Aleksandra Perz
Wren Lab
Oklahoma Medical Research Foundation

Overview

- MNEMONIC
  - Goals
  - Data acquisition and processing
  - Results and current state
- PubQC
  - Goals
  - Results so far

There is knowledge out there: objectives

Bring in the already available knowledge to interpret a new finding

Find samples that are similar:
learn something new about
the finding based on the literature

Compare shifts observed
between conditions:
find samples that exhibit similar
changes in taxonomic abundance

Predict characteristics
of the finding:
confirm the finding

Make and evaluate statements about global relationships between entities

Data

- Batch-download count data from EBI
- Represent counts as a fraction of total count within a sample (compositional data)
- Set low-count observations to zero
- TF-IDF transform counts
  (term frequency - inverse document frequency)
- Calculate cosine distances

Entities and entity hierarchy:

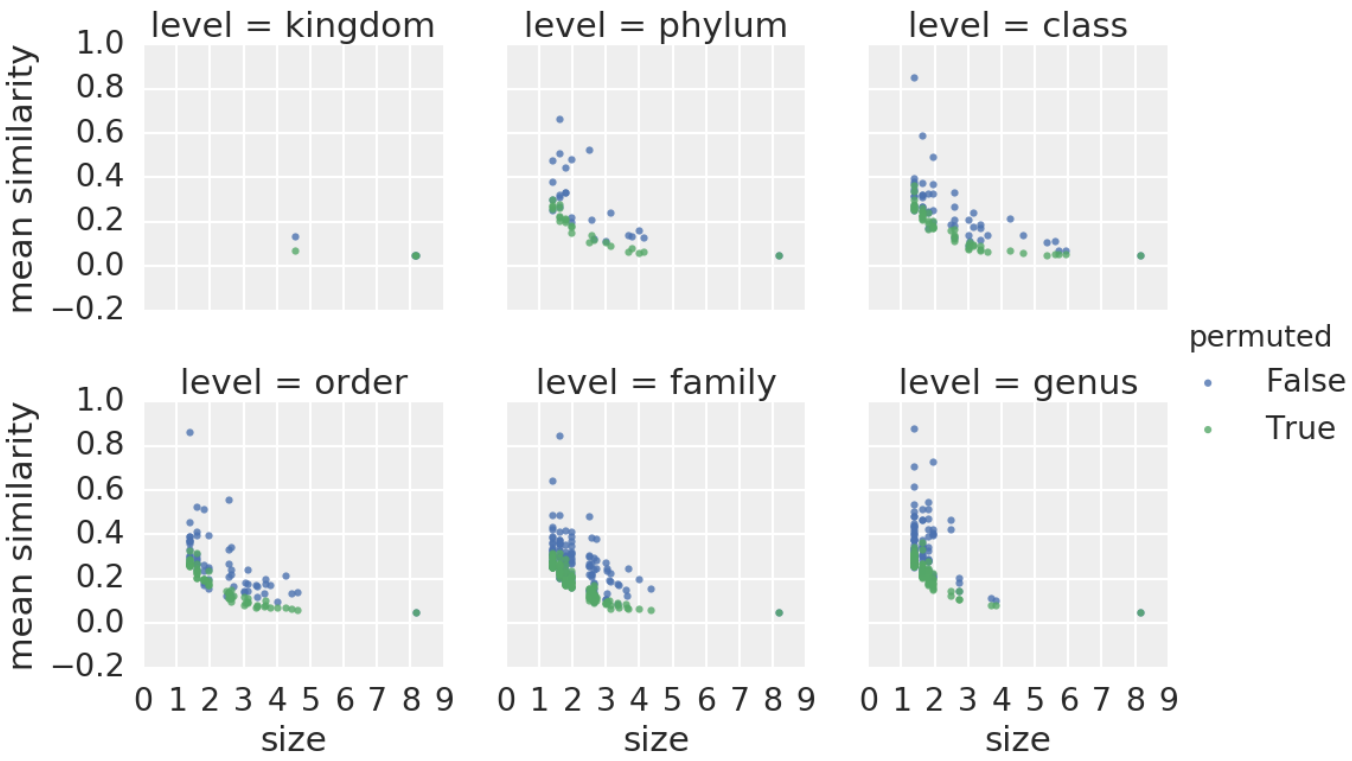
- Samples
  - Projects
  - Biomes
  - Keywords
  - Contrasts
- Taxa
  - Genera
  - Families
  - …
- Genes
  - Functions

Validation?

