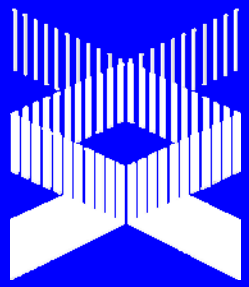


National Human Genome Research Institute

National Institutes of Health

U.S. Department of Health and Human Services

Cohort and Case-Control Studies in the Era of Genome-Wide Association

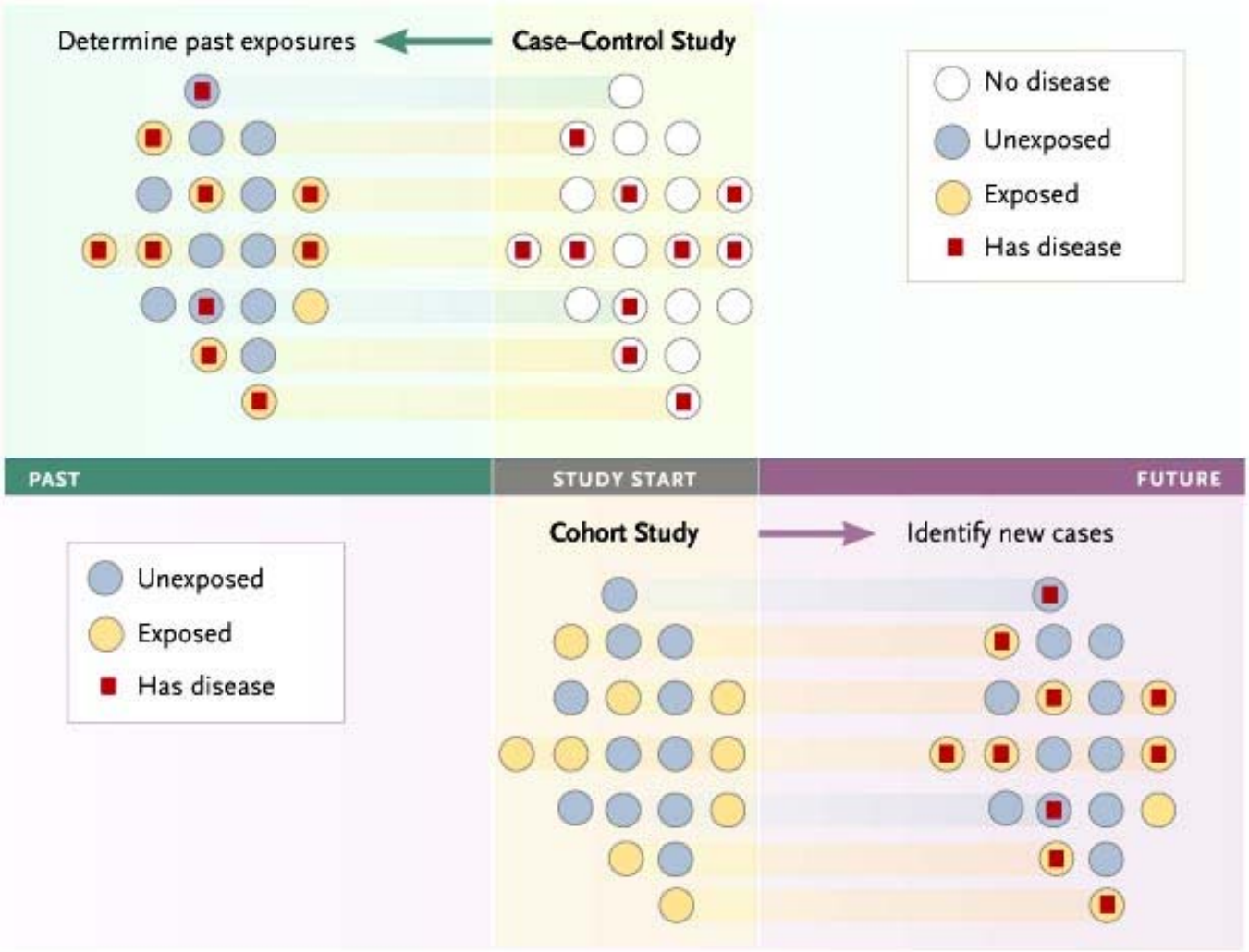
U.S. Department of Health and Human Services
National Institutes of Health
National Human Genome Research Institute

Teri A. Manolio, M.D., Ph.D.
Senior Advisor to the Director for Population Genomics
Director, Office of Population Genomics
February 15, 2007

Topics to be Covered

- Case-control and cohort studies
- Gene-environment interactions
- Genome-wide association studies

Risk Assessment in Case-Control and Cohort Studies



Manolio T, *N Engl J Med* 2003; 349:1587-1589.

Cohort Study

- Definition: investigation of representative sample of population followed forward in time for development of specified endpoints
- Purpose: identify risk factors predisposing to, or biomarkers predicting, development of disease in the population at large (not only among persons coming to medical attention)
- Value: detecting risk factors and risk markers that may be affected by disease, treatment, or lifestyle changes; subject to imperfect or biased recall; and/or having hypothesized early pathogenic effect

Pros and Cons of Cohort Studies

Disadvantages

- They are expensive.
- They take a long time.
- They are very broad-based.

Advantages

- They provide risk information obtainable through no other means.
- They are understandable to the public and media.
- They identify modifiable risk factors for potential preventive interventions.

Case-Control Study

- Definition: investigation of representative sample of disease cases compared with representative sample of disease-free controls, typically investigated backward in time for evidence of exposures existing prior to disease onset
- Purpose: identify associations between disease of interest and potential risk factors, particularly those that can be reported by participants or assessed from pre-existing records or specimens in an unbiased way
- Value: detecting risk factors for rare disease

Pros and Cons of Case-Control Studies

Advantages

- May be the only way to study rare diseases or those of long latency
- Existing records can be used if risk factor data collected independent of disease status
- Can study multiple etiologic factors simultaneously
- May be less time-consuming and expensive
- If assumptions met, inferences are reliable

Pros and Cons of Case-Control Studies

Disadvantages

- Relies on recall or records for information on past exposures; validation can be difficult or impossible
- Selection of appropriate comparison group may be difficult
- Multiple biases may give spurious evidence of association between risk factor and disease
- Usually cannot study rare exposures
- Temporal relationship between exposure and disease can be difficult to determine

"But," They Say, "*This* is Genetics!"

(You Dumb Epidemiologist)

"*This* is Different!"

- Genes are measured the same way in cases and controls
- Information on key exposure is easy to validate
- No recall or reporting involved
- Temporal relationship between genes and disease is piece of cake

"But," I Say,

- Bias-free ascertainment of cases and controls is still major concern; cases in most clinical series unlikely to be representative
- Assessment of risk modifiers or gene-environment interactions is likely to be incomplete or flawed

Case-Control Studies and Rare Diseases

- For a disease with incidence of 8 cases per 1,000 among unexposed, cohort study would require 3,889 exposed and 3,889 unexposed persons to detect two-fold increase in risk
- Case-control study would require 188 cases and 188 controls, assuming 30% exposure
- For disease with incidence of 2 cases per 1,000 among unexposed, would need 15,700 exposed and 15,700 unexposed to detect two-fold risk
- Case-control study would *still require only 188 cases and 188 controls*

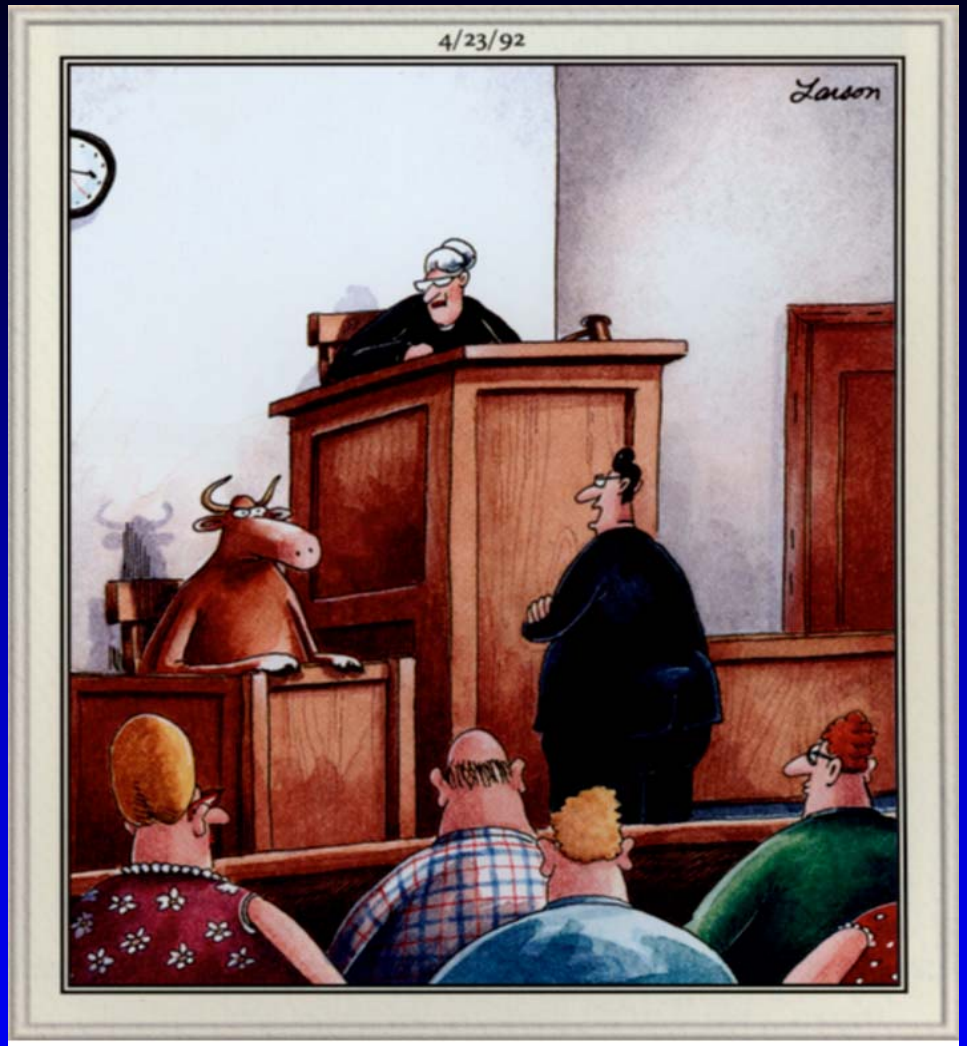
Schlessman JJ. *Case-Control Studies*, 1982.

So What's a Mother to Do?

- "Nesting" a case-control study within a prospective cohort may provide the best of both worlds
- Large proportion of cohort members who do not develop disease provide little incremental information
- If exposure information can be collected and stored for later measurement, can wait for cases to accrue and then measure exposures in limited sample of non-cases
    - stored biologic samples
    - stored images
- Can be expanded to "case-cohort" concept with representative sample of cohort, regardless of disease status, used for multiple comparisons

Comparison of Case-Control and Cohort Studies

Characteristic	Case-Control Studies	Cohort Studies
Temporal relationship of exposure to disease	May be hard to establish	Generally easy to establish
Types of associations studied	Single disease, multiple exposures	Multiple diseases, multiple exposures
Duration of study	Relatively short	Typically long
Cost of study	Low	High
Population size	Small	Large
Potential biases	Assessment of exposure	Assessment of outcome
Situation in which design is preferred	Disease is rare, exposure is frequent among diseased	Exposure is rare, disease is frequent among exposed



"Look. We know *how* you did it --*how* is no longer the question. What we now want to know is *why*. ... Why now, brown cow?"

