Federated Learning for Privacy-Preserving Immune Gene Typing and Cross-Cohort Immunogenomic Analysis from Long-Read Sequencing

Xavier Norales
Department of Computer Science, University of New Hampshire, Durham, NH, USA.
contactxavier@unh.edu

Kasper Hawkins
Department of Computer Science, George Mason University, Fairfax, VA, USA.
kasper219@gmu.edu

Brandon Karrett
Department of Computer Science, University of North Texas, Denton, TX, USA.
contactbrandon@unt.edu

Abstract

The rapid adoption of long-read sequencing technologies has enabled high-resolution typing of highly polymorphic immune genes, such as those in the major histocompatibility complex, yet the aggregation of such data across multiple cohorts for immunogenomic association studies introduces significant privacy risks. This paper proposes a federated learning framework designed to enable privacy-preserving immune gene typing and cross-cohort immunogenomic analysis from distributed long-read sequencing datasets. We conceptualize a system architecture that integrates local model training on cohort-specific sequencing repositories with secure aggregation protocols, differential privacy mechanisms, and decentralized governance structures. The framework addresses critical trade-offs between model fidelity, communication efficiency, statistical power, and protection against re-identification attacks. We examine the infrastructural demands of deploying such a system across heterogeneous clinical and research sites, including the need for harmonized variant calling pipelines, standardized immune gene annotations, and robust quality control measures that preserve privacy while ensuring biological validity. Furthermore, we analyze the governance and policy implications of federated immunogenomic analysis, including consent management, data sovereignty, and equitable access to derived models. By drawing parallels to existing federated learning deployments in medical imaging and electronic health records, we discuss sustainability, fairness, and robustness challenges specific to polymorphic gene typing. Our analysis concludes that while federated learning offers a compelling paradigm for multi-cohort immunogenomic discovery, its successful implementation requires careful orchestration of algorithmic, regulatory, and ethical dimensions.

Keywords

federated learning, privacy-preserving genomics, immune gene typing, long-read sequencing, cross-cohort analysis, differential privacy, secure aggregation, immunogenomics, data governance.

1. Introduction

The advent of long-read sequencing platforms has revolutionized the characterization of highly polymorphic immune genes, particularly the human leukocyte antigen (HLA) and killer-cell immunoglobulin-like receptor (KIR) gene families, whose complex structural variation and extreme polymorphism have historically resisted accurate typing with short-read technologies [1,2]. Long-read sequencing provides contiguous haplotypic information that enables complete resolution of allelic diversity, phase determination, and detection of novel variants [3]. However, the power to uncover associations between immune gene variation and disease susceptibility, drug response, or transplant outcomes depends on assembling large, diverse cohorts that collectively capture global genetic diversity [4]. Such cross-cohort aggregation poses fundamental privacy challenges because genomic data, especially the highly identifying HLA haplotypes, can be linked to individuals through re-identification attacks or inference of disease status [5]. Traditional approaches that centralize raw sequencing data from multiple institutions exacerbate these risks and often conflict with data protection regulations such as the General Data Protection Regulation (GDPR) and the Health Insurance Portability and Accountability Act (HIPAA).

Federated learning (FL) has emerged as a distributed machine learning paradigm that allows multiple data owners to collaboratively train a shared model without exchanging raw data [7]. Instead, only model updates—typically gradient vectors or model parameters—are communicated to a central server, where they are aggregated to improve the global model. When combined with differential privacy (DP) and secure multi-party computation (SMC), FL can provide provable privacy guarantees against both external adversaries and honest-but-curious participants [8]. Applying FL to immune gene typing from long-read data represents a novel synthesis of computational genomics, privacy engineering, and distributed systems research. Early efforts in federated genomics have focused on variant calling and association studies using short-read data [9], but the unique challenges posed by long-read sequencing—higher data volumes, longer runtimes, and the need for haplotype-aware alignment—require tailored architectural solutions.

This paper presents a comprehensive system-level framework for privacy-preserving immunogenomic analysis via federated learning. We describe the architectural components necessary to support cross-cohort immune gene typing, including local training modules, secure communication channels, and aggregation servers that enforce DP budgets. We then examine the structural trade-offs inherent in such a system: between model accuracy and privacy loss, between communication efficiency and convergence speed, and between statistical validity and cohort heterogeneity. Governance considerations are also central: federated immunogenomics demands new models of consent that allow participants to opt in to model training while retaining control over their data, and it raises questions about who owns the resulting models and how they are deployed equitably. By drawing on lessons from federated learning in healthcare and bioinformatics, we identify critical gaps and propose directions for future research.

