

FAST  
DATABASE  
SEARCHES

# Overview

- The challenges
  - Lots o' sequences
  - Changing databases
- Local search methods
  - BLAST: seeded searches
    - Plain old BLAST
    - Discontiguous MEGABLAST!!!
    - PSI-BLAST
  - Burrows-Wheeler alignment

# Sequence Databases

Store several different types of sequence data:

## DNA sequences

(individual genes, genome fragments, complete genomes)

## Protein sequences

Usually inferred from corresponding gene sequence

## RNA sequences

Snapshot of what cell(s) are doing - splicing complexity

# Considerations

Data type (duh!), size and **provenance**

**Modes of access**: queries, browsing, APIs

Documentation / stability / support / **persistence**

**Reliability** of information

- PubMed
- Protein
- Nucleotide
- CoreNucleotide
- EST
- GSS
- Structure
- Genome
- Books
- CancerChromosomes
- Conserved Domains
- 3D Domains
- Gene
- Genome Project
- dbGaP
- GENSAT
- GEO Profiles
- GEO Datasets
- HomoloGene
- Journals

**What does NCBI do?**

Founded in 1988 as a national resource for biology information, NCBI creates databases, conducts research in molecular biology, develops software tools for genome data, and disseminates information - all for the bettering of molecular processes and human health and disease. [More...](#)

- Hot Spots**
- ▶ Assembly Archive
  - ▶ Clusters of orthologous groups
  - ▶ Coffee Break, Genes & Disease, NCBI Handbook
  - ▶ Electronic PCR

**Bank vs. RefSeq**

... about the distinctions between **Bank** and **UniProt**? [Click here for a](#) of the databases and their dif

SITE MA  
Alphabetic  
Resource  
About N  
An introdu  
NCBI  
GenBan  
Sequence  
submissio  
and softw  
Literatur  
databas  
PubMed,  
Books, ar  
Central

Molecular  
databases  
Sequences,  
structures, and  
taxonomy  
Genomic biology  
The human genome,  
whole genomes,  
and related  
resources

**New Protein Clusters**  
Entrez Protein Clusters database

The new Entrez Protein Clusters database is a collection of Reference Sequence (RefSeq) proteins, from the complete genomes of prokaryotes, plasmids, and organelles, that have been grouped and annotated based on sequence similarity to predict protein function. [Click here to find out more about the Protein Clusters database.](#)

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

All Databases

**Welcome to NCBI**  
The National Center for Biotechnology Information advances science and health by providing access to biomedical and genomic information.

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

**Submit**

Deposit data or manuscripts into NCBI databases

**Download**

Transfer NCBI data to your computer

**Learn**

Find help documents, attend a class or watch a tutorial

**Develop**

Use NCBI APIs and code libraries to build applications

**Analyze**

Identify an NCBI tool for your data analysis task

**Research**

Explore NCBI research and collaborative projects

**COVID-19 Information**  
Public health information (CDC) | Research information (NIH) | SARS-CoV-2 data (NCBI) | Prevention and treatment information (HHS) | [Espace](#)

**Popular Resources**  
PubMed  
Bookshelf  
PubMed Central  
BLAST  
Nucleotide  
Genome  
SNP  
Gene  
Protein  
PubChem

**NCBI News & Blog**  
NCBI-NAID Beyond Phylogenies Codeathon was a success! 23 Jan 2023  
SARS-CoV-2 genomic data is critical for monitoring the viral spread and evolution of the COVID-19 pandemic, identifying newly emerging variants, and developing...  
Full-scale access to microbial Pathogen Detection data in the Cloud! 18 Jan 2023  
NCBI's Pathogen Detection resource now provides selected data on the Google Cloud Platform (GCP) allowing you better access to over 1 million bacterial...  
RefSeq Release 216 17 Jan 2023

## National Center for Biotechnology Information (GenBank)

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

EMBL-EBI

The home for big data in biology

27 million  
Average requests per day to EMBL-EBI websites.

More about EMBL-EBI's impact in our annual report >  
Data from 2016

Our unique Search service helps you explore dozens of biological data resources.  
[More about EBI Search >](#)

Find a tool for your data analysis. [\( Find a tool >](#)

Share your scientific data with the world. [Deposit data >](#)

Find a gene, protein or chemical

Searches: [blast](#) [keratin](#) [bfl1](#)

EMBL-EBI and our mission

EMBL-EBI shares data from life science experiments, performs basic research in computational biology and offers an extensive user training programme for participating researchers in academia and industry. We are part of EMBL, Europe's flagship laboratory for the life sciences. [More about our impact >](#)

**R** Research  
We contribute to the advancement of biology through basic investigator-driven research >

**T** Training  
We provide advanced bioinformatics training to scientists at all levels >

**I** Industry  
We help disseminate cutting-edge technologies to industry >

**% ELIXIR**  
We support, as an ELIXIR node, the coordination of biological data provision throughout Europe >

Latest news  
[Research highlights, service updates and more](#)

Our events  
Tues 8th Mar - Thurs 8th | [Course](#)  
[Bioinformatics Resources for Protein Biology](#)

Nucleotides,  
Genomes,  
**Protein function,**  
Protein-protein  
interactions

# European Molecular Biology Laboratory – European Bioinformatics Institute

# CARD

Use or Download Copyright & Disclaimer  
Help Us Curate #AMRCuration #WorkTogether

[Browse](#) [Analyze](#) [Download](#) [About](#)

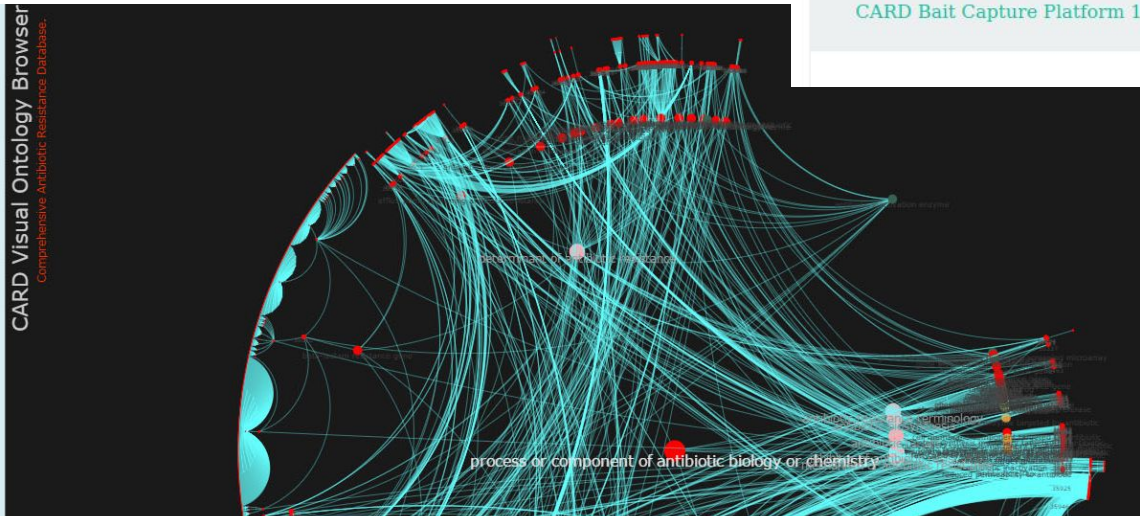
## The Comprehensive Antibiotic Resistance Database

