# MetaOmics-10T: The Foundational Dataset to Unlock Causal Modeling of Microbial Ecosystems

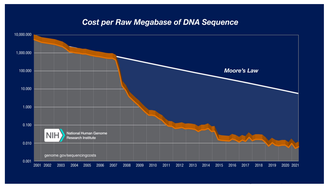
Arvid E. Gollwitzer, Deepak A. Subramanian, Isaac Tucker, Giovanni Traverso

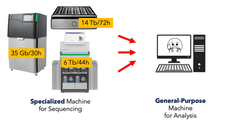
Massachusetts Institute of Technology · MITMECHE · KOCH INSTITUTE for Integrative Cancer Research at MIT · BROAD INSTITUTE · ETH zürich

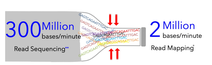
Full Paper
MetaOmics-10T

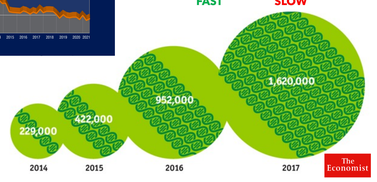
All Useful Links
LinkedIn, email…

## The Data We Need is Available - But Not Usable

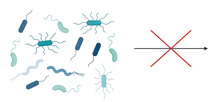


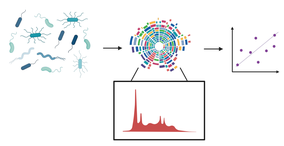

Number of Genomes Sequenced




### 100+ PB Exist, Yet 95% Is Unusable
Microbial data suffers from high noise and variability

### Mostly Noise
Standard models fail: they cannot distinguish between biological signal noise

### No Causal Structure
- Data cannot be property interpreted to identify causal relationships between microbial data and downstream tasks
- Current models achieve <60% accuracy on basic tasks



## Our Solution: MetaOmics-10T

### Quality-Aware Tokenization
**Unlocks Unprecedented Data**
- *Expands usable training data by 15%*
- *Lifts usable fraction from 5% to 40% (+35pp, 8× data)*

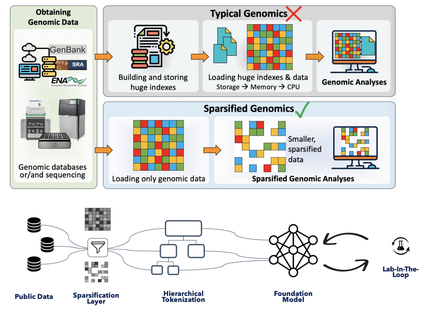
### Intelligent Sparsification
**Extracts useful data**
*Taxonomic classification as representative example: always achieve a superior accuracy-cost trade-off*

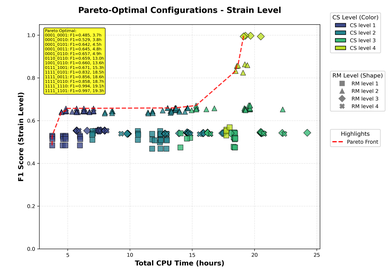
### 100,000+ Causal Trajectories

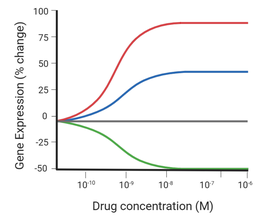
**Systematic perturbation-response enables counterfactual inference**

CRISPR knockouts + compound screens

Model-Guided Experimental Design (MGED)




## Pipeline Phases: From Noisy Archives to Causal Signal

### Phase 1: Mining Metagenomic Data ($10M)

**100 PB of microbial data**

↓

**QA-Tokenizer**

↓

**Quality-aware dataset**

**Mine 100+ PB**
SRA/ENA/RefSeq

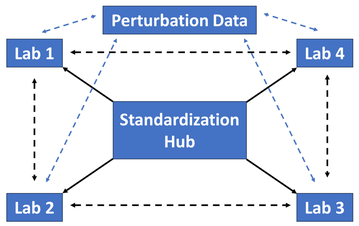
**Major efforts in sparsification and automated quality scoring**
and reinforcement-based vocabulary training

