Imperial College London

Royal Brompton & Harefield NHS
NHS Foundation Trust

GWAS, meta-analysis and imputation

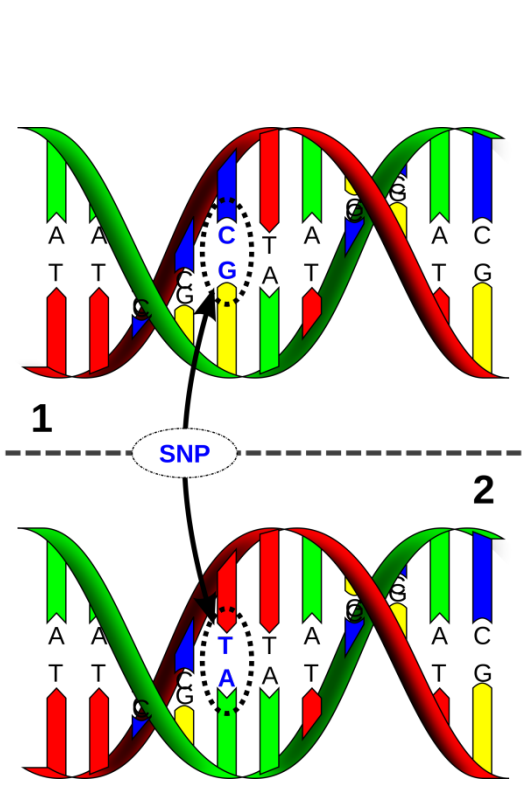
Dr Nicola Whiffin

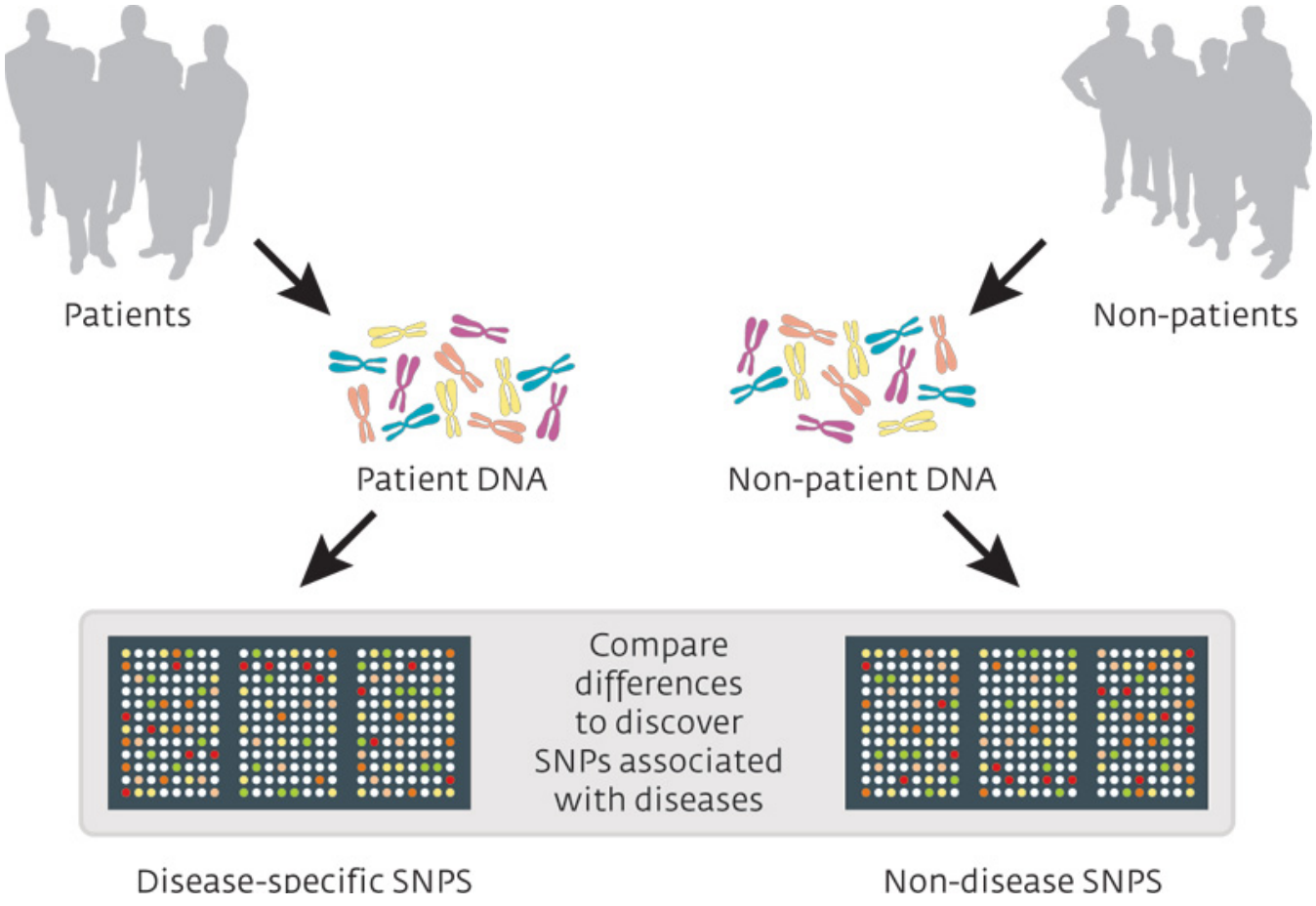
n.whiffin@imperial.ac.uk

Outline

- GWAS
  - Linkage disequilibrium
  - Quality control
  - Analysis methods
  - Limitations
- Meta analysis
- Imputation

