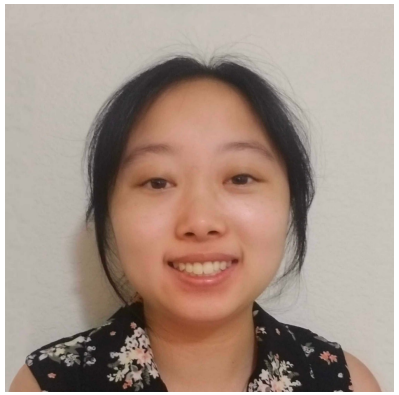
UCDAVIS
Department of Animal Science
Quantitative Genetics Lab

Homomorphic Encryption to Enable Sharing of Confidential Data in Agricultural Genome to Phenome

Hao Cheng
Department of Animal Science
University of California, Davis

AGBT, Mar 28 2023
(15 + 5 mins)

UCDAVIS
Department of Animal Science
Quantitative Genetics Lab



New Data

Big Data

Safe Data

Intermediate omics data
High-throughput phenotypes

UCDAVIS Department of Animal Science Quantitative Genetics Lab

Data Sharing is Essential

- Large data sets are needed to answer many genome to phenome questions
  - Sharing the extensive proprietary phenotypic and genetic data generated by industry

- Data sharing is essential for both Academia and Private Industry

UC DAVIS Department of Animal Science Quantitative Genetics Lab

Data Sharing in Academia

Journal

GENETICS

Issues | More Content ▾ | Series ▾ | Submit ▾ | Purchase | About ▾

Data Policy

When you publish in GENETICS, you help to catalyze scientific advances by sharing your experimental reagents, results and interpretations. For these articles to have the greatest impact, authors need to make unique research materials and data freely available to other investigators (see GENETICS, 184: 1).

Data

All data that are not represented fully within the tables and figures and necessary for confirming the conclusions presented in your manuscript must be made publicly available. Data should be archived in a public repository or database managed by a third party. Please note that journal policy does not allow for data to be available upon request.

Funding Agencies

EXECUTIVE OFFICE OF THE PRESIDENT
OFFICE OF SCIENCE AND TECHNOLOGY POLICY
WASHINGTON, D.C. 20502

August 25, 2022

MEMORANDUM FOR THE HEADS OF EXECUTIVE DEPARTMENTS AND AGENCIES

FROM: Dr. Alondra Nelson
Deputy Assistant to the President and Deputy Director for Science and Society
Performing the Duties of Director
Office of Science and Technology Policy (OSTP)

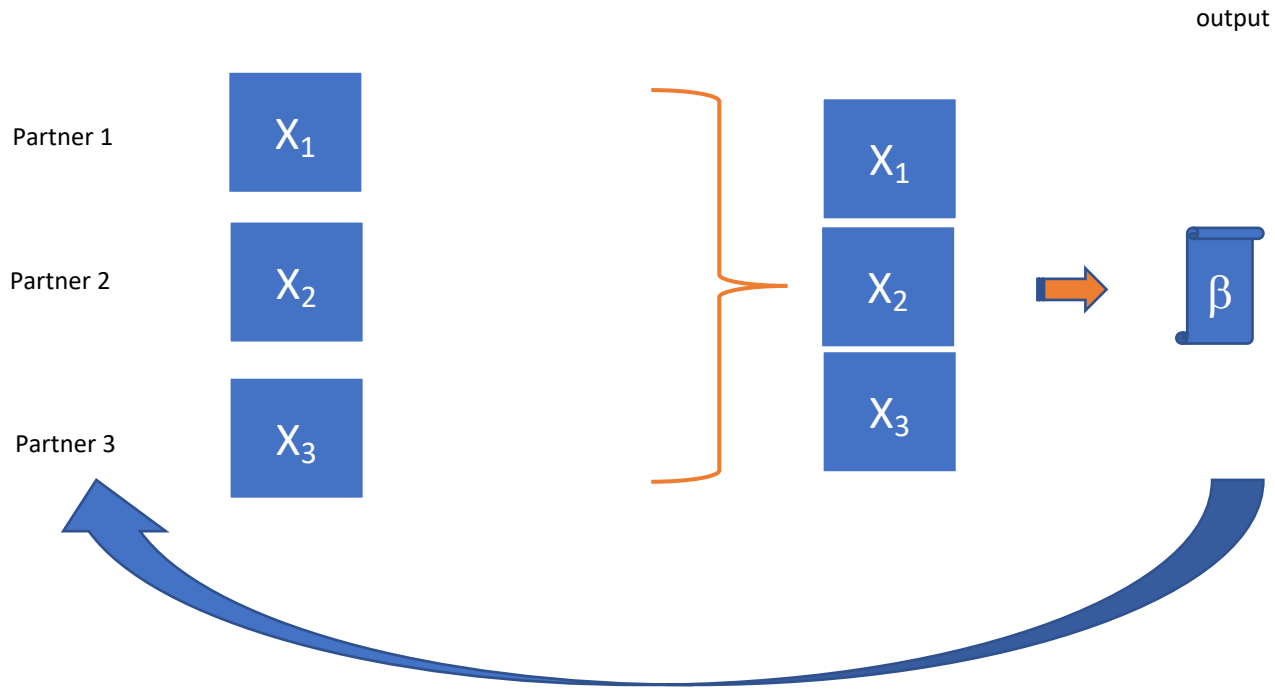
SUBJECT: Ensuring Free, Immediate, and Equitable Access to Federally Funded Research