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

6657 Ontology Terms, 5031 Reference Sequences, 1931 SNPs, 3013 Publications, 5078 AMR Detection Models

Resistome predictions: 377 pathogens, 21079 chromosomes, 2662 genomic islands, 41828 plasmids, 155606 WGS assemblies, 322710 alleles

[CARD Bait Capture Platform 1.0.0](#) | [State of the CARD 2021 Presentations & Demonstrations](#)



CARD  
the Comprehensive Antibiotic Resistance  
Database

Genes (>5000)  
Custom homology tool  
Carefully curated **ontology**




Studies <sup>i</sup>	<a href="#">23,258</a>
Biosamples <sup>i</sup>	<a href="#">79,418</a>
Sequencing Projects <sup>i</sup>	<a href="#">79,471</a>
Analysis Projects <sup>i</sup>	<a href="#">65,042</a>

[Download Excel Data file](#)  
File last generated: 22 Feb, 2016

## Welcome to the Genomes OnLine Database

**GOLD Release v.5**

**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.

<h3>1. Register</h3>  <p>Register your project information and Metadata in the Genomes Online Database</p> <p><a href="#">Register</a></p>	<h3>2. Annotate</h3>  <p>Annotate your microbial genome or metagenome with IMG/ER or IMG/MER</p> <p><a href="#">Annotate</a></p>	<h3>3. Publish</h3>  <p>Standards in Genomic Sciences</p> <p>Publish your genome or metagenome in open access standards-supportive journal.</p> <p><a href="#">Publish</a></p>
---	---	---

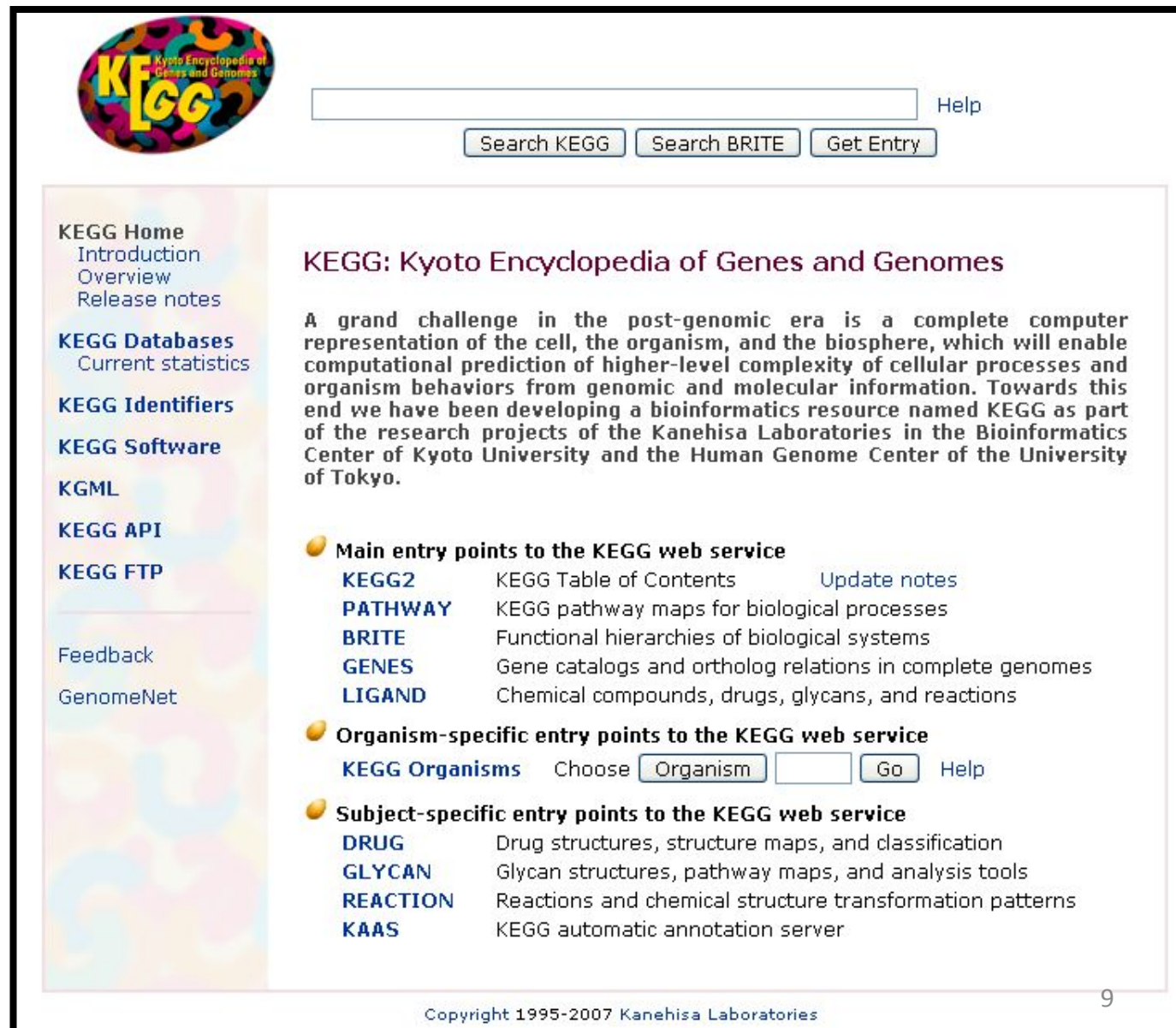
<h4>Studies</h4> <ul style="list-style-type: none"> <li><a href="#">Metagenomic 637</a></li> <li><a href="#">Non-Metagenomic 22,621</a></li> </ul>	<h4>Biosamples</h4> <ul style="list-style-type: none"> <li><a href="#">Classification</a></li> <li>Ecosystems                     <ul style="list-style-type: none"> <li>Host-associated <a href="#">13,350</a></li> <li>Engineered <a href="#">2,480</a></li> <li>Environmental <a href="#">9,470</a></li> </ul> </li> </ul>	<h4>Sequencing Projects</h4> <ul style="list-style-type: none"> <li><a href="#">Complete Projects 8,015</a></li> <li><a href="#">Permanent Drafts 33,298</a></li> <li><a href="#">Incomplete Projects 35,607</a></li> <li><a href="#">Targeted Projects 1,565</a></li> </ul>	<h4>Analysis Projects</h4> <ul style="list-style-type: none"> <li><a href="#">Genome Analysis 48,429</a></li> <li><a href="#">Metagenome Analysis 5,627</a></li> <li><a href="#">Combined Assembly 101</a></li> <li><a href="#">Genome from Metagenome 1,499</a></li> <li><a href="#">Metatranscriptome Analysis 1,280</a></li> <li><a href="#">Single Cell (Screened) 1,681</a></li> <li><a href="#">Single Cell (Unscreened) 781</a></li> <li><a href="#">Transcriptome Analysis 0</a></li> </ul>
<h4>Organisms</h4> <ul style="list-style-type: none"> <li><a href="#">Organisms 72,826</a></li> <li><a href="#">Archaea 1,198</a></li> <li><a href="#">Bacteria 55,157</a></li> </ul>	<h4>Special Projects</h4> <ul style="list-style-type: none"> <li><a href="#">Type Strain Projects 5,328</a></li> <li><a href="#">GEBA Projects 2,517</a></li> <li><a href="#">HMP Projects 2,921</a></li> </ul>	<h4>JGI Projects</h4> <ul style="list-style-type: none"> <li><a href="#">JGI Studies 1,111</a></li> <li><a href="#">JGI Biosamples 19,723</a></li> <li><a href="#">JGI Sequencing Projects 30,927</a></li> </ul>	<h4>Projects with Genbank Data</h4> <ul style="list-style-type: none"> <li><a href="#">Seq. Projects 42,394</a></li> <li><a href="#">Archaeal Projects 564</a></li> <li><a href="#">Bacterial Projects 35,732</a></li> </ul>

