Genetic epidemiology for complex diseases

John P.A. Ioannidis

The revolution of molecular genetics

- Apocalyptic promises of bio-information
- Reductionism
- Discovery-oriented approaches
- Massive data
- Globalization of research
- Analysis still largely based on traditional epidemiological principles

Multifactorial diseases

- For many common, important diseases, it is estimated that 20-80% of the risk at the population level is attributed to genetic risk factors
- Such diseases include, but are not limited to: Alzheimer's disease, schizophrenia, Parkinson's disease, diabetes, many cancers, coronary artery disease, osteoporosis, etc.
- In the large majority of these cases, it is speculated that there are many (2->30) genetic variants that each contributes a small risk towards the disease
- Some of these variants may be impossible to detect with linkage studies.

Typical association studies

- Case-control
- Retrospective cohorts
- Prospective cohorts
- Cross-sectional studies
- Nested studies

It usually boils down to:

- A disease group with specific allele and genotype frequencies
- A control group with specific allele and genotype frequencies

Measures of risk

- Study-level: Odds ratio OR=a*c/b*d; for diseases that are not very common in the population, it is an excellent estimate of population-level risk ratio
- Population: attributable fraction AF= Prev(OR-1)/[1+[Prev(OR-1)]], where Prev=prevalence of the allele of interest

Genome scans vs. association studies of candidate genes

- Genome scan: screening of very large areas of the genome or typically the entire genome
- Identification of relatively extended areas with evidence of linkage based on LOD score
- Further trimming of the candidate area is possible, but arriving at single level is not easy

- Association study: a candidate gene approach, targeting only one or limited number of gene(s) and variants thereof, often SNPs (single nucleotide polymorphisms)
- Target gene may or may not give strong linkage signal in linkage analyses

Whole genome association studies

- Screening of very large number of SNPs covering the whole genome
- Tested SNPs may be hundreds of thousands
- Requires extensive replication across several datasets
- Nominal significance up to p=10(-8)

Family-based vs. association studies

- Family-based
- Related individuals in pedigrees
- Typical design involves parents and children or only siblings
- Typical tests for analysis are the TDT (transmission disequilibrium test) and the sib-TDT or equivalents for quantitative traits (QTDT)

- Population-based
- Unrelated individuals with random selection or per specific eligibility criteria
- Typical design involves cases with the disease and controls without disease
- Typical tests for analysis are the chi-square and variants and ANOVA for quantitative traits with the possibility also for adjusted analyses

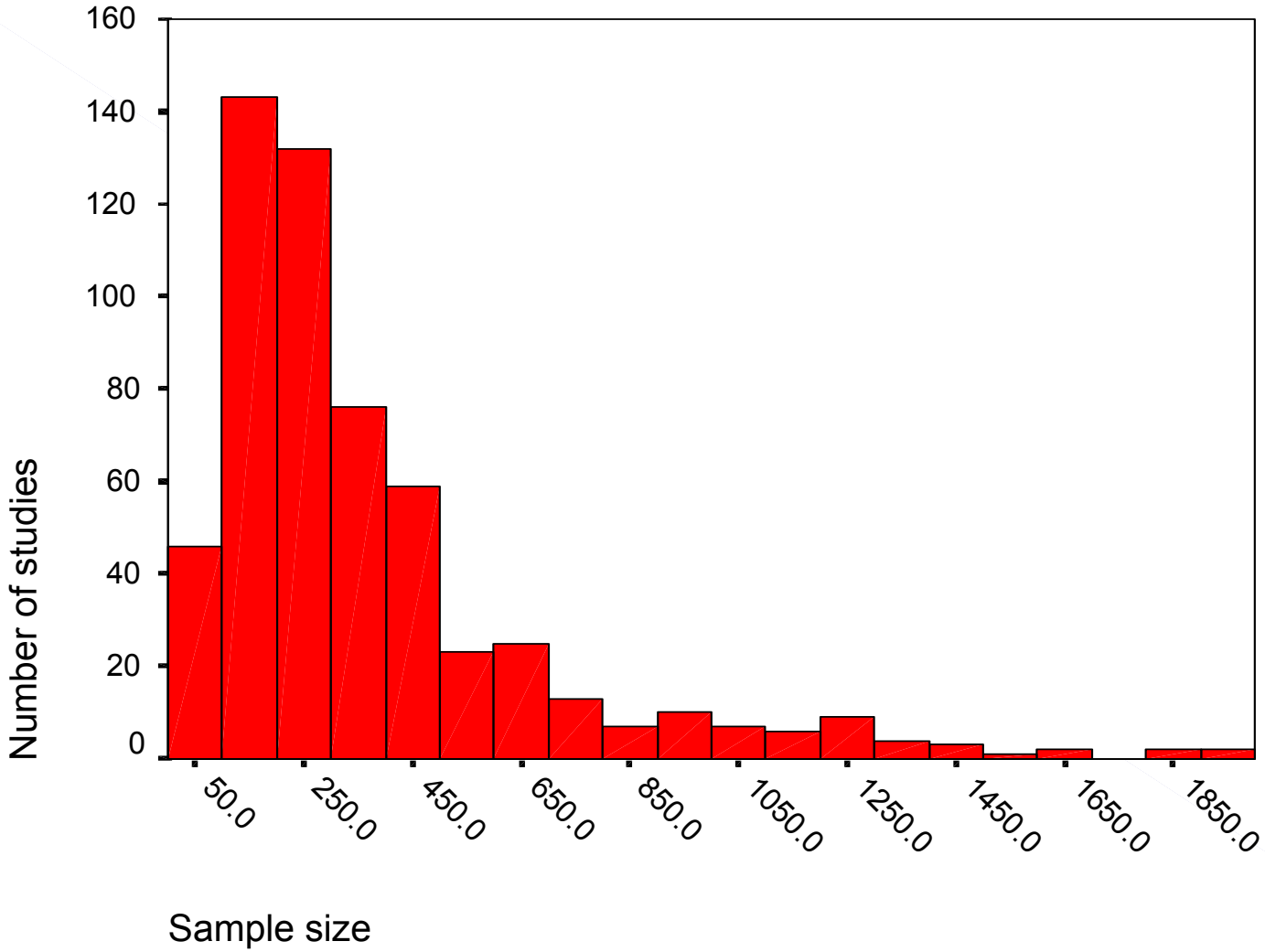
Finding an association could mean different things

- True association
- Linkage disequilibrium with a different, true culprit gene polymorphism in the same gene or a different gene
- Spurious finding due to chance
- Spurious finding due to bias (systematic errors)
- Spurious finding due to both chance and bias

