# Caring Without Sharing: GWAS in a Decentralized Setting

Armin Pourshafeie

Bustamante Lab

Joint work with:

Carlos D. Bustamante, Snehit Prabhu

# Running Decentralized GWAS

- Why and what else?
- Methods
- Simulations/experiments

# Introduction

- Goal: discover variants associated with a particular phenotype
- Discovering variants with small effect sizes requires large datasets
  - Data sharing can help
- Centralizing data is difficult (Hardware, policy, etc.)

# Meta-studies

- Combine the results of previous studies on the same phenotype

Find compatible studies to combine

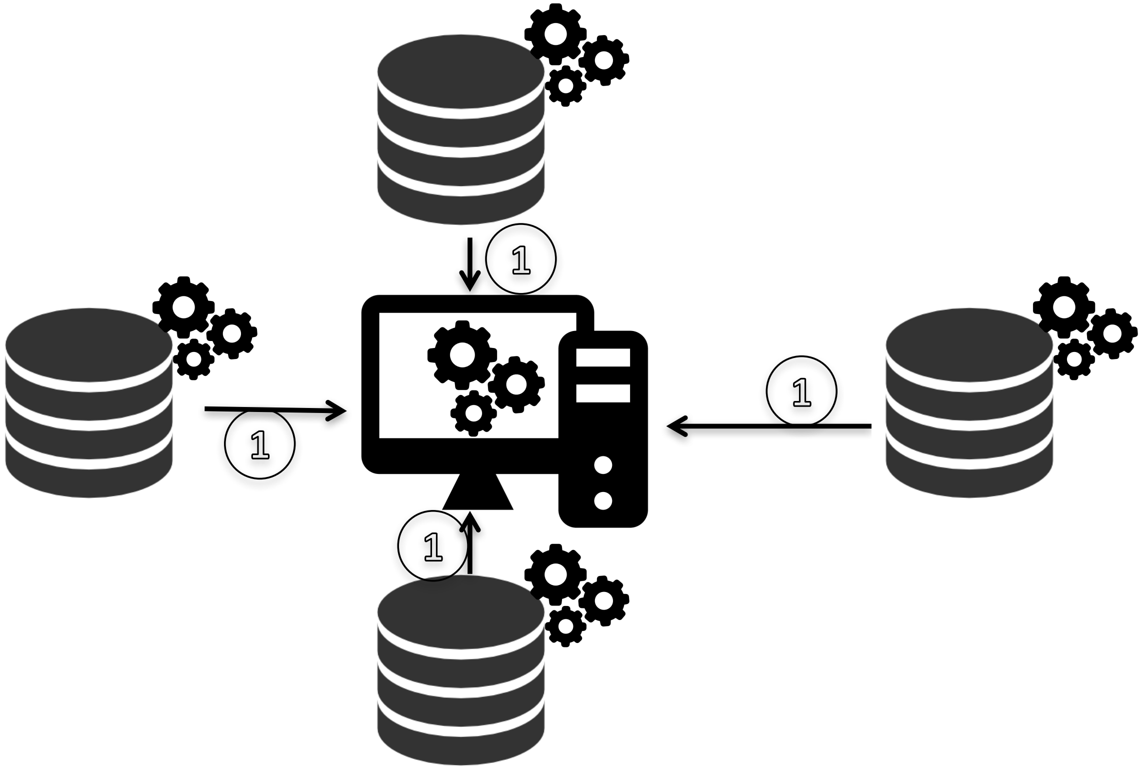
Choose a model (fixed effect vs. random effect) and combine the estimates