Comparison of results for sample-sample cosine distance between taxonomic data and functional data

- Mantel test: 0.28
  Permuted: ~0.00
  (10439 samples)

Comparison of mean similarity within taxonomic levels against permuted values



Where I stand

- Metadata browsing

  http://10.84.146.16:9000/home

- Sample-sample (also project-project, biome-biome)

  http://10.84.146.16:9000/projects/ERP010458

  http://10.84.146.16:9000/biomes/Fecal

- Taxon-taxon

  http://10.84.146.16:9000/taxa/k__Bacteria;%20p__Firmicutes;%20c__Clostridia;%20o__Clostridiales;%20f__Peptostreptococcaceae;%20g__[Clostridium];%20s__difficile

- Contrasts (differential abundance)

  http://10.84.146.16:9000/contrasts/soil-fecal

- Data management, optimization: getting results for > 300 samples and 3000 taxons is very slow!

Sample, project, biome results

- Precalculated distance matrix for all sample-sample relationships
- Collapsing and averaging for a higher-level entity (e.g. project)

Biomes most similar to Fecal

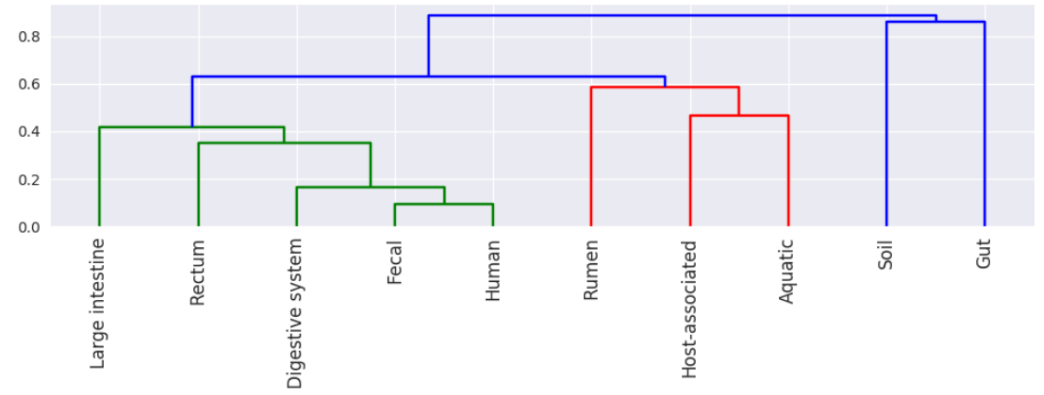
For agglomerative entities like project,

Biome 2	Metric
Fecal	0.719689
Rectum	0.731858
Human	0.752949
Digestive system	0.766736
Large intestine	0.858444
Soil	0.876893
Aquatic	0.889793
Host-associated	0.905403
Gut	0.917661
Rumen	0.917746

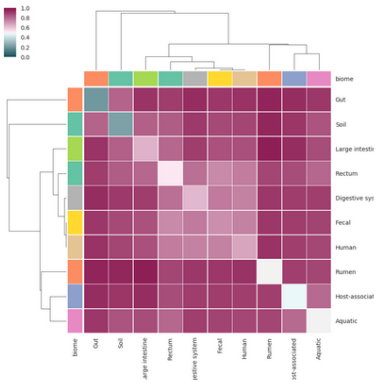
Samples most similar to ERS940246

Sample 2	Metric
ERS940246	0.000000
ERS939804	0.261820
ERS939485	0.324476
ERS941546	0.346890
ERS939653	0.348082
ERS920506	0.353270
ERS941537	0.355168
ERS915909	0.356859
ERS915910	0.358304
ERS918534	0.358816
ERS940404	0.360702

Biomes dendrogram for Fecal



Biomes heatmap for Fecal



TODO:

- Include more metrics (already implemented in taxon comparison)

