie et al., 2009, PLOS Genetics

[http://mathgen.stats.ox.ac.uk/impute/impute\\_v2.html](http://mathgen.stats.ox.ac.uk/impute/impute_v2.html)

# Fine mapping

## TCF7L2 region WTCCC T2D GWAS

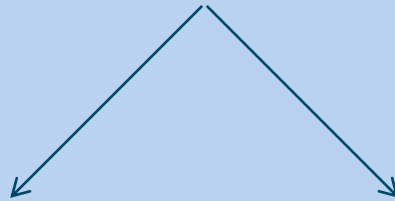


Taken from Marchini *et al.* (2007) Nat.Genet.

# Replication of top signals

## Signal prioritisation

## Validation in one or more independent studies



### *in silico* replication

existing GWAS data  
for the same phenotype  
in independent samples

### *de novo* replication

genotype SNPs of interest  
in independent samples using  
a different genotyping method

# *In silico* Replication

Study	SNP 1	SNP 2	SNP 3	SNP 4
Discovery GWAS Affy 500k	typed	typed	typed	typed
Replication set 1 Affy 500k	typed	Failed QC	typed	typed
Replication set 2 Illumina 610k	Not in chip	Not in chip	Not in chip	typed

# Replication of top signals from a GWAS

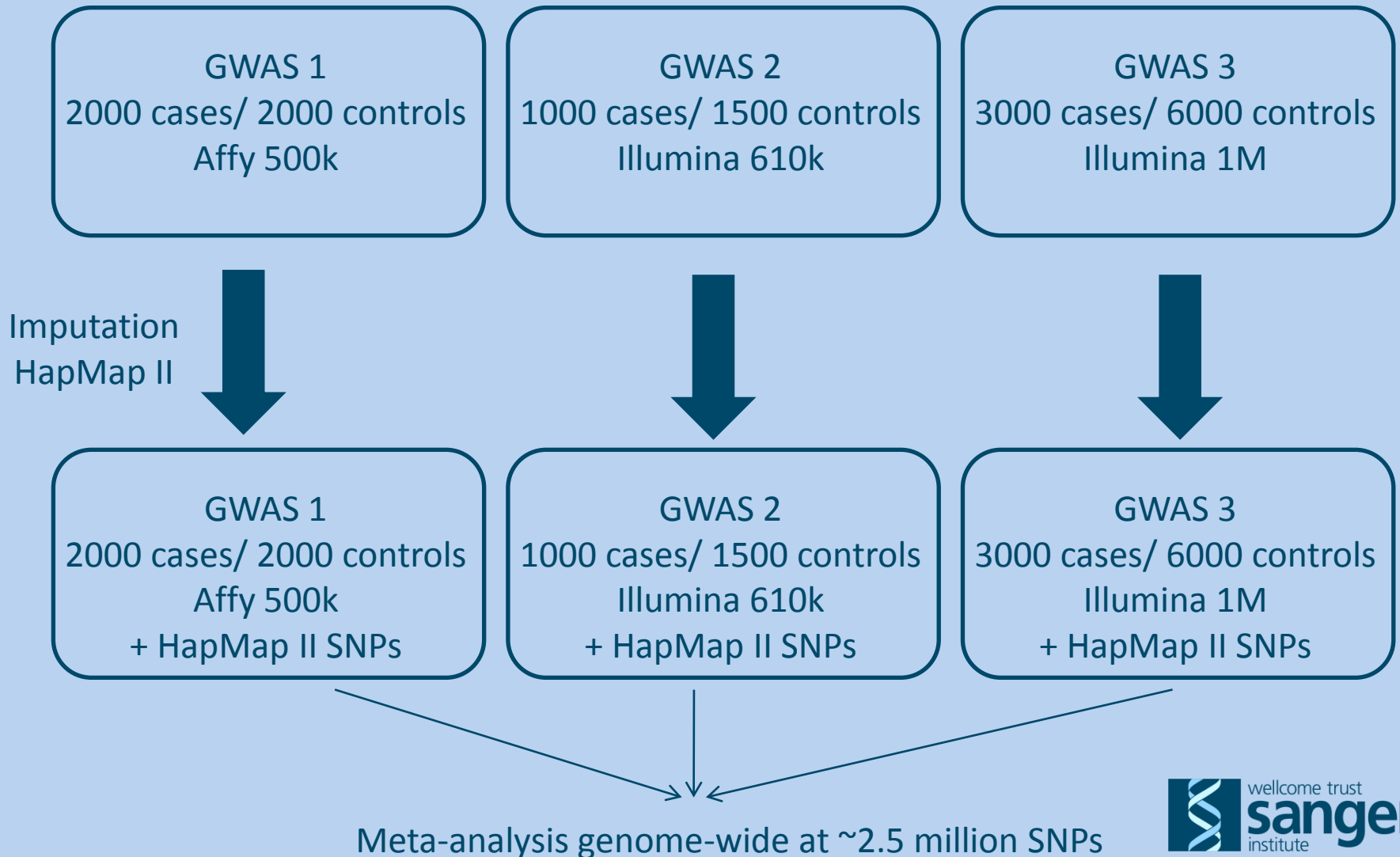
Following imputation in the replication sets

Study	SNP 1	SNP 2	SNP 3	SNP 4
Discovery GWAS Affy 500k	typed	typed	typed	typed
Replication set 1 Affy 500k	typed	imputed	typed	typed
Replication set 2 Illumina 610k	imputed	typed	imputed	typed



Obtain summary results for all SNPs from each replication study  
Combine in a meta-analysis framework

# Genome-wide meta-analysis



# Association Testing of Imputed Data

Genotype	0	1	2
Probability	0.01	0.18	0.81

- Accounting for uncertainty in imputed genotypes – an observed data likelihood is used in which the contribution of each possible genotype is weighted by its imputation probability followed by a Score test
- Test using expected genotype counts (i.e. the sum of the probabilities across all individuals in the sample) using logistic or linear regression
- Test using best guess genotype or imputed genotypes with posterior probability above some threshold is not recommended

# Post imputation QC

## Assess quality of imputation

- 1/ Correlation of imputed genotypes to directly typed genotypes
- 2/ Imputation info score  
[0 complete uncertainty, 1 no uncertainty]  
An information measure of  $\alpha$  on a sample of  $N$  individuals indicates that the amount of data at the imputed SNPs is equivalent to a perfect set of observed genotype data in a sample size of  $\alpha N$
- 3/ Rare SNPs in HapMap based imputation do not impute very well

# Factors that affect imputation accuracy

Data quality control (QC)

