Introduction to GWAS Data

Robbee Wedow, University of Colorado Boulder

@robbeewedow

Add Health

The National Longitudinal Study of Adolescent to Adult Health

Outline

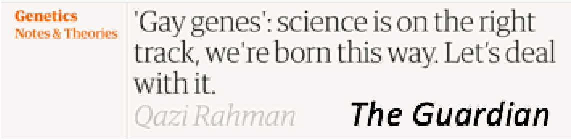
- Introduction to genome-wide association studies (GWAS)
- GWAS research (Educational Attainment)
- Polygenic Scores (David Braudt)
- Genetic data format (PLINK)
- Obtaining Add Health GWAS data
- Other considerations

INTRODUCTION TO GWAS

GWAS allows us to gain leverage on the public fascination with “nature or nurture?”

TIME
WHY YOUR DNA ISN'T YOUR DESTINY
The new science of epigenetics reveals how the choices you make can change your genes —and those of your kids

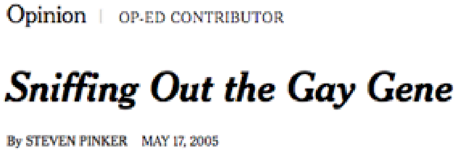
TIME
What scientists have uncovered about HOW MEMORY WORKS and how to improve it
THE I.Q. GENE?


Genetics
Notes & Theories

'Gay genes': science is on the right track, we're born this way. Let's deal with it.

Qazi Rahman

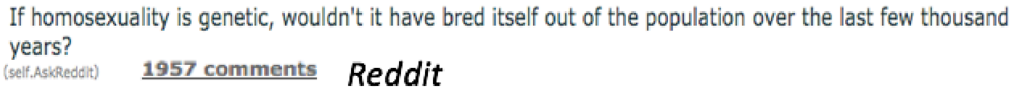
The Guardian


Opinion | OP-ED CONTRIBUTOR

Sniffing Out the Gay Gene

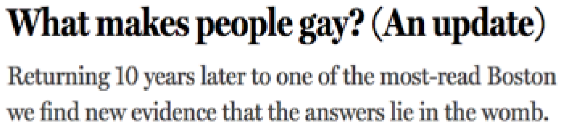
By STEVEN PINKER MAY 17, 2005

New York Times


If homosexuality is genetic, wouldn't it have bred itself out of the population over the last few thousand years?

(self.AskReddit) 1957 comments

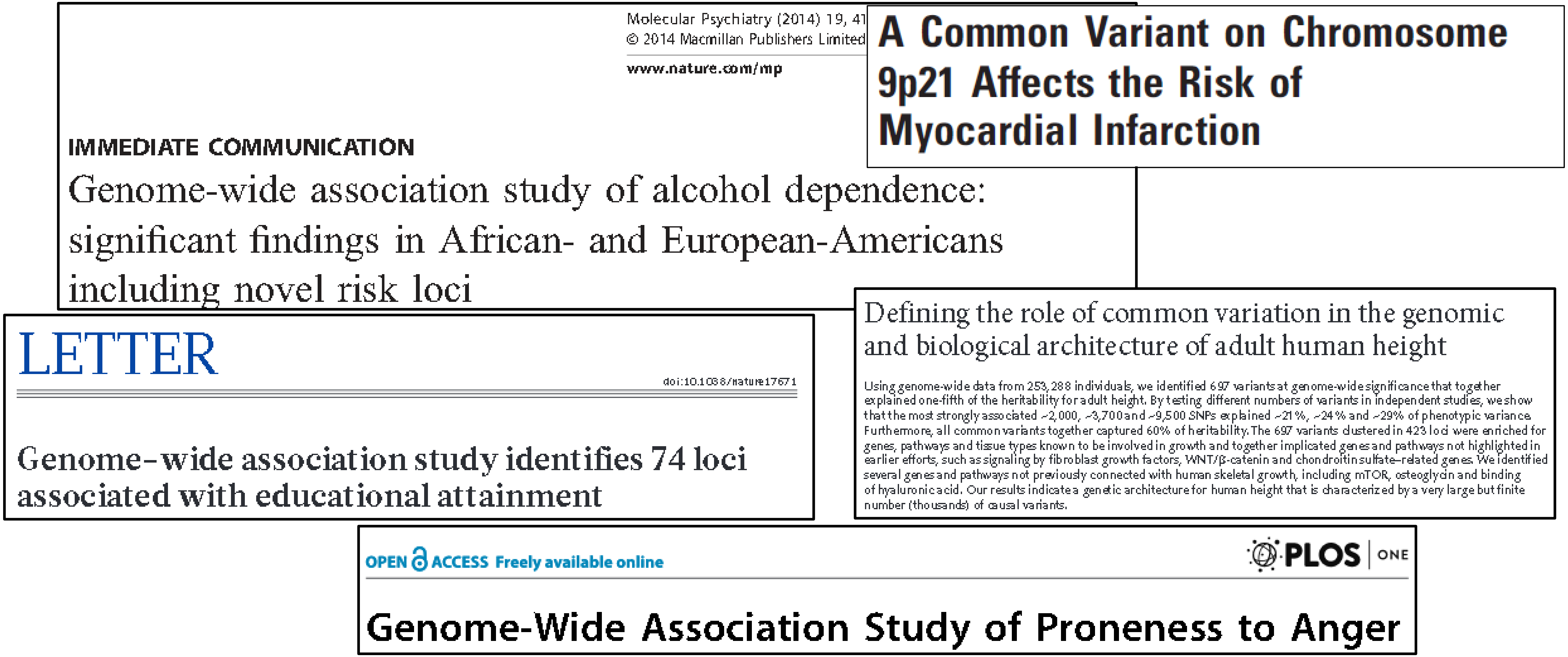
Reddit


What makes people gay? (An update)

Boston globe

Returning 10 years later to one of the most-read Boston Globe stories, we find new evidence that the answers lie in the womb.

“Inherited Disorders” Encompasses a Broad Spectrum of Diseases and Traits



Molecular Psychiatry (2014) 19, 41
© 2014 Macmillan Publishers Limited
www.nature.com/mp

IMMEDIATE COMMUNICATION

Genome-wide association study of alcohol dependence: significant findings in African- and European-Americans including novel risk loci

A Common Variant on Chromosome 9p21 Affects the Risk of Myocardial Infarction

LETTER

doi:10.1038/nature17671

Genome-wide association study identifies 74 loci associated with educational attainment

Defining the role of common variation in the genomic and biological architecture of adult human height

Using genome-wide data from 253,288 individuals, we identified 697 variants at genome-wide significance that together explained one-fifth of the heritability for adult height. By testing different numbers of variants in independent studies, we show that the most strongly associated ~2,000, ~3,700 and ~9,500 SNPs explained ~21%, ~24% and ~29% of phenotypic variance. Furthermore, all common variants together captured 60% of heritability. The 697 variants clustered in 423 loci were enriched for genes, pathways and tissue types known to be involved in growth and together implicated genes and pathways not highlighted in earlier efforts, such as signaling by fibroblast growth factors, WNT/β-catenin and chondroitin sulfate–related genes. We identified several genes and pathways not previously connected with human skeletal growth, including mTOR, osteoglycin and binding of hyaluronic acid. Our results indicate a genetic architecture for human height that is characterized by a very large but finite number (thousands) of causal variants.

OPEN ACCESS Freely available online

PLOS | ONE

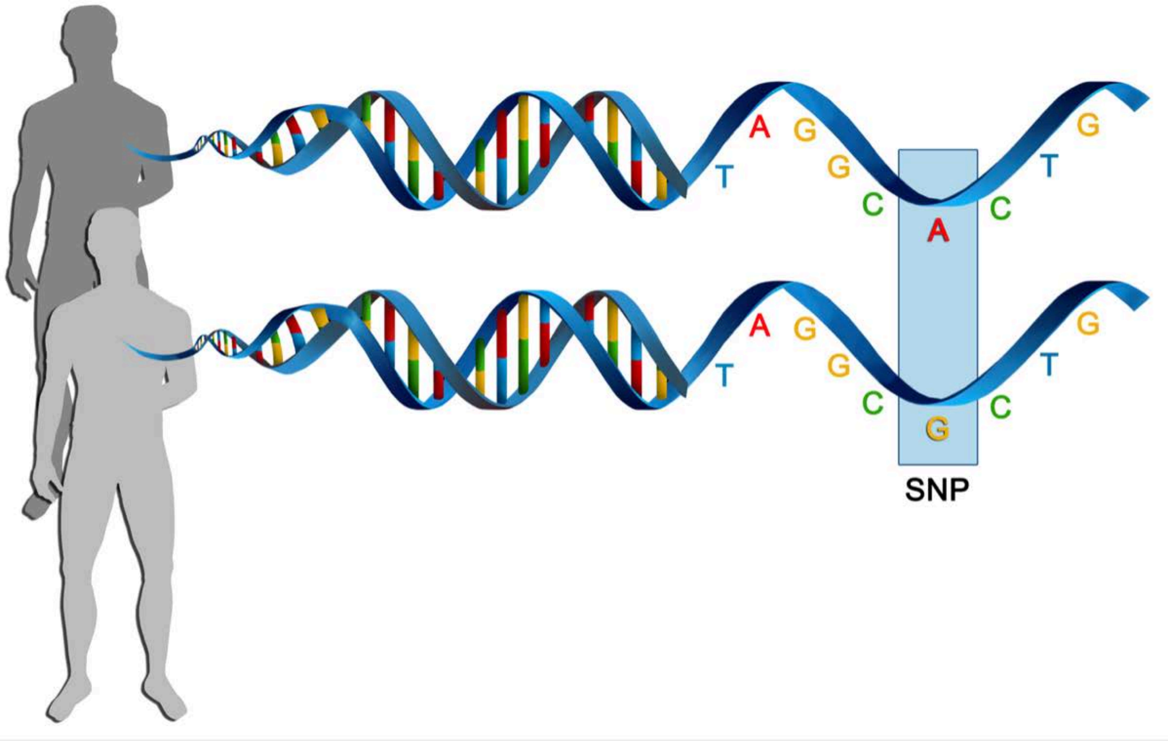
Genome-Wide Association Study of Proneness to Anger

Definition: Genome-Wide Association Study (GWAS)

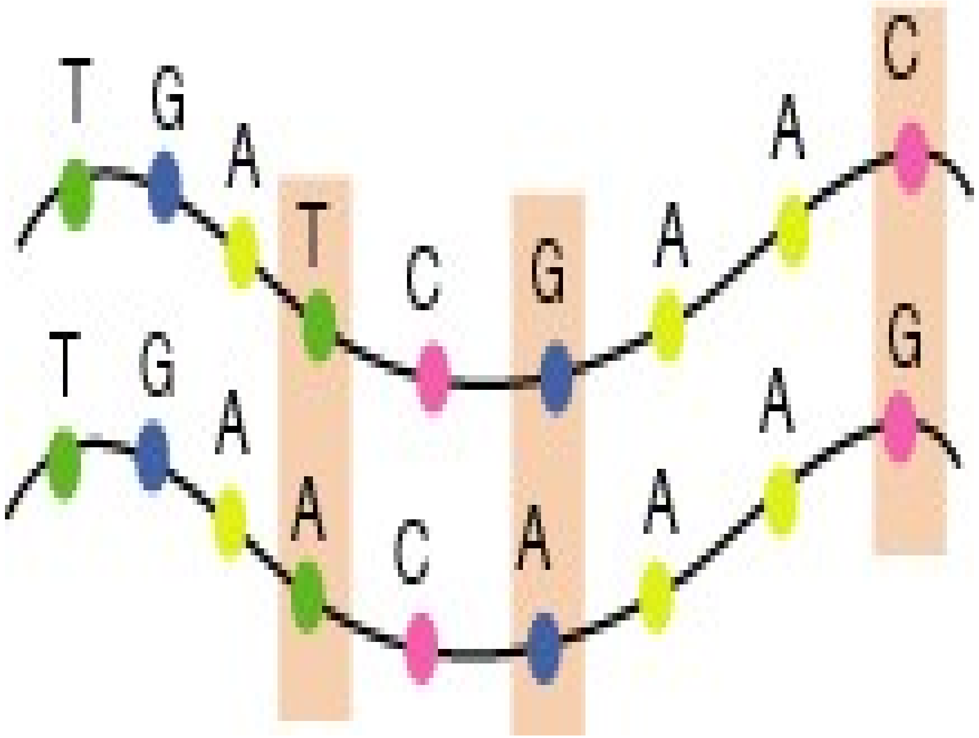
- A tool to evaluate the genetic basis of disease/phenotypes
- Study that surveys the genome for highly associated genetic variants
- Enables testing of multiple, genome-wide (~40-100 million) variants without any prior hypothesis
- GWAS genetic metric: the SNP

Genome-Wide Association Study (GWAS)

- The workhorse of gene discovery and much follow-up work (including gene-by-environment interaction studies) in modern statistical and social science genetics
- An atheoretical approach to the discovery of genetic associations across the base unit of molecular analyses, the single nucleotide polymorphism (SNP)
- The effects of SNPs across the genome are small and additive → therefore need enormous sample sizes to have the power to find these effects



Single Nucleotide Polymorphisms (SNPs)



- Single nucleotide polymorphisms (SNPs) are DNA sequence variations that occur when a single nucleotide (A,T,C,or G) in the genome sequence is altered
- Millions of SNPs in the genome!

