Power of
Genomewide Association Studies

Biostatistics 666

A Simple Disease Model

- Risk allele frequency p
- Background allele frequency f
- Increase in disease risk per allele r

- Examples:
  - *HLA-C* risk allele for psoriasis, $p=.15$, $f=.0065$, $r=2.6$
  - *TNIP1* risk allele for psoriasis, $p=.05$, $f=.0095$, $r=1.8$
  - *TCF7L2* risk allele for type 2 diabetes, $p=.35$, $f=.08$, $r=1.4$
  - *R1210C* risk allele for macular degeneration, $\mathrm{p}=10^{-4}$, $\mathrm{f}=.05$, $\mathrm{r}=25$

  - f selected so overall risk of disease is about 1%

What Happens in Cases …

$$P(case\ \&\ low\ risk) = (1-p)^2 f$$
$$P(case\ \&\ med\ risk) = 2p(1-p)fr$$
$$P(case\ \&\ high\ risk) = p^2 f r^2$$

$$P(case) = \left((1-p)^2 + 2p(1-p)r + p^2 r^2\right) f$$

$$P(low\ risk|case) = (1-p)^2 f / P(case)$$
$$P(med\ risk|case) = 2p(1-p)fr / P(case)$$
$$P(high\ risk|case) = p^2 f r^2 / P(case)$$

$$P(risk\ allele|case) = (p(1-p)r + p^2 r^2)/P(case)$$

What Happens in Screened Controls …

$$P(control\ \&\ low\ risk) = (1-p)^2(1-f)$$
$$P(control\ \&\ med\ risk) = 2p(1-p)(1-fr)$$
$$P(control\ \&\ high\ risk) = p^2(1-fr^2)$$

$$P(control) = (1-p)^2(1-f) + 2p(1-p)(1-fr) + p^2(1-fr^2)$$

$$P(low\ risk|control) = (1-p)^2(1-f)/P(control)$$
$$P(med\ risk|control) = 2p(1-p)(1-fr)/P(control)$$
$$P(high\ risk|control) = p^2(1-fr^2)/P(control)$$

$$P(risk\ allele|control) = (p(1-p)(1-fr) + p^2(1-fr^2))/P(control)$$

Today

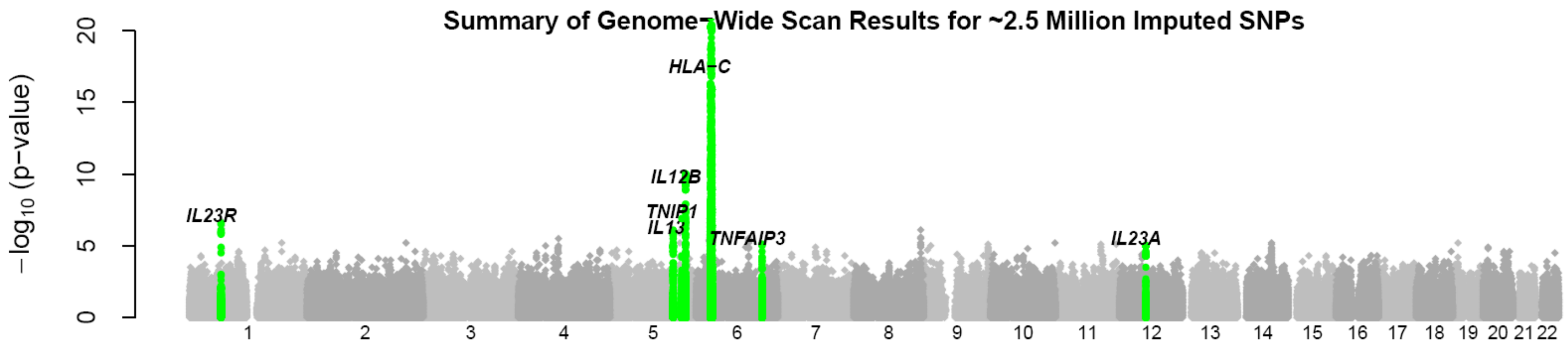
- A simple genetic model: frequency + risk
- A typical genomewide association study
- Power for genomewide association study
- Designing a two stage genomewide study
- Choices for analysis of two stage studies

Genomewide Association Studies

- Survey ~500,000 SNPs in a large set of cases and controls
  - Subset of SNPs is typically followed up in more samples
- Comprehensively survey common variants across genome
  - Via linkage disequilibrium, most common variants assessed
- Successful: many loci implicated in common disorders
  - Especially in contrast to results of candidate gene studies

Collaborative Association Study of Psoriasis: Example of a Successful GWAS

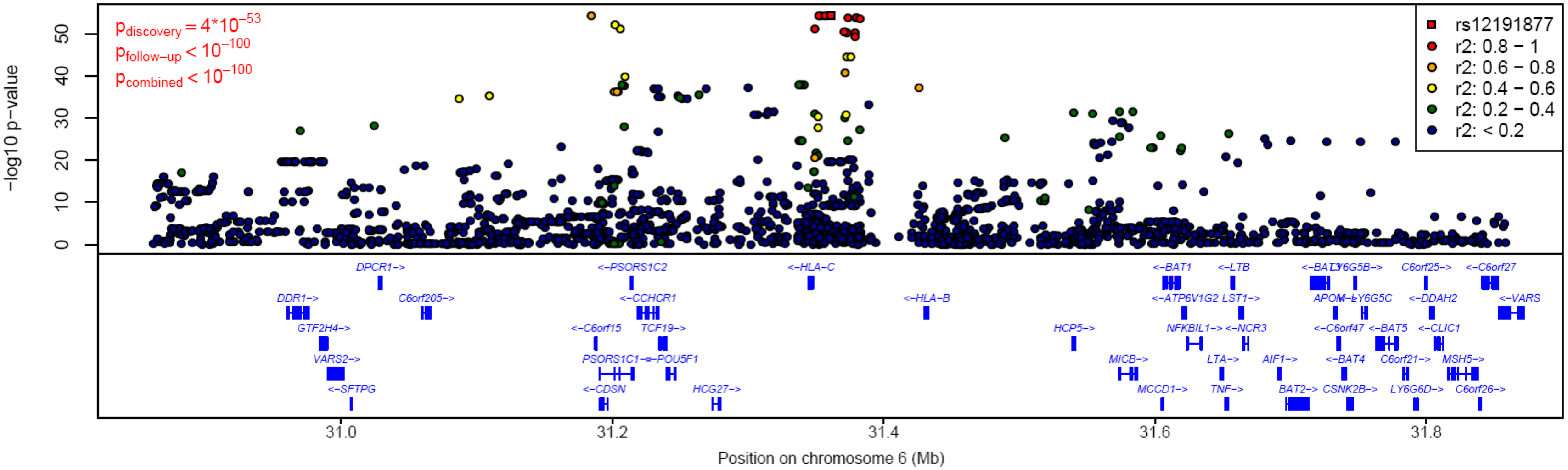
- Examined ~1,500 cases / ~1,500 controls at ~500,000 SNPs
- Examined 20 promising SNPs in extra ~5,000 cases / ~5,000 controls
- Outcome: 7 regions of confirmed association with psoriasis



