ONE CODEX

A platform for highly accurate, reproducible metagenomics

MetaSUB Meeting – New York Genome Center

June 20, 2015

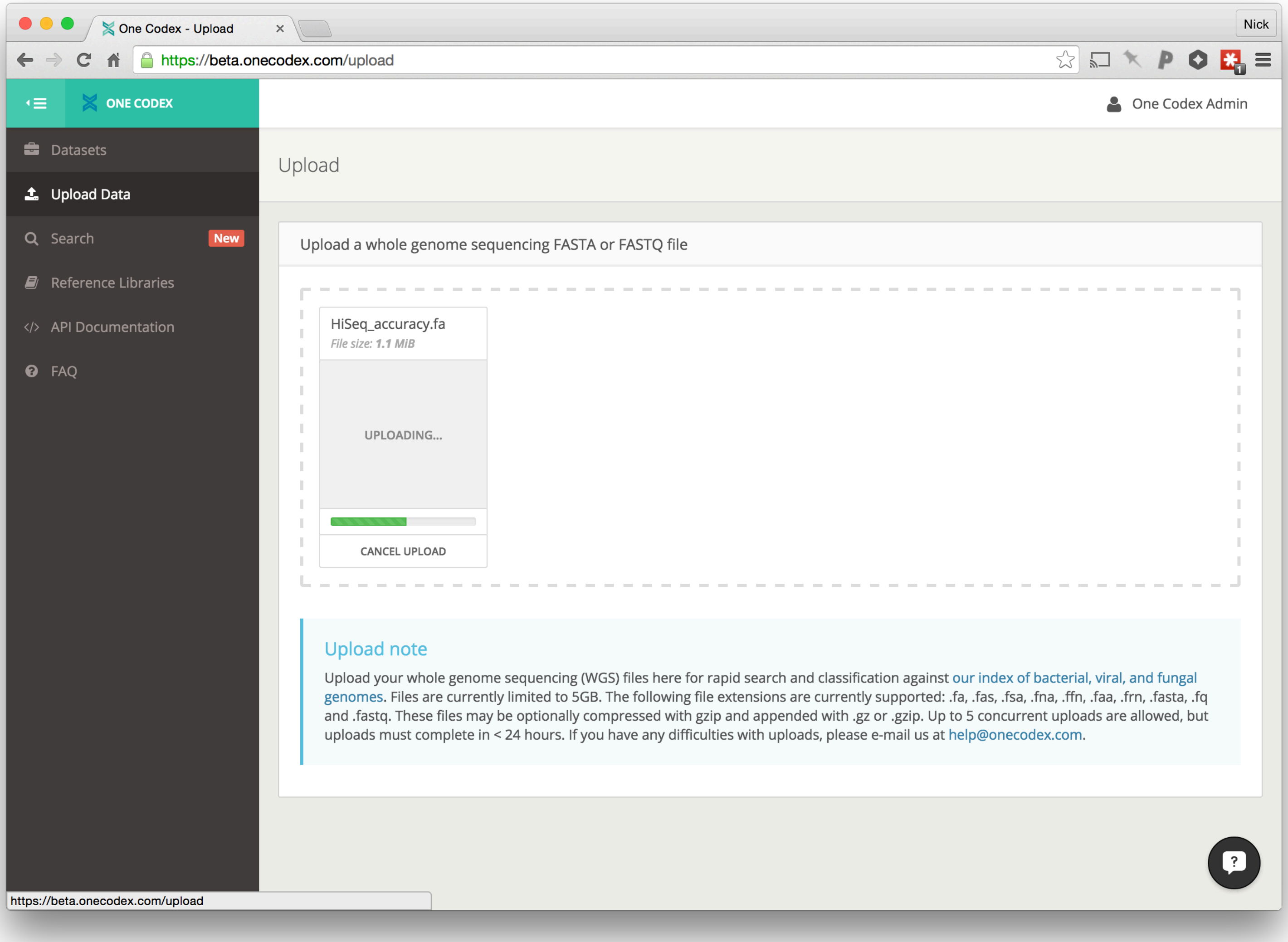
Quick overview

- One Codex
  - – Microbial genomics software and data platform
  - – Based in San Francisco

- How can we make better applied genomics software?
  - – Usable + intuitive
  - – Open + extensible (e.g., APIs)
  - – Move beyond analyzing a single sample (enable "databanking")

Towards applied microbial genomics

Technology should "just work"…



Upload a whole genome sequencing FASTA or FASTQ file

HiSeq_accuracy.fa

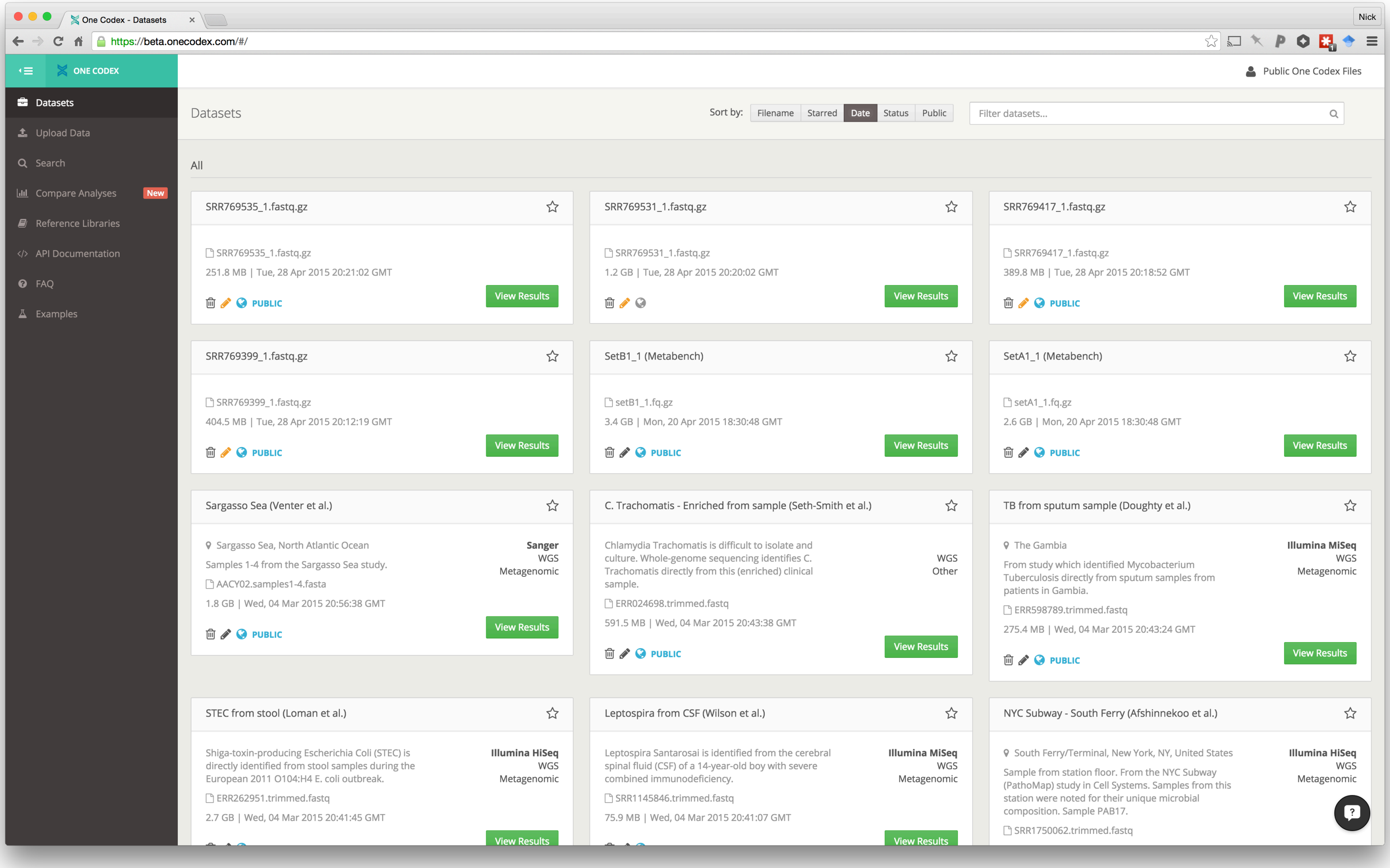
File size: ***1.1 MiB***

Upload note

Upload your whole genome sequencing (WGS) files here for rapid search and classification against our index of bacterial, viral, and fungal genomes. Files are currently limited to 5GB. The following file extensions are currently supported: .fa, .fas, .fsa, .fna, .ffn, .faa, .frn, .fasta, .fq and .fastq. These files may be optionally compressed with gzip and appended with .gz or .gzip. Up to 5 concurrent uploads are allowed, but uploads must complete in < 24 hours. If you have any difficulties with uploads, please e-mail us at help@onecodex.com.