# GOLD Genomes Online Database

## Genome projects Standards-compliant metadata



# Kyoto Encyclopedia of Genes and Genomes



The screenshot shows the KEGG website homepage. At the top left is the KEGG logo, a colorful oval with the letters 'KEGG' and 'Kyoto Encyclopedia of Genes and Genomes'. To the right is a search bar with a 'Help' link. Below the search bar are three buttons: 'Search KEGG', 'Search BRITE', and 'Get Entry'. On the left side, there is a navigation menu with links to 'KEGG Home', 'KEGG Databases', 'KEGG Identifiers', 'KEGG Software', 'KGML', 'KEGG API', and 'KEGG FTP'. Below these are 'Feedback' and 'GenomeNet' links. The main content area features the title 'KEGG: Kyoto Encyclopedia of Genes and Genomes' and a paragraph describing the project. Below this are three sections of entry points: 'Main entry points to the KEGG web service', 'Organism-specific entry points to the KEGG web service', and 'Subject-specific entry points to the KEGG web service'. Each section lists various categories with brief descriptions. At the bottom, there is a copyright notice for 1995-2007 Kanehisa Laboratories and a page number '9'.

**KEGG Home**  
Introduction  
Overview  
Release notes

**KEGG Databases**  
Current statistics

**KEGG Identifiers**

**KEGG Software**

**KGML**

**KEGG API**

**KEGG FTP**

Feedback  
GenomeNet

## KEGG: Kyoto Encyclopedia of Genes and Genomes

A grand challenge in the post-genomic era is a complete computer representation of the cell, the organism, and the biosphere, which will enable computational prediction of higher-level complexity of cellular processes and organism behaviors from genomic and molecular information. Towards this end we have been developing a bioinformatics resource named KEGG as part of the research projects of the Kanehisa Laboratories in the Bioinformatics Center of Kyoto University and the Human Genome Center of the University of Tokyo.

- **Main entry points to the KEGG web service**
  - KEGG2** KEGG Table of Contents Update notes
  - PATHWAY** KEGG pathway maps for biological processes
  - BRITE** Functional hierarchies of biological systems
  - GENES** Gene catalogs and ortholog relations in complete genomes
  - LIGAND** Chemical compounds, drugs, glycans, and reactions
- **Organism-specific entry points to the KEGG web service**

KEGG Organisms Choose   Help
- **Subject-specific entry points to the KEGG web service**
  - DRUG** Drug structures, structure maps, and classification
  - GLYCAN** Glycan structures, pathway maps, and analysis tools
  - REACTION** Reactions and chemical structure transformation patterns
  - KAAS** KEGG automatic annotation server

Copyright 1995-2007 Kanehisa Laboratories 9

Genomes  
Orthology information  
Protein functions  
**Biochemical pathways**

Limited access now  
(#%#!)

# A word about “metadata”

\$@\*#(\*!)!!

**nature**  
**biotechnology**

PERSPECTIVE

## Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MlxS) 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 Haneln<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 Kyrpides<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 Nakielnny<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

# 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 182 (February 2011): 124,277,818,310 bases, from 132,015,054 reported sequences

Release 200 (February 2014): 157,943,793,171 bases, from 171,123,749 reported sequences

Release 212 (February 2016): “We’re sorry, but the page cannot be found”

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

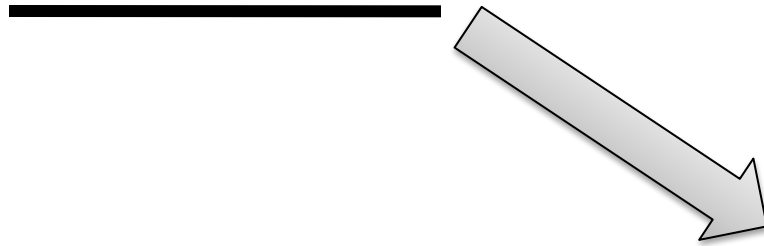
**Whole-genome shotgun: 15,089,161,465,959 bases**

Release 252 (October 2022): 1,562,963,366,851 bases from 240,539,282 sequences

**Whole-genome shotgun: 18,787,298,109,534 bases**

# Sequence of Interest...

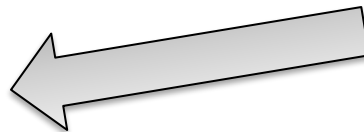
*argH* gene



GenBank

**Homologous sequences:**

- Evolutionary conservation
- Annotated functions
- Presence / absence in other organisms  
(phylogenetic profiles)