Larson, G. *The Complete Far Side*. 2003.

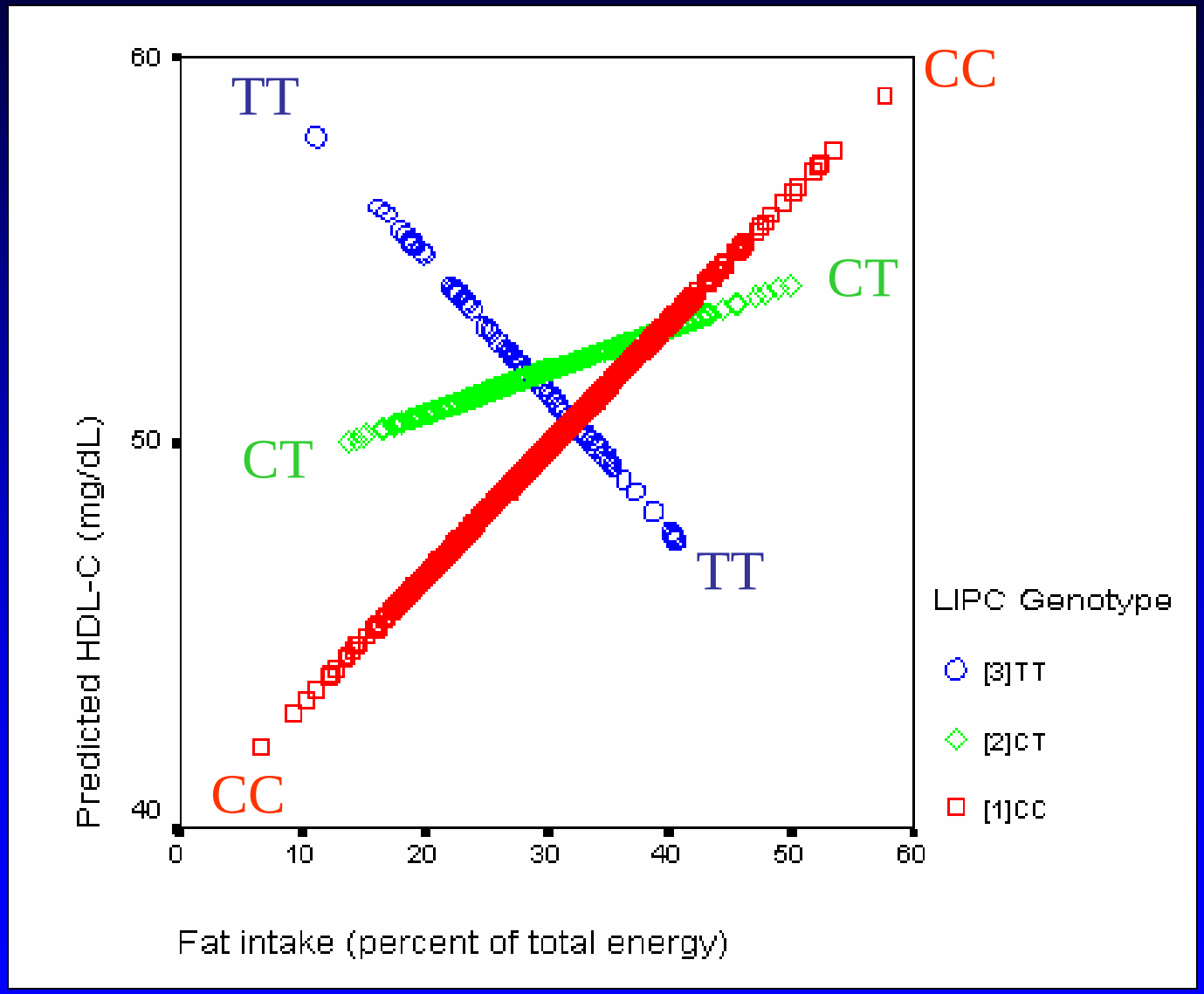
Why are Gene-Environment Interactions so Important to Public Health?

- Environmental and behavioral changes interacting with genetic predisposition have likely produced most of the recent epidemics of chronic diseases
- GxE may be key in reversing their course, by suggesting approaches for modifying effects of deleterious genes
- Future public health measures may focus on avoiding deleterious environmental exposure, especially in genetically susceptible persons

Why are Gene-Environment Interactions so Important to Research?

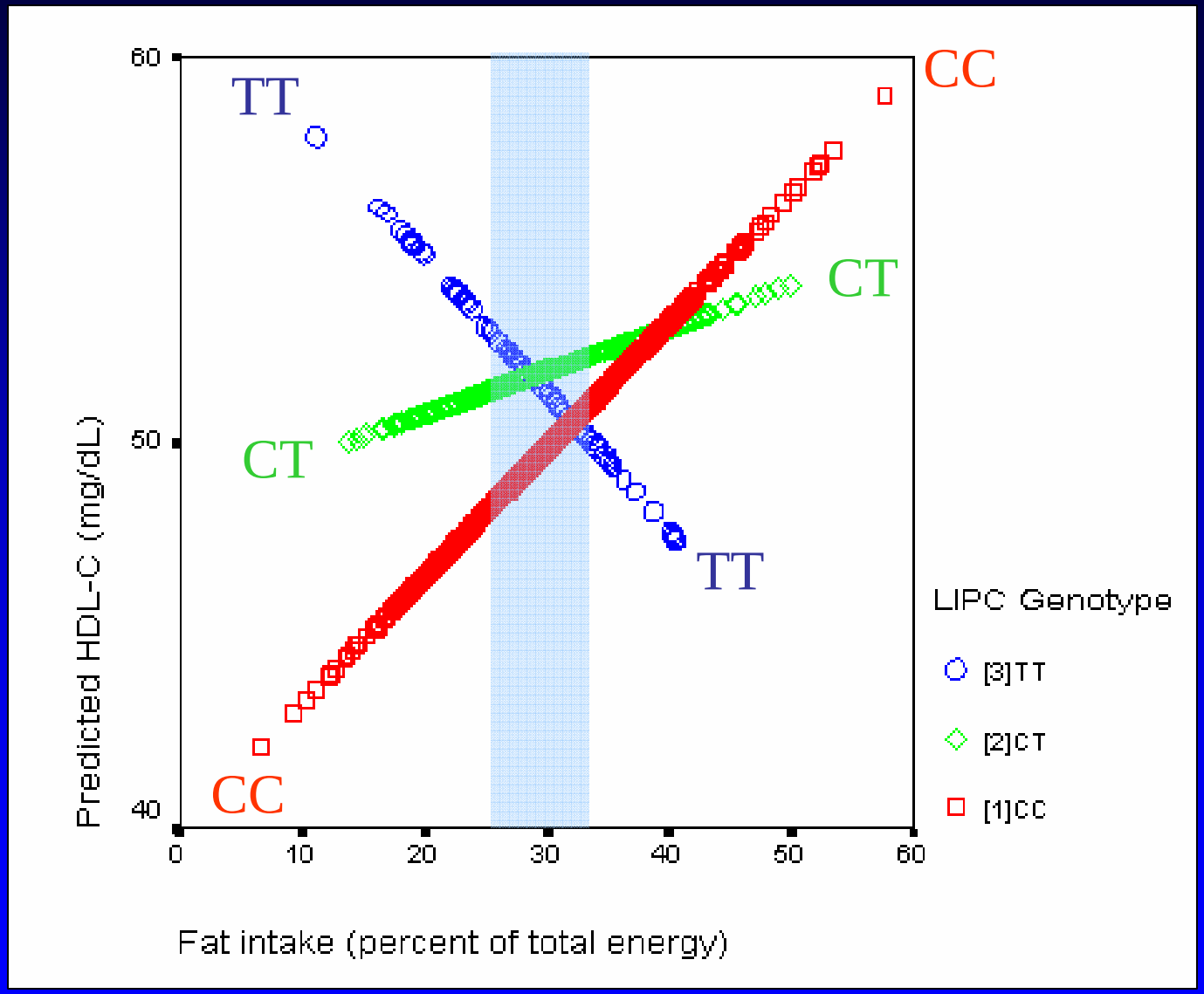
- Can mask detection of genetic (or environmental) effect if they are not identified and controlled for
- Can lead to inconsistencies in disease associations in different populations with:
  - Different environmental exposures that modify the effect of a genetic variant
  - Different prevalences of genetic variants that modify the effect of an environmental exposure

Is LIPC Genotype Related to HDL-C?



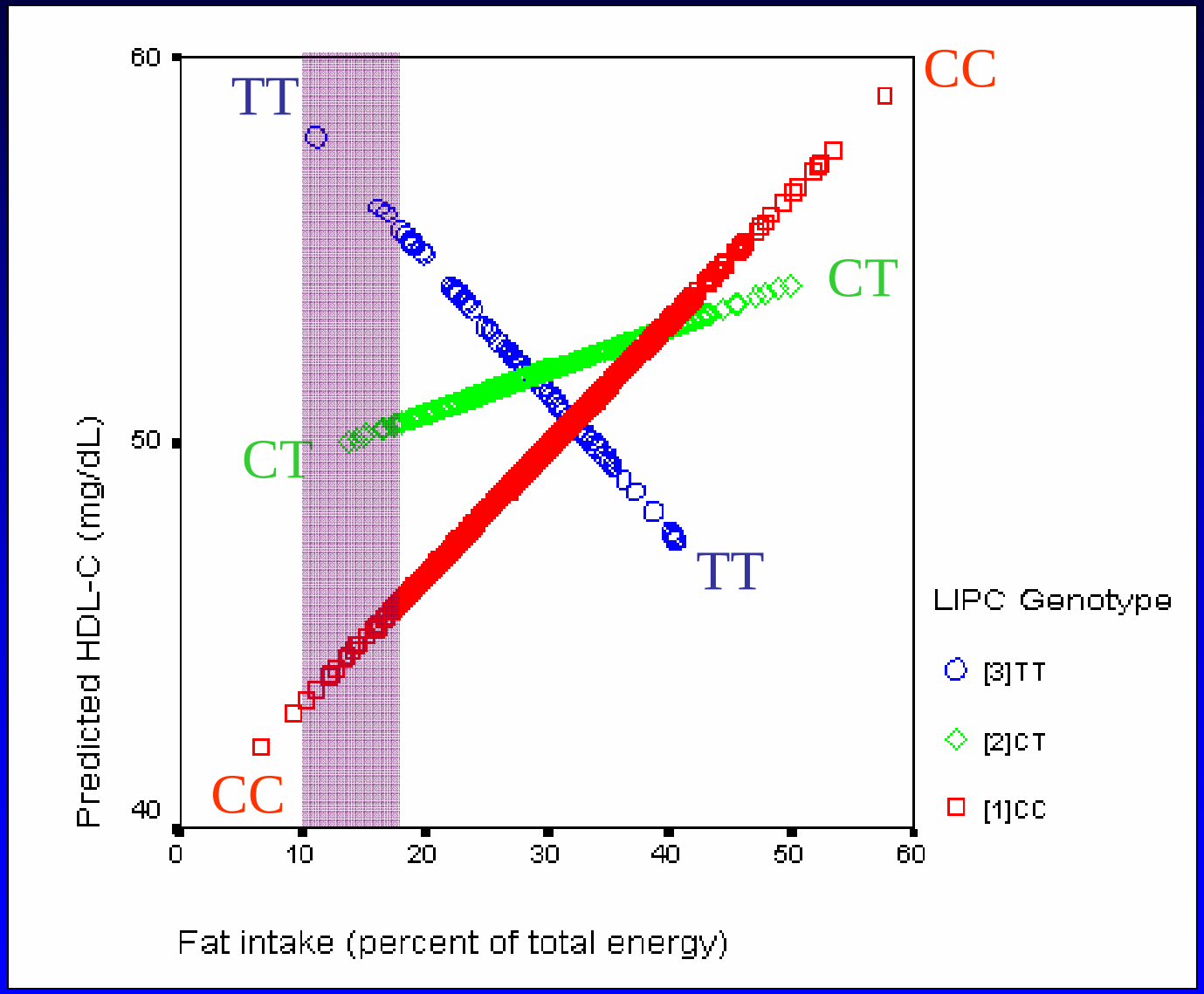
Ordovas et al, *Circulation* 2002; 106:2315-2321.