Study population

Properties of the reference panel

Genotyping chip

Common vs rare alleles

# QC before imputation

## Sample QC

Call rate

Excess heterozygosity

Sex discrepancies

Ethnic outliers

Related/Duplicated

## SNP QC

Call rate

HWE  $<10^{-4}$

MAF  $<1\%$

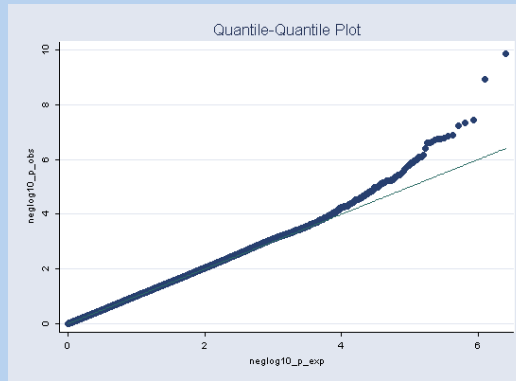
Intensity plots (where possible)

# Effect of QC pipeline on imputation with illustrative examples:

## UKBS vs 58BC directly typed and imputed SNPs

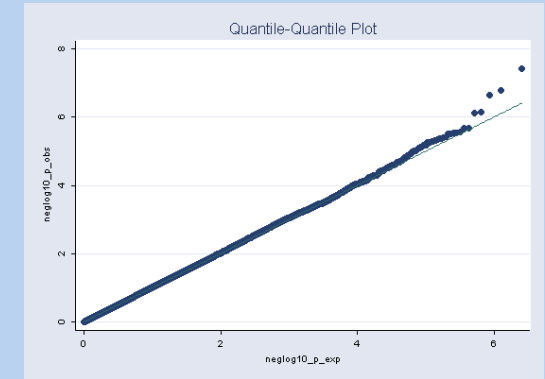
### WTCCC QC criteria

HWE exact  $p < 5.7 \times 10^{-7}$   
Intensity plots



### Our QC criteria

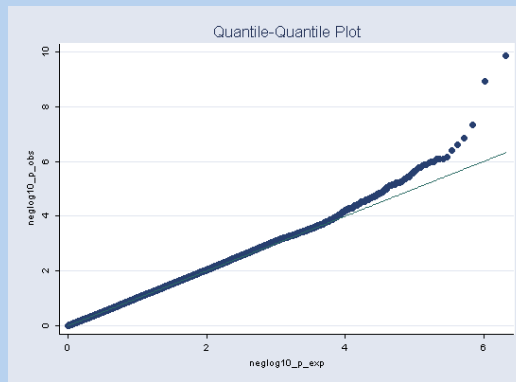
HWE exact  $p < 10^{-4}$   
+ intensity plots  
Re-imputation based  
on a cleaner set



### WTCCC QC criteria

+

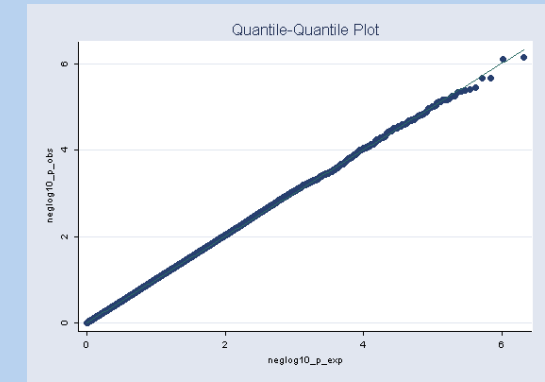
Post-imputation filters  
MAF < 0.01, info < 0.8



### Our QC criteria

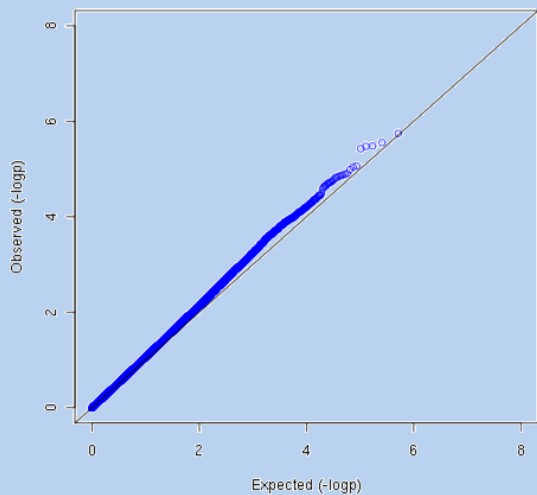
+

Post-imputation filters  
MAF < 0.01, info < 0.8

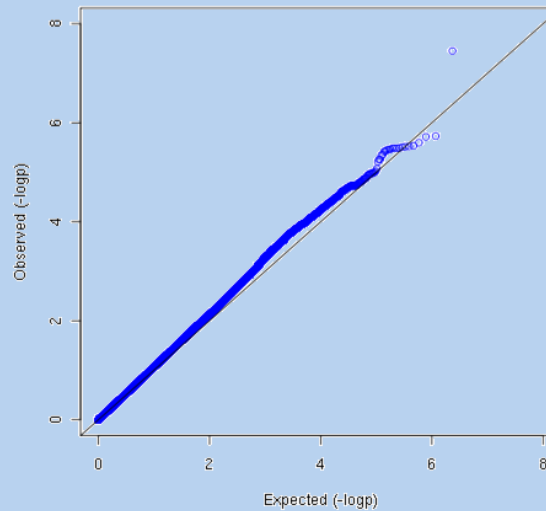


# Osteoarthritis GWAS

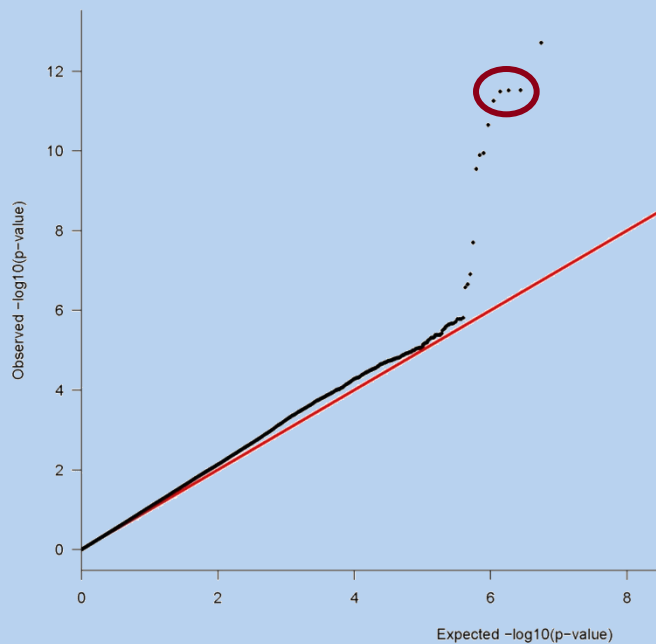
Directly typed SNPs (Illumina 610k)



