Objectives of Symposium

- To identify common, critical issues that have been encountered in applying genomic technologies to population studies at NIH and creative approaches to`solving them;
- To develop approaches for prioritizing and conducting population studies using`genomic technologies for use by individual ICs as desired
- To identify new tools for genomics, categorization of phenotypes, and database standardization required for genome-wide association and sequence-based studies.

Panel 1

- Beena Akolkar — NIDDK
- Stephen Chanock — NCI
- Luigi Ferrucci — NIA
- Daniela Gerhard — NCI
- Eric Green — NHGRI
- Jim Mullikin — NHGRI

Design Field Study	$1,500,000
Conduct Field Study	$2,500,000
DNA Extraction Request	$75,000
Genotyping WGS	$2,500,000
Data Analysis	$200,000
Follow-up Genotype	$1,400,000
Publication	*Priceless….(8.175M)*

Genomics: Different Paths

Wide sweep

Microarray

Looks at all transcripts in one assay

Uses oligo-dT to capture 'transcripts'

Provides snap-shot of genes

Focused analysis

Target each unique region

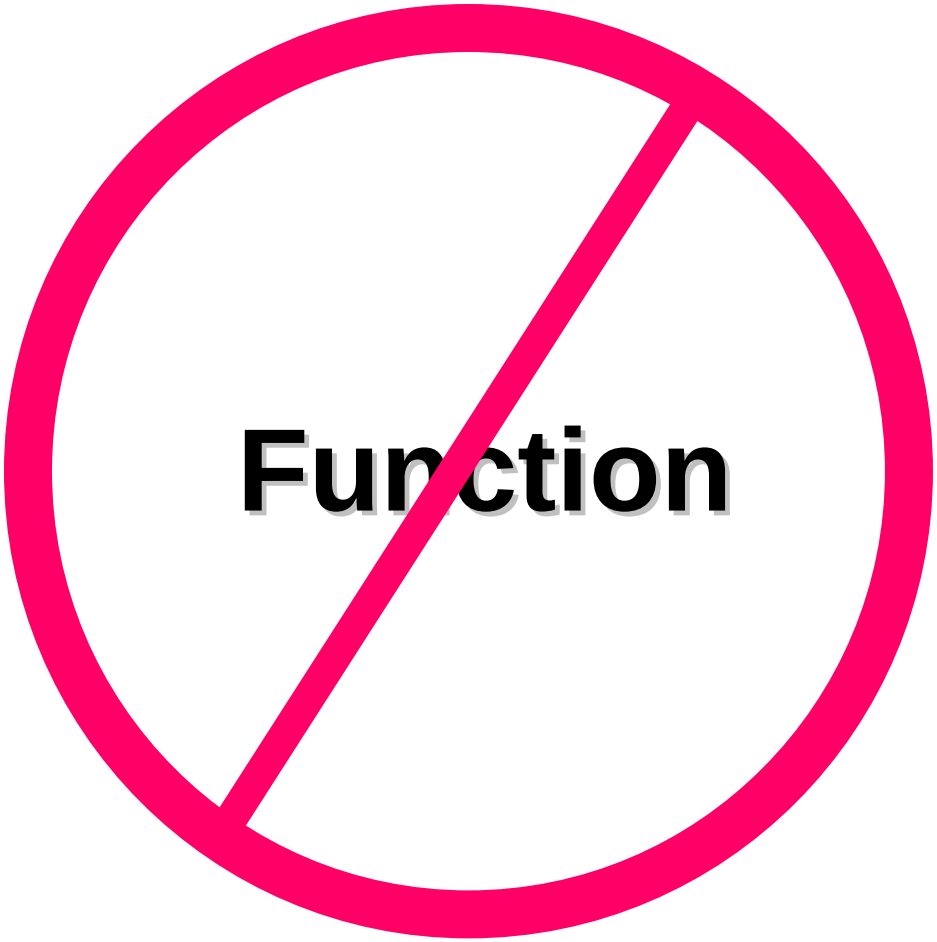
Sequence read (~500 bp)

Genotype (1 key bp)

Requires many assays

Issues in design & analysis

Whole Genome, or Partial Genome Scans Are Designed to Identify Genetic Markers



What Tools Do We Have?

- Extensive data base of common SNPs (MAF>5%)
- Technologies for small to large (1 to 10^6 SNPs)
- Analytical programs for simple analyses
  - Main effect
  - Population structure
- Sequencing technology for ‘targeted regions’

