National Cancer Institute

Probability of Detecting Disease-Associated SNPs in Case-Control Genome-Wide Association Studies

Ruth Pfeiffer, Ph.D.

Mitchell Gail
Biostatistics Branch
Division of Cancer Epidemiology&Genetics
National Cancer Institute, NIH, USA

William Wheeler, David Pee
Information Management Systems, Silver Spring, USA

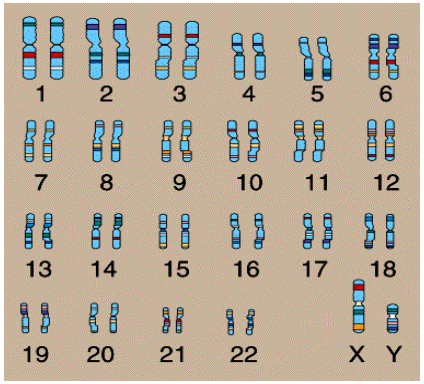
U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES
National Institutes of Health

Outline

- Genetics background
- Case-control Genome-Wide Association Study (GWAS)
- Ranking and selection procedures
- Performance criteria
  - Detection probability
  - Positive proportion
- Analytic results & simulations
- Extnesions: two stage designs
- Conclusions

The Human Genome

- Four DNA bases (nucleotides): A (adenine), T (thymine), G (guanine), C (cytosine)
- 3 billion base pairs
- 22+2 chromosomes
- > 99% of human DNA sequences identical



- ~8,000,000 single nucleotide polymorphims (**SNPs**):

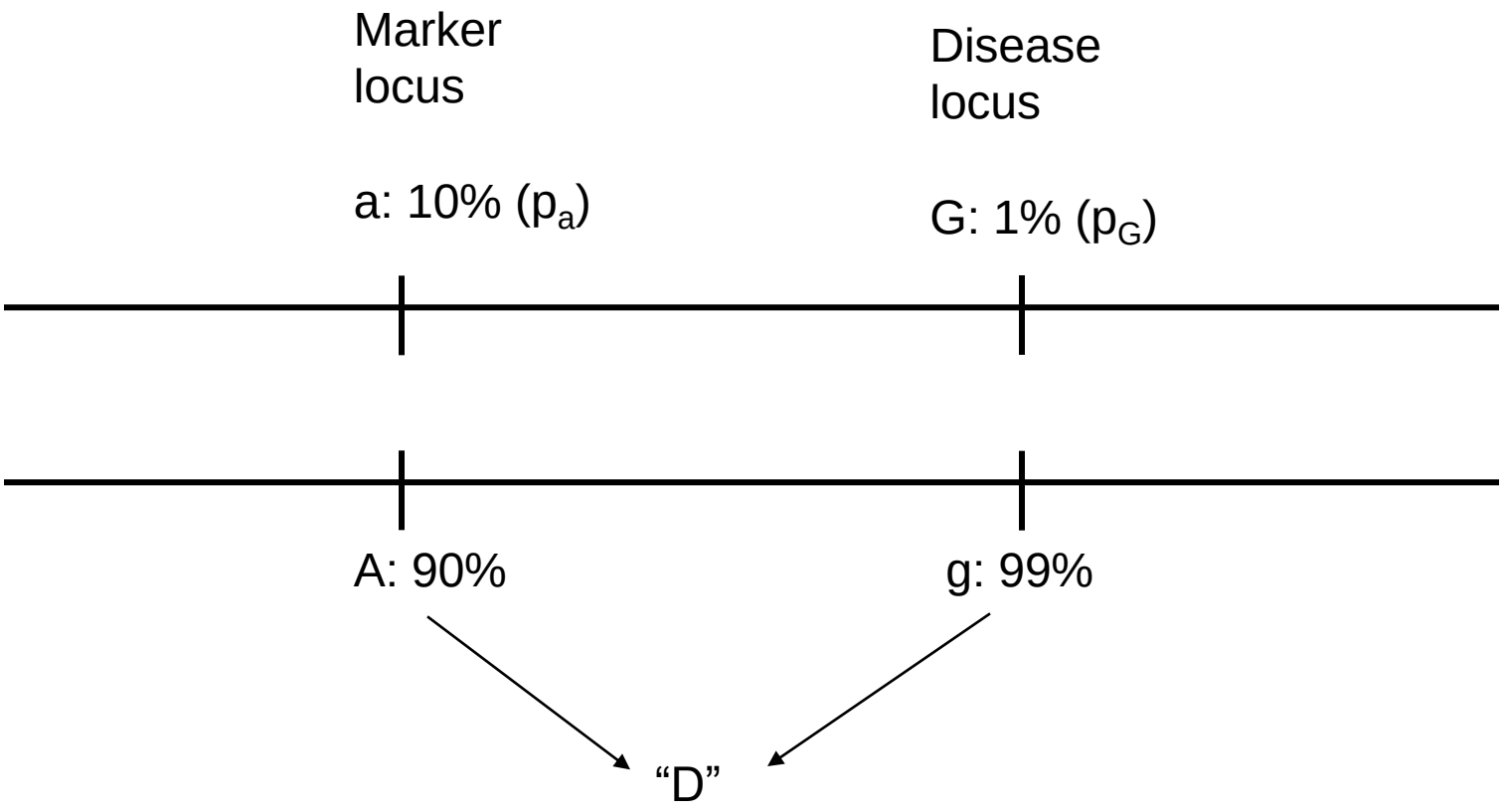
alterations in single nucelotide e.g. A**A**GGC -> A**T**GGC

For variation to be considered a SNP, it must occur in >= 1% of population.
SNPs occur every 100-300 bases, in coding (gene) and noncoding regions

Find disease genes: study variation in SNPs

Linkage Disequilibrium

Genotyping all loci is not possible (not yet!)
=> Utilization of correlation of alleles at two loci



