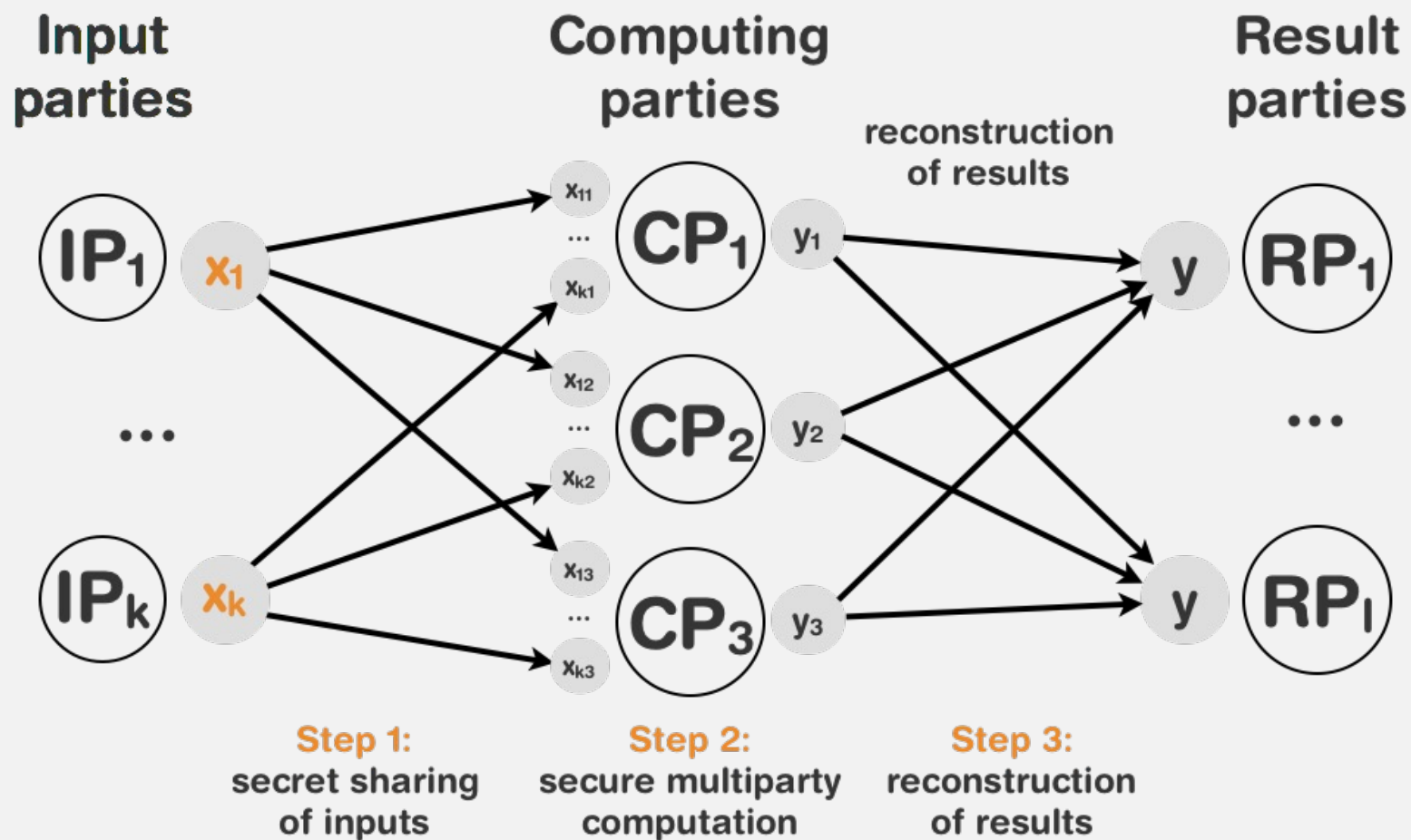


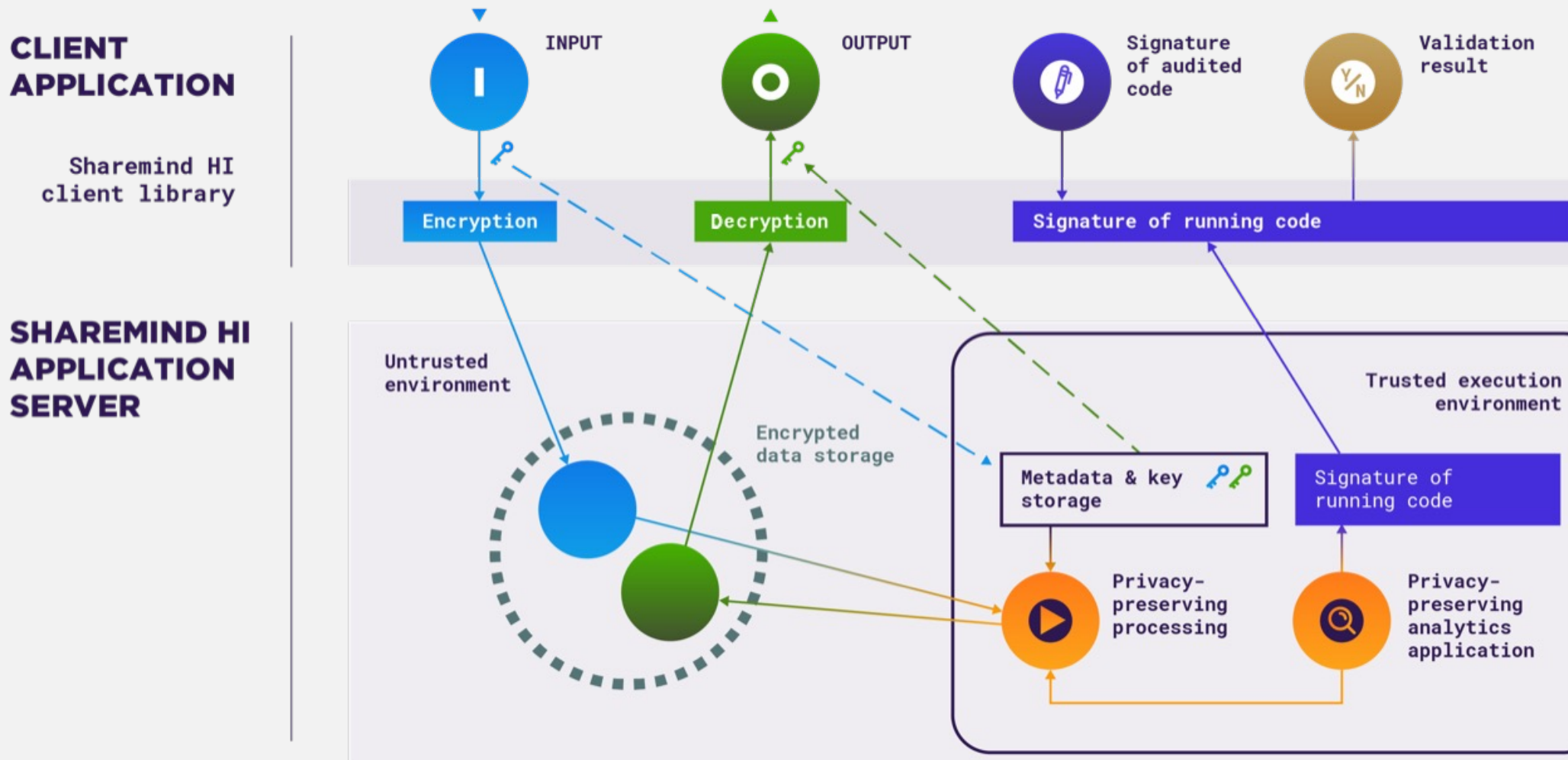
# Privacy-Preserving Precision Medicine

Liina Kamm  
Senior researcher @ Cybernetica

# Secure Multy-party Computation – Sharemind MPC



# Trusted Execution Environment – Sharemind HI



# Implementing Privacy-Preserving Genotype Analysis with Consideration for Population Stratification

Andre Ostrak, Jaak Randmets, Ville Sokk, Sven Laur, Liina Kamm  
MDPI Cryptography (Special Issue Secure Multiparty Computation)