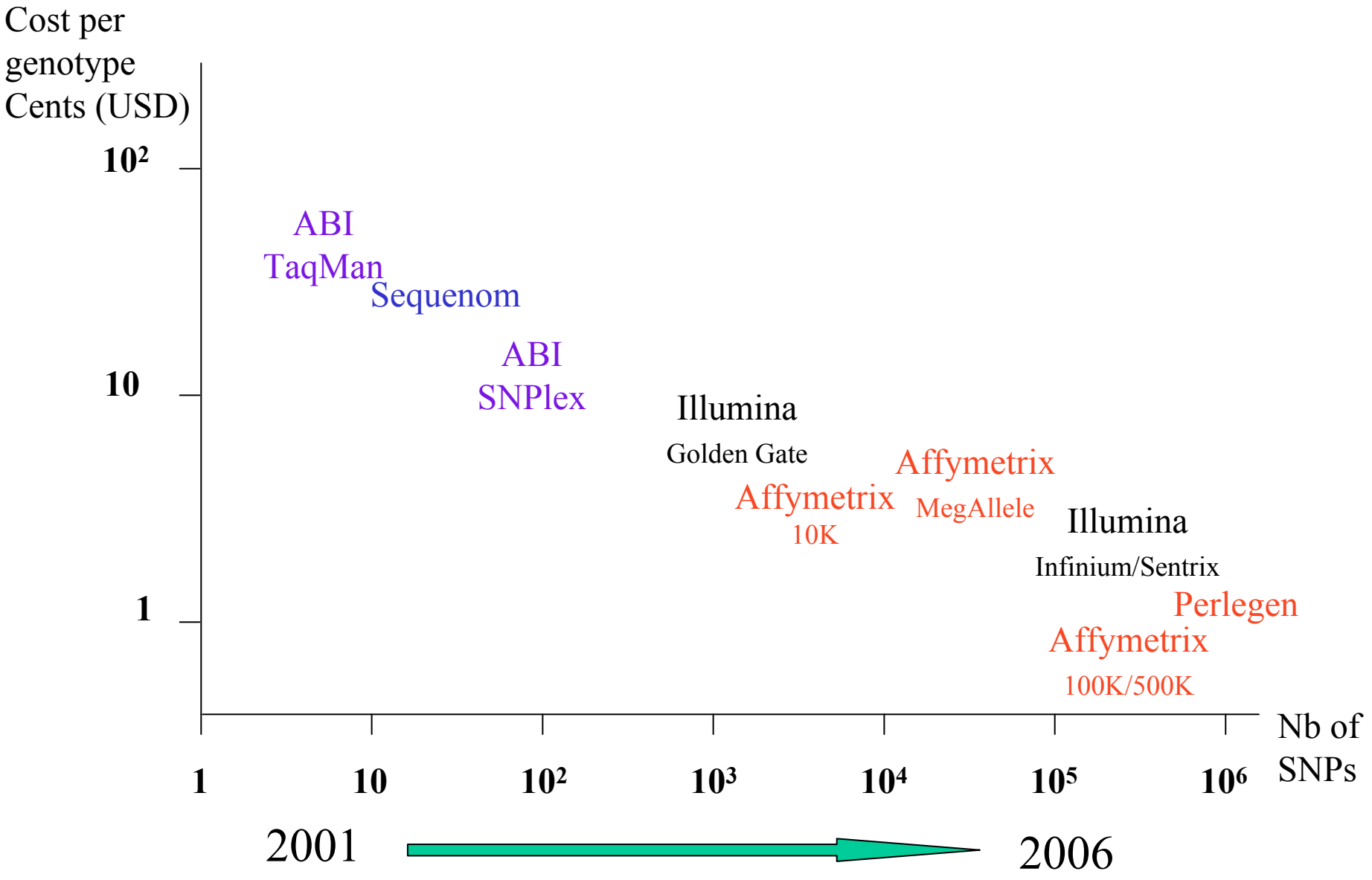
2006

What Tools Do We Need?

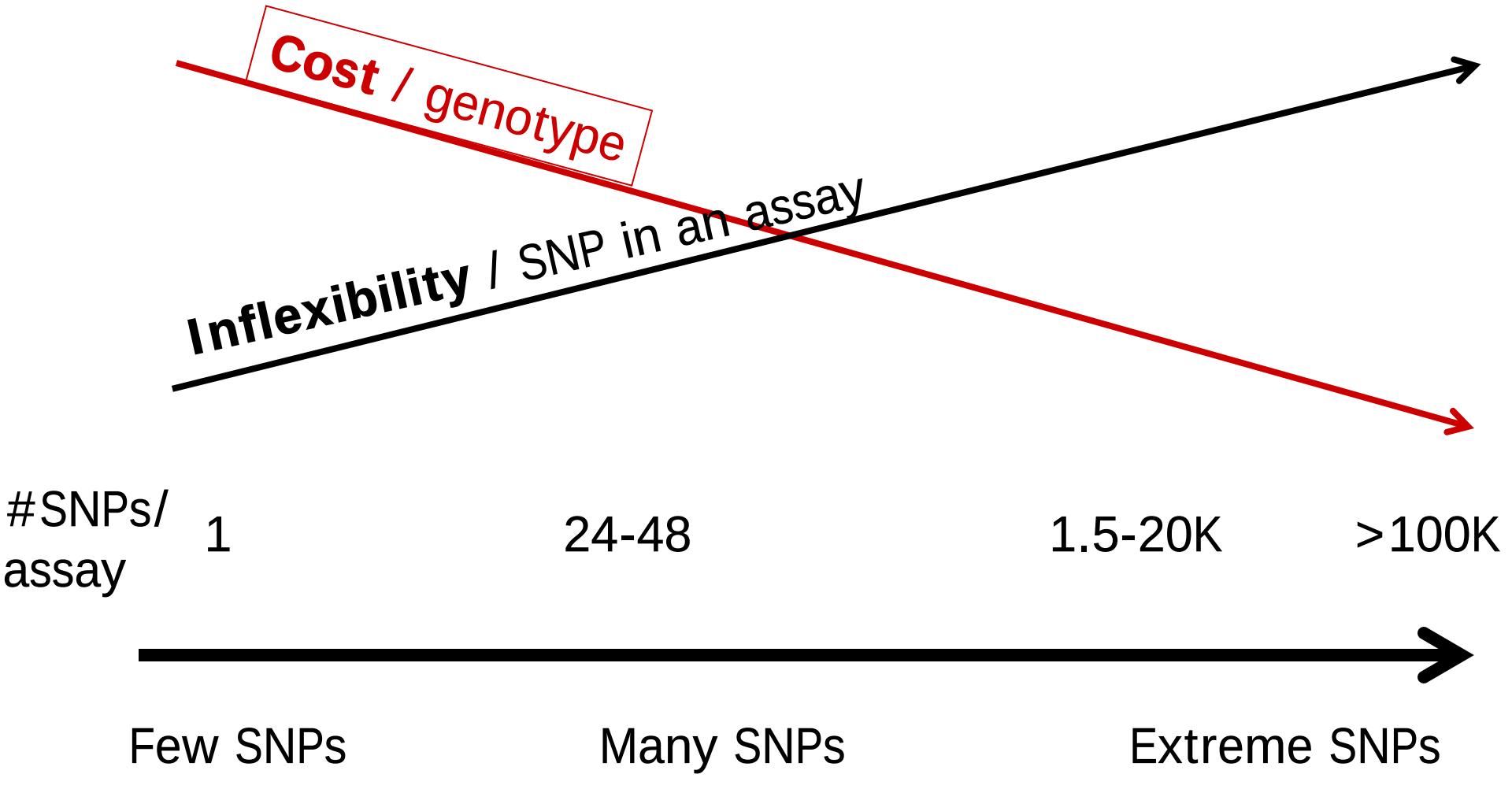
- Extensive data base of uncommon SNPs (MAF<5%)
- *Flexible* Technologies for small to large (1 to 10^6 SNPs)
  - Targeted to different populations
- Analytical programs for complex analyses
  - Gene-gene interaction
- Environmental measurements
- Complete genome sequence technology

Post 2006

Progress in Genotyping Technology



Genotype Opportunities



2006 What is Available for Whole Genome SNP Scans

Coverage analysis based on HapMap II Data

Build 20 MAF $\geq$5%, $r^2 \geq 0.8$ (pair-wise)

		CEU	YRI	JPT/CHB
Illumina	**HumanHap300**	**80%**	**35%**	**40%**
Illumina	**HumanHap500**	**91%**	**58%**	**88%**
Affymetrix*	**500k Mapping**	**63*%**	**41%**	**63%**
Perlegen	**"Custom Choice"**	**"Set by amount paid…."**		