Linkage Disequilibrium

Bi-allelic disease locus: disease allele G (p_G), wild type allele g (p_g); Bi-allelic marker: a, A (p_a, p_A)

Linkage disequilibrium (LD) defined as

$$D = P(A, G) - p_A p_G$$

$$D' = D/D_{max} = D / \min(p_G p_a, p_g p_a) \text{ for } D>0$$

$$r^2 = (D')^2 p_a p_G / p_A p_g$$

D' is upper bound of r^2

Genome wide approaches

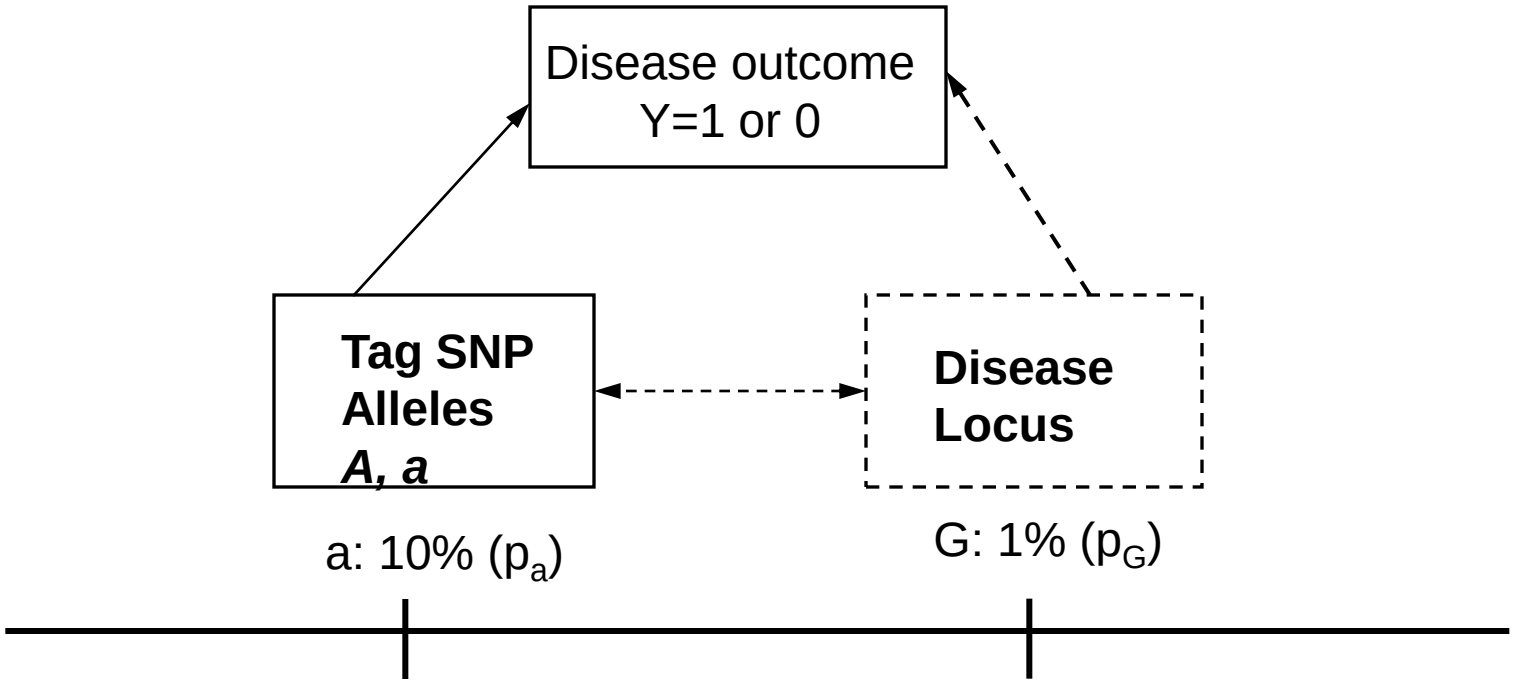
- Try to get closer to disease locus by high density SNP coverage: **tag SNPs**
- **tag SNP:** representative SNP in region of genome with high LD (r^2>0.8) with untyped SNPs in that region (HapMap project)
- Typically: **550,000-650,000** tag SNPs/subject

- *Still estimated that 25% of SNPs not captured adequately by tag SNPs*

Case-control Genome Wide Association Studies

- Retrospective sample of unrelated cases (diseased) & controls (non-diseased)
- Exposure: tagSNPs covering genome
- Assess correlation between SNP genotypes and disease status
- Common SNPs (minor allele frequencies>5%)
- Complex diseases (e.g. cancer)
- Small/moderate genetic effects (OR~1.3)

Genetic Association



Observed data for cases and controls

	SNP i Genotype			
	aa	**aA**	**AA**	**total**
Cases	r_0	r_1	r_2	R
Controls	s_0	s_1	s_2	S
Total counts	n_0	n_1	n_2	N

Assumed Penetrance Model

$$Y=\begin{cases}1, \text{ diseased individual}\\ 0, \text{ healthy individual}\end{cases}$$

$$P(Y=1|\ X_i)=\frac{\exp[\mu+\beta X_i]}{1+\exp[\mu+\beta X_i]}$$

X_i is a score attached to genotype of SNP i

additive model: X_i=# of A alleles (0,1,2)

dominant model: X_i=1 if {aA, AA}; =0 if {aa}

recessive model: X_i=1 if {AA}; =0 if {aa,aA}

Tests for Association based on tag SNP

Test H_0: $\beta = 0$ using

Wald Test : $W_i = \hat{\beta}_i^2 / \hat{V}ar(\hat{\beta}_i)$

Score Test : $S_i = U_i^2 / \hat{V}ar_0(U_i)$

where $U_i = \phi \sum_{cases}^{n} X_i - (1-\phi) \sum_{controls}^{n} X_i$

Remark on the null hypothesis

$H_{0:}$ ”no association between SNP i and disease”

is true if either one of the following

1. Disease has no genetic component
2. D=0 (true disease locus not in LD with observed SNP)

Factors causing false positive associations

- Population Stratification
  - Confounding by ethnicity (bias)
- Cryptic relatedness
  - Variance distortion
  - Genomic control methods (Devlin & Roeder, 1999)
- Differential genotyping error (Clayton et al, Nature Genetics, 2005)
  - Failure to call genotype not independent of case-control status
- **Multiple comparison: test 550,000 hypotheses**