# Best 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

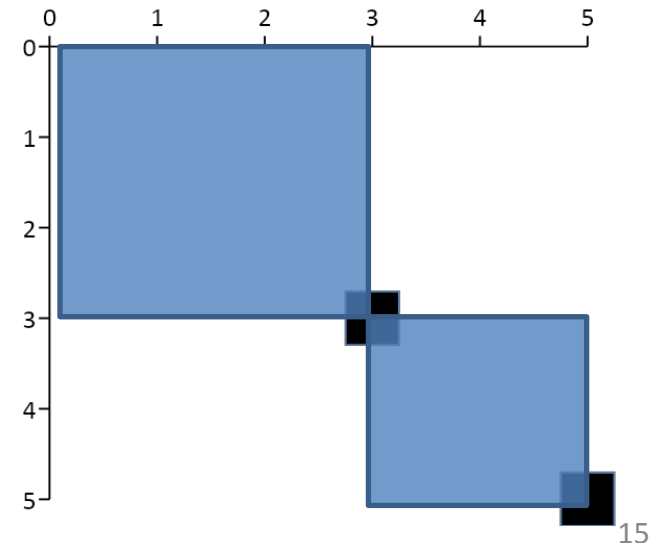


Score each pair of residues –  
consider every possible  
alignment

	D	E	A	K	P
N					
Q					
A					
R					
P					

Require an exact match of length  
 $L$  to “seed” the alignment

OR



# FASTA

(Pearson and Lipman, 1988)

- 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

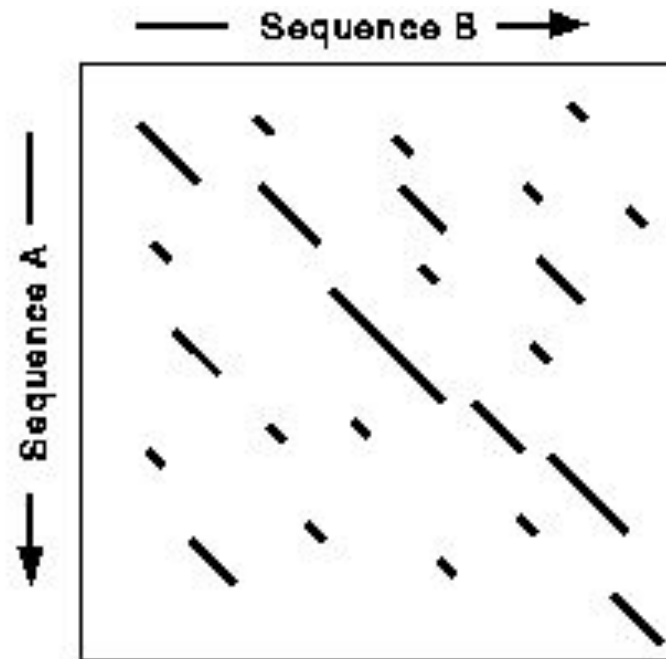
DQATS

AT	3
DQ	1
QA	2
TS	4

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

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)



WW = woo!  
AA = meh

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

# BLAST

(Altschul et al., 1990)

(Altschul et al., 1997)

- **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?

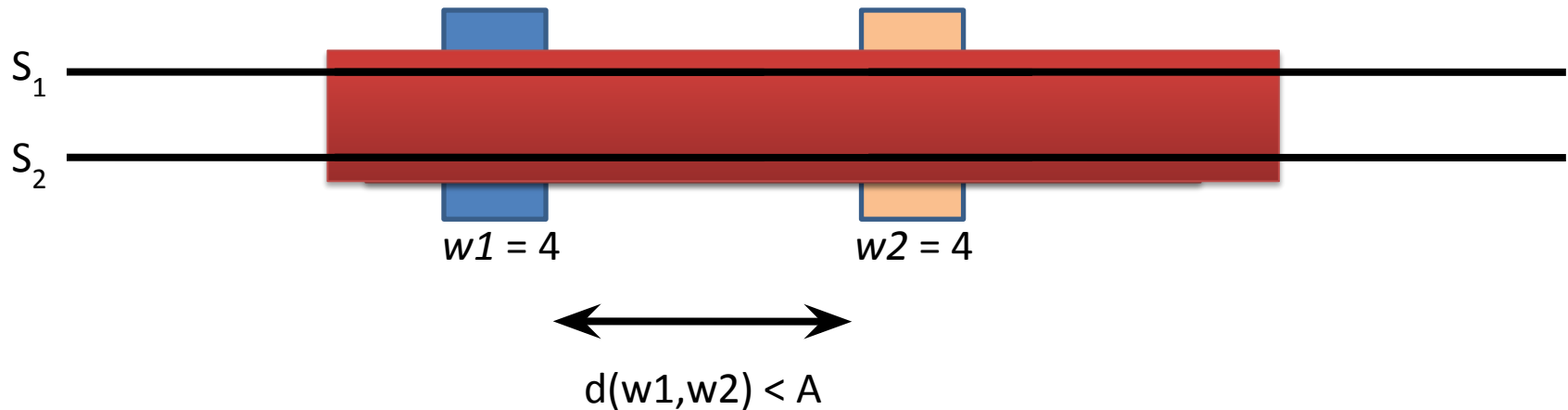
# Basic Principles of BLAST

- Exact matches are great and all, but they're not perfect
- 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 (".". maximal)

# Basic Principles of BLAST

- Instead of running full DP (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

