

BioSpect

the business of Bio

24 years of leveraging the business of Bio
www.biospectrumindia.co

Whose Genome Is It?

04 June 2026 | Features | By Ankit Kankar | ankit.kankar@mmactiv.com

Data Sovereignty vs. the AI Data-Hunger: Asia's Genomic Tug-of-War



An analysis of how national biobanks, security politics, and the scaling logic of modern AI are pulling Asia's genomic future in opposite directions.

A note on sourcing: the expert "voices" quoted below — a biobank director, a data-governance lawyer, an AI researcher, and a bioethicist — are illustrative composites. They are written to represent positions that researchers, regulators, and ethicists actually hold in this debate, drawn from the public record, rather than to reproduce statements from specific named individuals. All institutional facts, biobank figures, and statutes are real and sourced.

In January 2025, the Indian government announced that it had finished sequencing the genomes of just over ten thousand citizens — 10,074, drawn from 99 distinct communities — and deposited the data in a government repository called the Indian Biological Data Centre. The prime minister called it a defining moment for the country's biotechnology landscape. For a nation of more than 4,600 endogamous population groups, the ambition was explicit: build a genetic reference that actually reflects Indians, who remain badly underrepresented in the global databases that train the world's most powerful biomedical models.

The achievement was real. So was the contradiction buried inside it. The same logic that makes a sovereign Indian genome bank valuable — that the world's reference data is too European, too narrow, and too incomplete to serve everyone — is the logic that makes the data scientifically useless if it stays locked inside one country's servers. Diversity is precisely the thing that has to be pooled to matter. And pooling is precisely the thing that data-sovereignty law, national-security politics, and a decade of accumulated mistrust are now built to prevent.

That is the tug-of-war playing out across Asia. On one side: the conviction that a population's genomes are a strategic national asset, to be guarded like a power grid or a semiconductor fab. On the other: the brute statistical reality that AI-driven biology gets better with scale and diversity, and that the most useful models are trained on data that crosses borders. Patient privacy sits in the middle, invoked by everyone and protected by no one in particular.

The biobank boom

The first thing to understand is how much genomic infrastructure Asia has built, and how fast.

Japan got there earliest. BioBank Japan, founded at the University of Tokyo in 2003, now holds DNA from roughly 270,000 patients across 51 common diseases — one of the largest disease-oriented biobanks in the world, and a deliberate corrective to a field where most genome-wide studies were run on people of Western descent. A second Japanese effort, the Tohoku Medical Megabank, was launched after the 2011 earthquake and tsunami as both a public-health project and an act of regional reconstruction; it has recruited about 150,000 participants and stores millions of biospecimen tubes alongside multi-omic and clinical data.

Singapore moved later but with characteristic precision. Its National Precision Medicine program began in 2017, sequenced 10,000 genomes as a proof of concept, then completed a second phase — PRECISE-SG100K — that sequenced 100,000 whole genomes from the city-state's Chinese, Malay, and Indian populations. A third phase, launched in late 2025 and running to 2031, aims to enroll up to 450,000 people and push genomic results into routine clinical care. Singapore's pitch is unusual: its three founding ethnicities are estimated to capture something like 80 percent of Asian genetic variation, which lets a small country claim outsized scientific relevance and position itself as a bridge between Western regulatory models and Asian precision medicine.

China built at a different scale entirely. Beyond the half-million-participant China Kadoorie Biobank — itself a long-running collaboration with Oxford — China has poured state resources into genomic sequencing capacity, and its domestic champions in the sector grew into global players. India's Genome India project, though small in raw numbers at around 10,000 genomes, is best understood as a first reference catalogue for a population that the rest of the world's databases barely see.

Put together, the map looks like a boom. But each of these programs was built inside a national container, and the walls of those containers are getting higher.

"Every health minister in the region now understands the same two sentences," says a composite biobank director who has overseen one of these national programs. "First, our population's data is scientifically precious because nobody else

has it. Second, that is exactly why we cannot let it leave. The trouble is those two sentences point in opposite directions, and I'm the one standing where they collide."

