



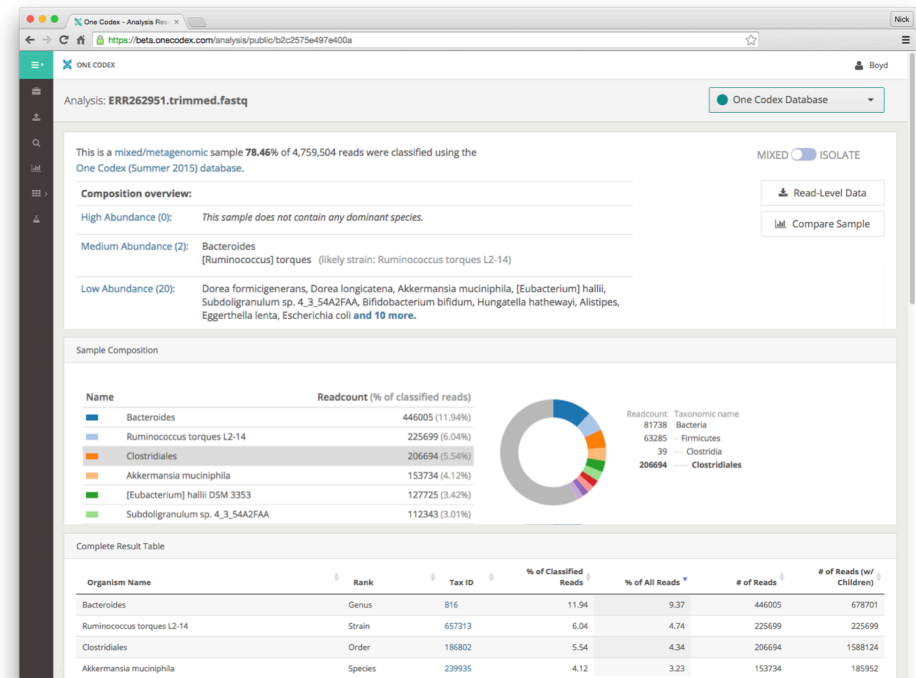
From biobanks to databanks?

*Exploring the PathoMAP project
on the One Codex data platform*

Nick Greenfield, Founder & CEO
MetaSUB Summit, Shanghai, July 1st 2016

Brief background (really!)

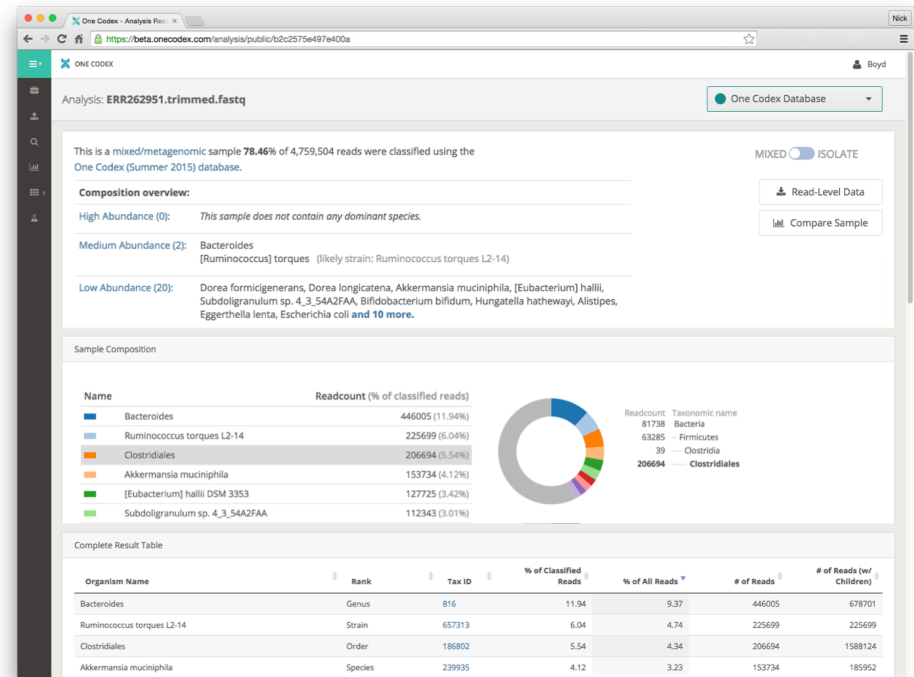
- San Francisco-based software company
- Data platform offering “sequence to answer” solution for microbial and metagenomics
- Design goals include: scalability, reproducibility, and ease-of-use



One Codex metagenomic analysis view

Brief background (really!)

- San Francisco-based software company
- Data platform offering “sequence to answer” solution for microbial and metagenomics
- Design goals include: scalability, reproducibility, and ease-of-use



One Codex metagenomic analysis view

Biobanks...



Biobanks...



