Advanced Association Analysis

David Balding
Professor of Statistical Genetics
University of Melbourne, and
University College London

2 Feb 2016

1. Population Structure and Cryptic Relatedness

2. Multiple testing and Genome-Wide Significance

3. Replication

4. Meta Analysis

1. Population Structure and Cryptic Relatedness
2. Multiple testing and Genome-Wide Significance
3. Replication
4. Meta Analysis

Population structure refers to a systematic pattern of mating within a population;

- this results in systematic patterns of (perhaps distant) relatedness, or **kinship**
- populations without structure are called “panmictic”.

E.g. mating can be influenced by social or religious groups, or spatial distance or geographic boundaries.

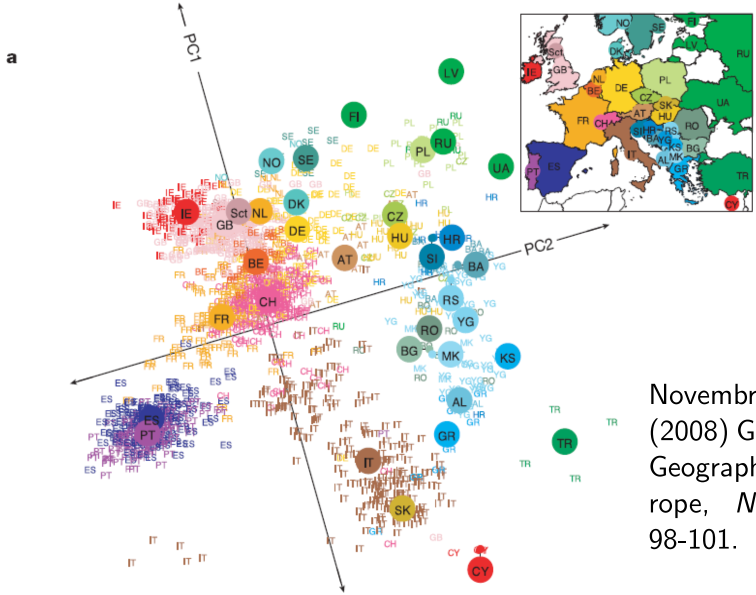
Simple models of population structure include

1. K-subpopulations, or ”island model”
   - preferential mating within islands
2. Continuous cline, or ”isolation by distance”
   - preferential mating with close-by individuals.

Admixture: previously distinct populations begin to inter-breed.

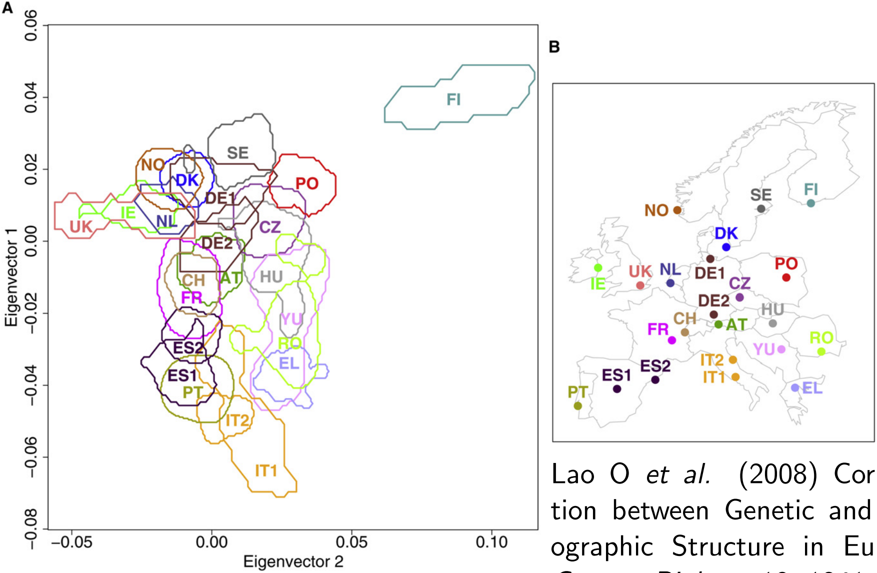
Cryptic relatedness: apparently unrelated individuals actually have some unsuspected relatedness.

Clinal pattern of humans in Europe: Novembre (08)



Novembre P *et al.* (2008) Genes Mirror Geography in Europe, *Nature* 456: 98-101.

