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

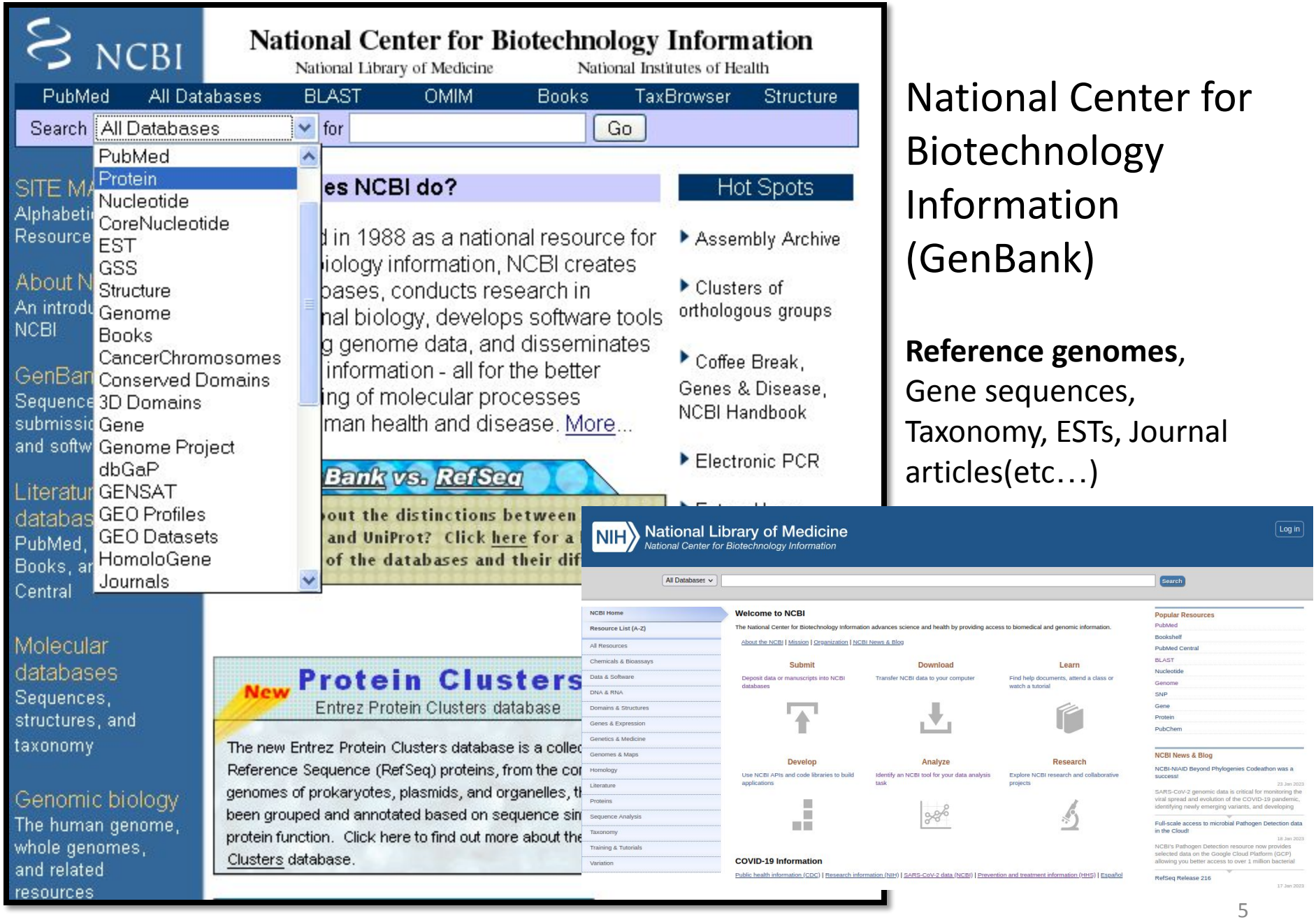
Considerations

Data type (duh!), size and provenance

Modes of access: queries, browsing, APIs

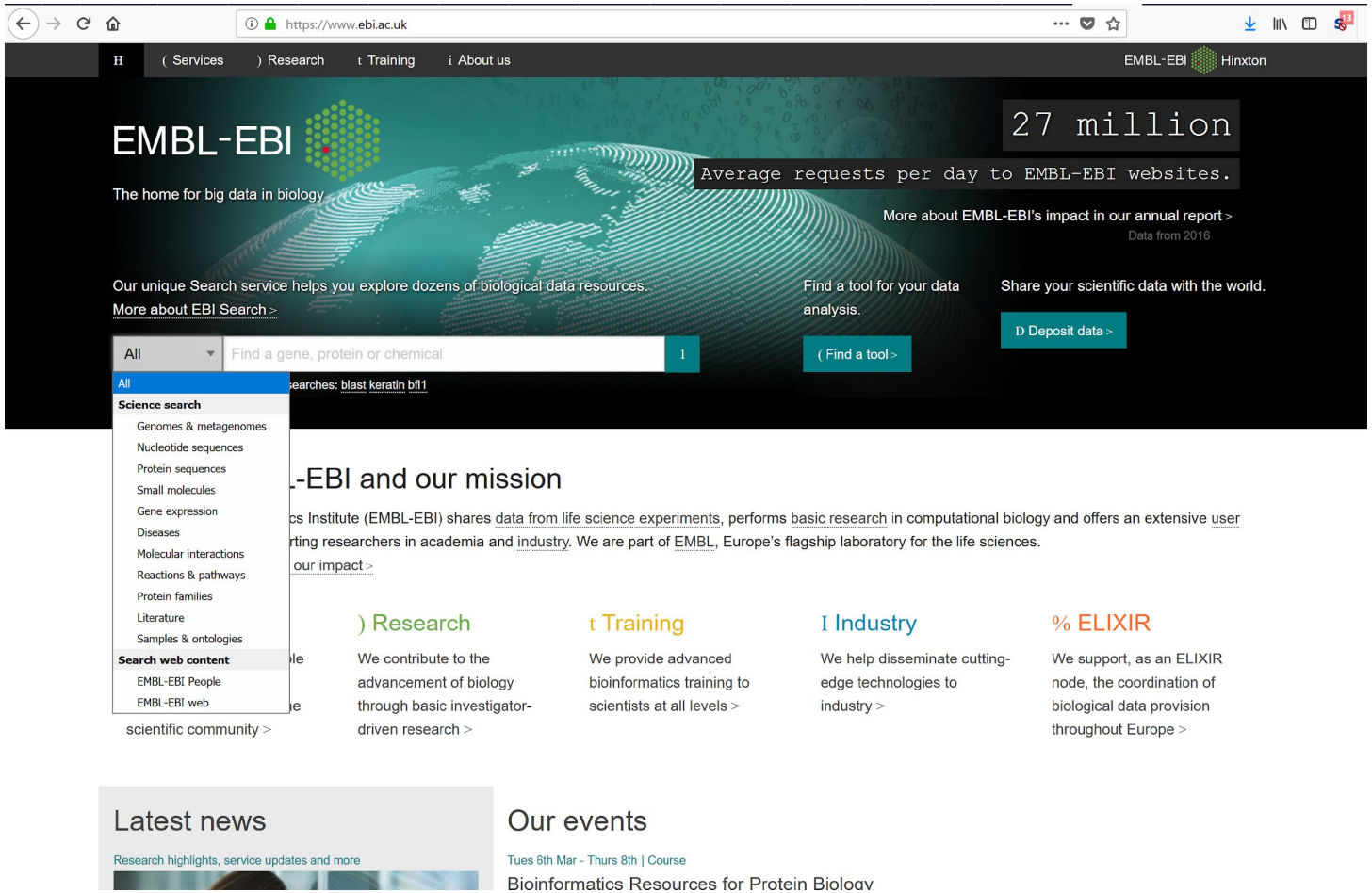
Documentation / stability / support / persistence

