
FAST
DATABASE
SEARCHES

Overview

- The challenges
  - So many sequences!
  - Different types of information
  - Constantly updating databases

- Local search methods
  - BLAST: seeded searches
    - Plain old BLAST
    - Discontiguous MEGABLAST!!!
    - PSI-BLAST

Sequence Databases

Store several different types of sequence data:

DNA sequences

(nanopore signal data, individual reads, individual genes, genome fragments, complete genomes)

Protein sequences

Usually inferred from corresponding gene sequence

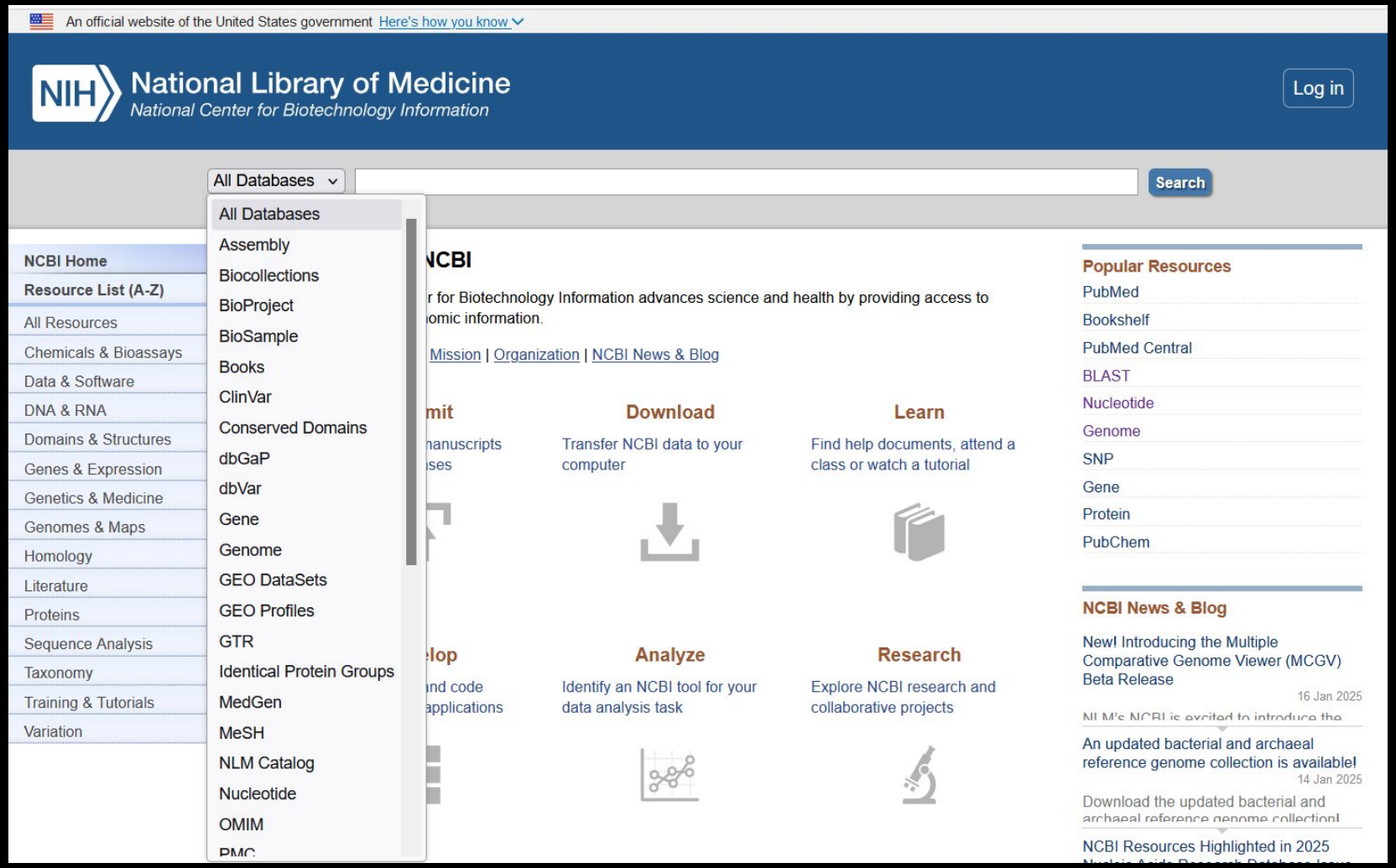
RNA sequences

(nanopore signal data, individual reads, expressed sequence tags, transcripts)

Considerations

- Data type, size and provenance
- Modes of access: queries, browsing, APIs
- Documentation / stability / support / persistence
- Reliability of information
- Data Use regulations / agreement

National Center for Biotechnology Information (NCBI)



Reference genomes,
Gene sequences,
Taxonomy,
ESTs
Journal articles
(etc…)

https://www.ncbi.nlm.nih.gov/datasets/genome/?taxon=561

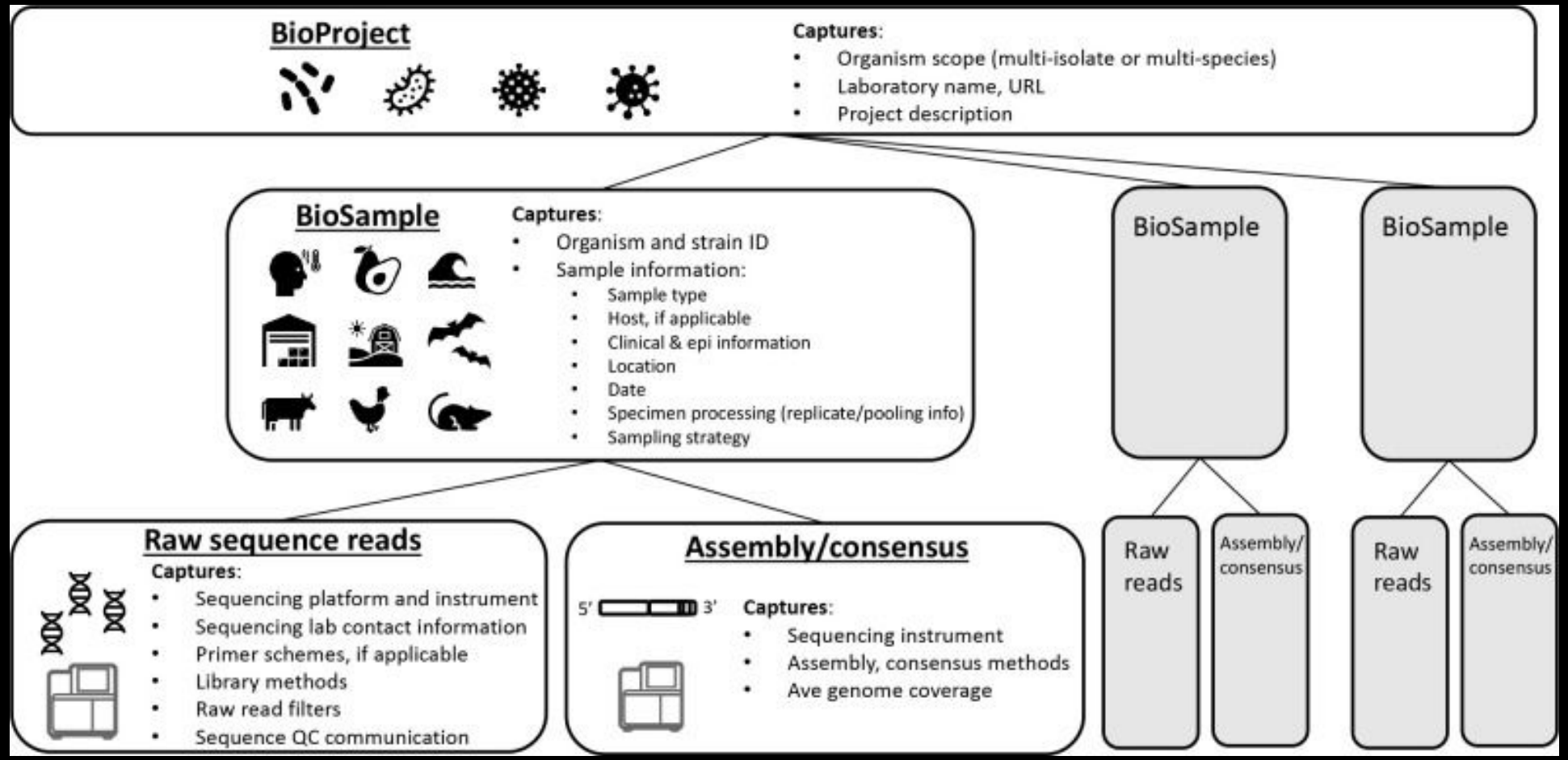
Filters

REST-like URL

Download | Select columns | 314,368 Genomes | Rows per page 20 | 1-20 of 314,368

Assembly	GenBank	RefSeq	Scientific name	Modifier	Annotation	Action
ASM2009747v1	GCA_020097475.1	GCF_020097475.1	Escherichia fergusonii	FDAARGOS_1499 (...	NCBI RefSeq Submitter	
ASM584v2	GCA_000005845.2	GCF_000005845.2	Escherichia coli str. K-12 substr...	K-12 substr. MG165...	NCBI RefSeq Submitter	
ASM886v2	GCA_000008865.2	GCF_000008865.2	Escherichia coli O157:H7 str. S...	Sakai substr. RIMD ...	NCBI RefSeq Submitter	
ASM2862233v1	GCA_028622335.1	GCF_028622335.1	Escherichia albertii	BIA_5-2 (strain)	NCBI RefSeq Submitter	
ASM2996246v1	GCA_029962465.1	GCF_029962465.1	Escherichia marmotae	YF8 (strain)	NCBI RefSeq Submitter	
ASM1483671v1	GCA_014836715.1	GCF_014836715.1	Escherichia whittamii	Sa2BVA5 (strain)	NCBI RefSeq Submitter	
ASM3132397v1	GCA_031323975.1	GCF_031323975.1	Escherichia ruysiae	AB136 (strain)	NCBI RefSeq Submitter	
ASM285371v1	GCA_002853715.1	GCF_002853715.1	Escherichia coli (E. coli)	14EC020 (strain)	NCBI RefSeq Submitter	
ASM1326v1	GCA_000013265.1	GCF_000013265.1	Escherichia coli UTI89	UTI89 (strain)	NCBI RefSeq Submitter	

NCBI (Pathogen) Data Object Model



EMBL-EBI

Unleashing the potential of big data in biology

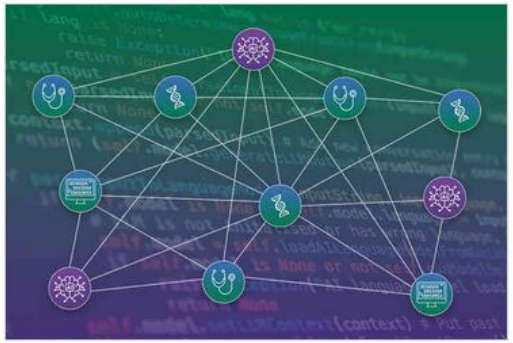
Find a gene, protein or chemical | All | Search

Example searches: blast keratin bfl1 | About EBI Search

Nucleotides,
Genomes,
Protein function,
Protein-protein interactions

Find data resources | Submit data | Explore our research | Train with us

Latest news



BioChatter: making large language models accessible for biomedical research

Researchers uncover what drives aggressive bone cancer

DECIPHER v11.29 released

09 Jan 2025

Fuelling discovery together: 2024 user survey learnings

European Molecular Biology Laboratory – European Bioinformatics Institute (EBI)

Example Record – "BlaZ"

UniProt BLAST Align Peptide search ID mapping SPARQL UniProtKB Advanced | List Search

Function
Names & Taxonomy
Subcellular Location
Phenotypes & Variants
PTM/Processing
Expression
Interaction
Structure
Family & Domains
Sequence
Similar Proteins

P06548 · BLA3_BACCE

Protein	Beta-lactamase 3	Amino acids	316 (go to sequence)
Gene	blaZ	Protein existence	Evidence at protein level
Status	UniProtKB reviewed (Swiss-Prot)	Annotation score	3/5
Organism	Bacillus cereus		

Entry | Variant viewer | Feature viewer | Genomic coordinates | Publications | External links | History

BLAST | Download | Add | Add a publication | Entry feedback

Function

Catalytic activity

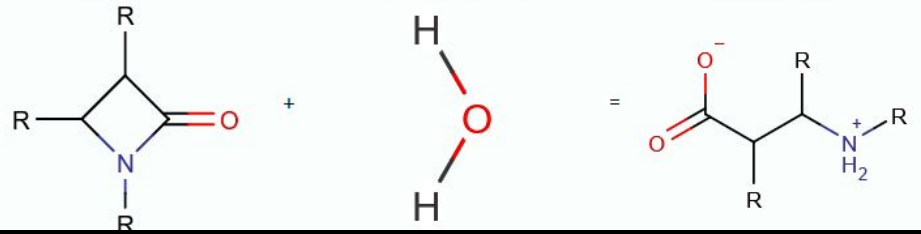
Rhea 20401

a beta-lactam + H2O = a substituted beta-amino acid PROSITE-ProRule Annotation

EC:3.5.2.6 (UniProtKB | ENZYME | Rhea)