Genome-wide association studies



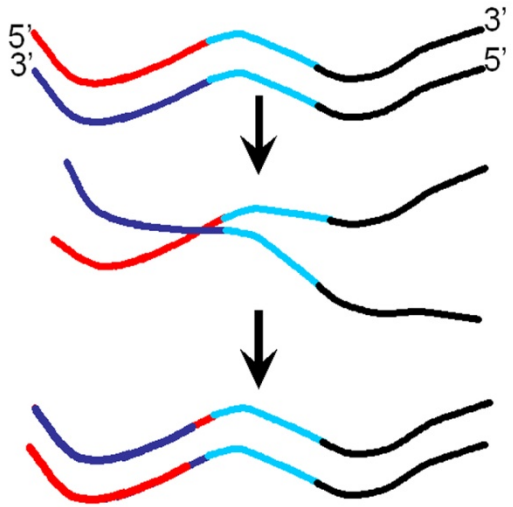


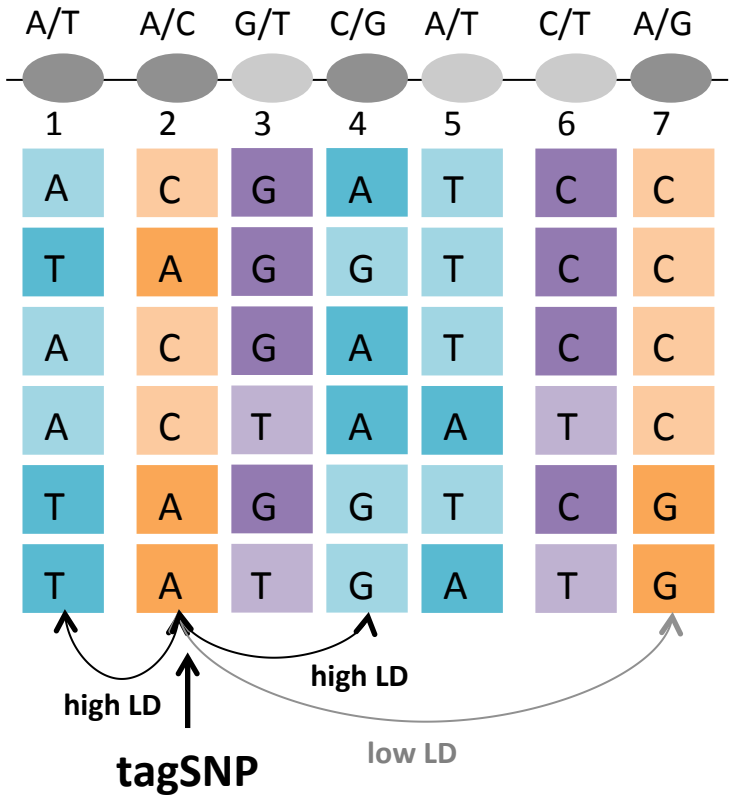
Genome-wide association studies

- 300,000 to 1,000,000 SNPs genome-wide
- Agnostic – no prior knowledge of location or function
- Thousands of samples
- Common variants, common diseases

Linkage disequilibrium

- SNPs adjacent in the genome are not randomly inherited
- Hotspots of recombination create haplotypes





Measures of LD – *D*’

2 loci A and B with alleles A1 and B1

Freq (F) of A1 = p_1, F of B1 = q_1 ,F of A1B1 = x_{11}

Linkage equilibrium $x_{11} = p_1 q_1$ (D=0)

$$D = x_{11} - p_1 q_1$$

$$D' = D/D_{MAX}$$

D_{MAX} is the smaller of $p_1 q_2$ and $p_2 q_1$

Measures of LD – r^2

$$r^2 = \frac{D^2}{p_1 p_2 q_1 q_1}$$

Takes into account the allele frequencies at each locus

plink...

*Last original PLINK release is **v1.07** (10-Oct-2009); **PLINK 1.9** is now available for beta-testing*

Whole genome association analysis toolset

Introduction | Basics | Download | Reference | Formats | Data management | Summary stats | Filters | Stratification | IBS/IBD | Association | Family-based | Permutation | LD calcualtions | Haplotypes | Conditional tests | Proxy association | Imputation | Dosage data | Meta-analysis | Result annotation | Clumping | Gene Report | Epistasis | Rare CNVs | Common CNPs | R-plugins | SNP annotation | Simulation | Profiles | ID helper | Resources | Flow chart | Misc. | FAQ | gPLINK

New (15-May-2014): **PLINK 1.9 is now available for beta-testing!**

PLINK is a free, open-source whole genome association analysis toolset, designed to perform a range of basic, large-scale analyses in a computationally efficient manner.

The focus of PLINK is purely on *analysis* of genotype/phenotype data, so there is no support for steps prior to this (e.g. study design and planning, generating genotype or CNV calls from raw data). Through integration with gPLINK and Haploview, there is some support for the subsequent visualization, annotation and storage of results.

PLINK (one syllable) is being developed by Shaun Purcell at the Center for Human Genetic Research (CHGR), Massachusetts General Hospital (MGH), and the Broad Institute of Harvard & MIT, with the support of others.

New in 1.07: meta-analysis, result annotation and analysis of dosage data.

Quick links

PLINK tutorial

gPLINK

Join e-mail list

Resources

FAQs | PDF

Citing PLINK

Bugs, questions?

Data management

- Read data in a variety of formats

SNPTEST

v2.5

SNPTEST

SNPTEST is a program for the analysis of single SNP association in genome-wide studies. The tests implemented include

- Binary (case-control) phenotypes, single and multiple quantitative phenotypes
- Bayesian and Frequentist tests
- Ability to condition upon an arbitrary set of covariates and/or SNPs.
- Various different methods for the dealing with imputed SNPs.

The program is designed to work seamlessly with the output of our genotype imputation software **IMPUTE** [1] and the programs **QCTOOL** and **GTOOL**. This program was used in the analysis of the 7 genome-wide association studies carried out by the Wellcome Trust Case-Control Consortium (WTCCC) [2]. Much of the theory behind the implemented tests is described in this paper [3].