Clinal pattern of humans in Europe: Lao (08)

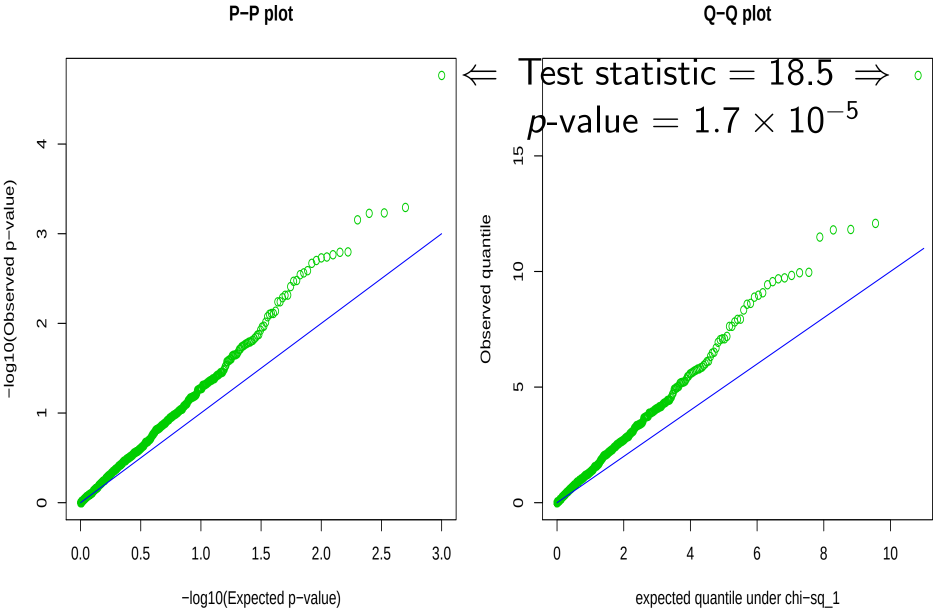


Lao O *et al.* (2008) Correlation between Genetic and Geographic Structure in Europe *Current Biology*, 18, 1241-8.

Population structure and genetic association studies

- Because of patterns of relatedness, genetic allele frequencies can vary across populations (more than sampling variation).
- Phenotypes can also reflect population structure, because
  1. they are controlled by many loci ("polygenic" model) that tend to vary with the population structure, and/or
  2. they vary with climate, diet or other environmental factors that differ across populations, and/or
  3. **ascertainment bias**: recruitment of phenotypic groups differs across populations

$\Rightarrow$ genome-wide tendency for genetic associations reflecting these effects rather than direct causal effect of a SNP.



⇐ Test statistic $= 18.5$ ⇒

p-value $= 1.7 \times 10^{-5}$

Association test results for 1K SNPs under H_0, with population structure. Here and below we assume that the test statistics are χ^2_1 under H_0.

How big is the problem for human studies?

- this question has been the topic of controversy for over 20 years
- the answer depends on several aspects of study design, e.g.
  1. the size of the study
     - larger studies can detect smaller effects
     - so a small effect of pop. structure can be relatively important;
  2. the demographic histories and environments of the populations studied;
  3. the pattern of recruitment.

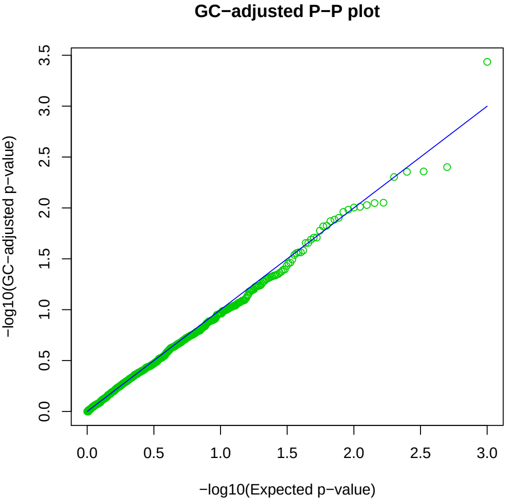
Effect can be important e.g. for differing levels of admixture among Native- or African-Americans.

WTCCC study of UK Caucasians: only a small effect of population structure overall, ~ 20 genes showing strong association with geography.

Cryptic relatedness:

Most studies of apparently unrelated individuals do include some close relatives, and current practice is to remove one of each relative pair although impact on results is usually small. Mixed regression models can allow for effects of kinship so that relatives do not need to be removed.

Genomic Control (GC)



GC: divide all test statistics by λ, the Genomic Inflation Factor (GIF), defined as the empirical median of the test statistics divided by the theoretical χ^2_1 median (=`qchisq(0.5,1)`).

After GC adjustment, empirical median $= H_0$ value. This makes sense only if very few SNPs tag causal variants - not usually true.[a]

GC now mainly used to measure the problem, not to remedy it.

[a] Yang J *at al.*, Genomic inflation factors under polygenic inheritance, *Eur J Hum Genet* (2011).