This memorandum provides policy guidance to federal agencies with research and development expenditures on updating their public access policies. In accordance with this memorandum, OSTP recommends that federal agencies, to the extent consistent with applicable law:

1. Update their public access policies as soon as possible, and no later than December 31st, 2025, to make publications and their supporting data resulting from federally funded research publicly accessible without an embargo on their free and public release;
2. Establish transparent procedures that ensure scientific and research integrity is maintained in public access policies; and,
3. Coordinate with OSTP to ensure equitable delivery of federally funded research results and data.

Data Sharing in Private Industry

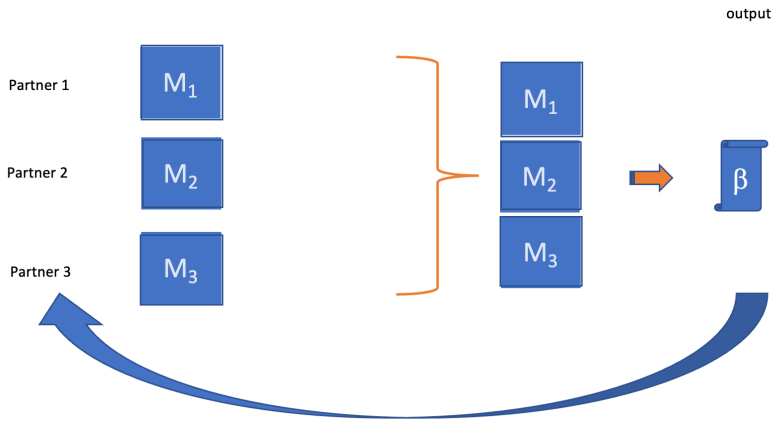
- Many joint analyses have demonstrated that the outcome of genetic analyses is significantly improved by combining multiple genetic studies into a single analysis



UCDAVIS Department of Animal Science Quantitative Genetics Lab

Data Sharing is Essential, But

- privacy concerns
- intellectual property, trade secrets
- it could be used to undermine a company's competitive advantage



UCDAVIS Department of Animal Science Quantitative Genetics Lab

FAIR

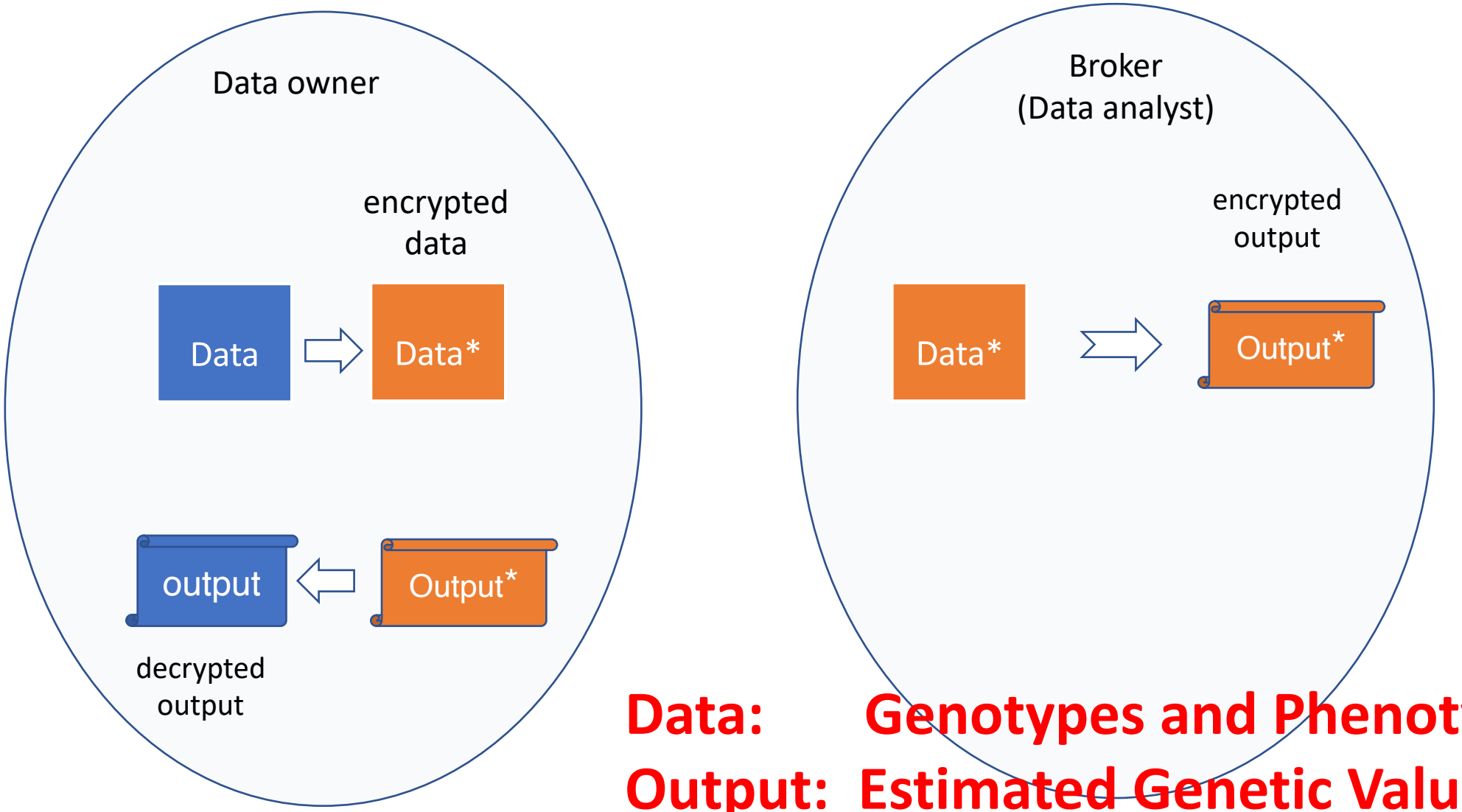
Findable, Accessible, Interoperable, Reusable

