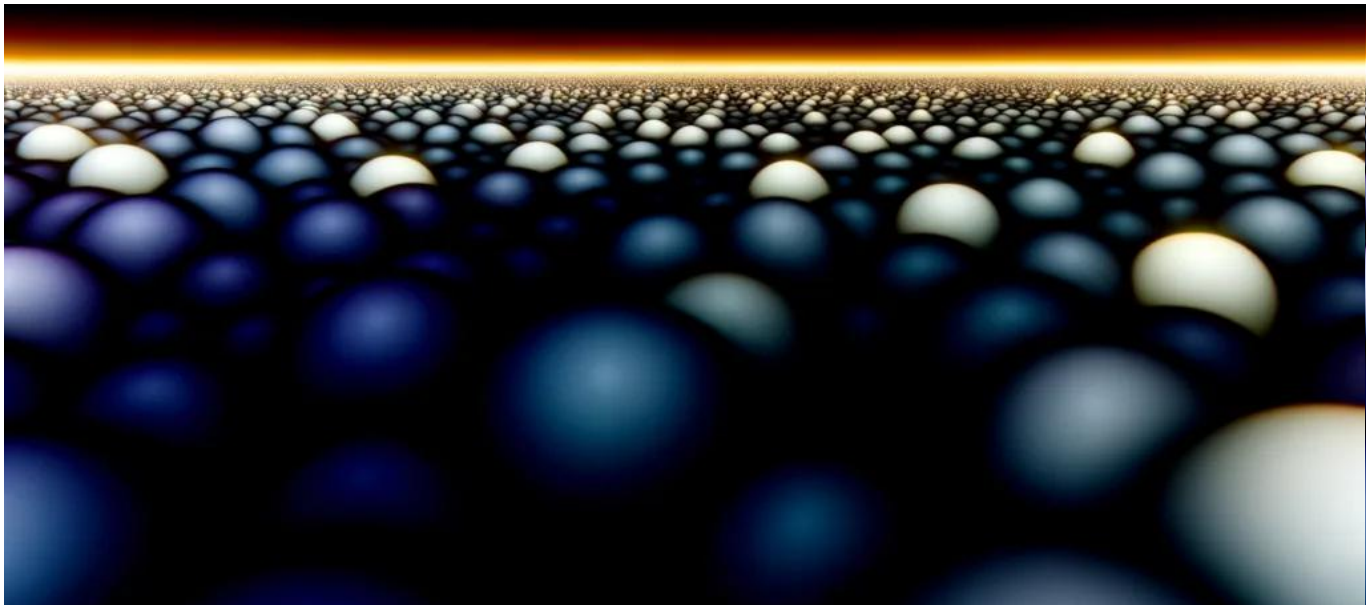


Illumina Launches 1 Billion Cell Atlas: 20 Petabytes of CRISPR Data Built with AstraZeneca, Merck, and Lilly to Train Next-Gen Drug Discovery AI

Illumina just gave AI drug discovery something it's never had: ground truth data for what happens when you toggle every human gene on and off across a billion cells. Three pharma giants are betting this solves the genetic validation problem that's been quietly killing drug candidates.

The Announcement: What Illumina Actually Built

On January 13, 2026, at the J.P. Morgan Healthcare Conference in San Francisco, Illumina unveiled the Billion Cell Atlas—the world's largest genome-wide genetic perturbation dataset. The numbers are staggering: 1 billion individual cells profiled, 20 petabytes of single-cell transcriptomic data, and all 20,000 human genes systematically perturbed

using CRISPR across 200+ disease-relevant cell lines.

This isn't an academic proof-of-concept. AstraZeneca, Merck (MSD), and Eli Lilly signed on as founding pharma partners, funding what Illumina describes as the first phase of a planned 5-billion cell resource to be completed over three years. The dataset covers immunology, oncology, cardiometabolic conditions, neurological disorders, and rare genetic diseases—essentially the full spectrum of where pharma R&D dollars flow.

The Billion Cell Atlas represents the first commercial product from Illumina's new BioInsight business unit, which focuses specifically on building multiomics datasets for pharmaceutical AI applications. The technical stack runs on Illumina's Single Cell 3 RNA prep platform, processes through the DRAGEN pipeline with hardware acceleration, and lives on Illumina Connected Analytics for cloud-based analysis at scale.

Why Pharma Paid for This: The Target Validation Crisis

Here's the dirty secret of AI-driven drug discovery: the algorithms work, but they're training on incomplete data. Machine learning models can predict protein structures, identify binding sites, and generate novel molecular candidates with impressive accuracy. What they can't do is tell you whether hitting a particular target will actually treat the disease.

