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

A researcher's
interface to database

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:

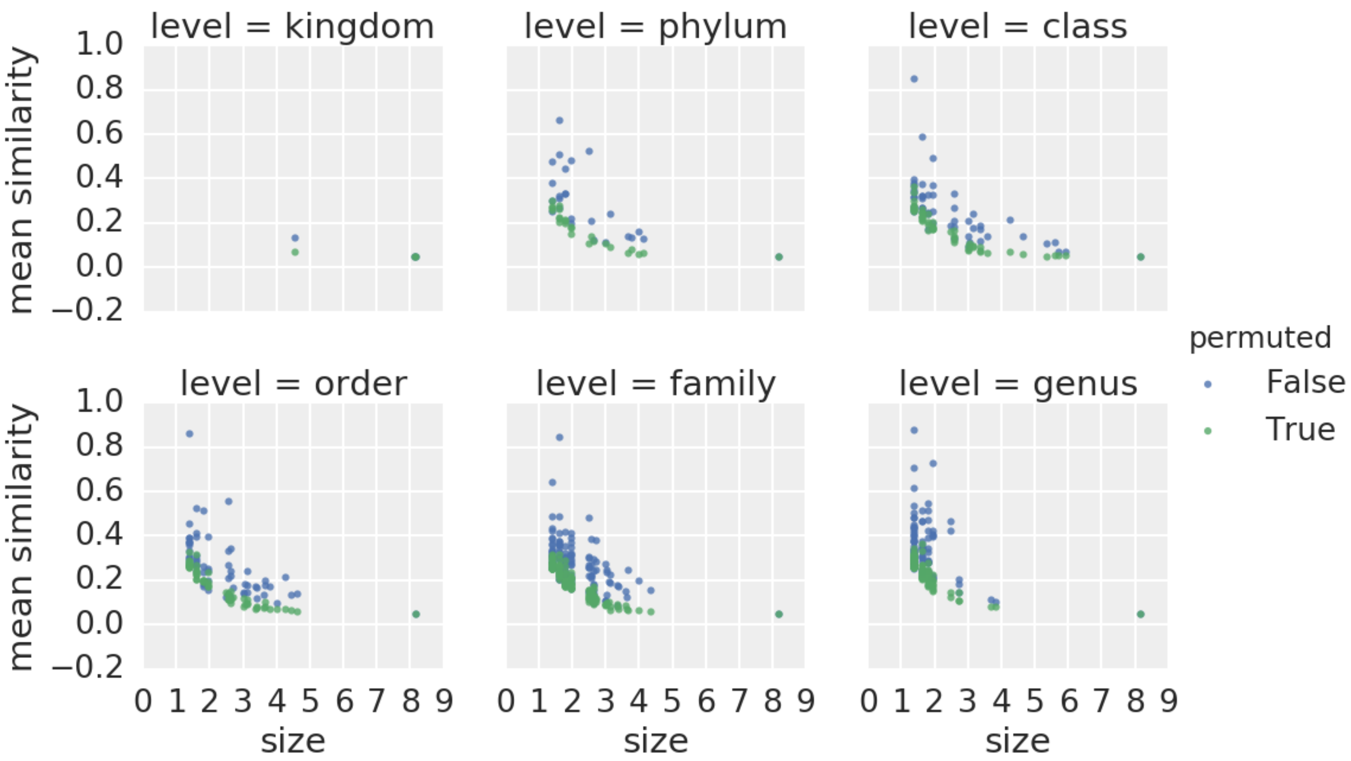
- Samples
  - Projects
  - Biomes
  - Keywords
  - Contrasts
- Taxa
  - Genera
  - Families
  - …
- Genes
  - Functions

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

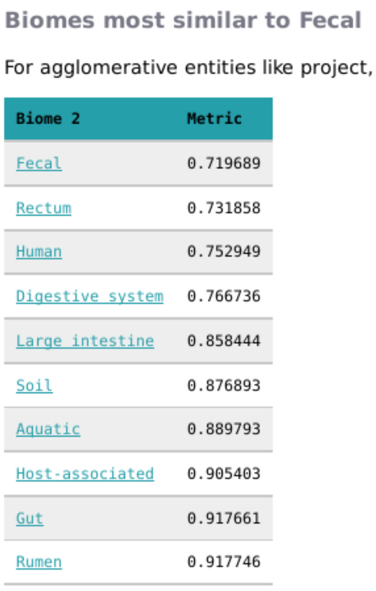
- Contrasts (differential abundance)

  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

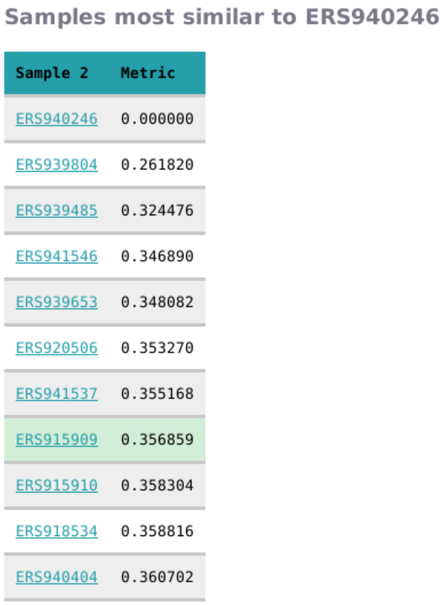
- Precalculated distance matrix for all sample-sample relationships
- Collapsing and averaging for a higher-level entity (e.g. project)



