National Cancer Institute

STUDY DESIGN

Facilitating Collaboration in Genome-Wide Association Studies

Robert N. Hoover, M.D., Sc.D.

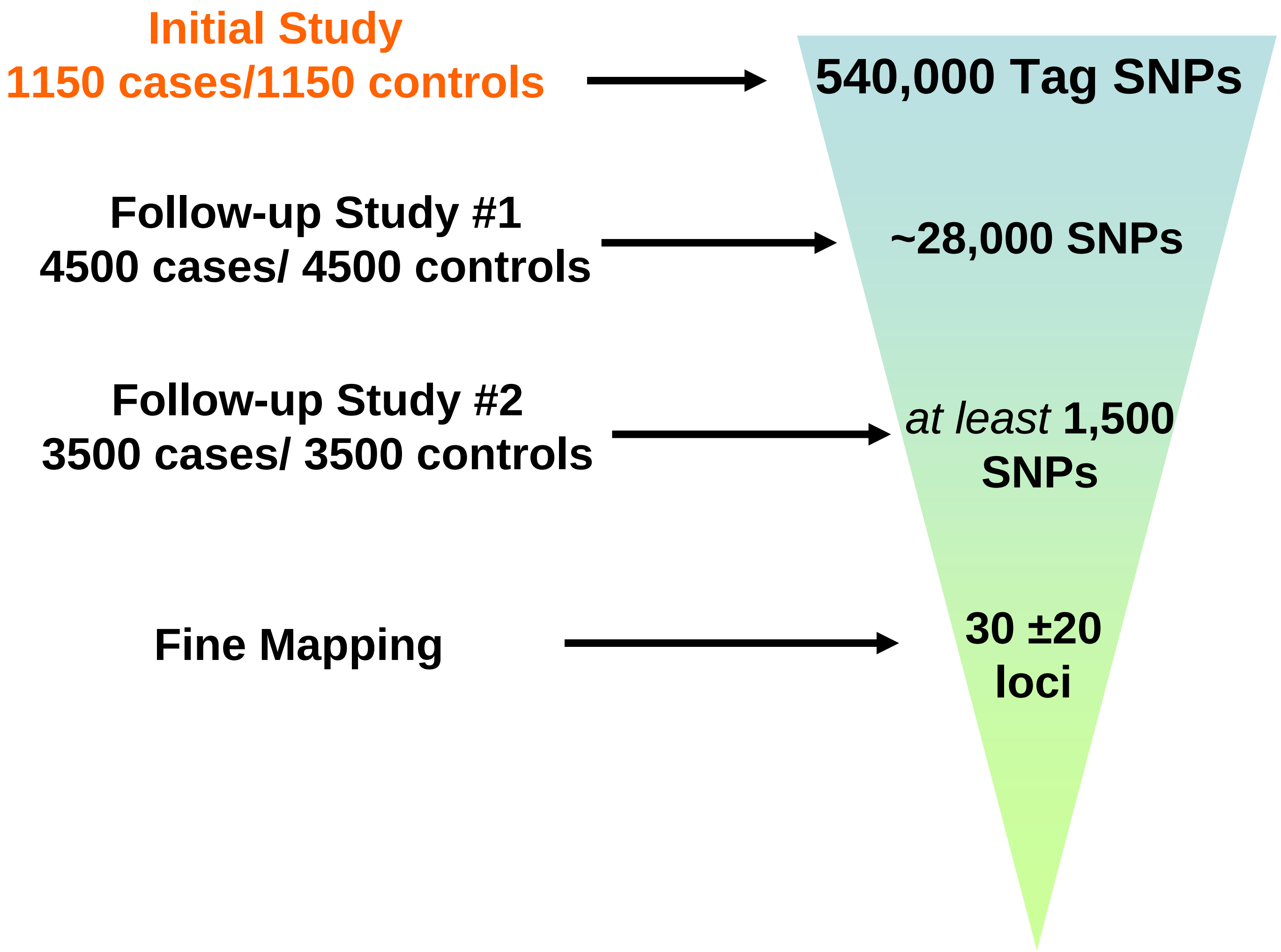
Division of Cancer Epidemiology and Genetics