Green hits have $p < 5x10^{-8}$ in final analysis

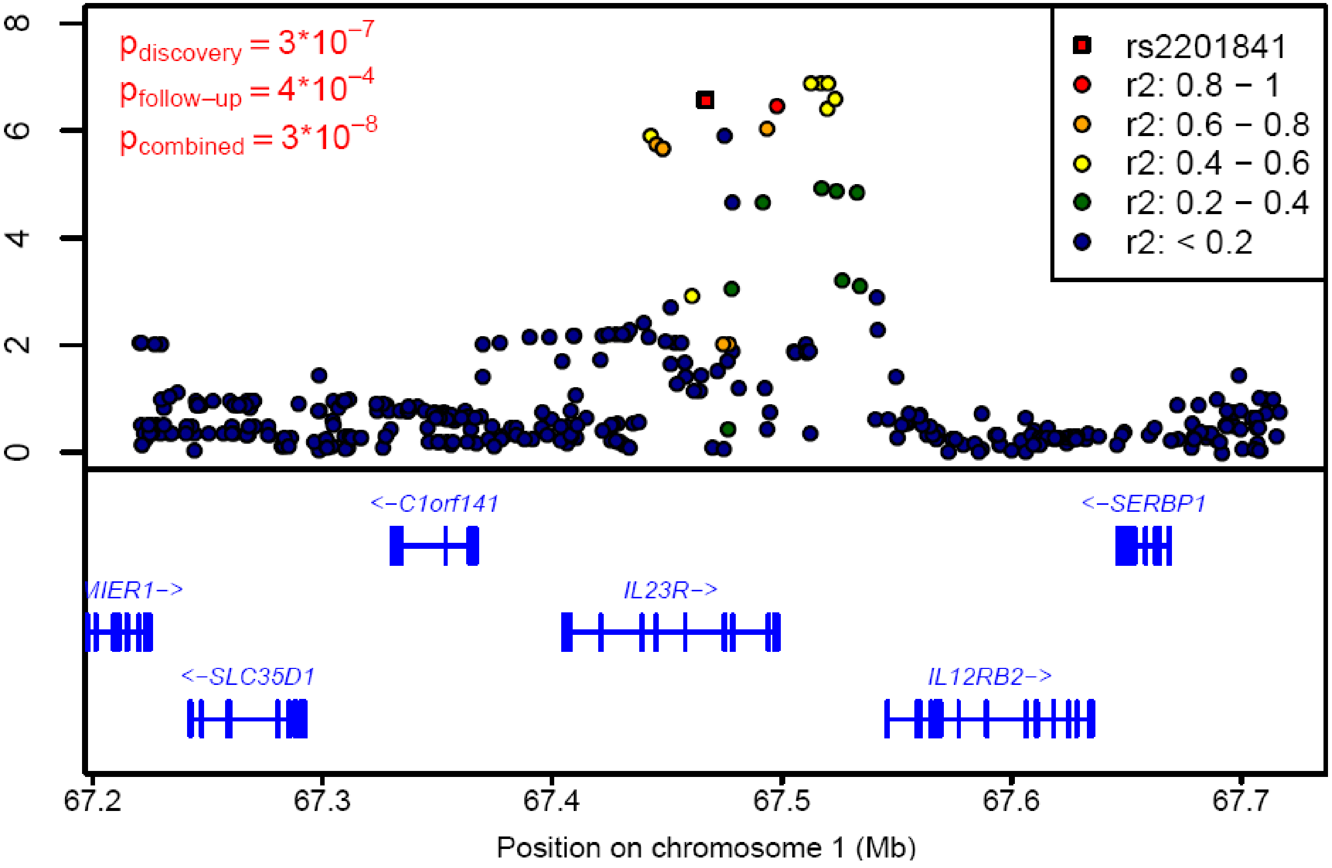
Nair et al, 2009

HLA-C



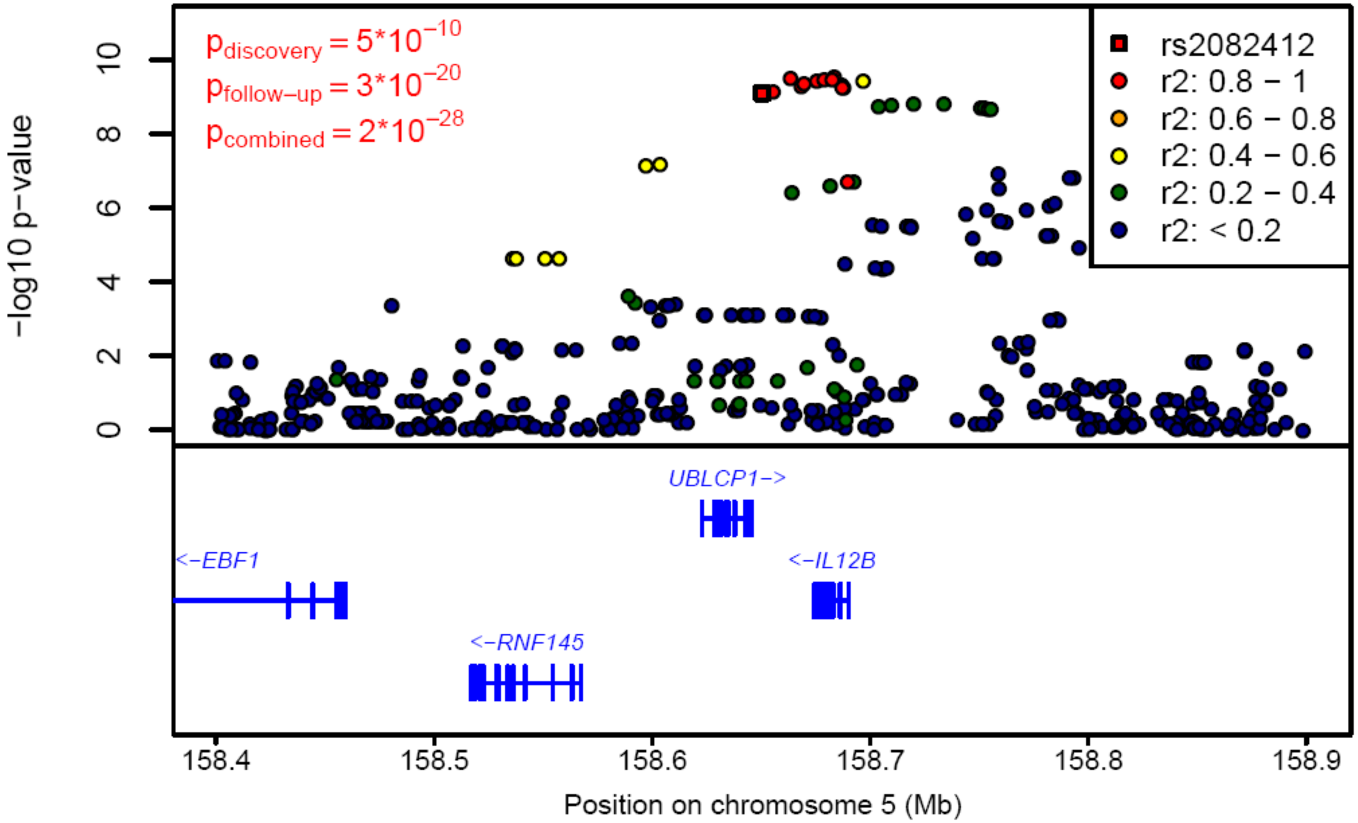
Top psoriasis associated SNPs in **strong linkage disequilibrium with HLA-Cw6**.
Evidence for psoriasis associated SNPs that are far from HLA-Cw6.

