MetaOmics-10T

Making the Microbiome Computable

A Foundational Dataset to Unlock Causal Modeling of Microbial Ecosystems

arvidg@mit.edu | *arvid@anto.bio*

Arvid E. Gollwitzer
arvidgollwitzer.com

From Observation to Intervention: The Formal Contract for Digital Twins

Modelling Microbial Ecosystems as controlled dynamical systems enabling three core AI tasks

1. Forecasting
Forecasting ecosystem dynamics (no intervention)

2. Prediction
Predicting counterfactual outcomes of interventions

3. Safe Inverse Design
We inverse-design of microbial therapies under safety constraints

MetaOmics-10T combines 10 trillion base pairs reclaimed from public archives using a Quality-Aware Tokenization (QA-Token)
with 100,000+ interventional trajectories generated via model-guided experimental design

All The Microbiome Data We Need is Available

100+ PB Exist, Yet 95% Is Unusable

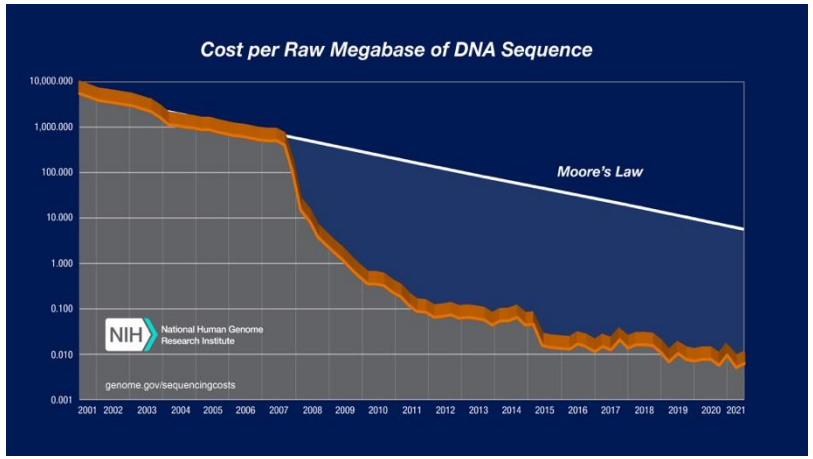
Microbial data suffers from high noise and variability, reducing the predictive power of models trained on this data

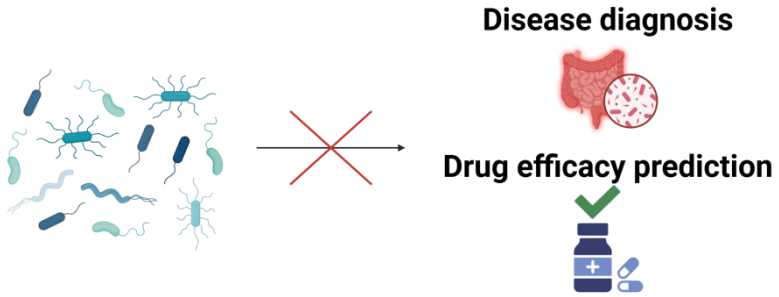
Most od the Data is Noise

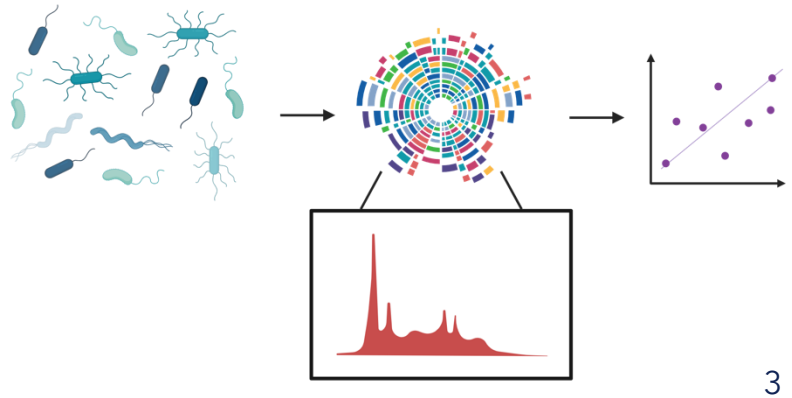
Standard models fail because they cannot distinguish between biological signal noise

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







From Noisy Archives to Causal Signal: The Missing Substrate

Quality-Aware Tokenization

RL + Sparsification

To remove irrelevant data incorporate quality directly into vocabulary construction

Unlocks Unprecedented Data

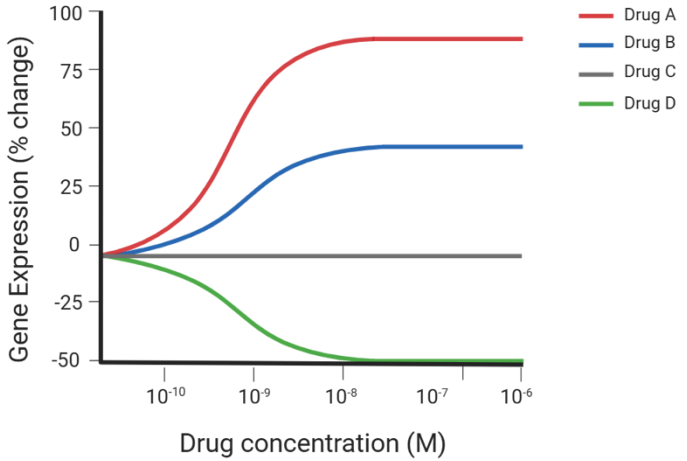
- *Expands usable training data by 15%*
- *Lifts usable fraction from 5% to 40% (+35pp, 8× data)*

Full Results: Paper
MetaOmics-10T

100,000+ Causal Trajectories

Systematic perturbation-response enables counterfactual inference

- CRISPR knockouts + compound screens
- Model-Guided Experimental Design (MGED)



From Archive to Causality: A Two-Phase Pipeline

Phase 1: Data Reclamation

Months 1–12 | $10M

Mine 100+ PB
across SRA/ENA/RefSeq…

100+ PB of microbial data

↓

QA-Tokenizer

↓

Quality-aware dataset

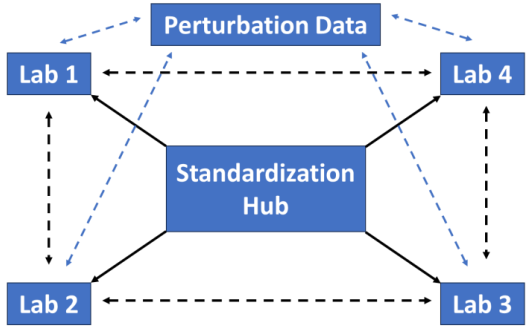
Major efforts in sparsification and automated quality scoring
and reinforcement-based vocabulary training

In-storage processing
to eliminate data movement (similar to GenStore, MegIS)

Total of ~6.8M hours, $10M

Phase 2: Causal Trajectories

Months 13–36 | $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 Screening with Microbiome-on-a-Chip Arrays)

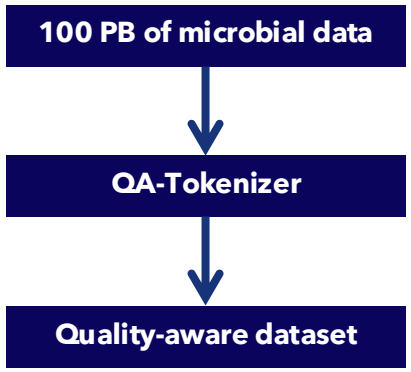
Tier 2 (Mechanistic Insights on High-Potential Interactions)

Tier 3 (Pre-Clinical Validation)

From Archive to Causality: A Two-Phase Pipeline

Phase 1: Data Mining

Months 1–12 | $10M



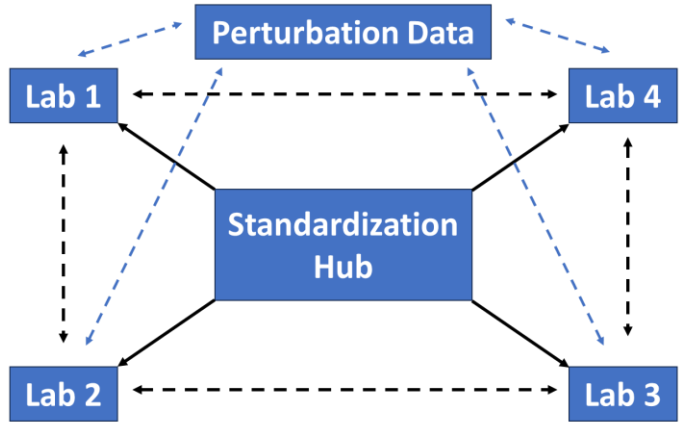
NCBI

uk biobank
Enabling scientific discoveries that improve human health

ENA
European Nucleotide Archive

Phase 2: Causal Trajectories

Months 13–36 | $40M

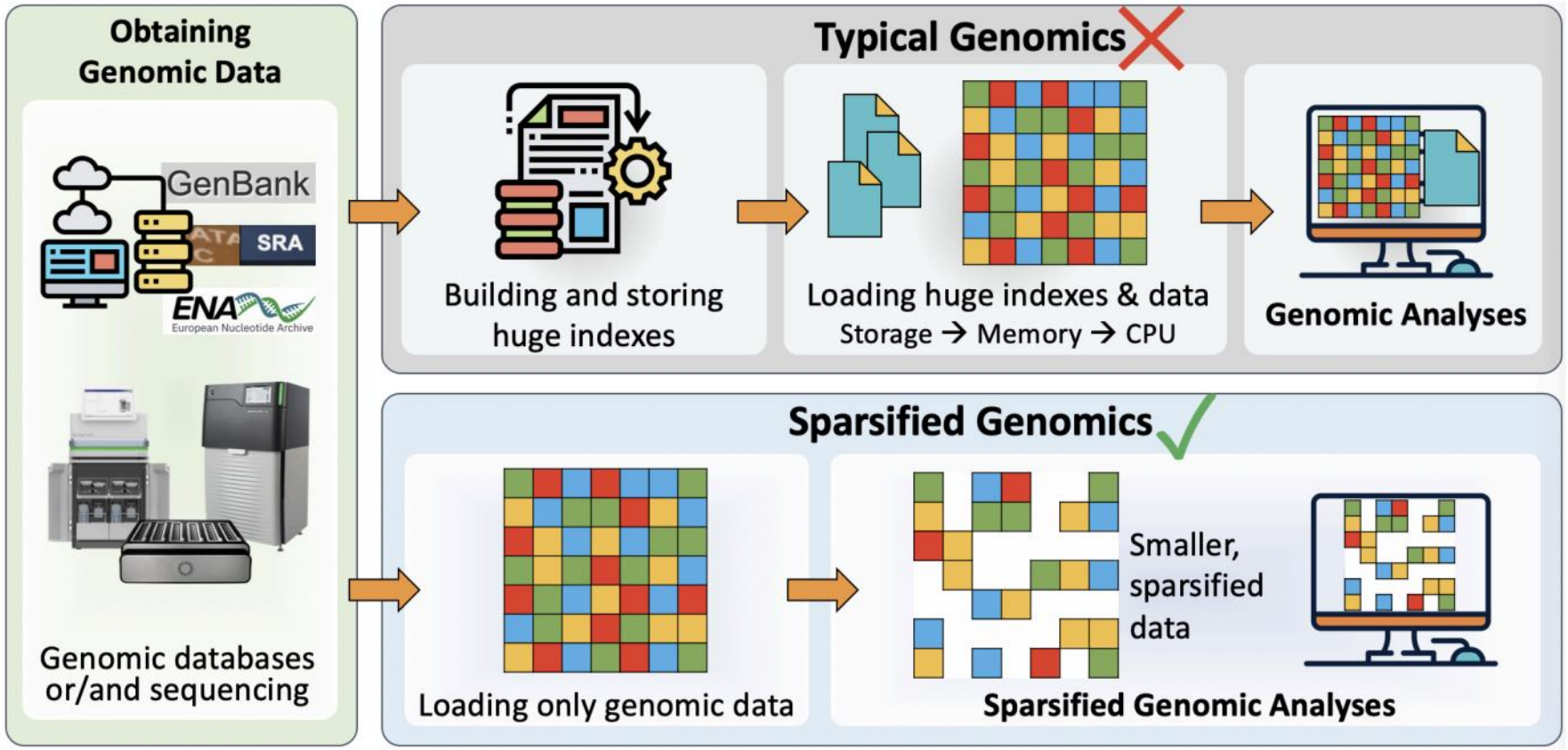


$50M investment yields equivalent of $1B+ dataset

A blueprint for foundational predictive models of microbial ecosystems

From Archive to Causality: Sparsified Data

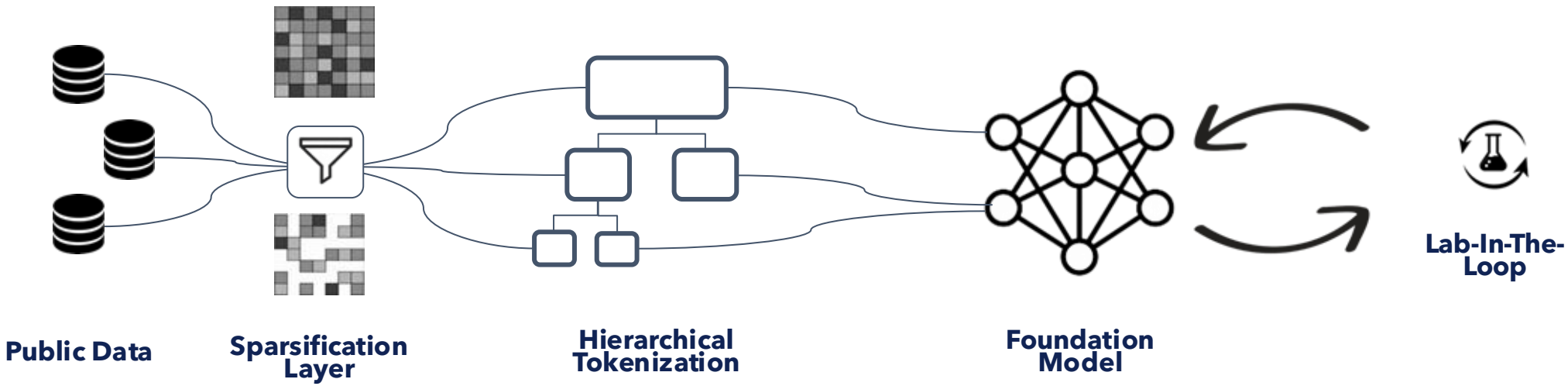
State-of-the-art computational methods analyzing genomic sequences fail to cope with the exponential growth of genomic sequencing data



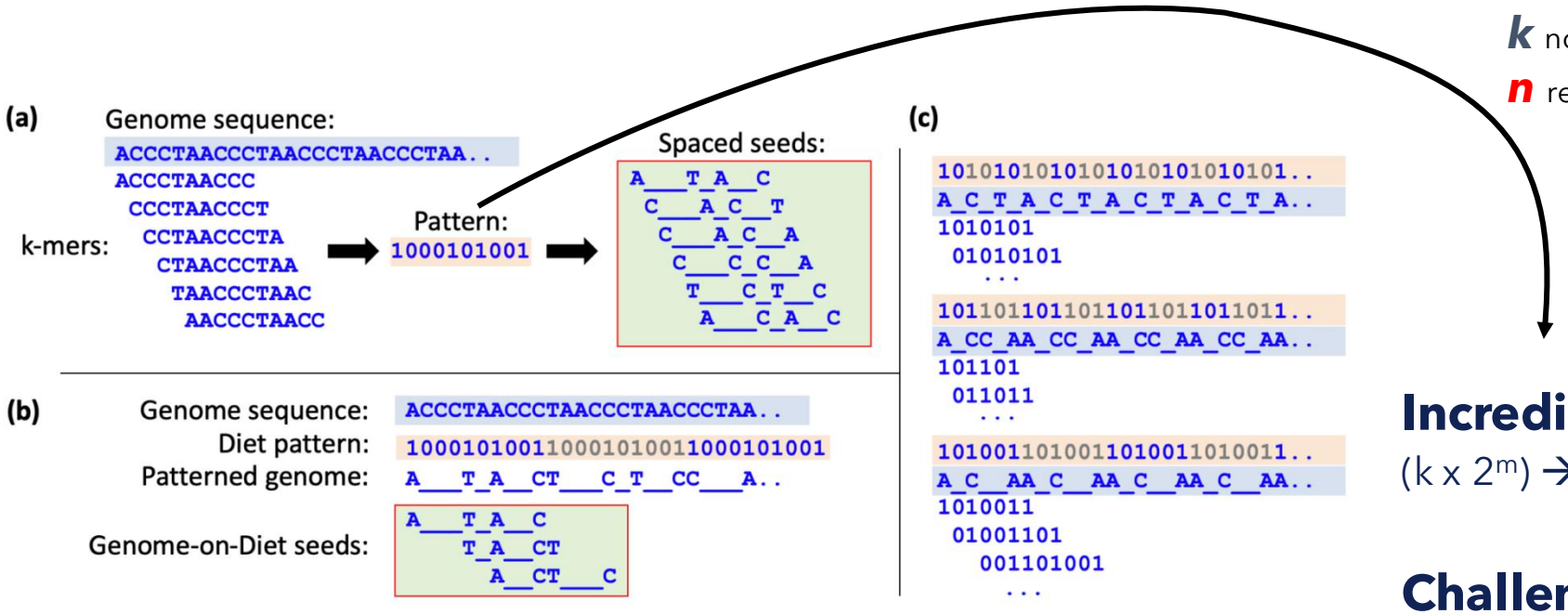
Reclaiming Data: Sparsification is Key

Most public data was previously unusable

We condense it into a high-quality meta-vocabulary



Sparsified Genomics: 10× Speedup at no Accuracy Loss



m read length

k no. of reads

n reference genome length

Incredibly Large Search Space

$(k \times 2^m)$ → Trillions of choices!

Challenge

Find the optimal choice for each read!

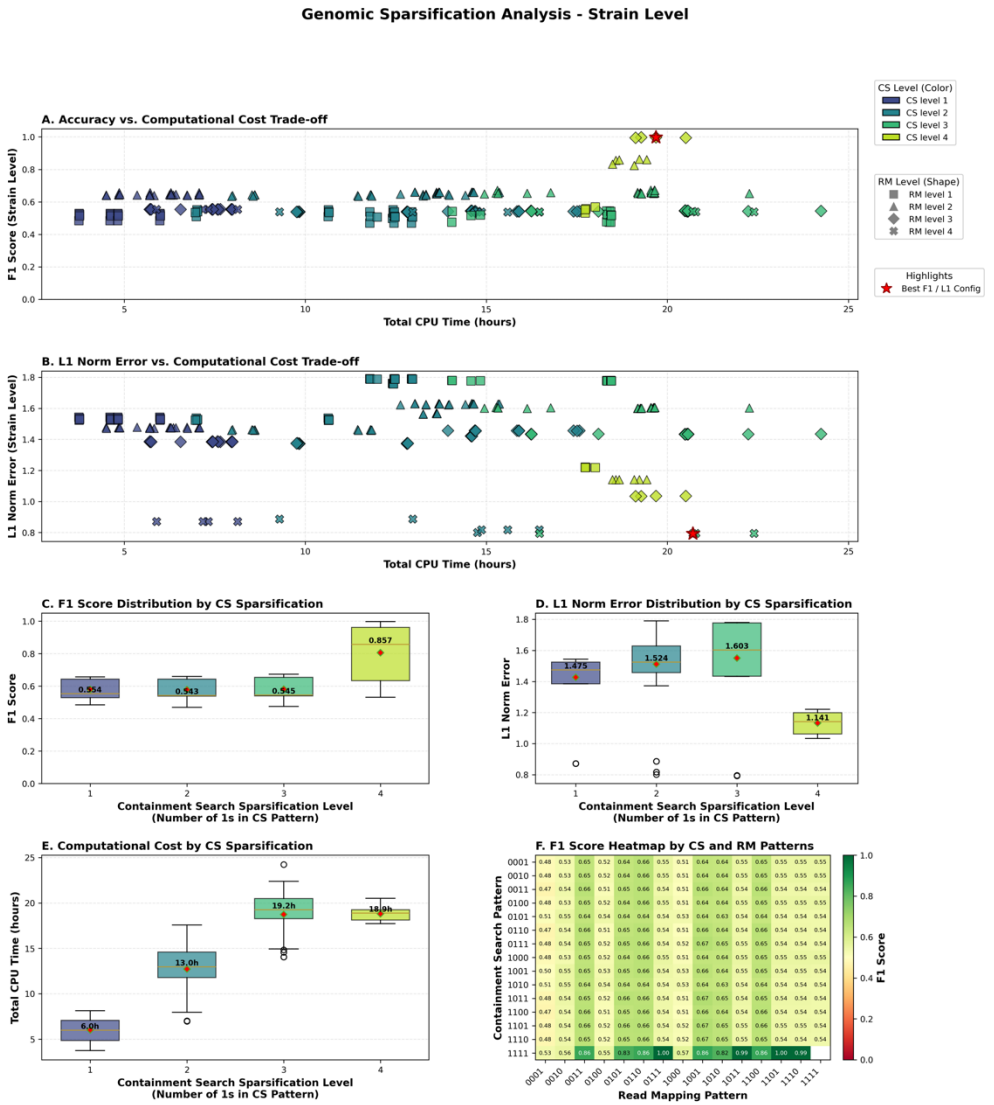
Data Sparsification: 10× Speedup at no Accuracy Loss

224 sparsification patterns

evaluated on the CAMI benchmark (taxonomic profiling)

Key Results:

- **F1 = 0.994 at 5.1× speedup**
- 17× faster containment search
- 3.4× faster read mapping
- 2.8× faster index construction
- Generalizes across datasets for strain-level classification



Data Sparsification: 10× Speedup at no Accuracy Loss

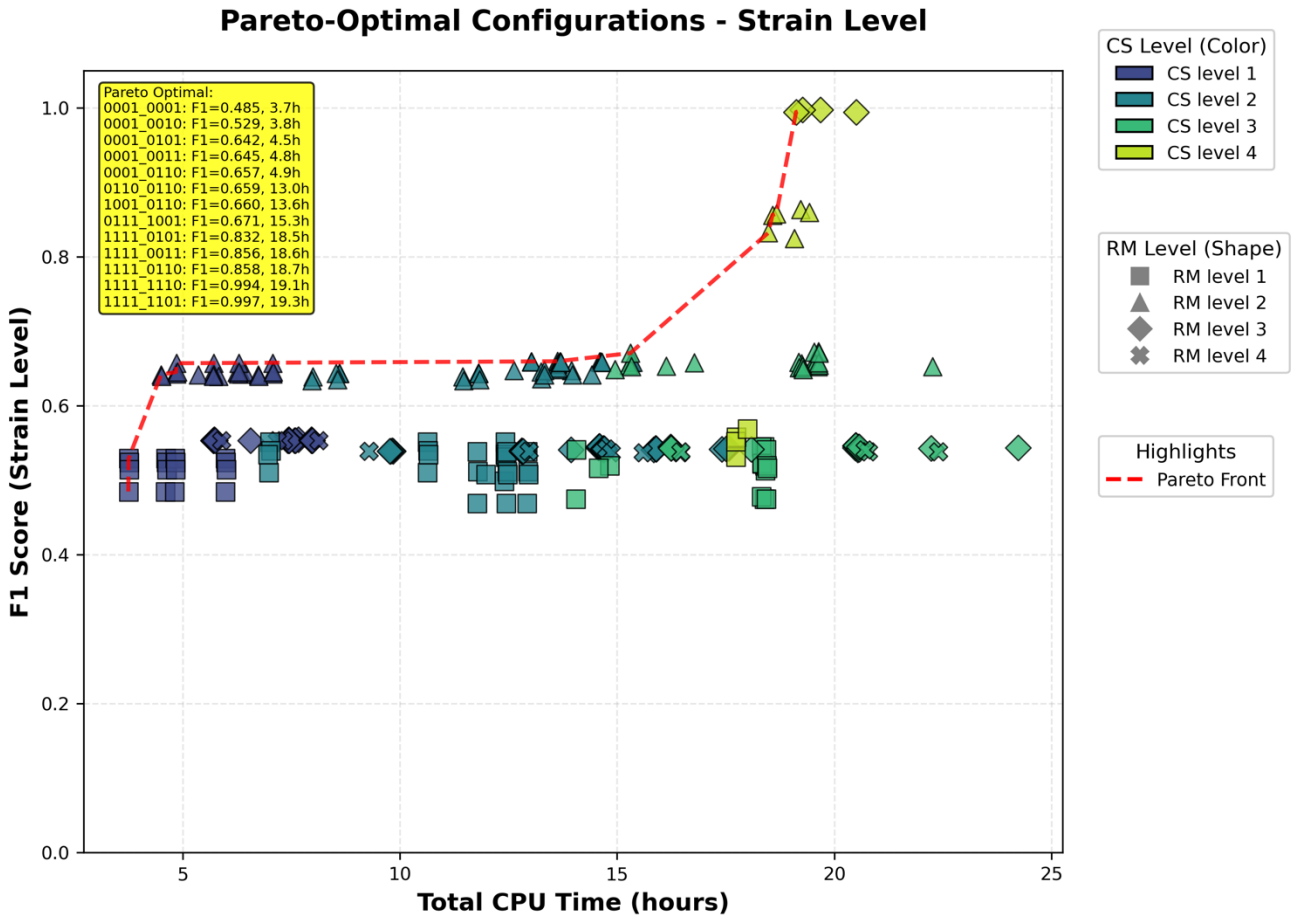
2240 sparsification patterns

evaluated on the CAMI benchmark (taxonomic profiling)

Key Results:

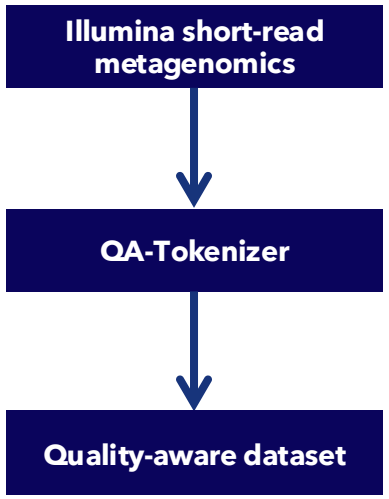
- **F1 = 0.994 at 5.1× speedup**
- 17× faster containment search
- 3.4× faster read mapping
- 2.8× faster index construction
- *12-13 Pareto-optimal configurations identified*

Paper
The Thinking Microscope



Archive to Model: A Pilot Dataset

Mining 10 TB SRA Data



25K diverse microbiome samples
10M reads

Data passed through QA-Tokenizer
500M token foundation model

Validation performance
12% better than held-out perplexity

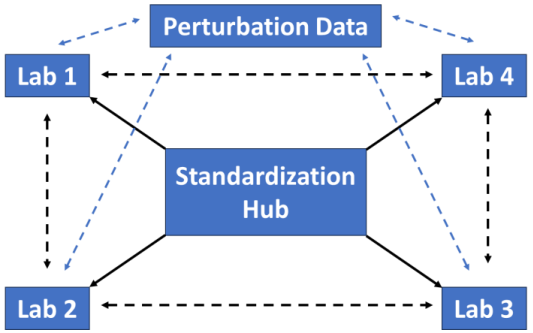
Unlocked Unprecedented Data
- Expands usable training data by 15%
- Increases usable fraction from 5% to 40% (+35pp, 8× data)

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

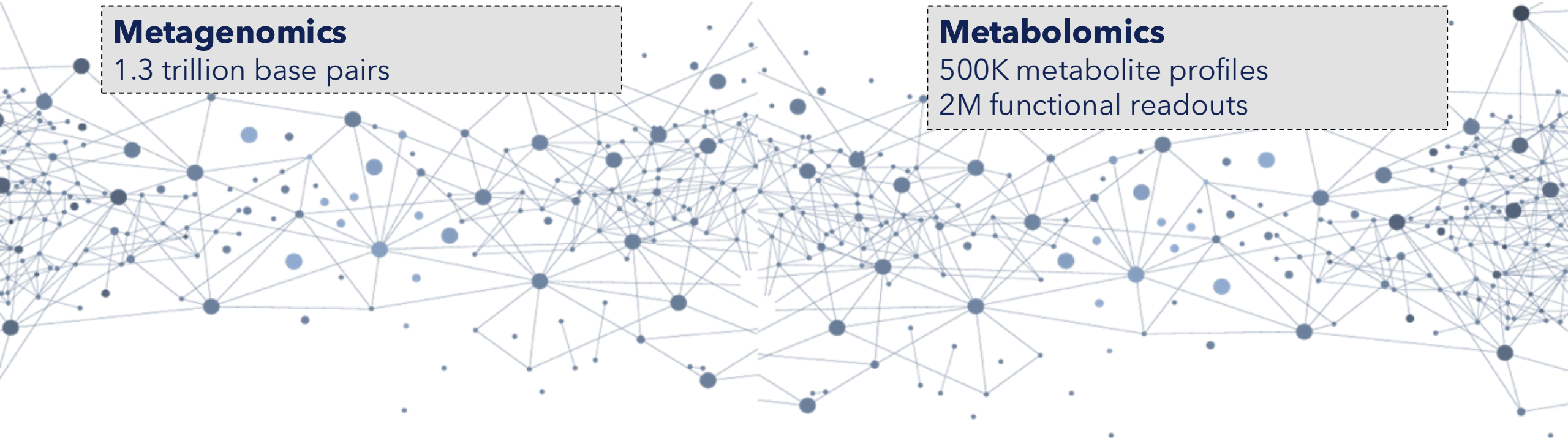
Batch effects between labs:
35% of variance -> need harmonization



New Foundation Models

26 x faster and 10x cheaper: The only practical model at the performance frontier

Evo2 (40B):	Assembled genomes – clean, single-organism, no environmental context
METAGENE-1 (7B):	Metagenomic reads – environmental samples, no causal structure
Darwin-7B:	**Multi-omic data + causal trajectories**



Metagenomics
1.3 trillion base pairs

Metabolomics
500K metabolite profiles
2M functional readouts

Darwin-7B at the Performance Frontier

Darwin-7B achieves 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)

Full Results: Paper
MetaOmics-10T

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

Arvid E. Gollwitzer*
Broad Institute of MIT and Harvard
Cambridge, MA, USA
arvidg@mit.edu

Deepak A. Subramanian
Dept. of Chemical Engineering, MIT
Koch Institute for Integrative Cancer Research, MIT
Broad Institute of MIT and Harvard
Cambridge, MA, USA

Isaac Tucker
Broad Institute of MIT and Harvard
Cambridge, MA, USA

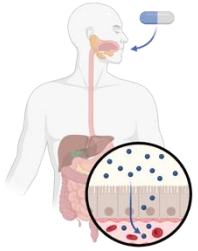
Giovanni Traverso*
Dept. of Mechanical Engineering, MIT, Cambridge, MA, USA
Div. of Gastroenterology, Hepatology and Endoscopy,
Brigham & Women's Hospital, Harvard Medical School, Boston, MA, USA
Koch Institute for Integrative Cancer Research, MIT, Cambridge, MA, USA
Broad Institute of MIT and Harvard, Cambridge, MA, USA
cgt20@mit.edu

Abstract

We propose **MetaOmics-10T**—an openly shareable, foundational dataset to unlock AI-accelerated discovery in microbial ecosystems. The dataset directly enables three high-impact AI tasks: (1) forecasting ecosystem dynamics, (2) predicting counterfactual outcomes of interventions, and (3) inverse-design of microbial therapies under safety constraints. MetaOmics-10T combines **10 trillion base pairs** reclaimed from public archives using a Quality-Aware Tokenization (QA-Token) framework with **100,000+ interventional trajectories** generated via model-guided experimental design. The result is a first-of-its-kind, probabilistic, intervention-ready corpus that addresses the principal bottleneck for causal modeling in microbiome science and provides an empirical testbed to assess the reach and limits of causal inference at scale.