Advantages: Biological & “close to truth”; “multi-media”; vast historical archives

Biobanks...



Disadvantages: Fragile, finite/depletable, expensive to use and maintain

[illegible]

... to databanks?



```
@HWI-D00151:75:H9AQQADXX:2:1101:11980:2239 2:N:0:GGAATCCTTATCCTCT
AATTCAACGATACGCCAGCCCTTCGACAAACGGCTGTACCGATCTACCGCTTTACGTGGCCCGAGGCCATTGCCGCACCAGAGAGATGCGCTTAAGGTAC
+
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00151:75:H9AW7ADXX:2:1101:14056:2237 2:N:0:GGAATCCTTATCCTCT
GTATTGTAGAGCTGTCGATTAGCGTGTCGNTGATGGCCAAATGAAGGGCGACCGCGTCTGGCGTCGACGCTGACGCGTAACGCTTCGTGGTGCTGCAATT
+
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00151:75:H9AQQADXX:2:1101:18423:2248 2:N:0:GGAATCCTTATCCTCT
GGCAATGGACAATGTATCCGCGGTGCGCGTAGTTGAAGGATGTGACGATAAACAGCATCACCCTATTTTCTCTTATACACATCTGACGCTGCGACGAAG
+
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00151:75:H9AQQADXX:2:1101:1667:2337 2:N:0:GGAATCCTTATCCTCT
CGCTATTAGTTGCCTGATTGAGCGTCTTCGCCATGCGGGCATTGCTGCCTGGAGTGGCAACAGTGTCCGGAACAGGAAGTGAATCGGCTAGGGAATGCA
+
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00151:75:H9AQQADXX:2:1101:1848:2341 2:N:0:GGAATCCTTATCCTCT
CAGAGCCGAAGACAAAAAGATGTCTGCAATAATGTAATGTACATATAGATCGATTGCGAGCATCATCGCAACTATGTGCTCATGGATTCTTCAACGTAAT
+
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00151:75:H9AQQADXX:2:1101:1849:2248 2:N:0:GGAATCCTTATCCTCT
GGCAATGGACGTCAAATGTGGCGCGCCGTAGTTGAAGGATGTACGATAAACAGCATCACCCTATTTGTCTCACGTTAACTCTGACGCTGCCGACGAA
+
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Complement biobanks: (Relatively) durable, distributable,
(nearly) free analysis & re-analysis

[illegible]

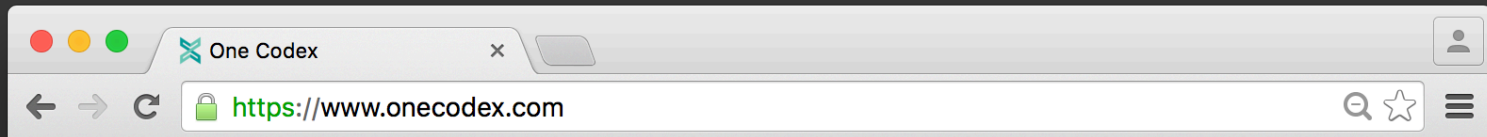
... to databanks?