Taxon results

HOME HELP ABOUT

K__BACTERIA; P__FIRMICUTES; C__CLOSTRIDIA; O__CLOSTRIDIALES; F__PEPTOSTREPTOCOCCACEAE; G__[CLOSTRIDIUM]; S__DIFFICILE

DOWNLOAD TABLE

Tf-idf cosine; ascending=True

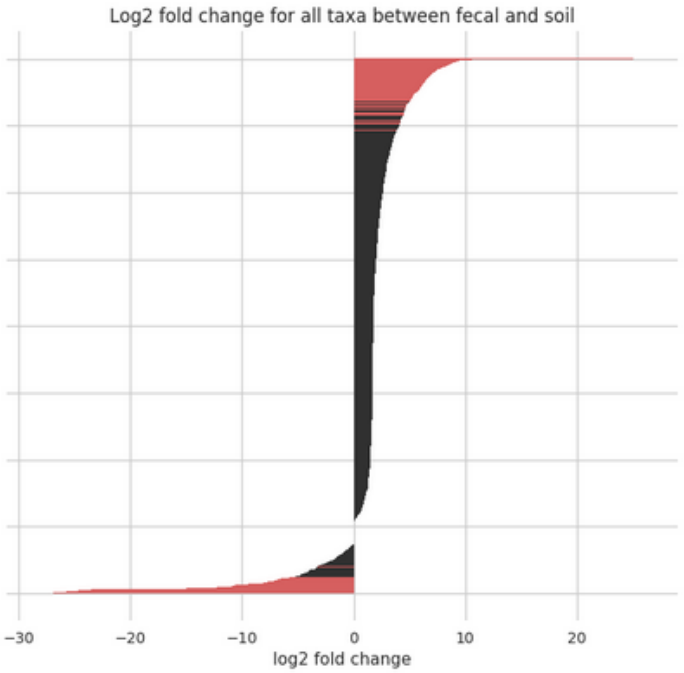
Taxon 2	Tf-idf cosine	N overlapping	Fisher exact p-value	Correlation	Chisquare kernel
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Peptostreptococcaceae; g__[Clostridium]; s__difficile	0.000000	2596.0	0.000007	1.000000	1.0
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Peptostreptococcaceae; g__; s__	0.283319	2452.0	0.000010	0.642032	0.0
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium; s__neonatale	0.297549	1176.0	0.000007	0.133426	0.0
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium; s__perfringens	0.329332	1937.0	0.000009	0.139407	0.0
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__Clostridium; s__butyricum	0.387502	1207.0	0.000006	0.086361	0.0
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Eubacteriaceae; g__Pseudoramibacter_Eubacterium; s__	0.439118	619.0	0.000006	0.158153	0.0
k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Carnobacteriaceae; g__; s__	0.466686	771.0	0.000007	0.195472	0.0
k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__Coprobacillus; s__cateniformis	0.467144	657.0	0.000005	0.121383	0.0
k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Enterococcaceae; g__Enterococcus; s__	0.475901	2300.0	0.000009	0.031046	0.0
k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Enterococcaceae; g__Enterococcus; s__casseliflavus	0.484032	1258.0	0.000008	0.112064	0.0
k__Bacteria; p__Firmicutes; c__Erysipelotrichi; o__Erysipelotrichales; f__Erysipelotrichaceae; g__[Eubacterium]; s__dolichum	0.494058	1383.0	0.000010	0.141750	0.0
k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__; g__; s__	0.497414	1666.0	0.000010	0.032788	0.0
k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Enterococcaceae; g__Vagococcus; s__	0.500240	1556.0	0.000008	0.030739	0.0
k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Leuconostocaceae; g__Weissella; s__	0.505225	1012.0	0.000008	0.105376	0.0
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Clostridiaceae; g__; s__	0.523169	2566.0	0.000007	0.137969	0.0
k__Bacteria; p__Actinobacteria; c__Coriobacteriia; o__Coriobacteriales; f__Coriobacteriaceae; g__Eggerthella; s__lenta	0.529286	1162.0	0.000010	0.104832	0.0
k__Bacteria; p__Firmicutes; c__Bacilli; o__Lactobacillales; f__Leuconostocaceae; g__; s__	0.554177	897.0	0.000008	0.023938	0.0

TODO:

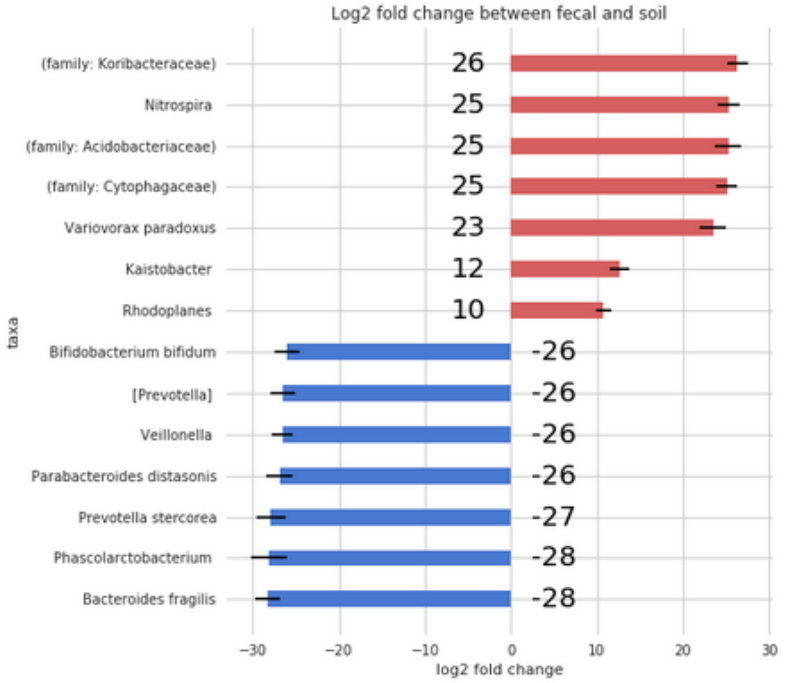
- Add genus, family, … collapsing
- Draw plots for taxon results similar to sample results

Contrasts

SOIL-FECAL



Comparison of fecal to soil (reference).