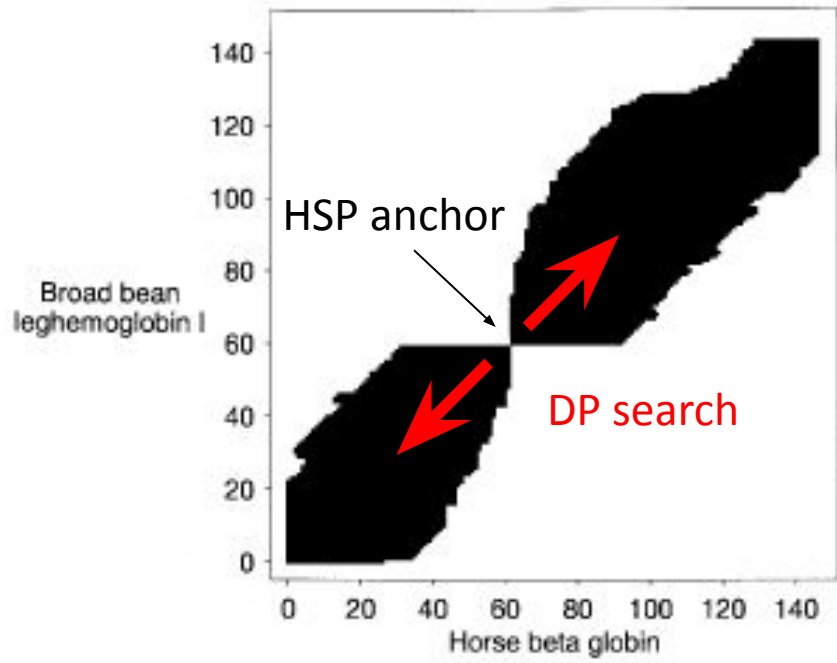
# Extending to high scoring pairs



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

# Gaps

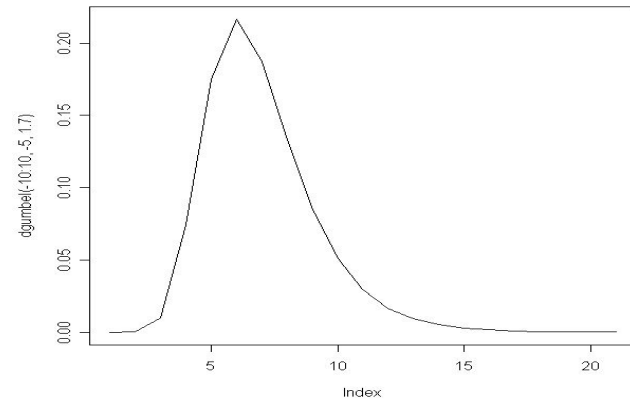
- Start from the middle of the high-scoring pair, and proceed with DP forward and backward until the path falls below a threshold
- DP is expensive, but we've saved ourselves a lot of time!
  - Most sequence pairs are *not* homologous
  - Anchored DP will be a *lot* faster





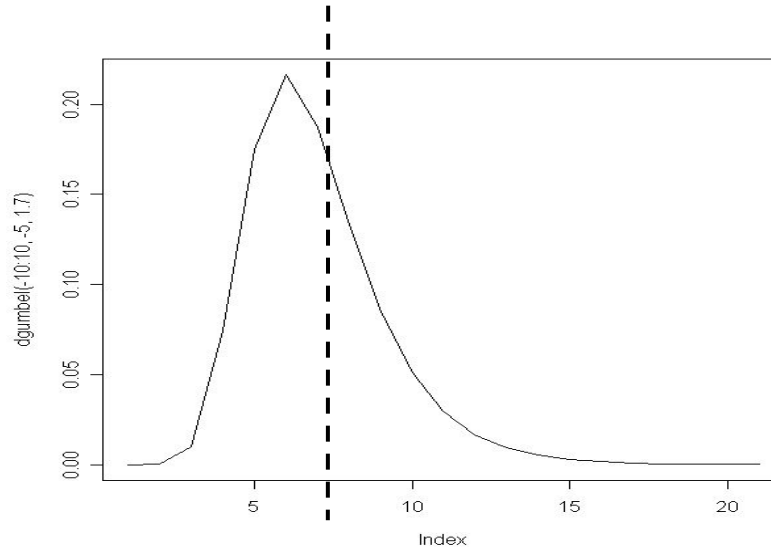
# Local alignment significance

- 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 **extreme value distribution**



# Karlin-Altschul statistics

aka **no permutations, thanks**



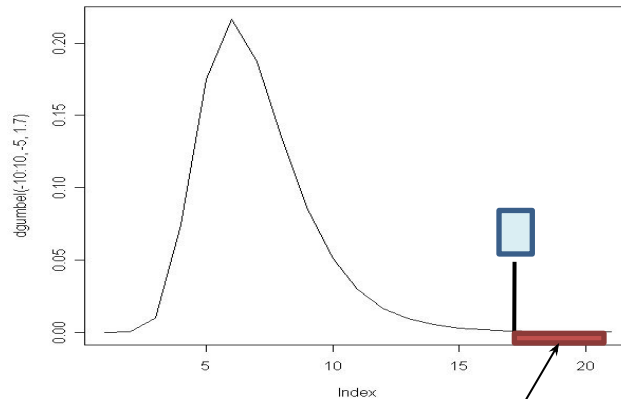
- The expected (=mean) score between a pair of random sequences is the mean of an extreme value distribution
- Given a scoring matrix (such as PAM250) and a set of amino acid frequencies, we can compute the parameters  $\lambda$  and  $K$  that define this distribution

# Karlin-Altschul statistics

Score from  
EVD

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

Observed score



$P(S > \square)$

# 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

blastn

blastp

blastx

tblastn

tblastx

## Enter Query Sequence

Enter accession number(s), gi(s), or FASTA sequence(s) [Clear](#)Query subrange [?](#)

From To 

Or, upload file

 No file selected. [?](#)

Job Title

Enter a descriptive title for your BLAST search [?](#) Align two or more sequences [?](#)

## Choose Search Set

Database

Non-redundant protein sequences (nr) [?](#)

Organism

Optional

 [?](#) excludeEnter organism common name, binomial, or tax id. Only 20 top taxa will be shown [?](#)

Exclude

Optional

 Models (XM/XP)  Non-redundant RefSeq proteins (WP)  Uncultured/environmental sample sequences

## Program Selection

Algorithm

 Quick BLASTP (Accelerated protein-protein BLAST) blastp (protein-protein BLAST) PSI-BLAST (Position-Specific Iterated BLAST) PHI-BLAST (Pattern Hit Initiated BLAST) DELTA-BLAST (Domain Enhanced Lookup Time Accelerated BLAST)Choose a BLAST algorithm [?](#)

BLAST

Search database nr using Blastp (protein-protein BLAST)

 Show results in a new window

## Algorithm parameters

## General Parameters

Max target sequences

100 [?](#)Select the maximum number of aligned sequences to display [?](#)

Short queries

 Automatically adjust parameters for short input sequences [?](#)

Expect threshold

0.05 [?](#)

Word size

6 [?](#)

Max matches in a query range

0 [?](#)

## Scoring Parameters

Matrix

BLOSUM62 [?](#)

Gap Costs

Existence: 11 Extension: 1 [?](#)

Compositional adjustments

Conditional compositional score matrix adjustment [?](#)

## Filters and Masking

Filter

 Low complexity regions [?](#)

Mask

 Mask for lookup table only [?](#) Mask lower case letters [?](#)

## Protein-protein BLAST (BLASTP):

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

Query

Database

Algorithm

Match parameters

Scoring

BLAST

Search database nr using Blastp (protein-protein BLAST)

 Show results in a new window

# PSI-BLAST (1997)

- Replace trusty old PAM or BLOSUM with a **position-specific** scoring matrix
- Iterate query – Position-specific scoring matrix (**PSSM**) procedure