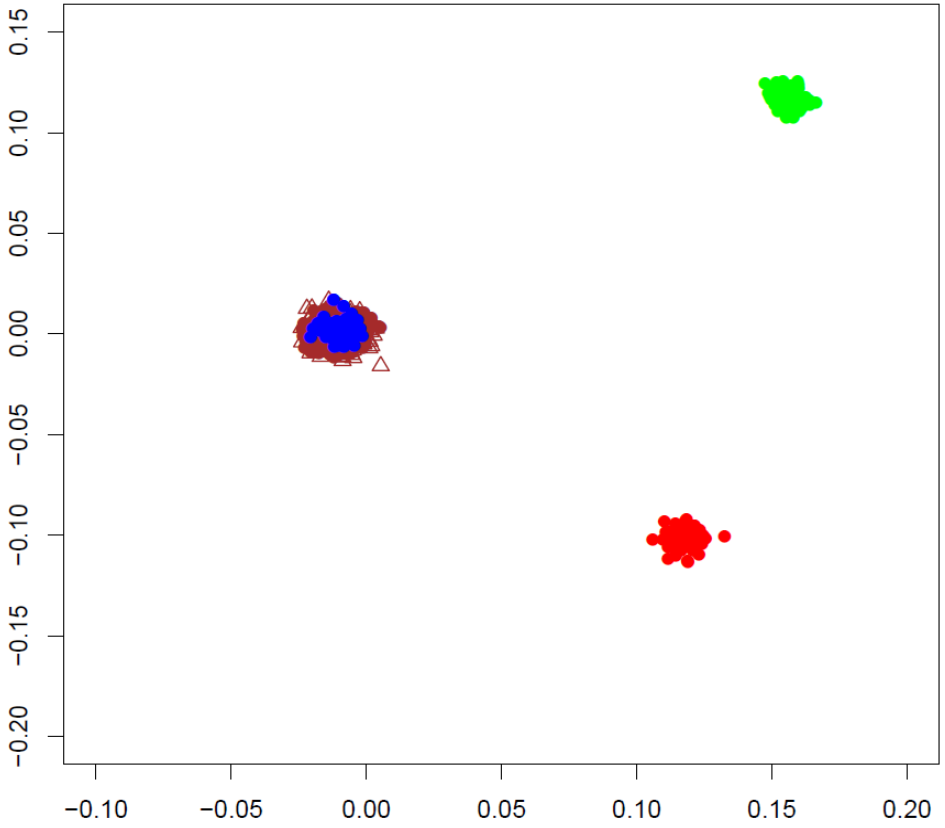
SNPTEST has many different features which are illustrated below through a number of different examples that use the datasets provided with the software in the directory `example/`. These files contain data at 200 SNPs on 1000 individuals that are split into a control cohort and a case cohort. These datasets can be used to try out the tests using both binary (case-control) and quantitative phenotypes.

The latest version of SNPTEST is v2.5. This release has several new features as documented here. To get started, download a pre-built binary for your platform from the download page and run an example command.

Quality control - General

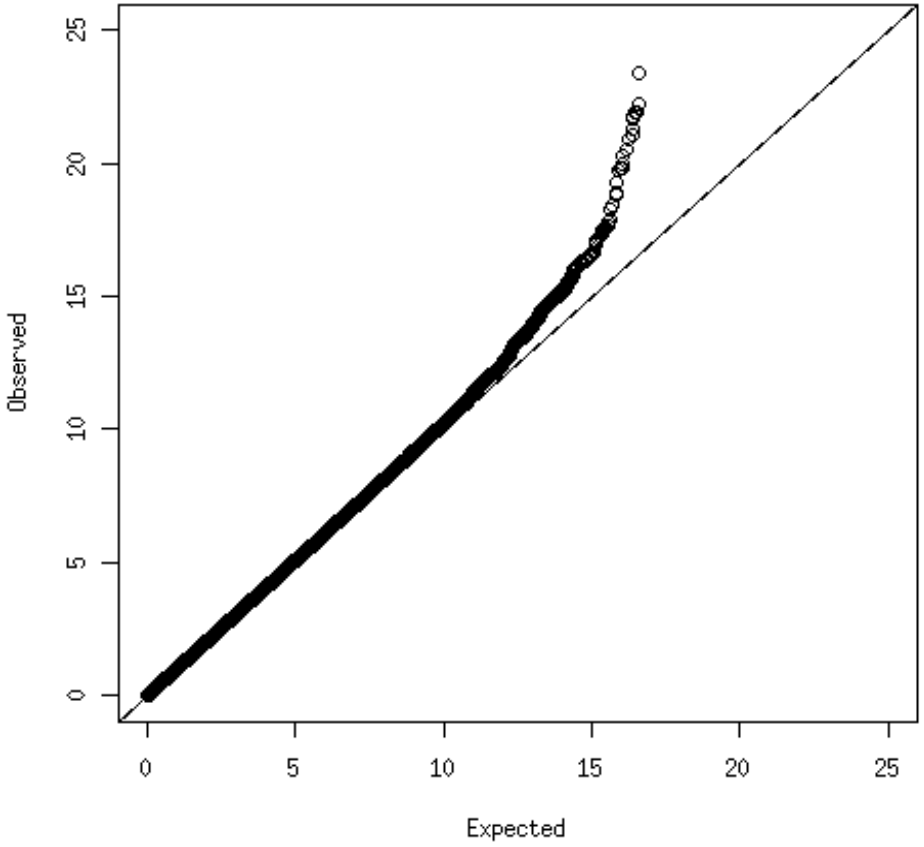
- Discordant sex
- Call rates per SNP and per sample
- SNPs with differing call rates between cases and controls
- Identical or related samples (IBS)

QC – principal components analysis



Ethnicities of cases and controls need to match to avoid population stratification

QC – Quantile-quantile plots



$$\lambda = \frac{\bar{x} \text{ lower 90\% } O(x)}{\bar{x} \text{ lower 90\% } E(x)}$$

Check adequate matching of cases and controls – no inflation of bottom 90% of test statistic

QC – Hardy Weinberg equilibrium

In the absence of evolutionary influences genotype frequencies will remain constant

Alleles at locus denoted A and a

$$f(\mathrm{A})=p \text{ and } f(\mathrm{a})=q$$

$$f(\mathrm{AA})=p^2,\ f(\mathrm{aa})=q^2 \text{ and } f(\mathrm{Aa}) = 2pq$$

Sum of all genotype frequencies must equal 1

$$\therefore\ p^2 + 2pq + q^2 = 1$$

Chi-squared (χ^2) test

Is there a statistical difference between observed and expected?