# Meta-studies

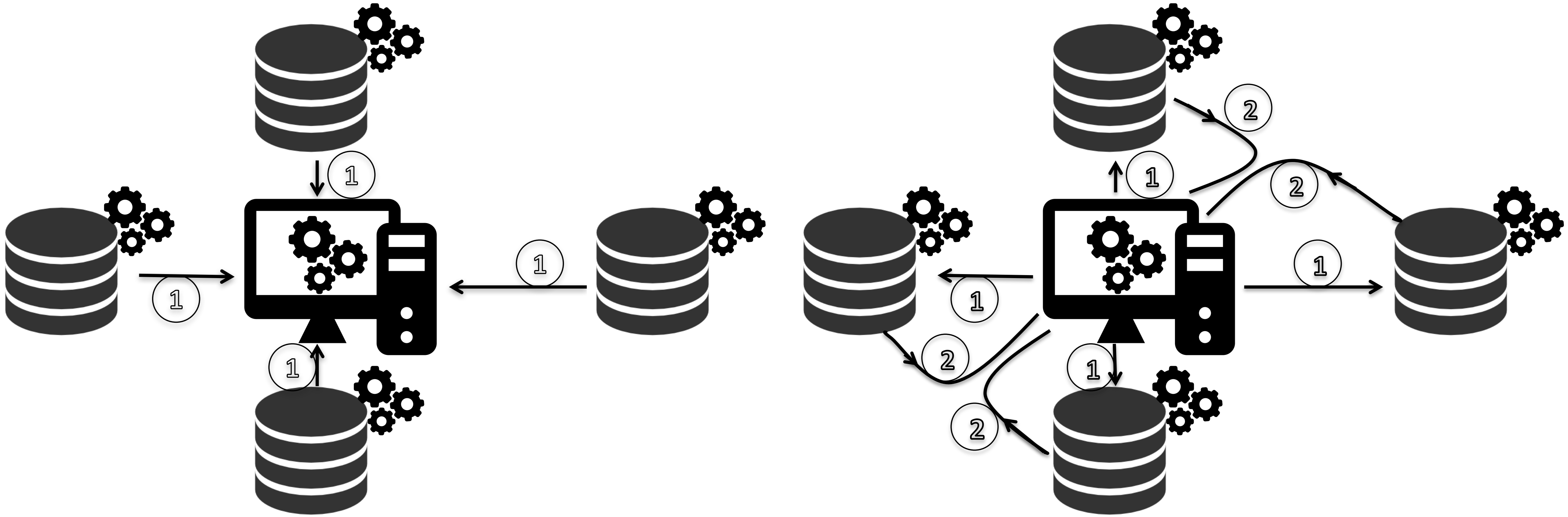
Find compatible studies to combine → Choose a model (fixed effect vs. random effect) and combine the estimates

- Pros:
  - Familiar
  - Readily available data
  - Computationally efficiency
  - Asymptotically statistical efficiency (Lin and Zeng 2010) (asymptotic in each study)
  - Some level of privacy
- Cons:
  - Unable to use small datasets
  - Difficult/non-existent quality control
  - Multiple regression is not possible
  - Difficult to control for structure/duplicates

# Different paradigms

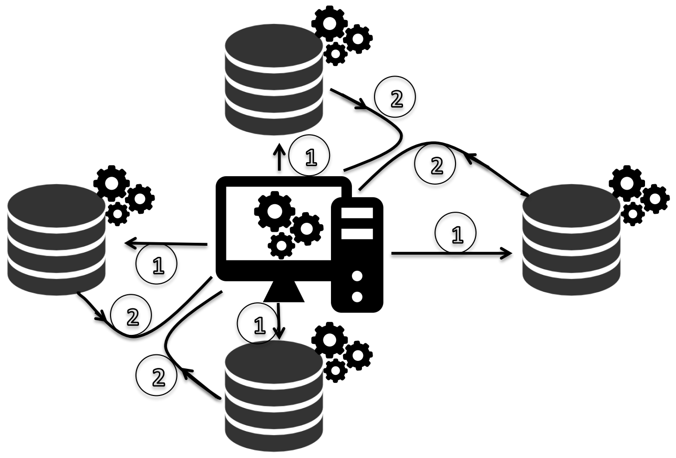


# Different paradigms

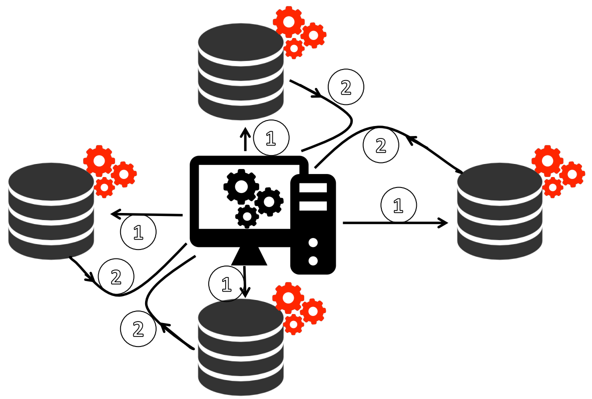


# Decentralized GWAS

- Quality control
- Population structure control (PCA)
- Imputation
- Association (logistic regression)



