

# 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)

An official website of the United States government [Here's how you know](#) 

**NIH** National Library of Medicine  
National Center for Biotechnology Information

[Log in](#)

[All Databases](#)  [Search](#)

[All Databases](#)

- NCBI Home
- Resource List (A-Z)
- All Resources
- Chemicals & Bioassays
- Data & Software
- DNA & RNA
- Domains & Structures
- Genes & Expression
- Genetics & Medicine
- Genomes & Maps
- Homology
- Literature
- Proteins
- Sequence Analysis
- Taxonomy
- Training & Tutorials
- Variation

**NCBI**  
for Biotechnology Information advances science and health by providing access to genomic information.

[Mission](#) | [Organization](#) | [NCBI News & Blog](#)

**mit** **Download** **Learn**

Transfer NCBI data to your computer

Find help documents, attend a class or watch a tutorial

**Develop** **Analyze** **Research**

Find code and applications

Identify an NCBI tool for your data analysis task

Explore NCBI research and collaborative projects

**Popular Resources**

- PubMed
- Bookshelf
- PubMed Central
- BLAST
- Nucleotide
- Genome
- SNP
- Gene
- Protein
- PubChem

**NCBI News & Blog**

New! Introducing the Multiple Comparative Genome Viewer (MCGV) Beta Release 16 Jan 2025  
NIH's NCBI is excited to introduce the [An updated bacterial and archaeal reference genome collection is available!](#) 14 Jan 2025  
Download the [updated bacterial and archaeal reference genome collection](#)  
NCBI Resources Highlighted in 2025 [Nucleic Acids Research Database Issue](#)

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

REST-like URL



Download

Select columns

314,368 Genomes

Rows per page

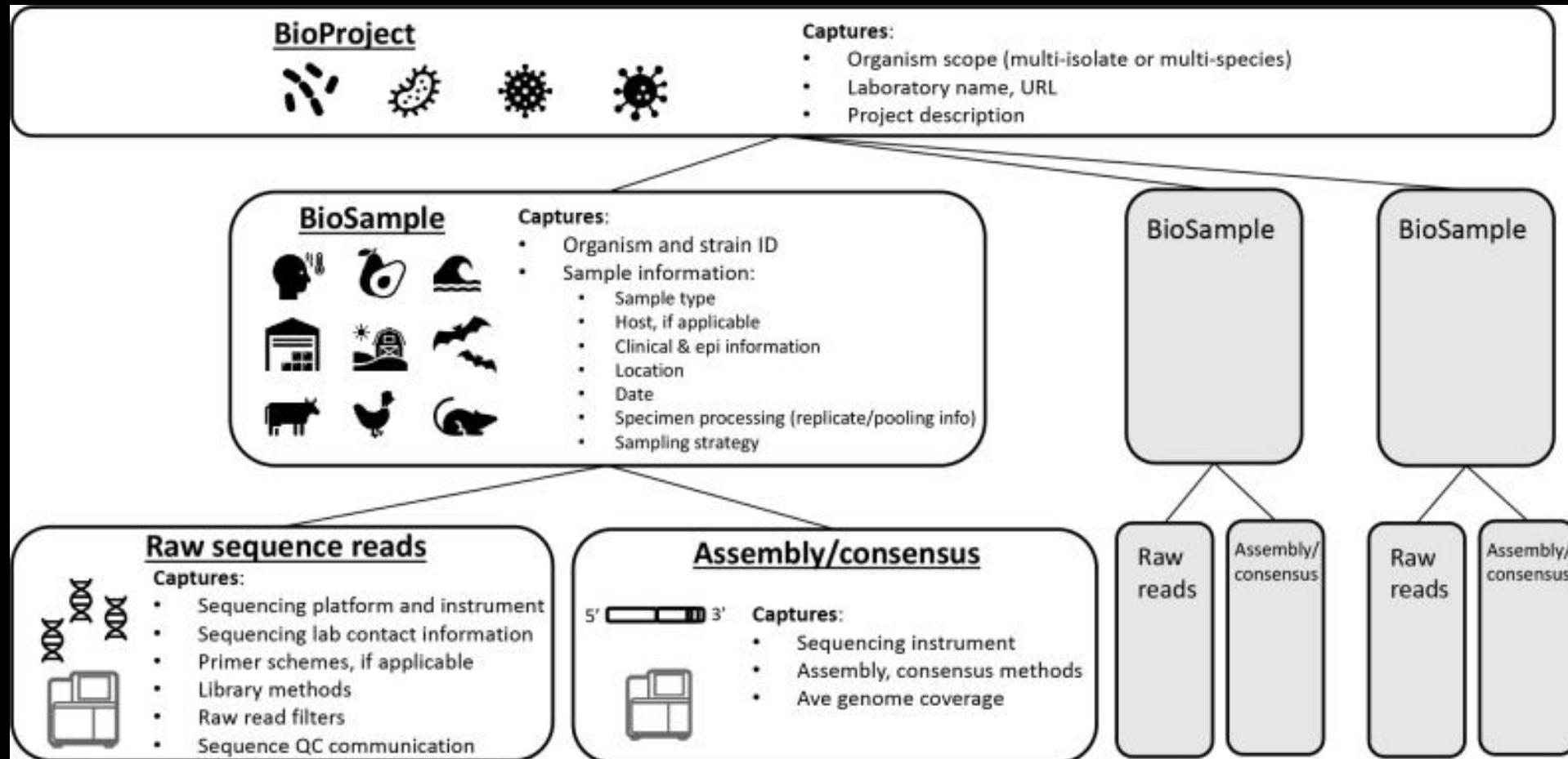
20

1-20 of 314,368



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

# NCBI (Pathogen) Data Object Model



Find a gene, protein or chemical

All



Search

Example searches: [blast](#) [keratin](#) [bfl1](#) | [About EBI Search](#)

Nucleotides,

Genomes,

**Protein function,**

Protein-protein interactions

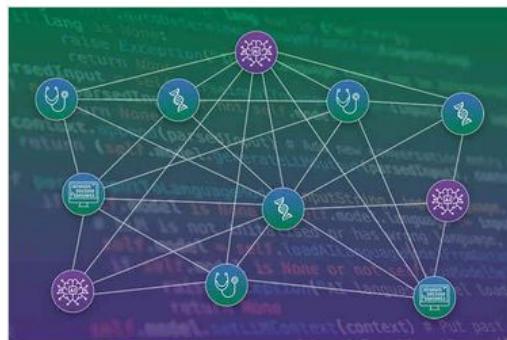
[Find data resources](#) →

[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



Results of EMBL-EBI's 2024 user survey

Fuelling discovery together: 2024 user survey learnings

# Example Record – “BlaZ”

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

**P06548 · BLA3\_BACCE**

**Function**

Names & Taxonomy

Subcellular Location

Phenotypes & Variants

PTM/Processing

Expression

Interaction

Structure

Family & Domains

Sequence

Similar Proteins

Protein<sup>i</sup> Beta-lactamase 3

Gene<sup>i</sup> blaZ

Status<sup>i</sup> UniProtKB reviewed (Swiss-Prot)

Organism<sup>i</sup> *Bacillus cereus*

Amino acids 316 (go to sequence)

Protein existence<sup>i</sup> Evidence at protein level

Annotation score<sup>i</sup> 3/5

Entry Variant viewer Feature viewer Genomic coordinates Publications External links History

BLAST Download Add Add a publication Entry feedback

**Function<sup>i</sup>**

**Catalytic activity<sup>i</sup>**

Rhea 20401 a beta-lactam + H<sub>2</sub>O = a substituted beta-amino acid PROSITE-ProRule Annotation

EC:3.5.2.6 (UniProtKB | ENZYME | Rhea )

Hide Rhea reaction ^

a  $\beta$ -lactam CHEBI:35627

R[C@H]1C(=O)N1R

H<sub>2</sub>O CHEBI:15377

H2O

a substituted  $\beta$ -amino acid CHEBI:140347

RC(=O)C1C(R)N(R)N1H

# International Nucleotide Sequence Database Consortium



- + China National Center for Bioinformatics (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)

CARD

Use or Download Copyright & Disclaimer

Help Us Curate #AMRCuration #WorkTogether

[Browse](#)

[Analyze](#)

[Download](#)

[About](#)

Search

## The Comprehensive Antibiotic Resistance Database

A bioinformatic database of resistance genes, their products and associated phenotypes.