Discovering genetic effects

In a **Genome-Wide Association Study (GWAS)**, we run a separate regression for every SNP *j* measured:

$$Phenotype_i = \mu + \beta_j x_{ij} + \gamma \cdot Controls_i + u_{ij}$$

- ➢ x_{ij}: genotype of individual i for SNP j
- ➢. β_j : predictive effect of SNP j

- Several methodological challenges arise. In particular:

1. Multiple hypothesis testing ➔ Apply stringent significance threshold

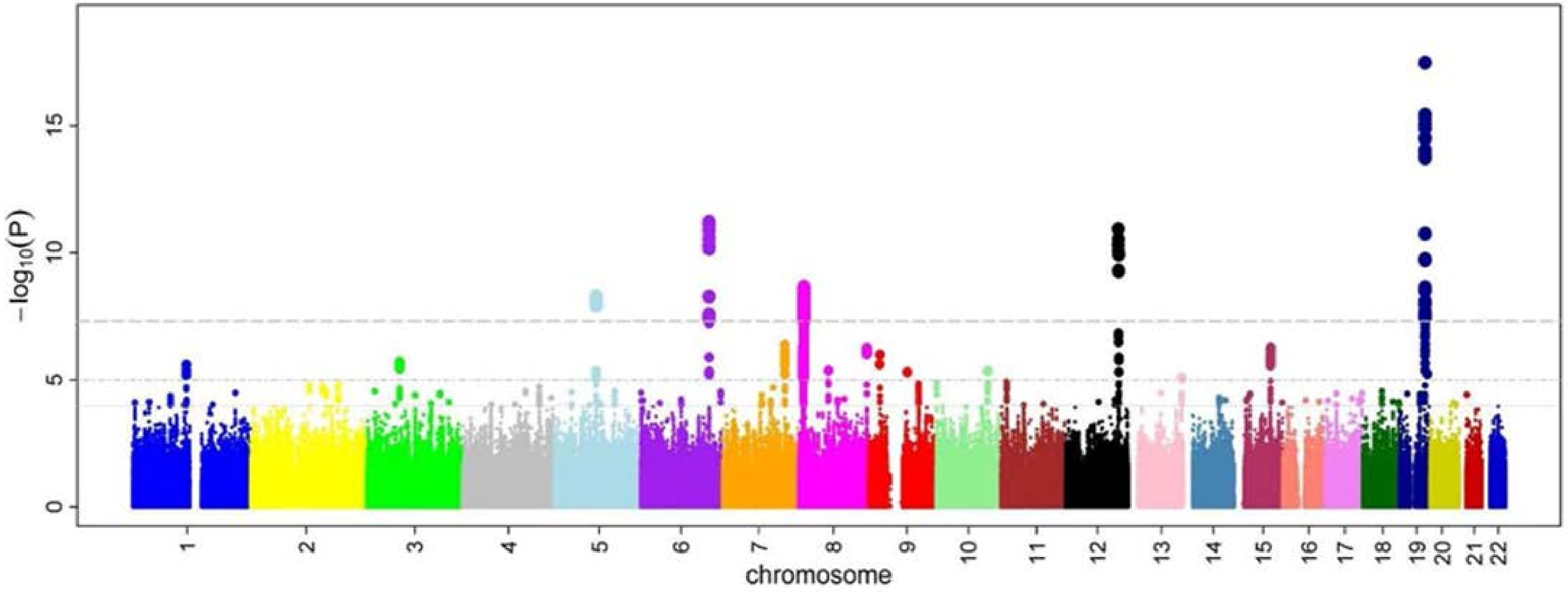
2. Need statistical power (due to small effect sizes) ➔ Use very large sample

3. Population stratification ➔ Use ethnically homogenous sample (to date, Europeans)

The Sample Size Tradeoff

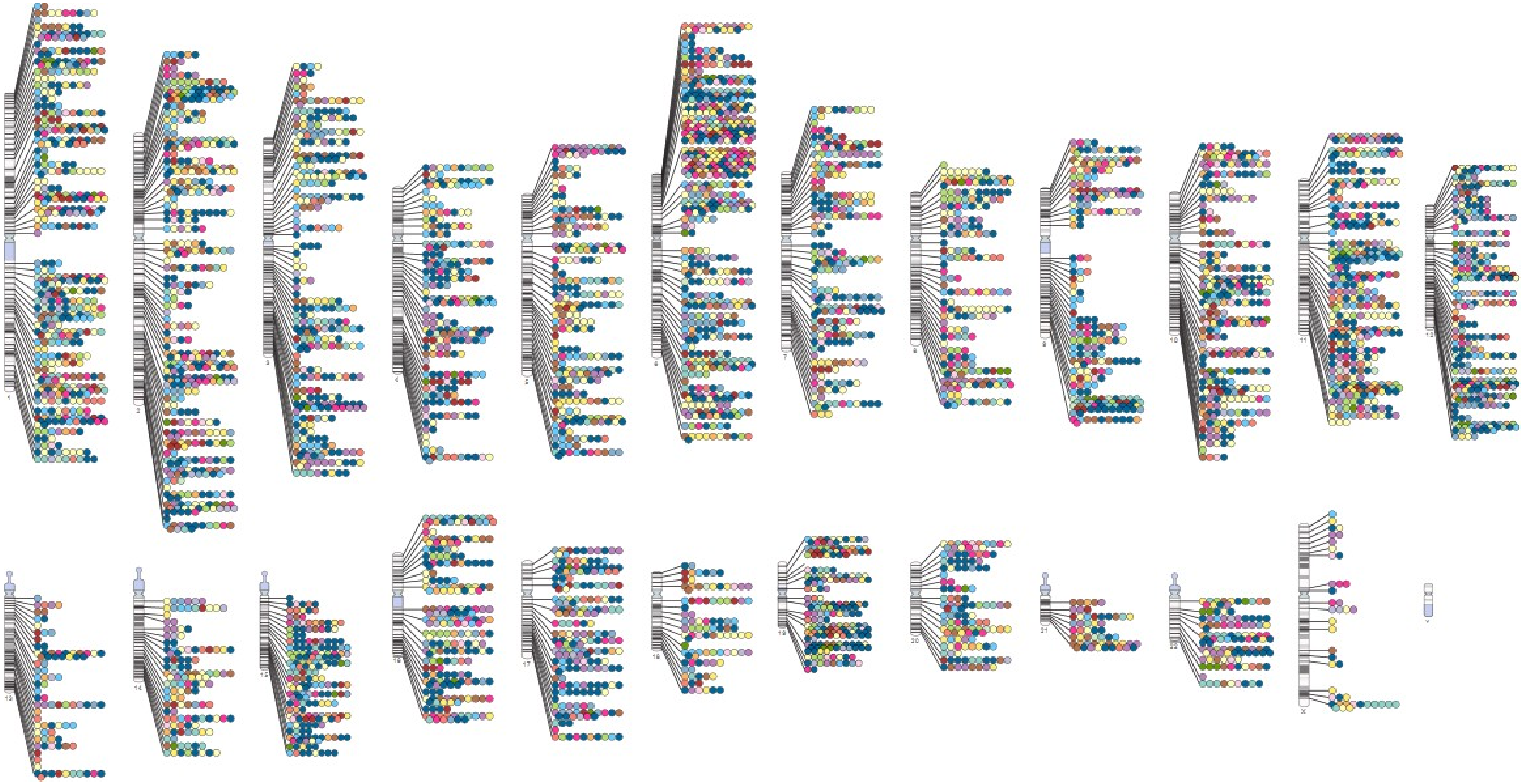
- Enormous sample size means a trade-off in outcome (phenotype) measurement
- Phenotypes are necessarily noisy
- GWAS casts a “wide net” with a limited set of control variables:
    - Measure genetic effects that are in fact “total effects”
    - Genetic effects include not only direct genetic effects, but also effects that occur through mediating and moderating environmental mechanisms, the focus of interesting follow-up gene-by-environment interaction studies

Manhattan Plots



- SNPs on *x*-axis
- -log of *p*value on *y*-axis
- Lowest *p*values = tallest peaks
- Horizontal line at genome-wide significant (*p*=5e-8)