Major postulated problems of molecular genetic studies

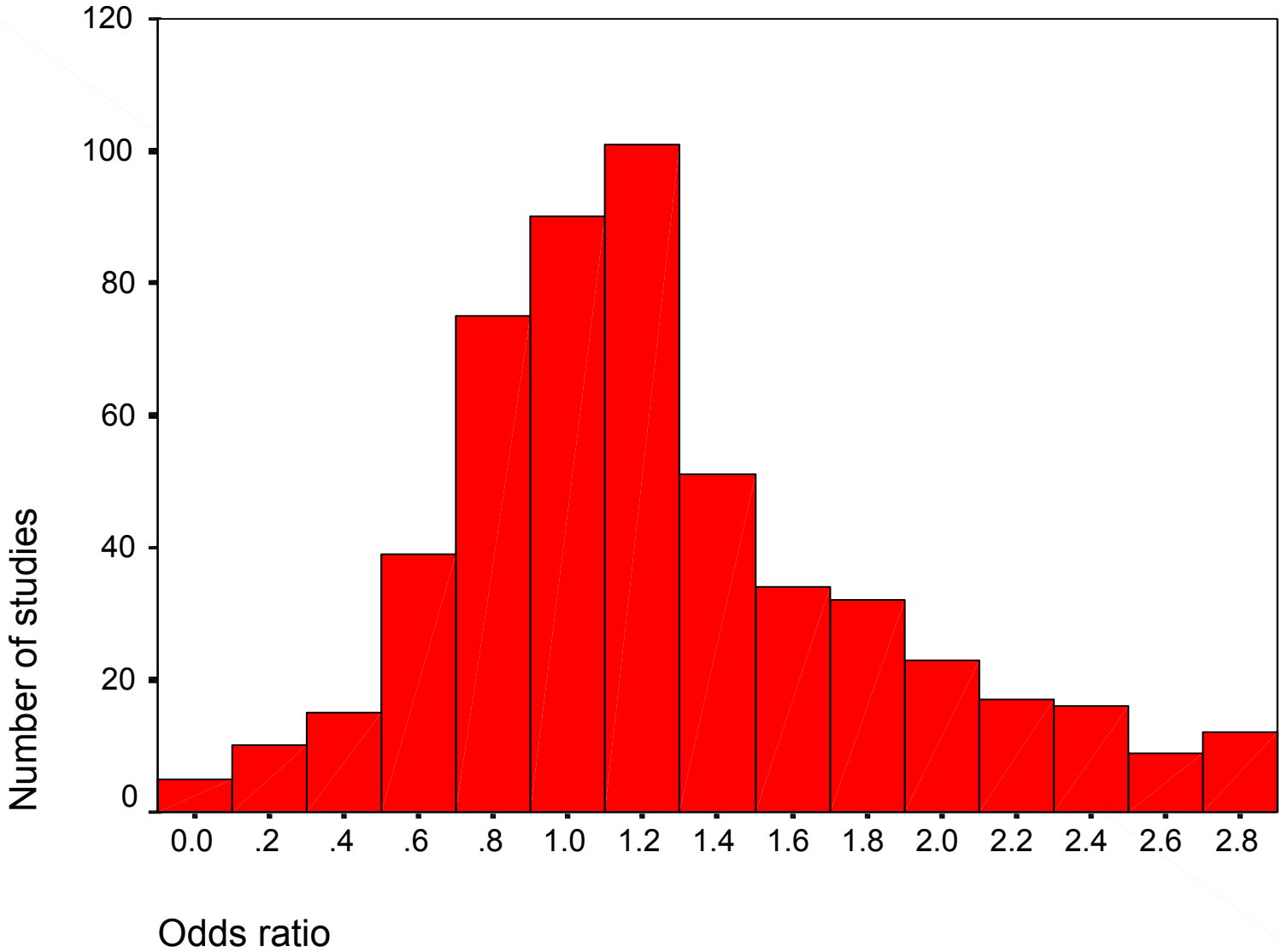
- Small sample sizes
- Small effect sizes
- Large number of genetic variants
- Old-epidemiology problems: confounding, misclassification
- Questionable replication validity

Most studies assessing genetic risk factors are small in terms of sample size



Ioannidis, Trends Mol Med 2003

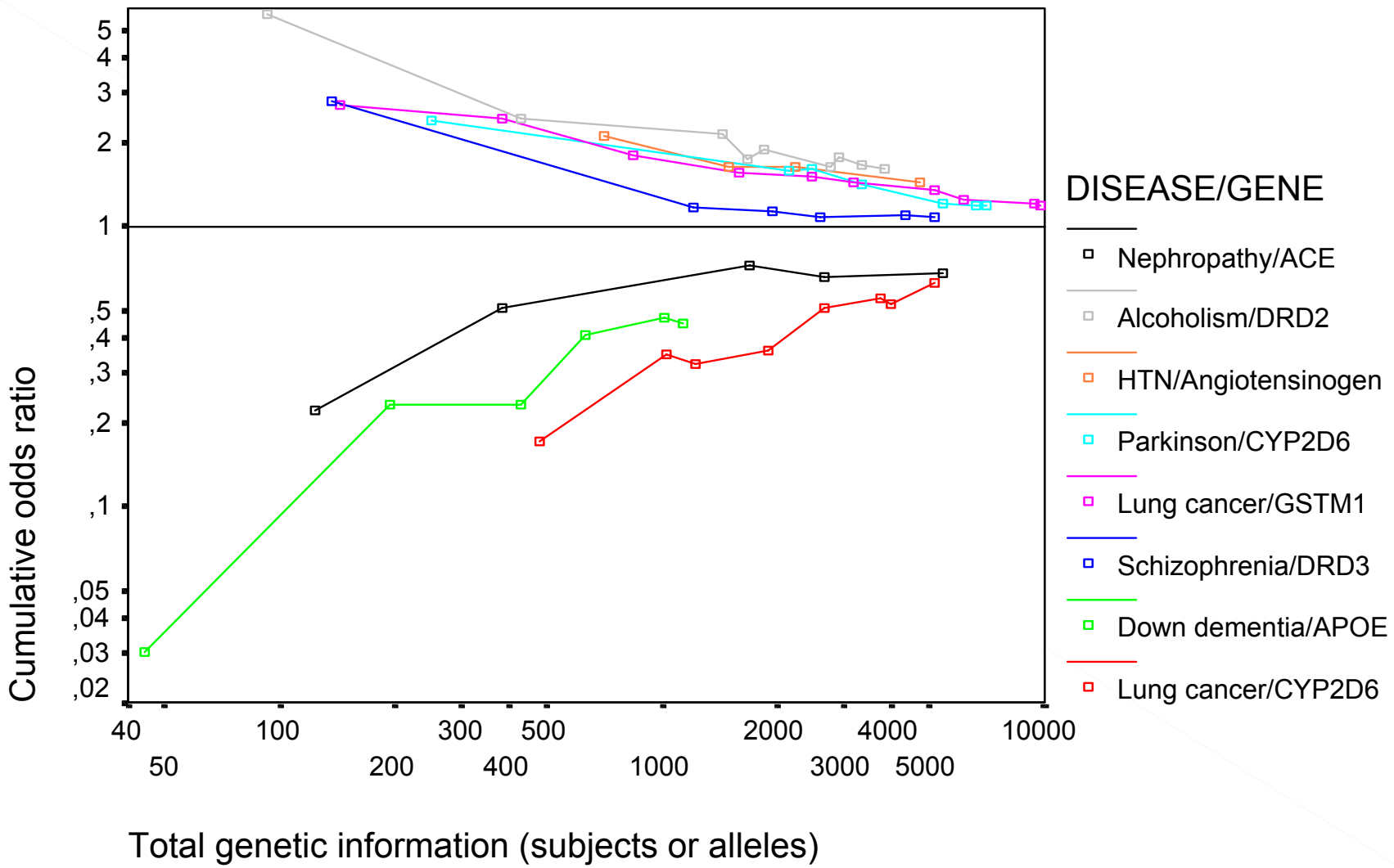
Most genetic effects in multigenetic diseases are small



