Public data resources for metagenomics

TGAC workshop:
Metagenomics: From Bench to Data Analysis
15/10/2015

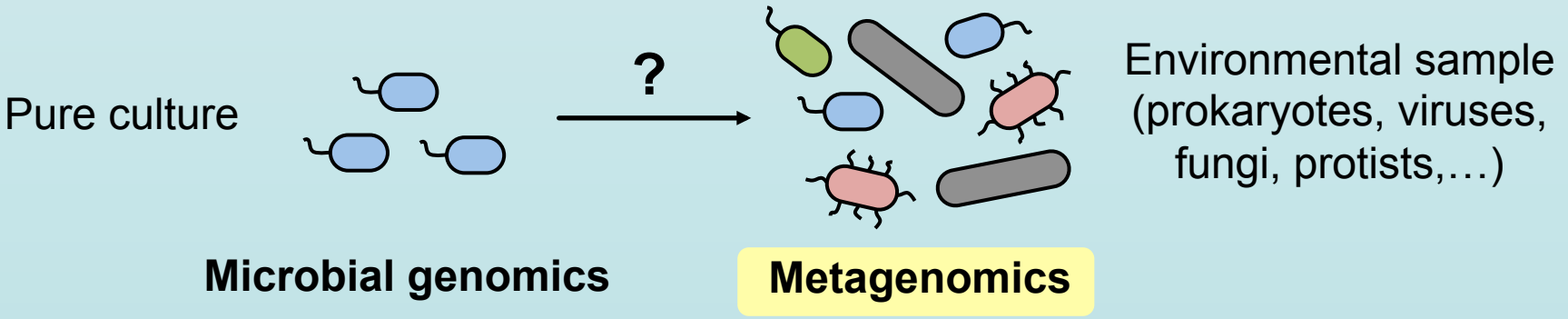
Jenny Pratscher
UEA
j.pratscher@uea.ac.uk

Metagenomics

University of East Anglia

What?

Problem: Majority (~99%) of microbial biodiversity have not been isolated yet!



→ Term first appeared in publication in 1998 (Handelsman *et al.*)

Direct genetic analysis of genomes contained with an environmental sample

Metagenomics

University of East Anglia

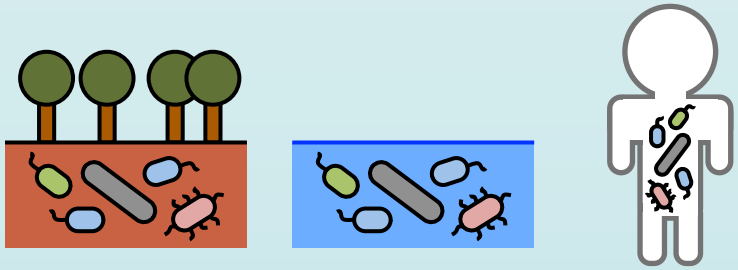
Why?

→ 1 g of soil contains 10^9-10^{10} microbial cells, most of those uncultured

Huge resource of so far unknown gene/protein functional information, for finding new biomarkers, etc.

Metagenomics

How does it work?



Environmental sample

DNA extraction & sequencing

Assembly ⟷ Binning

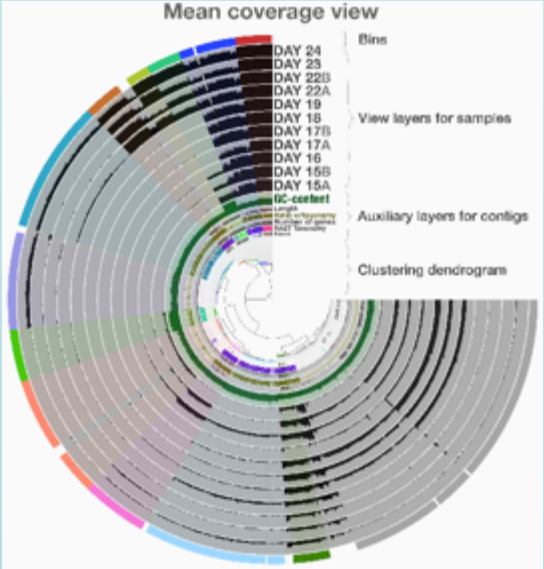
Annotation

Statistical analysis ⟷ Metadata

Metagenomics

What to do with your metagenome?

- Gene discovery
- Phylogenetic analyses
- Functional metagenomics → Metatranscriptomics, -proteomics
- Comparative metagenomics
  - → Changes in phylogenetic diversity
  - → Functional comparison (e.g. pathways)
- Extraction of genomes from metagenomes



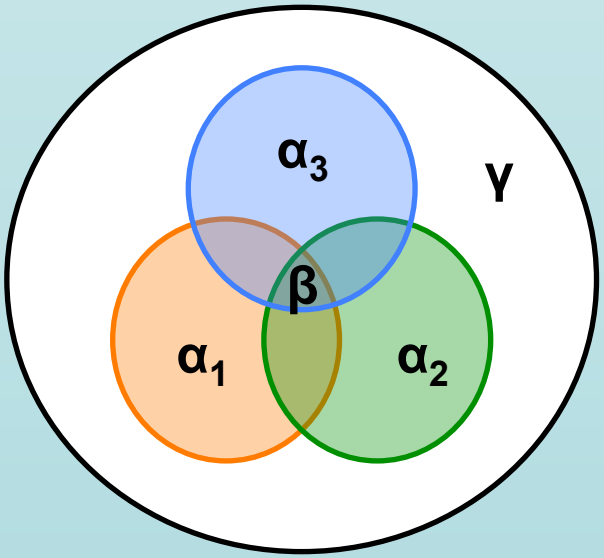
Comparison of infant gut metagenomes (Eren *et al*. 2015, Anvi'o platform)

Metagenomics

University of East Anglia

Comparative Metagenomics: Diversity analyses

- **Alpha diversity**:
  Diversity within a sample/ecosystem (species richness)
- **Beta diversity**:
  Comparison of diversity between samples/ecosystems
- **Gamma diversity**:
  Measure of the overall diversity within a large region



Metagenomics

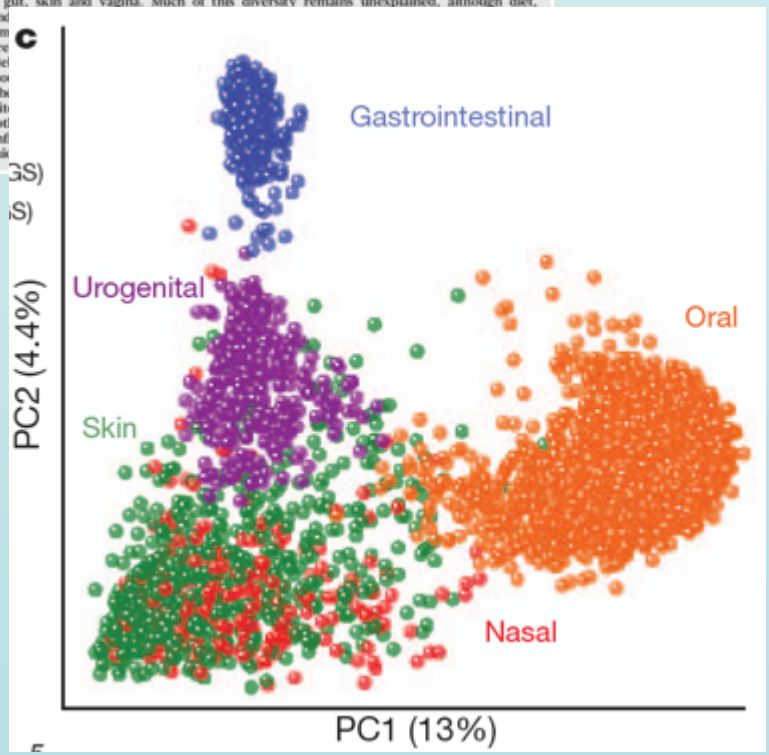
Comparative Metagenomics: Diversity analyses