Is LIPC Genotype Related to HDL-C?



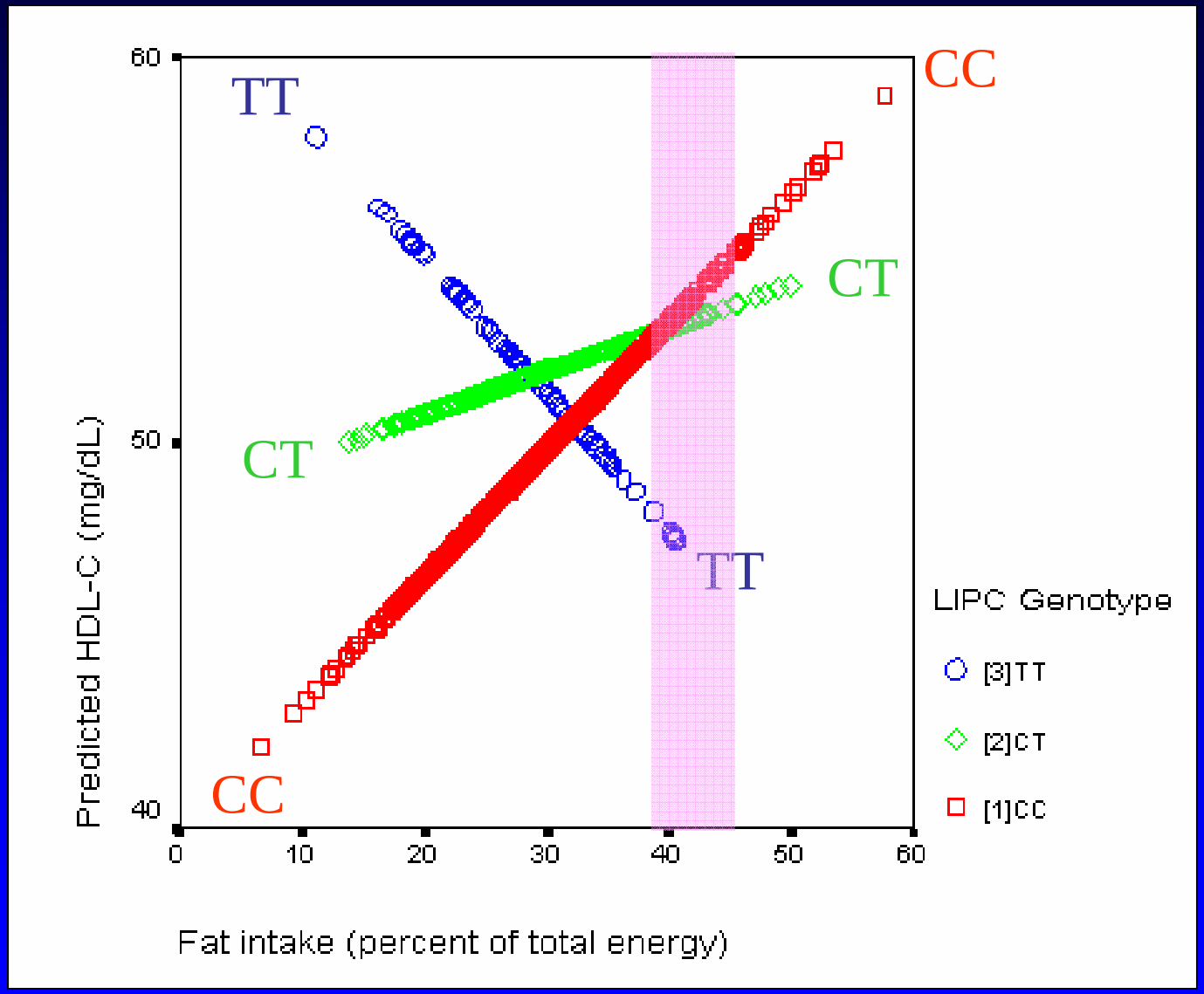
Ordovas et al, *Circulation* 2002; 106:2315-2321.

Is LIPC Genotype Related to HDL-C?



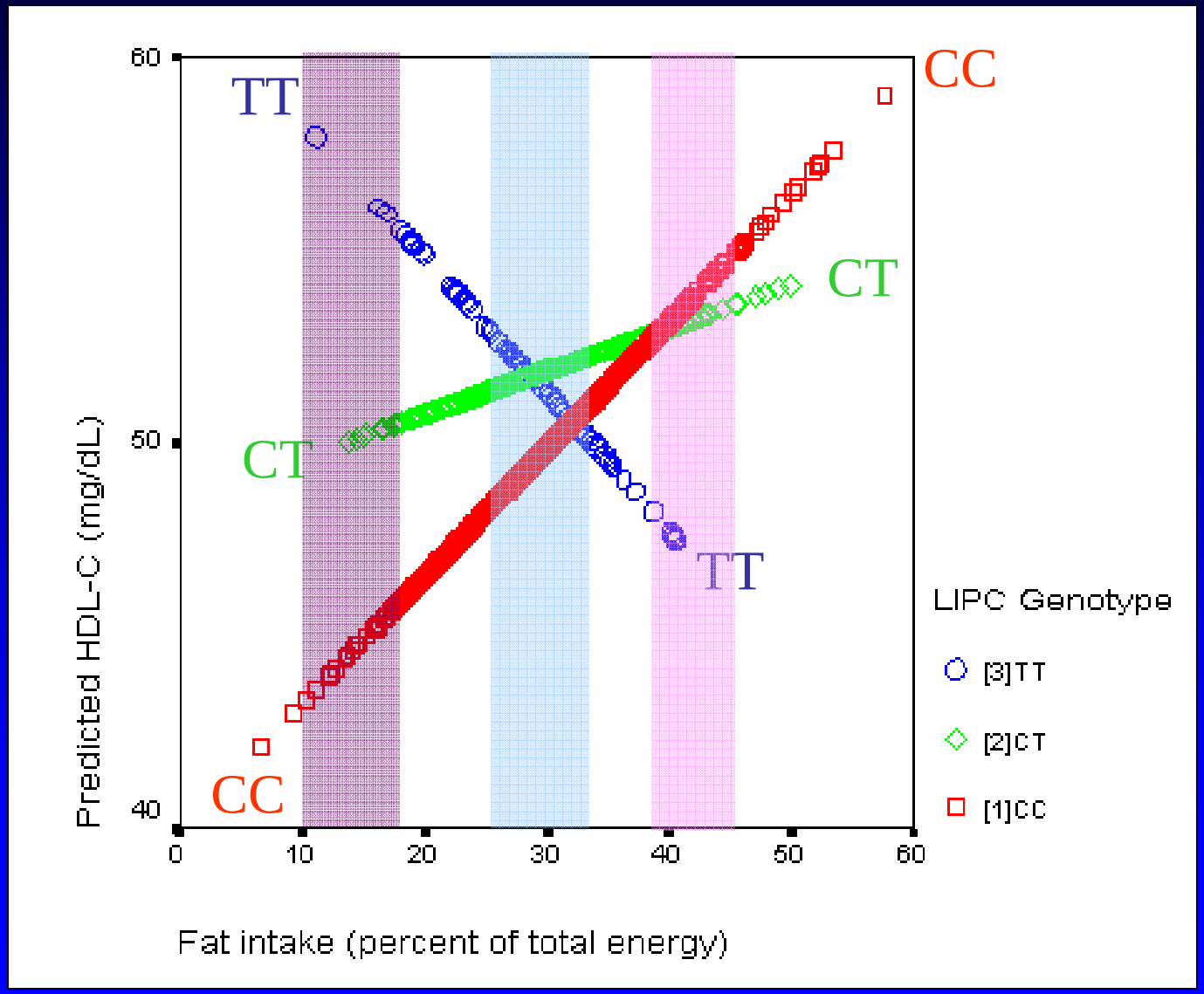
Ordovas et al, *Circulation* 2002; 106:2315-2321.

Is LIPC Genotype Related to HDL-C?



