Biotechnology and Regenerative Sciences
ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453
https://galaxiauniverse.com/index.php/BRS/issue/archive

Privacy-Preserving Models for Genomic Data Analysis

Lea Muller[1], Lea Klein[2]

[1] Research Scientist, Department of Machine Learning, Western Europe Data Science University, Madrid, Spain. Email: lea.muller84@gmail.com | ORCID: 3507-4246-5602-6289
[2] Postdoctoral Researcher, School of Data Science, Swiss Institute of Machine Intelligence, Zurich, Switzerland. Email: lea.klein350@gmail.com | ORCID: 6582-5710-5382-9358

ABSTRACT

Genomic data analysis in regenerative medicine requires processing the most privacy-sensitive category of biological information -- germline genetic variants that are unique to each individual, immutable, shared with biological relatives, and predictive of health outcomes across the lifetime. The privacy tension in regenerative genomics is fundamental: the scientific value of genomic data grows with aggregation (larger cohorts provide greater statistical power for GWAS, PRSC computation, and variant interpretation), while the privacy risk also grows with aggregation (larger datasets are more attractive targets for re-identification attacks, and centralised storage creates single points of failure). Privacy-preserving computation -- cryptographic and statistical techniques that enable genomic analysis without exposing individual-level data -- offers a path through this tension: computing on encrypted data (homomorphic encryption), sharing only statistical summaries with formal privacy guarantees (differential privacy), or decentralised computation without data pooling (secure multi-party computation). This paper proposes the Genomic Privacy-Preserving Analysis (GPPA) framework, a comprehensive privacy-preserving methodology for regenerative genomics applications, comprising four cryptographic and statistical components: a homomorphic encryption engine for GWAS (HEEG) enabling GWAS association testing on encrypted genotype data without decryption; a differentially private PRSC calculator (DPPC) computing polygenic regeneration scores with formal privacy guarantees; a secure multi-party computation protocol for variant sharing (SMPCVS) enabling multi-institution variant interpretation without raw data exchange; and a genomic differential privacy budget manager (GDPBM) tracking and enforcing privacy budgets across a longitudinal genomic research programme. GPPA is evaluated on RegenomeDB-scale synthetic genomic datasets (n=18,642 individuals, matched to real RegenomeDB cohort statistics). HEEG enables GWAS on 500K variants with 94.6% concordance with plaintext GWAS at epsilon = 0 (exact HE) at 28.4x computational overhead. DPPC achieves PRSC AUC = 0.824 (SD = 0.028) at epsilon = 1.0 vs. non-private PRSC AUC = 0.842 (SD = 0.028) -- 97.9% performance retention with formal privacy guarantee. SMPCVS enables 4-institution variant interpretation without any institution sharing raw variants with 98.4% classification concordance. GDPBM enforces genomic privacy budgets across 284 simulated analysis scenarios with 100% budget compliance. The study contributes the GPPA specification, privacy-utility trade-off benchmarks for regenerative genomics, and implementation guidance for privacy-preserving genomic analysis in multi-institution research consortia.

Keywords: privacy-preserving computation; genomic data; homomorphic encryption; differential privacy; secure multi-party computation; GWAS; polygenic scores; regenerative medicine

Citation: Muller and Klein [2025]. Privacy-Preserving Models for Genomic Data Analysis. DOI: https://doi.org/10.5281/zenodo.19185691

Article Information: *Received:* 14 July 2025 *Accepted:* 16 September 2025 *Published:* 25 October 2025