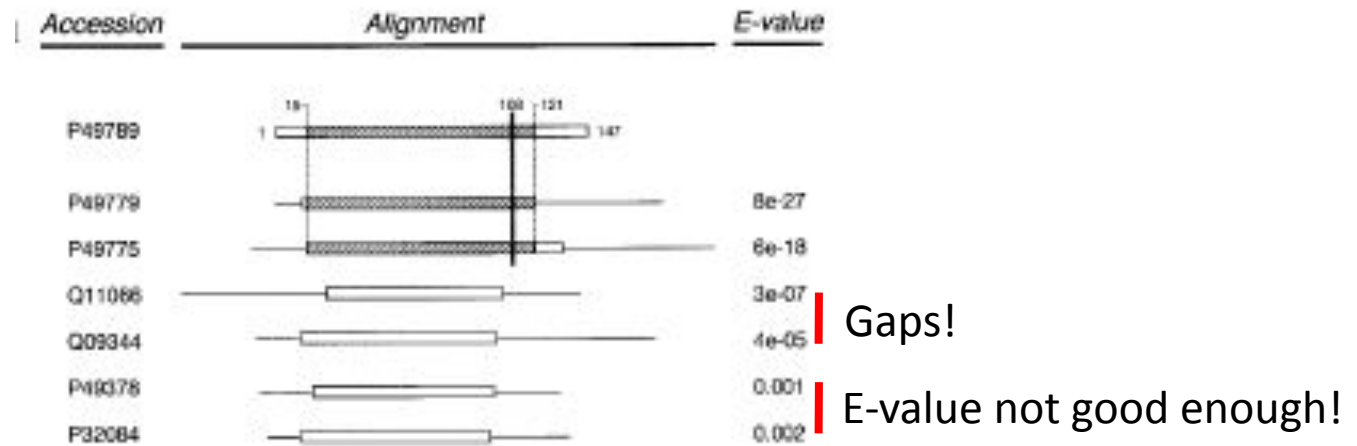
# Step 1

Run BLAST!



# Step 2

- Collapse significant local alignments into a multiple alignment



# 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

# 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

# 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)

# Discontiguous MEGABLAST

and PatternHunter

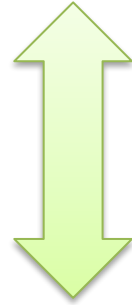
- BLAST isn't fast enough!
- Can we (etc...)

The background features a central, bright white and yellow explosion that radiates outwards. Numerous thin, golden-yellow lines of light extend from the center towards the corners of the frame, creating a sense of intense energy and motion. The overall color palette is dominated by warm tones of orange, yellow, and black.

**MEGABLAST!!!**



BLASTN is good for distant-ish sequences  
(but why not use BLASTP?) but kinda slow



Happy medium???

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



# Continuous Words

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

AA**ACGATCCGAAA**GTTT

GC**ACGATCCGAAA**ATCC

# Discontinuous Words

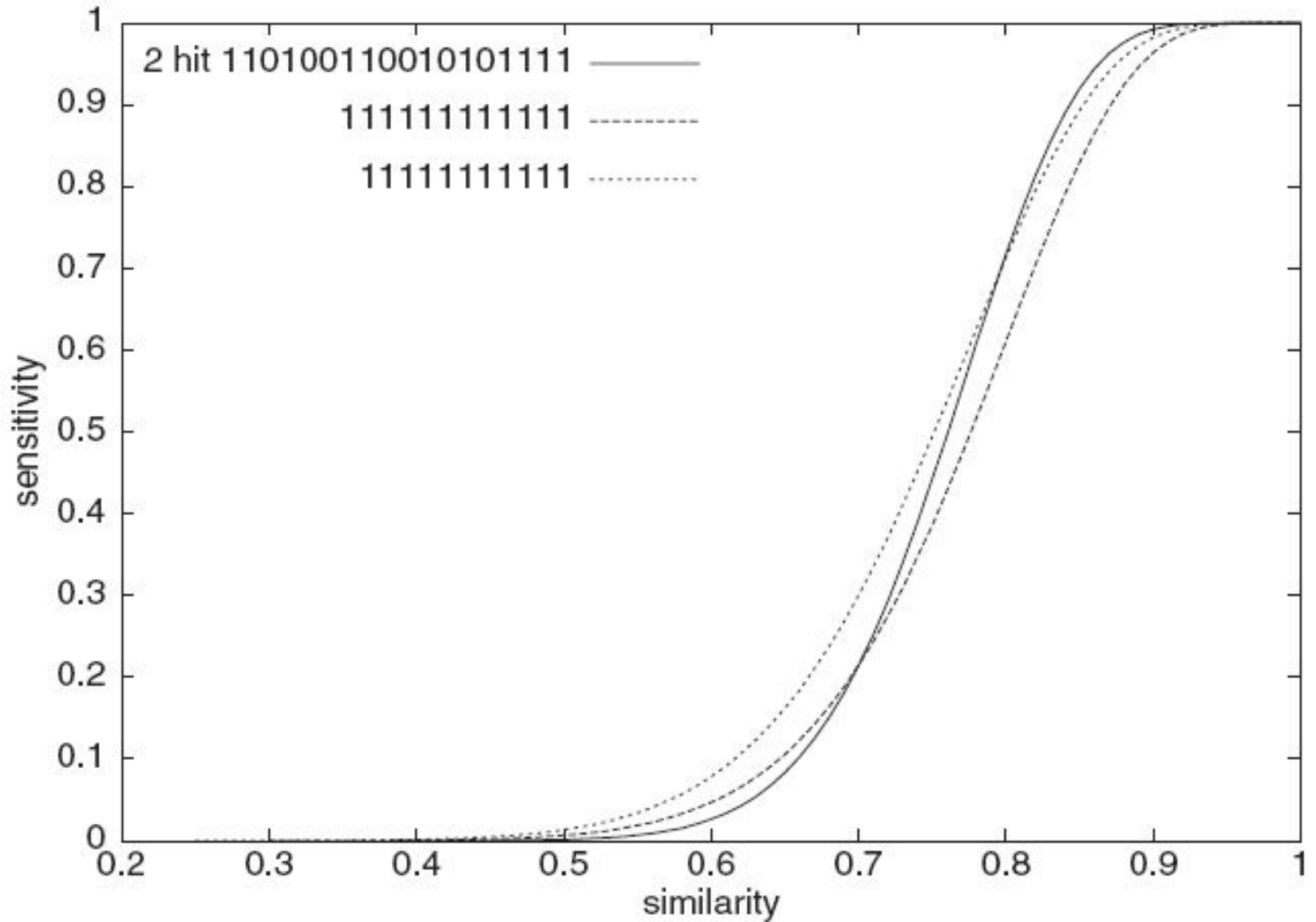
Search for words defined by a 'model':

