

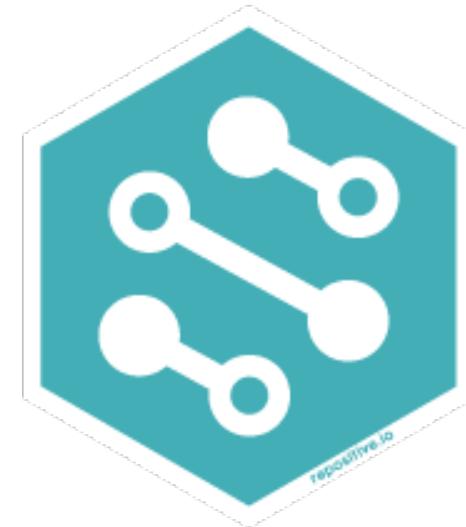
Organisers



Community



Sponsor



Research group



Venue



Friends



http://hack.bionode.io

Please join the *bionode-hack* github organisation

http://github.com/bionode-hack

github organisation homepage

http://gitter.im/bionode-hack/discussions

Instant messaging and general discussions

#bionodehack

#genomics

#nodejs

#campuslondon

@DNADigest

@bionode

@repository

@CampusLondon



Network: Campus Community

Password: dotheimpossible

#bionodehack
@DNADigest

#genomics
@bionode

#nodejs
@repositoryio

#campuslondon
@CampusLondon



<http://hack.bionode.io>

Please join the *bionode-hack* github organisation

<http://github.com/bionode-hack>

github organisation homepage

<http://gitter.im/bionode-hack/discussions>

Instant messaging and general discussions

#bionodehack

@DNADigest

#genomics

@bionode

#nodejs

@repositoryio

#campuslondon

@CampusLondon

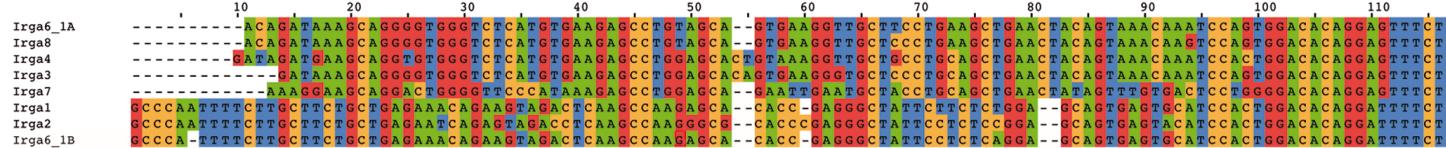
Who and what is DNAdigest?



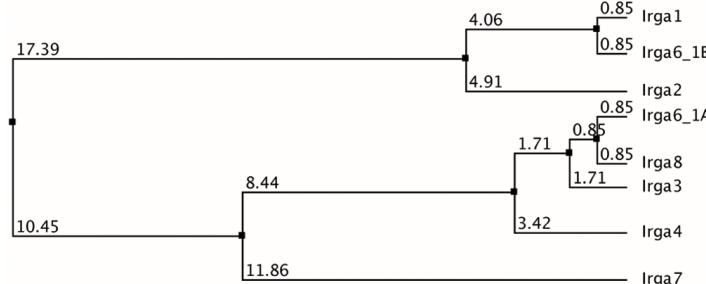
#BionodeHack

What is DNA data? Why care?

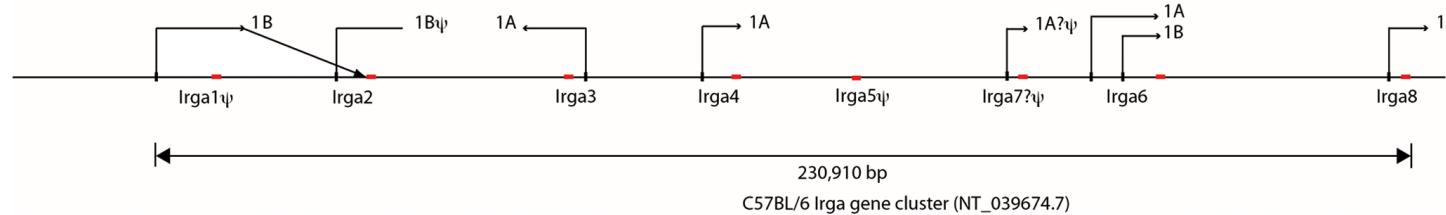
A



B



C



DNA interpretation is hard

All information comes from comparison:

- Cancer mutations
- Evolutionary trees
- Functions of the cell

Human DNA data is key for research in cancer and genetic diseases

We do events, workshops, hackathons



Latest events: #hackDNA

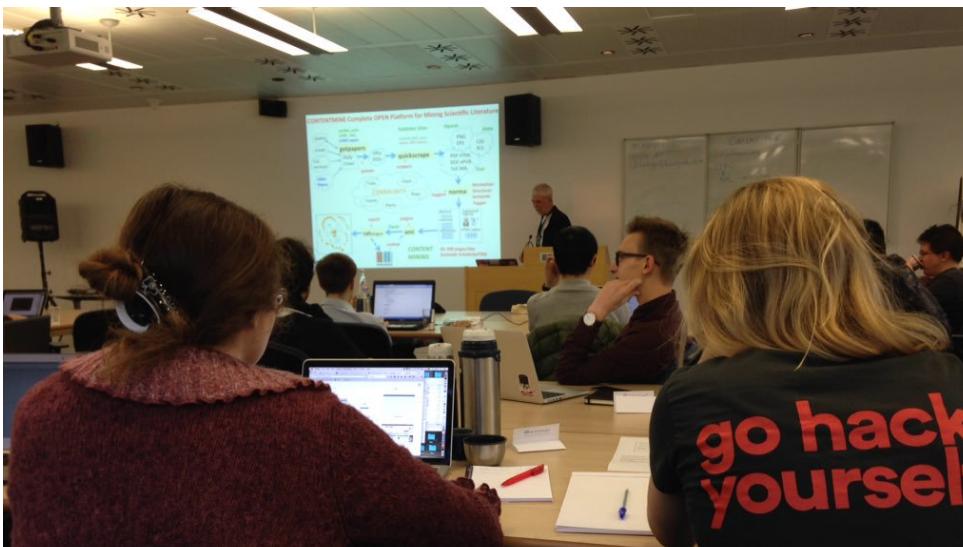
- Symposium on challenges of data sharing in genomics research
- ContentMine hackday
- Workshop Uni of Cambridge on how to find and access genomics data



Today: #BionodeHack

Aim of today

- Community learning experience
- Share best practices for data access
- Have fun! ☺



```
info: Searching using eupmc API
info: Found 277 open access results
retrieving results [=====
info: Done collecting results
Info: Duplicate records found: 275 unique results
Info: Saving result metadata
Info: Full EUPMC result metadata written to eupmc
info: Extracting fulltext HTML URL list (may not
info: Fulltext HTML URL list written to eupmc_full
info: Got XML URLs for 275 out of 275 results
info: Downloading fulltext XML files
Downloading files [=====
info: All XML downloads succeeded!
workshop@crunchbang ~/workshop/eupmc/eupmc %
```

What will focus on today



1. Improve the NCBI module - bionode-ncbi.
2. Prototyping new bionode modules for selected data sources.
3. Architecture challenges in bionode.io for designing a library of modules for fetching genomic data.

Data sources to work with today



Name	Repo name	Organism	Host	Data type	Access
Expression Atlas	bionode-gxa	Multiple	EBI	Expression data	Table / API
European variation archive	bionode-eva	Multiple	EBI	Variant data	API
Gene expression omnibus	bionode-ncbi (geo)	Multiple	NCBI	Expression data	API
Database of Genotypes and Phenotypes	bionode-ncbi (dbgap)	Human	NCBI	Controlled access genomic data	API
Sequence read archive	bionode-ncbi (sra)	Multiple	NCBI	Raw sequence data	API

Data sources to work with today



Name	Repo name	Organism	Host	Data type	Access
ENCODE	bionode-encode	Human	Encode	NGS of functional DNA elements	Table
MethyCancer	bionode-mc	Human	Chinese academy of sciences	Methylation data	FTP
Mouse tumor biology	bionode-mtb	PDX / Mouse	Jackson Labs	Patient derived xenograft data	Table
Expression Atlas	bionode-gxa	Multiple	EBI	Expression data	Table / API
Personal genome project	bionode-pgp	Human	Harvard	Personal genotypes	Table
Database of variants	bionode-ncbi (dbvar)	Multiple	NCBI	Variant data	FTP / API