Published GWAS through 01/2018



NHGRI-EBI Catalog, http://www.ebi.ac.uk/gwas/

- ● Abdominal aortic aneurysm
- ● Acute lymphoblastic leukemia
- ● Adhesion molecules
- ● Adverse response to carbamapezine
- ● Adiponectin levels
- ● Age-related macular degeneration
- ● AIDS progression
- ● Alcohol dependence
- ● Alopecia areata
- ● Alzheimer disease
- ● Amyloid A levels
- ● Amyotrophic lateral sclerosis
- ● Angiotensin-converting enzyme activity
- ● Ankylosing spondylitis
- ● Arterial stiffness
- ● Asparagus anosmia
- ● Asthma
- ● Atherosclerosis in HIV
- ● Atrial fibrillation
- ● Attention deficit hyperactivity disorder
- ● Autism
- ● Basal cell cancer
- ● Behcet's disease
- ● Bipolar disorder
- ● Biliary atresia
- ● Bilirubin
- ● Bitter taste response
- ● Birth weight
- ● Bladder cancer
- ● Bleomycin sensitivity
- ● Blond or brown hair
- ● Blood pressure
- ● Blue or green eyes
- ● BMI, waist circumference
- ● Bone density
- ● Breast cancer
- ● C-reactive protein
- ● Calcium levels
- ● Cardiac structure/function
- ● Carnitine levels
- ● Carotenoid/tocopherol levels
- ● Celiac disease
- ● Cerebral atrophy measures
- ● Chronic lymphocytic leukemia
- ● Cleft lip/palate
- ● Cognitive function
- ● Conduct disorder
- ● Colorectal cancer
- ● Corneal thickness
- ● Coronary disease
- ● Creutzfeldt-Jakob disease
- ● Crohn's disease
- ● Cutaneous nevi
- ● Dermatitis
- ● Drug-induced liver injury
- ● Endometriosis
- ● Eosinophil count
- ● Eosinophilic esophagitis
- ● Erectile dysfunction and prostate cancer treatment
- ● Erythrocyte parameters
- ● Esophageal cancer
- ● Essential tremor
- ● Exfoliation glaucoma
- ● Eye color traits
- ● F cell distribution
- ● Fibrinogen levels
- ● Folate pathway vitamins
- ● Follicular lymphoma
- ● Fuch's corneal dystrophy
- ● Freckles and burning
- ● Gallstones
- ● Gastric cancer
- ● Glioma
- ● Glycemic traits
- ● Hair color
- ● Hair morphology
- ● Handedness in dyslexia
- ● HDL cholesterol
- ● Heart failure
- ● Heart rate
- ● Height
- ● Hemostasis parameters
- ● Hepatic steatosis
- ● Hepatitis
- ● Hepatocellular carcinoma
- ● Hirschsprung's disease
- ● HIV-1 control
- ● Hodgkin's lymphoma
- ● Homocysteine levels
- ● Hypospadias
- ● Idiopathic pulmonary fibrosis
- ● IgA levels
- ● IgE levels
- ● Inflammatory bowel disease
- ● Intracranial aneurysm
- ● Iris color
- ● Iron status markers
- ● Ischemic stroke
- ● Juvenile idiopathic arthritis
- ● Keloid
- ● Kidney stones
- ● LDL cholesterol
- ● Leprosy
- ● Leptin receptor levels
- ● Liver enzymes
- ● Longevity
- ● LP (a) levels
- ● LpPLA(2) activity and mass
- ● Lung cancer
- ● Magnesium levels
- ● Major mood disorders
- ● Malaria
- ● Male pattern baldness
- ● Matrix metalloproteinase levels
- ● MCP-1
- ● Melanoma
- ● Menarche & menopause
- ● Meningococcal disease
- ● Metabolic syndrome
- ● Migraine
- ● Moyamoya disease
- ● Multiple sclerosis
- ● Myeloproliferative neoplasms
- ● N-glycan levels
- ● Narcolepsy
- ● Nasopharyngeal cancer
- ● Neuroblastoma
- ● Nicotine dependence
- ● Obesity
- ● Open angle glaucoma
- ● Open personality
- ● Optic disc parameters
- ● Osteoarthritis
- ● Osteoporosis
- ● Otosclerosis
- ● Other metabolic traits
- ● Ovarian cancer
- ● Pancreatic cancer
- ● Pain
- ● Paget's disease
- ● Panic disorder
- ● Parkinson's disease
- ● Periodontitis
- ● Peripheral arterial disease
- ● Phosphatidylcholine levels
- ● Phosphorus levels
- ● Photic sneeze
- ● Phytosterol levels
- ● Platelet count
- ● Polycystic ovary syndrome
- ● Primary biliary cirrhosis
- ● Primary sclerosing cholangitis
- ● PR interval
- ● Progranulin levels
- ● Prostate cancer
- ● Protein levels
- ● PSA levels
- ● Psoriasis
- ● Psoriatic arthritis
- ● Pulmonary funct. COPD
- ● QRS interval
- ● QT interval
- ● Quantitative traits
- ● Recombination rate
- ● Red vs.non-red hair
- ● Refractive error
- ● Renal cell carcinoma
- ● Renal function
- ● Response to antidepressants
- ● Response to antipsychotic therapy
- ● Response to hepatitis C treat
- ● Response to metaformin
- ● Response to statin therapy
- ● Restless legs syndrome
- ● Retinal vascular caliber
- ● Rheumatoid arthritis
- ● Ribavirin-induced anemia
- ● Schizophrenia
- ● Serum metabolites
- ● Skin pigmentation
- ● Smoking behavior
- ● Speech perception
- ● Sphingolipid levels
- ● Statin-induced myopathy
- ● Stroke
- ● Systemic lupus erythematosus
- ● Systemic sclerosis
- ● T-tau levels
- ● Tau AB1-42 levels
- ● Telomere length
- ● Testicular germ cell tumor
- ● Thyroid cancer
- ● Tooth development
- ● Total cholesterol
- ● Triglycerides
- ● Tuberculosis
- ● Type 1 diabetes
- ● Type 2 diabetes
- ● Ulcerative colitis
- ● Urate
- ● Venous thromboembolism
- ● Ventricular conduction
- ● Vertical cup-disc ratio
- ● Vitamin B12 levels
- ● Vitamin D insuffiency
- ● Vitiligo
- ● Warfarin dose
- ● Weight
- ● White cell count
- ● YKL-40 levels

GWAS RESEARCH

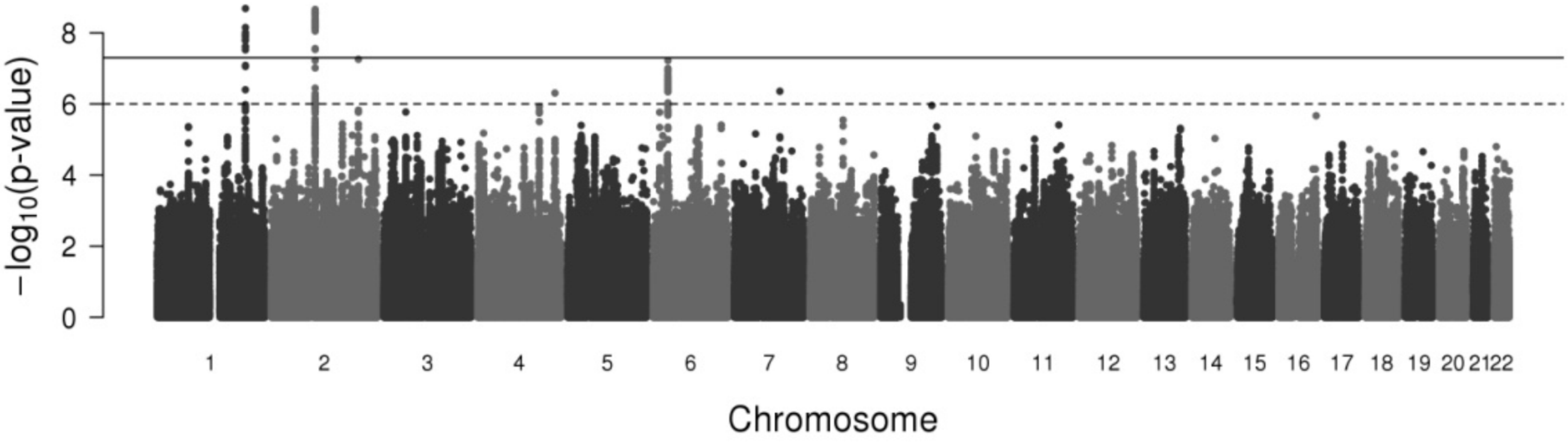
Social
Science
Genetic
Association
Consortium

EA1

- Discovery phase: 41 datasets with total sample size of $N \approx 100{,}000$.
  - Each cohort ran GWAS of *EduYears* (years of schooling)
  - One genome-wide significant association with *EduYears* and two with *College*.
- Replication phase: 12 independent datasets with total sample size of $N \approx 25{,}000$.

GWAS of 126,559 Individuals Identifies Genetic Variants Associated with Educational Attainment

Cornelius A. Rietveld *et al.*
Science **340**, 1467 (2013);
DOI: 10.1126/science.1235488


−log10(p-value)
0
2
4
6
8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

EA2

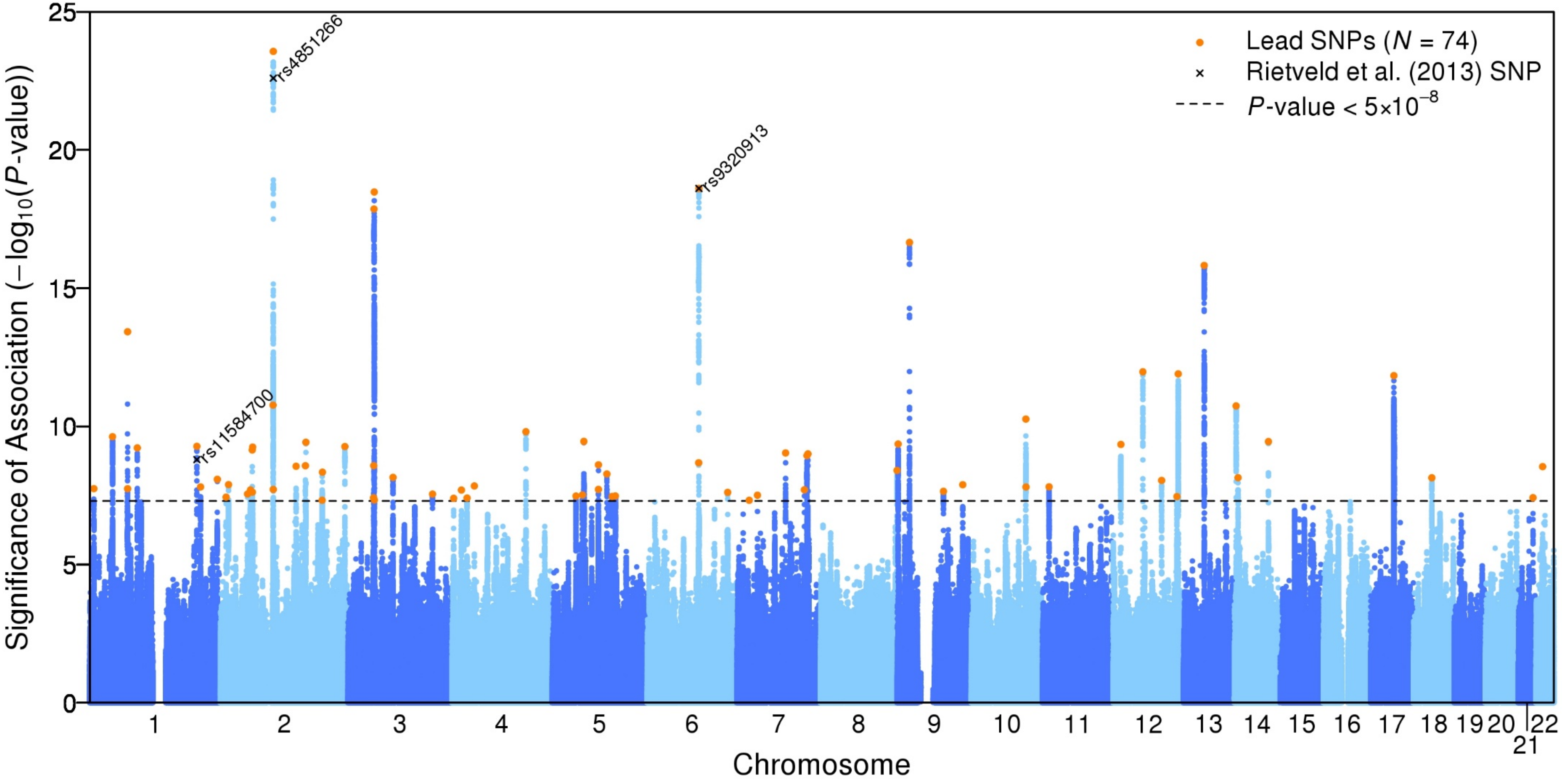
- 63 datasets with sample size of N = 293,723.
- Similar analysis plan as EA1, except focused exclusively on *EduYears* (not *College*).
- Found 74 genome-wide significant SNPs.
- After submission, first release of UK Biobank became available ($N \approx$ 110,000); used for replication.