Complicating factors

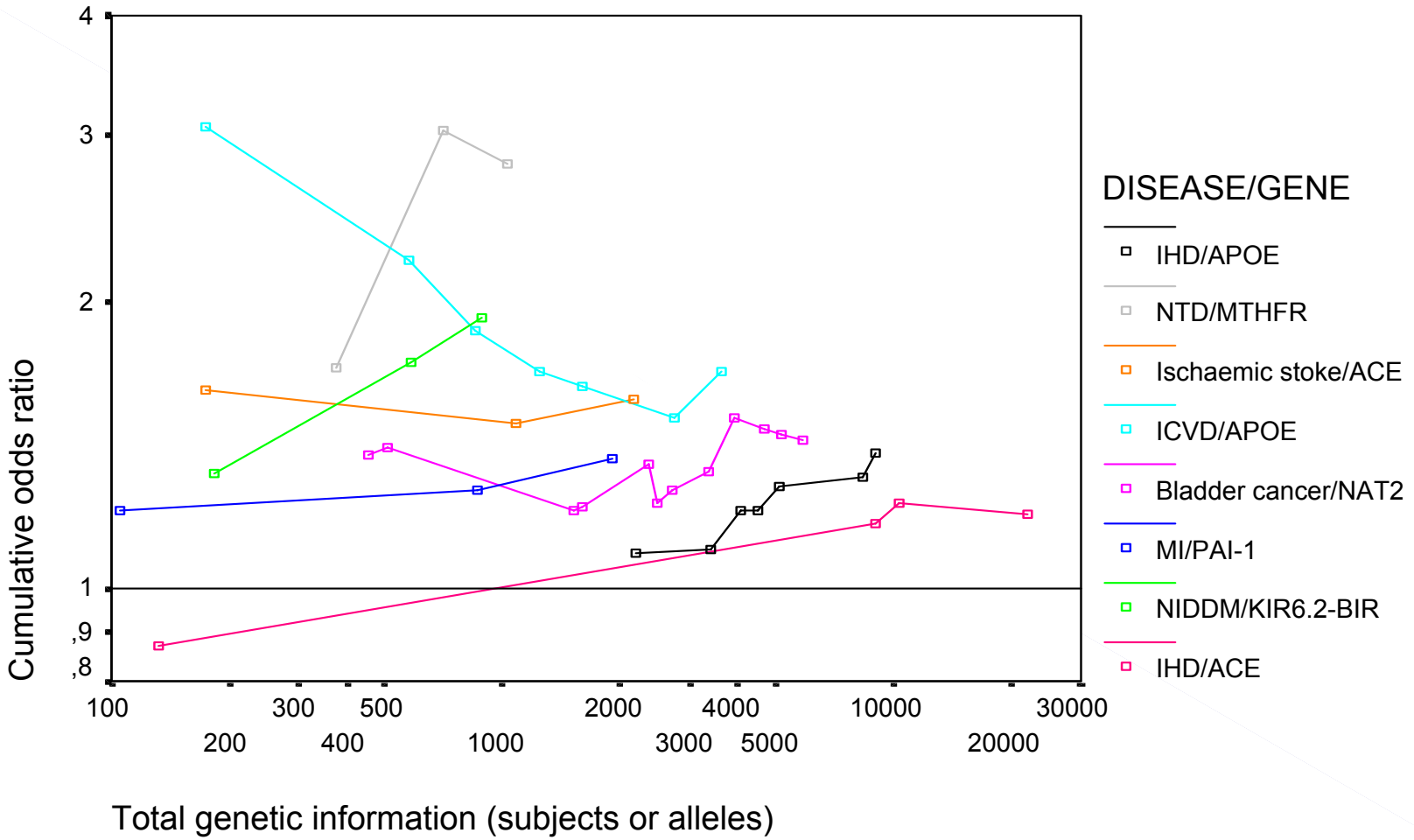
- Too many genes to consider
- Dominant/recessive/co-dominant effects
- Gene-gene interactions
- Gene-environment interactions
- Time-dependent effects
- Measurement errors for genotyping and for clinical and laboratory phenotype
- Unconscious bias
- Conscious bias

Diminishing effects



Ioannidis et al, Nature Genetics 2001

Late-established effects



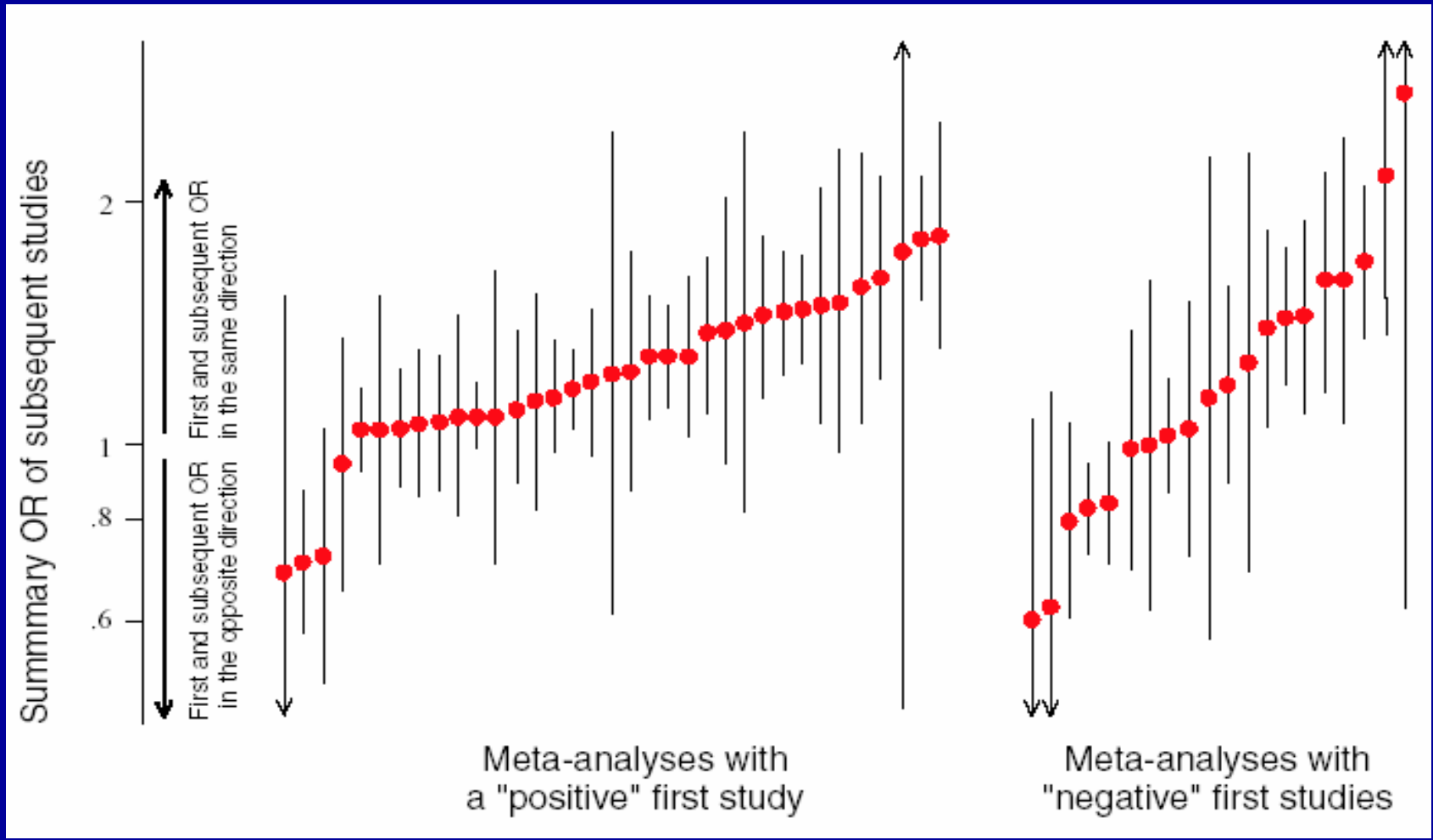
Counting fish in the sea of association analyses