Towards applied microbial genomics

… if we're going to sequence everything, we should be able to easily index, explore, and compare samples



Datasets

SRR769535_1.fastq.gz
SRR769535_1.fastq.gz
251.8 MB | Tue, 28 Apr 2015 20:21:02 GMT

SRR769531_1.fastq.gz
SRR769531_1.fastq.gz
1.2 GB | Tue, 28 Apr 2015 20:20:02 GMT

SRR769417_1.fastq.gz
SRR769417_1.fastq.gz
389.8 MB | Tue, 28 Apr 2015 20:18:52 GMT

SRR769399_1.fastq.gz
SRR769399_1.fastq.gz
404.5 MB | Tue, 28 Apr 2015 20:12:19 GMT

SetB1_1 (Metabench)
setB1_1.fq.gz
3.4 GB | Mon, 20 Apr 2015 18:30:48 GMT

SetA1_1 (Metabench)
setA1_1.fq.gz
2.6 GB | Mon, 20 Apr 2015 18:30:48 GMT

Sargasso Sea (Venter et al.)
Sargasso Sea, North Atlantic Ocean
Samples 1-4 from the Sargasso Sea study.
AACY02.samples1-4.fasta
1.8 GB | Wed, 04 Mar 2015 20:56:38 GMT
Sanger
WGS
Metagenomic

C. Trachomatis - Enriched from sample (Seth-Smith et al.)
Chlamydia Trachomatis is difficult to isolate and culture. Whole-genome sequencing identifies C. Trachomatis directly from this (enriched) clinical sample.
ERR024698.trimmed.fastq
591.5 MB | Wed, 04 Mar 2015 20:43:38 GMT
WGS
Other

TB from sputum sample (Doughty et al.)
The Gambia
From study which identified Mycobacterium Tuberculosis directly from sputum samples from patients in Gambia.
ERR598789.trimmed.fastq
275.4 MB | Wed, 04 Mar 2015 20:43:24 GMT
Illumina MiSeq
WGS
Metagenomic

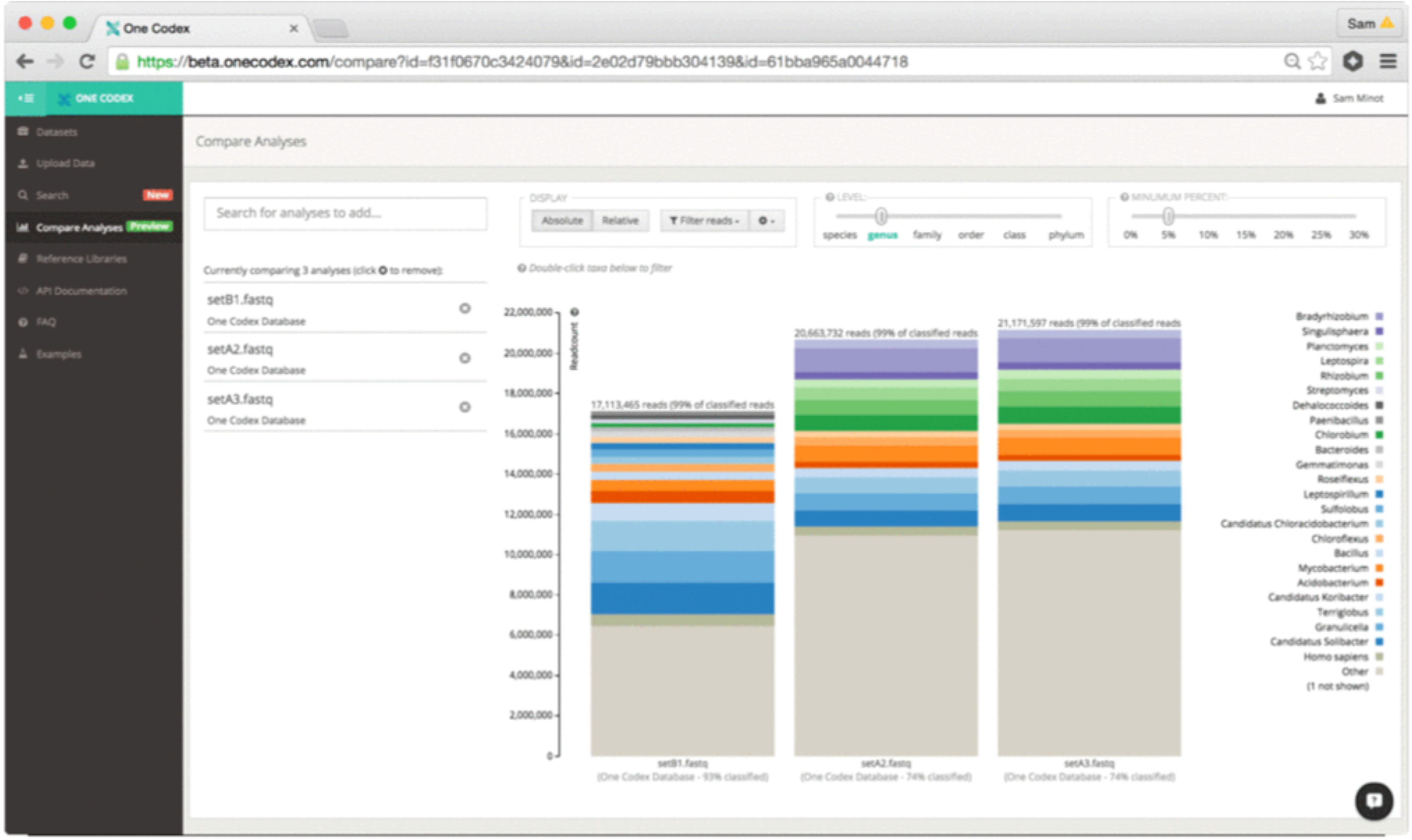
STEC from stool (Loman et al.)
Shiga-toxin-producing Escherichia Coli (STEC) is directly identified from stool samples during the European 2011 O104:H4 E. coli outbreak.
ERR262951.trimmed.fastq
2.7 GB | Wed, 04 Mar 2015 20:41:45 GMT
Illumina HiSeq
WGS
Metagenomic

Leptospira from CSF (Wilson et al.)
Leptospira Santarosai is identified from the cerebral spinal fluid (CSF) of a 14-year-old boy with severe combined immunodeficiency.
SRR1145846.trimmed.fastq
75.9 MB | Wed, 04 Mar 2015 20:41:07 GMT
Illumina MiSeq
WGS
Metagenomic

NYC Subway - South Ferry (Afshinnekoo et al.)
South Ferry/Terminal, New York, NY, United States
Sample from station floor. From the NYC Subway (PathoMap) study in Cell Systems. Samples from this station were noted for their unique microbial composition. Sample PAB17.
SRR1750062.trimmed.fastq
Illumina HiSeq
WGS
Metagenomic

Towards applied microbial genomics

… if we’re going to sequence everything, we should be able to easily index, explore, and compare samples



https://beta.onecodex.com/compare?id=f31f0670c3424079&id=2e02d79bbb304139&id=61bba965a0044718

Compare Analyses

Search for analyses to add...