Hide Rhea reaction

a β-lactam	H_2O	a substituted β-amino acid
CHEBI:35627	CHEBI:15377	CHEBI:140347



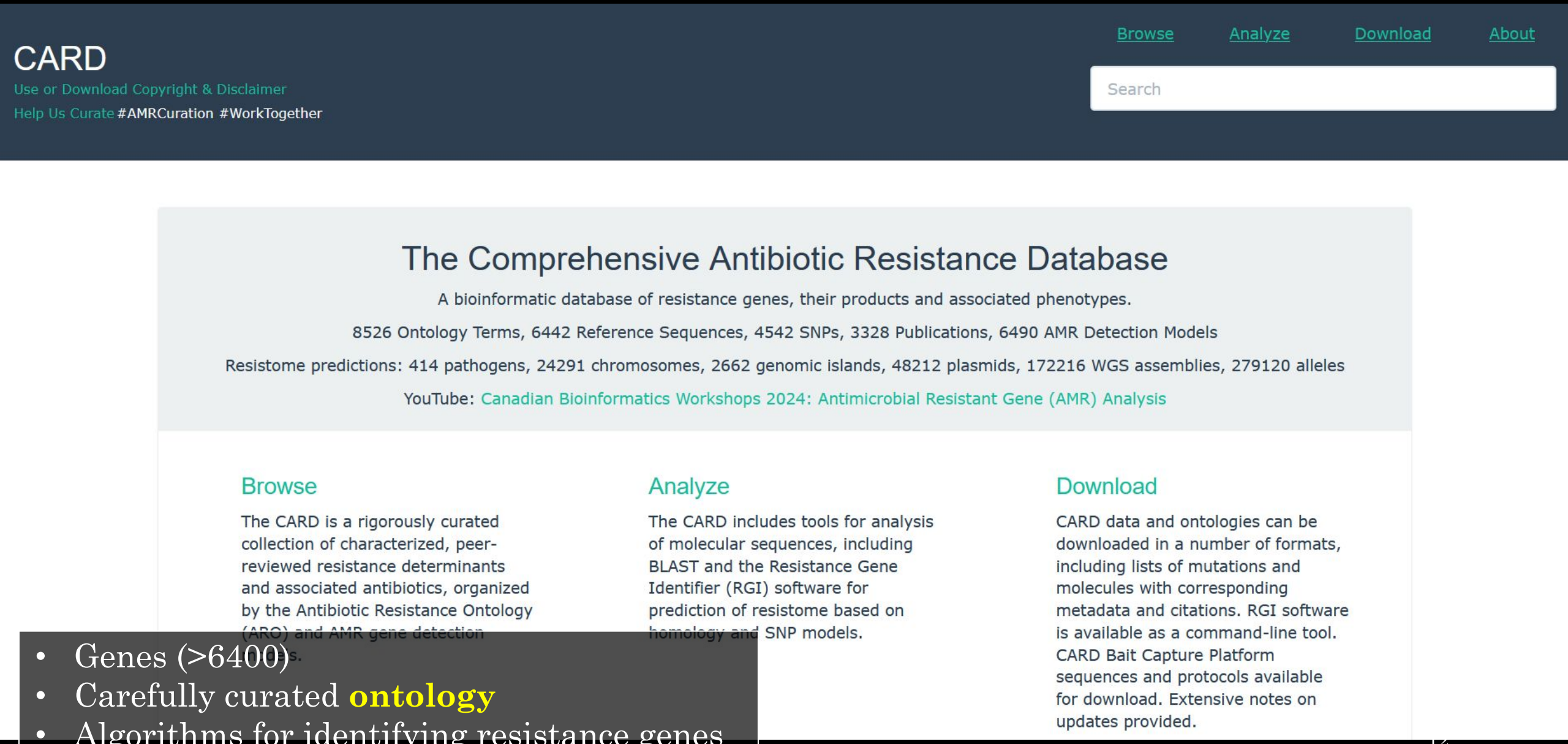
International Nucleotide Sequence Database Consortium

+ China National Center for Bioinformation (CNCB) - sort of
+ African node via AfricaCDC?

Challenge of unrestricted open data

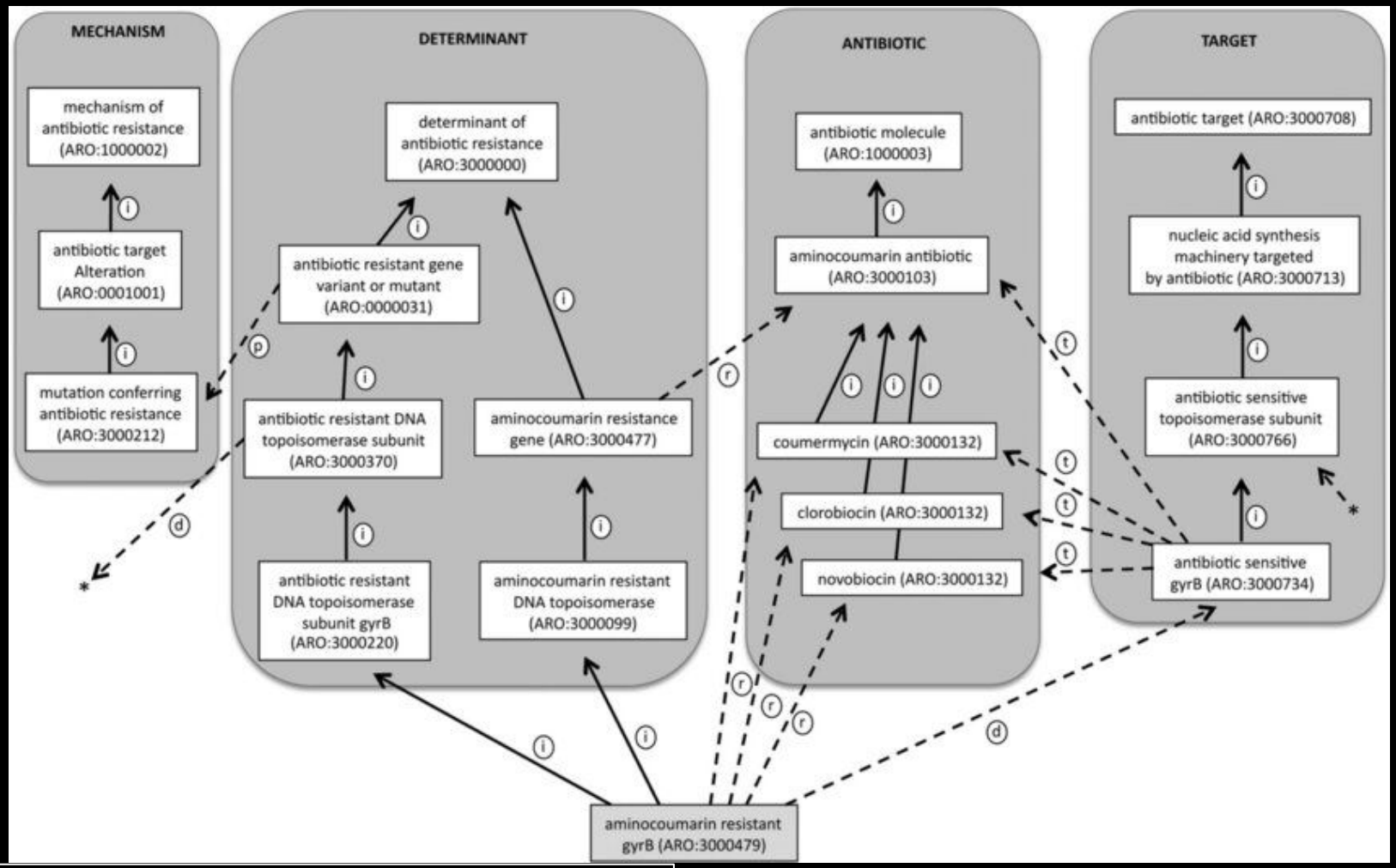
- Human Data
  - Anonymisation of individual records challenging
  - Medical ethics (consent/withdrawal of consent)
  - Commercial use/abuse
- Pathogen Data
  - Tension between public health and academic incentives
  - Commercial use/abuse e.g.,Indonesia Influenza Vaccine)
- Access and Benefit Sharing of Biodiversity Resources (CBD)
- Maintaining Databases
  - Expensive to maintain resources
  - Funding challenging to receive

The Comprehensive Antibiotic Resistance Database (CARD)



- Genes (>6400)
- Carefully curated **ontology**
- Algorithms for identifying resistance genes

The Comprehensive Antibiotic Resistance Database (CARD)



- Genes (>6400)
- Carefully curated **ontology**
- Algorithms for identifying resistance genes

BlaZ beta-lactamase [AMR Gene Family]

Download Sequences

Accession	ARO:3004197
Definition	BlaZ beta-lactamases are Class A beta-lactamases. These beta-lactamases are responsible for penicillin resistance in Staphylococcus aureus.
Drug Class	penicillin beta-lactam
Resistance Mechanism	antibiotic inactivation
Classification	**10 ontology terms** | Hide + process or component of antibiotic biology or chemistry + mechanism of antibiotic resistance + determinant of antibiotic resistance + antibiotic inactivation [Resistance Mechanism] + antibiotic inactivation enzyme + hydrolysis of antibiotic conferring resistance + antibiotic molecule + hydrolysis of beta-lactam antibiotic by serine beta-lactamase + beta-lactam antibiotic + beta-lactamase
Parent Term(s)	**2 ontology terms** | Hide + *confers_resistance_to_drug_class* penicillin beta-lactam [Drug Class] + class A beta-lactamase
Sub-Term(s)	**3 ontology terms** | Hide + PC1 beta-lactamase (blaZ) + mecC-type BlaZ + PC1
Publications	McLaughlin JR, et al. 1981. J Biol Chem 256(21): 11283-11291. Unique features in the ribosome binding site sequence of the gram-positive Staphylococcus aureus beta-lactamase gene. (PMID 6793593) Pence MA, et al. 2015. PLoS ONE 10(8):e0136605 Beta-Lactamase Repressor BlaI Modulates Staphylococcus aureus Cathelicidin Antimicrobial Peptide Resistance and Virulence. (PMID 26305782)

Antimicrobial Resistance Gene Predictions (*Klebsiella pneumoniae* plasmid)

Summary (summary counts and figures only include Loose hits of e-10 or better)

Filename	Date (UTC)	RGI Criteria	# Perfect Hits	# Strict Hits	# Loose Hits	Download
JN420336.1	February 01, 2024 15:02:45	Perfect, Strict, complete genes only	5	1	0	Download

Results (all Loose hits shown)

Search:

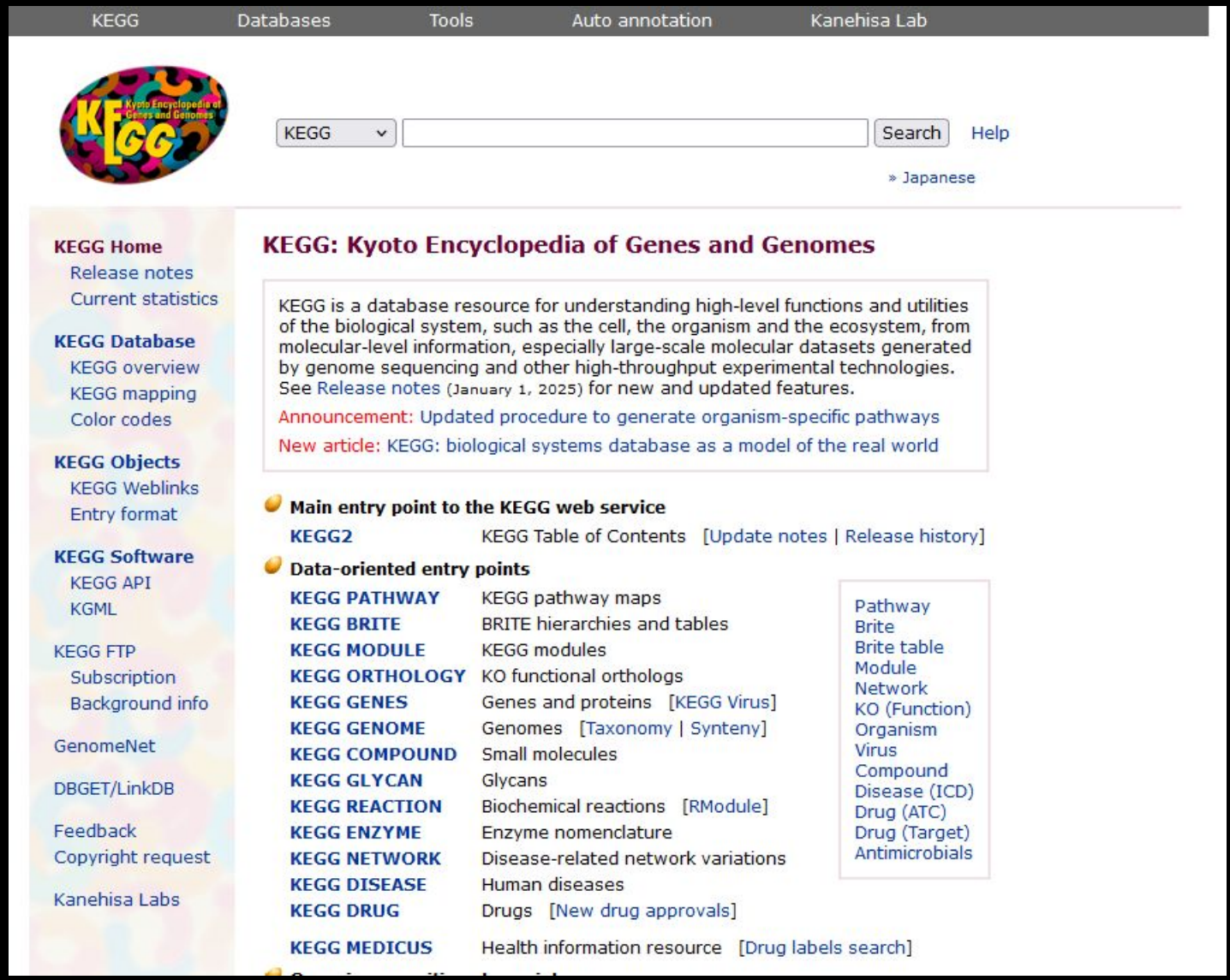
RGI Criteria	ARO Term	SNP	Detection Criteria	AMR Gene Family	Drug Class	Resistance Mechanism	% Identity of Matching Region	% Length of Reference Sequence
Perfect	OXA-1		protein homolog model	OXA beta-lactamase	carbapenem, cephalosporin, penam	antibiotic inactivation	100.0	105.43
Perfect	NDM-1		protein homolog model	NDM beta-lactamase	carbapenem, cephalosporin, cephamycin, penam	antibiotic inactivation	100.0	100.00
Perfect	QnrB1		protein homolog model	quinolone resistance protein (qnr)	fluoroquinolone antibiotic	antibiotic target protection	100.0	100.00
Perfect	catA1		protein homolog model	chloramphenicol acetyltransferase (CAT)	phenicol antibiotic	antibiotic inactivation	100.0	100.00
Perfect	CTX-M-15		protein homolog model	CTX-M beta-lactamase	cephalosporin, penam	antibiotic inactivation	100.0	100.00
Strict	AAC(6')-Ib-cr6		protein homolog model	AAC(6'), AAC(6')-Ib-cr	fluoroquinolone antibiotic, aminoglycoside antibiotic	antibiotic inactivation	98.99	100.00

