bitBiome

Data Engineer / Biology × Machine Learning

Unique Bio Data, Next-Gen AI

Job Information

Hiring Company
bitBiome, Inc.

Job ID
1621941

Division
R&D

Industry
Other

Company Type
Small/Medium Company (300 employees or less)

Non-Japanese Ratio
Majority Japanese

Job Type
Permanent Full-time

Location
Tokyo - 23 Wards, Shinjuku-ku

Train Description
Tozai Line, Waseda Station

Salary
6 million yen ~ 12 million yen

Work Hours
基本9：00 ~ 18：00 / 専門業務型裁量労働制：1日の労働時間 8時間とみなす

Holidays
完全週休2日制

Refreshed
November 11th, 2026 16:20

General Requirements

Minimum Experience Level
Over 3 years

Career Level
Mid Career

Minimum English Level
Business Level (Amount Used: English usage about 50%)

Minimum Japanese Level
Business Level

レポートラインのUSチームとは英語 / 日本のチームメンバーとは日本語を使ってのコミュニケーションになります

Minimum Education Level
Post Grad Degree (PHD/MBA etc)

Visa Status
Permission to work in Japan required

Job Description

Position Highlights & What Makes This Opportunity Unique

- Work with one-of-a-kind, proprietary large-scale biological data, comprising 3.5 billion genes.
- Experience the intersection of biological experiments and AI—something you won't find at a typical IT company. See the real-world impact of your work as the data and models you develop help create new enzymes and microbial strains with real-world functions.
- Go beyond building data pipelines and get involved in AI modeling.
- Work in a small, highly skilled team with a global mindset and international standards.

Compensation

Annual Salary: JPY 6,000,000–12,000,000

Mission

You will be responsible for the end-to-end design and operation of the data layer that underpins our machine learning models.

On the genomics side, you will track genotype changes across natural strains and engineered variants. On the laboratory side, you will transform wet-lab data generated through bit-EVO's DBTL (Design-Build-Test-Learn) cycles—including sequencing reads, plate assay results, and omics data—into labels suitable for model training.

You will then integrate these two data sources to build and maintain high-quality datasets for model training and evaluation, continuously updating them as new experimental results become available.

You will also play a key role in advancing our modeling capabilities. Working closely with laboratory researchers and AI scientists, you will identify data-related bottlenecks that limit model performance, propose better ways to represent genomic information and laboratory measurements, and contribute to the design and validation of new model architectures.

Key Responsibilities

1. Data Pipeline Development

- Track genotypic differences across strains and engineered variants, including SNPs, indels, gene insertions/deletions, and plasmid configurations. Maintain accurate links between each phenotype and the exact genotype responsible for it.
- Collect and curate phenotypic labels from internal assays, such as target compound production levels and titers, enzyme activity, and growth rates.
- Collect and curate omics data (transcriptomics and metabolomics) with consistent processing, metadata management, and quality control, and link these datasets to genomic data.
- Expand internal benchmark suites covering DNA, protein, and whole-genome tasks.

2. Collaboration with Laboratory Researchers

- Work closely with laboratory researchers to define naming conventions, recording practices, and linking rules for samples, strains, plates, and assay results, ensuring full traceability of every data point back to its original experiment.
- Design metadata standards and data templates for wet-lab experiments, and support their adoption in the laboratory's LIMS (Laboratory Information Management System).
- Develop simple, user-friendly tools that enable researchers to upload, validate, and explore data independently, without requiring engineering support.

3. Infrastructure and Data Quality Assurance

- Manage integrations between the LIMS and data platform, building and maintaining infrastructure that ensures end-to-end ID traceability for samples, plates, strains, and assay results as data flows from the laboratory into training datasets.
- Ensure the reproducibility of all datasets through version-controlled inputs, pinned tool environments, data cards, and data lineage tracking from raw files to training-ready datasets.

- Execute and manage optimized workloads on cloud-based GPU/CPU clusters.