Biomes most similar to Fecal

For agglomerative entities like project,

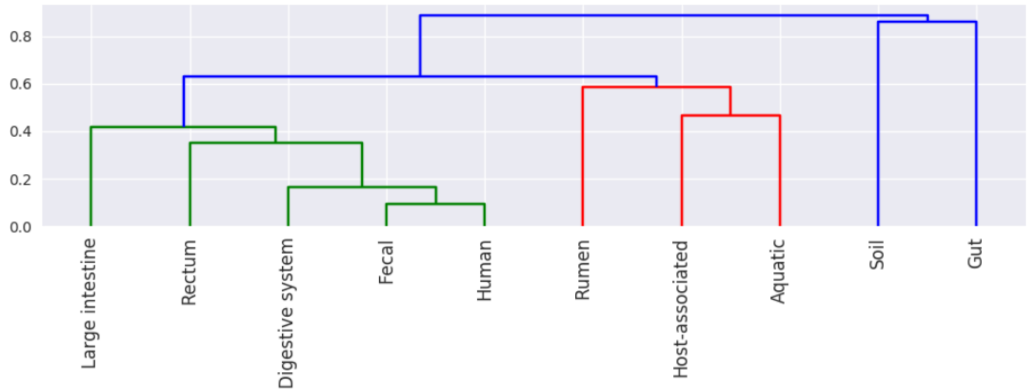
Biome 2	Metric
Fecal	0.719689
Rectum	0.731858
Human	0.752949
Digestive system	0.766736
Large intestine	0.858444
Soil	0.876893
Aquatic	0.889793
Host-associated	0.905403
Gut	0.917661
Rumen	0.917746



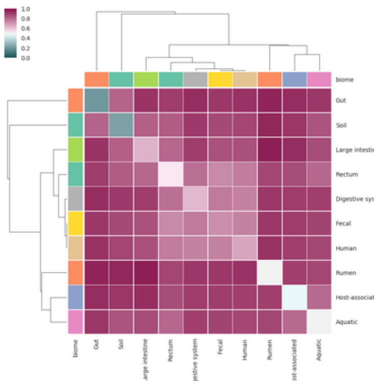
Samples most similar to ERS940246

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



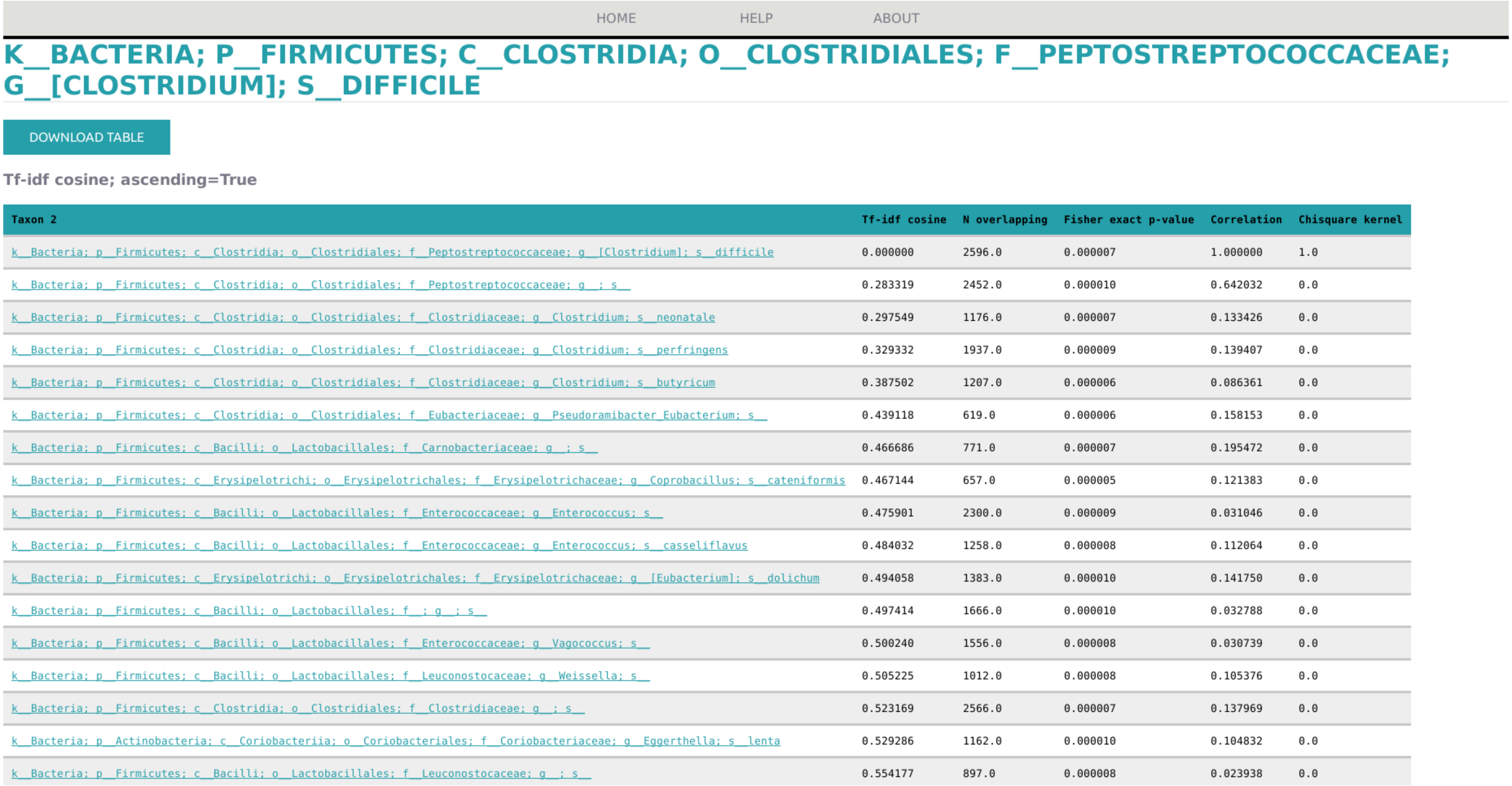
Biomes heatmap for Fecal



TODO:

- Include more metrics (already implemented in taxon comparison)