HapMap 2 based imputation

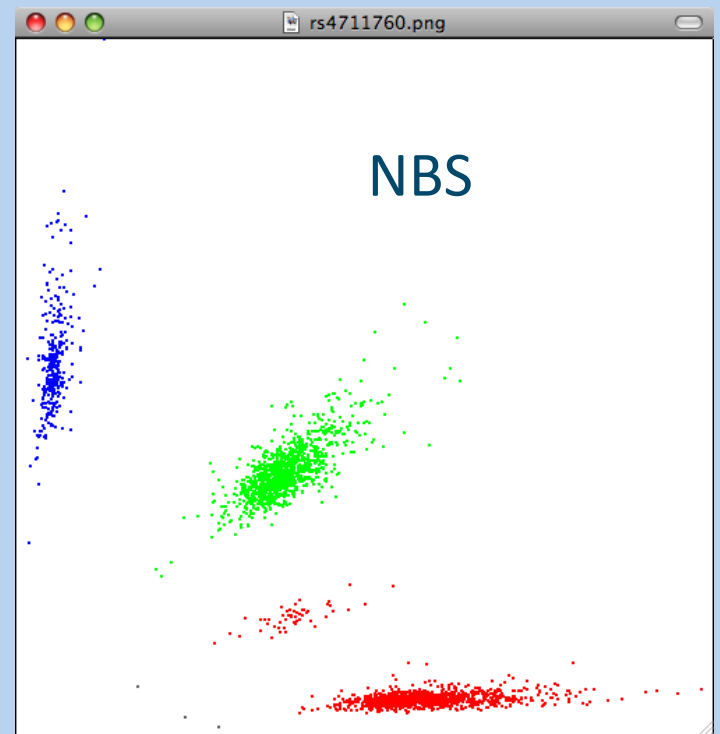
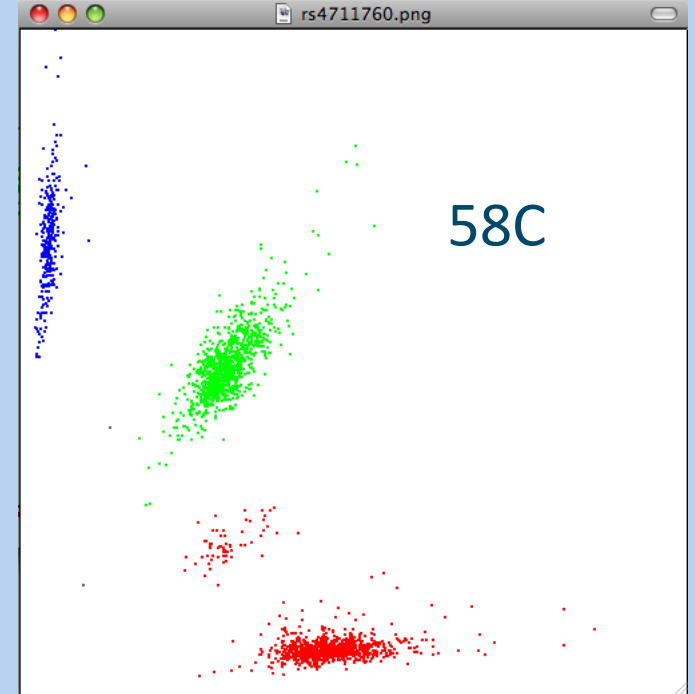
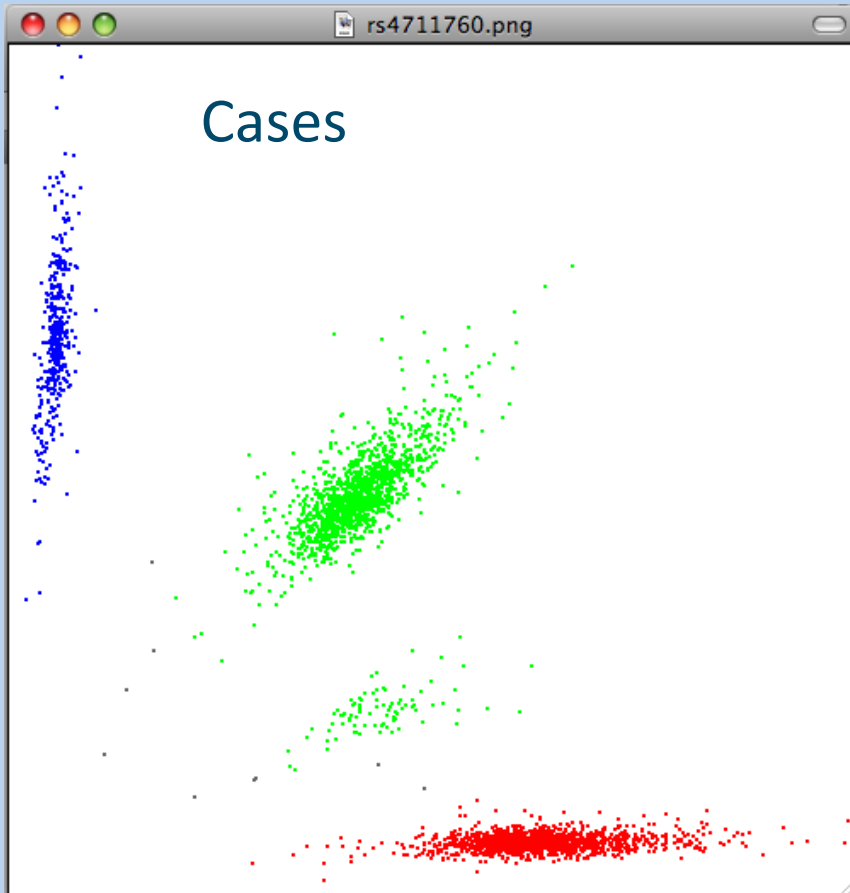


1 KG based imputation



rsID*	Distance to next (bp)	MAF	r <sup>2</sup>	Allele count in 1000G	OR	P value
-	156	0.015	0.64	3	3.09	5.6x10 <sup>-12</sup>
-	493	0.019	0.72	4	2.51	2.3x10 <sup>-11</sup>
-	745	0.019	0.71	4	2.43	1.3x10 <sup>-10</sup>
-	1461	0.015	0.61	3	2.93	1.1x10 <sup>-10</sup>