***77% (with 50k MegA)**

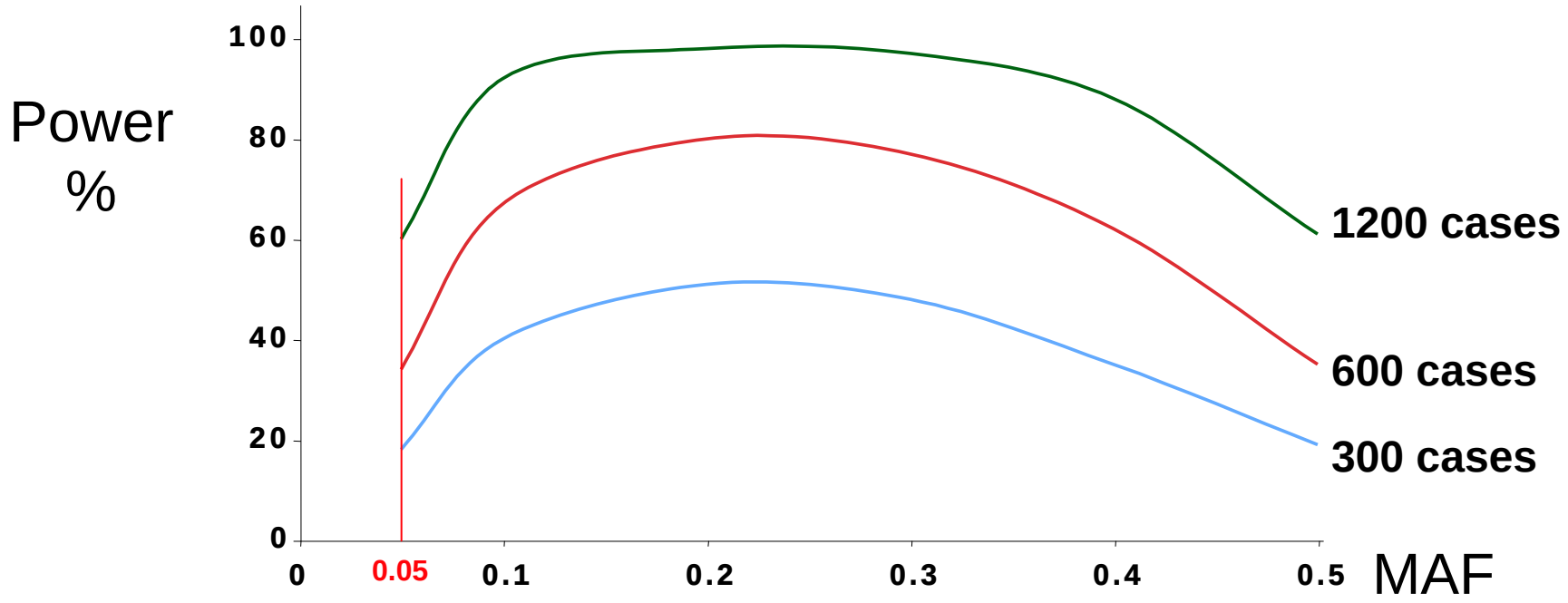
Quantums of Genotype Cost

Scope	Cost/SNP	Total
Singleplex	$0.25	$0.25
Multiplex (6-48)	$0.10	$5.00
Maxiplex (1500)	$0.04	$60.00
Super-plex (24,000)	$0.01	$250.00
Extreme-plex (>10^5)	$0.0013	$750.00

