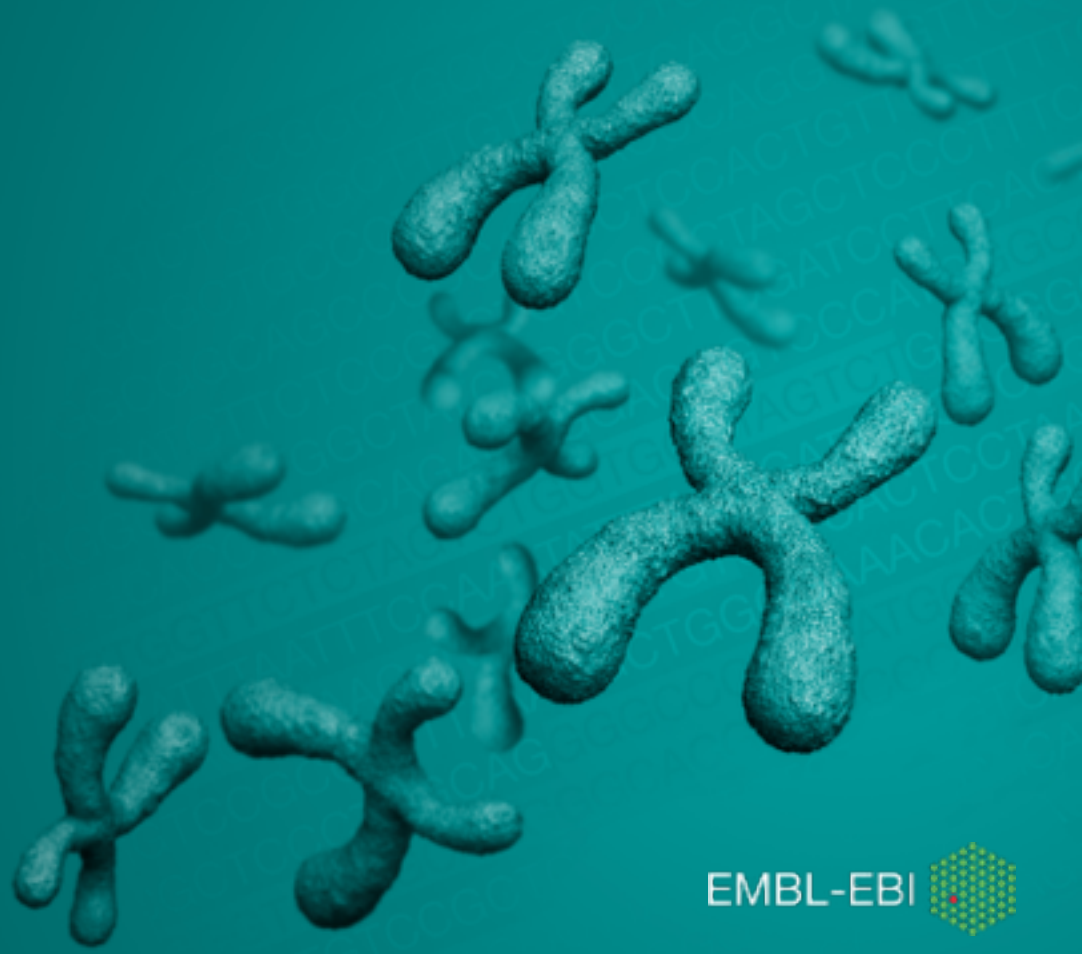
Towards understanding the functional and taxonomic repertoire of microbial communities using the EBI metagenomics portal

Rob Finn (rdf@ebi.ac.uk),

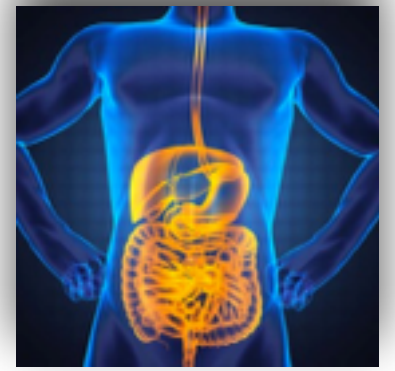
Open Door Workshop, Hinxton

13th May 2015

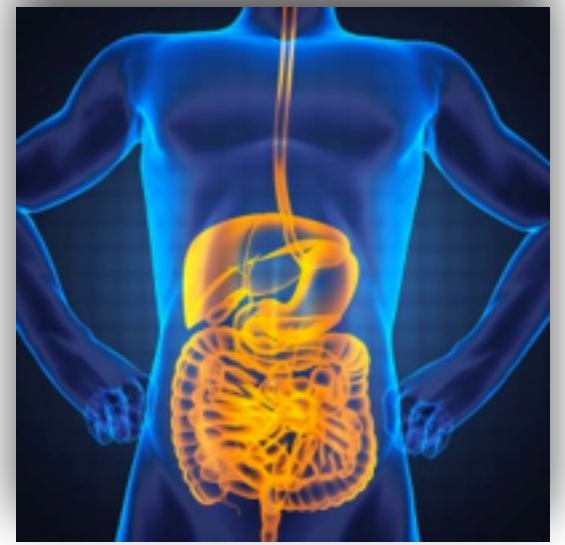


EMBL-EBI

Metagenomics - a broad range of applications



Metagenomics - a broad range of applications



ARTICLE

doi:10.1038/nature11234

Structure, function and diversity of the healthy human microbiome

The Human Microbiome Project Consortium*

OPEN ACCESS Freely available online PLOS | PATHOGENS

Targeted Restoration of the Intestinal Microbiota with a Simple, Defined Bacteriotherapy Resolves Relapsing *Clostridium difficile* Disease in Mice

Trevor D. Lawley[1]*, Simon Clare[1], Alan W. Walker[1], Mark D. Stares[1], Thomas R. Connor[1], Claire Raisen[1], David Goulding[1], Roland Rad[1], Fernanda Schreiber[1], Cordelia Brandt[1], Laura J. Deakin[1], Derek J. Pickard[1], Sylvia H. Duncan[2], Harry J. Flint[2], Taane G. Clark[3], Julian Parkhill[1], Gordon Dougan[1]

LETTER

doi:10.1038/nature11582

Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease

OPEN ACCESS Freely available online PLOS | ONE

Novel Gut-Based Pharmacology of Metformin in Patients with Type 2 Diabetes Mellitus

Antonella Napolitano[1]*, Sam Miller[2], Andrew W. Nicholls[3], David Baker[3], Stephanie Van Horn[4], Elizabeth Thomas[4], Deepak Rajpal[5], Aaron Spivak[5], James R. Brown[5], Derek J. Nunez[6]

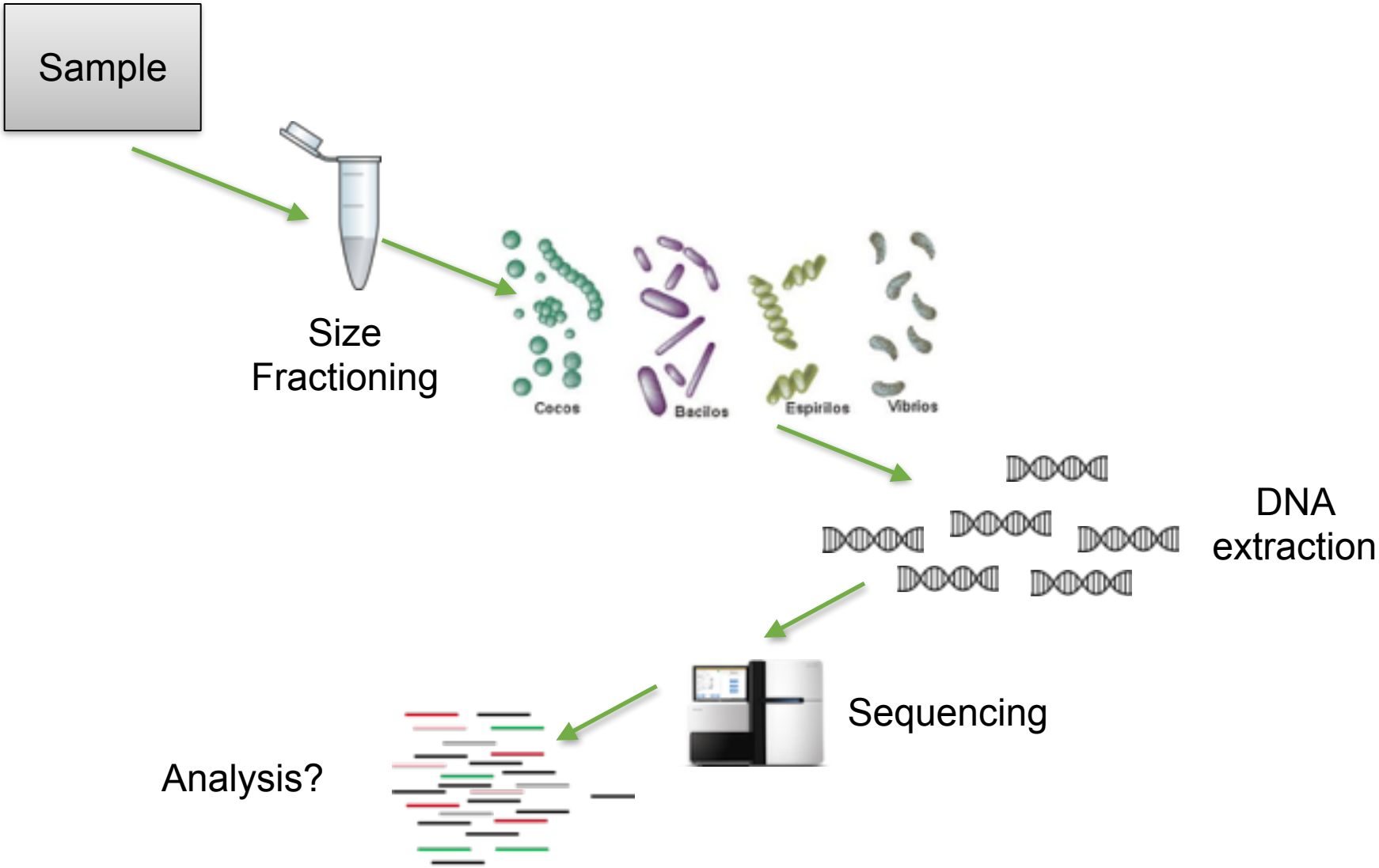
1 Immuno-Inflammation Unit, GSK R&D, Stevenage, Herts, United Kingdom, **2** Quantitative Sciences, GSK R&D, Stevenage, Herts, United Kingdom, **3** Safety Assessment, GSK R&D, Ware, Herts, United Kingdom, **4** Target and Pathways Validation, GSK R&D, Upper Providence, Pennsylvania, United States of America, **5** Computational Biology, GSK R&D, Upper Providence, Pennsylvania, United States of America, **6** Enteroendocrine Discovery Unit, GlaxoSmithKline R&D, GSK R&D, Research Triangle Park, North Carolina, United States of America

CMAJ RESEARCH

Gut microbiota of healthy Canadian infants: profiles by mode of delivery and infant diet at 4 months

Meghan B. Azad PhD, Theodore Konya MPH, Heather Maughan PhD, David S. Guttman PhD, Catherine J. Field PhD, Radha S. Chari MD, Malcolm R. Sears MB, Allan B. Becker MD, James A. Scott PhD, Anita L. Kozyrskyj PhD, on behalf of the CHILD Study Investigators

From environment to DNA sequence



From environment to DNA sequence

http://www.ebi.ac.uk/metagenomics

Overview: EBI Metagenomics Portal

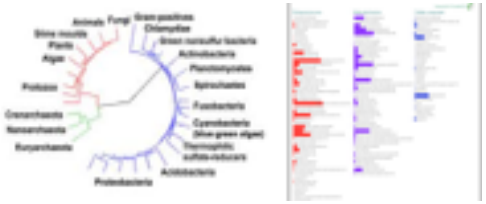
Easy submission

Submission of sequence data for archiving and analysis

Powerful analysis

Data **analysis** using selected EBI and external software tools

Visualisation



Data presentation and **visualisation** through web interface

Submitting to EBI Metagenomics

- EBI Metagenomics want to encourage people to supply as much detailed **metadata** as possible, but with the lowest possible overhead

who where, when, what how

name
institute
country …
contact

latitude and longitude,
depth,
salinity,
temperature,
time …

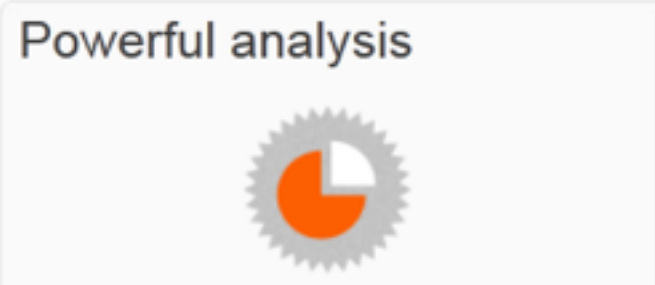
quantity,
conservation process,
storage conditions,
treatments,
extraction methods …

platform,
protocol,
filtering and QC,
analysis,
tool versions …

EBI Metagenomics

- Development of intuitive web-based tools : **ENA Webin** and **ISA tools**
- Use of templates and check-lists (MIGS/MIXS standards)
- Tutorial and direct support

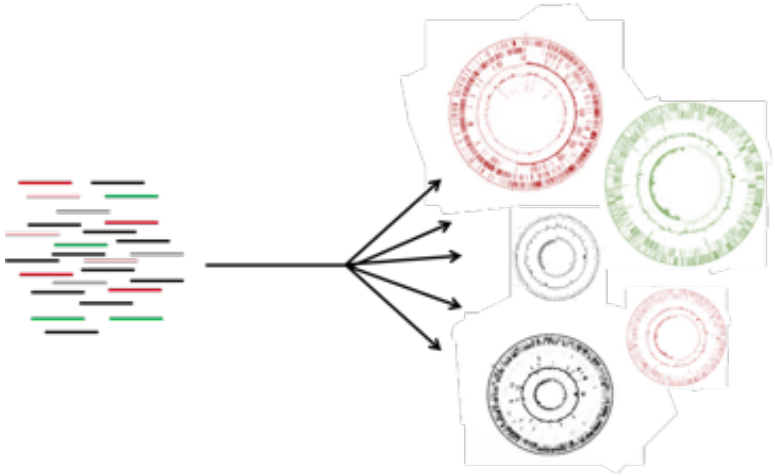
Overview: EMG Portal analysis



Data analysis using selected EBI and external software tools

- Provide robust sequence analysis services to all metagenomic researchers
  - ➢ Understand species diversity and functional potential of a community

Metagenomics assembly?