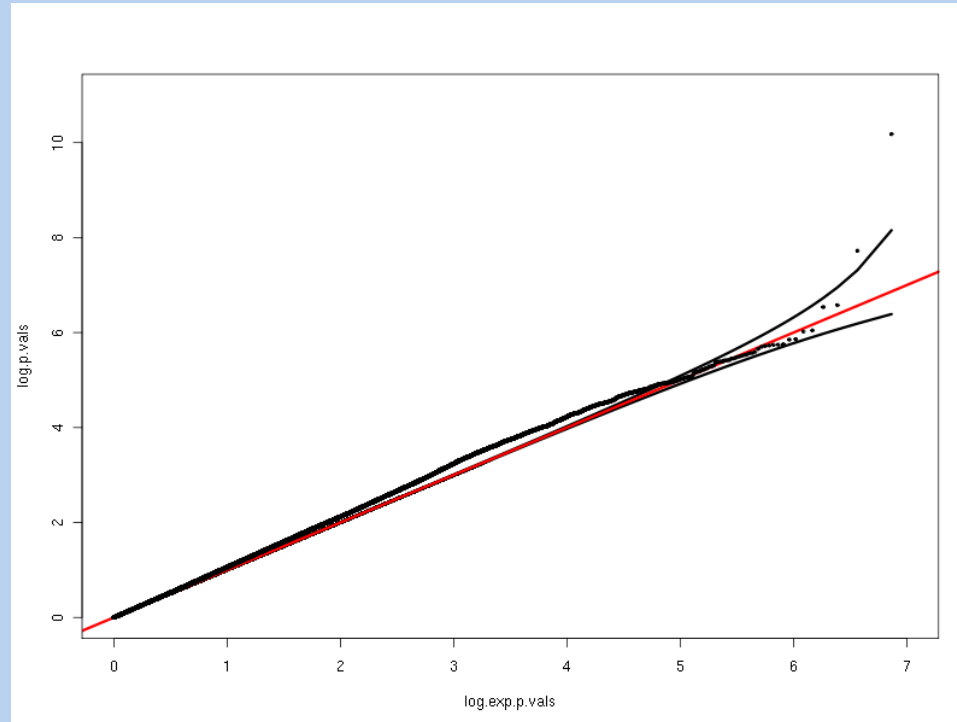
\*non-HapMap, non-dbSNP variants



# Inspect and remove bad intensity plots

## Re-impute based on a cleaner set

Imputed SNPs: 1 KG



# Imputation performance of various chips in HapMap 2

Imputation error rate is higher as MAF decreases

Imputation accuracy is higher for a larger reference panel

Imputation performs better when using a combination of CEU, YRI and JPT+CHB reference panels especially at rare SNPs compared to a single panel

Imputation in CEU performs better for Illumina chips rather than Affymetrix chips.

# Imputation for populations worldwide

Imputation based on a single most appropriate reference panel from HapMap 2 depending on study population (n=29)

Europe  
Asia  
America  
Oceania  
Middle East  
Africa



Imputation Accuracy

Considerable variation in imputation accuracy for populations within Africa

# HapMap Phase 3 reference panel

HapMap 3 has 10 distinct sets of haplotypes and larger number of haplotypes in each set e.g. 330 CEU

This allows more accurate imputation of rare SNPs

Imputation using all HapMap 3 populations is recommended even if the study population is CEU

# 1000 Genomes Project

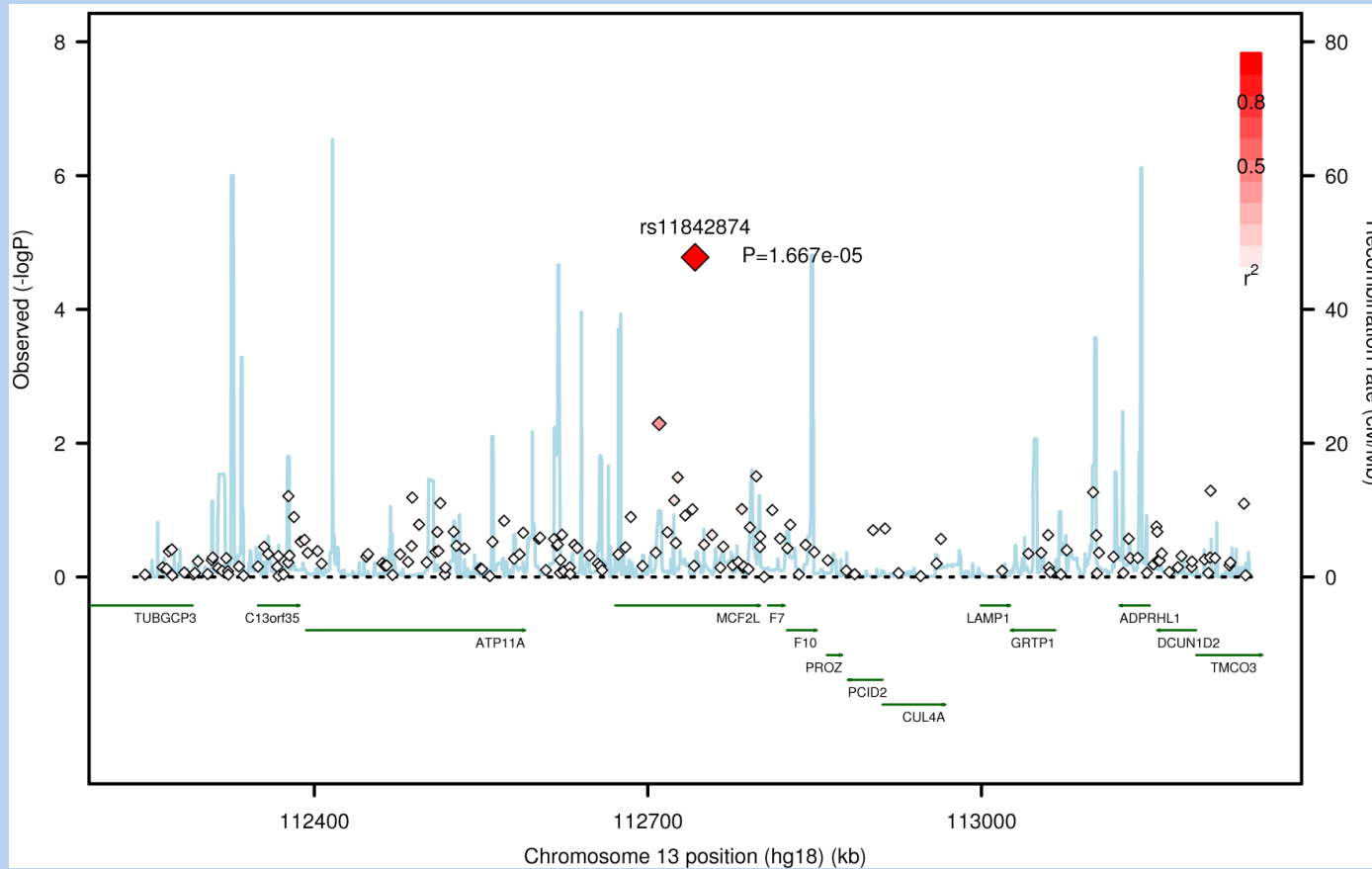
**Initiated in 2008**

**Aim: To create a public reference database for DNA polymorphism that is 95% complete at allele frequency 1% and more complete for common variants and exonic variants in multiple human populations**

**Large increase in terms of number of SNPs and number of samples will allow more accurate imputation for the majority of SNPs above 1% frequency**

