Can Big (Microbiome) Data Save the World? – Using AI and Microbiome Data to Predict Health Outcomes

Sirota Lab

Marina Sirota, PhD

Associate Professor

UCSF Bakar Computational Health Sciences Institute

BURROUGHS WELLCOME FUND

MARCH OF DIMES

Why Now?

1000 GENOMES FROM 14 ETHNIC GROUPS

GEO Gene Expression Omnibus

ENCODE ENCYCLOPEDIA OF DNA ELEMENTS

ExAC

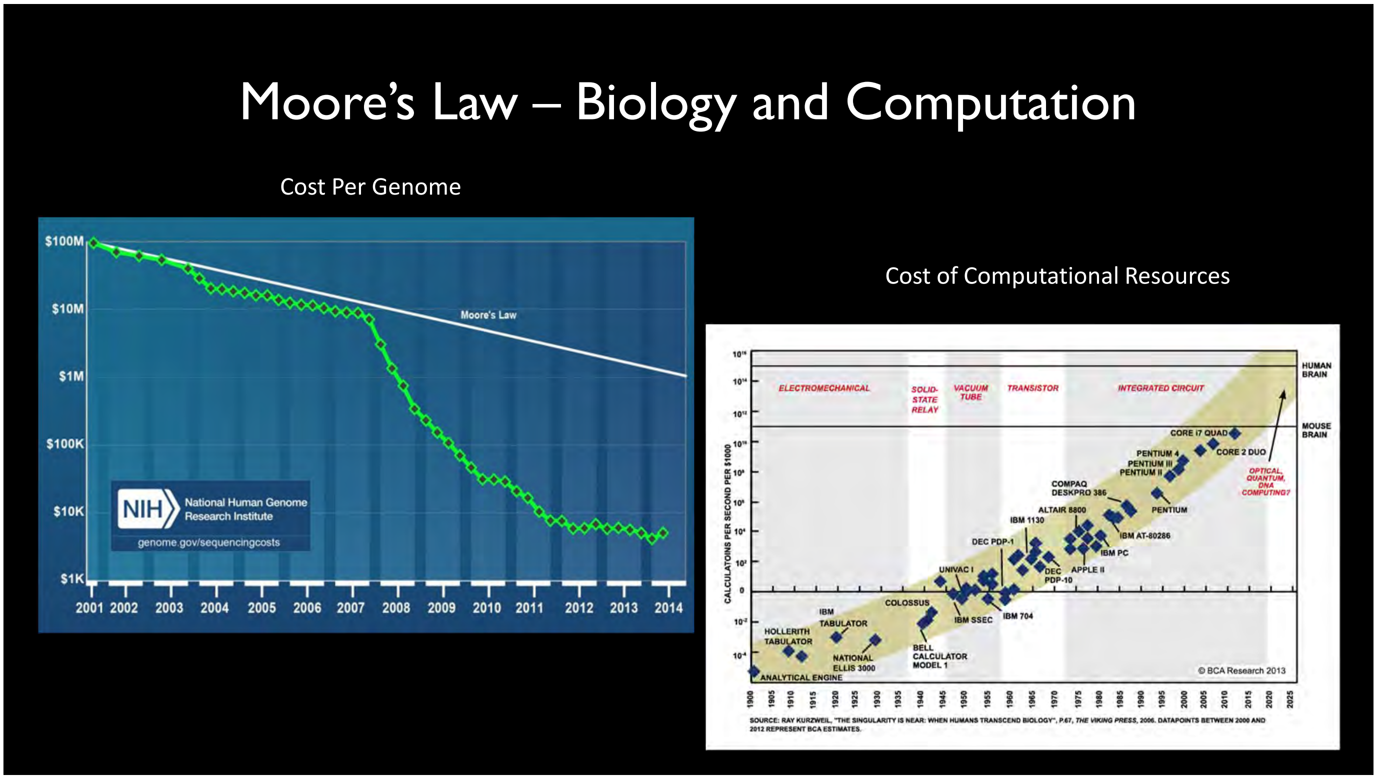
NIH LINCS PROGRAM

The Cancer Genome Atlas

Understanding genomics to improve cancer care

IMMPORT

BIOINFORMATICS FOR THE FUTURE OF IMMUNOLOGY

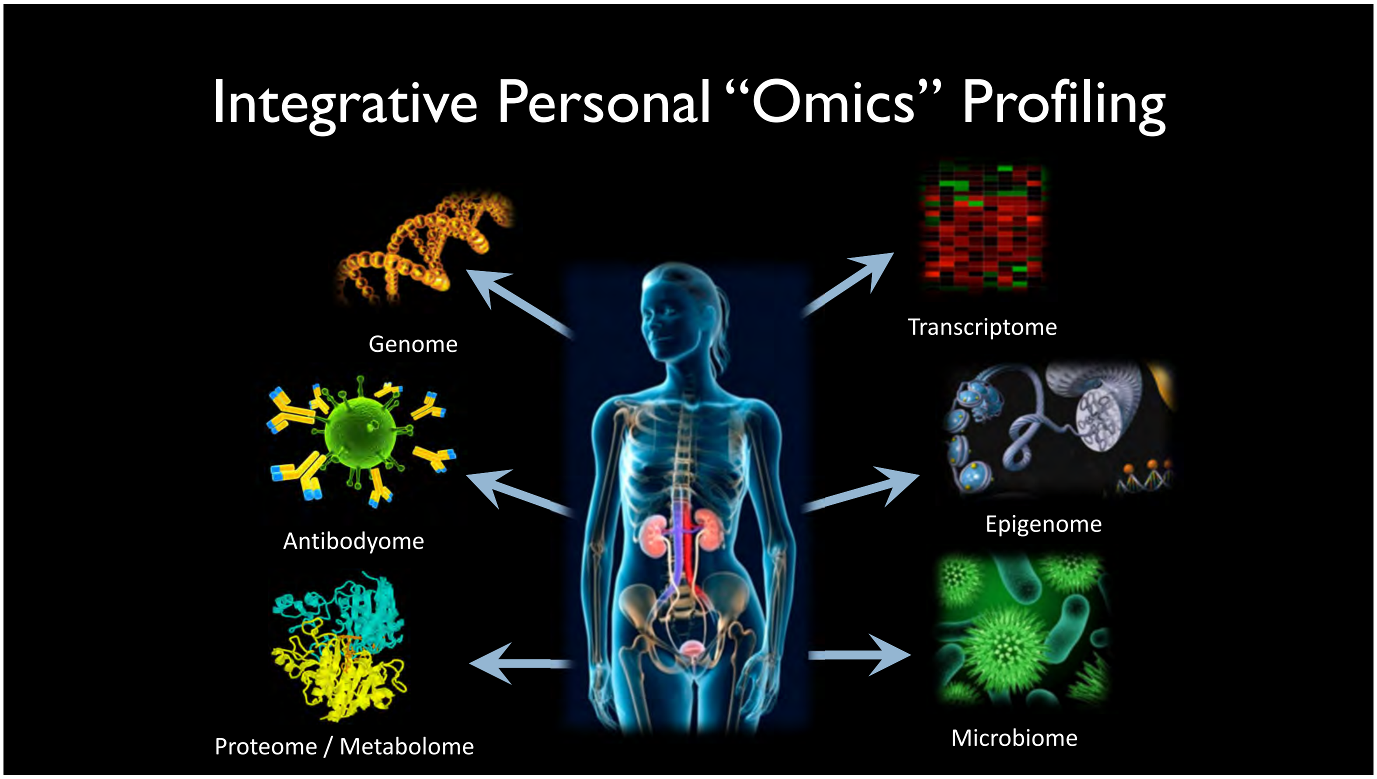


Moore's Law – Biology and Computation

Cost Per Genome

Cost of Computational Resources

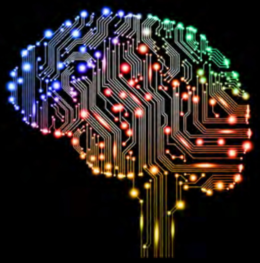
SOURCE: RAY KURZWEIL, "THE SINGULARITY IS NEAR: WHEN HUMANS TRANSCEND BIOLOGY", P.67, *THE VIKING PRESS*, 2006. DATAPOINTS BETWEEN 2000 AND 2012 REPRESENT BCA ESTIMATES.



Integrative Personal "Omics" Profiling

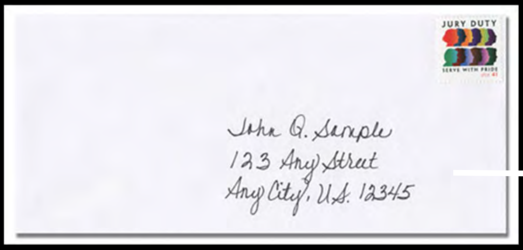
Artificial Intelligence

- A branch of computer science dealing with the simulation of intelligent behavior in computers – making predictions and decisions
- The capability of a machine to imitate intelligent human behavior
- Examples:
  - Visual perception
  - Speech recognition
  - Translation between languages
  - Decision-making



Machine Learning

- Computers are good at following instructions
- <u>Example</u> - run a spell checker on a document
- But what if the task is really complicated to describe?



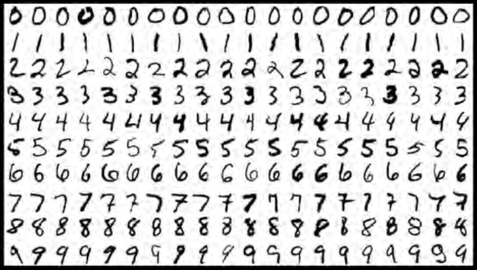
John Q. Sample
123 Any Street
Any City, US, 12345

Eugene Davydov, PhD

Example: Digit Recognition

- Many tasks aren’t easy to describe at the level of individual instructions
- We (humans) can recognize handwritten digits because we’ve seen many examples in our lifetimes

How can we write computer programs that can learn and get better from data?



Eugene Davydov, PhD

Machine Learning is Transforming Our Lives

- Machine learning (ML) is a type of artificial intelligence (AI) that provides computers with the ability to learn without being explicitly programmed
- Focuses on the development of computer programs that can change when exposed to new data

Recommender Systems

NETFLIX

amazon

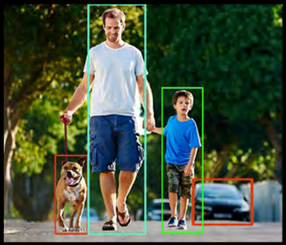


Image Processing

Biomedical Applications

Eugene Davydov, PhD


Searching...
13:25
Google Maps

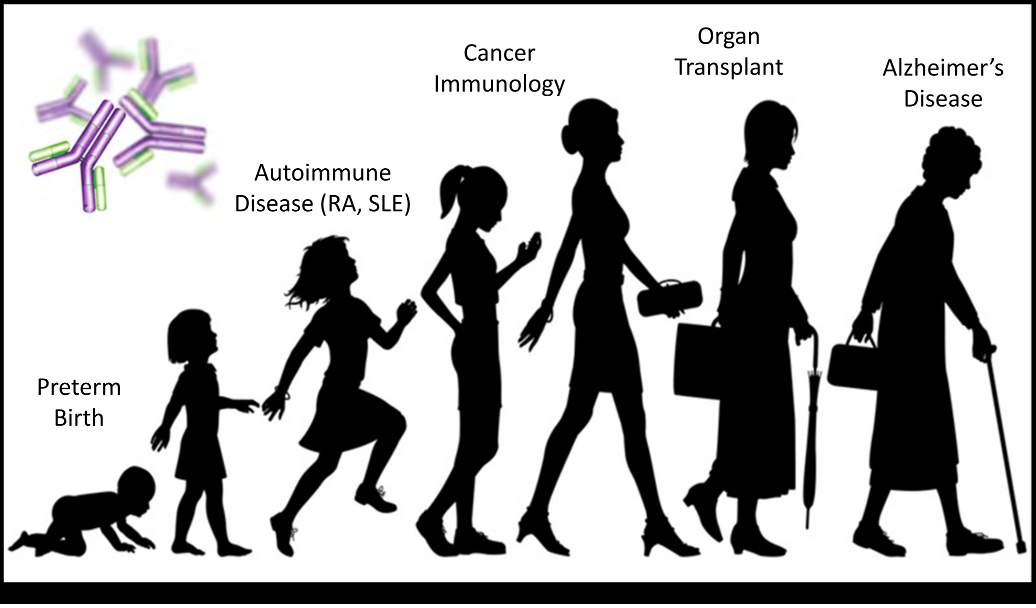
9


SAMSUNG
10:10
Yelp
best things nearby
best reviewers
Palmetto Bay Sunrise Cafe
Vinnie Van Go-Go's
Sakura Sushi
Rudi's Steak House
keywords
search
nearby
review
photos
check-ins

10

We have been so successful in transforming many everyday industries, but how can these data-driven approaches be applied in biomedicine?

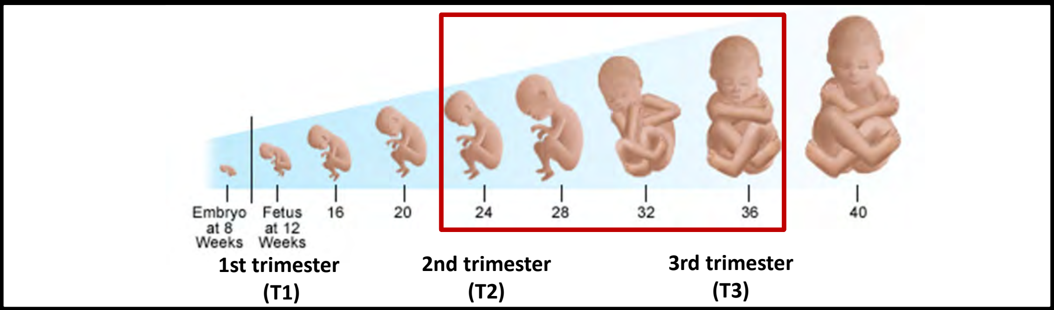
Leveraging Computation to Understand Health and Disease Across the Lifespan…



Pregnancy and Preterm Birth

Why Study Preterm Birth?

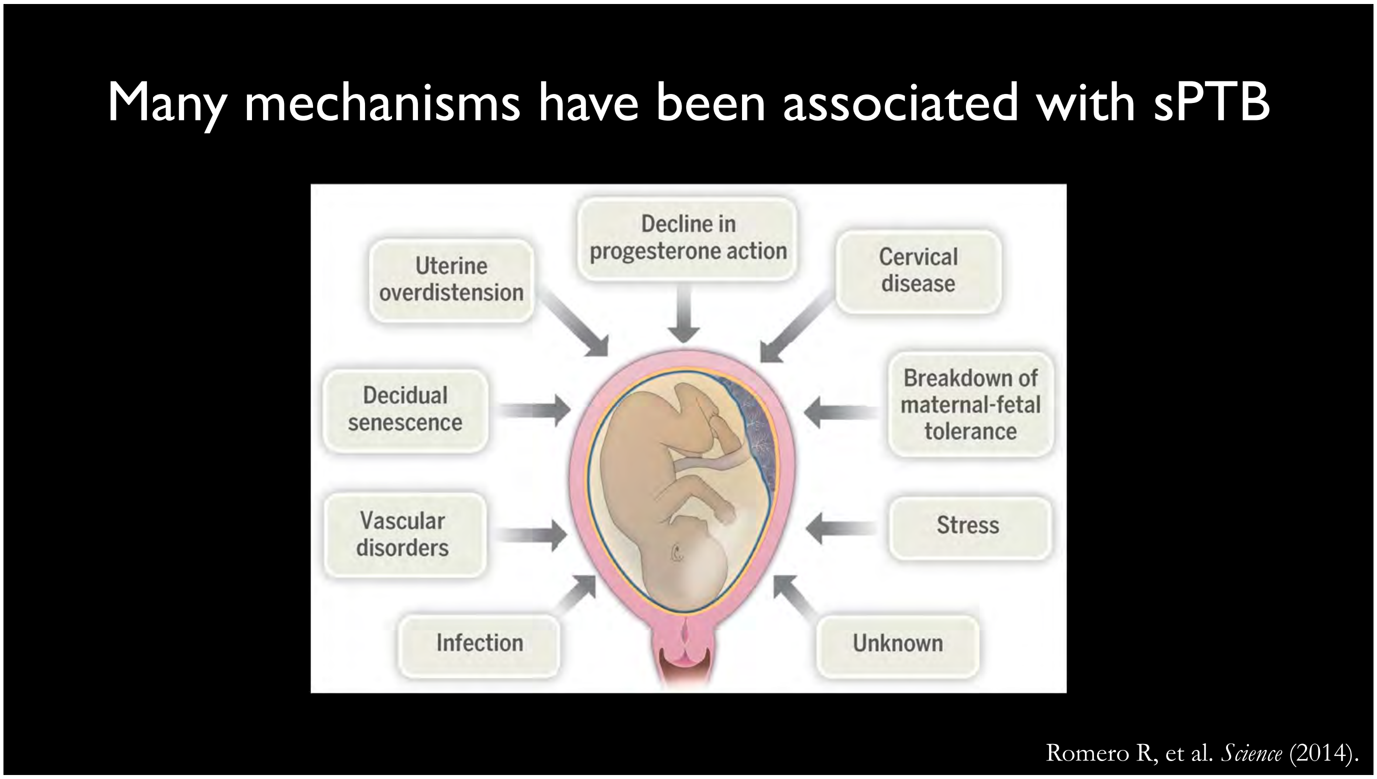
Preterm birth (PTB) - live birth before week 37 of gestation



Worldwide, 15 million babies are born premature each year.