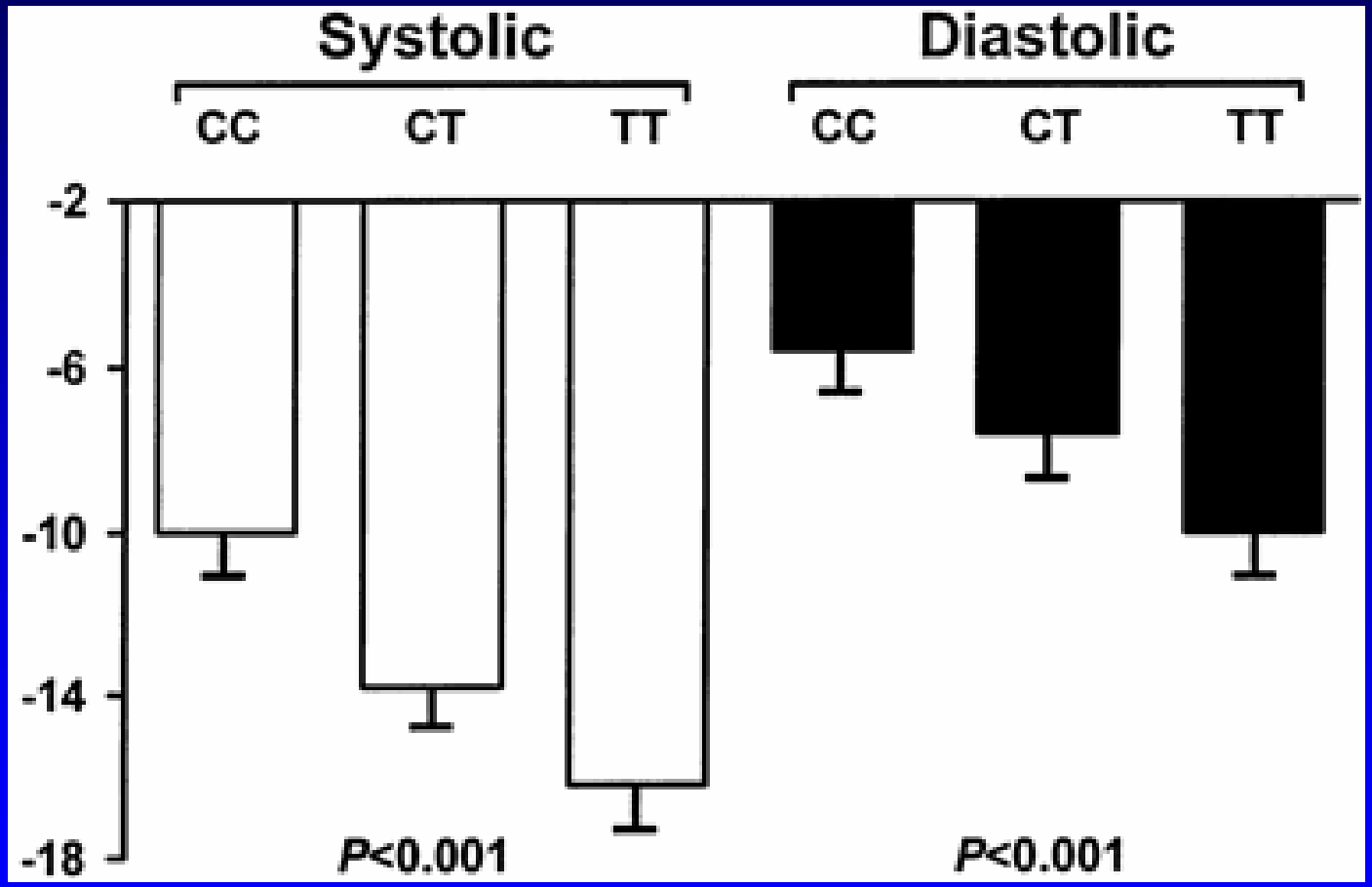
Ordovas et al, *Circulation* 2002; 106:2315-2321.

Is LIPC Genotype Related to HDL-C?

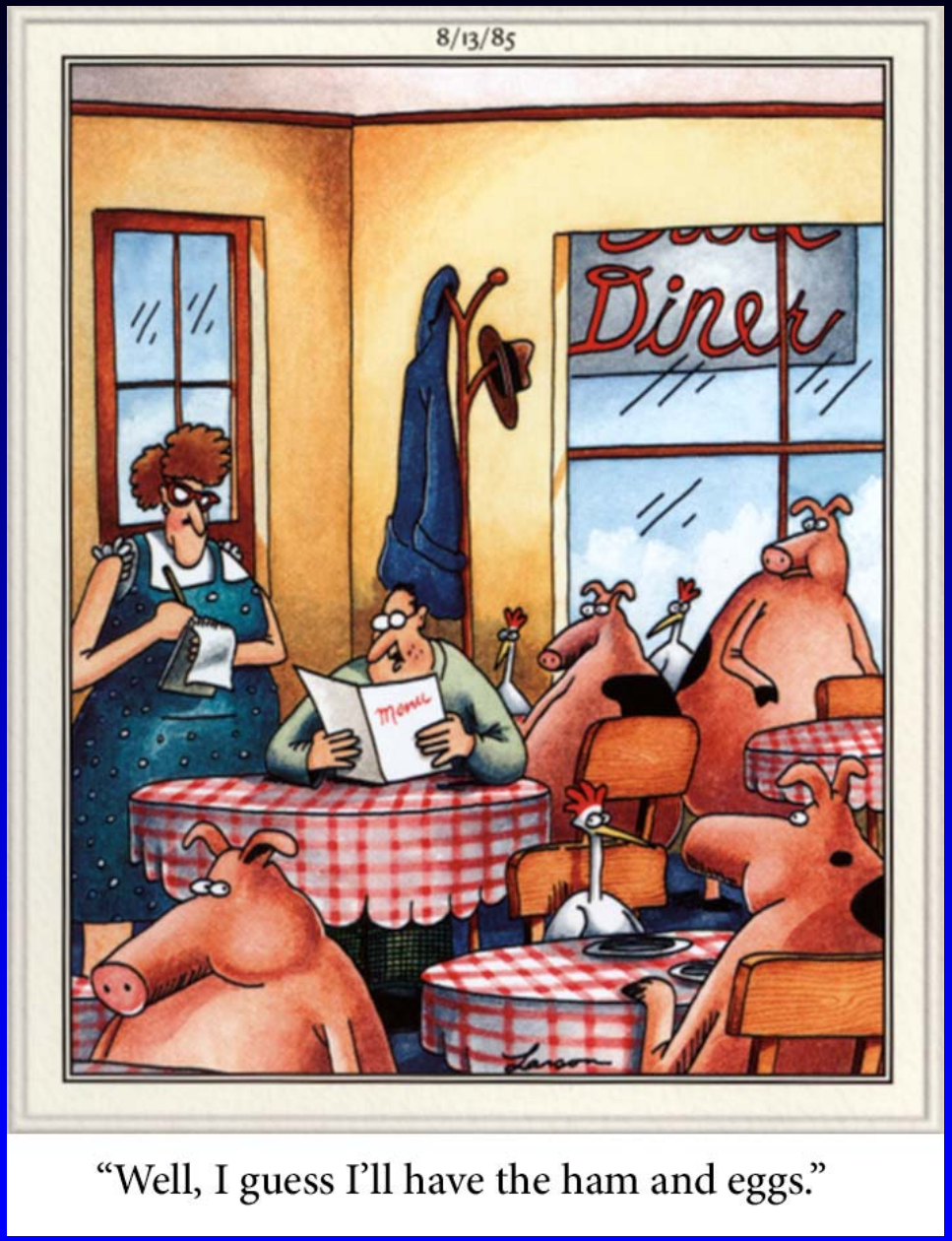


Ordovas et al, *Circulation* 2002; 106:2315-2321.

Blood Pressure Response to Thiazide by G Protein Genotype



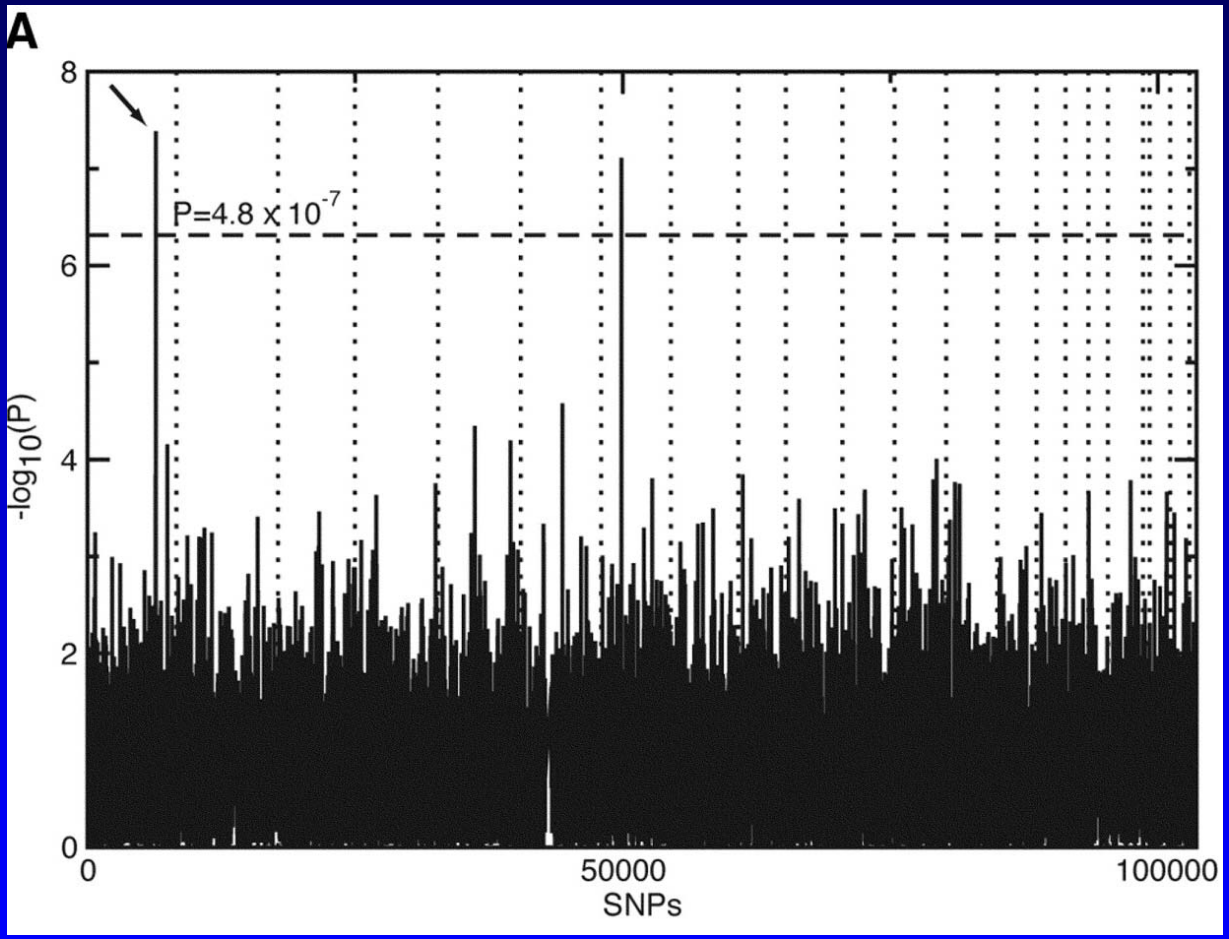
Turner S et al, *Hypertension* 2001; 37:739-743.



"Well, I guess I'll have the ham and eggs."

Larson, G. *The Complete Far Side*. 2003.

P Values of GWA Scan for Age-Related Macular Degeneration



Klein et al, *Science* 2005; 308:385-389.

Odds Ratios and Population Attributable Risks for AMD

Attribute (SNP)	rs380390 (C/G)	rs1329428 (C/T)
Risk allele	C	C
Allelic association χ^2 P value	4.1×10^{-8}	1.4×10^{-6}
Odds ratio (dominant)	4.6 [2.0-11]	4.7 [1.0-22]
Frequency in HapMap CEU	0.70	0.82
Population Attributable Risk	70% [42-84%]	80% [0-96%]
Odds ratio (recessive)	7.4 [2.9-19]	6.2 [2.9-13]
Frequency in HapMap CEU	0.23	0.41
Population Attributable Risk	46% [31-57%]	61% [43-73%]

Klein et al, *Science* 2005; 308:385-389.

