Introducing Microbiome Bioinformatics

Part 11.

Introducing sequence databases.

Recap: Aims

- **Microbiome analysis**
  - with particular regard to **sequence informatics concepts**
- "Top down" – putting analysis tools and resources in context
- No highly detailed technicalities (yet)
  - No instructions on how to run particular programs
- Why you are using the bioinformatics approaches you use; pros, cons; alternatives

Series of talks

- 10 so far
- Open ended… as long there is demand
- Expected to be every 2 weeks
  - Notwithstanding some larger gaps for various reasons…
  - all dates will be confirmed in advance
  - *Please refer to:* ***Bite-size bioinformatics mailing list***
    - *Contact* ***Mark Fernandes,*** *or me*
- Informal and flexible
  - Please interrupt and ask questions
  - **Suggestions for topics for further focus**
- Previous talks will be repeated, starting this Autumn

Topics, top-down

Introduction

Types of environments studied

Variety of microbiomes, and domains of Life

Aims of studies:

"Communities"

Functions

Aims of DNA/ RNA sequencing approaches

"barcoding"- use of marker genes

"shotgun" metagenomics

metatranscriptomics

16S rRNA

Other markers

aims of whole-metagenome sequencing

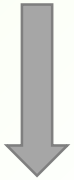
Dealing with the data: Informatics concepts

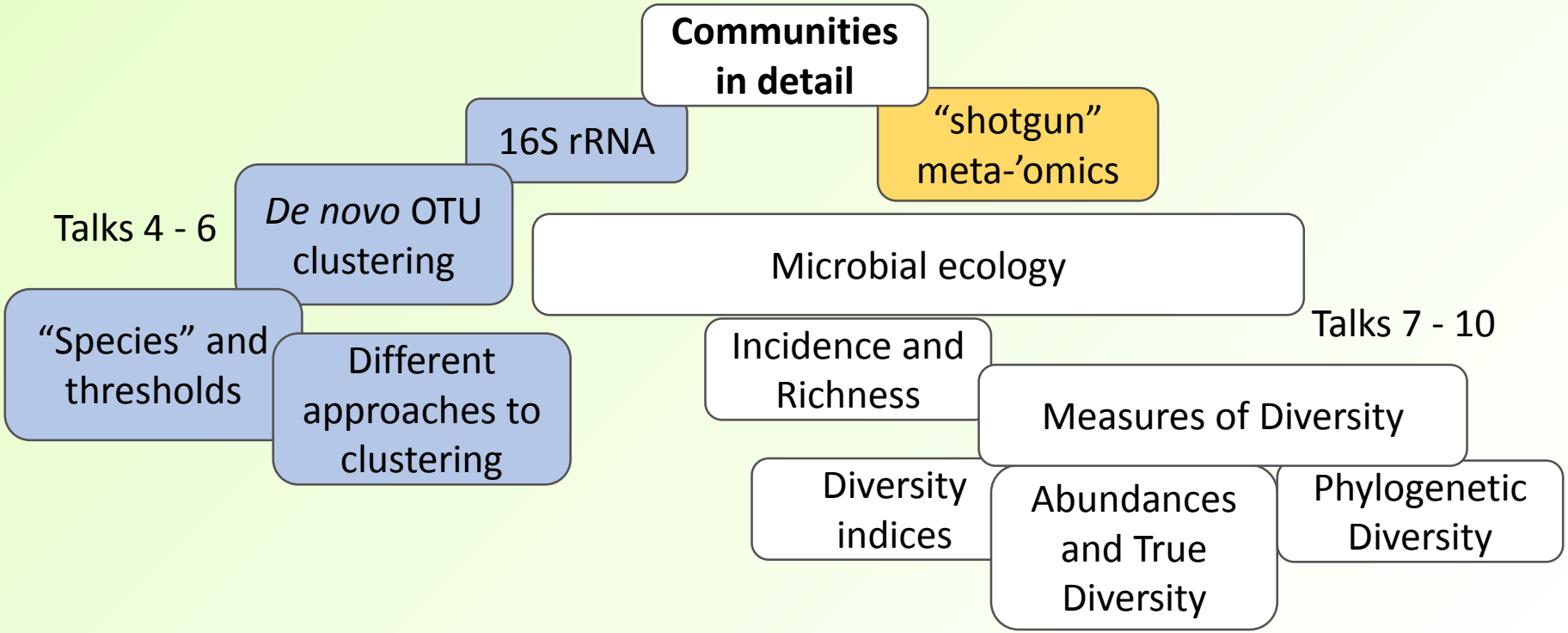
16S rRNA

"shotgun" meta-'omics

Communities in detail

Later: more technical details



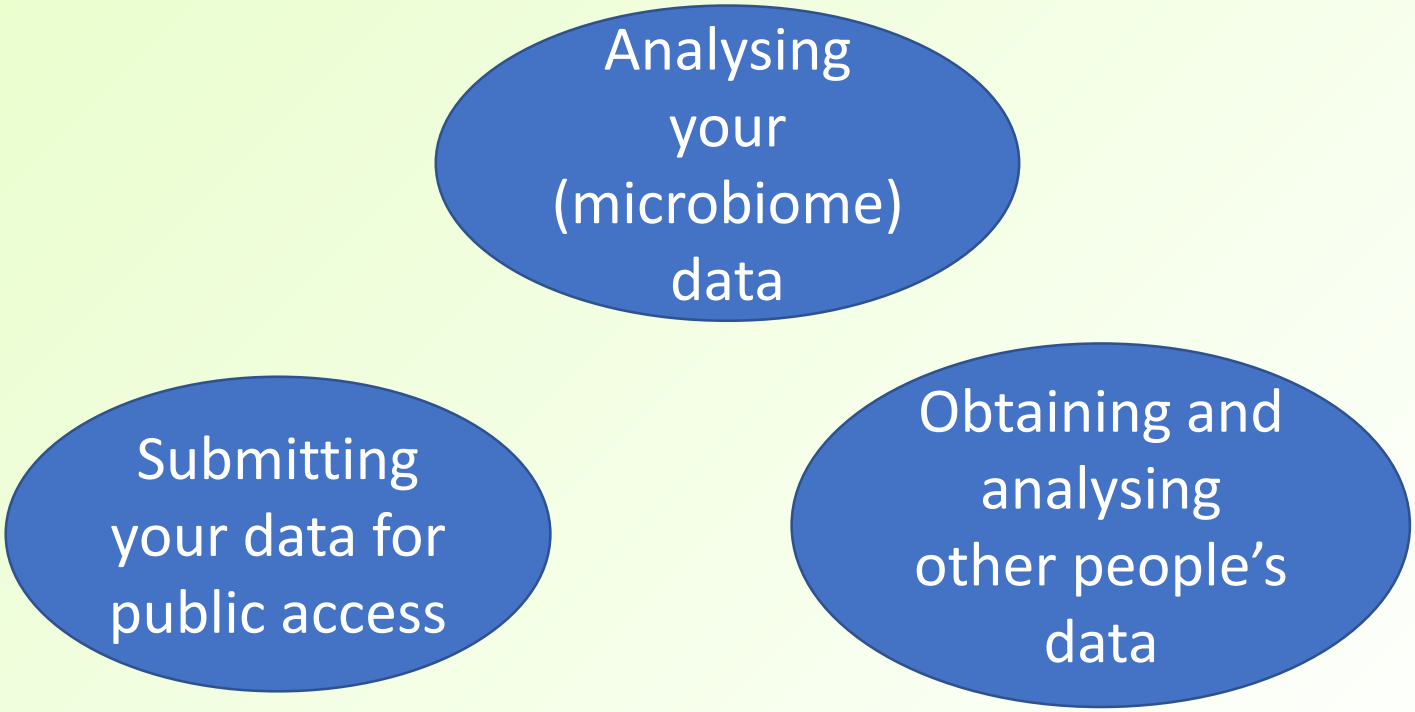

Communities in detail
16S rRNA
"shotgun" meta-'omics
Talks 4 - 6
De novo OTU clustering
Microbial ecology
"Species" and thresholds
Different approaches to clustering
Talks 7 - 10
Incidence and Richness
Measures of Diversity
Diversity indices
Abundances and True Diversity
Phylogenetic Diversity