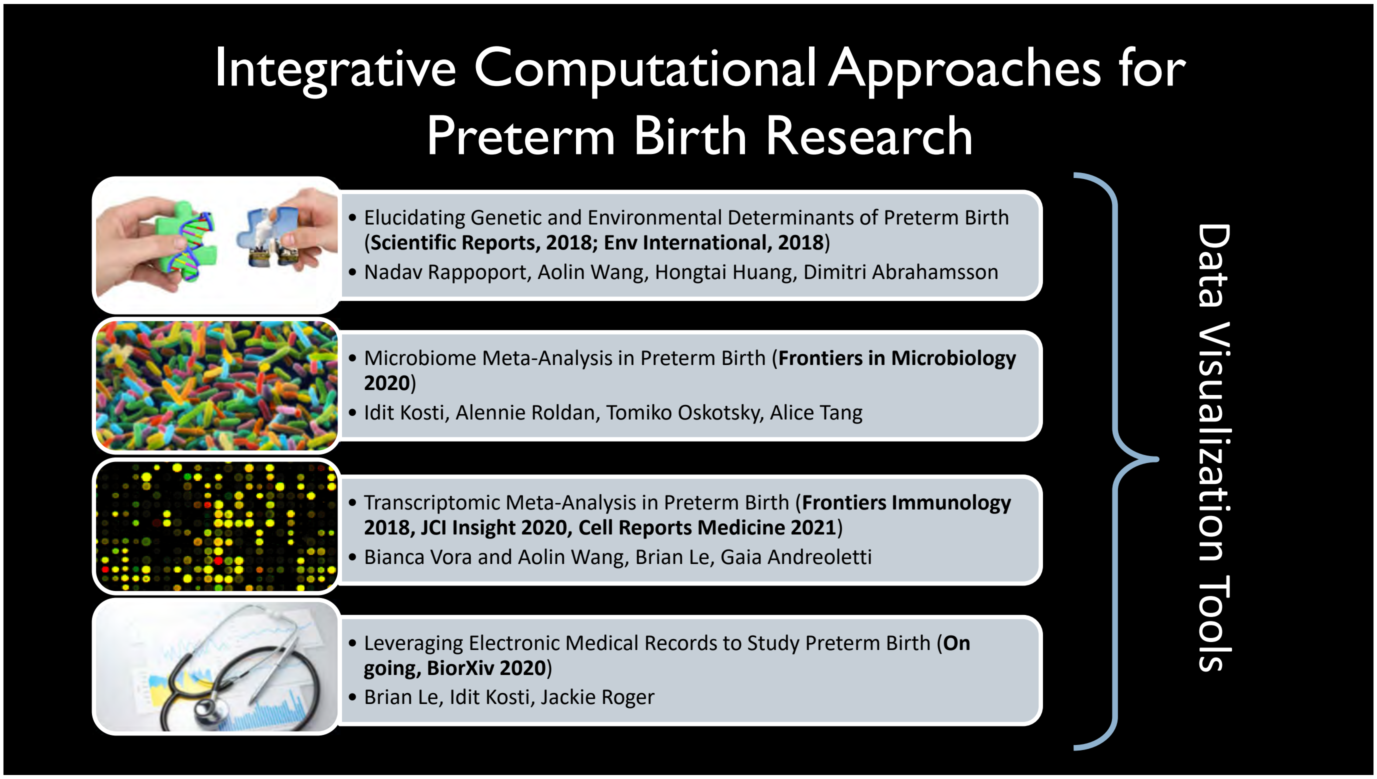
One million of these infants die within the first 28 days of life.

In nearly half of the cases of premature birth, there is no clear cause.



Many mechanisms have been associated with sPTB

Romero R, et al. *Science* (2014).



Integrative Computational Approaches for Preterm Birth Research

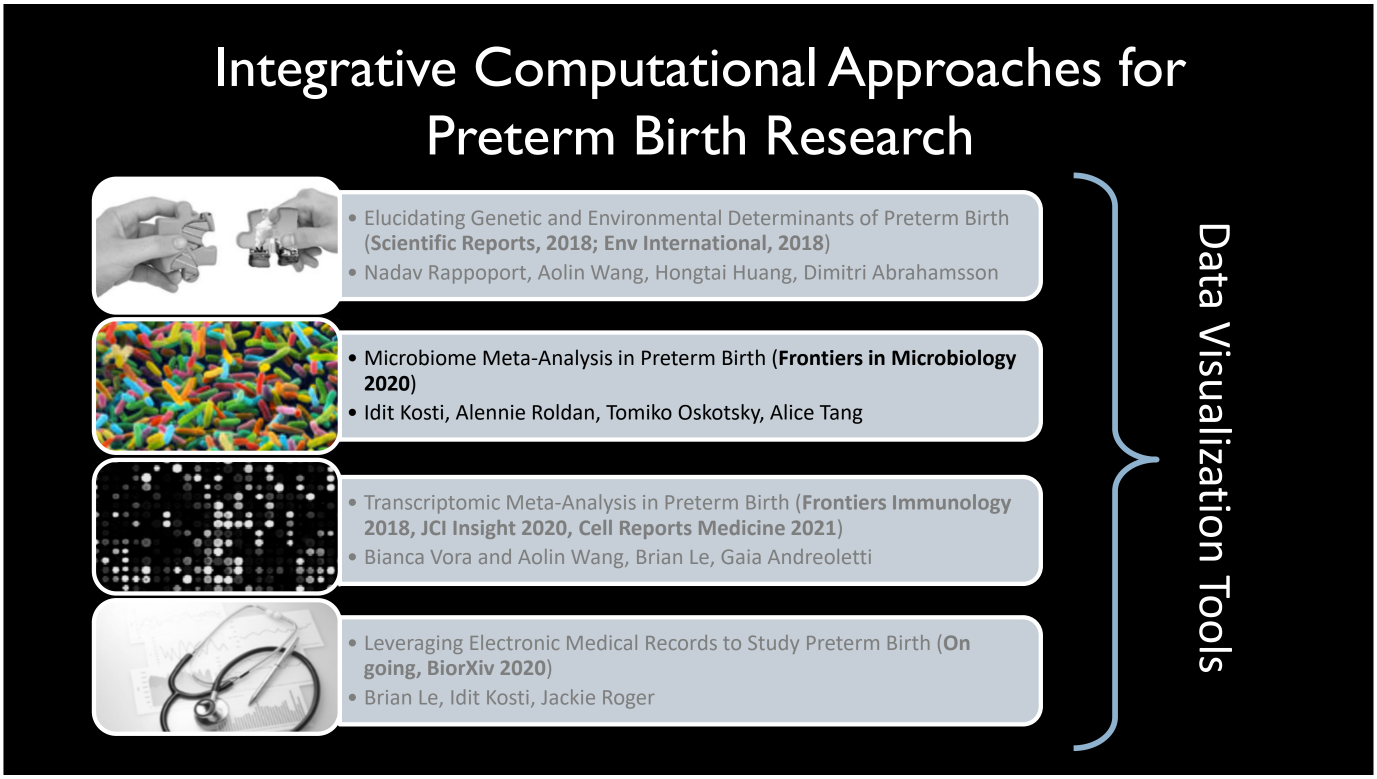
- Elucidating Genetic and Environmental Determinants of Preterm Birth (**Scientific Reports, 2018; Env International, 2018**)
- Nadav Rappoport, Aolin Wang, Hongtai Huang, Dimitri Abrahamsson

- Microbiome Meta-Analysis in Preterm Birth (**Frontiers in Microbiology 2020**)
- Idit Kosti, Alennie Roldan, Tomiko Oskotsky, Alice Tang

- Transcriptomic Meta-Analysis in Preterm Birth (**Frontiers Immunology 2018, JCI Insight 2020, Cell Reports Medicine 2021**)
- Bianca Vora and Aolin Wang, Brian Le, Gaia Andreoletti

- Leveraging Electronic Medical Records to Study Preterm Birth (**On going, BiorXiv 2020**)
- Brian Le, Idit Kosti, Jackie Roger

Data Visualization Tools



Integrative Computational Approaches for Preterm Birth Research

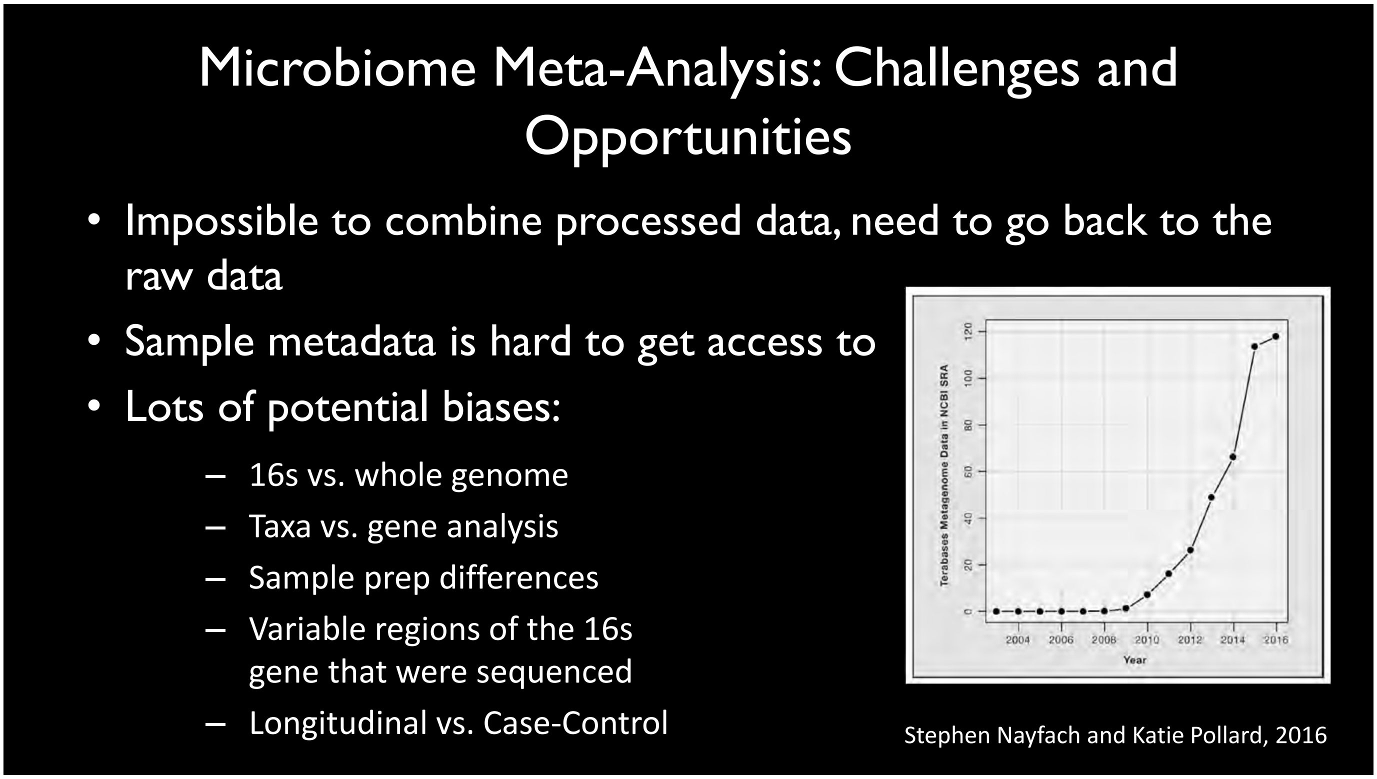
- Elucidating Genetic and Environmental Determinants of Preterm Birth (**Scientific Reports, 2018; Env International, 2018**)
- Nadav Rappoport, Aolin Wang, Hongtai Huang, Dimitri Abrahamsson

- Microbiome Meta-Analysis in Preterm Birth (**Frontiers in Microbiology 2020**)
- Idit Kosti, Alennie Roldan, Tomiko Oskotsky, Alice Tang

- Transcriptomic Meta-Analysis in Preterm Birth (**Frontiers Immunology 2018, JCI Insight 2020, Cell Reports Medicine 2021**)
- Bianca Vora and Aolin Wang, Brian Le, Gaia Andreoletti

- Leveraging Electronic Medical Records to Study Preterm Birth (**On going, BiorXiv 2020**)
- Brian Le, Idit Kosti, Jackie Roger

Data Visualization Tools



Microbiome Meta-Analysis: Challenges and Opportunities

- Impossible to combine processed data, need to go back to the raw data
- Sample metadata is hard to get access to
- Lots of potential biases:
  - 16s vs. whole genome
  - Taxa vs. gene analysis
  - Sample prep differences
  - Variable regions of the 16s gene that were sequenced
  - Longitudinal vs. Case-Control

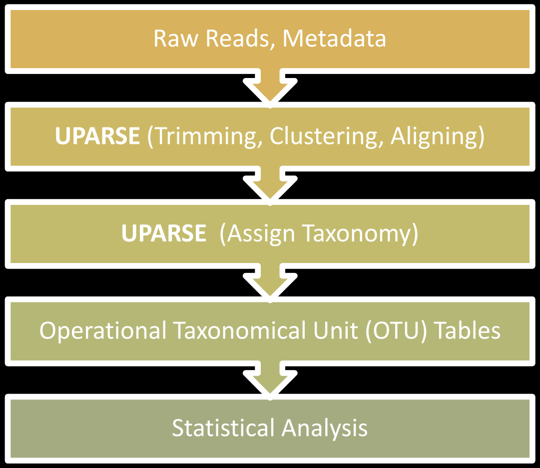
Stephen Nayfach and Katie Pollard, 2016

Public Pregnancy Microbiome Data

	# of samples	Mother/ Baby	# of Subjects	# of body sites	# of Species	# of reads	Reference	Body sites	Type of Data	Time points
HMP	5,298	Controls	242	18	5385	30,336,944	The Human Microbiome Project Consortium	Mouth, Nose, Skin, Gut and Vagina	16s	Controls
Stanford PTB Longitudinal (MOD)	4,399	Mother	51	11	4694	656,308	DiGiulio et al.	Vagina, Gut, Saliva, and Tooth/Gum	16s, WGS	Weekly
Placenta Penn	69	Mother	6	5	2108	207,5881	Lauder et al.	Placenta, Vagina	16s	Delivery
Vaginal PTB	349	Mother	100	1	550	2,213,608	Romero et al.	Vagina	16s	Every 4 weeks until 24, every 2 weeks until the last prenatal visit
Placenta PTB	48	Mother	48	1	TBD	1,348,416	Aagard et al.	Placenta	16s, WGS	Delivery
Gut Microbiome in Pregnancy	972	Mother/ Baby	91	1	2339	112,8572	Koren et al.	Gut	16s	T1,T3,PP
Vaginal Pregnancy	TBD	Mother	42	1	TBD	TBD	MacIntyre et al.	Vagina	16s	T1,T2,T3,PP