Genetic Studies in Unrelated Individuals *(pre-2005)*: Candidate Gene Studies

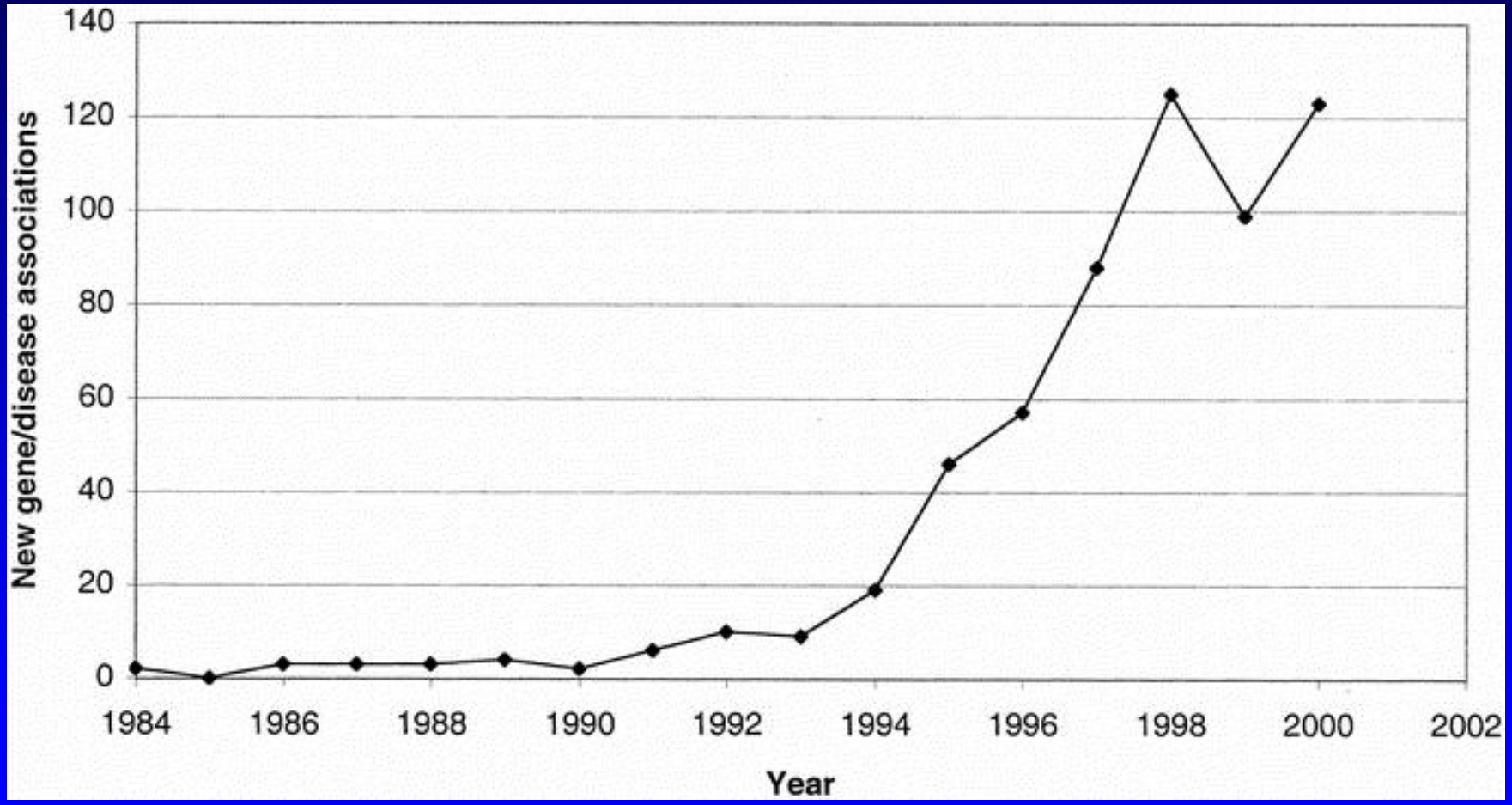
- Goal: characterize candidate genes and variants related to disease
- Not typically intended to "find genes," generally begun *after* disease-related variants identified
- Assess generalizability of family-based observations (genetic heterogeneity)
- Assess importance of allelic variation at population level (PAR, penetrance)
- Identify modification of genetic association by environmental factors (GxE interaction)

Age-Adjusted Odds on Hypertension by ACE ID/DD Genotype and Sex

	DD	ID	II	P-value
Men: % HTN	53.1	45.8	44.4	
Men: OR	1.67	1.19	1.00	0.004
Women: % HTN	43.3	41.8	44.4	
Women: OR	1.01	0.80	1.00	0.15

O'Donnell C et al, *Circulation* 1998; 97:1766-1772.

Number of New, Significant Gene-Disease Associations by Year, 1984 - 2000



Hirschhorn J et al, *Genet Med* 2002; 4:45-61.

Of 600 Gene-Disease Associations, Only 6 Significant in ≥ 75% of Identified Studies

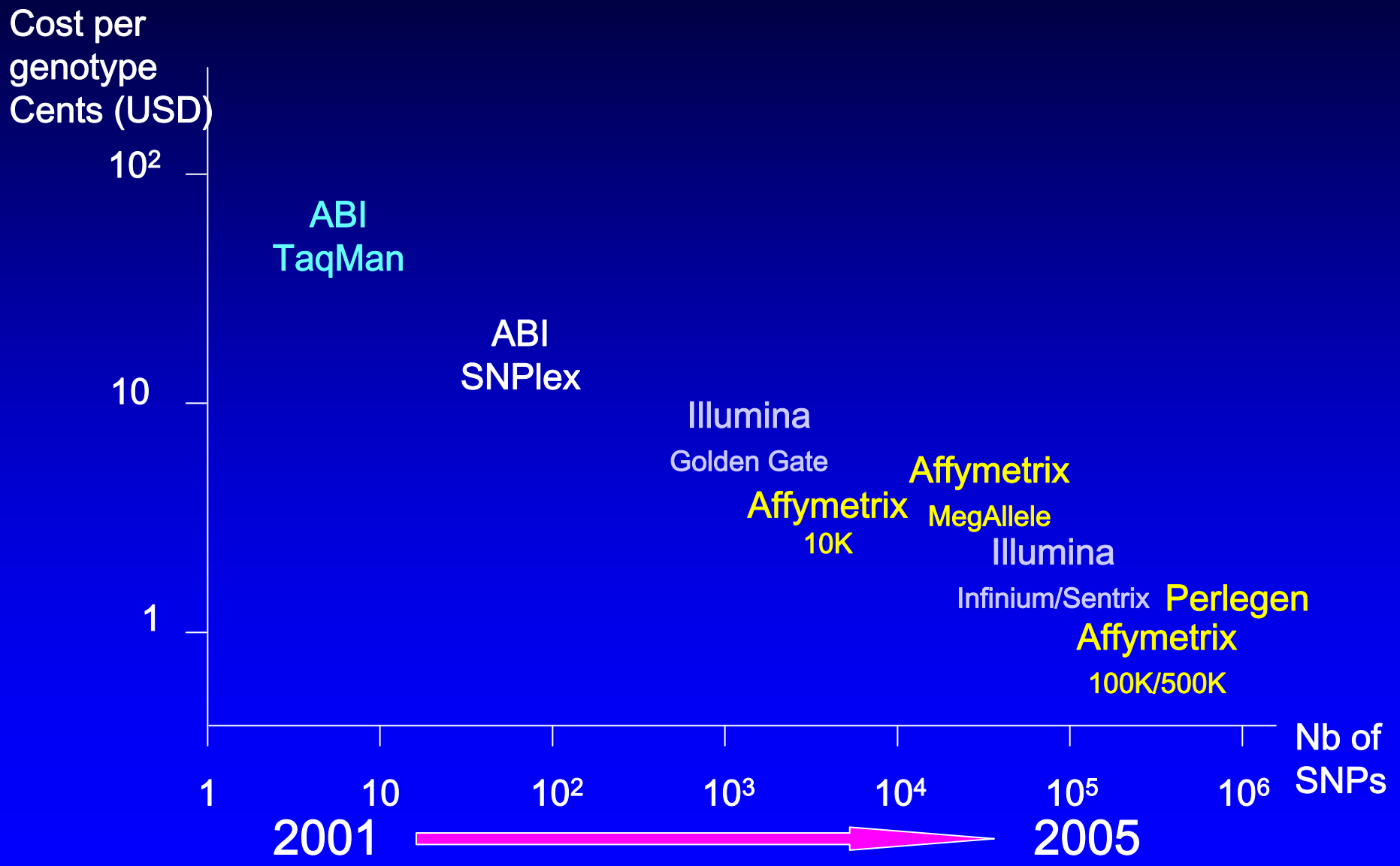
Disease/Trait	Gene	Polymorphism	Frequency
DVT	F5	Arg506Gln	0.015
Graves' Disease	CTLA4	Thr17Ala	0.62
Type 1 DM	INS	5' VNTR	0.67
HIV/AIDS	CCR5	32 bp Ins/Del	0.05-0.07
Alzheimer's	APOE	Epsilon 2/3/4	0.16-0.24
Creutzfeldt-Jakob Disease	PRNP	Met129Val	0.37

Hirschhorn J et al, *Genet Med* 2002; 4:45-61.

Application of Genomic Technologies to Well-Characterized Individuals: Genome-Wide Association and Sequencing

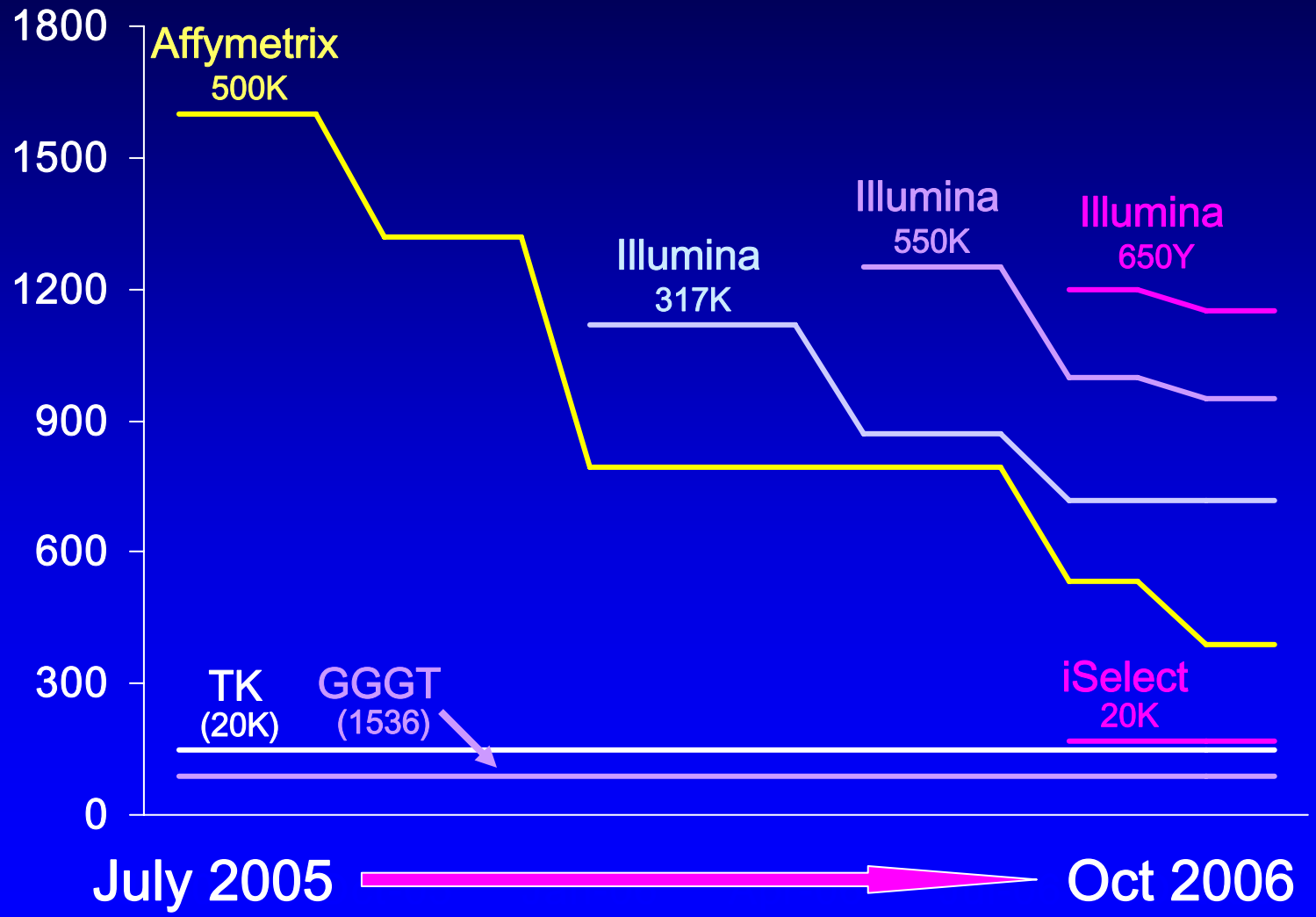
- Leverage large number of existing cohort and case-control studies of complex diseases
- Complex diseases: caused by multiple genes of small effect, not amenable to family studies
- Genome-wide: interrogate all variation throughout genome, 300-500K SNPs in thousands of unrelated individuals