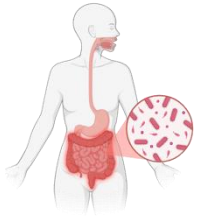
Applications & Models Unlocked by MetaOmics-10T

Predictive & Therapeutic Engineering



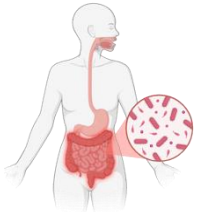
Drug-Microbiome Interactions
Prediction of drug response

Design of Interventions
In silico design of microbiome therapies

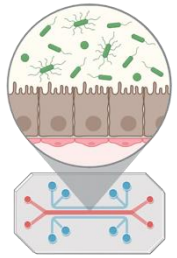


Universal Perturbation Engine
Zero-shot prediction of novel compound/genetic modification effects

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

Key Initiatives to Build MetaOmics-10T

Anto Biosciences

anto.bio

Backed by Y Combinator

Anto Bio @ YC

Our YC launch

Broad/MIT: FINGERPRINT

FINGERPRINT.bio

Massachusetts Institute of Technology

BROAD INSTITUTE

Davos Alzheimer's Collaborative

WORLD ECONOMIC FORUM

COMMITTED TO IMPROVING THE STATE OF THE WORLD

$1M Alzheimer's Insights AI Prize

Paper
MetaOmics-10T

All useful links!
LinkedIn, email…

MetaOmics-10T

Making the Microbiome Computable

Arvid E. Gollwitzer

arvidgollwitzer.com

arvidg@mit.edu

Appendix

Arvid E. Gollwitzer

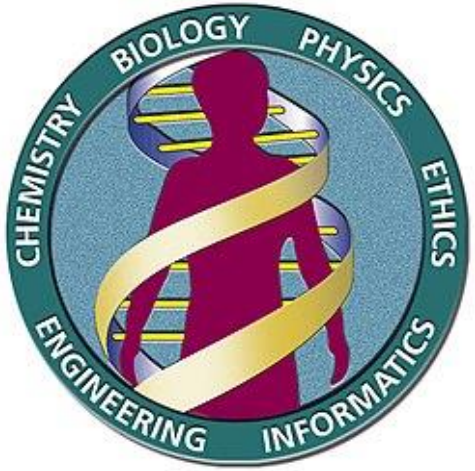
arvidg@mit.edu | arvid@anto.bio

Biggest Missing Layer In Drug Development And Human Health

Microbiome has been the biggest missing layer in drug development and human health

We learned to read the human genome
it reshaped medicine

Microbiome is still a black box
not because it is unimportant, but because it was not computable



2/3 Of Drugs Are Heavily Affected My The Microbiome

That is why most drugs only work for some people

Over 1 Billion people
are on drugs where the microbiome quietly decides treatment succeeds or failure

ARTICLE
https://doi.org/10.1038/s41586-019-1291-3

Mapping human microbiome drug metabolism by gut bacteria and their genes

Michael Zimmermann[1,3], Maria Zimmermann-Kogadeeva[1,3], Rebekka Wegmann[1,2] & Andrew L. Goodman[1]*

Individuals vary widely in their responses to medicinal drugs, which can be dangerous and expensive owing to treatment delays and adverse effects. Although increasing evidence implicates the gut microbiome in this variability, the molecular mechanisms involved remain largely unknown. Here we show, by measuring the ability of 76 human gut bacteria from diverse clades to metabolize 271 orally administered drugs, that many drugs are chemically modified by microorganisms. We combined high-throughput genetic analyses with mass spectrometry to systematically identify microbial gene products that metabolize drugs. These microbiome-encoded enzymes can directly and substantially affect intestinal and systemic drug metabolism in mice, and can explain the drug-metabolizing activities of human gut bacteria and communities on the basis of their genomic contents. These causal links between the gene content and metabolic activities of the microbiota connect interpersonal variability in microbiomes to interpersonal differences in drug metabolism, which has implications for medical therapy and drug development across multiple disease indications.

Following administration, drug molecules typically undergo chemical modification(s); the resulting metabolites can have functional and toxicological properties that are distinct from those of their parent drug[1]. Most drugs are delivered orally and can encounter commensal microorganisms in the small and large intestine. These microorganisms collectively encode 150-fold-more genes than the human genome; this genetic diversity encompasses a rich enzyme repository with drug-metabolizing potential. Anecdotal examples of interactions between the gut microbiome and drugs or drug metabolites, with intestinal and systemic pharmacological effects, have previously been reported. Such compound modifications by gut microorganisms can lead either to their activation (for example, sulfasalazine[2]), inactivation (for example, digoxin[3]) or toxification (for example, sorivudine and brivudine[4,5], and irinotecan[6]). For a few drugs, microbial biotransformation has been assigned to specific bacterial strains and gene products[3,5,7]. However, these examples are the exception, as there is little systematic understanding of the scope, specificity or microbial and/or chemical determinants of microbiome–drug interactions[8].

We set out to systematically assay interactions between drugs and microorganisms by measuring the ability of representative human gut bacteria to metabolize structurally diverse drugs, and by identifying drug-metabolizing microbial gene products. We establish that these drug-metabolizing microbial proteins can contribute to the in vivo drug metabolism of gnotobiotic mice, and provide evidence that metagenomics and genomics sequence data can explain the capacity of both isolated gut bacteria and complete communities to convert specific drugs. This could provide a means to mechanistically connect information about the microbiome to interpersonal variation in drug metabolism and toxicity.

Drug-metabolizing bacteria from the gut microbiome

We first assessed the capacity of 76 bacterial species and/or strains—which represent the major phyla of the human gut microbiome—to chemically modify medical drugs in vitro (Fig. 1a, Supplementary Table 1). We used a previously established combinatorial pooling strategy[9] to assign 271 drugs across 21 pools, such that each drug is represented in quadruplicate but shares a pool with any other drug twice at most (Extended Data Fig. 1a). The 271 drugs were selected to span chemical drug space, which resulted in a selection of diverse clinical indications (excluding antibiotics), physicochemical properties and predicted intestinal concentrations (Fig. 1b, Extended Data Fig. 1b–d, Supplementary Table 2). We incubated each gut species or strain with each drug pool and three vehicle controls under anaerobic conditions, and measured drug concentrations before and after a 12-h incubation by liquid-chromatography-coupled mass spectrometry (LC–MS). The 3,840 samples we analysed comprise a total of 20,596 bacteria–drug interactions, measured in quadruplicate.

We discovered that, for two thirds (176/271) of the assayed drugs, the level of the drug after incubation was significantly reduced (>20%, FDR-corrected P value ≤ 0.05) by at least one bacterial strain, and that each strain metabolizes 11–95 drugs (Extended Data Fig. 1e–g, Supplementary Table 3). Drug levels were largely unchanged in no-bacteria controls that were buffered to pH 4–7, controlling for acidification of the culture medium. By contrast, levels of positive-control drugs that were expected to be metabolized by gut bacteria[10]—such as sulfasalazine, lovastatin, omeprazole and risperidone—were significantly decreased over time (>20%, FDR-corrected P value ≤ 0.05) (Extended Data Fig. 2a). Clustering the bacterial isolates according to their drug-metabolizing activities recapitulates their phylogenetic relationships to the strain level, and reveals phylum-specific metabolic activities (Fig. 1c, Extended Data Fig. 3a). Clustering the drugs on the basis of these data revealed groups of compounds that share structural features, as shown by functional group and maximum common substructures analysis (Fig. 1d, e, Extended Data Fig. 3b, Supplementary Table 4). This suggests possible chemical targets for metabolic modifications by bacteria. For example, drugs that are specifically metabolized by Bacteroidetes (cluster I in Fig. 1c) contain ester or amide groups that can be hydrolysed, whereas the compounds that are metabolized by most bacteria (except Proteobacteria) (cluster II in Fig. 1c) all contain a nitro or azo group, which is prone to reduction in anaerobic metabolism. Functional group analysis suggests that particular chemical substructures (such as lactones, and nitro, azo and urea groups) predispose compounds for microbial metabolism (Extended Data Fig. 2b). Chemical groups that

[1]Department of Microbial Pathogenesis and Microbial Sciences Institute, Yale University School of Medicine, New Haven, CT, USA. [2]Present address: Department of Biology, Institute of Molecular Systems Biology, ETH Zurich, Zurich, Switzerland. [3]These authors contributed equally: Michael Zimmermann, Maria Zimmermann-Kogadeeva. *e-mail: andrew.goodman@yale.edu

Foundation Model To Understand The Microbiome Metabolism Of Drugs

The Engine: In-Silico Metabolism Correction

1. Scan: Identification of Liability
Scan pharmacopeia to identify approved or failed drugs with high "gut-drain" (metabolic degradation).

2. Mechanism: Atomic-Level Resolution
Pinpoints the exact strain/enzyme and the specific chemical bond responsible for degradation

3. Solve: Generative Armoring
We generate novel chemical analogs that retain target potency but are invisible to the bacterial enzyme.

Result
A proprietary New Chemical Entity (NCE) with superior efficacy/toxicity.

Foundation Model To Understand The Microbiome Metabolism Of Drugs

We redesign existing molecules

We optimize them for broader efficacy
so they work for far more people

We turn niche drugs into blockbusters

We finally understand the failure mechanisms

That enables us to fix the drug so they work with every gut.

Everyone has tried to fix the gut to make a drug work. But you can't change an entire population's gut microbiome.

Proof: Blinded Retrospective Prediction of Trial Failure

Using only data available prior to the trial

Darwin-7B correctly predicted the specific patient sub-populations that would fail, based on their microbiome signatures

Mechanism Identified

We identified the exact bacterial strain responsible for metabolizing the drug into an inactive form.

Crossing the Scaling Wall: Causal Foundation Models

Current State. Microbiome AI is stuck in the pre-ImageNet era: starved of quality and causality

Immediate Model Development Acceleration:

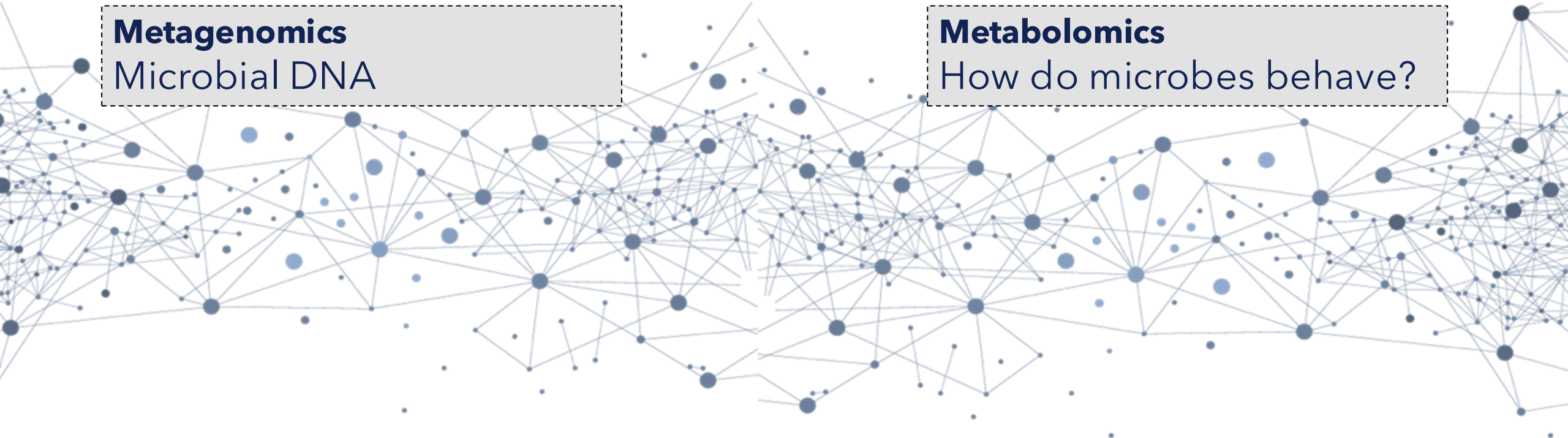
- **10× Larger Models**: Enables 10B+ parameter foundation models with emergent capabilities.
- **Causal Reasoning**: First dataset for counterfactuals—predicting intervention outcomes, not just observations.
- **Generalization**: Train once; transfer across organisms, environments, and interventions.

IP Engine

We own the most important assets to make drugs work across populations

Darwin-7B: 60x faster and 40% more accurate

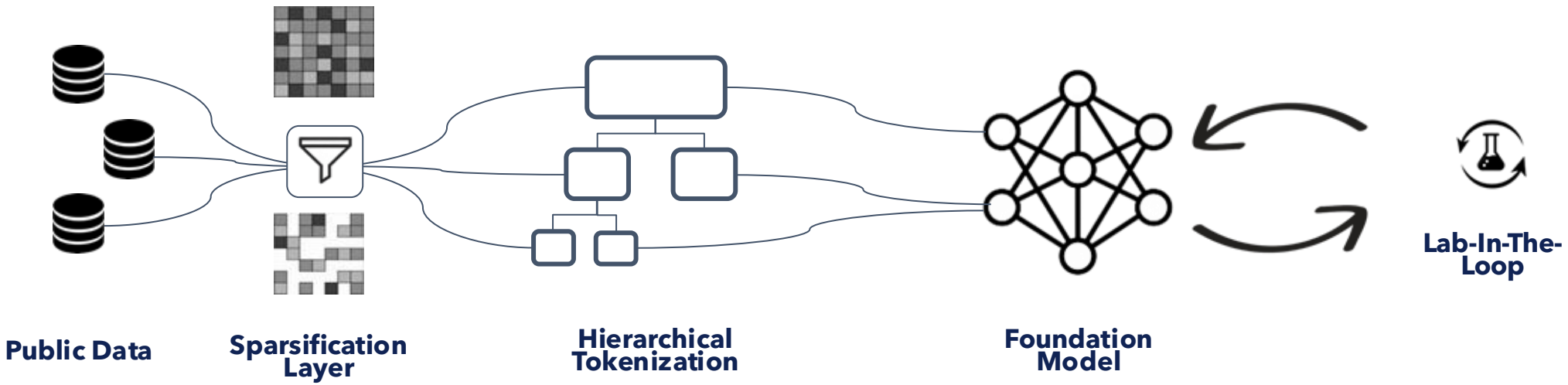
Than state-of-art models in genome understanding, taxonomic and functional profiling



Sparsification is Key

Most public data was previously unusable

We condense it into a high-quality meta-vocabulary



Microbiome Is The Missing Layer In Trillion Dollar Industries

assets with "unexplained" variability or toxicity

Scan

Darwin-7B

Patch

Chemically "armor" the molecule.

Validate

Lab | Central

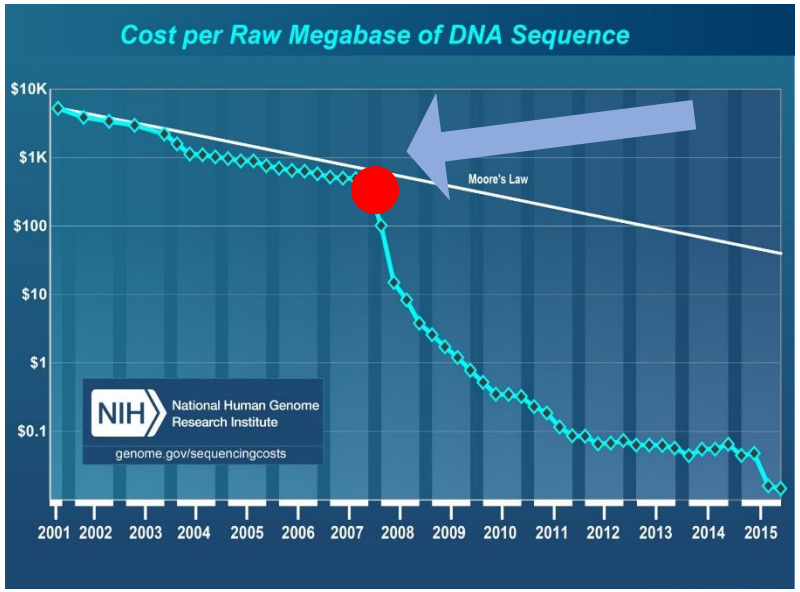
drug with larger accessible market.

The Data-Compute Gap

Anto Biosciences

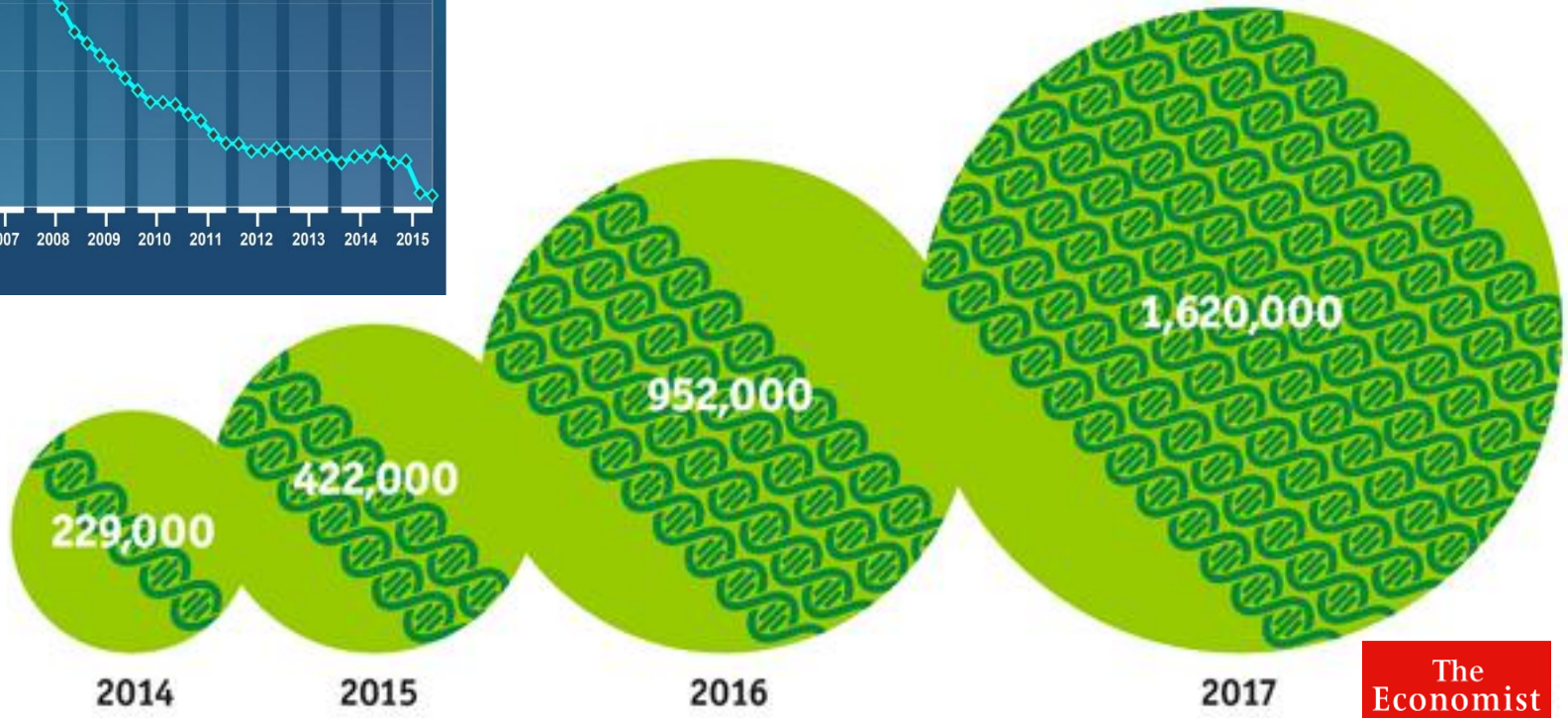
founders@anto.bio

Sequencing Cost Is Reducing



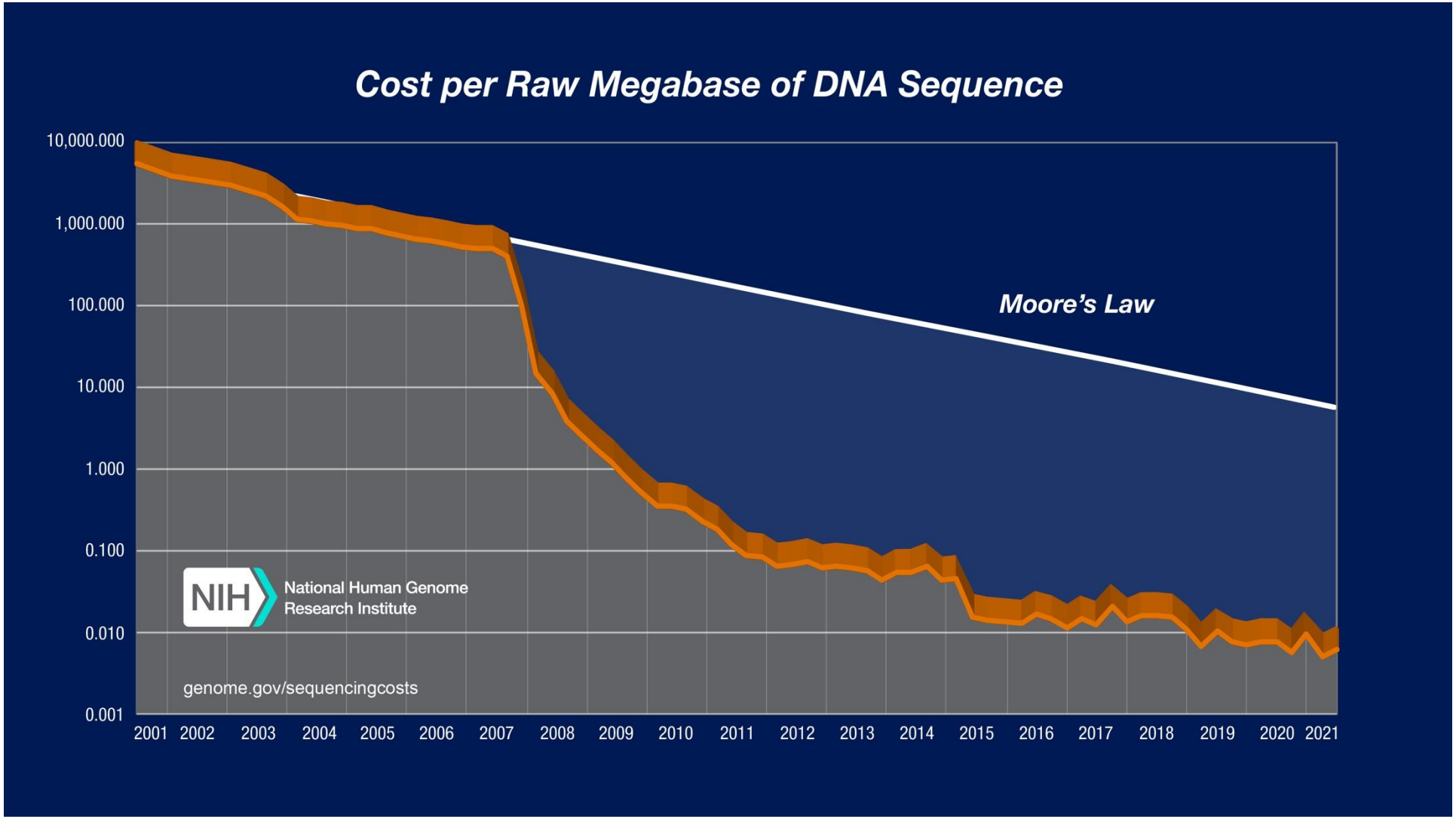
development of high-throughput sequencing (HTS) technologies

Number of Genomes Sequenced



http://www.economist.com/news/21631808-so-much-genetic-data-so-many-uses-genes-unzipped

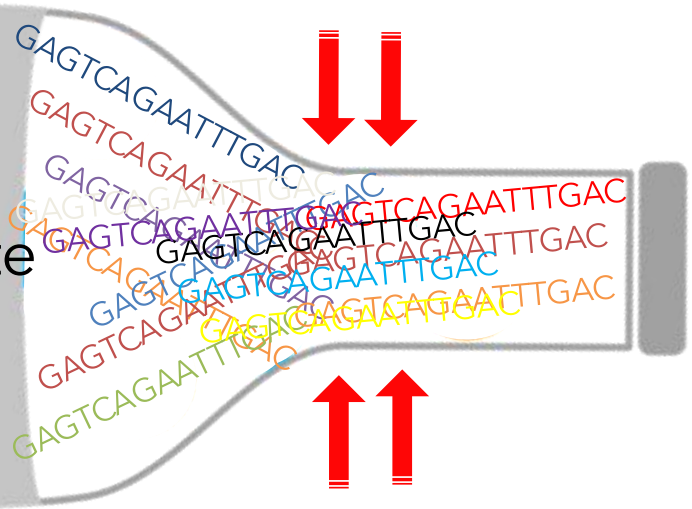
Genome Sequencing Cost Is Reducing



One Problem: Computation

300 Million bases/minute

Read Sequencing**



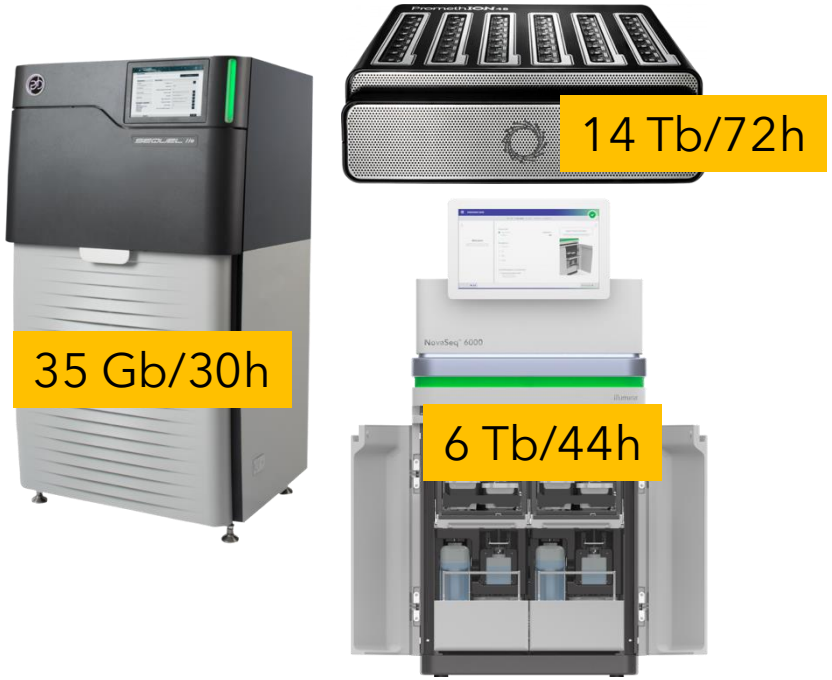
2 Million bases/minute

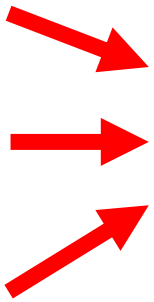
Read Mapping*

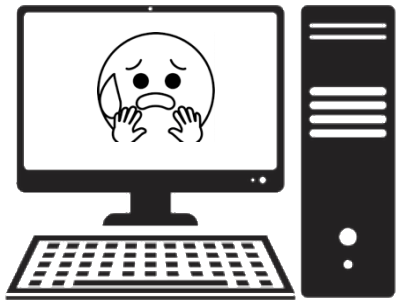
150x slower

* BWA-MEM
** HiSeqX10, MinION

Genome Sequencing Cost Is Reducing







Specialized Machine
for Sequencing

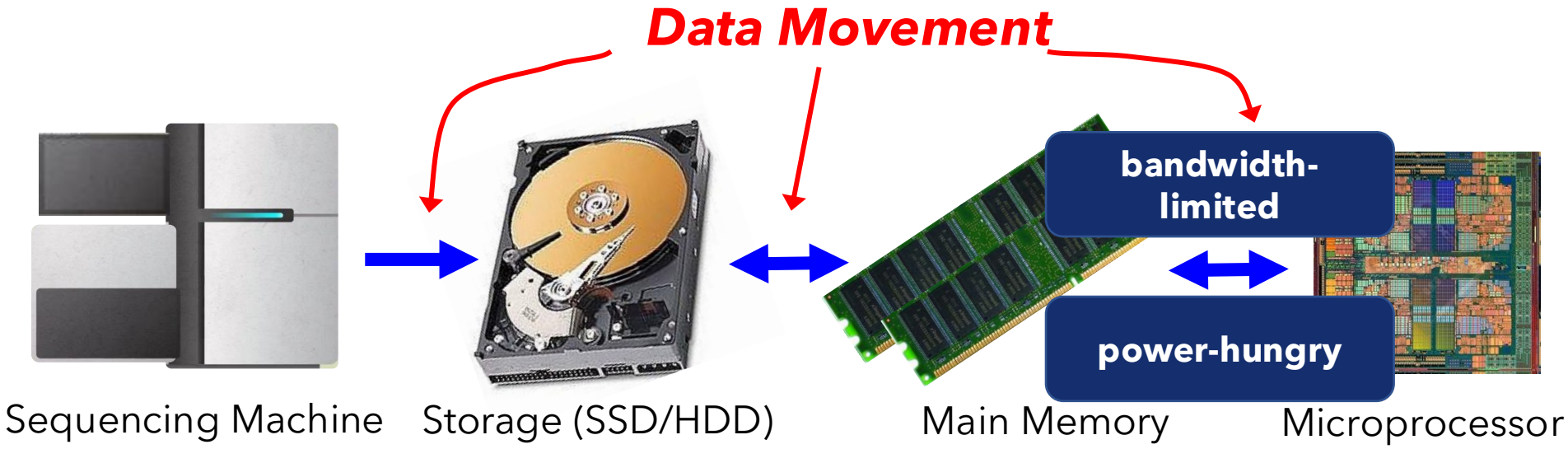
General-Purpose
Machine
for Analysis

FAST

SLOW

Data Movement Bottleneck

Data movement is a major bottleneck
in modern computer architectures



Fixing the Bottleneck: Special Purpose Systems

GenStore: A High-Performance In-Storage Processing System for Genome Sequence Analysis