- Metagenomics: Not clear how you avoid assembling sequences from different species together : chimaera

- No reference sequence to align against

raw reads

ENA
European Nucleotide Archive

QC

discarded reads

processed reads

rRNAselector

reads without rRNA

reads with rRNA

FragGeneScan

predicted CDS

Amplicon -based data

Qiime

InterProScan

Unknown function pCDS

Taxonomic analysis

Function assignment

EMG portal does not used blast based homology methods

Instead reads are compared to models (signatures) generated from multi-sequences alignments:

- more specific and meaningful annotations
- faster annotation

- rRNASelector identify 5, 16 and 28s rRNA (profile HMM models)
  => 16s-based Qiime taxonomy annotations

- FragGenScan predict CDSs (HMM models)
  => InterProScan functional annotations (profiles and models)

Profile Hidden Markov Models

Input multiple alignment:

```
seq1   ACG-LD
seq2   SCG--E
Seq3   NCGgFD
Seq4   TCG-WQ
       123-45
```

Consensus columns assigned,
Defining inserts and deletes:

Profile Hidden Markov Models

Input multiple alignment:

Consensus columns assigned,
Defining inserts and deletes:

```
seq1   ACG-LD
seq2   SCG--E
Seq3   NCGgFD
Seq4   TCG-WQ
       123-45
```

Profile Hidden Markov Models

Input multiple alignment:

Consensus columns assigned,
Defining inserts and deletes:

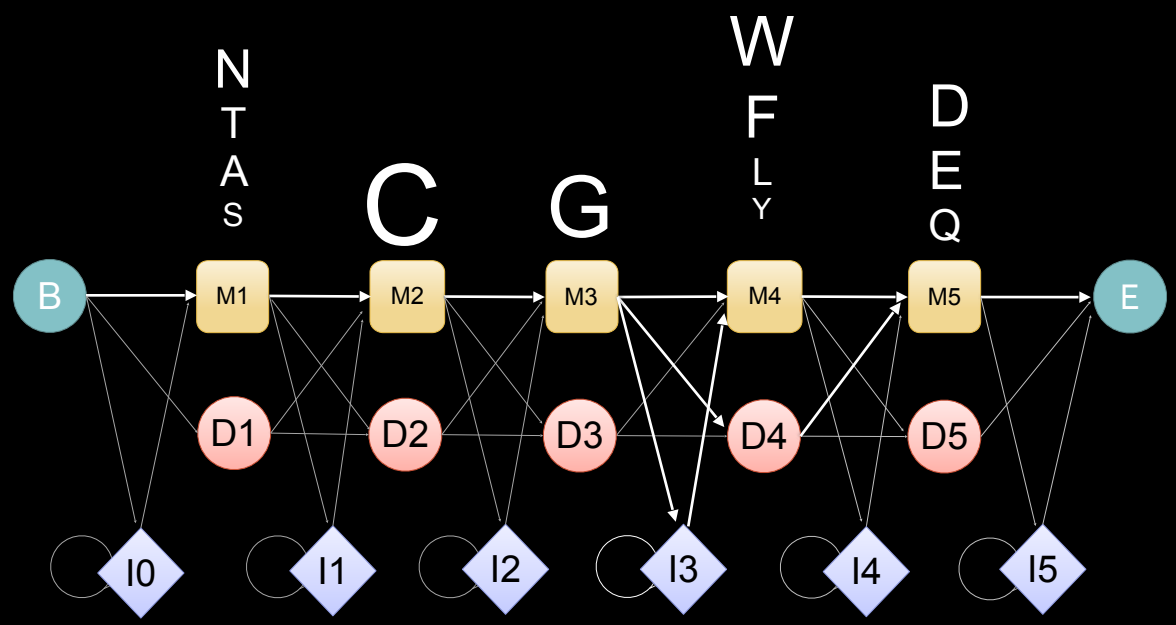
```
seq1   ACG-LD
seq2   SCG--E
Seq3   NCGgFD
Seq4   TCG-WQ
       123-45
```

Profile Hidden Markov Models

Input multiple alignment:

```
seq1   ACG-LD
seq2   SCG--E
Seq3   NCGgFD
Seq4   TCG-WQ
       123-45
```

Consensus columns assigned,
Defining inserts and deletes:



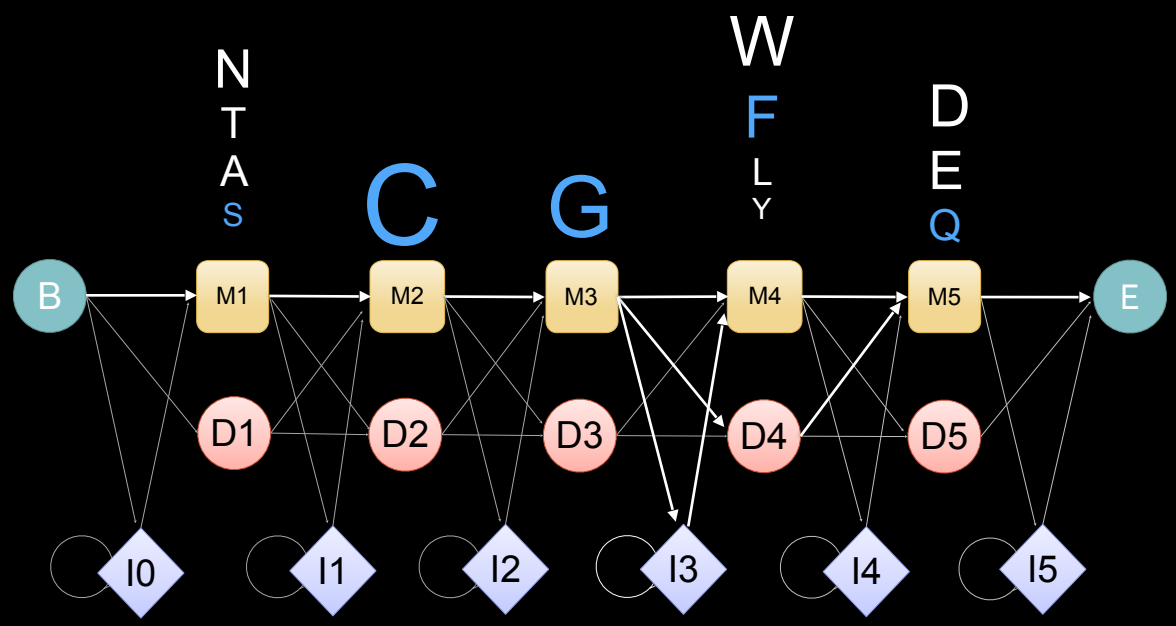
Plan7
core
model

Profile Hidden Markov Models

Input multiple alignment:

```
seq1   ACG-LD
seq2   SCG--E
Seq3   NCGgFD
Seq4   TCG-WQ
       123-45
```

Consensus columns assigned,
Defining inserts and deletes:



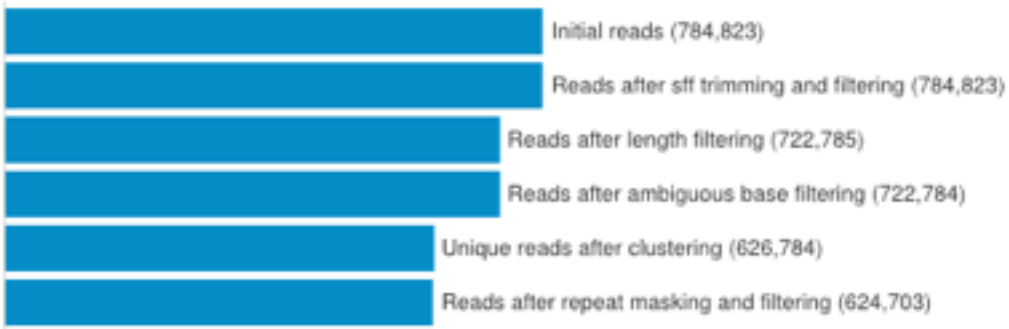
Plan7
core
model

QC processes

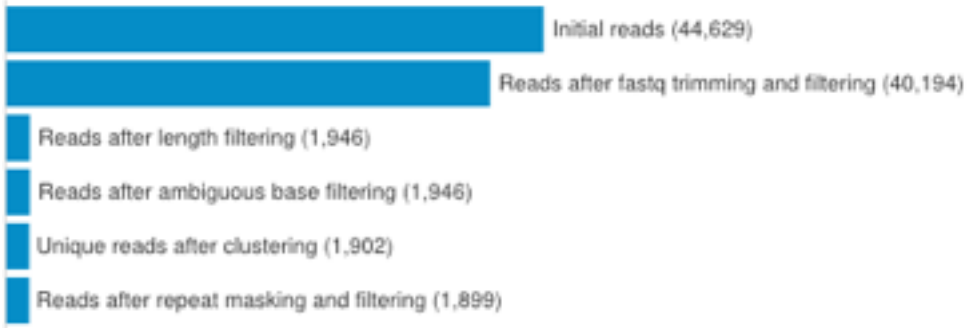
- Clipping - low quality ends trimmed and adapter sequences removed
- Quality filtering - sequences with > 10% undetermined nucleotides removed
- Read length filtering - short sequences (< 100 nt) are removed
- Duplicate sequences removal – clustered (99% identity UCLUST or 50 nt similarity Prefix) and representative sequence chosen
- Repeat masking - RepeatMasker (open-3.2.2), removes reads with 50% or more nucleotides masked (low complexity regions)

QC effects by sequencing platform

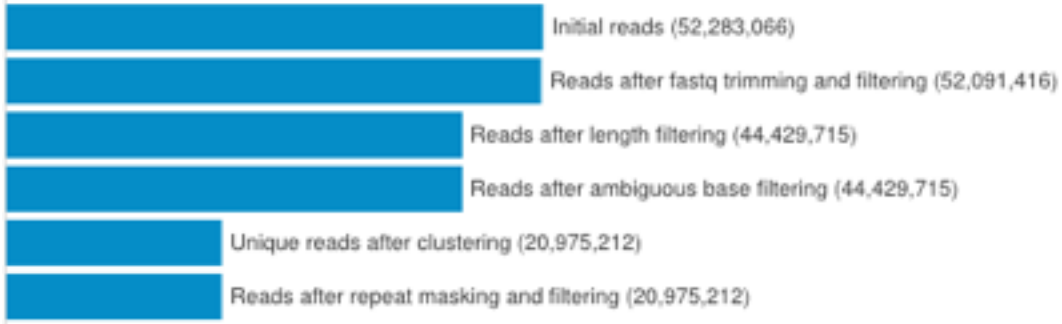
Roche 454



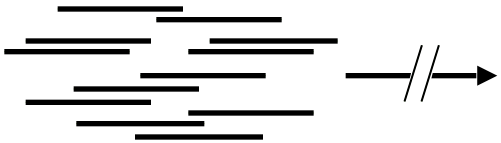
Ion Torrent



Illumina



EBI Metagenomics: taxonomic analysis



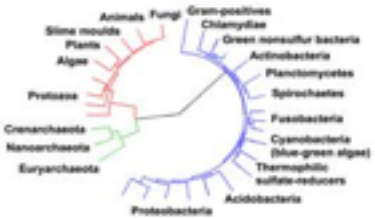
processed reads

rRNAselector

reads with rRNA

Amplicon-based data → Qiime

Taxonomic analysis



Common approaches to taxonomic analysis

- Identification of reads with 16S sequence (e.g. using rRNASelector) and closed-reference OTU picking in **QIIME**
- **Blast-based** analysis.
  - E.g. blasting reads against the NCBI non-redundant nucleotide or protein data databases and inferring taxonomic lineage from the best hit
  - The tool **MEGAN** requires Blast output. A major drawback is that without preprocessing of NGS datasets and access to a major computational resource, this is not an option for most.
- **MetaPhlAn** approach
  - (http://huttenhower.sph.harvard.edu/metaphlan)
  - relies on unique clade-specific marker genes identified from 3,000 reference genomes
  - fast, but limited to certain types of study (mainly human microbiome)