National Cancer Institute

U.S. DEPARTMENT
OF HEALTH AND
HUMAN SERVICES

National Institutes
of Health

General Strategy for Prostate & Breast Cancer GWAS



Initial Study
1150 cases/1150 controls → 540,000 Tag SNPs

Follow-up Study #1
4500 cases/ 4500 controls → ~28,000 SNPs

Follow-up Study #2
3500 cases/ 3500 controls → *at least* 1,500 SNPs

Fine Mapping → 30 ±20 loci

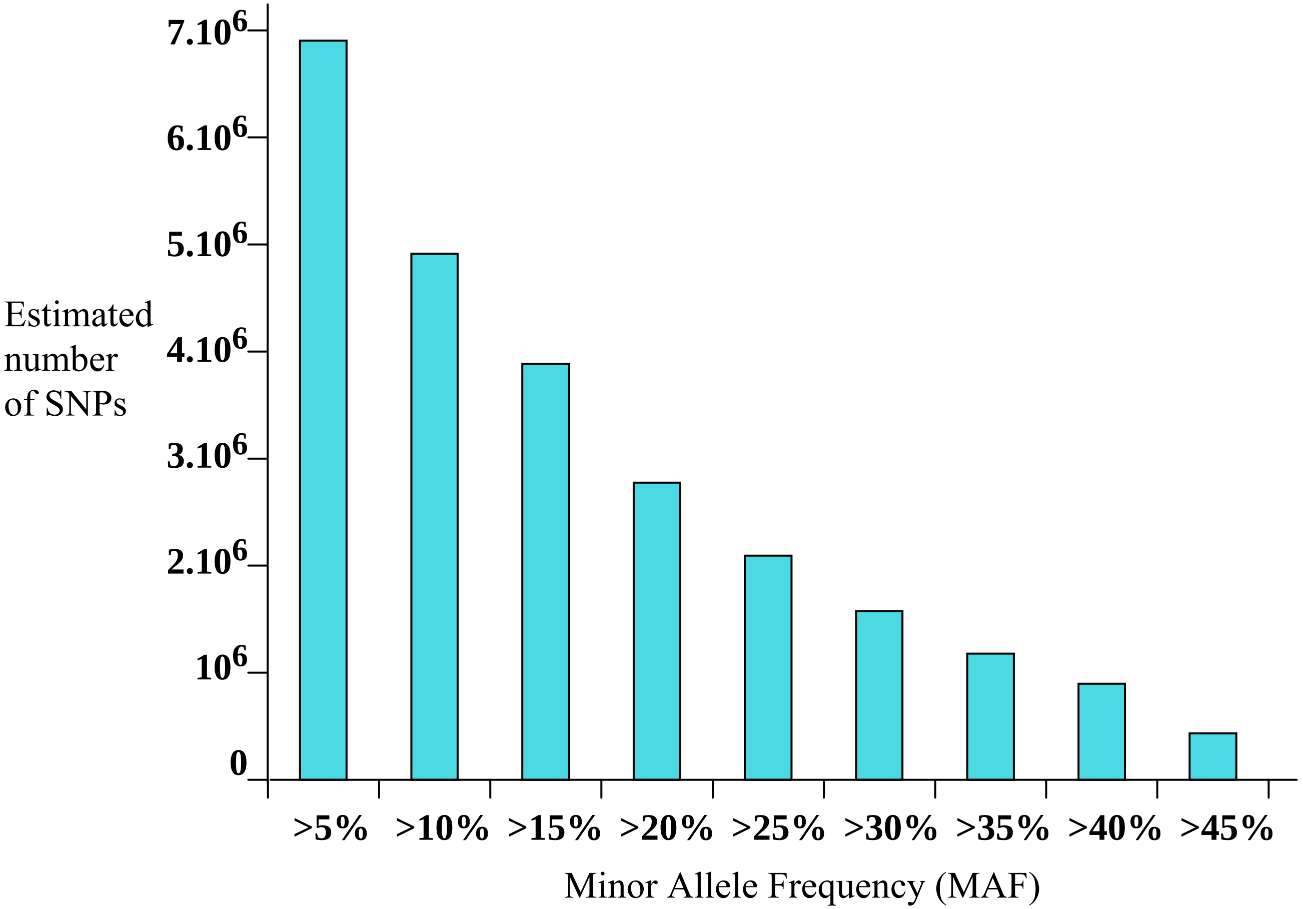
Considerations in Whole Genome Scans in Cancer

- Extent of Coverage of Genome
- Primary Scan
  - Adequate Size
  - Trade-off with effect
  - Study Design

Replication Strategy

  - Power calculations for how many stages
  - Joint vs consecutive analysis *(Skol Nat Genet 2006)*
  - Study Design

Estimated number of SNPs in the human genome as a function of their minor allele frequency



Common SNP : a SNP with MAF > 0.05 ; frequency of heterozytotes >≈ 10%

Adapted from Reich et al. Nat Genet (2003)

National Cancer Institute

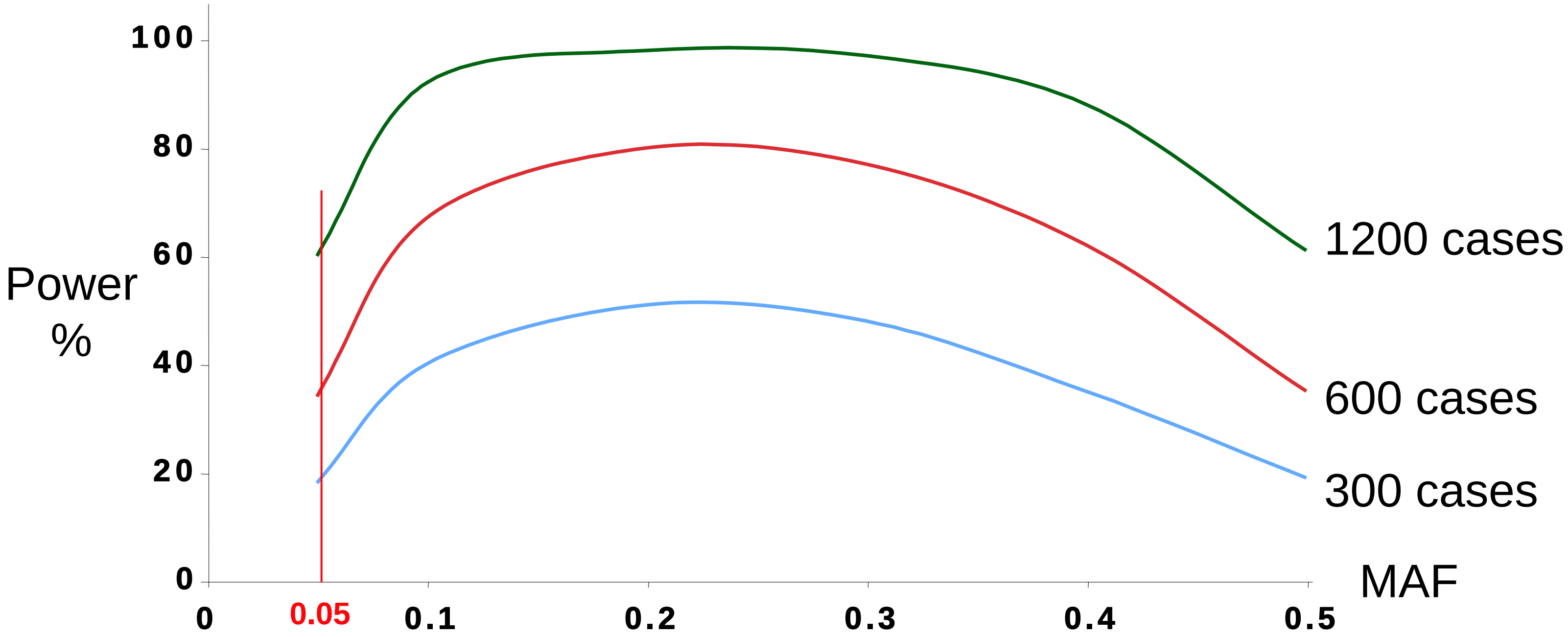
DESIGN ISSUES

DESIGN ISSUES

- Study Size
- Chance
- Bias

2-Stage WGS Strategy

Power as a function of MAF and sample sizes typed in the first stage



Disease model

- Prevalence 1%
- Single susceptibility SNP with a linkage disequilibrium r^2 = 0.8 with 1 genotyped SNP
- Dominant transmission
- Genotype relative risk : 1.5

