Introducing Microbiome Bioinformatics

Part 12.

Sequence databases (continued).

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

- 11 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

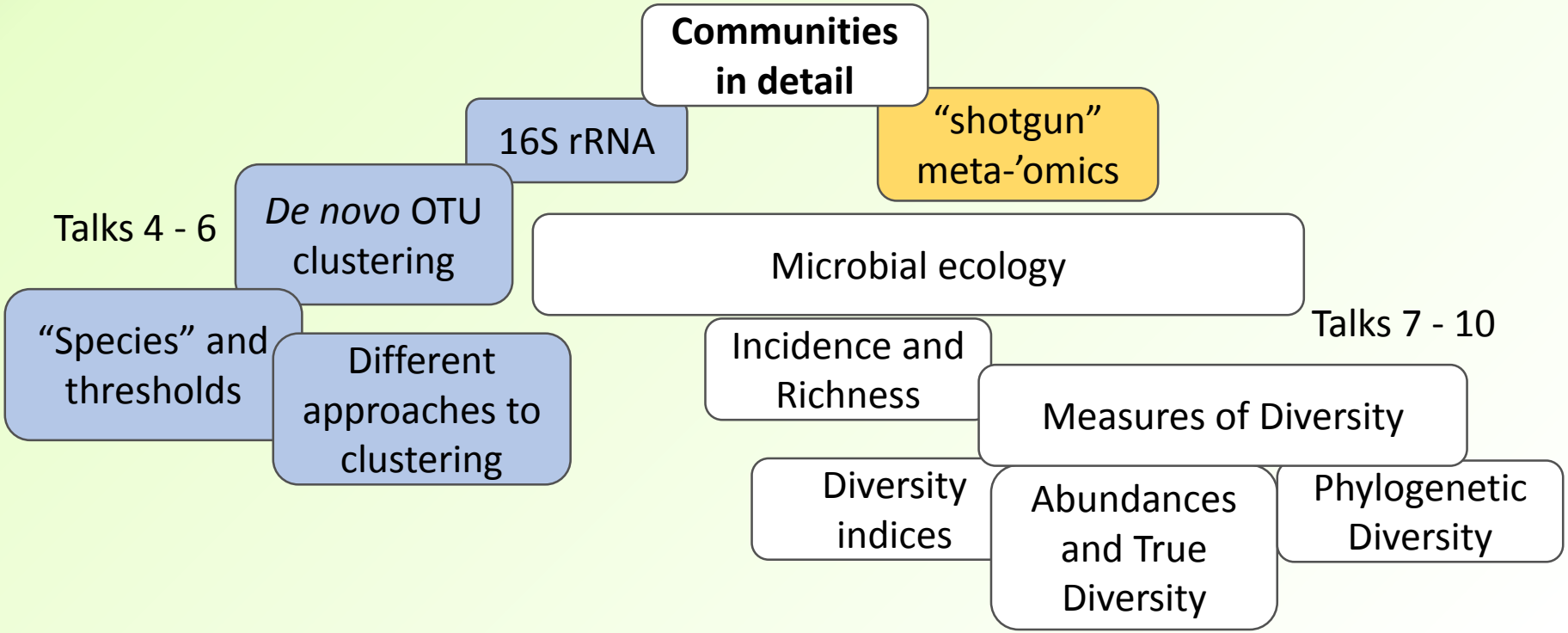
- Types of environments studied
- Variety of microbiomes, and domains of Life
- Aims of studies:
  - "Communities"
  - Functions

Aims of DNA/ RNA sequencing approaches

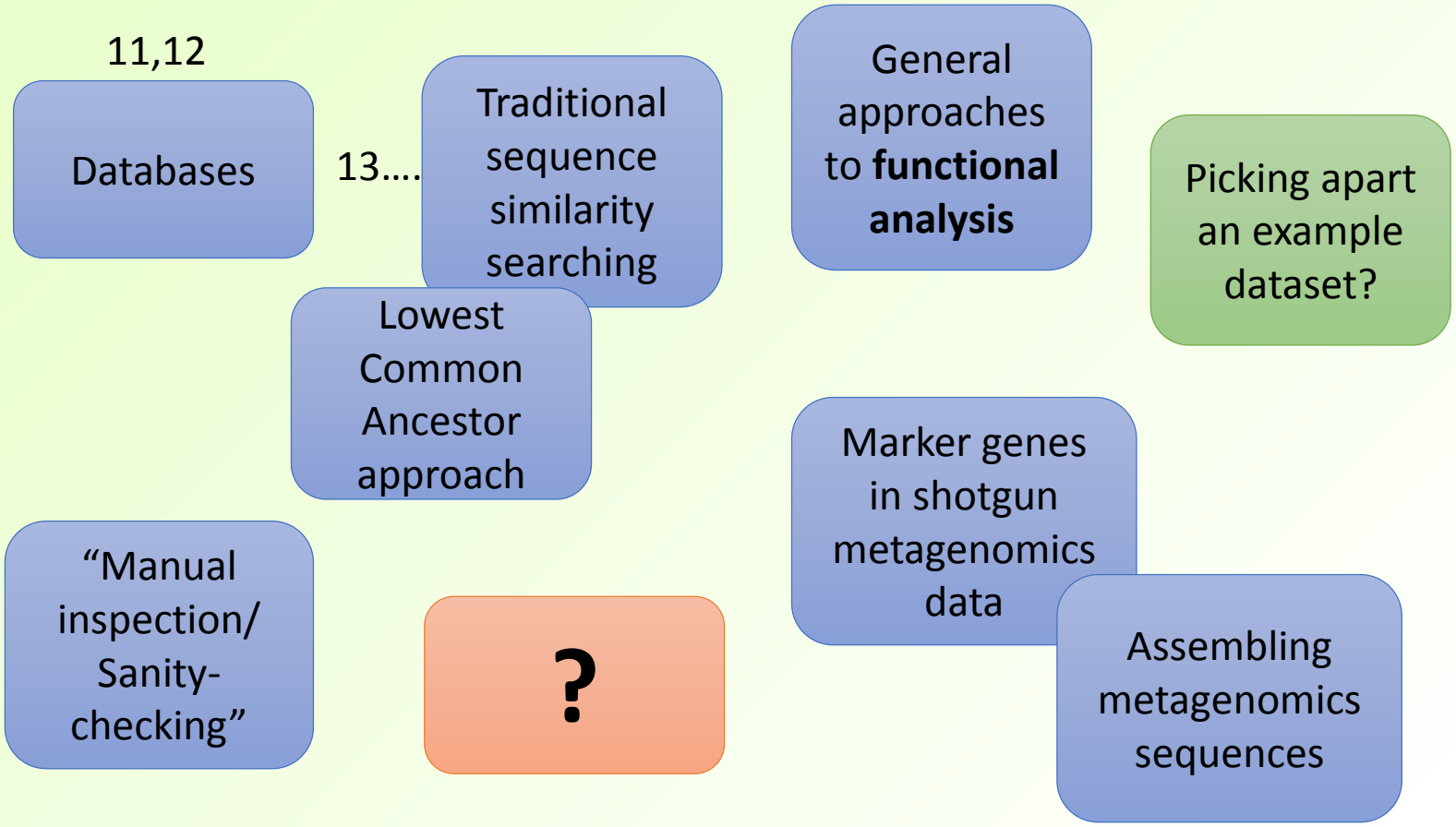
- "barcoding"- use of marker genes
  - 16S rRNA
  - Other markers
- "shotgun" metagenomics
  - metatranscriptomics
  - aims of whole-metagenome sequencing

Dealing with the data: Informatics concepts

- 16S rRNA
  - Communities in detail
- "shotgun" meta-'omics


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

What next?