LETTER

doi:10.1038/nature17671

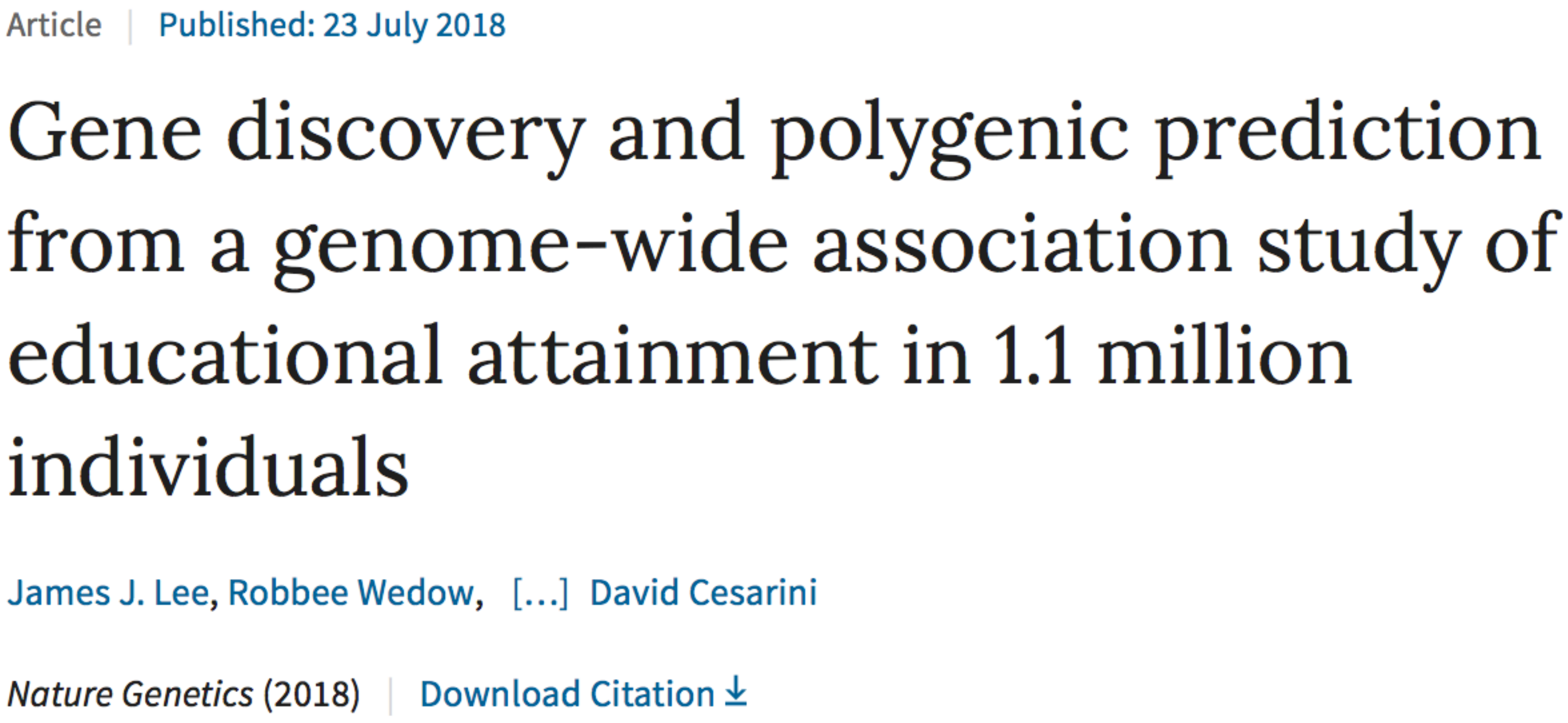
Genome-wide association study identifies 74 loci associated with educational attainment

A list of authors and their affiliations appears in the online version of the paper.


Significance of Association (−log10(P-value))
25
20
15
10
5
0
rs4851266
rs9320913
rs11584700
Lead SNPs (N = 74)
Rietveld et al. (2013) SNP
P-value < 5×10−8
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Chromosome

EA3

- HOT OFF THE PRESS!!!



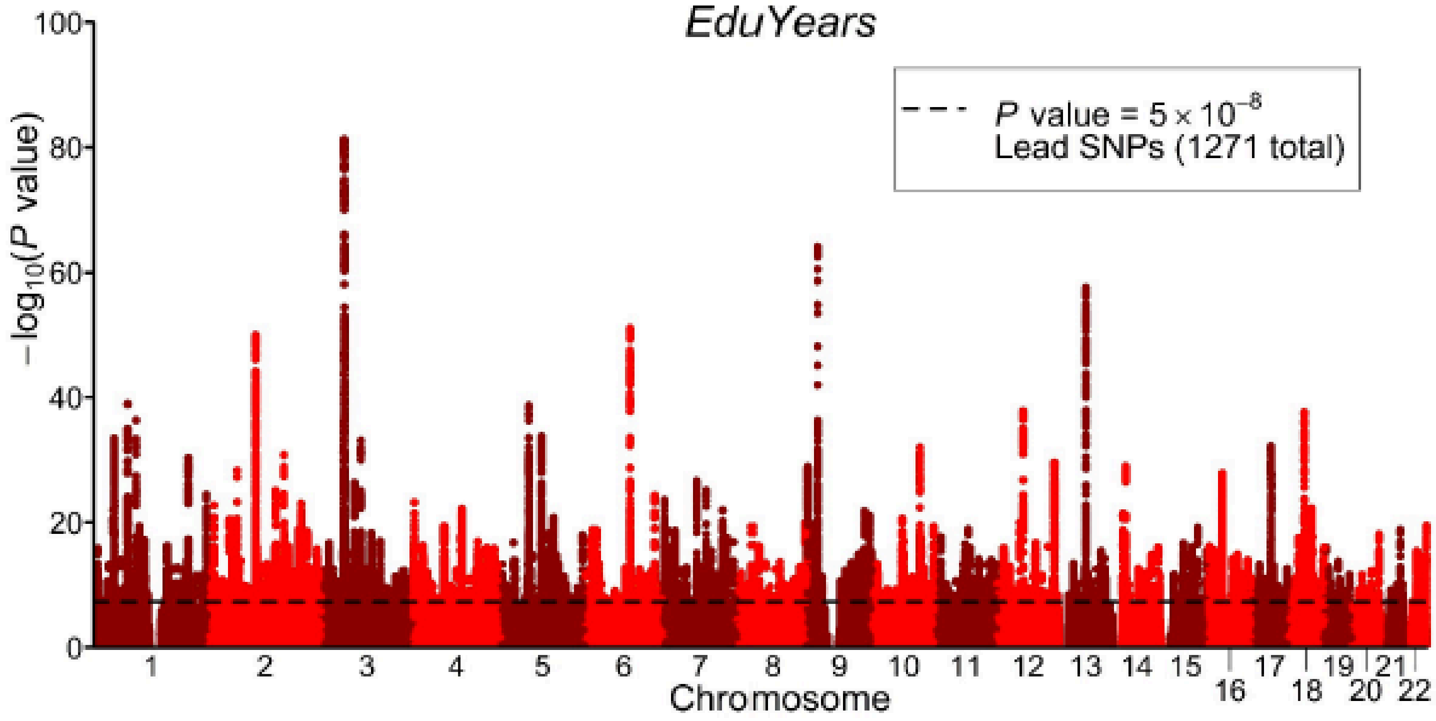
Article | Published: 23 July 2018

Gene discovery and polygenic prediction from a genome-wide association study of educational attainment in 1.1 million individuals

James J. Lee, Robbee Wedow, […] David Cesarini

Nature Genetics (2018) | Download Citation

70 cohorts (65 EA2 cohorts + 5 new cohorts), $N = 1,131,881$, 1,271 approximately independent ($r^2 < 0.1$) SNPs.



Adjusted for winners curse, the median effect size = 1.7 weeks of schooling per allele.

PLINK GENETIC DATA FORMAT

PLINK primer

- PLINK is one of the most common (and relatively universal) pieces of genetic analytics software
- Most commonly used for cleaning genetic data
- Everything you could need: https://www.cog-genomics.org/plink2

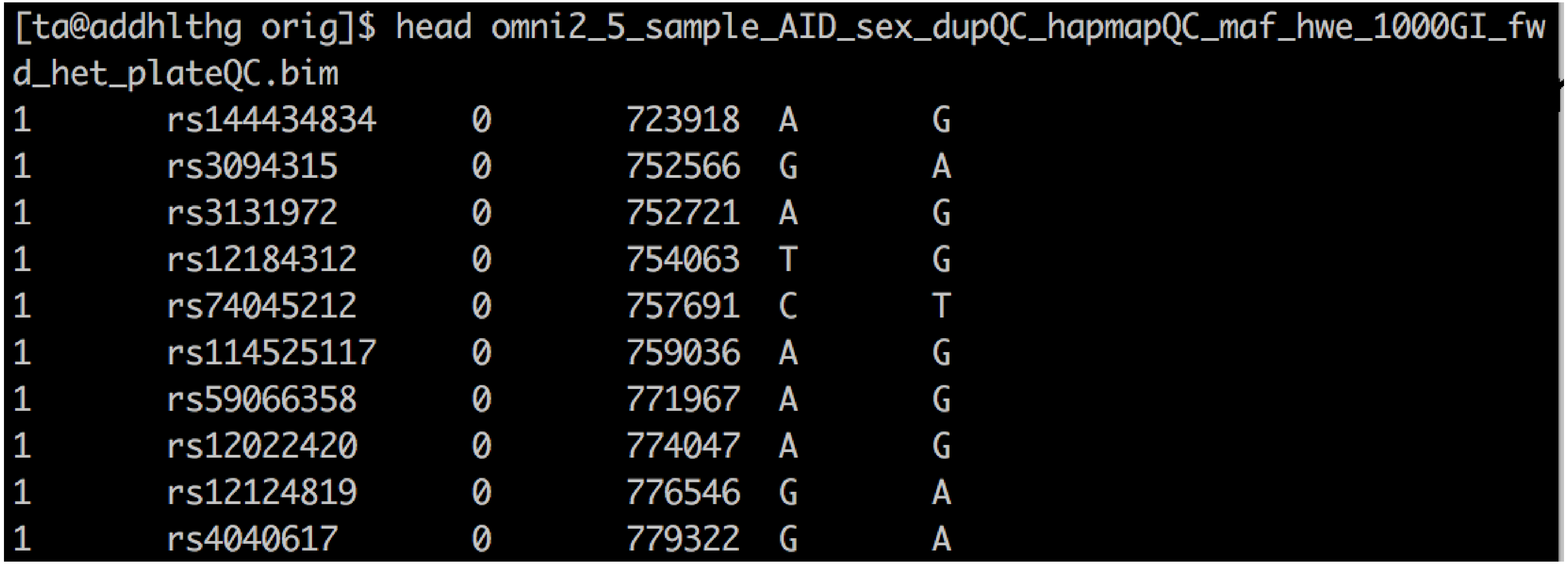
PLINK BINARY FORMAT: BIM/BED/FAM

- BIM/BED/FAM format is one of the most common formats genetic data are expressed in- this is the format Add Health data is stored in
- Three file types, each containing different information
  - BIM – information about genetic markers
  - BED – individual genetic data (compressed)
  - FAM – information about individuals (e.g., family identifiers, sex)

BIM/BED/FAM: BIM

- BIM, a text file with no header line, and one line per variant:
  - Chromosome code (usually an integer)
  - Variant identifier (usually rs number)
  - Allele 1 (usually minor allele)
  - Allele 2 (usually major allele)

BIM/BED/FAM: BIM



```
[ta@addhlthg orig]$ head omni2_5_sample_AID_sex_dupQC_hapmapQC_maf_hwe_1000GI_fw
d_het_plateQC.bim
1       rs144434834     0       723918  A       G
1       rs3094315       0       752566  G       A
1       rs3131972       0       752721  A       G
1       rs12184312      0       754063  T       G
1       rs74045212      0       757691  C       T
1       rs114525117     0       759036  A       G
1       rs59066358      0       771967  A       G
1       rs12022420      0       774047  A       G
1       rs12124819      0       776546  G       A
1       rs4040617       0       779322  G       A
```