Privacy concerns
Intellectual property
Trade secret

Safe Data: secure encryption of the data

- Confidential information is protected/obscured
- Allow further validation and research using encrypted data only
- Obtain same outcomes/results using the encrypted data (and the key)

Homomorphic Encryption



Data: Genotypes and Phenotypes

Output: Estimated Genetic Values

Homomorphic Encryption for Genotypes and Phenotypes (HEGP)

SNP Genotypes

0	1	1	2	2	0
1	2	1	2	0	0
2	2	1	2	0	1
0	0	2	1	2	0

Phenotypes

a	1.9
b	2.7
c	2.1
d	1.3

$\boldsymbol{P}\boldsymbol{X}$ $n \times nn \times p$

0.066	0.502	0.855	-0.113
-0.611	-0.659	0.438	0.030
0.735	-0.549	0.222	-0.331
-0.287	0.112	-0.168	-0.936

$\boldsymbol{P}\boldsymbol{y}$ $n \times nn \times 1$

Encrypted Genotypes

2.21	2.78	1.20	2.73	-0.09	0.86
0.22	-1.05	-0.77	-1.63	-1.16	0.44
-0.11	0.08	-0.25	0.48	0.81	0.22
-0.22	-0.4	-2.22	-1.62	-2.45	-0.17