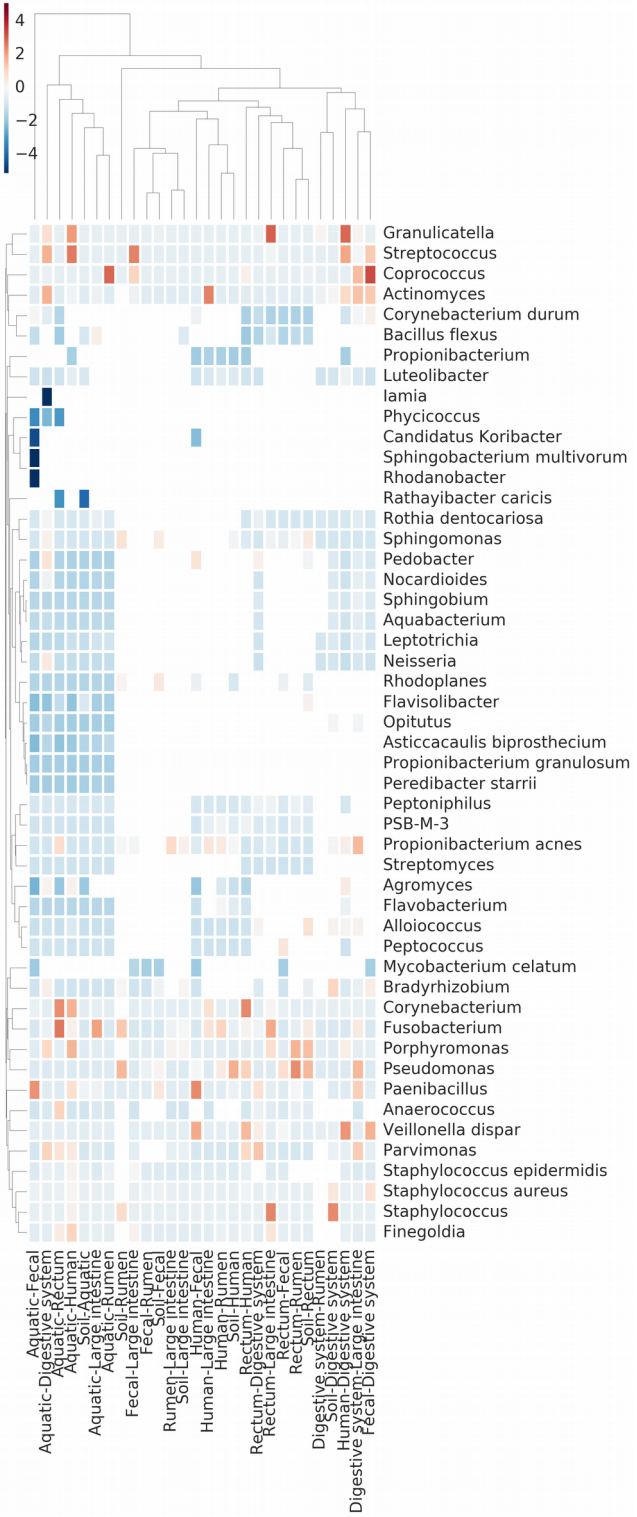


Comparison of fecal to soil (reference).
Showing 14 out of 282 taxa for which adjusted p-value < 0.01.
Standard errors are shown.

TODO:

- Add a possibility to specify arbitrary groups of samples
- Log fold change plots for higher taxonomy levels