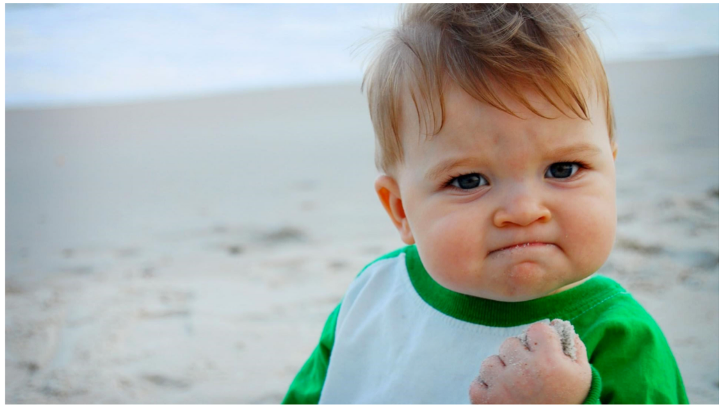
# QC




- No communications (almost):
  - Calling quality, missing per individual
- Few communications
  - #(Missing, Homo-ref, Hetro, Homo-alt)
    - Missing-per loci, Allele Freq, Hardy-Weinberg
  - Relatedness:
    - Hashing (Dan He, et al. 2014)
- LD-pruning.
  - Hard :(
  - Pass in a matrix after thinning (very local pruning)

# Population Structure Control (PCA)

- Easy:
  - Project everyone on dimensions discovered from a public dataset (1KG, Hapmap, etc.)
    - No need for LD pruning
    - Cheap, and fast



  - Biased, not applicable to underrepresented populations

# Population Structure Control (PCA)

- Hard:
  - ind-ind covariance matrix won't work

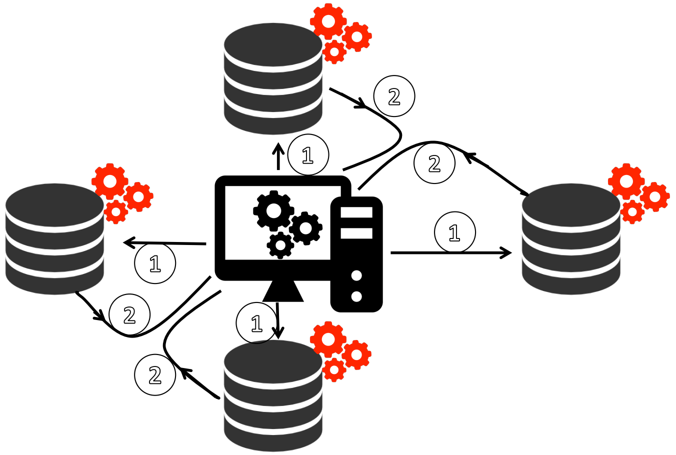
- $G_1 = \begin{bmatrix} ind_{1,1} \\ \vdots \\ ind_{1,N_1} \end{bmatrix}$, $G_2 = \begin{bmatrix} ind_{2,1} \\ \vdots \\ ind_{2,N_2} \end{bmatrix}$, … , $G_k = \begin{bmatrix} ind_{k,1} \\ \vdots \\ ind_{k,N_k} \end{bmatrix}$

Missing

  - $G^T G = \begin{bmatrix} block1 & & \\ & block2 & \\ & & block3 \end{bmatrix}$

# Population Structure Control (PCA)

- Use the LD-matrix (gene-gene) instead
  - Compute Gene-gene covariance matrix. All the genotypes of each individual is in a single dataset. The overall LD-matrix is simply the sum of these LD-matrices
- Pseudo-algorithm
  1. Compute the local LD-matrix
  2. Average the local LD-matrices at the center
  3. Perform eigen-decomposition
  4. Back solve for loadings at each silo



# Population Structure Control (PCA)

- Pros:
  - This is impossible to do in meta studies
  - Can implement with differential privacy
- Cons:
  - The LD-matrix is very large
    - This method is inefficient with many small size silos