Progress in Genotyping Technology



Courtesy S. Chanock, NCI

Continued Progress in Genotyping Technology



Courtesy S. Gabriel, Broad/MIT

www.hapmap.org

Vol 437|27 October 2005|doi:10.1038/nature04226 nature

ARTICLES

A haplotype map of the human genome

The International HapMap Consortium*

Inherited genetic variation has a critical but as yet largely uncharacterized role in human disease. Here we report a public database of common variation in the human genome: more than one million single nucleotide polymorphisms (SNPs) for which accurate and complete genotypes have been obtained in 269 DNA samples from four populations, including ten 500-kilobase regions in which essentially all information about common DNA variation has been extracted. These data document the generality of recombination hotspots, a block-like structure of linkage disequilibrium and low haplotype diversity, leading to substantial correlations of SNPs with many of their neighbours. We show how the HapMap resource can guide the design and analysis of genetic association studies, shed light on structural variation and recombination, and identify loci that may have been subject to natural selection during human evolution.

International HapMap Consortium, *Nature* 2005; 437:1299-1320.

Genetic Studies in Unrelated Individuals *post-2005*: Genome-Wide Association

- Find genes related to complex diseases
- Complex diseases: caused by multiple genes of small effect, not amenable to family studies
- Whole genome: interrogate all variation throughout genome, two main approaches
  - Family linkage study with 400 microsatellite markers, assumes ~10mb regions of LD
  - Unrelated case-control study with 300-500K SNPs, assumes ~10kb regions of LD

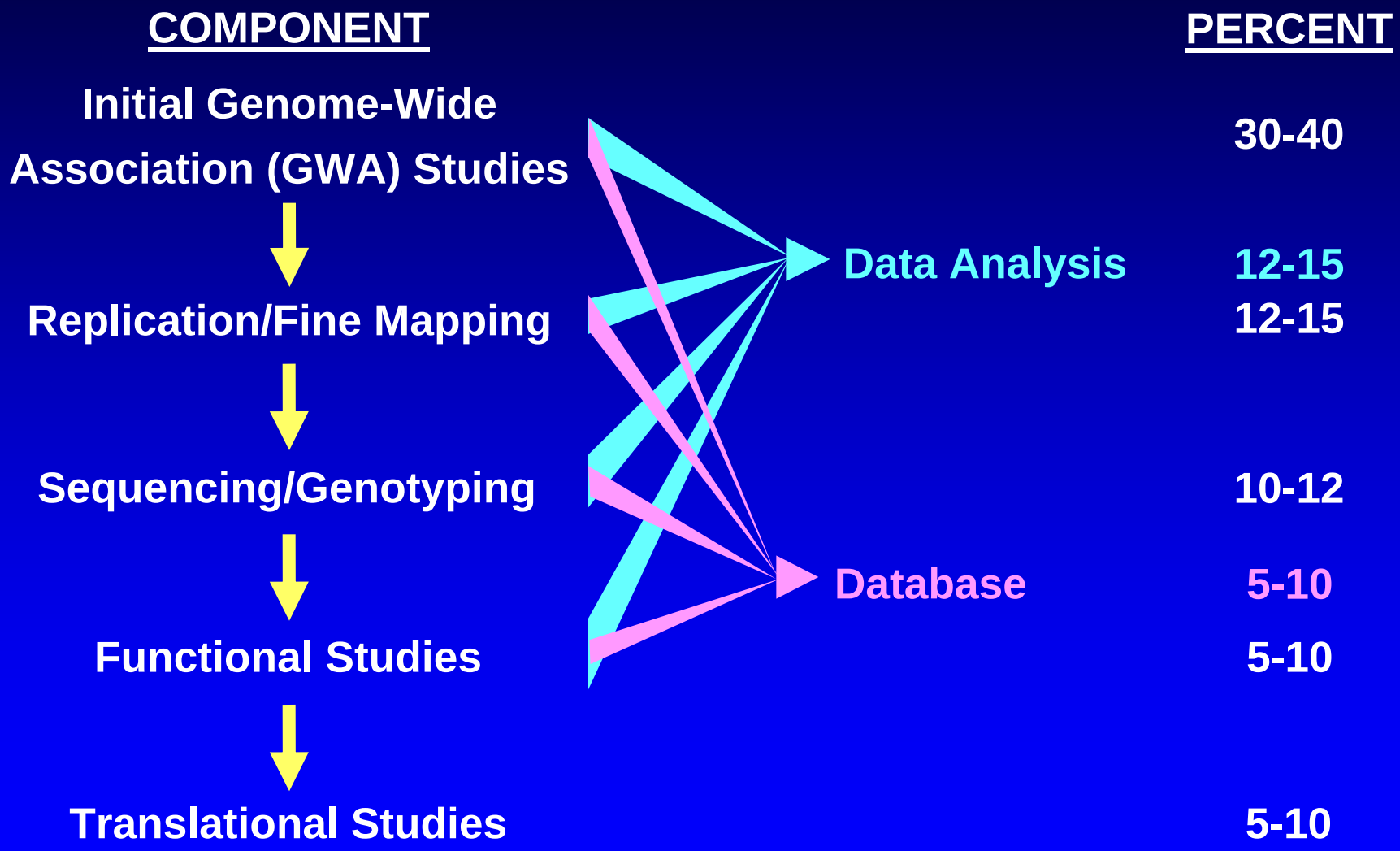
Value of GWA Studies in Unrelated Individuals

- Easier to study
- Many existing collections
  - Extremely well-characterized
  - Followed for long periods
  - Diverse in origin, exposures
- Large families remain very valuable
  - Not so common anymore
  - Confounding by shared environment

Current NIH-Affiliated GWA Studies

- Genetic Association Information Network (GAIN): public-private partnership for GWA studies of ~ 7 common diseases
- Genes and Environment Initiative (GEI)
  - RFA HG-06-014: Genotyping Facilities
  - RFA HG-06-032: Coordinating Center
  - RFA HG-06-033: Study Investigators
- NCI Cancer Genetic Markers of Susceptibility
- NHLBI Framingham SHARe Project
- NHLBI STAMPEED Program
- NIDDK Diabetes and Diabetes Complications

Flow of Investigation: From Genome-Wide Association to Clinical Translation



Need for Consensus on What Constitutes Replication

- Avalanche of GWA and candidate gene studies now and in near future
- Replication held as *sine qua non*
- Likelihood of single study establishing an association is low until sample sizes increase sufficiently and analytical methods improve substantially
- Common problem of how to interpret confusing and spurious findings

Proposed Criteria for Positive Replication

- Sufficient sample size to distinguish proposed effect from no effect convincingly
- Same or very similar trait (extension to related trait may increase confidence in finding, such as consistent finding for both dichotomized obesity and continuous BMI)
- Same or very similar population (extension to other populations may also increase confidence in finding, such as consistent association in populations of European, Asian, or even recent African ancestry

Proposed Criteria for Positive Replication

- Same inheritance model (dominant, co-dominant, recessive), though not necessarily same analytic method)
- Same gene, same SNP (or SNP in complete LD with prior SNP, $r^2 = 1$), same direction as original finding
- Highly significant association
- N.B.: Initial study must adequately describe these parameters

Proposed Criteria for True Non-Replication or "Meaningful Negativity"

- Same as for positive replication (same trait, same gene, same SNP, same direction, same genetic model)
- Must be identical trait and population to claim non-replication
- Powered to appropriate effect size (account for "winner's curse")

Importance of Genotyping Quality

- Report results of known study sample duplicates, HapMap or other standard duplicates
- Replicate small number of “significant” SNPs with second technology at some late stage
- May not be needed if nearby SNPs in strong LD show same results
- Strong caveats are needed regarding fallibility of genotyping
  - Results can change based on genotype calling algorithm
  - QC filters and consistency of results after applying them must be described

insight commentary

The case for a US prospective cohort study of genes and environment

Francis S. Collins

National Human Genome Research Institute, National Institutes of Health, Building 31, Room 4B09, MSC 2152, 31 Center Drive, Bethesda, Maryland 20892-2152, USA (e-mail: fc23a@nih.gov)

Information from the Human Genome Project will be vital for defining the genetic and environmental factors that contribute to health and disease. Well-designed case–control studies of people with and without a particular disease are essential for this, but rigorous and unbiased conclusions about the causes of diseases and their population-wide impact will require a representative population to be monitored over time (a prospective cohort study). The time is right for the United States to consider such a project.

Identification of the genetic and environmental factors that contribute to health, disease and response to treatment is essential for the reduction of illness. This, of course, is the primary goal of biomedical research. Several auspicious recent developments suggest that progress in this area could be quite rapid. The sequence of the human genome[1,2] and increasing information about the genome's function have provided a robust foundation for the investigation of human health and disease. Likewise, results from the exploration of human genetic environmental exposure have improved. These techniques promise to extend the range of epidemiological investigation[5]. There is growing recognition that a change in the environment, in combination with genetic disposition, has produced most recent epidemics of chronic disease, and may hold the key for reversing the course of some diseases[6]. For example, consider the interaction of presumed famine-protective genetic predispositions with a modern environment in which there is a ready availability of excess calories. This has probably contributed to the current obesity epidemic

Desirable Characteristics of Large US Cohort Study

- Large sample size
- Full representation of minority groups
- Broad range of ages
- Broad range of genetic backgrounds and environmental exposures
- Family-based recruitment for at least part of the cohort to control for population stratification
- Broad array of clinical and laboratory data, regular follow up for events, additional exposure assessment

After Collins FS, *Nature* 2004; 429:475-477.