Encrypted Phenotypes

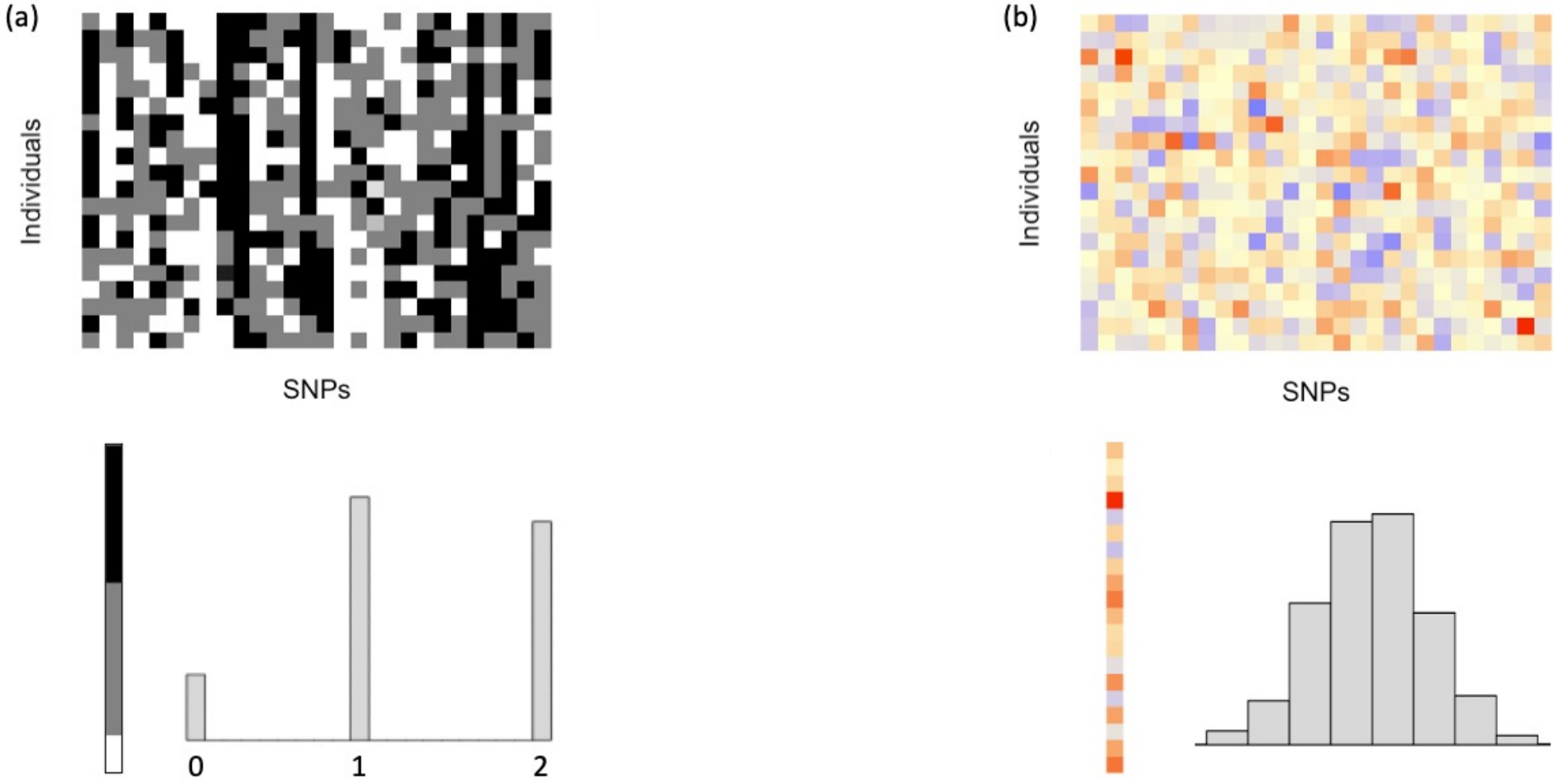
3.13
-1.98
-0.05
-1.81

linear transformation of
raw phenotypes (y) and genotypes (X)
by multiplying a randomly generated orthogonal matrix (P)

The encrypted data will be shared
The key (P) is not shared

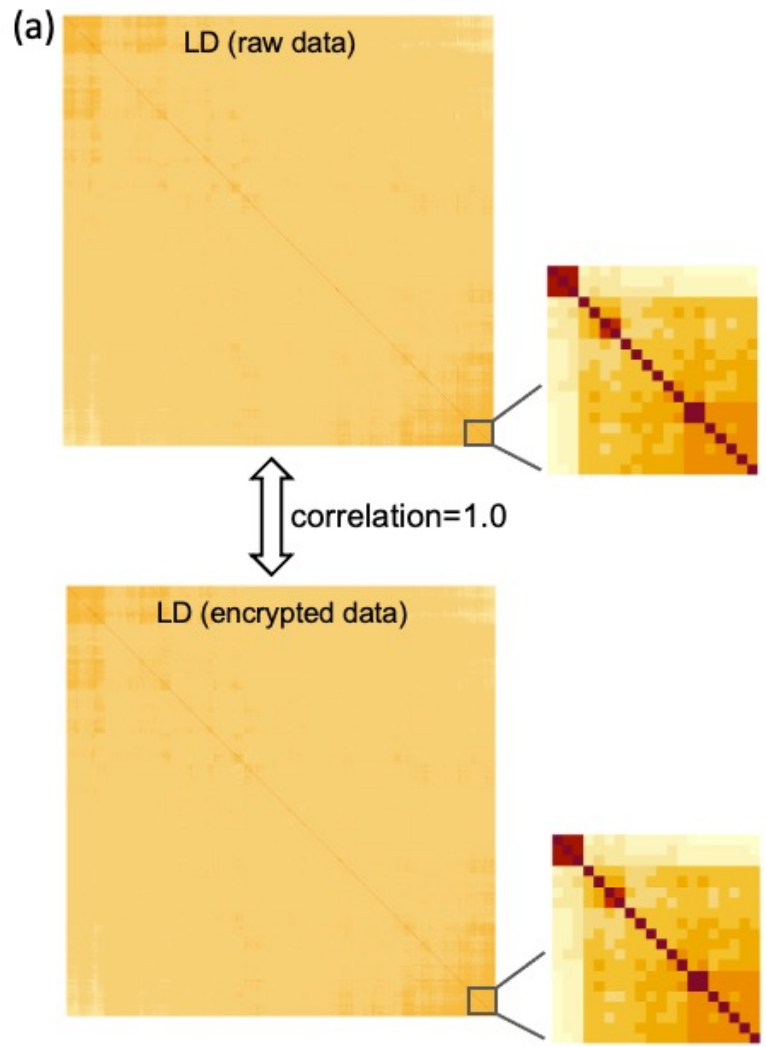
Richard Mott et al. Private Genomes and Public SNPs: Homomorphic Encryption of Genotypes and Phenotypes for Shared Quantitative Genetics, 2020

How do encrypted genotypes look like?