BIM/BED/FAM: BED

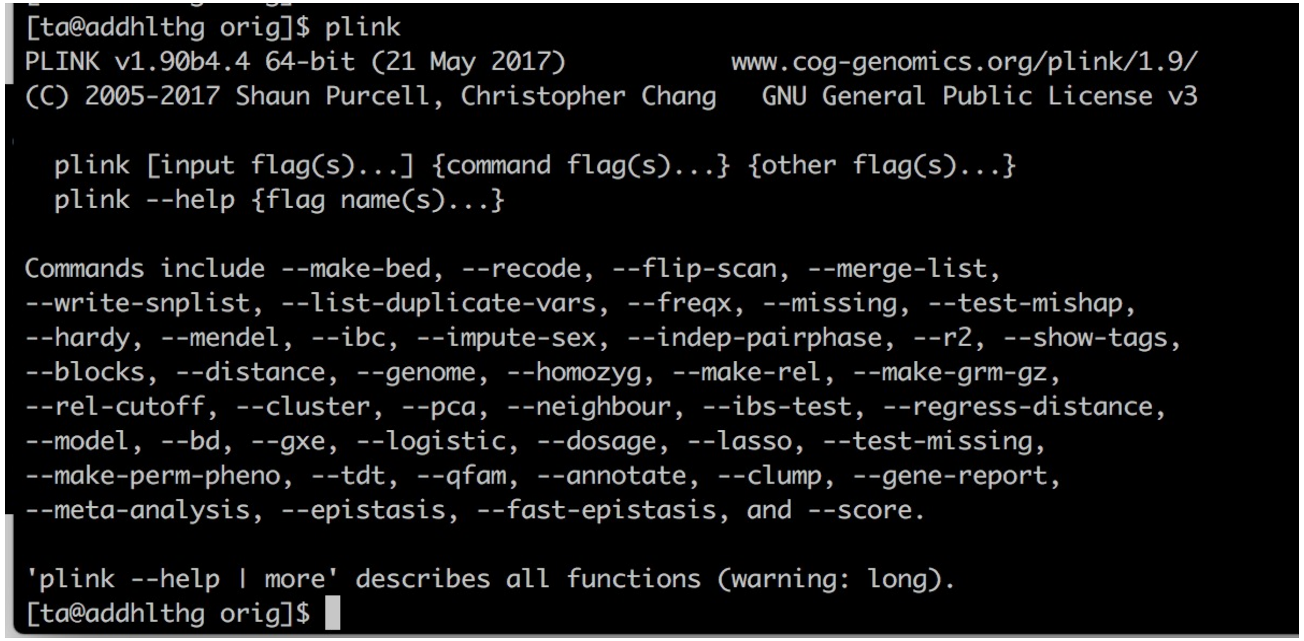
- BED, a condensed binary version of what’s called a PED file, that contains all genotype information for each person
- Don’t try to open or view it!

BIM/BED/FAM: FAM

- FAM, a text file with no header line, and one line per sample with the following six fields:
  - Family ID (‘FID’)
  - Within-family ID (‘IID’)
  - Within-family ID of father
  - Within-family ID of mother
  - Sex code (‘1’ = male, ‘2’ = female, ‘0’ = unknown)
  - Phenotype value (‘1’ = control, ‘2’ = case, ‘-9’ if missing or no phenotypes are present)

PLINK EXAMPLE

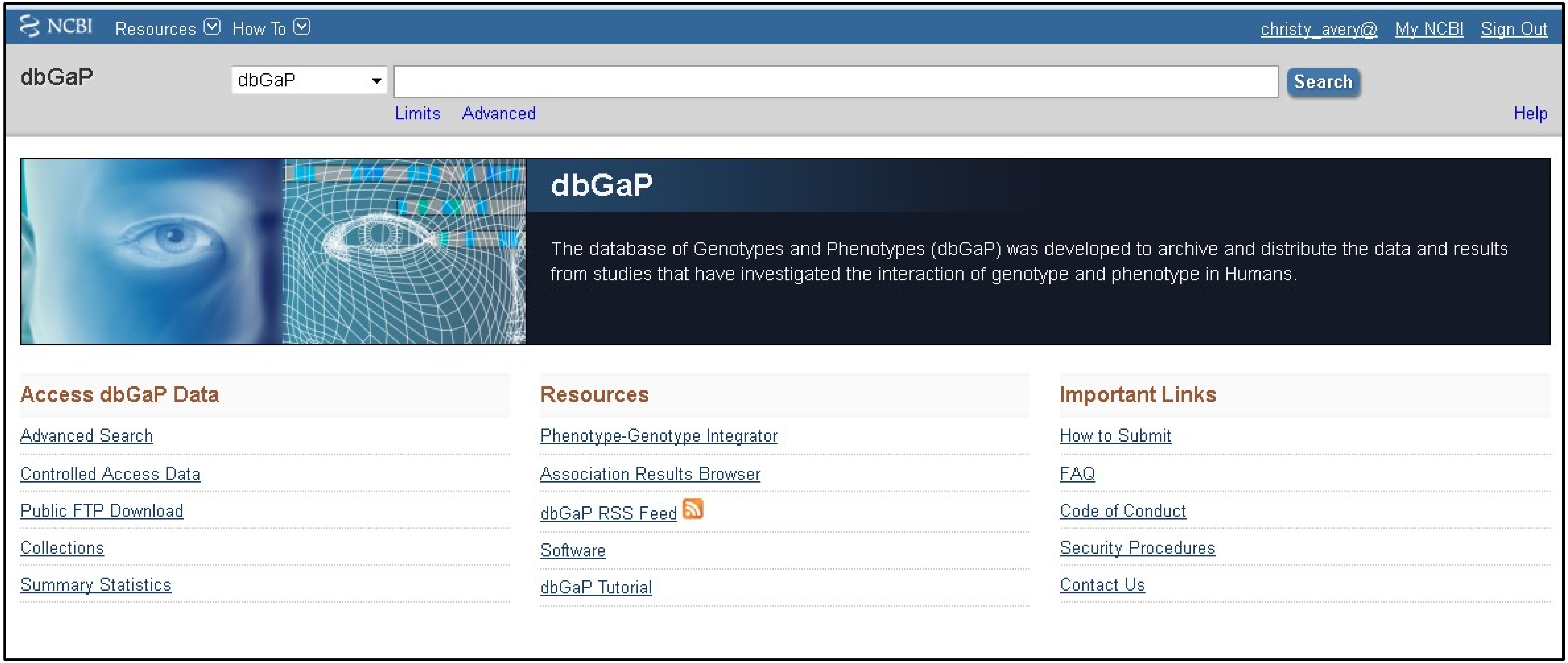
- PLINK will read in .bim/.bed/.fam and then allow analyses on these files
- With most software on a Linux server, calling the software is as simple as typing the name of the software
- In a Linux, simply type “plink”



```
[ta@addhlthg orig]$ plink
PLINK v1.90b4.4 64-bit (21 May 2017)          www.cog-genomics.org/plink/1.9/
(C) 2005-2017 Shaun Purcell, Christopher Chang   GNU General Public License v3

  plink [input flag(s)...] {command flag(s)...} {other flag(s)...}
  plink --help {flag name(s)...}

Commands include --make-bed, --recode, --flip-scan, --merge-list,
--write-snplist, --list-duplicate-vars, --freqx, --missing, --test-mishap,
--hardy, --mendel, --ibc, --impute-sex, --indep-pairphase, --r2, --show-tags,
--blocks, --distance, --genome, --homozyg, --make-rel, --make-grm-gz,
--rel-cutoff, --cluster, --pca, --neighbour, --ibs-test, --regress-distance,
--model, --bd, --gxe, --logistic, --dosage, --lasso, --test-missing,
--make-perm-pheno, --tdt, --qfam, --annotate, --clump, --gene-report,
--meta-analysis, --epistasis, --fast-epistasis, and --score.

'plink --help | more' describes all functions (warning: long).
[ta@addhlthg orig]$ █
```

OBTAINING ADD HEALTH GWAS DATA

dbGaP: Add Health Genotype Warehouse



dbGaP

The database of Genotypes and Phenotypes (dbGaP) was developed to archive and distribute the data and results from studies that have investigated the interaction of genotype and phenotype in Humans.

Access dbGaP Data

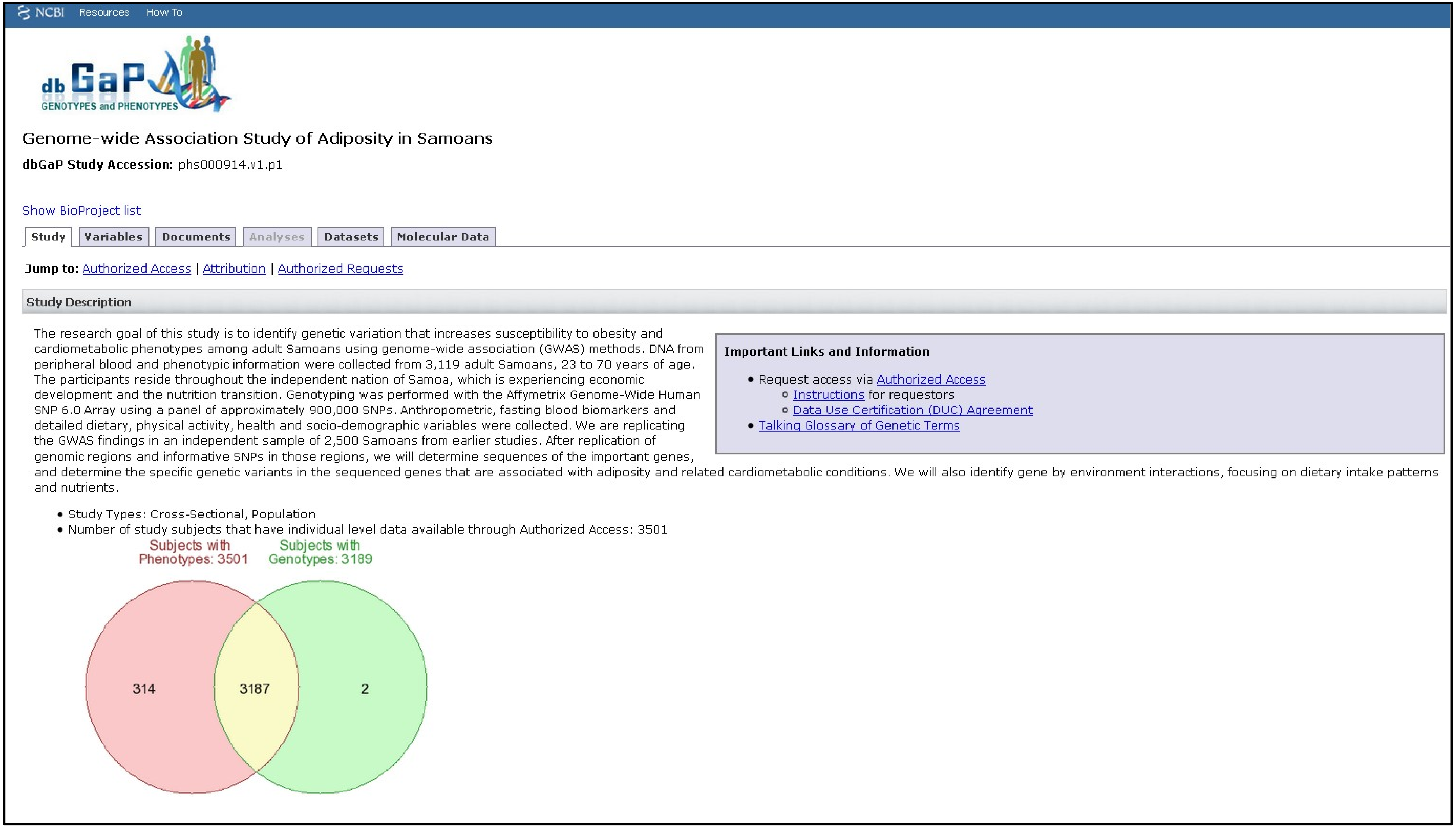
- Advanced Search
- Controlled Access Data
- Public FTP Download
- Collections
- Summary Statistics

