200 billion sequences and counting: analysis, discovery and exploration of datasets with EBI Metagenomics

Alex Mitchell

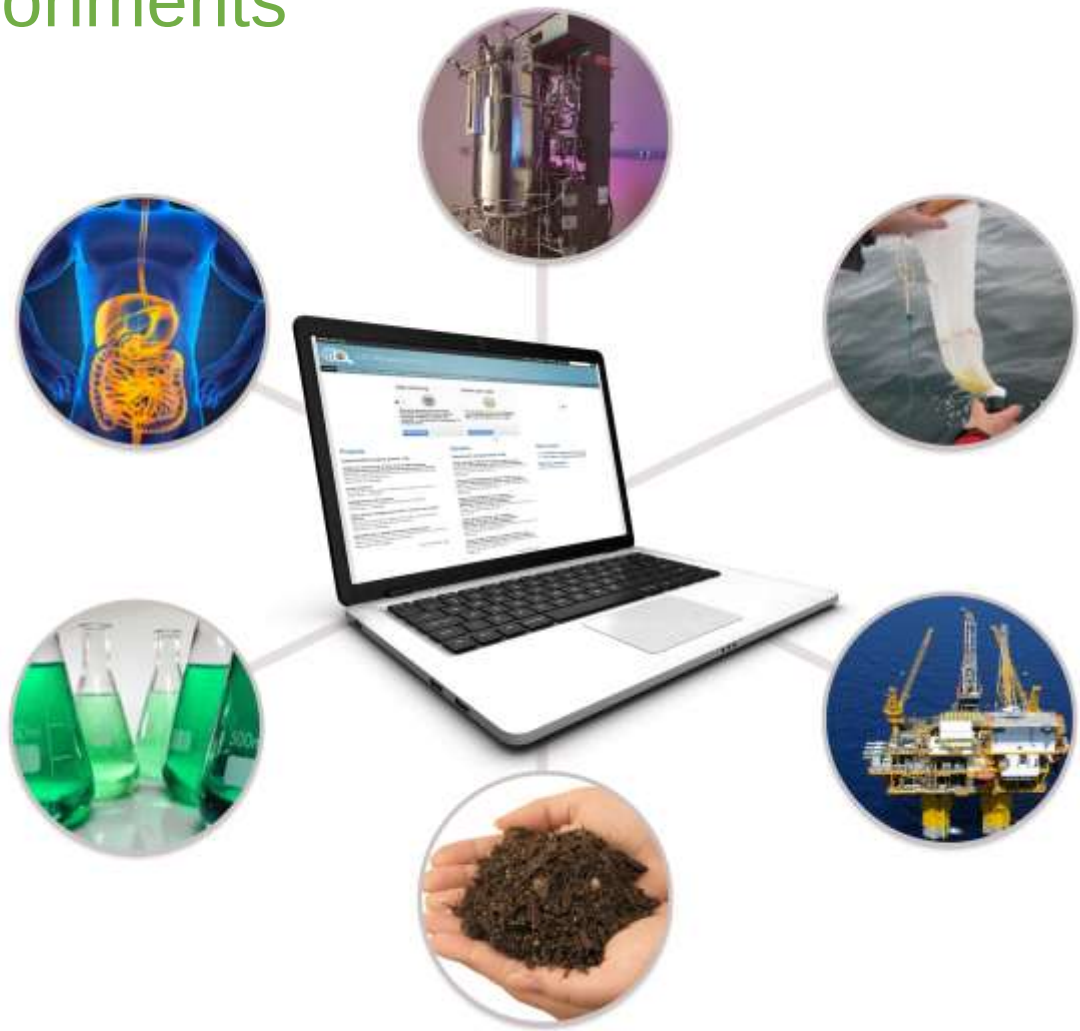
mitchell@ebi.ac.uk

EMBL-EBI

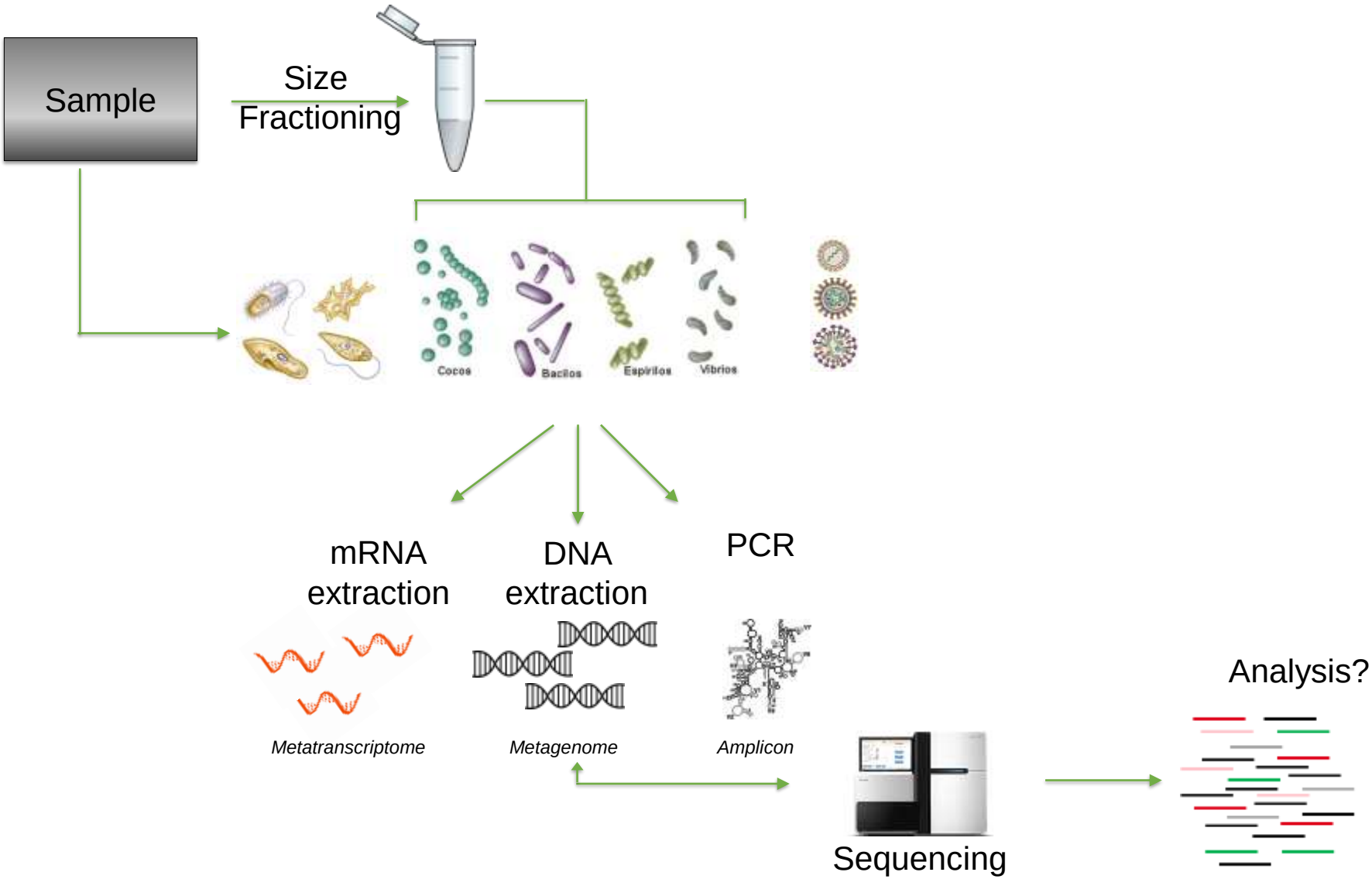
Overview

- Challenges of metagenomic data analysis
- The EBI Metagenomics analysis pipeline & portal
- Recent & forthcoming developments

‘Metagenomics’: a broad range of environments

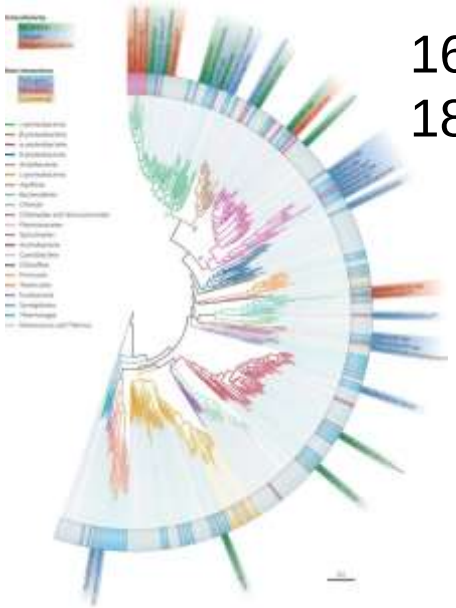


Different experimental design

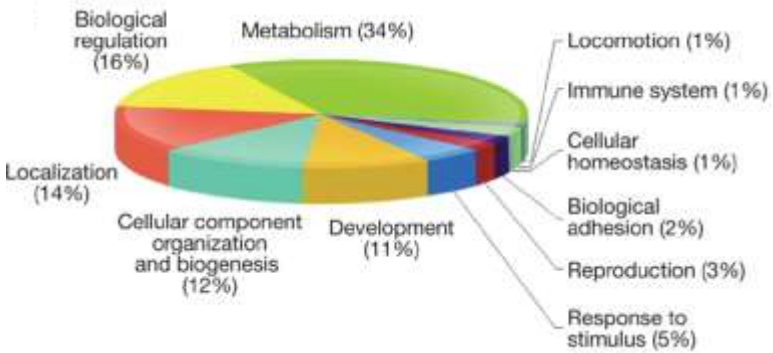


Taxonomic analysis

Analysis



16S rRNA
18S rRNA
ITS
k-mer
etc



Functional analysis

Identification and characterisation of protein coding sequences

Why is metagenomics challenging?

- Vast number of sequences
- Most are missing from the reference databases
- Short sequence fragments are hard to characterise
- Assembly can lead to chimeras
- Iddo Friedberg: ‘Metagenomics is like a disaster in a jigsaw shop’

- Millions of different pieces
- Thousands of different puzzles
- All mixed together
- Most of the pieces are missing
- No box art to refer to

Analysis pitfalls

- Different analysis tools can give different results
- The same tools can give different results, depending on the version and underlying algorithm (*e.g.,* HMMER2 *vs* HMMER3)
- The same version of the same tools can give different results depending on the reference database used

Reference databases

NCBI/ BLAST/ blastp suite

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query.

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

```
>PLANT1
MGERFFRNEMPEFVPEDLSGEEETVTECKDSLTKLLSLPYKSFSEKLHRYALSIKDKVVW
ETWERSGKRVRDYNLYTGVLGTAYLLFKSYQVTRNEDDLKLCLENVEACDVASRDSERVT
FICGYAGVCALGAVAAKCLGDDQLYDRYLARFRGIRLPSDLPYELLYGRAGYLWACLFLN
KHIGQESISSERMRSVVEEIFRAGRQLGNKGTCPLMYEWHGKRYWGAAHGLAGIMNVLMH
```

Or, upload file Choose File no file selected

Job Title PLANT1

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database Non-redundant protein sequences (nr)

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
G-protein coupled receptor 2 [Arabidopsis thaliana]	850	850	100%	0.0	100%	NP_175700.2
Chain A, Crystal Structure Of Arabidopsis Gcr2 [Arabidopsis thaliana]	849	849	100%	0.0	100%	3T33_A
putative G protein-coupled receptor; 80093-78432 [Arabidopsis thaliana]	830	830	97%	0.0	100%	AAG52264.1
predicted protein [Arabidopsis lyrata subsp. lyrata]	778	778	100%	0.0	92%	XP_002894411.1

Reference databases

NCBI/ BLAST/ blastp suite

Standard Protein BLAST

blastn | blastp | blastx | tblastn | tblastx

BLASTP programs search protein databases using a protein query.

Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) Clear

Query subrange

From

To

```
>PLANT1
MGERFFRNEMPEFVPEDLSGEEETVTECKDSLTKLLSLPYKSFSEKLHRYALSIKDKVVW
ETWERSGKRVRDYNLYTGVLGTAYLLFKSYQVTRNEDDLKLCLENVEACDVASRDSERVT
FICGYAGVCALGAVAAKCLGDDQLYDRYLARFRGIRLPSDLPYELLYGRAGYLWACLFLN
KHIGQESISSERMRSVVEEIFRAGRQLGNKGTCPLMYEWHGKRYWGAAHGLAGIMNVLMH
```

Or, upload file Choose File no file selected

Job Title PLANT1

Enter a descriptive title for your BLAST search

Align two or more sequences

Choose Search Set

Database UniProtKB/Swiss-Prot(swissprot)

Sequences producing significant alignments:

Select: All None Selected:0

Alignments Download GenPept Graphics Distance tree of results Multiple alignment

Description	Max score	Total score	Query cover	E value	Ident	Accession
RecName: Full=LanC-like protein GCR2; AltName: Full=G-protein coupled receptor 2 [Arabidopsis thaliana]	850	850	100%	0.0	100%	F4IEM5.1
RecName: Full=LanC-like protein GCL2; AltName: Full=G protein-coupled receptor 2-like protein 2; Short=Protein GCR2-like 2 [Arabidopsis thaliana]	541	541	100%	0.0	64%	Q8VZQ6.1
RecName: Full=LanC-like protein GCL1; AltName: Full=G protein-coupled receptor 2-like protein 1; Short=Protein GCR2-like 1 [Arabidopsis thaliana]	330	330	96%	1e-107	45%	Q9FJN7.1
RecName: Full=LanC-like protein 2; AltName: Full=Testis-specific adriamycin sensitivity protein [Mus musculus]	269	269	100%	9e-84	39%	Q9JJK2.1

Reference databases

Plant Physiology

HOME | ABOUT | SUBMIT | SUBSCRIPTIONS | ADVERTISE | ARCHIVE |

Institution: Wellcome Trust Genome Campus Sign In as Member / In

EXPAND

"Round up the usual suspects" A Comment on Nonexistent Plant GPCRs

Daisuke Urano (urano@email.unc.edu) **and Alan M. Jones**[1] (alan_jones@unc.edu)

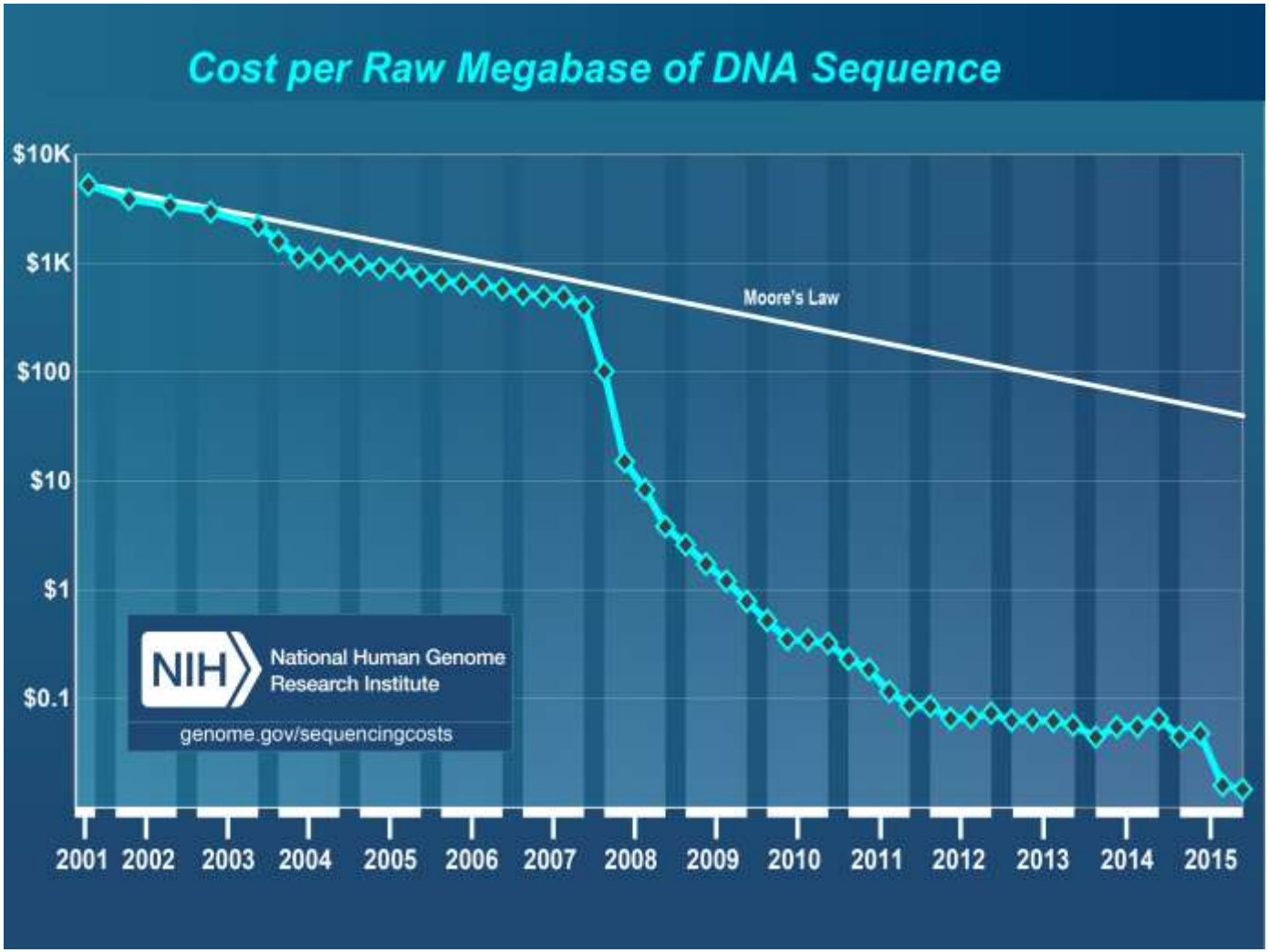
+ Author Affiliations

[1]* Corresponding author; email: alan_jones@unc.edu

Published online before print January 2013, doi: http://dx.doi.org/10.1104/pp.112.212324

Plant Physiology January 2013 pp.112.212324

Data volumes



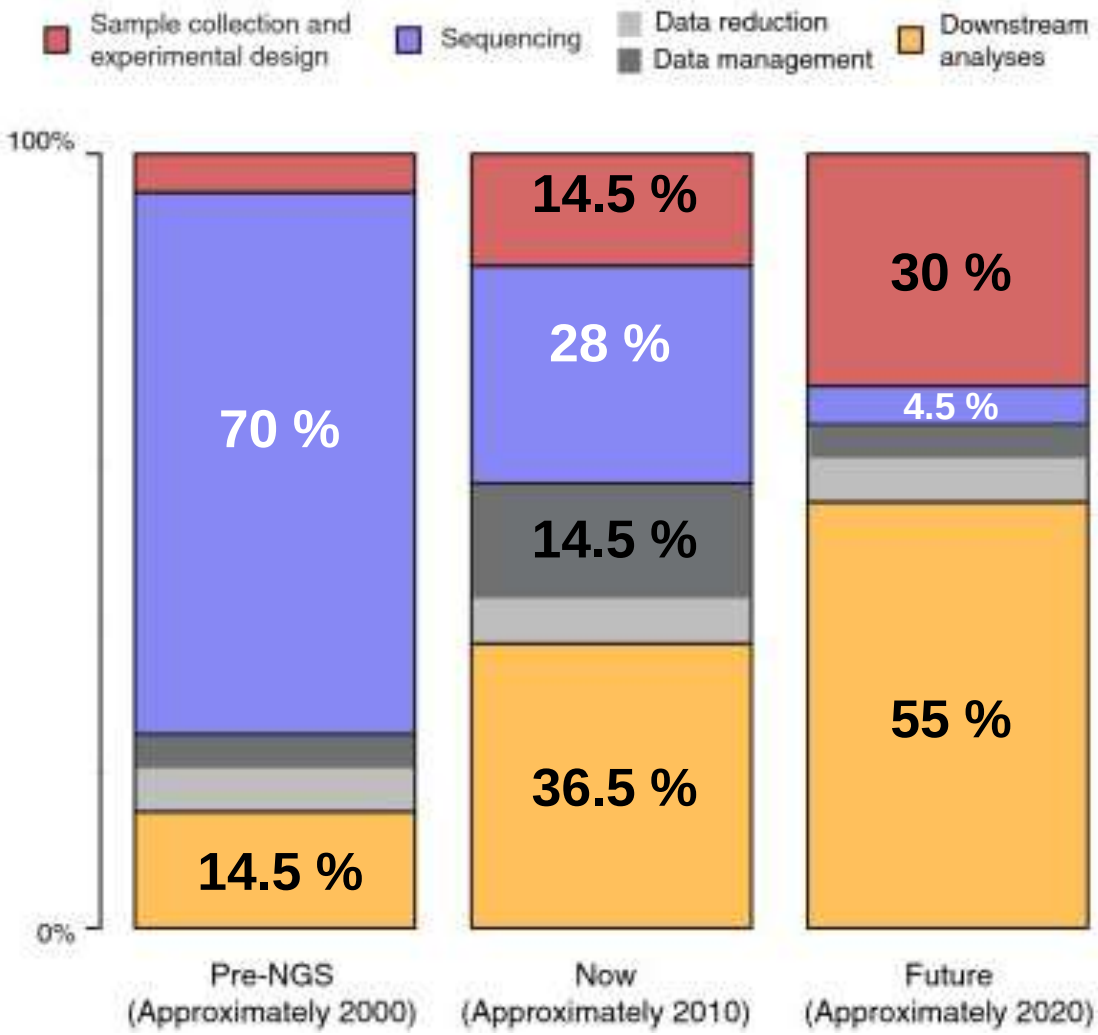
Wetterstrand KA. DNA Sequencing Costs: Data from the NHGRI Large-Scale Genome Sequencing
Available at: www.genome.gov/sequencingcosts.