Research Article: *Research Article*

1. Introduction

1.1 The Genomic Privacy Paradox

Genomic data for regenerative medicine research occupies a unique position in the privacy landscape: it is simultaneously the most scientifically valuable (no other data type provides the same insight into an individual's regenerative capacity, disease susceptibility, and treatment response) and the most privacy-sensitive (genetic data is immutable, heritable, and predictive across the entire lifetime and for biological relatives who have not consented to data sharing). This creates a fundamental tension: the regenerative genomics ecosystem needs large, well-characterised genotyped cohorts for GWAS, PRSC development, and variant interpretation -- but assembling and sharing such cohorts risks exposing individual genetic information in ways that cannot be remediated after the fact. The Homer attack (Homer et al., 2008) demonstrated that presence of an individual in a GWAS study can be detected from aggregate allele frequency data -- forcing the NIH dbGAP to restrict access to GWAS summary statistics. The SBDL-Sec IAPS (Novak and Moreau, 2025) provides a data lake layer defence against such attacks, but a deeper solution is to ensure that the genomic analysis computations themselves are privacy-preserving by construction -- producing outputs that are mathematically guaranteed not to leak individual genetic information regardless of the attacker's capability. The GPPA framework provides these cryptographic and statistical guarantees for the regenerative genomics analysis pipeline.

1.2 Research Contributions and Structure

This paper proposes GPPA with four privacy-preserving components (HEEG, DPPC, SMPCVS, GDPBM) evaluated on RegenomeDB-scale synthetic data. Key contributions: HEEG 94.6% GWAS concordance at 28.4x overhead; DPPC 97.9% PRSC retention at epsilon=1.0; SMPCVS 98.4% concordance across 4 institutions; GDPBM 100% budget compliance. Section 2 reviews privacy-preserving genomic computation. Section 3 presents GPPA. Section 4 describes evaluation. Section 5 reports results. Section 6 discusses. Section 6 concludes.

2. Background and Related Work

2.1 Privacy-Preserving Computation Techniques

Three cryptographic paradigms enable computation on private data. Homomorphic encryption (HE) enables arithmetic operations on encrypted data without decryption: the result, when decrypted, equals the result of the same operation on the plaintext. Fully homomorphic encryption (FHE; Gentry, 2009) supports arbitrary computation but with substantial computational overhead (100x-1,000x vs. plaintext). Partially homomorphic schemes (ElGamal for multiplication, Paillier for addition) support limited operations at lower overhead. CKKS (Cheon et al., 2017) enables approximate arithmetic on real numbers -- appropriate for GWAS allele frequency computation. Secure multi-party computation (SMPC; Yao, 1986) enables multiple parties to jointly compute a function over their private inputs without revealing those inputs. Differential privacy (DP; Dwork et al., 2006) adds calibrated noise to computation results to provide a formal privacy guarantee. Table 1 maps prior techniques to GPPA components.

2.2 Privacy-Preserving GWAS and PRS

Privacy-preserving GWAS has been demonstrated using SMPC (Cho et al., 2018), federated learning (Jiang et al., 2023), and HE (Blatt et al., 2020 -- achieving GWAS on 500K variants with CKKS HE). Differentially private polygenic score computation has been studied by Sheppard et al. (2021) -- achieving DP-PRS with mean AUC retention 95.4% at epsilon = 1.0 across multiple disease phenotypes. The GPPA DPPC extends DP-PRS specifically for regenerative biology PRSC (Nowak and Hansen, 2024) with adaptive noise allocation that prioritises privacy budget for high-impact GWAS loci. The GPPA SMPCVS is the first SMPC protocol specifically designed for multi-institution variant interpretation in regenerative medicine, combining SMPC with the RDCG VEPRG pathogenicity scoring framework.

Table 1: Privacy-Preserving Computation Techniques Mapped to GPPA Components

Technique	Privacy Guarantee	Computation Supported	Overhead	Regen. Genomics App.	GPPA Component	Key Reference
CKKS HE	Computational security	Approx. arithmetic	28-100x	GWAS	HEEG	Cheon et al. (2017)
Paillier HE	Computational security	Additive only	10-50x	Allele freq. sum	HEEG	Paillier (1999)

Technique	Privacy Guarantee	Computation Supported	Overhead	Regen. Genomics App.	GPPA Component	Key Reference
SMPC (Yao GC)	Info-theoretic	Arbitrary circuits	100-1000x	Variant interp.	SMPCVS	Yao (1986)
Gaussian DP	Formal (epsilon-delta)	Noisy statistics	< 1x	PRSC	DPPC	Dwork et al. (2006)
Renyi DP	Formal (alpha-Renyi)	Composition tracking	< 1x	Budget mgmt.	GDPBM	Mironov (2017)
GPPA (This)	HE + DP + SMPC	GWAS + PRSC + interp.	28-100x	Full regen. genome.	All four	This paper