The laws that build the walls

The sovereignty instinct has been written into statute across the region, in overlapping and not always consistent layers.

China has the most developed regime. Its 2019 Regulation on Human Genetic Resources requires a national security review before genetic information can be exported or opened to a foreign party — with particular scrutiny for data drawn from important genetic families, from designated geographic populations, or from the exome or genome sequencing of more than 500 individuals. Layered on top are the 2021 Personal Information Protection Law (PIPL), which treats genetic data as sensitive personal information and demands a "consent-plus" model for cross-border transfer, and the Data Security Law, which classifies and protects data according to its importance to national security. The Cyberspace Administration's 2024 provisions on cross-border data flows added yet another gate. The cumulative effect, as legal scholars have noted, is a system where the dominant feature is centralized state control, with individual privacy a genuine but secondary concern.

India's framework is younger and still settling. The Digital Personal Data Protection Act of 2023 governs personal data broadly, while Genome India's output sits inside the government-run Indian Biological Data Centre and is released to researchers through a managed access framework rather than open download. Japan governs health data under its Act on the Protection of Personal Information; Singapore under its Personal Data Protection Act, with the state acting as the central steward of the national genomic database. Each country has its own definition of consent, its own export controls, its own list of what counts as "sensitive."

"From a lawyer's chair, the problem is not that any one of these laws is unreasonable," says a composite data-governance attorney who advises life-sciences clients across Asia. "It's that they don't interoperate. A consortium that wants to train one model across four jurisdictions has to satisfy four incompatible consent standards, four export-review processes, and four definitions of de-identification — and in China the distinction between de-identified and truly anonymous data matters enormously, because under good-clinical-practice coding the data is merely de-identified. You can spend two years on the paperwork and never touch the science."

Genomes as a national-security asset

What turned a privacy conversation into a security one was the growing conviction, in capitals well beyond Asia, that aggregated human biological data is a strategic resource on the order of advanced chips or AI models themselves.

The clearest signal came from Washington. On December 18, 2025, the United States enacted the BIOSECURE Act as part of the fiscal-year 2026 National Defense Authorization Act. The law restricts federal agencies, contractors, and grant recipients from procuring equipment or services from "biotechnology companies of concern," using the Defense Department's existing list of Chinese military-linked companies as its trigger; China's leading genomics firms appear on that list, and reporting suggests prominent contract-research organizations may be added. An earlier draft had named specific companies outright. The reframing is the point: as one analysis put it, lawmakers are signaling that human biological data should be treated like semiconductors or AI weights — something whose control determines geopolitical trustworthiness.

The same month, the European Commission proposed an EU Biotech Act framing biotechnology in terms of strategic autonomy, and the US Department of Justice stood up a Data Security Program targeting bulk sensitive data flows to adversary states. Multi-omic health data — genome, proteome, metabolome, linked to clinical records — sits squarely inside this new category of "strategic data."

The security logic is not paranoid. Genomic data is uniquely sticky: it identifies not just an individual but their relatives, it cannot be revoked or reissued like a password, and at population scale it can reveal things about a nation's health vulnerabilities. A government that decides this data is a national asset is making a defensible call. The difficulty is that "defensible" and "scientifically optimal" are not the same thing, and the gap between them is where the cost lands.

The science cost

Here is the part that rarely makes the policy debate: localization has a measurable price, and AI is what makes the price visible.

Modern AI-driven biology — variant-effect predictors, polygenic risk models, the large genomic foundation models now being trained — improves with two things above all: the sheer volume of data and its diversity. A model trained overwhelmingly on European genomes makes worse predictions for everyone else, and the underrepresentation is not marginal. The entire rationale for Genome India, for BioBank Japan's East Asian focus, and for Singapore's multi-ancestral cohort is to fix that imbalance. But fixing it requires the data to be *combined* — and combination is exactly what sovereignty law throttles.