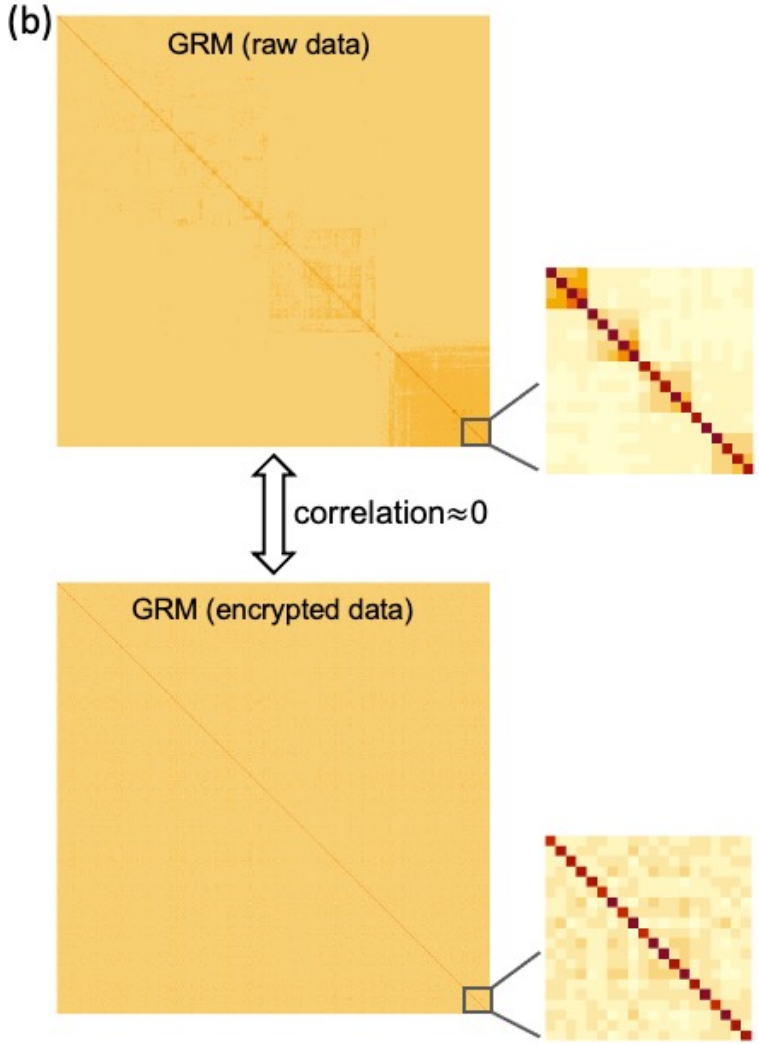


UC DAVIS Department of Animal Science Quantitative Genetics Lab

HEGP preserves relationships between SNPs



HEGP scrambles relationships between individuals



UCDAVIS Department of Animal Science Quantitative Genetics Lab

Data Analysis

- pig genotypes from *Cleveland et. al (2012)*
  - n=3534; p=50,436 (MAF>0.01)
  - Centered, each marker has zero mean

- simulated phenotypes
  - h^2=0.1; 0.3; 0.5; 0.7
  - QTL%= 1%; 10%; 50%; 100% (QTLs are included in markers)
  - group effects: individuals were randomly split into 4 groups
  - 10 replicates were applied

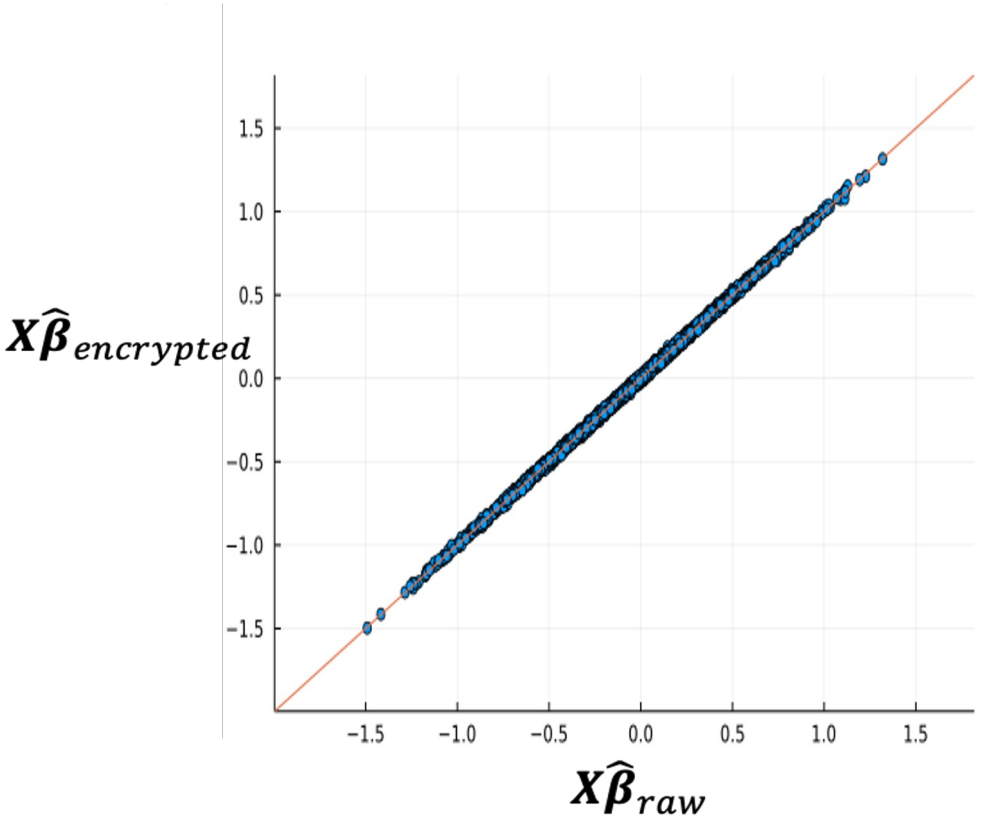
- BayesCπ (Bayesian regression method with mixture priors)