```
@HWI-D00151:75:H9AQQADXX:2:1101:11980:2239 2:N:0:GGACTCCTTATCCTCT
AATTCAACGATACGCCAGCCCTTCGACAAACGGCTGTACCGATCTACCGCTTTACGTGGCCCGAGGCCATTGCCGCACCAGAGAGATGCGCTTAAGGTAC
+
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00151:75:H9AW7ADXX:2:1101:14056:2237 2:N:0:GGACTCCTTATCCTCT
GTATTGTAGAGCTGTCGATTAGCGTGTCGNTGATGGCCAAATGAAGGGCGACCGCGTCTGGCGTCGACGCTGACGCGTAACGCTTCGTGGTGCTGCAATT
+
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00151:75:H9AQQADXX:2:1101:18423:2248 2:N:0:GGACTCCTTATCCTCT
GGCAATGGACAATGTATCCGCGGTGCGCGTAGTTGAAGGATGTGACGATAAACAGCATCACCCTATTTTCTCTTATACACATCTGACGCTGCGACGAAG
+
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00151:75:H9AQQADXX:2:1101:1667:2337 2:N:0:GGACTCCTTATCCTCT
CGCTATTAGTTGCCTGATTGAGCGTCTTCGCCATGCGGGCATTGCTGCCTGGAGTGGCAACAGTGTCCGGAACAGGAAGTGAATCGGCTAGGGAATGCA
+
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00151:75:H9AQQADXX:2:1101:1848:2341 2:N:0:GGACTCCTTATCCTCT
CAGAGCCGAAGACAAAAAGATGTCTGCAATAATGTAATGTACATATAGATCGATTGCGAGCATCATCGCAACTATGTGCTCATGGATTCTTCAACGTAAT
+
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
@HWI-D00151:75:H9AQQADXX:2:1101:1849:2248 2:N:0:GGACTCCTTATCCTCT
GGCAATGGACGTCAAATGTGGCGCGCCGTAGTTGAAGGATGTACGATAAACAGCATCACCCTATTTGTCTCACGTTAACTCTGACGCTGCCGACGAA
+
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
```

Requirements: Security, scalability, ease-of-use (*and* extensibility), and reproducibility

[illegible]

[illegible]



ONE CODEX

[Sign up](#)[Log in](#)

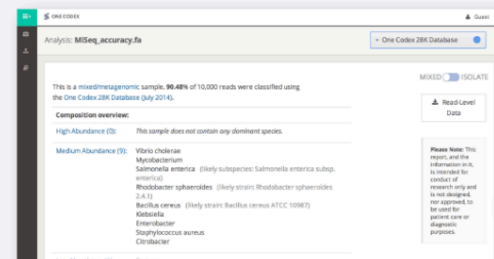
One Codex is a **data platform for applied microbial genomics**, enabling new and valuable applications in clinical diagnostics, food safety, and biosecurity

Best-in-class metagenomics

One Codex helps applied microbiologists rapidly assess their samples, with a platform that is:



Fast – Process isolates or complex metagenomic samples in minutes, not days



One Codex - Analysis Results