Data analysis speed

- The cost of sequencing has really gone down
- Now I can do metagenomics!
- ***Awesome!***

- Amount of sequence generated has increased **5,000-fold**
- Computational speed has increased only **10-fold**
- Time taken to analyse has increased **500-fold**
- ***$@%*!!!***

Data analysis cost



Sboner *et al.*
Genome Biology (2011) **12**:125

EMBL-EBI

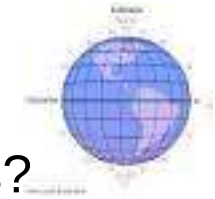
Longevity & usefulness of data

For data to have longevity and be useful to the scientific community, sequences need to be archived with contextual metadata

- How was it sampled? How was it extracted? How was it stored? What sequencing platform was used?



- Where did it come from? What were the environmental conditions (lat/long, depth, pH, salinity, temperature…) or clinical observations?



Contextual data

If contextual data is adequately described, querying and interpretation across projects becomes possible

- Show the microbial species found in the North Pacific
  - … at depths of 50 – 100 m
    - … in samples taken May-June
      - … compared to the Indian Ocean, under the same conditions

Why standardised vocabularies are important:
how many ways to say “female”?

18-day pregnant females
2 yr old female
400 yr. old female
adult female
asexual female
castrate female
cf.female
cystocarpic female
dikaryon
dioecious female
diploid female
f
famale
femail
female
female - worker
female (alate sexual)
female (calf)
hen
female (lactating)
female (pregnant)
female (outbred)
female parent
female plant
female with eggs
female worker
female, 6-8 weeks old
female, virgin
female, worker
female(gynoecious)
femele
female, pooled
femalen
females
females only
gynoecious
healthy female
probably female (based on morphology)
individual female
lgb*cc females
mare
female (worker)
monosex female
ovigerous female
oviparous sexual females
worker bee
female enriched
pseudohermaprhoditic female
remale
semi-engorged female
sexual oviparous female
sterile female worker
strictly female
tetraploid female
thelytoky
female (gynoecious)
worker caste (female)
sex: female
female, other
female child
femal
3 female
female (phenotype)
female mice
female, spayed
femlale
metafemale
sterile female
normal female
sf
vitellogenic replete female
worker
hexaploid female
female (f-o)
female (note: this sample was originally provided as a \"male\" sample to us and therefore labeled this way in the brawand et al. paper and original geo submission; however, detailed data analyses carried out in the meantime clearly show that this sample stems from a female individual)",

Considerations: storing data

Where are you going to store this?

- Locally : back-up ?

  long term ?

  sharing ?

  access ?

- Amazon, Google or specialist research clouds

- **Public repositories, such as ENA, NCBI or DDBJ**

ENA European Nucleotide Archive

NCBI

DDBJ DNA Data Bank of Japan

EBI Metagenomics

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

A ***free*** resource for the analysis, archiving & browsing of amplicon, metagenomic and metatranscriptomic data

Powerful analysis

Data archiving

EBI Metagenomics

Submit data

Tools to help capture & store metadata

Tools to help transfer data

Quality filtering of sequences

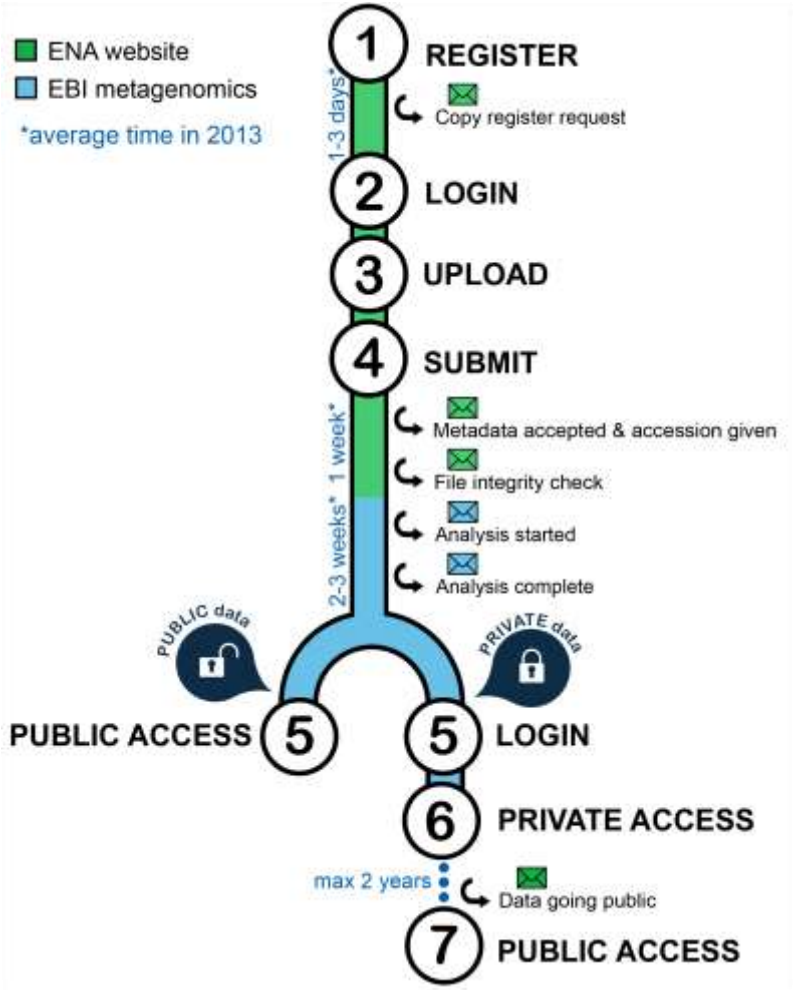
Sequence archiving

ENA
European Nucleotide Archive

Sequence analysis (prebuilt workflows)

Visualisation/Interpretation/Discovery

The data submission process



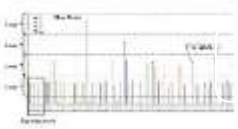
(1) Register for an account

(2) Upload sequence data and metadata

(3) Sequence data is archived in ENA and accessioned

(4) Sequence data is analysed by the metagenomics pipeline

(5) Projects, metadata and results are made available on the website for private or public browsing / download



raw reads

ENA
European Nucleotide Archive

QC

discarded reads

processed reads

rRNAselector

reads with rRNA

reads without rRNA

FragGeneScan

predicted CDS

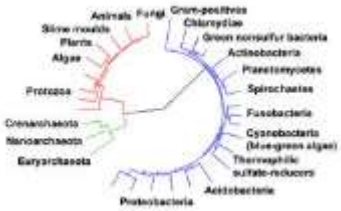
InterProScan

Unknown function pCDS

QIIME & GreenGenes

Taxonomic analysis

Function assignment



Pairwise sequence analysis approaches?

- **BLAST, BLAT, etc?**

- Scalability *(BLAST of 150 million seqs vs UniProt requires ~15 billion pairwise comparisons)*
- Sensitivity (sequences not in reference databases)

BLAST for annotation transfer

- BLAST of 2 proteins:

  60S acidic ribosomal protein P0 from 2 closely-related species

```
8     QKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGKATILMGKNTRIRTALKKNLQAVPQIEK    77
      ++K ++IEK + L   Y K+++   D VGS+Q+  +RKS+RG   +LMGK T IR  ++    + P+++
7     KRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGIGAVLMGKKTMIRKVIRDLADSKPELDA    76

78    LLPLVKLNMGFVFCKDDLSEIRNIILDNKSSSHPARLGVIAPIDVFIPPGPTGMDPSHTSFLESLGISTK   147
      L   +K N   +FCKD+++E++ +I   +  + PA+ GV AP DV IP GPTGM+P+ TSFL+ L I+TK
77    LNTYLKQNTCIIFCKDNIAEVKRVINTQRVGA-PAKAGVFAPNDVIIPAGPTGMEPTQTSFLQDLKIATK   145

148   IVKGQIEIQEHVHLIKQGEKVTASSATLLRKFNMNP-SYGVDVRTVYDDGVIYDAKVLDITDEDILEKFS   216
      I +GQI+I   VH+IK G+KV AS ATLL+K N+ P +YG++ + +YD G  Y      I++ED++ KF
146   INRGQIDIVNEVHIIKTGQKVGASEATLLQKLNIKPFTYGLEPKIIYDAGACYSP---SISEEDLINKFK   213

217   KGVSNVAALSRATGVITEASYPHVFVEAFKNIVALIIDSDYTFPLMKILKKWVENPEAFAAVAAPASAA-   286
      +G+ N+AA+S   G  T AS PH  + AFKN++A+  ++ YTF   +  K        AA AAP +AA
213   QGIFNIAAISLEIGYPTVASIPHSVMNAFKNLLAISFETSYTFDAAEKFKS--------AAAAAPVAAAP   278

286   KADEPKKEEAKKVEEEEEEEDGFMGFGMFD   318 Q94660
       A  P     K V EE++EE D  MG G+FD
275   SAAAPAAAAKKVVVEEKKEESDDDMGMGLFD   317 P22685
```

60S acidic ribosomal protein P0: multiple sequence alignment

```
Q5E940_BOVIN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_HUMAN -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_MOUSE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
    RLA0_RAT -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_CHICK -----------MPREDRATWKSNYFMKIIQLLDDYPKCFVVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_RANSY -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQQIRMSLRGK-AVVLMGKNTMMRKAIRGHLENN--SALE
Q7ZUG3_BRARE -----------MPREDRATWKSNYFLKIIQLLDDYPKCFIVGADNVGSKQMQTIRLSLRGK-AVVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_ICTPU -----------MPREDRATWKSNYFLKIIQLLNDYPKCFIVGADNVGSKQMQTIRLSLRGK-AIVLMGKNTMMRKAIRGHLENN--PALE
  RLA0_DROME -----------MVRENKAAWKAQYFIKVVELFDEFPKCFIVGADNVGSKQMQNIRTSLRGL-AVVLMGKNTMMRKAIRGHLENN--PQLE
  RLA0_DICDI -----------MSGAG-SKRKKLFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
Q54LP0_DICDI -----------MSGAG-SKRKNVFIEKATKLFTTYDKMIVAEADFVGSSQLQKIRKSIRGI-GAVLMGKKTMIRKVIRDLADSK--PELD
  RLA0_PLAF8 -----------MAKLSKQQKKQMYIEKLSSLIQQYSKILIVHVDNVGSNQMASVRKSLRGK-ATILMGKNTRIRTALKKNLQAV--PQIE
  RLA0_SULAC -----MIGLAVTTTKKIAKWKVDEVAELTEKLKTHKTIIIANIEGFPADKLHEIRKKLRGK-ADIKVTKNNLFNIALKNAG-----YDTK
  RLA0_SULTO ----MRIMAVITQERKIAKWKIEEVKELEQKLREYHTIIIANIEGFPADKLHDIRKKMRGM-AEIKVTKNTLFGIAAKNAG-----LDVS
  RLA0_SULSO ----MKRLALALKQRKVASWKLEEVKELTELIKNSNTILIGNLEGFPADKLHEIRKKLRGK-ATIKVTKNTLFKIAAKNAG-----IDIE
  RLA0_AERPE MSVVSLVGQMYKREKPIPEWKTLMLRELEELFSKHRVVLFADLTGTPTFVVQRVRKKLWKK-YPMMVAKKRIILRAMKAAGLE---LDDN
  RLA0_PYRAE -MMLAIGKRRYVRTRQYPARKVKIVSEATELLQKYPYVFLFDLHGLSSRILHEYRYRLRRY-GVIKIIKPTLFKIAFTKVYGG---IPAE
  RLA0_METAC ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVGIEGILATKMQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ETIP
  RLA0_METMA ------MAEERHHTEHIPQWKKDEIENIKELIQSHKVFGMVRIEGILATKIQKIRRDLKDV-AVLKVSRNTLTERALNQLG-----ESIP
  RLA0_ARCFU ------MAAVRGS---PPEYKVRAVEEIKRMISSKPVVAIVSFRNVPAGQMQKIRREFRGK-AEIKVVKNTLLERALDALG-----GDYL
  RLA0_METKA MAVKAKGQPPSGYEPKVAEWKRREVKELKELMDEYENVGLVDLEGIPAPQLQEIRAKLRERDTIIRMSRNTLMRIALEEKLDER--PELE
  RLA0_METTH -------------MAHVAEWKKKEVQELHDLIKGYEVVGIANLADIPARQLQKMRQTLRDS-ALIRMSKKTLISLALEKAGREL--ENVD
  RLA0_METTL -------MITAESEHKIAPWKIEEVNKLKELLKNGQIVALVDMMEVPARQLQEIRDKIR-GTMTLKMSRNTLIERAIKEVAEETGNPEFA
  RLA0_METVA -------MIDAKSEHKIAPWKIEEVNALKELLKSANVIALIDMMEVPAVQLQEIRDKIR-DQMTLKMSRNTLIKRAVEEVAEETGNPEFA
  RLA0_METJA --------METKVKAHVAPWKIEEVKTLKGLIKSKPVVAIVDMMDVPAPQLQEIRDKIR-DKVKLRMSRNTLIIRALKEAAEELNNPKLA
  RLA0_PYRAB -------------MAHVAEWKKKEVEELANLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAQELGKPELE
  RLA0_PYRHO -------------MAHVAEWKKKEVEELAKLIKSYPVIALVDVSSMPAYPLSQMRRLIRENGGLLRVSRNTLIELAIKKAAKELGKPELE
  RLA0_PYRFU -------------MAHVAEWKKKEVEELANLIKSYPVVALVDVSSMPAYPLSQMRRLIRENNGLLRVSRNTLIELAIKKVAQELGKPELE
  RLA0_PYRKO -------------MAHVAEWKKKEVEELANIIKSYPVIALVDVAGVPAYPLSKMRDKLR-GKALLRVSRNTLIELAIKRAAQELGQPELE
  RLA0_HALMA -----MSAESERKTETIPEWKQEEVDAIVEMIESYESVGVVNIAGIPSRQLQDMRRDLHGT-AELRVSRNTLLERALDDVD-----DGLE
  RLA0_HALVO -----MSESEVRQTEVIPQWKREEVDELVDFIESYESVGVVGVAGIPSRQLQSMRRELHGS-AAVRMSRNTLVNRALDEVN-----DGFE
  RLA0_HALSA -----MSAEEQRTTEEVPEWKRQEVAELVDLLETYDSVGVVNVTGIPSKQLQDMRRGLHGQ-AALRMSRNTLLVRALEEAG-----DGLD
  RLA0_THEAC -------------MKEVSQQKKELVNEITQRIKASRSVAIVDTAGIRTRQIQDIRGKNRGK-INLKVIKKTLLFKALENLGD----EKLS
  RLA0_THEVO -------------MRKINPKKKEIVSELAQDITKSKAVAIVDIKGVRTRQMQDIRAKNRDK-VKIKVVKKTLLFKALDSIND----EKLT
  RLA0_PICTO -------------MTEPAQWKIDFVKNLENEINSRKVAAIVSIKGLRNNEFQKIRNSIRDK-ARIKVSRARLLRLAIENTGK----NNIV
```

EMBL-EBI

Protein signatures

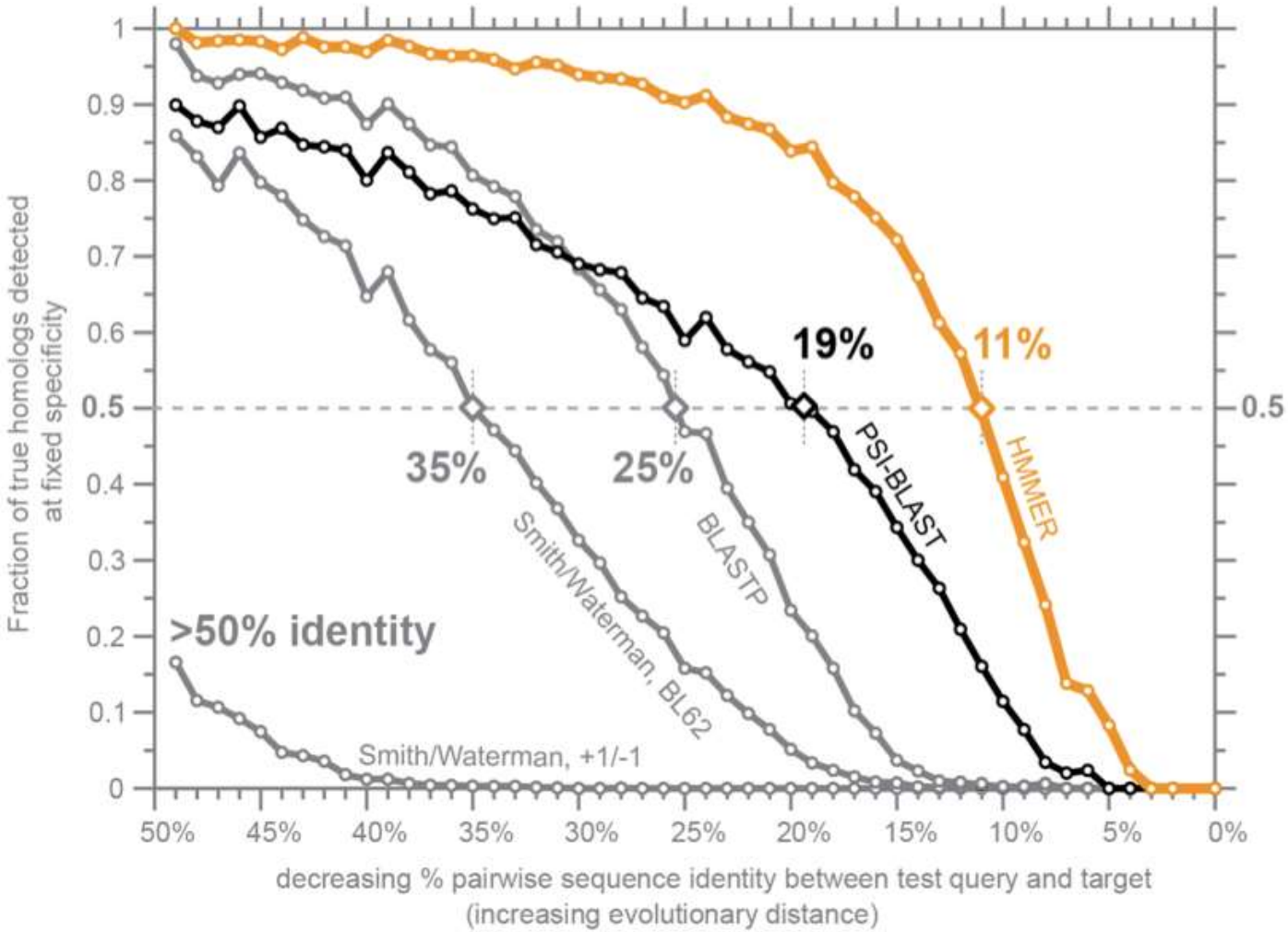
Alternatively, model the pattern of conserved amino acids at specific positions within a multiple sequence alignment