8526 Ontology Terms, 6442 Reference Sequences, 4542 SNPs, 3328 Publications, 6490 AMR Detection Models

Resistome predictions: 414 pathogens, 24291 chromosomes, 2662 genomic islands, 48212 plasmids, 172216 WGS assemblies, 279120 alleles

YouTube: [Canadian Bioinformatics Workshops 2024: Antimicrobial Resistant Gene \(AMR\) Analysis](#)

### [Browse](#)

The CARD is a rigorously curated collection of characterized, peer-reviewed resistance determinants and associated antibiotics, organized by the Antibiotic Resistance Ontology (ARO) and AMR gene detection

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

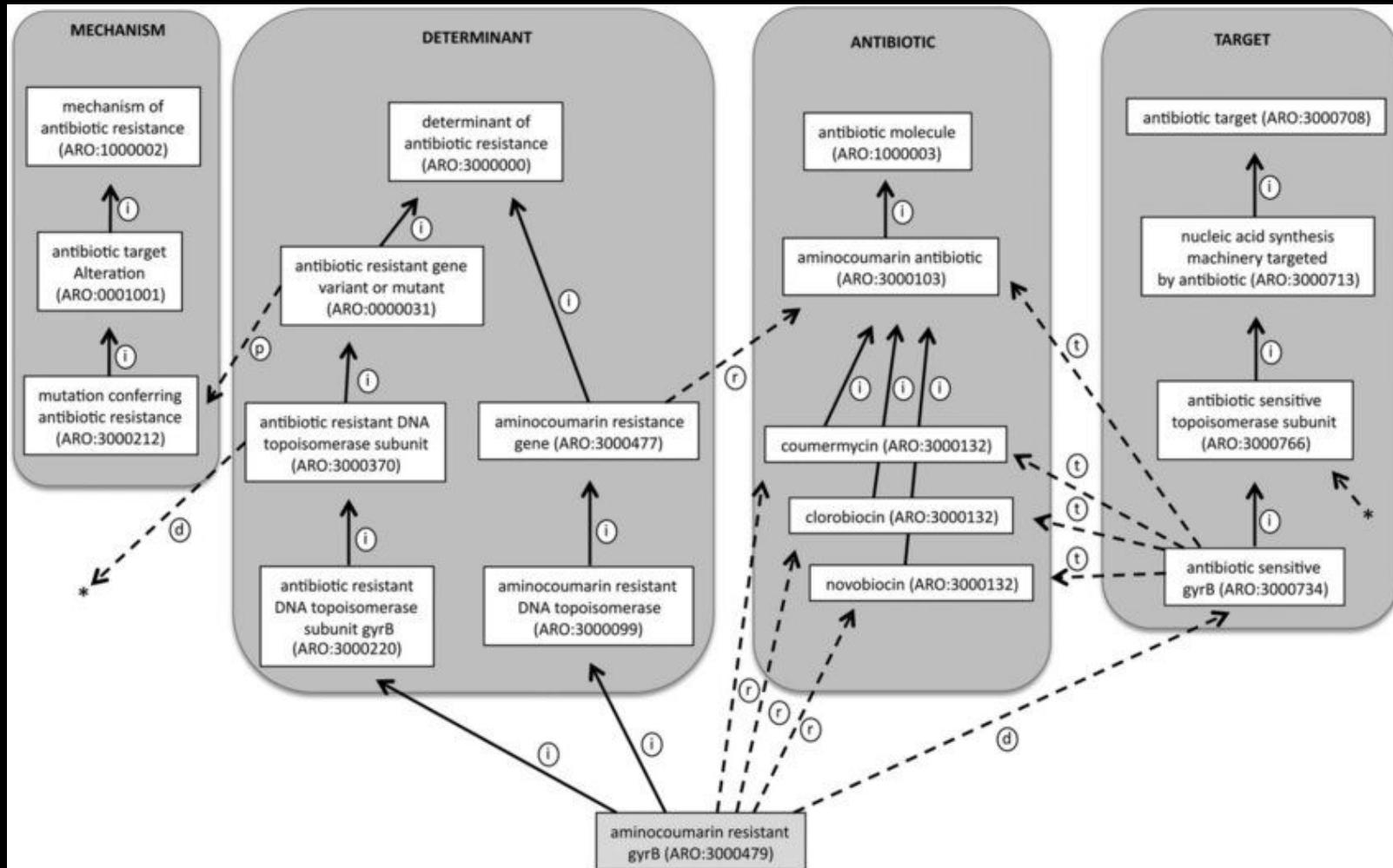
### [Analyze](#)

The CARD includes tools for analysis of molecular sequences, including BLAST and the Resistance Gene Identifier (RGI) software for prediction of resistome based on homology and SNP models.

### [Download](#)

CARD data and ontologies can be downloaded in a number of formats, including lists of mutations and molecules with corresponding metadata and citations. RGI software is available as a command-line tool. CARD Bait Capture Platform sequences and protocols available for download. Extensive notes on updates provided.

# 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 <i>Staphylococcus aureus</i> .
Drug Class	<a href="#">penicillin beta-lactam</a>
Resistance Mechanism	<a href="#">antibiotic inactivation</a>
Classification	<b>10 ontology terms</b>   <a href="#">Hide</a> + <a href="#">process or component of antibiotic biology or chemistry</a> + <a href="#">mechanism of antibiotic resistance</a> + <a href="#">determinant of antibiotic resistance</a> + <a href="#">antibiotic inactivation</a> [Resistance Mechanism] + <a href="#">antibiotic inactivation enzyme</a> + <a href="#">hydrolysis of antibiotic conferring resistance</a> + <a href="#">antibiotic molecule</a> + <a href="#">hydrolysis of beta-lactam antibiotic by serine beta-lactamase</a> + <a href="#">beta-lactam antibiotic</a> + <a href="#">beta-lactamase</a>
Parent Term(s)	<b>2 ontology terms</b>   <a href="#">Hide</a> + <a href="#">confers_resistance_to_drug_class</a> <a href="#">penicillin beta-lactam</a> [Drug Class] + <a href="#">class A beta-lactamase</a>
Sub-Term(s)	<b>3 ontology terms</b>   <a href="#">Hide</a> + <a href="#">PC1 beta-lactamase (blaZ)</a> + <a href="#">mecC-type BlaZ</a> + <a href="#">PC1</a>
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 <i>Staphylococcus aureus</i> beta-lactamase gene. ( <a href="#">PMID 6793593</a> ) Pence MA, et al. 2015. PLoS ONE 10(8):e0136605 Beta-Lactamase Repressor BlaI Modulates <i>Staphylococcus aureus</i> Cathelicidin Antimicrobial Peptide Resistance and Virulence. ( <a href="#">PMID 26305782</a> )

# Antimicrobial Resistance Gene Predictions (*Klebsiella pneumoniae* plasmid)

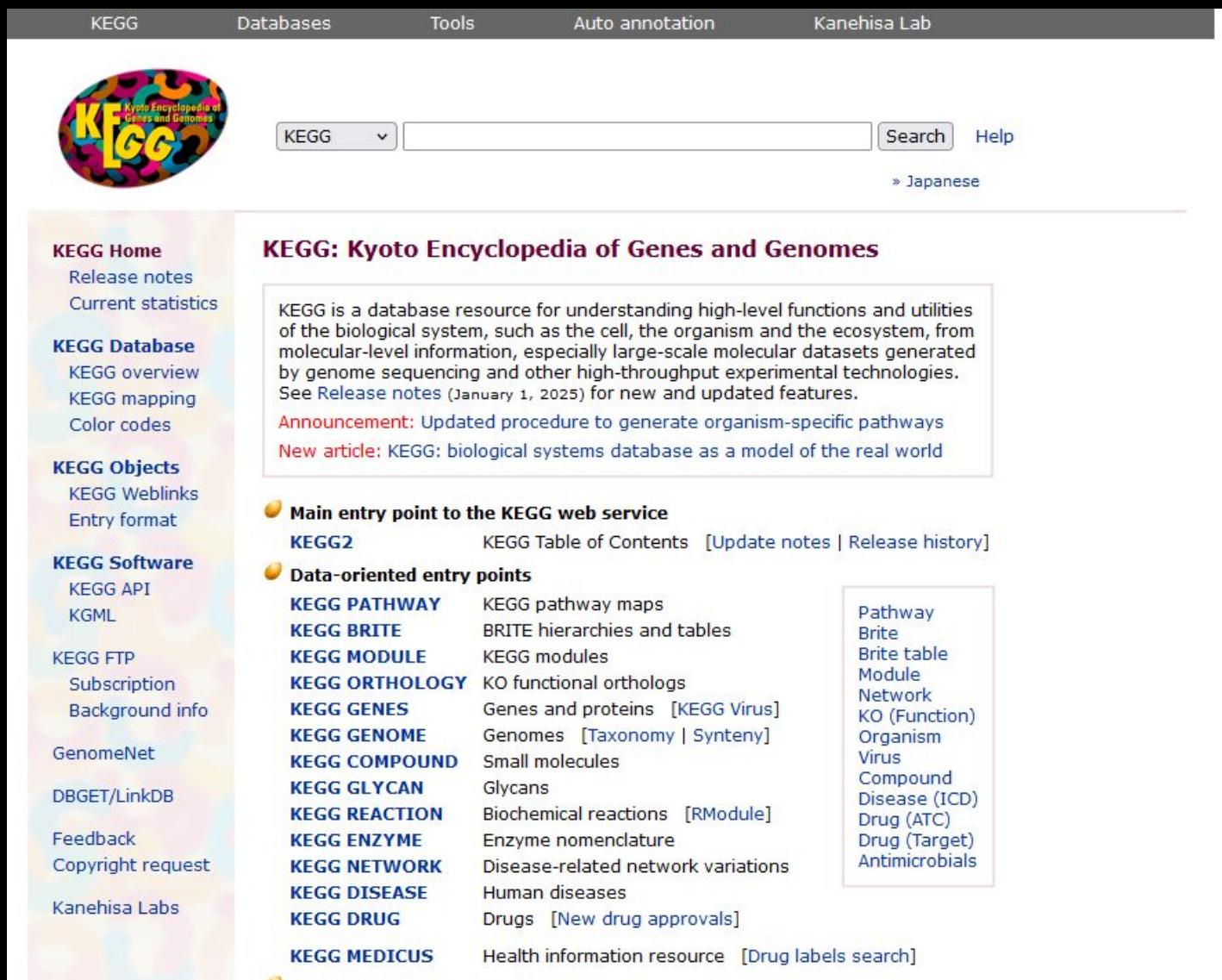
Summary (summary counts and figures only include Loose hits of e-10 or better)								
Filename	Date (UTC)	RGI Criteria		# Perfect Hits	# Strict Hits	# Loose Hits	Download	
JN420336.1	February 01, 2024 15:02:45	Perfect, Strict, complete genes only		5	1	0	<a href="#">Download</a>	
Results (all Loose hits shown)								
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')-lb-cr6		protein homolog model	AAC(6'), AAC(6')-lb-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



The screenshot shows the KEGG website homepage. The header includes links for KEGG, Databases, Tools, Auto annotation, and Kanehisa Lab. The main content area features the KEGG logo and a search bar. A sidebar on the left lists various KEGG resources: KEGG Home, KEGG Database, KEGG Objects, KEGG Software, KEGG FTP, GenomeNet, DBGET/LinkDB, Feedback, Copyright request, and Kanehisa Labs. The main content area contains a brief introduction to KEGG, an announcement about updated procedures, and a new article. Below this, there are two sections: 'Main entry point to the KEGG web service' and 'Data-oriented entry points'. The 'Data-oriented entry points' section lists various KEGG databases and resources, each with a link. A box on the right lists these resources with their corresponding KEGG identifiers: Pathway (KEGG PATHWAY), Brite (KEGG BRITE), Module (KEGG MODULE), Orthology (KEGG ORTHOLOGY), Genes (KEGG GENES), Genome (KEGG GENOME), Compound (KEGG COMPOUND), Glycan (KEGG GLYCAN), Reaction (KEGG REACTION), Enzyme (KEGG ENZYME), Network (KEGG NETWORK), Disease (KEGG DISEASE), Drug (KEGG DRUG), and Medicus (KEGG MEDICUS).

