UK Biobank data to fuel drug discovery

Jonathan Marchini
Executive Director, Head of Statistical Genetics and Machine Learning

REGENERON
SCIENCE TO MEDICINE®

Regeneron Genetics Center (RGC)

Established in 2014 and is now one of the largest operational human sequencing efforts

SAMPLE BIOBANKING

LIBRARY PREPARATION AND EXOME CAPTURE

ILLUMINA-BASED SEQUENCING

CLOUD BASED INFORMATICS & ANALYSIS

Mission:

Taking large scale human genetics to the next level for target discovery, support existing targets and identify novel indications

Regeneron Genetics Center:
Unprecedented Speed, Scale & Integration

~500,000 additional exomes to be sequenced annually

~2M exomes sequenced to date

120+ research collaborations

RGC
Regeneron Genetics Center

NUMEROUS existing targets & development programs validated

MULTIPLE potential new drug targets identified

>95% of genes with identified LOF carrier(s)

All accomplished in just the first 9 years!

RGC has the most diverse collection and catalogue of human coding variation to date

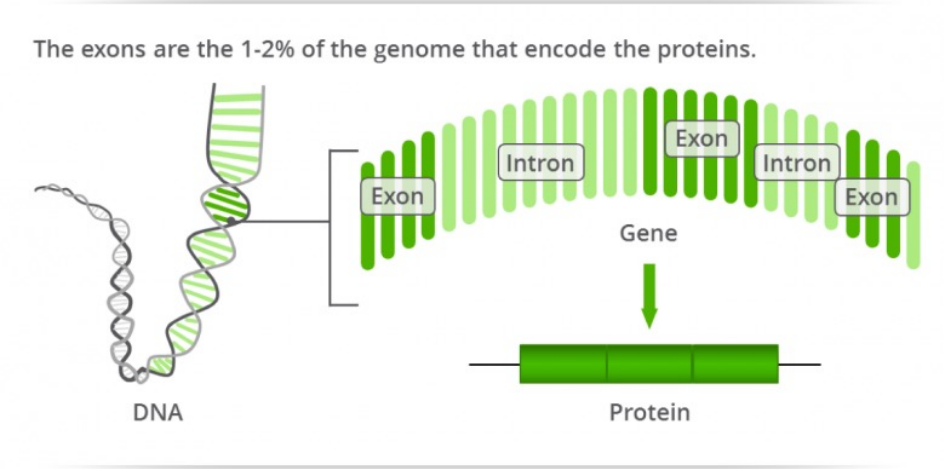
The New York Times

Hospital and Drugmaker Move to Build Vast Database of New Yorkers' DNA

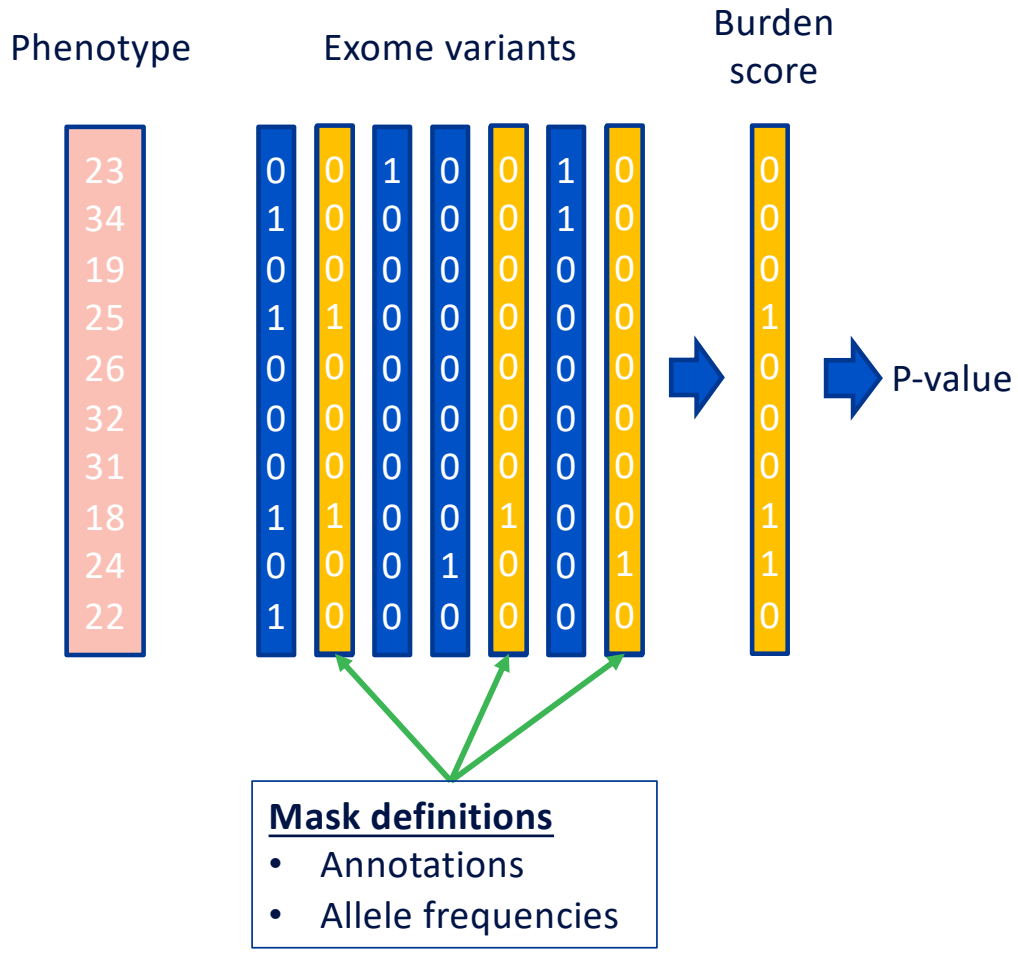
Patients will be asked if their genetic sequence can be added to a database — shared with a pharmaceutical company — in a quest to cure a multitude of diseases.

Wilbert Gibson is a Mount Sinai patient who agreed to let the hospital system use his genetic information in research for treatment of a variety of diseases. Hiroko Masuike/The New York Times

Exome sequencing of all 500,000 UK Biobank participants



- Regeneron led consortium of commercial companies.
- All 500,000 sequences made available in mid-2022.