- Patterns
- Profiles
- Profile HMMs

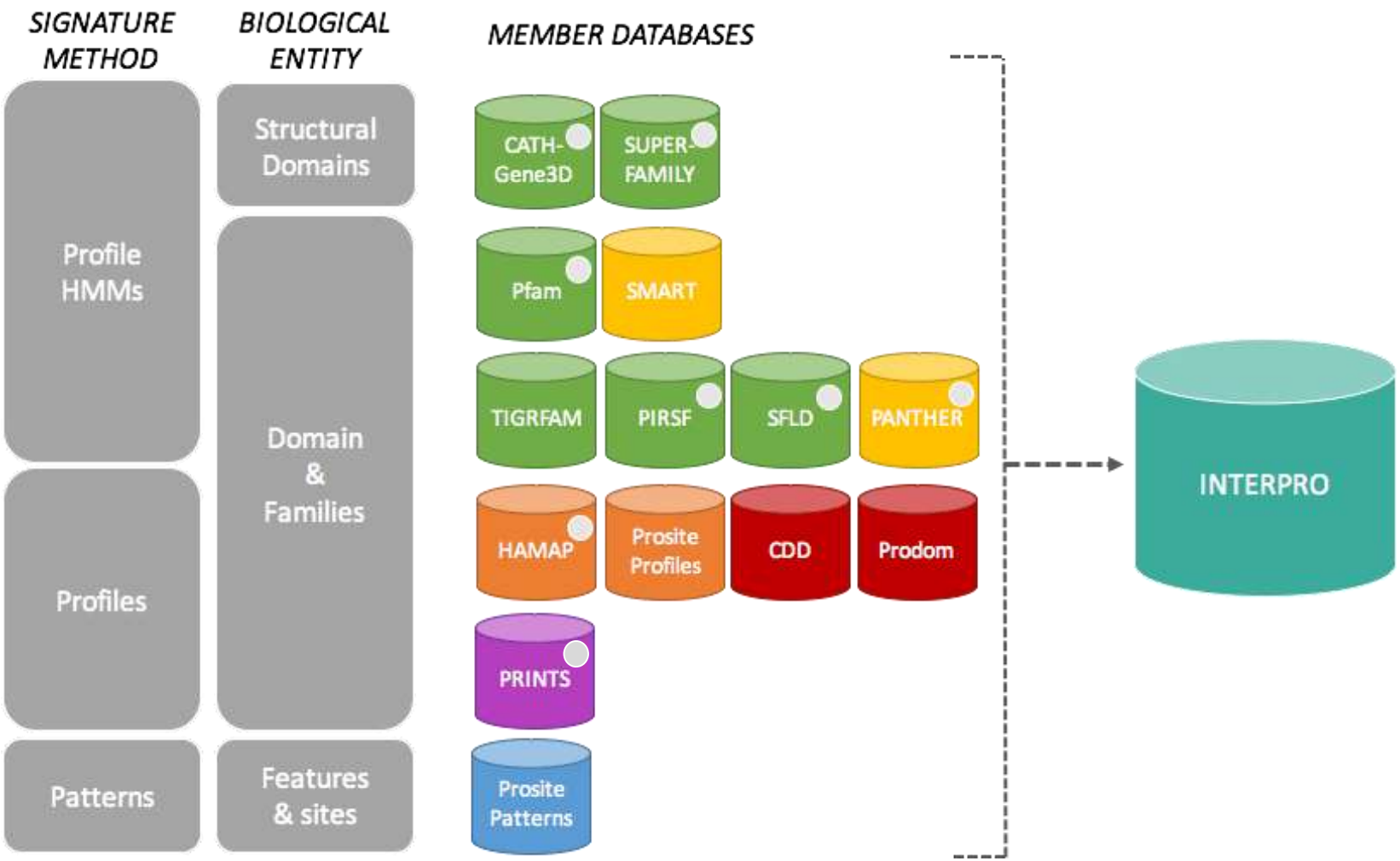
Use these models to infer relationships with the characterised sequences from which the alignment was constructed

Approach used by a variety of databases: Pfam, TIGRFAMs, PANTHER, Prosite, etc

Homology search sensitivity



InterPro - integrated classification of protein families

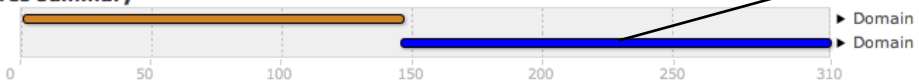


InterPro for functional annotation

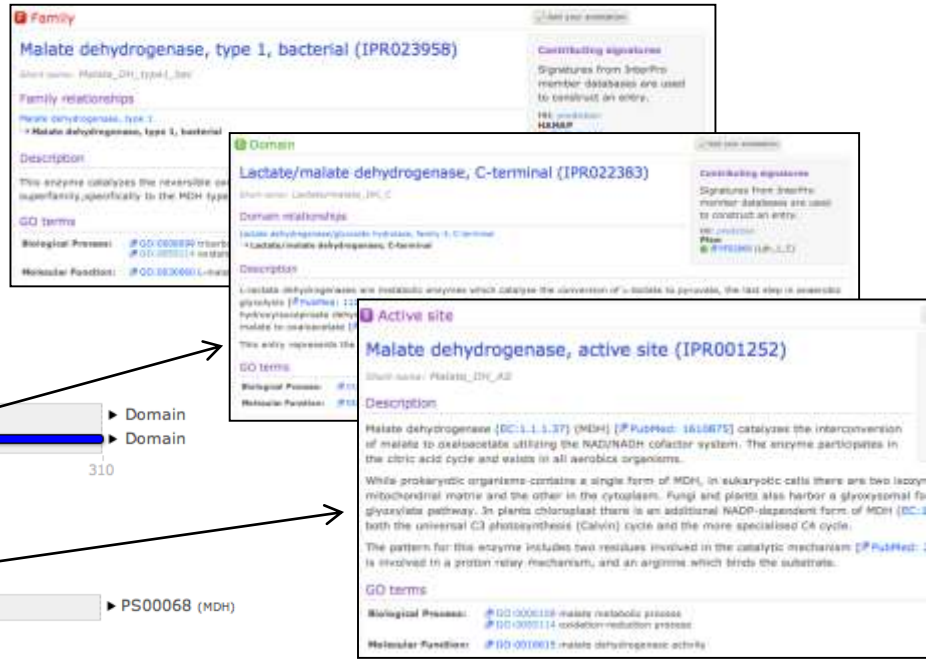
```
>Seq1
MNGTEGPNFYVPFSNKTGVVRSPFEAPQYYLAEP
WQFSMLAAYMFLLIVLGFPINFLTLYVTVQHKKLR
TPLNYILLNLAVADLFMVFGGFTTTLYTSLHGYFVF
GPTGCNLEGFFATLGGEIALWSLVVLAIERYVVVC
KPMSNFRFGENHAIMGVAFTW
```

- **Protein family membership:**
  - L-lactate/malate dehydrogenase (IPR001557)
    - Malate dehydrogenase, type 1 (IPR010097)
      - Malate dehydrogenase, type 1, bacterial (IPR023958)

- **Sequence features summary**



- **Sequence features**



Using InterPro for annotation

- Underlies the system that adds annotation to UniProtKB/TrEMBL
- Provides matches to ~50 million proteins (approx 80% of UniProtKB)
- Source of ~120 million GO terms for ~40 million distinct UniProtKB sequences

Annotation consistency:

- Using InterPro and GO for annotation allows *direct comparison* proteins in UniProtKB

Annotation is kept up to date

- Protein matches to database entires are checked every release
- Entries are updated, errors are fixed in response to:
    - Underlying databases changes
    - Sequence data changes
    - Gene Ontology changes
    - Biological knowledge changes

GO annotation in InterPro: why stability does not indicate accuracy in a sea of changing annotations
Sangrador-Vegas et al., Database (2016)
doi: 10.1093/database/baw027

EBI Metagenomics

Home | Submit data | Projects | Samples | Comparison tool beta | About EBI Metagenomics | Contact | Not logged in | Login

Submit, analyse, visualize and compare your data.

SUBMIT DATA

7438 data sets

4037 metagenomics
389 metatranscriptomics
2945 amplicons
67 assemblies

Public: 5451 runs, 4749 samples, 132 projects

Private: 1987 runs, 1926 samples, 82 projects

Browse projects

By selected biomes

Soil (16) | Marine (28) | Forest (4) | Non-human host (19) | Engineered (11)

Freshwater (9) | Grassland (5) | Human gut (23) | Air (1) | Wastewater (2)

View all biomes

Latest projects 132

Gut microbiota community and functions of mouse exposed to single As, single Fe, and combined As and Fe
Mice were fed with pure water, 3mg/L As, 5mg/L Fe, and 3mg/L As + 5mg/L Fe for 90 days. Then, the fecal samples were collected, and genomic DNA was extracted by FastDNA SPIN Kit for Soil ...
View more - 4 samples

Influence of soil properties on Archaeal diversity and distribution in the McMurdo Dry Valleys, Antarctica
Archaea are the least studied members of the microbial community in Antarctic soils. Their occurrence in coastal mineral soils has been documented, however, less is known about their ...
View more - 4 samples

View all projects

EBI Metagenomics

Home | Submit data | Projects | Samples | Comparison tool | About EBI Metagenomics | Contact | Not logged in | Login

EBI Metagenomics > Projects

Projects list

Text:

Biomes: All

Search | Clear

1 - 1 of 1

Download detailed info (CSV) Download table (CSV)

Biome	Project name	Samples	Last updated
	Metagenomes of hydrogen producing, xylan fed termites	3	28-Aug-2015

EBI Metagenomics

Home | Submit data | Projects | Samples | Comparison tool | About EBI Metagenomics | Contact | Not logged in | Login

EBI Metagenomics > Project: Metagenomes of hydrogen producing, xylan fed ter...

Project (ERP009615)

Metagenomes of hydrogen producing, xylan fed termites

Overview | Analysis summary

Last updated: 28-Aug-2015

Description

Metagenome study with samples from the guts of termites (Nasutitermes exitiosus sp). The termites had xylan as their only carbon source and were raised at temperatures of 33, 40 and 45 degrees respectively while producing highly concentrated hydrogen.

Classification: Host-associated > Arthropoda > Digestive system > Gut

Contact details

Institute: CSIRO
Name: Data Integration BRAEMBL
Email: Datasubs@ebi.org.au

Associated runs

Sample Name	Sample ID	Run ID	Experiment type	Version	Analysis results
Nasutitermes exitiosus gut metagenome incubated at 33	ERS662122	ERR762516	Metagenomic	2.0	Taxonomy I Function I
Nasutitermes exitiosus gut metagenome incubated at 40	ERS662123	ERR762517	Metagenomic	2.0	Taxonomy I Function I
Nasutitermes exitiosus gut metagenome incubated at 45	ERS662124	ERR762518	Metagenomic	2.0	Taxonomy I Function I

EBI Metagenomics

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

EBI Metagenomics > Project: ERP009615 > Sample: ERS662122 > Run: ERR762516

Run (ERR762516)

Overview | Quality control | Taxonomic analysis | Functional analysis | Download

Description

Nasutitermes exitiosus incubated at 33 degrees Celsius

Classification: Host-associated > Arthropoda > Digestive system > Gut

Sample name: Nasutitermes exitiosus gut metagenome incubated at 33 (ERS662

Project name: Metagenomes of hydrogen producing, xylan fed termites (ERP009

Data analysis

Experiment type: Metagenomic

Pipeline version: 2.0

Analysis date: 15/04/2015

Host associated

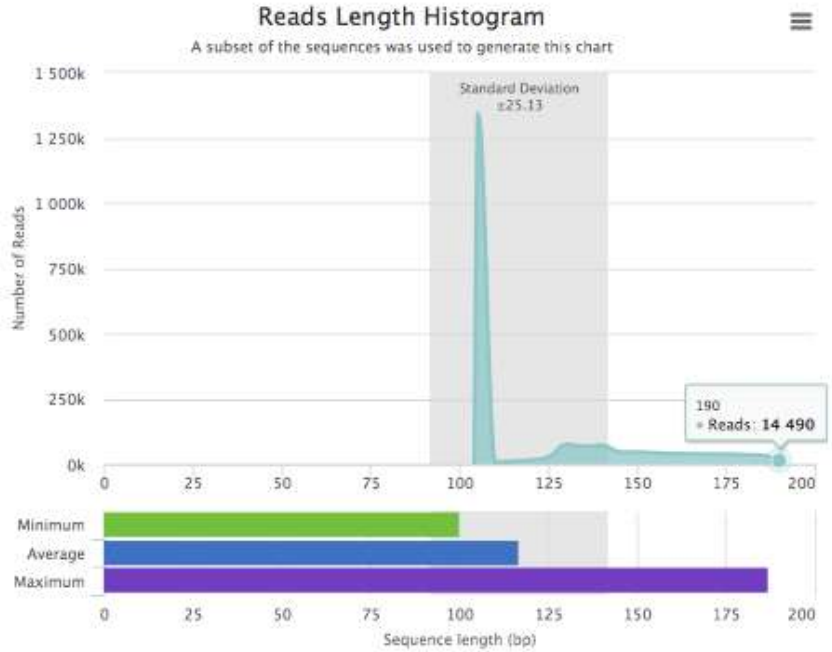
Species: *Termite*

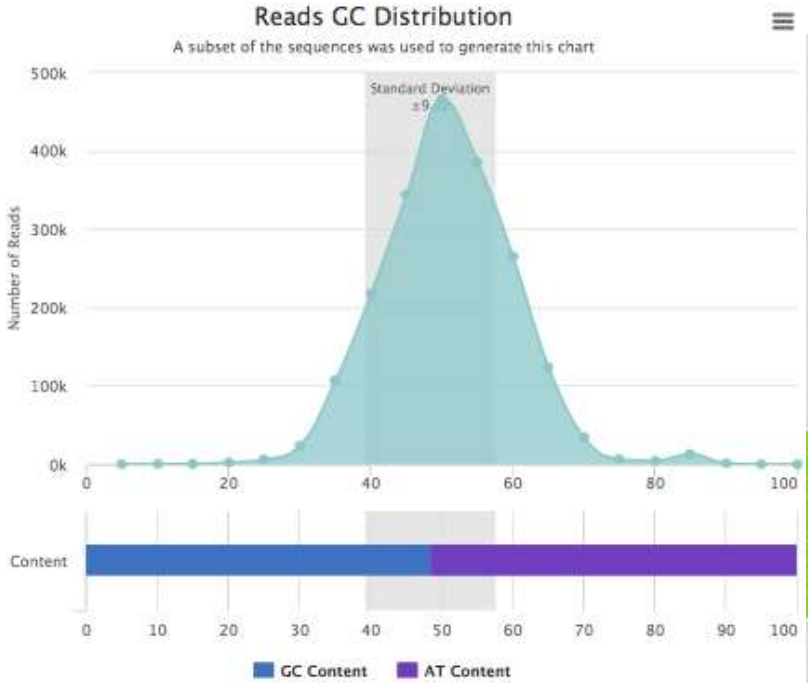
Localisation

Geographic location: Baulkham Hills, NSW, Australia

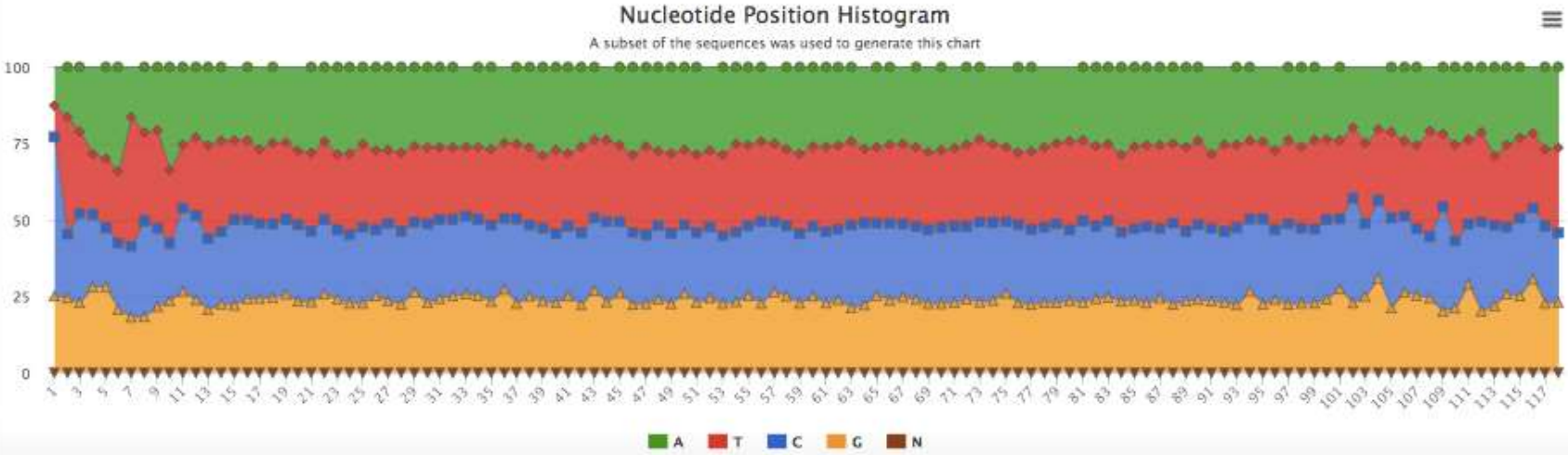
Other information

Investigation type	Metagenomic
Geographic location (country and/or sea,region)	Australia
Collection date	2013-06-02
Sequencing method	Illumina
NCBI sample classification	433724
Instrument model	Illumina HiSeq 2000
ENA checklist	ENA default sample checklist (ERC000011)
Host	Termite
Isolation source	Gut of termite Nasutitermes exitiosus sp. soldiers





The graph below show the relative abundance of nucletotides (A, C, G, T, or ambiguous base "N") at each position starting from the beginning of each read up to the first 500 base pairs.



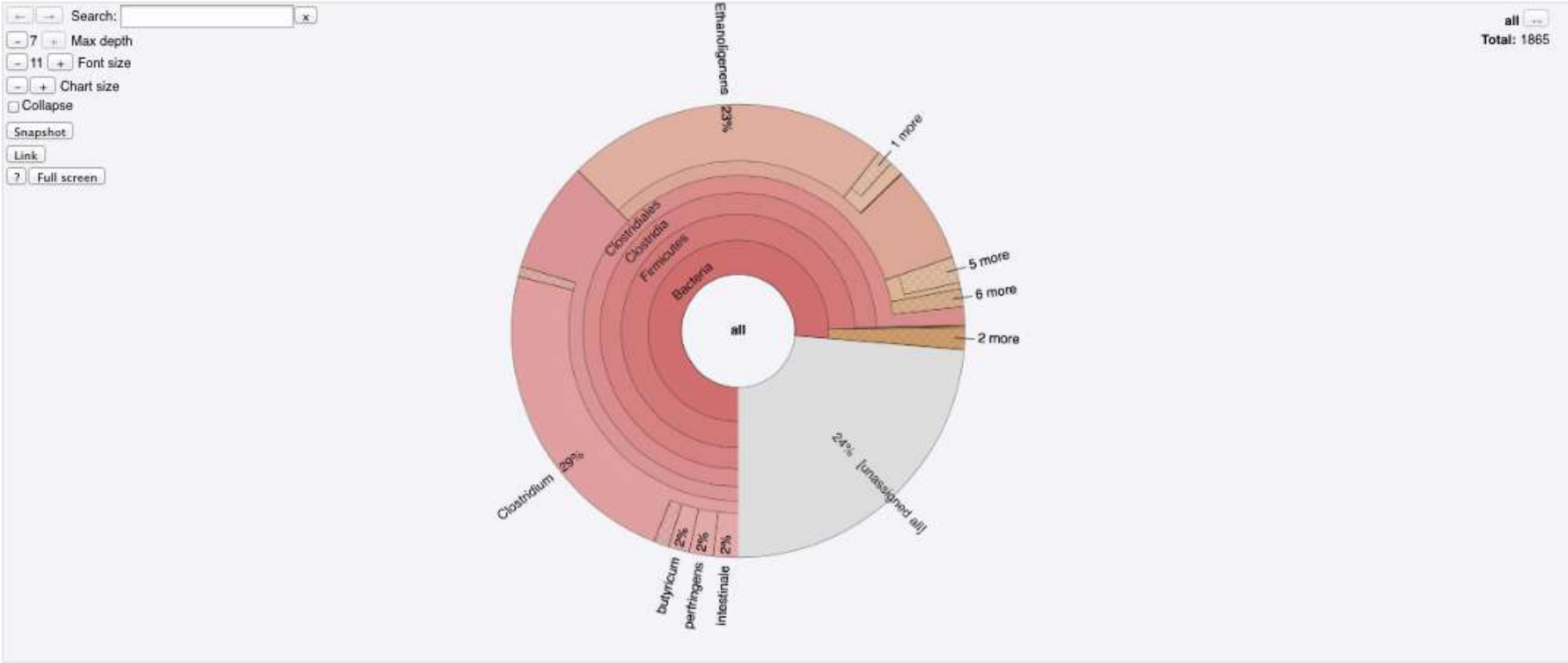
Run (ERR762516)

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

These are the results from the taxonomic analysis steps of our pipeline. You can switch between different views of the data using the menu of icons below (pie, bar, stacked and interactive krona charts). If you wish to download the full set of results, all files are listed under the "Download" tab.

Top taxonomy Hits

Switch view:



Run (ERR762516)

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

Functional analysis has 3 main outputs: a sequence features summary, showing the number of reads with predicted coding sequences (pCDS), the number of pCDS with InterPro matches, and so on; the matches of pCDS to the InterPro database and a chart of the GO terms that summarise the functional content of the sample's sequences. If you wish to download the full set of results, all files are listed under the "Download" tab.

