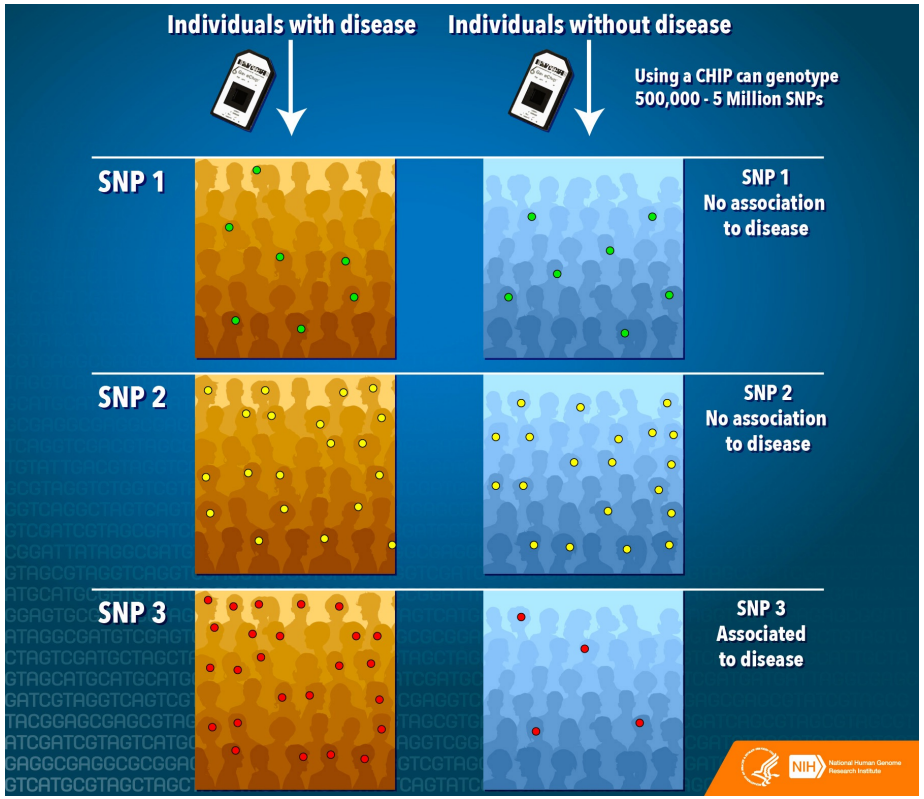
Secure large-scale genome-wide association studies using homomorphic encryption

Alexander Gusev

Dana-Farber Cancer Institute
Harvard Medical School

GWAS: Genome-Wide Association Studies



Goal:
Identify genetic mutations causal for disease

Input:
Disease case/control patients and cofactors
~1M genotyped common polymorphisms

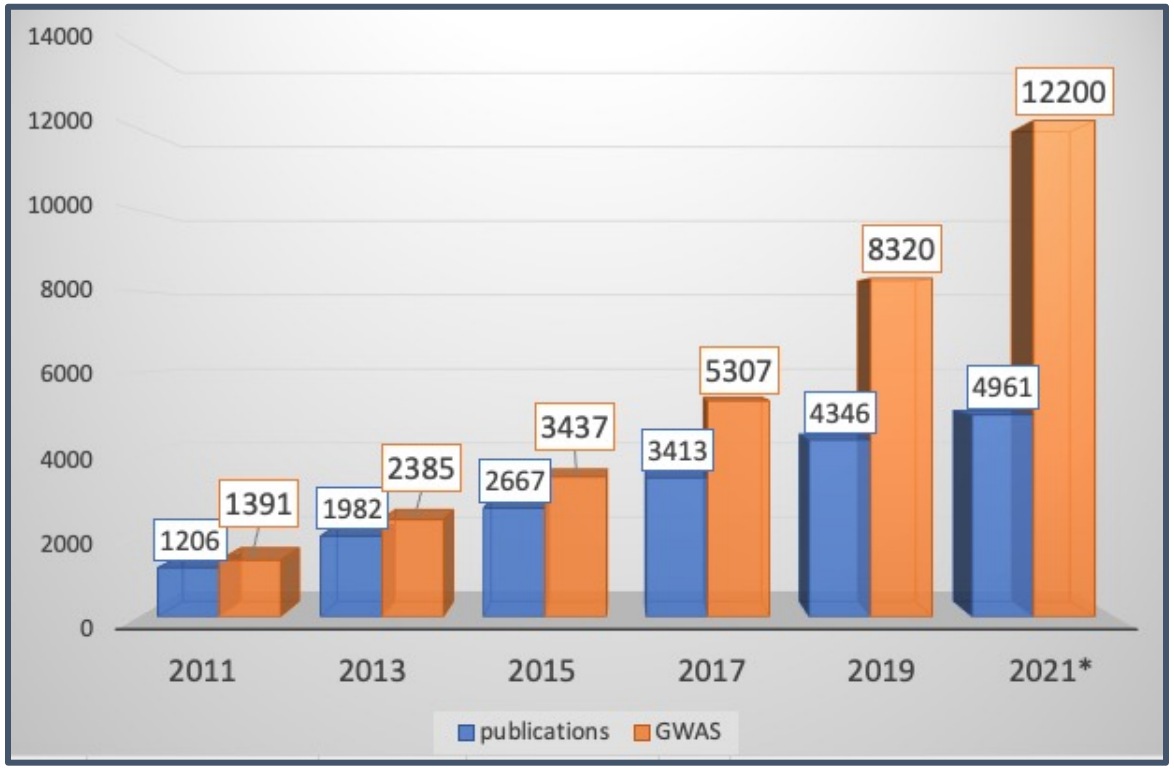
Model:
Test each polymorphism against disease status

Output:
Variant-disease association

GWAS associations for complex traits

- Thousands of reported associations
- Consistent replication across cohorts
- Together explaining a large fraction of heritable disease

- Genetic discovery is now mostly a matter of sample size



GWAS associations explain clinical outcomes

GWAS associations for clinical outcomes

nature genetics

A coding variant in *RARG* confers susceptibility to anthracycline-induced cardiotoxicity in childhood cancer