Taxonomic analysis

Currently only taxonomy analysis for Prokaryotes

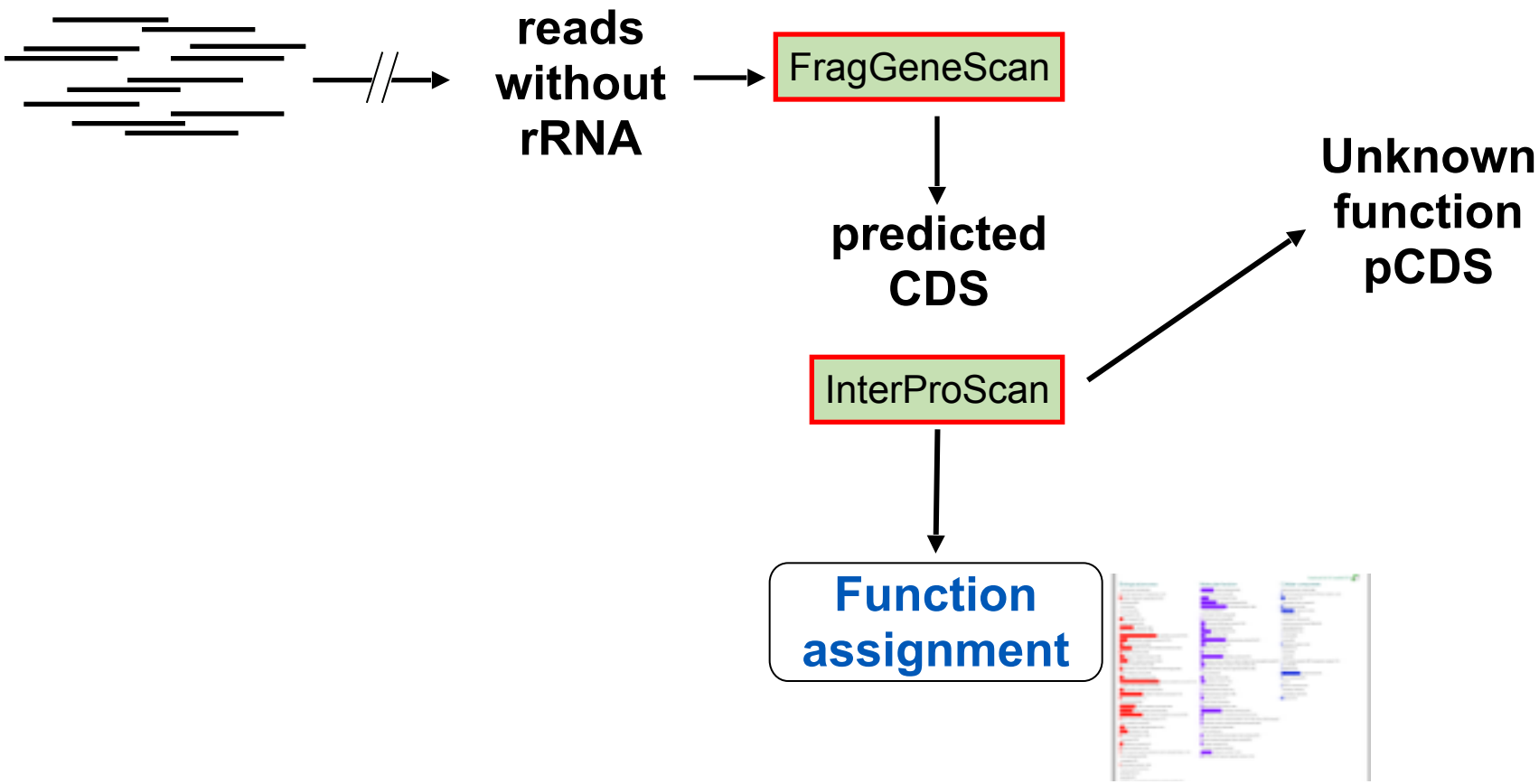
rRNA sequences are identified using rRNASelector:

- hidden Markov models to identify rRNA sequences
- 60 bp minimum overlap with curated HMM model
- E-value < 10^{-5}

Annotations are associated using Qiime:

- rRNA are annotated using the Greengenes reference database

EBI Metagenomics: functional analysis



EBI Metagenomics: functional annotation

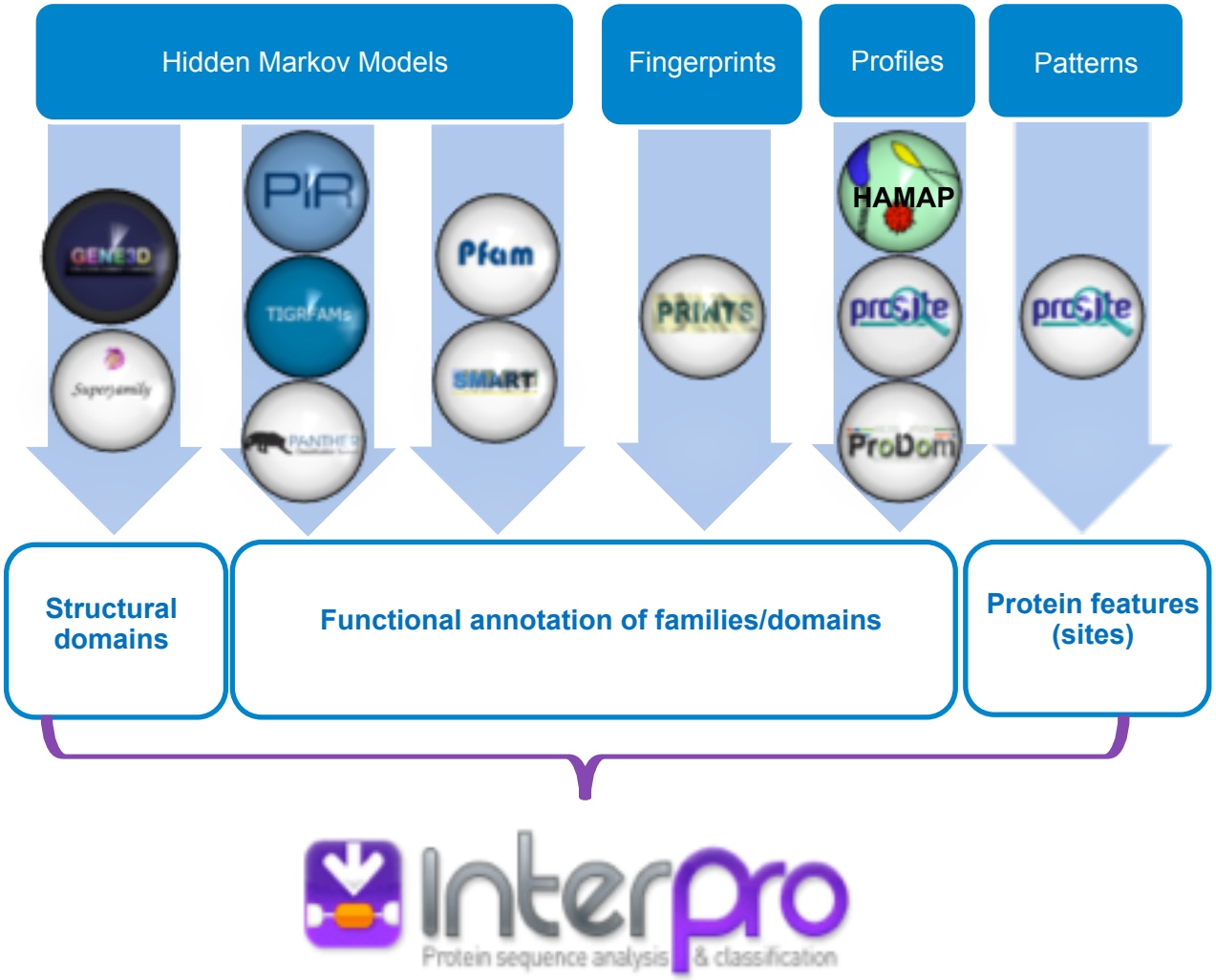
The pipeline uses FragGeneScan to predict CDSs directly from the reads:

- hidden Markov models to correct frame-shift using codon usage
- probabilistic identification of start and stop codons
- 60 bp minimum ORF

Annotation is carried out using InterProScan with a subset of InterPro's databases

- analysis speed
- ability to cope with sequence fragments

The benefits of InterPro



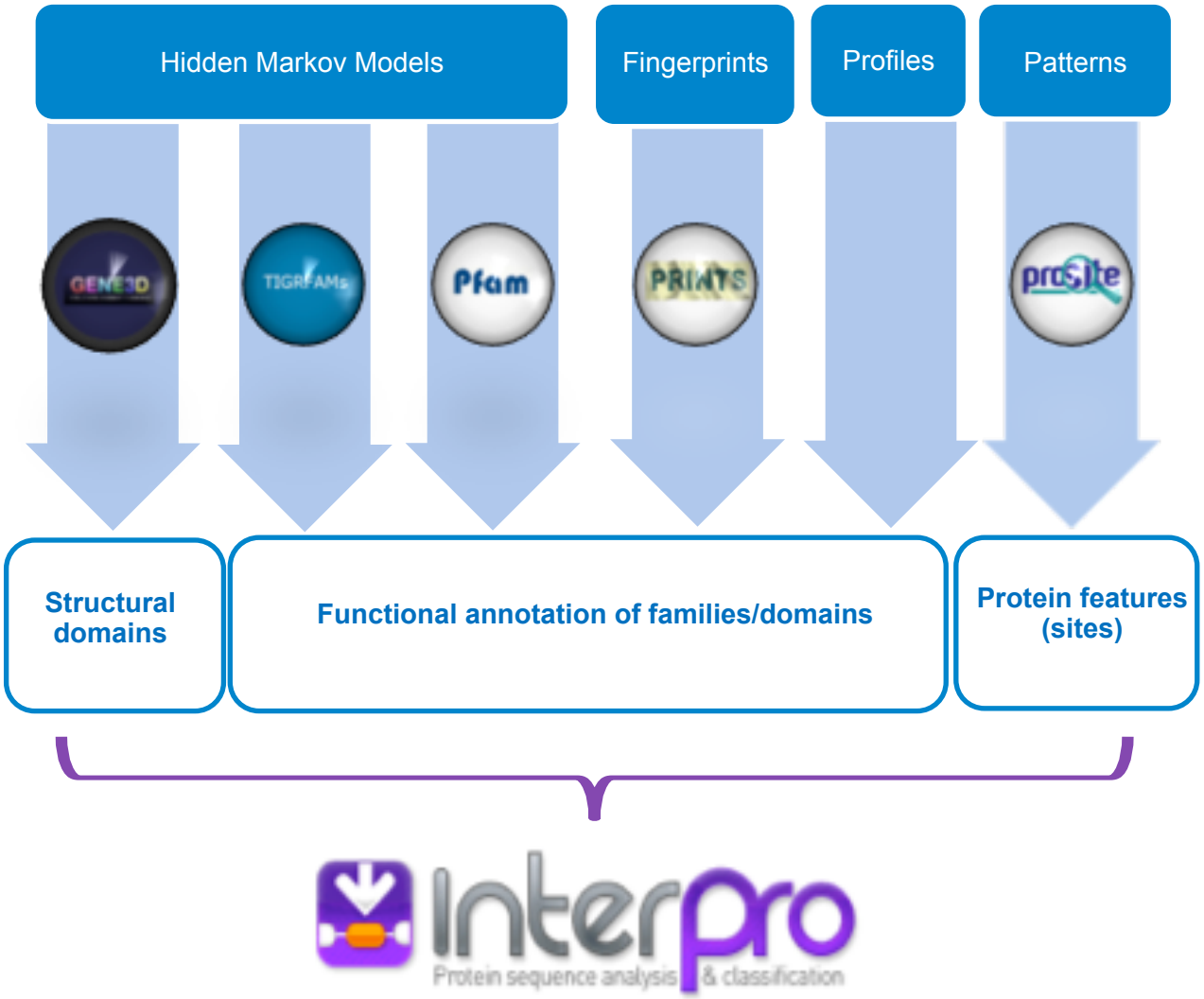
Using InterPro for annotation

- Underlies automated systems that annotate UniProtKB/TrEMBL
- Provides matches to **90 million** proteins - **over 80%** of UniProtKB
- Source of **~ 170 million** GO mappings for **~ 50 million** distinct UniProtKB sequences

Annotation consistency

- Using InterPro and GO allows direct comparison with proteins in UniProtKB

InterPro in the Metagenomics Portal

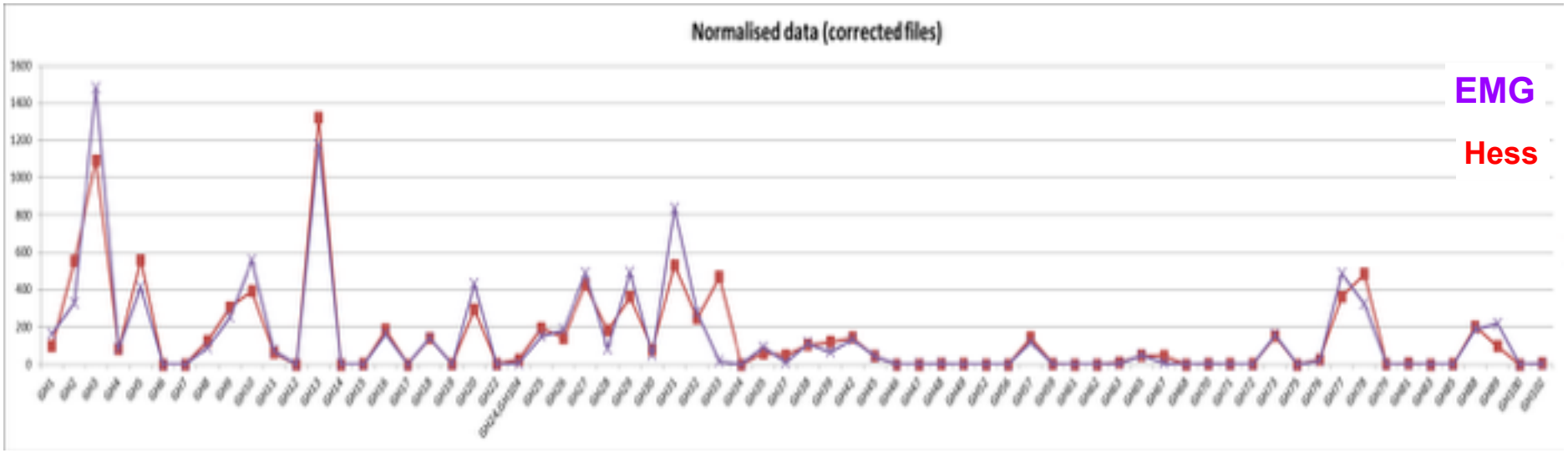


Annotations without assembly

Re-analysis of Hess et al, Science (2011) 331:463

Metagenomic Discovery of Biomass-Degrading Genes and Genomes from *Cow Rumen*

Comparison of the normalised number of genes / reads corresponding to CAZy Glycoside Hydrolase Family from the Hess et al paper and from the EMG pipeline.