Science

Current Issue | First release papers | Archive | About | Submit manuscript

HOME > SCIENCE > SEQUENCING OF 640,000 EXOMES IDENTIFIES GPR75 VARIANTS ASSOCIATED WITH PROTECTION FROM OBESITY

Science

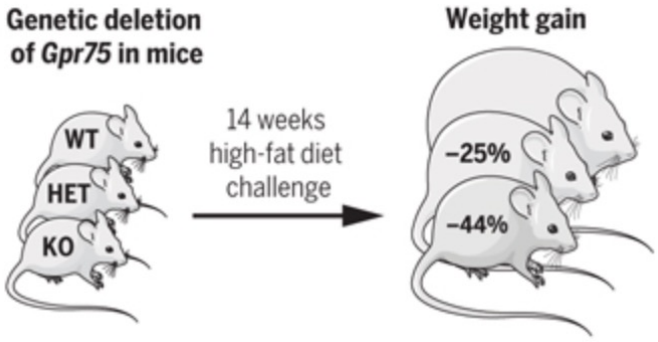
RESEARCH ARTICLE

Sequencing of 640,000 exomes identifies *GPR75* variants associated with protection from obesity

PARSA AKBARI, ANKIT GILANI, OLUKAYODE SOSINA, JACK A. KOSMICKI, LORI KHRIMIAN, YI-YA FANG, TRIKALDARSHI PERSAUD, VICTOR GARCIA, DYLAN SUN, ALEXANDER LI, JOELLE MBATCHOU, ADAM E. LOCKE, CHRISTIAN BENNER, NIEK VERWEIJ, NAN LIN, SAKIB HOSSAIN, KEVIN AGOSTINUCCI, JONATHAN V. PASCALE, ERCUMENT DIRICE, MICHAEL DUNN, REGENERON GENETICS CENTER‡, DISCOVEHR COLLABORATION‡, WILLIAM E. KRAUS, SVATI H. SHAH, YII-DER I. CHEN, JEROME I. ROTTER, DANIEL J. RADER, OLLE MELANDER, CHRISTOPHER D. STILL, TOORAJ MIRSHAHI, DAVID J. CAREY, JAIME BERUMEN-CAMPOS, PABLO KURI-MORALES, JESUS ALEGRE-DÍAZ, JASON M. TORRES, JONATHAN R. EMBERSON, RORY COLLINS, SUGANTHI BALASUBRAMANIAN, ALICIA HAWES, MARCUS JONES, BRIAN ZAMBROWICZ, ANDREW J. MURPHY, CHARLES PAULDING, GIOVANNI COPPOLA, JOHN D. OVERTON, JEFFREY G. REID, ALAN R. SHULDINER, MICHAEL CANTOR, HYUN M. KANG, GONCALO R. ABECASIS, KATIA KARALIS, ARIS N. ECONOMIDES, JONATHAN MARCHINI, GEORGE D. YANCOPOULOS, MARK W. SLEEMAN, JUDITH ALTAREJOS, GIUSY DELLA GATTA, ROBERTO TAPIA-CONYER, MICHAL L. SCHWARTZMAN, ARIS BARAS, MANUEL A. R. FERREIRA, AND LUCA A. LOTTA

fewer | Authors Info & Affiliations

- Rare predicted loss of function coding variants in *GPR75* for heterozygous carriers found to be associated with
  - Lower BMI (-1.8 kg/m^2)
  - Lower body weight (~5.3 kg or 11.6 lbs lower)
  - Protection against obesity (54% lower odds)



B

Study	Beta (95% CI) in SDs of BMI	Genotypes (RR|RA|AA)	p-value	Beta (95% CI) in kg/m^2 of BMI	Beta (95% CI) in kg of body weight
Discovery Cohorts					
MCPS	-0.48 (-0.82, -0.13)	95,816|30|0	7.1E-03	-2.6 (-4.4, -0.7)	-7.4 (-12.8, -2)
GHS_EUR	-0.27 (-0.52, -0.02)	121,010|51|0	3.6E-02	-1.4 (-2.8, -0.1)	-4.2 (-8, -0.3)
UKB_EUR	-0.34 (-0.49, -0.19)	428,572|147|0	6.6E-06	-1.8 (-2.6, -1)	-5.3 (-7.6, -3)
Subgroup meta-analysis results	-0.34 (-0.46, -0.22)	645,398|228|0	2.6E-08	-1.8 (-2.5, -1.2)	-5.3 (-7.1, -3.4)

RGC

REGENERON

Diverse RGC cohorts: Novel genetic discoveries and therapeutic targets

nature communications

Article

https://doi.org/10.1038/s41467-022-32398-7

Multiancestry exome sequencing reveals *INHBE* mutations associated with favorable fat distribution and protection from diabetes

Received: 8 February 2022

Accepted: 28 July 2022

Published online: 23 August 2022

Check for updates

Parsa Akbari[1,13], Olukayode A. Sosina[1,13], Jonas Bovijn[1,13], Karl Landheer[2], Jonas B. Nielsen[1], Minhee Kim[1], Senem Aykul[1], Tanima De[1], Mary E. Haas[1], George Hindy[1], Nan Lin[1], Ian R. Dinsmore[3], Jonathan Z. Luo[3], Stefanie Hectors[2], Benjamin Geraghty[1], Mary Germino[2], Lampros Panagis[2], Prodromos Parasoglou[2], Johnathon R. Walls[2], Gabor Halasz[2], Gurinder S. Atwal[2], Regeneron Genetics Center*, DiscovEHR Collaboration*, Marcus Jones[1], Michelle G. LeBlanc[1], Christopher D. Still[4], David J. Carey[4], Alice Giontella[5,6], Marju Orho-Melander[5], Jaime Berumen[7], Pablo Kuri-Morales[7,8], Jesus Alegre-Díaz[7], Jason M. Torres[9,10], Jonathan R. Emberson[9,10], Rory Collins[10], Daniel J. Rader[11], Brian Zambrowicz[2], Andrew J. Murphy[2], Suganthi Balasubramanian[1], John D. Overton[1], Jeffrey G. Reid[1], Alan R. Shuldiner[1], Michael Cantor[1], Goncalo R. Abecasis[1], Manuel A. R. Ferreira[1], Mark W. Sleeman[2], Viktoria Gusarova[2], Judith Altarejos[2], Charles Harris[2], Aris N. Economides[1,2], Vincent Idone[2], Katia Karalis[1], Giusy Della Gatta[1], Tooraj Mirshahi[4], George D. Yancopoulos[2], Olle Melander[5,12], Jonathan Marchini[1], Roberto Tapia-Conyer[8,13], Adam E. Locke[1,13], Aris Baras[1,13] ✉, Niek Verweij[1,13] & Luca A. Lotta[1,13] ✉

Akbari et al. (2022) Nat Commun 13, 4844

Association with favorable fat distribution ($p = 1.8 \times 10^{-09}$), favorable metabolic profile and protection from type 2 diabetes (~28% lower odds; $p = 0.004$) for heterozygous protein-truncating mutations in INHBE

The NEW ENGLAND JOURNAL of MEDICINE

ORIGINAL ARTICLE

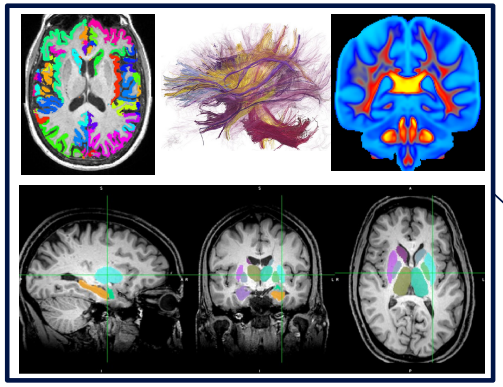
Germline Mutations in *CIDEB* and Protection against Liver Disease