"People think the bottleneck for AI in genomics is algorithms or compute," says a composite AI researcher who builds population-scale models. "It isn't. It's access to enough diverse genomes in one analyzable place. I can have ten national biobanks, each beautifully curated, and still be unable to train the model the science needs, because I'm not allowed to put them in the same room. A model that could have learned from two million diverse genomes instead learns from a hundred thousand from one ancestry. The accuracy gap that produces doesn't show up as a headline. It shows up later, as worse risk scores for the populations the data was supposed to help."

The cost is not hypothetical. Cross-border genomic collaborations have stalled or shrunk under the weight of export review and mismatched consent regimes; researchers in the field have warned for years that China's tightened human-genetic-resources rules, whatever their merits, deter the very international projects that would surface East Asian-specific variants. The irony is sharp. The data-sovereignty movement is partly a reaction to legitimate grievances — decades of "helicopter research" in which Western institutions extracted samples from the Global South and kept the benefits. But the remedy, taken to its logical end, reproduces the underrepresentation it set out to cure, just with the walls relocated.

And the stakes are rising, not falling, because the data itself is getting richer. The newest national programs are no longer just sequencing DNA. Singapore's SG100K has layered population-scale proteomics on top of its 100,000 genomes; Japan's megabanks have built multi-omic reference panels spanning the metabolome, proteome, and transcriptome. This multi-omic, clinically linked data is far more powerful for AI — and far more sensitive, and far more tightly regulated. Every increment in scientific value is an increment in security anxiety. The richer the dataset becomes, the more a government wants to wall it off, and the more a model would gain from combining it with others. The two curves diverge exactly as the science gets interesting.

"There's a real ethical claim under the sovereignty argument, and we shouldn't wave it away," says a composite bioethicist. "Communities have been mined for data and given nothing back. Sovereignty is, in part, a demand for benefit-sharing and dignity. But sovereignty can curdle into something else — a state asserting that *it* owns its citizens' genomes, which is a very different claim from a patient controlling their own. When the security framing takes over, the patient in the middle stops being a rights-holder and starts being a resource. That's the slide we have to watch."

Three use cases, three pressures

Consider how these forces play out concretely.

A national program. Genome India and Singapore's SG100K are sovereignty done well: state-funded, ethically reviewed, designed to correct genuine gaps, with data held in managed national repositories. They demonstrate capacity and produce real reference value. What they cannot do alone is reach the scale at which the most ambitious AI models become possible. A national reference catalogue is a foundation, not a finished building.

A cross-border collaboration. Now imagine a consortium that wants to combine Indian, Japanese, and Singaporean cohorts to study a disease with strong genetic components across Asian populations — the kind of project where pooling is the whole point. It immediately hits a wall of incompatible export rules and consent definitions. In China's case, any genome-sequencing dataset above 500 subjects triggers security review before it can be shared abroad at all. The collaboration doesn't get refused so much as slowed into irrelevance, its scientific window closing while the lawyers work. This is the documented pattern that has chilled international genomic research in the region.

A federated consortium. The third path tries to escape the binary. Instead of moving the data to the model, it moves the model to the data. Picture the same Asian-disease study, rebuilt: the Indian, Japanese, and Singaporean genomes never leave their respective national repositories. A shared model travels to each site, trains locally behind each country's firewall, and sends back only mathematical updates — no raw sequences, nothing that crosses an export-review threshold. Each government can audit exactly what leaves its borders, because what leaves is gradients, not genomes. The science

gets something close to the combined dataset; the regulators get to keep custody. It is not frictionless, but it is the rare design where the scientist and the security official can both sign off.

Models for safe sharing

The technical community has spent the better part of a decade building ways to collaborate without surrendering raw data, and the toolkit is now mature enough to be policy-relevant.