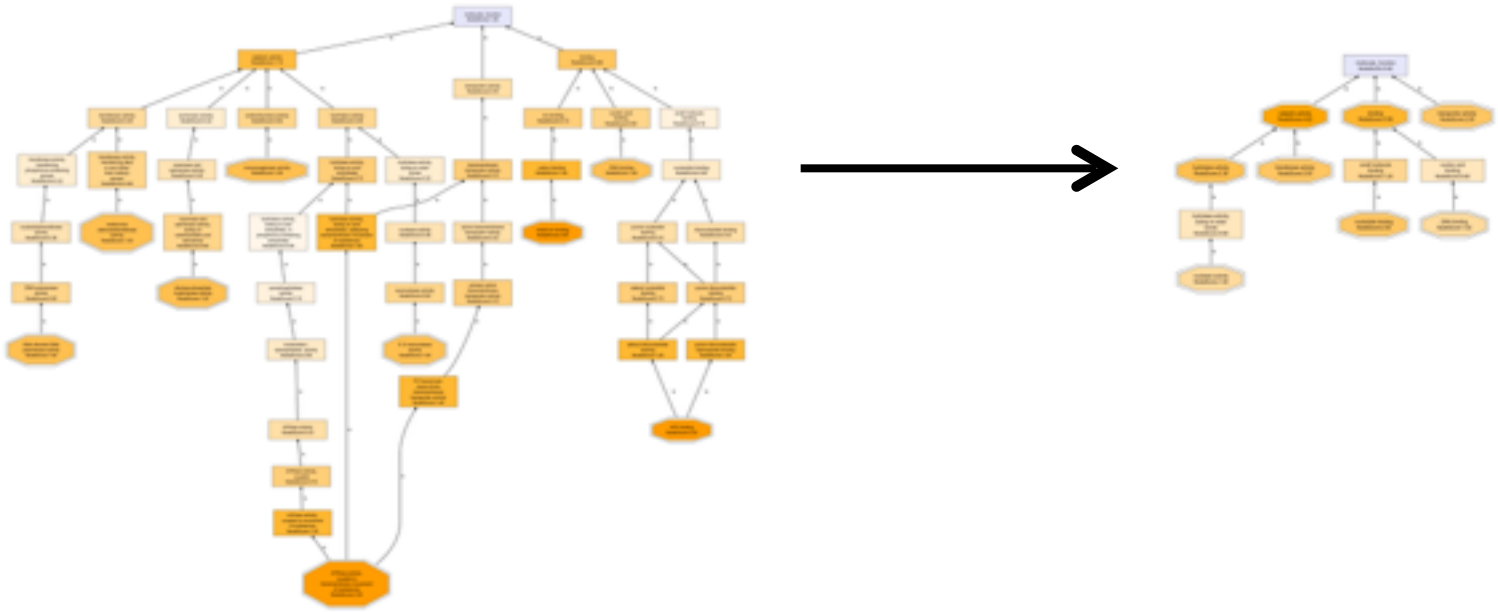


Hess et al: genome assembly then gene prediction using a subset of Pfam.
EMG pipeline: no assembly and gene prediction using InterPro.

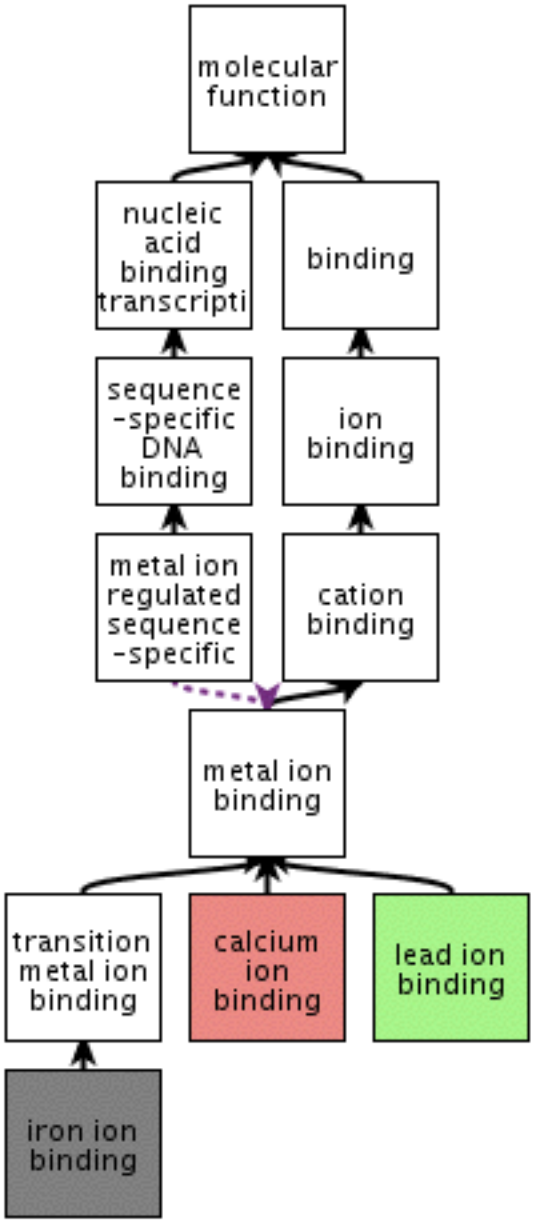
Discrepancies are due to the different ways in which significance cut-off are calculated.

Visualising data: GO Slims

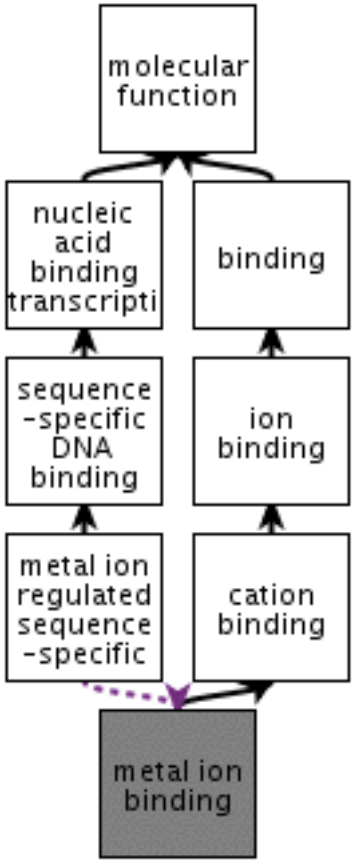
- Cut-down versions of the GO containing a subset of terms
- Give a broad overview of the ontology content without the detail of the specific fine-grained terms



GO Slims


molecular function
nucleic acid binding transcripti
binding
sequence -specific DNA binding
ion binding
metal ion regulated sequence -specific
cation binding
metal ion binding
transition metal ion binding
calcium ion binding
lead ion binding
iron ion binding

GO Slims



Slimmed term:

Metagenomics - Big Data

- Speed is really important! MAY

- Submitted nucleotide sequences: **27,509,856,436**
- Average length per sequence: **120 nt**
- Predicted CDS: **8,167,600,355**
- Total InterProScan matches: **1,866,871,818**
- Number of different samples: **2,808**

- **99.8%** of this data arrived and was processed in the last two years.
  96.5% of it is publicly available via the website.

Metagenomics - Big Data

- Speed is really important! MONDAY

- Submitted nucleotide sequences: **43,315,534,332**
- Average length per sequence: **120 nt**
- Predicted CDS: **17,301,862,307**
- Total InterProScan matches: **3,276,195,744**
- Number of different samples: **4,330**

- **99.8%** of this data arrived and was processed in the last two years.
  96.5% of it is publicly available via the website.

Downstream analysis: download options

Overview | Quality control | Taxonomy analysis | Functional analysis | **Download**

You can download in this section the full set of analysis results files and the original raw sequence reads.

Sequence data

- Submitted nucleotide reads (ENA website)
- Processed nucleotide reads (FASTA) - 2 MB
- Processed reads with pCDS (FASTA) - 2 MB
- Processed reads with InterPro matches (FASTA) - 1 MB
- Processed reads without InterPro match (FASTA) - 835 KB
- Predicted CDS (FASTA) - 710 KB
- Predicted CDS with InterPro matches (FASTA) - 451 KB

Functional Analysis

- InterPro matches (TSV) - 1 MB
- Complete GO annotation (CSV) - 44 KB
- GO slim annotation (CSV) - 7 KB

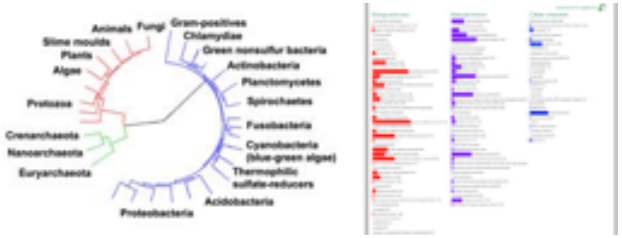
Taxonomic Analysis

- Reads encoding 5S rRNA (FASTA) - 565 bytes
- Reads encoding 16S rRNA (FASTA) - 21 KB
- Reads encoding 23S rRNA (FASTA) - 37 KB
- OTUs and taxonomic assignments (BIOM) - 6 KB
- Phylogenetic tree (Newick format) - 289 bytes
- OTUs and taxonomic assignments (TSV) - 2 KB

relatively small result files: can be used for downstream analysis with other tools

Overview: EMG Portal output

Visualisation



Data presentation and visualisation through web interface

- Assist laboratory researchers handle and make sense of massive volumes of sequence data
  - ➢ Do this by designing intuitive, user-friendly web interfaces
  - ➢ Browse, visualise and download

EMBL-EBI

Services | Research | Training | About us

EBI Metagenomics

Home | Submit data | Projects | Samples | About EBI Metagenomics | Contact

Hubert DENISE (edit) | logout

EBI Metagenomics > Projects

Projects list

Text:

Privacy: All projects

Search | Clear

Public projects can be browsed and searched using names and keywords

1 - 10 of 41

Download detailed info (CSV) | Download table (CSV)

Project name	Samples	Submitted date
A core gut microbiome in obese and lean twins	18	03-Mar-2010
A metagenomics transect into the deepest point of the Baltic Sea reveals clear stratification of microbial functional capacities	6	27-Sep-2013
Antarctica Aquatic Microbial Metagenome	18	12-Jul-2011
Arctic Winter marine ecosystem	1	19-Mar-2013
BGI Type 2 Diabetes study (SRP008047)	145	14-Mar-2013
BGI Type 2 Diabetes study (SRP011011)	218	26-Mar-2013
Beta Lactam Antibiotics and Human Gut Microbiota	12	23-Sep-2013
Brazos-Trinity Basin Sediment Metagenome: IODP Site 1320	1	04-Dec-2013
Buffalo rumen metagenomics	1	20-Mar-2013
Comparative freshwater metagenomics of Swedish and American lakes	0	04-Dec-2013

EMBL-EBI | Services | Research | Training | About us

EBI Metagenomics

Home | Submit data | Projects | Samples | About EBI Metagenomics | Contact

EBI Metagenomics » Project: A core gut microbiome in obese and lean twins

Project overview (SRP000319)

A core gut microbiome in obese and lean twins

BioProject ID: PJ32089

Submitted date: 03-Mar-2010

Description

We have characterized the fecal microbial communities of adult female monozygotic and dizygotic twin pairs concordant for leanness or obesity, and their mothers. The results demonstrate that a diversity of organismal assemblages can nonetheless yield a core microbiome at a functional level, and that deviations from this core are associated with different physiologic states (obese versus lean)

Experimental factor: obese vs lean

Related Publications

- A core gut microbiome in obese and lean twins. Turnbaugh PJ, Hamady M, Yatsunenko T. 2009 Nature. 2009 Jan 22;457(7228):480-4. Epub 2008 Nov 30.

Contact details

Institute: CGS-GL

Name: Turnbaugh PJ

Associated samples

Sample name	Sample ID	Collection date	Source	Analysis results
Obese human gut (patient TS29)	SRS009826	-	Host associated	Taxonomy | Function |
Obese human gut (patient TS28)	SRS009825	-	Host associated	Taxonomy | Function |
Lean human gut (patient TS2)	SRS000999	-	Host associated	Taxonomy | Function |
Lean human gut (patient TS1)	SRS000998	-	Host associated	Taxonomy | Function |
Obese human gut (patient TS19)	SRS001007	-	Host associated	Taxonomy | Function |
Overweight human gut (patient TS9)	SRS001006	-	Host associated	Taxonomy | Function |
Lean human gut (patient TS8)	SRS001005	-	Host associated	Taxonomy | Function |
Lean human gut (patient TS7)	SRS001004	-	Host associated	Taxonomy | Function |
Obese human gut (patient TS6)	SRS001003	-	Host associated	Taxonomy | Function |
Lean human gut (patient TS5)	SRS001002	-	Host associated	Taxonomy | Function |

Overall project description

Project-associated publication(s)

Samples in the study

Direct link to analysis results

EMBL-EBI

Services | Research | Training | About us

EBI Metagenomics

Home | Submit data | Projects | Samples | About EBI Metagenomics | Contact | Logout

EBI Metagenomics » Project: A core gut microbiome in obese and lean twins » Sample: Obese human gut (patient TS28)

Sample (SRS009825)

Obese human gut (patient TS28)

Overview | Quality control | Taxonomy analysis | Functional analysis | Download

Description

Sample Characteristics - The MOAFTS twin cohort, comprised of female likesex twin pairs, were identified from Missouri birth records over the period 1994-1999,when the twins had a median age of 15 years. A total of 350 twins from the larger MOAFTS cohort completed screening interviews for the present study. We were able to take advantage of the wave five interview of the MOAFTS twin cohort (which has 90% retention of wave four participants) to identify pairs most likely to meet study criteria.Eligibility was then confirmed at screening interview.

