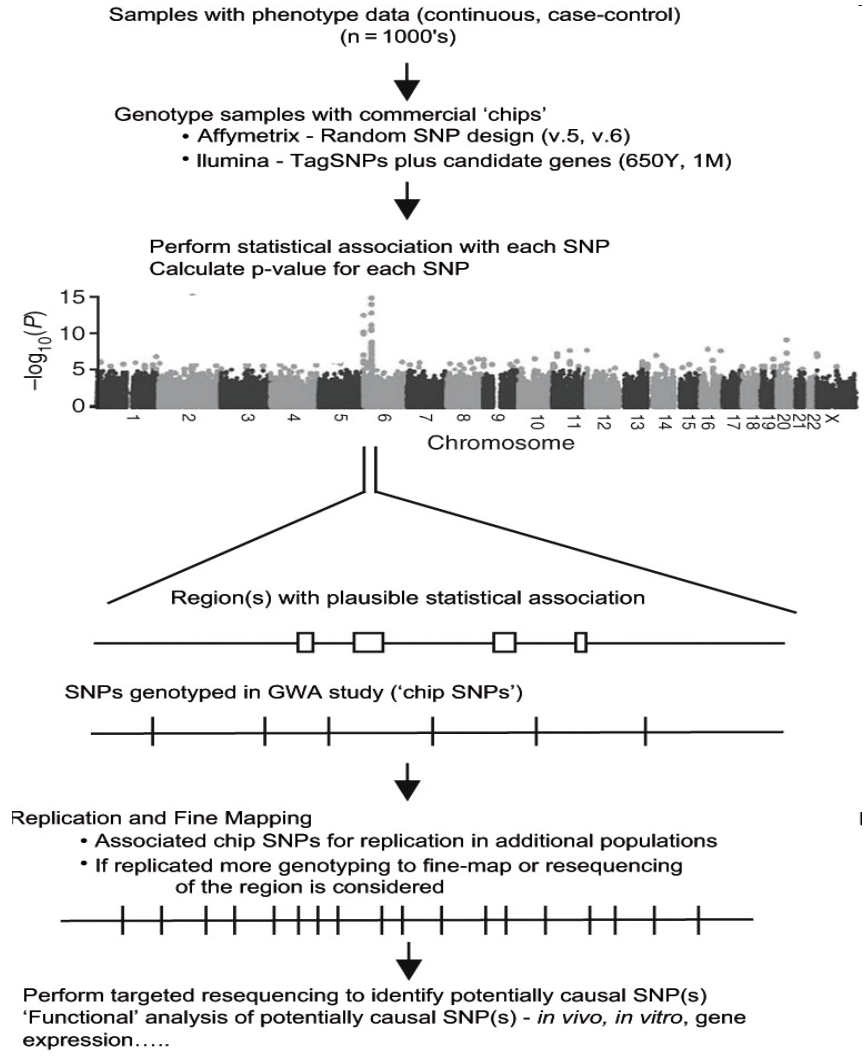
Analysis of common variants (Genome-wide association studies - GWAS)

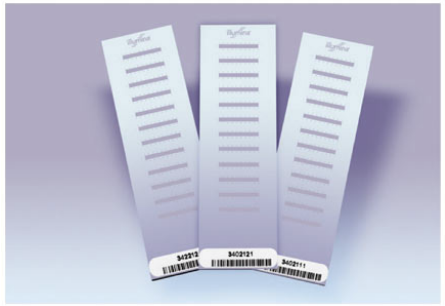
Analysis of rare variants (Rare variant association studies - RVAS)

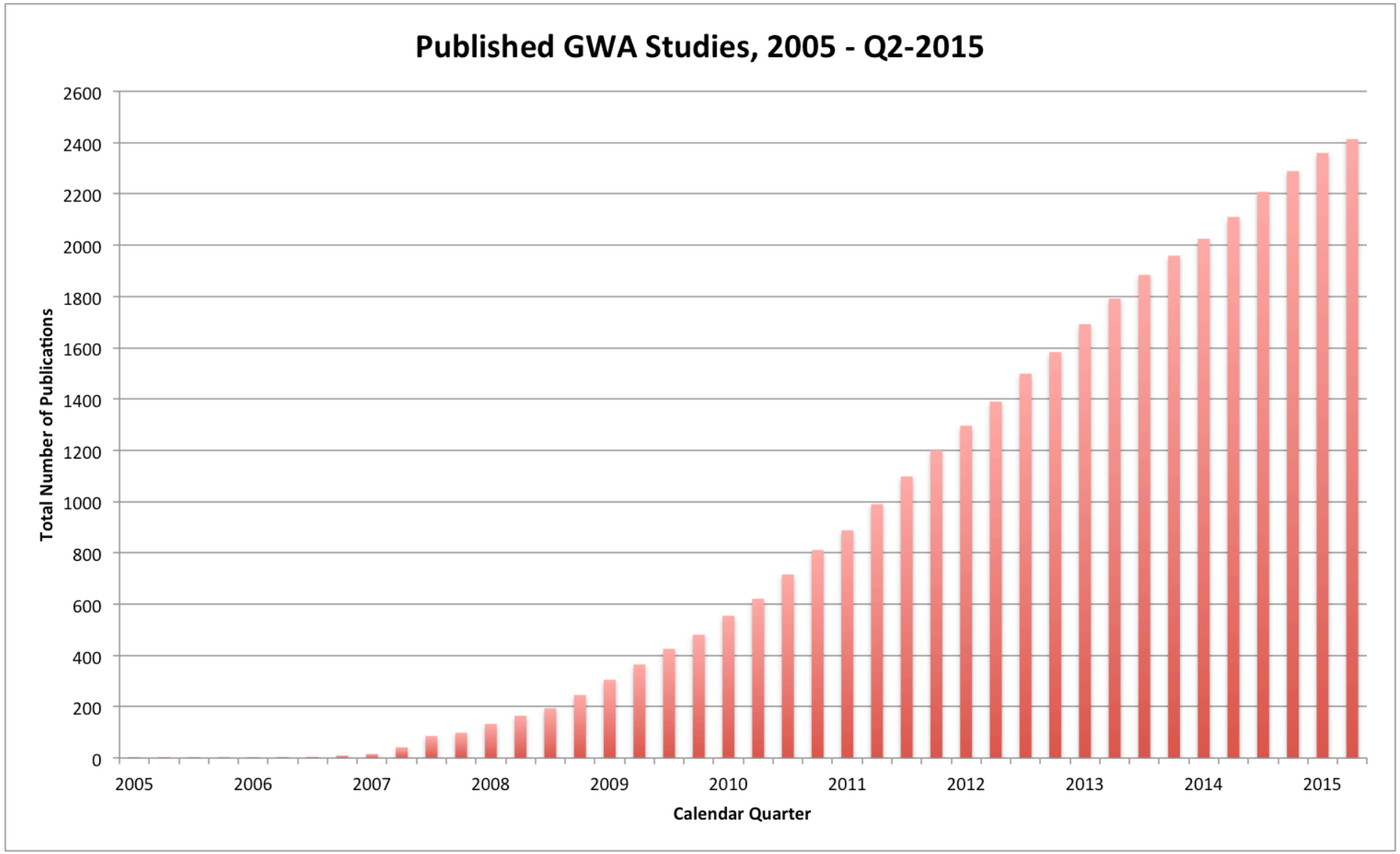
Genome-wide association studies (GWAS)



Rieder et al. 2008

Screen across the genome for SNPs that are associated with trait (agnostic approach)




Published GWA Studies, 2005 - Q2-2015
Total Number of Publications
2600
2400
2200
2000
1800
1600
1400
1200
1000
800
600
400
200
0
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015
Calendar Quarter

December 2005

Complement Factor H Polymorphism in Age-Related Macular Degeneration

Robert J. Klein,[1] Caroline Zeiss,[2]* Emily Y. Chew,[3]* Jen-Yue Tsai,[4]* Richard S. Sackler,[1] Chad Haynes,[1] Alice K. Henning,[5] John Paul SanGiovanni,[3] Shrikant M. Mane,[6] Susan T. Mayne,[7] Michael B. Bracken,[7] Frederick L. Ferris,[3] Jurg Ott,[1] Colin Barnstable,[2] Josephine Hoh[7]†

Age-related macular degeneration (AMD) is a major cause of blindness in the elderly. We report a genome-wide screen of 96 cases and 50 controls for polymorphisms associated with AMD. Among 116,204 single-nucleotide polymorphisms genotyped, an intronic and common variant in the complement factor H gene (*CFH*) is strongly associated with AMD (nominal *P* value $<10^{-7}$). In individuals homozygous for the risk allele, the likelihood of AMD is increased by a factor of 7.4 (95% confidence interval 2.9 to 19). Resequencing revealed a polymorphism in linkage disequilibrium with the risk allele representing a tyrosine-histidine change at amino acid 402. This polymorphism is in a region of CFH that binds heparin and C-reactive protein. The *CFH* gene is located on chromosome 1 in a region repeatedly linked to AMD in family-based studies.

Age-related macular degeneration (AMD) is the leading cause of blindness in the developed world. Its incidence is increasing as the elderly population expands (*1*). AMD is characterized by progressive destruction of the retina's central region (macula), causing central field visual loss (*2*). A key feature of AMD is the formation of extracellular deposits called drusen concentrated in and around the macula behind the retina between the retinal pigment epithelium (RPE) and the choroid. To date, no therapy for this disease has proven to be broadly effective. Several risk factors have been linked to AMD, including age, smoking, and family history (*3*). Candidate-gene studies have not found any genetic differences that can account for a large proportion of the overall prevalence (*2*). Family-based whole-genome linkage scans have identified chromosomal regions that show evidence of linkage to AMD (*4*–*8*), but the linkage areas have not been resolved to any causative mutations.

Like many other chronic diseases, AMD is caused by a combination of genetic and environmental risk factors. Linkage studies are not as powerful as association studies for the identification of genes contributing to the risk for common, complex diseases (*9*). However, linkage studies have the advantage of searching the whole genome in an unbiased manner without presupposing the involvement of particular genes. Searching the whole genome in an association study requires typing 100,000 or more single-nucleotide polymorphisms (SNPs) (*10*). Because of these technical demands, only one whole-genome association study, on susceptibility to myocardial infarction, has been published to date (*11*).

Study design. We report a whole-genome case-control association study for genes involved in AMD. To maximize the chance of success, we chose clearly defined phenotypes for cases and controls. Case individuals exhibited at least some large drusen in a quantitative photographic assessment combined with evidence of sight-threatening AMD (geographic atrophy or neovascular AMD). Control individuals had either no or only a few small drusen. We analyzed our data using a statistically conservative approach to correct for the large number of SNPs tested, thereby guaranteeing that the probability of a false positive is no greater than our reported *P* values.

We used a subset of individuals who participated in the Age-Related Eye Disease Study (AREDS) (*12*). From the AREDS

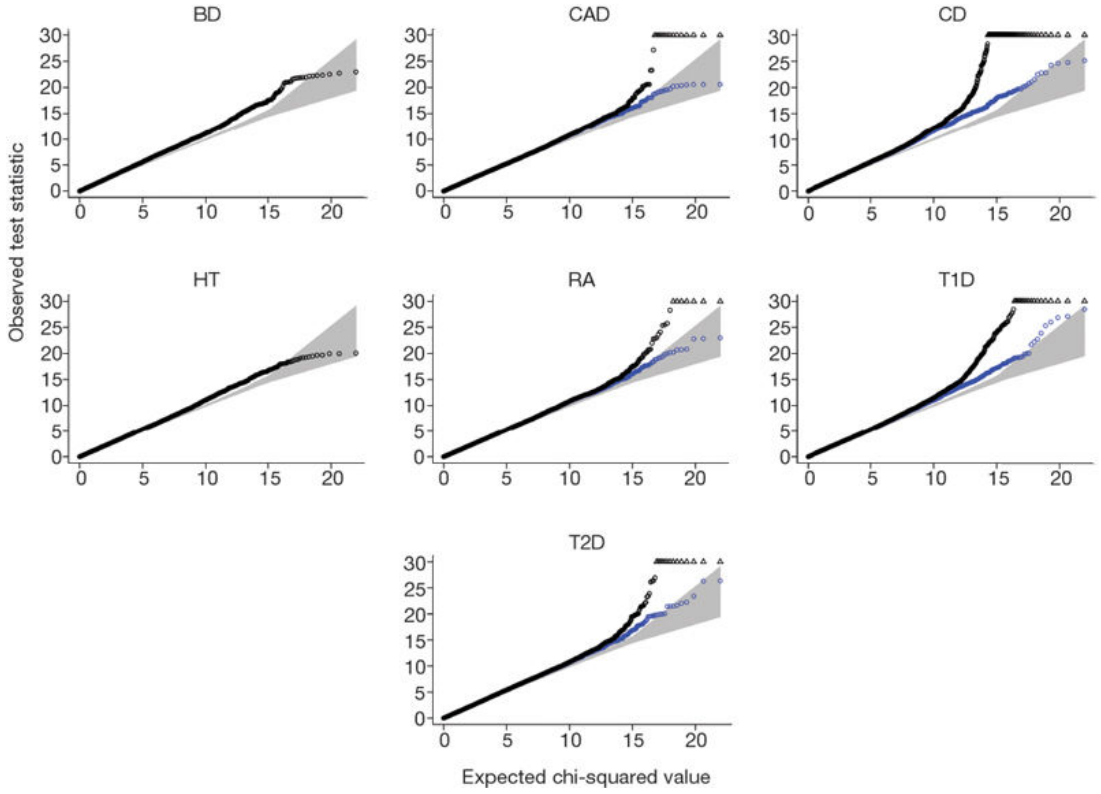
[1]Laboratory of Statistical Genetics, Rockefeller University, 1230 York Avenue, New York, NY 10021, USA. [2]Department of Ophthalmology and Visual Science, Yale University School of Medicine, 330 Cedar Street, New Haven, CT 06520, USA. [3]National Eye Institute, Building 10, CRC, 10 Center Drive, Bethesda, MD 20892–1204, USA. [4]Biological Imaging Core, National Eye Institute, 9000 Rockville Pike, Bethesda, MD 20892, USA. [5]The EMMES Corporation, 401 North Washington Street, Suite 700, Rockville MD 20850, USA. [6]W. M. Keck Facility, Yale University, 300 George Street, Suite 201, New Haven, CT 06511, USA. [7]Department of Epidemiology and Public Health, Yale University School of Medicine, 60 College Street, New Haven, CT 06520, USA.

