



Center for Bioinformatics and Computational Biology

Using Secure Computation for Statistical Analysis of Quantitative Genomic Assay Data

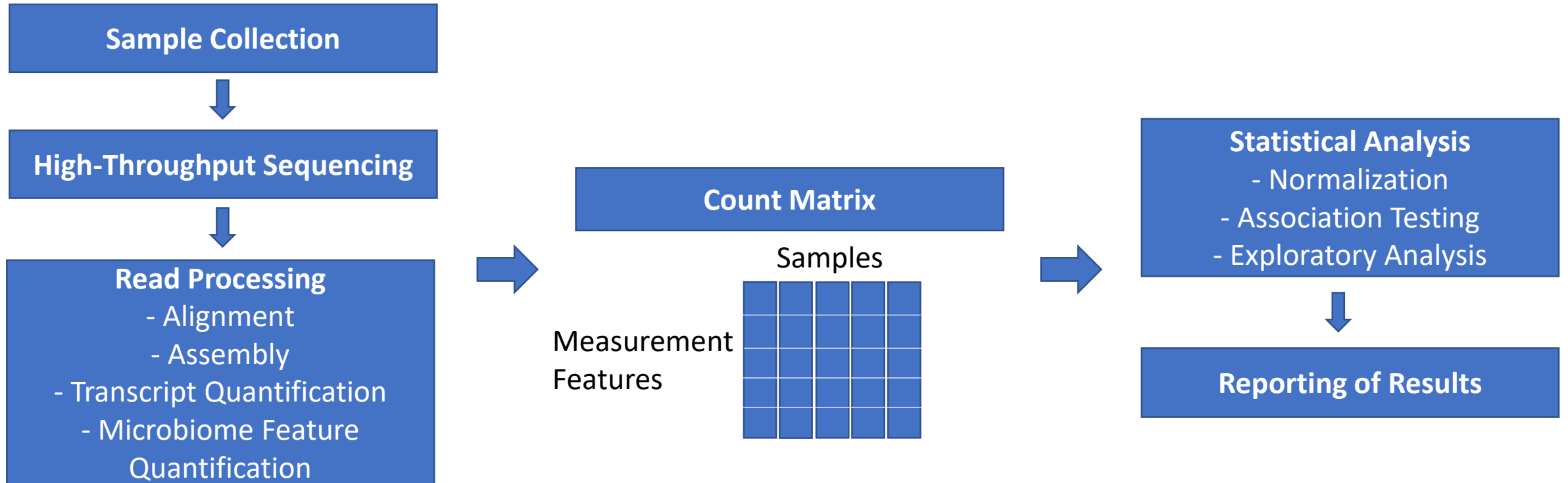
Justin Wagner

Ph.D. Candidate

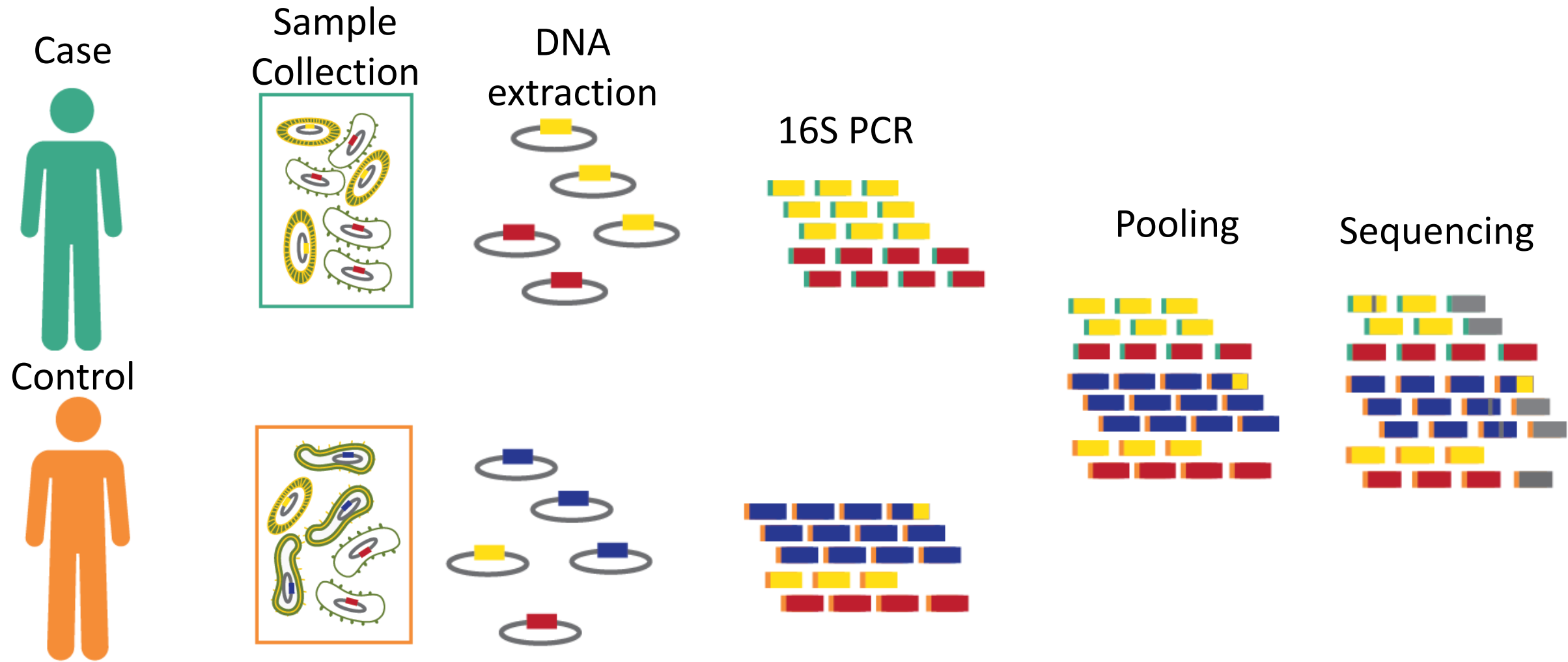
University of Maryland, College Park

Advisor: Hector Corrada Bravo

Genomic Assay Analysis Pipeline



Microbiome Sequencing



Used with Permission from Nathan Olson

Microbiome Sequencing (Continued)

Sequencing



Feature Inference

		
	0	8
	3	3
	3	3
	1	0
	0	1

Feature Annotation

		
	0	8
	3	3
	3	3
	1	0
	0	1

Features with
Hierarchy

Count Table Generation

Metagenomic Association Studies

- Presence/Absence
 - Tests if presence or absence of specific bacteria is associated phenotype
- Differential Abundance
 - Tests for difference in mean abundance of bacteria between phenotypes
- Alpha Diversity
 - Test for association with diversity between phenotypes

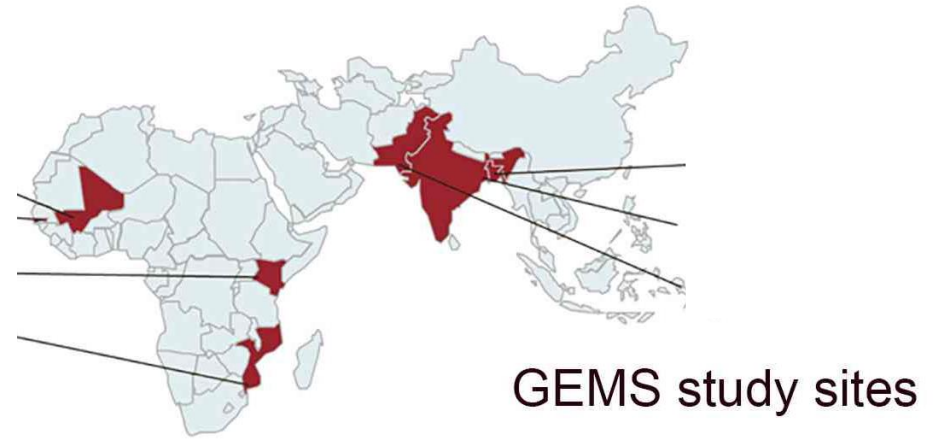
Bacteria Counts



		
	0	8
	3	3
	3	3
	1	0
	0	1

Integrative Analysis

- Large epidemiological projects are vital to gain insights of microbiome between populations
- Gathered across institutional domains and countries
- Examples
 - Global Enteric Multicenter Study
 - Human Microbiome Project