Study design

# Cases = # Controls
# Cases in stage 1 : as indicated
# SNPs in stage 1 : 500,000
# Cases in stage 2 : 2,000
# SNPs in stage 2 : 25,000
Significance level 0.00002

Note: Significance level = 0.00002 => 10 false positives

National Cancer Institute

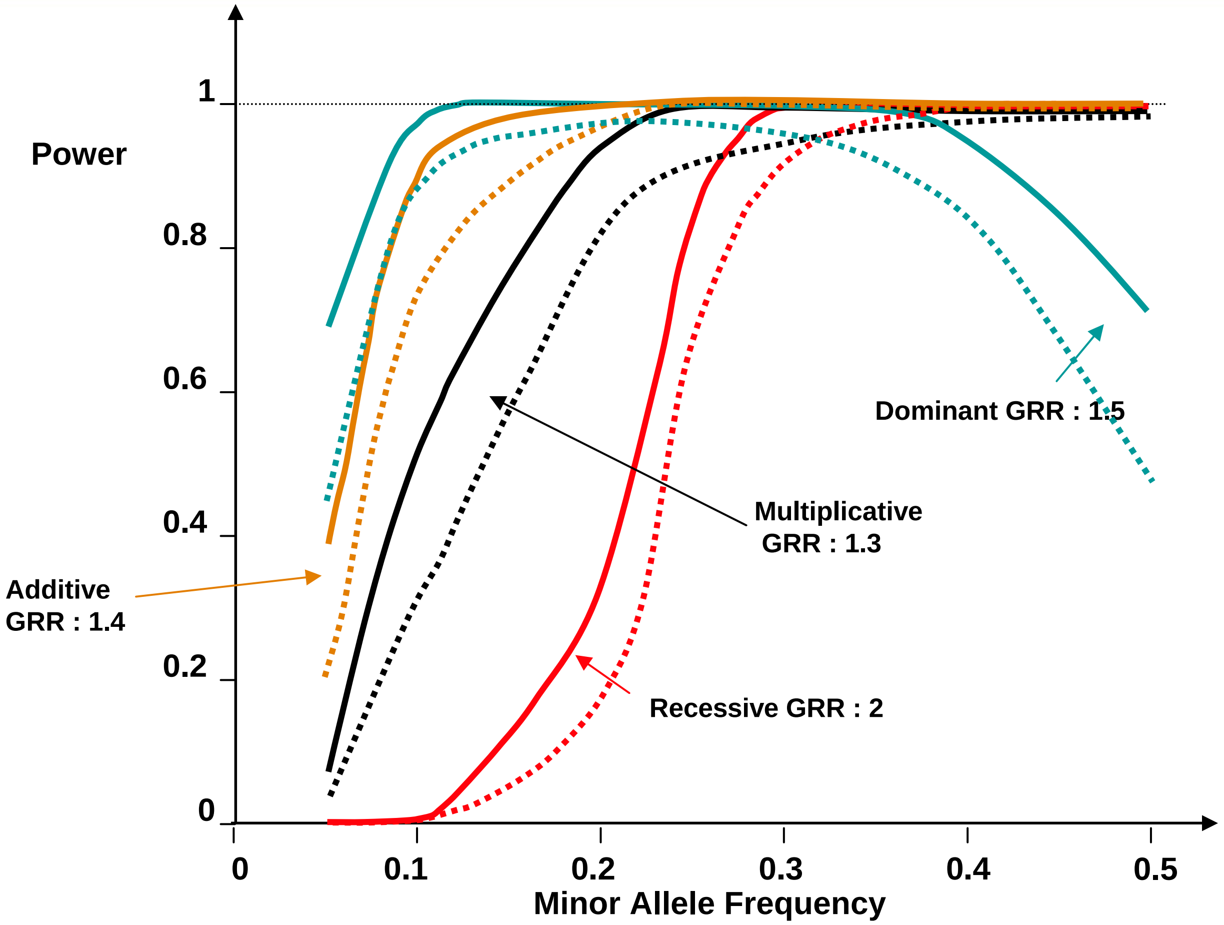
A quick note on ‘ideal’ power

- r^2 represents the statistical correlation between two loci
- It is a useful measure for association between susceptibility loci and SNPs
- Suppose SNP1 is involved in disease susceptibility and we genotype cases and controls at a nearby site SNP2
- To achieve the same power to detect associations at SNP2 as we would have at SNP1, sample size must increase by a factor of $1/r^2$

r^2	Additional Samples Required
0.50	100%
0.64	56%
0.70	43%
0.80	25%
0.90	11%
0.95	5%
1.00	0%

