TGAC - AllBio 2014

bmpvieira.com/allbio14

Who

Bruno Vieira | @bmpvieira

Phd Student @ Queen Mary University of London

Bioinformatics and Population Genomics

Supervisor:

Yannick Wurm | @yannick__

Before

2004-2009

Master in Human Biology and Environment
Licentiate in Cell Biology and Biotechnology

2009-2013

Bioinformatician and SysAdmin

Computational Biology & Population Genomics Group

2012-2013

Full Stack Web Developer - Built everything with Node.js, Express.js, Bootstrap, MongoDB and Redis

2013

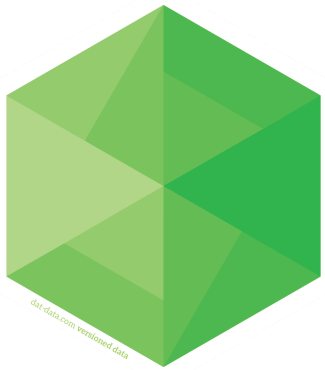
Full Stack Web Developer - Worked on integration with LinkedIn API

What

Bionode.io - *Modular and universal bioinformatics*

Pipeable UNIX command line tools and JavaScript / Node.js APIs for bioinformatic analysis workflows on the server and browser.

Collaborates with BioJS - *Represent biological data on the web*



Dat - *Build data pipelines*

Provides a streaming interface between every file format and data storage backend. *"git for data"*

Why Bionode / Node.js?

- Reusable, small and tested modules
- Same language everywhere (JavaScript)
- JavaScript is fast enough
- Package Manager that works (NPM)
- Huge number modules (93327, 199/day)
- Use other JS projects (Dat, BioJS, NoFlo)
- Streams applies well to Bioinformatics
- Easy to write CLI wrappers for Streams
- Possible to write Desktop GUI apps in JS

Reusable, small and tested

bionode-ncbi

> Node.js module for working with the NCBI API (aka e-utils).

Install

Install `bionode-ncbi` with npm:

```
$ npm install bionode-ncbi
```

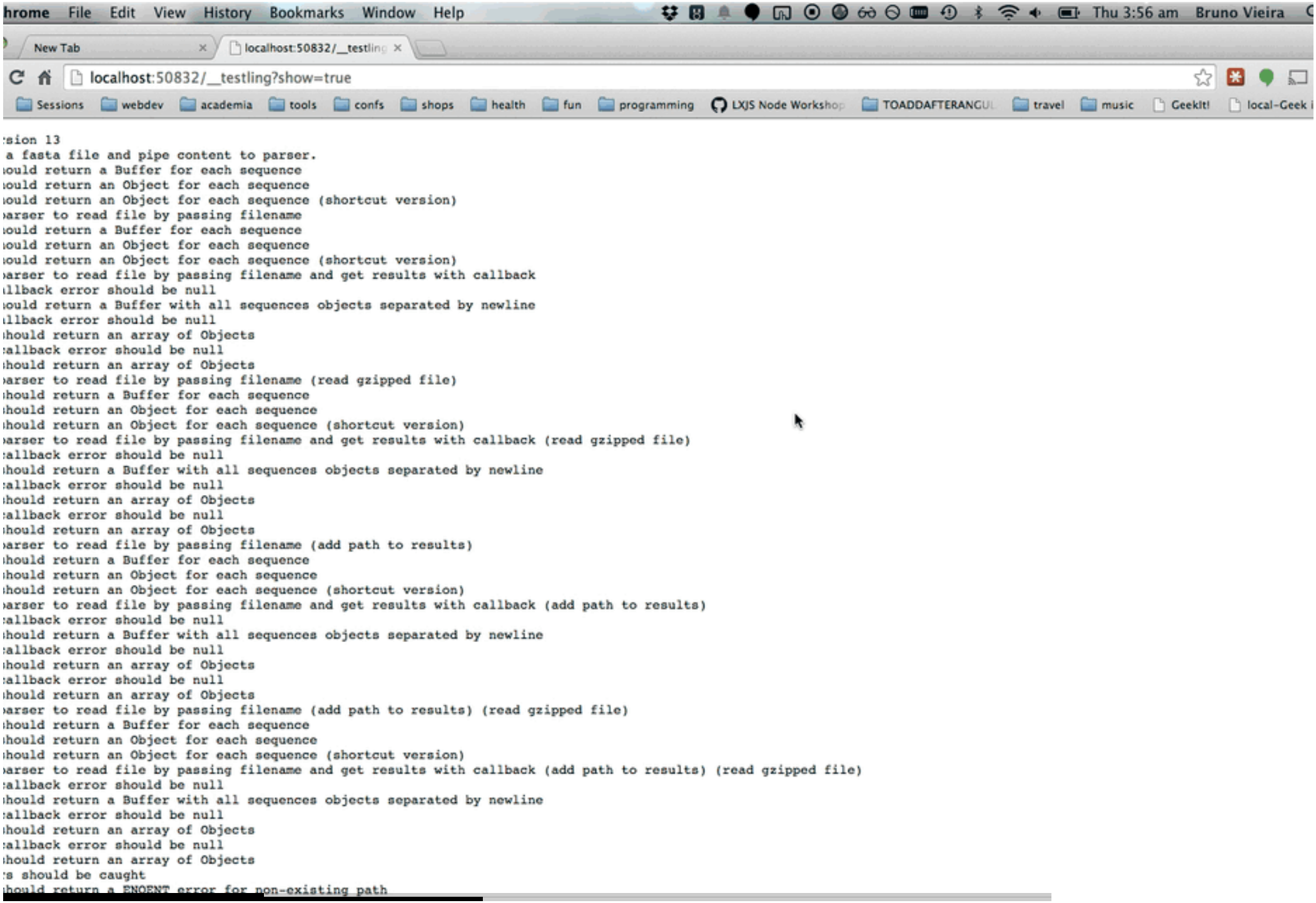
To use it as a command line tool, you can install it globally by adding `-g` .