IL23R



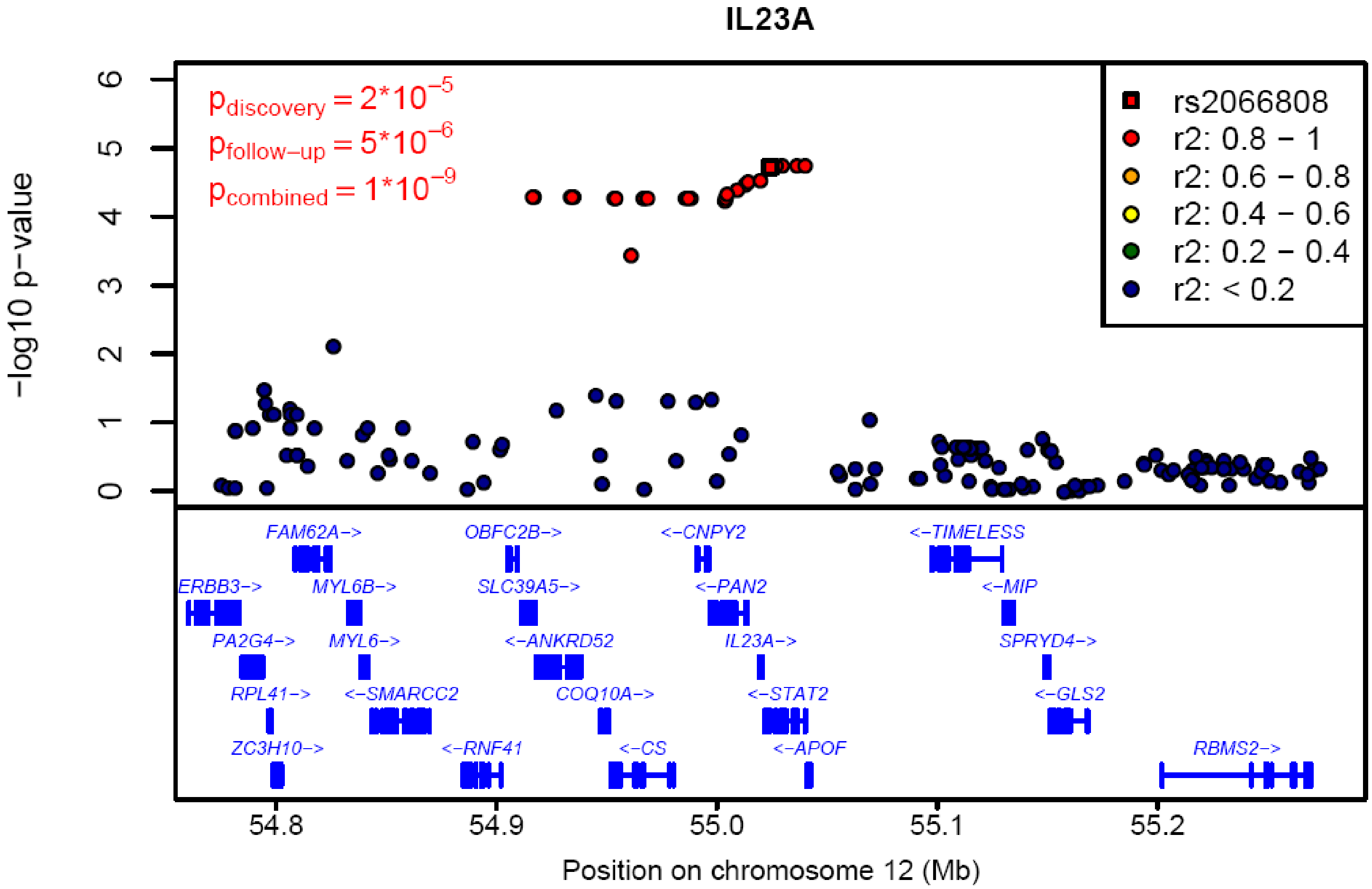
Previously identified locus, psoriasis associated SNPs also **associated with Crohn's**.

IL12B



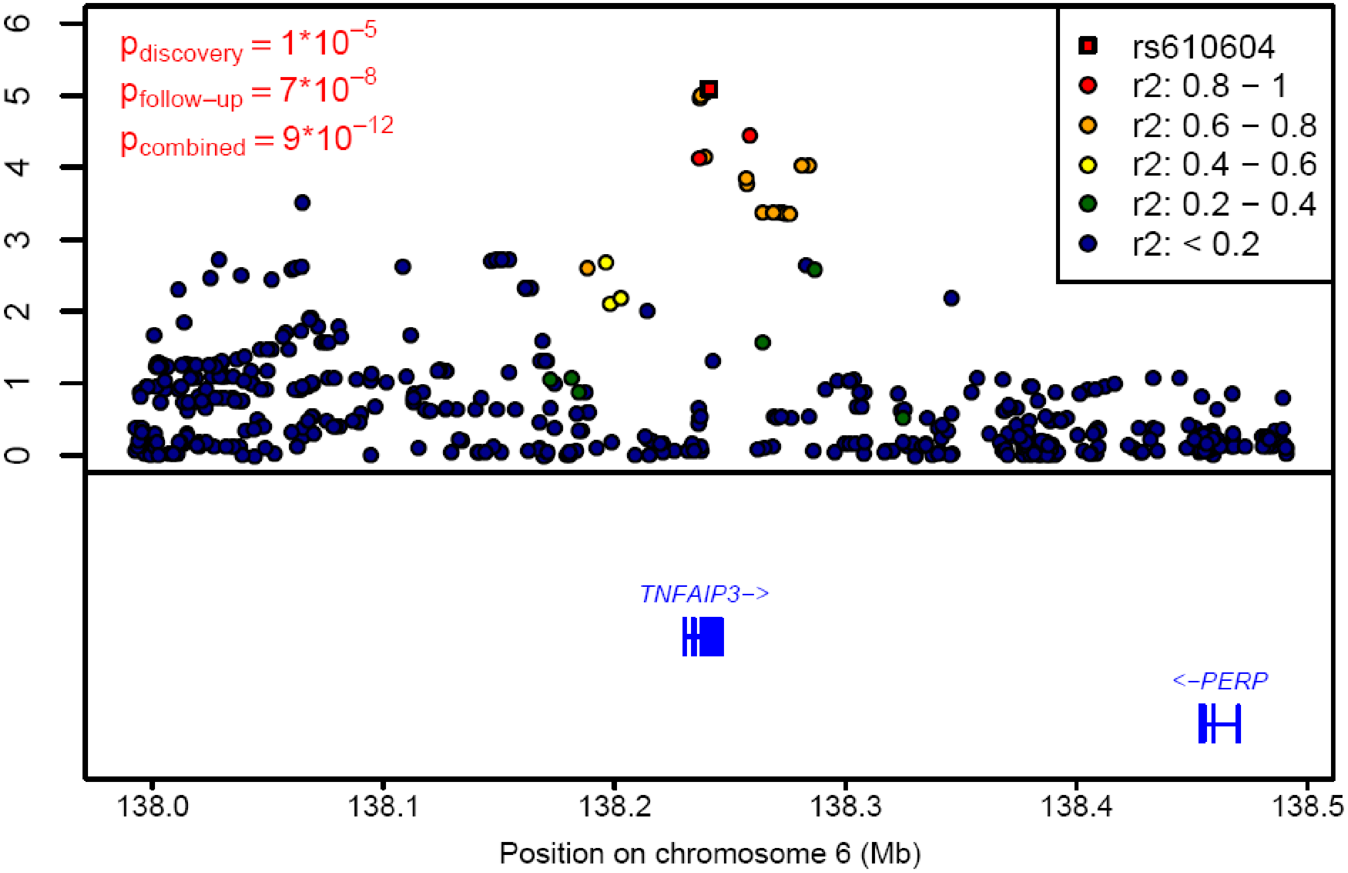
Previously identified locus, psoriasis associated SNPs **associated with Crohn's**.