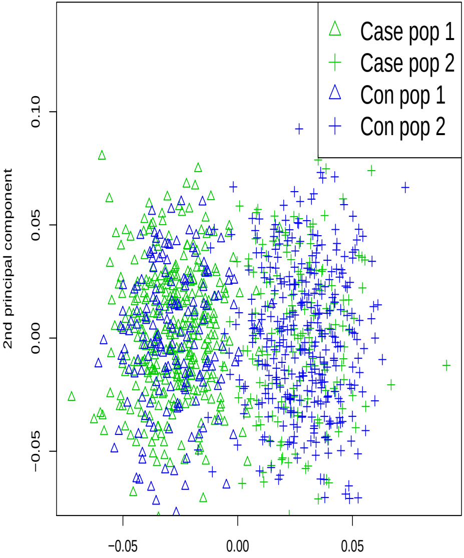
Principal components

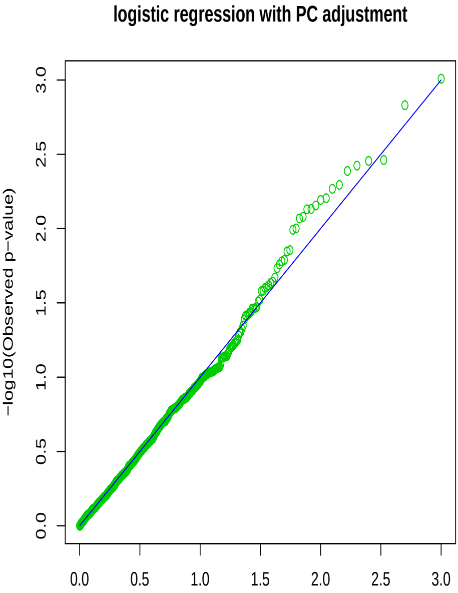
- Because n (# individuals) $\ll$ p (# variables, here SNPs), we eigen-decompose the n $\times$ n matrix $XX^T (= K)$ rather than $X^T X$, where X is n $\times$ p and is column standardised;
- K = average allelic correlation, viewed as a kinship coefficient;
- 1st PC: linear combinations of individuals with maximal variance; the closer the kinship of two individuals, the more similar their 1st PC scores tend to be.
  - If there are 2 subpopulations, 1st PC usually distinguishes them
  - admixed individuals have intermediate scores;
  - similarly, $k-1$ PCs distinguish k subpopulations (including admixture).
- But PCs also strongly influenced by patterns of LD: MHC, inversions.

PCs can be used as regression covariates.[1] Typically 2 – 15 PCs are used, no easy way to decide best number.

[1] Price A *et al.*, Principal components analysis corrects for stratification in genome-wide association studies. Nat Genet 38, 904-9, 2006.

Principal components





Mixed Regression Models and Kinship

Key idea is to eliminate from phenotypes any correlation that can be attributed to kinship, reflecting genome-wide polygenic effects. PC adjustment uses just the first few eigenvectors of the kinship matrix $\hat{K}$, linear mixed models (LMM) use the whole matrix:

$$\begin{array}{ccccccccc} Y = & \theta Z & + & \beta X_j & + & \gamma & + & \epsilon \\ & \text{covariate effect} & + & \text{SNP effect} & + & \text{random effect} & + & \text{residual} \end{array}$$

where $\gamma \sim N(0, \sigma_g^2 K)$ and $\epsilon \sim N(0, \sigma_e^2 I)$.

- Random effect γ corresponds to genome-wide additive polygenic effects, with correlation structure K, assumed known (more later).
- σ_g^2 measures the relative importance of polygenic effects, and is related to narrow-sense heritability via $h^2 = \sigma_g^2/(\sigma_g^2+\sigma_e^2)$.
- LMM can be sensitive to ascertainment, which can invalidate the assumption that phenotype correlation = genotype correlation.
- Possible solution reverse the regression: X_j is the response variable; Haseman-Elston regression: regress $K_{ii'}$ against $(Y_i - Y_{i'})^2 \quad \forall i, i'$.

Methods overview: population structure/cryptic kinship[2]

- **Genomic Control**: simple; fast; handles cryptic kinship + population structure; some loss of power, can be severe under ascertainment bias or polygenic inheritance; can work with $\sim 10^2$ SNPs.
- **PC adjustment**
  - uses only first few PCs of kinship matrix which (usually) measures large-scale population structure;
  - cannot handle cryptic kinship or complex forms of population structure.
  - problem of choosing # PCs to use.
- **Mixed regression models** use whole kinship matrix
  - adjust for cryptic kinship as well as population structure;
  - computational issues now essentially resolved;
  - doesn't allow for confounding role of selection;
  - can be affected by ascertainment for binary data.

[2] Astle W, Balding D, Population structure and cryptic relatedness in genetic association studies, *Stat Sci* 24(4), 451-471, 2009; Price A *et al.*, New approaches to population stratification in genome-wide association studies. Nat Rev Genet 11(7), 459-63, 2010; Loh P *et al.* Efficient Bayesian mixed-model analysis increases association power in large cohorts. *Nat Genet* 47.3 (2015): 284-290.