KEGG: Kyoto Encyclopedia of Genes and Genomes

KEGG is a database resource for understanding high-level functions and utilities of the biological system, such as the cell, the organism and the ecosystem, from molecular-level information, especially large-scale molecular datasets generated by genome sequencing and other high-throughput experimental technologies. See [Release notes \(January 1, 2025\)](#) for new and updated features.

**Announcement:** Updated procedure to generate organism-specific pathways

**New article:** KEGG: biological systems database as a model of the real world

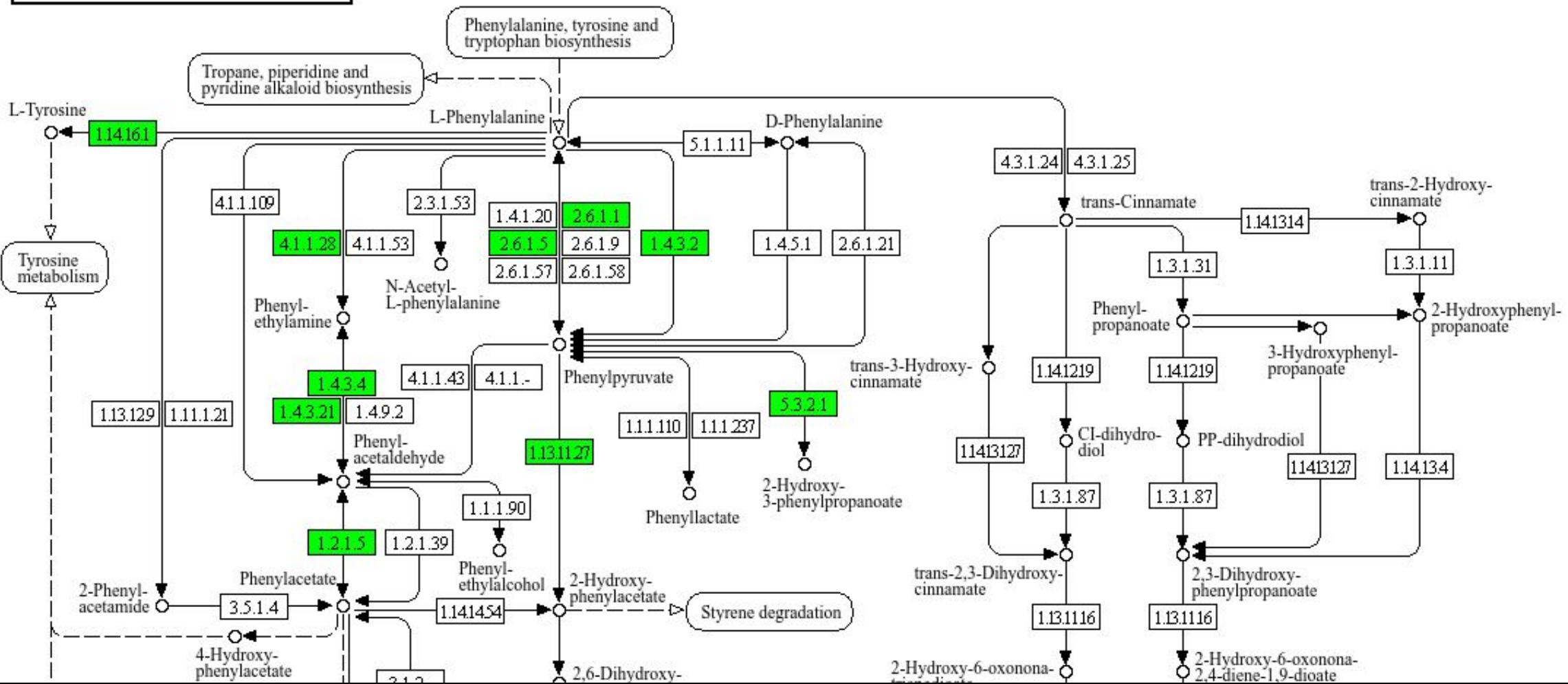
**Main entry point to the KEGG web service**

**KEGG2** KEGG Table of Contents [Update notes | Release history]

**Data-oriented entry points**

<b>KEGG PATHWAY</b>	KEGG pathway maps	Pathway
<b>KEGG BRITE</b>	BRITE hierarchies and tables	Brite
<b>KEGG MODULE</b>	KEGG modules	BRITE table
<b>KEGG ORTHOLOGY</b>	KO functional orthologs	Module
<b>KEGG GENES</b>	Genes and proteins [KEGG Virus]	Network
<b>KEGG GENOME</b>	Genomes [Taxonomy   Synteny]	KO (Function)
<b>KEGG COMPOUND</b>	Small molecules	Organism
<b>KEGG GLYCAN</b>	Glycans	Virus
<b>KEGG REACTION</b>	Biochemical reactions [RModule]	Compound
<b>KEGG ENZYME</b>	Enzyme nomenclature	Disease (ICD)
<b>KEGG NETWORK</b>	Disease-related network variations	Drug (ATC)
<b>KEGG DISEASE</b>	Human diseases	Drug (Target)
<b>KEGG DRUG</b>	Drugs [New drug approvals]	Antimicrobials
<b>KEGG MEDICUS</b>	Health information resource [Drug labels search]	

## PHENYLALANINE METABOLISM



Phenylalanine metabolism (GREEN = active in primates)

# A word about “metadata”

**nature  
biotechnology**

**PERSPECTIVE**

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

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

(2011)

# Genomes Online Database (GOLD)

**JGI** **GOLD**  
GENOMES ONLINE DATABASE

JGI HOME LOG IN

Home Search Distribution Graphs Biogeographical Metadata Statistics References Team Help News

**Welcome to the Genomes OnLine Database**

**GOLD**: Genomes Online Database, is a World Wide Web resource for comprehensive access to information regarding genome and metagenome sequencing projects, and their associated metadata, around the world.

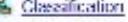
**GOLD Release v.5**

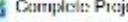
**1. Register**  
  
Register your project information and Metadata in the Genomes Online Database  
[Register](#)

**2. Annotate**  
  
Annotate your microbial genome or metagenome with IMG/ER or IMG/MER  
[Annotate](#)

**3. Publish**  
  
Standards in Genomic Sciences  
Publish your genome or metagenome in open access standards-supportive journal.  
[Publish](#)

**Studies**  
Metagenomic: 637  
Non Metagenomic 22,621

**Biosamples**  
  
Classification  
Ecosystems:  
Host-associated 13,350  
Engineered 2,180  
Environmental 9,470

**Sequencing Projects**  
  
Complete Projects: 0,015  
Permanent Drafts: 33,200  
Incomplete Projects 35,607  
  
Targeted Projects 1,565

**Analysis Projects**  
Genome Analysis: 40,429  
Metagenomic Analysis: 5,627  
Combined Assembly 101  
Genome from Metagenome 1,499  
Metatranscriptome Analysis 1,280  
Single Cell (Screened) 1,681  
Single Cell (Unscreened) 781  
Transcriptome Analyses 0

**Organisms**  
Organisms 72,826  
Archaea 1,198  
Bacteria 55,157

**Special Projects**  
Type Strain Projects 5,328  
GEBA Projects 2,517  
HMP Projects 2,921

**JGI Projects**  
JGI Studies 1,111  
JGI Biosamples 19,723  
JGI Sequencing Projects 30,821

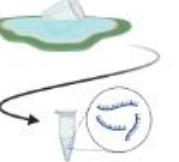
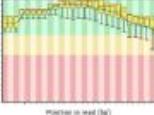
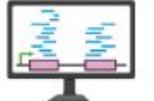
**Projects with Genbank Data**  
Seq. Projects 42,394  
Archaeal Projects 564  
Bacterial Projects 36,732

**Download Excel Data file**  
File last generated: 22 Feb, 2016

Genome projects  
Standards-compliant metadata

PROJECT INFORMATION			
GOLD Project ID	Gp0786898		
Project Name	Esch	ORGANISM NAME	
Other Names		GOLD Organism ID	Go0668578
Legacy GOLD ID		Organism Name	Escherichia coli ATCC BAA-196
NCBI BioProject Name	Antib	Other names (alias, synonyms, short names, common names)	
NCBI BioProject Accession	PRJN	Organism Domain	BACTERIAL
NCBI Locus Tag	BAA	Phylogeny	PSEUDOMONADOTA
NCBI BioSample Accession	SAM	Genus	Escherichia
Source Sample ID		Genus Synonyms	
PI	Unkn	Species	Escherichia coli
Added By	JGI a	Subspecies	
Last Modified By	Supr	Species Synonyms	
Project Comments		Strain	ATCC BAA-196
Project Status	Com	Strain Synonyms	
Project Relevance	Medi	Serovar/Cultivar	
Sequencing Center	Scienc	Culture Collection ID	
Collaborating Institute		Type Strain	
Funding Agency		Exemplar DOI	
Project Description	Esch	Exemplar Name	
Is JGI Project	No	Taxon DOI	
Project Information Visibility	Publi	Biosafety Level	
PROJECT TYPE		GENERAL PROPERTIES	
Specimen	Orga	Serovar/Cultivar	
Sequencing Strategy	Whol	Culture Collection ID	
Nucleic Acid	DNA	Type Strain	
EXTERNAL PROJECT LINKS			
JGI Data Utilization Status		Exemplar DOI	
JGI Award DOI		Exemplar Name	
ORGANISM TAXONOMY		GENERAL PROPERTIES	
NCBI Taxonomy ID	562	Oxygen Requirement	
NCBI Superkingdom	Bac	Cell Shape	
NCBI Kingdom		Motility	
NCBI Phylum	Pset	Sporulation	
NCBI Class	Gam	Temperature Range	
NCBI Order	Ente	Salinity	
NCBI Family	Ente	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”