Controlling for Multiplicity

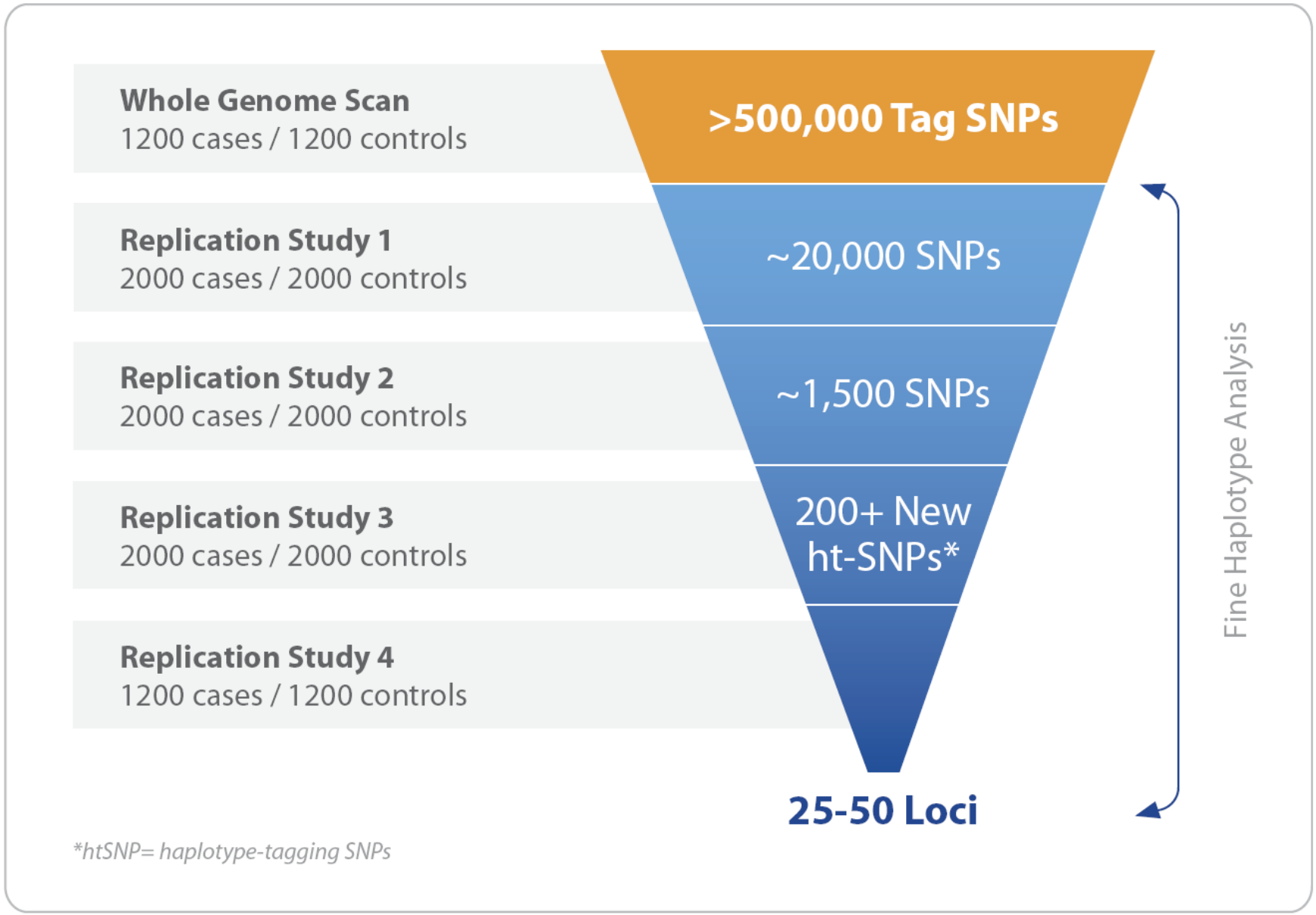
- **Control experiment-wise Type I error**
  - **Bonferroni correction: $\alpha/N=10^{-7}$**
  - **Refinements (e.g. Simes Test)**
  - **Adaptations for two-phase designs** (Skol et al., 2006)
- **Control false discovery rate (FDR)** (Benjamini and Hochberg, 1995)
- **Ranking and Selection**

Motivating Example: Cancer Genetic Markers of Susceptibility (CGEMS)

- **http://cgems.cancer.gov/**
- A three-year, $14 million initiative
- **Outcome:** Prostate cancer (finished), breast cancer (ongoing)
- Combine data from 5 large studies
- **550K chip**: measures N=550,000 tag SNPs

Yeager M, et al. *Nature Genetics,* 2007.

Replication Strategy for Prostate Cancer



*htSNP= haplotype-tagging SNPs

GWAS Designs

- **Single stage design**: all markers measured on all samples

- **Two stage design:**

  Stage 1: Proportion of available samples genotyped on large number of markers

  Stage 2: Proportion of these markers are followed up by genotyping them on remaining samples

Ranking and Selection for Single Stage Designs

- Sample n unrelated **cases (Y=1)** and n **controls (Y=0)**
- For each subject, measure X_i=0,1,2, the number of minor alleles at SNP i for i=1,. . N.