Central point: *Think cost per sample*

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

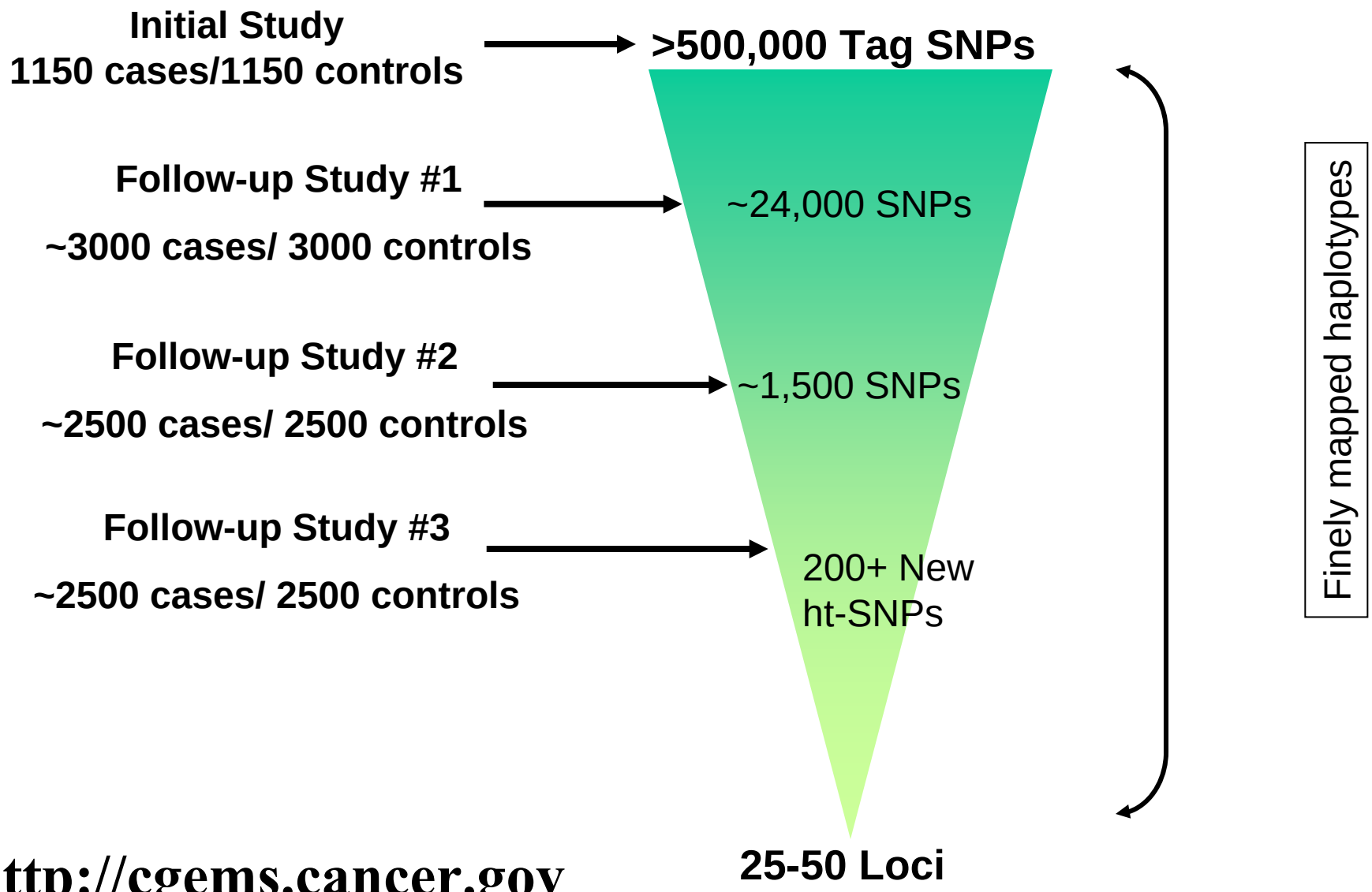
Study design

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

Skol Nat Genet 2006

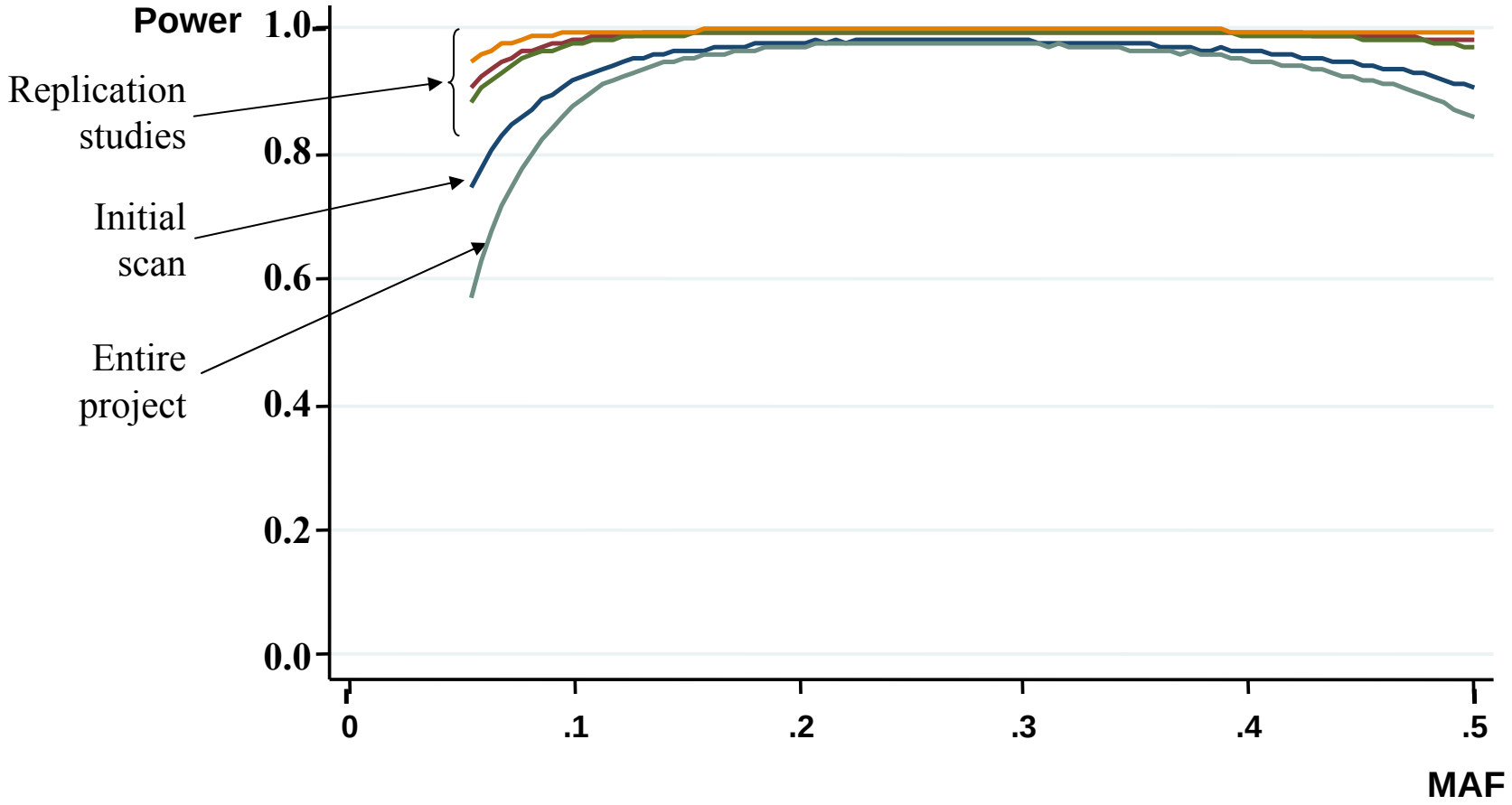
Replication Strategy for Prostate Cancer in CGEMS



http://cgems.cancer.gov

CGEMS: Detection Probability for 3 Stage Model

Dominant, odds ratio 1.5 ; r2 = 0.8 with the functional SNP

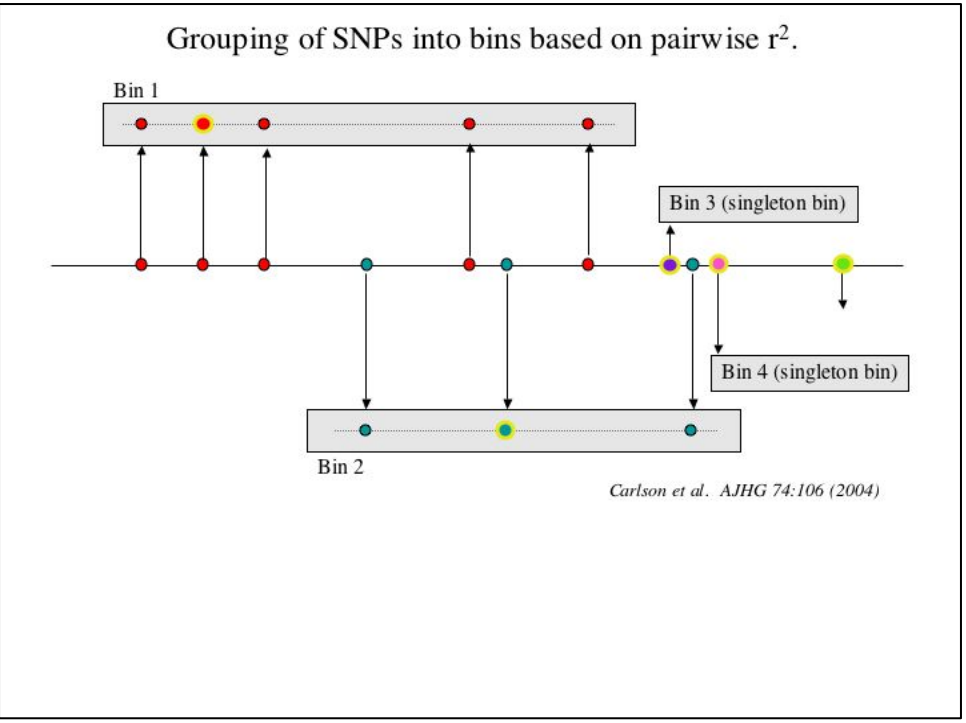


Scan in 1200 cases and 1200 controls

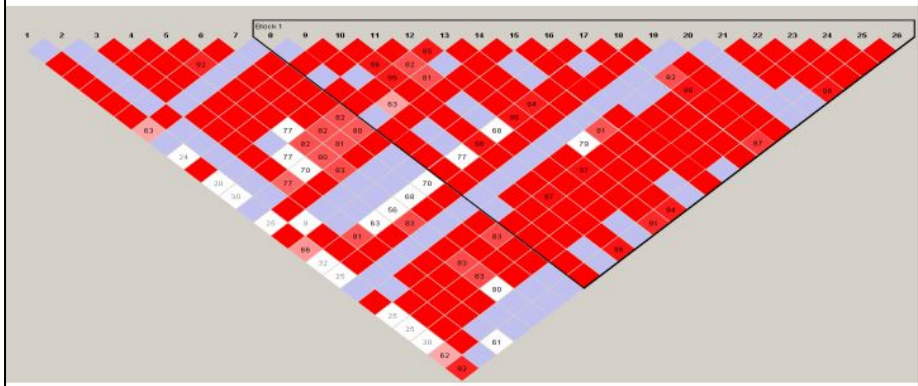
Validation in 3 studies each 2000/2000

Strategy for SNP Selection for Whole Genome Studies in Prostate Cancer

- To test all SNPs is presently too costly
- Utilize a strategy that capitalizes on linkage disequilibrium between SNPs