Federated learning is the headline approach. Rather than copying genomes into a central pool, each biobank keeps its data behind its own walls; a shared model is sent out, trained locally on each dataset, and only the model updates — not the underlying genomes — travel back to be combined. Studies on the UK Biobank and the 1000 Genomes Project have shown federated training can approach the accuracy of centralized training in the "cross-silo" setting that biobanks naturally occupy: a few institutions, each holding a lot of data, each unwilling or legally unable to hand it over. For a region of walled national programs, this maps almost perfectly onto the political reality.

Federation alone is not a privacy guarantee — model updates can leak information about the data that produced them — so it is increasingly paired with stronger protections: *differential privacy*, which adds calibrated noise so no individual can be reverse-engineered from the output; *secure multi-party computation* and *homomorphic encryption*, which let parties compute joint results over data none of them can see in the clear; and *trusted execution environments*, secure enclaves where analysis runs without exposing inputs. The Global Alliance for Genomics and Health has built a federated discovery layer — the Beacon network — that lets researchers ask whether a given variant exists in a dataset without exposing identifiable records.

Beyond the cryptography sit governance models. *Data trusts* and managed-access repositories — the model India already uses with its Biological Data Centre — let an independent steward grant vetted researchers controlled access under enforceable terms, rather than choosing between open download and total lockdown. And underneath all of it, *consent frameworks* have to evolve from one-time, all-or-nothing checkboxes toward dynamic, revocable consent that lets participants see and shape how their genomes are used, including in cross-border analysis.

"Federated learning is not magic, and anyone selling it as a silver bullet is overselling," cautions the composite AI researcher. "It's slower, it's harder to debug, and you still need real privacy guarantees on top. But it changes the political question. Today a minister has to choose between 'protect our data' and 'do the science.' Federation lets the answer be 'both,' and that's the only answer that survives contact with the politics."

The bioethicist puts the governance point more bluntly: "The technology can make sharing safe. It cannot make it *fair*. If a federated model trained on a dozen countries' genomes turns into a drug that those countries can't afford, we've solved the privacy problem and kept the colonialism. Benefit-sharing has to be written into the consortium agreement, not bolted on after the paper publishes."

Can sovereignty and science coexist?

The honest answer is that they can, but only if the region stops treating them as a zero-sum choice — and the current trajectory does not encourage optimism.

The momentum is toward higher walls. Security framing is ascendant; the BIOSECURE era has made "biological data as strategic asset" the default vocabulary in Washington and increasingly in Brussels, and Asian governments that built sovereign biobanks have every incentive to read that vocabulary back into their own export controls. Each new statute is individually reasonable and collectively suffocating. Left alone, the likely outcome is a continent of well-curated, scientifically isolated national datasets — sovereignty achieved, scale forfeited, and the underrepresentation that justified the whole enterprise quietly preserved.

The alternative requires three things to land at once. The technical layer is the easy part: federated learning, plus differential privacy and secure computation, is ready. The governance layer is harder: interoperable consent standards, mutual recognition of de-identification rules, and data-trust structures that let stewards say "yes, under conditions" instead of "no." And the political layer is hardest of all: governments have to decide that participating in a federated, privacy-preserving consortium counts as keeping their data sovereign — that sovereignty means *control over the terms of use*, not physical custody of every byte.

That redefinition is the whole game. If sovereignty is understood as custody, it is fundamentally incompatible with the collaborative scale that AI-driven biology needs, and the science loses. If sovereignty is understood as control — the right to set the terms, to enforce benefit-sharing, to revoke access, to keep raw genomes home while still contributing to a shared model — then it is not just compatible with the science but could become the framework that finally makes diverse, global genomics both possible and just.

Asia is the place where this will be decided, because Asia is where both the data and the walls are being built fastest. The biobanks are real. The laws are real. The security anxieties are real, and not always unfounded. Whether the region can hold all three without throttling the science depends on a single unresolved question — the one the patient in the middle has the most stake in and the least say over.

Whose genome is it, exactly? Until that question has a better answer than "the state's," the tug-of-war continues, and the data stays home.