Reliability of information



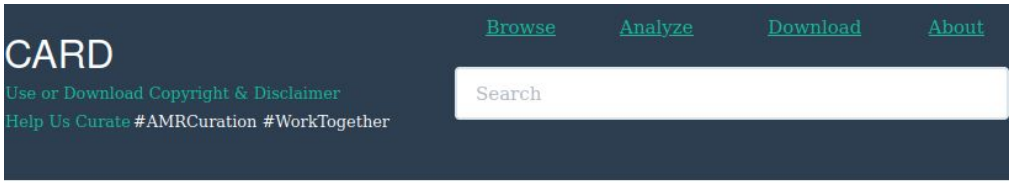
National Center for Biotechnology Information (GenBank)

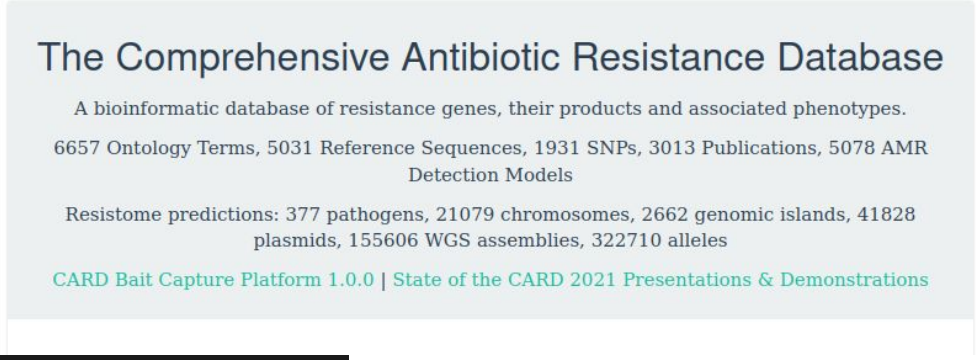
Reference genomes, Gene sequences, Taxonomy, ESTs, Journal articles(etc…)

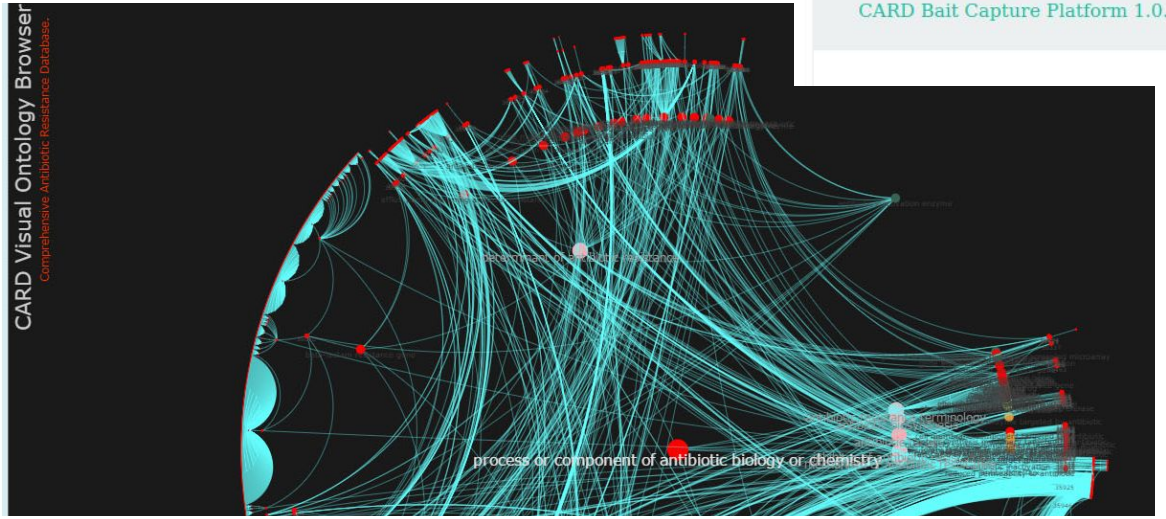


Nucleotides,
Genomes,
Protein function,
Protein-protein interactions

European Molecular Biology Laboratory – European Bioinformatics Institute







CARD
the Comprehensive Antibiotic Resistance Database

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

JGI GOLD GENOMES ONLINE DATABASE

JGI HOME LOG IN

Home | Search | Distribution Graphs | Biogeographical Metadata | Statistics | References | Team | Help | News

Studies	23,258
Biosamples	79,418
Sequencing Projects	79,471
Analysis Projects	65,042

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.

