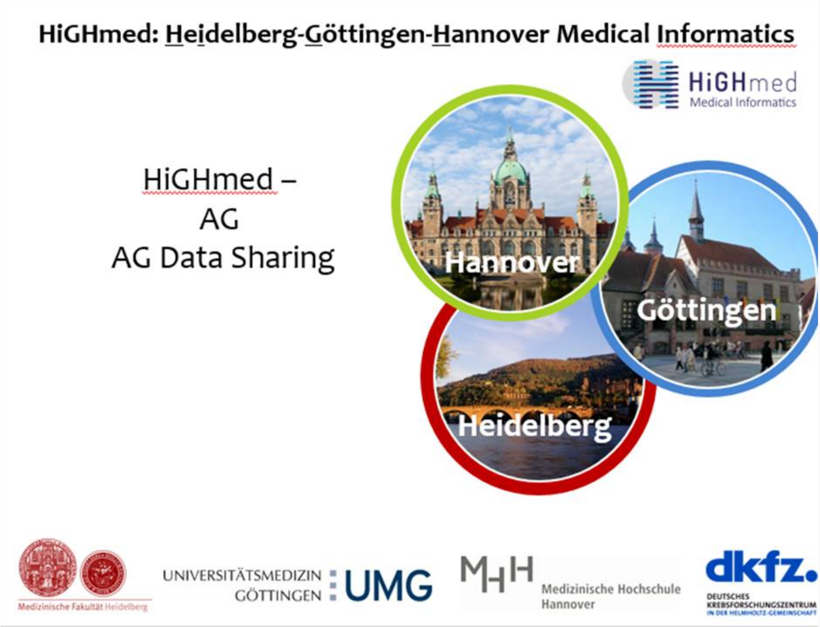
Getting Real: Merging Bioinformatics Standards and Crypto-State-of-the-Art

Workshop Omics in Medical Research – TMF, Berlin, 2017

Prof.Dr. Kay Hamacher

TU Darmstadt
http://www.kay-hamacher.de

Embedded into real practice . . .



HiGHmed: Heidelberg-Göttingen-Hannover Medical Informatics

HiGHmed
Medical Informatics

HiGHmed –
AG
AG Data Sharing

Medizinische Fakultät Heidelberg

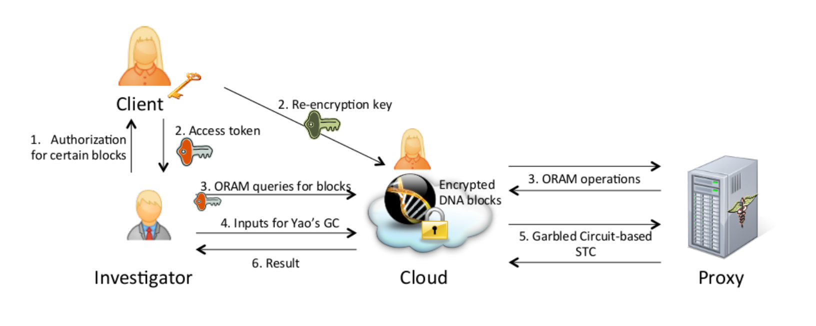
UNIVERSITÄTSMEDIZIN GÖTTINGEN : UMG

MHH Medizinische Hochschule Hannover

dkfz. DEUTSCHES KREBSFORSCHUNGSZENTRUM IN DER HELMHOLTZ-GEMEINSCHAFT

Examples (1): Oblivious RAM (ORAM)

write several times; shuffle data chunks; hide query $\Rightarrow$ protect queries

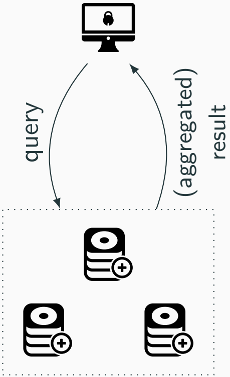


- pattern matching $\surd$
- DNA fingerprint(s) $\equiv$ description vector$\surd$
- Bayesian learning/updates $\surd$