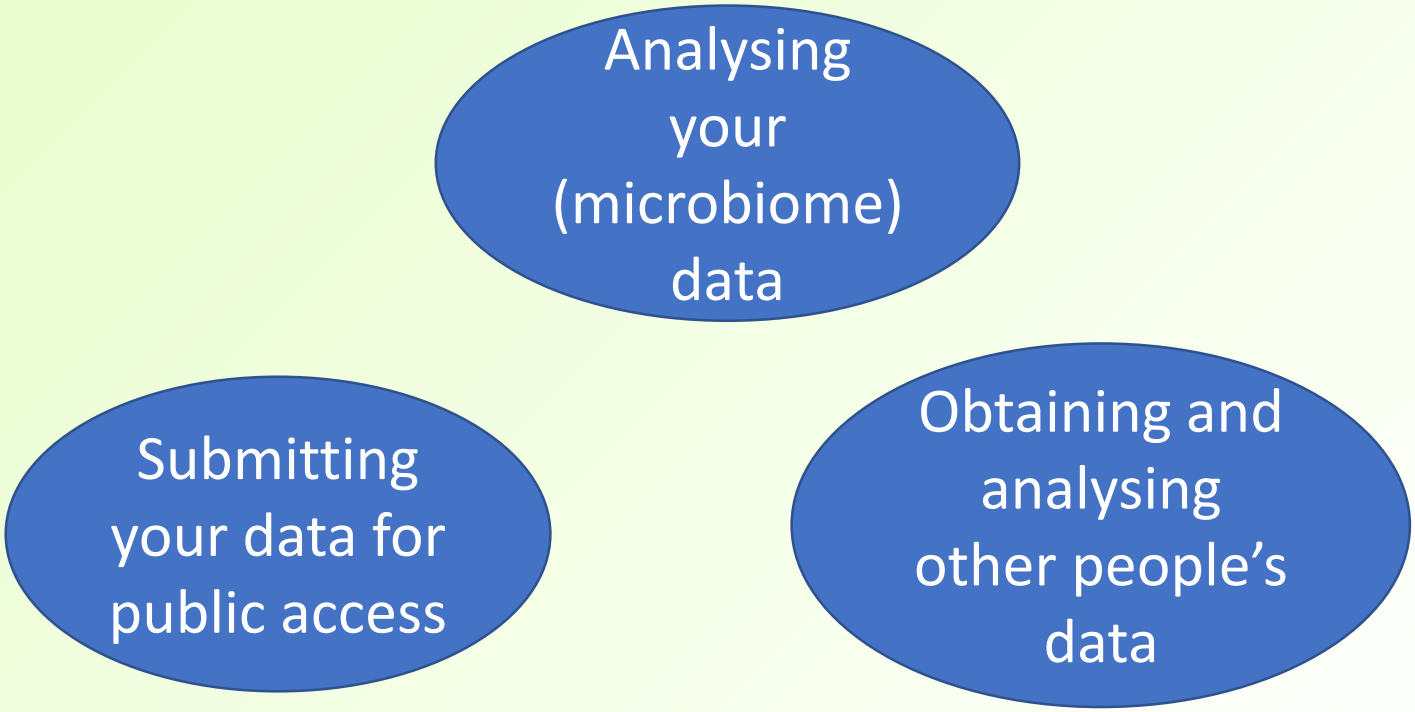

11,12
Databases
13….
Traditional sequence similarity searching
General approaches to **functional analysis**
Picking apart an example dataset?
Lowest Common Ancestor approach
Marker genes in shotgun metagenomics data
"Manual inspection/ Sanity-checking"
?
Assembling metagenomics sequences

Series of talks

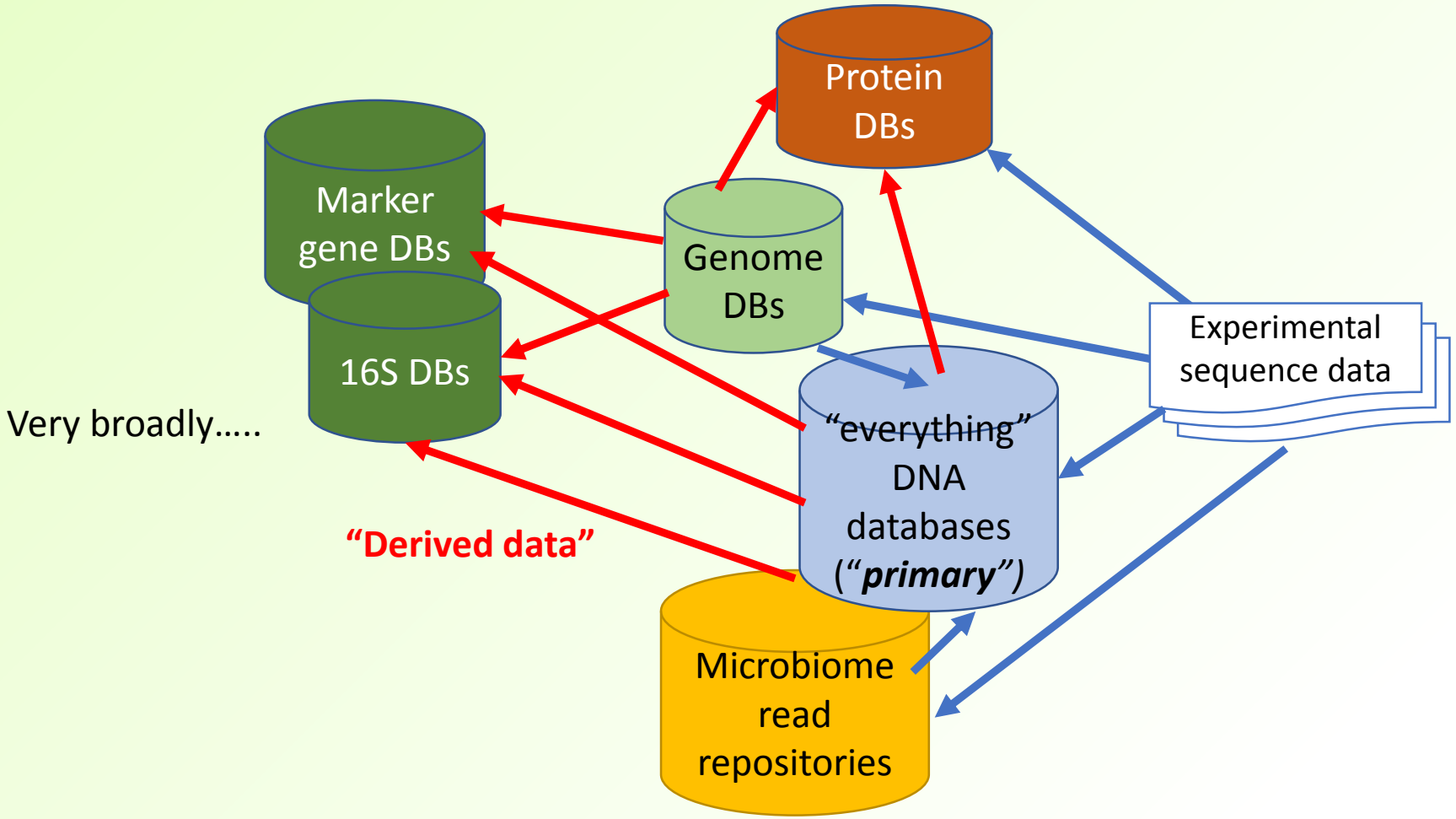
Slideshows - http://ghfs1.quadram.ac.uk/ghfs/

- 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: 8/9/2017 – Introducing sequence databases
- Part 12: today – Sequence databases (continued);