# Association

- Notation:
  - $.^{(k)} := k^{th}$ silo
  - $\ell^{(k)}(\beta)$ := -log-likelihood function evaluated on silo $k$ with parameter $\beta$

Centralized

$$\hat{\beta} = \text{argmin}_\theta \textstyle\sum_k \ell^{(k)}(\theta)$$

Meta-study (FE)

$$z^{(k)} = \text{argmin } \ell^{(k)}(x)$$
$$\hat{\beta} = \sum w^{(k)} z^{(k)}$$
$$\text{Assumption: } z^{(k)} = \beta + \varepsilon^{(k)}$$

$$\hat{\beta} = \text{argmin}_\theta \textstyle\sum_k \min_x \ell^{(k)}\left(x^{(k)}\right) \text{ s.t. } x^{(k)} = \theta$$

$$L_\rho(\theta, \lambda, x) = \sum_k \ell^{(k)}\left(x^{(k)}\right) + \lambda^T\left(x^{(k)} - \theta\right) + \frac{\rho}{2}\left\|x^{(k)} - \theta\right\|_2^2$$

Lagrange Multiplier

Augmented Lagrangian
(Hestenes, Powell 1969)

# Association

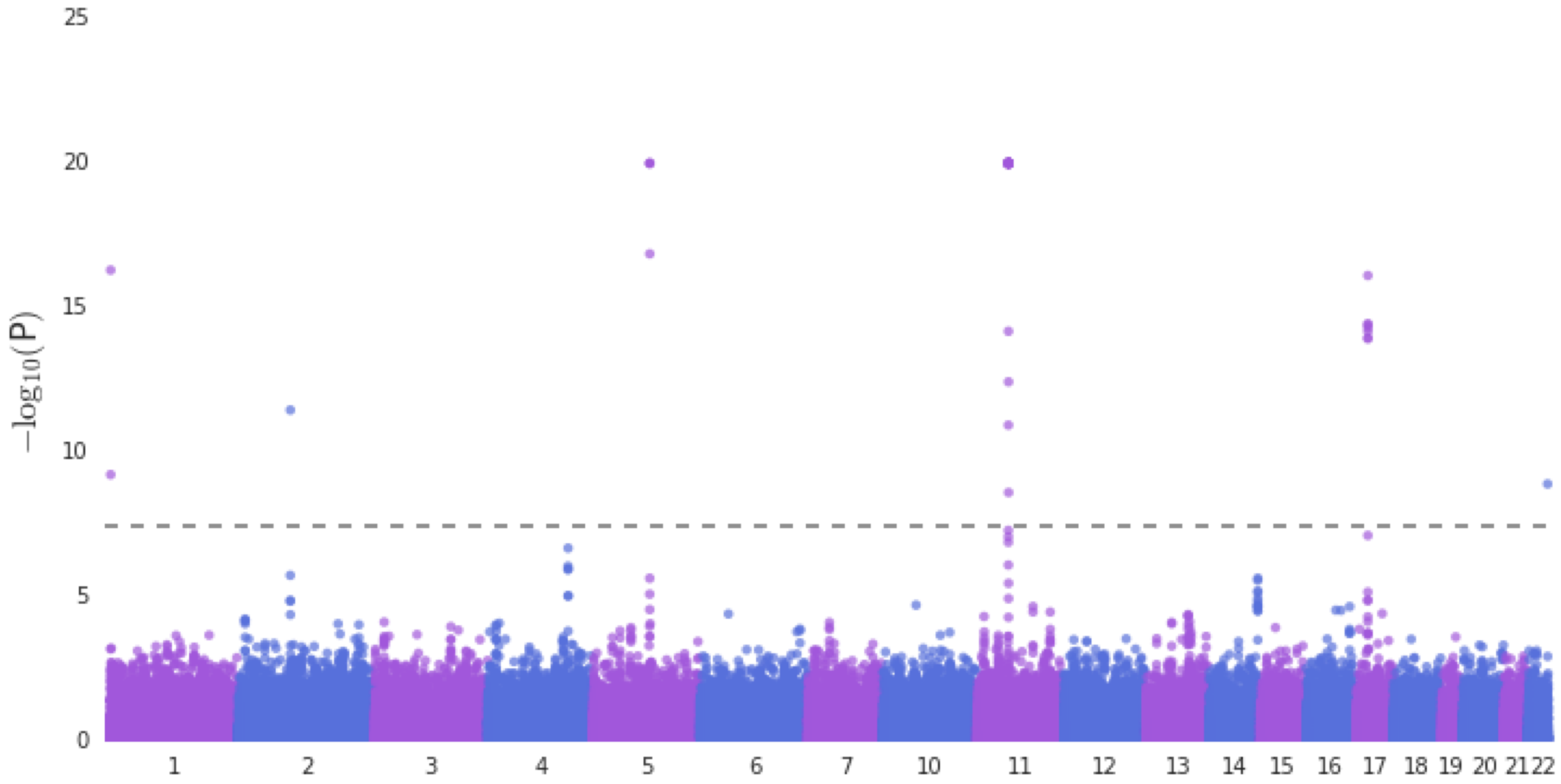
- $L_\rho(\theta, \lambda, x) = \sum_k \ell^{(k)}\left(x^{(k)}\right) + \lambda^T\left(x^{(k)} - \theta\right) + \frac{\rho}{2}\left\|x^{(k)} - \theta\right\|_2^2$
- Updates:
  - $z^{(k)} \leftarrow \mathrm{argmin}_x \ell^{(k)}\left(x^{(k)}\right) + \frac{\rho}{2}\left\|x^{(k)} - \theta + \lambda^{(k)}\right\|_2^2$ At each silo
  - $\theta \leftarrow \frac{1}{K}\sum_k z^{(k)}$ At the center
  - $\lambda \leftarrow \lambda^{(k)} + z^{(k)} - \theta$