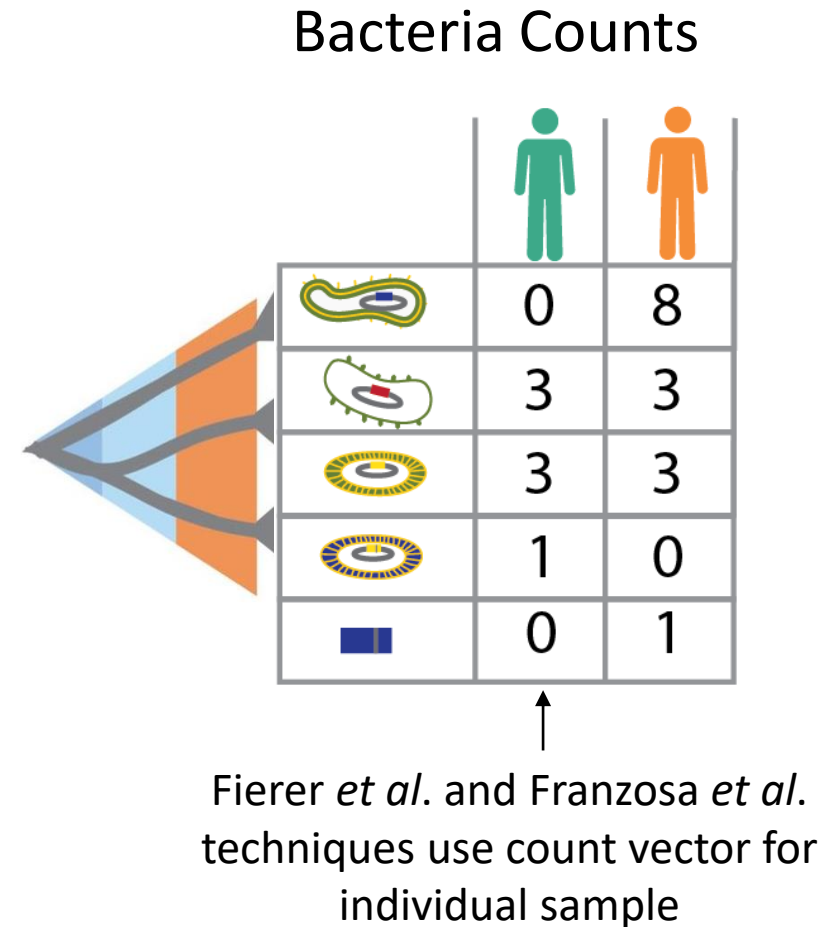


<http://www.medschool.umaryland.edu/CVD/Projects/Global-Enteric-Multicenter-Study-GEMS/>

<https://www.hmpdacc.org/>

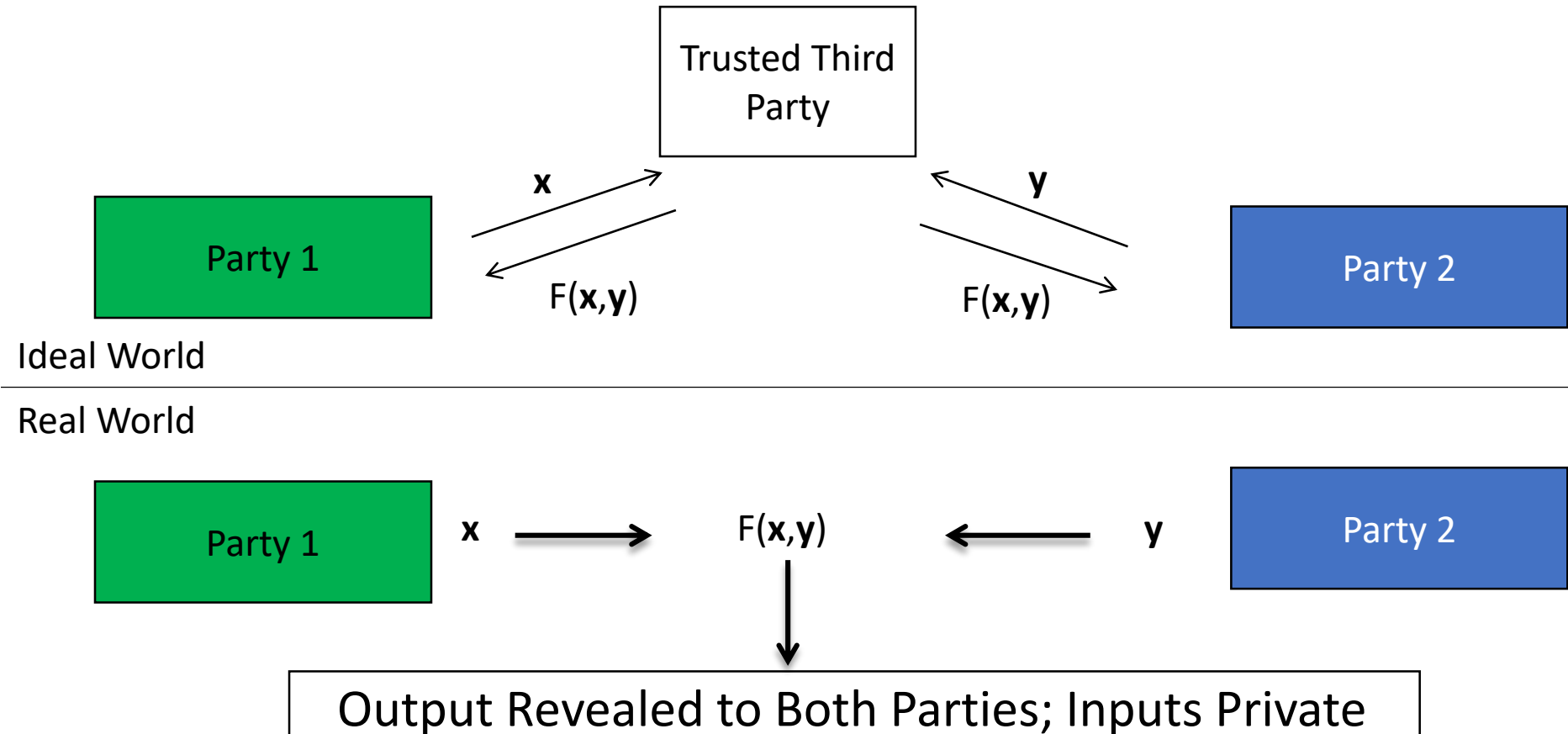
Microbiome-Based Identification

- Fierer *et al.* 2010
 - Forensic identification of objects touched by an individual
- Franzosa *et al.* 2015
 - Unique identification of ~30% of individuals in Human Microbiome Project at later time points
- Human Read Contamination



Secure Multi-Party Computation

- Two parties have input $\mathbf{x}$, $\mathbf{y}$, and want to compute $F(\mathbf{x}, \mathbf{y})$