Classification: Host-associated > Human > Digestive system > Large intestine > Fecal

Project: A core gut microbiome in obese and lean twins (SRP000319)

Host associated

Species: *Homo sapiens* Tax ID 9606

Sex: Female

Phenotype: Obese

Localisation

Geographic location: United States of America; Missouri

Other information

Sex	female
Geographic location (longitude)	not available
Geographic location (country and/or sea,region)	USA; Missouri
Collection date	not available
Environment (biome)	human microbiome
Environment (feature)	human-associated habitat ENVO:00009003
Environment (material)	faeces ENVO:00002003
Sequencing method	pyrosequencing
Family relationship	mono-zygotic twin
Host common name	Homo sapiens
Host subject id	F10T1Ob1
Host taxid	9606
Miscellaneous parameter	period without antibiotics; >6 months
Phenotype	obese
Geographic location (latitude)	not available
Alt id - submitted sample id	TS28
GOLD sample classification	Host-associated > Human > Digestive system > Large intestine > Fecal
NCBI sample classification	77133

QC, analysis results and download tabs

Data in related resources

Descriptive meta-data

EBI Metagenomics: QC tab

EMBL-EBI

Services | Research | Training | About us

EBI Metagenomics

Home | Submit data | Projects | Samples | About EBI Metagenomics | Contact

Hubert DENISE (edit) logout

EBI Metagenomics > Project: A core gut microbiome in obese and lean twins > Sample: Obese human gut (patient TS28)

Sample (SRS009825)

Obese human gut (patient TS28)

Overview | Quality control | Taxonomy analysis | Functional analysis | Download

The chart below shows the number of sequence reads which pass each of the quality control steps we have implemented in our pipeline. Note that, for paired-end data, sequence merging may have occurred and so the initial number of reads may differ from what is in the ENA. For more details about the data processing we employ, please see the about page.

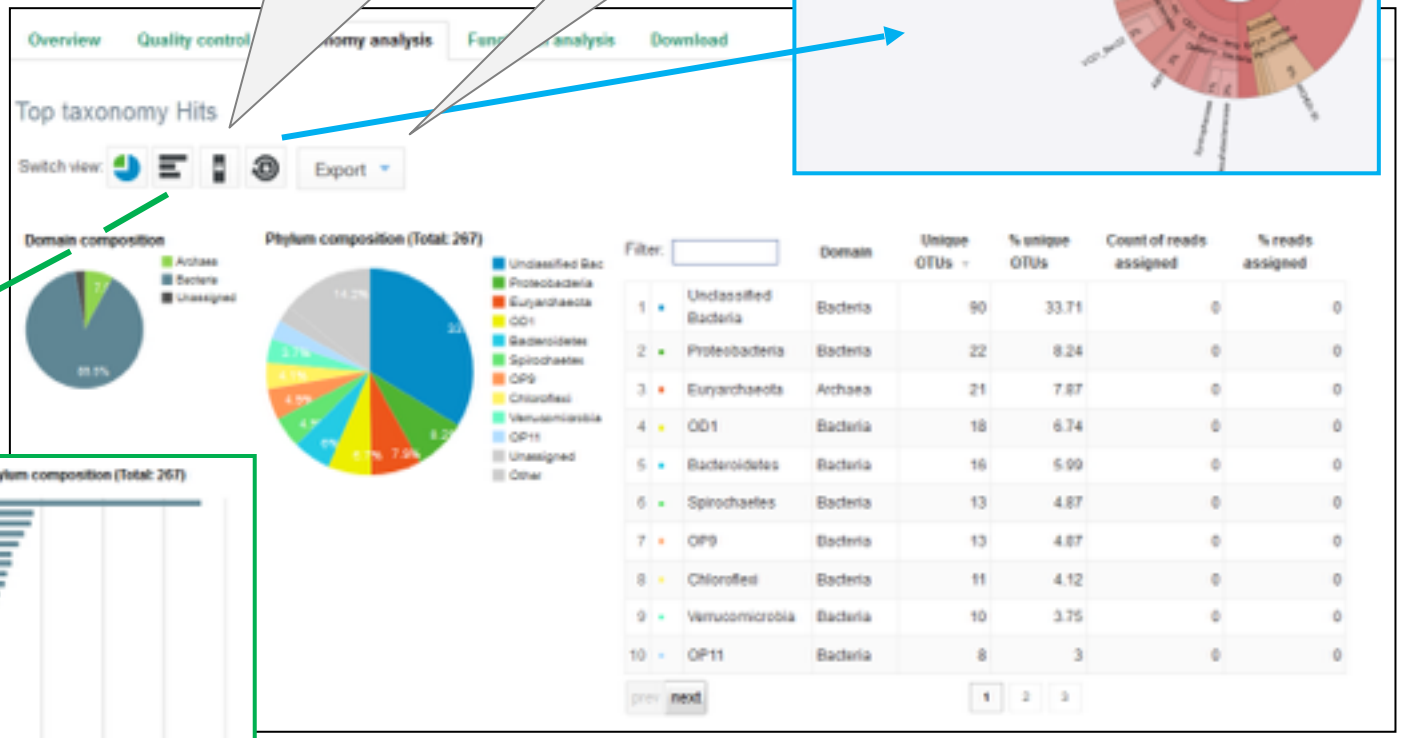
Sequence counts

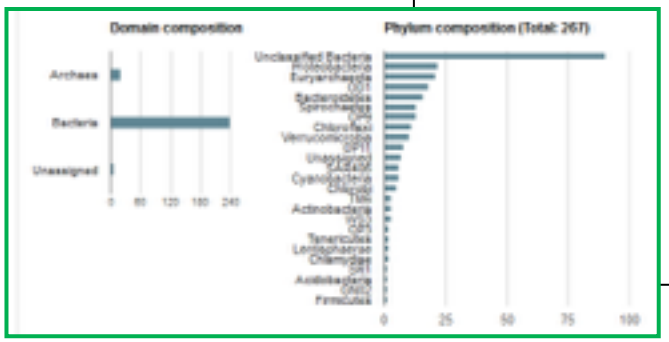
EBI Metagenomics: taxonomy analysis tab

Google charts dynamic representation

Switch to bar chart, column or Krona interactive views

Export charts





EBI Metagenomics: functional analysis tab

Google charts dynamic representation

Links to InterPro website

Export charts

Switch to bar chart, view

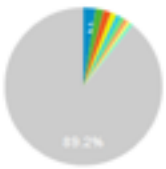
Overview | Quality control | Taxonomy analysis | Functional analysis | Download

InterPro match summary

Most frequently found InterPro matches to this sample:

Export

InterPro matches summary (Total: 5452)



- NAD(P)-binding domain
- Outer membrane insertion C-terminal signal, omp85 target
- Aldolase-type TIM barrel
- Rossmann-like alpha/beta/alpha sandwich fold
- Pyridoxal phosphate-dependent transferase, major region, subdo
- Tetratricopeptide-like helical
- Winged helix-turn-helix transcription repressor DNA-binding
- ATPase-like, ATP-binding domain
- ABC transporter-like
- Other

Filter:

	Entry name	ID	Hits
1	NAD(P)-binding domain	IPR016040	5546
2	Outer membrane insertion C-terminal signal, omp85 target	IPR017690	3572
3	Aldolase-type TIM barrel	IPR013785	3400
4	Rossmann-like alpha/beta/alpha sandwich fold	IPR014729	2674
5	Pyridoxal phosphate-dependent transferase, major region, subdomain 1	IPR015421	2075
6	Tetratricopeptide-like helical	IPR011990	1855
7	Winged helix-turn-helix transcription repressor DNA-binding	IPR011991	1631
8	ATPase-like, ATP-binding domain	IPR003594	1493
9	ABC transporter-like	IPR003439	1464
10	Signal transduction response regulator, receiver domain	IPR001789	1354

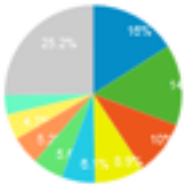
prev next 1 2 10 100 500 540 546

GO Terms annotation

A summary of Gene Ontology (GO) terms derived from InterPro matches to your sample is provided in the charts below.

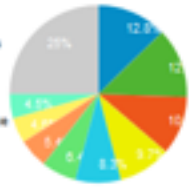
Switch view: Export

Biological process



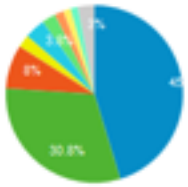
- nitrogen compound metabolic proc
- biosynthetic process
- oxidation-reduction process
- small molecule metabolic process
- protein metabolic process
- transport
- RNA metabolic process
- cellular amino acid metabolic proce
- DNA metabolic process
- Other

Molecular function



- nucleotide binding
- oxidoreductase activity
- hydrolase activity
- transferase activity
- nucleic acid binding
- cofactor binding
- transporter activity
- ligase activity
- metal ion binding
- Other

Cellular component

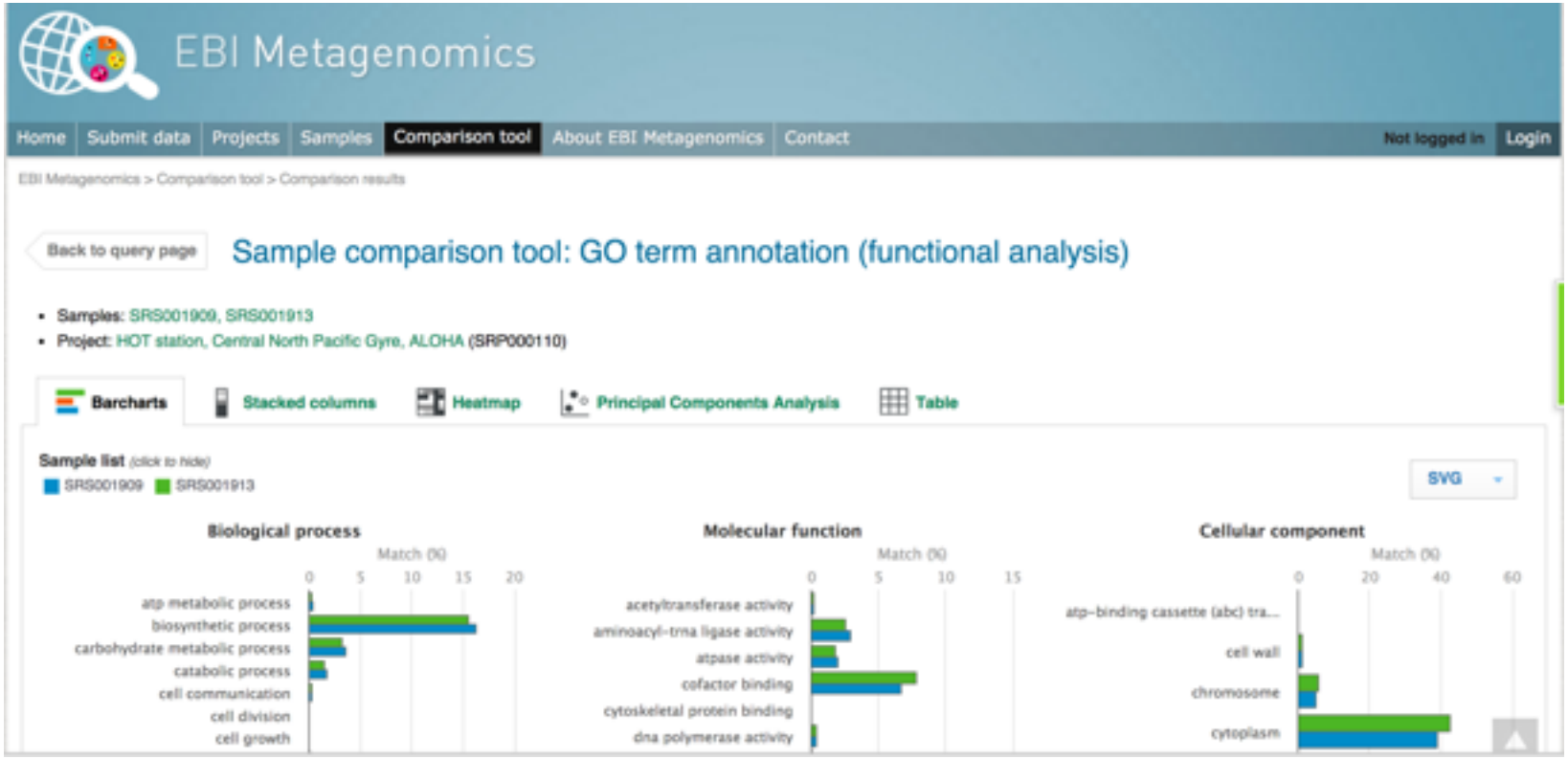


- membrane
- cytoplasm
- ribosome
- proton-transporting two-sector ATPa
- chromosome
- periplasmic space
- flagellum
- outer membrane
- plasma membrane
- Other