N. Verweij, M.E. Haas, J.B. Nielsen, O.A. Sosina, M. Kim, P. Akbari, T. De, G. Hindy, J. Bovijn, T. Persaud, L. Miloscio, M. Germino, L. Panagis, K. Watanabe, J. Mbatchou, M. Jones, M. LeBlanc, S. Balasubramanian, C. Lammert, S. Enhörning, O. Melander, D.J. Carey, C.D. Still, T. Mirshahi, D.J. Rader, P. Parasoglou, J.R. Walls, J.D. Overton, J.G. Reid, A. Economides, M.N. Cantor, B. Zambrowicz, A.J. Murphy, G.R. Abecasis, M.A.R. Ferreira, E. Smagris, V. Gusarova, M. Sleeman, G.D. Yancopoulos, J. Marchini, H.M. Kang, K. Karalis, A.R. Shuldiner, G. Della Gatta, A.E. Locke, A. Baras, and L.A. Lotta

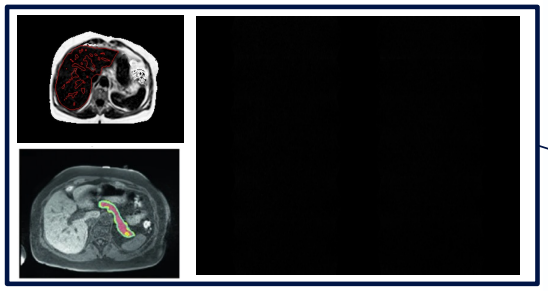
Verweij et al. (2022) N Engl J Med 387:332-344

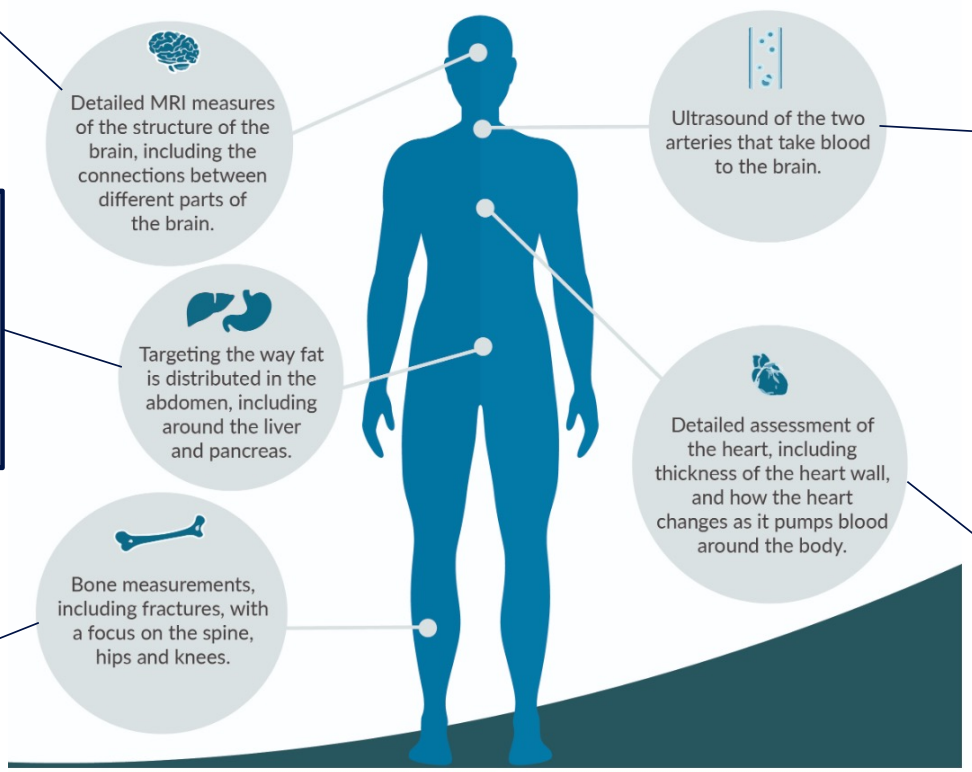
Rare predicted loss-of-function variants plus missense variants in CIDEB associated with 33% lower odds of liver disease of any cause

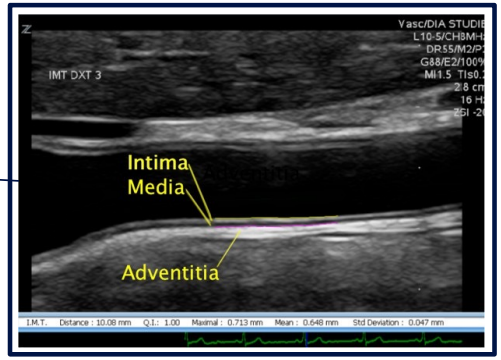
The UK Biobank Imaging Study

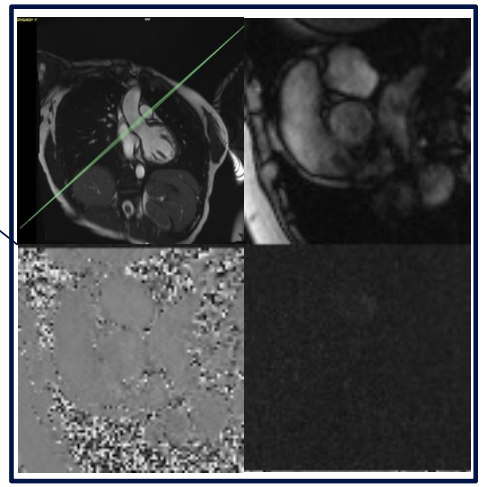


- Aims to collect brain, heart, and abdomen scans from 100,000 participants.
- Repeat set of imaging on 60,000 participants.
- Raw images available, but derived phenotypes appear more slowly





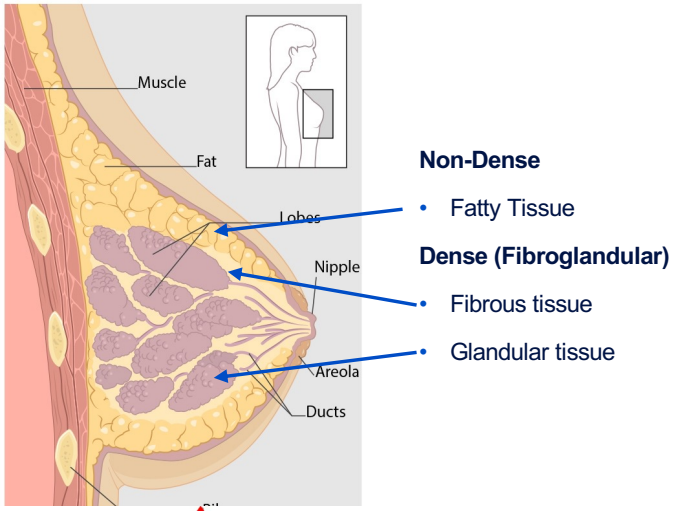




Can UKB whole body MRI be used to assess breast density?