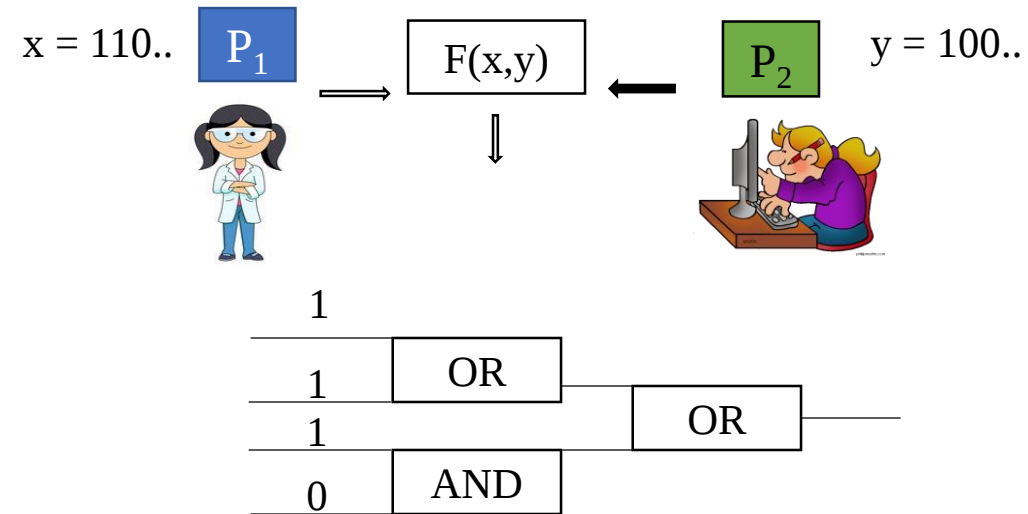


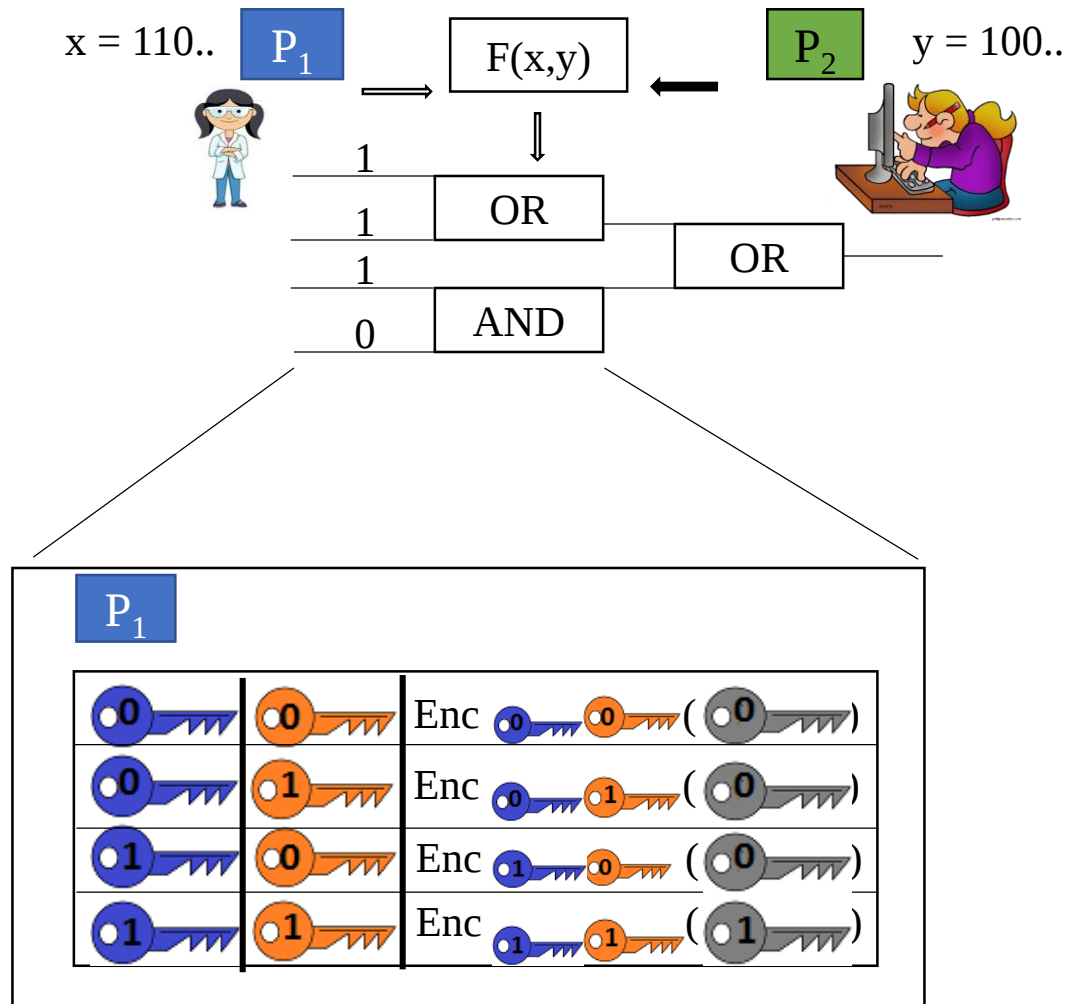
Yao's Garbled Circuits Protocol

- Parties agree on representation of function
- Input values are
 - **encrypted and exchanged**
 - such that, **neither party learns anything about the inputs**
 - **except what can be inferred by the output**

Garbled Circuits Protocol

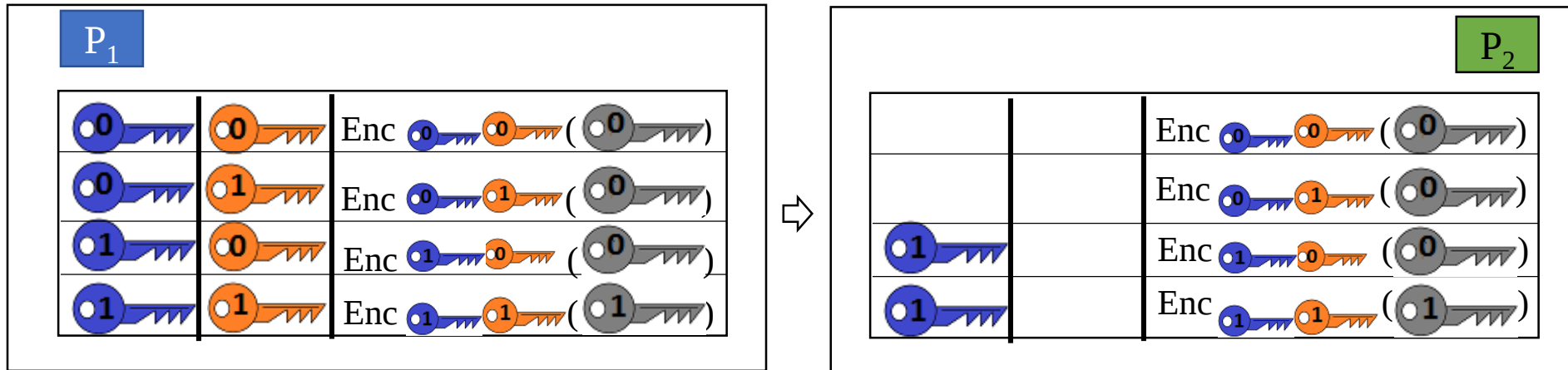
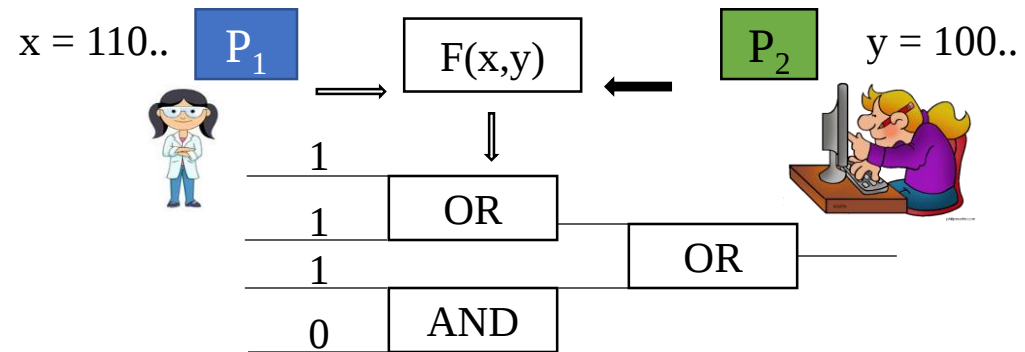
Parties P_1 and P_2 agree on a function $F(x,y)$ to compute