90% of drug candidates fail in clinical trials, and poor target selection remains the leading cause. Computational approaches can identify thousands of potential drug targets from genomic data, but validating which ones actually matter in disease biology requires experimental evidence that mostly doesn't exist at scale.

According to coverage from Pharmaceutical Technology, this is precisely the gap the Billion Cell Atlas addresses. By systematically turning each of the 20,000 human genes on and off via CRISPR and measuring the transcriptomic response across disease-relevant cell types, the dataset creates a functional map of genetic cause and effect.

Think of it this way: previous approaches told you that Gene X is associated with Disease Y based on statistical correlation in patient populations. The Billion Cell Atlas shows you what actually happens to cellular behavior when you manipulate Gene X directly. Association becomes mechanism.

For pharma companies, this translates to a simple value proposition: fewer expensive failures in late-stage clinical trials. If you can validate or invalidate a target hypothesis before committing $500 million to Phase III trials, even a modest improvement in success rates pays for the dataset many times over.

The Technical Architecture: How You Profile a Billion Cells

Generating 20 petabytes of meaningful biological data in one year requires solving several hard engineering problems simultaneously. The technical stack Illumina deployed deserves examination because it illustrates where the field is heading.

CRISPR Perturbation at Scale

Traditional CRISPR screens work by creating a library of guide RNAs, each targeting a different gene, and introducing them into cell populations. You then sequence the surviving cells to see which genetic knockouts affected some phenotype of interest. This approach has been standard for years, but it treats each cell as a binary data point—the gene was knocked out, and the cell either survived some selection pressure or it didn't.

The Billion Cell Atlas takes a different approach. Instead of just measuring survival, Illumina captures the full single-cell transcriptome—the expression levels of all genes—for each perturbed cell. This creates a high-dimensional readout showing how disrupting one gene affects the expression of thousands of others.

The combinatorial math is brutal. 20,000 genes × 200+ cell lines × multiple cells per condition × ~20,000 genes measured per cell = datasets that would have been computationally intractable a few years ago. Illumina's solution relies heavily on their DRAGEN pipeline, which uses field-programmable gate arrays (FPGAs) for hardware-accelerated sequence alignment and variant calling. **DRAGEN reduces processing time by roughly 90% compared to software-only pipelines, making this scale of analysis economically viable.**

Single-Cell Resolution Matters

Why single-cell? Because bulk sequencing—measuring average gene expression across

millions of cells—masks the biological heterogeneity that determines drug response. A tumor isn't one cell type; it's a ecosystem of cancer cells, immune cells, stromal cells, and others. A drug that affects the average expression profile might work brilliantly on one subpopulation and have no effect on another.

Single-cell transcriptomics preserves this information. When you perturb Gene X and measure responses at single-cell resolution, you can identify which cell types respond, how much variation exists within responding populations, and whether certain cells escape the perturbation entirely. For drug discovery, this means predicting not just whether a target matters, but in which cellular context it matters.

The 3□ RNA capture approach Illumina uses (Single Cell 3□ RNA prep) offers a practical tradeoff: you sacrifice full-length transcript information for higher throughput and lower cost per cell. For most target validation applications, this tradeoff makes sense—you're measuring gene expression levels, not transcript isoforms.

Cloud Infrastructure at Petabyte Scale

20 petabytes is not a number you can casually analyze on local hardware. Illumina hosts the dataset on their Connected Analytics platform, which provides cloud-based access for partner organizations. This raises immediate questions about data governance, query costs, and computational reproducibility that enterprise users will need to evaluate.

The platform supports both pre-computed analyses (what did knocking out BRCA1 do across all cell lines?) and custom queries (show me all perturbations that upregulated IL-6 in CD4+ T cells by more than 2-fold). The former is straightforward; the latter could become expensive quickly at this data volume. Illumina hasn't publicly detailed the pricing model, but pharma partners presumably negotiated access terms as part of their founding agreements.

What Most Coverage Gets Wrong

The press releases and initial coverage frame this primarily as an "AI training dataset," and while that's not inaccurate, it undersells the more immediate value and oversells the AI angle.

The Immediate Value Is Lookup, Not Training

BioPharma Trend's coverage emphasizes the AI training applications, but the most straightforward use case requires no machine learning at all: direct lookup. A pharma company with a target hypothesis can query the database to see exactly what happens when that gene is perturbed in relevant cell types. This is experimental validation by database query, not algorithmic inference.

That's not a small thing. Running a comparable CRISPR screen in-house takes months and costs hundreds of thousands of dollars. If the Billion Cell Atlas already contains your gene of interest in your cell type of interest, you get the answer immediately. The 200+ cell lines cover enough disease-relevant biology that many common target validation questions will have direct hits.