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 2,480

Environmental 9,470

Sequencing Projects

Complete Projects 8,015

Permanent Drafts 33,298

Incomplete Projects 35,607

Targeted Projects 1,565

Analysis Projects

Genome Analysis 48,429

Metagenome 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 Analysis 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 30,927

Projects with Genbank Data

Seq. Projects 42,394

Archaeal Projects 564

Bacterial Projects 35,732

GOLD
Genomes Online Database

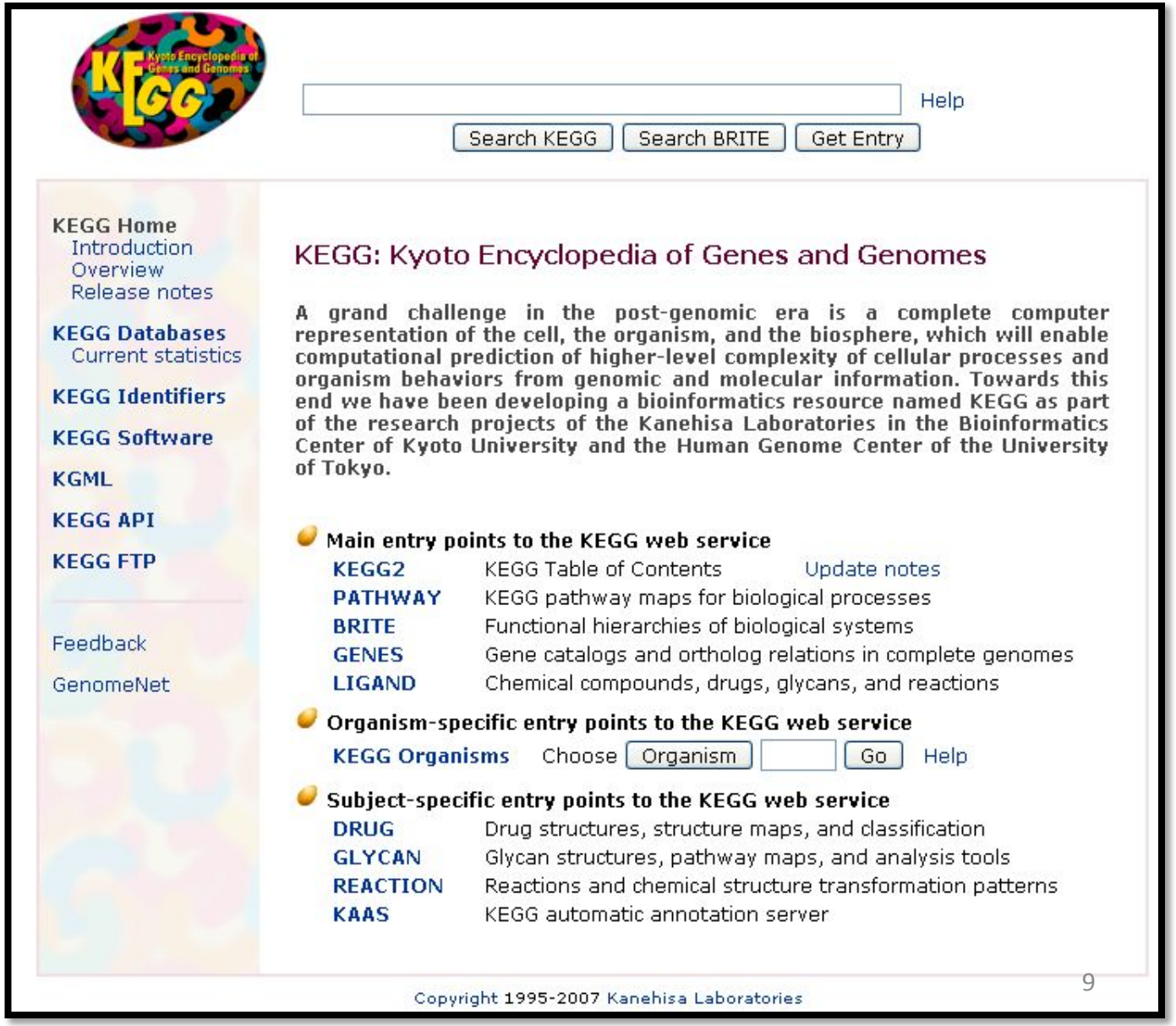
Genome projects

Standards-compliant metadata

Kyoto Encyclopedia of Genes and Genomes

Genomes
Orthology information
Protein functions
Biochemical pathways

Limited access now
(#%#!)



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

206293625 loci, 249722163594 bases, from 206293625 reported sequences

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

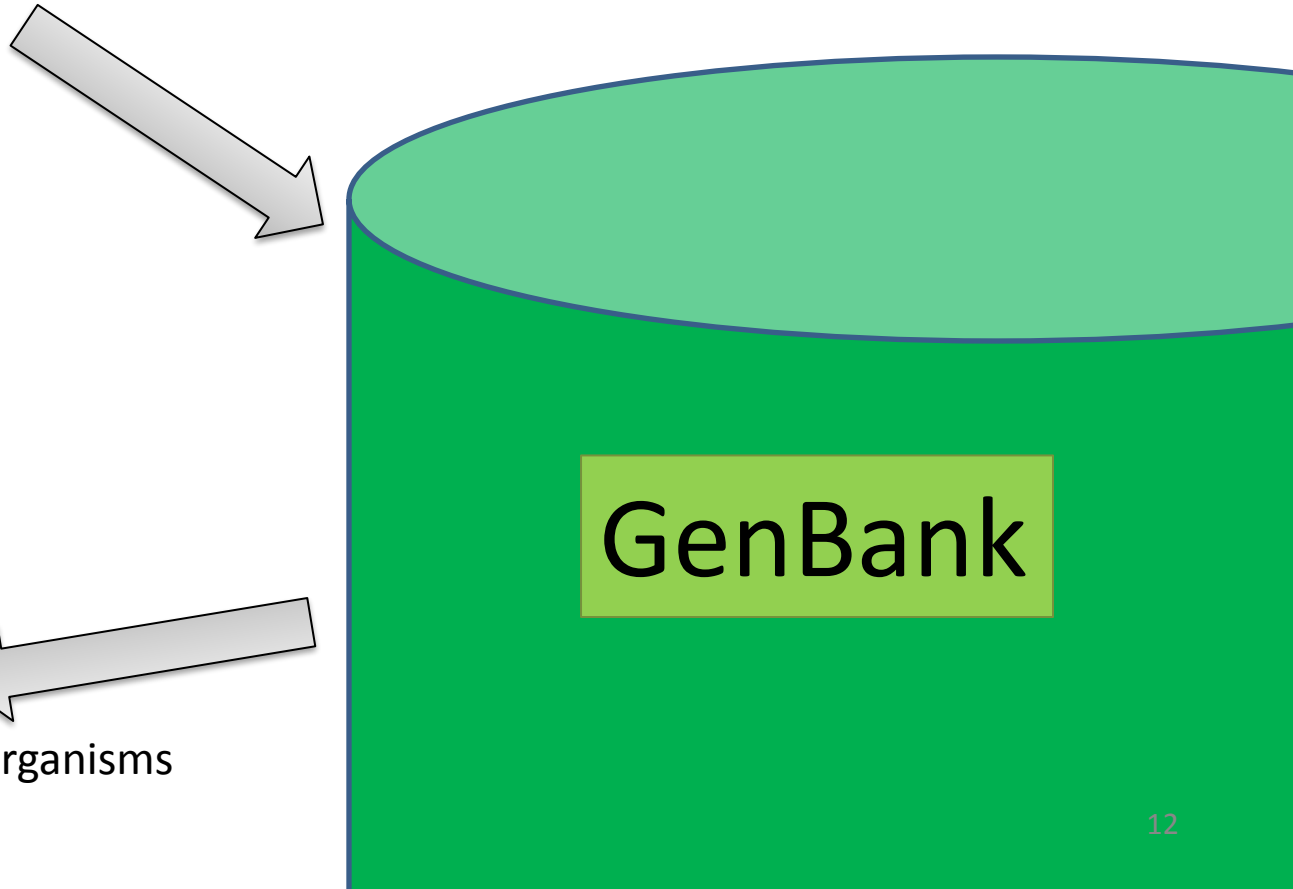
Uncompressed, the Release 158 flatfiles are 252 Gigabytes and the ASN.1 version is about 217 Gigabytes. The data can also be downloaded at a mirror site:

bio-mirror.net/biomirror/genbank

Release 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



<u>Homologous</u> 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

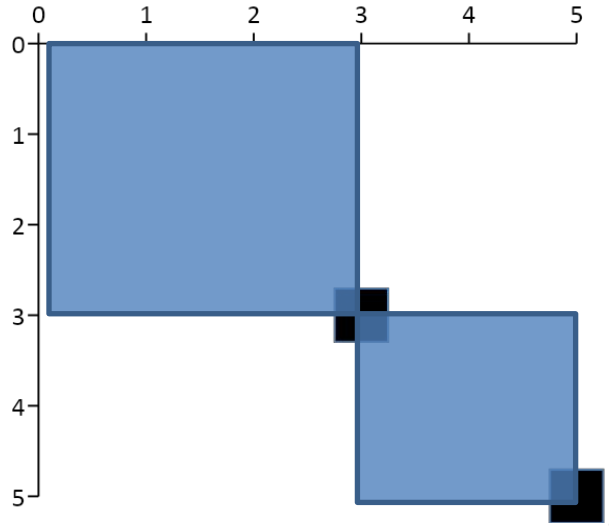
DEAKP

Score each pair of residues – consider every possible alignment

	D	E	A	K	P
N					
Q					
A					
R					
P					

OR

Require an exact match of length *L* to "seed" the alignment



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

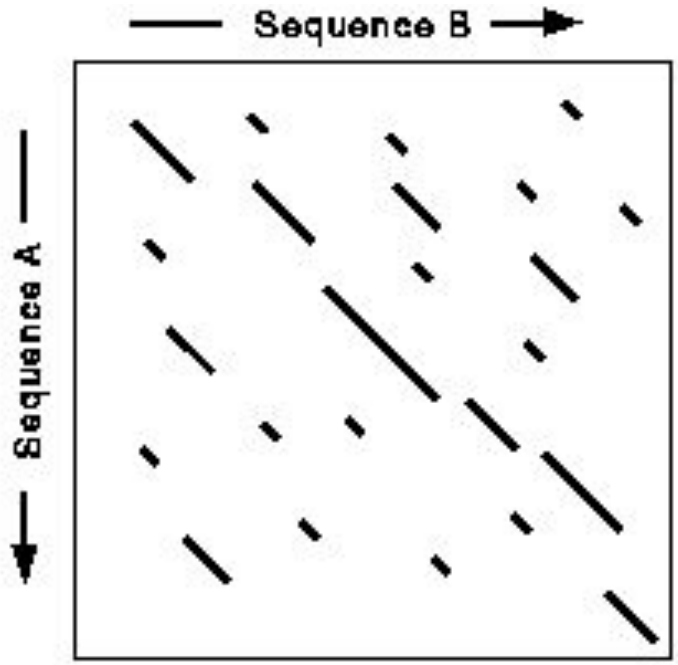
DQATS

AT	3
DQ	1
QA	2
TS	4

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

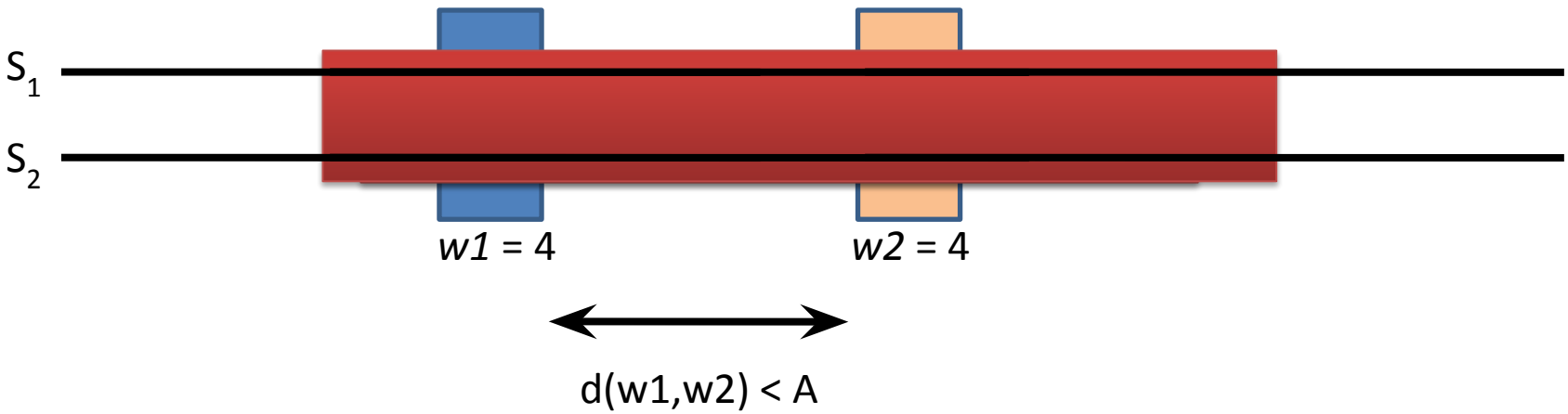
- Exact matches are great and all, but they're not perfect
- Find maximal <u>high-scoring pairs</u>: for a query / database sequence pair, find the best region(s) where:
  - The local alignment score (no gaps allowed!) is above a threshold *S*, and
  - The score cannot be increased by extending or trimming 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. <u>If word pairs are found</u>, use these to seed the high-scoring pairs
3. <u>If HSPs are found</u>, perform dynamic programming anchored with HSPs to complete the alignment