Taxon results



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

Tf-idf cosine; ascending=True

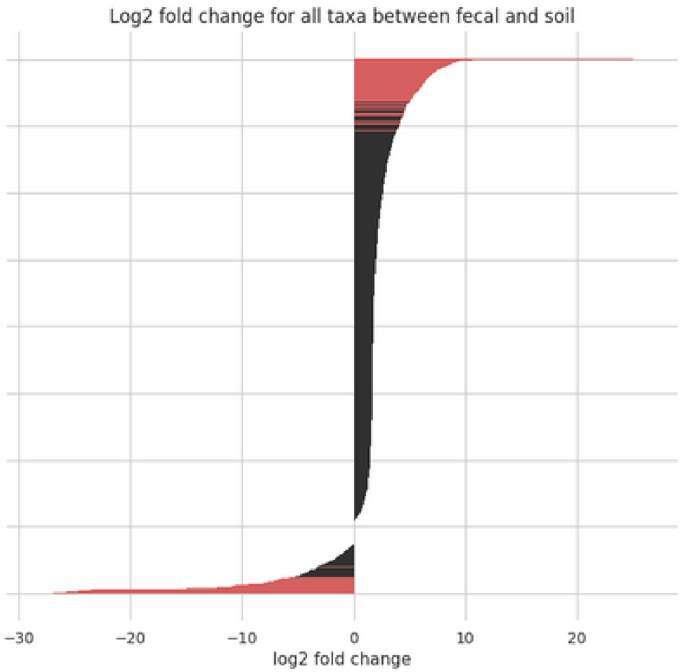
Taxon 2	Tf-idf cosine	N overlapping	Fisher exact p-value	Correlation	Chisquare kernel
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; f__Peptostreptococcaceae; g__[Clostridium]; s__difficile	0.000000	2596.0	0.000007	1.000000	1.0
k__Bacteria; p__Firmicutes; c__Clostridia; o__Clostridiales; 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:

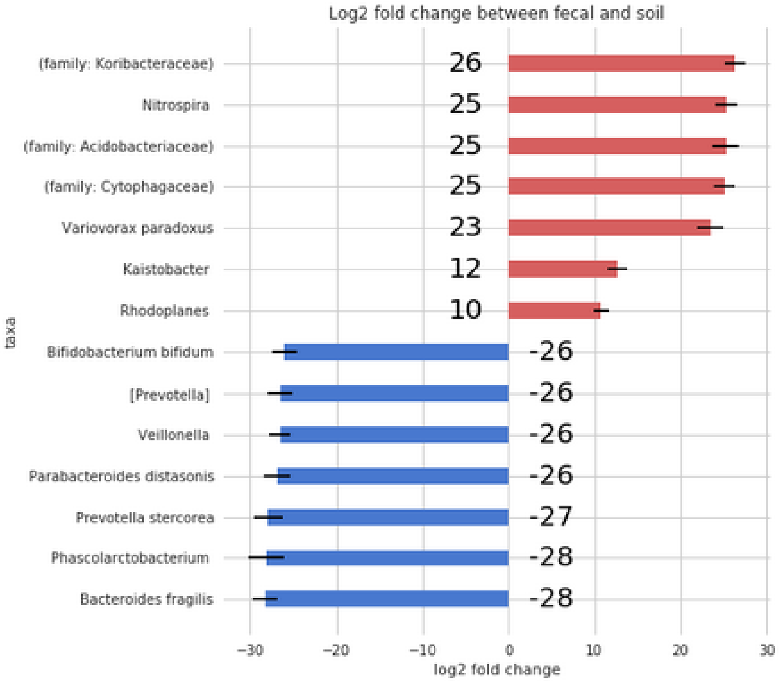
- Add genus, family, … collapsing
- Draw plots for taxon results similar to sample results

Contrasts

SOIL-FECAL



Comparison of fecal to soil (reference).