Folefac Aminkeng[1,2,13], Amit P Bhavsar[2,3,13], Henk Visscher[1,4], Shahrad R Rassekh[2,5], Yuling Li[2,3], Jong W Lee[1,2], Liam R Brunham[6], Huib N Caron[7], Elvira C van Dalen[7], Leontien C Kremer[7], Helena J van der Pal[7,8], Ursula Amstutz[2,3,12], Michael J Rieder[9], Daniel Bernstein[10], Bruce C Carleton[2,3,11,14], Michael R Hayden[1,2,6,14], Colin J D Ross[1–3,11,14] & The Canadian Pharmacogenomics Network for Drug Safety Consortium[15]

GWAS associations for clinical outcomes

ARTICLE

Received 24 Jun 2014 | Accepted 10 Mar 2015 | Published 5 May 2015

DOI: 10.1038/ncomms7889

Two susceptibility loci identified for prostate cancer aggressiveness

Sonja I. Berndt[1,*], Zhaoming Wang[1,2,*], Meredith Yeager[1,2], Michael C. Alavanja[1], Demetrius Albanes[1], Laufey Amundadottir[1], Gerald Andriole[3], Laura Beane Freeman[1], Daniele Campa[4], Geraldine Cancel-Tassin[5], Federico Canzian[6], Jean-Nicolas Cornu[1], Olivier Cussenot[5], W. Ryan Diver[7], Susan M. Gapstur[7], Henrik Grönberg[8], Christopher A. Haiman[9], Brian Henderson[9], Amy Hutchinson[2], David J. Hunter[10], Timothy J. Key[11], Suzanne Kolb[12], Stella Koutros[1], Peter Kraft[10], Loic Le Marchand[13], Sara Lindström[10], Mitchell J. Machiela[1], Elaine A. Ostrander[14], Elio Riboli[15], Fred Schumacher[9], Afshan Siddiq[16], Janet L. Stanford[12,17], Victoria L. Stevens[7], Ruth C. Travis[11], Konstantinos K. Tsilidis[18], Jarmo Virtamo[19], Stephanie Weinstein[1], Fredrik Wilkund[8], Jianfeng Xu[20], S. Lilly Zheng[20], Kai Yu[1], William Wheeler[21], Han Zhang[1], African Ancestry Prostate Cancer GWAS Consortium†, Joshua Sampson[1], Amanda Black[1], Kevin Jacobs[1], Robert N. Hoover[1], Margaret Tucker[1] & Stephen J. Chanock[1]

GWAS associations for clinical outcomes

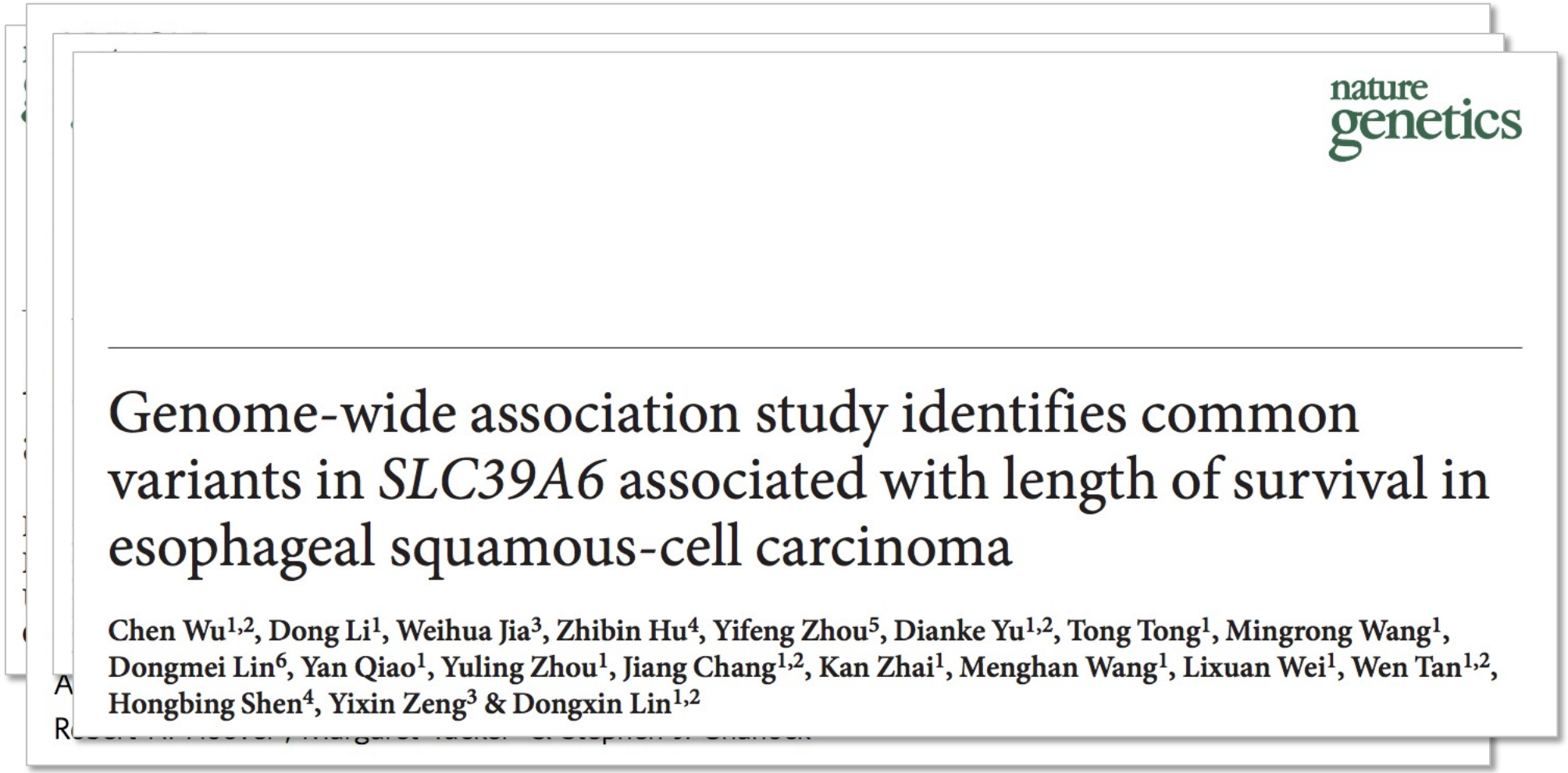
nature genetics

A three-stage genome-wide association study identifies a susceptibility locus for late radiotherapy toxicity at 2q24.1