1. Population Structure and Cryptic Relatedness
2. Multiple testing and Genome-Wide Significance
3. Replication
4. Meta Analysis

- Up until about 2007 there was frequently a failure to replicate reports of genetic association - the problem is now much reduced but has not entirely gone away.
- One reason was inadequate criteria for deciding when an association should be regarded as established.
- As the number of tests increased with improved marker technology, the possibilities for false positives also increased: called the **problem of multiple testing**.
- Traditionally a significance level of $\alpha = 0.05$ has been used in science, which allows on average one false positive per twenty tests under the null hypothesis.
- This is unacceptable for testing a million SNPs - it could generate 50 000 false positives.

Family-Wise Error Rate (FWER)

One solution is to control the FWER – the probability of making at least one type-1 error (false positive) in a "family" of N tests.

- That is, we control $\mathrm{P}(H_0 \text{ rejected for } \geq 1\ j | H_0 \text{ holds } \forall j)$.
- If each of N *independent* SNPs is tested at significance level α_{SNP}, then the probability under H_0 of ≥ 1 significant result is
$$\alpha_{FWER} = 1 - (1{-}\alpha_{SNP})^N.$$
- So to achieve a desired α_{FWER} we set
$$\alpha_{SNP} = 1 - (1 - \alpha_{FWER})^{1/N}.$$
- **Bonferroni approximation:** $\alpha_{SNP} = \alpha_{FWER}/N$.
- If $N = 10^6$ and we choose $\alpha_{FWER} = 5\%$, then we need
$$\alpha_{SNP} = 0.05/10^6 = 5 \times 10^{-8}.$$

Problems with use of FWER

- SNPs are *dependent* (LD), which makes the Bonferroni correction conservative; difficult to estimate α_{SNP} accurately.
- Calculated correctly, α_{SNP} varies with many factors (see below).
- Genome-wide H_0 is implausible: we always expect some positives, otherwise we wouldn't have performed the GWAS.
- We should consider all SNPs, even if not typed in current GWAS (there are likely to be other GWAS by other researchers).
- In reaction to early problems of non-replication, it is arguable that the genetic epidemiology community has over-reacted, imposing too-strict control of FWER leading to too many false negatives.

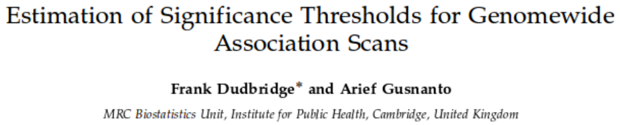
Permutation tests

- Allow approximation of α_{FWER} corresponding to different values of α_{SNP} overcoming the problem of LD.
- Randomly shuffle the phenotype values among subjects, with genotype data fixed.
- Analyse the randomised data and observe whether or not any SNP is significant.
- Repeat this procedure M times, and estimate

$$\hat{\alpha}_{FWER} = \frac{R+1}{M+1},$$

where $R = \#$ permuted datasets for which ≥ 1 SNP was significant.
- The 1 in numerator and denominator make the procedure tend to be conservative (estimate is biased slightly upward, $\mathbb{E}[\hat{\alpha}_{FWER}] > \alpha_{FWER}$) and ensures that we never obtain an estimate of 0. However the value 1 is arbitrary and does not guarantee that $\hat{\alpha}_{FWER} > \alpha_{FWER}$.

Genome-wide significance level



Estimation of Significance Thresholds for Genomewide Association Scans

Frank Dudbridge* and Arief Gusnanto

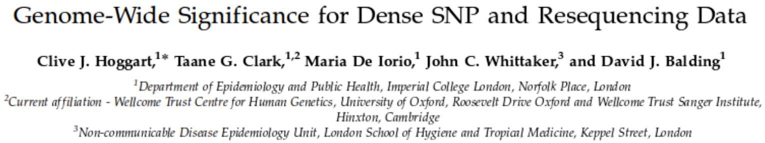
MRC Biostatistics Unit, Institute for Public Health, Cambridge, United Kingdom

Dudbridge & Gusnanto (2008)[3] took real GWAS genotype data and extrapolated the p-value threshold to an infinite density of SNPs.

- They found that $\hat{\alpha}_{FWER} = 2 \times 10^{-7}$ for the observed data, but this decreased to $\hat{\alpha}_{FWER} = 7 \times 10^{-8}$ if SNPs with the same statistical properties were infinitely dense in the genome.

[3] Estimation of significance thresholds for genomewide association scans. *Genetic Epidemiology*, 32(3), pp 227-234.

Genome-wide significance level



Genome-Wide Significance for Dense SNP and Resequencing Data

Clive J. Hoggart,[1*] Taane G. Clark,[1,2] Maria De Iorio,[1] John C. Whittaker,[3] and David J. Balding[1]

[1]Department of Epidemiology and Public Health, Imperial College London, Norfolk Place, London
[2]Current affiliation - Wellcome Trust Centre for Human Genetics, University of Oxford, Roosevelt Drive Oxford and Wellcome Trust Sanger Institute, Hinxton, Cambridge
[3]Non-communicable Disease Epidemiology Unit, London School of Hygiene and Tropical Medicine, Keppel Street, London

Hoggart *et al.* (2008)[4] used computer simulations of entire genomes under realistic population genetics models for three large populations. They then estimated $\hat{\alpha}_{FWER}$ if different classes of SNPs were tested.