IL23A



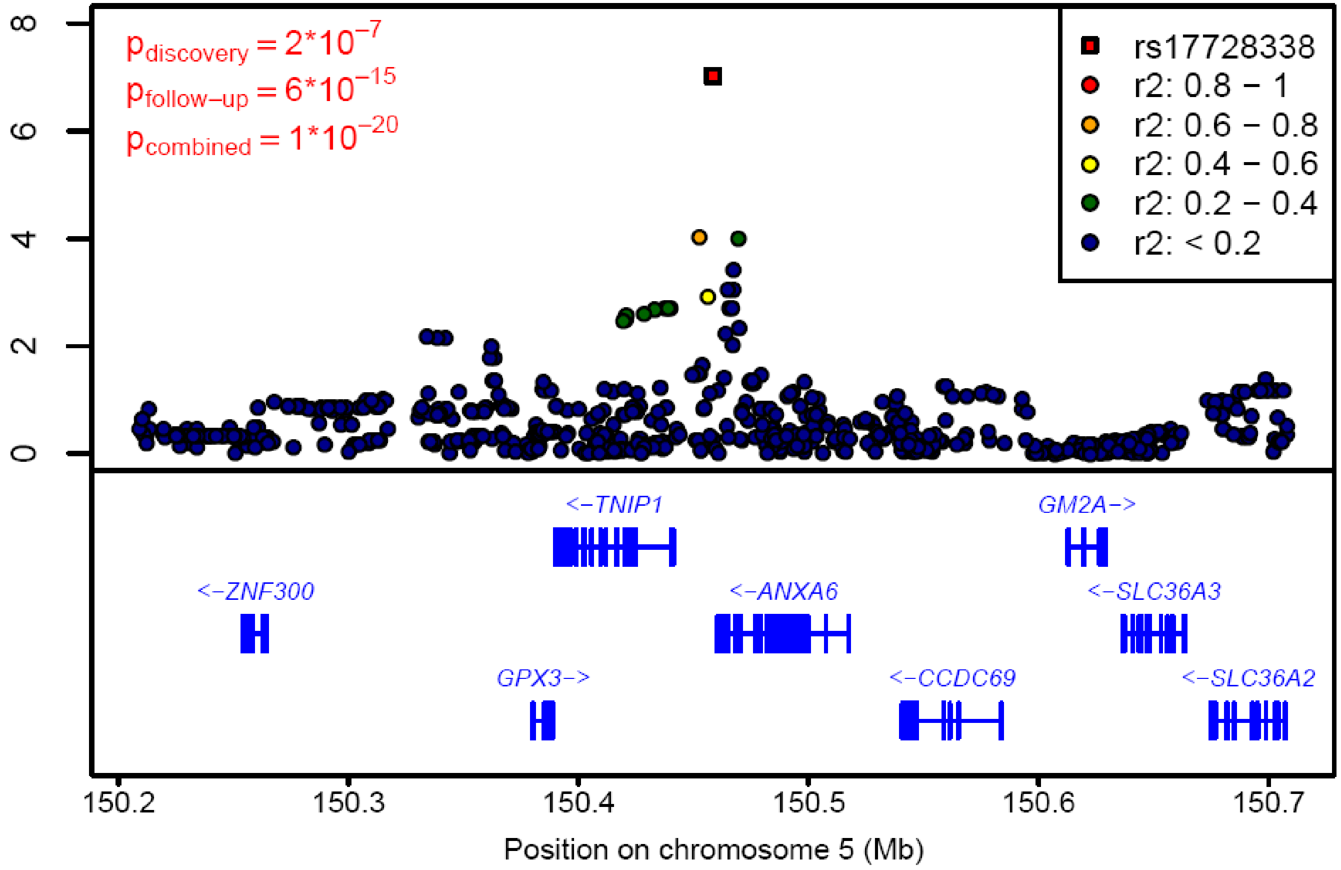
New locus, psoriasis associated SNPs **not associated** with Crohn’s.

TNFAIP3



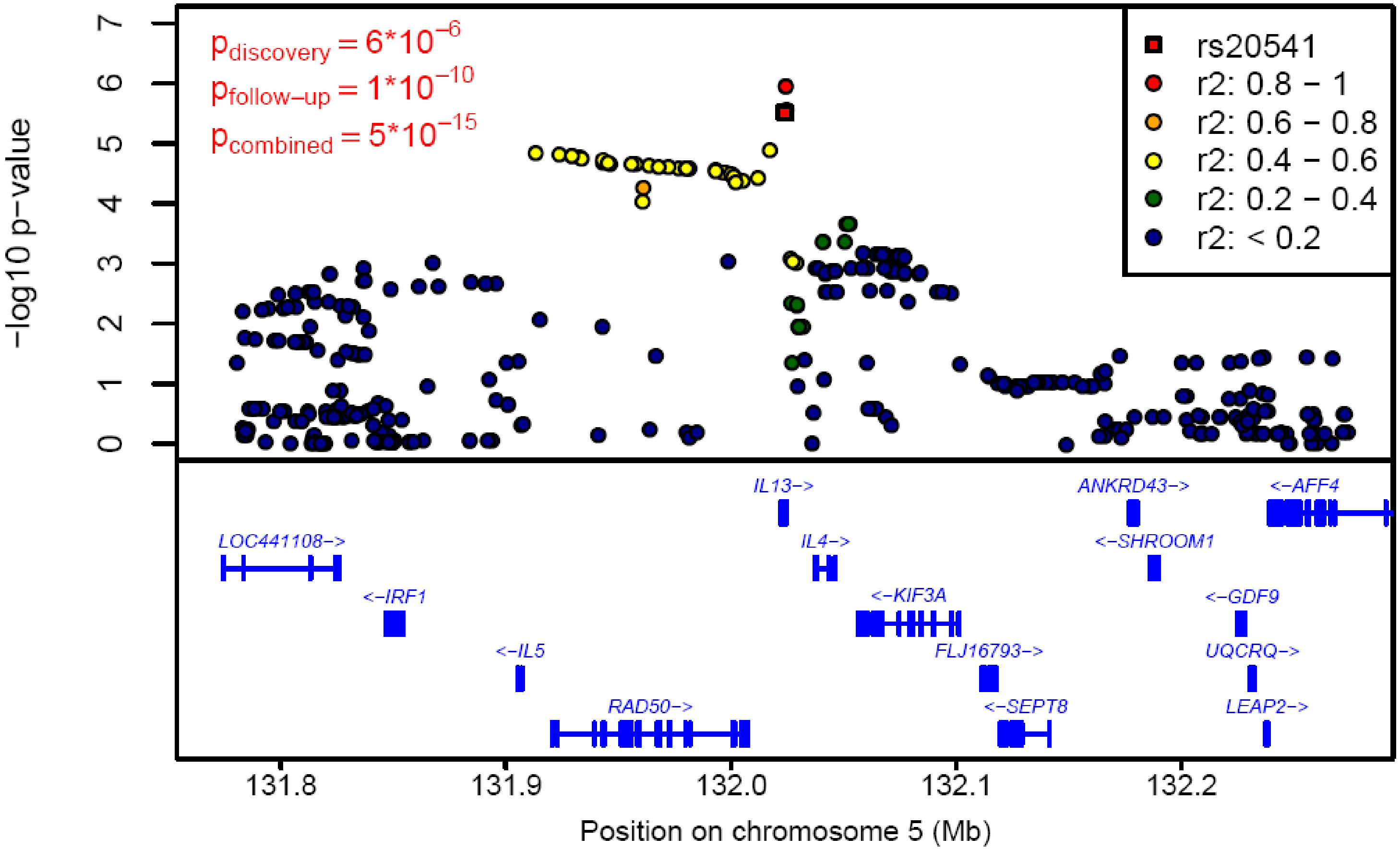
New locus; other SNPs in the locus are associated with lupus and rheumatoid arthritis.