Usage

If you are using `bionode-ncbi` with Node.js, you can require the module:

```
var ncbi = require('bionode-ncbi')
ncbi.search('sra', 'solenopsis').on('data', console.log)
```


JAVASCRIPT
JAVASCRIPT EVERYWHERE!
makeameme.org

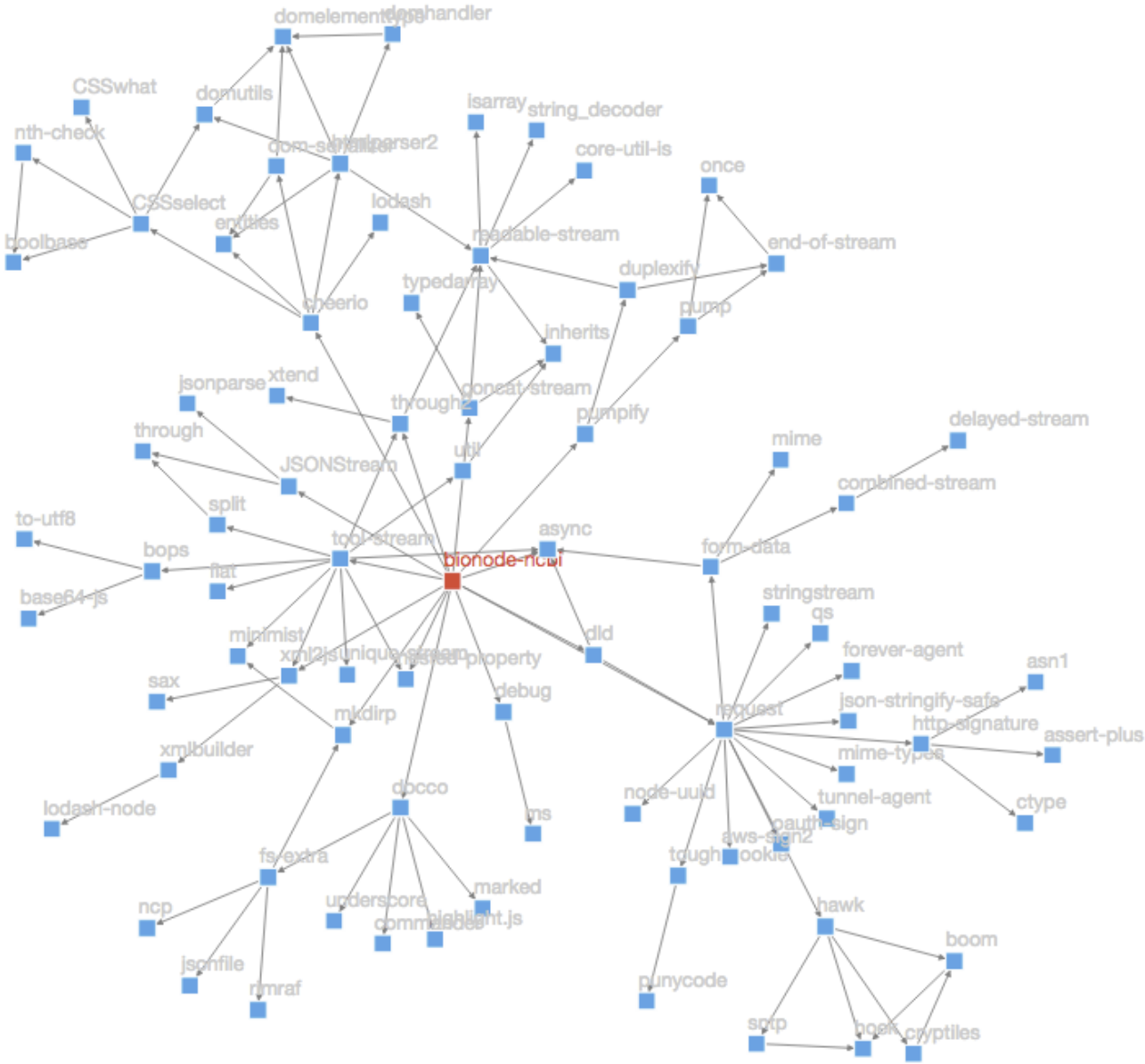

hrome File Edit View History Bookmarks Window Help
Thu 3:56 am Bruno Vieira
New Tab
localhost:50832/__testling
localhost:50832/__testling?show=true
Sessions webdev academia tools confs shops health fun programming LXJS Node Workshop TOADDAFTERANGUL travel music Geekit! local-Geek
sion 13
a fasta file and pipe content to parser.
ould return a Buffer for each sequence
ould return an Object for each sequence
ould return an Object for each sequence (shortcut version)
arser to read file by passing filename
ould return a Buffer for each sequence
ould return an Object for each sequence
ould return an Object for each sequence (shortcut version)
arser to read file by passing filename and get results with callback
llback error should be null
ould return a Buffer with all sequences objects separated by newline
llback error should be null
hould return an array of Objects
allback error should be null
hould return an array of Objects
arser to read file by passing filename (read gzipped file)
hould return a Buffer for each sequence
hould return an Object for each sequence
hould return an Object for each sequence (shortcut version)
arser to read file by passing filename and get results with callback (read gzipped file)
allback error should be null
hould return a Buffer with all sequences objects separated by newline
allback error should be null
hould return an array of Objects
allback error should be null
hould return an array of Objects
arser to read file by passing filename (add path to results)
hould return a Buffer for each sequence
hould return an Object for each sequence
hould return an Object for each sequence (shortcut version)
arser to read file by passing filename and get results with callback (add path to results)
allback error should be null
hould return a Buffer with all sequences objects separated by newline
allback error should be null
hould return an array of Objects
allback error should be null
hould return an array of Objects
arser to read file by passing filename (add path to results) (read gzipped file)
hould return a Buffer for each sequence
hould return an Object for each sequence
hould return an Object for each sequence (shortcut version)
arser to read file by passing filename and get results with callback (add path to results) (read gzipped file)
allback error should be null
hould return a Buffer with all sequences objects separated by newline
allback error should be null
hould return an array of Objects
allback error should be null
hould return an array of Objects
s should be caught
hould return a ENOENT error for non-existing path

JavaScript is fast

Package Manager that works

```
npm install bionode
npm install bionode -g
npm test
npm start
npm run test-browser
npm run build-docs
npm init
npm publish
```

search bionode-ncbi Go



Package info : **bionode-ncbi**

Graph Info

Dependencies graph of **bionode-ncbi** has nodes and **109** edges.