Microbiome Meta-Analysis Pipeline

Is time consuming and computationally intensive



Raw Reads, Metadata

UPARSE (Trimming, Clustering, Aligning)

UPARSE (Assign Taxonomy)

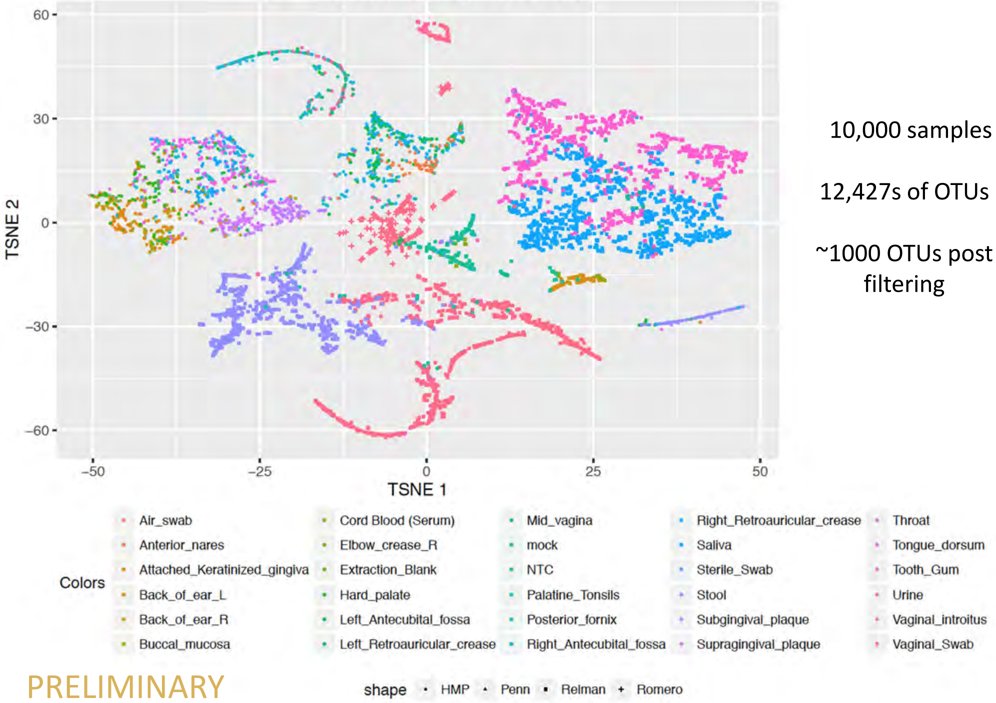
Operational Taxonomical Unit (OTU) Tables

Statistical Analysis

Study Name	Raw Reads Size
HMP	19.03GB
Stanford PTB Longitudinal (MOD)	8.98GB
Placenta Penn	991.2MB
Vaginal PTB	1.28GB
Placenta PTB	490.9MB
Gut Microbiome in Pregnancy	427.2MB

Idit Kosti, PhD

Microbiome Meta-Analysis Across Body Sites



10,000 samples

12,427s of OTUs

~1000 OTUs post filtering

PRELIMINARY

Idit Kosti, PhD

Can we identify new microbial species associated with PTB in a longitudinal analysis leveraging public data?

Bacterial Vaginosis is a Known Risk Factor for PTB

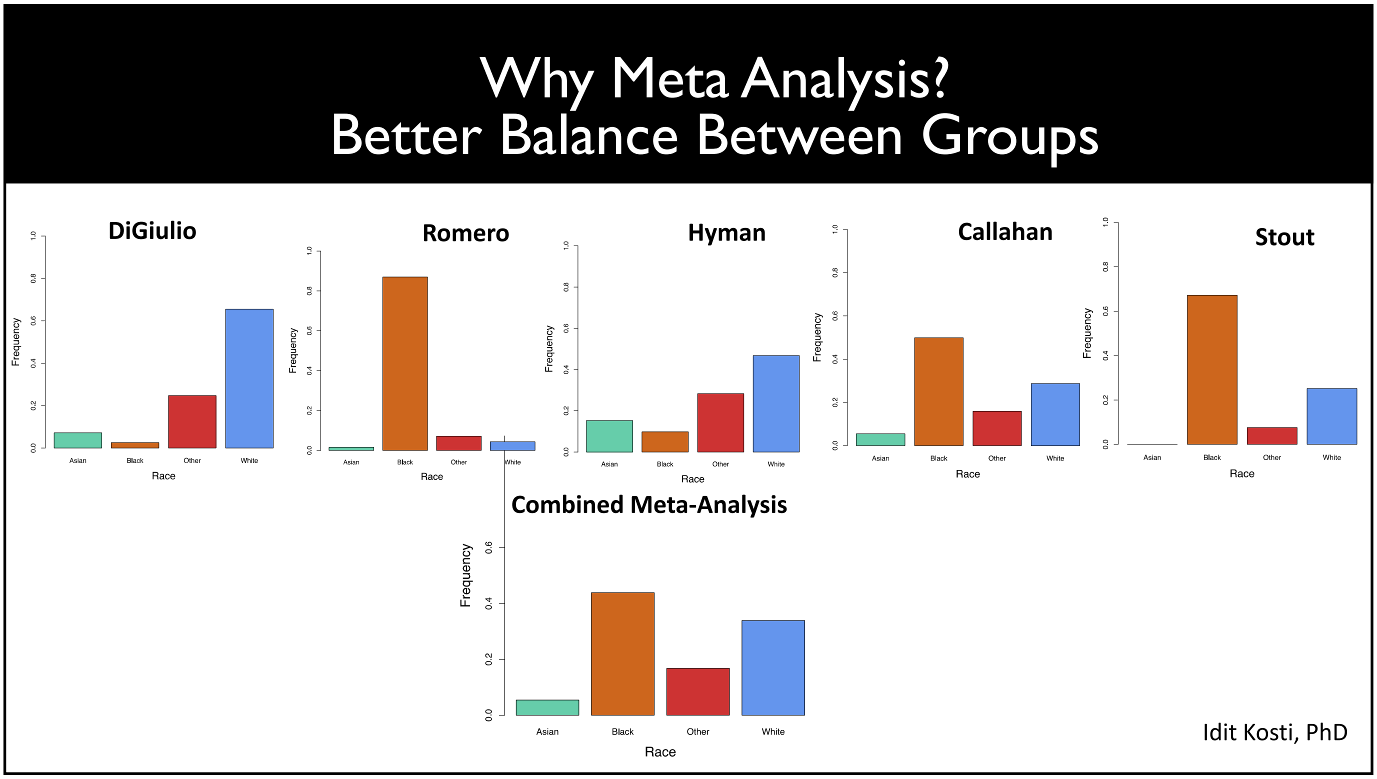
- Bacterial vaginosis is a state of an overgrowth of anaerobic bacteria, replacing the normal vaginal Lactobacillus.
- Bacterial vaginosis has been shown to increase the risk for preterm birth.
- Several microbiome studies have been carried out but no meta-analysis to date.

Manns-James L., J Midwifery Womens Health, 2011

Meta-Analysis Allows Greater Power

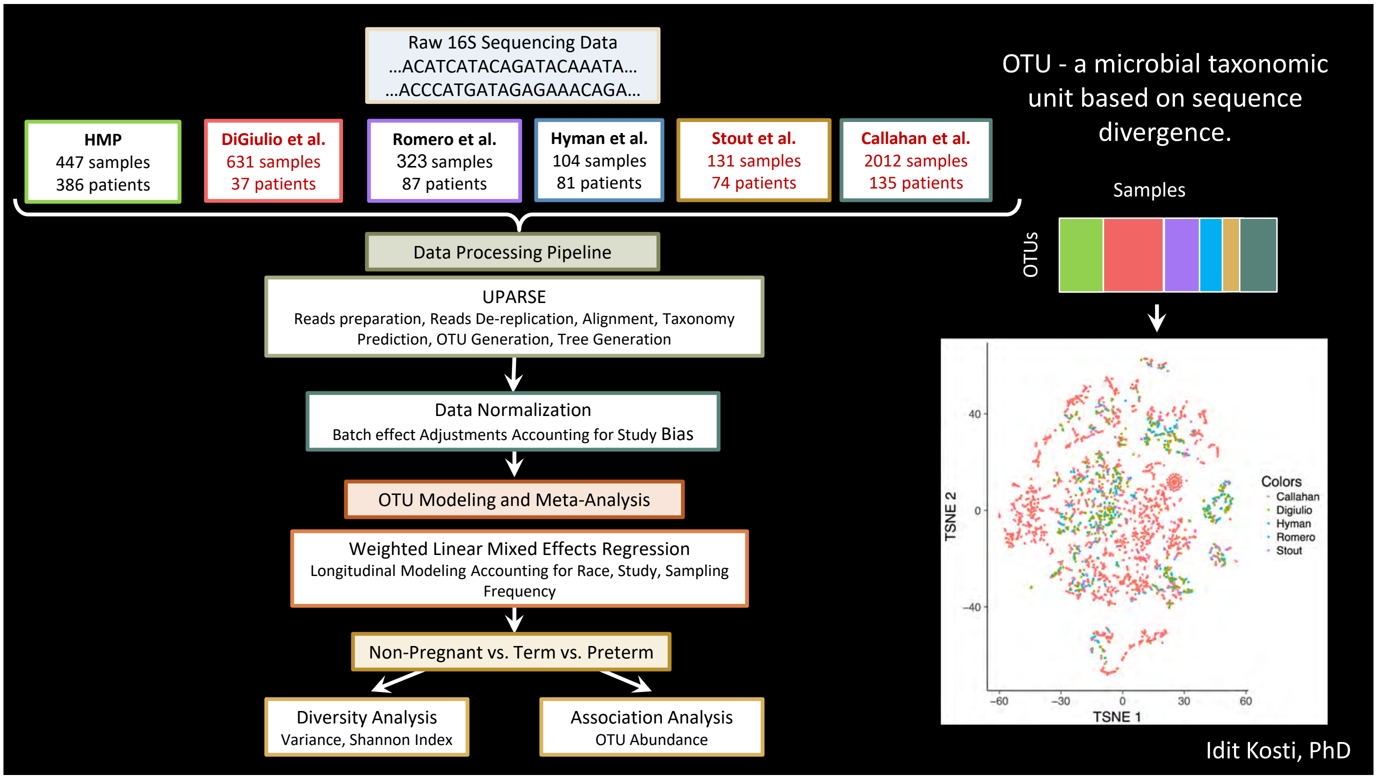
	Callahan et al.	**DiGiulio et al.**	**Hyman et al.**	**Romero et al.**	**Stout et al**
Total Number of Patients	135	37	82	87	74
PTB Patients	50	5	16	18	23
T1 Samples	42 (10 PTB)	21 (4 PTB)	37 (5 PTB)	6 (2 PTB)	14 (4 PTB)
T2 Samples	135 (50 PTB)	31 (4 PTB)	50 (10 PTB)	76 (17 PTB)	55 (18 PTB)
T3 Samples	123 (39 PTB)	36 (4 PTB)	46 (9 PTB)	60 (5 PTB)	59 (17 PTB)
Overall PTB Ratio	37%	12.5%	21%	17.33%	31.1%
Sampling Time Points	Once per week	Once per week	One per trimester	Once every 4 weeks (< 24 gestation weeks) Once every 2 weeks (> 24 gestation weeks)	Once per trimester

Over 3,000 samples and 350+ women



Why Meta Analysis?
Better Balance Between Groups

Idit Kosti, PhD



Raw 16S Sequencing Data
…ACATCATACAGATACAAATA…
…ACCCATGATAGAGAAACAGA…

HMP
447 samples
386 patients

DiGiulio et al.
631 samples
37 patients

Romero et al.
323 samples
87 patients

Hyman et al.
104 samples
81 patients

Stout et al.
131 samples
74 patients

Callahan et al.
2012 samples
135 patients

Data Processing Pipeline

UPARSE
Reads preparation, Reads De-replication, Alignment, Taxonomy Prediction, OTU Generation, Tree Generation

Data Normalization
Batch effect Adjustments Accounting for Study Bias

OTU Modeling and Meta-Analysis

Weighted Linear Mixed Effects Regression
Longitudinal Modeling Accounting for Race, Study, Sampling Frequency

Non-Pregnant vs. Term vs. Preterm

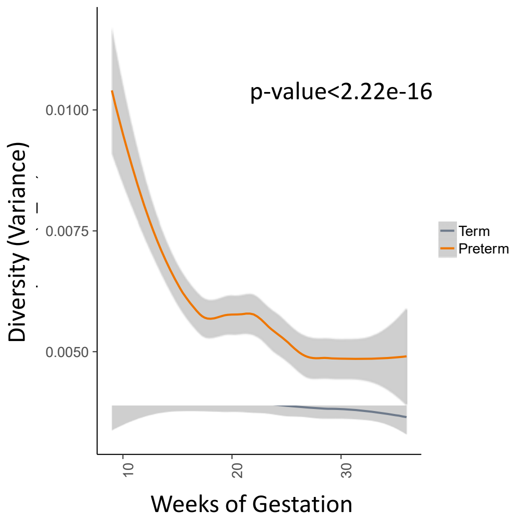
Diversity Analysis
Variance, Shannon Index

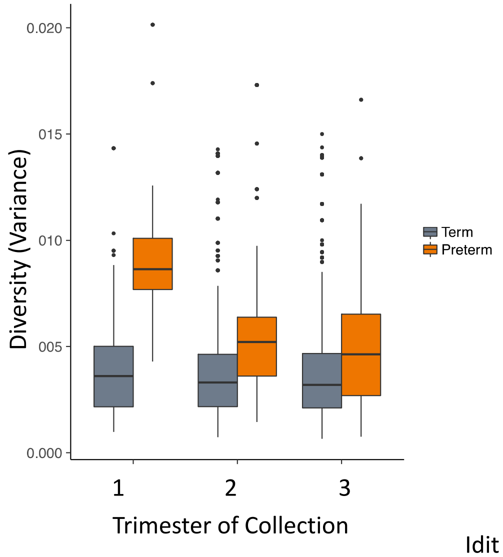
Association Analysis
OTU Abundance

OTU - a microbial taxonomic unit based on sequence divergence.