One step higher: cluster contrasts

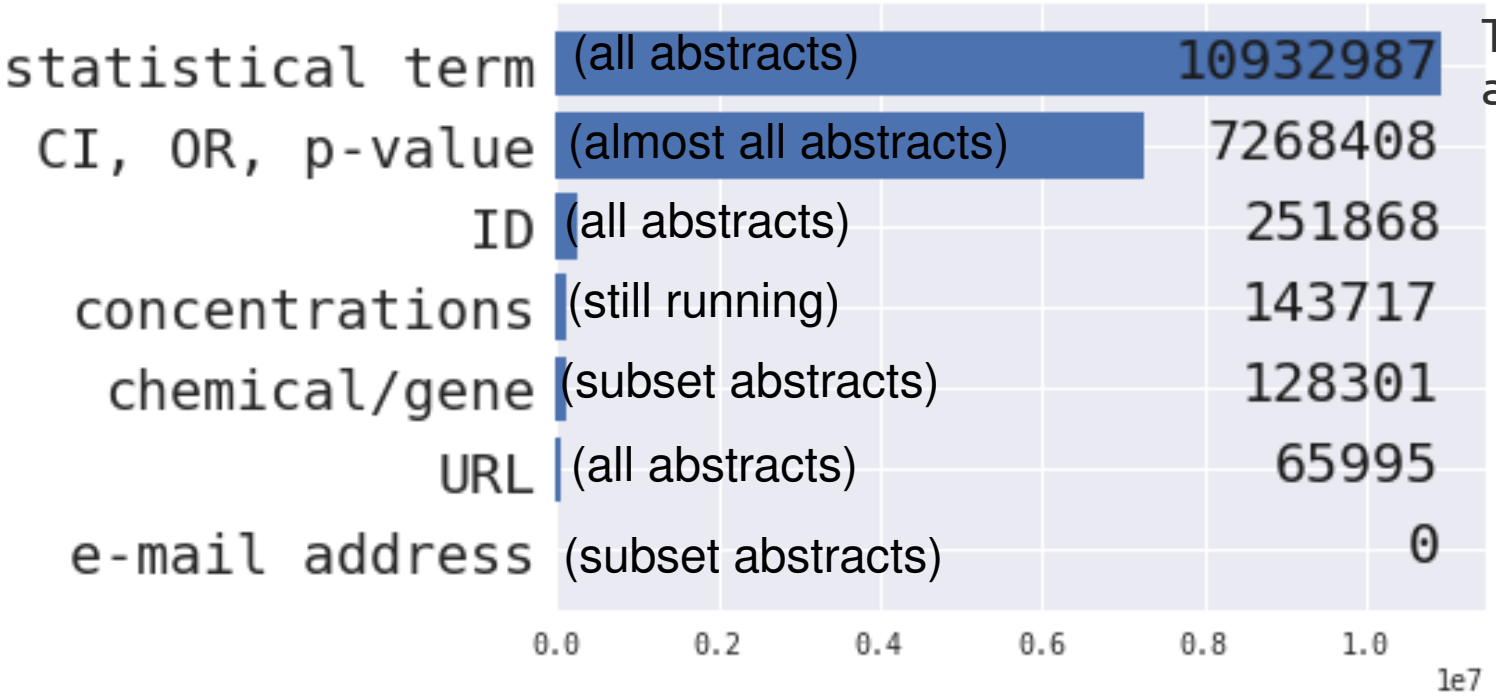


PubQC

Correct errors in scientific publications submitted to journals (“Reviewer n+1” scheme)

Analyze the errors in scientific literature in the wild

PubQC



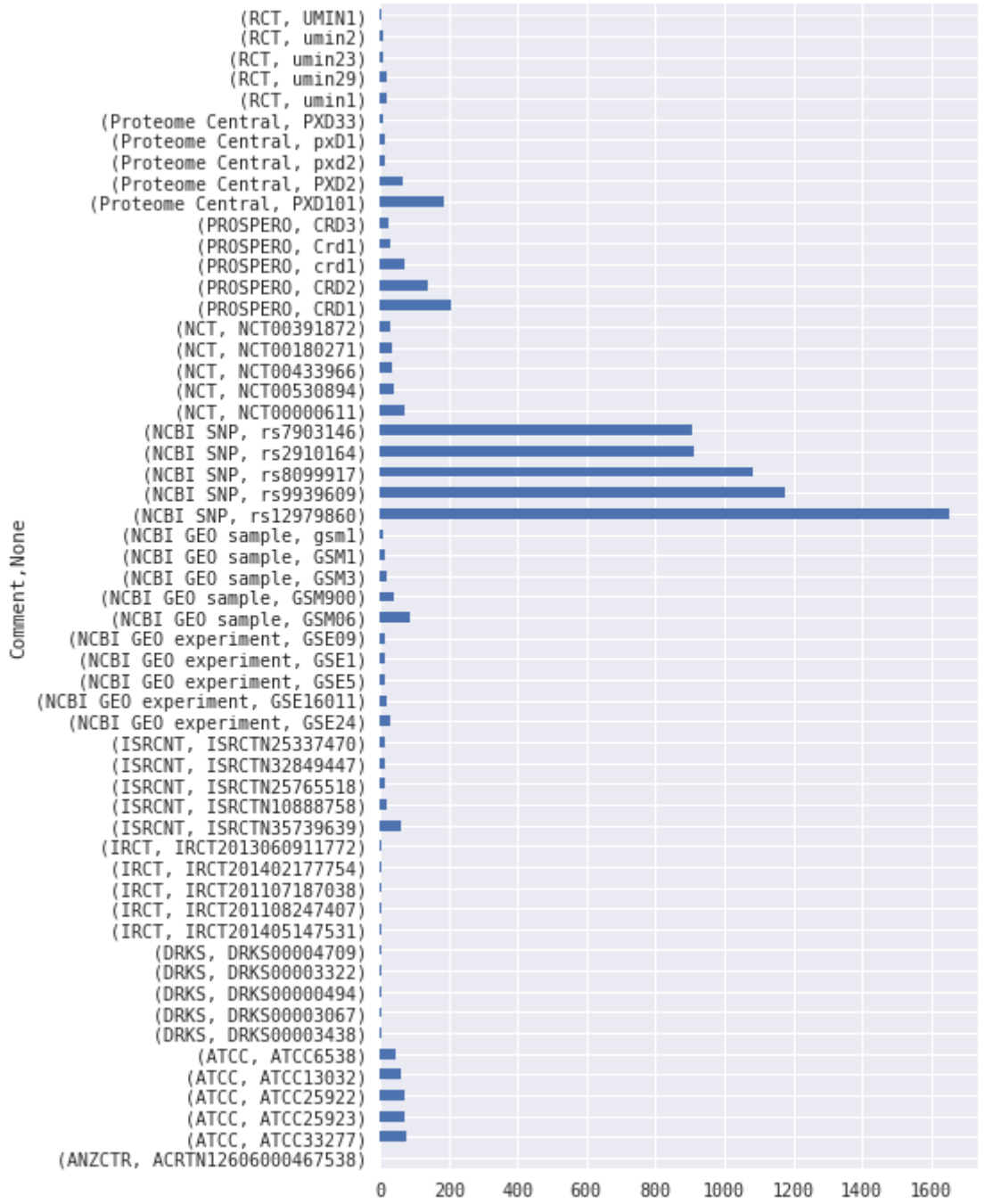
Total accessible abstracts:

http://10.84.146.16:9001/home

http://10.84.146.16:2460/notebooks/pubqc_test_results.ipynb

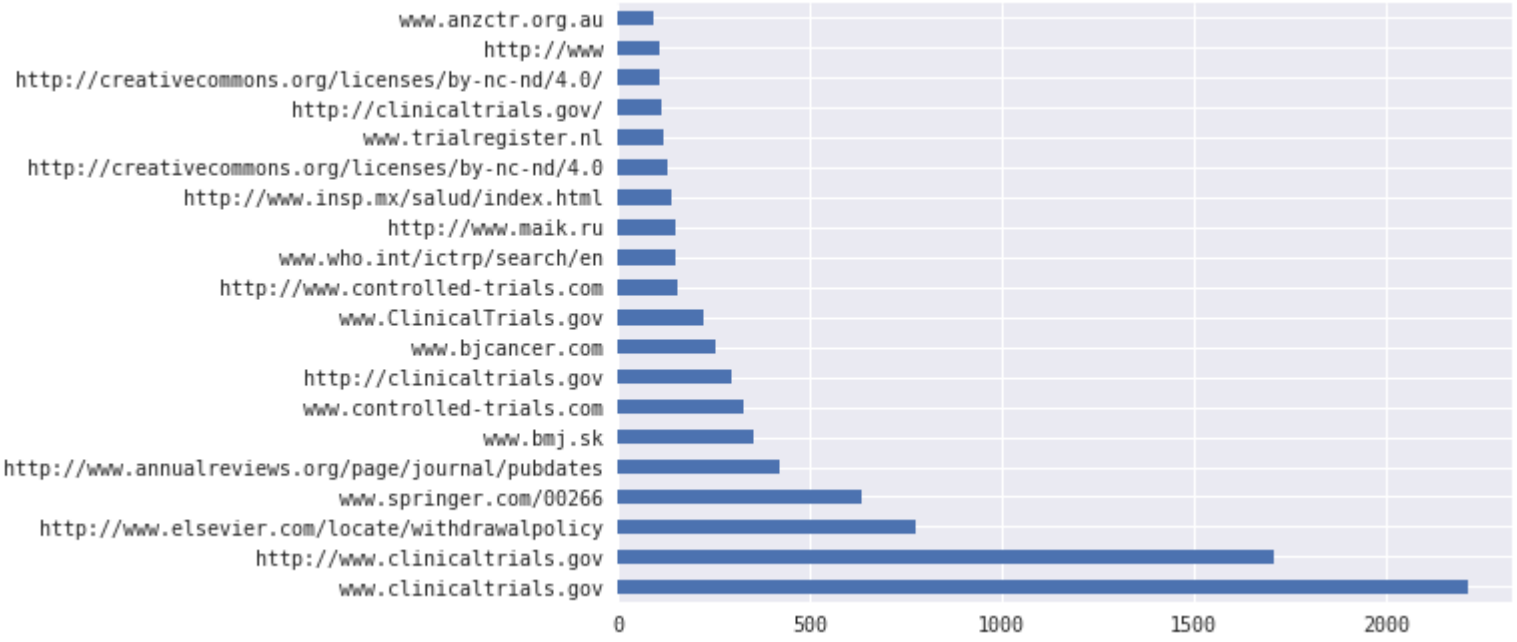
PubQC – IDs (regex)