**In-storage processing**
to eliminate data movement (like GenStore, MegIS…)

***Total of ~6.8M hours, $10M***

### Phase 2: Perturbation Trajectories for Simulation ($40M)

**Perturbation Trajectories for Simulation**
AI-in-the-loop, Model-Guided Experimental Design




**Standardized reagents/protocols**
overlap experiments for cross-lab calibration

**Tier 1** (Perturbation with Microbiome-on-a-Chip Arrays)

**Tier 2** (Mechanistic Insights on High-Potential Interactions)

**Tier 3** (Pre-Clinical Validation)

## Pilot Dataset: Acquisition and Results

### Mining 1 TB SRA Data

**1 TB of sequence read analysis data**

↓

**QA-Tokenizer**

↓

**Quality-aware dataset**

**25K diverse microbiome samples**
10M reads

**Data passed through QA-Tokenizer**
500M token foundation model

**Validation performance**
12% better than held-out perplexity

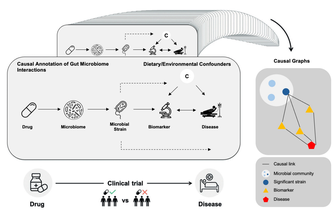
**Unlocked Unprecedented Data**
Expands usable training data by 15%
Increases usable fraction from 5% to 40% (+35pp, 8× data)

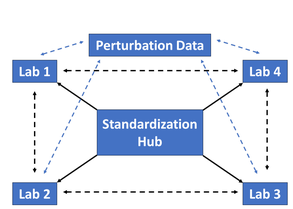
### Causal Trajectory Pilot

**Variety of trajectories sampled**
2 species, 2 compounds, 5 doses, 12 time points

**Cost $2100/trajectory at pilot scale**
10x higher than projected at scale

**Batch effects between labs**:
35% of variance -> need harmonization




## Key Results & Applications Unlocked by MetaOmics-10T

### Predictive & Therapeutic Engineering

**Drug-Microbiome Interactions**
Prediction of drug response

**Design of Interventions**
In silico design of microbiome therapies

**Universal Perturbation Engine**
Zero-shot prediction of novel compound/genetic modification effects

### Fundamental Biological Principles

**Host-Microbe Interactions**
Map molecular dialogue: protective microbes, immune shaping

**Mapping Microbiome Biogeography**
Spatial organization design principles and environmental reconfiguration

**Dark Matter**
Assign functions to unannotated genes/metabolites

**Darwin-7B: state-of-the-art across key benchmarks**

| Benchmark | Darwin-7B | METAGENE-1 | Evo2-7B |
|---|---|---|---|
| Pathogen Detection (MCC) | **94.5** | 93.0 | 87.0 |
| Metagenomic Profiling (F1) | **0.98** | – | 0.89 |
| Metabolic Pathway (wF1) | **0.91** | 0.84 | 0.79 |
| IBD Prediction (AUC) | **0.947** | – | – |
| T2D Prediction (AUC) | **0.883** | – | – |
| Antibiotic Resistance (AUC) | **0.910** | – | – |

*All comparisons statistically significant ($p<0.05$, two-sided t-test)*

**$50M investment yields equivalent of $1B+ dataset**
A blueprint for foundational predictive models of microbial ecosystems

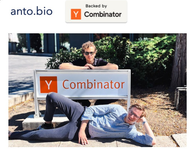
## Conclusion & Future Work

**MetaOmics-10T can unlock causal relationships in human and ecological microbiomes**
using quality-aware tokenization, subsequent foundation model training, and perturbation trajectory analysis

**We achieve superior validation performance and reduces batch-to-batch variation** resulting in 4x improvement in performance and 100x improvement in total cost

### Initiatives to Build MetaOmics-10T

**Anto Biosciences**
anto.bio · Backed by Y Combinator




Anto Bio @ YC

Our YC launch

**Broad/MIT: FINGERPRINT**
FINGERPRINT.bio


Davos Alzheimer's Collaborative · World Economic Forum

$1M Alzheimer's Insights AI Prize