EMBL-EBI

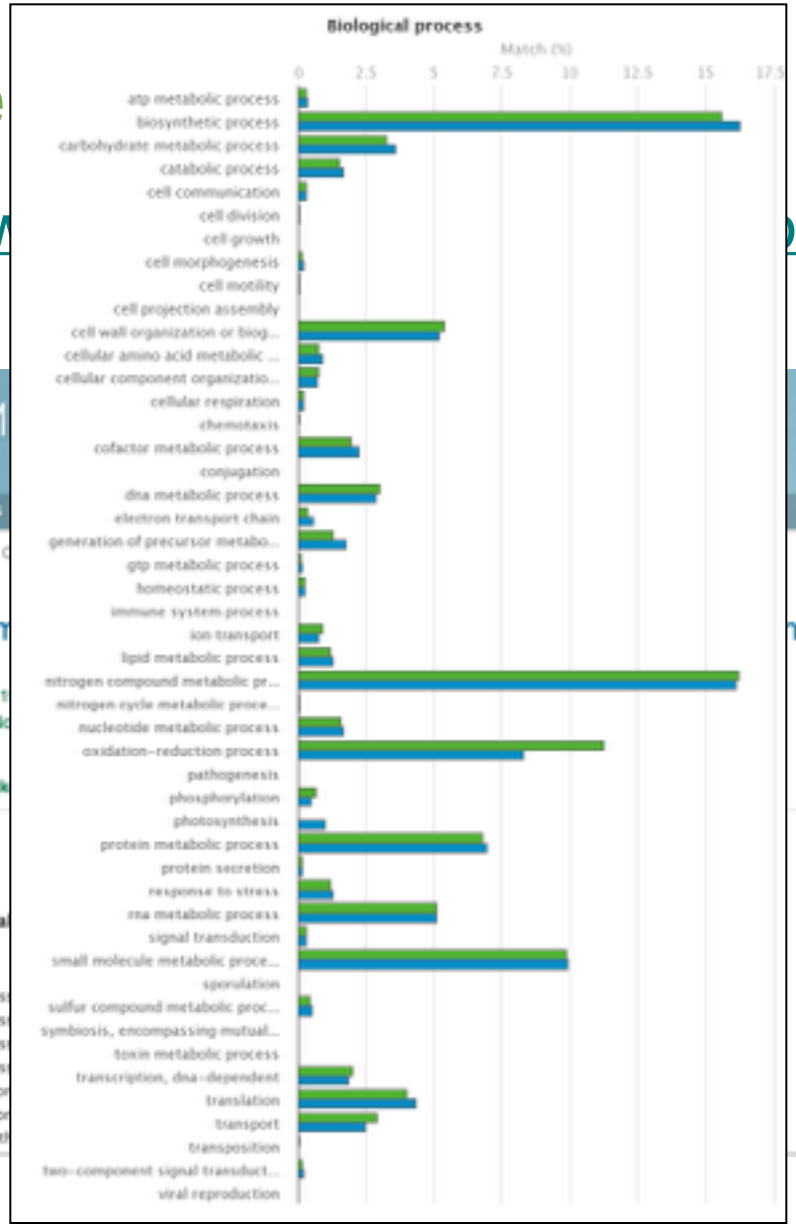
Sample Comparisons

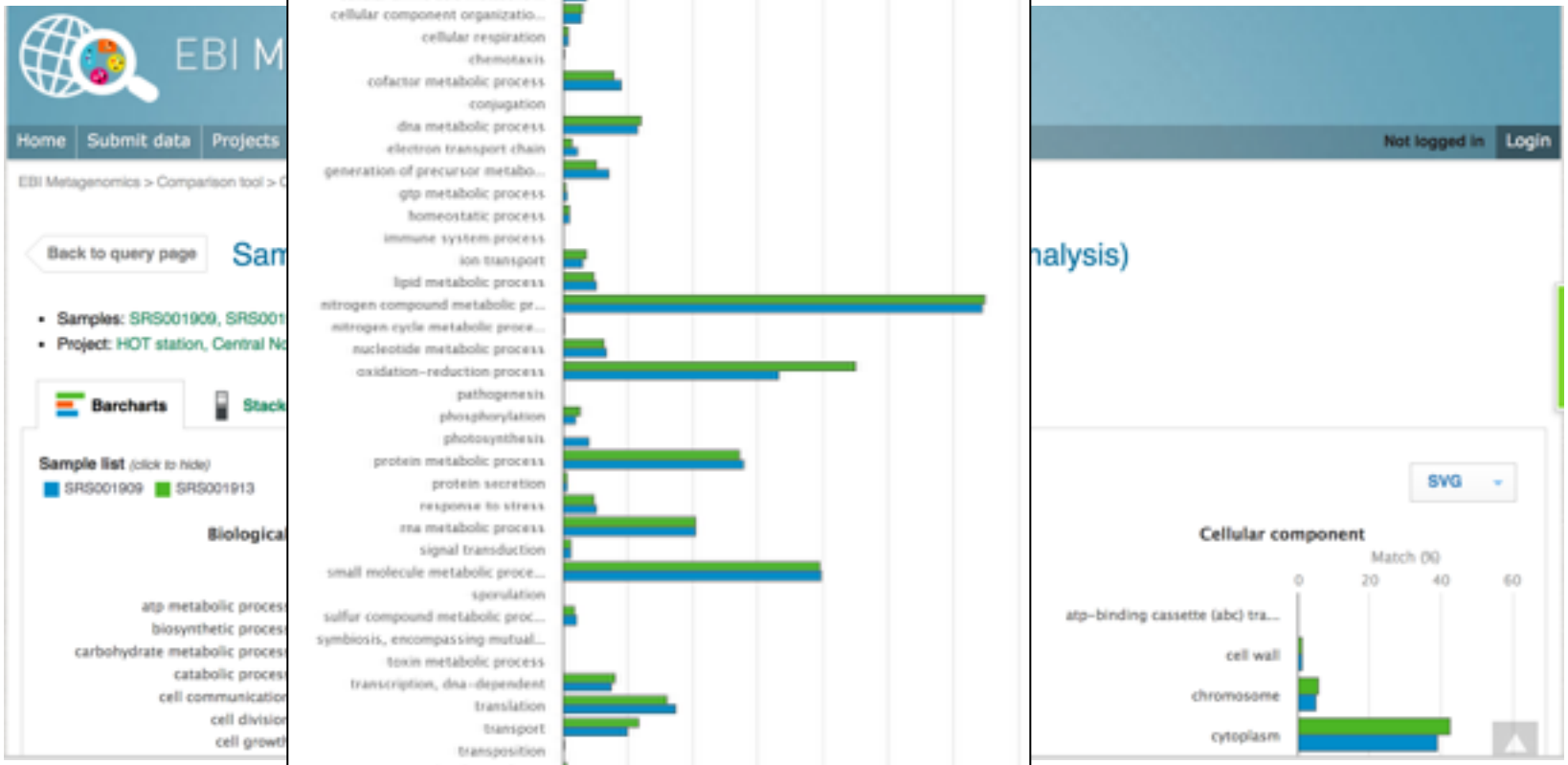
- https://www.ebi.ac.uk/metagenomics/compare