Multiplier	**Parameter**
>10000000	Gene variants
>1000	Diseases
>10	Outcomes
>10	Subgroups
>10	Genetic contrasts
>10	Investigators
1 quadrillion	**Candidate analyses**

The legend of focusing "based on biological plausibility"

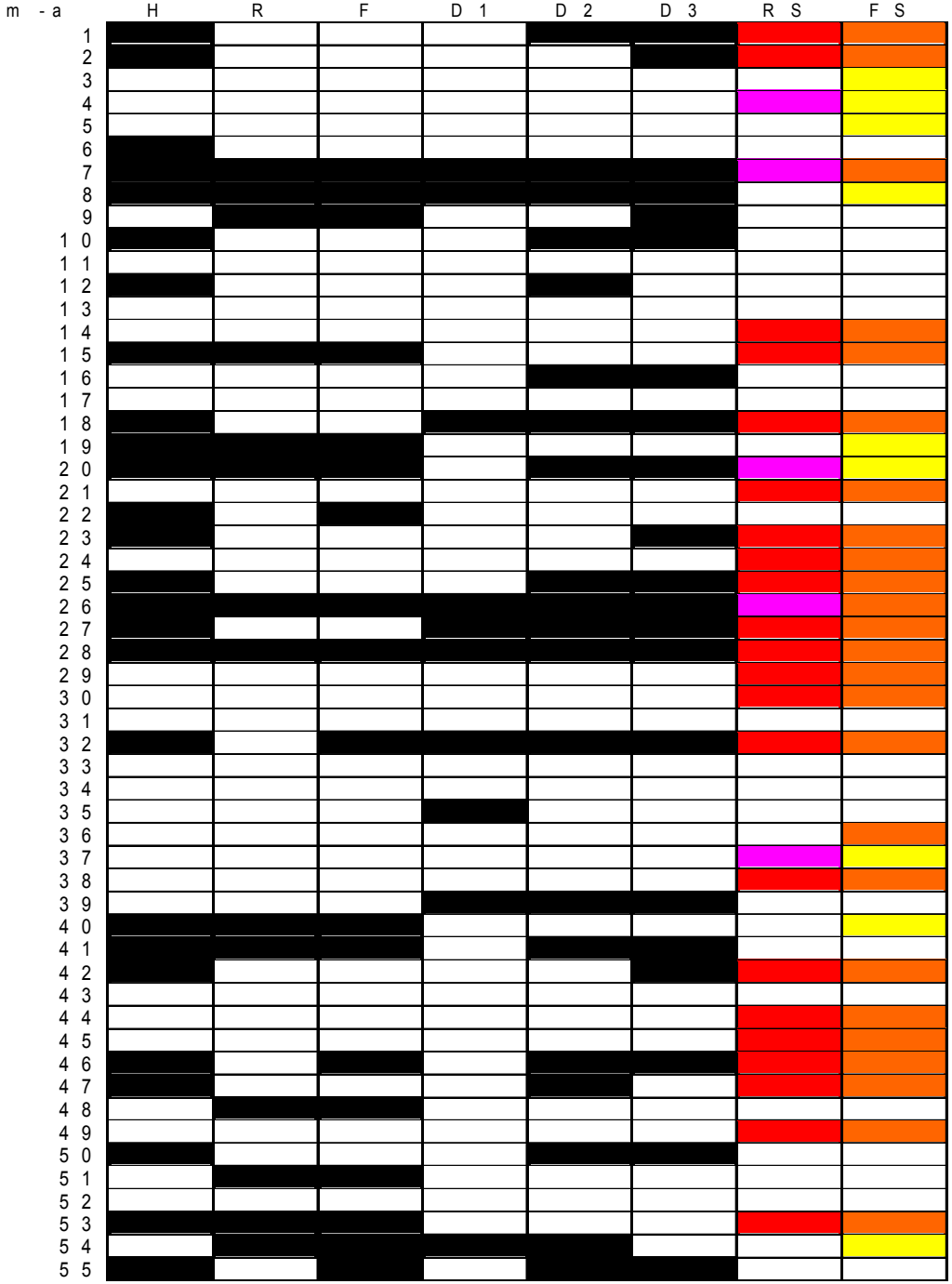
- Just in the year 2002 studies were published addressing the relationship of the APOE epsilon polymorphism with familial Alzheimer's disease; sporadic Alzheimer's disease; colorectal cancer; fatty liver; atherosclerosis; hyperlipidemia; acute ischemic stroke; spina bifida; coronary artery disease; normal tension glaucoma; hypertension; Parkinson's disease, diabetic nephropathy; pre-eclampsia; hepatitis C-related liver disease; cerebrovascular disease; coronary artery disease post-renal transplantation; non-specified cognitive impairment; childhood nephrotic syndrome; spontaneous abortion; multiple sclerosis; alcohol withdrawal; cognitive dysfunction after coronary artery surgery; alcoholic chronic pancreatitis; alcoholic cirrhosis; macular toxicity from chloroquine; macular edema; aortic valve stenosis; vascular dementia; type II diabetes mellitus; and migraine.

Early results mean little



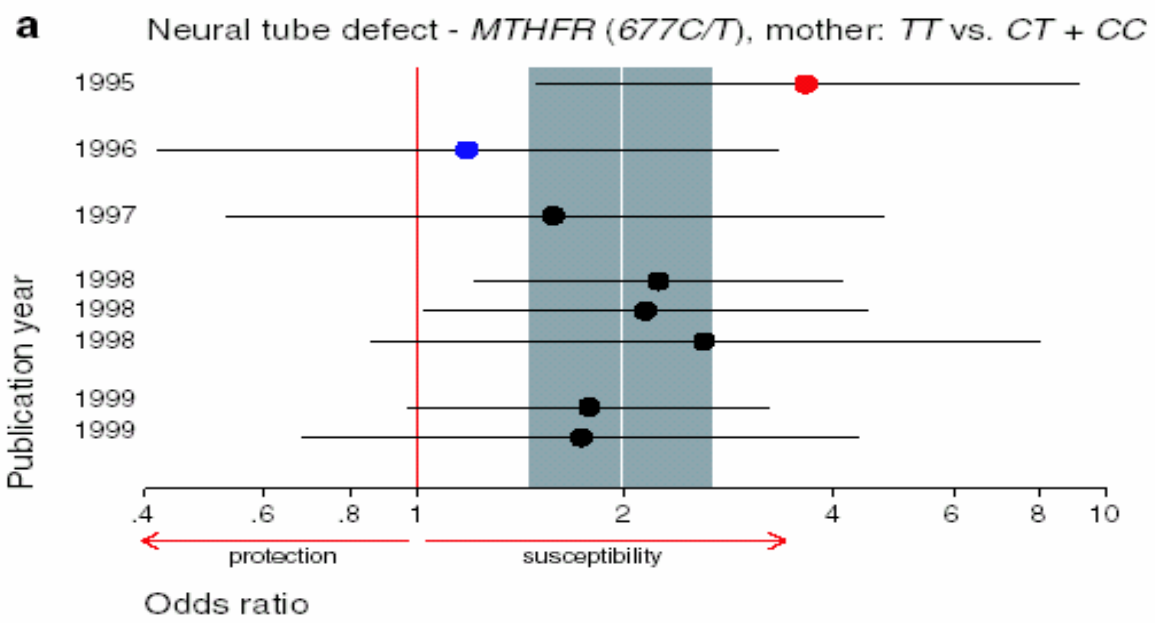
Predictors of statistically significant discrepancies between the first and subsequent studies on the same genetic association.