2. Background and Related Work

Immune gene typing from long-read sequencing has matured rapidly in recent years, with tools such as HLALA, Kourami, and the scalable framework described in [6] demonstrating the ability to produce full-length, allele-level calls from Pacific Biosciences and Oxford Nanopore reads. These methods exploit the long reads to phase across exons and introns, resolving ambiguous haplotypes that short reads cannot distinguish [3]. Despite these technical advances, the clinical and research utility of immune gene typing is often limited by

cohort size. Large-scale consortia such as the UK Biobank, All of Us, and the International HLA and Immunogenetics Workshop have demonstrated that cross-population analyses can reveal ancestry-specific allele frequencies, linkage disequilibrium patterns, and disease associations that are invisible in homogeneous cohorts [10]. Yet the logistical and regulatory barriers to sharing raw sequencing data across jurisdictions remain formidable.

Federated learning was originally proposed for mobile keyboard prediction [7], but has rapidly been adopted in healthcare domains where data sensitivity is paramount. In medical imaging, FL has enabled multi-institutional training of diagnostic models for chest X-rays, mammography, and histopathology without centralizing patient images [11]. In genomics, early federated approaches have focused on variant calling from whole-genome sequencing data. For example, the OpenFL platform and the iDASH secure genome analysis competition have demonstrated that FL can achieve comparable accuracy to centralized training for single-nucleotide variant detection [12]. However, immune gene typing presents additional complexities: the gene regions are extremely polymorphic, requiring de novo alignment against a large reference panel of known alleles; the output is categorical (allele calls) rather than continuous predictions; and the dimensionality of the haplotypic state space is enormous. Consequently, standard FL algorithms designed for convex optimization or deep neural networks may not be directly applicable.

Differential privacy provides a formal framework for quantifying and bounding the information leakage from model updates [2]. By adding calibrated noise to aggregated gradients, DP ensures that the contribution of any single individual cannot be distinguished with high confidence. In genomic FL, DP has been applied to protect summary statistics and model parameters, but the noise required for strong privacy guarantees can degrade model accuracy, especially when the number of participants is small [13]. The trade-off between privacy and utility is further complicated by the fact that immune gene variants are often rare—some alleles exist only in a handful of individuals—so the signal-to-noise ratio is inherently low. Adaptive DP mechanisms that allocate privacy budgets across training rounds, combined with secure aggregation that prevents the server from seeing individual updates, offer a path toward acceptable utility [14].

Secure multi-party computation (SMC) techniques, such as secret sharing and garbled circuits, can protect the raw values of model updates from the aggregation server, but they introduce substantial communication and computational overhead [15]. In the context of long-read sequencing, where the local training process itself is computationally intensive, the additional cost of SMC may be prohibitive for resource-constrained sites. Recent hybrid protocols that combine DP with lightweight secure aggregation have shown promise for reducing overhead while maintaining strong privacy [16]. Our framework adopts such a hybrid approach, leveraging the fact that the aggregated gradient across sites is often sufficient for convergence without revealing site-level statistics.

3. System Architecture and Design

The proposed federated immunogenomic analysis system comprises three main layers: local sites, a coordination server, and a governance layer. At each participating site, long-read sequencing data from consented individuals are processed through a standardized pipeline that performs base calling, alignment to a reference genome augmented with immune gene reference sequences, and genotyping using a tool such as the one described in [6]. This local pipeline produces per-individual consensus genotypes, which are then aggregated into site-level allele frequency tables or embedded into a latent representation suitable for model

training. The local model takes the form of a neural network or a statistical model that predicts phenotype associations from genotype features; however, the exact architecture depends on the downstream task (e.g., disease risk prediction, allele imputation, or population structure inference). To ensure cross-cohort compatibility, all sites must adopt a common data model with harmonized contig naming, variant normalization, and allele nomenclature (e.g., HLA allele codes from the IPD-IMGT/HLA database [5]).

During each training round, every site computes an update to its local model based on its own data and the current global model parameters, which are distributed by the coordination server. These updates are encrypted and sent to a secure aggregation server that computes the average update without decrypting individual contributions. The aggregation server may operate under a trusted execution environment (TEE) or use cryptographic protocols such as threshold Paillier encryption to ensure that even the aggregation server cannot infer site-level information [15]. After aggregation, the global model is updated and broadcast back to all sites. This process repeats until convergence or until a predefined privacy budget is exhausted.