Series of talks

Slideshows - http://ghfs1.quadram.ac.uk/ghfs/

- Part 1: 27/1/2017
  - "Biological and Experimental Stuff that a microbiome bioinformatician needs to know"
  - Overview of marker gene sequencing for community analysis
- Part 2: 10/2/2017
  - Overview of whole-metagenome sequencing
- Part 3: 24/2/2017
  - Focus on metatranscriptomics
- Part 4: 10/3/2017
  - Different bioinformatics approaches to processing 16S read data
- Part 5: 24/3/2017
  - *De novo* OTU clustering: sequence identities and how thresholds have been determined historically; relationships to taxonomic levels
- Part 6: 7/4/2017
  - The clustering problem: different approaches, and what can go wrong; the influence of amplification artefacts, sequencing errors and sequence lengths; computational OTUs versus species
- Part 7: 21/4/2017
  - Introducing microbial ecology: using observed abundances of OTUs (or species, or functions) to estimate the richness of the community (number of different OTUs, species etc)
- Part 8: 2/6/2017 – continuing microbial ecology: community diversity : diversity indices
- Part 9: 16/6/2017 – continuing microbial ecology: community diversity : true diversity
- Part 10: 28/7/2017 – concluding diversity (for now);
- Part 11: today – Introducing sequence databases

Sequence databases

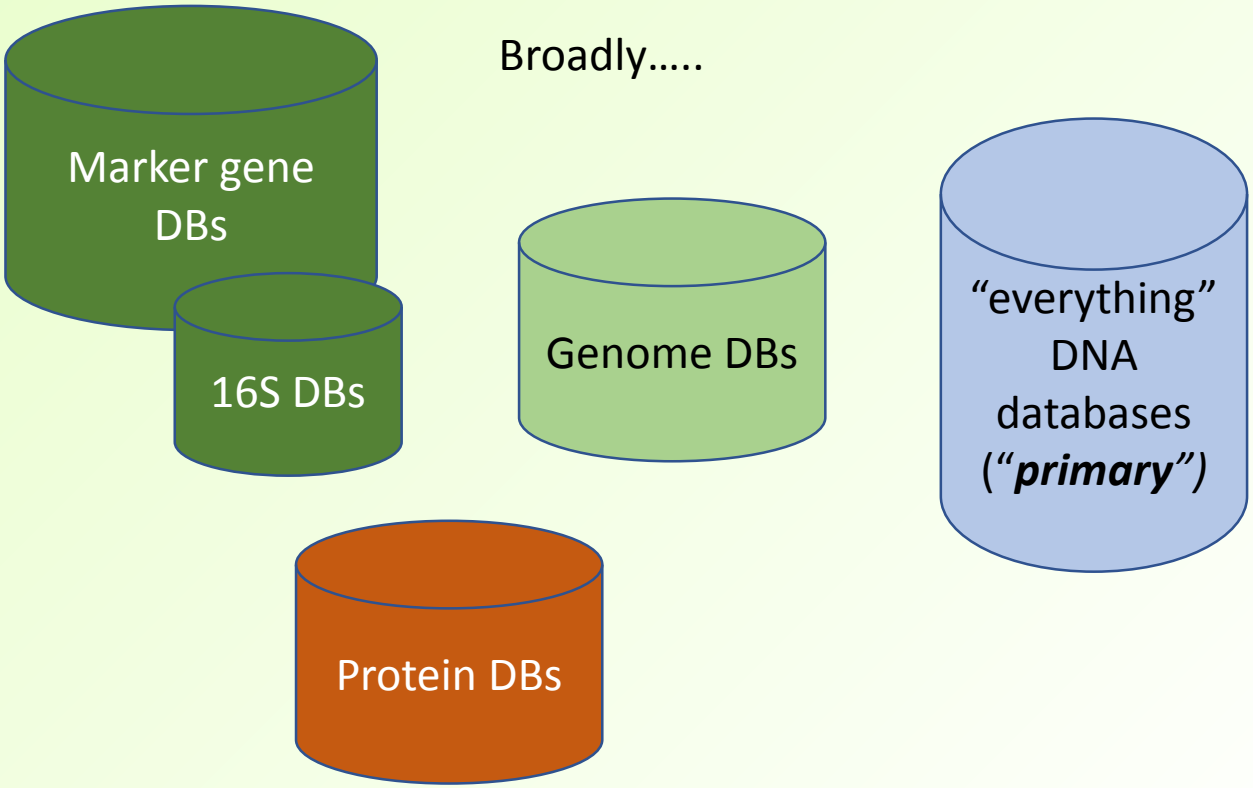


Sequence databases and microbiome analysis

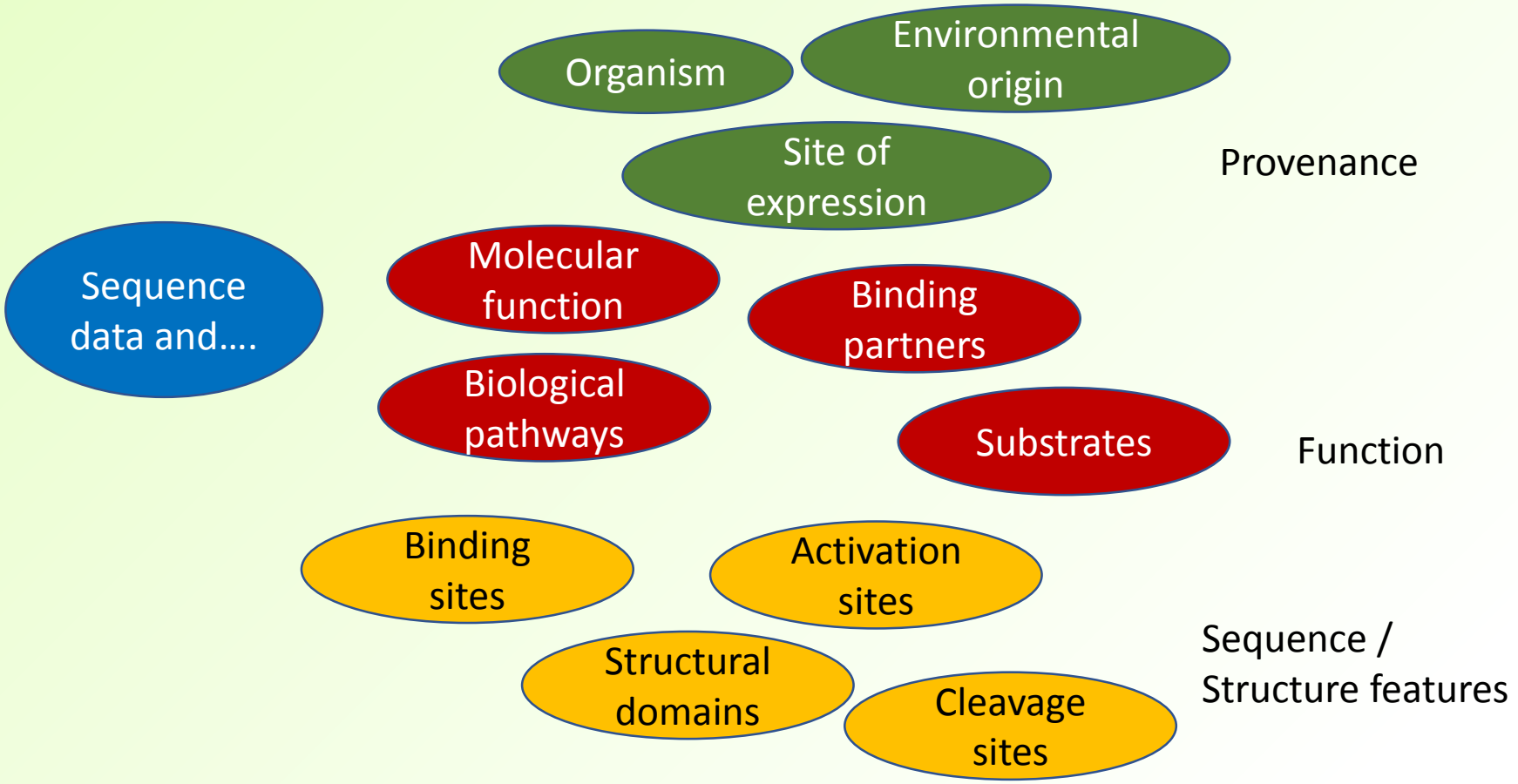
- Today – overview of potentially useful databases
  - Nucleic acid (nucleotide) sequence databases
  - Protein / peptide sequence databases
- Next time – principles of how these can be used to analyse 16S and whole-metagenome shotgun data

  (….part I)
  - In principle, you can usefully use any sequence database to find similar sequences to your sequence data
  - This direct approach is used in some methods of microbiome data analysis
  - But has pros/cons, and so various other methods can be used.There are also “indirect”methods. More next time.