Currently comparing 3 analyses (click ⊗ to remove):

setB1.fastq
One Codex Database

setA2.fastq
One Codex Database

setA3.fastq
One Codex Database

DISPLAY: [✓ Absolute] Relative

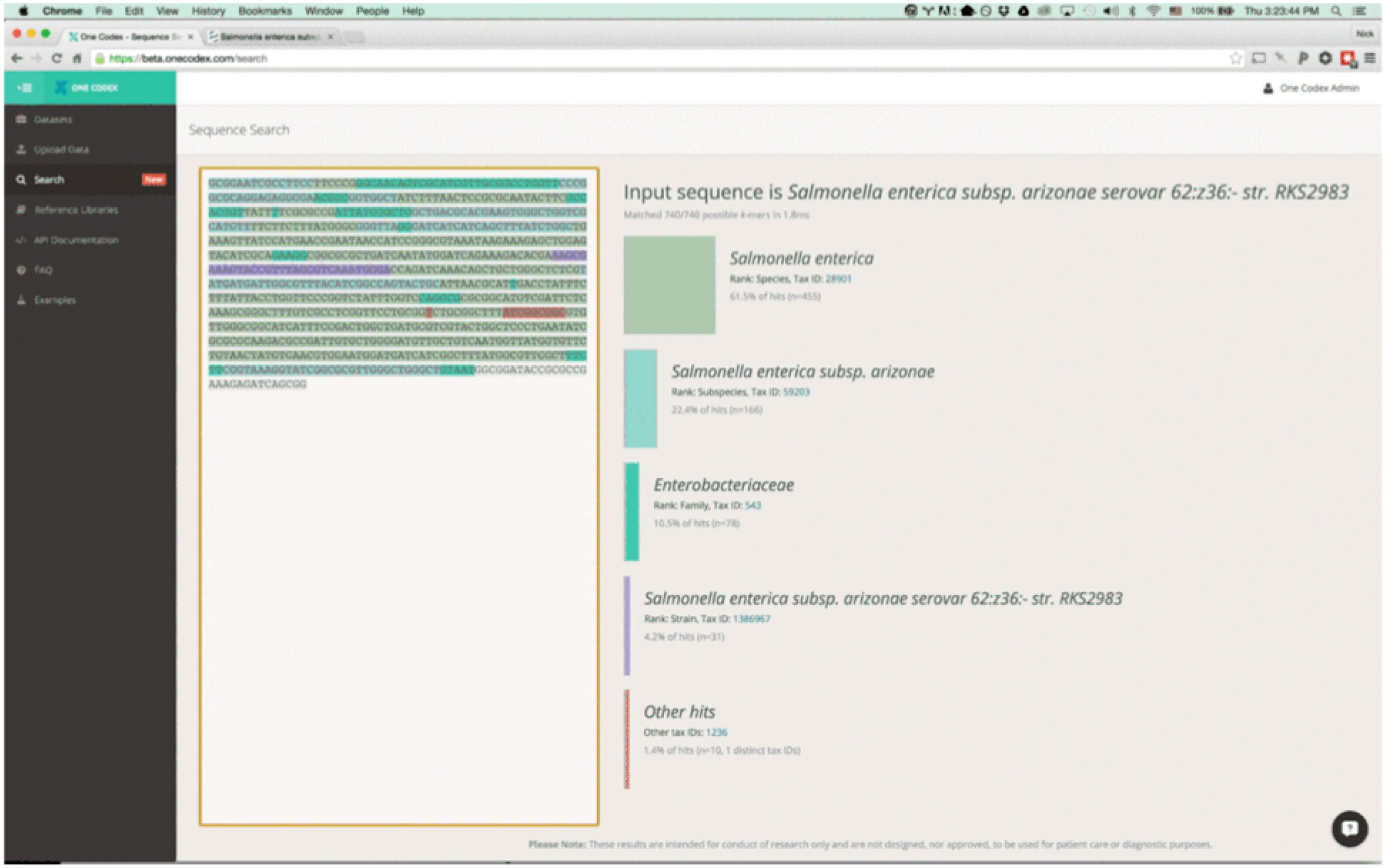
LEVEL: species [✓ genus] family order class phylum

MINUMUM PERCENT: 0% [✓ 5%] 10% 15% 20% 25% 30%

Double-click taxa below to filter

Towards applied microbial genomics

… if we're going to sequence everything, we should be able to easily index, explore, and compare samples



Sequence Search

```
GCGGAATCGCCTTCCTTCCCGGGCAACAGTCGCATCGTTGCGGCCTGGTTCCCG
GCGCAGGAGAGGGGAACGGCGGTGGCTATCTTTAACTCCGCGCAATACTTCGCC
ACGGTTATTTTCGCGCCGATTATGGGCTGGCTGACGCACGAAGTGGGCTGGTCG
CATGTTTTTCTTCTTTATGGGCGGGTTAGGGATCATCATCAGCTTTATCTGGCTG
AAAGTTATCCATGAACCGAATAACCATCCGGGCGTAAATAAGAAAGAGCTGGAG
TACATCGCAGAAGGCGGCGCGCTGATCAATATGGATCAGAAAGACACGAAAGCG
AAAGTACCGTTTAGCGTCAAATGGGACCAGATCAAACAGCTGCTGGGCTCTCGT
ATGATGATTGGCGTTTACATCGGCCAGTACTGCATTAACGCATTGACCTATTTC
TTTATTACCTGGTTCCCGGTCTATTTGGTCCAGGCGCGCGGCATGTCGATTCTC
AAAGCGGGCTTTGTCGCCTCGGTTCCTGCGGTCTGCGGCTTTATCGGCGGCGTG
TTGGGCGGCATCATTTCCGACTGGCTGATGCGTCGTACTGGCTCCCTGAATATC
GCGCGCAAGACGCCGATTGTGCTGGGGATGTTGCTGTCAATGGTTATGGTGTTC
TGTAACTATGTGAACGTGGAATGGATGATCATCGGCTTTATGGCGTTGGCTTTC
TTCGGTAAAGGTATCGGCGCGTTGGGCTGGGCTGTAATGGCGGATACCGCGCCG
AAAGAGATCAGCGG
```

Input sequence is *Salmonella enterica subsp. arizonae serovar 62:z36:- str. RKS2983*

Matched 740/740 possible k-mers in 1.8ms

Salmonella enterica
Rank: Species, Tax ID: 28901
61.5% of hits (n=455)

Salmonella enterica subsp. arizonae
Rank: Subspecies, Tax ID: 59203
22.4% of hits (n=166)

Enterobacteriaceae
Rank: Family, Tax ID: 543
10.5% of hits (n=78)

Salmonella enterica subsp. arizonae serovar 62:z36:- str. RKS2983
Rank: Strain, Tax ID: 1386967
4.2% of hits (n=31)

Other hits
Other tax IDs: 1236
1.4% of hits (n=10, 1 distinct tax IDs)

Please Note: These results are intended for conduct of research only and are not designed, nor approved, to be used for patient care or diagnostic purposes.