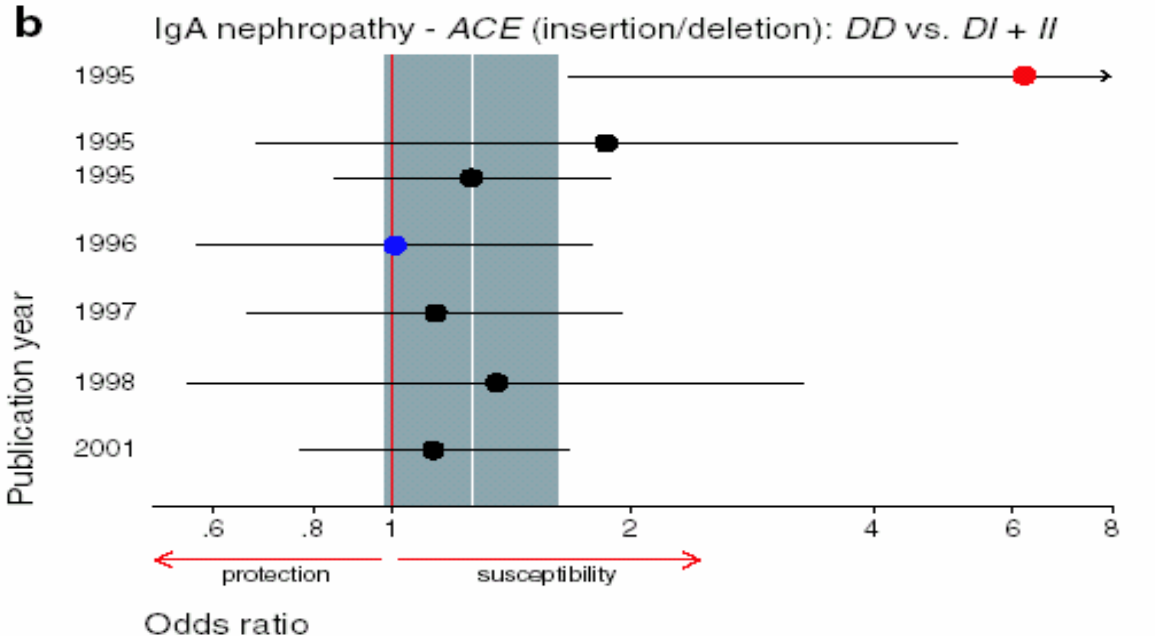
Predictor of discrepancy	Univariate regressions	
	OR (95% CI)	*P*-value
Total number of studies (per study)	1.17 (1.03-1.33)	.020
Sample size of first study(ies) (doubling)	0.42 (0.17-0.98)	.046
Single first study with clear genetic contrast	9.33 (1.01-86.3)	.044



H: heterogeneity
R/F: difference in first vs. subsequent
D1-D3: publication bias diagnostics
RS/FS: significant findings (with/without first studies)