Nika Mansouri Ghiasi
ETH Zürich
Switzerland

Jisung Park
ETH Zürich
Switzerland

Harun Mustafa
ETH Zürich
Switzerland

Jeremie Kim
ETH Zürich
Switzerland

Ataberk Olgun
ETH Zürich
Switzerland

Arvid Gollwitzer
ETH Zürich
Switzerland

Damla Senol Cali
Bionano Genomics
USA

Can Firtina
ETH Zürich
Switzerland

Haiyu Mao
ETH Zürich
Switzerland

Nour Almadhoun Alserr
ETH Zürich
Switzerland

Rachata Ausavarungnirun
KMUTNB
Thailand

Nandita Vijaykumar
University of Toronto
Canada

Mohammed Alser
ETH Zürich
Switzerland

Onur Mutlu
ETH Zürich
Switzerland

ABSTRACT

Read mapping is a fundamental step in many genomics applications. It is used to identify potential matches and differences between fragments (called *reads*) of a sequenced genome and an already known genome (called a *reference genome*). Read mapping is costly because it needs to perform *approximate string matching (ASM)* on large amounts of data. To address the computational challenges in genome analysis, many prior works propose various approaches such as accurate *filters* that select the reads within a dataset of genomic reads (called a *read set*) that *must* undergo expensive computation, efficient heuristics, and hardware acceleration. While effective at reducing the amount of expensive computation, all such approaches still require the costly movement of a large amount of data from storage to the rest of the system, which can significantly lower the end-to-end performance of read mapping in conventional of read mapping processes of reads with different properties and degrees of genetic variation, we meticulously design low-cost hardware accelerators and data/computation flows inside a NAND flash-based solid-state drive (SSD). Our evaluation using a wide range of real genomic datasets shows that GenStore, when implemented in three modern NAND flash-based SSDs, significantly improves the read mapping performance of state-of-the-art software (hardware) baselines by 2.07-6.05× (1.52-3.32×) for read sets with high similarity to the reference genome and 1.45-33.63× (2.70-19.2×) for read sets with low similarity to the reference genome.

CCS CONCEPTS

• **Computer systems organization** → **Special purpose systems**; • **Hardware** → **External storage**.

2024 ACM/IEEE 51st Annual International Symposium on Computer Architecture (ISCA)

MegIS: High-Performance, Energy-Efficient, and Low-Cost Metagenomic Analysis with In-Storage Processing

Nika Mansouri Ghiasi[1] Mohammad Sadrosadati[1] Harun Mustafa[1] Arvid Gollwitzer[1]
Can Firtina[1] Julien Eudine[1] Haiyu Mao[1] Joël Lindegger[1] Meryem Banu Cavlak[1]
Mohammed Alser[1] Jisung Park[2] Onur Mutlu[1]
[1]ETH Zürich [2]POSTECH

Metagenomics, the study of the genome sequences of diverse organisms in a common environment, has led to significant advances in many fields. Since the species present in a metagenomic sample are not known in advance, metagenomic analysis commonly involves the key tasks of determining the species present in a sample and their relative abundances. These tasks require searching large metagenomic databases containing information on different species' genomes. Metagenomic analysis

medicine [9,10], urgent clinical settings [11], understanding microbial diversity of an environment [12, 13], discovering early warnings of communicable diseases [14–16], and outbreak tracing [17]. The pivotal role of metagenomics, together with rapid improvements in genome sequencing (e.g., reduced cost and improved throughput [18]), has resulted in the fast-growing adoption of metagenomics [10, 19, 20].

Given a metagenomic sample, a typical workflow consists of

Sparsified Data

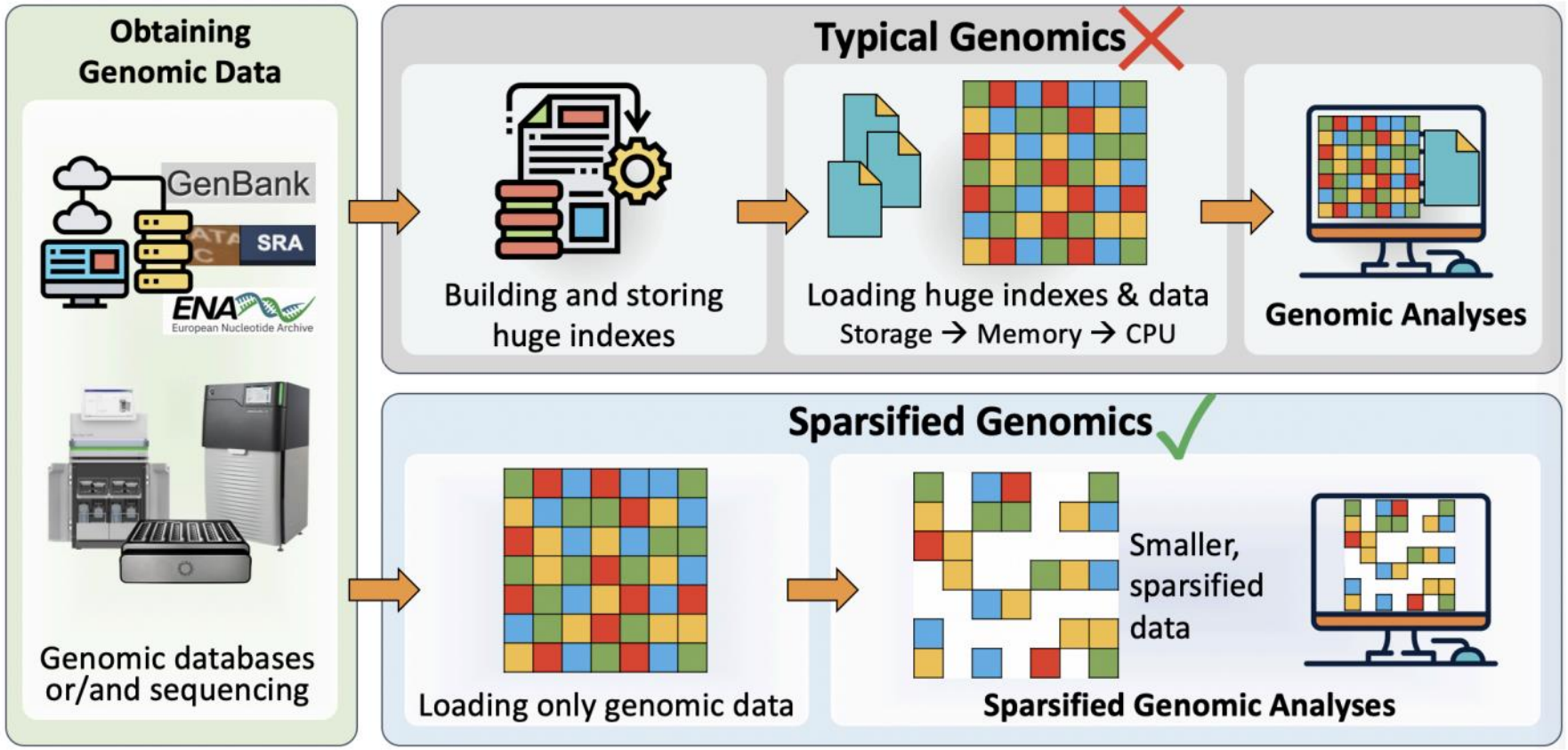
AI Controlled Sparsification

arvidg@ethz.ch

ETH Zurich

Sparsified Data

State-of-the-art computational methods analyzing genomic sequences fail to cope with the exponential growth of genomic sequencing data



References

Alser et al., "Genome-on-Diet: Taming Large-Scale Genomic Analyses via Sparsified Genomics", (2024)

Key Concept: Sparsified Genomics

Sparsified Genomes
= Genome on Diet

Idea: Systematically exclude bases from genomic sequences

- Use a fixed (Genome on diet) pattern or variable pattern (AI-controlled metagenomics) to determine which bases to include or exclude
- For example, a pattern of '10' means every alternate base is included

Goal: Create shorter, sparsified sequences that maintain the essential information for analysis.

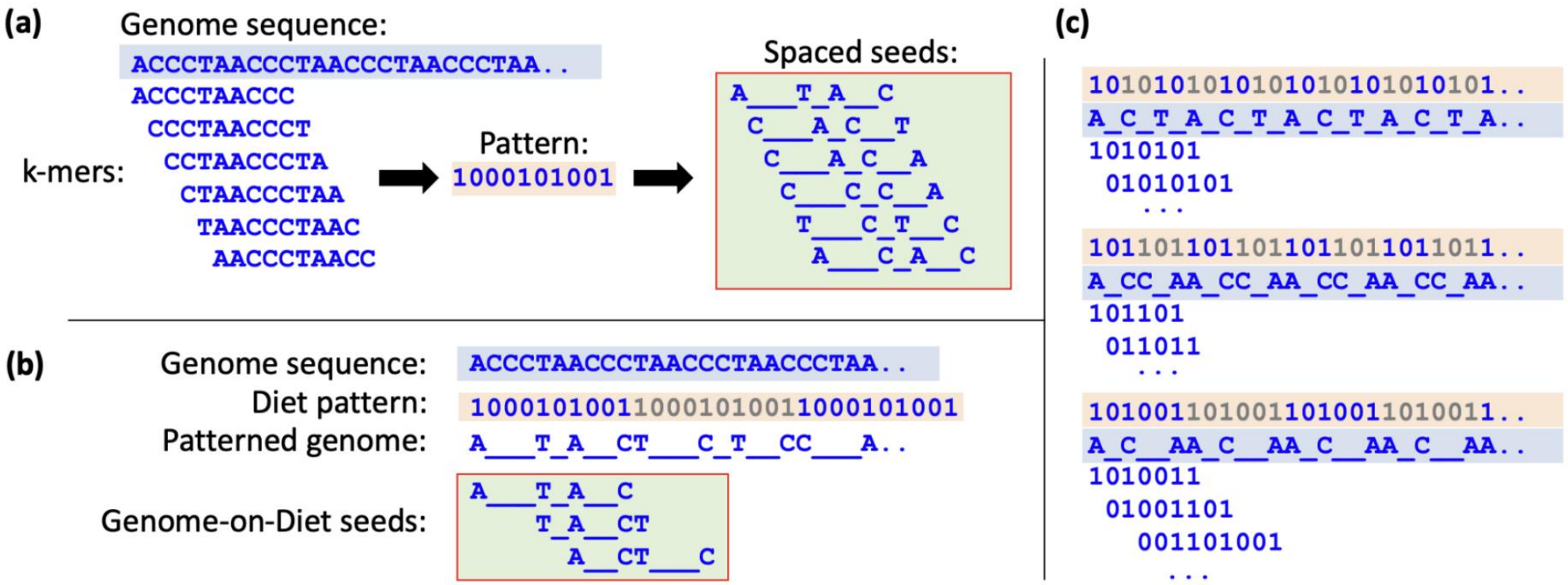
Benefits

- Faster processing: By reducing the input size, genomic analyses are faster
- Reduced memory usage: Smaller genomic sequences lead to reduced memory requirements for indexing and searching
- Higher accuracy: Sparsified sequences enable better accuracy in detecting genomic variations in some cases

References
Alser et al., "Genome-on-Diet: Taming Large-Scale Genomic Analyses via Sparsified Genomics", (2024)

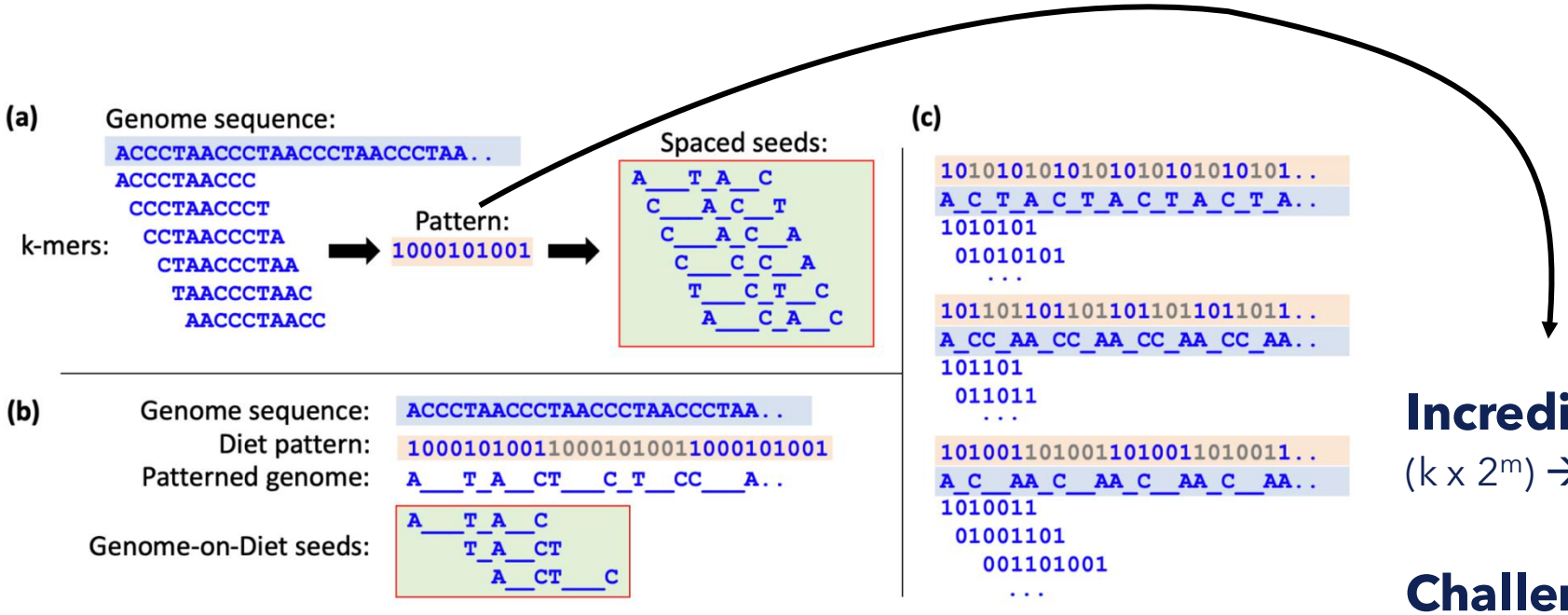
Sparsified Genomics

Genome on Diet



AI Controlled Sparsification

A Novel Mechanism



m read length

k no. of reads

n reference genome length

Incredibly Large Search Space

$(k \times 2^m)$ → Trillions of choices!

Challenge

Find the optimal choice for each read!

Memory Management Strategies

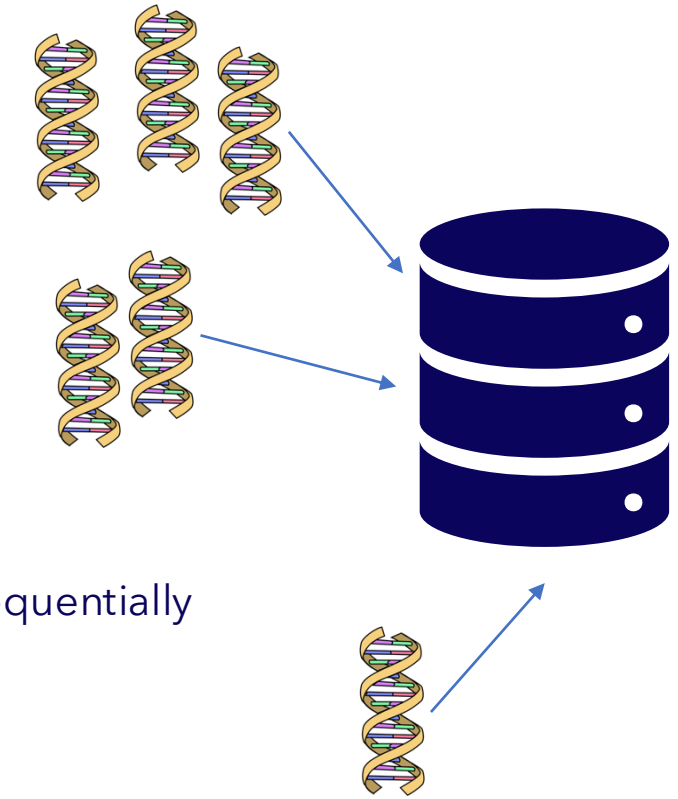
Memory-frugal
reference database index structure

Index Construction
divides the database into batches for efficient parallel processing

Batches
are processed independently, allowing for parallelized index access.

Memory Usage
during the reference database query stage is optimized by processing batches sequentially

Temporary Auxiliary Files
are minimized compared to other tools such as KMC3+CMash



Benefits of AI-Controlled Sparsification

Reduced total execution time

By processing less workload (smaller number of included bases)

Reduced peak memory footprint

Smaller number of extracted seeds and hence a smaller index

No need to pre-build genome indices

Now possible to build it during the analysis with low performance-overhead

Limitation

Might lead to finding a large number of sequences in the reference genome that are similar to the given read sequence possibly some falsely detected variations

Sparsification of data and Computation

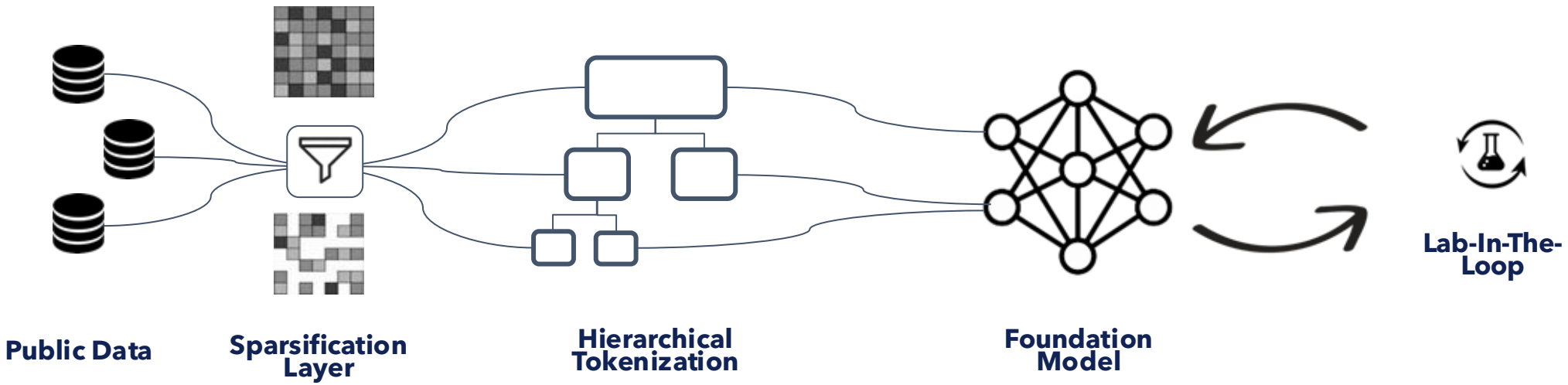
Anto Biosciences

founders@anto.bio

Sparsification is Key

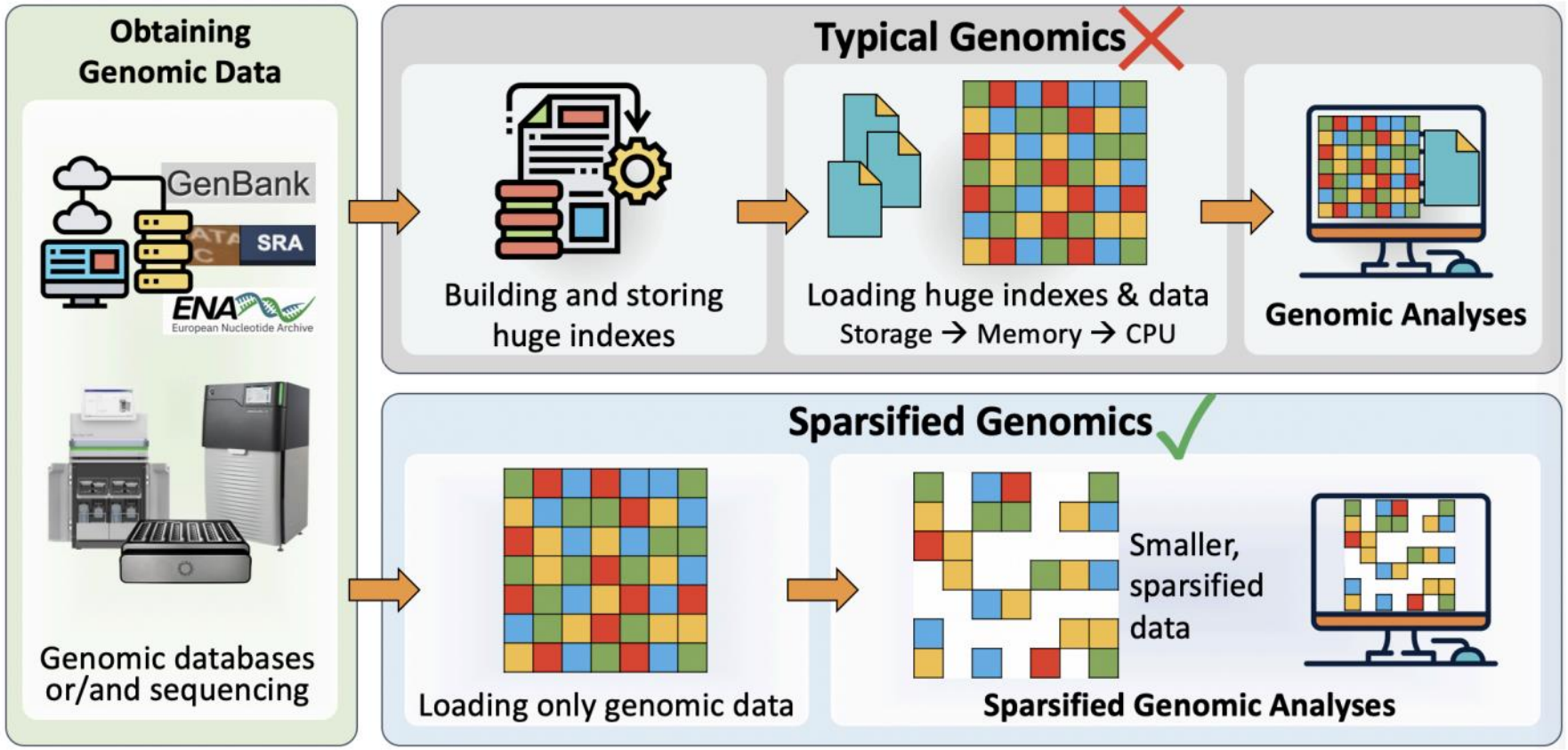
Most public data was previously unusable

We condense it into a high-quality meta-vocabulary



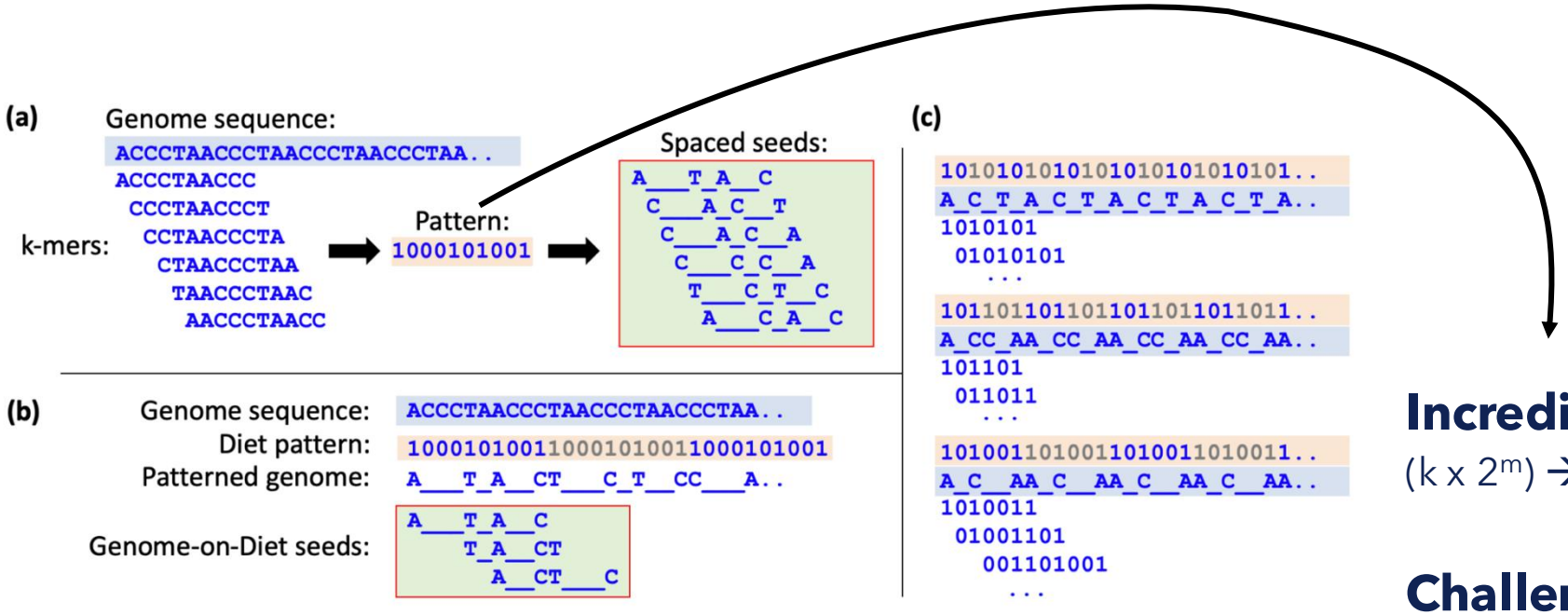
Sparsified Data

State-of-the-art computational methods analyzing genomic sequences fail to cope with the exponential growth of genomic sequencing data



AI Controlled Sparsification

A Novel Mechanism



m read length

k no. of reads

n reference genome length

Incredibly Large Search Space

$(k \times 2^m)$ → Trillions of choices!