- They found strong dependence of $\hat{\alpha}_{FWER}$ on many factors, including:
  - Population
  - MAF threshold
  - Choice of statistical test
  - Numbers of cases and controls.

[4] Genome-wide significance for dense SNP and resequencing data. *Genetic Epidemiology*, 32(2), pp 179-185.

Some results from Hoggart *et al.* (2008)

α_{FWER} **using 5K cases and 5K controls.**

Population:	European	East Asian	West African
MAF>0.05	3.1, 10	2.7, 12	1.5, 6.0
MAF>0.005	1.3, 5.2	1.3, 5.2	0.66, 2.6
All SNPs	0.69, 3.5	0.86, 3.5	0.65, 2.6

The values in each cell $\times 10^{-8}$ are for FWER $= 5\%$ and FWER $= 20\%$.

- A consensus has developed that α_{SNP} must be 5×10^{-8}.
- From results above, this corresponds to FWER $\approx 20\%$ in Europeans and East Asians.
- West Africans: need $\alpha_{FWER} \approx 2.5 \times 10^{-8}$ to achieve FWER $= 20\%$.

Power Calculations

Power is the probability that a SNP that is associated with Y in the population will be detected in the sample:

$$\text{Power} = \mathbb{P}[H_0 \text{ rejected} \mid H_0 \text{ false}].$$

Many tests have a χ^2_1 null distribution, in which case genome-wide significance corresponds to a test statistic $T >$ `qchisq(1-5e-8,1)`$=29.7$

Suppose we have sample size n and the true (population) proportion of variance explained by a variant is h^2. Its test statistic from single-SNP association testing will be distributed χ^2 with $\text{df} = 1$ and non-centrality parameter (ncp) $= nh^2/(1-h^2)$, sometimes written $\chi^2_1(nh^2/(1-h^2))$.

Power Calculations

Given n and h^2, we can compute the power in R as

```
pchisq(29.7,1,ncp=nh^2/(1-h^2),lower=F)
```

E.g., with $h^2 = 0.01, n = 1000$ and $n = 5000$,

pchisq(29.7,1,ncp=1000*.01/0.99,lower=F) = 0.012

pchisq(29.7,1,ncp=5000*0.01/0.99,lower=F) = 0.951

The h^2 used in these calculations is that of the SNP, which equals $r^2 h_c^2$ where h_c^2 is the heritability of the underlying causal variant and r^2 measures the LD between them. It follows that many true signals will not be detected if the sample size is not large, and/or the heritability of the causal variant is not high and/or it is poorly tagged by the SNPs tested.[5]

[5] See, for case/control traits: How informative is a negative finding in a small pharmacogenetic study?, *Pharmacogenomics (2012)*

False Discovery Rate (FDR)

An alternative to monitoring FWER is to control FDR: the fraction of false +ves among all +ves.

	Significant?	
True effect?	No	Yes
No	U	V
Yes	Q	R

- Type-1 error = $\mathbb{E}[\mathrm{V}/(\mathrm{U}+\mathrm{V})]$
- FDR = $\mathbb{E}[\mathrm{V}/(\mathrm{V}+\mathrm{R})]$ if $V+R > 0$, otherwise 0.

- We know which SNPs are significant; we don't know which of these are not true effects, but their number can be controlled or estimated.
- Controlling FDR is preferred over controlling FWER because it is more interpretable and more relevant to the investigator's problem of deciding which SNPs to follow up. FWER assumes an unrealistic H_0.

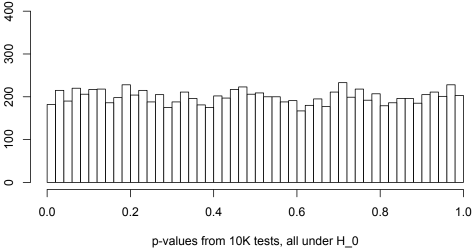
FDR: Benjamini-Hochberg (1995) procedure[6]

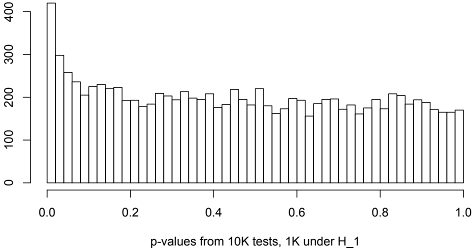
- Order the p-values for all N SNPs: $p_{(1)} \leq p_{(2)} \leq \ldots \leq p_{(N)}$.
- Find the largest j such that $p_{(j)} \leq j\alpha/N$
  - All p-values $\leq p_{(j)}$ are significant.
  - The FDR is guaranteed not to exceed α.
- Same as Bonferroni adjustment for the smallest p-value, but the threshold gets more liberal with subsequent p-values (by a factor of k for the kth smallest p-value).
- B-H procedure thus allows more significant SNPs than controlling the FWER at the same α level.
  - but this is not a good reason to prefer the FDR.