Power of the first two phases of CGEMS

Point wise significance 10^{-7} ; "genome wide" significance 0.05



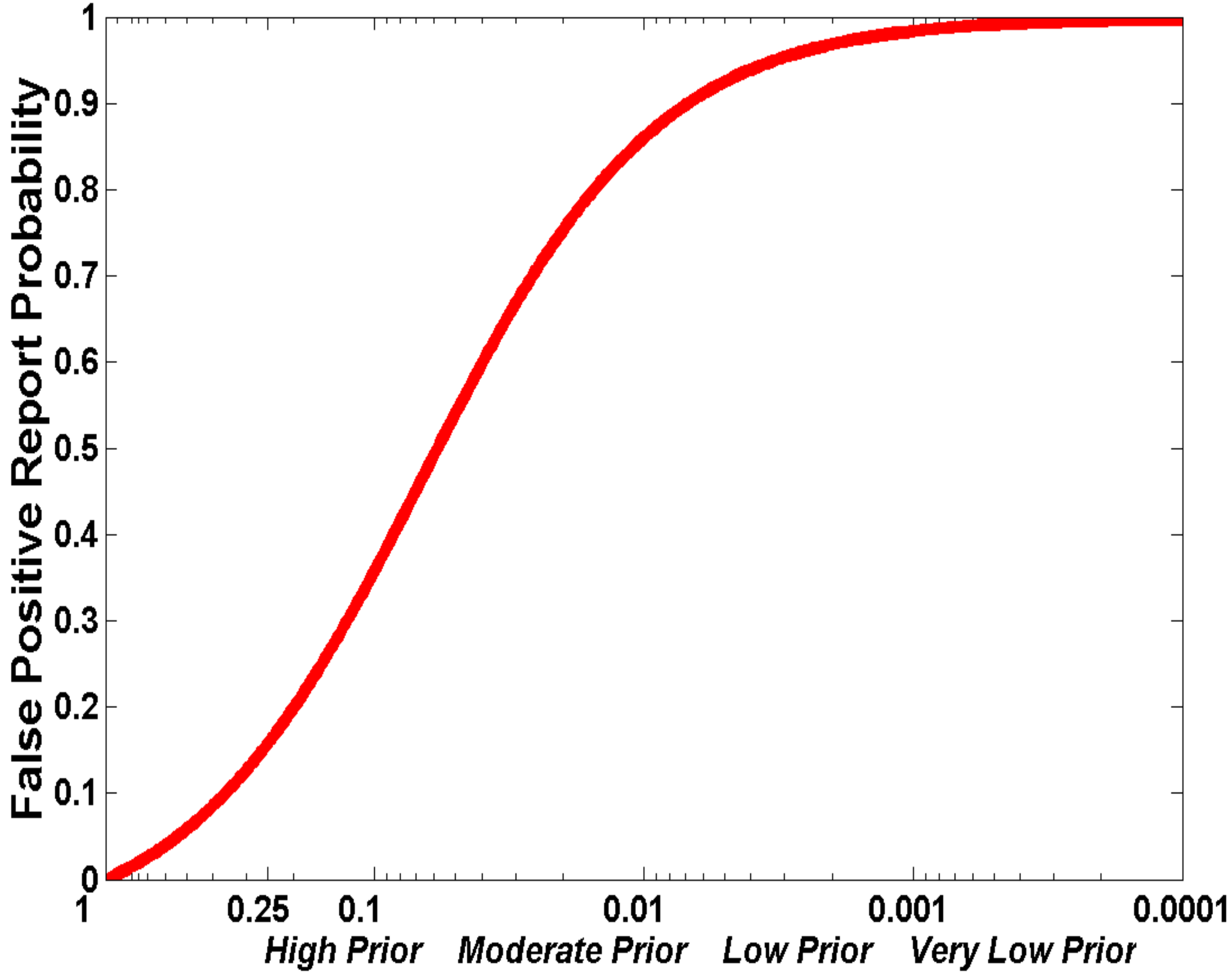
	GRR	AA	Aa	aa
Recessive	2.0	1.0	1.0	2.0
Dominant	1.5	1.0	1.5	1.5
Additive	1.4	1.0	1.4	1.8
Multiplicative	1.3	1.0	1.3	1.69

Continuous line : power for direct detection ($r^2 = 1$)
Dashed line : power for $r^2 = 0.8$

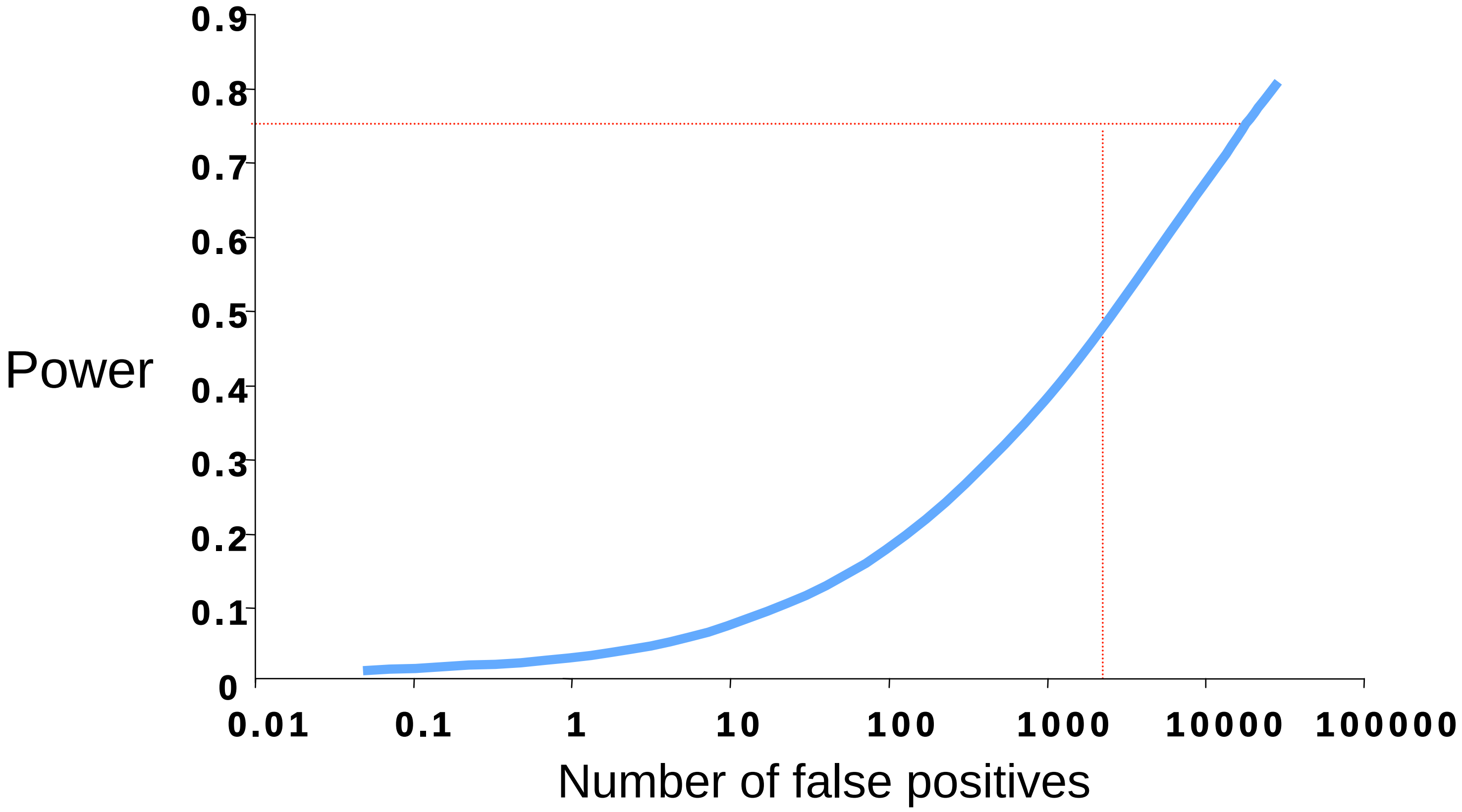
Skol et al. Nat Genet (2006)