<b>Database identifiers</b>  <p>IDs to describe how the sample was collected sample ID, subsample ID, pooled sample ID, MAG ID, project ID, site ID, event ID</p> <p>Accessions for where the sample is shared INSDC (ENA, NCBI, DDBJ), GISAID, GSA, Enterobase</p>	<b>Sequence information</b>  <p>Sequencing approach Purpose of sequencing Chain of custody Contact information, date of sequencing</p> <p>Lab procedures Sequencing platform and instrument, library preparation and sequencing protocols/scheme, genomic target enrichment</p>																																																
<b>Sample collection &amp; processing</b>  <p>Description of the site &amp; relevant external factors geographic location, watershed shapefile, presampling activity, descriptions of wastewater system and site</p> <p>Chain of custody contact information, dates and times for sample collection, storage, receipt, and processing</p> <p>Sampling approach organism targeted, purpose and scale of sampling</p> <p>Lab procedures sample volume; methods for sample collection and storage, specimen processing and extraction; controls</p>	<b>Bioinformatics &amp; QC metrics</b>  <p>QC overview QC method, issues, determination</p> <p>Procedures for processing genomic data Overall bioinformatics protocol; methods for raw sequence processing, dehosting, assembly, deduplication, read mapping; reference genome accession and/or taxonomic reference database</p> <p>QC details Breadth and depth of coverage, genome completeness and length, number of base pairs / reads / contigs, read length, percent or number of Ns, percent contamination</p>																																																
<b>Environmental conditions &amp; measurements</b>  <p>Human-related metrics Catchment population; populated area type</p> <p>Weather-related metrics Sampling &amp; pre-sampling weather, precipitation, temperature</p> <p>Physico-chemical metrics pH, alkalinity, dissolved oxygen, ORP, COD, CBOD, conductivity, salinity, TN, TP, sample temperature</p> <p>Water quality metrics Flow rate, turbidity, TSS, TDS, TS, fecal and urinary contamination</p>	<b>Pathogen diagnostic testing</b>  <table border="1"> <thead> <tr> <th>isolate name</th> <th>gene</th> <th>target presence</th> <th>value</th> <th>unit</th> <th>method</th> </tr> </thead> <tbody> <tr> <td>Sample A</td> <td>SARS-CoV2</td> <td>+</td> <td>present</td> <td>22</td> <td>ct</td> <td>qPCR</td> </tr> <tr> <td>Sample B</td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td>qPCR</td> </tr> <tr> <td>Sample C</td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td>bacteria culture</td> </tr> <tr> <td>Sample D</td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td>qPCR</td> </tr> <tr> <td>Sample E</td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td>qPCR</td> </tr> <tr> <td>Sample F</td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td>-</td> <td>bacteria culture</td> </tr> </tbody> </table>	isolate name	gene	target presence	value	unit	method	Sample A	SARS-CoV2	+	present	22	ct	qPCR	Sample B	-	-	-	-	-	qPCR	Sample C	-	-	-	-	-	bacteria culture	Sample D	-	-	-	-	-	qPCR	Sample E	-	-	-	-	-	qPCR	Sample F	-	-	-	-	-	bacteria culture
isolate name	gene	target presence	value	unit	method																																												
Sample A	SARS-CoV2	+	present	22	ct	qPCR																																											
Sample B	-	-	-	-	-	qPCR																																											
Sample C	-	-	-	-	-	bacteria culture																																											
Sample D	-	-	-	-	-	qPCR																																											
Sample E	-	-	-	-	-	qPCR																																											
Sample F	-	-	-	-	-	bacteria culture																																											
<b>Lineage / clade information</b>  <p>Analysis results lineage/clade name and analysis report filename</p> <p>Software used lineage/clade software name and version</p>	<b>AMR detection information</b>  <p>Analysis results analysis report filename</p> <p>Software and database used analysis software name and version, reference database name and version</p>																																																
<b>Strain and isolation information</b>  <p>IDs to describe the isolate isolate ID, alternative isolate ID, progeny isolate ID</p> <p>Isolation method / process microbiological method, strain, isolated by, isolation date, date of receipt</p> <p>Serotyping process and results serovar, serotyping method, phagotype</p>	<b>Taxonomic identification information</b>  <p>Software and database used read mapping software name and version, taxonomic reference database name and version</p> <p>Analysis results taxonomic analysis report filename</p> <p>Methods read mapping criteria, taxonomic analysis date</p>																																																

# Nucleic Acids Research Database Issue



## Nucleic Acids Research

Issues Section browse ▾ More Content ▾ Submit ▾ Purchase About ▾ Nucleic Acids Research

  
Volume 53, Issue D1  
6 January 2025

**JOURNAL ARTICLE**

### The 2025 Nucleic Acids Research database issue and the online molecular biology database collection

Daniel J Rigden , Xosé M Fernández

*Nucleic Acids Research*, Volume 53, Issue D1, 6 January 2025, Pages D1–D9, <https://doi.org/10.1093/nar/gkae1220>

Published: 10 December 2024 Article history ▾

 PDF  Split View  Cite  Permissions  Share ▾

**Article Contents**

- Abstract
- New and updated databases
- NAR online molecular biology database collection
- Acknowledgements
- Funding
- References

 Comments (0)

[Next >](#)

**Abstract**

The 2025 Nucleic Acids Research database issue contains 185 papers spanning biology and related areas. Seventy three new databases are covered, while resources previously described in the issue account for 101 update articles. Databases most recently published elsewhere account for a further 11 papers. Nucleic acid databases include EXPRESSO for multi-omics of 3D genome structure (this issue's chosen Breakthrough Resource and Article) and NAIRDB for Fourier transform infrared data. New protein databases include structure predictions for human isoforms at ASpdb and for viral proteins at BFVD. UniProt, Pfam and InterPro have all provided updates: metabolism and signalling are covered by new descriptions of STRING, KEGG and CAZy, while updated microbe-oriented databases include Enterobase, VFDB and PHI-base. Biomedical research is supported, among others, by ClinVar, PubChem and DrugMAP. Genomics-related resources include Ensembl, UCSC Genome Browser and dbSNP. New plant databases cover the Solanaceae (SolR) and

# 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](http://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

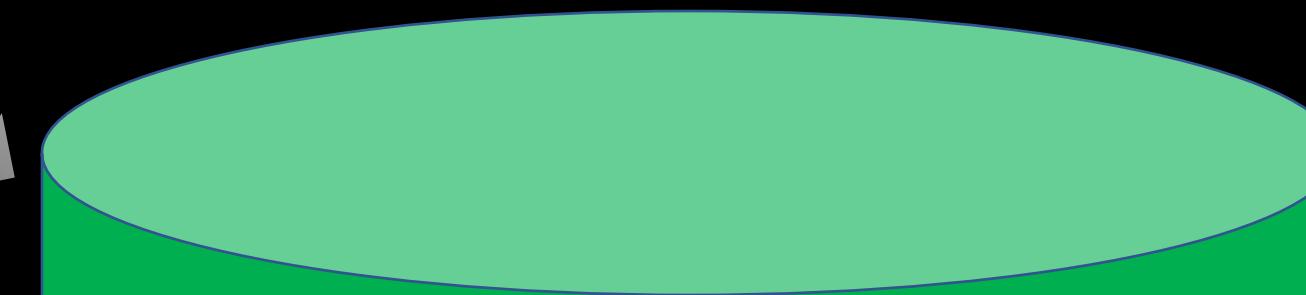
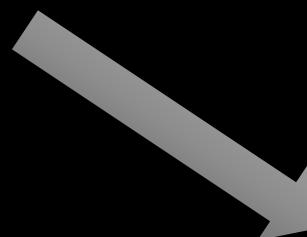
**Whole-genome shotgun: 24,310,993,199,448 bases**

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...