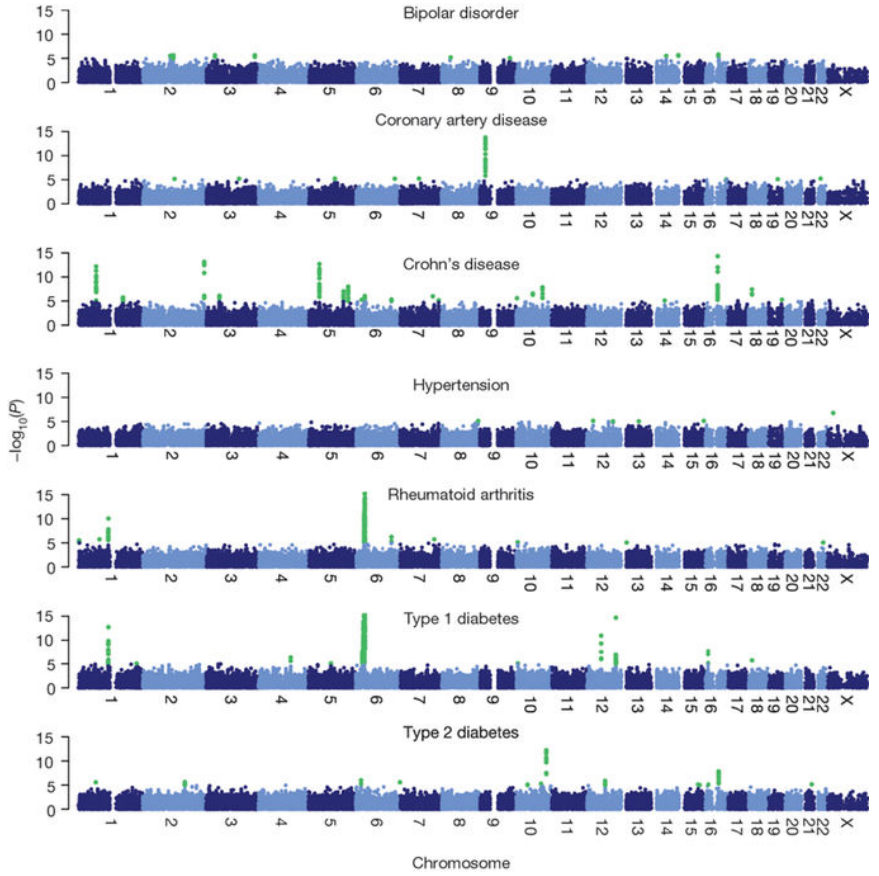
*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: josephine.hoh@yale.edu

Klein, Science 2005

March 2018

Presentation of results from large-scale genetic association studies



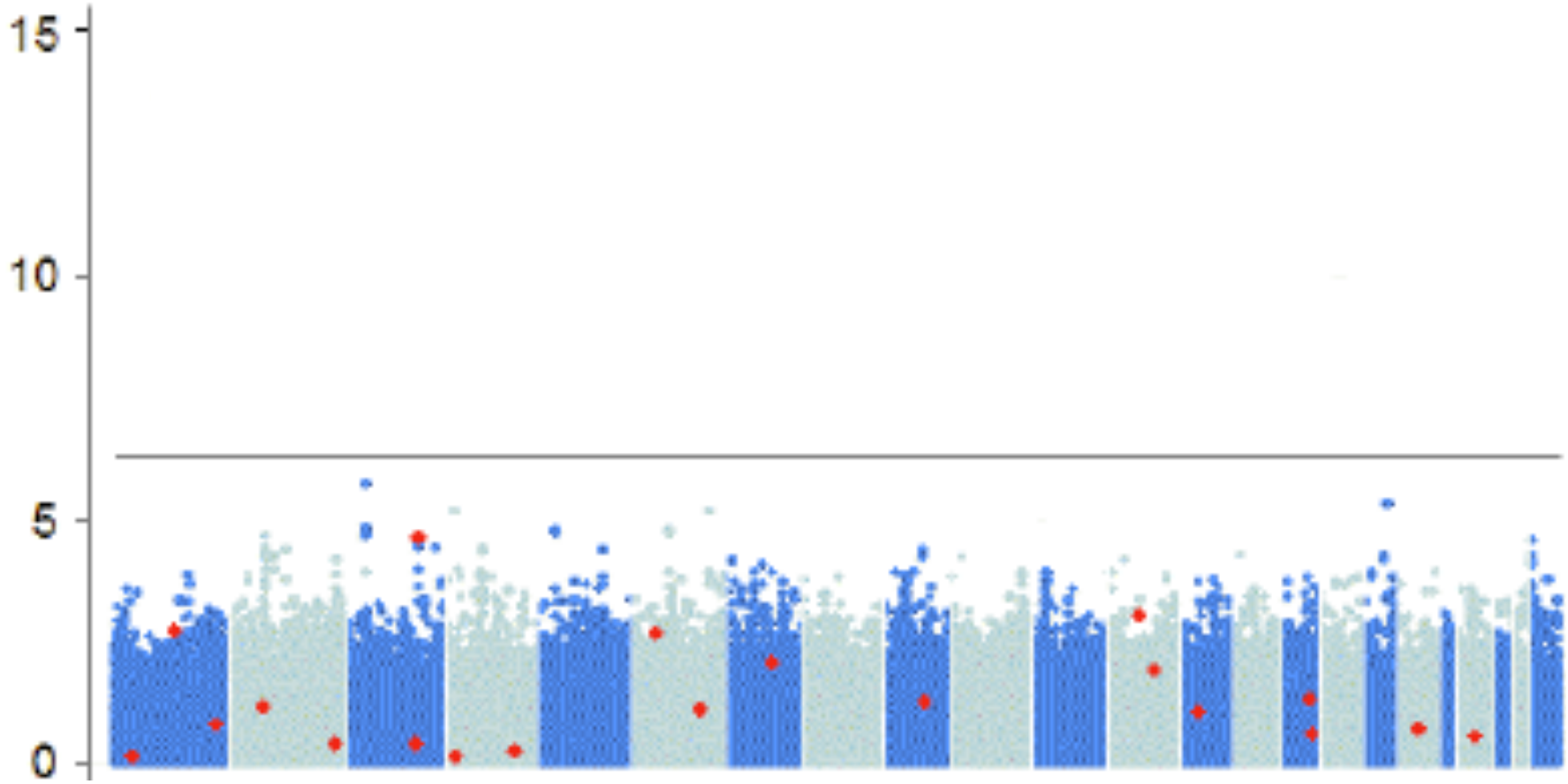


An association with p-value <5x10^{-8} is considered genome-wide significant

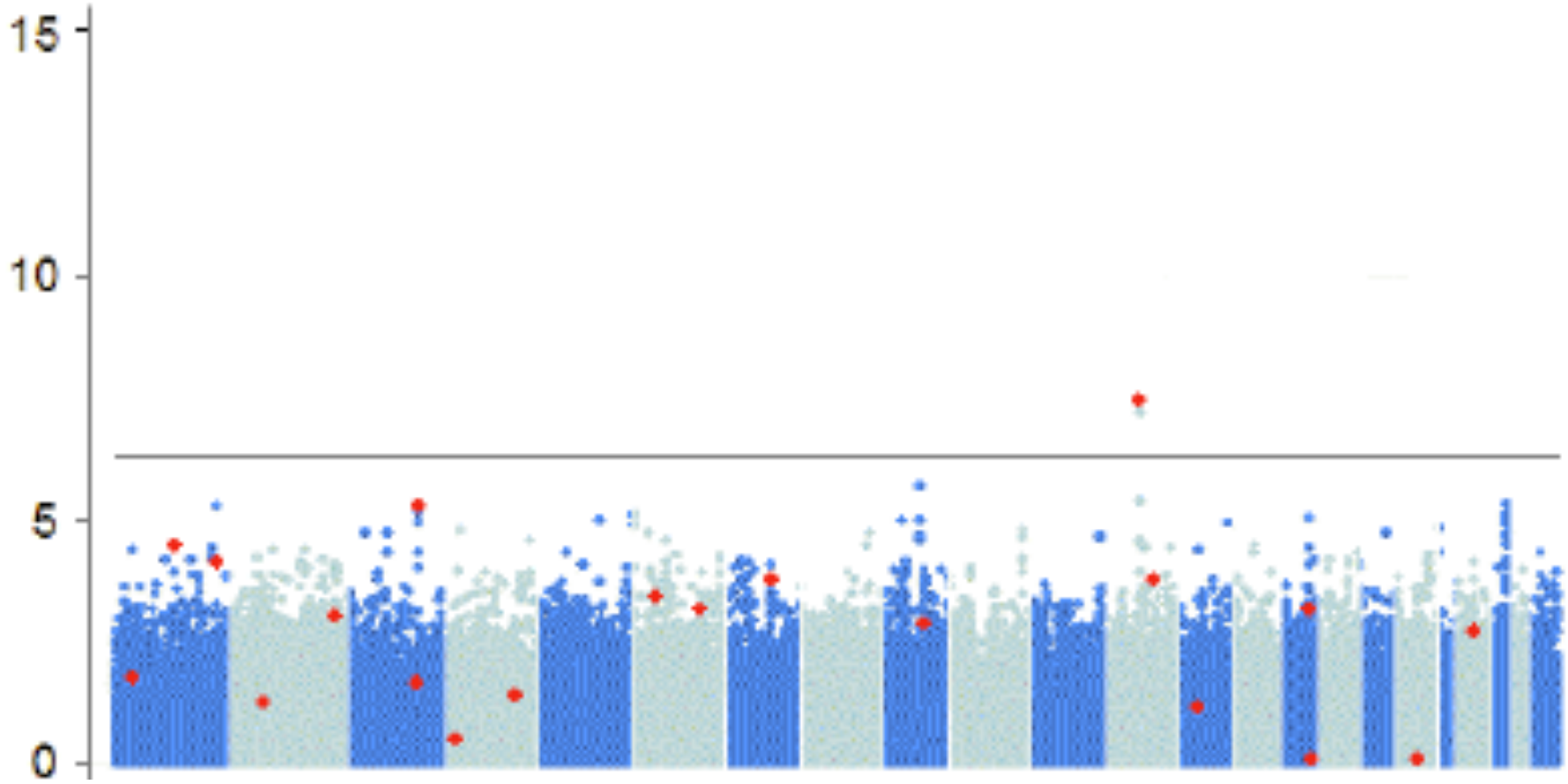
WTCCC, Nature 2007

The importance of sample size in GWAS

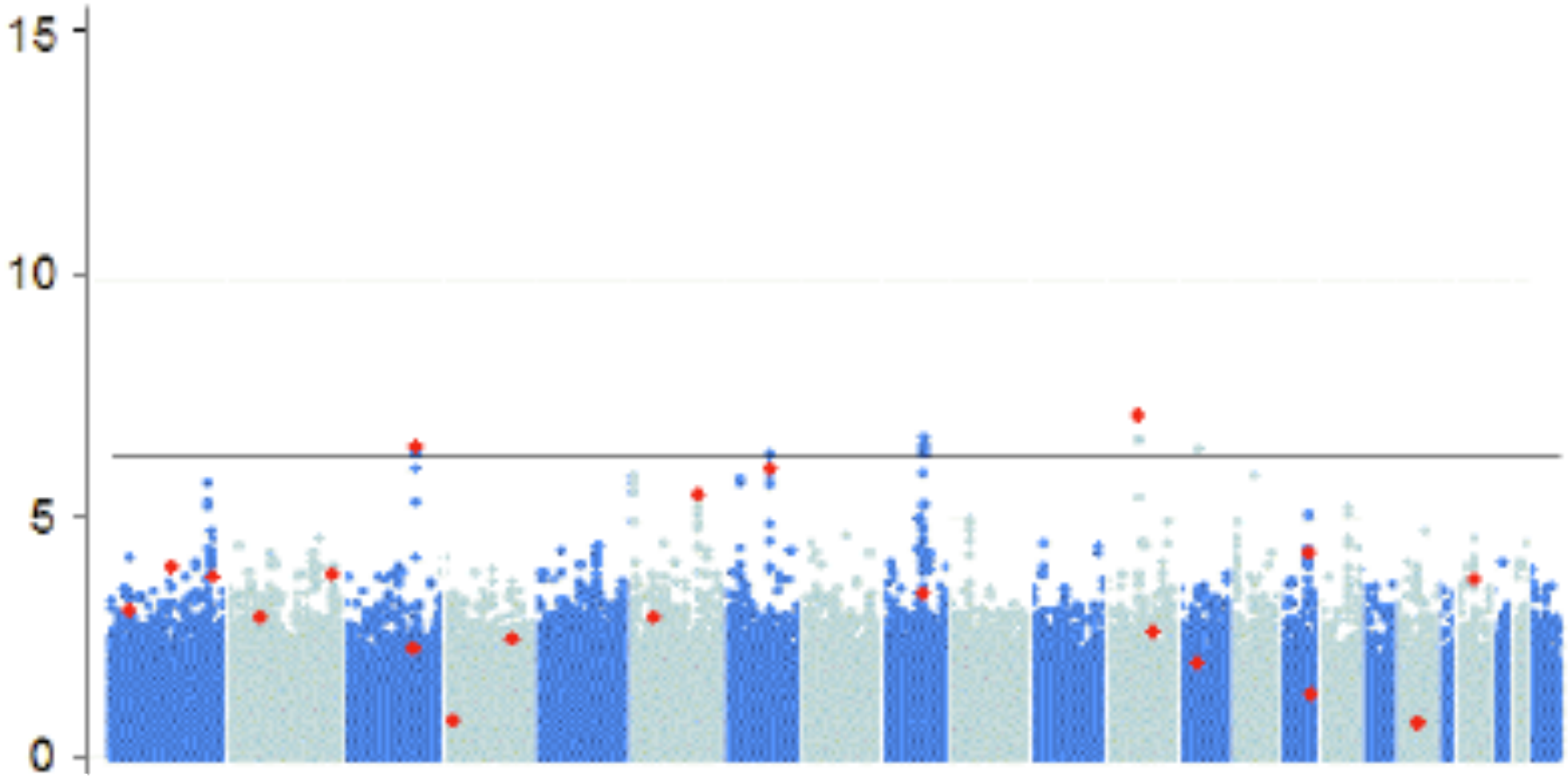
One of the first GWAS of height (N=1,914). Red dots represent SNPs that achieved a $P<5x10^{-7}$ in the joint analysis with stage 2 samples. The solid black horizontal line is the $P=5x10^{-7}$ line.