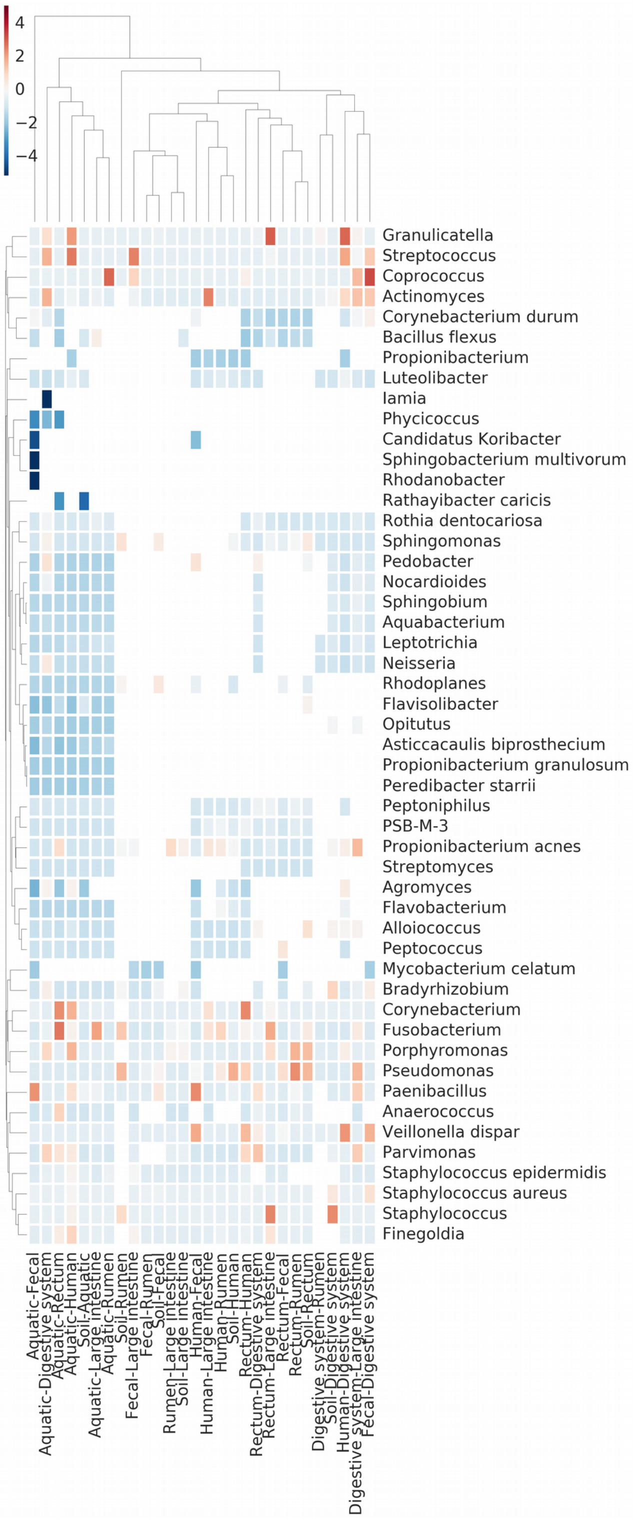


Comparison of fecal to soil (reference).
Showing 14 out of 282 taxa for which adjusted p-value < 0.01.
Standard errors are shown.

TODO:

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

One step higher: cluster contrasts



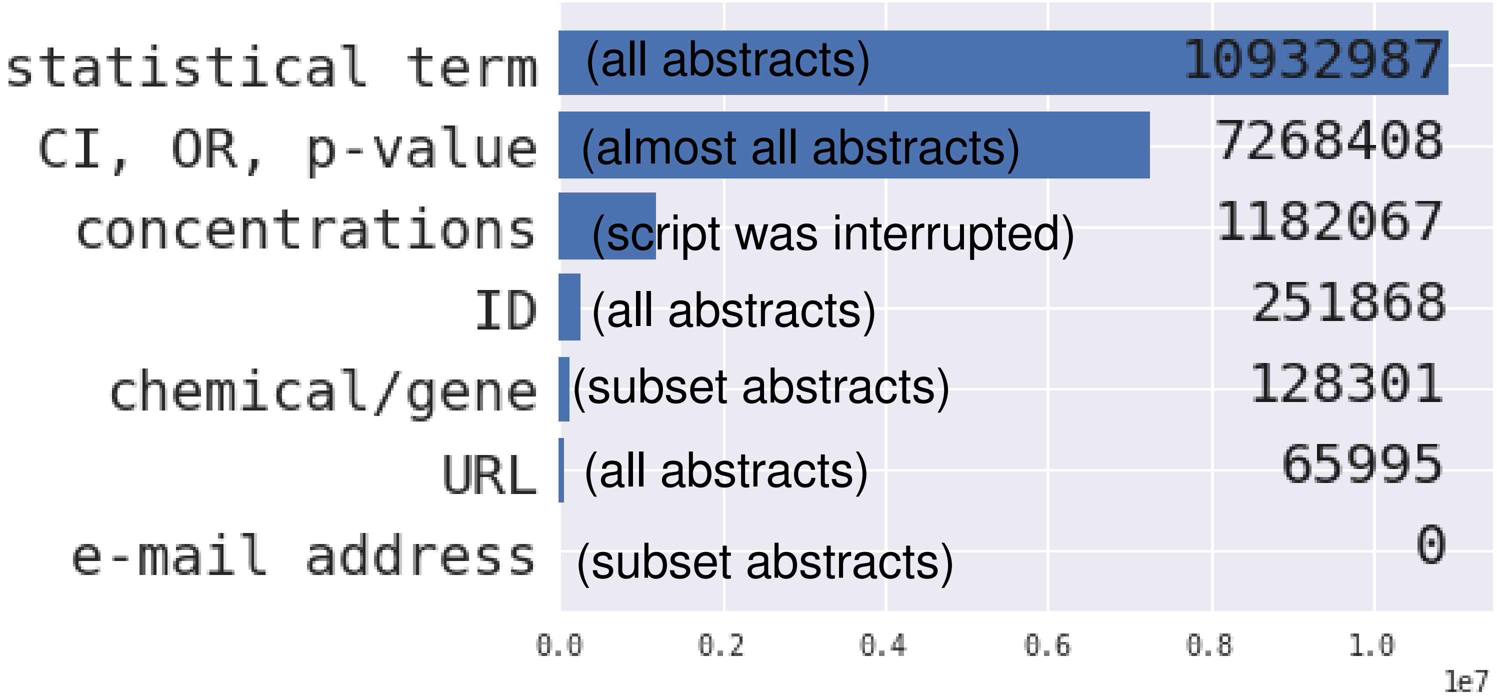
PubQC

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: 27,444,507

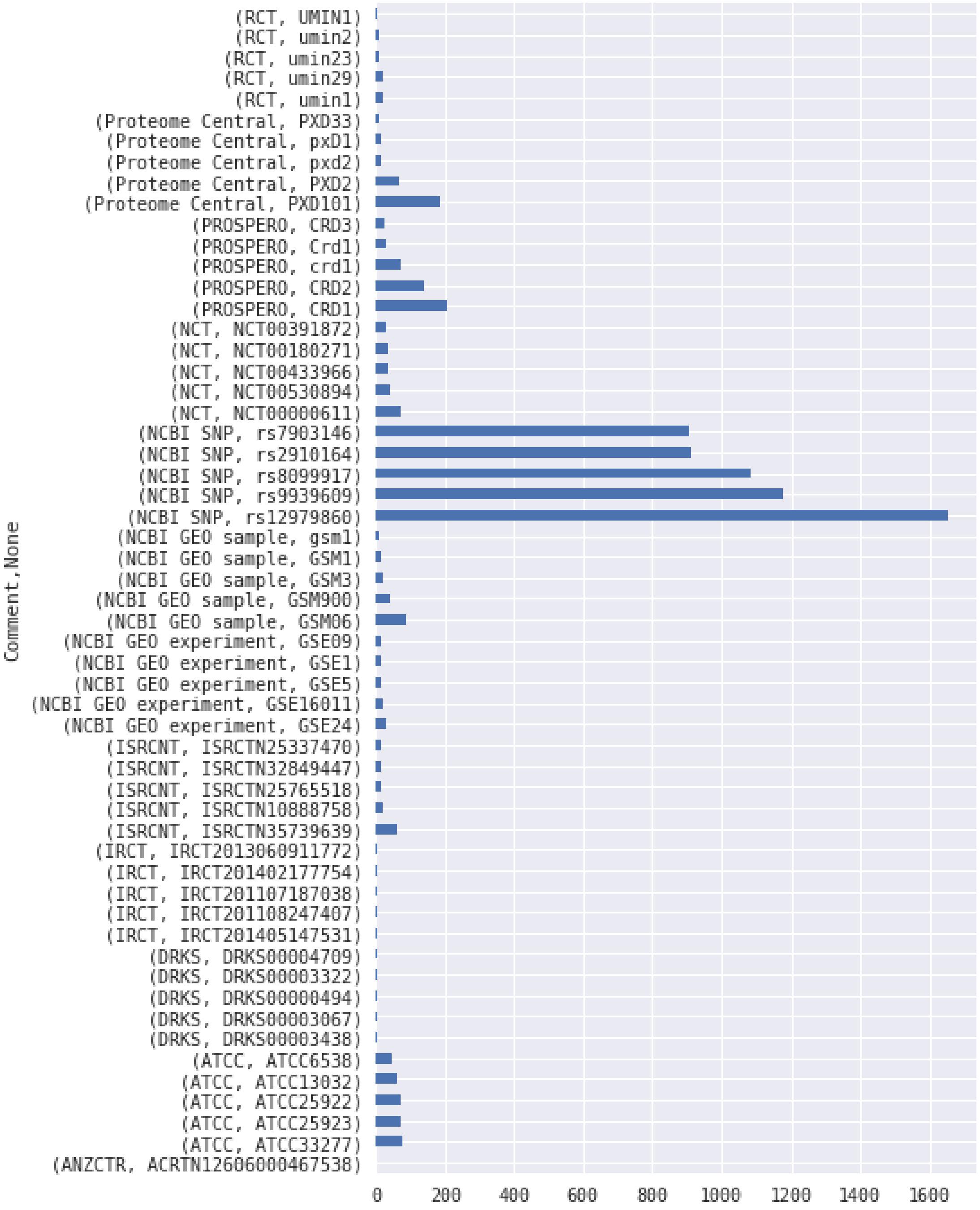


http://10.84.146.16:9001/home

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

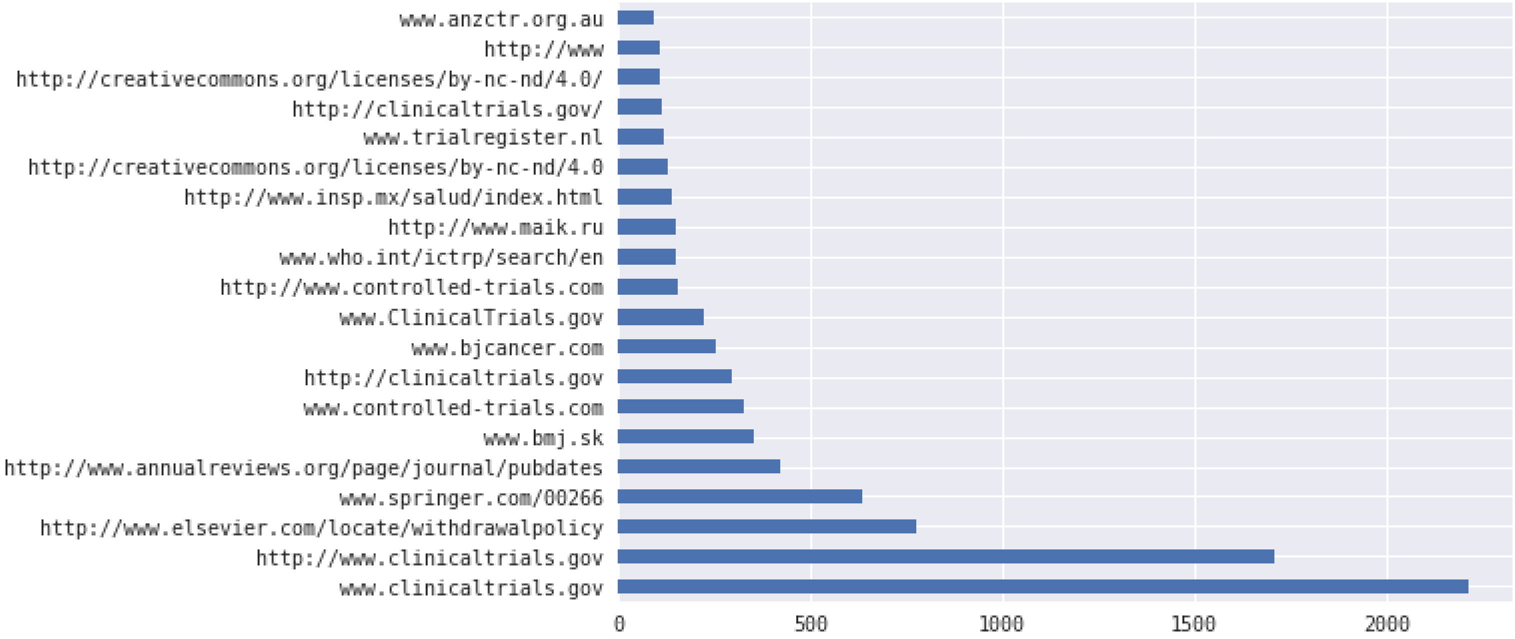
PubQC – IDs (regex)

- 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.