Idit Kosti, PhD

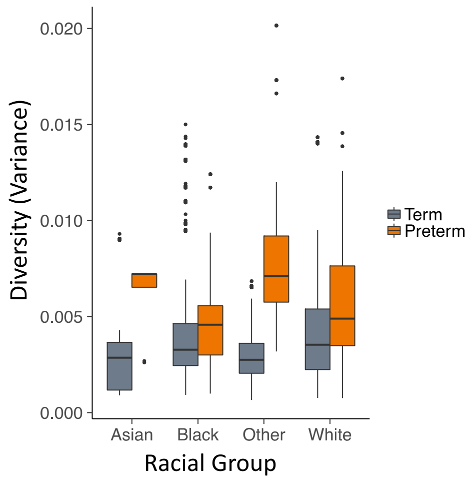
Higher Diversity is Associated with PTB in the First Trimester (T1)

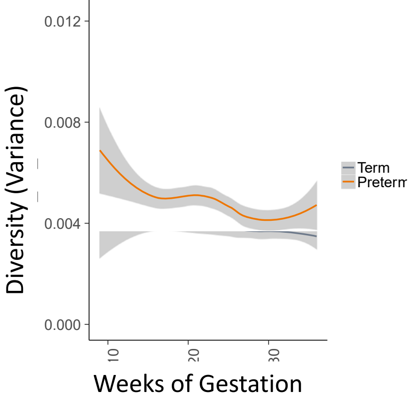




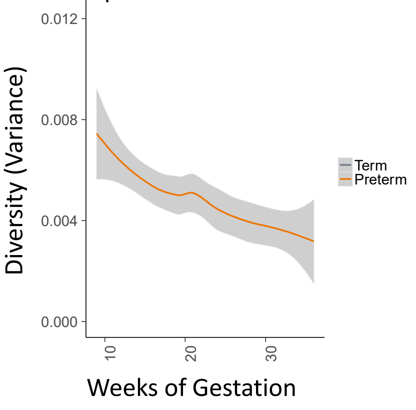
Idit Kosti, PhD

The Association Holds Across Racial Groups



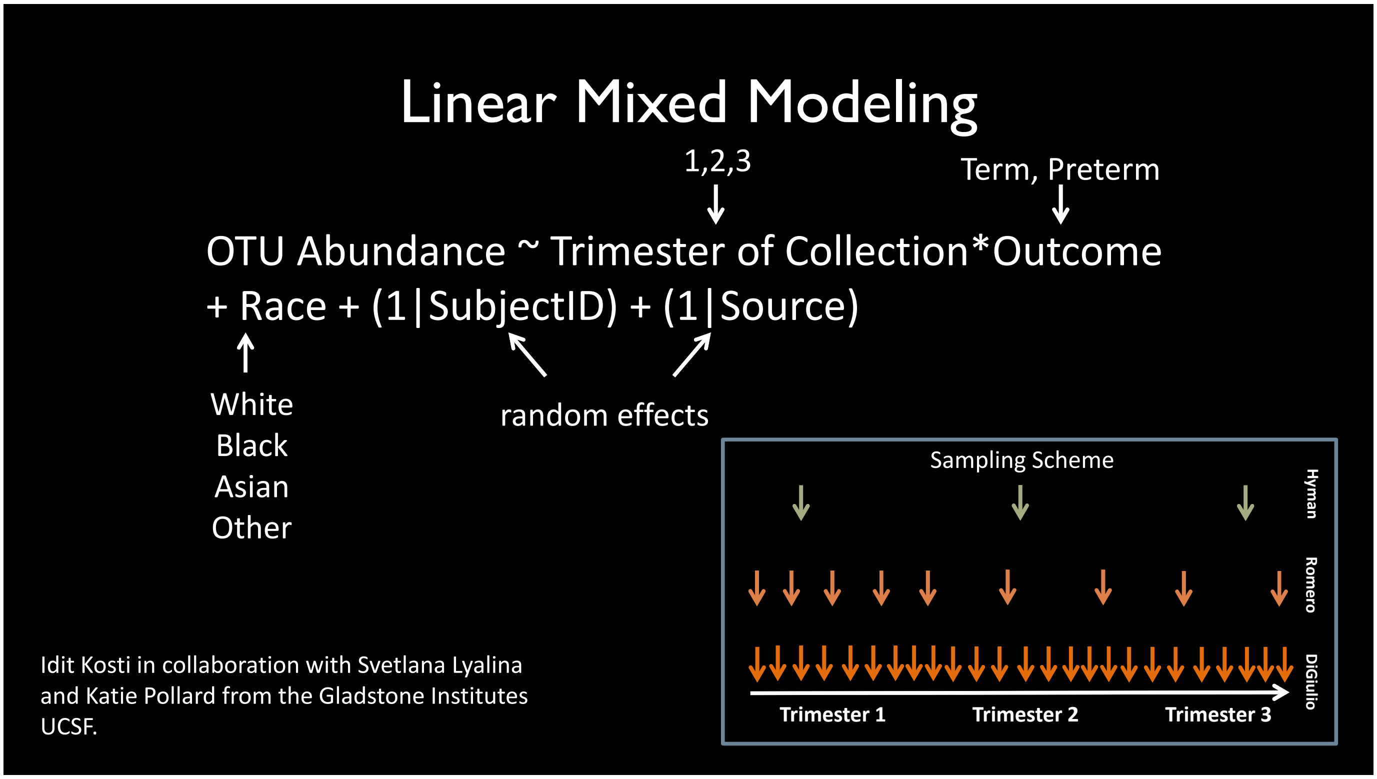

Black

p-value = 2.6e-4


White

p-value = 1.7e-6

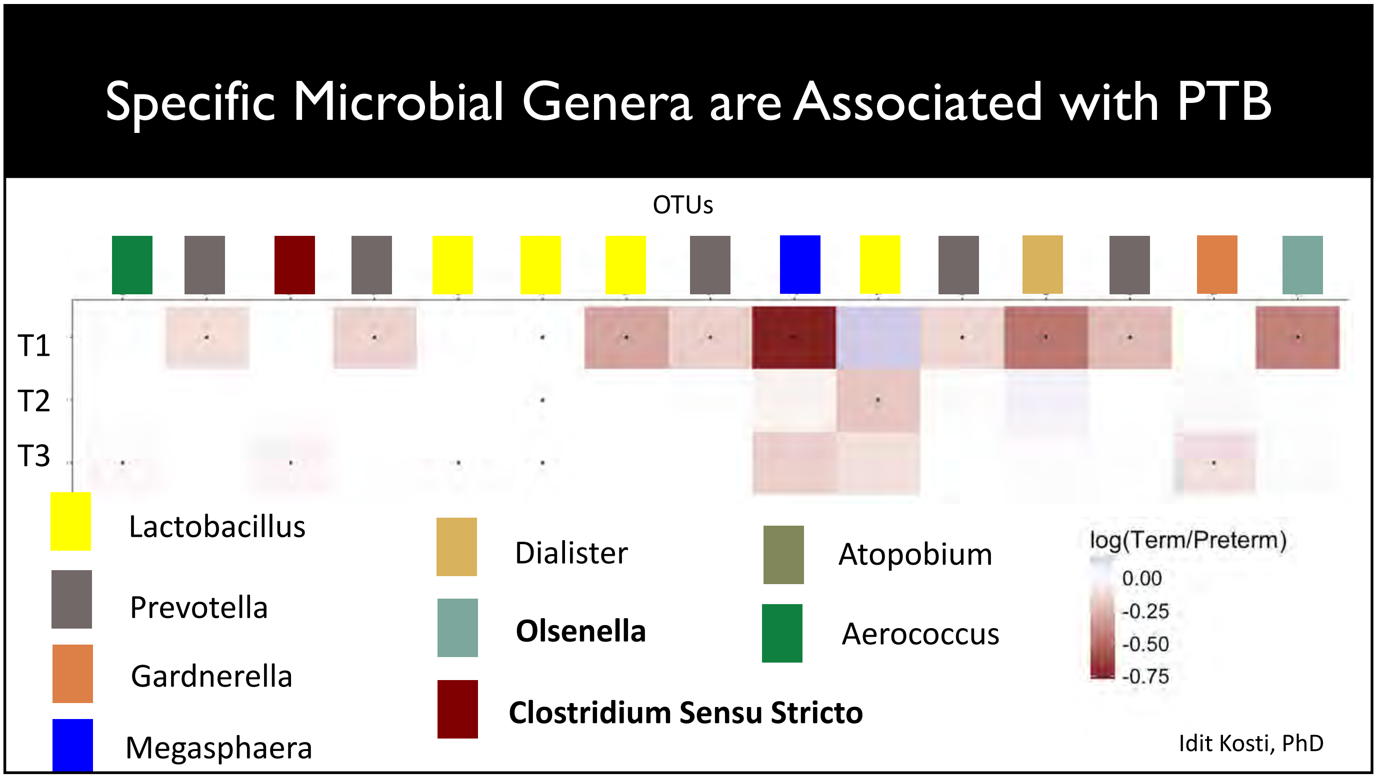
Idit Kosti, PhD



Linear Mixed Modeling

OTU Abundance ~ Trimester of Collection*Outcome + Race + (1|SubjectID) + (1|Source)

Idit Kosti in collaboration with Svetlana Lyalina and Katie Pollard from the Gladstone Institutes UCSF.



Specific Microbial Genera are Associated with PTB

Idit Kosti, PhD

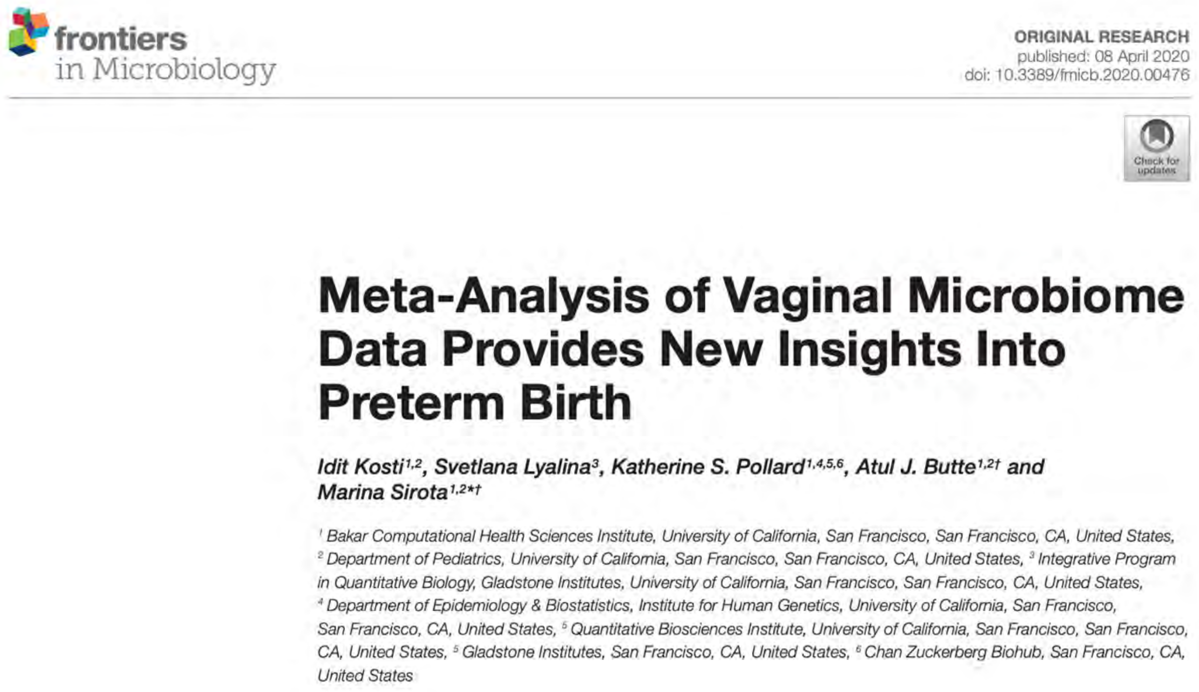
Novel Bacterial Genera Identified by Meta-Analysis

OTU	Genera	Trimester	Adj. P-value	Bacterial Vaginosis	Associated with PTB
Otu875	Olsenella	1	<2.2e-16	✓	✗
Otu1238	Clostridium sensu stricto	3	1.58E-03	✗	✗

Summary

- We observe higher microbial diversity in women who deliver preterm, especially in the first trimester.
- We identify novel bacteria associated with preterm birth using a meta-analysis approach leveraging public data.
- First trimester vaginal sampling may help identify those at risk for preterm birth.

Kosti I, Lyalina S, … Pollard K, Butte AJ, **Sirota M**. Meta-Analysis of Vaginal Microbiome Data Provides New Insights In Preterm Birth. *Frontiers in Microbiology, 2020.*



frontiers in Microbiology

ORIGINAL RESEARCH
published: 08 April 2020
doi: 10.3389/fmicb.2020.00476

Check for updates

Meta-Analysis of Vaginal Microbiome Data Provides New Insights Into Preterm Birth

Idit Kosti[1,2], Svetlana Lyalina[3], Katherine S. Pollard[1,4,5,6], Atul J. Butte[1,2†] and Marina Sirota[1,2†]*

[1] Bakar Computational Health Sciences Institute, University of California, San Francisco, San Francisco, CA, United States, [2] Department of Pediatrics, University of California, San Francisco, San Francisco, CA, United States, [3] Integrative Program in Quantitative Biology, Gladstone Institutes, University of California, San Francisco, San Francisco, CA, United States, [4] Department of Epidemiology & Biostatistics, Institute for Human Genetics, University of California, San Francisco, San Francisco, CA, United States, [5] Quantitative Biosciences Institute, University of California, San Francisco, San Francisco, CA, United States, [5] Gladstone Institutes, San Francisco, CA, United States, [6] Chan Zuckerberg Biohub, San Francisco, CA, United States

Kosti et al, Frontiers in Microbiology, 2020.

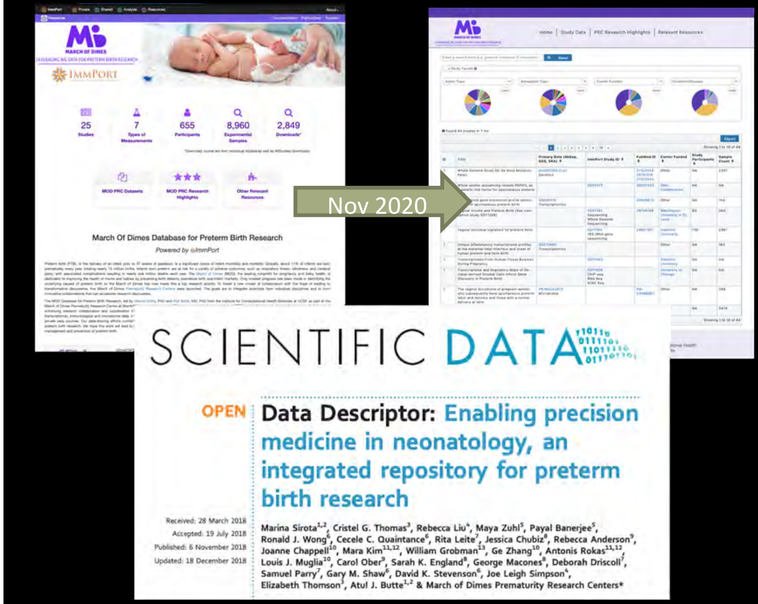
The Goals of Creating a PTB Data Repository

- Serve as a central data repository for omics data across the 6 MOD Transdisciplinary Centers
  - Transcriptomics, genomics, microbiome, proteome, CYTOF, methylation, metabolome, etc.
- Enable new scientific questions
- Enhance collaboration and coordination among centers
- Accelerate the pace of discovery

Tomiko Oskotsky, MD

MARCH OF DIMES

MOD Biorepository Database: Brief Overview



SCIENTIFIC DATA

OPEN Data Descriptor: Enabling precision medicine in neonatology, an integrated repository for preterm birth research