A critical design choice is the granularity of the model. In standard FL, the model is a deep neural network with millions of parameters; communicating these parameters across many rounds incurs high bandwidth costs. For immune gene analysis, a more efficient approach is to train a linear or logistic regression model on compressed features derived from the genotype matrix, such as principal components of allele frequencies or kernel similarity matrices. Such models have far fewer parameters and are less sensitive to the non-i.i.d. distribution of data across sites [17]. Moreover, because the feature space is inherently categorical (allele counts), the training objective is convex, enabling faster convergence with fewer communication rounds. However, this simplicity comes at the cost of reduced expressiveness; nonlinear interactions between alleles and environmental factors may be missed. Hybrid models that combine a small neural network for feature extraction with a linear head can balance expressiveness and communication efficiency.

4. Privacy and Security Considerations

The privacy guarantees of the federated framework depend on the interplay between local DP, secure aggregation, and access control policies. We adopt a model of local DP where each site adds controlled noise to its model update before encryption, providing a mechanism against collusion between the aggregation server and other sites. The noise is calibrated to the sensitivity of the update—that is, the maximum possible change in the update due to the inclusion or exclusion of a single individual's data. For immune gene typing, sensitivity can be high because rare alleles may contribute disproportionately to the gradient, especially in small cohorts. To mitigate this, we recommend using a subsampling technique: each site randomly selects a subset of its data for each training round, which amplifies privacy and reduces sensitivity [18]. The privacy budget is tracked across rounds using a composition theorem, and the training terminates when the budget is depleted.

Secure aggregation eliminates the need for the server to see individual updates, but it does not protect against inference attacks that use the final global model to infer membership. Even after training, the global model parameters can encode information about the training data. For example, if a particular allele is predictive of a rare phenotype, the corresponding model coefficient may be large enough to indicate the presence of individuals with that allele in the training set. Therefore, we apply a final round of DP to the global model before release, ensuring that the trained model itself satisfies epsilon-differential privacy with respect to the entire coalition of training datasets [19]. This post-processing step is compatible with secure

aggregation because the aggregate update can be noisy, and the noise composition can be accounted for in the overall privacy analysis.

Beyond algorithmic protections, organizational privacy safeguards are necessary. Each participating site must implement strict data access controls, maintain audit logs of all model update transmissions, and undergo periodic privacy auditing. The governance layer includes a multi-stakeholder committee that reviews research proposals, approves the use of the global model for specific analyses, and ensures that no site can query the model for individual-level predictions without proper authorization. Such governance structures are analogous to data use agreements in biobanks but must be adapted to the dynamic and decentralized nature of FL.

5. Cross-Cohort Analysis and Governance

The fundamental value of federated immunogenomics lies in its ability to combine information from diverse cohorts without moving the data. However, this benefit introduces new challenges for statistical analysis and data governance. Heterogeneity across cohorts—differences in sequencing platforms, read depth, coverage bias, population ancestry, and phenotype definitions—can confound the federated model if not properly addressed. For immune gene typing, platform-specific errors, such as systematic undercalling of certain alleles due to low mappability of long reads, must be harmonized before training. A common approach is to compute calibration curves for each site using a held-out reference dataset and to adjust the local loss function accordingly [20].

Cross-cohort analysis also raises questions of fairness and representativeness. If the global model is trained predominantly on data from populations of European descent, its predictions may be biased for individuals from other ancestries, leading to inequitable clinical outcomes [21]. FL does not automatically mitigate this bias; in fact, it can exacerbate it if certain sites with minority populations have smaller sample sizes and thus contribute less to the aggregate model. To promote fairness, the aggregation algorithm can be weighted to amplify the influence of underrepresented cohorts, or the training objective can incorporate fairness constraints that penalize disparities in predictive performance across groups. These methods must be implemented carefully to avoid violating privacy guarantees, as they require access to group membership metadata.

Governance of the federated system involves not only the technical architecture but also the legal and ethical frameworks that enable data sharing. Traditional informed consent models, which ask individuals to consent to specific uses of their data, are ill-suited to the open-ended nature of federated learning, where the downstream analyses are not known at the time of consent. A dynamic consent platform, allowing individuals to grant or revoke permission for their data to be used in model training, is desirable but requires a mechanism to exclude a participant's data from the trained model after it has been aggregated—a non-trivial problem known as the "right to be forgotten" in federated settings [22]. Solutions include unlearning techniques that approximate the removal of a data point's influence from the model, but these remain an active research area.

6. Deployment and Sustainability

Deploying a federated learning system across heterogeneous research institutions demands robust infrastructure and ongoing operational support. Each site must have sufficient computational resources to run the local processing pipeline (including alignment and genotyping) and to perform local model training. Many clinical genomics labs operate on

high-performance computing clusters or cloud instances, but they may lack the specialized software dependencies required for FL. Containerization (e.g., using Docker and Kubernetes) can mitigate this by packaging the entire environment, including the genotyping tool from [6], the FL client library, and the differential privacy module. Network connectivity must be reliable, with low latency for communication rounds; institutional firewalls that block external connections may need to be configured with secured tunnels.

