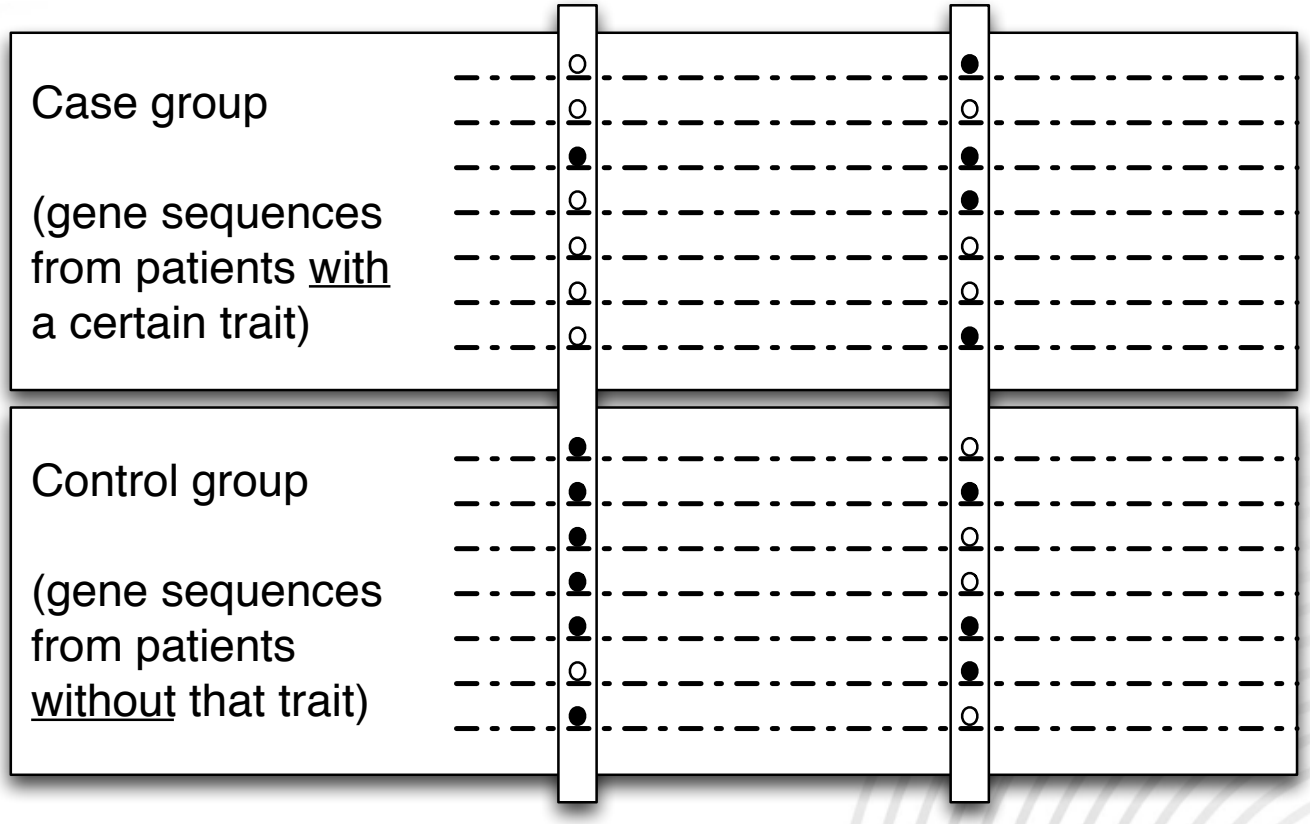
Applying MPC for Genotype Analysis While Considering Population Stratification

Andre Ostrak, Jaak Randmets, Ville Sokk, Sven Laur, Liina Kamm

Genome Wide Association Studies (GWAS)

- Genotype data presented as single nucleotide polymorphisms (SNPs)
- Phenotype data
- Case and control groups