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

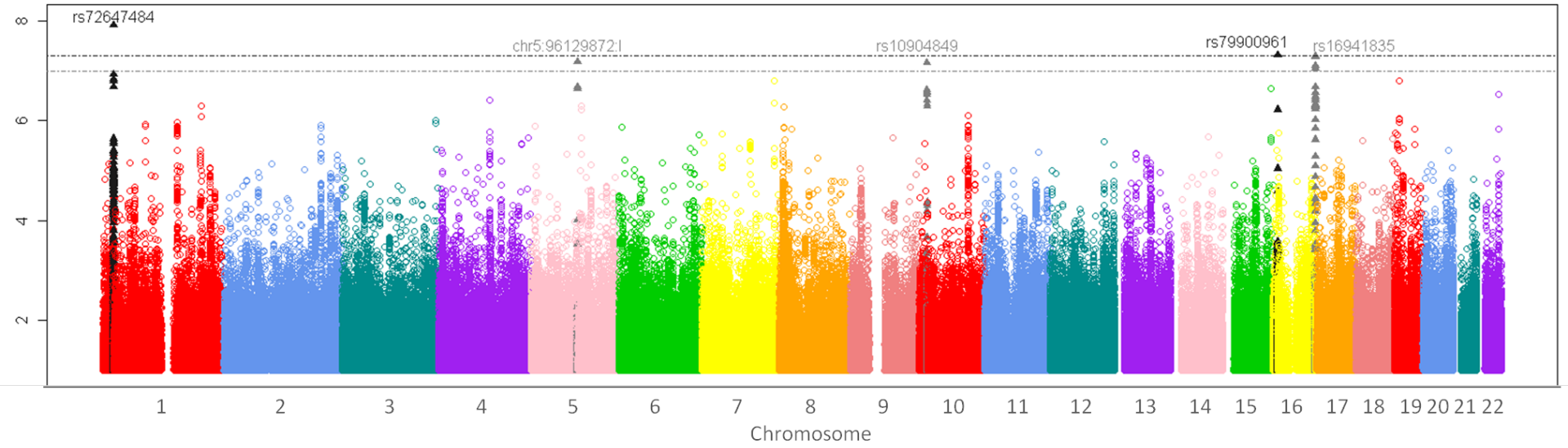
	Data type 1	Data type 2
Category 1	a	b
Category 2	c	d

$$\chi^2 = \frac{(ad - bc)^2(a+b+c+d)}{(a+b)(c+d)(b+d)(a+c)}$$

Testing for an association

- Cochran-Armitage trend test – extension to χ^2 incorporating category ordering
- For small sample sizes use Fisher's exact test
- Logistic regression

Manhattan Plot



P-values

Probability of obtaining a value that is at least as extreme the observed test statistic when H_0 is true

$$P = P(T \geq t \mid H_0)$$

$P < 0.05$ generally taken as significant evidence to reject the null hypothesis

Assessing statistical significance

- *P*<0.05 - the probability of 1 FP is 5%
- In GWAS this would lead to a large number of FPs (if 1,000,000 SNPs are tested then 5% are expected to have *P*<0.05 by chance -> 50,000)
- Bonferroni correction -> ***P* = 0.05/n**

  where n = number of genotyped SNPs
- *P* <5×10^{-8} generally accepted in GWAS

Odds ratios

How strongly is a SNP associated with a trait

OR = odds of an individual with A having B
odds of an individual without A having B