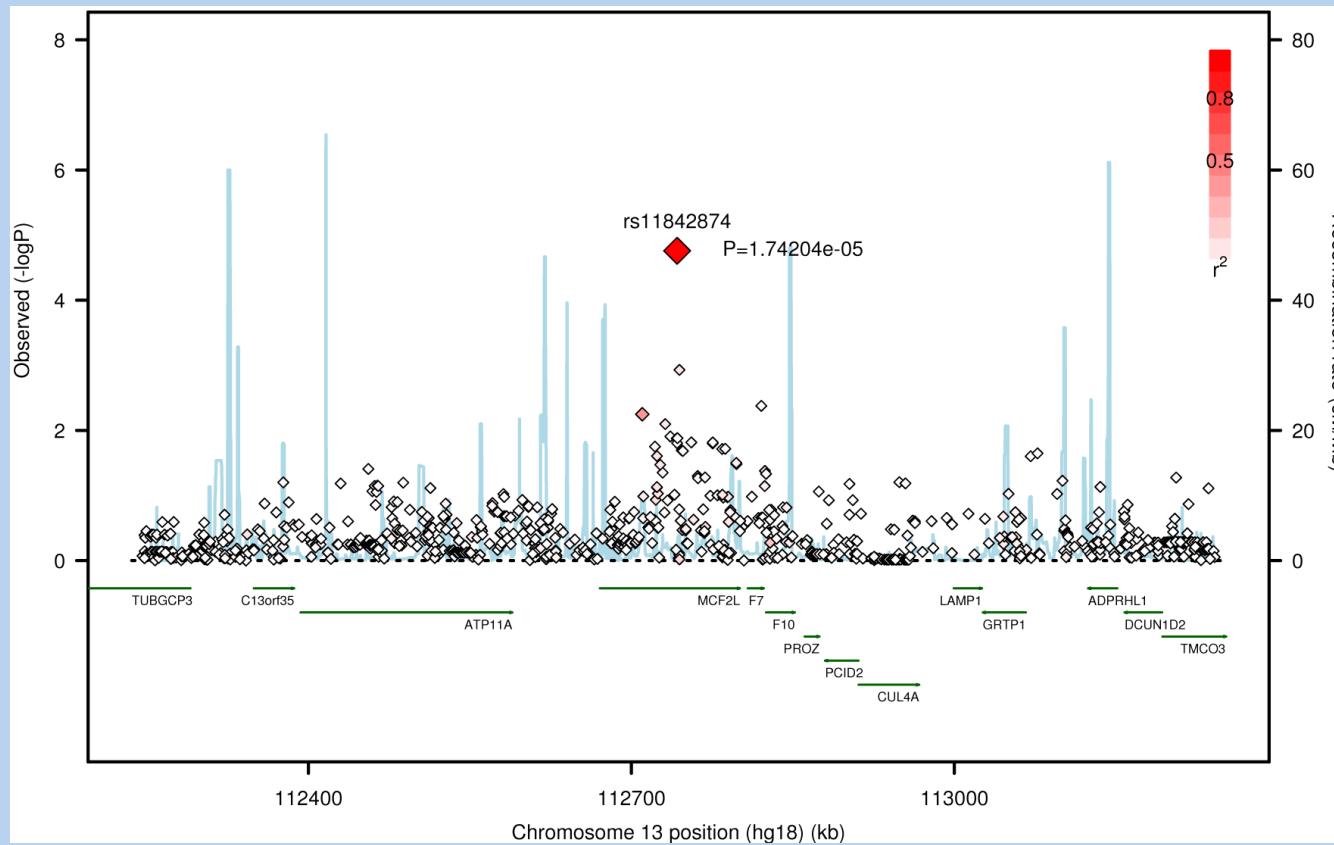
# Osteoarthritis GWAS

## Directly typed SNPs



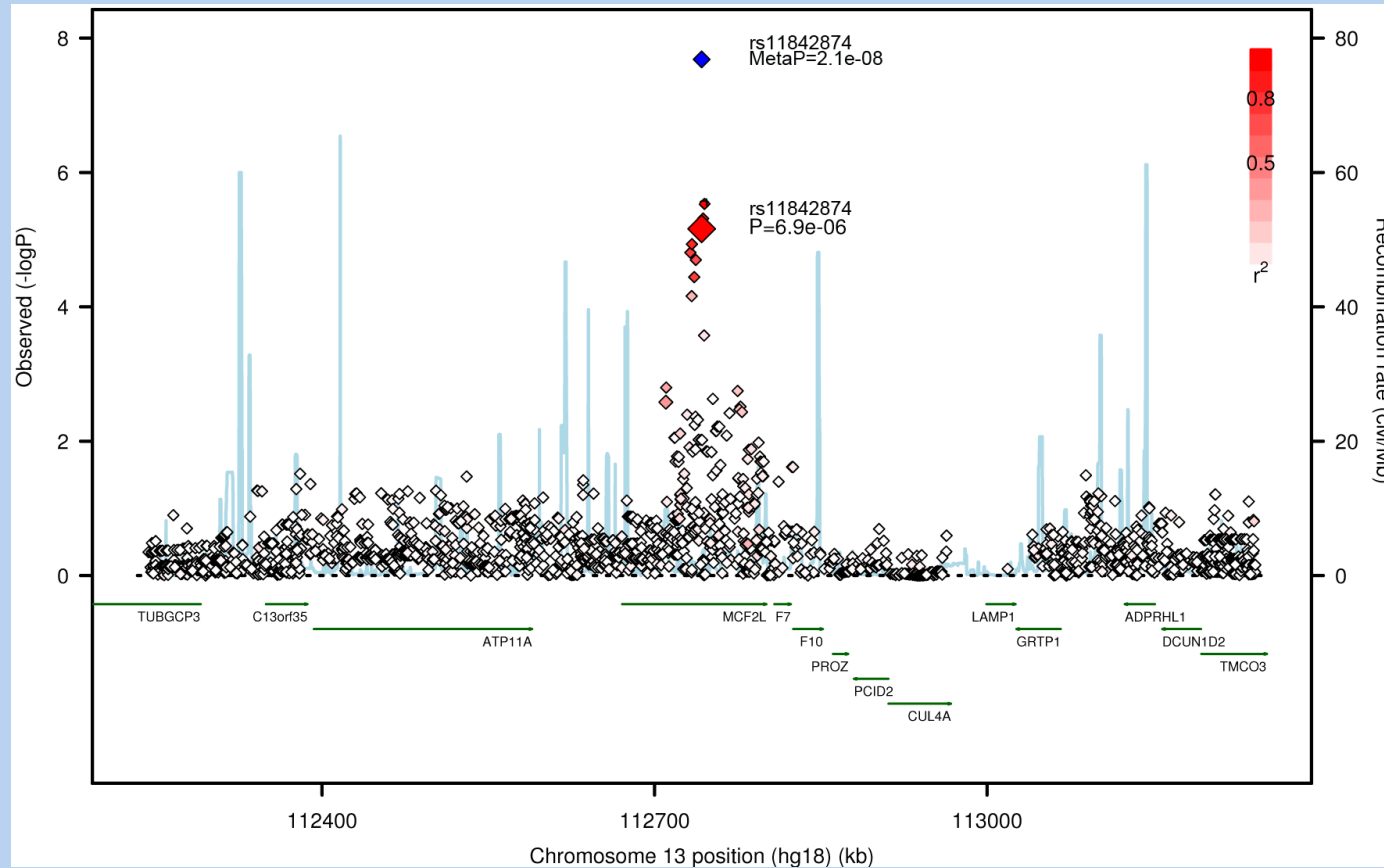
# Osteoarthritis GWAS

Directly typed + HapMap 2 based imputation



# Osteoarthritis GWAS

Directly typed + 1KG based imputation



# Imputation can boost power

Case control pairs (or population cohorts)