Sample

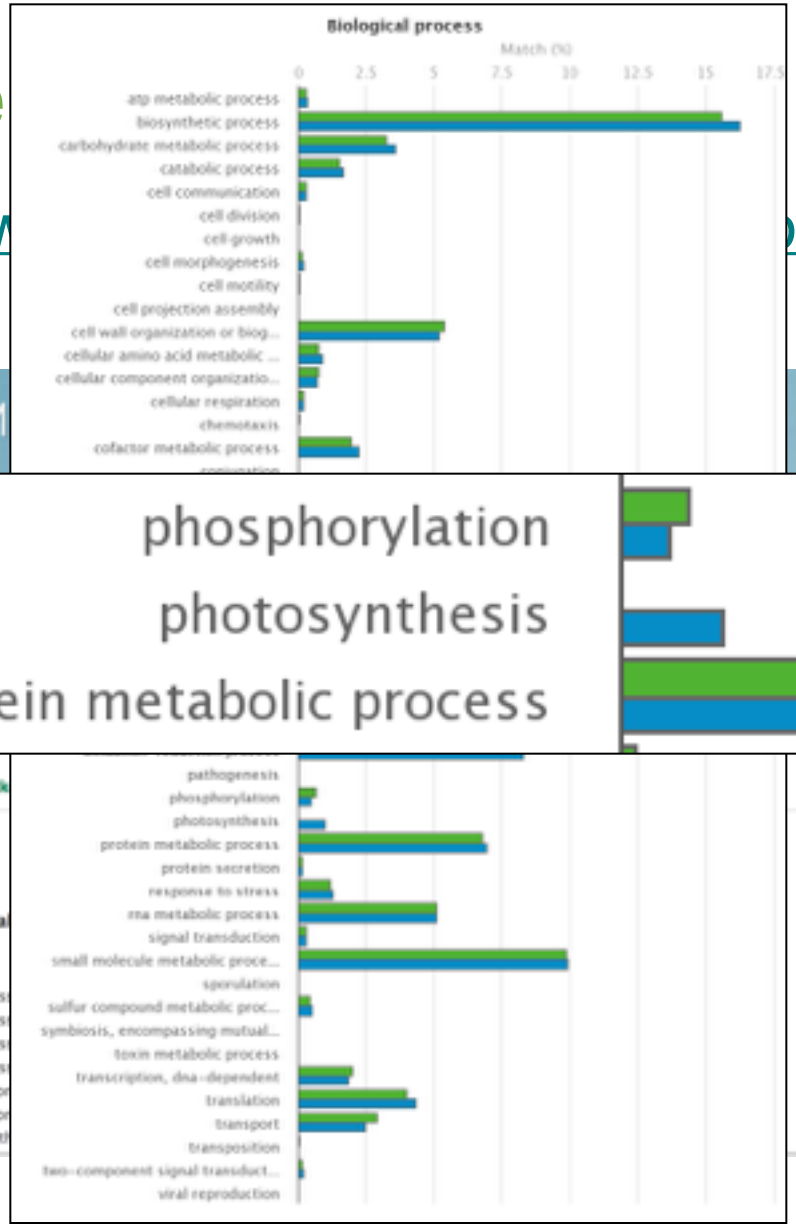
- https://w omics/compare

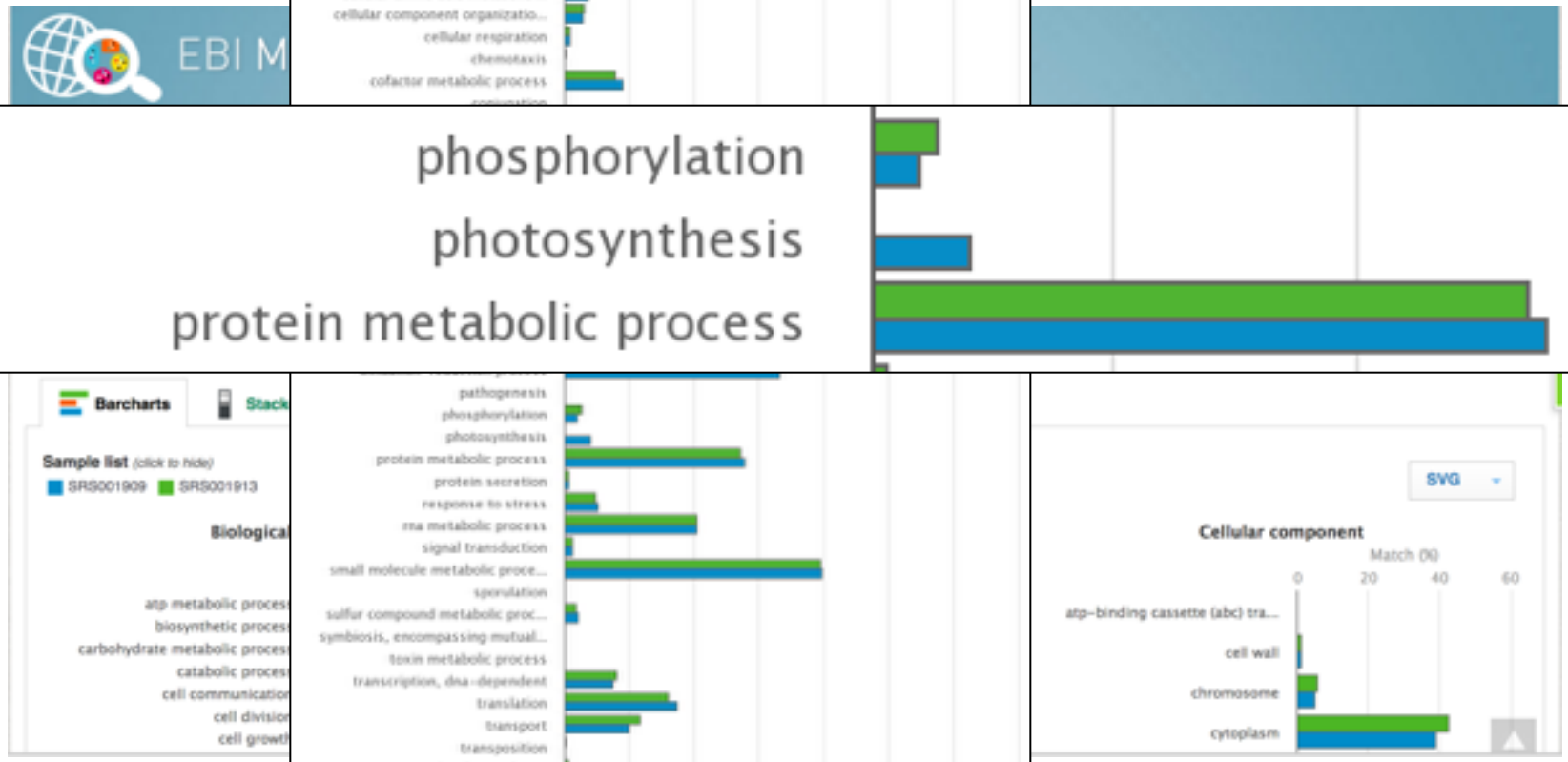




Sample

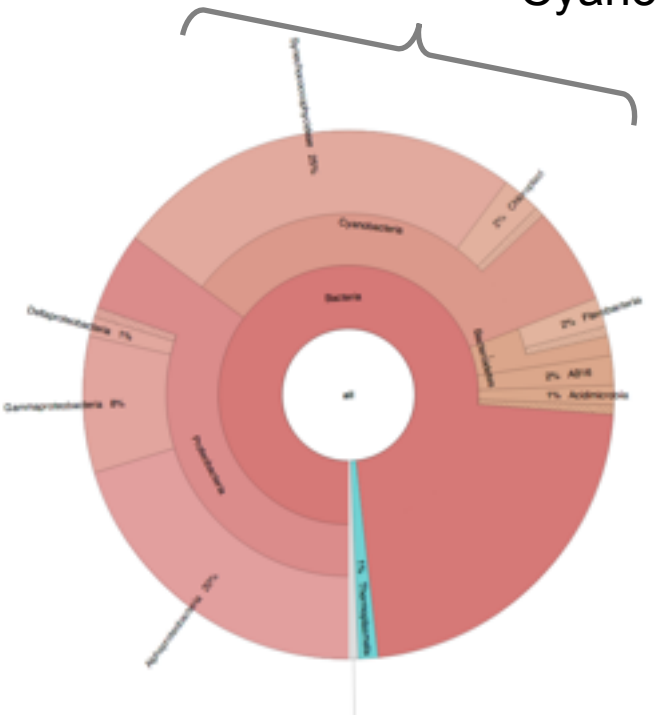
- https://w omics/compare





Comparison of two Marine Biomes - Taxonomic distributions

Cyanobacteria



Larger archaeal proportion

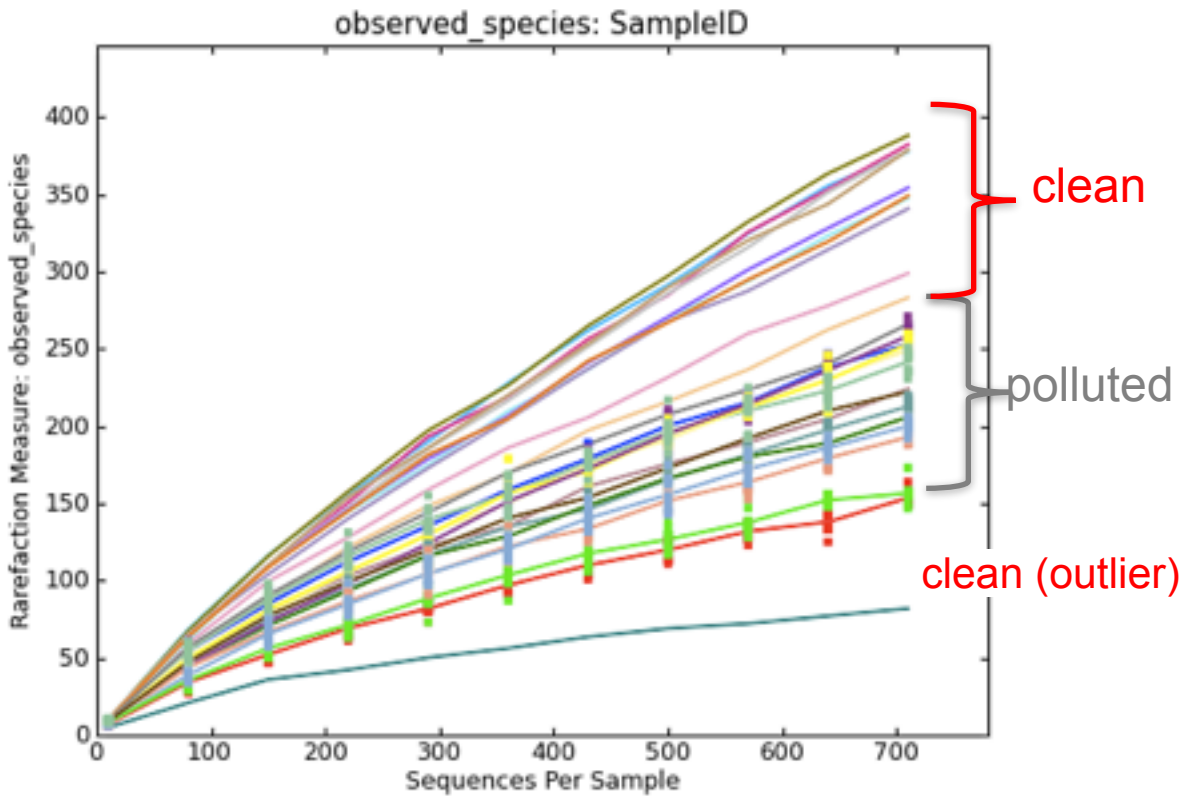
25m Depth

500m Depth

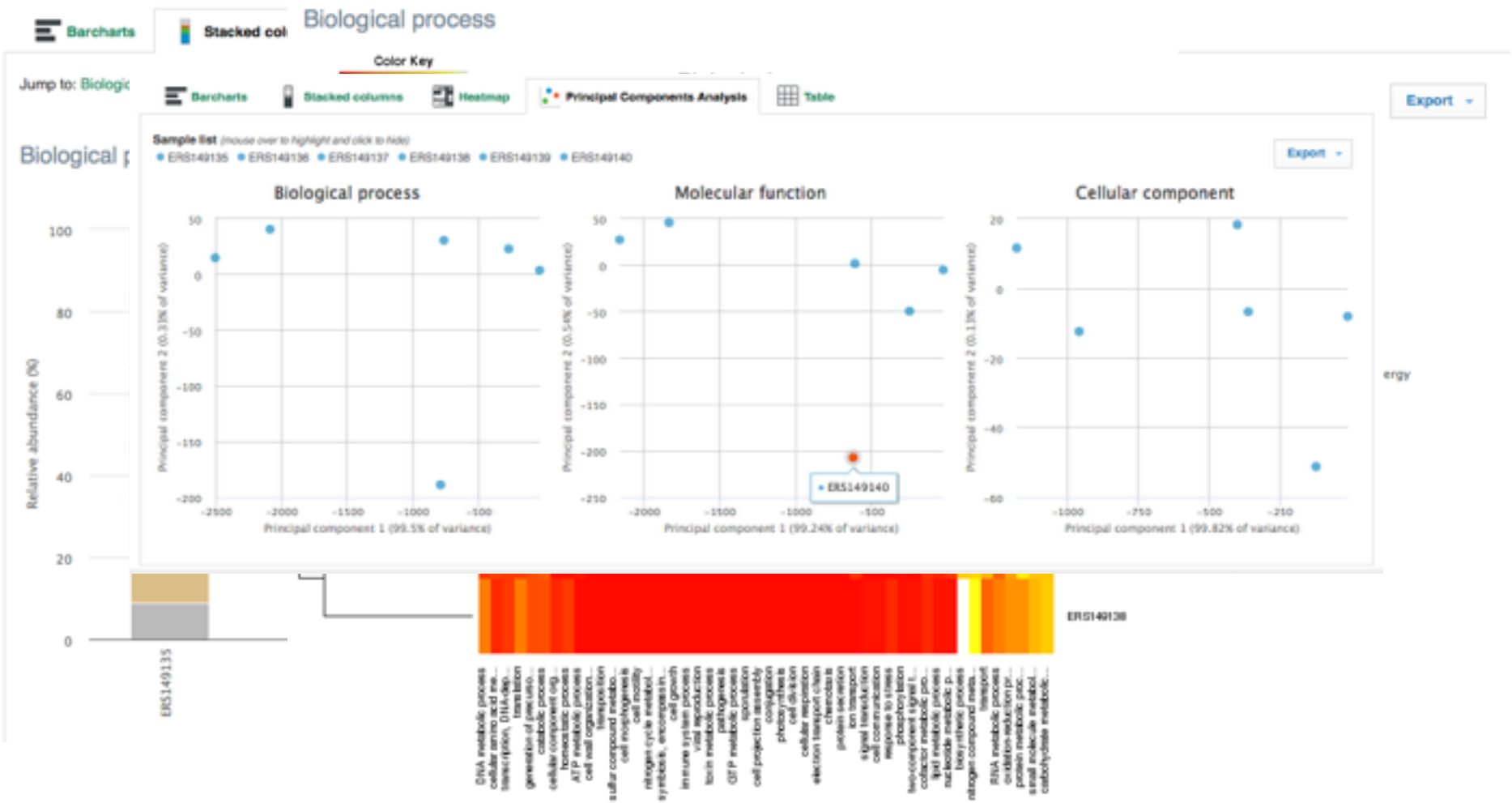
EBI Metagenomics: application of taxonomy analysis

Sutton et al, Appl. Environ. Microbiol (2013), 79(2):619
Impact of Long-Term Diesel Contamination on Soil Microbial Community Structure.