	Diseased	Healthy
Genotype A	D_A	H_A
Genotype a	D_a	H_a

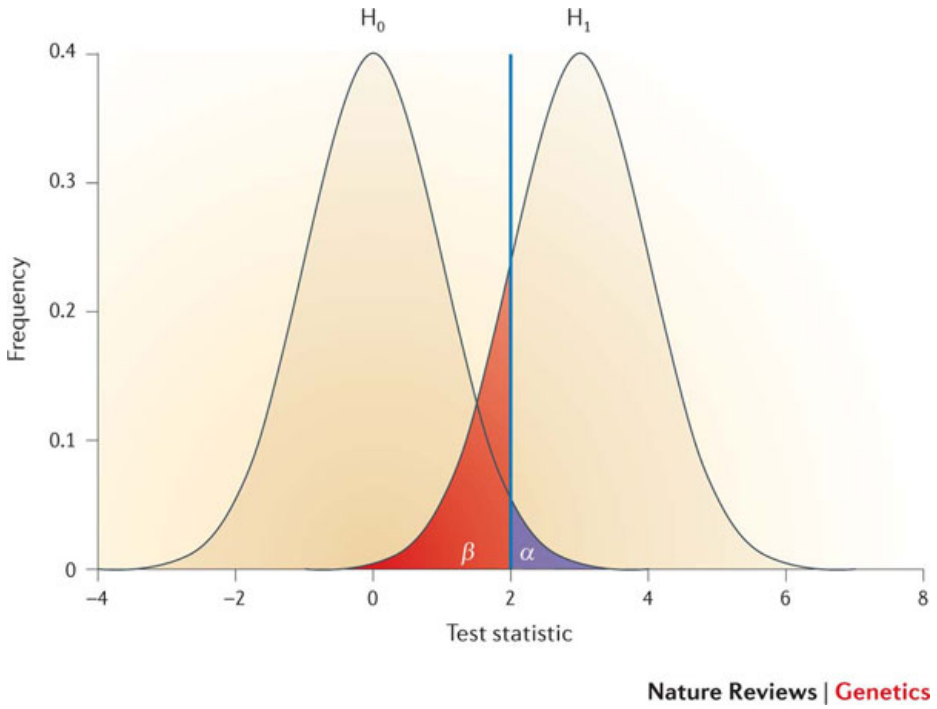
$$OR = \frac{D_A/H_A}{D_a/H_a}$$

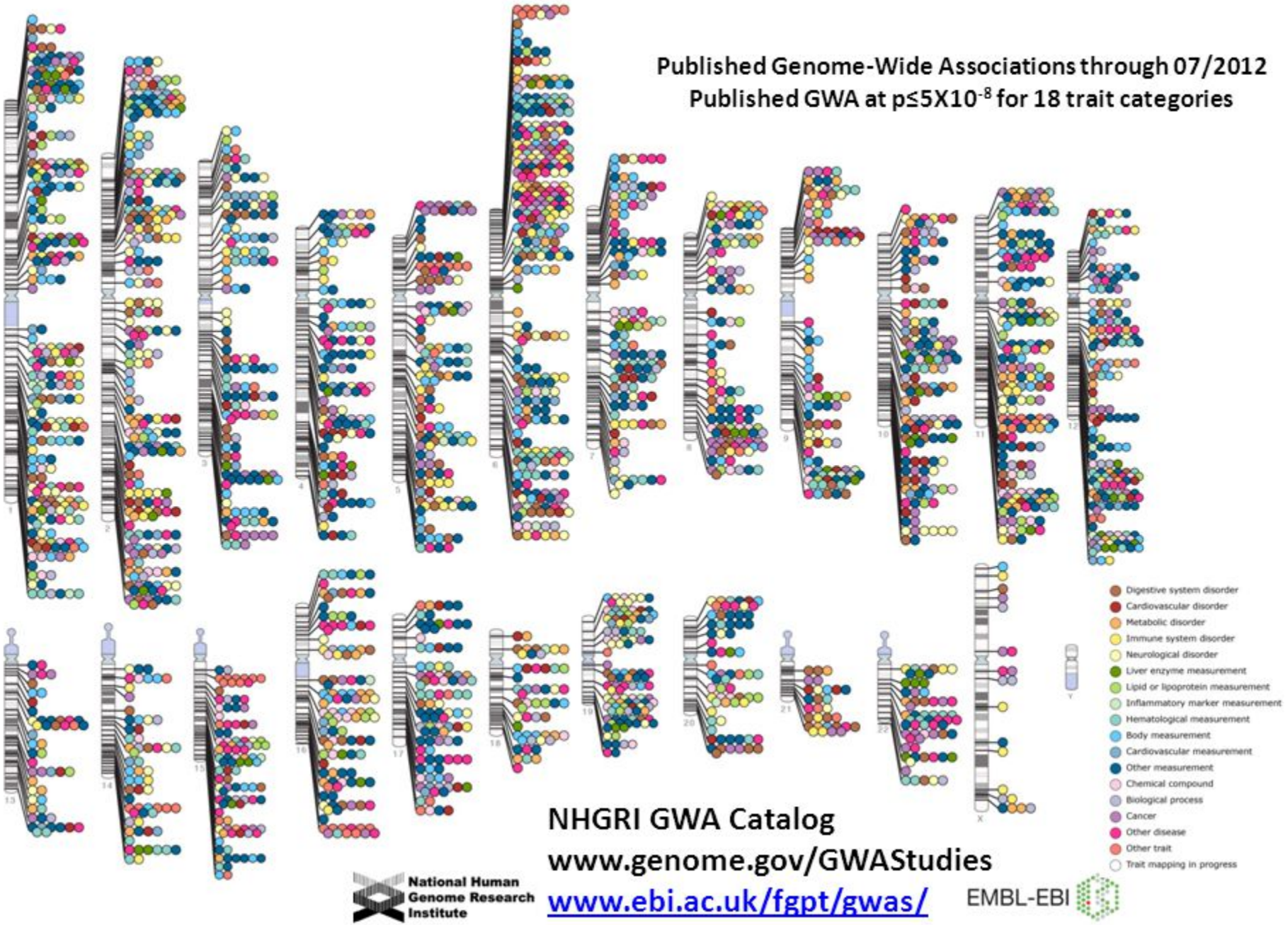
Power

***P* = 1 – β** (the probability of correctly rejecting H_0 when an association exists)

β is subject to factors outside investigator's control: effect size, frequency, and accuracy and completeness of data

Power increases with sample size




Published Genome-Wide Associations through 07/2012
Published GWA at p≤5X10-8 for 18 trait categories
Digestive system disorder
Cardiovascular disorder
Metabolic disorder
Immune system disorder
Neurological disorder
Liver enzyme measurement
Lipid or lipoprotein measurement
Inflammatory marker measurement
Hematological measurement
Body measurement
Cardiovascular measurement
Other measurement
Chemical compound
Biological process
Cancer
Other disease
Other trait
Trait mapping in progress
NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/
National Human Genome Research Institute
EMBL-EBI

Science. 2005 Apr 15;308(5720):419-21. Epub 2005 Mar 10.

Complement factor H variant increases the risk of age-related macular degeneration.

Haines JL[1], Hauser MA, Schmidt S, Scott WK, Olson LM, Gallins P, Spencer KL, Kwan SY, Noureddine M, Gilbert JR, Schnetz-Boutaud N, Agarwal A, Postel EA, Pericak-Vance MA.

Author information

Abstract

Age-related macular degeneration (AMD) is a leading caus
largely unknown. Previous studies identified chromosome 1
polymorphisms to interrogate this region and identified a st
the complement factor H gene within this haplotype reveale
with odds ratios between 2.45 and 5.57. This common varia

nature genetics

Genome-wide association scan identifies a colorectal cancer susceptibility locus on 11q23 and replicates risk loci at 8q24 and 18q21

Albert Tenesa[1,31], Susan M Farrington[1,31], James GD Prendergast[1], Mary E Porteous[2], Marion Walker[1], Naila Haq[1], Rebecca A Barnetson[1], Evropi Theodoratou[1,3], Roseanne Cetnarskyj[2], Nicola Cartwright[1], oussanakis[4], Thibaud Koessler[5], tefan Schreiber[7,9], Henry Völzke[10], eister[12], Hermann Brenner[12], an J Deary[16], John M Starr[17], ona[18], Emily Webb[19], ert[20], Dennis Ballinger[21], Yusuke Nakamura[23], Kustra[24], Alexandre Montpetit[29], Dunlop[1]

Vol 445|22 February 2007|doi:10.1038/nature05616

nature

ARTICLES

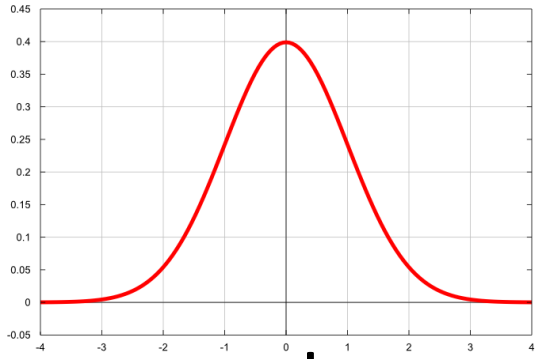
A genome-wide association study identifies novel risk loci for type 2 diabetes

Robert Sladek[1,2,4], Ghislain Rocheleau[1*], Johan Rung[4*], Christian Dina[5*], Lishuang Shen[1], David Serre[1], Philippe Boutin[5], Daniel Vincent[4], Alexandre Belisle[4], Samy Hadjadj[6], Beverley Balkau[7], Barbara Heude[7], Guillaume Charpentier[8], Thomas J. Hudson[4,9], Alexandre Montpetit[4], Alexey V. Pshezhetsky[10], Marc Prentki[10,11], Barry I. Posner[2,12], David J. Balding[13], David Meyre[5], Constantin Polychronakos[1,3] & Philippe Froguel[5,14]