Algorithms/methods at One Codex

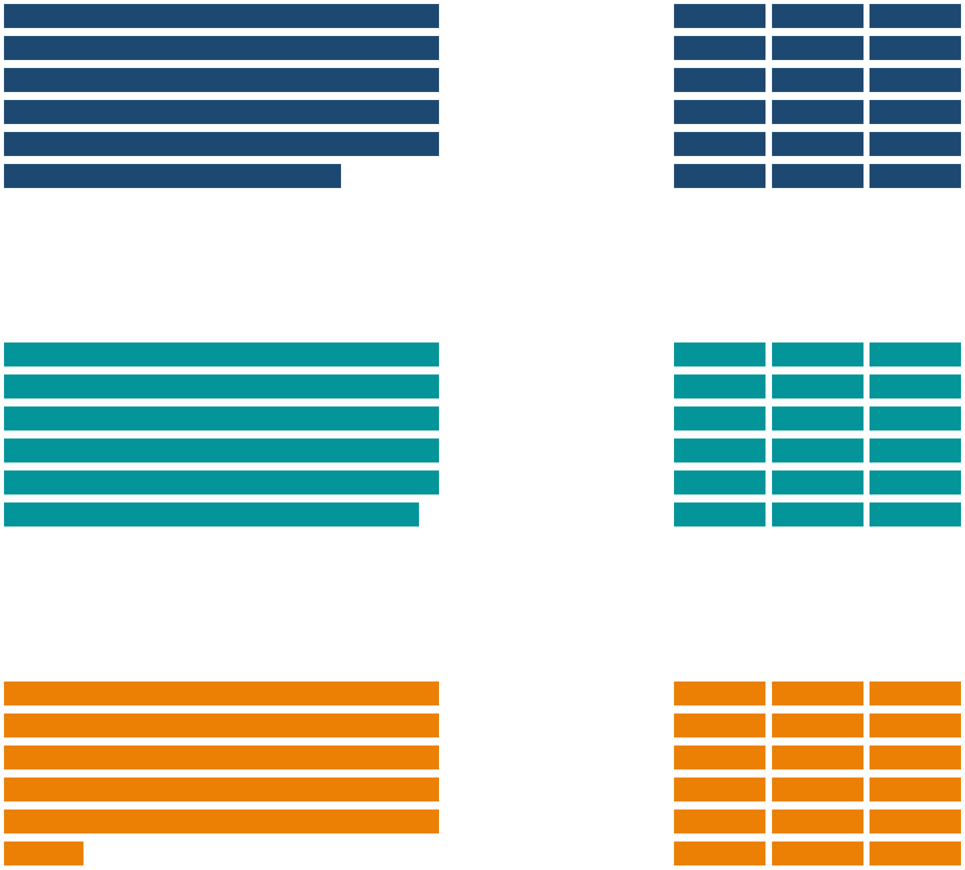
- “Exact alignment” or “alignment-free” *k*-mer based methods
- Scale well to huge reference sets (have 35k today)
- Dive into two today:
  a. Metagenomic classification
  b. Strain-typing

#1 Build a reference library

Reference Genomes

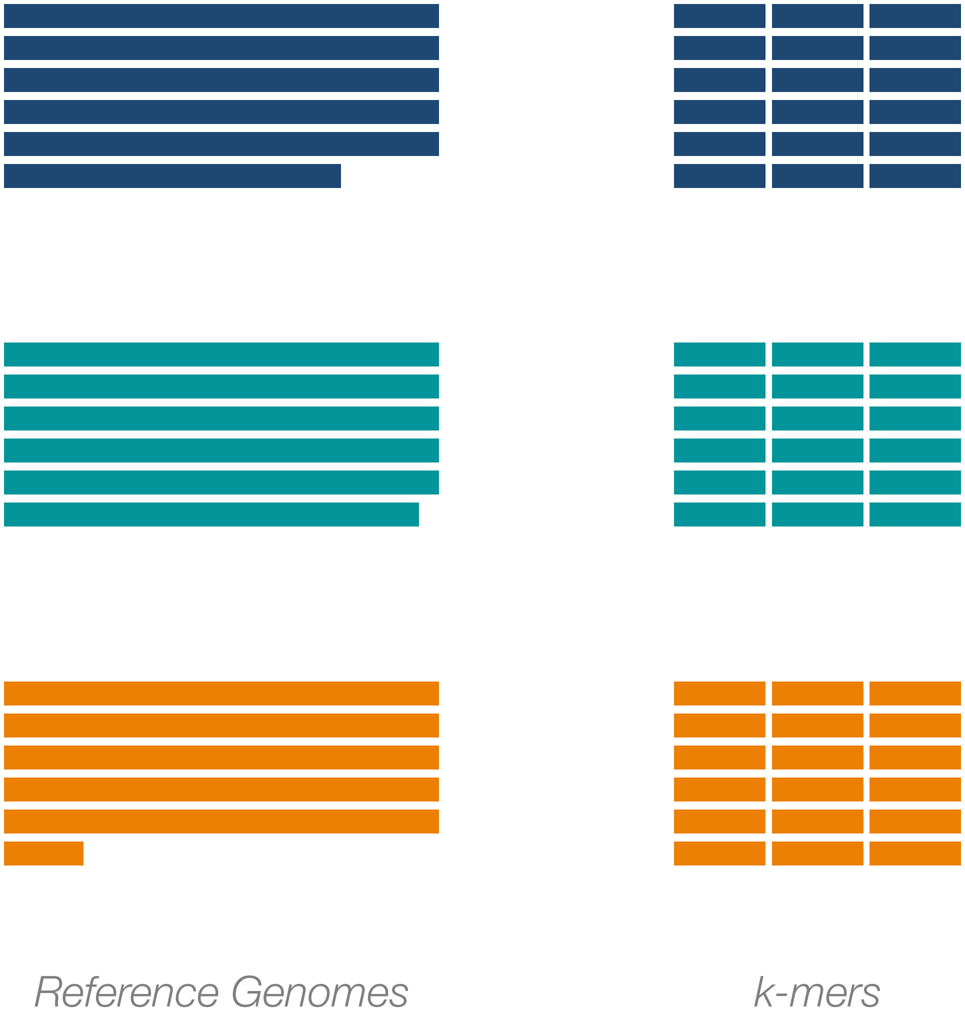
- Our latest database has ~35K bacterial, viral, fungal, and archaeal genomes
- Bigger tends to be better