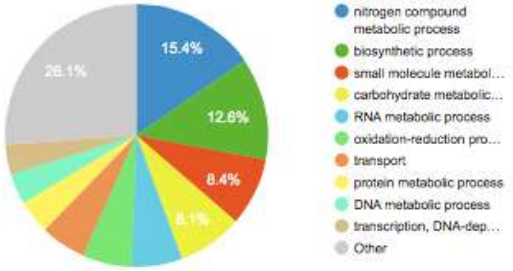
GO Terms annotation

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

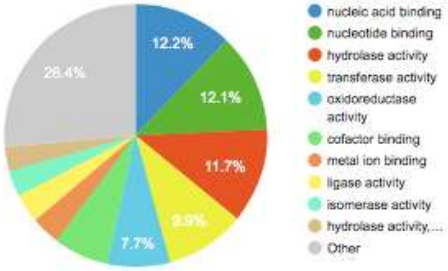
Switch view:

Export

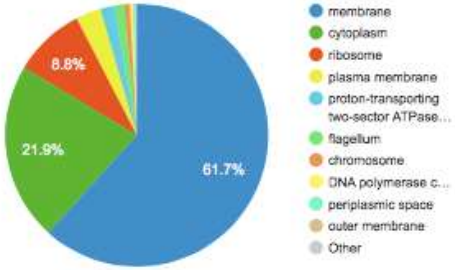
Biological process



Molecular function



Cellular component



Functional annotation: The Gene Ontology

GENEONTOLOGY
Unifying Biology

- Grew out of the model organism community
- Aims to unify the representation of gene and gene product attributes across species
- Allows cross-species and/or cross-database comparison

Inconsistency in naming of biological concepts

English is not a very precise language*

- Same name for different concepts
- Different names for the same concept

An example …

Tactition

Taction

Tactile sense

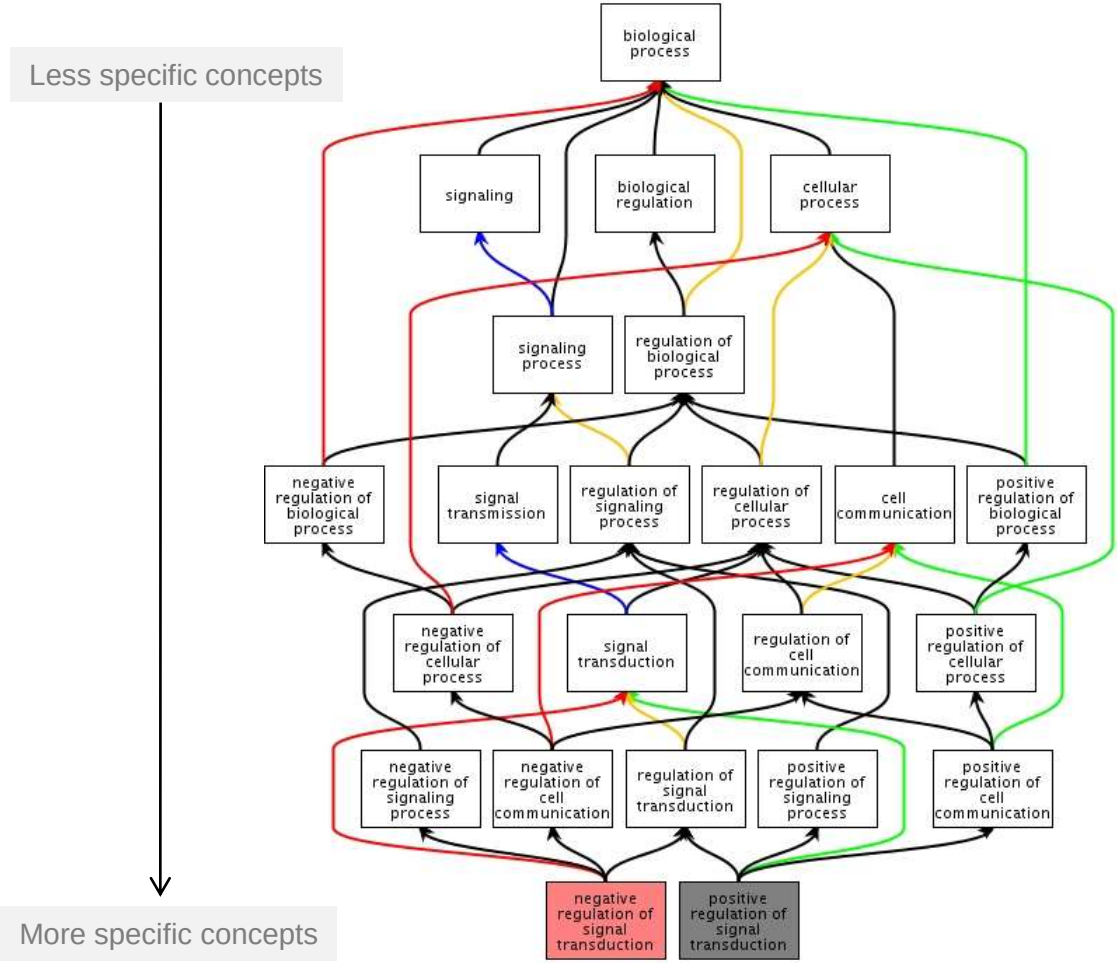
Sensory perception of touch ; GO:0050975

* this is a lie!

The Gene Ontology

- A way to capture biological knowledge in a written and computable form

- A set of concepts and their relationships to each other arranged as a hierarchy

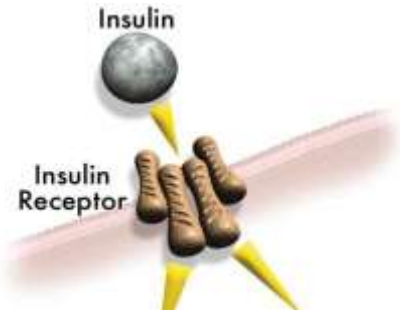


www.ebi.ac.uk/QuickGO

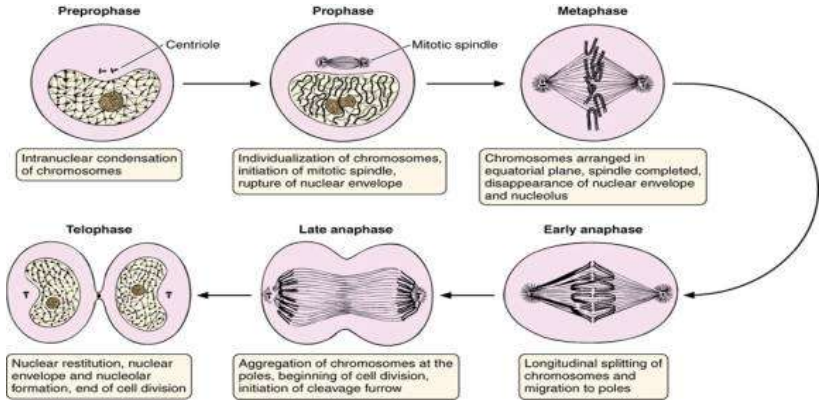
The Concepts in GO

1. Molecular Function

An elemental activity or task or job



- protein kinase activity
- insulin receptor activity



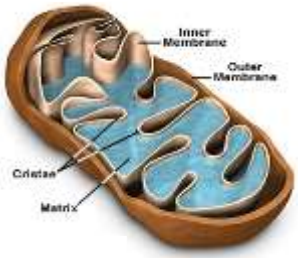
2. Biological Process

A commonly recognised series of events

- cell division

3. Cellular Component

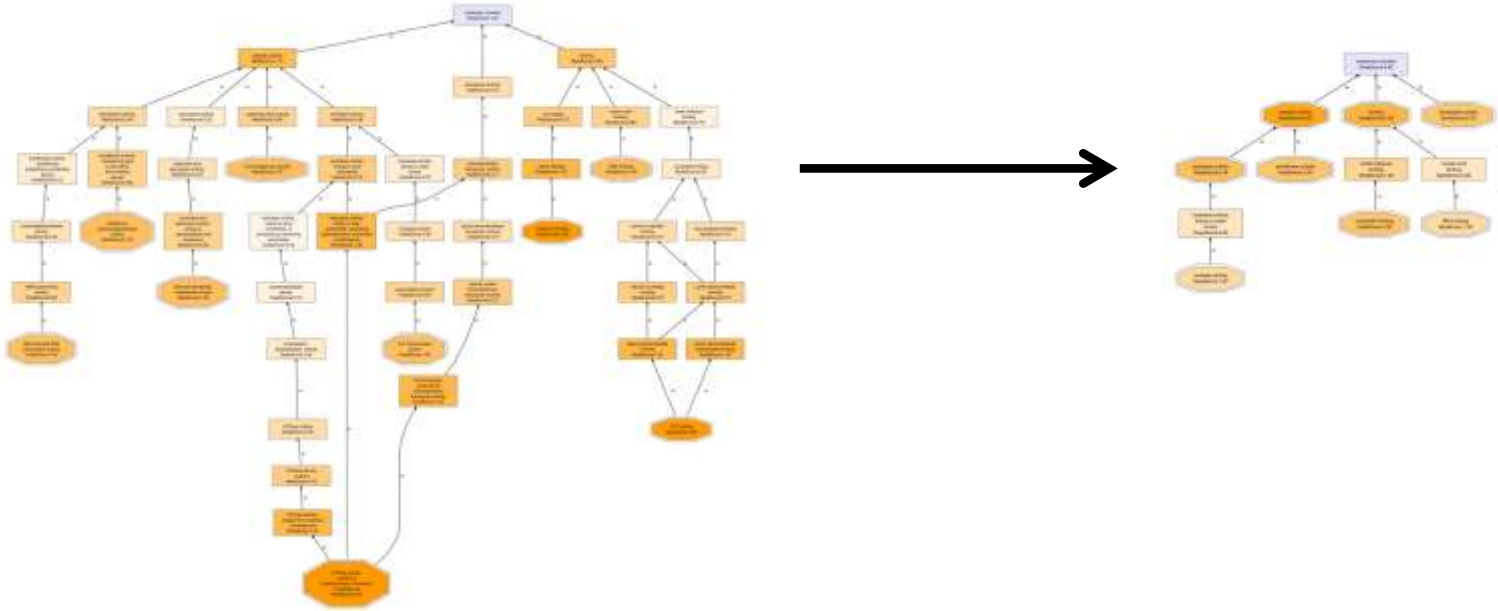
Where a gene product is located



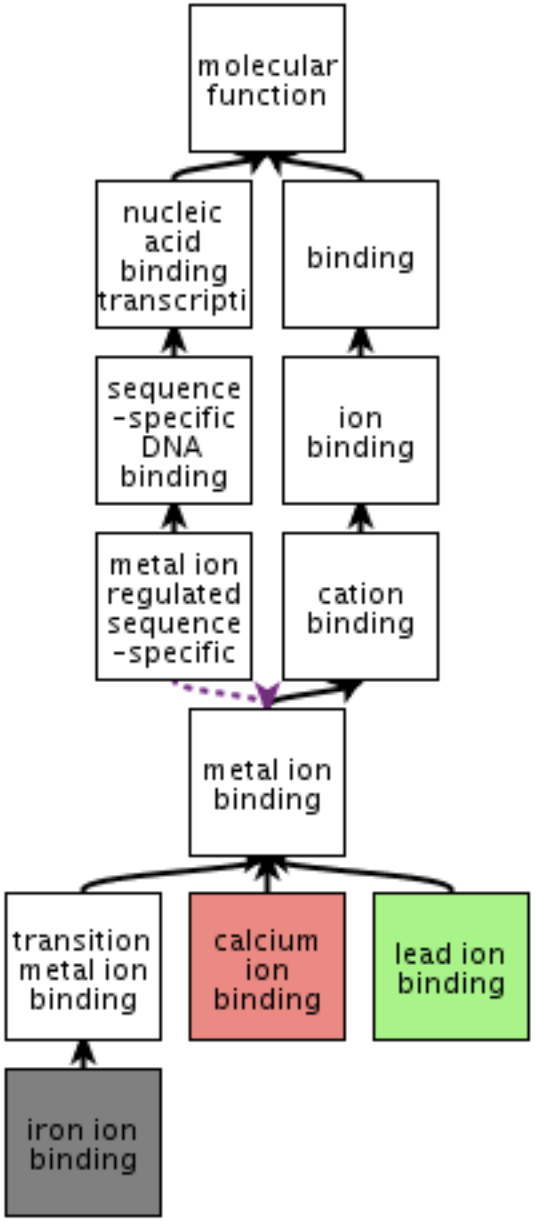
- mitochondrion
- mitochondrial matrix
- mitochondrial inner membrane

Visualising the data: GO Slims

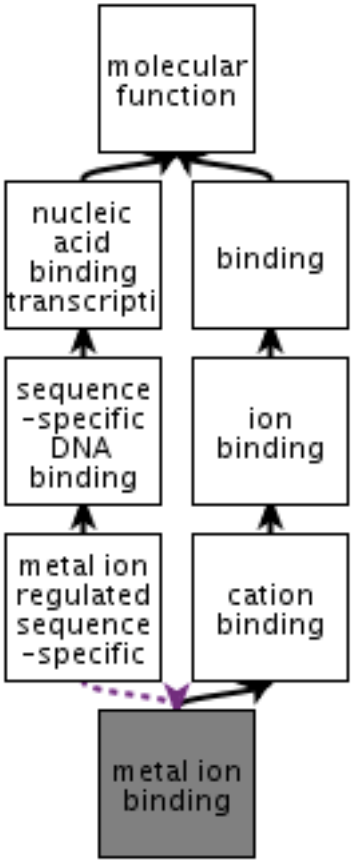
- GO slims are cut-down versions of the GO, containing a subset of terms
- Give a broad overview of the ontology content without the detail of the specific fine-grained terms



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

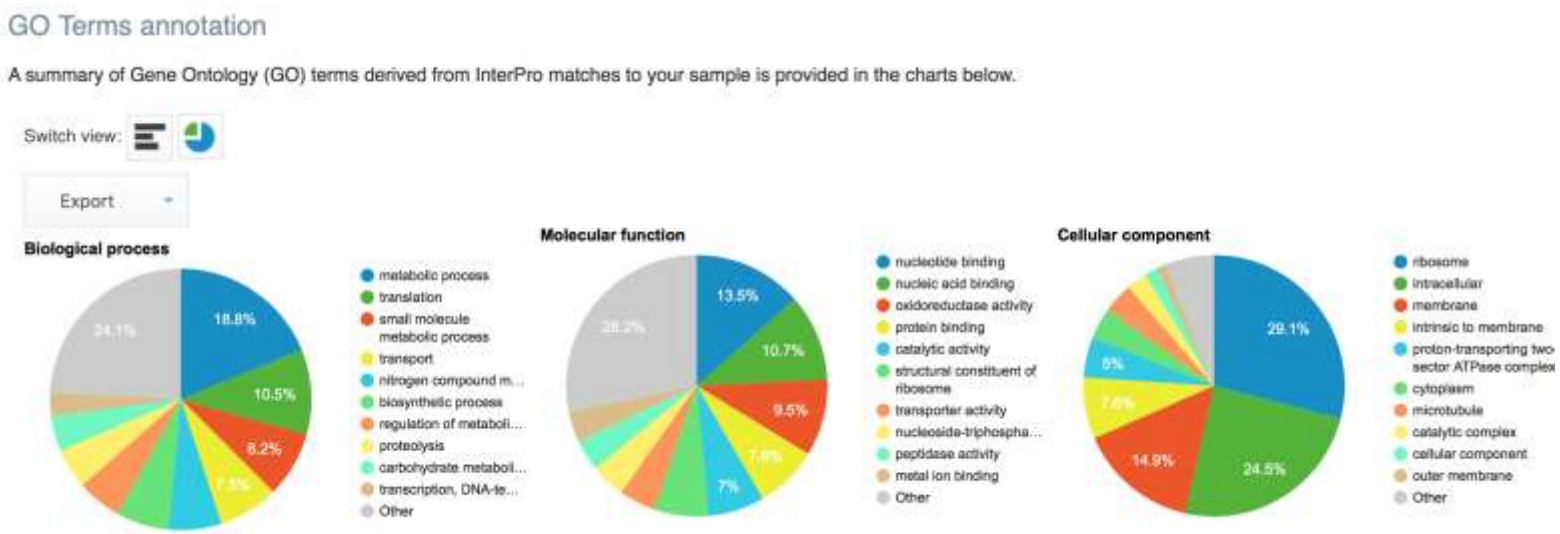
GO Slims



Slimmed term:

Visualising the data: GO slims

- For visualisation, EMG uses a GO slim, developed in-house for metagenomic data sets
- Recently rebuilt for pipeline v3.0, based on annotation of ~ 20 billion predicted coding sequences



Run (ERR762516)

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

Functional analysis has 3 main outputs: a sequence features summary, showing the number of reads with predicted coding sequences (pCDS), the number of pCDS with InterPro matches, and so on; the matches of pCDS to the InterPro database and a chart of the GO terms that summarise the functional content of the sample's sequences. If you wish to download the full set of results, all files are listed under the "Download" tab.

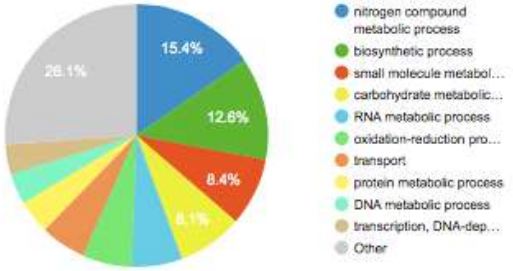
GO Terms annotation

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

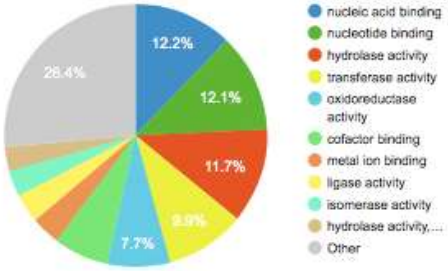
Switch view:

Export

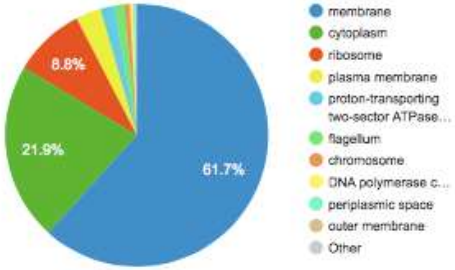
Biological process



Molecular function



Cellular component



Online comparison tool

EBI Metagenomics

Home | Submit data | Projects | Samples | Comparison tool | About EBI Metagenomics | Contact | Not logged in | Login

EBI Metagenomics > Comparison tool

Comparison tool

The comparison is currently based on a summary of Gene Ontology (GO) terms derived from InterPro matches to the selected runs.

Project list

- Long insert human faecal metagenomic library.
- Making and breaking DMS by salt marsh microbes (Illumina HiSeq 100bp)
- Meta-transcriptomic analysis of rumen microbiome of Mehsani buffalo
- MetaSoil
- Metagenome of a microbial consortium obtained from the Tuna oil field in the Gippsland ...
- Metagenome of grass carp intestinal contents and mucosa
- Metagenome sequencing of biogas plant operating wet fermentation
- Metagenomes and metatranscriptomes from the diffuse hydrothermal vents of Axial Seamoun...
- **Metagenomes of hydrogen producing, xylan fed termites**
- Metagenomic Characterisation of Opaque Beer Industry Wastewater
- Metagenomic Characterisation of Opaque Beer Industry Wastewater
- Metagenomic analysis of Ruminal Microbes
- Metagenomic analysis of sediments along a uranium gradient

More info about selected project

Run list (3 selected out of 3)

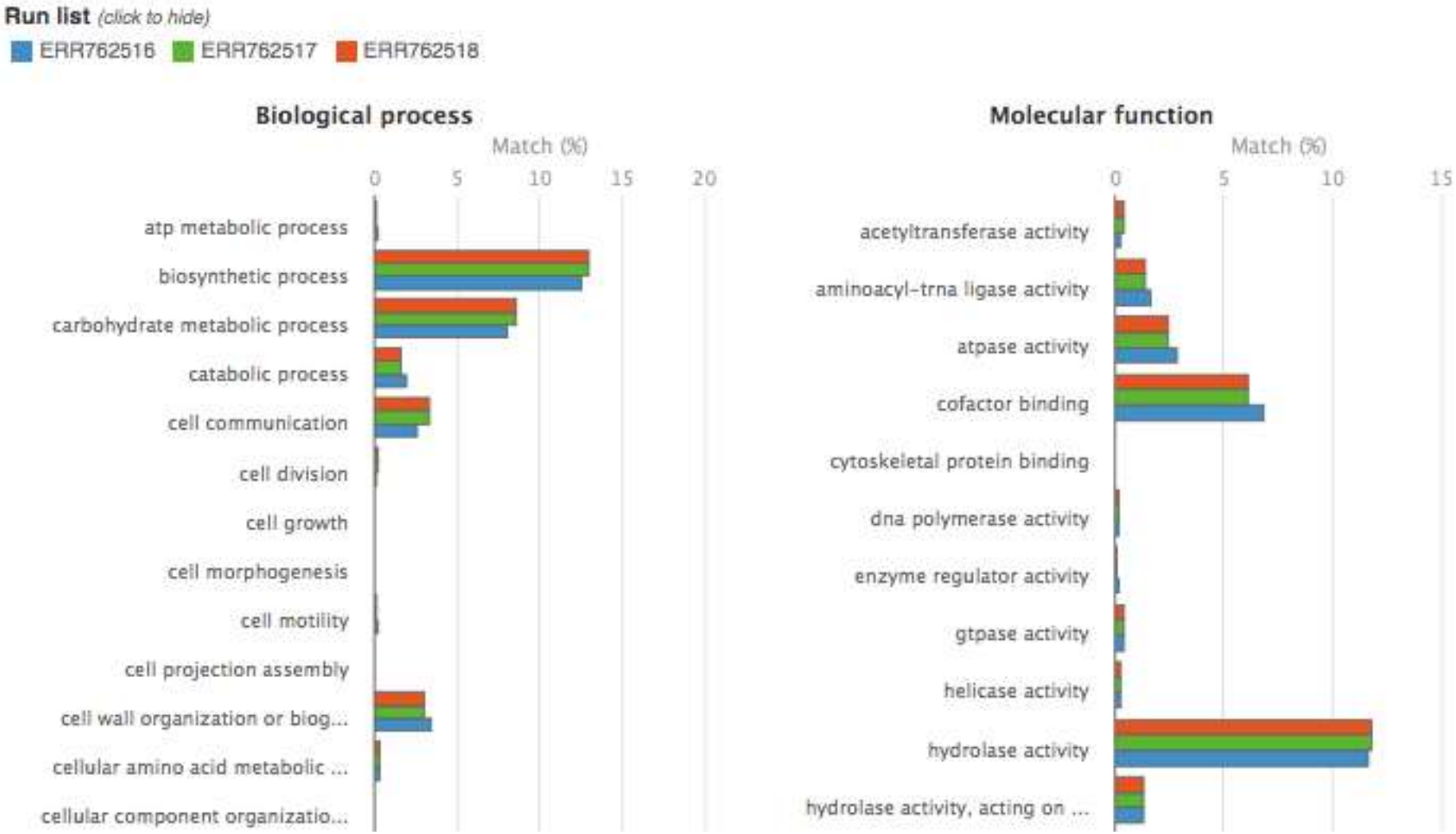
- Nasutitermes exitiosus gut metagenome incubated at 33 - ERR762516
- Nasutitermes exitiosus gut metagenome incubated at 40 - ERR762517
- Nasutitermes exitiosus gut metagenome incubated at 45 - ERR762518

Select all | Unselect all

Advanced settings

Show / hide advanced settings

Compare | Clear all


Run list (click to hide)
ERR762516
ERR762517
ERR762518
Biological process
Match (%)
0
5
10
15
20
atp metabolic process
biosynthetic process
carbohydrate metabolic process
catabolic process
cell communication
cell division
cell growth
cell morphogenesis
cell motility
cell projection assembly
cell wall organization or biog...
cellular amino acid metabolic ...
cellular component organizatio...
Molecular function
Match (%)
0
5
10
15
acetyltransferase activity
aminoacyl-trna ligase activity
atpase activity
cofactor binding
cytoskeletal protein binding
dna polymerase activity
enzyme regulator activity
gtpase activity
helicase activity
hydrolase activity
hydrolase activity, acting on ...