#BionodeHack

This repository Search

Pull requests Issues Gist

Watch 1 Star 0 Fork 0

bionode-hack / bionode-eva

Code Issues 0 Pull requests 0 Wiki Pulse Graphs Settings

No description or website provided. — Edit

1 commit

1 branch

0 releases

0 contributors

Branch: master

New pull request

Create new file

Upload files

Find file

Clone or download

Richard Shaw High-level description of dataset metadata retrieval

Latest commit b0e9cf5 a day ago

example_responses

High-level description of dataset metadata retrieval

a day ago

README.md

High-level description of dataset metadata retrieval

a day ago

README.md

bionode-eva

European Variant Archive (EVA)

Homepage: <http://www.ebi.ac.uk/eva/> API notes: <https://github.com/ebivariation/eva/wiki>

(The instructions below are for EVA but, according to the wiki, if the argument 'structural=true' is added to the initial queries DGVA is searched instead.)

Steps to metadata

1: How do I find the set of (human) datasets that are available?

First list the Studies for human (also have to specify reference version but currently only grch37 returns results; there are none for grch38):- http://www.ebi.ac.uk/eva/webservices/rest/v1/meta/studies/list?species=hsapiens_grch37 results in a list of (currently 13) studies: https://github.com/bionode-hack/bionode-eva/blob/master/example_responses/studies_list_hsapiens_grch37.json

Within the 'result' value, iterate over the studies extracting 'studyId'.

2: How do I request information on an individual dataset (or maybe a batch of datasets)?



#BionodeHack

What's in the README.md



- Brief description and main URL for the data source
- How to find the set of datasets that are available (and relevant, i.e. human) from this data source.
- How to request information on individual datasets (or maybe a batch of datasets)
- Where to find the dataset attributes and how to get them
- Accession number, URL, Title, Description, Technology, Assay type

Goal #1



The bionode module should provide functions to get all metadata (description) of datasets from the respective data source.

```
{  
    "experimentType": "MICROARRAY_ANY",  
    "experimentAccession": "E-GEO-10070",  
    "experimentDescription": "Gene Expression in MCF10A cells through  
    "lastUpdate": "03-07-2014",  
    "numberOfAssays": 13,  
    "numberOfContrasts": 3,  
    "species": [  
        "Homo sapiens"  
    ],  
    "kingdom": "animals",  
    "ensemblDB": "ensembl",  
    "experimentalFactors": [  
        "phenotype"  
    ],  
    "arrayDesigns": [  
        "A-AFFY-44"  
    ],  
    "arrayDesignNames": [  
        "Affymetrix GeneChip Human Genome U133 Plus 2.0 [HG-U133_Plus_2]  
    ]  
},
```

Goal #2



Extract from the the following entities into a JSON:

```
{  
    external_id : "E-GEO-10233",  
    title : "Transcription profiling by array ...",  
    description : "Transcription profiling by array of human PMA-activated MM6 cells ...",  
    tech : "A-AFFY-44",  
    assay : "Expression-Array",  
    url : "https://www.ebi.ac.uk/gxa/experiments/E-GEO-10233",  
    access : "open"  
}
```

Goal #3



Once the descriptions are fetched and mapped, develop:

- **Methods to download the raw datasets**
- **Methods for wrapping up API calls to raw data**

EVA supports querying alleles from a single position (returns JSON):

www.ebi.ac.uk/eva/webservices/rest/v1/segments/11:128446-128446/variants?species=hsapiens_grch37

EVA also supports variant query by gene:

www.ebi.ac.uk/eva/webservices/rest/v1/genes/PAX6/variants?species=hsapiens_grch37&limit=10

Teaming up



- Around 5-6 people per team (potentially subdivide in sub-teams)
- Who works in science
- Who worked with node.js and Streams
- Bioinformatics / Computational biology
- Who would like to learn node.js / streams

Feedback



- Before you leave today, we'd be grateful to get your feedback on this event. We want to provide the best events we can, and your input is essential. It only takes **15 seconds**, and the feedback form is available at: campus.co/thx1q
- In question 3 of the feedback survey, we would be very grateful if you could tell us what you liked and disliked about this hackathon, and why?
- This feedback goes both to Campus and DNAdigest



Thanks for joining

#BionodeHack