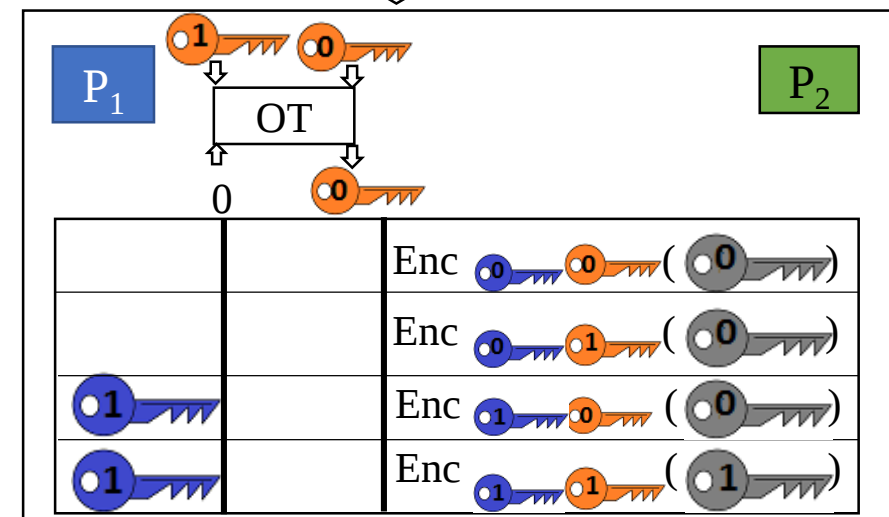
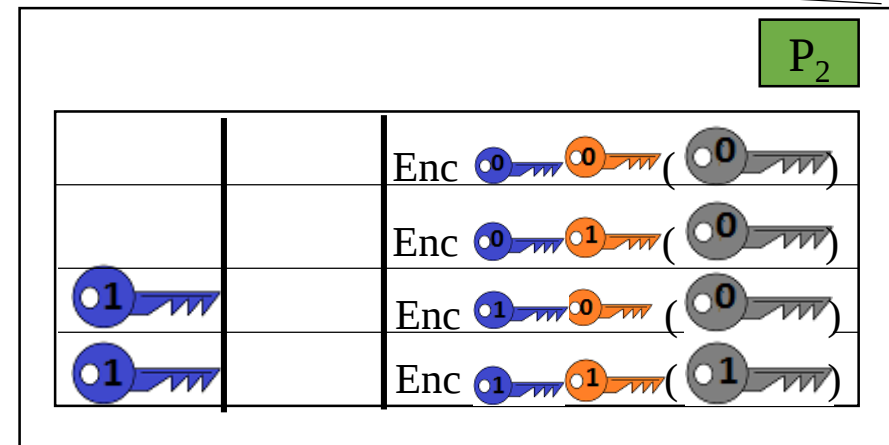
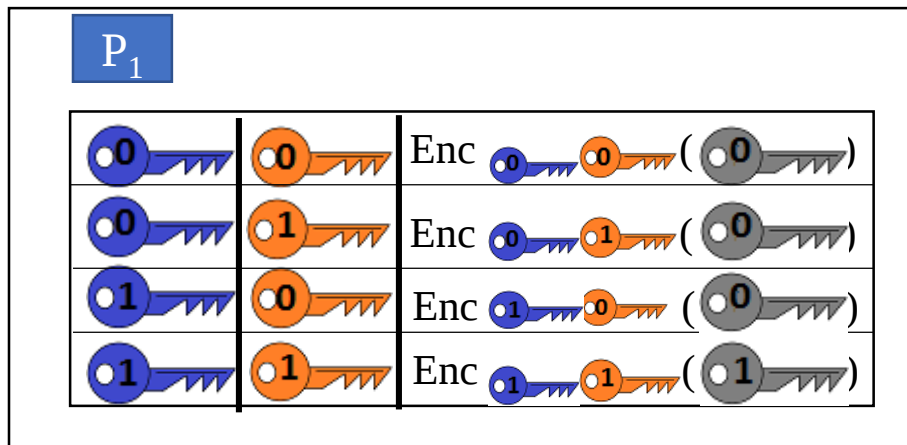
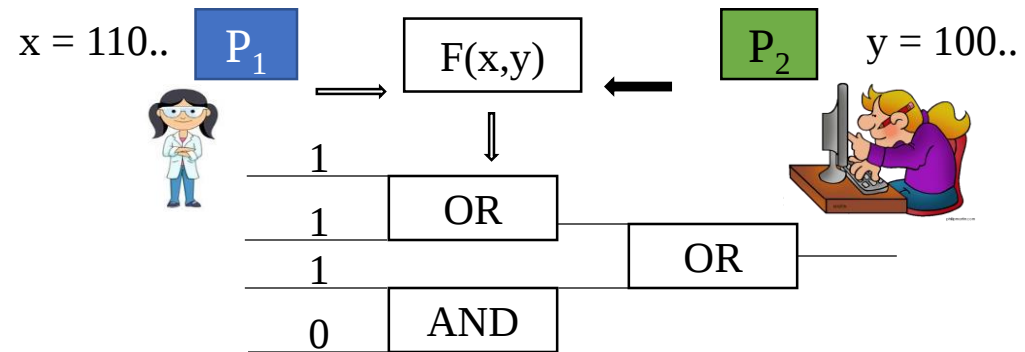
P_1 garbles the circuit:

- Maps each input bit to a random string which is as an encryption key
- Encrypts each output gate using appropriate keys



P2 receives garbled circuit:

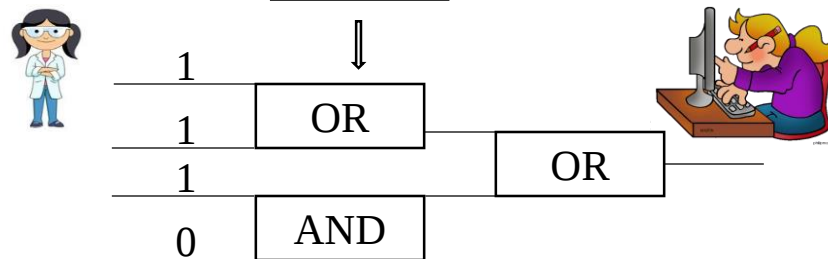
- P1 sends encrypted truth table entries
- P1 also sends encryption keys mapping to its input



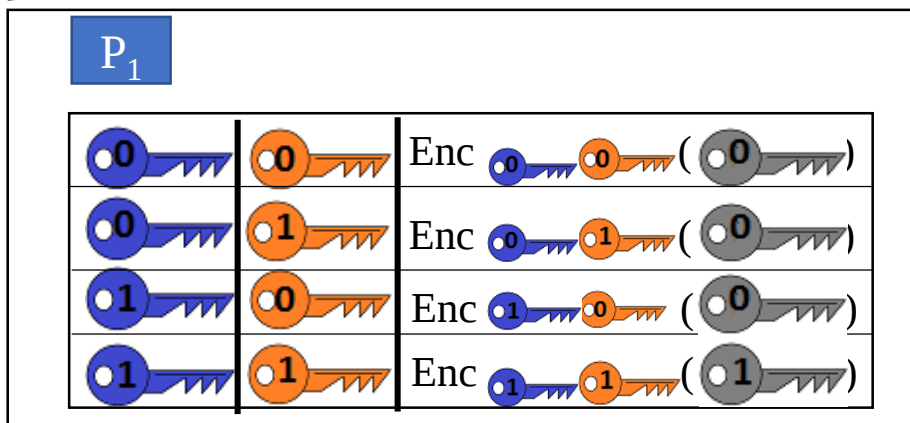
P1 and P2 run Oblivious Transfer:

- P2 receives key corresponding to its input
- P1 does not learn P2's input

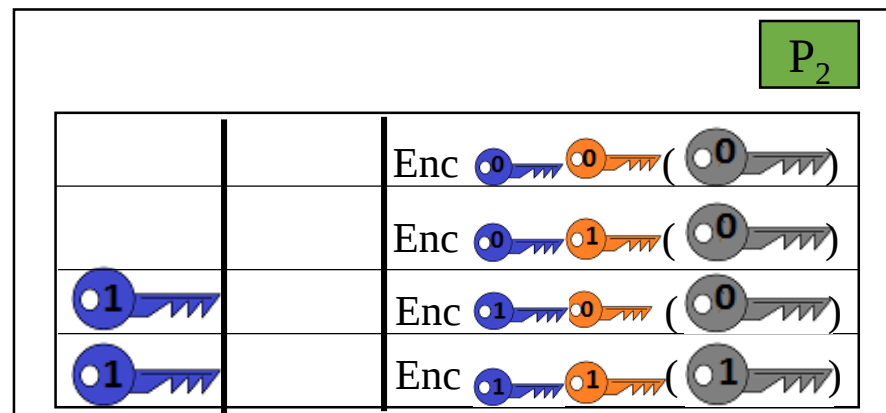
A. $x = 110..$ $P_1 \Rightarrow F(x,y) \leftarrow P_2$ $y = 100..$



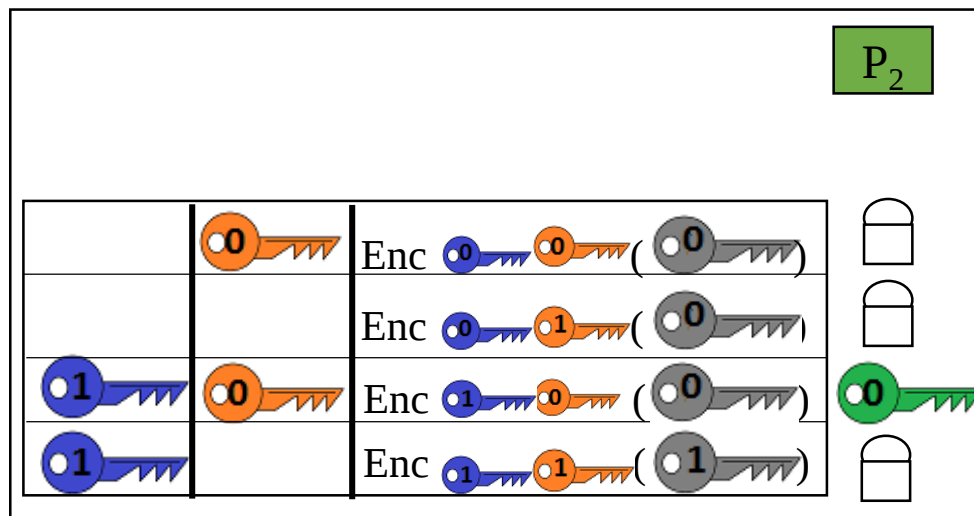
B.



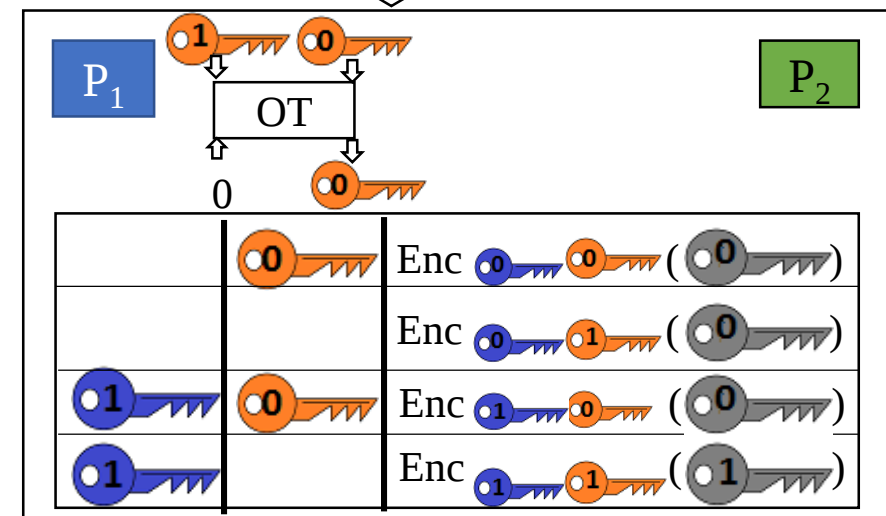
C.