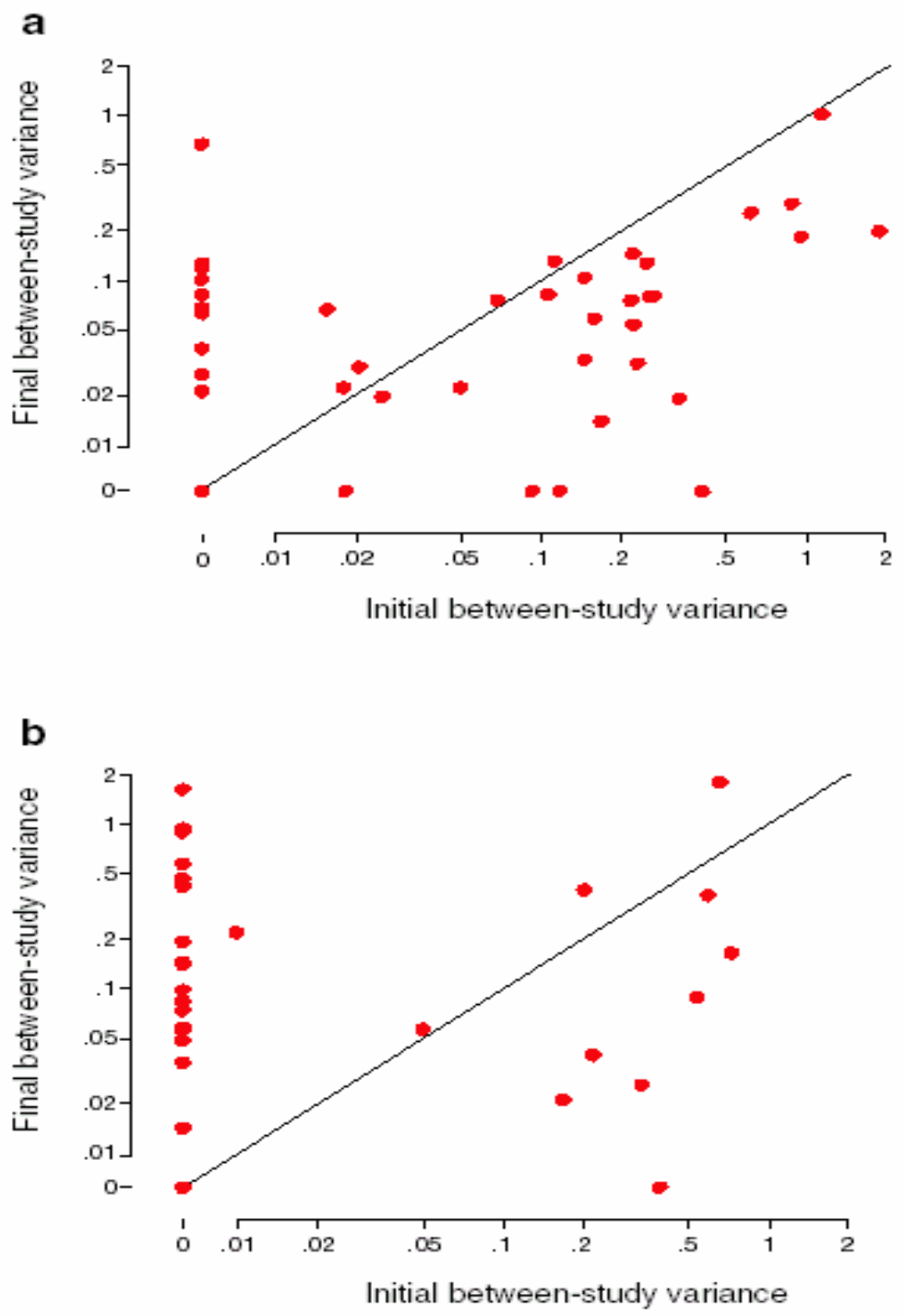
Ioannidis et al, Lancet 2003





Succession of early extremes: the Proteus phenomenon

Ioannidis and Trikalinos, J Clin Epi 2005



Genetics

Changes in between-study variance over time

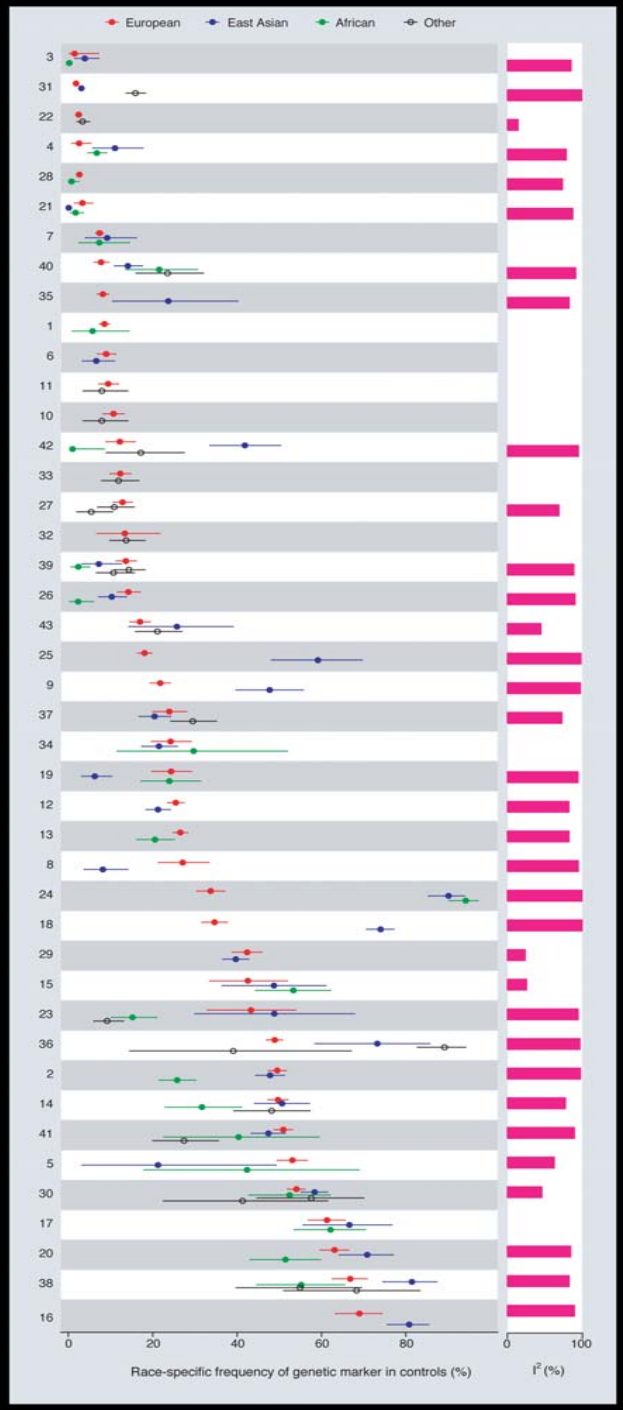
Health care interventions

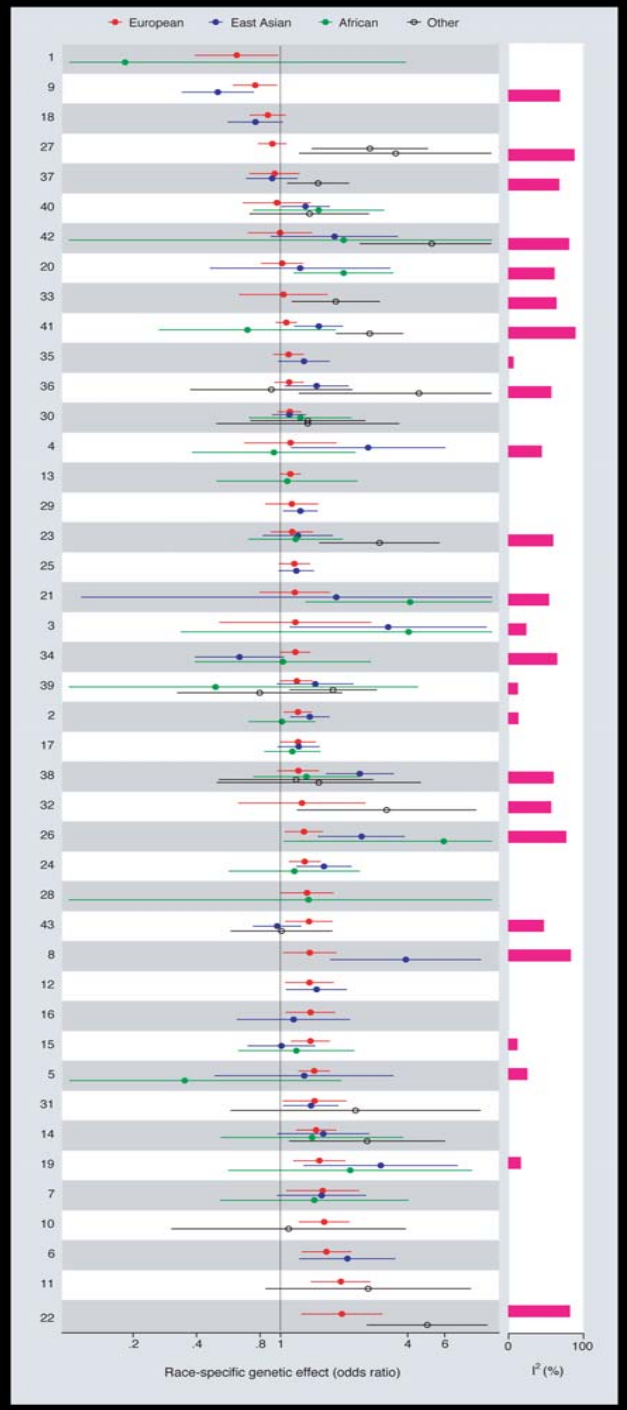
Racial (or other subgroup) differences?

- Empirical evidence suggest that while allele frequencies differ a lot (I-squared≥75%) in 58% of postulated gene-disease associations, differences in the effect sizes (odds ratios) occur in 14%.
- No differences in race-specific odds ratios have been recorded once we have exceeded a total sample size of N=10,000