TNIP1



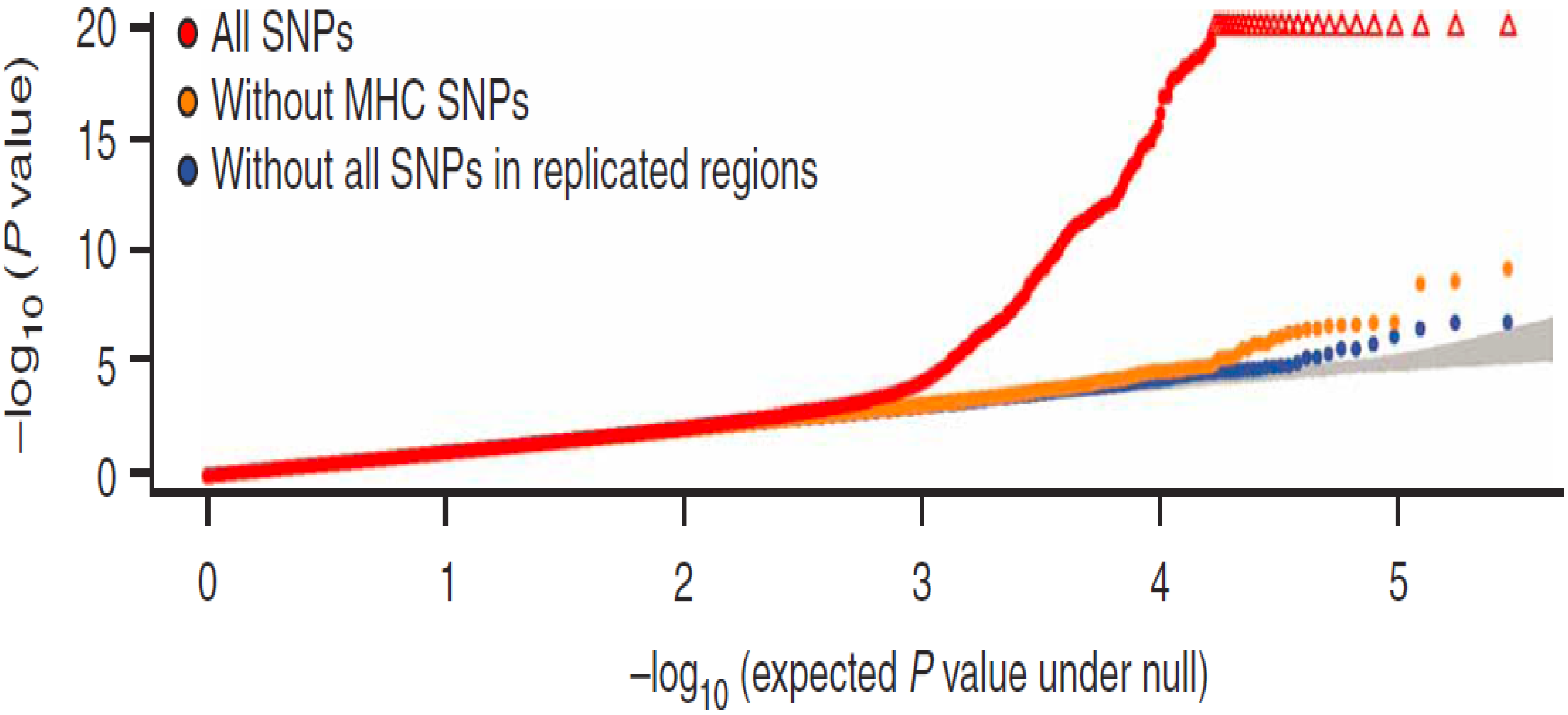
New locus; note potential evidence for independently associated alleles.

IL4/IL13



New locus; IL4 and IL13 are excellent functional candidates.

Q-Q Plot



Genomic control = 1.03

Multiple hits within a pathway…

- Three of the top replicated hits are for:

– IL23R (IL-23 receptor)	3×10^{-8}
– IL23A (IL-23 subunit)	9×10^{-10}
– IL12B (IL-23/IL-12 subunit)	1×10^{-28}

- Two other replicated hits at:

– TNFAIP3 (TNFα-inducible protein 3)	$9x10^{-12}$
– TNIP1 (TNFAIP3 interacting protein 1)	$1x10^{-20}$

- Evidence for epistasis among these SNPs?
  - None.

Summary of Results

SNP	Stage 1			Stage 2			P-value	Nearby Genes
	f_{cases}	$f_{controls}$	OR	f_{cases}	$f_{controls}$	OR		
rs12191877	.31	.14	2.79	.30	.15	2.64	$<10^{-100}$	HLA-C
rs2082412	.86	.79	1.56	.85	.80	1.44	2x10^{-28}	IL12B
rs17727338	.09	.06	1.72	.09	.05	1.59	1x10^{-20}	TNIP1
rs20541	.83	.78	1.37	.83	.79	1.27	5x10^{-15}	IL13
rs610604	.37	.32	1.28	.36	.32	1.19	9x10^{-12}	TNFAIP3
rs2066808	.96	.93	1.68	.95	.93	1.34	1x10^{-9}	IL23A
rs2201841	.35	.29	1.35	.32	.30	1.13	3x10^{-8}	IL23R

Notice how estimated effect size is consistently higher in Stage 1.
The “Winner’s Curse” is a common feature of genomewide studies.

Power Calculations

- For a given genetic model, evaluate alternative study designs

- For a given study design, identify genetic models that are likely to be detected

- Typically deal with many uncertainties…
  - What is an appropriate genetic model?
  - What is a desirable level of power?

Test Statistic

$$z = \frac{\hat{p}' - \hat{p}}{\sqrt{[\hat{p}'(1 - \hat{p}') + \hat{p}(1 - \hat{p})]/2N}}$$

Where:

$\hat{p}$’ is the observed case allele frequency

$\hat{p}$ is the observed control allele frequency

N is the number of cases and controls

Distribution Under the Null

- Under the null hypothesis p = p’

- Z is distributed as Normal(0, 1)

- Derive P-value thresholds for target significance level α
- Using Inverse Normal Cumulative Distribution Function
  - $\alpha = 0.05$ leads to $C = -\Phi^{-1}\left(\frac{0.05}{2}\right) = 1.96$
  - $\alpha = 5 \cdot 10^{-8}$ leads to $C = -\Phi^{-1}\left(\frac{5 \cdot 10^{-8}}{2}\right) = 5.45$

Distribution Under The Alternative

- For a specific set of expected case and control allele frequencies, …

- …we can calculate expected value of test statistic

$$\mu = \frac{p' - p}{\sqrt{[p'(1-p') + p\,(1-p)]/2N}}$$

- Under the alternative, statistic is Normal(μ, 1).

Power

- To calculate power, we first calculate:
  - Significance threshold ***C***
  - Expected test statistic $\boldsymbol{\mu}$

- Use normal cumulative distribution function Φ

- $P(|Z| > C)$

$$= P(Z > C) + P(Z < -C)$$
$$= 1 - \Phi(C - \mathrm{\mu}) + \Phi(-C - \mathrm{\mu})$$

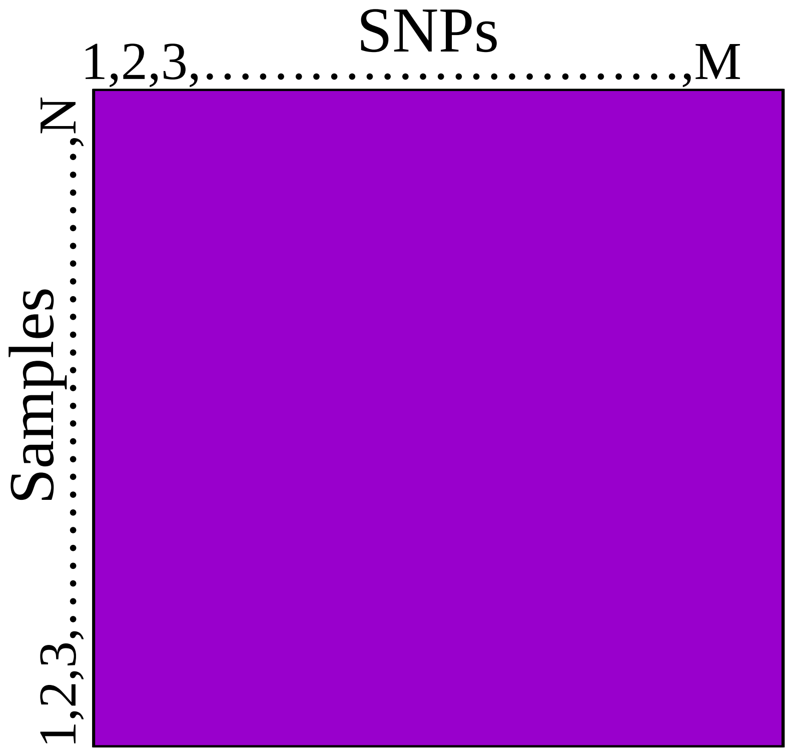
Example

- Test 1,000,000 independent markers
  - $\alpha = 0.05/1{,}000{,}000 = 5\text{x}10^{-8}$
  - C = 5.45

- Case allele frequency p' = 0.55
- Control allele frequency p = 0.45
- $N_{cases} = N_{controls} = 1{,}000$
- $\mu = 6.35$