Model: 111010010100110111

**AAACGAACAGAGAGTTTC**

**AAATGATCCGAAAGCTTC**

# Similar accuracy



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	∞
<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	∞	∞

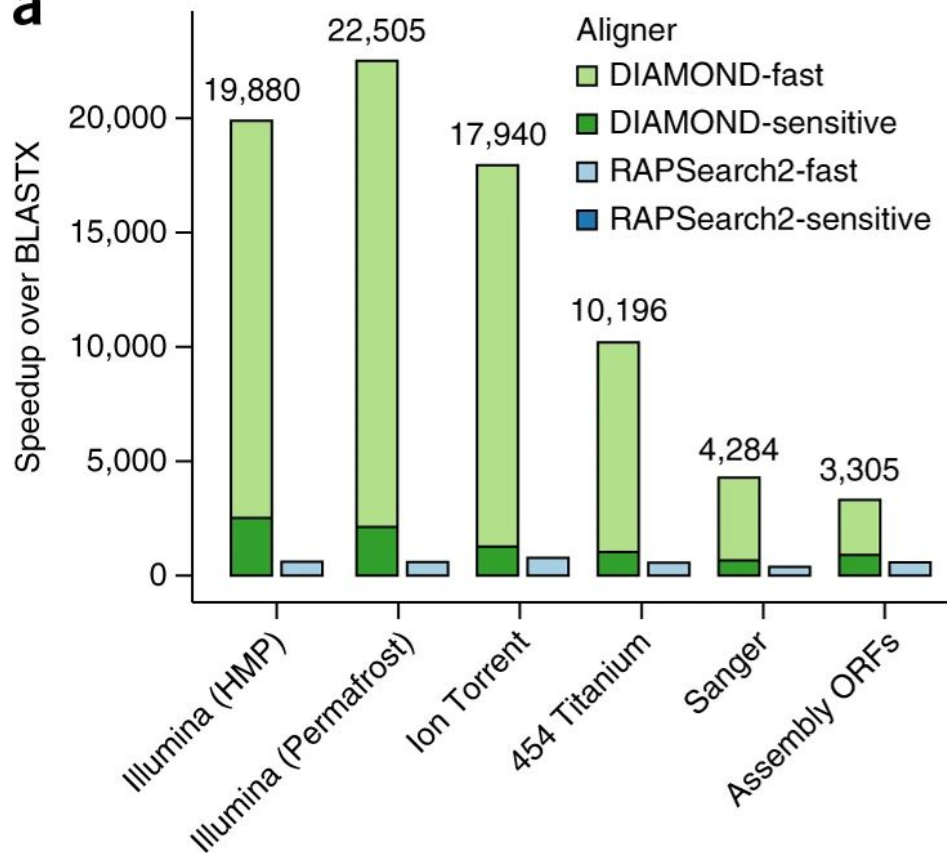
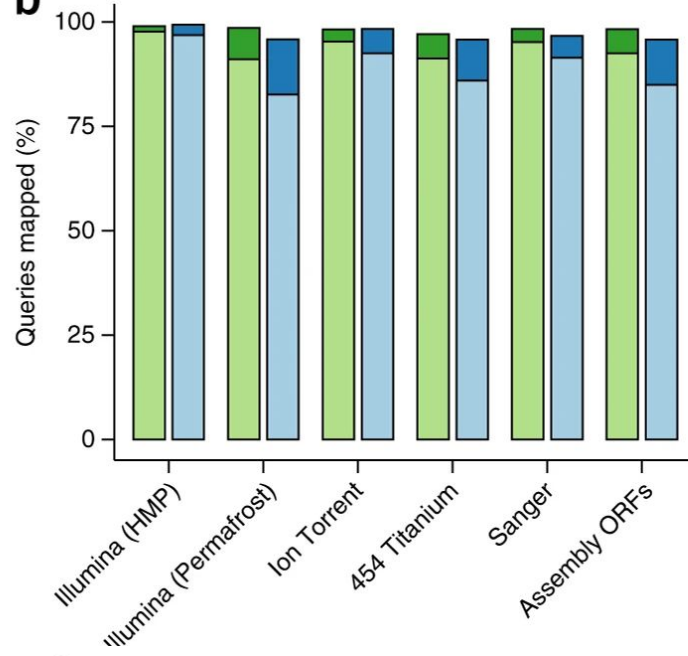
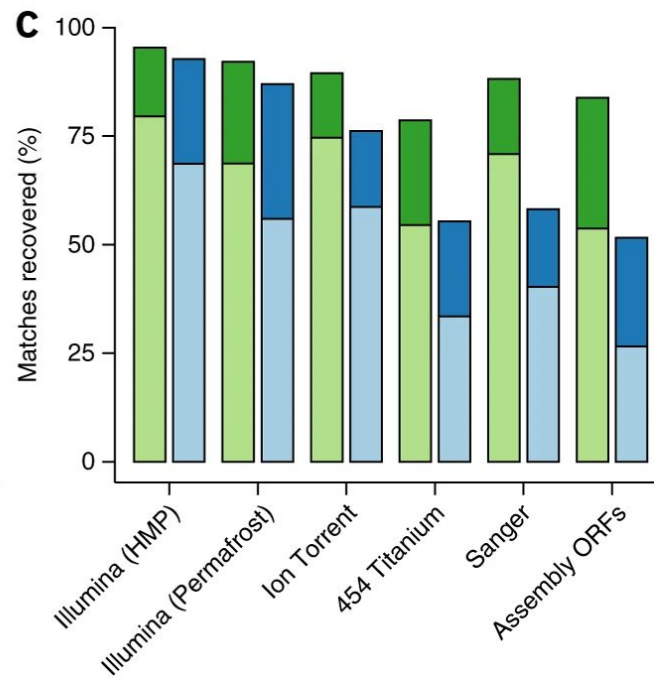
But it costs money!

# 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: faster BLAST with several tricks

- Double indexing: precompute all “seeds” in the database *and* query sequences, compare in lexicographical order (memory cache efficient)
- “Shaped” seeds (similar to discontinuous MEGABLAST, but for proteins)
- Reduced amino acid alphabet!  
[KREDQN] [C] [G] [H] [ILV] [M] [F] [Y] [W] [P] [STA]
- Other stuff

**a****b****c**

# Non-pretty example

---

- 1273 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_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)

???