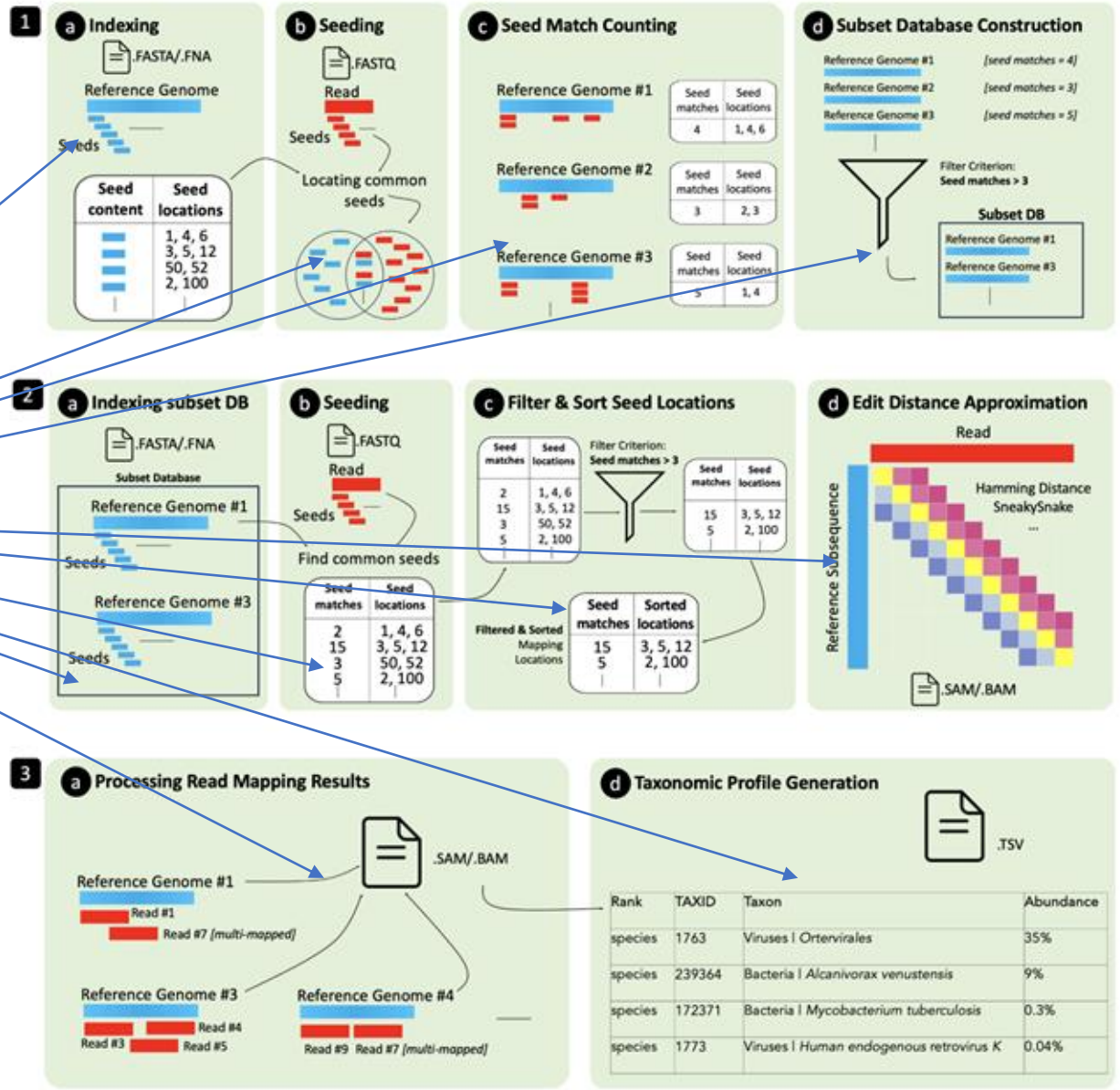
Challenge

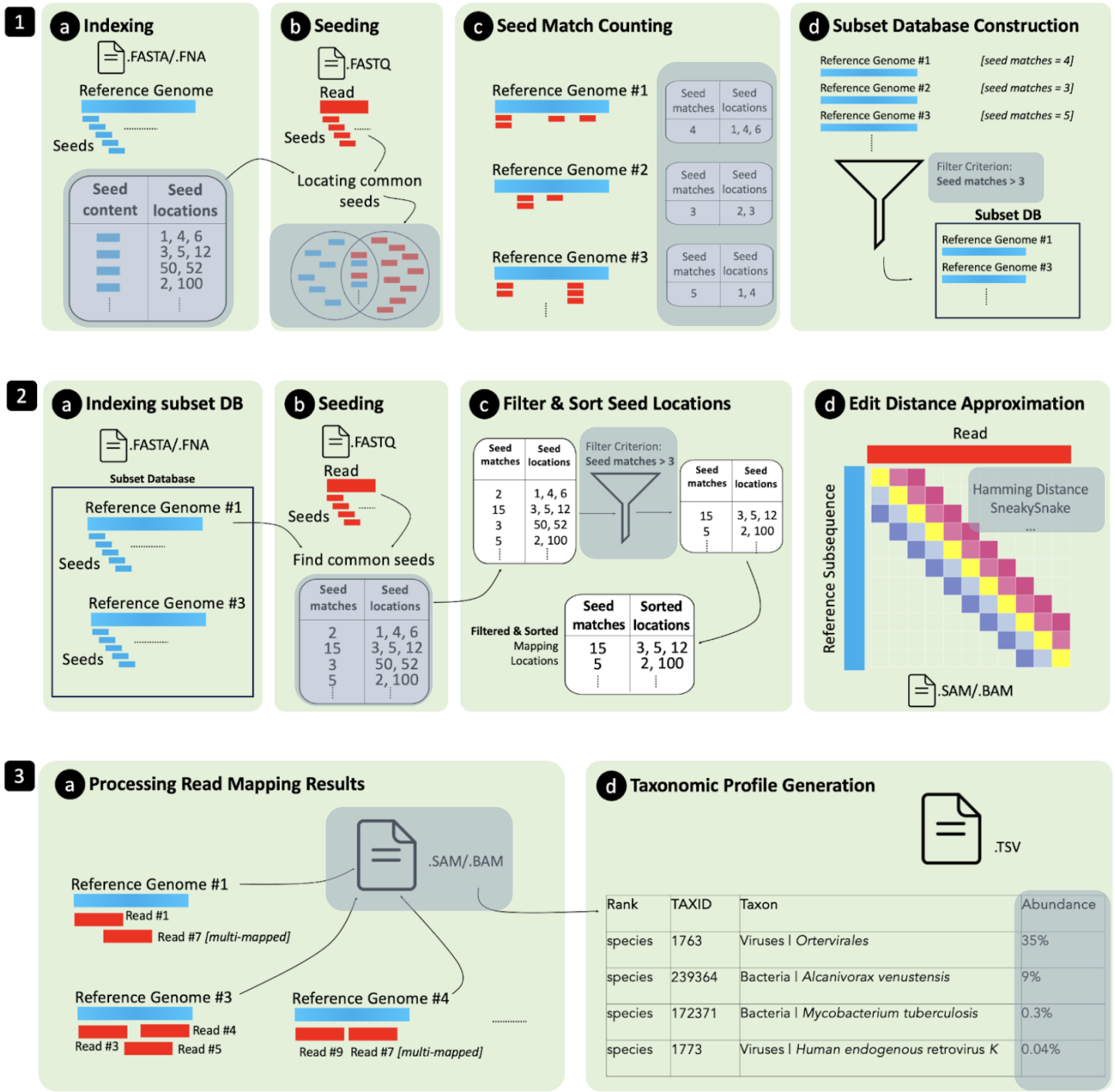
Find the optimal choice for each read!

AI-Controlled Metagenomics

Focusing on what's Important

Control Unit





Control Points (CPs)

Parameters and algorithms chosen on-the-fly

A different choice for every read sequence

Microbiome Biomarkers

Discovering a new class of biomarkers

founders@anto.bio

Altered gut microbiome composition by appendectomy contributes to colorectal cancer

Feiyu Shi, Gaixia Liu, Yufeng Lin, Cosmos liutao Guo, Jing Han, Eagle S. H. Chu, Chengxin Shi, Yaguang Li, Haowei Zhang, Chenhao Hu, Ruihan Liu, Shuixiang

Oncogene, 2022
https://doi.org/10.1038/s41388-022-02569-3

Key Problem
No reliable preventive treatment for colorectal cancer (CRC), one of the most common cancers worldwide

Goal
Identify early diagnosis and prevention measures

Key Idea
There might be a direct link between gut microbial dysbiosis and CRC

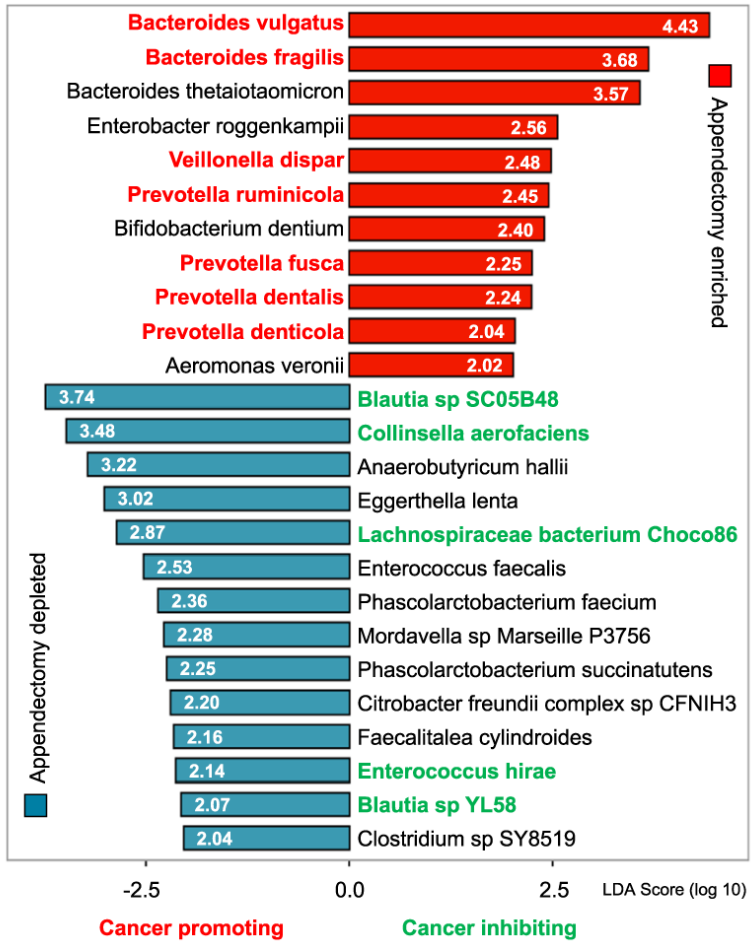
Key Results
Identified seven bacteria potentially causing CRC

Appendectomy and CRC

Appendectomy induces enrichments of CRC-associated species

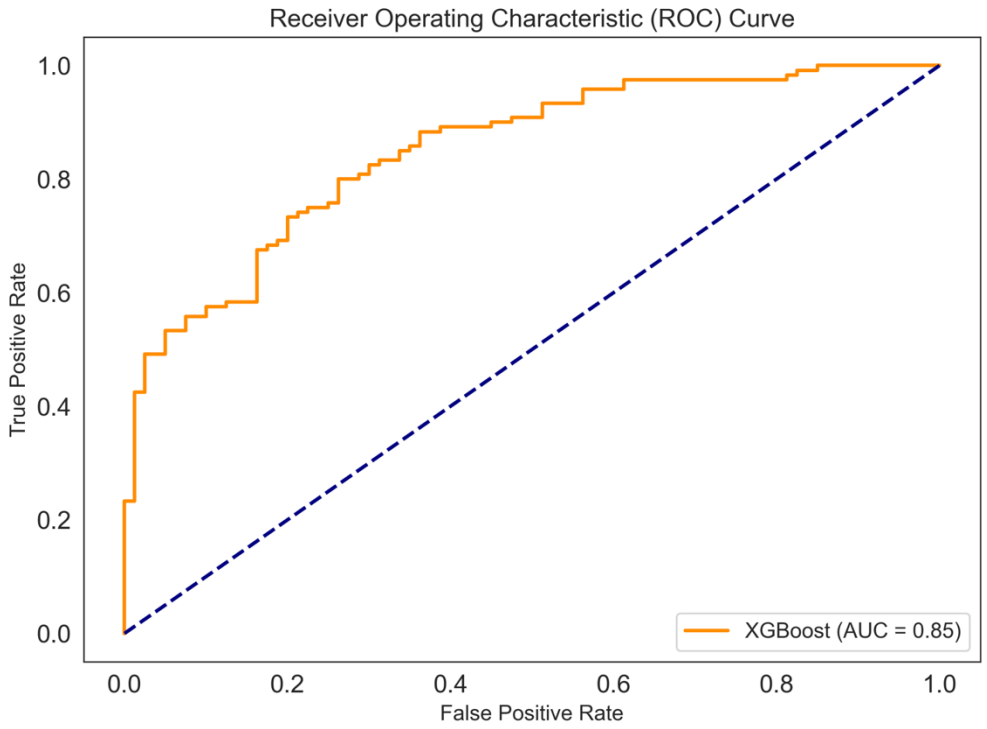
- Evaluating difference in gut microbiome at **species-level**
- **25 bacterial species** identified with significant difference in abundances (11 enriched and 14 depleted)

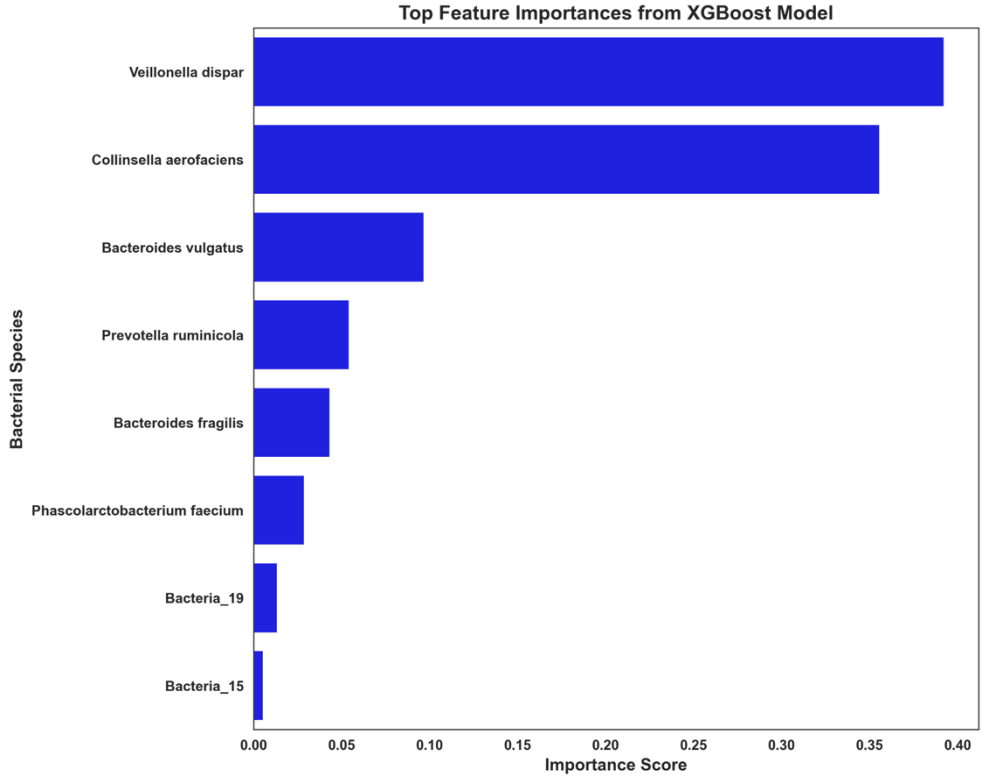
- **11 enriched species**, only 7 CRC- promoting
- Exclude some species, effect known from other studies (assumed not cancer-promoting)

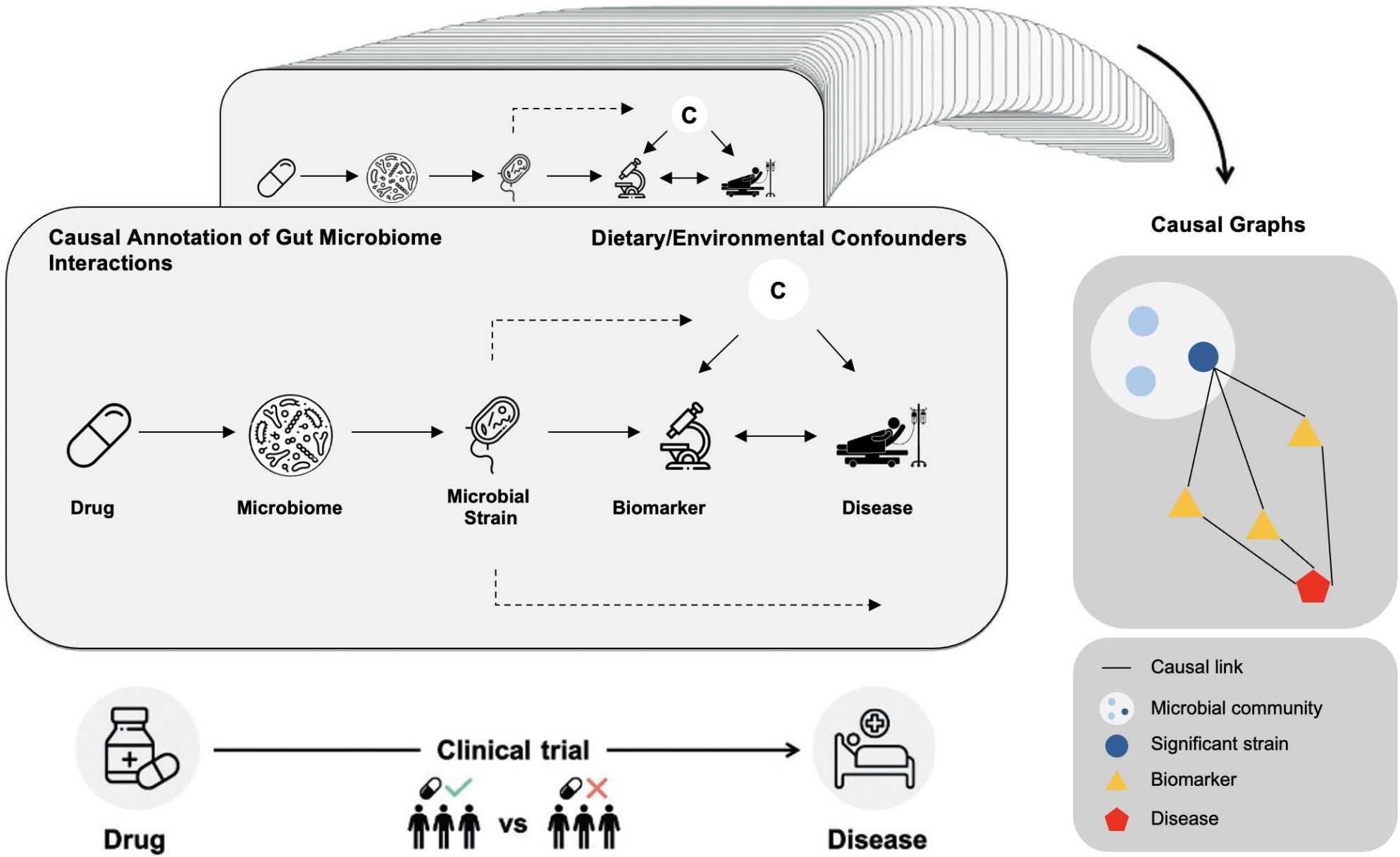


A Bar plot for the differential abundant species
Red: enriched species in appendectomy patients
Blue: depleted species in appendectomy subjects
promoting cancer: red font
cancer-inhibiting species green font

AI Controlled Metagenomics (CRC)






C
Causal Annotation of Gut Microbiome Interactions
Dietary/Environmental Confounders
C
Drug
Microbiome
Microbial Strain
Biomarker
Disease
Causal Graphs
Causal link
Microbial community
Significant strain
Biomarker
Disease
Clinical trial
vs
Drug
Disease

Most Recent Work: Key Results

More at *anto.bio* and *ycombinator.com/companies/anto-biosciences*

Table 1: Downstream task performance for genomic tokenization. Values are means with 95% confidence intervals over $n = 10$ runs.

Method	Variant F1	Taxa F1	Recon. Loss	Time (ms)
Standard BPE	.824±.004	.856±.005	.317±.010	10.0
SentencePiece	.837±.004	.872±.005	.301±.009	10.1
WordPiece	.829±.005	.863±.006	.308±.011	10.0
BPE-dropout	.841±.004	.878±.005	.295±.009	10.2
ByT5	.812±.006	.845±.007	.338±.012	25.3
CANINE	.818±.005	.852±.006	.325±.011	22.7
DNABERT-k	.851±.003	.889±.004	.287±.008	9.8
SuperBPE	.858±.003	.895±.004	.275±.008	10.3
GenTokenizer	.863±.003	.901±.003	.268±.007	10.5
QA-BPE-seq	**.891±.004**	**.917±.003**	**.241±.007**	**10.2**
Hedges' g	*8.2*	*4.3*	*3.5*	-

Table 2: Ablation Study for QA-BPE-seq (Variant F1 Score). Values are means with 95% confidence intervals over $n = 10$ runs.

Configuration	Variant F1	Rel. Change (%)
QA-BPE-seq (Full)	**0.891± 0.004**	**-**
w/o RL Framework (Greedy w_{ab})	0.862± 0.005	−3.3
w/o Quality Component ($R_Q = 0$)	0.825± 0.004	−7.4
w/o Information Reward ($R_I = 0$)	0.872± 0.005	−2.1
w/o Adaptive Params (α, β fixed)	0.857± 0.006	−3.8
w/o R_{bio} (Optional component)	0.885± 0.004	−0.7
QualTok (Ablation Baseline)	0.840± 0.005	−5.7

Table 5: Pathogen Detection benchmark results (MCC scores). QA-Token achieves state-of-the-art.

Model	Task-1	Task-2	Task-3	Task-4	Task-5	Avg
DNABERT	82.15	81.43	83.27	84.62	82.88	82.87
DNABERT-2	86.73	86.90	88.30	89.77	87.90	87.92
DNABERT-S	85.43	85.23	89.01	88.41	86.02	87.02
NT-2.5B-Multi	83.80	83.53	82.48	79.91	81.43	82.43
NT-2.5B-1000g	77.52	80.38	79.83	78.37	78.99	79.02
HyenaDNA	78.65	79.12	80.44	81.23	79.88	79.86
METAGENE-1	92.14	90.91	93.70	95.10	93.96	92.96
+QA-Token	**93.81**	**92.95**	**95.12**	**96.24**	**94.53**	**94.53**
Improvement	+1.67	+2.04	+1.42	+1.14	+0.57	+1.57

Table 6: Genome Understanding Evaluation (GUE): Multi-species benchmark spanning regulatory, structural, and variant analysis tasks.

Task Category	**METAGENE-1**	**QA-Token**	**Δ**	**p-value**
Regulatory Element Prediction				
TF-Mouse (4 tasks, avg. MCC)	71.4 ± 0.8	**72.8 ± 0.7**	+1.4	0.002
TF-Human (4 tasks, avg. MCC)	68.3 ± 0.9	**69.9 ± 0.8**	+1.6	0.001
Promoter Detection (MCC)	82.3 ± 0.5	**85.5 ± 0.4**	+3.2	<0.001
Enhancer Activity (AUC)	0.876 ± 0.012	**0.892 ± 0.010**	+0.016	0.003
Epigenetic Modifications				
H3K4me3 (MCC)	65.2 ± 0.6	**66.8 ± 0.5**	+1.6	0.002
H3K27ac (MCC)	66.8 ± 0.7	**68.2 ± 0.6**	+1.4	0.003
DNA Methylation (AUC)	0.823 ± 0.015	**0.841 ± 0.013**	+0.018	0.004
Structural Features				
Splice Site Detection (F1)	87.8 ± 0.4	**89.5 ± 0.3**	+1.7	<0.001
RNA Secondary Structure	72.1 ± 0.8	**73.9 ± 0.7**	+1.8	0.002
Variant Analysis				
COVID Variant (F1)	72.5 ± 0.6	**73.3 ± 0.5**	+0.8	0.018
SNP Effect Prediction	0.684 ± 0.021	**0.712 ± 0.018**	+0.028	0.001
Global Win Rate	46.4%	**57.1%**	**+10.7%**	-
Token Efficiency	370B tokens	**315B tokens**	**-15%**	-

Full Paper

https://openreview.net/pdf?id=UDvXiEngkX

Most Recent Work

More at *anto.bio* and *ycombinator.com/companies/anto-biosciences*

From Noise to Signal: Enabling Foundation-Model Pretraining on Noisy, Real-World Corpora via Quality-Aware Tokenization

Anonymous authors
Paper under double-blind review

Abstract

Current tokenization methods process sequential data without accounting for signal quality, limiting their effectiveness on noisy real-world corpora. We present *QA-Token (Quality-Aware Tokenization)*, which incorporates data reliability directly into vocabulary construction. Our framework introduces three technical contributions: (i) a bilevel optimization formulation that jointly optimizes vocabulary construction and downstream performance (proven NP-hard), (ii) a reinforcement learning approach that learns merge policies through quality-aware rewards with convergence guarantees, and (iii) an adaptive parameter learning mechanism via Gumbel-Softmax relaxation for end-to-end optimization.

We show that QA-Token achieves information-theoretic optimality under noisy conditions, with convergence guarantees for both policy and parameter learning. Experiments demonstrate consistent improvements: *genomics* (8.9% absolute F1 gain in variant calling, Hedges' $g = 8.2$), *finance* (30% Sharpe ratio improvement). At foundation scale, re-tokenizing METAGENE-1's 1.7 trillion base-pair corpus achieves state-of-the-art pathogen detection (94.53 MCC) while reducing token count by 15%. A 1.2B parameter financial model trained with QA-Token shows 12-27% improvements across forecasting tasks. These results demonstrate that quality-aware tokenization enables effective training on noisy corpora that standard methods cannot handle.

1 Introduction

Tokenization serves as the interface between raw data and neural computation. Current methods such as Byte-Pair Encoding (BPE) Sennrich et al. (2016) rely exclusively on frequency statistics, assuming that occurrence frequency correlates with semantic importance. This assumption fails when data quality varies significantly—from sequencing errors in genomics Ewing et al. (1998) to microstructure noise in financial markets Andersen et al. (2001). Models trained on noisy corpora using frequency-based tokenization inherit these errors, resulting in degraded performance.

The problem is substantial: error rates in third-generation sequencing exceed 10% Wenger et al. (2019), yet current tokenizers treat high-confidence and error-prone regions identically. In finance, over 40% of high-frequency data contains microstructure noise Hansen & Lunde (2006), but tokenization methods do not distinguish signal quality. This limitation constrains foundation model training on real-world data.

We present **Quality-Aware Tokenization (QA-Token)**, a framework that incorporates data quality into vocabulary construction. QA-Token introduces three technical contributions:

1. Bilevel Optimization with Complexity Analysis: We formalize tokenization as a bilevel optimization problem (Definition 1) that jointly optimizes vocabulary construction and downstream performance. We show this problem is NP-hard (Theorem 1) and develop a principled approximation scheme with theoretical guarantees.

2. Reinforcement Learning with Convergence Guarantees: We cast vocabulary construction as a Markov Decision Process (Definition 2) and employ reinforcement learning to discover optimal

https://openreview.net/pdf?id=UDvXiEngkX

HighClass: Efficient Metagenomic Classification via Quality-Aware Token Mapping and Sparsified Indexing

Anonymous authors
Paper under double-blind review

Abstract

Metagenomic classification requires both high accuracy and computational efficiency to process the exponentially growing volume of sequencing data. We present *HighClass*, a novel classification framework that fundamentally transforms the computational paradigm through variable-length token indexing, quality-aware scoring, and learned sparsification.

Our key innovation replaces alignment operations with hash-based token mapping, achieving $O(|\mathcal{T}|)$ complexity while maintaining competitive accuracy. We establish rigorous theoretical foundations: (1) generalization bounds proving $O(\sqrt{|V|\,|\mathcal{Y}|/n})$ convergence for vocabulary size V and $|\mathcal{Y}|$ taxa; (2) concentration inequalities under exponential α-mixing with explicit dependency factors; (3) consistency guarantees for maximum likelihood classification under identifiability conditions.

HighClass achieves 85.1% F1 on CAMI II—within 1.5% of state-of-the-art—while delivering 4.2× speedup and 68% memory reduction. Variable-length tokens provide 6.8 percentage points improvement over fixed k-mers through superior pattern capture. Quality-aware scoring with learned sensitivity $\eta = 1.8$ optimally weights sequencing evidence. Gradient-based sparsification retains 32% of genomic regions while preserving 94% accuracy.

Beyond empirical gains, our work establishes the first comprehensive theory of token-based genomic classification, providing uniform convergence guarantees and explicit characterization of dependency effects through α-mixing analysis. These results transform sequence classification from heuristic approaches to principled methods with provable guarantees.

1 Introduction

Metagenomic sequencing generates unprecedented volumes of data requiring classification at rates exceeding 10^{10} reads per day in clinical and environmental applications (Lloyd-Price et al., 2019; Thompson et al., 2017; Gardy & Loman, 2018). The fundamental challenge is determining the taxonomic origin of each read $X \in \mathcal{X}$ with respect to a reference database $\mathcal{D}$ while maintaining both accuracy and computational tractability.

1.1 The Fundamental Trade-off

Current metagenomic classifiers fall into two paradigmatic categories, each with inherent limitations:

Alignment-based methods solve classification through explicit sequence alignment, typically employing seed-and-extend strategies. For a read $X \in \Sigma^m$ and database $\mathcal{D}$, practical implementations achieve high accuracy but spend most time in seed-and-extend steps with effective per-read cost $\mathcal{O}(m \log n + k \log k)$ where n is index size and k the number of k-mer matches. The computational burden becomes prohibitive for modern datasets exceeding 10^{10} reads.

Alignment-free methods bypass explicit alignment using k-mer indexing or minimizer schemes. While achieving $\mathcal{O}(m)$ query complexity, they sacrifice accuracy through information loss during the fixed-length decomposition, particularly problematic for: (i) reads with heterogeneous quality

https://openreview.net/pdf?id=wkVsKDnl4s

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

Anonymous Author(s)
Affiliation
Address
email

Abstract

We propose **MetaOmics-10T**—an openly shareable, foundational dataset to unlock AI-accelerated discovery in microbial ecosystems. The dataset directly enables three high-impact AI tasks: (1) forecasting ecosystem dynamics, (2) predicting counterfactual outcomes of interventions, and (3) inverse-design of microbial therapies under safety constraints. MetaOmics-10T combines **10 trillion base pairs** reclaimed from public archives using a Quality-Aware Tokenization (QA-Token) framework with **100,000+ interventional trajectories** generated via model-guided experimental design. The result is a first-of-its-kind, probabilistic, intervention-ready corpus that addresses the principal bottleneck for causal modeling in microbiome science and provides an empirical testbed to assess the reach and limits of causal inference at scale.