Rejection of H_0 based on an alpha of 0.05

Power=0.8



Power of genome wide screen as a function of the number of retained false positive



Model :
One susceptibility allele : MAF = 0.1 , Odds Ratio = 1.4
LD of typed marker with susceptibility marker : r^2 = 0.8
Number of cases/control pairs : 1,200
Number of markers types : 500,000

Design Considerations

- Disease:
  - Incident
  - Prevalent
- Type:
  - Cohort
  - Case-control
    - Population-based
    - Hospital-based
- Quality:
  - Diagnosis (phenotype)
  - Study base
  - Biases

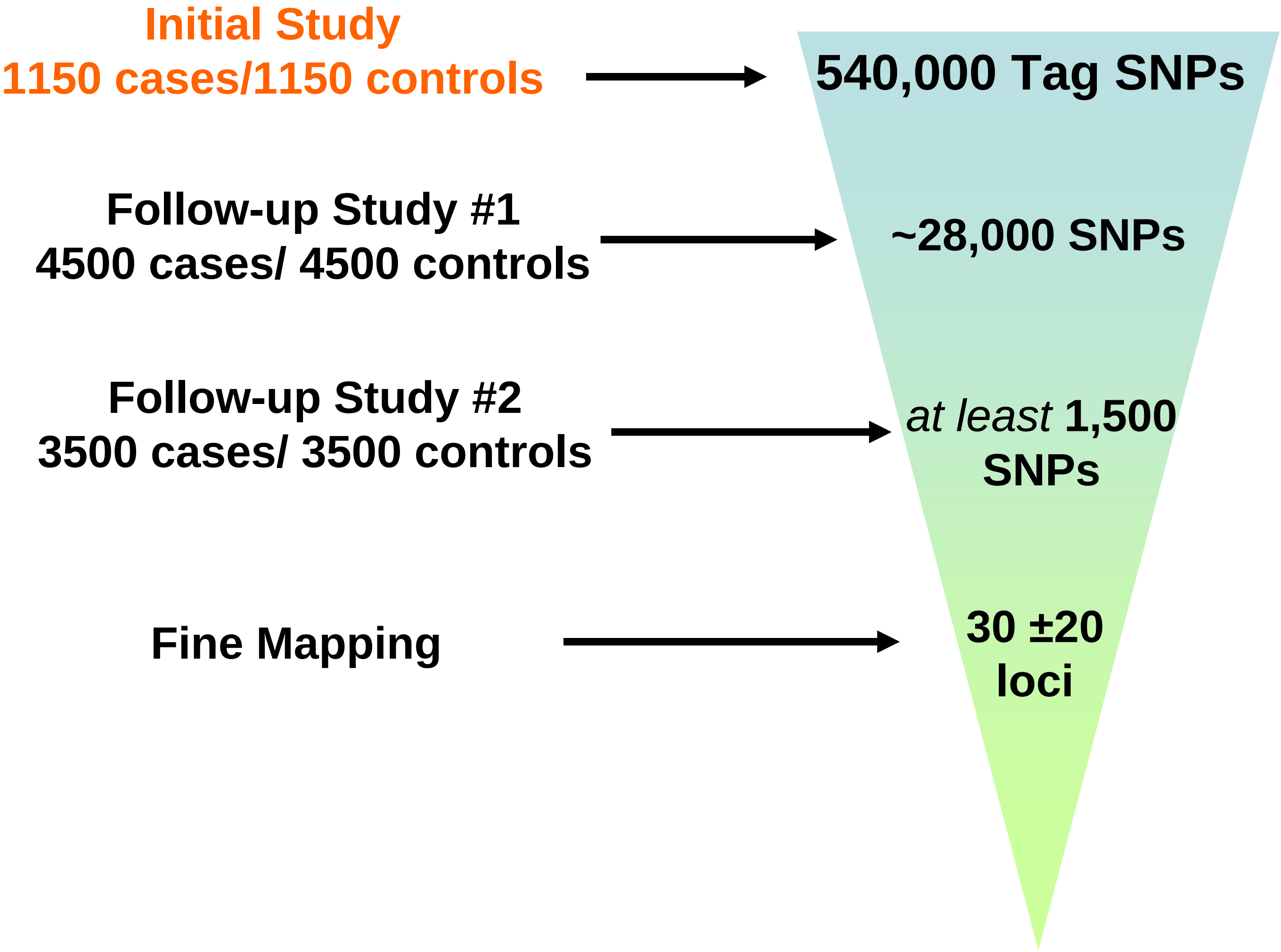
BIAS

Lung Cancer Risk and CYP2D6*

	Study 1	Study 2	Study 3
Relative Risk	**15. 6 (4.8 – 55.9)**	**6.1 (2.2 – 17.1)**	**0.6 (0.3 – 1.2)**
Epidemiologic Quality	**Low**	**Intermediate**	**High**
(% participation)	**(?)**	**(26%)**	**(80%)**