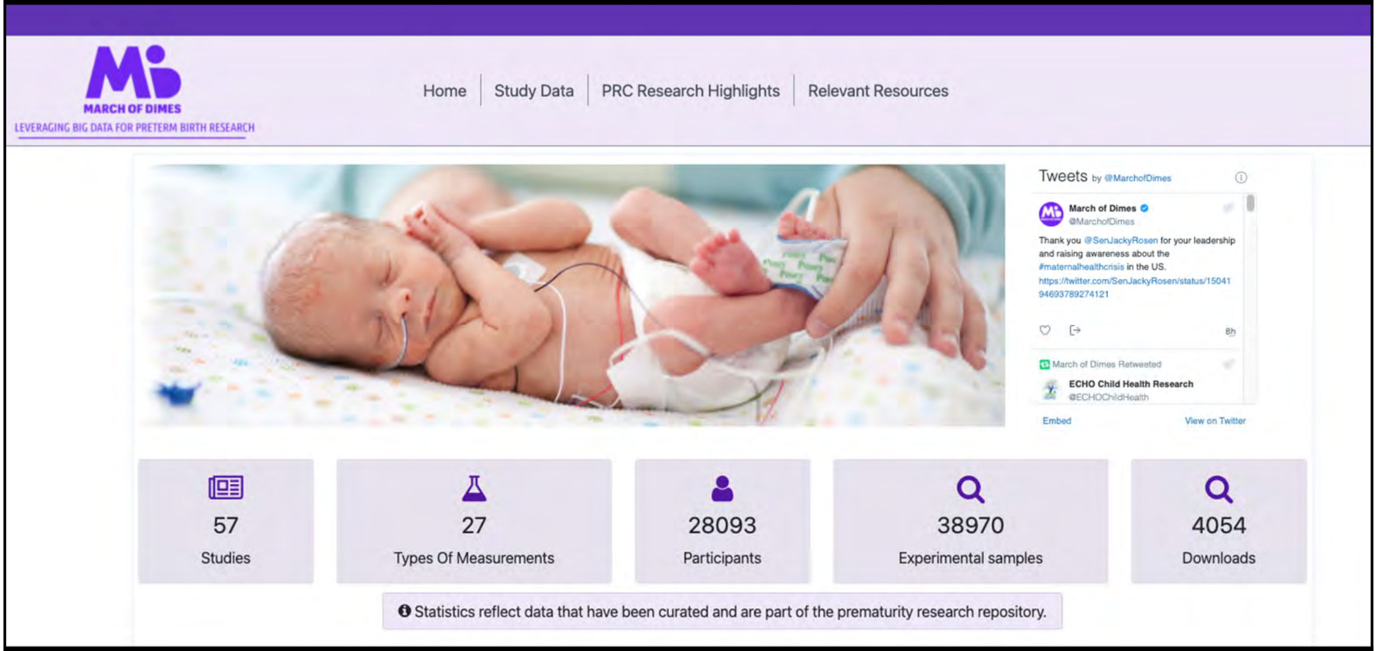
Received: 28 March 2018
Accepted: 19 July 2018
Published: 6 November 2018
Updated: 18 December 2018

Marina Sirota[1,2], Cristel G. Thomas[3], Rebecca Liu[4], Maya Zuhl[5], Payal Banerjee[5], Ronald J. Wong[6], Cecele C. Quaintance[6], Rita Leite[7], Jessica Chubiz[8], Rebecca Anderson[9], Joanne Chappell[10], Mara Kim[11,12], William Grobman[13], Ge Zhang[10], Antonis Rokas[11,12], Louis J. Muglia[10], Carol Ober[9], Sarah K. England[8], George Macones[8], Deborah Driscoll[7], Samuel Parry[7], Gary M. Shaw[6], David K. Stevenson[6], Joe Leigh Simpson[4], Elizabeth Thomson[3], Atul J. Butte[1,2] & March of Dimes Prematurity Research Centers*

- Launched in 2017, spun out into an independent resource in 2020
- Regularly updated with publications, study data, and other preterm birth resources from MOD Prematurity Research Centers (PRCs) and other groups
- https://pretermbirthdb.org

Sirota et al, Scientific Data 2018.

https://pretermbirthdb.org



March of Dimes @MarchofDimes

Thank you @SenJackyRosen for your leadership and raising awareness about the #maternalhealthcrisis in the US. https://twitter.com/SenJackyRosen/status/1504194693789274121

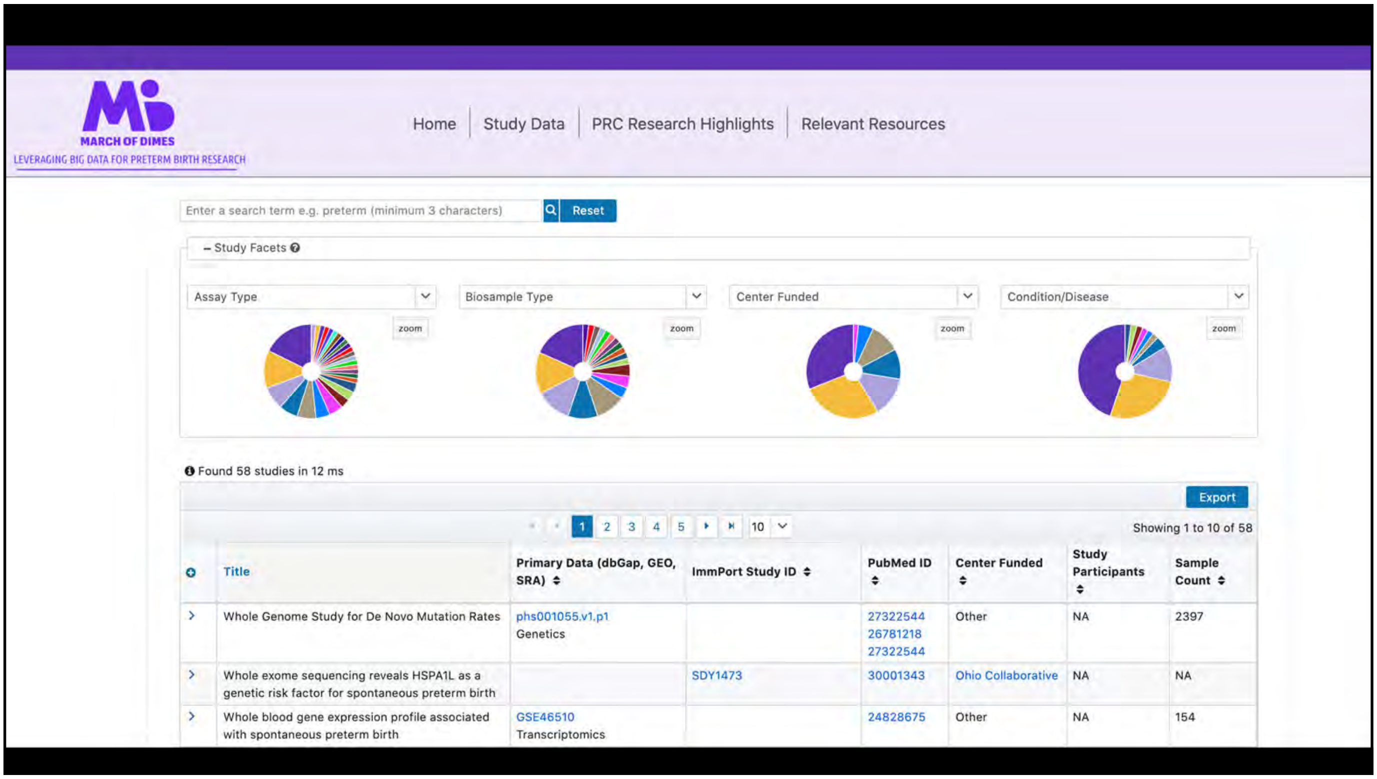
8h

March of Dimes Retweeted

ECHO Child Health Research @ECHOChildHealth

57	27	28093	38970	4054
Studies	Types Of Measurements	Participants	Experimental samples	Downloads

Statistics reflect data that have been curated and are part of the prematurity research repository.



Found 58 studies in 12 ms

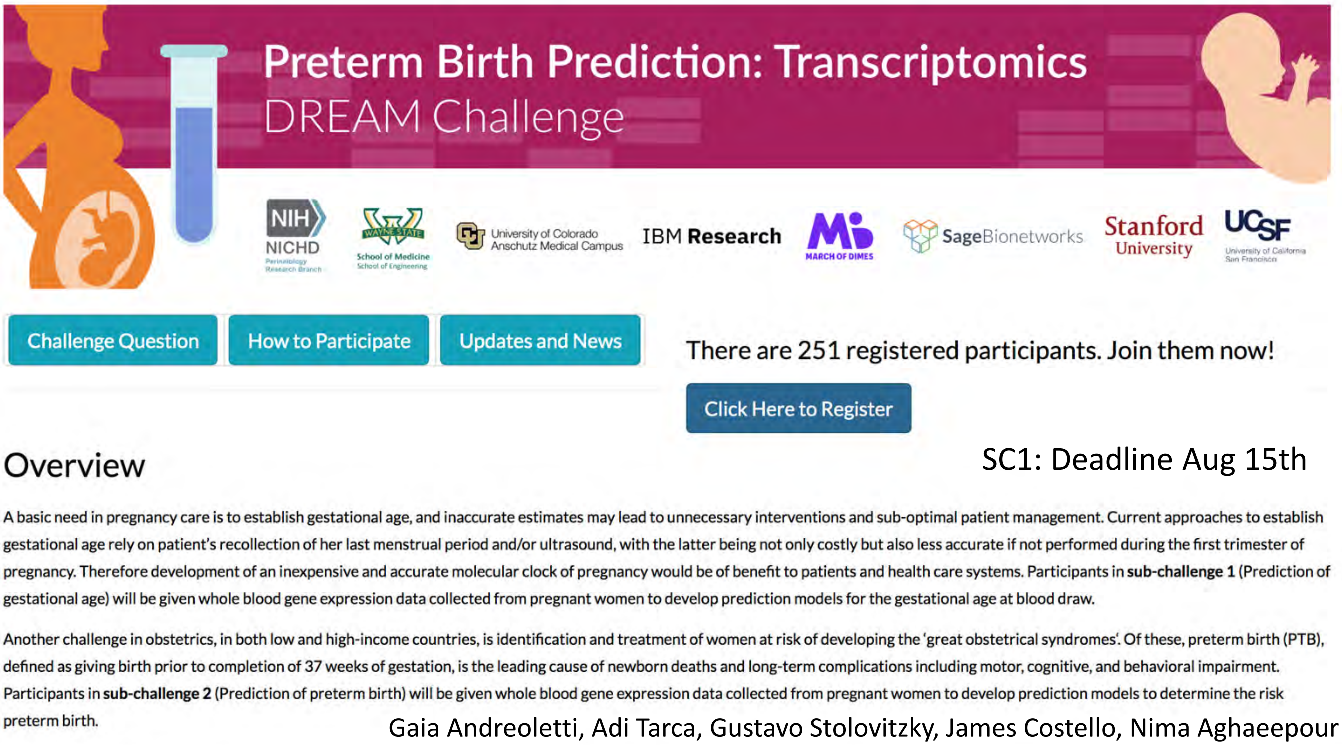
Showing 1 to 10 of 58

	Title	Primary Data (dbGap, GEO, SRA)	ImmPort Study ID	PubMed ID	Center Funded	Study Participants	Sample Count
>	Whole Genome Study for De Novo Mutation Rates	phs001055.v1.p1 Genetics		27322544 26781218 27322544	Other	NA	2397
>	Whole exome sequencing reveals HSPA1L as a genetic risk factor for spontaneous preterm birth		SDY1473	30001343	Ohio Collaborative	NA	NA
>	Whole blood gene expression profile associated with spontaneous preterm birth	GSE46510 Transcriptomics		24828675	Other	NA	154

DREAM Challenges for PTB

- The DREAM Challenges are crowdsourcing challenges examining questions in biology and medicine
- **Goal:** Bring computational community to MOD data resulting in rigorous, reproducible science and innovation
- **Question:** How can we predict those at higher risk of PTB (term vs. preterm or week of gestation) based on:
  - Transcriptomics Data, Microbiome Data, Multi-omics
  - Can use data in the repository for training, need datasets for testing!

Tomiko Oskotsky, Alennie Roldan, Jonathan Golob, Jake Albrecht, Gaia Andreoletti, Gustavo Stolovitzky, James Costello, Nima Aghaeepour, Adi Tarca, Roberto Romero, David MacIntyre, Phil Bennett, Yooni Li



Preterm Birth Prediction: Transcriptomics DREAM Challenge

There are 251 registered participants. Join them now!

SC1: Deadline Aug 15th

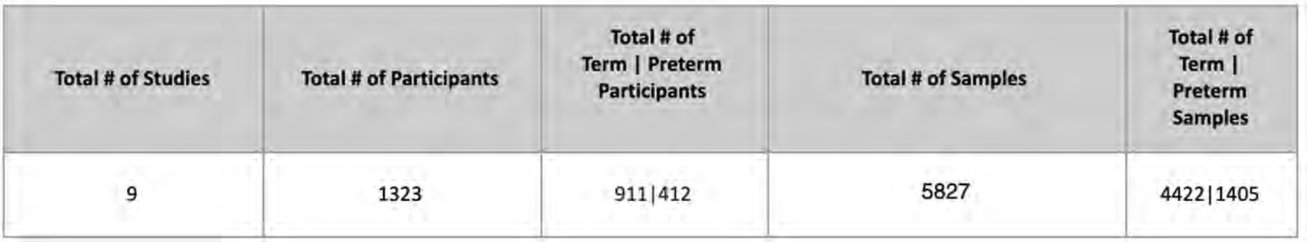
Overview

A basic need in pregnancy care is to establish gestational age, and inaccurate estimates may lead to unnecessary interventions and sub-optimal patient management. Current approaches to establish gestational age rely on patient's recollection of her last menstrual period and/or ultrasound, with the latter being not only costly but also less accurate if not performed during the first trimester of pregnancy. Therefore development of an inexpensive and accurate molecular clock of pregnancy would be of benefit to patients and health care systems. Participants in **sub-challenge 1** (Prediction of gestational age) will be given whole blood gene expression data collected from pregnant women to develop prediction models for the gestational age at blood draw.

Another challenge in obstetrics, in both low and high-income countries, is identification and treatment of women at risk of developing the 'great obstetrical syndromes'. Of these, preterm birth (PTB), defined as giving birth prior to completion of 37 weeks of gestation, is the leading cause of newborn deaths and long-term complications including motor, cognitive, and behavioral impairment. Participants in **sub-challenge 2** (Prediction of preterm birth) will be given whole blood gene expression data collected from pregnant women to develop prediction models to determine the risk preterm birth.