Type for 500k-1M SNPs



Data Quality Control



Test association at each SNP with the trait of interest



Follow up prioritised signals by replication



Combine your results with replication sets in a meta-analysis



Imputation can bridge the gap between different platforms

Imputation will increase number of SNPs that can be tested



Imputation can fine-map the signal



# Meta-analysis

**Meta-analysis is a set of methods that allows the quantitative combination of data from multiple studies**

**The synthesis of different datasets leads to a results summary based on evidence from the combined data**

**These methods also allow the quantitative evaluation of the consistency or inconsistency/heterogeneity of the results across multiple datasets**

# Motivation for meta-analysis of GWAS

Data from diverse studies examining the same trait can be synthesised within a meta-analysis framework

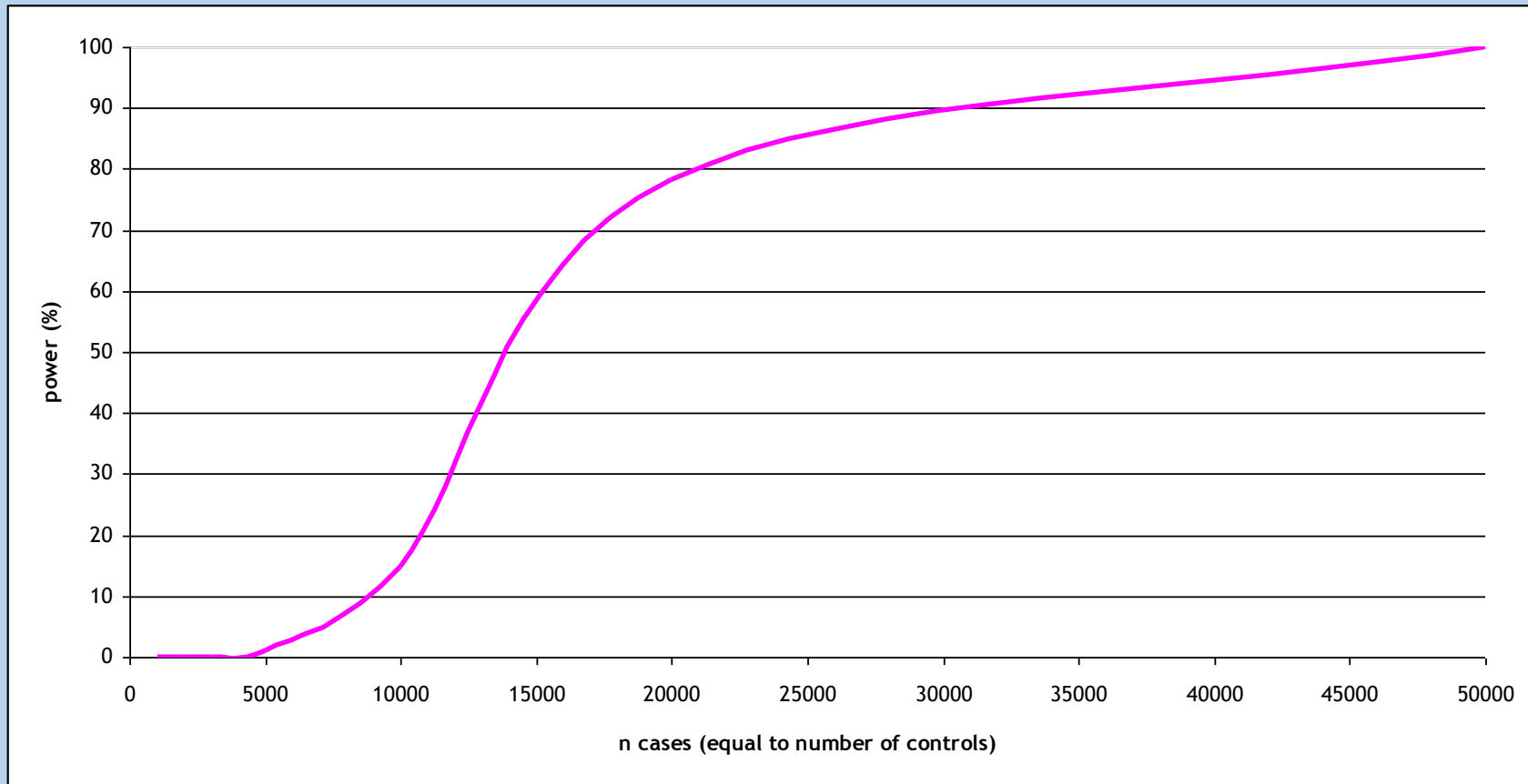
Increase in power through:

- Increased sample size

- Imputation of untyped variants

# Motivation for meta-analysis of GWAS

Power to detect association ( $p=5 \times 10^{-8}$ ) at a variant with risk allele frequency 0.30 and allelic OR 1.10