The AI Training Story Is More Complicated

Yes, this dataset will train better AI models for drug discovery. But treating single-cell perturbation data as just another training corpus misses some nuances.

First, the data is high-dimensional but narrow in certain ways. All measurements come from in vitro cell lines, not primary patient cells or in vivo models. Cell lines are immortalized, often aneuploid, and don't perfectly recapitulate disease biology. An AI model trained purely on cell line data may learn artifacts that don't translate to clinical reality.

Second, the perturbations are all genetic. Drugs don't work by perfectly knocking out or overexpressing genes—they partially inhibit proteins, have off-target effects, and vary in bioavailability across tissues. The mapping from "gene knockout" to "drug effect" is informative but not identical.

Third, foundation models for biology are still immature compared to language or vision. We don't have the equivalent of GPT-4 for cellular biology, and it's unclear whether scale alone will produce one. The Billion Cell Atlas provides better training data, but the architectures that can effectively use this data are still evolving rapidly.

The honest framing: this dataset is necessary but not sufficient for AI-driven drug discovery. It fills a critical gap in experimental evidence, but it won't magically make AI

predict successful drugs.

The Competitive Dynamics Are Interesting

Illumina's move here is partly defensive. The company has dominated sequencing hardware for two decades, but the center of value in genomics is shifting from raw sequencing to data analysis and interpretation. By creating a unique dataset that only Illumina can generate at this scale (thanks to their installed base and platform integration), they're building a moat in the data layer that doesn't depend solely on continued hardware dominance.

The founding partner structure is also notable. AstraZeneca, Merck, and Lilly get early access and presumably preferential terms, but they're also funding a resource that their competitors will eventually access. This suggests the pharma companies believe the rising tide of better target validation will lift all boats more than it creates competitive advantage for dataset contributors. That's a revealing bet about where differentiation in drug discovery actually comes from.

Practical Implications for Technical Leaders

If you're building or evaluating AI systems for life sciences, here's what the Billion Cell Atlas means for your architecture and vendor decisions.

Data Integration Becomes Critical

The Billion Cell Atlas is one dataset among many. Useful AI systems for drug discovery will need to integrate this perturbation data with:

- Protein structure predictions (AlphaFold, ESMFold)
- Small molecule binding data (ChEMBL, PubChem)
- Clinical outcomes data (electronic health records, trial results)
- Literature-derived knowledge graphs
- Proprietary internal datasets from pharma partners

None of these data sources share common identifiers, ontologies, or schemas. The

unglamorous work of entity resolution, data normalization, and maintaining provenance tracking will determine whether organizations can actually use resources like the Billion Cell Atlas effectively.

If you're building in this space, invest in your data engineering layer. The models are increasingly commoditized; the data pipelines are not.

Compute Costs Will Surprise You

Running meaningful analyses against 20 petabytes of data requires significant cloud compute resources. Before committing to Illumina Connected Analytics or any similar platform, model your actual query patterns and estimate costs carefully.

Some questions to ask:

- What's the cost per query for different analysis types?
- Can you export subsets of data to your own infrastructure?
- What's the latency for interactive versus batch analyses?
- How are compute costs allocated in multi-tenant environments?

The platform economics of biological databases often look attractive at small scale and become painful at production scale. Get clarity on pricing before building dependencies.

Benchmark Your Models Against This Data

If you're developing AI models for target identification, gene function prediction, or cellular response modeling, the Billion Cell Atlas provides a ground truth benchmark that didn't previously exist at this scale.

Consider designing evaluation protocols that test whether your models can predict the transcriptomic effects of gene perturbations before looking at the experimental data. If your model claims to understand gene function, it should be able to predict what happens when genes are knocked out. If it can't, you've learned something important about its limitations.

This is harder than it sounds—most current models were trained on datasets that predate

systematic perturbation screens, so they may have memorized correlational patterns without learning causal mechanisms. The Billion Cell Atlas provides a way to distinguish these failure modes.

Watch the Open-Source Response

Illumina is making this dataset available through their commercial platform, but the life sciences community has a strong tradition of open data sharing. Within 12-18 months, expect to see:

- Academic groups generating comparable (if smaller) datasets and releasing them publicly
- Open-source tools for analyzing perturbation data that don't require Illumina's platform
- Foundation models pre-trained on combinations of public and newly generated perturbation data

The Human Cell Atlas and related consortia have already produced significant single-cell datasets. Adding perturbation data to these resources is a natural extension that funding agencies will support.

For technical strategy, this means avoiding over-investment in proprietary platform lock-in while the landscape is still evolving. The first-mover advantage of accessing the Billion Cell Atlas early may be significant for pharma companies racing to validate specific targets, but for platform builders, maintaining flexibility is worth more than early access.

