Discovering Variants Conferring Risk for Common Diseases

Michael Boehnke
University of Michigan

Future Opportunities for
Genome Sequencing and Beyond:
A Planning Workshop for the NHGRI

July 28-29, 2014

Introduction

- Central goal of genomics: understand the genetic basis of human disease and use this knowledge to improve human health
- Serious advance toward this goal a reasonable aim for the next 5 years
- Mendelian diseases (Rod McInnis)
  - great importance in their own right
  - can be immensely instructive to understand genetic basis of common diseases
  - Mendelian and common two ends of a continuum

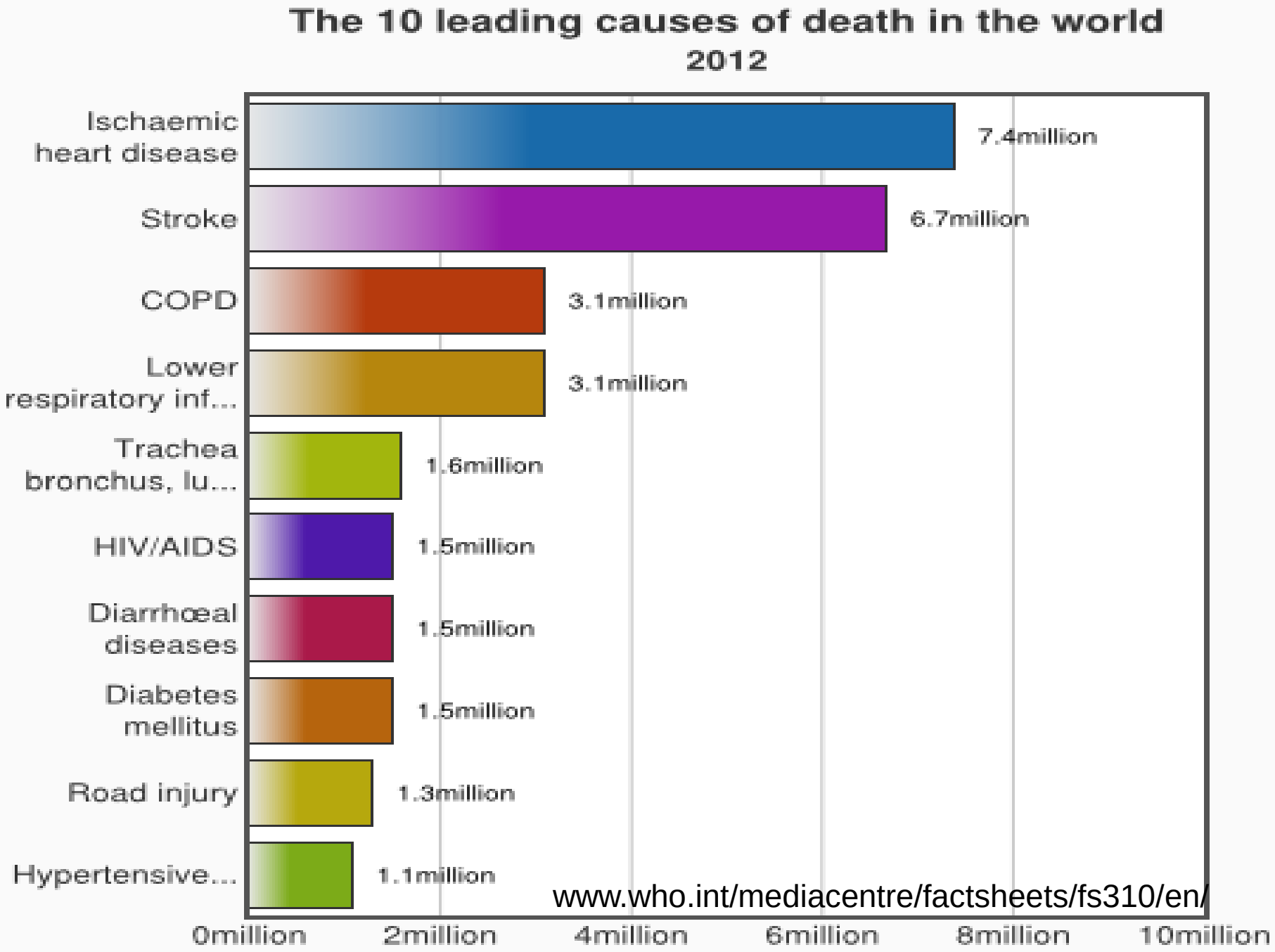
Introduction (continued)

- Common diseases are responsible for the large majority of human morbidity and mortality
- Now is a good time for planning
  - substantial experience with common variant GWAS
  - results of large-scale sequencing studies beginning to emerge

Questions posed by organizers

- What are the big problems that can be solved?
- What will it take to solve these problems comprehensively?
- What will happen if NHGRI decides not to pursue this area?

What are the big problems that can be solved?


The 10 leading causes of death in the world
2012
Ischaemic heart disease
7.4million
Stroke
6.7million
COPD
3.1million
Lower respiratory inf...
3.1million
Trachea bronchus, lu...
1.6million
HIV/AIDS
1.5million
Diarrhœal diseases
1.5million
Diabetes mellitus
1.5million
Road injury
1.3million
Hypertensive...
1.1million
www.who.int/mediacentre/factsheets/fs310/en/
0million
2million
4million
6million
8million
10million

What will it take to do this comprehensively?

- Explore the full spectrum of human variation for all common human diseases to
  - provide a better understanding of human biology and disease etiology
  - suggest targets for therapies and allow better targeting of therapies
  - improve risk prediction

Progress in identifying gene variants for common traits

Cholesterol
Obesity
Myocardial infarction
QT interval
Atrial fibrilliation
Type 2 diabetes
Prostate cancer
Breast cancer
Colon cancer
Height

Age related macular degeneration
Crohn's disease
Type 1 diabetes
Systemic lupus erythematosus
Asthma
Restless leg syndrome
Gallstone disease
Multiple sclerosis
Rheumatoid arthritis
Glaucoma