Limitations

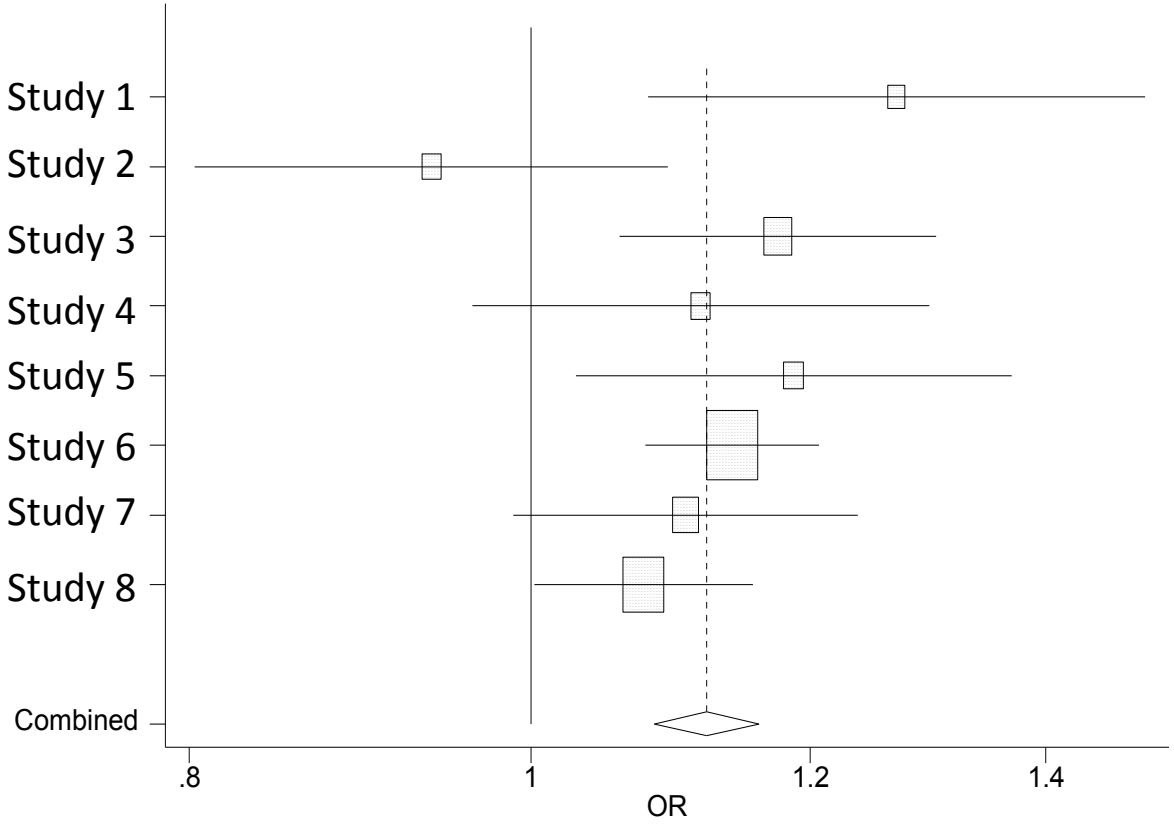
- tagSNPs were not chosen as good candidates for causation – further functional studies needed
- Identified variants have small effect sizes (ORs) – so far interactions between variants have not been identified
- Still a lot of missing heritability
- Currently mainly European studies

Meta-analysis



- Combine multiple GWAS to increase sample size -> greater power
- Fixed effects – true effect size (OR) is shared between all studies
- Random effects – effect size varies between studies (age, population differences) – studies are weighted ($W = 1/se^2$)
- Mantel-Haenszel method to combine data – compared to Normal distribution

Forest plots



Measures of heterogeneity – Q & I^2

$$Q = \Sigma\, W(E - E_c)^2$$

where E = effect size and $E_c = \Sigma\, WE/\Sigma W$

χ^2 with k-1 *d.f.*

Underpowered for <20 studies

$$I^2 = \frac{Q - d.f.}{Q} * 100$$

Values >75% characteristic of heterogeneity

iCOGS

ARTICLES

nature genetics

Large-scale genotyping identifies 41 new loci associated with breast cancer risk

Breast cancer is the most common cancer among women. Common variants at 27 loci have been identified as associated with susceptibility to breast cancer, and these account for ~9% of the familial risk of the disease. We report here a meta-analysis of 9 genome-wide association studies, including 10,052 breast cancer cases and 12,575 controls of European ancestry, from which we selected 29,807 SNPs for further genotyping. These SNPs were genotyped in 45,290 cases and 41,880 controls of

LETTERS

nature genetics

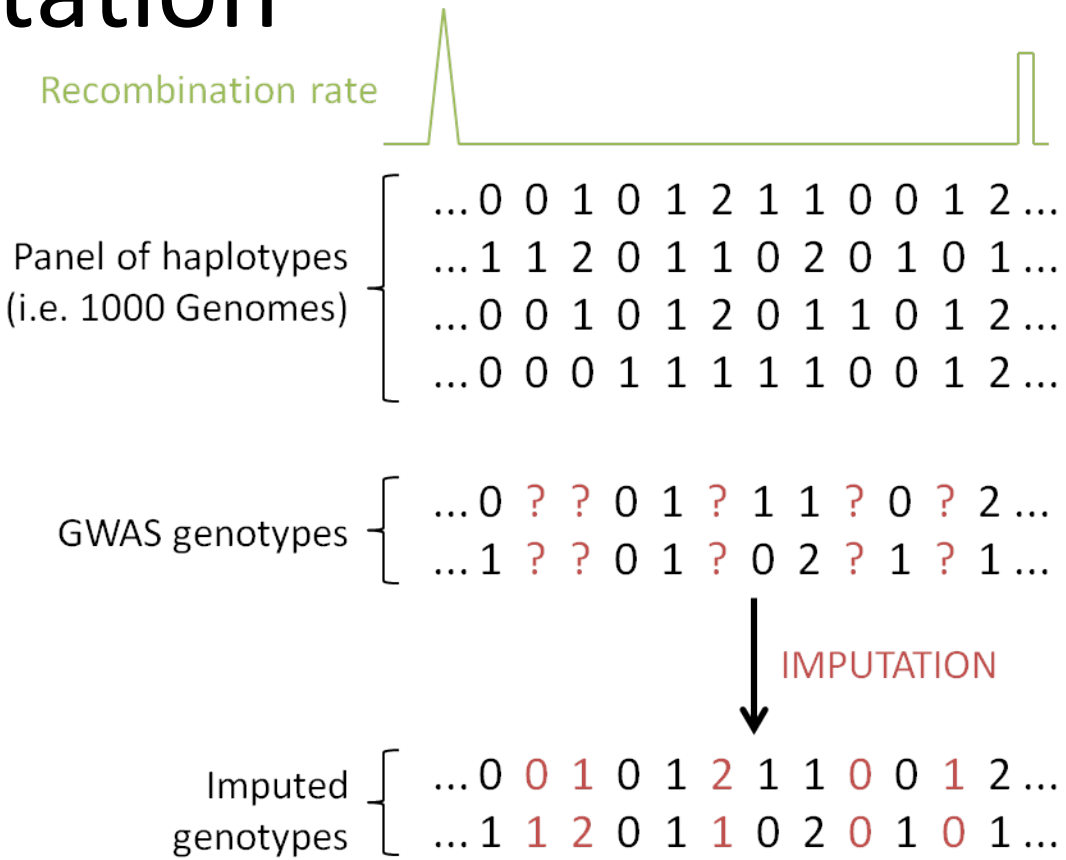
Identification of 23 new prostate cancer susceptibility loci using the iCOGS custom genotyping array