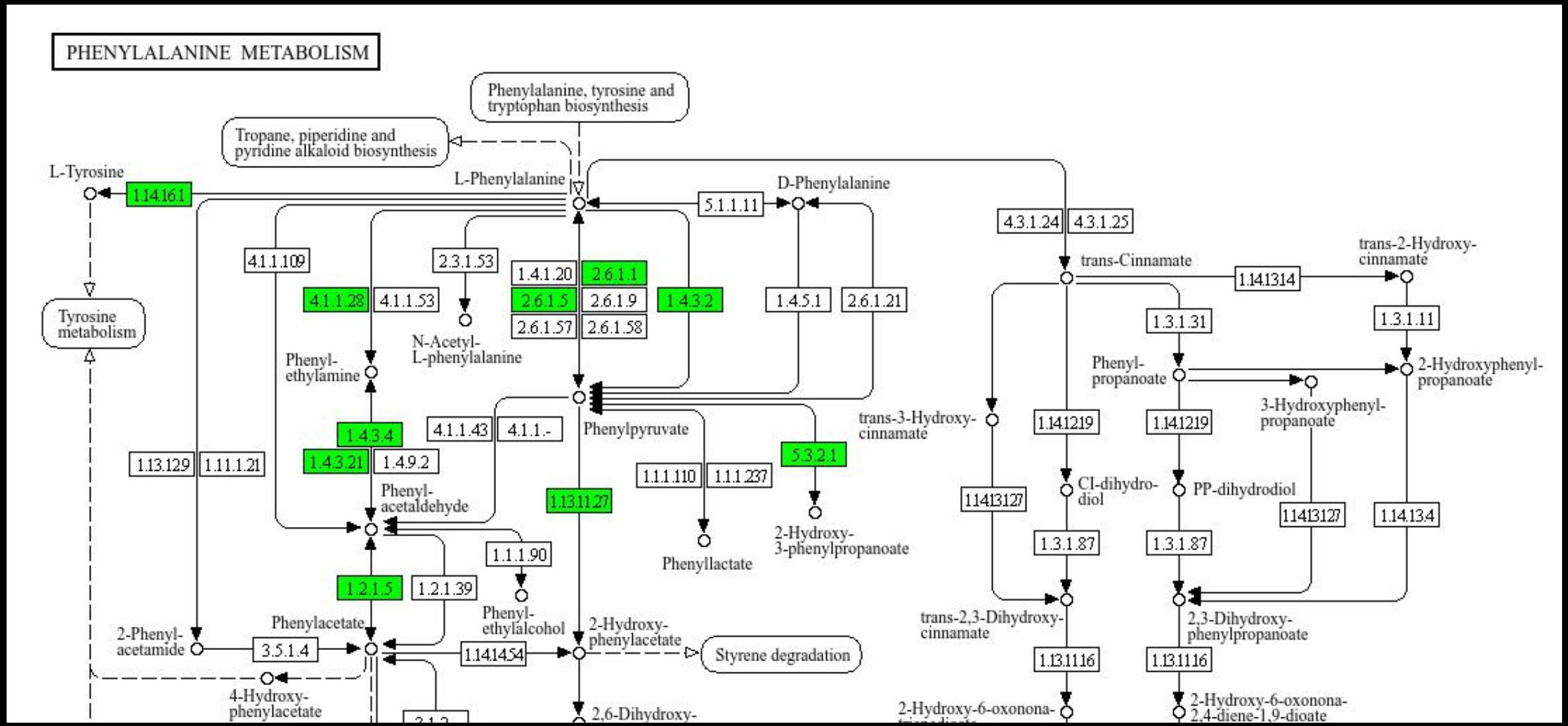
Previous 1 Next

Kyoto Encyclopedia of Genes and Genomes

Genomes
Orthology information
Protein functions
Biochemical pathways

Limited access now





Phenylalanine metabolism (GREEN = active in primates)

https://www.genome.jp/pathway/mapno=00360&category=Primates

A word about “metadata”

PERSPECTIVE

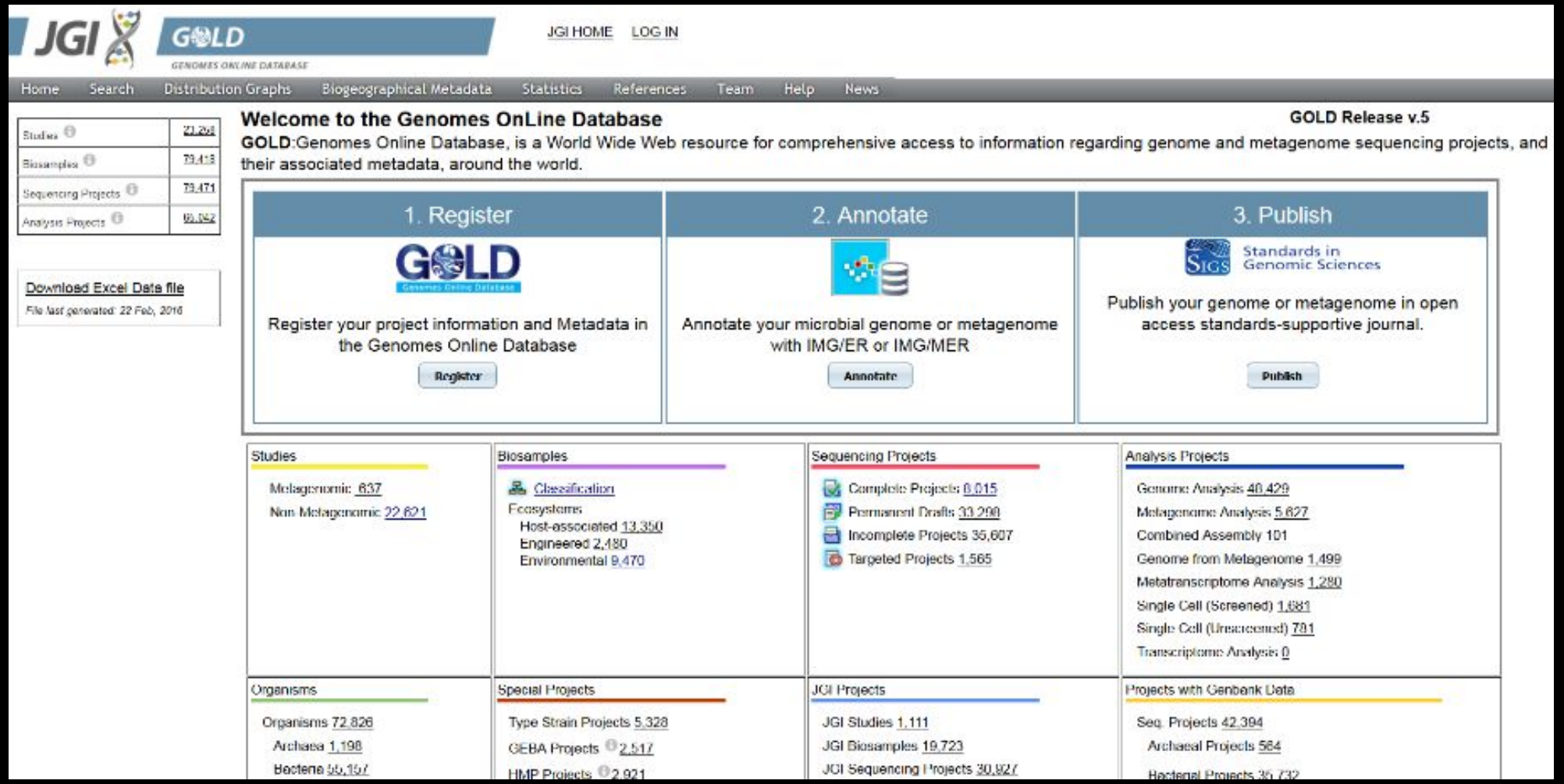
nature biotechnology

Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications

Pelin Yilmaz[1,2*], Renzo Kottmann[1], Dawn Field[3], Rob Knight[4,5], James R Cole[6,7], Linda Amaral-Zettler[8], Jack A Gilbert[9–11], Ilene Karsch-Mizrachi[12], Anjanette Johnston[12], Guy Cochrane[13], Robert Vaughan[13], Christopher Hunter[13], Joonhong Park[14], Norman Morrison[3,15], Philippe Rocca-Serra[16], Peter Sterk[3], Manimozhiyan Arumugam[17], Mark Bailey[3], Laura Baumgartner[18], Bruce W Birren[19], Martin J Blaser[20], Vivien Bonazzi[21], Tim Booth[3], Peer Bork[17], Frederic D Bushman[22], Pier Luigi Buttigieg[1,2], Patrick S G Chain[7,23,24], Emily Charlson[22], Elizabeth K Costello[4], Heather Huot-Creasy[25], Peter Dawyndt[26], Todd DeSantis[27], Noah Fierer[28], Jed A Fuhrman[29], Rachel E Gallery[30], Dirk Gevers[19], Richard A Gibbs[31,32], Inigo San Gil[33], Antonio Gonzalez[34], Jeffrey I Gordon[35], Robert Guralnick[28,36], Wolfgang Hankeln[1,2], Sarah Highlander[31,37], Philip Hugenholtz[38], Janet Jansson[23,39], Andrew L Kau[35], Scott T Kelley[40], Jerry Kennedy[4], Dan Knights[34], Omry Koren[41], Justin Kuczynski[18], Nikos Kyrpides[23], Robert Larsen[4], Christian L Lauber[42], Teresa Legg[28], Ruth E Ley[41], Catherine A Lozupone[4], Wolfgang Ludwig[43], Donna Lyons[42], Eamonn Maguire[16], Barbara A Methé[44], Folker Meyer[10], Brian Muegge[35], Sara Nakielny[4], Karen E Nelson[44], Diana Nemergut[45], Josh D Neufeld[46], Lindsay K Newbold[3], Anna E Oliver[3], Norman R Pace[18], Giriprakash Palanisamy[47], Jörg Peplies[48], Joseph Petrosino[31,37], Lita Proctor[21], Elmar Pruesse[1,2], Christian Quast[1], Jeroen Raes[49], Sujeevan Ratnasingham[50], Jacques Ravel[25], David A Relman[51,52], Susanna Assunta-Sansone[16], Patrick D Schloss[53], Lynn Schriml[25], Rohini Sinha[22], Michelle I Smith[35], Erica Sodergren[54], Aymé Spor[41], Jesse Stombaugh[4], James M Tiedje[7], Doyle V Ward[19], George M Weinstock[54], Doug Wendel[4], Owen White[25], Andrew Whiteley[3], Andreas Wilke[10], Jennifer R Wortman[25], Tanya Yatsunenko[35] & Frank Oliver Glöckner[1,2]