Extending to high scoring pairs

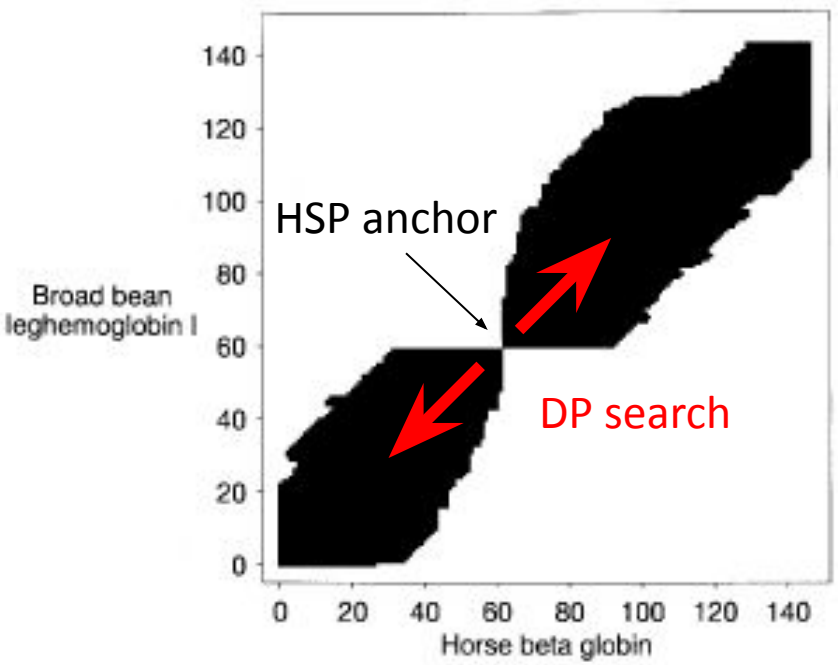


Try to extend matches, stop trying when a move drops the score below a given threshold

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

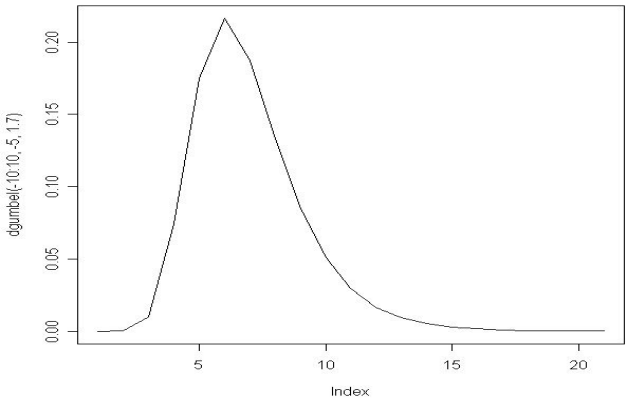

HSP anchor
DP search
Broad bean leghemoglobin I
Horse beta globin
0
20
40
60
80
100
120
140

Local alignment significance

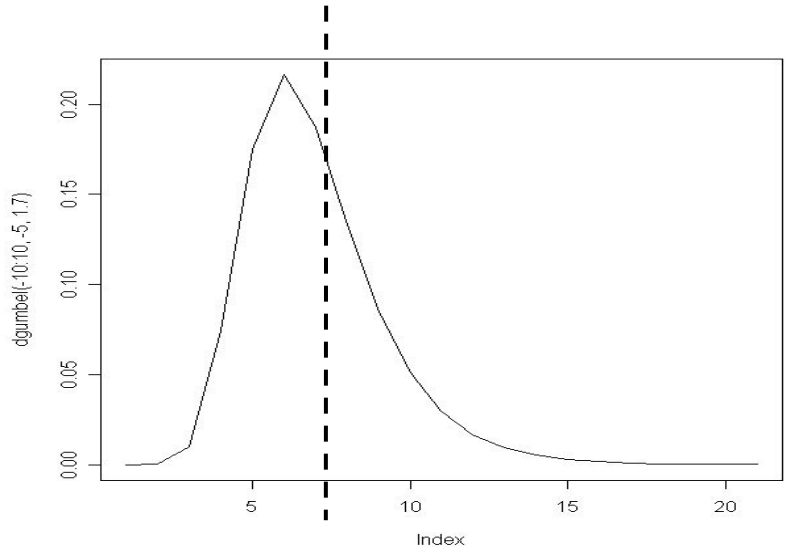
- How are alignment scores distributed?

- 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**

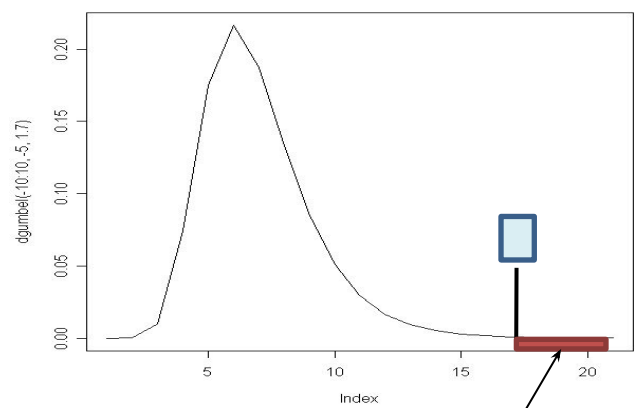


- The expected (=mean) score between a pair of random sequences is the mean of an extreme value distribution
- Given a scoring matrix (such as PAM250) and a set of amino acid frequencies, we can compute the parameters λ and K that define this distribution

Karlin-Altschul statistics

Score from EVD

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



P(S >)

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

Standard Protein BLAST

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 Browse... No file selected.