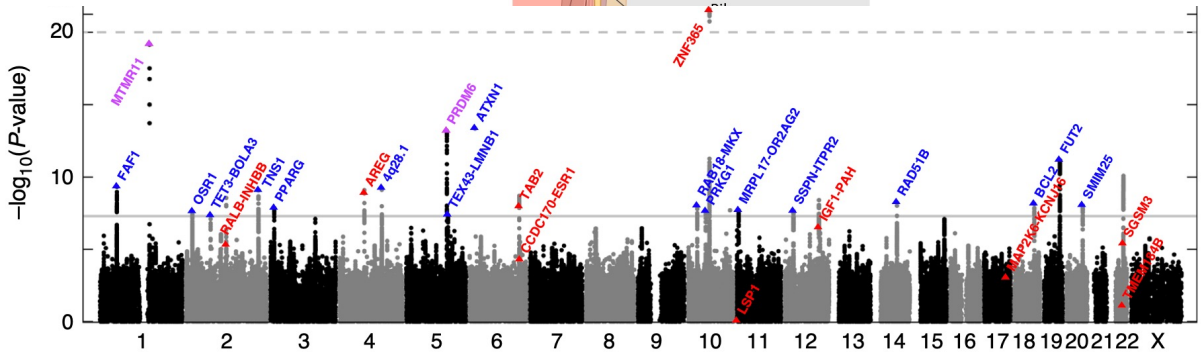
Breast density is a risk factor for breast cancer, independent of other known risk factors [1]

- Women with density > 75% have an increased risk (range across 4 studies: 2.82-5.99) compared to women with < 10% density [2]
- Approximately 1/3 of breast cancer risk may be attributable to density (2 studies, attributable risk percent = 28% and 33%) [2]
- Breast density is modifiable—tamoxifen-induced reduction of density decreases the risk of subsequent breast cancer [3]



Breast density has a strong genetic component

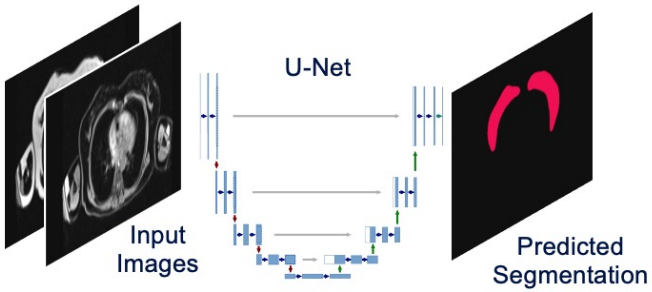
2 GWAS (BCAC and Kaiser) have identified 46 independent genome-wide significant breast density loci for three measures of breast density: dense area, non-dense area, and percent dense area, from mammograms. No ExWAS has been conducted to date

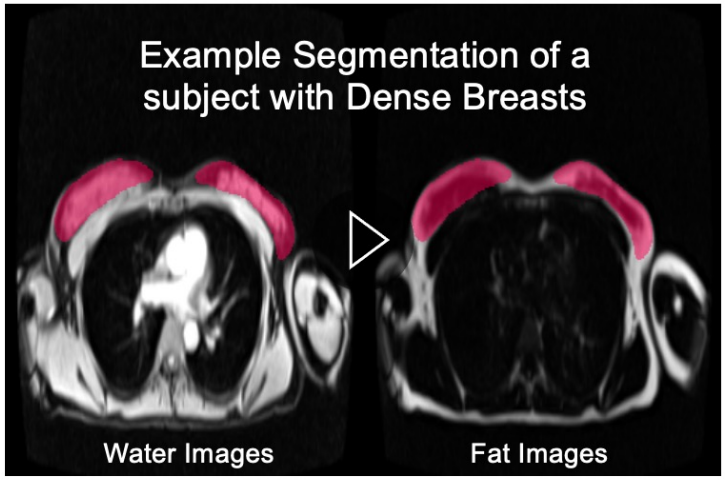


Sieh, W, et al. *Nat Comms.* 2020 11(1) :1-11 Mammography GWAS (N=24,192)

[1]McCormack and dos Santos Silva. *CEPB* 2006 15:1159-1169
[2]Boyd, NF et al. *Lancet Oncology* 2005 6:798-808
[3]Cuzick, J et al. *JNCI* 2011 103:744-752
[4]Boyd, NF et al. NEJM 2002 347:886-894
[5]Lindström, S, et al. *Nat Gen.* 2011 43(3):185-187
[6]Lindström, S, et al. *Nat Comms.* 2014 5(1):1-8
[7]Sieh, W, et al. *Nat Comms.* 2020 11(1) :5116

Step 1 : **Segment breasts** Ground truth breast segmentations from 122 subjects (~5,000 2D images) were used to train a U-Net model

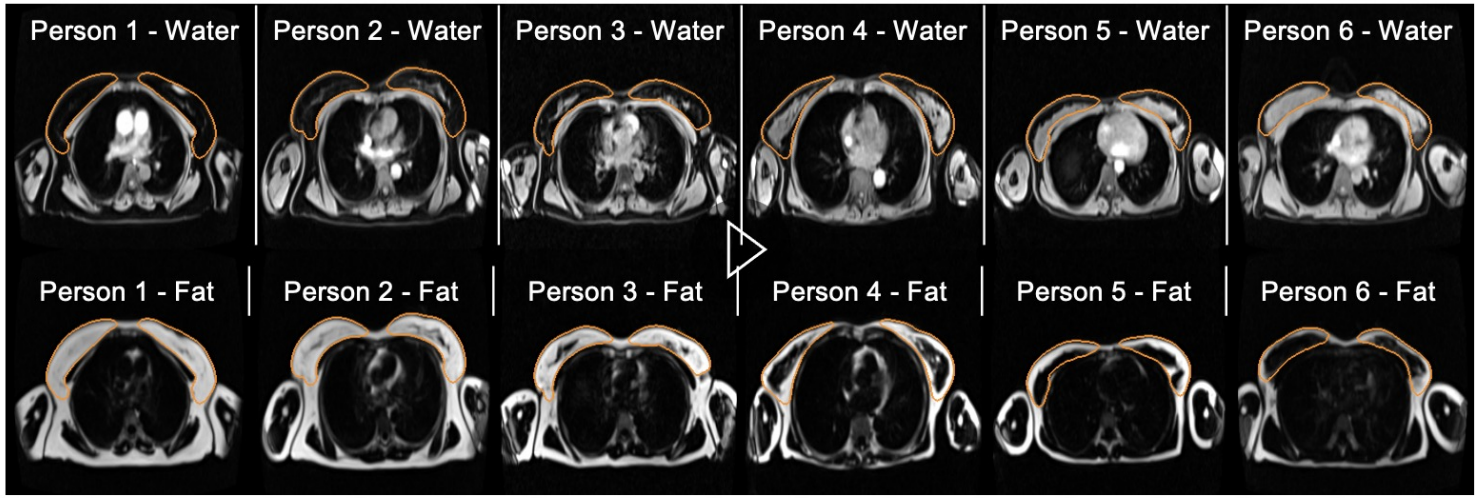




Step 2 : Measure density within breasts Water and fat-fraction maps were generated and used to derive:

Dense Volume (DV) = $\frac{Water}{Water+Fat}$

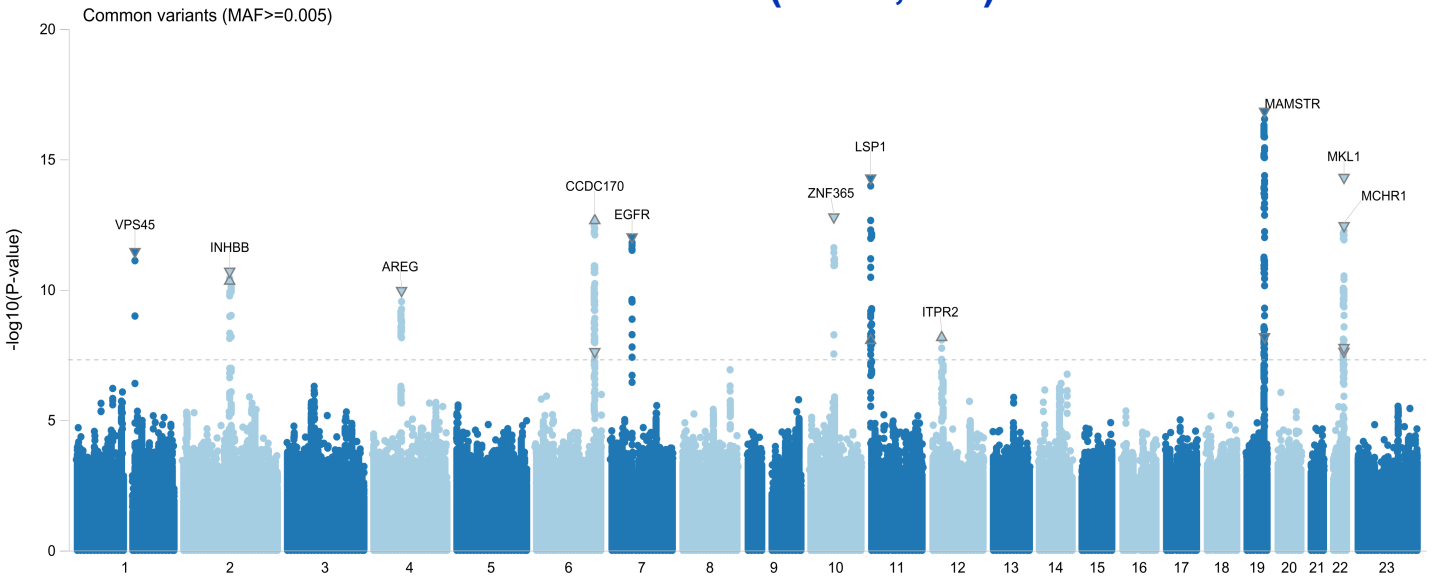
Non-Dense Volume (NDV) = $\frac{Fat}{Water+Fat}$



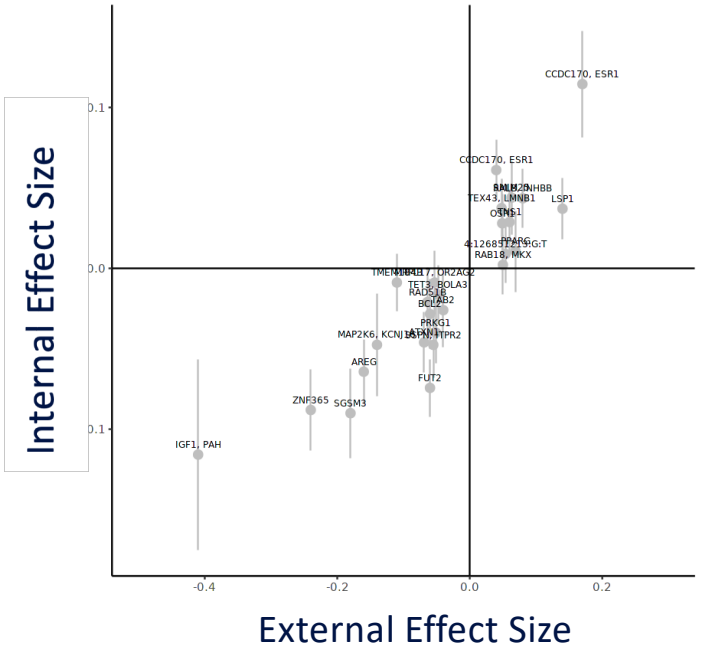
GWAS with MRI-derived Breast Density Phenotypes

Genetic associations with MRI-derived dense volume recapitulate associations with known breast density loci

Dense Volume (N=18,129)



Internal (y-axis) and external (x-axis) effect estimates are **directionally consistent at 26 variants across previously published loci**



11

Data at this scale requires highly efficient and flexible analysis

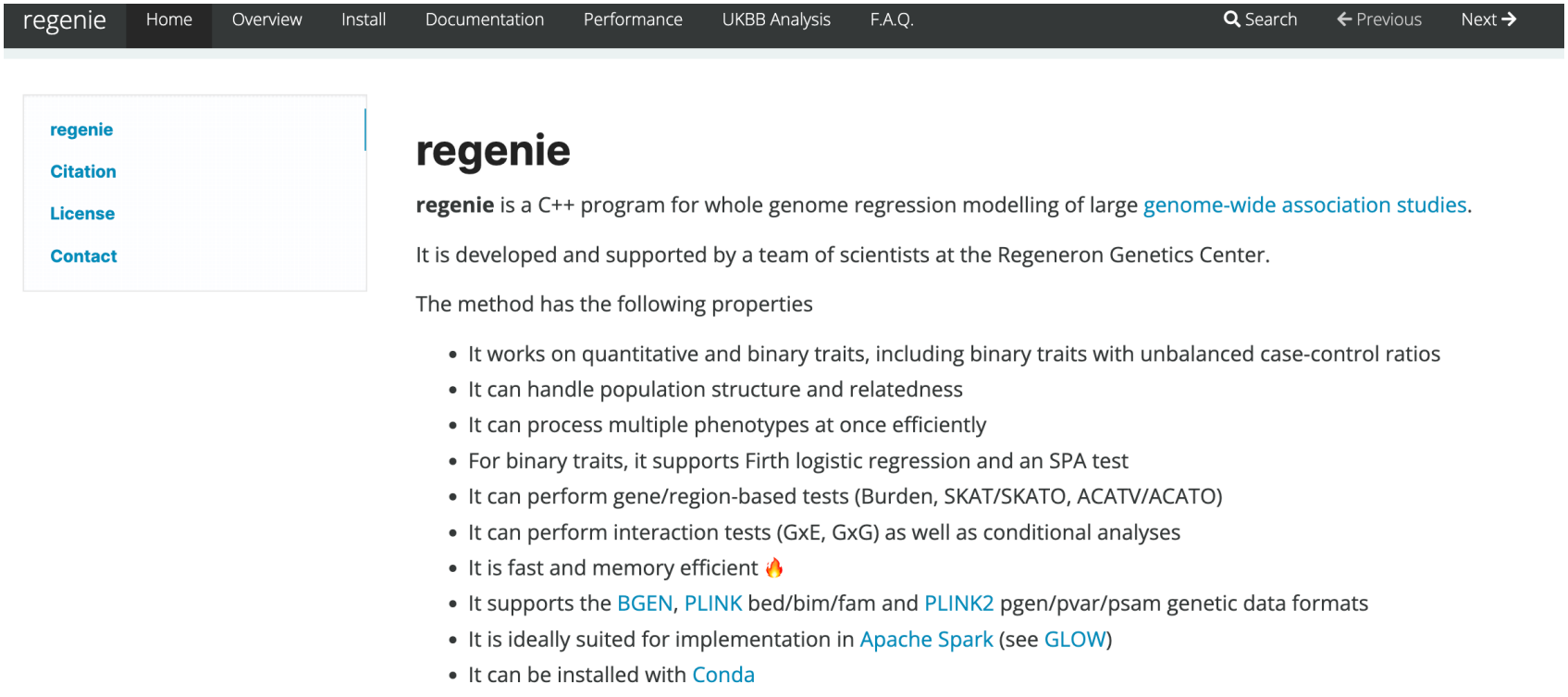
Technical Report | Published: 20 May 2021