CYBERNETICA

Data Quantity and Privacy

- Traits can be affected by multiple genetic locations
- Large volumes of heterogenous data are needed
- Privacy becomes an issue when sharing data between biobanks

CYBERNETICA

Population Stratification

- Geographic isolation of subpopulations during several generations
- Confounded associations
- The lactase gene (LCT) was shown to be connected to height in a European American cohort with significance $p < 10^{-6}$
- Both height and the LCT gene have wide variations across populations in Europe
- The spurious association was reduced, when individuals were rematched on the basis on European ancestry

Privacy-Preserving GWAS

- Extract-transform-load (ETL) and contingency table computation
- Hypothesis testing
- Correction for stratification

- Which algorithms to choose?

Correcting for Stratification Using PCA

- **Algorithm 1: Principal component analysis (PCA)**
  - Top eigenvectors of the sample kinship matrix
  - Eigendecomposition is used to infer population stratification
  - Its results are used to adjust the genotypes and phenotypes for stratification
  - Cochran-Armitage test for trend used on the adjusted results
- **Algorithm 2: FastPCA**
  - Uses recent advances in random matrix theory to reduce the computational effort in finding the top eigenvectors of the kinship matrix

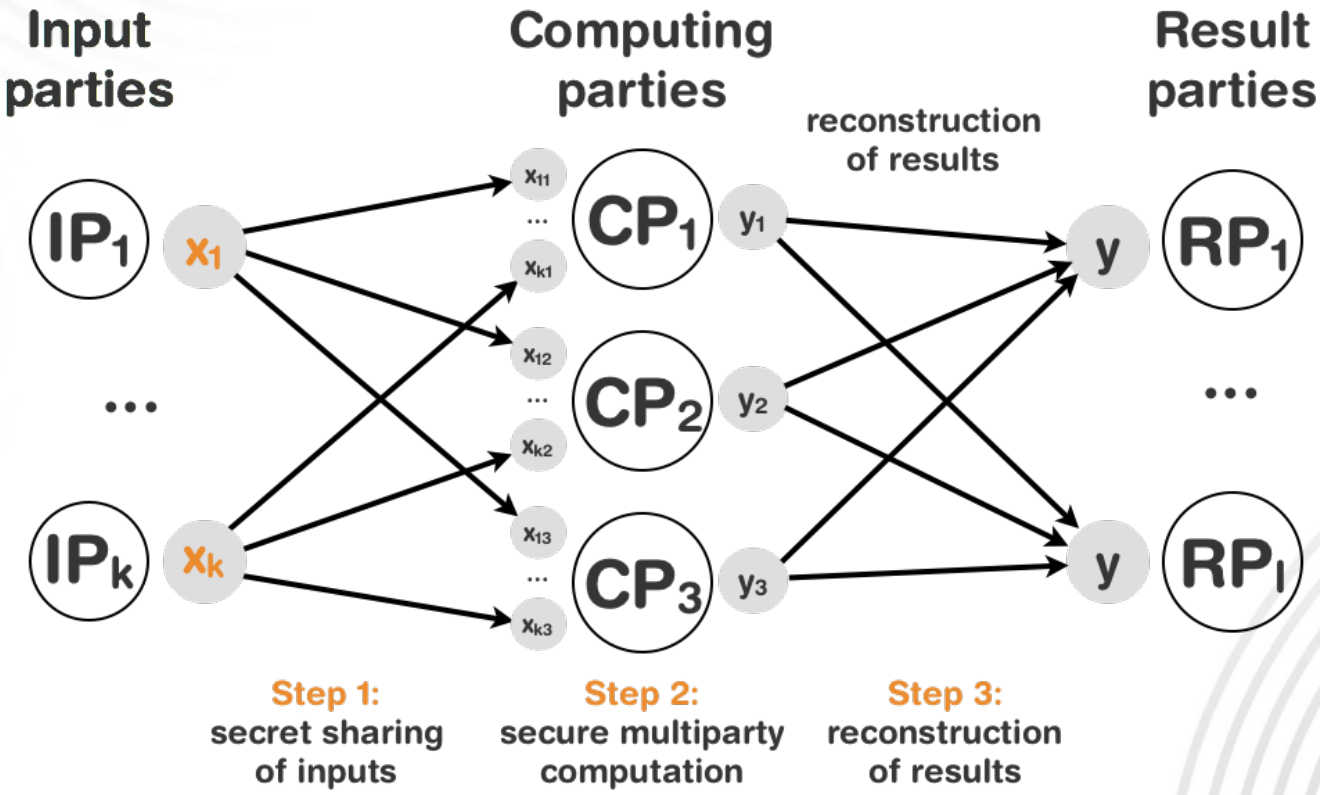
Correcting for Stratification

- **Algorithm 3: Genomic control**
  - Cochran-Armitage trend statistic for each SNP
  - Robust estimation for the variance inflation factor (median of trend statistics)
  - Divide the trend statistics with the estimation
- **Algorithm 4: EMMAX**
  - Linear mixed effect model for each SNP with the SNP as the fixed effect
  - Approximate the random coefficients giving the maximum likelihood estimates for the variance component factors
  - Find estimates for the regression coefficients
  - Compute the *t* statistics

What do we need?

- Database operations (oblivious join, data aggregation)