Haplotype blocks defined by Gabriel et al
Based on D' values for linkage disequilibrium



A quick note on ‘ideal’ power

- r^2 represents the statistical correlation between two loci
- 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 approximately $1/r^2$

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%

Justification of Cost
Based on what you are looking for

Size of Effect

- **Odds ratio 1.3 -> 2.5**

Sufficiently high allele frequency

Population attributable risk

True Negative

- ***Alternatively, tells you to look no more…***

Issues in Extreme Genotyping

- Assay optimization
  Errors in mapping, design & primers
- Software calling algorithm ‘in silico faith’
  - Reliance on programs
  - Impossible to check 800,000,000 genotypes
- DNA Source (blood, buccal, other)
  Quantity
  Quality
  Whole genome amplified- (aka previously WGA)
    Results in LOH
    97-98% Representation

Issues with Pooling Studies

- Accuracy
  - DNA quantification- Haque BMC Biotech 2003, 3:20.
  - Restriction of additional analyses
    - Pools defined by case/control
  - False negatives
  - False positives
    - ? Increase by what proportion
- Substantial cost savings

Current Conundrums of WGS

- Marker Selection
  - Representation of variation across genome
    - Blocks, bins and tags…..
    - Effect of Copy Number Variation (CNV)
- Number of scans per disease
  - Disease and Sub-type
  - Distinct populations
  - Survival
  - Pharmacogenomics
- Population genetic issues
  - Stratification
  - Admixed populations

What Do We Look For In New Technologies?

- Inflexion points: Cost shifts
- Flexibility of technology
  - Cosmopolitan target set
  - Tailor to study population (prior knowledge of structure)
- Efficient use of DNA
- Accurate software for data management and analysis

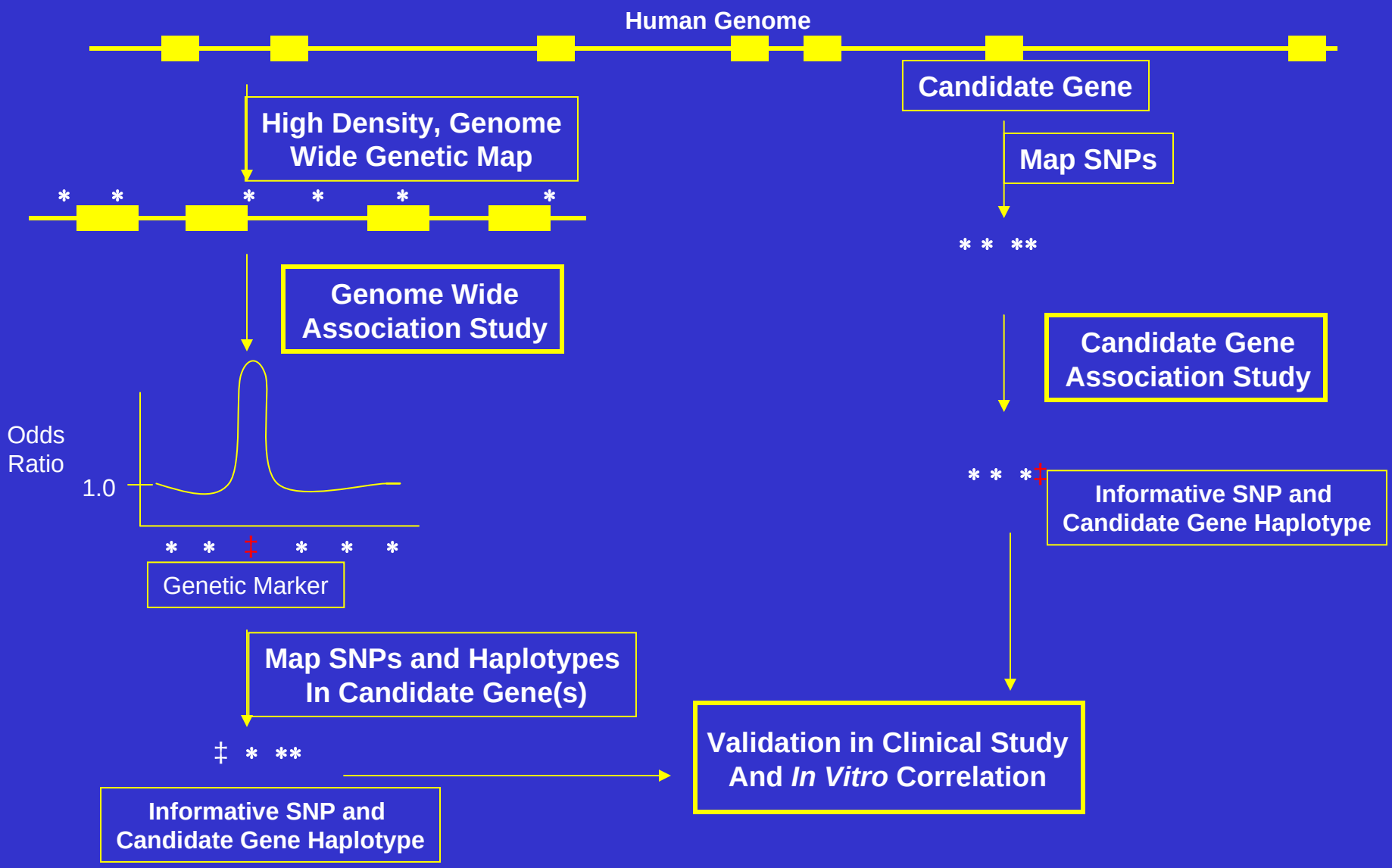
Central Issues: Panel 1

1. Current standards for genotyping technology: data completeness and reproducibility, genomic coverage, comparability across platforms, turnaround time, cost
2. Current standards for sequencing technology: data completeness and reproducibility, comparability across platforms, turnaround time, cost
3. Adopting new technologies
4. Proposals for continued sharing of experience NIH-wide
5. IP Issues and their impact on scientific decisions

Value Added Analysis in CGEMS

- Opportunity to investigate
  - Gene:environment
    - Covariates: BMI, smoking, serum levels
  - Gene:gene interactions
    - Explore pathways
  - Follow-up in cohort studies in CGEMS

Parallel Approaches To Identifying Genetic Determinants of Disease



Whole Genome Scans:SNPs

Illumina

- tagSNPs based on HapMapII
- 2 parts (317k + 240k)
- New 1 chip (540k)

Affymetrix

- Designed pre-HapMapII
- Spaced 500k markers
- Genic enrichment
- Redundancy
  - Useful
- 'Enrich' with Megallele
  - 3K *(90% Smith AJHG)*
  - 100k

Sequence Analysis

- Germ-line
  - Susceptibility/outcome
- Somatic analysis
  - Cancer
- Comparative analysis
  - Molecular evolution
  - Insight into sequences of signficance

Shift in Sequence Technology

Target Amplicons
Small to Large
Diagnostic to Genome

Assess:
Unique Regions of

Genome
Annotate variation

Highly Parallel
"Whole Genome"

Assess:
Complete genome

Assembly
Computationally Challenging

Issues in Sequence Analysis

Rare Variants	
Family Studies	Are There Enough?
Functional Analysis	Very Slow!
Annotation issues	Database?
Population-specific issues	Database?
Comparison with altered tissue	
Duplicate effort	Parallel analysis
Copy Number Variation	
Annotation issues	Database?