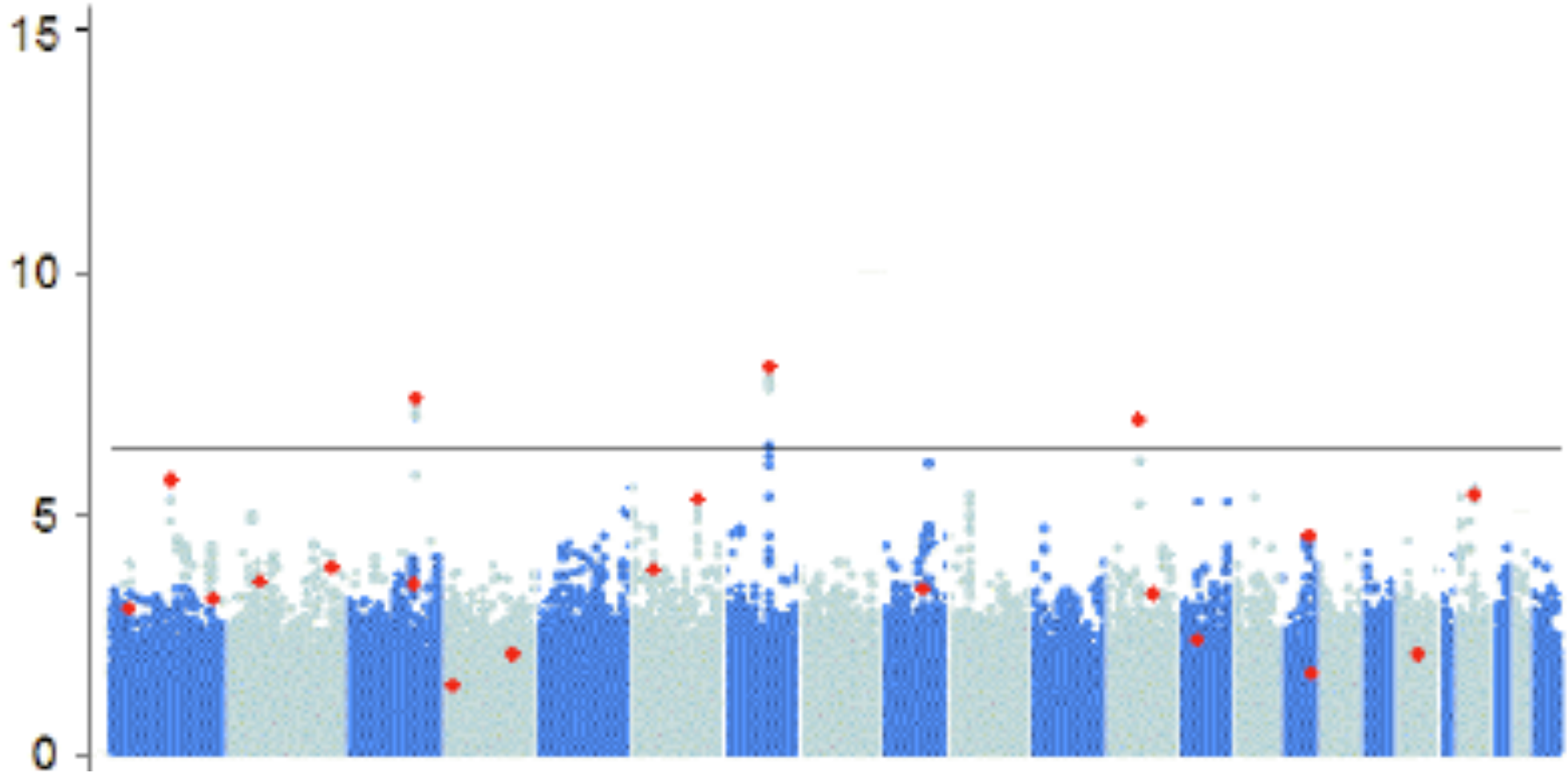
N=4,892



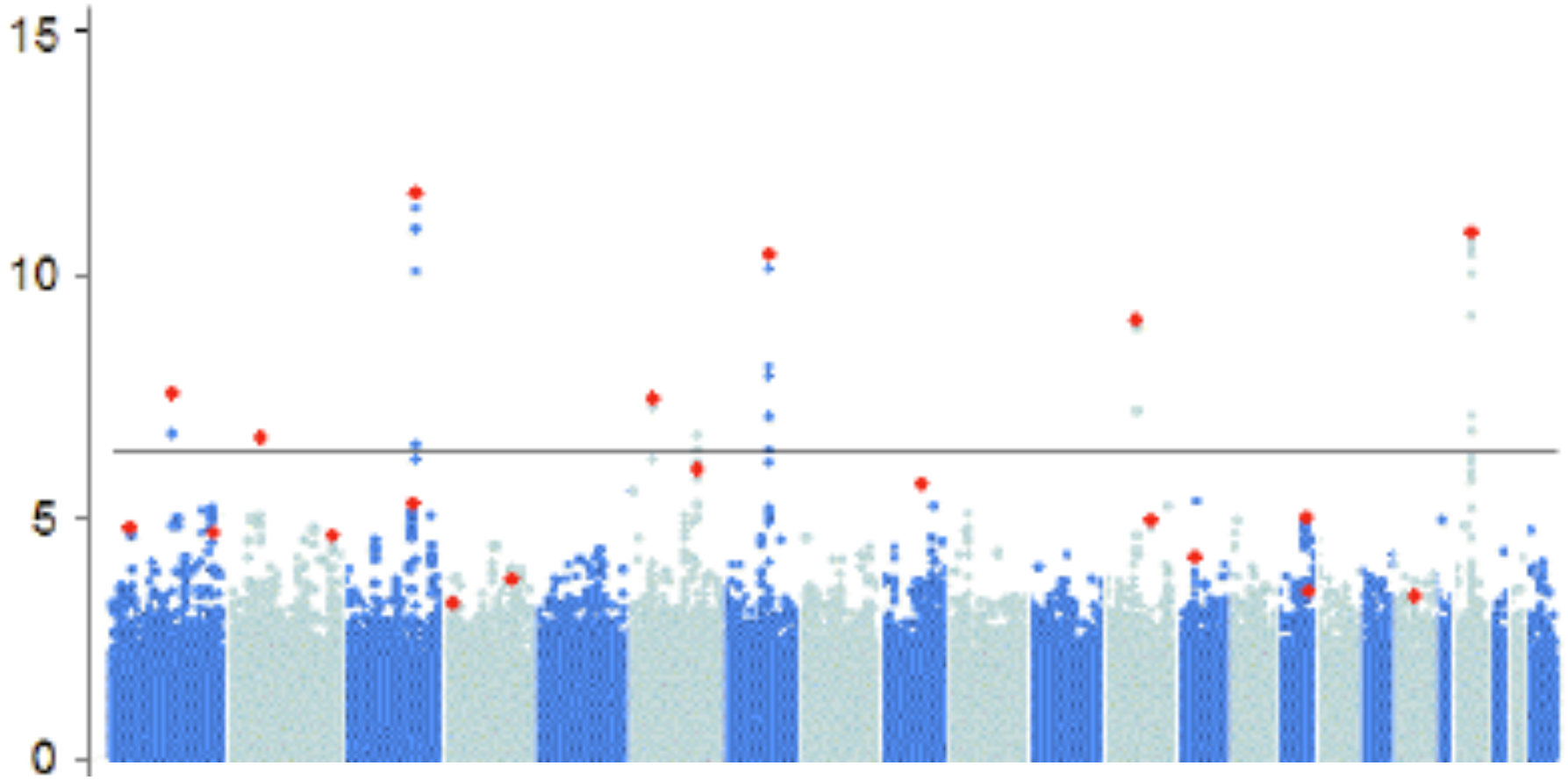
N=6,788



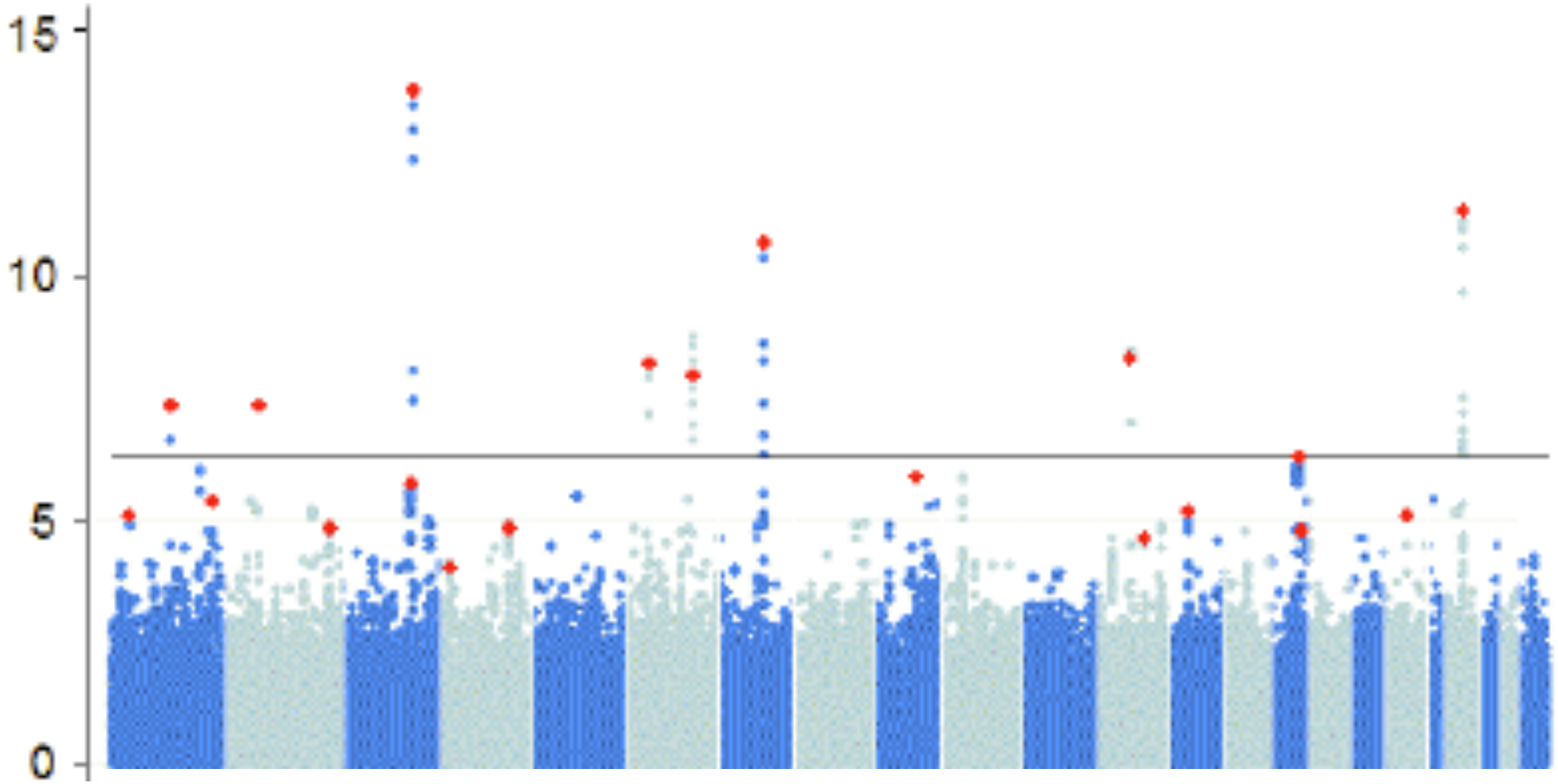
N=8,668



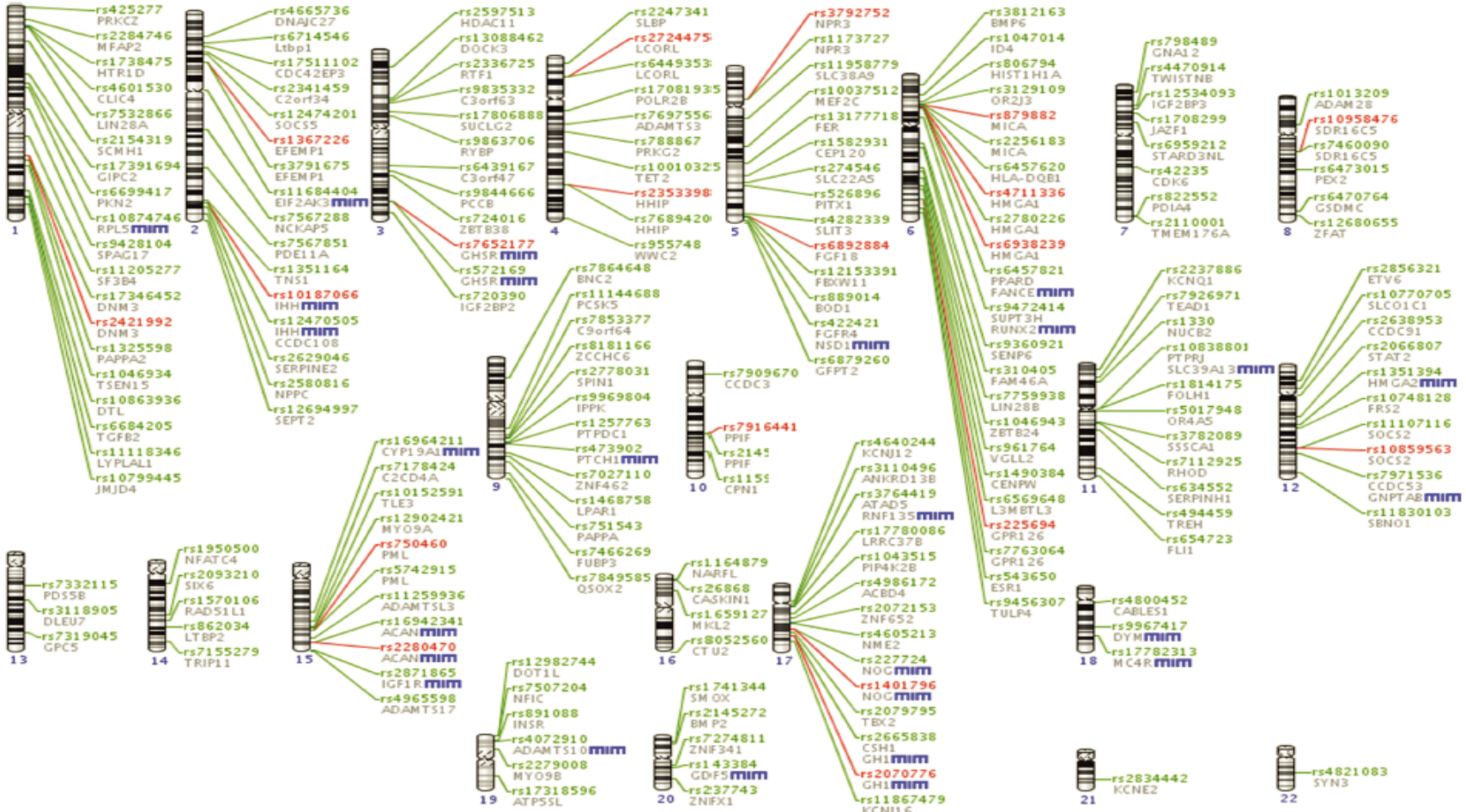
N=12,228



N=13,665



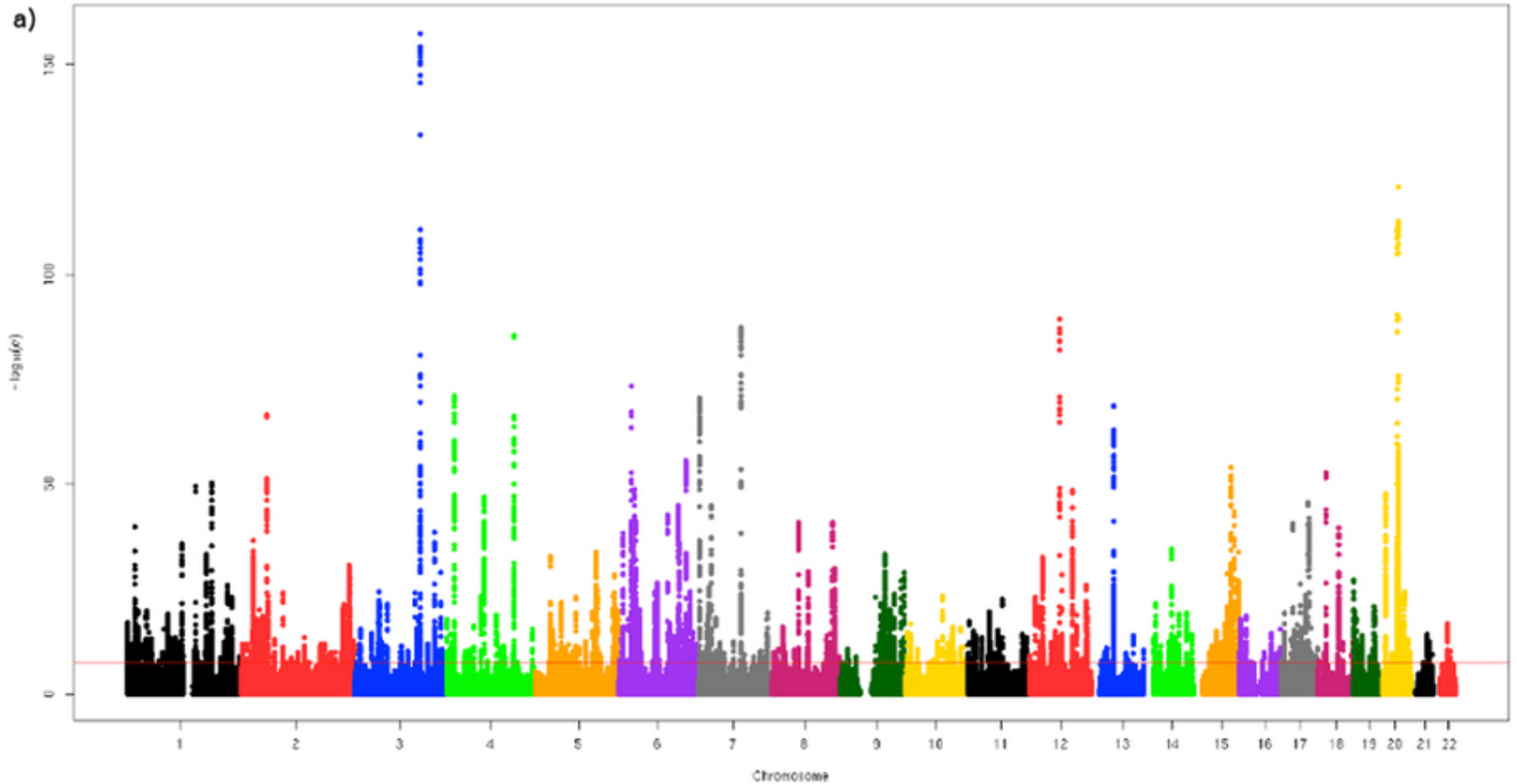
Supplementary Figure 2. 199 loci associated with adult height variation. Karyogram displaying the genome location of the 180 height SNPs identified from the primary meta-analysis (green) and the 19 secondary signals (red) discovered in the conditional analysis to be associated with height. The closest genes to the SNPs (gray) are followed by a MIM (blue) label if the gene underlies a skeletal growth-related Mendelian disorder described in OMIM. The plot was created using Affyrmation (http://genepipe.ngc.sinica.edu.tw/affyrmation/).