1 From Observation to Intervention: The Formal Contract for Digital Twins

Proposal at a Glance. *AI task:* forecasting, counterfactual prediction, and safe inverse design. *Rationale:* lack of interventional, quality-aware, multi-omic time series is the core bottleneck for causal modeling. *Dataset:* 10T bp reclaimed from archives + 100,000+ interventional trajectories with full metadata (protocols, doses, timings, quality). *Shareability:* weekly open releases with standardized schemas and ontologies. *Impact:* enables identifiable digital twins, robust counterfactuals, and principled policy synthesis. *Feasibility:* detailed cost, throughput, and experimental SOPs in Appendix D. **Digital Twin Definition.** We model microbial ecosystems as controlled dynamical systems $(\mathcal{S}, \mathcal{U}, \mathcal{T}_\theta, \mathcal{M})$ where $\mathcal{S} \subseteq \mathbb{R}^{n_s}$ is the state space encoding genomic abundances ($g_t \in \mathbb{R}^{n_g}$, $n_g \approx 10^6$) and metabolite concentrations ($m_t \in \mathbb{R}^{n_m}$, $n_m \approx 10^4$), $\mathcal{U} \subseteq \mathbb{R}^{n_u}$ the intervention space (CRISPR edits, compound doses), $\mathcal{T}_\theta : \mathcal{S} \times \mathcal{U} \to \Delta(\mathcal{S})$ the learned stochastic transition kernel parameterized by deep neural networks, and $\mathcal{M} : \mathcal{S} \to \mathcal{Y}$ the measurement map accounting for technical noise. To avoid symbol collisions later with model classes, we denote any *frozen proxy model* used for evaluation by $\mathcal{F}$, never by $\mathcal{M}$. The three core tasks with formal specifications:

- **Forecasting:** Learn $\hat{F}_\theta$ s.t. $\mathbb{E}[\|x_{t+\tau} - \hat{F}_\theta(x_{\le t})\|^2] \le \epsilon_F$ under autonomous dynamics $u_t = 0$.
- **Counterfactuals:** Estimate $p(x_{t+\tau} | do(u), x_{\le t})$ via backdoor adjustment when confounders Z are measured.
- **Inverse design:** Solve constrained optimization $u^* = \arg\min_{u \in \mathcal{U}} C(u) + \lambda\, d(\mathbb{E}[x_{t+\tau} | do(u), x_t], x^*)$ subject to safety constraints $g(u) \le 0$ and an uncertainty-aware trust region $D(\pi_{\text{beh}}, u) \le \rho$ (with D a divergence, e.g., Wasserstein or KL, ensuring safe extrapolation from observed actions), together with a chance constraint $\mathbb{P}(g(u) \le 0) \ge 1 - \alpha$ under model uncertainty.

https://openreview.net/pdf?id=IsJPrLpbvf

Understanding The Microbiome

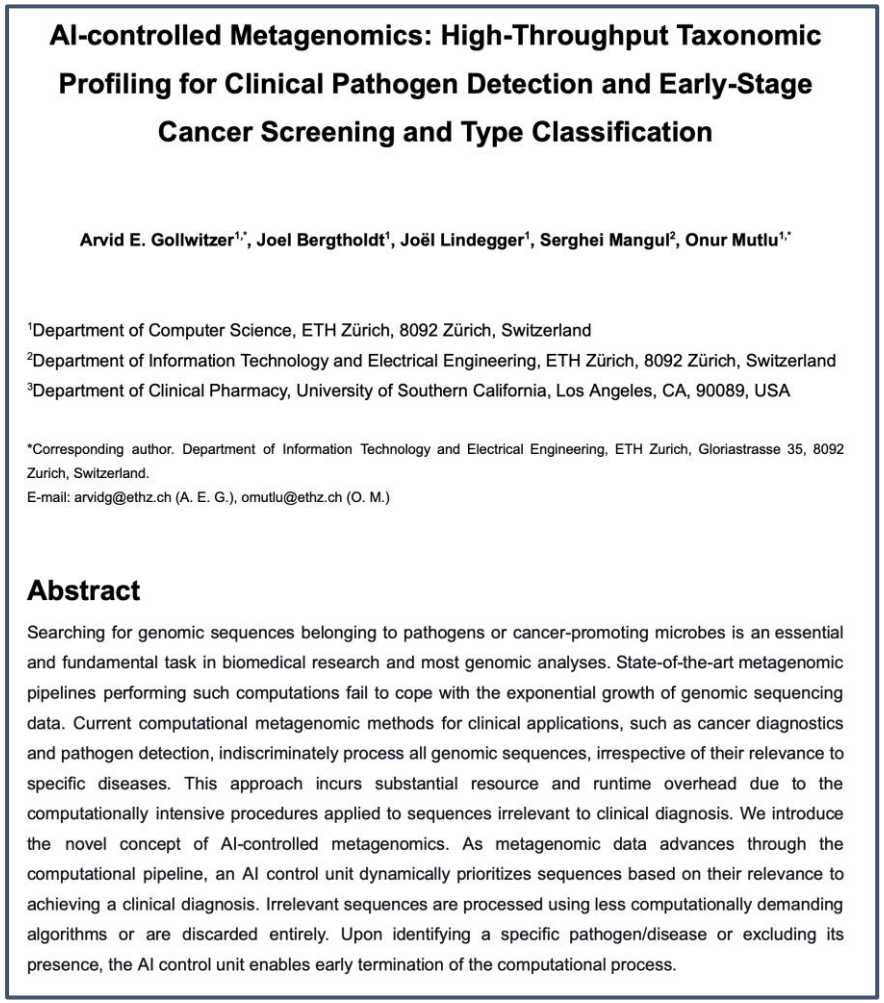
Enabling High-throughput Screening For Novel Biomarkers

arvid@anto.bio

End-to-end Data Processing Pipeline

From Patient Samples to Digital Biomarker

FASTQ File
From lab



AI-controlled Metagenomics: High-Throughput Taxonomic Profiling for Clinical Pathogen Detection and Early-Stage Cancer Screening and Type Classification

Arvid E. Gollwitzer[1,*], Joel Bergtholdt[1], Joël Lindegger[1], Serghei Mangul[2], Onur Mutlu[1,*]

[1]Department of Computer Science, ETH Zürich, 8092 Zürich, Switzerland
[2]Department of Information Technology and Electrical Engineering, ETH Zürich, 8092 Zürich, Switzerland
[3]Department of Clinical Pharmacy, University of Southern California, Los Angeles, CA, 90089, USA

*Corresponding author. Department of Information Technology and Electrical Engineering, ETH Zurich, Gloriastrasse 35, 8092 Zurich, Switzerland.
E-mail: arvidg@ethz.ch (A. E. G.), omutlu@ethz.ch (O. M.)

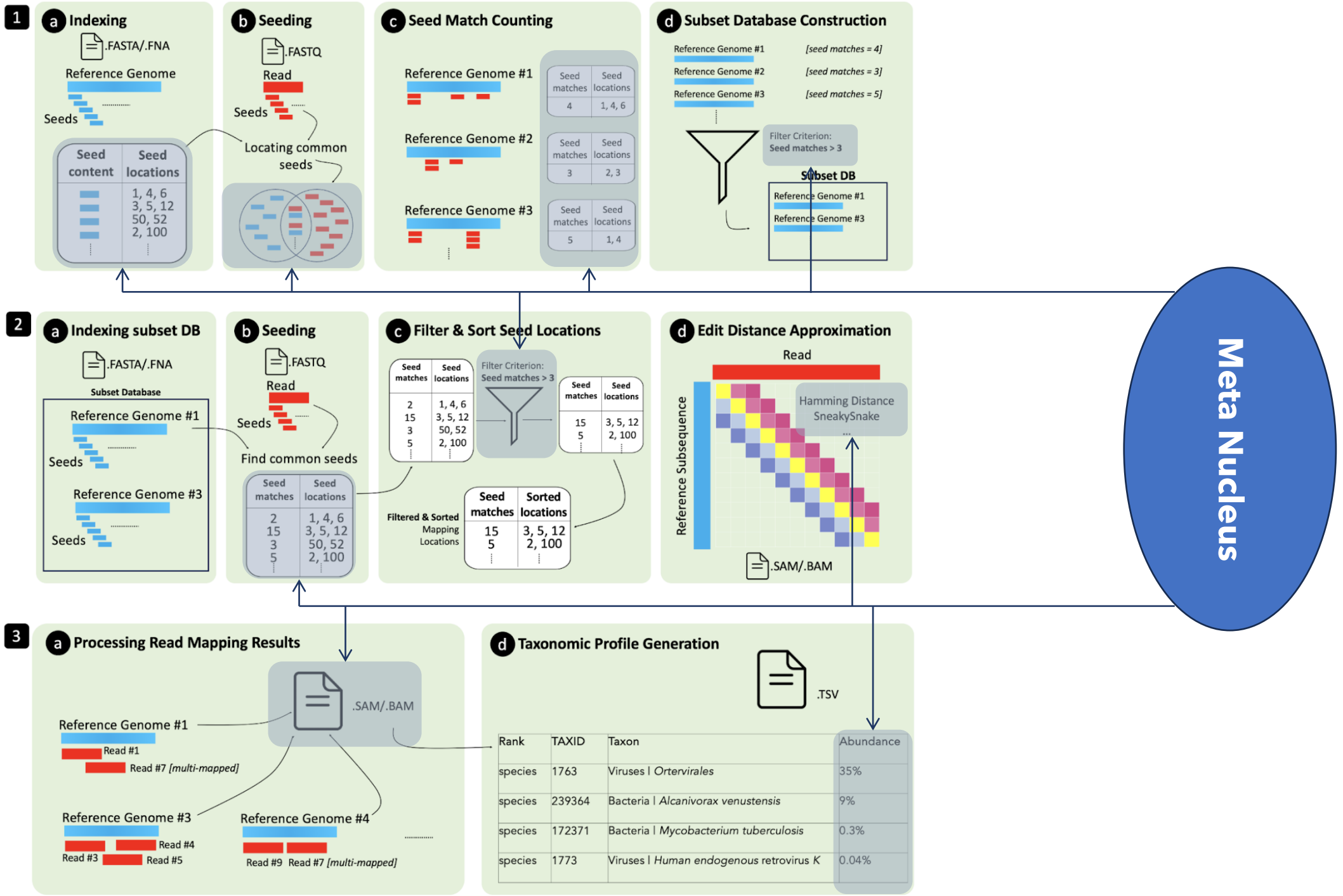
Abstract

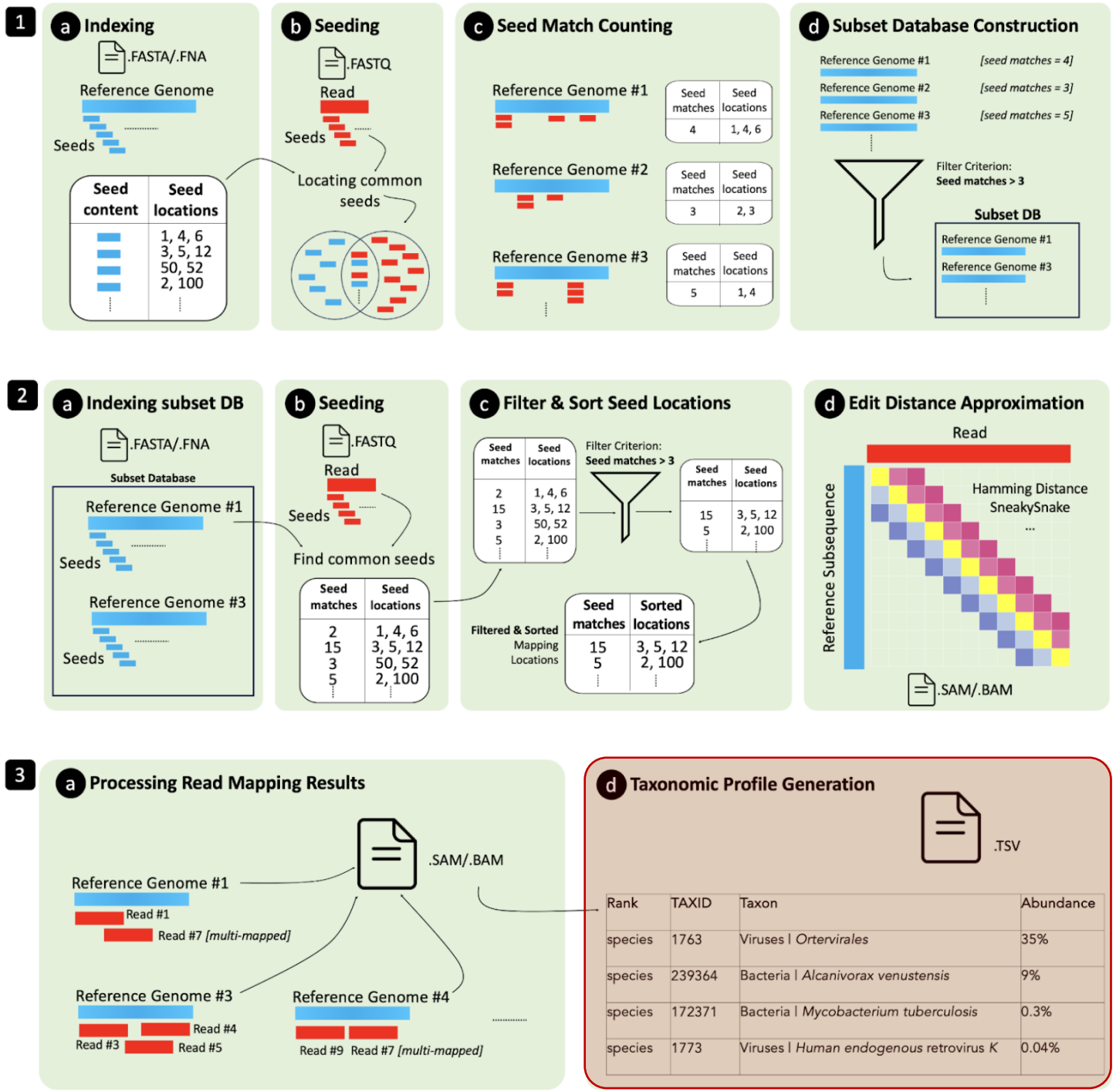
Searching for genomic sequences belonging to pathogens or cancer-promoting microbes is an essential and fundamental task in biomedical research and most genomic analyses. State-of-the-art metagenomic pipelines performing such computations fail to cope with the exponential growth of genomic sequencing data. Current computational metagenomic methods for clinical applications, such as cancer diagnostics and pathogen detection, indiscriminately process all genomic sequences, irrespective of their relevance to specific diseases. This approach incurs substantial resource and runtime overhead due to the computationally intensive procedures applied to sequences irrelevant to clinical diagnosis. We introduce the novel concept of AI-controlled metagenomics. As metagenomic data advances through the computational pipeline, an AI control unit dynamically prioritizes sequences based on their relevance to achieving a clinical diagnosis. Irrelevant sequences are processed using less computationally demanding algorithms or are discarded entirely. Upon identifying a specific pathogen/disease or excluding its presence, the AI control unit enables early termination of the computational process.

Taxonomy
Basis for new microbial Biomarkers

ML
Quantify Disease Intensity + Stage of cancer

Identify ‘digital’ microbial Biomarkers


1
a Indexing
.FASTA/.FNA
Reference Genome
Seeds
Seed content
Seed locations
1, 4, 6
3, 5, 12
50, 52
2, 100
b Seeding
.FASTQ
Read
Seeds
Locating common seeds
c Seed Match Counting
Reference Genome #1
Reference Genome #2
Reference Genome #3
Seed matches
Seed locations
4
1, 4, 6
3
2, 3
5
1, 4
d Subset Database Construction
Reference Genome #1 [seed matches = 4]
Reference Genome #2 [seed matches = 3]
Reference Genome #3 [seed matches = 5]
Filter Criterion:
Seed matches > 3
Subset DB
Reference Genome #1
Reference Genome #3
2
a Indexing subset DB
.FASTA/.FNA
Subset Database
Reference Genome #1
Seeds
Reference Genome #3
Seeds
b Seeding
.FASTQ
Read
Seeds
Find common seeds
Seed matches
Seed locations
2
15
3
5
1, 4, 6
3, 5, 12
50, 52
2, 100
c Filter & Sort Seed Locations
Filter Criterion:
Seed matches > 3
15
5
3, 5, 12
2, 100
Filtered & Sorted Mapping Locations
Sorted locations
d Edit Distance Approximation
Read
Reference Subsequence
Hamming Distance
SneakySnake
.SAM/.BAM
Meta Nucleus
3
a Processing Read Mapping Results
.SAM/.BAM
Reference Genome #1
Read #1
Read #7 [multi-mapped]
Reference Genome #3
Read #4
Read #3
Read #5
Reference Genome #4
Read #9
Read #7 [multi-mapped]
d Taxonomic Profile Generation
.TSV
Rank
TAXID
Taxon
Abundance
species
1763
Viruses | Ortervirales
35%
species
239364
Bacteria | Alcanivorax venustensis
9%
species
172371
Bacteria | Mycobacterium tuberculosis
0.3%
species
1773
Viruses | Human endogenous retrovirus K
0.04%

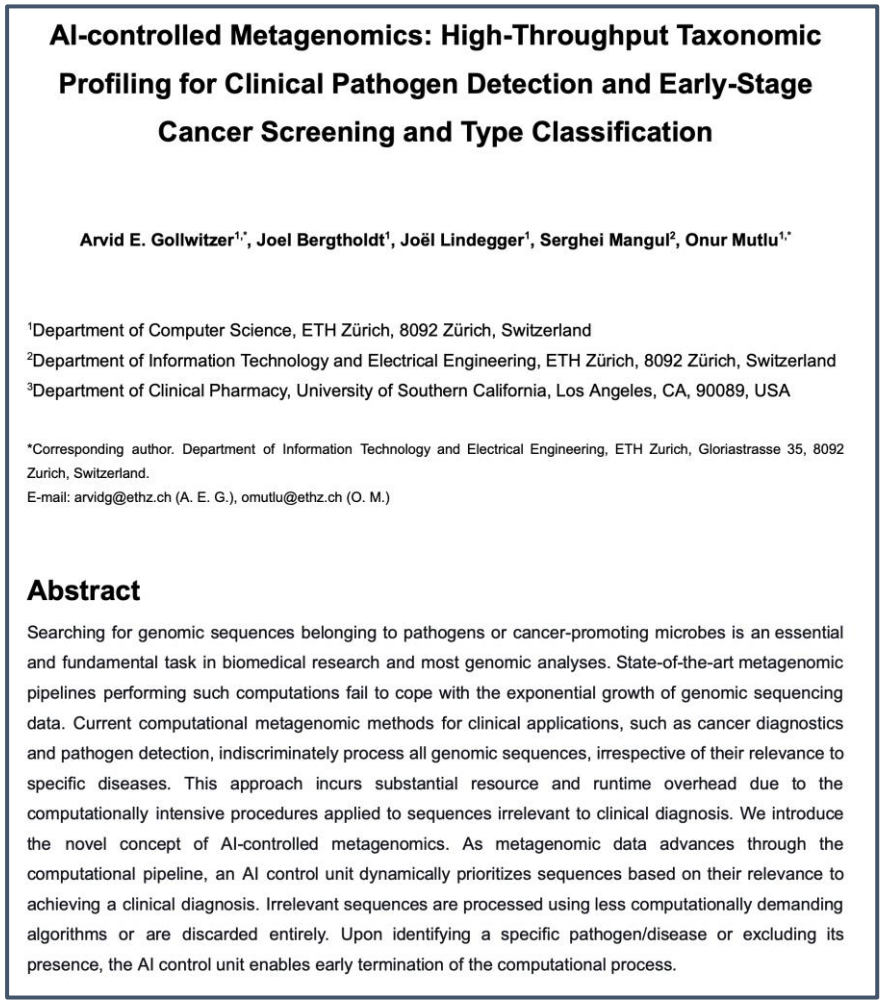

1
a Indexing
.FASTA/.FNA
Reference Genome
Seeds
Seed content
Seed locations
1, 4, 6
3, 5, 12
50, 52
2, 100
b Seeding
.FASTQ
Read
Seeds
Locating common seeds
c Seed Match Counting
Reference Genome #1
Seed matches
Seed locations
4
1, 4, 6
Reference Genome #2
3
2, 3
Reference Genome #3
5
1, 4
d Subset Database Construction
Reference Genome #1
[seed matches = 4]
Reference Genome #2
[seed matches = 3]
Reference Genome #3
[seed matches = 5]
Filter Criterion:
Seed matches > 3
Subset DB
Reference Genome #1
Reference Genome #3
2
a Indexing subset DB
.FASTA/.FNA
Subset Database
Reference Genome #1
Seeds
Reference Genome #3
Seeds
b Seeding
.FASTQ
Read
Seeds
Find common seeds
Seed matches
Seed locations
2
15
3
5
1, 4, 6
3, 5, 12
50, 52
2, 100
c Filter & Sort Seed Locations
Filter Criterion:
Seed matches > 3
15
5
3, 5, 12
2, 100
Filtered & Sorted Mapping Locations
Sorted locations
d Edit Distance Approximation
Read
Reference Subsequence
Hamming Distance
SneakySnake
...
.SAM/.BAM
3
a Processing Read Mapping Results
.SAM/.BAM
Reference Genome #1
Read #1
Read #7 [multi-mapped]
Reference Genome #3
Read #4
Read #3
Read #5
Reference Genome #4
Read #9
Read #7 [multi-mapped]
d Taxonomic Profile Generation
.TSV
Rank
TAXID
Taxon
Abundance
species
1763
Viruses | Ortervirales
35%
species
239364
Bacteria | Alcanivorax venustensis
9%
species
172371
Bacteria | Mycobacterium tuberculosis
0.3%
species
1773
Viruses | Human endogenous retrovirus K
0.04%

```
172045.11.1 strain  2|976|117743|200644|49546|308865|172045|172045.11.1 Bacteria|Bacteroidetes|Flavobacteriia|Flavobacteriales|Flavobacteriaceae|Elizabethkingia|Elizabethkingia miricola|  65.62351    172045.11.1 172045
172045.4.1  strain  2|976|117743|200644|49546|308865|172045|172045.4.1  Bacteria|Bacteroidetes|Flavobacteriia|Flavobacteriales|Flavobacteriaceae|Elizabethkingia|Elizabethkingia miricola|  31.87161    172045.4.1  172045
65700.0.1   strain  2|1224|1236|91347|1903409|551|65700|65700.0.1   Bacteria|Proteobacteria|Gammaproteobacteria|Enterobacterales|Erwiniaceae|Erwinia|Erwinia tracheiphila|  1.18743 65700.0.1   65700
172045.10.1 strain  2|976|117743|200644|49546|308865|172045|172045.10.1 Bacteria|Bacteroidetes|Flavobacteriia|Flavobacteriales|Flavobacteriaceae|Elizabethkingia|Elizabethkingia miricola|  0.4113  172045.10.1 172045
1044999.1   strain  2|1224|1236|91347|1903409|551|65700|1044999.1   Bacteria|Proteobacteria|Gammaproteobacteria|Enterobacterales|Erwiniaceae|Erwinia|Erwinia tracheiphila|Erwinia tracheiphila PSU-1    0.20906 1044999.1   1044999
29320.0.1   strain  2|201174|1760|85006|1268|1742992|29320|29320.0.1    Bacteria|Actinobacteria|Actinobacteria|Micrococcales|Micrococcaceae|Paenarthrobacter|Paenarthrobacter nicotinovorans|   0.10492 29320.0.1   29320
1736310.0   strain  2|201174|1760|85007|85025|1827|1736310|1736310.0    Bacteria|Actinobacteria|Actinobacteria|Corynebacteriales|Nocardiaceae|Rhodococcus|Rhodococcus sp. Leaf258|Rhodococcus sp. Leaf258 unknown strain    0.09178 1736310
1646373.0   strain  2|1224|1236|91347|1903411|1565532|1646373|1646373.0 Bacteria|Proteobacteria|Gammaproteobacteria|Enterobacterales|Yersiniaceae|Rouxiella|Rouxiella silvae|Rouxiella silvae unknown strain    0.08023 1646373.0   1646373
1736266.0   strain  2|1224|28216|80840|75682|75654|1736266|1736266.0    Bacteria|Proteobacteria|Betaproteobacteria|Burkholderiales|Oxalobacteraceae|Duganella|Duganella sp. Leaf126|Duganella sp. Leaf126 unknown strain    0.05876 1736266
1736357.0   strain  2|1224|28211|356|82115|379|1736357|1736357.0    Bacteria|Proteobacteria|Alphaproteobacteria|Rhizobiales|Rhizobiaceae|Rhizobium|Rhizobium sp. Leaf383|Rhizobium sp. Leaf383 unknown strain   0.05303 1736357.0   1736357
1947487.0   strain  2|976|117747|200666|84566|28453|1947487|1947487.0   Bacteria|Bacteroidetes|Sphingobacteriia|Sphingobacteriales|Sphingobacteriaceae|Sphingobacterium|Sphingobacterium sp. UBA2074|Sphingobacterium sp. UBA2074 unknown s
1947513.0   strain  2|976|117747|200666|84566|28453|1947513|1947513.0   Bacteria|Bacteroidetes|Sphingobacteriia|Sphingobacteriales|Sphingobacteriaceae|Sphingobacterium|Sphingobacterium sp. UBA6746|Sphingobacterium sp. UBA6746 unknown s
65700.1.1   strain  2|1224|1236|91347|1903409|551|65700|65700.1.1   Bacteria|Proteobacteria|Gammaproteobacteria|Enterobacterales|Erwiniaceae|Erwinia|Erwinia tracheiphila|  0.03024 65700.1.1   65700
279824.0    strain  2|976|768503|768507|563798|246875|279824|279824.0   Bacteria|Bacteroidetes|Cytophagia|Cytophagales|Cyclobacteriaceae|Algoriphagus|Algoriphagus alkaliphilus|Algoriphagus alkaliphilus unknown strain    0.01962 279824.
170623.4.1  strain  2|1224|1236|72274|135621|352|170623|170623.4.1  Bacteria|Proteobacteria|Gammaproteobacteria|Pseudomonadales|Pseudomonadaceae|Azotobacter|Azotobacter beijerinckii|  0.01662 170623.4.1  170623
29320.1.1   strain  2|201174|1760|85006|1268|1742992|29320|29320.1.1    Bacteria|Actinobacteria|Actinobacteria|Micrococcales|Micrococcaceae|Paenarthrobacter|Paenarthrobacter nicotinovorans|   0.0166  29320.1.1   29320
1736356.0   strain  2|201174|1760|1643682|85030|88138|1736356|1736356.0 Bacteria|Actinobacteria|Actinobacteria|Geodermatophilales|Geodermatophilaceae|Modestobacter|Modestobacter sp. Leaf380|Modestobacter sp. Leaf380 unknown strain  0.0
1909294.62.1    strain  2|1224|28211|356|||1909294|1909294.62.1 Bacteria|Proteobacteria|Alphaproteobacteria|Rhizobiales|||Rhizobiales bacterium|    0.01293 1909294.62.1    1909294
1909294.487.1   strain  2|1224|28211|356|||1909294|1909294.487.1    Bacteria|Proteobacteria|Alphaproteobacteria|Rhizobiales|||Rhizobiales bacterium|    0.01113 1909294.487.1   1909294
1294142.1   strain  2|1239|186801|186802|31979|1485|36845|1294142.1 Bacteria|Firmicutes|Clostridia|Clostridiales|Clostridiaceae|Clostridium|Clostridium intestinale|Clostridium intestinale URNW    0.01073 1294142.1   1294142
1123498.1   strain  2|1224|1236|135623|641|662|184755|1123498.1 Bacteria|Proteobacteria|Gammaproteobacteria|Vibrionales|Vibrionaceae|Vibrio|Vibrio ruber|Vibrio ruber DSM 16370 0.01059 1123498.1   1123498
1121405.0.1 strain  2|1224|28221|213118|213119|896|897|1121405.0.1  Bacteria|Proteobacteria|Deltaproteobacteria|Desulfobacterales|Desulfobacteraceae|Desulfococcus|Desulfococcus multivorans|Desulfococcus multivorans DSM 2059 0.01043 112
1736520.0   strain  2|1224|28211|204458|76892|75|1736520|1736520.0  Bacteria|Proteobacteria|Alphaproteobacteria|Caulobacterales|Caulobacteraceae|Caulobacter|Caulobacter sp. Root343|Caulobacter sp. Root343 unknown strain 0.00773 1736520
1121345.1   strain  2|1239|186801|186802|186803|1843210|100134|1121345.1    Bacteria|Firmicutes|Clostridia|Clostridiales|Lachnospiraceae|Anaerocolumna|Anaerocolumna xylanovorans|Anaerocolumna xylanovorans DSM 12503  0.00675 1121345.1
49280.0.1   strain  2|976|117743|200644|49546|49279|49280|49280.0.1 Bacteria|Bacteroidetes|Flavobacteriia|Flavobacteriales|Flavobacteriaceae|Gelidibacter|Gelidibacter algens|  0.00642 49280.0.1   49280
1909294.429.1   strain  2|1224|28211|356|||1909294|1909294.429.1    Bacteria|Proteobacteria|Alphaproteobacteria|Rhizobiales|||Rhizobiales bacterium|    0.00525 1909294.429.1   1909294
49280.1.1   strain  2|976|117743|200644|49546|49279|49280|49280.1.1 Bacteria|Bacteroidetes|Flavobacteriia|Flavobacteriales|Flavobacteriaceae|Gelidibacter|Gelidibacter algens|  0.00516 49280.1.1   49280
198092.0    strain  2|1224|28211|204441|433|125216|198092|198092.0  Bacteria|Proteobacteria|Alphaproteobacteria|Rhodospirillales|Acetobacteraceae|Roseomonas|Roseomonas rosea|Roseomonas rosea unknown strain   0.00486 198092.0    198092
1736519.0   strain  2|1224|28211|204458|76892|75|1736519|1736519.0  Bacteria|Proteobacteria|Alphaproteobacteria|Caulobacterales|Caulobacteraceae|Caulobacter|Caulobacter sp. Root342|Caulobacter sp. Root342 unknown strain 0.00449 1736519
1736225.0   strain  2|1224|1236|91347|1903409|551|1736225|1736225.0 Bacteria|Proteobacteria|Gammaproteobacteria|Enterobacterales|Erwiniaceae|Erwinia|Erwinia sp. Leaf53|Erwinia sp. Leaf53 unknown strain   0.00397 1736225.0   1736225
1736600.0   strain  2|201174|1760|85009|85015|1839|1736600|1736600.0    Bacteria|Actinobacteria|Actinobacteria|Propionibacteriales|Nocardioidaceae|Nocardioides|Nocardioides sp. Root79|Nocardioides sp. Root79 unknown strain  0.00389 173
1561024.0   strain  2|1224|1236|91347|1903409|551|1561024|1561024.0 Bacteria|Proteobacteria|Gammaproteobacteria|Enterobacterales|Erwiniaceae|Erwinia|Erwinia sp. B116|Erwinia sp. B116 unknown strain   0.00295 1561024.0   1561024
1736590.0   strain  2|1224|28216|80840|80864|12916|1736590|1736590.0    Bacteria|Proteobacteria|Betaproteobacteria|Burkholderiales|Comamonadaceae|Acidovorax|Acidovorax sp. Root70|Acidovorax sp. Root70 unknown strain 0.00183 1736590.0
1736431.0   strain  2|201174|1760|85009|85015|1839|1736431|1736431.0    Bacteria|Actinobacteria|Actinobacteria|Propionibacteriales|Nocardioidaceae|Nocardioides|Nocardioides sp. Root122|Nocardioides sp. Root122 unknown strain    0.00175
429344.0    strain  2|976|117743|200644|49546|252356|429344|429344.0   Bacteria|Bacteroidetes|Flavobacteriia|Flavobacteriales|Flavobacteriaceae|Maribacter|Maribacter polysiphoniae|Maribacter polysiphoniae unknown strain    0.00143 429
1736272.0   strain  2|1224|28216|80840|75682|149698|1736272|1736272.0   Bacteria|Proteobacteria|Betaproteobacteria|Burkholderiales|Oxalobacteraceae|Massilia|Massilia sp. Leaf139|Massilia sp. Leaf139 unknown strain   0.00142 1736272.0
598467.0    strain  2|1224|1236|91347|1903410|71655|598467|598467.0 Bacteria|Proteobacteria|Gammaproteobacteria|Enterobacterales|Pectobacteriaceae|Brenneria|Brenneria sp. EniD312|Brenneria sp. EniD312 unknown strain 0.00135 598467.0
728066.0    strain  2|201174|1760|85006|1268|1742993|728066|728066.0    Bacteria|Actinobacteria|Actinobacteria|Micrococcales|Micrococcaceae|Pseudarthrobacter|Pseudarthrobacter equi|Pseudarthrobacter equi unknown strain  0.00107 728066.
1736409.0   strain  2|201174|1760|85009|85015|1839|1736409|1736409.0    Bacteria|Actinobacteria|Actinobacteria|Propionibacteriales|Nocardioidaceae|Nocardioides|Nocardioides sp. Soil777|Nocardioides sp. Soil777 unknown strain    0.00094
1587522.0   strain  2|201174|1760|85007|85025|1827|1587522|1587522.0    Bacteria|Actinobacteria|Actinobacteria|Corynebacteriales|Nocardiaceae|Rhodococcus|Rhodococcus sp. MEB064|Rhodococcus sp. MEB064 unknown strain  0.00093 1587522.0
1909294.908.1   strain  2|1224|28211|356|||1909294|1909294.908.1    Bacteria|Proteobacteria|Alphaproteobacteria|Rhizobiales|||Rhizobiales bacterium|    0.00091 1909294.908.1   1909294
1188249.1   strain  2|1224|28211|204455|31989|1653176|402884|1188249.1  Bacteria|Proteobacteria|Alphaproteobacteria|Rhodobacterales|Rhodobacteraceae|Cereibacter|Cereibacter changlensis|Cereibacter changlensis JA139  0.00087 1188249.1
```