https://app.onecodex.com/analysis/1bd321353efa45dd

ONE CODEXMetaSUB

Analysis - P01295

One Codex Database

This is a **mixed/metagenomic** sample **44.77%** of 1,127,256 reads were classified using the **One Codex database**.

Composition overview:

High Abundance (1):

Leclercia adecarboxylata (likely strain : Leclercia adecarboxylata ATCC 23216 = NBRC 102595)

Medium Abundance (1):

Stenotrophomonas maltophilia

Low Abundance (9):

Sphingobium yanoikuyae, Lachnoclostridium phytofermentans, Stenotrophomonas rhizophila, Stenotrophomonas sp. SKA14, Stenotrophomonas pavanii, Stenotrophomonas chelatiphaga, Lysinibacillus xylanilyticus, bacterium MS4, Pantoea sp. PSNIH1

MIXEDISOLATE

Read-Level Data

Compare Sample

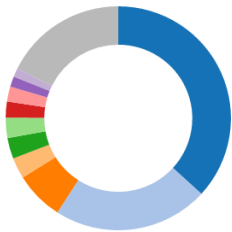
You are viewing the filtered One Codex results. Switch to view the unfiltered results.

Sample Composition

Name	Estimated Abundance
Leclercia adecarboxylata	36.79%
Stenotrophomonas maltophilia	22.30%
[Pseudomonas] geniculata	7.03%
Lysinibacillus xylanilyticus	3.03%
Stenotrophomonas sp. RIT309	3.01%
Acinetobacter sp. UNC436CL71CviS28	2.93%
Pantoea sp. PSNIH1	2.31%

DISPLAY LEVEL

Abundance



One Codex - Marker panel

https://app.onecodex.com/markerpanel/sample2

ONE CODEXMetaSUB

Panel: AMR Panel - Sample.fa

AMR Panel

Antamycins

Showing all rows

Name	Description	Status	Marker Support		
arr-3	AY038837-555:1007 Marker length: 453 bp	Absent	79.2 % coverage	100 % identity	1.5 x depth

Glycopeptides

Expand

Name	Description	Status	Marker Support		
van(C3)	AY033764-26:1078 Marker length: 1053 bp	Absent	0.0 % coverage	0.0 % identity	0.0 x depth
van(Y)	AF516335-5821:6366 Marker length: 546 bp	Absent	0.0 % coverage	0.0 % identity	0.0 x depth
van(D4)	AF277571-1262:2293 Marker length: 1032 bp	Absent	0.0 % coverage	0.0 % identity	0.0 x depth
van(B)	U00456-62:1090 Marker length: 1029 bp	Absent	0.0 % coverage	0.0 % identity	0.0 x depth