Models for Genetic Effects in Source Population

X_i=0,1,2, - number of minor alleles at ith SNP

SNPs 1,2, . . ., M disease-associated
SNPs M+1, . . . ,N non-disease associated

Probability of disease in source population given by

$$P(Y=1|X_1,...,X_N)=\frac{\exp(\mu+\sum_{i=1}^{M}\beta_i X_i)}{1+\exp(\mu+\sum_{i=1}^{M}\beta_i X_i)}$$

Models for the genetic effects

- **Fixed Effects Model:**

  $\beta_i = \beta$ for i=1,2, . . ., M

  $\beta_i = 0$ for i=M+1, . . ., N

- **Random Effects Model:**

  $\beta_i \sim N(0,\tau^2)$ for i=1,2, . . ., M

  $\beta_i = 0$ for i=M+1, . . ., N

  $E|\beta_i| = \tau(2/\pi)^{1/2} \approx 0.798\tau$

Chi-Square Statistics for Detecting Disease Association

Wald Test : $W_i = \hat{\beta}_i^2 / \hat{V}ar(\hat{\beta}_i)$

Score Test : $S_i = U_i^2 / \hat{V}ar_0(U_i)$

where $U_i = 0.5(\sum_{cases}^{n} X_i - \sum_{controls}^{n} X_i)$

Use 2-sided test with additive scores, X =0,1,2, both tests unchanged by assignment of “minor” allele.

Definitions for Ranking and Selection in One Stage Design

- **SNP i "T-selected" if W_i in top T values, i.e. rank(W_i) >N-T**
- **Detection probability (DP): probability that a disease SNP will be T-selected**
- **Proportion positive (PP): proportion of true disease SNPs among T selected SNPs**

Show: use of marginal model in cases and controls appropriate

Assumption: tagSNPs independent in source population, i.e. X_1 independent of $X_2,\ldots,X_N$

For rare disease,

$$P(Y=1|\mathbf{X}) \approx \exp(\mu + \sum_{i=1}^{M} \beta_i X_i)$$

For a single disease SNP, say SNP 1,

$$P(Y=1|X_1) \approx \exp(\mu^* + \beta_1 X_1)$$

$$where\ \mu^* = \mu + \log\{E \exp(\sum_{i=2}^{M} \beta_i X_i)\}$$

Marginal model in cases and controls

In case-control population,

$$\text{logit}\{P(Y=1|\,X_1)\} = \mu^{**} + \beta_1 X_1$$

$$\text{where } \mu^{**} = \mu^{*} + \log(\pi_1 / \pi_0)$$

$$\pi_i = P(\text{sampled}\,|\,Y=i),\ i=0,1$$

If tagSNPs independent in source population, then tagSNPs independent in cases and controls

Let $\rho_{ki} = P(X_i = k)$

$g_{ki} \equiv P(X_i = k \mid Y = 0) \approx \rho_{ki}$

$g_{klih} \equiv P(X_i = k, X_h = l \mid Y = 0) \approx \rho_{ki}\rho_{lh}$

$$f_{ki} \equiv P(X_i = k \mid Y = 1) = \rho_{ki} \exp(\beta_i k)\{\sum_{l=0}^{2} \rho_{li} \exp(\beta_i l)\}^{-1}$$

$f_{klih} \equiv P(X_i = k, X_h = l \mid Y = 1) =$

$$\rho_{ki}\rho_{lh} \exp(\beta_i k + \beta_h l)\{\sum_{s_1=0}^{2}\sum_{s_2=0}^{2} \rho_{s_1 i}\rho_{s_2 h} \exp(\beta_i s_1 + \beta_h s_2)\}^{-1} = f_{ki}\, f_{lh}$$

Properties of test statistics W_i

Based on independence of genotypes in cases and controls, we compute expected values of prospective estimating equations and their cross products using the retrospective sampling distributions and show that scores are uncorrelated

Estimating each β_i using separate logistic models yields independent estimates of β_i, thus W_i are independent

Similar results for score test S_i

Analytic Calculation of DP and PP

Special Case: *disease SNPs have same allele frequency and β*
Then W_i ~ G for i=1,…M: non-central chi-square with non–centrality $\beta^2/Var(\beta)$ for fixed effects model
W_i ~ F, central chi-square (1df), i=M=1,….N

$$DP = \int_0^{\infty} [\sum_{m=0}^{\min(M-1,T-1)} \binom{M-1}{m} g(c)G(c)^{M-1-m}\{1-G(c)\}^m \sum_{s=0}^{T-m-1} \binom{N-M}{s} \{1-F(c)\}^s \{F(c)\}^{N-M-s}]dc$$

$$PP=(M/T)(DP)$$

Generalizes to different G_i for various disease SNPs with differing allele frequencies and β_i

Approximation for M=1 Disease SNP

$$DP \approx 1 - G\{F^{-1}(1 - T / N)\}$$

In this case, DP equivalent to power of a test with same non-centrality but with type I error (alpha-level) T/N

Simulations

Brute force simulation not feasible; use analytic results including independence

- For each simulation (ISIM) NSIM=10,000
- For SNPs i=1,2, M obtain fixed or random β_i
- For SNPs i=M+1, . . . N let $\beta_i=0$
- For each SNP:
  - Draw random minor allele frequency (MAF) from distribution of MAFs in CGEMS controls (MAF≥0.05): mean=0.276, median=0.26, IQR: 0.15-0.38
  - Solve for μ^{**} such that P(Y=1)=0.5
  - Compute information matrix from case-control logistic model, using retrospective sampling distributions
  - Compute Var($\hat{\beta}_i$)
  - Sample $\hat{\beta}_i$ from N(β_i, Var($\hat{\beta}_i$))
  - Compute $W_i = \hat{\beta}_i^2 / Var(\hat{\beta}_i)$

