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
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
Biobanks...
Biobanks...

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

Disadvantages: Fragile, finite/depletable, expensive to use and maintain
... to databanks?
Complement biobanks: (Relatively) durable, distributable, (nearly) free analysis & re-analysis
... to databanks?
... to databanks?

Requirements: Security, scalability, ease-of-use (and extensibility), and reproducibility
... to databanks?
...to databanks?

@HWI-D00151:75:H9AQQADXX:2:1101:11980:2239 2:N:0:GGACTCTTATCCTCT
AATTCAACGATACGCCAGCCCTTGCAGCAAACGGCTGTACCAGATCTACCCGTCTTTACGTGCGCCGAGGCCATTGCACCAGAAGATGCAGCTTAAGG
+  
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

@HWI-D00151:75:H9AW7ADXX:2:1101:14056:2237 2:N:0:GGACTCTTATCCTCT
GTATTGTAGAGCTGTGATTAGCGTGNTGATGGCCAAATGAAGGCGACCGCGCTCGCAGGCTGCAGGACGGCTGAGCTGCAGCCTGCGTCTGCAGCTG
+  
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

@HWI-D00151:75:H9AQQADXX:2:1101:18423:2248 2:N:0:GGACTCTTATCCTCT
GGCAATGGACAATGTATCCGCGGGTCGCCGTAGTTGGCAATGGACAATGTATCCGCGGGTCGCCGTAGTTGGCAATGGACAATGTATCCGCGGGTCGCC
+  
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

@HWI-D00151:75:H9AQQADXX:2:1101:1849:2248 2:N:0:GGACTCTTATCCTCT
GGCAATGGACAATGTATCCGCGGGTCGCCGTAGTTGGCAATGGACAATGTATCCGCGGGTCGCCGTAGTTGGCAATGGACAATGTATCCGCGGGTCGCC
+  
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF

@HWI-D00151:75:H9AQQADXX:2:1101:1849:2248 2:N:0:GGACTCTTATCCTCT
GGCAATGGACAATGTATCCGCGGGTCGCCGTAGTTGGCAATGGACAATGTATCCGCGGGTCGCCGTAGTTGGCAATGGACAATGTATCCGCGGGTCGCC
+  
BBBFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFFF
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
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 cheiliphaga, Lysinibacillus xylanilyticus, bacterium MS4, Pantoea sp. PSNIH1

**Sample Composition**

<table>
<thead>
<tr>
<th>Name</th>
<th>Estimated Abundance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Leclercia adecarboxylata</td>
<td>36.79%</td>
</tr>
<tr>
<td>Stenotrophomonas maltophilia</td>
<td>22.30%</td>
</tr>
<tr>
<td>[Pseudomonas] geniculata</td>
<td>7.03%</td>
</tr>
<tr>
<td>Lysinibacillus xylanilyticus</td>
<td>3.03%</td>
</tr>
<tr>
<td>Stenotrophomonas sp. RIT309</td>
<td>3.01%</td>
</tr>
<tr>
<td>Acinetobacter sp. UNC436CL71CviS28</td>
<td>2.93%</td>
</tr>
<tr>
<td>Pantoea sp. PSNIH1</td>
<td>2.31%</td>
</tr>
</tbody>
</table>
## Panel: AMR Panel - Sample.fa

### Ansamycins

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Status</th>
<th>Marker Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>arr-3</td>
<td>AY038837-555:1007 Marker length: 453 bp</td>
<td>Absent</td>
<td>79.2 % coverage</td>
</tr>
</tbody>
</table>

### Glycopeptides

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Status</th>
<th>Marker Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>van(C3)</td>
<td>AY033764-26:1078 Marker length: 1053 bp</td>
<td>Absent</td>
<td>0.0 % coverage</td>
</tr>
<tr>
<td>van(Y)</td>
<td>AF516335-5821:6366 Marker length: 546 bp</td>
<td>Absent</td>
<td>0.0 % coverage</td>
</tr>
<tr>
<td>van(D4)</td>
<td>AF277571-1262:2293 Marker length: 1032 bp</td>
<td>Absent</td>
<td>0.0 % coverage</td>
</tr>
<tr>
<td>van(B)</td>
<td>U00456-62:1090 Marker length: 1029 bp</td>
<td>Absent</td>
<td>0.0 % coverage</td>
</tr>
<tr>
<td>van(C1)</td>
<td>AF162694-1411:2442 Marker length: 1032 bp</td>
<td>Absent</td>
<td>0.0 % coverage</td>
</tr>
</tbody>
</table>