Resources

- Phenotype-Genotype Integrator
- Association Results Browser
- dbGaP RSS Feed
- Software
- dbGaP Tutorial

Important Links

- How to Submit
- FAQ
- Code of Conduct
- Security Procedures
- Contact Us



Genome-wide Association Study of Adiposity in Samoans

dbGaP Study Accession: phs000914.v1.p1

Show BioProject list

Jump to: Authorized Access | Attribution | Authorized Requests

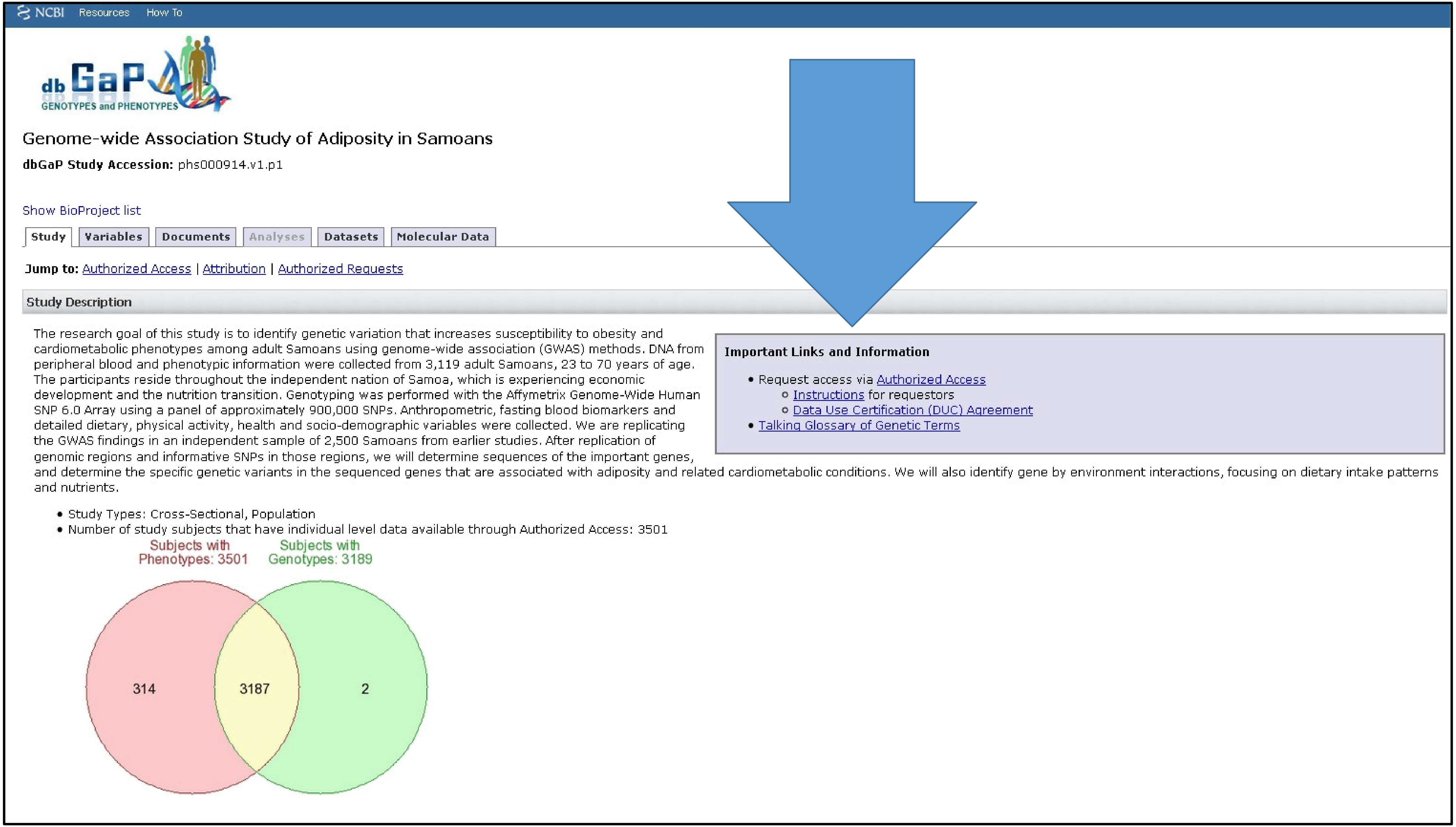
Study Description

The research goal of this study is to identify genetic variation that increases susceptibility to obesity and cardiometabolic phenotypes among adult Samoans using genome-wide association (GWAS) methods. DNA from peripheral blood and phenotypic information were collected from 3,119 adult Samoans, 23 to 70 years of age. The participants reside throughout the independent nation of Samoa, which is experiencing economic development and the nutrition transition. Genotyping was performed with the Affymetrix Genome-Wide Human SNP 6.0 Array using a panel of approximately 900,000 SNPs. Anthropometric, fasting blood biomarkers and detailed dietary, physical activity, health and socio-demographic variables were collected. We are replicating the GWAS findings in an independent sample of 2,500 Samoans from earlier studies. After replication of genomic regions and informative SNPs in those regions, we will determine sequences of the important genes, and determine the specific genetic variants in the sequenced genes that are associated with adiposity and related cardiometabolic conditions. We will also identify gene by environment interactions, focusing on dietary intake patterns and nutrients.

Important Links and Information

- Request access via Authorized Access
  - Instructions for requestors
  - Data Use Certification (DUC) Agreement
- Talking Glossary of Genetic Terms

- Study Types: Cross-Sectional, Population
- Number of study subjects that have individual level data available through Authorized Access: 3501



Genome-wide Association Study of Adiposity in Samoans

dbGaP Study Accession: phs000914.v1.p1

Show BioProject list

Jump to: Authorized Access | Attribution | Authorized Requests

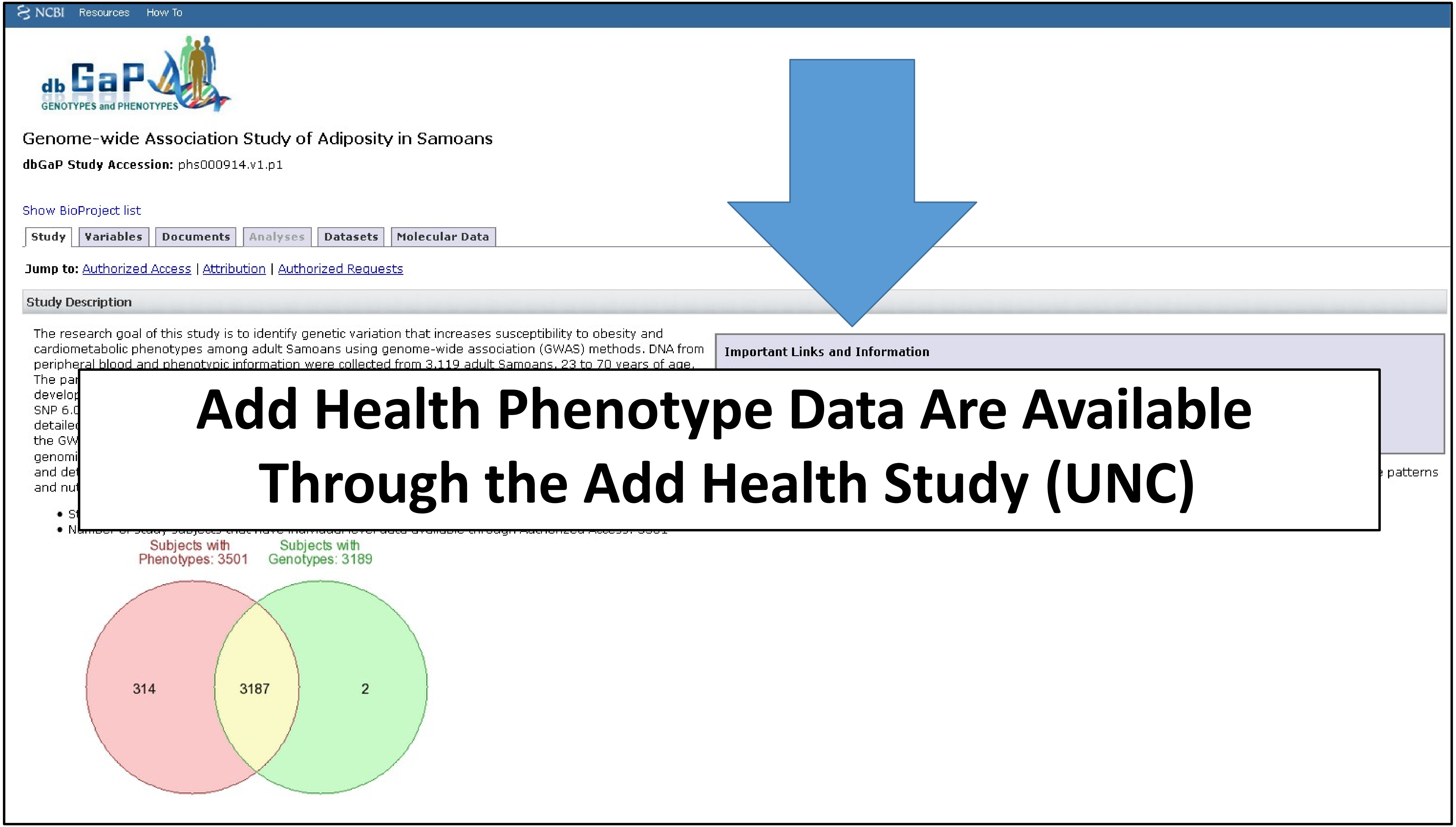
Study Description

The research goal of this study is to identify genetic variation that increases susceptibility to obesity and cardiometabolic phenotypes among adult Samoans using genome-wide association (GWAS) methods. DNA from peripheral blood and phenotypic information were collected from 3,119 adult Samoans, 23 to 70 years of age. The participants reside throughout the independent nation of Samoa, which is experiencing economic development and the nutrition transition. Genotyping was performed with the Affymetrix Genome-Wide Human SNP 6.0 Array using a panel of approximately 900,000 SNPs. Anthropometric, fasting blood biomarkers and detailed dietary, physical activity, health and socio-demographic variables were collected. We are replicating the GWAS findings in an independent sample of 2,500 Samoans from earlier studies. After replication of genomic regions and informative SNPs in those regions, we will determine sequences of the important genes, and determine the specific genetic variants in the sequenced genes that are associated with adiposity and related cardiometabolic conditions. We will also identify gene by environment interactions, focusing on dietary intake patterns and nutrients.

Important Links and Information

- Request access via Authorized Access
  - Instructions for requestors
  - Data Use Certification (DUC) Agreement
- Talking Glossary of Genetic Terms

- Study Types: Cross-Sectional, Population
- Number of study subjects that have individual level data available through Authorized Access: 3501



Genome-wide Association Study of Adiposity in Samoans

dbGaP Study Accession: phs000914.v1.p1

Show BioProject list

Jump to: Authorized Access | Attribution | Authorized Requests

Study Description

The research goal of this study is to identify genetic variation that increases susceptibility to obesity and cardiometabolic phenotypes among adult Samoans using genome-wide association (GWAS) methods. DNA from peripheral blood and phenotypic information were collected from 3,119 adult Samoans, 23 to 70 years of age.