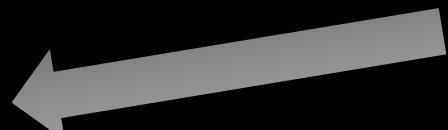
mystery gene



GenBank

## Homologous 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...

Score each pair of residues – consider every possible alignment

NQARP

DEAKP



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

	D	E	A	K	P
N					
Q					
A					
R					
P					

	D	E	A	K	P
N					
Q					
A			seed		
R					
P					

# 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

FASTA uses a **lookup table** to store  $k$ -tuple values

NQARP

AR	3
NQ	1
QA	2
RP	4

DQATS

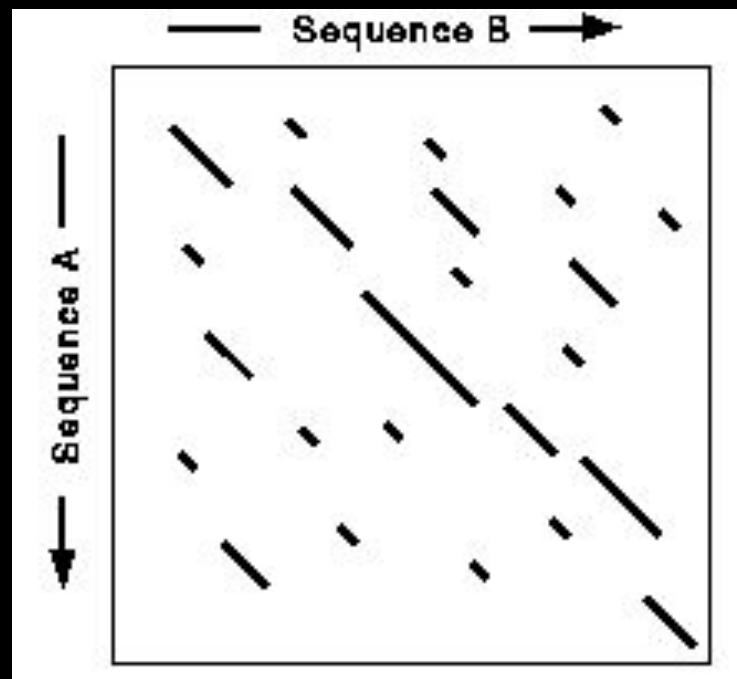
AT	3
DQ	1
QA	2
TS	4

$$\text{Offset} = \text{start(QA, NQARP)} - \text{start(QA, DQATS)} = 0$$

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)



$W+W = \text{high score}$   
 $A+A = \text{low score}$

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

# BLAST

- Basic Local Alignment Search Tool
- FASTA isn't fast enough!
- Can we trade away small amounts of optimality for further gains in performance?