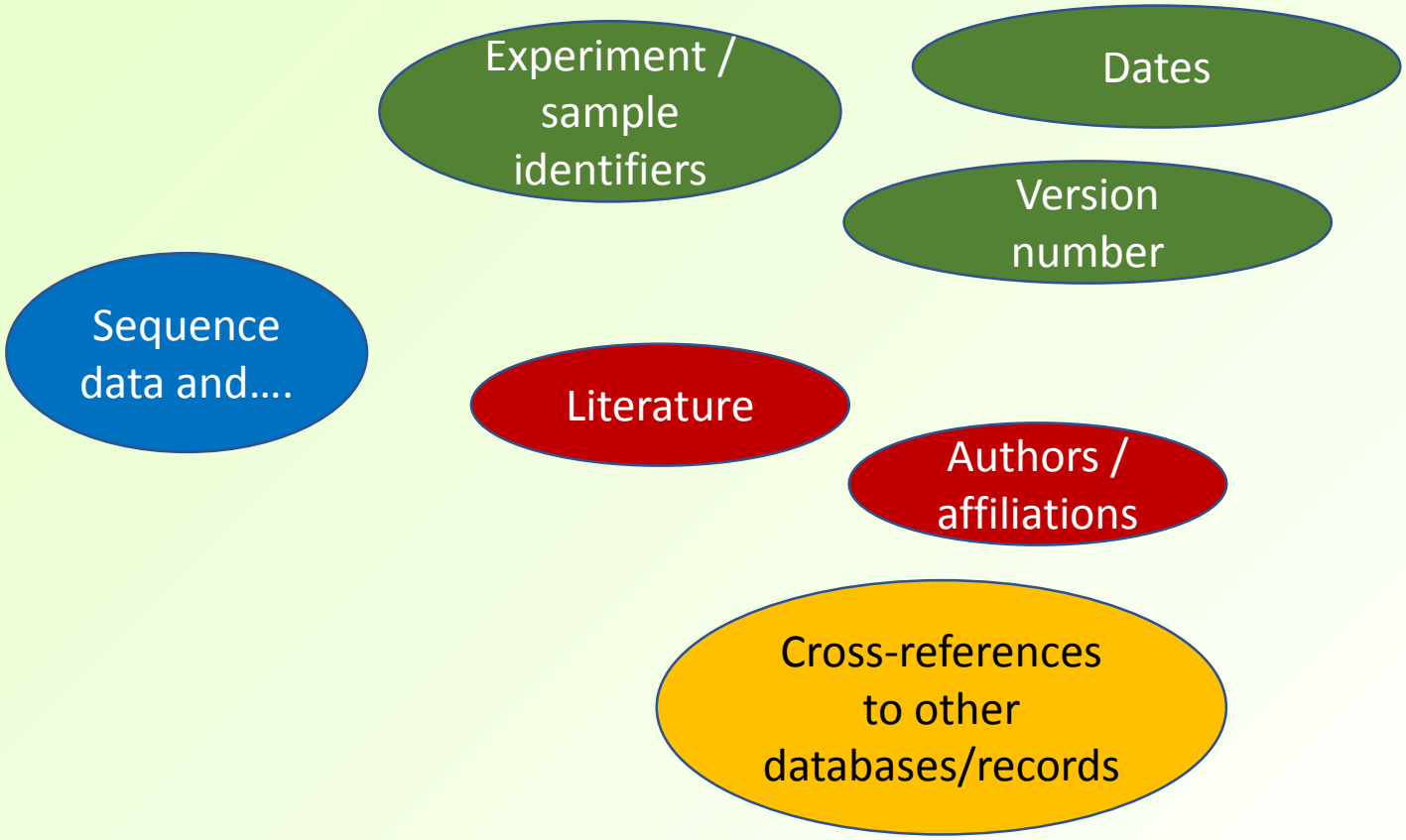
Sequence databases and microbiome analysis



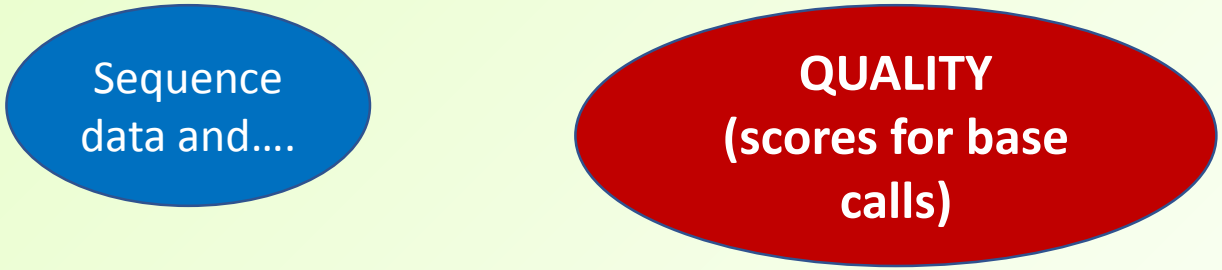
Annotations



“Metadata”


Experiment / sample identifiers
Dates
Version number
Sequence data and….
Literature
Authors / affiliations
Cross-references to other databases/records

Sequence **reads**



FASTQ format

General Database Concepts

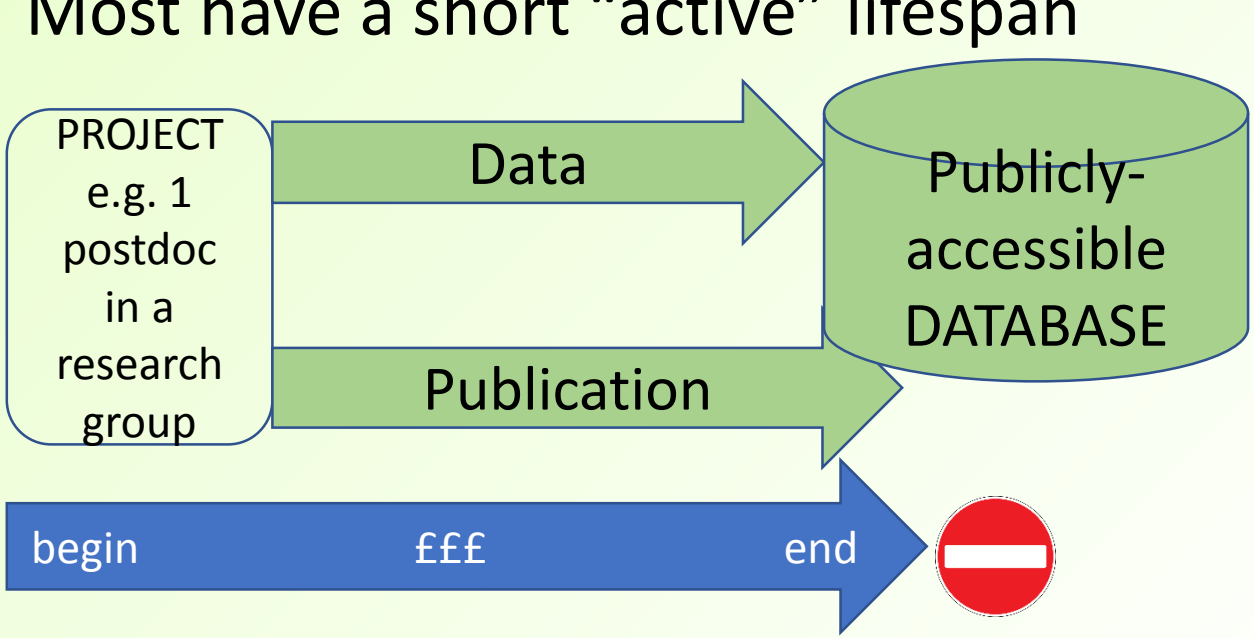
- Characteristics of any formalized database:
  - a collection of **records** (database '**entries**')
  - each record is uniquely identified by a **key** (a unique name/number)
    - in many databases, key is referred to as the "**Accession (number)**"
    - E.g. **'Q8X696**' (an example from the UniProt database)
    - a.k.a. "primary ID", "primary key"
  - a key is a special type (i.e. it is unique within the database) of **field**
  - other information about the record is held in other fields, e.g.
    - sequence
    - date
    - literature reference
    - functional annotation
    - etc

General Database Concepts

- Characteristics of any formalized database:
  - A given record may change over time:
    - be **updated** (same key, same record, different content in record)
    - be **retired** (deleted from the database)
  - Some records may be **merged into one** record (when the curators realize that two records represent the same thing)
  - Some databases will include **dates of records** (e.g. date of creation; most recent update; all updates)
  - Some databases make older versions of records available in an archive (including those now deleted)
- Some very large databases (e.g. ENA) consist of **regular full releases** (on a quarterly basis); the **Release-proper** is supplemented by an **Update** section, which changes daily and contains changes relative to the most recent Release.