N=183,727

Variance explained = 10%

Lango Allen et al, Nature 2010



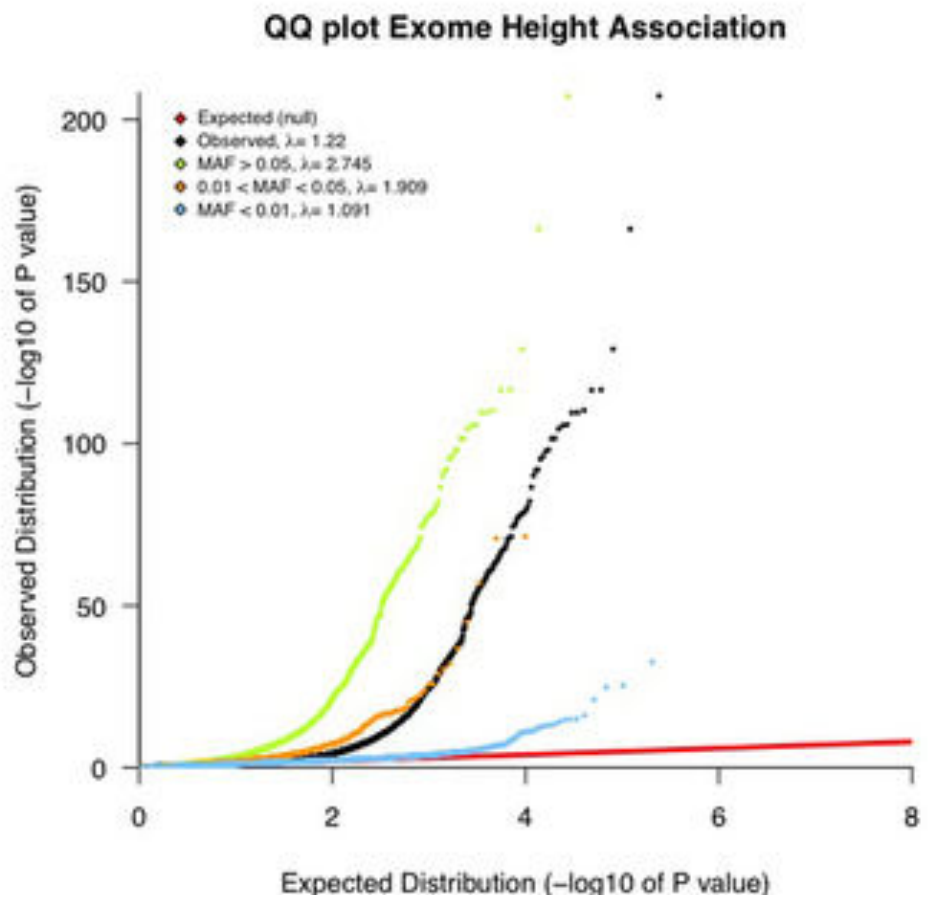
N= 253,288

697 SNPs (423 regions)

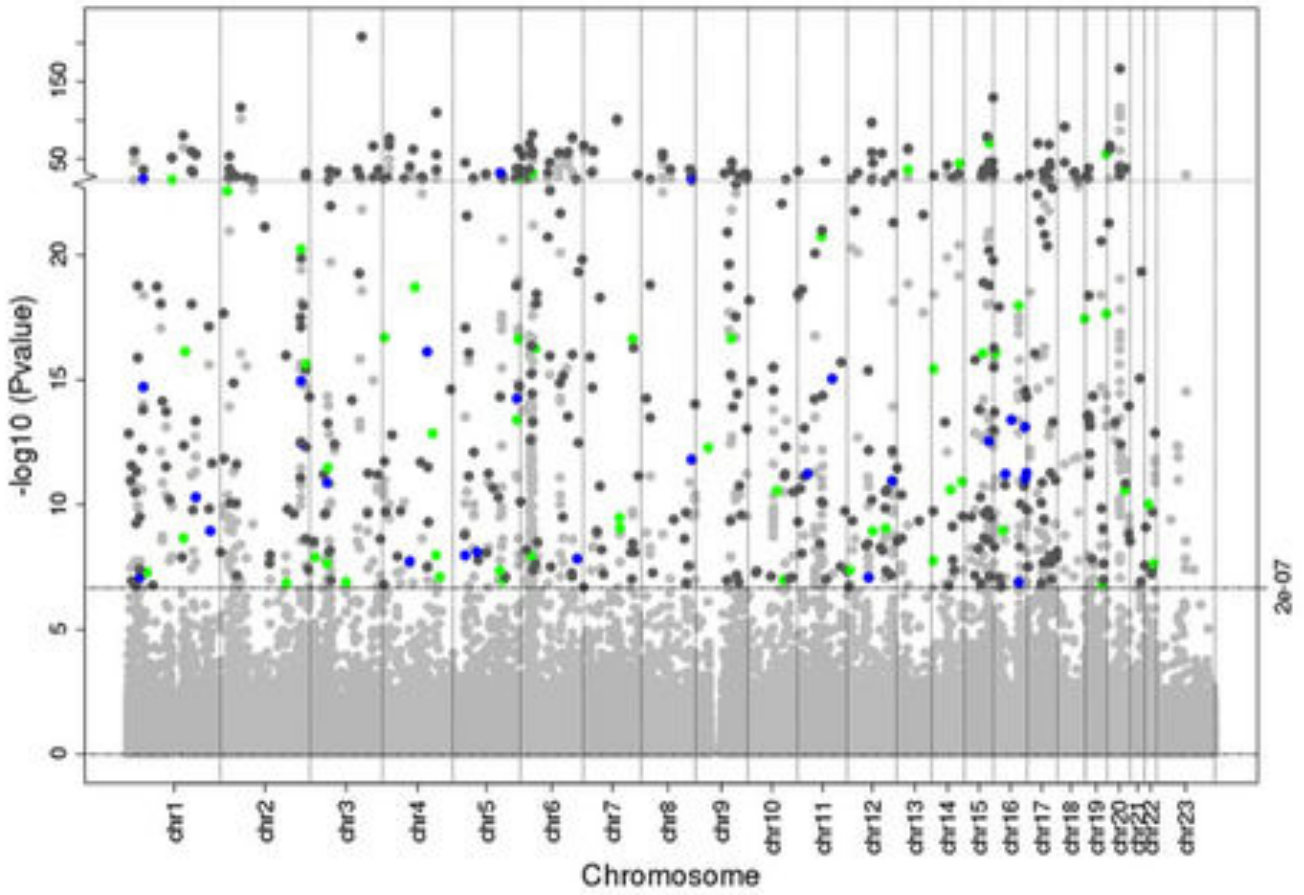
Variance explained ~ 20%

Wood et al, Nat Genet 2014

A



B

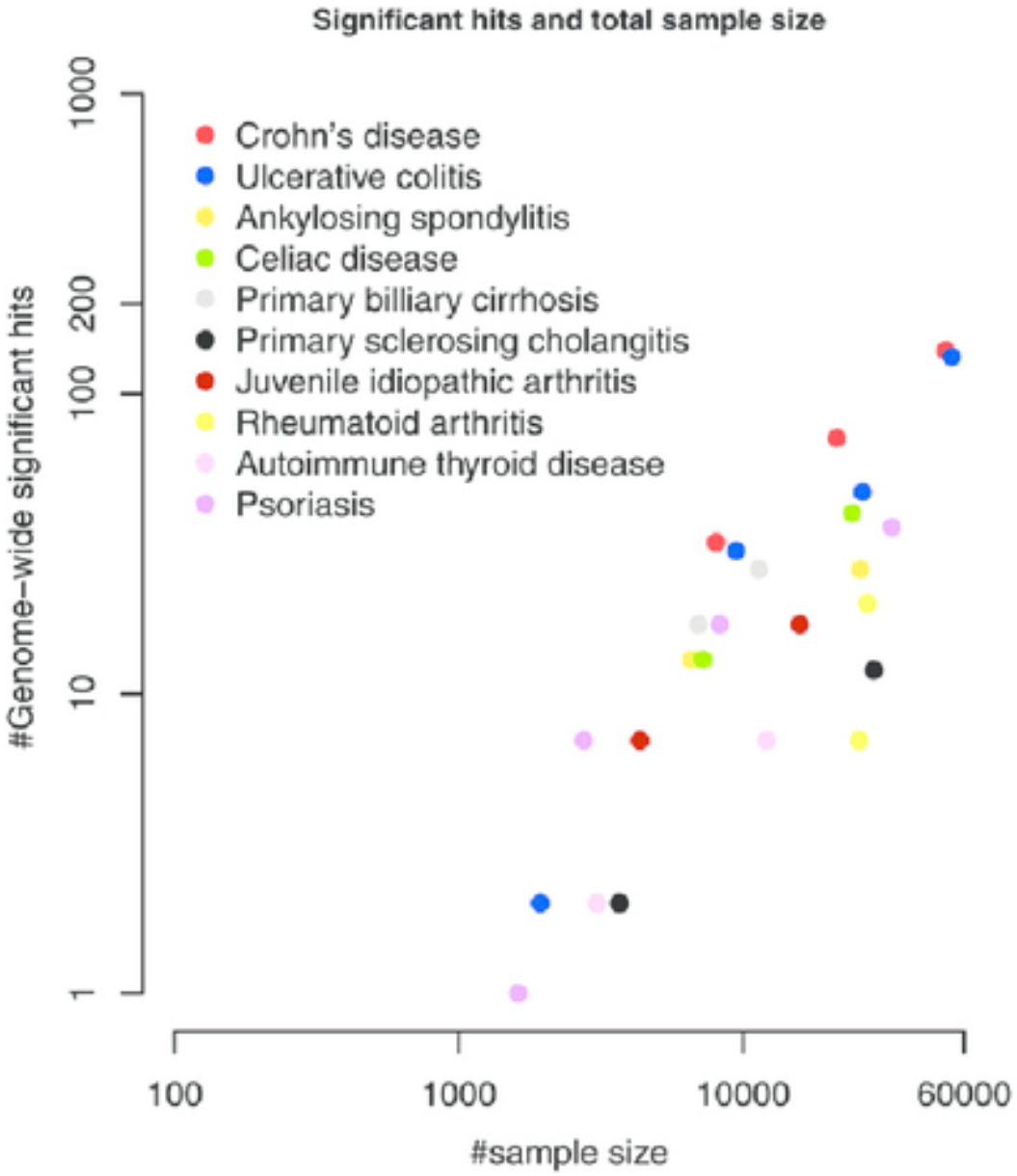


N= 458,927 (discovery) and N=252,501 (replication)

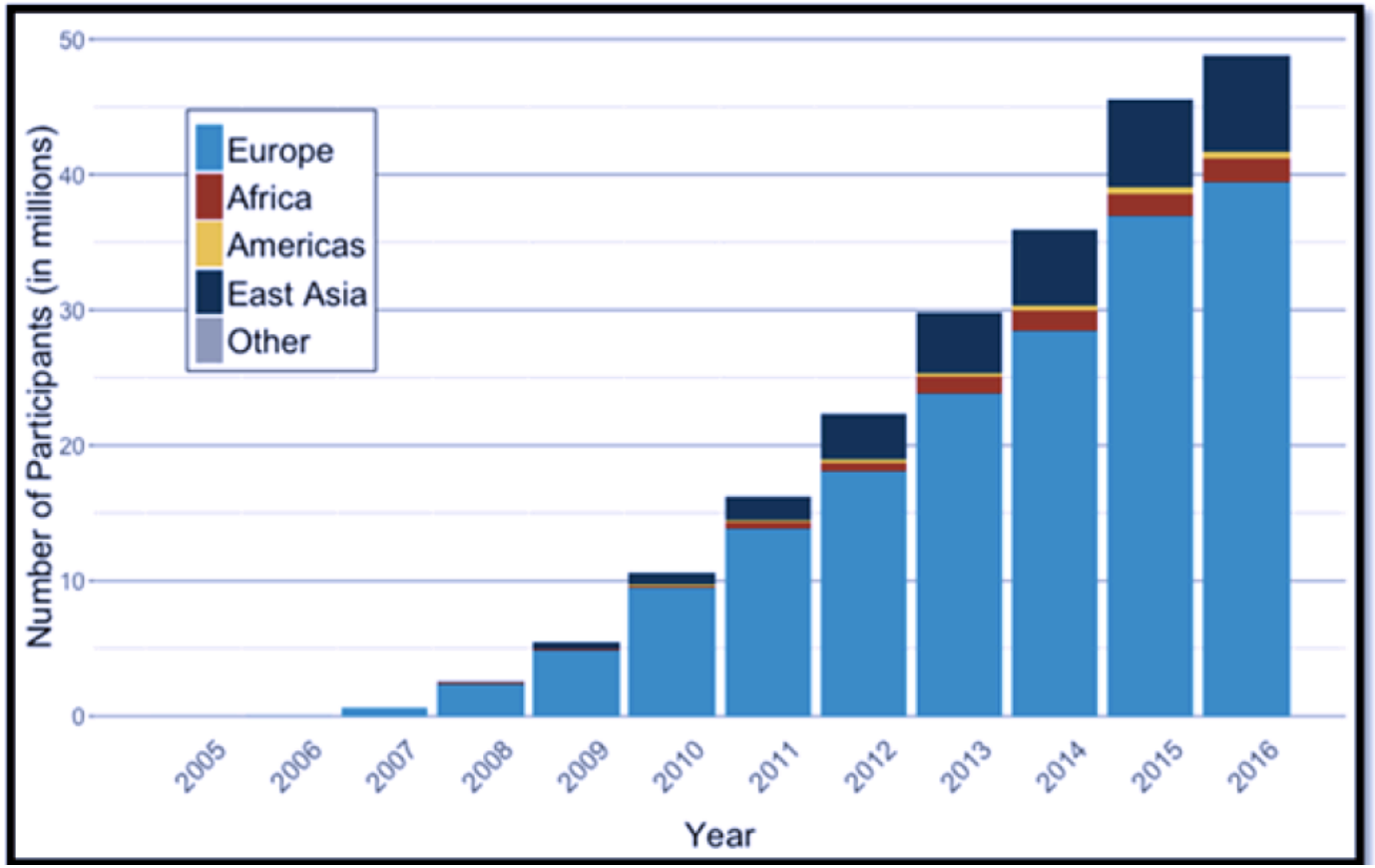
120 novel SNPs (83 with MAF <5%)

Variance explained = 27.4%

Marouli, Nature 2017



Chen, HMG 2014



Wojcik, https://www.biorxiv.org/content/early/2017/09/15/188094

Bias in the context of genetic epidemiology

- Ascertainment bias
  - Secondary phenotypes, e.g. Type 2 diabetes and BMI
- Survival bias
  - Might lead to a subtype analysis (milder form of disease)
- Respondent bias
  - Response rate has to differ by case-control status and genotype
- Diagnosis bias
  - Only a problem if the physician knows the genotype
- Recall bias
  - Not applicable in genetic epidemiology