Sustainability of the federated network depends on funding models and incentives for participation. Hospitals and research institutes bear the computational and personnel costs of hosting a local FL node, yet they may not directly benefit from the global model if their primary interest is in local analyses. Developing a value proposition—such as access to a more accurate global model for their own clinical decision support, or co-authorship on publications from cross-cohort analyses—is essential to maintain engagement. Additionally, the federated system must be designed for long-term maintenance, including software updates, security patches, and migrations to newer sequencing technologies. A modular architecture that abstracts the genotyping pipeline from the FL framework allows each component to evolve independently.

7. Fairness and Robustness

Fairness in federated immunogenomics extends beyond population-level bias to include algorithmic fairness across sites. If one site contributes data from a highly homogeneous population, while another contributes data from a highly admixed population, the second site's data may be disproportionately affected by the noise added for privacy, leading to a higher error rate in that site's local model updates. Differential privacy budgets should therefore be allocated in a site-aware manner, possibly with larger noise for larger sites to equalize the per-sample privacy guarantee [23]. Robustness against adversarial attacks is also a concern. A malicious site could introduce corrupted updates to steer the global model toward a desired outcome (a model poisoning attack) or could attempt to extract information about other sites' data through the global model via inference attacks. Defenses such as robust aggregation (e.g., coordinate-wise median or trimmed mean) and anomaly detection on update norms can mitigate poisoning, but they may conflict with privacy because they require inspecting the raw updates. Secure protocols that allow verification of update integrity without revealing update values remain an open challenge.

8. Case Illustrations and Future Directions

To ground the discussion, consider two illustrative scenarios. In the first, a consortium of transplant centers across Europe and Africa wishes to build a predictive model for graft-versus-host disease (GVHD) risk based on HLA mismatches. Each center possesses long-read HLA typing data from hundreds of donor-recipient pairs, but local regulations prohibit sharing raw sequences across borders. Using our federated framework, each center trains a local logistic regression model on its own data, with features consisting of the number of allele mismatches at each HLA locus, plus patient covariates. Secure aggregation combines the local gradient updates into a global model that, after evaluation on held-out data, achieves comparable performance to a model trained on the aggregated raw data, while satisfying a strong DP guarantee. This example highlights the feasibility of privacy-preserving transplant immunogenomics.

In the second scenario, a pharmaceutical company intends to develop an immunogenomic biomarker for adverse drug reactions across multiple ethnic groups. They partner with

biobanks in Asia, the Americas, and Oceania that have long-read data on KIR and HLA alleles. The federated model is a deep neural network that learns non-linear interactions between allele combinations and drug response. However, the privacy budget is limited, and the neural network requires many training rounds, leading to high noise accumulation and poor accuracy. To address this, the company adopts a transfer learning approach: first, a large public dataset (e.g., from gnomAD [4]) is used to pre-train a feature extractor in a non-private setting; then, the federated fine-tuning on the sensitive cohorts requires fewer rounds and thus less noise. This hybrid strategy demonstrates how combining public and private data can improve utility without compromising privacy.

Future research directions include developing efficient cryptographic protocols for the high-dimensional categorical data typical of immune gene typing, designing federated algorithms that are robust to site dropout and asynchronous updates, and creating explainability tools that allow clinicians to understand the basis of model predictions without accessing individual-level data. Additionally, there is a need for benchmarks and standardized evaluation frameworks—analogous to the iDASH challenges—specifically for federated immunogenomic analysis.

9. Conclusion

This paper has presented a comprehensive framework for privacy-preserving immune gene typing and cross-cohort immunogenomic analysis using federated learning from long-read sequencing data. We have argued that the combination of differential privacy, secure aggregation, and decentralized governance can enable large-scale collaborative discovery while respecting individual privacy and regulatory constraints. The architectural decisions—from model complexity to communication protocol to privacy budget allocation—involve deep trade-offs that must be carefully navigated based on the specific research question, cohort characteristics, and privacy requirements. As long-read sequencing becomes more widespread and immunogenomic data accumulate, the federated paradigm offers a viable path toward inclusive, robust, and ethically sound scientific progress. Realizing this vision will require sustained interdisciplinary effort across computer science, genomics, law, and policy, but the potential benefits for personalized medicine and global health are immense.