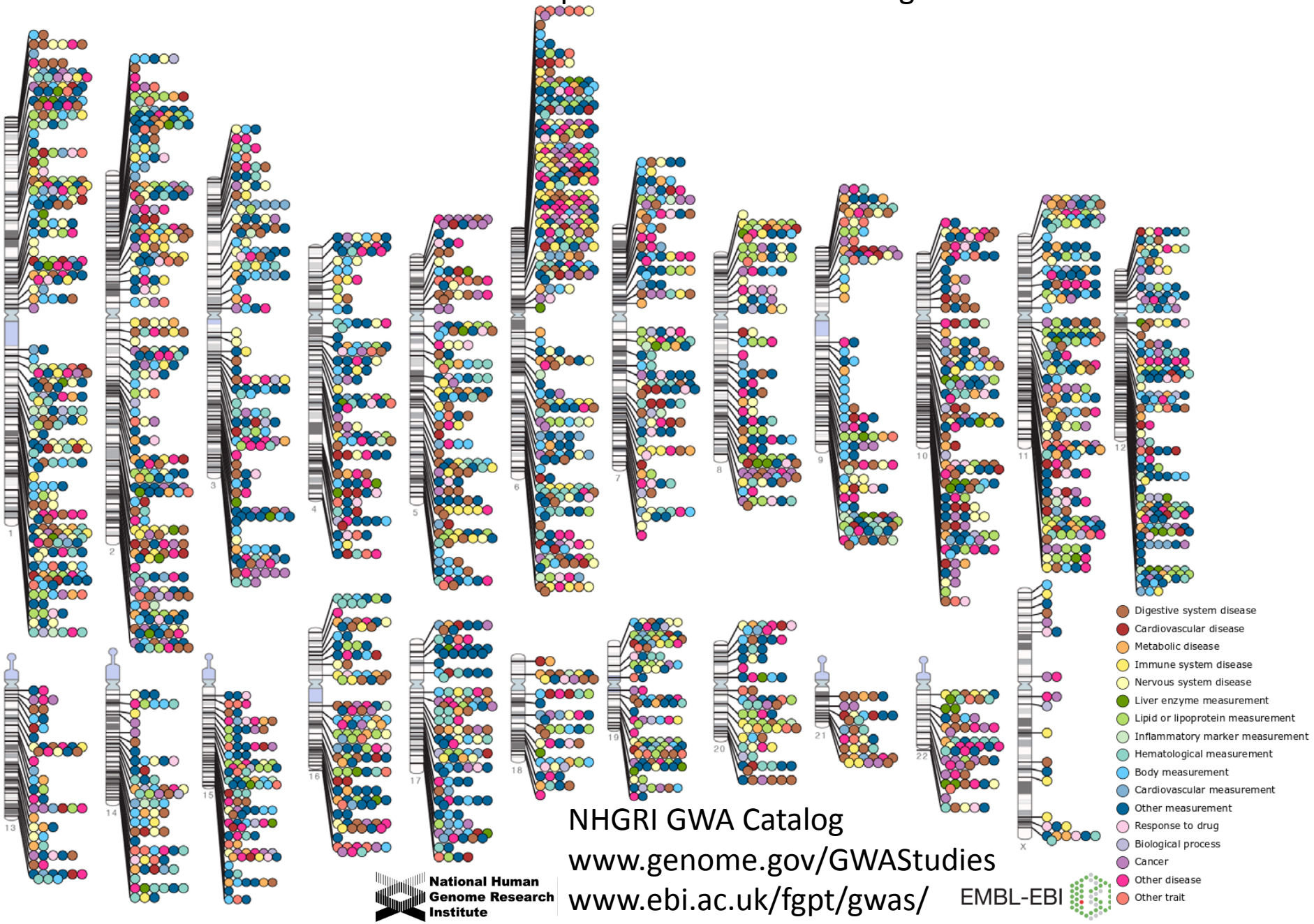
PPAR
IBD5
NOD2
CTLA4
KCNJ11
PTPN22
CD25
IRF5
PCSK9
CFH
NOS1AP
IFIH1
PCSK9
CFB/C2
LOC3877
15
8q24
IL23R
TCF7L2
8q24 #2
8q24 #3
8q24 #4
8q24 #5
8q24 #6
ATG16L1
5p13
10q21
IRGM
NKX2-3
IL12B
3p21
1q24
PTPN2
CDKN2B/A
TCF2
IGF2BP2
CDKAL1
HHEX
SLC30A8
MEIS1
LBXCOR1
BTBD9
C3
8q24
ORMDL3
4q25
TCF2
GCKR
FTO
C12orf30
ERBB3
KIAA0350
CD226
16p13
PTPN2
SH2B3
FGFR2
TNRC9
MAP3K1
LSP1
8q24
HMGA2
GDF5-
UQCC
HMPG
JAZF1
CDC123
ADAMTS9
THADA
WSF1
LOXL1
IL7R
TRAF1/C5
STAT4
ABCG8
GALNT2
PSRC1
NCAN
TBL2
TRIB1
KCTD10
ANGLPT3
GRIN3A

2000 2001 2002 2003 2004 2005 2006 2007

Slide courtesy of David Altshuler

Published Genome-Wide Associations through 12/2012

Published GWA at $p \leq 5X10^{-8}$ for 17 trait categories



NHGRI GWA Catalog
www.genome.gov/GWAStudies
www.ebi.ac.uk/fgpt/gwas/

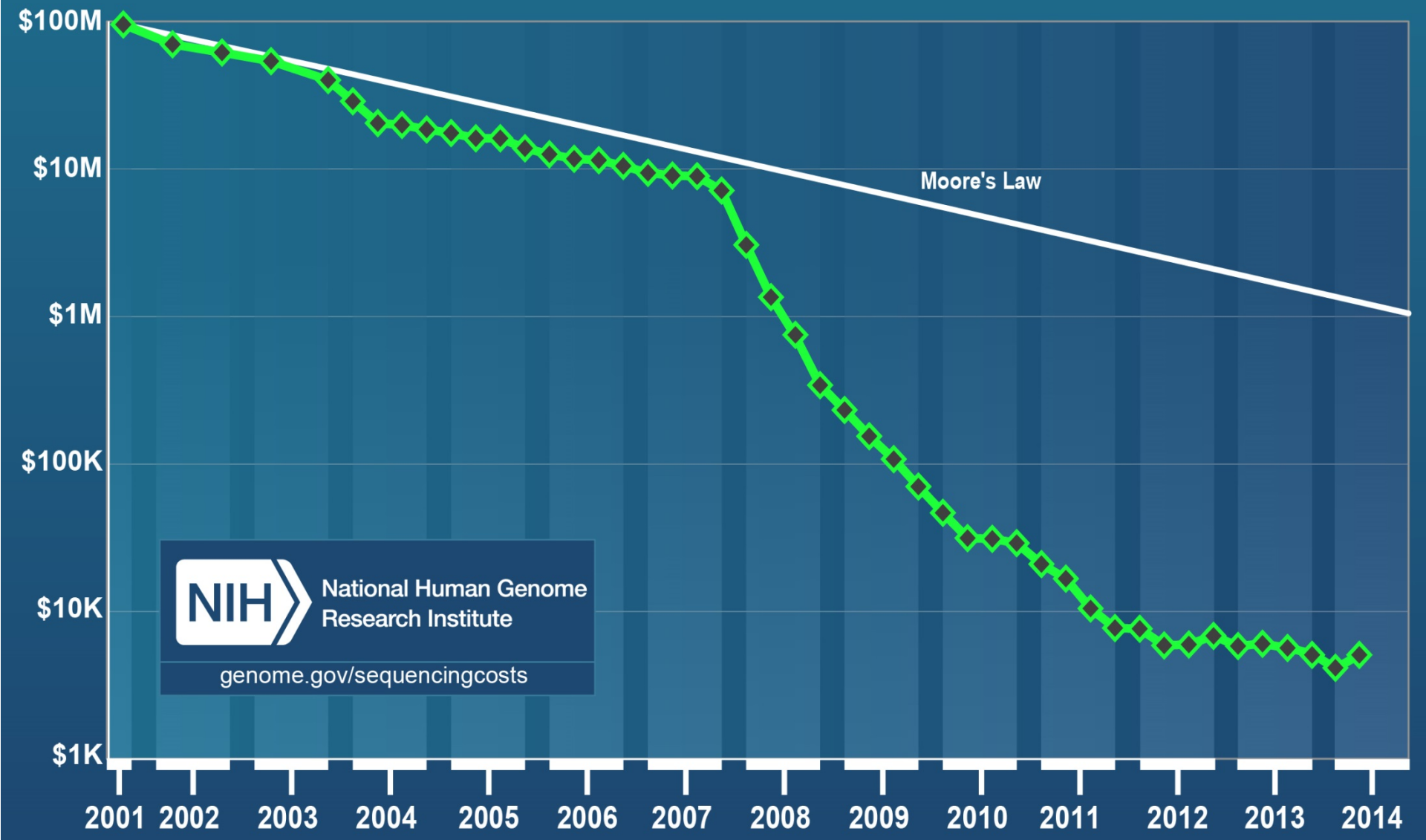
National Human Genome Research Institute

EMBL-EBI

Explore the full allele frequency spectrum

- We have made a start, but much more to do in discovery genomics
- Common variants explain only portion of disease heritability; for most diseases, $h^2 < 50\%$
- At only a few risk loci is gene, direction of effect, mechanism, impact on physiology identified
- Low-frequency variants will help understand many of these loci and remainder of genome
  - extent, effect size distribution now being revealed
  - potential to suggest function, lead to druggable targets, clinical action

Cost per Genome



Key lesson: sample size

- Sample size has been the key determinant of success in common disease genetics to date