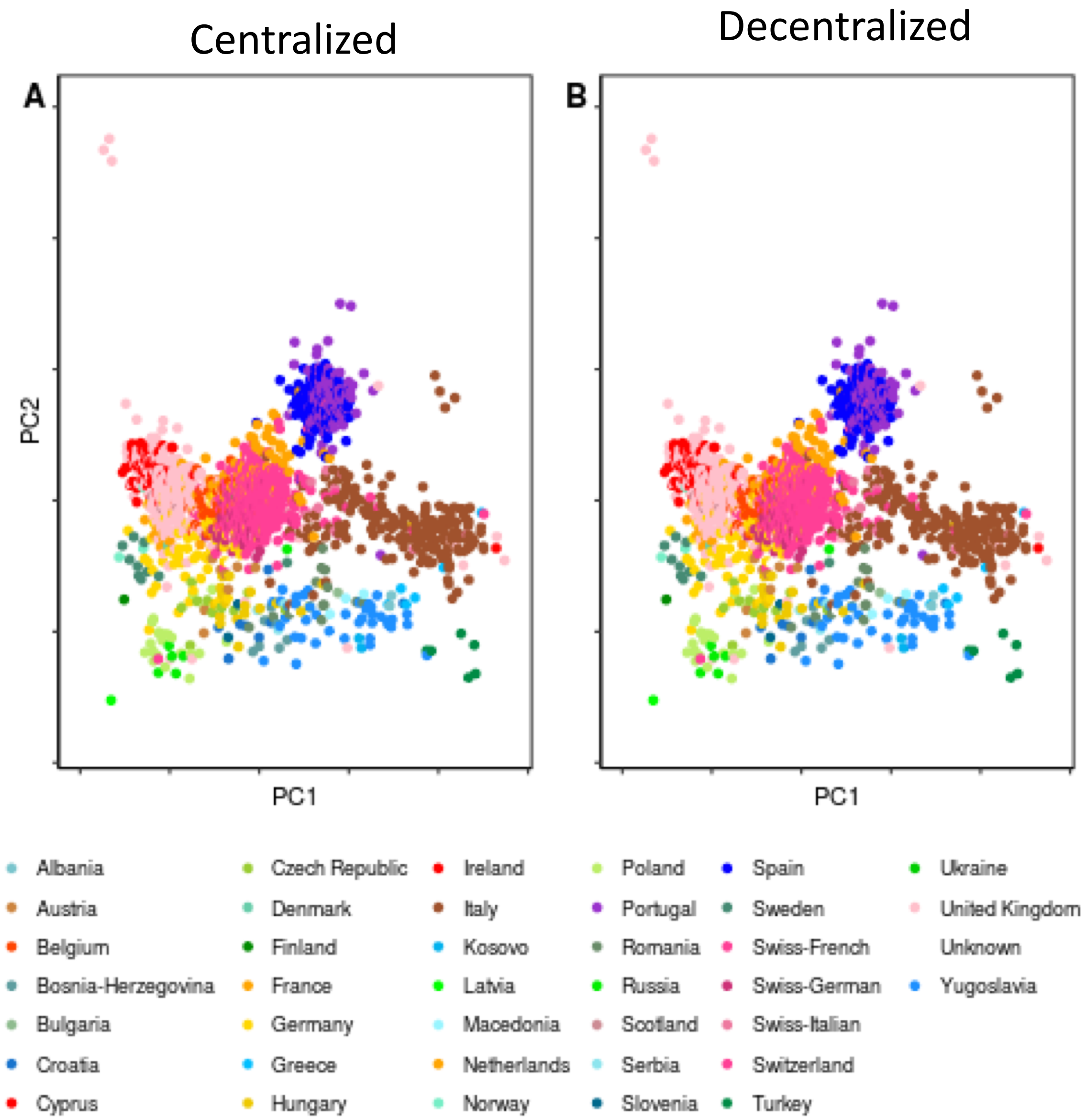
See Boyd, Stephen, et al. "Distributed optimization and statistical learning via the alternating direction method of multipliers." *Foundations and Trends® in Machine Learning* 3.1 (2011): 1-122.