Sequence databases

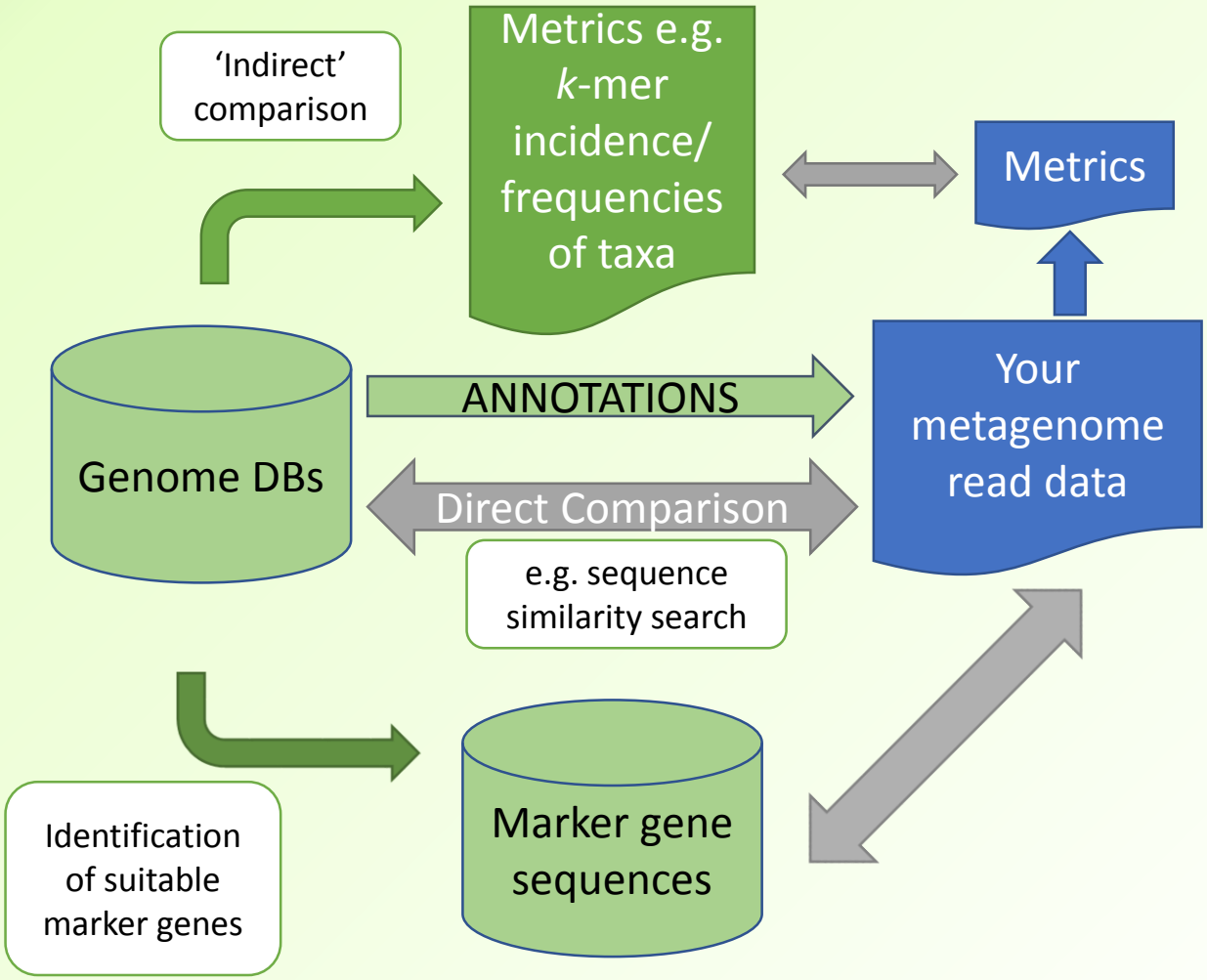


Sequence databases and microbiome analysis

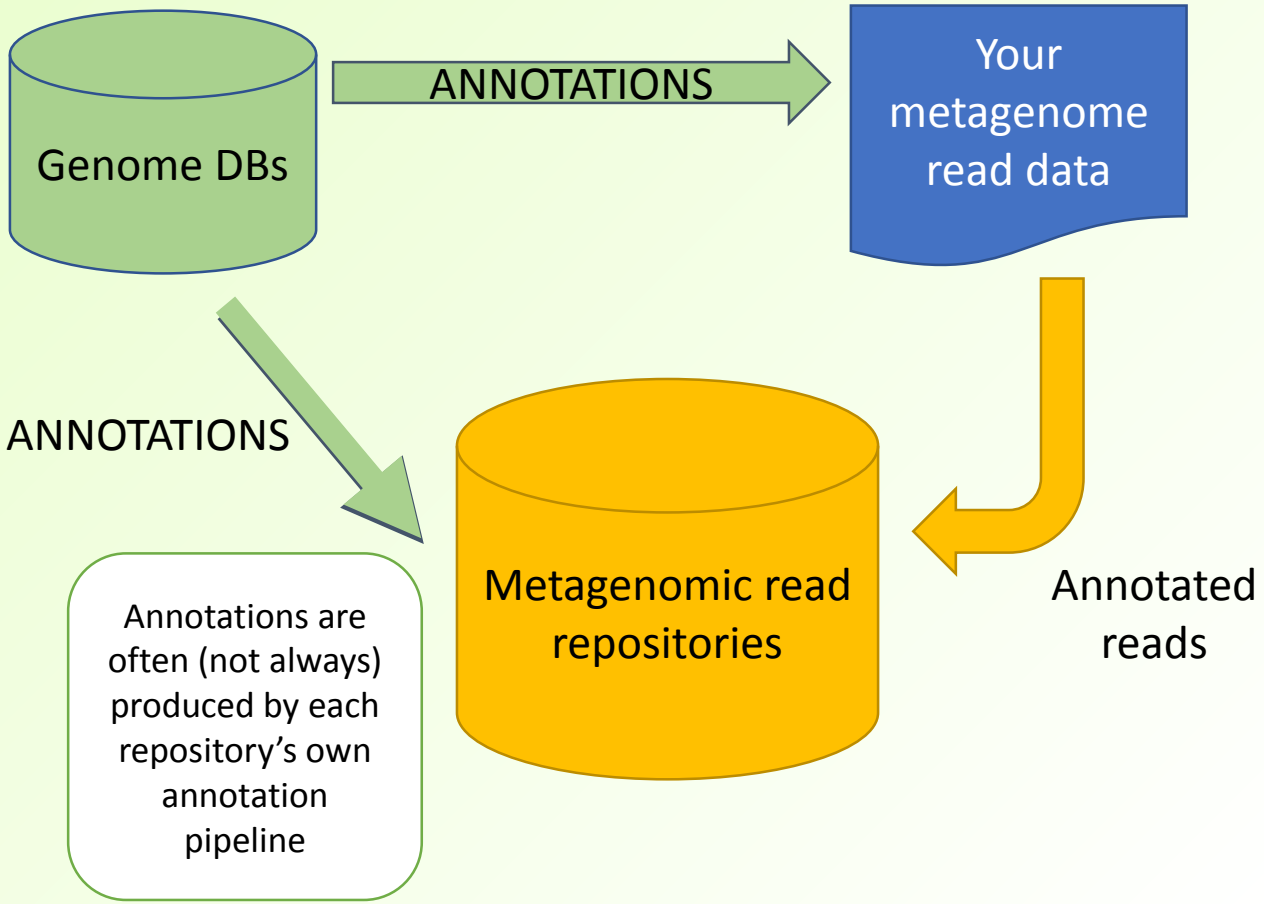


A very brief ‘how and why’ of genome sequence databases for analysing your data

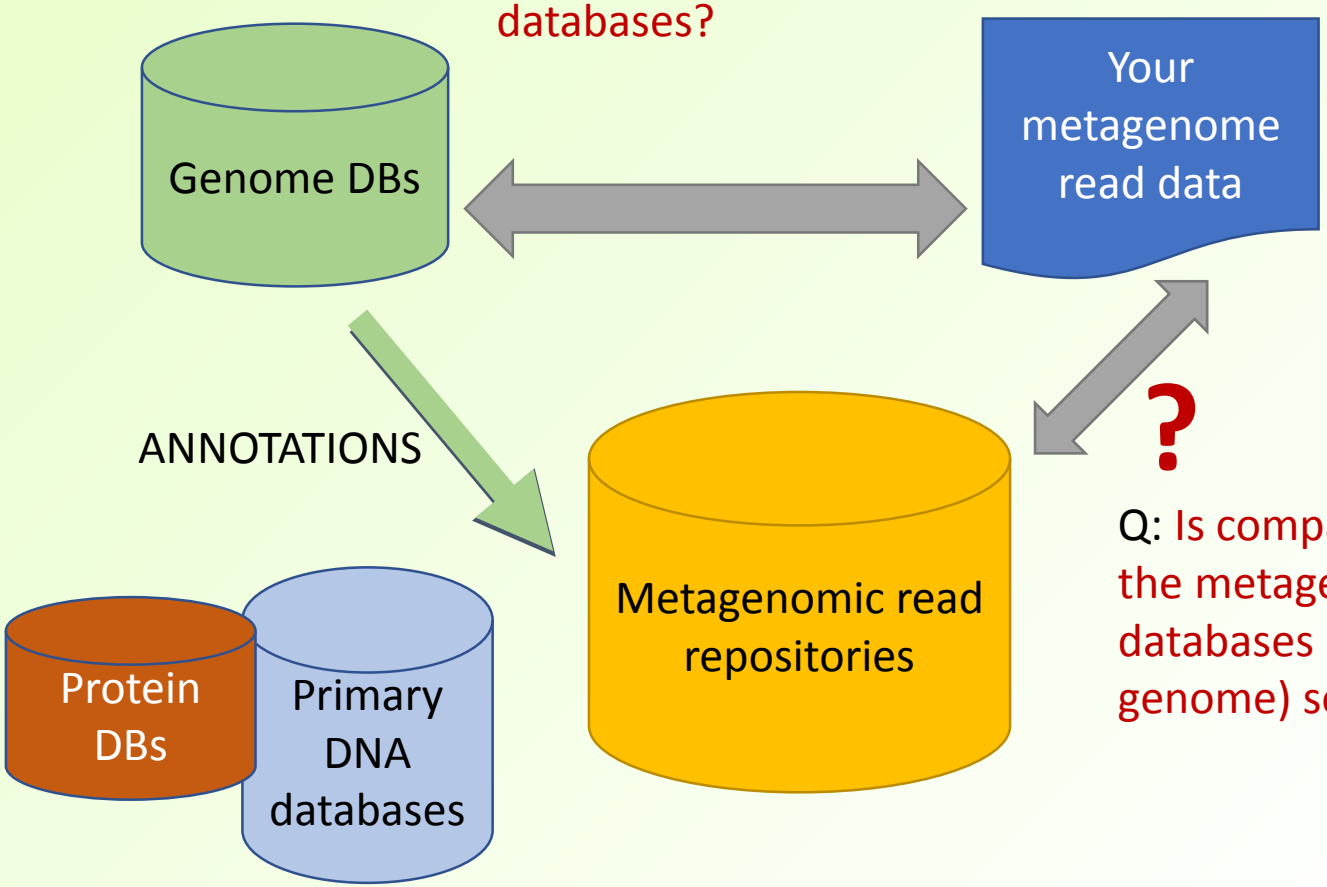
More details next time



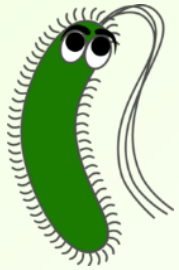
Similar principles apply to both metagenomic and 16S read data (with some caveats)


Genome DBs
ANNOTATIONS
Your metagenome read data