E.



D.

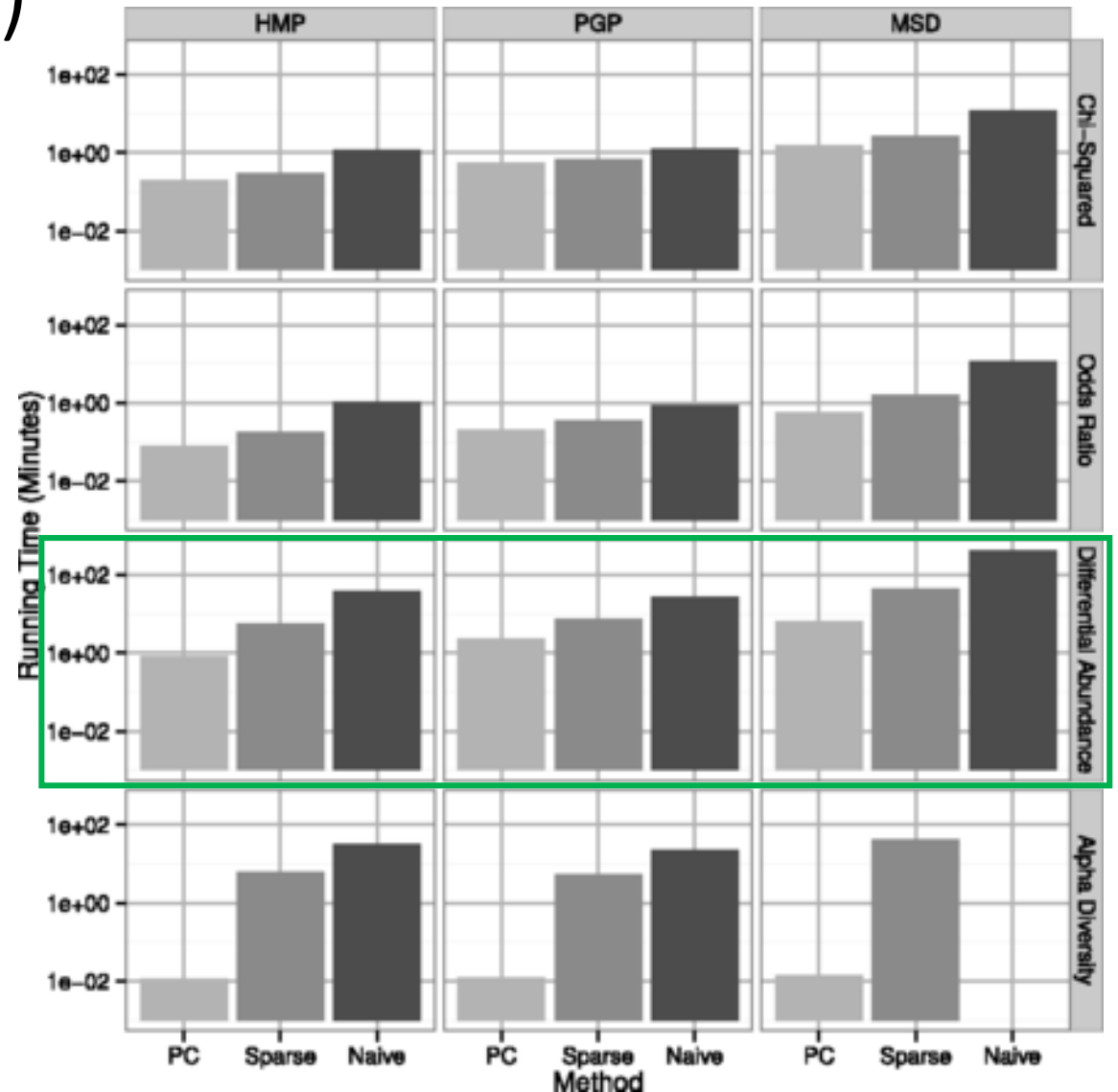


Implementation

- Used FlexSC library which is part of OblivM project
- Coded analysis functions
 - Chi-square
 - Odds Ratio
 - Two-sample t-test
 - Alpha diversity
- Ran analysis on three datasets
 - Moderate to Severe Diarrheal Disease (754 Features, 992 Samples)
 - Personal Genome Project (277 Features, 168 Samples)
 - Human Microbiome Project (97 Features, 694 Samples)

Running Time (Minutes)

- Split datasets up with mix of case and controls for each party
- Ran computation between two EC2 instances
 - r3.2xLarge instances with 2.5 GHz processors and 61 GB RAM
- Sparse matrix implementation
Two-sample t-test ran in under one hour for largest dataset



Network Usage (MB)

- Large data transfer overhead
- Transferred data within an AWS EC2 Region

Sparse From Garbler	HMP	PGP	MSD
Alpha Diversity	5862.87	5305.95	41568.3
Chi Square	203.64	543.32	1644.5
Odds Ratio	79.91	194.89	718.15
Differential Abundance	5055.84	6593.47	41020.37

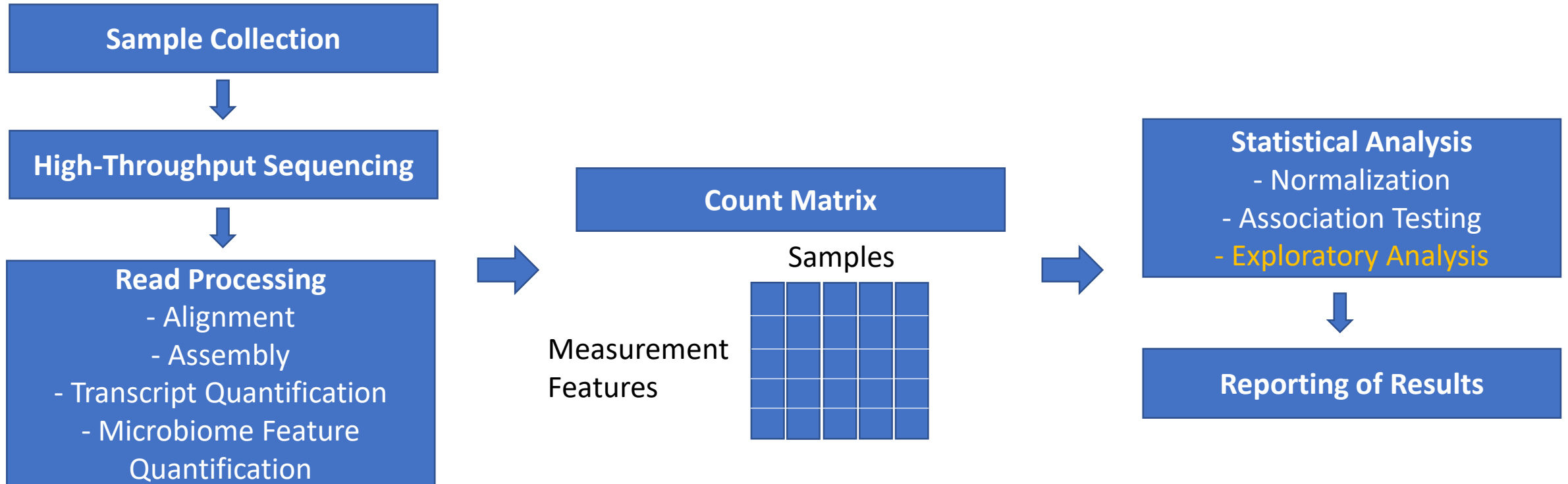
Sparse From Evaluator	HMP	PGP	MSD
Alpha Diversity	5.87	5.22	43.78
Chi Square	15.78	17.07	145.2
Odds Ratio	15.78	17.07	145.2
Differential Abundance	18.26	19.84	167.9

Accuracy

- Normalized Mean Squared Error: $\|x-y\|^2/\|x\|^2$ with x as the value output by R and y the value from our implementation