[6] Benjamini Y, Hochberg Y, 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Roy Statist Soc B*, 57(1), pp 289-300.

Estimating the fraction of null SNPs: Storey 02 method

Two simulations of association test statistic *p*-values for 10^4 SNPs





Top plot: no excess of low-*p*-values;

- This is expected since all were simulated under H_0, and so *p*-values should be uniformly distributed in (0,1).

Bottom plot: For a given λ such that $0 < \lambda \ll 1$ we can compute the fraction of *p*-values $> \lambda$. Dividing by the expected fraction $(= 1-\lambda)$ gives an estimate of the fraction of SNPs that follow H_0.

- The estimate should be stable for different values of λ, say between 0.1 and 0.25.

Computing the FDR (Storey 02 method)[7]

- First choose a significance threshold α. Here, let's say $\alpha = 10^{-3}$
- For the bottom histogram on previous slide, $\lambda = 0.15, 0.2$, and 0.25 all give estimate of 94% of SNPs following the null.
  - The true value used for the simulation was 90%, but some have low effect sizes so cannot be distinguished from null SNPs.
- Therefore the number of null SNPs with *p*-value $< \alpha$ is estimated to be $0.94 \times 10\,000 \times \alpha = 9.4$.
- We observe 47 SNPs with *p*-value $< \alpha$.
- $\Rightarrow$ among these 47 SNPs we expect 9.4 to be false positives,
  - giving an FDR estimate of $9.4/47 = 0.2$.
- If there are few, weak true +ves in a GWAS, precision of the estimate may be poor. LD also reduces precision.

[7] Storey J, 2002. A direct approach to false discovery rates. *J Roy Statist Soc B*, 64(3), pp 479-98.

1. Population Structure and Cryptic Relatedness

2. Multiple testing and Genome-Wide Significance

3. Replication

4. Meta Analysis

- Replication is usually regarded as a mandatory step to establish that an association is real.
- One goal of replication is to eliminate the possibility that the association found in the primary study was due to some undiscovered bias or error.
- However there is much confusion and some bad practice around what constitutes adequate replication.
- It is more powerful to put all resources into a single study, rather than split into two sub-studies to claim replication
  - such artificial replication should be avoided.

Replication

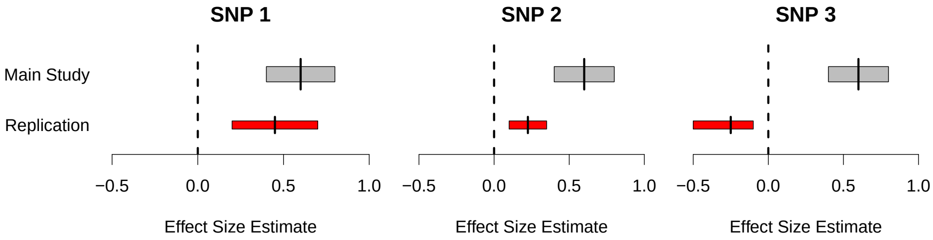
- Some researchers seek both the increased power of a single study and the protection from bias of a replication by performing a meta-analysis to combine two sub-studies
  - this is unacceptable, you have to choose one or the other goal.
- GWAS analyses have reached a mature stage, so the opportunities for bias or error are now limited and well understood; the need for replication should be reduced e.g. for rare phenotypes.

Technical replication

- Seek to obtain the same result using the same study samples
- Use a different genotyping technology, applied to the top hits or to non-hits with *a priori* support.
- Can eliminate false positives due to genotyping errors, and confirm sample identity.

Direct replication

- Requires independent study samples from the same population as the original ("discovery") study
- Must involve the same genetic variant, or a very good proxy ($r^2 \approx 1$)
- Association should be in the same direction and be consistent with the same genetic model (dominant, recessive, codominant). Odds ratios should be similar (may be reduced by "winner's curse", see below).



Indirect replication

Indirect replication tries to combine two different goals:

- to check that the original report of association was correct, and
- to investigate its generalisability, or the mechanism of association, by varying one or more factors from the discovery study.

The factors that are varied could include

- different alleles in the same gene/region
- different population
- different (but closely related) phenotype
  - e.g. obesity rather than type 2 diabetes
  - an intermediate quantitative trait, such as bone mineral density rather than a binary osteoporosis classification.