Simulated Estimates of DP and PP

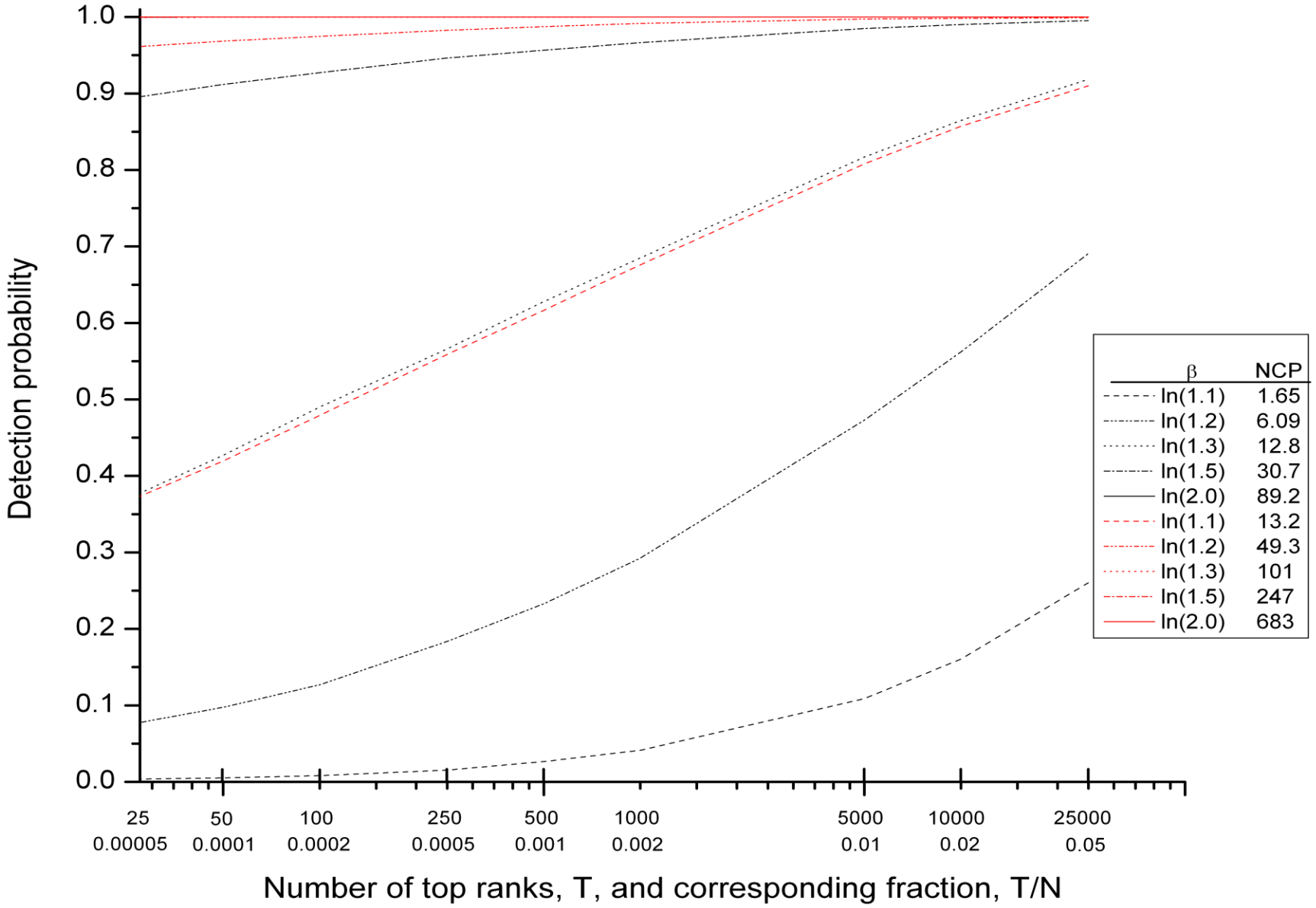
Define *I(m, ISIM,T)=1* if rank(W_m)> N-T , 0 otherwise

$$\hat{\mathrm{DP}} = NSIM^{-1}M^{-1}\sum_{m=1}^{M}\sum_{ISIM=1}^{NSIM} I(m, ISIM, T)$$

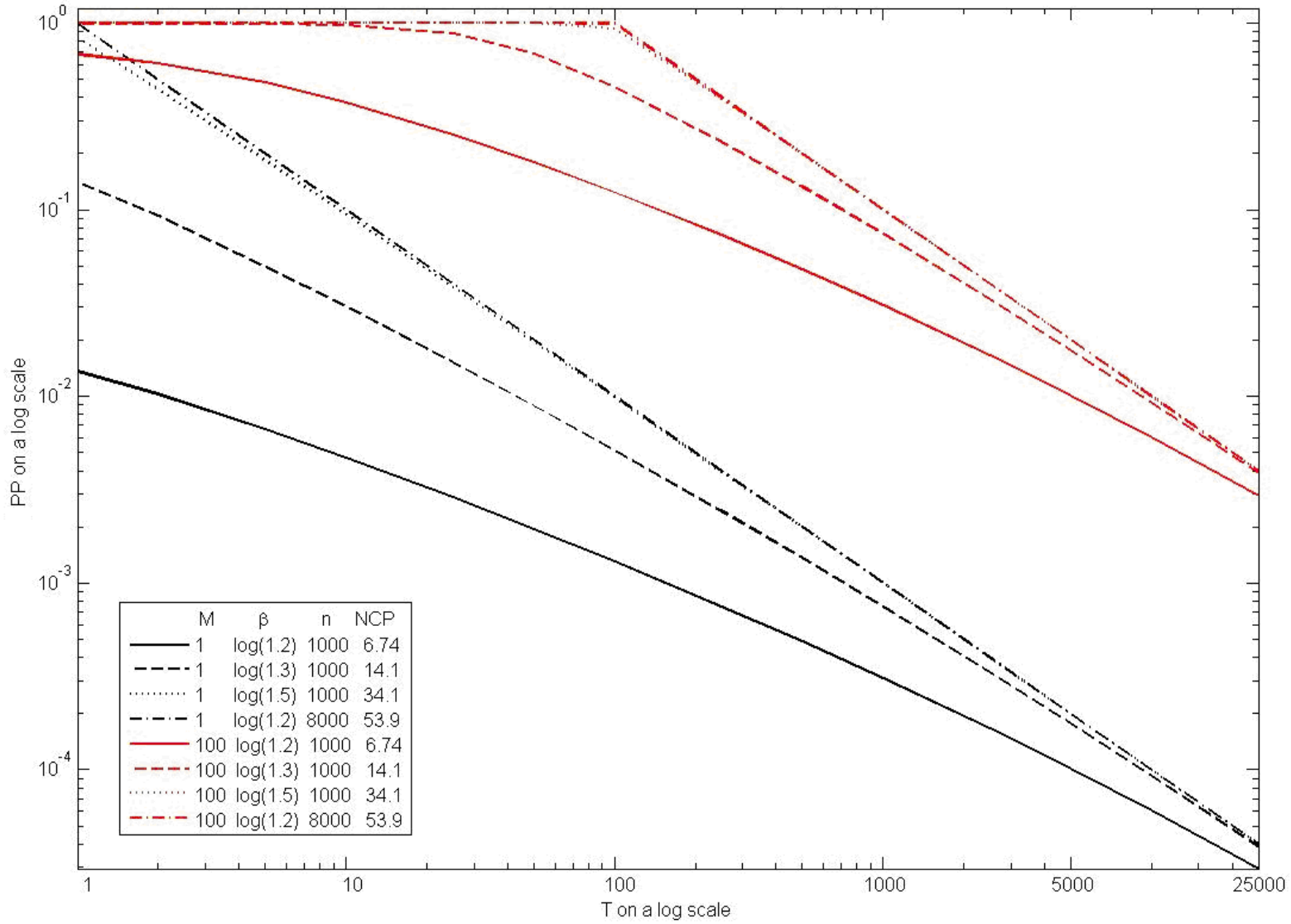
- **Probability a given disease SNP is T-selected**
- **Proportion of disease SNPs selected**