Laura Fachal[1,2], Antonio Gómez-Caamaño[3], Gillian C Barnett[4], Paula Peleteiro[3], Ana M Carballo[3], Patricia Calvo-Crespo[3], Sarah L Kerns[5], Manuel Sánchez-García[6], Ramón Lobato-Busto[6], Leila Dorling[4], Rebecca M Elliott[7], David P Dearnaley[8], Matthew R Sydes[9], Emma Hall[10], Neil G Burnet[11], Ángel Carracedo[1,2,12], Barry S Rosenstein[5], Catharine M L West[7], Alison M Dunning[4] & Ana Vega[1,2]

African Ancestry Prostate Cancer GWAS Consortium†, Joshua Sampson[1], Amanda Black[1], Kevin Jacobs[1], Robert N. Hoover[1], Margaret Tucker[1] & Stephen J. Chanock[1]

GWAS associations for clinical outcomes



nature genetics

Genome-wide association study identifies common variants in *SLC39A6* associated with length of survival in esophageal squamous-cell carcinoma

Chen Wu[1,2], Dong Li[1], Weihua Jia[3], Zhibin Hu[4], Yifeng Zhou[5], Dianke Yu[1,2], Tong Tong[1], Mingrong Wang[1], Dongmei Lin[6], Yan Qiao[1], Yuling Zhou[1], Jiang Chang[1,2], Kan Zhai[1], Menghan Wang[1], Lixuan Wei[1], Wen Tan[1,2], Hongbing Shen[4], Yixin Zeng[3] & Dongxin Lin[1,2]

GWAS associations inform drug targets

GWAS associations support drug targets

RESEARCH ARTICLE

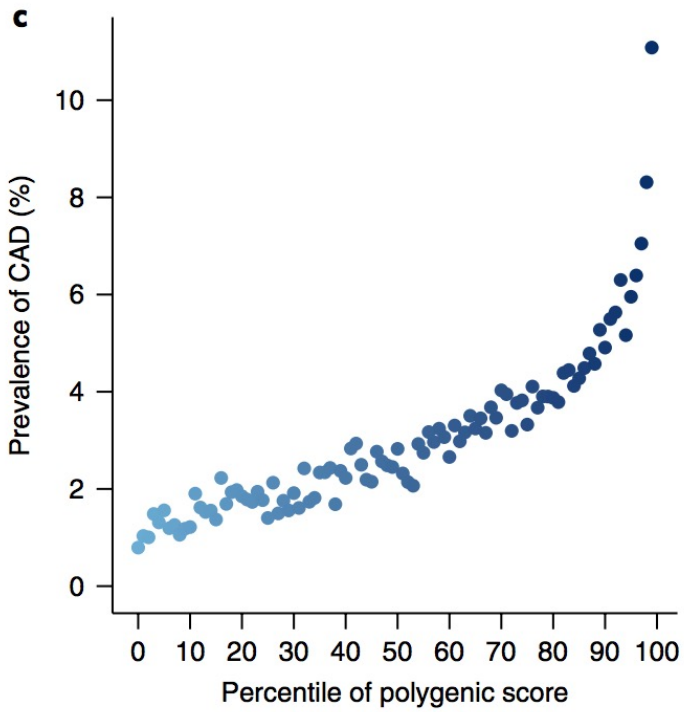
Are drug targets with genetic support twice as likely to be approved? Revised estimates of the impact of genetic support for drug mechanisms on the probability of drug approval

Emily A. King*, J. Wade Davis, Jacob F. Degner

"we find the use of human genetic evidence increases approval from Phase I by greater than two-fold, and, for Mendelian associations, the positive association holds prospectively"