Computationally efficient whole-genome regression for quantitative and binary traits

Joelle Mbatchou, Leland Barnard, Joshua Backman, Anthony Marcketta, Jack A. Kosmicki, Andrey Ziyatdinov, Christian Benner, Colm O'Dushlaine, Mathew Barber, Boris Boutkov, Lukas Habegger, Manuel Ferreira, Aris Baras, Jeffrey Reid, Goncalo Abecasis, Evan Maxwell & Jonathan Marchini

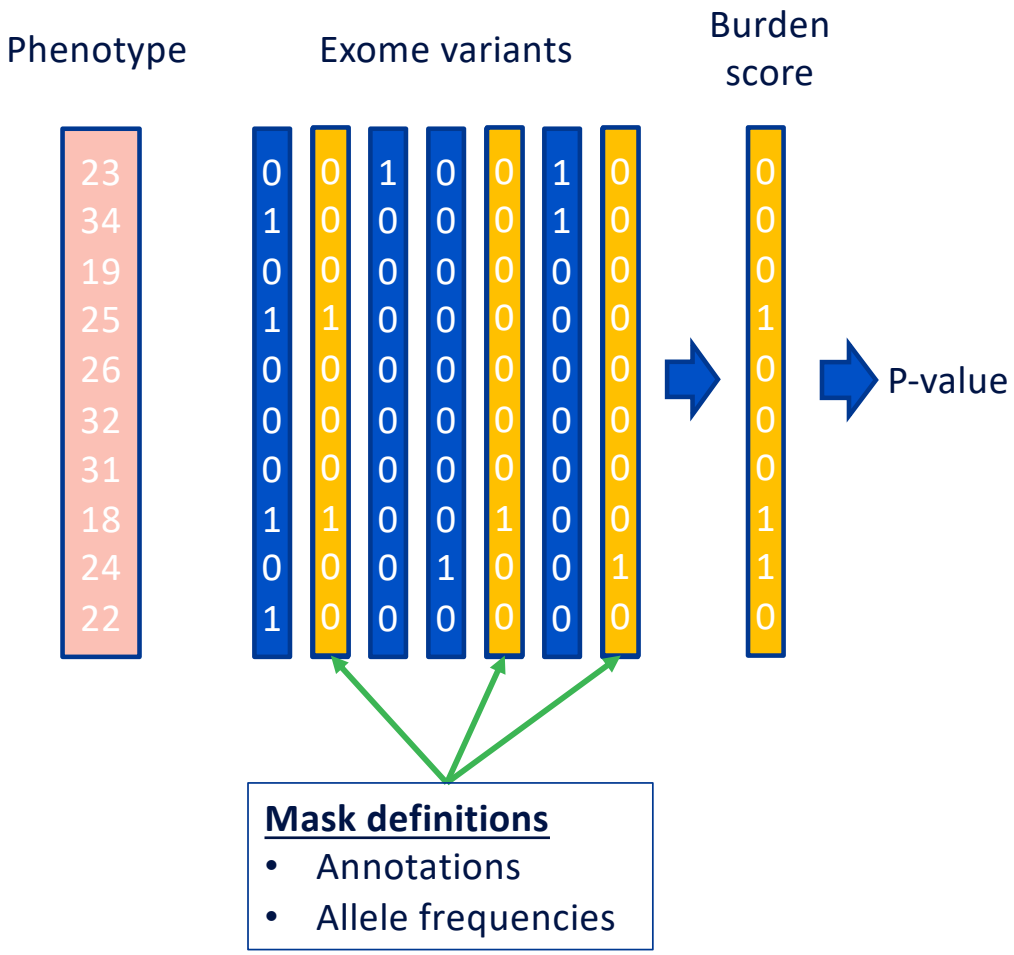
Nature Genetics **53**, 1097–1103 (2021) | Cite this article

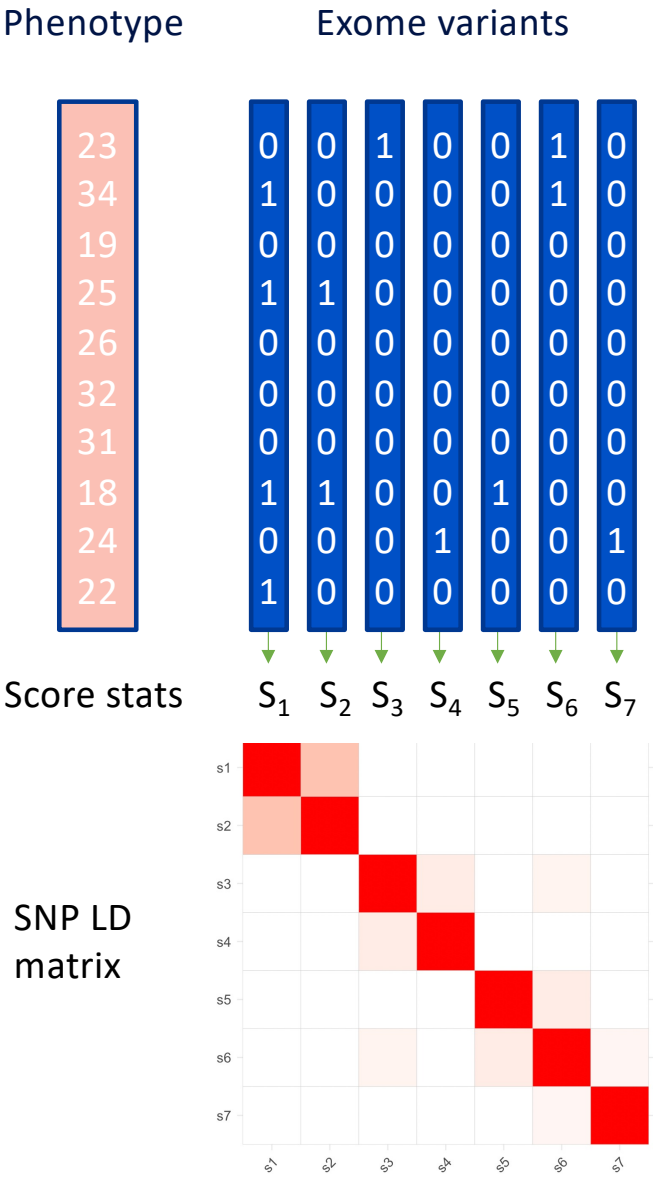
https://rgcgithub.github.io/regenie/



RGC

Efficient meta-analysis of gene-based tests (REMETA)