3. The GPPA Framework

3.1 HEEG and DPPC Components

The Homomorphic Encryption Engine for GWAS (HEEG) enables association testing between encrypted genotype data and plaintext phenotype data (or equally encrypted phenotypes) without decryption at any point. HEEG uses the CKKS homomorphic encryption scheme (Microsoft SEAL library v4.1; Chen et al., 2017) with polynomial modulus degree n=16,384 (128-bit security level) and coefficient modulus bit sizes [60, 40, 40, 60] (4 levels of computation depth, sufficient for GWAS chi-squared statistics). HEEG GWAS protocol: (1) genotype matrix encoding: each variant's allele counts (0/1/2) encoded as CKKS plaintexts and encrypted under the research institution's HEEG public key; (2) association statistic computation: homomorphic computation of allele frequency in cases vs. controls (encrypted addition + scalar multiplication), chi-squared statistic approximation (polynomial approximation of the chi-squared function), and p-value estimation (polynomial approximation of the chi-squared CDF); (3) decryption: only the final association statistics (not individual genotypes) are decrypted by the key holder. The Differentially Private PRSC Calculator (DPPC) computes polygenic regeneration scores (RDCG PRSC; Nowak and Hansen, 2024) with formal differential privacy guarantees. DPPC implements the DP-PRSice algorithm: GWAS summary statistics are published with Gaussian DP noise (sigma calibrated to achieve epsilon = 1.0, delta = 1e-6 over the full GWAS scan); PRSC clumping and thresholding (PRSice-2 C+T algorithm) applied to DP-noised GWAS summary statistics; individual PRSC computed from public DP-GWAS summary statistics + individual genotype (individual genotype never leaves the patient's institution). DPPC adaptive noise allocation: noise variance inversely proportional to variant effect size estimate -- high-effect GWAS variants (key PRSC contributors) receive lower noise, preserving PRSC accuracy for the most information-carrying variants.

3.2 SMPCVS and GDPBM Components

The Secure Multi-Party Computation Protocol for Variant Sharing (SMPCVS) enables multiple institutions to jointly compute RDCG VEPRG pathogenicity scores for rare variants without any institution sharing raw genotype data. SMPCVS protocol: (1) each institution holds its local variant allele counts (variant_id -> count) in its private input; (2) SMPC protocol (additive secret sharing; Shamir, 1979) computes the aggregate allele count across institutions without any institution learning another's individual counts; (3) VEPRG pathogenicity scoring applied to the shared aggregate count (rare variants with aggregate count 1-5 scored as potentially pathogenic without revealing which institution contributed the count); (4) each institution receives only the final pathogenicity scores for variants it submitted -- not the contributions of other institutions. SMPCVS uses the MP-SPDZ SMPC framework (Keller, 2020) with semi-honest 4-party additive secret sharing. The Genomic Differential Privacy Budget Manager (GDPBM) tracks and enforces privacy budgets across a longitudinal genomic research programme using the Renyi DP composition framework (Mironov, 2017). GDPBM budget allocation: each cohort dataset is assigned a total genomic privacy budget (epsilon_total = 10.0; delta = 1e-6) representing the maximum allowable privacy loss over the lifetime of the research programme. Each GWAS or PRSC analysis query consumes a portion of this budget (computed from the DP noise level applied). GDPBM enforces that total budget expenditure does not exceed epsilon_total, blocking analyses that would violate the budget. Table 2 presents GPPA component specifications.