Replication examples

Direct

- Wellcome Trust Case-Control Consortium (2007) found association of the G allele of rs17696736 with Type-1 diabetes (OR $= 1.37$, $p = 7 \times 10^{-14}$) in UK subjects
- Todd *et al.* (Nat Genet 2007) replicated this association in a larger sample of UK subjects (OR $= 1.16$, $p = 2 \times 10^{-6}$)

Indirect

- Stacey *et al.* (Nat Genet 2008) found association of the G allele of rs10941679 with breast cancer (OR $= 1.19$, $p = 3 \times 10^{-11}$)
- Turnbull *et al.* (Nat Genet 2010) found association with other SNPs in the same region (chr 5p12): rs7716600 (OR $= 1.11$, $p = 0.0034$) has $r^2 = 0.75$ with rs10941679.

Replication significance

Question: what significance level is required to achieve replication?

- No consensus on this. It is agreed that a less stringent threshold is required for replication than for the primary association, but that it should be more stringent than 0.05. Sometimes $\alpha_{rep} = 10^{-3}$ or $\alpha_{rep} = 10^{-4}$ are suggested as reasonable thresholds.

Question: how big should my replication study be?

- Bigger is always better.
- Replication study should in general be no smaller than the primary study.
- When calculating power for a planned replication study, remember to use a smaller effect size than was observed – because of winner's curse bias.

Winner's curse (Beavis effect)

The "winner's curse" bias can arise in sports performance, accident rates at "black spots", genetics and many other fields.

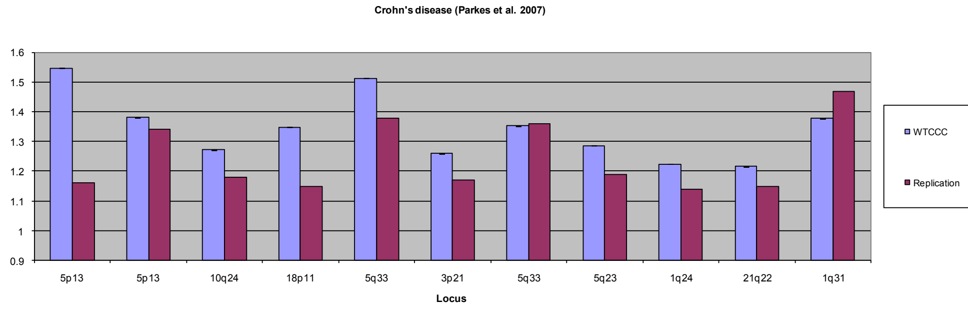
If there are many items (sports players, intersections, genetic loci) and the observation at each is due to a combination of

1. true effect: randomly-distributed over items; fixed over time,
2. "noise": randomly-distributed over items; varying over time,

then in any period of study the items with largest scores will have among the biggest values of both *true effects* AND *noise*.

In any follow-up study the *true effect* will be the same but the *noise* is likely to be much smaller (because it just happened to be unusually large in the primary study).

- Noted by Galton over 100 years ago; came to be known as "regression to the mean": the heights of children of tall parents are above-average, but $<$ parental mid-height.



- Odds-ratio estimates for 11 SNPs identified by the WTCCC (2007) as associated with Crohn's disease, estimated from the original study and a follow-up replication study.
- In most cases the estimate in the replicate study is lower: this is the "winner's curse":
    - There are many more than 11 SNPs with true but weak association. The 11 that were discovered were due to a combination of true effect + "luck". In the replicate study the true effect may be the same but we are usuaully not so "lucky".

Unbiased OR estimates and publication bias

Replication studies give unbiased estimates

- but inefficient as they ignore the discovery data.
- Various methods available to adjust GWAS data for selection by a p-value threshold
- Bowden & Dudbridge (Genet Epidemiol 2009) combine GWAS and replication data into an unbiased estimate

“Reporting” or “Publication” bias

- significant results are more likely to be published than non-significant;
- this can bias meta-analyses and systematic reviews;
- every well-conducted study should be published, irrespective of its outcome.