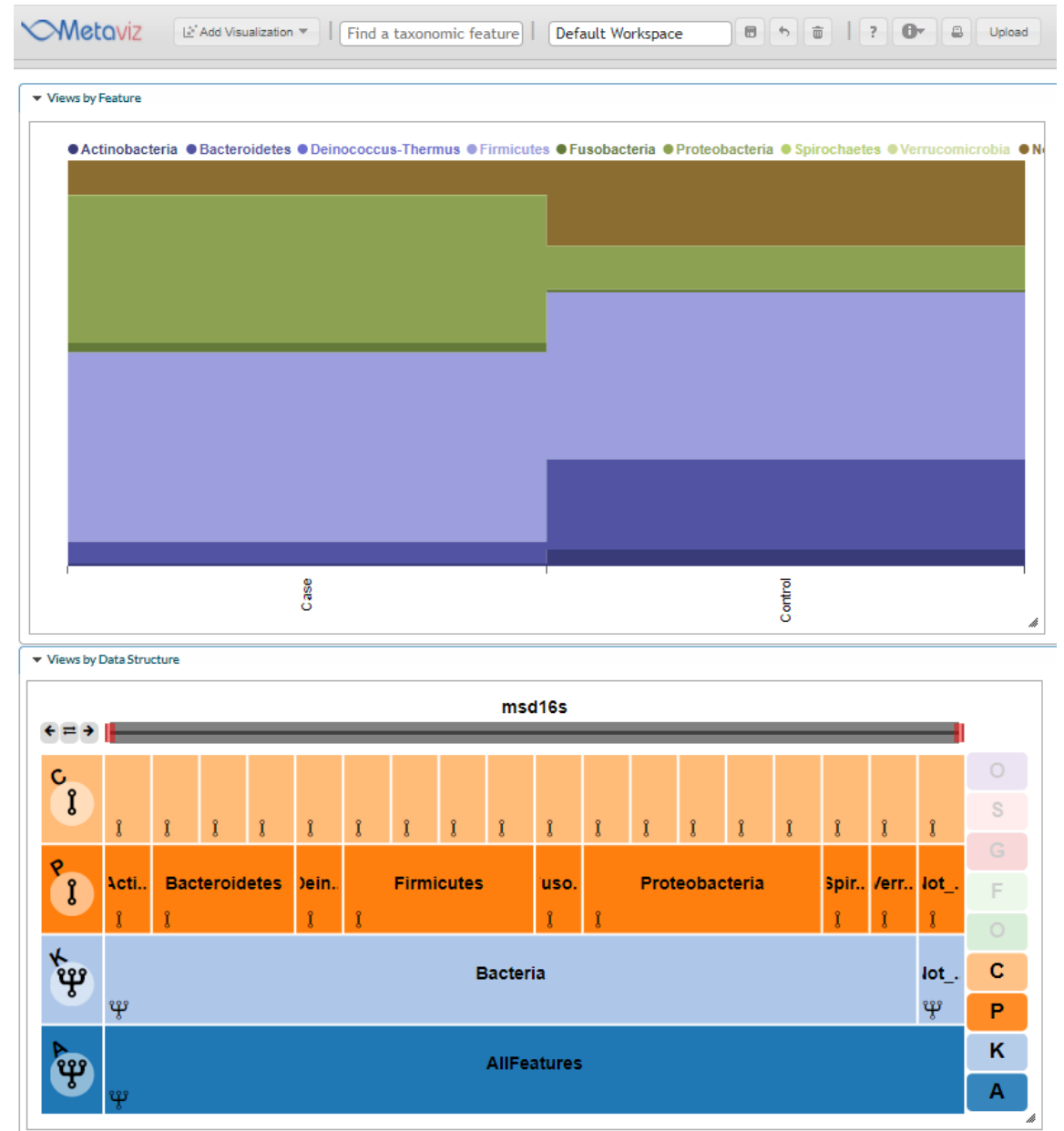
	PGP	HMP	MSD
Chi-square statistic	7.84e-07	7.48e-06	7.02e-08
Chi-square P-value	2.00e-07	2.14e-06	9.72e-08
Odds Ratio	1.60e-13	5.42e-13	2.44e-13
Differential abundance			
t-statistic	0.023	0.0017	0.0012
Differential abundance			
Degrees of Freedom	2.7e-4	2.5e-4	0.0028
Differential abundance			
P-value	0.0024	0.0026	0.0011

Proposed Solutions for Other Pipeline Components

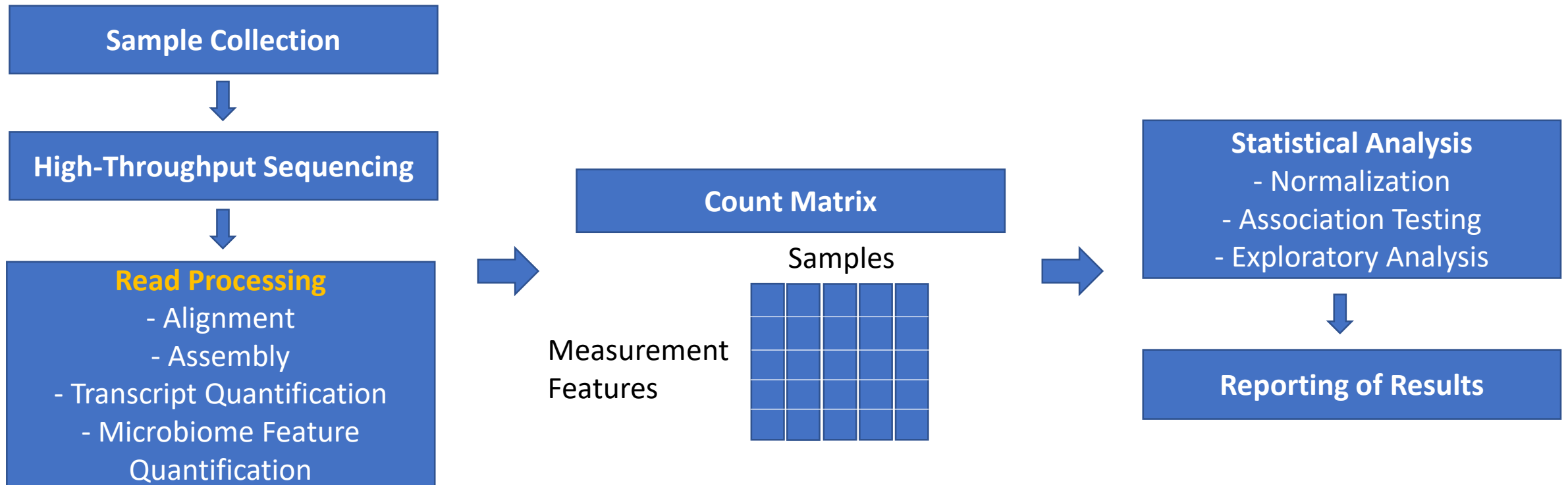


Visualization

- Community Proportion with Samples Grouped By Phenotype
- Visualized as a Stacked Bar Plot
- Proportion could be calculated using secure computation or differential privacy

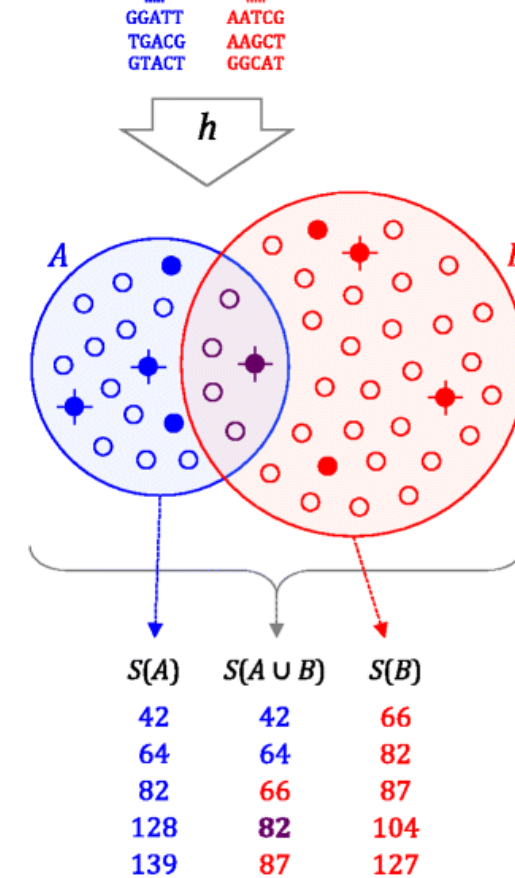


Proposed Solutions for Other Pipeline Components



Jaccard Index Estimate using MinHash

- Probabilistic technique
- Used in MASH distance measure and MHAP for assembly
- Would work with Private Set Intersection



$$J(A, B) = \frac{|A \cap B|}{|A \cup B|} \approx \frac{|S(A \cup B) \cap S(A) \cap S(B)|}{|S(A \cup B)|}$$