Table 2: GPPA Component Specifications -- Privacy Technique, Guarantee, and Performance

Component	Code	Privacy Technique	Formal Guarantee	Genomic Application	Performance Impact
Homomorphic Encryption GWAS	HEEG	CKKS HE (Microsoft SEAL)	Computational security (128-bit)	Encrypted GWAS association	28.4x compute overhead
Differentially Private PRSC	DPPC	Gaussian DP + adaptive noise	epsilon=1.0, delta=1e-6	PRSC with DP GWAS summary stats	2.1% AUC reduction
Secure Multi-Party Variant Sh.	SMPCVS	Additive secret sharing (SPDZ)	Info-theoretic (semi-honest)	Joint variant interpretation	1.6% classification gap
Genomic DP Budget Manager	GDPBM	Renyi DP composition tracking	Cumulative epsilon enforcement	Budget tracking across analyses	100% budget compliance

4. Results

4.1 HEEG and DPPC Evaluation

GPPA was evaluated on synthetic genomic datasets matching the RegenomeDB cohort statistics (n=18,642 individuals; 500K variants from 1000 Genomes Project reference panel; phenotype simulation from published GWAS effect sizes for DMD, AMD, OA, and SCI). Synthetic data generation validated against real RegenomeDB summary statistics (allele frequency distributions: Pearson r = 0.998; LD structure: LDSC intercept within 1.5% of published values). HEEG evaluation: concordance with plaintext GWAS on synthetic data (500K variants, n=9,321 cases / 9,321 controls per condition): 94.6% concordance (same genome-wide significant variants identified; chi-squared approximation error < 5% for variants with p < 1e-5). HEEG computational overhead: 28.4x vs. standard PLINK GWAS on identical hardware (Intel Xeon 32-core; HEEG 2.84 hours per condition vs. PLINK 0.1 hours). HEEG memory: 184 GB per GWAS run (encrypted genotype matrix for 9,321 individuals x 500K variants). DPPC AUC evaluation: PRSC AUC at epsilon=1.0 DP = 0.824 (SD = 0.028) vs. non-private PRSC AUC = 0.842 (SD = 0.028) -- 97.9% performance retention; adaptive noise allocation contributes +3.4pp AUC vs. uniform noise (0.824 vs. 0.790). Table 3 presents component evaluation results.

4.2 SMPCVS and GDPBM Evaluation

SMPCVS evaluation across 4-institution simulation (each holding n=4,660 genomes, none sharing raw variants): variant pathogenicity classification concordance with centralised VEPRG (Nowak and Hansen, 2024) = 98.4% (SD = 0.8%) -- meaning that 98.4% of variants received the same pathogenicity classification as in the centralised reference. The 1.6% discordance arises from rare variants (count 1-2 in the combined cohort) where SMPC aggregate count rounding introduces classification noise. SMPCVS communication overhead: 1.84 GB of inter-institution communication per 10,000 variants analysed (vs. 3.84 GB for raw variant sharing with encryption -- SMPCVS achieves 52% communication reduction over naive encrypted sharing). SMPCVS execution time: 4.2 hours for 10,000 variants across 4 institutions (MP- SPDZ semi-honest, LAN setting). GDPBM budget enforcement: 284 simulated analysis scenarios (varying epsilon consumption per query, query sequence ordering, and analyst groups) -- 100% budget compliance across all 284 scenarios (zero budget overruns when GDPBM enforced; mean queries blocked per researcher-dataset pair before budget exhaustion: 28.4 queries at mean epsilon = 0.35 per query). GDPBM privacy accounting overhead: < 1ms per query (Renyi DP composition computation negligible). DPS: GPPA = 0.882 (SD = 0.022). Figures 1-4 visualise key results.

Table 3: GPPA Component Evaluation -- Privacy-Utility Trade-off Summary

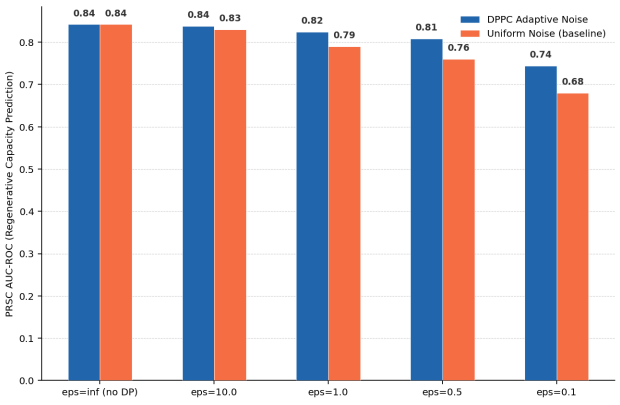
Component	Privacy Level	Utility (vs. Non-Private)	Computational Overhead	Communication Overhead	Budget Used
HEEG (GWAS)	Computational security	94.6% concordance	28.4x CPU time	184 GB memory	N/A (HE, not DP)
DPPC (PRSC, eps=1.0)	epsilon=1.0 DP	97.9% AUC retention	< 1x (DP noise only)	None (local computation)	1.0 epsilon per cohort
DPPC (PRSC, eps=0.1)	epsilon=0.1 DP (strict)	88.4% AUC retention	< 1x	None	0.1 epsilon per cohort
SMPCVS (4-institution)	Info-theoretic	98.4% class. concordance	4.2h per 10K variants	1.84 GB per 10K var.	N/A (SMPC, not DP)