van(C1)	AF162694-1411:2442 Marker length: 1032 bp	Absent	0.0 % coverage	0.0 % identity	0.0 x depth

Beta-lactams

Expand

Name	Description	Status	Marker Support		
blaTEM-1	AF309824-119:979 Marker length: 861 bp	Present	100 % coverage	99.8 % identity	7.1 x depth
blaLAT-1-LAT-4	X78117-142:1287 Marker length: 1146 bp	Probable	99.4 % coverage	97.8 % identity	5.8 x depth
blaOXA-1	J02967-1359:2189 Marker length: 831 bp	Absent	91.5 % coverage	99.9 % identity	2.3 x depth

One Codex

https://app.onecodex.com

ONE CODEX

MetaSUB

Samples

Sort:

Date

Name

Starred

Size

Public

Project

Search

PathoMAP

All

P01295

☆

SRR1749852.fastq.gz

PathoMAP

134.4 MB | 08/08/2015

city

metal

PathoMAP

...

View Results

GCSS-12

☆

SRR1750104.fastq.gz

PathoMAP

113.8 MB | 08/08/2015

USA:New York City

PathoMAP

aquatic

canal

...

View Results

P01607

☆

SRR1750039.fastq.gz

PathoMAP

111.1 MB | 08/08/2015

USA:New York City

city

...

View Results

P00379

☆

SRR1748967.fastq.gz

PathoMAP

3.0 GB | 08/08/2015

USA:New York City

city

...

View Results

One Codex

https://app.onecodex.com

MetaSUB

Samples

Sort:

Date

Name

Starred

Size

Public

Project

Search

PathoMAP

Edit

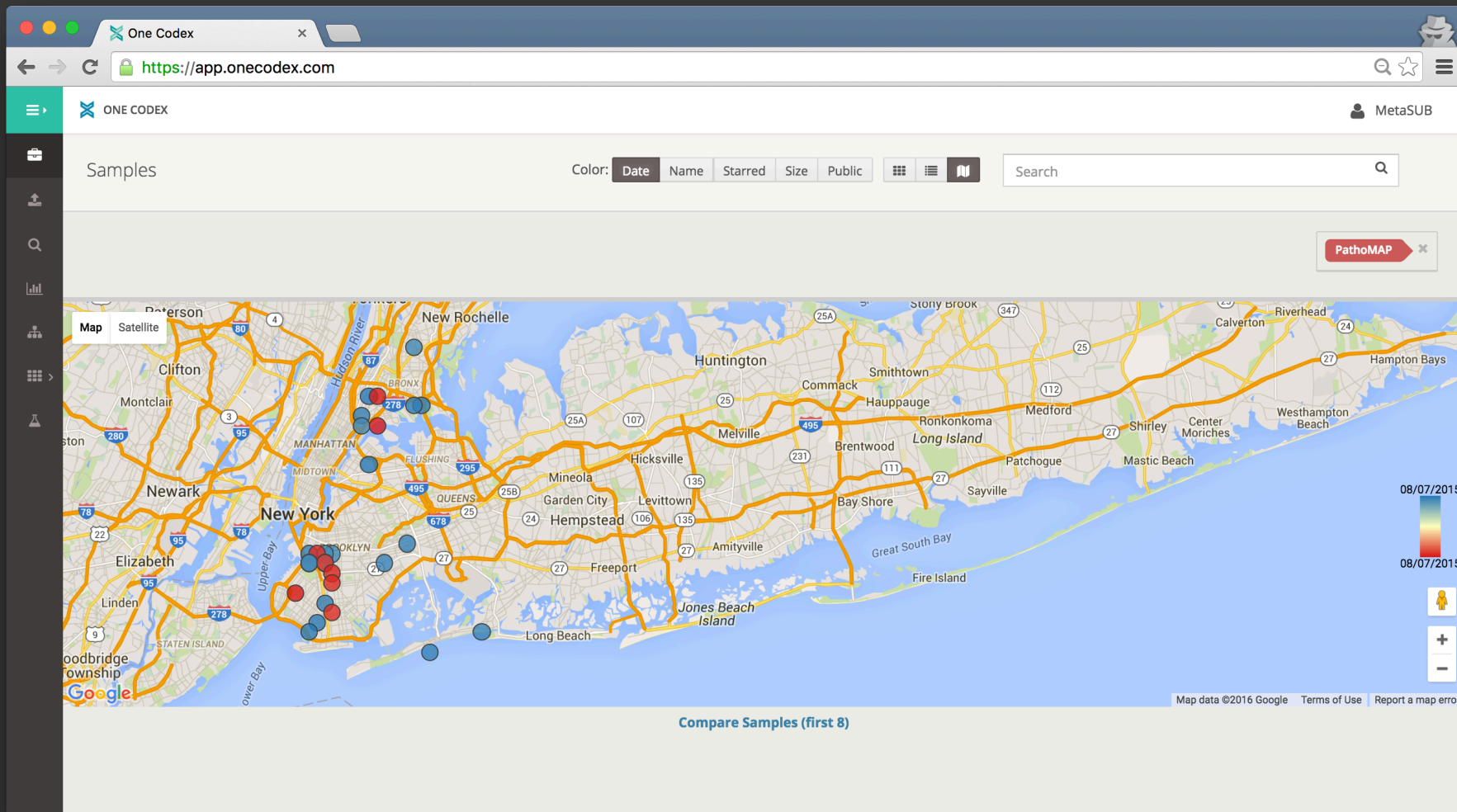
0 selected

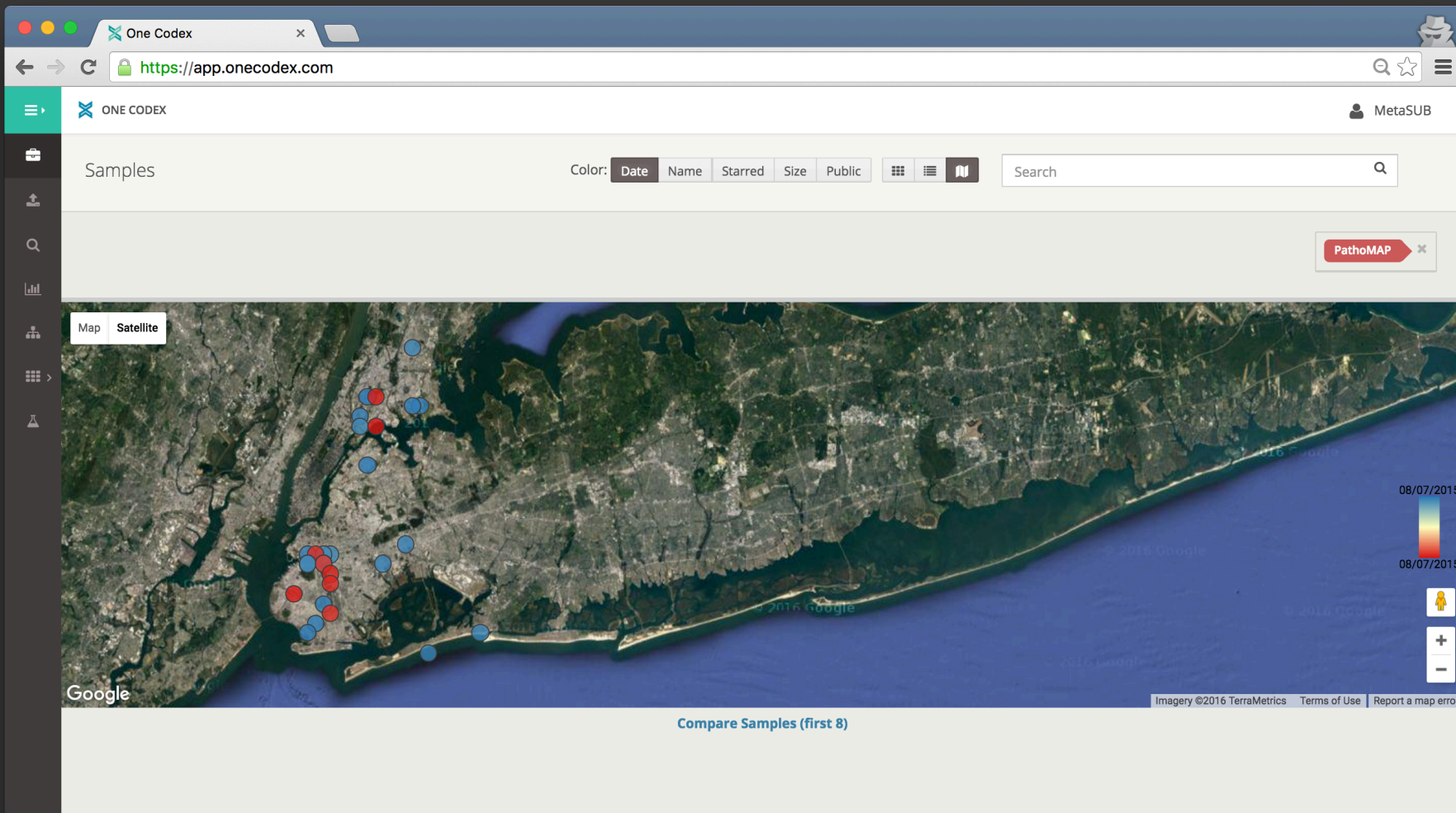
Page 1 of 48

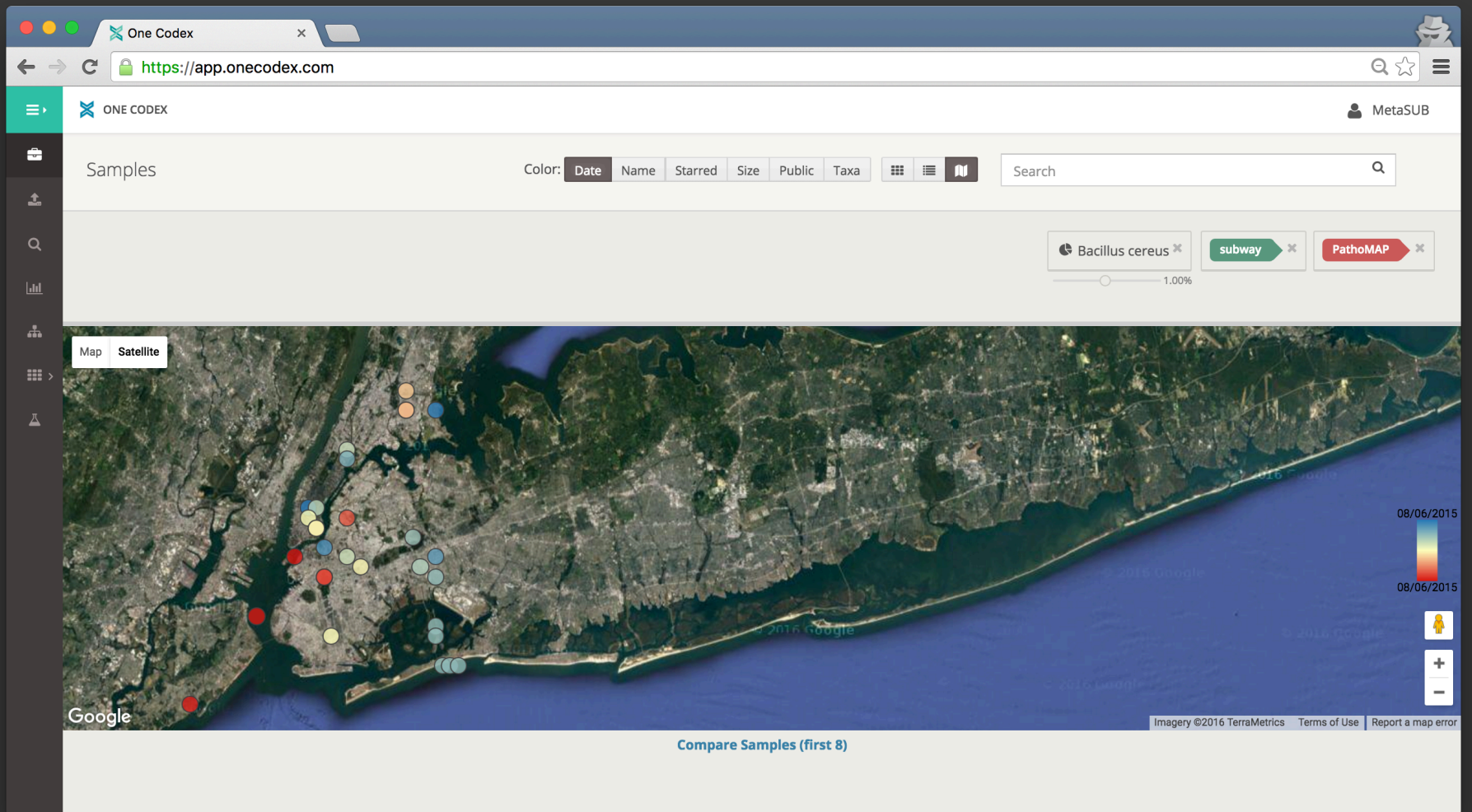
Download Sample Sheet (.xlsx)

Upload Sample Sheet

☆ P01295	8/8/15	134.4 MB	city metal PathoMAP subway		View Results
☆ GCSS-12	8/8/15	113.8 MB	PathoMAP aquatic canal unclassified water		View Results
☆ P01607	8/8/15	111.1 MB	city PathoMAP subway wood		View Results
☆ P00379	8/8/15	3.0 GB	city PathoMAP subway wood		View Results
☆ P01482	8/8/15	81.4 MB	city PathoMAP subway metal-plastic		View Results
☆ P01398	8/8/15	90.2 MB	city metal PathoMAP subway		View Results









Total items: 1584
Page 1/80, page size is 20
Links: [next](#) [last](#)

v1 API