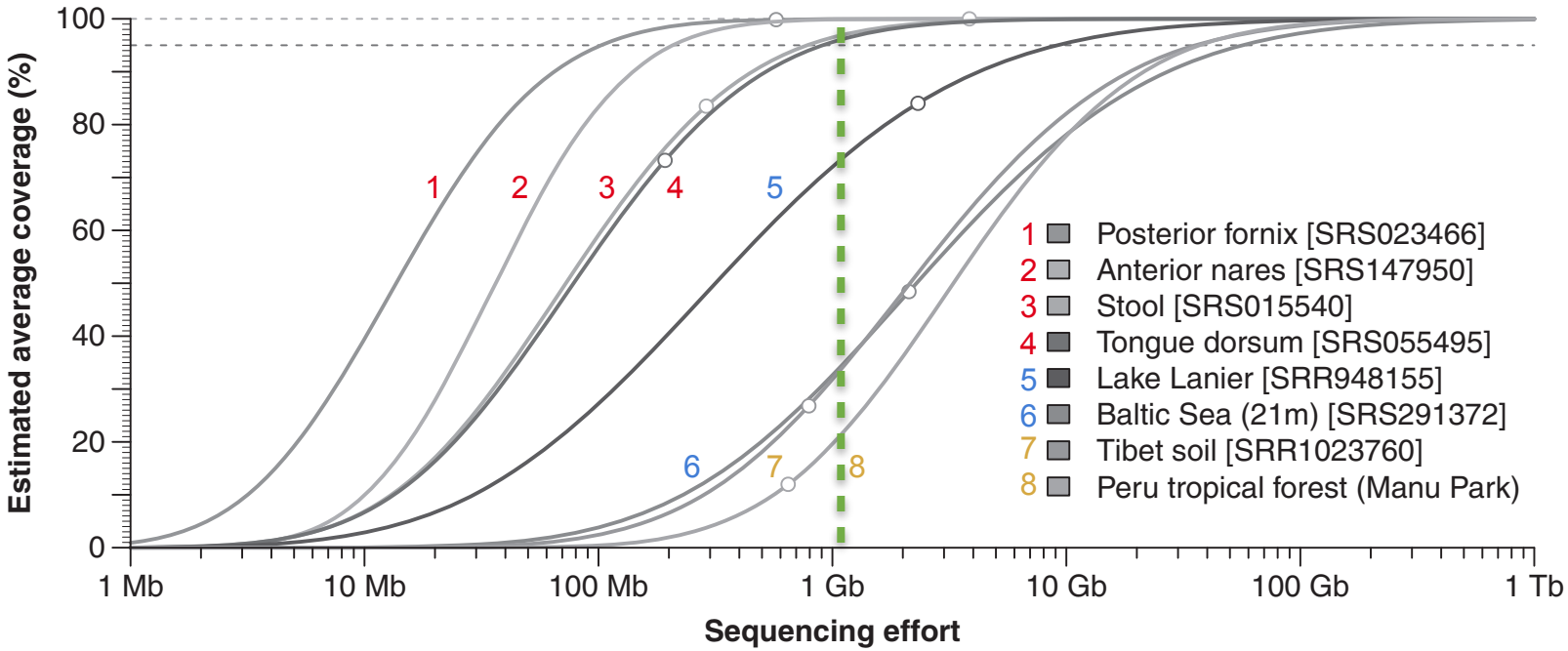
Alpha diversity analysis



Comparison tool

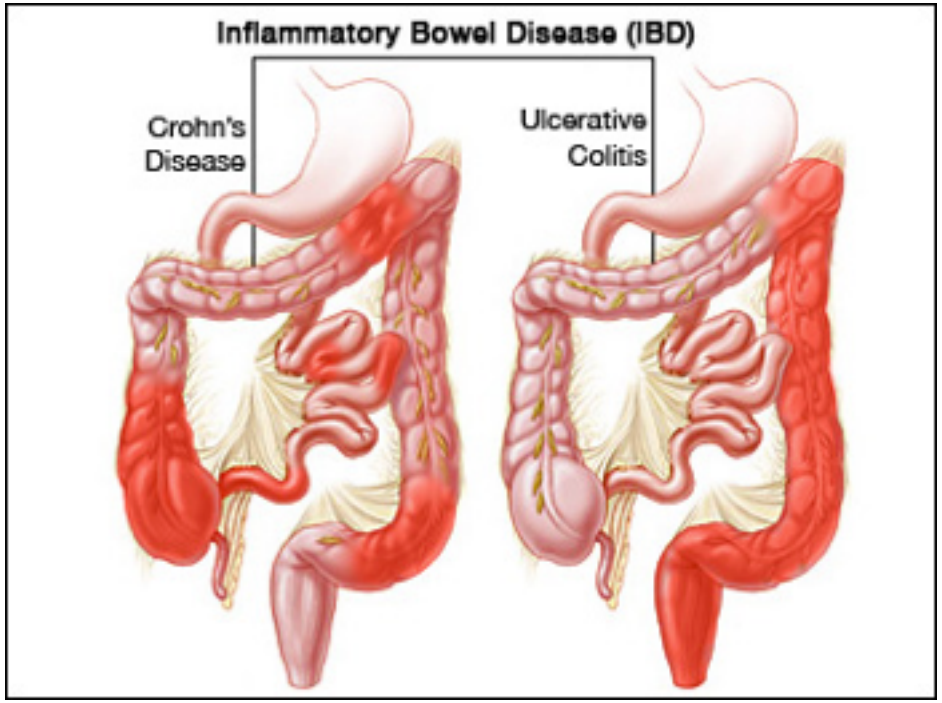


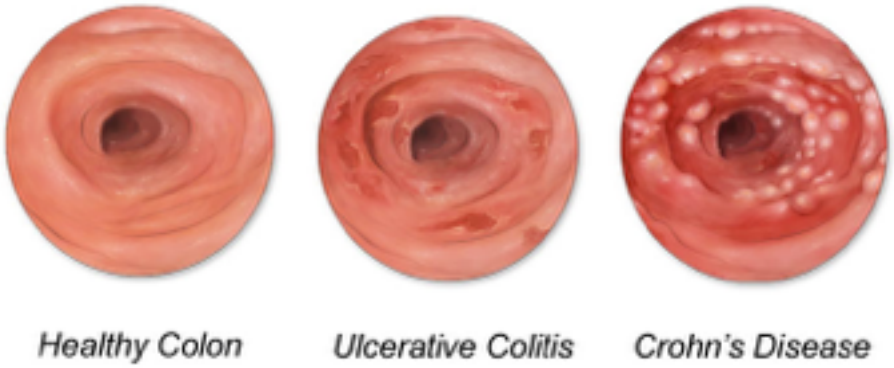
Show much of the microbial community has been sequenced?



Rodriguez-R and Konstantinidis, *ISMEJ*, 2014 1-3

Metagenomics - Human gut microbiome





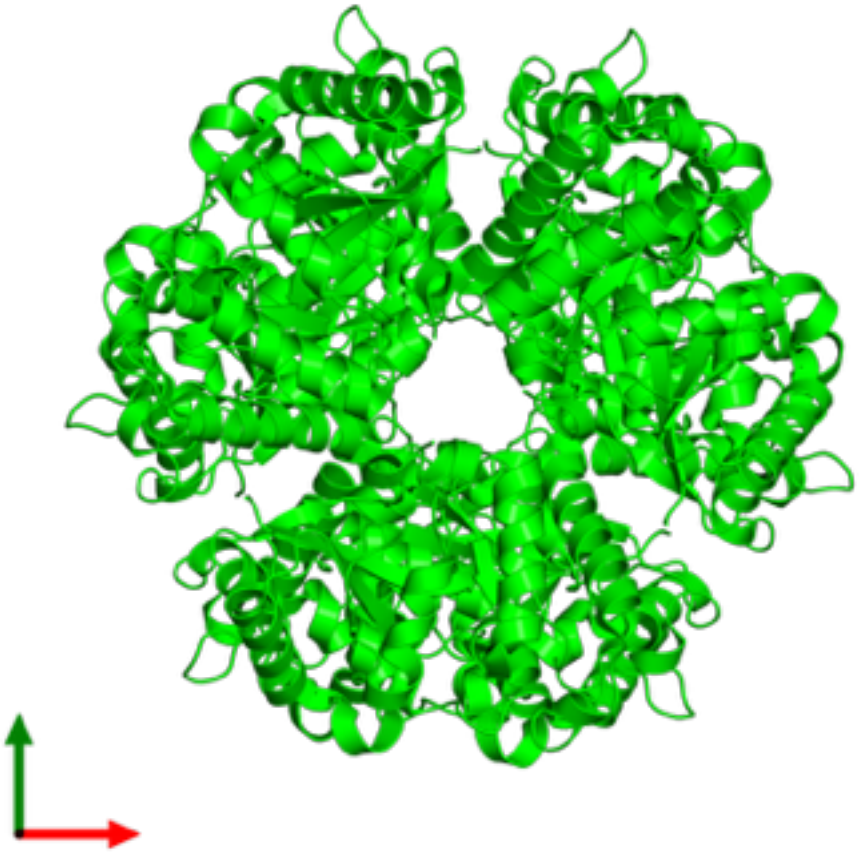
Susceptibility to inflammatory bowel disease (IBD) :

Friedreich ataxia (frataxin; fxn)

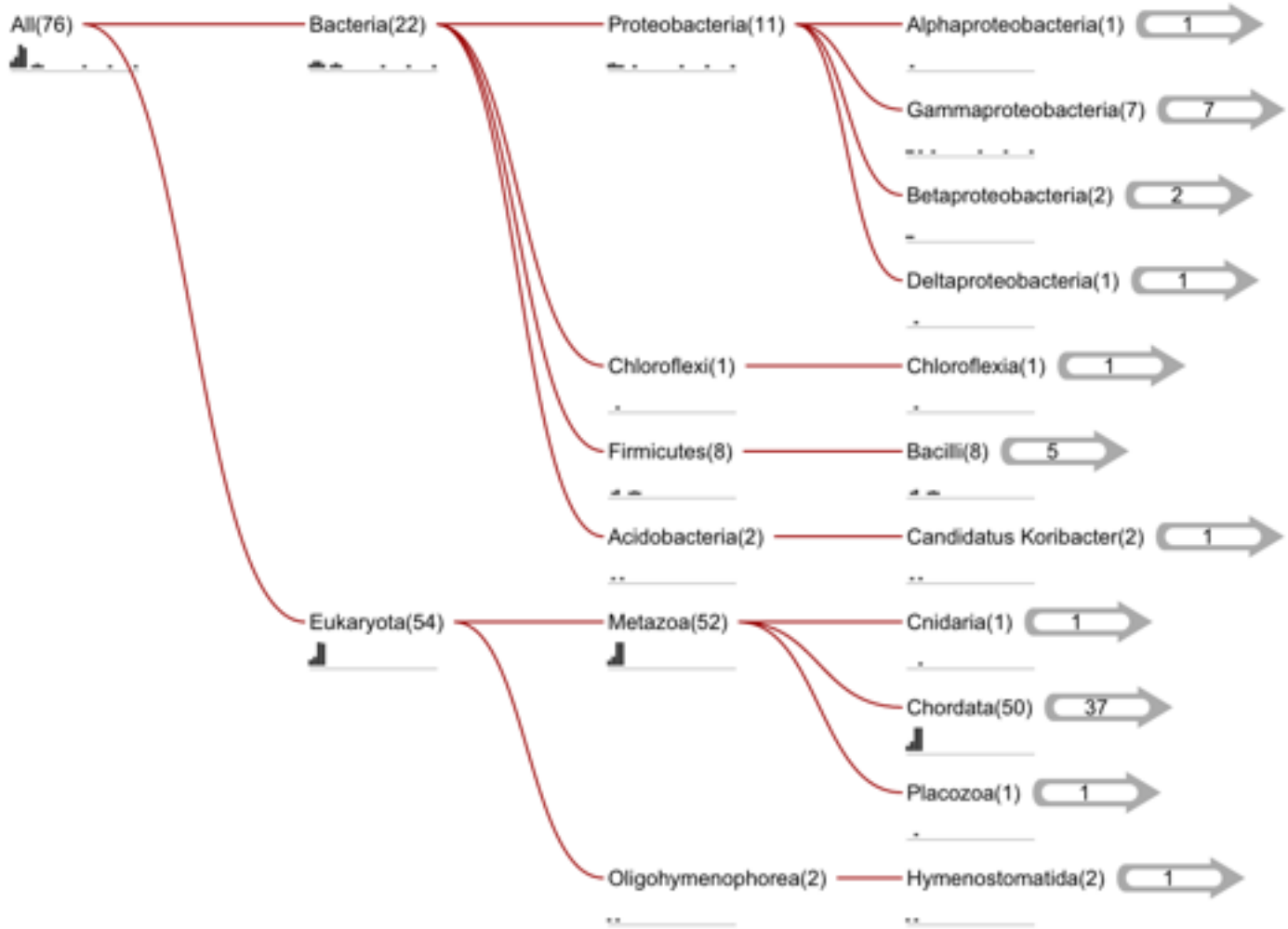
YdjC-like (unpublished data from Lawley Lab)

Homology and evolution

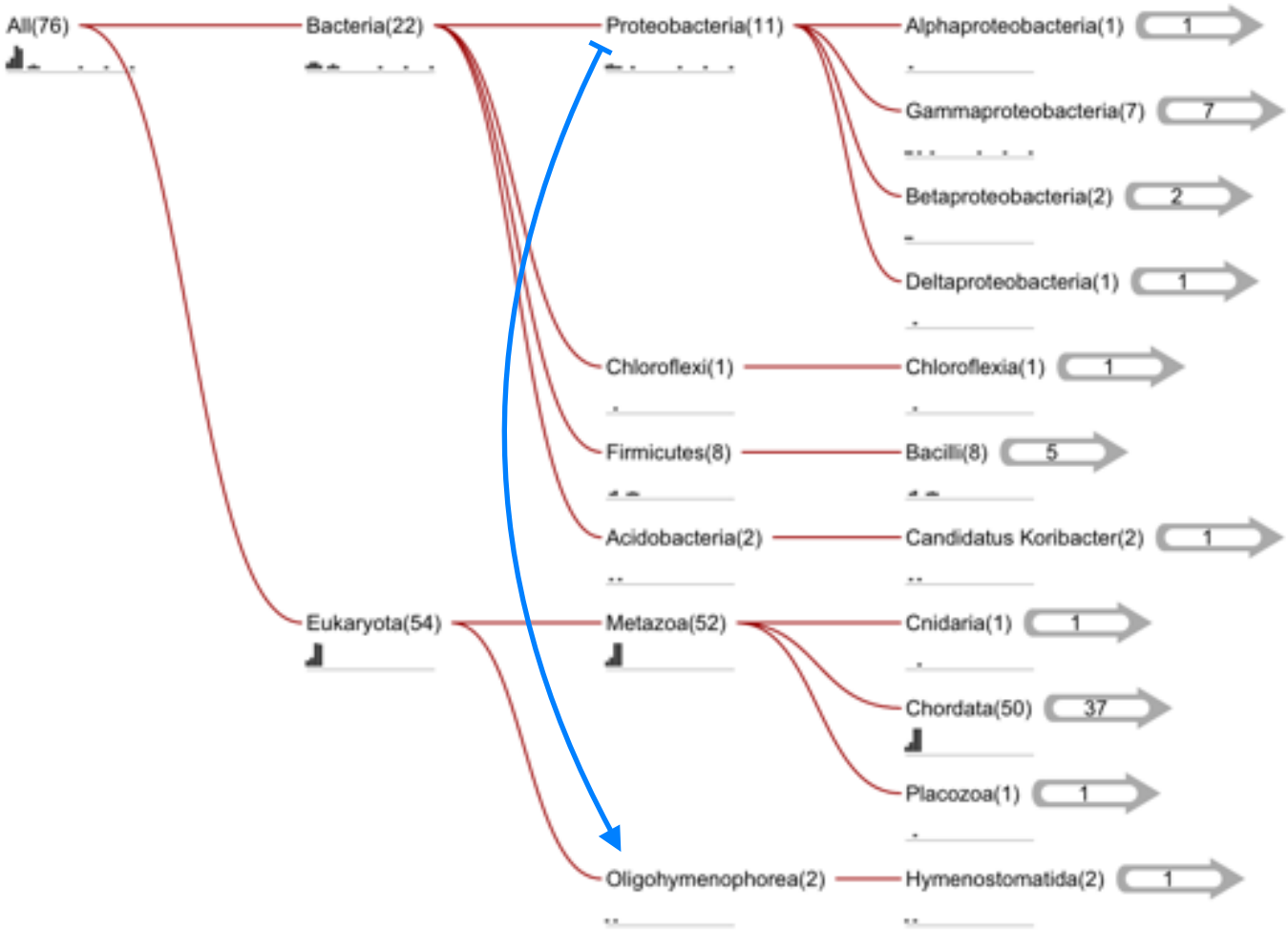
- YdjC gene from *E.coli*
  - Chitooligosaccharide deacetylase



Homology and evolution



Homology and evolution



PANTHER - YdjC

Acknowledgements

EMBL-EBI
Alex Mitchell
Hubert Denise
Matthew Fraser
Gift Nuka
Sebastien Pesseat
Maxim Scheremetjew
Francois Bucchini
Craig McAnulla
Sarah Hunter

Guy Cochrane
Rasko Leinonen
Rajesh Radhakrishnan
Petra Ten Hoopen

OeRC
Dawn Field
Peter Sterk

Susanna Sansone
Eamonn Maguire
Alejandra Gonzalez-Beltran
Philippe Rocca-Serra

External Collaborators
Sean Eddy
Eric Nawrocki

Trevor Lawley

Sterghios Moschos

Nils-Peder Willassen

EBI metagenomics - a new resource for the analysis and archiving of metagenomic data
Hunter *et al, NAR,* 2014 42:D600-D606