### Beta-lactams

<table>
<thead>
<tr>
<th>Name</th>
<th>Description</th>
<th>Status</th>
<th>Marker Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>blaTEM-1</td>
<td>AF309824-119:979 Marker length: 861 bp</td>
<td>Present</td>
<td>100 % coverage</td>
</tr>
<tr>
<td>blaLAT-1-LAT-4</td>
<td>X78117-142:1287 Marker length: 1146 bp</td>
<td>Probable</td>
<td>99.4 % coverage</td>
</tr>
<tr>
<td>blaOXA-1</td>
<td>J02967-1359:2189 Marker length: 831 bp</td>
<td>Absent</td>
<td>91.5 % coverage</td>
</tr>
</tbody>
</table>
v1 API

```json
[
  {
    "studi": "/api/v1/samples/404d876f78a74f5b",
    "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/235f286b0e07045ed"
    },
    "primary_analysis": {
      "$ref": "/api/v1/analyses/54edd12877664bb8"
    },
    "project": {
      "$ref": "/api/v1/projects/4e69ca548784edb"
    },
    "public": false,
    "size": 340164724,
    "starred": false,
    "tags": [
      {
        "$ref": "/api/v1/tags/b93e3abc3021420f"
      },
      {
      }
    ]
  }
]
```
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_sequence": "2015-01-30T00:00:00+00:00",
      "description": "city - subway - metal",
      "external_sample_id": null,
      "library_type": "WGS",
      "latitude": 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

```json
{
    "$uri": "/api/v1/projects/4e69cae548784ed8",
    "description": "PathoMAP Sequencing Project",
    "name": "PathoMAP",
    "owner": {
        "$ref": "/api/v1/users/235f286b0c7456d"
    },
    "permissions": [
        "can_administer",
        "can_download_files",
        "can_incur_charges",
        "can_edit_metadata",
        "can_add_files",
        "can_view_files"
    ],
    "project_name": "pathomap",
    "public": false
}
```

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.
v1 API

[ ]

"url": "/api/v1/samples/84c873edd2b84c25",
"created_at": "2015-08-07T01:26:23.196565+08:00",
"filename": "SRR1748569.fastq.gz",
"indexed": false,
"metadata": {
  "$ref": "/api/v1/metadata/e1c6a6279ff24440"
},
"owner": {
  "$ref": "/api/v1/users/235f286b8e7043ed"
},
"primary_analysis": {
  "$ref": "/api/v1/analyses/68edcf6c62964b54"
},
"project": {
  "$ref": "/api/v1/projects/4e69cae548784eddb"
},
"public": false,
"size": 436780487,
"starred": false,
"tags": [
  {
    "$ref": "/api/v1/tags/b93e3abc3021420f"
  },
  {
    "$ref": "/api/v1/tags/33c3abc3021420f"
  }
]
```python
# Load the One Codex client library
from onecodex.api import Api
import pandas as pd
def = 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]:

<table>
<thead>
<tr>
<th></th>
<th>SRR1748740</th>
<th>SRR1749835</th>
<th>SRR1748643</th>
<th>SRR1748647</th>
<th>SRR1748650</th>
</tr>
</thead>
<tbody>
<tr>
<td>name</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Pseudomonas putida</td>
<td>5.55%</td>
<td>6.45%</td>
<td>0.0%</td>
<td>0.1%</td>
<td>0.0%</td>
</tr>
</tbody>
</table>
```
# 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 dataset
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 = df[df['material'].isin(['material'])]

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

Clinical Diagnostics

- **Leptospira from CSF**
  - *Leptospira santarosai*, identified from the cerebral spinal fluid (CSF) of a 14-year-old boy with severe combined immunodeficiency
  - Wilson et al., NEJM (2014)

- **M. tuberculosis from sputum**
  - *Mycobacterium Tuberculosis* identified directly from sputum samples from patients in The Gambia
  - Doughty et al., PeerJ (2014)

- **STEC from Stool**
  - Shiga-toxin-producing Escherichia Coli (STEC) directly identified from stool samples during the European 2011 O104:H4 *E. coli* outbreak
  - Loman et al., JAMA (2013)

Clinical Diagnostics

- **Enriched C. trachomatis**
  - Seth-Smith et al., Genome Research (2013)

Sargasso Sea Study

- **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