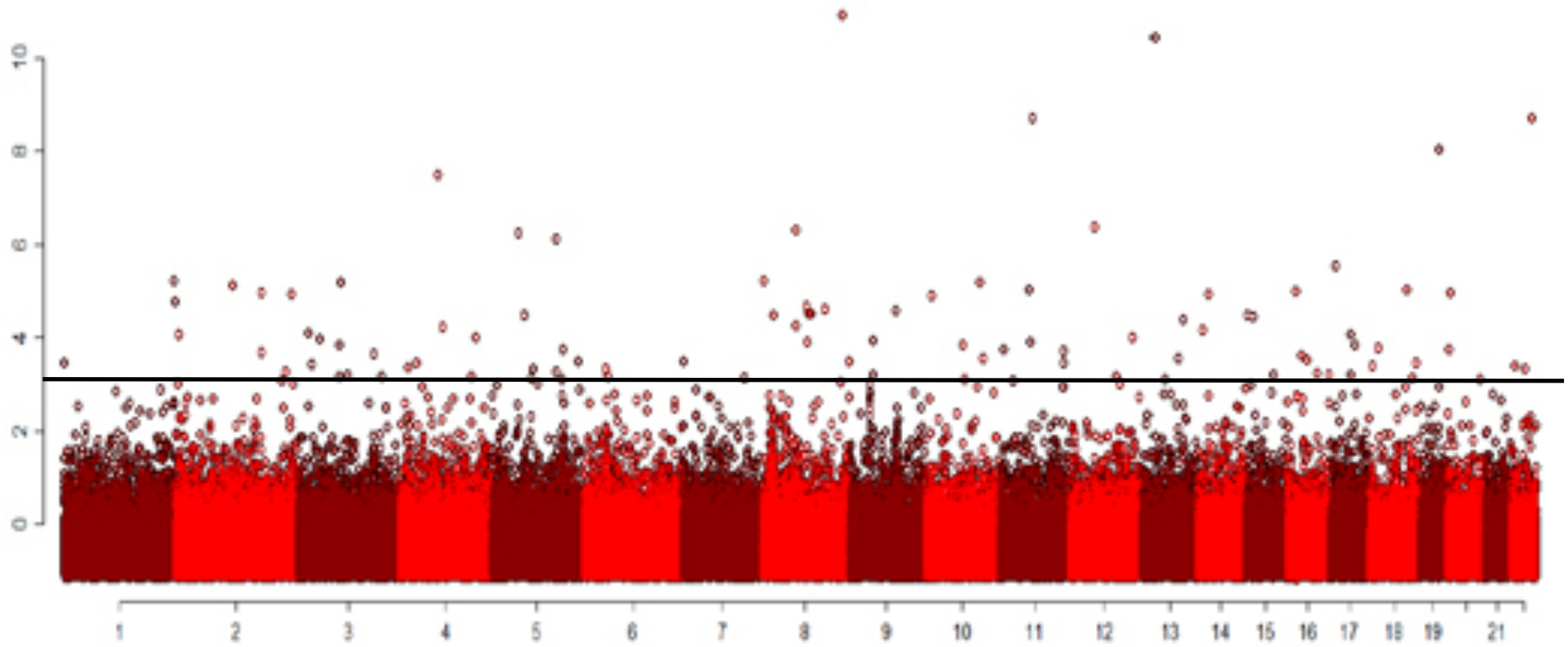
Note: This does not hold up for gene-environment interactions!

Differential genotyping error/missingness

- Systematic differences in how case and control samples were collected, handled, or genotyped can lead to spurious associations
  - DNA was collected from blood samples for cases and from cheek swabs for controls
  - Case samples have been sitting in the freezer for 15 years, control samples are new
  - Cases and controls were genotyped in different genotyping labs or by different platforms

There is one more dominating source of bias in genetic association studies – population stratification.

Genetic signatures of exceptional longevity in humans



Sebastiani, Science 2010

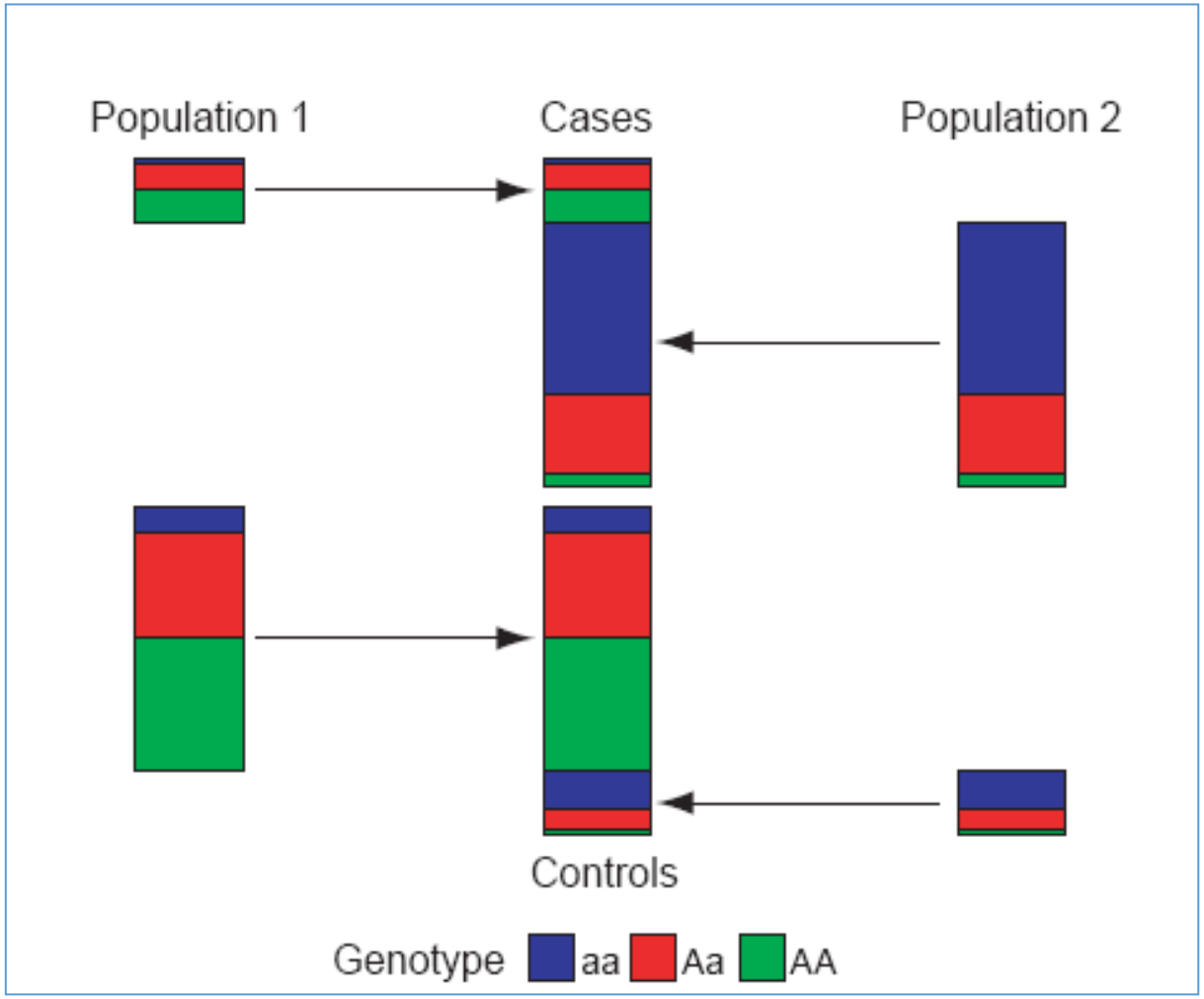
Retraction

AFTER ONLINE PUBLICATION OF OUR REPORT "GENETIC SIGNATURES OF EXCEPTIONAL LONGEVity in humans" (*1*), we discovered that technical errors in the Illumina 610 array and an inadequate quality control protocol introduced false-positive single-nucleotide polymorphisms (SNPs) in our findings. An independent laboratory subsequently performed stringent quality control measures, ambiguous SNPs were then removed, and resultant genotype data were validated using an independent platform. We then reanalyzed the reduced data set using the same methodology as in the published paper. We feel the main scientific findings remain supported by the available data: (i) A model consisting of multiple specific SNPs accurately differentiates between centenarians and controls; (ii) genetic profiles cluster into specific signatures; and (iii) signatures are associated with ages of onset of specific age-related diseases and subjects with the oldest ages. However, the specific details of the new analysis change substantially from those originally published online to the point of becoming a new report. Therefore, we retract the original manuscript and will pursue alternative publication of the new findings.

PAOLA SEBASTIANI,[1*] NADIA SOLOVIEFF,[1] ANNIBALE PUCA,[2] STEPHEN W. HARTLEY,[1] EFTHYMIA MELISTA,[3] STACY ANDERSEN,[4] DANIEL A. DWORKIS,[3] JEMMA B. WILK,[5] RICHARD H. MYERS,[5] MARTIN H. STEINBERG,[6] MONTY MONTANO,[3] CLINTON T. BALDWIN,[6,7] THOMAS T. PERLS[4*]

Science, 2011

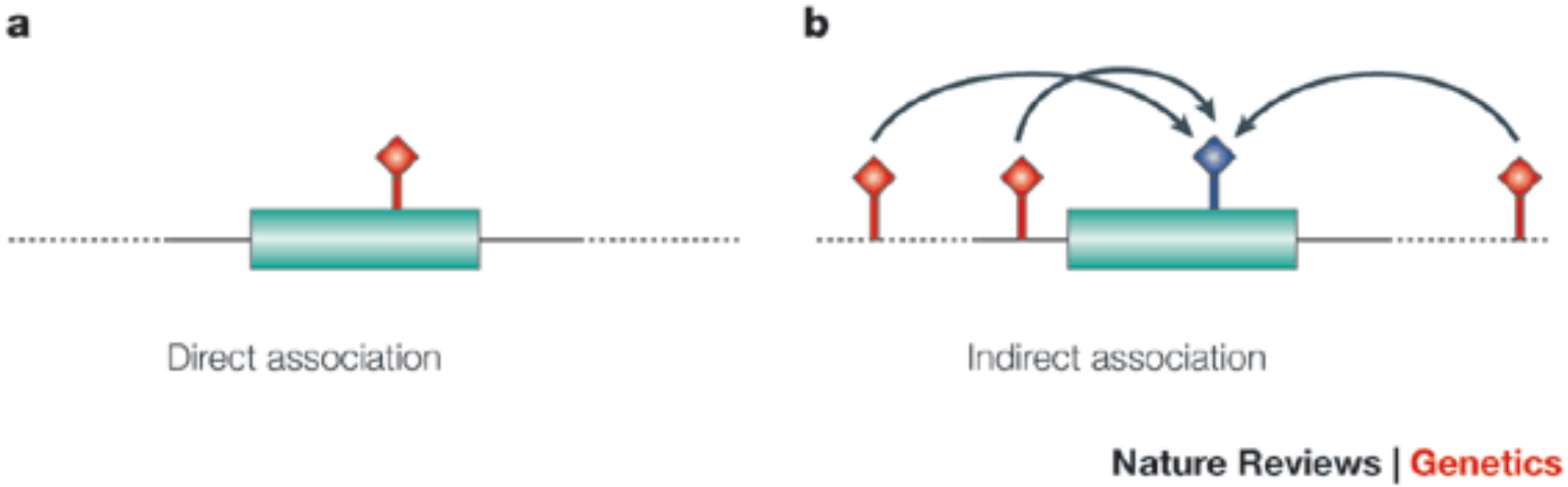
Population Stratification - Confounding by ancestry



Group differences in ancestry AND outcome

Marchini, Cardon et al. 2004; Price, Patterson et al. 2006

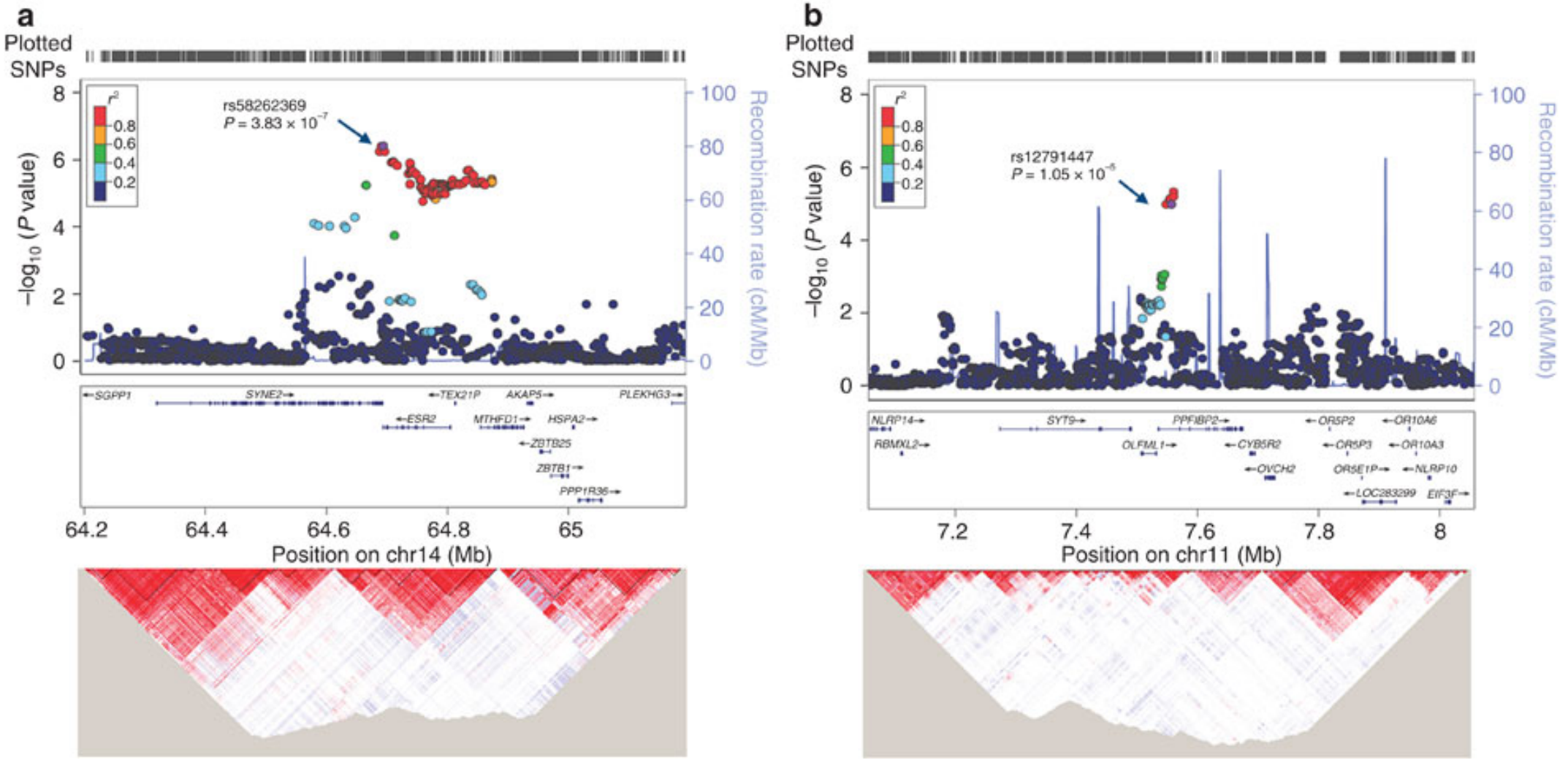
Fine-mapping



LD complicates things: Which SNP(s) is the causal SNP?