Future Issues

- Proteomics
- Epigenomics
- Metabolomics

Search for Genetic Contribution to Complex Diseases

- **Well positioned for**

  Common SNPs (>5%)

  High throughput technology
- **Not as well positioned for**

  Uncommon variants

  Structural variants (copy number variants)

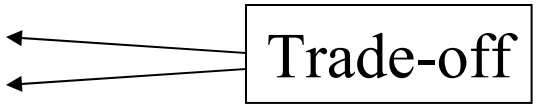
  Populations not in the “BIG 3”
    - CEU, Yoruba, East Asia

Whole Genome Scans (WGS=WGA)

- Public Health Impact
- Specific Aim(s)
  - Etiology
  - Survival
  - Pharmacogenomics
- Value-added Analyses
  - Co-variates
  - Biomarkers
  - Gene-environment interactions

Considerations in Whole Genome Scans

- Extent of Coverage of Genome
- Primary Scan
  - Adequate Size
  - Expected measured effect
  - Ascertainment of Population Structure
  - Study Design
    - Single study vs combined (heterogeneity)
- Replication Strategy
  - Power calculations for how many stages
  - Joint vs consecutive analysis *(Skol Nat Genet 2006)*
  - Design
    - Prospective vs. Retrospective



(www.hapmap.org)

- **Goal: To construct a haplotype map across the entire genome**
  - **270 individuals (Nigerians, Japanese, Chinese and whites)**

- **Phase 1: completed 03/01/2005**
  - **1,000,000 common SNPs (≥ 5%) genotyped: 1 per ~5 kb**
- **Phase 2: completed 10/28/05**
  - **~4,000,000 common SNPs (>5%) genotyped: 1 per ~1.5 kb**

- **A few hundred thousand SNPs will be needed to capture common variation across the entire genome (2005-2006)**
  - **A framework for comprehensive candidate gene and genome-wide association studies**
  - **Between 500,000 and 1.000,000**

http://cgems.cancer.gov

National Cancer Institute | U.S. National Institutes of Health | www.cancer.gov

CGEMS

Cancer Genetic Markers of Susceptibility

Contact Us | Site Map | Search

Core Genotyping Facility | Office of Cancer Genomics | Division of Cancer Epidemiology and Genetics

About CGEMS | News & Announcements | Resource Room | CGEMS Intranet

Cancer Genetic Markers of Susceptibility Project

The Cancer Genetic Markers of Susceptibility (CGEMS) is a three-year, $14 million initiative that will identify genetic alterations that make people susceptible to prostate and breast cancer. Scientists involved will use DNA available from five large studies of prostate cancer and five large studies of breast cancer to "scan" the genome for common genetic variations between patients who have these cancers and controls who do not have cancer.

Learn more >>

Background

Spotlight

Cancer Genetic Markers of Susceptibility (CGEMS)

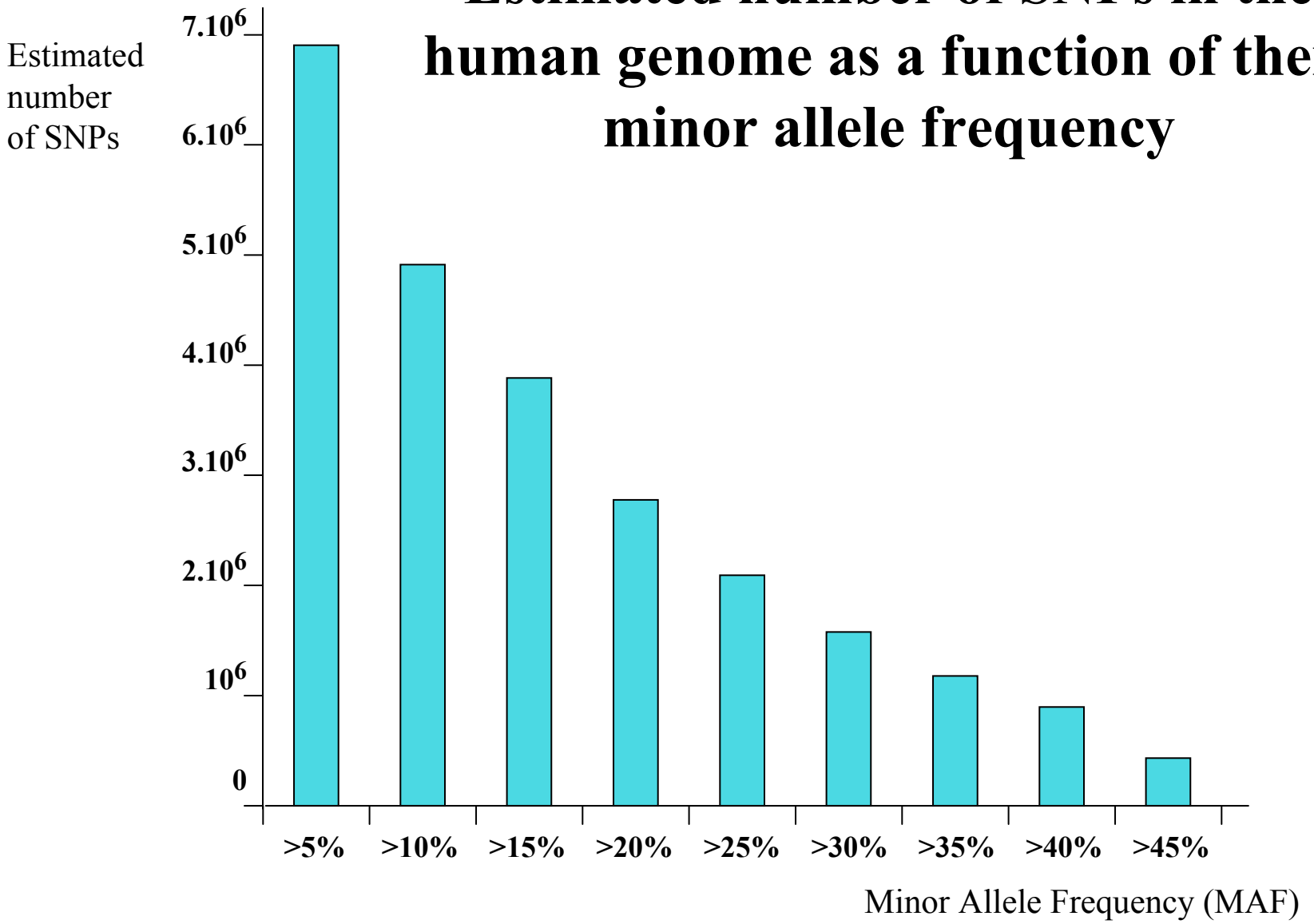
February 13, 2006

NCI begins studies to identify genetic risk factors for prostate and breast cancer.