The par
develop
SNP 6.0
detailed
the GW
genomi
and det
and nut

patterns

Important Links and Information

- St
- Number of study subjects that have individual level data available through Authorized Access: 3501

Add Health Phenotype Data Are Available Through the Add Health Study (UNC)

dbGaP Data Application

- Must be a tenure-track faculty or research scientist to apply
- Many datasets require IRB approval or IRB documentation of exempt status
- More than one dataset can be applied for on a single project, and datasets can be added later to an existing project
- Research use statement and non-technical project summary are required; non-technical summary will be publicly available online
- Will need to demonstrate how data will be securely stored and provide contact information for your institution’s IT director
- Will be reviewed by a committee; usually requires some revision in my experience
- About 1-2 months from start to finish for the entire process in my experience

dbGaP Research Statement

- 2,200 characters maximum
- Need to indicate:
    - Objectives of project
    - Study design
    - Analysis plan, including careful articulation of which outcomes (phenotypes) will be used
    - Explanation of how project is consistent with data use requirements of a particular dataset
    - Explanation of any planned collaboration with other researchers or institutions

OTHER CONSIDERATIONS

Team Science/Consortia

- Joining a consortium is often the first step in ***<u>GWAS</u>***

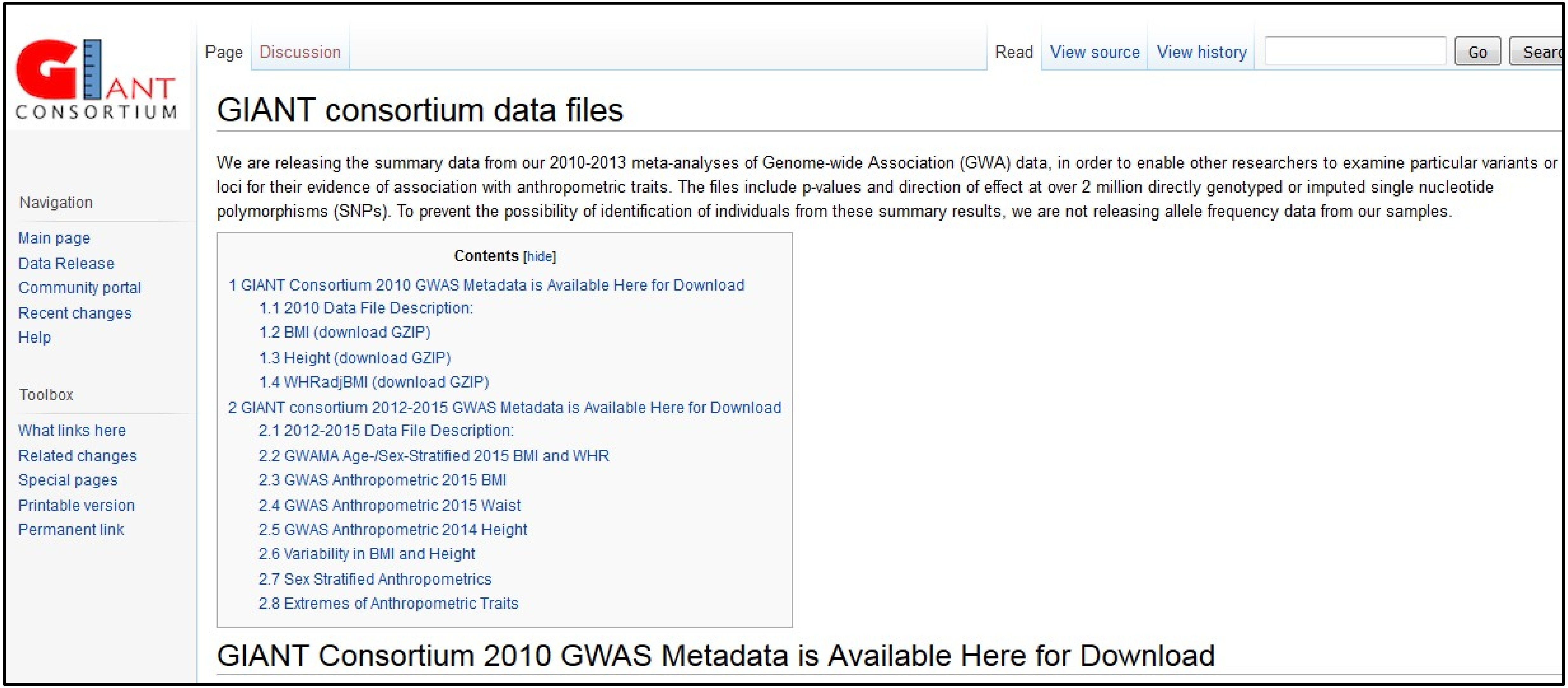
Statistical Power: Why we work in teams

We need to correct for 1,000,000 statistical tests when interrogating genome!

$\alpha = 0.05/1M$ or $5x10^{-8}$

Correction for only 1M tests given correlation in human genome

Summary Results from *Many Large Consortia* Are Available Online



GIANT consortium data files

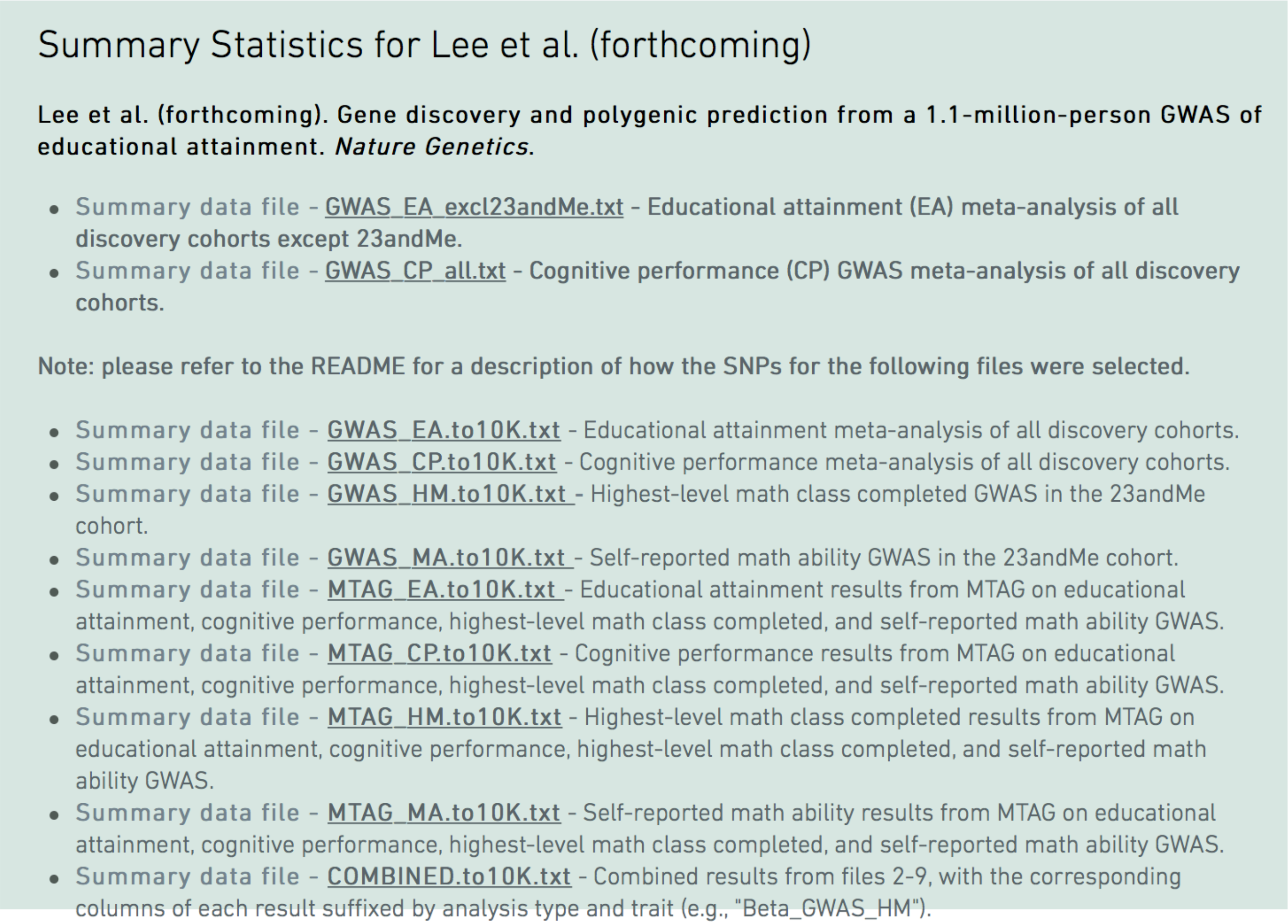
We are releasing the summary data from our 2010-2013 meta-analyses of Genome-wide Association (GWA) data, in order to enable other researchers to examine particular variants or loci for their evidence of association with anthropometric traits. The files include p-values and direction of effect at over 2 million directly genotyped or imputed single nucleotide polymorphisms (SNPs). To prevent the possibility of identification of individuals from these summary results, we are not releasing allele frequency data from our samples.

Contents [hide]

GIANT Consortium 2010 GWAS Metadata is Available Here for Download

Summary Results from ***Many Large Consortia*** Are Available Online

Social
Science
Genetic
Association
Consortium



Summary Statistics for Lee et al. (forthcoming)

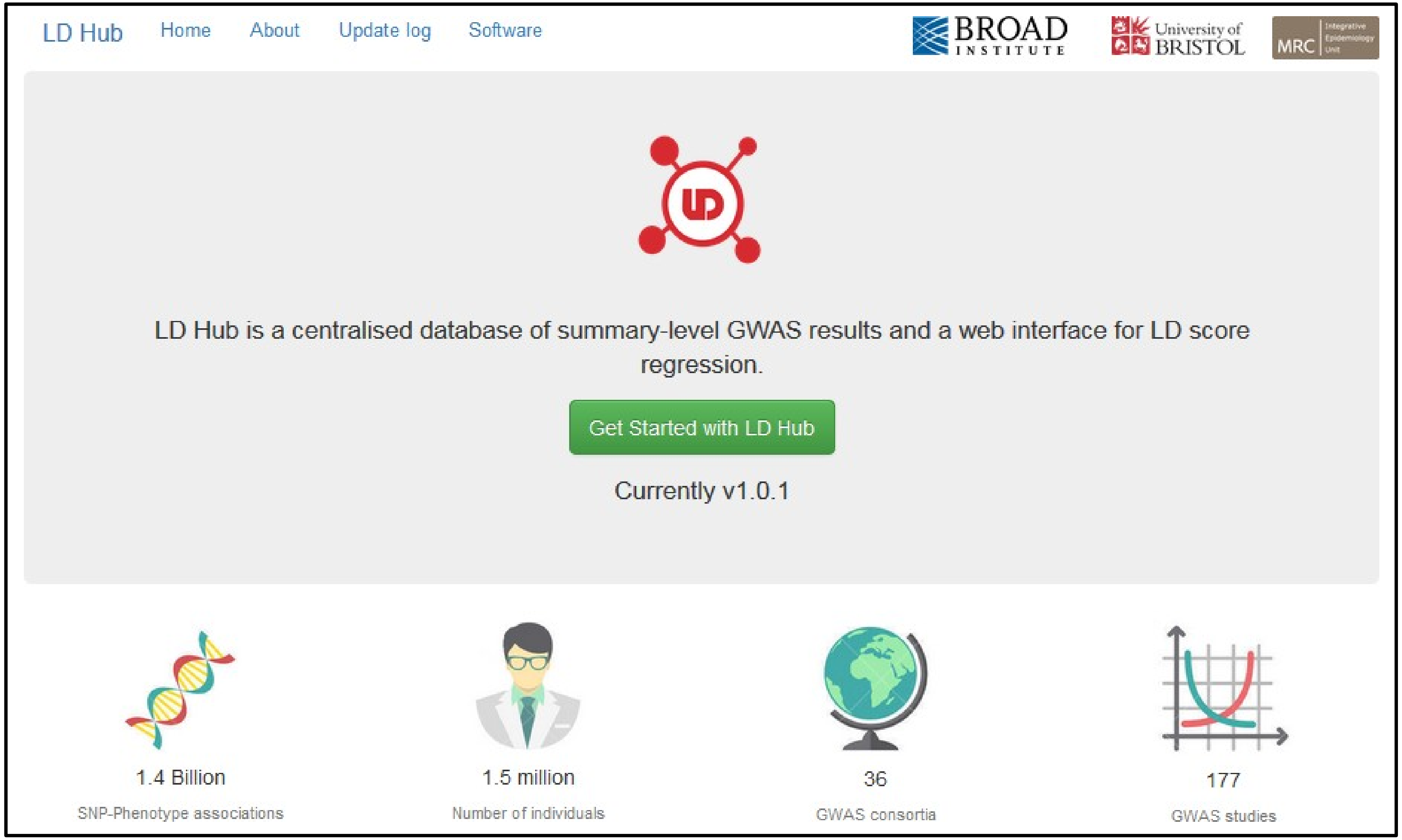
Lee et al. (forthcoming). Gene discovery and polygenic prediction from a 1.1-million-person GWAS of educational attainment. *Nature Genetics*.