Where This Leads: The 6-12 Month Horizon

The Billion Cell Atlas is phase one of a three-year roadmap to 5 billion cells. By early 2027, expect Illumina to announce expanded cell line coverage, additional disease areas, and potentially new data modalities beyond transcriptomics.

Multimodal Becomes Mandatory

Transcriptomics captures gene expression, but cells have other measurable properties:

protein levels (proteomics), chromatin accessibility (ATAC-seq), metabolite concentrations (metabolomics), and spatial organization within tissues (spatial transcriptomics). Illumina's BioInsight unit is explicitly focused on "multiomics," suggesting future Atlas expansions will layer additional measurement types onto the existing transcriptomic foundation.

For AI model development, multimodal biological data presents similar challenges to multimodal AI in other domains—different data types have different scales, noise characteristics, and semantic meanings. Architectures that effectively fuse transcriptomic, proteomic, and spatial data will outperform single-modality approaches, but designing such architectures remains an open research problem.

Competitive Datasets Will Emerge

Illumina has a significant head start, but they won't have the only large-scale perturbation dataset for long. Potential responses include:

- **10x Genomics**: Already a major player in single-cell analysis, 10x has the technical capability to generate comparable datasets and might do so to protect their position in the pharma data market.
- **Academic Consortia**: NIH, Wellcome Trust, and other funding bodies may support open-access alternatives, particularly if the Billion Cell Atlas pricing becomes prohibitive for academic researchers.
- **Pharma Internal Efforts**: Large pharma companies generate substantial internal CRISPR screening data. If data sharing norms shift, pooled datasets from multiple companies could rival Illumina's scale.

Regulatory Frameworks Will Lag

Using AI-derived insights for drug target selection raises regulatory questions that agencies haven't fully addressed. If a company validates a target primarily through AI analysis of the Billion Cell Atlas (rather than traditional wet-lab experiments), how should regulators evaluate the evidence quality? Current FDA guidance on AI in drug development focuses primarily on clinical decision support, not preclinical target selection.

Expect 2026-2027 to bring increased regulatory attention to AI-driven target validation, potentially including new guidance documents or pilot programs. Companies building in

this space should engage with regulatory affairs teams early to understand how AI-derived evidence will be treated in IND applications.

The Business Model Implications

Illumina's strategic bet with BioInsight is that data products can become a meaningful revenue stream alongside hardware and consumables. If the Billion Cell Atlas gains traction, expect Illumina to expand the model—potentially offering customized datasets, exclusive early access tiers, or API-based pricing for algorithmic queries.

For pharma companies, the interesting question is whether shared foundational datasets like this reduce R&D costs enough to offset the loss of proprietary data advantages. The willingness of AstraZeneca, Merck, and Lilly to co-fund a shared resource suggests the answer is yes, at least for certain data types. Target validation is expensive enough that collective investment makes sense even among competitors.

For AI-native drug discovery companies (Recursion, Insitro, and the growing list of startups), the Billion Cell Atlas is double-edged. It provides valuable training data, but it also potentially commoditizes one source of competitive advantage—proprietary perturbation screens. Companies that differentiated on data scale may need to shift differentiation toward model architecture, vertical integration, or speed of iteration.

The Bottom Line

The Billion Cell Atlas doesn't solve AI drug discovery. It solves one specific problem—the absence of systematic experimental evidence for what individual genes do across disease-relevant cell types—and it solves that problem at unprecedented scale.

For CTOs evaluating AI investments in life sciences: this dataset exists, it's accessible (for a price), and it will become a standard benchmark for any model claiming to understand cellular biology. Your architecture decisions should assume this data exists and plan for how you'll integrate or compete with it.

For senior engineers building biological AI systems: the data engineering challenges here are substantial and underappreciated. Entity resolution, provenance tracking, and efficient

querying against petabyte-scale biological databases are not solved problems. Whoever solves them will capture significant value.

For founders in the drug discovery space: the landscape is shifting toward shared foundational resources with proprietary advantages concentrated in model development and clinical translation. Pure data plays are getting harder; pure algorithm plays were always hard. The winners will integrate across the stack.

The Billion Cell Atlas is a $20-petabyte proof that AI drug discovery has matured from "we need better algorithms" to "we need better data"—and that pharmaceutical companies are now willing to fund the infrastructure to get it.



WRITTEN BY

Artur Markus

Software engineer in Basel building AI infrastructure and agentic systems, with a background in pharmaceutical automation, GxP compliance, and validation. I help teams turn AI from a chatbot into a reliable operational layer.

Book a meeting