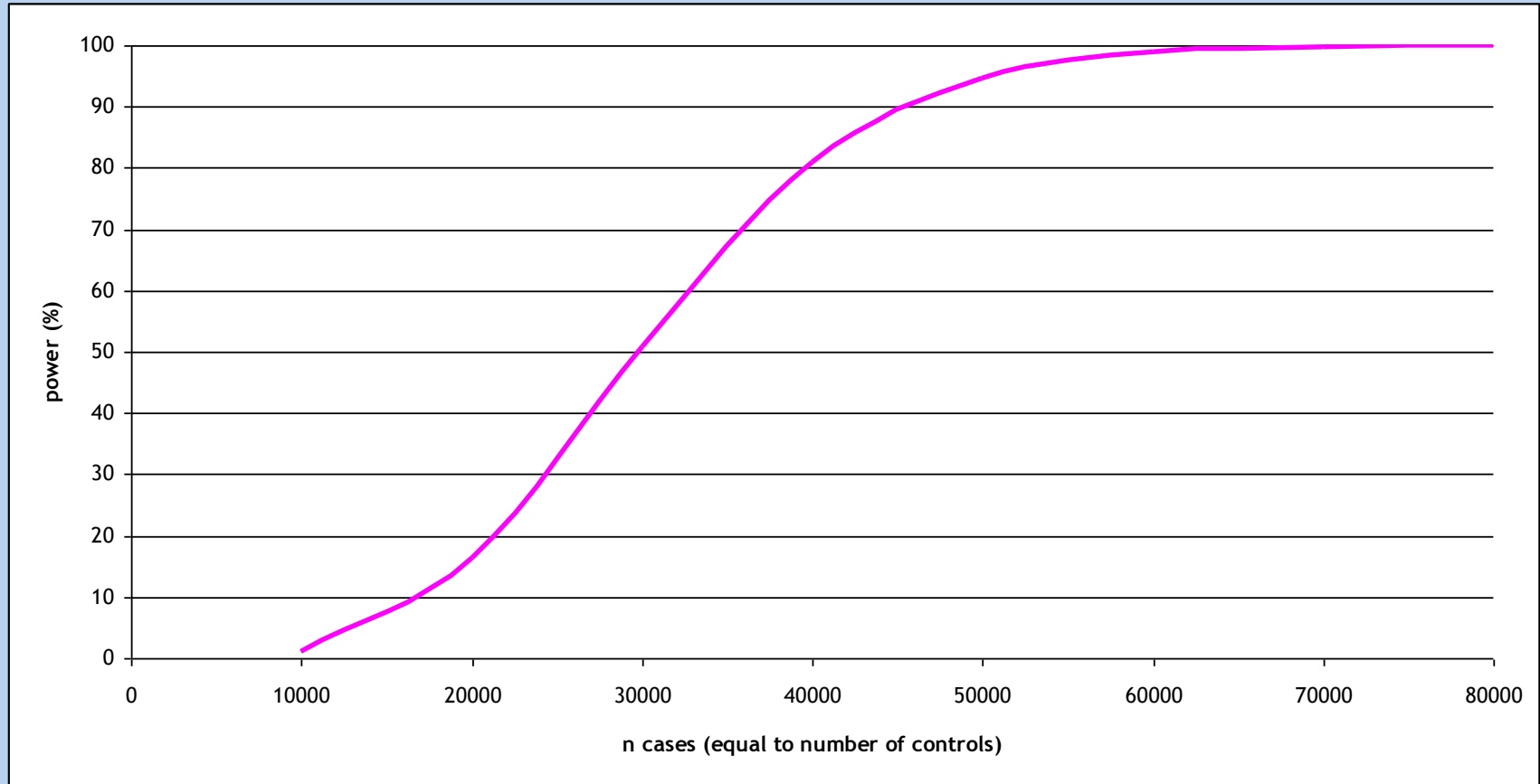
# Motivation for meta-analysis of GWAS

Recent shift of the complex trait genetics field towards low frequency (between 1% and 5%) and rare (less than 1%) variation

The issue of sample size and power is much more pronounced in the study of rare variation, especially because as yet unidentified very large effect sizes are unlikely to exist for polygenic disorders

# Motivation for meta-analysis of GWAS

Power to detect association ( $p=5 \times 10^{-8}$ ) at a variant with risk allele frequency 0.005 and allelic OR 1.50



# Meta-analysis -principles

**Synthesis of association summary statistics from different datasets examining the same phenotype to obtain a summary based on evidence from the combined data**

**In epidemiological terms, meta-analyses provide a better estimate of effect size**

**In the GWAS setting, meta-analysis is usually initially carried out to help the discovery of further susceptibility variants of moderate/ small effect size that would have otherwise escaped detection due to low power**

# Meta-analysis -principles

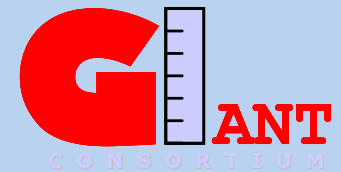
Facilitated by imputation, which enables the combination of data across different genotyping platforms

Reference datasets (e.g. [www.hapmap.org](http://www.hapmap.org); [www.1000genomes.org](http://www.1000genomes.org)), can be used to impute genotypes for all variants at untyped positions in the target dataset of interest, using the GWAS genotypes as a scaffold

Meta-analyses can be carried out sequentially, and can be updated when new GWAS datasets for the same trait emerge

# Forming consortia

The first step in many GWAS meta-analyses involves setting up consortia to study specific traits of interest



DIAbetes Genetics  
Replication And Meta-analysis



Robust GWAS meta-analyses require a robust predefined protocol, specifying basic analytical options, such as genetic model examined, strategy for covariate adjustments etc

The majority of GWAS meta-analyses combine data retrospectively, making harmonisation of study design extremely difficult



# Meta-analysis of GWAS

Requires summary statistics at each variant

Information on analysis method and covariates used

Information on the size of the study

Information on the independence of samples

Information on approaches taken to adjust for any population stratification (for example genomic control)

Information on strand and build of the human genome, on which allele coding has been based

# Meta-analysis of GWAS

**Genotyping platforms**

**Limited overlap between variants**

**Imputation of untyped SNPs**

**Analysis of directly typed and imputed variants  
for each scan separately**

**Combination of summary results across studies**

**Importance of quality control**

# Standardising QC and analysis method

Phenotype definition

Genotype level QC (call rate, HWE, MAF)

Imputation (reference, analysis)

Association testing

Correcting for population structure prior to the meta-analysis

Centralised v distributed meta-analysis

# Typical data sharing table format

<b>STUDY TITLE</b>	
<b>General information</b>	
<b>Name of study</b>	
<b>Name of analyst</b>	
<b>Email of analyst</b>	
<b>Study design</b>	population-based, family-based –please give details
<b>Sample information</b>	
<b>Number of cases (females)</b>	
<b>Number of controls (males)</b>	
<b>Ethnic composition</b>	
<b>Possible relatedness issues</b>	are individuals related (how?)
<b>Possible structure issues</b>	mixed population?
<b>Genotyping and imputation information</b>	
<b>Genotyping platform</b>	
<b>Summary of key QC metrics</b>	
<b># SNPs passed QC</b>	
<b>Imputation method</b>	
<b>Imputation settings</b>	
<b>Reference data used for imputation</b>	including build
<b>Analytical information</b>	
<b>Association analysis method for imputed genotypes</b>	accounting for uncertainty using SNPTEST or other (which?) program, using only genotypes with $P(\text{call}) > X$ (which threshold?) as hard calls, using best guess genotypes
<b>Calculated GC lambda (typed SNPs)</b>	
<b>Calculated GC lambda (imputed SNPs)</b>	
<b>Covariates included</b>	PCA, GC, none
<b>Genetic model</b>	