Results: ~ identical estimated breeding values

- $corr(\boldsymbol{X}\hat{\boldsymbol{\beta}}_{raw}, \boldsymbol{X}\hat{\boldsymbol{\beta}}_{encrypted}) \approx 0.9996$

  Given raw genotypes $\boldsymbol{X}$

		h^2			
		0.1	0.3	0.5	0.7
QTL%	1%	0.9993	0.9997	0.9998	0.9998
	10%	0.9992	0.9996	0.9997	0.9997
	50%	0.9992	0.9996	0.9997	0.9997
	100%	0.9992	0.9995	0.9997	0.9997

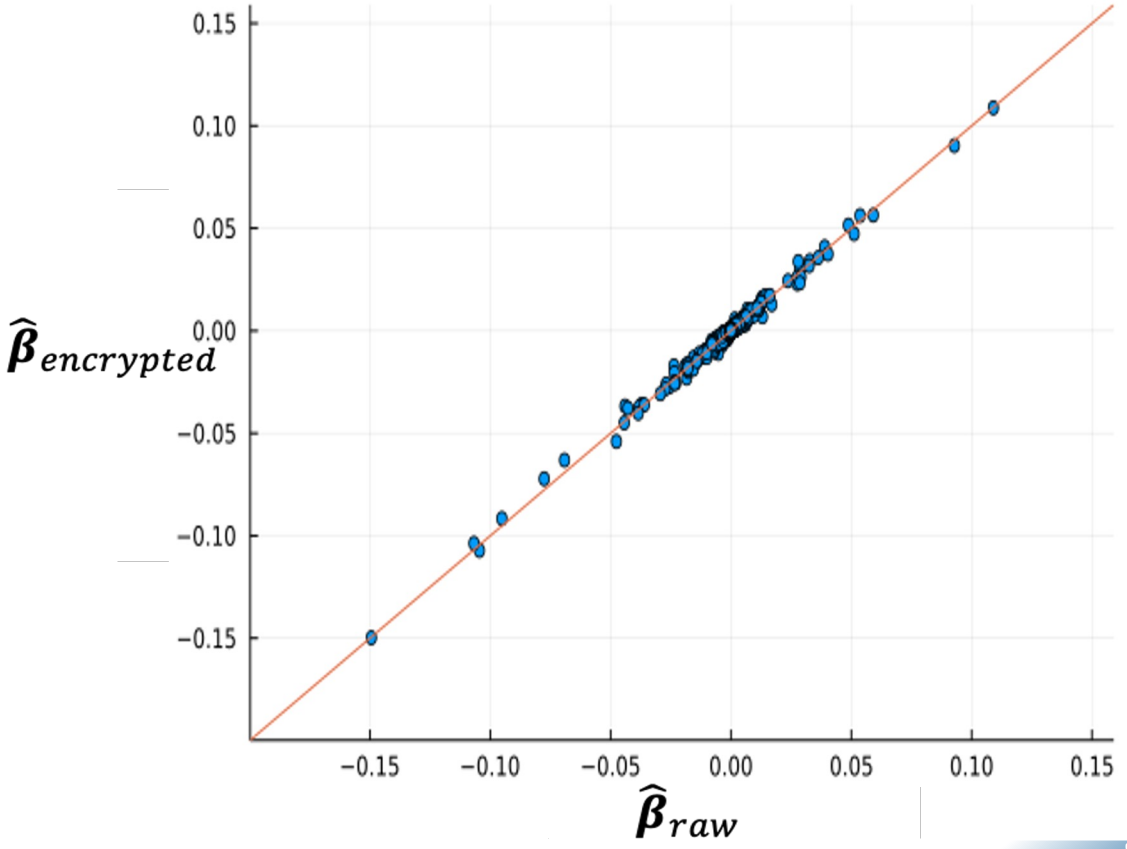


Results: ~ identical estimated marker effects

- $corr(\hat{\boldsymbol{\beta}}_{raw}, \hat{\boldsymbol{\beta}}_{encrypted}) \approx 0.9929$

(i.e., correlation between estimated marker effects using raw/encrypted data)