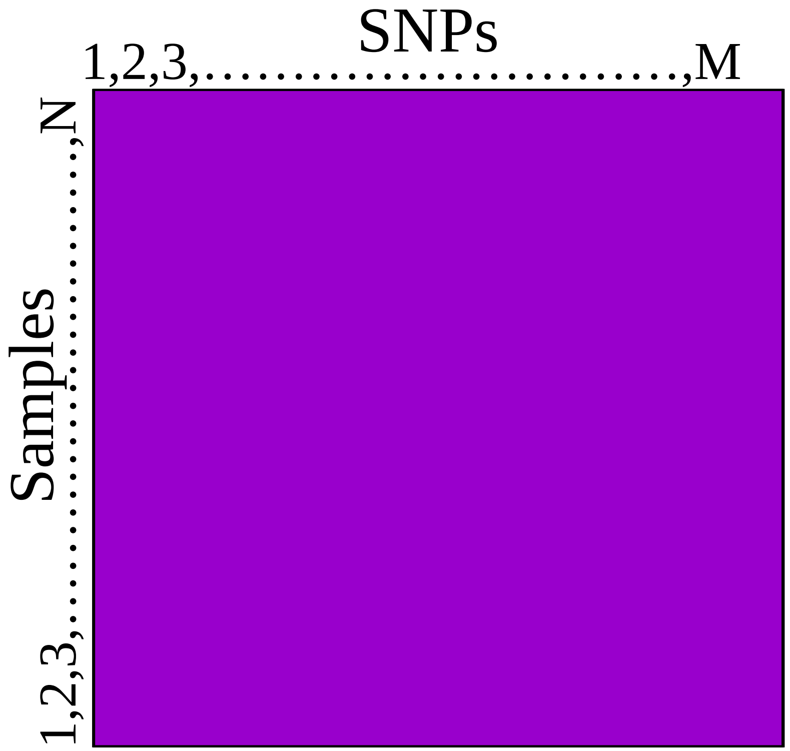
- Power = 81%
  - If N = 500, power = 17%
  - If N = 2000, power = 100%

One Stage Genomewide Study


SNPs
1,2,3,.........................,M
Samples
1,2,3,.........................,N

A comprehensive study might examine all M SNPs in all N samples.

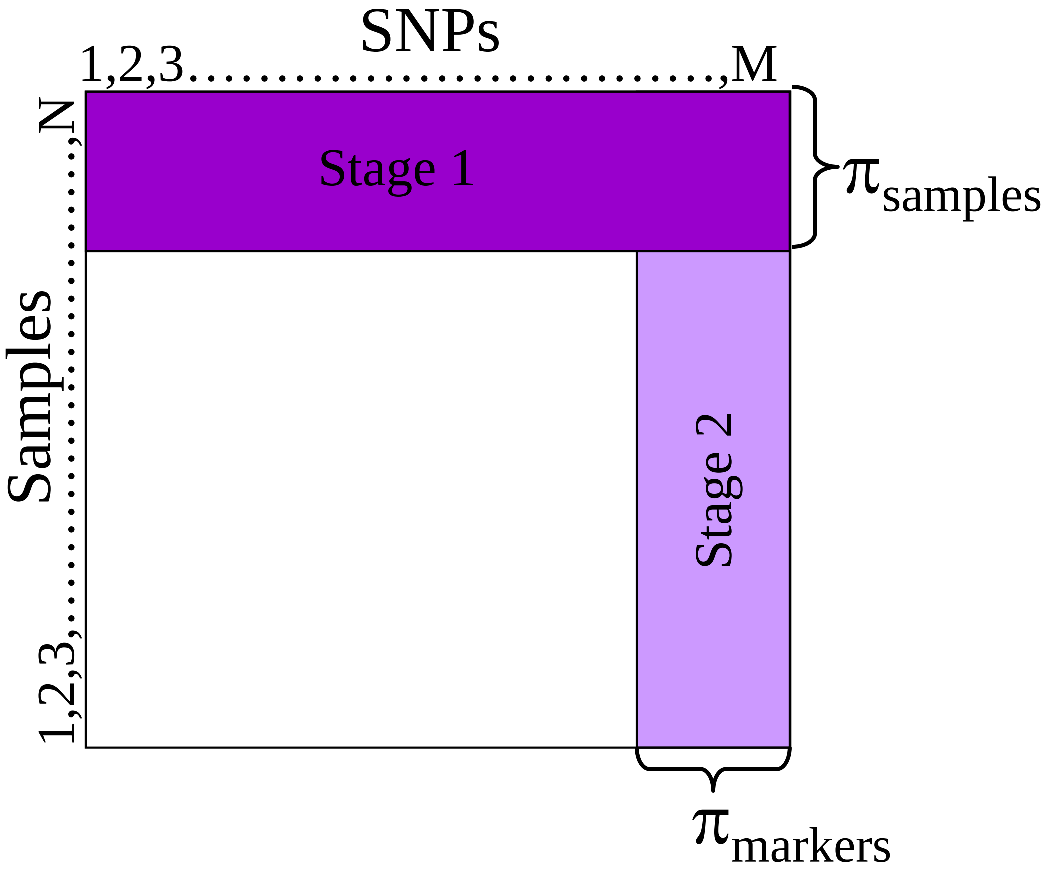
Analysis of One Stage Study



Declare significance using p-value threshold of 0.05 / M.
Threshold of $5x10^{-8}$ is typical, assumes 1 million independent tests.

Two Stage
Genomewide Association Studies

Two Stage Genomewide Study



A more cost effective study might only examine:

- All SNPs in a fraction of samples, $\pi_{samples}$
- All individuals for a fraction of markers, $\pi_{markers}$

Relative Genotyping Effort

- The total number of genotypes required in a two stage study is …
- $N_{genotypes} = MN\pi_{samples} + MN(1 - \pi_{samples})\pi_{markers}$
- For example, if we …
  - Genotype 30% of samples in Stage 1
  - Follow-up 0.1% of markers in Stage 2

  - Total number of genotypes will be reduced 69.93%

Relative Cost

- The reduction in cost is typically less dramatic …
- … but still substantial

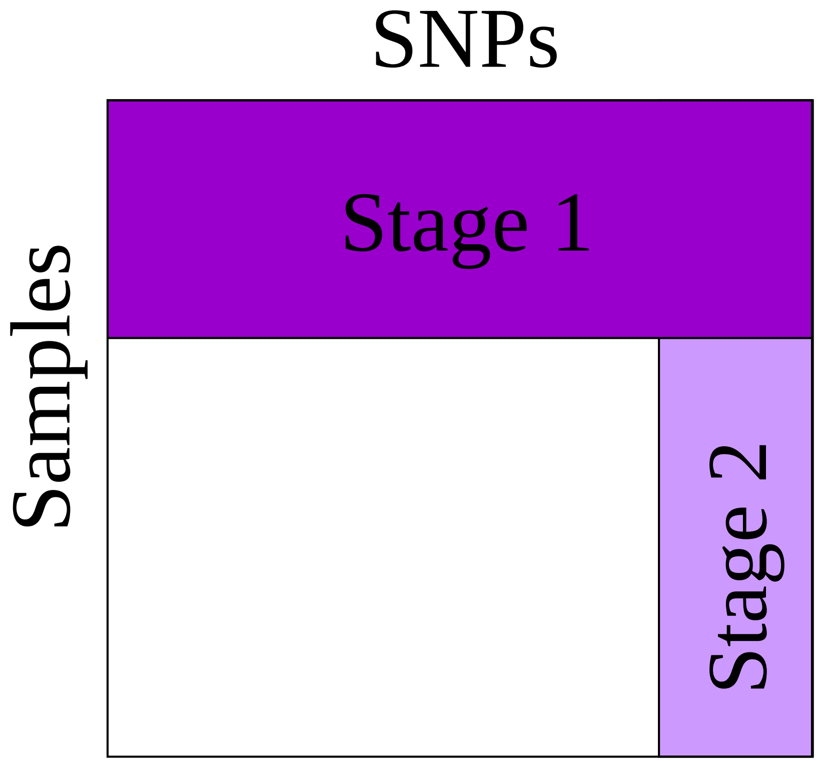
- Main limitation is that genotyping is cheaper “in bulk”
  - τ is ratio of stage 1 to stage 2 costs on a per genotype basis