Desirable Characteristics of Large US Cohort Study (Continued)

- Technologically advanced dietary, lifestyle, and environmental exposure data
- Collection and storage of biological specimens
- Sophisticated data management system
- Access to materials and data by all researchers
- Goals should not be “hypothesis-limited”
- Comprehensive community engagement from the outset
- State of the art (?dynamic) consent to allow multiple uses of data and regular feedback to participants

After Collins FS, *Nature* 2004; 429:475-477.

"Now stay calm. ... Let's hear what they said to Bill."

Larson, G. *The Complete Far Side*. 2003.

OPINION

Genes, environment and the value of prospective cohort studies

Teri A. Manolio, Joan E. Bailey-Wilson and Francis S. Collins

Abstract | Case–control studies have many advantages for identifying disease-related genes, but are limited in their ability to detect gene–environment interactions. The prospective cohort design provides a valuable complement to case–control studies. Although it has disadvantages in duration and cost, it has important strengths in characterizing exposures and risk factors before disease onset, which reduces important biases that are common in case–control studies. This and other strengths of prospective cohort studies make them invaluable for understanding gene–environment interactions in complex human disease.

The sequencing of the human genome and increased investigation of its function are providing powerful research tools for identifying

when populations are subject to different environmental exposures that modify the effect of a given genetic variant (or the

Merging and emerging cohorts

How best to study the effects of genes and environment on US health? In the first of two commentaries, **Walter C. Willett** and his co-authors argue that investing in existing studies is the most efficient approach. In the second, **Francis S. Collins** and **Teri A. Manolio** explain their support for a new national cohort.

Not worth the wait

In 2006, the United Kingdom initiated a national long-term health study of 500,000 middle-aged adults that will involve collecting DNA and other biological specimens[1]. Further cohorts are being considered elsewhere in Europe and Asia. Francis Collins (ref. 2 and page xx) has proposed a similar national cohort of several hundred thousand North Americans to enable future studies of the genetic basis of human diseases and individual susceptibility to environmental factors. The cost is estimated to reach $3 billion or more[3].

We are concerned that results from a new cohort would not be available for at least ten years, as five years would be needed for funding, planning and enrolment, and another five for following up even the earliest analyses of the most common diseases; results for most cancers would take longer. We believe that a

limited to adult experiences, but lifetime histories are solicited for some exposures such as smoking.

crosshead

Moreover, participants have already provided informed consent for the use of their biographical data and biospecimens for research. Institutional review boards have approved each study and monitor new ethical issues as they arise. Many investigators already have data-sharing policies in place for collaborations. Reconsent is sometimes needed with existing or new cohorts to accommodate new ethical issues, but this is more easily obtained when a trusting relationship already exists between participants and investigators.

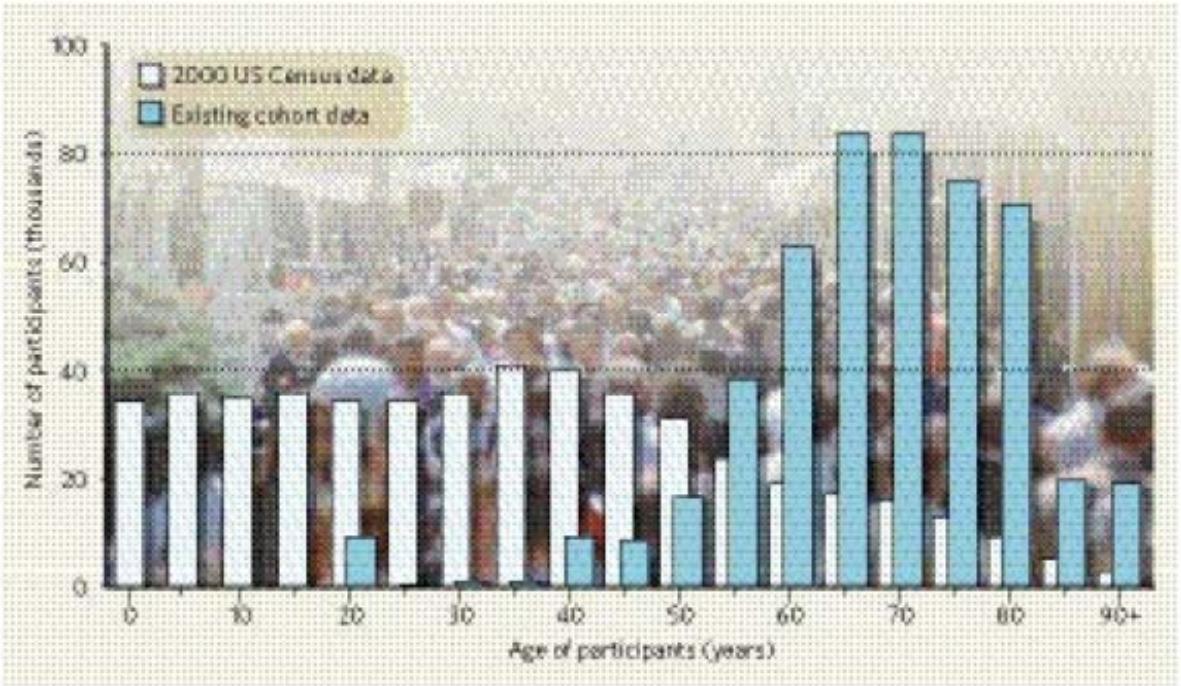
Individual cohorts have already provided key clues to disease etiology and prevention, but

Necessary but not sufficient

The proposal advocated in the preceding Commentary by Willett *et al.*[1], namely to extend existing cohort studies rather than start a new large-scale prospective study from scratch, has many merits. Indeed, a National Institutes of Health (NIH) study group that assessed the pros and cons of various models in 2004 considered this option in some depth, and their report[2] made many of the same points.

Certainly, assembling existing cohorts into a large consortium would provide a powerful resource for investigating genetic and environmental factors in health and disease. The argument that this method is likely to be less costly than a new cohort, and would yield results more quickly, carry considerable weight. But Willett *et al.* do not address all of the suboptimal aspects of this approach. Those should be clearly noted, lest expectations of such a consortium exceed what it is likely to deliver.

First, there is the issue of standardization. Phenotypic measures used by the existing

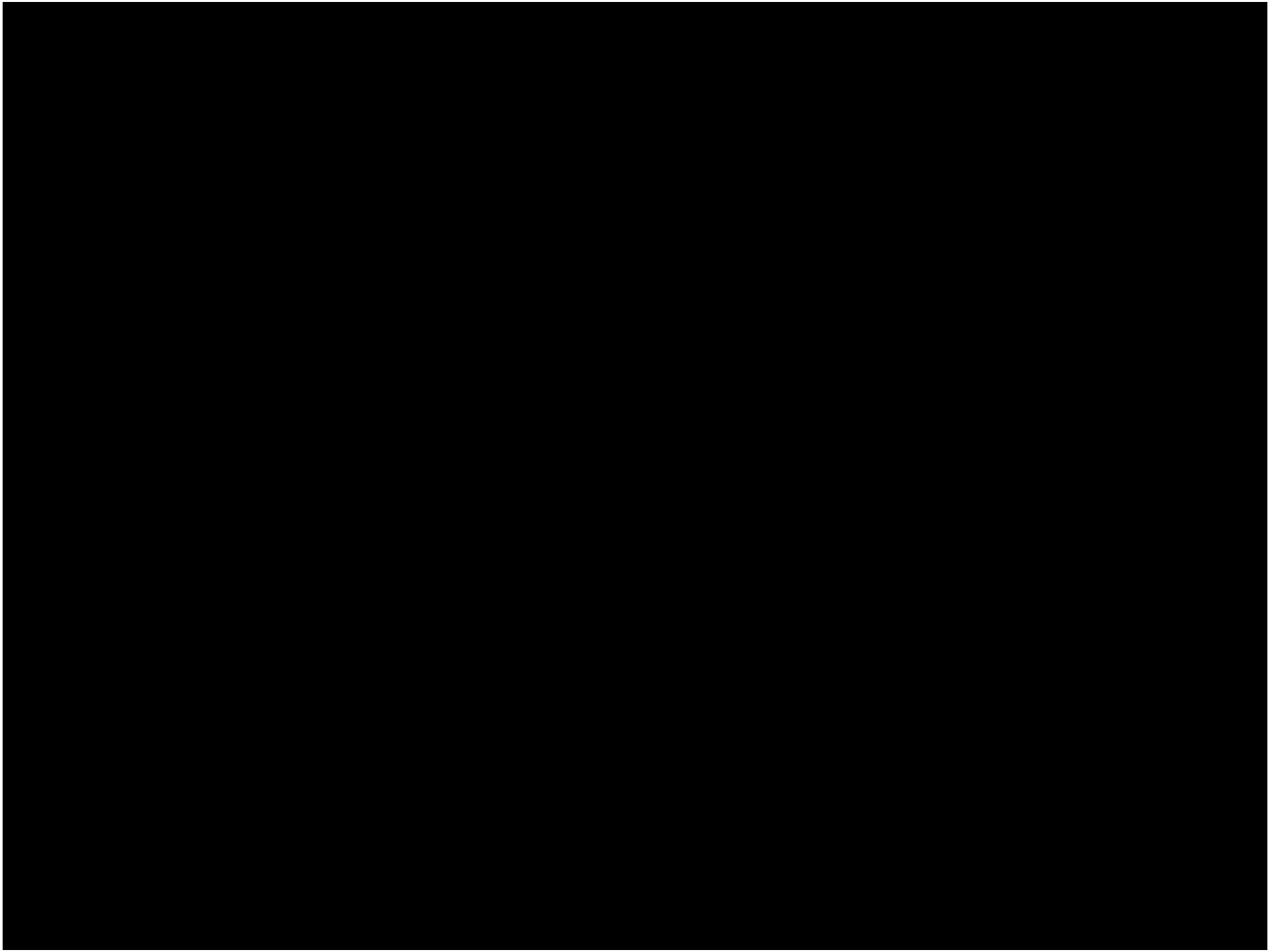


Comparison of an estimated distribution of a 500,000-person cohort, based on existing cohort data[2], with the 2000 US census.