```
[
  {
    "$uri": "/api/v1/samples/404d876f78a74f58",
    "created_at": "2015-08-06T19:22:15.646906+00:00",
    "filename": "SRR1748642.fastq.gz",
    "indexed": false,
    "metadata": {
      "$ref": "/api/v1/metadata/517f888cdd7d4526"
    },
    "owner": {
      "$ref": "/api/v1/users/235f286b0e7045ed"
    },
    "primary_analysis": {
      "$ref": "/api/v1/analyses/54edd12877664bb8"
    },
    "project": {
      "$ref": "/api/v1/projects/4e69cae548784edb"
    },
    "public": false,
    "size": 340104724,
    "starred": false,
    "tags": [
      {
        "$ref": "/api/v1/tags/b93e3abc3021420f"
      },
      {

```

200 OK
</> COPY JSON





Total items: 1584
Page 1/80, page size is 20
Links: [next](#) [last](#)

v1 API

```
[
  {
    "$uri": "/api/v1/samples/404d876f78a74f58",
    "created_at": "2015-08-06T19:22:15.646906+00:00",
    "filename": "SRR1748642.fastq.gz",
    "indexed": false,
    "metadata": {
      "$uri": "/api/v1/metadata/517f888cdd7d4526",
      "date_collected": "2013-07-11T00:00:00+00:00",
      "date_sequenced": "2015-01-09T00:00:00+00:00",
      "description": "city - subway - metal",
      "external_sample_id": null,
      "library_type": "WGS",
      "location_lat": 40.76,
      "location_lon": -73.96,
      "location_string": "USA:New York City",
      "name": "P00088",
      "platform": "Illumina HiSeq",
      "sample": {
        "$ref": "/api/v1/samples/404d876f78a74f58"
      },
      "sample_type": "Metagenomic"
    },
    "owner": {
      "$uri": "/api/v1/users/235f286b0e7045ed",
      "username": "metasub"
    }
  }
]
```

200 OK
[</>](#) [COPY](#) [JSON](#)



v1 API