more

DCEG and CGF Collaborate on CGEMS Initiative

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)

CGEMS

Conduct whole genome SNP scans

- **Prostate**
- **Breast**

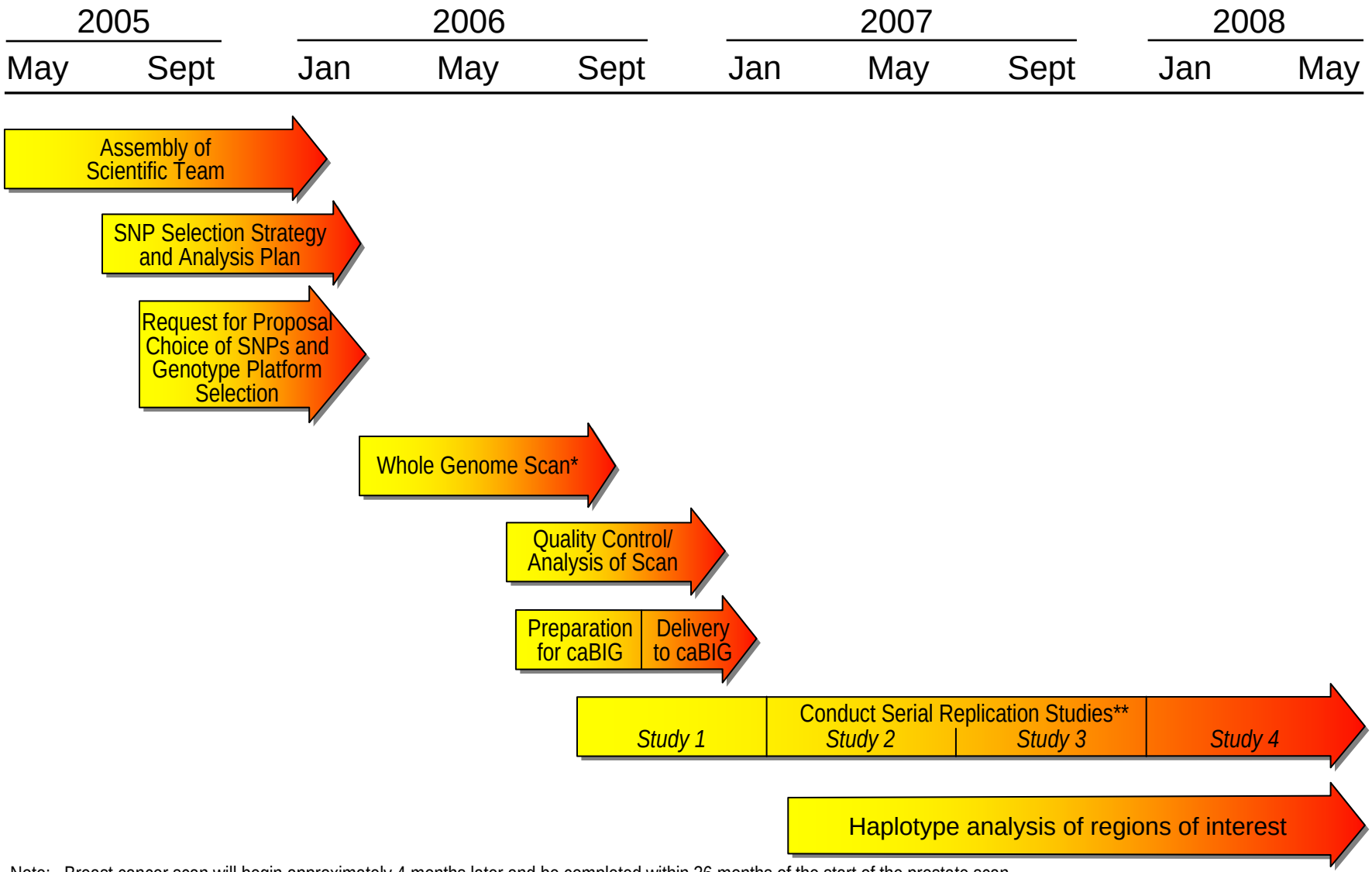
Rapid sequential replication studies

Aggressive time-line

Initial Scan in a Cohort Study

- **PLCO- Prostate Cancer**
- **Nurses Health- Breast Cancer**

Milestones for CGEMS Prostate Cancer Scan



Note: Breast cancer scan will begin approximately 4 months later and be completed within 36 months of the start of the prostate scan

* Whole genome scan of prostate will be performed in two parts

** Timing and specific studies will depend upon technical throughput and cost- Executive summaries will be posted within 4 months of completion

Whole Genome Scans

- Statistical Issues
  - Primary scan
    - Trade-off between size and detectable effect
  - Replication plan
    - Sufficiently powered to retain true positives
- Data availability
  - Public access policy
- Public Tools
- Common Database Structure
- Consortial/Collaborative Efforts