1. Population Structure and Cryptic Relatedness

2. Multiple testing and Genome-Wide Significance

3. Replication

4. Meta Analysis

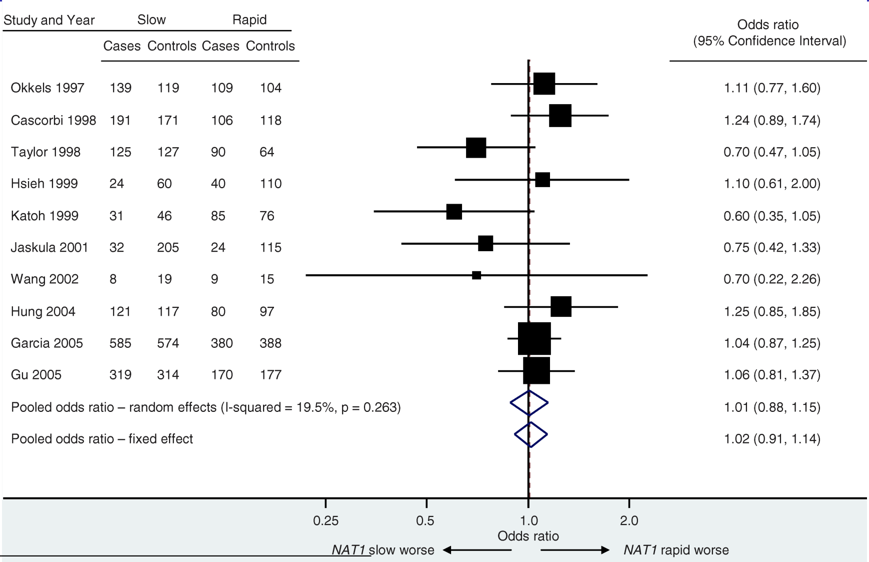
Meta Analysis

- Meta Analysis (MA) involves combining the results of several studies to obtain an overall conclusion. The MA usually has more power than any individual study.
- MA does not replace the need for replication – it retains any biases present in the individual studies.

Results from an MA and all its component studies can be represented in a forest plot (next slide).

- The lines indicate the confidence interval (CI, usually 95%) for the parameter of interest (usually additive model OR for a binary trait).
- The square represents the point estimate and its area reflects the study size.
- The diamond shows point estimate and CI for the meta-analysis.

Forest Plot[8]



Study and Year	Slow		Rapid		Odds ratio (95% Confidence Interval)
	Cases	Controls	Cases	Controls	
Okkels 1997	139	119	109	104	1.11 (0.77, 1.60)
Cascorbi 1998	191	171	106	118	1.24 (0.89, 1.74)
Taylor 1998	125	127	90	64	0.70 (0.47, 1.05)
Hsieh 1999	24	60	40	110	1.10 (0.61, 2.00)
Katoh 1999	31	46	85	76	0.60 (0.35, 1.05)
Jaskula 2001	32	205	24	115	0.75 (0.42, 1.33)
Wang 2002	8	19	9	15	0.70 (0.22, 2.26)
Hung 2004	121	117	80	97	1.25 (0.85, 1.85)
Garcia 2005	585	574	380	388	1.04 (0.87, 1.25)
Gu 2005	319	314	170	177	1.06 (0.81, 1.37)
Pooled odds ratio – random effects (I-squared = 19.5%, p = 0.263)					1.01 (0.88, 1.15)
Pooled odds ratio – fixed effect					1.02 (0.91, 1.14)

[8] Sagoo G, Little J, Higgins J (2009) Systematic Reviews of Genetic Association Studies. PLoS Med 6(3): e1000028

Meta analysis: fixed effects model

- assumes that the effect size is the same in all studies contributing to the meta-analysis
  - unrealistic assumption, as LD and environmental effects may differ between populations
  - contribution of heterogeneity across studies to estimation variance is measured by I^2.
- Then the effect size estimate Y_i from the ith study, $i = 1, \ldots, k$, has $\mathbb{E}[Y_i] = \mu$ and $\text{Var}[Y_i] = \sigma_i^2$.
- The inverse-variance weighted average is given by

$$\bar{Y} = \sum_{i=1}^{k} w_i Y_i \qquad \text{where} \qquad w_i = \frac{1/\sigma_i^2}{\sum_i 1/\sigma_i^2}.$$

  Then $\mathbb{E}[\bar{Y}] = \mu$ and $1/\text{Var}[\bar{Y}] = \sum_i 1/\sigma_i^2$ which is the minimum variance among weighted averages with weights that sum to 1. Note that if $\sigma_i^2 = \sigma^2 \; \forall i$ then $w_i = 1/k$ and $\text{Var}[\bar{Y}] = \sigma^2/k$.

Meta analysis: random effects model

- Assumes that the true odds ratio in each study is drawn independently from an $N(\mu, \sigma^2)$ distribution.
  - So estimates are shrunk towards a global average.
- It is a more realistic model, but the wrong hypothesis is usually tested:
  - $H_0 : \mu = 0$ is tested treating σ^2 as a nuisance parameter
  - should test $H_0 : \mu = \sigma = 0$.
- This error can lead to a dramatic loss of power, because if the effect size varies over populations this tends to *weaken* the significance of the MA when it should *strengthen* it.
- Despite this it has often been recommended by statisticians as more conservative than fixed effects MA: fortunately, this advice has generally been ignored.

Meta analysis: further issues

- Best practice is for a consortium to plan a meta-analysis in advance
    - This allows harmonisation of QC procedures and analysis methods, and sharing of expertise.
- Ideally individual-level data would be pooled for joint analysis, which allows more sophisticated analyses
    - adjustment for individual covariates
    - fine control of population stratification
    - pooling of information about covariates across studies.
- However for reasons of convenience and to avoid issues around permissions for data use, MA usually proceeds only by combining over studies their test statistics, such as OR estimates, together with a measure of precision such as sample size or CI.