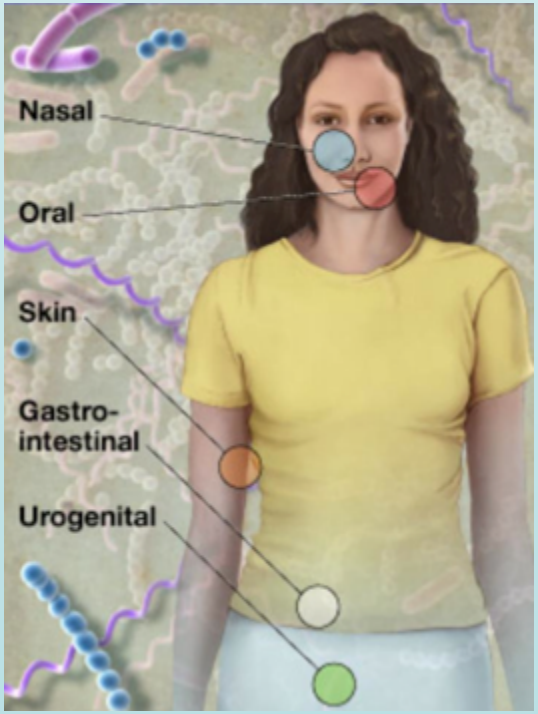
Structure, function and diversity of the healthy human microbiome

The Human Microbiome Project Consortium*

Studies of the human microbiome have revealed that even healthy individuals differ remarkably in the microbes that occupy habitats such as the gut, skin and vagina. Much of this diversity remains unexplained, although diet, environment, host genetics and ecology of human-associated m and set of distinct, clinically re signature microbes to vary widel individuals. The project enco configurations occupied by th stable among individuals despit the strongest associations of bot of structural and functional conf characterization of the epidemi

Human Microbiome Project





→ Specific communities in different habitats

Nature 2012

Metagenomics

University of East Anglia

Public data resources - What for?

- Initial analysis of your metagenomes (gene finding, annotation, etc.)
- Comparison of your metagenome with others (e.g. same environment)
- Mining for genes of interest in other metagenomes

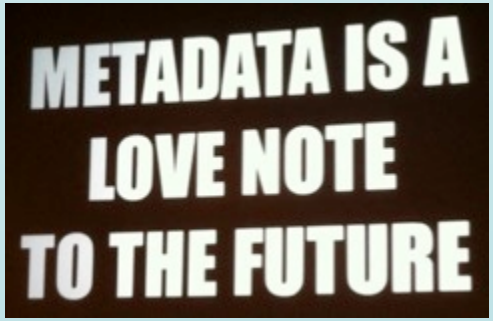
Best of all: it‘s free!

Metagenomics - Public data resources

University of East Anglia

Metadata

- **Essential**: what, where, who, how and when?
- Metadata = Context → Context enables connections
- Without metadata, raw data have very limited usefulness!



- **MIxS** = Minimum information about any (x) sequence
- Developed by the Genomic Standards Consortium (GSC)