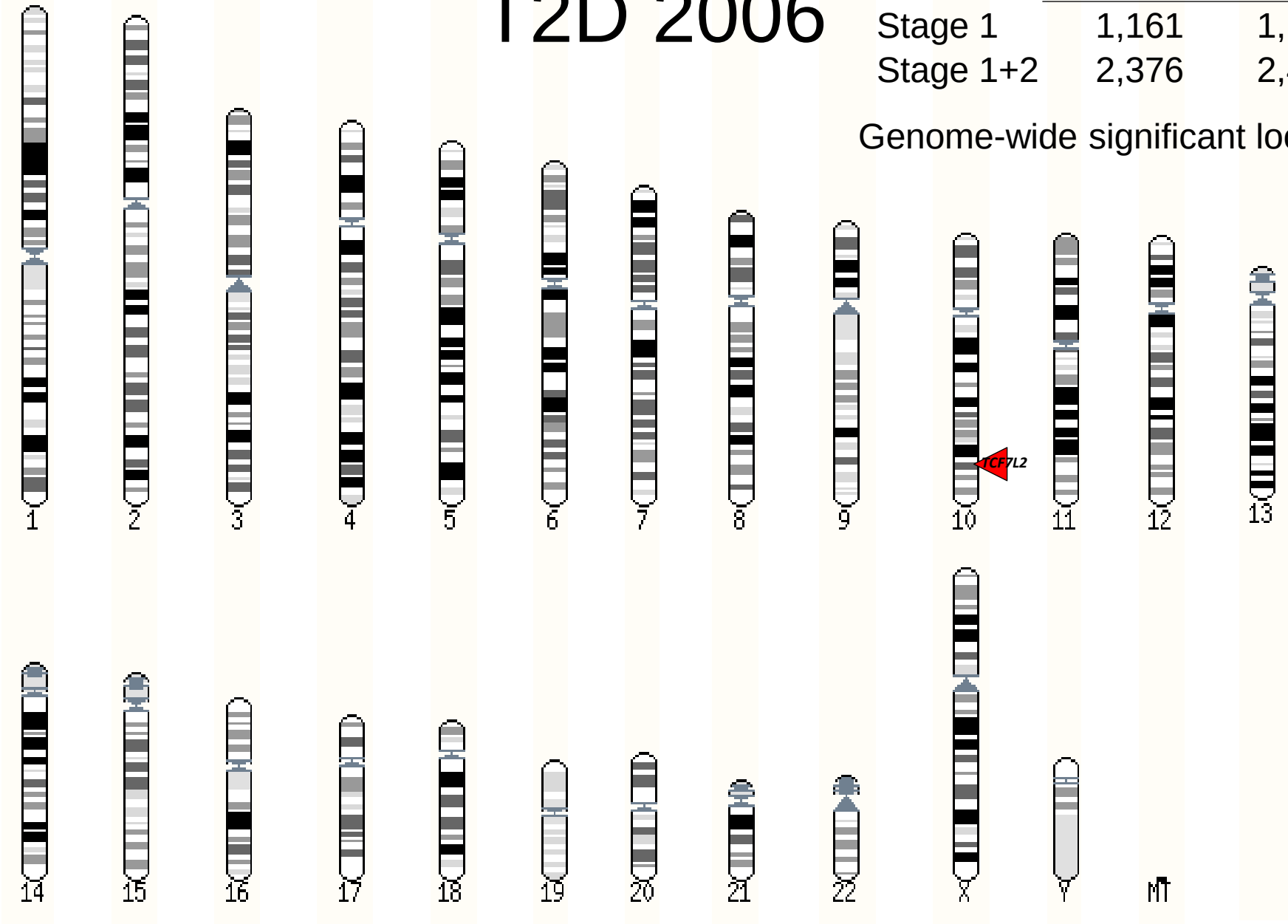
  "Location, location, location"

- Example: type 2 diabetes (T2D)