Department: R&D – Computational Biology

Reporting Line: Principal AI Scientist (based in California, USA)

Team Structure:

- 1 Head of Innovation (based in the US)
- 1 Principal AI Scientist (based in the US)
- 2 team members (based at the Japan headquarters)

About bitBiome

Most microorganisms on Earth have never had their DNA sequenced. bitBiome, a startup originating from Waseda University, brings together proprietary machine learning, genomics, and wet-lab capabilities under one roof to unlock this unexplored microbial diversity and turn it into sustainable products.

Why Join bitBiome?

- An open and collaborative working environment where employees can share ideas and take initiative, characteristic of a growing startup.
- An opportunity to experience both business growth and personal development in a dynamic environment.
- A welcoming and focused workplace with an average employee age of 38 and a gender ratio of approximately 60:40.

Our Core Technologies & Platforms

- **bit-MAP®:** A proprietary technology that sequences microorganisms at the single-cell level, enabling highly accurate genomic analysis beyond the capabilities of conventional methods.
- **bit-GEM:** A database containing over 3 billion microbial genes assembled through bit-MAP, many of which have never been identified in existing public databases. We use this resource to discover novel genes and valuable enzymes, and to train models that link genotypes with phenotypes.
- **bit-EVO:** A large-scale, highly scalable DBTL (Design-Build-Test-Learn) cycle that continuously incorporates experimental results from the wet lab to improve the accuracy of subsequent designs.

You will join a small, agile team of researchers and computational scientists with a fast decision-making process. The data infrastructure and workflows you build will serve as the foundation for training our AI models.

Working Hours

- Standard working hours: 9:00 AM–6:00 PM (8 working hours per day)
- Discretionary Labor System for Professional Work (deemed working hours: 8 hours per day)
- Break: 60 minutes
- Overtime: May be required

Initial Work Location:

Waseda University Research Innovation Center, Building 121
513 Waseda-Tsurumakicho, Shinjuku-ku, Tokyo, Japan

Access: 2-minute walk from Waseda Station (Tokyo Metro Tozai Line)

Potential Changes to Work Location: Employee's home

Remote Work Policy:

Employees are expected to work from the office during their first approximately three months, while becoming familiar with their role and responsibilities. After this period, remote work arrangements can be discussed with the team.

Ideally, candidates should be able to work from the office 1–3 days per week.

Holidays & Leave

- 128 days of annual holidays (for the current fiscal year: January–December 2026; subject to calendar variations)

- Two days off per week (Saturday and Sunday)
- Summer vacation
- Year-end and New Year holidays
- Bereavement and special leave
- Paid annual leave (10 days granted after six months of employment)

Benefits & Perks

- Comprehensive social insurance coverage
- Commuting allowance (up to JPY 50,000 per month)
- PhD Support Program (financial assistance with tuition fees for employees pursuing a doctoral degree)
- Support for internal study groups and academic conference participation
- Company events held several times a year

Required Skills

Required Qualifications

- At least 2 years of hands-on experience building data pipelines in a corporate or research institution setting.
- Strong proficiency in Python, practical SQL skills, and solid command-line experience with Linux.
- Practical understanding of common bioinformatics file formats, including FASTA, GFF/GTF, GenBank, and sequencing/alignment output formats.
- Hands-on experience with cloud object storage and computing services (e.g., AWS, GCP, Nebius), Docker, and Git.
- Business-level proficiency in either Japanese or English, with the ability to work effectively in a bilingual team.

Preferred Qualifications

- Experience with microbial or prokaryotic genomics.
- Experience preparing data for deep learning, and training and running inference with those models (for example in PyTorch).
- Familiarity with genomic or protein language models (for example ESM, ProGen, Evo, GenomeOcean, Bacformer) and their input formats.
- Experience working with lab teams to register and track sample and assay data in a LIMS.
- Experience with workflow managers (Nextflow, Snakemake, Airflow, Prefect or Dagster).

Number of Openings: 1

Selection Process:

1. Resume Screening
2. First Interview (HR)
3. Second Interview (Head of Innovation and Principal AI Scientist)
4. Reference Check
5. Final Interview (CEO / CSO / CTO)
6. Final Hiring Decision

Additional Information:

- All interviews except the final interview are planned to be conducted online.
- Additional interviews or meetings may be arranged depending on the circumstances.
- We are happy to discuss scheduling adjustments to accommodate other ongoing recruitment processes.