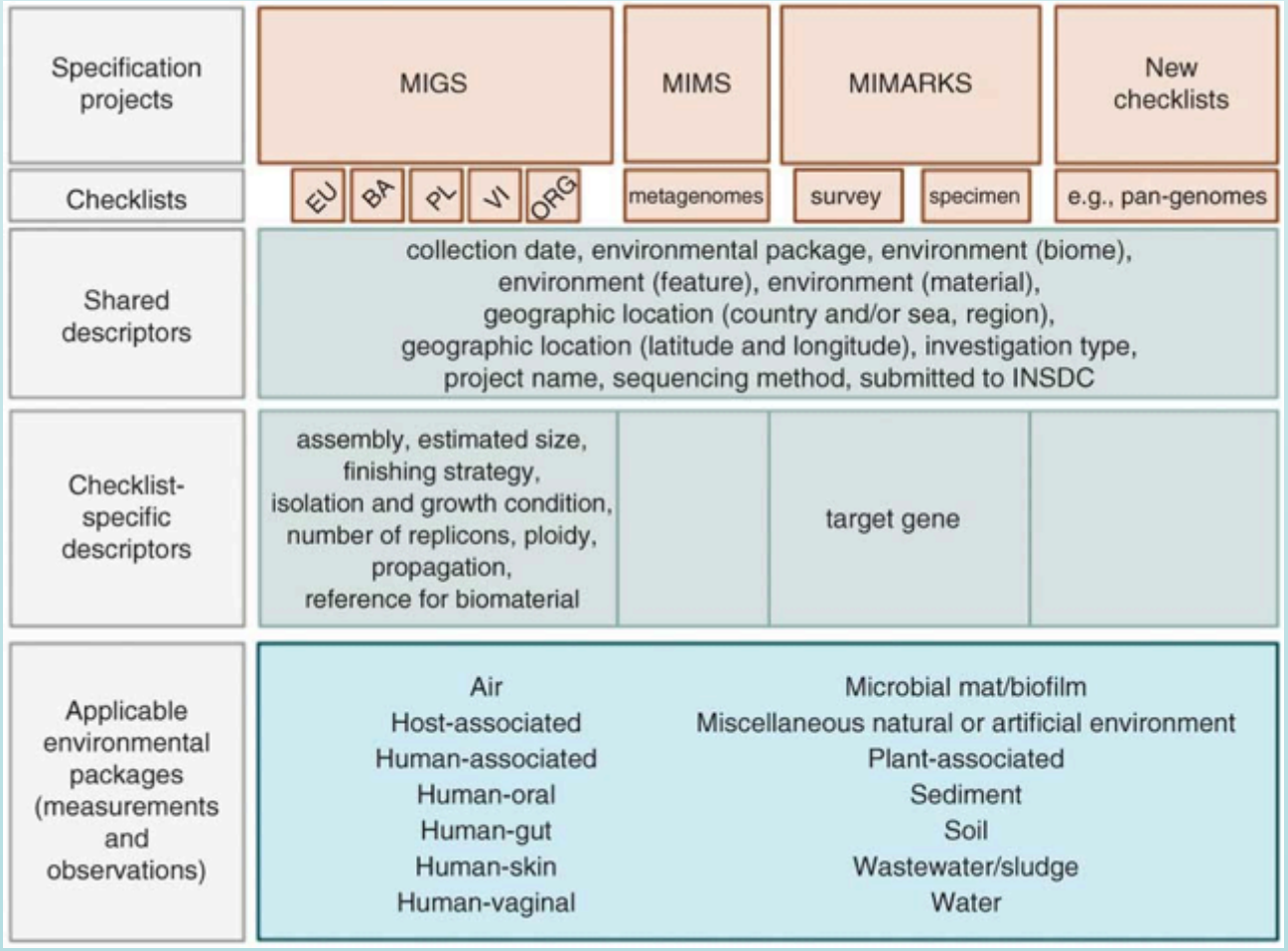
http://twitpic.com/6ry6ar

Metagenomics - Metadata

University of East Anglia

MIxS

- **MIMS** = Minimal Information about a Metagenomic Sequence



Specification projects	MIGS	MIMS	MIMARKS	New checklists
Checklists	EU, BA, PL, VI, ORG	metagenomes	survey, specimen	e.g., pan-genomes
Shared descriptors	collection date, environmental package, environment (biome), environment (feature), environment (material), geographic location (country and/or sea, region), geographic location (latitude and longitude), investigation type, project name, sequencing method, submitted to INSDC			
Checklist-specific descriptors	assembly, estimated size, finishing strategy, isolation and growth condition, number of replicons, ploidy, propagation, reference for biomaterial		target gene	
Applicable environmental packages (measurements and observations)	Air, Host-associated, Human-associated, Human-oral, Human-gut, Human-skin, Human-vaginal		Microbial mat/biofilm, Miscellaneous natural or artificial environment, Plant-associated, Sediment, Soil, Wastewater/sludge, Water	

Yilmaz *et al*. 2011

Metagenomics - Public data resources

University of East Anglia

rRNA sequence databases

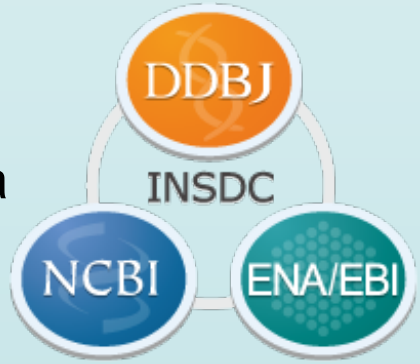
- **SILVA rRNA Database Project**: www.arb-silva.de
- **RDP** Ribosomal Database Project: http://rdp.cme.msu.edu
- **Greengenes**: http://greengenes.lbl.gov/

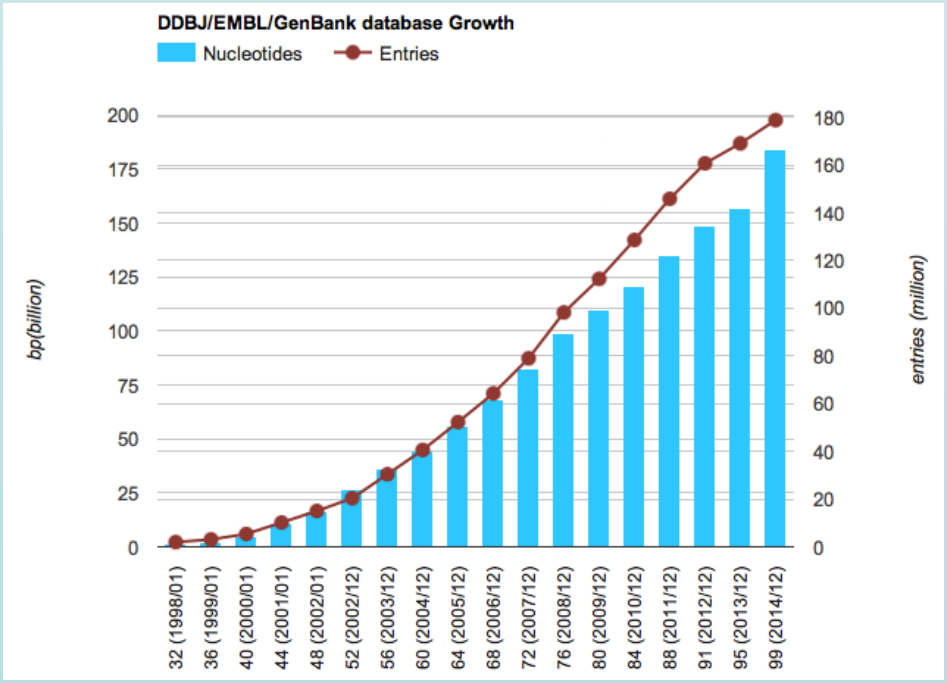
Metagenomics - Public data resources

University of East Anglia

International Nucleotide Sequence Database Collaboration

- **NCBI Genbank**: www.ncbi.nlm.nih.gov/genbank/
- **ENA** European Nucleotide Archive: www.ebi.ac.uk/ena
- **DDBJ** DNA Databank of Japan: www.ddbj.nig.ac.jp





Data exchange on daily basis
→ contain the same data!

www.ddbj.nig.ac.jp/breakdown_stats/dbgrowth-e.html#graph1

Metagenomics - Public data resources

University of East Anglia

Genbank & BLAST

- National Centre for Biotechnology Information (NCBI), established 1982
- Release 209 (15/08/15):

  187,066,846 sequence entries
  199,823,644,287 nucleotides

- **Functions:**

  Search & Browse (text and sequence search)
  Download
  Submit & Update
  Also stand-alone, command line version → blast+

