Secure Gene Mining

Liina Kamm

STACC, University of Tartu

STACC Software Technology and Applications Competence Center

TARTU ÜLIKOOL 1632 UNIVERSITAS TARTUENSIS

Estonian Computer Science Theory Days
Nelijärve 2011

Overview

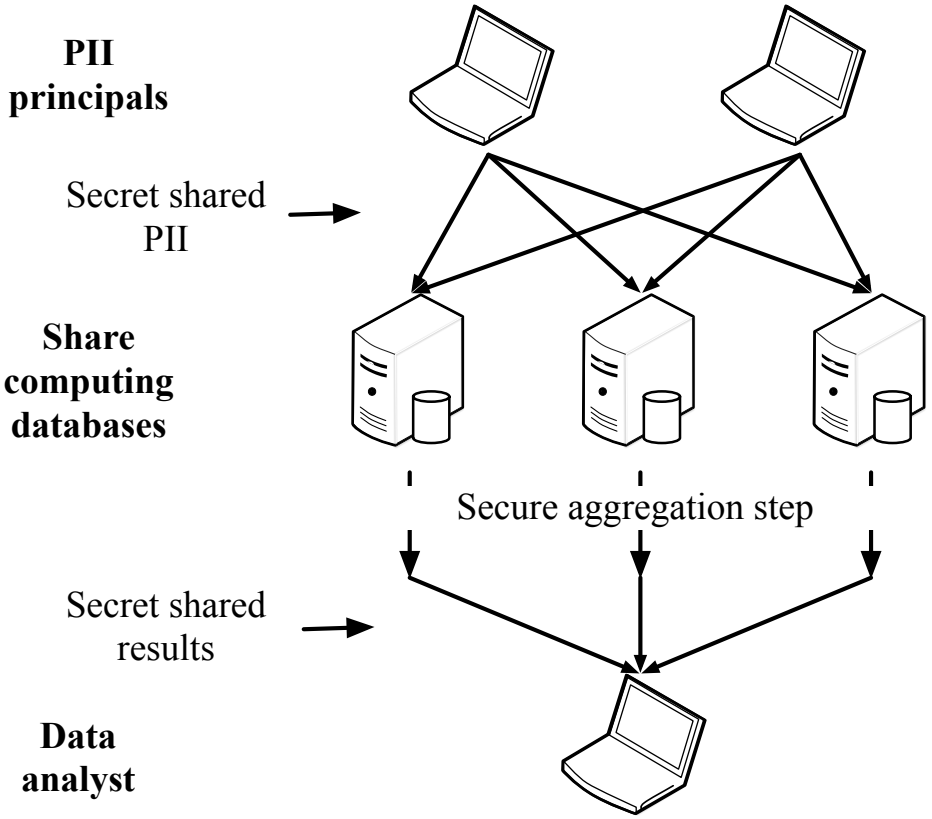
- Purpose
- Secure multi-party computation
- Data
- Experiments
- Results

Purpose

- Combine data from different biobanks
- Data integration and analysis
- Sensitive data

Secure Multi-Party Computation

- Data entry
- Data sharing
- Data aggregation
- Published results



STACC Software Technology and Applications Competence Center

Example Scenarios

- Three different biobanks want to share data for analysis
- NIH wants to share data to analysts
- 23andMe project

Data Analysis

- Genome-wide association studies
- Cases and controls
- Hypothesis testing

Data Description

- Affymetrix Mapping 500K Array Set on the 270 samples typed by the International HAPMAP project

	NA06985	NA06991	NA06993	NA06994	…
SNP_A-1780270	BB	AB	AA	AA	…
SNP_A-1780272	AB	AB	AB	BB	…
SNP_A-1780285	BB	BB	BB	BB	…

Data Structure (1)

- We have:

	D1	D2	D3
SNP1	BB	AB	AA
SNP2	AB	AB	NN

- We need:
  - Count of alleles A and B in the case group
  - Count of alleles A and B in the control group

Data Structure (2)

- We had:

	D1	D2	D3
SNP1	BB	AB	AA
SNP2	AB	AB	NN

- We get:

	D1	D2	D3
SNP1_A	0	1	2
SNP1_B	2	1	0
SNP2_A	1	1	0
SNP2_B	1	1	0

Compulsory Formula Slide

- Consider the following allele counts:

	Cases	Controls
Allele 1	a	b
Allele 2	c	d

- We compute the standard χ^2 test statistic of independence based on these observed allele counts

$$T_1 = \frac{(ad - bc)^2(a + b + c + d)}{[(a + b)(c + d)(a + c)(b + d)]}$$

Data Privacy

- All the computations with gene data are secure multi-party computations
- Analyst provides the significance threshold
- Analyst receives a boolean value stating whether the SNP was significant or not