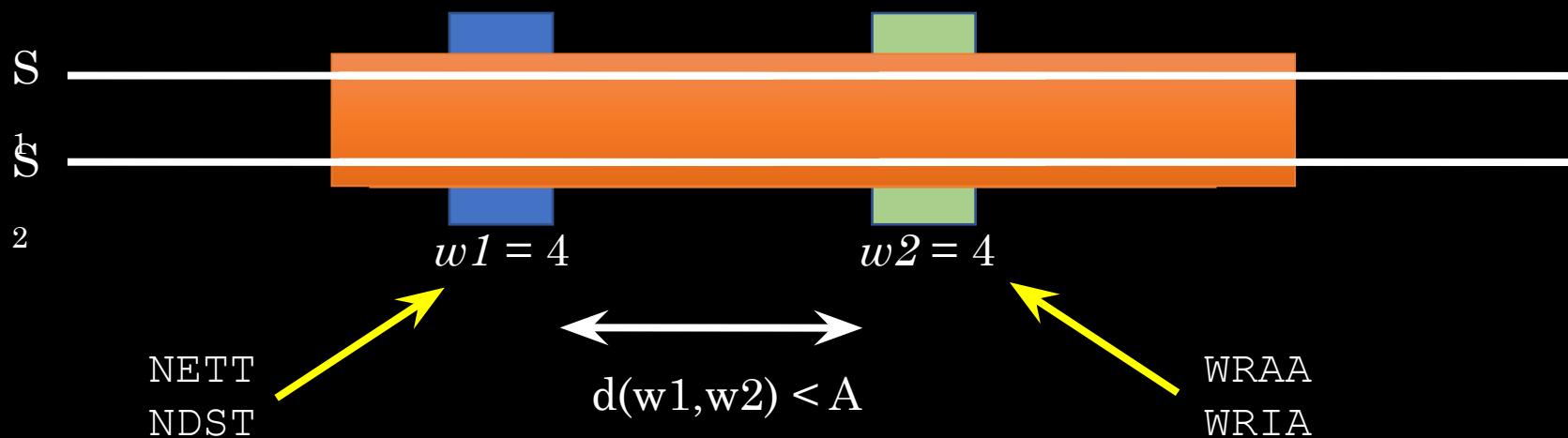
# Basic Principles of BLAST

- 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

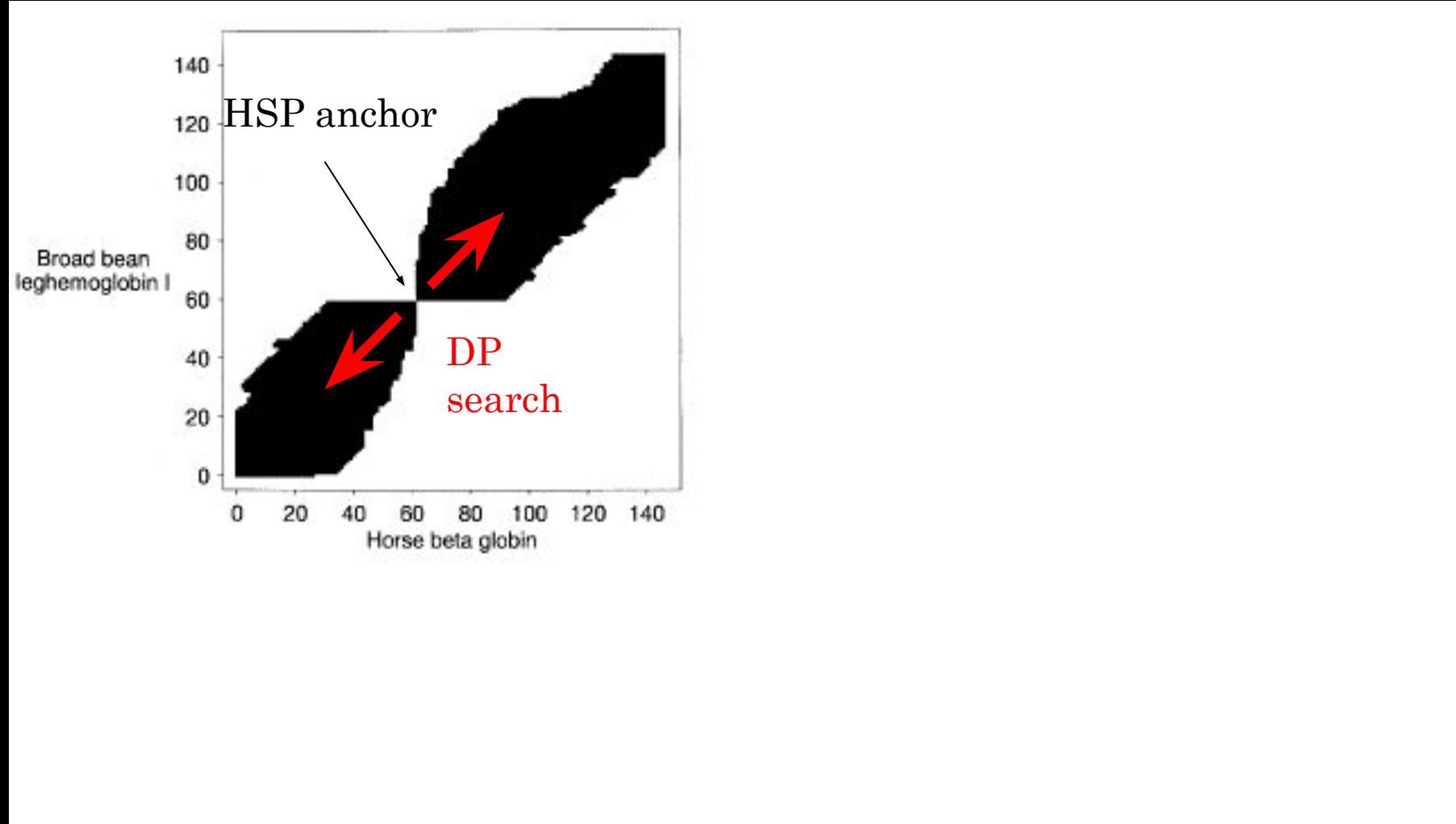
# Extending to high scoring pairs



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

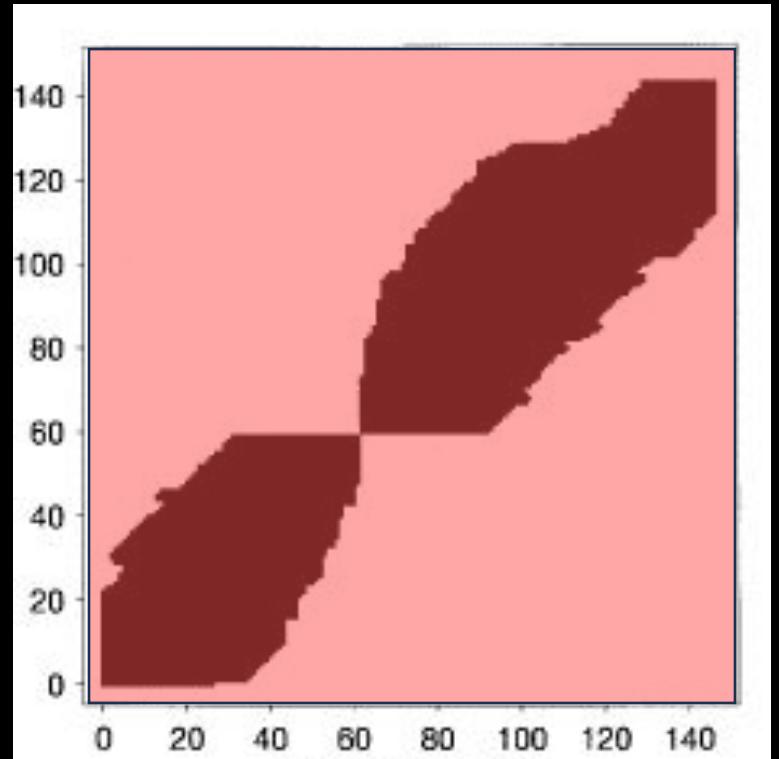
# Gaps

- 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



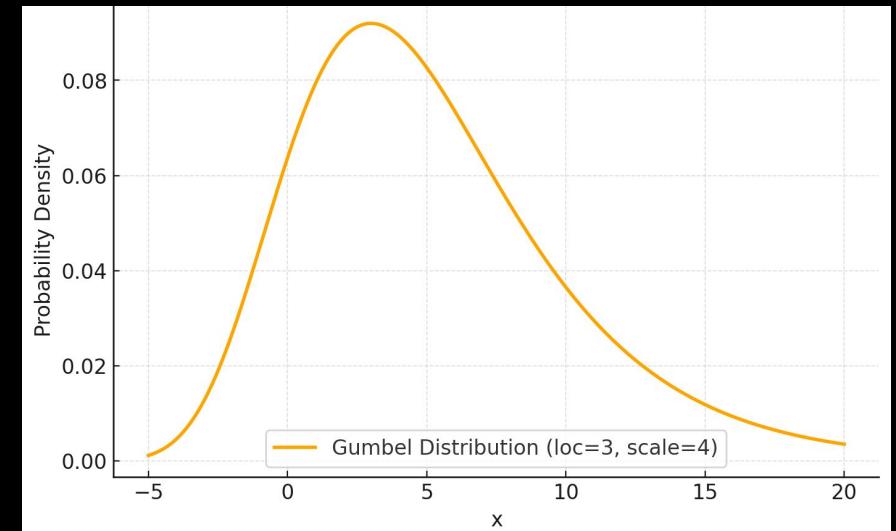
# What have we achieved?

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



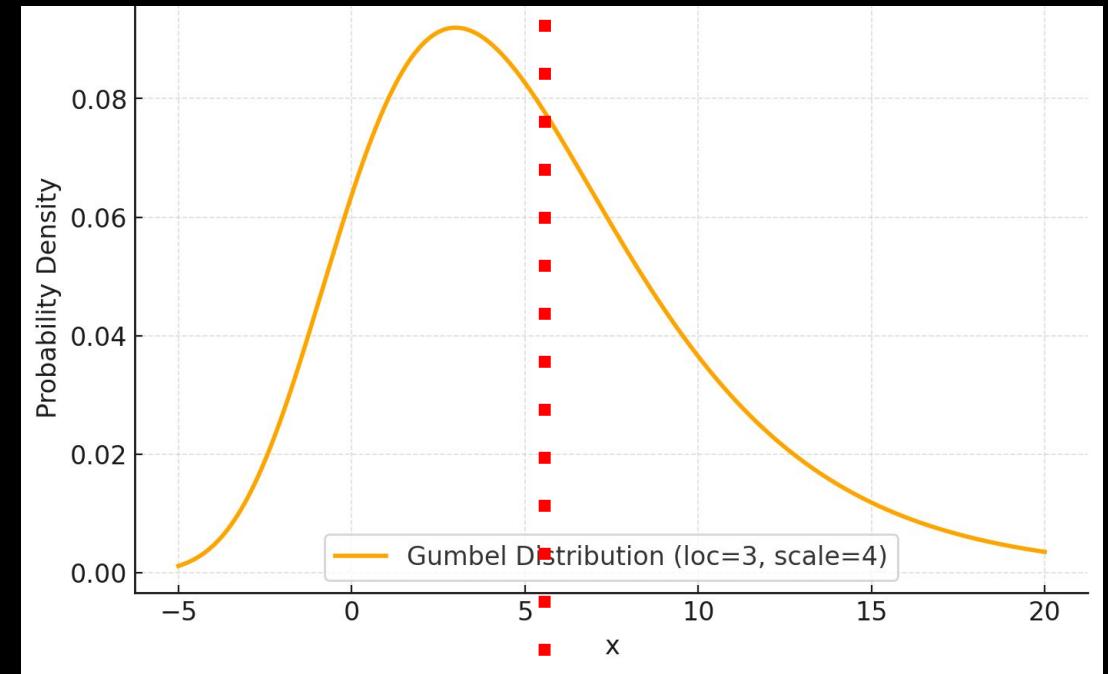
# Local alignment significance

- 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  $\lambda$  and  $K$  that define this distribution



# Karlin-Altschul statistics

The probability of getting a score greater than this:

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

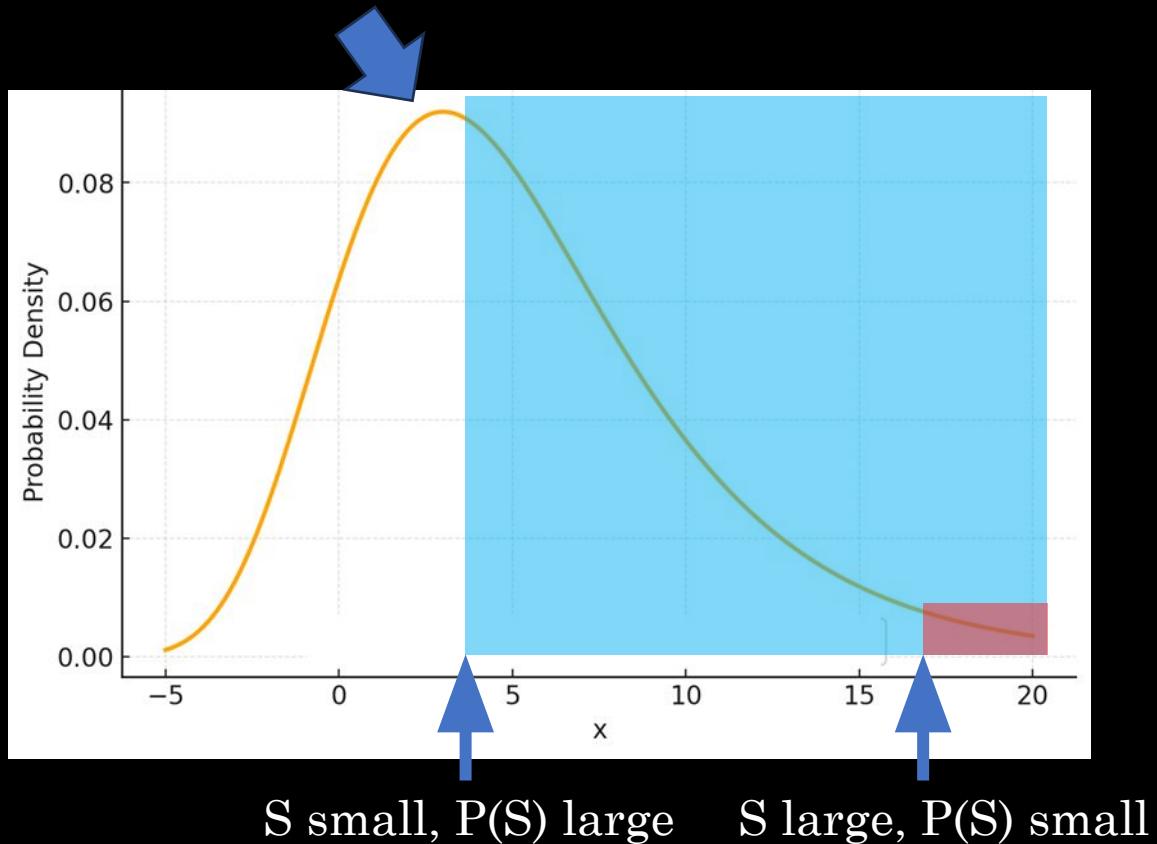
Is roughly equal to this

Where:

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

# Karlin-Altschul statistics

The substitution matrix determines the shape of the curve



# Karlin-Altschul statistics

- Different matrices (PAM, BLOSUM, etc.) define different EVDs – different  $K$  and  $\lambda$
- 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>

The screenshot shows the NCBI BLASTP search interface. The top navigation bar includes tabs for blastn, blastp (which is selected), blastx, tblastn, and tblastx. The main search form is titled 'Enter Query Sequence' and includes fields for 'Enter accession number(s), gi(s), or FASTA sequence(s)', 'Query subrange', 'Or, upload file', 'Job Title', and 'Align two or more sequences'. Below this is the 'Choose Search Set' section, which allows selecting a 'Database' (Non-redundant protein sequences (nr)), 'Organism' (Optional), and 'Exclude' (Models (XM/XP), Non-redundant RefSeq proteins (WP), Uncultured/environmental sample sequences). The 'Program Selection' section contains a list of algorithms: Quick BLASTP, blastp (selected), PSI-BLAST, PHI-BLAST, and DELTA-BLAST. The 'Algorithm parameters' section includes fields for 'Max target sequences' (set to 100), 'Short queries' (checked for 'Automatically adjust parameters for short input sequences'), 'Expect threshold' (0.05), 'Word size' (6), and 'Max matches in a query range' (0). The 'Scoring Parameters' section includes 'Matrix' (BLOSUM62), 'Gap Costs' (Existence: 11 Extension: 1), and 'Compositional adjustments' (Conditional compositional score matrix adjustment). The 'Filters and Masking' section includes 'Filter' (Low complexity regions) and 'Mask' (Mask for lookup table only, Mask lower case letters). At the bottom is a 'BLAST' search button and a link to 'Search database nr using Blastp (protein-protein BLAST)'.