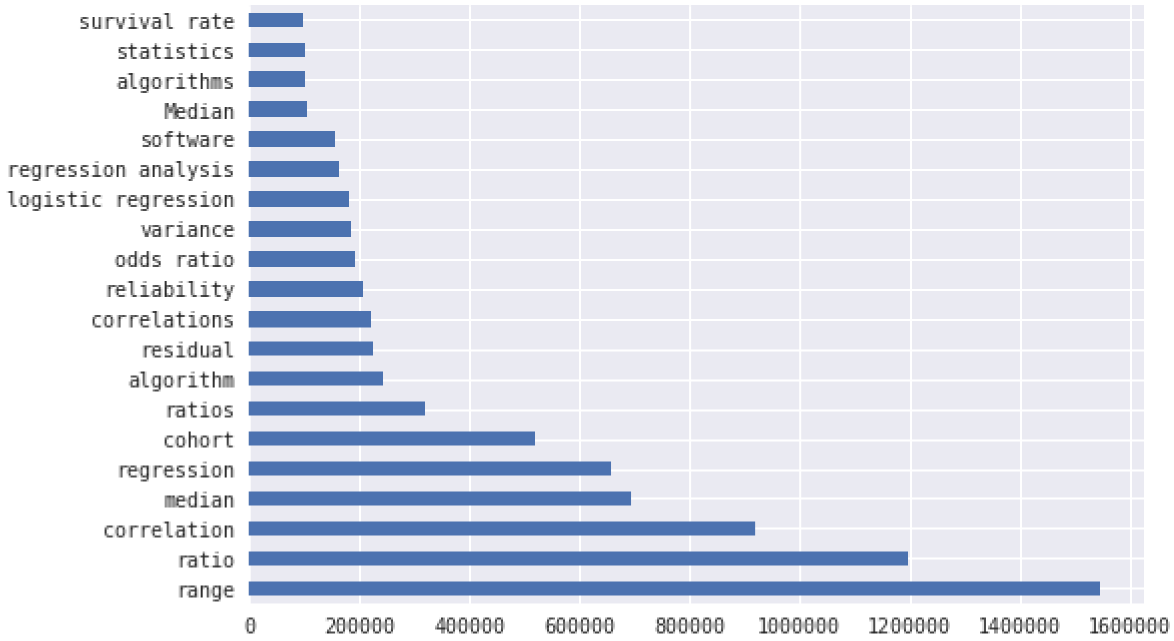
PubQC – URLs (regex)

- 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



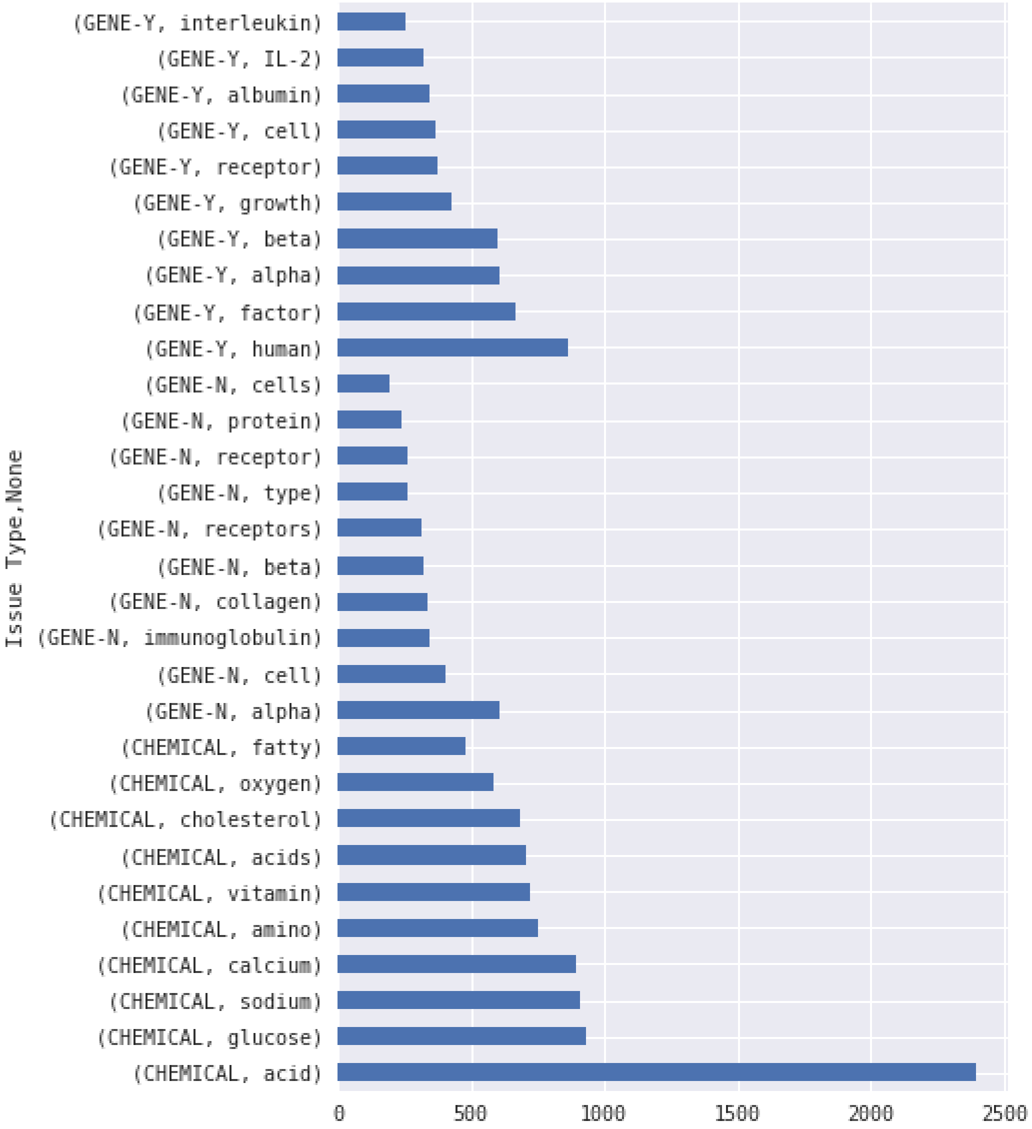
PubQC – terms (simple pattern match)