- 251,868 found
- Belinostat (**PXD101**) is a novel HDAC inhibitor with IC50 of 27 nM in a cell-free assay, with activity demonstrated in cisplatin-resistant tumors
- **rs12979860** is a SNP near the IL28B gene, encoding interferon-lambda-3 (IFN-lambda-3)
- **rs9939609** is a SNP in the fat mass and obesity associated FTO gene, aka the "Fat Gene"
- **ATCC 33277** Porphyromonas gingivalis genome - human pathogen
- **rs8099917** is related to hepatitis C treatment response
- **GSE16011** is a glioma cohort
- (TODO) **GSM1** is a protein, **GSM06** is a cell line, **CRD1** a protein domain, etc.



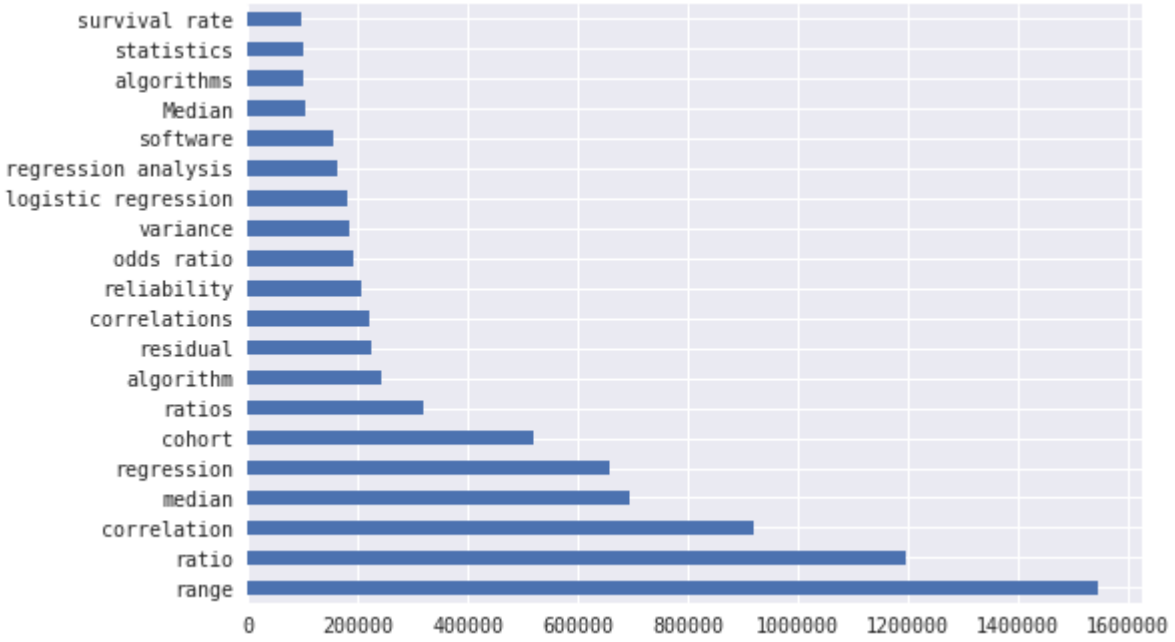
PubQC – URLs (regex)

- 65,995 found
- **www.springer.com/00266** - aesthetic plastic surgery (??)
- (TODO) '**http://www**' is almost always matched when there's a space after the 'www.'
- (TODO) spaCy results were rubbish – why?
- (TODO) programatically get respone from the server