End-to-end Data Processing Pipeline

From Patient Samples to Digital Biomarker

FASTQ File
From lab



AI-controlled Metagenomics: High-Throughput Taxonomic Profiling for Clinical Pathogen Detection and Early-Stage Cancer Screening and Type Classification

Arvid E. Gollwitzer[1,*], Joel Bergtholdt[1], Joël Lindegger[1], Serghei Mangul[2], Onur Mutlu[1,*]

[1]Department of Computer Science, ETH Zürich, 8092 Zürich, Switzerland
[2]Department of Information Technology and Electrical Engineering, ETH Zürich, 8092 Zürich, Switzerland
[3]Department of Clinical Pharmacy, University of Southern California, Los Angeles, CA, 90089, USA

*Corresponding author. Department of Information Technology and Electrical Engineering, ETH Zurich, Gloriastrasse 35, 8092 Zurich, Switzerland.
E-mail: arvidg@ethz.ch (A. E. G.), omutlu@ethz.ch (O. M.)

Abstract

Searching for genomic sequences belonging to pathogens or cancer-promoting microbes is an essential and fundamental task in biomedical research and most genomic analyses. State-of-the-art metagenomic pipelines performing such computations fail to cope with the exponential growth of genomic sequencing data. Current computational metagenomic methods for clinical applications, such as cancer diagnostics and pathogen detection, indiscriminately process all genomic sequences, irrespective of their relevance to specific diseases. This approach incurs substantial resource and runtime overhead due to the computationally intensive procedures applied to sequences irrelevant to clinical diagnosis. We introduce the novel concept of AI-controlled metagenomics. As metagenomic data advances through the computational pipeline, an AI control unit dynamically prioritizes sequences based on their relevance to achieving a clinical diagnosis. Irrelevant sequences are processed using less computationally demanding algorithms or are discarded entirely. Upon identifying a specific pathogen/disease or excluding its presence, the AI control unit enables early termination of the computational process.

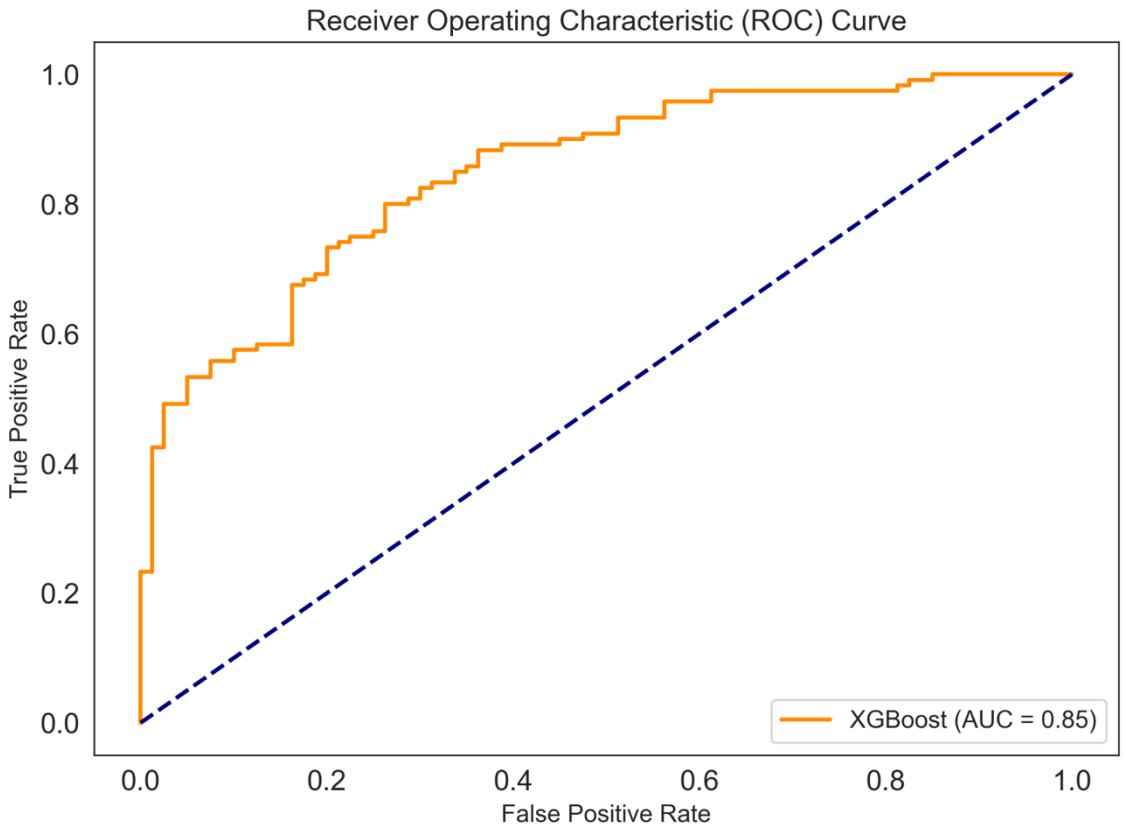
Taxonomy
Basis for new microbial Biomarkers

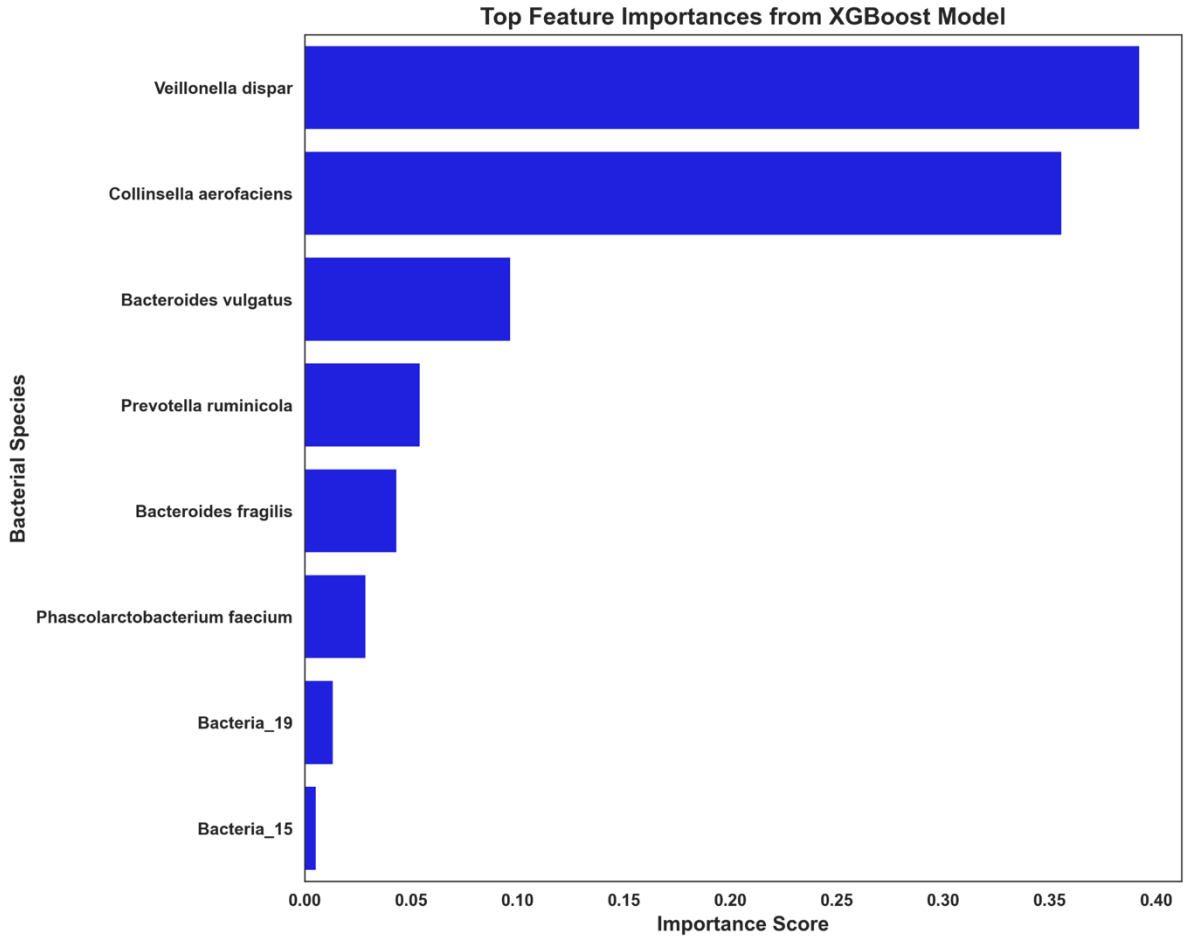
ML
Quantify Disease Intensity + Stage of cancer

Identify ‘digital’ microbial Biomarkers

ML – 0.85 AUC

Detection of AJCC Stage I CRC





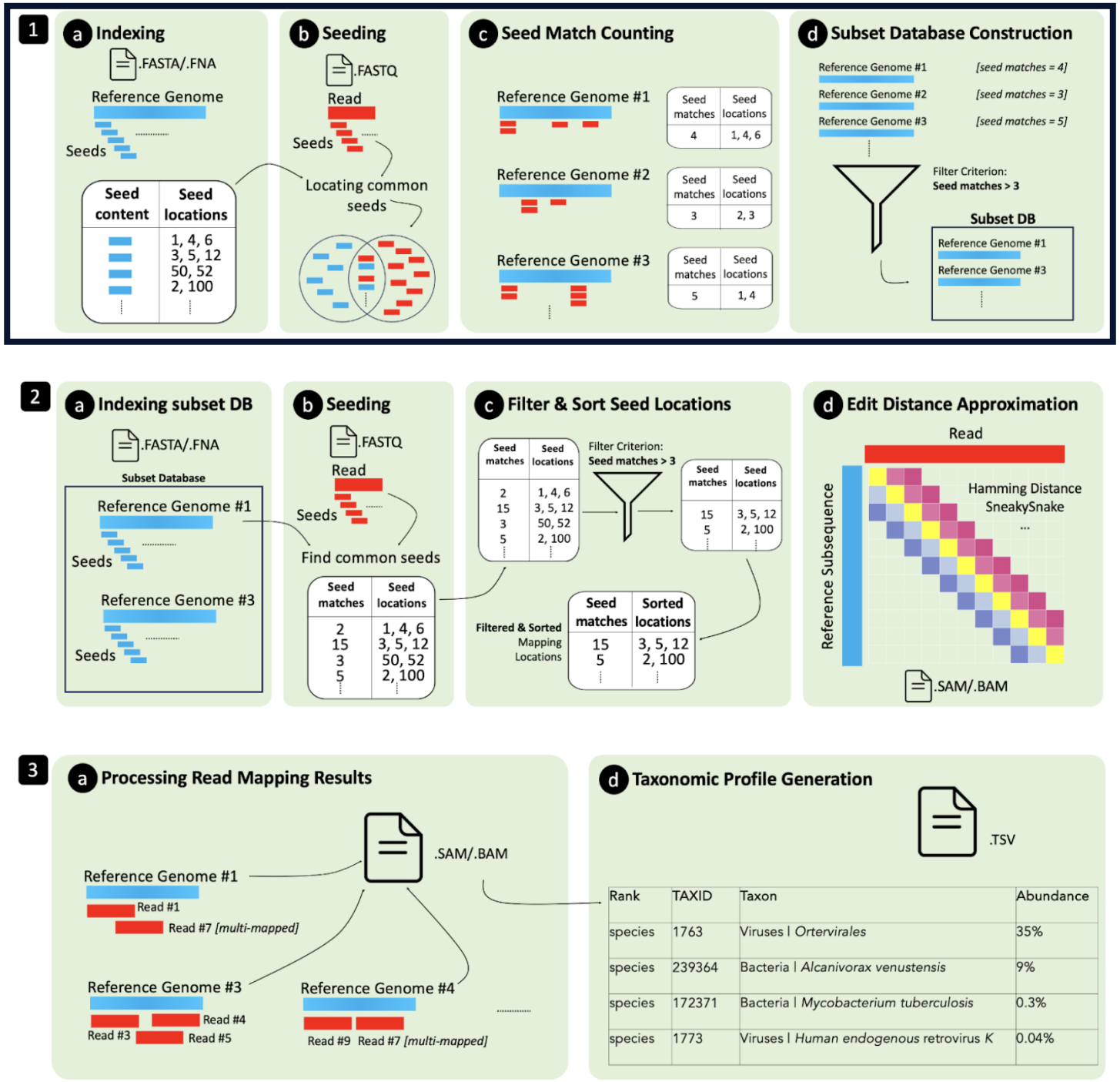
Benchmarking Results

Comprehensive Comparison against the most important state-of-the-art tools

arvidg@mit.edu

Containment Search

Sparsifying Genomic Sequences Enables Large-scale Containment Search



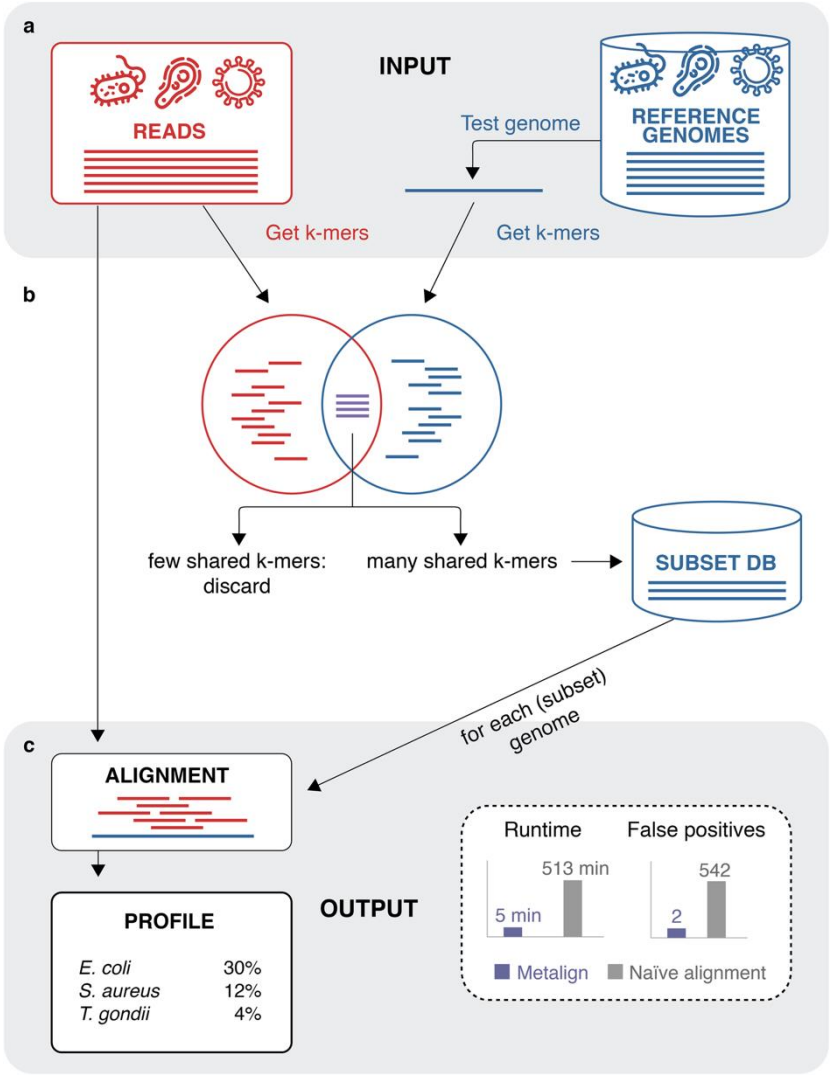
Containment Indexing with KMC3 and CMash

KMC3
used to enumerate k-mers in reads, creating a k-mer database

CMash
calculates the containment index using MinHash, selecting relevant reference genomes

containment index
determines the intersection of k-mers between reads and reference genomes

Indexing
is memory intensive, but MetaTrinity optimizes indexing to reduce runtime by 5.5x compared to KMC3+CMash



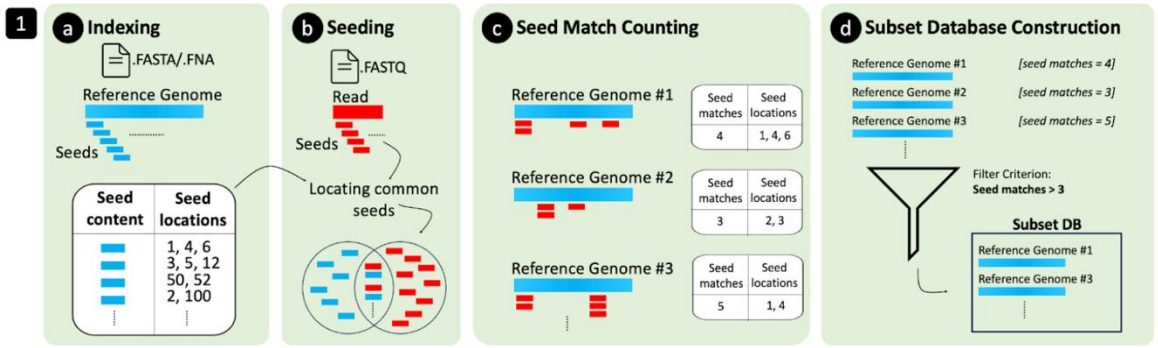
Stage 1: Containment Search

Filters reference database
to create a smaller subset DB

Uses memory-frugal seeding
approach to estimate the similarity between reads and reference genomes

Reduces unnecessary computations
by filtering highly dissimilar references

4x speedup
over Metalign's reference database filtering procedure which uses KMC3 and CMash for reference genome indexing and containment estimation



Benchmarking: Containment Search

Containment Search Overview

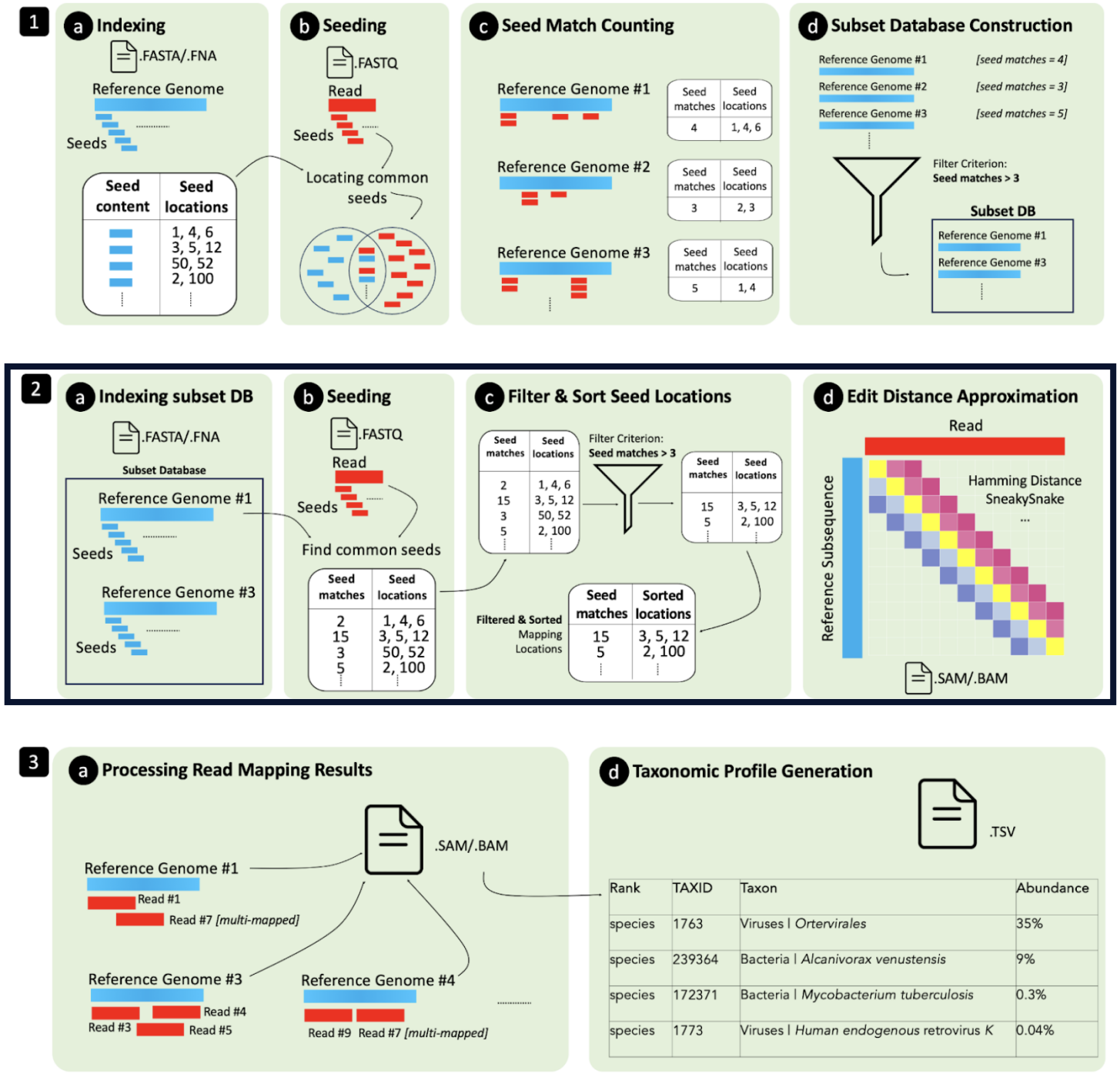
Measures similarity between two genomic datasets by calculating k-mer intersections.

Sparsified Genomes vs KMC3+CMash:

- Speedup: 72.7-75.88x faster for large-scale containment search compared to KMC3+CMash.
- Storage Efficiency: Uses significantly less storage (723.3x more efficient) by avoiding pre-built large indexes and dynamically generating required data.
- On-the-Fly Indexing: Genome-on-Diet builds the containment index during analysis, removing the need for massive pre-built databases and allowing faster execution.

Read Mapping

Sparsifying Genomic Sequences Significantly Accelerates Read Mapping



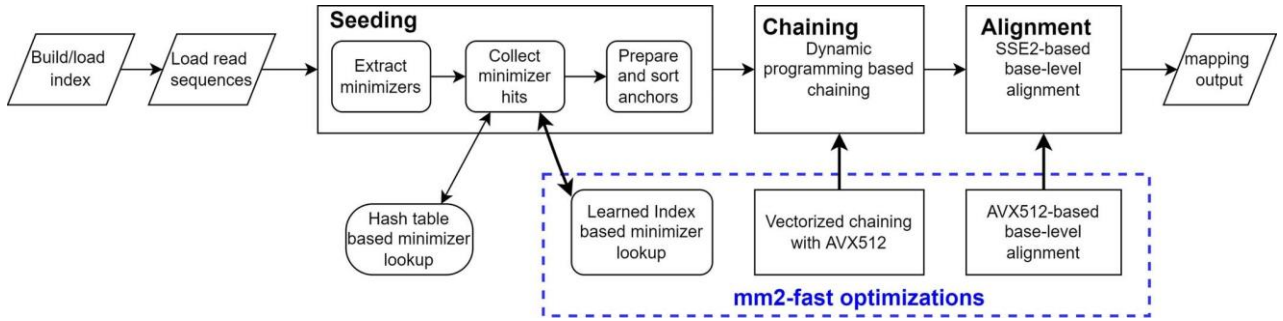
Read Mapping

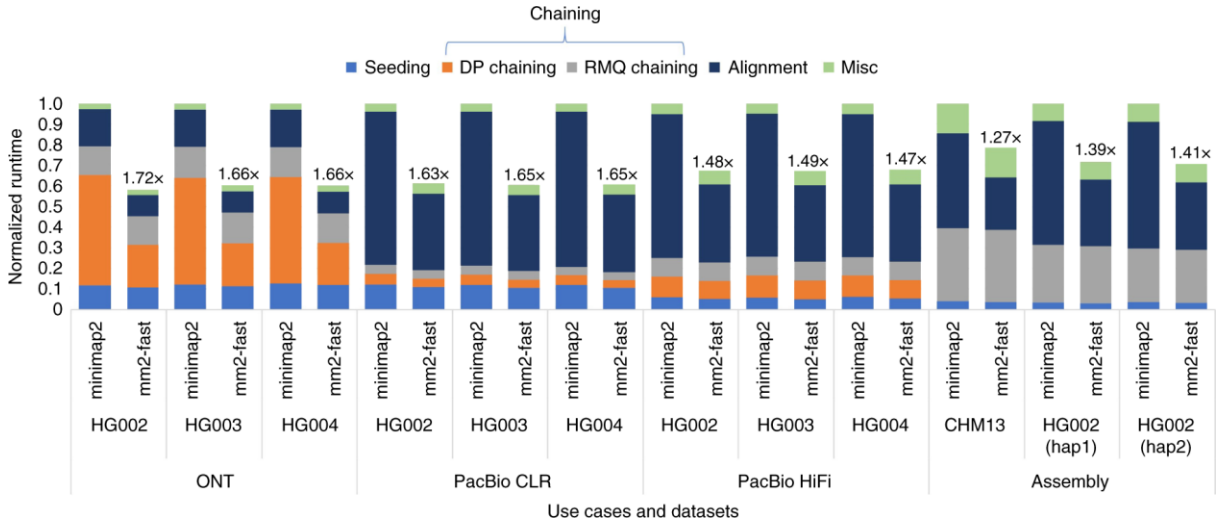
Brief Communication | Published: 28 February 2022

Accelerating minimap2 for long-read sequencing applications on modern CPUs

Saurabh Kalikar, Chirag Jain, Md Vasimuddin & Sanchit Misra

Nature Computational Science **2**, 78–83 (2022) | Cite this article





Methodology: Heuristic Read Mapping

Filters
Reads based on the minimum number of seed hits

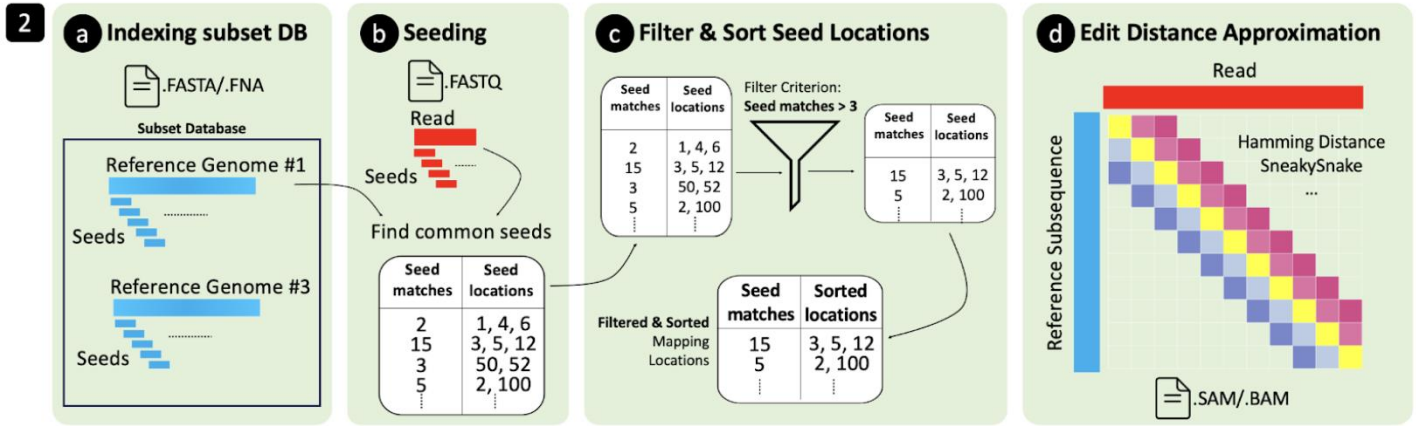
Heuristic Algorithms
To approximate edit distance.