Acknowledgements

Corrada Bravo Lab

UMD Center for Bioinformatics and
Computational Biology

Joseph N. Paulson

Xiao Wang

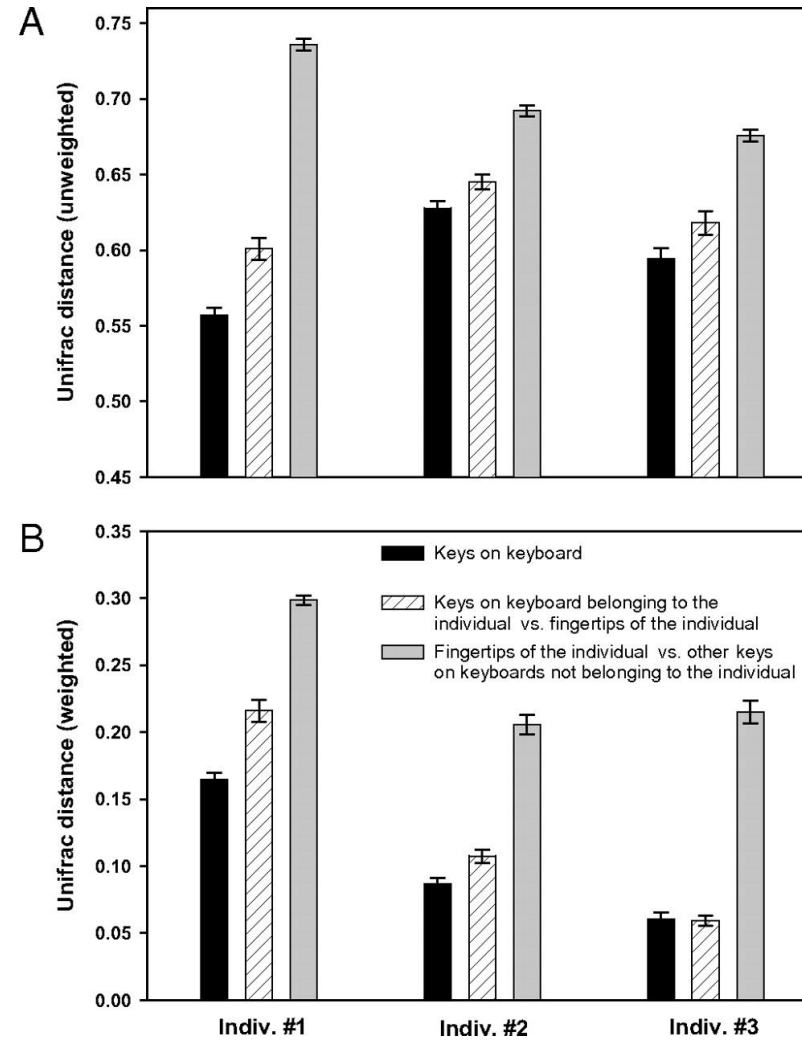


Thank You

- Questions?

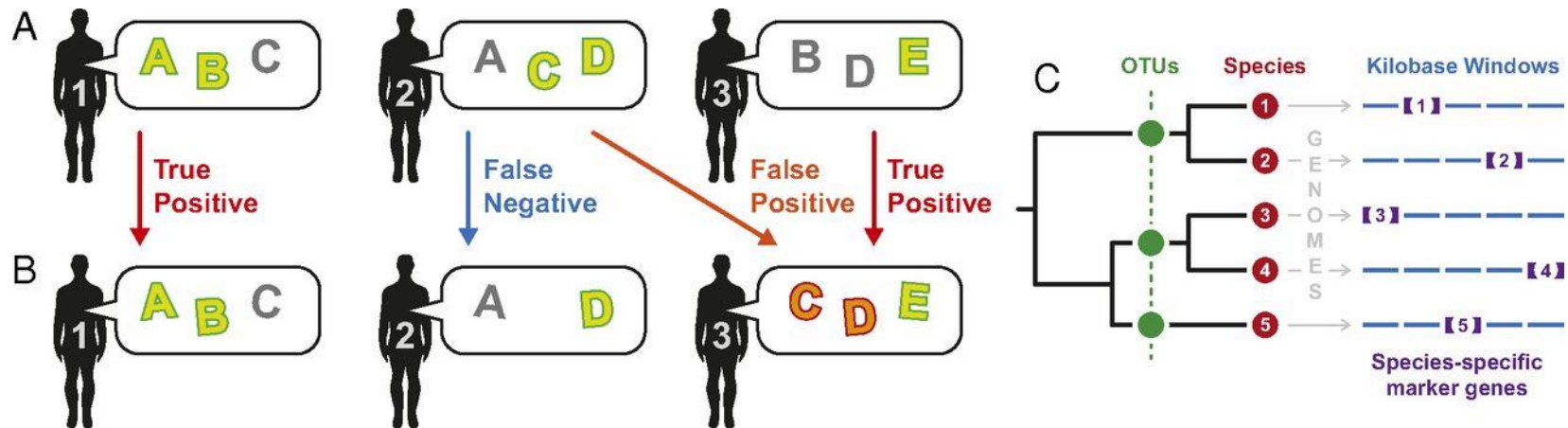
Backup Slides

Bacterial community distances between keyboard keys and fingertips.



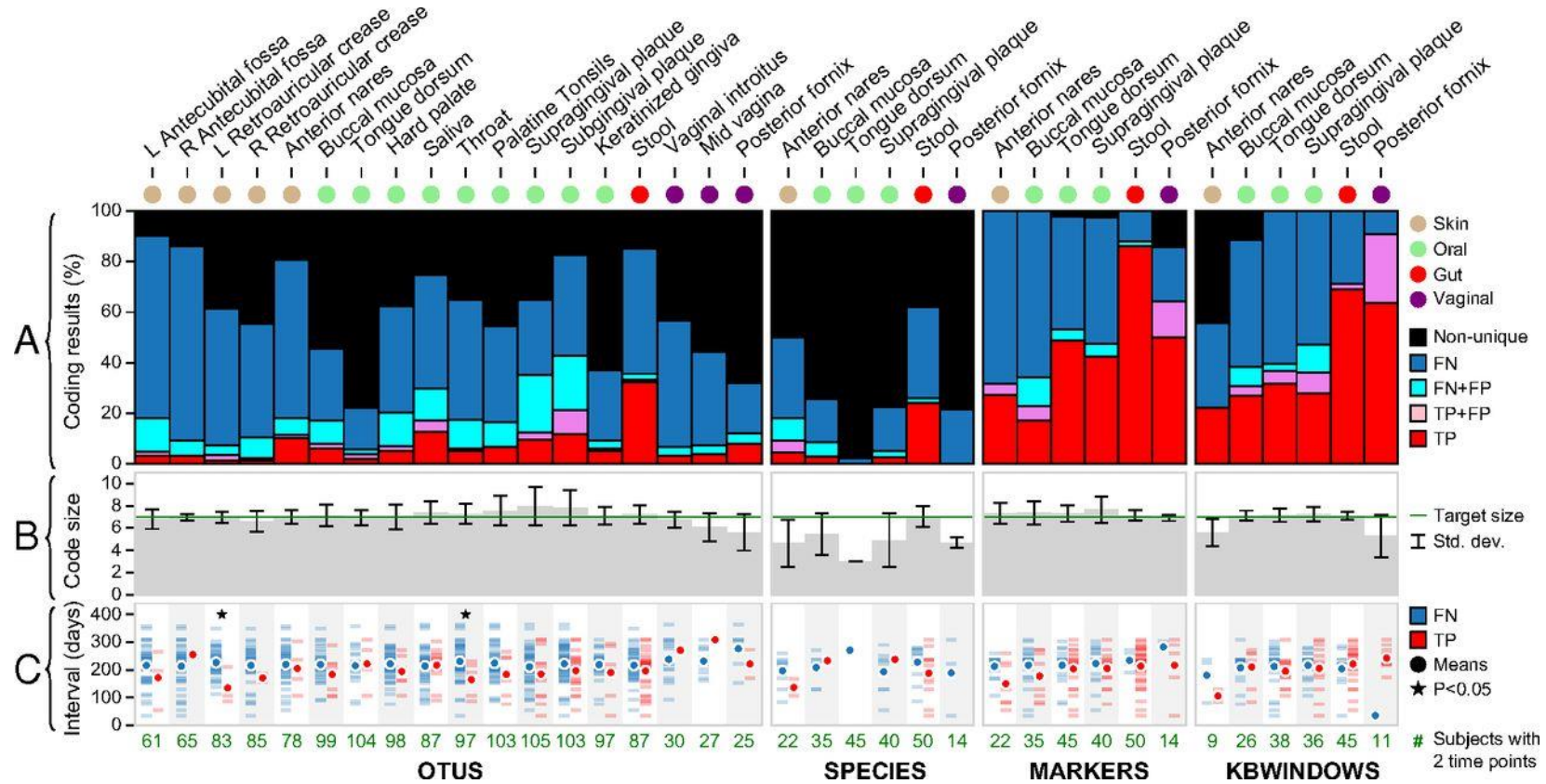
Noah Fierer et al. PNAS 2010;107:6477-6481

Metagenomic codes (overview).