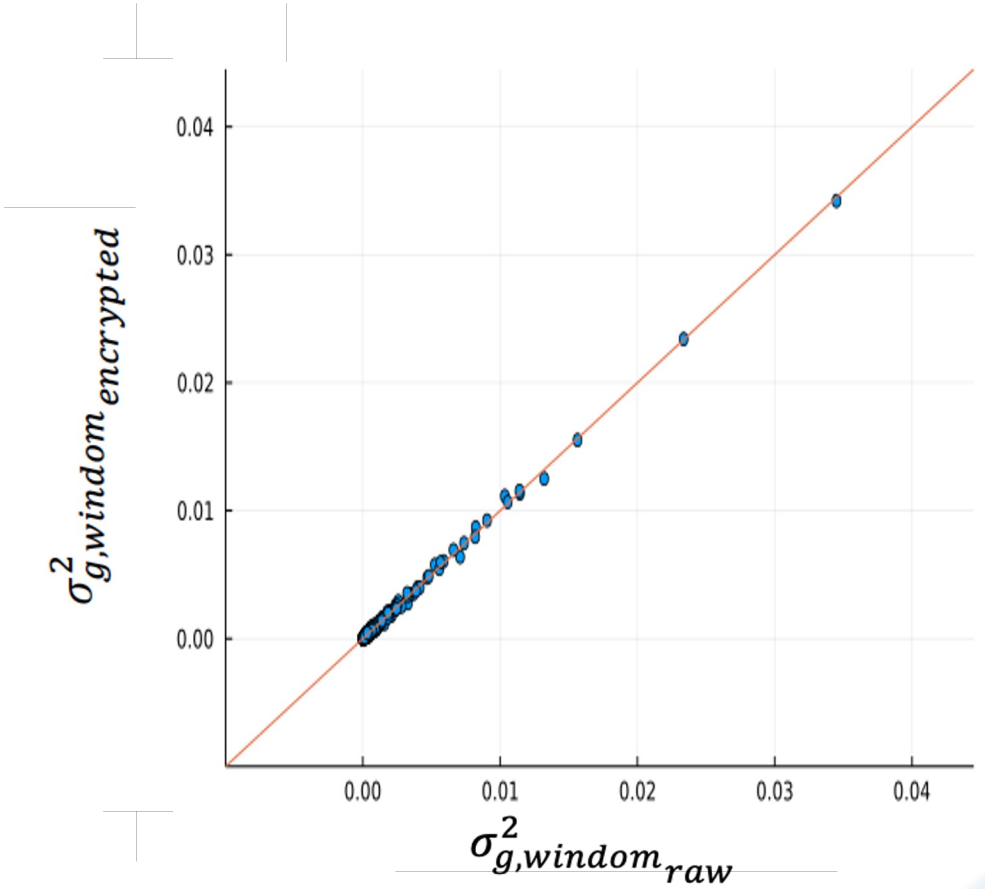
		h^2			
		0.1	0.3	0.5	0.7
QTL%	1%	0.9937	0.995	0.9957	0.9961
	10%	0.9928	0.9927	0.9926	0.9928
	50%	0.9898	0.9937	0.9916	0.9911
	100%	0.9906	0.9932	0.9925	0.992



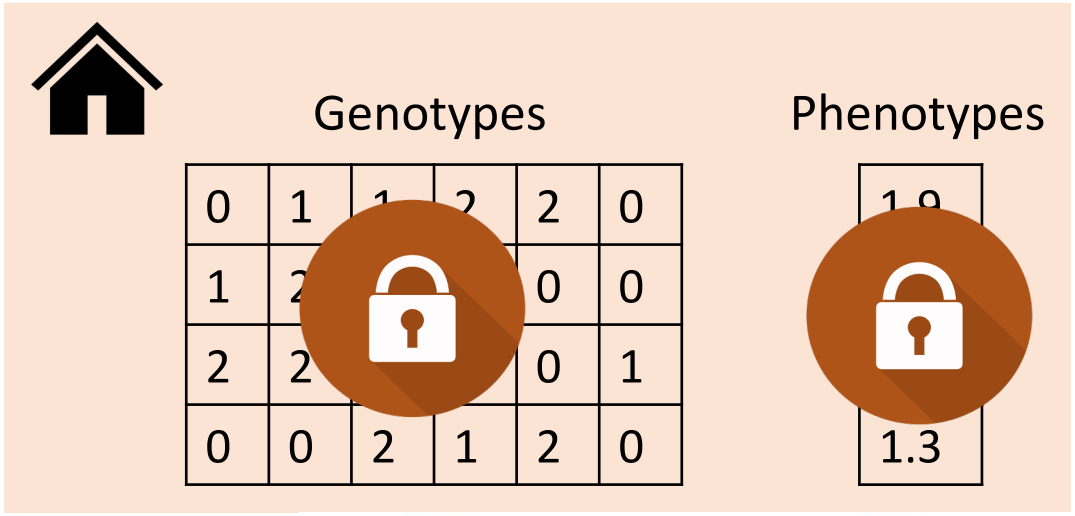
Results: ~ identical local genetic variance