Prostate cancer is the most frequently diagnosed cancer in males in developed countries. To identify common prostate cancer susceptibility alleles, we genotyped 211,155 SNPs on a custom Illumina array (iCOGS) in blood DNA from 25,074 prostate cancer cases and 24,272 controls from the international PRACTICAL Consortium. Twenty-three new prostate cancer susceptibility loci were identified at genome-wide significance ($P < 5 \times 10^{-8}$). More than 70 prostate cancer susceptibility loci, explaining ~30% of the familial risk for this disease, have now been identified. On the basis of combined risks conferred by the new and previously known risk loci, the top 1% of the risk distribution has a 4.7-fold higher risk than the average of the population being profiled. These results will facilitate population risk stratification for clinical studies.

significant association at $P < 0.01$ for overall prostate cancer. These SNPs were genotyped as part of a custom array that included 211,155 SNPs (the iCOGS chip), 85,278 of which were specifically chosen for their potential relevance to prostate cancer (74,001 were from GWAS top hits as described, 13,739 were from fine mapping of known susceptibility regions at the time of the chip design and 1,398 were from candidate gene studies in key pathways (for example, hormone metabolism, HOX genes, the cell cycle and DNA repair; **Fig. 1** and Online Methods); some SNPs were in more than one category). The results of the GWAS component are presented here. The details of the iCOGS array can be found on the COGS website (see URLs).

The iCOGS array was used for the genotyping of 25,074 prostate cancer cases and 24,272 controls from 32 studies participating in the PRACTICAL Consortium (Online Methods). Of these, 39,337

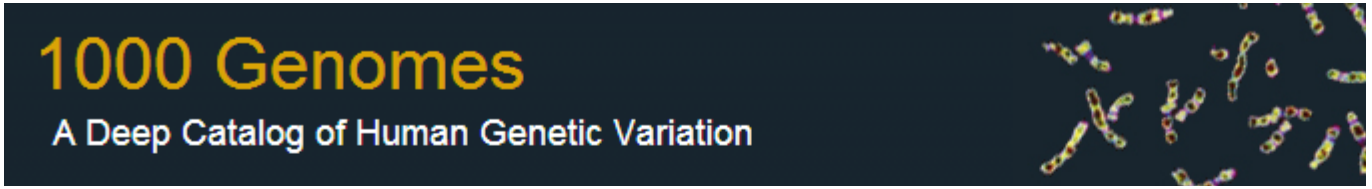
Imputation



Computationally predict genotypes of un-typed SNPs based on haplotypes of a reference dataset – more samples are better, ethnicity is important

Imputation reference data

Dense array genotyping
HapMap 3 - 1,011 individuals from Africa, the Americas and Europe
Some related individuals



Whole genome sequencing at 4-6x coverage
Phase 1 – 1,092 individuals (246 AFR, 181 AMR, 286 ASN, 379 EUR) – ~40 million SNVs and Indels
Phase 3 – 2,577 individuals (691 AFR, 355 AMR, 1,017 ASN, 514 EUR) – ~82 million SNVs and Indels