Phenotype
Exome variants
Burden score
23
34
19
25
26
32
31
18
24
22
P-value
Mask definitions
• Annotations
• Allele frequencies



Advantages

- Works for Burden and SKAT tests [1]
- Easy to change mask definition
- Amenable to meta-analysis [2,3,4] i.e. metaSKAT and metaSTARR

Disadvantages

Calculating and keeping track of LD matrices per trait

Mask definitions

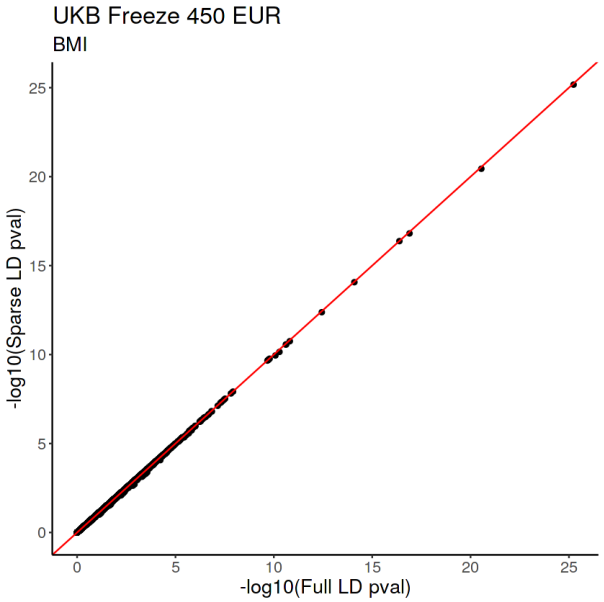
- Annotations
- allele frequencies

P-value [1]

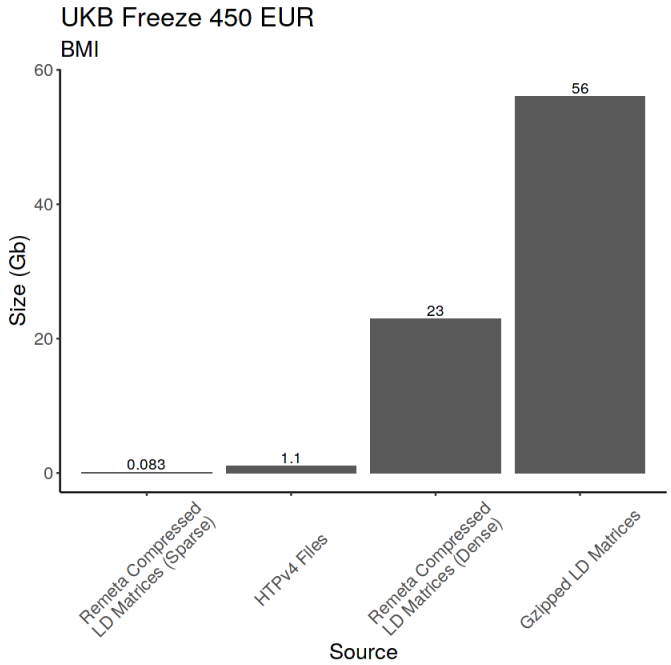
[1] Lee et al. (2012) AJHG 91(2): 224–237
[2] Lee at al. (2013) AJHG 93(1):42–53
[3] Liu et al. (2014) NatGen 46(2):200-4
[4] Li et al. (2023) NatGen 55(1): 154–164.

Gene-based meta-analysis on an industrial scale

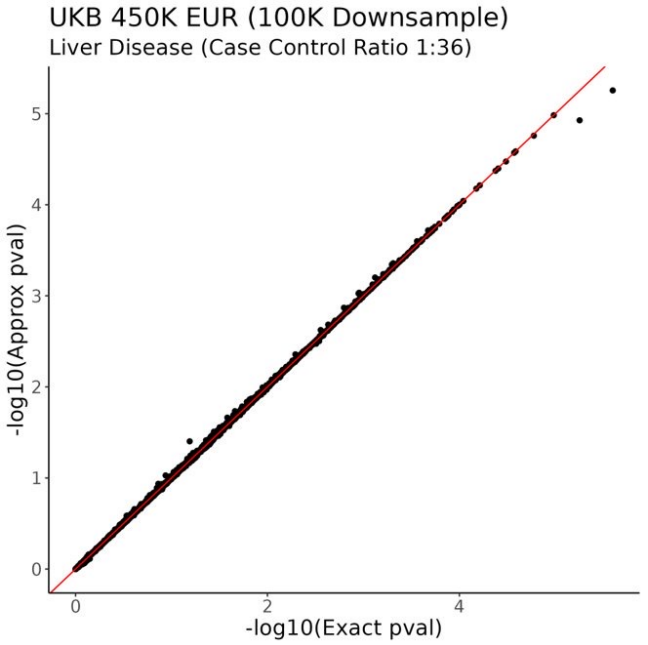
Challenge scale/ease-of-use when there are many cohorts, phenotypes/sub-phenotypes…