#2 Enumerate all possible *k*-mers

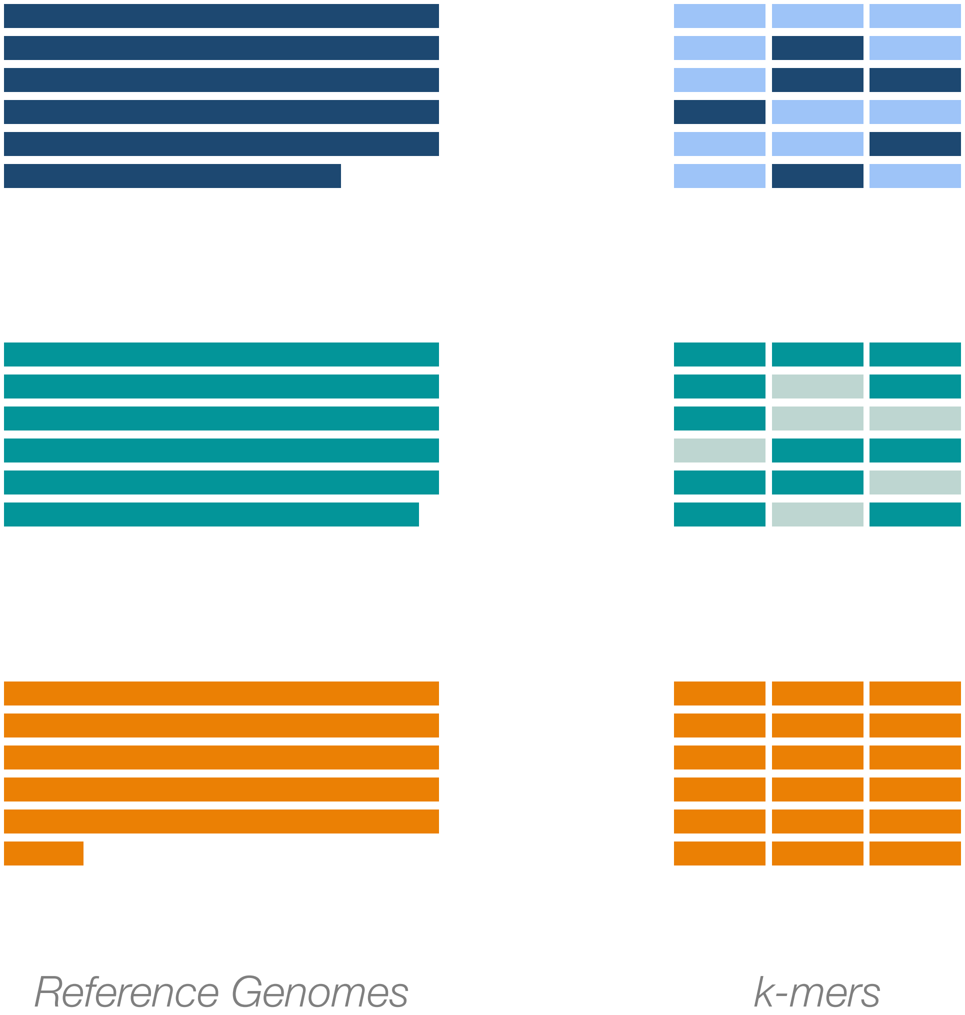


Reference Genomes *k-mers*

#3 Associate *k*-mers and tax IDs

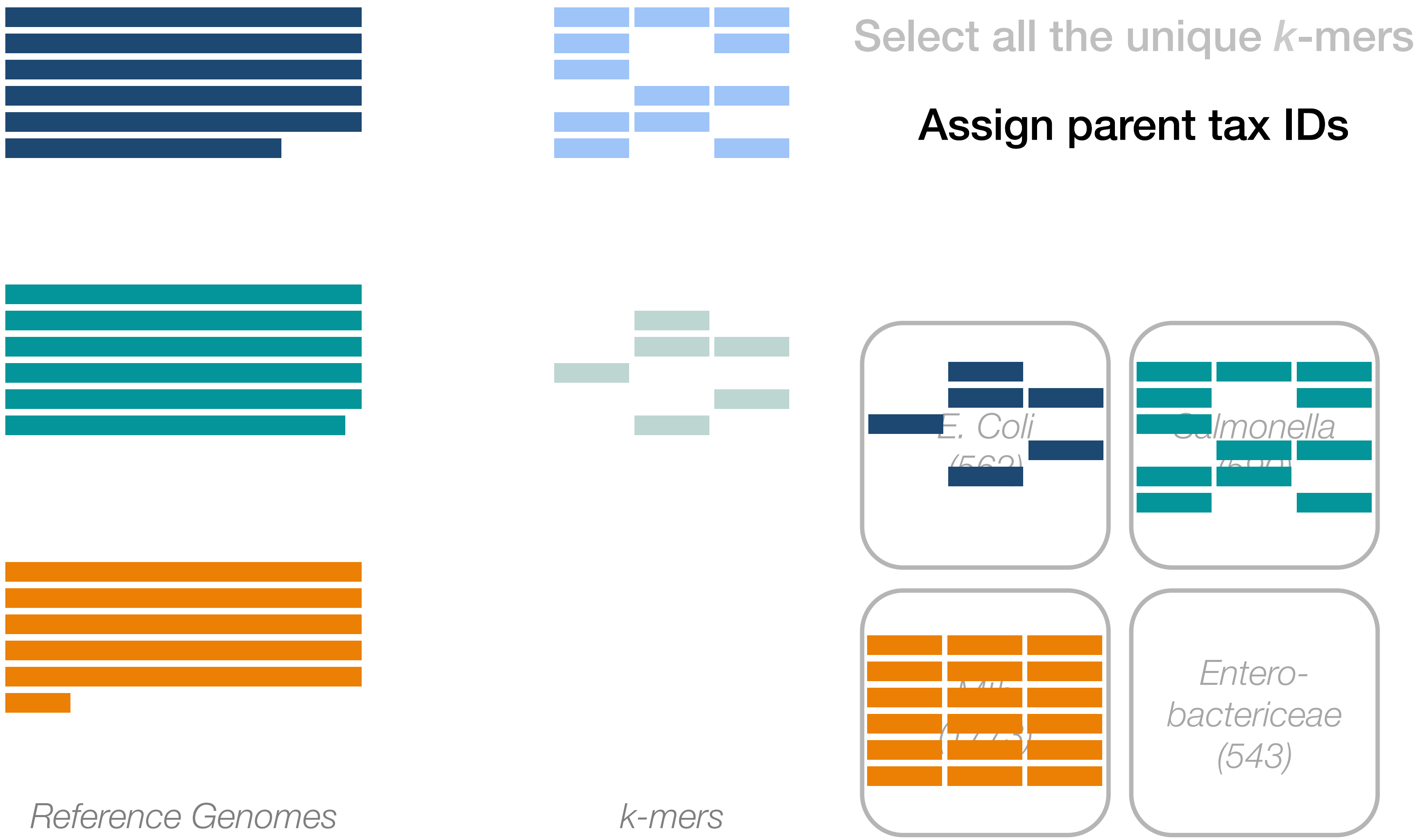


#3 Associate *k*-mers and tax IDs



Select all the unique *k*-mers

#3 Associate *k*-mers and tax IDs

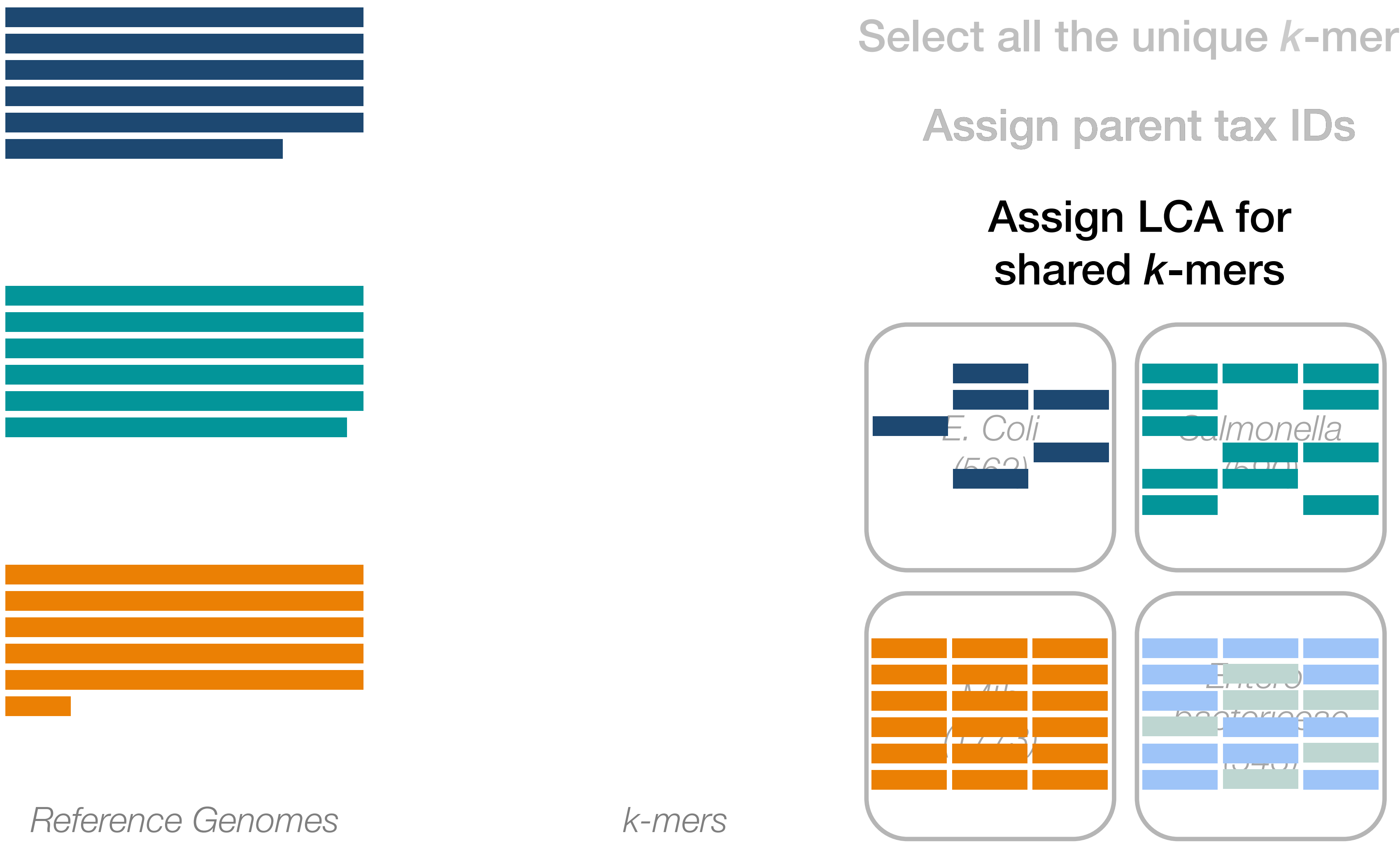


#3 Associate *k*-mers and tax IDs

Select all the unique *k*-mers

Assign parent tax IDs

Assign LCA for shared *k*-mers



#4 Lookup all *k*-mers for an input

```
@Sample.FASTQ length=151
TAGAGAATGTTAAGTCAGTTCCAGGACCAGTTAGGCGCACTTTATCTGTTTTATT...
+
BBBBBFFFFFFFGGGFGGBGGGGHGHHCGHHHGGHAGFHHDFHFGHHHFHHFGGE...
```