Karvelas, Peter, Katzenbeisser, Tews, Hamacher *"Privacy Preserving Whole Genome Sequence Processing through Proxy-Aided ORAM"* Workshop on Privacy in the Electronic Society (WPES2014):1-10

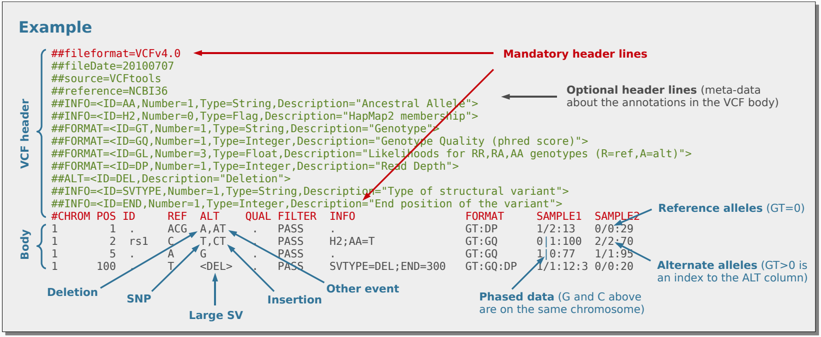
Examples (2): Privacy-Preserving VCF Queries

- Genomic data is sensitive data
- Medical treatment and research profits from access to large genomic datasets
- Trivially aggregating multiple genomic databases . . .
    - allows for more detailed analysis
    - but risks privacy of each dataset and entry
    - might not be desirable by the database provider
- **Solution:** Secure Multi-Party Computation (SMPC) for private aggregation and database queries



Federation of genomic databases

Variant Call Format (VCF)



Example

```
##fileformat=VCFv4.0
##fileDate=20100707
##source=VCFtools
##reference=NCBI36
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele">
##INFO=<ID=H2,Number=0,Type=Flag,Description="HapMap2 membership">
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=GQ,Number=1,Type=Integer,Description="Genotype Quality (phred score)">
##FORMAT=<ID=GL,Number=3,Type=Float,Description="Likelihoods for RR,RA,AA genotypes (R=ref,A=alt)">
##FORMAT=<ID=DP,Number=1,Type=Integer,Description="Read Depth">
##ALT=<ID=DEL,Description="Deletion">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=END,Number=1,Type=Integer,Description="End position of the variant">
#CHROM POS ID     REF  ALT    QUAL FILTER  INFO                 FORMAT     SAMPLE1   SAMPLE2
1        1 .      ACG  A,AT    .   PASS    .                    GT:DP      1/2:13    0/0:29
1        2 rs1    C    T,CT    .   PASS    H2;AA=T              GT:GQ      0|1:100   2/2:70
1        5 .      A    G       .   PASS    .                    GT:GQ      1|0:77    1/1:95
1      100 .      T    <DEL>   .   PASS    SVTYPE=DEL;END=300   GT:GQ:DP   1/1:12:3  0/0:20
```

VCF header
Body
Mandatory header lines
Optional header lines (meta-data about the annotations in the VCF body)
Reference alleles (GT=0)
Alternate alleles (GT>0 is an index to the ALT column)
Deletion
SNP
Large SV
Insertion
Other event
Phased data (G and C above are on the same chromosome)

Figure 1: VCF format example – from vcftools website.

Beacon Project

Beacon Network

Global Alliance
for Genomics & Health
Collaborate. Innovate. Accelerate.

Beacon Network (GA4GH Project) established to evaluate eagerness of institutions world-wide to engage in distributed variant query service.

“Do you have a genome with mutation X in your database?”