Gaia Andreoletti, Adi Tarca, Gustavo Stolovitzky, James Costello, Nima Aghaeepour

DREAM Challenge for PTB: Microbiome

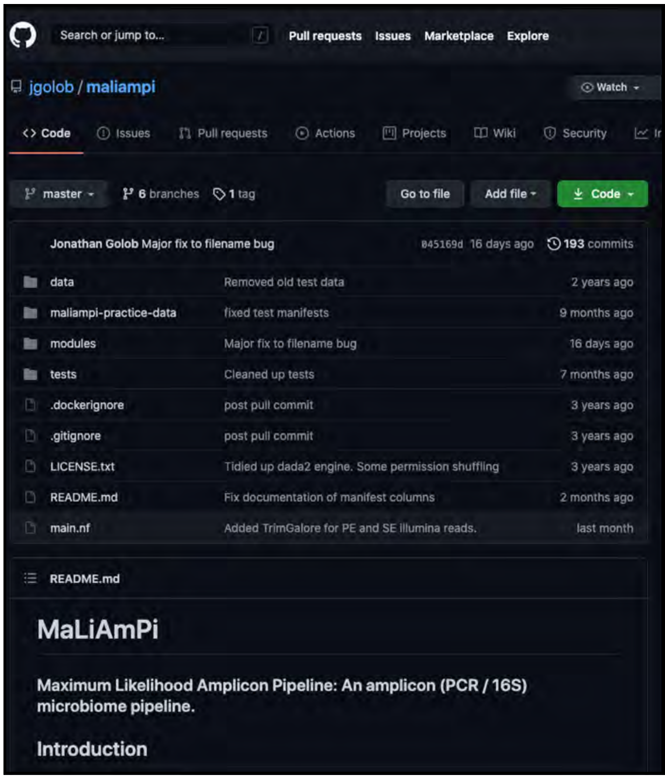


Total # of Studies	Total # of Participants	Total # of Term | Preterm Participants	Total # of Samples	Total # of Term | Preterm Samples
9	1323	911|412	5827	4422|1405

- **9 studies** incorporated into metadata analyses
- **1,323 participants** ~ 2:1 term:preterm participants (911 term / 412 preterm)
- **5827 samples** ~ 3:1 term:preterm samples (4422 term / 1405 preterm)

Possible Sub-challenges
1: Prediction of gestational age
2: Prediction of preterm birth

Study Accession ID	Center	Title	# of Participants	# of Term | PTB Participants	# of Samples	# of Term|PTB Samples	V Region Sequences
PRJEB21325	**Imperial College London**	Vaginal dysbiosis increases risk of preterm fetal membrane rupture, neonatal sepsis and is exacerbated by erythromycin. (Brown 2018)	116	21 | 95	158	23 | 135	V1 - V2
PRJEB30642	**Imperial College London**	Establishment of vaginal microbiota composition in early pregnancy and its association with subsequent preterm prelabor rupture of the fetal membranes.	98	35|63	254	78|176	V1 - V2
PRJNA242473	University of Maryland	The vaginal microbiota of pregnant women who subsequently have spontaneous preterm labor and delivery and those with a normal delivery at term	88	70|18	340	284|56	V1 - V3
PRJNA294119	**Washington University**	Early pregnancy vaginal microbiome trends and preterm birth	74	51|23	145	99|46	V1 - V3
PRJNA393472	**Stanford University**	Replication and Refinement of a Vaginal Microbial Signature of Preterm Birth	133	83|50	2134	1586|548	V4
PRJEB11895 & PRJEB12577	**Imperial College London**	The interaction between vaginal microbiota, cervical length, and vaginal progesterone treatment for preterm birth risk	154	121|33	154	121|33	V1 - V3
SDY465	**Stanford University**	Temporal and spatial variation of the human microbiota during pregnancy	40	33|7	1005	858 | 147	V3 - V5
JX871219 & JX871316	UCSF	Diversity of the Vaginal Microbiome Correlates With Preterm Birth	82	65|17	134	110|24	*Nearly* Complete Gene
PRJNA504518 (phs001739.v1.p1.)	**University of Pennsylvania**	Cervicovaginal microbiota and local immune response modulate the risk of spontaneous preterm delivery	538	432|106	1503	1263 | 240	V3 - V4



Jonathan Golob Major fix to filename bug 045169d 16 days ago 193 commits

data	Removed old test data	2 years ago
maliampi-practice-data	fixed test manifests	9 months ago
modules	Major fix to filename bug	16 days ago
tests	Cleaned up tests	7 months ago
.dockerignore	post pull commit	3 years ago
.gitignore	post pull commit	3 years ago
LICENSE.txt	Tidied up dada2 engine. Some permission shuffling	3 years ago
README.md	Fix documentation of manifest columns	2 months ago
main.nf	Added TrimGalore for PE and SE illumina reads.	last month

README.md

MaLiAmPi

Maximum Likelihood Amplicon Pipeline: An amplicon (PCR / 16S) microbiome pipeline.

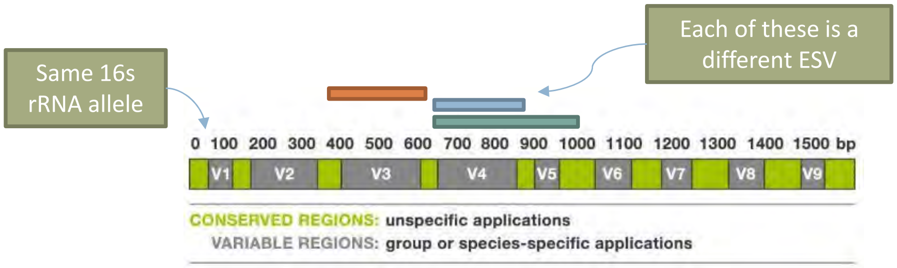
Introduction

MALIAMPI

- **Maximum Likelihood Amplicon Pipeline: An amplicon (PCR / 16S) microbiome pipeline.**
- Maliampi is a phylogenetic placement-based pipeline for handling 16S rRNA amplicons
- Leverage inherent ambiguity involved in any PCR / sequencing based approach, as well as rationally deal with the limits of the available references mapping between sequence-space and taxonomy.
- **A typical workflow:**
  - Create sequence variants
  - Make a reference package
  - Place on the reference package
  - Classify

Jonathan Golob, MD, PhD

Problem: When combining 16S rRNA gene variable region ESVs from *different* studies, using different primers / sequencers / etc, the same organism will be split into different ESVs.

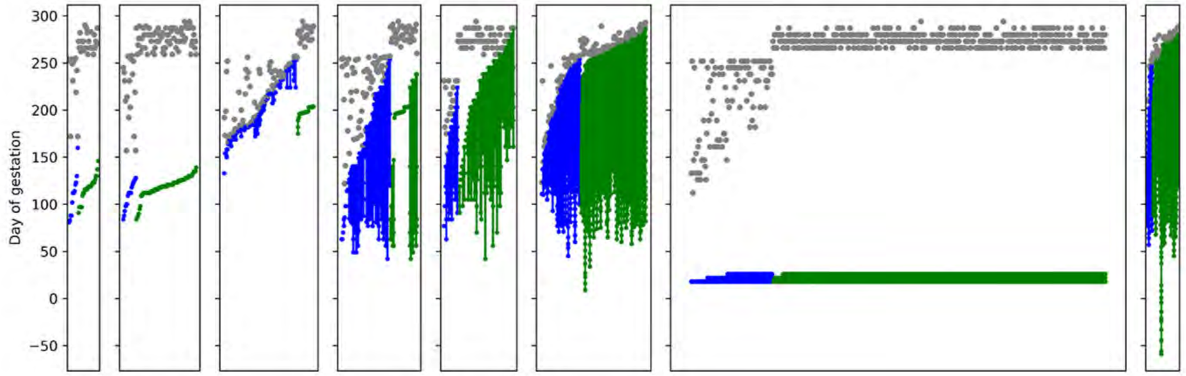

Same 16s rRNA allele

Each of these is a different ESV

Solution: Phylogenetically place the ESVs onto a phylogeny of full-length 16S rRNA alleles possibly in the collection of specimens being analyzed.

Compare the weighted phylogenetic trees across conditions

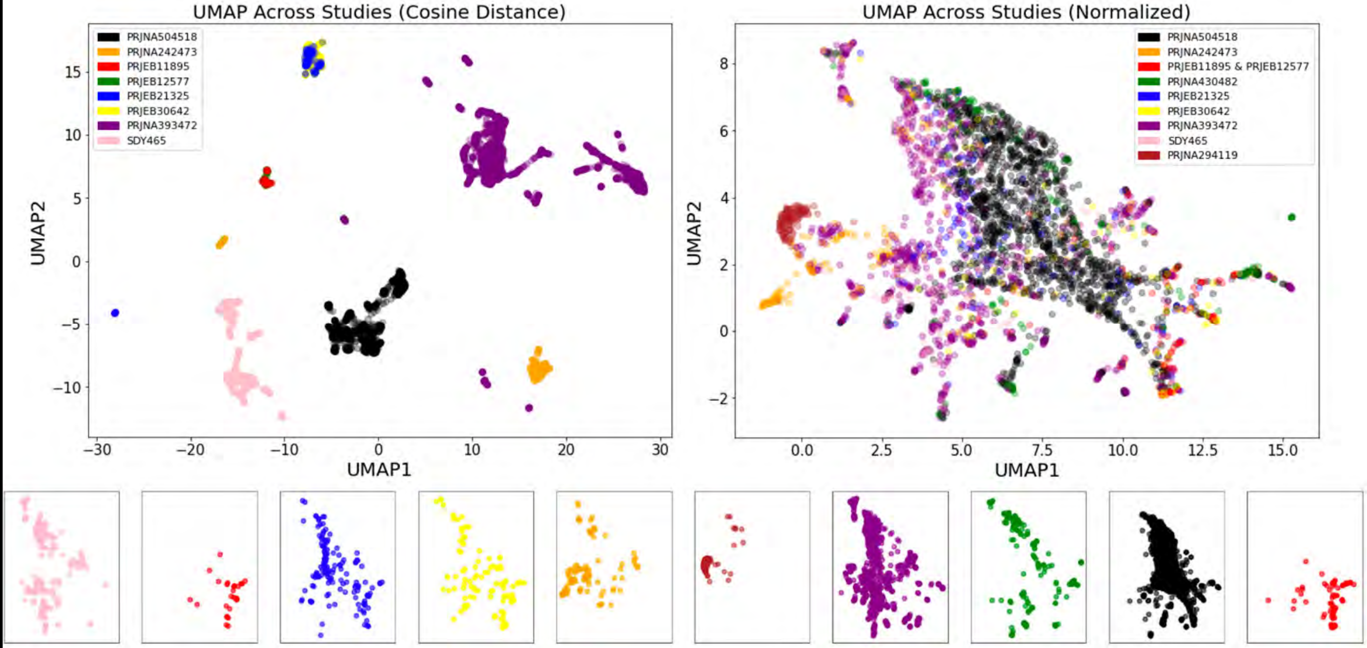
Most studies have longitudinal collection of specimens from participants. The time and frequency of collection vary by study. This is both an analytic challenge and opportunity.



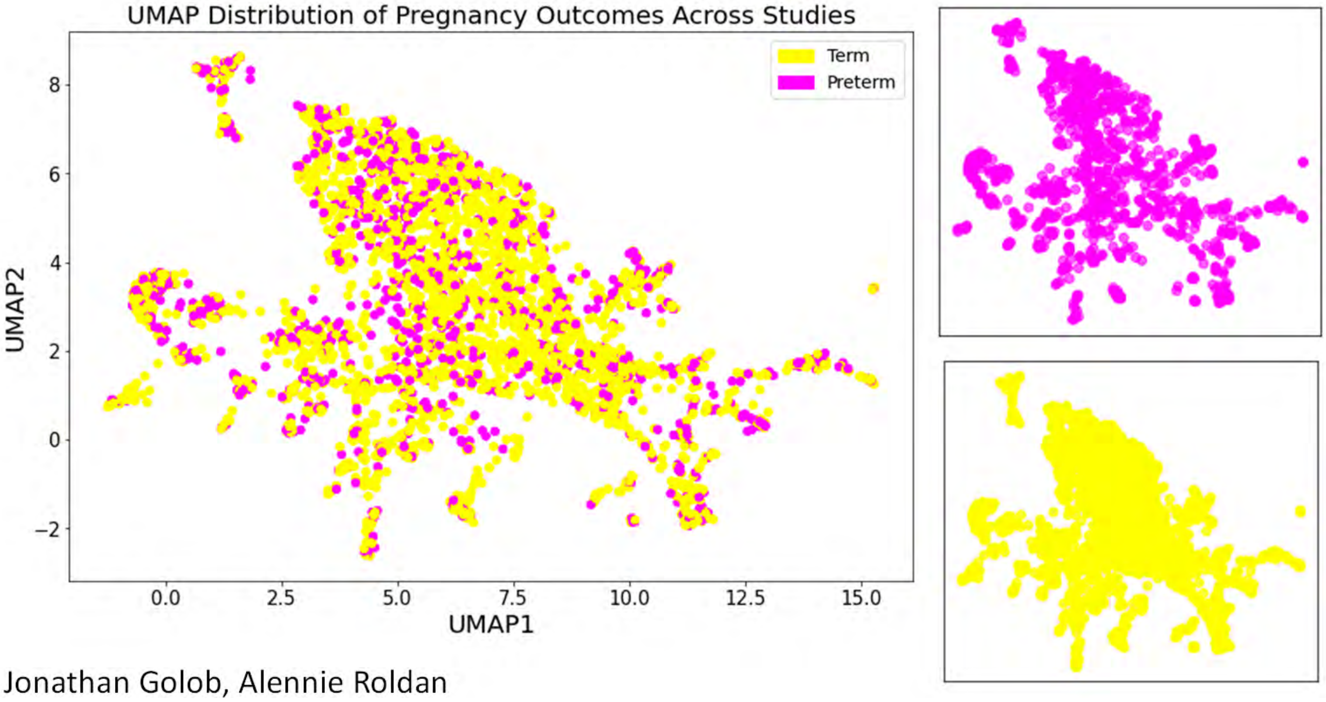
Each block is a *study*.
Each column is a participant / pregnancy.
Grey: Gestational day of delivery
Blue: Specimen collected from pre-term pregnancy
Green: Specimen collected from Term pregnancy