ANNOTATIONS
Metagenomic read repositories
Annotated reads
Annotations are often (not always) produced by each repository's own annotation pipeline

Q: Conversely is it sensible to use ONLY genome sequence databases?



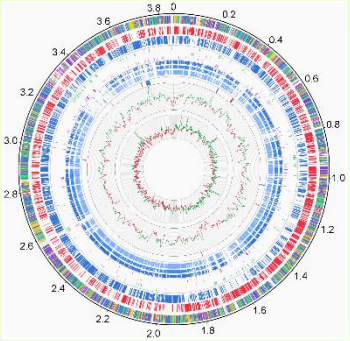
Q: Is comparison with the metagenome databases (instead of genome) sensible?



Microbial genome sequences

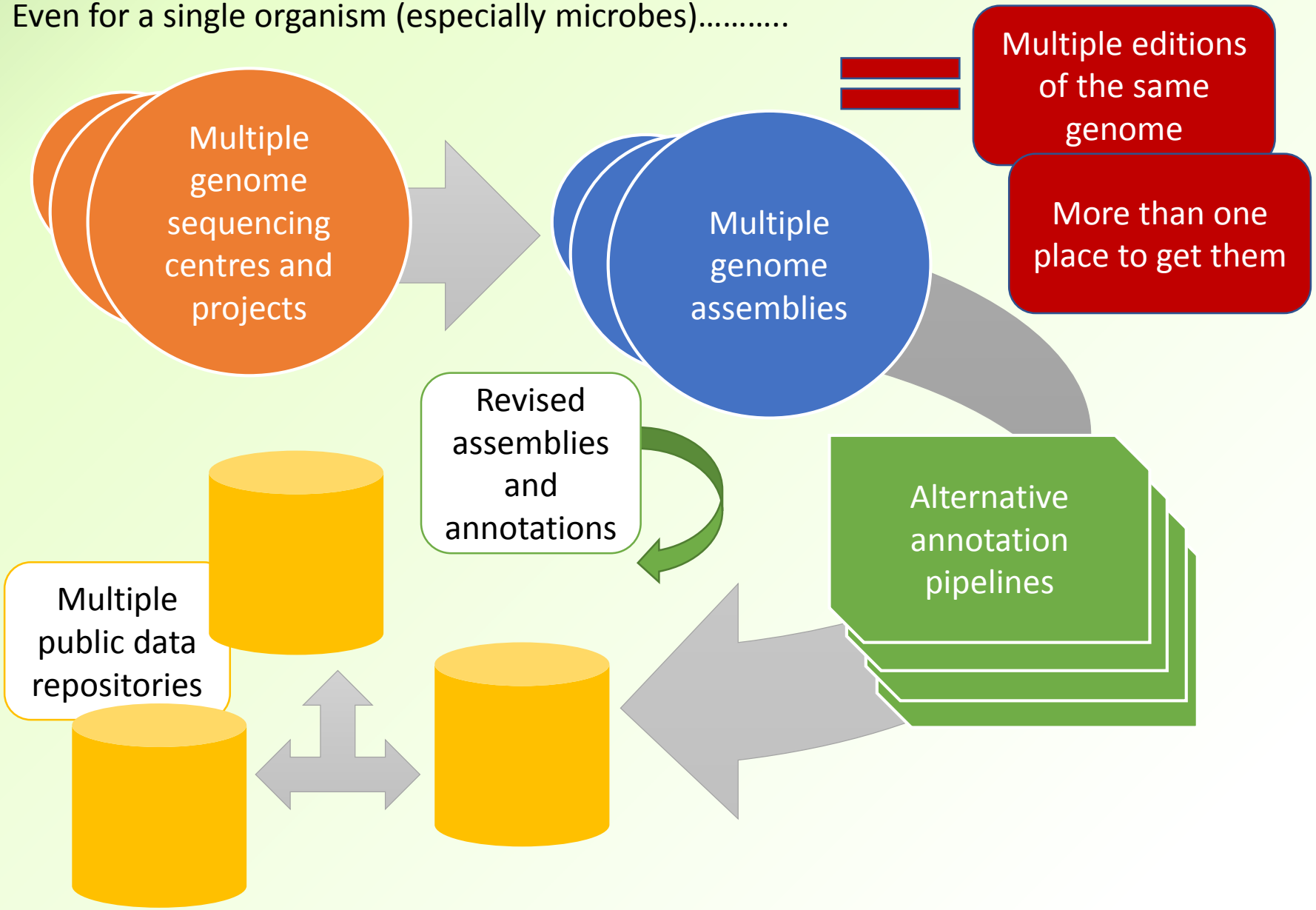
You can never have enough of them (?)