GWAS results can predict genetic risk

Polygenic Risk Prediction (PRS) from GWAS



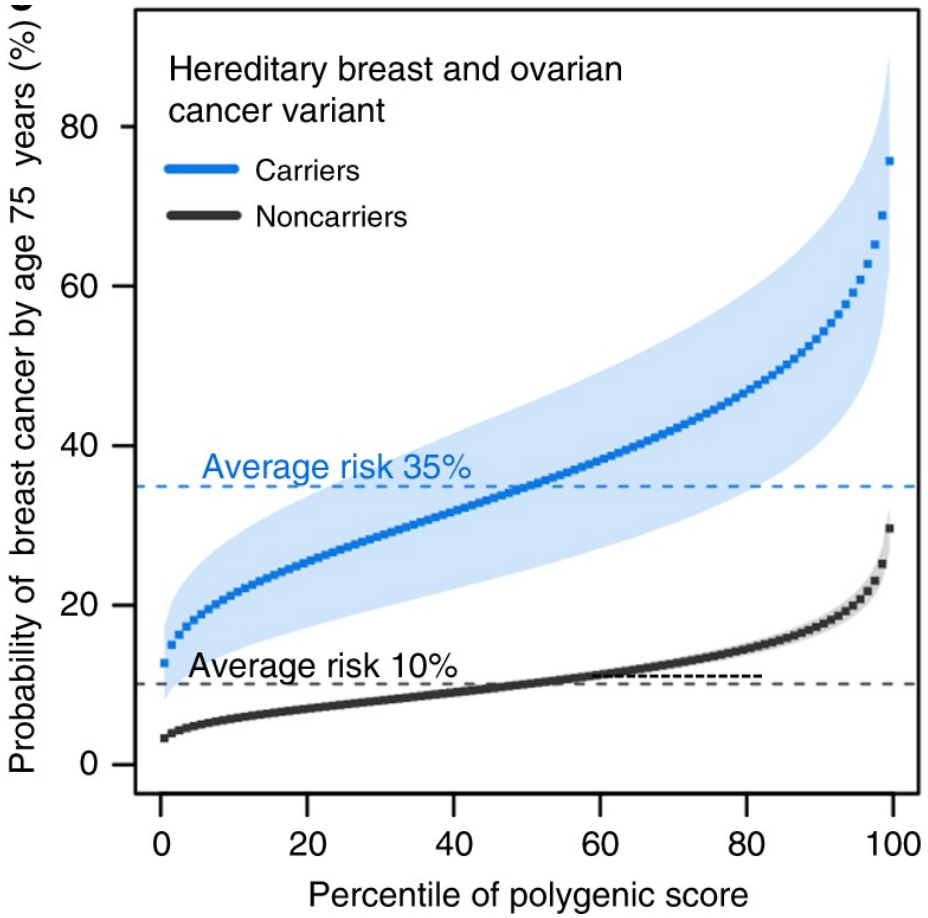
Score derived from GWAS

$$PRS_i = \sum_j^M \hat{\beta}_j \times dosage_{ij}$$

"For coronary artery disease, [high PRS] prevalence is 20-fold higher than the carrier frequency of rare monogenic mutations conferring comparable risk. We propose that it is time to contemplate the inclusion of polygenic risk prediction in clinical care, and discuss relevant issues."