Bioinformatics databases

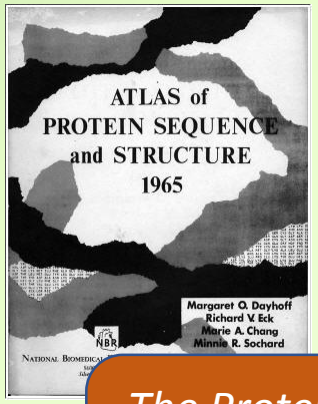
- Many thousands of them
- Many have a fairly short lifespan
- Most have a short “active” lifespan



NO LONGER MAINTAINED

- **Data not updated**
- **Problems not fixed**
- **“Zombie DB”**
- **or goes down forever**

Evolution of **Primary Protein Sequence Databases**



The Protein Sequence Database

Early 1960s
Dayhoff et al.
NBRF,
Washington D.C.

PIR, NBRF

MIPS collection, Munich

1990s

JIPID, Tokyo

The Protein Identification /Information Resource Protein Seq DB (PIR-PSD)

NBRF, 1984

Swiss-Prot

Bairoch et al.,
Univ. Geneva
1986

Swiss-Prot

TrEMBL

Bairoch, Apweiler
SIB Switzerland,
EBI, Hinxton

UNIPROT

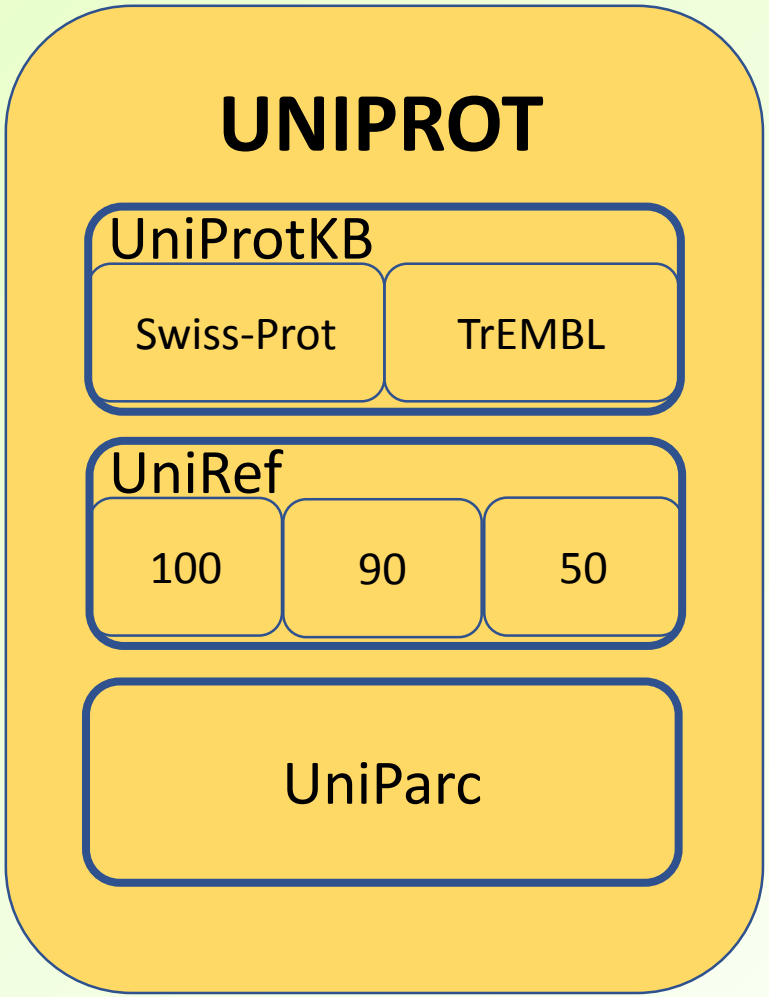
UniProtKB

UniRef

UniParc

2002

EBI / SIB / PIR collaboration



- Compartmentalised: different sections useful for different purposes
- Redundant and non-redundant sections
- Cross-references to other databases

Primary Sequence Databases

- Basically, role is to store all publicly-available DNA/RNA/protein sequences of all known life...
- "**Primary**" because they are the **principal** and **largest** databases; there aren't many of them
  - not the original sources of the data
- Include sequence data from many types of sources
- From many projects, large and small

Primary Sequence Databases

- Nucleic acid sequences:
- **ENA** (EMBL, Europe), **Genbank** (NCBI, USA), **DDBJ** (Japan)
  - Collaborate/duplicate (**INSDC**) – see http://www.insdc.org
- Protein sequences:
  - **UniProt**
    - Consortium: EMBL-EBI, SIB (Europe); PIR (USA)
  - **NCBI protein database** ('GenPept')
  - DDBJ Amino Acid Sequence DB (**DAD**)

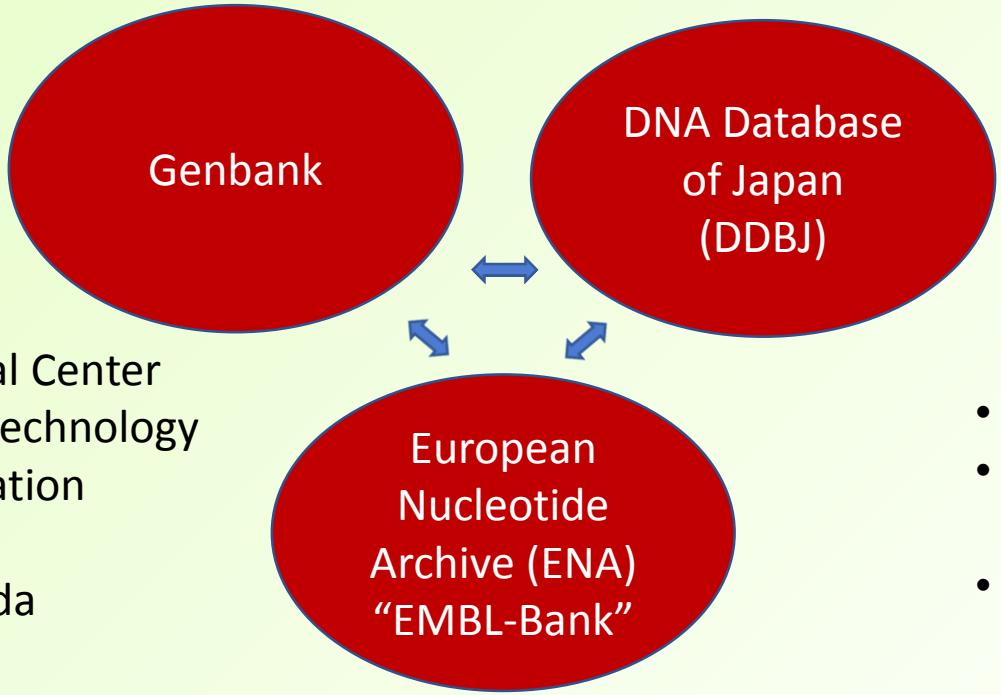
Primary Sequence Databases

- Include sequence data from many types of sources:
  - genome sequencing/resequencing
  - transcriptome sequencing
  - environmental samples (meta-'omes)
  - traditional low-throughput studies

Types of sequence in the DBs

- Proteins/peptides
- Whole chromosomes/pseudomolecules
- Plasmids
- Single genomic reads
- cDNA, ESTs
- various other types of sequence “tags”
- Constructs e.g. BACs, YACs etc
- Gene coding-region sequences
- Complete gene sequences
- Raw read data from all kinds of experiments
- genome-sequencing projects; marker-gene amplicons; metagenomes
- ...etc
- The primary databases consist of various DATA CLASSES
  - E.g. “**Standard**” sequences, i.e. assembled, annotated
  - See http://www.ebi.ac.uk/ena/about/formats

International Nucleotide Sequence Database Collaboration (INSDC)