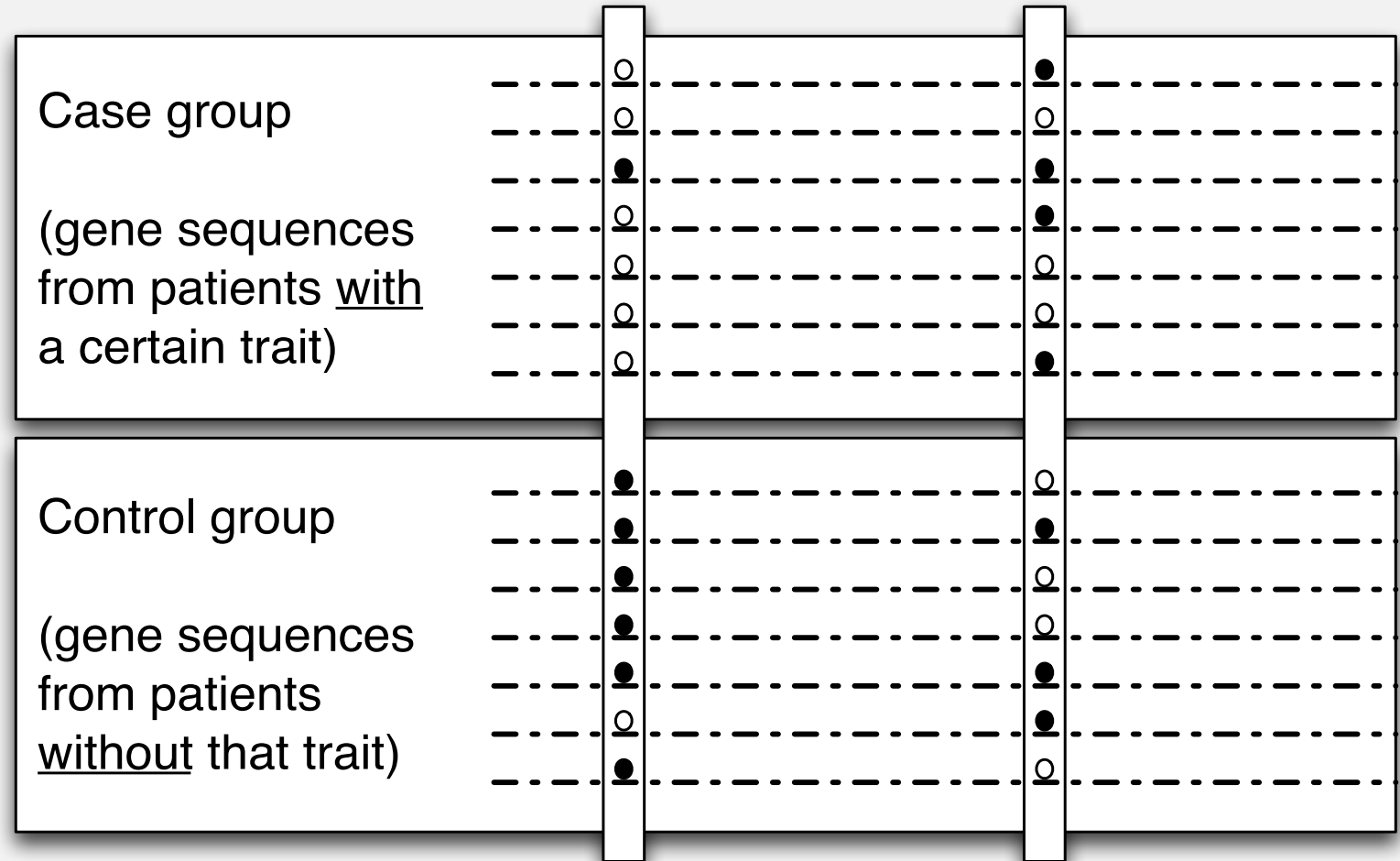
<https://www.mdpi.com/2410-387X/5/3/21>

This work has been supported by the EU H2020-SU-ICT-03-2018 Project No. 830929  
CyberSec4Europe ([cybersec4europe.eu](http://cybersec4europe.eu)).



# Genome Wide Association Studies (GWAS)

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



# 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

# Population Stratification

- Geographic isolation of subpopulations during several generations
- 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 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

# 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.





# Thank you!

[liina.kamm@cyber.ee](mailto:liina.kamm@cyber.ee)

 [cybernetica](#)

 [CyberneticaAS](#)

 [cybernetica\\_ee](#)

 [Cybernetica](#)