# Results

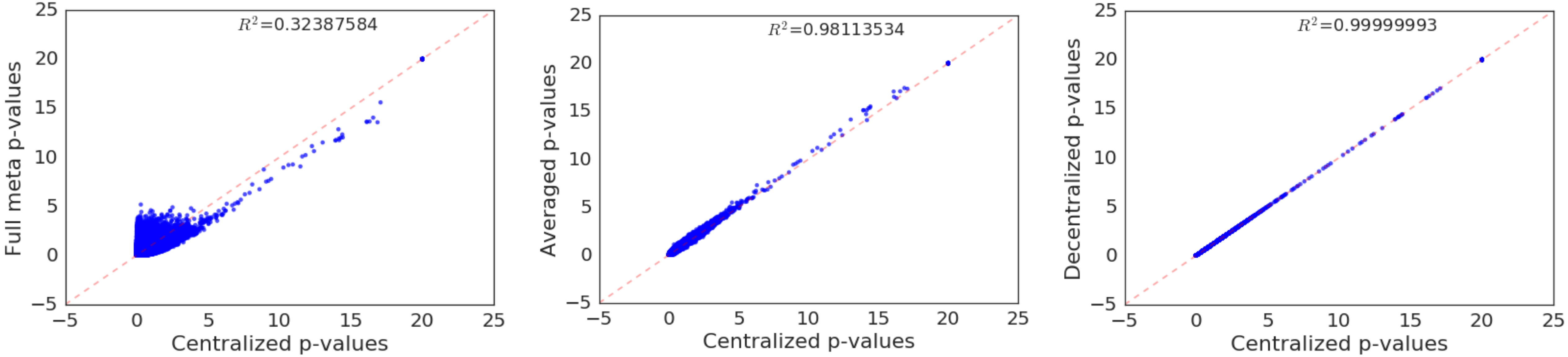
- Simulated GWAS on POPRES[1]
  - 2274 ind ~ 400k Loci
  - Simulated a case-control phenotype according to a logistic model
  - 50-50
  - 10 causal SNPS with effect size drawn from a gaussian + noise
- Two experiments:
  - 5 silos, random distribution ($n \approx 450$ per silo)
  - 2 silos, cases vs controls
- All regressions include 1 SNP + 5 PCs