Maintainers


feedic x 10


isaacs x 7


mikeal x 6


hueniverse x 6


substack x 5


mafintosh x 4


felixge x 3


dominictarr x 3


rvagg x 3

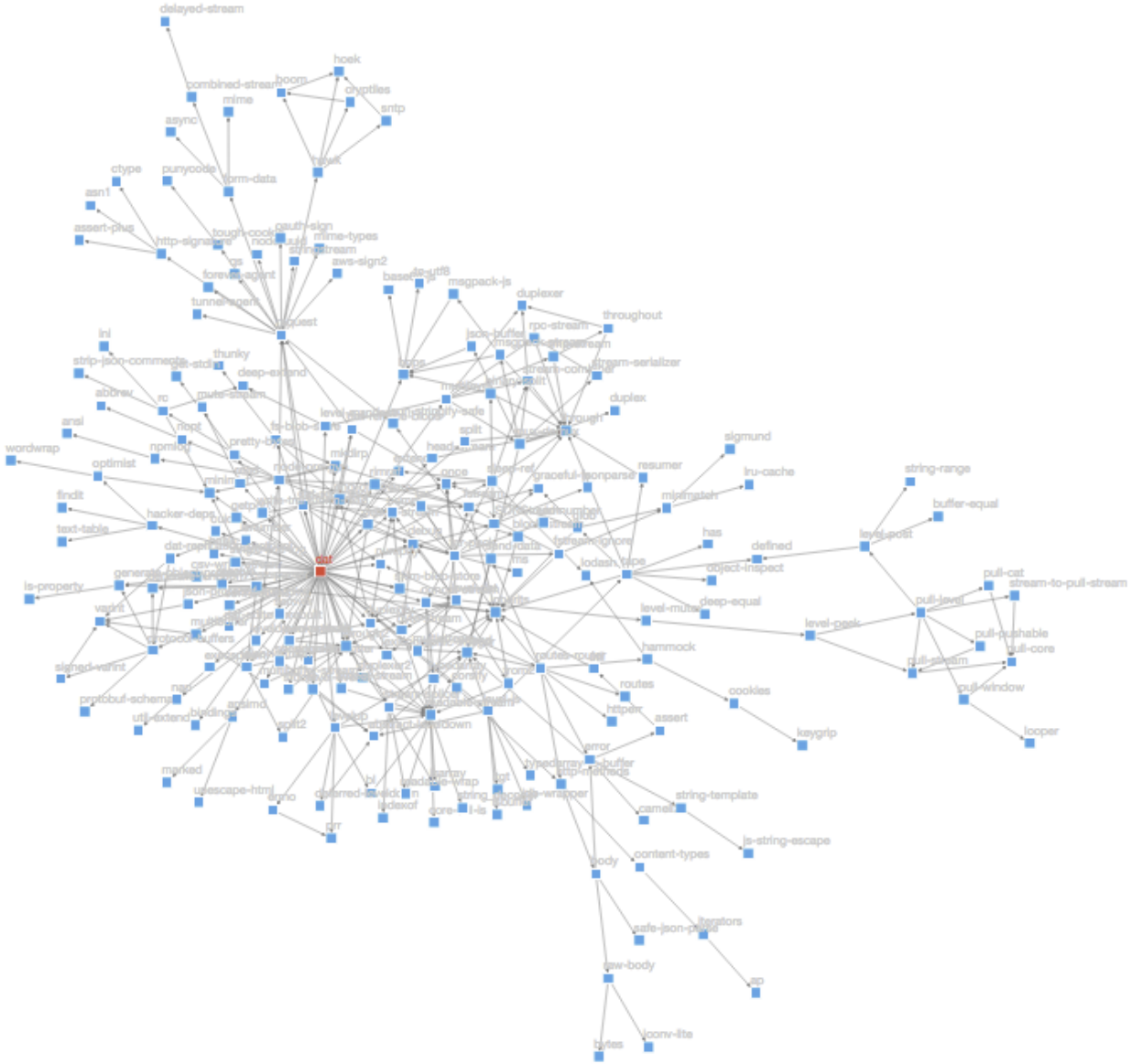

mathias x 3


mcavage x 3

CommonJS pattern

```
 // awesome-lib/index.js
module.exports = function() {
  return "Small modules everywhere"
}

// myscript.js
var awesome = require('awesome-lib')
awesome()
```

search dat Go



Package info : **dat**

Graph Info

Dependencies graph of **dat** has **207** nodes and **350** edges.