T2D 2006

	Cases	Controls
Stage 1	1,161	1,174
Stage 1+2	2,376	2,432

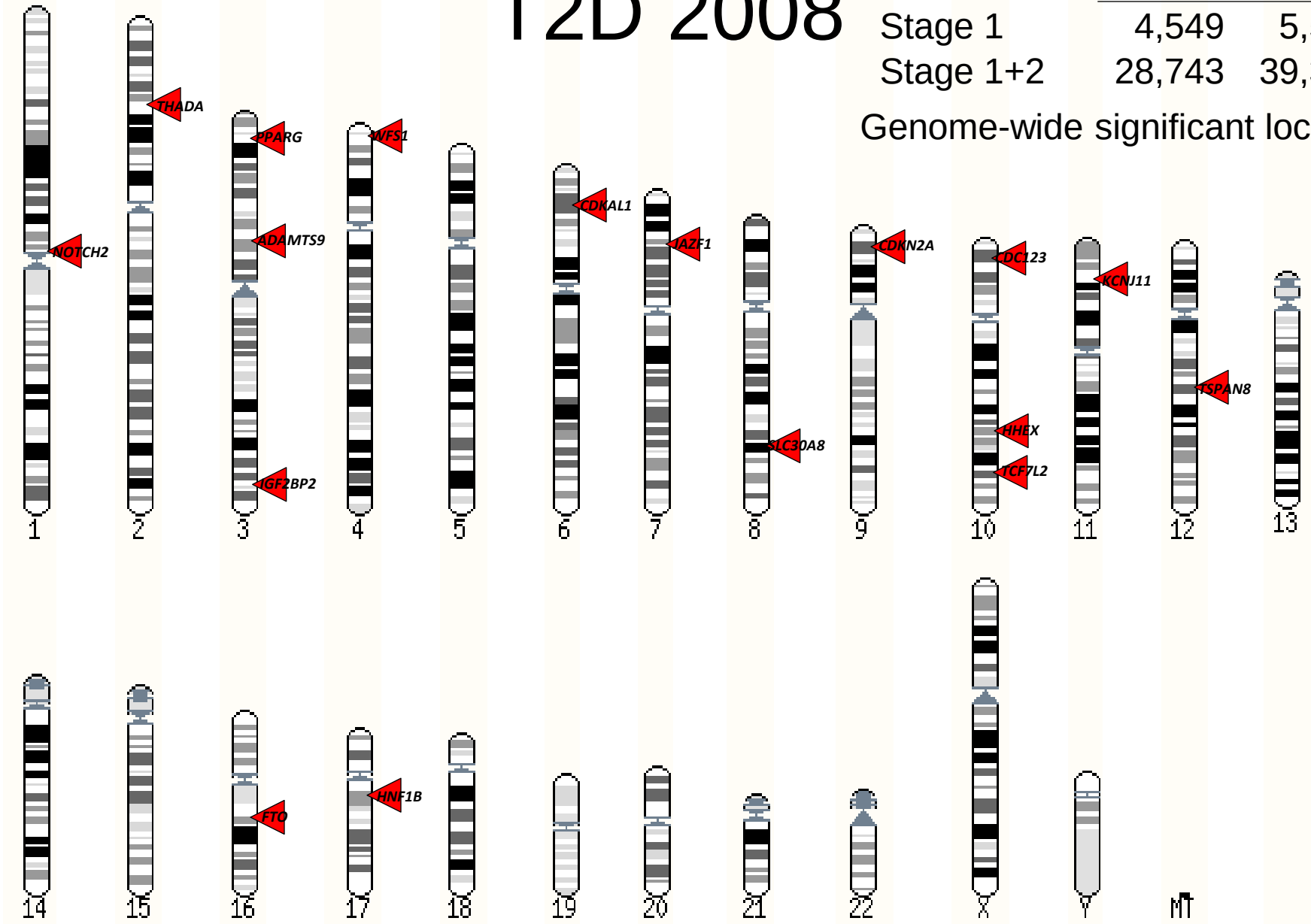
Genome-wide significant loci: 1



T2D 2008

	Cases	Controls
Stage 1	4,549	5,579
Stage 1+2	28,743	39,397

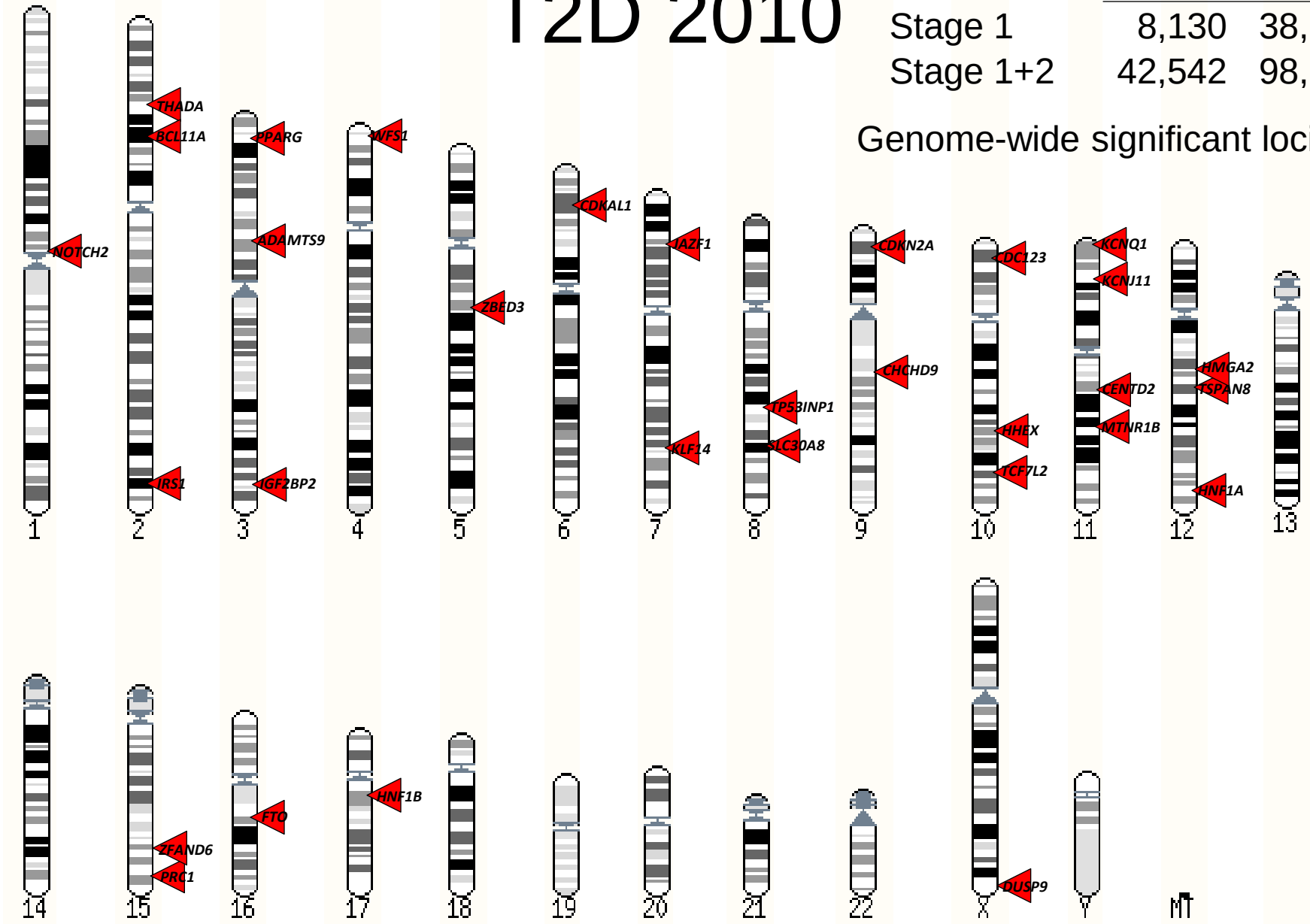
Genome-wide significant loci: 17



T2D 2010

	Cases	Controls
Stage 1	8,130	38,987
Stage 1+2	42,542	98,912

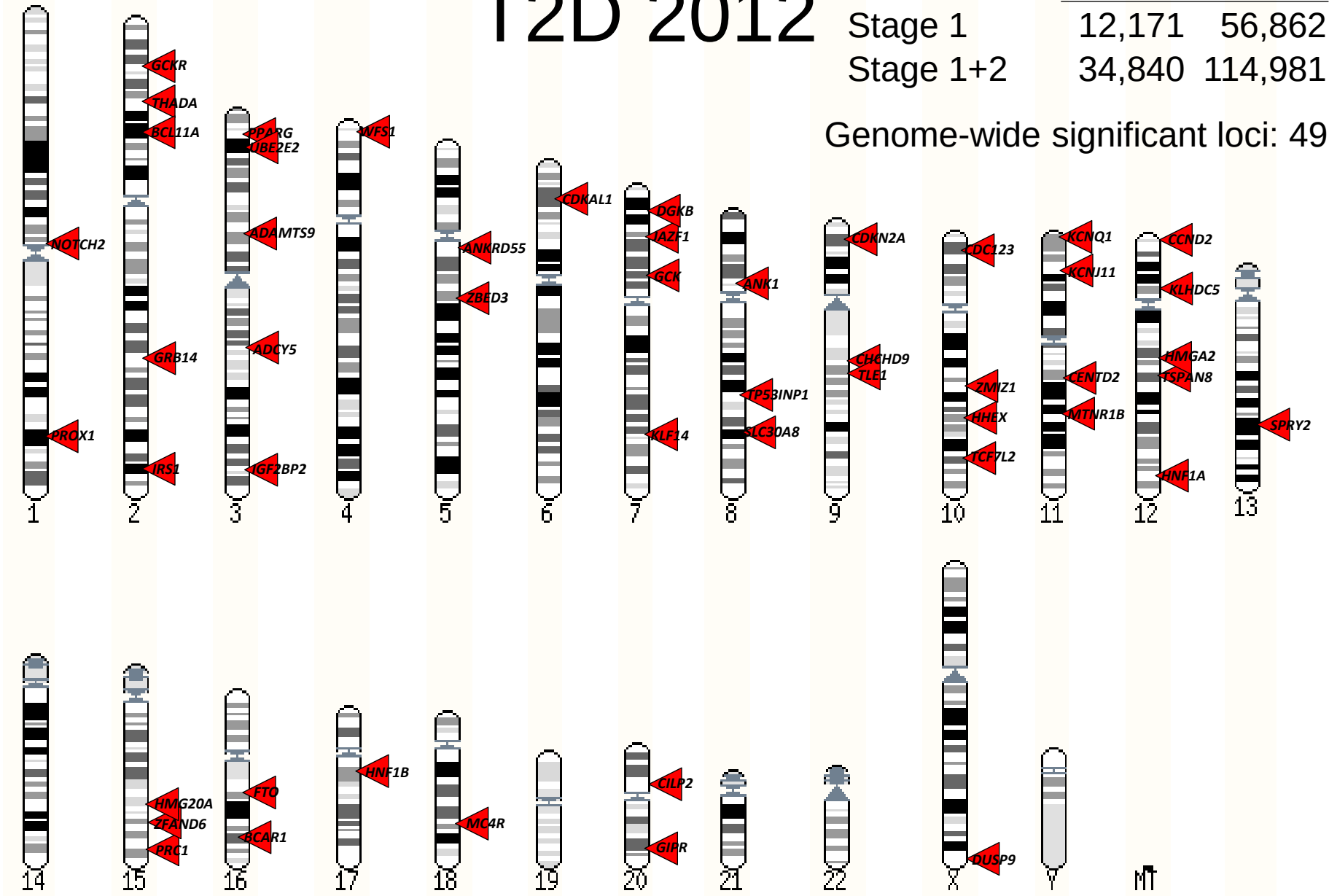
Genome-wide significant loci: 31



T2D 2012

	Cases	Controls
Stage 1	12,171	56,862
Stage 1+2	34,840	114,981

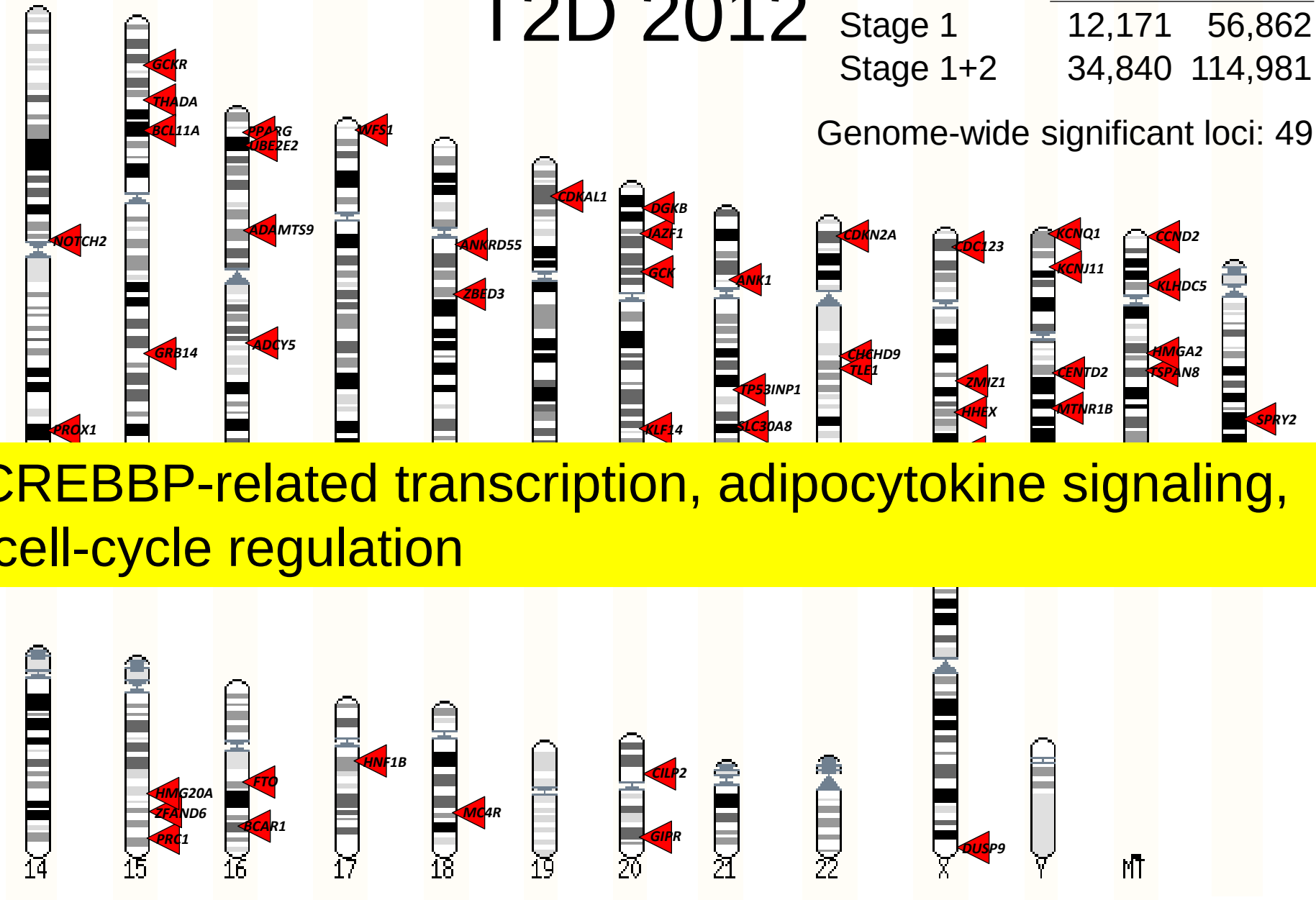
Genome-wide significant loci: 49



T2D 2012

	Cases	Controls
Stage 1	12,171	56,862
Stage 1+2	34,840	114,981

Genome-wide significant loci: 49



CREBBP-related transcription, adipocytokine signaling, cell-cycle regulation

Key lesson: sample size

- Sample size has been the key determinant of success in common disease genetics to date

  "Location, location, location"

- Technology and analysis tools also crucial: e.g. informative, low-cost genotype arrays, genotype imputation

- Collaboration for joint or meta-analysis across studies

Required sample size n to identify disease association

- n scales ~linearly with 1/MAF
  - MAF=.003 requires ~100x sample size as MAF=.3
- n scales ~linearly with 1 / $[\log OR]^2$ (fast)
  - OR = 1.2, 1.5, 3 require relative n's of 36, 7, 1
- n increases (slowly) with number of tests
  - 50M tests requires ~30% larger n than 1M tests

Study design matters too

- Designs: population cohort, case-control
- Both have advantages depending on trait, question; not mutually exclusive
- Case-control more powerful for genetic discovery for most diseases
  - “we are not doing many overpowered studies”