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

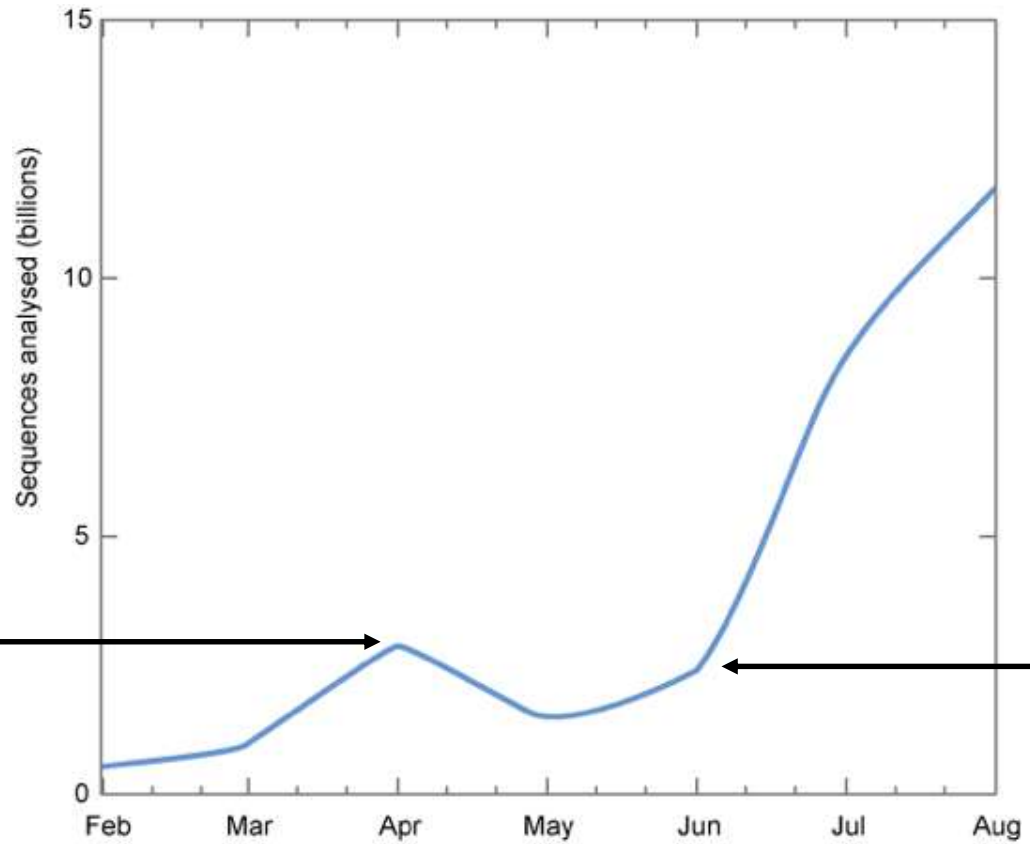
- InterPro matches (TSV) - 1 MB
- Complete GO annotation (CSV) - 44 KB
- GO slim annotation (CSV) - 7 KB

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

Pipeline Developments



Version 2.0 Max

HMMER fix, version 2.1

High throughput enables large-scale project analyses

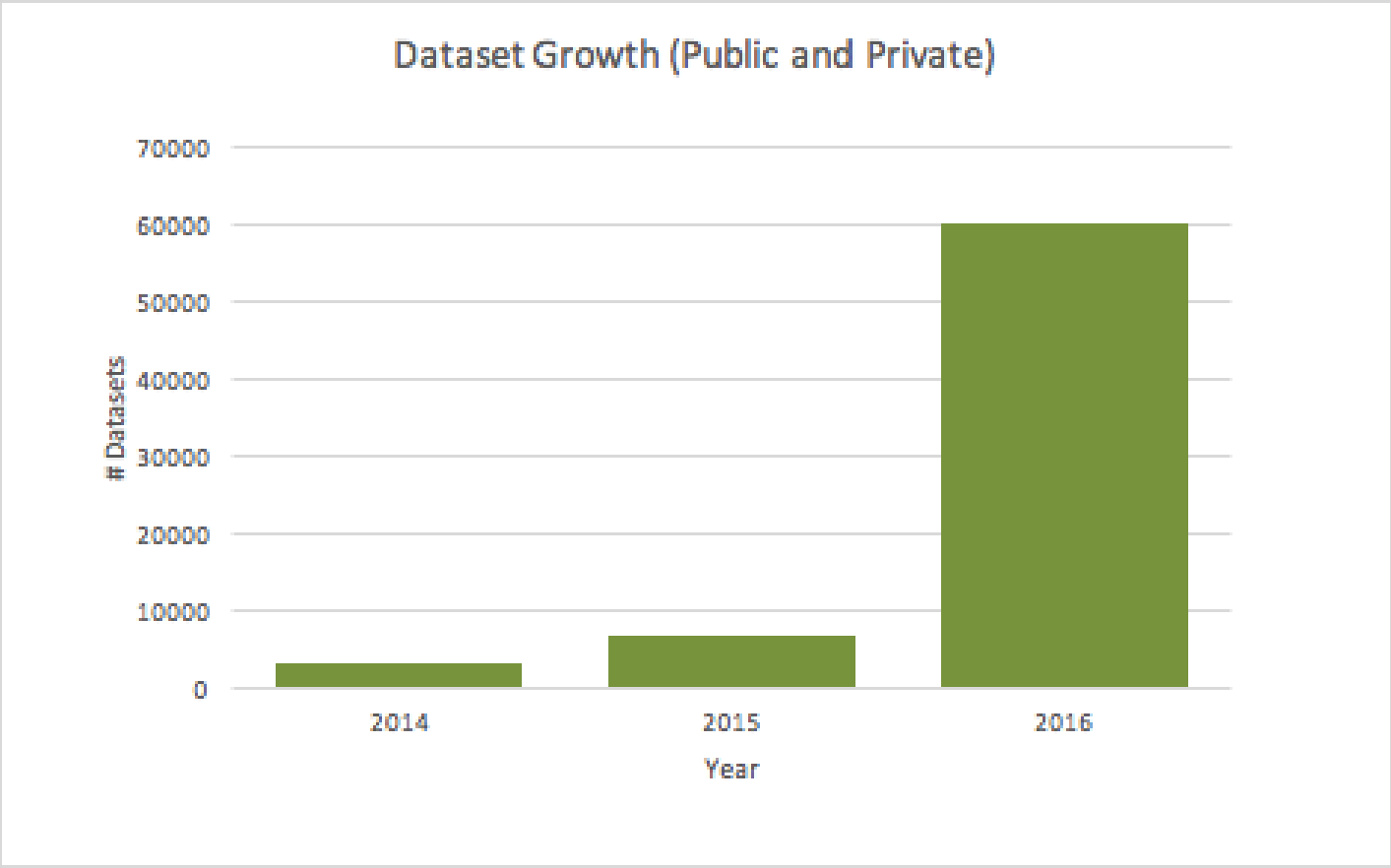
Over 8,000 16S amplicon data sets, aiming to shed light on the connections between human microbiome and health

Tara Oceans

7K samples from 210 stations sampled from around the globe

EBI analysis of Global Ocean microbiome set: 135 samples/248 runs, size-fractionated for prokaryotes = 10TB sequence data

EMG Metagenomics - Large Volumes of Data

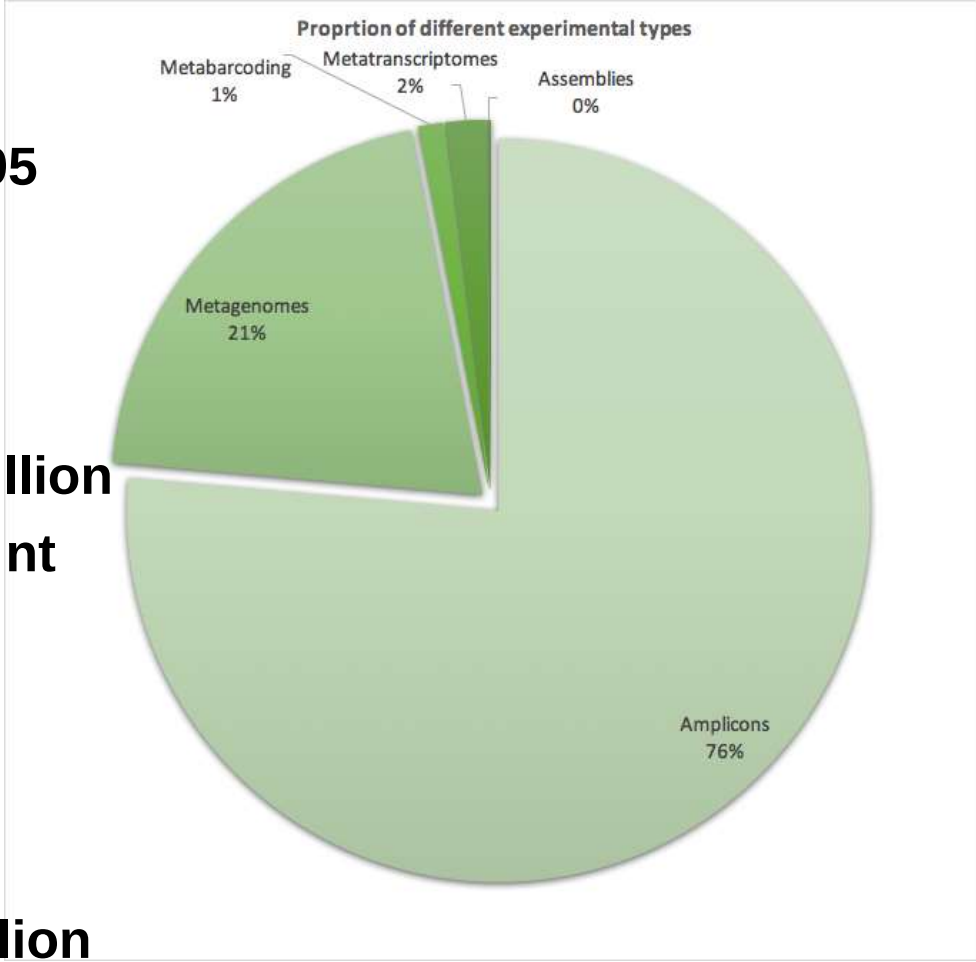


EMG Metagenomics - Large Volumes of Data

- Number of different projects: **681**
- Number of different samples: **39,795**
- Number of runs: **56,883**

- Nucleotide sequence reads: **230 billion**
- Average length per sequence: **120 nt**

- Predicted rRNAs**: 3.6 billion**
- Predicted CDS: **126 billion**
- Total InterProScan matches: **33 billion**



EBI Metagenomics

Home | Submit data | Projects | Samples | Comparison tool beta | About EBI Metagenomics | Contact | Not logged in | Login

Submit, analyse, visualize and compare your data.

SUBMIT DATA

59560 data sets

45503 amplicons
99 assemblies
688 metabarcoding
12217 metagenomes
1053 metatranscriptomes

Public: 56883 runs, 39795 samples, 681 projects

Private: 2605 runs, 2335 samples, 125 projects

Browse projects

By selected biomes

Soil (374)
Host-associated human (64)
Engineered (52)
Human digestive system (45)
Marine (44)
Host-associated mammals (43)
Host-associated plant (41)
Forest soil (26)
Freshwater (20)
Grassland (19)

Browse all biomes

Latest projects 681

Microbial community diversity in reactors digesting chicken manure
Microbial community diversity in reactors digesting chicken manure ...
View more - 9 samples - compare

Alterations of the human gut microbiome in multiple sclerosis
Alterations of the human gut microbiome in multiple sclerosis ...
View more - 210 samples - compare

A Catalogue of the Mouse Gut Metagenome
A Catalogue of the Mouse Gut Metagenome ...
View more - 184 samples - compare

Comparison of the fecal microbiota during variable collection and storage methods
Comparison of the fecal microbiota during variable collection and storage methods ...
View more - 112 samples - compare

DNA from FIT can replace stool for microbiota-based colorectal
DNA from FIT can replace stool for microbiota-based colorectal ...

View all projects

EBI Metagenomics

Home | Submit data | Projects | Samples | Comparison tool beta | About EBI Metagenomics | Contact | Not logged in | Login

Browse projects

By selected biomes


Soil (374)


Host-associated human (64)


Engineered (52)


Human digestive system (45)


Marine (44)


Host-associated mammals (43)


Host-associated plant (41)

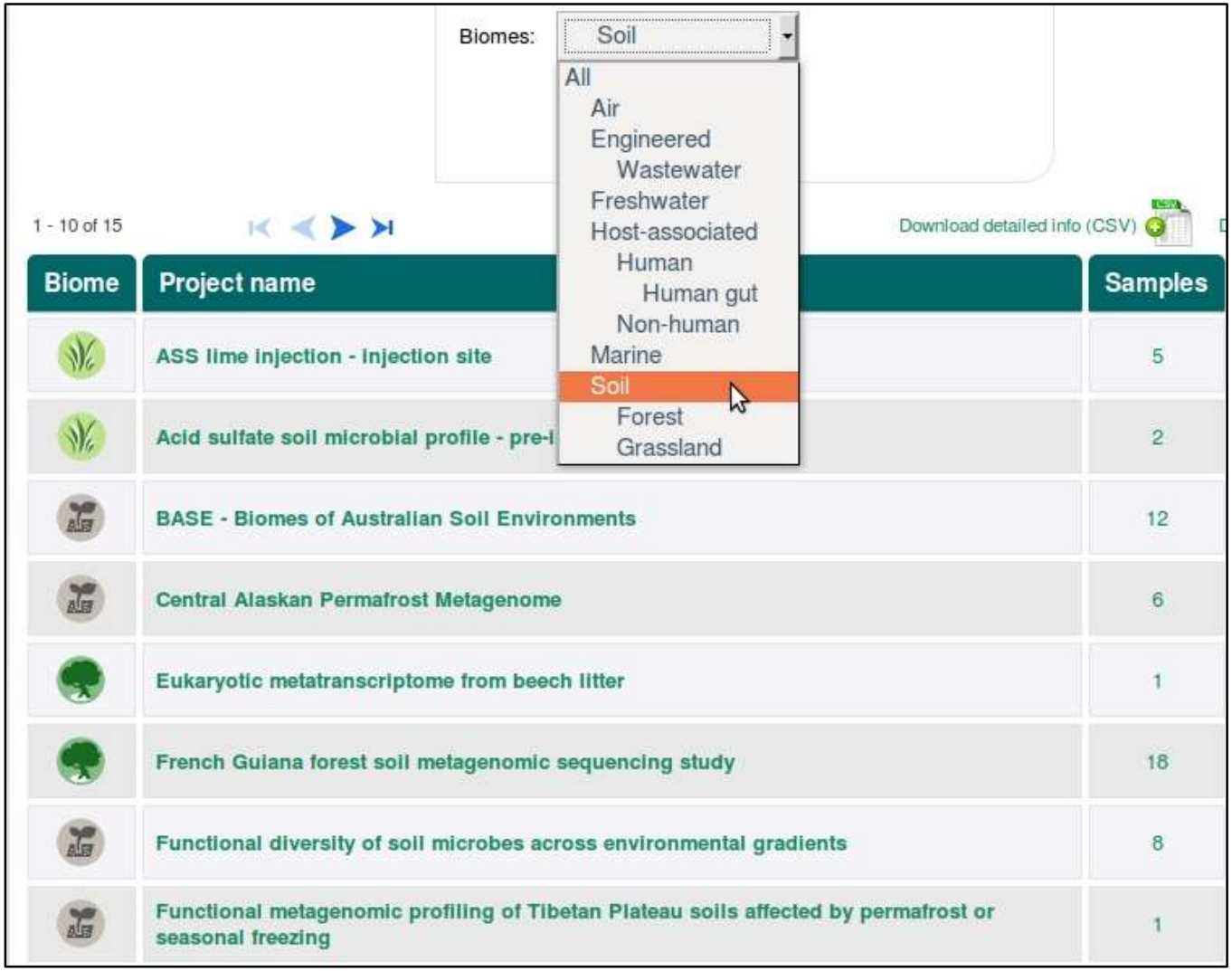

Forest soil (26)


Freshwater (20)


Grassland (19)

Browse all biomes

Biome search



Project summary files

EBI Metagenomics

Home | Submit data | Projects | Samples | Comparison tool beta | About EBI Metagenomics | Contact | Not logged in | Login

EBI Metagenomics > Project: Sheep rumen metagenomes and metatranscriptomes

Project (SRP022254)

Sheep rumen metagenomes and metatranscriptomes

Overview | Analysis summary

In this section you can download the different results matrix ... ns. To visualise and download the analysis results for individual ru

Pipeline version 1.0

Functional analysis for the project

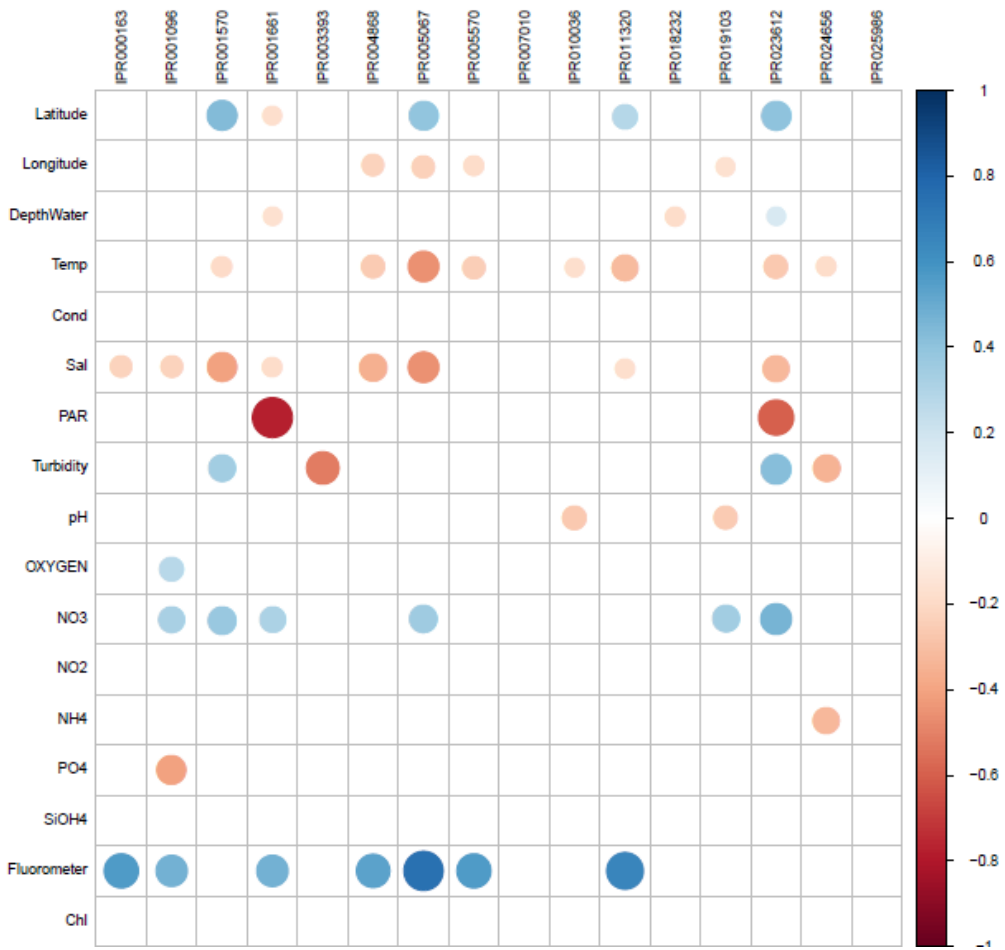
- InterPro matches (TSV) - 1 MB
- Complete GO annotation (TSV) - 435 KB
- GO slim annotation (TSV) - 29 KB

Taxonomic analysis for the project

- Phylum level taxonomies (TSV) - 3 KB
- Taxonomic assignments (TSV) - 31 KB

IPR	description	ERS223747	ERS223748	ERS223749	ERS223750	ERS223751	ERS223753	ERS223763	ERS223795	ERS223858
IPR000001	Kringle	20	33	33	25	40	30	33	1	34
IPR000003	Retinoid X recep	5	1	6	3	2	1	4	0	3
IPR000007	Tubby, C-termina	222	195	225	240	255	235	190	5	234
IPR000008	C2 domain	2116	1905	2191	1849	1951	2286	1690	47	2271
IPR000009	Protein phospha	15	20	31	15	28	30	17	0	16
IPR000010	Proteinase inhib	0	1	0	2	5	0	0	0	1
IPR000011	Ubiquitin/SUMO	1	0	0	0	0	2	0	0	2
IPR000012	Retroviral VpR/V	0	0	0	0	0	0	0	0	0
IPR000014	PAS domain	39	11	11	7	16	12	8	1	10
IPR000015	Outer membrane	96	0	0	0	0	0	0	0	1
IPR000020	Anaphylatoxin/f	1	2	0	0	1	0	0	0	0
IPR000021	Hok/gef cell toxi	3	0	0	0	0	0	0	0	0
IPR000022	Carboxyl transfe	426	332	405	280	275	386	294	8	383
IPR000023	Phosphofructoki	231	177	186	203	187	218	164	6	181
IPR000025	Melatonin recep	0	0	0	0	1	0	0	0	0
IPR000026	Guanine-specific	4	0	0	0	0	0	0	0	0
IPR000028	Chloroperoxidas	0	0	0	0	0	0	0	0	0
IPR000030	PPE family	0	0	0	0	0	0	0	0	0
IPR000031	Phosphoribosyla	94	66	54	59	55	78	63	3	87
IPR000032	Phosphocarrier H	27	0	0	1	0	1	0	0	0
IPR000033	LDLR class B repe	717	599	714	565	641	674	526	22	777
IPR000034	Laminin B type IV	163	137	176	122	149	187	130	7	154
IPR000035	Alkylbase DNA gl	1	0	0	0	0	0	0	0	0
IPR000036	Peptidase A26, o	0	0	0	0	0	0	0	0	0
IPR000037	SsrA-binding pro	8	0	1	0	0	0	0	0	0

Project-level analyses (OSD data)



Courtesy of Bernardo Duarte, João Canning-Clôde,Catarina Magalhães, Luís Torgo, Isabel Caçador. Manuscript in preparation.