```
{
  "$uri": "/api/v1/projects/4e69cae548784edb",
  "description": "PathoMAP Sequencing Project",
  "name": "PathoMAP",
  "owner": {
    "$ref": "/api/v1/users/235f286b0e7045ed"
  },
  "permissions": [
    "can_administer",
    "can_download_files",
    "can_incur_charges",
    "can_edit_metadata",
    "can_add_files",
    "can_see_files"
  ],
  "project_name": "pathomap",
  "public": false
}
```

200 OK
</> COPY JSON

Need help with the API? Click [here](#) to view our API documentation, or send us a message to the right!

Hint: A JSON ?sort query argument can be used to sort GET requests

Server time: 2016-06-29 02:45:03 / Logged in as
metasubway@gmail.com

[API Schema](#)



Total items: 1437
Page 1/72, page size is 20
Links: [next](#) [last](#)

v1 API

```
[
  {
    "$uri": "/api/v1/samples/04c873edd2b84c25",
    "created_at": "2015-08-07T01:26:23.186565+00:00",
    "filename": "SRR1748569.fastq.gz",
    "indexed": false,
    "metadata": {
      "$ref": "/api/v1/metadata/e1c6a6279ff24440"
    },
    "owner": {
      "$ref": "/api/v1/users/235f286b0e7045ed"
    },
    "primary_analysis": {
      "$ref": "/api/v1/analyses/68edcf6c62964b54"
    },
    "project": {
      "$ref": "/api/v1/projects/4e69cae548784edb"
    },
    "public": false,
    "size": 436780487,
    "starred": false,
    "tags": [
      {
        "$ref": "/api/v1/tags/b93e3abc3021420f"
      },
      {

```

200 OK
</> COPY JSON



One Codex

https://app.onecodex.com

ONE CODEX

Databank

jupyter

MetaSUB Client Library Demo (autosaved)

Python 2

FileEditViewInsertCellKernelHelp

Code

CellToolbar

In [1]:

```
# Load the One Codex client library
from onecodex.api import Api
import pandas as pd
df = pd.DataFrame({})

# Open a connection to the API
ocx = Api()

# Load in each dataset
for analysis in ocx.Classifications.where(success=True):
    # Convert the analysis to a DataFrame
    readcounts = analysis.table_df(key='tax_id', format='percent')
    df[analysis.sample.filename] = readcounts

print df.head()
```

Out[1]:

	SRR1748740	SRR1749835	SRR1748643	SRR1748647	SRR1748650
name					
Pseudomonas putida	5.55%	6.45%	0.0%	0.1%	0.0%


```
# Load the One Codex client library
from onecodex.api import Api
import pandas as pd
df = pd.DataFrame({})

# Open a connection to the API
ocx = Api()

# Load in each dataset
for analysis in ocx.Classifications.where(success=True):
    # Convert the analysis to a DataFrame
    readcounts = analysis.table_df(key='tax_id', format='percent')
    df[analysis.sample.filename] = readcounts

    # Add the 'material' metadata
    meta = analysis.sample.metadata
    df.ix['material', analysis.sample.filename] = meta['material']