- Cohorts useful for
  - estimating effect size, population impact
  - discovery for QTs, very common diseases e.g. T2D
  - select extremes for QTs; cases, controls for disease

Study design: general preferences

- Deep phenotyping
  - help interpret associations for primary trait(s)
  - more traits for which we might identify association
- Broad consent
  - maximize value of data
- Available for callback based on genotype
  - study impact of rare variants
  - participants
  - family members

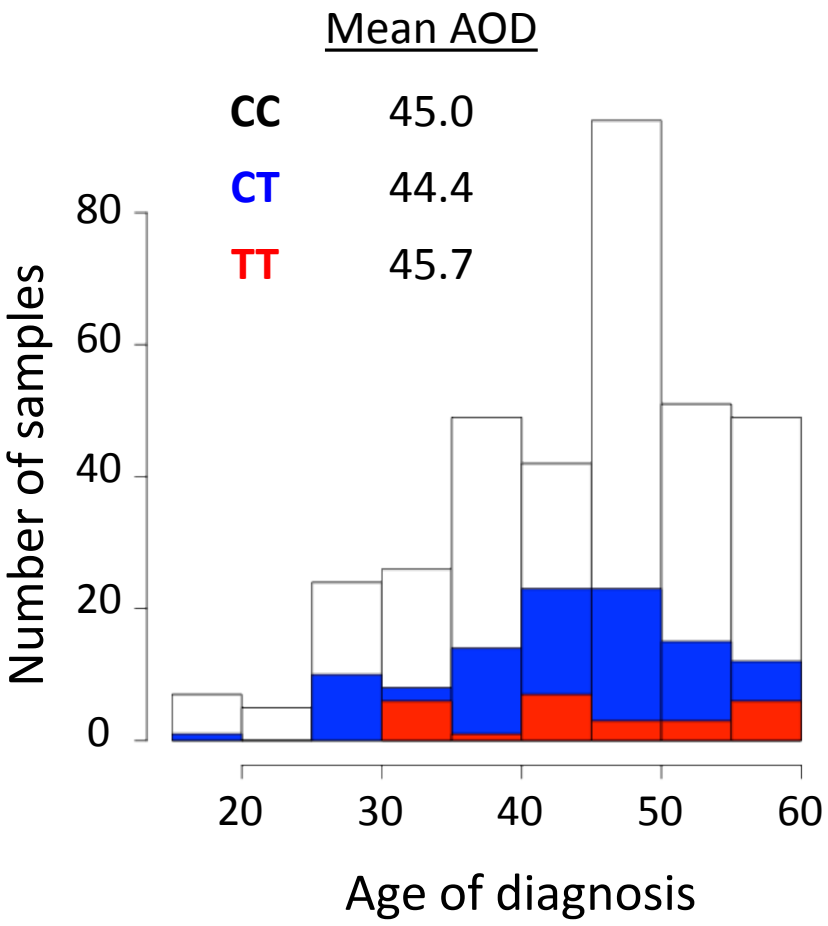
Study design approaches to increase power

- Careful study design and analysis can weight the dice towards increased power
- Assay multiple populations; variants rare in some may be (more) common in others
  - *TBC1D4* and T2D in Greenland (Moltke et al. 2014)
  - ***PAX4* and T2D in East Asians (T2D-GENES)**

PAX4 R192H is associated with T2D in East Asians (T2D-GENES)

Cohort	MAF	*p*-value
Korean	.077	1.0×10^{-4}
Singapore Chinese	.128	1.9×10^{-5}
Meta	.102	7.9×10^{-9}

- Only 3 copies of R192H present in n=10,775 of other ancestries
- *PAX4* mutations cause MODY, R192H not associated with age of diagnosis
- R192H impairs *PAX4* ability to repress transcription of insulin and glucagon



Slide courtesy of Tanya Teslovich

Study design approaches to increase power

- Careful study design and analysis can weight the dice towards increased power
- Assay multiple populations; variants rare in some may be (more) common in others
  - *TBC1D4* and T2D in Greenland (Moltke et al. 2014)
  - *PAX4* and T2D in East Asians (T2D-GENES)
- Group variants within functional units
  - *G6PC2* and fasting glucose (T2D-GENES)
  - ***SLC30A8* and T2D (Flannick et al. 2014)**