  → Example: Global Ocean Sampling

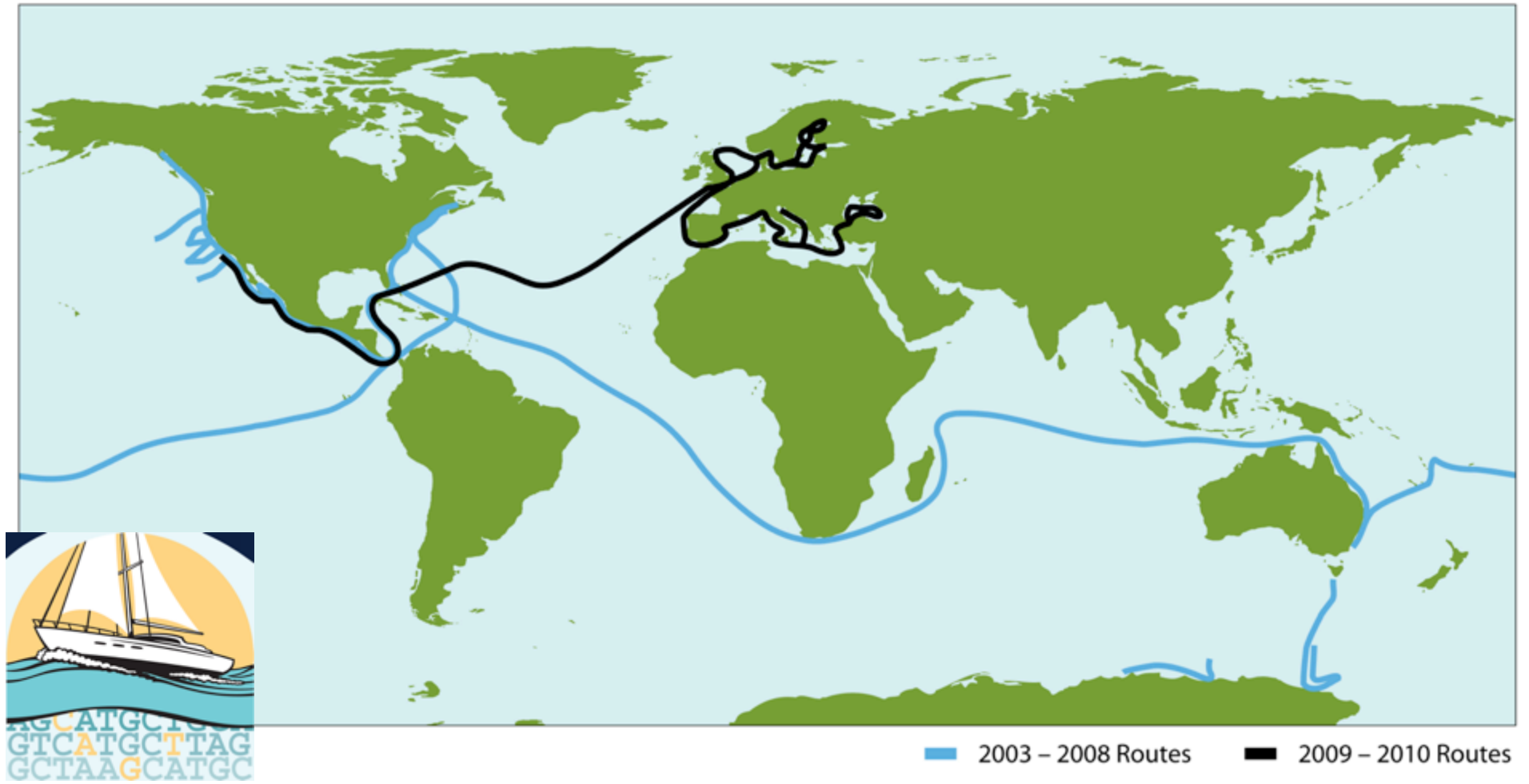
www.ncbi.nlm.nih.gov/genbank/ & http://blast.ncbi.nlm.nih.gov/Blast.cgi

Metagenomics - Public data resources

University of East Anglia

Global Ocean Sampling (GOS)

(e.g. Rusch *et* al./Yooseph *et al.*, *PLOS Biology* 2007)



www.jcvi.org/cms/research/projects/gos/

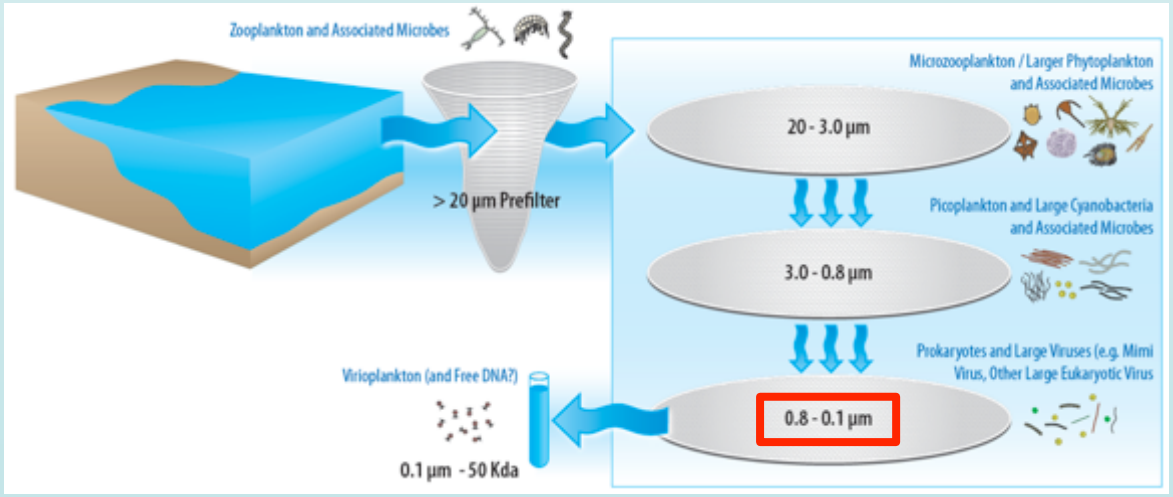
Metagenomics - Public data resources

University of East Anglia

Global Ocean Sampling (GOS)

(e.g. Rusch *et* al./Yooseph *et al.*, *PLOS Biology* 2007)

Water sampling apparatus



Sampling

- > 200 sampling points
- Sequenced with shotgun sequencing
- 12.6 Gb of sequence data

How to access this data? → e.g. NCBI

www.jcvi.org/cms/research/projects/gos/

Metagenomics - Public data resources

University of East Anglia

Genbank & BLAST - Example

- NCBI/Genbank: http://www.ncbi.nlm.nih.gov/genbank/
- GOS data→ Search in All Databases for “Global Ocean Sampling”
- Browse through (and/or download) samples of interest
- Biosamples → “MIMS Environmental/Metagenome…”
- http://www.ncbi.nlm.nih.gov/biosample/2954345
- Look at WGS project code → AACY (needed for blast)

- NCBI/BLAST: http://blast.ncbi.nlm.nih.gov/Blast.cgi
- Blast dmdA (DMSP demethylase) gene against GOS
- http://www.ncbi.nlm.nih.gov/nuccore/NC_003911.12
- blastn → WGS → AACY (“Marine metagenome, whole genome shotgun sequencing project”)

Metagenomics - Public data resources

University of East Anglia

European Nucleotide Archive (ENA)

ENA
European Nucleotide Archive

- European Bioinformatics Institute (EBI), established 1982
- Release 125 (01/06/15):

  629,681,877 sequence entries
  1,401,669,271,501 nucleotides
  → 3.2 TB of data

- **Functions:**

  Search & Browse (text and sequence search)
  Download
  Submit & Update

→ Example: “English Channel metagenome”

www.ebi.ac.uk/ena

Metagenomics - Public data resources

University of East Anglia

ENA - Example

- ENA: http://www.ebi.ac.uk/ena
- Search for “English channel L4 metagenome”
- http://www.ebi.ac.uk/ena/data/view/SRP001108
- Description of project and possibility to download samples

Metagenomics - Public data resources

University of East Anglia

DNA Databank of Japan (DDBJ)

- National Institute of Genetics (NIG), established 1986
- Release 102 (09/2015):

  187,785,897 sequence entries
  200,654,335,022 nucleotides

- **Functions:**

  Data submission
  Search/Analysis (e.g. BLAST, ClustalW, annotation)
  Download
  Supercomputer

www.ddbj.nig.ac.jp

Metagenomics - Public data resources

University of East Anglia

Ensembl

- Joint project between EMBL-EBI and the Wellcome Trust Sanger Institute
- Launched 1999 for Human Genome Project
- Only eukaryotic genomes, no metagenomes
- BUT extended to **Ensembl Genomes**