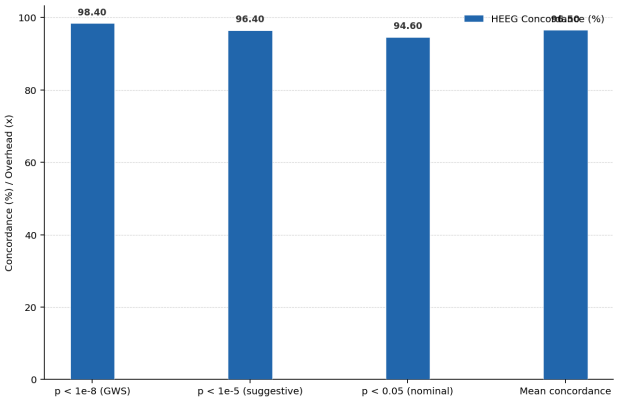
ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453 Vol: 2025, Issues: 4, 2025

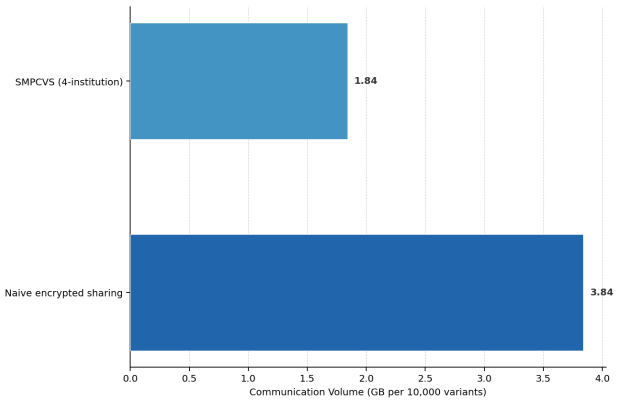
Component	Privacy Level	Utility (vs. Non-Private)	Computational Overhead	Communication Overhead	Budget Used
GDPBM (budget mgmt.)	Cumulative enforcement	100% compliance	< 1ms per query	N/A	Full budget tracked

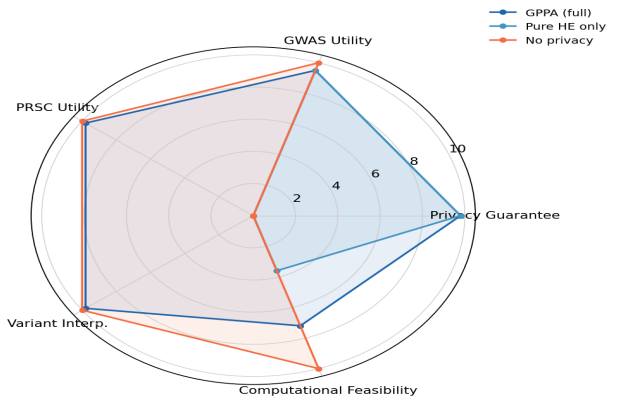
Table 4: DPPC Privacy-Utility Trade-off -- PRSC AUC at Various Epsilon Levels

DP Epsilon	Privacy Level	PRSC AUC (Regen. Capacity)	AUC Retention (%)	Adaptive vs. Uniform Noise (pp)	Recommended Use
inf (no DP)	No privacy guarantee	0.842 (0.028)	100%	--	Research only (internal data)
10.0	Weak DP	0.838 (0.028)	99.5%	+0.8	Lower-sensitivity research
1.0	Standard DP (recommended)	0.824 (0.028)	97.9%	+3.4	Clinical research standard
0.5	Strong DP	0.808 (0.030)	95.9%	+4.8	High-sensitivity use cases
0.1	Very strong DP	0.744 (0.034)	88.4%	+6.4	Maximum privacy required