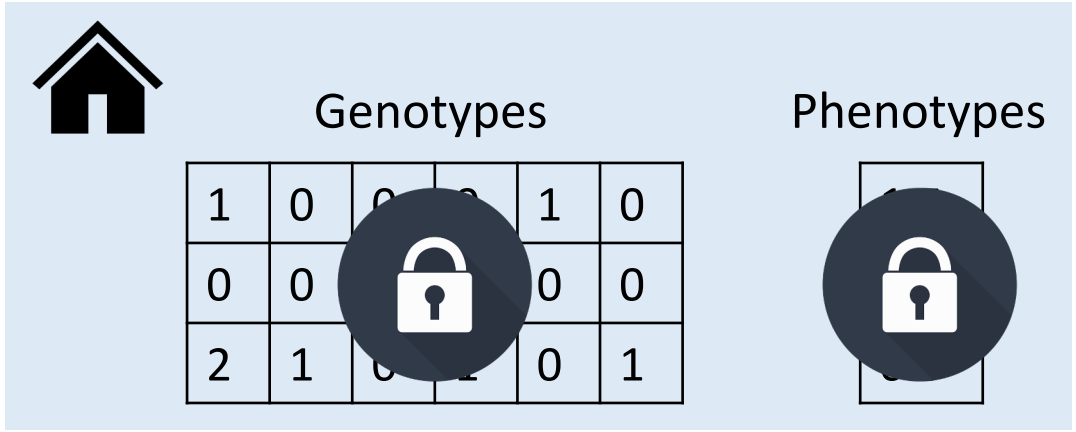
- 20 SNPs per-window, #window=2522
- $corr\left(\sigma^2_{g,windom_{raw}}, \sigma^2_{g,windom_{encrypted}}\right) \approx 0.9923$

		h^2			
		0.1	0.3	0.5	0.7
QTL%	1%	0.9954	0.995	0.9966	0.9968
	10%	0.9895	0.9923	0.9929	0.9943
	50%	0.9908	0.9946	0.9932	0.9924
	100%	0.9728	0.9934	0.9936	0.9933



Joint analysis using encrypted data from multiple contributors





P_1X_1 P_1y_1 P_2X_2 P_2y_2

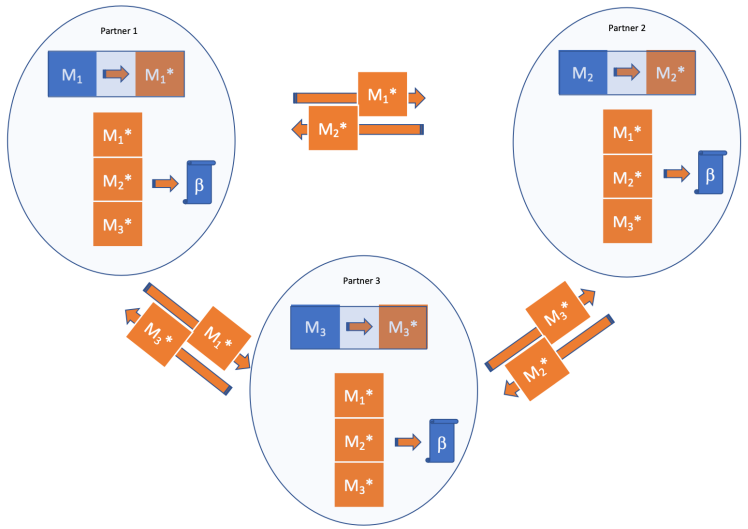
Joint analysis

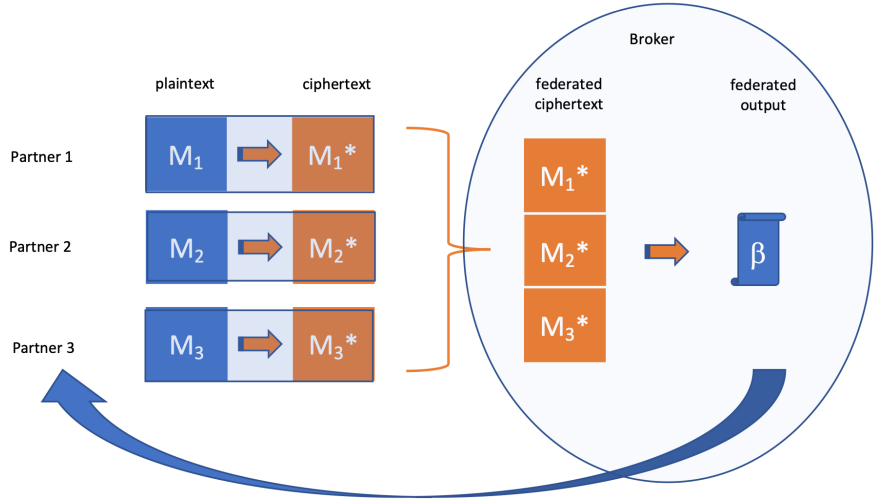
Allow each contributor to the joint data to use its own private key prior to sharing the data

Next Step

- Develop methods, protocols, software, and workflows to support data sharing using homomorphic encryption

- Work with data sharing stakeholders to facilitate implementation of homomorphic encryption





UCDAVIS Department of Animal Science Quantitative Genetics Lab

FAIR

Findable, Accessible, Interoperable, Reusable

Privacy concerns
Intellectual property
Trade secret

Homomorphic encryption

- Confidential information is protected/obscured
- Obtain same outcomes/results using the encrypted data (and the key)
- Joint analysis using encrypted data from multiple contributors

Questions?

USDA-NIFA-AFRI 2018-67015-27957
USDA-NIFA-AFRI 2021-67015-33412

AG2PI Agricultural Genome to Phenome Initiative Seed Grant