- Floating-point data type and operations (actually fixed-point is sufficient, with some adjustment)
  - For floating-point numbers, addition is especially slow
- Very many parallel matrix multiplications
- Natural logarithm

Sharemind shared3p



CYBERNETICA

Sharemind MPC

	Sharemind MPC
Secure storage	additive secret sharing of each individual value between three computing parties
Secure computing	Honest-but-curious MPC
Algorithm implementation language	SecreC 2

- Infrastructure:
  - 3 computers with Intel Xeon E5-2640 processors
  - 128 GB of memory
  - dedicated 10 Gb/s connections

CYBERNETICA

PCA (seconds)

Subtask	1500 SNPs 200 donors	5000 SNPs 200 donors	2000 SNPs 800 donors
Table preparation	65 s	210.6 s	347 s
Float to fixed	2.4 s	7.8 s	
GSPCA	182.8 s	183.3 s	2606.0 s
Stratification control	1365.9 s	4520.2 s	22112.0 s
Test statistics	136.2 s	436.6 s	675.4 s
SUM	**1752.4 s**	**5358.4 s**	**25740.4 s (7.15 hours)**

Genomic Control (seconds)

Subtask	300 000 SNPs 220 donors	300 000 SNPs 440 donors	300 000 SNPs 660 donors
Table preparation	2060.3 s	3992.9 s	5960.8 s
Cochran-Armitage	19.3 s	19 s	18.7 s
Stratification control	17.9 s	16.5 s	20.5 s
SUM	**2096.5 s**	**4028.4 s**	**6000.0 s**

EMMAX (seconds)

Subtask	500 SNPs 200 donors	1000 SNPs 200 donors	5000 SNPs 800 donors
Table preparation	21.9 s	22.6 s	221 s
Kinship matrix	1770.5 s	1781.1 s	17775.1 s
GSPCA	14210.7 s	14389.8 s	14480.6 s
Maximum likelihood	5557.0 s	5596.0 s	54923.1 s
Test statistics	0.3 s	0.2 s	2.3 s
SUM	**21560 s (~6 hours)**	**21789.5 s (~6 hours)**	**87400.2 s (~24.3 hours)**

Sharemind Hardware Isolation (HI)

	Sharemind MPC	Sharemind HI
Secure storage	additive secret sharing of each individual value between three computing parties	Full database encrypted with AES
Secure computing	Honest-but-curious MPC	Intel® Software Guard eXtensions (SGX) Trusted Execution Environment
Algorithm implementation language	SecreC 2	C++ with Sharemind HI SDK that enable large database processing and access control in enclaves