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)

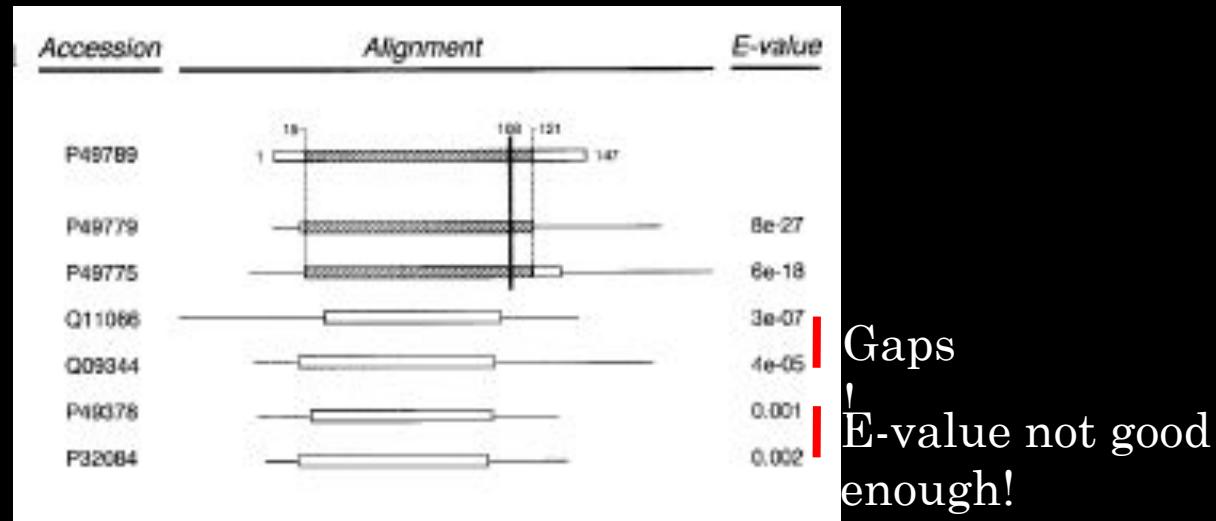
	1	2	3	4	5	6
A	0	0.01	0.01	0.02	0.05	0
C	0	0.8	0.03	0	0.09	0
D	0	0.03	0.6	0.01	0.01	0.2
E	0	0.01	0.1	0	0.15	0.02
...	...	...	...	...	...	...
$\Sigma$	1.0	1.0	1.0	1.0	1.0	1.0

# PSI-BLAST: Step 1

Run BLAST!

# PSI-BLAST: Step 2

- Collapse significant local alignments into a multiple alignment



# 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  $-\infty$  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

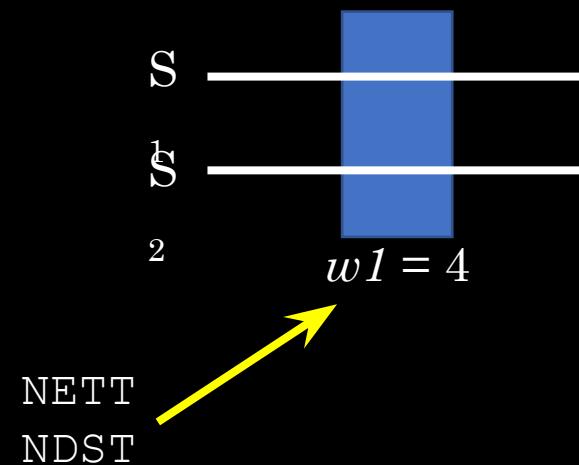
- BLAST isn't fast enough!
- Can we (etc...)



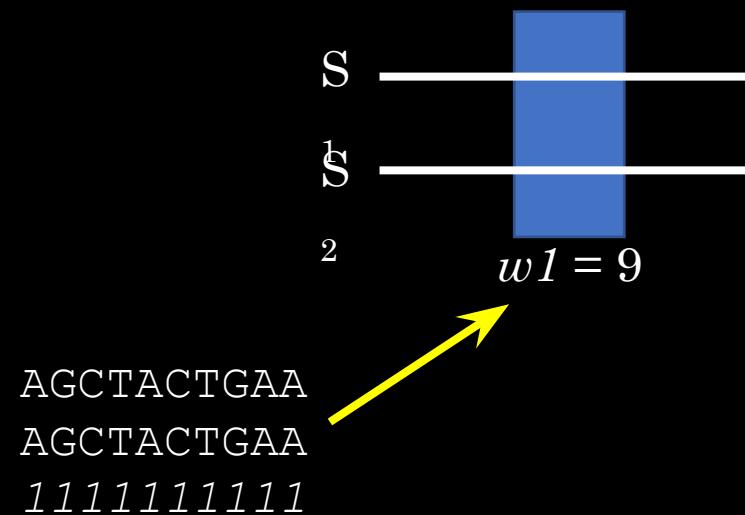
**MEGABLAST!!!**

# Reminder: seeds

Amino acids:  
Seeds with matrix score > threshold



Nucleotides:  
Seeds with literal-matching “model”