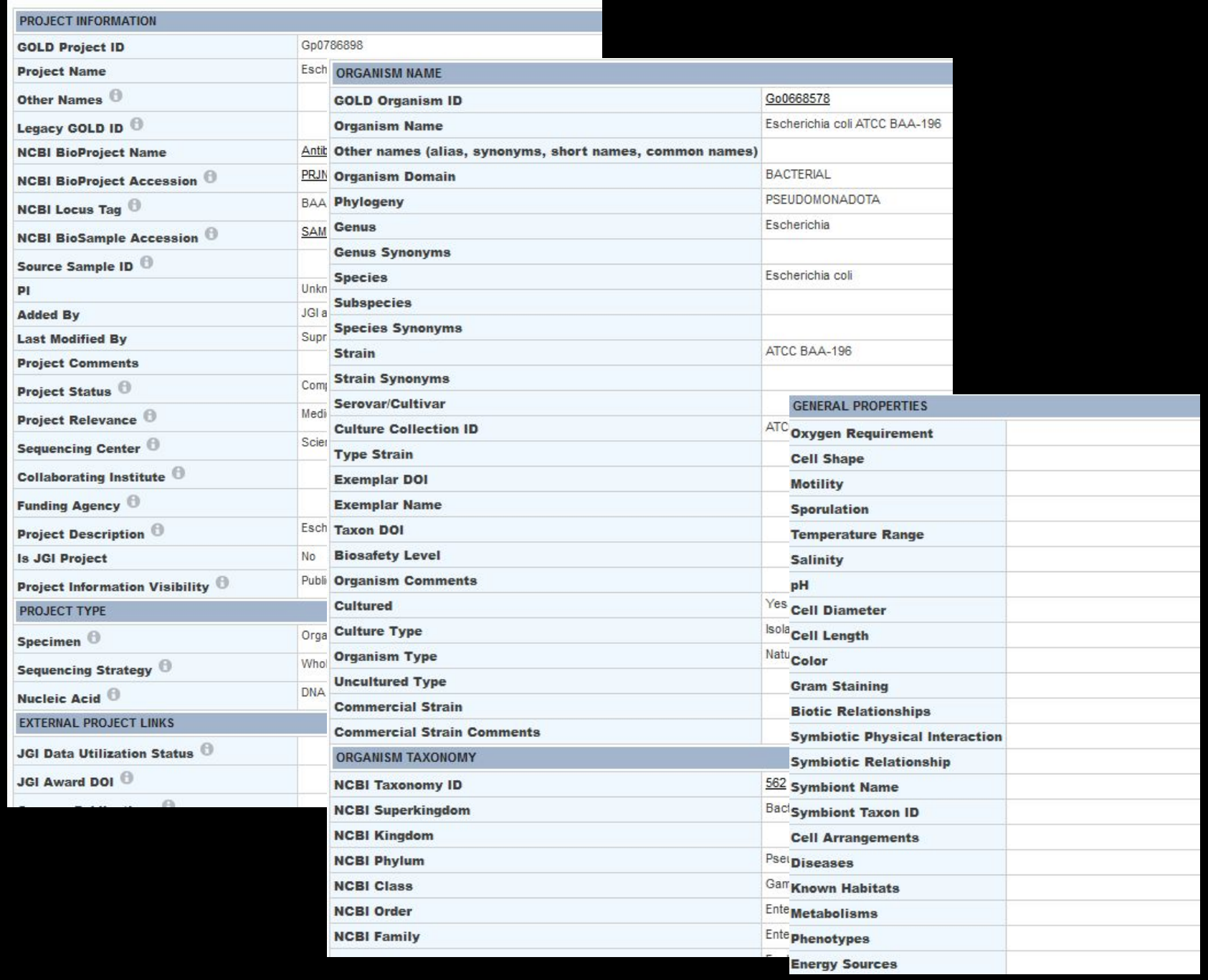
(2011)

Genomes Online Database (GOLD)



Genome projects

Standards-compliant metadata

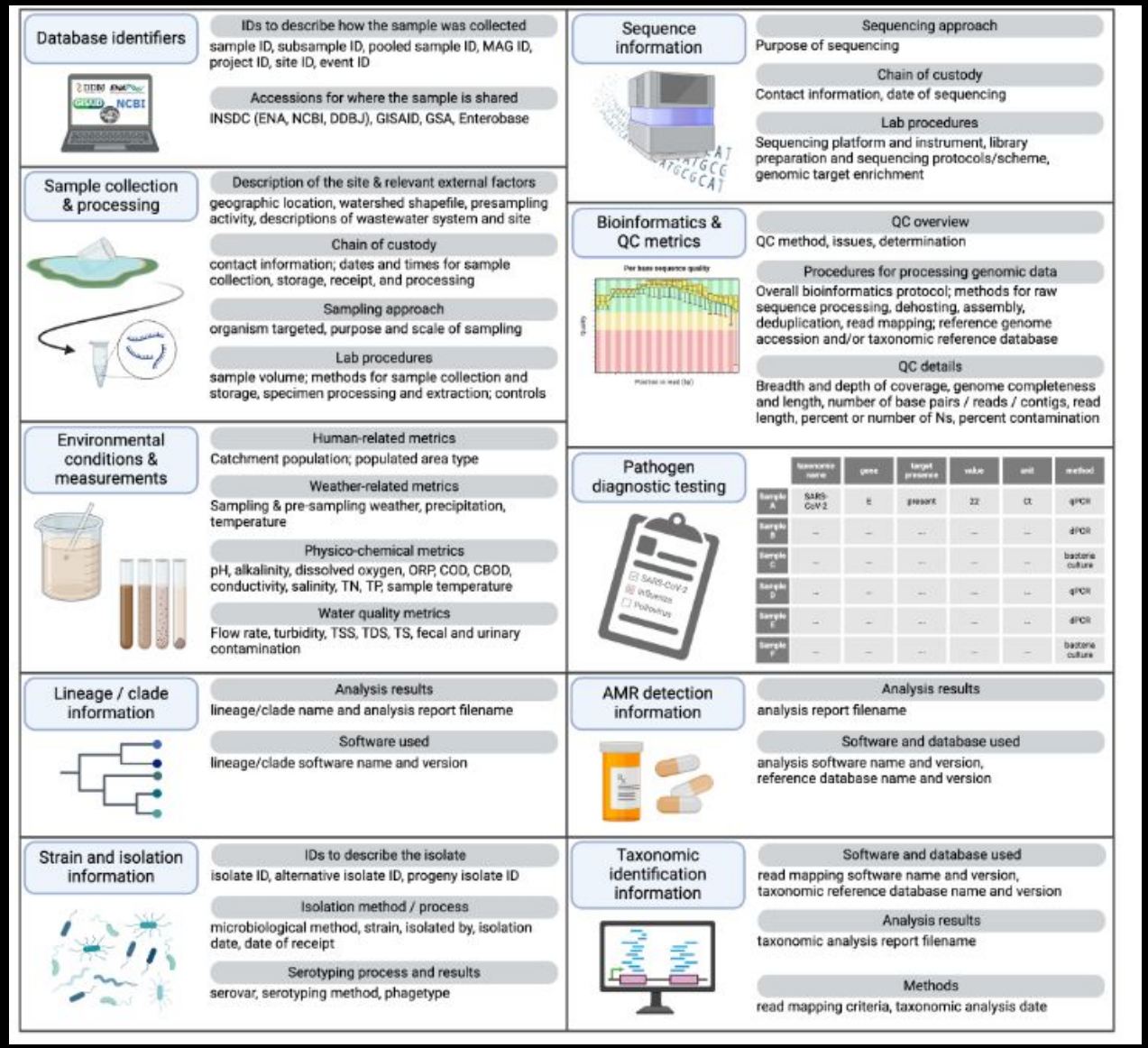


PROJECT INFORMATION	
GOLD Project ID	Gp0786898
Project Name	Esch
Other Names	
Legacy GOLD ID	
NCBI BioProject Name	Antib
NCBI BioProject Accession	PRJN
NCBI Locus Tag	BAA
NCBI BioSample Accession	SAM
Source Sample ID	
PI	Unkn
Added By	JGI a
Last Modified By	Supr
Project Comments	
Project Status	Comp
Project Relevance	Medi
Sequencing Center	Scier
Collaborating Institute	
Funding Agency	
Project Description	Esch
Is JGI Project	No
Project Information Visibility	Publi
PROJECT TYPE	
Specimen	Orga
Sequencing Strategy	Whol
Nucleic Acid	DNA
EXTERNAL PROJECT LINKS	
JGI Data Utilization Status	
JGI Award DOI	

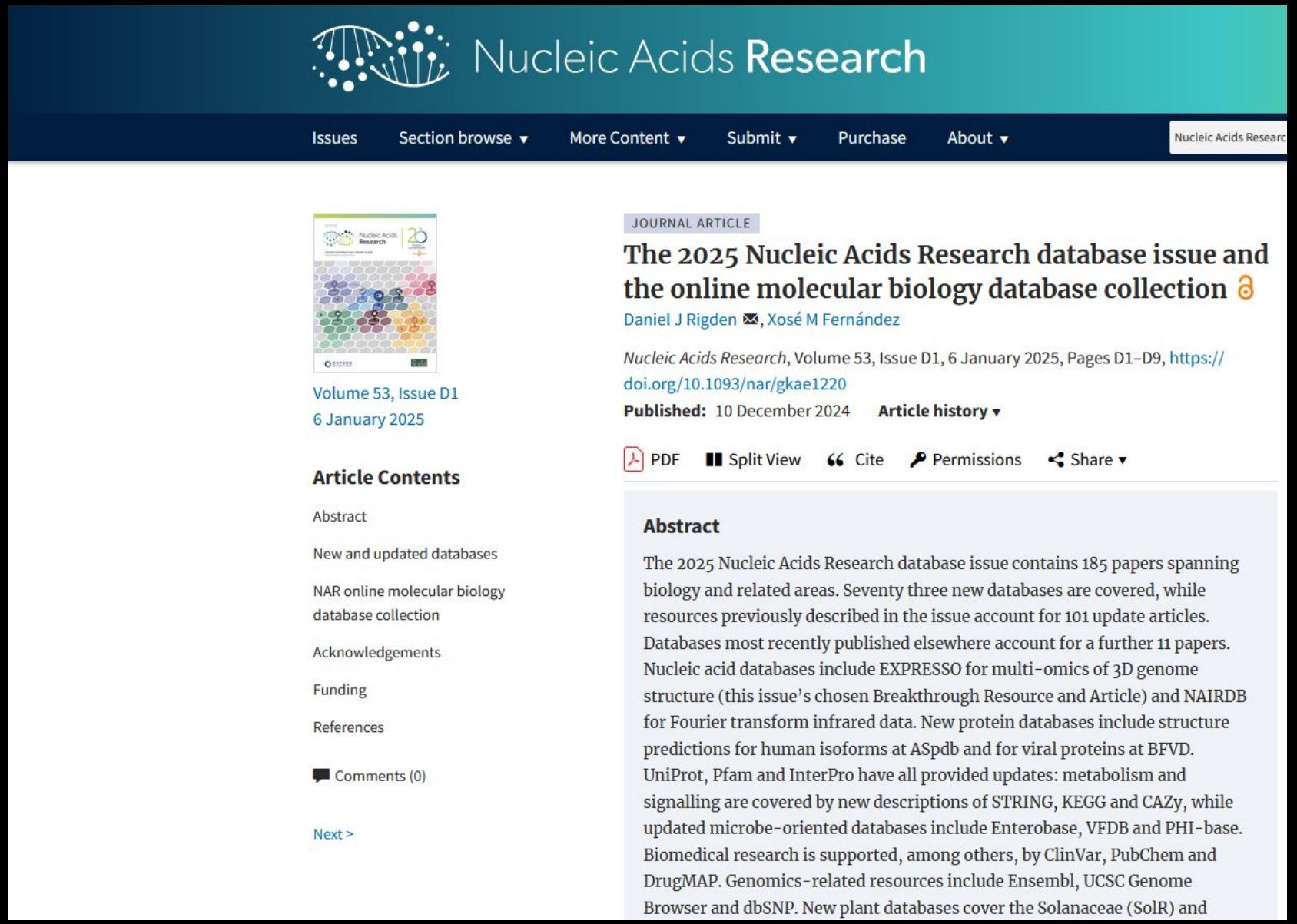
ORGANISM NAME	
GOLD Organism ID	Go0668578
Organism Name	Escherichia coli ATCC BAA-196
Other names (alias, synonyms, short names, common names)	
Organism Domain	BACTERIAL
Phylogeny	PSEUDOMONADOTA
Genus	Escherichia
Genus Synonyms	
Species	Escherichia coli
Subspecies	
Species Synonyms	
Strain	ATCC BAA-196
Strain Synonyms	
Serovar/Cultivar	
Culture Collection ID	ATC
Type Strain	
Exemplar DOI	
Exemplar Name	
Taxon DOI	
Biosafety Level	
Organism Comments	
Cultured	Yes
Culture Type	Isola
Organism Type	Natu
Uncultured Type	
Commercial Strain	
Commercial Strain Comments	
ORGANISM TAXONOMY	
NCBI Taxonomy ID	562
NCBI Superkingdom	Bact
NCBI Kingdom	
NCBI Phylum	Pseu
NCBI Class	Gam
NCBI Order	Ente
NCBI Family	Ente

GENERAL PROPERTIES	
Oxygen Requirement	
Cell Shape	
Motility	
Sporulation	
Temperature Range	
Salinity	
pH	
Cell Diameter	
Cell Length	
Color	
Gram Staining	
Biotic Relationships	
Symbiotic Physical Interaction	
Symbiotic Relationship	
Symbiont Name	
Symbiont Taxon ID	
Cell Arrangements	
Diseases	
Known Habitats	
Metabolisms	
Phenotypes	
Energy Sources	

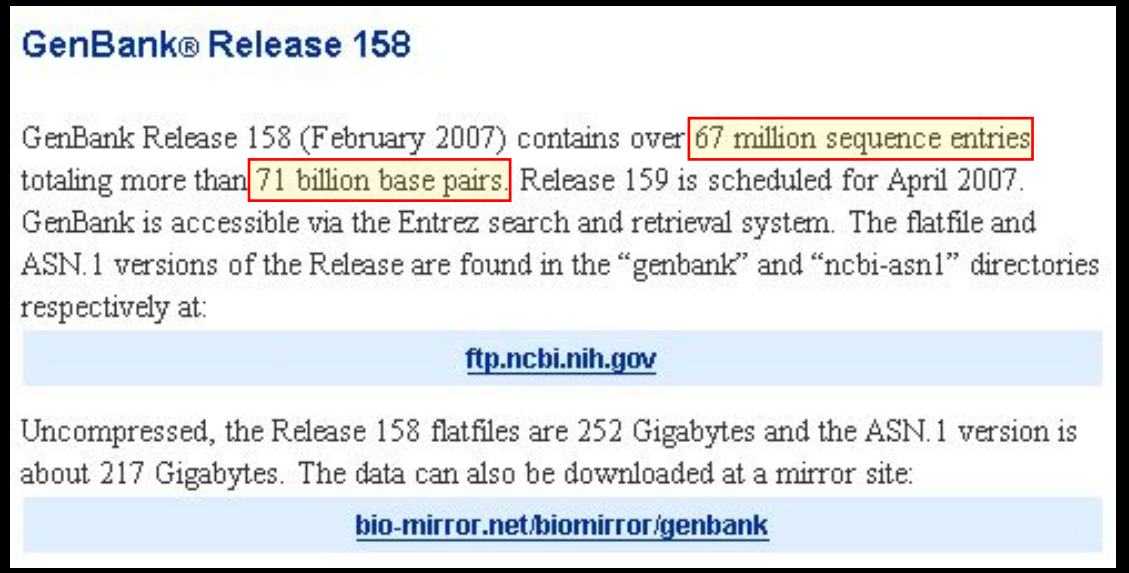
A word about “metadata”