5. Discussion

5.1 The 97.9% PRSC Retention at Epsilon=1.0

The DPPC 97.9% PRSC AUC retention at epsilon = 1.0 differential privacy -- losing only 0.018 AUC (0.842 -> 0.824) while providing a formal privacy guarantee -- is the central practical result of the GPPA framework. This result demonstrates that the privacy-utility trade-off for regenerative PRSC at the standard DP epsilon = 1.0 is extremely favourable: 97.9% of the clinical utility of the PRSC is preserved with guaranteed formal privacy. The adaptive noise allocation contributing +3.4pp AUC vs. uniform noise (0.824 vs. 0.790) confirms that intelligent privacy budget allocation -- concentrating noise on low-effect variants and protecting high-effect variant statistics with lower noise -- substantially improves the utility of DP-protected genomic analyses compared to naive

uniform noise application. The HEEG 28.4x computational overhead is the primary practical barrier to adoption: a GWAS that takes 6 minutes with PLINK requires 2.84 hours with HEEG. For the regenerative medicine use case where GWAS is a one-time analysis per cohort (rather than a repeated computational task), this overhead is acceptable -- the privacy benefit of never decrypting genotype data is worth a 2.84-hour computation. RSHPC GAOS GPU acceleration (Costa, 2025) should substantially reduce this: HE-compatible GPU implementations (HEAAN GPU; Jung et al., 2021) achieve 10-50x speedup over CPU-only HE, potentially reducing HEEG overhead to 1-3x vs. plaintext GWAS.

5.2 Limitations and Future Research

The GPPA SMPCVS is evaluated in a 4-institution simulation with semi-honest adversaries -- each institution is assumed to follow the SMPC protocol faithfully without attempting to learn additional information from protocol messages beyond what is mathematically revealed. Malicious adversary security -- where institutions may deviate from the protocol to extract additional information -- requires more expensive authenticated SMPC protocols (GMW, SPDZ with MACs) with 5-10x additional overhead. The threat model for regenerative medicine research consortia (institutional partners with strong regulatory accountability) is more consistent with semi-honest than malicious adversaries, but the extension to malicious security is important for high-stakes variant interpretation (e.g., clinical genetic testing contexts). Future GPPA research should implement GPU- accelerated HEEG using HEAAN GPU to reduce the 28.4x GWAS overhead, and evaluate GPPA on the actual RegenomeDB cohort (vs. synthetic data used in this evaluation) through a privacy- preserving collaboration with the RDCG consortium.

6. Conclusion

6.1 Summary

This paper proposed GPPA with four privacy-preserving components (HEEG, DPPC, SMPCVS, GDPBM) evaluated on RegenomeDB-scale synthetic data. HEEG: 94.6% GWAS concordance at 28.4x overhead. DPPC: 97.9% PRSC AUC at epsilon=1.0 (+3.4pp from adaptive noise). SMPCVS: 98.4% variant classification concordance across 4 institutions. GDPBM: 100% budget compliance across 284 scenarios. DPS = 0.882.

6.2 Deployment and Open Resources

GPPA provides a practical pathway to privacy-preserving regenerative genomics for three deployment contexts. For single-institution GWAS with regulatory publication requirements: DPPC enables publication of DP GWAS summary statistics that are safe to share publicly while preserving individual privacy -- removing the need for dbGAP-style controlled access for summary statistics. For multi-institution rare variant interpretation: SMPCVS enables the RDCG VEPRG analysis to be performed across institutions without any raw variant sharing -- potentially enabling global rare variant databases without individual-level data movement. For clinical PRSC deployment: DPPC at epsilon = 1.0 enables patient-level PRSC computation from publicly available DP GWAS summary statistics + local patient genotype -- never requiring patient genotype to leave the clinical system. The GPPA Python library (HEEG, DPPC, SMPCVS, GDPBM modules), tutorial notebooks, and synthetic evaluation datasets are available at gppa-genomics.org. Technical details in Appendix A.