Sparse LD has no effect on gene p-values

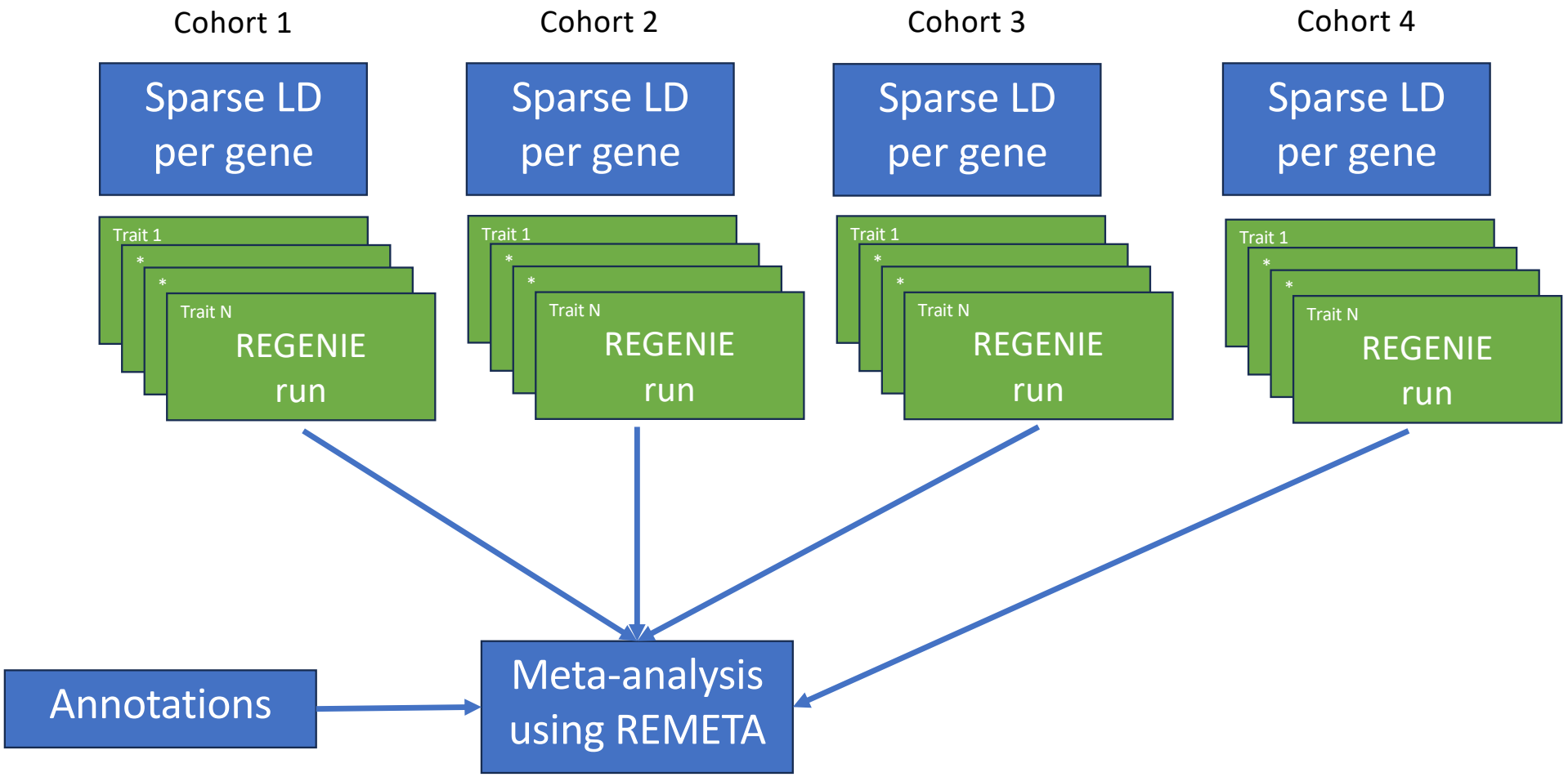


Sparse LD can be stored compactly



Using a reference LD file per cohort is accurate

Meta-analysis workflow



RGC presentations at ASHG

RGC Author	Assignment	Title	Presentation Date	Presentation Time	Location ***
Kuan-Han Wu	**Platform**	Rare variants associated with prostate cancer risk discovered from 255,640 male exomes influence risk of prostate cancer metastasis	02 November 2023	9:15 a.m. – 9:30 a.m.	Conv Ctr/Room 146B/Level 1
Joelle Mbatchou	Poster	Using protein language model annotations to improve the power of exome-wide association studies (PB4424)	02 November 2023	3:00 p.m. – 5:00 p.m.	Exhibit & Poster Hall, Halls A/B
Vijay Kumar	Poster	Population-scale analysis of the trinucleotide repeat expansion in the Huntingtin gene (HTT) from 854,251 human exomes (PB1723)	02 November 2023	3:00 p.m. – 5:00 p.m.	Exhibit & Poster Hall, Halls A/B
Yuxin Zou	Poster	Joint fine-mapping of single variants and gene-based tests from exome sequencing and genotype imputation (PB4238)	02 November 2023	3:00 p.m. – 5:00 p.m.	Exhibit & Poster Hall, Halls A/B
Veera Rajagopal	**Platform**	Discovering genes linked to both cognition and psychiatric disorders through analysis of 888,052 exomes	03 November 2023	2:15 p.m. – 2:30 p.m.	Conv Ctr/Ballroom B/Level 3
Rujin Wang	**Platform**	A large-scale meta-analysis of genome-wide association studies reveals genetics underlying Parkinson's Disease leveraging electronic health records	03 November 2023	2:30 p.m. – 2:45 p.m.	Conv Ctr/Room 207A/Level 2
Arden Moscati	Poster	Genetic ancestry-based case-control matching to improve power to trait-specific association analysis (PB4158)	03 November 2023	3:00 p.m. – 5:00 p.m.	Exhibit & Poster Hall, Halls A/B
Ariane Ayer	Poster	Phenome-wide genetic associations of educational attainment with mental and behavioral disorders (PB1706)	03 November 2023	3:00 p.m. – 5:00 p.m.	Exhibit & Poster Hall, Halls A/B
Liron Ganel	Poster	Rare variant analysis of MRI-derived fat distribution phenotypes strengthens detected effects compared to larger meta-analysis (PB1754)	03 November 2023	3:00 p.m. – 5:00 p.m.	Exhibit & Poster Hall, Halls A/B
Sahar Gelfman	Poster	A large meta-analysis identifies genes associated with Anterior Uveitis (PB1313)	03 November 2023	3:00 p.m. – 5:00 p.m.	Exhibit & Poster Hall, Halls A/B
Tyler Joseph	Poster	REMETA: Efficient meta-analysis of gene-based tests in large-scale genetic studies (PB4350)	03 November 2023	3:00 p.m. – 5:00 p.m.	Exhibit & Poster Hall, Halls A/B
Jack Kosmicki	**Platform**	Exome sequencing of >1 million individuals identifies 209 genes associated with adult human height	04 November 2023	10:45 a.m. – 11:00 p.m.	Conv Ctr/Room 202A/Level 2
Blair Zhang	Poster	Genetic risk score in age-related macular degeneration subtypes across electronic health record cohorts (PB4168)	04 November 2023	2:15 p.m. – 4:15 p.m.	Exhibit & Poster Hall, Halls A/B
Manav Kapoor	Poster	Genome-wide exploration of positively selected loci and their association to disease phenotypes in 30,000 individuals from Sri Lanka and Bangladesh (PB3072)	04 November 2023	2:15 p.m. – 4:15 p.m.	Exhibit & Poster Hall, Halls A/B
Sophia Praggastis	Poster	A genome-wide meta-analysis connects iron homeostasis to metabolic disease through poly-unsaturated fatty acid synthesis (PB1308)	04 November 2023	2:15 p.m. – 4:15 p.m.	Exhibit & Poster Hall, Halls A/B
Kathy Burch	**Platform**	Leveraging ~937K exomes to estimate cancer risk conferred by rare deleterious germline variants in hereditary cancer risk genes	05 November 2023	10:00 a.m. – 10:15 a.m.	Conv Ctr/Room 202A/Level 2

*** **Location:** Walter E. Washington Convention Center