Nucleic Acids Research Database Issue



These databases are *huge*



GenBank® Release 158

GenBank Release 158 (February 2007) contains over 67 million sequence entries totaling more than 71 billion base pairs. Release 159 is scheduled for April 2007. GenBank is accessible via the Entrez search and retrieval system. The flatfile and ASN.1 versions of the Release are found in the "genbank" and "ncbi-asn1" directories respectively at:

ftp.ncbi.nih.gov

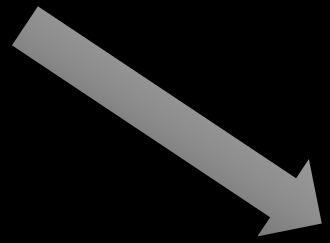
Uncompressed, the Release 158 flatfiles are 252 Gigabytes and the ASN.1 version is about 217 Gigabytes. The data can also be downloaded at a mirror site:

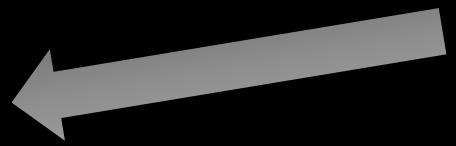
bio-mirror.net/biomirror/genbank

Release 3 (December 1982): 680,338 bases, from 606 sequences
Release 182 (February 2011): 124,277,818,310 bases, from 132,015,054 sequences
Release 188 (February 2012): 137,384,889,783 bases, from 149,819,246 sequences
Release 200 (February 2014): 157,943,793,171 bases, from 171,123,749 sequences
Release 223 (December 2017): 249,722,163,594 bases, from 206,293,625 sequences
Whole-genome shotgun: > 500,000,000,000 bases
Release 236 (December 2019): 399,376,854,872 bases, from 216,214,215 sequences
Whole-genome shotgun: 7,323,655,233,013 bases
Release 240 (October 2020): 698,688,094,046 bases from 219,055,207 sequences
Whole-genome shotgun: 9,627,627,030,647 bases
Release 246 (October 2021): 1,014,763,752,113 bases from 233,642,893 sequences
Whole-genome shotgun: 15,089,161,465,959 bases
Release 258 (October 2023): 2,433,391,164,875 bases from 233,642,893 sequences
Whole-genome shotgun: 24,310,993,199,448 bases
Release 269 (December 2025): 6,651,459,875,408 bases from 259,677,058 sequences
Whole-genome shotgun: 43,082,971,215,013 bases

Sequence of Interest…

mystery gene





<u>Homologous</u> sequences:
-Evolutionary conservation
-Annotated functions
-Presence / absence in other organisms
(phylogenetic profiles)

The Exact Approach

Use exact local alignment (i.e., Smith-Waterman) to find optimal matches between query sequence and all database sequences

This is impractical given S-W complexity (although hardware and software speedups exist)

We need heuristics!

What we *really* need

- Search methods that are not necessarily perfect, but maintain high levels of sensitivity and specificity relative to S-W

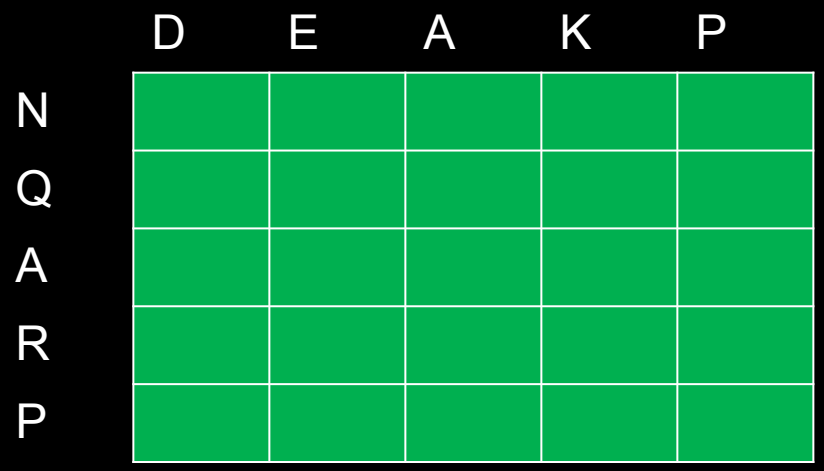
- Statistics to tell us when observed similarities are likely to be significant
  - the expectation value – how many matches to the database are expected by chance?

An important tradeoff…

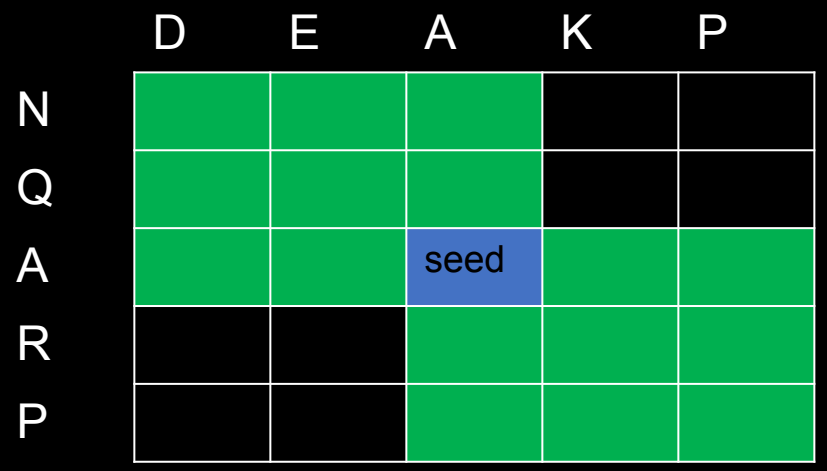
NQARP

DEAKP

Score each pair of residues – consider every possible alignment



Require an exact or at least awesome match of length L to "seed" the alignment and restrict the search space



FASTA

- Define the *ktup* parameter, which is the minimum length of exact match needed to seed an alignment

- Nucleotides: *ktup* typically 4-6
- Amino acids: *ktup* 1-2

Pearson and Lipman (1988) *Proc Natl Acad Sci*

FASTA uses a lookup table to store k-tuple values

NQARP

AR	3
NQ	1
QA	2
RP	4

DQATS

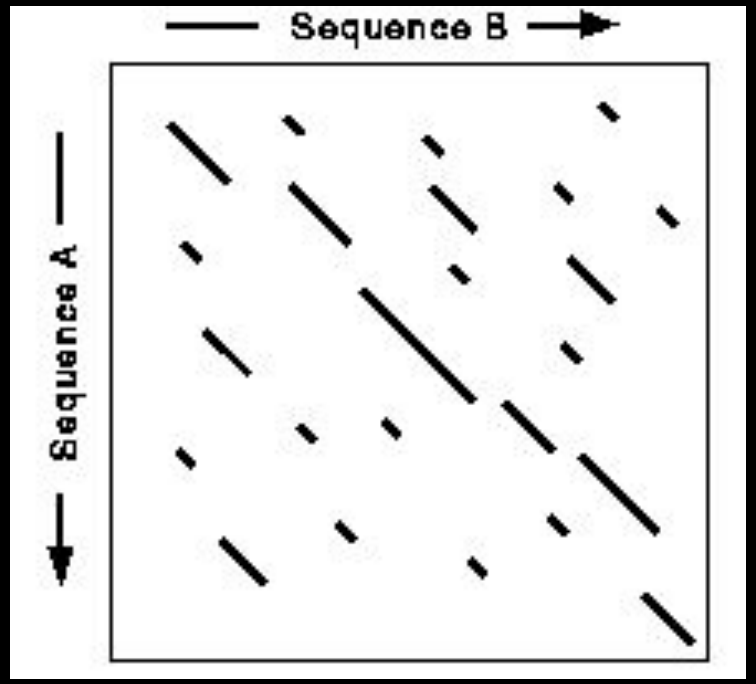
AT	3
DQ	1
QA	2
TS	4

Offset = start(QA, NQARP) – start(QA, DQATS) = 0

This is a k-mer decomposition!

Find ‘diagonals’ (no gaps!) in the sequence plot that have a high proportion of matching *k*-tuples

(PAM250 is used to weight matches of different *k*-tuples)



W+W = high score
A+A = low score

Additional steps: choose and rescore best diagonals
Statistics: randomization approach (many replicates)

BLAST

- **B**asic **L**ocal **A**lignment **S**earch **T**ool
- FASTA isn't fast enough!
- Can we trade away small amounts of optimality for further gains in performance?

Altschul et al. (1990) *J Mol Biol*
Altschul et al. (1997) *Nucleic Acids Res*

Basic Principles of BLAST

- Exact matches are great, but potentially too stringent

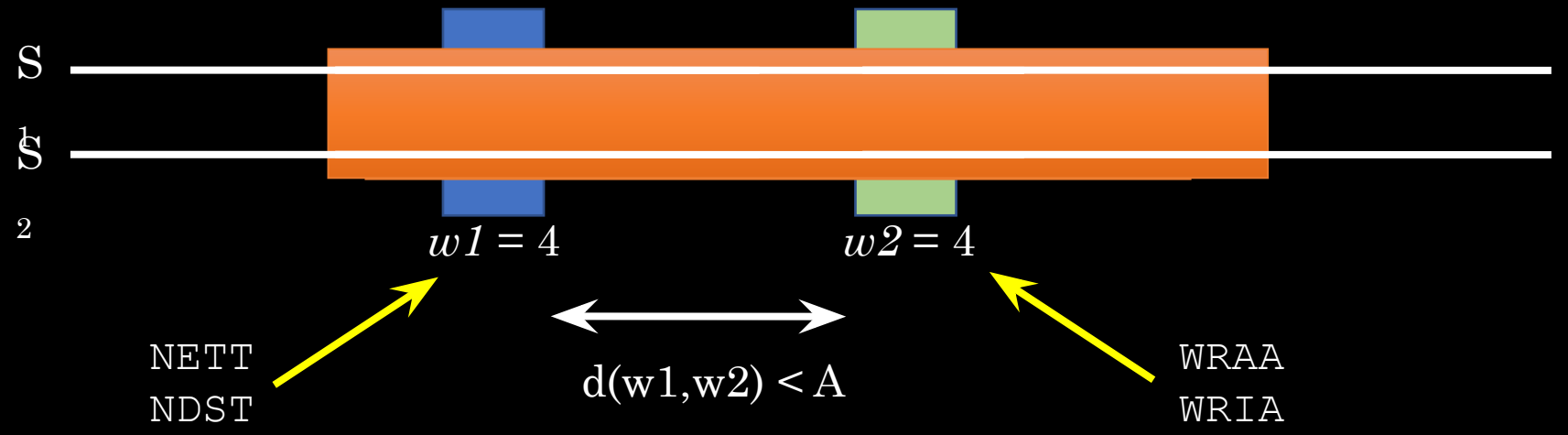
- Find maximal high-scoring pairs: for a query / database sequence pair, find the best region(s) where:
  - The local alignment score (no gaps allowed!) is above a threshold *S*, and
  - The score cannot be increased by extending or trimming the local alignment (therefore maximal)

Basic Principles of BLAST

- Instead of running full dynamic programming (à la S-W):

1. Identify matches that contain two word pairs (or *hits*) of length w, with a score of at least T, that are separated by no greater than A nucleotides
2. If word pairs are found, use these to seed the high-scoring pairs
3. If HSPs are found, perform dynamic programming anchored with HSPs to complete the alignment

Extending to high scoring pairs

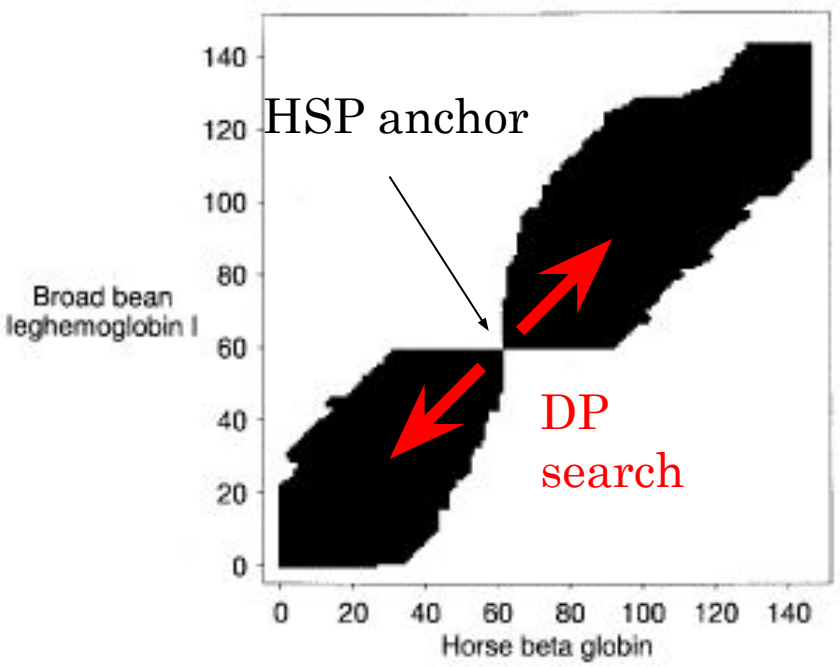


Try to extend matches, stop trying when a move drops the score below a given threshold