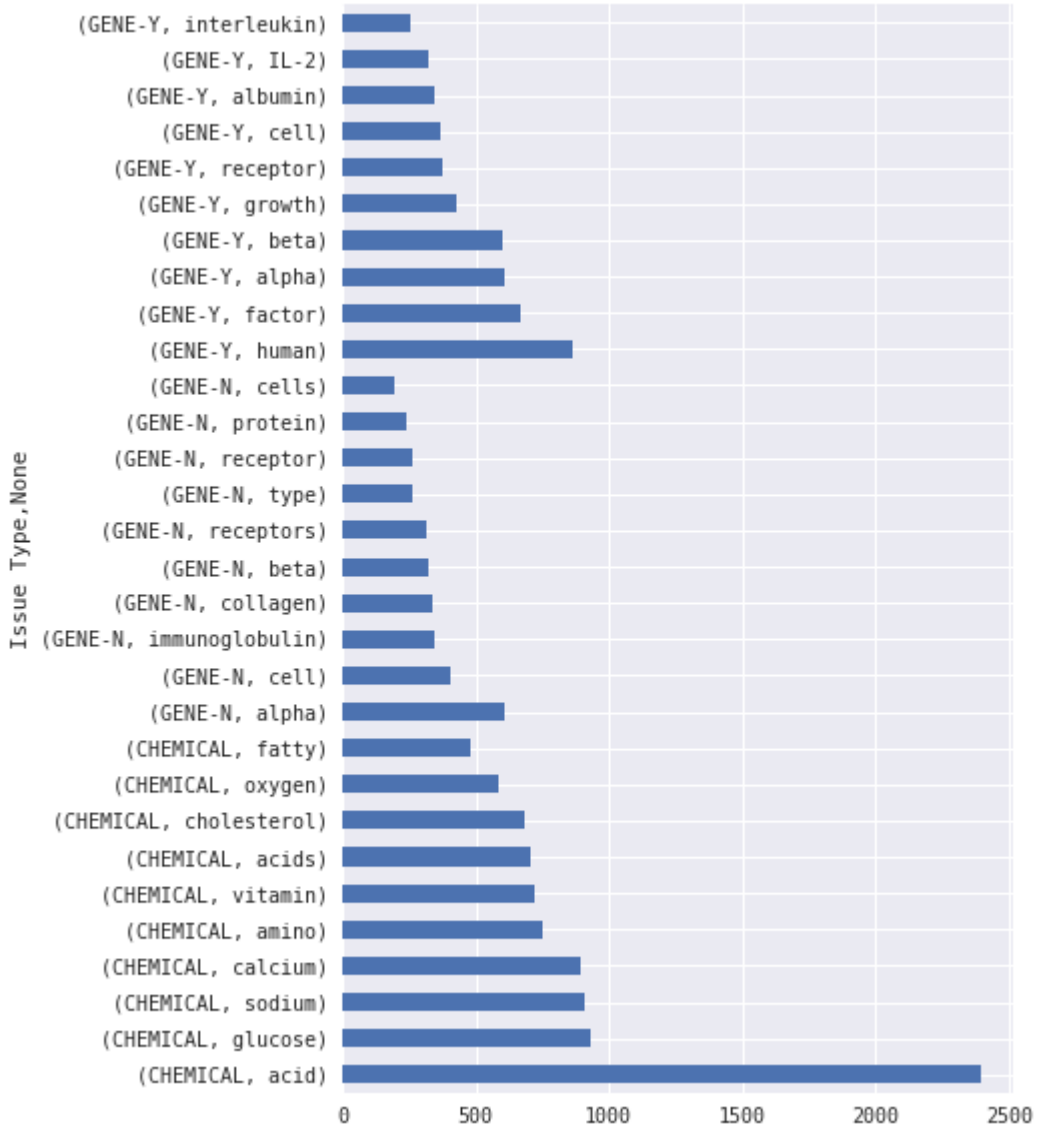
PubQC – terms (simple pattern match)

- 10,932,987 found
- (TODO) ignore case

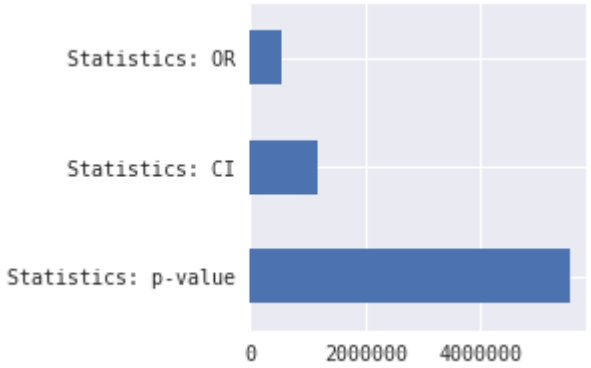


PubQC – chemicals, genes (spaCy)

- 128,301 found, but...



PubQC – CI, OR, p-value



- Number of values associated with each

Statistics: CI	3	12043
	4	545
	2	12
	5	5
	6	4
Statistics: OR	1	6208
	2	1
Statistics: p-value	1	63403

- (TODO) scientific notation
- (TODO) recalculate statistics

- Extracted p-values

0.050000	13836
0.001000	11933
0.010000	8188
0.000100	3201
0.020000	1857
0.000000	1543
...	
1.000000	184
...	
648.000000	1
6.540000	1

PubQC – concentrations, percentages

- (TODO) Get chemicals in the proximity

1:1	385
2:1	178
18:2	143
18:1	121
16:0	106
...	
100mg/kg)	8
50mg/kg)	8
12:0	8
200mg/kg)	7
5mg/kg)	7
19:00	7

95%	10130
50%	3683
10%	2510
20%	2378
100%	2207
5%	2033
30%	1892
90%	1857
80%	1794
40%	1744
25%	1550
60%	1487
70%	1465
1%	1369