*** Risk of homozygous extensive metabolizers compared to homozygous poor metabolizers.**

General Strategy for Prostate & Breast Cancer GWAS



Initial Study
1150 cases/1150 controls → **540,000 Tag SNPs**

Follow-up Study #1
4500 cases/ 4500 controls → **~28,000 SNPs**

Follow-up Study #2
3500 cases/ 3500 controls → *at least* **1,500 SNPs**

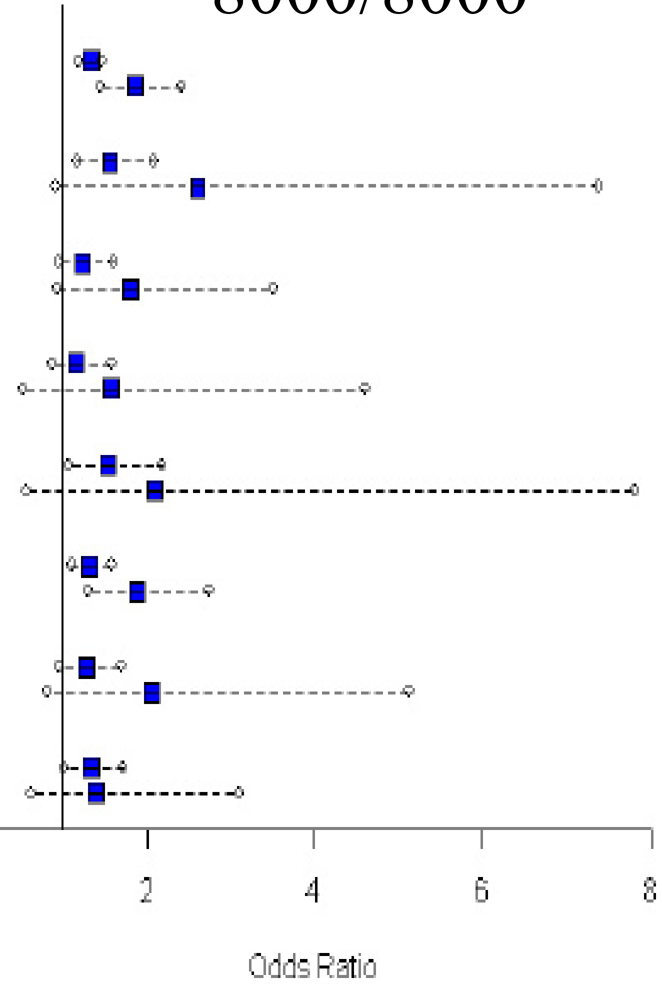
Fine Mapping → **30 ±20 loci**

Results: Overall

BPC3

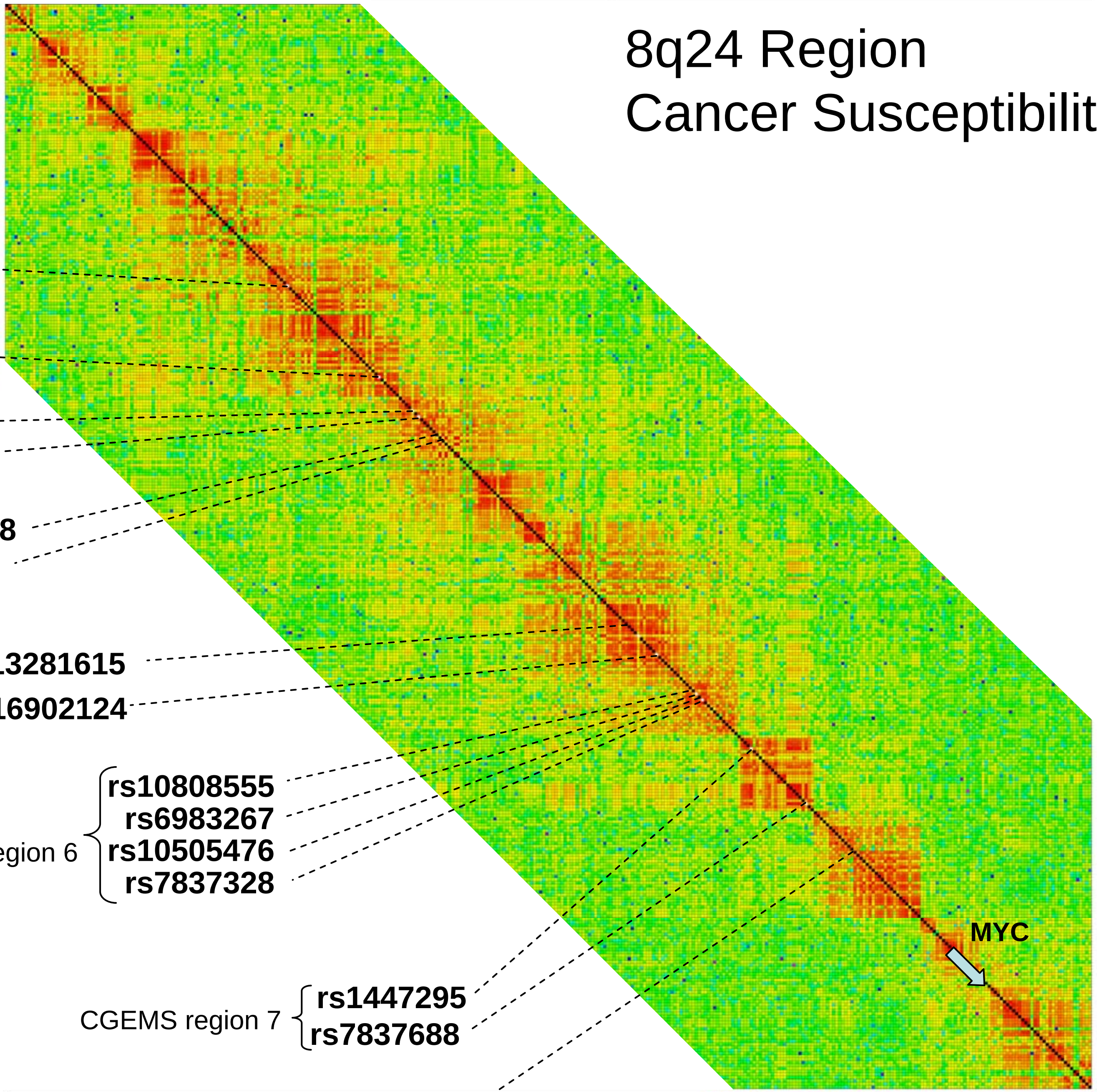
8000/8000

Cohort	Genotype	Cases / Controls	OR (99%CI)	P-value
All	CC	5,566 / 6,666	Ref.	4.00×10^{-19}
(p_{het}=0.483)	AC	2,064 / 1,842	1.33 (1.20-1.46)	
	AA	279 / 175	1.87 (1.44-2.42)	
ACS	CC	871 / 955	Ref.	2.63×10^{-5}
	AC	238 / 166	1.56 (1.17-2.08)	
	AA	21 / 9	2.61 (0.92-7.37)	
ATBC	CC	606 / 623	Ref.	0.012
	AC	312 / 260	1.23 (0.95-1.60)	
	AA	45 / 25	1.81 (0.94-3.51)	
EPIC	CC	551 / 869	Ref.	0.258
	AC	169 / 233	1.17 (0.87-1.58)	
	AA	12 / 12	1.57 (0.53-4.59)	
HPFS	CC	495 / 545	Ref.	3.63×10^{-3}
	AC	157 / 114	1.53 (1.07-2.19)	