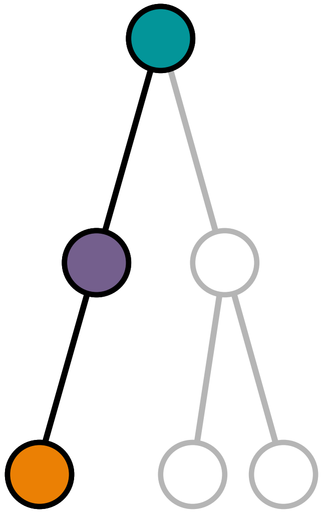
#4 Lookup all *k*-mers for an input

```
@Sample.FASTQ length=151
TAGAGAATGTTAAGTCAGTTCCAGGACCAGTTAGGCGCACTTTATCTGTTTTATT...
```

Firmicutes (1239)

Bacillus cereus group (86661)

Bacillus cereus (1396)



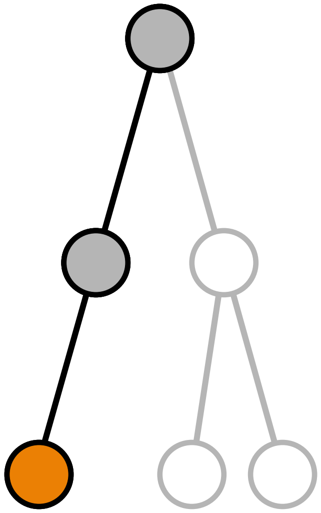
#5 Classify the input read

```
@Sample.FASTQ length=151
TAGAGAATGTTAAGTCAGTTCCAGGACCAGTTAGGCGCACTTTATCTGTTTTATT...
```

Firmicutes (1239)

Bacillus cereus group (86661)

Bacillus cereus (1396)



Classify result:
Highest weighted root-to-leaf path

Metagenomic classification summary

BUILD

1. Build a library of reference genomes
2. Enumerate all of the *k*-mers
3. Build a database of *k*-mers and associated taxonomy IDs

CLASSIFY

4. Look up all possible *k*-mers in a sample
5. Classify each read independently based on these *k*-mer “hits”

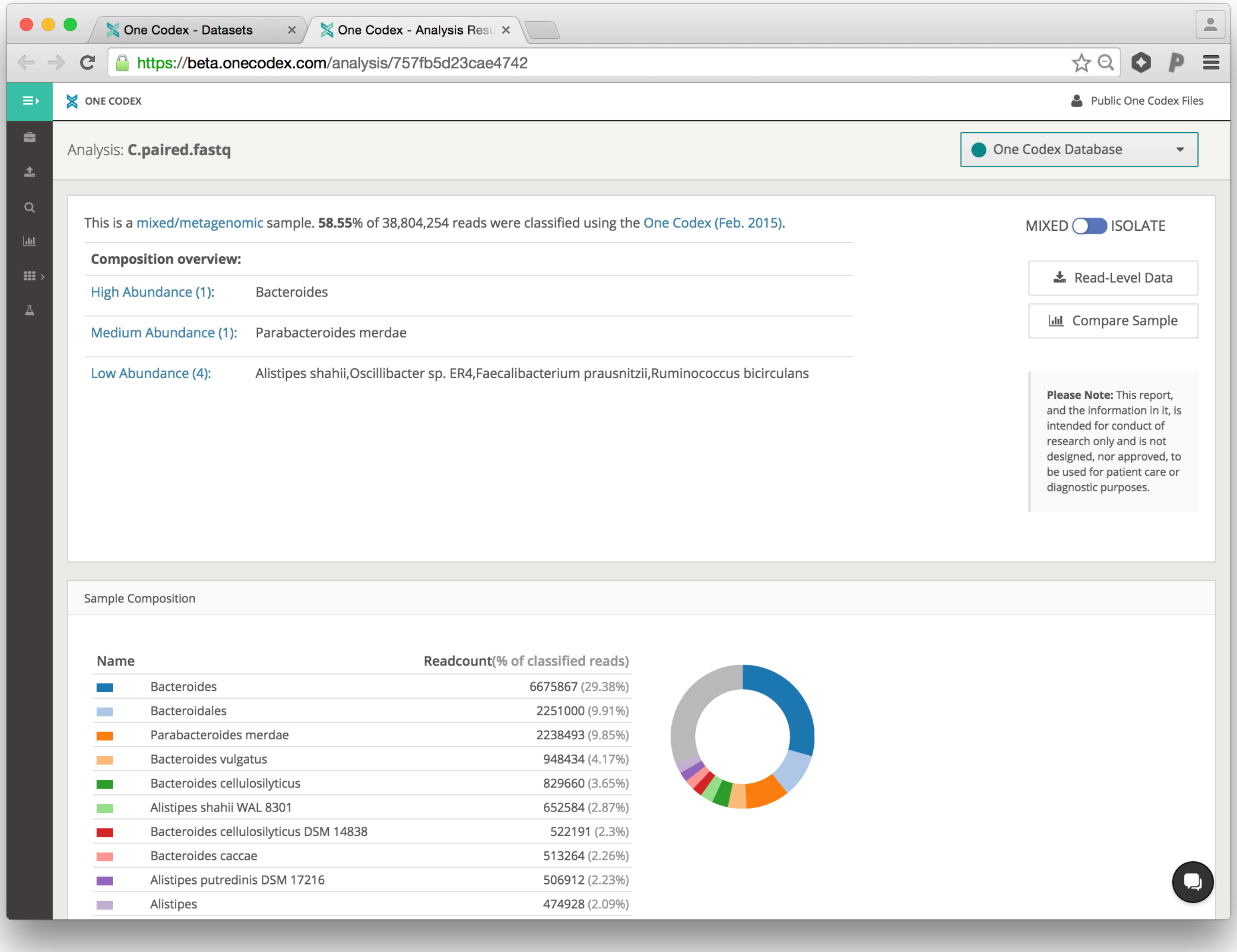
Metagenomic classification: accuracy

The One Codex platform delivers a comprehensive database – enabling substantial accuracy improvements

Classifier	HiSeq – Precision	HiSeq – Sensitivity	MiSeq – Precision	MiSeq – Sensitivity
PhymmBL	79.1	79.1	76.2	76.2
PhymmBL (conf. > 0.65)	99.1	73.9	92.5	73.0
Megablast	99.1	79.0	92.4	75.7
Naïve Bayes	82.3	82.3	77.8	77.8
Kraken	99.2	77.1	94.7	73.5
One Codex	**99.5**	**96.4**	**97.8**	**90.4**

Comparison of genus-level performance values adapted from Wood and Salzburg, 2014

Metagenomic classification: results



Analysis: **C.paired.fastq**

This is a mixed/metagenomic sample. **58.55**% of 38,804,254 reads were classified using the One Codex (Feb. 2015).