Type 2 Diabetes and *SLC30A8*

Flannick et al. *Nat Genet* 2014

Sequence: 750 from Finland and Sweden (DGI)

Look up in 2K from Iceland (deCODE)

Look up in 13K from 5 ancestry groups (T2D-GENES + GoT2D)

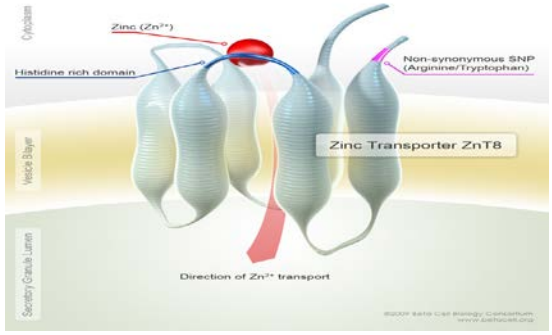
Genotype: 54K Europeans

Look up in 80K Europeans (multiple studies)

SLC30A8: Beta-cell-specific Zn^{++} transporter

Protein change	Annotation	case	control	OR	p
p.R138X	Nonsense	14/25,125	58/35,284	0.47	.0067
p.K34SfsX50	Frameshift	2/3,463	234/79,649	0.18	.0041

Protein change	Annotation	case	control	Population origin
p.Q174X	Nonsense	1	5	South Asian
p.Y284X	Nonsense	0	3	South Asian
p.S327TfsX55	Frameshift	0	2	African-American
p.I291FfsX2	Frameshift	0	1	African-American
p.W152X	Nonsense	0	1	Swedish
c.71+2T>A	Splice donor	1	1	African American
c.271+1G>A	Splice donor	0	2	South Asian, East Asian
c.419-1G>C	Splice acceptor	1	0	South Asian
c.572+1G>A	Splice donor	0	1	African American
p.M1I	Initiator codon	0	1	German
Total	-	3	17	p=.0021



For all LoF variants:
OR=0.34 $p=1.7x10^{-6}$

Attractive drug target

Based on slides courtesy of Jason Flannick

Type 2 Diabetes and *SLC30A8* (continued)

- Analysis of 12 *SLC30A8* LoF variants in 149,134 individuals across 22 studies
- Combined OR=0.34, p-value = 1.7×10^{-6}
- Importance of combining data across multiple variants and studies
- Enabled by analysis of data on multiple ancestries
- But what a job! Would be great if the data were in one place in a readily useable form

Data aggregation and knowledge

- Rapid data sharing arguably the most important legacy of the HGP
- We can do more than deposit data
- Aggregate data to maximize its utility, enable more powerful and efficient inference
- Facilitated by broad consent
- Next step: "knowledge portals" that operate on aggregated data, provide results to wide audience

How large a sample is required?

- Large samples required
- Actual numbers will differ based on genetic architecture (largely unknown)
- Zuk et al. *PNAS* 2014 suggest 25K cases and 25K controls per disease for exomes
- Reasonable starting point
- Are these large samples available? Starting point: GWAS samples

Large disease sample collections are available now

GWAS samples in 18 diseases: 400,000 cases

		GWAS Cases			GWAS Cases
Cardiometabolic	Early Myocardial Infarction	20,000	Psychiatric/ Neurologic	Schizophrenia	30,000
	Coronary Artery Disease	64,000		Bipolar	10,000
	Type 2 Diabetes	60,000		Autism	20,000
	Atrial Fib/Stroke	10,000		Alzheimer	10,000
Germline Cancer Risk	Breast Cancer	25,000	Autoimmune	Type 1 Diabetes	30,000
	Prostate Cancer	10,000		IBD/Crohn's	30,000
	Colon Cancer	13,000		Multiple Sclerosis	20,000
	Lung Cancer	20,000		Rheumatoid Arthritis	30,000
	Melanoma	13,000		Lupus	15,000

Slide courtesy of Eric Lander

Which diseases?

- Focus on some first, develop strategies, methods
- Given many common diseases, be opportunistic and initially advantage diseases with
  - large numbers of well-phenotyped, broadly-consented, callback-eligible samples
  - investigator groups that are well organized, collegial, strong record of data sharing
  - significant financial support from categorical institute or other funder
- Success of relevant GWAS consortia instructive

Key strategic issues/choices

- NHGRI vs categorical institutes
- Large centers vs distributed capacity
- Common vs Mendelian diseases
- Discovery vs translation
- Exomes vs genomes

Key strategic issues/choices

- NHGRI and categorical institutes
- Large centers and distributed capacity
- Common and Mendelian diseases and everything in between
- Discovery and translation
- Exomes and genomes

Avoid false dichotomies!

NHGRI and the categorical institutes
The role of the NHGRI is to …

- Advance paradigms
- Develop, evaluate, and harden methods/tools
  - many methods/tools general, NHGRI logical leader
  - provide scale, infrastructure, capacity
- Imagine and develop foundational resources: e.g. HGP, other genomes, HapMap, 1000G, …
- Enlist help of categorical institutes when possible
  - FY2013 NHGRI: $486M, several ICs >$3B

Large centers and distributed capacity

- Large centers
  - set standards, develop analysis paradigms and infrastructure
  - industrialize genomics, enable large studies
- More small centers
  - increase opportunity/competition
  - enable broader range of studies
- Both important for
  - training
  - innovation
  - expanding capacity

Discovery and translation

- Genetic discovery for common diseases has only begun, and for rare variants has barely started
- Translation for common diseases requires (much more) discovery
- Translation now can take advantage of what we know now, prepare us for when we know more
- Virtuous circle: discovery and translation can reinforce if we capture data/hypotheses/samples from translation and use them to inform discovery

Exomes and genomes

- Exomes
  - cost/sample size
  - interpretability
  - inherently limited
- Genomes
  - cost higher but likely coming down faster
  - surely the direction we will go eventually
  - time for a more significant investment so we can prepare
  - provide better exomes

What will happen if NHGRI decides not to pursue this area?

- Hard to imagine, but if not …
- Science/biomedicine: fragmented effort that will be slower, more costly, less efficient, and result in less interoperable data
- NHGRI: a huge lost opportunity

Opportunities (1)

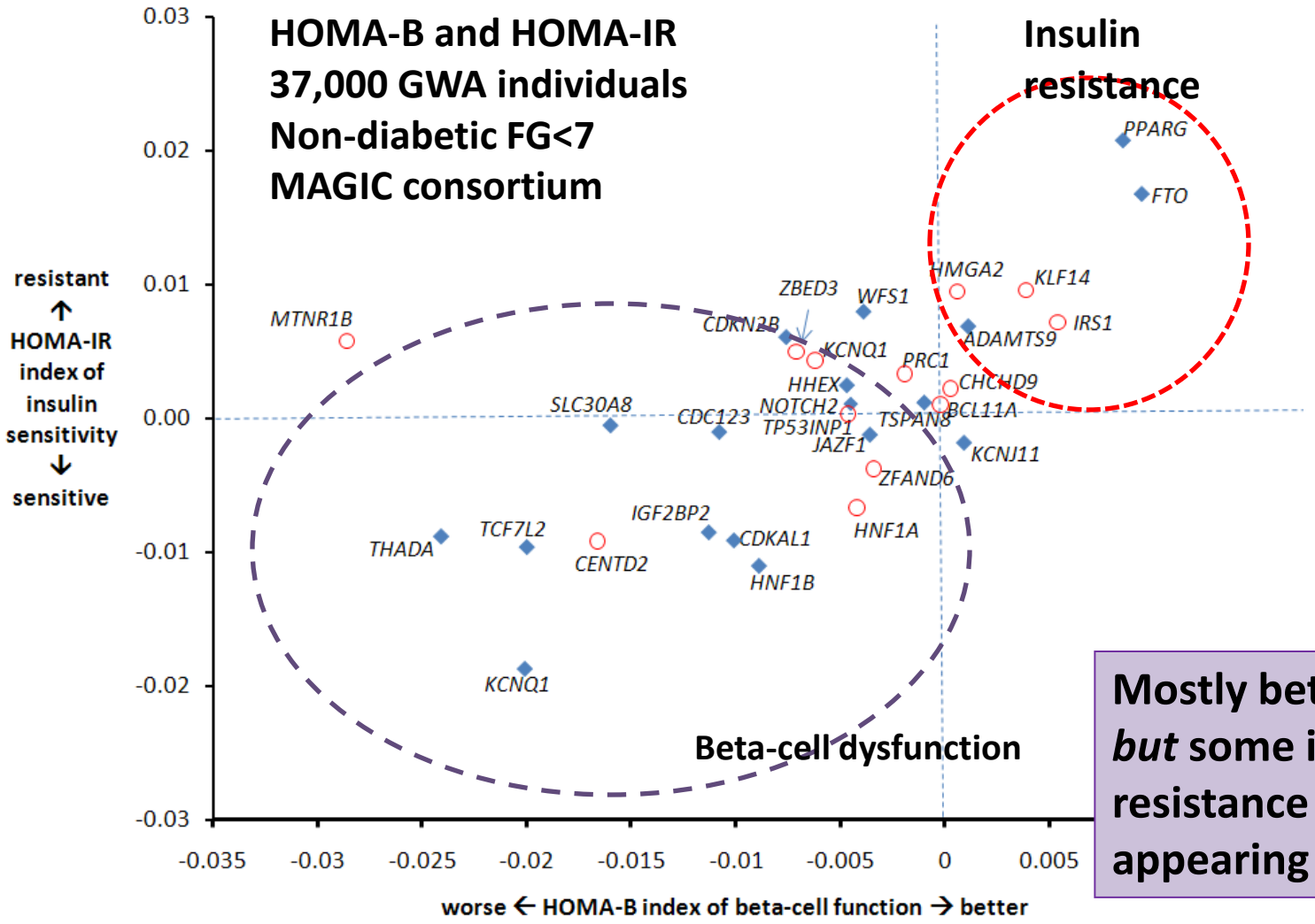
- Disease: focus on exemplar diseases
- Samples: encourage identification and aggregation of large, well-phenotyped, broadly consented samples
- Resource: set of recallable sequenced genomes, e.g. LoF carriers for every human gene
- Technology: continue focus on sequencing and statistical/computational methods and tools
- Whole genome sequencing: time to do more