Performance results (seconds)

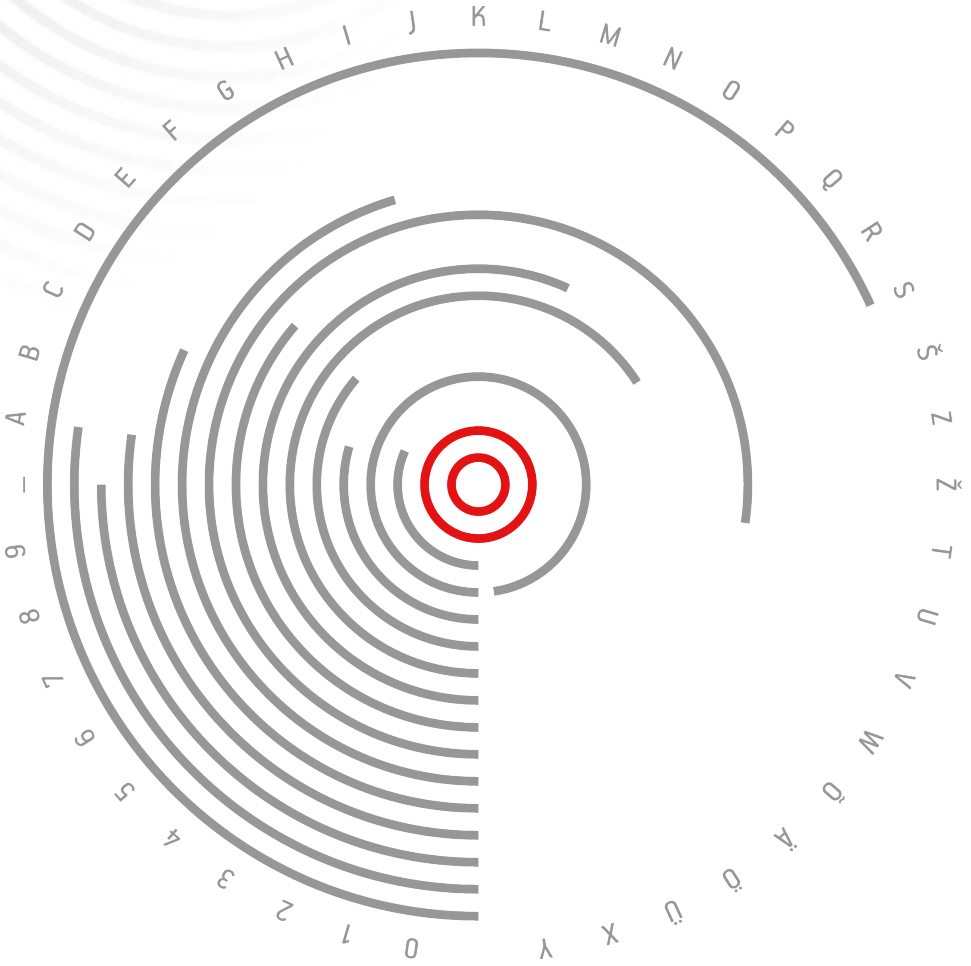
Experiment	Patient count	MPC	HI	If HI was 100x slower
PCA	5000 SNPs 200 donors	5358.43 s (89.3 minutes)	2.42 s	242 s
Genomic Control	300000 SNPs 200 donors	2096.5 s (35 minutes)	15.11 s (reading input data)	~20 s (reading input data)
EMMAX	5000 SNPs 200 donors	87400 s (24.3 hours)	7.14 s	714 s

Conclusions

- It is possible to perform the whole GWAS process in the privacy-preserving domain
- Optimisations from FastPCA
- Feasibility depends on the requirements of the study.

CYBERNETICA


CYBERNETICA


A B C D E F G H I J K L M N O P Q R S Š Z Ž T U V W Ö Ä Ü X Y 0 1 2 3 4 5 6 7 8 9 _