# Nucleotide seed length



# BLASTN!

## 1111111111

BLASTN is good for distant-ish sequences but kinda slow



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

AAACGATCCGAAAAGTTT  
GCACGATCCGAAAATCC  
11111111111

# Discontiguous Words

Specific model does **not** need to be contiguous

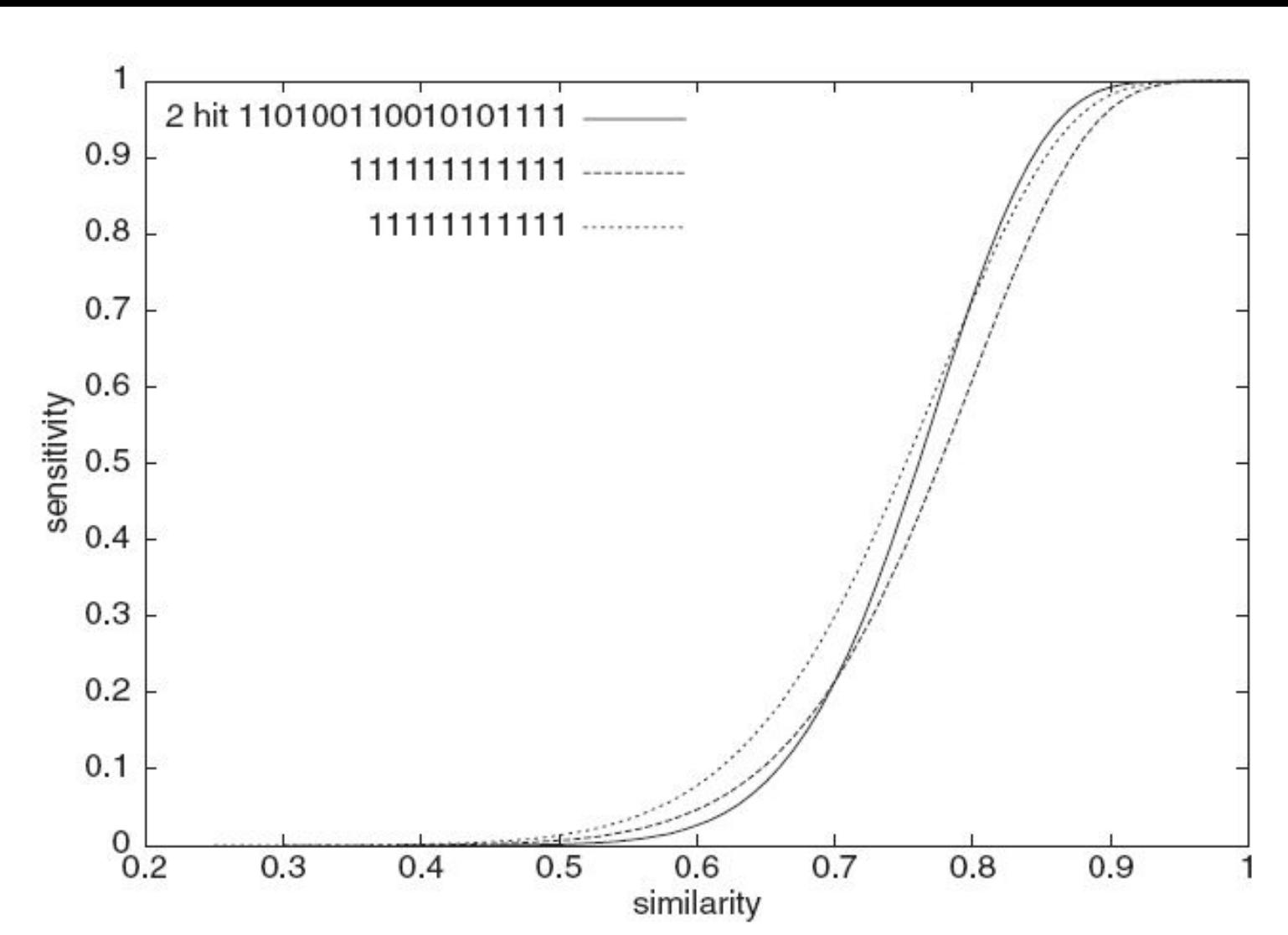
AAACGAAACAGAGAGTTTC

AAATGATCCGAAAGCTTC

111010010100110111

# Comparing models

## PatternHunter – commercial variant



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

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

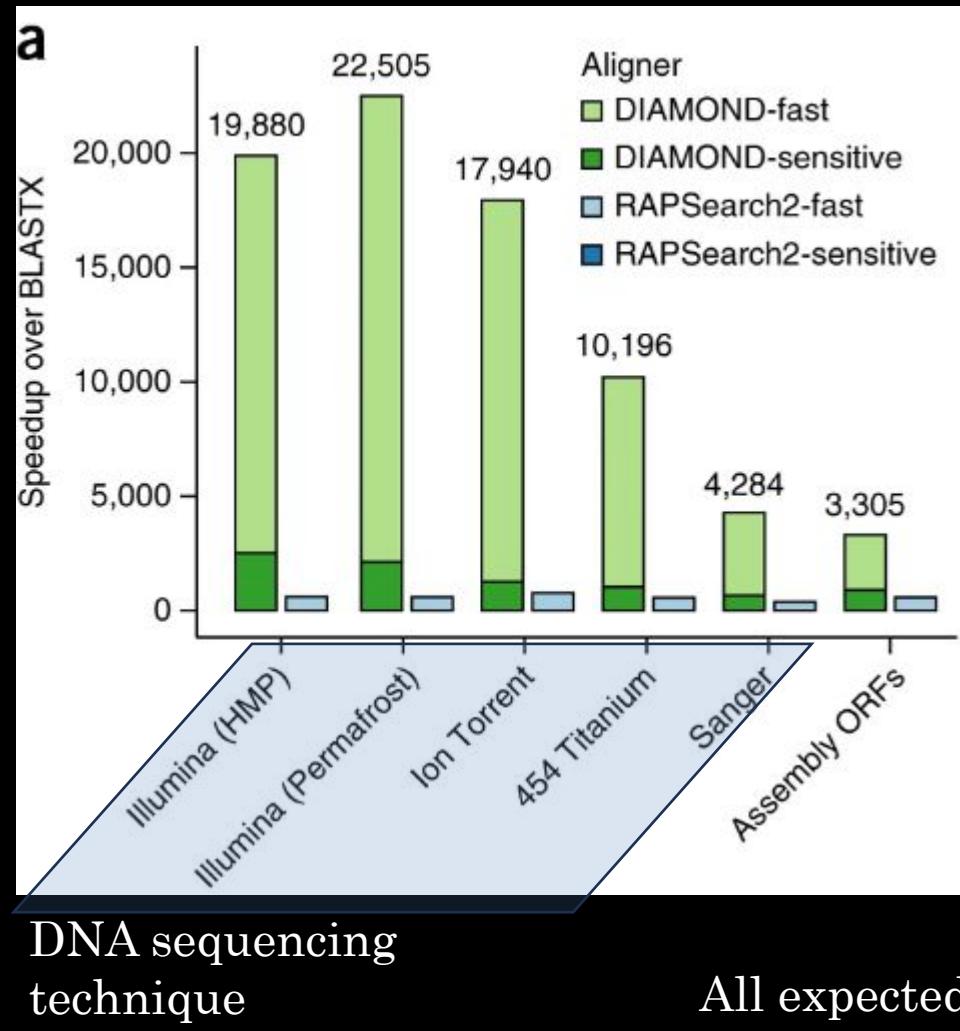
But it costs  
money!

# Other important issues

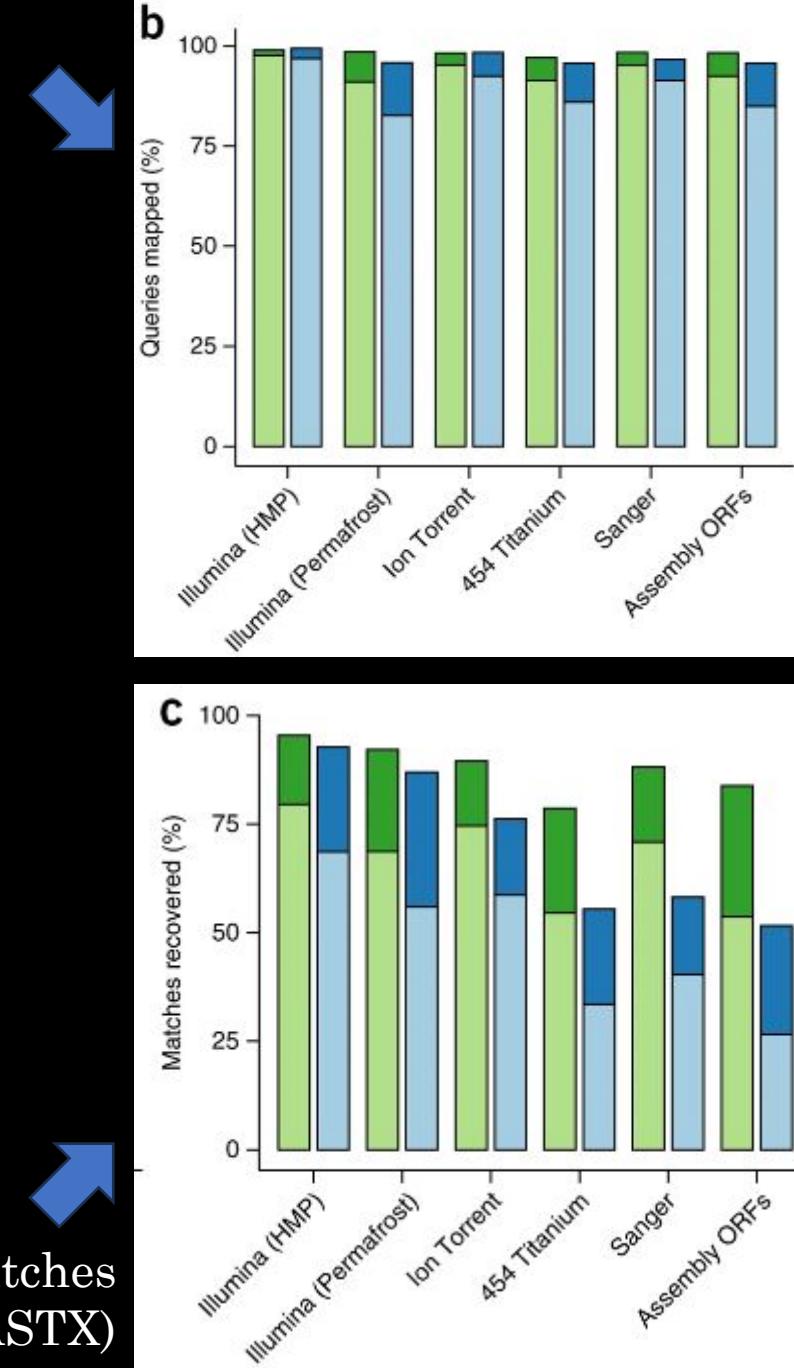
- Low complexity sequence (e.g., AGAGAGAG) can lead to inflated statistics and should be removed prior to the search
- We are still dependent on the choice of substitution matrix!

# 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

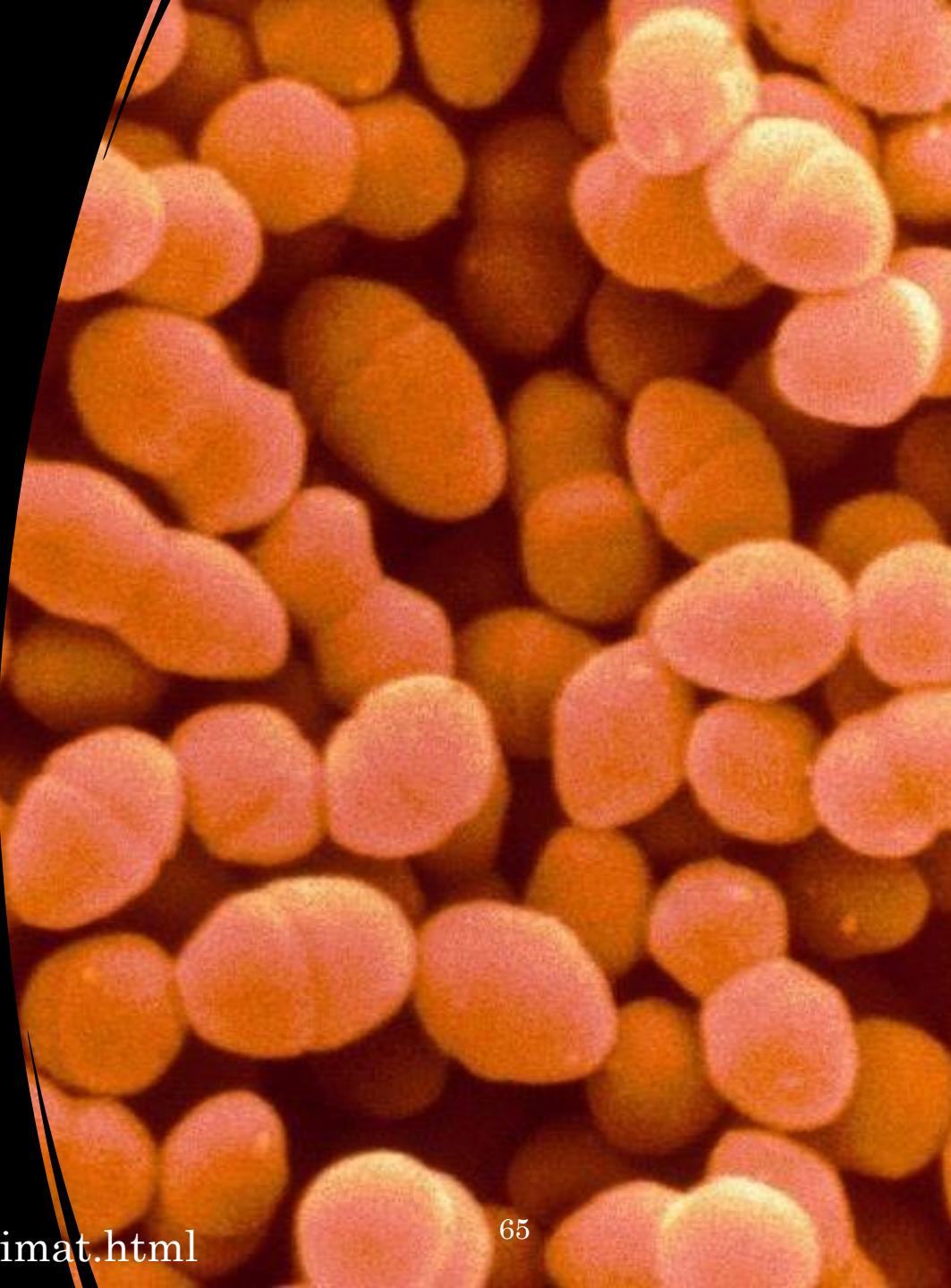


At least one match



# 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?



# 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_002233004.1	[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 luteae]	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