- $Cost\ ratio = \pi_{samples} + (1 - \pi_{samples})\pi_{markers}\tau$

- For example, if we …
  - Genotype 30% of samples in Stage 1
  - Follow-up 0.1% of markers in Stage 2
  - Relative cost ratio is 100

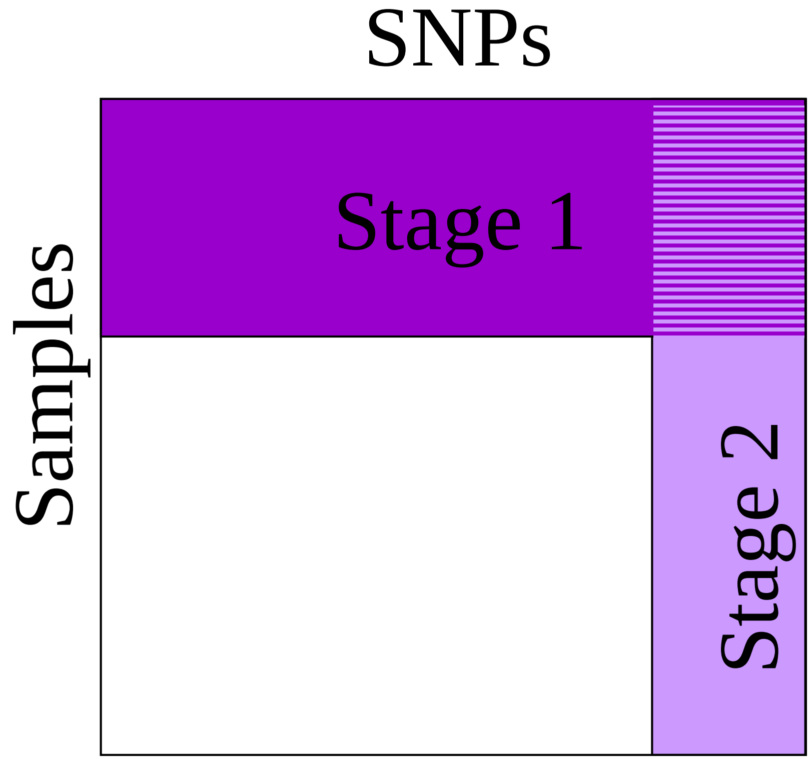
  - Total cost will be reduced 63.00%

Replication Based Analysis



Select markers to follow-up using p-value threshold of $\pi_{markers}$.
Declare significance using threshold of $0.05/(M \cdot \pi_{markers})$
Final analysis uses only stage 2 samples.

Joint Analysis



Select markers to follow-up using p-value threshold of $\pi_{markers}$.
Declare significance using threshold of approximately 0.05/M.
Final analysis uses stage 1 and stage 2 samples.

Power for Replication Based Analysis

- Simplest approach would be to calculate
  - C_1 and C_2 as the significance thresholds for each stage
  - μ_1 and μ_2 as the expected statistics for each stage
  - P_1 and P_2 as the power for each stage
  - $P_{replication} = P_1 P_2$ as the overall power

- Refined analysis might enforce that stage 1 and stage 2 statistics should have the same sign

$$P_2 = (1 - \Phi[C_2 - \mu_2]) \frac{1 - \Phi[C_1 - \mu_1]}{1 - \Phi[C_1 - \mu_1] + \Phi[-C_1 - \mu_1]} + \Phi[-C_2 - \mu_2] \frac{\Phi[-C_1 - \mu_1]}{1 - \Phi[C_1 - \mu_1] + \Phi[-C_1 - \mu_1]}$$

Power for Joint Analyses

- Simplest approach would be to calculate
  - C_1 and C as stage 1 and overall significance thresholds
  - μ_1 and μ as stage 1 and overall expected statistics
  - P_1 and P as stage 1 and single stage study power
  - $P_{\text{joint}} = P_1P$ as the overall power

- Refined analysis models joint distribution of stage 1 and overall test statistic

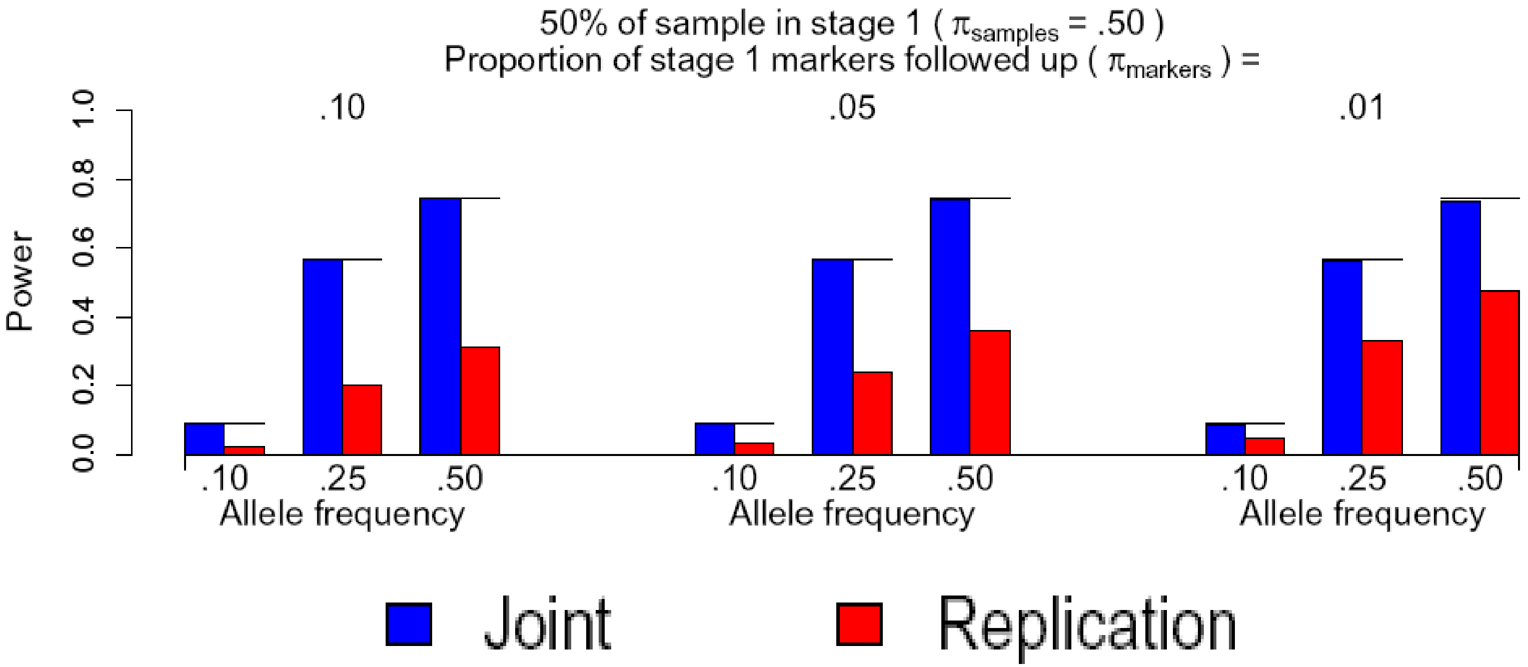
$$
\begin{aligned}
P_{\text{joint}} &= P(|z_{\text{joint}}| > C_{\text{joint}}|T) \\
&= \int_{-\infty}^{-C_1} [P(z_{\text{joint}} > C_{\text{joint}}|z_1 = x) + P(z_{\text{joint}} < -C_{\text{joint}}|z_1 = x)] f(x|T)dx \\
&+ \int_{C_1}^{\infty} [P(z_{\text{joint}} > C_{\text{joint}}|z_1 = x) + P(z_{\text{joint}} < -C_{\text{joint}}|z_1 = x)] f(x|T)dx
\end{aligned}
$$

$T: |Z| > C_1$

Replication or Joint Analysis?