Khera et al. 2018 Nat Genet

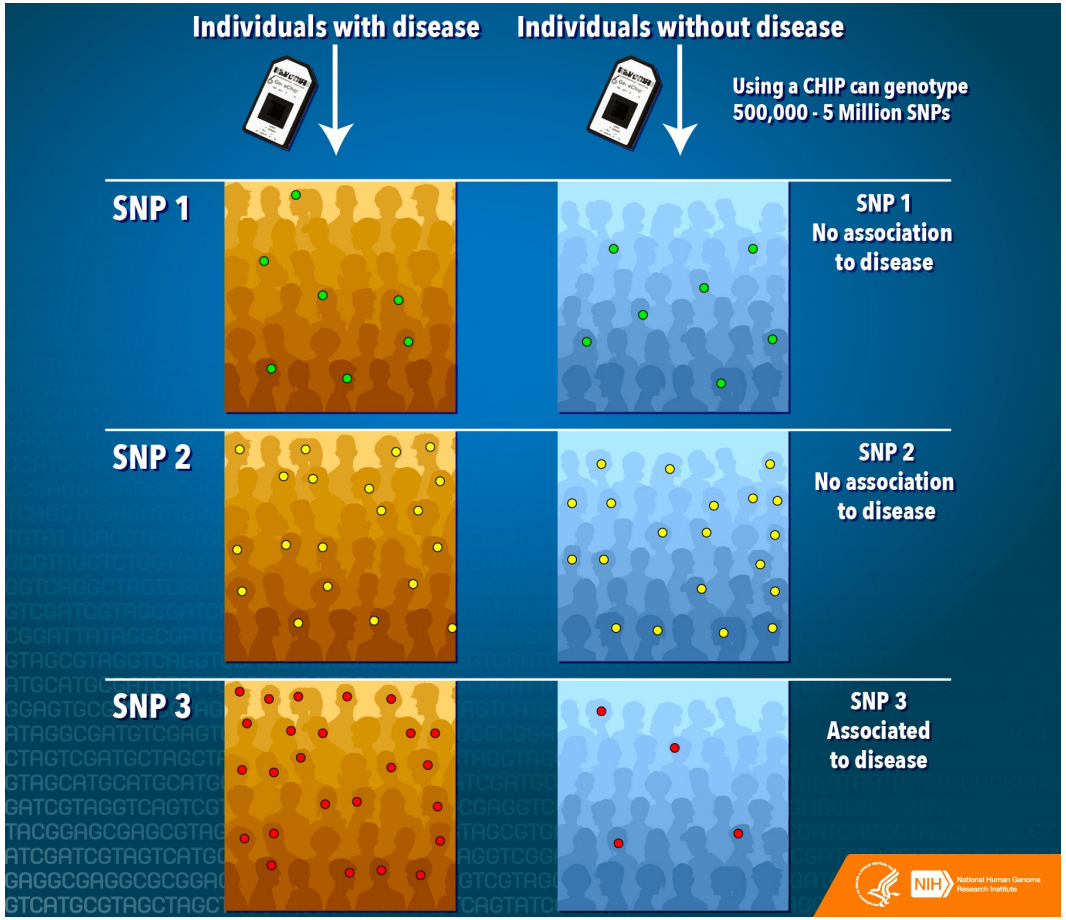
Polygenic score modifies monogenic risk



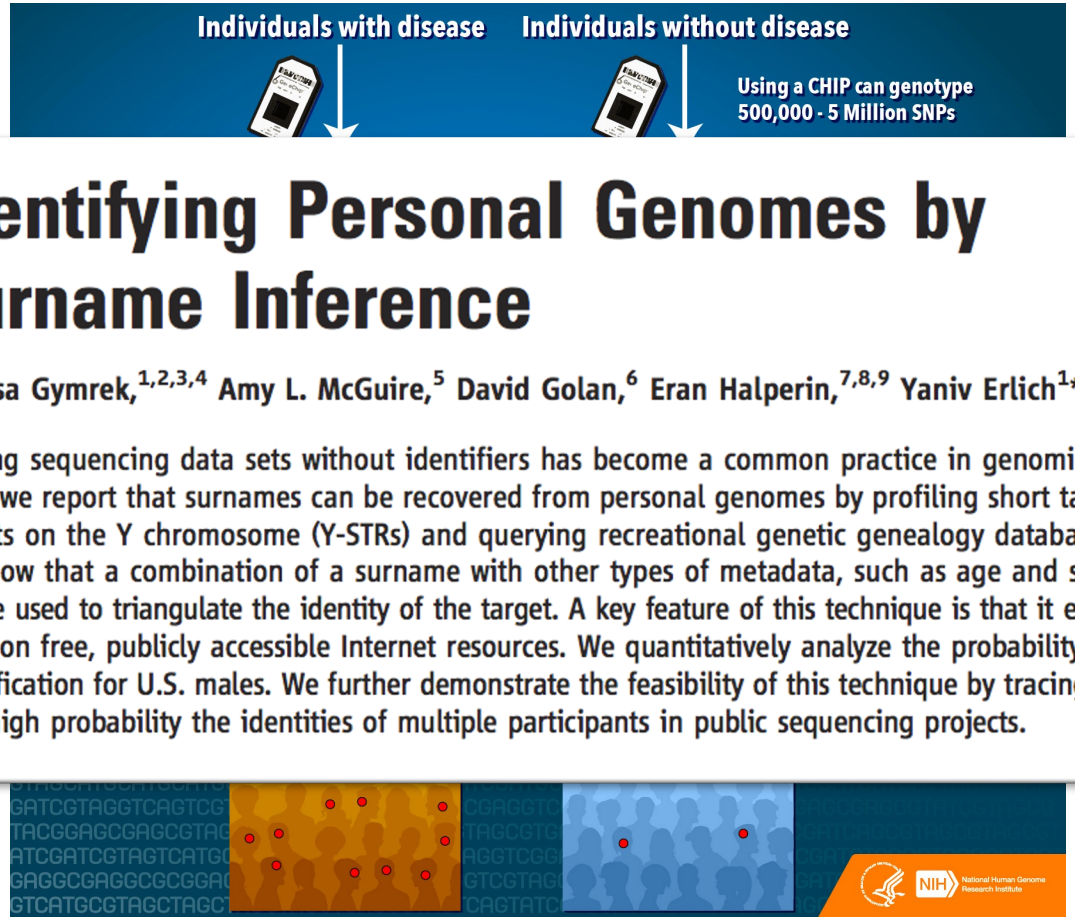
Fahed et al. 2020 Nat Comms

Barriers for GWAS

Barriers: Individual-level privacy



Barriers: Individual-level privacy



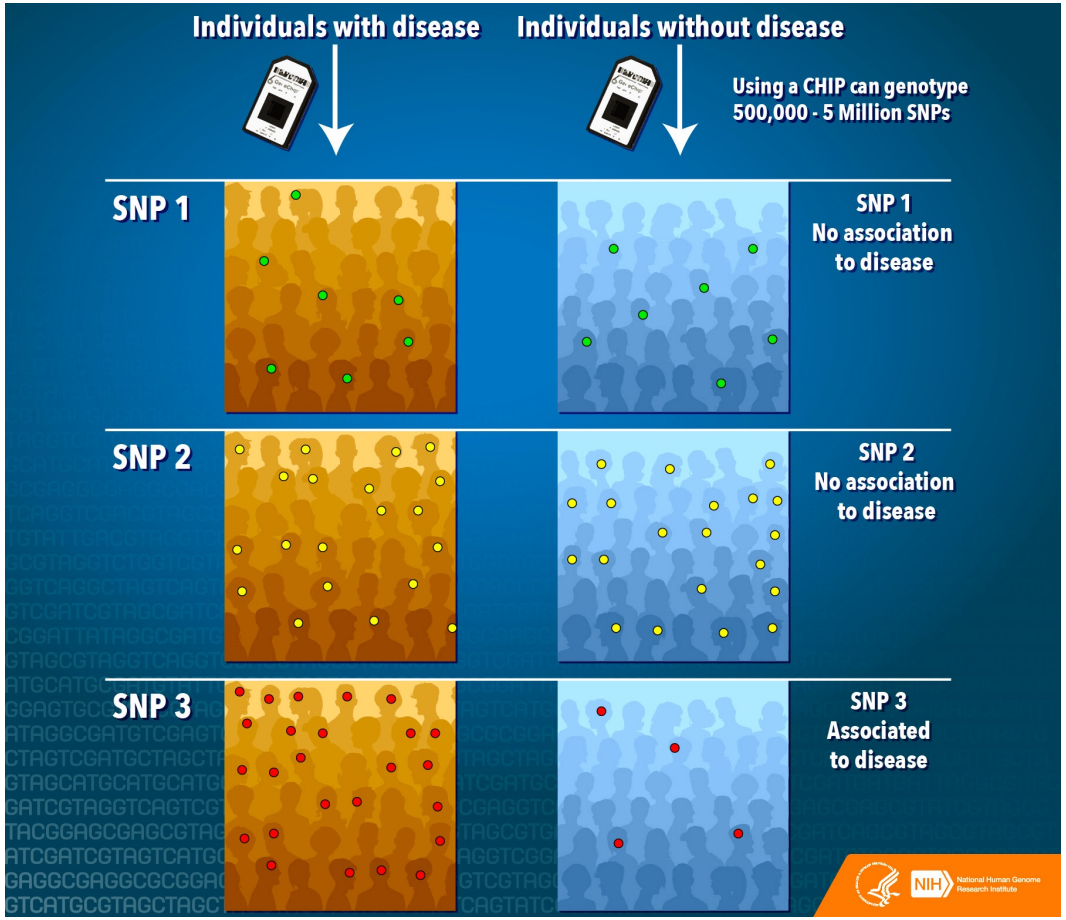
Identifying Personal Genomes by Surname Inference

Melissa Gymrek,[1,2,3,4] Amy L. McGuire,[5] David Golan,[6] Eran Halperin,[7,8,9] Yaniv Erlich[1]*

Sharing sequencing data sets without identifiers has become a common practice in genomics. Here, we report that surnames can be recovered from personal genomes by profiling short tandem repeats on the Y chromosome (Y-STRs) and querying recreational genetic genealogy databases. We show that a combination of a surname with other types of metadata, such as age and state, can be used to triangulate the identity of the target. A key feature of this technique is that it entirely relies on free, publicly accessible Internet resources. We quantitatively analyze the probability of identification for U.S. males. We further demonstrate the feasibility of this technique by tracing back with high probability the identities of multiple participants in public sequencing projects.