1. Nelson, Matthew R., et al. *AJHG* (2008):

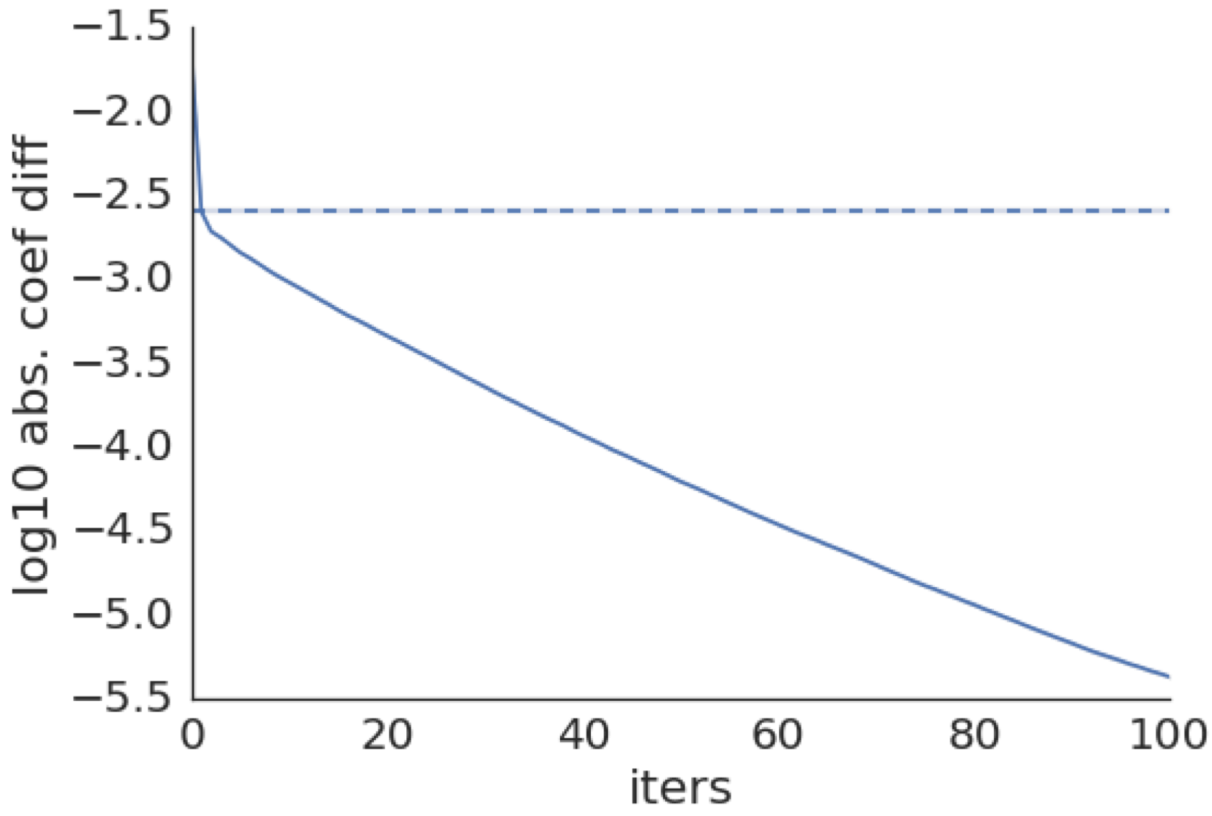

25
20
15
10
5
0
$-\log_{10}(P)$
1
2
3
4
5
6
7
10
11
12
13
14
15
16
17
18
19
20
21
22

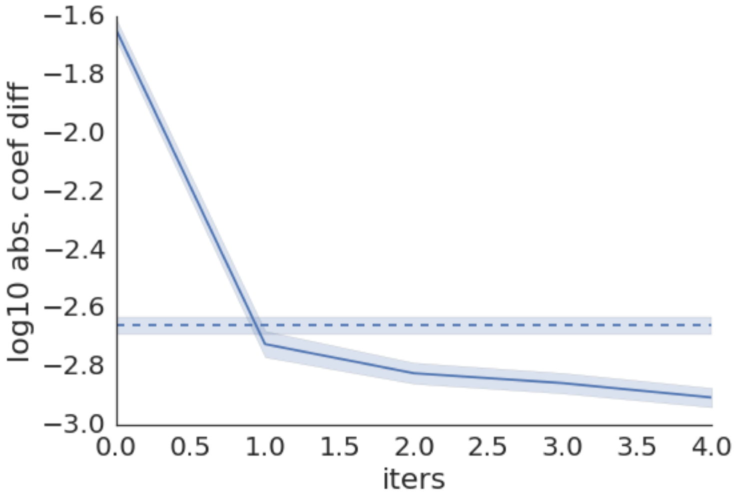

Centralized
Decentralized
A
B
PC2
PC1
PC1
Albania
Austria
Belgium
Bosnia-Herzegovina
Bulgaria
Croatia
Cyprus
Czech Republic
Denmark
Finland
France
Germany
Greece
Hungary
Ireland
Italy
Kosovo
Latvia
Macedonia
Netherlands
Norway
Poland
Portugal
Romania
Russia
Scotland
Serbia
Slovenia
Spain
Sweden
Swiss-French
Swiss-German
Swiss-Italian
Switzerland
Turkey
Ukraine
United Kingdom
Unknown
Yugoslavia