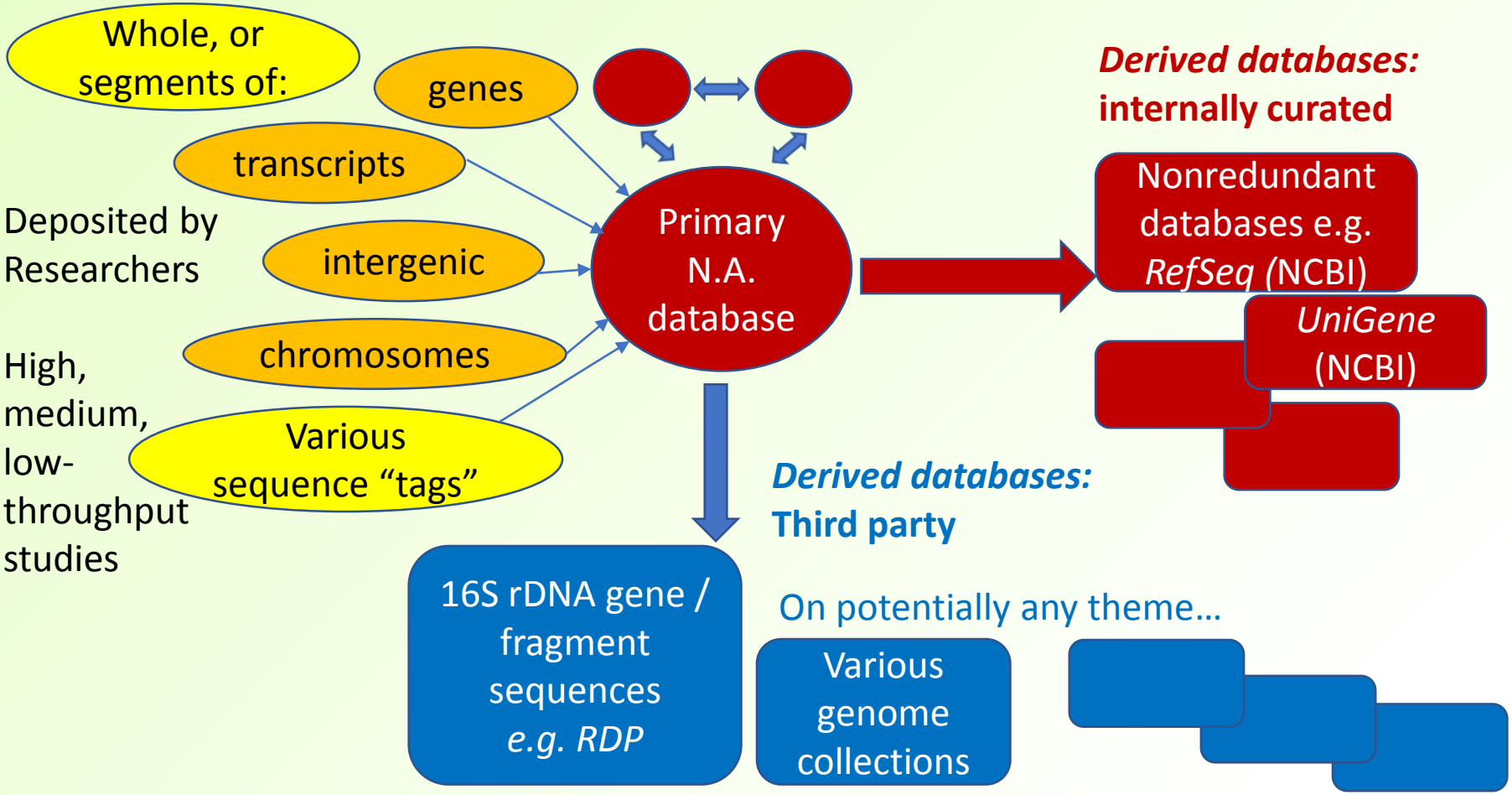
DDBJ Center, Natl. Inst. Genetics, Mishima

National Center for Biotechnology Information (NCBI), Bethesda

EMBL-European Bioinformatics Institute (EBI), Hinxton

- Same data
- synchronized daily
- Managed and presented in slightly different ways
- www.insdc.org

The ins and outs of primary sequence databases



What else is available from the primary database providers?

- Besides the annotated sequence records:
  - Sequence read data
    - **Sequence Read Archive (SRA)**: quality scores, in the form of FASTQ data
    - Trace Archive: equivalent for first-generation (Sanger) reads
  - Data describing **projects, experiments, samples**
    - BioProjects, BioSamples
- All the above are available at NCBI, DDBJ, EMBL-EBI
  - Again, presentation/modes of access at each site differ
  - Collectively, constitute ENA at EBI
  - (Organisation slightly different at NCBI, DDBJ)

Redundancy

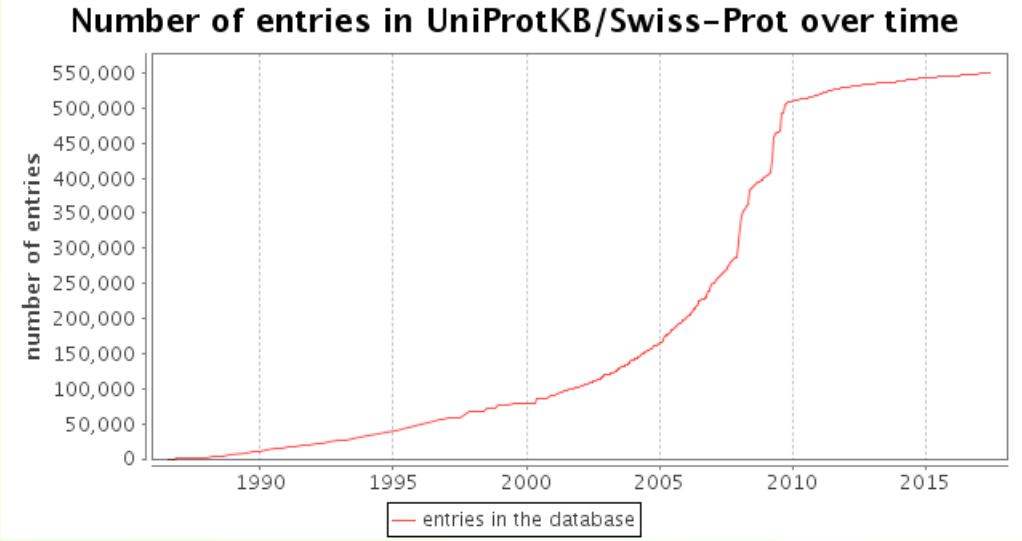
- Multiple research groups may sequence the same gene, chromosome etc
  - **and independently deposit the data in the same DB**
- Clearly this happens a lot with microbiome data
  - 16S sequences
  - Fragments of genomes from metagenomics reads
- Often it is important to be able to access **original data sets**, irrespective of redundancy
- For other purposes, it is essential to have access to **nonredundant** data

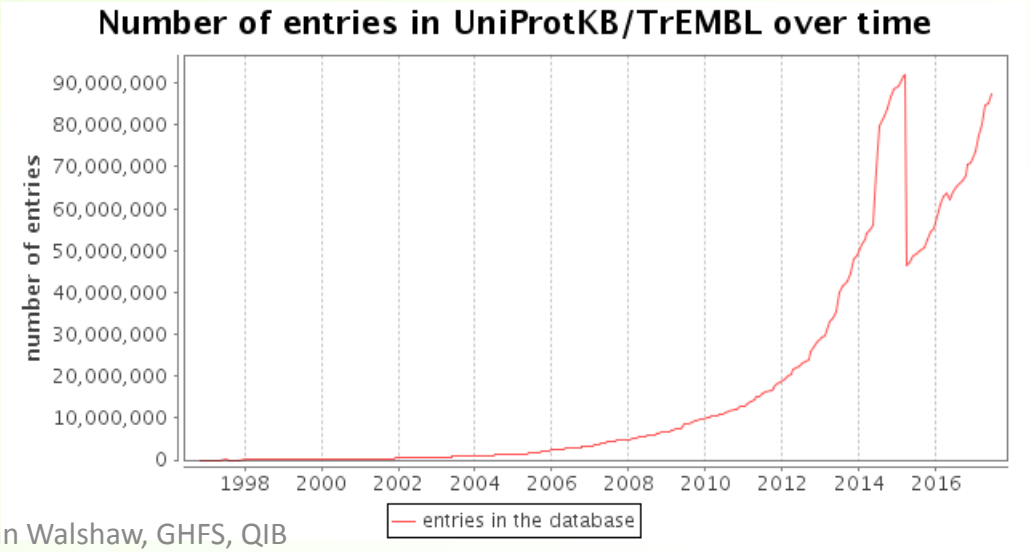
RefSeq : nonredundant sequences

- **RefSeq** – maintained by the NCBI
  - a database of sequences of:
    - Genomic DNA – including genes and whole chromosomes, and thus whole prokaryote genome sequences
    - Transcripts
    - Proteins
  - Non-redundant, i.e.: **Single standard reference sequence for each gene, chromosome, transcript, protein**
  - Cross-referenced
- Not to be confused with…
  - UniRef (nonredundant data sets from UniProt)
  - UniGene (NCBI) – which associates multiple fragments of **transcripts** with a single **gene sequence** (also achieves nonredundancy)