Job Title

Enter a descriptive title for your BLAST search

Align two or more sequences

Query

Choose Search Set

Database: Non-redundant protein sequences (nr)

Organism (Optional): Enter organism name or id--completions will be suggested — exclude — Add organism

Enter 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

Database

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

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

Match parameters

Scoring Parameters

Matrix: BLOSUM62

Gap Costs: Existence: 11 Extension: 1

Compositional adjustments: Conditional compositional score matrix adjustment

Scoring

Filters and Masking

Filter: Low complexity regions

Mask: Mask for lookup table only — Mask lower case letters

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

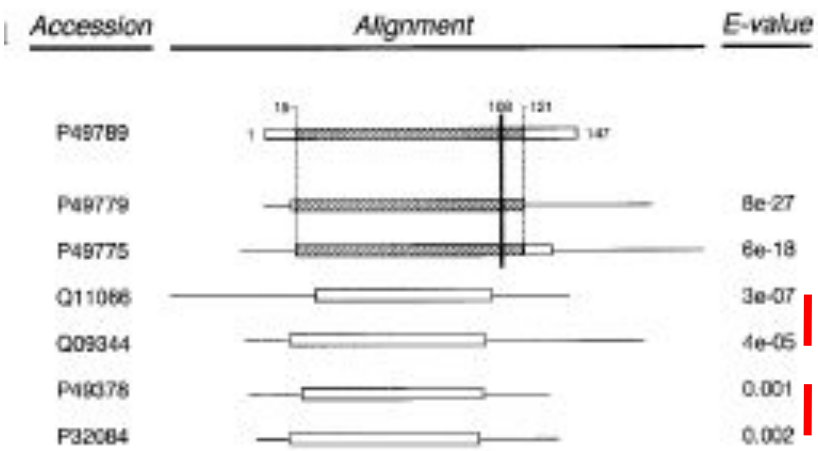
- Iterate query – Position-specific scoring matrix (PSSM) procedure

Step 1

Run BLAST!

Step 2

- Collapse significant local alignments into a multiple alignment



Gaps!

E-value not good enough!

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

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

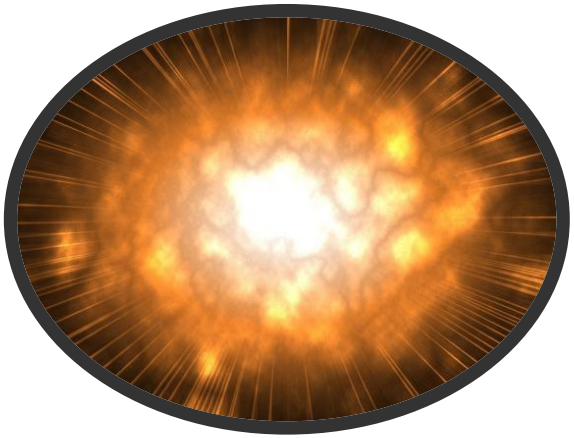
and PatternHunter

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

MEGABLAST!!!

Nucleotide seed length

BLASTN!: 11111111111



MEGABLAST!!!:
1111111111111111111111111111

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

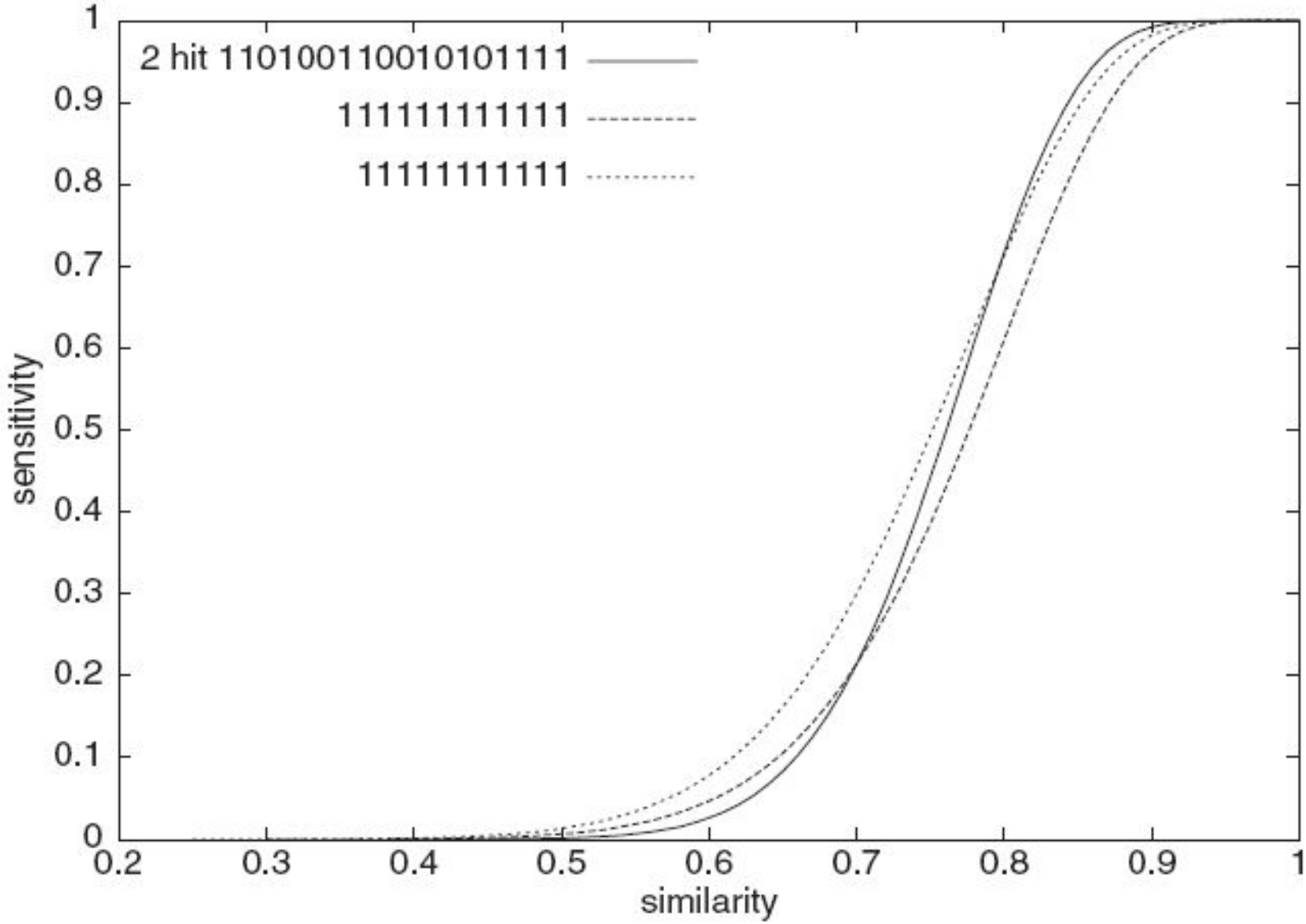
GC**ACGATCCGAAA**ATCC

Discontinuous Words

Search for words defined by a ‘model’:

```
Model: 111010010100110111
       AAACGAACAGAGAGTTTC
       AAATGATCCGAAAGCTTC
```

Similar accuracy



PatternHunter: 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!