Hirschhorn & Daly. *Nature Reviews Genetics 2005*

Results from a prostate cancer GWAS



Wang. *Nature Comm, 2015*

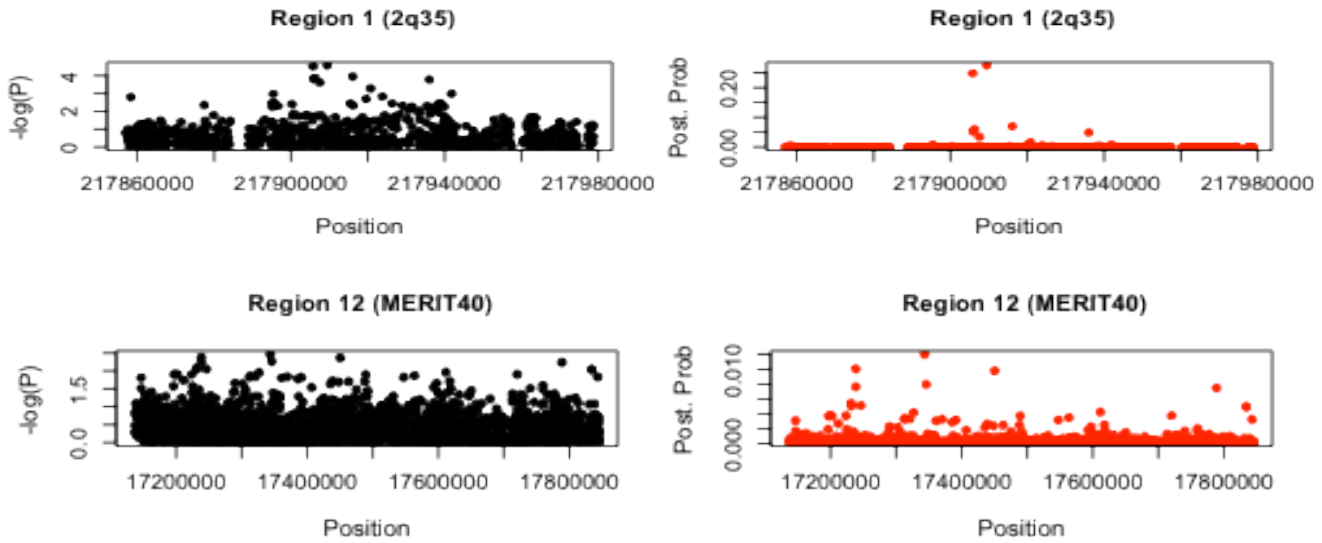
Fine-mapping approaches

- Conditional regression analysis
  - Rerun analysis and adjust for the most significant SNP, see if any other SNP remains significant. Keep going until no more significant SNPs
- Calculate posterior probabilities for each SNP
- Incorporate “functional” information to identify biological plausible SNPs
- Choose a set of “potentially causal variants” and take them forward for downstream analysis.

Fine-mapping

Approximate Bayesian analysis to estimate the posterior probability that a given SNP is causal

- Ratio of the likelihood from a logistic regression for SNP_i and the sum across all likelihoods for other SNPs in the region
- Assumes only one causal SNP in the region and that each SNP is equally likely *a priori* to be the causal variant



Maller, Nat Genet 2012

Lindstrom, Breast Can Res 2016

Incorporating functional annotation data in fine-mapping

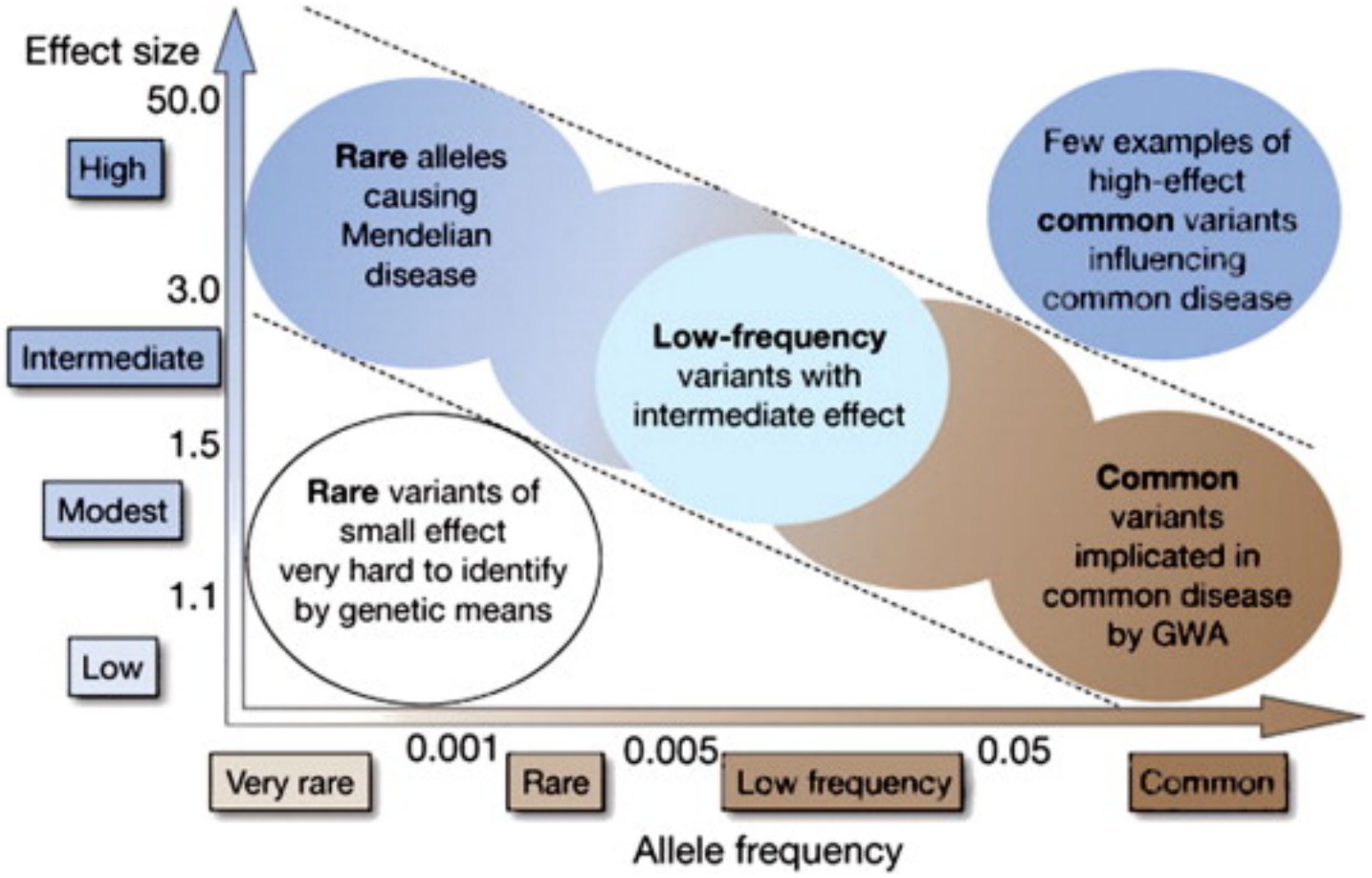
PAINTOR - **P**robabilistic **A**nnotation **INT**egrat**OR**

- Prioritizes SNPs based on posterior probabilities
- Allows for multiple causal variants at a locus
- Requires only summary association statistics and a reference population (e.g. 1000 Genomes)
- Integrates functional annotations (e.g. ENCODE)
- Estimates probability of causality in functional annotations from the data itself

Kichaev et al, PLoS Genetics 2014

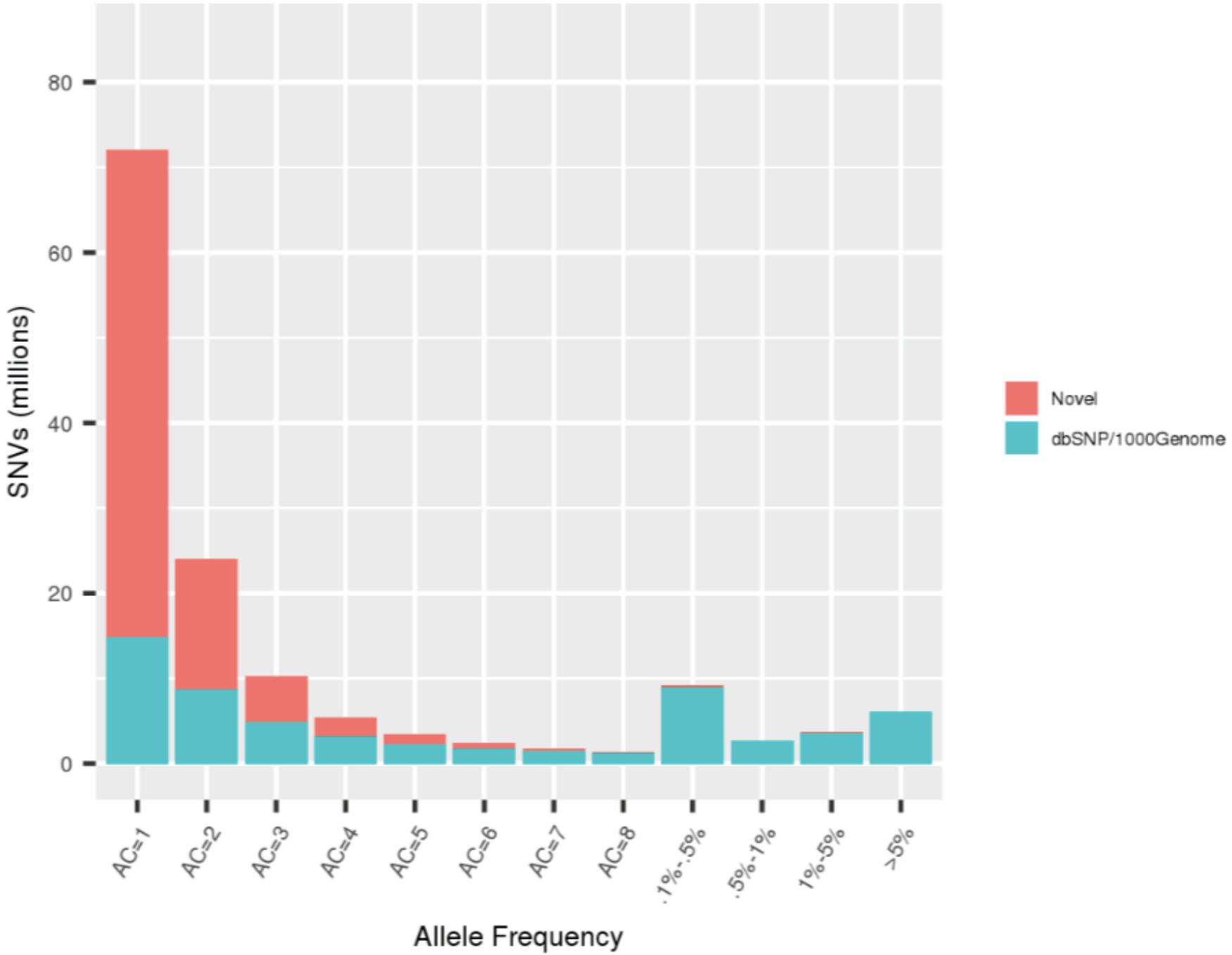
Rare variant association studies

Identifying genetic variation associated with disease



Manolio et al, Nature 2009

A recent study sequenced 10,545 human genomes and found more than 150 million variants



Telenti, PNAS 2016

Introduction – Rare variants

- Usually less than 1% (depending on who you ask)

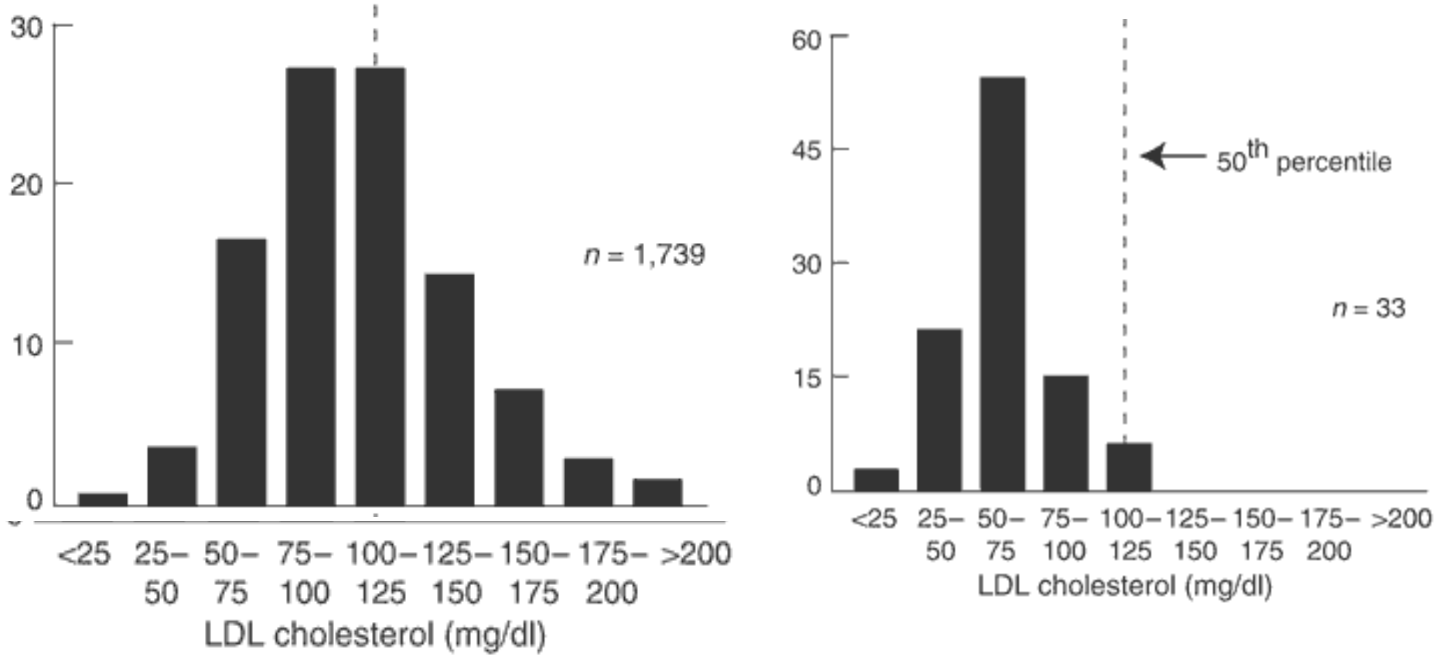
- Traditional single variant association analysis have low statistical power and/or are not valid
  - MAF=1% in 1,000 cases and 1,000 controls implies 40 minor alleles
  - Low cell counts lead to invalid statistical tests/low power

- Because the number of rare variants is much larger than the number of common variants, more stringent significance levels might be required, further reducing power

Why do we care about rare variants when they only affect a small proportion of the population?