Composition overview:

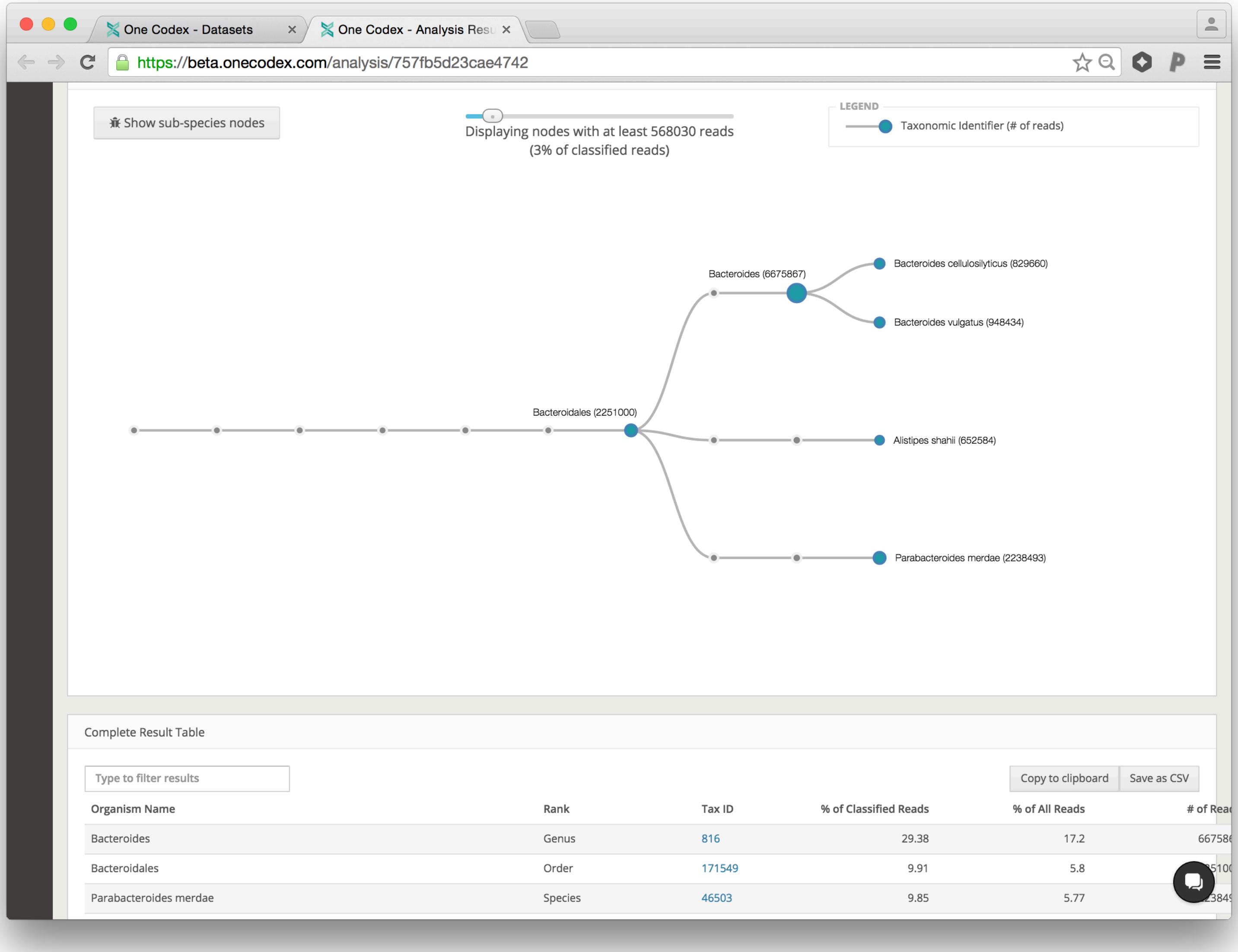
High Abundance (1):	Bacteroides
Medium Abundance (1):	Parabacteroides merdae
Low Abundance (4):	Alistipes shahii,Oscillibacter sp. ER4,Faecalibacterium prausnitzii,Ruminococcus bicirculans

Please Note: This report, and the information in it, is intended for conduct of research only and is not designed, nor approved, to be used for patient care or diagnostic purposes.

Sample Composition

	Name	Readcount(% of classified reads)
🟦	Bacteroides	6675867 (29.38%)
🟦	Bacteroidales	2251000 (9.91%)
🟧	Parabacteroides merdae	2238493 (9.85%)
🟧	Bacteroides vulgatus	948434 (4.17%)
🟩	Bacteroides cellulosilyticus	829660 (3.65%)
🟩	Alistipes shahii WAL 8301	652584 (2.87%)
🟥	Bacteroides cellulosilyticus DSM 14838	522191 (2.3%)
🟥	Bacteroides caccae	513264 (2.26%)
🟪	Alistipes putredinis DSM 17216	506912 (2.23%)
🟪	Alistipes	474928 (2.09%)

Metagenomic classification: results


One Codex - Datasets
One Codex - Analysis Resu
https://beta.onecodex.com/analysis/757fb5d23cae4742
Show sub-species nodes
Displaying nodes with at least 568030 reads
(3% of classified reads)
LEGEND
Taxonomic Identifier (# of reads)
Bacteroides (6675867)
Bacteroides cellulosilyticus (829660)
Bacteroides vulgatus (948434)
Bacteroidales (2251000)
Alistipes shahii (652584)
Parabacteroides merdae (2238493)
Type to filter results
Copy to clipboard
Save as CSV

Complete Result Table

Organism Name	Rank	Tax ID	% of Classified Reads	% of All Reads	# of Read
Bacteroides	Genus	816	29.38	17.2	667586
Bacteroidales	Order	171549	9.91	5.8	25100
Parabacteroides merdae	Species	46503	9.85	5.77	23849

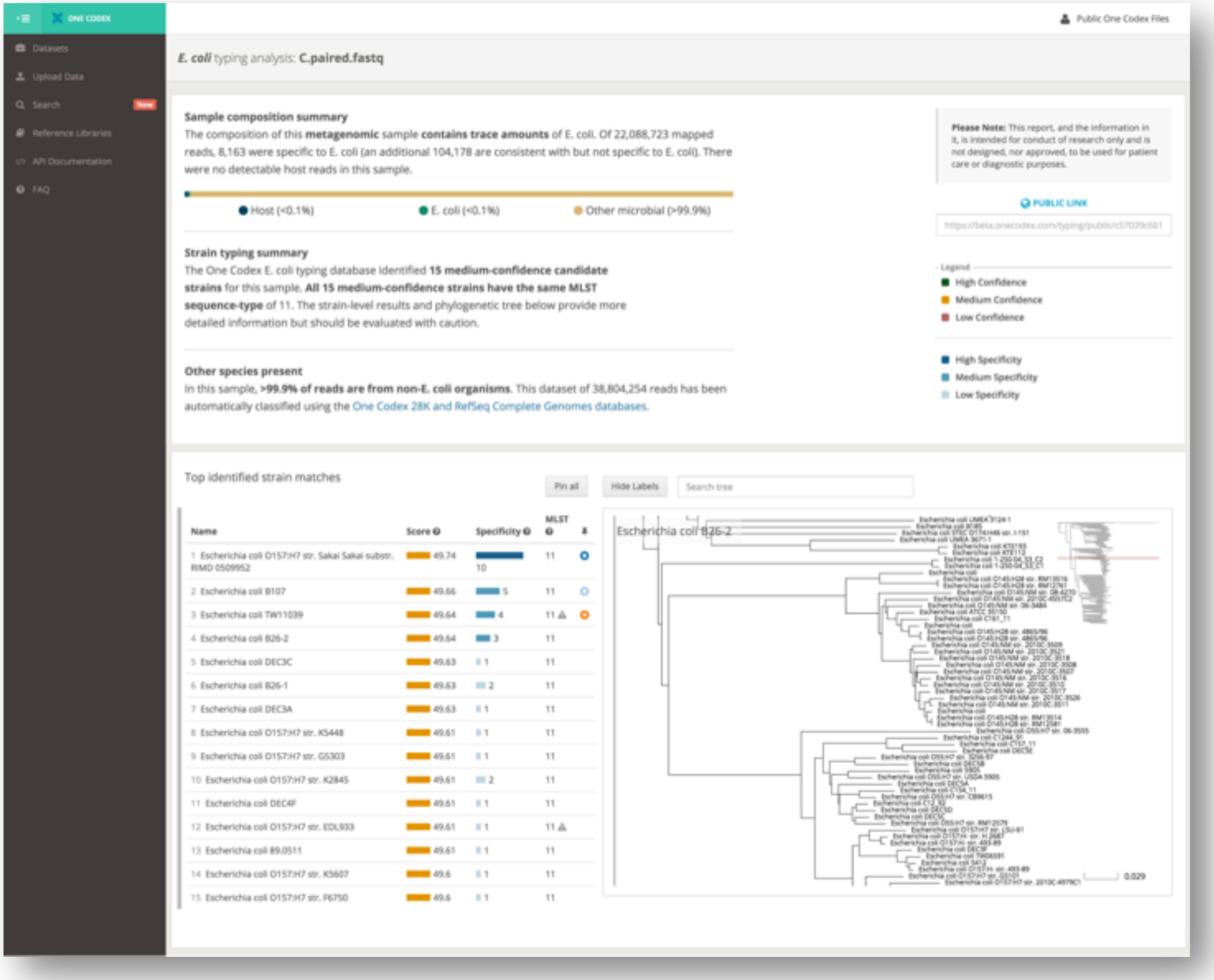
Company confidential

ONE CODEX

B) Strain-typing

- Won CDC “No Petri Dish” Challenge
- Works for isolates and mixed samples
- 99+% “serotype-level” accuracy
- Higher resolution than MLST at 10-20x lower coverage

B) Strain-typing



E. coli typing analysis: **C.paired.fastq**

Sample composition summary

The composition of this **metagenomic** sample **contains trace amounts** of E. coli. Of 22,088,723 mapped reads, 8,163 were specific to E. coli (an additional 104,178 are consistent with but not specific to E. coli). There were no detectable host reads in this sample.