$$\hat{\mathrm{PP}}=(\mathrm{M})(\hat{\mathrm{DP}})/\mathrm{T}$$

Results for Fixed Effects Model

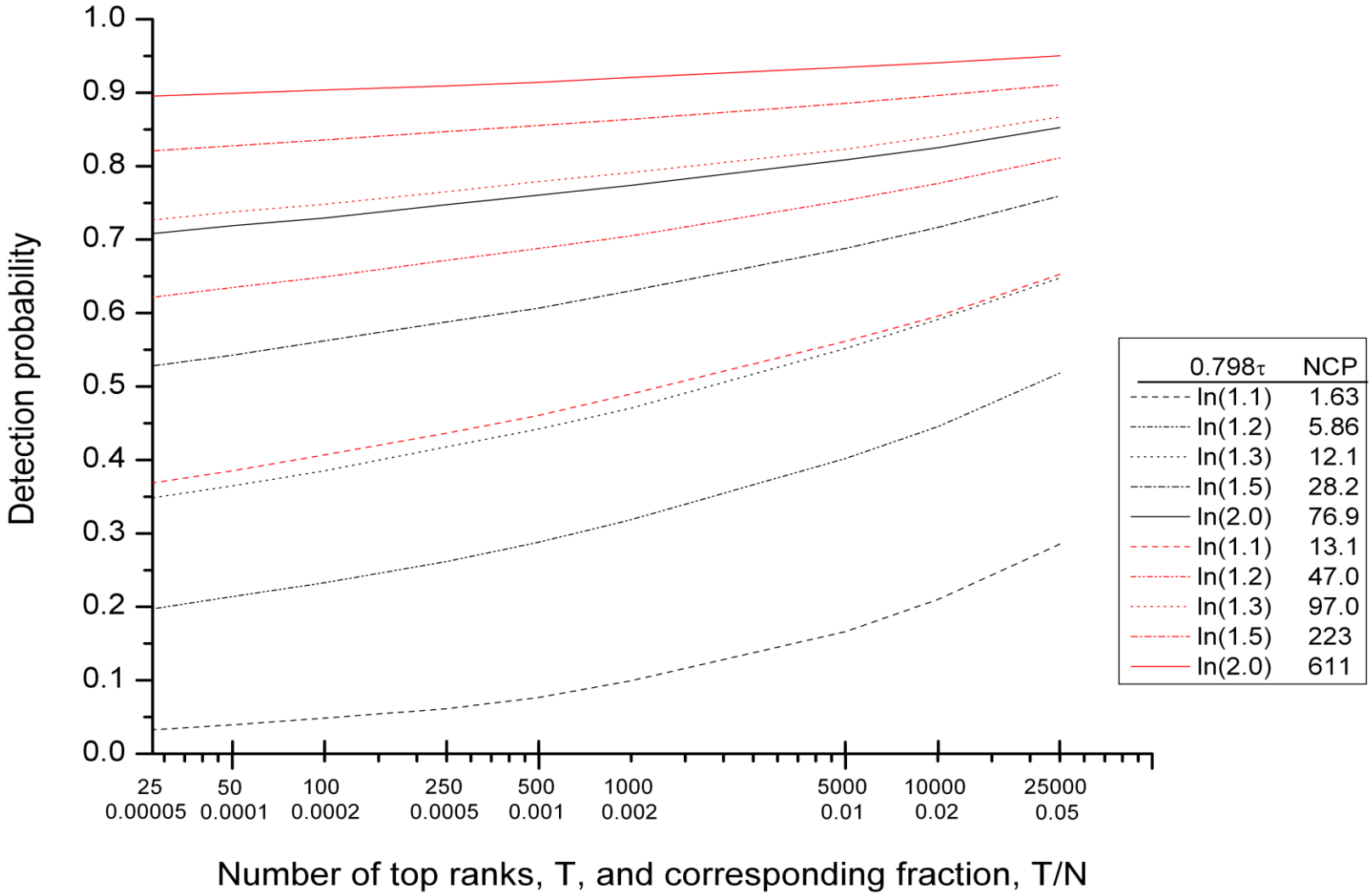


DP for Fixed Effects Model. n=1000 black; n=8000 red. Odds ratios 1.1, 1.2, 1.3, 1.5, 2.0. T on log scale.