Maintainers

 mafintosh x 28

 dominictarr x 26

 isaacs x 24

 substack x 24

 raynos x 13

 rvagg x 10

 mikeal x 9

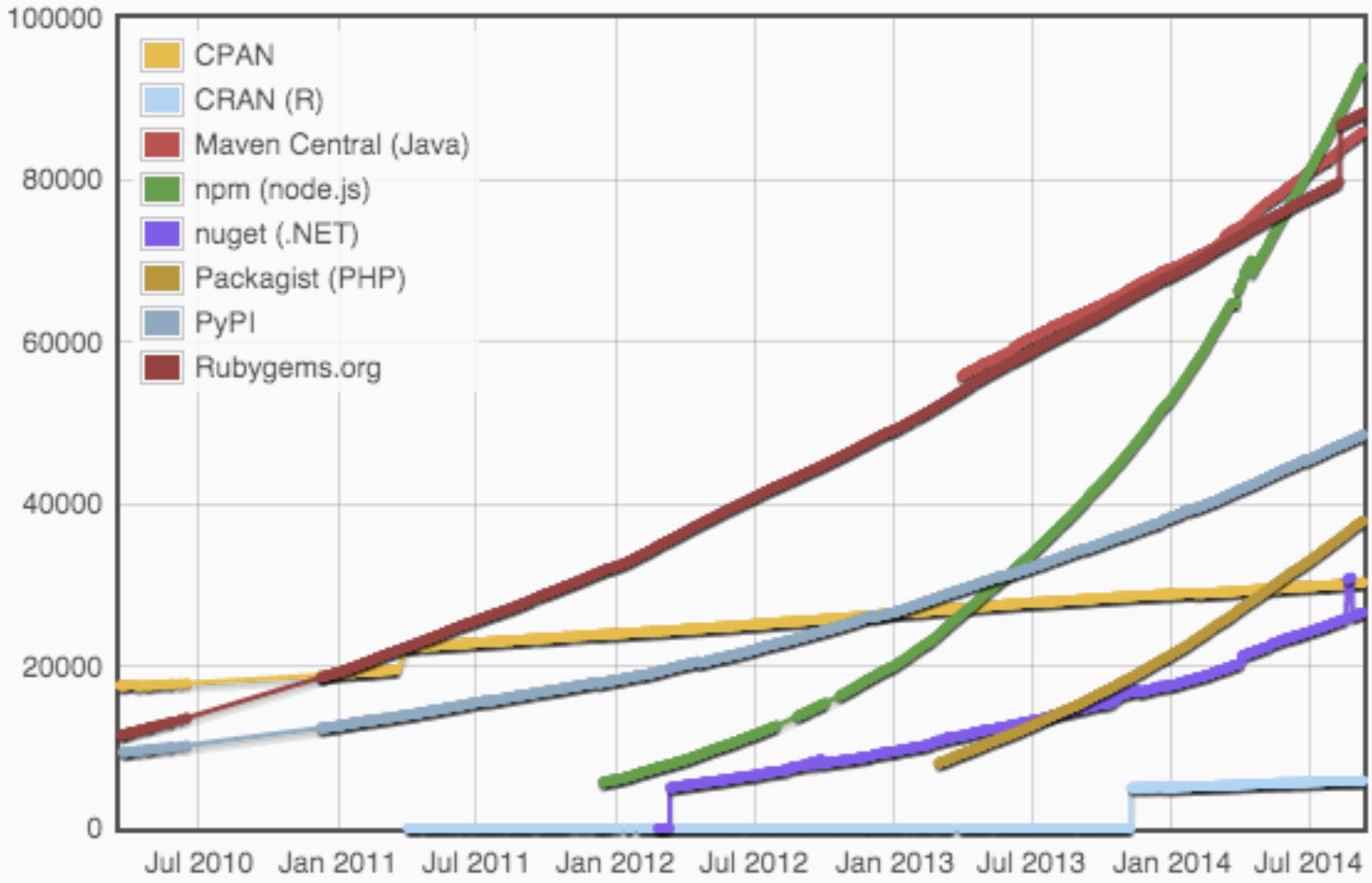
 maxogden x 7

 tootallnate x 6

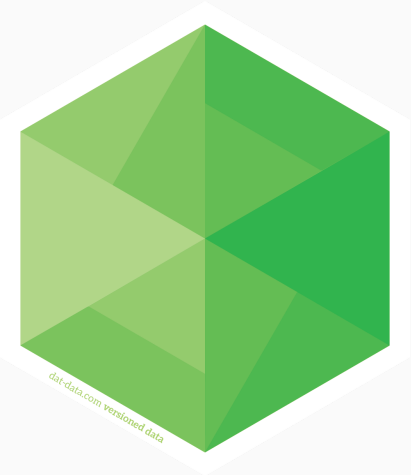
 hueniverse x 6

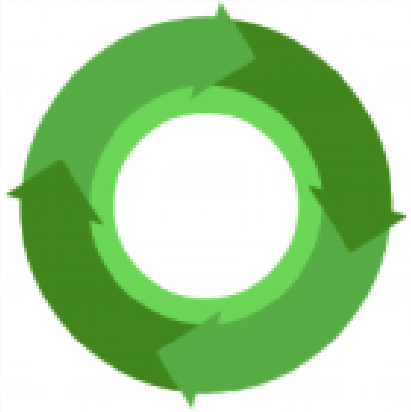
 maxogden x 6

Module counts



Benefit from other JS projects

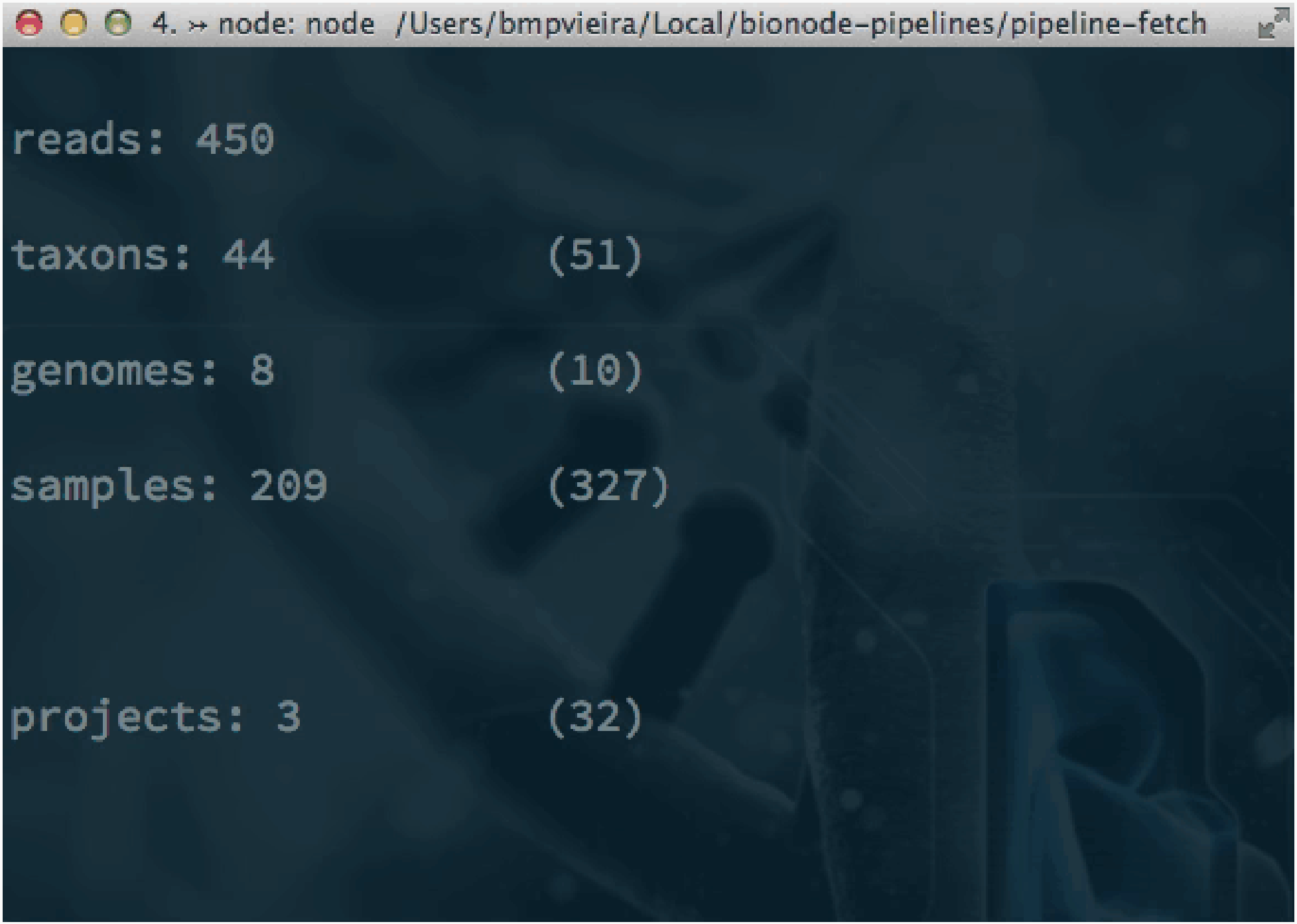