Heatmap of significant Spearman correlations between protein families and environmental conditions across 150 sites

Discoverability

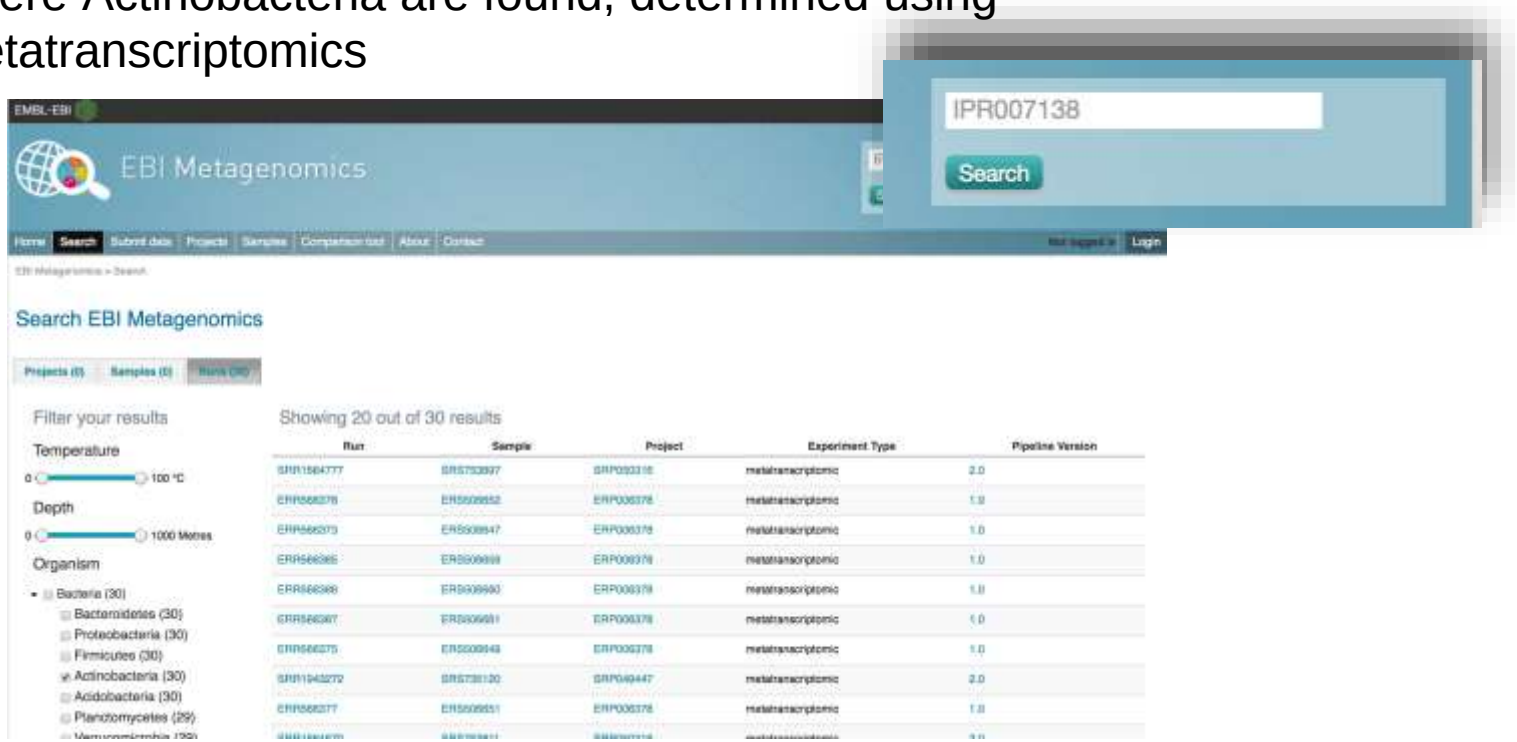
EBI-search underpins new search interface

- e.g. Give me all antibiotic biosynthesis monooxygenases in soil, where Actinobacteria are found, determined using metatranscriptomics

Discoverability

EBI-search underpins new search interface

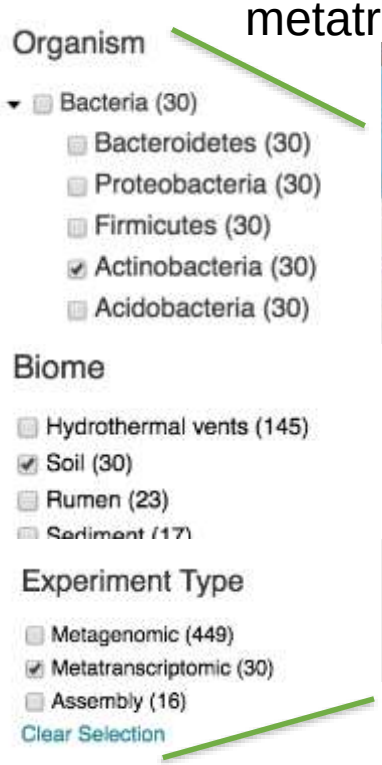
- e.g. Give me all antibiotic biosynthesis monooxygenases in soil, where Actinobacteria are found, determined using metatranscriptomics

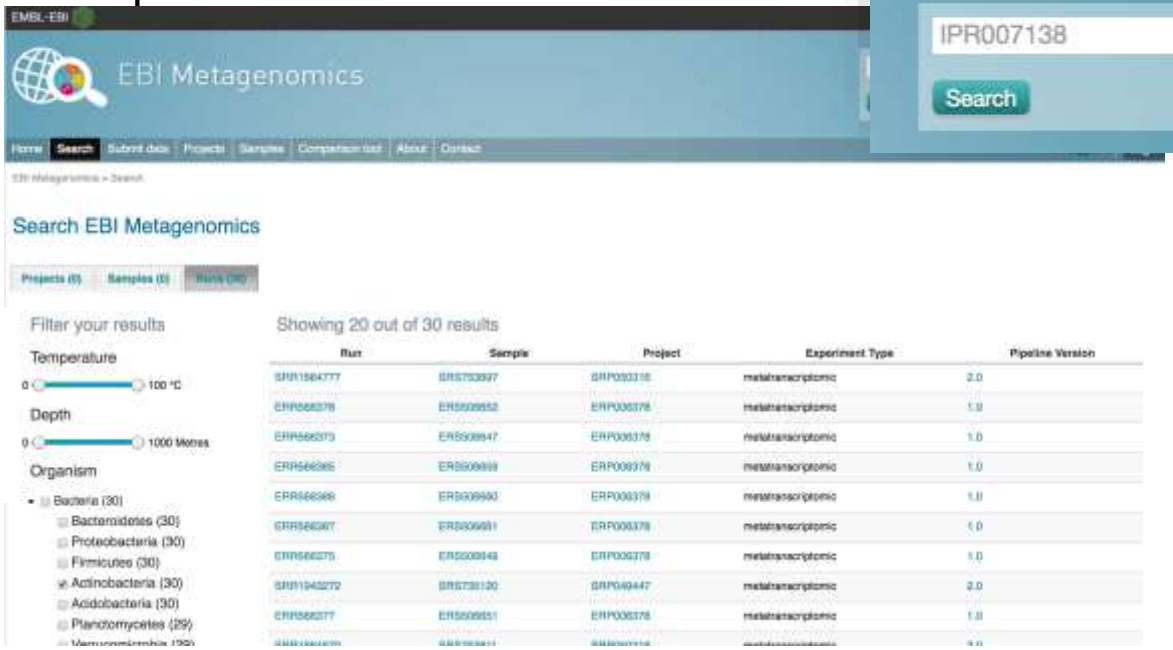


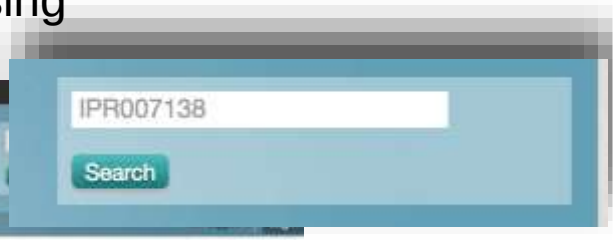
Discoverability

EBI-search underpins new search interface

- e.g. Give me all antibiotic biosynthesis monooxygenases in soil, where Actinobacteria are found, determined using metatranscriptomics



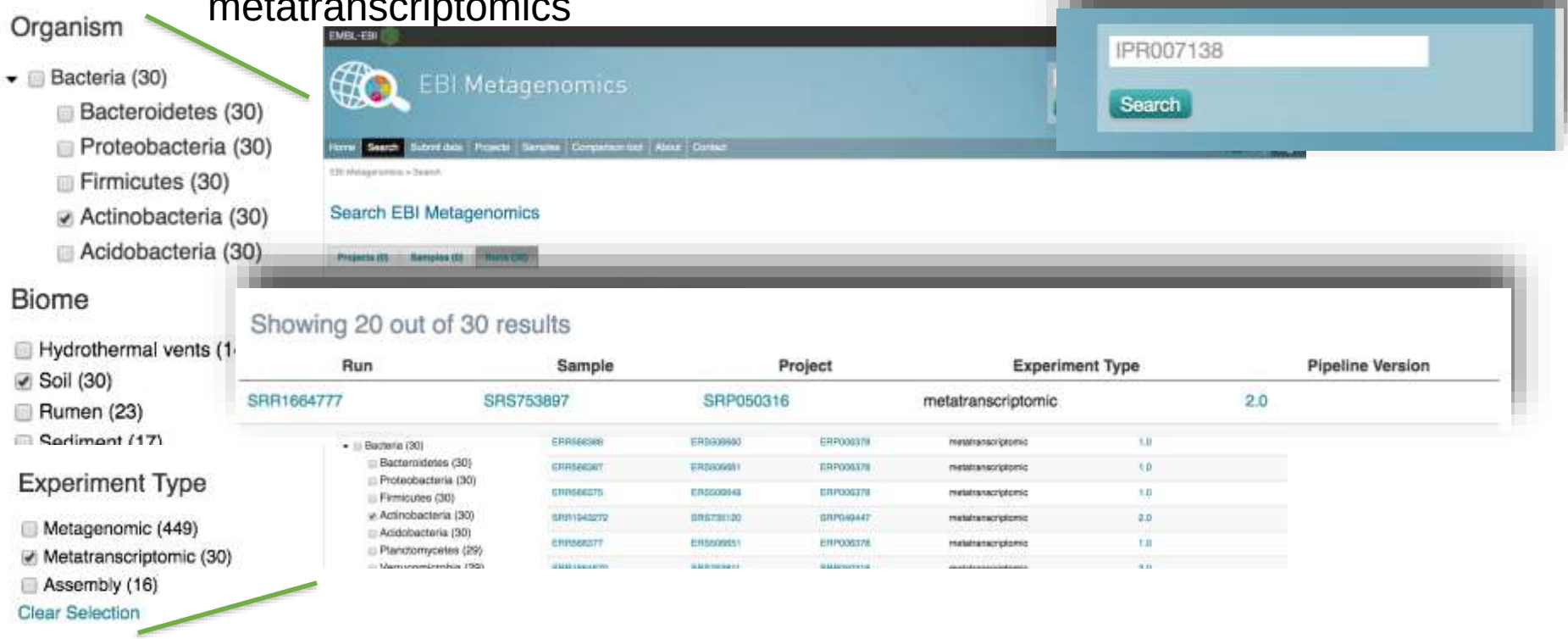




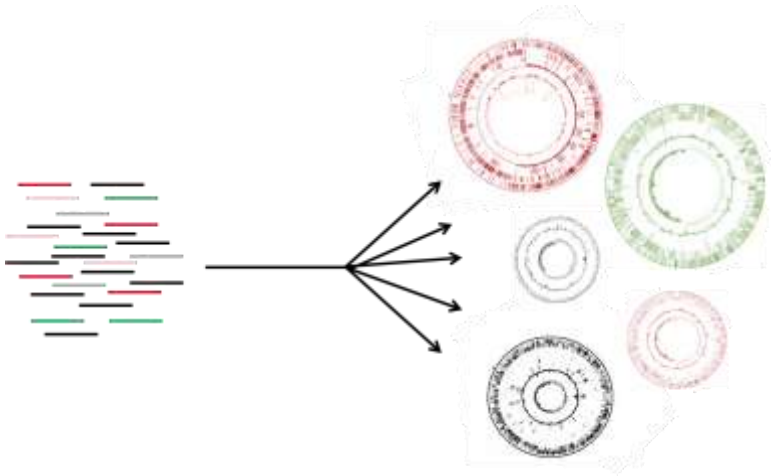
Discoverability

EBI-search underpins new search interface

- e.g. Give me all antibiotic biosynthesis monooxygenases in soil, where Actinobacteria are found, determined using metatranscriptomics



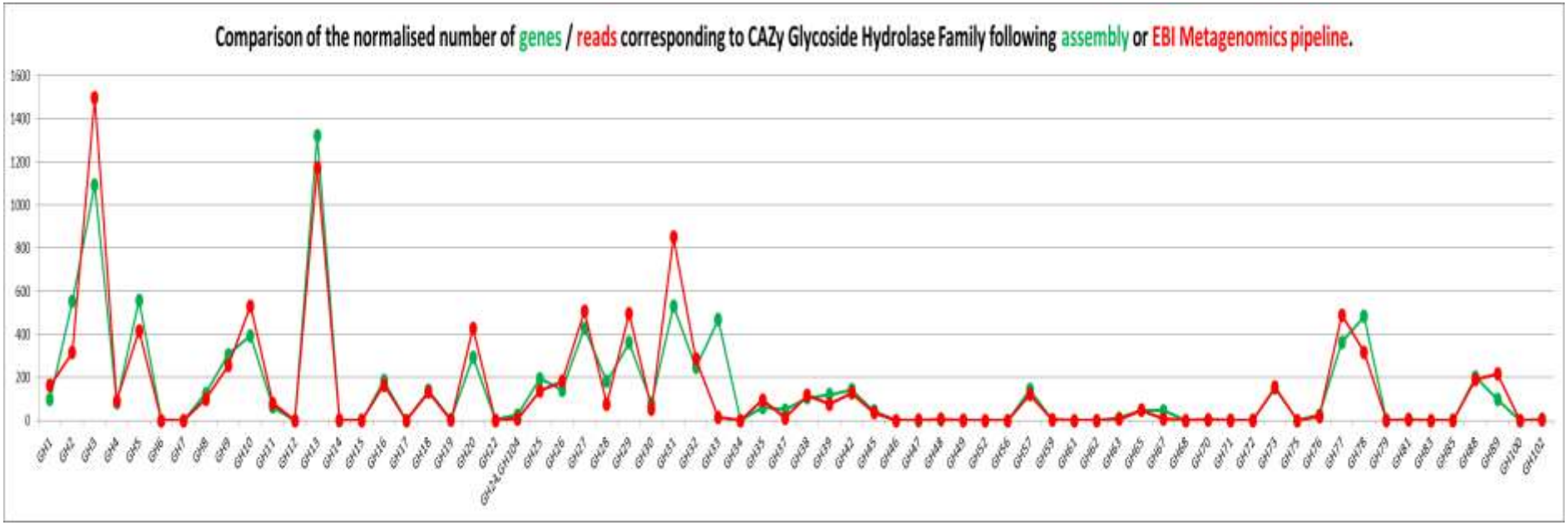
Assembly of metagenomics data?



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

Assembly is not part of standard analysis pipeline

We are still able to annotate metagenome data - re-analysis of rumen metagenomics by Hess et al, (2011)

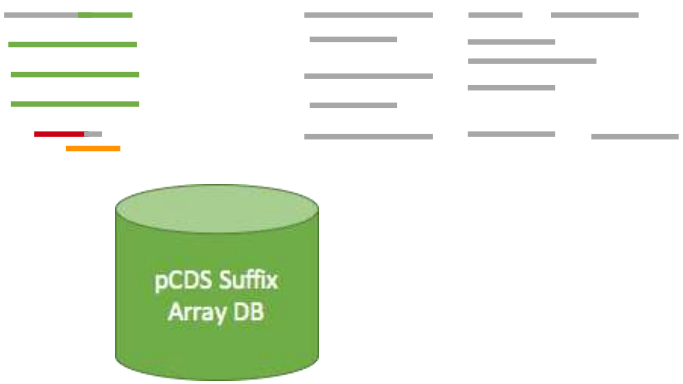


Can we perform targeted assembly?

BIOCATALYSTS
exceeding enzyme expectations

Stage 1 - Suffix Array Database Generation

Knowns Unknowns



Published online 20 November 2014 *Nucleic Acids Research, 2015, Vol. 43, No. 3 e18*
doi: 10.1093/nar/gku1210

GRASP: Guided Reference-based Assembly of Short Peptides

Cuncong Zhong, Youngik Yang and Shibu Yooseph*

Informatics Department, J. Craig Venter Institute, La Jolla, CA 92037, USA

Stage 2 - Peptide Assembly

Query

2.1 Identification of seed

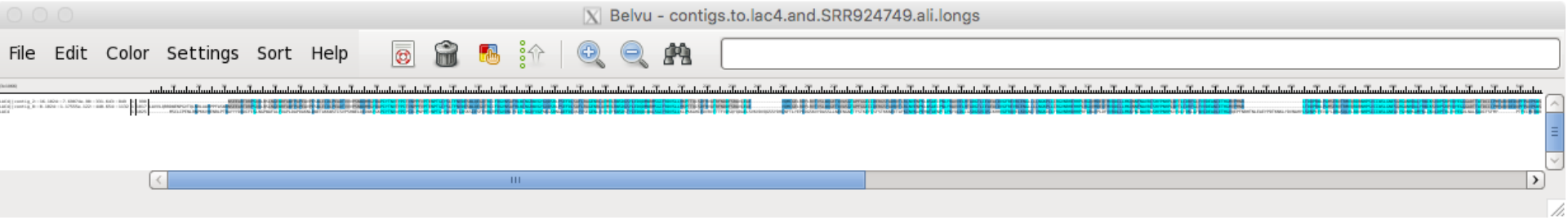
2.2 Extension of seed, N-C, C-N

2.3 Repeat for each seed

2.4 Merge seeds

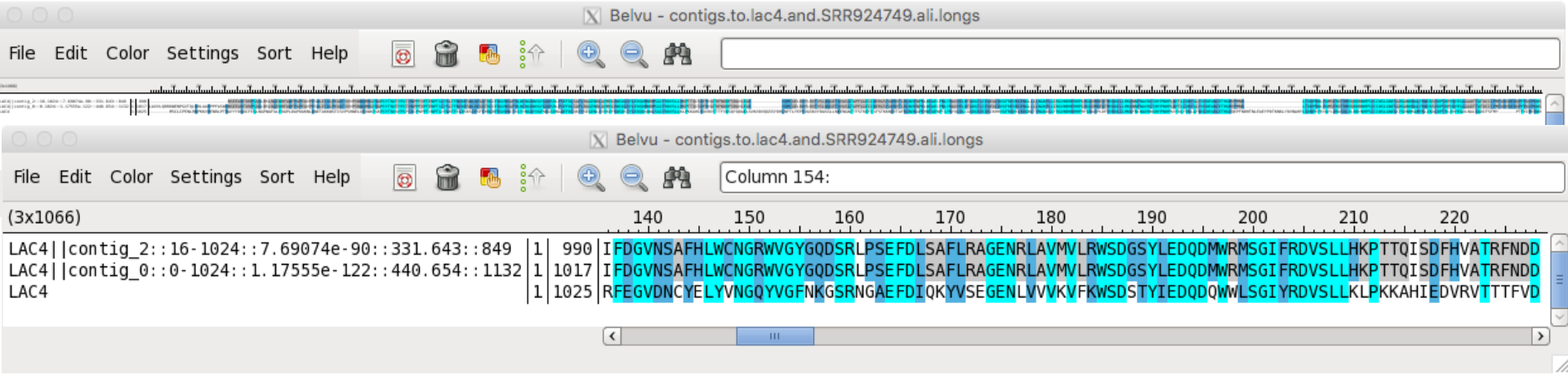
Can we perform targeted assembly?