Again Filters
reads with high edit distances.

Heuristic Algorithms
SneakySnake, Hamming Distance, etc.. to achieve the best accuracy-runtime tradeoff

Fourfold Runtime Reduction
Compared to minimap2

Edit distance approximation
Methods to avoid full sequence alignment



Benchmarking: Read Mapping

Compared to Minimap2

- We accelerate all steps of read mapping: indexing, seeding, voting, and alignment.
- **Indexing**: 1.79-1.85x faster indexing due to efficient base exclusion and optimized pattern usage.
- **Seeding**: 1.72-2.69x faster seeding due to reduced input size and parallelized computation.
- **Location Voting**: 65.57-651.19x faster location voting, which reduces computational overhead by eliminating unnecessary seed matches.

Overall Speedup 2.57-6.28x across Illumina, HiFi, and ONT reads

Memory Usage:

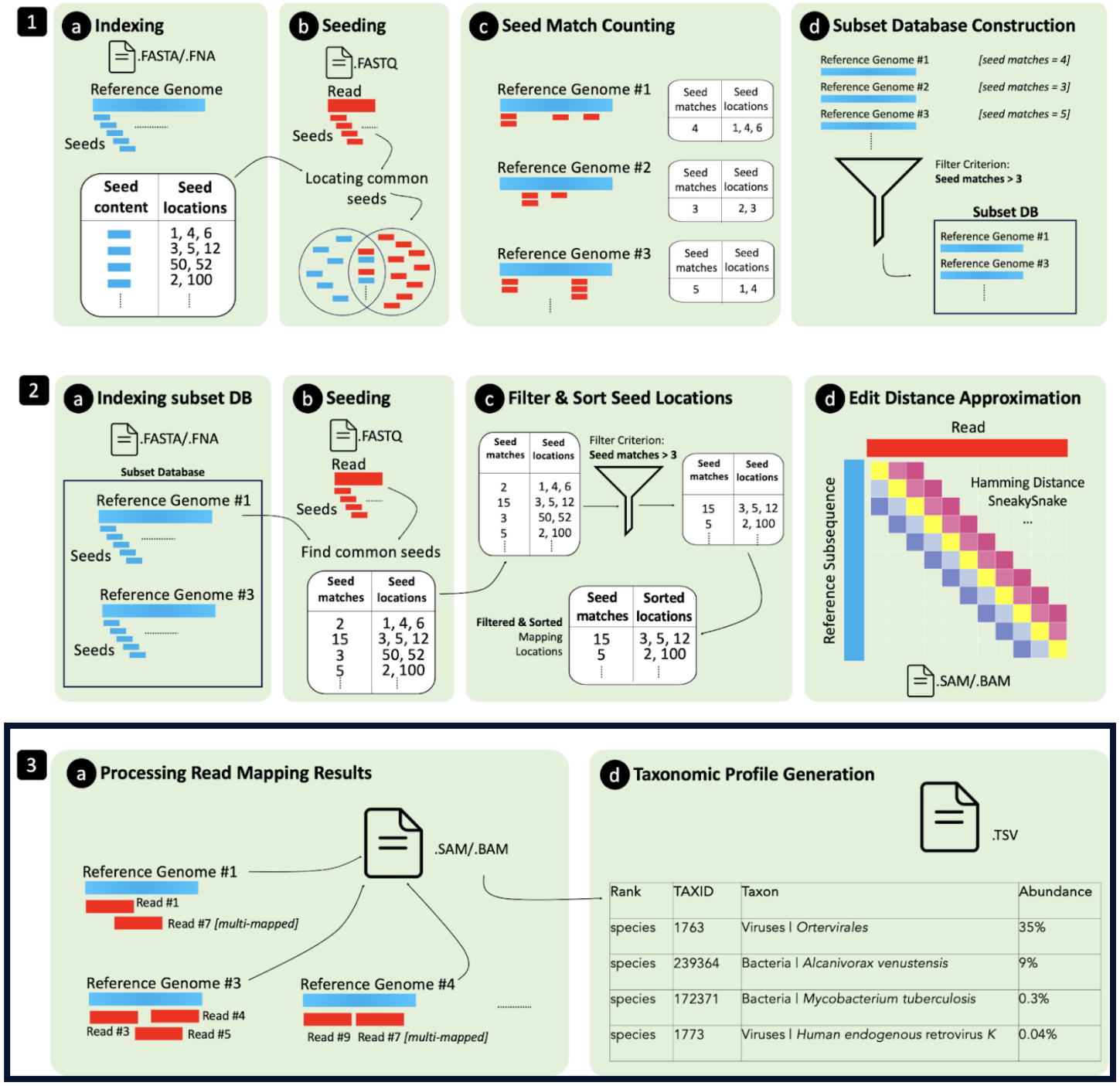
- Reduction of 1.6-2.1x for Illumina and HiFi reads compared to minimap2.
- Higher Peak Memory for ONT Reads: Due to the need to align ultra-long reads, peak memory can be higher, but the segmentation strategy reduces excessive memory use.

Impact of Patterns

- If using *fixed patterns*, '10' provides the best trade-off between speed and accuracy.
- **AI-controlled Metagenomics: Optimal pattern for each read!**

Taxonomic Profiling

Sparsifying Genomic Sequences Allows Robust Taxonomic Profiling



Taxonomic Profiling

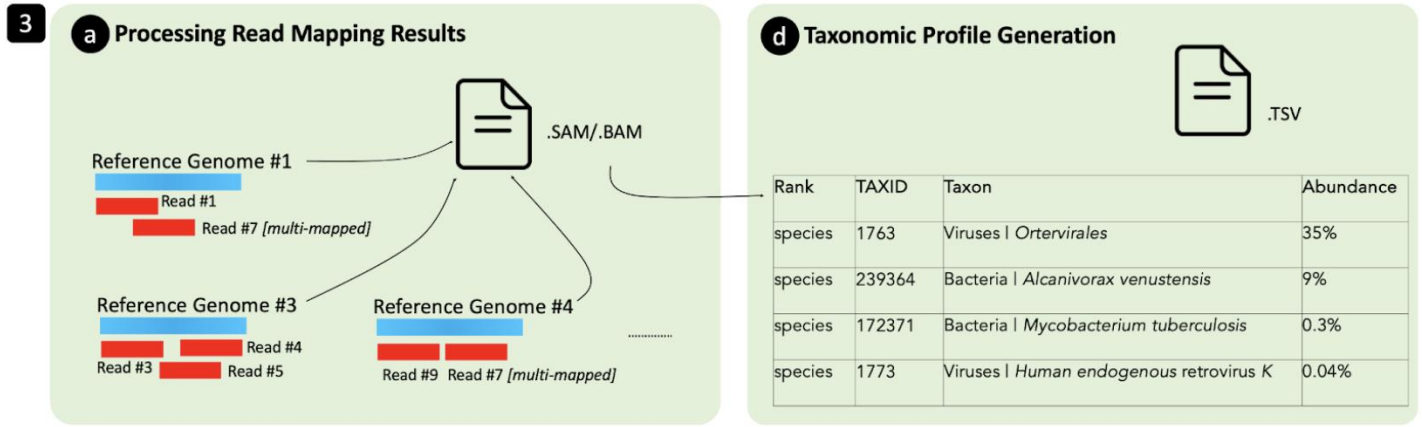
Determines presence
and relative abundance of taxa in a sample

False Positives
Much reduced by excluding organisms with low relative abundances

Profiling results
In CAMI format

Results Include
L1 norm error for relative abundance accuracy

Streamlined approach
Streaming in SAM file



Results: Taxonomic Profiling

Identifying microbes in metagenomic samples by comparing the genomic composition of the sample to known databases.

Efficient Candidate Selection

Narrows down the list of candidate organisms more efficiently compared to KMC3+CMash

Improvement Over Metalign

Genome-on-Diet improves the performance of Metalign by speeding up the identification of candidate genomes and reducing the need for subsequent steps.

Accuracy Metrics

Higher precision in identifying the presence and relative abundance of microbes, maintaining a high true accept rate and minimizing false positives.

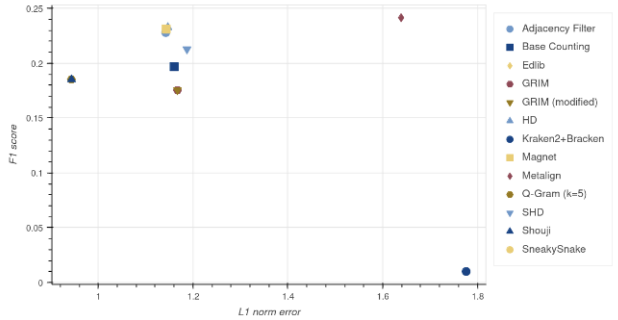
Accuracy Evaluation

Simulated Data
CAMI Challenge

A

	False negatives	False positives	Completeness	Purity	F1 score	L1 norm error
Kraken2+Bracken	210	6274	0.14	0.01	0.01	1.78
GRIM (modified)	91	1337	0.63	0.1	0.18	1.17
GRIM	91	1337	0.63	0.1	0.18	1.17
Shouji	91	1246	0.63	0.11	0.19	0.94
Q-Gram (k=5)	91	1246	0.63	0.11	0.19	0.94
Base Counting	91	1148	0.63	0.12	0.2	1.16
SHD	93	1016	0.62	0.13	0.21	1.19
SneakySnake	93	905	0.62	0.14	0.23	1.14
Magnet	93	905	0.62	0.14	0.23	1.14
HD	93	896	0.62	0.14	0.23	1.15
Edlib	93	896	0.62	0.14	0.23	1.15
Adjacency Filter	93	923	0.62	0.14	0.23	1.14
Metalign	92	856	0.62	0.15	0.24	1.64

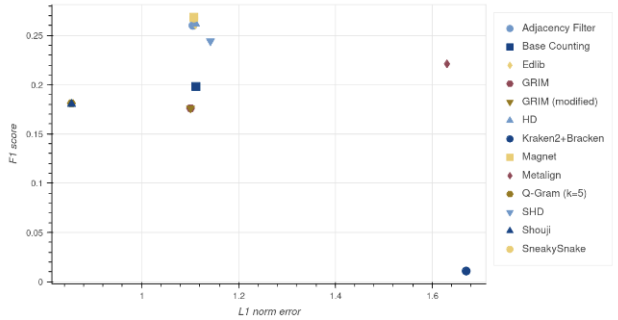
B



C

	False negatives	False positives	Completeness	Purity	F1 score	L1 norm error
Kraken2+Bracken	46	4601	0.36	0.01	0.01	1.67
Shouji	30	352	0.58	0.11	0.18	0.85
Q-Gram (k=5)	30	350	0.58	0.11	0.18	0.85
GRIM (modified)	30	363	0.58	0.1	0.18	1.1
GRIM	30	363	0.58	0.1	0.18	1.1
Base Counting	30	310	0.58	0.12	0.2	1.11
Metalign	30	266	0.58	0.14	0.22	1.63
SHD	30	230	0.58	0.15	0.24	1.14
HD	30	207	0.58	0.17	0.26	1.11
Edlib	30	208	0.58	0.17	0.26	1.11
Adjacency Filter	30	209	0.58	0.17	0.26	1.11
SneakySnake	30	200	0.58	0.17	0.27	1.11
Magnet	30	199	0.58	0.17	0.27	1.11

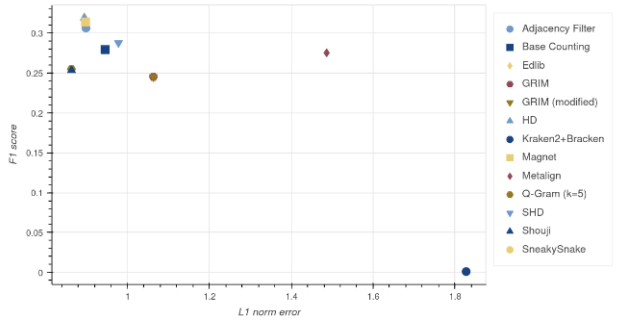
D



E

	False negatives	False positives	Completeness	Purity	F1 score	L1 norm error
Kraken2+Bracken	19	6934	0.17	0	0	1.83
Shouji	4	108	0.83	0.15	0.25	0.86
GRIM (modified)	4	113	0.83	0.14	0.25	1.06
GRIM	4	113	0.83	0.14	0.25	1.06
Q-Gram (k=5)	4	107	0.83	0.15	0.26	0.86
Metalign	4	96	0.83	0.17	0.28	1.49
Base Counting	4	94	0.83	0.17	0.28	0.95
SHD	4	90	0.83	0.17	0.29	0.98
SneakySnake	4	79	0.83	0.19	0.31	0.9
Magnet	4	79	0.83	0.19	0.31	0.9
Adjacency Filter	4	82	0.83	0.19	0.31	0.9
HD	4	77	0.83	0.2	0.32	0.89
Edlib	4	77	0.83	0.2	0.32	0.89

F



Accuracy Evaluation

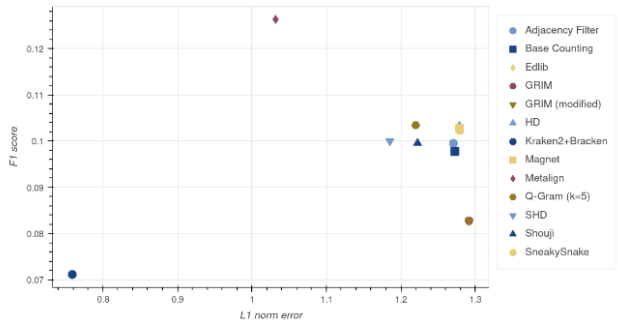
Real Data

Human Gut Samples

A

	False negatives	False positives	Completeness	Purity	F1 score	L1 norm error
Kraken2+Bracken	1	547	0.95	0.04	0.07	0.76
GRIM (modified)	10	256	0.55	0.04	0.08	1.29
GRIM	10	256	0.55	0.04	0.08	1.29
SneakySnake	11	182	0.5	0.06	0.1	1.28
Shouji	10	207	0.55	0.05	0.1	1.22
SHD	11	187	0.5	0.06	0.1	1.19
Q-Gram (k=5)	10	198	0.55	0.06	0.1	1.22
Magnet	11	181	0.5	0.06	0.1	1.28
HD	11	180	0.5	0.06	0.1	1.28
Edlib	11	181	0.5	0.06	0.1	1.28
Base Counting	11	192	0.5	0.05	0.1	1.27
Adjacency Filter	11	188	0.5	0.06	0.1	1.27
Metalign	10	156	0.55	0.07	0.13	1.03

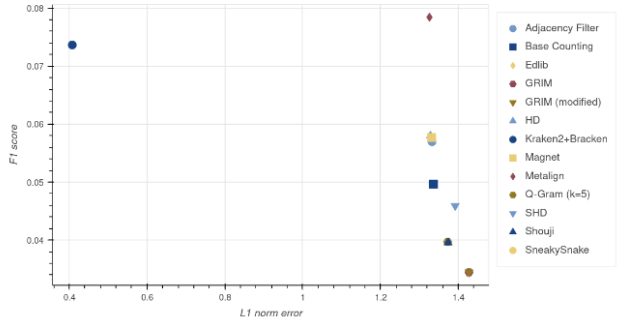
B



C

	False negatives	False positives	Completeness	Purity	F1 score	L1 norm error
GRIM (modified)	1	1008	0.95	0.02	0.03	1.43
GRIM	1	1008	0.95	0.02	0.03	1.43
Shouji	1	873	0.95	0.02	0.04	1.37
Q-Gram (k=5)	1	870	0.95	0.02	0.04	1.37
SHD	1	747	0.95	0.02	0.05	1.39
Base Counting	1	688	0.95	0.03	0.05	1.34
SneakySnake	1	587	0.95	0.03	0.06	1.33
Magnet	1	587	0.95	0.03	0.06	1.33
HD	1	584	0.95	0.03	0.06	1.33
Edlib	1	586	0.95	0.03	0.06	1.33
Adjacency Filter	1	595	0.95	0.03	0.06	1.33
Kraken2+Bracken	0	478	1	0.04	0.07	0.41
Metalign	3	373	0.84	0.04	0.08	1.33

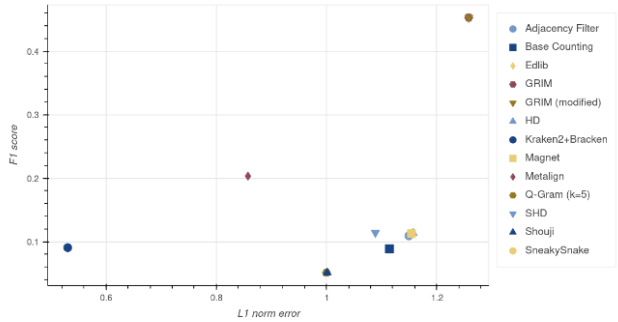
D



E

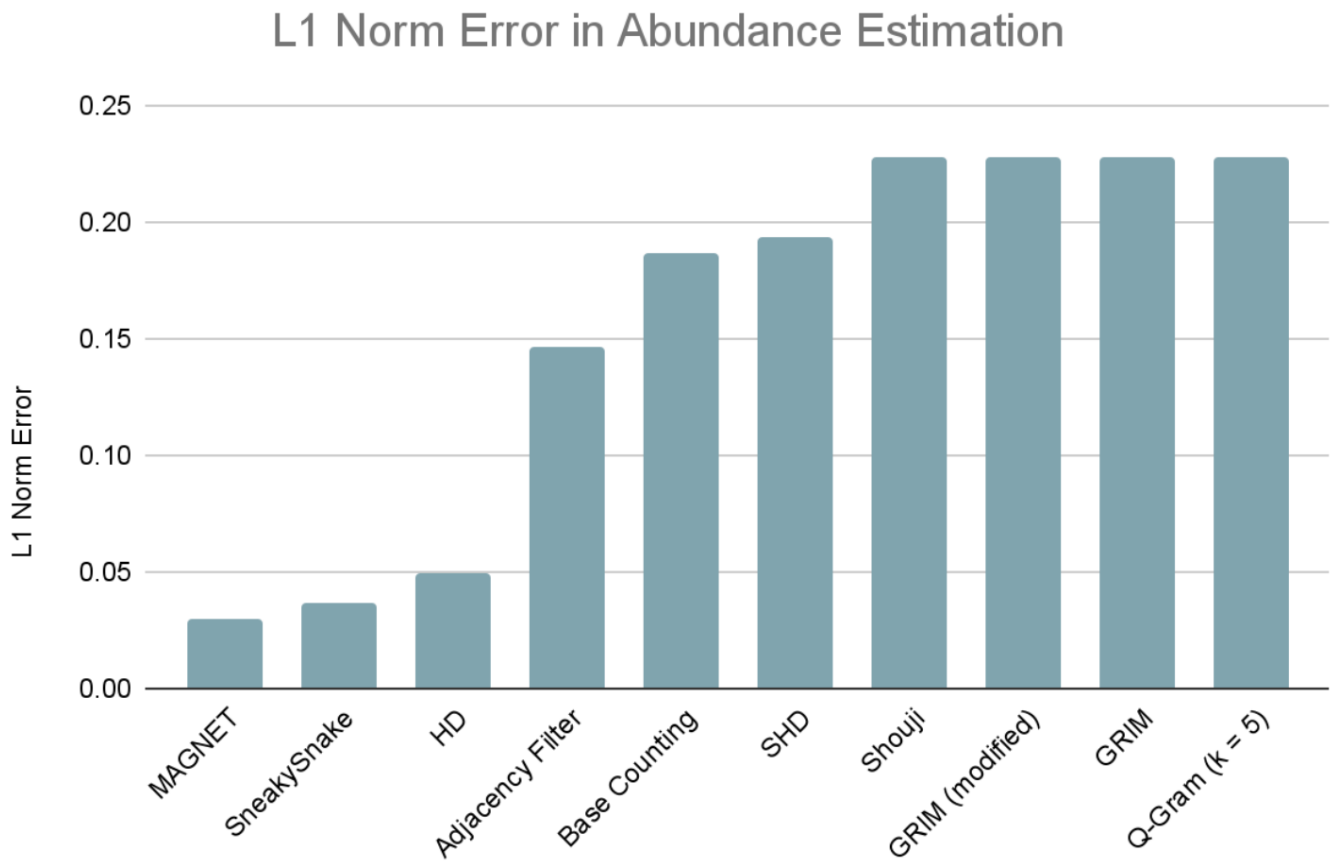
	False negatives	False positives	Completeness	Purity	F1 score	L1 norm error
Shouji	0	704	1	0.03	0.05	1
Q-Gram (k=5)	0	703	1	0.03	0.05	1
Kraken2+Bracken	0	381	1	0.05	0.09	0.53
Base Counting	0	390	1	0.05	0.09	1.11
SneakySnake	0	299	1	0.06	0.11	1.15
SHD	0	295	1	0.06	0.11	1.09
Magnet	0	299	1	0.06	0.11	1.15
HD	0	297	1	0.06	0.11	1.16
Edlib	0	297	1	0.06	0.11	1.16
Adjacency Filter	0	310	1	0.06	0.11	1.15
Metalign	0	149	1	0.11	0.2	0.86
GRIM (modified)	7	22	0.63	0.35	0.45	1.26
GRIM	7	22	0.63	0.35	0.45	1.26

F



Accuracy Evaluation

Simulated Data
CAMI Challenge

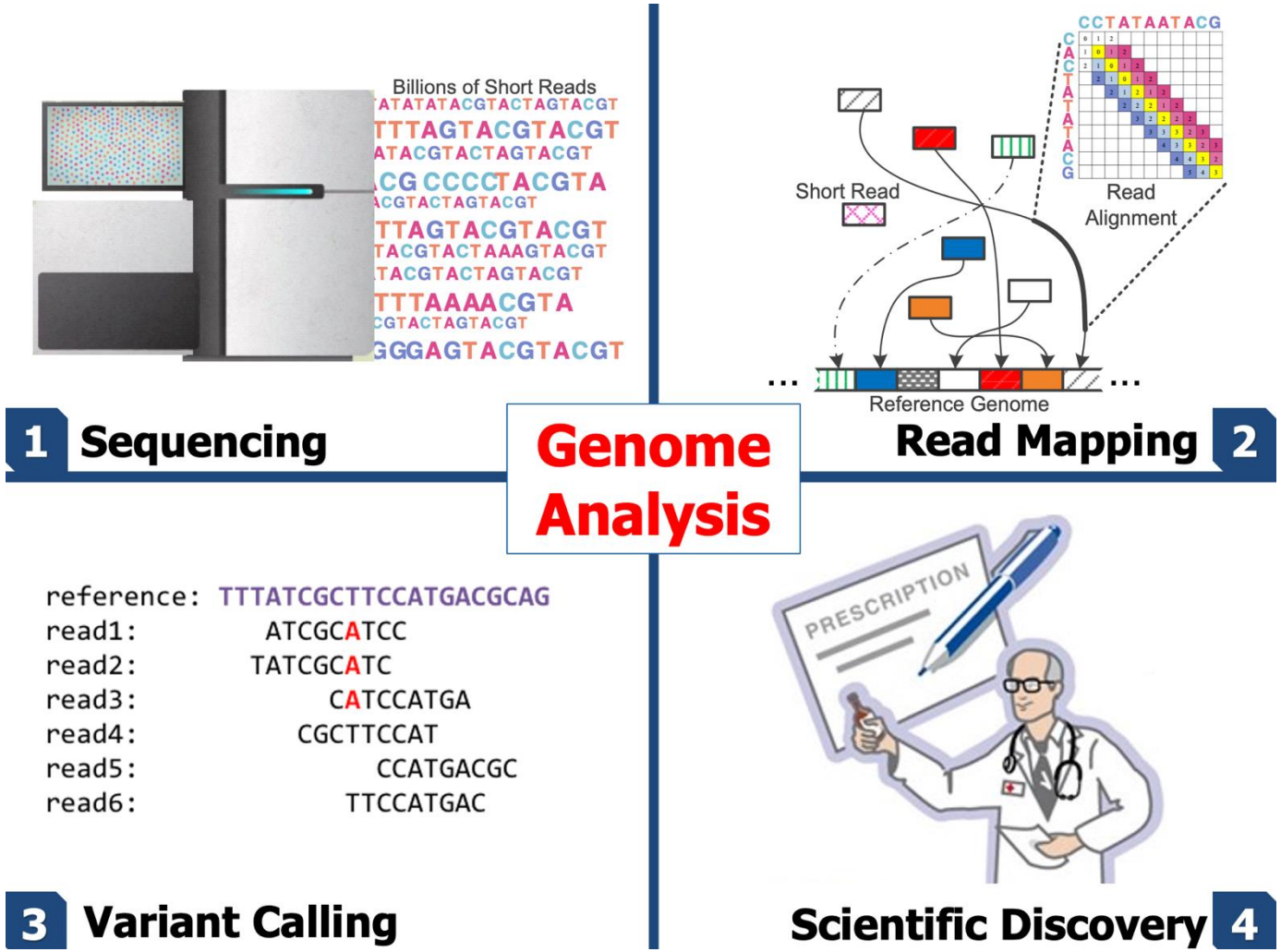


Related Literature

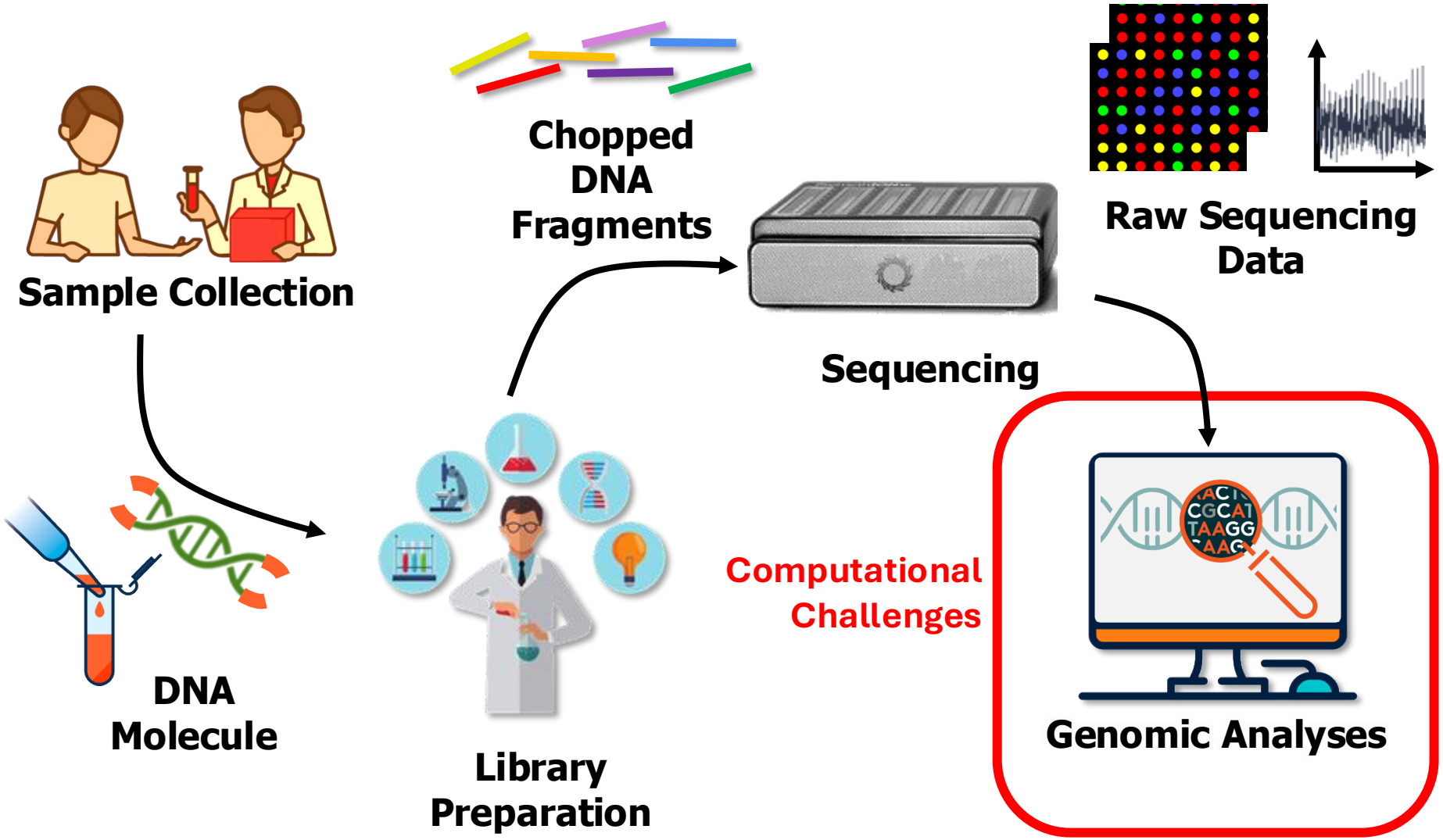
Related literature and extended benchmarking and performance evaluation results