“Yes” / **“No”**

Beacon Privacy Issues

Recent work (Thenen, N., Ayday, E., C. Ercument; bioRxiv; 200147; Sep. 2017) strongest challenge to privacy of Beacon networks yet

- Need only 5 queries to re-identify member in 65-individuals beacon (95 % confidence)
  - even when single-nucleotide polymorphisms (SNPs) with minor allele frequency (MAF) $< 5\%$ are hidden
  - exploiting linkage disequilibrium
  - and high-order Markov chains

Our work mitigates this risk by

- hiding which beacon contributed to total
- incurring a threshold on `count`-queries
- enabling a reverse follow-up scenario

Variant Query Format

- Compression/Encoding of variant data into $(\kappa = 32, \psi = 16)$ bit (key, value) space
- Key holds the position in a reference genome $(\log_2(3.2 \cdot 10^9) \approx 31.6 < 32)$
- Small value space must incur variant information loss
- But domain-specific knowledge makes a virtue of this necessity to let similar variants match
- Allow same key multiple times to store e.g. diploid variations or upper and lower length of deletion/copy number variation (CNV)/inversion

VQF Translation

Variant	Stored Information	#Values
SNP/SNV	Alternative nucleotide.	4
Deletion/CNV /Inversion	Up and down rounded logarithm (base $b = 2$) of length, up to log-length s. Plus frameshift bit for deletions.	$s = 16$
Insertion	Up to $s_{\text{ins}} = 7$ inserted nucleotides and a frameshift bit.	$2 \cdot 4^{s_{\text{ins}}=7}$
Other	Only flag as *other* variation. Captures more complex variations.	1

Privacy-Preserving Genomic Queries

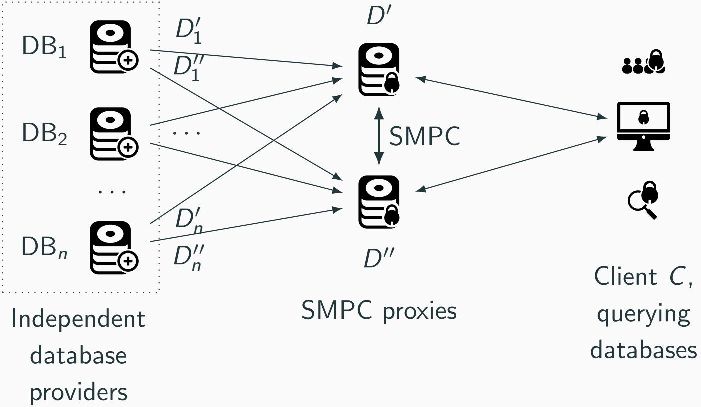
Supported queries are of the form:

SELECT f(*) FROM Variants
WHERE ((loc$_1$, var$_1$), ..., (loc$_m$, var$_m$)) IN Genome
AND cancer = X AND ... AND age$_{\mathrm{min}}$ $\leq$ age $\leq$ age$_{\mathrm{max}}$

f(*) can be an arbitrary function (e.g., aggregation, threshold).

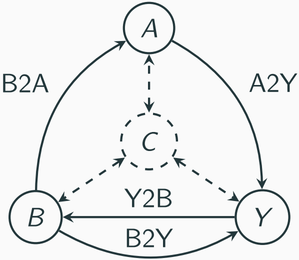
D. Demmler, K. Hamacher, T. Schneider, S. Stammler.
Privacy-Preserving Whole-Genome Variant Queries 16th Int. Conf. Cryptology And Network Security (CANS) accepted, 2017.

Schematic Overview



The ABY Framework

- ABY (Demmler, D., Schneider, T., Zohner, M. (NDSS'2015)) is a C++ framework for secure two-party computation
- 3 supported protocols:
  - Arithmetic sharing: $+$ secret-sharing
  - Boolean sharing (GMW): $\oplus$ secret-sharing
  - Yao's garbled circuits
- Open source on github: `http://encrypto.de/code/ABY`



Benchmarks

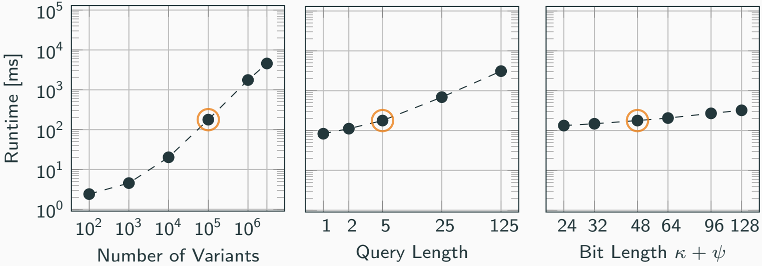


Figure 2: Online runtime in ms.