  → Including Ensembl Protists, Bacteria, Fungi, Metazoa, Plants

  → But also no metagenomes

www.ensembl.org & http://ensemblgenomes.org

Metagenomics - Public data resources

University of East Anglia

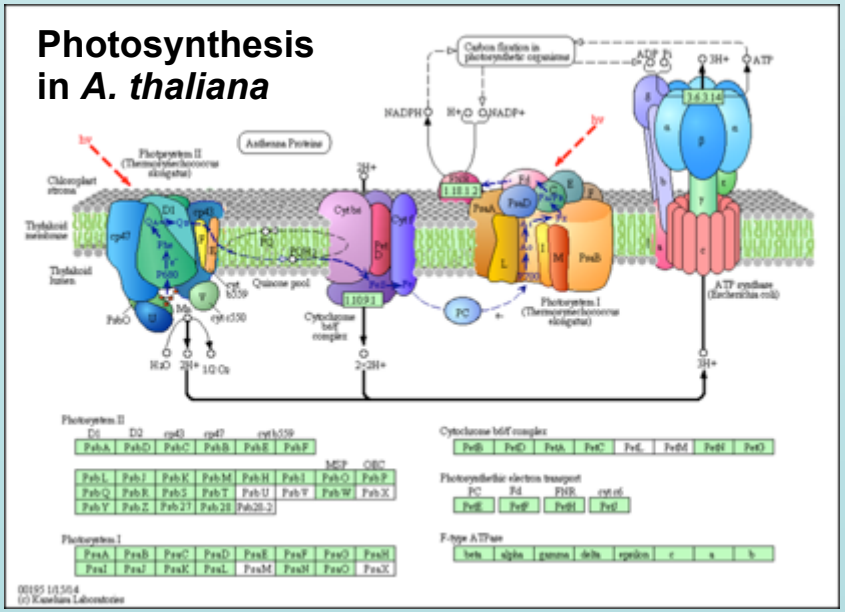
Kyoto Encyclopedia of Genes and Genomes (KEGG)

- Kyoto University, Japan (launched in 1995)
- PATHWAY maps (13/10/15) → 478 reference maps (414,789 total)
- **Functions:**

  KEGG PATHWAY database (e.g. metabolism, replication)

  Pathway prediction (KEGG maps)

  Comparison between genomes/ metagenomes

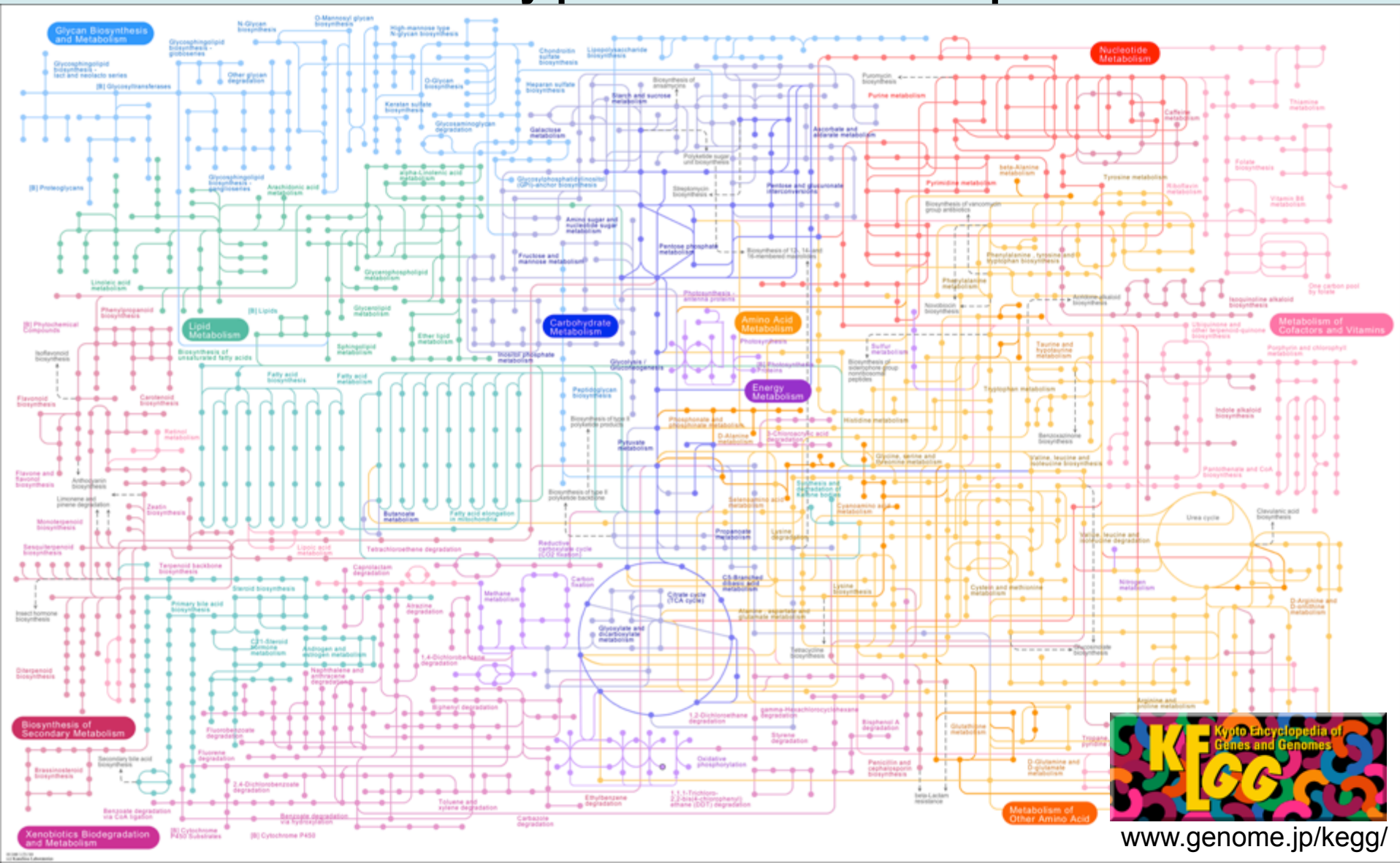


www.genome.jp/kegg/ & www.kegg.jp/kegg/mapper.html

Metagenomics - Public data resources

University of East Anglia

Pathway prediction: KEGG maps



www.genome.jp/kegg/

Metagenomics - Public data resources

University of East Anglia

Protein annotation databases

- **UniProt** Universal Protein Resource:
  www.uniprot.org

- **InterPro** Protein Sequence Analysis & Classification:
  www.ebi.ac.uk/interpro/