# Experiment 1: (iid distributed individuals, 5 Silos)

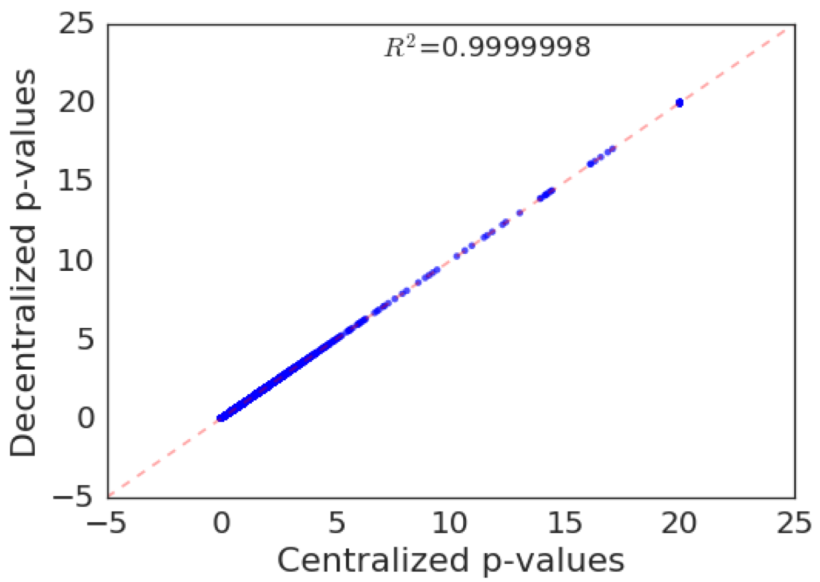


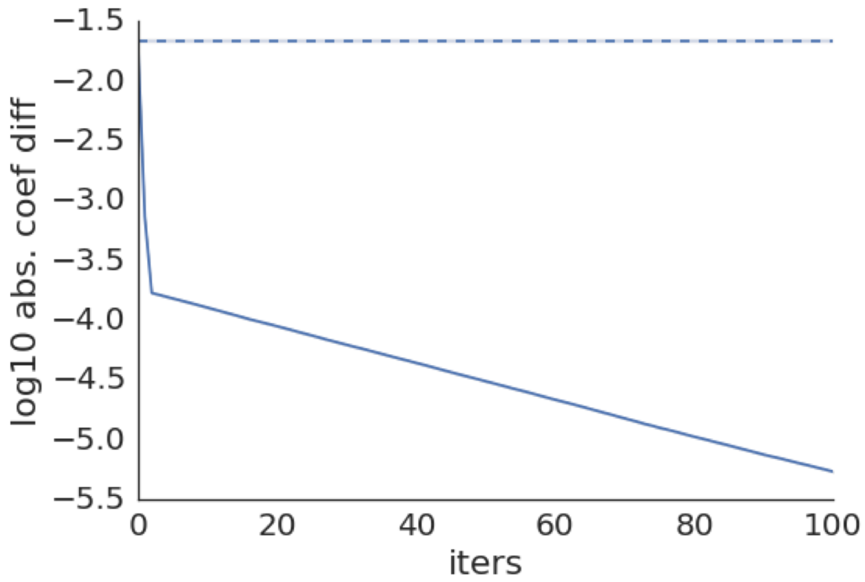
# Experiment 1: (iid distributed individuals, 5 Silos)



# Experiment 2: (Cases vs. Controls)



# Acknowledgments

- Collaborators
  - Carlos Bustamante
  - Snehit Prabhu

- Funding:
  - NHGRI SGTP

- **Thank you!**

# Questions?


https://imgs.xkcd.com/comics/efficiency.png