Opportunities (2)

- Information: more active data aggregation and sharing, knowledge sharing
- Discovery and translation: a virtuous circle if we take advantage
- Functional characterization of variants: prospective, high-throughput
- Training: invest more in genome science, and statistics and computational science
- Genotyping: genotype arrays on huge samples

Backup slides

T2D: Roles of Insulin Secretion and Action



Mostly beta-cell genes, *but* some insulin resistance genes appearing

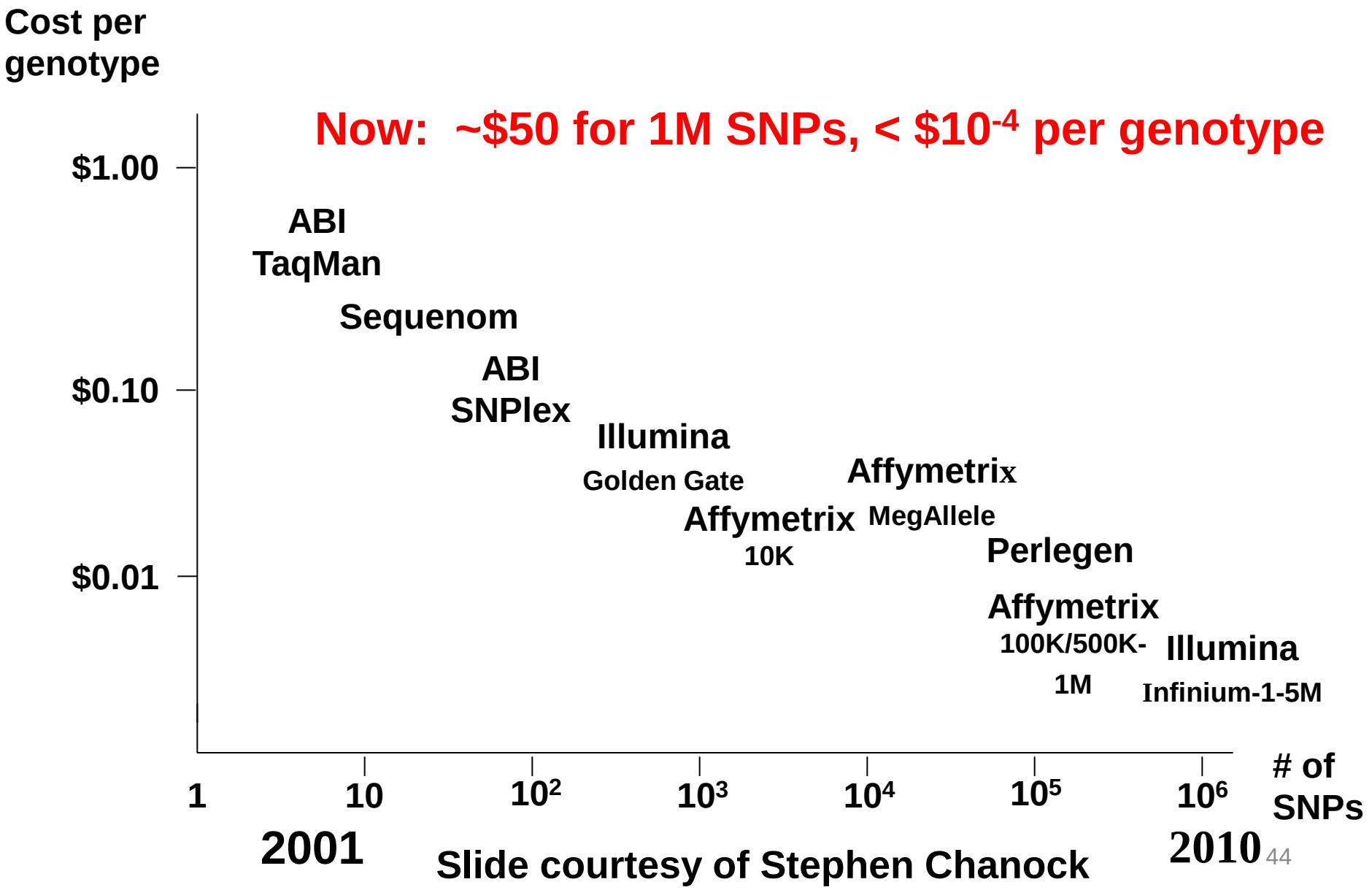
Slide courtesy of Mark McCarthy

Cardiometabolic GWAS Consortium Results

published through July 2014

Name	Studies	n_{GWAS}	Trait(s)	Loci
DIAGRAM	38	150K N_e=88K	T2D	49
GIANT	51	159K	BMI, WHR	66
MAGIC	63	133K	Glucose, Insulin	53
Global Lipids	60	189K	LDL, HDL, TG, CH	157
ICBP	29	69K	SBP, DBP	28