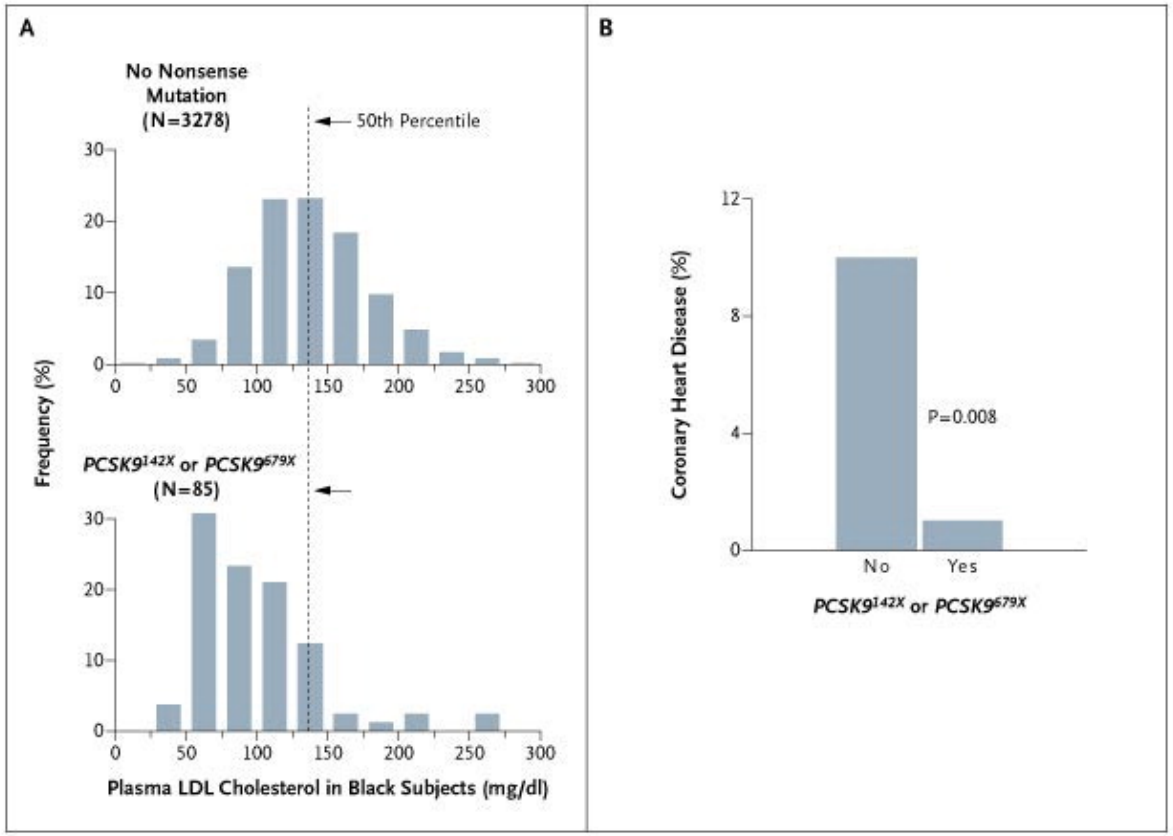
PCSK9 and LDL cholesterol

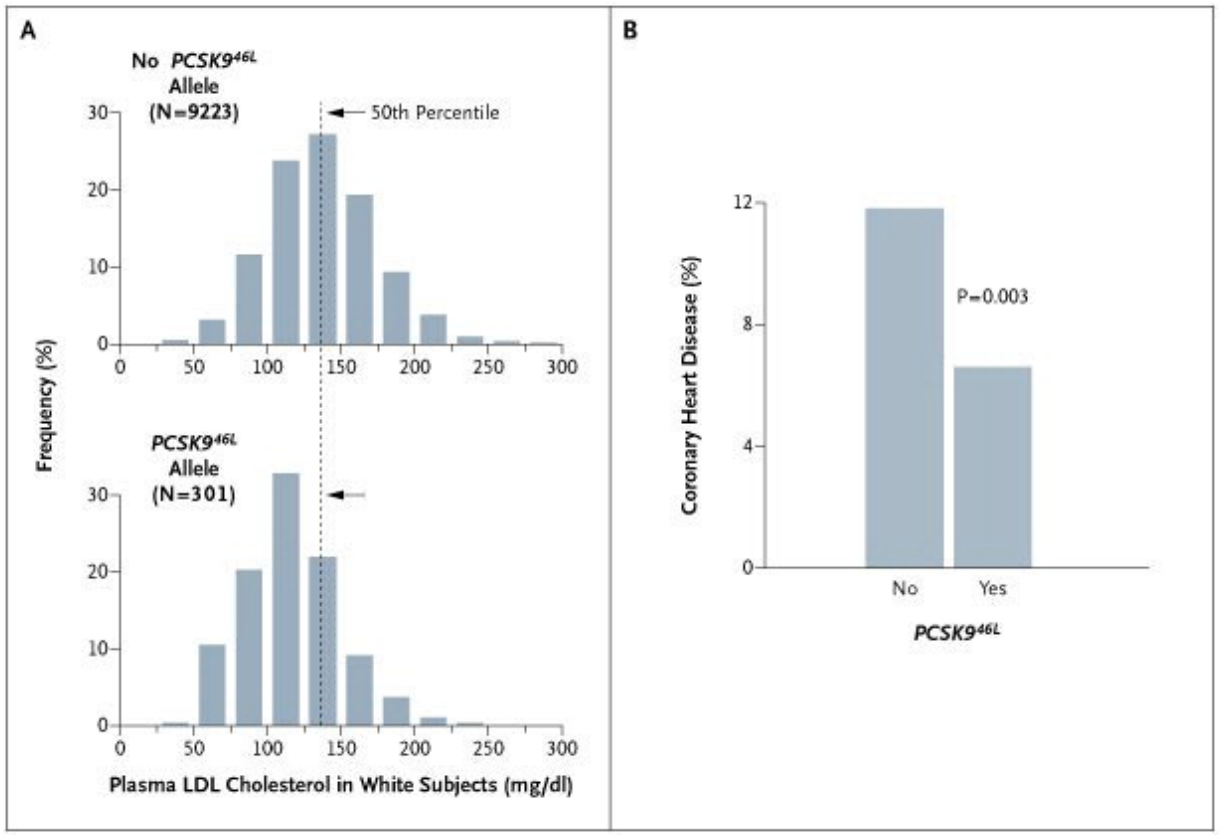
Plasma LDL-C levels in African American subjects without (left) and with (right) a nonsense mutation in *PCSK9*.



Cohen, Nat Genet 2005

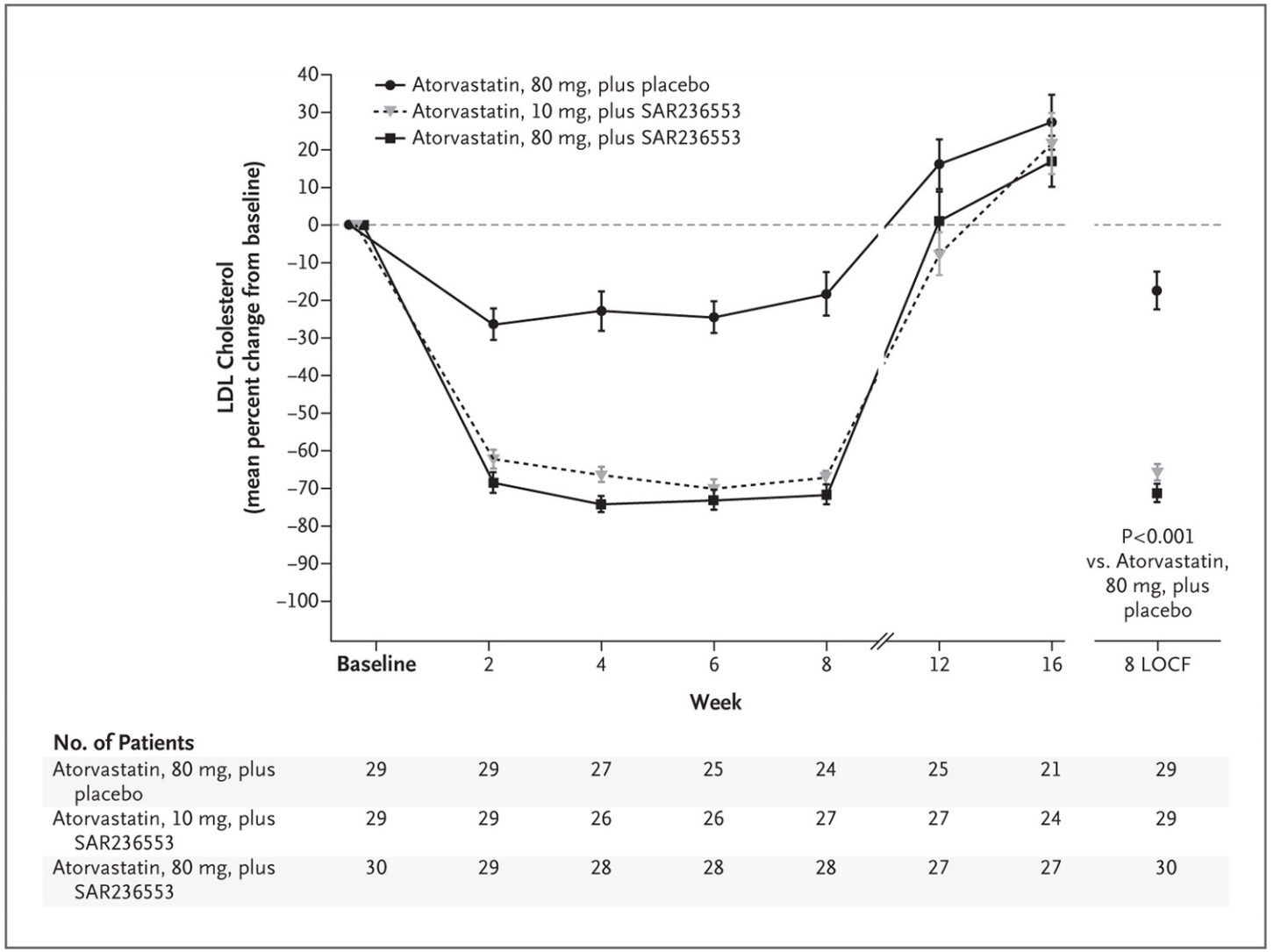
PCSK9 mutations and coronary heart disease





Cohen, NEJM 2005

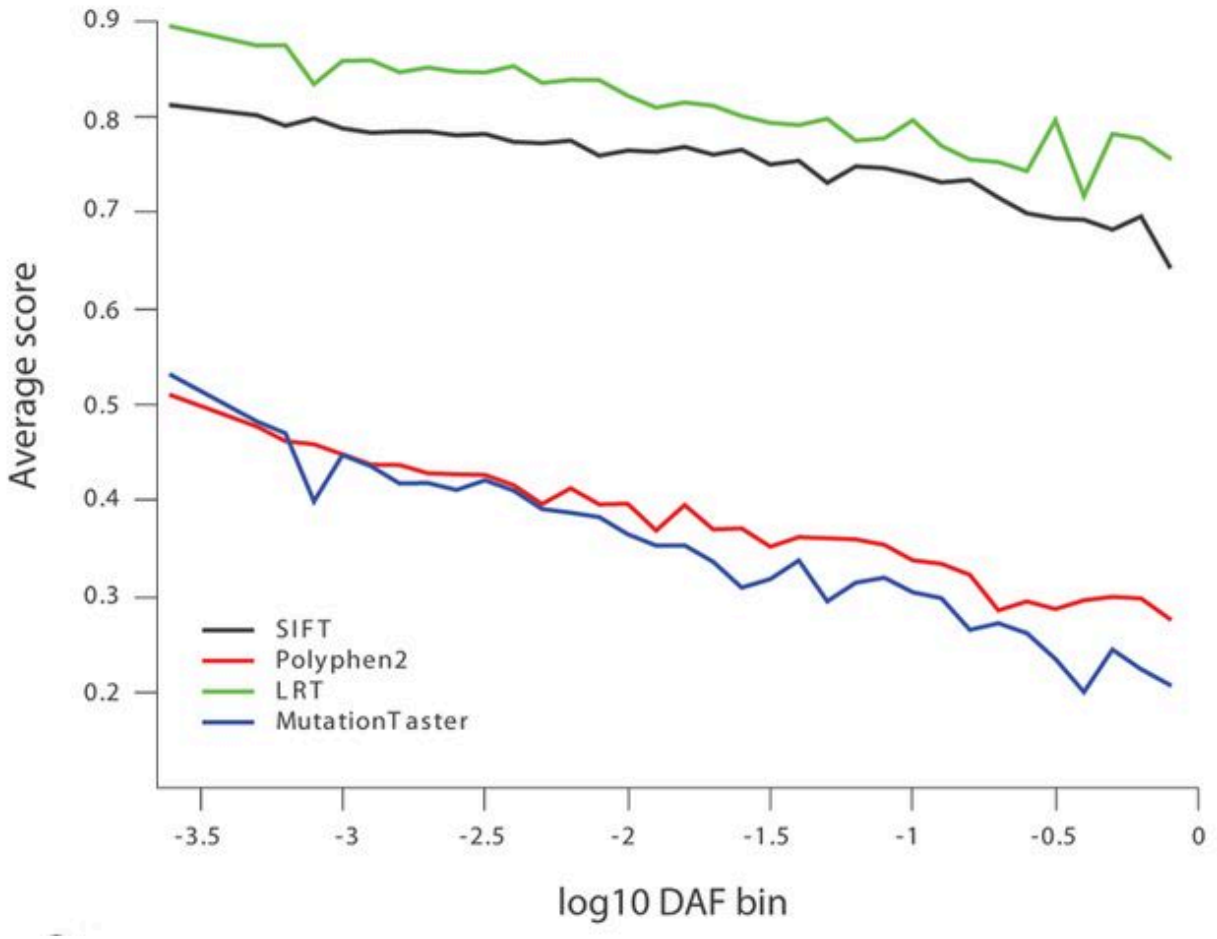
A *PCSK9* antibody decreases LDL (8-week trial)



No. of Patients								
Atorvastatin, 80 mg, plus placebo	29	29	27	25	24	25	21	29
Atorvastatin, 10 mg, plus SAR236553	29	29	26	26	27	27	24	29
Atorvastatin, 80 mg, plus SAR236553	30	29	28	28	28	27	27	30

Roth, NEJM 2012

Fig. 3 Signatures of purifying selection in protein-coding SNVs. Relationship between the evidence that a variant is functionally important and MAF for four different methods.



Jacob A. Tennessen et al. Science 2012;337:64-69

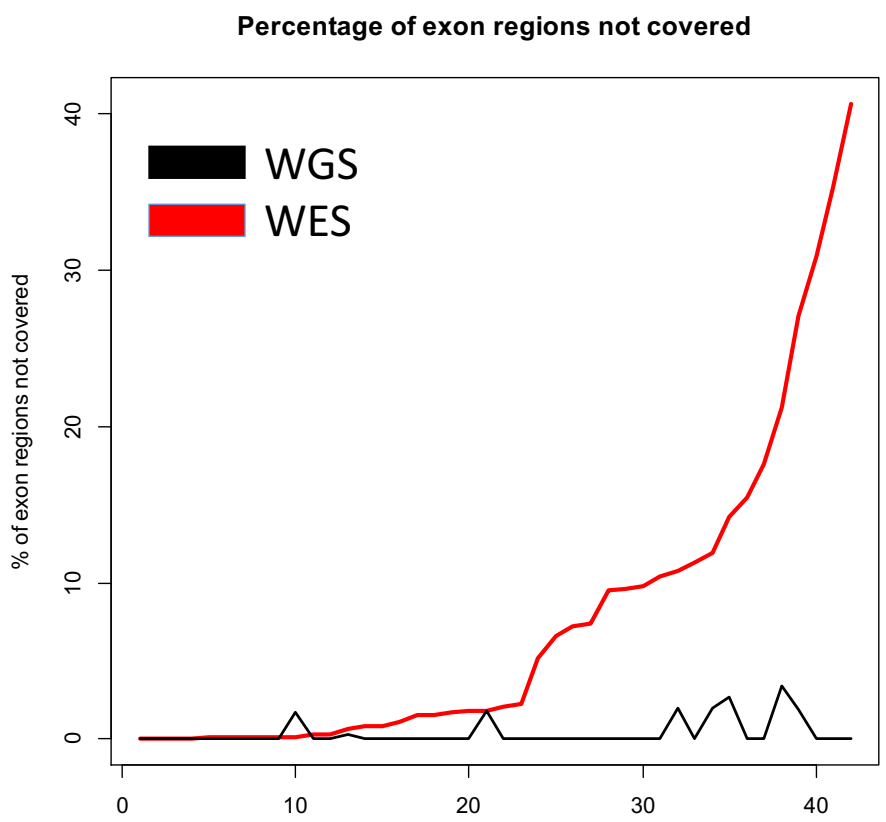
Study design for rare variant analysis

	Advantage	Disadvantage
High-depth WGS	can identify nearly all variants in the genome with high confidence	very expensive
Low-depth WGS	cost-effective and useful approach for association mapping	has limited accuracy for rare-variant identification and genotype calling; compared to deep sequencing, is subject to power loss if the same number of subjects is sequenced
Whole-exome sequencing	can identify all exomic variants; is less expensive than WGS	is limited to the exome
GWAS chip and imputation	inexpensive	has lower accuracy for imputed rare variants Will miss any variants unique to your sample
Exome chip (custom array)	much cheaper than exome sequencing	provides limited coverage for very rare variants and for non-Europeans is limited to target regions