Barriers: Sensitive data sharing

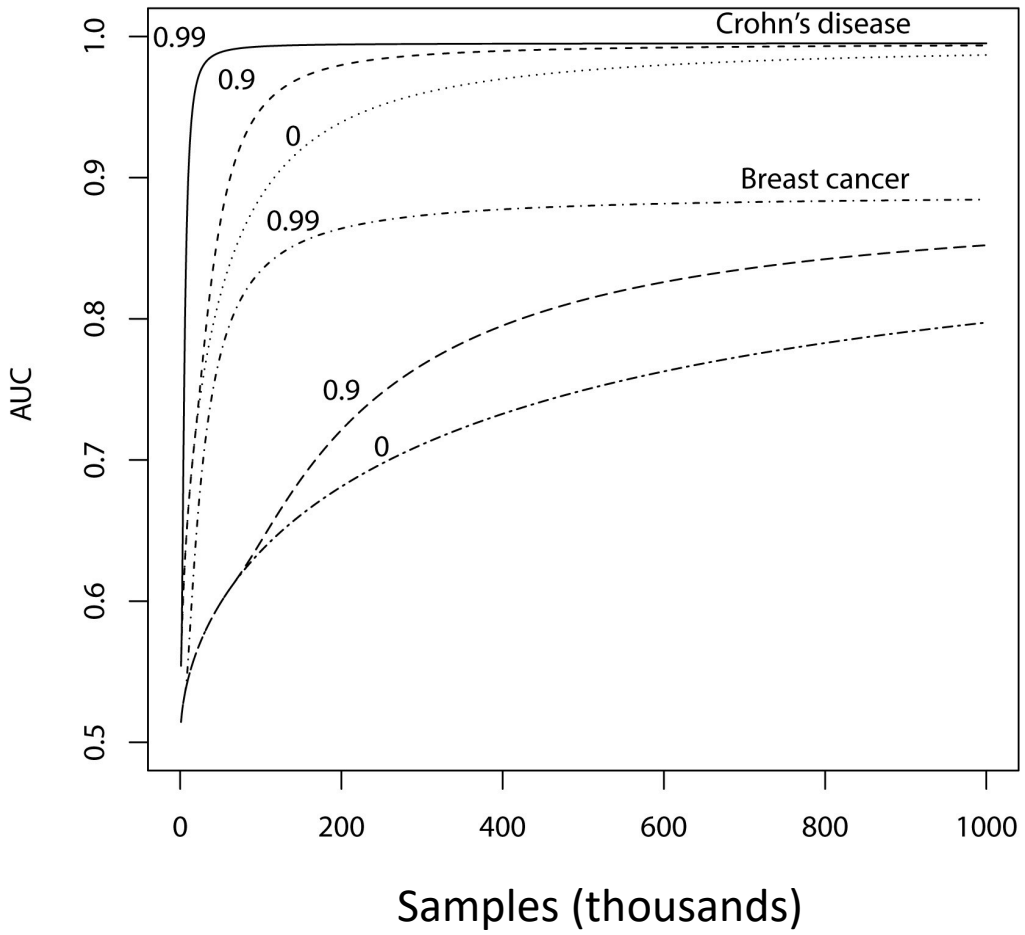
Dr. Jekyll's data



Dr. Hyde's data

Barriers: Scalability

Risk Prediction accuracy



Dudbridge 2013 PLoS Genet

Solution: Secure, Encrypted GWAS

Previous work: secure multi-party GWAS

Encrypted computing approach: secure multi-party computation[1]

- Statistical test: Cochran Armitage trend test
- Benchmark GWAS: 26k samples x 260k SNPs

Results:

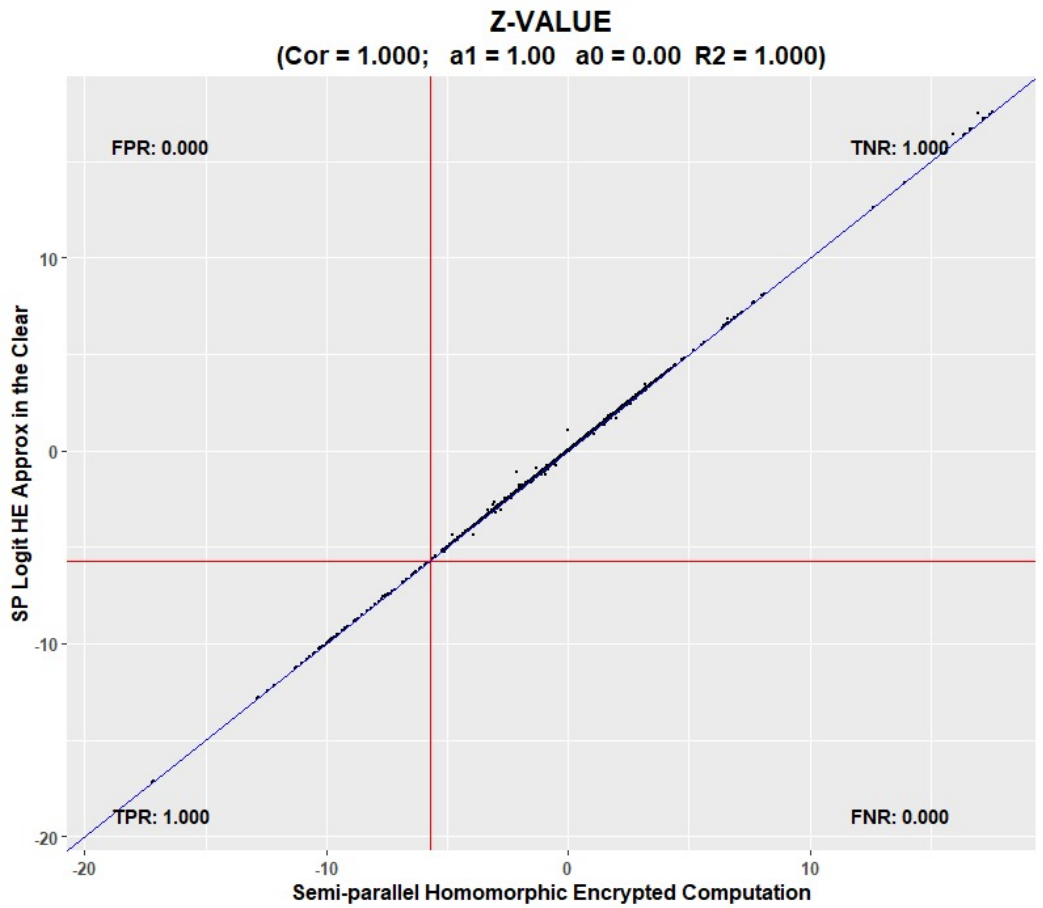
- Runtime on 100k samples x 500k SNPs: **193 hours**
- Requires live, interactive communication
- Logistic regression “does not yield a practical runtime”
- *Expect that HE would be 5,000-10,000x slower and infeasible*[2]