BEFORE NORMALIZATION

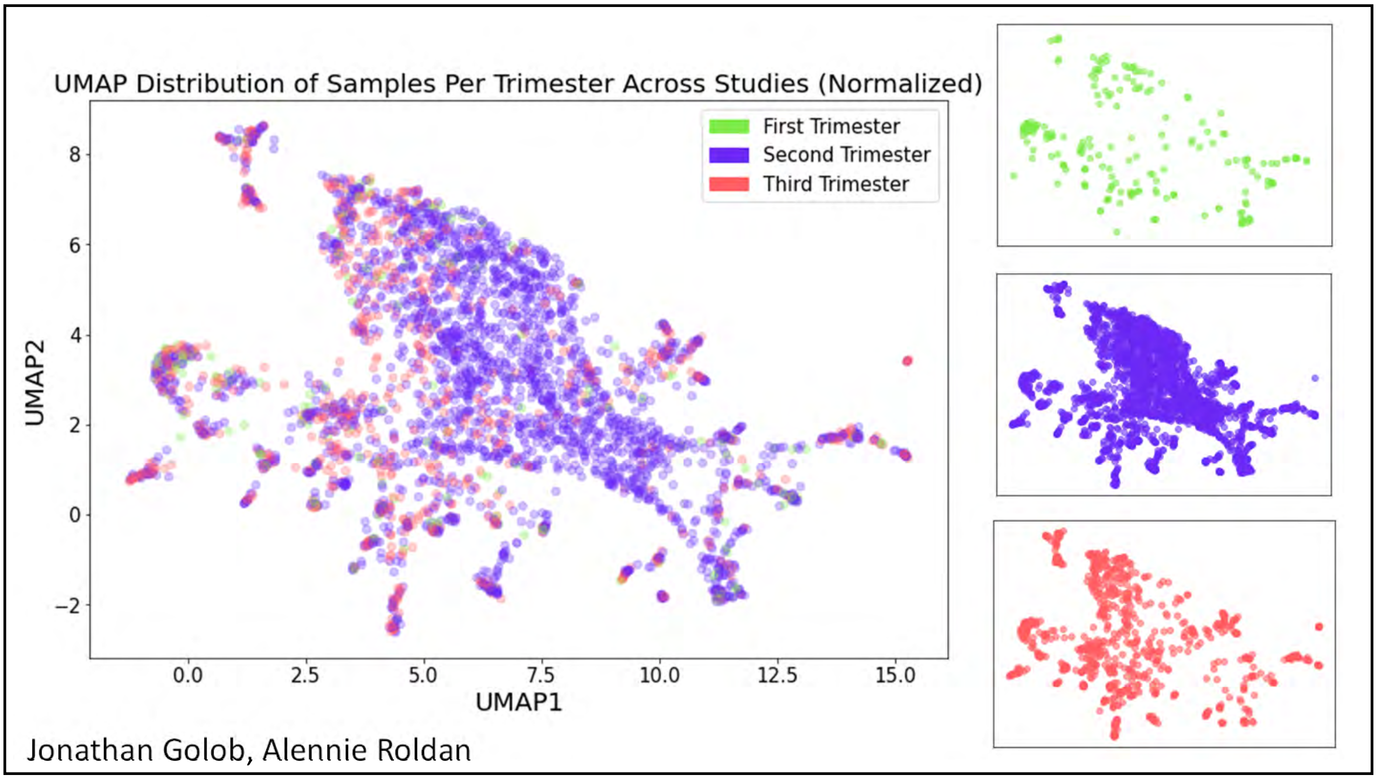
AFTER NORMALIZATION



Alennie Roldan, Jonathan Golob



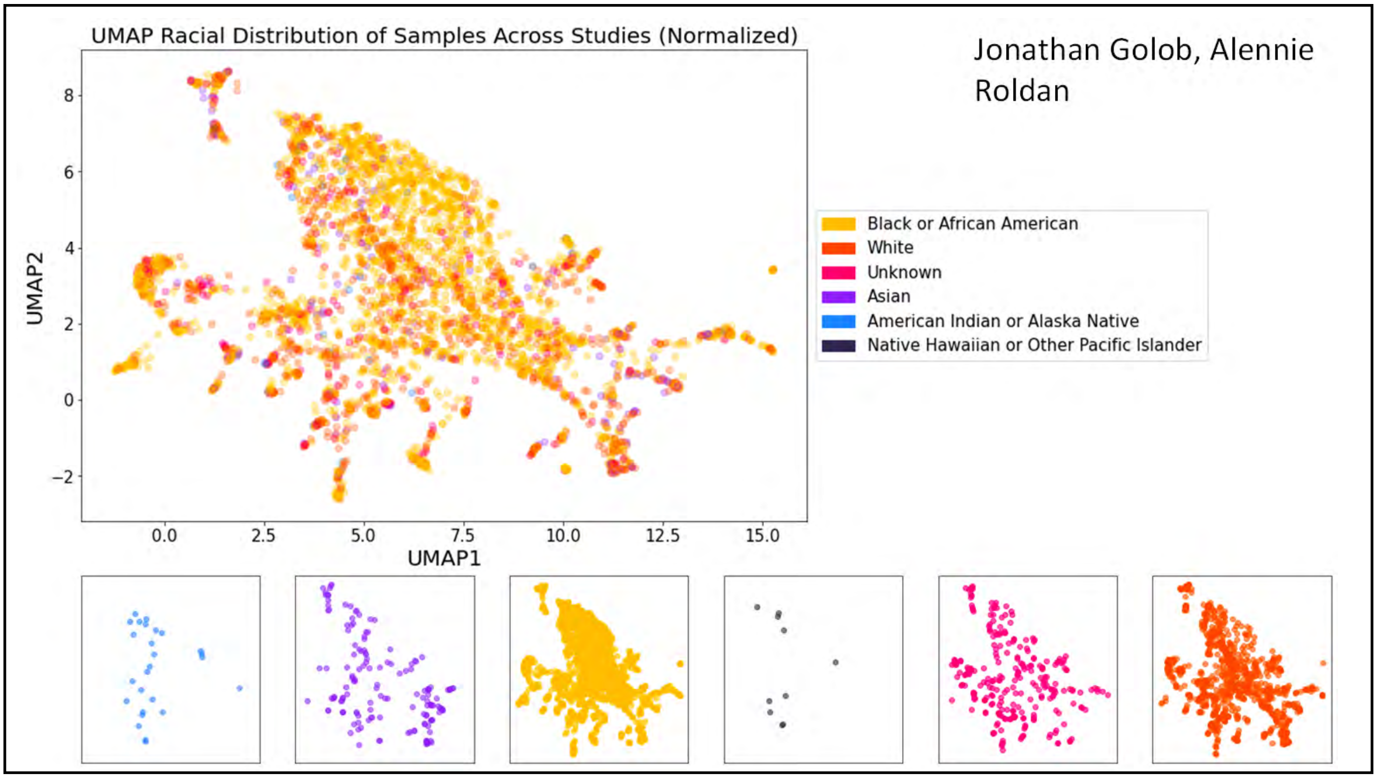
Jonathan Golob, Alennie Roldan



UMAP Distribution of Samples Per Trimester Across Studies (Normalized)

Jonathan Golob, Alennie Roldan

49

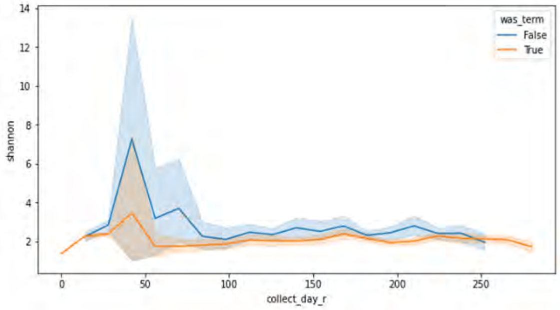


UMAP Racial Distribution of Samples Across Studies (Normalized)

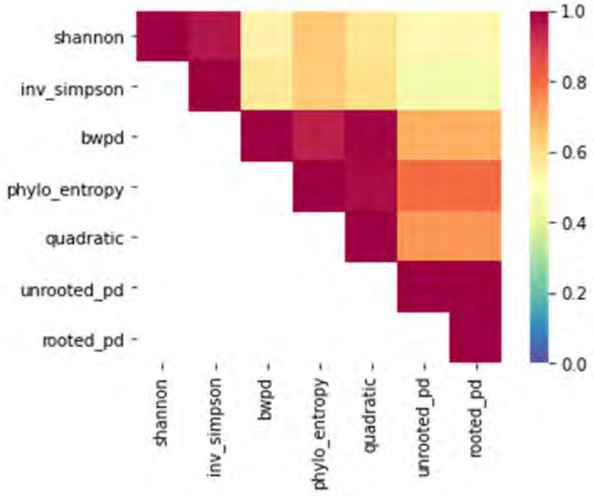
Jonathan Golob, Alennie Roldan

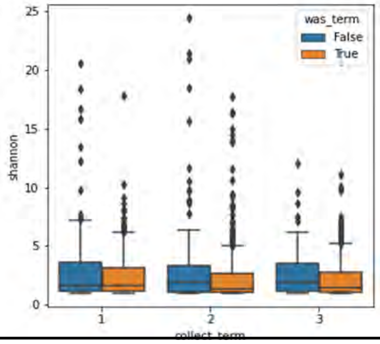
50

Diversity Analysis is Consistent With Previous Findings



Rounded to nearest 2 wks, 95% Confidence Interval



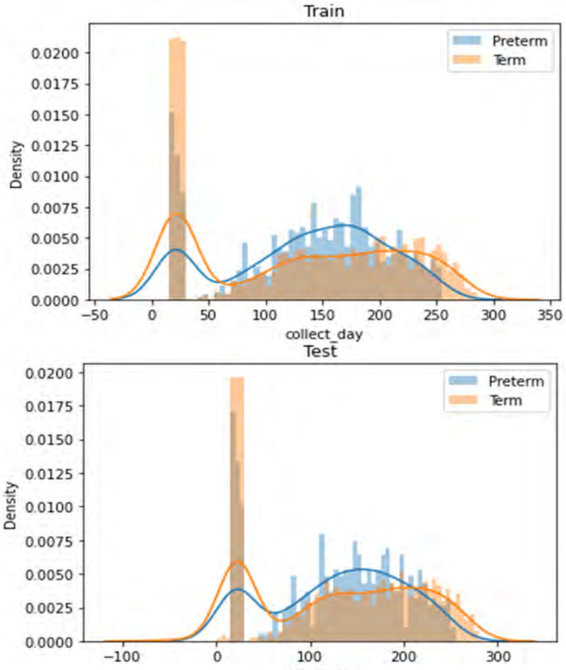


Mann Whitney U Test:
trim: 1 pval: 0.30
trim: 2 pval: 2.9e-07
trim: 3 pval: 0.00013

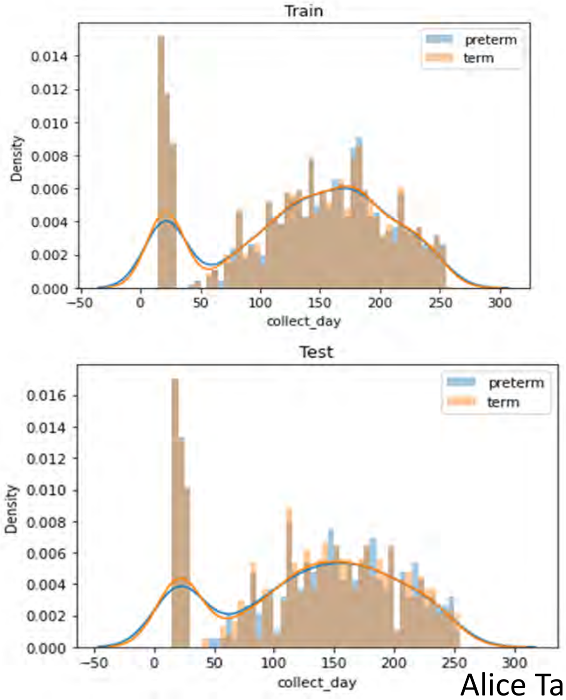
PRELIMINARY

Alice Tang

Unmatched (binsize = 5)

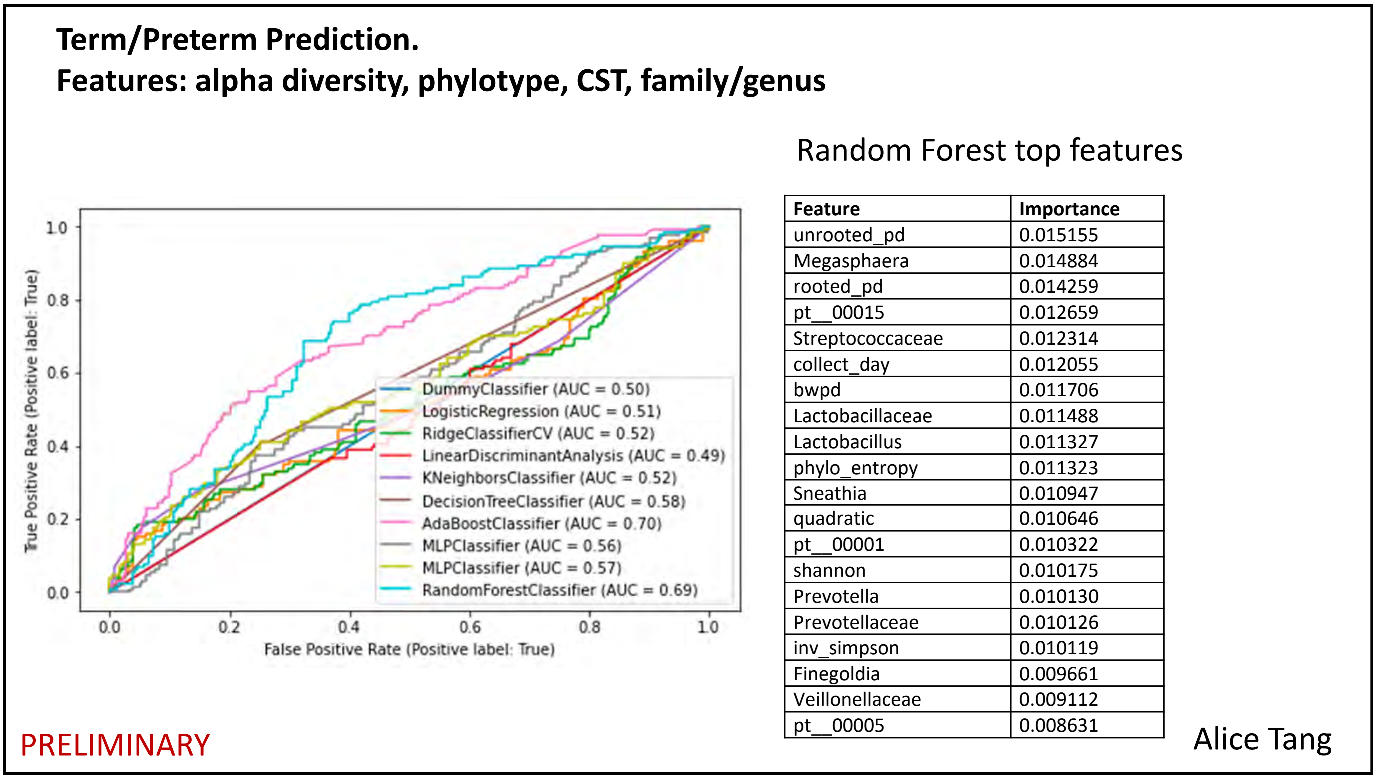


Matched (binsize = 5)



PRELIMINARY

Alice Tang



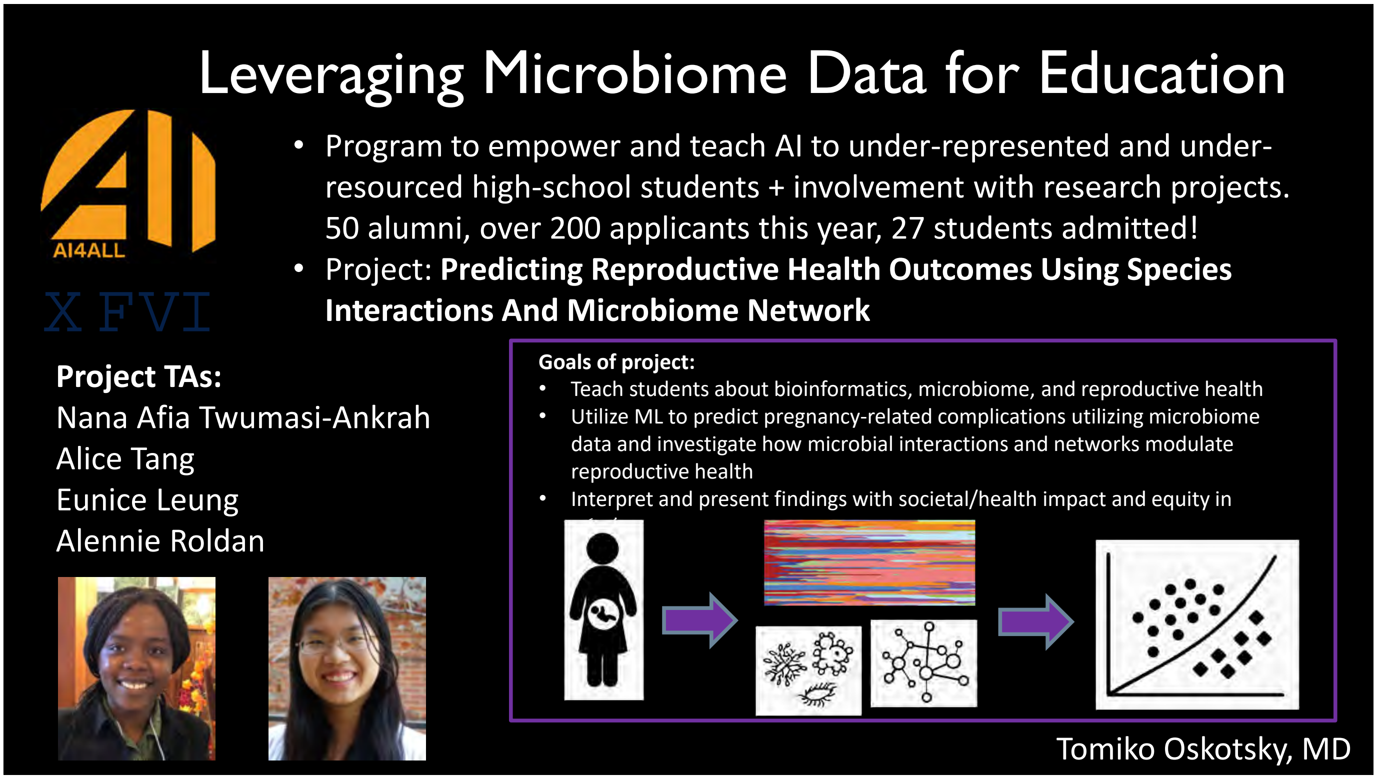
Term/Preterm Prediction.
Features: alpha diversity, phylotype, CST, family/genus

Random Forest top features

Feature	Importance
unrooted_pd	0.015155
Megasphaera	0.014884
rooted_pd	0.014259
pt__00015	0.012659
Streptococcaceae	0.012314
collect_day	0.012055
bwpd	0.011706
Lactobacillaceae	0.011488
Lactobacillus	0.011327
phylo_entropy	0.011323
Sneathia	0.010947
quadratic	0.010646
pt__00001	0.010322
shannon	0.010175
Prevotella	0.010130
Prevotellaceae	0.010126
inv_simpson	0.010119
Finegoldia	0.009661
Veillonellaceae	0.009112
pt__00005	0.008631

PRELIMINARY

Alice Tang



Leveraging Microbiome Data for Education

- Program to empower and teach AI to under-represented and under-resourced high-school students + involvement with research projects. 50 alumni, over 200 applicants this year, 27 students admitted!
- Project: **Predicting Reproductive Health Outcomes Using Species Interactions And Microbiome Network**

Project TAs:
Nana Afia Twumasi-Ankrah
Alice Tang
Eunice Leung
Alennie Roldan

Goals of project:

- Teach students about bioinformatics, microbiome, and reproductive health
- Utilize ML to predict pregnancy-related complications utilizing microbiome data and investigate how microbial interactions and networks modulate reproductive health
- Interpret and present findings with societal/health impact and equity in

Tomiko Oskotsky, MD

AI will change the world, Who will change AI?