Primary databases are big

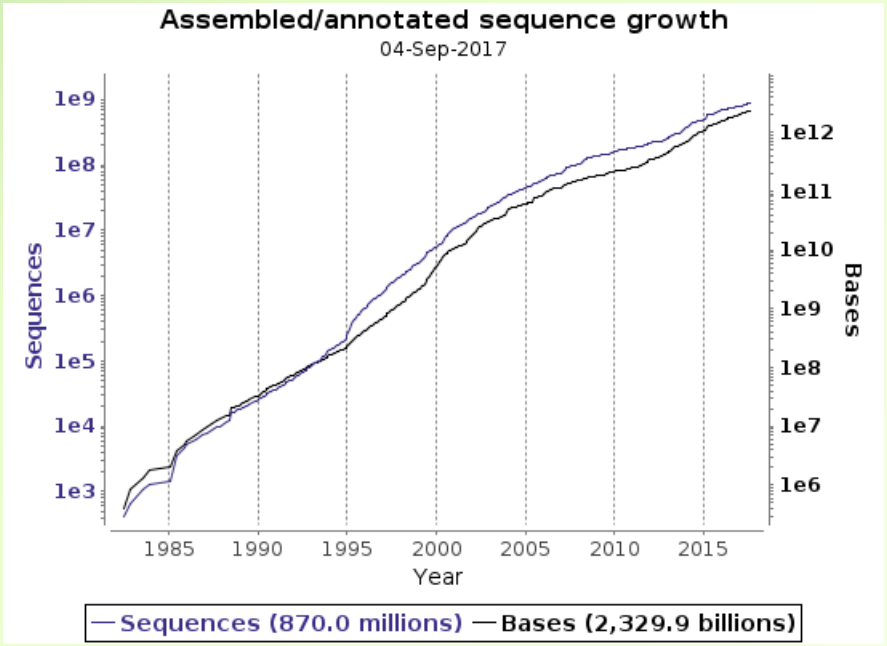
- UniProtKB - currently:
- 555,000 manually annotated/reviewed entries (Swiss-Prot)
- 89,000,000 auto-annotated/unreviewed entries (TrEMBL)
- www.uniprot.org/statistics/

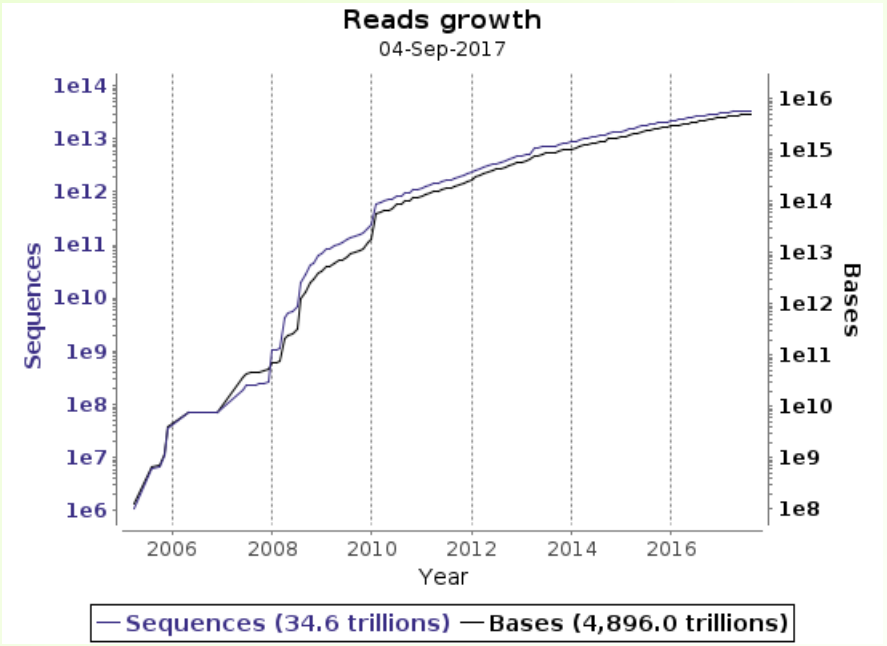


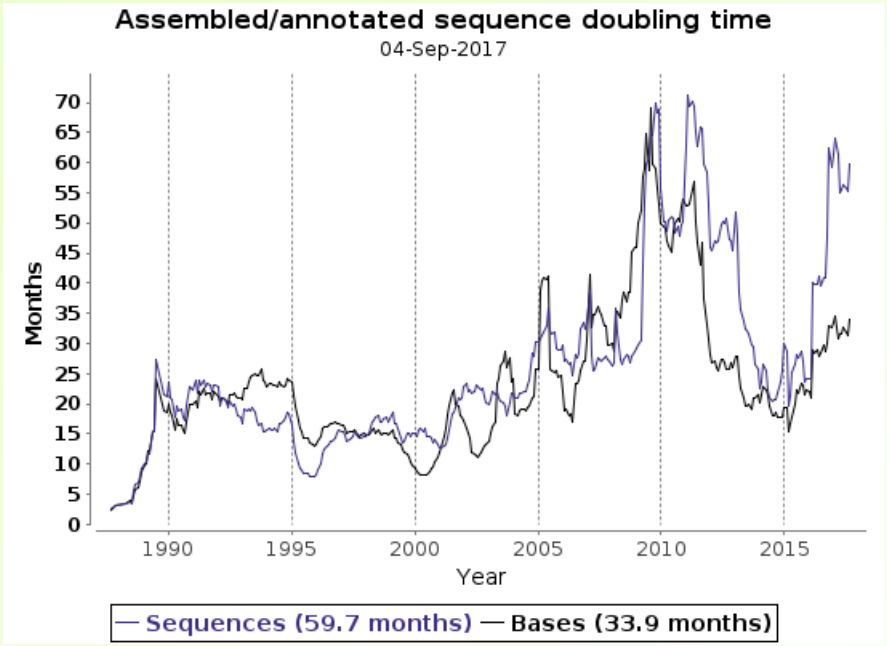


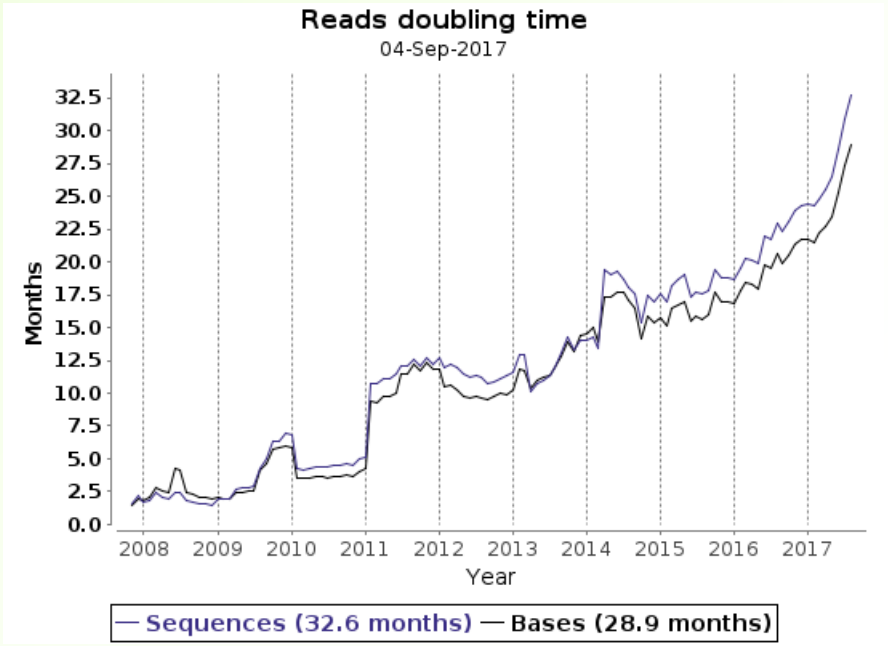
Primary nucleotide databases are VERY big

http://www.ebi.ac.uk/ena/about/statistics


Assembled/annotated sequence growth
04-Sep-2017
Sequences
Bases
Year
Sequences (870.0 millions)
Bases (2,329.9 billions)


Reads growth
04-Sep-2017
Sequences
Bases
Year
Sequences (34.6 trillions)
Bases (4,896.0 trillions)


Assembled/annotated sequence doubling time
04-Sep-2017
Months
Year
Sequences (59.7 months)
Bases (33.9 months)


Reads doubling time
04-Sep-2017
Months
Year
Sequences (32.6 months)
Bases (28.9 months)

Beyond primary DB providers

- Besides primary database maintainers-
  - numerous other providers of large datasets of many different kinds
  - Available for download, online interactive access and/or in conjunction with online analysis tools
- A few major examples (there are many others):
  - Kyoto University Bioinformatics Center
    - **Kyoto Encyclopedia of Genes and Genomes** (KEGG)
  - **Wellcome Trust Sanger Institute**, Hinxton
  - Swiss Institute of Bioinformatics, Lausanne
    - **ExPASy Bioinformatics Portal**
  - DOE-Joint Genome Institute (**JGI**), California
  - Human Microbiome Project Data Analysis and Coordination Centre (**HMP-DACC**)

Ribosomal RNA (gene) sequence databases

16S
Fungal 28S

Ribosomal Database Project (RDP)
Michigan State Univ.