(Or enough collections of them)



10.1371/journal.pone.0015489

Even for a single organism (especially microbes)…….



Some collections

- **Generally with their own analysis/annotation pipelines**
  → Added value
- **NCBI Genomes** https://www.ncbi.nlm.nih.gov/genome - all kingdoms of life
  - Sequences + annotations available via Genbank
  - *and* via RefSeq
  - Not necessarily identical annotations! (also reannotation projects)
- At EBI – **Ensembl (Genomes)** – bacteria.ensembl.org
  - fungi.ensembl.org , protists.ensembl.org (no viruses)
- **PATRIC**, Univ. Chicago www.patricbrc.org
  - Originally, Pathosystems Resource Integration Center
  - Prokaryotes, plus host eukaryotes
  - **107,086** Bacteria; **1,110** Archaea; **8** vertebrate + **2** invertebrate
  - Some features require registration

Some collections

- Integrated Microbial Genomes (**IMG**), DOE-JGI img.jgi.doe.gov/m
  - Prokaryotes and Eukaryote microbes
  - Also provides an annotation system
  - Distributes existing public genomes and user-deposited data

	Isolates		SAGs		MAGs	
Sequenced at:	JGI	All	JGI	All	JGI	All
Bacteria	6142	49192	1833	2214	3854	4347
Archaea	207	776	198	294	79	247
Eukarya	76	267	0	0	1	1
Viruses	0	7854	0	44	0	0

(Only data sets with GOLD metadata were counted.)

Some collections

- Sanger Inst. genomes www.sanger.ac.uk/science/data
  - **180** bacterial genomes; **32** protozoa; yeast- SGRP: **37** *S. cerevisiae*, **27** *S. paradoxus;* **5** other fungi; **12** virus including **5** bacteriophage
- **HMRGD** at HMP-DACC – reference prokaryote genomes assembled from human microbiome (about **1,400**) www.hmpdacc.org/hmp/HMRGD/
- Virus Pathogen Resource (ViPR) www.viprbrc.org
  - Large number (**28,164**) of complete and (> 50,000) partial genomes
- FungiDB http://fungidb.org **85** genomes + **20** oomycete

Some collections

- Many other places
- Some are primarily resources for visualising and analysing the data, which also provide the genome data for download
  - Interactively view genome structure and annotations of functions/pathways
  - (Not all such resources make the data available for download)

Where to get genome sequences?

- In general, any given public genome sequence will be available from several different places
  - Especially true for bacteria
- When is a genome sequence “complete”? Drafts,finishing,builds…
- Genomes may be available in complete or incomplete form – regarding both **sequence** and **annotation**
  - Whole chromosomes; Incomplete chromosomes in fragments (“scaffolds”)
  - Sometimes, associated plasmids too
  - With no annotations, or….
  - …partial annotations (e.g. just predicted coding regions)….
  - …more complete annotations : more detailed information on genes, gene products (maybe with protein sequences available)
  - And whether they are hypothetical or experimentally verified
  - **Alternative sets of annotations** may be distributed
    - By different repositories; Even by the same repository

Example:

- *Staphylococcus aureus* strain 1943STDY5573617
  - BioProject PRJEB2655 (“Diversity of MRSA”)
  - 29 fragments (0.5 kbp – 633 kbp); total **2.7 Mbp**
    - 7 fragments > 100 kbp
  - Sanger Inst. → ENA → DDBJ, NCBI

```
>FJNP01000001.1 Staphylococcus aureus strain 1943STDY5573617 genome assembly,
contig: ERS329598SCcontig000001, whole genome shotgun sequence
AGTACATCTATGTCTACTTTAGGTTTTATTGACATAAATAAAGCTCCCTTCAAAGTTTTC
ATTTTTTCAATGTCTACTTTGAAGGGAGCATTTCACTGAACTTTGTTCAGGCTCTTTTTA
AATGTATATCAGACATGGGCAGCGACTTGATAGTGAAAGTCCATATATGCTTTGTAGTCA
AAACTGCTAGCGGATATTGTTATCTTAACAAACGTGAAGCTCAAGCAGCAATTTAGTCAT
TTTATTTTTTATTGAAAGAAGTGGAAAACATGACAATGATATATAGAAATAATTTCATTG
TGTTCGTTTTATCATTTTTTATTAGTATTATATTGTATTCATCGCACGTATTACTCCCAT
TTATGTTTGGTCCTATTATCGCATCAATCATTTGTGTGAAAGTTTTCAAACTTGATATTA
AATGGCCATTCTTACTTAGTGAATTAGGGATTGTACTATTAGGTGTGCAAATCGGATCAA
CGTTTACGAAAAATGTCGTTATGGATATTAAAGACAATTGGCTTTCGATTATTGTTGTAT
CTATTTCGATATTATTAATTGCATTAGTAATGGCATTTATTTTCAAAAAAATTGCACGTA
TTAATACAGAAACAGCTATTTTAAGTGTTATACCAGGAGCACTAACACAAATGCTGGTCA
TGGCTGAACAAGACAAACGTGCTAATTTGTTAGTAGTTAGCTTAACGCAAACATCACGAA
TTATATTTGTTGTTGTTTTAGTACCGTTCATTTCATATTTTTTTCATGATGGTAACATGC
ATGCGAATGGTAAGTTAACAAAAGTCTTGCCTTTATCACAAGTATTAAACATAGGGCAAA
TAGTTATTTTAGTGATAGCTATCTTTATAGTTTATCTAATTATGTCTAAAATAAAGTTTC
CAACATTTCAATTATTAGCACCACTCATTGTATTAATTGTTTGGAATTTTTCTACAGGTT
TAACATTTACACTAGATCATTGGTTGTTGAACATGGCACAACTAATATATATGATTAGAA
TTGGAGTTCAAATAGCGCATTTATTGTCAGATTTAAAAGGTAGACTAGCAATCGCAATTA
CAATTCAAAATATTATGTTGATAATTGGTGCGCTAATCATGGTTTATGTCATACATTTCT
```