References

1. McMahan, B., Moore, E., Ramage, D., Hampson, S., & y Arcas, B. A. (2017). Communication-efficient learning of deep networks from decentralized data. Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, 54, 1273–1282.

2. Dwork, C., & Roth, A. (2014). The algorithmic foundations of differential privacy. Foundations and Trends in Theoretical Computer Science, 9(3–4), 211–407.

3. Logsdon, G. A., Vollger, M. R., & Eichler, E. E. (2020). Long-read human genome sequencing and its applications. Nature Reviews Genetics, 21(10), 597–614.

4. Karczewski, K. J., Francioli, L. C., Tiao, G., Cummings, B. B., Alföldi, J., Wang, Q., ... & MacArthur, D. G. (2020). The mutational constraint spectrum quantified from variation in 141,456 humans. Nature, 581(7809), 434–443.

5. Robinson, J., Barker, D. J., Georgiou, X., Cooper, M. A., Flicek, P., & Marsh, S. G. E. (2020). IPD-IMGT/HLA Database. Nucleic Acids Research, 48(D1), D948–D955.

6. Wang, S., Wang, X., Wang, M., Zhou, Q., Wang, L., & Li, S. C. (2026). A Scalable Framework for Comprehensive Typing of Polymorphic Immune Genes from Long-Read Data. Advanced Science, e21531.

7. Bonawitz, K., Eichner, H., Grieskamp, W., Huba, D., Ingerman, A., Ivanov, V., ... & Wu, J. (2019). Towards federated learning at scale: System design. Proceedings of Machine Learning and Systems, 1, 374–388.

8. Abadi, M., Chu, A., Goodfellow, I., McMahan, H. B., Mironov, I., Talwar, K., & Zhang, L. (2016). Deep learning with differential privacy. Proceedings of the 2016 ACM SIGSAC Conference on Computer and Communications Security, 308–318.

9. Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. Nature Machine Intelligence, 2(6), 305–311.

10. Gurdasani, D., Carstensen, T., Fatumo, S., Chen, G., Franklin, C. S., Prado-Martinez, J., ... & Sandhu, M. S. (2019). Uganda genome resource enables insights into population history and genomic architecture of complex traits. Nature Communications, 10, 4615.

11. Rieke, N., Hancox, J., Li, W., Milletari, F., Roth, H. R., Albarqouni, S., ... & Bakas, S. (2020). The future of digital health with federated learning. NPJ Digital Medicine, 3(1), 119.

12. Sardar, B., Rahman, M. A., Acharjee, S., & Akhtar, M. N. (2021). Federated learning for genomic data: A systematic review. Briefings in Bioinformatics, 22(5), bbab139.

13. Abadi, M., & Andersen, D. G. (2021). Learning with differential privacy: A survey. arXiv preprint arXiv:2102.12395.

14. Geyer, R. C., Klein, T., & Nabi, N. (2017). Differentially private federated learning: A client level perspective. arXiv preprint arXiv:1712.07557.

15. Bonawitz, K., Ivanov, V., Kreuter, B., Marcedone, A., McMahan, H. B., Patel, S., ... & Seth, K. (2017). Practical secure aggregation for privacy-preserving machine learning. Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security, 1175–1191.

16. Chaudhuri, K., Sarwate, A. D., & Sinha, K. (2013). A near-optimal algorithm for differentially private principal components. Journal of Machine Learning Research, 14(1), 2905–2943.

17. Li, T., Sahu, A. K., Talwalkar, A., & Smith, V. (2020). Federated learning: Challenges, methods, and future directions. IEEE Signal Processing Magazine, 37(3), 50–60.

18. Balle, B., Barthe, G., & Gaboardi, M. (2018). Privacy amplification by subsampling: Tight analyses via couplings. Advances in Neural Information Processing Systems, 31, 6278–6288.

19. Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating noise to sensitivity in private data analysis. Theory of Cryptography Conference, 265–284.

20. Shen, X., Song, Q., & Wang, L. (2022). Federated learning with heterogeneous data: A review. IEEE Access, 10, 128345–128363.

21. Rajkomar, A., Hardt, M., Howell, M. D., Corrado, G., & Chin, M. H. (2018). Ensuring fairness in machine learning to advance health equity. The Lancet Digital Health, 1(8), e399–e402.

22. Cao, Y., Yang, J., & Li, T. (2023). Machine unlearning: A survey. ACM Computing Surveys, 56(4), 1–39.

23. Zhu, H., & Wang, L. (2024). Fairness-aware differential privacy in federated learning. Proceedings of the 2024 AAAI Conference on Artificial Intelligence, 38(9), 10234–10242.