16S/18S
23S/28S

SILVA
Max Planck Inst. for Marine Microbiology / Jacobs University

- Experts using their own systems for classification / curation
- Gives rise to **non-identical taxonomies**
- Which may also differ from other taxonomies (e.g. NCBI)

EzBioCloud 16S Database (formerly EzTaxon)
Chunlab, Inc.

Greengenes
Univ. Colorado

16S

- Associated with their own software, e.g.
- RDP Classifier
- SINA-Aligner (SILVA)

- Can be used in principle with various other software
- (may be some limitations)
- QIIME, MOTHUR, MEGAN etc

```
LOCUS       AY779786                 598 bp    DNA     linear   ENV 08-NOV-2004
DEFINITION  Uncultured archaeon clone sw3a52 small subunit ribosomal RNA gene,
            partial sequence.
ACCESSION   AY779786
VERSION     AY779786.1
KEYWORDS    ENV.
SOURCE      uncultured archaeon
  ORGANISM  uncultured archaeon
            Archaea; environmental samples.
REFERENCE   1  (bases 1 to 598)
  AUTHORS   Siering,P.L. and Wilson,M.S.
  TITLE     Geochemical and biological diversity in acidic hot springs in
            Lassen Volcanic National Park
  JOURNAL   Unpublished
REFERENCE   2  (bases 1 to 598)
  AUTHORS   Siering,P.L. and Wilson,M.S.
  TITLE     Direct Submission
  JOURNAL   Submitted (12-OCT-2004) Biological Sciences, Humboldt State
            University, 1 Harpst Street, Arcata, CA 95521, USA
FEATURES             Location/Qualifiers
     source          1..598
                     /organism="uncultured archaeon"
                     /mol_type="genomic DNA"
                     /isolation_source="thermal acidic environment"
                     /db_xref="taxon:115547"
                     /clone="sw3a52"
                     /environmental_sample
                     /country="USA: California, Lassen Volcanic National Park"
     rRNA            <1..>598
                     /product="small subunit ribosomal RNA"
ORIGIN
        1 gtgccagccg ccgcggtaat accagccccg cgagtggtcg ggactcttgc tgggcctaaa
       61 gcgcccgtag ccggcccggt aagtccctcc ttaaagcccc gggctcaacc cggggagcgg
      121 ggggatactg ccgggctagg gggcgggaga ggccgggggt accccagggg taggggcgaa
      181 atccgataat ccctggggga ccaccagtgg cgaaggcgcc cggctggaac gcgcccgacg
      241 gtgaggggcg aaagccgggg gagcgaaccg gattagatac ccgggtagtc ccggctgtaa
      301 actatgcggg ccaggtgtcg ggcgggcgtt agagcccgcc cggtgccgca gggaagccgt
      361 taagcccgcc gcctggggag tacggccgca aggctgaaac ttaaaggaat tggcgggggg
      421 gcacacaagg ggtggagcct gcggctcaat tggagtcaac gccgggaacc tctaccgggg
      481 gcgacagcag gatgacggcc aggctaacga ccttgcccga cgcgctgagg ggaggtgcat
      541 ggccgtcgcc agctcgtgct gtgaagtgtc ctgttaagtc aggcaacgag cgagaccc
//
```

GenBank record: AY779786.1

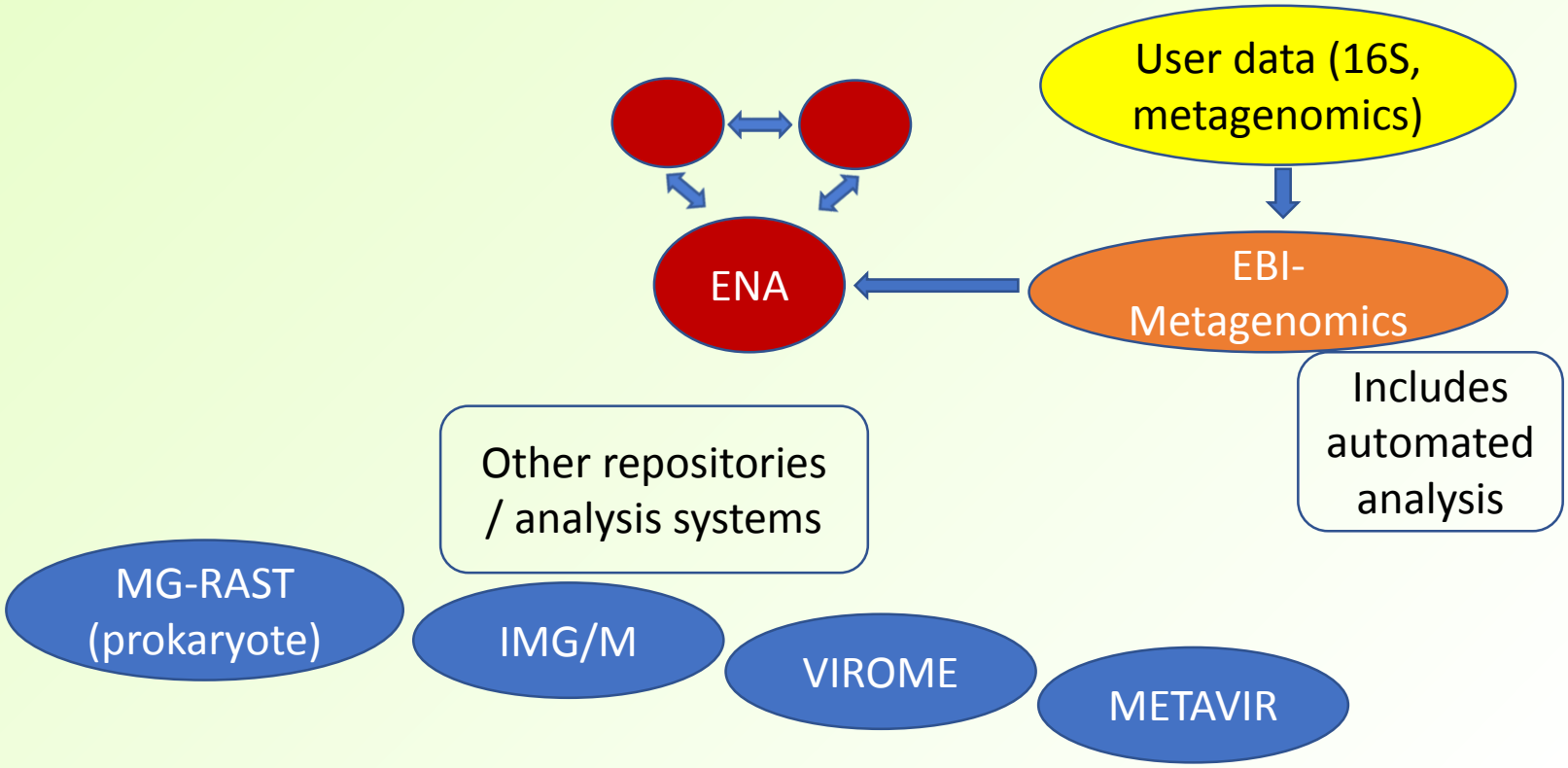
```
LOCUS       AY779786                 598 bp    DNA     linear   ENV 08-NOV-2004
DEFINITION  Uncultured archaeon clone sw3a52 small subunit ribosomal RNA gene,
            partial sequence.
ACCESSION   AY779786
VERSION     AY779786.1
KEYWORDS    ENV.
SOURCE      uncultured archaeon
  ORGANISM  uncultured archaeon
            Archaea; environmental samples.
```

GenBank record: AY779786.1

```
LOCUS       S000444337               598 bp    rRNA    linear   BCT 08-Nov-2004
DEFINITION  uncultured archaeon; sw3a52.
ACCESSION   AY779786 REGION: <1..>598
SOURCE      uncultured archaeon
  ORGANISM  uncultured archaeon
            Root; Archaea; "Crenarchaeota"; Thermoprotei; Acidilobales;
            Acidilobaceae; Acidilobus.
```

RDP record: S000444337

Microbiome-specific repositories



To be continued…

(Microbial) Genome Sequence Databases

References – Sequence Databases

Primary nucleotide sequence databases

- European Nucleotide Archive (ENA) http://www.ebi.ac.uk/ena
  - ("EMBL-Bank" is the ENA component equivalent to GenBank and DDBJ)
  - Leinonen, R. et al. (2011) The European Nucleotide Archive, *Nucleic Acids Res.* **39** (Database issue) D28-D31
- GenBank https://www.ncbi.nlm.nih.gov/genbank/
  - Benson *et al.* (2017) GenBank, *Nucleic Acids Res.* **45** (Database issue) D37-D42
  - Rindone, W.P. *et al.* (1983) GenBank™ - the Genetic Sequence Data-bank, *DNA-A Journal of Molecular & Cellular Biology* **2** (2) 173
  - Rindone, W.P. (1983) GenBank, *Trends in Pharmacological Sciences* **4** 326
- DNA Data Bank of Japan (DDBJ) http://www.ddbj.nig.ac.jp/
  - Mashima, J. *et al.* (2017) DNA Data Bank of Japan, *Nucleic Acids Res.* **45** (Database issue) D25-D31
  - Miyazawa, S. (1990) DNA Data-bank of Japan – Present Status and Future-plans, in *Computers and DNA*, pp. 47-61, eds G.I. Bell and T. Marr, Addison-Wesley, Reading, M.A.