Eric A. Franzosa et al. PNAS 2015;112:E2930-E2938

Temporal stability of metagenomic codes.



Eric A. Franzosa et al. PNAS 2015;112:E2930-E2938

Secure Microbiome Analysis Implementation

- Chi-squared test

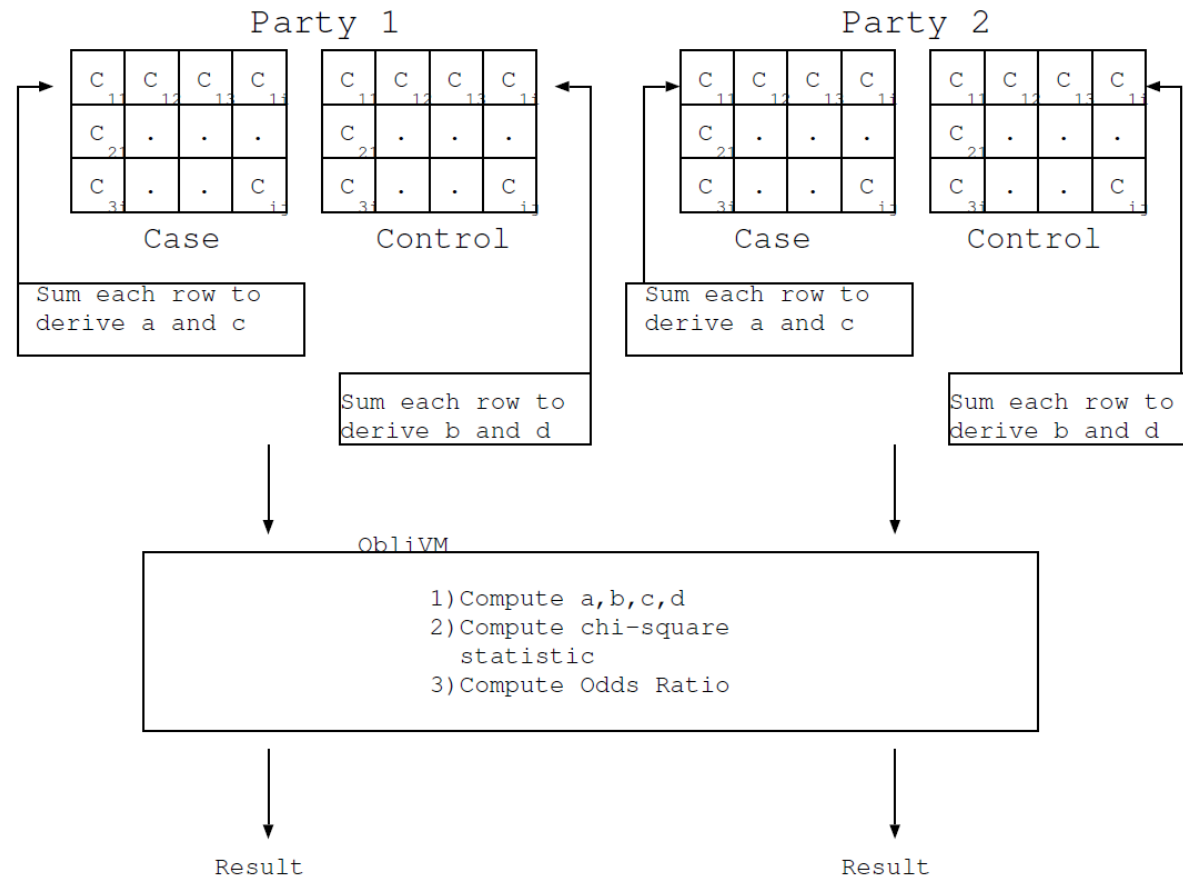
		Samples			
		Case		Control	
Features		c_{11}	c_{12}	..	c_{1i}
		c_{21}	..	..	..
		c_{31}	..	..	..
		..	..	..	..
		c_{i1}	..	..	c_{ij}

	Feature P resent	Feature A bsent
C ase	a	c
C ontrol	b	d

$$\chi^2 = \frac{(a + b + c + d)(bc - ad)^2}{(a + b)(a + c)(b + d)(c + d)}$$

$$OR = \frac{ad}{cb}$$

Secure Microbiome Analysis Implementation



Secure Microbiome Analysis Implementation

- Two sample t-test

		Samples			
		Case		Control	
Features		c_{11}	c_{12}	..	c_{1i}
		c_{21}	..	..	..
		c_{31}	..	..	..
		..	..	..	..
		c_{i1}	..	..	c_{ij}

Mean Abundance

$$\bar{x}_i = \frac{1}{n} \sum_1^j M_{ij}$$

Variance

$$s_i^2 = \frac{1}{n-1} \sum_1^j (M_{ij} - \bar{x}_i)^2$$

t-statistic

$$t = \frac{\bar{x}_{Feature_i case} - \bar{x}_{Feature_i control}}{\sqrt{\left(\frac{s_{Feature_i case}^2}{n_{case}}\right) + \left(\frac{s_{Feature_i control}^2}{n_{control}}\right)}}$$

Secure Microbiome Analysis Implementation

- Two sample t-test

		Samples			
		Case		Control	
Features		c_{11}	c_{12}	..	c_{1i}
		c_{21}	..	..	..
		c_{31}	..	..	..
		..	..	..	..
		c_{i1}	..	..	c_{ij}

Mean Abundance

$$\bar{x}_{OTU_{i,k}} = \frac{1}{n_{k_1} + n_{k_2}} \left(\sum_j^{n_{k_1}} M_{ij} + \sum_j^{n_{k_2}} M_{ij} \right)$$

Variance

$$s_{OTU_{i,k}}^2 = \frac{\sum_j^{n_{k_1}} M_{ij}^2 + \sum_j^{n_{k_2}} M_{ij}^2 + \frac{(\sum_j^{n_{k_1}} M_{ij} + \sum_j^{n_{k_2}} M_{ij})^2}{n_{k_1} + n_{k_2}}}{n_{k_1} + n_{k_2}}$$

Secure Microbiome Analysis Implementation

- Alpha Diversity

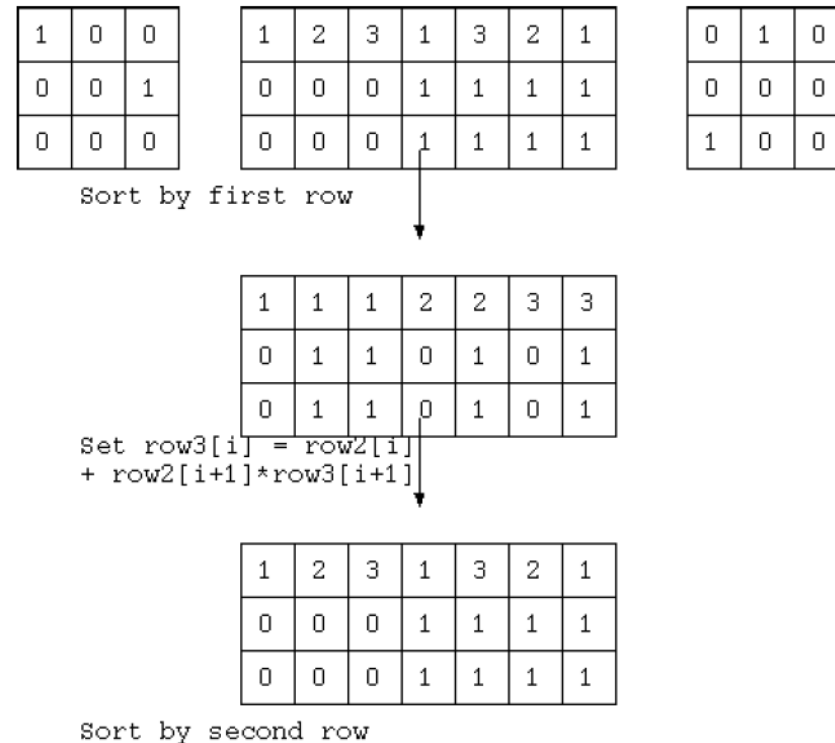
		Samples			
		Case		Control	
Features		c_{11}	c_{12}	..	c_{1i}
		c_{21}	..	..	..
		c_{31}	..	..	..
		..	..	..	..
		c_{i1}	..	..	c_{ij}

Simpson's index:
$$D = \frac{\sum n(n-1)}{N(N-1)}$$

n is number of feature counts for feature i

N is total number of counts observed in a sample

Secure Microbiome Analysis Implementation



Third row is used as a Counter for
Chi-Square Test and Odds Ratio

Add a fourth row to compute sum of
squares for t-test