- Query - Yeast Beta-galactosidase
- GRASP generated 79 contigs, 2 full length protein matches
- Contig_0 was 99.8% identical to the *E. coli* sequence N2H7R2_ECOLX

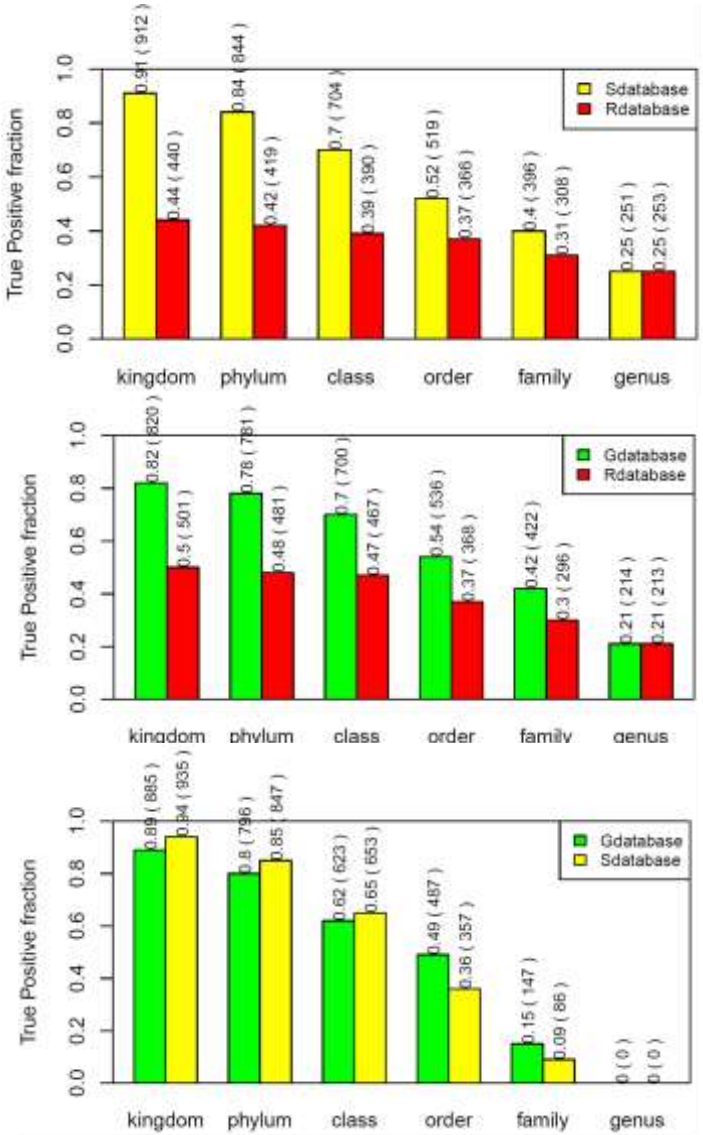


Can we perform targeted assembly?

- Query - Yeast Beta-galactosidase
- GRASP generated 79 contigs, 2 full length protein matches
- Contig_0 was 99.8% identical to the *E. coli* sequence N2H7R2_ECOLX



Comparison of different 16S reference databases



Closed Reference OTU assignment using QIIME

GreenGenes, Silva, RDP

Query: Novel Lineages

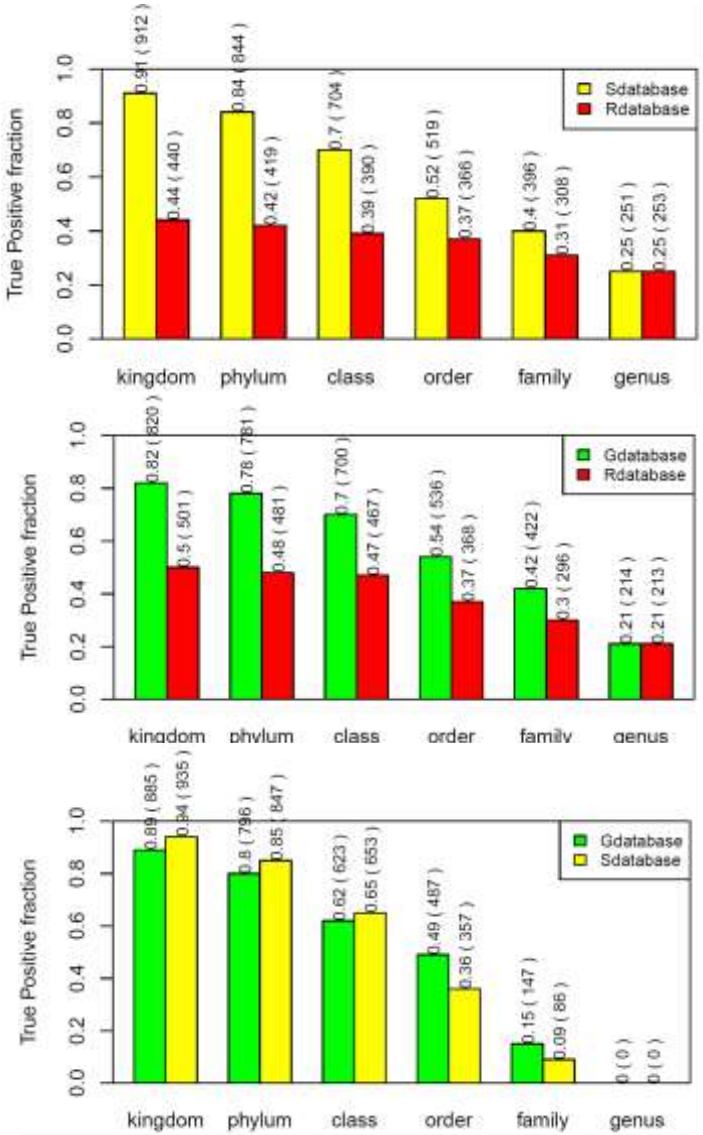
↓

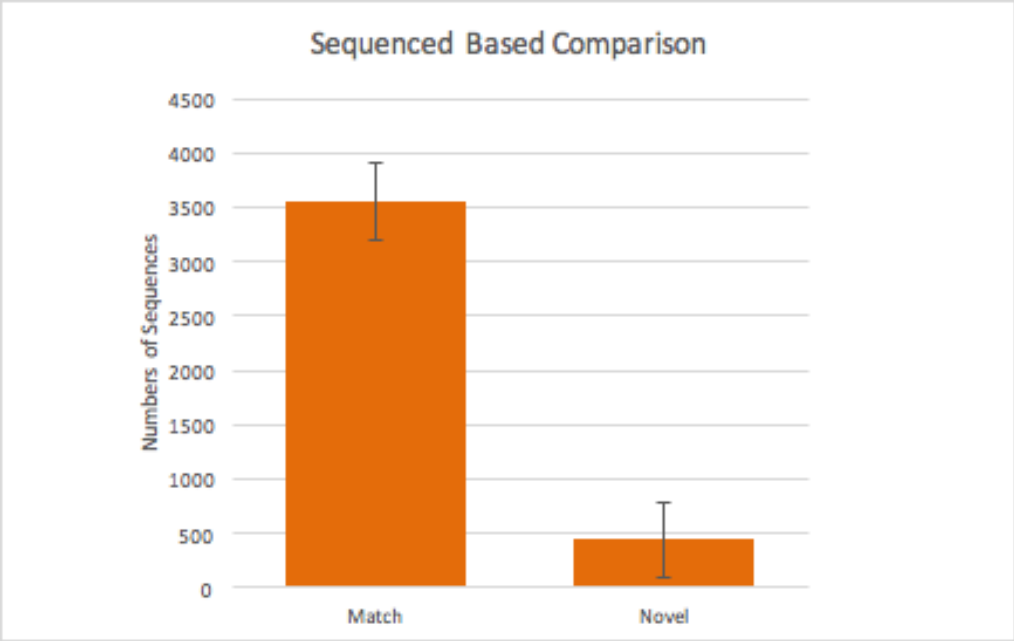
QIIME + Reference DB

↓

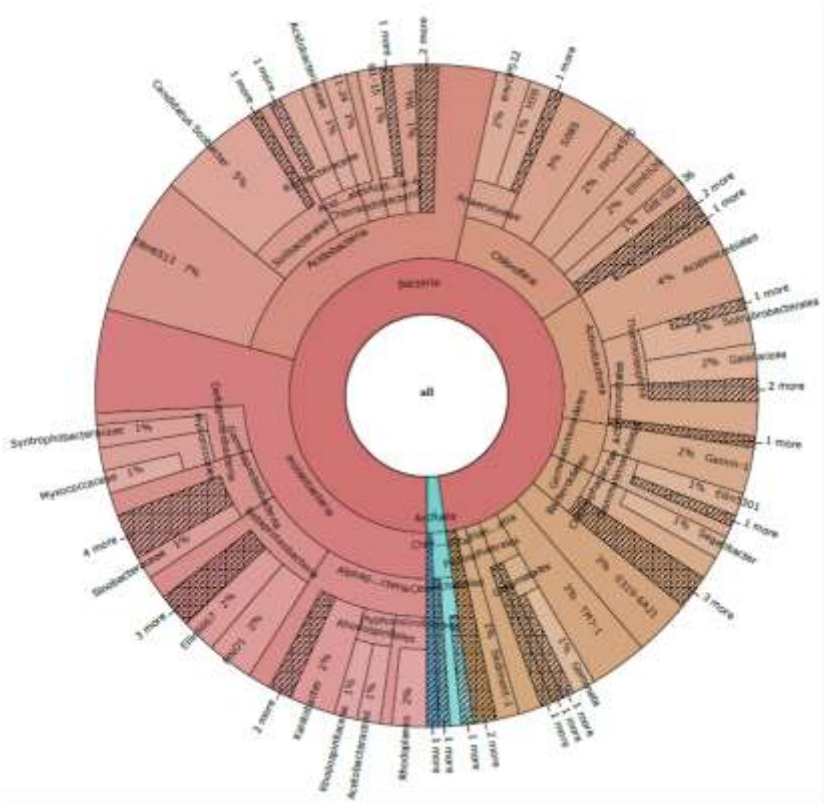
Compare OTU at different taxonomic levels

Comparison of different 16S reference databases

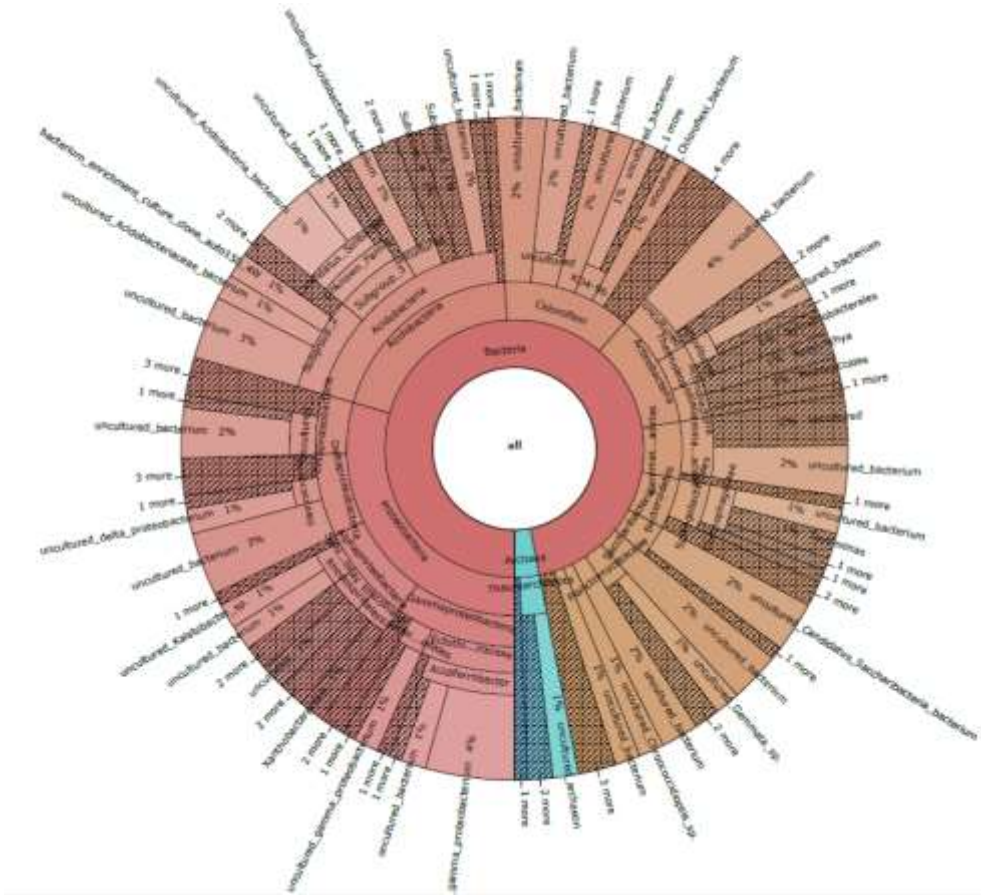




Japan tsunami-affected soil



GreenGenes



Silva

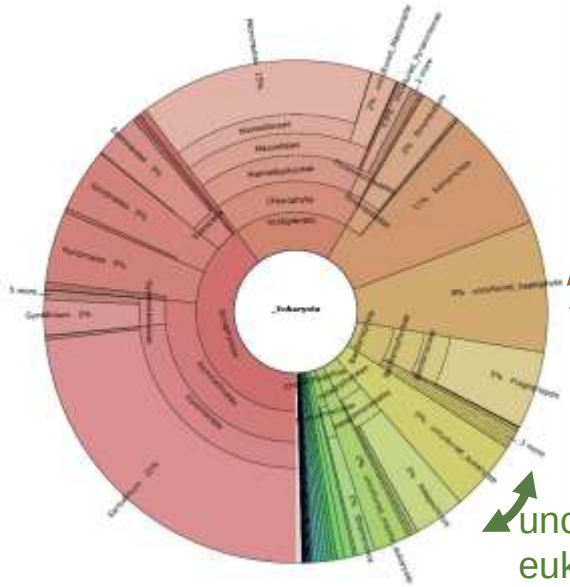
Hiraoka et al., 2016

Moving beyond 16S rRNA

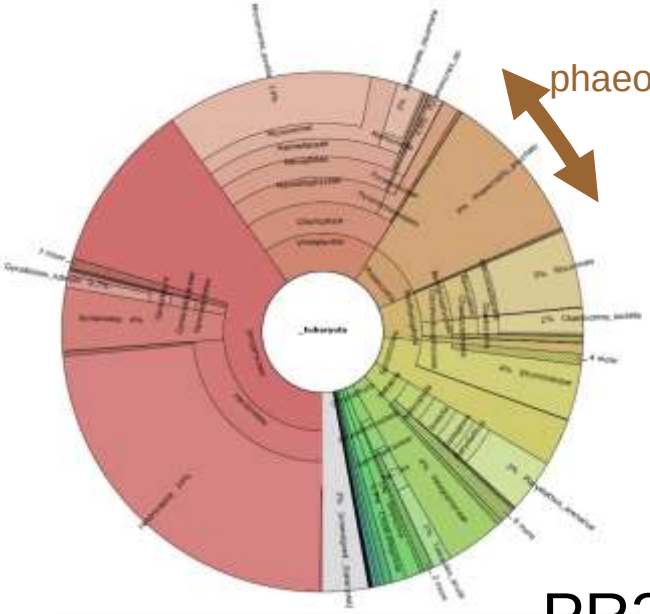
New analysis module to replace RNAselector

- Uses Rfam models for LSU, SSU and 5/5.8S
- Better discrimination of 16S/18S rRNAs
- Allows identification of 18S from WGS data sets

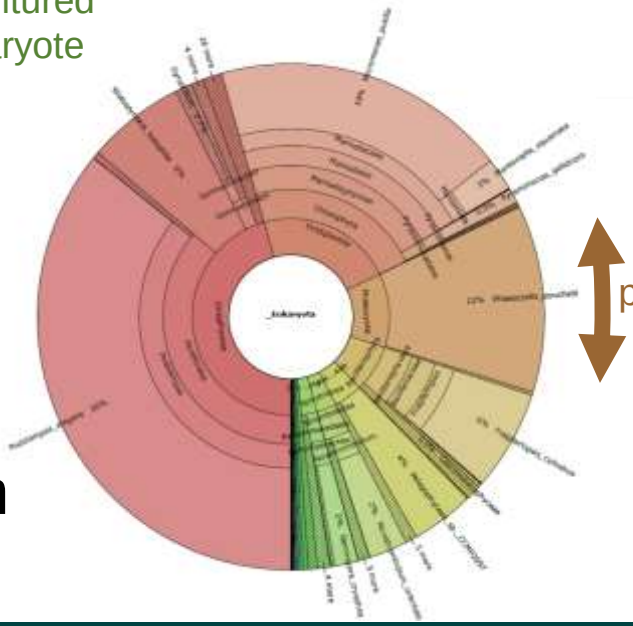
18S rRNA data: OSD Greenland Strait



Silva



PR2



Rfam

Other portals

MG-RAST
metagenomics analysis server

http://metagenomics.anl.gov/

img

http://img.jgi.doe.gov/

camera
Marine Microbial Ecology

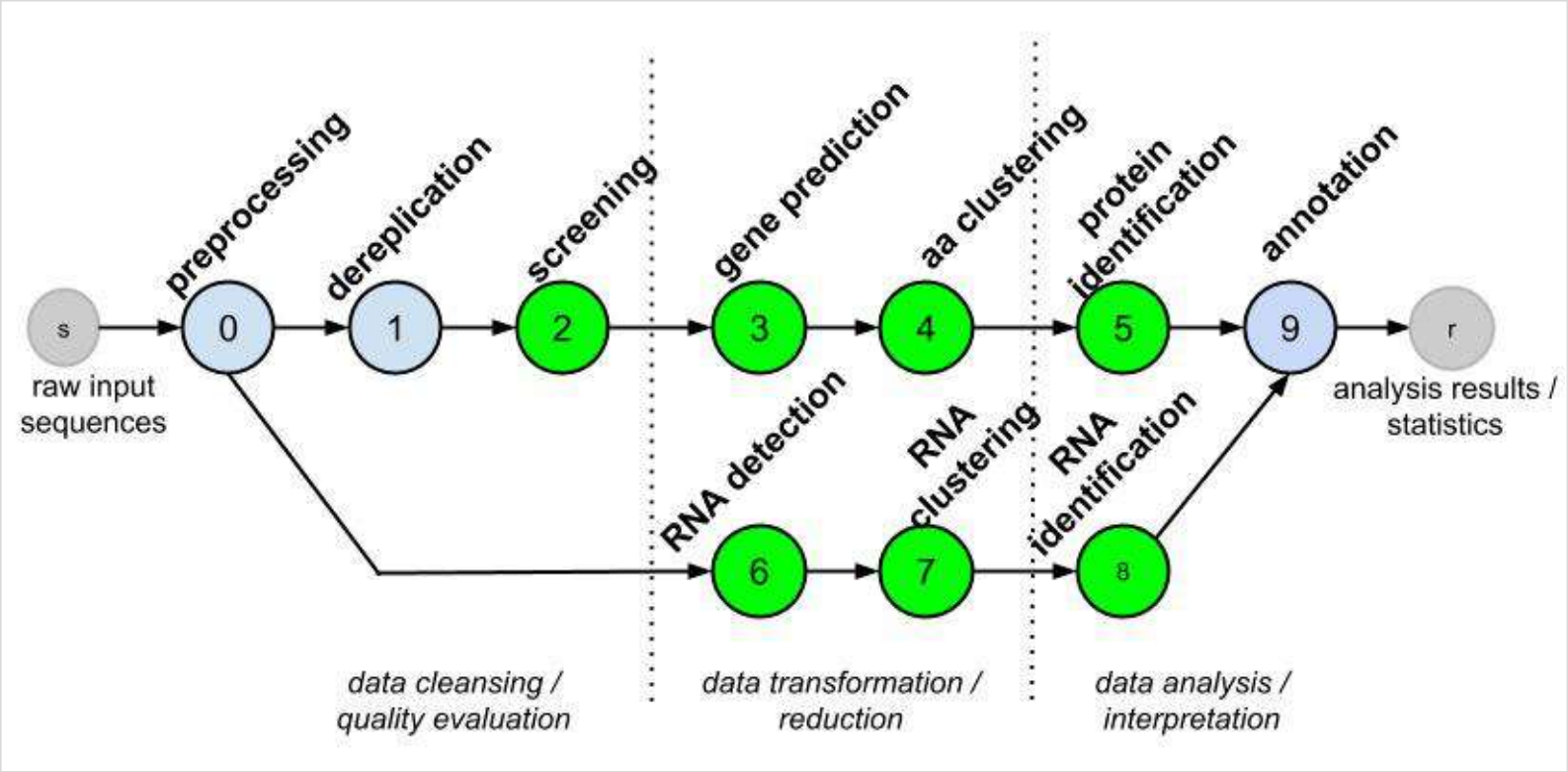
http://camera.calit2.net/

→

iMicrobe

http://imicrobe.us

Simplified MG-RAST workflow



Source: MG-RAST

High level summary of the MG-RAST workflow

- High performance quality control – tested with over 250,000 data sets
- Feature prediction
  - Intrinsic feature (gene) prediction for **CDS** – sequencing error tolerant
  - Extrinsic feature prediction for **rRNA**
  - No other features
- Clustering
  - 90% identity for CDSs
  - 97% for rRNAs
- superBLAT similarity vs. M5NR meta-database
  - **similarity based with protein and rRNA references**
  - Controlled hierarchies: SEED, KEGG pathways, eggNOGs, COGs
  - SILVA, Greengenes, RDP
  - Lowest-common-ancestor, best-hit, representative hit strategies
  - NCBI taxonomy; (soon also SILVA taxonomy)
- Storing abundance profiles
- Query based on profiles
  - No decision on e-value, % Identity or alignment length before query time

Search page is a good entry point

Metagenome Search

Search for Metagenomes

john.spears@mines.edu | Search

Match[?] metadata / MG-RAST id function organism

Find by metadata / mg-rast id

MG-RAST id, name, biome, project name... | Search

Find by function or functional category

BatE, 3.4.11.9, RNA Metabolism... | Search

Find by organism

Firmicutes, Mobiluncus curtisii... | Search

Note: To create a collection, first select the metagenomes in the first column, then click "create collection".