NCBI records:

```
##gff-version 3
#!gff-spec-version 1.21
#!processor NCBI annotwriter
#!genome-build 10900_2#28
#!genome-build-accession NCBI_Assembly:GCA_900070305.1
##sequence-region FJNP01000001.1 1 632980
##species http://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?id=1280
FJNP01000001.1  EMBL    region  1       632980  .       +       .
ID=id0;Dbxref=taxon:1280;collection-date=1999;country=United
Kingdom:Scotland;gbkey=Src;isolation-source=sub-cutaneous abscess;mol_type=genomic
DNA;note=contig: ERS329598SCcontig000001;serovar=NA;strain=1943STDY5573617
FJNP01000001.1  EMBL    gene    276     1343    .       +       .
ID=gene0;Name=ERS329598_00001;gbkey=Gene;gene_biotype=protein_coding;locus_tag=ERS329598_00001
FJNP01000001.1  EMBL    CDS     276     1343    .       +       0
ID=cds0;Parent=gene0;Dbxref=NCBI_GP:CZQ50841.1;Name=CZQ50841.1;gbkey=CDS;product=Abrb;protein_i
d=CZQ50841.1;transl_table=11
FJNP01000001.1  EMBL    gene    1566    2126    .       -       .
ID=gene1;Name=tag;gbkey=Gene;gene=tag;gene_biotype=protein_coding;locus_tag=ERS329598_00002
FJNP01000001.1  EMBL    CDS     1566    2126    .       -       0
ID=cds1;Parent=gene1;Dbxref=NCBI_GP:CZQ50864.1;Name=CZQ50864.1;gbkey=CDS;gene=tag;product=DNA-
3-methyladenine glycosylase;protein_id=CZQ50864.1;transl_table=11
FJNP01000001.1  EMBL    gene    2529    5159    .       +       .
ID=gene2;Name=valS;gbkey=Gene;gene=valS;gene_biotype=protein_coding;locus_tag=ERS329598_00004
FJNP01000001.1  EMBL    CDS     2529    5159    .       +       0
ID=cds2;Parent=gene2;Dbxref=NCBI_GP:CZQ50881.1;Name=CZQ50881.1;gbkey=CDS;gene=valS;product=Valy
l-tRNA synthetase;protein_id=CZQ50881.1;transl_table=11
FJNP01000001.1  EMBL    gene    5172    6443    .       +       .
ID=gene3;Name=folC;gbkey=Gene;gene=folC;gene_biotype=protein_coding;locus_tag=ERS329598_00005
FJNP01000001.1  EMBL    CDS     5172    6443    .       +       0
ID=cds3;Parent=gene3;Dbxref=NCBI_GP:CZQ50905.1;Name=CZQ50905.1;gbkey=CDS;gene=folC;product=Dihy
drofolate synthase / Folylpolyglutamate synthase;protein_id=CZQ50905.1;transl_table=11
FJNP01000001.1  EMBL    gene    6711    7418    .       +       .
ID=gene4;Name=comC;gbkey=Gene;gene=comC;gene_biotype=protein_coding;locus_tag=ERS329598_00006
FJNP01000001.1  EMBL    CDS     6711    7418    .       +       0
ID=cds4;Parent=gene4;Dbxref=NCBI_GP:CZQ50930.1;Name=CZQ50930.1;gbkey=CDS;gene=comC;product=Late
competence protein ComC%252C processing protease;protein_id=CZQ50930.1;transl_table=11
FJNP01000001.1  EMBL    gene    7415    8101    .       +       .
ID=gene5;Name=ERS329598_00007;gbkey=Gene;gene_biotype=protein_coding;locus_tag=ERS329598_00007
```

Genome annotations

(In general; virus, prokaryote, eukaryote)

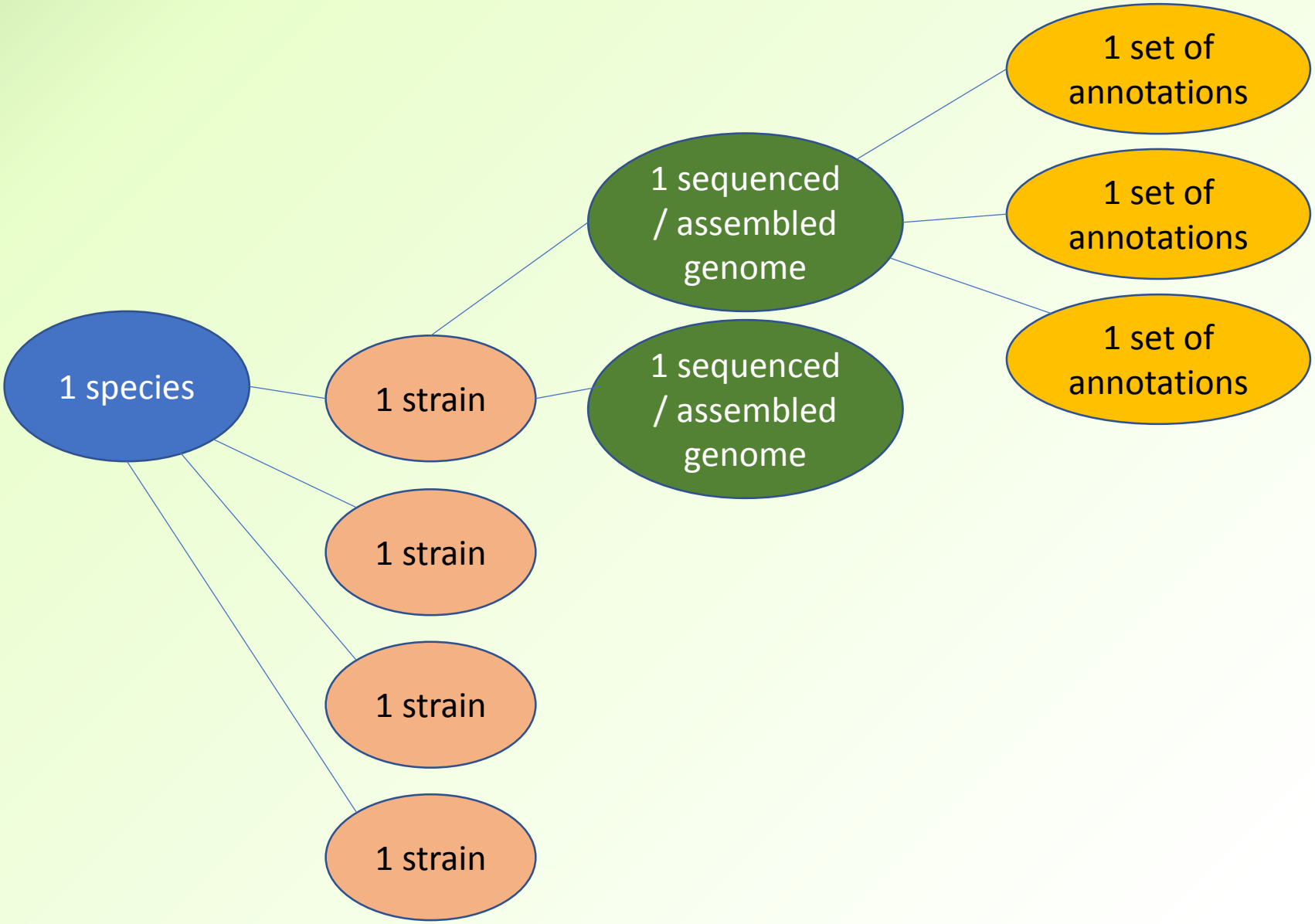
- Early genome sequencing projects involved teams of annotators
- Third parties may have provided alternative annotations
- Various factors can cause differences
  - E.g. different algorithms for gene-prediction ; Methods for inferring likely gene function
- Annotation is much more automated than it used to be
  - Especially true for prokaryote genomes
  - Easy to apply multiple methods
- Primary repositories (e.g. NCBI; Ensembl-Genomes at EBI) will generally **automatically annotate** deposited genome sequences (not always)
  - Sometimes using **more than one method**
  - And may provide the depositors' annotations as well (if any)
- Additionally, annotations may be revised and updated
  - Especially true of human and principal model organisms
  - Also: **NCBI Prokaryotic RefSeq Genome Reannotation Project**

➢ **Multiple assemblies/annotations may be available (even from one repository)**

Beyond traditional genome projects

- Traditional targeted single-genome projects:
  - Microbes: culture of a single strain
  - Metazoan or plant: Library of chromosome fragments from an individual specimen
    - or maybe several specimens
- Targeted 'pan-genome' projects
  - Multiple (different) strains of the same species
  - Multiple individuals, e.g.
    - 1,000 Genomes Project (human)
    - 100,000 Genomes Project (human); includes >1 genome from some individuals
- When genomes 'select you' (not the other way round)
  - Microbial pathogen isolates from disease outbreaks
  - Assembling microbial genomes from large-scale **metagenome** projects