Gaps

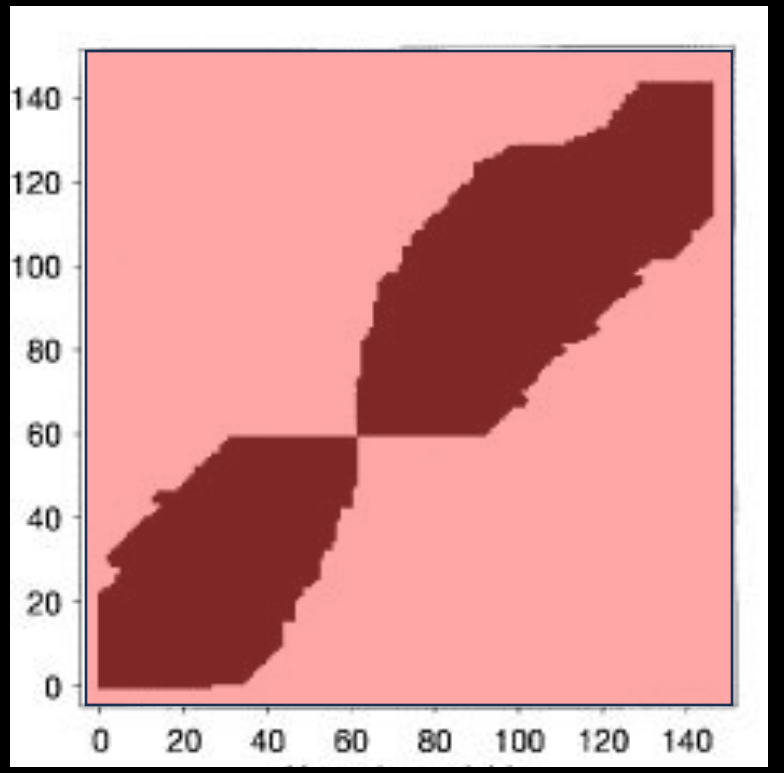
- Start from the middle of the high-scoring pair, and proceed with DP forward and backward until the path falls below a threshold

- DP is expensive, but we've saved ourselves a lot of time!
  - Most sequence pairs are *not* homologous and will drop out before we get to the DP step
  - Anchored DP will be a *lot* faster


HSP anchor
DP
search
Broad bean
leghemoglobin I
Horse beta globin
0
20
40
60
80
100
120
140

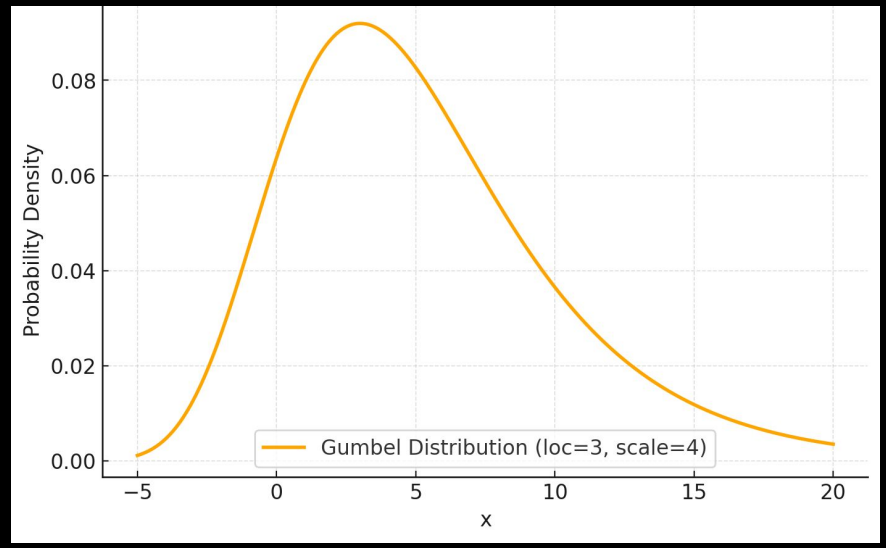
What have we achieved?

- Drastic reduction in the number of DP alignment processes AND reduced search space for gapped alignments
- Branch and bound, sort of



Local alignment significance

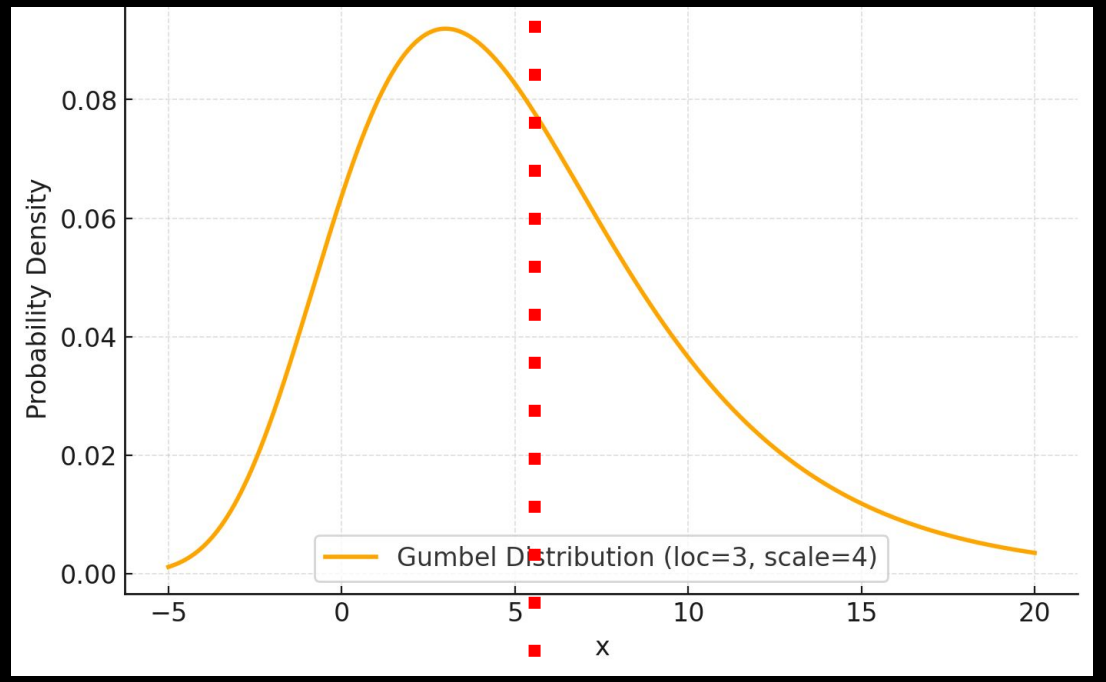
- How are alignment scores distributed?
- More to the point, what is the distribution of best alignment scores between a random pair of sequences?
- Follows the two-parameter Gumbel extreme value distribution – not Gaussian!



Karlin-Altschul statistics: no permutations, thanks

- The expected alignment score between a pair of random sequences is the mean of an extreme value distribution

- Given a scoring matrix (such as PAM250) and a set of amino acid frequencies, we can compute parameters λ and K that define this distribution



Karlin-Altschul statistics

The probability of getting a score greater than this:

$$P(S > \frac{\ln(nm)}{\lambda} + x) \approx 1 - \exp(-Ke^{-\lambda x})$$

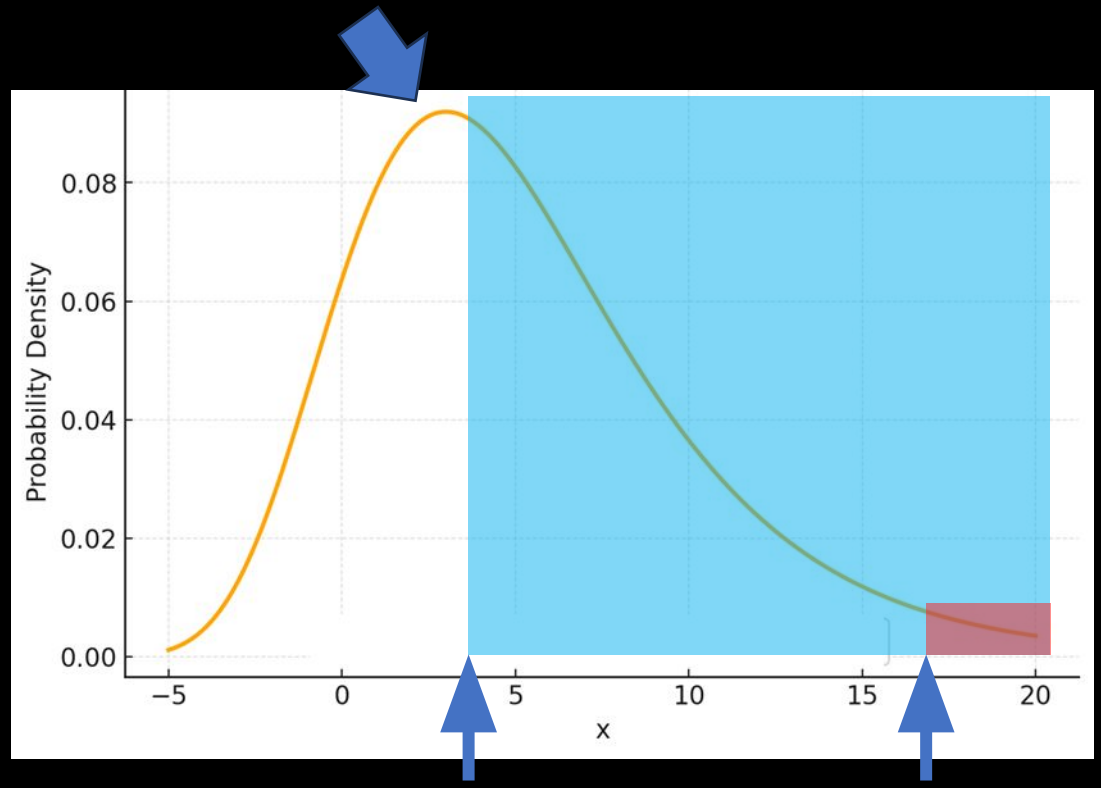
Is roughly equal to this

Where:

- n = length of sequence 1 (query)
- m = length of sequence 2 (database)
- K and λ are determined by the substitution matrix

Karlin-Altschul statistics

The substitution matrix determines the shape of the curve



S small, P(S) large S large, P(S) small

Karlin-Altschul statistics

- Different matrices (PAM, BLOSUM, etc.) define different EVDs – different K and λ

- We can normalize the search score S to equalize the effects of different matrices:

$$S' = \frac{\lambda S - \ln K}{\ln 2}$$

- So we can compare bitscores from different matrices directly

From P to E

Expectation value (e-value):

The expected number of hits to a database of random sequences of the **same total** length as the "real" sequence databases

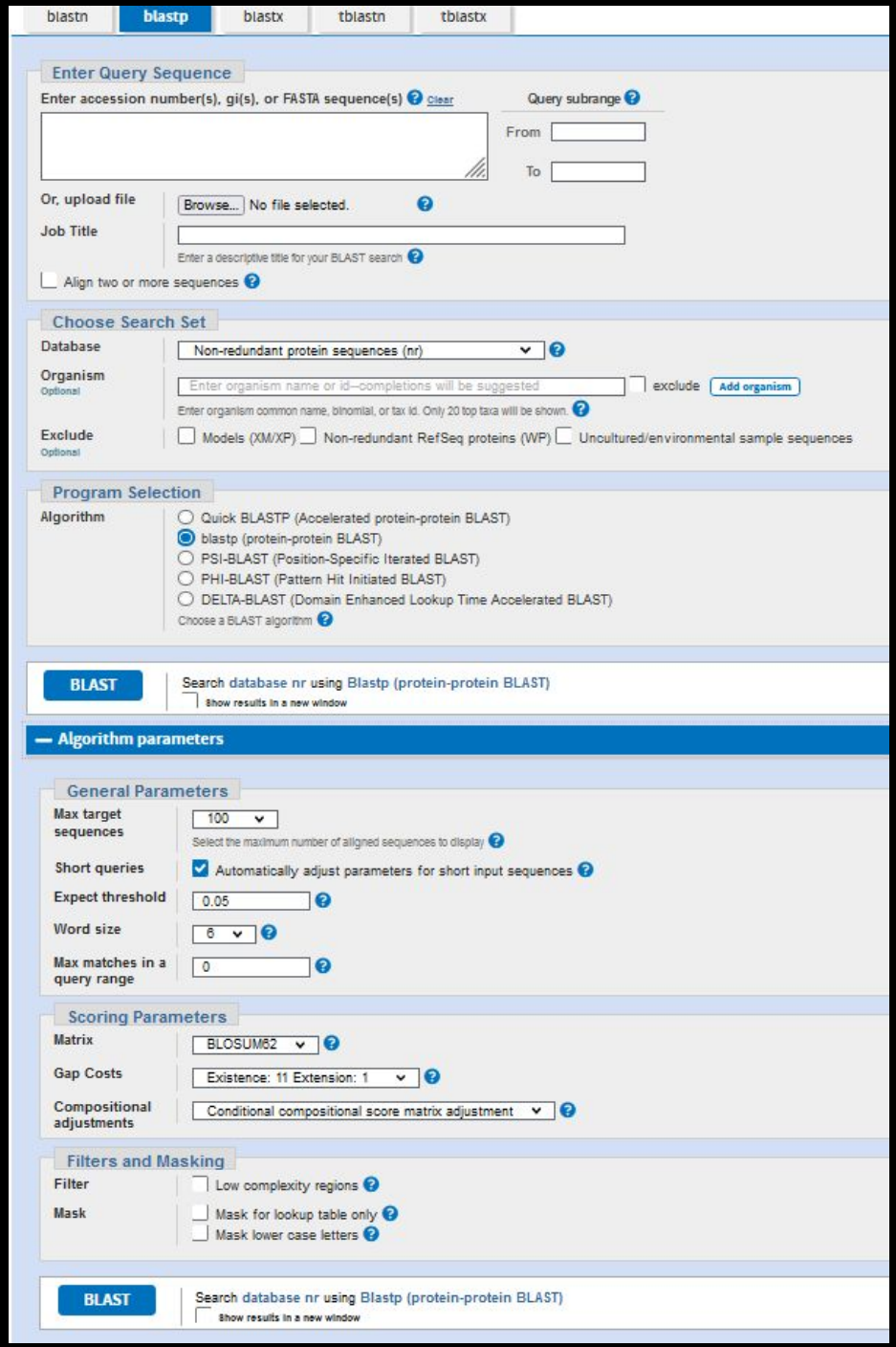
$$E = \frac{nm}{2^{S'}}$$

n = query sequence length

m = database length

Protein-protein BLAST (BLASTP):

https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE=Proteins



Query

Database

Algorithm

Match parameters

Scoring

The BLAST Family

Obvious:

- BLASTP – Protein query, protein DB
- BLASTN – Nucleotide query, nucleotide DB

Maybe less obvious:

- BLASTX – Translated nucleotide query, protein DB
- TBLASTN – Protein query, translated nucleotide DB
- TBLASTX – Translated nucleotide query, translated nucleotide DB (sloooow)

BLAST vs. FASTA

- In *very* rough terms, BLAST is about ten times faster than FASTA (but it depends on the data set and the specific tweaked version of the programs)
- FASTA is generally thought to be more sensitive than BLAST (although this again depends on the data set)
- No one reallllly uses FASTA anymore

PSI-BLAST (1997)

- Replace trusty old PAM or BLOSUM with a position-specific scoring matrix
- Remember: Frequencies and substitution probabilities can be different at different locations in a set of homologous proteins
- Iterative query – Position-specific scoring matrix (PSSM) procedure

PSSMs

Single sequence

1	2	3	4	5	6
M	C	D	N	L	K

Matrix constructed from multiple sequences (typically represented as log-odds ratios)

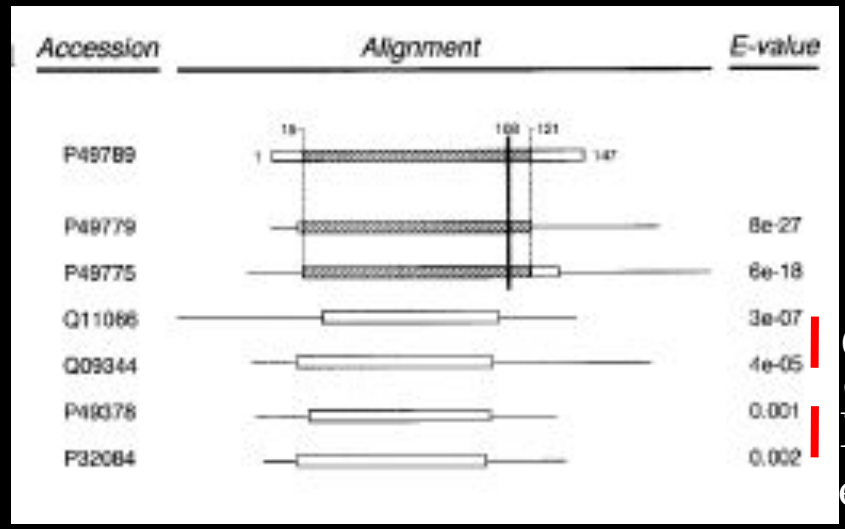
	1	2	3	4	5	6
A	0	0.01	0.01	0.02	0.05	0
C	0	0.8	0.03	0	0.09	0
D	0	0.03	0.6	0.01	0.01	0.2
E	0	0.01	0.1	0	0.15	0.02
…	…	…	…	…	…	…
Σ	1.0	1.0	1.0	1.0	1.0	1.0

PSI-BLAST: Step 1

Run BLAST!

PSI-BLAST: Step 2

- Collapse significant local alignments into a multiple alignment



Gaps

!

E-value not good enough!

PSI-BLAST: Step 3

- Build a column-specific matrix from the multiple alignment – this is similar to the PAM matrix

	Position 1	Position 2	Position 3
A	1.9	-4.0	-2.2
C	-5.0	-2.4	-3.1
D	-2.3	-0.5	0.1
…			

- Pseudocounts (based on substitution matrix) are added to avoid the embarrassing -∞ situation

PSI-BLAST: Step 4

- Iterate the search: BLAST using the profile rather than a single sequence, as the query

- When do we stop?
  - When no new hits are found
  - When we get tired of hitting the 'BLAST!' button

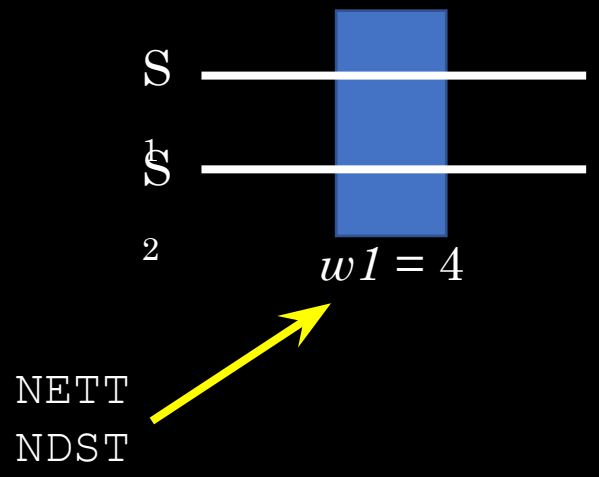
Discontiguous MEGABLAST

- BLAST isn't fast enough!
- Can we (etc…)

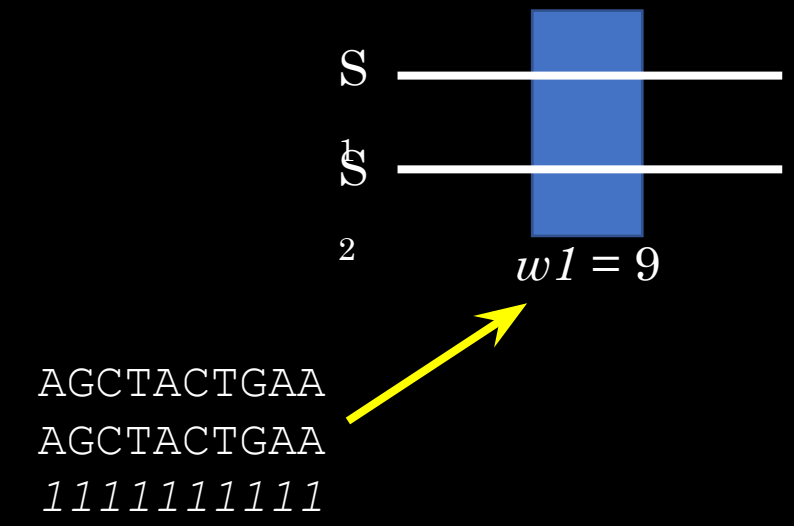

MEGABLAST!!!

Reminder: seeds

Amino acids:
Seeds with matrix score > threshold



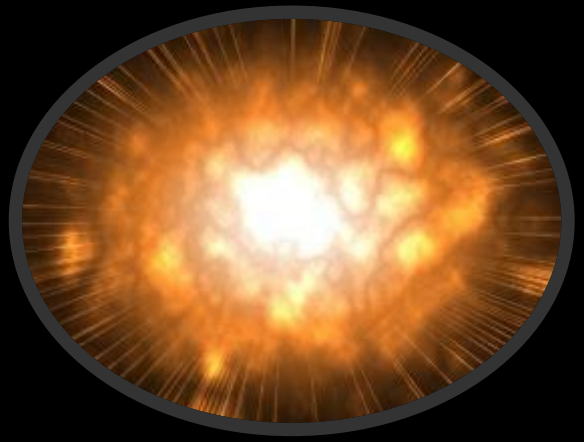
Nucleotides:
Seeds with literal-matching “model”



Nucleotide seed length

BLASTN!

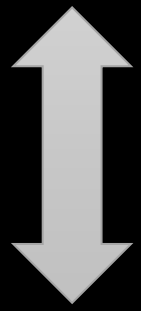
11111111111



MEGABLAST!!!

11111111111111111111111111

BLASTN is good for distant-ish sequences but kinda slow



Happy medium???

MEGABLAST!!! is good for very, very, very similar sequences and fast

Continuous Seeds

- BLASTN (for nucleotides) has a default word length of 11 to find the initial hits. This word must be contiguous

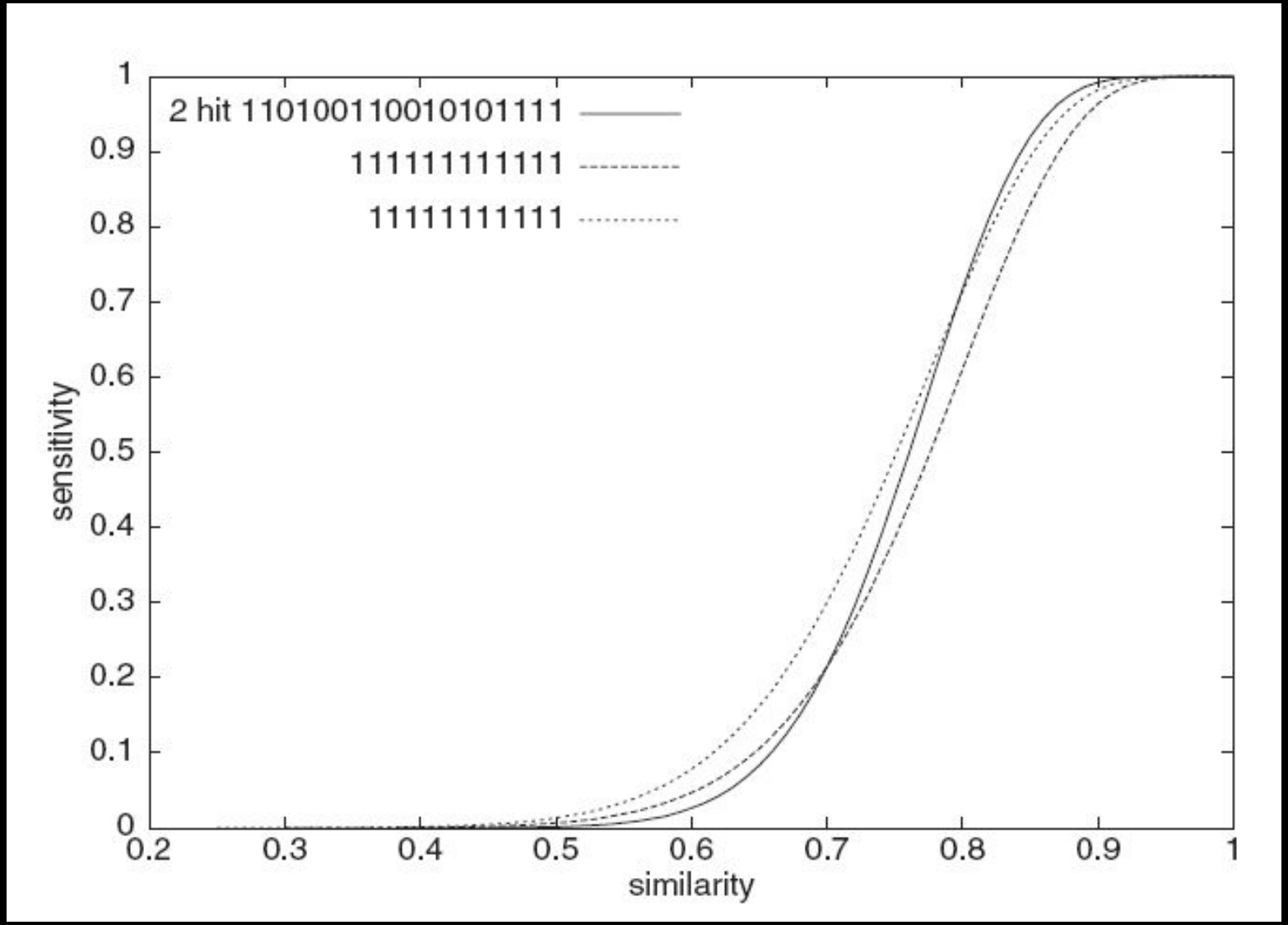
AAACGATCCGAAAGTTT
GCACGATCCGAAAATCC
11111111111

Discontiguous Words

Specific model does not need to be contiguous

AAACGAACAGAGAGTTTC
AAATGATCCGAAAGCTTC
11101001010011011 1

Comparing models
PatternHunter – commercial variant



PatternHunter: Ma et al., 2002

PatternHunter is quite a bit faster than the contiguous-word BLAST family

Seq1	Size	Seq2	Size	PH	PH2	MB28	Blastn
M. pneumoniae	828 K	*M. genitalium*	589 K	10 s/65 M	4 s/48 M	1 s/88 M	47 s/45 M
E. coli	4.7 M	*H. influenza*	1.8 M	34 s/78 M	14 s/68 M	5 s/561 M	716 s/158 M
A. thaliana chr 2	19.6 M	*A. thaliana* chr 4	17.5 M	5020 s/279 M	498 s/231 M	21 720 s/1087 M	∞
H. sapiens chr 22	35 M	*H. sapiens* chr 21	26.2 M	14 512 s/419 M	5250 s/417 M	∞	∞

But it costs money!

Li et al. (2002) *Bioinformatics*
Reviewed in Stojanov (2018) *Annals of West University of Timişoara*

Other important issues

- Low complexity sequence (e.g., AGAGAGAG) can lead to inflated statistics and should be removed prior to the search

- We are still dependent on the choice of substitution matrix!