Ioannidis et al, Nat Genet 2004

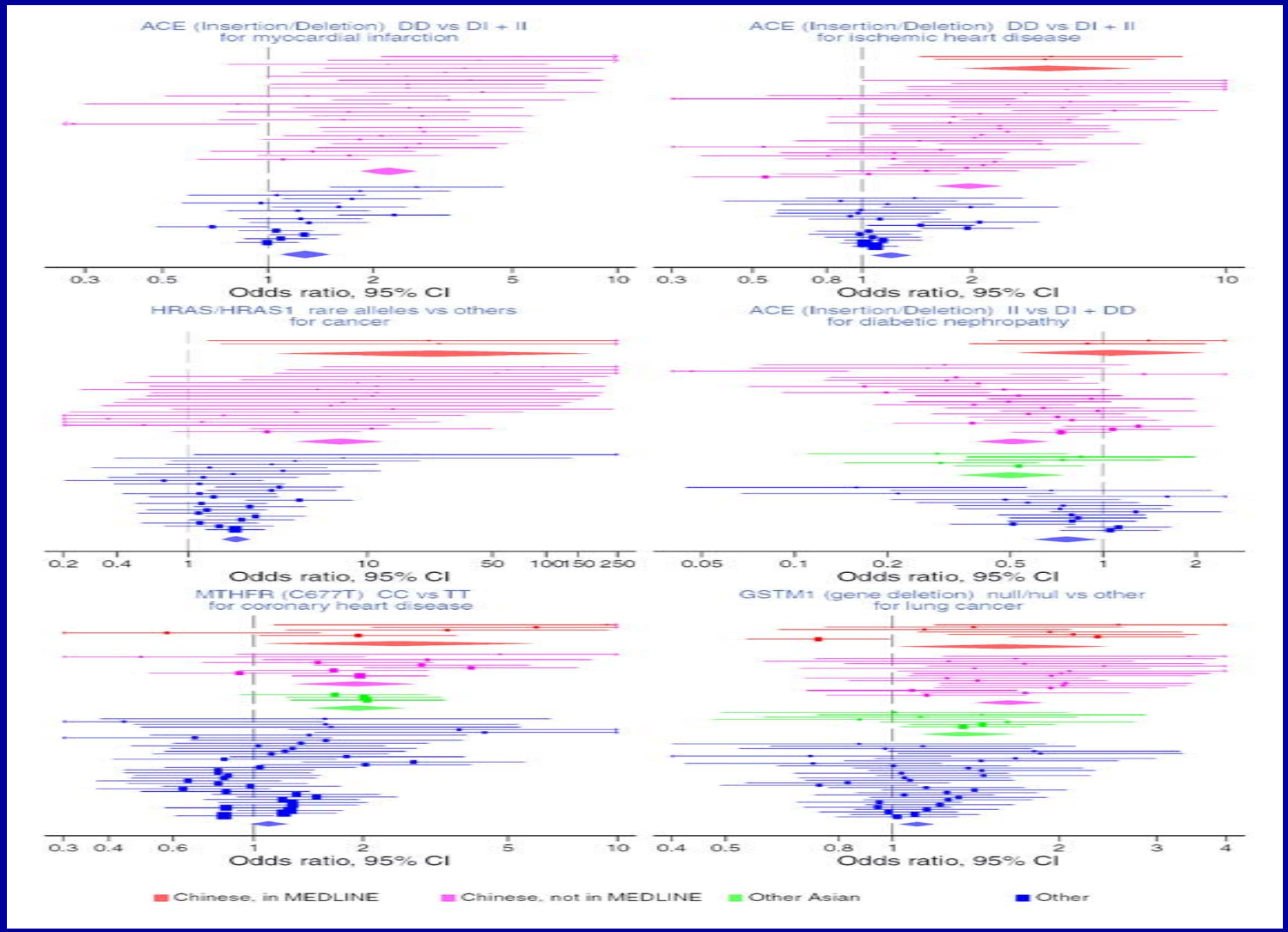
Control rates: $I^2 \geq 75\%$ in 58%





Odds ratios: $I^2 \geq 75\%$ in 14%

Global science?



Problems of standardization

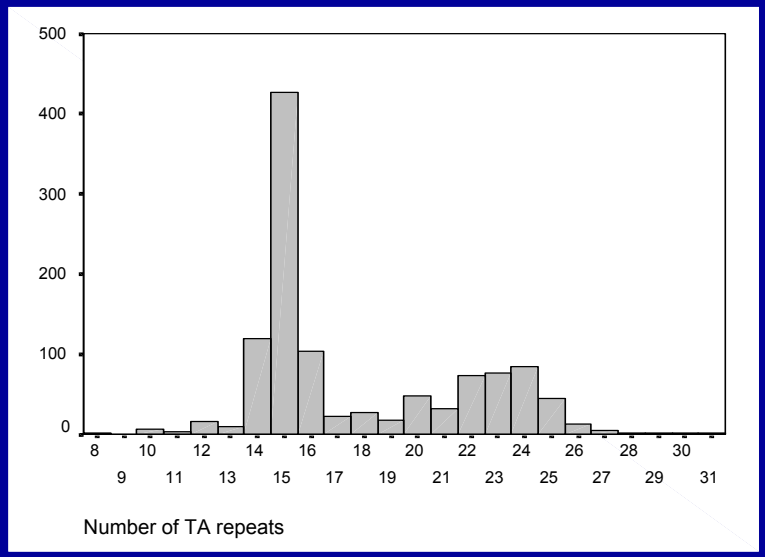
- Polymorphic markers
- Variable techniques
- Time-to-event outcomes
- Multivariate analyses
- Intermediate and surrogate outcomes

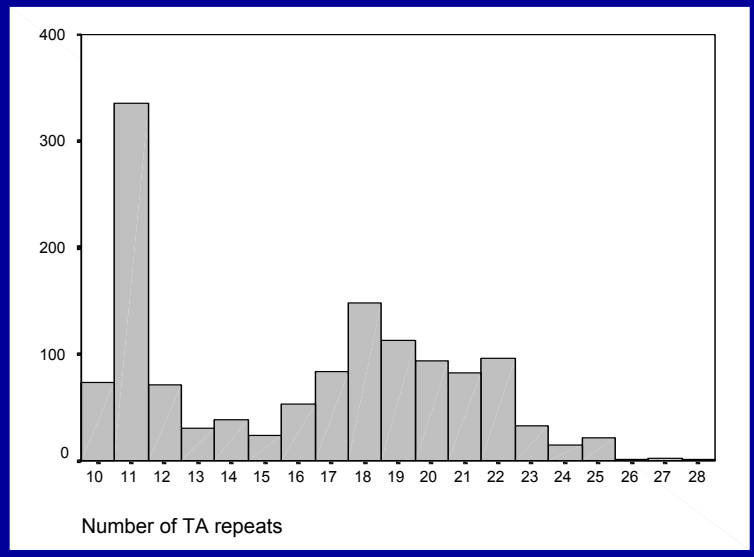
A prospective MIPD: GENOMOS

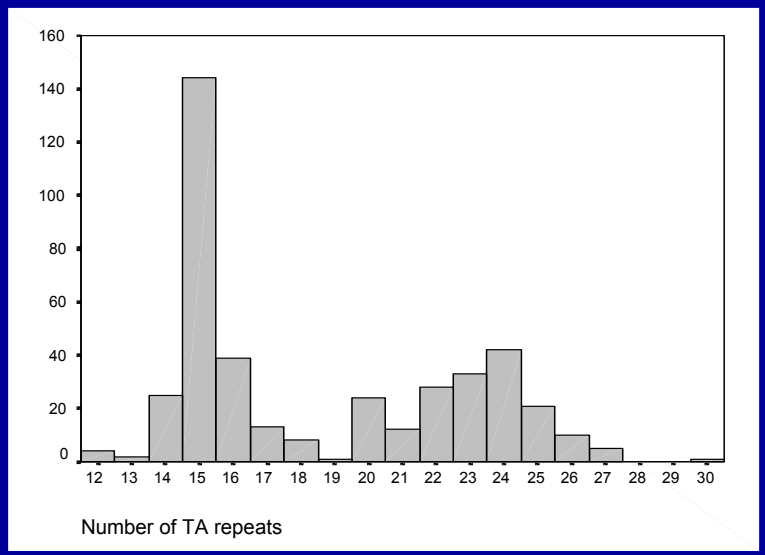
- Meta-analysis of individual-level data on osteoporosis on over 26,000 subjects with prospective genotyping
- 10 teams involved across Europe, several of them multicentric
- A unique opportunity to evaluate the genetics of osteoporosis with rigorous large scale evidence