	AA	11 / 6	2.09 (0.56-7.80)	
MEC	CC	1,426 / 1,565	Ref.	2.58×10^{-7}
	AC	728 / 614	1.32 (1.11-1.58)	
	AA	146 / 88	1.89 (1.30-2.75)	
PHS	CC	801 / 1,123	Ref.	0.013
	AC	200 / 220	1.27 (0.96-1.69)	
	AA	21 / 15	2.06 (0.83-5.12)	
PLCO	CC	816 / 986	Ref.	0.014
	AC	260 / 235	1.33 (1.02-1.72)	
	AA	23 / 20	1.39 (0.63-3.10)	



Schumacher FR et al., Cancer Res. 2007 Apr 1;67(7):2951-6.

8q24 Region
Cancer Susceptibility



127.6 M

CGEMS region 1 rs979200

CGEMS region 2 rs1456310

CGEMS region 3 rs6470494
rs1016343

CGEMS region 4 rs132544738
rs6983561

CGEMS region 5 rs13281615
rs16902124

CGEMS region 6 rs10808555
rs6983267
rs10505476
rs7837328

CGEMS region 7 rs1447295
rs7837688

CGEMS region 8 **rs7824074**

129.0 M

GWAS: What is Working

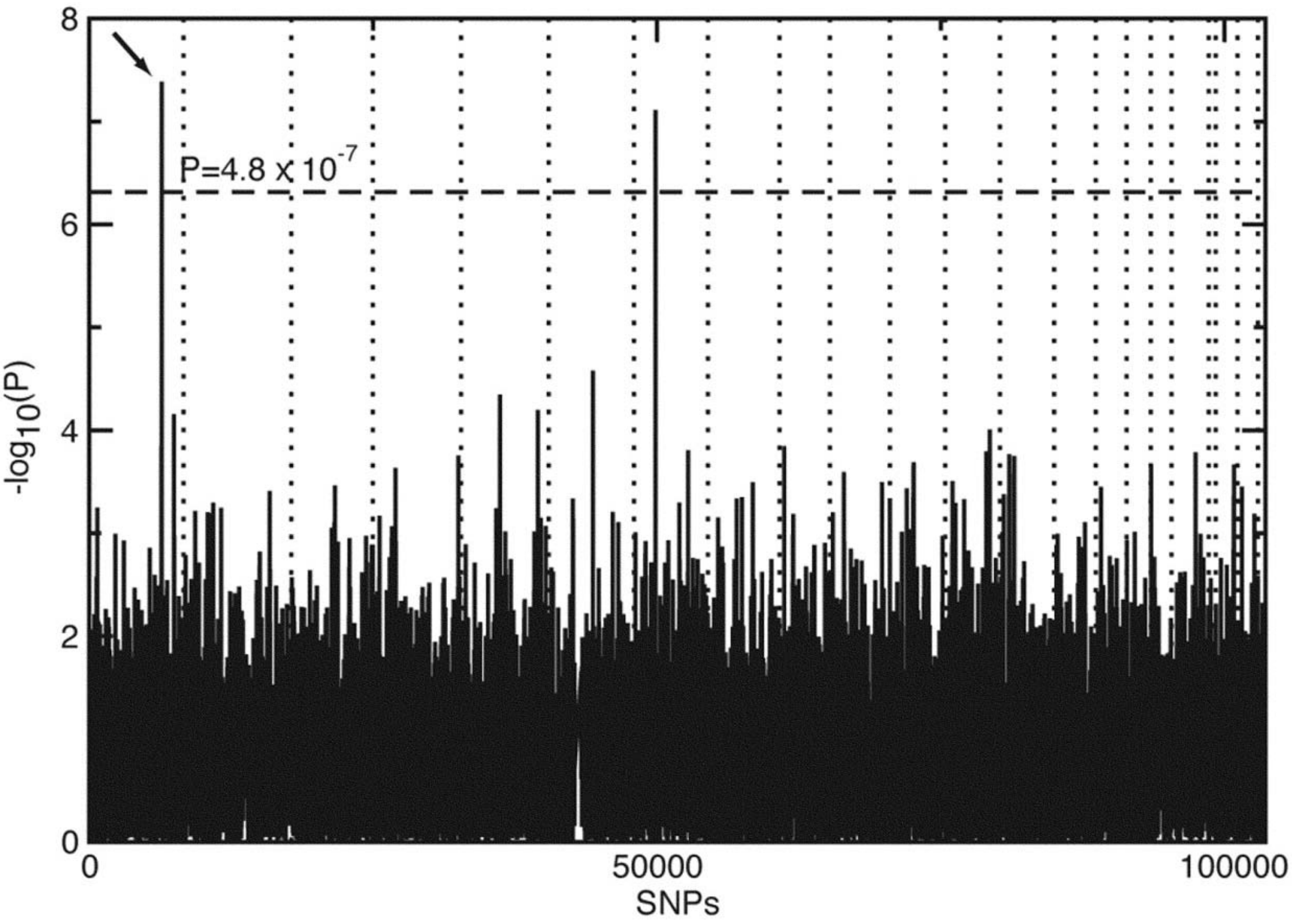
- Very large studies
- Replication, replication, replication (planned and coordinated)
- Rigorous, high-quality design, conduct, analysis
  - Genomics
  - Epidemiology
  - Statistics
  - Informatics
- Data sharing
- Accomplished Through Consortia

COMPROMISES?