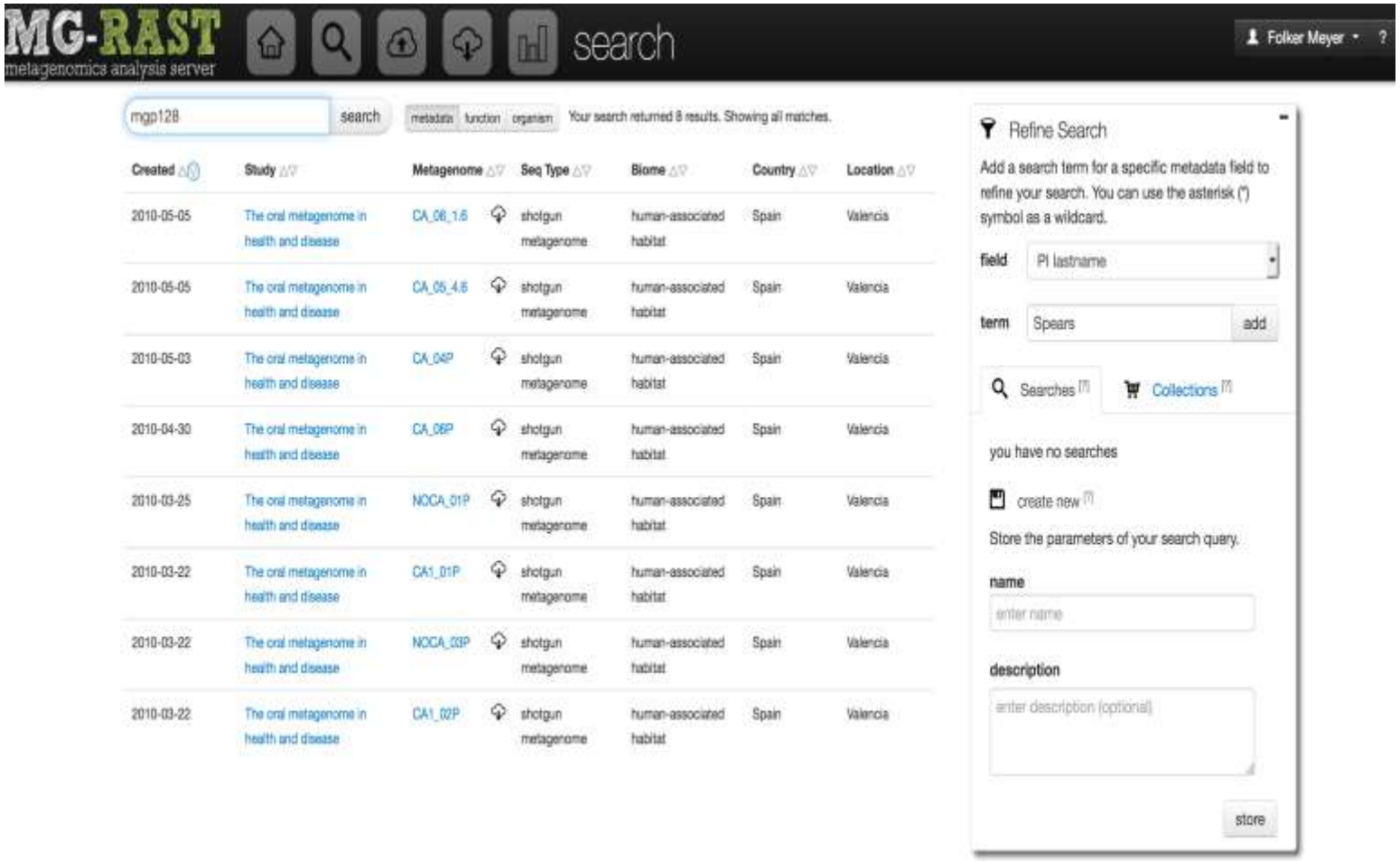
create collection

«first «prev Displaying 1-10 of 454 results next» last»

Select	Seq Type	Metagenome (ascending)	MG-RAST ID	Project	Biome	Feature	Material	Country	Location
	WGS	10K4-90	4517551.3	KAUST_SoilColumnT	terrestrial biome	subterrestrial habitat	Soil	Saudi Arabia	KAUST
	WGS	10K4-90	4518539.3	KAUST_SoilColumnT	terrestrial biome	subterrestrial habitat	Soil	Saudi Arabia	KAUST
	WGS	10K5-120	4517552.3	KAUST_SoilColumnT	terrestrial biome	subterrestrial habitat	Soil	Saudi Arabia	KAUST
	WGS	10K5-120	4518540.3	KAUST_SoilColumnT	terrestrial biome	subterrestrial habitat	Soil	Saudi Arabia	KAUST
	WGS	1K1-0	4517399.3	KAUST_SoilColumn	terrestrial	subterrestrial	Soil	Saudi	KAUST

Data is organised into studies

Unique IDs: e.g. **mgp128**



Study page

MG-RAST metagenomics analysis server | study | Folker Meyer ?

The oral metagenome in health and disease (mgp128)

principle investigator Alex Mira, CSISP
visibility public
static link http://metagenomics.anl.gov/linkin.cgi?project=mgp128

description

The oral cavity of humans is inhabited by hundreds of bacterial species and some of them have a key role in the development of oral diseases, mainly dental caries and periodontitis. We describe for the first time the metagenome of the human oral cavity under health and diseased conditions, with a focus on supragingival dental plaque and cavities. Direct pyrosequencing of eight samples with different oral-health status produced 1 Gbp of sequence without the biases imposed by PCR or cloning. These data show that cavities are not dominated by Streptococcus mutans (the species originally identified as the ethiological agent of dental caries) but are in fact a complex community formed by tens of bacterial species, in agreement with the view that caries is a polymicrobial disease. The analysis of the reads indicated that the oral cavity is functionally a different environment from the gut, with many functional categories enriched in one of the two environments and depleted in the other. Individuals who had never suffered from dental caries showed an over-representation of several functional categories, like genes for antimicrobial peptides and quorum sensing. In addition, they did not have mutans streptococci but displayed high recruitment of other species. Several isolates belonging to these dominant bacteria in healthy individuals were cultured and shown to inhibit the growth of cariogenic bacteria, suggesting the use of these commensal bacterial strains as probiotics to promote oral health and prevent dental caries.

funding source

Spanish MICINN: SAF2009-13032-C02-02 from the I+D program, BIO2008-03419-E from the EXPLORA program and MICROGEN CSD2009-00006 from the Consolider- Ingenio program.

contact

Administrative
Alex Mira (mira_ale@gva.es)
CSISP (http://www.csisp.gva.es/web/alex/home)
Avda. CataluÃƒÂ±a, 21. Valencia, Spain

Overview page

MG-RAST metagenomics analysis server

shotgun metagenome

NOCA_01P

processing receipt

This shotgun metagenome is part of the study 'The oral metagenome in health and disease'

by Alex Mira, CSISP - published in The ISME journal, 2012 Jan.

Visibility	public	**NCBI Project ID**	-
ID	mgm4447192.3	**GOLD ID**	-
Static Link	http://metagenomics.anl.gov/linkin.cgi?metagenome=mgm4447192.3	**PubMed ID**	21716308
Sample	mgs25820	**Library**	mgl52920

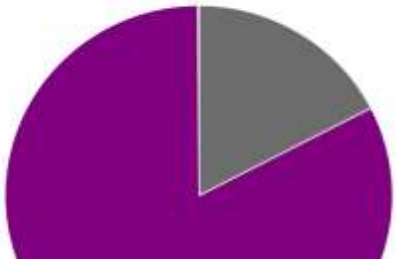
The data set NOCA_01P was uploaded on 2010-03-25 at 13:52:44 and contains 204,218 sequences totaling 77,538,485 basepairs with an average length of 380 bps.

Of the sequences tested, 35,703 sequences (17.48%) failed to pass the QC pipeline. Of those, dereplication identified 19,171 sequences as artificial duplicate reads.

Of the sequences that passed QC, 2,973 sequences (2%) contain ribosomal RNA genes, 151,581 sequences (89.95%) contain predicted proteins with known functions, and 13,961 sequences (8.28%) contain predicted proteins with unknown function.

The data on this page represents the automated analysis generated by the MG-RAST automated processing pipeline. Details on processing this data set are here. Data shown here is displayed as a quick way to assess the quality and contents of the data set. We note that the submitting authors may have performed their own analysis. The analysis page provides the best way to perform in-depth analyses of this data set.

Sequence Breakdown



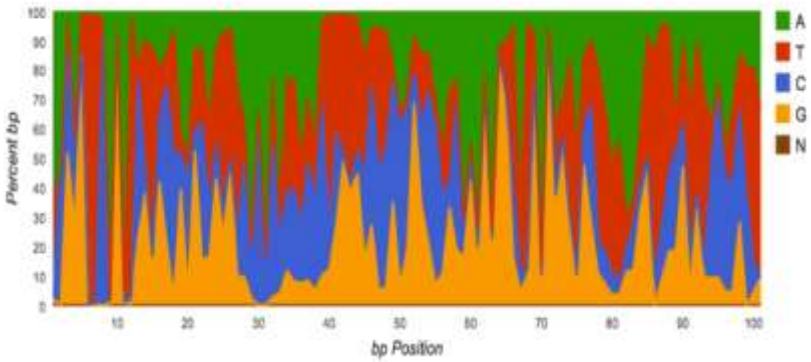
- failed QC - 17.48%
- unknown - 0.00%
- predicted feature - 82.52%

- Unique ID per data set
- E.g. **mgm4447192.3**

Overview page II: Nucleotide histogram

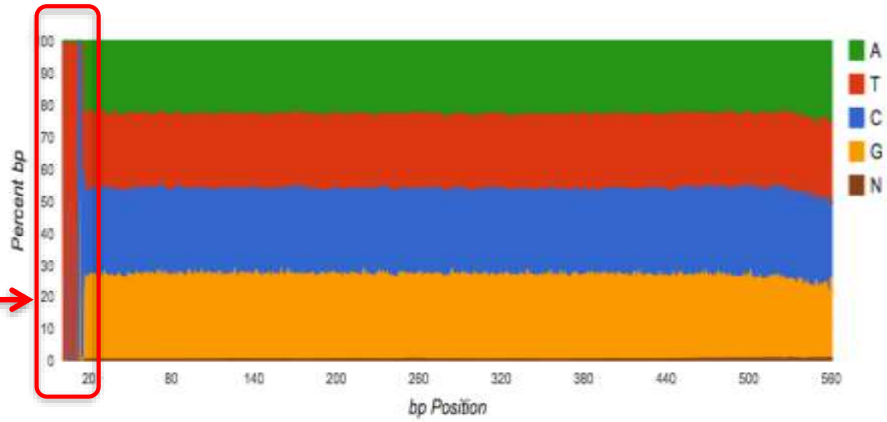
- Amplicon datasets should show biased distributions of bases at each position

Reflects conservation and variability in the recovered sequences:



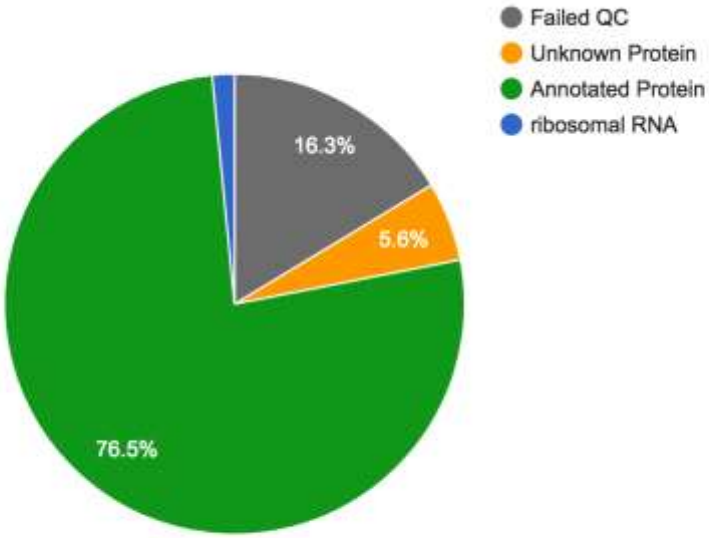
- WGS datasets should have roughly equal proportions of basecalls

An example of untrimmed barcodes



Overview page III: Piechart and flowchart

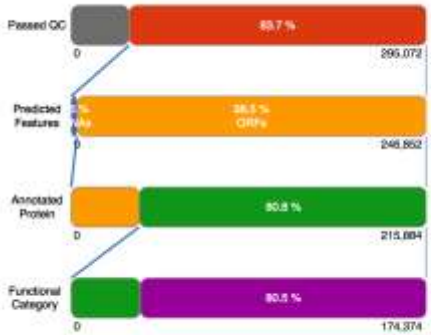
Sequence Breakdown



Note: Sequences containing multiple predicted features are only counted in one category.
Currently downloading of sequences via chart slices is not enabeled.

ANALYSIS FLOWCHART

48,220 sequences failed quality control. Of those, dereplication identified 24,870 sequences (8.4% of total) as artificial duplicate reads (ADRs). Of the 248,852 sequences (totaling 116,177,839 bps) that passed quality control, 243,055 (98.5%) produced a total of 215,884 predicted protein coding regions. Of these 215,884 predicted protein features, 174,374 (80.8% of features) have been assigned an annotation using at least one of our protein databases (M5NR) and 41,510 (19.2% of features) have no significant similarities to the protein database (orfans). 140,353 features (80.5% of annotated features) were assigned to functional categories.



ANALYSIS STATISTICS

Upload: bp Count	129,851,692 bp
Upload: Sequences Count	295,072
Upload: Mean Sequence Length	440 ± 120 bp
Upload: Mean GC percent	46 ± 12 %
Artificial Duplicate Reads: Sequence Count	24,870
Post QC: bp Count	116,177,839 bp
Post QC: Sequences Count	248,852
Post QC: Mean Sequence Length	470 ± 69 bp
Post QC: Mean GC percent	46 ± 12 %
Processed: Predicted Protein Features	215,884
Processed: Predicted rRNA Features	22,055
Alignment: Identified Protein Features	174,374
Alignment: Identified rRNA Features	1,588
Annotation: Identified Functional Categories	140,353

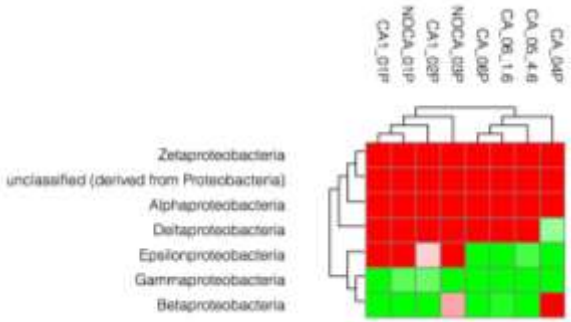
Overview and **Analysis** page

- Overview page provides rough overview of taxa and functions
  - Allows initial glance
  - Uses bad parameters for your data set
- Use the **Analysis page** for
  - **Comparison**
  - **Filtering**
  - **Subsetting**
  - **Data export**
  - **Parameter adjusting**

Adjust parameters on the analysis page for your question

This is typically missed by benchmarking papers and reviews

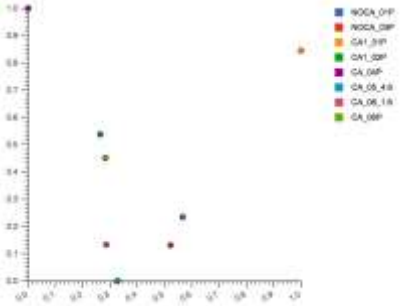
Range of visualisations / comparisons



Heatmap comparisons

Zoom to individual genus

Compare specific functions



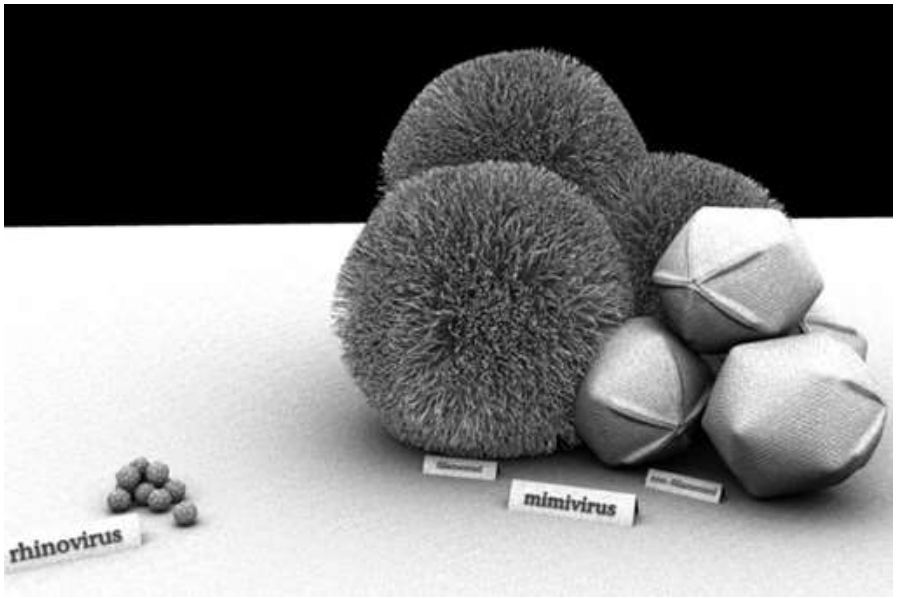
PCoA

Export results: images, data, sequence sets
Fully-featured API

Viral resources

VIROME: Viral Informatics Resource for Metagenome Exploration
http://virome.dbi.udel.edu

METAVIR: Annotation and comparison of viral metagenomic sequences
http://metavir-meb.univ-bpclermont.fr



[Source: http://www.weizmann.ac.il/Structural_Biology/research]

Gene/Genome Catalogues

MetaHit: Metagenomes of the human intestinal tract. 3.3 million genes.
http://www.metahit.eu

HMP reference genomes: data from several 1,000 reference genomes isolated from human body sites.
http://hmpdacc.org/reference_genomes/reference_genomes.php

Ocean Microbial Reference Catalog: gene catalogues from Tara Oceans* GOS, and other publicly available reference sets.
http://ocean-microbiome.embl.de/companion.html#OM-RGC

*New Euk set: 117 million genes. BLAStx comparison against UniRef90 + MMETSP took ~ **9 million** cpu hours on HPC infrastructure (nodes with 2*8 cores SandyBridge@2.7GHz and 64 Gb of RAM)

EBI Metagenomics

- Analysing large amounts of data via a standardised pipeline
- Systematically indexing results and metadata to support data discovery
- Improvements underway: better taxonomic analyses, targeted assembly, API
- Many other tools and resources out there – no ‘perfect’ tool covering all analyses

Follow/tweet us at:
@EBImetagenomics

Acknowledgements

EBI Metagenomics

- Robert Finn
- Hubert Denise
- Maxim Scheremetjew
- Sebastien Pesseat
- Simon Potter
- Matloob Qureshi

ENA European Nucleotide Archive

- Guy Cochrane
- Petra Ten Hoopen

MG-RAST metagenomics analysis server

- Folker Meyer

EMBL

BBSRC bioscience for the future

excelerate