worse with time. If we wish to address complex

Admittedly, this discussion remains hypo-

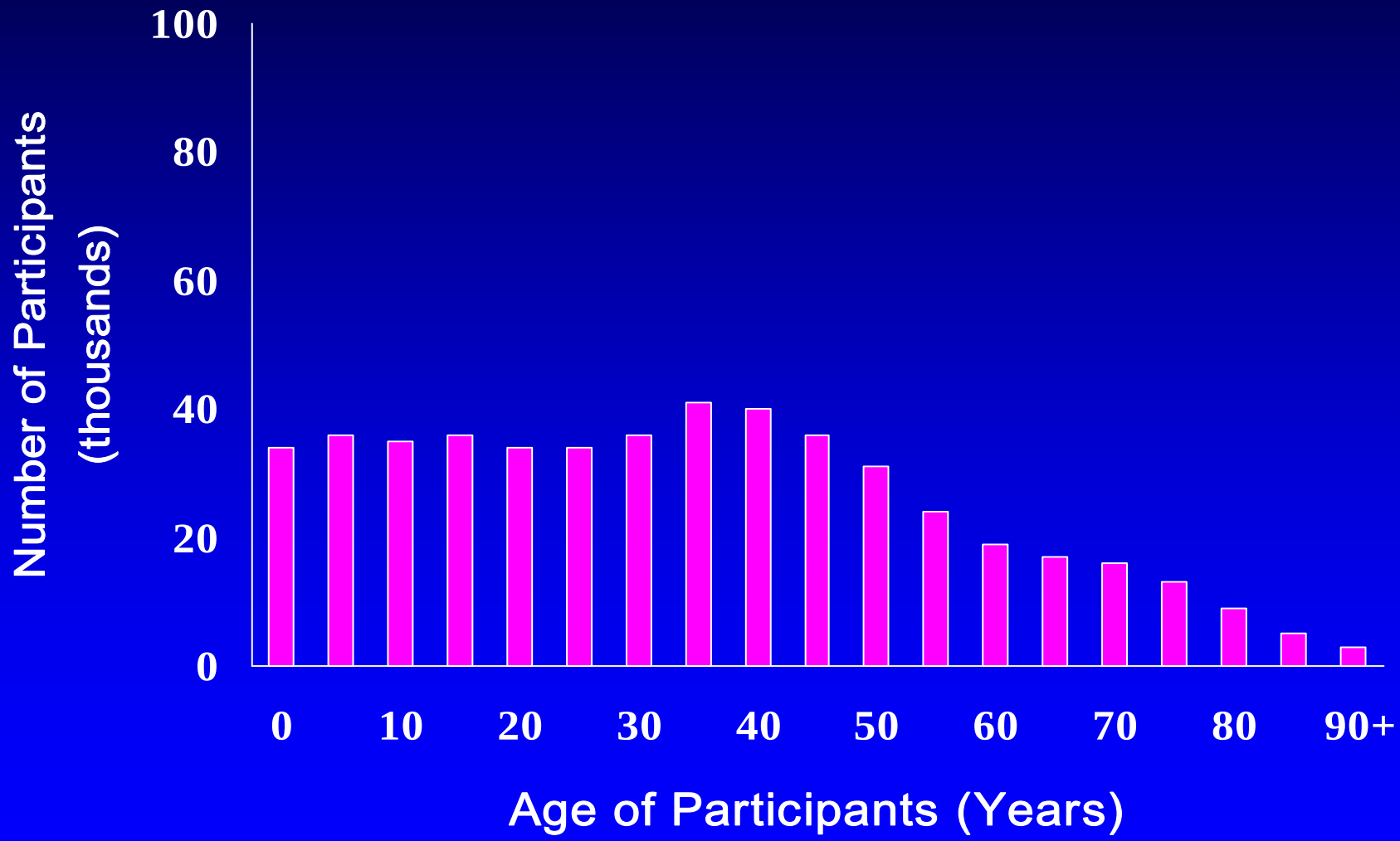
Larson, G. *The Complete Far Side*. 2003.



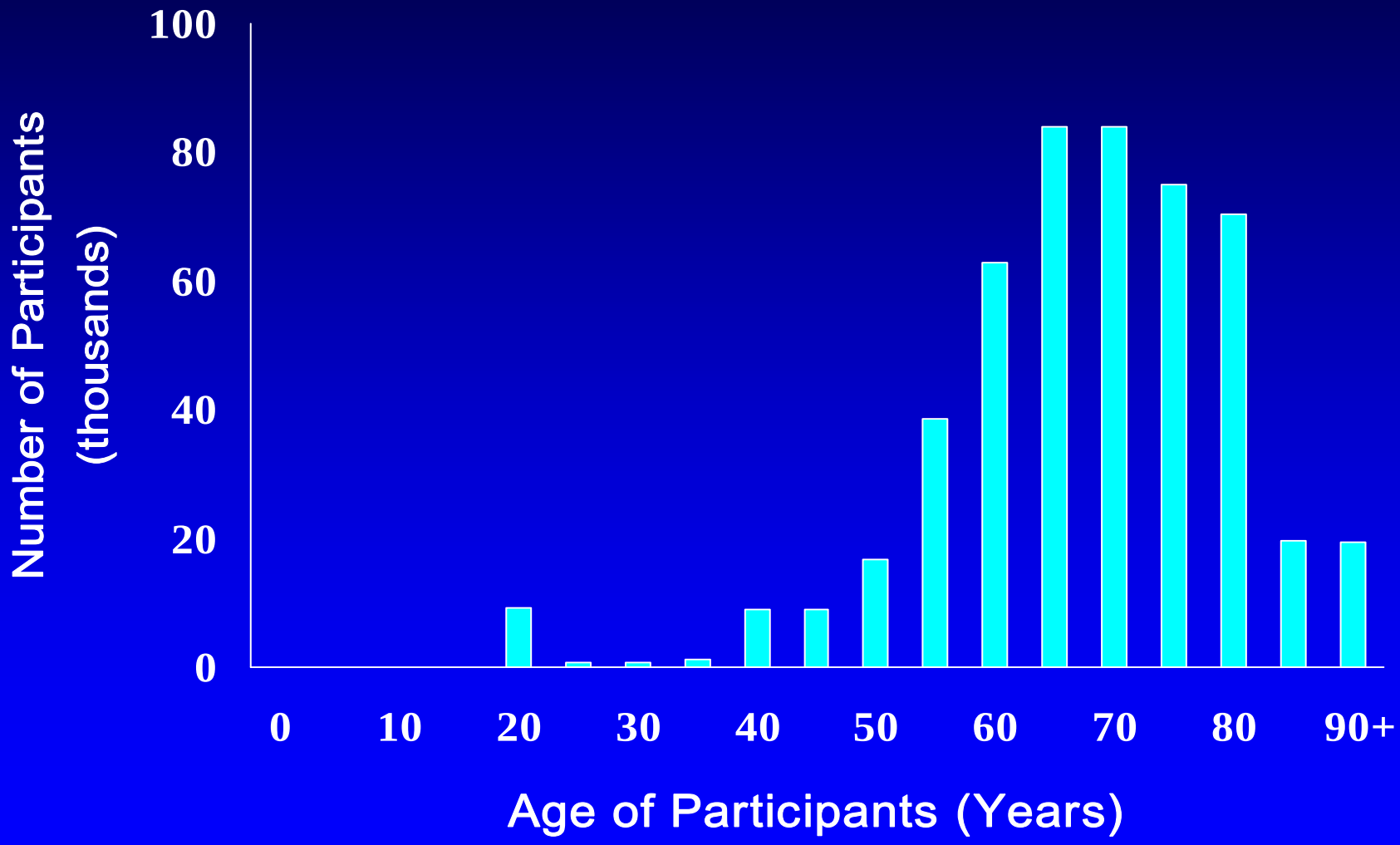
Studies Selected for GAIN Genotyping

PI	Institution	Condition	N (cases/controls)
Gonçalo Abecasis	U Michigan	Psoriasis	1,449/1,450
Steven Faraone	SUNY Syracuse	ADHD	956 (1,912)
Pablo Gejman	Northwestern	Schizophrenia	1,540/1,540 EA; 1,100/1,100 AA
John Kelsoe	UC San Diego	Bipolar I	1,158/ 0 EA; 380/ 0 AA
Patrick Sullivan	UNC Chapel Hill	Major Depression	1,860/1,860
James Warram	Joslin Diabetes Center	Type I DN	453/445

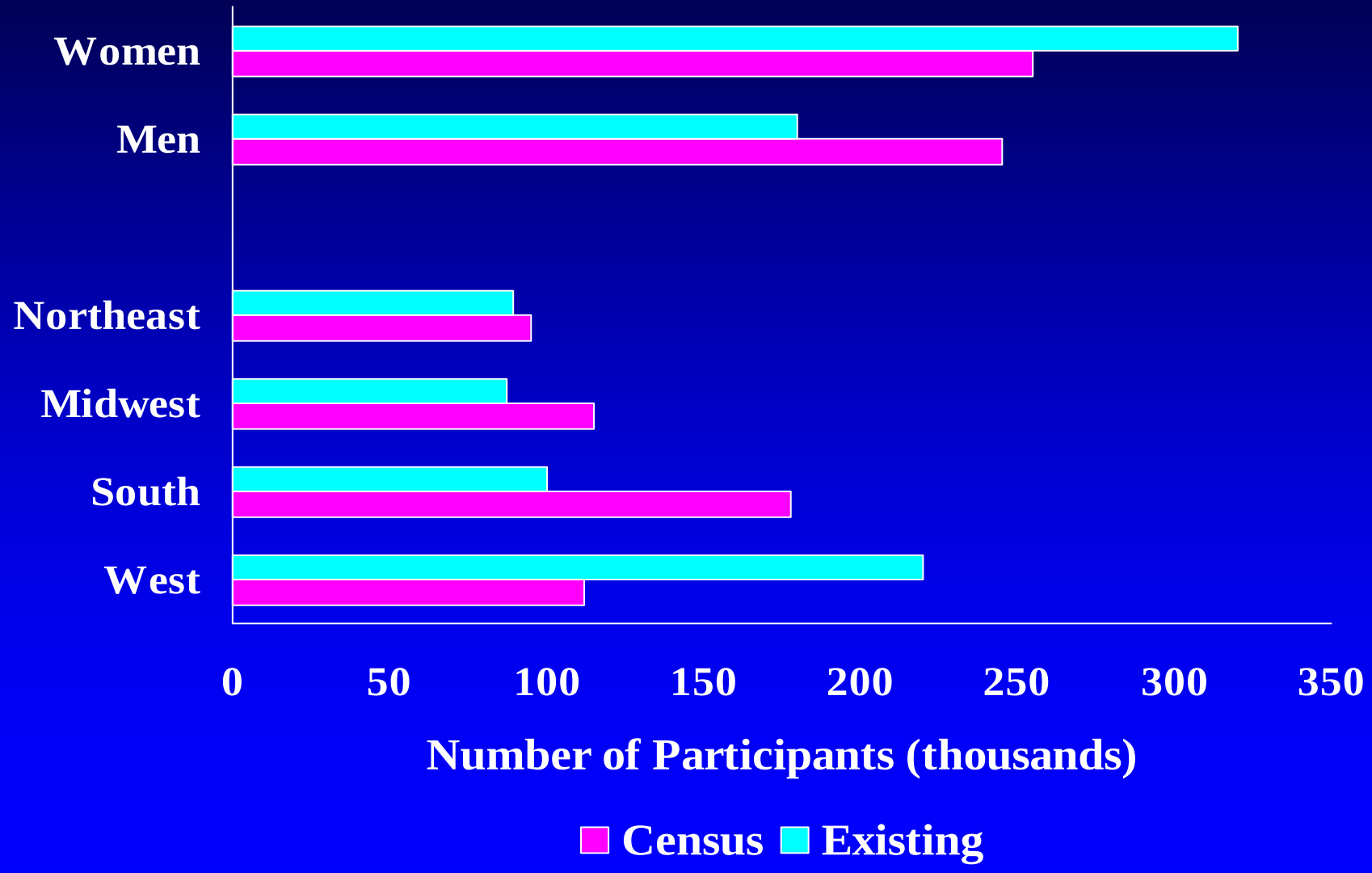
Estimated Age Distribution of Representative US Cohort (2000 Census)



Estimated Age Distribution of Existing NIH-Funded Cohorts



PROJECTED SEX AND REGIONAL DISTRIBUTION OF EXISTING COHORTS AND US CENSUS



PROJECTED EDUCATION DISTRIBUTION OF EXISTING COHORTS AND US CENSUS (Age ≥ 25)