PP on log scale versus T on log scale. M=1 black; M=100 red. Odds ratios 1.2, 1.3, 1.5, 2.0.

Results for Random Effects Model



DP for Random Effects Model. n=1000 black; n=8000 red. 0.798τ=log of 1.1, 1.2, 1.3, 1.5, 2.0. T on log scale.

Summary: single stage design

DP and PP are useful criteria for ranking

Related work: Zaykin & Zhivotovsky, 2005; Satagopan et al, 2004

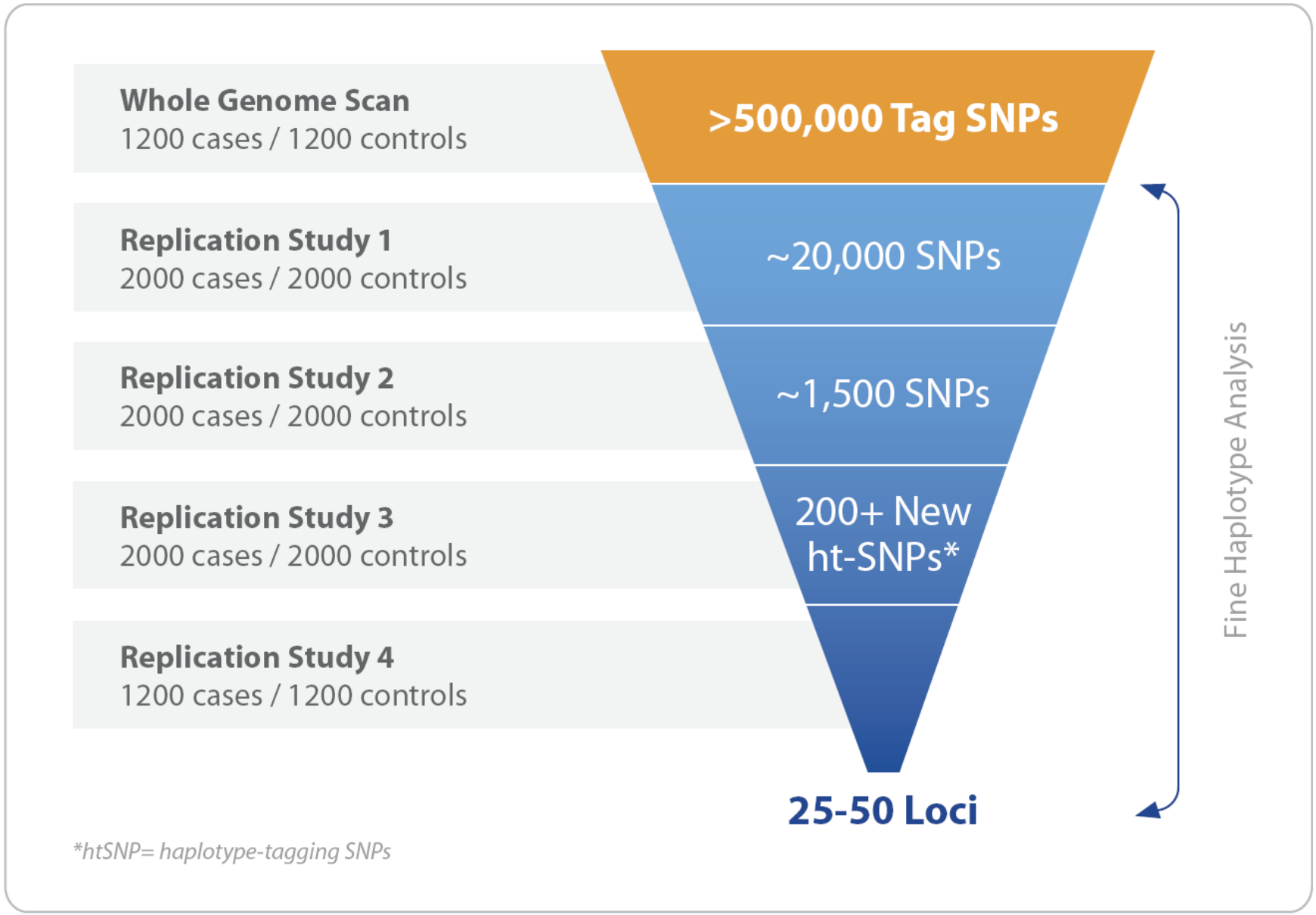
For OR=1.2, large T required to assure adequate DP. T=25,000 yields DP=0.69 (fixed effects model)

DP larger for fixed effects than random effects model

DP is decreased by increasing M for M>T

PP decreases with T for T>M. Increasing T to increase DP decreases PP and is futile if n is too small.

Replication Strategy for Prostate Cancer



*htSNP= haplotype-tagging SNPs

Extension: two stage designs

Stage 1: $\pi_{samples}$ genotyped on 550,000 SNPs

Stage 2: T_1 SNPs largest chi-square statistics followed up by genotyping on remaining samples.

Replication analysis: view stage 2 as a replication study; final selection of the SNPs depended on ranking only the chi-square statistics in stage2

Joint analysis: for each of the T_1 SNPs selected in stage 1, compute

$$\lambda C_1 + (1-\lambda)C_2 , \lambda = 0.0, 0.05, 0.1, \ldots, 1.0$$

C_i chi-square statistic observed in stage i

For each value of λ estimate DP, present maximal DP with corresponding λ.

Skol et al, Nature genetics, 2006

Two stage designs, cont.