**STILL NOT FAST ENOUGH!!!  
The Burrows-Wheeler Transform**



# Resequencing



UK

100K

RARE GENETIC VARIANTS IN HEALTH AND DISEASE

**DNA sequencing**



**Reference human assembly**



# BWA: The Burrows-Wheeler Aligner



X = googol

# BWA: The Burrows-Wheeler Aligner

l o \$ o o g g      BWT  
STRING

(6, 3, 0, 5, 2, 4, 1)      SUFFIX  
ARRAY

0	6	\$googo l	
1	3	gol\$go o	Finding 'go':
2	0	googol \$	Finding 'goo':
3	5	l\$goog o	
4	2	ogol\$g o	
5	4	ol\$goo g	
6	1	oogol\$ g	

# Recursive search by adding prefixes

$$\underline{R}(aW) = C(a) + O(a, \underline{R}(W) - 1) + 1$$

$$\bar{R}(aW) = C(a) + O(a, \bar{R}(W))$$

Min, max of word  $W$   
prefixed by character  $a$

Min, max rows that  
have word  $W$  as prefix

Count of non-\$ characters  
smaller than  $a$

# of occurrences of  $a$  in  
BWS up to  $\max(W)$

0	6	\$googo	l
1	3	gol\$go	o
2	0	googol	\$
3	5	l\$goog	o
4	2	ogol\$g	o
5	4	ol\$goo	g
6	1	oogol\$	g

$\min(\$) = \underline{0}$   
 $\max(\$) = 6$

Exact match exists only if  $\min \leq \max$   
for entire searched word (Ferragina and  
Manzini, 2000)

# Recursive search by adding prefixes

googol\$ ORIGINAL STRING

lo\$oogg BWT STRING

ool

0	6	\$googo	l
1	3	gol\$go	o
2	0	googol	\$
3	5	l\$goog	o
4	2	ogol\$g	o
5	4	ol\$goo	g
6	1	oogol\$	g

l

$$\min(l) = C(l) + O(l, \min(\$) - 1) + 1 = 2 + 0 + 1 = 3$$

$$\max(l) = C(l) + O(l, \max(\$)) = 2 + 1 = 3$$

ol

$$\min(ol) = C(o) + O(o, \min(l) - 1) + 1 = 3 + 1 + 1 = 5$$

$$\max(ol) = C(o) + O(o, \max(l)) = 3 + 2 = 5$$

ool

$$\min(ool) = C(o) + O(o, \min(ol) - 1) + 1 = 3 + 3 + 1 = 7$$

$$\max(ool) = C(o) + O(o, \max(ol)) = 3 + 3 = 6$$

$\min(ool) > \max(ool) = ???$



**Why this is awesome:** Sequence reads are effectively searched against different parts of the reference genome at the same time

0	6	\$googo l
1	3	gol\$go o
2	0	googol \$
3	5	l\$goog o
4	2	ogol\$g o
5	4	ol\$goo g
6	1	oogol\$ g

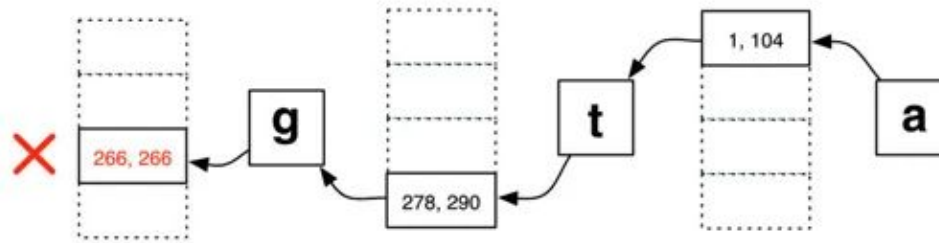
→

→

Also, notice that the formula only makes use of the BWT string – everything else can be forgotten

- **Why this is slightly less awesome:**  
Preprocessing requires many GB of memory
- What about mismatches?

Exact



Searching for "ggta" in a string that lacks "ggta" but has a one-mismatch alignment to "ggtg"

Langmead et al. (2009)  
*Genome Biol*

# BWA refinements

Allowing mismatches: maximum deviation from the search string

- Store searches in a heap to prioritize the lowest mismatches in the search so far

Custom penalties for mismatches, insertion and deletions

Double indexing: BWTs from both ends meet in the middle (avoids massive amounts of futile backtracing)

Memory refinements: store only parts of the BWT and O matrix, calculate the rest on the fly

Program	Single-end			Paired-end		
	Time (s)	Conf (%)	Err (%)	Time (s)	Conf (%)	Err (%)
bowtie-125	1966	88.0	0.07	1701	91.0	0.37
BWA-125	3021	93.0	0.05	3059	97.6	0.04
MAQ-125	17506	92.7	0.08	19388	96.3	0.02
SOAP2-125	555	91.5	0.17	1187	90.8	0.14

One million pairs of 32, 70 and 125 bp reads, respectively, were simulated from the human genome with 0.09% SNP mutation rate, 0.01% indel mutation rate and 2% uniform sequencing base error rate. The insert size of 32 bp reads is drawn from a normal distribution  $N(170,25)$ , and of 70 and 125 bp reads from  $N(500,50)$ . CPU time in seconds on a single core of a 2.5 GHz Xeon E5420 processor (Time), percent confidently mapped reads (Conf) and percent erroneous alignments out of confident mappings (Err) are shown in the table.

SOAP2: somewhere between 300x and 1200x faster than BLAST

**Table 2.** Evaluation on real data

Program	Time (h)	Conf (%)	Paired (%)
Bowtie	5.2	84.4	96.3
BWA	4.0	88.9	98.8
MAQ	94.9	86.1	98.7
SOAP2	3.4	88.3	97.5

The 12.2 million read pairs were mapped to the human genome. CPU time in hours on a single core of a 2.5 GHz Xeon E5420 processor (Time), percent confidently mapped reads (Conf) and percent confident mappings with the mates mapped in the correct orientation and within 300 bp (Paired), are shown in the table.

# Where to try

- BLAST
  - <http://www.ncbi.nlm.nih.gov/BLAST/>
    - Different variants are included in different options
    - MEGABLAST!!! and Discontiguous MEGABLAST!!! are options for BLASTN
  - BLAST+ package:
    - [http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE\\_TYPE=BlastDocs&DOC\\_TYPE=Download](http://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download)
- FASTA
  - <http://www.ebi.ac.uk/fasta33/>
- SSEARCH for Smith-Waterman alignment
  - Included in the FASTA package (<ftp://ftp.hgc.jp/pub/mirror/virginia/fasta/>)
- BWA:
  - <http://bio-bwa.sourceforge.net/>