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

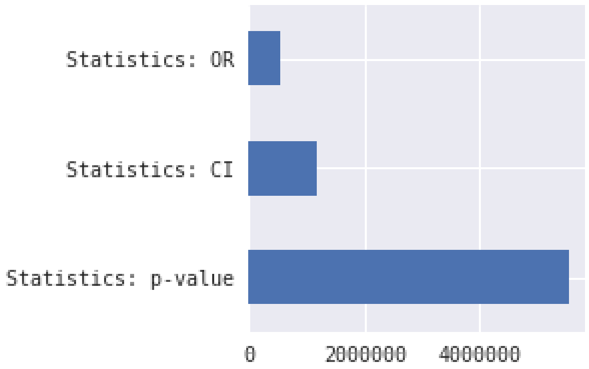


PubQC – chemicals, genes (spaCy)

- 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
```

PMID	Pattern	Location	value	Lower	Upper	Recalculated OR	Pattern	Location	Context
3384186	95% CI (CI)=(0.93, 3.17	(762, 785)	[95.0, 0.93, 3.17]	0.93	3.17		OR)=1.71	(752, 760)	ve to the NHWs [odds ratio (OR)=1.71; 95% CI (CI)=(0.93, 3.17)]. The risk of severe retinopathy (pr
3387899	95% CI, 1.5-3.5	(943, 958)	[95.0, 1.5, 3.5]	1.5	3.5		OR=2.3	(935, 941)	eir place of residence indicated (OR=2.3; 95% CI, 1.5-3.5). There was a trend for patients to repor
3387899	95% CI, 0.9-6.7	(1032, 1047)	[95.0, 0.9, 6.7]	0.9	6.7		OR=2.4	(1024, 1030)	s to report an unhappy childhood (OR=2.4; 95% CI, 0.9-6.7). Being unmarried, undergoing parental se
3389360	3.94, CI=1.52-10.20	(849, 868)	[3.94, 1.52, 10.2]	1.52	10.2		OR=3.94	(846, 853)	d an elevated risk for brain cancer (OR=3.94, CI=1.52-10.20). In addition, there was a linear relat
3389361	1.72, CI 1.09-2.97	(750, 768)	[1.72, 1.09, 2.97]	1.09	2.97		OR 1.72	(747, 754)	the reduction division of the plant (OR 1.72, CI 1.09-2.97) including, in particular, Soderberg (OR
3389361	1.71, CI 1.07-2.72	(810, 828)	[1.71, 1.07, 2.72]	1.07	2.72		OR 1.71	(807, 814)	including, in particular, Soderberg (OR 1.71, CI 1.07-2.72) and prebake (OR 2.26, CI. 1.27-4.02) po
3389361	2.26, CI. 1.27-4.02	(846, 865)	[2.26, 1.27, 4.02]	1.27	4.02		OR 2.26	(843, 850)	(OR 1.71, CI 1.07-2.72) and prebake (OR 2.26, CI. 1.27-4.02) potroom workers. The risk of IHD did n
3394698	95% CI=1.6-4.9	(877, 891)	[95.0, 1.6, 4.9]	1.6	4.9		OR)=2.8	(868, 875)	eral cancer (adjusted odds ratio (OR)=2.8, 95% CI=1.6-4.9). Further, this relation was modified by t
3394698	95% CI=1.2-3.0	(1003, 1017)	[95.0, 1.2, 3.0]	1.2	3		OR=1.9	(995, 1001)	initial cancer diagnosis (ratio of OR=1.9, 95% CI=1.2-3.0 for a five-year differential in time since
3394698	95% CI=0.3-1.2	(1191, 1205)	[95.0, 0.3, 1.2]	0.3	1.2		OR=0.6	(1183, 1189)	h a 36% decrease in risk (adjusted OR=0.6, 95% CI=0.3-1.2); this estimate excluded the magnitude of
3394707	95% CI=1.1-2.2	(520, 534)	[95.0, 1.1, 2.2]	1.1	2.2		OR)=1.6	(511, 518)	showed that male sex (odds ratio (OR)=1.6, 95% CI=1.1-2.2), maternal five-year age increase (OR=1.3,
3394707	95% CI=1.1-1.5	(578, 592)	[95.0, 1.1, 1.5]	1.1	1.5		OR=1.3	(570, 576)	, maternal five-year age increase (OR=1.3, 95% CI=1.1-1.5), plural birth (OR=3.0, 95% CI=1.2-7.1) an
3394707	95% CI=1.2-7.1	(617, 631)	[95.0, 1.2, 7.1]	1.2	7.1		OR=3.0	(609, 615)	.3, 95% CI=1.1-1.5), plural birth (OR=3.0, 95% CI=1.2-7.1) and black maternal race (OR=0.0, 95 perce
3395581	95% CI=1.0-2.6	(634, 648)	[95.0, 1.0, 2.6]	1	2.6		OR)=1.6	(625, 632)	icultural production (odds ratio (OR)=1.6, 95% CI=1.0-2.6). A concomitant increase was detected for
3395	95%	(708, 722)	[95.0, 1.0, 2.5]	1	2.5		OR=1.6	(700, 706)	increase was detected for farmers (OR=1.6, 95% CI=1.0-

PubQC – concentrations, percentages

- (TODO) Get chemicals in the proximity (extract noun?)