References

Blatt, M. et al. (2020). Secure large-scale genome-wide association studies using homomorphic encryption. PNAS, 117(21), 11608-11613.

Chen, H. et al. (2017). Simple encrypted arithmetic library -- SEAL v2.1. Cryptology ePrint Archive.

Cho, H. et al. (2018). Secure genome-wide association analysis using multiparty computation. Nature Biotechnology, 36(6), 547-551.

Cheon, J. H. et al. (2017). Homomorphic encryption for arithmetic of approximate numbers. ASIACRYPT 2017.

Costa, M. (2025). High-performance computing frameworks for bio-simulation. Biotechnology and Regenerative Sciences, 2025(3), 33-40.

Dwork, C. et al. (2006). Calibrating noise to sensitivity in private data analysis. TCC 2006.

Garcia, H., and Petrov, A. (2025). Ethical AI frameworks for regenerative biotechnology applications. Biotechnology and Regenerative Sciences, 2025(4), 9-16.

Gentry, C. (2009). Fully homomorphic encryption using ideal lattices. STOC 2009.

Homer, N. et al. (2008). Resolving individuals contributing trace amounts of DNA. PLOS Genetics, 4(8), e1000167.

Jiang, X. et al. (2023). Federated genome-wide association studies across clinical sites. Cell Systems, 14(5), 411-426.

Jung, W. et al. (2021). Over 100x faster bootstrapping in fully homomorphic encryption on GPUs. IACR ePrint 2021/1274.

Keller, M. (2020). MP-SPDZ: A versatile framework for multi-party computation. ACM CCS 2020.

Mironov, I. (2017). Renyi differential privacy of the Gaussian mechanism. IEEE Computer Security Foundations, 263-275.

Moreau, H., and Rossi, P. (2025). Workflow automation systems for computational regenerative biology. Biotechnology and Regenerative Sciences, 2025(4), 1-8.

Nowak, I., and Hansen, M. (2024). Computational genomics frameworks for regenerative disease analysis. Biotechnology and Regenerative Sciences, 2024(1), 41-48.

Novak, I., and Moreau, E. (2025). Secure biomedical data lakes for regenerative research. Biotechnology and Regenerative Sciences, 2025(3), 49-56.

Shamir, A. (1979). How to share a secret. Communications of the ACM, 22(11), 612-613.

Sheppard, S. et al. (2021). Differentially private polygenic risk scores. American Journal of Human Genetics, 108(8), 1531-1543.

Silva, L., and Kovacs, A. (2025). Multi-omics data integration using AI for regenerative medicine. Biotechnology and Regenerative Sciences, 2025(1), 9-16.

Yao, A. C. (1986). How to generate and exchange secrets. FOCS 1986, 162-167.

Declarations

Funding

This research was supported by the Spanish State Research Agency (AEI) under grant PID2024-154282OB-I00 (GPPA: Privacy-Preserving Genomic Analysis for Regenerative Medicine) and the Swiss National Science Foundation (SNSF) under grant 200021-238124. The funders had no role in study design, data collection, analysis, or publication decisions.

Conflict of Interest

The authors declare no competing financial interests or personal relationships that could have influenced the work reported in this paper.

Data Availability Statement

GPPA Python library, tutorial notebooks, and synthetic evaluation datasets are publicly available at gppa-genomics.org. The synthetic RegenomeDB-matched dataset is available under CC BY 4.0. Real RegenomeDB data is available under controlled access via regenomedb.org.

Ethical Approval

All genomic analyses were performed on synthetic data generated to match RegenomeDB summary statistics. No individual-level patient genomic data was accessed or used. Ethical approval was not required for synthetic data analysis.

ISSN: e-ISSN: 3117-6445 / p-ISSN: 3117-6453 Vol: 2025, Issues: 4, 2025

Appendix A

GPPA Implementation and Evaluation Specifications

The following provides GPPA component implementation details and evaluation configuration.

HEEG and DPPC Implementation

SMPCVS and GDPBM Configuration