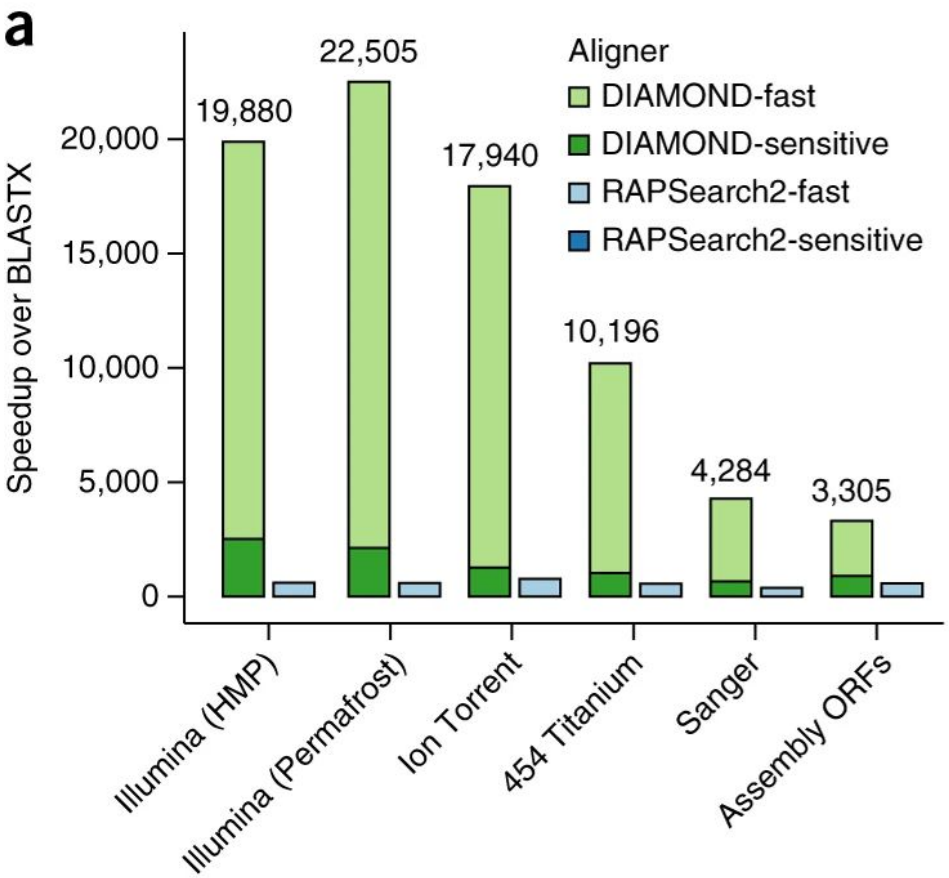
DIAMOND: faster BLAST with several tricks

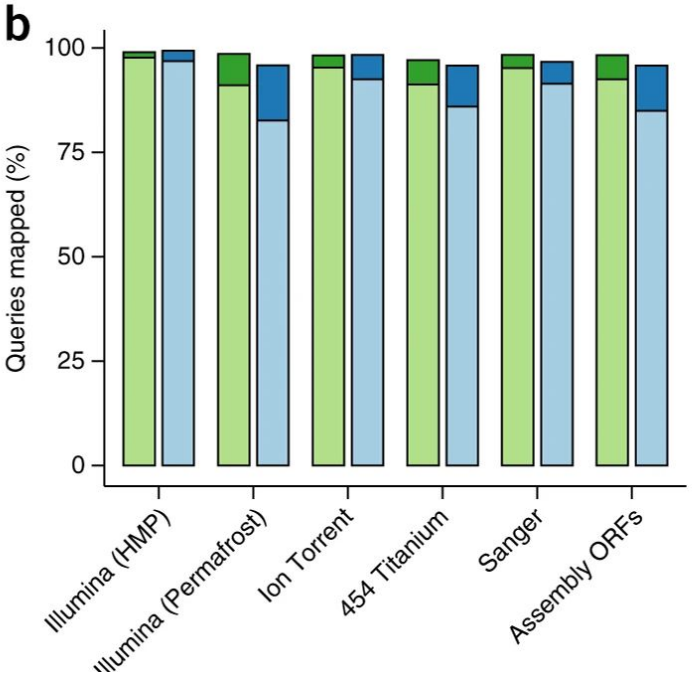
- Double indexing: precompute all “seeds” in the database *and* query sequences, compare in lexicographical order (memory cache efficient)
- “Shaped” seeds (similar to discontiguous MEGABLAST, but for proteins)
- Reduced amino acid alphabet!

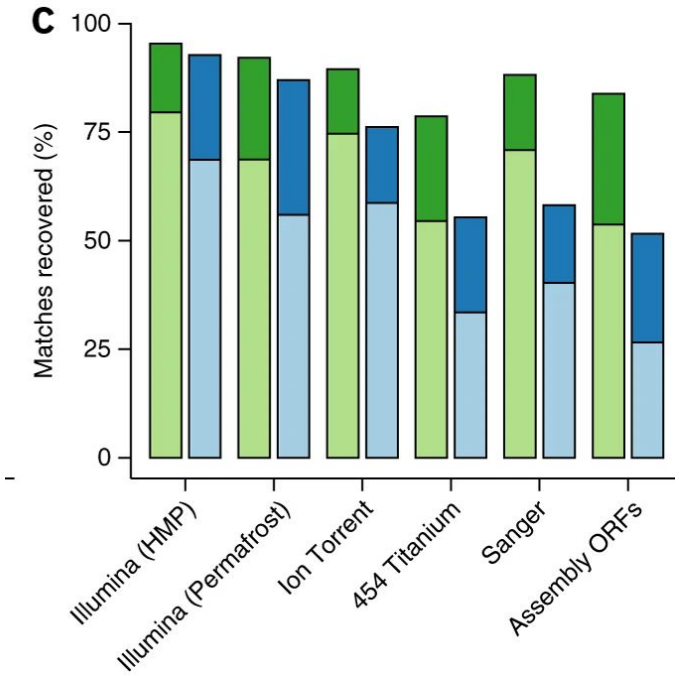
  [KREDQN] [C] [G] [H] [ILV] [M] [F] [Y] [W] [P] [STA]