- Summary data file - GWAS_EA_excl23andMe.txt - Educational attainment (EA) meta-analysis of all discovery cohorts except 23andMe.
- Summary data file - GWAS_CP_all.txt - Cognitive performance (CP) GWAS meta-analysis of all discovery cohorts.

Note: please refer to the README for a description of how the SNPs for the following files were selected.

- Summary data file - GWAS_EA.to10K.txt - Educational attainment meta-analysis of all discovery cohorts.
- Summary data file - GWAS_CP.to10K.txt - Cognitive performance meta-analysis of all discovery cohorts.
- Summary data file - GWAS_HM.to10K.txt - Highest-level math class completed GWAS in the 23andMe cohort.
- Summary data file - GWAS_MA.to10K.txt - Self-reported math ability GWAS in the 23andMe cohort.
- Summary data file - MTAG_EA.to10K.txt - Educational attainment results from MTAG on educational attainment, cognitive performance, highest-level math class completed, and self-reported math ability GWAS.
- Summary data file - MTAG_CP.to10K.txt - Cognitive performance results from MTAG on educational attainment, cognitive performance, highest-level math class completed, and self-reported math ability GWAS.
- Summary data file - MTAG_HM.to10K.txt - Highest-level math class completed results from MTAG on educational attainment, cognitive performance, highest-level math class completed, and self-reported math ability GWAS.
- Summary data file - MTAG_MA.to10K.txt - Self-reported math ability results from MTAG on educational attainment, cognitive performance, highest-level math class completed, and self-reported math ability GWAS.
- Summary data file - COMBINED.to10K.txt - Combined results from files 2-9, with the corresponding columns of each result suffixed by analysis type and trait (e.g., "Beta_GWAS_HM").

Summary Results from ***Many Large Consortia*** Are Available Online



LD Hub is a centralised database of summary-level GWAS results and a web interface for LD score regression.

Currently v1.0.1

1.4 Billion
SNP-Phenotype associations

1.5 million
Number of individuals

36
GWAS consortia

177
GWAS studies

Race/Ethnicity Heterogeneity

Why are genomic studies only in Europeans?

Medical genomics has focused almost entirely on those of European descent. Other ethnic groups must be studied to ensure that more people benefit, say **Carlos D. Bustamante, Esteban González Burchard** and **Francisco M. De La Vega.**

In the past decade, researchers have dramatically improved our understanding of the genetic basis of complex chronic diseases, such as Alzheimer's disease and type 2 diabetes, through more than 1,000 genome-wide association studies (GWAS). These scan the genomes of thousands of people for known genetic variants, to find out which are associated with a particular condition.

Yet the findings from such studies are likely to have less relevance than was previously thought for the world's population as a whole. Ninety-six per cent of subjects included in the GWAS conducted so far are people of European descent[1] (see 'Sampling bias'). And a recent *Nature* survey suggests that this bias is likely to persist in the upcoming efforts to sequence people's entire genomes[2].

Geneticists worldwide must investigate a much broader ensemble of populations, including racial and ethnic minorities. If we do not, a biased picture will emerge of which variants are important, and genomic medicine will largely benefit a privileged few. ▶

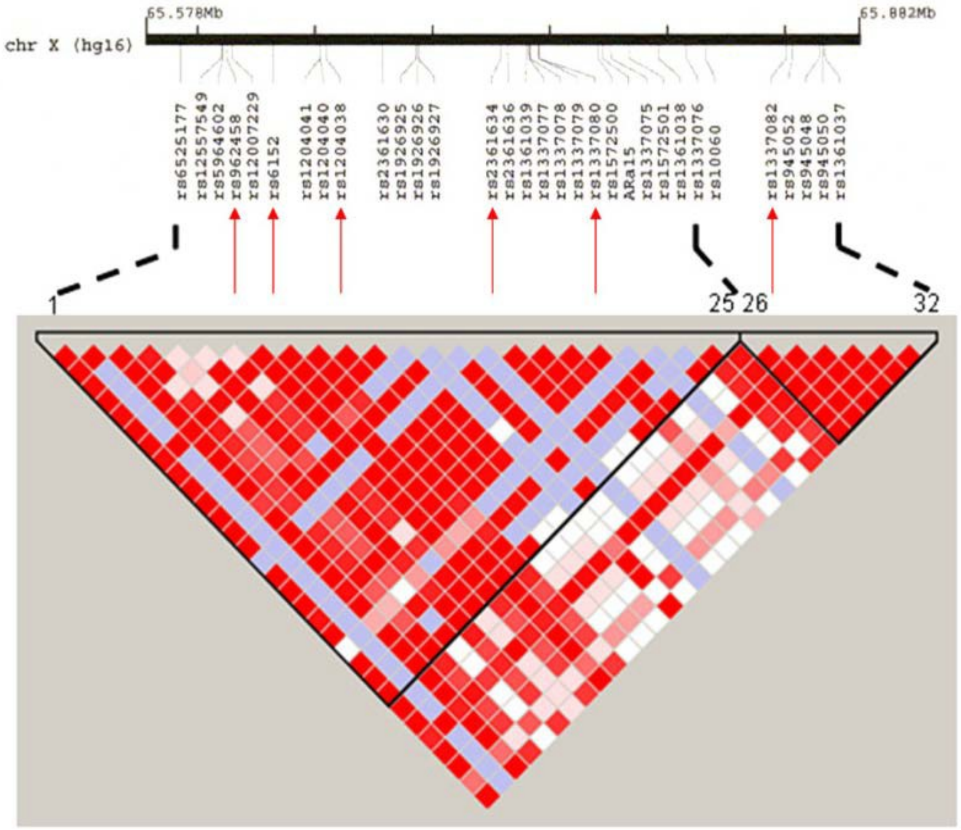
SUMMARY

- Those most in need must not be the last to benefit from genetic research
- Reviewers and granting bodies must demand racial and ethnic diversity in genome studies
- Global genomics needs the financial support of governments and non-profits

PMID:21753830

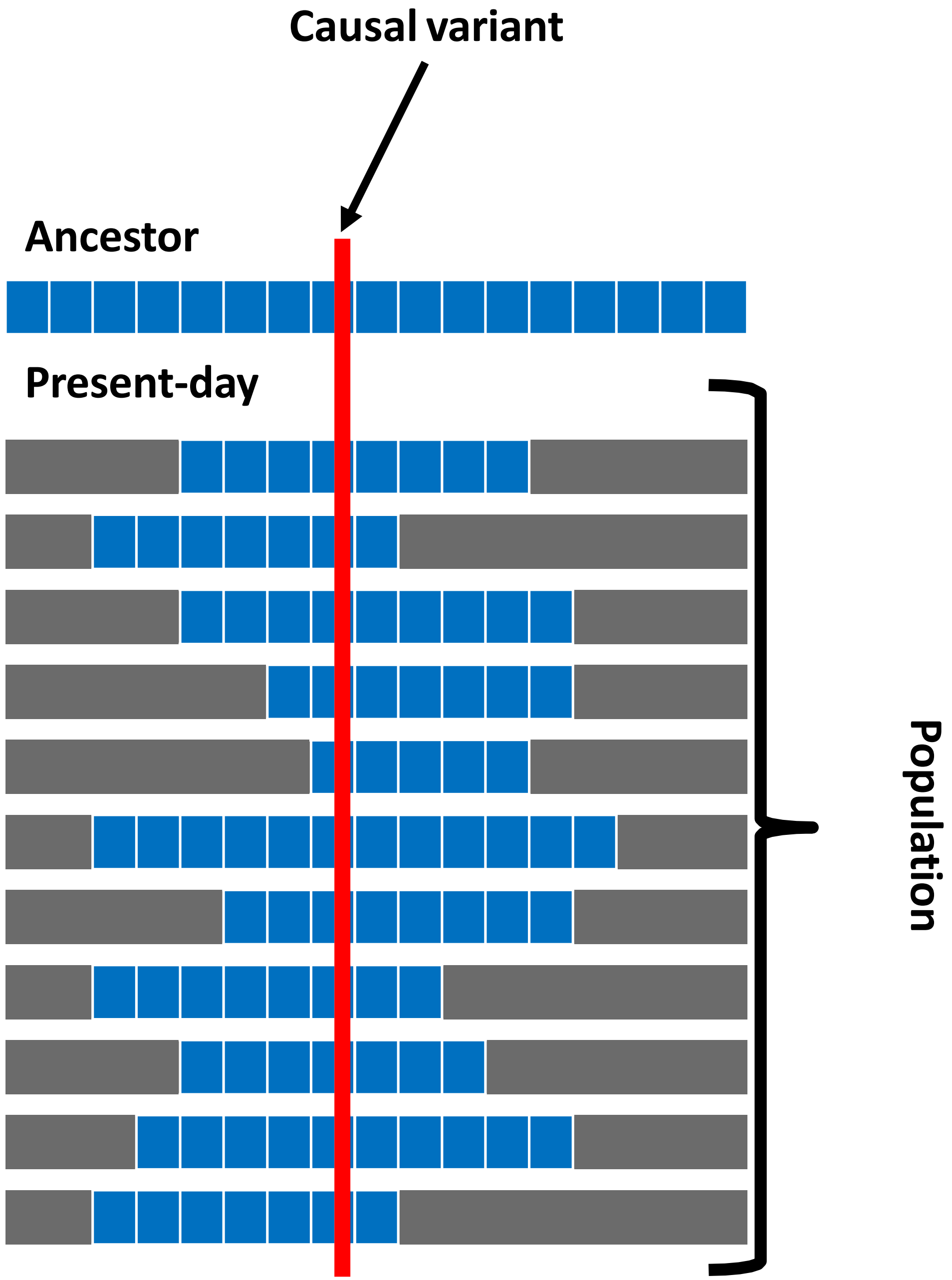
Linkage Disequilibrium (LD)

- SNPs in the genome that are closer to each other are inherited together
- SNPs are inherited in blocks
- Therefore, there is a “non-random” assortment of alleles in a given population



Linkage Disequilibrium (LD)

- LD patterns are population-specific
- Because of the population-specific pattern of LD, confounding by LD is expected to vary across populations (population stratification) and thus genetic analyses must be population specific



Race/Ethnicity Heterogeneity

Take-home messages:

1 – Genes generalize, but variation in SNPs exist.

2 – Studies in non-European populations are needed.

Challenges to analyzing GWAS data

- Many tools are available for analyzing GWAS data- for running GWAS, making polygenic scores, cleaning genetic data, etc.
- Implementation may be challenging if modest Unix/R/python expertise
- Storage: Easiest solution if inexperienced to storing genetic data is to get in touch with research computing at your university or institution

THANK YOU!

rwedow@alumni.nd.edu