arvidg@mit.edu

Processing Genomic Data



Our Focus: Better Algorithms

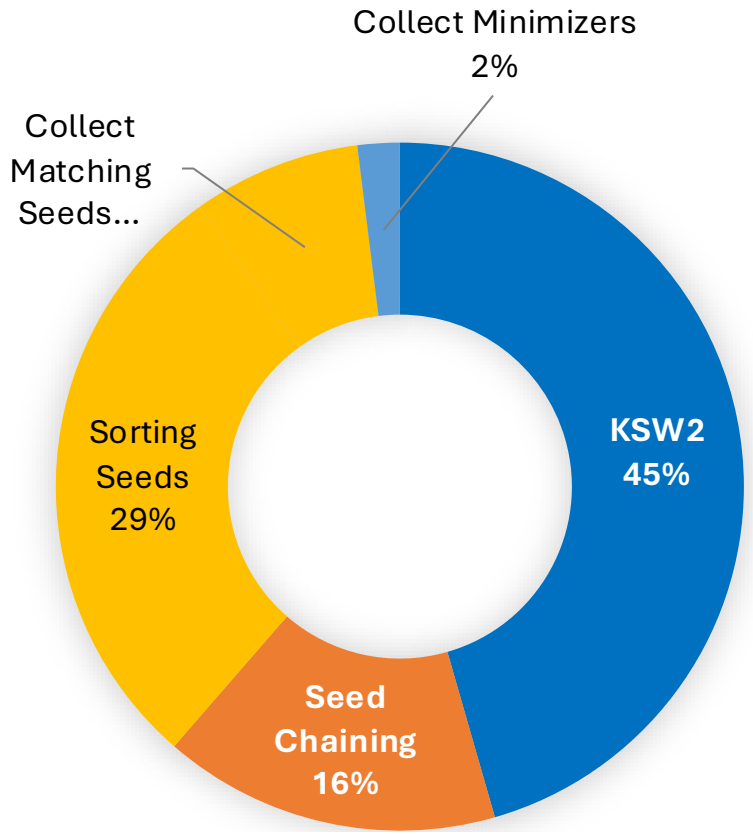


Execution Time Breakdown

>60%

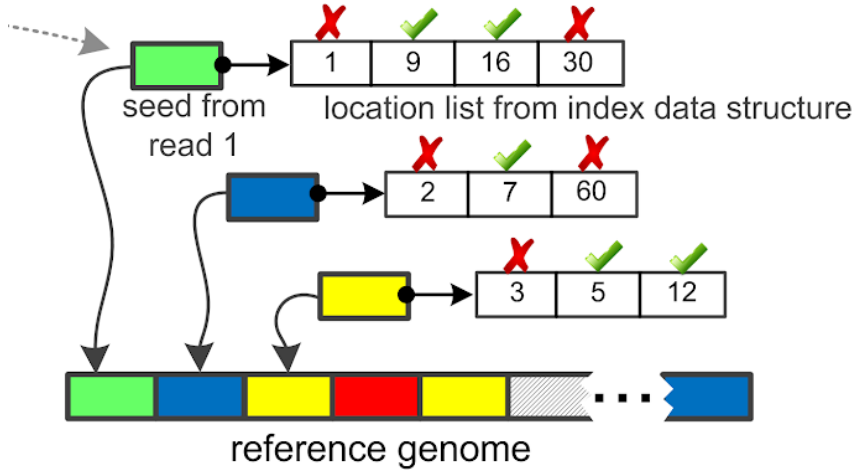
of the read mapper's execution time is spent in sequence alignment

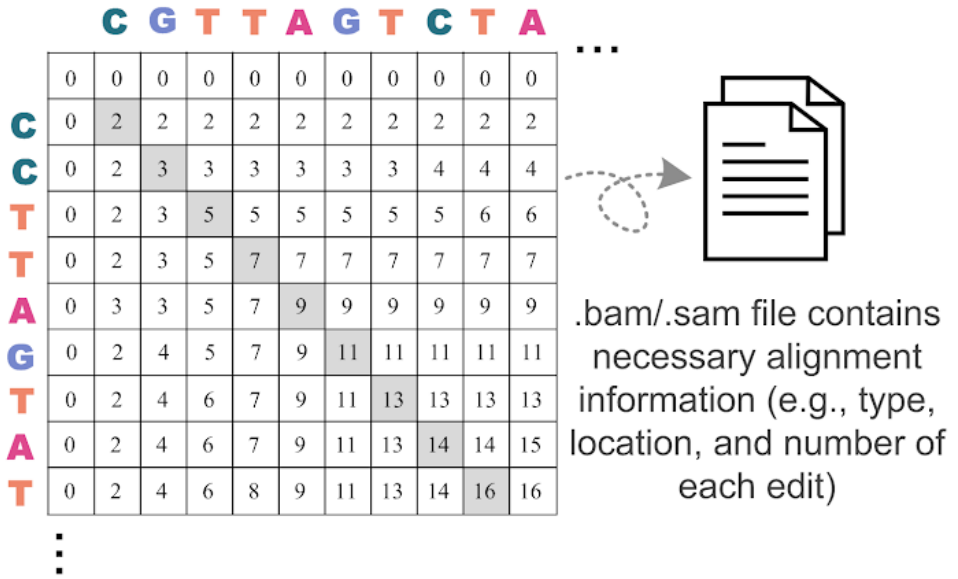
minimap2



ONT FASTQ size: 103MB (151 reads), Mean length: 356,403 bp, std: 173,168 bp, longest length: 817,917 bp

Two Main Paradigms





k-mer based
Fast, but *not accurate*

Alignment based
Accurate, but *slow*

Our Motivation

Metagenomics
Accuracy <-> Speed

Early Approaches
- Alignment-based, increasingly impractical due to the growth of HTS data
- Tools like *PathSeq* computationally infeasible on large datasets
- Alignment-free methods are generally faster

Need for high accuracy
- Distinguish cell-type-specific intracellular microbes from extracellular and contaminating microbes
- Small number of microbial reads per cell inadequate for alignment-free methods (use *PathSeq)*

Medical Applications
Alignment-free approaches

Oncology
- Tumors sequenced by The Cancer Genome Atlas Program (lots of new data available!)
- Used to build a classifier for cancer type, using alignment-free approach Kraken
- *Poore, G.D., Kopylova, E., Zhu, Q. et al.* ***Microbiome analyses of blood and tissues suggest cancer diagnostic approach****. Nature 579, 567–574 (2020).*
  https://doi.org/10.1038/s41586-020-2095-1

Tradeoff
- Alignment vs. alignment-free
- Trend: Alignment free & large data sets

Our Motivation

Metagenomics
Alignment-free

Existing Tools
- Typically rely on string or k-mer matching to obtain a taxonomic assignment for each read
- *E.g., Kraken & Centrifuge*
- Assign reads to the lowest taxonomic rank possible or to a pre-determined taxonomic level (i.e., genus, species, strain)

Clinical Application
- Often need to distinguish pathogenic strains from non-pathogenic strains
- *Kraken, Kraken2,* and *Centrifuge* often used in Hospitals (surveyed 50+ Hospitals)

Benchmarking
Comparing Metagenomic Tools

Data Sets
- Critical Assessment of Metagenome Interpretation (CAMI)
- International Microbiome and Multiomics Standards Alliance (IMMSA)

Recent benchmarking study
- Simon, H. Y., Siddle, K. J., Park, D. J. & Sabeti, P. C. **Benchmarking metagenomics tools for taxonomic classification**. Cell 178, 779–794 (2019).
- Covers 20 taxonomic classifiers, alignment-based (*PathSeq, MetaPhlAn2*) and alignment-free (*Kraken, CLARK, KrakenUniq, Centrifuge*)

Two Big Breakthroughs

(1) Sparsified Data

Genome on Diet

(2) Sparsified Data + Computation

AI Controlled Genomics

Sparsified Computation: ML Control

(1) Containment Search: Subset Database Construction

Index Querying

- Input set of sequencing reads and a reference genome database.
- Collect seeds from all reference genomes within the database (4 CPs).

Seeding

- Extract seeds from the input read set for comparison with reference genomes (4 CPs).
- Seed Match Counting:
- Estimate the number of seed matches between each reference genome and the read set using memory-efficient methods (2 CPs).

Subset Database Construction

- Select reference genomes to form a smaller subset database (2 CPs).

Sparsified Computation: ML Control

(2) Heuristic Read Mapping

Querying the Subset Database:

Load seeds from the subset database and extract seeds from the read set (3 CPs).

Candidate Mapping Location Identification:

Identify candidate mapping locations by locating seed matches (1 CP).

Filtering and Sorting Mapping Locations:

- Filtering: Discard candidate mapping locations that do not achieve a minimum seed match threshold (2 CPs).
- Sorting: Sort remaining locations by seed match counts (2 CPs).

Sparsified Computation: ML Control

(2) Heuristic Read Mapping

Edit Distance Approximation:

- Approximate the edit distance using heuristic methods (e.g., SneakySnake, Hamming Distance, SHD) (3 CPs).
- Thresholding: Exclude reads with a minimum edit distance exceeding a certain threshold of the read length (2 CPs).

(3) Taxonomic Profiling and Output Generation

- Include uniquely mapped reads and multi-mapped reads in the analysis.
- Abundance Levels: Quantify abundance based on mapped reads and their respective edit distances (1 CP).

- Taxonomic Profile Generation:
- Discard taxa with relative abundance levels below a threshold (1 CP).

Control Unit Architecture

Supervised Models

- **Decision Trees and Random Forest Classifiers**: To make probabilistic decisions at key stages
- **High-Likelihood Seed Matches**: Identifies likely seed matches and filters improbable mapping locations
- **Taxonomic Specificity**: dynamically refine subset databases to include only relevant reference genomes

Unsupervised Models

- Self-organizing maps (SOMs), Restricted Boltzmann Machines (RBMs), and Deep Stacking Networks (DSNs): Capture complex, latent structures in genomic data.
- SOMs: Understand microbial community structure.
- RBMs and DSNs: Capture hidden patterns in *sparsified* k-mer structures, optimizing heuristic read mapping and edit distance approximation.

Huge Design Space: Control Architecture

AI Control Unit

Fundamentally novel architecture that orchestrates the execution of metagenomic pipelines.

Meta-Nucleus Orchestration Layer

Acts as the command center of the metagenomic pipeline.

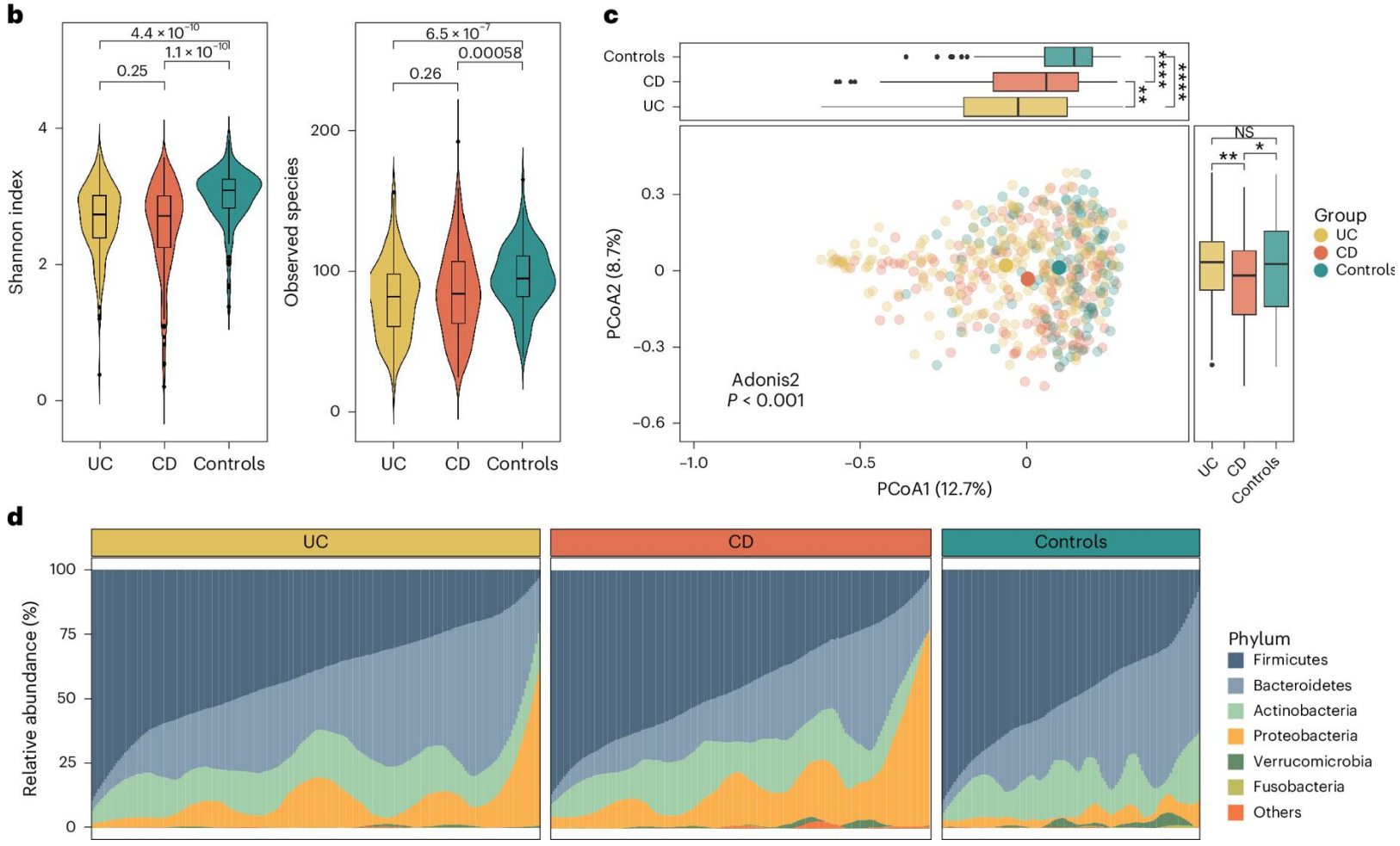
Combine Insights

Supervised and unsupervised learning models to make precise, real-time classification decisions.

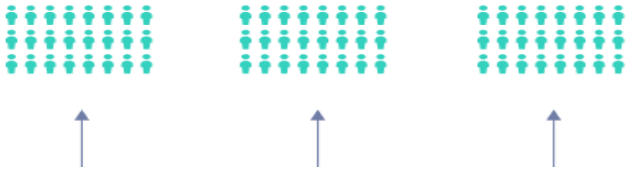
Multi-Level Decision Engine

Leverages contributions from supervised and unsupervised models for complementary perspectives on the metagenomic data as it moves through the pipeline.

Characterization of Microbial Alterations in IBD



Methods



Metagenomics

- Hiseq 2500 platform (**Illumina**)
- **2 × 150 bp paired-end reads** with an average data size of 15 Gb per sample
- **Contaminated** reads from host discarded by mapping against human genome using **Bowtie2**
- 0.1% of the reads from the 314 samples classified as human reads

Longitudinal Study

- **Repeatedly** check for **CRC** development
- Subjects were **followed up** from **recruitment** to the **date of CRC** diagnosis, death or until 2020
- During the entire follow-up period, repeatedly examine the same individuals to detect any changes that might occur over a period
- Clinical Data Analysis and Reporting System, CDARS in **Hong Kong**

Results

Compositional shift of gut microbiome after appendectomy

- Positive **clinical association** between **appendectomy** and **CRC** development

Gut microbiome profiling

- 314 fecal samples from 157 appendectomy cases and 157 normal controls
- Average of **34,168,657 paired reads** per sample
- Microbial composition observed at **phylum-level**
- Relative abundance **cutoff ≥ 1%**

Microbial signature influenced by sampling time

- Examined microbiome in specimens collected 6 months, 6–12 months, 12–18 months, 18–24 months, and 24 months **after appendectomy**
- Suggesting microbial community **changes** caused by appendectomy could **persist > 2 years**

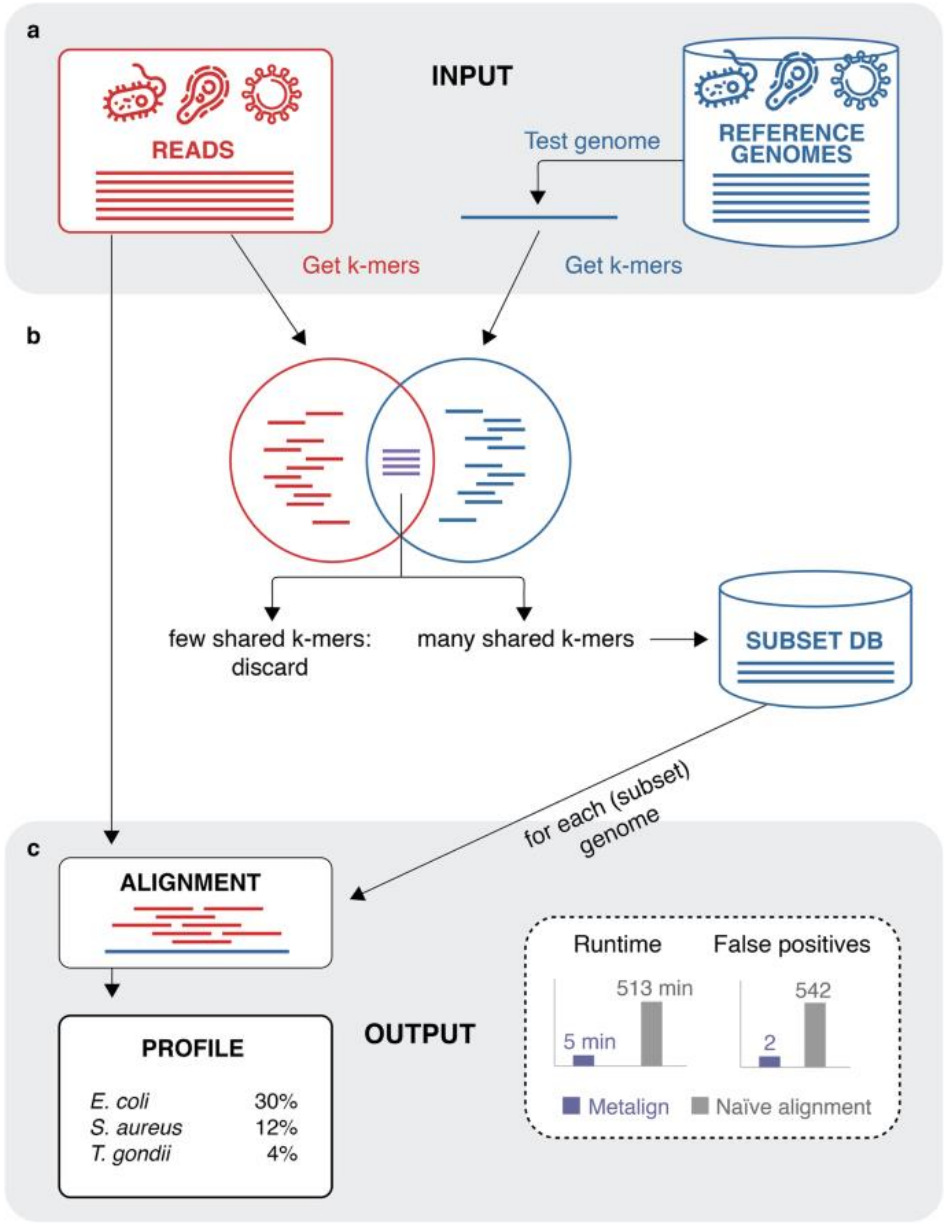
Metalign

LaPierre, N., Alser, M., Eskin, E. et al.

Metalign: efficient alignment-based metagenomic profiling via containment min hash.

Genome Biol 21, 242 (2020).

https://doi.org/10.1186/s13059-020-02159-0



Benchmarking Study

Simon, H. Y., Siddle, K. J., Park, D. J. & Sabeti, P. C.

Benchmarking metagenomics tools for taxonomic classification.

Cell 178, 779–794 (2019).

Table 1. A List of Benchmarked Classifiers and Their Various Characteristics

Type	Classifier	Custom Databases	Generates Abundance Profile	Memory Required	Time Required	Reference
DNA	Bracken	yes	yes	<1 Gb	<1 min	Lu et al., 2017
	Centrifuge	yes	yes	20 Gb	7 min	Kim et al., 2016
	CLARK	yes	yes	80 Gb	2 min	Ounit et al., 2015
	CLARK-S	yes	yes	170 Gb	40 min	Ounit and Lonardi, 2016
	Kraken	yes	yes	190 Gb	1 min	Wood and Salzberg, 2014
	Kraken2	yes	yes	36 Gb	1 min	Wood and Salzberg, 2014
	KrakenUniq	yes	yes	200 Gb	1 min	Breitwieser et al., 2018
	k-SLAM	yes	yes	130 Gb	2 h	Ainsworth et al., 2017
	MegaBLAST	yes	no	61 Gb	4 h	Morgulis et al., 2008
	metaOthello	no	no	30 Gb	1 min	Liu et al., 2018
	PathSeq	yes[a]	no	140 Gb	5 min	Walker et al., 2018
	prophyle	yes	no	40 Gb	40 min	Břinda et al., 2017
	taxMaps	yes	yes	65 Gb	25 min	Corvelo et al., 2018
Protein	DIAMOND	yes	no	110 Gb (varies)	10 min	Buchfink et al., 2015
	Kaiju	yes	yes	25 Gb	1 min	Menzel et al., 2016
	MMseqs2	yes	no	85 Gb (varies)	9 h	Steinegger and Söding, 2017
Markers	MetaPhlAn2	no	yes	2 Gb	1 min	Truong et al., 2015
	mOTUs2	no	yes	2 Gb	1 min	Milanese et al., 2019

"Custom databases" refers to the ability for the end user to create a custom database. The time and memory requirements are for a 5.7 million-read dataset with the database and input already cached in memory. Some methods (marked as "varies") have the ability to flexibly decrease their memory usage (at the cost of a massive increase in run time).

[a]The latest version of PathSeq now allows the user to create and specify a custom database, but this option was not available when benchmarking studies were performed; thus, it was excluded from those analyses.

Benchmarking Study

Simon, H. Y., Siddle, K. J., Park, D. J. & Sabeti, P. C.

Benchmarking metagenomics tools for taxonomic classification.

Cell 178, 779–794 (2019).

Table 1
A List of Benchmarked Classifiers and Their Various Characteristics

Type	Classifier	Custom Databases	Generates Abundance Profile	Memory Required	Time Required	Reference
DNA	Bracken	yes	yes	<1 Gb	<1 min	Lu et al., 2017
	Centrifuge	yes	yes	20 Gb	7 min	Kim et al., 2016
	CLARK	yes	yes	80 Gb	2 min	Ounit et al., 2015
	CLARK-S	yes	yes	170 Gb	40 min	Ounit and Lonardi, 2016
	Kraken	yes	yes	190 Gb	1 min	Wood and Salzberg, 2014
	Kraken2	yes	yes	36 Gb	1 min	Wood and Salzberg, 2014
	KrakenUniq	yes	yes	200 Gb	1 min	Breitwieser et al., 2018
	k-SLAM	yes	yes	130 Gb	2 h	Ainsworth et al., 2017
	MegaBLAST	yes	no	61 Gb	4 h	Morgulis et al., 2008
	metaOthello	no	no	30 Gb	1 min	Liu et al., 2018
	PathSeq	yes[a]	no	140 Gb	5 min	Walker et al., 2018
	prophyle	yes	no	40 Gb	40 min	Břinda et al., 2017
	taxMaps	yes	yes	65 Gb	25 min	Corvelo et al., 2018
Protein	DIAMOND	yes	no	110 Gb (varies)	10 min	Buchfink et al., 2015
	Kaiju	yes	yes	25 Gb	1 min	Menzel et al., 2016
	MMseqs2	yes	no	85 Gb (varies)	9 h	Steinegger and Söding, 2017
Markers	MetaPhlAn2	no	yes	2 Gb	1 min	Truong et al., 2015
	mOTUs2	no	yes	2 Gb	1 min	Milanese et al., 2019

Benchmarking Study

Simon, H. Y., Siddle, K. J., Park, D. J. & Sabeti, P. C.

Benchmarking metagenomics tools for taxonomic classification.

Cell 178, 779–794 (2019).

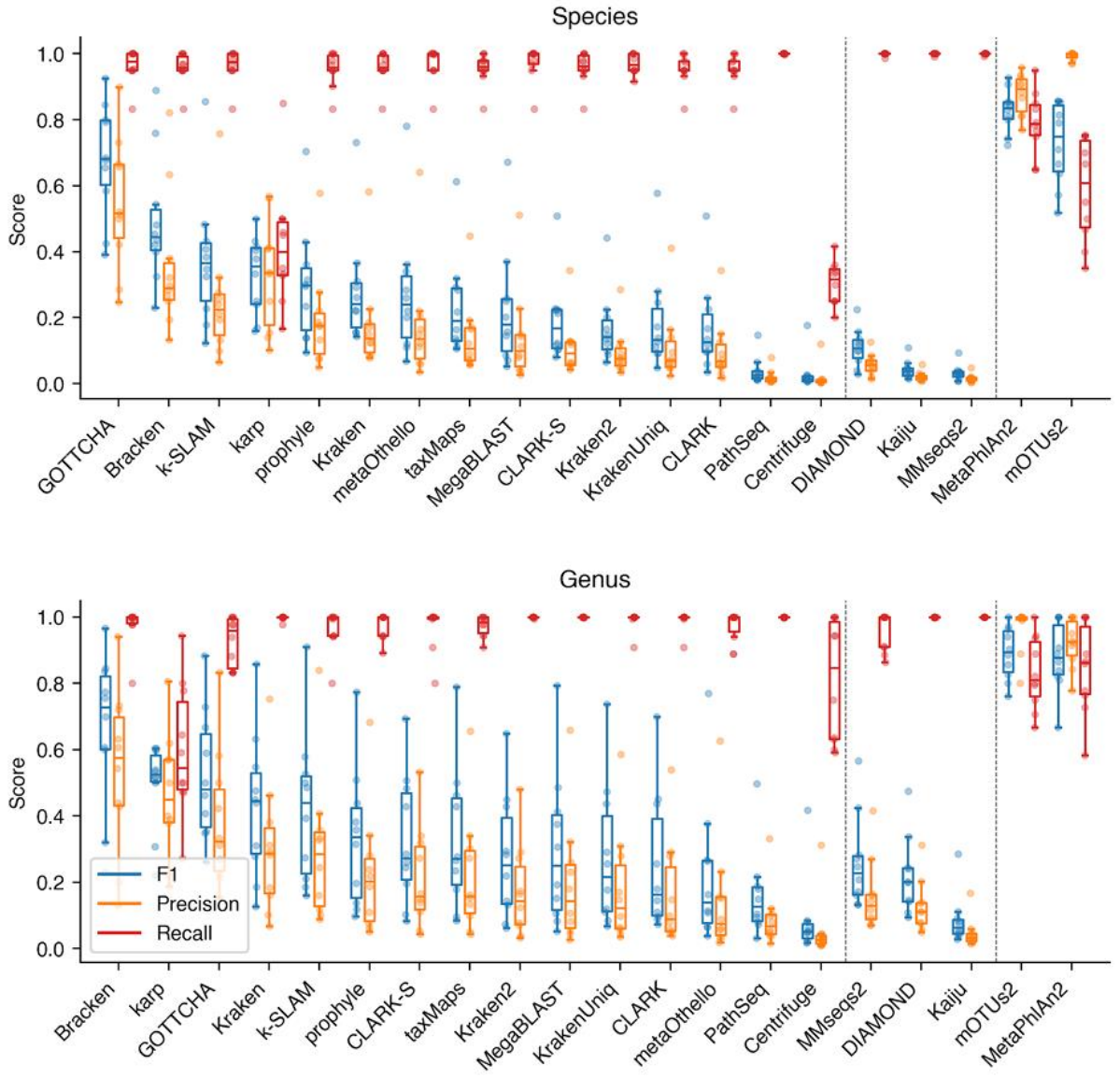


Figure S1. Baseline precision, recall, and F1 statistics on unfiltered abundance reports with no abundance threshold (considering all classified taxa regardless of abundance) at the species and genus levels using default databases. Related to Figure 3.

Benchmarking Study

Simon, H. Y., Siddle, K. J., Park, D. J. & Sabeti, P. C.

Benchmarking metagenomics tools for taxonomic classification.

Cell 178, 779–794 (2019).

Figure 7
Benchmark of Computational Resources