- Other stuff

Buchfink et al. (2015) *Nat Meth*

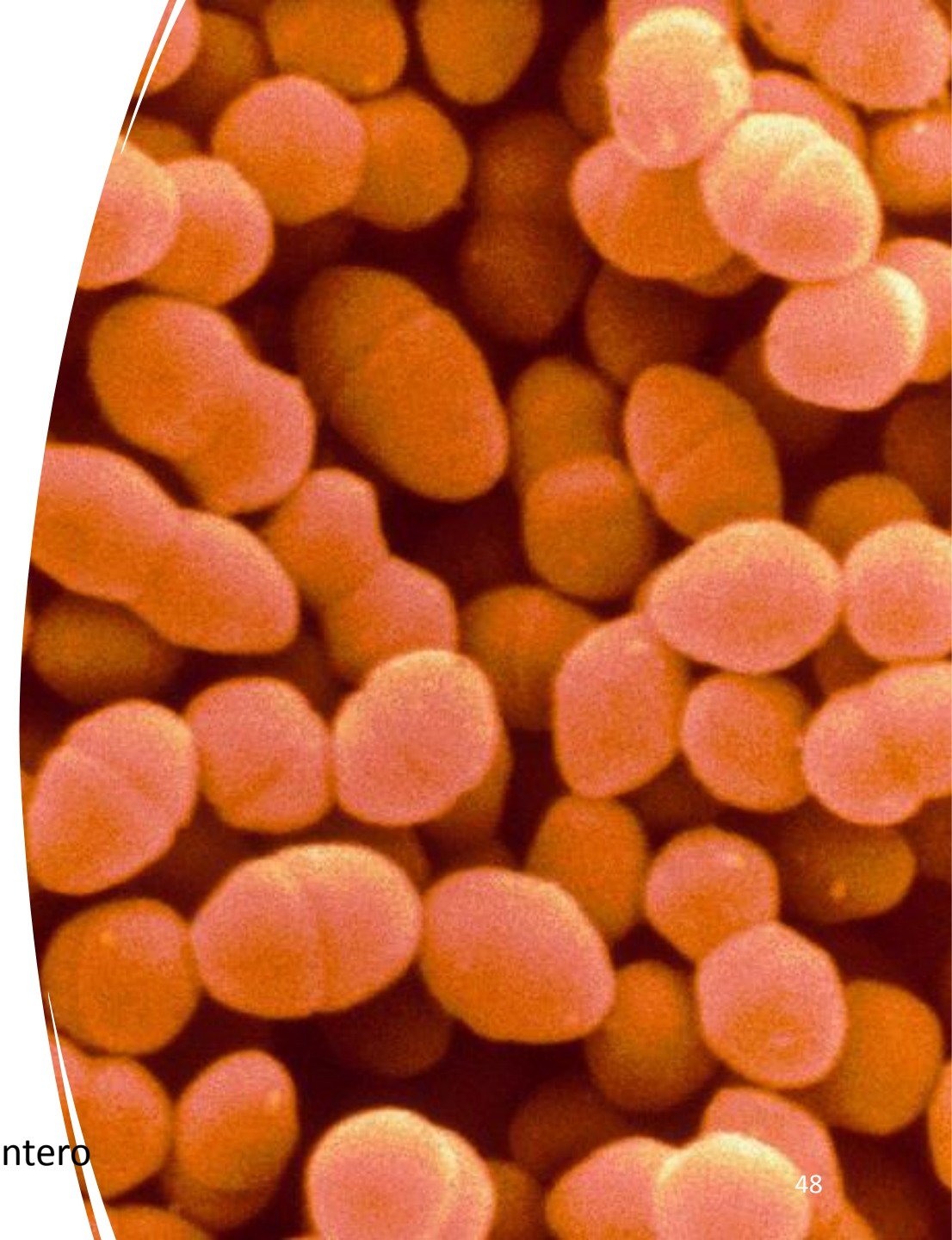

a
Speedup over BLASTX
20,000
15,000
10,000
5,000
0
19,880
22,505
17,940
10,196
4,284
3,305
Aligner
DIAMOND-fast
DIAMOND-sensitive
RAPSearch2-fast
RAPSearch2-sensitive
Illumina (HMP)
Illumina (Permafrost)
Ion Torrent
454 Titanium
Sanger
Assembly ORFs


b
Queries mapped (%)
100
75
50
25
0
Illumina (HMP)
Illumina (Permafrost)
Ion Torrent
454 Titanium
Sanger
Assembly ORFs


c
Matches recovered (%)
100
75
50
25
0
Illumina (HMP)
Illumina (Permafrost)
Ion Torrent
454 Titanium
Sanger
Assembly ORFs

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?



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)

???

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

Li and Durbin, *Bioinformatics* 2009

BWA:
The Burrows-Wheeler Aligner

lo$oogg **BWT STRING**

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

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

Finding ‘go’:

Finding ‘goo’:

Li and Durbin, *Bioinformatics* 2009

Recursive search by adding prefixes

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

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

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

Count of non-$ characters smaller than *a*

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

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

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($) = **0**
max($) = 6

Exact match exists only if min <= max for entire searched word (Ferragina and Manzini, 2000)

Recursive search by adding prefixes

ool

googol$ **ORIGINAL STRING**

lo$oogg **BWT STRING**

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) = ???

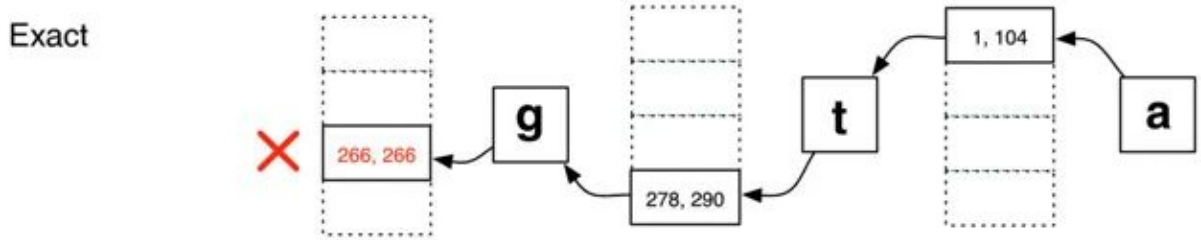
Why this is awesome: Sequence reads are effectively searched against different parts of the reference genome <u>at the same time</u>

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?



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

	Single-end			Paired-end		
Program	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

- 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/