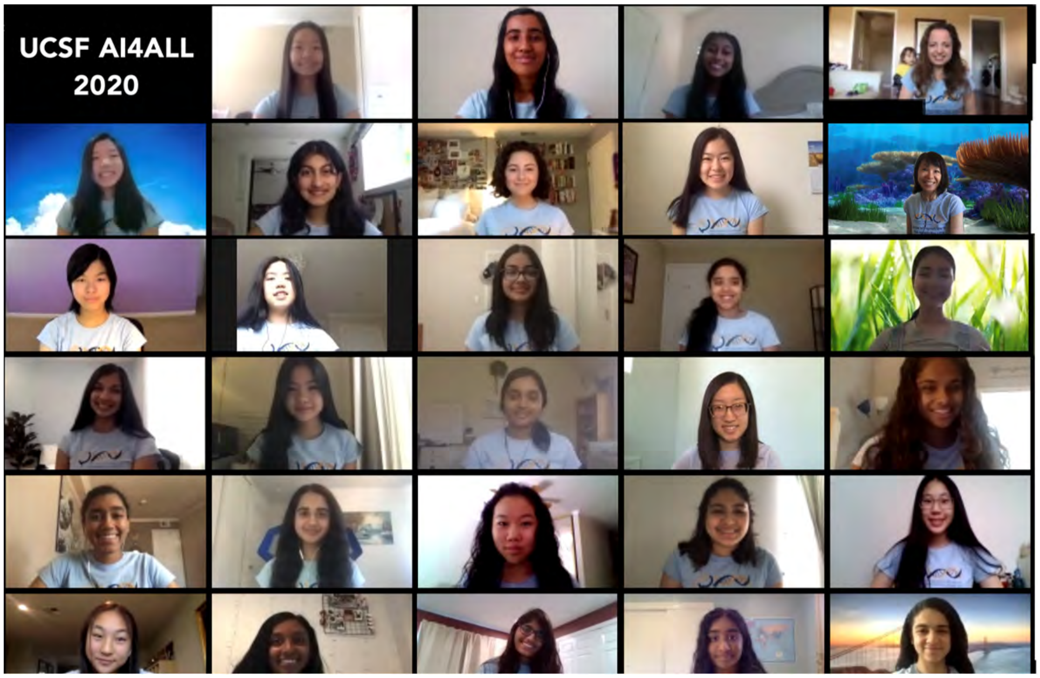
Marina Sirota, PhD
Associate Professor, Bakar Computational Health Sciences Institute, UCSF
Tomiko Oskotsky, MD
Research Scientist, Bakar Computational Health Sciences Institute, UCSF

AI4ALL

UCSF

Class of 2019

Class of 2020



Class of 2021

Over three weeks, our students…

- Learned about AI, Machine Learning, Clustering
- Learned how to code in Python
- Heard a number of faculty guest lecturers talk about their work applying AI in biomedicine
- Various Panels (Alumni, Undergrad, Grad, Industry)
- Attended a personal growth session focused on science communication
- Worked on 5 awesome research project

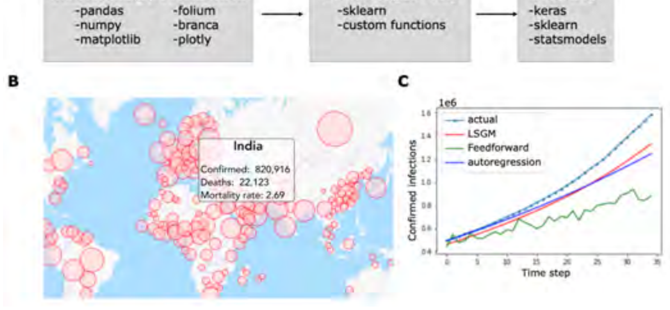
2020 Student Research Projects

- ***Project 1: AI for Global Health - AI and COVID-19 Time Series Data***
  TAs: Jon Wang, Ishan Kanungo, and Jillian Burchard
- ***Project 2: COVID-19 Protein-Protein Interactions (PPI) - AI and Proteomics***
  TAs: Snow Naing and Isha Karim
- ***Project 3: COVID-19 Chest X-ray Classification - Computer Vision***
  TAs: Ina Chen, Tianna Grant, and Janice Yang
- ***Project 4: Latent Variable Modeling of COVID-19 Host Transcriptomic Response***
  TAs: Will Connell, Ruchika Bajaj, and Claire Swadling
- ***Project 5: PredictCOVID - AI and Electronic Medical Record (EMR) Data***
  TAs: Alice Tang and Bianca Vora

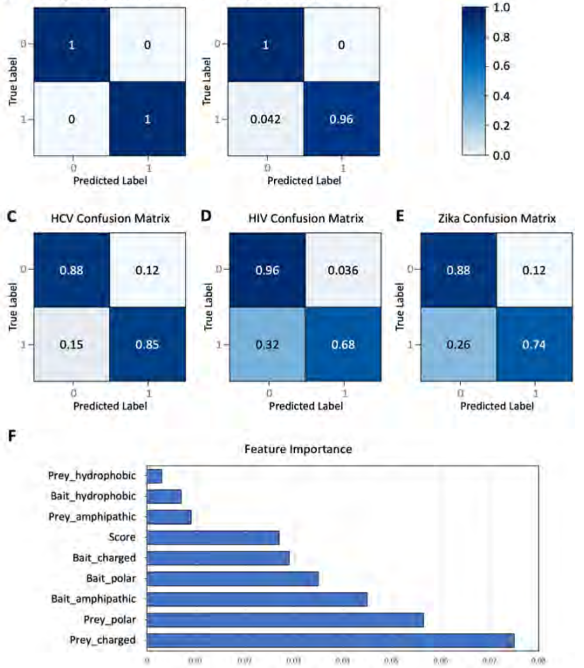
https://www.youtube.com/watch?v=uImjiHl7MDw&feature=youtu.be

0101010110100100100100 1
1110010001010101101010 1

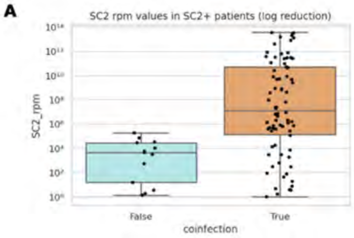
AI for Global Health



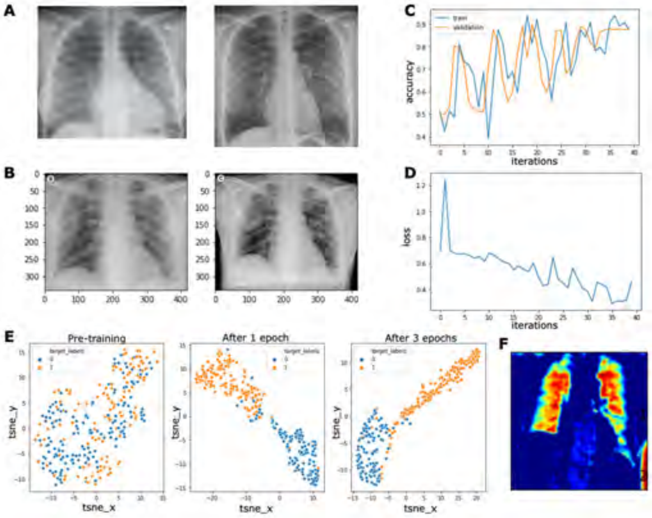
AI and Proteomics - COVID-19 Protein-Protein Interactions (PPI) predictions



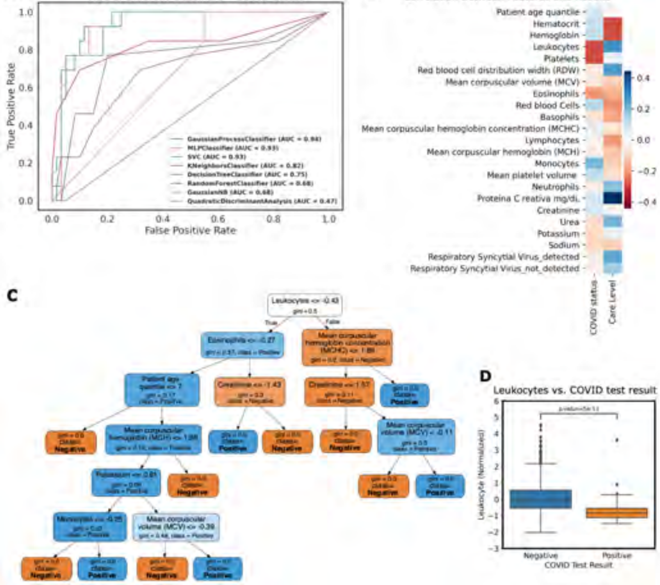
Latent Variable Modeling of COVID-19 Metagenome Transcriptomic Profiles

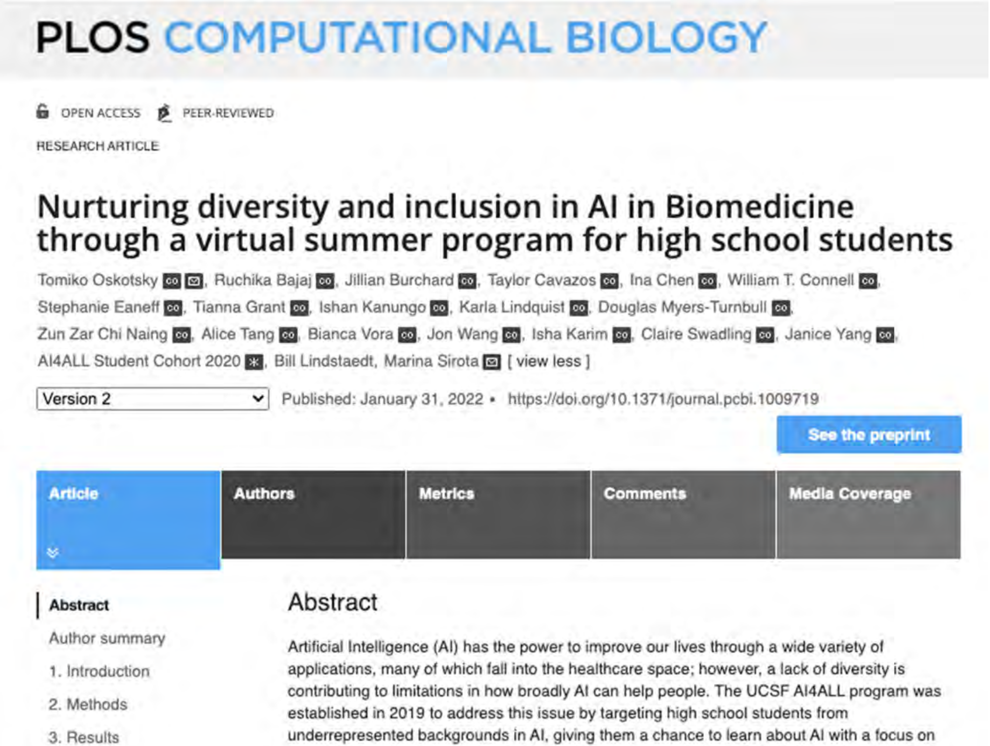


AI for Imaging Data - Computer Vision for COVID-19 Chest X-Ray Classification



PredictCOVID - AI and Electronic Medical Record (EMR) data





PLOS COMPUTATIONAL BIOLOGY

OPEN ACCESS PEER-REVIEWED

RESEARCH ARTICLE

Nurturing diversity and inclusion in AI in Biomedicine through a virtual summer program for high school students

Tomiko Oskotsky, Ruchika Bajaj, Jillian Burchard, Taylor Cavazos, Ina Chen, William T. Connell, Stephanie Eaneff, Tianna Grant, Ishan Kanungo, Karla Lindquist, Douglas Myers-Turnbull, Zun Zar Chi Naing, Alice Tang, Bianca Vora, Jon Wang, Isha Karim, Claire Swadling, Janice Yang, AI4ALL Student Cohort 2020, Bill Lindstaedt, Marina Sirota [view less]

Version 2 Published: January 31, 2022 • https://doi.org/10.1371/journal.pcbi.1009719

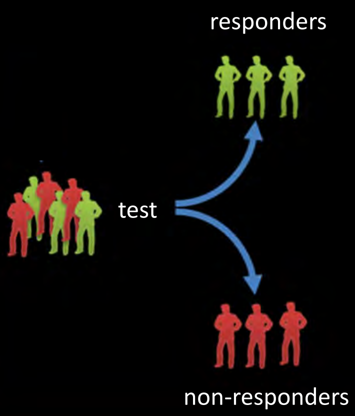
Abstract

Artificial Intelligence (AI) has the power to improve our lives through a wide variety of applications, many of which fall into the healthcare space; however, a lack of diversity is contributing to limitations in how broadly AI can help people. The UCSF AI4ALL program was established in 2019 to address this issue by targeting high school students from underrepresented backgrounds in AI, giving them a chance to learn about AI with a focus on

Summary

- Lots of (microbiome) data is available in the public domain – data sharing is vital!
- AI can be applied to build predictive models in disease
- Crowdsourcing approaches allow us to bring computational community to the data and interesting questions
- Case study: preterm birth
- Important to train the next generation of AI researchers!

Big Data → Precision Medicine


responders

test

non-responders

> *"Doctors have always recognized that every patient is unique, and doctors have always tried to tailor their treatments as best they can to individuals. You can match a blood transfusion to a blood type — that was an important discovery. What if matching a cancer cure to our genetic code was just as easy, just as standard? What if figuring out the right dose of medicine was as simple as taking our temperature?"*
>
> *- President Obama, January 30, 2015*

Acknowledgements

SIROTA LAB
Tomiko Oskotsky
Brian Le
Idit Kosti
Gaia Andreoletti
Silvia Pineda
Dmitry Rychkov
Kat Yu
Daniel Bunis
Stella Belonwu
Dimitri Abrahamsson
Alice Tang
Zachary Cutts
Yaqiao Li
Sarah Woldemariam
Alennie Rodan

UCSF
Tippi Mackenzie
Renan Sper
Scott Boyd
Sandra Nielson
Tony Capra
Abin Abraham
Tracey Woodruff
Amy Padula
Susan Lynch
Connie Ha

STANFORD
David Stevenson
Ron Wong
Sota Iwatani
Cele Quaintance
Gary Shaw
Ivana Maric
Dan DiGiulio
David Relman
Ben Callahan
Nima Aghaeepour
Brice Gaudilliere

PERATON
Elizabeth Thomson
Patrick Dunn
Kate Hypes
Ruth Monteiro
Zhiping Gu
Bryan Walters

DREAM
Gustavo Stolovitzky
James Costello
Adi Tarca
Roberto Romero
Jonathan Golob
Jacques Ravel

MOD TDC
Atul Butte
Sam Parry
Rita Leite
Aubrey Bailey
Louis Muglia
Ge Zhang
Joanne Chappell
Rebecca Anderson
Jessica Chubiz
Sarah England
Neal Sondheimer
Antonis Rokas
Phil Bennet
Sing Sing Way

Funding Sources
March of Dimes, NLM, NIA, NIAMS, NICHD, PTBi, Pfizer, BWF

Tech and Admin Support: Boris Oskotsky, Edna Rodas

69

70

Thanks!

marina.sirota@ucsf.edu

http://sirotalab.ucsf.edu

We are hiring!