- Numbers
  - Initial vs. subsequent stages of scan
  - Replication studies
- Quality
- Examples:
  - AMD
  - Cambridge breast cancer
  - PanScan
- Strategies for what to relax and in what order is complicated

National Cancer Institute

Complement Factor H Gene and Macular Degeneration



Science. 2005 April 15; 308:385

Cambridge University Breast Cancer GWAs

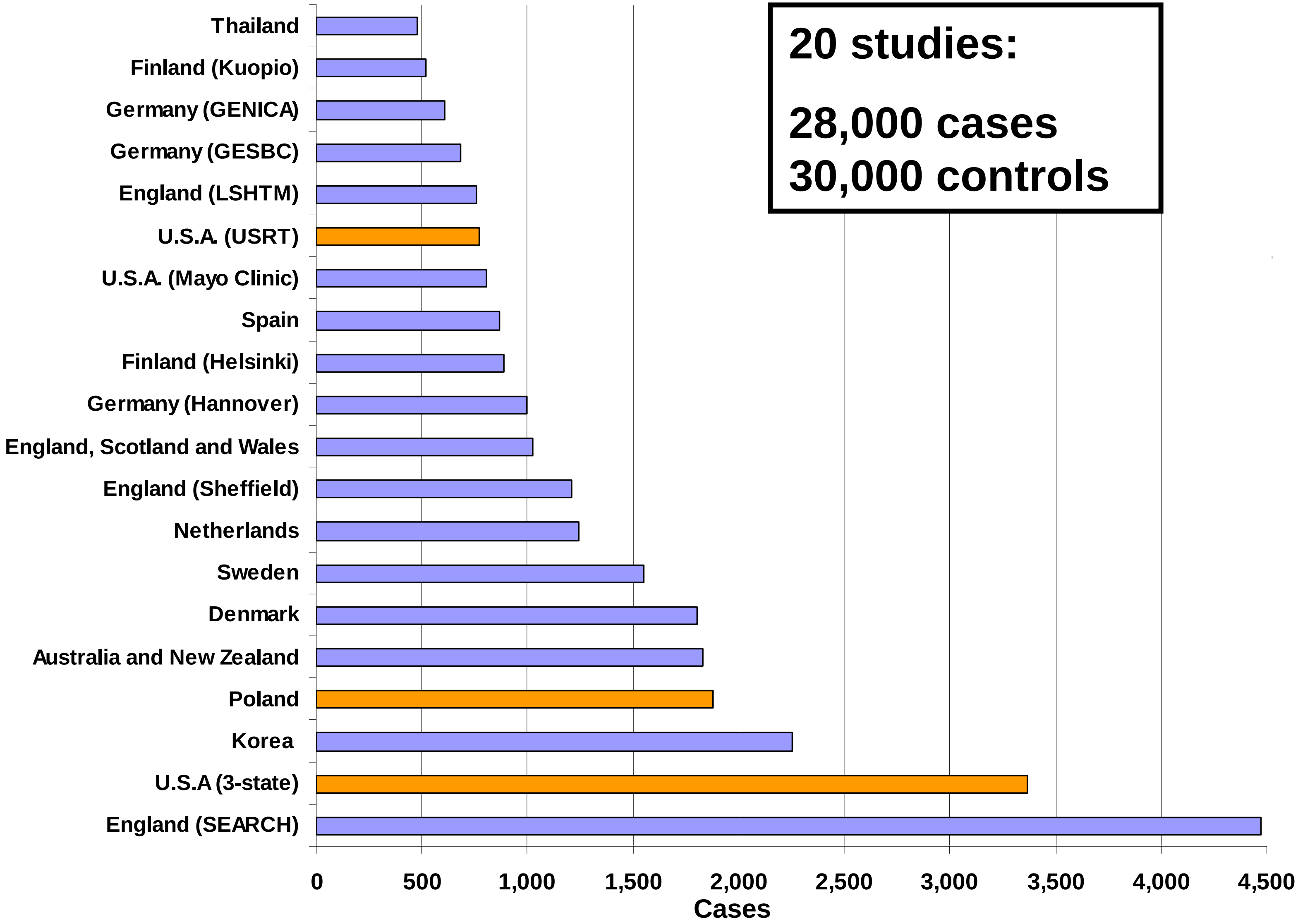
First Stage: 390 cases / 364 controls
267,000 SNPs

Second Stage: 4000 cases / 4000 controls
12,700 SNPs

Third Stage: 22,000 cases / 22,000 controls
30 SNPs

National Cancer Institute

Breast Cancer Association Consortium



20 studies:

28,000 cases
30,000 controls

“In this issue, four investigative teams …have sought to replicate the findings from a GWA study of PD by Maraganore et al. Taken together these four studies appear to provide substantial evidence that none of the SNPs originally featured as potential PD loci are convincingly replicated and that all may be false positives.”

	# of cases	# of SNPs
Tier 1	443	198,000
Tier 2	332	1800

“We identified 11 SNPs that were associated with PD (P<.01) in both tier 1 and tier 2 samples and had the same direction of effect.” (Maraganore et al)

GWAS: What is Working

- Very large studies
- Replication, replication, replication (planned and coordinated)
- Rigorous, high-quality design, conduct, analysis
  - Genomics
  - Epidemiology
  - Statistics
  - Informatics
- Data sharing
- Accomplished Through Consortia

COMPROMISES?

- Yes, BUT

- Strategies for what to relax and in what order is complicated