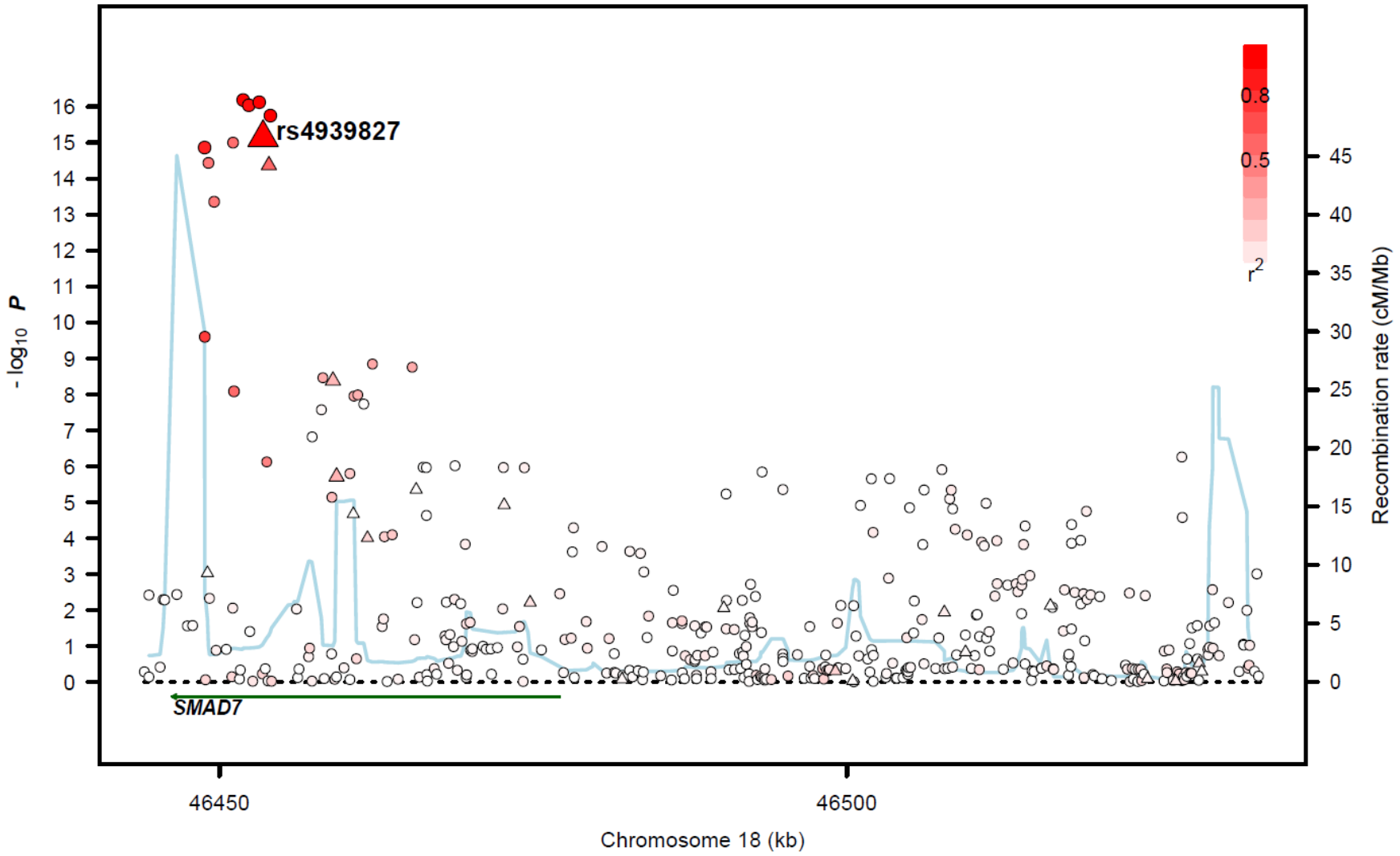
IMPUTE2

- Statistical program for imputing unobserved genotypes in GWAS data
- Takes into account all SNPs in LD with the SNP of interest
- Outputs probabilities of each possible genotype to take into account uncertainty in prediction

SNPTEST and META

- SNPTEST tests for association taking into account uncertainties from IMPUTE
- Outputs Information (INFO) metrics reflecting certainty
- META performs meta-analysis of multiple studies only including studies passing an INFO threshold
- Outputs Q and I^2

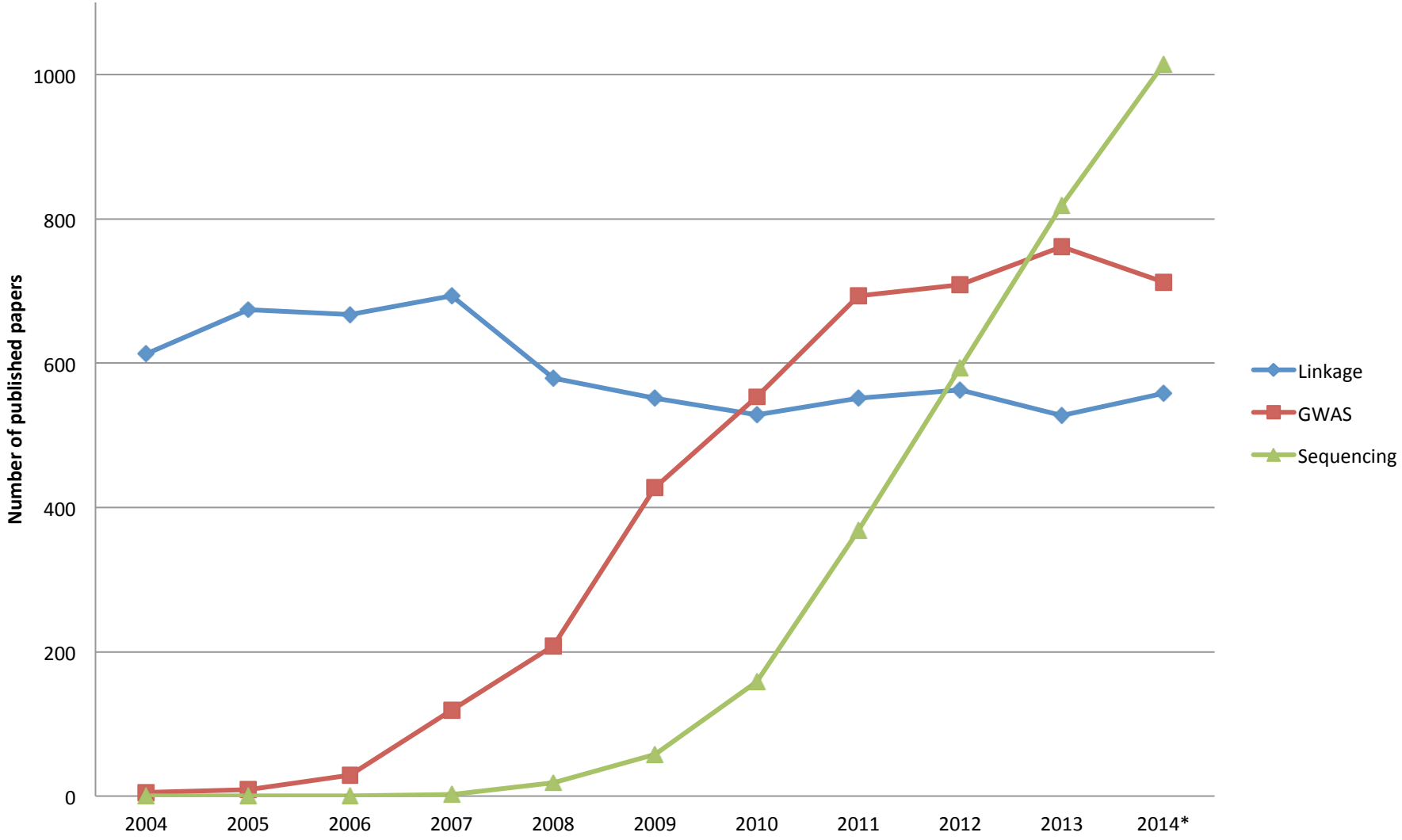
SNAP Plot


rs4939827
SMAD7
- log10 P
Recombination rate (cM/Mb)
r2
0.8
0.5
46450
46500
Chromosome 18 (kb)

Limitations

- Need to start with the same SNPs in cases and controls to avoid spurious associations
- Low frequency SNPs are hard to accurately impute – not adequately represented in reference data and SNPs on GWAS chips are common – rare SNPs are not tagged (MAF cutoff ~1% in controls)

The future of genetic studies



Any Questions?