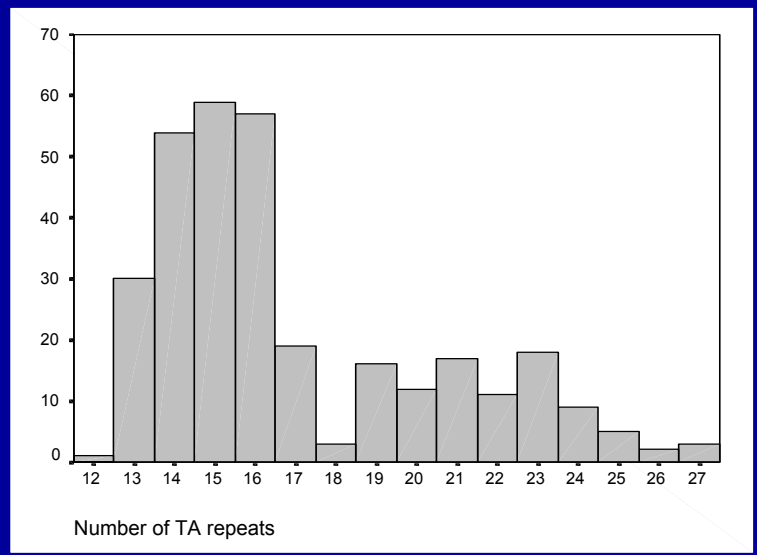
Ioannidis et al, JAMA 2004

Distribution of TA alleles of the ER alpha gene in 4 populations

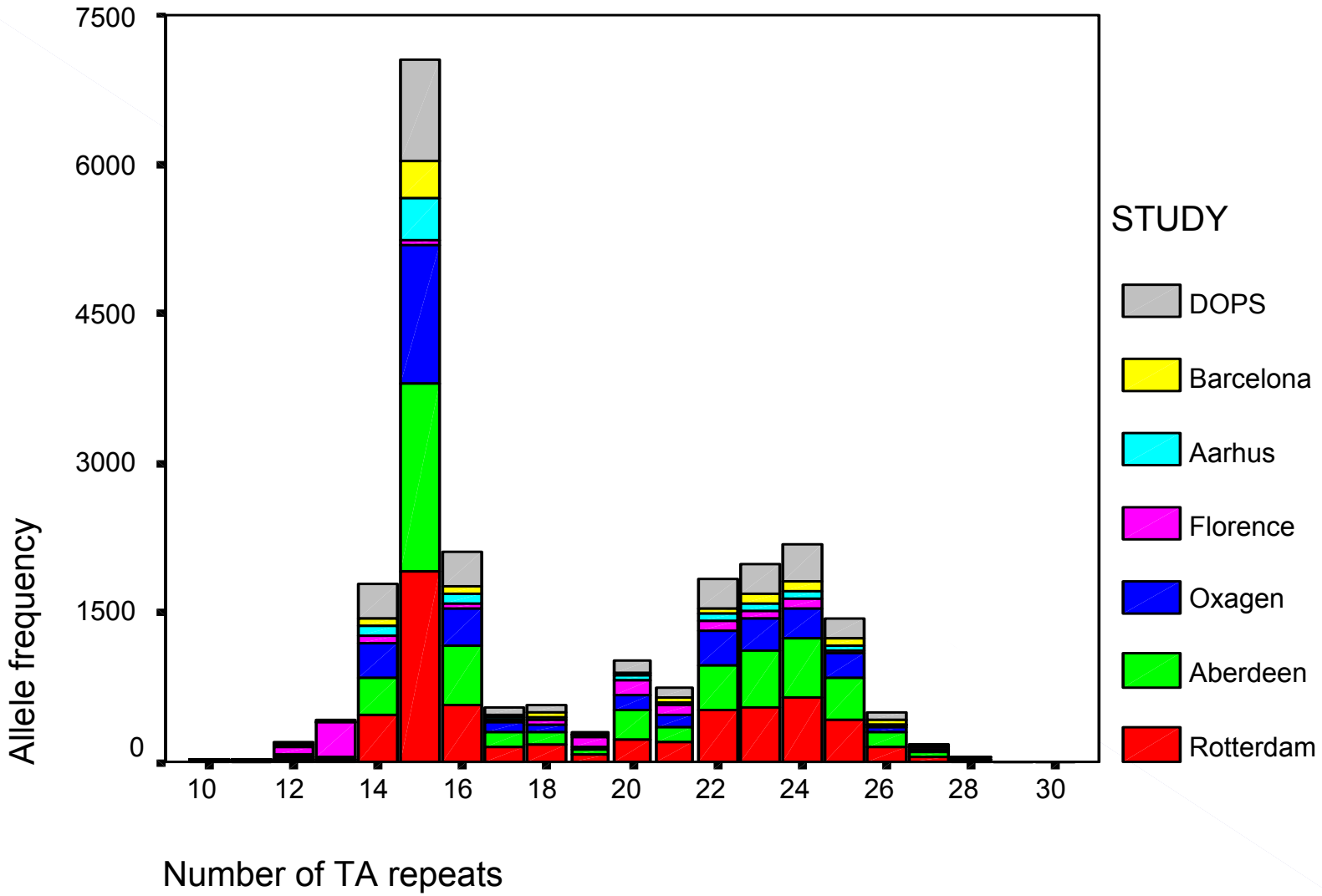








Standardization of genotypes in a prospective MIPD



Other challenges

- Whole genome association meta-analyses
- Whole genome searches meta-analysis

Science at low pre-study odds of true findings

Ioannidis. Why most published research findings are false. PLoS Medicine, 2005

Positive predictive value (PPV) of research findings for various combinations of power (1-β), ratio of true to no relationships (R) and bias (u)

1-β	R	u	Practical example	PPV
0.80	1:1	0.10	Adequately powered RCT with little bias and 1:1 pre-study odds	.85
0.95	2:1	0.30	Confirmatory meta-analysis of good quality RCTs	.85
0.80	1:3	0.40	Meta-analysis of small inconclusive studies	.41
0.20	1:5	0.20	Underpowered, phase I/II well-performed RCT	.23
0.20	1:5	0.80	Underpowered, phase I/II poorly performed RCT	.17
0.80	1:10	0.30	Adequately powered, exploratory epidemiological study	.20
0.20	1:10	0.30	Underpowered, exploratory epidemiological study	.12
0.20	1:1000	0.80	Discovery-oriented exploratory research with massive testing	.0010
0.20	1:1000	0.20	As above, but with more limited bias (more standardized)	.0015

The future: investigator or data specimen registration

- Upfront study registration has been adopted for randomized clinical trials, as a means for minimizing publication and reporting biases and maximizing transparency
- For molecular research, upfront registration in public of all ideas is counter-intuitive and goes against the individualistic spirit of discovery in basic research
- Instead one could aim for registries of investigators and data specimen collections

Ioannidis et al. Am J Epidemiol Aug 15, 2005.

Registries of data/sample collections

- Inclusive networks of investigators working on the same disease, set of genes or field
- Promotion of better methods and standardization
- Research freedom for individual participating teams
- Thorough and unbiased testing of proposed hypotheses with promising preliminary data on large-scale comprehensive databases
- Due credit to investigators for both “positive” and “negative” findings
- It is feasible to start from existing coalitions of investigators (“networks”) that work on specific diseases, genes or fields

Grading the credibility of molecular evidence

- First axis: Effect size
- 1.1 Very small or small effect size (relative risk<2)
- 1.2 Moderate effect size (relative risk 2-5)
- 1.3 Large effect size (relative risk>5)
- Second axis: Amount and replication of evidence
- 2.1. Single or scattered studies
- 2.2. Meta-analyses of group data
- 2.3. Large-scale evidence from inclusive networks
- Third axis: Protection from bias
- 3.1 Clear presence of strong bias in the evidence
- 3.2 Uncertain about the presence of bias
- 3.3 Clear strong protection from bias
- Fourth axis: Biological credibility
- 4.1 No functional/biological data or negative data
- 4.2 Limited or controversial functional data
- 4.3 Convincing functional data
- Fifth axis: Relevance
- 5.1 No clinical or public health applicability
- 5.2 Limited clinical or public health applicability
- 5.3 Considerable clinical/public health applicability

Ioannidis, Int J Epidemiol, in press