[1] Cho et al. 2018 Nat Biotechnol; [2] Jagadeesh et al. 2017 Science

Results

	Prior MPC work	Our HE work
Algorithm	Multi-party computation	Homomorphic encryption
Statistical test	Cochran Armitage Trend (CAT)	Allelic χ^2 (CAT equivalent) Logistic regression
Dataset	26k samples x 260k SNPs + extrapolation	
Accuracy of test	Nearly perfect	
Runtime on 100k samples x 500k SNPs	193 hours Practically impossible	**5.6 hours** **234 hours (log reg)**

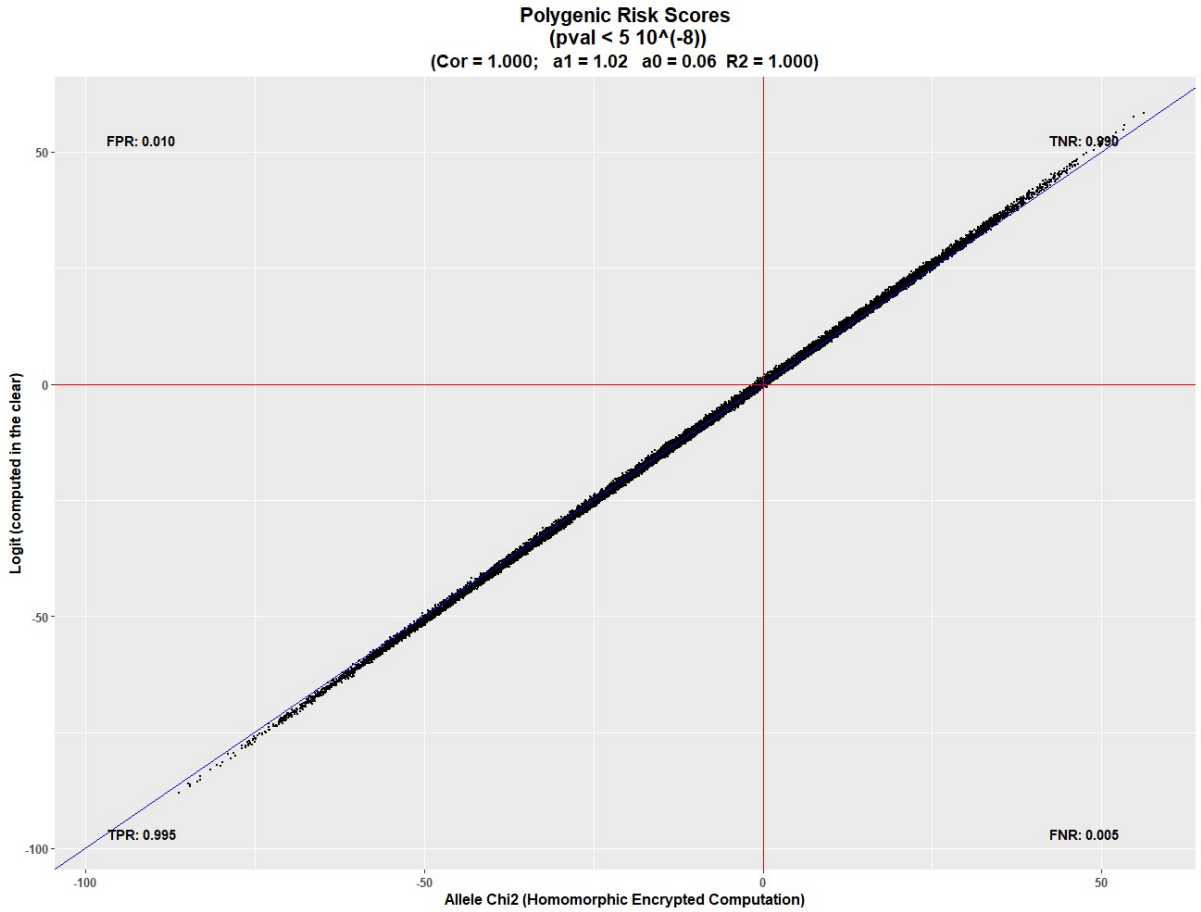
No loss in accuracy overall



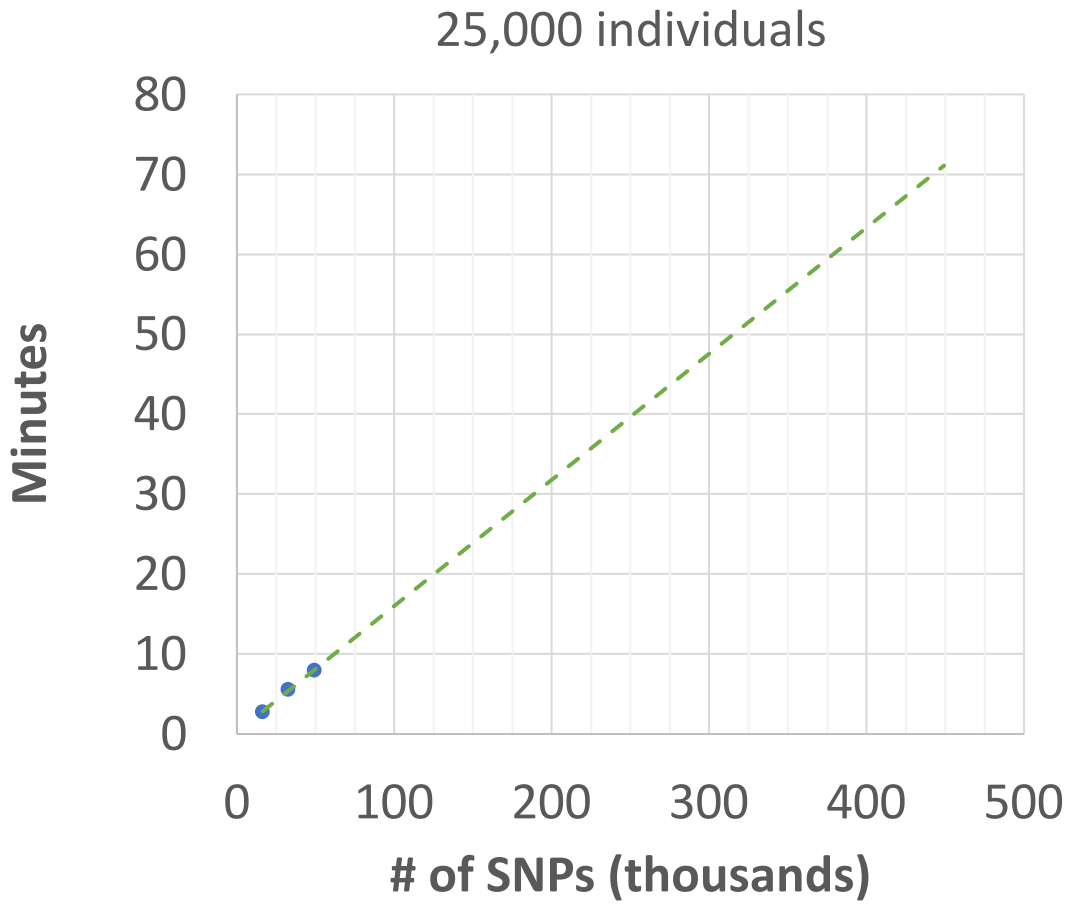
No loss in accuracy for top hits

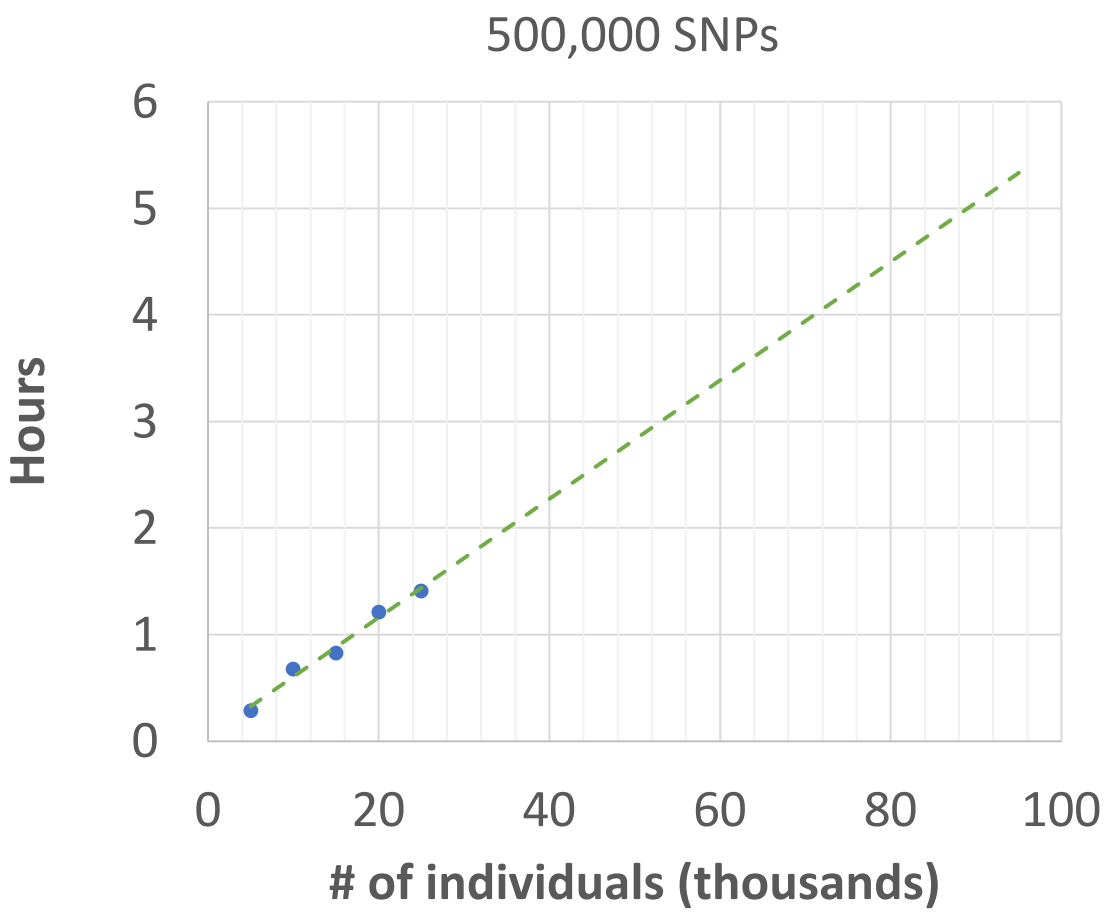
SNP	Clear OR	Encrypted OR	Clear Chi^2	Encrypted Chi^2
rs2230199_C	1.40	1.40	263.13	263.13
rs114203272_T	0.64	0.64	61.11	61.11
rs10033900_T	1.13	1.13	51.64	51.64
rs943080_C	0.89	0.89	41.76	41.76
rs2043085_T	0.89	0.89	41.40	41.40
rs8135665_T	1.13	1.13	33.96	33.96
rs79037040_G	0.92	0.92	25.35	25.35
rs114212178_T	0.82	0.82	6.72	6.72

No loss in accuracy for genomic prediction



Scalable beyond 100,000 individuals





PNAS Proceedings of the National Academy of Sciences of the United States of America

Secure large-scale genome-wide association studies using homomorphic encryption

Marcelo Blatt[a,1], Alexander Gusev[a,b,1], Yuriy Polyakov[a,1,2], and Shafi Goldwasser[a,c,1,2]

[a]Duality Technologies, Inc., Newark, NJ 07103; [b]Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02215; and [c]Simons Institute for the Theory of Computing, University of California, Berkeley, CA 94720

Secure-GWAS: Opportunities

GWAS identifies **causal mutations, drug targets, and risk/outcome predictors** … but effective GWAS is not possible without **data sharing**

Secure-GWAS for researchers:

- GWAS across institutions without data sharing
- Secure collaboration on sensitive phenotypes

Secure-GWAS for individuals:

- Participate in studies on-demand without sacrificing privacy