References – Sequence Databases

Sequence databases associated with or derived from the primary nucleotide sequence databases – a few examples

- RefSeq https://www.ncbi.nlm.nih.gov/refseq/
  - O'Leary, N.A. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation, *Nucleic Acids Res.* **44**(Database issue) D733-D745
- UniGene https://www.ncbi.nlm.nih.gov/unigene/
  - (see also NCBI Resource Coordinators (2016) Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res.* **44**(Database issue) D7-D19

- Sequence Read Archive (sequences and quality data)
  - SRA at NCBI https://trace.ncbi.nlm.nih.gov/Traces/sra/
  - SRA at DDBJ (DRA) http://trace.ddbj.nig.ac.jp/dra/index_e.html
  - EBI SRA (sometimes referred to as "ERA") is integrated into ENA
  - Leinonen *et al.* (2011) The Sequence Read Archive, *Nucleic Acids Res.* **39** (Database issue) D19-D21

References – Sequence Databases

Primary protein sequence databases

- UniProt www.uniport.org
  - The UniProt Consortium (2017) UniProt: the universal protein knowledgebase, *Nucleic Acids Res.* **45** D158-D169

See also

- NCBI Protein https://www.ncbi.nlm.nih.gov/protein
- DDBJ Amino Acid Sequence Database (DAD) – see DDBJ, and ftp://ftp.ddbj.nig.ac.jp/ddbj_database/dad/

References – Sequence Databases

"Historical" primary protein sequence database references:

- Relating to the original Protein Sequence Database:
  - Dayhoff, M.O. (1965) Atlas of protein sequence and structure [Vol. 1], Silver Spring, MD, U.S.A. http://www.worldcat.org/title/atlas-of-protein-sequence-and-structure-vol-1-1965-margaret-o-dayhoff-et-al/oclc/605459794
  - Dayhoff, M.O. *et al.* (1975) Evolution of Sequences within Protein Superfamilies *Naturwissenschaften* **62** 154-161
  - **PIR-PSD** (see UniProt); *final release December 2004; now integrated into UniProt* http://pir.georgetown.edu/pirwww/dbinfo/pir_psd.shtml
    - Barker W.C. *et al.* (1991) The PIR sequence database, *Nucleic Acids Res.* **19** (Suppl) 2231-2236
    - Wu, C.H. *et al.* (2003) The Protein Information Resource, *Nucleic Acids Res.* **31** (1) 345-347
  - **MIPS** (formerly Martinsried Institute for Protein Sequences; latterly Munich Information Center for Protein Sequences, and the name of the database it maintained)
    - Mewes, H.W. *et al.* (1998) MIPS: a database for protein sequences and complete genomes , *Nucleic Acids Res.* **26** (1) 33-37
    - Mewes, H.W. *et al.* (2010) MIPS: curated databases and comprehensive secondary data resources in 2010, *Nucleic Acids Res.* **39** (Database issue) D220-D224
    - MIPS is now integrated into the Institute of Bioinformatics and Systems Biology (IBIS), German Research Center for Environmental Health https://www.helmholtz-muenchen.de/ibis
- Swiss-Prot/TrEMBL (now components of UniProt)
  - Bairoch, A. & Boeckmann, B. (1991) The SWISS-PROT protein sequence data bank, *Nucleic Acids Res.* **19** (Suppl) 2247–2249
  - Bairoch, A. & Apweiler, R. (1996) The SWISS-PROT Protein Sequence Data Bank and Its New Supplement TREMBL, *Nucleic Acids Res.* **24** (1) 21–25

References – Sequence Databases

Some mentioned example non-sequence databases associated with primary database providers:

- NCBI
  - BioProject
    https://www.ncbi.nlm.nih.gov/bioproject
  - Samples https://www.ncbi.nlm.nih.gov/biosample
  - Assembly
    https://www.ncbi.nlm.nih.gov/assembly

References – Sequence Databases

Institutes, consortia and miscellaneous projects

- International Nucleotide Sequence Database Collaboration (**INSDC**) http://www.insdc.org
- European Bioinformatics Institute (**EBI**) http://www.ebi.ac.uk
  - *is an outstation of*: European Molecular Biology Laboratory (**EMBL**) http://www.embl.de
- DNA Database of Japan **(DDBJ) Center** http://www.ddbj.nig.ac.jp
- National Center for Biotechnology Information (**NCBI**) http://www.ncbi.nlm.nih.gov
- Swiss Institute of Bioinformatics (**SIB**) http://www.isb-sib.ch
- Protein Information Resource (**PIR**) http://pir.georgetown.edu
- Joint Genome Institute (**JGI**) http://jgi.doe.gov
- Wellcome Trust Sanger Institute http://www.sanger.ac.uk
- Kyoto Encyclopedia of Genes and Genomes (**KEGG**)http://www.kegg.jp
- NIH Human Microbiome Project (**HMP**) https://commonfund.nih.gov/hmp
  - Data Analysis and Coordination Centre http://www.hmpdacc.org

References – Sequence Databases

Ribosomal RNA gene sequence databases

- RDP (Ribosomal Database Project) http://rdp.cme.msu.edu
  - Cole, J.R. *et al.* (2014) Ribosomal Database Project: data and tools for high throughput rRNA analysis, *Nucl. Acids Res.* **42**(Database issue) D633-D642
- SILVA http://www.arb-silva.de
  - Quast, C. *et al.* (2013) The SILVA ribosomal RNA gene database project: improved data processing and web-based tools, *Nucl. Acids Res.* **41** (Database issue) D590-D596
- Greengenes http://greengenes.secondgenome.com
  - (previously http://greengenes.lbl.gov - data still available there)
  - McDonald, D. *et al.* (2011) An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea, *ISME J.* **6**(3) 610-608
- EzBioCloud (formerly EzTaxon) http://www.ezbiocloud.net/taxonomy
  - Chun, J. *et al.* (2007) Introducing EzTaxon-e: a prokaryotic 16S rRNA gene sequence database with phylotypes that represent uncultured species, *Int. J. Syst. Evol. Microbiol.* **57**(10) 2259-2261
  - Kim, O.S. *et al.* (2012) EzTaxon: a web-based tool for the identification of prokaryotes based on 16S ribosomal RNA gene sequences, *Int. J. Syst. Evol. Microbiol.* **62**(3) 716-721

References – Sequence Databases

Some metagenome sequence repositories/annotation resources

- EBI Metagenomics https://www.ebi.ac.uk/metagenomics
  - Mitchell, A. *et al.* (2016) EBI metagenomics in 2016 - an expanding and evolving resource for the analysis and archiving of metagenomic data, *Nucleic Acids Res.* **44** (Database issue) D595-D603
- MG-RAST http://metagenomics.anl.gov
  - Meyer, F. *et al.* (2008) The Metagenomics RAST server — A public resource for the automatic phylogenetic and functional analysis of metagenomes, *BMC Bioinformatics* **9** 386
- IMG/M https://img.jgi.doe.gov
  - Chen, I.A. *et al.* (2017) IMG/M: integrated genome and metagenome comparative data analysis system, *Nucleic Acids Res.* **45** (Database issue) D507-D516
- VIROME http://virome.dbi.udel.edu
  - Wommack, K. E. *et al.* (2012) VIROME: a standard operating procedure for analysis of viral metagenome sequences, *Stand. Genomic Sci.* **6**(3) 427-439
- METAVIR / METAVIR2 http://metavir-meb.univ-bpclermont.fr
  - Roux, S. et al. (2011) Metavir: a web server dedicated to virome analysis, *Bioinformatics* **27** (21) 3074-3075
  - Roux, S. et al. (2014) Metavir 2: new tools for viral metagenome comparison and assembled virome analysis, *BMC Bioinformatics* **15** 76