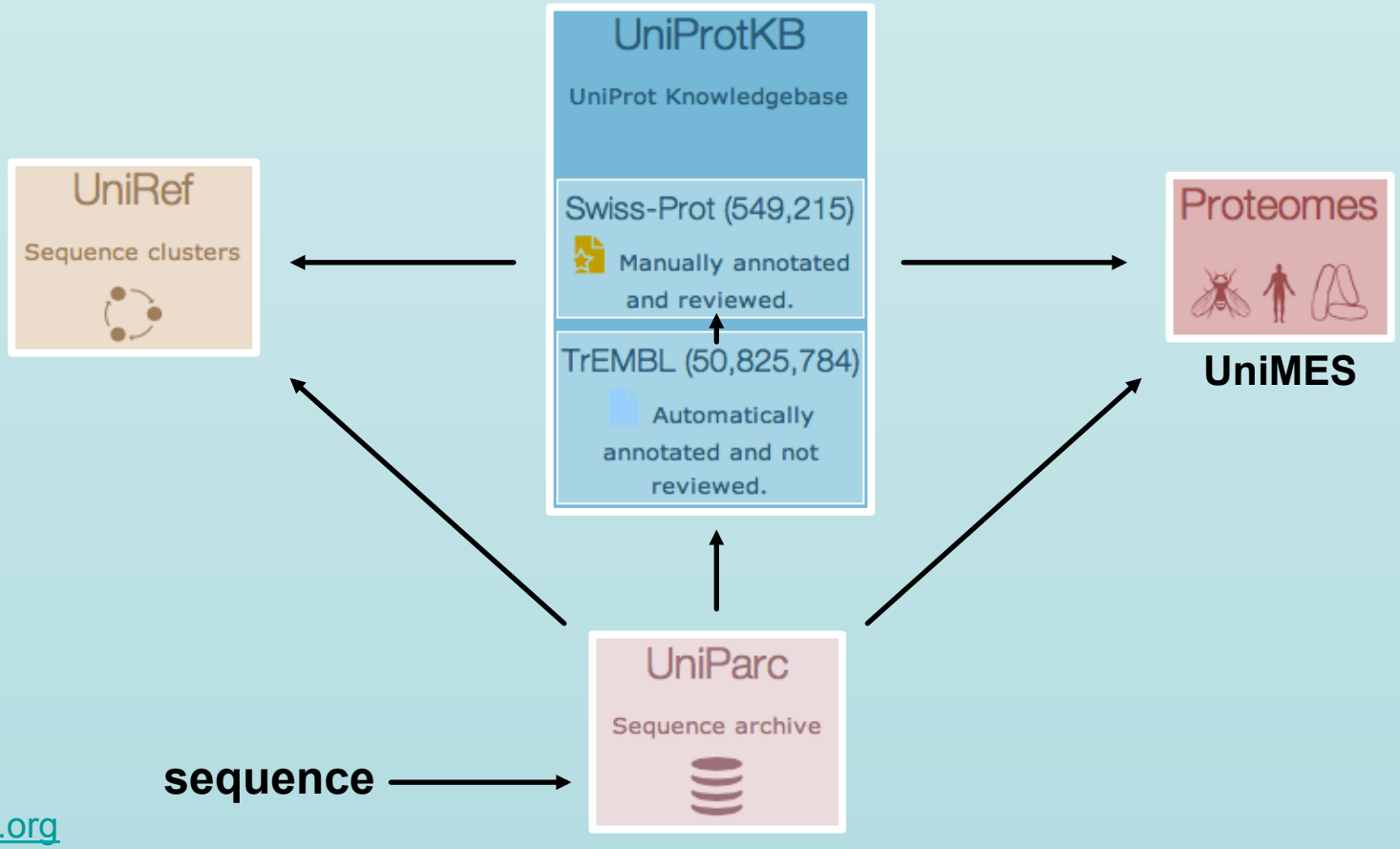
- **RCSB PDB** Protein Data Bank (crystallographic database for 3-dimensional structural data)
  www.rcsb.org

Metagenomics - Public data resources

University of East Anglia

UniProt

- **EBI** European Bioinformatics Institute, UK
- **SIB** Swiss Institute of Bioinformatics, Switzerland
- **PIR** Protein Information Resource, US



www.uniprot.org

Metagenomics - Public data resources

University of East Anglia

UniProt

- **EBI** European Bioinformatics Institute, UK
- **SIB** Swiss Institute of Bioinformatics, Switzerland
- **PIR** Protein Information Resource, US

- **Functions:**

Text search

BLAST

Sequence alignments

Retrieve/ID mapping

→ Example: BLAST dmdA sequence

www.uniprot.org

Metagenomics - Public data resources

University of East Anglia

UniProt - Example

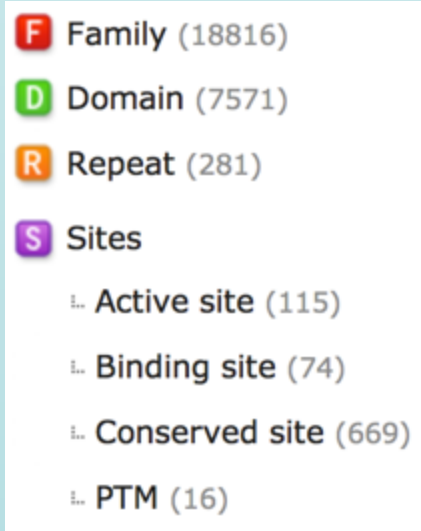
- UniProt: www.uniprot.org
- BLAST → paste dmdA (DMSP lyase) gene sequence → Run BLAST
- (Job identifier: B2015101415CHQOZ84V)
- Select best hit, look at alignment

Metagenomics - Public data resources

University of East Anglia

InterPro

- EBI-EMBL (launched 1999), linked to UniProt databases
- Diagnostic signatures consisting of models (prot. families, domains, sites)
- Release 53.0 (23/06/15): 27,542 entries



- **Functions:**

  Search/Analysis of protein sequences
  (of up to 40k amino acids)
  Classification into families
  Predicting domains & important sites
  InterProScan

www.ebi.ac.uk/interpro/ & www.ebi.ac.uk/interpro/search/sequence-search

Metagenomics - Public data resources

University of East Anglia

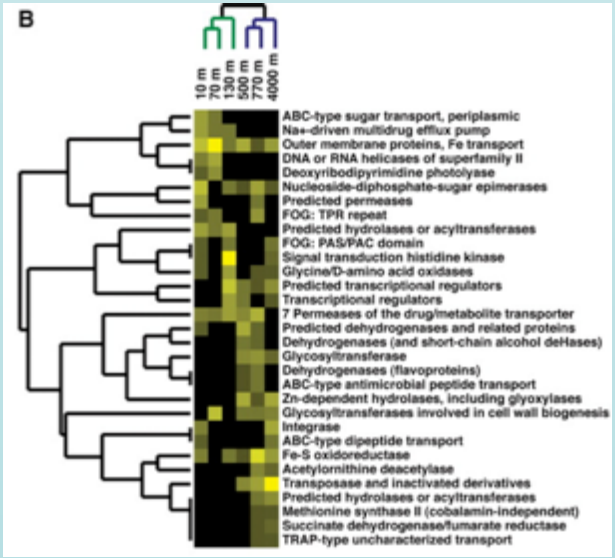
Function & Pathway prediction

→ Grouping of genes/proteins in metagenome by:

Orthology

conserved sequence features

COG, KO, FIGfam, …



DeLong, 2006

Structure

similar protein domains (e.g. catalytic)

Pfam, TIGRfam, …

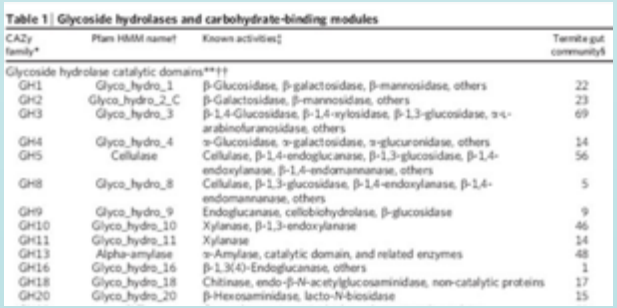


Table 1 | Glycoside hydrolases and carbohydrate-binding modules

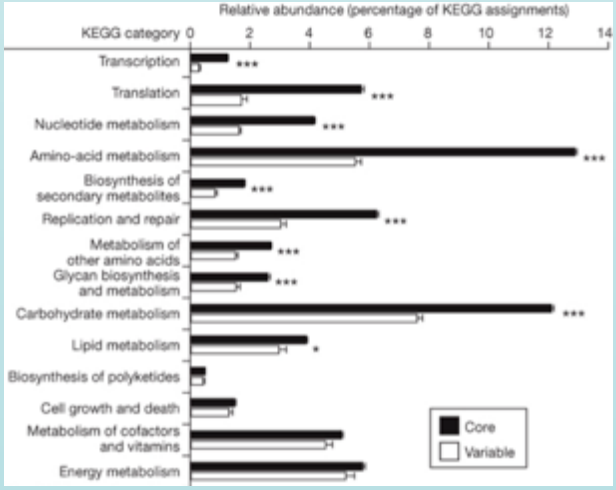
CAZy family*	Pfam HMM name†	Known activities‡	Termite gut community§
Glycoside hydrolase catalytic domains**††			
GH1	Glyco_hydro_1	β-Glucosidase, β-galactosidase, β-mannosidase, others	22
GH2	Glyco_hydro_2_C	β-Galactosidase, β-mannosidase, others	23
GH3	Glyco_hydro_3	β-1,4-Glucosidase, β-1,4-xylosidase, β-1,3-glucosidase, α-L-arabinofuranosidase, others	69
GH4	Glyco_hydro_4	α-Glucosidase, α-galactosidase, α-glucuronidase, others	14
GH5	Cellulase	Cellulase, β-1,4-endoglucanase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	56
GH8	Glyco_hydro_8	Cellulase, β-1,3-glucosidase, β-1,4-endoxylanase, β-1,4-endomannanase, others	5
GH9	Glyco_hydro_9	Endoglucanase, cellobiohydrolase, β-glucosidase	9
GH10	Glyco_hydro_10	Xylanase, β-1,3-endoxylanase	46
GH11	Glyco_hydro_11	Xylanase	14
GH13	Alpha-amylase	α-Amylase, catalytic domain, and related enzymes	48
GH16	Glyco_hydro_16	β-1,3(4)-Endoglucanase, others	1
GH18	Glyco_hydro_18	Chitinase, endo-β-N-acetylglucosaminidase, non-catalytic proteins	17
GH20	Glyco_hydro_20	β-Hexosaminidase, lacto-N-biosidase	15

Warnecke, 2007

Biological roles

pathway and process involvement

GO, KEGG, MetaCyc, SEED, …

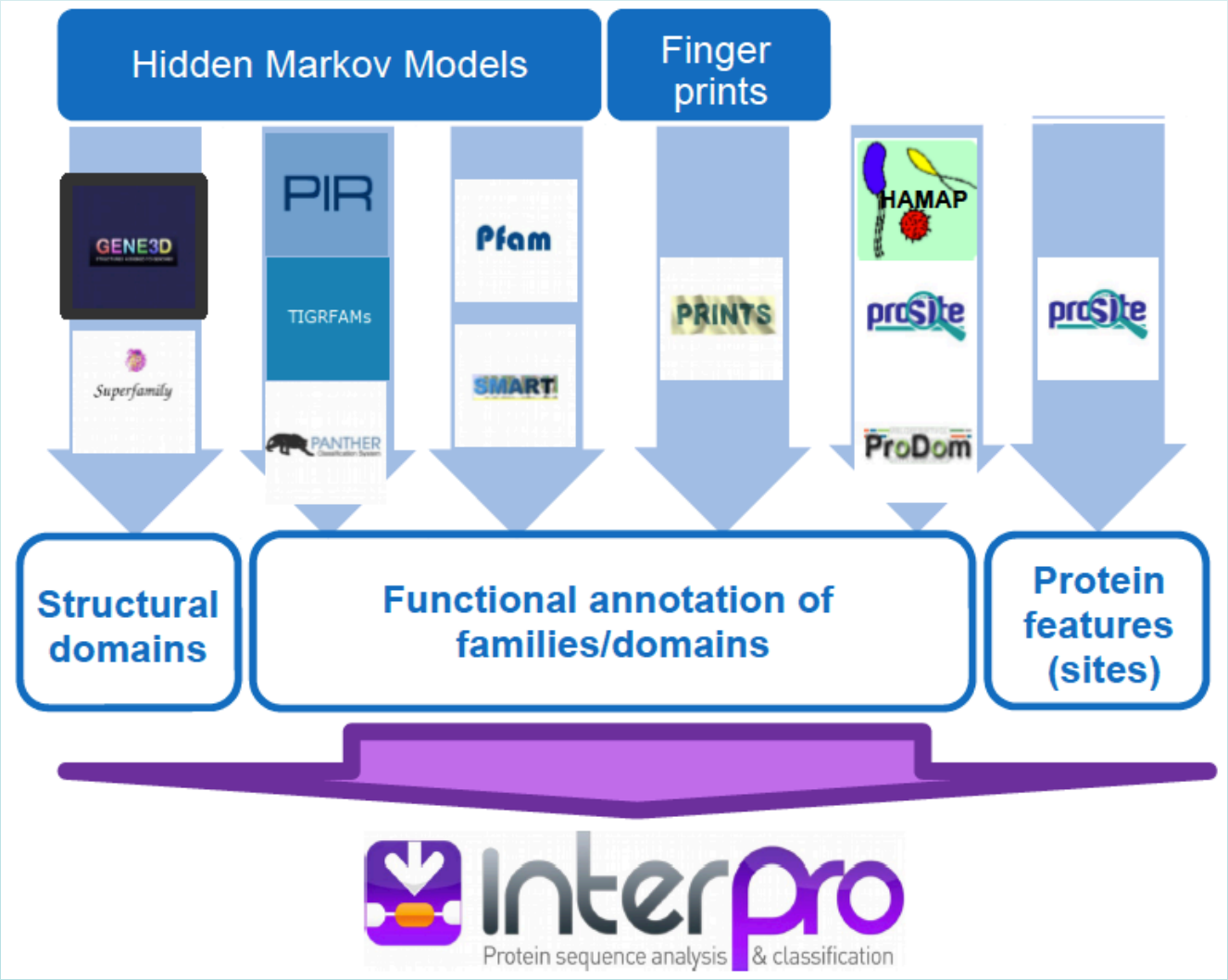


Turnbaugh, 2009

Metagenomics - Public data resources

University of East Anglia

InterPro



→ Example: “Example protein sequence”