DIAMOND: double index alignment of next-generation sequencing data

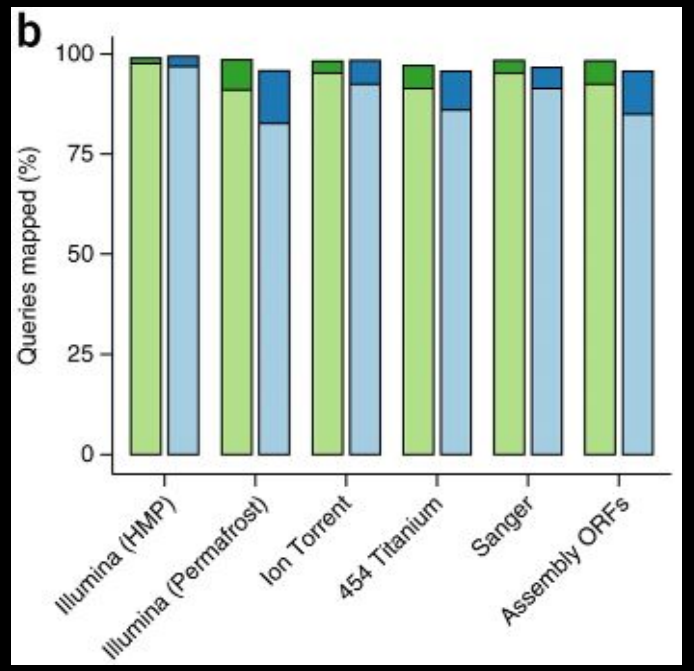
- Double indexing: precompute all "seeds" in the database *and* query sequences, compare in lexicographical order (memory cache efficient)
- "Shaped" seeds (similar to discontiguous MEGABLAST, but for proteins)
- Reduced amino acid alphabet!

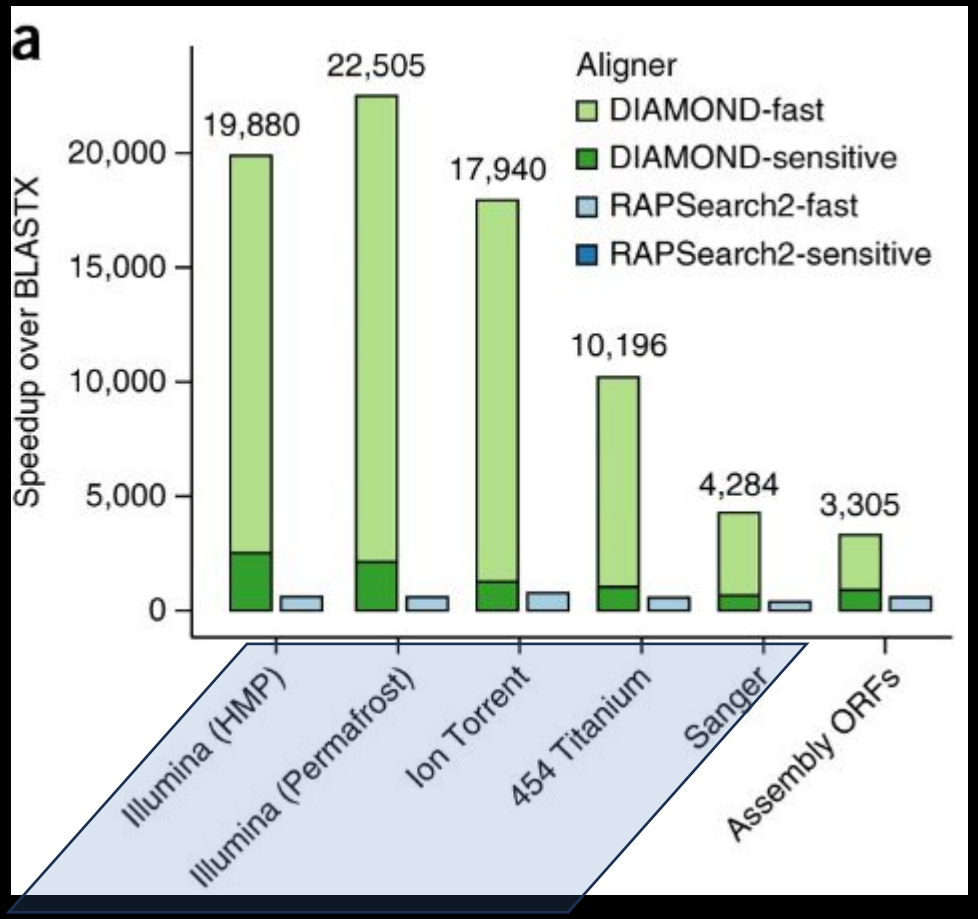
  [KREDQN] [C] [G] [H] [ILV] [M] [F] [Y] [W] [P] [STA]

- Other stuff

Buchfink et al. (2015) *Nat Meth*

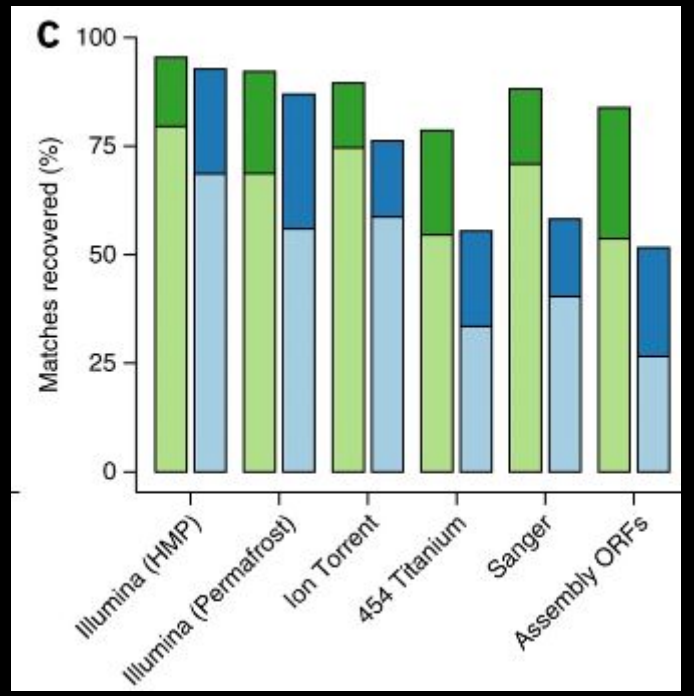
At least one match





DNA sequencing technique

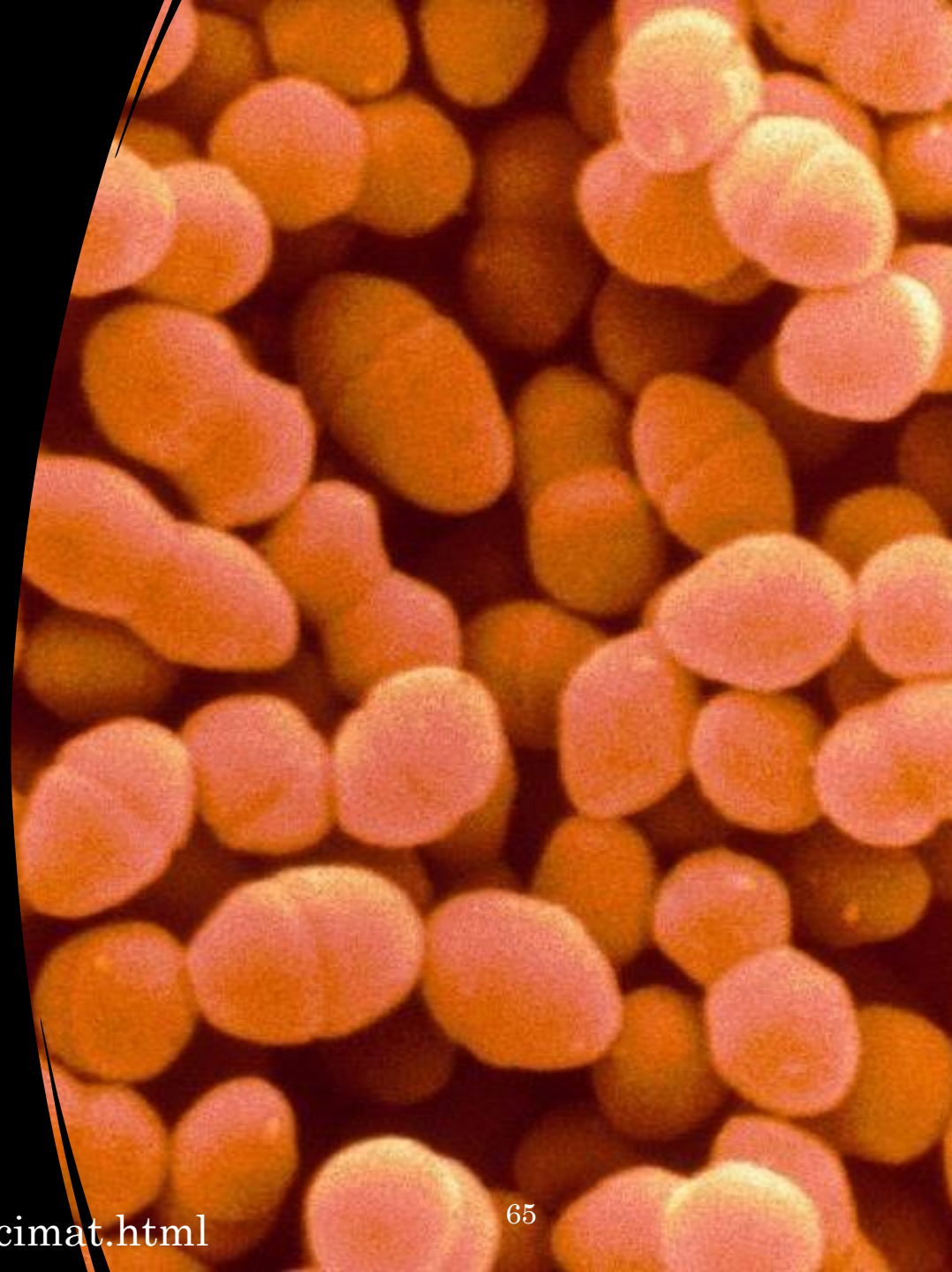
All expected matches (relative to BLASTX)



Non-pretty example

- 1,273 genomes of *Enterococcus faecium* vs. 21,000 reference genomes from RefSeq

- The big question: are there genes in *Enterococcus* with very, very, very similar homologs in distantly related groups of bacteria?



https://fineartamerica.com/featured/3-enterococcus-faecium-sem-scimat.html

DIAMOND-BLASTX

- Query: protein-coding genes from an *E. faecium* plasmid
- Database: predicted proteins from 21,000 genomes
- VERY stringent thresholds: minimum 99% identical, at least 90% of total length
- Run locally

Query	Subject	Taxonomic range	Function	% Identity	e-value	Query start	Query end	Length
18_length=47093_depth=1.75x	WP_000331160.1	[Bacteria]	MULTISPECIES: ATP-binding protein	100	0	34273	36717	2444
18_length=47093_depth=1.75x	WP_074371015.1	[Staphylococcus aureus]	ATP-binding protein	99.9	0	34273	36717	2444
18_length=47093_depth=1.75x	WP_116449323.1	[Streptococcus agalactiae]	ATP-binding protein	99.9	0	34273	36717	2444
18_length=47093_depth=1.75x	WP_001574271.1	[Bacilli]	MULTISPECIES: YtxH domain-containing protein	99.9	0.00E+00	36723	38897	2174
18_length=47093_depth=1.75x	WP_060649663.1	[Staphylococcus aureus]	YtxH domain-containing protein	99.7	0.00E+00	36723	38897	2174
18_length=47093_depth=1.75x	WP_041160410.1	[Clostridioides difficile]	YtxH domain-containing protein	99.2	0.00E+00	36723	38897	2174
18_length=47093_depth=1.75x	WP_001574275.1	[Bacteria]	MULTISPECIES: tetracycline resistance ribosomal protection protein Tet(M	100	0.00E+00	41204	43120	1916
18_length=47093_depth=1.75x	WP_012775613.1	[Streptococcus suis]	tetracycline resistance ribosomal protection protein Tet(M)	99.5	0.00E+00	41204	43120	1916
18_length=47093_depth=1.75x	WP_002333004.1	[Bacilli]	MULTISPECIES: hypothetical protein	99.4	0.00E+00	4822	3212	1610
18_length=47093_depth=1.75x	WP_000136908.1	[Bacilli]	MULTISPECIES: recombinase family protein	99.8	0.00E+00	26267	24708	1559
18_length=47093_depth=1.75x	WP_206918171.1	[Lactococcus sp. LG606]	recombinase family protein	99.8	0.00E+00	26249	24708	1541
18_length=47093_depth=1.75x	WP_002294513.1	[Bacteria]	MULTISPECIES: ABC-F type ribosomal protection protein Lsa(E)	100	0.00E+00	18264	16783	1481
18_length=47093_depth=1.75x	WP_074371031.1	[Staphylococcus aureus]	ABC-F type ribosomal protection protein Lsa(E)	99.8	0.00E+00	18264	16783	1481
18_length=47093_depth=1.75x	WP_222317233.1	[Vagococcus lutrae]	ABC-F type ribosomal protection protein Lsa(E)	99.8	0.00E+00	18264	16783	1481
18_length=47093_depth=1.75x	WP_000813488.1	[Bacteria]	MULTISPECIES: DUF87 domain-containing protein	100	1.35E-298	30162	31544	1382

Not super-informative

RefSeq ID!

Resistance to tetracycline (bad)

Resistance to multiple drug classes (very bad)

??? (DUF = “Domain of Unknown Function”)

Summary

- Full dynamic programming is too slow for large database searches

- Key guiding principles:
  - Avoid comparisons where possible
  - Reduce search spaces
  - Calculate appropriate statistics