Improvements in Genotyping Technology



rs2233580 (*PAX4* R192H) is associated with T2D in East Asians

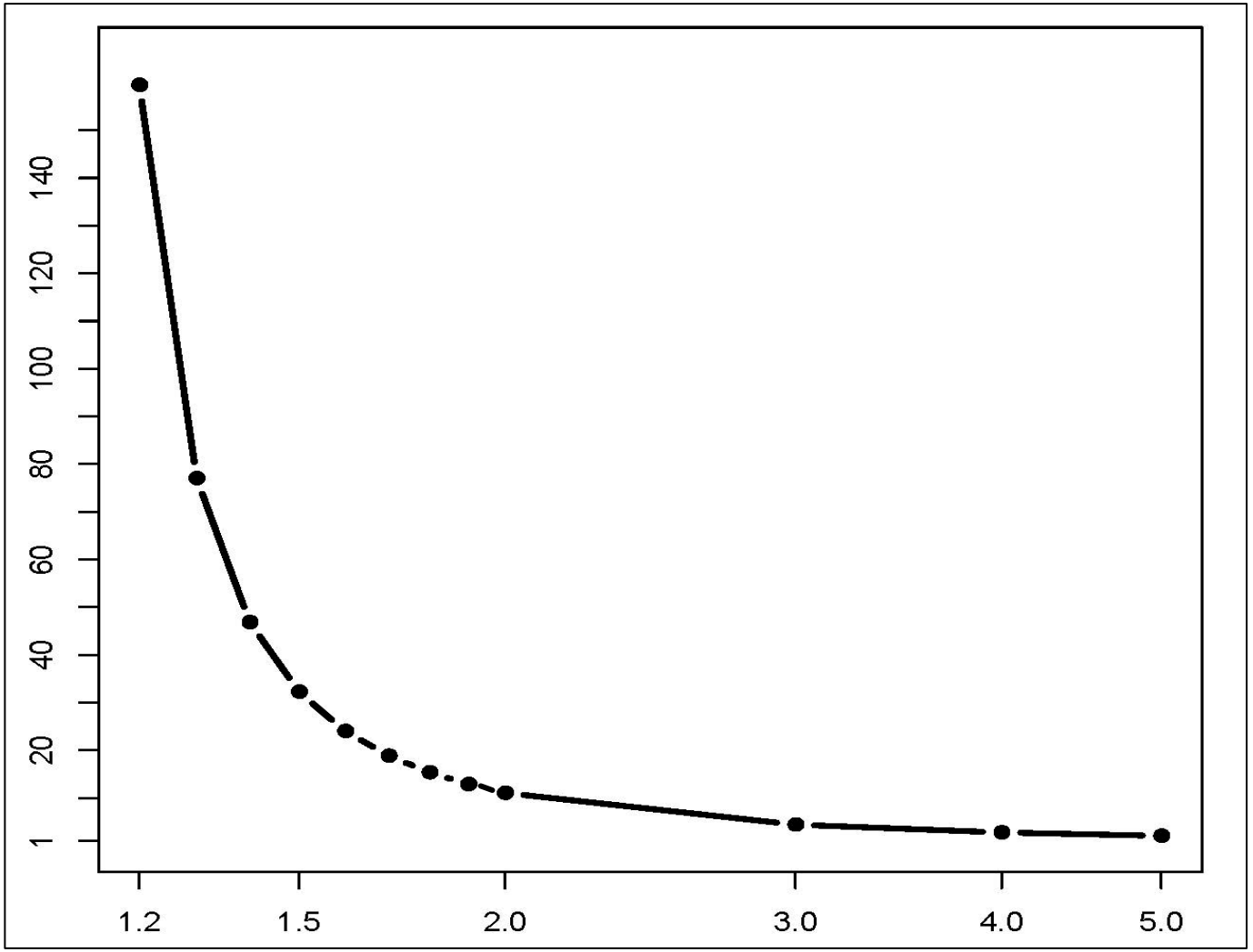
- rs2233580 encodes *PAX4* R192H
  - Variant absent in African Americans, Finns; seen once in South Asians
  - Independent of East Asian GWAS SNP rs6467136 (r^2 = .02)
  - Not genotyped or well-imputed using early GWAS arrays
- Paired Box Gene 4 *(PAX4)*
  - Essential for development of pancreatic islet beta cells
  - Mutations cause MODY

Testing for Association (Total Sample Size n)

MAF=	.05	.02	.01	.005	.002	.001
OR=1.5	7K	16.4K	32K	63K	156K	310K
OR=2.0	2.2K	4.8K	9.5K	19K	46K	92K
OR=3.0	0.8K	1.7K	3.1K	6K	15K	30K

Additive model, n/2 cases, n/2 controls, 80% power, $\alpha=5x10^{-8}$

Relative Sample Size as a Function of Odds Ratio

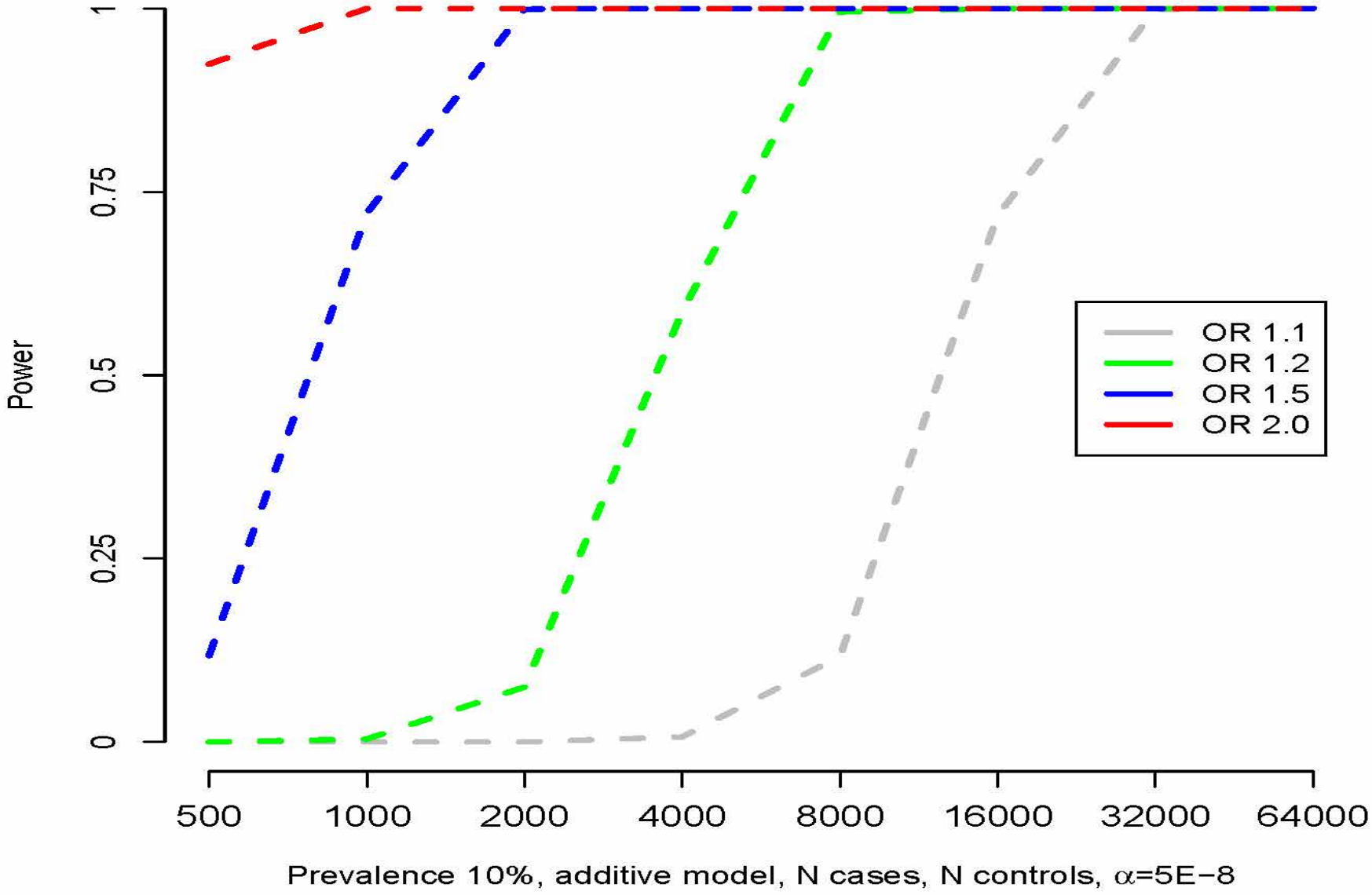


Loci Identified by GWAS Consortia

Name	Most recent publication as of July 2014
DIAGRAM	Morris et al. 2012 Nat Genet. Metabochip paper
GIANT	Berndt et al. 2013 Nat Genet (GWAS using samples in top and bottom 5% of trait distributions)
MAGIC	Scott et al. 2012 Nat Genet. Metabochip paper
Global Lipids	2013 GLGC Metabochip paper
Global BP Gen	Ehret G.B., Munroe P.B., Rice K.M., Bochud M., Johnson A.D., Chasman D.I., Smith A.V., Tobin M.D., Verwoert G.C., Hwang S.J., et al. Genetic variants in novel pathways influence blood pressure and cardiovascular disease risk. Nature. 2011;478:103–109

Power to Detect Association

MAF=.3



Power to Detect Association

MAF=.3 and MAF=.003