```
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
```

2.2%
Percent
(838, 842)
xed meat or chicken or pork or beef contributes 2.2%), the GI
emission is estimated to be 0.28 or 0
meat (CHEMICAL); chicken (CHEMICAL); pork (CHEMICAL)

0.5%
Percent (178, 182)
sound biomicroscopy. To determine the effect of 0.5%
tropicamide and the resultant mydriasis on the
tropicamide (CHEMICAL)

84%
percent
(635, 638)
c SBR became fully granulated and finished with 84% and 99
of nitrogen and phosphorus removal, res
nitrogen (CHEMICAL)

19.73ng/ml
concentration
(556, 567)
. Cutoff value for detecting stage B HF was 19.73ng/ml for
catestatin with 90% sensitivity and 50.9
catestatin (CHEMICAL)

cat0

Column information

Column name: cat0

Column type: object

Table: Summary by cat0

cat0	logistic								lognormal								normal		
	count	mean	std	min	25%	50%	75%	max	count	mean	std	min	25%	50%	75%	max	count	mean	std
a	189.0	0.253651	1.711486	-4.02	-0.880	0.10	1.2000	6.53	189.0	1.778730	2.267438	0.07	0.600	1.05	2.1100	18.72	189.0	-0.030159	1.032
b	160.0	-0.015062	1.716513	-5.16	-0.910	-0.21	0.9675	6.50	160.0	1.607500	1.690542	0.07	0.580	1.05	1.9625	10.46	160.0	0.009625	0.968
c	151.0	0.419073	2.052203	-8.76	-0.815	0.37	1.5350	5.83	151.0	1.766556	2.838467	0.09	0.515	1.09	1.9750	30.58	151.0	-0.012517	1.002

Titles: MANOVA results for all variables

-	Df	Pillai	approx F	num_Df	den Df	Pr(>F)
cat0	2	0.014649	0.91312	8	990	0.5046
Residuals	497	-	-	-	-	-

Titles: MANOVA results for response lognormal

-	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cat0	2	3.0	1.5021	0.2839	0.753
Residuals	497	2629.5	5.2908	-	-

Titles: MANOVA results for response logistic

-	Df	Sum Sq	Mean Sq	F value	Pr(>F)
cat0	2	15.04	7.5180	2.2633	0.1051
Residuals	497	1650.90	3.3217	-	-

lognormal

Column information

Column name: lognormal

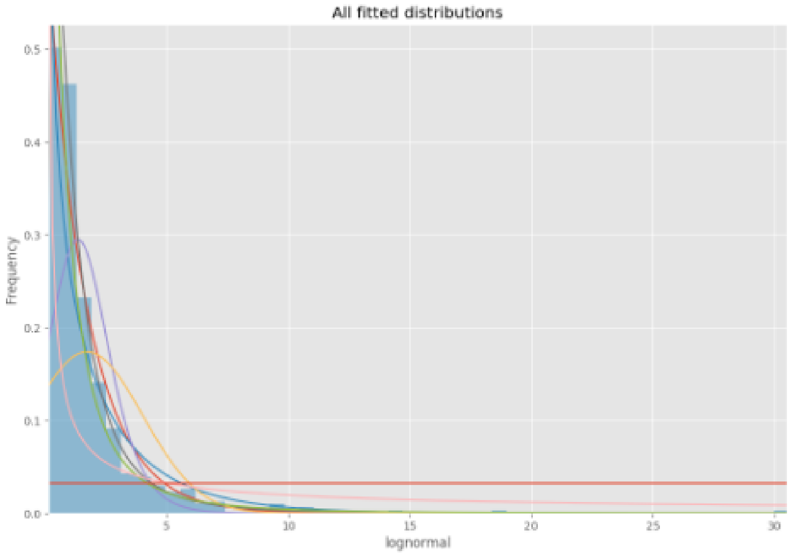
Column type: float64

Table: Column summary

count	mean	std	min	25%	50%	75%	max
500.0	1.72026	2.296861	0.07	0.58	1.065	2.0025	30.58

Table: Fitted distributions with best parameters and residual sum of squares

distribution	parameters	SSE
beta	a=0.98, b=113.42, loc=0.07, scale=192.26	0.185949
chi2	df=1.32, loc=0.07, scale=1.54	0.471846
logistic	loc=1.34, scale=0.85	0.626428
lognorm	s=0.97, loc=0.01, scale=1.06	0.0629841
norm	loc=1.72, scale=2.29	1.05756
pareto	b=2.02, loc=-1.83, scale=1.90	0.510913
powerlaw	a=0.27, loc=0.07, scale=33.09	1.124
uniform	loc=0.07, scale=30.51	2.09939



Article

Qualitative Inquiry
1–2
© The Author(s) 2017
Reprints and permissions:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1077800417727767
journals.sagepub.com/home/qix
SAGE

Who Cares?

Maria K. E. Lahman[1]

Keywords
methodology, poetry, research poetry

I fantasize saying,
“Who CARES if
YOU think poetry
can be research?”,
with a negligent
shrug of
shoulders.

Research IS poetry.

We are having a
methodological
ARGUMENT.

Tug
of
War,

Over
our DEAD fetuses’ bodies,[1]
a girl’s ANOREXIA,[2]
my postpartum DEPRESSION[3]
their HIV,[4]
her MURDER,[5]
his SUICIDE,[6]
a daughter’s racial IDENTITY,[7]
their COMING OUT.[8]

Experience,
story,
research,
poetry,
inquiry,
explorations,
should
fluster,
perplex,
unsettle
method…
METHODOLATRY.[9]

But I’d be lying.

I
know
I
care…

Declaration of Conflicting Interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) received no financial support for the research, authorship, and/or publication of this article.

Notes

1. Lahman (2013).
2. Chan (2003).
3. Lahman (2008).
4. Poindexter (2002).
5. Lahman (2011).
6. Teman (2010).
7. Davis (2007).
8. Teman (2011).
9. Janesick (1994).

References

Chan, Z. C. (2003). A poem: Anorexia. *Qualitative Inquiry*, *9*, 956-957.

Davis, A. M. (2007). SIP (School Induced Psychosis) poem for my daughter. *Qualitative Inquiry*, *13*, 919-924.

Janesick ,V. J.(1994). The dance of qualitative research design -- metaphor, methodolatry and meaning. In N. K. Denzin & Y. S. Lincoln (Eds.), Handbook of qualitative research. pp.209-219.Thousand Oaks, CA: Sage.

[1]University of Northern Colorado, Greeley, USA

Corresponding Author:
Maria K. E. Lahman, University of Northern Colorado, McKee Hall, Box 124, Greeley, CO 80631, USA.
Email: maria.lahman@unco.edu