➢ **Multiple assemblies/annotations may be available (even from one repository)**


1 species
1 strain
1 strain
1 strain
1 strain
1 sequenced / assembled genome
1 sequenced / assembled genome
1 set of annotations
1 set of annotations
1 set of annotations

Powerful but potentially confusing

- Comprehensive series of databases and access points for a variety of purposes
- = multiple ways of accessing the same data in different forms
- With corresponding records in many of these DBs cross-referenced accordingly
- Within one web-based resource (e.g. NCBI)
- (with cross-refs to DBs on other sites)

Redundancy

A quick recap from last time

RefSeq : nonredundant sequences

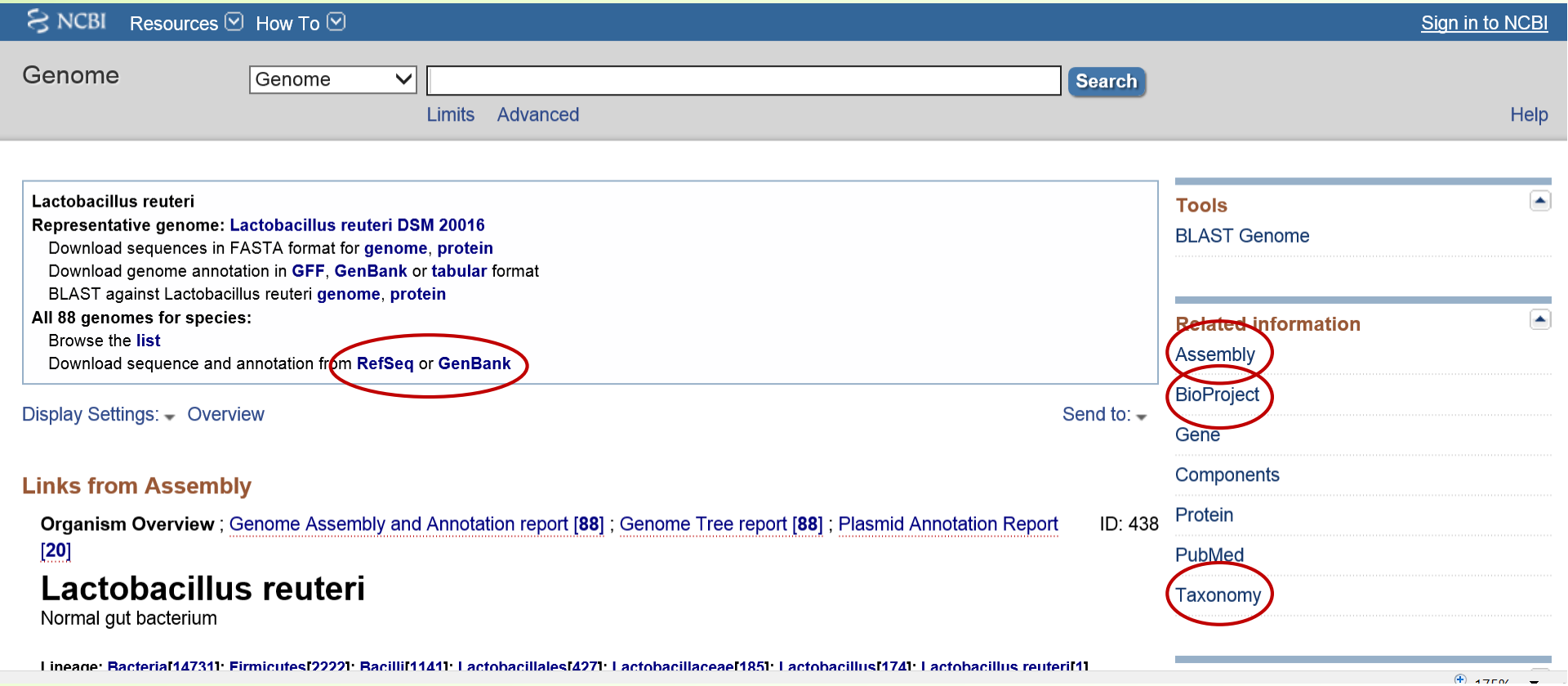
- **RefSeq** – maintained by the NCBI
  - a database of sequences of:
    - Genomic DNA – including genes and whole chromosomes, and thus whole prokaryote genome sequences
    - Transcripts
    - Proteins
  - Non-redundant, i.e.: **Single standard reference sequence for each gene, chromosome, transcript, protein**
  - Cross-referenced

Getting a genome

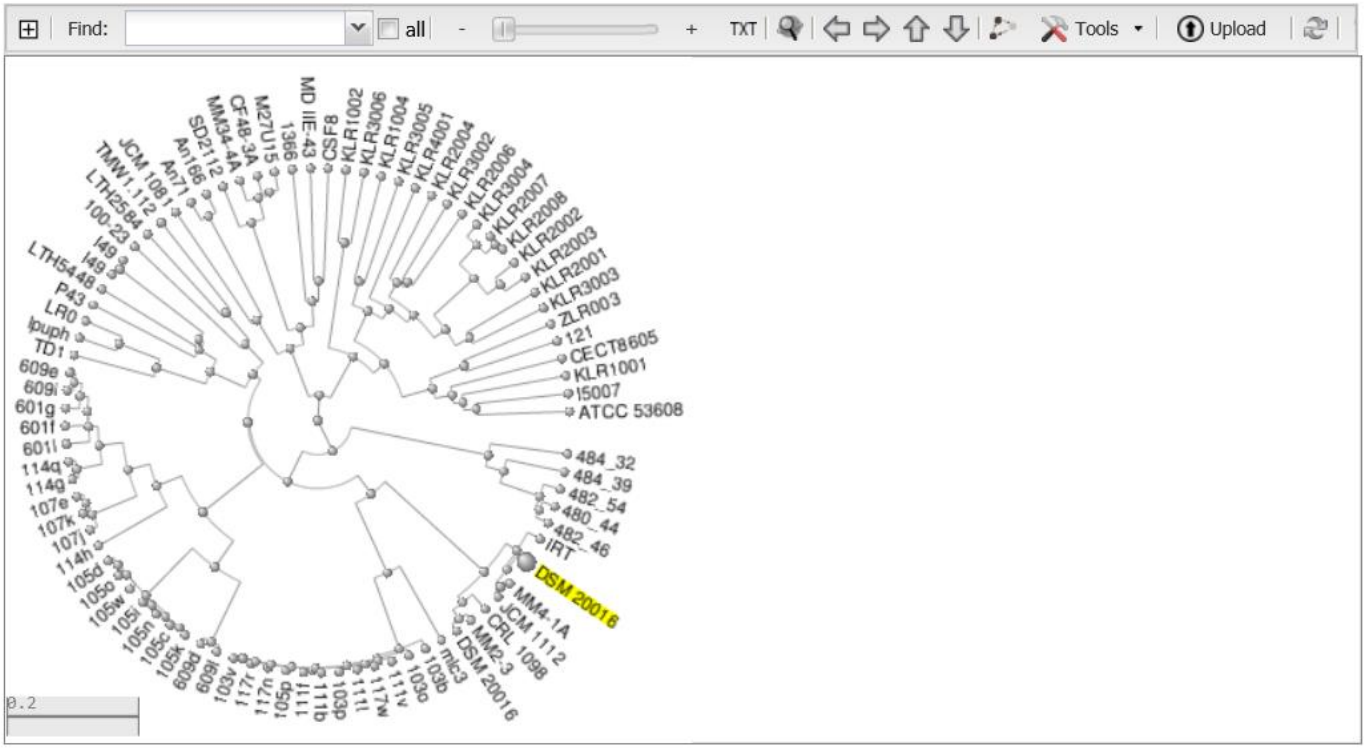
Example

Get the *Lactobacillus reuteri* genome sequences + annotations from NCBI

- *L. reuteri* has corresponding entries in Genome DB
- Each available strain has an entry in
  - BioProject database
  - Assembly database
  - Taxonomy database
  - Its genome sequence is available with annotations from Genbank
  - Alternatively from RefSeq (different files)