● Host (<0.1%) ● E. coli (<0.1%) ● Other microbial (>99.9%)

Strain typing summary

The One Codex E. coli typing database identified **15 medium-confidence candidate strains** for this sample. **All 15 medium-confidence strains have the same MLST sequence-type** of 11. The strain-level results and phylogenetic tree below provide more detailed information but should be evaluated with caution.

Other species present

In this sample, **>99.9% of reads are from non-E. coli organisms**. This dataset of 38,804,254 reads has been automatically classified using the One Codex 28K and RefSeq Complete Genomes databases.

Please Note: This report, and the information in it, is intended for conduct of research only and is not designed, nor approved, to be used for patient care or diagnostic purposes.

PUBLIC LINK

https://beta.onecodex.com/typing/public/c57039c661

Legend

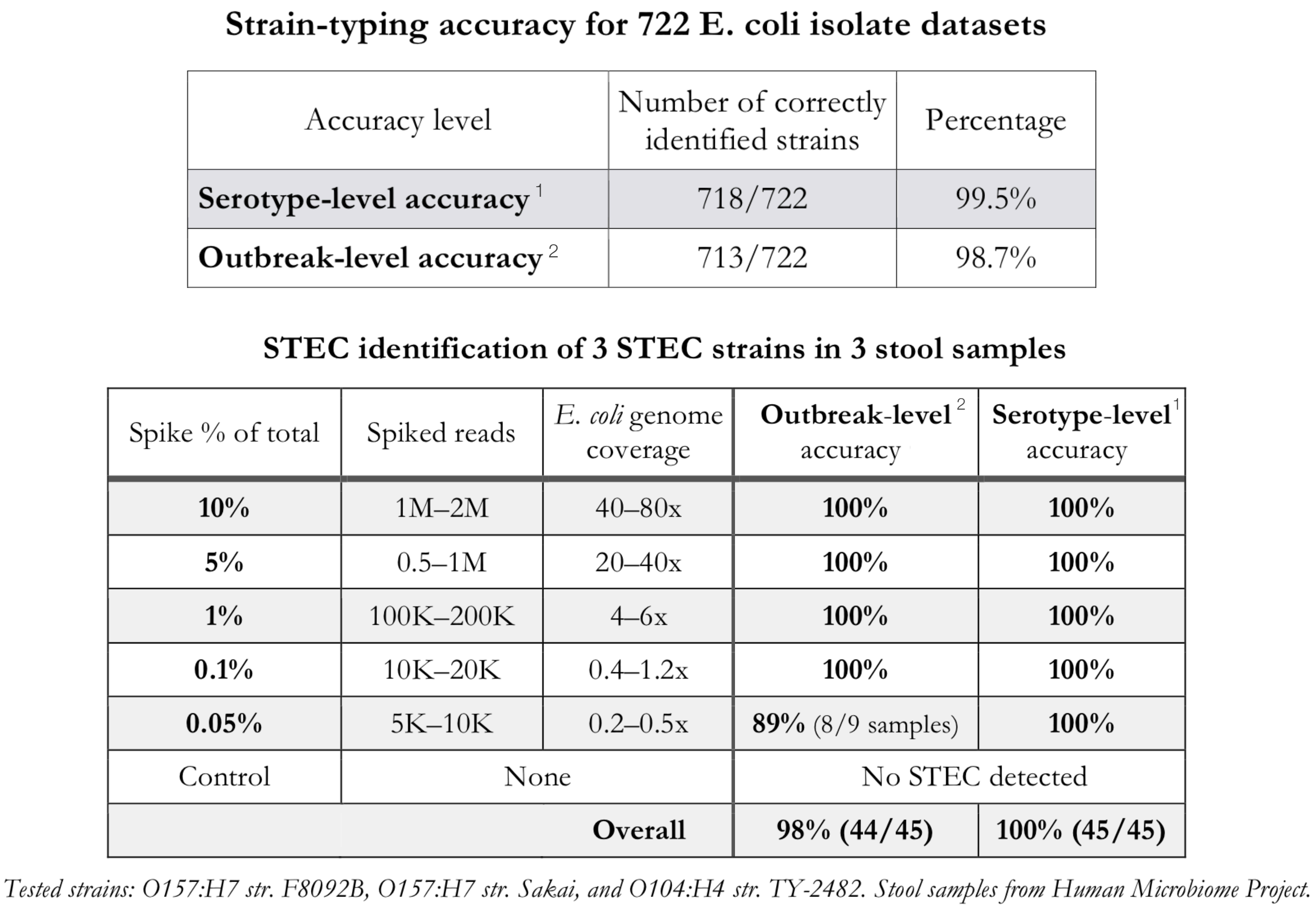
- 🟩 High Confidence
- 🟧 Medium Confidence
- 🟥 Low Confidence

- 🟦 High Specificity
- 🟦 Medium Specificity
- ⬜ Low Specificity

Top identified strain matches

Name	Score	Specificity	MLST	
1 Escherichia coli O157:H7 str. Sakai Sakai substr. RIMD 0509952	49.74	10	11	○
2 Escherichia coli B107	49.66	5	11	○
3 Escherichia coli TW11039	49.64	4	11 ⚠	○
4 Escherichia coli B26-2	49.64	3	11	
5 Escherichia coli DEC3C	49.63	1	11	
6 Escherichia coli B26-1	49.63	2	11	
7 Escherichia coli DEC3A	49.63	1	11	
8 Escherichia coli O157:H7 str. K5448	49.61	1	11	
9 Escherichia coli O157:H7 str. G5303	49.61	1	11	
10 Escherichia coli O157:H7 str. K2845	49.61	2	11	
11 Escherichia coli DEC4F	49.61	1	11	
12 Escherichia coli O157:H7 str. EDL933	49.61	1	11 ⚠	
13 Escherichia coli 89.0511	49.61	1	11	
14 Escherichia coli O157:H7 str. K5607	49.6	1	11	
15 Escherichia coli O157:H7 str. F6750	49.6	1	11	

B) Strain-typing: accuracy



Strain-typing accuracy for 722 E. coli isolate datasets

Accuracy level	Number of correctly identified strains	Percentage
Serotype-level accuracy [1]	718/722	99.5%
Outbreak-level accuracy [2]	713/722	98.7%

STEC identification of 3 STEC strains in 3 stool samples

Spike % of total	Spiked reads	*E. coli* genome coverage	**Outbreak-level** [2] accuracy	**Serotype-level** [1] accuracy
10%	1M–2M	40–80x	**100%**	**100%**
5%	0.5–1M	20–40x	**100%**	**100%**
1%	100K–200K	4–6x	**100%**	**100%**
0.1%	10K–20K	0.4–1.2x	**100%**	**100%**
0.05%	5K–10K	0.2–0.5x	**89%** (8/9 samples)	**100%**
Control	None		No STEC detected	
		Overall	**98% (44/45)**	**100% (45/45)**

Tested strains: O157:H7 str. F8092B, O157:H7 str. Sakai, and O104:H4 str. TY-2482. Stool samples from Human Microbiome Project.

[1] Identified strain within 0.06 genetic distance (median within-serotype distance) from known spike-in

[2] Identified strain within 95% confidence interval for intra-serotype genetic distance

Questions?

Open beta: https://beta.onecodex.com

Contact: nick@onecodex.com