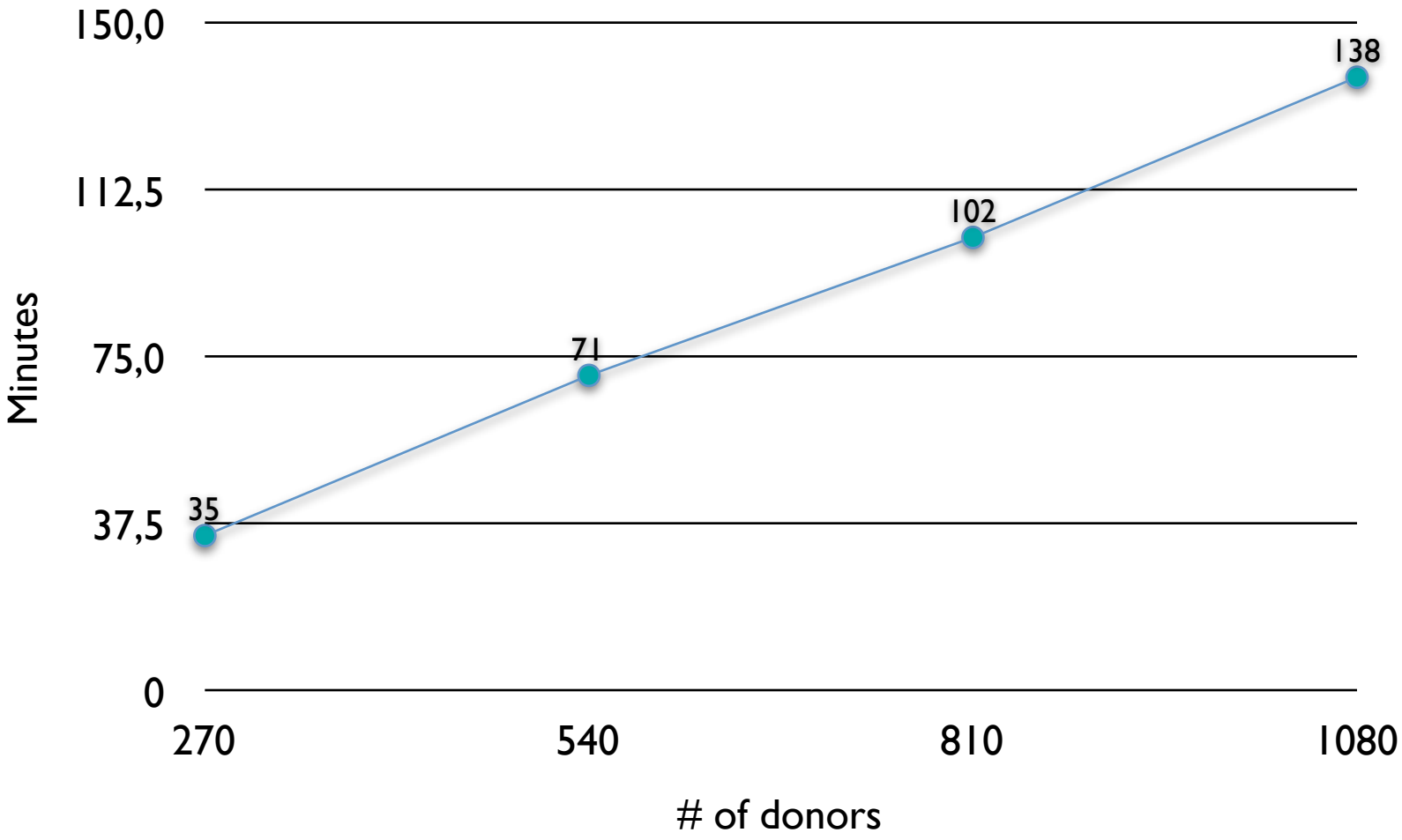
Tools

- Sharemind VM
- SecreC language
- Tokyo Cabinet DB engine
- Sharemind cluster
  - 48 GB RAM (less was used)
  - 12 cores (2 were used)
  - LAN network connection (current bottleneck)

Experiments

- Initial experiment:
  - 270 donors
  - 262 264 SNPs
- Additional experiments:
  - 540 donors
  - 810 donors
  - 1080 donors

Results



STACC Software Technology and Applications Competence Center

Implementation Experience

- Database issues
  - Database size
  - Querying columns vs. rows
- Optimising code
  - Manual vectorisation
- Running the experiments
  - Tuning the network layer configuration