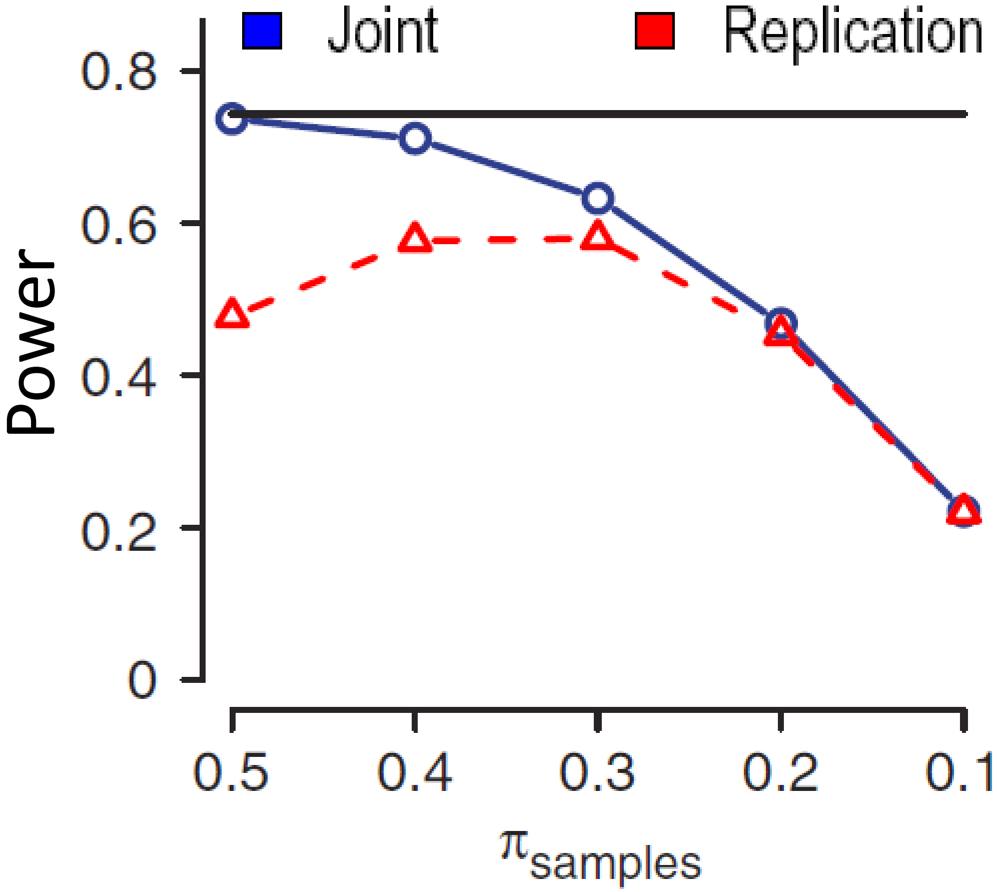
- Replication based analysis
  - Requires smaller multiple testing adjustment
- Joint analysis uses more data
  - We expect stronger signal using all available data
- Both analyses are compatible with the same experimental design

Replication of Joint Analysis?



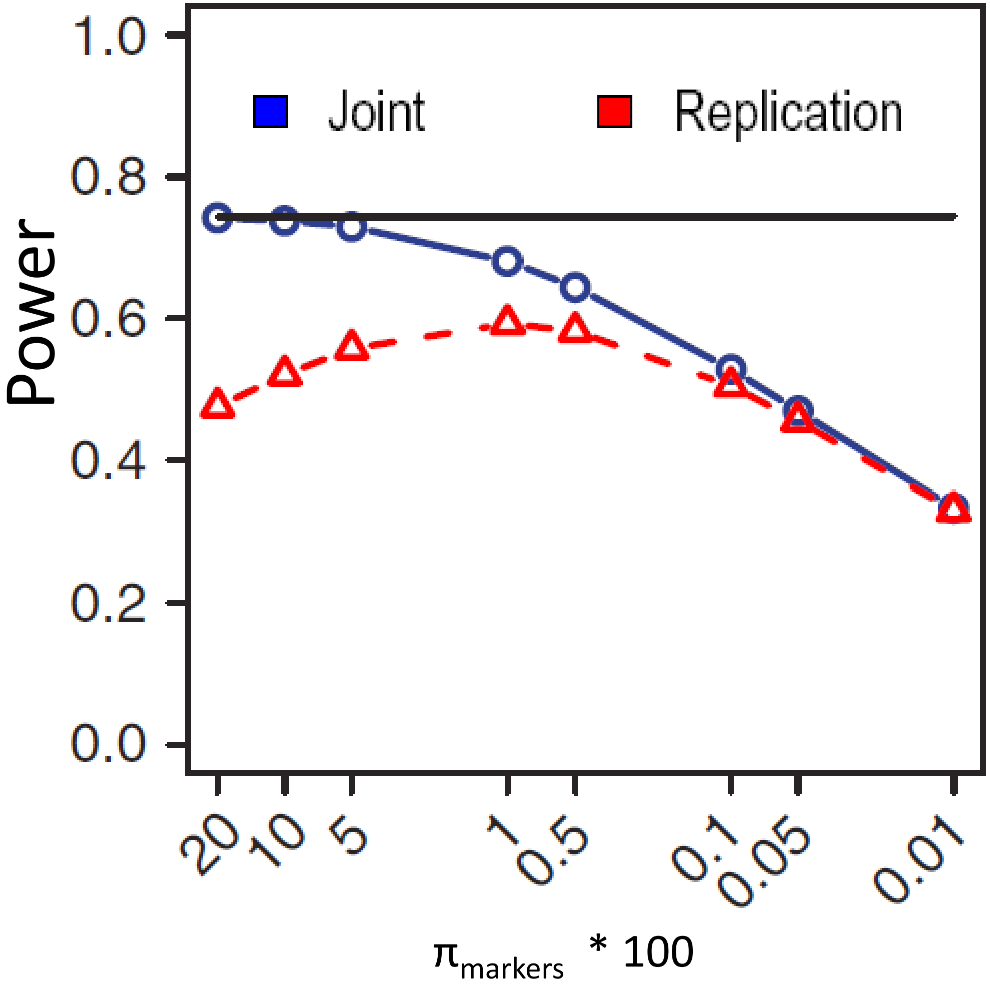
300,000 markers genotyped on 1000 cases, 1000 controls
Multiplicative model, prevalence 10%, GRR = 1.4

Replication or Joint Analysis? Effect of Varying $\pi_{samples}$



- α=0.05 / 300,000
- $\pi_{markers}$ = 0.01
- N = 1,000
- p = 0.50
- p‘= 0.66

Replication or Joint Analysis? Effect of Varying $\pi_{markers}$



- α=0.05 / 300,000
- $\pi_{samples}$ = 0.30
- N = 1,000
- p = 0.50
- p‘= 0.66

Refining Calculation

- Instead of setting p and p’ arbitrarily, use a genetic model

- Suppose that the relative risk of disease is:
  - Baseline for those with no risk alleles
  - r_1 for those with one risk allele
  - r_2 for those with two risk alleles

- Then:

$$p' = \frac{p(1-p)r_1 + p^2 r_2}{(1-p)^2 + 2p(1-p)r_1 + p^2 r_2}$$

Refining Calculation II

- Instead of setting p and p’ arbitrarily, use a genetic model

- Suppose that controls are known to be free of disease and *K* is the disease prevalence

- Then:

$$p_{control} = \frac{p - Kp'}{1 - K}$$

Some Important Messages

- Power calculations can help design study
  – How to best invest limited funds?

- Well designed two stage studies approximate power of more costly studies where all samples genotyped at all markers

- Joint analysis is much more efficient than replication based analyses

Recommended Reading

- Skol el al (2006) Joint analysis is more efficient than replication based analysis for two-stage genomewide association studies. *Nature Genetics* **38:**209-13

- Nair et al (2009) Genomewide scan reveals association of psoriasis with IL-23 and NF-kB pathways. *Nature Genetics* **41:**199-204