# Typical data sharing table format

Column header	Description
SNP	SNP rs number (if unknown, e.g. with some Affymetrix SNPs, report Affy SNP ID)
build	e.g. "36", human genome build used
strand	e.g. "+", human genome strand used
chromosome	chromosome on which SNP resides
position	position of SNP on chromosome in base pairs, based on human genome build used
imputed	"1" for imputed, "0" for directly-typed SNP passing QC
major_allele	e.g. "G", major allele at that SNP, based on control frequency
minor_allele	e.g. "A", minor allele at that SNP, based on control frequency
MAF_controls	e.g. "0.246", minor allele frequency in controls -provide 3 digits to the right of the decimal
OR_allele	e.g. "A", allele to which the OR has been estimated
call_rate	e.g. "0.985", call rate for this SNP across cases and controls -provide 3 digits to the right of the decimal
exact_HWE_cases	exact HWE p value in cases
exact_HWE_controls	exact HWE p value in controls
OR	e.g. "1.097", allelic odds ratio -provide 3 digits to the right of the decimal
lower_95%CI	e.g. "0.874", lower 95% confidence interval of the OR -provide 3 digits to the right of the decimal
upper_95%CI	e.g. "1.267", upper 95% confidence interval of the OR -provide 3 digits to the right of the decimal
additive_p_uncorr	additive model p value, uncorrected for genomic control
additive_p_corr	additive model p value, corrected for genomic control
impute_info_score	e.g. "0.98", metric for imputation accuracy (i.e. value for $r^2_{\text{hat}}$ or proper_info measures, depending on imputation programme used; if some other measure used, please specify)

# Assessment of heterogeneity

The first step in carrying out a meta-analysis involves assessing heterogeneity across the combined studies

## Statistics

Cochran's Q: is there heterogeneity?

$I^2$ : how much heterogeneity is there?

# Reasons underlying heterogeneity

**Chance**

**Errors and biases differently affecting the results of different datasets**

**Variable LD between the typed marker and the causal variant**

**Genuine differences in genetic effects across different populations**

# Interpreting heterogeneity

Profile of extreme signals emerging from GWA scans can be profoundly affected by ascertainment scheme –informative heterogeneity

Recognising the potential for heterogeneity (most easy to do when one has large signals and large sample sets) can “rescue” associations from being discarded as replication failures reveal useful facts about biology, in this case the mechanism through which the variant is acting on disease risk

The obesity-associated FTO locus represents a prime example of informative heterogeneity. Discovered through the Wellcome Trust Case Control Consortium (WTCCC) type 2 diabetes (T2D) GWAS, in which cases and controls had not been matched for body mass index (BMI) [WTCCC, 2007; Zeggini et al, 2007; Frayling et al, 2007], this signal was not replicated in further T2D GWAS [DGI, 2007; Scott et al, 2007], which had matched cases and controls for BMI.

# Meta-analysis

Based on estimate of effect size (e.g. OR-based)

Must have independent set of effect sizes

Larger studies should carry more weight

Weight each effect size by the inverse variance

Fixed effects

Random effects

Based on p value

Bayesian

# Specialised software for GWAS meta-analysis

## GWAMA

Genome-Wide Association Meta-Analysis;

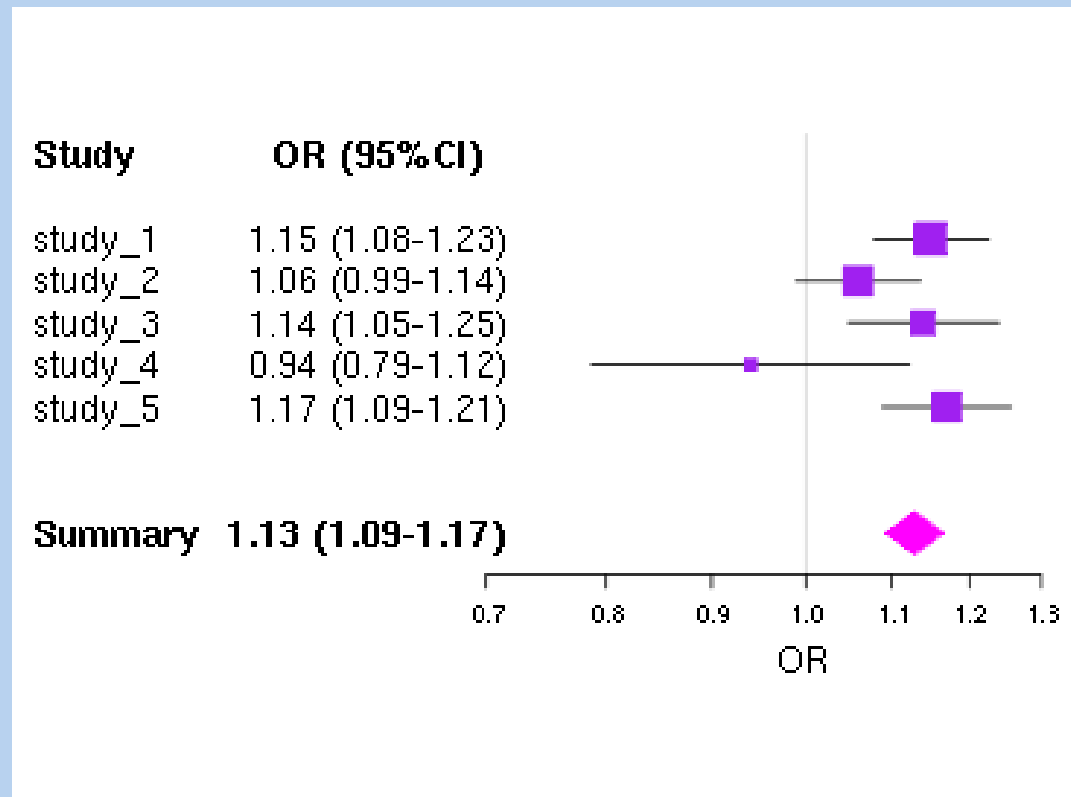
[www.well.ox.ac.uk/gwama/contact.shtml](http://www.well.ox.ac.uk/gwama/contact.shtml)

## METAL

Meta-Analysis Helper;

[www.sph.umich.edu/csg/abecasis/metal](http://www.sph.umich.edu/csg/abecasis/metal)

# Visualising meta-analysis results



# References

Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* 447, 661-678 (2007)

Marchini, J., Howie, B., Myers, S., McVean, G. & Donnelly, P. A new multipoint method for genome-wide association studies by imputation of genotypes. *Nat. Genet* 39, 906-913 (2007)

Li Y, Willer C, Sanna S, Abecasis G. Genotype imputation. *Annu Rev Genomics Hum Genet.* 10:387-406 (2009)

Browning SR. Missing data imputation and haplotype phase inference for genome-wide association studies. *Hum Genet*;124(5):439-50 (2008)

Zeggini E, Ioannidis JP. Meta-analysis in genome-wide association studies. *Pharmacogenomics.* 10(2):191-201 (2009)