Comparison of HapMap 1 and HapMap 2 for CEU MAF>5%

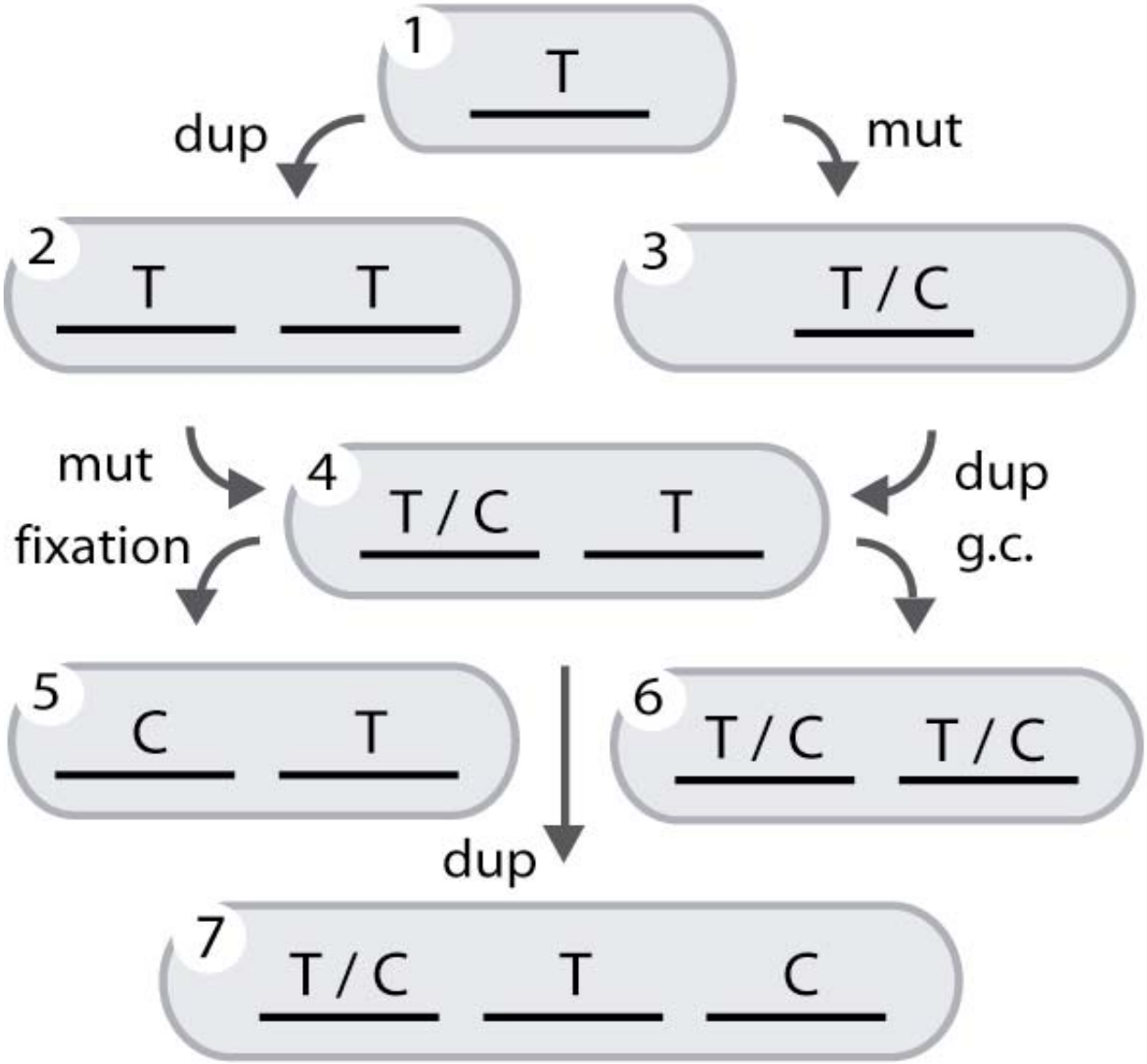
Phase I Bin statistics:

Size	Bins	% Bins
1	178312	58.76%
2	48752	16.07%
3	24312	8.01%
4	14201	4.68%
5	9245	3.05%
6	6402	2.11%
7	4426	1.46%
8	3324	1.10%
9	2542	0.84%
10	1936	0.64%
11	1590	0.52%
12	1177	0.39%
13	1026	0.34%
14	796	0.26%
> 14	5394	1.78%
Total	**303435**	**100.00%**

Phase II Bin statistics:

Size	Bins	% Bins
1	279577	52.13%
2	74165	13.83%
3	41403	7.72%
4	27210	5.07%
5	19716	3.68%
6	14594	2.72%
7	11321	2.11%
8	9223	1.72%
9	7485	1.40%
10	6187	1.15%
11	5210	0.97%
12	4365	0.81%
13	3792	0.71%
14	3262	0.61%
> 14	28818	5.37%
Total	**536328**	**100.00%**

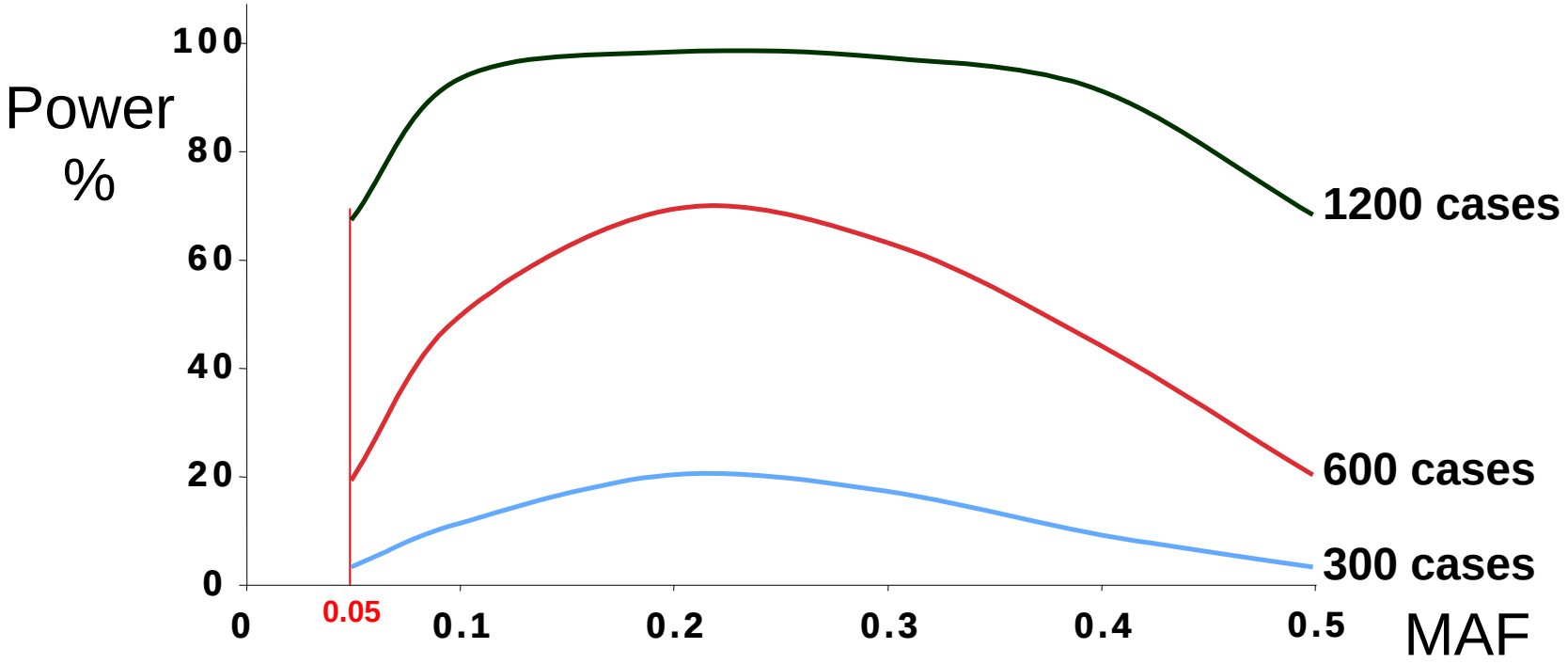
Thinking about Copy Number Polymorphisms…



C Lee 2005

2-stage WGS strategy

Power as a function of MAF and sample sizes



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 X # in stage 1**
SNPs in stage 2 : 25,000
Significance level 0.00002

Note: For significance level = 0.00002 => 10 false positives

Skol 2006

National Cancer Institute

Challenges of Keeping Pace with Evolving Genotyping and Sequencing Technologies

Stephen Chanock, M.D.

Senior Investigator, POB,CCR &

Director, Core Genotyping Facility, NCI

U.S. DEPARTMENT OF HEALTH AND HUMAN SERVICES

National Institutes of Health