# Make a boxplot
plot_species = ['Escherichia coli', 'Enterobacter cloacae']
df.boxplot(column=plot_species, by=['material'])
```

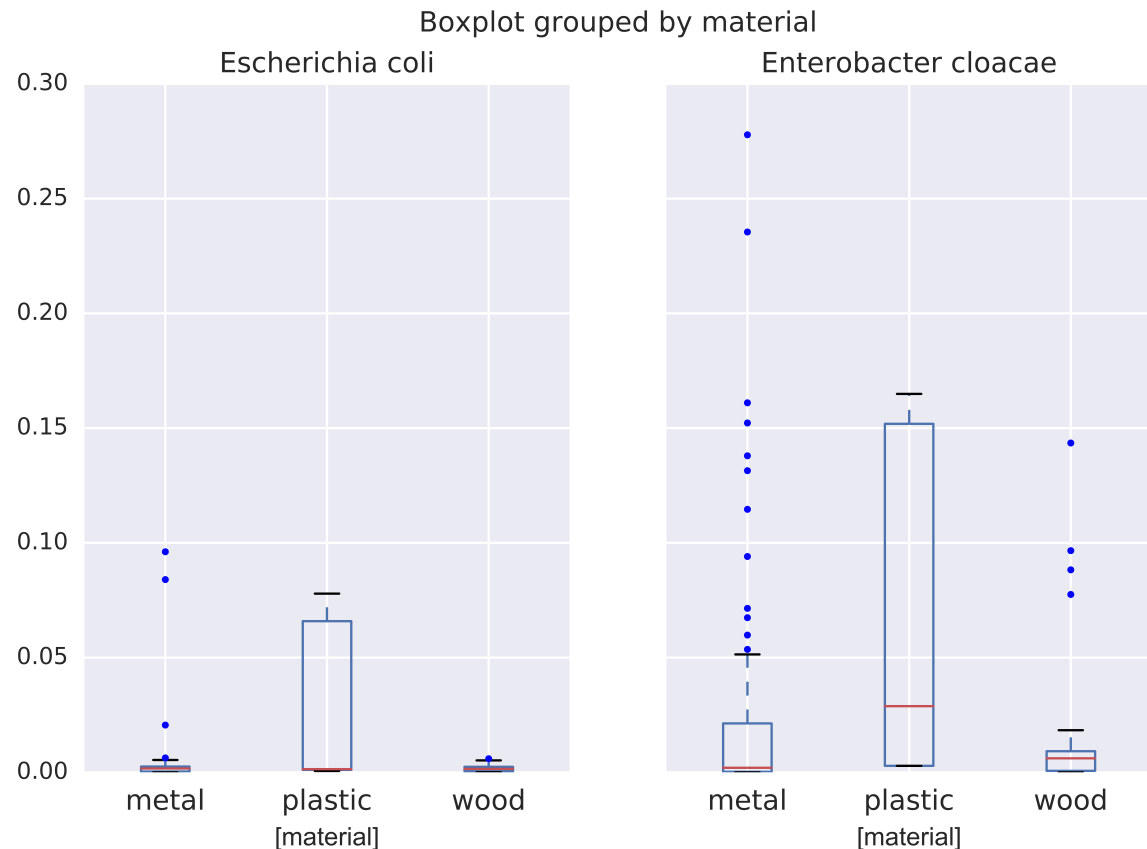
```
# Load the One Codex client library
from onecodex.api import Api
import pandas as pd
df = pd.DataFrame({})

# Open a connection to the API
ocx = Api()
```

```
# Load in each data
for analysis in ocx:
    # Convert the analysis to a DataFrame
    readcounts = analysis.readcounts
    df[analysis.sample_id] = readcounts

    # Add the 'material' column
    meta = analysis.metadata
    df.ix['material'] = meta['material']

# Make a boxplot
plot_species = ['Escherichia coli', 'Enterobacter cloacae']
df.boxplot(column='readcounts', by='material', facet_col=plot_species)
```



One Codex - Marker panel

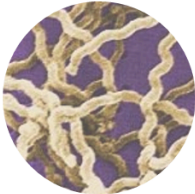
https://app.onecodex.com/examples

ONE CODEXMetaSUB

Sample Applications

Clinical Diagnostics

Leptospira from CSF
[Wilson et al., NEJM \(2014\)](#) »

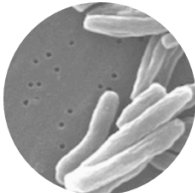


Metagenomic Results

Leptospira santarosai, identified from the cerebral spinal fluid (CSF) of a 14-year-old boy with severe combined immunodeficiency

Clinical Diagnostics

M. tuberculosis from sputum
[Doughty et al., PeerJ \(2014\)](#) »

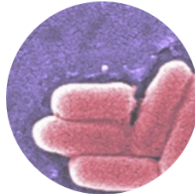


Metagenomic Results

Mycobacterium Tuberculosis identified directly from sputum samples from patients in The Gambia

Clinical Diagnostics

STEC from Stool
[Loman et al., JAMA \(2013\)](#) »

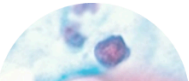


Metagenomic Results

Shiga-toxin-producing *Escherichia coli* (STEC) directly identified from stool samples during the European 2011 O104:H4 *E. coli* outbreak

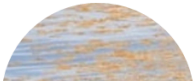
Clinical Diagnostics

Enriched *C. trachomatis*
[Seth-Smith et al., Genome Research \(2013\)](#) »



Marine Microbiology

Sargasso Sea Study
[Venter et al., Science \(2004\)](#) »



Built Environment

New York City Subway
[Afshinnekoo et al., Cell Systems \(2015\)](#) »



Try it out!

www.onecodex.com

hello@onecodex.com

[@onecodex](#)