Compute probability that exactly M_2 disease SNPs selected after stage 1, given that T_1 SNPs were selected and there were K_1 non-disease SNPs and M_1 disease SNPs

Use joint densities of the (M_1-M_2+1)th and (M_1-M_2)th order statistic of M_1 disease SNPs, and $(K_1-T_1-M_2+1)$th and $(K_1-T_1-M_2)$th order statistic of the K_1 non-disease SNPs.

$W^1_{(i)}$ - order statistics of disease SNPs

$W^0_{(i)}$ - order statistics of non-disase SNPS

Analytic expression

Fixed effects model, same allele frequencies for all disease SNPs

$$P(\text{exactly } M_2 \text{ disease SNPs chosen after stage1}) =$$

$$P(W^1_{(M_1-M_2+1)} > W^0_{(M_1-M_2+1)}; W^0_{(K_1-T_1+M_2+1)} > W^1_{(M_1-M_2)}) =$$

$$\int_{-\infty}^{\infty}\int_{-\infty}^{w1} P(W^1_{(M_1-M_2+1)} > w_2; w_1 > W^1_{(M_1-M_2)}) dG(w_1, w_2) =$$

$$\int_0^{\infty}\int_0^{w1}\int_{w2}^{\infty}\int_0^{\min(x1,w1)} \frac{K_1!}{(K_1-T_1+M_2-1)!(T_1-M_2-1)!}\frac{M_1!}{(M_1-M_2-1)!(M_2-1)!} G(w_2)^{M_1-M_2-1}$$

$$\{1-G(w_1)\}^{M_2-1} g(w_1) g(w_2) \{1-F(x_1)\}^{T_1-M_2-1} F(x_2)^{K_1-T_1+M_2-1} f(x_1) f(x_2) dx_2 dx_1 dw_2 dw_1$$

Special cases

Special case: $M_2=0$

$$P(\text{zero disease SNPs chosen after stage1}) = P(W^0_{(K_1-T_1+1)} > W^1_{(K_1-T_1+1)}) =$$

$$\int_0^{\infty}\int_0^{y} \frac{K_1!}{(K_1-T_1)!(T_1-1)!}\{1-F(y)\}^{T_1-1}F(y)^{K_1-T_1-1}f(y)M_1G(x)^{M_1-1}g(x)dxdy$$

Special cases: $M_2=M_1$

$$P(\text{all disease SNPs chosen after stage1}) = P(W^1_{(1)} > W^0_{(K_1-T_1+M_1)}) =$$

$$\int_0^{\infty}\int_y^{\infty} \frac{K_1!}{(K_1-T_1+M_1-1)!(T_1-M_1)!}\{1-F(y)\}^{K_1-T_1+M_1-1}F(y)^{T_1-M_1}f(y)M_1(1-G(x))^{M_1-1}g(x)dxdy$$

Special case $M_2=T_1$

$$P(\text{all disease SNPs chosen after stage1}) = P(W^1_{(M_1-M_2+1)} > W^0_{(K_1)}) =$$

$$\int_0^{\infty}\int_y^{\infty} K_1F(y)^{K_1-1}f(y)\frac{M_1!}{(M_1-M_2)!(M_2-1)!}G(x)^{M_1-M_2}\{1-G(x)\}^{M_2-1}g(x)dxdy$$

Probability of detecting a disease SNP (DP) and optimal stage1 weight for fixed effects model with log odds ratio per allele β= log(1.2) for 8000 cases and 8000 controls

π_{sample}	**Analysis**	**Number of disease SNPs,** $M_0 = 1$			
		$T_1 = 1000$		$T_1 = 25,000$	
		$T_2 = 1$	$T_2 = 100$	$T_2 = 1$	$T_2 = 100$
0.125	**Replicate**	.266	.269	.630	.664
	Joint	.267	.269	.635	.664
	λ_{opt}	.25	.27	.35	.25
0.25	**Replicate**	.612	.626	.803	.881
	Joint	.613	.626	.825	.885
	λ_{opt}	.32	.15	.40	.25
0.50	**Replicate**	.769	.897	.821	.912
	Joint	.843	.900	.858	.953
	λ_{opt}	.40	.20	.45	.40
1.00	**One-stage**[a]	.882	.966	.882	.966

Summary, two stage design

- A small first stage can only partially be compensated for by selecting a large number of SNPs for further testing
- Even joint analysis does not improve DP much if $\pi_{sample} \leq 0.25$
- As the cost per genotype for stage1 drops, compared to later stages, economic incentives for multi-stage designs decrease
- Advantage of one-stage design: offers unbiased estimates of genetic effects that can be combined easily with those from other studies; multi-stage designs introduce selection biases that complicate meta-analyses.

References

Gail MH, **Pfeiffer RM**, Wheeler W, Pee D, Probability of Detecting Disease-Associated Single Nucleotide Polymorphisms in Case-Control Studies with Whole Genome Scans, in press, Biostatistics.

Pfeiffer R, Gail MH, Genetic Epidemiology, 25 (2), pp 136--48, 2003.
Skol et al. 2006. Nature Genetics, 38:209-13.
Dudridge 2006, Am J of Hum Genetics. 78:1094-95.
Marchini et al. 2006. Nature Genetics, 2005: 413-7.
Freedman et al. 2004. Nature Genetics, 36: 388-395
Marchini et al. 2004. Nature Genetics, 36: 512-517
Campbell et al. 2005. Nature Genetics, 37: 868-872
Devlin and Roeder, 1999. Biometrics, 55: 997-1004
Price et al. 2006. Nature Genetics, 38: 904-909.
Pritchard JK et al. 2000. Genetics, 155: 945-959
Reich and Goldstein, 2001. Genetic Epidemiology, 20: 4-16
Satten GA et al. 2001. Am J. Hum. Genet. 68: 466-477

Platforms & SNP Chips

- Affymetrix 100K
- Affymetrix 500K

} essentially random set of SNPs

- Illumina 317K
- Illumina 550K

} designed using Hapmap

- Illumina 650Y (550K+100K YRI fill in)