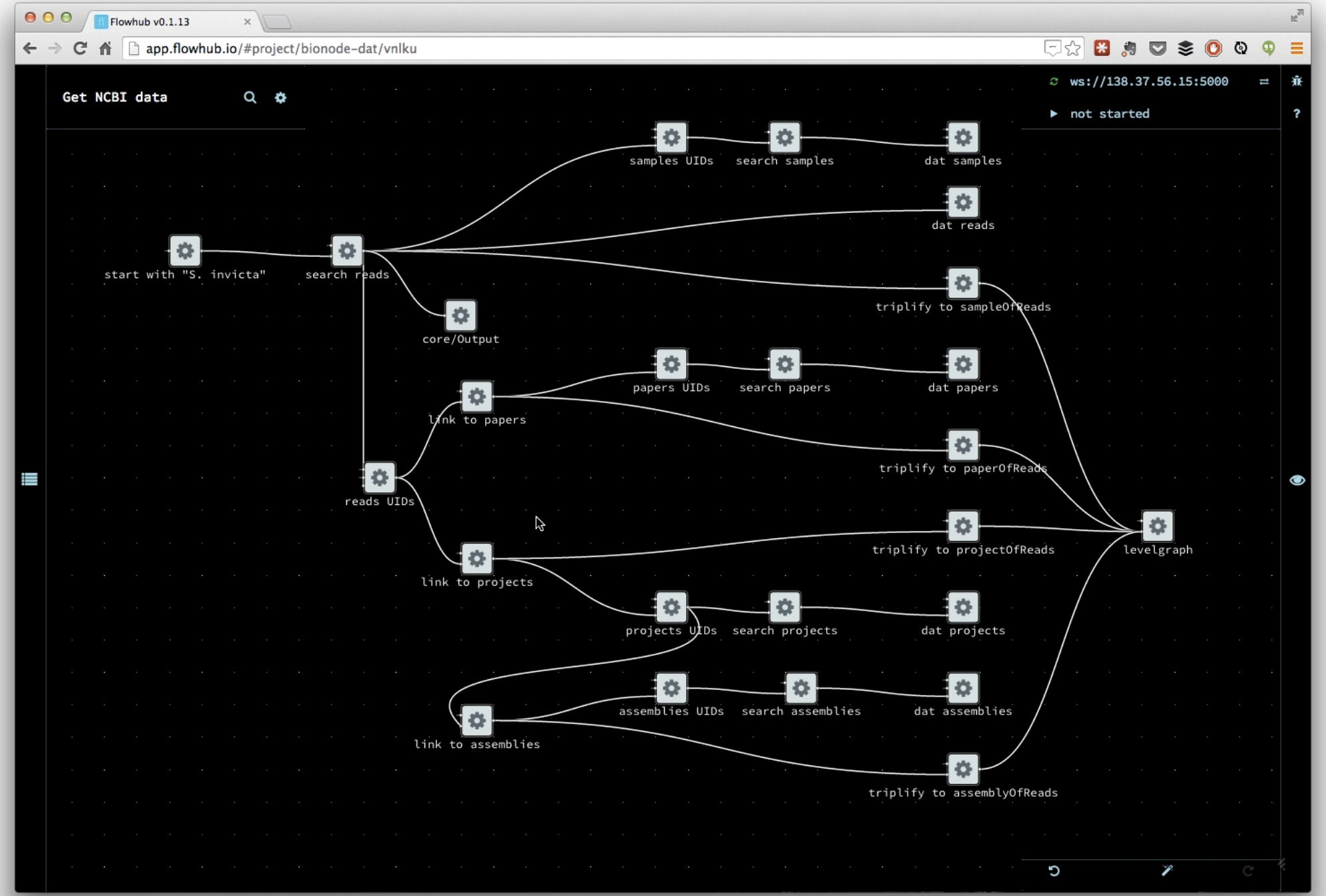


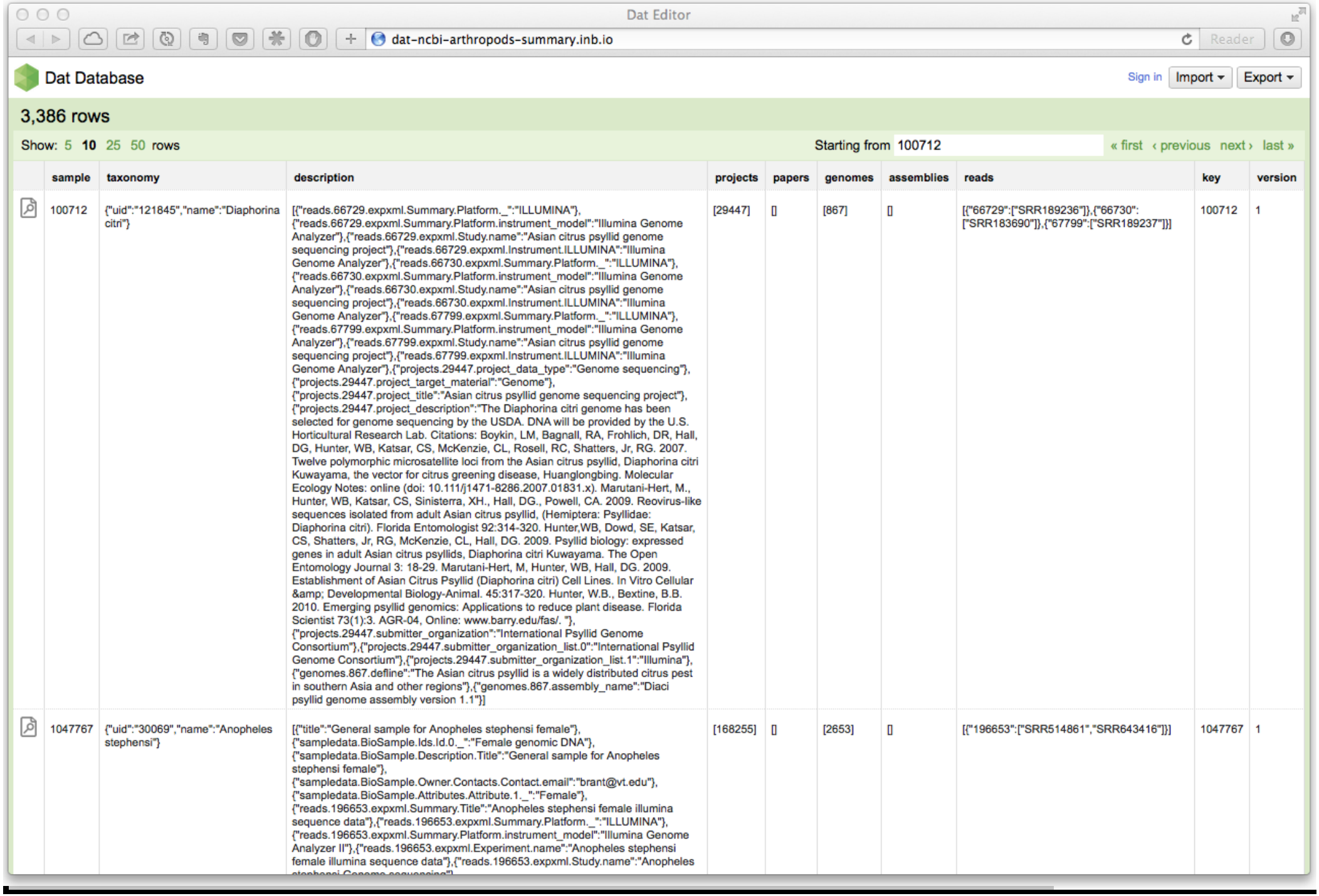
Streams

@substack: "It's turtles all the way down!"

```
var fork1 = through.obj()
var fork2 = through.obj()

ncbi
  .search('sra', 'Solenopsis invicta')
  .pipe(fork1)
  .pipe(dat.reads)

fork1
  .pipe(tool.extractProperty('expxml.Biosample.id'))
  .pipe(ncbi.search('biosample'))
  .pipe(dat.samples)

fork1
  .pipe(tool.extractProperty('uid'))
  .pipe(fork2)


fork2
  .pipe(ncbi.link('sra', 'pubmed'))
  .pipe(ncbi.search('pubmed'))
  .pipe(dat.papers)
```


4. ↣ node: node /Users/bmpvieira/Local/bionode-pipelines/pipeline-fetch
reads: 450
taxons: 44 (51)
genomes: 8 (10)
samples: 209 (327)
projects: 3 (32)


Flowhub v0.1.13
app.flowhub.io/#project/bionode-dat/vnlku
Get NCBI data
ws://138.37.56.15:5000
not started
samples UIDs
search samples
dat samples
dat reads
start with "S. invicta"
search reads
triplify to sampleOfReads
core/Output
papers UIDs
search papers
dat papers
link to papers
triplify to paperOfReads
reads UIDs
triplify to projectOfReads
levelgraph
link to projects
projects UIDs
search projects
dat projects
assemblies UIDs
search assemblies
dat assemblies
link to assemblies
triplify to assemblyOfReads