Metagenomics - Public data resources

University of East Anglia

InterProScan - Example

- InterPro: http://www.ebi.ac.uk/interpro
- Search → Example protein sequence

- http://www.ebi.ac.uk/interpro/sequencesearch/iprscan5-S20151014-115127-0023-86869032-oy

Metagenomics - Public data resources

University of East Anglia

Data and analysis platforms

- **IMG/M** Integrated Microbial Genomes with Microbiome Samples:
  https://img.jgi.doe.gov/cgi-bin/mer/main.cgi
- **EBI Metagenomics**: www.ebi.ac.uk/metagenomics/
- **MG-RAST**: http://metagenomics.anl.gov
- **MicroScope** Micobial Genome Annotation & Analysis Platform:
  www.genoscope.cns.fr/agc/microscope/home/
- **iPlant** Discovery Environment (former CAMERA data):
  https://preview.iplantcollaborative.org/de/
  & iMicrobe: http://data.imicrobe.us

Metagenomics - Public data resources

University of East Anglia

Integrated Microbial Genomes (IMG)

JGI IMG INTEGRATED MICROBIAL GENOMES

- U.S. Department of Energy (DOE) - Joint Genome Institute (JGI)
- Integrated Microbial Genomes with Microbiome Samples (**IMG/M**)
- IMG content (21/09/15):

	Datasets
Bacteria	30326
Archaea	596
Eukarya	220
Plasmids	1196
Viruses	3905
Genome Fragments	1192
Metagenome	4210
Total Datasets	41645

- **Functions:**

  Gene, genome, function search & download
  Analysis cart
  Genome comparison
  etc.

→ Example: Global Ocean Sampling

https://img.jgi.doe.gov/cgi-bin/mer/main.cgi

Metagenomics - Public data resources

University of East Anglia

IMG/M- Example

- IGM/M: https://img.jgi.doe.gov/cgi-bin/mer/main.cgi
- Find GOS data: Find Genomes → Genome Search → study name → "global ocean sampling"→ search (56 metagenomes)
- You can also search by metadata categories, e.g. altitude, ecosystem!
- Select All → Add Selected to Genome Cart → Analysis cart

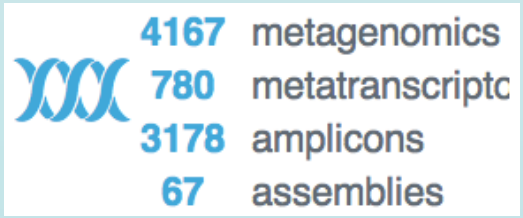
- Find Genes → BLAST (blastn, e-value 1e-5) → paste dmdA gene
- Select metagenome (Domain "Genome Cart") → select up to 20
- Run BLAST → Select All (9) → Add Selected to Scaffold Cart
- You can also export and re-import genome/gene carts etc.!

Metagenomics - Public data resources

University of East Anglia

EBI Metagenomics

- EBI-EMBL
- Projects (12/10/15):



- **Functions:**

  Search and access public metagenomic or metatranscriptomic projects and samples

  Download raw data, visualise and analyse results

  InterProScan

  → Example: Global Ocean sampling taxonomy and functions

www.ebi.ac.uk/metagenomics/

Metagenomics - Public data resources

University of East Anglia

EBI - Example

- EBI: https://www.ebi.ac.uk/metagenomics/
- Projects → Text search “global ocean sampling”
- Project name “Gobal Ocean Sampling Expedition”
- https://www.ebi.ac.uk/metagenomics/projects/SRP003580
- Select first sample
- Look at Overview (metadata), and taxonomic and functional analysis

Metagenomics - Public data resources

University of East Anglia

MG-RAST

MG-RAST
metagenomics analysis server

- Metagenomic Rapid Annotations using Subsystems Technology
- Maintained by Argonne National Laboratory, U.S.
- Metagenomes (12/10/15):

# of metagenomes	212,523
# base pairs	86.02 Tbp
# of sequences	684.49 billion
# of public metagenomes	30,104

- **Functions:**

  Automated annotation and analysis for prokaryotic metagenomic samples

  Comparison & download

  Con: No proper BLAST function, very slow…

http://metagenomics.anl.gov

Metagenomics - Public data resources

University of East Anglia

MicroScope - MaGe

- Microbial Genome Annotation & Analysis Platform
- French National Sequencing Center (Genoscope), by LABGeM
- **Functions:**
  - Microbial comparative genome analysis
  - Manual functional annotation
  - BLAST searches
  - Comparison tools (e.g. pan/core-metagenome)
  - Con: Only for very small metagenomes (< 20 Mb of contigs per bin)
  - Con: Upload/Analysis takes quite long (~ 6-8 weeks)

www.genoscope.cns.fr/agc/microscope/home/

Metagenomics - Public data resources

University of East Anglia

iPlant – iMicrobe

- iPlant: virtual organisation, US National Science Foundation (NSF) funded
- iMicrobe: University of Arizona (Gordon & Betty Moore Foundation funds)
- Former CAMERA data

- **Functions:**

  Discovery environment → tools for data analysis (etc. BLAST)

  Pro: Access to data previously on CAMERA

  Con: Not very user friendly, a lot of functions don't seem to work yet…

www.iplantcollaborative.org & http://data.imicrobe.us

Metagenomics - Public data resources

University of East Anglia

Human Microbiome Project (HMP)

- launched in 2008, $170 million budget
- 200 scientists at 80 institutions
- sequenced microbiota from 4,788 samples from 242 healthy individuals (129 m, 113 w)
- samples from 18 (w) & 15 (m) body habitats

→ HMP project catalog (hmpdacc.org/catalog/)

→ NCBI: Bioproject ID 43021

→ IMG: IMG/M HMP

www.hmpdacc.org

Metagenomics - Public data resources

University of East Anglia

Data and analysis platforms for viromes

- **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

Metagenomics

University of East Anglia

Some tools for metagenomics to try:

- **VizBin** Reference-independent visualization and binning:
  http://claczny.github.io/VizBin/

- **GroopM & CheckM** Recovery of genomes from metagenomes & quality analyses:
  http://ecogenomics.github.io/GroopM/
  http://ecogenomics.github.io/CheckM/

- **Anvi'o** Analysis and visualization platform for 'omics data:
  http://merenlab.org/projects/anvio/

- **MetAMOS** Modular framework for assembly, analysis and validation:
  http://metamos.readthedocs.org

- **MEGAN** Metagenome Analyzer:
  http://ab.inf.uni-tuebingen.de/software/megan5/
  → **see talk and hands-on tomorrow by Suparna Mitra!**