NCBI Resources How To
Sign in to NCBI
Genome
Genome
Search
Limits Advanced
Help
Lactobacillus reuteri
Representative genome: Lactobacillus reuteri DSM 20016
Download sequences in FASTA format for genome, protein
Download genome annotation in GFF, GenBank or tabular format
BLAST against Lactobacillus reuteri genome, protein
All 88 genomes for species:
Browse the list
Download sequence and annotation from RefSeq or GenBank
Tools
BLAST Genome
Related information
Assembly
BioProject
Gene
Components
Protein
PubMed
Taxonomy
Display Settings: Overview
Send to:
Links from Assembly
Organism Overview ; Genome Assembly and Annotation report [88] ; Genome Tree report [88] ; Plasmid Annotation Report [20]
ID: 438
Lactobacillus reuteri
Normal gut bacterium
Lineage: Bacteria[14731]; Firmicutes[2222]; Bacilli[1141]; Lactobacillales[427]; Lactobacillaceae[185]; Lactobacillus[174]; Lactobacillus reuteri[1]

Dendrogram (based on genomic BLAST)



Numbers of genomes available from **NCBI** FTP site, 19-9-2017.

Lower numbers of RefSeq genomes are not solely due to redundancy.

E.g. Bacteria, Archaea - see

https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/reannotation/

- Some genomes are suppressed, due to failing assembly/annotation quality checks

	GenBank	RefSeq
Bacteria	106,355	94,940
Archaea	1,260	676
Fungi	2,557	238
Plant	450	94
Protozoa	507	79
Invertebrate	682	149
Vertebrate – mammalian	298	112
Vertebrate - other	245	124
Viral	7,499	7,497
Other	4	N/A

NCBI GenBank genomes – some illustrative bacterial tallies

- September 2017
- Most bacterial **strains** are represented by a single genome. Some exceptions:
- *Escherichia*: 7,070 genomes
  - *E. coli:* 7,008
    - ***E. coli* strain K-12:** **30**
      - *E. coli* strain K-12 substrain MG1655: 11
    - *E. coli* O157:H6: 156
    - *E. coli* O104:H4: 58
- *Salmonella:* 7,595
  - *S. enterica:* 7,582
    - *S. enterica* subsp. *enterica:* 6,493
      - *S. enterica* subsp. *enterica* serovar Typhimurium: 953
        - *S. enterica* subsp. *enterica* serovar Typhimurium **str. DT104:** **364**
      - *S. enterica* subsp. *enterica* serovar Typhi: 2,000

These 2 species account for 14% of all the genomes

	NCBI GenBank genomes, 21 Sep. 2017
Streptococcus pneumoniae	8,289
Streptococcus agalactiae	947
Pseudomonas aeruginosa	2,487
Campylobacter coli	817
Campylobacter jejuni	1,085
Listeria monocytogenes	1,528
Helicobacter pylori	695
Yersinia pestis	321
Clostridium difficile*	1,090
Clostridium botulinum	202
Enterococcus faecalis	531
Bacteroides fragilis	115
Lactobacillus reuteri	89
Bifidobacterium bifidum	34
Akkermansia muciniphila	18

Ensembl:
project, system, data and browser

- *"The goal of Ensembl was … to **automatically annotate** the [human] genome, **integrate this annotation with other available biological data** and make all this **publicly available via the web**.*
- *Since the website's launch in July 2000, many more genomes have been added to Ensembl and the range of available data has also expanded to include comparative genomics, variation and regulatory data."*
- http://www.ensembl.org/info/about/index.html

Ensembl

- Original Ensembl project http://www.ensembl.org went on to encompass:
  - Other vertebrate genomes (now > 100)
  - 2 non-vertebrate chordates (sea squirts)
  - And genomes of 3 early non-vertebrate model organisms
    - Worm, Fly, Brewers' yeast
  - *Nucleic Acids Research*(2014) **42** D749-D755 doi: 10.1093/nar/gkt1196
- Since then:
  - Ensembl Genomes http://ensemblgenomes.org
    - Ensembl Bacteria http://bacteria.ensembl.org 43,552 genomes
      - (**database includes Archaea**; currently 494 genomes)
    - Ensembl Fungi http://fungi.ensembl.org 811
    - Ensembl Metazoa http://metazoa.ensembl.org 68
    - Ensembl Plants http://plants.ensembl.org 47
    - Ensembl Protists http://protists.ensembl.org 200

Ensembl:
project, system, data and browser

- Interactive graphical browsing
- Bulk downloads
  - All data for the genome, particular regions, etc
- Programmatic access to data (Application Programming Interface)
  - Perl
  - C (Ensembl API within EMBOSS AJAX library)
- Software tools available for your own use:
  - Ensembl annotation pipeline components
  - Ensembl website

Where did we go wrong…

- Abstract

"Motivation: It is only a matter of time until ***a user will see not many but one integrated database of information for molecular biology.***

Is this true?

Is it a good thing? Why will it happen? Where are we now?

What developments are fostering and what developments are impeding progress towards this end?"

- Frishman *et al.* (1998) Comprehensive, comprehensible, distributed and intelligent databases: current status, *Bioinformatics* **14**(7) 551-561

Next session

Friday 20th October, Barton Room

References

- Ensembl http://www.ensembl.org
  - Hubbard, T. *et al.* (2002) The Ensembl genome database project *Nucleic Acids Res.* **30** (1) 38-41
  - Yates, A. *et al.* (2016) Ensembl 2016 *Nucleic Acids Res.* **44** (D1) D710-D716
- Ensembl Genomes http://ensemblgenomes.org
  - Kersey, P.J. *et al.* (2010) Ensembl Genomes: Extending Ensembl across the taxonomic space, *Nucleic Acids Res.* **38** (suppl_1) D563-D569
  - Kersey, P.J. *et al.* (2016) Ensembl Genomes 2016: more genomes, more complexity *Nucleic Acids Res.* **44** (D1) D574-D580
- FungiDB http://fungidb.org
  - Stajich, J.E. *et al.* (2012) FungiDB: an integrated functional genomics database for fungi, *Nucleic Acids Res.* **40** (D1) D675-D681
- IMG/M – Integrated Microbial Genomes and Microbiome Samples http://img.jgi.doe.gov/m
  - Markowitz, V.M. *et al.* (2009) IMG/M: A data management and analysis system for metagenomes *Nucleic Acids Res.* **36** (D1) D534-D538

References

- NCBI Genomes https://www.ncbi.nlm.nih.gov/genome
- PATRIC http://www.patricbrc.org
  - Wattam, A.R. *et al.* (2014) PATRIC, the bacterial bioinformatics database and analysis resource, *Nucleic Acids Res.* **42** (D1) D581-D591
  - Wattam, A.R. *et al.* (2017) Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center, *Nucleic Acids Res.* **45** (D1) D535-D542
- RefSeqhttps://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
- ViPR http://www.viprbrc.org
  - Pickett *et al.* (2012) ViPR: an open bioinformatics database and analysis resource for virology research, *Nucleic Acids Res.* **40** (D1) D593-D598
  - Pickett *et al.* (2012) Virus Pathogen Database and Analysis Resource (ViPR): A Comprehensive Bioinformatics Database and Analysis Resource for the Coronavirus Research Community *Viruses* **4** (11) 3209-3226