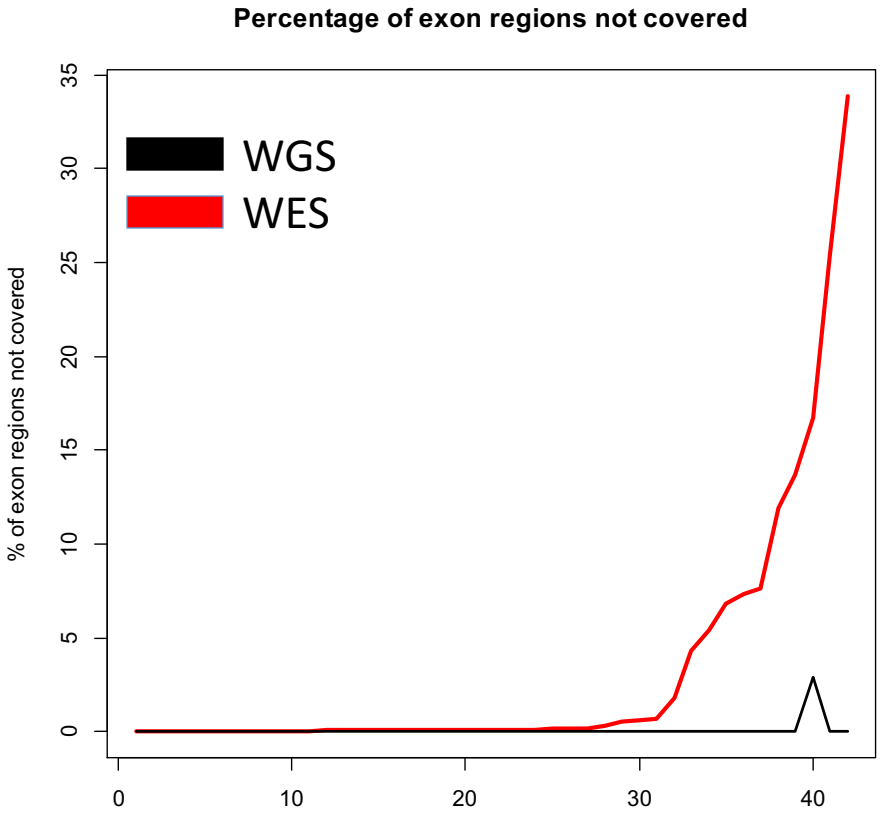
Lee, AJHG 2014

WGS has consistently good coverage across all of the exons whilst ESP exome coverage is more variable, missing up to 40% of exon regions for some genes.

>20x threshold



>10x threshold



Brad Chapman, Rory Kirchner

What to do?

- Many different rare variant tests are available.
  - Some are based on aggregating variants (“burden” tests)
    - CMC (Li and Leal, 2008)
    - WSS (Madsen and Browning, 2009)
    - Variable Threshold approach (Price, 2010)
  - Some are based on studying the distribution of variants
    - C-alpha (Neale, 2011)
    - SKAT (Wu, 2011)

Burden tests

- Collapse many variants into a single risk score
  - Combine minor allele counts into one variable
- Collapsing approach
  - Gene, pathways, functional annotations, etc
  - Much more straight-forward for coding regions
- Weighing
  - Variant type (predicted function)
  - Variant frequency

The Cohort Allelic Sums Test - CAST

Main Idea: Combine rare variants according to some (arbitrary) feature (gene, genetic region, functional category) and assess the new variable

Step 1: Create an indicator variable X for individual *j:*

$$X_j = \begin{cases} 1 \; if \; rare \; variants \; are \; present \\ 0 \; otherwise \end{cases}$$

Step 2: $\ln\left(\frac{p}{1-p}\right) = \alpha + \beta X$ (logistic regression)

Morgenthaler, Mutat Res 2007

Variant Collapsing – 2 approaches

i)

Subject	V1	V2	V3	V4	X
1	1	0	0	0	**1**
2	0	1	0	0	**1**
3	0	0	0	0	**0**
4	0	0	0	0	**0**
5	0	0	0	0	**0**
6	0	0	0	0	**0**
7	0	0	1	1	**1**
8	0	0	0	1	**1**

ii)

Subject	V1	V2	V3	V4	X
1	1	0	0	0	**1**
2	0	1	0	0	**1**
3	0	0	0	0	**0**
4	0	0	0	0	**0**
5	0	0	0	0	**0**
6	0	0	0	0	**0**
7	0	0	1	1	**2**
8	0	0	0	1	**1**

Drawback with burden tests

- Assume all variants in a set are causal and associated with a trait in the same direction. The common assumption is often that the rare allele increases disease risk

- If this is untrue, power is lost.

- Solution: Tests that look at the distribution of rare variants

The C-alpha test

- Main idea: Test whether observed variants either increase or decrease risk (or have no effect). Risk variants are expected to be more common in cases; protective variants more common in controls.

Position	Annotation	High Lipid Level	Low Lipid Level
21078358	Ala4481Thr	2	5
21078359	Ile4314Val	3	0
21078990	Arg4270Thr	6	3
21079417	Val4128Met	1	7
21083082	Thr3388Lys	2	1
21083637	Ser3203Tyr	6	0
21086035	Leu2404Ile	2	3
21086072	Glu2391Asp	2	2
21086127	Thr2373Asn	2	2
21086308	Val2313Ile	2	1
21087477	His1923Arg	6	12
21087504	Asn1914Ser	0	5
21087634	Asp1871Asn	2	0
21091828	Pro1143Ser	0	6
21091872	Arg1128His	0	3
21091918	Asp1113His	1	3
21106140	Thr498Asn	2	0
Singletons		6	4

Nonsynonymous variants discovered via targeted pooled sequencing in 192 individuals with extreme triglyceride levels. High counts represent the number of copies of the variant discovered in 96 individuals who have high triglycerides (defined as exceeding the 5% upper tail of the population distribution). Low counts represent the number of copies of the variant discovered in 96 individuals who have low triglycerides (lower 5% tail). The singletons are grouped together and listed as the penultimate row because its total count is second largest (10, versus 18 for the His1923Arg). For details about pooled sequencing, see Text S1.
doi:10.1371/journal.pgen.1001322.t001

***APOB* variant counts in individuals with high/low triglyceride levels.**

Neale, PLoS Genetics 2011

C-alpha test

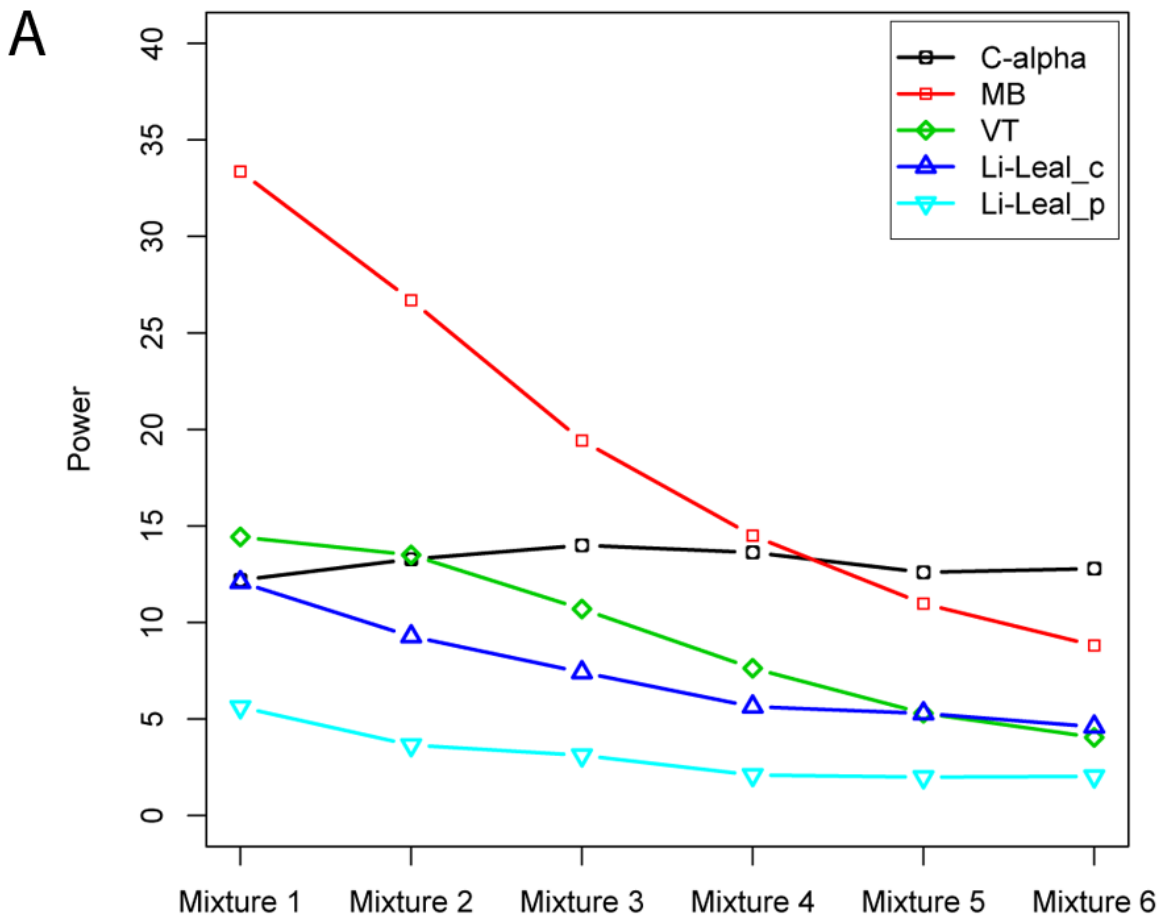
- If there is no association, variants are distributed randomly between cases and controls following a binomial (*n*,*p*) distribution. For example, if the case:control ratio is 1:1, a variant seen twice (doubleton) would be observed in cases *y* times where *y* is either 0, 1 and 2 with probability ¼, ½ and ¼, respectively.

- If there is an association, we typically will observe a higher proportion of doubletons with $y = 2$ and/or $y = 0$ than expected.

- C-alpha can be used to detect a pattern across the full set of rare variants. Under the null hypothesis, $p_i = p_0$. The alternative hypothesis is that p_i follows a mixture distribution across all *variants*, with some variants being detrimental ($p_i > p_0$), some neutral, and some protective ($p_i < p_0$).

Neale, PLoS Genetics 2011

Power comparisons for C-alpha, Madsen-Browning (MB), Variable threshold (VT), and CMC (binary: Li-Leal_p and count of rare variants: Li-Leal_c). As the mixing proportions between risk and protective variants increase (moving from 0, 10, 20, 30, 40 and 50% chance of any of the phenotypically relevant variants being protective), C-alpha maintains power, while other tests lose power.



Neale, PLoS Genetics 2011

SKAT: sequence kernel association test

- In contrast to the C-alpha test, SKAT is regression-based and thereby allows for adjustment of covariates.
- Uses a variance-component score test in a mixed-model framework to assess regression coefficients for rare variants.

$$logit\ P(y_i = 1) = \alpha_0 + \boldsymbol{\alpha}' \boldsymbol{X}_i + \boldsymbol{\beta}' \boldsymbol{G}_i$$

$\boldsymbol{y_i}$: case-control status; α_0: intercept; $\boldsymbol{\alpha}$ = [α_1,…, α_m]' is the vector of regression coefficients for the *m* covariates; $\boldsymbol{X_i}$: fixed effects of covariates; $\boldsymbol{\beta}$ = [β_1,…,β_p]' is the vector of regression coefficients for the *p* observed gene variants in the region; $\mathbf{G_i}$: (G_{i1}, G_{i2}, …, G_{ip}) genotypes for the *p* variants within the region

$$H_0: \boldsymbol{\beta = 0} \text{ or } \beta_1 = \beta_2 = \ldots = \beta_p = 0$$

Wu, AJHG 2011

Combined tests

- SKAT-O
  - Picks the best combination of SKAT and a burden test, and then corrects for the flexibility afforded by this choice. Specifically, if the SKAT statistic is Q1, and the squared score for a burden test is Q2, SKAT-O considers tests of the form (1-rho)*Q1 + rho*Q2, where rho is between 0 and 1.

Lee, AJHG 2012

Table 2. Summary of Statistical Methods for Rare-Variant Association Testing

	Description	Methods	Advantage	Disadvantage	Software Packages[a]
Burden tests	collapse rare variants into genetic scores	ARIEL test,[50] CAST,[51] CMC method,[52] MZ test,[53] WSS[54]	are powerful when a large proportion of variants are causal and effects are in the same direction	lose power in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	EPACTS, GRANVIL, PLINK/SEQ, Rvtests, SCORE-Seq, SKAT, VAT
Adaptive burden tests	use data-adaptive weights or thresholds	aSum,[55] Step-up,[56] EREC test,[57] VT,[58] KBAC method,[59] RBT[60]	are more robust than burden tests using fixed weights or thresholds; some tests can improve result interpretation	are often computationally intensive; VT requires the same assumptions as burden tests	EPACTS, KBAC, PLINK/SEQ, Rvtests, SCORE-Seq, VAT
Variance-component tests	test variance of genetic effects	SKAT,[61] SSU test,[62] C-alpha test[63]	are powerful in the presence of both trait-increasing and trait-decreasing variants or a small fraction of causal variants	are less powerful than burden tests when most variants are causal and effects are in the same direction	EPACTS, PLINK/SEQ, SCORE-Seq, SKAT, VAT
Combined tests	combine burden and variance-component tests	SKAT-O,[64] Fisher method,[65] MiST[66]	are more robust with respect to the percentage of causal variants and the presence of both trait-increasing and trait-decreasing variants	can be slightly less powerful than burden or variance-component tests if their assumptions are largely held; some methods (e.g., the Fisher method) are computationally intensive	EPACTS, PLINK/SEQ, MiST, SKAT
EC test	exponentially combines score statistics	EC test[67]	is powerful when a very small proportion of variants are causal	is computationally intensive; is less powerful when a moderate or large proportion of variants are causal	no software is available yet

Lee, AJHG 2014

Issues in rare variant analysis (i)

- Which variants to include?
  - All variants
  - Some pre-selected (or empirically estimated) threshold
  - Predicted impact (SIFT, PolyPhen-2, CADD)

- How to test non-exonic regions?
  - Rare variants are often grouped by gene making variant grouping straight-forward in exome studies.
  - For whole-genome analysis, alternative approaches such as sliding window or additional functional annotations (conserved regions, regulatory regions etc) can be used

Issues in rare variant analysis (ii)

- Which association test to use
  - Performance of various tests will depend on the underlying genetic architecture of the trait of interest.
    - If we believe that there are multiple variants with risk-increasing effects, burden tests are most powerful
    - If we believe that there is a mixture of risk increasing and risk decreasing variants and/or most variants do not have an effect, variance-component methods are most powerful
  - If no prior information is available, multiple approaches can be used (e.g. both burden and variance component methods). Have to consider multiple testing.

Issues in rare variant analysis (iii)

- Population stratification
  - Emerging field – it is not clear how effective principal components (or linear mixed models) are for population stratification adjustment
  - Studies have suggested that it is not more effective to generate principal components on rare variants compared to common variants
  - However, principal components can be used to identify controls that are closely matched on ancestry to the cases

Issues in rare variant analysis (iv)

- In general, rare variants are more difficult to impute

- We have talked before about the danger of genotyping your cases on one array (or version of array) and the controls of another array. For rare variants (e.g. exome arrays), this might cause even larger issues!

Issues in rare variant analysis (v)

- Replication is more complex for rare variants:
  - Since the variants are by definition rare, they might be unique to the discovery population
  - For single variants, replication is fairly straightforward: genotype the variant in the replication population
  - For gene-based association tests: Sequencing the gene (or region) can identify additional variants
  - Choose whichever approach which allow you to maximize number of samples in your replication!