A typical query from 100 000 variants on 5 positions with $\kappa = 32$ and $\psi = 16$ needs 178 ms runtime and 11 MiB communication in the online phase. (Setup phase: 3.4 s, 733 MiB)

Summary & Acknowledgements

- Trust not essential: computational trust
  vs. social/legal construct
- work in security / protocol design / cryptography
  mature and applicable
- loss of (computational) efficiency $\surd$
  but trade-off might be worth it
- framework(s) do exist;
  but you <u>need</u> experts in crypto & bioinfo
  "Anyone, from the most clueless amateur to the best cryptographer, can create an algorithm that he himself can't break." — Bruce Schneier

`http://www.kay-hamacher.de`

CRISP
Center for Research
in Security and Privacy

CYSEC
Cybersecurity
TU Darmstadt

Bundesministerium
für Bildung
und Forschung

HiGHmed
Medical Informatics

DFG

HESSEN
Hessisches
Ministerium für
Wissenschaft
und Kunst

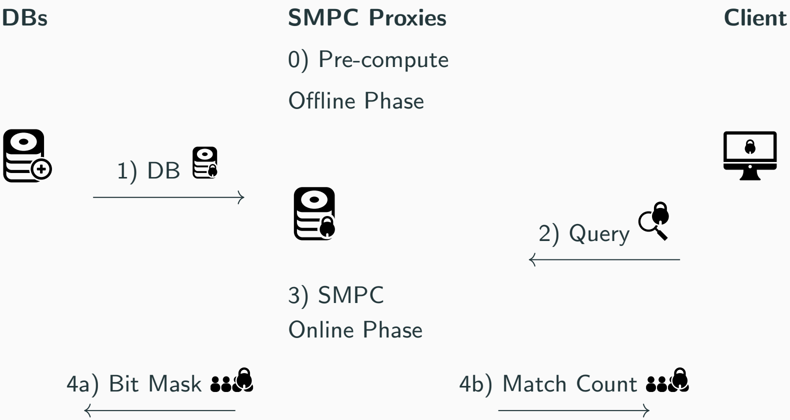
FCI
FONDS DER
CHEMISCHEN
INDUSTRIE

http://www.kay-hamacher.de

kay.hamacher@cysec.de

Protocol

Protocol Phases



DBs

SMPC Proxies

Client

0) Pre-compute

Offline Phase

1) DB

2) Query

3) SMPC

Online Phase

4a) Bit Mask

4b) Match Count

Circuit Design

- GMW protocol for inherent secret-sharing
- Efficient SIMD operations
- Variant matching: array of equality gates combined with OR tree
- Auxiliary information: combined with AND tree

Homomorphic Encryption (HE)

Starting with

C. Gentry. *"Fully Homomorphic Encryption Using Ideal Lattices"* 41$^{\mathrm{st}}$ ACM Symposium on Theory of Computing (STOC), 2009

$$\begin{tikzcd}
A, B \arrow[r, "+"] \arrow[d, "\mathrm{Enc}"] & A + B \\
\mathrm{Enc}(A), \mathrm{Enc}(B) \arrow[r, "\oplus"] & \mathrm{Enc}(A + B) \arrow[u, "\mathrm{Dec}"']
\end{tikzcd}$$

Several framworks, e.g., Damgård, et.al., *"Efficient and secure comparison for online-auctions"*, ACSIP2007, Lecture Notes in Computer Science (4586)