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

Folefac Aminkeng1,2,13, Amit P Bhavsar2,3,13, Henk Visscher1,4, Shahrad R Rassekh2,5, Yuling Li2,3, Jong W Lee1,2, Liam R Brunham6, Huib N Caron7, Elvira C van Dalen7, Leontien C Kremer7, Helena J van der Pal7,8, Ursula Amstutz2,3,12, Michael J Rieder9, Daniel Bernstein10, Bruce C Carleton2,3,11,14, Michael R Hayden1,2,6,14, Colin J D Ross1–3,11,14 & The Canadian Pharmacogenomics Network for Drug Safety Consortium15
GWAS associations for clinical outcomes

Two susceptibility loci identified for prostate cancer aggressiveness

Sonja I. Berndt, Zhaoming Wang, Meredith Yeager, Michael C. Alavanja, Demetrius Albanes, Laufey Amundadottir, Gerald Andriele, Laura Beane Freeman, Daniele Campa, Geraldine Cancel-Tassin, Federico Canzian, Jean-Nicolas Cornu, Olivier Cussenot, W. Ryan Diver, Susan M. Gapstur, Henrik Grönberg, Christopher A. Haiman, Brian Henderson, Amy Hutchinson, David J. Hunter, Timothy J. Key, Suzanne Kolb, Stella Koutros, Peter Kraft, Loic Le Marchand, Sara Lindström, Mitchell J. Machiela, Elaine A. Ostrander, Elio Riboli, Fred Schumacher, Afshan Siddiq, Janet L. Stanford, Victoria L. Stevens, Ruth C. Travis, Konstantinos K. Tsilidis, Jarmo Virtamo, Stephanie Weinstein, Fredrik Wilkund, Jianfeng Xu, S. Lilly Zheng, Kai Yu, William Wheeler, Han Zhang, African Ancestry Prostate Cancer GWAS Consortium, Joshua Sampson, Amanda Black, Kevin Jacobs, Robert N. Hoover, Margaret Tucker & Stephen J. Chanock
GWAS associations for clinical outcomes

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

Laura Fachal1,2, Antonio Gómez-Caamaño3, Gillian C Barnett4, Paula Peleteiro3, Ana M Carballo3, Patricia Calvo-Crespo3, Sarah L Kerns5, Manuel Sánchez-García6, Ramón Lobato-Busto6, Leila Dorling4, Rebecca M Elliott7, David P Dearnaley8, Matthew R Sydes9, Emma Hall10, Neil G Burnet11, Ángel Carracedo1,2,12, Barry S Rosenstein5, Catharine M L West7, Alison M Dunning4 & Ana Vega1,2

African Ancestry Prostate Cancer GWAS Consortium†, Joshua Sampson1, Amanda Black1, Kevin Jacobs1, Robert N. Hoover1, Margaret Tucker1 & Stephen J. Chanock1
GWAS associations for clinical outcomes

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

Chen Wu¹,², Dong Li¹, Weihua Jia³, Zhibin Hu⁴, Yifeng Zhou⁵, Dianke Yu¹,², Tong Tong¹, Mingrong Wang¹, Dongmei Lin⁶, Yan Qiao¹, Yuling Zhou¹, Jiang Chang¹,², Kan Zhai¹, Menghan Wang¹, Lixuan Wei¹, Wen Tan¹,², Hongbing Shen⁴, Yixin Zeng³ & Dongxin Lin¹,²
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

“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”

Emily A. King*, J. Wade Davis, Jacob F. Degner
GWAS results can predict genetic risk
"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."
Polygenic score modifies monogenic risk

Fahed et al. 2020 Nat Comms
Barriers for GWAS
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 Erlich1*

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

AUC

Samples (thousands)

Crohn's disease
Breast cancer

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

| Algorithm                  | Prior MPC work                  | Our HE work                      |
|----------------------------|---------------------------------|----------------------------------|
| Multi-party computation    |                                 | Homomorphic encryption           |
| Statistical test           | Cochran Armitage Trend (CAT)    | Allelic $\chi^2$ (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

Graphs showing the relationship between the number of Single Nucleotide Polymorphisms (SNPs) and the time taken for processing.

- **25,000 individuals**: The graph on the left shows a linear relationship between the number of SNPs (in thousands) and the number of minutes required for processing.
- **500,000 SNPs**: The graph on the right shows a similar linear relationship for the number of SNPs and the number of hours required for processing.

The diagrams illustrate that the processing time increases linearly with the number of SNPs, indicating scalability beyond 100,000 individuals.
Secure large-scale genome-wide association studies using homomorphic encryption

Marcelo Blatt\textsuperscript{a,1}, Alexander Gusev\textsuperscript{a,b,1}, Yuriy Polyakov\textsuperscript{a,1,2}, and Shafi Goldwasser\textsuperscript{a,c,1,2}

\textsuperscript{a}Duality Technologies, Inc., Newark, NJ 07103; \textsuperscript{b}Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02215; and \textsuperscript{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