	sample	taxonomy	description	projects	papers	genomes	assemblies	reads	key	version
	100712	{"uid":"121845","name":"Diaphorina citri"}	[{"reads.66729.expxml.Summary.Platform._":"ILLUMINA"}, {"reads.66729.expxml.Summary.Platform.instrument_model":"Illumina Genome Analyzer"},{"reads.66729.expxml.Study.name":"Asian citrus psyllid genome sequencing project"},{"reads.66729.expxml.Instrument.ILLUMINA":"Illumina Genome Analyzer"},{"reads.66730.expxml.Summary.Platform._":"ILLUMINA"}, {"reads.66730.expxml.Summary.Platform.instrument_model":"Illumina Genome Analyzer"},{"reads.66730.expxml.Study.name":"Asian citrus psyllid genome sequencing project"},{"reads.66730.expxml.Instrument.ILLUMINA":"Illumina Genome Analyzer"},{"reads.67799.expxml.Summary.Platform._":"ILLUMINA"}, {"reads.67799.expxml.Summary.Platform.instrument_model":"Illumina Genome Analyzer"},{"reads.67799.expxml.Study.name":"Asian citrus psyllid genome sequencing project"},{"reads.67799.expxml.Instrument.ILLUMINA":"Illumina Genome Analyzer"},{"projects.29447.project_data_type":"Genome sequencing"}, {"projects.29447.project_target_material":"Genome"}, {"projects.29447.project_title":"Asian citrus psyllid genome sequencing project"}, {"projects.29447.project_description":"The Diaphorina citri genome has been selected for genome sequencing by the USDA. DNA will be provided by the U.S. Horticultural Research Lab. Citations: Boykin, LM, Bagnall, RA, Frohlich, DR, Hall, DG, Hunter, WB, Katsar, CS, McKenzie, CL, Rosell, RC, Shatters, Jr, RG. 2007. Twelve polymorphic microsatellite loci from the Asian citrus psyllid, Diaphorina citri Kuwayama, the vector for citrus greening disease, Huanglongbing. Molecular Ecology Notes: online (doi: 10.111/j1471-8286.2007.01831.x). Marutani-Hert, M., Hunter, WB, Katsar, CS, Sinisterra, XH., Hall, DG., Powell, CA. 2009. Reovirus-like sequences isolated from adult Asian citrus psyllid, (Hemiptera: Psyllidae: Diaphorina citri). Florida Entomologist 92:314-320. Hunter,WB, Dowd, SE, Katsar, CS, Shatters, Jr, RG, McKenzie, CL, Hall, DG. 2009. Psyllid biology: expressed genes in adult Asian citrus psyllids, Diaphorina citri Kuwayama. The Open Entomology Journal 3: 18-29. Marutani-Hert, M, Hunter, WB, Hall, DG. 2009. Establishment of Asian Citrus Psyllid (Diaphorina citri) Cell Lines. In Vitro Cellular & Developmental Biology-Animal. 45:317-320. Hunter, W.B., Bextine, B.B. 2010. Emerging psyllid genomics: Applications to reduce plant disease. Florida Scientist 73(1):3. AGR-04, Online: www.barry.edu/fas/. "}, {"projects.29447.submitter_organization":"International Psyllid Genome Consortium"},{"projects.29447.submitter_organization_list.0":"International Psyllid Genome Consortium"},{"projects.29447.submitter_organization_list.1":"Illumina"}, {"genomes.867.defline":"The Asian citrus psyllid is a widely distributed citrus pest in southern Asia and other regions"},{"genomes.867.assembly_name":"Diaci psyllid genome assembly version 1.1"}]	[29447]	[]	[867]	[]	[{"66729":["SRR189236"]},{"66730":["SRR183690"]},{"67799":["SRR189237"]}]	100712	1
	1047767	{"uid":"30069","name":"Anopheles stephensi"}	[{"title":"General sample for Anopheles stephensi female"}, {"sampledata.BioSample.Ids.Id.0._":"Female genomic DNA"}, {"sampledata.BioSample.Description.Title":"General sample for Anopheles stephensi female"}, {"sampledata.BioSample.Owner.Contacts.Contact.email":"brant@vt.edu"}, {"sampledata.BioSample.Attributes.Attribute.1._":"Female"}, {"reads.196653.expxml.Summary.Title":"Anopheles stephensi female illumina sequence data"},{"reads.196653.expxml.Summary.Platform._":"ILLUMINA"}, {"reads.196653.expxml.Summary.Platform.instrument_model":"Illumina Genome Analyzer II"},{"reads.196653.expxml.Experiment.name":"Anopheles stephensi female illumina sequence data"},{"reads.196653.expxml.Study.name":"Anopheles	[168255]	[]	[2653]	[]	[{"196653":["SRR514861","SRR643416"]}]	1047767	1

Command Line Interface

```
# Subset a fasta file to a particular sequence

cat sequences.fasta
| bionode-fasta
| grep "contig123"
| bionode-fasta --write > contig123.fasta


# Find the reads datasets used for the Solenopsis invicta assembly

bionode-ncbi search assembly Solenopsis invicta |
tool-stream extractProperty uid                 |
bionode-ncbi link assembly bioproject           |
tool-stream extractProperty destUID             |
bionode-ncbi link bioproject sra                |
tool-stream extractProperty destUID             |
bionode-ncbi urls sra                           |
dat import --json
```

Project status: available

- Data access:
  - ncbi
- Parsing
  - fasta
  - bbi
- Wrangling
  - seq
- Wrappers
  - sra
  - sam

Project status: down the line

- Data access:
  - ebi
  - ensembl
- Parsing
  - fastq
  - sam
  - vcf
  - gff
- Wrangling
  - quality control/stats
- Wrappers
  - blast

Try

generalhenry.com/data-plumber

Install

Node

```
# OSX
brew install node
```

```
# Ubuntu
sudo apt-get install nodejs npm
```

Bionode

```
npm install bionode
```

Thanks!

Acknowledgements:

@yannick__
@maxogden
@mafintosh
@alanmrice